Methods & Modeling

Day 4 — Every choice for a reason

MATTHEW MCDERMOTT · COLUMBIA · JUNE 25, 2026

Who am I

I work on two sides of health AI: the data infrastructure beneath it, and representation learning on top.

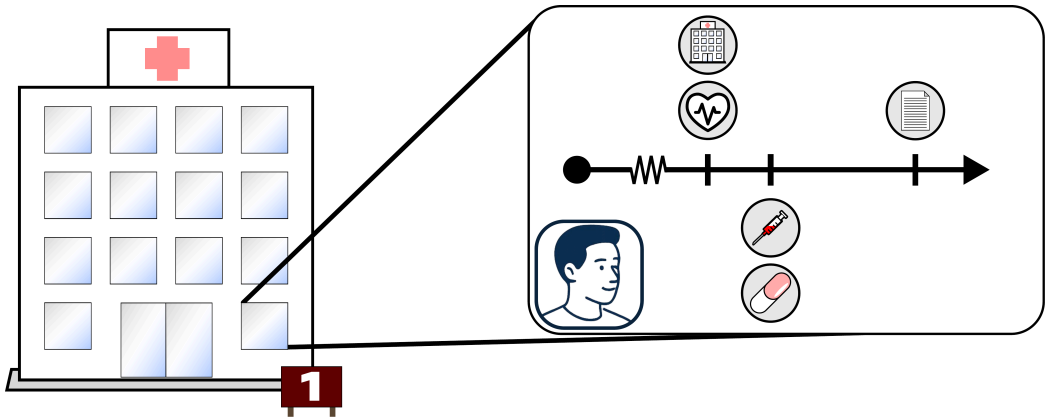
- Assistant Professor, Columbia DBMI; PhD at MIT with Pete Szolovits.
- Data infrastructure — MEDS, an open standard for the structure of health records (complementary to OMOP), and MEDS-DEV, a reproducible cross-dataset benchmark.
- Representation learning — structure-inducing pre-training and EveryQuery.
- A throughline: reproducibility in ML for health.

📣 **I'm hiring — recruiting PhD students & postdocs for my lab at Columbia DBMI. Come find me.**

We'll assume your data is an event stream

A patient is a **longitudinal stream of timestamped, typed events** — "MEDS-like."

- The **root form** of medical data: labs, meds, vitals, notes, images are all just *events with a time*.
- Carries **arbitrary modalities** — even ones we can't use yet — without throwing them away.
- Keeps the **generating process** in view, which is what lets us reason about choices.



An irregular, multimodal event stream — admissions, vitals, notes, meds.

AHLI HEALTH AI SUMMER CAMP 2026

FIRST — WHAT ARE WE EVEN DESIGNING?

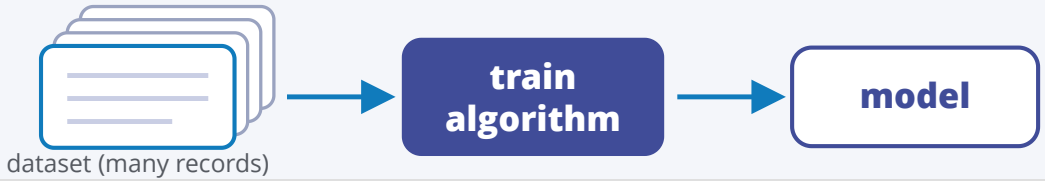
Model vs. algorithm

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Your algorithm is not your model

algorithm

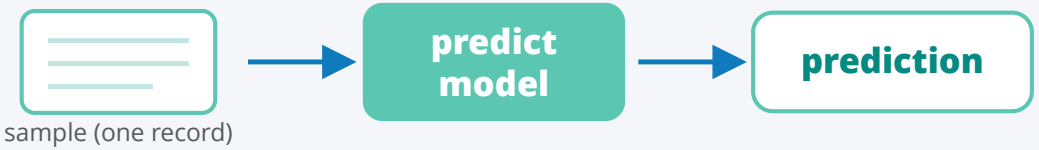
```
train(dataset, config) → model
```



- the shape of the input changes the API

model

```
predict(sample) → prediction
```



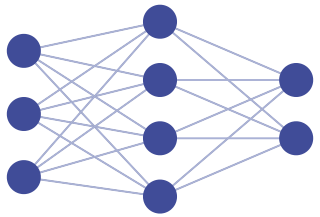
What must ship with the model?

`predict(sample)` only ever sees **one sample** — so anything the model needs must travel **inside** it.

What must ship with the model?

`predict(sample)` only ever sees **one sample** — so anything the model needs must travel **inside** it.

Every model

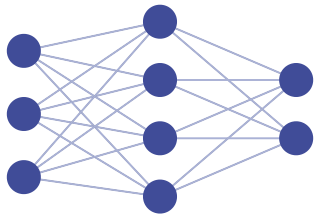


the weights

What must ship with the model?

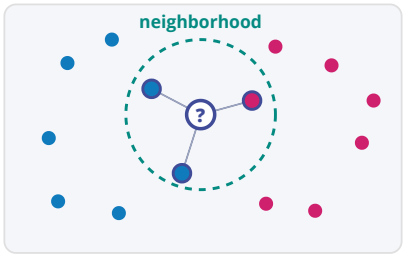
`predict(sample)` only ever sees **one sample** — so anything the model needs must travel **inside** it.

Every model



the weights

k-NN / retrieval

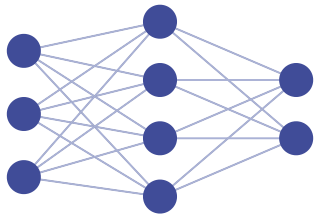


+ reference set

What must ship with the model?

