BIGGER IS BETTER. OR IS IT?

Lessons learned from applying a deep neural network to Twitter posts in order to estimate potentials of using big data to monitor the SDGs.

Master's Thesis

University of Bern, Faculty of Science

Institute of Geography

Submitted by

Benjamin Andrea Schüpbach

Matr.No. 14-100-564

Bern, January 2020

Supervisor

PD Dr. Andreas Heinimann

Mein lieber Sohn, fliege nicht zu tief, damit die Federn nicht ins Meerwasser tauchen, sonst werden sie feucht und ziehen dich in die Tiefe. Fliege aber auch nicht zu hoch, sonst schmilzt die Sonne das Wachs, die Flügel fallen auseinander, und du stürtzt ab. Fliege die Mittelstrasse zwischen Meer und Sonne immer nur hinter mir her!

- *Daidalos und Ikaros*, Gustav Schwab (1990)

Abstract

With the emergence of the Internet of Things (IoT) and the extensive amount of data produced by it, science's desire to investigate this vast amount of untapped data is growing, resulting in the paradigm of big data: data sets of exceedingly large volumes, growing at exceptional rates, consisting of enormous amounts of structured and unstructured data. At the same time, artificial intelligence needed to analyze data sets of these proportions continue to improve as well.

The potentials attributed to big data analyses are extensive, particularly in the context of efficiently generating reliable, up-to-date data to measure progress towards the Agenda 2030's Sustainable Development Goals (SDGs). However, many scientific contributions in this domain focusing on unexploited capacities rely on further technological progress and therefore project future potentials. Yet, the SDGs were designed to tackle *current* challenges.

Nonetheless, for some of the indicators of sustainability introduced with the SDGs, it is still unclear how reliable data can be generated. Therefore, this study examines current technological capabilities and their potential contribution to overcoming a lack of data. It does so with an example of a big data analysis: applying a deep neural network to geolocated media content posted to Twitter, in order to illustrate the challenges left to overcome if big data is to be used to generate useful information for measuring progress towards the SDGs.

The findings of this study show that current technological capabilities already facilitate real-time analyses of big data from social media on a global scale. Yet, biases resulting from uncertainties regarding the accuracy of geolocated social media posts, along with low internet penetration rates and a consequent lack of data - coupled with an unavailability of data from prime sources like Facebook and Instagram - render such analyses incomplete, thus diminishing the significance of information gained this way.

Better access to more data from diverse sources is needed to improve on our current capacities to generate significant and reliable data to monitor efforts of improving sustainability. However, especially analyses of data from social media are embedded in a debate over privacy and data protection. This debate is here to stay. Nevertheless, some of the reservations against artificial intelligence and big data analyses can be alleviated by a high degree of transparency (i.e. by making big data projects open source).

Contents

List of Figures

List of Tables

1 Introduction

With the emergence of the Internet of Things (IoT), a desire for real-time analytics of continuously growing datasets arose (Mohammadi et al., 2018). Through continuous advancements in the domain of artificial intelligence (AI), science's capability of analyzing such vast amounts of data continues to grow as well (Redmon and Farhadi, 2019; Redmon, 2019; Redmon et al., 2016). At the same time, concerns about the potential abuse of such techniques on datasets containing a vast amount of information exist (Barnes, 2013). Particularly in relation to personal information and big data from social media (Grout, 2019).

The debate about potential gains and drawbacks of combining big data and machine learning approaches happens in a time of great demand for fast and efficient ways to generate various kinds of data. Especially to support international efforts of sustainable development (SDSN, 2018). In 2015, as part of the *Agenda 2030*, the UN's Sustainable Development Goals (SDGs) were adopted (United Nations, 2018). With the intent to produce trackable, incremental success for these goals, indicators of progress were introduced (UN Statistics Division, 2019a). Yet some of these indicators were not yet fully functional at the time of adoption, because it was unclear how data for such indicators could be generated (IAEG-SDGs, 2017).

As a possible solution to combat data shortages, big data is often attributed with great potentials to alleviate issues of data availability (Andries et al., 2019; Albert et al., 2019; Anejionu et al., 2019). However, some of these attributions assume further technological progress and therefore project future potentials (Chen and Zhang, 2014; Benjelloun et al., 2015).

Against the backdrop of global challenges like climate change, one soon realizes the urgent need for reliable data to monitor efforts of sustainable development. Hence, this study examines *current* technological capabilities and their potential contribution to overcoming a lack of data. It explores potentials of using big data and machine learning techniques for an efficient, near real-time, monitoring of the UN's Sustainable Development Goals (United Nations, 2018). It does so by diving into a specific example; training and subsequently applying an image classification algorithm which detects all-season roads (information needed for the *rural access index*, indicator 9.1.1 of the SDGs) in digtial images, to media content posted to Twitter (Roberts et al., 2006; Transport & ICT, 2016). The findings of this analysis are then discussed in a wider context, in order to contribute to the debate about potential benefits and drawbacks of using big data to monitor the SDGs.

This study's design offers a broad understanding of the context in which this debate takes place, subsequently dives deeper and illustrates different perspectives within said debate using a specific example, before discussing insights gained thusly in a broader context again.

After establishing the theoretical context (big data, sustainable development and image classification), as well as the goals of this study in section 1, methods (image classification) used in the analysis are explained in section 2. Section 3 presents results from this analysis. The following section 4 offers an interpretation as well as a discussion of these findings. Section 5 then summarizes the conclusions of this study and provides an outlook for future potential future research.

1.1 Big Data

Wu Xindong et al. (2014:97) define big data as "large-volume, heterogeneous, autonomous sources with distributed and decentralized control", including "complex and evolving relationships among data". Susanne Schnorr-Bäcker (2016:172) explains that factual and temporal variability further characterize big data. As explained by Patrick Mikalef et al. (2018) another common way of defining big data follows the *muliptle V's* consisting of the *volume*, *velocity* and *variety* (sometimes *veracity* and *value* are also included) of data. They refer to the sheer size of data sets, the speed at which data can be collected or becomes obsolete and the plurality of information coming from big data sources (including text, images and video), respectively. Mikalef et al. further remark that the growing rate, at which big data sets are increasing in size, is often represented in definitions as well.

1.1.1 Benefits of using Big Data

Schnorr-Bäcker (2016) sees strong scientific potential in big data, especially in connecting different data sources. She identifies the monitoring of political goals (e.g. Sustainable Development Goals) through statistical indicators as one of the most meaningful domains in which big data can have a significant impact. Furthermore, she acknowledges the importance of up-to-date data on such endeavours, that big data sources like social media provide.

In accordance with Schnorr-Bäcker, Ahmed Oussous et al. (2018) suggest that the analysis of big data sources can yield knowledge about complex and dynamic systems, supporting decision making processes through automatic detection of anomalies and trends. Phillip Chen and Chun-Yang Zhang (2014:317) underline this suggestion further by claiming that "taking advantage of valuable knowledge beyond big data will become the basic competition for today's enterprises". They further argue that for researchers and policy-/ decision makers, a new wave of growth awaits to be uncovered through harnessing the potentials of big data. Chen and Zhang (2014:337) even see big data as "the next frontier for innovation".

Fatima-Zahra Benjelloun et al. (2015) conclude in their paper that an additional benefit of big data comes from the potential of integrated analyses which arise through inter-disciplinary research. They mention opportunities in domains like healthcare, trading, agriculture, politics and tourism. In their study, one dominant recurring theme is the focus on patterns of customer behaviour, as well as behaviour based recommendations (e.g. recommendations of restaurants, hotels).

As shown above and in the following sections, the potentials - good or bad - of big data are manifold. Most likely, though, we are not yet able to clearly estimate the impact big data will have on human life. Because of the overwhelmingly large quantities of data which are produced every day, big data is strongly tied to artificial intelligence, which enables analyses at these scales. Rafael Reif (2019:9), President of the Massachusetts Institute of Technology (MIT), recently wrote: "Artificial Intelligence is an enabling technology.[...] It will help humanity learn more, waste less, work smarter, live longer – and better understand, predict and make decisions about almost anything that can be measured."

1.1.2 Criticism of Big Data

However, there are voices raising concerns over big data's potential to do more harm than good. Trevor Barnes (2013), for instance, sees parallels between potential future trajectories for the paradigm of big data with historical criticism against the quantitative revolution in Geography.

First, he sees a disconnection of data from what is important. Or, in other words, disconnecting data from *knowledge*. Barnes (2013:299) argues that "techniques and numbers become fetishized, put on a pedestal, prized for what they are rather than for what they do". According to him, this can be viewed in parallel to the critical notion (against the quantitative revolution in Geography) that, because outputs from certain techniques come in mathematical form, they will be touted as knowledge, since mathematics is often viewed as the language of science.

Second, he criticises the notion that, in order for big data to gain legitimacy, information needs to be converted into numbers. With this, he argues, important context is lost. This stands in accordance with, and elaborates further on, his first point. During the 1970s critics of the quantitative revolution argued that through the same process in geographic studies, contextual information that could not be expressed in numbers was left out. Barnes explains how through omission of context, distorted, misleading or even tragic outcomes can be the result.

Third, he asserts that in a big data perspective, "numbers are the story, shorn of the need of any interpretation" (Barnes, 2013:300). He further argues that to the question of why something happens, correlation coefficients alone do not provide an answer. Rather, causal explanatory frameworks are needed to find the answers to more important questions. This connects to Barnes' fourth and final point of making the link between knowledge and the desire to change the world with it. Without it, he argues, big data possesses a built-in conservatism.

Jeremy Crampton et al. (2013:132) convey it bluntly: "big data will not replace thinking". They add that studies, especially when drawing upon big data from social media, are often naive in the way insights are extrapolated to make blanket statements about society as a whole. Additionally, they are concerned that, when dealing with geotagged information, some crucial reservations are ignored. Namely, that a piece of geotagged information may not necessarily have been produced at, or represent the specified location it references. Jose Ramon Albert et al. (2019) highlight this by stressing the importance of a good understanding of the limitations inherent in big data.

Another, heavily debated topic in the domain of big data (and machine learning) is privacy. This debate will not be covered in-depth, for it is not central to this thesis. However, it is at the very core of a broader discussion about big data and as such should be briefly mentioned. Vic Grout (2014) takes privacy concerns in the age of IoT, big data and machine learning to the extreme in suggesting a "global Shazam for People". He foreshadows the use of the IoT in gathering vast amounts of incomplete, noisy sample data about people, effectively generating a personal fingerprint which can be compared to an online database, returning a wealth of information about the person (ranging from innocent facts like their surname to compromising information with the possibility of damaging the person's reputation) - much in the same way the popular smartphone application Shazam uses incomplete, noisy sample data about rhythm, frequencies, etc., to find the song titles ambient music as accurately as possible. Grout (2019) forecasts a significant reduction in personal privacy, considering people who would use such an application might be able to learn more about another person than they themselves already know.

David Edelman (2019:13) suggests that societies' role for making sure this never happens lies in the alignment of benefits of artificial intelligence with the obligation of public trust. His remarks on how to deal with privacy concerns in the age of machine learning are of particular importance in the domain of big data.

In his opinion, processes to find clear, fair and uniform design principles should be pursued simultaneously to technological advancements, for no side in this debate can afford to wait until the other presents a perfect solution.

1.1.3 Big Data for Sustainability

As outlined in the Report of the Global Working Group on Big Data for Official Statistics (2014), many potentials for big data in the context of the Sustainable Development Goals (see section 1.3) were still largely unexplored five years ago. In the time since, sustainability-specific potentials were incrementally being investigated. Exemplary contributions focused on institutional preparedness to facilitate even starting to include big data into measuring progress on the Sustainable Development Goals, integration of conventional data sources and big data, remote sensing, social and economic urban analytics and smart cities (Kharrazi et al., 2016; Albert et al., 2019; Andries et al., 2019; Anejionu et al., 2019; Allam and Dhunny, 2019).

However, as with big data in general (see section 1.1.2), in order to unlock its true potential for sustainable development, significant investments are still required. Or, as Ali Kharrazi, Hua Quin and Yi Zhang (2016:1297) put it: "big data is not a 'magic bullet' and requires investment in 'data infrastructure' before realizing [...] potential benefits". Albert et al. (2019) additionally note that only a limited number of countries currently have ongoing big data projects, highlighting the need for further investments - not necessarily only in data infrastructure but also the in the associated skill sets needed to work with big data and with that in educational programs to boost analytical capacities.

Another essential factor in leveraging big data for sustainability is internet coverage. This is especially important for a global monitoring of the Sustainable Development Goals using big data, as the SDGs aim to include the entire world population (United Nations, 2018). Steve McFeely (2019) acknowledges possible uncertainties that arise with generating SDG Indicators based on data which includes only half of the world's population. Indeed, internet penetration rates were estimated only at around 41% in 2017, meaning more than half of the people on earth did not have access to the internet at that time (ITU, 2017). His reservations are effectively the same as concerns from Crampton et al. and Albert et al. in the previous section (1.1.2).

Overall it seems like an integration of big data for sustainability is currently seen equally as dependent on technological advancements in the domains of data gathering and data processing as it is on institutional readiness and societal acceptance. But the consensus seems to be that significant improvements in terms of cost reduction and automated data gathering will be made in the future. One question will be at what cost, in terms of privacy reduction, these improvements come for human beings. According to Edelman (2019) this question is here to stay. Another question will focus on how to interpret the massive amounts of data produced this way - for if terrestrial data sources are used, internet penetration rates may prove to be an additional primary obstacle to overcome in order to unlock big data's potentials for sustainable development.

One possibility to overcome this obstacle is to find ways in which data interpretation can be automated as well. Recent advancements in artificial intelligence hint at currently still uncovered, vast potentials for interpretation automation, as the example of automated solid waste detection by Anjali Anadkat et al. (2019) shows. They developed an image classification algorithm which detects different kinds of garbage in digital images. Others, like Krishna Karthik Gadiraju et al. (2018) have applied image classification techniques to remote sensing data in order to detect informal settlement areas. Automating the interpretation of data sets which are too big to be analyzed conventionally by human beings is an exciting prospect for leveraging big data in efforts of improving sustainability.

