ICML 2026 Tutorial

Calibration, Decisions, and Collaboration in Learning

Natalie Collina
Penn → MIT

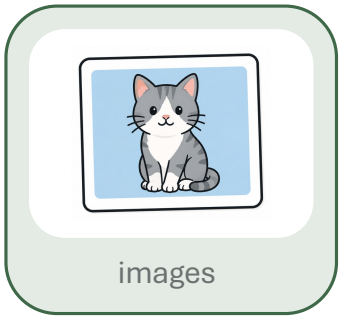
Ira Globus-Harris
Cornell

Aaron Roth
Penn

https://calibration-tutorial.github.io/

Machine learning is about making predictions

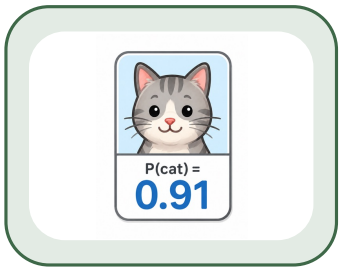
Given some context $\mathcal{X}$...



Machine learning is about making predictions

Given some context $\mathcal{X}$...

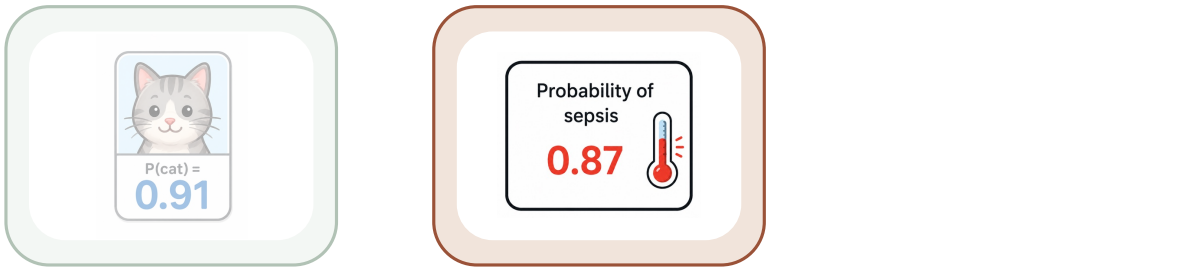
can we predict



...the probability the image depicts a cat?

Machine learning is about making predictions

Given some context $\mathcal{X}$...

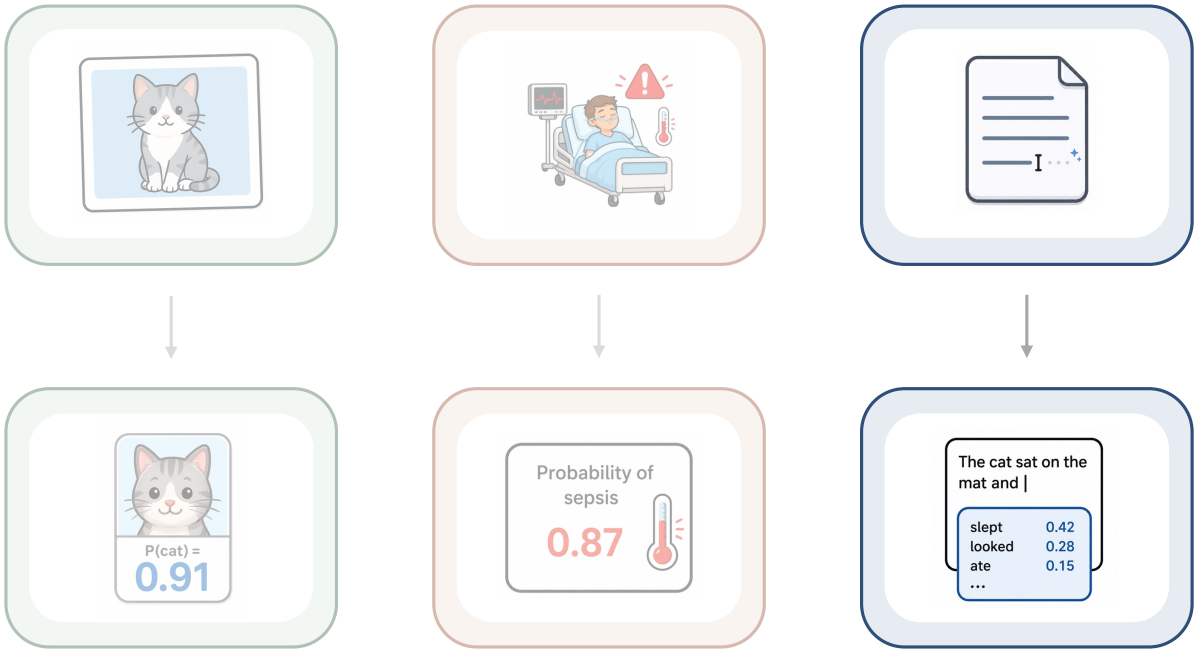


can we predict

...the probability the patient develops sepsis today?

Machine learning is about making predictions

Given some context $\mathcal{X}$...



can we predict

...the probability the next word in the text is “slept”?

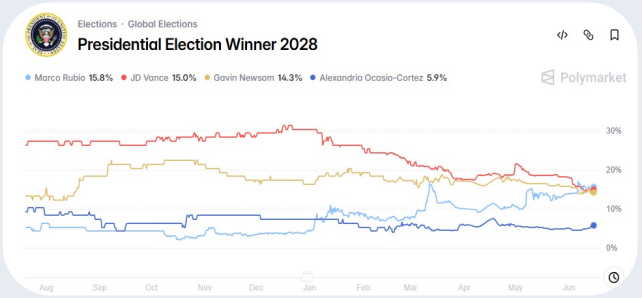
How can we make sense of probabilities?

What is the probability that if I flip a fair coin 16 times I get exactly 9 heads?

- We have a mathematical model that maps well onto reality; we can compute this in closed form.
- We can conduct the experiment repeatedly and empirically estimate probabilities.

How can we make sense of probabilities?

What is the probability that JD Vance will win the 2028 US Presidential Election?”



- If we posit a probabilistic model of the universe, this is perhaps philosophically coherent, but it is not a repeatable event; we can’t get empirical estimates.

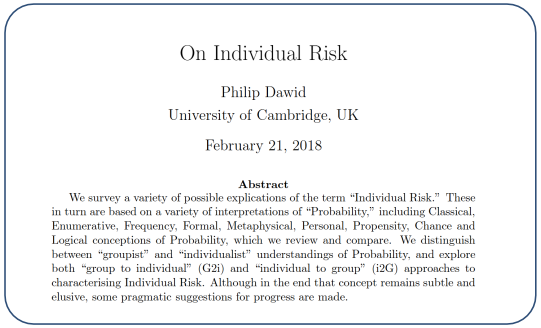
How can we make sense of probabilities?

- Many things we want to predict are closer to the 2nd example.
  - What is the probability of the next token given 50 pages of text?
    - There is no two extant texts with the same 50 page prefix…
  - What is the probability that Alice develops sepsis in the next 12 hours?
    - There are no two patients with exactly the same medical history and sequence of vital readings
  - What is the probability of rain in Seoul tomorrow?
    - There are no two days with exactly the same atmospheric readings…
  - …

How can we make sense of probabilities?

In general, no way to verify "**correctness**" of forecasts of "**individual probabilities**" and unclear they are even philosophically coherent.

This is the **Reference Class Problem** from Philosophy of Science, and it pervades machine learning.



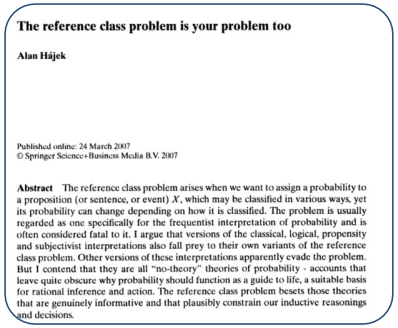
On Individual Risk

Philip Dawid
University of Cambridge, UK

February 21, 2018

Abstract

We survey a variety of possible explications of the term "Individual Risk." These in turn are based on a variety of interpretations of "Probability," including Classical, Enumerative, Frequency, Formal, Metaphysical, Personal, Propensity, Chance and Logical conceptions of Probability, which we review and compare. We distinguish between "groupist" and "individualist" understandings of Probability, and explore both "group to individual" (G2i) and "individual to group" (i2G) approaches to characterising Individual Risk. Although in the end that concept remains subtle and elusive, some pragmatic suggestions for progress are made.



The reference class problem is your problem too

Alan Hájek

Published online: 24 March 2007

Abstract The reference class problem arises when we want to assign a probability to a proposition (or sentence, or event) A, which may be classified in various ways, yet its probability can change depending on how it is classified. The problem is usually regarded as one specifically for the frequentist interpretation of probability and is often considered fatal to it. I argue that versions of the classical, logical, propensity and subjectivist interpretations also fall prey to their own variants of the reference class problem. Other versions of these interpretations apparently evade the problem. But I contend that they are all "no-theory" theories of probability - accounts that leave quite obscure why probability should function as a guide to life, a suitable basis for rational inference and action. The reference class problem besets those theories that are genuinely informative and that plausibly constrain our inductive reasonings and decisions.

So we can’t be sure we’ve learned “true probabilities”…

What are the properties of “true probabilities” that make them useful?

Maybe we can learn those.

What can we do?

We can always measure averages over *subsets* of outcomes…

What's the average patient

- outcome
- prediction error
- prediction bias

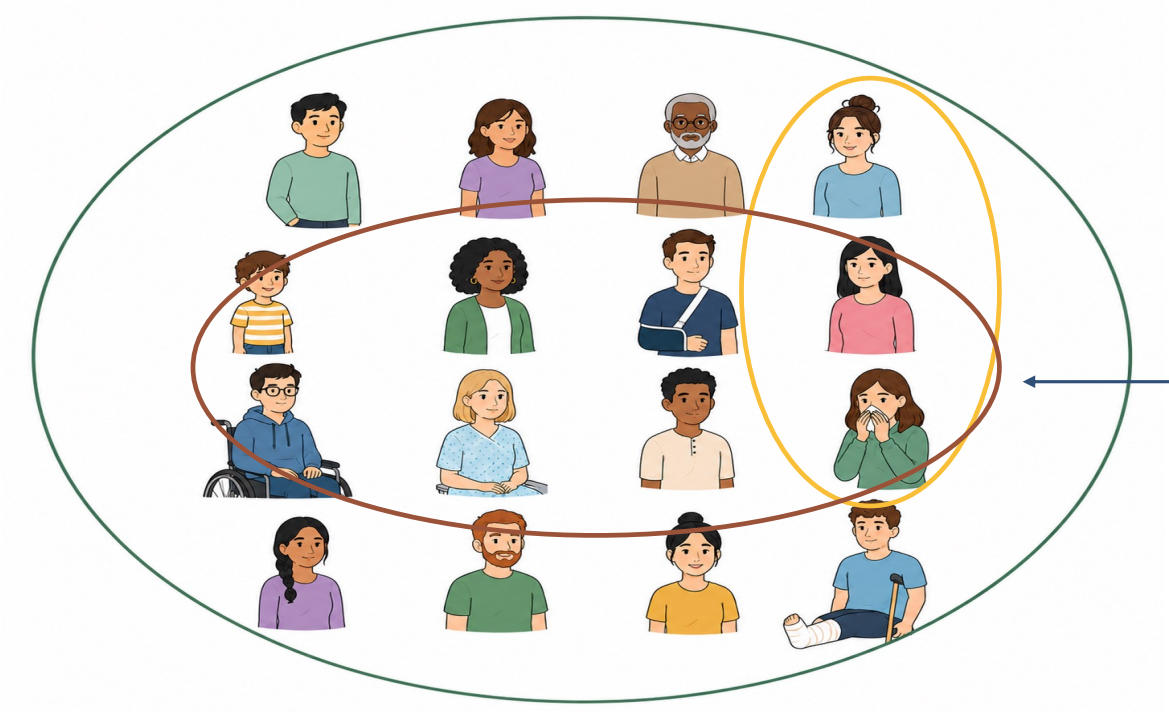
What can we do?

We can always measure averages over *subsets* of outcomes...

What about for these patients?

What can we do?

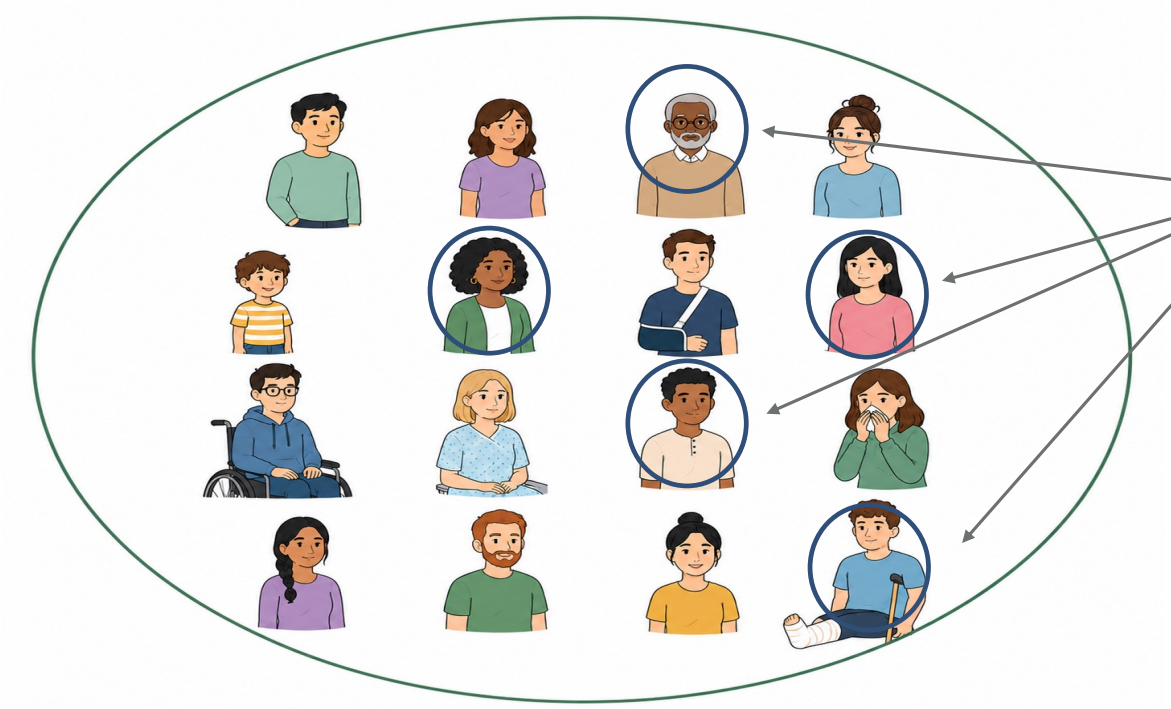
We can always measure averages over *subsets* of outcomes…



Or these?

Are we back to predicting individual probabilities?

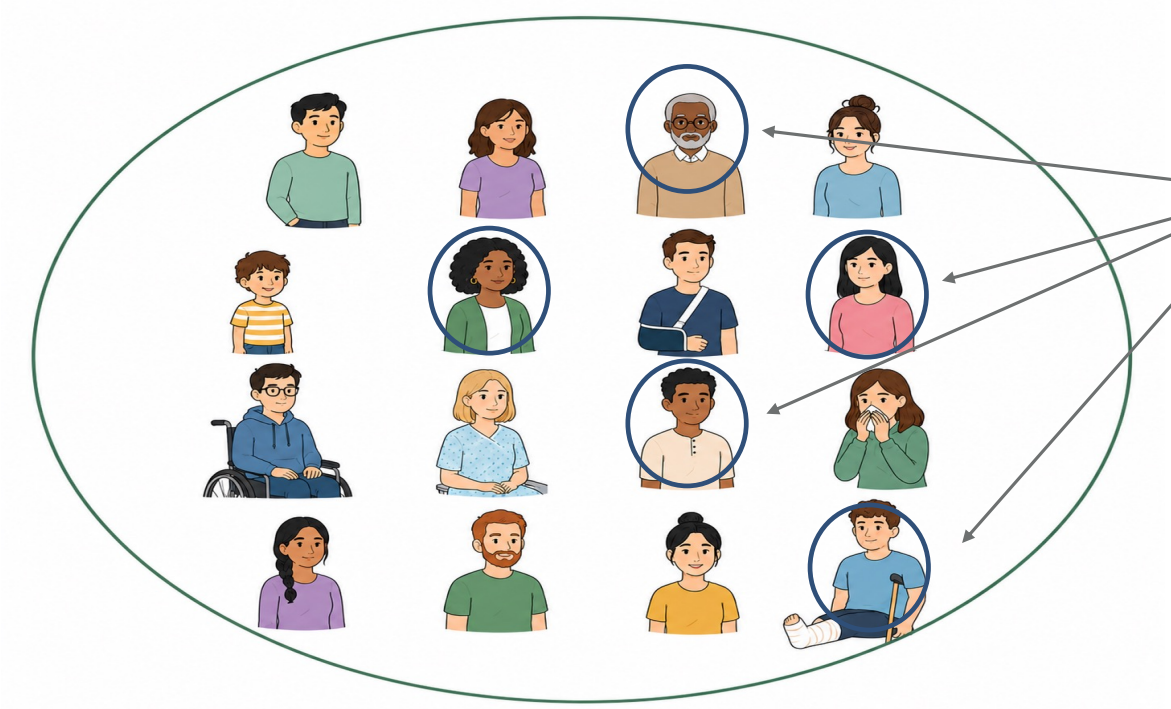
If we take our subsets to be too small then yes…



Let’s learn on each of these subsets!

Are we back to predicting individual probabilities?

If we take our subsets to be too small then yes…



Let’s learn on each of these subsets!

Luckily, we won’t have to this!

Goal: Predictions which have same useful properties as true probabilities

Real world

Indistinguishable!

World where our predictions are correct

The trick: clever debiasing

Clever debiasing

Make predictions unbiased on the right subsequences

We use clever debiasing to **interpret probabilities**...

Interpretable probabilities

Meaningful, trustworthy predictions.

We use clever debiasing to **interpret probabilities**...

Interpretable probabilities

Meaningful, trustworthy predictions.

...to give guarantees for downstream decisions...

...to give guarantees for downstream decisions...

Good decision-making

Make predictions suited for downstream decisions.

...and to reason about collaborative settings!

Clever debiasing is an efficient post-hoc process

Post-hoc + do-no-harm

Already trained your model? That's okay—you can still use these tools!

Lightweight postprocessing

Efficiently applied to the model you've already trained.

Do-no-harm guarantee

The new model will only be better than the one you started with.

Overarching Goal

- Define a collection of simple empirical tests that real probabilities would pass, and are auditable from data.
- Learn predictors that pass the tests.
- Choose the tests so that they are sufficient for downstream applications.
  - i.e. test for the properties of probabilities that are useful.

Roadmap

Multicalibration (and related)

A simple hierarchy of tests we can ask for

- Why we can pass these tests (a clever minimax argument)
- Efficient algorithms for passing these tests.

Prediction for downstream decision-making

Learn a single predictor that can be used to simultaneously optimize many downstream utility functions.

Fast multi-party agreement and information aggregation

Let two parties with mutually unintellegible observations (e.g. an AI and a person) interact to get complementary benefits.

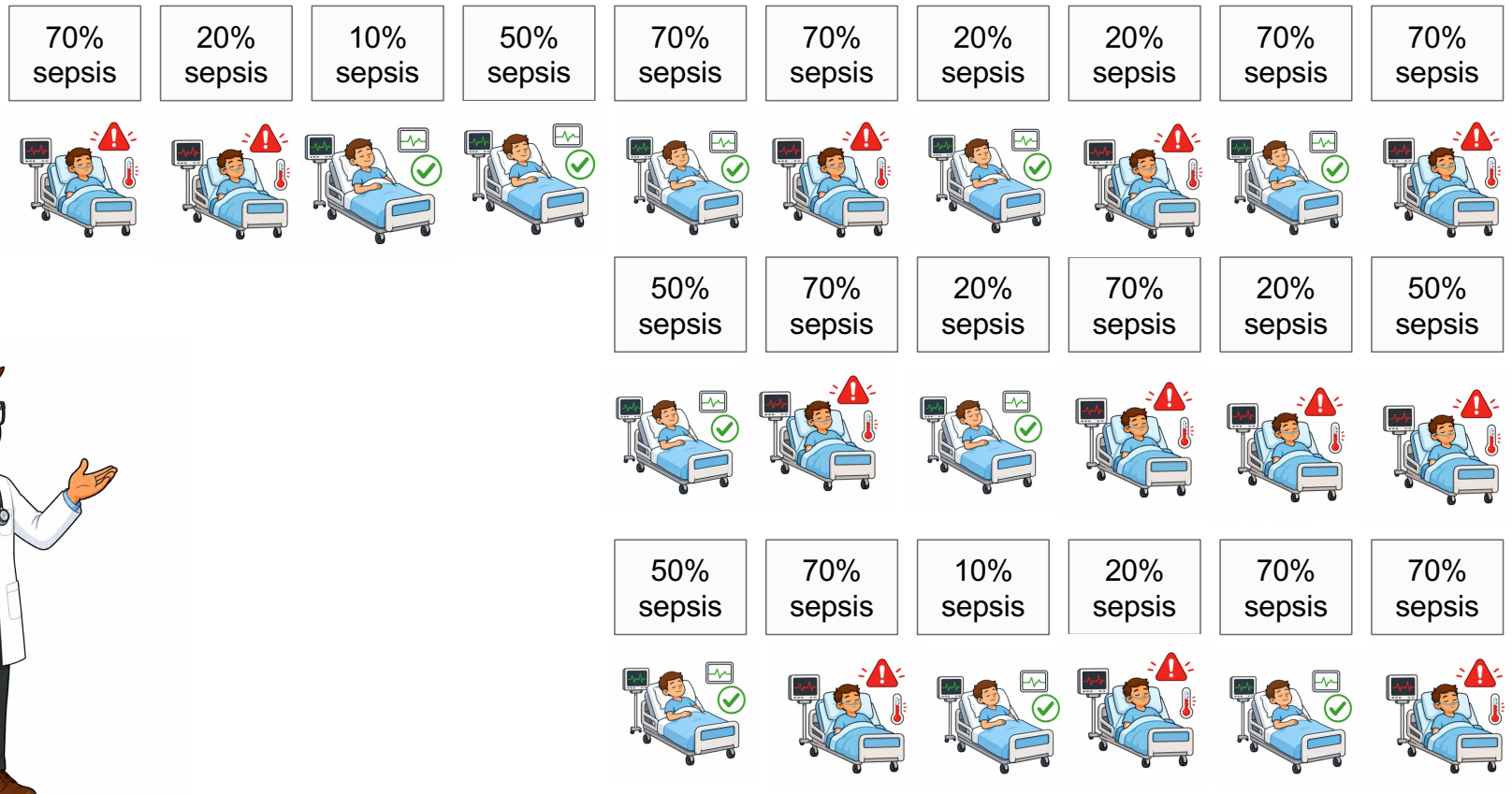
Getting Started

Calibration

What is Calibration [Dawid '82]?

70%
sepsis

What is Calibration [Dawid '82]?



What is Calibration [Dawid '82]?

70% sepsis

70% sepsis

70% sepsis

70% sepsis

70% sepsis

70% sepsis

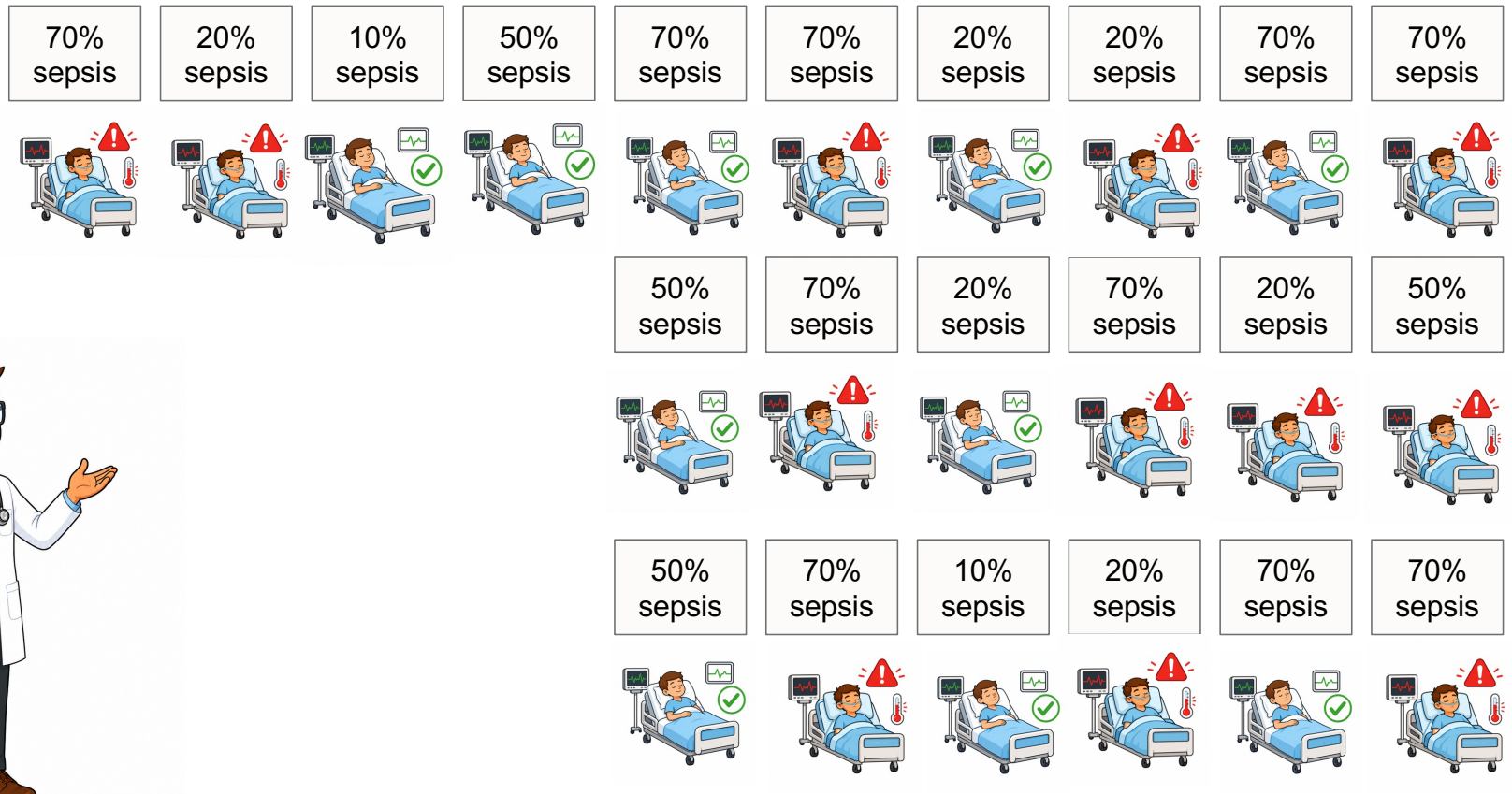
70% sepsis

70% sepsis

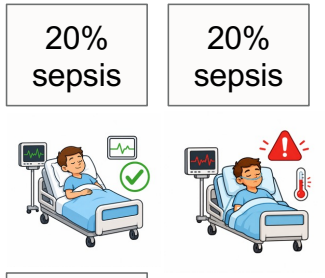
70% sepsis

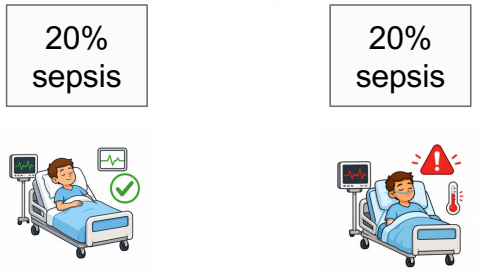
70% sepsis

What is Calibration [Dawid '82]?



What is Calibration [Dawid '82]?





Calibration: Self-Referential Consistency

[Dawid '82]

- Fix a distribution $\mathcal{D}$ on labeled examples $\mathcal{D} \in \Delta(X \times \{0,1\})$.
- Calibration asks that predictions be statistically unbiased conditional on the predictions themselves. In one dimension:

$f: X \to [0,1]$ is calibrated if for all $v \in [0,1]$:

$$\mathbb{E}[f(x) - y | f(x) = v] = 0$$

Approximate Calibration Error?

