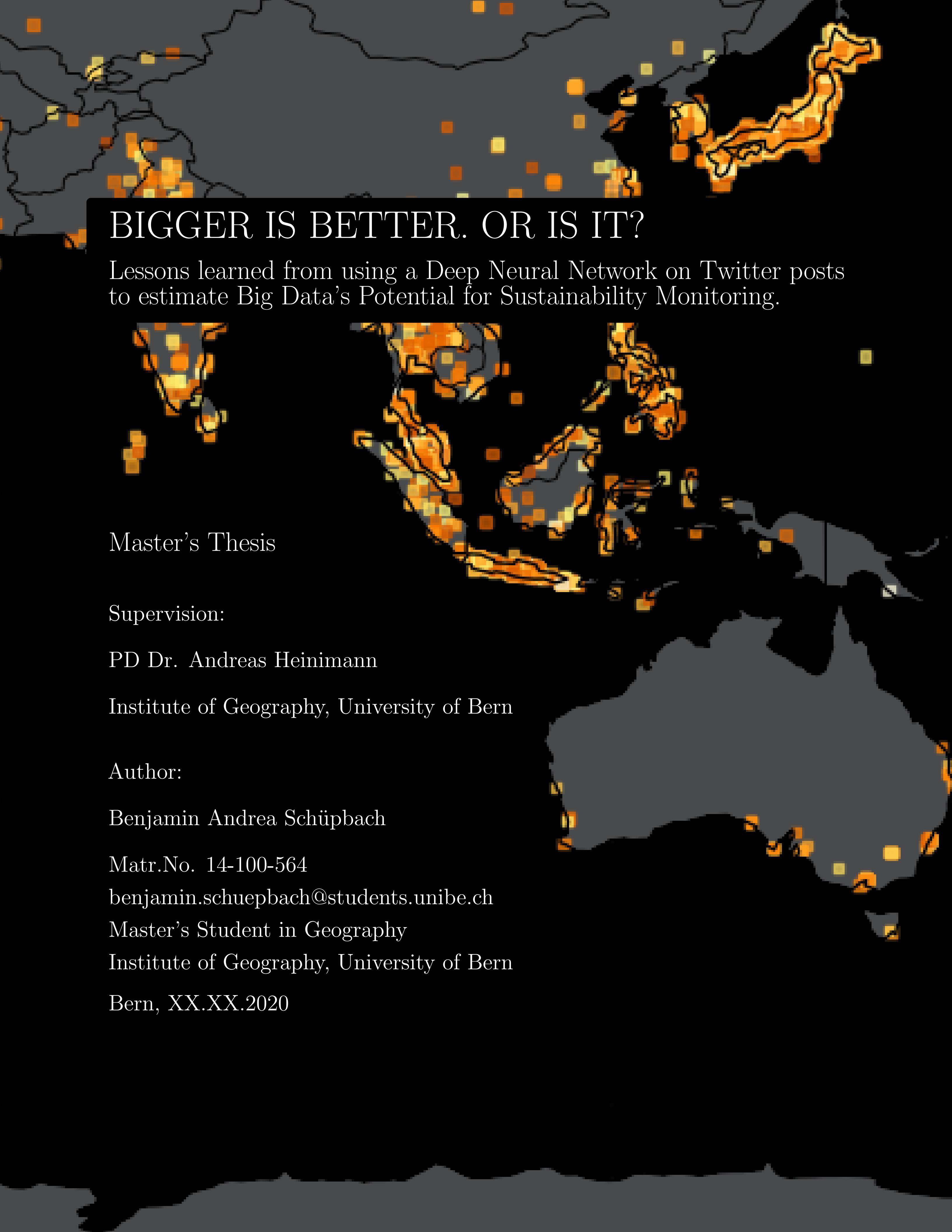

BIGGER IS BETTER. OR IS IT?

Lessons learned from using a Deep Neural Network on Twitter posts to estimate Big Data's Potential for Sustainability Monitoring.

Master's Thesis

Supervision:

PD Dr. Andreas Heinimann

Institute of Geography, University of Bern

Author:

Benjamin Andrea Schüpbach

Matr.No. 14-100-564
benjamin.schuepbach@students.unibe.ch
Master's Student in Geography
Institute of Geography, University of Bern

Bern, XX.XX.2020

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*, Schwab (1990)

Abstract

List of Abbreviations here please

Contents

1 Introduction

With the emergence of the Internet of Things (IoT), the need for real-time analytics of continuously growing datasets arose (Mohammadi et al., 2018).