1.2 Sustainable Development

This section introduces the concept of sustainable development (SD). After a brief overview of its origins, some of the many important steps in the adoption of SD into today's global political agenda by the United Nations are highlighted, before the focus of this section is shifted towards development disparities and the rural access index.

1.2.1 Origins of Sustainable Development

Ulrich Grober (2007) describes how today's notion of sustainable development originated from the concept of sustainability. Grober further elaborates on how the term "Sustainability" was first introduced to the domain of forestry through Hanns-Carl von Carlowitz (1732) with his *magnum opus* "Sylvicultura Oeconomica" in which he described the necessity of a controlled and sustained use of timber. He explains that timber was an essential resource at the time that could not be substituted. According to Grober (2007:18), von Carlowitz criticized "the contemporary short-termed way of thinking which was centred solely on making money", thus emphasizing that society should assure a steady supply of timber through conservation and reforestation efforts in order to guarantee the continual and sustained use of the resource.

The following centuries saw authors like Thomas Robert Malthus (1926) and George Perkins Marsh (1965) as well as the Club of Rome (1972) publish their concerns about human overpopulation, resource shortages and a possible system collapse of the world as it was. In *The Limits to Growth*, Donella Meadows and the Club of Rome (1972:23) concluded that "if the present growth trends in world population, industrialization, pollution, food production, and resource depletion continue unchanged, the limits to growth on this planet will be reached sometime within the next one hundred years. The most probable result will be a rather sudden and uncontrollable decline in both population and industrial capacity". Jacobus A. du Pisani (2006) gives a comprehensive and detailed overview of this period in the history of the idea of sustainable development and the various theories on development and progress that preceded it.

More than 200 years would pass after Carlowitz' concerns until the modern notion of sustainable development was introduced formally into global politics. Michael Redclift (2005) explains that through the report on global environment and development by the *Brundtland Commission*, or "World Commission on Environment and Development" (1987), the term "Sustainable Development" was introduced into political vocabulary. Gro Harlem Brundtland (1987:292), who headed the commission, defines sustainable development as development that meets "[...] the needs and aspirations of the present generation without compromising the ability of future generations to meet their needs".

1.2.2 Sustainable Development as a Geopolitical Paradigm

The Brundtland definition is the cornerstone of sustainable development as it is known today. And while it was the Brundtland Report that introduced SD into political agendas arond the world, the need for specific, quantifiable goals to work towards arose (Du Pisani, 2006). Shantayanan Devarajan et al. (2002) as well as David Hulme (2009) illustrate the progression from just the idea of SD, through major stepping-stones like the *United Nations Conference on Environment and Development* in Rio de Janeiro (in 1992), the *International Conference on Population and Development* in Cairo (in 1994) and the *World Summit on Social Development* in Copenhagen (in 1995), to the first major global development framework: the Millennium Development Goals (MDGs).

The MDGs were introduced in September 2000 at the *United Nations Millennium Summit* in New York City (UN General Assembly, 2000). These Goals were aimed at issues of poverty, hunger, primary education, gender equality, child mortality, maternal health, preventable diseases, environmental sustainability and a global partnership for development. Towards the end of the 15 year period of the MDGs, Jeffrey David Sachs (2012:2206) states that "developing countries have made substantial progress towards achievement of the MDGs, although the progress is highly variable across goals, countries, and regions". He further explains how the world has entered a new geological epoch in which human activity has become the most dominant force in fundamental earth dynamics. While this notion is not universally accepted, it illustrates a further shift in societal consciousness towards Sustainable Development (Heikkurinen et al., 2019). Sachs (2012:2207) further argues that "in view of [...] dire and unprecedented challenges, the need for urgent, high-profile, and change-producing global goals should be obvious". The MDGs brought sustainability onto the global political main stage and turned sustainable development into a geopolitical paradigm. Yet most of the challenges the MDGs were addressing persisted at least to some degree past their expiration date (Sachs, 2012).

1.2.3 17 Goals to Transform Our World

Because of the persistent challenges mentioned above, the UN General Assembly adopted the new *2030 Agenda for Sustainable Development* on 25 September 2015 (United Nations, 2018). The new agenda consisted of 17 goals and originally included 169 subordinate targets, making it the most extensive global development framework to date (UN Statistics Division, 2019c). Compared to the MDGs, the SDGs thus cover more dimensions of development more specifically (see table 1). In March of 2018, as well as one year later in March of 2019, the list of indicators for the SDGs was expanded. Today it consists of 232 indicators (UN Statistics Division, 2019a). All indicators are classified in tiers that determine their conceptual clarity and progress towards international methodological standards for data collection:

Goal	SDGs (2015-2030)	MDGs (2000-2015)
1	No Poverty	Eradicate Extreme Poverty and Hunger
2	Zero Hunger	Achieve Universal Primary Education
3	Good Health and Well-Being	Promote Gender Equality and Empower Women
4	Quality Education	Reduce Child Mortality
5	Gender Equality	Improve Maternal Health
6	Clean Water and Sanitation	Combat HIV/AIDS, Malaria and other Diseases
7	Affordable and Clean Energy	Ensure Environmental Sustainability
8	Decent Work and Economic Growth	Global Partnership for Development
9	Industry, Innovation and Infrastructure	
10	Reduced Inequalities	
11	Sustainable Cities and Communities	
12	Responsible Consumption and Production	
13	Climate Action	
14	Life Below Water	
15	Life on Land	
16	Peace, Justice and Strong Institutions	
17	Partnerships	

Table 1: Comparison between SDGs and MDGs (United Nations, 2017).

Box 1: Tier Classification of SDG Target Indicators

Tier I: [The] indicator is conceptually clear, has an internationally established methodology and standards are available, and data are regularly produced by countries for at least 50 per cent of countries and of the population in every region where the indicator is relevant.

Tier II: [The] indicator is conceptually clear, has an internationally established methodology and standards are available, but data are not regularly produced by countries.

Tier III: No internationally established methodology or standards are yet available for the indicator, but methodology/standards are being (or will be) developed or tested.

(UN Statistics Division, 2019b)

1.2.4 Development Disparities

However, there sometimes are significant disparities in regional and national development under the SDGs (United Nations Statistics Division, 2019c). As outlined by Peter Frankenfeld (2005), disparities describe inequalities. Frankenfeld (2005:190), who focuses on *spatial* disparities, sees inequalities as "structure characteristics" of today's societies. He argues, that in order to analyze spatial disparities, a reasonable delimitation of space has to be performed first. He further asserts, that while often the focus of studies concerning spatial disparities is of economic nature, they are present in all aspects of society, including education, mobility and healthcare (see also Burzynski et al., (2019) and Mokdad et al., (2019)). Frankenfeld (2005:185) additionally acknowledges, that addressing disparities is always closely related to normative questions. One of which might be; *To what extent do we want to tolerate certain disparities?*

As seen in table 1, goal 10 of the SDGs (*"Reduce inequalitites within and among countries"*) specifically calls for a reduction of disparities. In fact, one main objective of the SDGs as a whole, is to alleviate some of the most pressing issues resulting from (development and/or spatial) disparities, including issues of unequal access to political, social and economic processes for all people (United Nations, 2017).

As outlined in more detail below, this study is focused in part on disparities of infrastructure granting access to particularly economic opportunities. Its focus is on SDG target indicator 9.1.1: *Proportion of the rural population who live within 2 km of an all-season road* (see sections 1.2.5 and 1.2.6). Mobility and access to urban centers is crucial for the rural population. For farmers in rural areas, stable access to markets in more densely populated areas is paramount to their economic success, for their ability to access markets within reasonable time frames and using modern modes of transport positively correlates with their productivity, as indicated in the Transport & ICT Report (2016).

1.2.5 The Rural Access Index

Until December 31 of 2018, indicator 9.1.1 was classified as a tier III indicator (SDSN, 2015). Today, it is classified as a tier II indicator, even though the methodological approach to gather data for indicator 9.1.1 has been around since 2006 (UN Statistics Division, 2019b). The indicator was first introduced by Peter Roberts et al. (2006) as Rural Access Index (RAI) in the context of the Results Measurement System of the International Development Association.

"In practice the RAI *measures the number of rural people who live within two kilometers* (typically equivalent to a walk of 20-25 minutes) *of an all-season road* *as a proportion of the total rural population.* An “all-season road” is a road that is motorable all year round by the prevailing means of rural transport (typically a pick-up or a truck which does not have four-wheel-drive). Occasional interruptions of short duration during inclement weather (e.g. heavy rainfall) are accepted, particularly on lightly trafficked roads” (Roberts et al. (2006:2). However, for the purposes of this study only asphalt roads were considered *all-season roads* (ASRs).

According to Roberts et al. (2006:4), RAI should be measured ”by analysis of household surveys that include appropriate questions about access to transport. The aim is to integrate this with the measurement of household characteristics such as income and access to services such as education, health and clean water supply”. Although this methodological approach has since been updated for the application as SDG target indicator 9.1.1, RAI is still considered to be among the most important global development indicators of mobility in the metadata-repository of the SDGs (UN Statistics Division, 2019c).

Today’s official methodological approach suggested by the Transport & ICT Report (2016), titled *Measuring Rural Access: using new technologies*, uses a combination of geospatial data (more specifically population distribution data, urban extent data, vectorized road data and measurements of road utility status) with a final spatial resolution of 100m x 100m, as opposed to data from household surveys. Requirements for the calculation of RAI are structured into three separate data requirement domains. It is in the domain of measuring road condition (data requirement 3) where the example provided in this thesis aims to make a contribution (see section 1.4). Therefore, data requirement 1 (population distribution data) as well as data requirement 2 (urban extent data and vectorized road data) are not introduced further in this section.

1.2.6 Challenges in Measuring Road Condition

The following section elaborates on remaining challenges concerning costs, information delay and expenditure of human labour mentioned in the Transport & ICT Report (2016). Table 2 gives an overview of what the Transport & ICT Report suggests as suitable sources for road condition data, along with some of their respective advantages and disadvantages.

Apart from ”Free Apps for Road Assessment”, all of the potential data sources are liable to pay costs. The Transport & ICT Report (2016:22) states that ”it is always possible to collect the necessary condition data with reasonable accuracy, although at a cost”. The costs for data collection are directly linked to the availability of data and the processing steps needed to extract relevant information, as well as initial investments for equipment (e.g. high initial investment costs for unmanned aerial

Data Source	Advantage	Disadvantage
Road Inventory Survey	Technically solid, consistent with government responsibility	Costly, Irregular updates, country-specific assessment standards
Satellite Imagery	Consistency across countries, potential for high frequency data collection	Costs, Technically challenging to identify road condition in detail, significant computational process required
Unmanned Aerial Drones	Good mobility	Technically challenging, computational process required
Call Detail Record	Consistency across countries, potential for high frequency data collection	Access to data, noise in data
Free Apps for Road Assessment	Cost effective, Potential contribution through crowd-sourcing	Statistical errors between measured IRI and actual roughness
Commercial Apps for Road Assessment	Relevant analytical tools provided together	Statistical errors between measured IRI and actual roughness

Table 2: Summary of possible Sources for Road Condition Data (Transport & ICT, 2016:23).

drones).

In terms of information delay, satellite imagery and unmanned aerial drones offer fast information. Meanwhile, data from road inventory surveys, call detail records (georeferenced information about calls made/received, owned by cell phone carriers) and mobile applications for road surface assessment rely on the frequency of surveys, recorded drives or calls made for timeliness. Depending on these factors, data availability may be good or lagging behind.

All potential data sources for modern RAI calculation are to some degree labour intensive. International, standardized procedures using satellite imagery, unmanned aerial drones, call detail records or data gathered through mobile applications can reduce initial-/ and upkeep costs, however.

1.3 Image Classification

This section covers the basics of image classification using convolutional deep neural networks. Considering deep neural networks themselves make up a wide field of research, the following focuses only on aspects relevant in the context of this thesis. First, the concept of deep neural networks is briefly explained. Second, basics of training and validating such models are mentioned. Third, the image classifier used in this study is introduced. For a more profound understanding of deep learning, the book *Deep Learning, A Practicioner's Approach* by Josh Patterson and Adam Gibson (2017) is recommended.

1.3.1 Mimicking the Animal Brain

Conceptually, deep neural networks attempt to mimic the functionality of the animal brain. Animal brains consist of a large amount of interconnected biological neurons (see figure 1). For example, it is estimated that there are around 1 billion neurons in a human brain (Barha et al., 2016).
Josh Patterson and Adam Gibson (2017) give a detailed overview of how the conceptualization of artificial neurons progressed over time. They explain how the earliest version of an artificial neuron was created in 1957 by Frank Rosenblatt. It is called *single layer perceptron* (see figure 2) and it is a binary classifier. They further explain that, comparison with a biological neuron, inputs to an artificial neuron (x_1 through x_n) represent information inflows from previous neurons. Each input is multiplied with a weight (w_1 through w_n) and each such result represents information being transferred along a dendrite in biological neurons. Weights scale information that enters the core of the perceptron, either amplifying or minimizing its impact on further processing of information. All of the products of inputs and weights are summed up.