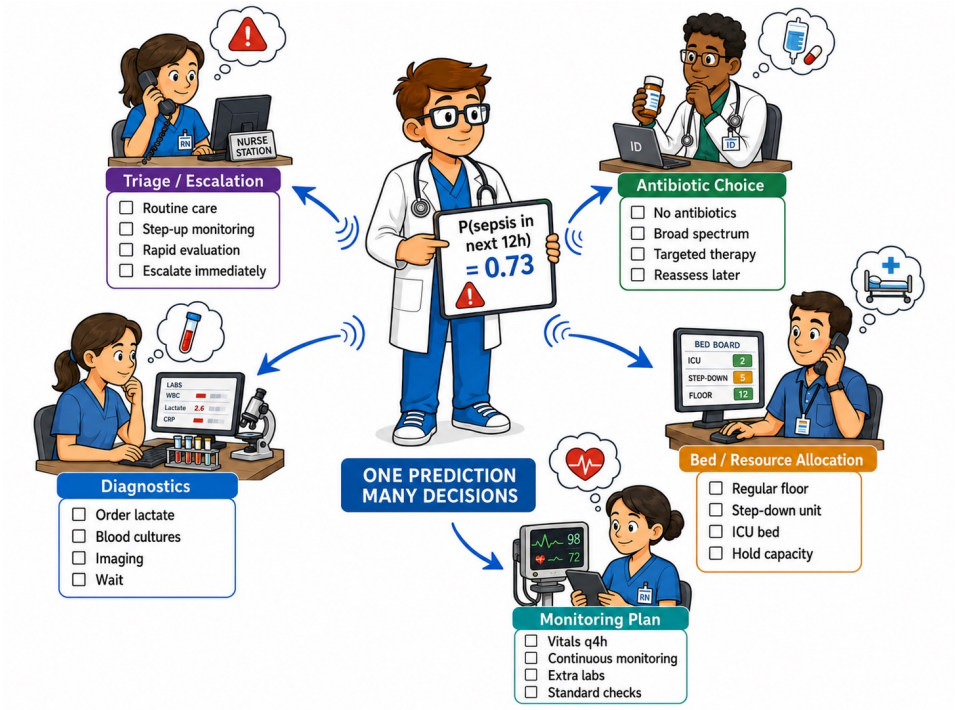
Many ways to measure (topic of active research).

A classic: Expected Calibration Error

$$ECE(f, \mathcal{D}) = \sum_{v} \Pr[f(x) = v] \cdot |\mathbb{E}[f(x) - y \mid f(x) = v]|$$

(Some disadvantages: Discontinuous in f, inconsistent with good rates in online prediction, …)

Calibration Instills Trust in Predictions



Calibration Instills Trust in Predictions

Theorem: Fix any $u: A \times [0,1] \to [0,1]$. If forecaster $f(x) = p$ has $ECE(f, \mathcal{D}) \leq \epsilon$, then following the best-response policy $BR(u, p)$ obtains payoff almost as high as any other policy $\pi: [0,1] \to A$:

$$\mathbb{E}_{p,y}[u(BR(u,p), y)] \geq \mathbb{E}_{p,y}[u(\pi(p), y)] - 2\epsilon$$

$$u(a,p) = p \cdot u(a,1) + (1-p) \cdot u(a,0)$$

$$BR(u,p) = \arg\max_a u(a,p)$$

"Can't do better than to trust the predictions"

Calibration Instills Trust in Predictions

Theorem: Fix any $u: A \times [0,1] \to [0,1]$. If forecaster $f(x) = p$ has $ECE(f, \mathcal{D}) \leq \epsilon$, then following the best-response policy $BR(u,p)$ obtains payoff almost as high as any other policy $\pi: [0,1] \to A$:

$$\mathbb{E}_{p,y}[u(BR(u,p),y)] \geq \mathbb{E}_{p,y}[u(\pi(p),y)] - 2\epsilon$$

Proof:

$$\begin{aligned}
&\mathbb{E}_p\left[\mathbb{E}_y[u(BR(u,p),y)|p]\right] \\
&= \mathbb{E}_p[u(BR(u,p),\mathbb{E}[y|p])] \\
&\geq \mathbb{E}_p[u(BR(u,p),p)] - \epsilon \\
&\geq \mathbb{E}_p[u(\pi(p),p)] - \epsilon \\
&= \mathbb{E}_{p,y}[u(\pi(p),y)] - 2\epsilon
\end{aligned}$$

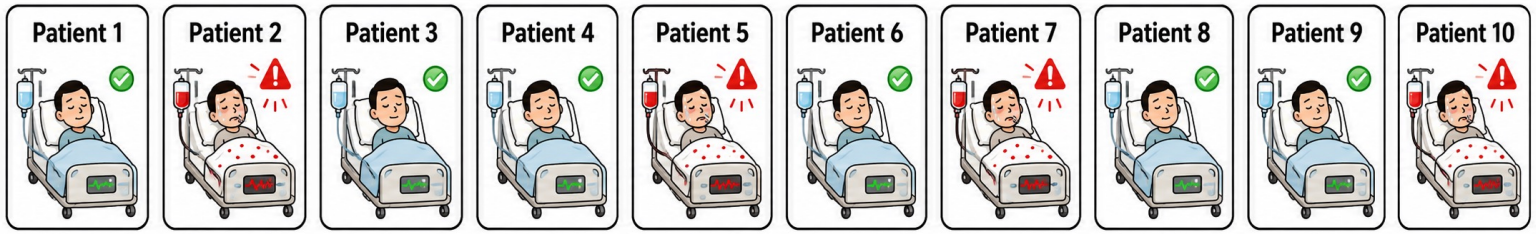
Real world

Optimistica

Real world

But wait! Calibration Alone is Weak.

Average Frequency of Sepsis: 40%



There are many ways to be calibrated. Here are two.

A+

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
0%	100%	0%	0%	100%	0%	100%	0%	0%	100%

A+

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
40%	40%	40%	40%	40%	40%	40%	40%	40%	40%

The Catch

- The benchmark class in the previous theorem was policies $\pi(f(x))$ that take as input *only* f(x). So if $f(x)$ is constant so are the comparators.
  - i.e. it doesn't apply to decision makers who have other observations.
- Calibration also does not imply accuracy (better than a constant predictor).

So calibration is a *minimal* desirable property, but we want more.

Multicalibration

[Hebert-Johnson, Kim, Reingold, Rothblum, ICML '18]

- We can ask for calibration conditional on external context.
- Fix a collection of grouping functions $\mathcal{G} = \{g_1, \dots, g_k\}$, $g_i : X \to \{0,1\}$
- Let $\mathcal{D}_g = \mathcal{D}_{|(g(x)=1)}$.

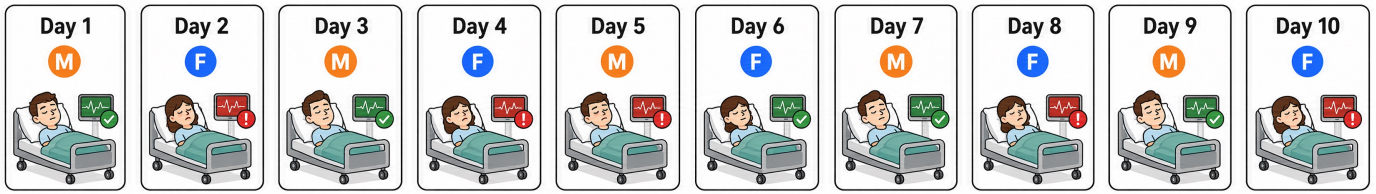
A predictor f is ϵ-multicalibrated with respect to $\mathcal{D}$ and $\mathcal{G}$ if :

$$ECE(f, \mathcal{D}, \mathcal{G}) = \max_{g \in \mathcal{G}} \Pr_{\mathcal{D}}[g(x) = 1] \cdot ECE\left(f, \mathcal{D}_g\right) \leq \epsilon$$

Multicalibration

Suppose we can now observe clinically relevant features.

Overall frequency of sepsis: 50%. For female patients: 80%. For male patients: 20%



A+

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
20%	80%	20%	80%	20%	80%	20%	80%	20%	80%

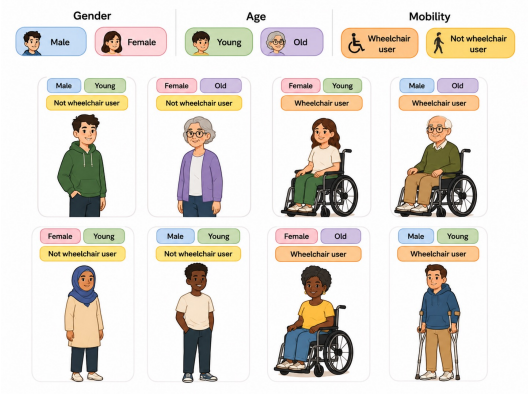
F

Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
50%	50%	50%	50%	50%	50%	50%	50%	50%	50%

Multicalibration

[Hebert-Johnson, Kim, Reingold, Rothblum, ICML '18]

- The "groups" $\mathcal{G}$ can be arbitrary and overlapping. For example:
  - Original fairness motivation --- intersecting demographic groups (indicators of gender, ethnicity, age, etc)



  - Level-sets of benchmark policies
  - Functions of a history of interaction with an interlocutor…
  - …

More generally… Outcome Indistinguishability [Dwork, Kim, Reingold, Rothblum, Yona '21]

- Fix a collection of tests $\mathcal{T}$, where each test a $\in \mathcal{T}$ has the form: a: $\mathcal{X} \times [0,1] \to [-1,1]$

A predictor $f: \mathcal{X} \to [0,1]$ satisfies *ϵ-outcome indistinguishability* (OI) relative to $\mathcal{T}$ if:

$$\max_{a \in \mathcal{T}} \left| \mathbb{E}\left[a\left(x, f(x)\right)\left(f(x) - y\right) \right] \right| \leq \epsilon$$

A world in which $\Pr[y = 1 \mid x] = f(x)$ *would be indistinguishable from the real world to a test that computes* $\mathbb{E}\left[a\left(x, f(x)\right) \cdot y \right]$.

Outcome Indistinguishability

In the real world, (x, y) are jointly drawn according to some complicated process.

In **optimistica**, every day $y \sim Ber(f(x))$.
i.e. predictions are real probabilities.

Real world

Optimistica

Goal: The Real World and Optimistica should be indistinguishable to tests $\mathbb{E}[a(x, f(x))y]$.

Outcome Indistinguishability is statistically/computationally feasible.

Theorem: For any collection of tests $\mathcal{T}$, we can learn a predictor with OI error ϵ:

$$\max_{a \in \mathcal{T}} \left| \mathbb{E}\left[a\big(x, f(x)\big)\big(f(x) - y\big) \right] \right| \leq \epsilon$$

from only $n \leq O\left(\frac{\log |\mathcal{T}|}{\epsilon^2}\right)$ many samples ("the statistically optimal rate").

And we can do this even in the sequential adversarial setting.

- Modelling, e.g., arbitrary distribution shift.

The Sequential/Adversarial Prediction Setting

In rounds $t = 1, \ldots, T$:

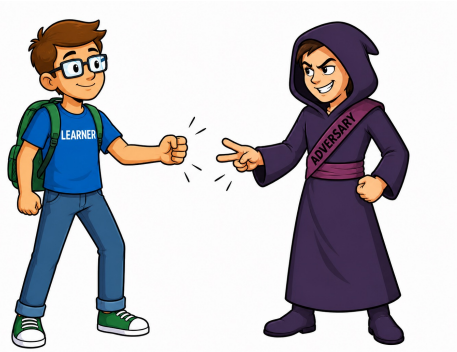
- The learner observes some context $x_t \in X$.
- The learner produces a prediction $p_t \in [0,1]$
- The learner observes outcome $y_t \in [0,1]$ chosen by an adaptive adversary.

Can evaluate outcome indistinguishability error on the transcript $\pi = (x_t, p_t, y_t)_{t=1}^{T}$:

$$OI(\pi, \mathcal{T}) = \max_{a \in \mathcal{T}} \frac{1}{T} \sum_{t} |a(x_t, p_t)(y_t - p_t)|$$

(i.e. OI on the empirical distribution)

An Interlude: Zero Sum Games



- A Zero Sum Game is defined by:
  1. A *minimization player* (the learner) with finite strategy space A_1
  2. A *maximization player* (the adversary) with finite strategy space A_2
  3. A utility function u: $A_1 \times A_2 \to \mathbb{R}$.

     Extended to distributions in the natural way. For $Q_1 \in \Delta A_1, Q_2 \in \Delta A_2$:

     $$u(Q_1, Q_2) = \mathbb{E}[u(a_1, a_2)]$$

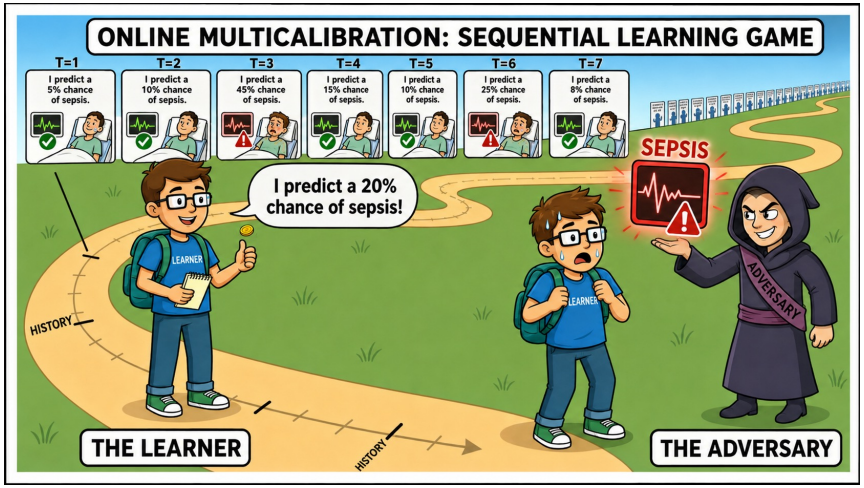
Von Neumann's Minimax Theorem:

$$\min_{Q_1 \in \Delta A_1} \max_{a_2 \in A_2} u(Q_1, a_2) = \max_{Q_2 \in \Delta A_2} \min_{a_1 \in A_1} u(a_1, Q_2)$$

"Order of play doesn't matter"

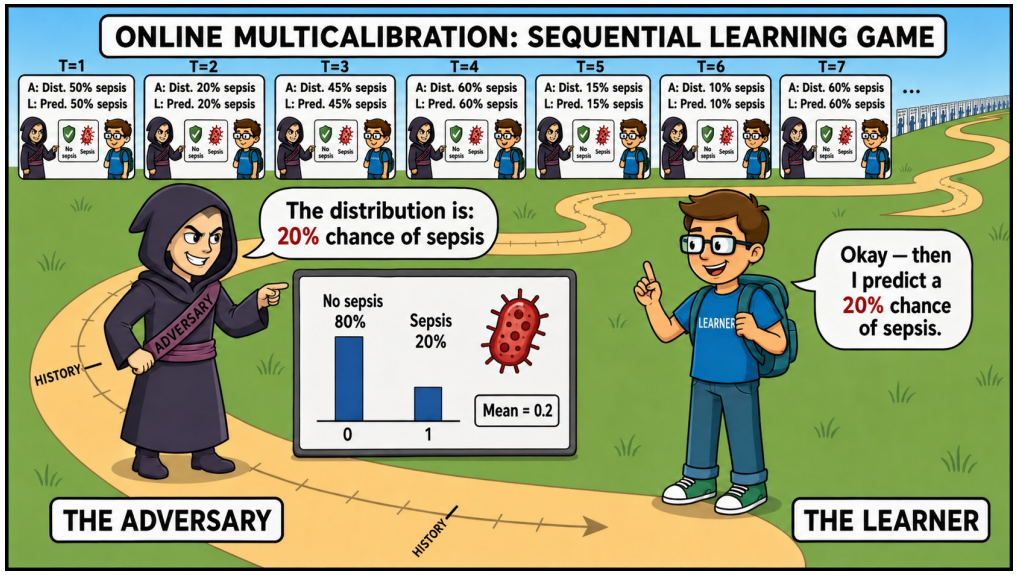
Where does the optimal rate come from?

- Sequential learning is a zero-sum game played against an adversary.
- The learner must go first (commit to an algorithm) and give guarantees against all adversary strategies.



Where does the optimal rate come from?

- But by the minimax theorem (must verify conditions), the value of the game is the same as in the (seemingly) much easier scenario in which the adversary goes first and the learner gets to respond.



Where does the optimal rate come from?

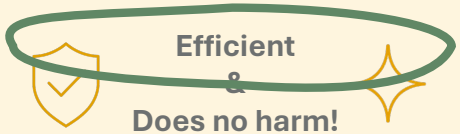
- Unsurprisingly, when the learner knows the distribution, they can satisfy OI constraints at the statistically optimal rate.
- Amazingly (!) by the minimax theorem, there is an algorithm that guarantees they do just as well in adversarial environments.

Nonconstructive!
The game we are applying the minimax argument to has strategy spaces corresponding to *all possible algorithms*.

The trick: clever debiasing

Clever debiasing

Make predictions unbiased on the right subsequences



Interpretable probabilities

Meaningful, trustworthy predictions.

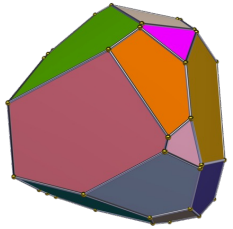
Good decision-making

Use predictions well for downstream decisions.

Collaborative learning

Learn efficiently from distributed information.

Making the Algorithm Efficient



- OI asks that we satisfy a collection of linear constraints, on average, over the transcript. For each $a \in \mathcal{T}$:

$$\sum_t a(x_t, p_t)(y_t - p_t) \leq 0$$

The time average of $a(x_t, p_t)(y_t - p_t)$ should *approach a polytope*.

A "Blackwell Approachability" problem.

Making the Algorithm Efficient

- Say a hater is betting against you, hoping you fail.
- They are running a "no-regret" algorithm maintaining a distribution w^t over tests $a \in \mathcal{T}$.
- Each round, if your hater bet on test $a \in T$ their joy would be $a(x_t, p_t)(y_t - p_t)$, your violation of that test constraint.
- Since the hater hedges:

Your Hater's Joy at round t = $\mathbb{E}_{a \sim w^t}[a(x_t, p_t)(y_t - p_t)]$

Making the Algorithm Efficient

- Since they are running a no regret algorithm they are guaranteed near optimal joy.

$$\text{Your Biggest Failure} = \max_{a \in \mathcal{T}} \frac{1}{T} \sum_t a(x_t, p_t) \cdot (y_t - p_t)$$

$$\text{Your Hater's Average Joy} \geq \text{Your Biggest Failure} - \sqrt{\frac{\log |\mathcal{T}|}{T}}$$

Algorithmic Strategy:
Play so as to minimize your hater's joy

Making the Algorithm Efficient

Let:

$$q_t = \arg\min_{q \in \Delta[0,1]} \max_{y} \mathbb{E}_{p_t \sim q}[\text{Your Haters Joy}]$$

$$\text{Your Hater's Joy} = \mathbb{E}_{a \sim w^t}[a(x_t, p_t)(y_t - p_t)]$$

If you knew $\mathbb{E}[y]$ you could always set $p = \mathbb{E}[y]$

and guarantee your hater 0 expected joy....

Selecting $p_t \sim q_t$ also guarantees your hater 0 expected joy.

Making the Algorithm Efficient

Since

$$0 = \mathbb{E}[\text{Your hater's joy}]$$

$$\geq \text{Your Biggest Failure} - \sqrt{\frac{\log|\mathcal{T}|}{T}}$$

It must be that after T rounds...

$$OI(\pi, \mathcal{T}) \leq \sqrt{\frac{\log|\mathcal{T}|}{T}}$$

Or equivalently, $OI(\pi, \mathcal{T}) \leq \epsilon$ if $T \geq \frac{\log|\mathcal{T}|}{\epsilon^2}$

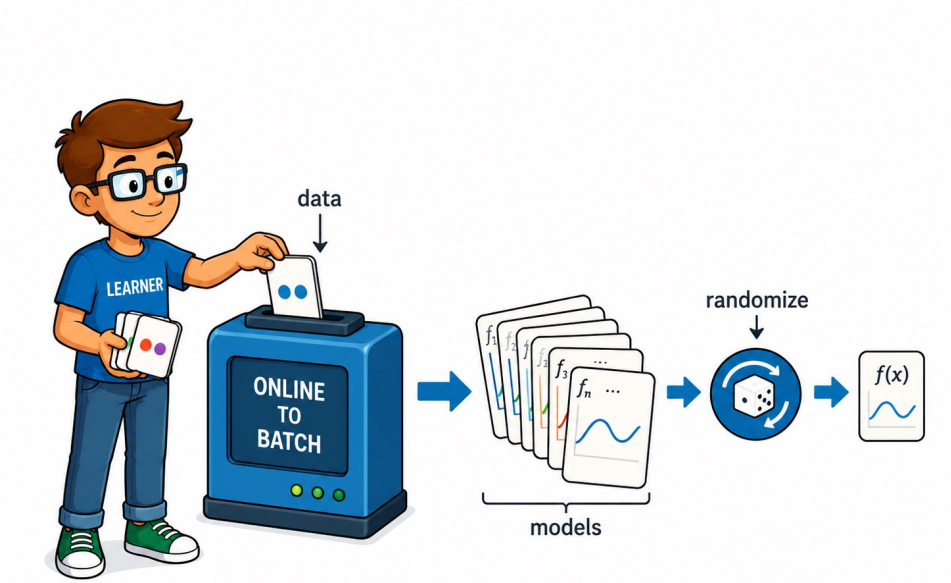
The final online adversarial algorithm

Given: A set of tests $\mathcal{T}$

Initialize a no regret algorithm with actions $\mathcal{T}$.
For $t = 1\ to\ T$:

1. Let w^t be the distribution maintained by the no regret algorithm.
2. Let $\mathrm{J(p, y)} = \mathbb{E}_{a \sim w^t}[a(x_t, p) \cdot (y - p)]$
3. Let $q_t = \arg \min_{q \in \Delta[0,1]} \mathbb{E}_{p \sim q}[J(p, y)]$ (A linear program)
4. Play $p_t \sim q_t$
5. Feed the algorithm gain $a(x_t, p_t) \cdot (y_t - p_t)$ for each test $a \in \mathcal{T}$.

Online to Batch Reduction



This obtains optimal sample complexity $O\left(\frac{\log|\mathcal{T}|}{\epsilon^2}\right)$

How does multicalibration fit?

- Suppose our predictions lie on a grid $\Lambda = \left\{0, \frac{1}{m}, \frac{2}{m}, \dots, 1\right\}$

$$ECE(\pi, \mathcal{G}) = \max_{g \in \mathcal{G}} \frac{1}{T} \sum_{v \in \Lambda} \left| \sum_{t} 1[p_t = v] \cdot 1[g(x_t) = 1](y_t - p_t) \right|$$

Dual view:

$$ECE(\pi, \mathcal{G}) = \max_{g \in \mathcal{G}, \sigma: \Lambda \to \{-1,1\}} \frac{1}{T} \sum_{t} \sigma(p_t) \cdot 1[g(x_t) = 1](y_t - p_t)$$

This is OI for the class $\mathcal{T}_{mcal(\mathcal{G})} = \left\{a_{g,\sigma}(x_t, p_t) = g(x) \cdot \sigma(p_t)\right\}_{g,\sigma}$

How does multicalibration fit?

Lower Bound [Collina, Liu, Noarov, Roth ‘26]

Learning a multicalibrated predictor requires $\widetilde{\Theta}\left(\frac{1}{\epsilon^3}\right)$ samples.

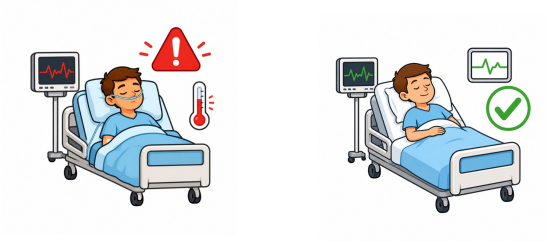
Worse: for k-class outcomes it requires $\widetilde{\Theta}\left(\frac{1}{\epsilon^{k+1}}\right)$ samples.

The issue: too many tests!

Fortunately, fewer OI tests often suffice.

Complicating the Running Example

Sepsis/no sepsis



Patient's future clinical trajectory

Vitals, laboratory values, organ function, and infection indicators over the next 24 hours

This requires being OI with respect to a *lot* of tests!

Downstream Decisions

But even if the prediction is very complex, the *actions* it may induce can be simpler

If the main goal is good *downstream decisions*, then can we use less tests?

The trick: clever debiasing

Clever debiasing

Make predictions unbiased on the right subsequences

Efficient
&
Does no harm!

Interpretable probabilities

Meaningful, trustworthy predictions.

Good decision-making

Use predictions well for downstream decisions.

Collaborative learning

Learn efficiently from distributed information.

The trick: clever debiasing

Clever debiasing

Make predictions unbiased on the right subsequences

Efficient
&
Does no harm!

Interpretable probabilities

Meaningful, trustworthy predictions.

Good decision-making

Use predictions well for downstream decisions.

Collaborative learning

Learn efficiently from distributed information.

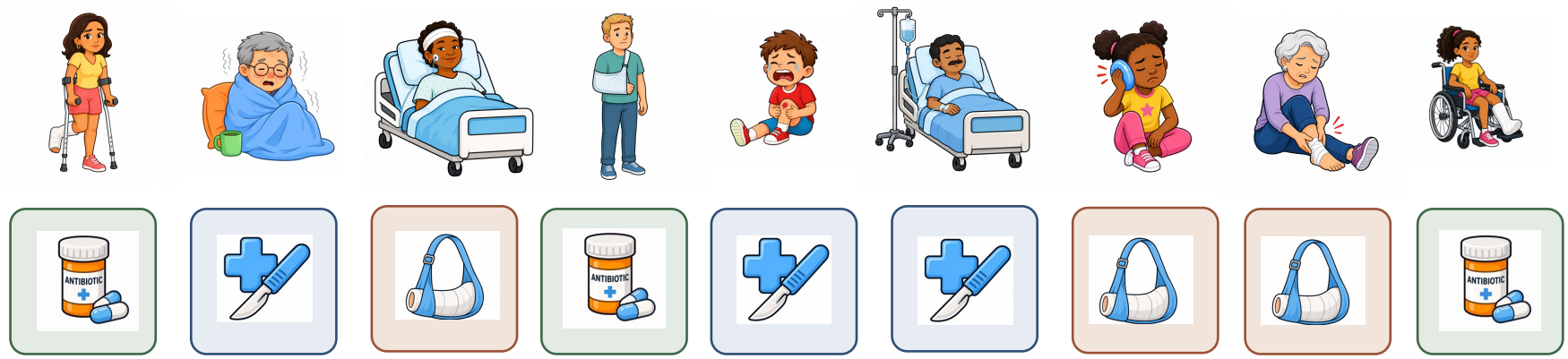
What Makes a Decision “good?”