Explain the relevance and necessity of this study: good hook!

Explain how the introduction is structured, explain the motivation of the study and the dive from big data, through sustainable development to the RAI and back up to big data

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 included also) 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 Potentials of Big Data

Schnorr-Bäcker (2016) sees strong potential for 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 that raise concerns over Big Data's potential to do more harm than good. Trevor Barnes (2013), for instance, sees parallels between potential 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 form 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 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, but elaborates further, 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 the Big Data view, "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. 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 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 names to 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 their 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 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 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.

1.2 Development Disparities

1.3 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 politics by the United Nations are highlighted before various key models are introduced. Finally some of the current measures used to globally advance efforts of SD are presented.

1.3.1 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. Timber was an essential resource at the time, and 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 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".

1.3.2 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).

1.3.3 17 Goals to Transform Our World

Because of these persistent challenges, the UN General Assembly adopted the new 2030 Agenda for Sustainable Developments on 25 September 2015 (United Nations, 2018). The new agenda consists 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 X). In March of 2018 as well as one year later in March of 2019, the list of indicators for the SDGs was expanded to 232 total indicators (UN Statistics Division, 2019a). All indicators are classified in tiers that determine their conceptual clarity and progress towards 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.

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)

1.3.4 The Rural Access Index

The main focus of this thesis is on SDG 9, indicator 1.1: *Proportion of the rural population who live within 2 km of an all-season road* (see chapters X, X and X). 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, eventhough 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 the 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). For the purposes of this study only asphalt roads were considered *all-season roads*.

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 the transport 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, measurements of road utility status) with a final spatial resolution of 100mx100m as opposed to data from household surveys. The 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 this thesis aims to make a contribution (see section X). 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.

WHAT IS RAI GOOD FOR, WHAT DOES IT SAY AND WHY DO WE NEED TO KNOW ABOUT IT?

1.3.5 Challenges in Measuring Road Condition

The following section elaborates on remaining challenges concerning costs, information delay and expenditure of human labour in the Transport & ICT Report. Table X 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 drones).

In terms of information delay, satellite imagery and unmanned aerial drones offer a lot of flexibility. Meanwhile, data from road inventory surveys, call detail records (georeferenced information about calls made/received, owned by cell phone carriers) and 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.

Dobermann and Nelson 2013: Dobermann, A. and Nelson, R. et al. (2013). Solutions for Sustainable Agriculture and Food Systems. Technical report of the Thematic Group on Sustainable Agriculture and Food Systems. Paris, France and New York, USA: SDSN.

1.4 Image Classification

1.4.1 Deep Neural Networks

Good paper: Mohammadi et al 2018

Data Source	Advantage	Disadvantage
Road Inventory Survey	Technically solid, consistent with government responsibility	Costsly, 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).

1.4.2 Training Image Classifiers: Optimization Problems

1.4.3 Validating Image Classifiers: Mean Average Precision

1.4.4 YOLO & Darkflow

Mention modification to darkflow to log loss at each step directly as output!

1.5 Goals of this Study

Show potentials of big data in combination with machine learning for Geographic Information Science on the example of SDG Indicator 9.1.1.

1.5.1 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 X). Research question 2 is oriented towards the potential overall contribution of Big Data for Sustainability.

Research Question 1: Can georeferenced data for indicator 9.1.1 (RAI) of the SDGs be generated using a Deep Neural Network on the Twitter Streaming API?

(If answer to RQ1 is yes:) **Research Question 1.1:** Are these data comparable to conventional data for indicator #58 of the SDGs in terms of quality and accuracy?

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

2 Methods

This section gives an overview of the methods used to conduct this study. It is structured chronologically in the sense of data processing steps (see also figure X). First, the data source is introduced. Second, the process of harvesting the images used to train the *Image Recognition Algorithm for Road Utility Status* (ICARUS) is explained, as well as the methods used to train the algorithm. In the third subsection, validation procedures used in this process are shown. 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 and documented on 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 longitude/latitude format. It is estimated that the total amount of tweets including lat/lon geotags adds up to around 0.1% (Crampton et al., 2013). A second filter was implemented to additionally filter the stream for Tweets which also had 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 May 12 and June 12 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 (Redmon and Farhadi, 2018). The decision to use Darkflow-YOLO depended on the necessity to use YOLO with Python.

