PRIOR LABS

Accurate predictions on small data with a tabular foundation model

September 2025

Outline

Why tabular foundation models should be a priority

“Accurate predictions on small data with a tabular foundation model”

Our roadmap

Article | Open access | Published: 08 January 2025

Accurate predictions on small data with a tabular foundation model

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister & Frank Hutter

Nature **637**, 319–326 (2025) | Cite this article

Foundation models have transformed text & images

PRIOR LABS

But our most valuable data is organized in tables

Text

OpenAI

DeepMind

deepseek

Spreadsheets & Databases

Images

OpenAI

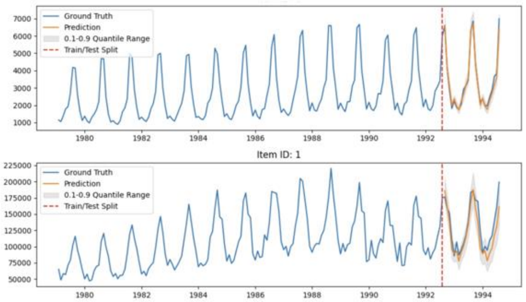
black forest labs.

Tabular models power 1000s of use-cases

PRIOR LABS

male	age	education	currentSmoker	cigsPerDay	BMI	heartRate	glucose	TenYearCHD
1	39	4	0	0	26.97	80	77	0
0	46	2	0	0	28.73	95	76	0
1	48	1	1	20	25.34	75	70	0
0	61	3	1	30	28.58	65	103	1
0	46	3	1	23	23.1	85	85	0
0	43	2	0	0	30.3	77	99	0
0	63	1	0	0	33.11	60	85	1

tabular prediction problem



time-series forecasting problem

Tabular data is the statisticians familiar modality

Tabular prediction examples

- Responder / Non-responder
- Genomic risk prediction
- Cancer risk prediction based on biomarkers
- Designing clinical trials cohorts, predicting dropout
- Predicting hospital readmission
- Generating highly accurate synthetic survey data based on existing samples

How will X develop over time? X could, e.g., be:

- Disease prevalence
- Product demand
- Transaction patterns
- Stock prices
- Call center volume
- ...

Why Tabular Foundation Models?

Bojan Tunguz
@tunguz

Neural networks for tabular data is Machine Learning equivalent of wearing socks with sandals.

PRIOR LABS

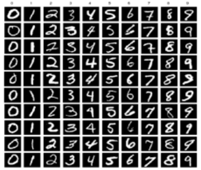
Why Tabular Foundation Models?

male	age	currentSmoker	education	cigsPerDay	BMI	TenYearCHD
1	39	0	Advanced	N/A	26.97	0
0	46	0	Limited	0	28.73	0
1	48	1	Basic	N/A	25.34	0
0	61	1	Secondary	30	28.58	1
0	46	1	Secondary	23	23.1	0

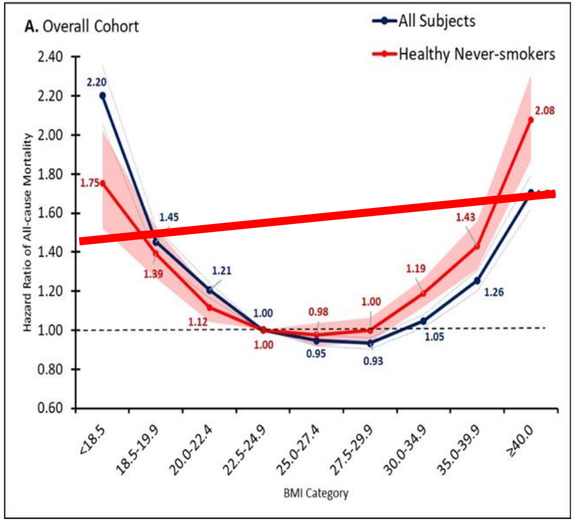
Tabular Predicton Task

1. Free text & concepts not leveraged
2. Missing values not treated natively
3. No semantic problem understanding
4. Poor transfer of knowledge across datasets
5. Poor out-of-distribution results
6. Difficulty integrating with LLMs
7. No tabular search and retrieval
8. Learning on multiple tables
9. Not possible to integrate domain knowledge
10. Adding or removing data requires retraining

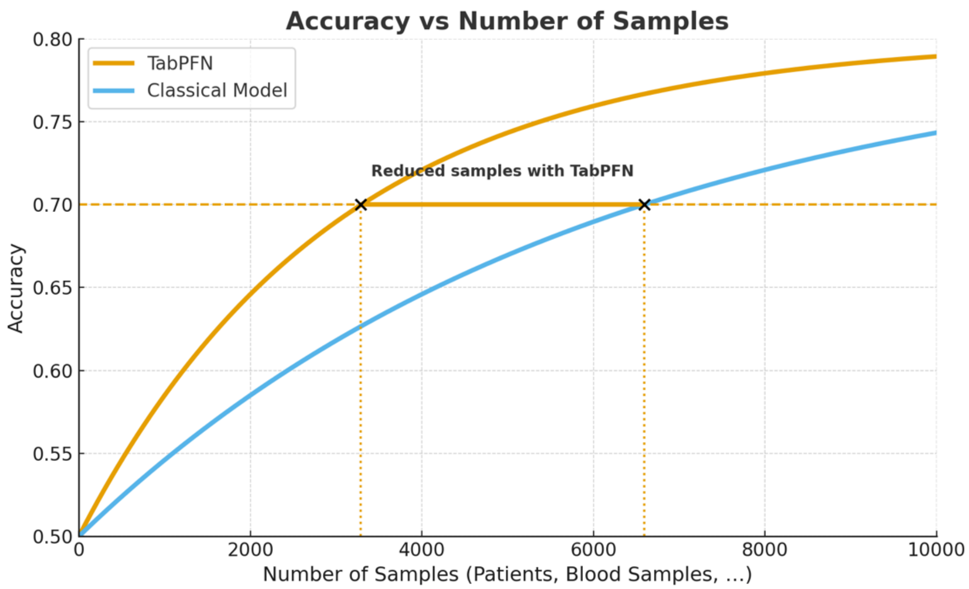
Why Tabular Foundation Models?



Q: What are five strategies to increase sales for a small business?
A:



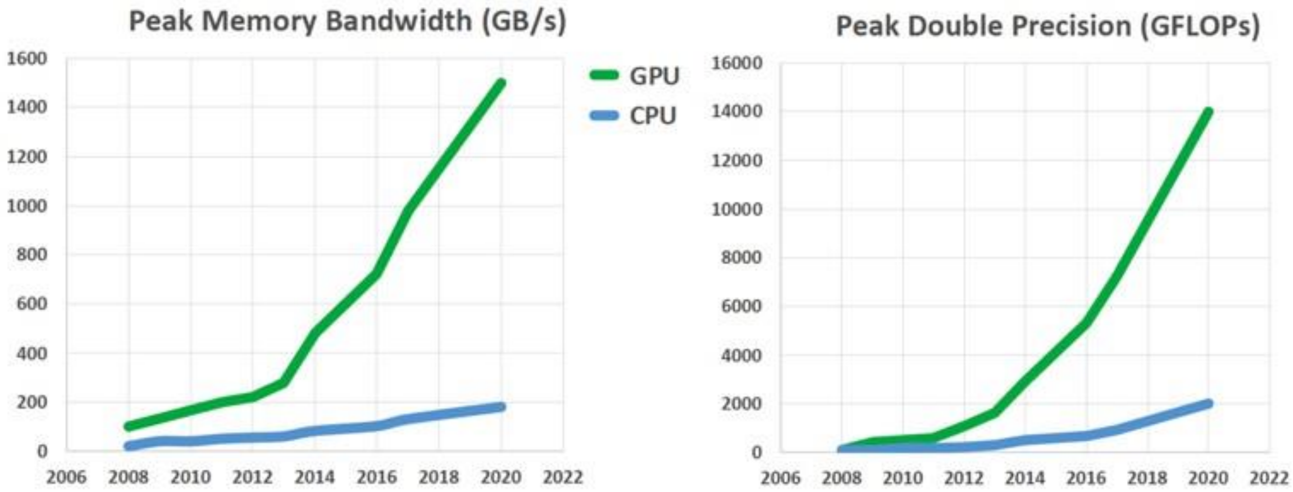
Why Tabular Foundation Models?