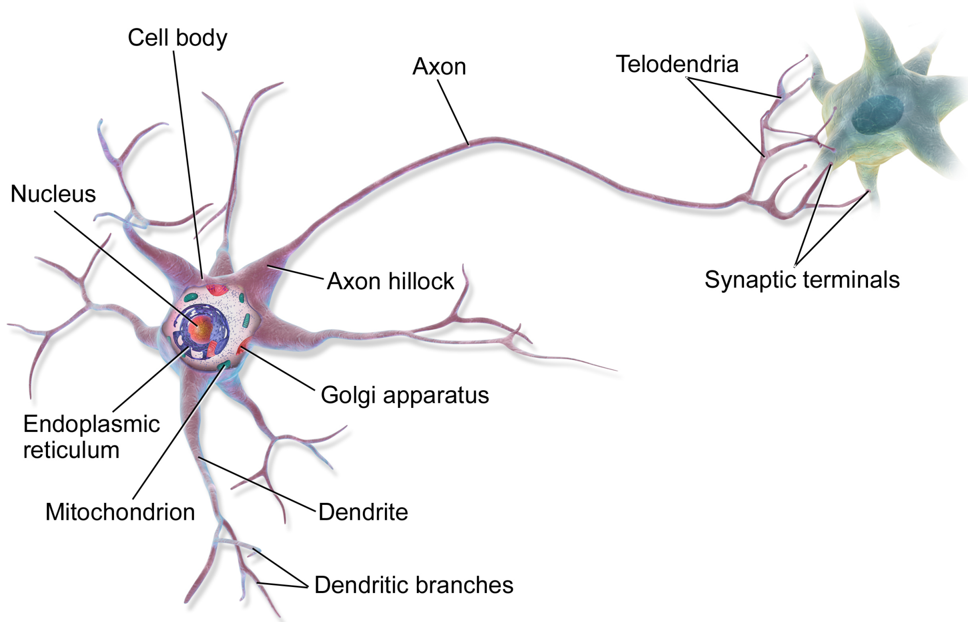

Figure 1: Biological Neuron. Courtesy of BruceBlaus, commons.wikimedia.org, CC BY 3.0.

Afterward, this sum is passed to a threshold function (e.g. a step function), which determines if the perceptron passes on information or not.

This is similar to the way biological neurons behave when information is processed in an animal's brain, where action potentials are created and information is passed on when the depolarization of a cell's membrane reaches a critical threshold (Barha et al., 2016:54).

Patterson and Gibson (2017) explain that a perceptron learns through randomly adjusting the individual values of its weights, until all inputs are correctly classified. If there is no linear function which can separate the input dataset, a perceptron will continuously try - and fail - to improve. Changing from a simple threshold function to a more flexible activation function (e.g. a *softmax* activation function) solves this problem. It converts a perceptron into an artificial neuron. Feeding the output of such a neuron to another creates a *multilayer feed-forward network*, this transmission of information is called *forward propagation*. Each stage of a feed-forward network is called a *layer* and can have one or more artificial neurons. Layers are labelled as input layers (the first layers to receive input information), output layers (layers that output results) and hidden layers (all of the layers in between) (Patterson and

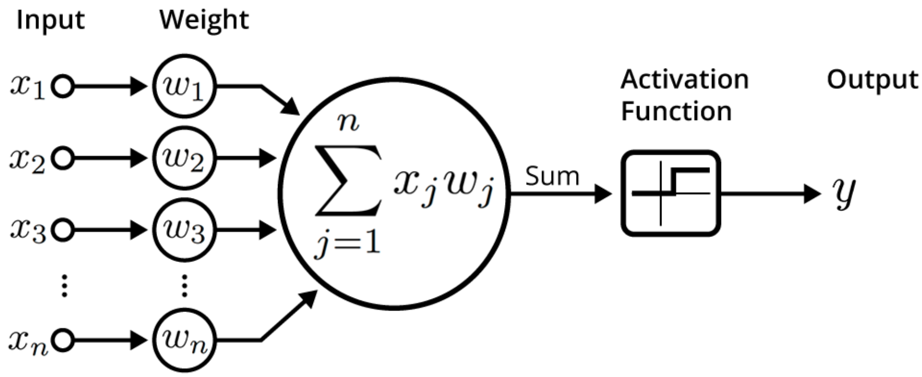

Figure 2: Single Layer Perceptron. Shubh Saxena, becominghuman.ai.

Gibson, 2017). If a network has multiple hidden layers, it is considered a deep neural network (DNN) (Mohammadi et al., 2018). If all neurons from a previous layer are propagating information to all neurons of the next layer it is regarded as a *fully connected layer*. Going beyond regular deep neural networks, *convolutional* DNNs are capable of feature detection. Convolutions (mathematical operations) make it possible to segregate an image into different features, which can then be assessed individually (Patterson and Gibson, 2017).

Similarly to the perceptron, a deep neural network learns by continuously and randomly adjusting its weights. An optimization algorithm (e.g. stochastic gradient descent) is used to reward the network for adjustments that improve its overall performance and to penalize adjustments that make it perform worse. To distribute penalties for bad performances, a process called *backpropagation* is used. Backpropagation essentially is a backwards pass-through of the network. First, error rates are calculated and the weights of the output layer are updated. Then, this process is repeated for each hidden layer, until the input layer is reached (Patterson and Gibson, 2017).

The optimization algorithm bases its judgment on the quantification of how well a network performed a given task. This value is called *loss*. Loss can be calculated in various ways. One example of a loss function is the *sum-squared error loss* used by Joseph Redmon et al. (2016).

1.3.2 Hyperparameter Tuning

Li Deng (2014), explains that backpropagation struggles to learn models containing more than a limited amount of hidden layers. This is why the learning optimization process needs to be controlled - or tuned - with further parameters, which are called *hyperparameters* (Patterson and Gibson, 2017). Patterson and Gibson (2017:78) explain that hyperparameters help "ensuring that the model neither underfits or overfits the training dataset, while learning the structure of the data as quickly as possible". Underfitting refers to a model that has not reached its potential, while overfitting refers to a model that is trained so specifically that it loses the ability to generalize and only performs well on the training dataset (Dietterich, 1995; Stephen et al., 2014:163)

Hyperparameters include *learning rate* (this governs by how much individual weights are adjusted during backpropagation), learning *batch size* (this regulates the amount of input features that are used to learn at each step) or the number of *epochs* (one epoch represents one full pass over all learning features).

Deng (2014:24) acknowledges the significant difficulty of sensible hyperparameter tuning as "one major barrier to the application of DNNs" because "it currently requires considerable skills and experience to choose sensible values" for these parameters. Due to the difficulty of hyperparameter tuning, it is sometimes seen as more of an art than anything else (Brecque, 2019). Or as Will Koersen (2018) put it: "Hyperparameter tuning relies more on experimental results than theory". Deng (2014) explains that there is more research needed to find sensible ways of hyperparameter tuning in order to achieve the best possible results with deep learning. He mentions random sampling, where hyperparameters are randomly generated and tested, after which point all of the hyperparameter settings are compared to one another.

1.3.3 Training and Validating Deep Neural Networks

Expanding on the previous section, the following offers a shallow introduction into the training and validation process of deep neural networks. It is intended to cover, briefly, only what is relevant in the context of this study. For a deeper dive into the training process of neural networks, contributions by Deng (2014) and Shaoqing Ren et al. (2016) are recommended.

Training a Deep Neural Network boils down to repeated and guided random adjustments of its weights (as mentioned in the previous section) until the model makes desirable predictions. As described in the previous section, validation of hyperparameter settings also takes place during training, as training a DNN is an iterative process. One important step to take when validating DNNs is the separation

	Positive' (predicted)	**Negative' (predicted)**
Positive (actual)	true positive	false negative
Negative (actual)	false positive	true negative

Table 3: Confusion Matrix (Patterson and Gibson, 2017:37).

of training data and validation data. A model should not be validated on the same data it has trained on, as this increases the probability of overfitting the model (Koehrsen, 2018).

As Patterson and Gibson (2017:36) point out, context is key in validating DNNs. Using a confusion matrix, they illustrate their argument (see table 3): "Sometimes, we only care how often a model gets any prediction correct; other times, it's important that the model gets a certain type of prediction correct more often than the others". Applying their point in the context of image classification, one can imagine that sometimes, it is important that an image classifier finds every instance of a given object, a toothbrush for instance, in an image (resulting in a high score for true positives), other times it might be more important that it does not label anything other than a toothbrush as such (resulting in a low score for false positives).

The effectiveness of neural networks can be measured using tools that keep the ideas from the previous paragraph in mind: recall and precision (Singhal, 2001). The following gives a brief but unfinished overview of these concepts. Recall measures if a model finds all relevant cases (or all of the ground truth bounding boxes) while precision measures the proportion of found cases that are relevant (i.e. correct positive cases):

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

$$Precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Using precision and recall, another metric, average precision (AP, representing the area under a precision/recall curve) can be calculated for each prediction (Henderson and Ferrari, 2017):

$$AP = \int_0^1 p(r)\, dr$$

AP can be calculated for all classes an image classifier can predict. The mean average precision (mAP) over all classes is then used to gauge the quality of predictions made by the classifier (Henderson and Ferrari, 2017). It is calculated as follows:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i$$

Predictions from image classifiers are made using bounding boxes, depicting in which area of an image an object is situated. For a prediction to be considered correct, it usually has to cover at least 50% of the ground truth bounding box - a threshold made popular through the PASCAL VOC Challenge (Everingham et al., 2010). This is illustrated using a metric called intersection-over-union (IoU), which measures the area of overlap between a prediction bounding box and a ground-truth bounding box (see figure 3).

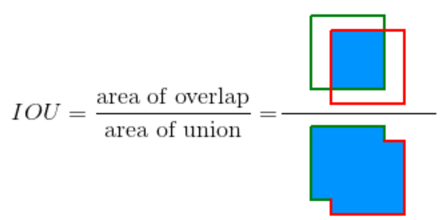

Figure 3: Intersection-over-Union (IoU). Courtesy of rafaelpadilla, github.com, MIT License.

1.3.4 YOLO, Darknet and Darkflow

This section introduces the neural network chosen for this study, YOLO (see also section 2). YOLO (You Only Look Once) is an image classifier published by Joseph Redmon, Santosh Divvala, Ross Girshick and Ali Farhadi (2016). Their network allows for classifications of images with a single pass-through of information, whereas other networks assess bounding boxes multiple times before outputting a final classification (Girshick et al., 2014). Redmon and Farhadi (2018:1) explain that this ability to assess images on a single pass of information makes YOLO "significantly faster than other detection methods with comparable performance". To train a YOLO network, the *Darknet* framework is used. Redmon describes Darknet as "an open source neural network framework written in C and CUDA" (Redmon, 2019). For this study, a Python implementation of Darknet - *Darkflow* - was used, as outlined in more detail in section 2.2 (thtrieu, 2019).

1.4 Goals of this Study

From this introduction it becomes apparent that there is a clear and urgent need for research in the domains of sustainable development, big data and machine learning, especially with a focus on combining big data and machine learning to generate much needed information on the progress of sustainable development as measured by the Agenda 2030 indicator catalogue.

This study therefore aims to make two important contributions to these efforts: First, using data supplied by the Twitter streaming API, it provides an example case which illustrates some of the challenges to overcome in order to realize the full potential of georeferenced big data and machine learning techniques for SDG monitoring. Second, it assesses potentials and limitations of the combination of big data and machine learning techniques for SDG monitoring, once the aforementioned challenges can be overcome.

1.4.1 Research Questions

In this section, research questions based on the goals of this study are formulated. Research questions 1 and 1.1 are directly linked to target indicator 9.1.1 of the SDGs (see section 1.3.4). Research question 2 is oriented towards the potential overall contribution of Big Data for Sustainability:

Research Question 1
To what extent can georeferenced road surface information for indicator 9.1.1 of the SDGs (RAI) be generated using a deep neural network on the Twitter Streaming API?

Research Question 1.1
Are these data (from RQ 1) comparable to conventional data for indicator 9.1.1 (RAI) of the SDGs in terms of quality, coverage and accuracy?

Research Question 2
What are potentials and limitations of Big Data analyses for monitoring the SDGs?

Hardware Specifications	
Processor	AMD FX(tm)-6100 @ 3.30 GHz
RAM	8.00 GB
Graphics Card	NVIDIA GeForce GTX 760
	1152 CUDA Cores
	2.00 GB of GDDR5 VRAM

Table 4: Hardware Specifications of the Computer used to train ICARUS.

2 Methods

This section gives an overview of the methods used to conduct this study. It is structured in chronological order of data processing steps (see also figure 6). First, the data source is introduced. Second, the process of harvesting images for training the *Image Recognition Algorithm for Road Utility Status* (ICARUS) is explained, as well as other methods used to train the algorithm. In the third subsection, validation procedures used during this process are described. The fourth part explains how ICARUS was run. The final section covers how the results form running ICARUS were mapped. All of the scripts, resources and outputs generated throughout this process are available from the author's GitHub page.

2.1 Data Source

For this thesis, the Twitter streaming application programming interface (Twitter streaming API) was used as the single data source, as at the time it was the only streaming API for social media which was accessible without request limitations or prior submission of an application to use it with. The streaming API offers an inherent option to filter for tweets with attached geographical coordinates in latitude/longitude format. It is estimated that the total amount of tweets including lat/lon geotags adds up to approximately 0.1% (Crampton et al., 2013). A second filter was implemented to additionally percolate the stream for tweets which include media appended to them. Thus, a CSV file could be generated that recorded coordinates, the URL of appended media and information on the date and time of a published tweet. Such data was collected globally, encompassing every tweet published through the API between 12 May and 23 September 2019.

2.2 Training the Classifier

In this study, a YOLO image classifier was used, implemented in the Darkflow environment. YOLO was chosen for its fast detection speeds, which were required to deal with real-time data flows in testing and the large harvest data set. The decision to use Darkflow-YOLO depended on the necessity to use YOLO with Python (see also section 1.4.4).

To generate a set of training images, the Twitter streaming API was combined with the Google Cloud Vision API. Upon signing up with the Vision API, users are granted a free credit to test the API, which was used to filter out images from Twitter containing asphalt roads. Upon expiration of this free credit, a set of 5000 training images and a further set of 200 validation images were generated and manually checked for mistakes. Once checked, both sets were classified by manually drawing bounding boxes around areas in the images containing segments of asphalt road. As mentioned in the introduction, only asphalt roads were considered *all-season roads* (see section 1.2.5). This simplification was made due to time constraints associated with a master's thesis. In total, more than 18 000 labels were generated.

