Computational Challenges for Artificial Intelligence and Machine Learning in Environmental Research

Martin Werner[1] Gabriel Dax[2] Moritz Laass[3]

Abstract: In the last decades, environmental research has started to adopt a data-driven perspective enabled by huge sensor networks, satellite-based Earth observation, and almost ubiquitous Internet access. Some of these data-driven approaches are expected to make visions of a sustainable future come true. For example, by enabling societies to live in sustainable smart cities, or to feed the world with precision agriculture. Or by fighting environmental pollution or global deforestation with increased observational power. However, there is a serious gap between some of the current expectations put into data-driven techniques and the maturity of the field of spatial machine learning and artificial intelligence or computer science in general. We give a few examples of open research issues that computer science has to solve in order to make data-driven approaches to environmental sciences successful.

Keywords: Environmental Sciences; Computer Sciences

1 Introduction

We live in a world that is in severe danger due to issues with the environment. The human population has been abusing the planet for a very long time, taking more than giving back to the ecosystem. But in the last decades, our understanding of the impact that humankind has on the planet has increased and we are on the way trying to mitigate some of these effects. For example, the world has been able to agree on certain goals and time frames for concrete reductions in the amount of CO2 emission. In a more general way, we see a trend that the broad public starts taking care of environmental questions more than ever before. Not only international movements like “Fridays For Future” or small countries proceedings against industry countries, as well global political institutions like the United Nations have clearly identified the problem of sustainable development.

The United Nations Sustainable Development Goals identify 17 areas depicted in Figure 1 in which action needs to be taken for a “shared blueprint for peace and prosperity for people and the planet, now and into the future.” [Na].

[1] Technical University of Munich, Professorship of Big Geospatial Data Management, TUM Faculty Aerospace and Geodesy, Ottobrunn, Germany, martin.werner@tum.de

[2] Technical University of Munich, Professorship of Big Geospatial Data Management, TUM Faculty Aerospace and Geodesy, Ottobrunn, Germany, gabriel.dax@tum.de

[3] Technical University of Munich, Professorship of Big Geospatial Data Management, TUM Faculty Aerospace and Geodesy, Ottobrunn, Germany, moritz.laass@tum.de

doi:10.18420/inf2020_95

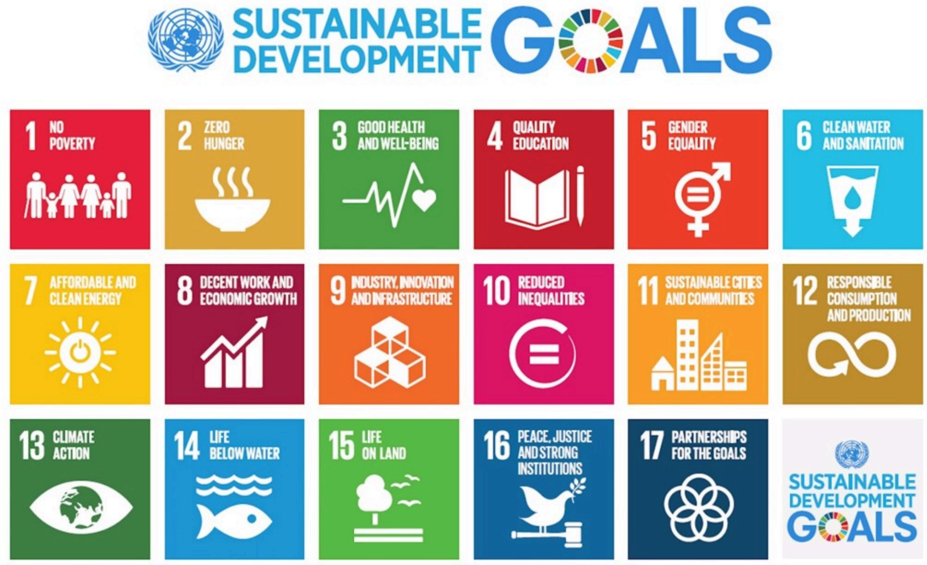
SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES 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 FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

