

NORTHWESTERN UNIVERSITY

Understanding and Mitigating New Risks in Location-Aware Technologies

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Computer Science

By

Allen Yilun Lin

EVANSTON, ILLINOIS

June 2021

© Copyright by Allen Yilun Lin 2021

All Rights Reserved

ABSTRACT

Understanding and Mitigating New Risks in Location-Aware Technologies

Allen Yilun Lin

Location-aware technologies, such as personal navigation applications, location-based AR games, and artificial intelligence systems that learn from data about places, increasingly mediate our understanding of and interactions with the world. However, a number of risks associated with location-aware technologies have emerged, jeopardizing the welfare of its users. This dissertation seeks to understand and mitigate two such emerging risks associated with location-aware technologies - safety and geographic biases. The first part of this dissertation focuses on understanding and mitigating the safety issues in one of the most popular location-aware technologies - personal navigation systems. We analyzed catastrophic incidents that involved the use of personal navigation applications, identified technological causes that contributed to these incidents, and designed solutions to mitigate the most prominent technological causes. The second part of this dissertation focuses on mitigating the geographic biases in two types of location-aware technologies that leverage data-driven artificial intelligence - location-aware technologies based on computer vision and location-aware technologies based on knowledge base.

Acknowledgements

I owe appreciations to many people who helped me tremendously throughout this journey. First, I wanted to thank my advisor Dr. Brent Hecht. I have always told my friends that, as a PhD student I could not have imagined a better advisor than you. Academically, you taught me the way of scientific inquiry, lent me your wisdom and encouragement when I hit the wall, and supported me to grow on every skill that I need (including English writing and communications). Personally, you demonstrated how an extremely busy professor could still be patient, considerate, and caring to their students. Working with you is the wisest academic choice I have made!

My committee - Dr. Darren Gergle and Dr. Nell O'Rourke. Thank you for your incredible support on this thesis and throughout my PhD study. Thank you for being so patient during my last year of dissertation writing while still giving timely and incredibly helpful comments. Special thanks to Nell for the insightful questions and feedback that made me reflect on what I really wanted to communicate in this thesis. Special thanks to Darren for pushing me to think harder on the big picture and overall narrative, which tremendously strengthened this thesis, and for hosting Isaac and me in your lab during our first days at Northwestern so that we had the warmth of home.

Thank you to my mentors at Microsoft Research and Google for giving me two amazing internships. Justin and Scott, thanks for so many insightful discussions and continued support on finishing our paper after my internship. Jan and Inna, thank you for hosting

me for a sunny summer in the Bay Area after a drizzling winter in Seattle. I am grateful for the collaborations and extremely happy that we can keep collaborating!

Thank you to my fellows in PSA Research, to Isaac, Nick, Hanlin, and Jack - you have probably seen too many practice talks from me in the last year and I owe you many thanks for your constructive feedback. Thank you to the wider HCI community at Northwestern for welcoming me into the group and providing a place to learn, exchange ideas, and grow together.

Thank you to GroupLens lab at University of Minnesota for supporting the start of my computer science research journey. To Dr. Loren Terveen, Dr. Joseph Konstan, Dr. Lana Yarosh, Dr. Haiyi Zhu for providing a lovely and flexible place for us to pursue our research interests. To Bowen, Qian, Steven, Jake, Andrew, Morten, Toby, Daniel for many serious discussions, fun chats, and lunch specials together at Little Szechuan Restaurant. I will never forget the year I spent in GroupLens.

Thank you to my family in Wuhan - I couldn't have asked for a better one. Thanks for your unconditional love and support during these years of overseas study.

My wife, Charlene, thank you for being my cheerleader and comforter, sharing both hard times and good times during this journey. Thanks you for making meals while I was on paper deadlines and offering final feedback on the night before my important presentations. Nothing could have been accomplished without you.

Above all, thank you Jesus Christ for leading me on and throughout this journey. I am grateful that You used these years to let me experience more of Your faithfulness, kindness, and love.

Table of Contents

ABSTRACT	3
Acknowledgements	4
Table of Contents	6
List of Tables	10
List of Figures	13
Chapter 1. Introduction	16
1.1. Motivation	16
1.2. Background	19
1.3. Organization of the Thesis and Summary of the Chapters	28
Part 1. Safety Issues in Personal Navigation Technologies	33
Chapter 2. Understanding the Catastrophic Incidents Associated with Personal Navigation Technologies	34
2.1. Introduction	34
2.2. Related Work	38
2.3. Methods	40
2.4. Results	47

	7
2.5. Implications for Research and Design	57
2.6. Discussion	62
2.7. Conclusion	65
Chapter 3. Mitigating the Risks of Personal Navigation Systems by Adding Road	
Safety Awareness	66
3.1. Transition	66
3.2. Introduction	67
3.3. Related Work	70
3.4. Defining Road Safety for Navigation Systems	71
3.5. Predicting Road Safety	73
3.6. Discussion	85
3.7. Conclusion	88
 Part 2. Geographic Inequalities in Location-Aware Artificial Intelligence	
Systems	90
Chapter 4. Mitigating Geographic Inequalities in Computer Vision-Based	
Location-Aware AI Technologies	95
4.1. Motivation	95
4.2. Introduction	99
4.3. Related Work	104
4.4. Producing ImageNet-Compatible Tags	108
4.5. Validating that Wikimedia Commons Images Can Reduce Geographic Biases	118
4.6. Additional Benefits: Increasing the Size and Breadth of ImageNet	129

4.7. Discussion	134
4.8. Future Work and Limitations	137
4.9. Conclusion	138
Chapter 5. Addressing the Geographic Biases in Location-Aware Technologies based on Commonsense Knowledge Base	140
5.1. Introduction	140
5.2. Background	146
5.3. Datasets Development	147
5.4. Modeling	157
5.5. Discussion	168
5.6. Conclusion	172
Chapter 6. Supporting Projects	173
6.1. Project 1: Mining Wikimedia Commons to Enrich News Articles	173
6.2. Project 2: Understanding the Geographic Inequalities in Pokemon Go	174
Chapter 7. Future Work	176
7.1. For the Safety Issues in Location-Aware Computing	176
7.2. Geographic Inequalities in Location-Aware Artificial Intelligence Technologies	178
Chapter 8. Conclusion	181
References	184
Appendix A. Detailed Classification Accuracies in Figure 5.4	212

Appendix B. The Mathematical Relationship Between Likelihood and the Loss Function	215
---	-----

List of Tables

2.1	Distribution of the <i>Seriousness</i> of these 158 incidents.	48
2.2	Distribution of the <i>Incident Types</i> of these 158 incidents.	49
2.3	Distribution of <i>Technological Cause</i> . Note: The # does not add up to 158 because coders did not enter a code when there was not enough information in given news story to make a certain type of assessment.	53
3.1	Features for our road safety classifier	77
3.2	The summary statistics for road safety class labels in four ground truth datasets. Our statistics show that two road safety classes are highly imbalanced. (HR: Croatia; GR: Greece; SL: Slovenia; NL: The Netherlands)	78
3.3	The missing values rates for OSM features. Features that are not shown in the table do not have missing values problem. Our statistics show that the missing values rates differ among different countries. (HR: Croatia; GR: Greece; SL: Slovenia; NL: The Netherlands)	78
3.4	Our road safety classifier's performance on the Croatia dataset. The performance of the baseline approach is in the bracket. (HR: Croatia; GR: Greece; SL: Slovenia; NL: The Netherlands)	83

3.5	The percentage of unsafe routes among randomly sampled routes in Croatia by types of area and types of route options. The criteria for unsafe routes specify the minimum percentage of high-risk road segments a route needs to include for this route to be considered an unsafe route. For example, “>10%” means that routes contain at least 10% of high-risk road segments are unsafe routes.	84
4.1	The ten most-represented countries in the Commons.	121
4.2	Countries that are overrepresented in ImageNet are less dominant in the Commons. ImageNet’s country representation are from Shankar et la. [258].	121
4.3	Mean likelihood of the pre-training models predicting a ground truth bridegroom image as a bridegroom image. The mean likelihood is averaged across all data in the test sets.	127
5.1	The percent of articles and page views associated with potential sub-articles	152
5.2	Feature importance averaged across all languages and all definitions of sub-articles on the High-Interest dataset. Feature importance is computed using a Random Forest model.	166
5.3	The impact of applying our model on the top 1000 most-viewed articles in English Wikipedia.	168

A.1	Detailed classification accuracies for <i>High-Interest</i> dataset (Figure 5.4 (a))	212
A.2	Detailed classification accuracies for <i>Random</i> dataset (Figure 5.4 (b))	213
A.3	Detailed classification accuracies for <i>Ad-Hoc</i> dataset (Figure 5.4 (c))	213
A.4	Detailed overall classification accuracies (Figure 5.4 (d))	214

List of Figures

2.1	A Google Street View image depicting the complex geography of the location of the incident in article 146.	55
3.1	Without road safety information, Google Maps recommended a faster but more dangerous route (highlighted). By contrast, the alternative route (grey) is slower but safer. Images are from Google Maps and Google StreetView.	68
3.2	The propagation of geographic inequalities from the large-scale training dataset to the AI-based location-aware technologies	93
4.1	The propagation of geographic inequalities from the large-scale ImageNet dataset to the computer-vision-based location-aware technologies	96
4.2	An image in the Commons and its corresponding category labels (indicated by red dotted lines)	109
4.3	Flow chart of the YAGO3 Mapper	113
4.4	Flow chart of the Content-Based Tags Filter	115
4.5	Geolocatable images in the Commons. Log-scaled classification, with adjustment for small numbers of images.	120

		14
4.6	The technique for summarizing the number of images	130
4.7	The reverse cumulative histogram of the # images the Commons could provide to the difficult classes in ILSVRC. The 0.21 (y-value) at “1200” tick on x-axis means that the Commons could provide 21% of difficult classes at least 1200 images, which double the size of training data for these classes.	131
4.8	The reverse cumulative histogram of the # images the Commons could provide to the previously-uncovered ImageNet classes. The red dotted line is the average number of images per previously-uncovered class (650 images).	134
5.1	The propagation of geographic inequalities from Wikipedia to the knowledge base location-aware technologies	141
5.2	An example of the ambiguity inherent in sub-article indicators, in this case the “ <code>{{main article}}</code> ” template in the English Wikipedia. Links in red indicate potential sub-article relationships	149
5.3	The “See also” section of the English article about the Canadian territory of Nunavut. Some links here are clearly sub-articles, while others are more distantly related to the concept of Nunavut (e.g. “Arctic policy of Canada”).	150
5.4	Classification accuracies across datasets, language editions and different thresholds of sub-article ratings. Each colored vertical bar	

shows the best accuracies among the machine learning algorithms considered, and the black line indicates baseline performance.

CHAPTER 1

Introduction

1.1. Motivation

Location-aware technologies, which encompass computing systems and devices that leverage location data, are bringing significant benefits to both individuals and societies. For many individual users, location-aware technologies are indispensable for them to understand and navigate the physical world, answering simple, yet fundamental questions such as “where am I” and “what is there”. For an increasingly data-driven society, the expected economic values of the location-aware technologies are enormous. For example, studies have estimated that more than 60% of user-generated digital information in our society are associated with spatial information (e.g., [110, 94]), opening a billion-dollar market [197] for location-aware technologies to operate in.

Despite these immense benefits, location-aware technologies can expose their users to a variety of risks. Previous research on the risks of location-aware technologies has almost exclusively focused on privacy issues [260] as the most significant user concern [50, 90]. Through decades of research, researchers have gained a comprehensive understanding of different types of location privacy attacks (e.g., [168, 32, 192]) and have mitigated these attacks through developing well-known countermeasures (e.g., [33, 109, 15]).

However, as location-aware technologies evolved, new risks beyond the location privacy. These new risks brought previously unseen negative effects to both individual users and societies and called for research efforts to understand and mitigate them.

One type of new risks is the risks to users' physical safety. Compared to other computing technologies, location-aware technologies have a unique power to influence where users would physically be, such as the restaurant to go, the park to visit, or the route to drive on. As such, if location-aware technologies are not carefully designed to ensure the safety of the places they recommend, users who take these recommendations might throw themselves into dangerous, sometimes even deadly, environments. Sadly, such catastrophes have already occurred. On one of the most popular location-aware technologies - personal navigation systems, users who followed the directions have been led to go directly into the Atlantic Ocean [115], to get stranded in Australian outback [195], and to drive off the cliff in England [240].

Another type of new risk is the geographic inequalities on the performance of the location-aware technologies. While these geographic biases can be produced through a variety of mechanisms, this dissertation is mostly concerned with the geographic biases rooted in the data that power the location-aware technologies. The presence of digital divide resulted in uneven production of digital content across different socioeconomic classes. [287, 283, 55]. Geographically, higher quantity and quality digital content are produced for locations with privileged socioeconomic status [278, 106, 105]. Data-driven location-aware technologies rely on these digital content about places to learn knowledge and offer services. As such, geographically biases in the digital content are inherited by

the location-aware technologies, contributing to differential service qualities for users from different areas, which in turn reinforces existing socioeconomic inequalities.

The goal of this dissertation is to understand and mitigate the risks of safety and geographic inequalities in location-aware technologies. The first part of this dissertation will focus on understanding and mitigating the safety issues related to one of the most widely used location-aware computing technologies - personal navigation systems. We employed a novel methodology to collect catastrophic incidents associated with the use of personal navigation applications, analyzed the major themes in these incidents, and identify the contributing technological causes. To mitigate one of the major technological causes we identified - that personal navigation technologies usually function without the critical knowledge of road safety information, we built a low-cost machine-learning based system to support adding large-scale road safety information for the use of personal navigation systems. The second part of this dissertation concentrates on understanding and addressing the geographic biases in the location-aware technologies that leverage data-driven artificial intelligence. Specifically, we focused on two types of such technologies - location-aware technologies that are based on computer-vision techniques and location-aware technologies that are based on knowledge base techniques.

The rest of this chapter is structured as follows. First, we situate the topics of safety and geographic biases broadly in the computing research and highlight how these background literature relate to the context of location-aware technologies. We then summarize the overall structure of this dissertation and offer a high-level summary for each chapter.

1.2. Background

Below, we summarize two areas of work in computing research that provide background and context for this thesis.

1.2.1. Safety Concerns of Computing Technologies

Computing machinery forms a critical backbone of our society today. As such, how to ensure the safety of the people who operate, or are subject to, these computing devices becomes a critical focus of the computing research community. Different from the concept of *security*, which puts primary concerns on protecting the information in computing technologies [298], the concept of *safety* requires that computing devices are "free from accidents or losses" [183], thus putting more emphases on human and society. Most of the existing research about the safety of computing systems centered around the so-called safety-critical systems. Also known as life-critical systems, these computing systems are defined as "systems whose failure could result in loss of life, significant property damage, or damage to the environment" [160].

1.2.1.1. Traditional Safety-Critical Computing Systems. Traditional safety-critical computing systems are primarily developed to help humans control complex machinery in high-stake industries. In aviation industry, fly-by-wire computing systems have been developed to reduce the workload of pilots. In the control rooms of nuclear power plants, computer programs monitor the temperature and pressure of the reactors' cooler systems through digital sensors and issue alarms when changes happen. In medical industry, computer-controlled robotic devices help doctors perform gastrointestinal surgery [139], coronary artery bypass [155], and minimally invasive spine surgery [263].

To understand the safety concerns of these high-stake computing systems, a common approach is to gather as much as possible information about their failures. To this end, data loggers or recorders have been widely deployed on these traditional safety-critical computers on aircraft, trains, and power plant control room to reliably record what happens before, during, and after accidents. More recently, surveillance cameras have been installed to augment these data recorders with visual and audio descriptions. This information has been critical for the postmortems of a number of high-profile disasters involving computing systems. For example, in the investigations of the recent two Boeing 737 MAX crashes, black box data, including both the conversation in the cockpit and the instrument readings, helped investigators dissect the two fatal accidents [152].

Researchers who studied these disastrous failures usually found that the root cause is a combination of human errors and technical failures. For example, in the Boeing 737 Max case above, four factors, including "poor documentation, rushed release, delayed software updates, and human out of the loop" [152], have been identified as causes to the tragedy. Another example that is more familiar to the computing community is the Three Mile Island accident in 1979, in which a reactor in a nuclear power plant partially melted down and brought wide and significant environmental and health consequences to the local community. While the technical failure that led to deficiency of coolant is the direct cause, the postmortem also pointed to the human negligence of the initial warnings due to the ambiguous indicators on the control panel [181].

Accordingly, the general approach to mitigate safety concerns of these safety-critical computing systems also involve two aspects - increasing the reliability of the computing systems and improving how these systems interact with their operators. Work to improve

the dependability of computing systems is usually conducted by researchers in software engineering and aims to enhance the fault avoidance, fault tolerance, fault removal, and fault forecasting of software systems [178]. One typical approach to achieve such goals is formal methods [40], which involves defining unambiguous (usually mathematical) requirements for computer systems and then relentlessly verify that systems adhere to these standards. Human-computer interaction researchers works on the improvement of user interface to reduce human errors during system operations [38]. Example work on this aspect includes developing dependable touch screens for cockpits [228], inventing mechanisms to support multiple displays in control rooms [272], and engineering new simulators for evaluating the safety of different arrangement of user interfaces [273].

1.2.1.2. Ubiquitous Safety-Critical Computing Systems. While traditional safety-critical computing systems are only used in limited high-stake scenarios, modern safety-critical computing systems have become ubiquitous in everyday life. Early in 2002, a highly-cited paper about safety-critical systems published on the International Conference on Software Engineering [160] predicted that the number and types of safety-critical computing systems will increase significantly as computers are more embedded in our society. This is precisely what happens almost 20 years later. Despite never called so, many personal computing technologies we used daily, such as navigation systems or augmented reality devices, fit squarely under the definition of safety-critical computing systems - “systems whose failure could result in loss of life, significant property damage, or damage to the environment” [160].

For these ubiquitous safety-critical technologies, researchers need new data collection method to gather information about their safety concerns. Unlike traditional safety-critical computing systems which are usually equipped with specialized recording devices, ubiquitous safety-critical technologies are limited by their relatively low cost and privacy concerns and thus can not record detailed descriptions of their user interactions. To address this challenge, researchers usually employ lab studies or ethnographic studies. In either case, researchers observe how participants use the safety-critical technologies and identify scenarios where it is presents a safety concern (e.g. [44].) For example, to study the safety concerns of driving navigation systems, researchers have conducted ethnographic studies where they sat in the car with the drivers, recorded how drivers interacted with the technologies, and recorded assorted safety problems [44]. Other researchers have leveraged lab studies where they observed participants using driving simulators and identified when and for how long user are distracted by navigation systems (e.g. [77, 169, 294]). Similarly, to study the safety issues of virtual or augmented reality devices, researchers have asked participants to engage with VR devices and monitored their physiological attributes identify potential health risks (e.g., [176, 46, 58]).

However, one major limitation of the lab and ethnographic studies is that the safety concerns are usually observed under normal usages, making it difficult to establish their strong causal relationship with disastrous failures. One main reason for this limitation is that relative to the number of experiments researchers could run, disastrous failures are extremely rare. For example, even if researchers study 200 participants, a very large sample size for an ethnographic study [184], they are still unlikely to observe any disastrous failure. In addition, human subject studies are regulated by strict ethical protocols that

prohibits putting the participants in dangerous or unhealthy situations like disastrous technology failures. For example, it would be unethical to treat users with extremely violent VR content for understanding what leads to high-stake physiological reactions. As such, the safety concerns observed from the lab and ethnographic studies can be more precisely described as *potentially dangerous, yet manageable*. However, it is still unknown what role they play in the disastrous failures that actually "result in loss of life, significant property damage, or damage to the environment".

The Part 1 of this dissertation introduces a new data collection method adapted from communication studies that helps address the above limitations of the lab and ethnographic studies. Most importantly, this method enables researchers direct access to rich descriptions of disastrous failures from an unlikely data source - news articles. Successful applying our method to studying the disastrous failures of personal navigation technologies, we offer not only the insights about the safety concerns that directly contribute to the disastrous failures of personal navigation technologies, but also an example of implementing this new method in HCI context which can be replicated for studying safety concerns of other ubiquitous computing technologies.

1.2.2. Fairness and Geographic Biases in Computing Technologies

More and more frequently, decisions about individuals are completely or partially delegated to computing systems. Many decision making processes such as resume screenings, school admissions, and car insurance quotes now involve complex predictive models that

decision makers heavily rely on. As computers start to make decisions for and about individuals, a frequently asked question is - are these decisions from algorithms and models fair?

1.2.2.1. Fairness in Machine Learning. Fairness is largely regarded as a social construct [175, 190, 291]. As such, its specific definition varies in different context. In computer science, the definitions of fairness are mainly discussed in the context of predictive models such as machine learning [215]. In this context, researchers have broadly classified fairness definitions into the following categories [26, 215]:

- **Statistically Unbiased.** Statistical bias refers to the differences between the expected value of the model's predictions and the ground truth value. A statistically unbiased model will give estimates which on average equals to the true value. For example, if a model that predicts the sun set time is 10 minutes early one day and 10 minutes late the other day but the average error is zero, this model is said to be statistically unbiased. While this definition of bias was used in early discussions of machine learning biases, research community have now considered it insufficient [216]. One of its major limitation is that this definition fails to consider widely accepted social justice values, making it particularly inappropriate for describing the predictive systems that make decisions about humans.
- **Group fairness.** To address this limitation, group fairness, a umbrella term for a set of related fairness definitions, asks machine learning models to conform to one of the most widely acknowledged social values - no systematical performance differences among different populations. For example, one of the more popular definitions under group fairness is demographic parity, which requires that the subjects selected by the

models are demographically balanced and has close relationship to the affirmative action concept in policy making. Another frequently mentioned definition is predictive parity, which requires that the positive predictive value, or precision, are the same for different groups and has been strongly argued for in the high-profile case of recidivism prediction [56]. As we will show later, this concept is particularly relevant to the geographic biases that Part 2 of this dissertation focuses on. Other definitions of group fairness include the criteria to equalize false positive rate, false negative rate, and accuracy among different populations [215].

- **Individual fairness.** This fairness definition requires that similar individuals should be treated similarly. While this definition seems to be straightforward, it defers all nuances how to measure the similarity between individuals. According to the authors of this definition [80], similar individuals are context dependent and thus the similarity metric should be developed on a case-by-case basis in a form of distance metric. In the context of measuring the biases of recommendation systems, for example, similar users could be defined by the similarity of their ratings to the items in the systems, a widely used similarity metric in user-user collaborative filtering technique [163].
- **Process fairness.** While the previous definitions mainly focus on the fairness in the outcome, process fairness emphasizes that the process of building machine learning model is fair, resembling the procedural justice notion in legal proceedings. For example, focusing on a key step in building predictive systems - feature selection, Grgic-Hlaca et. al. [3] conducted an intriguing experiment in which they democratized the decision-making about what features to be included in the classification models to the crowdworkers.

Using the above definitions, many specific fairness issues in machine learning have been identified. In computer vision and image classification, researchers have repeatedly found that image classifiers have group fairness issues along a variety of demographic dimensions such as gender [248], race and skin tone [45], and geography [258, 72]. Similarly in natural language processing, gender biases [305, 39] and age-related biases [74] have been verified in some of the most popular language models.

Researchers have proposed various methods to mitigate the fairness issues above. While there exists methods relying more on human intervention, such as enhancing model’s interpretability so that human can vet, identify, and manually fix fairness issues in the models (e.g. [174, 180, 233]), most existing mitigation methods focus on building a less biased AI model itself. To this end, researchers, broadly speaking, take two approaches - building fairer training datasets and/or debiasing trained machine learning models. To build fairer training datasets, researchers have maintained the importance of increasing the diversity of the training datasets [285]. To this end, researchers have leveraged methods such as data augmentation to artificially increase the data variability [207], combining multiple datasets [157], and re-balancing the data distribution of biased datasets [302]. We highlight that the second part of this thesis which focuses mitigating the geographic inequalities in location-aware AI technologies, largely falls into this approach. Compared with the first approach, techniques to debias machine learning models are more context-dependent. For example, in the context of mitigating biases in word embeddings, popular debiasing techniques center around different ways to isolate gender-related information into subset of dimensions [306, 39].

1.2.2.2. Geographic Biases in Computing Technologies. Geographic biases can be largely regarded as one specific type of fairness issues. In the context of machine learning, it is usually possible to adapt specific fairness definitions to define geographic biases. Let us use the predictive parity definition under the group fairness as an example. The formal definition of predictive parity requires that for group a and b , a binary classifier ¹ has the same positive predictive rate, i.e., $P_a\{Y = 1|\hat{Y} = 1\} = P_b\{Y = 1|\hat{Y} = 1\}$ where Y refers to the true label and \hat{Y} refers to the predicted label. Adapting this formal definition for geographic biases, it prescribes that a binary classifier achieves the same positive predictive rate on samples from type of location g and type of locations h i.e., $P_g\{Y = 1|\hat{Y} = 1\} = P_h\{Y = 1|\hat{Y} = 1\}$.

Recent empirical studies have found geographic biases in some widely used machine learning systems. For example, using the above definition of predictive parity, Shankar et al. [258] showed that, for the task of recognizing bridegroom images, widely used pre-training models in computer vision have better positive predictive rates in developed countries than in developing countries. Similarly but independently, de Vries et al. [72] also found that the positive predictive rates of some of the most popular commercial image recognition models decrease as the medium income of the country where the image is from increase. More importantly, both studies also argued that the root cause of the geographic inequalities in computer vision models is the geographic biases in the training data - these training data predominantly focus on developed countries and include too few samples from developing countries.

¹Although fairness definition for non-binary classifier exist, we follow the convention in the literature and only consider binary classifier here.

To explain the the mechanism through which the geographic biases enter into the training dataset, computing researchers draw relationship to the wider representation issues in the digital content in general [149]. Given that it has been well known in social science that certain regions and communities (e.g., urban) are better represented than others (e.g., rural) in cultural, political, and economic aspects of our society, researchers argued that this representation biases across geography have been reproduced in the digital content that are generated by human and ultimately used to train computer algorithms [149]. Specifically, researchers have investigated the geographic coverage gaps in large digital knowledge repository such as Wikipedia [150, 106], in vast amount of social media content such as Tweets [151], and in the availability and quality of peer-economy services [284]. The general consensus is that, while the location-aware computing technologies come with great promises of empowerment, the digital content or services created by these platforms and later used to train many machine learning models still replicate, and in some cases even reinforce, the existing geographic representation biases in society.

1.3. Organization of the Thesis and Summary of the Chapters

This thesis is organized into two part. Th first part of this thesis focuses on understanding and mitigating the safety issues in personal navigation systems - a ubiquitous safety-critical technologies in the location-aware computing domain. Two projects are included in this first part:

Chapter 2: Understanding the Catastrophic Incidents Associated with Personal Navigation Technologies

Work in Chapter 2 addresses the previously mentioned gap that no research has analyzed the failures in these ubiquitous safety-critical technologies and aims to understand the safety concerns of personal navigation systems that will lead to catastrophic consequences. Borrowing techniques from public health research and communication studies, we construct a corpus of 158 detailed news reports of unique catastrophic incidents associated with personal navigation technologies. We then identify key themes in these incidents and the roles that navigation technologies played in them, e.g. missing road safety attributes such as lane-width and clearance height contributed to over 24% of these incidents. With the goal of reducing casualties associated with personal navigation technologies, we outline implications for design and research that emerge from our results, e.g. advancing “space usage rule” mapping, incorporating weather information in routing, and improving visual and audio instructions in complex situations.

Chapter 3: Mitigating the Risks of Personal Navigation Systems by Adding Road Safety Awareness

Work in Chapter 2 identified a number of catastrophic safety concerns of personal navigation systems. Chapter 3 begins the process to mitigate one of these concerns - the personal navigation systems are usually unaware of road safety attributes. To this end, the key idea behind our work is to add road-safety-awareness into personal navigation systems. To do so, we first created a definition for road safety that navigation systems can easily understand by adapting well-established safety standards from transportation studies. Based on this road safety definition, we then developed a machine learning-based road safety classifier that predicts the safety level for road segments using a diverse feature

set constructed only from publicly available geographic data. Evaluations in four different countries show that our road safety classifier achieves satisfactory performance.

The second part of this thesis shifts the focus to another type of risks of location-aware technologies - geographic biases. Following prior work in the fairness of machine learning, we focus on the machine learning or artificial intelligence based location-aware technologies. We identify that the source of the geographic biases of these location-aware AI technologies are rooted in the underlying core artificial intelligence models, such as computer vision, natural language processing, and knowledge base technologies, which these location-aware AI technologies are built upon. As such, we take on this root cause to address the geographic biases in the underlying core artificial intelligence models. To this end, we focus on two types of core AI technologies, computer vision and knowledge base, that are widely used in location-aware computing.

Chapter 4: Mitigating Geographic Biases in Computer Vision-Based Location-Aware AI Technologies

Prior research have identified that the geographic biases of image classification models can be attributed to the geographic biases in large image datasets on which these models are trained. Specifically, these datasets sourced most of the images from developed countries and underrepresented developing countries. Traditional approaches for enhancing the geodiversity of these datasets require large-scale market-based crowdsourcing, which is prohibitively expensive and subject to geodiversity issues per se. In this paper, we explore the potential of using Wikimedia Commons - the largest dataset of peer-produced encyclopedic imagery - as a low-cost mechanism to reduce the geographic biases of ImageNet. We detail how we produce ImageNet-compatible tags for Wikimedia Commons' images

by exploiting the Commons' peer-produced categories. Using these ImageNet-compatible tags, we show that the Commons can provide substantial number of images from developing countries for ImageNet and demonstrated that these geographically diverse images can significantly increase ImageNet pre-training models performance in developing countries which are previously underserved.

Chapter 5: Mitigating Geographic Biases in Computer Vision-Based Location-Aware AI Technologies

Another core AI technologies location-aware computing leveraged frequently is knowledge base which contains rich facts about and relationships between places. Many knowledge base technologies extracted these facts and relationships from Wikipedia. However, we show that geographic biases with respect to information access exist in these Wikipedia-based knowledge bases due to a popular yet erroneous assumption - all facts and relationships related to a certain geographic concept are exclusively contained in the Wikipedia article that has this geographic concept as title. We further demonstrate that knowledge bases that adopt this problematic assumption will have poorer information access to Wikipedia content on more prominent geographic concepts. To mitigate this problem and ensure that knowledge base have equal information access to Wikipedia content on all geographic concepts, we designed and developed the first system to gather all Wikipedia content that are relevant to one geographic concept. We show that, using a diverse feature set and standard machine learning techniques, our system can achieve good performance on most of our ground truth datasets, significantly outperforming baseline approaches.

Chapter 6: Supporting Projects

The final chapter of Part 2 summarizes two additional projects that provide important

context or key motivations to the core content in Part 2. Both projects were led or co-led by me and published during the course of my PhD studies. The first project built a system that mines data visualizations on Wikimedia Commons and recommends relevant data visualizations to news articles. This project is the one of the first works that taps into the enormous potentials of Wikimedia Commons images and inspired our work in Chapter 4 for using Wikimedia Commons to augment ImageNet. The second project analyzed the geographic biases in Pokemon Go - a location-based game that was wildly popular in 2016 and found that the spatial distribution of key resources in this game disadvantage rural areas and predominately minority neighborhoods. This project, which demonstrated geographic biases in a different type of location-aware computing, offer critical context and motivations for the our work that mitigate the geographic biases in the location-aware AI technologies.

Part 1

Safety Issues in Personal Navigation Technologies

CHAPTER 2

Understanding the Catastrophic Incidents Associated with Personal Navigation Technologies

Note: Much of this work was originally published in the Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems [187]. Necessary modifications have been made to situate the original content in the context of this thesis.

2.1. Introduction

A tourist drives his rental car across a beach and directly into the Atlantic Ocean [115]. A person in Belgium intending to drive to a nearby train station ends up in Croatia [289]. A family traveling on a dirt road gets stranded for four days in the Australian outback [195]. These incidents all have one major factor in common: playing a key role in each incident was a personal navigation technology, i.e. a GPS device, a mobile map app (e.g. Google Maps, Apple Maps) or a “SatNav”.

Catastrophic incidents associated with personal navigation technologies are sufficiently common that they have come to be associated with a colloquial name: “Death by GPS” [213]. While thankfully not all of these incidents involve the loss of life, it is not uncommon to see media reports of people endangering themselves or others and/or causing extensive property damage due in part to their interaction with a personal navigation technology.

It is tempting to blame these incidents on users and users alone. Indeed, reports of these incidents are often peppered with comments from witnesses and observers inquiring

as to why drivers “wouldn’t question driving into a puddle that doesn’t seem to end” [213] and did not notice “multiple-language traffic signs” [289]. However, it is our responsibility as HCI researchers to design better systems that help people avoid making “user errors” [220], especially when these errors involve such extensive human and financial costs.

The geographic human-computer interaction (“GeoHCI”) [124] literature includes a relatively large body of work that examines how people use GPS-based navigation technologies in standard scenarios and in the course of their everyday lives (e.g. [44, 129, 145, 179, 182]). However, no work has focused on the increasingly large number of catastrophic incidents associated with these technologies. In other words, the “Death by GPS” phenomenon has yet to be studied in a rigorous fashion.

This chapter seeks to begin the process of addressing this gap in the literature. As has been pointed out in the work on typical interactions with GPS devices [44], a major obstacle to the systematic analysis of “Death by GPS” incidents is that no database of these incidents exists. Additionally, methods that have been used to study interaction with GPS devices in the past (e.g. lab studies, field studies) are not valid for this type of analysis.

To overcome these obstacles, we turned to an unlikely source of data: news articles. This approach is adapted from the public health literature, where news articles are used as sensors when the research topic is of sufficient significance but no authoritative dataset is available. Using rigorous best practices for building a minimally biased-corpus of news stories and expert-led qualitative coding, we collected and analyzed a dataset of 158 news stories about unique catastrophic incidents associated with personal navigation technologies.

In our analyses of this corpus, we had two cascading research goals:

Goal 1: Identify the patterns that characterize catastrophic incidents associated with personal navigation technologies.

Goal 2: Use the identified patterns to generate implications for research and design that can help build safer personal navigation technologies.

More specifically, for our first goal, we sought to ascertain themes in both the basic properties of these incidents (e.g. Who was involved? What happened?) and themes in the roles that navigation technologies played in the incidents (i.e. How did the navigation technology specifically fail the user?). Based on the identified patterns, our second research goal involved outlining a series of concrete steps that researchers and practitioners can take to prevent the reoccurrence of common types of catastrophic incidents (and save lives).

We find, for instance, that a large number of “Death by GPS” incidents are single-vehicle collisions (likely far more than accidents caused by other factors), that stranding events were the next most common type of incident, and that distraction by a navigation device was significantly associated with more serious incidents. With regard to the roles of technology, we observed that missing road characteristics attributes (e.g. road surface types and current condition) had a substantial effect, as did the failure to correctly infer routing preferences (among a series of other factors). The implications for research and design that emerge from our findings span the spectrum of “GeoHCI” topical domains. For example, we discuss how our results highlight the importance of (1) incorporating vehicle type and weather information into routing algorithms, (2) improving navigation guidance in the face of complex geographies, and (3) developing separate interfaces for

tourists and locals. More generally, our results show that navigation devices can be more intelligent about safety than their current state-of-the-art: telling users to pay attention to their environment when the device is turned on. Blanket warnings like these are known to be ineffective in HCI [220], and our results show a path forward towards improved approaches.

In summary, this chapter makes the following contributions:

- (1) We perform the first research that systematically characterizes catastrophic incidents associated with personal navigation technologies and the role that these technologies played in these incidents. We identify major themes in the incidents themselves and in the roles played by technology.
- (2) With the goal of preventing the patterns we identified in these catastrophic incidents from reoccurring, we outline a series of implications for research and design that can help us develop safer personal navigation technologies.

To further research on this topic, we are also releasing the core dataset we developed for the work in this chapter ¹. This dataset consists of the complete corpus of 158 news stories along with all the codes we applied to each story in the process described below. To make our findings more accessible, we are also releasing an interactive web map version of the corpus, which allows users to see the exact location of each incident along with the Google StreetView imagery at the location and further information about the incident ².

A Note on Terminology: The subject of this research resulted in several terminological challenges. The core technologies of interest to this chapter – GPS devices, SatNav devices, and mobile map applications like Google Maps and Apple Maps – are

¹The dataset can be downloaded here: <https://goo.gl/8vE14V>

²The interactive map is available here: <https://goo.gl/j1Q8S4>

often referred to using the term “GPS”. This term ignores the diverse positioning techniques (e.g. Wi-Fi positioning), routing algorithms, and cartography built into these technologies, so we felt it was imprecise to use this more casual language. As such, we use the term “personal navigation technology” (sometimes shortened to “navigation technology” or “navigation device”). Similarly, given the diversity of the types of incidents in our corpus, assigning this class of incidents a formal name was not straightforward. We chose the term “catastrophic incidents” in accordance with the “extremely unfortunate or unsuccessful” definition of “catastrophic” [271].

2.2. Related Work

This work’s core motivation primarily emerges from two areas in the “GeoHCI” literature: (1) work that has examined the use of personal navigation technologies in standard scenarios and (2) research that has looked at the long-term behavioral and cognitive effects of using these technologies.

2.2.1. Navigation Technologies in Standard Scenarios

Researchers began to investigate HCI issues associated with in-car navigation systems almost as soon as these technologies were first commercialized [76, 77, 294]. This thread of research covers a diverse set of topics including attention demands [77, 169, 294], cartography [179, 208, 249], different modes of output [71, 145] and age-related variation [7], all with a focus on everyday usage scenarios. For instance, Kun et al. [169] conducted a lab simulation study and found that graphical GPS interfaces distracted users from the primary task of driving. Medenica et al. [208] coupled augmented reality with in-car GPS

navigators and showed that this combination reduced drivers' distractions. Jensen et al. [145] compared different interaction modes of in-car GPS navigators and concluded that the combination of audio-visual output is preferred by drivers, but did not significantly reduce device-related distractions.