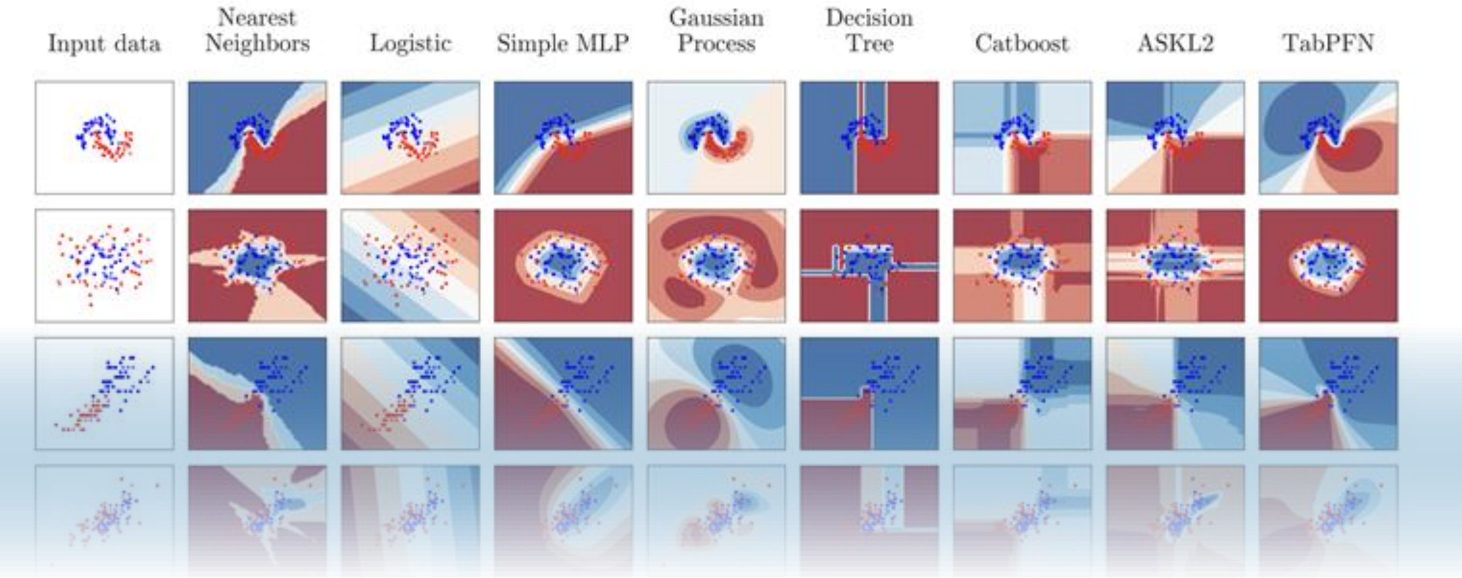
Billions invested in data collection,
better analysis methods can supercharge all these efforts.

Why Tabular Foundation Models?

Hardware and Ecosystem Lottery: Progress and capabilities in GPUs and Attention-based DL methods currently far outsizes CPU



Source: https://www.nextplatform.com/2019/07/10/a-decade-of-accelerated-computing-augurs-well-for-gpus/

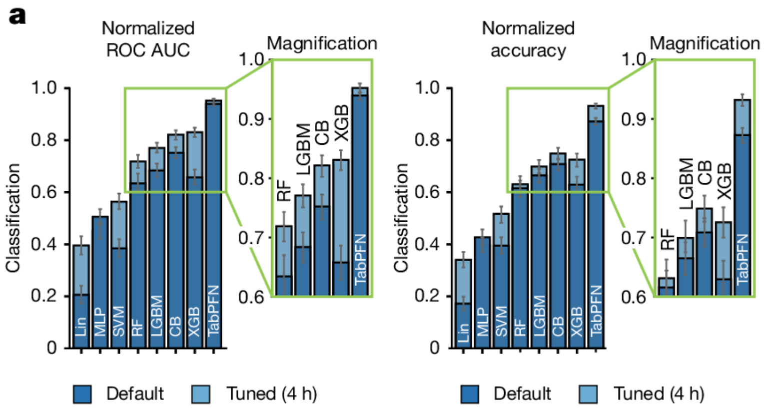


Accurate predictions on small data with a tabular foundation model

Plot based on scitkit-learn example by
Gaël Varoquaux & Andreas Müller

TLDR;

We develop an easy-to-use foundation model for tabular data prediction - using algorithmic in-context learning - outperforming in a second baselines even trained for hours.



Limitations:

Size: up to 10000 data points, 500 features, 10 classes

Article | Open access | Published: 08 January 2025

Accurate predictions on small data with a tabular foundation model

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister & Frank Hutter

Nature **637**, 319–326 (2025) | Cite this article

“Accurate predictions on small data with a tabular foundation model”, Noah Hollmann, Samuel Müller et al. Nature 2025

Helping doctors in decision-making

PRIOR LABS

NHS

New AI Tool to support high-stakes decisions around patient intubation

University of Warwick researchers have led the development of a new AI tool that can help doctors make the difficult and high-stakes decision of whether to intubate a patient in acute respiratory failure.

Acute respiratory failure occurs when the respiratory system cannot provide oxygen to, and/or remove carbon dioxide from, the body. Treatment is primarily based on providing external respiratory support, such as non-invasive ventilation (NIV) through a facemask, but around 40% of patients fail NIV and subsequently require endotracheal intubation and invasive mechanical ventilation.

Published in **Intensive Care Medicine**, University of Warwick researchers have developed a new AI model that can help clinicians identify patients who will need intubation much earlier in their treatment, which improves outcomes for those patients with acute respiratory failure.

https://warwick.ac.uk/news/pressreleases/new_ai_tool_to_support_decisions_around_patient_intubation

Improving outcomes for cancer patients

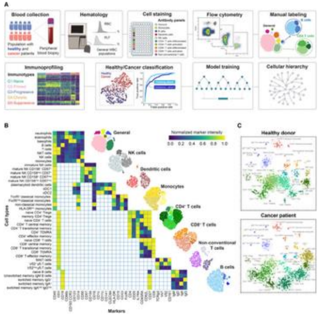
PRIOR LABS

BostonGene

Predicting immunotherapy response in cancer patients is challenging because existing methods overlook the broader immune system. BostonGene developed a **machine-learning-powered blood test and analysis platform** that identifies immune profiles and enables better treatment prediction using TabPFN. Published in "Cancer Cell".

Challenge

- **Limited and expensive data** from complex immune cell analyses
- **High dimensional data** from multiparameter flow cytometry
- **Extensive time** required for manual model tuning as there is a constant flow of new data



53%
error reduction

Differentiating blood samples of cancer patient vs healthy individuals (ROC 0.91 vs 0.81)

90%
time saved

TabPFN reduced model development times by 90% as no hyperparameter tuning was required even with new data.

Enhanced efficiency, allowing scientists to concentrate on meaningful clinical analysis rather than machine learning

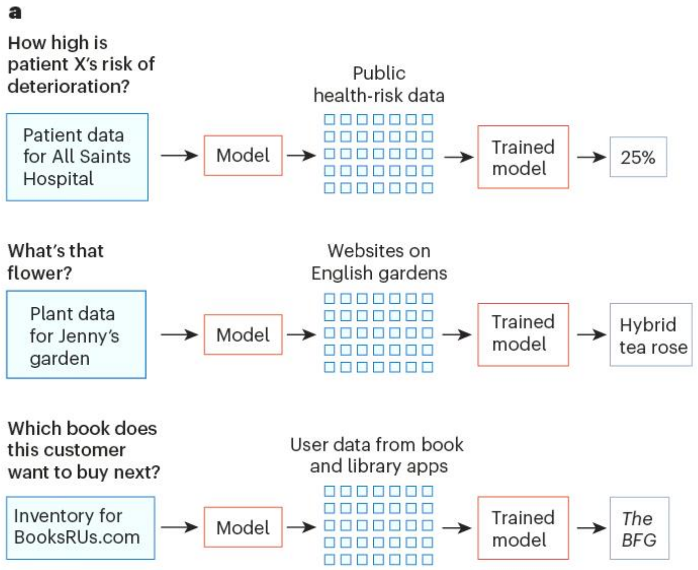
Michael Goldberg, PhD, VP R&D at BostonGene

Accurately classifying cancer patients and healthy individuals based on the distribution of immune cells in the peripheral blood was a remarkably challenging task — TabPFN made it a reality.