Fig. 1: U.N. Sustainable Development Goals

Many of these United Nations Sustainability Goals are related to data. For example, the goal “Zero Hunger” is of course related to global food production and sharing, which is a complex process whose efficiency is largely depending on data. Similarly, it is expected that sustainable cities and communities will need data-intensive services to solve some issues related to density in urban regions like smart transport or air pollution. As a final example, the goal “Climate Action”, spelled out “Take urgent action to combat climate change and its impacts”, will depend largely on big data provided by Earth observation satellites in order to quantify effects related to climate. In this context, the European Space Agency (ESA) is running research and development in the climate change initiative[4] worth exploring. It currently contains global data products related to important climate variables such as aerosols, biomass, clouds, fire, glaciers, ice shields, land cover, land surface temperature, sea level, sea ice, soil moisture and much more. Without going into details, one realizes that these global big data products are important to the environment and should be exploited as much as possible in data-intensive systems. This links environmental research to computer science, statistics, and machine learning or more general to data science.

For this paper, data science is the culmination of knowledge from three important aspects: mathematics and statistics, computer science, and domain knowledge [WF17]. In this paper, we want to discuss some computer science issues that must be addressed in the context of environmental research for an overall increasing ability to understand, control, manage, and solve issues related to the environment.

[4] see http://cci.esa.int/

2 Data Acquisition, Representation & Compression

In order to come up with an overview of how computer science contributes to environmental research, we first discuss two different types of data because they pose very different challenges to the computer science community. The first type captures all types of measured or scientific data, for example, collected by professional controlled surveying methods with reliable and high-quality sensors. The second type of data subsumes all sorts of user-generated data including volunteered geographic information (VGI) as well as social media or web sources. We will shortly characterize these two types of data in this section.

2.1 Measured Data and Scientific Data

Due to the development of cheap sensors and computing, all domains of engineering generate increasing amounts of data. And in addition, the domains of engineering that design and build sensors are increasing both accuracy and quantity of information gathered from sensors every year. One domain very relevant to environmental research is the domain of remote sensing, especially from space. In remote sensing from space, satellites sense information about the Earth and this information can be used to quantitatively measure on the Earth surface. One very important type of satellites employs multispectral cameras to capture images of the Earth surface. These satellites are conceptually similar to RGB cameras, except that they can capture more different spectral bands (e.g., colors) and that advanced techniques are needed to georeference images taken from space and to mitigate atmospheric distortions including cloud cover and cloud shadows. These satellites typically orbit around Earth with slowly rotating orbits and are, therefore, capable of taking images everywhere on the planet from time to time. Similarly, synthetic aperture radar (SAR) satellites are routinely used in Earth observation. They have advantages such as that they are not affected by clouds and generate more continuous observations and they are capable of surprising measurements using a technique called interferometry in which changes in small scale on the Earth surface can be observed. In both cases, engineers have made sure and are continuously working on high-quality data processing including proper estimation and calibration of all system parameters. The resulting data is huge, but highly accurate.

Another important domain generating measured and scientific data is the area of autonomous driving currently under development. In this area, cars are employed with high-quality, certified sensors to allow vehicles to navigate unknown environments. These sensors include cameras, radar systems, and laser scanners. Together, these systems are assumed to enable safe vehicle operation in unknown environments. However, the amount of data that current development systems capture is extremely huge and complicated as well.

A third domain generating huge amounts of data are our communities and businesses. For example, cities are routinely measuring traffic information, bus locations, taxi data, electricity information, ranging up to large installations of video surveillance in some cases.

In addition, mobile networks are capturing presence and mobility information of mobile devices almost everywhere in our complex cities.

In general, these data sources are well-designed, well-understood and pose questions mainly related to ethics, data size, and data ownership.

3 Personal, Human-generated and Societal Data

Complementary to such measured sources of big data for environmental science are sources originating more in human and societal contexts. This includes news streams, social media messages, human-curated knowledge such as OpenStreetMap and Wikipedia, opinionated data sources such as blog posts from certain platforms, or blind web scale data collections such as common crawl. This type of data has significant limitations with respect to its quality measured in aspects such as bias, risk of abusive publication of information (fake news, etc.), data quality, and data completeness. Though the Internet and social media act as a mirror of aspects of our societies due to discussions taking place on major platforms such as Facebook, Twitter and major news platforms, one is tempted to forget that the Internet does only represent a small fraction of our society and that people behave and communicate differently from real life.

