EBERHARD KARLS UNIVERSITÄT TÜBINGEN

machine learning new perspectives for science

Seminar Winter Term 2024/2025

(Auto-)ML for tabular data

Katharina Eggensperger

katharina.eggensperger@uni-tuebingen.de

AutoML for Science

October 17th, 2024

Story Line Today

[20min] **Big Picture / Tabular ML**

[?] Your *Questions*

[10min] **Organization**

[?] Your *Questions*

[15min] **Topics/Papers**

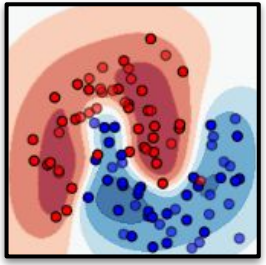
[?] Your *Questions*

[30min; if time left; unlikely] **How to give a good presentation**

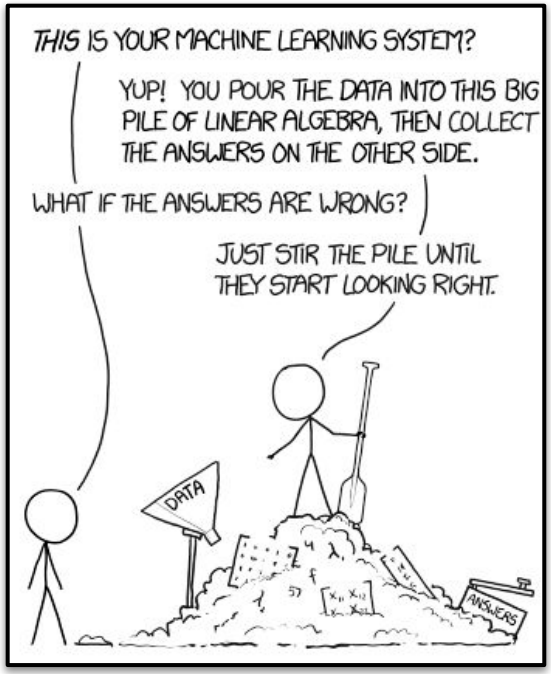
The Big Picture

```
>> What is this about?
```

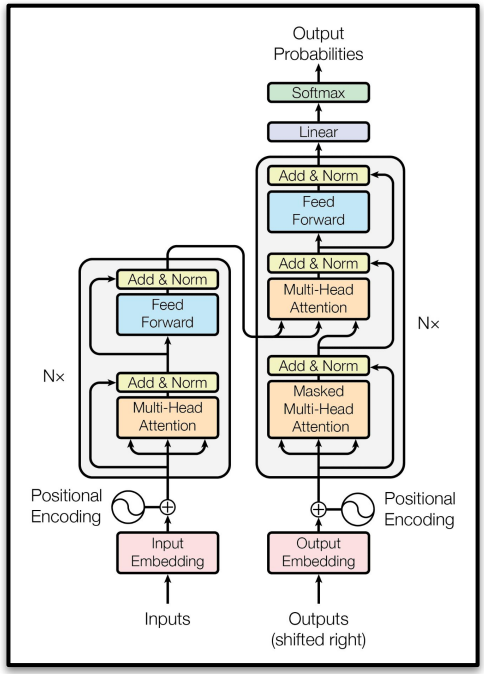
Machine learning is …



source: scikit-learn



source: XKDC



"Attention is all you need" paper by Vaswani, et al., 2017

Tabular data is everywhere!

🩺⛰️🌦️📖🔬

in science: healthcare, biology, geoscience, climate science, psychology, economics, …

💵🦾🚗📦🏢

in industry: finance, manufacturing, e-commerce, marketing, insurance, …

Prediction
"Provide a label given a record"
→ supervised machine learning

Focus of this seminar

Data Generation
"Synthetize more data given some examples"
→data augmentation for data-scarce applications

Data Understanding
"Answer a question based on information provided in a table"
→ extract human-readable information

* based on: Fang et al. "Large Language Models (LLMs) on Tabular Data: Prediction, Generation, and Understanding - A Survey" in TMLR'24

