

Master's Thesis Supervision Plan

Academic Year 2024-2025

MIA/MPP/MDS

1. Supervisor information

Name	Lynn Kaack
Professor	Professor of Computer Science and Public Policy
Office	3.22
E-Mail	kaack@hertie-school.org
Faculty Assistant	Alex Karras <karras@hertie-school.org>
Office Hours	On demand

2. Supervision Profile

General Academic Field: *How would you summarise your academic background?*

My work is interdisciplinary, and involves the academic fields of policy analysis, data science, and machine learning. I will supervise theses that focus on applications of data science and ML, more theoretical machine learning method development, method development in policy analysis, as well as public policy more generally (if it is related to my research interests). Refer to my [website](#) and [Google Scholar](#) for past work.

Main Research Areas: *On what research areas are you especially keen to supervise students?*

My group is currently working on the following research topics:

- The greenhouse gas emissions impacts of machine learning
- Machine learning approaches for improving bike infrastructure
- Data-driven approaches for assessing charging infrastructure needs
- Machine learning methods for public policy
- Municipal governance models of urban AI for sustainability
- Social and environmental impacts of mineral deep-sea mining

I am particularly interested in developing machine learning-based approaches to help design climate policy. I also supervise theses on the policy and governance questions at the intersection of artificial intelligence and climate change.

Below, you find a number of project ideas for your thesis that are related to the research conducted in my research group. Some projects are proposed by external partners. Basing your thesis on ongoing work can be rewarding and has a higher chance of resulting in a publication. The short proposals below are grouped by 1) data science projects 2) policy and governance projects.

If you are interested in any of those topics, please contact me before you begin to prepare your proposal. Also contact me if you have an idea that relates to the [research in my group](#), and you would like to iterate on it.

Relevant resources and datasets that could help identifying research ideas related to my work:

- Climate Change AI's searchable database of workshop papers on machine learning and climate change: <https://www.climatechange.ai/papers>

- Bike Sharing data for Berlin covering six months <https://citylab-berlin.org/en/projects/bikesharing/> as well as cycling counter data <https://www.berlin.de/sen/uvk/verkehr/verkehrsplanung/radverkehr/weitere-radinfrastruktur/zaehlstellen-und-fahrradbarometer/>
- Voluntarily provided cycling data (including routes, speed, and experienced vibration) <https://www.ecosense.mein-dienstrad.de/>
- Crowd source bicycle data for various cities <https://www.movebis.org/>
- ACMs SIGEnergy collection of energy datasets and models: <https://energy.acm.org/resources/>
- Traffic data such as <https://utd19.ethz.ch/index.html>; <https://berlinmobil-app.de>; <https://www.mdm-portal.de>
- Open source wind power data <https://onlinelibrary.wiley.com/doi/10.1002/we.2766>
- ENTSO-E Transparency Platform <https://transparency.entsoe.eu/>
- Radiant ML Hub <https://mlhub.earth>
- I also have EU Laws annotated for policy design, annotated annual corporate reports, and US municipal codes.

Data science projects

Machine learning and urban cycling [additional supervision by Silke Kaiser]

Promoting carbon-free transport in cities has emerged as a significant topic of discussion not only in German cities but across Europe. Numerous scientific publications highlight the benefits associated with this initiative, including a reduction in noise and air pollution, improved individual and public health, and a substantial decrease in carbon gas emissions resulting from the shift away from motorized traffic. However, one of the current challenges faced by city planners is the limited and fragmented information available regarding these transport modes. Most cities rely on only a handful of cycling and pedestrian counting stations. A striking example is the city of Berlin, which possesses approximately ten times more motorized traffic counting stations compared to those dedicated to monitoring cyclists.

Machine Learning offers an innovative solution to related issues. Two key inquiries arise: Firstly, what methodologies can be employed to comprehend the spatial distribution of cycling and pedestrian counters within a city? What underlying patterns arise (mainly located at big streets or intersections, close to stores, rather central, next to above average cycling lanes, ...)? Can one propose another distribution/location? Secondly, can machine learning in combination with time series effectively predict cycling and/or pedestrian counts, and how far in advance can these predictions be made?