The projects in this research thread that most directly motivated our work are those of Hipp et al. [129] and Brown and Laurier [44]. Both studies considered the “troubles” drivers encountered with in-car GPS devices in typical driving situations. Hipp et al. [129] conducted a traditional user interface evaluation to compare the performances of different types of in-car navigation systems on the same route. They identified unexpressed routing preferences, failure to understand intentional detours from planned routes, and the lack of real-time traffic information as the common interaction weakness of commercial navigators (with the latter now being fixed in most modern navigation technologies). Brown and Laurier [44] carried out an interaction analysis in which they observed and interviewed drivers about their daily uses of in-car GPS to understand their navigation practices. They outlined five types of “normal troubles” of using in-car GPS navigators in everyday driving: destination, routing, maps and sensors, timing of instructions and inflexibility of the technology.

This work is distinguished from that above in that instead of studying the use of personal navigation technologies in standard scenarios, we focus on catastrophic incidents that involved these technologies. Some of the roles that these technologies play in catastrophic incidents are similar to those identified in the literature on standard scenarios, and other roles are new to the literature (as are the resulting design implications). We

discuss the relationship between our findings and the findings from prior work in detail below.

2.2.2. Long-term Impact of Navigation Technology Use

Another class of relevant research focuses on understanding the behavioral and cognitive changes produced by personal navigation technologies. For instance, Leshed et al. [182] conducted an ethnography-based study and showed that drivers using GPS-based navigation technologies are disengaged from their surrounding environment. Aporta and Higgs [14] examined the long-term impact of navigation technology at a larger scale, arguing that the adoption of navigation technologies has alienated many Inuit hunters from the traditional wayfinding skills they have depended on for thousands of years. Other studies have looked at the cognitive impact of navigation systems. For instance, Gardony et al. [93] conducted a lab-based simulation study and demonstrated that these devices may impair users' ability to record information about the environment and their spatial orientation. The findings of this line of work inform this chapter's research and design implications, specifically those related to the multifaceted relationships between a navigation technology, its user, and the environment.

2.3. Methods

Although catastrophic incidents associated with personal navigation technologies are sufficiently noteworthy to have been given a moniker – “Death by GPS” – no authoritative dataset of these incidents exists. The high stakes of these incidents make them worthy of study, but the lack of available data and relative rarity of these incidents make it difficult

to analyze them. Additionally, lab experiments or other simulations are not currently well-suited to this research area.

Fortunately, the domain of public health has significant experience studying phenomena with the same core properties as “Death by GPS” incidents, i.e. relatively rare phenomena of media interest for which no authoritative dataset is available and for which simulations are not currently tractable. Specifically, to examine these phenomena, researchers in this domain have followed a two-step pipeline: (1) build a corpus of news stories describing these incidents and (2) analyze the corpus using expert-led qualitative coding techniques. For example, in the absence of a national surveillance system for homicide-suicide in the United States, Malphurs and Cohen [196] collected and coded related news articles from 191 national newspapers to identify the number and subtypes of such incidents. This approach of relying on newspapers to summarize the characteristics of homicide-suicide incidents has also been applied in the Netherlands [185] and Italy [241]. Similarly, to study the collisions between wheelchairs and motor vehicles, a type of accident that is not distinguished in police reports, LaBan and Nabity [171] gathered 107 news articles using LexisNexis. They analyzed this corpus to understand gender incidence ratios, proportion of different types of motor vehicles, the time of incidents, and other characteristics of these incidents.

In our work, we adopt this approach from the public health literature. To do so, we first verified that no relevant authoritative dataset exists by contacting several local police departments and national agencies, including the Minneapolis Police Department (USA), the Aachen Police Department (Germany), National Transportation Safety Board (USA)

and National Highway Traffic Safety Administration (USA). We then implemented the pipeline from public health, using the process described in more detail below.

2.3.1. Phase 1: Corpus Development

One of the key challenges in the public health-based approach is gathering the corpus of news articles. Most prior work has relied on one of two methods: (1) an exhaustive search in the local newspaper of a specific study site (e.g. [27, 235]) or (2) unstructured but extensive querying of news search engines (e.g. [171, 196]). Since our work is not well-suited to a specific study site, we implemented a more robust version of the latter approach using best practices from communication studies for sampling news stories with minimal bias [173, 276].

The first step in this minimal bias sampling approach involves leveraging prior research in this space (i.e. the literature covered in the Related Work section) to seed a set of keywords, which is then grown using a variety of structured strategies (e.g. synonym generation). These keywords are then used to iteratively query a news database (in our case LexisNexis), with the set of keywords refined at each step. Achieving acceptable precision for one's keywords is particularly important given that these databases often have strict limits on the total number of news stories that can be returned for a given query (e.g. LexisNexis's is set to 1000 stories). We were able to achieve a precision of 40.8%, which is within acceptable parameters [276]. This left us with 408 articles that were related to catastrophic incidents associated with personal navigation technologies. We note that we restricted our search to stories published in 2010 or later to ensure that our findings are relevant to modern personal navigation technologies (e.g. smartphone map

apps) rather than early-generation devices. Additionally, due to language constraints with respect to the database and the coders, we only searched for stories written in English, a subject we cover in more detail in the limitation section.

Two challenges remained before we could begin the next stage of the pipeline. First, many of the articles were opinion columns about the “Death by GPS” phenomenon and did not describe specific incidents. Second, some incidents were described by two different journalists in two different publications. To remove these articles from our dataset, two researchers conducted an exhaustive search, reading each article and evaluating its validity for our study and matching duplicates (we kept the more detailed of any two stories on the same incident; disagreements were resolved through discussion).

In the end, we were left with a corpus that contains 158 news stories, each describing a unique catastrophic incident associated with personal navigation technologies. For replication purposes and for researchers who may want to apply this method in other contexts, we have included additional detail about how we implemented our corpus-building procedure in the documentation of our coded dataset. We also discuss the one minor change we had to make to the standard procedure to adapt it to the goals of our research project: we could not simply use the keywords from prior research on interaction with navigation technologies because these only focused on standard scenarios. As such, we used a slightly more iterative keyword generation method in which researchers identified keywords from small samples of actual relevant news stories.

2.3.2. Phase 2: Coding by Domain Experts

The second stage of the public health pipeline involves employing a relatively standard qualitative coding procedure with an important exception: coders are domain experts. This expertise enables coders to map properties of incidents reported in news articles to pre-existing topics in the literature of interest, or to new challenges when relevant. In our case, our two coders (members of our research team) had extensive expertise in both geography and HCI, the two fields most associated with our research questions. More specifically, each coder had both a Masters' degree in geography or geoinformatics and a Masters' degree in computer science (with a focus on HCI).

The specifics of the coding process were as follows: using a small seed corpus, knowledge of our research goals, and expertise in the research domain, our coders jointly established a series of coding dimensions. Next, using a random sample of 10 articles, the coders jointly developed a list of codes for each dimension and developed a corresponding codebook (this is included in our dataset). Both coders then evaluated a set of 40 overlapping articles to assess each dimension for interrater reliability. Importantly, when it was not possible to assess an article for a particular coding dimension, coders left the value blank.

The Cohen's Kappa of coders' results on all dimensions ranged from 0.69 to 0.95, which indicates "substantial agreement" [177]. Of particular note, we achieved a Cohen's Kappa of 0.79 for the *Technological Cause* dimension, which is the basis for a set of key findings below. As the Cohen's Kappa was sufficiently high for all dimensions, coders evaluated the remaining articles on an individual basis.

Beyond *Technological Cause*, other major coding dimensions that were considered included *Seriousness of the incident* (e.g., was death involved?), *Incident Type* (e.g., was it a single-vehicle collision? Did a vehicle get stranded?), *Weather, Road Surface* (e.g., was it on a dirt road or a paved road?), *Distraction* (i.e., was distraction explicitly noted as an issue in the article) and *Local Driver* (i.e., whether the driver lived in the area of the incident). A complete list of dimensions and their corresponding specific codes is included in our public dataset. For the major coding dimensions, coders were able to assign codes for over 90% of incidents with the exception of *Local Driver*, in which 37% of incidents could not be coded.

2.3.3. A Note for Interpreting Our Results

As described above, there is a consensus (e.g. [165, 171, 185, 196, 235, 241]) in the public health literature that when no authoritative data is available, the news article-based pipeline we employ here can provide valuable early insight about a phenomenon of interest. However, news article-derived data has its share of limitations, as is the case with many datasets considered in public health (e.g. even authoritative crime datasets have been criticized for potentially strong racial biases [238], an issue the computing community has been facing in association with predictive policing technologies [222]). To the best of our knowledge, our use of news article-derived data is novel to the HCI literature. As such, we believe that highlighting the known limitations of this type of data early is important so that our results can be interpreted in proper context.

The most significant limitation of news article-derived data is a risk of “newsworthiness” bias, or an overrepresentation of incidents that are in alignment with the incentive

structures of news organizations. While at least one study has found no such bias (e.g. [235]), others have found newsworthiness bias to manifest as an overrepresentation of (1) accidental incidents (e.g. fewer suicides, more unusual events) or (2) more fatal incidents (e.g. more murders, fewer assaults) [85]. All incidents that we examine are accidental in nature, making the accidental bias less relevant [235, 85]. However, a potential bias towards fatal incidents is important to consider when examining our results below.

To minimize further risk of bias, we employ robust statistical tests when making comparisons between types of incidents. In most cases, we are able to simply use Pearson's Chi-squared test of independence. However, in circumstances where the assumptions of Chi-squared distribution are violated due to relatively small sample size, we used a likelihood ratio G test of independence, a best practice suggested by [5, 206]. Bonferroni correction has been applied to all p-values reported in this chapter.

Newsworthiness bias mainly affects proportional results (i.e. comparisons between incident types), which are a small percentage of the results we present below. The bulk of our results are either qualitative descriptions of incidents or absolute values (e.g. raw counts of incidents of certain types). Our absolute results should be interpreted in the context of a limited understanding of the size of the incident population, i.e. we do not know what share of catastrophic incidents associated with personal navigation technologies are included in our news corpus. However, even if the incidents in our sample are close to the entire population, the aggregate devastation to blood and treasure of just these incidents make them worthy of analysis and discussion in the HCI literature, which

does not often examine such high-cost interactions with technology. In order to add additional context and further this discussion, we provide qualitative descriptions of incidents wherever space allows.

2.4. Results

In this section, we provide an overview of the major results that emerged from our coding process. In doing so, we seek to address our first research goal: characterizing patterns in catastrophic incidents associated with personal navigation technologies. We organize our thematic findings into two groups (1) themes in the basic properties of these incidents and (2) themes in the technological causes of these incidents. We discuss each group of findings in turn below.

2.4.1. Basic Properties

2.4.1.1. Many People Have Died in Incidents Associated with Personal Navigation Technologies. Table 2.1 shows the results of our coding for the Seriousness of the incidents with respect to human and financial cost. Clear in Table 2.1 is that navigation technologies have been associated with some truly tragic events: our corpus describes the deaths of 52 people in total, including two children. These deaths occurred across 45 incidents, or 28% of our corpus. Additionally, our corpus contains 23 incidents (15%) that resulted in significant bodily harm, but not death.

Although the proportion of fatal incidents in our corpus may be exaggerated due to the aforementioned newsworthiness bias, the absolute number of deaths (and injuries) associated with navigation technologies that we have identified is alarming. GPS devices,

Seriousness of Incidents	#(%)
Major	67(43%)
↔ Deaths	44(28%)
↔ Injuries	23(15%)
Medium (e.g., property damage, legal consequences)	52(33%)
Low (e.g., significant detour)	39(25%)

Table 2.1. Distribution of the *Seriousness* of these 158 incidents.

mobile maps, and other navigation technologies provide us with tremendous benefits, but these results indicate that they also have a set of costs that had not yet been systematically enumerated. These results also highlight the importance of better understanding catastrophic incidents like those studied here, as well as using this understanding to design safer technologies.

Table 2.1 also shows that “Death by GPS” is not the ideal term to describe incidents associated with navigation technologies that have serious implications. Over 50% of the incidents in our corpus did not involve death or significant injury, with the damage in these cases being primarily of a financial or other nature. Examples of these incidents include a group of skiers who intended to go to La Plagne, a famous ski resort in the Alps, but ended up arriving at Plagne, a town in southern France that is 715 km away (article #46). Another example involved five men who drove onto a nuclear power plant’s property at the behest of their navigation device and were suspected of terrorism (article #95).

2.4.1.2. The Most Common Incident Type is a Single-Vehicle Crash, but There is Substantial Incident Type Diversity. Table 2.2 depicts the results of our coding for Incident Type and shows that the most common type of incident in our corpus is

Types of Incidents	#(%)
Crashes	90(57%)
↔ Single-vehicle collisions	51(32%)
↔ Crashes with vehicles	26(17%)
↔ Crashes with pedestrians/bikes	13(8%)
Stranded/stuck (e.g., in the wildness)	31(20%)
Significant detour (e.g., wrong addresses)	25(16%)
Wrong way (i.e., opposite side)	7(4%)
Trespass (e.g., violate space usage rules)	5(3%)

Table 2.2. Distribution of the *Incident Types* of these 158 incidents.

car crashes. However, the table also shows that crashes are far from the only type of incident we encountered. For instance, almost 20% of incidents resulted in cars being stranded in very rural areas and over 15% involved people going on substantial detours. We were also surprised by the number of reports (7) of people driving on the wrong side of the road for an extended distance. Such examples include a person who drove 48km on the wrong side of a highway after following her device’s instructions to enter the wrong freeway ramp (article #90) and a 37-year-old man who was caught driving the wrong way on an Australian highway for more than 10 km and attributed the error to his navigation device (article #12).

In Table 2.2 , we also show subtypes of the Crashes incident type. We found that single-vehicle collisions comprised the majority of crashes (51 cases, 32% of overall incidents), with crashes with other vehicles (26 cases, 17%) and crashes with pedestrians and bikes (13 cases, 8%) making up the remainder of crash incidents. To understand single-vehicle collisions in more detail, we did an additional round of coding to identify more detailed themes (this was done by a single expert coder). Here we found that vehicles colliding with

buildings, walls, and guardrails due to excessively narrow roads were the most common type of single-vehicle incident. Crashing with low overhead bridges is another common occurrence in our corpus, with a diverse array of other objects in the environment being the subject of the remainder of the single-car crashes.

Personal Navigation Technology-related Crashes Appear to Be Proportionally Different Than Typical Crashes

To put the above results in context, we utilized The National Automotive Sampling System General Estimates System (NASS GES) dataset from the U.S. National Highway Traffic Safety Administration [83]. The NASS GES dataset contains a representative sample of vehicle crashes of all types as reported by police. While not directly comparable to our corpus, the NASS GES can provide a sense of whether personal navigation technology-related crashes are proportionally similar to the population of car crashes, or whether the role played by navigation technology manifests in different types of crash outcomes.

Our results suggest that the latter is the case: crashes associated with personal navigation technologies appear to be different in type relative to typical crashes. For instance, only 15% of car crashes in the NASS GES dataset are single-vehicle collisions, whereas the same type accounts for 58% of crashes in our corpus (Table 2.2). Moreover, crashes associated with building/walls/guardrails and overhead bridges are much less common in the NASS GES dataset, comprising less than 2% of crashes overall, while in our corpus they account for 42% of all crashes. Among other implications, this result provides further evidence that simply adopting standard recommendations from traditional traffic

safety research will not be able to address the safety concerns associated with personal navigation technologies.

2.4.1.3. Unfamiliarity with One’s Surroundings Plays a Key Role. A substantial percentage of the incidents in our corpus occurred when users of personal navigation technologies were outside of their home region. Specifically, 78% percent of the incidents involved non-locals and only 22% percent involved locals. Some examples of incidents involving non-locals include one story in which a person drove her car into a swamp (article #23). The driver was quoted as saying “This was the road it told me to take . . . I don’t know the area at all, so I just thought it was oka”. Another incident consisted of a driver hitting and killing a pedestrian, with the corresponding news article reporting that “the driver was unfamiliar with the area and was adjusting her GPS navigational syste” (article #71).

While the use of navigation technologies likely increases outside of one’s home region, this result does suggest that user interfaces for navigation technologies may want to encourage more caution and support users in different ways when they are in their home region and when they are traveling. We discuss these implications for design in more detail below.

2.4.1.4. Distraction Leads to More Serious Incidents. We identified a significant association between our Distraction coding dimension and our Seriousness dimension, with distraction leading to many deadly incidents ($\chi^2(2)=19.2$, $p<.05$). Examining the 21 deadly incidents that involved distraction in more detail, we found that in five cases, people were using non-critical features of their navigation device. For instance, a driver

killed a cyclist while “using the zoom-in function” (article #33) and another driver from Springfield injured a bicyclist while “looking for place to eat on GPS” (article #40).

2.4.1.5. Stranding Risk Increases with Dirt Roads, Bad Weather, and Especially Both at the Same Time. We observed significant associations between our Road Surface coding dimension and our Incident Type dimension. In particular, if vehicles were traveling on a dirt road, there were more than the expected number of stranding incidents ($G^2(12)=53.0, p <.05$). This was especially the case when weather was a factor. Examples include a medical student from Quebec who followed GPS and got stranded on a logging road for three days in the snow (article #4) and a British couple and their children who were stranded for four days on an unsealed road that was made muddy by torrential rain (article #115). Interestingly, the latter family thought that their in-car GPS device was suggesting a significant shortcut and followed its instructions as a result, a point we return to later.

More generally, we found significant interaction between disaster type dimension and the weather dimension ($G^2(4)=21.1, p <.05$). Specifically, there are more than the expected number of stranding incidents under severe weather, as one might anticipate.

2.4.2. Technological Causes

Table 2.3 shows the results of our coders assessing each article for its Technological Cause. In this section, we discuss the themes in the distribution of these results, as well as the findings from a more detailed coding that we conducted to understand important trends.

2.4.2.1. Attributes that are Missing or Incorrect Are a Major Problem. Geographic information, like that which is used in routing for personal navigation technologies,

Technological Causes	#(%)
Missing or incorrect attributes	64(53%)
↔ Physical attributes of the road (e.g., road surface)	30(25%)
↔ Clearance height	17(14%)
↔ Traffic rules (e.g., no left turn)	5(4%)
↔ Temporary blockage	3(3%)
↔ Ferry line as road	3(3%)
↔ Geopolitical boundary (e.g., country border)	2(2%)
↔ Private area	2(2%)
↔ Bridge limitation	2(2%)
Non-transparent/wrong route preference	18(16%)
Instructions/visualizations	18(16%)
Incorrect toponym disambiguation (i.e. Salem, MA vs. Salem, OR)	8(7%)
Geocoding	7(6%)
Missing and incorrect geographic objects	5(4%)

Table 2.3. Distribution of *Technological Cause*. Note: The # does not add up to 158 because coders did not enter a code when there was not enough information in given news story to make a certain type of assessment.

consists of two components: spatial information and attributes of that spatial information [101]. Broadly speaking, in a routing context, the spatial information is the location of a road and the attributes consist of key properties of the road (e.g. the speed limit).

Our results in Table 2.3 suggest that missing and incorrect attributes play a major role in the catastrophic incidents in our corpus, being in part responsible for 64 (53%) of these incidents. To better understand the types of attributes involved, one expert conducted a second round of coding to determine the types of attributes that were most often missing or incorrect and the results are also included in Table 2.3. The physical characteristics of the road (e.g. width, surface) (30 incidents) and clearance height (17 incidents) were by

far the most common type of attributes that were missing or incorrect. Indeed, stories about routing algorithms neglecting the road characteristics and the heights of overpasses are pervasive in our corpus. For example, as noted above, failure to incorporate road surface information led multiple sedans to be stranded on unpaved roads (often in a very dangerous fashion) (e.g. article #4, #115) and multiple trucks ran into serious trouble due to low-clearance roads (e.g., article #6, #34, #36). Indeed, we found trucks were more susceptible to suffer from attribute related issues due to this problem as evidenced by the significant interaction between our Vehicle Type coding dimension and the Technological Cause dimension ($G^2(15)=67.4, p <.05$).

Another theme present in the attribute types in Table 2.3 is the notion of “space usage rules” (SURs) [246], or regulations associated with the use of a certain area (in this case, a road). For instance, in one incident, a truck that traveled on truck-prohibited road killed a father and a daughter in a sedan (article #27). In another, an in-car GPS device guided a driver up a private driveway, and the driver ended up in a physical confrontation with the owners of the property (article #102).

2.4.2.2. Cartographic and Audio Instructions Are Not Capable of Handling Complex Geographic Contexts. Table 2.3 shows that almost 15 percent of our incidents involved an issue with routing guidance, either in visual (cartographic) or audio form. Past work on the use of GPS devices in standard scenarios identified that excessive instructions are a significant problem with GPS usability [7, 44]. While we did observe this problem in our corpus, many of the incidents given this code by our experts related to a different issue: the inability of the personal navigation technology to help drivers navigate complex geographic contexts.



Figure 2.1. A Google Street View image depicting the complex geography of the location of the incident in article 146.

For example, in one story in our corpus, a person who was driving at night was faced with a freeway on-ramp that was immediately parallel to a railroad track (article #146). Figure 2.1 shows a Street View image of the exact location of the incident. When the driver’s navigation device asked him to turn right, the driver turned onto the railroad tracks as the instructions were ambiguous. Ten kilometers later, the driver’s car was destroyed by an oncoming train, but fortunately the driver survived by jumping out of the car. Similarly, article #66 tells a tragic story in which a bicyclist was hit by a driver who ignored a “Yield” sign at a non-typical intersection because the driver’s navigation device screen simply instructed her to “go straight”. Wrong-way driving was particularly (and significantly; $G^2(24)=100.1$, $p < .05$) associated with cartographic and navigation instruction issues, and complex geographies were common in these cases. For instance, one report in our corpus (article #39) describes the story of a driver who followed her

navigation device’s instructions to “take the first left turn” at a roundabout. However, the actual first left turn (not the first legal left turn) was the exit ramp of a freeway, and the driver – who was on the road at night – entered the freeway driving in the wrong direction. This driver sadly lost her life.

2.4.2.3. Standard Scenarios versus Catastrophic Incidents. As noted above, past work has done a rigorous job of identifying and categorizing problems encountered by users of personal navigation technologies in standard usage scenarios. While the issues discussed above have not been highlighted in prior work, one additional contribution of the results in Table 2.3 is to add gravity to many of the previously identified issues. For instance, in a study of challenges encountered in standard GPS device usage, Brown and Laurier [44] found that route preferences, out-of-date spatial data, the timing of navigation guidance, and positioning errors were key sources of user frustration. Some of these issues appear in Table 2.3, meaning that they were in part responsible for a number of catastrophic incidents in addition to more everyday usability issues.

Of particular note are Brown and Laurier’s findings with respect to route preference. Route preference issues played a role in 15% of the news stories in our corpus, indicating they are a significant issue in catastrophic incidents as well as everyday usage scenarios. However, the route selection issues present in our corpus are of a substantially different character than those identified by Brown and Laurier. Specifically, while participants in Brown and Laurier’s study wanted more route choice, people in our corpus were given too many choices (i.e. at least one was dangerous). For example, in one incident a Canadian couple got lost in rural Nevada after selecting the “shortest path” route option suggested by their navigation device, which included a little-maintained road. They were stranded

in Nevada for 49 days, during which time the husband sadly lost his life (article #9). We return to this case and the issue of strict “shortest path” routing and route selection in the implications section.

With respect to prior work, it is also interesting to examine Table 2.3 for what is not common or present at all in our corpus. It appears that some issues with everyday use of navigation technologies do not play a role in catastrophic incidents associated with these technologies. For instance, positioning inaccuracies and the lack of adaptability to intentional “detour” were the sources of major usability challenges in the work of Brown and Laurier. However, neither appeared in our corpus. Similarly, missing spatial data was not a major issue in our corpus – it played a role in only 5% of incidents – but has been identified as a significant issue in standard usage scenarios. For catastrophic incidents, the issue appears to be attributes rather than the spatial data itself, a subject we discuss immediately below.

2.5. Implications for Research and Design

In this section, we turn our attention to our second research goal: helping to identify solutions to the problems we found in our results section by enumerating a series of implications for both research and design. Some of these implications suggest improvements to the design of existing systems, while other present important new challenges for the GeoHCI research community. We have organized these implications into two high-level categories corresponding to two broad areas of the GeoHCI research space: implications related to *spatial computing* (e.g. routing algorithms, missing attributes) and implications related *user interaction* issues.

2.5.1. Spatial Computing Implications

2.5.1.1. Geometries without Attributes Can Be Dangerous. A major finding above is that missing attributes play a substantial role in the catastrophic incidents in our corpus. This suggests that road network geometries may be “getting ahead” of the corresponding attributes. That is, data providers are adding road segments to their networks faster than they are adding the attributes to those segments that are necessary to facilitate safe routing.

These results suggest that data providers may not want to integrate road segments into their networks unless those segments have high-quality data for a core set of attributes. Based on our findings, these attributes should include the type of the road (e.g. dirt, asphalt) and the clearance height of the road (as defined by any overpasses, tunnels, and other obstacles) at minimum.

2.5.1.2. Incorporate Vehicle Type into Routing Decisions. Even when high-quality attributes are included, however, they must be used intelligently by routing algorithms. Returning to Table 2.3, a key theme emerges in this respect: many of the incidents included in this table could have been prevented if routing algorithms can understand the limitations of the vehicle that they are routing. For instance, it is often not safe for sedans to drive down rough country roads, and trucks should not drive down roads with low clearance heights. Coupled with good coverage of attributes, incorporating vehicle type information would be a straightforward and effective way to maintain good coverage of possible routes (e.g. allowing SUVs to drive down rough country roads), while at the same time increasing safety.

2.5.1.3. Extend Space Usage Rule Mapping Efforts to Road Networks. We identified that the lack of space usage rules (i.e. usage regulations) is a common missing attribute associated with the catastrophic incidents in our corpus. Space usage rules (SURs) have been a topic of growing interest in the GeoHCI research community in the past few years (e.g., [132, 246, 268]), but this literature has focused on mapping rules associated with regions rather than roads. For example, a common research challenge in SUR mapping is identifying regions in which smoking is legal or illegal [246].

Our research suggests that more effort should be spent on the identification of SURs for road networks. In particular, improving data related to the maximum clearance of roads, whether roads are public or private, and improved recognition of traffic rules are particularly important. Fortunately, unlike many SUR mapping challenges that require multifaceted approaches (e.g. natural language processing, crowdsourcing), it is likely that much of the work here can be done using computer vision (CV) approaches. The automated detection of traffic rules in this fashion is already underway [21]. It is likely that private property signs would present unique challenges for CV algorithms due to their diversity, but this is a contained problem that can likely be at least partially addressed with current state-of-the-art CV techniques.

2.5.1.4. Weather Information Matters for Routing. Our results suggest that routing algorithms should consider weather information when generating routes, and should do so in concert with vehicle type information. A substantial number of the stranding incidents in our corpus would have been avoided with relatively straightforward weather- and vehicle-aware routing approaches. For instance, if it has rained 20 centimeters in the past day, routing algorithms should not send drivers of sedans down dirt roads. Similarly,

if it has snowed 20 centimeters and it has stayed below freezing, routing algorithms should recommend that sedan drivers stick to main thoroughfares, which are plowed more quickly and more often (and should perhaps consider increasingly available information in many cities about which roads have been plowed since the last major snow).

2.5.1.5. Unintended Consequences of Map Matching. We observed in our corpus that map matching techniques [107] can backfire. These techniques are designed to mitigate GPS noise by “snapping” vehicle locations to the closest road network geometry. However, they were likely involved in the three incidents in which a person drove on a train track parallel to a road (article #17, #32, #116) and also a few incidents in which people drove on the wrong side of the divided road (e.g. article #12, #90) (all cases happened in evening). In these cases, map matching algorithms likely “snapped” the driver’s position to the nearest or the correct side of the road, making the driver believe that they were on right track (which may be difficult to assess at night).

Although more work is needed to understand this issue in detail, one potential improvement is to make map matching algorithms more error-sensitive in situations in which the distance between geometries is smaller than the error tolerance. Specifically, when an algorithm notices that there are multiple parallel linear geometries (e.g. a divided highway or a railroad parallel to a road), it can reduce the tolerance of its map matching radius. When observing a small, persistent mismatch for a short period, GPS devices could immediately prompt users about this mismatch and ask the driver to look at the environment to confirm that the vehicle is on a legal road.

2.5.2. User Interaction Implications

2.5.2.1. Route Preference Must Be Accompanied with Adequate Information

to Make an Educated Choice. Past research on the use of navigation technology in standard scenarios has advocated for providing greater route preference for users. Our results suggest that this preference must be accompanied with adequate information for users to make safe decisions. Current navigation devices often offer multiple routing preferences such as “fastest”, “shortest”, or “eco mode”. At the very least, these technologies should warn users that certain choice may involve traversing unsafe territory, as was the case with the Canadian couple that chose the “shortest path” through Nevada without understanding the consequences of doing so.

As mentioned above, in addition to the usability problem of excessive instructions with bad timing found by previous studies, we identified a new type of guidance-related problem: instructions that are too simple for the spatial decisions that the user has to make. Two research challenges emerge from this issue: (1) automatically detecting complex geographies and (2) developing interfaces to better support users in these contexts. With regard to the first challenge, public crash datasets (e.g. [83]) can provide ground truth information to help develop regression models that assess the complexity of a routing context based on the topology of the surrounding road network (and likely other information, such as railroads). The second challenge might be at least partially addressed through the use of image-based navigation, i.e. by annotating Street View imagery with arrows and labels. Image-based navigation is known to have benefits over most other approaches [293] but needs to be updated frequently to reflect any potential changes in the environment.

2.5.2.2. Local Mode and Non-Local Mode. Our results suggest that non-local drivers are at substantially greater risk for catastrophic incidents associated with navigation technologies than local drivers. These findings advocate for the development of customized features for each of these populations, i.e. a “local mode” and a “non-local mode”. For instance, neuroscience research has shown that more attention is required when driving in an unfamiliar environment [193]. As such, designers should investigate strategies for reducing interaction with drivers when drivers are outside their home region(s). Additionally, routing algorithms could provide non-local drivers with an “easiest” route that prioritizes highways and avoids complex intersections to minimize the turn-by-turn instructions and general information load. Similarly, GPS devices could disable non-essential functionality (e.g., searching for local restaurants) while in unfamiliar territory and re-enable those functions only when drivers come to a complete stop (or return to their home areas).

2.6. Discussion

In this chapter, we provided the first characterization of the patterns in catastrophic incidents associated with the use of personal navigation technologies. We have also outlined a series of implications for design and research that emerge from these patterns. Below, we highlight several discussion points associated with this research.

First, it is interesting to reflect on the design implications in the context of automated vehicles. Some of the implications will clearly become moot if a human is not behind the wheel (e.g. those related to improved instructions), as will be the case for many of the core functionalities of navigation devices [28]. However, other implications may become significantly more important. For instance, adding attribute information to

geometries, improving understanding of space usage rules and incorporating weather information will be critical to helping automated cars avoid dangerous navigation decisions. The same would likely apply in the nearer term with semi-automated cars, as recent work suggests that there may be excessive deference to automated routing approaches given the attentional challenges of partial automation [48]. Similarly, the research community has pointed out the need to keep drivers engaged when behind the wheel of a mostly-automated vehicle. Prompting users in the case of persistent map matching issues and engaging them in other difficult navigation-related tasks may be one way to accomplish this goal.

Second, the news article-based pipeline we use here may be able to help HCI researchers examine other difficult-to-study phenomena. As noted above, our public health-based approach is best suited to phenomena that share three properties: (1) no authoritative dataset is available, (2) instances are too rare to observe in large numbers in the wild and cannot be replicated in a lab setting, and (3) instances are frequently covered by journalists. Some additional HCI phenomena that share these properties include criminal events in the sharing economy and safety concerns related to location-based games like Pokémon GO [61]. To make it easier for researchers to employ our methodology, we have provided a step-by-step description of our approach in the documentation that is included with our coded dataset.

It is important to note that our coded dataset contains much more data than we could fully describe in this chapter. While we have highlighted what we as researchers in the geographic HCI domain believe to be the most important themes in our results, other researchers may benefit from examining our data from a different perspective. One

particularly interesting avenue of exploration that we are working to investigate is using the spatial locations of each incident (available in the dataset) to try to develop predictive models of the types of areas in which the use of navigation technologies might be particularly risky.

While we believe it is important for the HCI community to examine and learn from catastrophic incidents associated with the use of computing technologies, it is also important to put the relative incidence of these catastrophes in context. While we identified that GPS devices and related technologies played a role in at 158 catastrophic incidents involving 52 deaths, these technologies have also likely played a role in saving the lives of many people (e.g. guiding people to emergency resources, preventing people from getting lost). With this in mind, the design and research suggestions we make above are careful to be augmentative of existing navigation technology functionality rather than substantially altering current functionality.

2.6.0.1. Limitations. In addition to the drawbacks of the news article-based pipeline discussed above, our work is also subject to several additional limitations. For instance, while our incident corpus is the first agglomeration of its type of any scale, future work should seek to increase this size by either finding more news stories or collecting data on incidents that are not reported in the news. With respect to identifying unreported incidents, crowdsourcing has been proven effective for building databases of technology failures in the domain of aviation [35]. This may be an approach that is feasible in this domain as well. Similarly, a related limitation of our dataset is that it that 97% of our articles came from either the U.S., the U.K., Canada, New Zealand, or Australia (due to

the focus on English articles). It is reasonable to assume that patterns in other countries might be different, and future work should examine these patterns.

The issue of survivor bias should also be considered. It is likely that navigation technologies have played a role in a significant number of deadly accidents for which there was no witness or exogenous information to identify the role of the technology (the 44 deadly incidents considered here had one or both of these). Interestingly, survivor bias could counteract the fatality bias discussed above.

2.7. Conclusion

In this chapter, we have extended prior work on user interaction with navigation technologies to consider catastrophic incidents associated with these technologies. We have characterized key patterns that exist in these incidents and enumerated implications for research and design that emerge from these patterns. This research increases our understanding of how the navigation technologies that we design cause serious harm, as well as provides a path towards developing safer navigation technologies.

CHAPTER 3

Mitigating the Risks of Personal Navigation Systems by Adding Road Safety Awareness

Note: The work in this chapter was submitted to MobileHCI 2020 but was rejected. The reviewers did not question validity of the methodology or the results but maintained that the topic of this work is unfit for MobileHCI. Necessary modifications have been made to situate the original content in the context of this thesis. Special thanks to Runsheng Xu, a previous master student of Computer Science at Northwestern University who contributed significantly to the design and implementation of the machine learning experiments in this chapter.

3.1. Transition

Work in Chapter 2 analyzed the catastrophic incidents related to the use personal navigation systems and identified a series of technological causes. Chapter 3 begins the process of addressing of these technological causes. In particular, Chapter 3 focuses on one of the most prominent technological causes - many road-safety related attributes are missing or incorrect in personal navigation systems. In other words, personal navigation systems are not aware of if the roads they routed users unto are safe or not. As such, work in Chapter 3 takes the first step towards adding road-safety-awareness into personal navigation systems.

3.2. Introduction

While reliable in most cases, personal navigation technologies (e.g., Google Maps) occasionally route drivers to dangerous roads, resulting in catastrophic incidents [48]. As we just mentioned, work in Chapter 2 identified 47 catastrophic incidents in which navigation systems directed drivers to unpaved roads, narrow lanes, and roads with low clearance, causing drivers to get stranded, to crash with roadside objects, and to hit overpasses. In these incidents, drivers bore financial losses, suffered severe injuries, and, in extreme cases, lost their lives.

One key reason behind these catastrophic incidents is that personal navigation systems usually lack road safety information and thus assume all roads are equally safe [187]. As such, navigation systems might generate routes that prioritize other parameters such as travel time over travel safety, resulting in short-yet-dangerous routes. Indeed, many short-yet-dangerous route recommendations that potentially lead to crashes can be found on Google Maps, the most popular personal navigation system used by over 50 million people every day [18]. For instance, Figure 3.1 compares two routes with the same origins and destinations generated by Google Maps in a rural area of Hungary. Without road safety information, Google Maps recommended a route that is faster but more dangerous (highlighted). The roads on this route are narrow and poorly surfaced, which, according to transportation research [29], are more dangerous for drivers. In comparison, the alternative route (gray) is slower but safer - it passes through roads that are wide and well-maintained.