Core idea: tabular foundation model

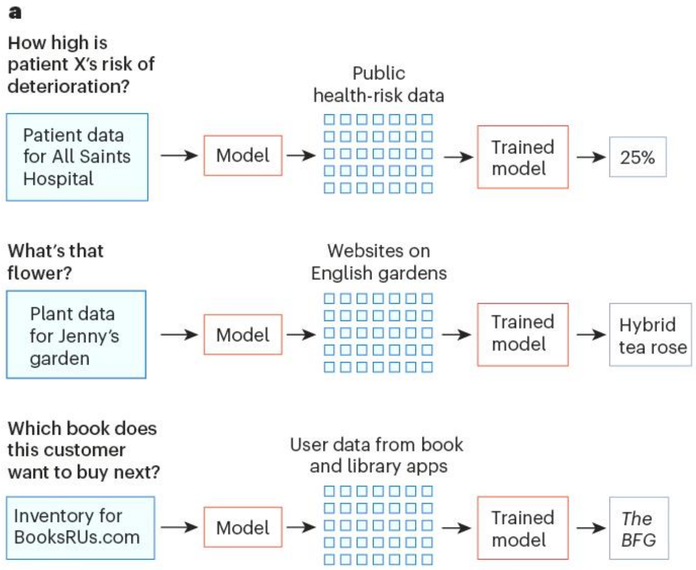
PRIOR LABS

Traditional ML

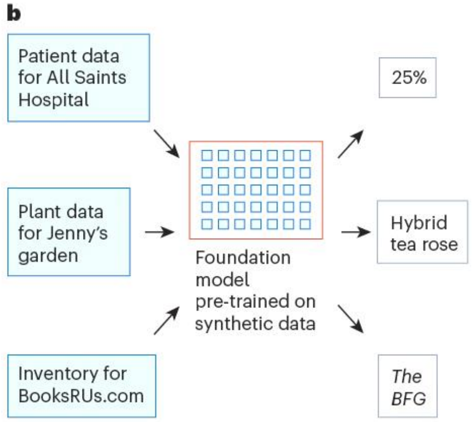


Core idea: tabular foundation model

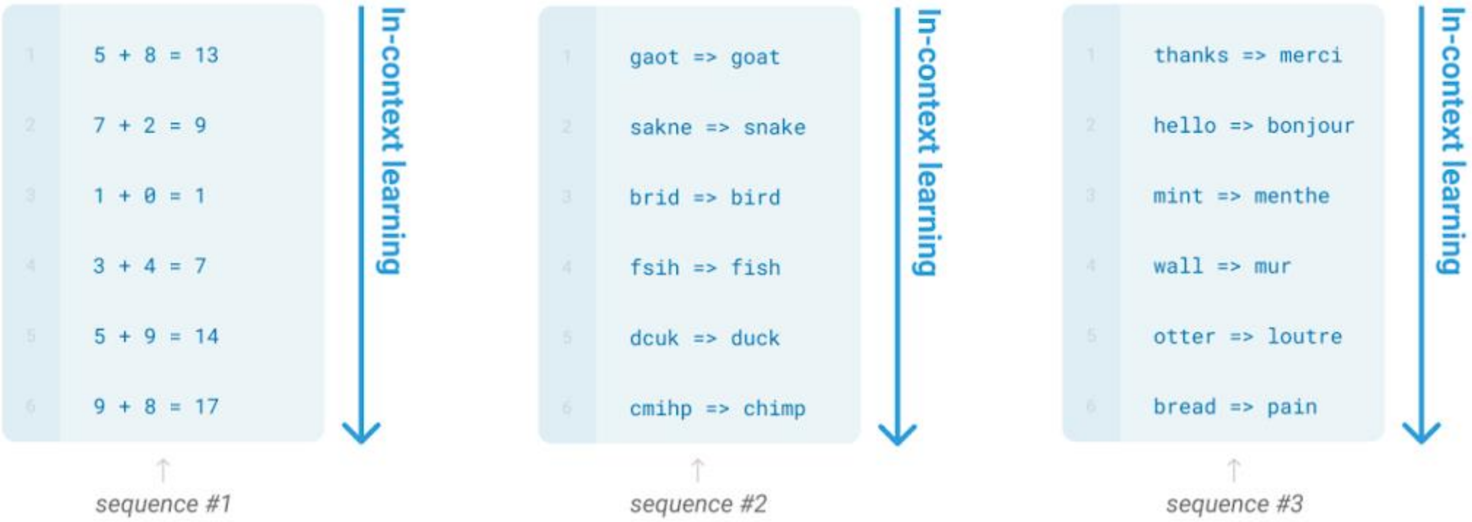
Traditional ML



Foundation models

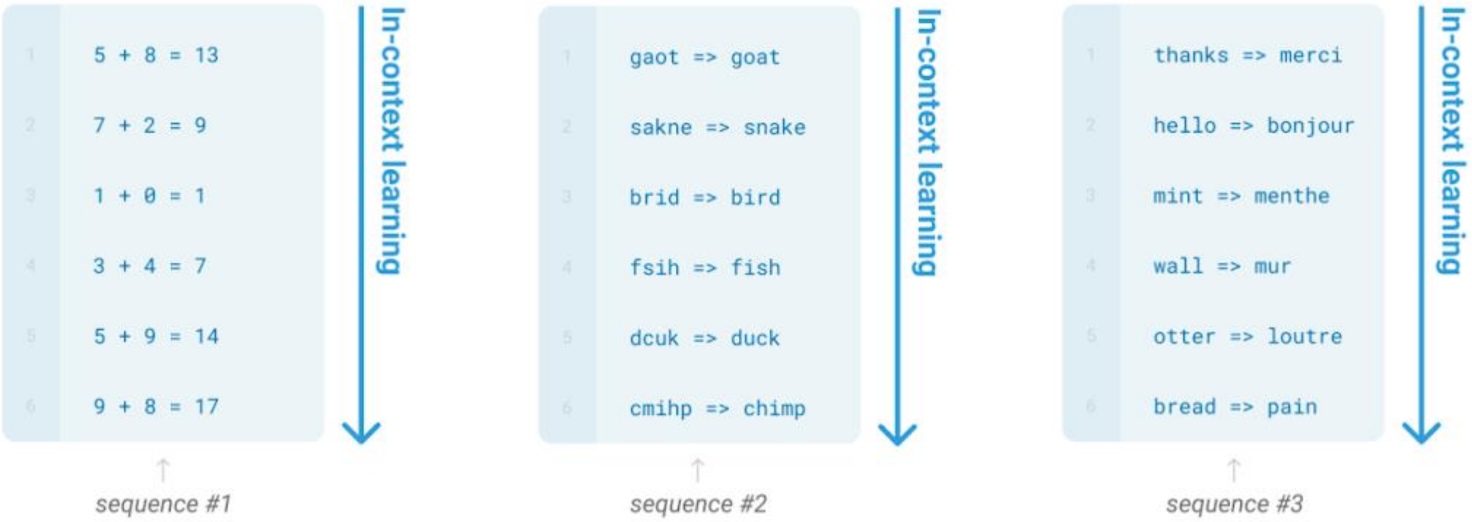


Tabular foundation models use in-context learning to make predictions on a dataset.



In-context learning (ICL) has emerged as a new paradigm for natural language processing (NLP), where LLMs make predictions based on contexts augmented with a few examples.

Image source: https://vitalflux.com/in-context-learning-gpt-3-concepts-examples/



In-context learning (ICL) has emerged as a new paradigm for natural language processing (NLP), where LLMs make predictions based on contexts augmented with a few examples.