4 The Evolution of Big Data for Environmental Research

With these preparations on two different types of data relevant to environmental research, we focus on computer science aspects in this chapter and first match the difference between these two types of data with concepts of big data. Big data is a family of techniques evolving in order to solve the challenges posed by current data collections on computational infrastructures. One widely accepted definition of big data relies on three aspects Volume, Variety, and Velocity (3V). Though there are many additional Vs such as veracity, value or vagueness being used (even in standards), the advantage of the 3V definition is that all three aspects can be quantitatively measured to some extent. With volume, one refers to the sheer amount of data. Big data is a situation in which traditional computational techniques (relational database management systems, file systems, personal computers) fail due to the amount of data being too large. With velocity, one refers to the situation in which data arrives or needs to be moved faster than feasible in a single computer. Volume and velocity are tightly interwoven as processing high-volume data sets in distributed computing leads to high volumes of data being communicated in short time, that is high velocity. And vice versa, storing high velocity data leads to high volume collections. The third aspect of big data, variety, is represented by the high number of ways, data can be organized, modeled, and stored. This includes aspects such as file formats or data representation as well as organizational aspects such as data ordering and data distribution. Some general aspects of big data are that pre-processing is usually considered impossible or ineffective. That is, most big data systems deal with “raw” data in their “natural” ordering and format.

In environmental research, big data issues are currently limiting the exploitation of data by scientists. For example, even the open access satellite data published by ESA (e.g., Sentinel satellites) or U.S.G.S. (e.g., Landsat satellites) is difficult to acquire, store and use in practice due to the size of the data. This practically limits very large (e.g., global analysis) to well-funded research agencies and companies.

Similarly, the data generated by mobile sensors in the public (cameras, cars, etc.) is very difficult to share in a research community due to legal issues related to privacy. In practice, this means that mainly car manufacturing companies or large research institutes can do their own data acquisitions to base their research on. There are some ongoing activities to share benchmark data and lately even car manufacturers support research with publishing data, but the availability of data remains limited. Due to recent abuse such as during democratic elections all over the world, social media APIs are limited. That is, the whole source of data related to social media is difficult to access without cooperating with social network companies. And there are significant ethical and legal challenges to keep in mind when working with or on social media data. Because even if you think that a given data collection is ethically sound for a certain research purpose, one must make sure that the data collection is not being used in another way now or in the future. In addition, one must think of mechanisms to curate such datasets with updated information on the user's intent. For example, when observing a social media message on one day, one should not use it anymore as soon as the original user has deleted the post. In fact, one should delete the message as well.

5 Open Computer Science Challenges

In the context of challenges related to Big Data, there are several computer science aspects that need further research in order to increase the adoption and reduce the barriers inherent to the amount of data.

The first domain of research is **compression**. Currently, the most valuable data for environmental sciences is generated by high-quality sensors and the data distribution mechanisms do their best not to compromise any of this quality. For example, satellite images are published with lossless compression, GPS data is published in non-simplified form, car sensor information logs are published without any form of processing in order to not sacrifice value. However, the authors argue that compromising volume and data quality is essential to sustainable application and open science. Research is needed on how to compress data such that (1) the amount of researchers that are financially able to study data at large scale is increased and (2) that the application of computing to the data is not wasting too much electrical energy for results that would have been possible from highly-compressed data extracts as well.

Data compression is coupled with **information theory** and information theory can be used as a tool to integrate compression and algorithms and there is some research already going on. One starting point is possibly the area of compression distance and feature-free data mining [CV05] or some work on spatial data representation with probabilistic sketches [We15; We19].

Another direction of computer science research for environmental research and beyond is to research **efficient algorithms**. By reducing the computational complexity of algorithms, one decreases the overall need for computational capacity and can consume larger amounts of data in shorter time either reducing the amount of parallelization or increasing on the limits of a given infrastructure.