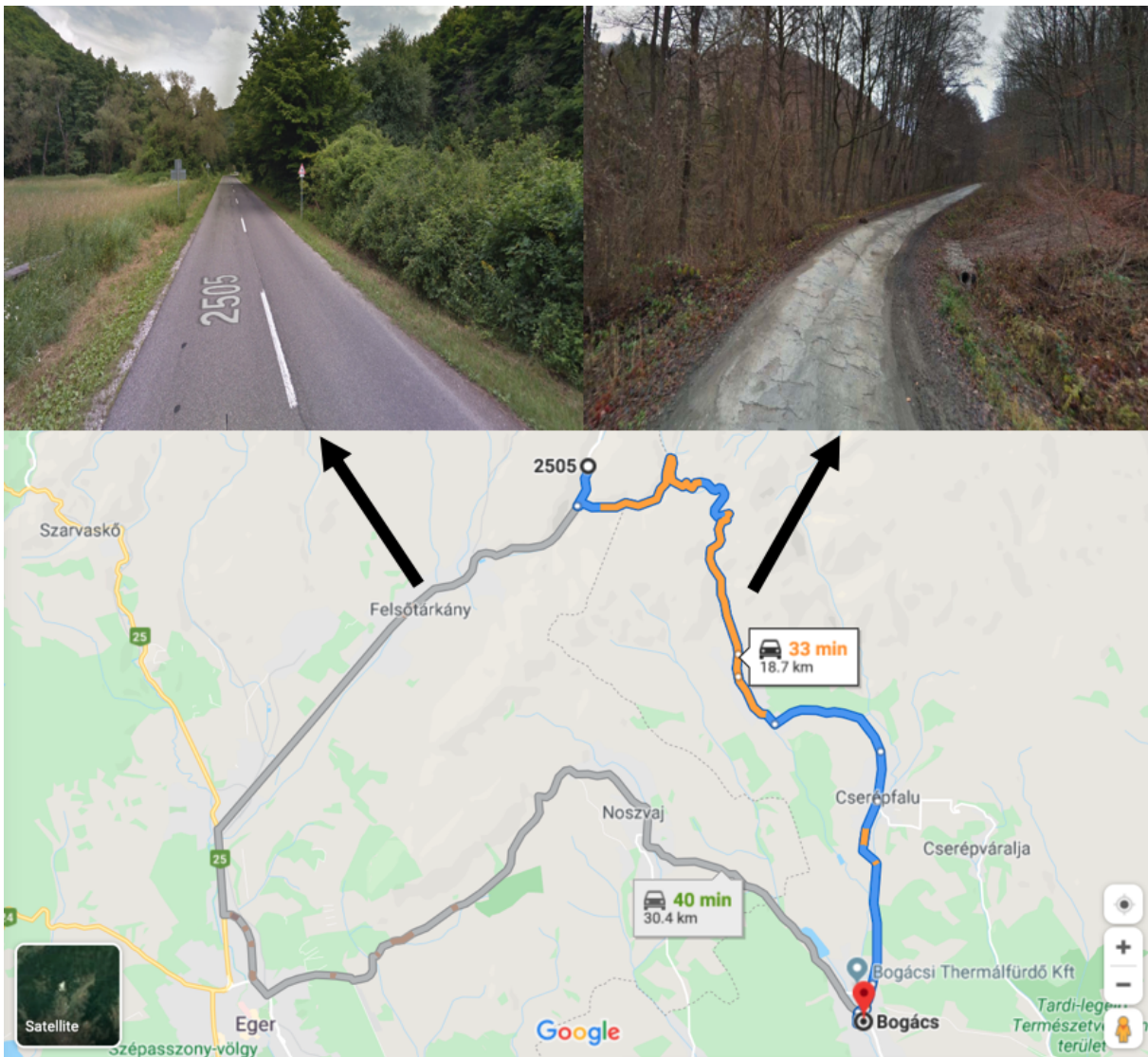


Figure 3.1. Without road safety information, Google Maps recommended a faster but more dangerous route (highlighted). By contrast, the alternative route (grey) is slower but safer. Images are from Google Maps and Google StreetView.

To help navigation systems generate safer routes for drivers, we sought to begin the process of adding road safety awareness into personal navigation devices. We accomplished this goal in two stages. First, we developed a straightforward road safety definition that personal navigation systems can easily understand by adapting well-established

safety standards from transportation studies. Specifically, we focused on the Star Ratings standard from International Road Assessment Programme (iRAP), which gives a single measure of the "built-in" safety for every 100-meter road segment (i.e., safety measured by a road design alone) [143]. Based on the Star Ratings standard, we created a straightforward binary definition for high-risk roads and safe roads, which navigation systems can smoothly incorporate.

In the second stage of our work, we utilized machine learning to produce large-scale predictions using our binary road safety definition. Although the road safety labels we defined can be directly converted from the Star Ratings scores, existing Star Ratings scores only cover a small percentage of roads, which is insufficient for navigation systems. To increase the coverage of road safety labels, we developed a classifier that automatically predicts high-risk roads and safe roads. Our classifier relied on a key insight: many road attributes that are highly relevant to road safety are with open access and can be used as features to predict the road safety labels. Leveraging open data from geographic peer-production projects and government transportation agencies as features, we built and tested our road safety classifier with straightforward machine learning algorithms in four different countries (Greece, Slovenia, Croatia, and the Netherlands). Our results show that our model achieved useful performance, consistently outperforming baseline approaches.

We continue below with an overview of related work, followed by a description of our road safety definition in the context of navigation systems. We then describe the construction and evaluation of our road safety classifier, which predicts the road safety labels according to our definition. Finally, we discuss how our road safety classifier may

be extended to regions beyond the four countries studied in this chapter, and how road safety awareness can enable new technology designs that significantly enhance the safety of drivers who use navigation systems.

3.3. Related Work

3.3.1. Risks Associated with Navigation Systems

Many HCI studies have investigated the risks associated with personal navigation devices. While early work primarily focused on identifying (e.g., [169, 19, 145]) and addressing (e.g., [282, 25, 232]) risks related to distraction introduced by the navigation systems, recent studies have highlighted other types of risks, including the risks of recommending unsafe routes. Most notably, Lin et al. [187] analyzed 158 catastrophic incidents associated with personal navigation systems and found that 47, or about 1/3 of the events they collected, were attributed to dangerous routes recommended by the navigation systems. For example, Lin and colleagues found that personal navigation systems sometimes routed drivers through roads that are unpaved, narrow, or with low clearance, causing vehicles to get stranded, to crash with roadside objects, or to hit overpasses. Furthermore, Lin et al. also highlighted a set of road attributes, including road surface, lane width, and vertical clearance, that are critical for navigation systems to consider in order to generate safe routes for drivers.

Our work has a twofold relationship with the work by Lin et al. First, our work directly addressed the risks associated with dangerous routes identified in their work by adding road safety awareness into navigation systems. Second, as we show later, the road

attributes that Lin et al. recognized as critical to road safety informed how we defined and predicted road safety for navigation systems.

3.3.2. Computational Approaches to Road Safety Assessment

Various sub-fields of computer science have worked to automate aspects of road safety assessment. One popular approach is the image-based approach, which applies computer vision techniques on road images to estimate individual road features needed for safety assessment [161, 188, 304]. Our work is most directly motivated by Song et al. [269], who applied convolutional neural networks and multi-task learning to automate the Star Ratings road assessment. By analyzing street-level panorama images, their system can estimate a variety of road safety features such as lane width, number of lanes, and road curvature, and can predict the overall safety score defined in Star Ratings standard. Although our work is not directly comparable to Song et al. due to data licensing issues and different problem formulations, these two sets of results together highlight the potential to combine imagery and crowdsourced data for even better road risk prediction in the future (we return to this point in Discussion).

3.4. Defining Road Safety for Navigation Systems

In this section, we detail how we defined road safety for navigation purposes. We first describe how we selected an appropriate safety standard from transportation studies and then explain how we simplified this standard to create a straightforward definition that navigation systems can easily understand.

Road safety standards developed by transportation researchers can be broadly classified into two types. The first type of standard is based on crash statistics (e.g., [262, 6, 111]), which computes a road safety score as the ratio of the number of crashes to some measure of exposure such as the population in a region. The second type of standard is road attribute-based (e.g., [117, 143]) and computes the road safety score as a polynomial of many design attributes of the roads, including but not limited to width, surface, and vertical clearance.

We selected the most widely used road attribute-based standard - the Star Ratings standard [143] - as the basis for our road safety definition because this standard is particularly appropriate for the context of navigation systems. Proposed and maintained by the International Road Assessment Programme (iRAP) ¹, the Star Ratings standard has been partially implemented in more than 100 countries around the world. The Star Ratings standard calculates a single score ranging from 1-star (most dangerous) to 5-star (safest) using a formula that involves an extensive list of safety-related road design attributes. The Star Ratings standard is particularly attractive to our purpose because its definition incorporates all road attributes that Lin et al. identified as the major contributors to GPS-related catastrophic incidents [187], such as width, surface, and vertical clearance of the roads.

However, Star Ratings standard can not be directly used by navigation systems because the five-star systems are difficult to interpret for navigation purposes. As a standard created for transportation studies, one of the main purposes of the Star Ratings is to offer detailed and accurate information to evaluate the cost of road improvement projects

¹<https://www.irap.org/>

[142]. As such, while the five-star system in the Star Ratings is perfect for differentiating the amounts of money required for the road improvement projects, it does not offer interpretations about which roads are desirable for drivers to travel - a critical piece of information for navigation systems.

To address the lack of interpretability issues, we simplified the Star Ratings standard to create our road safety definition. Among the five levels in the Star Ratings system, the three-star level is a critical threshold. Most notably, three-star is considered as an economical criterion to ensure a desirable safety level by iRAP [143] and has been adopted by the World Bank and other government agencies as a requirement for the roads construction projects [223] around the world. In other words, roads with three stars and above are considered to have reached the minimum safety for traveling purposes.

Leveraging this critical threshold, we collapsed the five-star system in the Star Ratings into two classes and created the following binary definitions for road safety:

- **Safe roads:** road segments with three stars and above
- **High-risk roads:** road segments with one or two stars.

3.5. Predicting Road Safety

Using the above binary definition of road safety, it is trivial to convert the existing Star Ratings data into road safety labels. However, the resulting road safety labels have one significant limitation - they do not cover road segments without existing Star Rating scores. In fact, due to the high implementation cost of the Star Ratings standard which requires manually labeling road design attributes from images of the road, the existing Star Ratings scores are extremely sparse. For example, according to iRAP's official statistics

[143], even in regions that have the best implementation such as Europe, only an average of 5.4% of road segments are rated with Star Ratings scores. As such, any binary road safety labels generated from these Star Ratings scores do not have sufficient coverage for personal navigation technologies.

To scale up the coverage of our road safety labels, we proposed and evaluated a machine learning-based road safety classifier that automatically predicts our binary road safety labels for road segments. As such, the structure of this section follows the best practices of applied machine learning research (e.g., [49, 53, 9]). We first define the feature set used in our road safety classifier. We then summarize how we collected the ground truth datasets, including both the features and the road safety labels, and highlight the unique characteristics of these data. Finally, we detail the training and evaluation strategies and find that our classifier achieves useful performances that outperform the baseline approaches.

3.5.1. Feature Definitions

Our road safety classifier is only as powerful as the features it leverages. Our choice of features is grounded in well-established theories from transportation studies that demonstrated these features' high relevance to road safety. In addition, to facilitate other researchers to replicate our results and leverage our approach, we purposeful only selected features that come from two publicly accessible sources - geographic peer-production projects and government transportation agencies.

3.5.1.1. OpenStreetMap Features. The most important features used in our road safety classifier are the features from OpenStreetMap (OSM). As known as the "Wikipedia

of maps” [63], OpenStreetMap is a highly successful geographic peer-production project in which volunteers gather to produce free and accurate data for all geographic entities, including roads, for the entire world. Geographic data from OSM have been widely used in a variety of commercial-level map products such as Bing Maps and Facebook Live Maps [64].

The 12 OSM features we selected, shown in the top section of Table 3.1, can be broadly classified into two categories. The first category includes the OSM data that directly map to road attributes used in the Star Ratings standard. Because our road safety label is defined based on the Star Rating standard, we hypothesized that these OSM features would be highly relevant to our road safety labels. We manually checked the definitions of over 100 road attributes defined by OSM community [62] and identified eight such OSM features such as maximum speed, road surface, and road lighting condition.

The second category of OSM features that our road safety classifier used characterizes the accessibility from the location of a road segment. Accessibility in transportation planning describes how easy it is to reach other areas from a given location [191]. We selected features about accessibility because prior research in transportation studies has established that places with lower accessibility are associated with poorer road safety [114, 13, 224]. In this work, we focused on accessibility metrics that solely depend on roads’ topological structure. Specifically, we used isochrones, which show the area a person can reach from a given location within a given period [225]. We computed the 15-minute isochrone from the midpoint of the road segment and included the area of the isochrone, the total population within the isochrone, and the reach factor (detailed in Table 3.1) as our features.

3.5.1.2. Traffic Statistics Features. In addition to features from OSM, we also included traffic statistics as features. We do so because prior research in transportation studies found that traffic statistics, such as the operating speed and traffic volume, are highly correlated to road safety [100]. The bottom section of Table 3.1 details the traffic statistics features, including annual average daily traffic, the mean operating speed of cars, the 85% operating speed of cars.

However, compared with the OSM features, traffic statistics features suffered from a major drawback. While OSM data around the world are accessible through a centralized repository maintained by the OSM community², the accessibility of traffic statistics highly varies - local and national government agencies make own decision on whether to publish the data or not.

We took this ecological validity concern into account when designing our machine learning experiments. Specifically, we trained and evaluated two versions of our classifier - one with all features and the other with OSM features only - in order to gain a comprehensive understanding of the ecological validity of our model. As we show later, parsimonious models with only OSM features performed almost as good as the models trained on all features, highlighting that our model is robust in areas where traffic statistics are unavailable.

3.5.2. Ground Truth Datasets

3.5.2.1. Building Ground Truth Datasets. We prepared four ground truth datasets, each representing a sample of road segments from the following four countries - Croatia,

²https://wiki.openstreetmap.org/wiki/Downloading_data

Source	Feature Name	Definition
OSM Features	highway	one-hot encoding indicating the type of the road
	oneway	a binary value of whether this road segment is one way
	maxspeed	an integer for the maximum speed of this segment
	surface	one-hot encoding of the surface material
	smoothness	an integer for road surface quality level
	lit	a binary value of whether street lights equipped
	bridge	a binary value of whether there is a bridge above or below this road segment
	lanes	an integer for the number of lanes
	nodes	an integer for the number of nodes nearby the road segment
	iso area reach factor	a float value that measures the complexity of the road network around this road segment [108]
Traffic Statistics Features	totpop	an integer for the population covered by the isochrone
	direction	the driving direction of this segment
	aadt	an integer for the annual average daily traffic of the road segment
	mean speed of cars	an integer for vehicle mean operating speed
	85% speed of cars	an integer for the vehicle 85th percentile

Table 3.1. Features for our road safety classifier

Slovenia, Greece, and the Netherlands. As we show later, these four countries were selected because they represent countries with relatively different OSM feature qualities. By conducting experiments on ground truth datasets with varying feature qualities, we gain a deeper understanding of how well our road safety classifier will generalize to other regions.

To build these four ground truth datasets, we first generated binary road safety labels. To do so, we downloaded the existing Star Ratings data in these four countries, including both the coordinates of the road segments and the Star Rating scores of these road segments, from iRAP website [143]. We then converted these Star Rating scores to binary road safety labels.

Class	HR	GR	SL	NL
High-risk	11.71%	8.39%	16.31%	8.00%
Safe	88.29%	91.61%	83.69%	92.00%
Total	6320	41582	22873	55985

Table 3.2. The summary statistics for road safety class labels in four ground truth datasets. Our statistics show that two road safety classes are highly imbalanced. (HR: Croatia; GR: Greece; SL: Slovenia; NL: The Netherlands)

Feature	HR	GR	SL	NL
oneway	9.2%	23.3%	12.03%	19.7%
maxspeed	33.2%	62.7%	41.0%	51.4%
surface	35.7%	43.1%	44.9%	50.3%
smoothness	65.3%	78.5%	68.2%	73.5%
lit	41.3%	47.2%	45.8%	47.6%
average	36.94%	50.96%	42.38%	48.5%

Table 3.3. The missing values rates for OSM features. Features that are not shown in the table do not have missing values problem. Our statistics show that the missing values rates differ among different countries. (HR: Croatia; GR: Greece; SL: Slovenia; NL: The Netherlands)

For each road segment in our ground truth datasets, we obtained the features data from a variety of sources. To obtain the OSM feature, we queried the OSM-related APIs [234, 256] using the coordinates of the road segments. Concerning traffic statistics features, we obtained them from government websites and iRAP websites (which pre-compiles traffic data from the government.)

3.5.2.2. Data Characteristics and Preprocessing. Our ground truth datasets have certain characteristics that require additional preprocessing. First, the two classes for road safety are imbalanced. Table 3.2 shows that, consistently across all four datasets, only a small fraction of the road segments (8% - 17%) are in the high-risk class, presenting challenges for some machine learning algorithms [118]. As such, we applied a data re-sampling technique, a common countermeasure for handling class-imbalance problems

in applied machine learning research (e.g., [10, 242, 218]). Specifically, we applied the commonly-used Synthetic Minority Over-sampling Technique (SMOTE) [51] on our training set, which over-samples minority class by creating "synthetic" samples that fall within the proximity of real samples.

Second, as shown in Table 3.3, OSM feature qualities vary among four ground truth datasets. Like many other peer-production projects such as Wikipedia, OSM does not stipulate mandatory attributes. Instead, OSM gives contributors the freedom to add as many, or as few attributes. Our analyses show that, while most of the OSM features we defined have high data completeness, a handful of features, especially features that represents road attributes, suffer from missing values problem. More importantly, analysis presented in Table 3.3 indicates that, for these OSM features, the specific rates of missing values are different for different countries we studied, a trend that is consistent with prior studies on OSM data quality [150, 12, 112]. For example, while the rates of missing values are lower in the Croatia (avg. 37%) dataset and the Slovenia dataset (avg. 42%), they are considerably higher for Greece dataset (avg. 51%) and the Netherlands dataset (avg. 49%). To partially address the varying degree of missing data, we applied the SOFT-IMPUTE algorithm [205], a data imputation technique that employs a soft-thresholded singular value decomposition to calculate the missing elements in the feature matrix iteratively.

3.5.3. Training and Evaluation

We trained our road safety classifier with a variety of machine learning algorithms, including both ensemble models and conventional models. With respect to ensemble models, we

used an XGBoost model [54], a gradient boosting tree framework that has been shown to be effective in a variety of applied machine learning tasks (e.g. [127, 52, 251]). In addition to the ensemble model, we also employed a number of traditional classifiers, including Logistic Regression, Random Forest, and Decision Tree. Our experiments revealed that XGBoost consistently outperformed other machine learning algorithms. As such, we report only the classification accuracies for XGBoost in the Results section.

To build training and testing sets, we can not use standard k-fold cross-validation due to the particular nature of our dataset. Road segments, like many other types of spatial data, are adjacent to each other. As such, training and test sets generated by standard k-fold cross-validation will be correlated, violating the independent identically distribution (i.i.d) assumption fundamental to many statistical machine learning algorithms [261]. As such, we adopted a strategy from Song et al. [269] to maximize the independence between the samples in the training set and the samples in the test set. Specifically, after building the initial training and test sets with the standard 10-fold cross-validation, we swapped a handful of samples between training sets and test sets to ensure that samples in the test set are at least 500 meters away from the samples in the training set.

For evaluation, we report the average precision, recall, and F1 score using 10-fold cross-validation with the above spatially-aware strategy. We selected precision, recall, and F1 score as performance metrics because these metrics are more informative for evaluating performance on imbalanced data [244]. In addition, we report the performance for high-risk road class and safe road class separately to highlight the different extent of performance improvements for each class. Because this work is the first to define and to attempt to predict road safety labels in the context of navigation systems, we

cannot compare our classifier’s performance with prior work. This is a relatively common scenario when applying machine learning to HCI research due to HCI’s nature as a problem-defining discipline (e.g., [123, 230, 189]). As such, we followed the best practice to compare the performance of our classifier with that of the straightforward baseline approaches [123, 230, 189]. In this work, we leveraged the most powerful baseline approach, which randomly generates predictions by respecting the class distributions in ground truth dataset.

3.5.4. Results

Results in Table 3.4 show our model’s ability to predict safety labels for road segments in our four country-defined ground truth datasets. Our results highlight three high-level trends. First, across the board, our road safety classifier consistently outperformed the baseline approach on all metrics. In particular, the performance increase is most significant for high-risk road class - the minority class. For example, on the Netherlands dataset, our road safety classifier that used all features can increase the F1 score on the high-risk road class from 0.07 to 0.62, a nine-fold increase. Similarly, considerable performance increase (at least five-fold increase) can be found in the results for high-risk road class in other datasets. Even for the safe road class, which already has very high baseline accuracies, our road safety classifier still consistently make smaller improvements. For example, on the Greece dataset and the Netherlands dataset on which our classifier can achieve over 0.9 baseline F1 scores for safe road class, our classifier trained on all features still further increase the F1 scores by about 0.05. Overall, these results demonstrate the clear benefits of applying our road safety classifier - for areas without existing road safety

labels, applying our road safety classifiers offers predictions that are consistently better than baseline approaches. On the Croatia dataset, our classifier even reaches near-perfect predictions.

The second trend in Table 3.4 is that although our classifier consistently outperformed the baseline, we notice that our classifier’s absolute accuracies, especially for the high-risk road class, vary among different ground truth datasets. For example, our classifier achieves high accuracies with over 0.8 F1 scores for high-risk road class on the Croatia dataset and on the Slovenia dataset. In comparison, F1 scores for high-risk road class on the Greece dataset and the Netherlands dataset are moderate (0.55). One possible explanation is the different OSM feature qualities in these four countries - recall Table 3.3 shows that the Greece dataset and the Netherlands dataset have high rates of missing values (50%) for some OSM features, while the Croatia dataset and the Slovenia dataset have lower rates of missing values (40%). Although our classifier’s performance on the lower-end still outperformed the baseline significantly, we suspected that the performance could further decrease when applied to regions that have even higher missing values rates. We return to this point in the discussion section.

Table 3.4 also shows that the parsimonious models using only OSM features perform almost as well as the models using all features. Our results show only a negligible decrease in the F1 score for most of the experiments. The fact that parsimonious models are able to perform equally well demonstrates the robustness of our approach - even in areas where local government does not provide traffic statistics, the performance of our road safety classifier will not significantly decrease.

Table 3.4. Our road safety classifier’s performance on the Croatia dataset. The performance of the baseline approach is in the bracket. (HR: Croatia; GR: Greece; SL: Slovenia; NL: The Netherlands)

	Class	Features	Precision	Recall	F1-Score
HR	High-risk	All	0.95 (0.13)	0.94 (0.12)	0.94 (0.12)
		OSM	0.93 (0.13)	0.94 (0.12)	0.93 (0.12)
	Safe	All	0.99 (0.88)	0.99 (0.89)	0.99 (0.88)
		OSM	0.97 (0.88)	0.98 (0.89)	0.98 (0.88)
GR	High-risk	All	0.58 (0.08)	0.45 (0.08)	0.51 (0.08)
		OSM	0.57 (0.08)	0.44 (0.08)	0.50 (0.08)
	Safe	All	0.95 (0.92)	0.97 (0.92)	0.97 (0.92)
		OSM	0.93 (0.92)	0.96 (0.92)	0.96 (0.92)
SL	High-risk	All	0.79 (0.16)	0.83 (0.16)	0.81 (0.16)
		OSM	0.78 (0.16)	0.81 (0.16)	0.80 (0.16)
	Safe	All	0.95 (0.83)	0.95 (0.83)	0.95 (0.83)
		OSM	0.94 (0.83)	0.95 (0.83)	0.94 (0.83)
NL	High-risk	All	0.74 (0.07)	0.53 (0.08)	0.62 (0.07)
		OSM	0.71 (0.07)	0.44 (0.08)	0.54 (0.07)
	Safe	All	0.96 (0.92)	0.97 (0.92)	0.97 (0.92)
		OSM	0.93 (0.92)	0.96 (0.92)	0.94 (0.92)

3.5.5. Estimating Exposure to Unsafe Routes

While prior studies (e.g., [187]) have discovered the risks associated with unsafe routes given by navigation systems, the prevalence of the unsafe routes remains unknown. With our road safety classifier, we can, for the first time, begin to quantify how often users of navigation systems will be exposed to unsafe routing results.

To do so, we conducted a small case study in Croatia, where our road safety classifier has the best performance. To account for the potential differences between urban and rural areas, we selected two regions - Croatia’s capital Zagreb (roughly 900 km²) and an equal-sized rural region in Croatia. We then generated a large number of routes for these two areas. Specifically, we followed the best practices in routes evaluation studies (e.g.,

Table 3.5. The percentage of unsafe routes among randomly sampled routes in Croatia by types of area and types of route options. The criteria for unsafe routes specify the minimum percentage of high-risk road segments a route needs to include for this route to be considered an unsafe route. For example, “>10%” means that routes contain at least 10% of high-risk road segments are unsafe routes.

Area	Option	Criteria for Unsafe Routes			
		>10%	>30%	>50%	>70%
Urban	Shortest	90.4%	67.4%	30.0%	7.6%
Urban	Fastest	90.6%	66.5%	32.3%	8.7%
Rural	Shortest	98.8%	85.6%	52.9%	25.2%
Rural	Fastest	99.1%	85.4%	52.3%	26.1%

[147, 116, 307]) which first randomly sample origin-destination pairs and then generate routes between these origin-destination pairs. Although this approach has its limitations, it has been used in lieu of actual trip data, which is difficult to obtain outside of the context of large personal navigation technology companies. In our case, we sampled 10,000 origin-destination pairs for each region and used Mapzen Valhalla [1], a popular open-sourced routing service, to generate the shortest routes and the fastest routes between these origin-destination pairs. Finally, we applied our road safety prediction model on the road segments included these 20,000 routes to quantify the unsafe routes.

Table 3.5 shows the estimated percentages of unsafe routes by type of regions and type of route options. In addition, we tested different criteria for defining a unsafe route by varying the minimum percentage of high-risk road segments the route contains. The results in Table 3.5 suggest that users of navigation systems have high exposure to potentially unsafe routing results. Specifically, if routes that have at least 10% of their road segments as high-risk road segments are considered unsafe routes, over 90% of routing results given by navigation systems are potentially dangerous in Croatia, regardless of the

type of regions and the type of route options. This number might look startling, but it is unsurprising given that we found that 11.71% of road segments in Croatia are high-risk (see Table 3.2.) Even if we use a lenient criteria for unsafe routes - routes with at least 70% of high-risk road segments - experiment results still show that as much as a quarter of routing results are potentially dangerous in rural areas of Croatia. While driving on these potentially dangerous route will not necessarily result in accidents, users' high exposure to potentially unsafe routes still highlights the prevalence of the risks.

A second trend in Table 3.5 is that while route options have little influence over the prevalence of unsafe routes, urbanness/ruralness do affect the percentage of unsafe routes. Specifically, using navigation systems in rural areas of Croatia will be exposed to 8% to 20% more unsafe routes than doing so in urban areas. While the higher percentage in rural areas can be partially attributed to the fact that randomly generated routes might be biased towards more local roads over highways (which draws most of the traffic in the rural area), our results still highlight the higher risks of driving on rural roads which is consistent with crash statistics from National Highway Traffic Safety Administration [290].

3.6. Discussion

Our results demonstrate that using straightforward machine learning algorithms and publicly available geographic data as features, we can successfully predict road safety labels with accuracies that are significantly higher than the baseline approach. In the following, we first discuss the factors to consider when extending our road safety classifier to regions beyond the areas studied in this chapter. With an eye towards future work, we

then discuss various technology designs through which our road safety predictions could help personal navigation technologies reduce the risks of dangerous routes.

3.6.1. Extending our approach to other regions

Our results show that, while the performance of our classifier consistently beat the baseline, its performance in a region might be affected by the OSM feature quality in that region.

As such, when applying our road classifier to other regions, one important factor to consider is the quality of OSM features in this region. Researchers have found that OSM data quality is usually better in urban/developed areas than in rural/undeveloped areas [148, 308, 202]. As such, our road safety classifier might have higher performances for more developed regions such as Europe and have lower performances for less developed regions, such as South Asia and Africa. Similarly, urban areas, which have higher OSM qualities, might observe higher performance than the rural areas. As such, future work that extends our classifier to other regions should examine the specific OSM data quality, especially the rates of the missing value, before applying our approach. Moreover, in order to improve the accuracy in the areas with low OSM data quality, future work should explore additional features beyond OSM and traffic statistics or more advanced machine learning models.

3.6.2. Integrating Road Safety Predictions into Navigation Systems

Our work can produce large-scale reliable road safety predictions which navigation systems can directly leverage. These road safety predictions present many opportunities for

designing new algorithms and user interactions for protecting drivers' safety. Below, we briefly mention the opportunities for new algorithm designs and then focus our attention on new user interactions that ensure driver safety.

For algorithm innovation, one clear opportunity is building new routing algorithms that prioritize road safety. As a starting point, one may simply adopt the existing algorithms designed for generating safe routes that pass through low-crime areas (e.g., [158, 89, 257]). For example, the crime-index score in the cost functions of these safe routing algorithms can be replaced by an aggregated road risk score of the candidate route. Alternatively, instead of adding waypoints in low-crime areas, these safe routing algorithms can add waypoints that detour through a safe road. However, one potential externality of this new type of routing algorithms is their impact to the local communities that are excluded from (or included in) the alternative routes. As such, when developing these routing algorithms that prioritize road safety, developers should conduct experiments to understand and measure this externality. Work from Johnson et al. [147] illustrates different approaches for how to conduct such experiments.

Our road safety prediction also enable new safety-oriented users interaction of the personal navigation technologies. Even with new routing algorithms that prioritize road safety, drivers might still unavoidably travel on high-risk roads given their time constraints or their choices of destinations. In these scenarios, road safety awareness-enabled user interactions serve as the critical last layer of protection. These essential user interactions can happen in two phases - during route recommendation and while driving. During route recommendations, navigation systems can put users in control of the trade-off between road safety and other parameters. To begin with, as a preference setting, navigation

systems can explicitly ask for the maximum distance or portion of high-risk roads that users are willing to tolerate within a route and incorporate this constraint when running the routing algorithms. In addition, navigation systems can explicitly communicate the trade-off between time and route safety among alternative routes through routes overview. For example, in addition to informing users of the distance and time of the candidate routes, navigation systems can highlight the distance or the portion of high-risk road segments through textual or visual communications.

During the trip, when passing through high-risk roads, drivers must focus their entire attention on the road [133]. As such, HCI researchers have called for navigation systems that can timely alert drivers for abnormal road environments (e.g., [48]). With our road safety predictions, navigation systems will know precisely where the high-risk road segments are. Given this knowledge, navigation systems can generate audio and visual alerts when drivers are approaching the dangerous road segments to call drivers' attention to the road. With cars becoming increasingly automated and drivers delegating more responsibility to the technology, navigation systems with road safety awareness will be even more vital to ensure user safety.

3.7. Conclusion

A dangerous assumption adopted by existing navigation systems is that all roads are equally safe. Navigation systems with this wrong assumption have routed drivers to dangerous routes, threatening driver safety. In this chapter, we sought to address this risk by adding road safety awareness to the navigation systems. Specifically, we defined road safety in the context of the navigation systems and proposed a road safety classifier

to automatically predict the road safety for many roads using out-of-the-box machine learning algorithms and diverse features from public geographic data. We demonstrate we can predict the road safety labels with accuracies that are significantly higher than baseline approaches. Finally, we discussed the factors to consider when extending our road safety classifier to other regions and the novel safety designs on navigation systems enabled by our road safety predictions.

Part 2

Geographic Inequalities in Location-Aware Artificial Intelligence Systems

Part 1 of this thesis described our effort to understand and mitigate safety issues in personal navigation systems. In Part 2, we focus on a different type of risk – geographic inequalities in the location-aware artificial intelligence systems.

Location-aware artificial intelligence (AI) technologies, also referred to as “GeoAI technologies” by the research community [135, 198, 136, 144], are AI technologies applied to the locational data. Similar to other application domains of artificial intelligence, a key characteristic of location-aware AI technologies is that they usually heavily rely on state-of-the-art models or methods in *core artificial intelligence fields* such as natural language processing, computer vision, and commonsense knowledge base. In the specific context of location-aware AI technologies, these technologies leveraged advancements in core artificial intelligence fields to acquire, process, and serve a variety of locational data, such as geo-referenced images, videos, and articles. With the large-scale available of these locational data, location-aware AI technologies have been widely applied to help solve critical tasks in society. Such examples include identifying forest fires from geo-referenced satellite images [239], extracting socioeconomic attributes about neighborhoods by analyzing street view images [97], and updating official gazetteer [247] from user-generated textual content containing places.

However, an emerging risk of these highly beneficial location-aware AI technologies is their geographic biases. Recall in Section 1.2.2.2, we defined the geographic biases of location-aware technologies as the phenomena that the performance of these technologies, usually measured as the precision in classification tasks, varies when the technologies are applied to locational data from different types of geographies. In practice, GeoAI

researchers and practitioners have reported a number such geographic biases on location-aware AI technologies. For example, it has been found that remote sensing image classification models have poorer performance on identifying object in rural areas like farmland than in other places [250]. Similarly, models to recognizing buildings from street view images have lower performance in developing countries than developed countries [258]. Given that these location-aware AI technologies have been applied to solve important societal problems, the geographic biases in these technologies can potentially can resulted in unfair regional inequalities.

While the geographic inequalities in AI-based location-aware technologies can be produced through many different mechanism, we will focus on one mechanism that is particularly relevant to the Part 2 of this dissertation - the propagation of geographic inequalities from the large-scale general training data.

To understand this mechanism, it is helpful to review how most of the AI-based location-aware technologies are built. Like most of the application domains of artificial intelligence, location-aware AI technologies are often built by taking an domain-agnostic artificial intelligence model trained on general tasks, such as a pre-trained convolutional neural network model widely used in computer vision research, and adapt it with additional domain specific data or expertise. To trace back one more step, these domain-agnostic artificial intelligence models are often built by training a model on large-scale labeled training datasets. For example, pre-trained convolution neural networks in computer visions are trained on large-scale human-labeled image datasets such as ImageNet [73] and OpenImages [170].

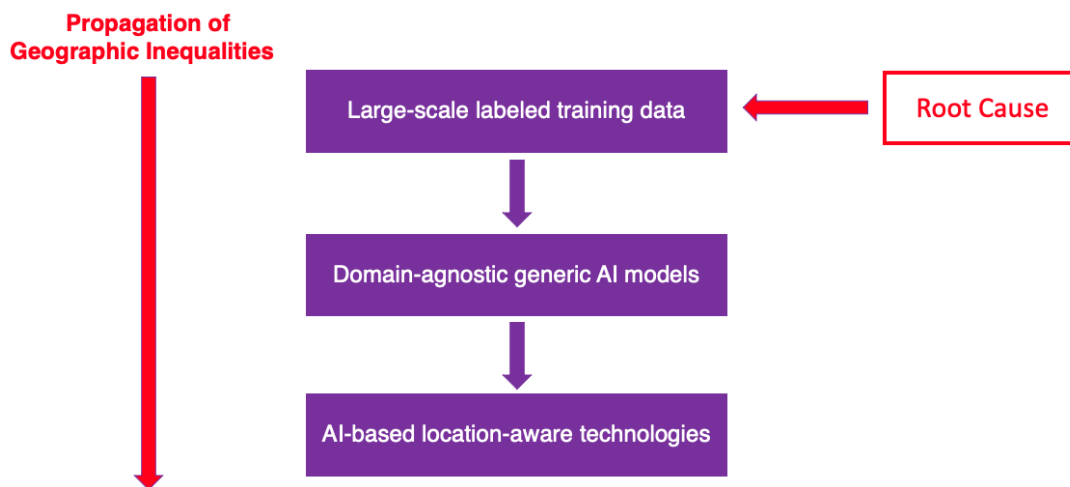


Figure 3.2. The propagation of geographic inequalities from the large-scale training dataset to the AI-based location-aware technologies

This approach of building AI-based location-aware technologies makes it possible that geographic inequalities could enter these technologies by a propagation process as shown in Figure 3.2. That is, if geographic inequalities exist in the large-scale training data at the beginning, they could be propagated, first to the domain-agnostic generic AI models such as pre-trained convolutional neural networks, and finally to the application domains including AI-based location-aware technologies.

Unfortunately, research in ethical AI has shown that this is indeed the case. As we detail later, researchers have shown that large-scale training dataset used in computer vision contain significant geographic inequalities regarding where the images are sourced and have shown that pre-training models trained on these large-scale training datasets inherit these geographic inequalities [258, 72]. Likewise, in natural language processing models, many studies have found similar geographic biases[146, 125] in the widely used pre-trained models such as word embeddings.

Understanding the mechanism through which the geographic inequalities are produced in AI-based location-aware technologies also points out one approach to mitigate this type of geographic inequalities. That is, if we could mitigate the geographic inequalities in the root - the large-scale training dataset, in theory, geographic inequalities will follow the same propagation process and be reduced in the domain-agnostic AI models and finally in the AI-based location-aware technologies.

Part 2 of this dissertation adopts this top-down approach for mitigating the risk of geographic biases in the location-aware AI technologies. In Chapter 4, we describe our effort to mitigate the geographic inequalities in computer-vision-based location-aware technologies. In Chapter 5, we focus on mitigating the geographic inequalities in location-aware technologies that are based on commonsense knowledge base.

CHAPTER 4

Mitigating Geographic Inequalities in Computer Vision-Based Location-Aware AI Technologies

Note: The work in this chapter was submitted to SIGCHI 2020 but was rejected. The reviewers did not question the validity of the methodology or the results but maintained that the topic of this work is unfit for SIGCHI and suggested us to submit to Computer Vision or Machine Learning conferences. Necessary modifications have been made to situate the original content in the context of this thesis.

4.1. Motivation

Many location-aware AI systems leveraged computer vision to understand geo-referenced images [135, 198, 136]. For example, tourism apps have used computer vision models to identify landmarks through phone cameras in order to provide clearer guidance and richer information to users (e.g., [128, 199, 229]). Urban planning systems might use computer vision models to process high-resolution satellite images to identify geographic features such as buildings and road networks (e.g., [303, 280, 270]). In some innovative scenarios, location-aware census tools have used computer vision to classify street-parked cars for demographic estimation (e.g., [97, 95, 96]).