Properties of tabular data? Why is it challenging for ML?*

* Talk with your neighbor for 5mins, we will collect results

Hint: Think about differences between tables and images

Culmen Length	Culmen Depth	Flipper Length	Weight	Sex	Species
39.1	18.7	181	3750	♂	Adelie
39.5	17.4	186	3800	♀	Adelie
40.3	18.0	195	3250	♀	Adelie
35.3	18.9	187	3800	♀	Adelie
40.6	18.6	183	3550	♂	Adelie
40.5	17.9	187	3200	♀	Adelie
42.3	21.2	191	4150	♂	Adelie
45.2	17.8	198	3950	♀	Chinstrap
46.1	18.2	178	3250	♀	Chinstrap
49.8	15.9	229	5950	♂	Gentoo
43.5	15.2	213	4650	♀	Gentoo
51.5	16.3	230	5500	♂	Gentoo
46.2	14.1	217	4375	♀	Gentoo
55.1	16.0	230	5850	♂	Gentoo

Properties/Characteristics of tabular data

Heterogeneity
- different feature types: categorical, numerical (and sometimes text)
- each feature has its own value range
- feature often correlate / are irrelevant

Sparsity
- imbalanced labels
- missing values
- extreme values and long-tailed distributions

Dependence on pre-processing
- encoding of categoricals
- transformation of numericals
- feature engineering
- often requires domain knowledge

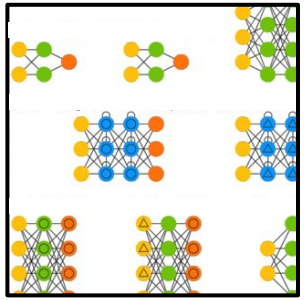
Order invariance
- order of features and samples do not matter

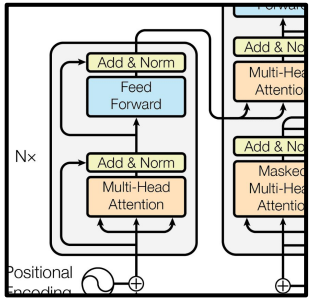
No prior-knowledge
- lack of prior knowledge about structure

All of these provide unique challenges for ML!
(= there's a lot to research)

All of these make the application of tabular ML challenging
(= we need AutoML)

The state of ML for tabular data





Tree-based methods are great

- robust to hyperparameters
- interpretability
- fast training

but

- no gradients
- hard to scale

DL methods are great

- benefit from progress in DL research
- pre-trained models for small data
- fast inference

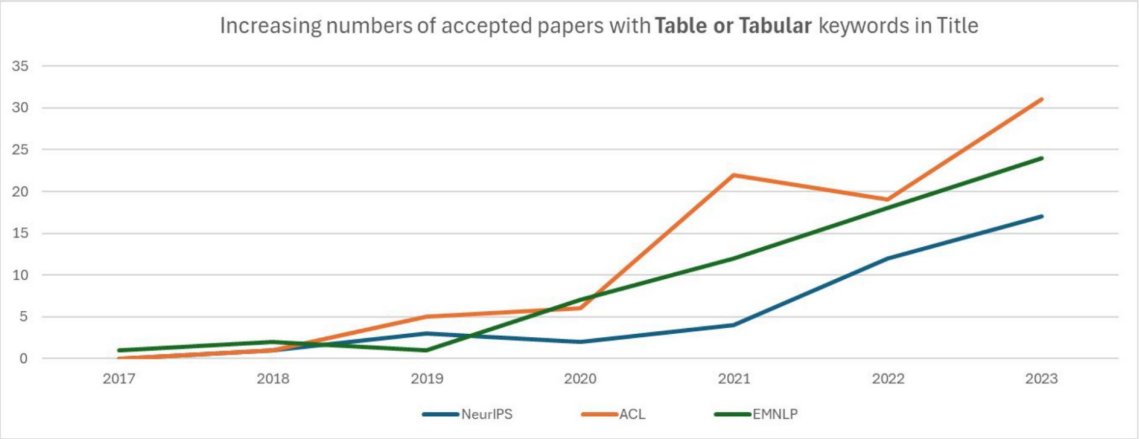
but

- black-box models
- many hyperparameters

Which one is better?
→ not the most interesting research question

Better: AutoML perspective
Which models exists? → How to apply them in practice? | When do they (not) perform well?
→ Can we learn the “how” and “when”?