A promising avenue for research involves designing a master's thesis focused on the city of Berlin or any other urban area of interest. Data for Berlin is available in our research group, and additional data from diverse sources like OpenStreetMaps, weather data, infrastructure data, or bike-sharing data could be crucial and are to be gathered by the student.

Predicting building energy demand across Europe [Partner: MCC and TU Berlin]

Buildings are responsible for about 40% of the EU's energy consumption. To design decarbonization policies, spatially resolved energy consumption needs to be predicted at scales that go beyond district- or city-level case studies. In a large [research project](#) with researchers from TU Berlin, Mercator Research Institute on Global Commons and Climate Change, and ETH Zürich, we are generating a building-level database for the entire EU residential building stock, which will include more than 200 million buildings. We are also developing a model that combines large scale and high granularity to estimate the energy demand for heating in residential buildings and simulate retrofit scenarios. Potential thesis projects can involve machine learning techniques to fill data gaps, gathering and working with energy supply data for buildings, and working on the energy model. Different topics would require different levels of data science, machine learning or engineering expertise.

What resolution should satellite imagery have to detect electricity infrastructure? [Partner: University of Edinburgh]

In the effort to keep global warming below 1.5°C, gaps in the availability of reliable data emerges as one of the roadblock in the path of sustainable planning. Data on the locations of electricity infrastructure is often not available, particularly in low-income countries. Machine learning has proven itself as a reliable method to infer energy infrastructure from satellite imagery but it is unclear what quality of imagery is required for a neural network to be able to detect for instance transmission or the even smaller distribution towers. Knowing the answer to that question is crucial for projects that aim to fill gaps in global energy datasets using automated and hence scalable tools.

In this project, the student will use state of the art object detection models and apply them to a [labelled dataset of electric infrastructure](#) in high-resolution satellite imagery. Then, the imagery will gradually be sampled down (i.e. artificially coarsened) to observe how the network's performance decreases in the process. Ideally, a threshold of required resolution will be deduced.

During the thesis, you will acquire skills in one of the most intricate and useful fields of machine learning, along with experience with essential methods of data manipulation that are tremendously helpful in conducting data-driven research. Prerequisites for this project are coding experience in Python and some understanding of basic machine learning terminology.

Energy technology dictionary

In order to achieve ambitious climate targets, all sectors of the economy need to move away almost entirely from fossil fuels within the next three decades. This means that a mix of many different low-carbon technologies needs to be deployed at scale, many of which are not market-ready yet or face significant deployment barriers. Policy makers need to understand path-dependencies in their strategies, avoid carbon lock-in, navigate political feasibility, and need timely data to learn from other jurisdictions. This thesis will develop an approach for automated detection of low-carbon technologies in texts, to monitor technological trends across a large number of documents, and identify policy priorities in regards to technology options. A central research component will involve developing a dictionary and taxonomy of low-carbon technology terms that can be used on example text corpus of Nationally Determined Contributions. A large annotated dataset of legislative texts is available.

Understanding sustainability impacts of multimodal mobility schemes, analysing open data on mobility platforms [Partner: German Environment Agency / Umweltbundesamt – UBA]

Mobility platforms are becoming increasingly widespread, which are built based on open source and open data principles. The German Environment Agency (Umweltbundesamt) and its Environment and Transport Department is interested in cooperating on thesis projects making use of these data to analyse sustainability impacts, and would provide guidance on thesis' research aims. Research ideas may focus on how users and user choices can be characterized, how the platforms may contribute to urban sustainability, deriving actionable insights for legislators from these data, and informing new policy initiatives. The findings and recommendations may feed into a larger research project carried out by the German Environment Agency and create the basis for evaluating ongoing and planned legislative initiatives on mobility data and platforms at national and EU level, and the impact of these mobility platforms and multimodal mobility schemes onto sustainable mobility.