Say the doctor has some *utility function* $u(a, y)$ mapping from the state y and action a

Let $d\big(f(x)\big)$ be the decision rule which chooses action $a \in \arg\max u(a, f(x))$

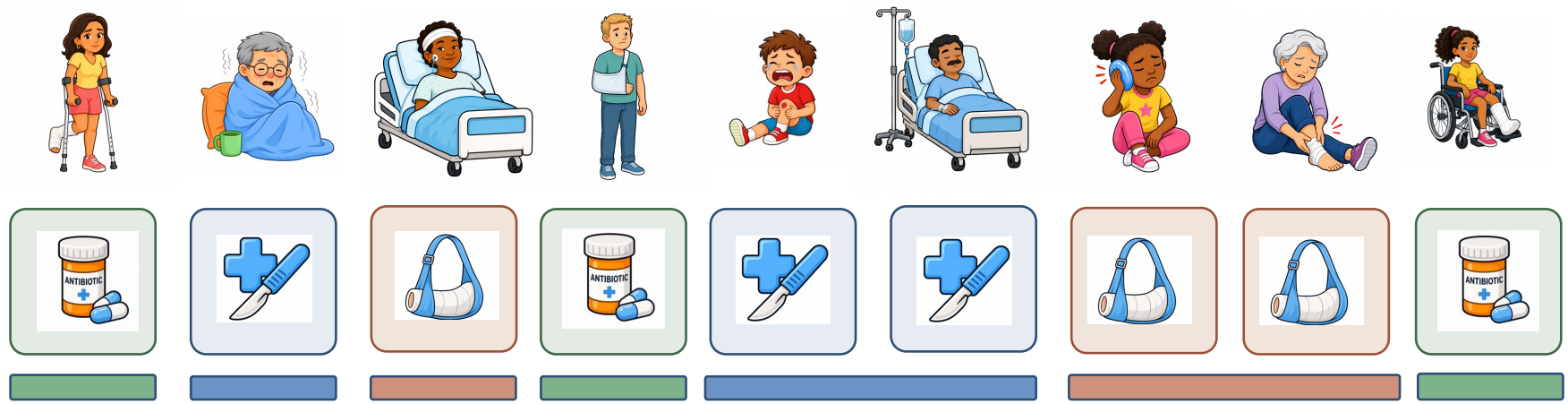
Goal: ensure that the decision maker has **no swap regret** if they follow the predictions

Interlude: Swap Regret



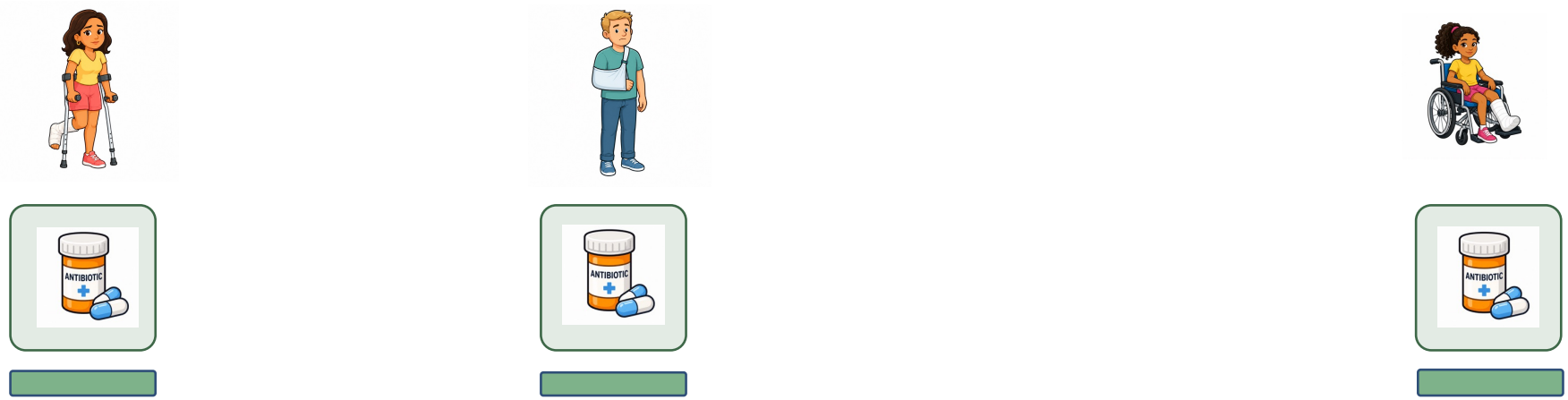
Measures how much a decision-maker can improve by swapping **all times they played one action with another action**

Interlude: Swap Regret



Measures how much a decision-maker can improve by swapping **all times they played one action with another action**

Interlude: Swap Regret



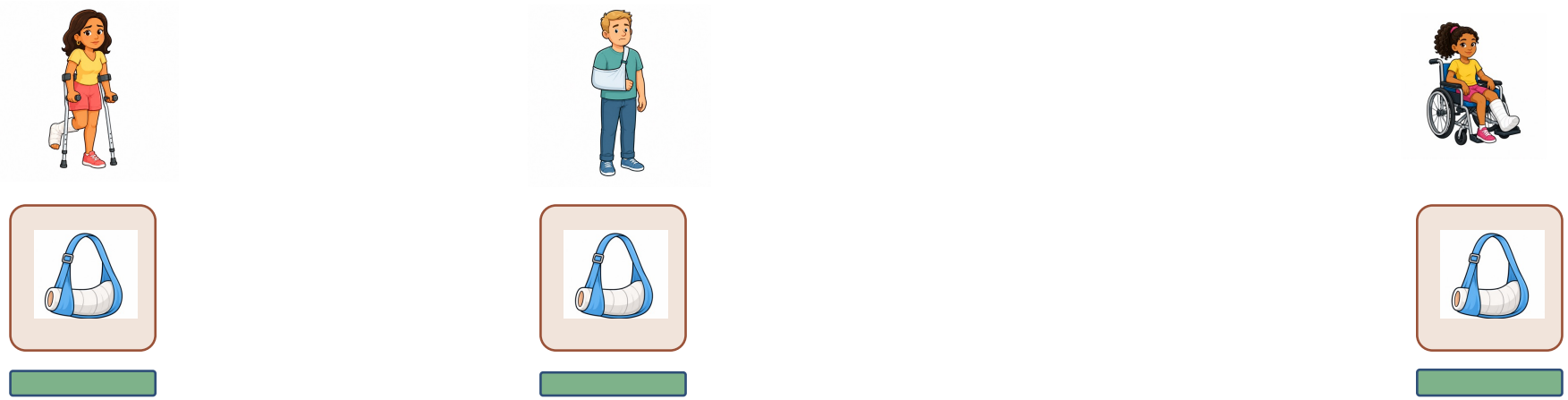
If a decision maker has no swap regret, then on average on the instances they prescribed antibiotics, **this was the best action**

Interlude: Swap Regret



If a decision maker has no swap regret, then on average on the instances they prescribed antibiotics, **this was the best action**

Interlude: Swap Regret

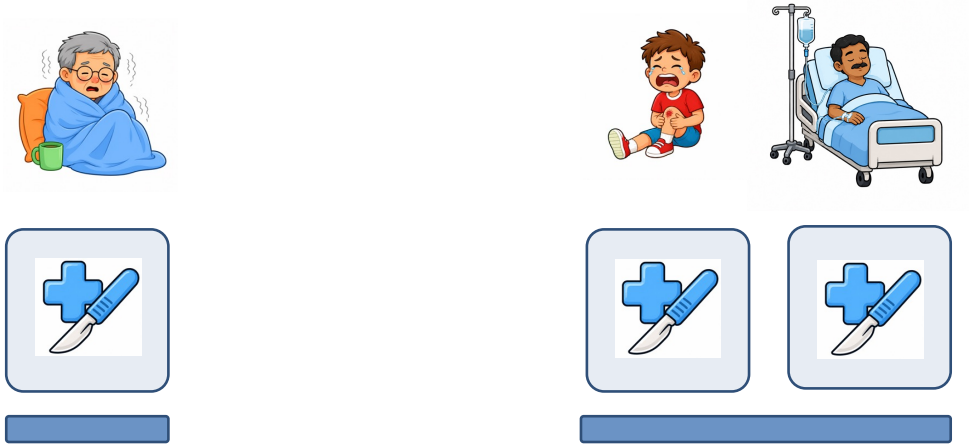


If a decision maker has no swap regret, then on average on the instances they prescribed antibiotics, **this was the best action**

Interlude: Swap Regret

If a decision maker has no swap regret, then on average on the instances they applied a cast, **this was the best action**

Interlude: Swap Regret



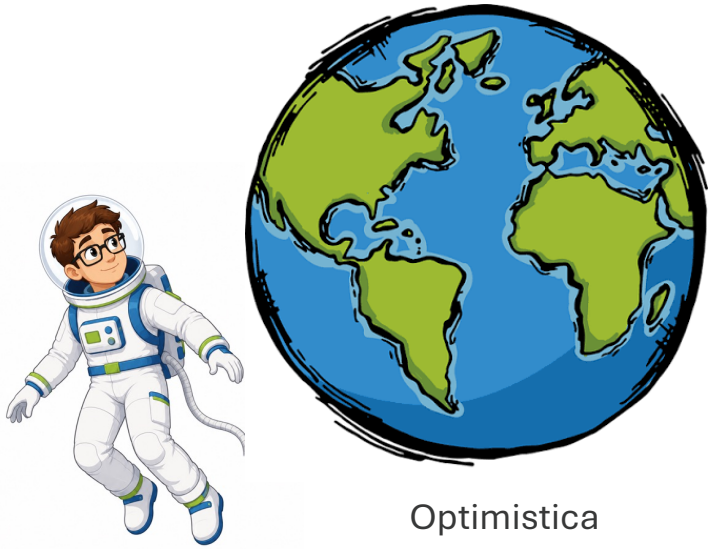
If a decision maker has no swap regret, then on average on the instances they performed surgery, **this was the best action**

What OI tests do we really need to pass here?

In the real world, (x, y) are jointly drawn according to some complicated process.

In **optimistica**, every day $y \sim Ber(f(x))$.
i.e. predictions are real probabilities.

Real world



Optimistica

Goal: The Real World and Optimistica should be Indistinguishable to Statistics $\mathbb{E}\big[a\big(x, f(x)\big)y\big]$.

What OI tests do we really need to pass here?

In the real world, (x, y) are jointly drawn according to some complicated process.

In **optimistica**, every day $y \sim Ber(f(x))$.
i.e. predictions are real probabilities.

Real world

$d(f(x))$ is the decision rule which chooses action $a \in \arg\max u(a, f(x))$

In other words, we don't require our predictions to be fully calibrated overall anymore—just *unbiased* on the doctor's best-response regions

Optimistica

For no swap regret on downstream decisions: The Real World and Optimistica should be indistinguishable to $\mathbb{E}[\mathbf{1}\{d(f(x)) = a\}y]$ for all a.

Decision Calibration

For no swap regret on downstream decisions: The Real World and Optimistica should be indistinguishable to $\mathbb{E}[\mathbf{1}\{d(x, f(x)) = j\}y]$ for all j.

Based on the information the algorithm has given me, the best choice of action is to increase antibiotics ($d(x, f(x)) = antibiotics$)

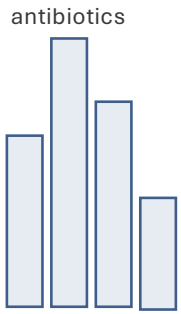
Does this mean I should increase antibiotics?

Well, in Optimistica, I certainly should.

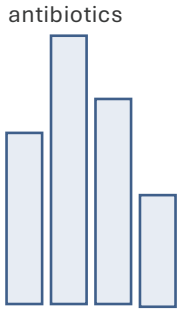
And in aggregate over instances where $d(x, f(x)) = antibiotics$, we cannot distinguish between Optimistica and the real world

So the expected utility of antibiotics is the *same* in both settings

Furthermore, the expected utility of the other actions are also the same!



Optimistica



Real World

Getting Decision Calibration

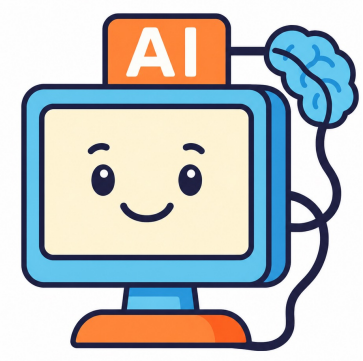
For no swap regret on downstream decisions: The Real World and Optimistica should be indistinguishable to $\mathbb{E}[\mathbf{1}\{d(x, f(x)) = j\}y]$ for all j.

Simply an application of the OI algorithm, with the |A| tests above!

But wait…

- I already have a great model that gets pretty good accuracy and uses lots of complicated techniques. Are you telling me to throw it away?

- No! You can get all of our guarantees *on top* of your own model's performance



The trick: clever debiasing

Clever debiasing

Make predictions unbiased on the right subsequences

Efficient
&
Does no harm!

Interpretable probabilities

Meaningful, trustworthy predictions.

Good decision-making

Use predictions well for downstream decisions.

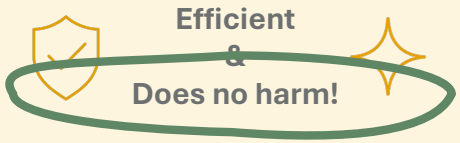
Collaborative learning

Learn efficiently from distributed information.

The trick: clever debiasing

Clever debiasing

Make predictions unbiased on the right subsequences



Interpretable probabilities

Meaningful, trustworthy predictions.

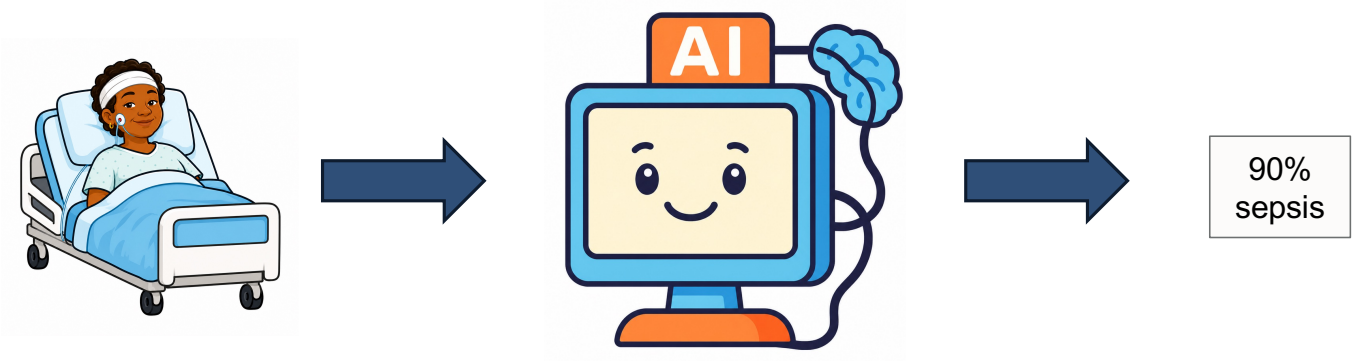
Good decision-making

Use predictions well for downstream decisions.

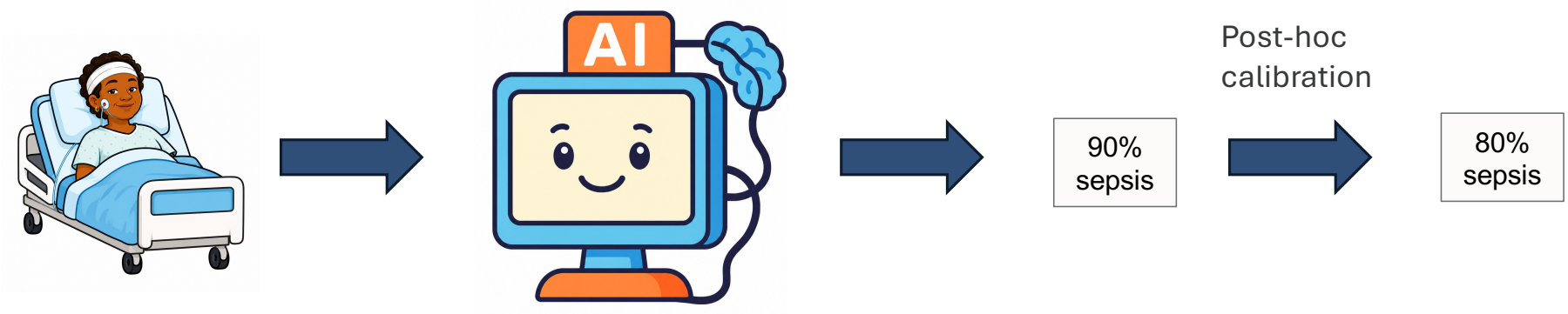
Collaborative learning

Learn efficiently from distributed information.

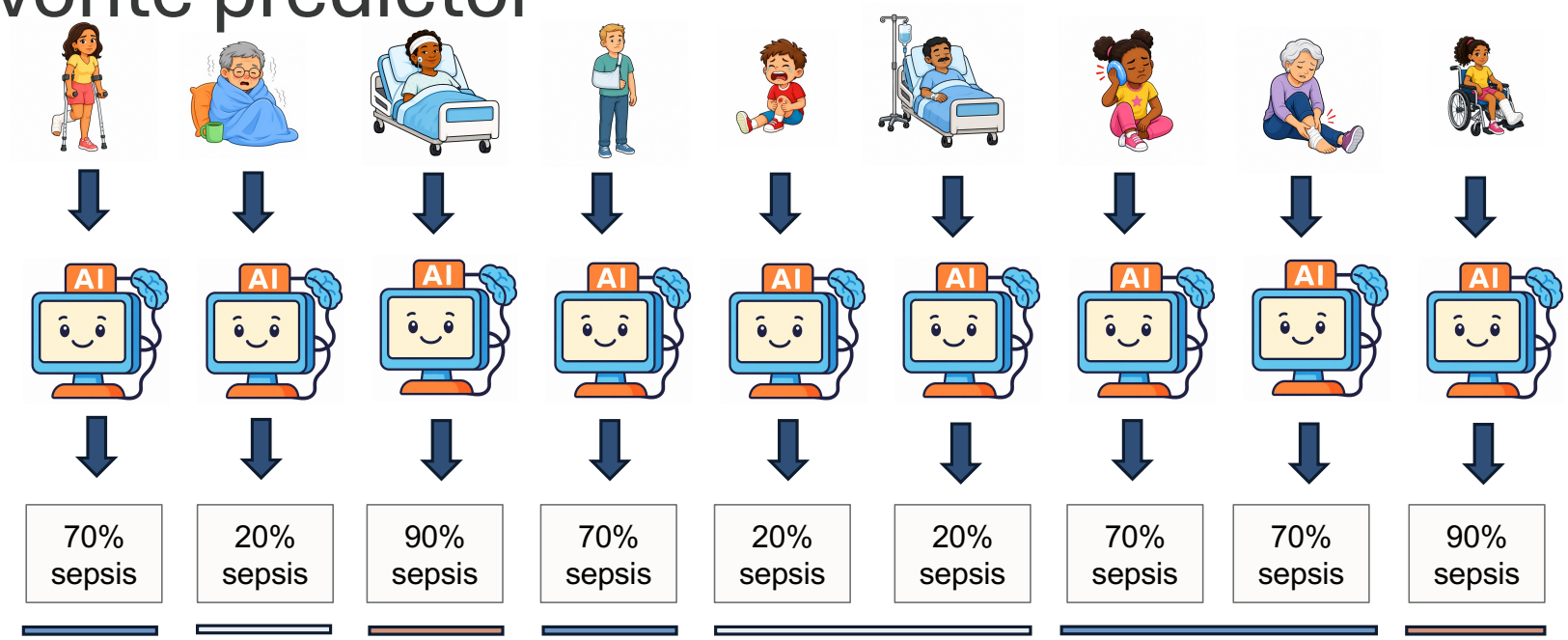
Warm-up: Post-hoc calibration on your favorite predictor



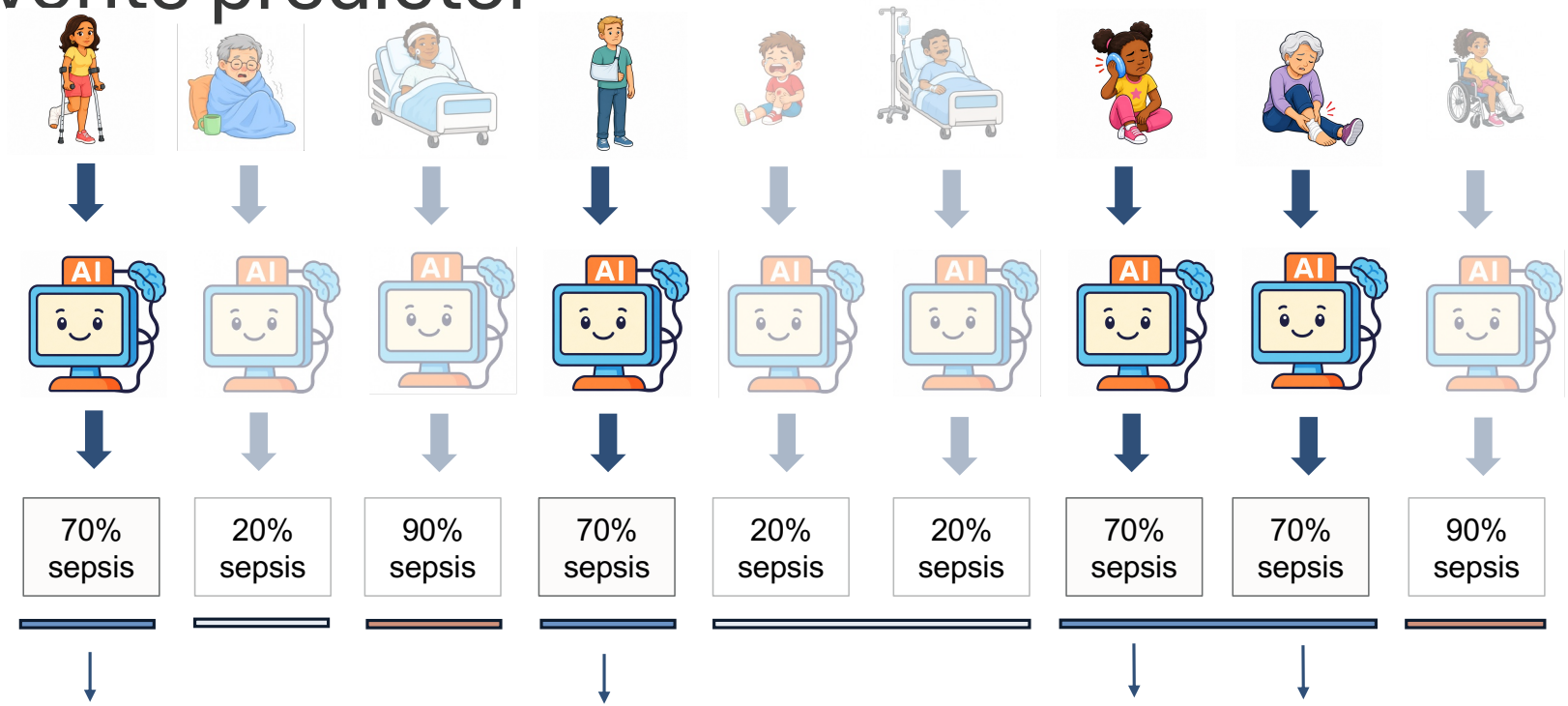
Warm-up: Post-hoc calibration on your favorite predictor



Warm-up: Post-hoc calibration on your favorite predictor

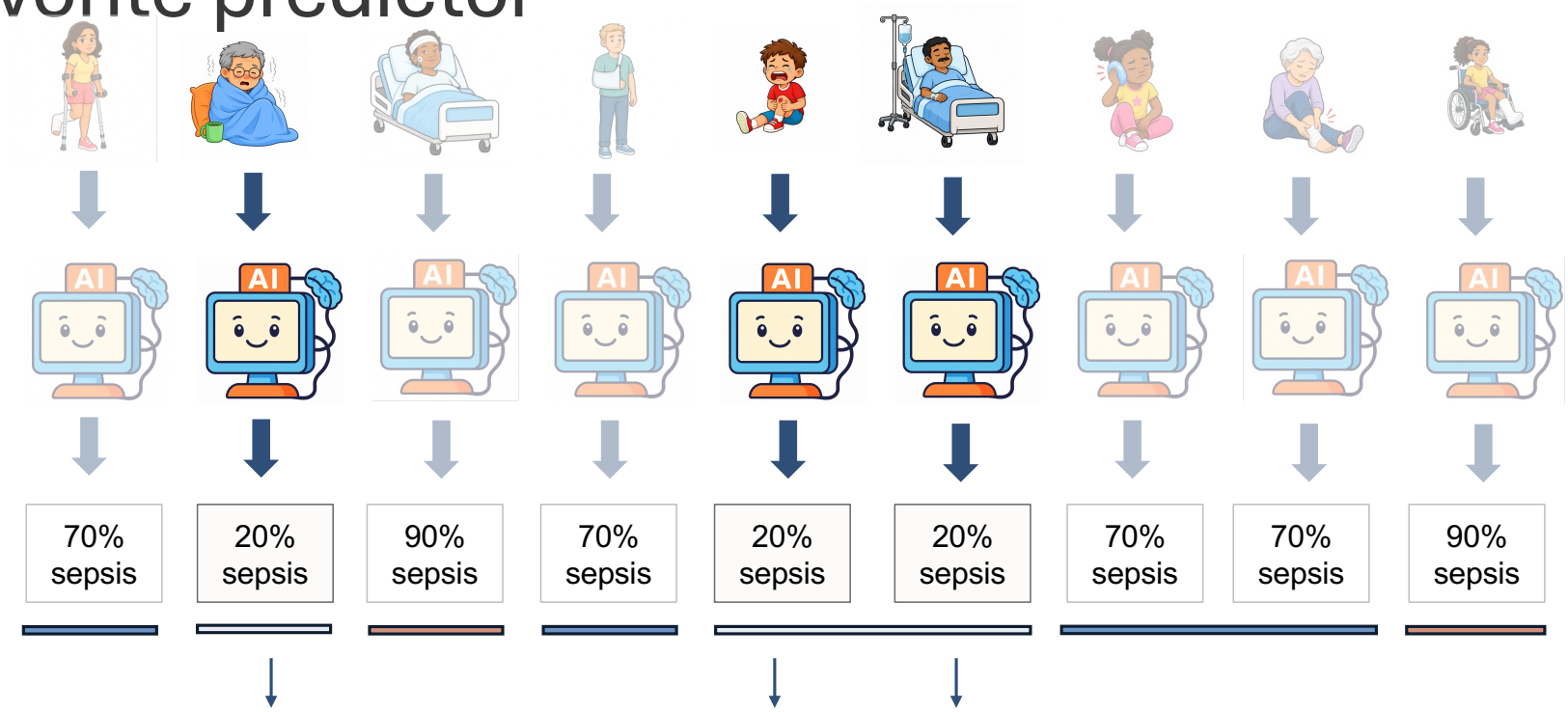


Warm-up: Post-hoc calibration on your favorite predictor

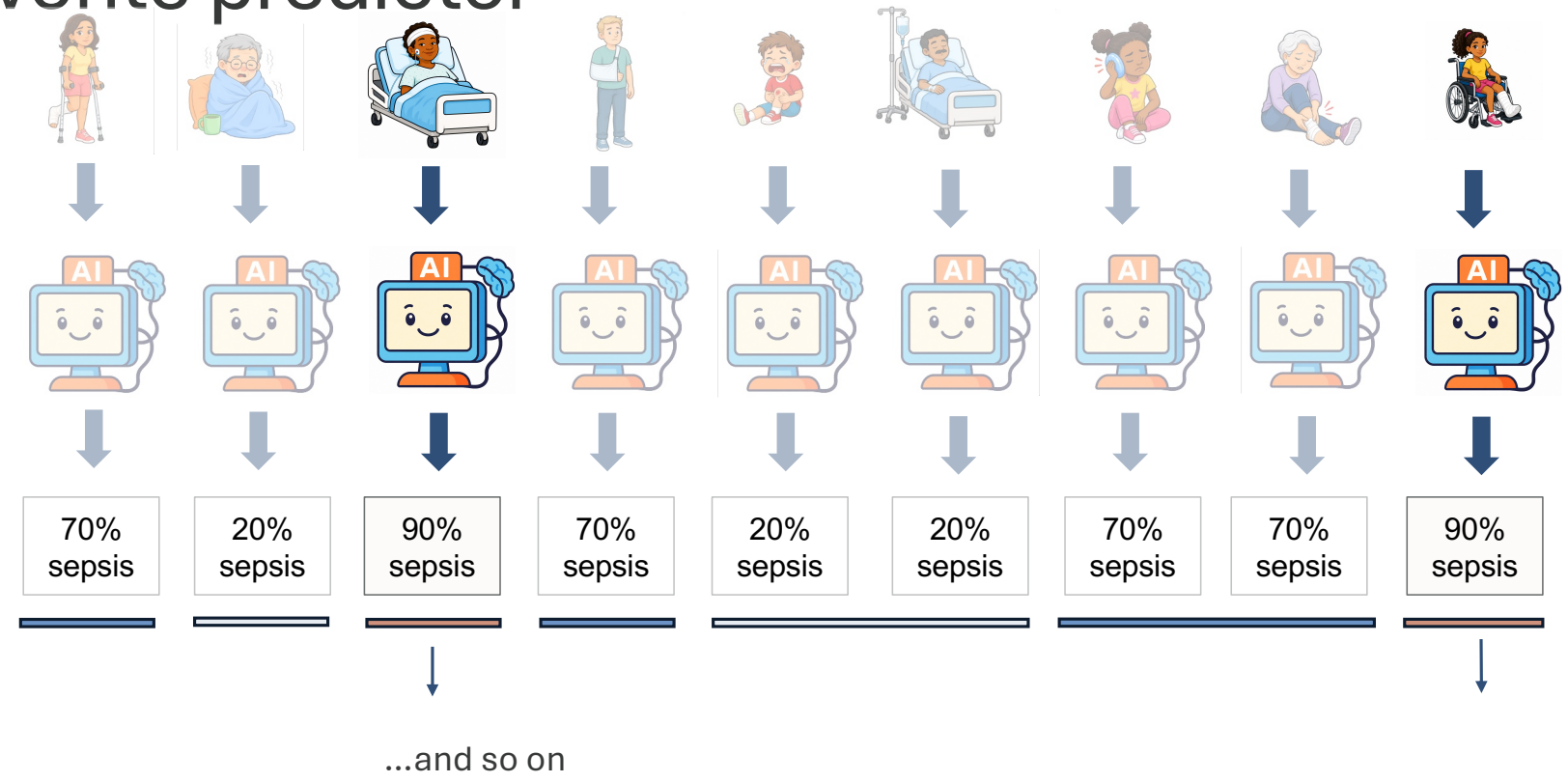


Feed these instances into one calibration algorithm

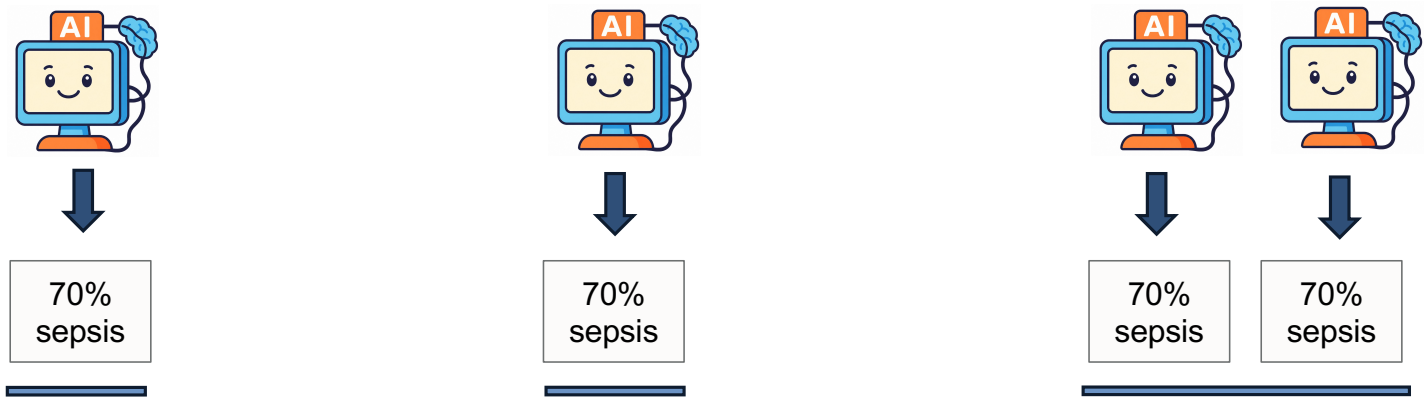
Warm-up: Post-hoc calibration on your favorite predictor