Tabular ML research is exploding



- very fast expanding research community
- NeurIPS workshop series on "Table Representation Learning"
- regular competitions (Kaggle AutoML Grand Prix; numer.ai)

Sidenote: Main data modality for research on AutoML methods*

* excluding Neural Architecture Search

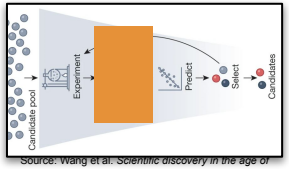
Source: Dong et al. "Large Language Models for Tabular Data: Progresses and Future Directions" SIGIR'24 Tutorial

My motivation: AutoML and tabular data (or why I offer this seminar)

ML is a great tool for research (in science)

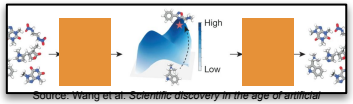
- should be accurate
- should be easy-to-use
- should be reproducible
- should be systematic
- should generalize

Screen solutions



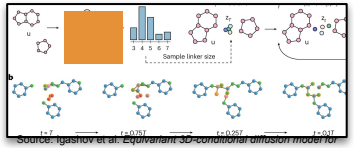
Source: Wang et al. *Scientific discovery in the age of artificial intelligence*. Nature 2023

Latent spaces



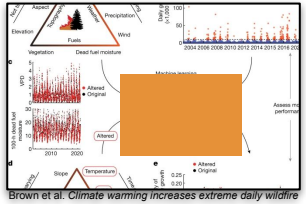
Source: Wang et al. *Scientific discovery in the age of artificial intelligence*. Nature 2023

Design new solutions



Source: Igashov et al. *Equivariant 3D-conditional diffusion model for molecular linker design*. Nature Machine Intelligence 2023

Model relationships



Brown et al. *Climate warming increases extreme daily wildfire growth risk in California*. Nature 2023

Requires AutoML!

- Model Selection
- Hyperparameter Tuning
- Neural Architecture Search
- Meta-Learning
- …

Developing AutoML is **most effective** if

- there is a **large design space**
- there are **many similar tasks**

→ Tabular data has all of it (and is crucial for many domains)

Questions?

Organization

```
>> When, where and how?
```

🌟 Features of this seminar

Main Feature ❤️

🔥 **Broad overview of state-of-the-art tabular ML**
(focusing on supervised learning)

Topics:

- DL vs classic ML
- Specialized DL architectures
- Interpretable models
- Foundation models
- Feature engineering
- Synthetic data generation
- LLMs for tabular data

Bonus Features 💎

🔥 **A lot interaction!**

- Active participation is necessary
- Discussion is part of your presentation

→You will gain a much better understanding

🔥 **A lot of feedback!**

- Feedback from me (and my team) before presentation
- Practice and prepare with your *study buddy*
- Anonymous feedback after the presentation

→High-quality presentations for everyone
→Improve your presentation skills

😢 Not featured: databases, mandatory practical exercises, introduction to ML, indepth AutoML methods

Organizational information

📅 **When?** Thursday 14:00 (c.t.) - 16:00 (actually 14:15-15:45)

📍 **Where?** MvL6, seminar room ground level

🧍 **How many?** 14 students

📚 **Expected Background Knowledge**

- Machine Learning
- Deep Learning (this includes transformers)
- (optional) Practical experience with ML/DL

🎯 **Grades:** Presentation / Slides[1] / Report / Participation

!!Note: Attendance in all sessions is expected & filling out the paper assignment form is binding

All info is on my website: **Seminar: (Auto-)ML for tabular data**

[1] **Send me your slides (as pdf) right after your presentation (latest by the end of the week)**
Note: If you change anything (fixed equations, corrected typo, add explanation) in the slides, please add a short statement in the email.