Possible datasets include:

- Communal data such as bbnavi in Brandenburg (<https://bbnavi.de>) or BerlinMobil (<https://berlinmobil-app.de>)
- Platforms by the German ministry of transportation mCloud (<https://www.mcloud.de>) and MDM (<https://www.mdm-portal.de>)

Policy and governance projects

Adverse effects of artificial intelligence on climate change mitigation

How AI is applied can affect decarbonization strategies at large, and it adds uncertainty to what are feasible pathways for reaching net zero. Given the diversity of application areas, and lack of systematic studies as many areas that are nascent, it is currently very difficult to understand the overall impact of AI on GHG emission (positive and negative) and to include AI in forward looking analyses. Providing more systematic impact assessments on the applications of AI systems is therefore urgently needed to anticipate and shape the technology's effect on climate change. A particular challenge could be to estimate system-level indirect effects of AI on energy consumption and greenhouse gas emissions, including adverse effects that could counteract decarbonization efforts. Examples here include increased fossil fuel production efficiency, cost-effectiveness and longevity of emissions-intensive infrastructure, and effects of targeted advertising on consumption. This thesis will be a narrative analysis, based on interviews and literature review, about AI-driven technological developments that might have an adverse effect on climate change mitigation.

Mobility platforms and multimodal mobility schemes – Regulatory framework conditions and policy options to foster a more sustainable platform economy across the transport sector [Partner: German Environment Agency / Umweltbundesamt - UBA]

Mobility platforms are becoming increasingly widespread, which bundle private, shared and public mobility offers and schemes on digital platforms, and integrate additional infrastructure and transport-related data. These platforms are aimed at marketing different travel options to promote multimodal mobility based on the Mobility-as-a-Service (MaaS) principle. However, the actual transport and socio-ecological effects of this expanding platform economy in the transport sector are uncertain. They depend on the organizational and business models of the platform-based mobility concepts and the underlying regulatory frameworks. Economic and legal issues regarding data access, sovereignty and data protection are important, and significantly influence the potential of the platform economy to promote social and environmental sustainability in transportation.

For example by way of systematic literature review and expert interviews, the thesis may examine which regulatory framework conditions are necessary for the business models, operating concepts and data architectures in order to minimize the negative social and environmental effects of the mobility platforms. Moreover, the thesis can analyse governance and policy requirements regarding the data that is made available by the users of these mobility platforms.

The thesis will be written in cooperation with the German Environment Agency (Umweltbundesamt) and its Environment and Transport Department, which provides further info and guidance on thesis' research aims. The findings and recommendations will feed into a larger research project carried out by the German Environment Agency. Moreover, this project will create the basis for evaluating ongoing and planned legislative initiatives on mobility data and platforms at national and EU level and the impact of these mobility platforms and multimodal mobility schemes onto sustainable mobility.

Methods Expertise: *What type of research methods can you advise on?*

Statistics (with focus on prediction), data science, machine learning, policy analysis

Supervision Style: *What is distinctive about your approach to supervision?*

Supervision will be conducted by regular colloquia and in individual meetings. I am always happy to set up individual meetings as they are very important for developing the research. Team projects are also welcome.

Examples of Previously Supervised Thesis Topics: *(if applicable)*

The preprint of a thesis that I have co-supervised at ETH Zürich and that resulted in a publication can be found here: <https://arxiv.org/pdf/2108.01415.pdf>

Thesis topics I have supervised at Hertie include:

- Tracking Differences in Word Usage Across Party Lines: A Word Embedding Approach
- Autonomous People-Movers Are Almost Ready for Germany, but Is Germany Ready for Them?
- How effective are interventions in energy-efficient appliance use and purchase? – A Meta-Analysis

Recommended General Readings: *(if applicable)*

- Rolnick et al. (2022) [Tackling Climate Change with Machine Learning](#)
- Morgan (2017) [Theory and Practice in Policy Analysis: Including Applications in Science and Technology](#)
- Kaack et al. (2022) [Aligning artificial intelligence with climate change mitigation](#)

Plagiarism:

Plagiarism is an infringement of § 11 Good Academic Conduct, 2a: "Infringements of the standards of good academic conduct include for instance to use wordings, ideas or other intellectual work of others in an academic work without clearly indicating the author. The obligation to indicate the authorship of others shall apply irrespective of whether or not the sources used are protected by copyright" (See:

Exam Rules, § 11 Good Academic Conduct for more information). It is vital to keep track of your sources and to cite all material properly. The Library will offer a session available to all students on resource management and proper citation.