To generate a set of training images, the Twitter streaming API was combined with the Google Cloud Vision API. Upon signing up with the

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 3: Hardware Specifications of
the Computer used to train ICARUS.

Vision API, users are granted a free credit with which 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 image sets were classified by manually drawing bounding boxes around areas in the images containing asphalt roads. As mentioned in the introduction, only asphalt roads were considered *all-season roads* (see section 1.3.4). This was due to time constraints associated with a master's thesis. In total, more than 18 000 labels were generated.

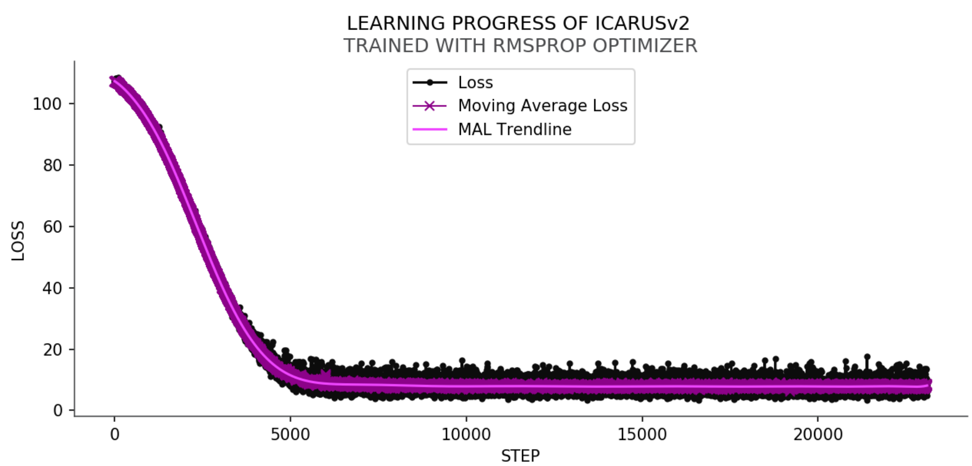

Figure 1: Learning Progress of ICARUS (Steps 0 - 20 000).

Because of hardware limitations (see table X) the DNN architecture of tiny-YOLO was chosen for ICARUS (see also section introduction YOLO).

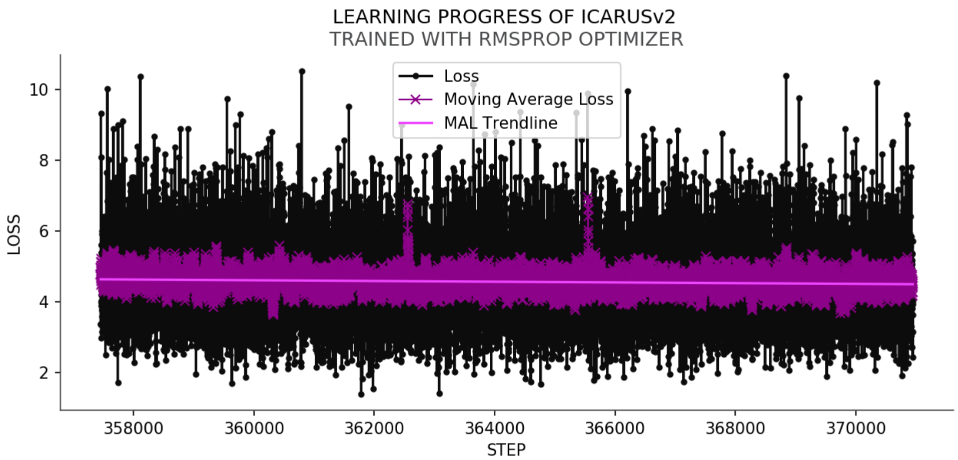

Figure 2: Learning Progress of ICARUS (Steps 358 000 - 370 000).