Furthermore, a subdomain of database research is research on **reasoning and managing data under uncertainty**. Uncertain data is very common in big data and it is a pity that though theoretical models of uncertain data management have been researched for a long time, the amount of practicable techniques from this field is small. More research needs to be done with respect to handling uncertainty in a scalable way.

As many of the current environmental challenges are rooted in human behavior, it is essential that we are observing human activities. However, collecting precise data about human behavior limits the freedom of individuals and is – for good reasons – not allowed in democratic countries. One has to make sure that the individual rights of people being observed stay protected as much as possible and that the environmental research application's societal value is higher than the privacy loss of individuals. It is a very difficult ethical area and one contribution from computer science is to design **privacy-preserving algorithms**. These are distributed multi-party algorithms in which the data contribution of individuals (e.g., people, sensors, etc.) is protected while sufficient aggregate results or results of computation is still possible. One widely accepted framework is the framework of differential privacy in which data is perturbed when it is produced (e.g., at a sensor) in a way that is statistically cancelling in aggregates.

It is clear that we will be using more and more measured data in decision-making processes as a consequence of "digitization". But we need to be very careful as it is unclear how resilient against fake data such decision-making processes are. This opens a research area of **resilient distributed algorithms** in which algorithms are required to work even in case of attackers. This includes topics ranging from **resilient computer vision to the cyber security** of individual nodes and components in a future Internet of Things.

Tightly coupled with resilient distributed algorithms is the domain of **correctness and verification**. How can we organize all data processes in a way such that the behavior of a system adheres to a given specification. In fact, assessing the correctness of distributed algorithms is a very challenging topic and also less strict aspects such as debugging and testing are complicated due to the huge amount of possible situations the system might face at runtime.

6 The Evolution of Machine Learning and AI for Environmental Research

Big data forms the foundation of machine learning and artificial intelligence and complex methodologies from these fields usually consume a lot of data leading to a strong link between big data and modern machine learning. The current state of artificial intelligence leads to extreme expectations in general as it has been solving hard problems in narrow domains quite successfully. However, it is unclear whether these expectations can be fulfilled outside a few narrow domains. Deep learning and modern artificial intelligence has unquestionably had significant impact on computer vision and image recognition, on text understanding and speech recognition, on dialogue systems and in handwriting recognition, and in playing games. It has as well been applied to subdomains of environmental science such as in remote sensing and cartographic applications [Sa20]. However, traditional methods such as decision trees or support vector machines still dominate these fields for the reason that the amount of available training data is very small compared to the variety and complexity of real-world classification tasks in these data domains. Therefore, a lot of research on combating overfitting and reducing the information-theoretic complexity of input data (data simplification, dimensionality reduction) is still needed in order to see significant adoption of deep learning in these complicated real-world tasks.

This is where the workshop on “Künstliche Intelligenz in der Umweltinformatik” (artificial intelligence for environmental sciences) will put its emphasis: with which contributions from the computer science domain can we enable environmental scientists to apply the unquestionable power of deep learning and artificial intelligence in a promising, scientifically rigid, valuable and sustainable way? Personally, the authors fear that the discrepancy between narrow AI and research papers based on narrow AI in fields related to environmental sciences and real-world deployability and generalization capabilities of such models might lead to the next AI depression in which people realize that artificial intelligence is not a magic tool unless you understand and model your data and your tasks rigorously and come up with techniques related to solving the most pressing problems of applying deep learning in the spatial domain.

7 Open Computer Science Challenges

Let us again list some fields of computer science that will play an essential role in the further development of artificial intelligence for environmental sciences. One important aspect holding back the wide adoption of artificial intelligence is the fear that something goes wrong. As long as we cannot assess the decisions made in artificial intelligence, we can never be sure to base decisions and political policies in such results. That is, without **explainable artificial intelligence**, adoption of artificial intelligence in environmental sciences remains an academic exercise. Related to this explainability is the question of dependability, which is also linked to the computer science challenge of verification. A dependable artificial intelligence system is a system where one can safely base decisions in. Though dependability and explainability are interrelated concepts, they differ very much. While explainability just asks for human-understandable reasons for decisions, the decisions might be wrong or might be based on tampering. With dependable AI, one goes one step further that the system makes sure that the data is real data and not the result of tampering with an AI system.