Warm-up: Post-hoc calibration on your favorite predictor

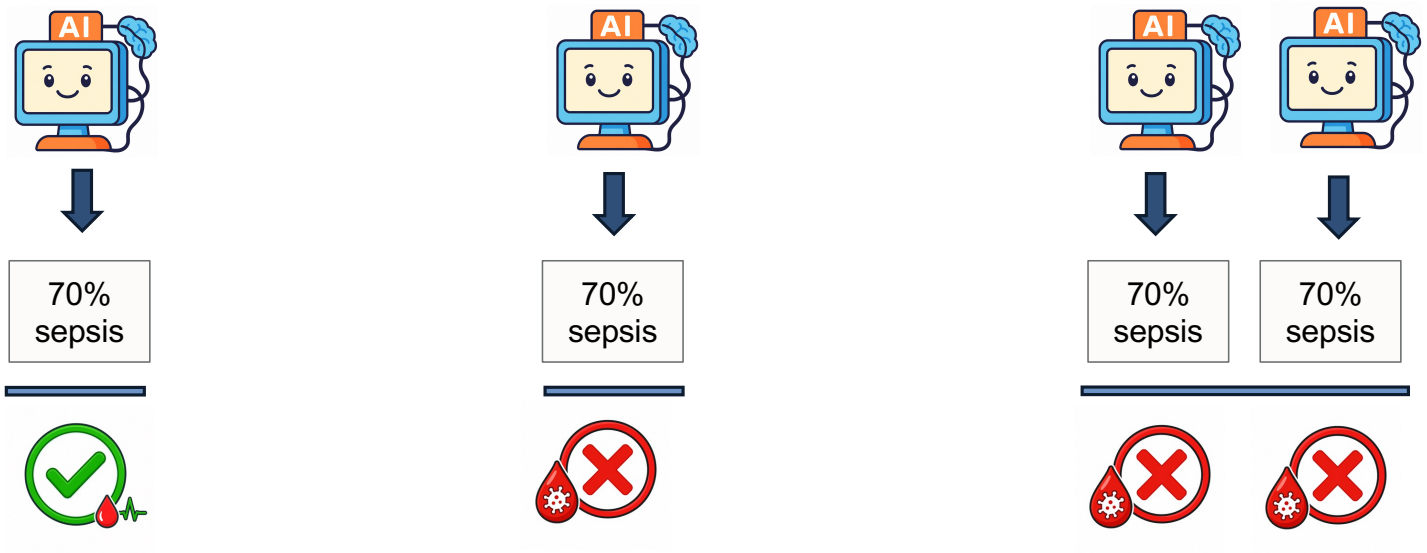


Post-hoc calibration only decreases squared error



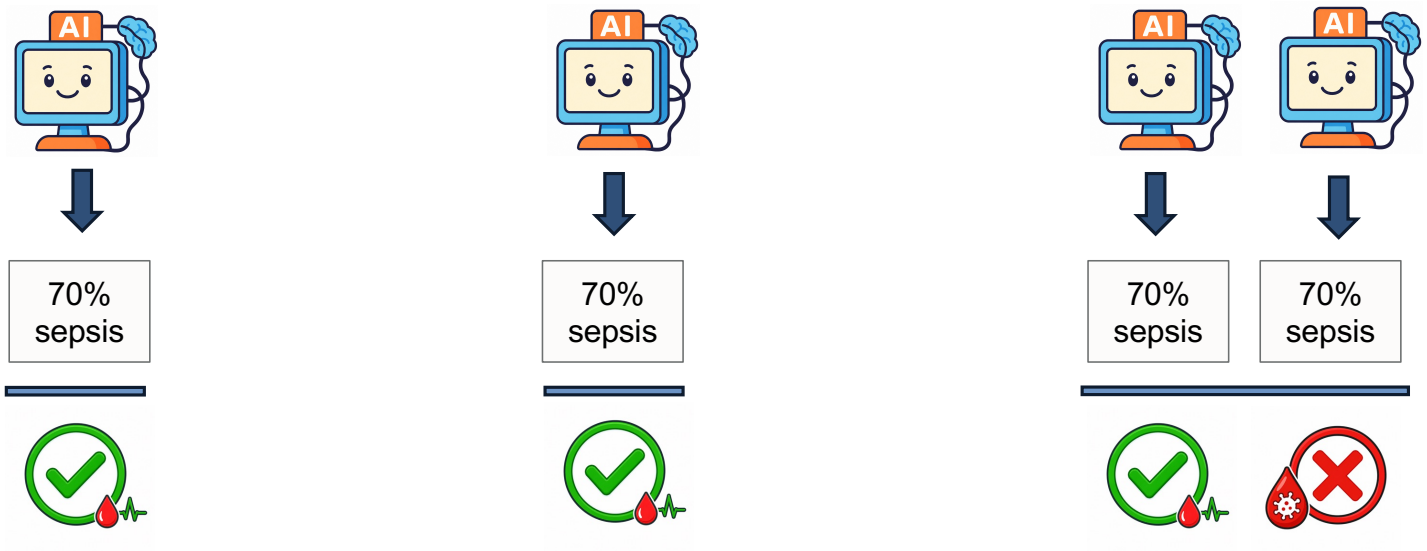
We can compute total squared error by looking separately at all the level sets of your favorite predictor

Post-hoc calibration only decreases squared error



On all the instances where your favorite predictor predicted 70% sepsis, the *best case scenario* for squared error is that 70% of the patients develop sepsis

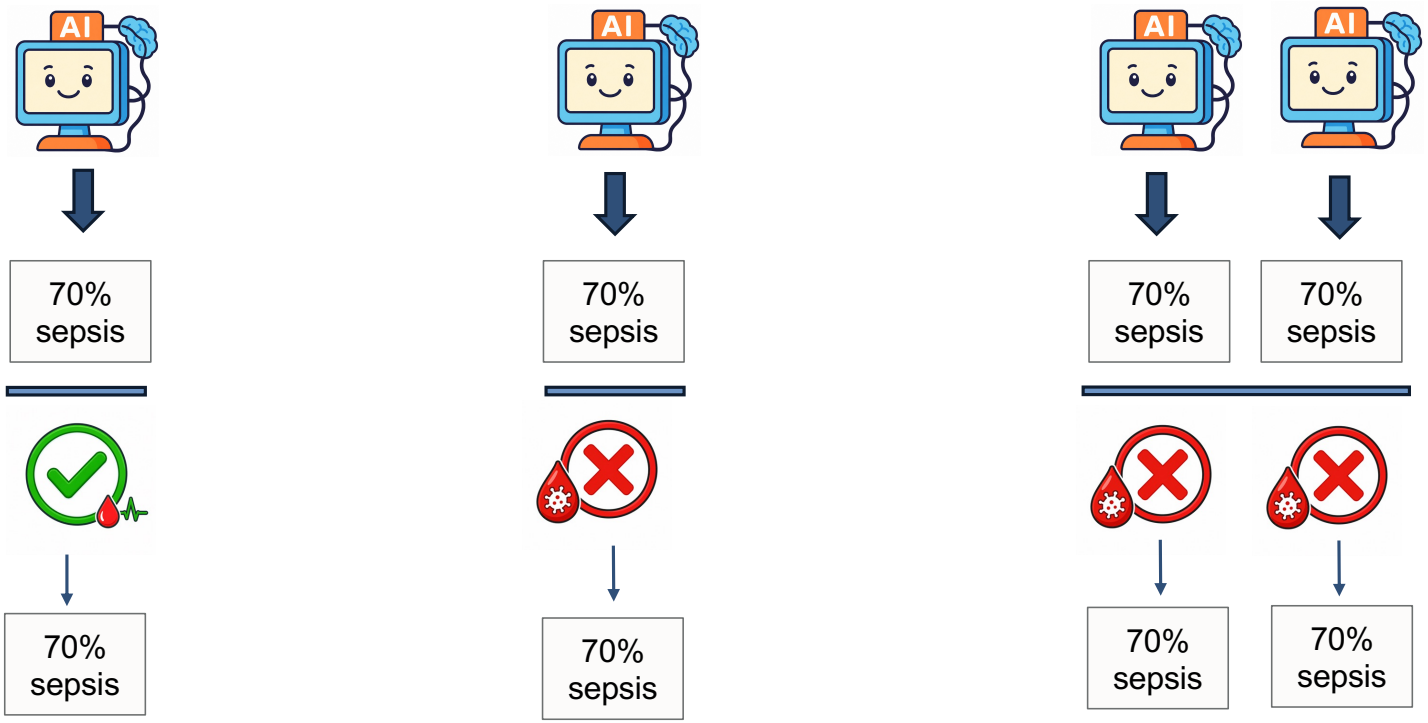
Post-hoc calibration only decreases squared error



On all the instances where your favorite predictor predicted 70% sepsis, the *best case scenario* for squared error is that 70% of the patients develop sepsis

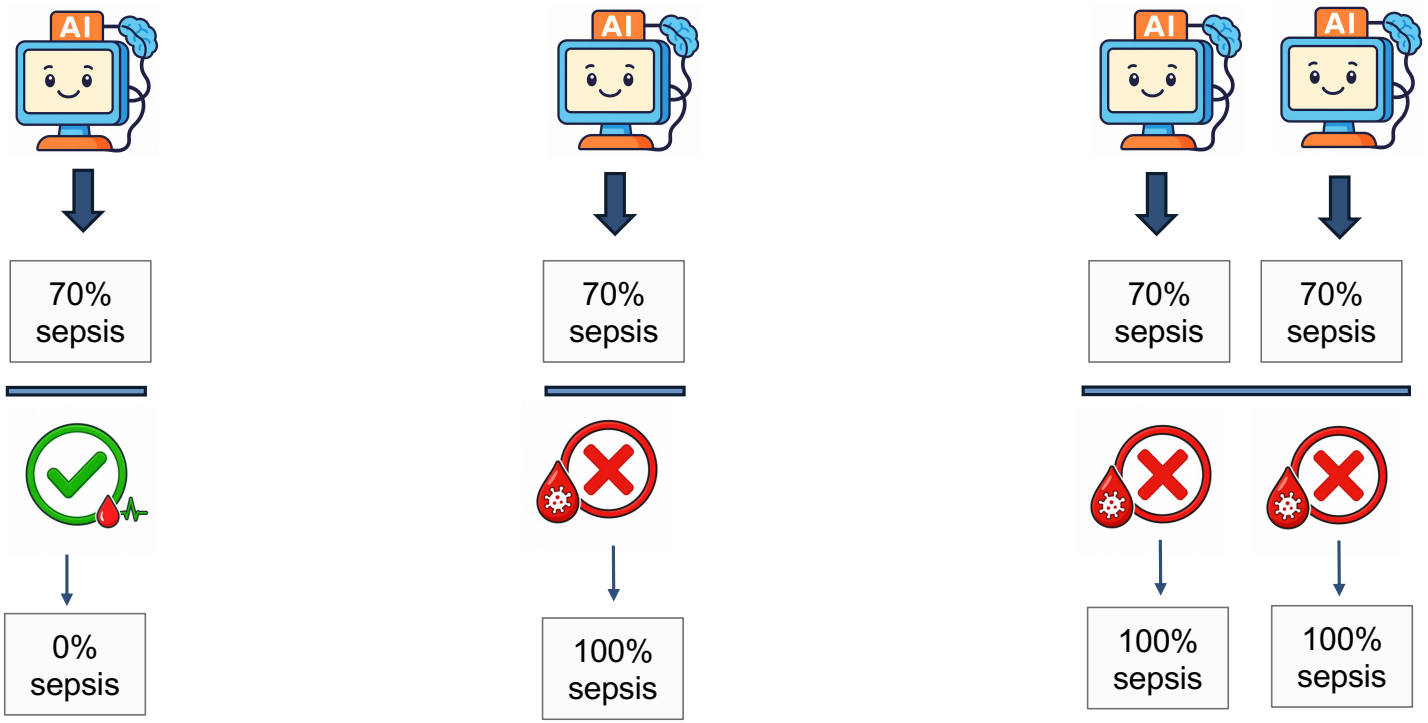
Could be worse—maybe your predictor is biased here—but let's be generous

Post-hoc calibration only decreases squared error



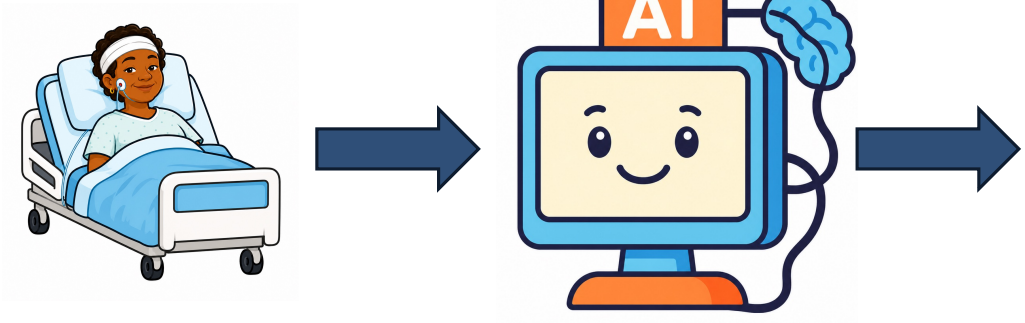
Meanwhile, on this same set of instances when using post-hoc calibration, the *worst case scenario* for squared error is that 70% of the patients develop sepsis

Post-hoc calibration only decreases squared error

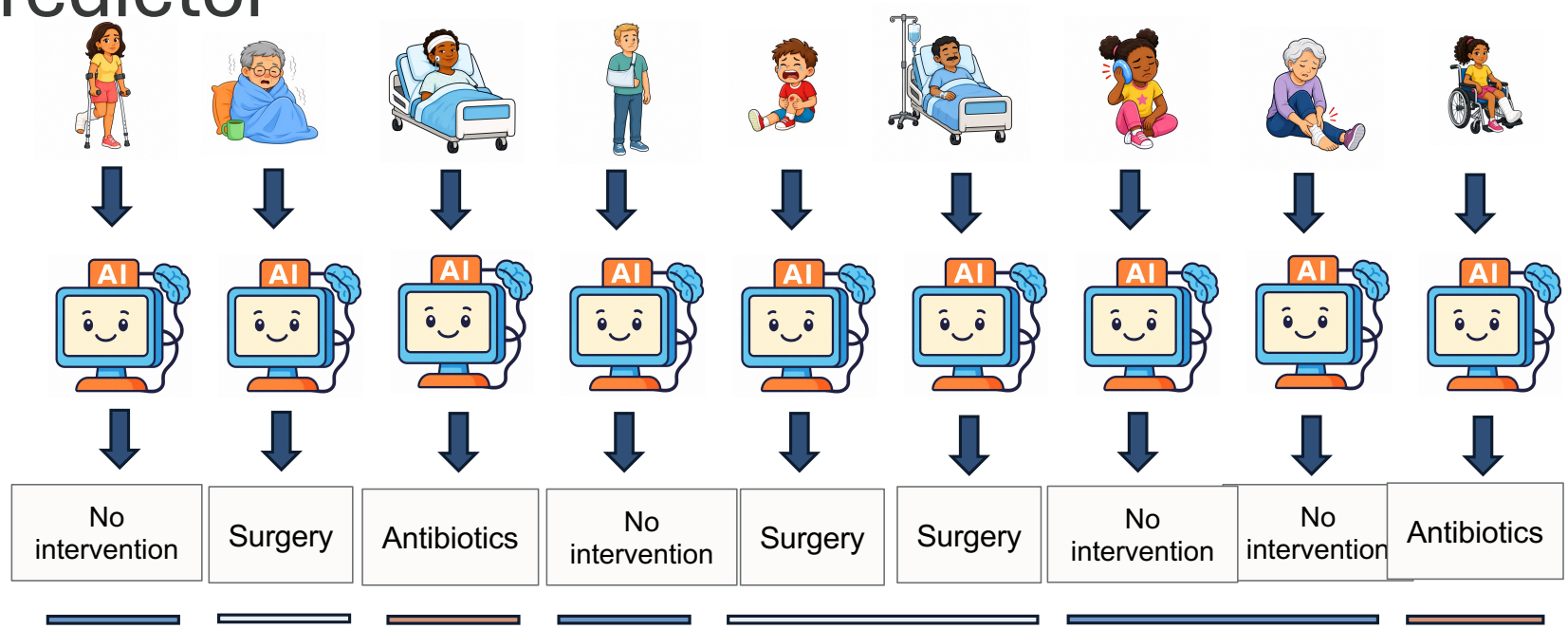


Could be better! Maybe we predict perfectly! But calibration guarantees, at the very least, *unbiasedness*.

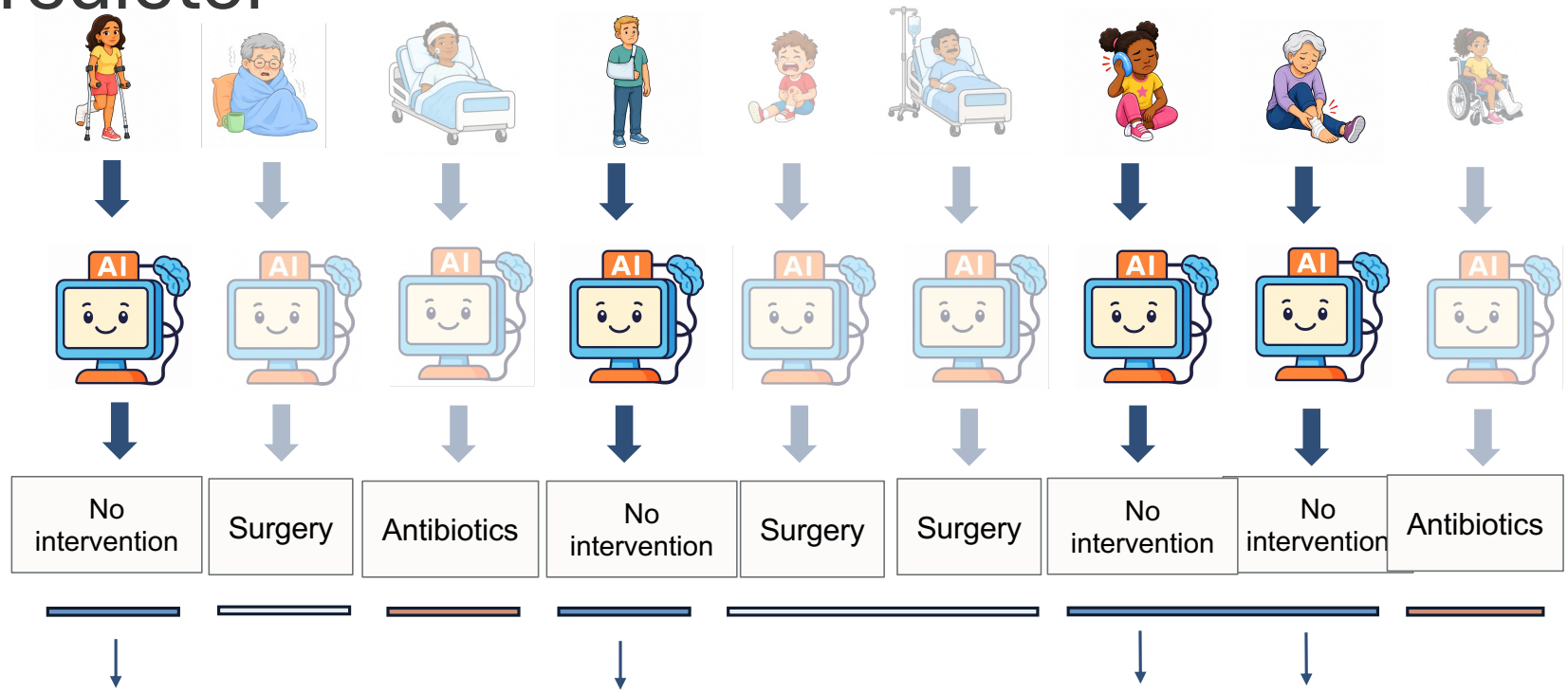
Post-hoc *decision* calibration on your favorite predictor



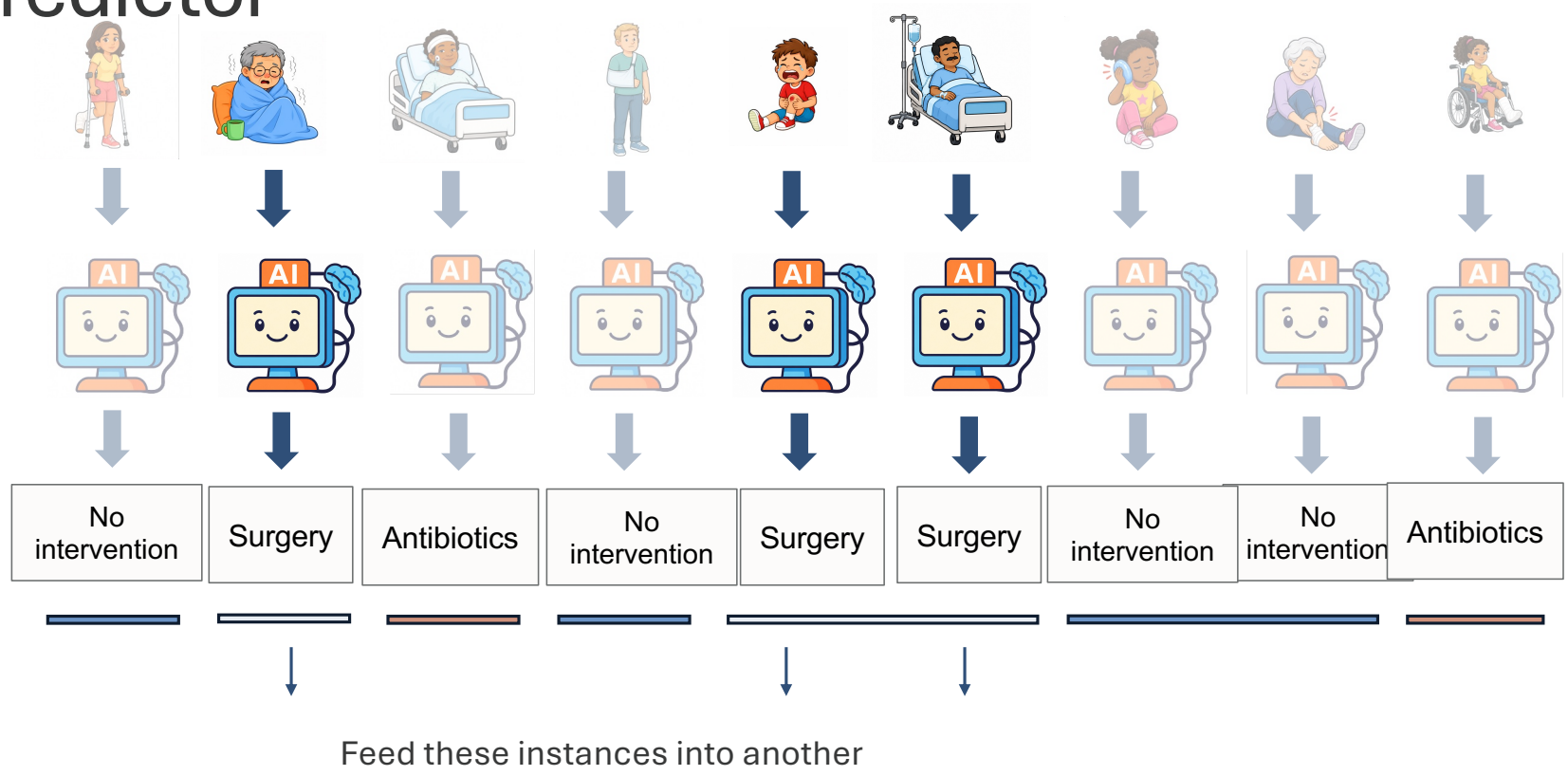
Post-hoc decision calibration on your favorite predictor



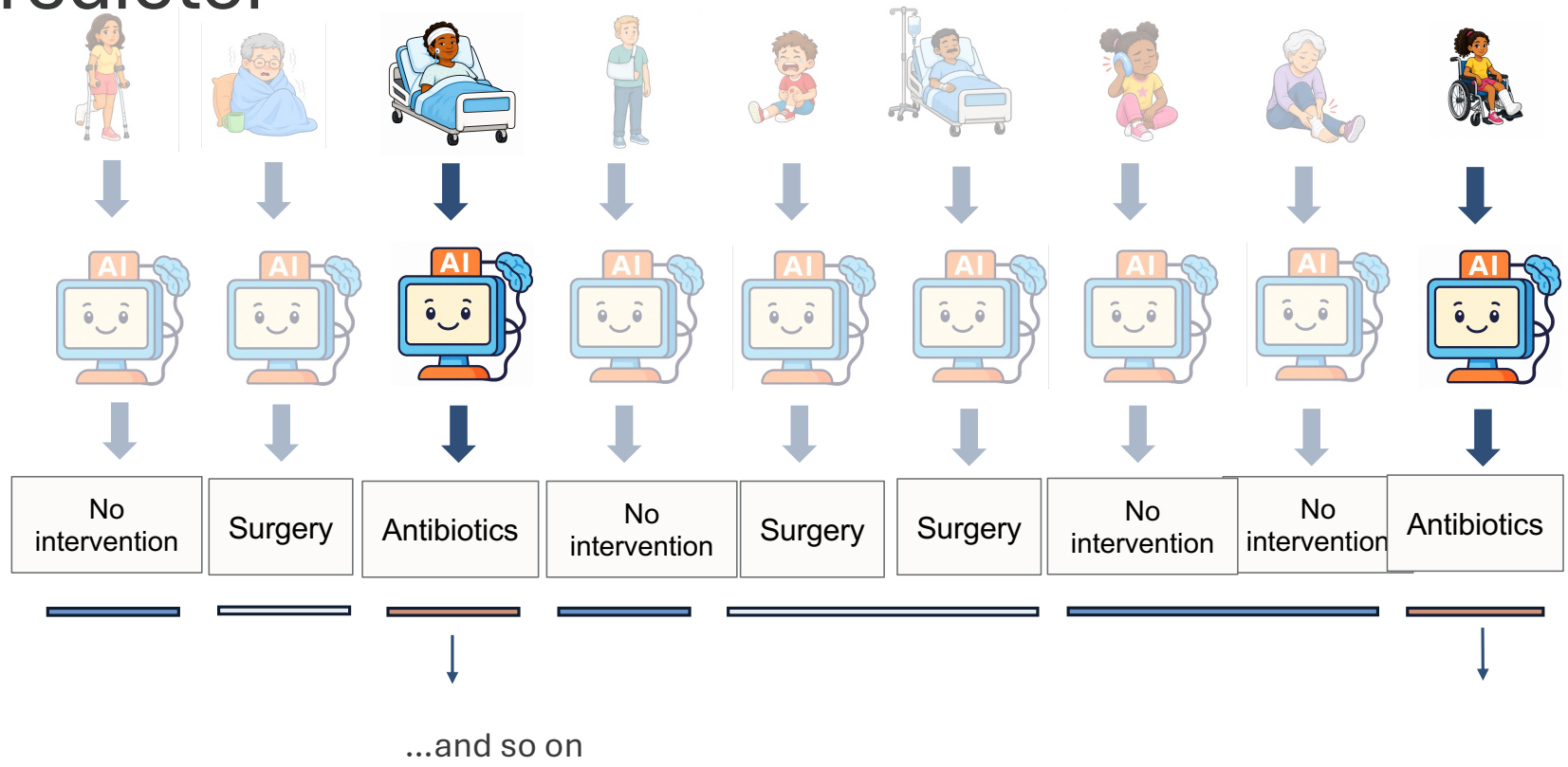
Post-hoc decision calibration on your favorite predictor



