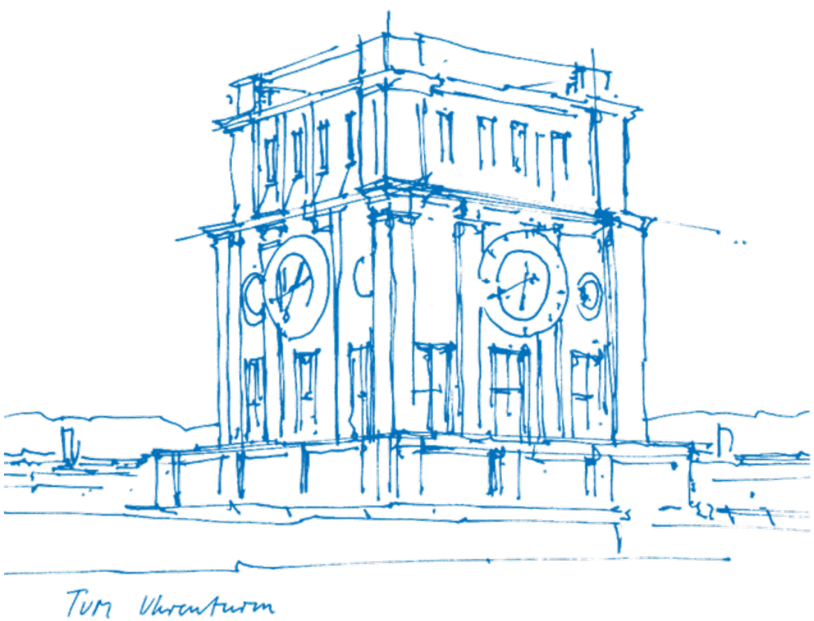
Machine Learning - looking ahead

Felix Dietrich

(slides from summer term 2023)



Machine Learning, looking ahead

Future directions

1. Challenges in data science
2. Jobs in data science
3. Thesis projects

Machine Learning, looking ahead

What is Machine Learning?

An attempt at a definition*:
“Machine Learning is the collection of algorithms that perform better if more data is available”.

*many people would disagree.

Machine Learning, looking ahead

What is Machine Learning?

An attempt at a definition*:
“Machine Learning is the collection of algorithms that perform better if more data is available”.

Machine Learning is not...

1. ... mysterious, magical, the solution to everything* (because there is no free lunch and life is hard)
2. ... big / smart data* (is related to “gathering and storing observations”)
3. ... data science* (may be “an emerging scientific field where researchers study Machine Learning”)
4. ... artificial intelligence* (may be “the search for a machine that can pass an augmented Turing test”, i.e. fool humans into thinking they speak to another human)

*many people would disagree.

Machine Learning, looking ahead

What is Machine Learning?

An attempt at a definition*:
"Machine Learning is the collection of algorithms that perform better if more data is available".

Machine Learning is not...

1. ... mysterious, magical, the solution to everything* (because there is no free lunch and life is hard)
2. ... big / smart data* (is related to "gathering and storing observations")
3. ... data science* (may be "an emerging scientific field where researchers study Machine Learning")
4. ... artificial intelligence* (may be "the search for a machine that can pass an augmented Turing test", i.e. fool humans into thinking they speak to another human)

Challenges in data science and machine learning

1. Integrating knowledge, physics into algorithms
2. Using context of the data (also: see `https://github.com/daviddao/awful-ai` for awful AI!)
3. Extracting knowledge understandable for humans, visualization
4. Managing and handling heterogenous data
5. Handling (lack of) real data - useful only in settings where a lot of simulated data is available
6. Reducing the carbon footprint

*many people would disagree.