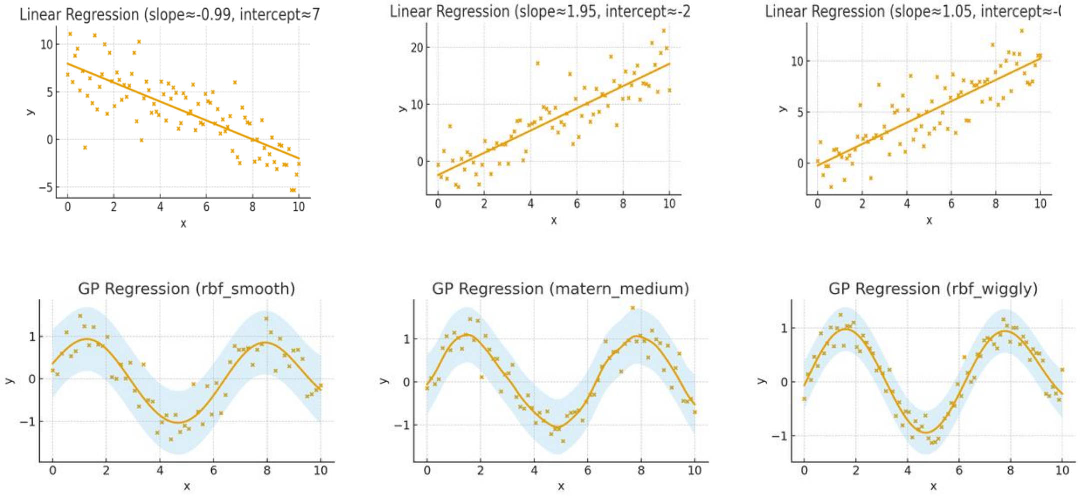
Can we use ICL to learn numeric algorithms?

Image source: https://vitalflux.com/in-context-learning-gpt-3-concepts-examples/

Prior work

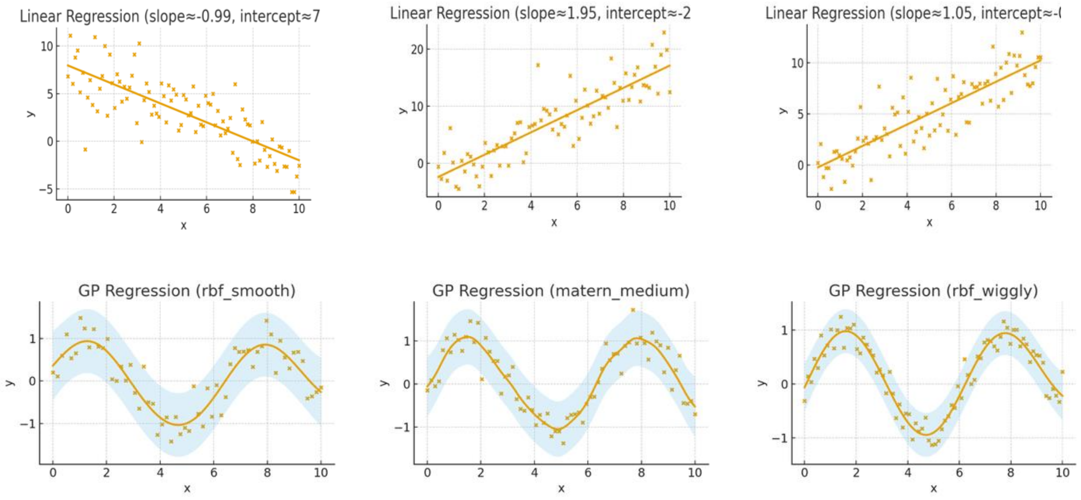
Transformers Can Do Bayesian Inference

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, Frank Hutter
ICLR 2022



Yes! Just show millions of examples to your model

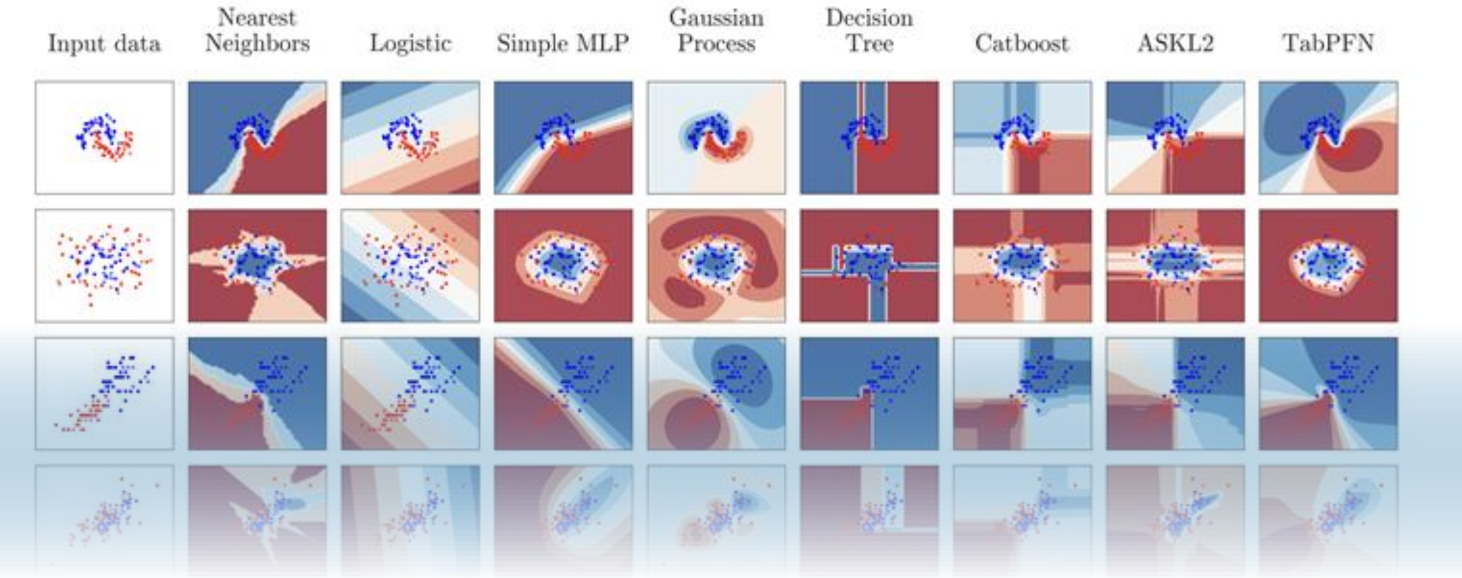
Image source: https://vitalflux.com/in-context-learning-gpt-3-concepts-examples/



Yes! Just show millions of examples to your model

$$p(y|x, D) = \int p(y|x, t)p(t|D)dt$$

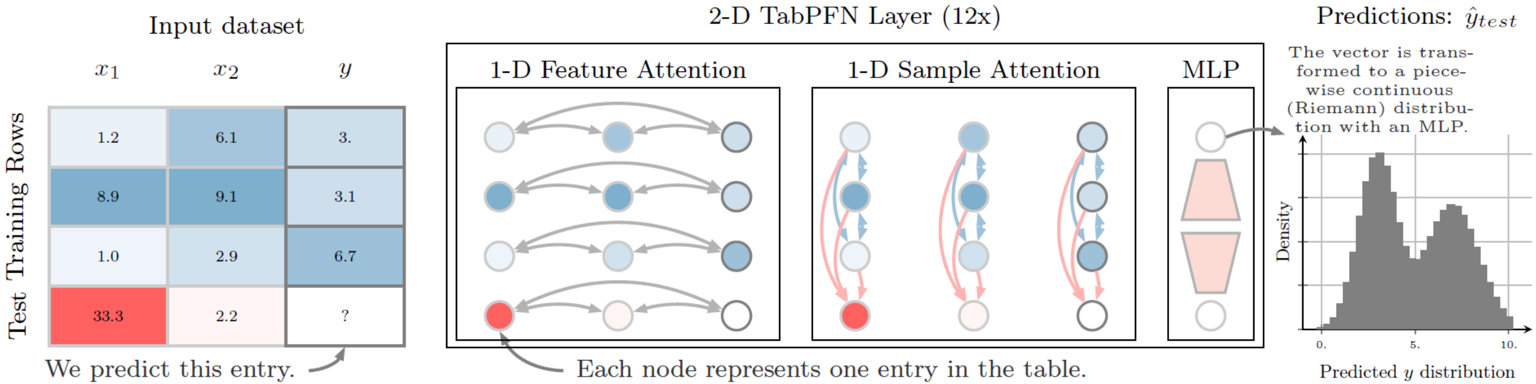
Image source: https://vitalflux.com/in-context-learning-gpt-3-concepts-examples/