Post-hoc decision calibration on your favorite predictor



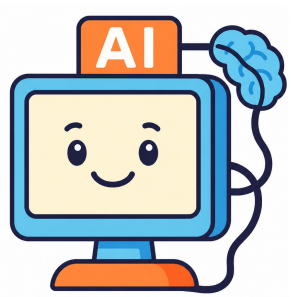
Post-hoc decision calibration on your favorite predictor



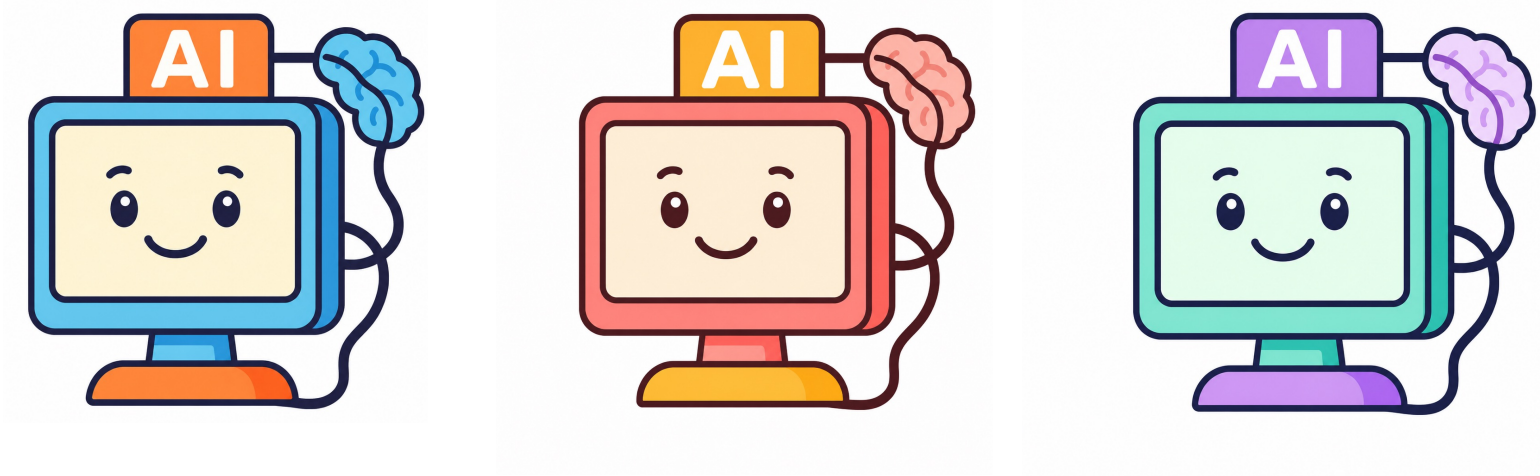
Post-hoc decision calibration on your favorite predictor

On average over the times I take a given action, it was the *most* effective action...

And my medical interventions are at least as effective as those proposed by my favorite model!

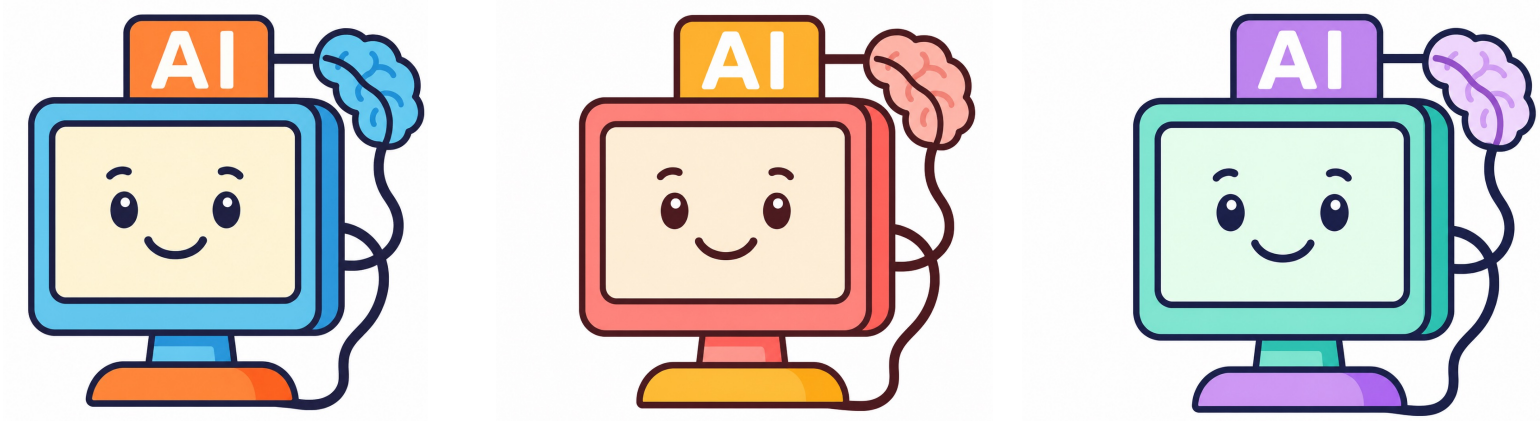


But what if I want to compete with *multiple* predictors?



Could do the same trick with *intersections* of outputs...
but the # of intersections scales **exponentially** with # of predictors

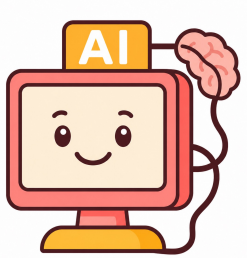
But what if I want to compete with *multiple* predictors?



More general trick: outperforming any given predictor is just another **outcome indistinguishability test**

But what if I want to compete with *multiple* predictors?

Real world



$\mathbb{E}[(M\ (x) - y)^2 - (f(x) - y)^2]$ should look the same in both worlds

Optimistica

In our standard OI form:

For n predictors, requires n tests

For outperforming a model M: The Real World and Optimistica should be indistinguishable with respect to $\mathbb{E}[(M(x) - f(x))y]$

Calibeating

- Note that the previous argument was quite lossy…
- There are formal ways in which the described procedure is even better than just *do no harm*
- Specifically, for our new predictor p,

$$\text{SQErr}(p) \leq \text{SQErr}(\text{AI}) - \text{CalErr}(\text{AI})$$

- The new predictor *improves* over the old one by at least the old predictor's calibration error
- This is a notion called *calibeating*, introduced in [Foster, Hart 2023]

Tractable Agreement and Information Aggregation

Clever debiasing

Make predictions unbiased on the right subsequences

Interpretable probabilities

Meaningful, trustworthy predictions.

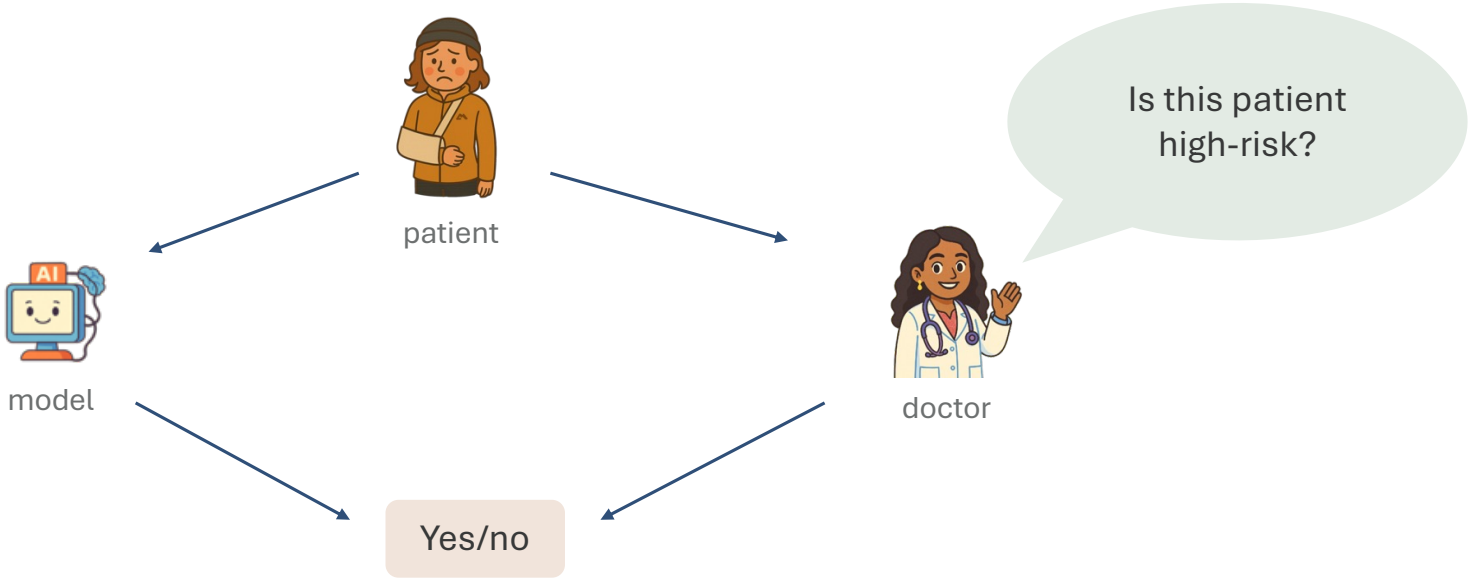
Good decision-making

Use predictions well for downstream decisions.

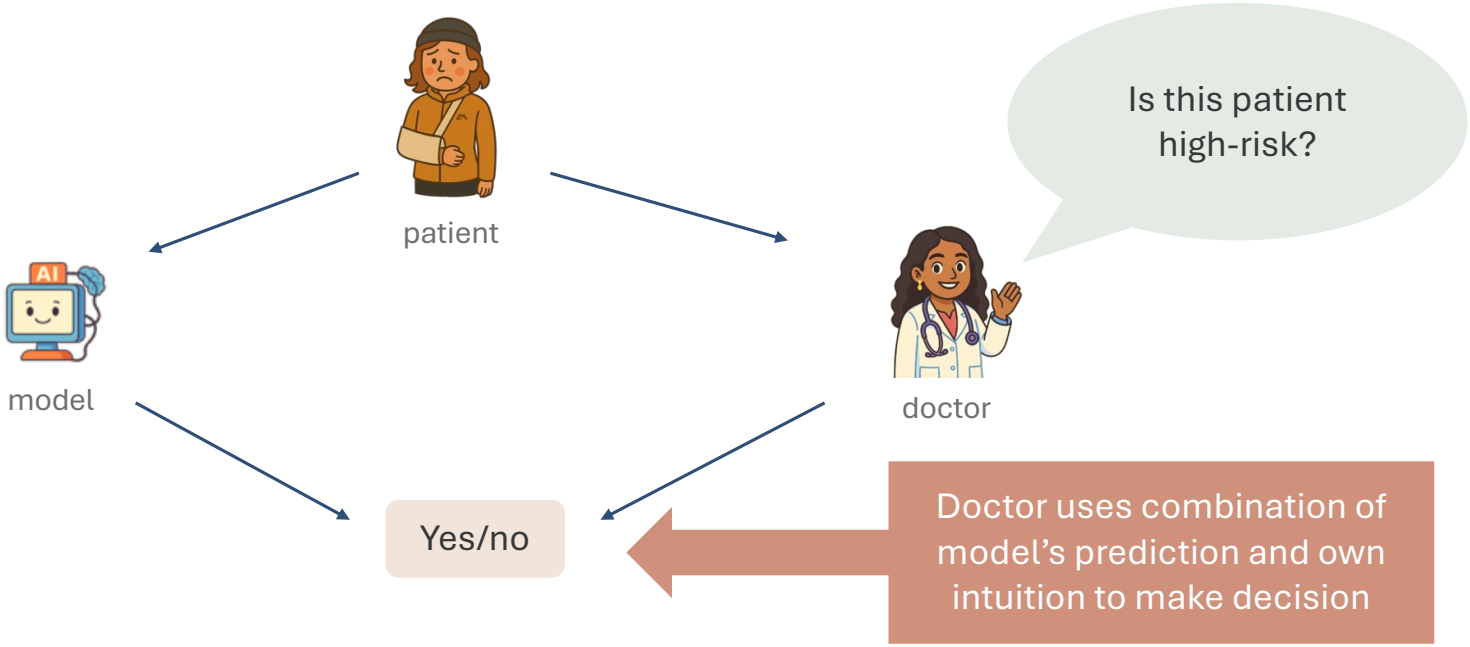
Collaborative learning

Learn efficiently from distributed information.

In reality, we use ML tools *collaboratively*

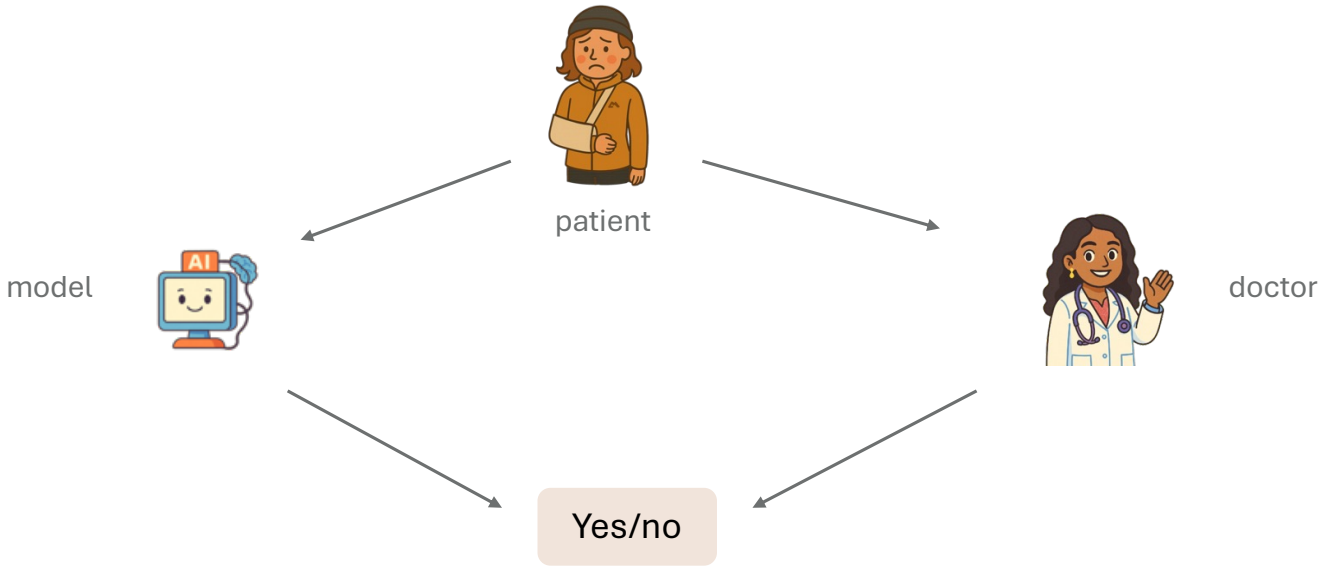


In reality, we use ML tools *collaboratively*



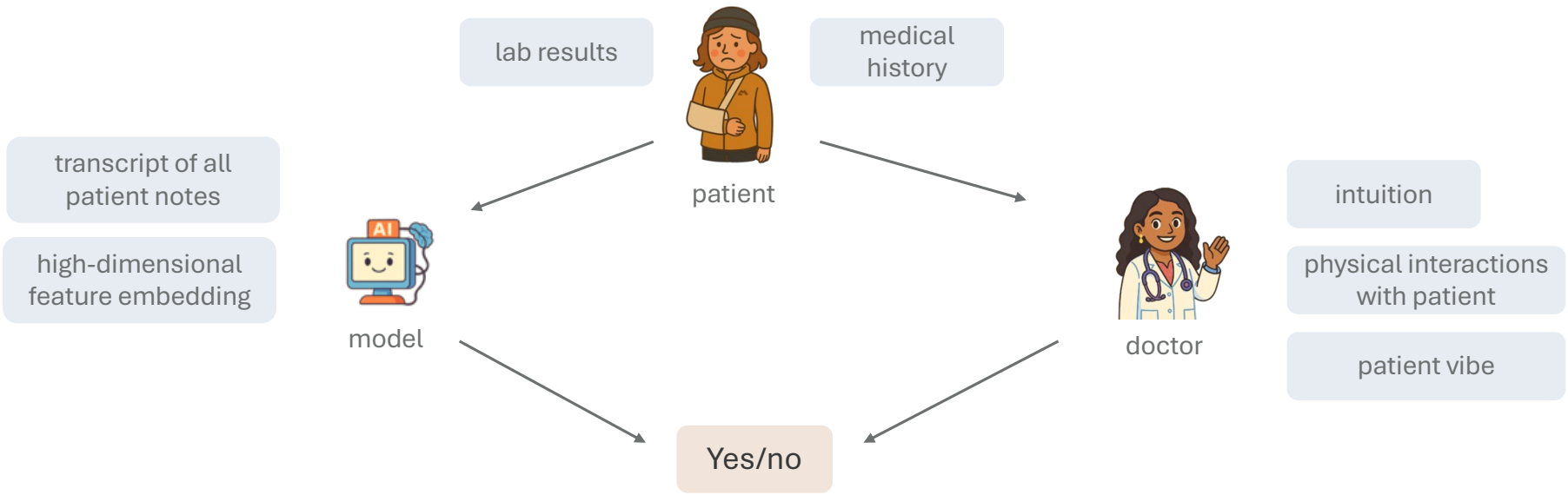
A Model for Human-AI Collaboration

How should we model these collaborative settings, where human insights matter?



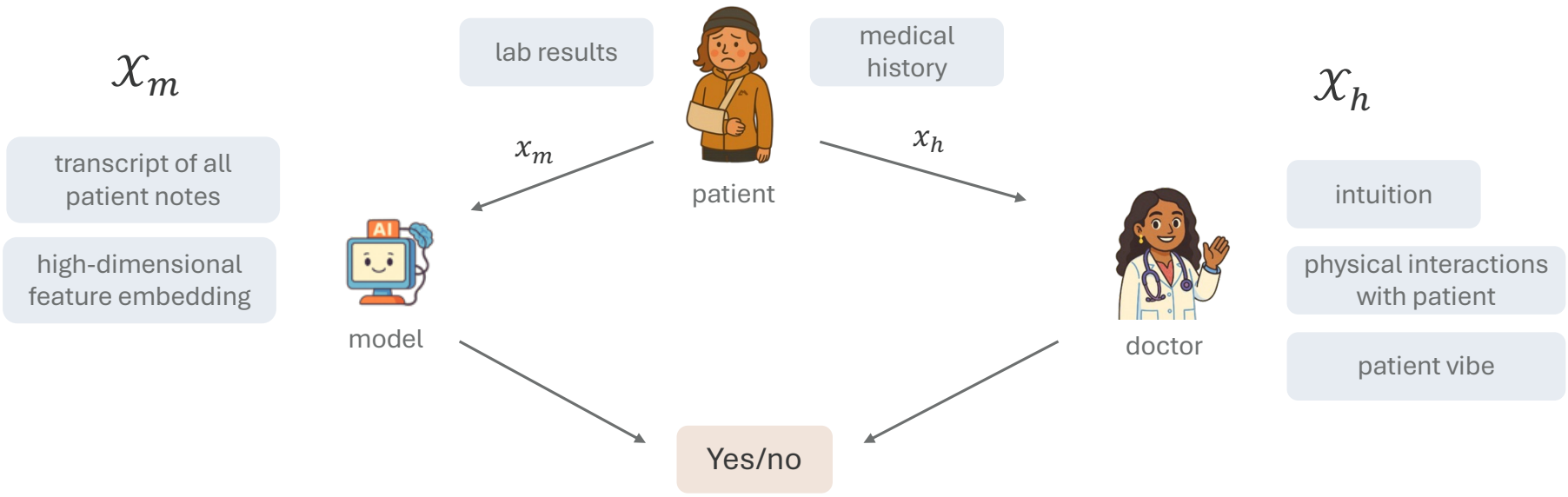
A Model for Human-AI Collaboration

Here, the patient's features are **distributed** between the human and model.



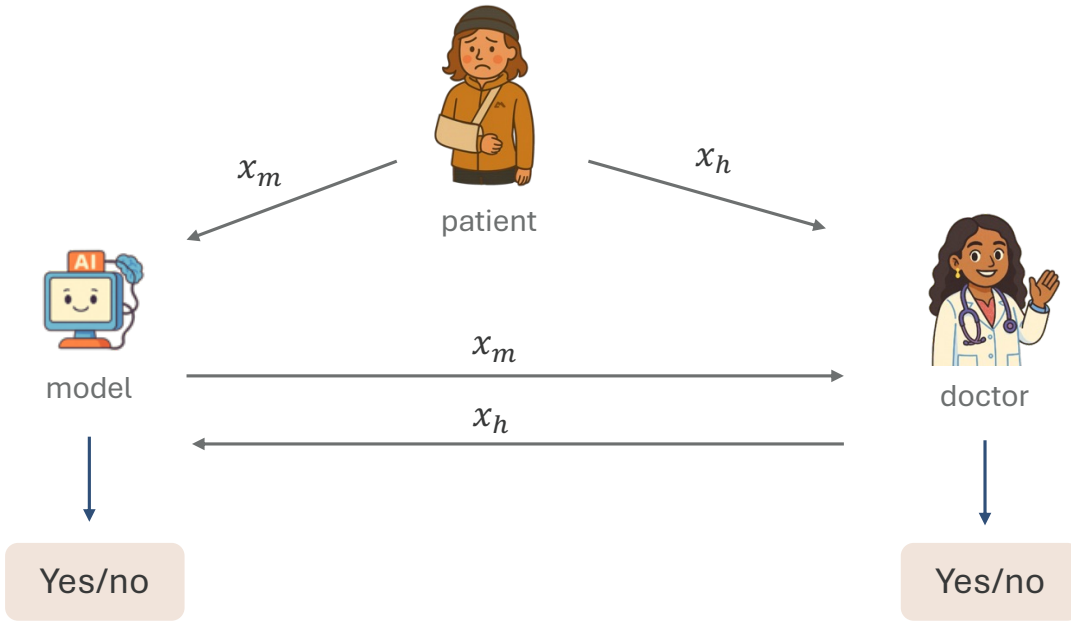
A Model for Human-AI Collaboration

Here, the patient's features are **distributed** between the human and model.



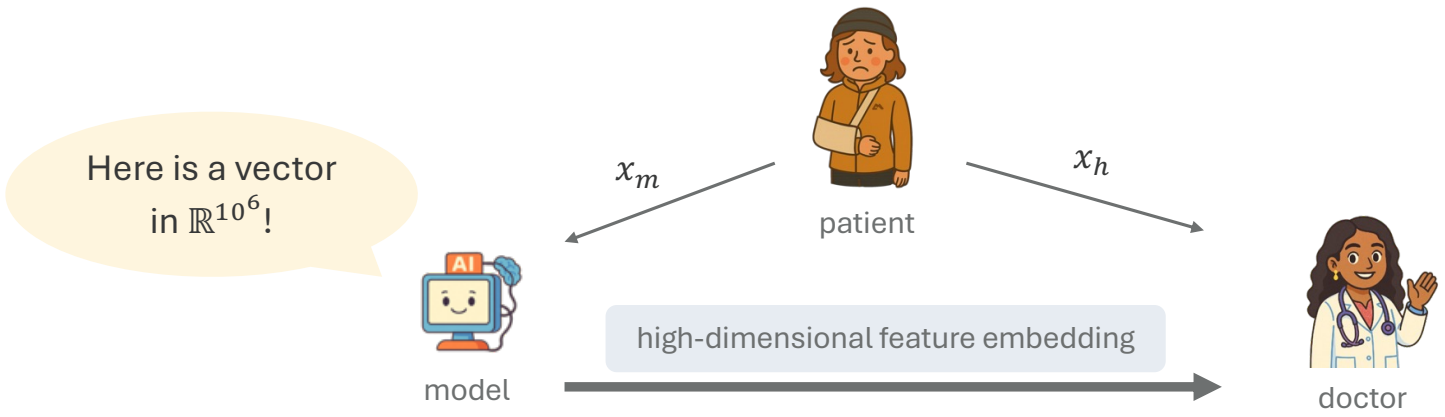
A Model for Human-AI Collaboration

In a perfect world, the doctor and model could access all the features.