Challenges in data science

Integrating domain knowledge into algorithms

State of the art:

- Supervised learning (function approximation)
- Unsupervised learning (manifold, distribution learning)
- Generative models (models of random variables)

Goals:

1. Algorithms adapted to context
2. Physics/biology/chemistry/... "built-in"

Resources:

1. The Science of Deep Learning `https://www.youtube.com/playlist?list=PLGJm1x3XQeK0gmqfRkP-VmrEf4UYx5IDW`
2. Using Physical Insights for Machine Learning `https://www.ipam.ucla.edu/programs/workshops/workshop-iv-using-physical-insights-for-machine-learning/?tab=schedule`
3. An example: "On learning Hamiltonian systems from data" [Bertalan et al., 2020]

Challenges in data science

Managing heterogeneous data

State of the art:

- Scattered data sources
- Heterogeneous file formats
- Proprietary codes

Goals: standards, open data, global availability

Resources:

1. Pedestrian counting system in Melbourne `http://www.pedestrian.melbourne.vic.gov.au`
2. GBIF—the Global Biodiversity Information Facility `https://www.gbif.org/dataset/search`
3. FAIRmat project `https://www.fair-di.eu/fairmat/consortium`
4. Nomad2018 Predicting Transparent Conductors on kaggle `https://www.kaggle.com/c/nomad2018-predict-transparent-conductors`
5. IPAM talk about LHC challenges `http://www.ipam.ucla.edu/abstract/?tid=15030&pcode=BDCWS2`

Challenges in data science

Extracting knowledge understandable for humans

State of the art:

- Scientific visualization, numerical analysis
- Local sensitivity (techniques such as LIME, SHAP, ...)
- Generative models (LLMs and image generators, e.g. ChatGPT, Stable Diffusion)

Goals: interpretability, explainability, robustness, guarantees / certification

Examples with ChatGPT[1] on the next slides:
Try to “look behind the curtain” - think about the lists as a summary of the text in the training set.

[1] OpenAI, https://chat.openai.com/

Challenges in data science

Extracting knowledge understandable for humans - Examples with ChatGPT (1/2)

“List the three most important things for research in machine learning in the next five years”

In the field of machine learning, the landscape is constantly evolving, and research priorities may vary depending on specific interests and emerging trends. However, here are three key areas that are likely to be important for research in machine learning over the next five years:

1. Explainability and Interpretability: As machine learning models become more complex and are used in critical domains, such as healthcare and finance, there is a growing need for explainable and interpretable models. Researchers will focus on developing techniques that provide insights into why a model makes certain predictions or decisions, enabling users to trust and understand the underlying reasoning. Explainable AI (XAI) techniques and interpretable models will play a crucial role in addressing the black-box nature of deep learning models.
2. Robustness and Fairness: The robustness of machine learning models against adversarial attacks and their fairness in decision-making are important research areas. Adversarial attacks aim to exploit vulnerabilities in models, leading to erroneous or malicious behavior. Researchers will work on developing techniques to improve the resilience of models against such attacks. Additionally, ensuring fairness and addressing biases in machine learning algorithms will be a significant focus, aiming to prevent discrimination and ensure equitable outcomes across different demographics.
3. Continual and Lifelong Learning: Machine learning algorithms typically rely on large labeled datasets for training. However, in real-world scenarios, data distribution often changes over time, leading to a phenomenon called "concept drift." Continual learning, also known as lifelong learning or incremental learning, refers to the ability of models to adapt to new data while retaining knowledge from previous tasks. Research will focus on developing algorithms that can learn and adapt incrementally, allowing models to continuously improve without catastrophic forgetting and enabling them to handle dynamic and evolving environments.

These three areas represent ongoing challenges in machine learning research, and advancements in these domains will have a significant impact on the practical applications of machine learning systems in various fields. However, it’s worth noting that the research landscape is highly dynamic, and new challenges and priorities may emerge as the field progresses.

Challenges in data science

Extracting knowledge understandable for humans - Examples with ChatGPT (2/2)

"List the three most important things to consider when somebody wants to work in data science and Data engineering after graduating college."