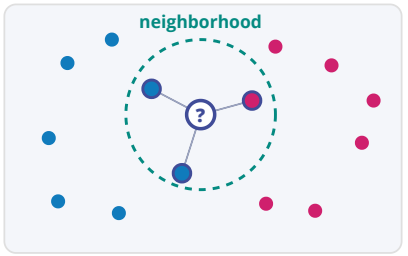
`predict(sample)` only ever sees **one sample** — so anything the model needs must travel **inside** it.

Every model



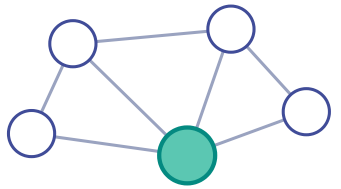
the weights

k-NN / retrieval



+ reference set

Population graph



+ graph neighborhood

AHLI HEALTH AI SUMMER CAMP 2026

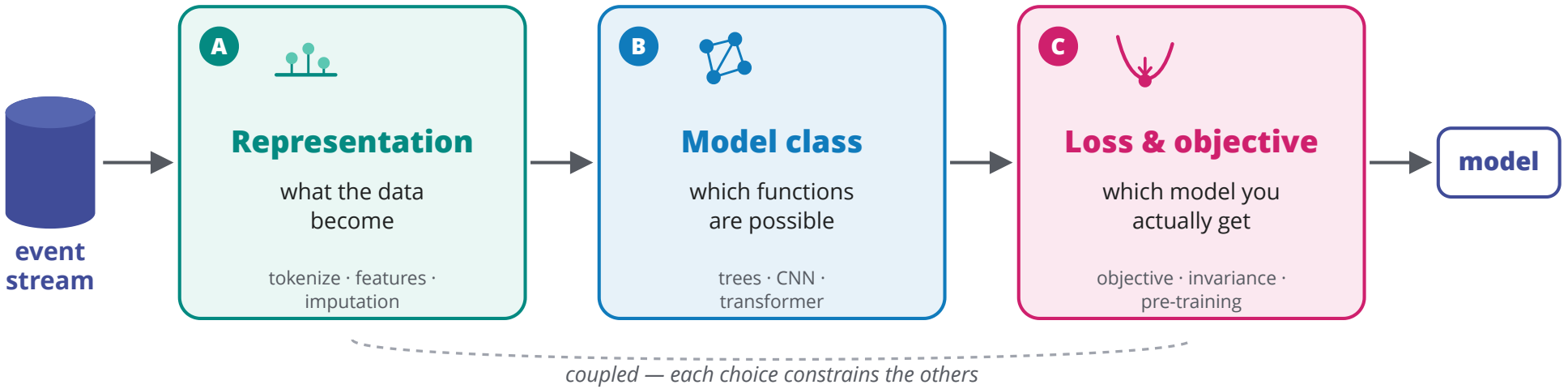
DESIGNING A MODEL =

Three coupled choices

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Three coupled choices

Data representation · model class · loss/objective — **not independent**: each constrains the others.



For **every** choice — representation, model class, loss — you should be able to say **why**, and the ***why*** should depend on your **data** and **problem**. Methodological work is **justifying, testing, and showing** those choices.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Honestly? We can't yet say which choices matter

Ask an expert *"what should I train for **my** EHR population and task?"* — the honest answer is **"it depends,"** and we mostly **can't say on what**.[1]

- Health data is high-dimensional, biasedly sampled, and **every site is unique.**
- So principled, data-dependent design is more **aspiration** than established science.

That gap is the opportunity — and the reason to **test** your choices, not assume them.

[1] McDermott, M. "The (lack of?) science of machine learning for healthcare." *ML4H* 2024 (PMLR 259:19–29).

A · Representation — expose the signal, or bury it

The right representation does the model's work for it; the wrong one hides or fakes the signal.

A · Representation — expose the signal, or bury it

The right representation does the model's work for it; the wrong one hides or fakes the signal.

- **Classic win:** a **spectrogram** turns a raw waveform into a time–frequency image where the pattern is separable (ECG/EEG); **eGFR** normalizes creatinine by age & sex so one number means one thing.

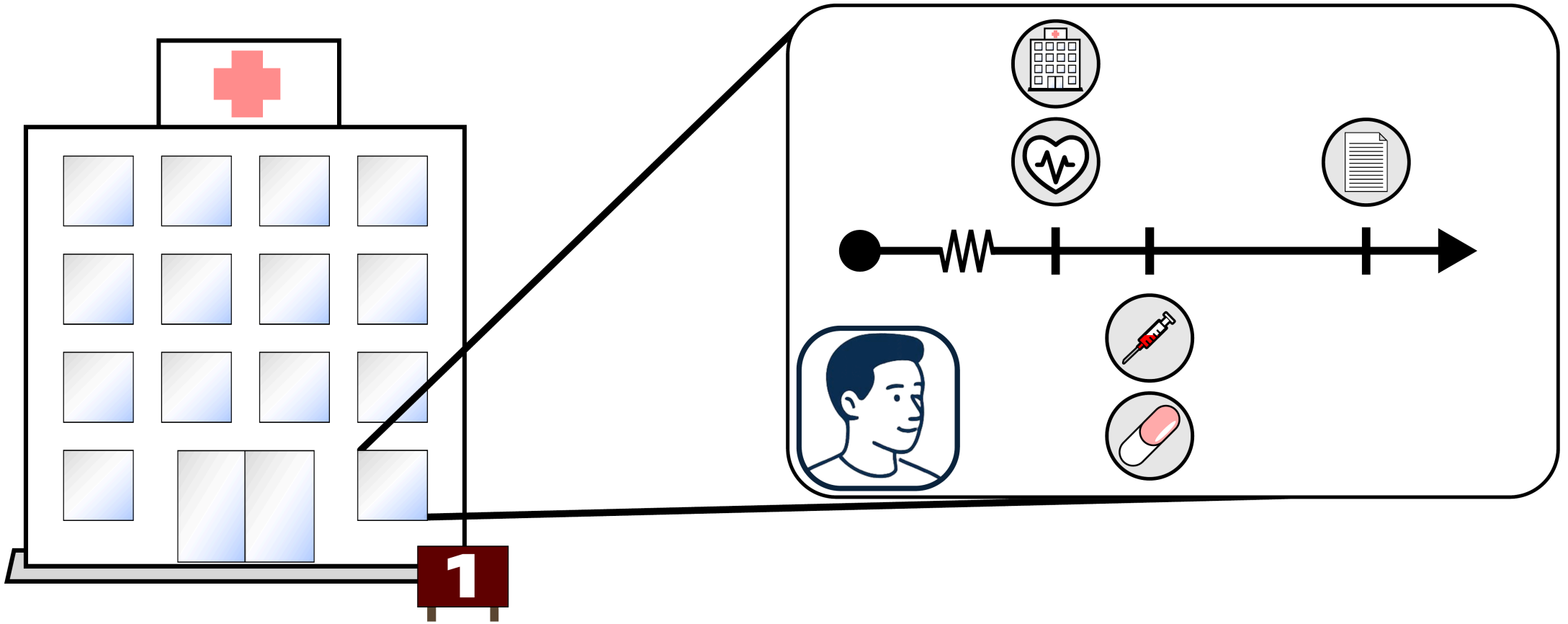
A · Representation — expose the signal, or bury it