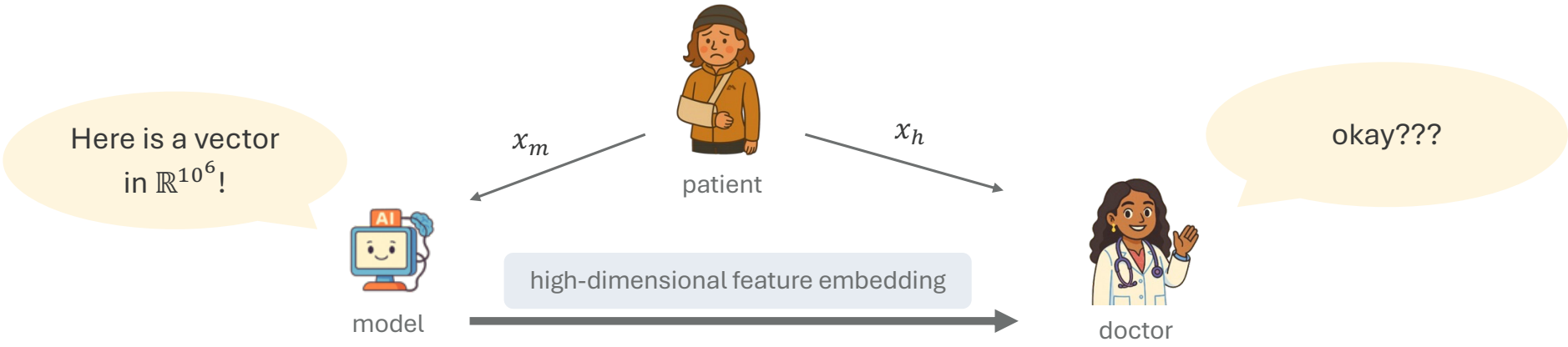
A Model for Human-AI Collaboration

In a perfect world, the doctor and model could access all the features. **This is too strong an assumption.**



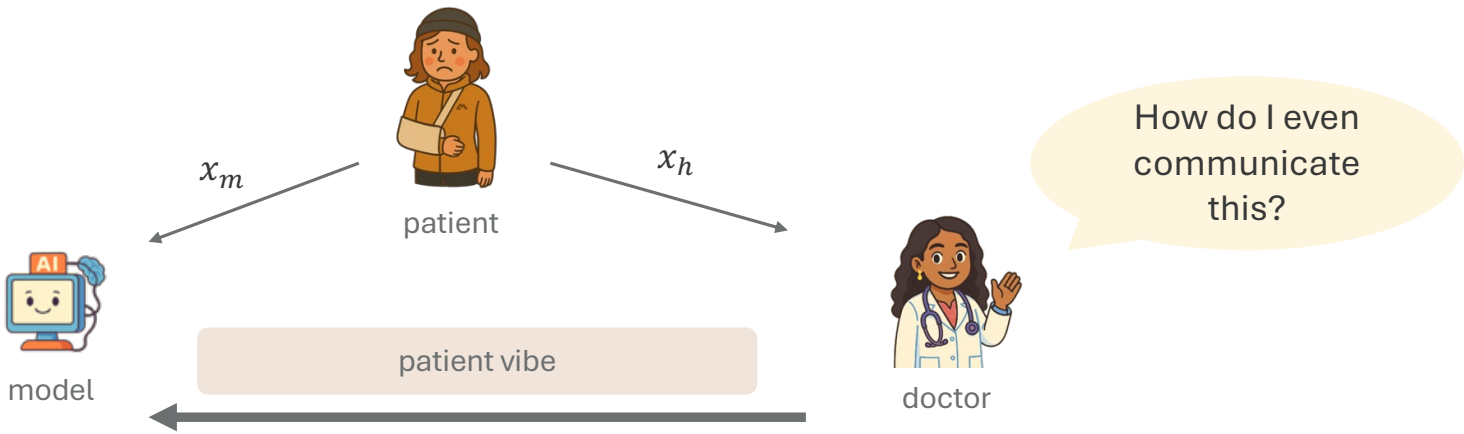
A Model for Human-AI Collaboration

In a perfect world, the doctor and model could access all the features. **This is too strong an assumption.**



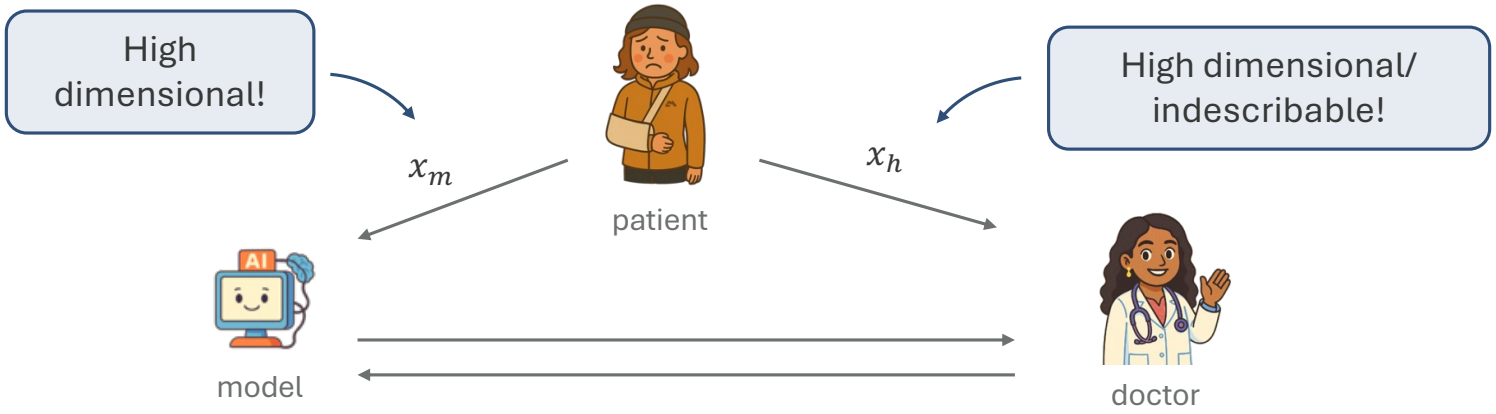
A Model for Human-AI Collaboration

In a perfect world, the doctor and model could access all the features. **This is too strong an assumption.**



A Model for Human-AI Collaboration

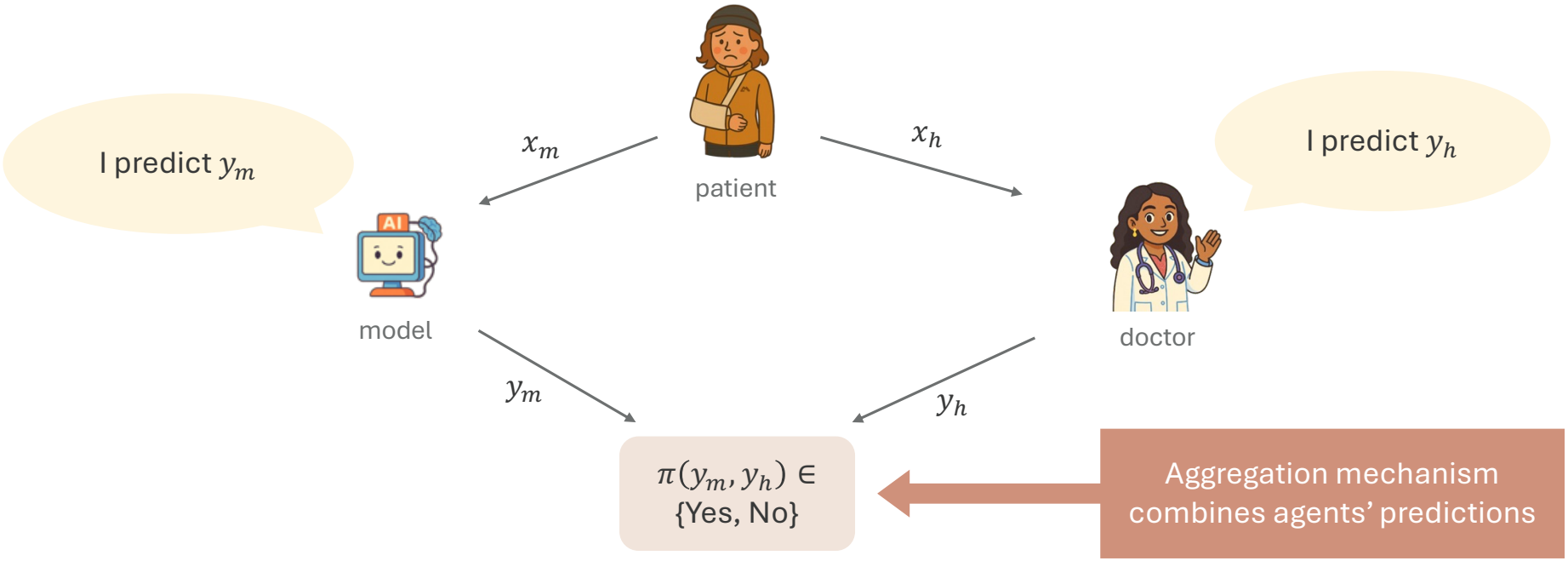
In a perfect world, the doctor and model could access all the features. **This is too strong an assumption.**



Even if we *could* do this, we'd have to communicate a **huge amount** of information per patient. **So what can we do?**

A Model for Human-AI Collaboration

If each agent makes a prediction, could we just have them share those and then aggregate?



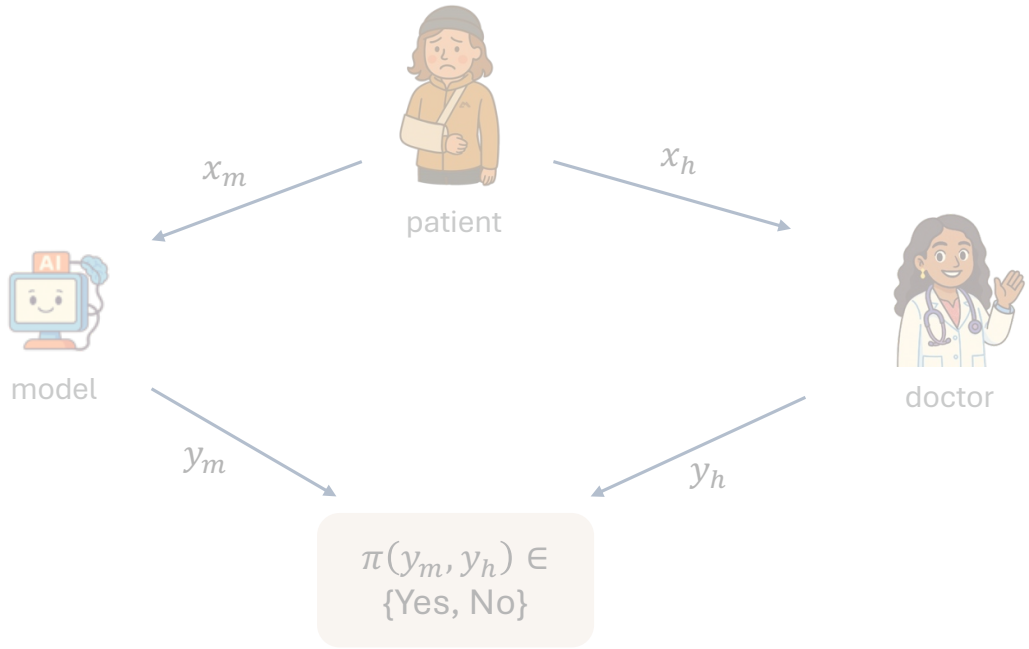
A Model for Human-AI Collaboration

If each agent makes a prediction, could we just have them share those and then aggregate? **Too weak.**

“No free lunch” for collaboration

No aggregation rule (voting, averaging predictions, etc.) exists such that for any distribution, you can guarantee that the agents are better off combining their predictions than they would have been otherwise.

[Peng, Garg, Kleinberg ‘23]



A Model for Human-AI Collaboration

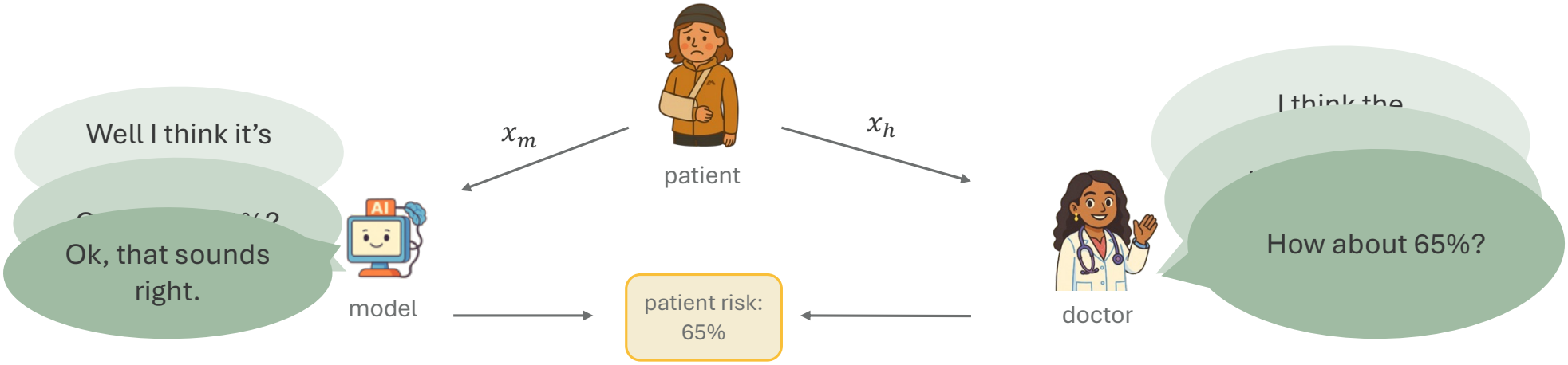
If each agent makes a prediction, could we just have them share those and then aggregate? **Too weak.**

So, is there no hope for collaborative learning?

There is! Huzzah!

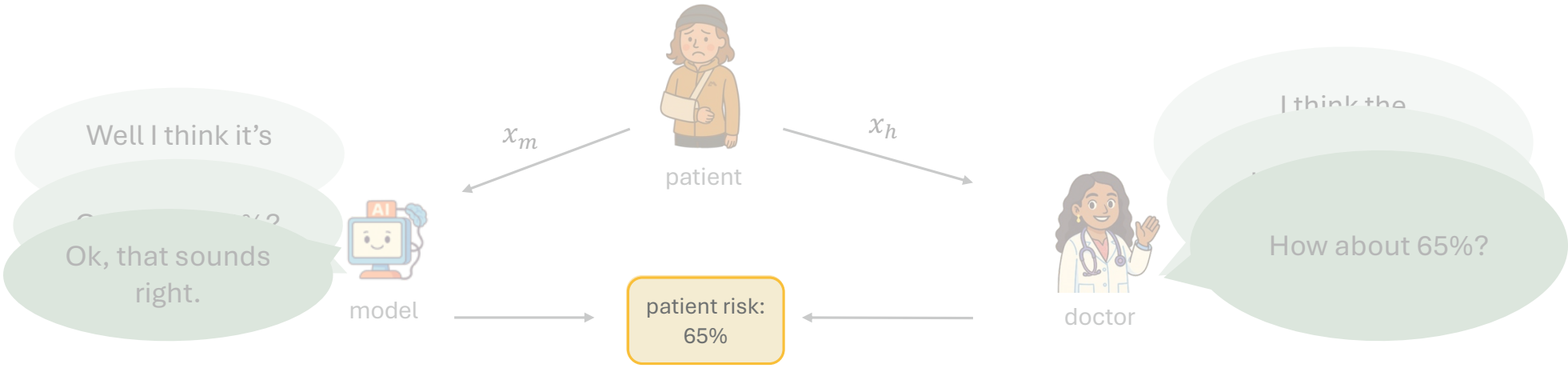
An *Interactive* Model for Human-AI Collaboration

We will avoid this intractability by letting the human and machine **interact** until they **agree**.



An *Interactive* Model for Human-AI Collaboration

We will avoid this intractability by letting the human and machine **interact** until they **agree**.

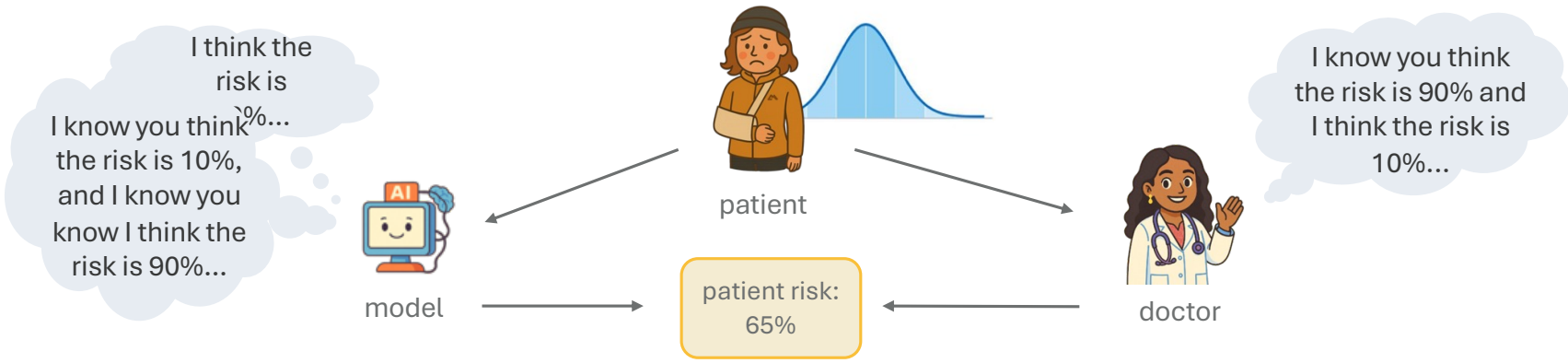


Note: we'll talk about this as if they are agreeing on a **predicted label** $y \in [0,1]$, but we can also think about communicating best-response actions.

Why might interaction until agreement help?

Theorem [Aumann '76]

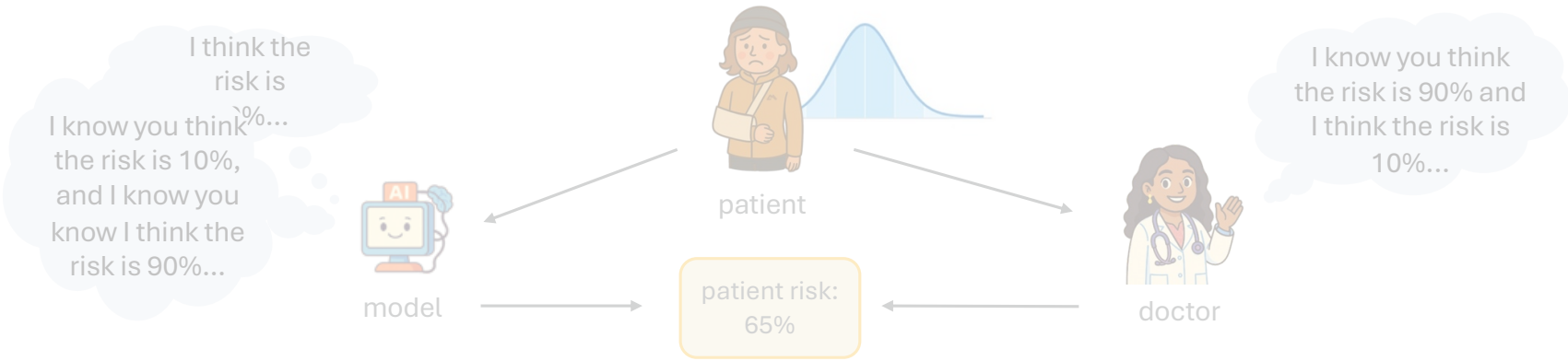
If two Bayesian agents have a common prior and common knowledge of each others' posterior expectation, the posterior expectations will be equal.



Why might interaction until agreement help?

Theorem [Aumann '76]

If two Bayesian agents have a common prior and common knowledge of each others' posterior expectation, the posterior expectations will be equal.



Bayesian agents with common knowledge **cannot agree to disagree.**

Why might interaction until agreement help?

Bayesian agents with common knowledge **cannot agree to disagree.**

But will they reach agreement quickly?

Why might interaction until agreement help?

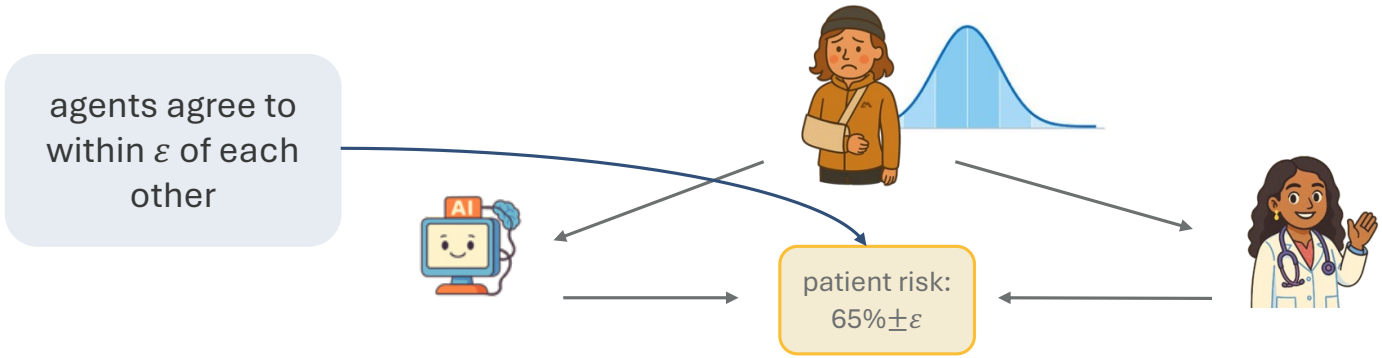
Theorem [Aaronson '04]

If the underlying state space is finite and the predictions are scalars, then with probability $1 - \delta$, the agents reach ε-agreement in $1/\varepsilon^2\delta$ rounds.

Why might interaction until agreement help?

Theorem [Aaronson '04]

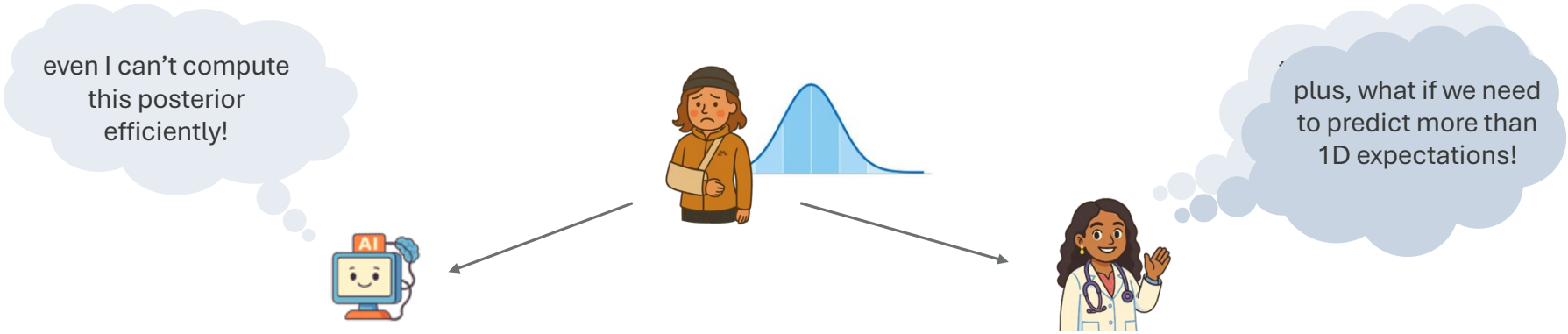
If the underlying state space is finite and the predictions are scalars, then with probability $1 - \delta$, the agents reach ε-agreement in $1/\varepsilon^2\delta$ rounds.



Bayesian agents with common knowledge **cannot agree to disagree** after only a polynomial number of rounds of communication of **one bit of information.**

Limitations of the Bayesian setting

In practice, will our agents have a common prior?



Limitations of the Bayesian setting

In practice, will our agents have a common prior?

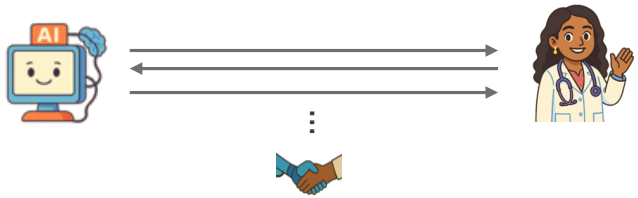
Can we **relax** these behavioral assumptions while still maintaining strong convergence guarantees? Can we guarantee they'll agree to something **good**?

Yes! Via clever conditional calibration!

What should our goals be?

Goal 1: Tractable Agreement

"On most days, the doctor and computer should agree (within some margin ε) on a patient's sepsis risk after few rounds of communication."



What should our goals be?

We'll do this in an **online learning** setting

Goal 1: Tractable Agreement

"On most days, the doctor and computer should agree (within some margin ε) on a patient's sepsis risk after few rounds of communication."

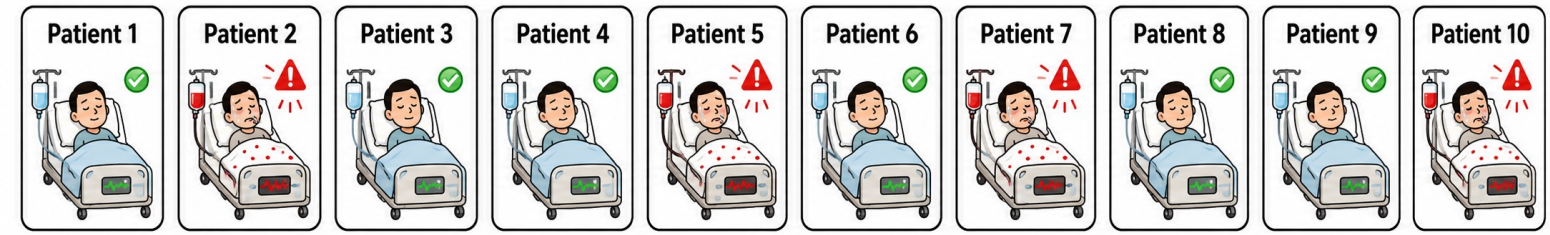
Goal 2: Information Aggregation

"The thing they agree on should be as good as it could be."

Online Agreement and Information Aggregation

Like in the first parts of this tutorial, we want results to hold in a general online and adversarial setting.

Every day, a new patient arrives

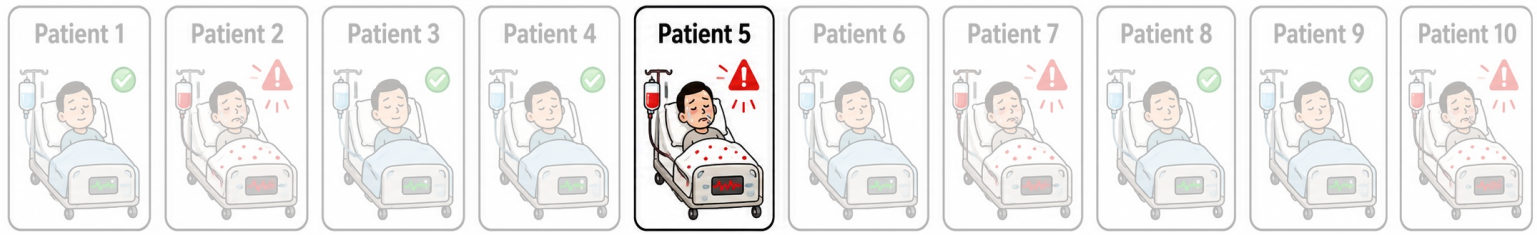


time t

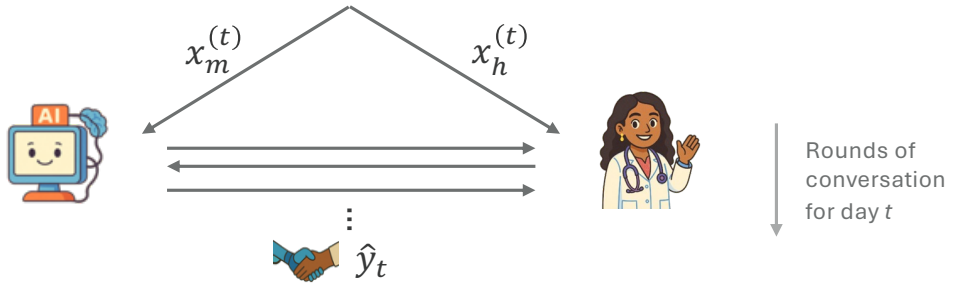
Online Agreement and Information Aggregation

Like in the first parts of this tutorial, we want results to hold in a general online and adversarial setting.

Every day t a patient arrives



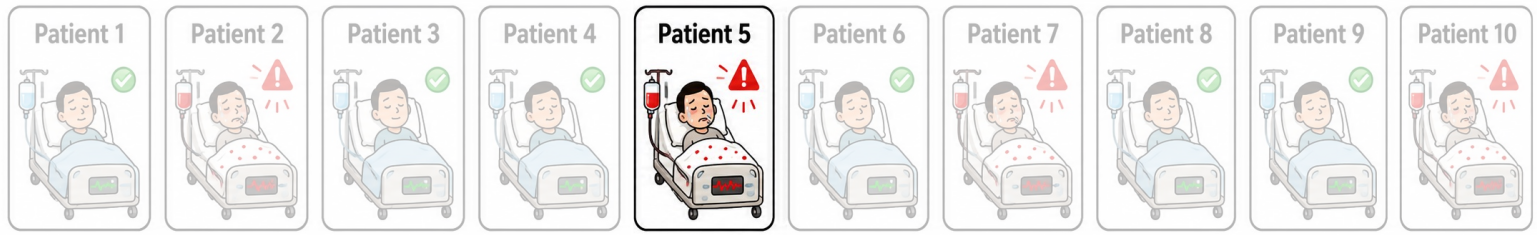
On each day t the doctor and model converse for k_t rounds until they agree on a prediction $\hat{y}_t$



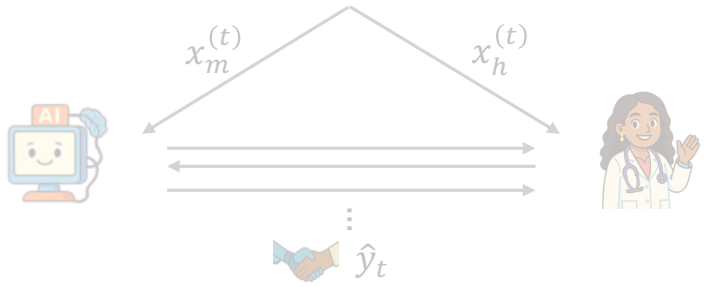
Online Agreement and Information Aggregation

Like in the first parts of this tutorial, we want results to hold in a general online and adversarial setting.