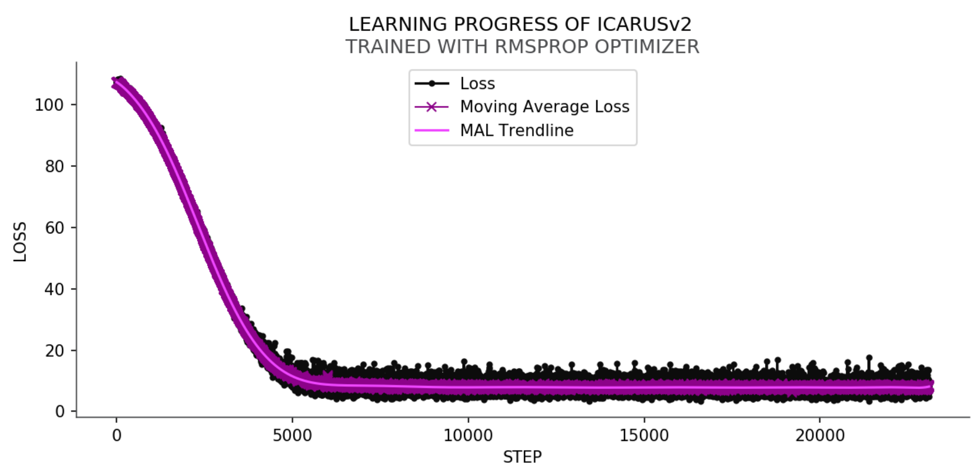

Figure 4: Learning Progress of ICARUS (Steps 0 - 20 000).

Because of hardware limitations (see table 4) the DNN architecture of tiny-YOLO was chosen for ICARUS. With Darkflow, ICARUS was trained using a RMSPROP optimizer (Tieleman and Hinton, 2012). Through adjustments of the learning rate whenever the step loss reached a plateau, a final loss of around 4-5 was reached (see fig. 4 and 5). This procedure was done multiple times to test different optimizers as well as hyperparameters, hence the versioning (ICARUSv2) in figures 4 and 5. For all

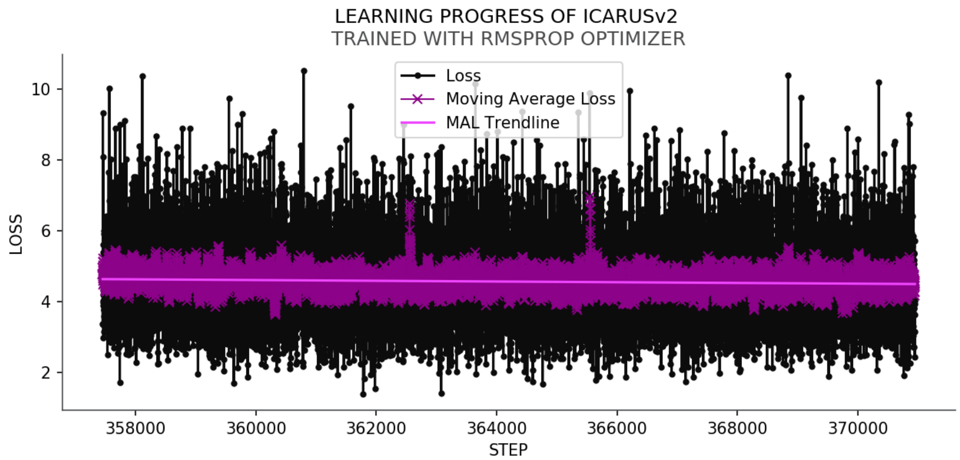

Figure 5: Learning Progress of ICARUS (Steps 358 000 - 370 000).

of this thesis, the acronym ICARUS references this ICARUSv2 version. It should be noted that because of aforementioned hardware limitations, the maximum batch size possible during training was 16. In order to log values for loss and moving average loss, the `"darkflow/net"` script was modified slightly.

2.3 Validating the Classifier

As a contextual experiment during the early stages of development, the Google Streetview API was used to test ICARUS. The underlying reasoning for this was the assumption that ICARUS might classify bare rocks as all-season roads due to possible similarities between color and texture of these two media.

To test this, bounding box lat/lon coordinates were used to set up a 10x10m mesh grid. Coordinates of each intersection were queried with the API, then the image response from the API was passed through an early version of ICARUS. Locations for each prediction were then mapped. This procedure was done first for a sparsely populated area south of the river Allenbach in the municipality of Adelboden, Switzerland. It was then repeated for a more densely populated area in Mossel Bay, South Africa.

The results of these tests were satisfactory, in that all predicted all-season roads were in close proximity of known streets when viewed on OpenStreetMap. This early experiment was not conducted to test the quality of ICARUS as an image classifier. It did, however, confirm that the development of ICARUS was heading in a reasonable direction.

A program was written to gauge how many images with all-season roads are missed entirely by ICARUS. This script was run every time ICARUS produced a checkpoint in training. It outputs the percentage of hits on the validation set. Checkpoints 344150, 344750 and 357450 had the best hit rates at 71, 68 and 80%, respectively.

For a standardized assessment of the quality of ICARUS, mAP was calculated (see section 1.4.3). As the algorithm was *trained* on the 5000 training images, these could not be used for validation. The additional, manually classified validation set consisting of 200 images (ground truth) was used for this step. To calculate mAP, a script by GitHub user João Cartucho was used. This process was repeated for the 13 best performing checkpoints in terms of hit rate, showing that checkpoint 344750 had the highest mAP score out of all of them. Therefore, this checkpoint was ultimately used to run ICARUS.

With reference to the context-dependent nature of DNN validation mentioned by Patterson and Gibson (2017) in section 1.3.3, the IoU threshold during mAP calculation was set to 0.3. This value was chosen because of the amorphous nature of road segments in most images on which the algorithm trained.

2.4 Running ICARUS

This section covers the process of conducting the analysis outlined in previous sections, as illustrated in figure 6. In a first step, the Twitter streaming API (see section 2.1) was filtered for tweets containing appended geodata and media. Metadata (coordinates, URL of attached media, date and time, as well as a unique identifier) for each tweet matching these criteria were saved as raw data. Afterward, these data were assessed in bulk using ICARUS with a minimal *prediction threshold* of 0.5 and later visualized. The following sections provide a more detailed overview over this procedure.

2.4.1 Harvesting Data

Because ICARUS could not run indefinitely on the author's desktop PC, and no server was used to execute all of the steps in figure 6, the processes of data gathering and data assessment were split up. As a low-cost and resource efficient solution, a Raspberry Pi 3 (RasPi) was deployed to filter the Twitter streaming API. It has a significantly lower energy consumption than a regular desktop PC.

With the RasPi deployed, the API could be scanned indefinitely. Through remote access to the RasPi, data gathering was monitored. Sporadically, data harvesting was manually suspended for a short period of time, which meant the RasPi had to generate a new savefile to write into. This was done to avoid total data loss in case

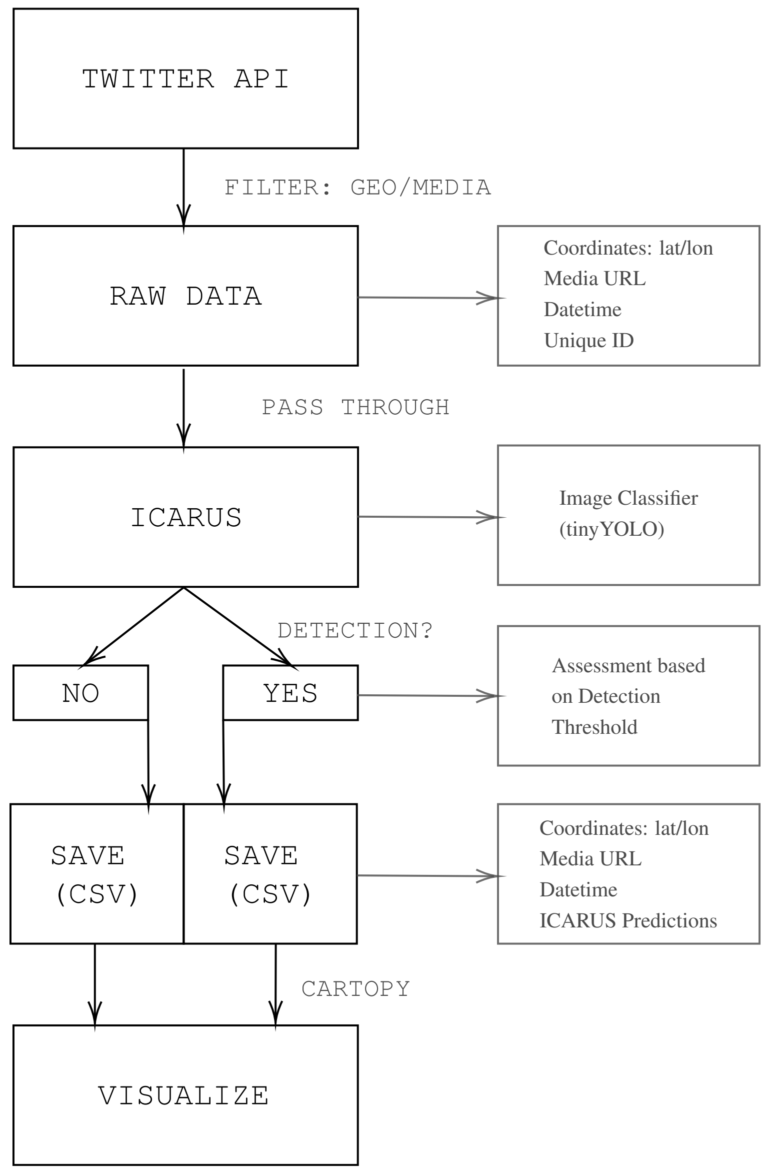

Figure 6: Flowchart of Data Processing Steps.

severe complications occurred and a savefile got corrupted. For convenient further use, all savefiles were later re-joined.

2.4.2 Image Classification with ICARUS

As mentioned above, the harvest savefile contains a URL for media appended to each tweet. For the classification step, ICARUS iterates over each URL, downloads the associated media and classifies its contents. To manage the toll this takes on disk space, media attachments are overwritten with information from the next Tweet each time a new classification step begins, effectively only saving one image at a time, resulting in minimal disk space requirements. For each iteration, ICARUS saves coordinates, URL, date and time, average prediction confidence and a dictionary of all predictions made (including bounding box coordinates for each prediction) into a separate output CSV file (see figure 6).

Due to the aforementioned separation of processing steps into filtering tweets on a RasPi and image classification with ICARUS on a more powerful desktop PC, the time needed to analyze harvested tweets was condensed. Analyzing 500 000 tweets in bulk takes 4 days and 6 hours on the machine used in this study. It was therefore decided that one month worth of harvests (12 May - 12 June 2019, consisting of 552 004 tweets) would be analyzed, for the computer was needed in other projects too often to allow for further classifications of harvests. An integrated version of the stream listener and ICARUS exists, however, which would make an indefinite, real-time assessment possible using dedicated hardware.

2.5 Visualization

This section explains how results from both the Twitter streaming API and ICARUS were visualized. All visualizations in this thesis were made using the matplotlib and numpy libraries for python, or with ArcGIS Pro. For mapping purposes, the cartopy library was used. All geographically projected visualizations are projected in EPSG 32662 with a central longitude of 0.0.

2.5.1 Visualizing Harvests

To visualize all harvested tweets, two approaches were used. First, all harvested tweets with geotag and appended media were drawn onto a worldmap as squares at 20% opacity. This approach was chosen to illustrate the absolute amount of harvested data. Even though features were drawn with opacity, there were so many of them

Parameter	**Input**
Input Feature	Consolidated Streaming API Harvests
Population Field	None
Output Cell Size	0.1
Search Radius	2
Area Units	Square Kilometers
Output Cell Values	Expected Counts
Method	Geodesic

Table 5: Parameters used to calculate Kernel Density of Harvests in ArcGIS Pro.

that the intended side-effect (visualizing density using opacity) was not achieved, as there is a lot of overlap of single features when mapped like this.

Therefore, a more direct approach of calculating kernel density (KDE) was used. This step was executed in ArcGIS Pro, as KDE calculation with matplotlib and scikit-learn resulted in a longitudinal distortion in the computed density layer. It should be noted, that while correcting this distortion would be possible in a python environment, ArcGIS Pro was used to save the time needed to implement such a correction. From ArcGIS Pro, a layer was then exported and fed back into the python script for further mapping. To adequately represent density hot spots and account for the uneven global distribution in the harvested dataset, a logarithmic *colormap* was used to illustrate data density. A summary of the parameters used to conduct the kernel density analysis is provided in table 5.

2.5.2 Visualizing Predictions

As mentioned above, ICARUS saves a dictionary of predictions (including confidence and bounding box information) for each assessed image. Additionally, a value for mean prediction confidence is calculated for each such data point. Using bounding box coordinates, exact parts of an image that a classification is based on can be visualized with their respective prediction confidence (see fig. 7 and 8). Multiple predictions can be drawn on the same image, resulting in multiple green bounding boxes as well as multiple prediction confidence indicators.

Predictions from ICARUS were visualized in the same python environment as

Figure 7: Visualization of Predictions with ICARUS.

mentioned in section 2.5. To color-map prediction confidences onto a world map, the aforementioned mean prediction confidences were used. The total amount of predictions form ICARUS (based on 1 month of harvests) did not warrant a similar procedure to calculate kernel density as implemented for harvest mapping. However, if ICARUS were integrated directly into the harvesting process, such a calculation would be possible and reasonable.