With darkflow, ICARUS was trained using a RMSPROP optimizer. Through adjustments of the learning rate whenever the step loss reached a plateau, a final loss of around 4-5 was reached (see fig. X and X). It should be noted that because of aforementioned hardware limitations, the maximum batch size possible was 16. In order to log values for loss and moving average loss, the darkflow/net script was modified slightly. This procedure was done multiple times to test different optimizers as well as training parameters, hence the versioning (ICARUSv2) in figures X and X. For all of this thesis, the acronym ICARUS references this ICARUSv2 version.

2.3 Validation

As a preliminary 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 for a sparsely populated area south of the river *Allenbach* in the municipality of Adelboden in 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 by no means 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 ran every time ICARUS produced a checkpoint in training (see section X, introduction). 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 an acceptable assessment of the quality of ICARUS, mean average precision was calculated (see section X, introduction). 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.

2.4 Running ICARUS

This section is about the process of actually running ICARUS. EXPLAIN FLOWCHART HERE! This was done for input data of one month, while data for X months was gathered.

2.4.1 Gathering Actual Data for Assessment

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 X, the processes of gathering data and assessing data were split up. As a cost and resource efficient solution, a Raspberry Pi 3 (RasPi) was deployed to filter the Twitter streaming API. It has a much lower energy consumption than a regular Desktop PC.

With the RasPi deployed, the API could be scanned around the clock. Through the *Remote Desktop Connection* application, data gathering was monitored. Sporadically, data harvesting was manually suspended for a few

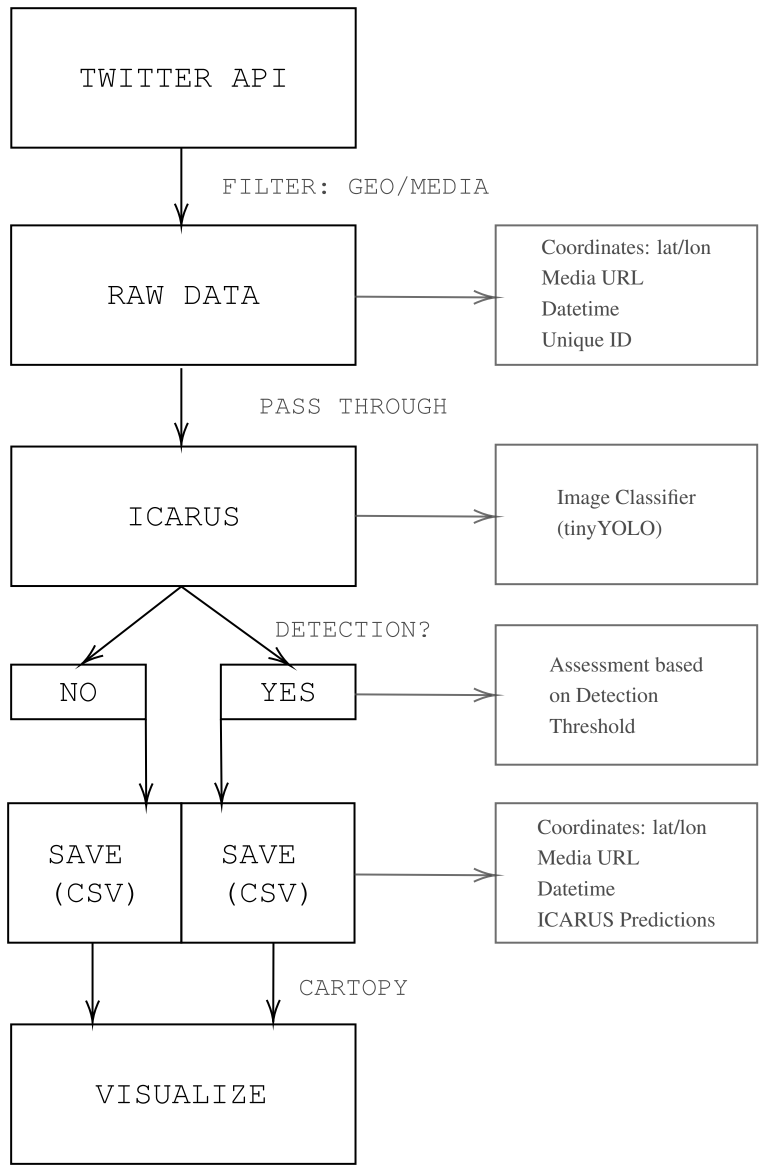

Figure 3: Flowchart of Data Processing Steps.

seconds which meant the RasPi had to generate a new savefile to write into. This was done to avoid total data loss in case severe complications occurred and a savefile got corrupted. For convenient further use, the split 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 classifying 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 step (each time a new Tweet is assessed), 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. This process is illustrated in figure X.