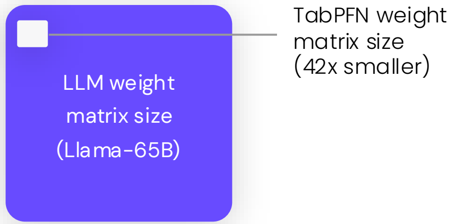
Accurate predictions on small data with a tabular foundation model

Plot based on scitkit-learn example by
Gaël Varoquaux & Andreas Müller

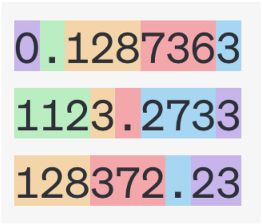
An architecture designed for tables



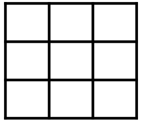
An architecture designed for tables



Small weight matrices

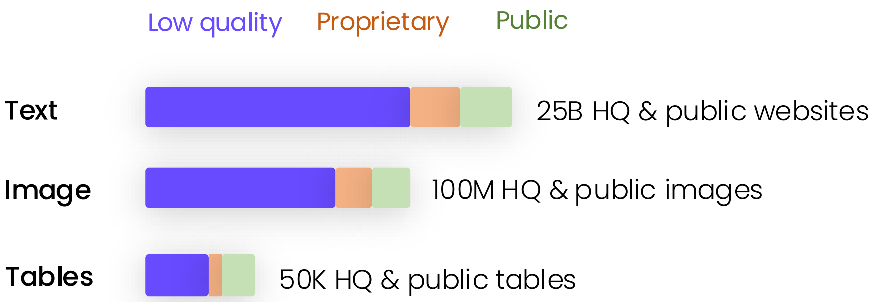


Numerical tokenization



KV-Caching for fast inference

Bridging data scarcity with synthetic data

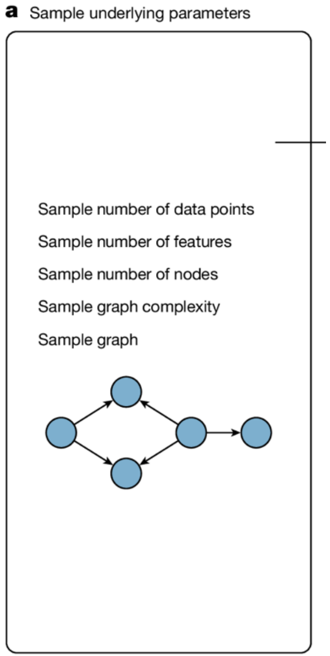


1.000.000x

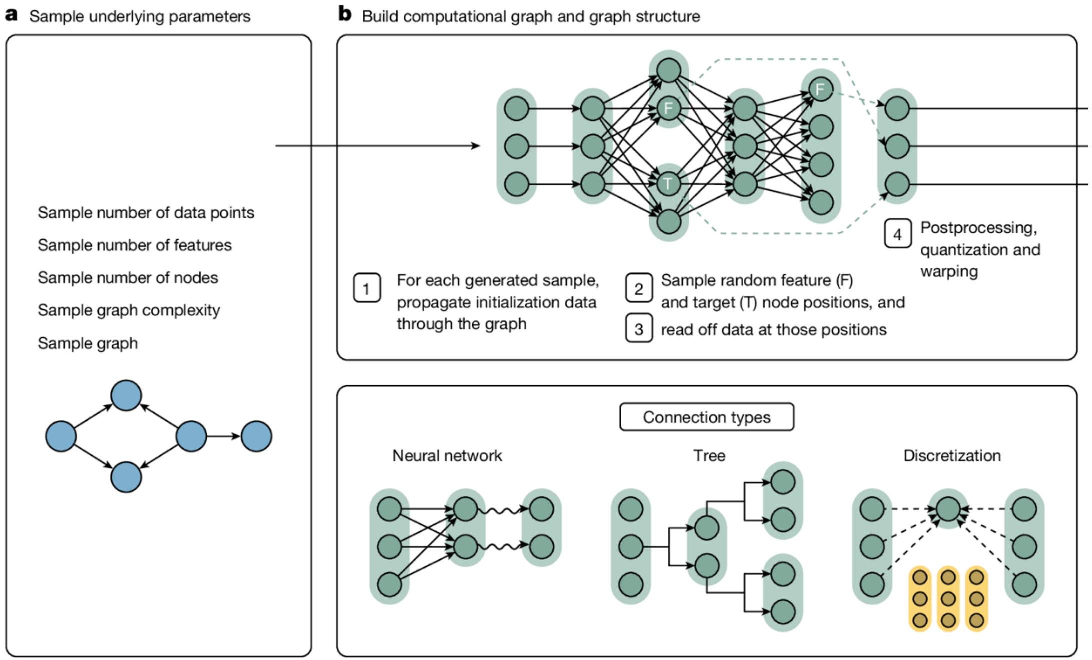
less data available for tables compared to webpages

We are training on synthetic proprietary tabular data

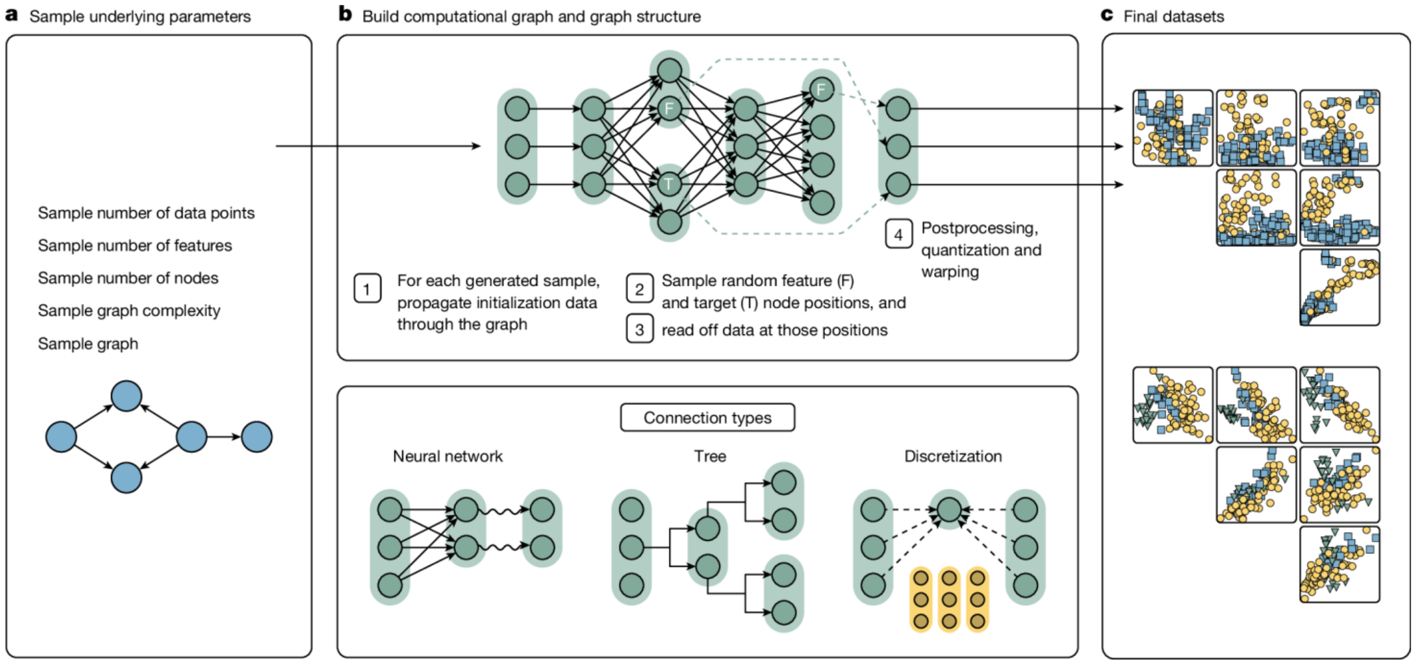
Synthetic data based on principles from causality