Every day t a patient arrives



On each day t the doctor and model converse for k_t rounds until they agree on a prediction $\hat{y}_t$



At the end of day t, we learn the true outcome y_t

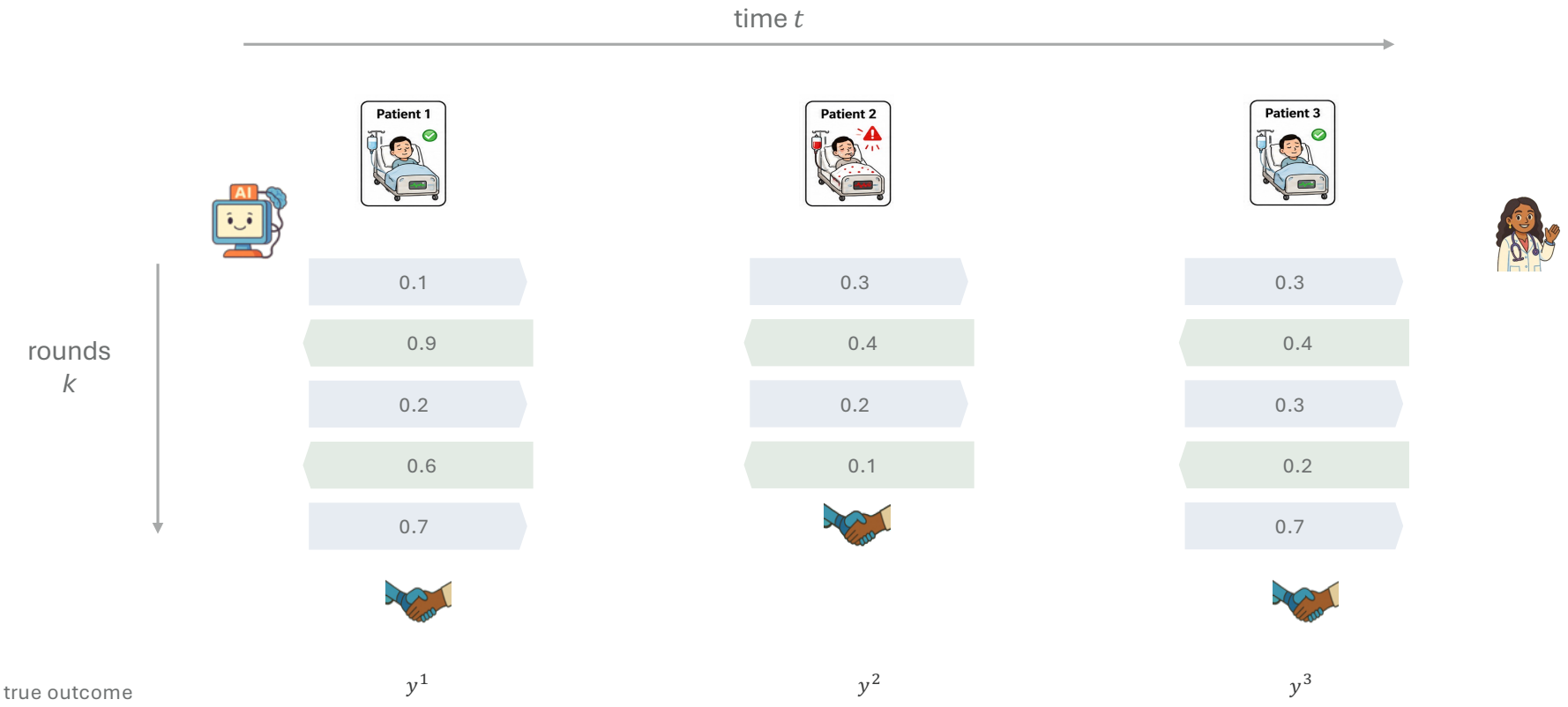
$$y_t \in \{sepsis, no\ sepsis\}$$

Goal 1:Tractable Agreement

Previously, we saw how we could cleverly enforce **conditional calibration** to get **interpretable probabilities** and to do well on **downstream decisions**. Can we do something similar here?

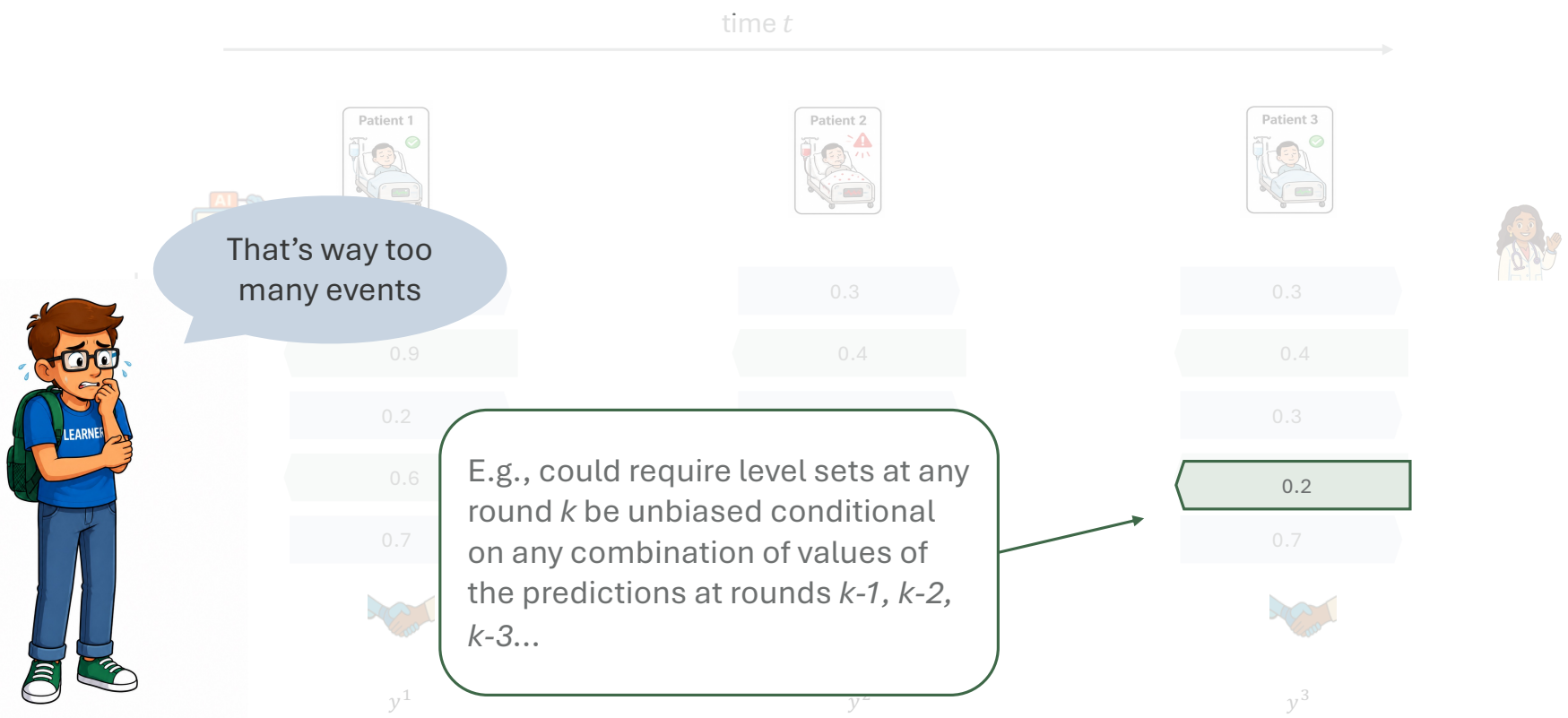
What if we required calibration conditional on the **transcript of conversations?**

What should our conditioning event be?

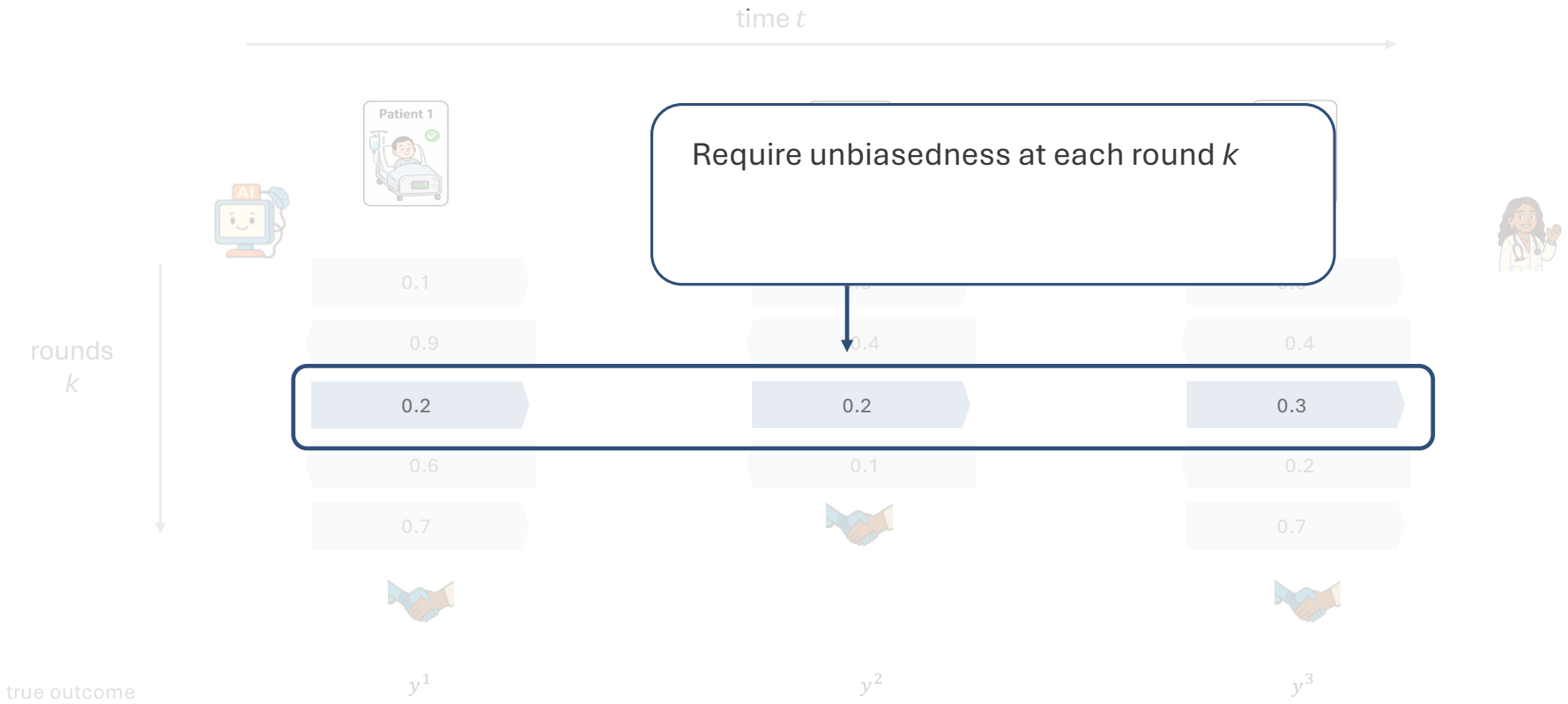


We could try to be unbiased conditional on all possible interactions in previous rounds…

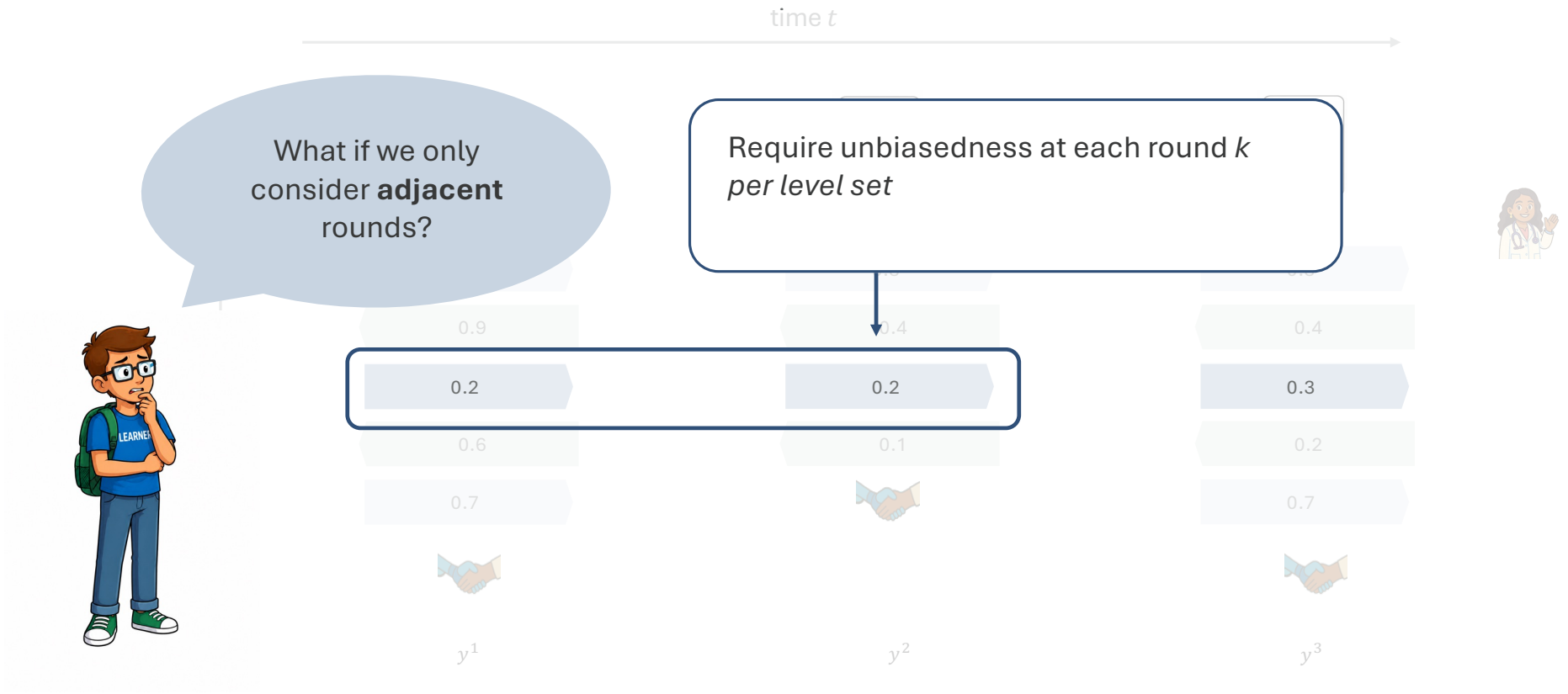
We could try to be unbiased conditional on all possible interactions in previous rounds…



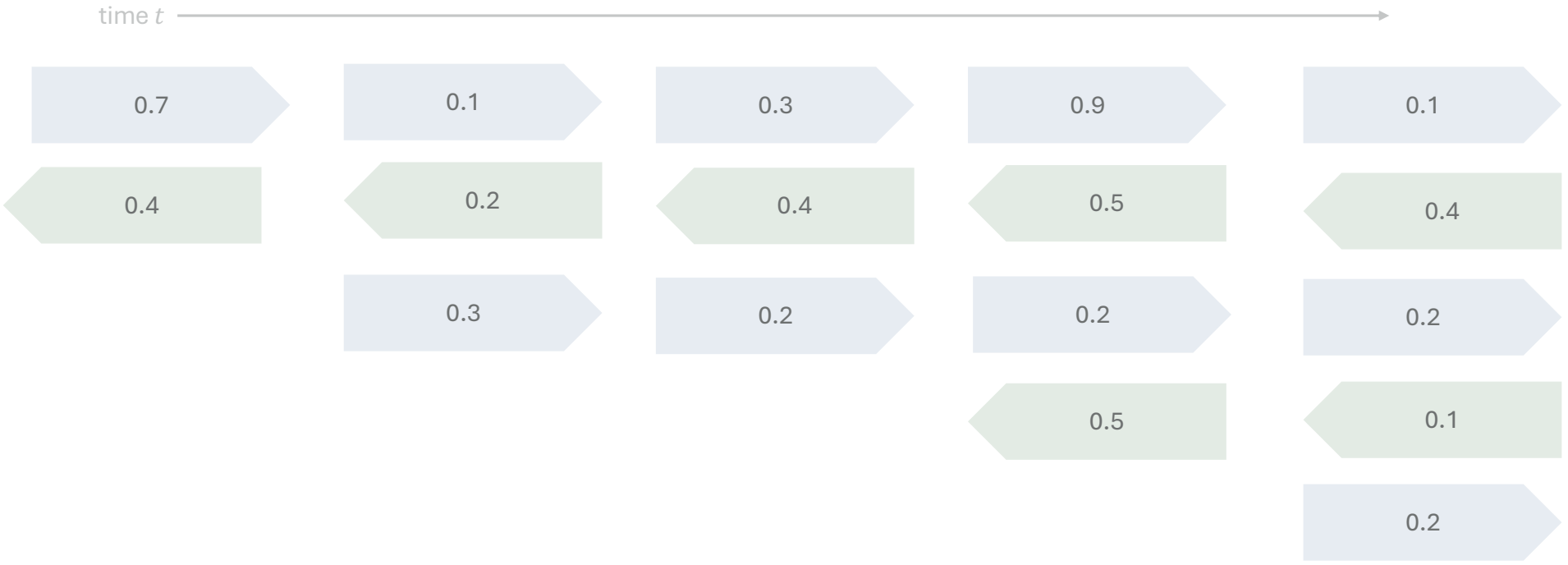
Or try to be unbiased conditional on our predictions per round...



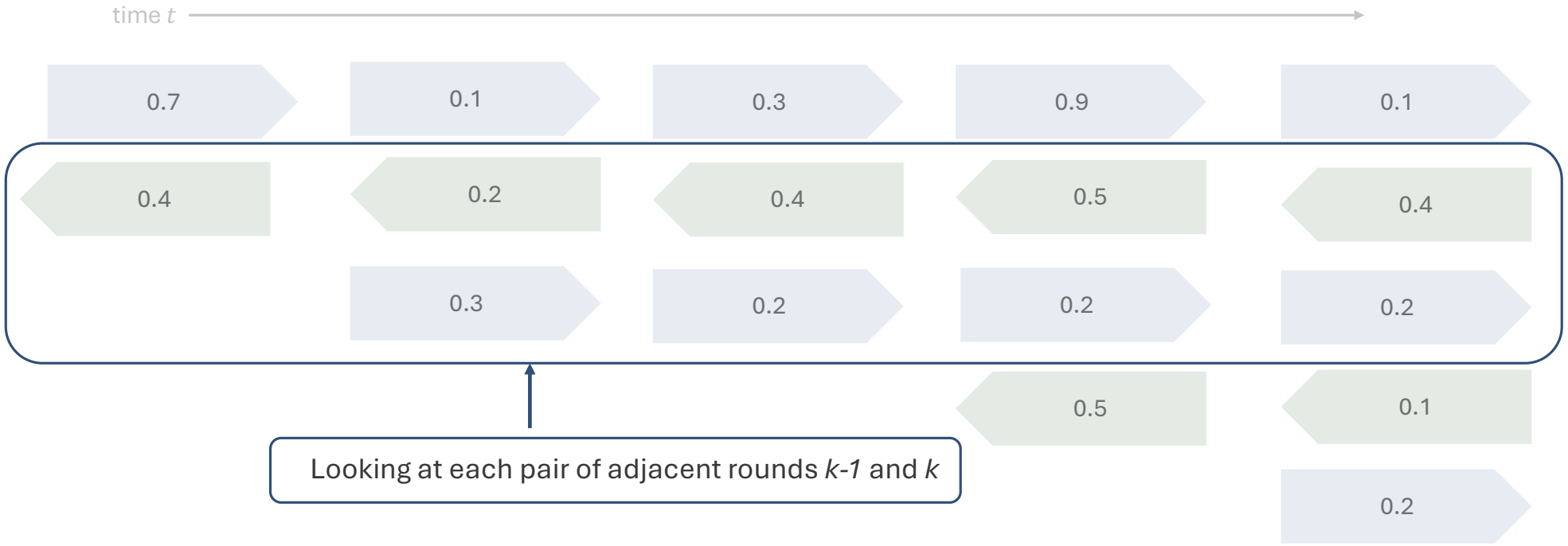
Or just try to be unbiased conditional on our predictions per round…



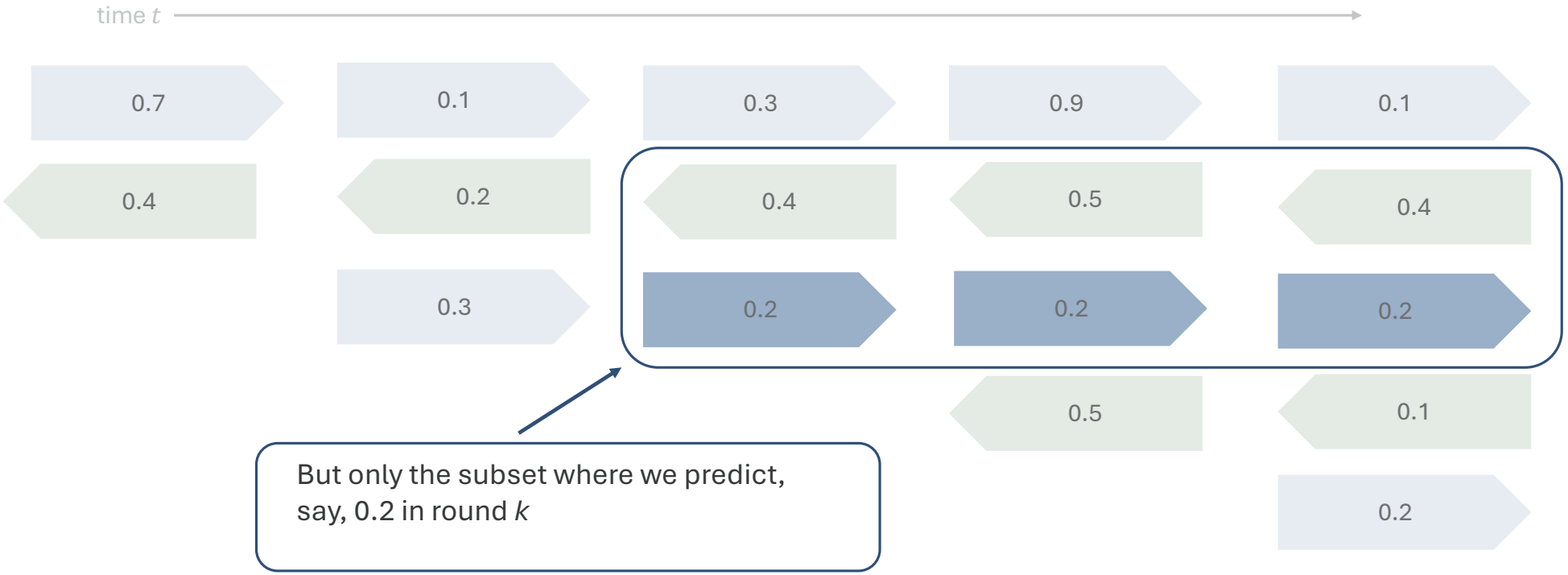
Conversation calibration conditions over *adjacent rounds*



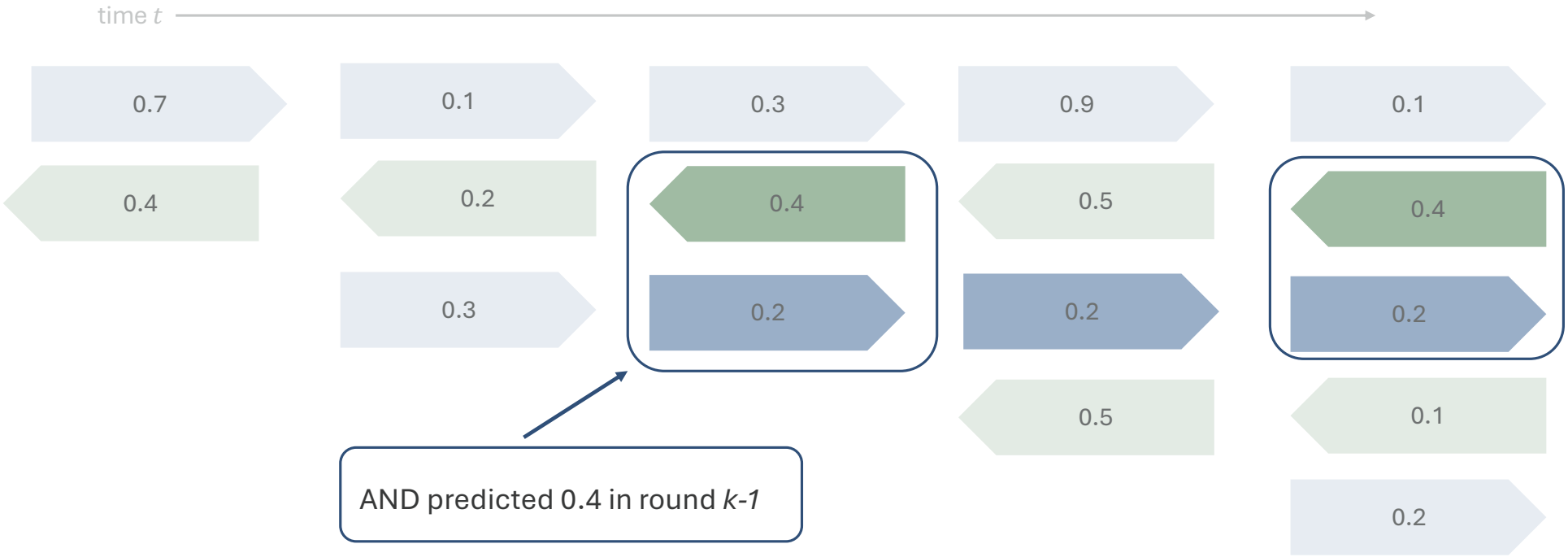
Conversation calibration conditions over *adjacent rounds*



Conversation calibration conditions over *adjacent rounds*



Conversation calibration conditions over *adjacent rounds*



(We'll require this for all adjacent rounds and possible combinations of predictions)

Conversation calibration

An agent is conversation calibrated if for every round of conversation k, they are calibrated *jointly on their prediction p and the other agent's prediction p' at round k − 1*.

E.g. the human is conversation calibrated if for all rounds k and predictions p, p'

$$\mathbb{E}[p - y | \, human\ predicts\ p\ at\ round\ k, model\ predicts\ p'\ at\ round\ k - 1] \approx 0$$

Conversation Calibration ⇒ Fast Agreement

Observation: If sequence 1 is calibrated conditional on sequence 2, it has weakly lower squared error.

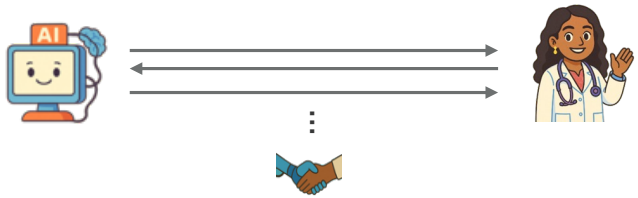
And, if subsequence 2 is substantially **different** than subsequence 1, it must **improve** squared error.

So, between adjacent rounds, must either get close to agreement, or substantially drop squared error. And you can't do that too many times!

So, we have obtained tractable agreement by enforcing **conversation calibration**!

Goal 1: Tractable Agreement

"On most days, the doctor and computer should agree (within some margin ε) on a patient's sepsis risk after few rounds of communication."



But did we agree on **something good**?

Goal 1: Tractable Agreement

"On most days, the doctor and computer should agree (within some margin ε) on a patient's sepsis risk after few rounds of communication."

Goal 2: Information Aggregation

"The thing they agree on should be as good as it could be."

Goal 2: Information Aggregation

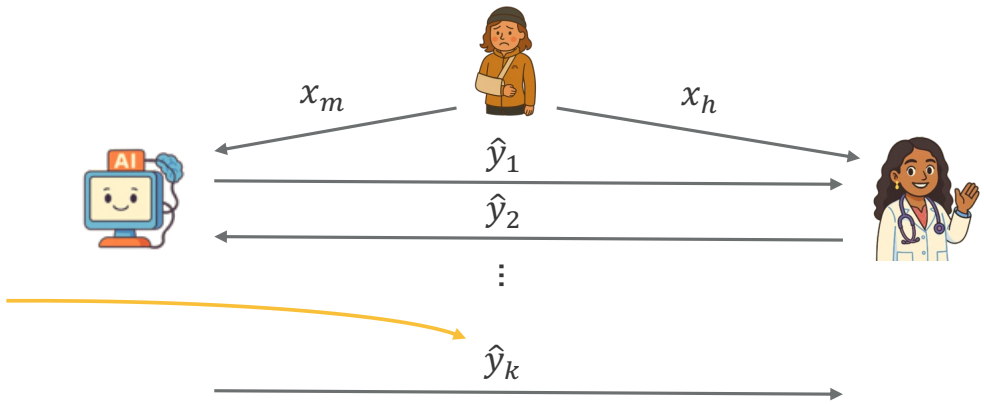
What does it mean for collaboration to be **worth it**?