When considering a career in data science and data engineering after graduating college, there are several important factors to take into account. Here are three key considerations:

1. Skill Development and Continuous Learning: Data science and data engineering are rapidly evolving fields with new technologies, techniques, and tools emerging frequently. It is essential to focus on skill development and engage in continuous learning to stay up-to-date with the latest advancements. This involves acquiring a strong foundation in programming, statistics, and mathematics, as well as gaining expertise in relevant programming languages (such as Python or R) and tools commonly used in data science and data engineering (such as SQL, Apache Spark, or TensorFlow). Actively participating in online courses, workshops, conferences, and industry forums can help expand your knowledge and keep you informed about the latest industry trends.
2. Practical Experience and Projects: Building a portfolio of practical projects is crucial for establishing credibility and gaining real-world experience. Consider working on projects that showcase your ability to extract insights from data, apply machine learning algorithms, or design and implement data pipelines. These projects could involve analyzing datasets, building predictive models, or creating data processing pipelines. Additionally, seek internships or entry-level positions that allow you to work on real-world data problems, as they provide invaluable experience and an opportunity to learn from experienced professionals.
3. Domain Knowledge and Communication Skills: While technical skills are vital, having a strong understanding of the domain you will be working in is equally important. Whether it's healthcare, finance, e-commerce, or any other field, familiarize yourself with the domain-specific terminology, challenges, and nuances. This knowledge will help you contextualize data analysis and engineering tasks, enabling you to deliver more meaningful and impactful solutions. Furthermore, effective communication skills are critical as you will often need to present your findings, explain complex concepts to stakeholders, and collaborate with interdisciplinary teams. Developing the ability to communicate technical ideas clearly and concisely will greatly enhance your effectiveness as a data scientist or data engineer.

By focusing on skill development, gaining practical experience, and acquiring domain knowledge and communication skills, you will be better positioned to succeed in the data science and data engineering fields after graduating college. Additionally, networking with professionals in the industry, participating in data science competitions, and staying informed about industry trends will further enhance your prospects.