The right representation does the model's work for it; the wrong one hides or fakes the signal.

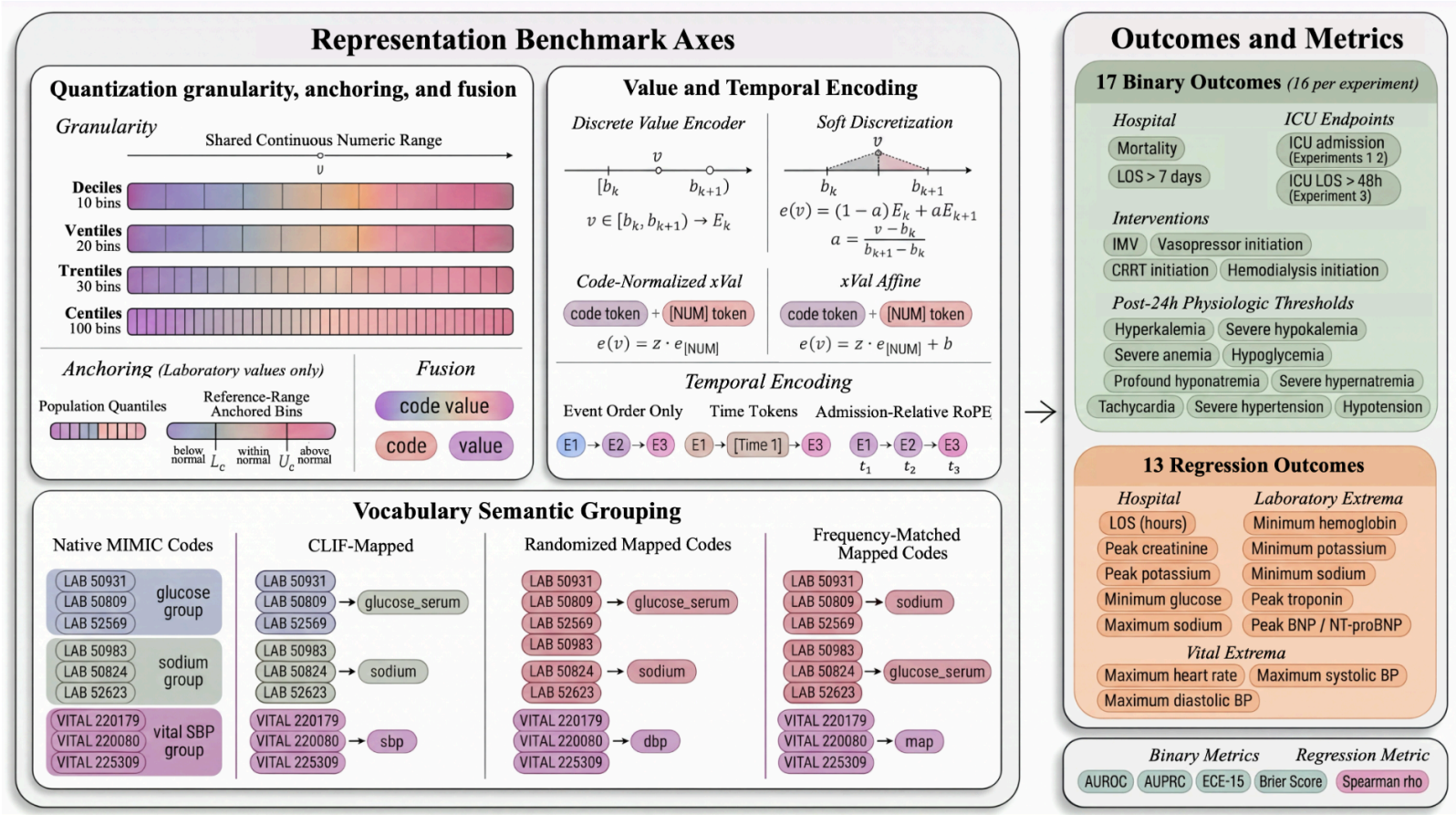
- **Classic win:** a **spectrogram** turns a raw waveform into a time–frequency image where the pattern is separable (ECG/EEG); **eGFR** normalizes creatinine by age & sex so one number means one thing.
- **Pitfall:** mean-impute a missing lab and you **discard informative missingness** — *when* a test was ordered can predict outcomes better than its value.[1]

[1] Agniel, D., Kohane, I. S. & Weber, G. M. "Biases in electronic health record data due to processes within the healthcare system." *BMJ* 361 (2018).