The presentation

Your presentation lasts ~40 minutes and should consist of:

- **20 minutes presentation**
  i.e. summary of your paper: motivation, methods, experiments, strengths / weaknesses
  → see also separate presentation on “How to give a good presentation”

- **10 minutes discussion**

- **5-10 minutes additional content (before or after questions)**
  i.e. something that is not part of the paper and you found interesting
  - methodological details on the evaluation (e.g. a metric or statistical analysis)
  - methodological detail on the method (e.g. a trick for data processing)
  - a follow-up paper that you’ve read
  - code demo
  - comparison to another paper in the seminar
  - well prepared discussion on connection to AutoML

concluded by a 5 minute break to collect anonymous feedback via a form

The report [tentative]

3-5 pages (format TBD; excluding references) covering the following

A **review** of your paper (1-2 pages)

- Motivation
- Main contribution
- 3 strengths
- 3 weaknesses
- 3 questions you would ask the authors after a conference talk

A broader **discussion of your topic** in the context of the seminar (1-2 pages)

- How does it relate to other topics discussed?
- Do we need AutoML to apply your method?
- Has this method been extended / used elsewhere

A **list with further material** that you've collected during preparation, e.g. (0.5 page)

- Code
- Public reviews
- Blogposts
- Video Tutorials

The ideal timeline

> -14d	-7d	-3d	0	<+2d	end of term
Read Paper	**Prepare Presentation**	**Get Feedback**	**Presentation & Feedback**	**Send Slides**	**Send Report**
Read and understand your topic (also take into account further info, e.g., videos, blogposts, posters, slides, reviews)	Also get feedback from your study buddy	(optional) Get Feedback from me			

Questions?

Dates

17.10.2024	Today	
24.10.2024	no meeting	
31.10.2024	Intro I (How to give a good presentation / Background on AutoML systems)	
07.11.2024	no meeting	
14.11.2024	Intro II (Tabular foundation models / Background on DL for tabular data)	
21.11.2024	#1 DL vs classic ML	[FTTransformer; Why?]
28.11.2024	#2 Interpretability	[GAM X LLM; TabNet]
05.12.2024	#3 In-Context Learning	[TabPFN; ForestPFN]
12.12.2024	#4 More DL	[MotherNet; TabR]
19.12.2024	#5 LLM I	[Elephant; FeatureLLM]
26.12.2024	no meeting	
02.01.2025	no meeting	
09.01.2025	no meeting	
16.01.2025	#6 Data generation	[GReaT; TabDDM]
23.01.2025	#7 LLM II	[TabLLM; Tabula8B]
30.01.2025	buffer / probably no meeting	
06.02.2025	buffer / probably no meeting	

Large models 4 AutoML

#1 Session: DL and Classical ML methods

Why does DL not work out-of-the-box? Why do tree-based methods work better?

1. **[FTTransformer]** Gorishniy et al. Revisiting Deep Learning Models for Tabular Data (NeurIPS'22)
2. **[Why?]** Grinsztajn et al. Why do tree-based models still outperform deeplearning on typical tabular data? (NeurIPS'22)

#2 Session: Interpretability

How can we build interpretable models for tabular model?

1. **[GAM X LLM]** Bordt et al. Data Science with LLMs and Interpretable Models XAI@AAAI'24, Lou et al. Accurate intelligible models with pairwise interactions (KDD'13)
2. **[TabNet]** Arik et al. TabNet: Attentive Interpretable Tabular Learning (AAAI'21)

Large models 4 AutoML (cont'd)

#3 Session: In-Context Learning

Can we train a model to instantly yield predictions for new datasets?

1. **[TabPFN]** Hollmann et al. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second (ICLR'23)
2. **[ForestPFN]** Breejen et al. Why In-Context Learning Transformers are Tabular Data Classifiers (arxiv'24)

#4 Session: More DL

Can we directly predict neural network weights and other methods

1. **[MotherNet]** Müller et al. MotherNet: A Foundational Hypernetwork for Tabular Classification (arxiv'23)
2. **[TabR]** Gorishniy et al. TabR: Tabular Deep Learning Meets Nearest Neighbors (ICLR'24)

AutoML 4 Large Models

#5 Session: LLMs I

Data contamination for LLM for tabular data and LLMs for feature engineering

1. **[Elephant]** Bordt et al. Elephants Never Forget: Memorization and Learning of Tabular Data in Large Language Models (arxiv'24)
2. **[FeatureLLM]** Han et al. Large Language Models Can Automatically Engineer Features for Few-Shot Tabular Learning (ICML'24)