The most difficult and most important area of research for a wide adoption of deep learning in environmental sciences and spatial information sciences in general is, however, more general the question of whether and how spatial processes can be captured in deep learning settings. This amounts to understanding, estimating and taking into account all sorts of **spatial autocorrelation** and also mitigate issues related to structures introduced by space subdivisions such as **gerrymandering**. These are very hard problems where only some initial success is known from spatial statistics and where there are quite a few negative results related to that current machine learning approaches are not able to learn spatial analysis tasks unless they have trivial autocorrelation structure. Given this hardness, this is an exciting area of research. Solving or mitigating some of these issues is really a major step forward for data-driven environmental science and beyond.

But even today, machine learning can be applied to spatial data and in order to reduce the risks of domain scientists in claiming wrong results, we – as computer scientists – should come up with software, techniques and analyses of best practices: how can you use machine learning in a spatial domain? How can you avoid many of the pitfalls of transferring from the narrow domains of image recognition or language analysis to the wide and complicated domain of modelling human behavior and the Earth system? Best practices are needed as well as some textbooks and major reference works to spread the word about how to correctly and carefully exploit the advances of machine learning in environmental science. And as the risks of doing something wrong or suboptimal when dealing with spatial data are so high, we call for a culture of reproducible research. We should not anymore accept any papers that just claim a certain achievement without a stringent and open scientific debate based in data and software. And for this, we should concentrate on topics such as compression or privacy-preserving data mining as enablers of open science in computational environmental science.

8 Conclusion

Machine Learning and Artificial Intelligence has shown great advances in certain domains in the last decade. Unfortunately, this leads to expectations in other domains with much more involved or complicated situations. In summary, this paper tried to remark that many of these complications are related to computer science problems and that significant contributions of computer scientists are needed aside domain expert knowledge and engineering in order to allow a responsible and sustainable exploitation of machine learning and artificial intelligence for environmental sciences.

As a conclusion, we see computer science research needs working on **(1) reducing the data footprint, (2) reducing the computational complexity, and (3) reducing abusive or wrong adoption of techniques** leading to claims that are too optimistic or even just wrong. To some extent, the computer science community is responsible for coming up with solutions and guidance in these three directions as spelled out in more detail in the chapters.

We are convinced, that artificial intelligence and machine learning are integral parts of solving the environmental challenges we are facing by understanding and monitoring the changes to the Earth system as well as by allowing more sustainable behavior in large scales through ideas such as smart cities. Therefore, we hope that this first iteration of the workshop “Künstliche Intelligenz in der Umweltinformatik” (artificial intelligence for environmental science) will serve as a focal point of research and development with respect to the computer science contributions needed for environmental sciences and, thereby, for a sustainable future.

References

[CV05] Cilibrasi, R.; Vitányi, P. M.: Clustering by compression. IEEE Transactions on Information theory 51/4, pp. 1523–1545, 2005.

[Na] Nations, U.: U.N. Sustainable Development Goals, URL: https://sustainabledevelopment.un.org/?menu=1300.

[Sa20] Salcedo-Sanz, S.; Ghamisi, P.; Piles, M.; Werner, M.; Cuadra, L.; Moreno-Martínez, A.; Izquierdo-Verdiguier, E.; Muñoz-Marí, J.; Mosavi, A.; Camps-Valls, G.: Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources. Information Fusion 63/, pp. 256–272, 2020.

[We15] Werner, M.: BACR: Set Similarities with Lower Bounds and Application to Spatial Trajectories. In: 23rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL 2015). 2015.

[We19] Werner, M.: GloBiMaps - A Probabilistic Data Structure for In-Memory Processing of Global Raster Datasets. In: 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL ’19). 2019.

[WF17] Werner, M.; Feld, S.: Successful Data Science Is a Communication Challenge. In: DIGITAL MARKETPLACES UNLEASHED. 2017.