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 such fashion takes 4 days and 6 hours on the Desktop PC used in this study. It was therefore decided that one month worth of harvests (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.

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. 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 Harvested Data

To visualize all harvested tweets, two approaches were used. First, all harvested tweets with geotag and appended media were drawn onto a worldmap

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

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

Therefore, the second 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 X.

2.5.2 Visualizing Predictions

As mentioned above, ICARUS saves a dictionary of predictions (including confidence and bounding box information) for each assessed data point. Additionally, a value for mean prediction confidence is calculated for each data point. Using bounding box coordinates, exact parts of an image that a

Figure 4: Visualization of Predictions with ICARUS.

classification is based on can be visualized with their respective prediction confidence (see fig.X & X). 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 environment as mentioned in section X.X.X. To color-map prediction confidence on the world maps, the aforementioned mean prediction confidence was used. As explained in section X.X.X, the total amount of predictions form ICARUS (based on 1 month of harvests) did not warrant a similar procedure to calculate kernel density as used for harvest mapping. However, if ICARUS were integrated directly into the harvesting process, such a calculation would be possible and reasonable.

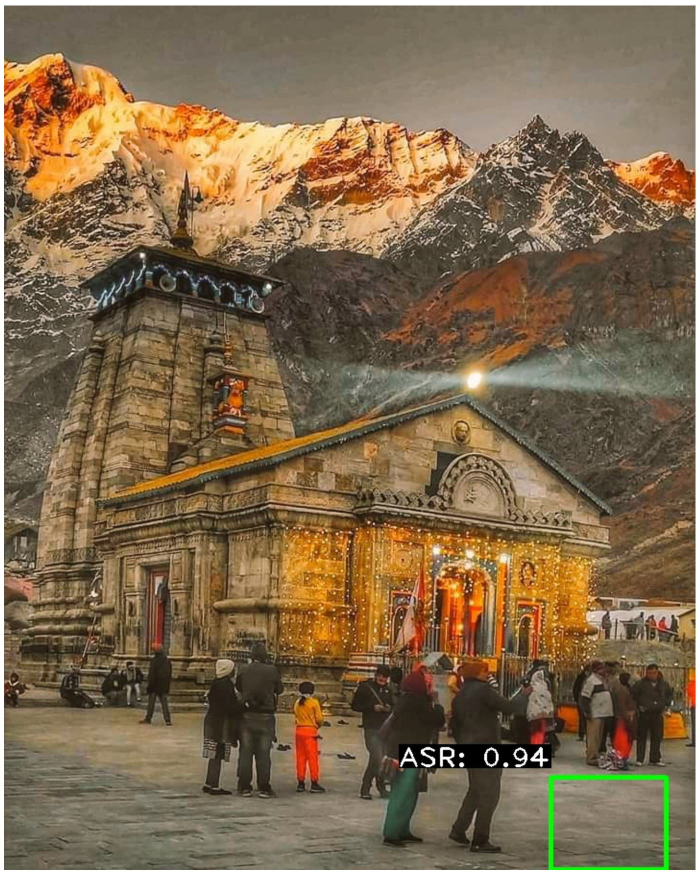

Figure 5: Visualization of Predictions with ICARUS.

3 Results

3.1 ICARUS as an Image Classifier

Visualizing bounding boxes of predictions yields additional insights into strengths and weaknesses of the algorithm. Upon closer inspection of figures X and X, it becomes apparent that in figure X ICARUS classified all areas containing an all-season road correctly, while in figure X it only classified a small part. Yet in the second example it returns a higher prediction confidence.

3.2 Twitter as a Datasource

3.3 1 Month Analysis

4 Discussion

As mentioned in section X.x.x, ICARUS was trained on the tiny-YOLO DNN architecture. It was developed on on a system containing limited hardware (see table x). Also, time constraints (maybe training it for longer could have yielded better results, but that was not possible within the limits of this study).

Discuss the technical difficulty of classifying things like roads (no real recurrent shapes, difficult to manually classify using rectangles.