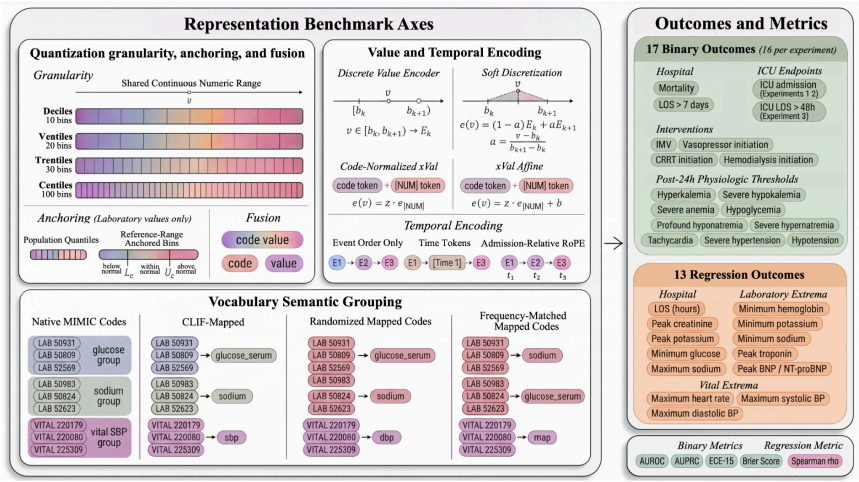
How should you *tokenize* the event stream?



How should you *tokenize* the event stream?



How should you *tokenize* the event stream?



Turning the event stream into the token sequence your model reads is itself a representation choice — a very active area:

- **Granularity, anchoring & fusion** — how finely to bin values, whether to anchor to the reference range, whether to fuse code with value.
- **Value & temporal encoding** — discrete vs. soft vs. continuous encoders; event order vs. time tokens vs. relative position.
- **Vocabulary grouping** — native codes vs. semantic groupings of labs & vitals.

[1] Lee et al. "Representation Before Training: A Fixed-Budget Benchmark for Generative Medical Event Models." *MLHC* 2026 (arXiv:2604.16775). [2] Guo et al. "Tokenization Tradeoffs in Structured EHR Foundation Models." arXiv:2603.15644 (2026). [3] Montgomery & Nielsen. "From Binning to Joint Embeddings." MLRH workshop 2025.
[4] Shickel et al. "Multi-dimensional patient acuity estimation with longitudinal EHR tokenization." *Front. Digit. Health* 2022. [5] Al Attrach et al. "Rethinking Tokenization for Clinical Time Series: When Less is More." *ML4H* 2025.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

B · Model class — match the inductive bias

The class you pick is a **claim about the data's structure** — make it on purpose.

B · Model class — match the inductive bias

The class you pick is a **claim about the data's structure** — make it on purpose.

- **Win:** on tabular data, **gradient-boosted trees** often beat deep nets — neural nets are hurt by uninformative features and non-smooth targets, and tabular data isn't rotation-invariant.[1] **CNNs** win on images by baking in translation-equivariance.

B · Model class — match the inductive bias

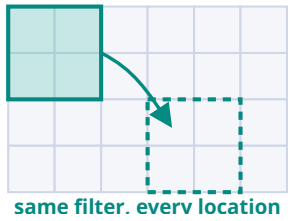
The class you pick is a **claim about the data's structure** — make it on purpose.

- **Win:** on tabular data, **gradient-boosted trees** often beat deep nets — neural nets are hurt by uninformative features and non-smooth targets, and tabular data isn't rotation-invariant.[1] **CNNs** win on images by baking in translation-equivariance.
- **Pitfall:** reach for a deep net on tabular *because it's fashionable* → it loses.

The reason is the bias, not the brand: a class wins when its assumptions match the data's structure.

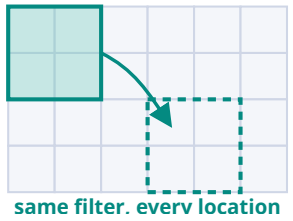
[1] Grinsztajn, L., Oyallon, E. & Varoquaux, G. "Why do tree-based models still outperform deep learning on typical tabular data?" *NeurIPS* 2022.

Quick hits: the architecture *is* an inductive bias

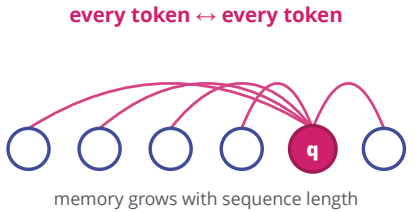


CNNs · regular grids — one filter slides everywhere (**translation equivariance**); a pointwise multiply in **Fourier** space.

Quick hits: the architecture *is* an inductive bias

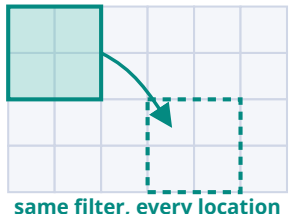


CNNs · regular grids — one filter slides everywhere (**translation equivariance**); a pointwise multiply in **Fourier** space.

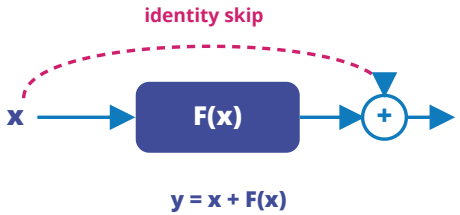


Transformers — every token directly attends to all others: a length-scaling memory, no fixed **bottleneck**.