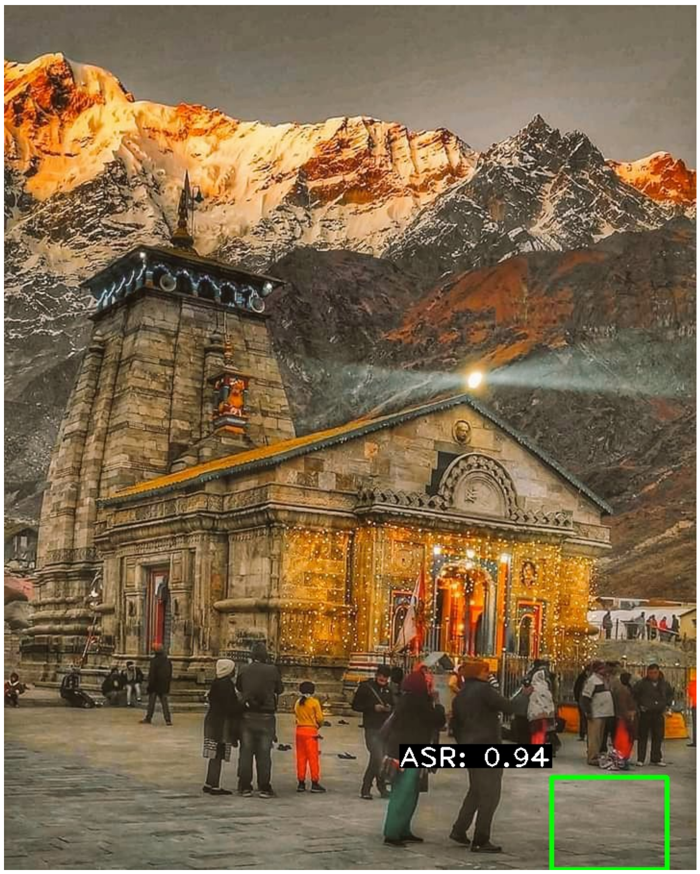

Figure 8: Visualization of Predictions with ICARUS.

3 Results

In this section, the results of this study are presented. It is split into three main parts: Big data from Twitter as a data source for sustainability monitoring, ICARUS as an image classifier as well as the fully integrated 1-month analysis of road surface condition using ICARUS.

3.1 Geotagged Images from Twitter as a Data Source

After the previous section presented the final version of ICARUS, this section is focused on the results of harvesting big data from Twitter. As figure 9 shows, the average amount of tweets being published per day including geodata and appended media is currently around 15 000 during weekdays and around 20 000 on weekends. In May, June and September there presumably were connection errors or internet outages, resulting in momentarily very low amount of harvests (represented by short and sharp drops in figure 9).

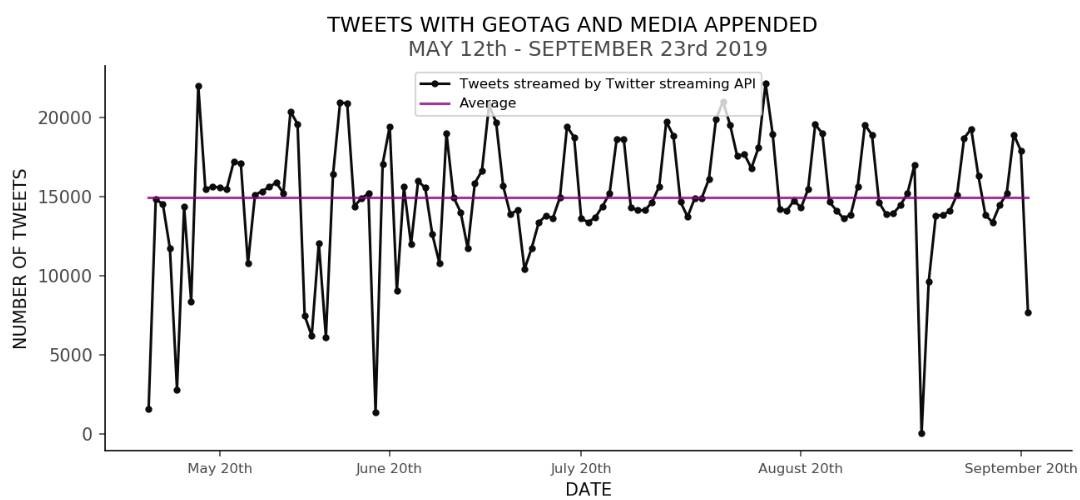

Figure 9: Twitter Streaming API Output Over Time.

The following pages include visualizations of all harvested tweets between May and September of 2019 (see figure 11), as well as a visual representation of tweeting density per square kilometer during the same 5 month period (see figure 12).

Figure 11 contains 1.8 million single points representing tweets including a geotag and appended media. As it was generated from harvests during a 5 month period, the extrapolated total amount of tweets per year meeting these criteria lies somewhere around 5.5 million. Horizontal lines across the whole map at approximately 25° and -25° latitude result from Twitter activity related to the New Horizons and DISCOVR spacecrafts.

From figure 12 it becomes apparent that Japan, some parts of Turkey, Europe and the United States of America produce the highest amounts of tweets in the 5 month harvesting period with around 50-300 tweets/km^2, followed by some parts of South East Asia and the Middle East with around 5-50 tweets/km^2. Rough extrapolation estimates result in 120-800 tweets/km^2/y, or 12-120 tweets/km^2/y for these respective areas. A comparison of human population density (see figure 14, Annex) and tweeting density yields further insights. Mainly, that some of the most densely populated areas (India; Northern India in particular, and Eastern China) produce little output of tweets containing geodata and appended media. Similarly, most populated areas in Africa show sparse densities. Possible reasons for these findings are discussed in more detail in section 4.

3.2 ICARUS as an Image Classifier

Visualizing bounding boxes of predictions yields valuable insights into the strengths and weaknesses of the classification algorithm. Upon closer inspection of figures 7 and 8, it becomes apparent that in figure 7 ICARUS classified all areas containing an all-season road correctly, while in figure 8 it only classified a small part. Yet in the second example it returns a higher prediction confidence. This behaviour was observed commonly and is represented in ICARUS' mAP score of 0.14 (see sections 1.4.3 and 2.3). Additionally, manual assessments further suggest that ICARUS struggles to identify asphalt roads in the foreground of images, but excels at classifications in the middle ground. Both of these issues are further discussed in section 4.

3.3 1 Month ICARUS Analysis

As explained in section 2.4.2, a total of 552 004 harvested tweets were analyzed with ICARUS. This is the equivalent of one month of harvests. Validation of ICARUS showed, that prediction thresholds of > 0.8 provided most reliable outcomes. Results of this analysis are visualized in figure 13. From the total amount of analyzed harvests, only 16 863 ($\approx 3\%$) were classified as *all-season roads*. Japan, Europe, the United States of America, as well as some parts of South-East Asia have the highest occurrences of ASRs. Meanwhile, some of the previously mentioned, densely populated areas in India and China show comparatively low amounts of ASR detections.

Figure 10 illustrates the amount of ASR detections using a prediction threshold of

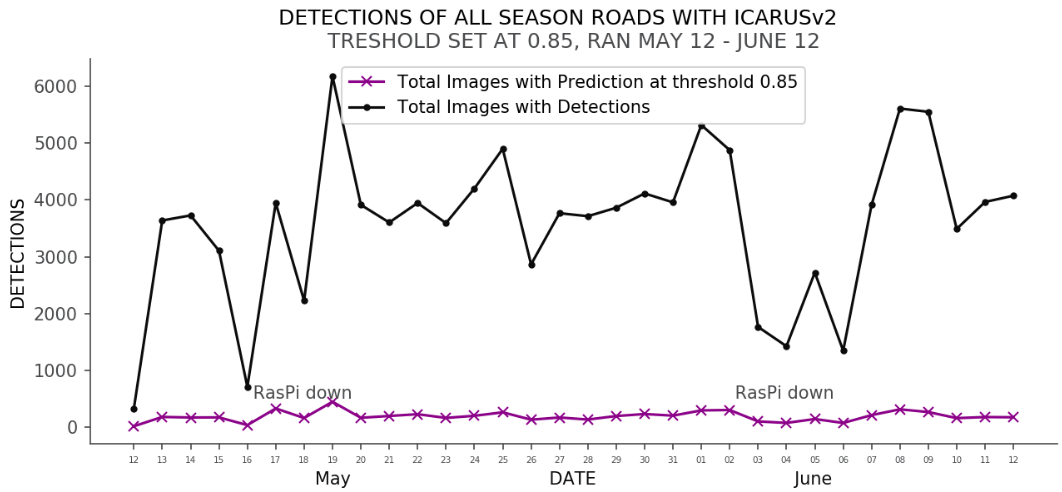

Figure 10: Discrepancies between Outputs at different Prediction Thresholds.

0.85 in relation to the number of assessed tweets which got at least one prediction at threshold the minimal threshold of 0.5. It shows that only a small part of predictions are at a level of 0.8 or higher, which (as mentioned above), was found to be necessary for reliable results.

To communicate the findings of this study more effectively, a website was created at https://taetscher.github.io/interactiveICARUS/, which contains an interactive map of ICARUS' 1 month analysis at prediction threshold 0.9. The jump from threshold 0.8 to 0.9 was made because of an even lower amount of predictions at that threshold, which benefits the website's performance.

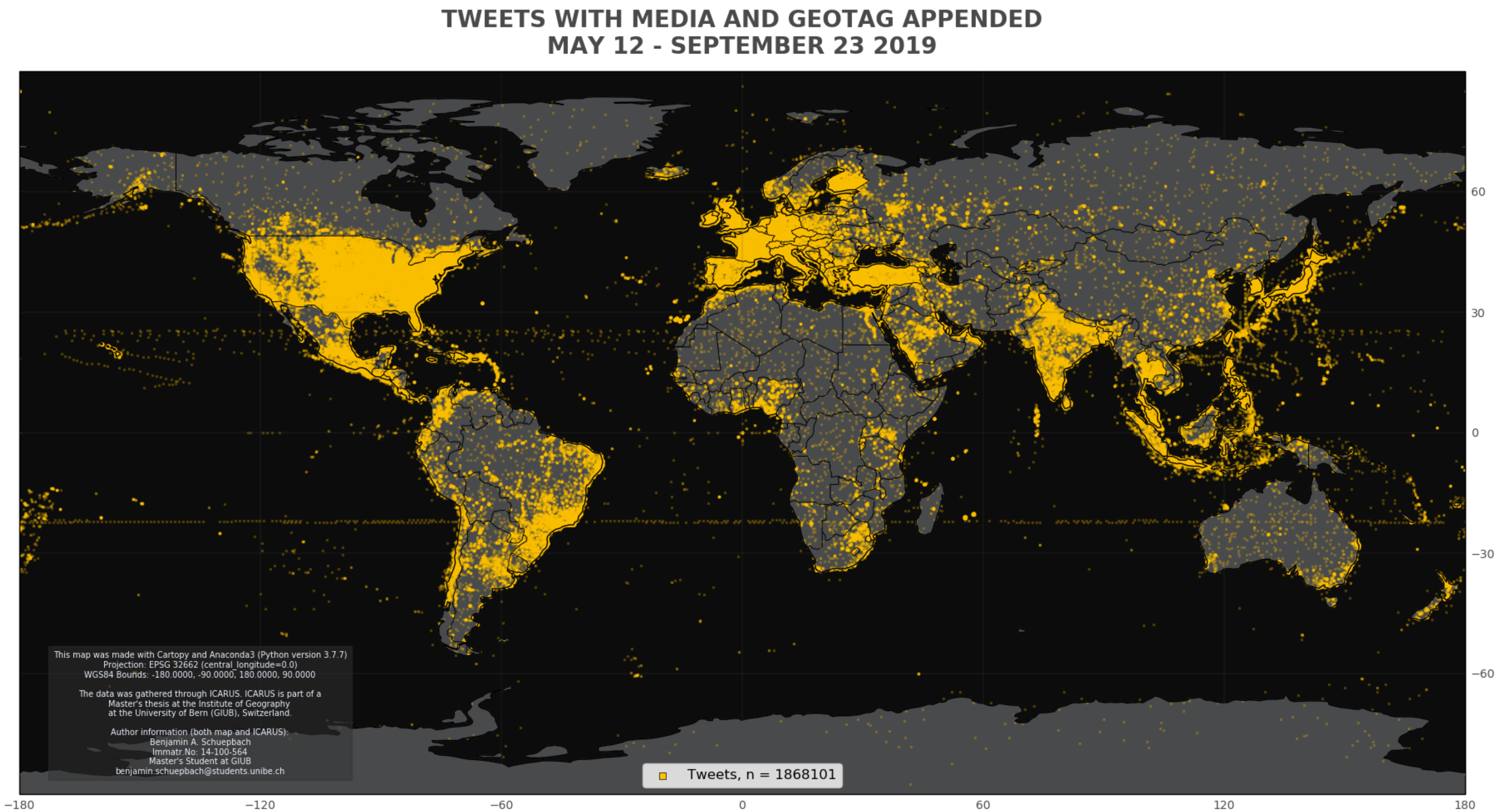

Figure 11: Tweets Harvested between May and September 2019.