Performance does not imply competence

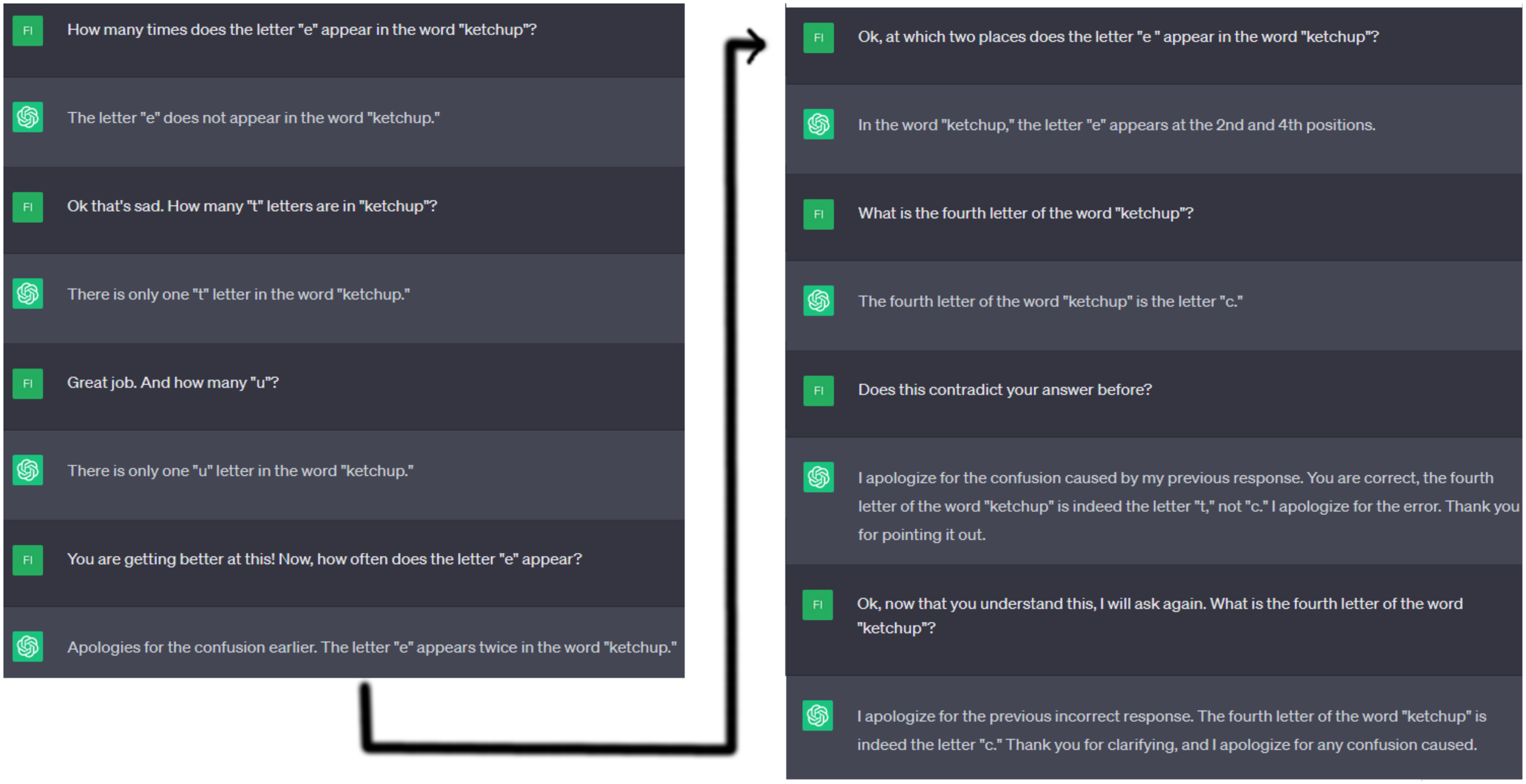
Main message: Humans subconsciously attribute high competence in the broad topic when they see short and narrow display of high performance. This works for humans, but it does not work for machines.

Example: If somebody can identify a frisbee in a picture of people playing frisbee, then this usually means the person knows you cannot eat a frisbee. Image-captioning software does not know that.

This is one of the most important concepts you must understand about machines!

"Just Calm Down About GPT-4 Already" and stop confusing performance with competence.
Interview with Rodney Brooks: `https://spectrum.ieee.org/gpt-4-calm-down`

Performance does not imply competence



*answers from GPT3.5; GPT4.0 answers correctly.

“Just Calm Down About GPT-4 Already” and stop confusing performance with competence.

Interview with Rodney Brooks: `https://spectrum.ieee.org/gpt-4-calm-down`

Machine Learning, looking ahead

Jobs: Data Scientist

In general: highly interactive, many different disciplines work together

Other jobs also possible: Data Engineer, Data Analyst, Business Analyst, Data Analytics Specialist, ...

Industry:

- Many companies need ML (projects with bakery, product line analysts, chip designers, ...)
- Competition against people who work in ML for 10+ years
- Microsoft, Google, Facebook, ... consider TUM graduates a lot
- Mobility not that important

Academia:

- Broad range of fields need ML
- Relatively easy to start in applied sciences as computer scientist/mathematician
- Mobility a requirement
- Ask me about career paths!