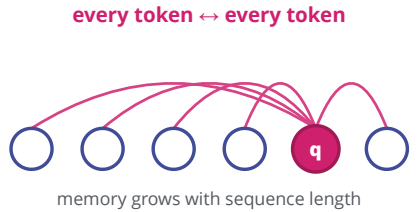
Quick hits: the architecture *is* an inductive bias



CNNs · regular grids — one filter slides everywhere (**translation equivariance**); a pointwise multiply in **Fourier** space.

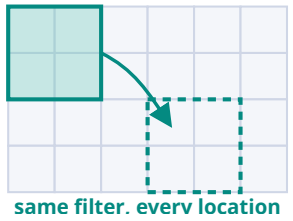


ResNets — `y = x + F(x)`: the skip eases optimization; layers learn a **correction**, not the whole map.

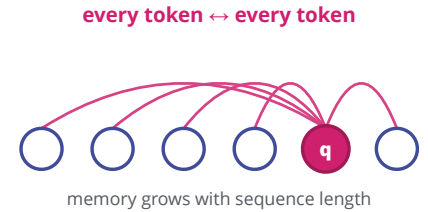


Transformers — every token directly attends to all others: a length-scaling memory, no fixed **bottleneck**.

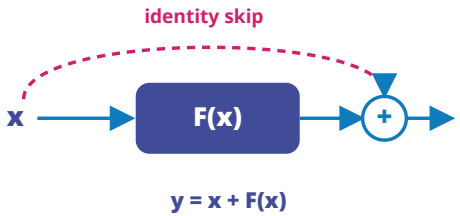
Quick hits: the architecture *is* an inductive bias



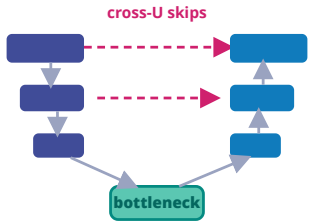
CNNs · regular grids — one filter slides everywhere (**translation equivariance**); a pointwise multiply in **Fourier** space.



Transformers — every token directly attends to all others: a length-scaling memory, no fixed **bottleneck**.



ResNets — `y = x + F(x)` : the skip eases optimization; layers learn a **correction**, not the whole map.



U-Nets — cross-U skips carry **high-res detail** across the bottleneck → sharp **segmentation**.

C · Loss & objective — encode an inductive bias

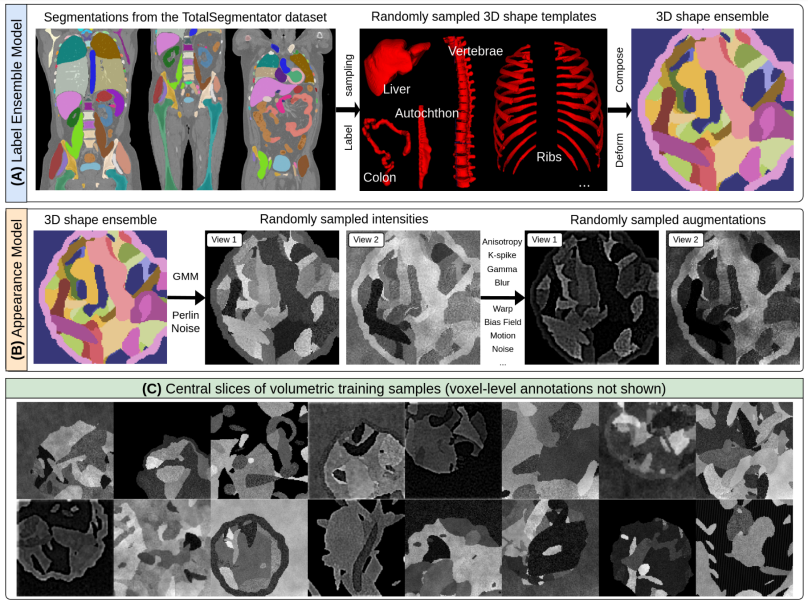
Your objective encodes an **inductive bias** about your task and data — so you should be able to **justify why that bias is likely to help.**

- **Win:** pick a bias you can defend — *shape, not appearance* for imaging; a *pre-training graph* for relational data. *(next two slides)*
- **Pitfall:** a fashionable objective that doesn't match your data may not stand the test of time.

Loss example: *shape, not appearance*

For imaging, **Dey et al. (Golland lab)** pretrain so features depend on **anatomical shape**, not intensity or appearance.

- A **data engine** composes random shape ensembles, then renders each with **randomized appearance** — intensities, noise, artifacts.
- A **dense contrastive loss** pulls together features of the *same structure* across appearances → the objective **is** the shape-invariance claim.
- Yields general 3D features that transfer to registration & few-shot segmentation **without any real pre-training data.**



Randomized synthesis: shared shapes, randomized appearance.

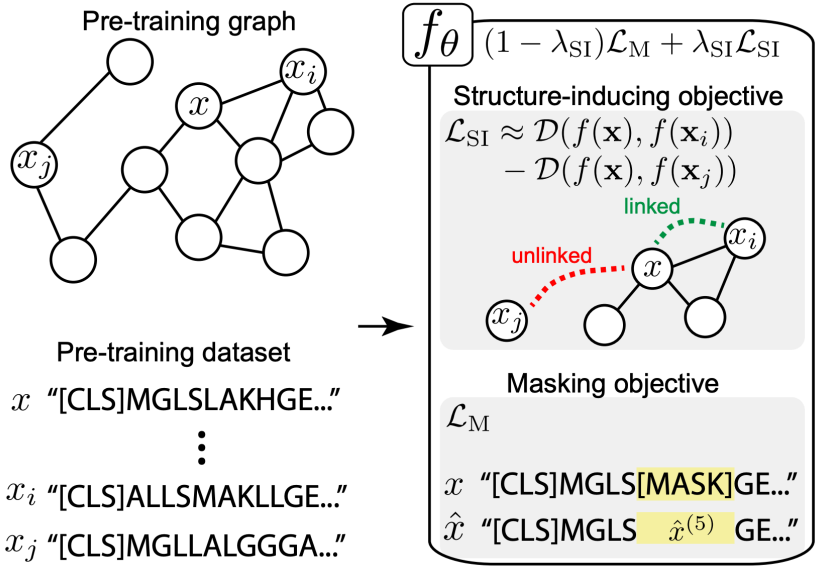
[1] Dey, N., ..., Golland, P. "Learning general-purpose biomedical volume representations using randomized synthesis." *ICLR* 2025.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Loss example: *a pre-training graph* (SIPT)

For relational data, **structure-inducing pre-training (SIPT)** adds a term so the **embedding geometry matches a known graph** of relationships.