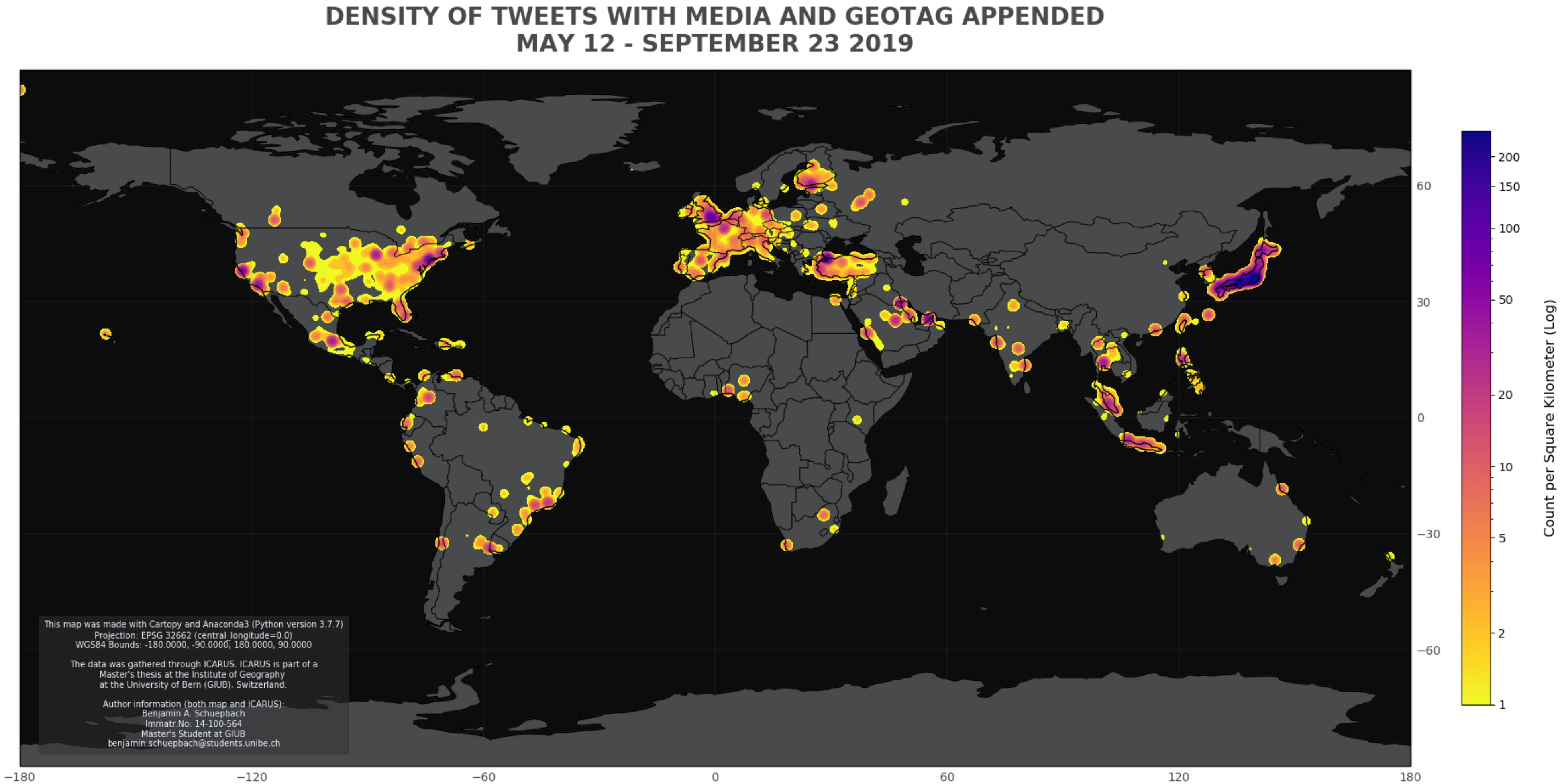

Figure 12: Density of Harvested Tweets.

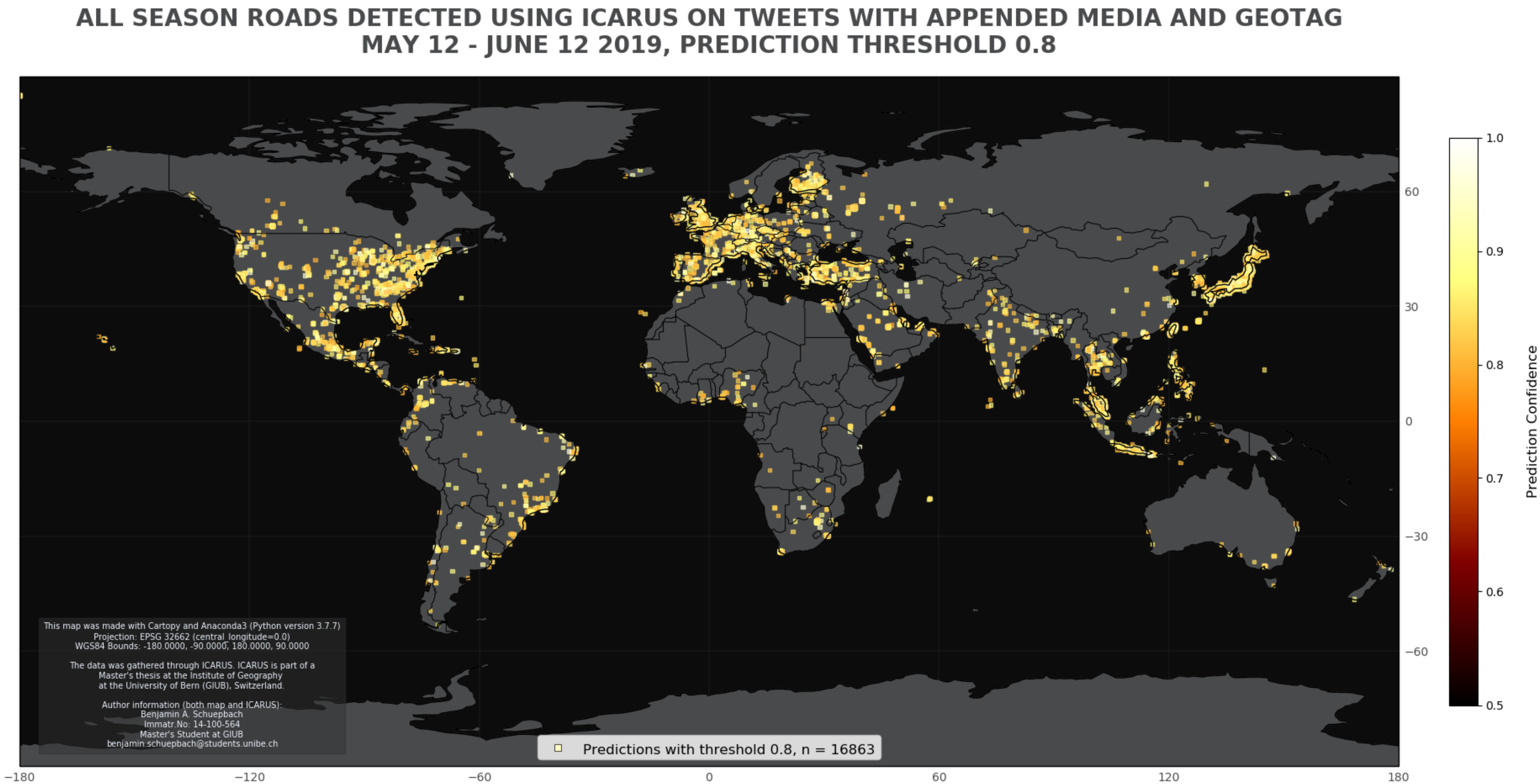

Figure 13: ICARUS Output @ Prediction Threshold 0.8.

4 Interpretation & Discussion

This section discusses previously presented results (see section 3) and answers the research questions posed in section 1.4. First, implications of choosing a data source for big data analyses are examined. Second, ICARUS and its flaws are discussed, answering research questions 1 and 1.1. Third, concerns about the use and potential harmfulness of big data are considered. Fourth, taking into account all of the above, current potentials of using big data to monitor the SDGs are explained, answering research question 2. Finally, general potentials in the field of Geographic Information Science, generated through access to big data and machine learning, are explored.

4.1 Sourcing Big Data

As Benjelloun et al. (2015) explain, the true potential of big data lies in integrating different sources of information into one analysis scheme (see section 1.1.1). This both highlights the difficulty of finding single data sources to cover an area of interest adequately, as well as potential gains from combining various big data sources.

In the context of this study, this relates to the impacts on map coverage arising through the (un-)availability of data. As outlined in section 2.1, Twitter was the only freely available, real-time data source for social media data at the time of conducting this study. However, as the main focus of Twitter lies on the transmission of newslets, other sources like Instagram or Facebook might yield more image data, more densely covering a larger area of the world. Particularly Instagram might be a gold mine for analyses similar to the one in this study, as its main focus lies on sharing images with the world. Furthermore, Instagram also has a feature which allows users to share a geotag alongside an image.

With restricted access to contents of such social media outlets and a simultaneous need for access to more data sources leading to better coverage for monitoring, a conflict of interests is inevitable. As outlined by contributions from Grout (2014; 2019) and Edelman (2019) in the introduction, this conflict is likely here to stay. For example, one could argue that making such vast amounts of valuable data available for SDG monitoring purposes should be part of *corporate social responsibilities*. Yet, with unrestricted access, there is always a possibility that someone could exploit the extensive amount of data in each social media post disingenuously.

Continuing this line of argumentation, however, results in the realisation that a restriction of access essentially leads to an accumulation of power: as corporations like Facebook or Instagram (which is also owned by Facebook) are the only ones with access to such data, and the people using their services may never know what happens to the posts they upload (Weber, 2012).

Along with access to data sources, the popularity of various social media outlets in different parts of the world is another significant factor connected to map coverage. While Twitter seems to be most popular in Japan, outlets like Instagram or Facebook might be more popular in other locations, again highlighting the potential of integrating multiple information sources into one analysis scheme.

Apart from social media, other promising sources with the potential to cover a significant part of the world include satellite imagery or drone footage, as well as car-mounted sensors (e.g. rear view cameras, dashcams, vibrations recorded by crash recorders, other sensors), as outlined in part by the Transport & ICT Report (2016). Especially the integration of satellite imagery and drone footage with image classification algorithms promises to produce a large amount of useful, spatially comprehensive data - available in near real-time - as the availability of up-to-date cloudless earth observation imagery continues to improve (EOS, 2020).

4.2 Analyzing Big Data: ICARUS' Flaws

From section 3.2 it is apparent that ICARUS' performance - in its current state - is inadequate to generate beneficial information for use with SDG target indicator 9.1.1 (RAI). Apart from an insufficient coverage with Twitter explained in the previous section, its *raw ability* to classify asphalt roads is unsatisfactory. Most likely, a variety of biases contained in the training data set, the shallow DNN-architecture of ICARUS as well as technical difficulties during training, are responsible for its deficiencies. The following explains ICARUS' flaws in more detail, before addressing research questions 1 and 1.1. Finally, general requirements for a successful application of machine learning in big data analyses are examined.

As both the training- and validation data sets used to train ICARUS were sourced from twitter during a two-week time period, and with the knowledge of locally varying tweeting densities, there likely is a sampling bias in both data sets. In other words, it is possible that ICARUS was trained on specifics of all-season roads in places where tweeting density was high during the two-week time period of sourcing training data. To alleviate concerns of sampling biases contained in training sets, strategic decisions during sourcing (like distributed time intervals, sourcing images from other data sources, specifying that images from different regions be included equally as often) need to be taken.

Further performance was likely lost due to the limited hardware available and thus through the choice of DNN-architecture (tiny-YOLO) of ICARUS (see table 4). A more elaborated, *deeper* DNN-architecture (including more hidden layers) like

YOLOv3 would most likely have yielded significantly better results (Redmon and Farhadi, 2018). Using the hardware available not only affected ICARUS' performance, but also meant that a real-time analysis was not possible. Moreover, the time constraints placed on training the classifier, by nature of this study being a master's thesis, may also have had a negative impact on ICARUS' capabilities. As figures 4 and 5 indicate, learning had to be stopped before reaching a loss value below 1 (see also section 2.2). Moreover, pre-training the neural network as done by Deng et al. (2014) as well as Apte et al. (2017) might have improved the algorithm further. Such a pre-training step was not included in this study.

Additionally, the technical difficulty of training an image classifier on amorphous shapes (like parts of roads between the legs of human beings, see fig 8) using classical, rectangular bounding boxes - while avoiding to include objects like cars, bicycles or feet of passing pedestrians - likely also diminished ICARUS' performance. It is possible, that using image segmentation methods to select areas of interest during the training phase, might yield appealing results to mitigate this problem.

As Patterson and Gibson (2017) indicate, though, context is important when evaluating deep neural networks. In the case of ICARUS, it might suffice if it predicts an image containing roads correctly - even though it may not find all instances of roads - i.e., making the the amount of false positives the primary factor to evaluate ICARUS' performance in the context of what it is supposed to accomplish. In other words, for the intended application of ICARUS, false positives are worse than false negatives, as images often contain multiple patches of roads, segmented by other objects (like people standing on the road, cars, etc.). So long as at least one true positive but no false positives are predicted, the result may suffice.

ICARUS was trained to predict ASRs only when encountering roads with an asphalt surface (see section 2.2), although in practice, it often also recognizes concrete surfaced roads (see fig. 8). Clearly this is a compromise, as not all ASRs fulfill this requirement. In very arid regions for example, a simple gravel road may suffice. Defining ASRs as asphalt roads was done because of time constraints, as explained above. For a full implementation of ICARUS for SDG target indicator 9.1.1, different definitions of ASR - e.g. depending on climatic variability - would be needed. Then the algorithm could compare lon/lat coordinates of the image with a map of climate zones and determine which regionally specific requirements for ASRs to use for predicitions.

With ICARUS' flaws outlined, as well as the issues stemming from Twitter as this study's only data source examined, research questions 1 and 1.1 can be addressed. Research question 1 asks *to what extent georeferenced road surface information for indicator 9.1.1 (RAI) of the SDGs can be generated using a deep neural network on* *the Twitter streaming API*. While this study shows that this endeavour is clearly possible, its implementation in the present study is not satisfactory enough to claim that ICARUS can perform this task to the full extent. As explained above, ICARUS' performance likely suffers because of a variety of biases, as well as unfinished training, the latter resulting from time constraints related to the nature of a master's thesis. Moreover, hardware limitations further restrict performance. Still, as DNNs are to be evaluated with respect to their specific context, ICARUS serves as a successful *proof of concept* and offers new insights into the vast unused potential of big data analyses.

Research question 1.1 asks *if data generated by ICARUS is comparable to conventional data for RAI in terms of quality, coverage and accuracy*. As explained in section 1.2.5, the official methodology for measuring road surface uses traditional road inventory surveys. ICARUS is unable to compete with the accuracy or quality of such measurements. However, in terms of coverage it can offer world-wide, real-time information. Yet currently, this information mostly comes from urban areas, only sparsely covering periurban or rural areas, reducing the amount of actually gained knowledge through better coverage - because densely populated urban centres have a higher likelihood of having good access to ASRs when compared to the periphery. Furthermore, in the context of SDG indicator 9.1.1, the focus lies on *rural* population.

Thus, ICARUS does not yet represent a viable alternative to conventional RAI measuring procedures. As most of ICARUS' flaws are a direct result of insufficient hardware, software or time, though, making improvements on the algorithm should be cheap and relatively straightforward. The most influential factor in determining whether or not a system like ICARUS might be feasible in the future will, however, be access to various data sources. In the context of image classification for the monitoring of SDGs, access to streaming APIs from companies like Instagram or Facebook is key. This is further discussed in the following section.