As we previously mentioned, one widely used approach to build location-aware AI technologies is to customize domain-agnostic generic AI models trained on large-scale labeled training data. Figure 4.1 illustrates how this approach has been applied to build

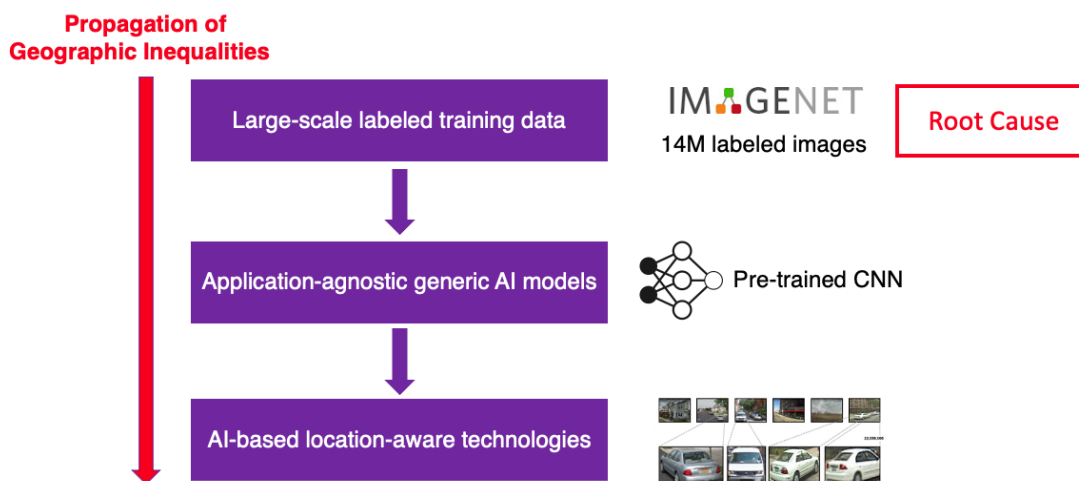


Figure 4.1. The propagation of geographic inequalities from the large-scale ImageNet dataset to the computer-vision-based location-aware technologies

many computer-vision-based location-aware AI technologies. The bottom level of Figure 4.1 represents computer-vision-based location-aware technologies, such as the location-aware census tool that analyzes street-parked cars for demographic estimation (e.g., [97, 95, 96]). The second level of Figure 4.1 indicates that the location-aware census tool is built on top of a domain-agnostic generic AI model - pre-trained Convolutional Neural Network (CNN), which is a deep learning model trained on image recognition tasks that has demonstrated its utility to serve as general image feature extractor for many different downstream computer vision tasks (e.g., [255, 265, 47]), including many in location-aware computing domains (e.g., [214, 8, 154, 200, 245, 300]). Finally, the top level shows that these pre-trained CNNs are built by training on large-scale image dataset, such as ImageNet [73] or Open Images [170], which contains over tens of millions of human-labeled images.

As we discussed before, this approach of building computer-vision location-aware technologies makes the propagation of geographic inequalities possible. That is, if there are

geographic inequalities in the large-scale image datasets, it is very likely that the pre-trained CNN and the computer-vision location-aware technologies will also exhibit geographic inequalities.

Researchers have indeed proved the propagation of geographic inequalities. In computer vision research, Shankar et al. [258] and DeVeries et la. [72] both independently found that large image datasets that are used to train pre-trained CNN models source most of the images from western and developed countries and contain very few examples from the developing countries. In particular, in Open Images [170] and ImageNet [73] - two of the most widely used large image datasets for pre-training purpose, 60% of images are from a limited set of developed countries, including United States, Great Britain, France, Italy, Canada, Spain, and Australia [258, 72].

More importantly, these researchers have also proved that CNN models pre-trained on ImageNet and Open Images inherited these geographic inequalities. For example, Shankar et al. [258] demonstrated that all widely-used pre-trained models have higher accuracy for recognizing "bridegroom" images from developed countries than for recognizing "bridegroom" images from developing countries. Similarly, DeVeries et al. showed that, in general, the accuracy of pre-trained CNN on the same object recognition tasks are higher in wealthier countries [72].

Following the diagram in Figure 4.1, the geographic inequalities may very likely propagate computer vision-based location-aware technologies given that many of these technologies are built on top of the pre-trained CNN models. For example, for a tourist app that relies on pre-trained models to identify church buildings through phone camera, the fact that the pre-trained models have geographic inequalities means that the tourist app

might not function well in less developed countries, where the church buildings look significantly different from western countries [65]. In fact, computer vision-based location-aware technologies developers have already noted some anecdotal evidence of such propagation in street view image classification [79] and remote sensing image classification [250]. Considering the fact location-aware AI are more likely than not to be used at different locations, the harm of the geographic inequalities on the pre-trained models will be maximized, severely impacting the usability of these computer vision-based location-aware AI systems.

Our work in this chapter takes the first step towards mitigating the geographic inequalities in computer vision-based location-aware technologies. As we discussed before, our approach is to directly mitigate the root cause of the problem - the geographic inequalities in the large-scale training dataset. More concretely, we seek to increase the geodiversity of these large image datasets by adding millions of images from developing countries to thousands of classes. To make our solution easily understandable, we describe our solution in the context of ImageNet, the most important and widely used dataset for pre-training purpose [138, 164, 119]. Because classes in different large image datasets have significant overlap[166], we note that our solution can be easily extended to other large image datasets such as Open Images.

Our work in this chapter also provides initial evidence that a more geographically equitable image dataset does lead to a more geographically equitable pre-trained CNN model. To do so, we carefully designed machine learning experiments that shows a CNN model which is pre-trained on an ImageNet enhanced with images from developing countries performs better than the original, gold standard ImageNet pre-trained model in previously

low-performing countries. Finally, we note that given the ImageNet pre-trained model is widely used for many downstream computer vision, the results from this chapter, especially the more geographically equitable ImageNet and the pre-trained CNN models, could potentially benefit a wide range of computer vision tasks beyond those in the location-aware technologies domain.

4.2. Introduction

ImageNet [73] is one of the most widely-used datasets in the computer vision community. With 14 million human-annotated images for 21,000 object classes in WordNet [211], is arguably the one with the greatest significance for research and widest impact for applications. For researchers, ImageNet serves as a gold standard dataset for training and testing hundreds of deep learning-based image classification models [73, 167, 266]. Even more importantly, pre-training on ImageNet to obtain general visual features is a common practice for solving a wide range of vision tasks [138, 259] beyond image classification.

Despite ImageNet’s tremendous success, recent studies have indicated that ImageNet is subject to substantial geographic representations biases, leading to fairness issues in downstream computer vision models. Specifically, Shankar et al. [258] and Misra et al. [72] both independently verified that ImageNet, like many other large image training datasets, contains significantly fewer images from developing countries and the corresponding cultures. These authors further demonstrated that, models pre-trained on ImageNet perform relatively poorly for recognizing images of certain concept from the underrepresented countries. For example, Shankar et al. showed that a classifier pre-trained on ImageNet has substantially lower performance recognizing bridegroom images

from Pakistan than bridegroom images from the United States. Given that pre-training models on ImageNet and then fine-tuning it on domain-specific datasets have become a paradigm to solve many computer vision tasks [118, 164, 302, 138], the lower performance of ImageNet pre-trained models in developing country will impacts thousands if not millions of downstream computer vision applications deployed in developing countries.

ImageNet was constructed using a well-known, standard crowdsourcing approach in HCI research, i.e. paid microtasks distributed on market places such as Amazon Mechanical Turk. However, leveraging this approach to enhance the geographic representations of ImageNet faces significant challenges. Most importantly, considering the uneven distribution of the Turkers geographic backgrounds [141, 75], it is unclear if such an approach could work at all. Using the task of labeling bridegroom images as an example, it may be impossible to recruit Turkers from *every* developing country in order to find and label the bridegroom images from their own cultures. In addition, even assuming one can recruit enough Turkers from diverse geographic backgrounds, crowdsourcing project on the ImageNet-scale would require significant amount of time, money, and human resources [98]. For example, adopting the Amazon Mechanical Turk approach, building the original ImageNet took researchers two and half years and millions of dollars of budget.

In this research, we sought to overcome these substantial barriers by relying on the product of a different crowdsourcing process: volunteer-based peer production. Specifically, we explore the hypothesis that Wikimedia Commons – the central source of images for all language editions of Wikipedia [120] – can offer millions of labeled images from developing countries for thousands of classes in ImageNet. Three characteristics of Wikimedia Commons make it especially attractive for enhancing ImageNet. First, while Turkers are

predominantly from English-speaking countries, the contributors on Wikimedia Commons are from around the world owing to Wikimedia Commons' multilingual nature. Second, it contains over 50 million freely usable images. This number is over three times the size of ImageNet and might already include enough images necessary for enhancing ImageNet, saving the time for running large-scale market-based crowdsourcing project. Third, since the Commons was built with an encyclopedic focus, the topics covered by the Commons' images are extremely diverse, containing over 9,642,912 different categories of images that is more than the 21,841 categories by magnitudes.

However, before we can quantitatively evaluate our hypothesis that images on Wikimedia Commons can enhance the representation of developing countries in ImageNet, we first had to overcome a major obstacle: producing ImageNet-compatible tags for over 50 millions images on the Wikimedia Commons. Without this step, the geographically diverse images from the Wikimedia Commons can not be integrated into the corresponding classes in ImageNet, making these Commons' images useless for our goal of training pre-training models with higher accuracy in developing countries.

To produce ImageNet-compatible tags for the Commons' images, we leveraged the noisy-but-rich category labels collaboratively produced by the Commons editors. Similar to the category labels on Wikipedia, category labels in Wikimedia Commons are user-generated tags that are used to group images of similar topics. Producing ImageNet-compatible tags for Wikimedia Commons images is difficult. First, the Commons' noisy categories need to be matched to WordNet class name. Second, not all categories can be used – tags in ImageNet exclusively describe the content of the images whereas the

Commons' categories often describe non-content information such as the authorship of the objects in the images.

To address these challenges, we developed a pipeline to produce ImageNet-compatible tags for Commons images. Our pipeline consists of two systems, each of which addresses one of the challenges. The first system, our YAGO3 Matcher, mines noisy categories labels and matches them to WordNet classes using YAGO3 [194], a well-known semantic ontology that combines entities from both Wikipedia and WordNet. The second system, our Content-Based Tags Filter, identifies the types of the matched tags and returns only the tags that describe the content of the images.

Using the ImageNet-compatible tags produced by our two-part pipeline, we find strong evidence in support of our hypothesis: images from the Commons can substantially enhance the representation of developing countries in ImageNet. We showed that among the 18.5 million Commons' images that have clear location information, 4.1 million, or 22%, are from developing countries. This means that Wikimedia Commons can *at least* add 4.1 million images from developing countries to ImageNet, which is roughly $\frac{1}{3}$ of the current size of ImageNet. If we assume that the Commons' images without location information has similar representations of developing countries, the Commons has potential to offer roughly 11 millions images from developing country, a number that is similar to the original size of ImageNet. To offer more convincing evidence to support our hypothesis, we used a case study to show that fine-tuning pre-training models on these Commons' images will reduce the geographic biases in these pre-training models by increasing their accuracies in previously underperforming developing countries. Given the wide popularity of the pre-training model, our Commons-enhanced ImageNet dataset can

lead enormous cumulative benefits for lots of downstream computer vision applications targeted for developing countries.

Finally, besides accomplishing our main goal, we opportunistically found two other benefits for combining Wikimedia Commons and ImageNet - extending the breadth and growing the size of ImageNet. For extending the breadth of ImageNet, the Commons could supply sufficient training data for 15% of previously uncovered WordNet classes, potentially enabling computer vision models to learn novel classes. For expanding the size of ImageNet, the Commons' images could double the number of training images for 25% of ImageNet classes with below-average classification accuracy, which likely increase pre-training models' performance on this class according to recent studies on the relationship between the size of training dataset and model's performance [279].

In summary, this chapter demonstrates that the peer-produced images and categories from Wikimedia Commons can offer substantial and critical enhancements to the geodiversity of widely used large image datasets. In addition, we used a case study to demonstrate experimentally that when trained on combining Commons and ImageNet dataset, pre-training models can have significantly higher accuracy for images from previously underperforming developing countries, bring extensive benefits to downstream computer vision that are targeted towards these regions. We are publishing with this chapter the 118 million ImageNet-compatible tags and 18.1 million location tags for these Commons' images that we developed through this research.

4.3. Related Work

4.3.1. Biases in Large Image Datasets

The geographic diversity issues of ImageNet is an example of the biases in large gold standard image datasets, which has draw increasing attention from research community.

Researchers identified two main types of biases in large image datasets. One type of biases is introduced by human who labeled the datasets (e.g., [24, 78, 252]). IN general, it has been shown that human labelers (often crowdworkers) with different cultural backgrounds [78], different gender views [226], and different knowledge specialties [252] will label the same item differently.

Another type of bias, which is what we aim to address in this chapter, is the biases from underlying data distributions. Also called representation bias [149], data distribution biases is the fact that different subgroups of data are unevenly sampled into the dataset (e.g., [74, 125, 285]). Specifically, as we previously mentioned, Shankar et al. [258] and de Vries et al. [72] both independently demonstrated that large image datasets like ImageNet and Open Images have more images from countries with better economic advancement. Similarly, ImageNet team [302] has identified that images in the People subtree of ImageNet has unequal representations on protected demographic attributes including age, gender, and color.

4.3.2. Mitigating Biases in Large Image Datasets

While many studies have focused on identifying the biases in large image datasets, very fewer studies have proposed solutions to mitigate these biases. For mitigating biases introduced by human labelers, one of few solutions is by Barbosa et al. [24], who proposed

a framework to ensure that tasks are allocated to crowdworkers with heterogenous cultural backgrounds.

For mitigating biases introduced by unequal representations, there are only two published work to the best our knowledge. The first project, conducted by ImageNet team [302], focused on mitigating the imbalance of the representations on protected demographic attributes (e.g., gender, race, etc.) They achieved such goal by first acquiring crowdsourced labels for the protected attributes for the images and then removing the images from the overrepresented groups. While effective for addressing the biases in demographic attributes, this approach is unfortunately not applicable for addressing geographic diversity. This is because our problem present a unique challenge - the original ImageNet have *extremely limited, if not zeros*, images for some of the underrepresented countries [258], making it impossible to achieve equality through re-balancing.

More similar to our project, the second project [103] to mitigate unequal representations in image datasets focused on enhancing the country diversity in Open Images, a large image dataset built by Google which suffers similar geographic biases problem as ImageNet. Called *Open Images Extended - Crowdsourced* [103], this project aimed to increase the number of images from developing countries by directly recruiting volunteers from developing countries to capture, annotate, and upload images from their everyday life. While Open Images Extended - Crowdsourced should be applauded for being the first concrete effort to solve geodiversity problem in large image datasets, this project has several limitations. Most importantly, it leveraged a market-based crowdsourcing approach similar to Amazon Mechanical Turk, which has well-known difficulties to recruit participants from a wide range of geographies [141]. Indeed, the statistics published by project

showed that 83% of the images were collected from one country - India. The second limitation is its insufficient data volume. Despite the project takes advantages of Google’s centralized crowdsourcing platform - Google Crowdsourcing¹ which has as many contributors as Amazon Mechanical Turks², the resulting dataset only contains 478k images, which can not bring meaningful geodiversity enhancement given that the original Open Images dataset has over 9 millions images. On a per-class basis, the resulting dataset contains an average of only 80 images for each class, a number that is again smaller by orders of magnitude than the average class size in the original ImageNet (1520).

Compared to the Open Images Extended - Crowdsourced, we used a completely different crowdsourcing approach - Wikipedia-based peer-production. The contributors from our approach are more geographically diverse than those on market-based crowdsourcing platform such as Amazon Mechanical Turks or Google Crowdsourcing [68, 75].

It is important to point out that, despite the previously mentioned limitations, Open Images Extended - Crowdsourced remained the only available images dataset that focus on geographic diversity. As such, our experiment relies on this dataset as the testing dataset (detailed later).

4.3.3. Research about Wikimedia Commons

Wikimedia Commons is a Wikimedia Foundation project that curates freely-licensed media files (videos, images, sound clips) with a goal of providing knowledge to the world

¹<https://crowdsourcing.google.com/>

²Although Google Crowdsourcing does not publish its number of contributors, it is possible to infer the number from its Leaderboard. One contribution will place a user at Rank 20,000 - 40,000 on the leaderboard for one task. Given that there are 10 different tasks, it is likely that the total number of contributors of Google Crowdsourcing is on the scale of 100,000, which is almost the same as the number of workers on Amazon Mechanical Turks as estimated by researchers [75]

[69]. Of the Commons' 52 million media files ³, 90% are images. Each image is associated with peer-produced categories, which can serve as image tags to organize images of similar subjects. Similar to categories on Wikipedia, the categories do form a hierarchy, but the hierarchy is ill-formed and difficult to utilize as a formal ontology [159].

While Wikipedia has been the subject of extensive studies in the HCI community and beyond, Wikimedia Commons has drawn relatively little attention [209]. Among the studies that have focused on the Commons, researchers have lever-aged Commons images to support image retrieval [186, 237], measured the Commons' economic value [81], and de-signed mechanisms to promote participation in the Commons [209].

Most relevant to our study, a series of efforts have focused on producing and improving structured metadata for Commons' images. For example, with the goal of extending the DBpedia [16], a Wikipedia-based knowledge database in Resource Description Framework (RDF) format, Vaidya et al. [286] built a dataset containing RDF triples for all files on Commons. Follow-up work by Ferrada et al. [84] enhanced this dataset with the visual descriptors for these images. While these studies also used peer-produced categories, their main goal was to accurately convert them into RDF triples without using the semantics of these labels. Our study differs from this past work by deeply engaging with the semantics of these labels by matching them to formal ontology YAGO3, classifying their types, and evaluating their potential values for enhancing ImageNet.

³As of September, 2019


4.4. Producing ImageNet-Compatible Tags

In this section, we describe the pipeline we used to produce ImageNet-compatible tags for Commons images by exploiting the peer-produced category labels attached to Commons images. We first provide background about category labels on the Commons. We then describe two high-level challenges associated with the Commons' category labels that motivate the many design choices in our pipeline. Finally, we detail how we implemented the two systems that constitute our pipeline – YAGO3 Matcher and Content-Based Tags Filter – which address these two high-level challenges.

4.4.1. Category Labels on the Commons

As is the case with category labels on Wikipedia, category labels in the Commons appear at the bottom of pages dedicated to each Commons media file. Commons category labels are also contributed and edited in the same fashion as Wikipedia category labels. Figure 4.2 shows an example of a Commons page dedicated to an image and the category labels the image has been assigned.

We obtained all Commons category labels from the Wikimedia Download website [87]. More specifically, we downloaded the `categorylinks` table from early 2019, which contains the mappings between categories labels and each media file on the Commons. We further filter out all mappings that are not associated with image files as defined by the Commons [86], which includes both bitmap and vector images. In the end, we extracted 299 million raw categories for 49.6 million images.



Not logged in [Talk](#) [Contributions](#)

[File](#) [Discussion](#) [View](#) [Edit](#) [History](#)

File:U.S. Sailors assigned to Construction Dive Detachment Alpha, Underwater Construction T dive over the remains of the battleship USS Arizona Joint Base Pearl Harbor-Hickam, Hawaii, March 2013 130321-N-WX059-135.jpg

From Wikimedia Commons, the free media repository

[File](#) [File history](#) [File usage on Commons](#) [File usage on other wikis](#) [Meta](#)



Size of this preview: 800 × 392 pixels. Other resolutions: 320 × 157 pixels | 640 × 314 pixels | 1,024 × 502 pixels | 1,280 × 635 pixels | 1,470 pixels.

[Categories: 2013 in Pearl Harbor | USS Arizona Memorial | Underwater diving in Hawaii](#)

Hidden categories: [File:US military images - CM-DOD uploaded by Red \(when needed\)](#) | [Images from DOD upload](#)

This page was last edited on 31 July 2019, at 18:02.

Figure 4.2. An image in the Commons and its corresponding category labels (indicated by red dotted lines)

4.4.2. Challenges

To produce ImageNet-compatible tags from the Commons category labels, we face two high-level challenges. The first high-level challenge is that peer-produced category labels are known to be very noisy [159, 277]. ImageNet labels, on the other hand, conform to the very-well-structured WordNet ontology. Specifically, we observed that Commons category labels can differ from the ImageNet tags in several ways. First, category labels can be noun phrases (e.g. <paintings of Van Gogh>) but ImageNet labels are strictly nouns (e.g. <painting>). Second, Commons categories can contain named entities while ImageNet labels are usually abstract concepts. This requires mapping, for examples, specific models of cars (e.g., <ZAZ-965A Zaporozhets>) to their hypernyms (e.g. <car>) in ImageNet labels. Third, the Commons categories are multilingual, including labels such as “<中华人民共和国政府工作报告>” (English: Report on the work of the government of the People’s Republic of China ⁴) while ImageNet tags are exclusively in English. We address the challenge of matching Commons categories to ImageNet with our YAGO3 Matcher system, which we detail later.

The second high-level challenge is that categories on the Commons describe much more diverse information than the labels in ImageNet do. According to the tag taxonomy literature [301], labels for images can be classified into the following types:

- (1) *Content-based*: labels that describe the object or the concepts in the images.
- (2) *Context-based*: tags that provide the context (time and location) in which the image was created.

⁴This is the Chinese version of the State of the Union Address.

- (3) *Attribute*: tags of specific human beings who are the authors of the images or the objects in the images. E.g., <Charles_Dickens> tag for an image of his book.
- (4) *Subjective*: tags that express the user’s opinion and emotion, e.g. “funny” or “cool”.
- (5) *Organizational*: tags that are reserved for administrative purposes, such as licensing information and community task organization.

The labels in ImageNet exclusively describe the objects and concepts in the images [73], thus corresponding to content-based tags. In contrast, Commons categories contain all five types of tags, four of which need to be removed to be compatible with ImageNet labels. While some of these four types of categories are easy to identify, others are more difficult. For subjective categories, we observed that they are rarely used due to the encyclopedic nature of the Commons. Removing organizational categories, can also be straightforward – following previous work [277], we black-list them according to a manually curated list (detailed below). However, removing attribute and context-based categories is less straightforward and requires understanding the semantics of categories. As such, to distinguish attribute and context-based categories from content-based categories, we designed a Content-Based Tags Filter which leverages the YAGO3 semantics provided by YAGO3 Matcher. In the following section, we first describe YAGO3 Matcher and then describe the Content-Based Tags Filter.

4.4.3. YAGO3 Matcher

Our YAGO3 Matcher, matches the category labels to ImageNet labels through the use of YAGO3. YAGO3 is a knowledge base that maps named entities from multilingual

Wikipedia articles (e.g. <Honolulu, HI>) to WordNet classes (e.g. <wordnet_city>). As such, it is well suited for our purpose of matching the Commons' categories, which contain many multilingual named entities, to ImageNet labels, which are WordNet classes. Specifically, YAGO3 Matcher leverages two particular datasets from YAGO3: the yago-Taxonomy table, which describes type-of relationship between WordNet classes (e.g., <wordnet_conference> is a type of <wordnet_event>), and yagoTypes table, which describes the type-of relationship between named entities from Wikipedia and WordNet classes (e.g., <SIGCHI> is a type of <wordnet_conference>).

To match the category labels to entities in YAGO3, our YAGO3 Matcher implemented the following logic (Figure 4.3). In Step 1, the Matcher cleans the labels by removing unwanted labels and preprocessing excessively complex labels. In particular, it removes the organizational tags by comparing them to a manually curated blacklist that includes 60 categories such as “license info required” and “media with locations”. The Matcher additionally skips the tags that are too generic for images, such as “photos”, “pictures”, “images”, etc. For excessively long labels, it splits them using obvious signs that indicate a composite of multiple components such as brackets or slashes. These categories are compared against all YAGO entities to find an exact match and unmatched categories are moved to the next step.

While the previous step matches named entities in foreign languages through multilingual YAGO3, Step 2 matches the generic concepts in other languages using machine translation services. Our Matcher tokenizes each input category and uses the English dictionary to identify foreign language words. These are then translated to English using

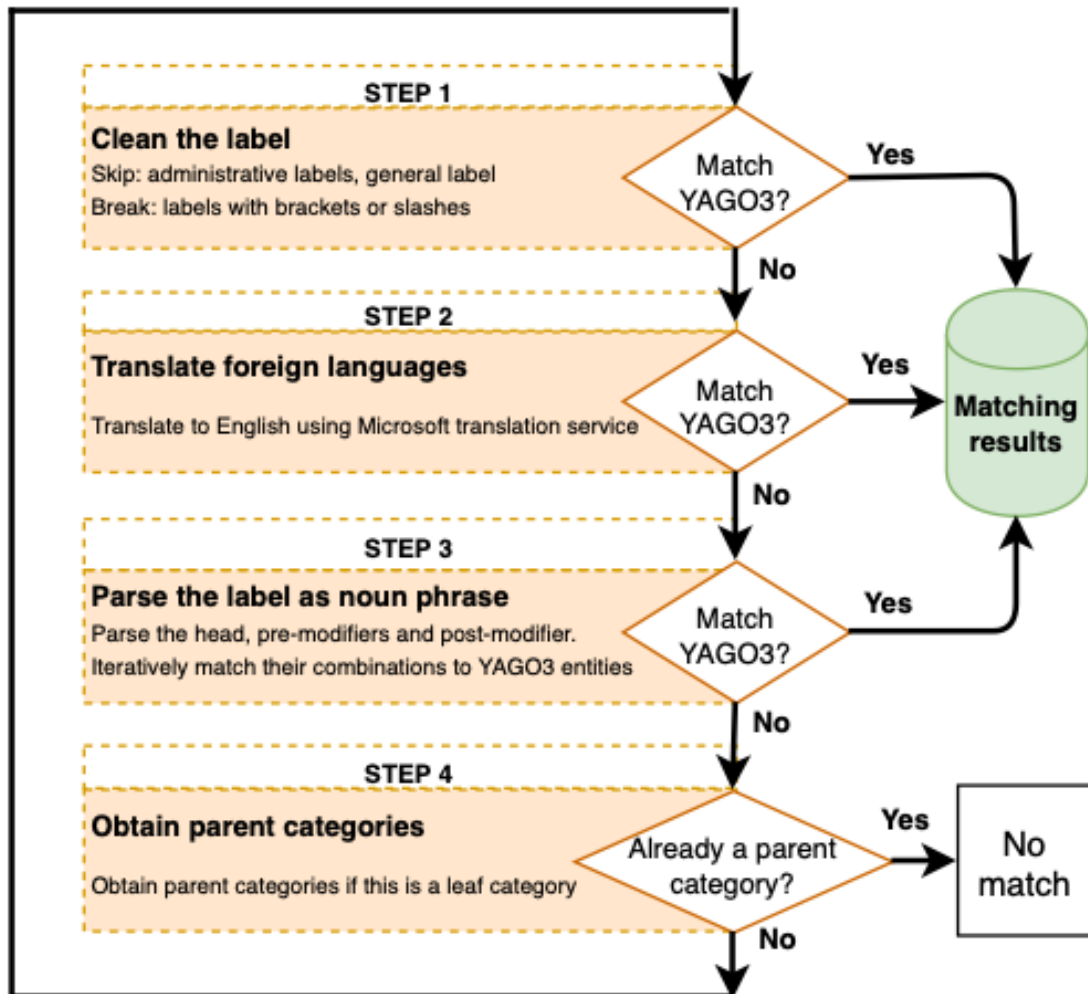


Figure 4.3. Flow chart of the YAGO3 Mapper

Microsoft Translation API [210]. The results are again matched to YAGO3's entities and unmatched labels are moved to Step 3.

If a category fails to be matched to YAGO3 in Step 1 and Step 2, the category is highly likely to be a noun phrase which needs to be further processed. As such, Step 3 parses and matches the core noun phrase in the categories using a system that is developed to match Wikipedia's category labels [277]. While details can be found in the original paper, at a

high-level, the parser divides the category into different grammatical components including premodifier, head and post-modifier and incrementally matches these components to entities in YAGO3. Similar to the previous step, unmatched categories are moved to Step 4.

Step 4 handles another unique problem for the Commons – handling named entities that are too specific. For labels that cannot match YAGO entities, Step 4 looks for their parent categories and matches them using the same procedure starting from Step 1. Following previous work [277], this process is not recursive – the Matcher only queries the direct parents to avoid loops in the category hierarchy.

4.4.3.1. Evaluating YAGO3 Matcher. We conducted a lightweight evaluation to assess the performance of the YAGO3 Matcher. We first developed a test set by randomly sampling 1000 images from the Commons. For this test set, our Mapper produced matching results for 2002 categories. Two researchers coded the 2002 categories using a custom user interface. Specifically, they coded the correctness of the matching result by checking if it directly matches the category label when the category label is a named entity, or if it captures the core nouns when the category label is a noun phrase. For samples with disagreement, these two researchers discussed and reached consensus. The full codebook and coding interface are available in the included Github repository.

Our evaluation shows that YAGO3 Matcher achieves high performance on both precision (0.93) and recall (0.91). We note that our accuracy is comparable with previous systems that match categories on Wikipedia to WordNet (e.g., [91]).

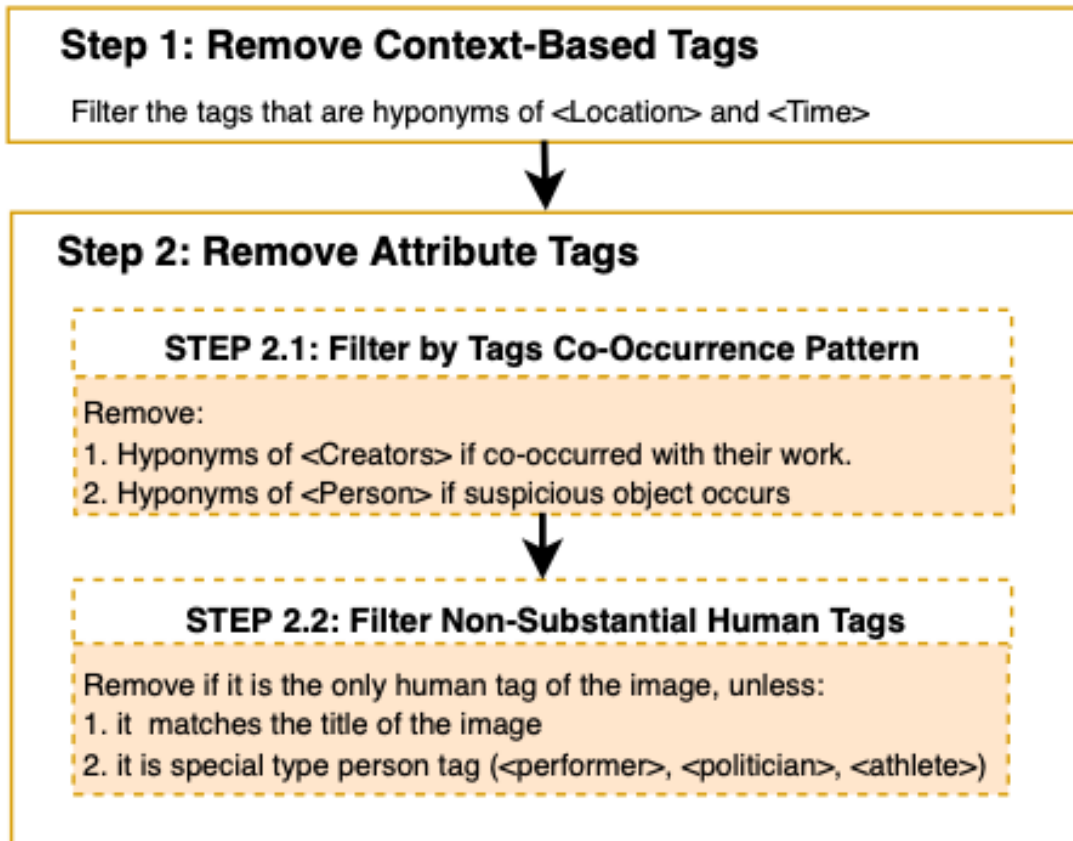


Figure 4.4. Flow chart of the Content-Based Tags Filter

4.4.4. Content-Based Tags Filter

As mentioned previously, only content-based tags are needed for ImageNet. To extract only content-based tags, our Content-Based Tags Filter leverages the YAGO3 semantics provided by the YAGO3 Matcher.

Figure 4.4 shows the flow chart for the logic in the Content-Based Tags Filter. Step 1 removes two important types of context-based tags that are identified by prior literature [36, 301] – time and location tags. For this, we followed a well-established approach from past work [227, 264] which, at its core, used the semantics provided by a knowledge base

to identify the time and location tags. Specifically, we identified time and location tags by examining whether they are in the `<wordnet_location>` or `<wordnet_time_period>` subtree in YAGO3 and, if they are, removed them from the matching results. We note that although these context-based tags cannot serve as labels for the images in ImageNet, the location context-based tags are helpful for assessing the geography the images. We used these tags in our later analysis.

Step 2 involves a more challenging task: removing attribute tags. Prior work identified that, in the context of image labeling, attribute tags mostly represent the authors of the image [301]. As such, Step 2 involves distinguishing between two types of usages of the tags related to a specific person (e.g., `<Charles_Dickens>`): the cases in which they are used for an image of the person (e.g., picture of Charles Dickens) and the cases in which they are used for an image of the person’s work (e.g., a book by Charles Dickens). This task is particularly difficult because, unlike context-based tags, attribute tags cannot be identified solely by their YAGO3 semantics.

Since no solution exists for filtering attribute tags, we implemented our own approach. We first experimented with computer vision-based approaches. We trained an image classifier that identifies the presence of humans in the images and classifies tags of specific person as attribute tags if no human exists in the image. Although we achieved over 90% of precision on this task, adopting this strategy would require downloading over 6 million Wikimedia Commons images that contain tags of specific person, which would abuse Wikimedia API resources. As such, we experimented with a rule-based approach that only leverages the tags semantics and co-occurrence patterns. As we show later, this rule-based approach achieved performance comparable to the computer vision-based

approach. We note that our approach prioritizes precision over recall because we are more interested in retrieving correct content-based tags than every possible content-based tag that exists in the Commons.

Our rule-based approach, illustrated at the bottom of Figure 3, is based on straightforward heuristics. Step 2.1 leverages tag co-occurrence patterns: it removes tags of humans who might be the creators if they co-occur with tags of their work. For example, it is straightforward that `<Charles_Dickens>` is an attribute tag if the image also has `<wordnet_book>` or `<A_Christmas_Carol>`. Specifically, we identified five types of co-occurrence patterns between the hyponyms of the author and the hyponyms of their work, including the co-occurrence between the hyponyms of writers and hyponyms of books, architects and buildings, illustrators and paintings, sculptors and sculpture, and engraver and engravings.

Step 2.1 is effective when these co-occurrences patterns exist. However, this is not always the case – an image with a Charles Dickens book might have `<Charles_Dickens>` but not `<wordnet_book>`. As such, Step 2.2 removes attribute tags based on another heuristic – if an image has multiple tags associated with human, it is more likely that an image actually describes a human instead of the object they created. For example, an image that truly has Charles Dickens as an object likely has multiple tags associated with humans such as `<Charles_Dickens>`, `<wordnet_writer>`. By contrast, an image about a book by Charles Dickens would most likely have no other tag associated with human beside `<Charles_Dickens>`. Using this heuristic, if a tag associated with humans does not co-occur with other tags associated with humans, we remove this tag unless it matches the title of the image, or it is a hyponym of a manually curated white list, which includes

human tags that have substantial usages as content-based tags on the Commons such as <Politician>, <Athlete>, and <Performer>.

4.4.4.1. Evaluating Content-Based Tags Filter. Since Step 1 leverages a well-established approach, our evaluation focused on Step 2 of our Content-Based Tags Filter. We constructed a test set by randomly sampling 200 human tags that are labeled as content-based tags by the Filter. Because evaluating whether a tag associated with humans is content-based or attribute is straightforward for a human labeler, only one researcher coded the test set. Our results show that 92% of content-based human tags that are retrieved are true content-based tags and only 8% are still attribute tags. We do note that the recall of human content-based tags is relatively low. However, this does not undermine the purpose of the Filter, which is to produce high quality content-based tags.

4.5. Validating that Wikimedia Commons Images Can Reduce Geographic Biases

Applying our pipeline on the entire Wikimedia Commons produced 103 million ImageNet-compatible tags for roughly 47 million images. Using these tags, we conducted two analyses to quantitatively demonstrate the benefits of combining Wikimedia Commons and ImageNet for the purpose reducing the geographic biases in computer vision. First, we showed that Wikimedia Commons images can indeed increase the overall representations of developing countries. Second, using a case study, we showed that it is straightforward to using these Wikimedia Commons images to make statistically significant improvements to ImageNet pre-training model’s performance in developing countries.

4.5.1. Analyses 1: Does Wikimedia Commons increase the geodiversity of ImageNet?

In this section, we investigate whether the Commons can increase the geodiversity of ImageNet, in particular with respect to the relative and absolute coverage of developing countries.

4.5.1.1. Extracting the Locations of Wikimedia Commons images. While inferring the location of an image remains a challenging vision task [92], the rich peer-produced metadata on the Commons made it possible to extract the locations for a non-trivial portion of images on Wikimedia Commons. We leveraged two types of metadata to extract the location of the Commons' images. First, many images on the Commons are associated with latitude and longitude coordinates, which are either directly imported from a camera's EXIF information or provided through geocoding by Commons editors. Second, as discussed above, our Content-Based Tags Filter also identified large numbers of context tags that describe images' locations.