Synthetic data based on principles from causality

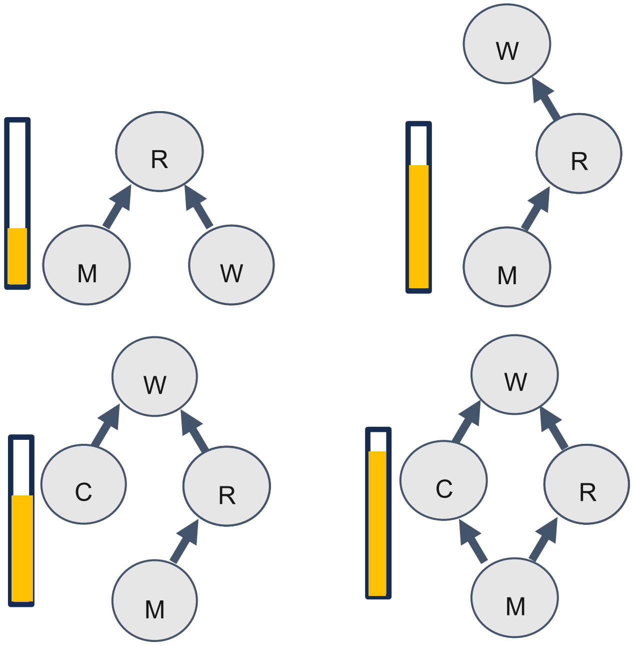


Synthetic data based on principles from causality



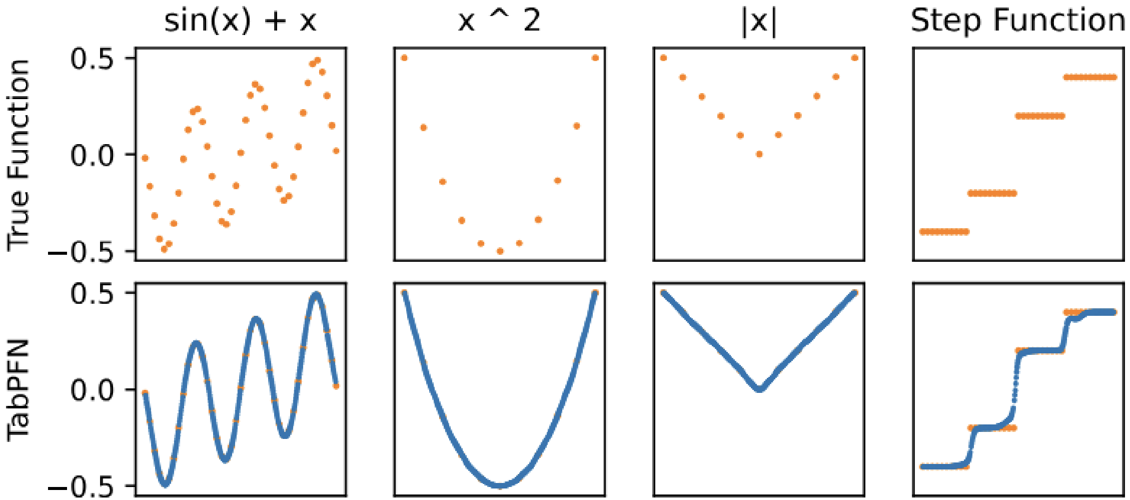
TabPFN Prior: Structural Causal Models

Month (M)	Wet? (W)	Raining (R; Target)
1	Y	Y
2	N	Y

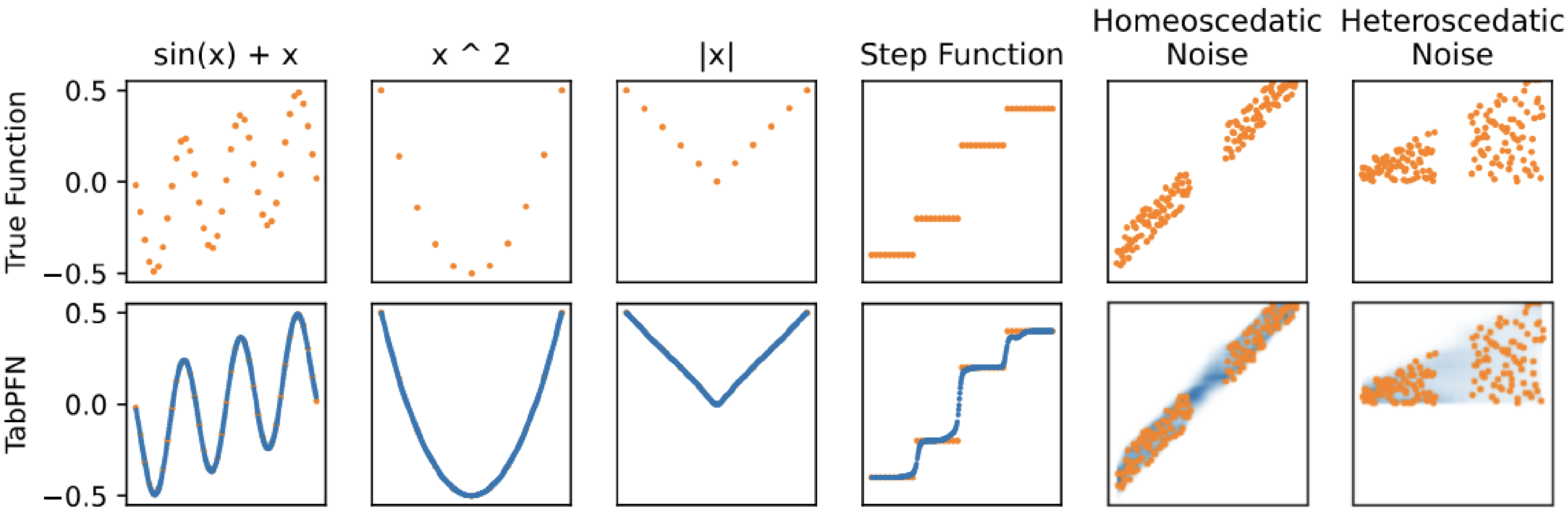


$$p(y|x, D) = \int p(y|x,t)p(t|D)dt$$

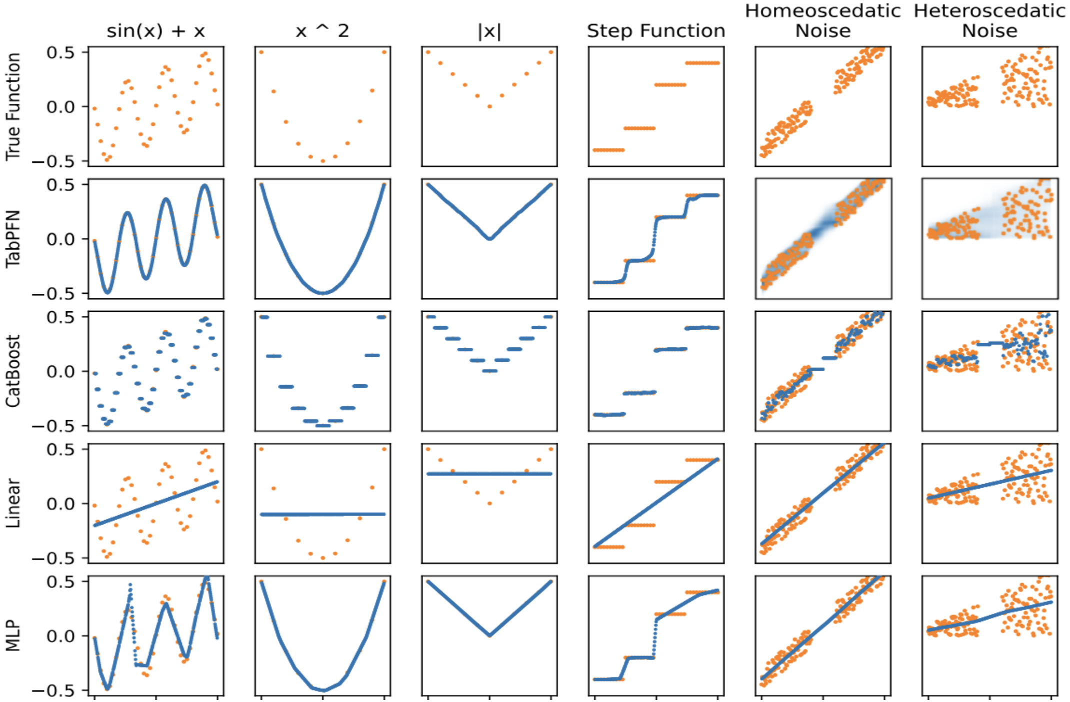
Fitting toy functions



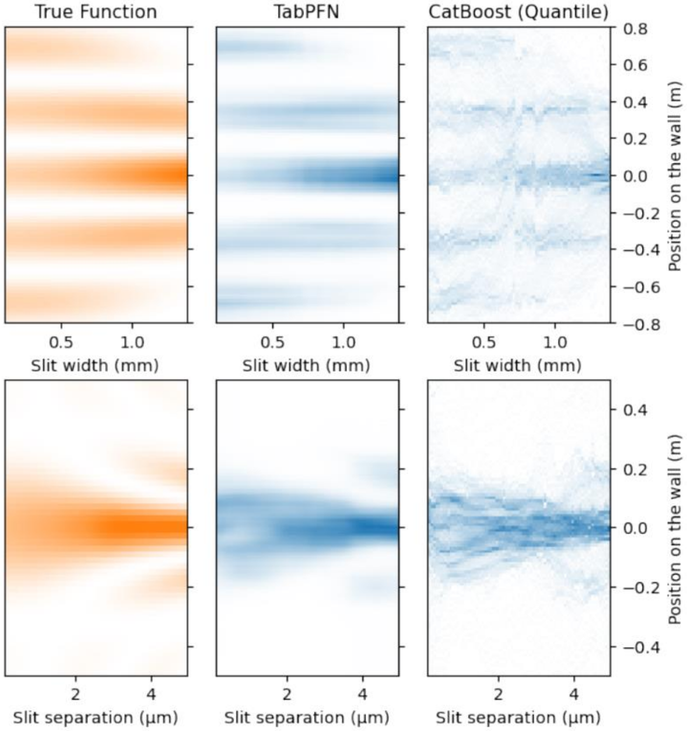
Fitting toy functions