As illustrated by interactively mapping results of ICARUS using a website, transparency is a key factor for effective communication of results from big data analyses - particularly when applying machine learning techniques. While the results still contain a lot of data to process for a single human being, the ability to visualize the basis on which an algorithm makes its decision yields valuable insights into its effectiveness. This effect would be amplified, if prediction bounding boxes were drawn on top of the image being loaded (as in figures 7 and 8), and if prediction confidence could manually be adjusted on-the-fly using a slider. Such features were not possible in this case, as the website was built using D3.js. D3 can read .csv files only if the separator is an actual *comma*. In the case of ICARUS, because Darkflow and yolo output multiple predictions in JSON format, a *semicolon* was used as a separator for the output output.csv file. Through adjusting ICARUS' output structure, however, these functionalities would be possible.

From the above, a set of general requirements for a successful application of machine learning in analyzing big data can be generated. These include: real-time access to multiple data sources with good coverage, fast and well-trained neural networks which can generalize well, transparent and interactive communication (visualization) of results, as well as the hardware to support such endeavors.

4.3 Potentials of vs. Concerns about Big Data in Monitoring the SDGs

Barnes' (2013:299) reservations about big data underline the importance of transparent scientific analyses: "Clearly, big data has information coming out of its ears, but is it generating useful knowledge? Do we now collect data for data's sake? Because it is there. Because we can. [...] My fear is that big data will increasingly produce noise. But because its output comes in mathematical form, and since this is the hallmark of science ('mathematics is nature's language' as Galileo said), it will be touted as knowledge. And all the while the world is going to hell in a handbasket". In fact, this thesis is a good example of his point. Without transparent communication of the methods used (as well as their remaining flaws), it would be easy to misinterpret the results presented in section 3 and thusly come to the conclusion, that ICARUS is ready to be applied in a real-world context.

Allowing others to interpret this study's results for themselves using sensible means of communication acknowledges Barnes' point and ties into what Crampton et al. (2013:132) express by stating, that "big data will not replace thinking": ICARUS is not ready to be applied in a real-world context due to remaining methodological shortcomings. Twitter as its only data source is insufficient for covering international development disparities definitively, as its spatial coverage is limited, and its usage varies regionally. Geotagged images from social media may not depict the places their appended coordinates represent. All-season roads are not unanimously made from asphalt.

With such a level of transparency, this study aims to avoid "extrapolated insights" and "sweeping statements about society as a whole" (Crampton et al., 2013:132). Nonetheless, it offers a valuable sense of what might still be possible in the field of image classification and big data analytics for monitoring the SDGs. Through integration of multiple data sources, an application of the methodology used in this study to other SDG target indicators is conceivable. Target indicators (other than 9.1.1) which might benefit from data acquisition using image classification on big data are shown in table 6. Particularly target indicators 11.1.1, 11.6.1 and 14.1.1 are promising candidates for such endeavours.

As a good starting point, using the existing algorithm by Anadkat et al. (2019) on geotagged images from social media may provide easily accessible, real-time information on plastic pollution of shorelines (SDG indicator 14.1.1), as well as urban solid waste collection via a proxy-estimation of levels of urban littering (SDG indicator 11.6.1). In the case of SDG indicator 11.1.1, where Gadiraju et al. (2018) have applied image classification to remote sensing data, the inclusion of a ground-based perspective from social media content may offer additional information to detect informal settlement areas.

In general, reservations against using big data analyses for a successful monitoring of the SDGs can be alleviated to a large degree by ensuring transparency. This means, first and foremost, making source code available to all. Using services like GitHub in combination with an open license ensures access to source code - allowing everyone to not only review what is analyzed or how analysis is conducted, but also to make a contribution themselves. Making such projects open-source is more trustworthy than sporadic publication of results acquired using unspecified algorithms on unknown data sources. Ensuring transparency also means that the origins of data used in such contexts is declared. Projects which may infringe on civil privacy rights (e.g. projects which focus on personal properties, like classification of indoor sanitation facilities as part of an effort to collect data for SDG indicator 1.4.1) either may need adjustments to these prerequisites in order to be generally accepted, or may require other data collection processes.

4.4 Potentials of Big Data for Geographic Information Science

When Frankenfeld (2005) wrote that analyses of spatial disparities require the setting of spatial boundaries (determining regions of interest), he most likely did not anticipate the possibilities georeferenced big data might bring a decade later. Against the background of the findings from this study, it may be reasonable to suggest, that soon the setting of spatial boundaries may lose its importance in the context of analyzing spatial disparities. Georefereced big data might improve our ability to study some of these disparities to such a degree, that thematic boundaries will become more important than spatial boundaries.

More data with better coverage allows for this shift. The benefit of such an approach is that studies which previously had to be conducted for individual regions could theoretically now be conducted on a global scale, adding comparability and improving our perception of beneficial or adverse effects relative to not only national or sub-national context, but using global references.

Indicators	UNSD Indicator Codes
1.4.1 Proportion of population living in households with access to basic services	C010401
5.b.1 Proportion of individuals who own a mobile telephone, by sex	C050b01
11.1.1 Proportion of urban population living in slums, informal settlements or inadequate housing	C110101
11.2.1 Proportion of population that has convenient access to public transport, by sex, age and persons with disabilities	C110201
11.6.1 Proportion of urban solid waste regularly collected and with adequate final discharge out of total urban solid waste generated, by cities	C110601
11.7.1 Average share of the built-up area of cities that is open space for public use for all, by sex, age and persons with disabilities	C110701
14.1.1 Index of coastal eutrophication and floating plastic debris density	C140101
17.8.1 Proportion of individuals using the Internet	C170801

Table 6: SDG target indicators for which data could be gathered using image classification on big data (United Nations Statistics Division, 2019c).

Apart from improved analysis of spatial disparities, geographic information science stands to gain much more from big data and automated classification. Even if only a fraction of social media contents are geotagged, the amount of information that could be gained by indefinite, automated analysis of spatially comprehensive data is bound to be extraordinary. Furthermore, as the example by Gadiraju et al. (2018) illustrates, merging machine learning and remote sensing produces additional, vast potentials.

Forecasting future specific developments resulting from this union of scientific fields is difficult. It is almost certain, however, that geography and geographic information science will be well advised to embrace these new possibilities which offer fast, efficient and continuously more reliable information.

5 Conclusion & Outlook

This section presents a summary of the main insights produced by this thesis. First, a short summary explains this study's main findings. The second part contains specific summaries for the different key domains in which these findings are applicable. It focuses on ICARUS as an image classifier, big data for sustainability, as well as big data in geography. Finally, possible avenues of further research are explored.

As explained in section 4.2, ICARUS is not yet ready to be applied in a real-world context. From the previous section, avenues of improvement are quite clear, however. The algorithm would have to be trained on better hardware (to allow for a deeper network architecture), using more well fitted bounding boxes to account for polymorphous shapes of road patches in images. Additionally, more advanced training could help performance as well. Pre-training a smaller model similarly to what Deng (2014) and Apte et al. (2017) propose, might yield better results.

Furthermore, the training data set should be carefully compiled to avoid various biases that likely are contained in the training set used for ICARUS, which in turn translate into the quality of the final model. These include a spatial bias (since most Twitter posts including images and coordinates seem to originate in Japan), a temporal bias (since the training set was compiled over a two week period), as well as a general bias stemming from inequalities in internet access across the world. Since ICARUS has a global scope, avoiding these biases is key to producing reliable data.

This study has illustrated some of the challenges left to overcome, if solving data availability problems for sustainable development is to be done using big data. One major challenge is data unreliability, which can result from biases explained above. Further uncertainty can arise because of the difficulty of verifying geoinformation appended to individual pieces of information, which is a concern consistent with findings by Crampton et al. (2013).

Big data analyses are embedded in a debate over privacy concerns and potentials for harmful misconduct. Some of these causes for concern can be alleviated with a high level of transparency and openness, for example by making research projects open-source.

If all of these challenges can be overcome, geolocated big data offer an unprecedented amount of information to be analyzed, boosting data availability for measuring progress towards the Sustainable Development Goals.

With continuously rising internet penetration rates, it is likely that the amount of geolocated data available, as well as the global coverage of this data will increase in the future. Geographers will want to harness this potential, particularly as the internet of things continues to expand. However, all of the challenges mentioned in previous sections particularly apply to geographic use of big data as well.

Further research in these areas should focus on assessing the reliability of geolocated big data with respect to spatial accuracy. Additionally, future big data analyses should incorporate as many different data sources as are available (i.e. from other social media outlets, satellite imagery, IoT sensors, official statistics) in order to maximize the area covered by such analyses.

Acknowledgements

While the process of conducting research itself is incredibly rewarding, I am lucky that on several occasions, when I needed external inputs on my work or maybe a small motivational push, many people were there for me. Most notably PD Dr. Andreas Heinimann, who is a great mentor to me and always took my ambitious goals seriously. I would like to express my deepest gratitude for his support and guidance. Furthermore, I would like to thank everyone who encouraged me along my way, because this thesis would have turned out much less exciting without their assistance.

References

Abuhamad, G., Reif, L. R., Edelman, R. D., O'Neil, C., Zittrain, J., Zuckerman, E., Beales III, J. H., Muris, T. J., Acquisti, A., and Downes, L. (2019). Internet, Big Data, and Algorithms: Threats to Privacy and Freedom or Gateway to a new Future. https://assets.aspeninstitute.org/content/uploads/2019/06/MIT-Conference-Report.pdf.

Albert, J. R., De Dios, C., Martinez Jr, A., Sebastian-Samaniego, I., Miradora, K., Lapuz, J. A., and Martillan, M. (2019). Readiness of National Statistical Systems in Asia and the Pacific for Leveraging Big Data to Monitor the SDGs. ADB Briefs, Asian Development Bank, Manila, Philippines.

Allam, Z. and Dhunny, Z. A. (2019). On big data, artificial intelligence and smart cities. *Cities*, 89:80–91.

Anadkat, A. P., Monisha, B. V., Puthineedi, M., and Patnaik, A. K. (2019). Drone based Solid Waste Detection using Deep Learning & Image Processing. page 8.

Andries, A., Morse, S., Murphy, R., Lynch, J., Woolliams, E., and Fonweban, J. (2019). Translation of Earth observation data into sustainable development indicators: An analytical framework. *Sustainable Development*, 27(3):366–376.

Anejionu, O. C., Thakuriah, P. V., McHugh, A., Sun, Y., McArthur, D., Mason, P., and Walpole, R. (2019). Spatial urban data system: A cloud-enabled big data infrastructure for social and economic urban analytics. *Future Generation Computer Systems*, 98:456–473.

Apte, M., Mangat, S., and Sekhar, P. (2017). YOLO Net on iOS.

Barha, C. K., Nagamatsu, L. S., and Liu-Ambrose, T. (2016). Chapter 4 - Basics of neuroanatomy and neurophysiology. In Aminoff, M. J., Boller, F., and Swaab, D. F., editors, *Handbook of Clinical Neurology*, volume 138 of *Neuroepidemiology*, pages 53–68. Elsevier.

Barnes, T. J. (2013). Big data, little history. *Dialogues in Human Geography*, 3(3):297–302.

Benjelloun, F.-Z., Lahcen, A. A., and Belfkih, S. (2015). An overview of big data opportunities, applications and tools. In *2015 Intelligent Systems and Computer Vision (ISCV)*, pages 1–6. IEEE.

Brecque, C. (2019). Demystifying Hyper-Parameter tuning. https://towardsdatascience.com.

Brundtland, G. H. (1987). Our Common Future—Call for Action*. *Environmental Conservation*, 14(4):291–294.

Burzynski, M., Deuster, C., and Docquier, F. (2019). Geography of skills and global inequality. *Journal of Development Economics*.

Chen, C. and Zhang, C. (2014). Data-intensive applications. *Challenges, techniques and technologies: A survey on big data, Inf. Sci*, 275:314–347.

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., and Zook, M. (2013). Beyond the geotag: Situating 'big data'and leveraging the potential of the geoweb. *Cartography and geographic information science*, 40(2):130–139.

Deng, L. (2014). A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, 3.

Devarajan, S., Miller, M. J., and Swanson, E. V. (2002). *Goals for Development: History, Prospects, and Costs*. The World Bank.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys*, 27(3):326–327.

Du Pisani, J. A. (2006). Sustainable development – historical roots of the concept. *Environmental Sciences*, 3(2):83–96.

EOS, Earth Observing System (2020). Land Viewer — EOS. https://eos.com/landviewer/.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338.

Frankenfeld, P. (2005). *Handwörterbuch der Raumordnung*. Akad. für Raumforschung und Landesplanung, Hannover, 4., neu bearb. aufl edition. OCLC: 76719551.

Gadiraju, K. K., Vatsavai, R. R., Kaza, N., Wibbels, E., and Krishna, A. (2018). Machine Learning Approaches for Slum Detection Using Very High Resolution Satellite Images. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1397–1404, Singapore, Singapore. IEEE.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587.

Grober, U. (2007). Deep roots-a conceptual history of'sustainable development'(Nachhaltigkeit).

Grout, V. (2014). Shazam for People. *J. Robot Autom*, 1:1–4.

Grout, V. (2019). No More Privacy Any More? *Information*, 10(1):19.

Heikkurinen, P., Ruuska, T., Wilén, K., and Ulvila, M. (2019). The Anthropocene exit: Reconciling discursive tensions on the new geological epoch. *Ecological Economics*, 164:106369.

Henderson, P. and Ferrari, V. (2017). End-to-end training of object class detectors for mean average precision. *arXiv:1607.03476 [cs]*. http://arxiv.org/abs/1607.03476.

Hulme, D. (2009). The Millennium Development Goals (MDGs): A short history of the world's biggest promise.

IAEG-SDGs, Inter Agency and Expert Group on SDG Indicators (2017). Tier Classification for Global SDG Indicators. https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/.

ITU, International Telecommunication Union (2017). ICT Indicators. https://idp.nz/Global-Rankings/ITU-Key-ICT-Indicators/6mef-ytg6.