To obtain EXIF coordinates data, we again leveraged the Wikimedia Commons' download website, which contains the EXIF coordinates for 8.2 million images. For location-based context tags, we note that we only used named-entity level tags (e.g., <Pearl Harbor>) and excluded the concept level location tags (e.g., <wordnet_city>). Some images are associated with multiple location tags, which usually refer to the same location at different scales. For example, an image of the USS Arizona Memorial might have the location tags <Pearl Harbor> as well as <Honolulu>. Since our analysis occurs at the country level, we simply used one location tag as all of these location tags are more local than country level. Since these location tags all correspond to named entities in

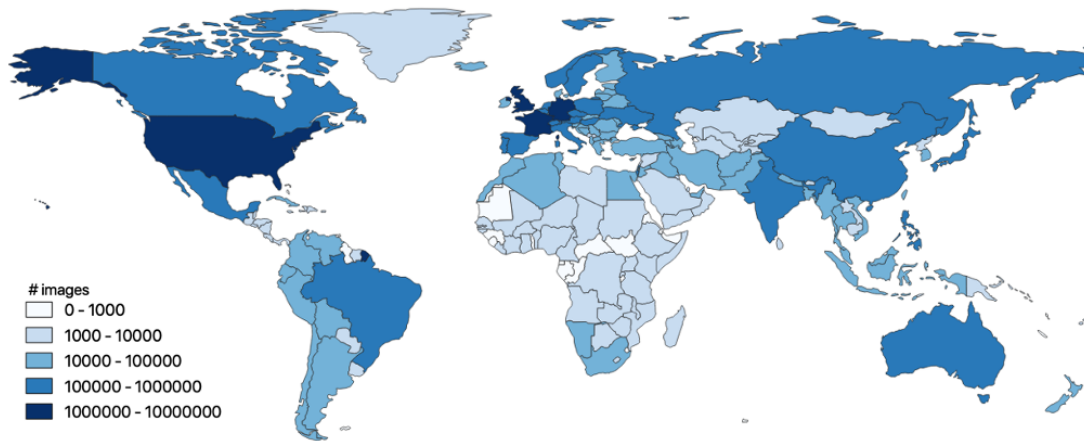


Figure 4.5. Geolocatable images in the Commons. Log-scaled classification, with adjustment for small numbers of images.

YAGO3, we geocoded each tag by querying the English Wikipedia API and requesting their coordinates from their corresponding Wikipedia article information box. In total, we extracted coordinates for 18.5 million images, 44% through EXIF coordinates and 54% through context-based tags.

4.5.1.2. Comparing the Commons and ImageNet on Geodiversity. Figure 4.5 maps the geographic distribution of these 18.5 million images and Table 4.1 lists the 10 most-represented countries. Immediately visible in the figure and the table is that while developed countries like the United States and the United Kingdom are quite prominent in the Commons, developing countries such as the Philippines are also well-represented.

Table 4.2 puts these results in the context of the equivalent numbers for ImageNet as reported by Shankar et al. [39]. The broader geographic distribution of the Commons clearly stands out in Table 2: while the top-five countries in ImageNet comprise 65% of ImageNet, the equivalent number is only 31.4% for our locatable Commons images.

Rank in the Commons	Country	% in the Commons
1	United Kingdom	13.8
2	United States	11.6
3	Germany	10.5
4	France	7.5
5	The Philippines	4.0
6	Czechia	3.3
7	Spain	3.3
8	Netherlands	3.3
9	Italy	3.0
10	Russia	2.8
Sum		63.0

Table 4.1. The ten most-represented countries in the Commons.

Rank in ImageNet	Country	% in ImageNet	% in Commons
1	United States	45.4	11.6
2	United States	7.6	13.8
3	Germany	6.2	3.0
4	France	3.0	1.5
5	The Philippines	2.8	1.5
Sum		65.0	31.4

Table 4.2. Countries that are overrepresented in ImageNet are less dominant in the Commons. ImageNet’s country representation are from Shankar et al. [258].

This includes a dramatic reduction in the relative representation of the United States (ImageNet: 45.4%, the Commons: 11.6%).

Finally, we focus on how our core question - how many images in the Commons are from developing countries? Using the IMF’s definition of developing countries [140], we

found that the Commons can contribute at least 4.1 million new images from developing countries to ImageNet. This is a substantial number: ImageNet currently has 14 million images in total. These additions would include 724k images from the Philippines, 306k from China, 281k from India, 281k from Brazil, and 123k from Mexico. However, we should note that these number are only for Wikimedia Commons images with verifiable locations and there are additional 32 millions images on Wikimedia Commons without location information. If we assume an equal percentage of images from developing countries, that is 4.1 millions images from developing countries among 18.5 millions images with location information (22%), in the rest of 32 millions images without location information, Wikimedia Commons can offer a staggering number of 10.4 millions images from developing countries, almost the same size of ImageNet itself (14.2 millions.)

4.5.2. Analysis 2: Does Wikimedia Commons lead to better ImageNet pre-training models in developing countries? A Case Study on Bridegroom Images

In Analysis 1, we answered a fundamental question - can the geographically diverse Wikimedia Commons be used to address the underrepresentation of developing countries in ImageNet? Our results showed that Wikimedia Commons can indeed provide millions of images from the developing countries, substantially increasing the geodiversity of ImageNet. However, an equally important question remains - will the greater geodiversity in dataset brought by Wikimedia Commons lead to pre-training models' better performance in developing countries?

The ideal approach to exhaustively answer such a question is to test if Wikimedia Commons could lead to better performance in developing countries in *all 1000 classes* in ImageNet pre-training models. However, to get such exhaustive answer faces several challenges. The most important challenge is that we lack the complete knowledge about ImageNet pre-trained models' baseline performance in developing countries for all 1000 classes. While prior research [258, 72] demonstrated the geographic inequalities in the ImageNet dataset, they have only shown that the ImageNet pre-trained model has lower performance in developing countries *on a very limited number classes*. As such, if we conduct experiments on all 1000 classes and do not observe improvement in developing countries on a class for which geographic biases have not been verified, we do not know if the null result indicates that Wikimedia Commons are not helpful or the null result is simply due to the fact that the geographic biases on this class is negligible. An additional challenge is the high-demand of computing resources. Testing on all 1000 classes requires the storage, the pre-processing, and, most importantly, the training of a combination of several millions of images, which require potentially terabytes of storage and multiple hundreds of hours of GPU/CPU time.

Facing the challenges above, we used a case study to get some initial evidence for the Wikimedia Commons' benefits on improving pre-training models in developing countries. Our case study used an ImageNet class for which obvious geographic biases have been verified in [258, 72] - the *bridegroom* class. By using this class, we seek to provide not only initial evidence to support the general utility of Wikimedia Commons' geographically diverse images, but also a working solution for mitigating the geographic biases in a high-profile ImageNet classes that has well-known geographic biases. However, we do note

that our case study is essentially a viability test - if Wikimedia Commons images can improve the accuracy in developing countries, we should be able to observe this effect on the bridegroom class. We lay out future steps necessary for getting an exhaustive answer for all 1000 ImageNet classes in Section 4.8.

To obtain a Wikimedia Commons-enhanced pre-training model for the bridegroom class, we leveraged a transfer learning process called *fine-tuning*, which takes a deep learning model that has been pre-trained on original ImageNet dataset and trains it with new data. Specifically, we took the InceptionV3 model [281] from Tensorflow 2.0 [4] that is pre-trained on original ImageNet task and trained it with additional bridegroom images from Wikimedia Commons. Doing so will, in theory, allow the original pre-training model to learn what "bridegroom" looks like in developing countries through the geographically diverse Commons images.

Our fine-tuning process was carefully designed in order to account for the unique characteristic of our task. While fine-tuning process can be applied to any vision task, our fine-tuning task is relatively unique because recognizing bridegroom images from developing countries needs to focus on the subtle differences in the images (e.g. different type of costumes) compared to the pre-training task of recognizing bridegroom images from developed countries. Because the differences between original and new task are subtle, recent computer vision research on fine-tuning suggests that the fine-tuning process should be designed in a way to prevent pre-training models from learning substantially different features in new data but instead to focus on fine-grained differences [57]. To do so, we adopted a number of best practices for handling such a scenario from [57]. First, we froze most of the layers in the original pre-trained InceptionV3 model and made only

the top two layers changeable during fine-tuning. Second, compared with the original bridegroom dataset in ImageNet which has 1000 images, we created a random sample of only 100 bridegroom images from Wikimedia Commons, a small fine-tuning dataset following the suggestions from the best practices. Finally, we fine-tuned the original pre-trained model with only one epoch, meaning that the model only learned from the fine-tuning dataset once.

To show that this Wikimedia Commons-enhanced pre-training models have higher accuracy on the bridegroom images from developing countries, we compared this model with two baseline models. The first baseline model is an original, unmodified InceptionV3 ImageNet pre-training model, which has been shown to have relatively low performance in the developing countries. However, only using this baseline is insufficient due to a key confounding factor - the additional bridegroom images used by the Commons-enhanced model. Compared with the first baseline model, the enhanced model has been trained on an additional 100 bridegroom images (from Wikimedia Commons). As such, it is possible that the better performance of the Commons-enhanced model in developing countries can be simply attributed to the fact that the model is trained on *more* bridegroom images, but not necessarily due to the *more geographic diversity* introduced by the Wikimedia Commons images. To control for the effect of additional training data, we trained a second baseline model. We again took a InceptionV3 ImageNet pre-training model, but instead of fine-tuning it on 100 Wikimedia Commons images, we fine-tuned it with 100 human-labeled bridegroom images from developed countries. Specifically, we randomly sampled 100 bridegroom images from a third dataset – Open Images dataset, whose images are primarily from developed world [258, 72] and, at the same time, can minimize overfitting

issues when training ImageNet pre-trained models. To control for any additional confounding factors, we leveraged the the same training parameters, including the number of epoch and the number of layers frozen, for fine-tuning the second baseline model.

We evaluated the performance of our Commons-enhanced pre-training model and two baseline models using a test set that includes only bridegroom images from developing countries. To build such a test set, we leveraged bridegroom images from Open Images Extended - Crowdsourced, which are exclusively sourced from developing countries. Even though this dataset has a number of deficiencies (mentioned above), it remains the only available image dataset that focuses on geographic inclusiveness. More importantly, a test set constructed from Open Images Extended - Crowdsourced, which is independent from both ImageNet and Wikimedia Commons, provides the fairest comparison between the original and Commons-enhanced pre-training models. Specifically, we sampled 200 bridegroom images from the Open Images Extended - Crowdsourced dataset. Finally, in order to provide a baseline for the models performance on the original ImageNet dataset which are mainly images from developing countries, we created a second test set by randomly sampling 200 bridegroom images from the original ImageNet dataset.

Following previous work [258], the main performance measurement we used is the mean likelihood of the pre-training model for predicting ground truth bridegroom images as bridegroom images. In other word, this measurement indicates, on average, how "confident" the pre-training model is to correctly predict the bridegroom image in the test set as a bridegroom image. The mean likelihood is preferred over more straightforward measurement such as accuracy because it offers more fine-grained measurement of the

Table 4.3. Mean likelihood of the pre-training models predicting a ground truth bridegroom image as a bridegroom image. The mean likelihood is averaged across all data in the test sets.

Model \ Test Set	test-developed	test-developing
	baseline-original	0.815
baseline-control	0.855	0.273
enhanced-geodiversity	0.866	0.288

errors made by the models ⁵. As such, it helps us not only understand the performance of the pre-training models on the original task (i.e. predicting bridegroom images per se) but also the potential downstream effect to the widely popular pre-training/fine-tuning computer vision paradigm.

Table 4.3 compares the performance of our Commons-enhanced model and two baseline models on the test set. Immediately observed in Table 4.3 is that Wikimedia Commons-enhanced model (*enhanced-geodiversity*) has higher mean likelihood than both baseline models on the test set that include exclusively bridegroom images from developing countries (i.e., results in *test-developing*). Most importantly, compared with the baseline *baseline-control* that controls for the effect of additional training data, the mean likelihood of *enhanced-geodiversity* is improved from 0.273 to 0.288, or a 5% increase. A Wilcoxon signed-rank test (non-parametric test for two paired groups) showed that this increase is significant ($Z=159$, $p=1.5 \times 10^{-33}$) and the effect size as computed by Cohen’s d is 0.742 which is between a “medium” (close to “large”) effect [60].

⁵For most of the ImageNet pre-training models, because the last layer is usually a softmax function, cross entropy loss between the predicted likelihood and the true likelihood is most frequently used. See Appendix for the mathematical relationship between the loss function the the likelihood of each data point.

While the above analyses prove that the performance increase is statistically significant and non-trivial, its practical implications remains slightly intangible. As such, we designed a novel experiment to put the effect size of this performance increase in context. In this experiment, we additionally fine-tuned several variations of *baseline-control* with incrementally more training data sampled from the same data source and compared these variations’ performance with *enhanced-geodiversity*. Our experiments showed that 10% to 20% more training data are needed in order to increase the performance of *baseline-control* to the level of *enhanced-geodiversity*. In other words, on this specific task, the benefits of training on geographically diverse Wikimedia Commons dataset roughly equal to training on 10% to 20% additional non-geographically diverse ground truth images, which has important implications for data collection process as we discuss later. Finally, we highlight that, while we only applied this experiment on the bridegroom class in our case study, it could be extended as a general approach to explain the benefits of geographically diverse training data to models for other classes.

In addition to supporting our main hypothesis that Wikimedia Commons-enhanced models has significant higher performance in developing countries on the bridegroom class, Table 4.3 highlights another important observation - *enhanced-geodiversity* model even outperformed the two baseline models on the test set that includes images primarily from developed countries. Specifically, Wilcoxon signed-rank tests showed that the likelihood of *enhanced-geodiversity* model is significantly higher than that of the *baseline-control* ($Z=162$, $p=1.6 \times 10^{-33}$) with a effect size of Cohen’s $d = 0.856$ which is a “large” effect size [60]. This result seems to suggest that geodiversity in the training samples even

brings benefits to the samples from developed countries where the models have previously learnt well. Future research should further explore this hypothesis.

4.6. Additional Benefits: Increasing the Size and Breadth of ImageNet

Analysis 1 and Analysis 2 confirmed our main hypotheses that Wikimedia Commons images from geographically diverse and these images are helpful at improving pre-training models' performance in previously underperforming countries. During this process, we opportunistically found that the Wikimedia Commons dataset we produced can bring two additional long-anticipating benefits to ImageNet: increasing the size and breadth of ImageNet. In this section, we explain why these two benefits are important and describe how Wikimedia Commons images could enable these benefits.

4.6.1. Increasing the Number of Images

ImageNet has approximately 500-1000 images for each class. While this number is sufficient for learning some classes, it is not enough in other cases. Error analyses of 1000 ImageNet classes show that a subset of classes are more difficult to learn [73, 243], with accuracy as low as only 60% with state-of-the-art deep learning models for these classes. Recent advances in deep learning theory have suggested that more training data in difficult classes could be effective at improving classification accuracies on these classes [279]. As such, we assessed if the Commons could increase the number of images in for these more difficult-to-learn classes.

We summarized the number of images for each ImageNet class in the Commons using a straightforward technique that involves traversing the YAGO3 ontology. Figure 4.6

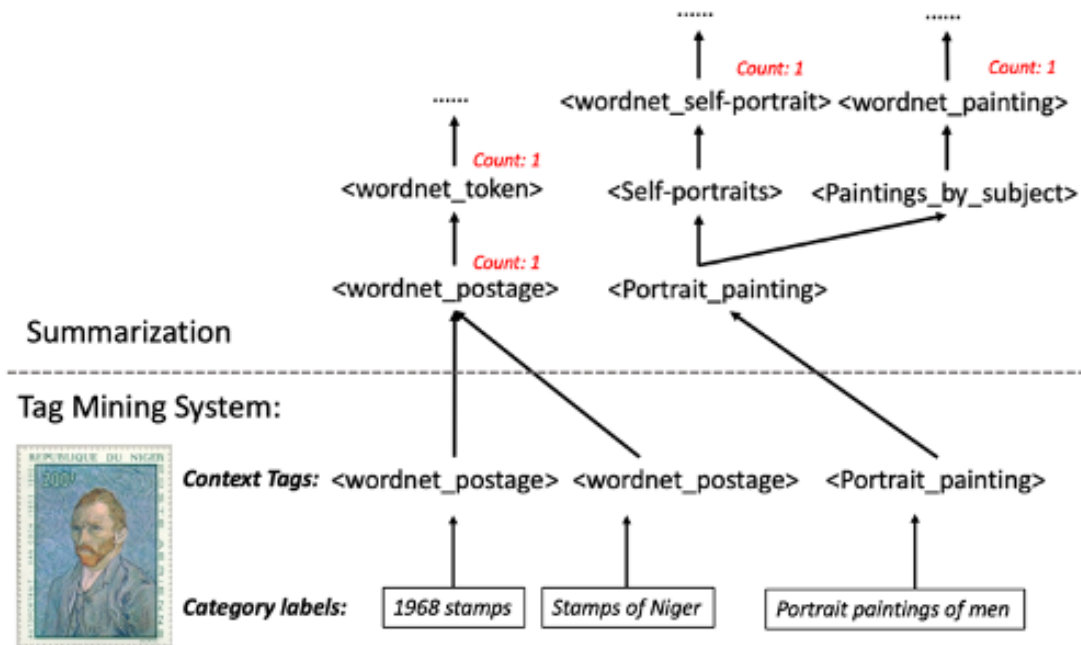


Figure 4.6. The technique for summarizing the number of images

illustrates this technique using an image of a Nigerien stamp which contains a self-portrait of Vincent van Gogh. When summarizing for `<Portrait_painting>`, our technique searches for all ancestors of this class and increments their counts by one. When one tag has been associated with the same image multiple times (e.g., `<wordnet_postage>`), the technique only counts the tag once. As such, the count associated with each WordNet class effectively describes the number of images that are in the subtree of this WordNet class.

4.6.2. Increasing the Data Size for Difficult Classes

Using the summary statistics produced in previous section, we demonstrate that the Commons could be used to increase the data size for two types of difficult classes in ImageNet: those with low classification accuracies and those that are semantically similar.

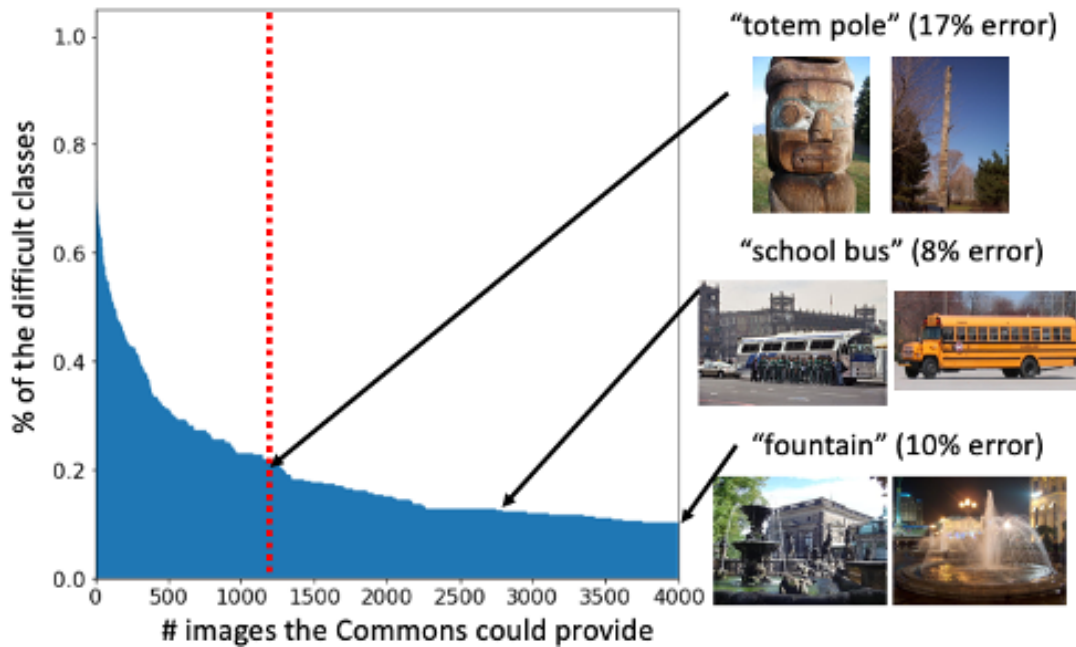


Figure 4.7. The reverse cumulative histogram of the # images the Commons could provide to the difficult classes in ILSVRC. The 0.21 (y-value) at “1200” tick on x-axis means that the Commons could provide 21% of difficult classes at least 1200 images, which double the size of training data for these classes.

We identified classes with low classification accuracies using the per-class accuracies reported by ImageNet Large Scale Vision Recognition Challenge (ILSVRC). ILSVRC contains a 1000-class (1.2 million images) subset of ImageNet and has been widely used for benchmarking object recognition models and producing ImageNet pre-trained features. More specifically, we obtained the per-class performance for the best performing team in the ILSVRC 2015, the latest competition that provides per-class accuracy information. Although average error rate is low (3.5%), the per-class error rates actually vary greatly, ranging from 0% for easy-to-classify classes such as wine bottle, zebra, and barber chair, to 36% for more difficult classes such as guenon, paddle, and hard disk. We focused on the 352 classes that have an greater than average error rate ($>3.5\%$).

Examining the counts for these 352 difficult-to-classify classes, it is clear that the Commons can substantially add to the number of images available for many of these classes. Figure 4.7 shows a reverse cumulative histogram of number of images the Commons can provide for these 352 classes. Note the red dotted line in Figure 4.7 that intersects the x-axis at 1200, which is the average number of images for each class in the ILSVRC dataset. This line intersects with the distribution at 0.21, highlighting that the Commons can provide 21% of the 352 classes with at least 1200 images. In other words, the Commons can roughly double the size of the training data available for at least 21% of difficult-to-classify classes. The median number of images that the Commons can offer to each of these classes is 130, which is 10% of the average number of images per ILSVRC class.

In addition to the classes with low classification accuracies in ILSVRC, Deng et al. [73] identified another type of class in ImageNet that is particularly challenging for computer vision algorithms – semantically similar classes. Analyzing the classification accuracies for the 7404 leaf classes in the ImageNet ontology, they found that computer vision models are easily confused with the classes that are siblings in the WordNet semantic hierarchy (e.g. different types of buildings). The challenge of classifying semantically similar classes has become known as Fine-Grained Visual Recognition [34, 137, 156] and has drawn increasing attention from the research community.

We investigated whether the Commons can provide new images to help tackle the fine-grained classification problem. Following Deng et al. [73], we focused on the leaf node classes under <Building>. Our results show that the Commons can double the number of the images in 50% of all <Building> leaf node classes.

Overall, we observed that the Commons can be used to grow the number of images in difficult-to-classify ImageNet classes, including the classes that have low classification accuracies and the classes that are semantically similar. Given that more training data leads to more accurate deep learning models on a variety of vision tasks [279], our results suggest potential improvements to ImageNet baseline accuracies.

4.6.2.1. Enhancement 3: Expanding ImageNet’s Coverage. While ImageNet aims to cover a majority of WordNet’s over 80,000 classes, it has so far only covered 27.3%. As noted above, improving ImageNet’s coverage could enable new computer vision applications such as fine-grained recognition [25].

We investigated whether the Commons could extend the coverage of ImageNet by providing a sufficient number of images for WordNet classes that are not in ImageNet. To define the sufficient number of images, we refer to the average number of images in ImageNet which is 650 (this is lower than the average number in the ILSVRC dataset). We focus our analysis on the WordNet leaf-node classes; for the non-leaf-node classes, one can aggregate the images from the child leaf-node classes.

Figure 4.8 shows the reverse cumulative histogram of the number of images the Commons can provide to the previously uncovered WordNet leaf-node classes. The Commons can provide 1591 of all previous uncovered classes with at least 650 images. We note that some of these new classes are semantically related – including different types of mineral and artwork (examples in Figure 4.8) – making them well-suited to new fine-grained classification tasks. We also note that at the tail of the histogram (which is not shown), some new leaf-node classes such as <wordnet_portrayal> and <wordnet_carnival> have a truly enormous number of images in the Commons (225k and 80k respectively). This indicates

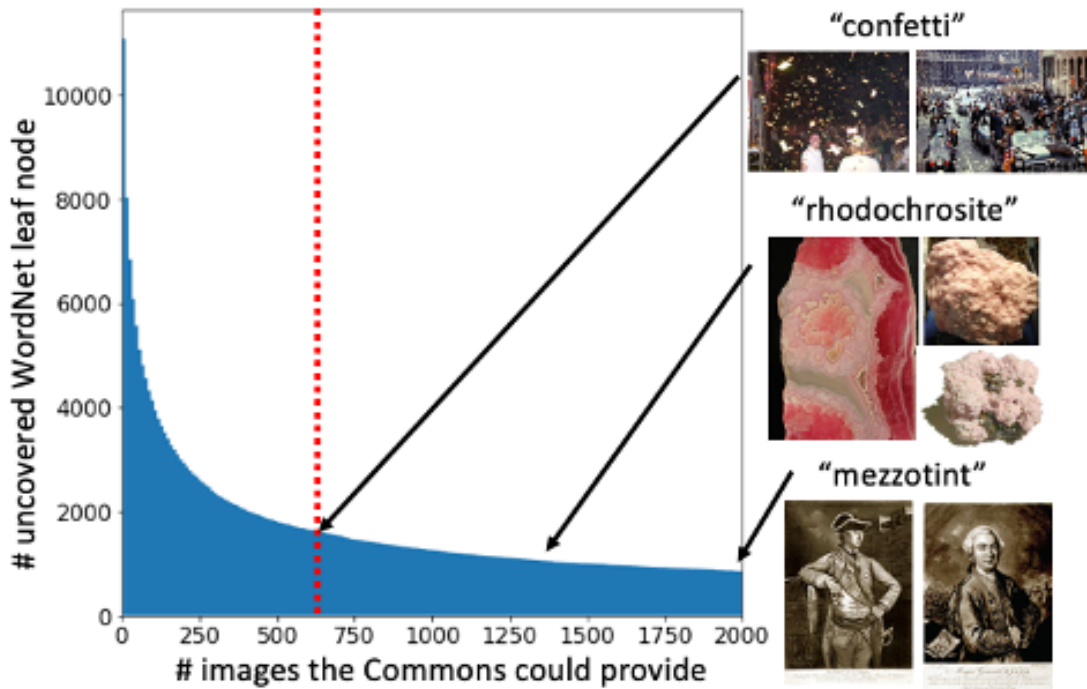


Figure 4.8. The reverse cumulative histogram of the # images the Commons could provide to the previously-uncovered ImageNet classes. The red dotted line is the average number of images per previously-uncovered class (650 images).

that sub-WordNet-class-level fine-grained classification opportunities might exist for these classes by leveraging our YAGO3 entity tags.

4.7. Discussion

Our results offer initial evidence that the peer-produced Wikimedia Commons has tremendous potential to enhance the geodiversity of ImageNet, a critical gold standard dataset that directly or indirectly affects thousands of computer vision models. Besides this main contribution, our findings have additional implications for research that enhances other large gold standard datasets and studies around Wikimedia Commons. Below, we unpack these implications in more detail.

4.7.1. Reducing Training Data Size with Geographically Diverse Data

Our experiment results demonstrate that on the task of recognizing bridegroom images, the benefits of training on geographically diverse training data equals to training on 10% -20% more non-geographically diverse training data. An alternative interpretation of this result is that, if machine learning models are trained with geographically diverse training data, it is likely that these models can achieve the same accuracy with fewer (non-geographically diverse) training data. Given cost of obtaining human-labeled, high-quality training data as well as training, purposefully designing the data collection to prioritize geographic diversity could potentially reduce the amount of data collection, saving money of the dataset developer as well as helping protecting the privacy of contributors. In addition, smaller training data also indicates lower financial and environmental cost of model training process, which has been the subject of ethical concerns of developing large deep learning models [172, 126, 11].

4.7.2. Using Peer Production to Enhance Large Gold Standard Datasets

While market-based crowdsourcing such as Amazon Mechanical Turk is commonly used for collecting data for large gold standard datasets, we showed that a different crowdsourcing mechanism, peer production, can be leveraged as a low-cost approach to enhance these datasets. Successful peer production datasets are often large and knowledge-driven [30, 31], making them a natural candidate for gold standard datasets.

While peer production data are still unable to equally represent all regions of the world [122], we show that peer production has greater geodiversity than some market-based crowdsourcing approaches like the one used in ImageNet. Given that geodiversity has

been a recurring issue for large gold standard datasets (e.g., [72, 258]), our results suggest a promising approach to improve geographic representation in gold standard datasets, if there exists a parallel peer-production dataset. Peer production is far from a panacea when it comes to geodiversity, but it appears to be better than market-based crowdsourcing on the geodiversity front.

It is important to reiterate a number of considerations when integrating peer-produced data into gold standard datasets. First, research has identified a number of societal biases with peer-production (aside from geodiversity), including gender bias [104, 236] and self-focus bias [121]. As such, integrating peer-production data into gold standard datasets should be carefully conducted in light of these issues. Second, the quality of peer-production is known to be better for certain type of projects than for others. For example, Benkler et al. [31] argued that peer-production is most effective for building knowledge-based project such as Wikipedia-like projects, but less effective for creative works. In this regard, the quality of certain types of peer-produced data might be insufficient. Finally, not all peer-production projects have license as friendly as Wikimedia Commons' or the community might be opposed to the idea of using their data as gold standard datasets for machine learning. As such, licensing and community relationship are some of the other concerns that researchers need to consider if using peer-produced data.

4.7.3. Untapped Research Potential of Wikimedia Commons

In addition to facilitating the three ImageNet enhancements we describe above, our work highlights that the Commons might offer many other research opportunities for those in HCI and computer vision alike. For computer vision research, we discovered that

Commons contains more than 10 million context-based and attribute tags that remain largely untapped in the vision community. These non-content-based tags could enable classifications such as styles of buildings, authorship/genre of artwork, and styles of cloth – tasks that demand models to pick up extremely high-level and subtle visual features.

For HCI researchers, Wikimedia Commons has exciting research potential. For instance, comparing Wikimedia Commons to Wikipedia will bring invaluable knowledge about the similarities and differences on the topic coverage between visual and textual peer-produced knowledge. The tags generated and the pipeline developed by our work could help enable this research.

4.8. Future Work and Limitations

The work in this chapter is only the initial step towards the ultimate goal of fully addressing the geographic inequalities in the computer vision-based location-aware AI technologies. To provide readers with the full picture, we use this section to outline all necessary steps for achieving this ultimate goal and to situate the work in this chapter.

Referring back to the propagation of geographic inequalities illustrated in Figure 4.1, the work in this chapter focuses on mitigating the geographic inequalities in the large-scale dataset (i.e. ImageNet) as well as providing initial evidence for mitigating the geographic inequalities in the generic AI models (i.e. ImageNet pre-trained CNN.) Future work should provide more comprehensive evidence for the reduced geographic inequalities in the generic AI models by replicating our machine learning experiments on more ImageNet classes. In addition, while we strongly believe that the reduced geographic

inequalities will follow the same propagation process to finally benefit computer vision-based location-aware technologies, future work should still seek experimental evidence and, more importantly, quantify the effect size.

While our results demonstrate the huge potential of leveraging the Commons to address the geographic inequalities in computer vision-based location-aware technologies, it is important to point out a few limitations of our work. One important limitation is related to the accuracy and biases of the peer-produced category labels on the Commons. An assumption of our work is that these peer-produced category labels are of high quality. While the wide popularity of the Wikimedia Commons (the second most popular projects among all Wikimedia Foundations projects) is generally an strong indication of its high quality [295], no research has directly measured the accuracy of the category labels or the gender and racial biases of the content on the Commons.

A second, more minor limitation is that precision for our YAGO Matcher (93%) is slightly lower than ImageNet’s human labeling (99.7%). However, we note that the tags we produced is still far more accurate than many other large image datasets (e.g., 20% for Open Images [279]).

4.9. Conclusion

To address the geographic biases that exist in the computer vision pre-training models, we explored the possibility of leveraging Wikimedia Commons, a large peer-produced encyclopedic dataset, to substantially increase the representation of developing countries in large image datasets such as ImageNet. We showed that the Commons can 1) provide

a substantial number images from the developing world, and 2) has the potentials to lead to better pre-training models for developing countries.

CHAPTER 5

Addressing the Geographic Biases in Location-Aware Technologies based on Commonsense Knowledge Base

Note: the work in this chapter was originally published in the Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017) [189]. Necessary modifications have been made to situate the original content in the context of this thesis.

5.1. Introduction

In this chapter, we focus on mitigating the geographic inequalities in another type of AI-based location-aware technologies - knowledge base location-aware technologies. As the bottom level of Figure 5.1 shows, an example of knowledge base location-aware technology might be a geographic question & answering (Q&A) chat bot that answers geographic questions such as “What is the population of the United States.” As we discussed previously, one widely used approach to develop AI-based location-aware technologies on the top of a generic AI model. In the case of geographic Q&A chat bot, this generic AI model might be a generic knowledge base technologies, such as Google’s Knowledge Graph [267] (see the middle level of Figure 5.1), which are a type of technology that stores the facts and relationships about concepts in human knowledge. Finally, the generic AI model was based on a large-scale labeled training data - in our example, this is Wikipedia and its

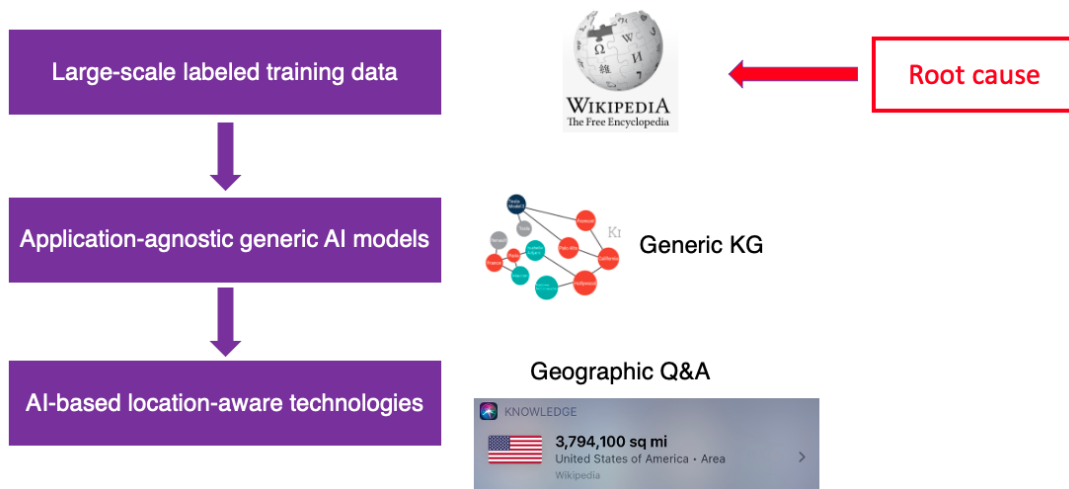


Figure 5.1. The propagation of geographic inequalities from Wikipedia to the knowledge base location-aware technologies

associated data, which have been the foundation of numerous successful knowledge bases (e.g., [277, 130, 16].)

As Figure 5.1 illustrates, similar to how geographic inequalities propagate to computer vision-based location-aware technologies, if geographic inequalities exist in the first level - Wikipedia data, it is likely that these geographic inequalities will propagate through the diagram and finally enter the knowledge base location-aware technologies.

However, we wanted to emphasize that the geographic inequalities in Wikipedia data that this chapter focuses on are different from the geographic inequalities in the ImageNet. The geographic inequalities in ImageNet and other large image datasets are inequalities of data quantity - some geographies are represented with more samples than other geographies. While the inequalities of data quantity do exist in Wikipedia data, they are unfortunately hard to be mitigated with technical solutions - since Wikipedia per se is already arguably the most geographically equitable encyclopedia [106], it is almost impossible to

find another data source that could make up for the knowledge underrepresentations of certain geographies in Wikipedia.

Instead, this chapter focuses on mitigating a different type of geographic inequalities - the inequalities that some geographic concepts have more complete Wikipedia data than other geographic concepts.

To understand this type of geographic inequalities, it is necessary to understand a special type of Wikipedia articles called “sub-article”. The Wikipedia community strongly encourages editors to divide lengthy articles into multiple articles out of a desire to facilitate readability, maximize ease of editing, sustain contributions, and make Wikipedia suitable for diverse technological contexts (e.g. slow connections) [66]. According to these guidelines, the original article (parent article) should contain a summary of the concept, and each split-off article (sub-article) should have a full treatment of an important subtopic. As such, while article-as-concept assumption is valid for geographic concepts with one corresponding Wikipedia articles, it is problematic for geographic concepts with more Wikipedia content which have multiple corresponding Wikipedia articles (i.e. one parent article and several sub-articles). For example, Lauderdale in Minnesota, the first U.S. city where the author lived, has only one Wikipedia article “Lauderdale, Minnesota”. By contrast, in addition to the parent article “Portland, Oregon”, Wikipedia contains many other sub-articles about Portland, including “History of Portland, Oregon”, “Downtown Portland”, “Tourism in Portland, Oregon” (among over a dozen other sub-articles).

While sub-articles bring convenience for Wikipedia readers, they lead to data completeness issue for Wikipedia-based knowledge base. Wikipedia-based knowledge base

provides geographic knowledge under, what we called, the “article-as-concept” assumption. That is, Wikipedia-based knowledge base assumes there is one-to-one mapping between geographic concepts and the Wikipedia articles. Under the “article-as-concept” assumption, it assumes that the entirety of the English description of a geographic concept should be in the article with the same title of this geographic concept, *and that article alone*. As such, when a Wikipedia-based knowledge base requests information about, for example, Portland, Oregon, the knowledge base will only retrieve incomplete data about this concept, including only content from “Portland, Oregon” yet missing out all content in the sub-articles associated with “Portland, Oregon”.