Fitting toy functions



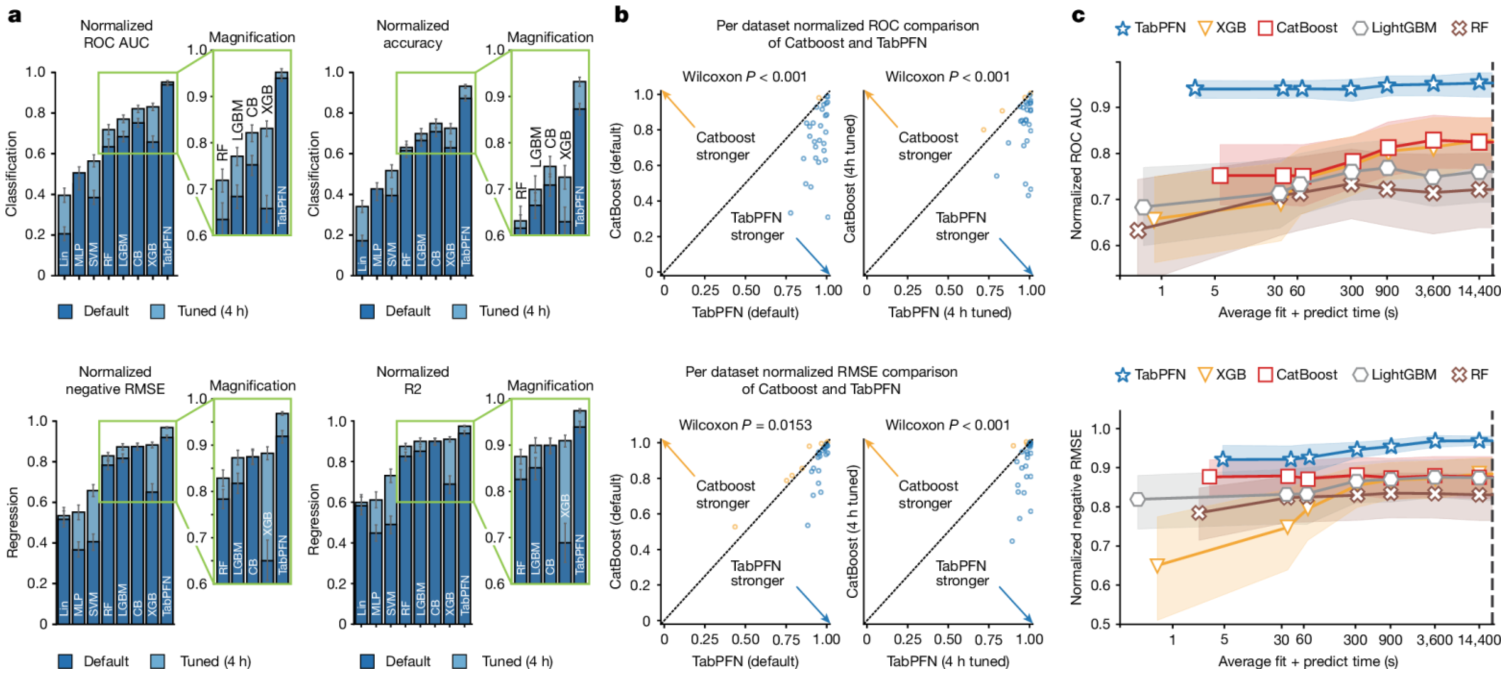
Photon Double-slit Experiment



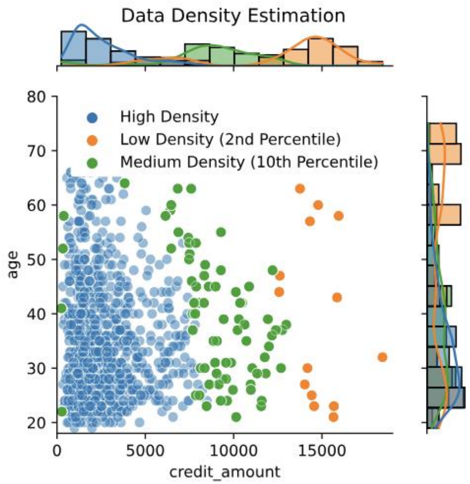
Light intensity pattern in a double-slit experiment after observing the positions of 1 000 photons.

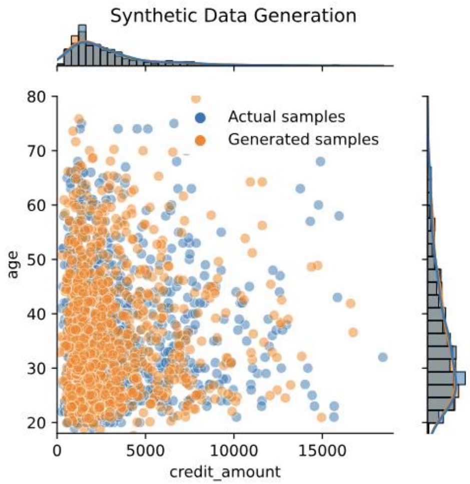
Results

Aggregated Benchmark Results

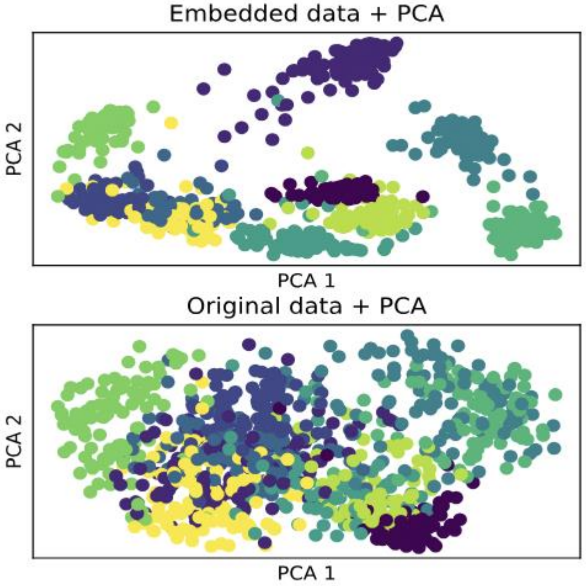


Foundation Model Abilities: data density estimation and generation of new synthetic samples on the German Credit Dataset



