Sociotechnical Foundations of GeoAI and Spatial Data Science

Date and Time
October 25, 2024 - arrival day
October 26-27, 2024 - meeting
October 28, 2024 - departure day

Location
Springer Schlössl, Vienna, Austria

Motivation

We are excited to announce a specialist meeting on "Sociotechnical Foundations of GeoAI and Spatial Data Science" which will take place at the Springer Schlössl in Vienna, Austria on October 26-27, 2024. We will offer accommodation and travel support for around 30 participants across career stages, geographic regions, and academic/industry backgrounds. The meeting will provide an opportunity to discuss the sociotechnical and ethical foundations of GeoAI and Spatial Data Science in the context of recent advances in generative AI and foundation models.

With recent breakthroughs in foundation models, such as large language models and text-to-image models in AI and GeoAI, there is an urgent need to develop a community-driven roadmap. This roadmap will help us to positively and actively shape the next five years by providing our (geo)spatial perspective to the broader AI community. Otherwise, we risk being passively shaped by those next five years.

We are already seeing first research on developing geo-foundation models. However, the costs and computational resources required for training, tuning, and deploying these models may exceed what most individual labs (or even universities) can handle. For instance, training large language models or text-to-image models is known to cost hundreds of thousands of dollars. While the science community is currently trying to catch up by closing the performance gap between super-large and smaller models, we believe that future GeoAI and Spatial Data Science research may benefit from a similar approach as implemented by the Physics community. Such an approach involves forming a joint community-wide consensus to inform funding agencies and donors about the future long-term research roadmap, e.g., for missions such as the Kepler space telescope, that benefits the entire community.

The specialist meeting (together with related activities such as online seminars and the "GeoMachina" workshop) aims at identifying and discussing the sociotechnical foundation of GeoAI, Spatial Data Science, and geo-foundation models from an ethical perspective in order to prepare and positively shape the AI-based disruptions ahead of us. Reflecting on our profession's ethical implications will assist us in conducting this potentially disruptive research more responsibly. It will assist us in identifying pitfalls in designing, training, and deploying GeoAI-based systems, and developing a shared understanding of the benefits but also potential dangers of artificial intelligence and machine learning research across academic fields, all while sharing our unique (geo)spatial perspective with others.

To give just one example why such bi-directional exchange is important, it is worth mentioning that currently AI teams from Google Brain, Sony, and others are trying to understand potential coverage and representational biases in their training, validation, testing datasets, and models. They do so by studying their 'geo-diversity' and using terms, e.g., the Modifiable Areal Unit Problem, and technologies that originated in geography and GIScience. Put differently, our skills and methods benefit the broader AI community.

To provide structure to our discussions, we aim at covering the following key topics:

AI Sustainability: Training of an AI system can cause carbon emissions equivalent to hundreds of flights across the US. This does not take into account the cost of adjusting and deploying these systems nor the cradle-to-grave emissions generated through manufacturing, transporting, and recycling the required hardware (which are substantially higher yet). Given that geo-foundation models may need substantially more frequent retraining, our community should progress using GreenAI methods instead of RedAI, where progress is essentially bought through research consumption thereby excluding most competition. Additionally, the metrics used to measure how environmentally friendly current machine learning systems are, rely on (over)simplistic models of space and geography, e.g., by ignoring the population affected by negative environmental impacts in relation to the benefiting population. Put differently, it needs geospatial analysis methods to properly quantify how sustainable current progress is as well as a better understanding of ownership and governance structures. Interestingly, there are first signs that foundation models don't necessarily have to be very large (and, thus, resource intensive) to provide good results if the underlying architectures are improved.

Bias and Debiasing: Are training datasets, pre-processing, neural architectures, evaluation criteria, prompt engineers, and users (feeding back into the system) biased? What types of bias are specific to (geo)spatial data and models? How do researchers and practitioners in the broader AI community think about geo-diversity, and can we contribute new perspectives? What types of biases affect geographic data, e.g., VGI, and how can we mitigate them? Given that debiasing will be done algorithmically, how biased will debiasing be? These are just some of the questions that the GeoAI (and broader AI) community is currently facing, and that cannot be resolved just by the technical community alone. For instance, debiasing on the data level may lead to models that more accurately reflect social aspirations at the cost of masking realities expressed by the original data, which do not (yet) reflect these social and political aspirations. If training data sources for text-to-image models contain only 3.1% of images from China and India combined, is this reflected in the way foundation models represent geographic space? Can we as geographers and spatial data scientists contribute measures of geo-diversity back to the global AI community?