Additional notes on plagiarism [optional!]

--

3. Master's Thesis Colloquium

Joint data science thesis colloquia sessions:

The colloquium sessions are partly run together with faculty from the Data Science Lab (Prof. Lynn Kaack and Prof. Slava Jankin). The first 2 joint colloquia sessions take place before the master thesis proposal submission deadline. The main purpose of these sessions is to guide you in the process of identifying a research topic, developing an appropriate research question, and preparing the proposal in the context of data science-based research. The 2 remaining joint colloquia session take place in the second half of the fall semester. There will also be colloquia sessions that are just run by me, just for my supervisees. The schedule below clarifies on the timing and sequencing of the joint and individual sessions.

If you are an MDS student, please consider the following additional information:

1. You are strongly encouraged to attend all the joint colloquia. You are going to be supervised by one of the data science professors, and they will guide you in identifying and developing and appropriate research topic and question for your thesis.
2. You do not have to sign up for the joint sessions. We will inform you about the exact dates and times in due time.
3. You will be assigned to one of the data science professors after having submitted your thesis proposal. Before that, we don't expect you to reach out to us individually outside of the joint colloquia.

If you are an MPP/MIA student, please consider the following additional information:

1. If you strongly prefer to be supervised by one of us and you plan to work quantitatively/data science-based, we also encourage you to attend the first two sessions of the joint colloquia.
2. In that case, please contact datasciencelab@hertie-school.org if you would like to attend these optional sessions. We will inform you about the exact dates and times in due time.
3. We cannot guarantee that you will be assigned to one of the data science professors. However, if that is the case, the third and fourth sessions of the joint colloquia are mandatory.

Participation & Milestones:

There are four main assignments to be delivered at some point over the course of the colloquium:

- Thesis proposal
- Project presentation
- Pre-analysis plan
- Research report

Overview:

Colloquium dates	Format	Session content
Session 1: Mid-September 2024	Joint session (optional for MPP/MIA)	Exploring the landscape of data science thesis topics and ideas
Session 2: Mid-September 2024	Joint session (optional for MPP/MIA)	From thesis idea to proposal: the art of developing a research question
Master's thesis proposal supervision – Deadline October 1, 2024, via Moodle		
Session 3: Mid-October 2024	Supervisor session (mandatory)	Presentation and discussion of thesis proposals
Session 4: Mid-November 2024	Joint session (mandatory)	Developing the structure and research design of your thesis
Session 5: Early-December 2024	Joint session (mandatory)	Notes on scientific writing for publication
Session 6: February 2025	Supervisor session (mandatory)	Research plan
Session 7: March 2025	Supervisor session (mandatory)	Results
Session 8: Mid-April 2025	Supervisor session (mandatory)	Final draft

Sessions:

Session 1 (joint): Exploring the landscape of data science thesis topics and ideas	
Aim	We will begin with an introduction to the different types of data science research themes and ideas that you can explore to identify potential topics for your master's thesis. The supervisors will also offer a number of project ideas and inspiration from their own area of research and practice partners. The goal of the session is to give you an impression of the broad variety of interesting directions that your research can take before deciding for one particular area.
Assignment	Thesis proposal draft containing a description of (several) research ideas

Session 2 (joint): From thesis topics and ideas to proposal: the art of critical questioning	
Aim	You have found an idea, now what? This session will delve into the process of turning a vague idea and topic into a researchable problem for a thesis project. We will look at some examples, strategies and tips for forming research questions, which can then be used to build your proposal.

