STAT 992: Foundation Models for Biomedical Data

Ben Lengerich

Lecture 15: Foundation Models & Tabular Data

April 13, 2026

Previously: Imaging, Genomics, EHRs

Foundation models succeed when the **data admit a canonical representation with exploitable structure**.

- Inductive biases:
  - Imaging → Locality + Translation Invariance → Convolutions / ViTs
  - Natural Language → Sequential Structure → Transformers (Attention)
  - Genomics → Discrete Sequence + Motifs + Long-range Dependencies → Transformers / Sequence Models
  - EHRs → Temporal + Multimodal Event Sequences → Sequence Models / Transformers

Today: Foundation Models & Structured Data

- Let's consider **structured data**:
  - relational tables
  - knowledge graphs
  - databases

In structured data, **semantics are relational and schema-dependent**, not intrinsic to the tokens.

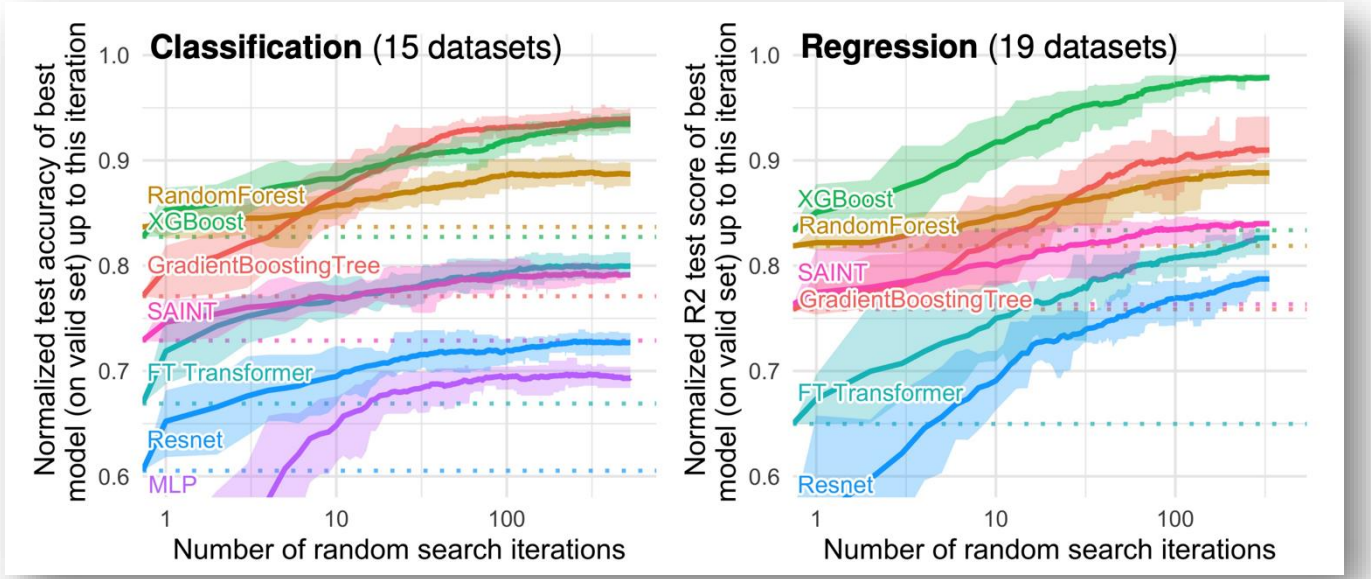
- In images, a pixel means the same thing everywhere.
- In language, a word means roughly the same thing across documents.
- **In tabular data, a column means something different in every dataset.**

The challenge of Tabular Data

- Key properties:
  - heterogeneous columns
  - permutation invariance over rows
  - weak or unknown structure
- Ironically, structured data used to be easier than unstructured data.
  - Can bake in statistical inductive biases to parametric models:
    - Montonicity
    - Sparsity
    - Additivity
    - etc.

The challenge of Tabular Data

Why do tree-based models still outperform deep learning on tabular data? [Grinsztajn et al. (2022)]



- Three answers:
  - **Smooth vs. irregular functions** – Neural networks improve when labels are smoothed, indicating a bias toward smooth functions; tree ensembles handle irregular, piecewise patterns better.
  - **Uninformative features** – Neural networks are more sensitive to noisy/irrelevant features, while trees are more robust due to local feature selection at splits.
  - **Feature representation / rotations** – Neural networks' approximate rotation invariance can hurt performance when feature axes have semantic meaning; trees benefit from axis-aligned splits that respect this structure.

Competing Paradigms for Tabular Foundation Models

Competing Paradigms for Tabular Data

- How should a model represent a dataset D to answer queries?

(x,D)→y

- Paradigm 1: Parameterized Models

$$\theta(D) = \text{argmax}_\theta \sum_i \log p(y_i \mid x_i, \theta)$$

- Paradigm 2: Tabular Foundation Models

$$p(\, y \mid x, D\,) \approx f_\phi(D, x)$$

- Paradigm 3: LLM + Tools

$$\hat{y} = LLM\left(x, Tool(D)\right)$$

- Paradigm 4: Structured Representations

$$S(D) = \{s_1(D), \ldots, s_k(D)\}, \qquad p(\, y \mid x, D\,) = g\big(x, S(D)\big)$$

Parameterized Models

- Paradigm 1: Parameterized Models

$$\theta(D) = \text{argmax}_\theta \sum_i \log p(y_i \mid x_i, \theta)$$

- **Strengths:**
  - strong inductive bias
  - efficient
  - interpretable
- **Limits:**
  - fixed structure
  - limited compositional reuse

Tabular Foundation Models

- Paradigm 2: Tabular Foundation Models

$$p(y \mid x, D) \approx f_\phi(D, x)$$

- **Strengths:**
  - prior over datasets
  - no training per dataset
  - single forward pass
- **Limits:**
  - no decomposition
  - no reuse across queries
  - limited interpretability

TabPFN

PRIOR

Fundamental

A good resource:
https://tabular-foundation.christophmolnar.com/

LLM + Tools

- Paradigm 3: Tabular Foundation Models

$$\hat{y} = LLM\left(x, Tool(D)\right)$$

- Example Tools:
  - text-to-SQL
  - Python
  - agents
- **Strengths:**
  - flexible
  - general
- **Limits:**
  - inefficient
  - brittle
  - no memory

Structured Representations

- Paradigm 4: Structured Representations

$$S(D) = \{s_1(D), \ldots, s_k(D)\}, \qquad p(\, y \mid x, D\,) = g\big(x, S(D)\big)$$

- Example Representations:
  - $\mathbb{E}[y \mid x_j]$
  - subgroup statistics
  - feature effects
  - interactions
- **Strengths:**
  - Compositional reuse across queries
  - Interpretability
  - Amortized compute
- **Limits:**
  - Potential information loss, Difficult systems design

Tradeoffs

Method	Efficiency	Flexibility	Compositionality
Classical	high	low	medium
TabPFN	high	medium	low
LLM tools	low	high	low
Structured	medium	medium	high

The big picture: Foundation Models & Tabular Data

- Foundation models succeed when meaning is shared across is shared.
- Tabular data requires meaning to be *constructed*.
- **Takeaways:**

1. No canonical representation. Unlike images or language, tabular datasets do not share a common structure.

2. Different paradigms = different representations of data: statistical models (trees, NN), weights (TabPFN-style), computation (LLM + tools), structure (explicit summaries)

3. No dominant solution (yet). The right abstraction for tabular data is still an open problem.

Questions?