More importantly, this data completeness issue is unequal among geographic concepts. Depending on the amount of their Wikipedia content, geographic concepts have varying number of sub-articles. For example, while a significant geographic concept, such as United States, might have more than 50 sub-articles, a less significant geographic concept, such as Jieyang which is my father’s hometown, have zero sub-articles. As such, for geographic concepts with no sub-articles, Wikipedia-based knowledge base will provide *complete* Wikipedia information. However, for geographic concepts that have multiple sub-articles, Wikipedia-based knowledge base will only provide *partial* information, retrieving only the content in the parent articles but missing out important details in the sub-articles.

In this chapter, we proposed one approach to address the unequal data completeness issue among geographic concepts on Wikipedia. That is, we developed a generalizable solution to match parent articles of geographic concepts with all their corresponding sub-articles so that Wikipedia-based knowledge base systems can easily retrieve complete data

for all geographic concepts. To achieve this goal, this chapter introduces the sub-article matching problem. The sub-article matching problem describes the following task: for a potential parent article p in a given language edition, accurately identify all corresponding sub-articles p_s in the same language edition. For instance, solving the sub-article problem involves connecting the parent article “Portland, Oregon” with its “History of Portland, Oregon” and “Portland Fire & Rescue” sub-articles (and others), and repeating for all articles in multilingual Wikipedia. As we will show, this will mean determining whether millions of potential sub-articles are indeed sub-articles of corresponding parent articles.

In addition to defining and motivating the sub-article matching problem, this chapter presents the first system that addresses the problem. We collected sub-article ground truth corpora consisting of pairs of parent articles and candidate sub-articles ($\langle p, p_{cs} \rangle$ pairs) in three languages (English, Spanish, and Chinese). We then used that data to train a model that can achieve 84% classification accuracy on average, outperforming baseline approaches by 17%. Further, we show that our model works best on the articles that attract the most reader interest: it outperforms baseline accuracy by 50% on a dataset consisting only of high-interest articles.

The model’s performance is achieved through the use of heterogeneous features ranging from Wikipedia editing practices to the output of semantic relatedness algorithms. Moreover, we were careful to avoid language-specific features in our model, meaning that the model should work equally well in most (if not all) major language editions.

This chapter also situates the challenges to the article-as-concept assumption in a broader theoretical framework. Specifically, we discuss how this issue can be interpreted as a human-machine variation of author-audience mismatch [134], which was originally

conceptualized for human authors and human audiences. We argue that the author-audience mismatch framework – adapted to machine audiences and human authors – is useful for understanding a growing number of problems associated with intelligent technologies’ use of semi-structured peer-produced datasets like Wikipedia.

In the interest of furthering progress on the sub-article matching problem, we have made our full sub-article ground truth dataset publicly available and we have released our entire sub-article model as an extension to the WikiBrain Wikipedia software library [253]. This will make our model immediately usable by Wikipedia researchers and practitioners. Additionally, including our model in WikiBrain will also allow researchers to make straightforward comparisons with our model’s performance and, hopefully, improve upon it.

Finally, we note that, although the examples we used above are exclusively geographic concepts, the article-as-concept assumption in Wikipedia-based knowledge base also affects non-geographic concepts. To make a broader contribution, while the remainder of this chapter continues using geographic concepts as illustrations, our solution is designed in a way that can tackle the sub-article matching problems for both geographic and non-geographic concepts.

Below, we first address related work. We then discuss our efforts to build reliable ground truth datasets for the sub-article matching problem. Third, we address how we constructed our classification models and interpret model performance. Finally, we highlight the theoretical implications of our work associated with author-audience mismatch.

5.2. Background

In this section, we expand on how knowledge base uses Wikipedia and how the sub-article matching problem will be problematic for these scenarios.

The article-as-concept assumption is problematic for projects that to integrate Wikipedia into structured knowledge base. Besides well-known projects such as Google’s Knowledge Graph [267], YAGO [277, 130], and DBPedia [16], one recent high-profile project is Wikimedia Foundation’s very own effort - Wikidata [288]. Broadly speaking, Wikidata consists of items that correspond to Wikipedia articles connected via semantically-labeled properties. Wikidata has become a critical resource for many intelligent technologies (e.g., [82, 217, 299]), which potentially makes the article-as-concept assumption more problematic.

Wikidata’s reliance on the article-as-concept assumption dates back to its launch in 2012, when it used Wikipedia articles to seed its corpus of concepts. As a result of this approach, sub-articles are considered to be separate concepts relative to their parent articles. In other words, in Wikidata, “History of Portland” (English) and “Portland” (English) are treated as describing entirely different concepts, as is the case for hundreds of thousands of other parent article/sub-article relationships.

The problems associated with the article-as-concept assumption in Wikidata are quite apparent in the first large technology to use Wikidata information: Wikipedia itself. For instance, Wikidata is now the backend for the “Other languages” links in the Wikipedia sidebar. These links refer readers to other articles about the same concept in different languages. Because of the article-as-concept assumption in Wikidata, a reader of the

“History of Portland, Oregon” (English) article (or a study using the Wikidata’s connections between language editions) will not be exposed to the large “History” section in the German article on “Portland (Oregon)” that happens to not be split into a separate article. The same problems are occurring with other Wikidata integrations in Wikipedia. For instance, the Wikipedia templates that now draw information from Wikidata will struggle to look up information about a single concept that is split across multiple Wikidata items.

One potential application of our work is leveraging the open-source software package we have released to power a Wikidata editing bot to help propagate “facet of” properties to more Wikidata items. In addition, our software package can also be used to help Wikidata-based applications dynamically integrate information from parent articles and sub-articles.

5.3. Datasets Development

5.3.1. Wikipedia Corpus

We downloaded XML Wikipedia data dumps in January 2015 and processed these dumps using WikiBrain [253], which is a Java software framework that provides access to a range of Wikipedia datasets and Wikipedia-based algorithms (including the semantic relatedness algorithms we use below). We focused on three language editions: Chinese, English and Spanish. These were selected because they (1) are widely spoken and (2) span the East/West spectrum, which has proven to be an important consideration across a large body of HCI literature (e.g., [23, 122, 131, 203]). We additionally processed 22 other language editions as we required additional language editions to operationalize language-neutral features (see below). Overall, our Wikipedia datasets contains the 25 language

editions with the most articles (as of January, 2015), excluding largely bot-generated editions like Cebuano and Waray-Waray. It is important to note that, in addition to utilizing WikiBrain, we also contributed back to WikiBrain by developing a sub-article matching problem extension.

5.3.2. Sub-article Candidates

An essential dataset in our study is our large sub-article candidate dataset. In this subsection, we first define sub-article candidates and then describe how we mine them from the Wikipedia corpora described above.

5.3.2.1. Defining Sub-article Candidates. Without additional information, any Wikipedia article could be a sub-article of any other Wikipedia article in the same language edition. As such, a potentially intractable computational problem emerges. For instance, with its over 5,210,165 articles, the English Wikipedia alone would require examining more than 25 trillion potential sub-article relationships ($5210165 * 5210165 - 5210165$).

Fortunately, Wikipedia editors use a number of indicators that allow us to prune away a huge portion of these potential sub-article relationships a priori, resulting in a much smaller (but still large) group of sub-article candidates. Editors employ these indicators – which vary from language edition to language edition – to highlight for readers when content from a parent article has been split off into a sub-article.

To identify indicators for sub-article candidates, a single investigator fluent in English and Spanish accessed thousands of pages in all 25 language editions in our corpus, focusing on concepts that typically had sub-articles in many language editions (e.g. countries, major historical events, large cities). Although context is usually sufficient to identify a

compact the [urban growth boundaries](#) of the city.^[178]

Sports [edit]

`{{Main article|Sports in Portland, Oregon}}`

Main article: [Sports in Portland, Oregon](#) Wiki markup

Portland is home to two major league sports franchises: the [Portland Trail Blazers](#) of the [NBA](#) and the [Portland Timbers](#) of [Major League Soccer](#). The [Portland Thorns](#) of the

Snippet from “Portland, Oregon” (English)

- [Ya ba](#) contains a combination of [methamphetamine](#) and [caffeine](#).

History

`{{main article|History of chocolate|History of coffee|History of tea|History of yerba mate}}`

Discovery and spread of use Wiki markup

Main articles: [History of chocolate](#), [History of coffee](#), [History of tea](#) and [History of yerba mate](#)

According to Chinese legend, the [Chinese emperor Shennong](#), reputed to have reigned

Snippet from “Caffeine” (English)

Figure 5.2. An example of the ambiguity inherent in sub-article indicators, in this case the “`{{main article}}`” template in the English Wikipedia. Links in red indicate potential sub-article relationships

sub-article relationship, the investigator used Google Translate as an aid when necessary. Whenever the investigator encountered a potential sub-article relationship, he recorded the parent article (e.g. “Portland, Oregon”), the potential sub-article (e.g. “History of Portland, Oregon”), and, most importantly, the Wiki markup that was used to encode the relationship (Wiki markup is the markup language used by Wikipedia editors to write articles). The final dataset consists of 3,083 such records, and is included in the release of our ground truth dataset.

See also [edit]

- [Chemetco](#), U.S. company that produced air-borne dioxin inferred to be the source of contamination in Nunavut
- [Archaeology in Nunavut](#)
- [Scouting and Guiding in Nunavut](#)
- [Symbols of Nunavut](#)
- [Arctic policy of Canada](#)

Figure 5.3. The “See also” section of the English article about the Canadian territory of Nunavut. Some links here are clearly sub-articles, while others are more distantly related to the concept of Nunavut (e.g. “Arctic policy of Canada”).

Using the above process, we identified two general types of sub-article indicators. The first type is the template-based indicator that resembles the appearance of Figure 5.2, although the specific markup and prompt varies within and across languages (e.g. the “`{{main article}}`” template in Spanish is “`{{AP}}`” for artículo principal, and similar English templates include `{{see also}}` and `{{further}}`). Templates in Wikipedia are a type of wiki markup that editors can use to generate complex HTML just by entering a few parameters, which is illustrated in Figure 5.2,.

The second type of indicator is significantly different. Potential sub-article relationships encoded through this type of indicator are listed at the bottom of articles under a header titled “See also” or its equivalent in other languages. Interestingly, using see also-based indicators for sub-articles is explicitly contrary to community guidelines in some language editions (e.g. the English Wikipedia’s Manual of Style, which state that “See

also” sections should be reserved for links to peripherally-related content). However, our candidate dataset reveals that editors frequently violate these guidelines (e.g. Figure 5.3).

More generally, while both template-based and see-also based indicators are often used to indicate sub-article relationships, they are also used for a variety of other purposes, causing large numbers of false positives to emerge. Figure 5.2 illustrates this phenomenon with an example highlighting one of the more common sub-article indicators in the English Wikipedia: the main article template-based indicator. The top of Figure 5.2 shows a situation in which this indicator is used to demarcate a true sub-article relationship (“Portland, Oregon” and “Sports in Portland, Oregon”), but the bottom shows a situation in which this is clearly not the case (“Caffeine” and History of chocolate”, “History of coffee”, “History of tea”, and “History of yerba mate”).

To summarize, although sub-article indicators like “main article” are ambiguous, mining them from the article text of each language edition is an essential pre-processing step. This is because (1) they can be considered to be a broad superset of all sub-article relationships and (2) they prevent us from having to compare all articles in every language edition in a pairwise fashion (a potentially intractable brute force approach). Below, we describe the procedures we use to execute this multi-faceted mining process.

5.3.2.2. Mining Sub-article Candidates. After developing our dataset of sub-article indicators, we used these indicators to write a script that parsed out all sub-article candidates across all 25 languages. In most cases, this script utilizes straightforward regular expressions, although other cases were more complicated. Our script is included in our open-source WikiBrain sub-article software library.

	English Wiki	Chinese Wiki	Spanish Wiki
% of articles (with templates + see also section)	20.5%	11.6%	25.2%
% of articles (with tem- plates)	4.9%	2.3%	7.1%
% of pageviews (with tem- plates)	24.7%	11.9%	25.6%

Table 5.1. The percent of articles and page views associated with potential sub-articles

A quick examination of this dataset was our first indication that separating the signal (true sub-articles) from the noise (false sub-articles) will be difficult, even among the much narrower class of sub-article candidates. We originally expected that many sub-articles would follow the pattern “something of parent article” such as “Geography of the United States” (sub-article of “United States”), and the equivalent in each of the 25 languages we considered (e.g. “Geografía de Estados Unidos” in Spanish). However, it became clear in this preliminary dataset that a significant portion of sub-articles violate this pattern. For instance, this preliminary dataset contains potential sub-article relationships between parent articles p and candidate sub-articles p_{cs} such as $p =$ “List of Chinese Inventions” and $p_{cs} =$ “Four Great Inventions”, $p =$ “United States” and $p_{cs} =$ “American Literature” and $p =$ “The Silmarillion” and $p_{cs} =$ “Valaquenta” (all from the English Wikipedia.)

Overall, we found sub-article candidates for a substantial proportion of Wikipedia articles. For instance, over a quarter of articles in English and Spanish Wikipedia contained sub-article candidates. More details about these percentages, including the share of template-based and see also-based indicators, is available in Table 5.1.

5.3.3. Ground Truth Datasets

As is typical in many types of machine learning, our sub-article models require extensive ground truth data for both training and testing. Because no prior work has defined, let alone attempted to solve, the sub-article matching problem, it was necessary to both generate our own ground truth datasets as well as to define their character and structure. By developing these datasets and making them publicly available, we also hope to make it easier for other researchers to build on our results.

In this sub-section, we first define the high-level structure of our ground truth datasets, focusing on how they were sampled from the larger set of overall sub-article candidates. We then describe how we labeled each potential sub-article relationship in these datasets. This is an important process that, as is often the case when developing the first ground truth dataset for a new problem, led to a formalization of the definition of sub-articles.

5.3.3.1. Sampling and Structure: Selecting Sub-article Candidates. We generated three ground truth datasets, each using a different strategy to sample from the sub-article candidate population. All three datasets consist of a series of $\langle p, p_{cs} \rangle$ pairs (i.e. \langle parent article, potential sub-article \rangle pairs). For each dataset, we randomly selected 200 pairs from English, and 100 each from Spanish and Chinese. Each of the three sampling strategies was designed to maximize ecological validity for a different class of sub-article matching use cases. These strategies are described in more detail below:

- *High-Interest*: This dataset consists of $\langle p, p_{cs} \rangle$ pairs whose parent articles are sampled from the 1,000 most-viewed articles for a given language edition. We gathered page view data from the Wikimedia’s Page View API¹ and aggregated it

¹https://wikimedia.org/api/rest_v1/

from August 2015 to February 2016. This is a user-centered dataset that is driven by the actual demand for Wikipedia content. Performance on this dataset will be a good proxy for the model’s performance on most real-life systems, especially those that are directly user-facing (e.g. Omnipedia [23], Manypedia [204], etc.). This means that *High-Interest* is likely the most important of the three datasets.

- *Random*: The parent articles in this dataset were sampled randomly from all articles in each language edition. Corresponding sub-article candidates were then randomly selected from the set of candidates available for each parent article (potential parent articles without sub-articles were ignored). Given the long-tail quality distribution of Wikipedia articles [2], this dataset includes large numbers of short, low-quality articles. It also contains many articles that are of very limited reader interest (relatively speaking). This dataset will give a lower-level understanding of performance across entire language editions.
- *Ad-Hoc*: This dataset was generated by sampling from the 3,083 $\langle p, p_{cs} \rangle$ pairs generated during the indicator identification process, focusing only on candidates from English, Spanish, and Chinese. This was the first dataset we generated, and it served as a successful feasibility test. It also provides a useful triangulation of our models’ performance relative to the other two ground truth datasets.

5.3.3.2. Labeling: Evaluating Sub-article Candidates. The Wikipedia community has no precise formal definition for what separates a “true” sub-article from a “false” one. This is because the sub-article construct was invented for human readers, and human readers do not need such binary distinctions: if a human is interested in a link, s/he clicks

on it, regardless of whether the link is a sub-article or not. (We return to this issue when covering audience-author mismatch in the Discussion section.)

Wikipedia-based studies and systems, on the other hand, must make this binary distinction, and must do so frequently and explicitly. As such, a major challenge becomes finding a way to codify the definition of a sub-article relationship in our ground truth datasets in a fashion that Wikipedia-based systems and studies can understand, while at the same time respecting the variability of sub-article relationships encoded by human Wikipedia editors.

To address this challenge, we adopted the approach of past Wikipedia research that has also sought to codify fluid Wikipedia constructs (e.g., [23]). Specifically, we utilized a two-part method that allows the researcher or system designer to decide how broadly or narrowly they want to define the sub-article construct, according to the needs of their application or study. The first step in this process involved coding potential sub-article relationships on an ordinal spectrum in recognition of the non-binary nature of the sub-article relationships. We then proposed several reasonable thresholds and examined our results using each of these thresholds. This multi-faceted approach allows us to understand the performance of our models at each breadth level and allows the users of our system to flexibly choose from broader and stricter definitions of sub-articles.

With regard to the ordinal coding stage, for each language of each dataset (English, Spanish, Chinese), we recruited two coders who were fluent in the corresponding language. Coders were asked to assign each $\langle p, p_{cs} \rangle$ a code along an ordinal scale from 0 (definitely not a sub-article) to 3 (definitely a sub-article), with each code being defined as follows²

²Our full coding instructions and codebook are included in Appendix.

- **3**: The only reason the sub-article candidate exists is to split the corresponding parent article into more manageable subtopics. The potential sub-article really does not deserve its own page, and the corresponding parent article is the best place to put the sub-article’s content.
- **2**: Same as above, but the topic of the sub-article candidate is significant enough to warrant its own page.
- **1**: The sub-article candidate contains information that would be useful to have in the parent article, but also contains its own, unrelated (non-overlapping) content.
- **0**: The sub-article candidate is about a topic that is trivially related to the parent article or has a large amount of non-overlapping content.

The inter-rater reliability on our datasets as computed by Weighted Cohen’s Kappa [59] ranged from 0.56 to 0.78, which is considered a “moderate” to “substantial” agreement [177]. We used Weighted Cohen’s Kappa since it is the most appropriate for our ordinal codes [20].

After examining our ground truth ratings data, we determined three reasonable thresholds that researchers and practitioners may want to use to separate sub-articles from non-sub-articles. The strictest definition requires an average score of 3.0 from two coders, meaning that both gave the relationship a ‘3’. Next, we considered a threshold at an average rating of 2.5, which is more flexible but still required one coder to give a candidate relationship a ‘3’ and the other to give it a ‘2’. Finally, we also considered an average score of 2.0 as a threshold, which is the broadest definition and can come from various rating configurations.

5.4. Modeling

5.4.1. Overview

The goal of our modeling exercise was to accurately address the sub-article matching problem using machine learning techniques. In other words, our aim was to build classification models that can accurately predict whether a parent article/sub-article candidate pair $\langle p, p_{cs} \rangle$ represents a true sub-article relationship (i.e. a $\langle p, p_s \rangle$). As we discussed above, we defined this classification problem along three dimensions: (1) dataset $\{high-interest, random, ad-hoc\}$, (2) sub-article definition/threshold average rating = 2.0, 2.5, 3.0 and (3) language English, Spanish, Chinese. Our machine learning experiments, described below, allow us to assess our models' performance along each dimension.

We experimented with a variety of well-known machine learning algorithms including SVM, Random Forest, Decision Tree, Naïve Bayes, Logistic Regression, and Adaboost. In the body of this section, we report results from the algorithm with the best performance. Our emphasis here is to demonstrate that one can successfully address the sub-article matching problem using popular machine learning algorithms instead of providing a detailed performance analysis of each specific algorithm. However, we supplement this discussion with detailed classification accuracies for each machine learning algorithm and dataset configuration in Appendix A. Interestingly, in most situations, the choice of the best machine learning algorithm is consistent across languages and definitions of sub-articles within a given dataset type (i.e. *high-interest*, *random* or *ad-hoc*).

For evaluation, we followed standard practice [37] and conducted 10-fold cross validation, reporting the average accuracy across all folds. Because work in this chapter

is the first to define and attempt to solve the sub-article matching problem, we cannot compare our models' accuracies with any prior work. This is a relatively common situation when applying machine learning in HCI research [123, 230]. When it occurs, the best practice is to one compares one's results to straightforward baseline approaches (e.g., [113, 231, 296]). In this chapter, we utilized the baseline approach that we found to be most powerful: always predicting the most frequent label in the training set.

Because our models are only as powerful as the features they leverage, before describing our results, we first describe each feature in detail. The features we use are diverse, drawing from techniques ranging from simple syntax comparisons, to network metrics, to advanced natural language processing algorithms. However, nearly all of our features have one key property in common: language neutrality. This means that they can be utilized to help predict whether a sub-article candidate is really a sub-article of a given parent article, regardless of the language edition of the parent and candidate.

A frequent technique we use to make a feature that would be otherwise language-specific into one that is language neutral is immediately converting language-specific $\langle p, p_{cs} \rangle$ pairs to language-neutral concepts using Wikidata cross-language mappings. For instance, comparing the number of characters or tokens shared by the parent and sub-article candidate article titles is a feature whose output and effectiveness varies extensively across language editions (i.e. it is more useful in Western languages than Eastern languages). However, by using cross-language mappings, when considering articles from Eastern language editions, our models can take advantage of the power of this approach in Western languages by examining the titles of the equivalent concepts in English, Spanish, and so on. Our typical approach to implementing this at scale is to calculate the value of

each feature in a language-specific fashion for all articles about the same concepts as the input $\langle p, p_{cs} \rangle$ pair. We then aggregate the output of these values, e.g. using maximums, averages, ratios, or summaries.

5.4.2. Features

PotSubLangsRatio

Since assigning sub-article indicators is a manual process, we expect that if an article is a sub-article candidate for a given parent article in many different language editions, this will increase the likelihood that the candidate is a true sub-article. For instance, “History of the United States” (English) is a sub-article candidate (as indicated by the “main article” template) of “United States” (English), and the same relation is true for their corresponding articles in Spanish Wikipedia (with the `{{AP}}` template). We operationalize this feature by calculating the ratio between the number of languages in which there is a potential sub-article relationship and the number of languages in which the parent articles and sub-article candidate both have corresponding articles.

MaxTokenOverlap

This feature focuses on article titles only and considers the percentage of tokens in the parent article’s title contained within the sub-article candidate’s title. It takes the maximum token overlap of the equivalent articles in all languages. A high value signifies that the parent article and sub-article candidate share a large portion of words in their titles (in at least one language) and we hypothesized that this would represent a higher-likelihood sub-article relationship. When computing this feature, we tokenized Eastern languages

that are written in a scriptio continua pattern (no spacing or other dividers) and to match characters between Traditional Chinese and Simplified Chinese.

NumLangsRatio

This feature measures the relative “globalness” of the parent article and the sub-article candidate across all 25 languages. It is computed as the number of language editions in which the parent article has foreign language equivalents, divided by the same number for the sub-article. For instance, for the pair “Portland, Oregon”, “Sports in Portland, Oregon” in English Wikipedia, “Portland, Oregon” has corresponding articles in all 25 languages while “Sports in Portland, Oregon” only has an article in the English Wikipedia. We hypothesized that a higher value would indicate a higher likelihood of a parent/sub-article relationship because other languages might not yet have split the relevant content into two articles.

MainTemplatePct

This feature leverages one of the most prominent sub-article indicators: the main template. Main templates can be seen in Figure 1, and all 25 language editions have a version of this template. Although ambiguously used in many cases, we hypothesized that main templates had the highest precision of all the indicators. Moreover, in many languages, the usage guide for this template corresponds well with the notion of sub-articles³. We calculated this feature as follows: the number of language editions in which the sub-article candidate appears in parent article’s main template divided by the number of language

³https://en.wikipedia.org/wiki/Template:Main_article

editions in which there is any sub-article indicator between the two articles. In other words, the feature is the share of the potential sub-article relationships between two concepts defined using a main template.

MaxSectionTokenOverlap

This feature specifically considers the template-based sub-article indicators. Note that in Figure 5.2, these indicators almost always appear below a section sub-head. This feature is the direct analogue to MaxTokenOverlap, but uses the title of the preceding section sub-head rather than the titles of the articles.

MaxMainTFInSub

In all language editions, most Wikipedia articles begin with a summary of the content of the article. This feature calculates the term frequency of the parent article’s title in the summary paragraph of the sub-article candidate and takes the maximum across all languages. We hypothesized that as part of the natural editing process, when editors spin off a sub-article, they refer back to the parent article in the introduction. As such, we expected a higher value would lead to a higher likelihood of a sub-article relationship.

IndegreeRatio

This feature describes the relative centrality of the parent article and sub-article candidate in the article graph of a given language edition. We hypothesized that true sub-article relationships would more often involve a central/important parent and a less central/important sub-article than vice versa. This feature is calculated by taking the ratio

of the indegree of the parent article (i.e. the number of Wikipedia articles that contain a hyperlink to this article) and the indegree of the sub-article, each of which is summed across all languages. Indegree is commonly used as a straightforward metric of network centrality/importance in large graphs like the Wikipedia article graph [153].

MilneWitten

This feature is the MilneWitten semantic relatedness (SR) measurement [297, 212] between the parent article and sub-article candidate. We hypothesized that a higher SR between the two articles would mean that these two articles are more likely to be in true sub-article relationship. For example, in the English Wikipedia, “History of chocolate” and “Caffeine” are less related than “Sports in Portland, Oregon” and “Portland, Oregon”.

Other Features

Besides the features described above, we also tested features that consider the structural complexity of p and p_{cs} . For example, the ratio between the number of templates in a parent article and candidate sub-article and the ratio between the number of references in a parent article and a candidate sub-article. These features provided only a marginal improvement to the classification accuracy. With parsimony in mind, we did not include them in the final model construction.

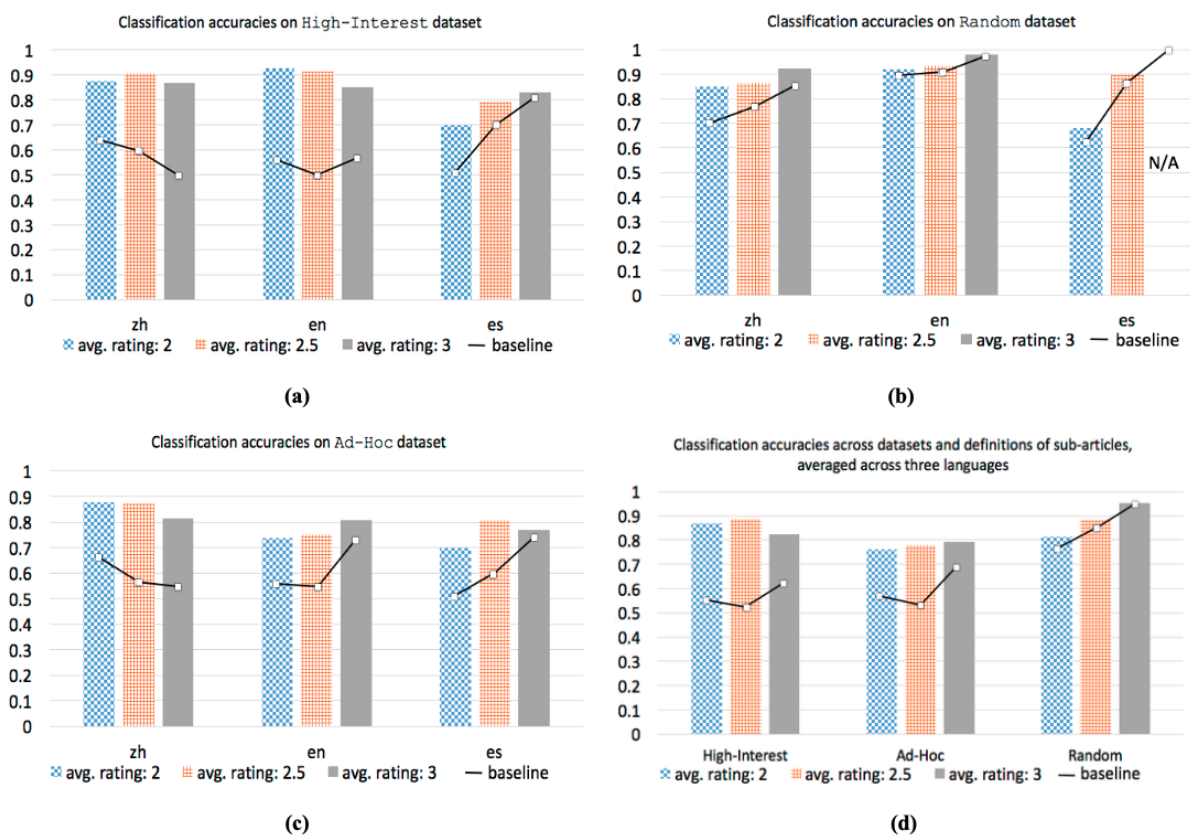


Figure 5.4. Classification accuracies across datasets, language editions and different thresholds of sub-article ratings. Each colored vertical bar shows the best accuracies among the machine learning algorithms considered, and the black line indicates baseline performance.

5.4.3. Results

5.4.3.1. On High-Interest Dataset. Figure 5.4 (a) shows the ability of our model to correctly distinguish true sub-article relationships from false ones in the High-Interest ground truth dataset according to all three definitions of sub-articles (2.0, 2.5, 3.0). Recall that the High-Interest dataset focuses on high-demand concepts that are frequently accessed by Wikipedia readers, making our model’s results on this dataset particularly important. For simplicity, although we tried the multiple machine learning algorithms

described above, we only report the one with the highest accuracy in the figure (more results are available in Appendix A). For High-Interest, Linear SVM and Random Forest alternated as the best techniques.

As can be seen in Figure 5.4 (a), in all cases, our models outperformed the baseline method by a substantial margin. On average, our model exceeded baseline performance by a factor of 1.45, and this number went up to around 1.8 in specific cases. The highest absolute accuracy was 90% (English, 2.5-average) and the lowest absolute accuracy was 70% (Spanish, 2.0-average). Overall, Figure 5.4 shows that for concepts that are in high demand, the features we have defined make it a relative straightforward task for a learned classification model to determine whether a sub-article candidate is a “true” sub-article.

5.4.3.2. On Random Dataset. Figure 5.4 (b) shows the classification results on the Random dataset. Immediately visible in the figure is that the baseline accuracies for this dataset are much higher than those for the other datasets. In the most extreme case – the 3.0-threshold Spanish dataset – the baseline accuracy reaches 100%, which makes the classification results meaningless in the context of this study. These higher baselines occur because in this dataset, the vast majority of sub-article candidates are negative examples. The necessary result of this situation is that a baseline approach that always guesses ‘no’ will be accurate most of the time, and this is a tough baseline for any algorithm to beat.

Indeed, while their absolute accuracies are quite high, our models in general only marginally improve upon the baselines with respect to this dataset. The highest classification accuracy relative to the baseline is on Chinese sub-article candidates with the 2.0-average threshold. English with a 3.0-average threshold was the only configuration in which the classifier failed to improve upon the baseline at all.

5.4.3.3. On Ad-Hoc Dataset. Figure 5.4 (c) shows the classification accuracies on the Ad-Hoc dataset. Among the machine learning algorithms, Linear SVM and Random Forest alternate as the best algorithm across different languages and definitions of sub-articles. While absolute accuracy levels on this dataset are comparable with those for High-Interest, the baseline accuracies are generally higher. This means that a smaller improvement was made by our models on this dataset relative to the baseline.

5.4.3.4. Summary of Results. Figure 5.4 (d) compares the average classification performance for each dataset across all sub-article thresholds and languages. This figure further reinforces several high-level trends mentioned above:

- On the *High-Interest* dataset that contains articles in high-demand by readers and the *Ad-Hoc* dataset that is sampled from more meaningful concepts, our classification results outperform the baseline consistently and substantially across all three languages (See Figure 5.4 (a), (c), and (d)).
- On the *Random* dataset that contains articles typically of lower interest, shorter length, and lower quality, our models generally do not make a substantial improvement compared to the baseline method (See Figure 5.4 (b) and (d)).

5.4.4. Feature Analysis

In order to understand which features were the most important to the success of our models, we examined the Random Forest versions of our models. These models have the advantage of (1) being the top or close to the top performing models in most configurations of our experiments (see Appendix) and (2) they afford straightforward analysis of feature importance. Specifically, in Random Forest models, feature importance can be evaluated

Feature	Avg. Rank of Importance
MaxMainTFInSub	2.7
MaxTokenOverlap	2.8
MaxSectionTokenOverlap	3.7
NumLangsRatio	3.8
PotSubLangsRatio	4.5
MilneWitten	5.4
IndegreeRatio	5.6
MainTemplatePct	7.3

Table 5.2. Feature importance averaged across all languages and all definitions of sub-articles on the High-Interest dataset. Feature importance is computed using a Random Forest model.

by adding up the weighted impurity decrease for all trees in the forest using an impurity function such as the Gini index or Shannon entropy [41, 42]. This approach has been used for feature selection in various domains including but not limited to bioinformatics [274, 275], image classification [99], and ecology [70].

Table 5.2 presents the importance rank for each feature on the *High-Interest* dataset (averaged across language and sub-article definition). Note that *MainTemplatePct* is the least important feature. Recall that this feature is motivated by the fact that it reflects community-defined rules for linking parent articles and sub-articles. As such, we originally expected *MainTemplatePct* to be a strong predictor of true sub-article relationships. However, even though we account for human error to some extent (by aggregating across all languages), *MainTemplatePct* remains relatively non-predictive. Closer examination of the data reveals that although these guidelines are properly documented, human editors failed to consistently follow the guidelines. For example, in the “Donald Trump” (English) article, which is one of the top ten most-viewed English articles in our page view

data, Wikipedia editors correctly tagged the true sub-article “Donald Trump presidential campaign, 2016” with the template `{{main article}}` while they incorrectly tagged the true sub-article “Donald Trump presidential campaign, 2000” with the template `{{see also}}`.

Table 5.2 also shows that *MaxSectionTokenOverlap*, *MaxTokenOverlap*, and *MaxMainTFInSub* are relatively important features. Unlike *MainTemplatePct*, which relies on editors explicitly indicating sub-article relationships as defined by community guidelines, these natural language features are implicit: they capture the lexical, linguistic and semantic relationships between parent articles and sub-articles that emerge through the natural editing process. For instance, when Wikipedia editors move content to a new sub-article, it is natural for them to add a lead sentence that points back to the parent article ⁴. This natural editing process is captured by *MaxMainTFInSub*. We believe that the variation in effectiveness between the explicit standard-based features and the implicit features may point to author-audience mismatch issues, which will be detailed in the Discussion section.

5.4.5. The Impact of the Article-as-Concept Assumption

Our trained models allow us to quantify the impact of the article-as-concept assumption. Specifically, the models allow us to ask: for how many articles and page views is the article-as-concept assumption invalid due to sub-articles?

To address this question, we deployed our High-Interest, 2.5-average threshold model on the 1,000 most-viewed articles in English Wikipedia. Table 5.3, which contains the

⁴https://en.wikipedia.org/wiki/Wikipedia:Splitting#How_to_properly_split_an_article

Impact Metric	Statistics
% of articles with sub-articles	70.8%
% of page views to articles with sub-articles	71.0%
Avg # of sub-article per article	7.5

Table 5.3. The impact of applying our model on the top 1000 most-viewed articles in English Wikipedia.

results of this analysis, shows that sub-articles cause violations of the article-as-concept assumption in a large percentage of cases. For instance, over 70% page views to this set of critical articles go to articles that contain at least one sub-article. Table 3 also reveals that on average, each of the top 1000 English Wikipedia articles has 7.5 sub-articles.

This result has important implications for user-centric Wikipedia-based technologies such as Omnipedia, Manypedia, and others. Based on the findings in Table 5.3, designers of these technologies should assume that users will frequently engage with articles that have sub-articles. Indeed, it appears that at least for the most popular articles, sub-articles are not the exception, they are the rule.

5.5. Discussion

5.5.1. Audience-Author Mismatch

Immediately above, we showed that the article-as-concept assumption, which is central to many Wikipedia-based studies and systems, fundamentally breaks down for a substantial proportion of high-value Wikipedia articles. In this section, we describe how this issue may be just one in a growing series of problems for Wikipedia-based studies and systems associated with author-audience mismatch [134].

The author-audience mismatch framework was originally intended to explain problems associated with human authors failing to sufficiently customize their content for a given audience (e.g. in a cross-cultural context). However, this framework may also be helpful for understanding the root cause of the article-as-concept assumption and its resultant problems. Namely, Wikipedia editors write Wikipedia articles for the needs of human audiences, but increasingly, Wikipedia has two additional audiences as well: Wikipedia-based studies and, in particular, Wikipedia-based systems. These audiences often have fundamentally different needs than human audiences.

It is important for Wikipedia's studies and systems audience that all content about a concept in a given language edition be contained in a single article. However, for Wikipedia editors' intended audience – other humans – doing so would violate Wikipedia's guidelines to break up long articles into parent and sub-articles. These guidelines emerge from clear human needs. For instance, lengthy articles take a long time to load with the slow Internet connections used by many Wikipedia readers [67]. Additionally, long-standing notions of web usability clearly establish that lengthy web pages result in poor user experiences [43, 221, 219].

The tension between the needs of human audiences and those of Wikipedia studies and systems is exacerbated by a few additional findings in this chapter. In our analysis of feature performance, we identified that while Wikipedia has developed a number of structured mechanisms to link sub-articles to their corresponding parent articles (e.g. main article templates), they are misused so extensively that our models found them only minimally helpful when separating true sub-article relationships from false ones. We also observed the inverse. For instance, structured constructs like “See also” sections are not

supposed to be used for strong relationships like those between parent articles and sub-articles, but editors use them this way anyway. It is not hard to see why these problems have emerged: human authors know that their human audience can click on a sub-article link if they are interested, regardless of whether it is properly encoded.