Assignment	A short presentation of your proposal (of 5 minutes) and the full proposal for the next meeting. (Due: next meeting)
------------	--

Session 3: Mid October Presentation and discussion of thesis proposal	
Aim	Discuss the proposal and develop a timeline for the project
Assignment for this session	For this session , please prepare a short 5-minute presentation on your proposal. The pitch should clearly explain what the research question is, and how are you going about addressing it.

Session 4 (joint): Developing the architecture and research design of your thesis	
Aim	A large part of writing the master's thesis is centered around the research methodology. This session will provide examples of some of the major quantitative research designs and key considerations such as data management, biases, validity, reliability, and ethical concerns.
Assignment	<p>Develop a pre-analysis plan (4-5 pages). The plan should contain the following sections and be sent it to me by 23:59, Dec 16, 2024.</p> <p>The proposal should contain the following sections:</p> <ol style="list-style-type: none"> 1. Summary: Describe the main goals of the project, what is the scientific question you are trying to answer, and how will you go about it in no more than 300 words. 2. Motivation and background: Why is this research question interesting to you, how does your background align with the topic, and what skills or knowledge do you want to acquire by doing the work that is valuable for your future career? 3. Introduction: Discuss what is the context for the research question and what previous work has been conducted in this area. Please discuss 1-3 key academic papers that are most relevant to your project. You should summarise the papers, the goal and achievement of the papers, data and methods used in the papers, and how they relate to your proposed topic. Also include references to other relevant papers in the introduction as needed. 4. Research question: What research question will you try to answer with the thesis and what is the main approach to address the question? 5. Data and methods: What data and methods are you planning to use? Please be as specific as possible in what you want to approach to achieve to facilitate further discussion whether it is appropriate for answering the research question. Also mention alternative methods, especially if you are unsure about which method is best.

	<p>6. Expected results: Describe what a successful outcome of your project would look like.</p> <p>7. Timeline of the project: A Gantt chart or a table that indicates key milestones.</p> <p>Please use Overleaf or another software to draft the proposal in LaTeX. We will use the session to discuss the project proposals, answer any questions that you may have, and discuss next steps.</p>
--	---

Session 5 (joint): Notes on scientific writing for publication	
Aim	<p>One of the goals of writing the master's thesis is to contribute to the scientific literature and thus, your writing and the structure of your thesis should be at the level of publishable research. This session will provide an overview of the process of scientific writing for the purpose of publication and some tips on how to prepare your manuscript for potential submission to a journal or conference.</p>
Assignment	<p>Spend your time during the winter break with focus on becoming an expert in the narrow area around your thesis, collecting data and refining your research design. Start thinking about what the main figure of your thesis should display, and develop draft figures.</p> <p>Research report (3-5 pages). The plan should include (1) an updated version of the pre-analysis plan (see below) and (2) a summary of the progress you made including a discussion of challenges and how you plan to address them. (Due: Early January)</p> <p>Now include a 3-page Related Work section of your thesis. The section should discuss papers that inspired your approach, papers that you use as baselines, papers proposing alternative approaches to the problem, papers applying your methods to different tasks, etc. You shouldn't go deep into detail on any single paper, instead you should explain how the papers relate to each other and how they relate to your work.</p> <p>You should take note of the datasets used in the papers, methods developed and used, evaluation strategies, and theoretical arguments.</p> <p>Use different resources to identify relevant papers, including Google Scholar, Scopus, ResearchGate, as appropriate. You will find that often references cited in articles and reports (especially review articles) can lead you to the most relevant literature in a field. Don't be shy about asking me for help or reaching out to people that have relevant expertise for recommendations.</p>

<p>Session 6: Week of January 9, 2022</p> <p>Research plan</p>
--

Aim	Discuss the research plan and progress
Assignment	<p>Please prepare a short 5-minute presentation on the research plan and progress. The pitch should clearly explain what the research question is, and how are you going about addressing it. If possible, include preliminary results that you have already achieved (this can also be descriptive, e.g. visualizing the data you are working with). Please include at least one original figure.</p> <p>You will be asked to use that pitch to present your work in this session, and engage in a short feedback round on the work.</p>