Discuss optimization possibilities (like having the RasPi make Savefiles for each day, then once per day letting a server run ICARUS on these)

Teaching another algorithm regionally, climatically dependent differences in "all-season road" would be doable, then pass it through desertICARUS if lat/lon values are from a desert, for instance.

For the purposes of this study only asphalt roads were considered *all-season roads*. Of course this is a compromise, as in different climates one would expect other kinds of roads to be motorable all year round. In a desert setting, a simple gravel road might meet this condition sufficiently, for instance. This level of detail was not possible in the context of this thesis, for going beyond asphalt roads would require multiple, specifically tuned instances of ICARUS.

Privacy and concerns over harmfulness of Big Data.

Coming back to Trevor Barnes (2013:299): "Clearly, big data has information coming out of its ears, but is it generating useful knowledge?"

- argument of disconnecting data from what is important and praising techniques and numbers for what they are rather than what they do

image classification is not just incredibly cool, it can generate data from data, which then might yield useful information.

Barnes (2013:299) "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".

Argument that this might be phrased a bit strong, but he might have a point nonetheless.

What does my work say about this?

- Argument of losing context when focusing on mathematics based nature of big data information drawing inspiration from human geography: mixed methods approach? - Argument of data reflecting the past while this may be true for aged data which is not very current, in a real-time analysis framework the data might do exactly what it should: reflect in real-time the measurable part of the world, so that in combination with non-measurable context, conclusions can be drawn and progress can be evaluated.

Relating to Jeremy Crampton et al. 2013: discuss the meaning of results of this study and the importance of avoiding to make blanket statements about a society as a whole when dealing with big data from social media. Illustrate this with Results from harvest mapping as points and as densities. Argue that these concerns are less pressing in the case of this study for it does not analyze a social phenomenon or the semantic content of images. However, since the use of Twitter is linked with only semi specific subset of the population, any analysis of its data yields information based on their acting and can never be seen as a representation of reality in its entirety.

Relating to the second point of Crampton et al. of ignoring crucial reservations about the reliability of big data from social media: further studies would need to assess the percentage of accurate representations of places through geotagged information from images produced through social media. If this percentage turns out to be high enough, one could argue that the sheer quantity of information would compensate for some errors within the data.

Relating to the Transport & ICT Report and the definition of Big Data in the introduction: Integration of different sources like free apps for road assessment could offer additional coverage, where social media cannot.

Discussion of privacy concerns with ICARUS.

5 Conclusion & Outlook

5.1 Improving ICARUS

train on better hardware, use image segmentation, then choose specific segments as training inputs instead of rectangular bounding boxes

include other data sources, if possible: - Other Social Media Platforms generating Images - Records from Open Source Apps, Phone Calls suggested by Roberts in RAI measurements - Satellite Imagery - Other Proxies for Rural Access (Semantic Analysis of Social Media for instance, Number of Car owners, Number of Car dealerships & repair workshops, same for other motorized vehicles, nearby bus stops, etc...) - IoT: Car Dashcams, front facing cameras etc. could easily provide much more data. In order to mitigate privacy concerns, data could be pre-processed in these devices themselves and uploaded monthly. This would prevent, to some extent, a reverse-engineering of driven routes and problems which would arise through real-time georeferenced data upload.

5.2 Big Data for Sustainability

Real-time analysis can have large impacts on efforts of sustainability, especially when combining multiple proxies for different indicators.

5.3 Big Data in Geography

ADD { } TO BIBLIOGRAPHY ENTRIES THAT AREN'T DISPLAYED CORRECTLY IN THE .BIB FILE

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. https://www.adb.org/publications/national-statistical-systems-big-data-sdgs.

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

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.

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

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.

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

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.

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

Division, U. N. S. (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/.

Division, U. N. S. (2019b). IAEG-SDGs — Tier Classification for Global SDG Indicators. https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/.

Division, U. N. S. (2019c). SDG Indicators — Metadata Repository. https://unstats.un.org/sdgs/metadata/.

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

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.

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

ICT, T. . (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.

ITU, I. T. U. (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.

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.

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.

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

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

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, S. D. S. N. (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/.

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

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.

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

WCED, W. C. o. E. a. D. (1987). *Our Common Future*. Oxford University Press, Oxford, New York.

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.