- Pre-training can be read as inducing a **structure over samples** — SIPT makes that structure an explicit **target**.
- You **write the relational prior into the loss** instead of hoping the model discovers it.
- Tighter latent structure → better downstream transfer where that structure matters.



SIPT: a pre-training graph shapes the embedding space.

[1] McDermott, M., Yap, B., Szolovits, P. & Zitnik, M. "Structure-inducing pre-training." *Nature Machine Intelligence* 5 (2023).

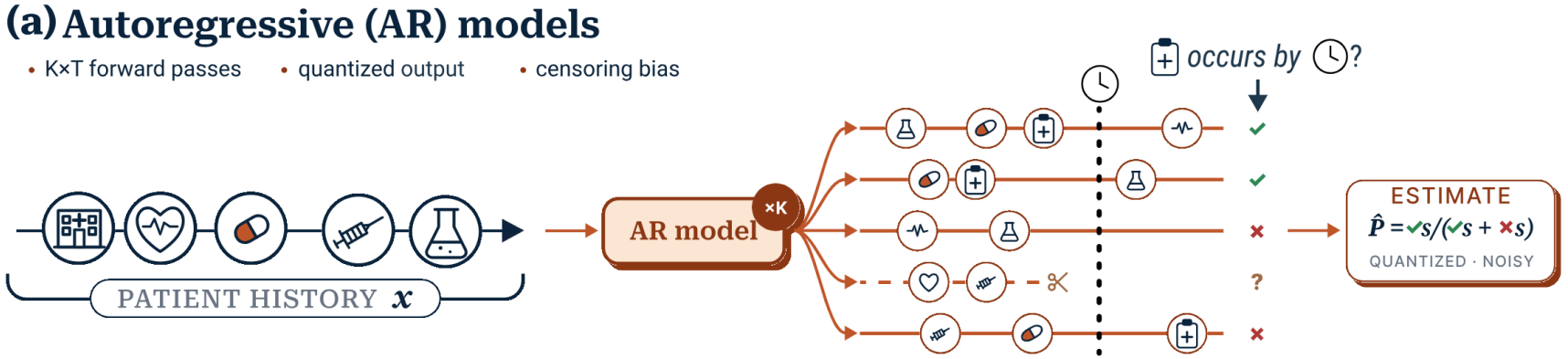
My favorite example: MIMIC wasn't "saturated"

For years we called **MIMIC** and the ICU benchmarks *saturated*. We were wrong — only recently did we find we can build **genuine foundation models** on exactly that data.[1,2]

[1] Renc, P., et al. ETHOS / ARES — generative EHR timelines. *npj Digital Medicine* 2024; *GigaScience* 2025. [2] Waxler et al. "Generative medical event models improve with scale" (Epic Cosmos / Curiosity). arXiv:2508.12104 (2025).

My favorite example: MIMIC wasn't "saturated"

For years we called **MIMIC** and the ICU benchmarks *saturated*. We were wrong — only recently did we find we can build **genuine foundation models** on exactly that data.[1,2]



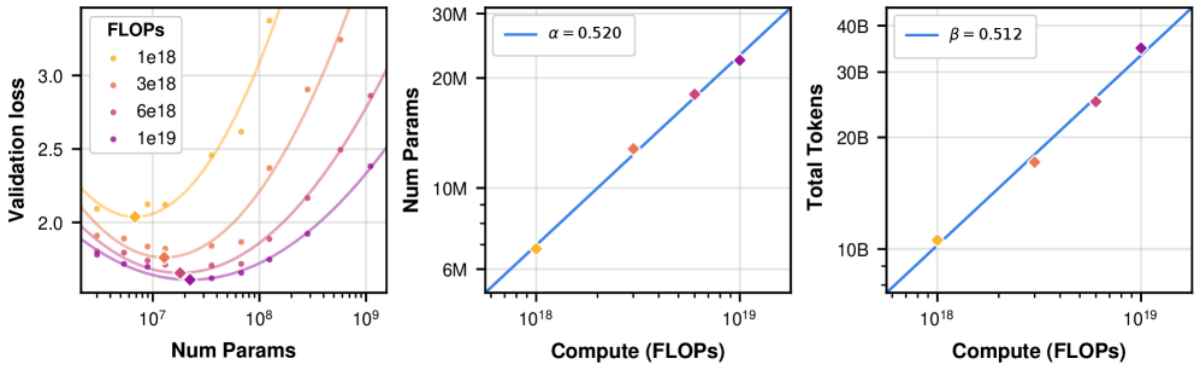
How they work: autoregressive FMs sample many full timelines, then estimate the answer.

[1] Renc, P., et al. ETHOS / ARES — generative EHR timelines. *npj Digital Medicine* 2024; *GigaScience* 2025. [2] Waxler et al. "Generative medical event models improve with scale" (Epic Cosmos / Curiosity). arXiv:2508.12104 (2025).