Midterm Milestone: Analysis	
Aim	Preliminary results and refine the plan for finalizing the work
Assignment	<p>This stage should help you make progress on your thesis, practice your technical writing skills, and receive feedback on both. You are expected to have started your analysis and can present some initial results. For example, if your project involves data analysis, you should present initial experimental results and establish a validation strategy to be performed at the end of experimentation. At the very minimum, you should show that you have setup your data, baseline model code and evaluation metric, and run experiments to obtain some results.</p> <p>By 23:59, February 24, 2025, please extend your draft to now include three more sections on Overleaf:</p> <ul style="list-style-type: none"> - Proposed method: Please describe your main approaches and methods, and you may want to include equations and figures. You should clearly state whether your approaches are original or provide necessary references. Note that you should be as detailed as possible on assumptions and decisions you have taken in your analysis, such that someone could replicate your work based on your description. In the contrary, you do not need to provide text-book descriptions of the methods you are using but enough context for someone slightly outside the field to follow along. To provide an example, I do not need to read another description of what a random forest is beyond a short summary and explanation of why it is the most appropriate method in your setting. Instead, I do want to know what you did to choose hyperparameters and what values you used based on that (in the Experiments section). - Experiments: This is the part where you describe how you used the methods to answer your research question. You may want to describe the dataset(s) used (provide references), discuss the evaluation metric(s), validation and test sets, and any other details needed to understand your evaluation strategy. Use this section to explain how you set up your analysis, and how robust and general your findings

	<p>will be. If applicable, also discuss how you ran the experiments, providing necessary technical details.</p> <ul style="list-style-type: none"> - Planned work: describe what you plan to do for the rest of the project and what results you expect to find.
--	---

Session 7: Week of March 20, 2023	
Results	
Aim	Discuss results and qualitative analysis of the findings.
Assignment	<p>By 23:59, March 17, 2025, please update the discussion of your analysis with what you have done in the meantime. You will also need to add a 2-page results section to your draft. Please also prepare a 5-min presentation about your work.</p> <p>The results section will contain a description of your results and a qualitative evaluation of the results. You should provide details that show you understand what the results mean based on how you set up your analysis. What are appropriate details will differ by the type of analysis you conducted. This draft should also include first ideas of appropriate visualisation strategies. This concretely means that you should have a draft of a main figure for your research and include that in your presentation to be discussed.</p> <p>We will use this session for the research presentations.</p>

Session 8: Mid April	
Final draft	
Aim	Discuss the final version of the thesis
Assignment	<p>Send the final version of the thesis with all the sections completed to me by 23:59, April 21, 2025. It should provide a complete discussion of the work undertaken for the project that allows a reader to understand what you have done, replicate the work if they want to, contextualize the results, and understand the significance of your work.</p> <p>Your final version of the thesis builds on your midterm milestone and should contain the following sections (the titles and divisions may change based on your topic):</p> <ol style="list-style-type: none"> 1. Abstract: should be a concise (less than 300 words) motivation of the problem, description of the aims and your contribution, and main findings. 2. Introduction: Explains the problem and the research question, the significance of the problem, an overview of the state of the literature and how your work is different, and explains key ideas of your approach and results. 3. Related Work

	<ol style="list-style-type: none"> 4. Research question and approach 5. Methods and experiments 6. Results 7. Discussion 8. Conclusion 9. Acknowledgements 10. Contribution (only for team projects): Describe the contributions of each team member. 11. References: You should use BibTeX referencing. 12. Appendix: Additional material that you would like to include that is not part of the main analysis but should be published alongside the master's thesis. <p>We will use this session to discuss the final version of your thesis and clarify any remaining questions before you are required to hand it in.</p>
--	--

Session 9: On demand Final check and poster preparation	
Aim	Check any outstanding issues in the write up stage and discuss the poster concept
Assignment	For the day of the session, you should prepare a digital poster mock-up identifying the key visuals from the thesis that you intend to use.