#6 Session: Synthetic data generation

Create more data using diffusion models and LLMs

1. **[GReat]** Borisov et al. Language Models are Realistic Tabular Data Generators (ICLR'24)
2. **[TabDDM]** Kotelnikov et al. TabDDPM: Modelling Tabular Data with Diffusion Models (ICML'23)

#7 Session: LLMs II

Leverage LLMs for supervised learning on tabular data

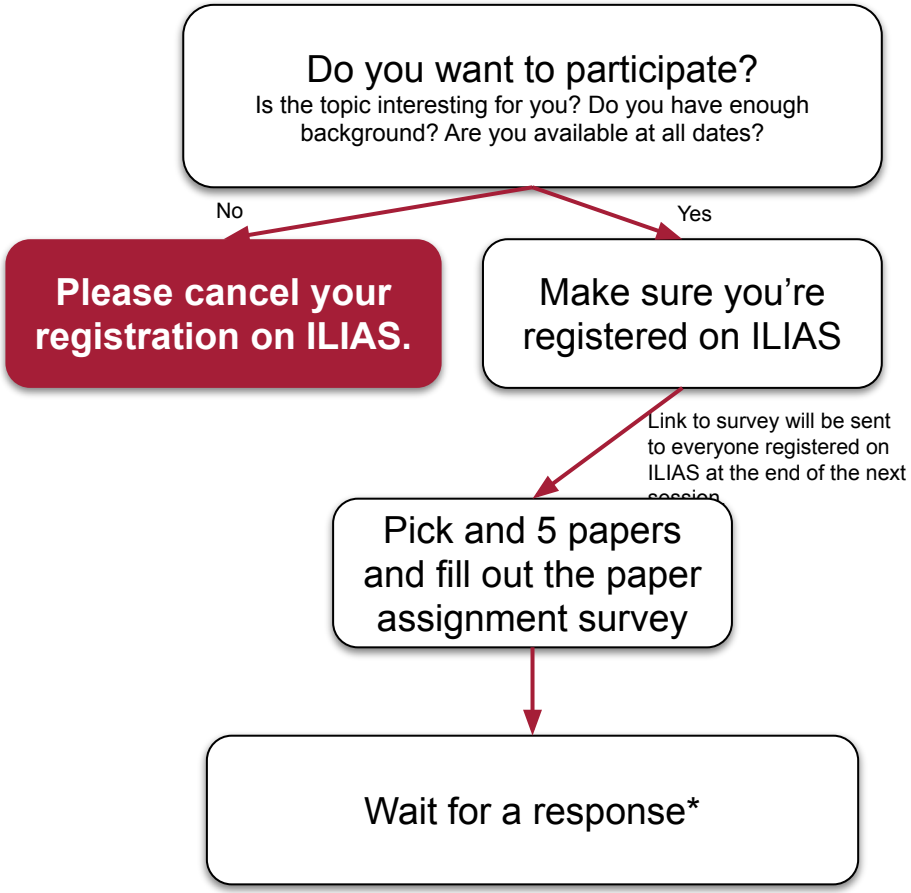
1. **[TabLLM]** Hegselmann et al. TabLLM: Few-shot Classification of Tabular Data with Large Language Models (ICML'23)
2. **[Tabula8B]** Gardner et al. Large Scale Transfer Learning for Tabular Data via Language Modeling (arxiv'24)

Questions?

What’s next?

```
>> What should I do now?
```

Your Next Task?



Important Dates

- **Before Thursday, 24.10.2024 (noon)**
  - (De-)Register on ILIAS
  - Look at topics
- **Thursday, 24.10.2024 - 29.10.2024 (noon)**
  - Fill out topic survey

→ You will hear back **before Thursday, 31.11.2024**

* Assignment policy:
- ILIAS registration is mandatory
- Presence today is required
- Higher priority for new students
  (!= not participated in prior versions)

Also: Filling out the paper survey is binding, i.e. dropping the seminar after that is not possible.