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

My favorite example: MIMIC wasn't "saturated"

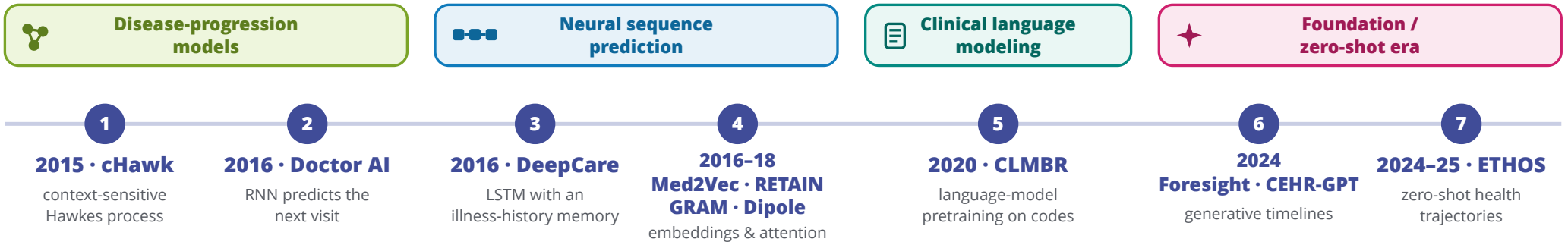
For years we called **MIMIC** and the ICU benchmarks *saturated*. We were wrong — only recently did we find we can build **genuine foundation models** on exactly that data.[1,2]



[1] Renc, P., et al. ETHOS / ARES — generative EHR timelines. *npj Digital Medicine* 2024; *GigaScience* 2025. [2] Waxler et al. "Generative medical event models improve with scale" (Epic Cosmos / Curiosity). arXiv:2508.12104 (2025).

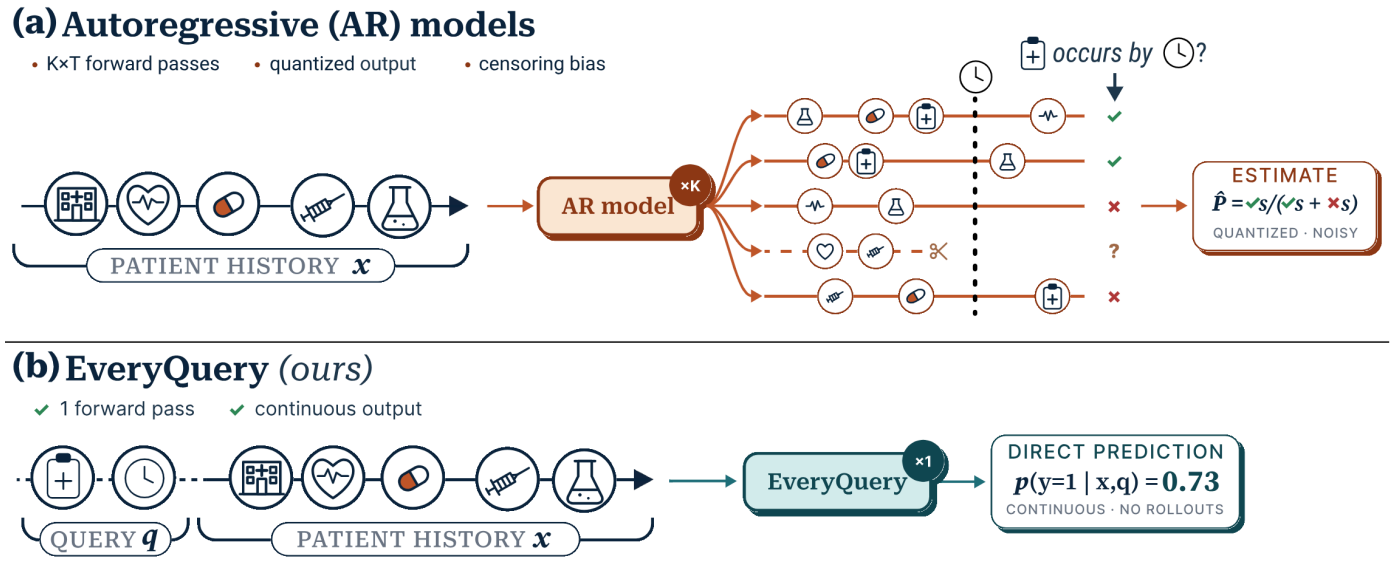
AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Evolution of autoregressive EHR modeling



Maybe we're not done — EveryQuery

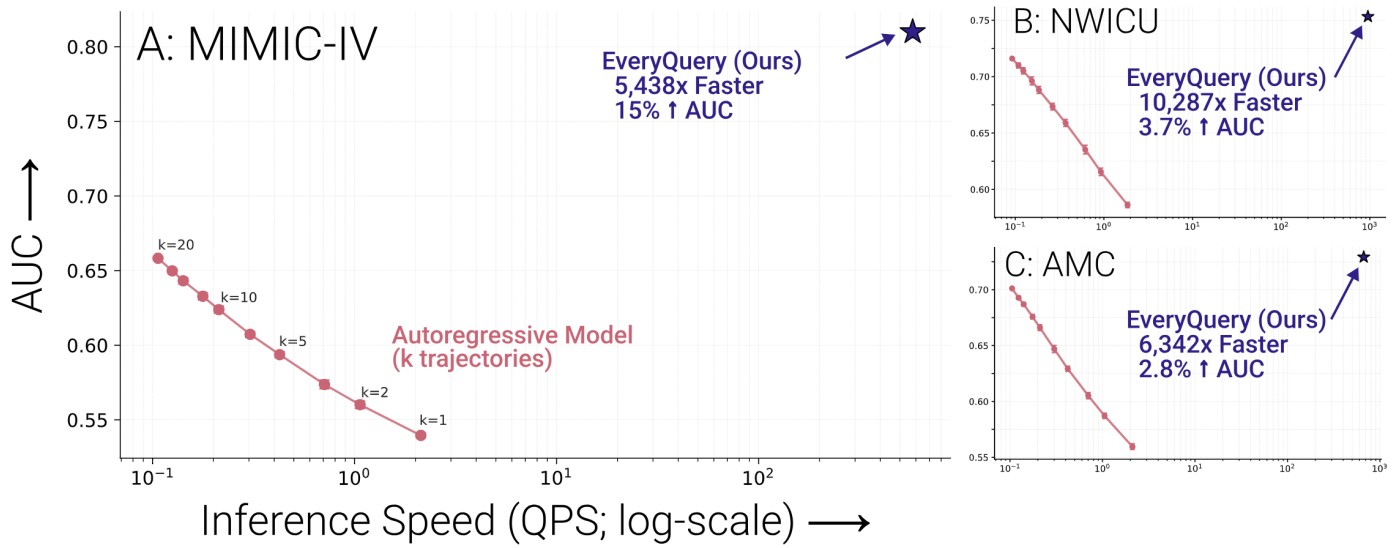
The whole timeline rests on one **inductive bias** — the **temporal structure** of EHR data as the defining signal for method development. Maybe we're **not done** exploiting it: **EveryQuery** (ours) swaps autoregressive rollouts for a single, query-conditioned **direct prediction**.