Schema and Data Diversity: Foundation models rest on the assumption that pre-trained models of sufficient size can be used across domains and downstream tasks. However, this may neglect regional variability and lead to less accurate results overall. It is important that models are trained on a diverse set of datasets across several data types (modes) and that diversity also includes variability in the schema knowledge underlying these datasets. So far, data diversity is purely approached from a perspective of representativeness, e.g., of a given data collection. Local/regional differences in schema knowledge are not broadly taken into account despite their importance, e.g., due to varying laws, being widely recognized. Given that most AI chatbots are now utilizing retrieval-augmented generation (RAG) to connect to knowledge graphs and other data sources to retrieve data instead of dreaming them up, how are these data sources prioritized? Where does their schema knowledge come from? How would we provide data for a GeoRAG?

GeoAI Neutrality: Given that geo-foundation models will impact how we learn about the world, and, in a second step, also how we act in the world, it is crucial to understand whether GeoAI methods are neutral, and if they are not, at which stages, e,.g., data curation, unbalanced results are introduced. The current lack of consensus within our own community about what algorithmic neutrality means and implies are posing substantial challenges to our ability to positively shape the disruptions of (Geo)AI that society will likely face over the next five years. In most cases, lack of neutrality arises from data curation, during prompt engineering, by selecting certain data and not others, by inadequately matching the training task to future downstream applications, but also due to issues of ownership and governance, and so on. Can we develop clear definitions of GeoAI neutrality and guidelines to achieve it?

Disruption: Ahead of us lie disruptions that make the invention of the Internet pale in comparison. We must shape these next years actively and positively instead of being shaped by them. Our community has a lot to offer to the broader AI community; however, the costs of contributing to the current state of the art, e.g., geo-foundation models, are very high, and the required hardware, storage, and deployment costs cannot easily be handled by single research groups and often not even by universities alone. Hence, it is important that we as a community jointly form a research agenda similar to how this has been done in (Astro)physics for decades. Agreeing on such a community goals—driven research agenda and approaching funding agencies with such proposals requires a clear understanding of the benefits and risks of developing and deploying the geo-foundation models of the near future. If we develop a joint and informed consensus, the benefits of current AI developments will far outweigh the drawbacks and potential risks. Communicating this optimism while informing about risks ahead is also key to educating the future Spatial Data Science workforce.

Format of the Specialist Meeting

The workshop will be held over two full days October 26-27, 2024 with arrival and departure days before and after. For the arrival day, we will also provide opportunities to jointly explore Vienna. We will keep the tradition of offering a morning hike alive.

We invite colleagues from all disciplinary backgrounds, career stages, geographic regions, genders, and ethnicities to apply. We kindly request all potential applicants to fill out the linked form [https://forms.gle/mauHoPHweJxAjoBF6], including a brief biography (up to 200 words), and a one-pager (400-600 words) detailing their motivation to participate in the meeting. The one-pagers and biographies will be published on the meeting's webpage and inform the discussion. The deadline for applications is **July 17, 2024.** We aim to provide accommodation and travel support for around 25-30 external attendees and, therefore, about 35 participants overall. While the meeting will focus on discussions, each participant will also have the opportunity to present a lightning talk during the opening session. Other roles will include panelists, keynote speakers, and so forth. All participants will be co-authors on the meeting report. Please also use the opportunity to participate in our additional activities that we will offer before and after the in-person meeting such as our open-access book on “Geography According to ChatGPT”, our “GeoMachina” autonomous GIS analyst workshop at SIGSPATIAL 2024, the Spatial Data Science Symposium SDSS 2024, and other webinars.

Please feel free to reach out to Krzysztof Janowicz (krzysztof.janowicz@univie.ac.at) for further questions and /or Daniela Woelfle (daniela.woelfle@univie.ac.at) for administrative requests, e.g., with respect to the application form.

We gratefully acknowledge support from Esri, AAG, and the University of Vienna.

Krzysztof Janowicz, University of Vienna, John Wertman, Esri, Mike Goodchild, University of California, Santa Barbara, Gary Langham, AAG, and Coline Dony, AAG