Author-audience mismatch problems may create major challenges for machines and studies beyond the article-as-concept assumption, and even in peer production datasets other than Wikipedia. For example, consider the tagging infrastructure in OpenStreetMap (OSM) – the “Wikipedia of Maps” [88]. OpenStreetMap has a robust and well-defined set of tagging practices, but recent evidence suggests that OSM editors do not follow standards when they are not perceived as necessary for their specific human audience [162]. For instance, if a human has to decide between tagging a restaurant that serves both coffee and donuts with either “coffee shop” or “donut shop”, it is unlikely they will spend time reading and strictly following the guidelines. Instead, the most likely thought process is “I’ll choose one and people know that they can probably get coffee and donuts at both types of places.” However, for an OSM-based location-based recommender system using this information as training data (e.g. [22]), this is a potentially serious source of noise if it generalizes across many tags.

Returning to Wikipedia, another example comes from how different language editions treat synonyms. Recent research by Wulczyn et al. [299] has found that different language editions opt to create articles for different synonyms of the same concept. For instance, the English Wikipedia decided to create article only for “Neoplasm” while German Wikipedia chose to create article only for “Tumor”. While this may suit the human audiences of each

individual language edition, it requires additional intelligence for a machine to establish a linkage.

5.5.2. Limitations and Future Work

While we were able to address the sub-article matching problem with good accuracy for most datasets, our solution has a few important limitations that serve as starting points for future work. First, our models were trained only on potential sub-article relationships that are explicitly encoded by Wikipedia editors. We believe this approach is appropriate as it respects the decisions of editors and almost certainly captures the vast majority of sub-article candidates. That said, it would be interesting to try to discover implicit sub-articles in an automated fashion. A system that can execute implicit sub-article discovery successfully may be useful to Wikipedia editors [292] in addition to system-builders and researchers who work with Wikipedia.

Another limitation is that our models are trained on only a small set of data from just three language editions. There are possible nuances in sub-article usage that might be missed with this limited view. A third limitation is that although our models work well on the articles that attract the most reader interest, they fail to work equally well on the large number of articles that are of lower quality and shorter length. Future work should involve designing a model focused on these articles.

Finally, since we made our features language-independent by aggregating over all language editions, we observed some scalability problems for structurally complex articles. For example, feature generation for the “United States” (English) page, which contains over 200 potential sub-articles, can take minutes. While this problem can be

easily addressed by caching the classification results, future work may want to improve the scalability of our approach to reduce pre-processing and updating time.

5.6. Conclusion

In this chapter, we analyzed how article-as-concept assumption will result in unequal information access in Wikipedia-based knowledge base, which is widely used to power location intelligence in location-aware technologies. We formulated the sub-article matching problem as a way to address the article-as-concept assumption. By developing models that draw on a diverse feature set, we addressed the sub-article matching problem with relatively high accuracy, especially for the Wikipedia articles that attract the most attention. Finally, in order to help researchers immediately address the sub-article matching problem in their own systems and studies and push this line of research forward, we have made our model and our gold standard sub-article datasets freely available for download⁵

⁵<http://z.umn.edu/WikiSubarticles>

CHAPTER 6

Supporting Projects

In this chapter, I briefly describe two additional projects led or co-led by me that support the main work in the Part 2 of this thesis. Although not directly addressing the topic of geographic inequalities in Location-Aware artificial intelligence technologies, these two projects provide important context and inspire key solutions for the work in Chapter 4 and Chapter 5.

For each of the two supporting project, I first explain their connections to this thesis and then offer a summary of their research questions, methodologies, and results.

6.1. Project 1: Mining Wikimedia Commons to Enrich News Articles

6.1.1. Connections to this Dissertation

Project 1 supports the work in Chapter 4 by offering a critical understanding of the diversity, magnitude, and quality of the images on Wikimedia Commons. Originally published as a full paper on the 2018 World Wide Web Conference [186], this project built a system that mines data visualizations in Wikimedia Commons and recommends suitable ones for contextualizing the facts and numbers referenced in news articles. As one of the first projects that leveraged Wikimedia Commons' images, this project demonstrated the great diversity, quality, and volume of this dataset, motivating us to use it to enhance the geodiversity of ImageNet.

6.1.2. Summary of Project 1

Data visualizations in news articles (e.g., maps, line graphs, bar charts) greatly enrich the content of news articles and result in well-established improvements to reading comprehension. However, existing systems that generate news data visualizations either require substantial manual effort or are limited to very specific types of data visualizations, thereby greatly restricting the number of news articles that can be enhanced. To address this issue, we define a new problem: given a news article, retrieve relevant visualizations that already exist on the web. We show that this problem is tractable through a new system, VizByWiki, that mines contextually relevant data visualizations from Wikimedia Commons, the central file repository for Wikipedia. Using a novel ground truth dataset, we show that VizByWiki can successfully augment as many as 48% of popular online news articles with news visualizations. We also demonstrate that VizByWiki can automatically rank visualizations according to their usefulness with reasonable accuracy (nDCG@5 of 0.82).

6.2. Project 2: Understanding the Geographic Inequalities in Pokemon Go

6.2.1. Connections to this Dissertation

Project 2 supports the Part 2 overall by understanding the geographic inequalities in a different type of location-aware computing technologies - location-based game. Published as a full paper on 2017 CHI Conference on Human Factors in Computing Systems [61], this project found that the location-based games also contain geographic inequalities similar to location-based AI technologies. Specifically, similar to the trend in computer-vision-based location-aware AI technologies that places with better economic status enjoy

higher performance, places with better economic status such as urban areas or areas whose predominant residence are of privileged race have substantially more resources in Pokemon Go, driving large crowd of players towards these geographies and potentially reinforce their economic advantages.

6.2.2. Summary of Project 2

The widespread popularity of Pokémon GO presents the first opportunity to observe the geographic effects of location-based gaming at scale. This paper reports the results of a mixed methods study of the geography of Pokémon GO that includes a five-country field survey of 375 Pokémon GO players and a large scale geostatistical analysis of game elements. Focusing on the key geographic themes of places and movement, we find that the design of Pokémon GO reinforces existing geographically-linked biases (e.g. the game advantages urban areas and neighborhoods with smaller minority populations), that Pokémon GO may have instigated a relatively rare large-scale shift in global human mobility patterns, and that Pokémon GO has geographically-linked safety risks, but not those typically emphasized by the media. Our results point to geographic design implications for future systems in this space such as a means through which the geographic biases present in Pokémon GO may be counteracted.

CHAPTER 7

Future Work

The work we described in this thesis should be the start, but not the end, towards the goal of understanding and addressing the safety issues and geographic biases in location-aware computing technologies. Inspired by the results presented in this thesis, this chapter envisions the short-term and long-term future work for tackling these two emergent risks.

7.1. For the Safety Issues in Location-Aware Computing

7.1.1. Short-Term: Further Improving the Safety of Personal Navigation Systems

In short term, future work should leverage the results in this thesis to further improve the safety of personal navigation systems. Below, we describe some specific directions.

The most straightforward future work is to leverage the road safety prediction model presented in Chapter 3 to build a new routing algorithms that prioritize road safety. As we briefly mentioned in 3, the binary safety labels predicted by our models are designed to be easy to use by navigation systems. For example, researchers of personal navigation systems have developed many alternative routing algorithms (e.g., [201, 257, 158]) which optimize routing results based on criteria other than time or distance. These algorithms can easily incorporate our binary safety labels as an additional term in their cost function for generating routes that prioritize road safety. During this process, researchers should also study what thresholds to use for considering a route safe/unsafe based on some initial

evidence we presented in Chapter 3. For example, should the algorithms label a route with 90% of its road segments unsafe segments a unsafe route? Or 80%?

Second direction of future work is to extend our work to investigate the safety issues with the navigation module in semi-autonomous or self-driving vehicles. As the vehicles become more intelligent, navigation and route planning that are currently performed by human will be increasingly delegated to the navigation module of the vehicle, making the safety and robustness of the navigation systems a fundamental need. While some design implications for enhancing the safety of navigation systems presented in Chapter 2 will be moot in the self-driving era (e.g., those that are related to visual and audio instructions), other implications related to the underlying geographic data and algorithms might be increasingly important. Future work on this topics should compare and contextualize the results Chapter 2 in the semi-autonomous vehicle scenario.

7.1.2. Long-Term: Study Safety Issues in Other Location-Aware Computing Systems

Location-aware computing is unique that their users don't simply use these technologies in front of their desks - they bring the technologies with them and use the technologies where ever they go. This unique characteristic of mobility makes users of location-aware technologies more susceptible to physical danger than many other types of computing devices.

As such, future work should investigate safety issues in other types of location-aware technologies beyond personal navigation technologies. One example might be location-based AR games such as Pokemon Go, which have already been reported to lead to bodily injuries due to distractions or guiding players to dangerous locations [17, 254].

7.2. Geographic Inequalities in Location-Aware Artificial Intelligence Technologies

7.2.1. Short-Term: Further Addressing Geographic Biases in Computer Vision-Based Location-Aware AI Technologies

In the short-term, future work should finish executing the research plan laid out in the Future Work section of Chapter 4.8. Specifically, the work in this thesis accomplished the first two steps in the research plan - Step 1, which is to enhance ImageNet with millions of geographically diverse images and Step 2, which is viability test to demonstrate that these geographically diverse images can improve pre-trained models' accuracies in developing countries using selected high-profile classes. Future work should continue to Step 3, which is to replicate the Step 2 study on greater number of ImageNet classes and Step 4, which ultimately tested the effect in computer vision-based location aware computing systems which heavily rely on pre-training models.

In addition to finishing the research plan of mitigating geographic inequalities of computer vision pre-training models, the ImageNet-compatible geodiversity data we generated also made it possible, for the first time, to explore a fundamental question in vision knowledge - which concepts have homogeneous visual appearances across the global and which concepts have heterogeneous visual appearances. From a limited examples we saw

in our work that, for example, the visual appearances of weddings and bridegroom have relatively large geographic differences on their visual appearance, making it easy for machine learning models to have unequal performances on these classes. However, are these classes unique? Are there any other concepts/classes also susceptible to the same issue? Such questions are critical for building comprehensive visual knowledge of our world and for gaining a fine-grained understanding of the geo-generalizability of computer vision pre-training models.

7.2.2. Long-Term: Developing Location-Aware Technologies for Developing Countries

In the long-term, future work should focus on developing better location-aware computing technologies for developing countries. As the emerging markets with billions of mobile users, developing countries is where many technology companies wanted to deploy their location-aware applications and products. However, while many of these applications and products enjoy great success in their home countries, they might not be directly deployable in emerging markets. The computer vision-based location-aware technologies is just one such example. One other prominent example is Google Maps mode of transportation for developing countries in Southeast Asia such as India [102]. While enormously useful in developed country, the traditional car-based routing is much less useful in India where the primary mode of transportation is motorcycle. To cope with the unique situation in India, Google Map developed new routing algorithms for motorcycles that go through narrow

paths and shortcuts and offered new landmark-based turn-by-turn navigation instructions. Future work should seek to better understand and summarize the different types of adaptation location-aware computing systems need to make for developing countries.

CHAPTER 8

Conclusion

Location-aware computing technologies have become increasingly integrated with our personal life and social activities. As the popularity and importance of these technologies grew, new types of negative impacts emerged where these technologies sometimes brought unintended consequences to its users and the society.

In this thesis, we seek to understand and address two new risks in location-aware computing technologies - safety and geographic inequalities. In the first part of the thesis which focuses on the topic of safety, we ground our work in one of the most popular location-aware technologies - personal navigation systems and focus on the factors that contribute to the catastrophic car incidents involving this technology. To build a reliable dataset for these catastrophic incidents which are difficult to collect, we introduced a novel data collection method from public health research into the human-computer interaction research. Using this novel dataset, we identified a number of key technological causes for these catastrophic incidents, including the lack of road safety attribute in the navigation systems, wrong geocoding, insufficient visual instructions etc.

We then focus on mitigating one of the key technological causes - the lack of road safety information in the navigation systems. Specifically, leveraging well-established standard in transportation studies, we defined binary road safety label that can be easily used in navigation systems and built machine learning-based systems to automatically predict these binary road safety labels for large-scale road segments for navigation systems to

use. Our machine learning-based system leverages diverse geographic user-generated road attribute data from OpenStreetMap and achieved desirable accuracy in different countries.

In the second part of the thesis, we turn our attention to addressing the geographic inequalities of AI powered location-aware technologies. We identified that the geographic inequalities in these location-aware technologies are actually rooted in the underlying core AI technologies, such as computer vision, natural language processing and knowledge base technologies. As such, we focused on mitigating the geographic inequalities in two core AI technologies that location-aware computing frequently rely on - computer vision and knowledge graph. For computer vision, we address the geographic inequalities in pre-training models through enhancing the geodiversity of these large image datasets these models are trained on. To do so, we extracted millions of images from diverse geographies and produced corresponding labels for these images. Using these geographically diverse images, we show that pre-training models can gain significant and non-trivial improvements on previously underperforming developing countries.

Finally, we address the geographic inequalities in location-aware technologies that rely on Wikipedia-based knowledge base. We described the problem, in which these knowledge base technologies have unequal access to the Wikipedia content about geographic concepts and identified an erroneous "article-as-concept" assumption which led to the unequal access. In response, we built machine-learning based systems to ensure that knowledge base will have equally access complete information for all geographic concepts.

As mentioned before, we believe our work should be the start, not the end, towards addressing the negative impact of location-aware technologies. In this regards, we hope future work will continue this endeavor to mitigate the individual and societal negative

impact of location-aware technology. We point out a number of future directions, including but not limited to extending the safety analyses to other location-aware computing technologies and further adapting the technologies to developing world.

References

- [1] Valhalla. <https://github.com/valhalla/valhalla>.
- [2] Wikipedia:100,000 feature-quality articles. https://en.wikipedia.org/wiki/Wikipedia:100,000_feature-quality_articles.
- [3] , N., ZAFAR, M. B., GUMMADI, K. P., AND WELLER, A. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law* (2016), vol. 1, p. 2.
- [4] ABADI, M., AGARWAL, A., BARHAM, P., BREVDO, E., CHEN, Z., CITRO, C., CORRADO, G. S., DAVIS, A., DEAN, J., DEVIN, M., ET AL. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).
- [5] AGRESTI, A. *An introduction to categorical data analysis*. John Wiley & Sons, 2018.
- [6] AHMAD, S. A. R. S. T. B. N. Road safety risk evaluation using gis-based data envelopment analysis—artificial neural networks approach. In *Appl. Sci* (2017).
- [7] AL MAHMUD, A., MUBIN, O., AND SHAHID, S. User experience with in-car gps navigation systems: comparing the young and elderly drivers. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services* (2009), pp. 1–2.
- [8] ALCANTARILLA, P. F., STENT, S., ROS, G., ARROYO, R., AND GHERARDI, R. Street-view change detection with deconvolutional networks. *Autonomous Robots* 42, 7 (2018), 1301–1322.
- [9] ALVES, L. G. A., RIBEIRO, H. V., AND RODRIGUES, F. A. Crime prediction through urban metrics and statistical learning. In *Physica A* (2018), vol. 505, p. 435–443.
- [10] ANDREW ESTABROOKS, TAEHO JO, N. J. A multiple resampling method for learning from imbalanced data sets. In *Computational Intelligence 20* (2004), pp. 18–36.

- [11] ANTHONY, L. F. W., KANDING, B., AND SELVAN, R. Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv:2007.03051* (2020).
- [12] ANTONIOU, V., AND SKOPELITI, A. Measures and indicators of vgi quality: An overview. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences 2* (2015), 345.
- [13] ANTUNES, A., SECO, A., AND PINTO, N. An accessibility–maximization approach to road network planning. *Computer-Aided Civil and Infrastructure Engineering* 18, 3 (2003), 224–240.
- [14] APORTA, C., HIGGS, E., HAKKEN, D., PALMER, L., PALMER, M., RUNDSTROM, R., PFAFFENBERGER, B., WENZEL, G., WIDLOK, T., APORTA, C., ET AL. Satellite culture: global positioning systems, inuit wayfinding, and the need for a new account of technology. *Current anthropology* 46, 5 (2005), 729–753.
- [15] ARDAGNA, C. A., CREMONINI, M., DAMIANI, E., DI VIMERCATI, S. D. C., AND SAMARATI, P. Location privacy protection through obfuscation-based techniques. In *IFIP Annual Conference on Data and Applications Security and Privacy* (2007), Springer, pp. 47–60.
- [16] AUER, S., BIZER, C., KOBILAROV, G., LEHMANN, J., CYGANIAK, R., AND IVES, Z. Dbpedia: A nucleus for a web of open data. In *The semantic web*. Springer, 2007, pp. 722–735.
- [17] AYERS, J. W., LEAS, E. C., DREDZE, M., ALLEM, J.-P., GRABOWSKI, J. G., AND HILL, L. Pokémon go—a new distraction for drivers and pedestrians. *JAMA internal medicine* 176, 12 (2016), 1865–1866.
- [18] BACA, M. C. Google maps now shows speed traps, potentially raising the ire of law enforcement, 2019.
- [19] BACH, K. M., JÆGER, M. G., SKOV, M. B., AND THOMASSEN, N. G. Interacting with in-vehicle systems: understanding, measuring, and evaluating attention. In *Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology* (2009), British Computer Society, pp. 453–462.
- [20] BAKEMAN, R., AND GOTTMAN, J. M. *Observing interaction: An introduction to sequential analysis*. Cambridge university press, 1997.

- [21] BALALI, V., RAD, A. A., AND GOLPARVAR-FARD, M. Detection, classification, and mapping of us traffic signs using google street view images for roadway inventory management. *Visualization in Engineering* 3, 1 (2015), 15.
- [22] BALLATORE, A., MCARDLE, G., KELLY, C., AND BERTOLOTTO, M. Recomap: an interactive and adaptive map-based recommender. In *Proceedings of the 2010 ACM Symposium on Applied Computing* (2010), pp. 887–891.
- [23] BAO, P., HECHT, B., CARTON, S., QUADERI, M., HORN, M., AND GERGLE, D. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), pp. 1075–1084.
- [24] BARBOSA, N. M., AND CHEN, M. Rehumanized crowdsourcing: a labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–12.
- [25] BARK, K., TRAN, C., FUJIMURA, K., AND NG-THOW-HING, V. Personal navi: Benefits of an augmented reality navigational aid using a see-thru 3d volumetric hud. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (2014), ACM, pp. 1–8.
- [26] BAROCAS, S., HARDT, M., AND NARAYANAN, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. <http://www.fairmlbook.org>.
- [27] BAULLINGER, J., QUAN, L., BENNETT, E., CUMMINGS, P., AND WILLIAMS, K. Use of washington state newspapers for submersion injury surveillance. *Injury Prevention* 7, 4 (2001), 339–342.
- [28] BEATTIE, D., BAILLIE, L., HALVEY, M., AND MCCALL, R. Adapting satnav to meet the demands of future automated vehicles. In *CHI 2015 workshop on experiencing autonomous vehicles: crossing the boundaries between a drive and a ride* (2015).
- [29] BEENSTOCK, M., AND GAFNI, D. Globalization in road safety: explaining the downward trend in road accident rates in a single country (israel). *Accident Analysis & Prevention* 32, 1 (2000), 71–84.
- [30] BENKLER, Y., AND NISSENBAUM, H. Commons-based peer production and virtue. *Journal of political philosophy* 14, 4 (2006), 394–419.
- [31] BENKLER, Y., SHAW, A., AND HILL, B. M. Peer production: A form of collective intelligence. *Handbook of collective intelligence* 175 (2015).

- [32] BERESFORD, A. R., AND STAJANO, F. Location privacy in pervasive computing. *IEEE Pervasive computing* 2, 1 (2003), 46–55.
- [33] BERESFORD, A. R., AND STAJANO, F. Mix zones: User privacy in location-aware services. In *IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004. Proceedings of the Second* (2004), IEEE, pp. 127–131.
- [34] BERG, T., LIU, J., WOO LEE, S., ALEXANDER, M. L., JACOBS, D. W., AND BELHUMEUR, P. N. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 2011–2018.
- [35] BILLINGS, C., LAUBER, J., FUNKHOUSER, H., LYMAN, E., AND HUFF, E. Nasa aviation safety reporting system.
- [36] BISCHOFF, K., FIRAN, C. S., NEJDL, W., AND PAIU, R. Can all tags be used for search? In *Proceedings of the 17th ACM conference on Information and knowledge management* (2008), pp. 193–202.
- [37] BISHOP, C. M. *Pattern recognition and machine learning*. springer, 2006.
- [38] BOLL, S., PALANQUE, P., MIRNIG, A. G., CAUCHARD, J., LÜTZHÖFT, M. H., AND FEARY, M. S. Designing safety critical interactions: Hunting down human error. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–7.
- [39] BOLUKBASI, T., CHANG, K.-W., ZOU, J. Y., SALIGRAMA, V., AND KALAI, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems* (2016), pp. 4349–4357.
- [40] BOWEN, J., AND STAVRIDOU, V. Safety-critical systems, formal methods and standards. *Software engineering journal* 8, 4 (1993), 189–209.
- [41] BREIMAN, L. Random forests. *Machine learning* 45, 1 (2001), 5–32.
- [42] BREIMAN, L. Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA* 1 (2002), 58.
- [43] BRINCK, T., GERGLE, D., AND WOOD, S. D. *Usability for the web: Designing web sites that work*. Elsevier, 2001.

- [44] BROWN, B., AND LAURIER, E. The normal natural troubles of driving with gps. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2012), pp. 1621–1630.
- [45] BUOLAMWINI, J., AND GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (2018), pp. 77–91.
- [46] CALVERT, S. L., AND TAN, S.-L. Impact of virtual reality on young adults' physiological arousal and aggressive thoughts: Interaction versus observation. *Journal of applied developmental psychology* 15, 1 (1994), 125–139.
- [47] CARREIRA, J., AGRAWAL, P., FRAGKIADAKI, K., AND MALIK, J. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 4733–4742.
- [48] CASNER, S. M., HUTCHINS, E. L., AND NORMAN, D. The challenges of partially automated driving. *Communications of the ACM* 59, 5 (2016), 70–77.
- [49] CHANCELLOR, S., KALANTIDIS, Y., PATER, J. A., DE CHOUDHURY, M., AND SHAMMA, D. A. Multimodal classification of moderated online pro-eating disorder content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), pp. 3213–3226.
- [50] CHANG, S. E., HSIEH, Y.-J., LEE, T.-R., LIAO, C.-K., AND WANG, S.-T. A user study on the adoption of location based services. In *Advances in web and network technologies, and information management*. Springer, 2007, pp. 276–286.
- [51] CHAWLA, N. V., BOWYER, K. W., HALL, L. O., AND KEGELMEYER, W. P. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.* 16, 1 (June 2002), 321–357.
- [52] CHEN, H., ENGKVIST, O., WANG, Y., OLIVECRONA, M., AND BLASCHKE, T. The rise of deep learning in drug discovery. *Drug discovery today* 23, 6 (2018), 1241–1250.
- [53] CHEN, L., DING, Y., LYU, D., LIU, X., AND LONG, H. Deep multi-task learning based urban air quality index modelling. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (5 2017).
- [54] CHEN, T., AND GUESTRIN, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (8 2016), ACM, p. 785–794.

- [55] CHEN, W., AND WELLMAN, B. The global digital divide—within and between countries. *IT & society* 1, 7 (2004), 39–45.
- [56] CHOULDECHOVA, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
- [57] CHU, B., MADHAVAN, V., BEJBOM, O., HOFFMAN, J., AND DARRELL, T. Best practices for fine-tuning visual classifiers to new domains. In *European conference on computer vision* (2016), Springer, pp. 435–442.
- [58] COBB, S. V. G. Measurement of postural stability before and after immersion in a virtual environment. *Applied ergonomics* 30, 1 (1999), 47–57.
- [59] COHEN, J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin* 70, 4 (1968), 213.
- [60] COHEN, J. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.
- [61] COLLEY, A., THEBAULT-SPIEKER, J., LIN, A. Y., DEGRAEN, D., FISCHMAN, B., HÄKKILÄ, J., KUEHL, K., NISI, V., NUNES, N. J., WENIG, N., ET AL. The geography of pokémon go: beneficial and problematic effects on places and movement. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), pp. 1179–1192.
- [62] CONTRIBUTORS, O. Map features - openstreetmap wiki. https://wiki.openstreetmap.org/wiki/Map_Features. (Accessed on 01/18/2019).
- [63] CONTRIBUTORS, O. Openstreetmap: 'it's the wikipedia of maps', 2012. <https://www.theguardian.com/theobserver/2012/feb/18/openstreetmap-world-map-radicals>.
- [64] CONTRIBUTORS, O. Major openstreetmap consumers, 2019. https://wiki.openstreetmap.org/wiki/Major_OpenStreetMap_Consumers.
- [65] CONTRIBUTORS, W. Wikipedia article: Church architecture - divergence of eastern and western church architecture. https://en.wikipedia.org/wiki/Church_architecture#Divergence_of_Eastern_and_Western_church_architecture.
- [66] CONTRIBUTORS, W. Wikipedia:article size. https://en.wikipedia.org/wiki/Wikipedia:Article_size.

- [67] CONTRIBUTORS, W. Wikipedia:article size. https://en.wikipedia.org/wiki/Wikipedia:Article_size.
- [68] CONTRIBUTORS, W. Wikipedia:wikipedians. <https://en.wikipedia.org/wiki/Wikipedia:Wikipedians>.
- [69] CONTRIBUTORS, W. C. Commons:project scope. https://commons.wikimedia.org/wiki/Commons:Project_scope#Aim_of_Wikimedia_Commons.
- [70] CUTLER, D. R., EDWARDS JR, T. C., BEARD, K. H., CUTLER, A., HESS, K. T., GIBSON, J., AND LAWLER, J. J. Random forests for classification in ecology. *Ecology* 88, 11 (2007), 2783–2792.
- [71] DALTON, P., AGARWAL, P., FRAENKEL, N., BAICHO, J., AND MASRY, A. Driving with navigational instructions: Investigating user behaviour and performance. *Accident Analysis & Prevention* 50 (2013), 298–303.
- [72] DE VRIES, T., MISRA, I., WANG, C., AND VAN DER MAATEN, L. Does object recognition work for everyone? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2019), pp. 52–59.
- [73] DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., AND FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (2009), Ieee, pp. 248–255.
- [74] DÍAZ, M., JOHNSON, I., LAZAR, A., PIPER, A. M., AND GERGLE, D. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), pp. 1–14.
- [75] DIFALLAH, D., FILATOVA, E., AND IPEIROTIS, P. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining* (2018), pp. 135–143.
- [76] DINGUS, T. A., AND HULSE, M. C. Some human factors design issues and recommendations for automobile navigation information systems. *Transportation Research Part C: Emerging Technologies* 1, 2 (1993), 119–131.
- [77] DINGUS, T. A., HULSE, M. C., ANTIN, J. F., AND WIERWILLE, W. W. Attentional demand requirements of an automobile moving-map navigation system. *Transportation Research Part A: General* 23, 4 (1989), 301–315.

- [78] DONG, Z., SHI, C., SEN, S., TERVEEN, L., AND RIEDL, J. War versus inspirational in forrest gump: Cultural effects in tagging communities. In *Sixth International AAAI Conference on Weblogs and Social Media* (2012).
- [79] DUBEY, A., NAIK, N., PARIKH, D., RASKAR, R., AND HIDALGO, C. A. Deep learning the city: Quantifying urban perception at a global scale. In *European conference on computer vision* (2016), Springer, pp. 196–212.
- [80] DWORK, C., HARDT, M., PITASSI, T., REINGOLD, O., AND ZEMEL, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (2012), pp. 214–226.
- [81] ERICKSON, K., PEREZ, F. R., AND PEREZ, J. R. What is the commons worth? estimating the value of wikimedia imagery by observing downstream use. In *Proceedings of the 14th International Symposium on Open Collaboration* (2018), pp. 1–6.
- [82] ERXLEBEN, F., GÜNTHER, M., KRÖTZSCH, M., MENDEZ, J., AND VRANDEČIĆ, D. Introducing wikidata to the linked data web. In *International Semantic Web Conference* (2014), Springer, pp. 50–65.
- [83] FACTS, T. S. A compilation of motor vehicle crash data from the fatality analysis reporting system and the general estimates system, nhtsa, us department of transportation, 2014. *Google Scholar* (2013).
- [84] FERRADA, S., BUSTOS, B., AND HOGAN, A. Imgpedia: a linked dataset with content-based analysis of wikimedia images. In *International Semantic Web Conference* (2017), Springer, pp. 84–93.
- [85] FINE, P. R., JONES, C. S., WRIGLEY, J. M., RICHARDS, J. S., AND ROUSCULP, M. D. Are newspapers a viable source for intentional injury surveillance data? *Southern medical journal* 91, 3 (1998), 234–242.
- [86] FOUNDATION, W. Media statistics. <https://commons.wikimedia.org/wiki/Special:MediaStatistics>.
- [87] FOUNDATION, W. Wikimedia downloads.
- [88] FOX, K. Openstreetmap: It’s the “wikipedia of maps”. <https://www.theguardian.com/theobserver/2012/feb/18/openstreetmap-world-map-radicals>.

- [89] FU, K., LU, Y.-C., AND LU, C.-T. Treads: A safe route recommender using social media mining and text summarization. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2014), ACM, pp. 557–560.
- [90] FUSCO, S. J., MICHAEL, K., MICHAEL, M., AND ABBAS, R. Exploring the social implications of location based social networking: an inquiry into the perceived positive and negative impacts of using lbsn between friends. In *2010 Ninth International Conference on Mobile Business and 2010 Ninth Global Mobility Roundtable (ICMB-GMR)* (2010), IEEE, pp. 230–237.
- [91] FÜR INFORMATIK, M.-P.-I. Evaluation of the yago3 accuracy. <https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/statistics/>.
- [92] GALLAGHER, A., JOSHI, D., YU, J., AND LUO, J. Geo-location inference from image content and user tags. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2009), IEEE, pp. 55–62.
- [93] GARDONY, A. L., BRUNYÉ, T. T., MAHONEY, C. R., AND TAYLOR, H. A. How navigational aids impair spatial memory: Evidence for divided attention. *Spatial Cognition & Computation* 13, 4 (2013), 319–350.
- [94] GARSON, G. D., BIGGS, R. S., AND BIGGS, R. S. *Analytic mapping and geographic databases*. No. 87. Sage, 1992.
- [95] GEBRU, T., KRAUSE, J., WANG, Y., CHEN, D., DENG, J., AIDEN, E. L., AND FEI-FEI, L. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences* 114, 50 (2017), 13108–13113.
- [96] GEBRU, T., KRAUSE, J., WANG, Y., CHEN, D., DENG, J., AND FEI-FEI, L. Visual census: Using cars to study people and society. *Bigvision* (2015).
- [97] GEBRU, T., KRAUSE, J., WANG, Y., CHEN, D., DENG, J., AND FEI-FEI, L. Fine-grained car detection for visual census estimation. In *Thirty-First AAAI Conference on Artificial Intelligence* (2017).
- [98] GERSHGORN, D. The data that transformed ai research—and possibly the world. *Quartz. Quartz. July 26* (2017), 2017.
- [99] GISLASON, P. O., BENEDIKTSSON, J. A., AND SVEINSSON, J. R. Random forests for land cover classification. *Pattern Recognition Letters* 27, 4 (2006), 294–300.

- [100] GOLOB, T. F., AND RECKER, W. W. Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions. *Journal of transportation engineering* 129, 4 (2003), 342–353.
- [101] GOODCHILD, M. F. A geographer looks at spatial information theory. In *International Conference on Spatial Information Theory* (2001), Springer, pp. 1–13.
- [102] GOOGLE. Designing google maps for motorbikes. <https://design.google/library/designing-google-maps-motorbikes/>.
- [103] GOOGLE. Open images extended - crowdsourced extension. <https://storage.googleapis.com/openimages/web/extended.html>.
- [104] GRAELLS-GARRIDO, E., LALMAS, M., AND MENCZER, F. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (2015), pp. 165–174.
- [105] GRAHAM, M. Internet geographies: Data shadows and digital divisions of labour. *Graham, M* (2014), 99–116.
- [106] GRAHAM, M., HOGAN, B., STRAUMANN, R. K., AND MEDHAT, A. Uneven geographies of user-generated information: Patterns of increasing informational poverty. *Annals of the Association of American Geographers* 104, 4 (2014), 746–764.
- [107] GREENFELD, J. S. Matching gps observations to locations on a digital map. In *81th annual meeting of the transportation research board* (2002), vol. 1, Washington, DC, pp. 164–173.
- [108] GROUP, G. R. Openrouteservice. <https://github.com/GIScience/openrouteservice>, 2020.
- [109] GRUTESER, M., AND GRUNWALD, D. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st international conference on Mobile systems, applications and services* (2003), pp. 31–42.
- [110] HAHMANN, S., AND BURGHARDT, D. How much information is geospatially referenced? networks and cognition. *International Journal of Geographical Information Science* 27, 6 (2013), 1171–1189.
- [111] HAJI, G. A. Towards a road safety development index (rsdi): Development of an international index to measure road safety performance. *Linköping University Electronic Press: Linköping, Sweden* (2005).

- [112] HAKLAY, M., BASIOUKA, S., ANTONIOU, V., AND ATHER, A. How many volunteers does it take to map an area well? the validity of linus' law to volunteered geographic information. *The cartographic journal* 47, 4 (2010), 315–322.
- [113] HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., AND WITTEN, I. H. The weka data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.
- [114] HANDY, S. L., AND CLIFTON, K. J. Evaluating neighborhood accessibility: Possibilities and practicalities. *Journal of transportation and statistics* 4, 2/3 (2001), 67–78.
- [115] HANSON, H. Gps leads japanese tourists to drive into australian bay. *The Huffington Post* (2012). http://www.huffingtonpost.com/2012/03/19/gps-tourists-australia_n_1363823.html.
- [116] HAQUE, S., KULIK, L., AND KLIPPEL, A. Algorithms for reliable navigation and wayfinding. In *International Conference on Spatial Cognition* (2006), Springer, pp. 308–326.
- [117] HARWOOD, D. W., SOULEYRETTE, R. R., FIELDS, M. A., AND GREEN, E. R. Comparison of countermeasure selection methods for use in road safety management 2. In *Presented at the 5th International Symposium on Highway Geometric Design* (2015).
- [118] HE, H., AND GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.
- [119] HE, K., GIRSHICK, R., AND DOLLÁR, P. Rethinking imagenet pre-training. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 4918–4927.
- [120] HE, S., LIN, A. Y., ADAR, E., AND HECHT, B. The_tower_of_babel.jpg: Diversity of visual encyclopedic knowledge across wikipedia language editions. In *Twelfth International AAAI Conference on Web and Social Media* (2018).
- [121] HECHT, B., AND GERGLE, D. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on Communities and technologies* (2009), pp. 11–20.
- [122] HECHT, B., AND GERGLE, D. The tower of babel meets web 2.0: user-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2010), pp. 291–300.

- [123] HECHT, B., HONG, L., SUH, B., AND CHI, E. H. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2011), pp. 237–246.
- [124] HECHT, B., SCHÖNING, J., ERICKSON, T., AND PRIEDHORSKY, R. Geographic human-computer interaction. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 2011, pp. 447–450.
- [125] HECHT, B., AND STEPHENS, M. A tale of cities: Urban biases in volunteered geographic information. In *Eighth International AAAI Conference on Weblogs and Social Media* (2014).
- [126] HENDERSON, P., HU, J., ROMOFF, J., BRUNSKILL, E., JURAFSKY, D., AND PINEAU, J. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research* 21, 248 (2020), 1–43.
- [127] HENGL, T., DE JESUS, J. M., HEUVELINK, G. B., GONZALEZ, M. R., KILIBARDA, M., BLAGOTIĆ, A., SHANGGUAN, W., WRIGHT, M. N., GENG, X., BAUER-MARSCHALLINGER, B., ET AL. Soilgrids250m: Global gridded soil information based on machine learning. *PLoS one* 12, 2 (2017).
- [128] HILE, H., GRZESZCZUK, R., LIU, A., VEDANTHAM, R., KOŠECKA, J., AND BORRIELLO, G. Landmark-based pedestrian navigation with enhanced spatial reasoning. In *International Conference on Pervasive Computing* (2009), Springer, pp. 59–76.
- [129] HIPPEL, M., SCHAUB, F., KARGL, F., AND WEBER, M. Interaction weaknesses of personal navigation devices. In *Proceedings of the 2nd International Conference on automotive user interfaces and interactive vehicular applications* (2010), pp. 129–136.
- [130] HOFFART, J., SUCHANEK, F. M., BERBERICH, K., AND WEIKUM, G. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence* 194 (2013), 28–61.
- [131] HONG, L., CONVERTINO, G., AND CHI, E. H. Language matters in twitter: A large scale study. In *Fifth international AAAI conference on weblogs and social media* (2011).
- [132] HOPF, K., DAGEFÖRDE, F., AND WOLTER, D. Identifying the geographical scope of prohibition signs. In *International Conference on Spatial Information Theory* (2015), Springer, pp. 247–267.