[1] Chandak, P., Kondas, G., Antwarg Friedman, L., Kohane, I. & McDermott, M. "EveryQuery: Zero-Shot Clinical Prediction via Task-Conditioned Pretraining over EHRs." arXiv:2603.07900 (2026).

Maybe we're not done — EveryQuery

The whole timeline rests on one **inductive bias** — the **temporal structure** of EHR data as the defining signal for method development. Maybe we're **not done** exploiting it: **EveryQuery** (ours) swaps autoregressive rollouts for a single, query-conditioned **direct prediction**.



[1] Chandak, P., Kondas, G., Antwarg Friedman, L., Kohane, I. & McDermott, M. "EveryQuery: Zero-Shot Clinical Prediction via Task-Conditioned Pretraining over EHRs." arXiv:2603.07900 (2026).

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

AHLI HEALTH AI SUMMER CAMP 2026

WE CAN'T JUST ASSUME

How do we know we're making good choices?

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Think stupider — then test

To know whether a choice matters, **shrink the problem until you can see it.**

- **Simplify the task:** make it binary; condition away censoring.
- **Synthesize the data:** plant the property you think matters (e.g. informative missingness) — or remove it.
- **Simplify the model:** make it stupid-simple so the *effect* is unambiguous.
- **Break it on purpose:** build cases that *can't* work or *can't fail*, then check they do.

A good synthetic experiment

A good synthetic experiment is **so clear that, if it doesn't work, there's a bug** — you plant the property you're testing *by construction*, so the right answer is known in advance.

- Build the data so the effect **must** appear; a stripped-down model should recover it.
- If it doesn't, the problem is in **your code**, not the world — a tight, fast debugging loop.
- And if you **can't design such an experiment**, that's the real signal: you probably **don't yet fully understand why your method should work** — and fixing *that* is the actual work.

Tests you can re-run

Generate **controlled cohorts where the answer is known**, assert **behavioral invariants**, and re-run them on every change:

- **EveryQuery** · `tests/training_validity` — plant a signal *by construction*, then assert the model **recovered** it.
- **MEDS-EIC-AR** · `tests/grammar` — train on a tiny **grammar**, then assert generations **stay grammar-valid**.

Re-running these turns "we can't be sure" into **behavior you've pinned down — and kept pinned.**

EveryQuery

MEDS-EIC-AR

Your turn — build a synthetic-experiment notebook

> **Deliverable:** pick **one** design choice for your project — a **representation**, a **model class**, or a **loss** — and build a **synthetic experiment** where the property you're betting on is present (or absent) *by construction*. Then **show** whether your choice helps.

Name the property → make the choice → test it → show the result.

Next: 1:30 guest (Walter Gerych) — trustworthy-by-construction · 2:45 build session · 4:15 share your experiment.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

References — autoregressive EHR models

1. **cHawk** — Choi, Du, Chen, Song, Sun. "Constructing Disease Network and Temporal Progression Model via Context-Sensitive Hawkes Process." *ICDM* 2015.
2. **Doctor AI** — Choi et al. "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks." *MLHC* 2016.
3. **DeepCare** — Pham et al. "DeepCare: A Deep Dynamic Memory Model for Predictive Medicine." *PAKDD* 2016.
4. **Med2Vec** — Choi et al. "Multi-layer Representation Learning for Medical Concepts." *KDD* 2016.
5. **RETAIN** — Choi et al. "RETAIN: An Interpretable Predictive Model via Reverse Time Attention." *NeurIPS* 2016.
6. **GRAM** — Choi et al. "GRAM: Graph-based Attention Model for Healthcare Representation Learning." *KDD* 2017.
7. **Dipole** — Ma et al. "Dipole: Diagnosis Prediction via Attention-based Bidirectional RNN." *KDD* 2017.
8. **CLMBR** — Steinberg et al. "Language models are an effective representation learning technique for EHR data." *J. Biomed. Inform.* 2021.
9. **EHRSHOT** — Wornow et al. "EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models." *NeurIPS Datasets & Benchmarks* 2023.
10. **Foresight** — Kraljevic et al. "Foresight — a generative pretrained transformer for modelling patient timelines using EHRs." *Lancet Digital Health* 2024.
11. **CEHR-GPT** — Pang et al. "CEHR-GPT: Generating EHRs with Chronological Patient Timelines." *arXiv* 2024.
12. **ETHOS** — Renc et al. "Zero-shot health-trajectory prediction using a transformer." *npj Digital Medicine* 2024.
13. **Curiosity / Epic Cosmos** — "Generative Medical Event Models Improve with Scale." *arXiv:2508.12104* (2025).