The afternoon breakout

Day 3 — simulate your evaluation *before* you model

EVALUATION & STUDY DESIGN · MATTHEW MCDERMOTT

Breakouts change gear today

- **Days 1–2:** discussion groups around a sub-modality of health data.
- **Days 3–4:** you **build** — a Colab / Jupyter notebook that simulates your **evaluation** (today) or your **methods** (tomorrow) setup.

If you'd like, drop your notebook into your folder in the **workbook GitHub**.

Why set up your evaluation first?

Design how you'll evaluate **before** you build the model. Doing it now helps you:

- **optimize for the right thing** as you develop,
- see where you're **vulnerable to noise** or to being **misled by a metric**,
- know up front **how to present and interpret** your results.

Methods are the easy case — evaluation is the interesting one

For **methods**, a synthetic experiment is usually straightforward: build fake data so the **factor of your model that matters is present (or absent) by construction.** *(More on this tomorrow — with examples.)*

For **evaluation**, "synthetic experiments" are trickier and far less common. Pick whichever of these **prototypes** fits your project:

1. You care about **X**, but work on **Y**
2. You can measure X — but only **noisily / through proxies**
3. The **interpretation** of X depends on hidden population properties
4. **Many metrics, many comparisons**

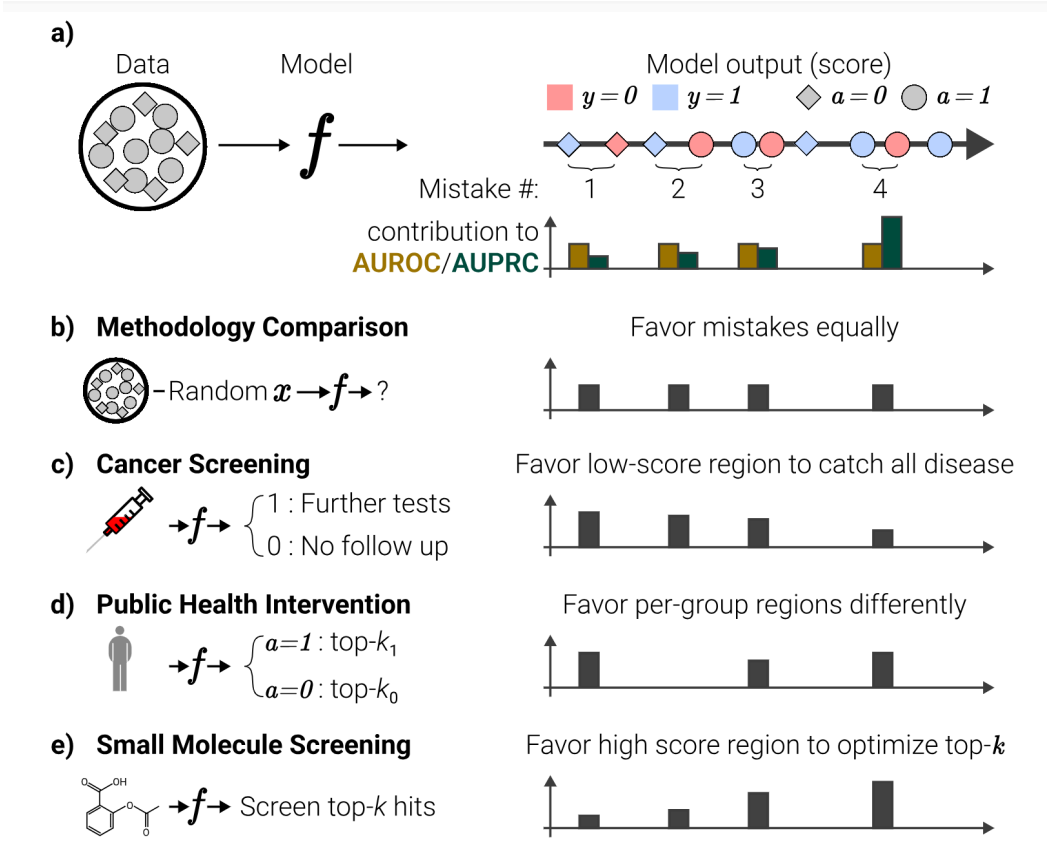
Prototype 1 · You care about X, but work on Y

You ultimately care about impact **X**; the problem in front of you is **Y**.

- **First, can you reframe?** Either *assume* Y is valuable because of X and just optimize Y — or work on **X** directly.
- **If not:** write a **simple mathematical model** of how Y turns into X.

Even a rough model **links what you can measure to what you care about** — it makes your assumptions explicit and checkable, and shows how much a gain in Y is actually worth.

Prototype 1, worked · which region of the score matters?



McDermott et al. "A Closer Look at AUROC and AUPRC under Class Imbalance." *NeurIPS* 2024.

Prototype 2 · You measure X, but only noisily / via proxies

You can measure what you care about — but **imperfectly**, through proxies that are vulnerable to noise.

Ask: **how sensitive is my evaluation to which kinds of noise?**

- **Simulate** different noise sources and see what flips your conclusion.
- Write a **sensitivity-analysis protocol** — and test it on *fake* model results.
- Track **additional proxy metrics** that help triangulate the true signal.

Prototype 3 · Interpretation depends on hidden population properties

You can measure X — but its **meaning** may hinge on properties of your population that aren't obvious.

Simulate different patient / sample properties, then ask:

- **How would my evaluation mislead me** in each case?
- **How would I detect** that it's happening?

Prototype 4 · Many metrics, many comparisons

You're comparing many options across many metrics, and need to **communicate and prioritize** them.

Less about simulating noise — more about **how you aggregate and report**:

- average metric across tasks? average **rank order**? **win-rate**?
- a **Pareto frontier** over competing metrics?

Decide **early** so you can tell when a model is *actually* winning.

Your task this afternoon

> Build a notebook that **simulates or sets up your evaluation**: pick the prototype(s) that fit, generate **fake data** or **fake model results**, and pressure-test how you'd measure success — *before* you start modeling.

Add it to your **workbook GitHub** folder if you'd like. We'll regroup to share what we found.