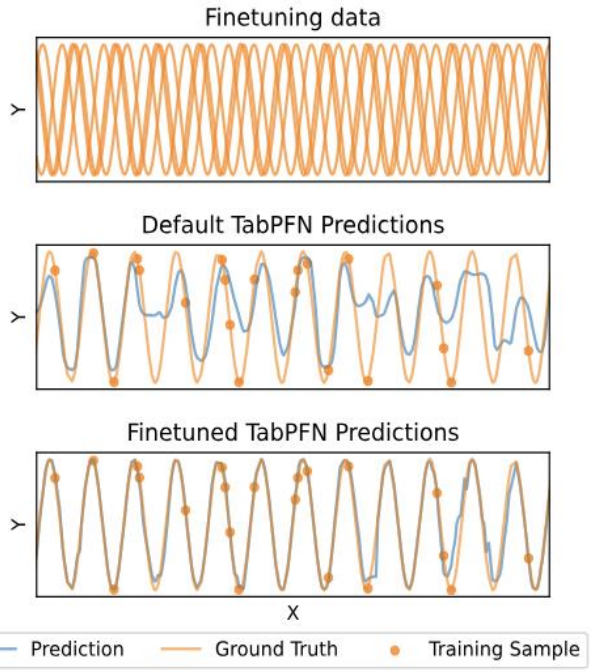


Foundation Model Abilities: learned embeddings are useful representations of each sample on the handwritten digits dataset



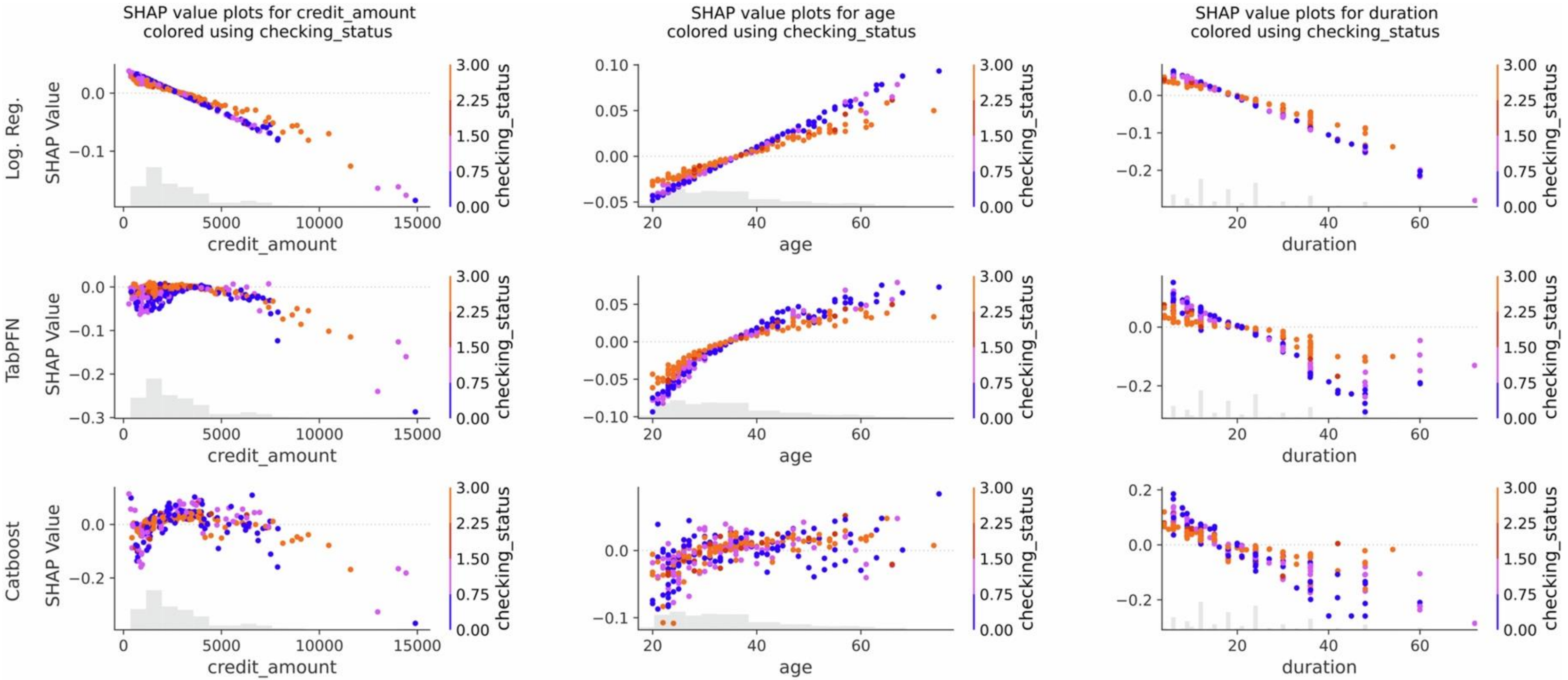
Results

Foundation Model Abilities: Finetuned on a dataset containing sine curves (top row), the model makes more accurate predictions on another sine curve dataset.



Interpreting non-linear models

Shapley Values

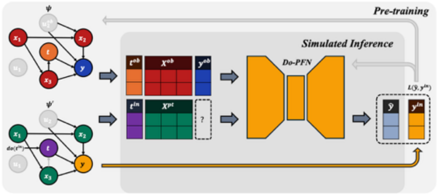


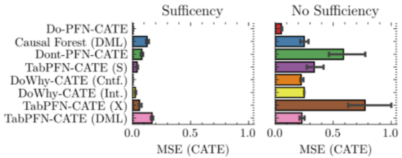
Predicting the effects of causal interventions

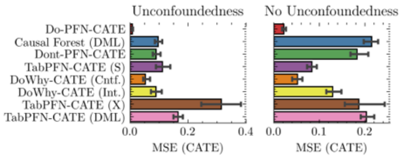
DO-PFN: IN-CONTEXT LEARNING FOR CAUSAL EFFECT ESTIMATION

Jake Robertson*[1,4], Arik Reuter*[2], Siyuan Guo[2,3], Noah Hollmann[5], Frank Hutter†[4,1,5], Bernhard Schölkopf†[2,1]

[1]ELLIS Institute Tübingen, Tübingen, Germany
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3]University of Cambridge, Cambridge, United Kingdom
[4]University of Freiburg, Freiburg, Germany
[5]Prior Labs, Freiburg, Germany



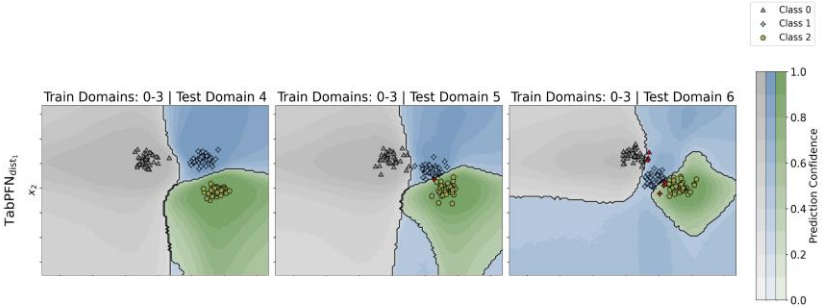


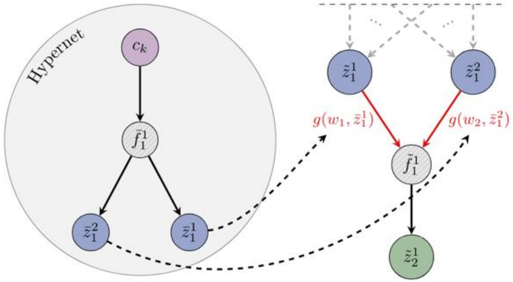


Robustness under distribution shift

Drift-Resilient TabPFN: In-Context Learning Temporal Distribution Shifts on Tabular Data

Kai Helli[*,1,2] **David Schnurr**[*,1,3] **Noah Hollmann**[1,4] **Samuel Müller**[1] **Frank Hutter**[5,1]
[1] University of Freiburg, [2] Technical University of Munich, [3] ETH Zurich,
[4] Charité University Medicine Berlin, [5] ELLIS Institute Tübingen, [*] Equal contribution.
Correspondence to kai.helli@tum.de





Looking forward to get in touch!

Website

priorlabs.ai

Contact

noah@priorlabs.ai