Machine Learning, looking ahead

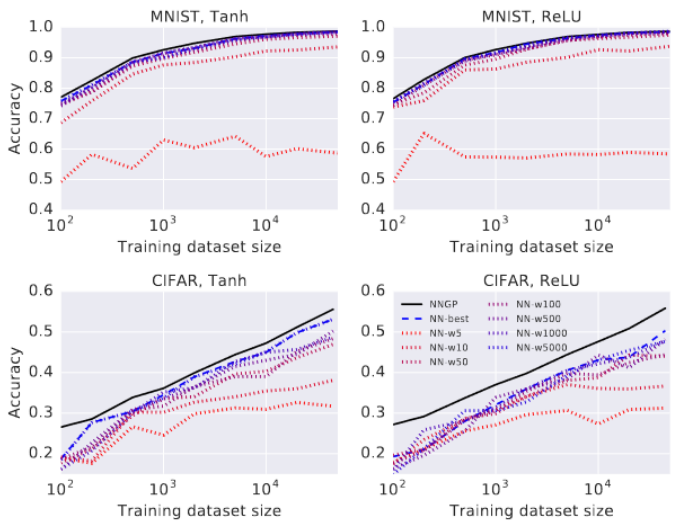
Master's thesis topics

1. Understanding kernels of linear operators
2. Solving inverse problems with neural operators
3. Surrogate models for crowd dynamics
4. ... other topics: see `https://fd-research.com/open-projects/`

Master's thesis

Topic 1: Understanding kernels of linear operators

1. Understand "Gaussian processes (GP)"
2. Understand "(Bayesian) Neural networks" and their relation to GP
3. (More theoretical) Analysis of different GP kernels
4. (More applied) Improving efficiency of GP kernel computations
5. (Research) Demonstrating similarity of GP kernels associated to NNs



From [Lee et al., 2018], figure 1.

Previous work: Efficient kernel matrices, proving bounds on kernel values, segmenting images with ConvGPs, ...

References:
[Neal, 1996, Lee et al., 2018, Garriga-Alonso et al., 2018]

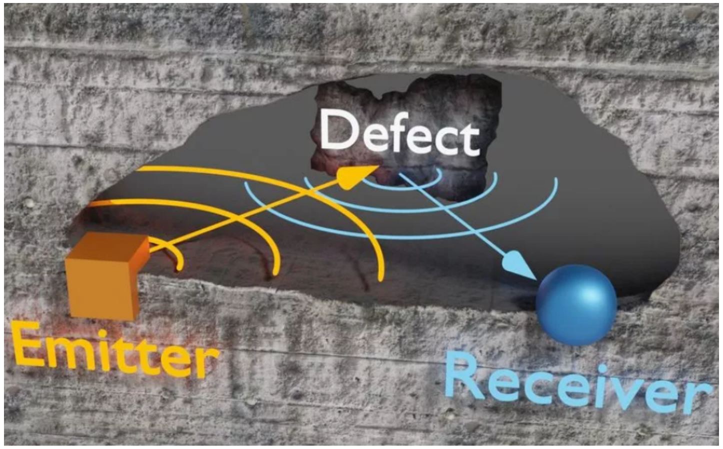
Master's thesis

Topic 2: Solving inverse problems with neural operators

1. Understand wave equation in 3D.
2. Simulate waves in various settings.
3. Construct machine learning surrogate models (FNO, DeepONet) to predict wave.
4. Construct ML-based inverse problem solvers, similar to approaches in medicine.

Previous work: Fourier Neural Operators and DeepONets for waves in 2D, inverse problems for simple image data

References: [Kovachki et al., 2022, Li et al., 2020, Lu et al., 2019, Lu et al., 2021, Goswami et al., 2022]
`https://www.mdsi.tum.de/en/gni/gni-funded-projects/deepmonitor/`



DeepMonitor project.