Kharrazi, A., Qin, H., and Zhang, Y. (2016). Urban Big Data and Sustainable Development Goals: Challenges and Opportunities. *Sustainability*, 8(12):1293.

Koehrsen, W. (2018). Hyperparameter Tuning the Random Forest in Python. https://towardsdatascience.com.

MacFeely, S. (2019). The Big (data) Bang: Opportunities and Challenges for Compiling SDG Indicators. *Global Policy*, 10(S1):121–133.

Malthus, T. R. and Bonar, J. (1926). *First Essay on Population, 1798*. Macmillan & Co. Ltd. (1926 edition with notes by James Bonar), London. OCLC: 2710982.

Marsh, G. P. (1965). *Man and Nature: Or, Physical Geography as Modified by Human Action*. Harvard University Press, Cambridge. OCLC: 952754139.

Meadows, D. H. and Club of Rome, editors (1972). *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind*. Universe Books, New York.

Mikalef, P., Pappas, I. O., Krogstie, J., and Giannakos, M. (2018). Big data analytics capabilities: A systematic literature review and research agenda. *Information Systems and e-Business Management*, 16(3):547–578.

Mohammadi, M., Al-Fuqaha, A., Sorour, S., and Guizani, M. (2018). Deep Learning for IoT Big Data and Streaming Analytics: A Survey. *IEEE Communications Surveys & Tutorials*, 20(4):2923–2960.

Mokdad, A. H., Mensah, G. A., Krish, V., Glenn, S. D., Miller-Petrie, M. K., Lopez, A. D., and Murray, C. J. (2019). Global, Regional, National, and Subnational Big Data to Inform Health Equity Research: Perspectives from the Global Burden of Disease Study 2017. *Ethnicity & Disease*, 29(Suppl 1):159–172.

Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., and Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, 30(4):431–448.

Patterson, J. and Gibson, A. (2017). *Deep Learning: A Practitioner's Approach*. O'Reilly, Sebastopol, CA, first edition edition. OCLC: ocn902657832.

Redclift, M. (2005). Sustainable development (1987–2005): An oxymoron comes of age. *Sustainable development*, 13(4):212–227.

Redmon, J. (2019). Darknet: Open Source Neural Networks in C. https://pjreddie.com/darknet/.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *arXiv:1506.02640 [cs]*. http://arxiv.org/abs/1506.02640.

Redmon, J. and Farhadi, A. (2018). YOLOv3: An Incremental Improvement. *arXiv:1804.02767 [cs]*. http://arxiv.org/abs/1804.02767.

Redmon, J. and Farhadi, A. (2019). YOLO: Real-Time Object Detection. https://pjreddie.com/darknet/yolo/.

Ren, S., He, K., Girshick, R., Zhang, X., and Sun, J. (2016). Object detection networks on convolutional feature maps. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1476–1481.

Roberts, P., KC, S., and Rastogi, C. (2006). Rural Access Index: A Key Development Indicator.

Sachs, J. D. (2012). From millennium development goals to sustainable development goals. *The Lancet*, 379(9832):2206–2211.

Schnorr-Bäcker, S. (2016). Politik begleitendes statistisches Monitoring und neue Datenquellen. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 10(2):163–185.

Schwab, G., editor (1990). *Die schönsten Sagen des klassischen Altertums*. Number 500 in Goldmann-Taschenbuch. Goldmann, München, 25. aufl edition. OCLC: 635974581.

SDSN, Sustainable Development Solutions Network (2015). Indicators and a Monitoring Framework for Sustainable Development Goals: Launching a data revolution for the SDGs. Technical report. http://unsdsn.org/resources/publications/indicators/.

SDSN, Sustainable Development Solutions Network (2018). Launching a data revolution for the Sustainable Development Goals. http://indicators.report/indicators/.

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.

Stephen, J., Rohil, H., Vasavi, S., Chaudhary, K., Kumar, N., and Saini, V. K. (2014). *Computer Science, Communication and Instrumentation Devices*. OCLC: 7377987186.

thtrieu (2019). Thtrieu/darkflow. https://github.com/thtrieu/darkflow.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

Transport & ICT (2016). Measuring Rural Access : Using New Technologies. Technical Report 107996, The World Bank. http://documents.worldbank.org/curated/en/367391472117815229/Measuring-rural-access-using-new-technologies.

United Nations (2017). Sustainable Development Goals. Sustainable Development Knowledge Platform. https://sustainabledevelopment.un.org/.

United Nations (2018). The Sustainable Development Agenda. https://www.un.org/sustainabledevelopment/development-agenda/.

United Nations General Assembly (2000). United Nations Millennium Declaration. https://undocs.org/A/RES/55/2.

United Nations Global Working Group on Big Data for Official Statistics (2014). Report of the Global Working Group on Big Data for Official Statistics. https://unstats.un.org/unsd/statcom/doc15/2015-4-BigData-E.pdf.

United Nations Statistics Division (2019a). Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development. https://unstats.un.org/sdgs/indicators/indicators-list/.

United Nations Statistics Division (2019b). IAEG-SDGs — Tier Classification for Global SDG Indicators. https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/.

United Nations Statistics Division (2019c). SDG Indicators — Metadata Repository. https://unstats.un.org/sdgs/metadata/.

von Carlowitz, H.-C. (1732). *Sylvicultura Oeconomica*, volume 1. Bey Johann Friedrich Brauns sel. Erben.

WCED, World Commission on Environment and Development (1987). *Our Common Future*. Oxford University Press, Oxford, New York.

Weber, T. (2012). Facebook buys Instagram for $1bn. *BBC News*. https://www.bbc.com/news/technology-17658264.

Wu, X., Zhu, X., Wu, G.-Q., and Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):97–107.

6 Annex

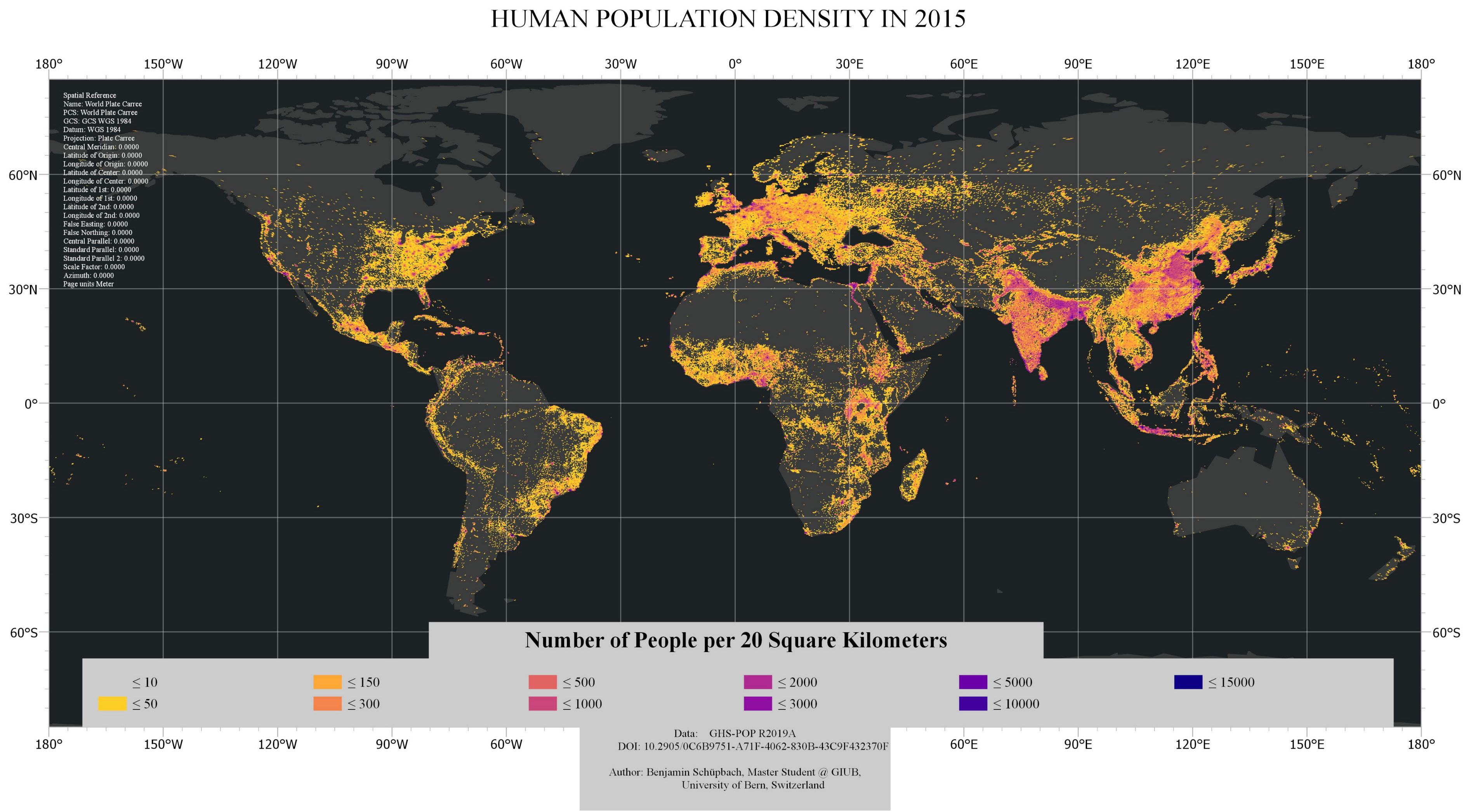

Figure 14: Human Population Density in 2015

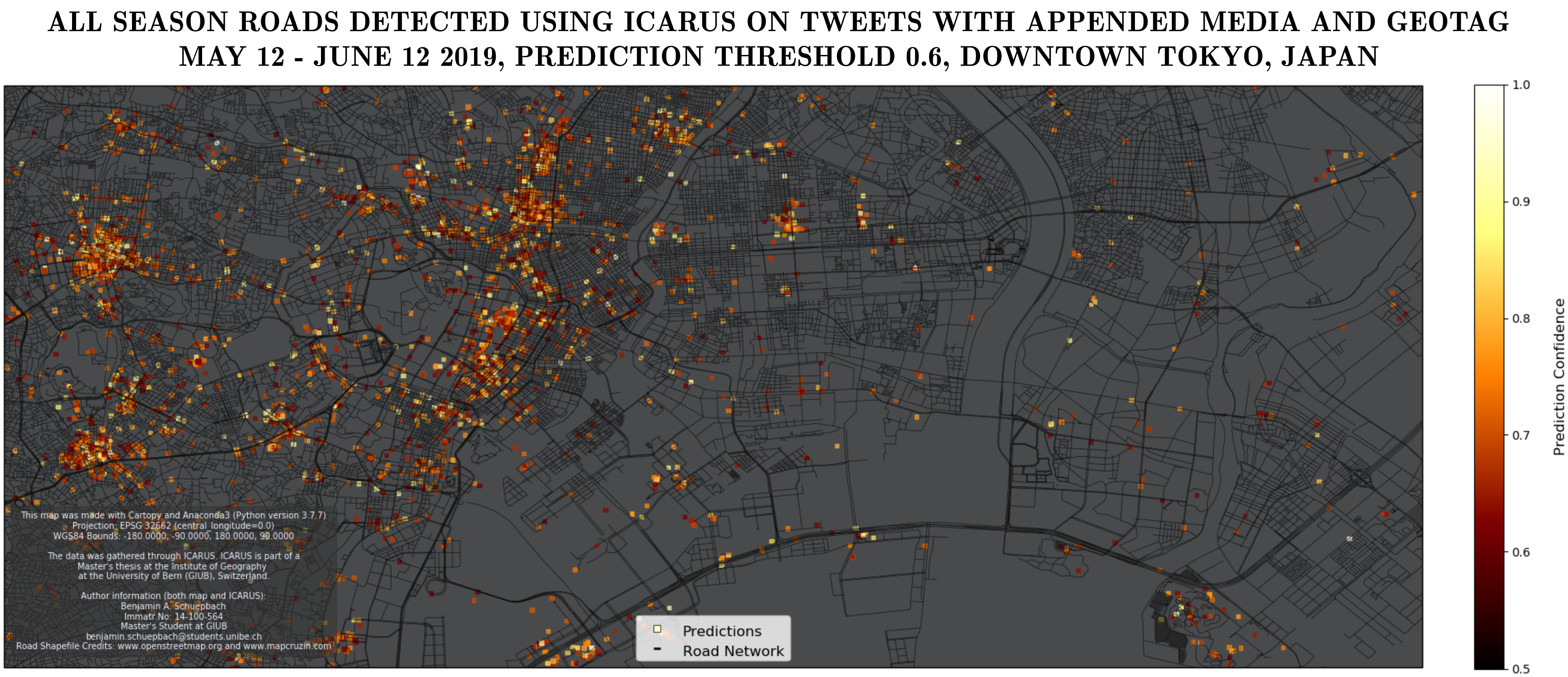

Figure 15: Zooming in on Downtown Tokyo indicates the Spatial Accuracy of Harvests.

ALL SEASON ROADS DETECTED USING ICARUS ON TWEETS WITH APPENDED MEDIA AND GEOTAG
MAY 12 - JUNE 12 2019, PREDICTION THRESHOLD 0.85, DOWNTOWN TOKYO, JAPAN

Figure 16: Zooming in on Downtown Tokyo indicates the Spatial Accuracy of Harvests.

Erklärung

gemäss Art. 30 RSL Phil.-nat. 18

Name/Vorname: Schüpbach Benjamin Andrea

Matrikelnummer: 14-100-564

Studiengang: Geography Master Mono 120 ECTS

Bachelor ☐ Master ☑ Dissertation ☐

Titel der Arbeit: Bigger is Better. Or is it?

Lessons learned from applying a deep neural network to Twitter posts in order to estimate potentials of using big data to monitor the SDGs.

LeiterIn der Arbeit: PD Dr. Andreas Heinimann

Ich erkläre hiermit, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe r des Gesetzes vom 5. September 1996 über die Universität zum Entzug des auf Grund dieser Arbeit verliehenen Titels berechtigt ist.
Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen.

20.01.2020

Ort/Datum

Unterschrift