- [133] HORBERRY, T., ANDERSON, J., REGAN, M. A., TRIGGS, T. J., AND BROWN, J. Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. *Accident Analysis & Prevention* 38, 1 (2006), 185–191.
- [134] HORNIKX, J., AND O’KEEFE, D. J. Conducting research on international advertising: The roles of cultural knowledge and international research teams. *Journal of Global Marketing* 24, 2 (2011), 152–166.
- [135] HU, Y., GAO, S., LUNGA, D., LI, W., NEWSAM, S., AND BHADURI, B. Geoai at acm sigspatial: progress, challenges, and future directions. *SIGSPATIAL Special* 11, 2 (2019), 5–15.
- [136] HU, Y., GAO, S., NEWSAM, S. D., AND LUNGA, D. Geoai 2018 workshop report the 2nd acm sigspatial international workshop on geoai: Ai for geographic knowledge discovery seattle, wa, usa-november 6, 2018. *SIGSPATIAL special* 10, 3 (2018), 16.
- [137] HUANG, S., XU, Z., TAO, D., AND ZHANG, Y. Part-stacked cnn for fine-grained visual categorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 1173–1182.
- [138] HUH, M., AGRAWAL, P., AND EFROS, A. A. What makes imagenet good for transfer learning? *arXiv preprint arXiv:1608.08614* (2016).
- [139] HYUN, M.-H., LEE, C.-H., KIM, H.-J., TONG, Y., AND PARK, S.-S. Systematic review and meta-analysis of robotic surgery compared with conventional laparoscopic and open resections for gastric carcinoma. *British Journal of Surgery* 100, 12 (2013), 1566–1578.
- [140] IMF. World economic outlook database—weo groups and aggregates information, 2018. <https://www.imf.org/external/pubs/ft/weo/2018/02/weodata/groups.htm>.
- [141] IPEIROTIS, P. G. Demographics of mechanical turk.
- [142] IRAP. Here’s how irap can help you save lives, 2017. https://www.irap.org/how-we-can-help/?et_open_tab=et_pb_tab_1#mytabs|1.
- [143] IRAP. International road assesment programme:, 2017. <https://www.irap.org/>.
- [144] JANOWICZ, K., GAO, S., MCKENZIE, G., HU, Y., AND BHADURI, B. Geoai: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond, 2020.

- [145] JENSEN, B. S., SKOV, M. B., AND THIRURAVICHANDRAN, N. Studying driver attention and behaviour for three configurations of gps navigation in real traffic driving. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), pp. 1271–1280.
- [146] JOHNSON, I., AND HECHT, B. Structural causes of bias in crowd-derived geographic information: Towards a holistic understanding. *Retrieved September 17* (2015), 2016.
- [147] JOHNSON, I., HENDERSON, J., PERRY, C., SCHÖNING, J., AND HECHT, B. Beautiful... but at what cost?: An examination of externalities in geographic vehicle routing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 15.
- [148] JOHNSON, I., LIN, A. Y., LI, T. J.-J., HALL, A., HALFAKER, A., SCHÖNING, J., AND HECHT, B. Not at home on the range: Peer production and the urban/rural divide. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), CHI, pp. 13–25.
- [149] JOHNSON, I. L. *Identifying and Addressing Structural Inequalities in the Representativeness of Geographic Technologies*. PhD thesis, Northwestern University, 2019.
- [150] JOHNSON, I. L., LIN, Y., LI, T. J.-J., HALL, A., HALFAKER, A., SCHÖNING, J., AND HECHT, B. Not at home on the range: Peer production and the urban/rural divide. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), pp. 13–25.
- [151] JOHNSON, I. L., SENGUPTA, S., SCHÖNING, J., AND HECHT, B. The geography and importance of localness in geotagged social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), pp. 515–526.
- [152] JOHNSTON, P., AND HARRIS, R. The boeing 737 max saga: lessons for software organizations. *Software Quality Professional* 21, 3 (2019), 4–12.
- [153] KAMPS, J., AND KOOLEN, M. Is wikipedia link structure different? In *Proceedings of the second ACM international conference on Web search and data mining* (2009), pp. 232–241.
- [154] KANG, J., KÖRNER, M., WANG, Y., TAUBENBÖCK, H., AND ZHU, X. X. Building instance classification using street view images. *ISPRS journal of photogrammetry and remote sensing* 145 (2018), 44–59.

- [155] KAPPERT, U., CICHON, R., SCHNEIDER, J., GULIELMOS, V., AHMADZADE, T., NICOLAI, J., TUGTEKIN, S.-M., AND SCHUELER, S. Technique of closed chest coronary artery surgery on the beating heart. *European journal of cardio-thoracic surgery* 20, 4 (2001), 765–769.
- [156] KHOSLA, A., JAYADEVAPRAKASH, N., YAO, B., AND LI, F.-F. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)* (2011), vol. 2.
- [157] KHOSLA, A., ZHOU, T., MALISIEWICZ, T., EFROS, A. A., AND TORRALBA, A. Undoing the damage of dataset bias. In *European Conference on Computer Vision* (2012), Springer, pp. 158–171.
- [158] KIM, J., CHA, M., AND SANDHOLM, T. Socroutes: safe routes based on tweet sentiments. In *Proceedings of the 23rd International Conference on World Wide Web* (2014), ACM, pp. 179–182.
- [159] KITTUR, A., CHI, E. H., AND SUH, B. What’s in wikipedia? mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2009), pp. 1509–1512.
- [160] KNIGHT, J. C. Safety critical systems: challenges and directions. In *Proceedings of the 24th international conference on software engineering* (2002), pp. 547–550.
- [161] KOCH, C., AND BRILAKIS, I. Improving pothole recognition through vision tracking for automated pavement assessment. In *Proc. of EGICE Workshop* (2011), pp. 1–8.
- [162] KOGAN, M., ANDERSON, J., PALEN, L., ANDERSON, K. M., AND SODEN, R. Finding the way to osm mapping practices: Bounding large crisis datasets for qualitative investigation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), pp. 2783–2795.
- [163] KONSTAN, J. A., MILLER, B. N., MALTZ, D., HERLOCKER, J. L., GORDON, L. R., AND RIEDL, J. Grouplens: applying collaborative filtering to usenet news. *Communications of the ACM* 40, 3 (1997), 77–87.
- [164] KORNBLITH, S., SHLENS, J., AND LE, Q. V. Do better imagenet models transfer better? In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2019), pp. 2661–2671.

- [165] KRAEMER, J. D., AND BENTON, C. S. Disparities in road crash mortality among pedestrians using wheelchairs in the usa: results of a capture–recapture analysis. *BMJ open* 5, 11 (2015), e008396.
- [166] KRASIN, I., AND DUERIG, T. Google ai blog - introducing the open images dataset. <https://ai.googleblog.com/2016/09/introducing-open-images-dataset.html>.
- [167] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105.
- [168] KRUMM, J. A survey of computational location privacy. *Personal and Ubiquitous Computing* 13, 6 (2009), 391–399.
- [169] KUN, A. L., PAK, T., MEDENICA, Ž., MEMAROVIĆ, N., AND PALINKO, O. Glancing at personal navigation devices can affect driving: experimental results and design implications. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (2009), pp. 129–136.
- [170] KUZNETSOVA, A., ROM, H., ALLDRIN, N., UIJLINGS, J., KRASIN, I., PONT-TUSET, J., KAMALI, S., POPOV, S., MALLOCI, M., DUERIG, T., ET AL. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982* (2018).
- [171] LABAN, M. M., AND NABITY JR, T. S. Traffic collisions between electric mobility devices (wheelchairs) and motor vehicles: Accidents, hubris, or self-destructive behavior? *American journal of physical medicine & rehabilitation* 89, 7 (2010), 557–560.
- [172] LACOSTE, A., LUCCIONI, A., SCHMIDT, V., AND DANDRES, T. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700* (2019).
- [173] LACY, S., WATSON, B. R., RIFFE, D., AND LOVEJOY, J. Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly* 92, 4 (2015), 791–811.
- [174] LAGE, I., ROSS, A., GERSHMAN, S. J., KIM, B., AND DOSHI-VELEZ, F. Human-in-the-loop interpretability prior. *Advances in neural information processing systems* 31 (2018), 10159–10168.
- [175] LAMERTZ, K. The social construction of fairness: Social influence and sense making in organizations. *Journal of Organizational Behavior* 23, 1 (2002), 19–37.

- [176] LAMPTON, D. R., RODRIGUEZ, M. E., AND COTTON, J. E. Simulator sickness symptoms during team training in immersive virtual environments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2000), vol. 44, SAGE Publications Sage CA: Los Angeles, CA, pp. 530–533.
- [177] LANDIS, J. R., AND KOCH, G. G. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.
- [178] LAPRIE, J.-C. Dependability: A unifying concept for reliable computing and fault tolerance. *Dependability of resilient computers* (1989), 1–28.
- [179] LAVIE, T., ORON-GILAD, T., AND MEYER, J. Aesthetics and usability of in-vehicle navigation displays. *International Journal of Human-Computer Studies* 69, 1-2 (2011), 80–99.
- [180] LAW, P.-M., MALIK, S., DU, F., AND SINHA, M. The impact of presentation style on human-in-the-loop detection of algorithmic bias. *arXiv preprint arXiv:2004.12388* (2020).
- [181] LE BOT, P. Human reliability data, human error and accident models—illustration through the three mile island accident analysis. *Reliability Engineering & System Safety* 83, 2 (2004), 153–167.
- [182] LESHED, G., VELDEN, T., RIEGER, O., KOT, B., AND SENEGERS, P. In-car gps navigation: engagement with and disengagement from the environment. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2008), pp. 1675–1684.
- [183] LEVESON, N. G. *Safeware: system safety and computers*. Addison-Wesley, 1995.
- [184] LEWIS, S. Qualitative inquiry and research design: Choosing among five approaches. *Health promotion practice* 16, 4 (2015), 473–475.
- [185] LIEM, M., AND KOENRAADT, F. Homicide-suicide in the netherlands: A study of newspaper reports, 1992–2005. *The Journal of Forensic Psychiatry & Psychology* 18, 4 (2007), 482–493.
- [186] LIN, A. Y., FORD, J., ADAR, E., AND HECHT, B. Vizbywiki: mining data visualizations from the web to enrich news articles. In *Proceedings of the 2018 World Wide Web Conference* (2018), pp. 873–882.

- [187] LIN, A. Y., KUEHL, K., SCHÖNING, J., AND HECHT, B. Understanding death by gps: A systematic study of catastrophic incidents associated with personal navigation technologies. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), ACM, pp. 1154–1166.
- [188] LIN, J., AND LIU, Y. Potholes detection based on svm in the pavement distress image. In *2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science* (2010), IEEE, pp. 544–547.
- [189] LIN, Y., YU, B., HALL, A., AND HECHT, B. Problematizing and addressing the article-as-concept assumption in wikipedia. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (2017), pp. 2052–2067.
- [190] LIND, E. A., KRAY, L., AND THOMPSON, L. The social construction of injustice: Fairness judgments in response to own and others' unfair treatment by authorities. *Organizational behavior and human decision processes* 75, 1 (1998), 1–22.
- [191] LITMAN, T. Measuring transportation: traffic, mobility and accessibility. *Institute of Transportation Engineers. ITE Journal* 73, 10 (2003), 28.
- [192] MACHANAVAJJHALA, A., KIFER, D., GEHRKE, J., AND VENKITASUBRAMANIAM, M. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 3–es.
- [193] MADER, M., BRESGES, A., TOPAL, R., BUSSE, A., FORSTING, M., AND GIZEWSKI, E. R. Simulated car driving in fmri—cerebral activation patterns driving an unfamiliar and a familiar route. *Neuroscience letters* 464, 3 (2009), 222–227.
- [194] MAHDISOLTANI, F., BIEGA, J., AND SUCHANEK, F. M. Yago3: A knowledge base from multilingual wikipedias.
- [195] MALKIN, B. Britons stranded in australian outback for four days after satnav error. *Telegraph* (2010).
- [196] MALPHURS, J. E., AND COHEN, D. A newspaper surveillance study of homicide-suicide in the united states. *The American Journal of Forensic Medicine and Pathology* 23, 2 (2002), 142–148.
- [197] MANYIKA, J. Big data: The next frontier for innovation, competition, and productivity. http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.

- [198] MAO, H., HU, Y., KAR, B., GAO, S., AND MCKENZIE, G. Geoai 2017 workshop report: the 1st acm sigspatial international workshop on geoai:@ ai and deep learning for geographic knowledge discovery: Redondo beach, ca, usa-november 7, 2016. *SIGSPATIAL Special 9*, 3 (2017), 25.
- [199] MARIMON, D., SARASUA, C., CARRASCO, P., ÁLVAREZ, R., MONTESA, J., ADAMEK, T., ROMERO, I., ORTEGA, M., AND GASCÓ, P. Mobiar: tourist experiences through mobile augmented reality. *Telefonica Research and Development, Barcelona, Spain* (2010).
- [200] MARMANIS, D., DATCU, M., ESCH, T., AND STILLA, U. Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters 13*, 1 (2015), 105–109.
- [201] MARSHALL, A. Crime alerts come to brazilian waze, just in time for the olympics — wired, 07 2016. <https://www.wired.com/2016/07/crime-alerts-come-brazilian-waze-just-time-olympics/>.
- [202] MASHHADI, A., QUATTRONE, G., AND CAPRA, L. Putting ubiquitous crowd-sourcing into context. In *Proceedings of the 2013 conference on Computer supported cooperative work* (2013).
- [203] MASSA, P., AND SCRINZI, F. Exploring linguistic points of view of wikipedia. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (2011), pp. 213–214.
- [204] MASSA, P., AND SCRINZI, F. Manypedia: Comparing language points of view of wikipedia communities. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (2012), pp. 1–9.
- [205] MAZUMDER, R., HASTIE, T., AND TIBSHIRANI, R. Spectral regularization algorithms for learning large incomplete matrices. 2287–2322.
- [206] MCHUGH, M. L. The chi-square test of independence. *Biochemia medica: Biochemia medica 23*, 2 (2013), 143–149.
- [207] MCCLAUGHLIN, N., DEL RINCON, J. M., AND MILLER, P. Data-augmentation for reducing dataset bias in person re-identification. In *2015 12th IEEE International conference on advanced video and signal based surveillance (AVSS)* (2015), IEEE, pp. 1–6.
- [208] MEDENICA, Z., KUN, A. L., PAEK, T., AND PALINKO, O. Augmented reality vs. street views: a driving simulator study comparing two emerging navigation aids. In

Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (2011), pp. 265–274.

- [209] MENKING, A., RANGARAJAN, V., AND GILBERT, M. "sharing small pieces of the world" increasing and broadening participation in wikimedia commons. In *Proceedings of the 14th International Symposium on Open Collaboration* (2018), pp. 1–12.
- [210] MICROSOFT. Translator text api. microsoft translator for business. <https://commons.wikimedia.org/wiki/Special:MediaStatistics>.
- [211] MILLER, G. A. Wordnet: a lexical database for english. *Communications of the ACM* 38, 11 (1995), 39–41.
- [212] MILNE, D., AND WITTEN, I. H. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management* (2008), pp. 509–518.
- [213] MILNER, G. *Pinpoint: how GPS is changing technology, culture, and our minds*. WW Norton & Company, 2016.
- [214] MOVSHOVITZ-ATTIAS, Y., YU, Q., STUMPE, M. C., SHET, V., ARNOUD, S., AND YATZIV, L. Ontological supervision for fine grained classification of street view storefronts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 1693–1702.
- [215] NARAYANAN, A. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA* (2018), vol. 1170.
- [216] NARAYANAN, A. Tutorial: 21 fairness definitions and their politics.(2018). URL <https://www.youtube.com/watch> (2018).
- [217] NICKEL, M., MURPHY, K., TRESP, V., AND GABRILOVICH, E. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE* 104, 1 (2015), 11–33.
- [218] NICKERSON, A., JAPKOWICZ, N., AND MILIOS, E. Using unsupervised learning to guide re-sampling in imbalanced data sets. In *In Proceedings of the Eighth International Workshop on AI and Statistics* (2001), p. 261–265.
- [219] NIELSEN, J. Be succinct!(writing for the web). *Jacob Nielsen's Alertbox* (1997).
- [220] NIELSEN, J. Error message guidelines. *Nielsen Norman Group 24* (2001).

- [221] NIELSEN, J. Mini-ia: Structuring the information about a concept. *Retrieved July 20* (2011), 2016.
- [222] OF THE PRESIDENT, E. O., MUNOZ, C., DIRECTOR, D. P. C., OF SCIENCE, M. U. C. T. O. S. O., POLICY)), T., FOR DATA POLICY, D. D. C. T. O., OF SCIENCE, C. D. S. P. O., AND POLICY)), T. *Big data: A report on algorithmic systems, opportunity, and civil rights*. Executive Office of the President, 2016.
- [223] OF THE UNITED STATES, C. Letters to world bank.
- [224] OLSSON, J. Improved road accessibility and indirect development effects: evidence from rural philippines. *Journal of Transport Geography* 17, 6 (2009), 476–483.
- [225] O’SULLIVAN, D., MORRISON, A., AND SHEARER, J. Using desktop gis for the investigation of accessibility by public transport: an isochrone approach. *International Journal of Geographical Information Science* 14, 1 (2000), 85–104.
- [226] OTTERBACHER, J., CHECCO, A., DEMARTINI, G., AND CLOUGH, P. Investigating user perception of gender bias in image search: the role of sexism. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (2018), pp. 933–936.
- [227] OVERELL, S., SIGURBJÖRNSSON, B., AND VAN ZWOL, R. Classifying tags using open content resources. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (2009), pp. 64–73.
- [228] PALANQUE, P., COCKBURN, A., DÉSSERT-LEGENDRE, L., GUTWIN, C., AND DELERIS, Y. Brace touch: A dependable, turbulence-tolerant, multi-touch interaction technique for interactive cockpits. In *International Conference on Computer Safety, Reliability, and Security* (2019), Springer, pp. 53–68.
- [229] PALETTA, L., FRITZ, G., SEIFERT, C., LULEY, P. M., AND ALMER, A. A mobile vision service for multimedia tourist applications in urban environments. In *2006 IEEE Intelligent Transportation Systems Conference* (2006), IEEE, pp. 566–572.
- [230] PANG, B., LEE, L., AND VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (2002), Association for Computational Linguistics, pp. 79–86.
- [231] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG,

- V., ET AL. Scikit-learn: Machine learning in python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.
- [232] PEISSNER, M., DOEBLER, V., AND METZE, F. Can voice interaction help reducing the level of distraction and prevent accidents. *Meta-Study Driver Distraction Voice Interaction* (2011), 24.
- [233] RAHWAN, I. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.
- [234] RAIFER, M. Overpass turbo, 2020. <https://overpass-turbo.eu>.
- [235] RAINEY, D. Y., AND RUNYAN, C. W. Newspapers: a source for injury surveillance? *American journal of public health* 82, 5 (1992), 745–746.
- [236] REAGLE, J., AND RHUE, L. Gender bias in wikipedia and britannica. *International Journal of Communication* 5 (2011), 21.
- [237] REDI, M. Visual enrichment of collaborative knowledge bases. https://commons.wikimedia.org/w/index.php?title=File:Visual_enrichment_of_collaborative_KB.pdf&page=1.
- [238] REGOLI, R. M., AND HEWITT, J. D. *Exploring criminal justice*. Jones & Bartlett Learning, 2008.
- [239] REICHSTEIN, M., CAMPS-VALLS, G., STEVENS, B., JUNG, M., DENZLER, J., CARVALHAIS, N., ET AL. Deep learning and process understanding for data-driven earth system science. *Nature* 566, 7743 (2019), 195–204.
- [240] REPORTER, B. Man follows sat nav to cliff edge. *BBC News* (2009). http://news.bbc.co.uk/2/hi/uk_news/england/bradford/7962212.stm.
- [241] ROMA, P., SPACCA, A., POMPILI, M., LESTER, D., TATARELLI, R., GIRARDI, P., AND FERRACUTI, S. The epidemiology of homicide–suicide in italy: A newspaper study from 1985 to 2008. *Forensic science international* 214, 1-3 (2012), e1–e5.
- [242] RUKSHAN BATUWITA, V. P. Efficient resampling methods for training support vector machines with imbalanced datasets. In *The 2010 International Joint Conference on Neural Networks* (2010).
- [243] RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A., BERNSTEIN, M., ET AL. Imagenet

- large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.
- [244] SAITO, T., AND REHMSMEIER, M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one* 10, 3 (2015).
- [245] SALBERG, A.-B. Detection of seals in remote sensing images using features extracted from deep convolutional neural networks. In *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (2015), IEEE, pp. 1893–1896.
- [246] SAMSONOV, P. A., TANG, X., SCHÖNING, J., KUHN, W., AND HECHT, B. You can't smoke here: Towards support for space usage rules in location-aware technologies. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (2015), pp. 971–974.
- [247] SANTOS, R., MURRIETA-FLORES, P., CALADO, P., AND MARTINS, B. Toponym matching through deep neural networks. *International Journal of Geographical Information Science* 32, 2 (2018), 324–348.
- [248] SCHEUERMAN, M. K., PAUL, J. M., AND BRUBAKER, J. R. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.
- [249] SCHREIBER, J. Bridging the gap between useful and aesthetic maps in car navigation systems. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services* (2009), pp. 1–4.
- [250] SEED, D. Geo-diversity for better, fairer machine learning, Aug. 2020.
- [251] SEGLER, M. H., KOGEJ, T., TYRCHAN, C., AND WALLER, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science* 4, 1 (2018), 120–131.
- [252] SEN, S., GIESEL, M. E., GOLD, R., HILLMANN, B., LESICKO, M., NADEN, S., RUSSELL, J., WANG, Z., AND HECHT, B. “Turkers, scholars,” “arafat” and “peace” cultural communities and algorithmic gold standards. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2015), pp. 826–838.
- [253] SEN, S., LI, T. J.-J., TEAM, W., AND HECHT, B. Wikibrain: democratizing computation on wikipedia. In *Proceedings of The International Symposium on Open Collaboration* (2014), pp. 1–10.

- [254] SERINO, M., CORDREY, K., MCLAUGHLIN, L., AND MILANAİK, R. L. Pokémon go and augmented virtual reality games: a cautionary commentary for parents and pediatricians. *Current opinion in pediatrics* 28, 5 (2016), 673–677.
- [255] SERMANET, P., EIGEN, D., ZHANG, X., MATHIEU, M., FERGUS, R., AND LECUN, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229* (2013).
- [256] SERVICE, O., 2018. <https://openrouteservice.org/>.
- [257] SHAH, S., BAO, F., LU, C.-T., AND CHEN, I.-R. Crowdsafe: crowd sourcing of crime incidents and safe routing on mobile devices. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2011), ACM, pp. 521–524.
- [258] SHANKAR, S., HALPERN, Y., BRECK, E., ATWOOD, J., WILSON, J., AND SCULLEY, D. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536* (2017).
- [259] SHARIF RAZAVIAN, A., AZIZPOUR, H., SULLIVAN, J., AND CARLSSON, S. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2014), pp. 806–813.
- [260] SHEKHAR, S., FEINER, S., AND AREF, W. G. From gps and virtual globes to spatial computing-2020. *GeoInformatica* 19, 4 (2015), 799–832.
- [261] SHEKHAR, S., FEINER, S., AND AREF, W. G. From gps and virtual globes to spatial computing-2020. In *GeoInformatica* (9 2015), vol. 19, ACM, pp. 799–832.
- [262] SHEN, Y., HERMANS, E., BRIJS, T., WETS, G., AND VANHOOF, K. Road safety risk evaluation and target setting using data envelopment analysis and its extensions. *Accident Analysis & Prevention* 48 (2012), 430–441.
- [263] SHWEIKEH, F., AMADIO, J. P., ARNELL, M., BARNARD, Z. R., KIM, T. T., JOHNSON, J. P., AND DRAZIN, D. Robotics and the spine: a review of current and ongoing applications. *Neurosurgical focus* 36, 3 (2014), E10.
- [264] SIGURBJÖRNSSON, B., AND VAN ZWOL, R. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web* (2008), pp. 327–336.

- [265] SIMONYAN, K., AND ZISSERMAN, A. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems* (2014), pp. 568–576.
- [266] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [267] SINGHAL, A. Introducing the knowledge graph: things, not strings. *Official google blog 16* (2012).
- [268] SOLL, M., NAUMANN, P., SCHÖNING, J., SAMSONOV, P., AND HECHT, B. Helping computers understand geographically-bound activity restrictions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), pp. 2442–2446.
- [269] SONG, W., WORKMAN, S., HADZIC, A., ZHANG, X., GREEN, E., CHEN, M., SOULEYRETTE, R., AND JACOBS, N. Farsa: Fully automated roadway safety assessment. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2018), IEEE, pp. 521–529.
- [270] SRIVASTAVA, S., VARGAS MUÑOZ, J. E., LOBRY, S., AND TUIA, D. Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data. *International Journal of Geographical Information Science* (2018), 1–20.
- [271] STAFF, M.-W. *Lexico.com*. Oxford University Press (OUP), 2019.
- [272] STRATMANN, T. C., BRAUER, D., AND BOLL, S. Supporting the perception of spatially distributed information on ship bridges. In *Proceedings of Mensch und Computer 2019*. 2019, pp. 475–479.
- [273] STRATMANN, T. C., GRUENEFELD, U., HAHN, A., BOLL, S., STRATMANN, J., AND SCHWEIGERT, S. Mobile bridge-a portable design simulator for ship bridge interfaces. *TransNav: International Journal on Marine Navigation and Safety of Sea Transportation 12* (2018).
- [274] STROBL, C., BOULESTEIX, A.-L., KNEIB, T., AUGUSTIN, T., AND ZEILEIS, A. Conditional variable importance for random forests. *BMC bioinformatics 9*, 1 (2008), 307.
- [275] STROBL, C., BOULESTEIX, A.-L., ZEILEIS, A., AND HOTHORN, T. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics 8*, 1 (2007), 25.

- [276] STRYKER, J. E., WRAY, R. J., HORNIK, R. C., AND YANOVITZKY, I. Validation of database search terms for content analysis: The case of cancer news coverage. *Journalism & Mass Communication Quarterly* 83, 2 (2006), 413–430.
- [277] SUCHANEK, F. M., KASNECI, G., AND WEIKUM, G. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web* (2007), pp. 697–706.
- [278] SUI, D., GOODCHILD, M., AND ELWOOD, S. Volunteered geographic information, the exaflood, and the growing digital divide. In *Crowdsourcing geographic knowledge*. Springer, 2013, pp. 1–12.
- [279] SUN, C., SHRIVASTAVA, A., SINGH, S., AND GUPTA, A. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 843–852.
- [280] SUN, T., DI, Z., AND WANG, Y. Combining satellite imagery and gps data for road extraction. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery* (2018), pp. 29–32.
- [281] SZEGEDY, C., VANHOUCHE, V., IOFFE, S., SHLENS, J., AND WOJNA, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2818–2826.
- [282] TARQUI, G., CASTRO, L. A., AND FAVELA, J. Reducing drivers’ distractions in phone-based navigation assistants using landmarks. In *Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction*. Springer, 2013, pp. 342–349.
- [283] TELECOMMUNICATIONS, N., AND INFORMATION ADMINISTRATION (DOC), WASHINGTON, D. *Falling through the net: A Survey of the “Have nots” in rural and urban America*. ERIC Clearinghouse, 1995.
- [284] THEBAULT-SPIEKER, J., TERVEEN, L., AND HECHT, B. Toward a geographic understanding of the sharing economy: Systemic biases in uberx and taskrabbit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 3 (2017), 1–40.
- [285] TORRALBA, A., AND EFROS, A. A. Unbiased look at dataset bias. In *CVPR 2011* (2011), IEEE, pp. 1521–1528.
- [286] VAIDYA, G., KONTOKOSTAS, D., KNUTH, M., LEHMANN, J., AND HELLMANN, S. Dbpedia commons: structured multimedia metadata from the wikimedia commons. In *International Semantic Web Conference* (2015), Springer, pp. 281–289.

- [287] VAN DIJK, J. A. Digital divide research, achievements and shortcomings. *Poetics* 34, 4-5 (2006), 221–235.
- [288] VRANDEČIĆ, D., AND KRÖTZSCH, M. Wikidata: a free collaborative knowledge-base. *Communications of the ACM* 57, 10 (2014), 78–85.
- [289] WATERFIELD, B. Gps failure leaves belgian woman in zagreb two days later. *Retrieved September 12* (2013), 2016.
- [290] WEBB, C. N. Geospatial summary of crash fatalities. Tech. rep., 2020.
- [291] WEBB, J., AND LIFF, S. Play the white man: the social construction of fairness and competition in equal opportunity policies. *The Sociological Review* 36, 3 (1988), 532–551.
- [292] WELD, D. S., WU, F., ADAR, E., AMERSHI, S., FOGARTY, J., HOFFMANN, R., PATEL, K., AND SKINNER, M. Intelligence in wikipedia. In *AAAI* (2008), vol. 8, pp. 1609–1614.
- [293] WENIG, D., STEENBERGEN, A., SCHÖNING, J., HECHT, B., AND MALAKA, R. Scrollinghome: bringing image-based indoor navigation to smartwatches. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services* (2016), pp. 400–406.
- [294] WIERWILLE, W. W., ANTIN, J. F., DINGUS, T. A., AND HULSE, M. C. Visual attentional demand of an in-car navigation display system. In *Vision in Vehicles II. Second International Conference on Vision in Vehicles Applied Vision Association-Ergonomics Society Association of Optometrists* (1988).
- [295] WILKINSON, D. M., AND HUBERMAN, B. A. Cooperation and quality in wikipedia. In *Proceedings of the 2007 international symposium on Wikis* (2007), pp. 157–164.
- [296] WITTEN, I. H., AND FRANK, E. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record* 31, 1 (2002), 76–77.
- [297] WITTEN, I. H., AND MILNE, D. N. An effective, low-cost measure of semantic relatedness obtained from wikipedia links.
- [298] WOLF, M., AND SERPANOS, D. Safety and security in cyber-physical systems and internet-of-things systems. *Proceedings of the IEEE* 106, 1 (2017), 9–20.

- [299] WULCZYN, E., WEST, R., ZIA, L., AND LESKOVEC, J. Growing wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web* (2016), pp. 975–985.
- [300] XIE, M., JEAN, N., BURKE, M., LOBELL, D., AND ERMON, S. Transfer learning from deep features for remote sensing and poverty mapping. In *Thirtieth AAAI Conference on Artificial Intelligence* (2016).
- [301] XU, Z., FU, Y., MAO, J., AND SU, D. Towards the semantic web: Collaborative tag suggestions. In *Collaborative web tagging workshop at WWW2006, Edinburgh, Scotland* (2006).
- [302] YANG, K., QINAMI, K., FEI-FEI, L., DENG, J., AND RUSSAKOVSKY, O. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), pp. 547–558.
- [303] ZHANG, F., DU, B., AND ZHANG, L. Scene classification via a gradient boosting random convolutional network framework. *IEEE Transactions on Geoscience and Remote Sensing* 54, 3 (2015), 1793–1802.
- [304] ZHANG, L., YANG, F., ZHANG, Y. D., AND ZHU, Y. J. Road crack detection using deep convolutional neural network. In *2016 IEEE international conference on image processing (ICIP)* (2016), IEEE, pp. 3708–3712.
- [305] ZHAO, J., WANG, T., YATSKAR, M., ORDONEZ, V., AND CHANG, K.-W. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).
- [306] ZHAO, J., ZHOU, Y., LI, Z., WANG, W., AND CHANG, K.-W. Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496* (2018).
- [307] ZIELSTRA, D., AND HOCHMAIR, H. H. Using free and proprietary data to compare shortest-path lengths for effective pedestrian routing in street networks. *Transportation Research Record* 2299, 1 (2012), 41–47.
- [308] ZIELSTRA, D., AND ZIPF, A. A comparative study of proprietary geodata and volunteered geographic information for germany. In *13th AGILE International Conference on Geographic Information Science* (2010).

APPENDIX A

Detailed Classification Accuracies in Figure 5.4

	zh			en			es		
	Rtg:2	Rtg:2.5	Rtg:3	Rtg:2	Rtg:2.5	Rtg:3	Rtg:2	Rtg:2.5	Rtg:3
Baseline	63.54%	59.37%	50%	56.03%	50.24%	58.93%	50.98%	70.58%	81.37%
Linear SVM	85.44%	90.81%	84.66%	88.92%	91.38%	83.09%	64.81%	67.45%	81.18%
Random Forest	85.33%	86.44%	86.77%	92.69%	90.80%	82.95%	70.00%	74.18%	81.18%
Naïve Bayes	85.44%	87.66%	83.66%	89.35%	89.88%	82.57%	65.90%	73.36%	70.45%
Logistic	85.33%	88.07%	83.44%	90.33%	89.90%	82.07%	67.72%	71.27%	78.27%
KNN	72.88%	74.00%	71.88%	80.54%	77.21%	78.16%	69.54%	79.18%	83.18%
Adaboost	83.33%	85.44%	82.66%	88.88%	90.83%	82.97%	59.18%	70.72%	75.18%
Decision Tree	79.00%	81.00%	75.88%	88.28%	86.42%	78.07%	67.72%	62.81%	72.27%

Table A.1. Detailed classification accuracies for *High-Interest* dataset (Figure 5.4 (a))

	zh			en			es		
	Rtg:2	Rtg:2.5	Rtg:3	Rtg:2	Rtg:2.5	Rtg:3	Rtg:2	Rtg:2.5	Rtg:3
Baseline	69.60%	76.47%	85.29%	89.10%	91.58%	98.01%	63.00%	86.00%	100%
Linear SVM	83.18%	82.27%	91.18%	91.52%	92.02%	98.02%	66.99%	85.00%	N/A
Random Forests	79.27%	86.18%	86.18%	91.04%	93.02%	97.52%	63.00%	90.00%	N/A
Naïve Bayes	73.45%	69.36%	85.18%	79.61%	75.26%	59.04%	67.99%	77.00%	N/A
Logistic	84.18%	83.27%	92.095	91.52%	92.02%	98.02%	66.99%	88.00%	N/A
KNN	81.45%	83.09%	86.09%	89.52%	92.02%	98.02%	52.00%	85.00%	N/A
Adaboost	81.18%	81.18%	90.09%	87.59%	90.02%	98.02%	62.99%	83.00%	N/A
Decision Tree	82.27%	76.36%	92.18%	84.09%	87.57%	97.04%	64.99%	86.00%	N/A

Table A.2. Detailed classification accuracies for *Random* dataset (Figure 5.4 (b))

	zh			en			es		
	Rtg:2	Rtg:2.5	Rtg:3	Rtg:2	Rtg:2.5	Rtg:3	Rtg:2	Rtg:2.5	Rtg:3
Baseline	67.96%	58.25%	55.33%	56.71%	55.72%	74.12%	51.15%	59.59%	75.75%
Linear SVM	85.63%	83.63%	80.45%	73.66%	73.66%	77.09%	69.44%	80.55%	71.88%
Random Forests	86.63%	82.54%	81.36%	70.57%	72.14%	80.61%	68.66%	73.66%	76.88%
Naïve Bayes	79.72%	78.72%	81.36%	73.14%	69.69%	37.30%	67.44%	71.66%	67.77%
Logistic	86.45%	85.54%	81.45%	71.19%	73.66%	77.57%	66.44%	73.55%	75.00%
KNN	78.81%	82.63%	71.72%	71.28%	66.28%	69.61%	59.55%	66.55%	66.77%
Adaboost	82.54%	73.90%	75.63%	70.07%	73.11%	78.11%	64.55%	70.44%	71.66%
Decision Tree	84.72%	81.36%	77.63%	64.66%	65.51%	71.09%	60.55%	67.55%	64.77%

Table A.3. Detailed classification accuracies for *Ad-Hoc* dataset (Figure 5.4 (c))

	<i>High-Interest</i>			<i>Ad-Hoc</i>			<i>Random</i>		
	Rtg:2	Rtg:2.5	Rtg:3	Rtg:2	Rtg:2.5	Rtg:3	Rtg:2	Rtg:2.5	Rtg:3
Baseline	56.04%	52.83%	62.46%	58.31%	53.10%	69.72%	77.77%	86.38%	95.27%
Linear SVM	86.09%	88.59%	81.44%	74.00%	77.70%	79.12%	81.10%	86.57%	94.78%
Random Forests	83.86%	85.87%	78.73%	73.46%	74.70%	74.62%	78.16%	88.34%	94.28%
Naïve Bayes	84.12%	85.62%	81.20%	73.50%	74.26%	74.21%	74.70%	80.18%	86.39%
Logistic	84.87%	87.60%	80.70%	75.46%	76.45%	79.37%	81.10%	88.32%	94.28%
KNN	77.47%	77.75%	74.82%	72.03%	72.28%	72.92%	80.60%	87.83%	94.52%
Adaboost	84.17%	84.66%	80.96%	74.99%	73.74%	75.63%	80.39%	86.59%	94.40%
Decision Tree	85.13%	80.17%	74.57%	67.03%	67.01%	72.42%	74.45%	79.90%	92.34%

Table A.4. Detailed overall classification accuracies (Figure 5.4 (d))

APPENDIX B

The Mathematical Relationship Between Likelihood and the Loss Function

In the context of deep learning model, especially the ImageNet pre-training models, the last layer is usually a softmax layer which has the form:

$$\sigma(f_k) = \frac{e^{f_k}}{\sum_{j=1}^K e^{f_j}}$$

where f is a K-length vector of class scores output from previous layers, f_k is the k-th value in the f vector, which represent the class score for k-th class. The softmax function $\sigma(f_k)$ produces a probability (between 0 and 1) that represents the *likelihood* that a particular sample belong to class k.

For a single sample i , cross-entropy loss has the form:

$$L_i = H_i(p, q) = - \sum_k p_i(k) \log q_i(k)$$

where $q_i(k)$ is the estimated likelihood of the sample i for k-th class, which is $\sigma_i(f_k)$. p_i denotes the true probability of the class membership. Because each sample has one true class membership, $p = [0, \dots, 1, \dots, 0]$ contains a single 1 at y_i position. As such,

$$L_i = H_i(p, q) = -\sigma_i(f_{y_i}) = -\text{likelihood}_i$$

Summing over the cross-entropy loss for all samples, we get:

$$L = H(p, q) = - \sum_i (\textit{likelihood}_i)$$