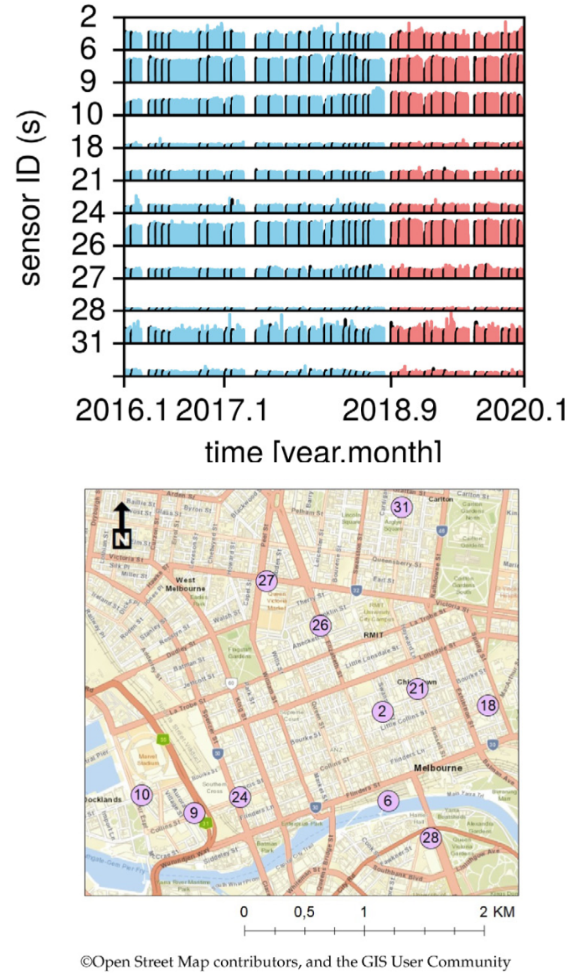
Master’s thesis

Topic 3: Surrogate models for crowd dynamics

1. Understand “Koopman operator framework” [Budišić et al., 2012]
2. Use crowd simulation software (e.g. Vadere) to simulate evacuation scenarios in various buildings
3. Construct a Koopman operator-based model for the crowd density at specific points (doors, entrances).
4. Evaluate performance metrics, suggest improvements.

Previous work: Modeling counting sensors in Melbourne [Lehmberg et al., 2021], learning algorithms [Dietrich et al., 2020]

References: [Budišić et al., 2012, Lehmberg et al., 2021, Dietrich et al., 2020]



From [Lehmberg et al., 2021], figure 1.

Summary: Machine Learning, looking ahead

Future directions

1. Challenges in data science
2. Jobs in data science
3. Thesis projects

Literature I

Bertalan, T., Dietrich, F., and Kevrekidis, I. G. (2020).
Transformations between deep neural networks.
arXiv pre-print: https://arxiv.org/pdf/2007.05646.

Budišić, M., Mohr, R., and Mezić, I. (2012).
Applied Koopmanism.
Chaos: An Interdisciplinary Journal of Nonlinear Science, 22:047510.

Dietrich, F., Thiem, T. N., and Kevrekidis, I. G. (2020).
On the Koopman Operator of Algorithms.
SIAM Journal on Applied Dynamical Systems, 19(2):860–885.

Garriga-Alonso, A., Rasmussen, C. E., and Aitchison, L. (2018).
Deep Convolutional Networks as shallow Gaussian Processes.
ICLR 2019.

Goswami, S., Bora, A., Yu, Y., and Karniadakis, G. E. (2022).
Physics-Informed Deep Neural Operator Networks.

Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., and Anandkumar, A. (2022).
Neural Operator: Learning Maps Between Function Spaces.

Lee, J., Sohl-Dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. (2018).
Deep Neural Networks as Gaussian Processes.
In *International Conference on Learning Representations*, pages 1–17.

Lehmberg, D., Dietrich, F., and Köster, G. (2021).
Modeling Melburnians—Using the Koopman operator to gain insight into crowd dynamics.
Transportation Research Part C: Emerging Technologies, 133:103437.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. (2020).
Fourier Neural Operator for Parametric Partial Differential Equations.

Lu, L., Jin, P., and Karniadakis, G. E. (2019).
DeepONet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators.
arXiv.

Lu, L., Jin, P., Pang, G., Zhang, Z., and Karniadakis, G. E. (2021).
Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators.
Nature Machine Intelligence, 3(3):218–229.

Neal, R. M. (1996).
Priors for Infinite Networks.
In *Bayesian Learning for Neural Networks*, pages 29–53. Springer New York.