One option: collaboration shouldn’t hurt either party.

Collaboration does no harm.

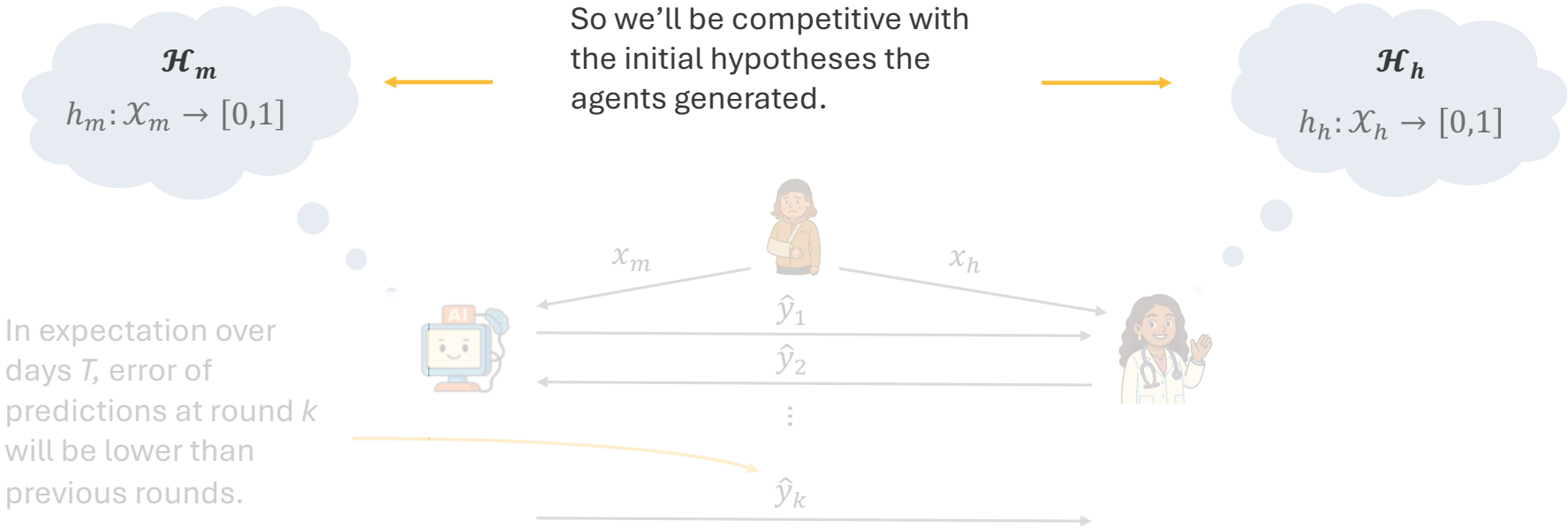
If the agents are conversation calibrated, collaboration only improves their expected prediction error.

In expectation over days *T,* error of predictions at round *k* will be lower than previous rounds.



Collaboration does no harm.

If the agents are conversation calibrated, collaboration only improves their expected prediction error.



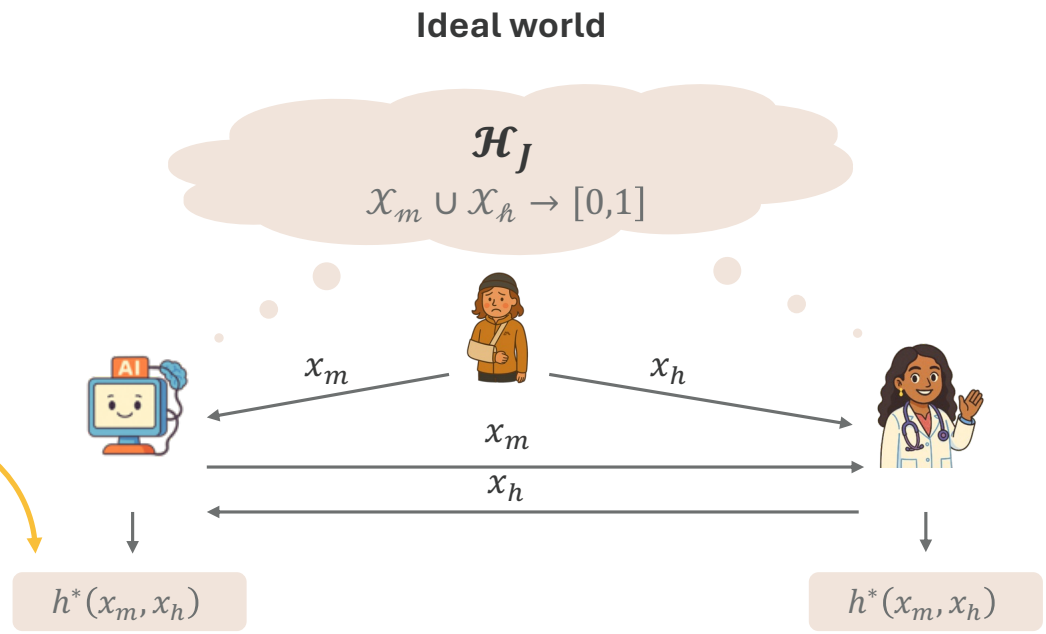
Goal 2: Information Aggregation

But can we do better?

Goal 2: Information Aggregation

Strongest benchmark: If everyone communicates their features, they could compute the best hypothesis over this full joint space.

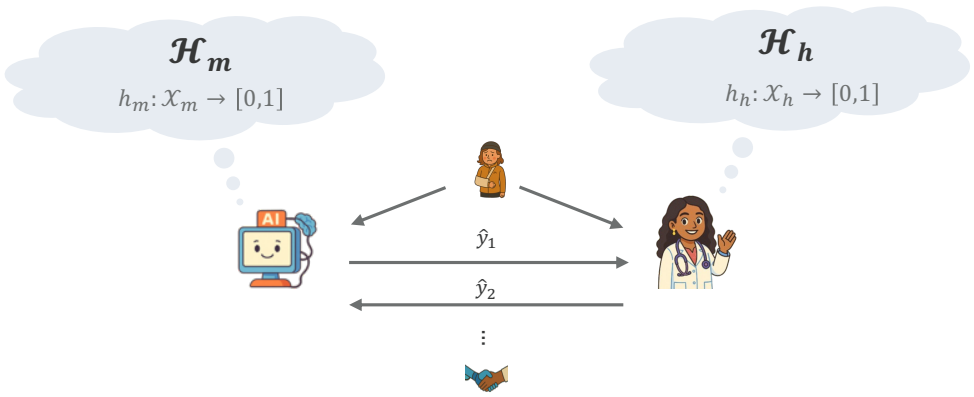
Can we ever compete with this best model if we are restricted to only sharing predictions?



Goal 2: Information Aggregation

Our goal: **compete with $\mathcal{H}_J$**, even though we don't have access to it!

A more realistic world



In real world, can only communicate predictions learned from own hypothesis class + shared predictions

Ideal world

In ideal world, can learn best $h(x_m, x_h)$

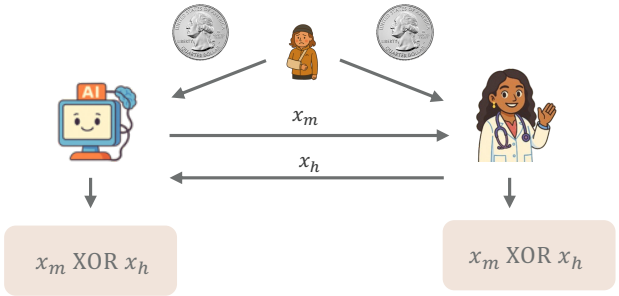
Should it even be possible to compete with $\mathcal{H}_J$?

Counter-example: Say $X_m, X_h \sim \text{Bern}(1/2)$, and $Y = X_m \text{ XOR } X_h$.

A more realistic world

In real world, everyone guesses ½.

Ideal world



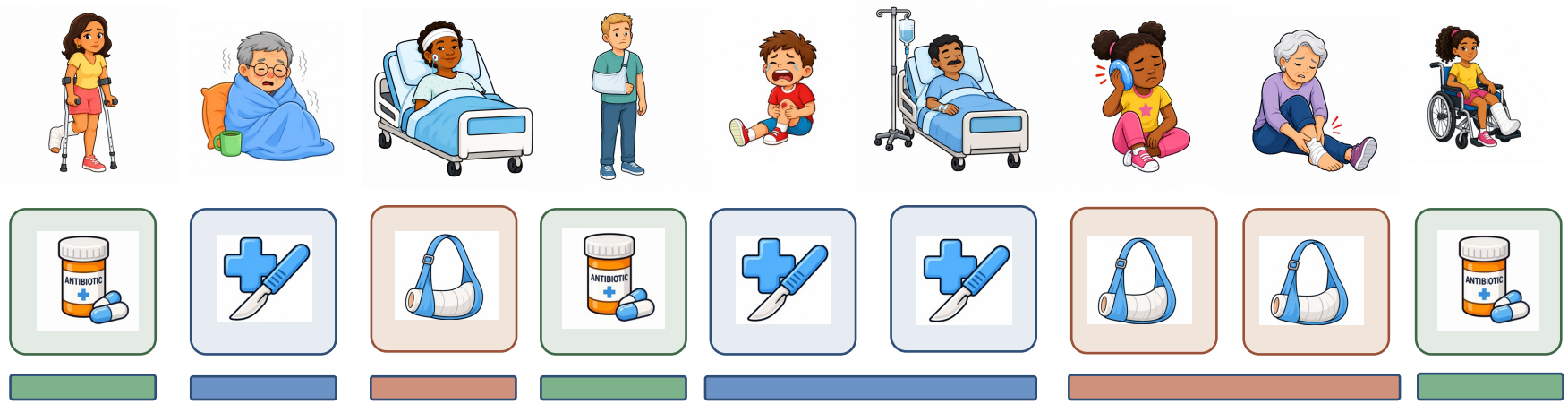
In ideal world, can learn outcome perfectly

Goal 2: Information Aggregation

New goal: characterize **when** we can compete with $\mathcal{H}_J$

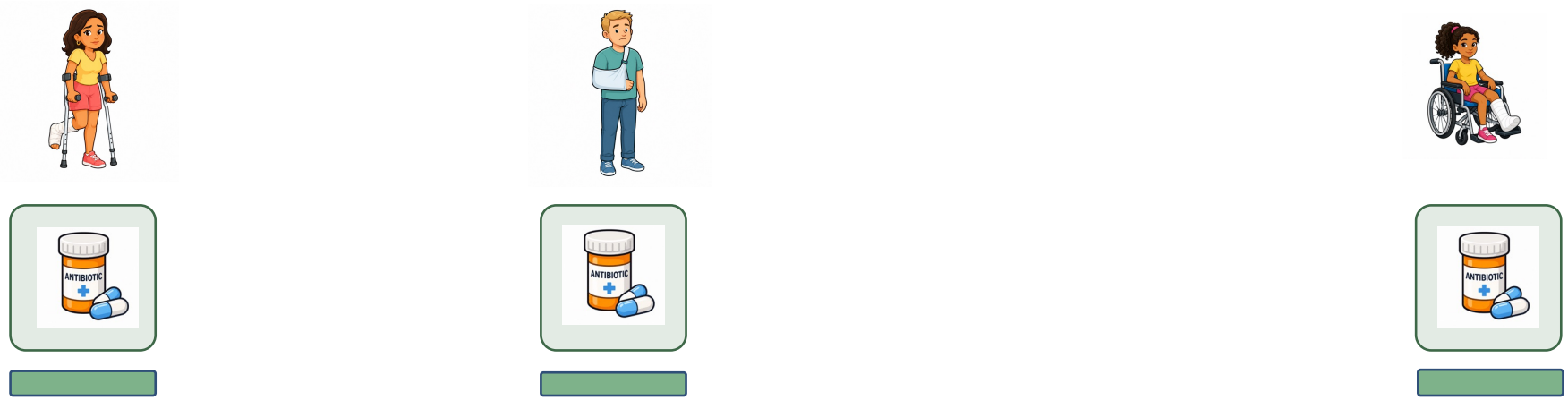
How we'll get there? **Swap regret!**

Remember Swap Regret?!



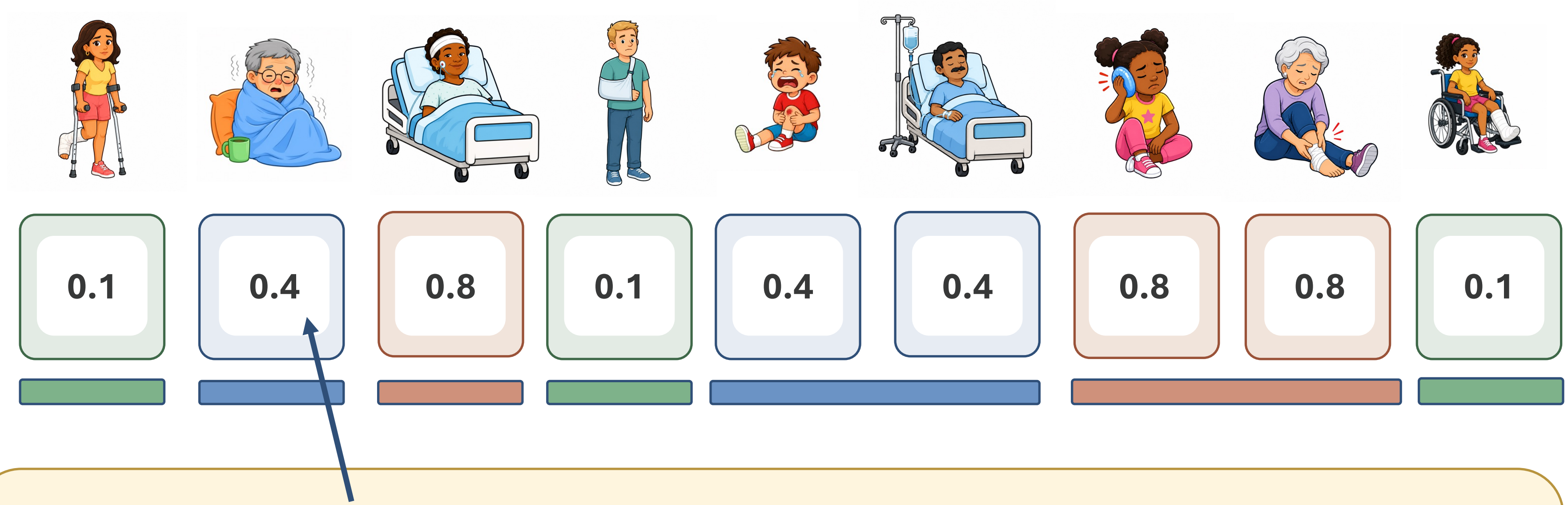
Measured how much a decision-maker can improve by swapping **all times they played one action with another action**

Remember Swap Regret?!



If a decision maker has no swap regret, then on average on the instances they prescribed antibiotics, **this was the best action**

Remember Swap Regret?!



Now we're in real-valued prediction land so what does swap regret mean?

An agent has no swap regret with **respect to class $\mathcal{H}$** if they can't improve by **swapping out** a level set of their predictions with a model $h \in \mathcal{H}$.

Remember Swap Regret?!

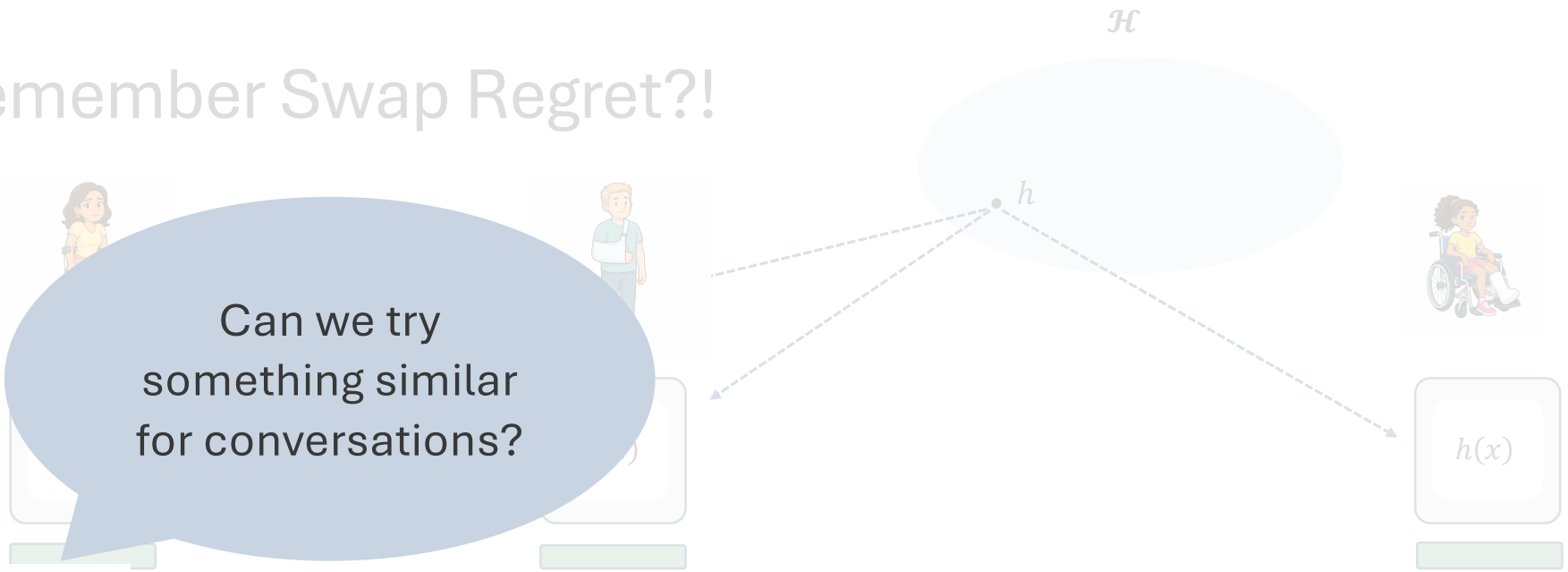
0.1

0.1

0.1

We'll take a single **level set** of our predictions, like all the times we predicted 0.1

Remember Swap Regret?!



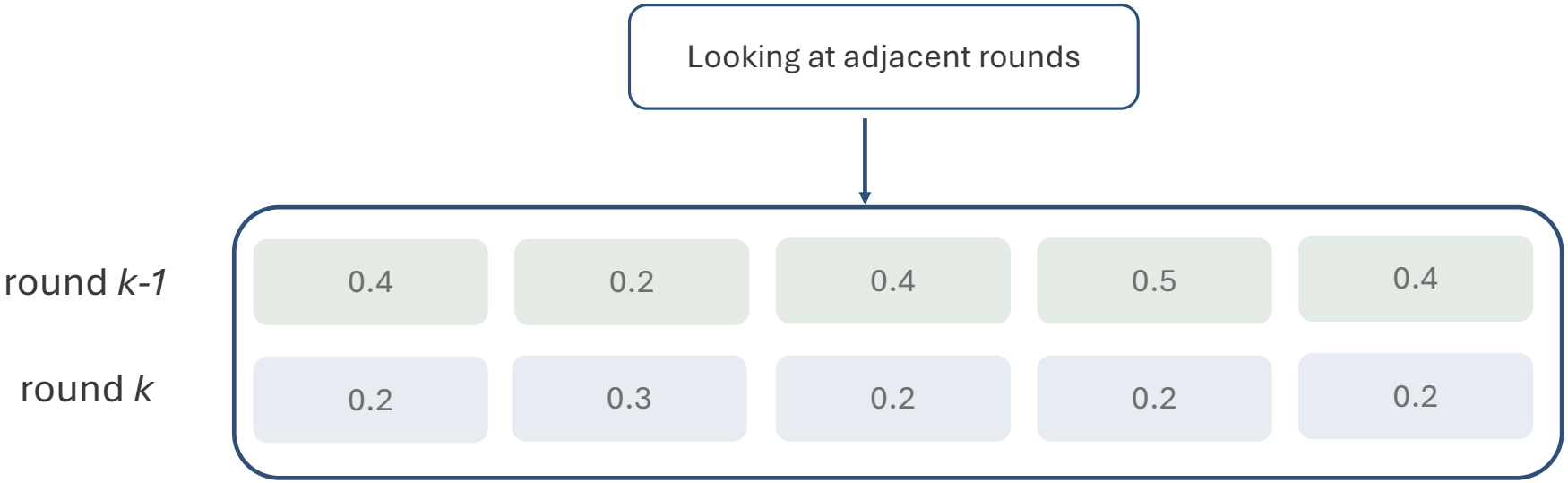
Can we try something similar for conversations?

take a single **level set** of our

and we'll check if there's any model

What if we use the same conditioning event as for conversation calibration?

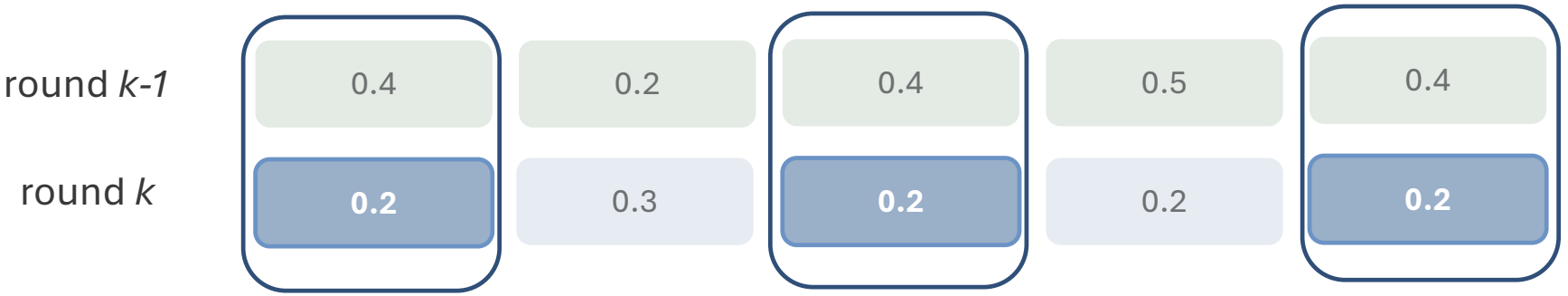
Conversation swap regret wrt $\mathcal{H}$



Can we take that same conditioning event as for conversation calibration?

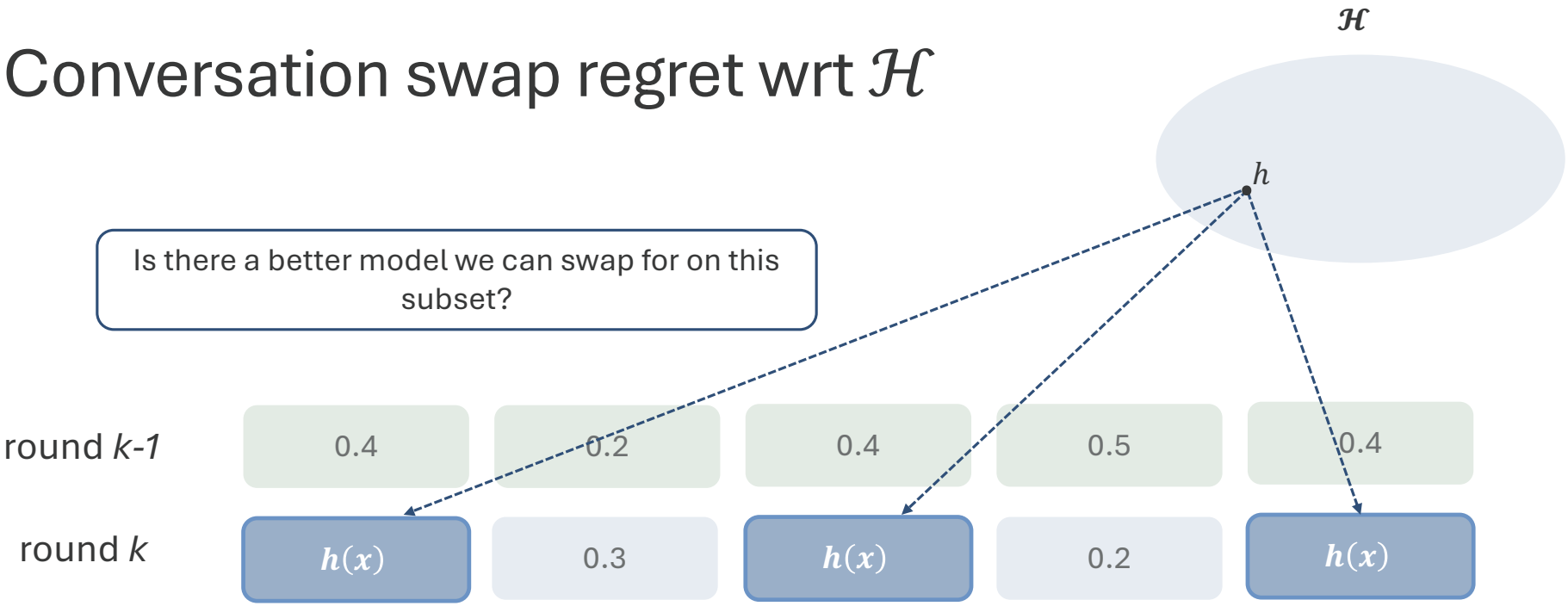
Conversation swap regret wrt $\mathcal{H}$

The days we predicted 0.2 in round *k* and 0.4 in round *k-1*



Can we take that same conditioning event as for conversation calibration?

Conversation swap regret wrt $\mathcal{H}$



Can we take that same conditioning event as for conversation calibration?

Conversation swap regret wrt $\mathcal{H}$

An agent has no conversation swap regret with respect to $\mathcal{H}$ if for all rounds k and predictions p jointly with the other agents' predictions p' at round $k-1$, **there is no $h \in \mathcal{H}$ which improves on p's squared error.**

E.g. the human has low swap regret if for all rounds k and predictions p, p',

$$\mathbb{E}[(p-y)^2 | \text{ human predicts } p \text{ at round } k, \text{model predicts } p' \text{ at round } k-1-1]$$

$$\leq \operatorname{argmin}_{h \in \mathcal{H}} \mathbb{E}[(h(x)-y)^2 | \text{ human predicts } p \text{ at round } k, \text{model predicts } p' \text{ at round } k-1]$$

Ok, cool definition,
but why were we doing
any of this again?

Remember, the goal is to have collaboration help: the agents should learn from their interaction when possible.

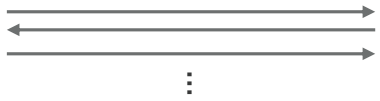
For instance….they should have

no swap regret with respect to $\mathcal{H}_m \cup \mathcal{H}_h$

… can conversation swap regret get us there? **Yes!**

Goal: Get no swap regret on $\mathcal{H}_m \cup \mathcal{H}_h$

I have no conversation swap regret wrt $\mathcal{H}_m$!



I have no conversation swap regret wrt $\mathcal{H}_h$!

If our agents have no conversation swap regret on $\mathcal{H}_m$ and $\mathcal{H}_h$ respectively and the clinician is calibrated for $\mathcal{H}_m \cup \mathcal{H}_h$,

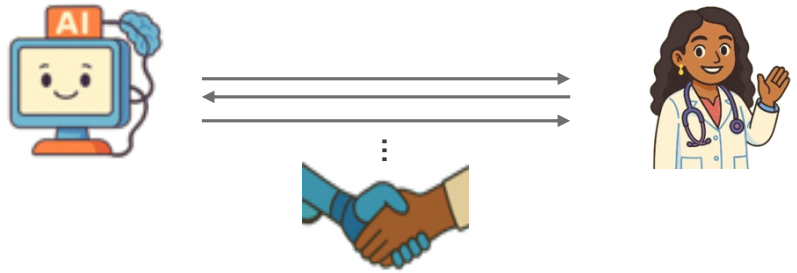
Each agent has access to their own hypothesis class, so we can make them have no conversation swap regret on those.

then our agents will be conversation calibrated.

*modulo MANY pages of details

Goal: get swap regret wrt $\mathcal{H}_m \cup \mathcal{H}_h$

I have no conversation swap regret wrt $\mathcal{H}_m$!



I have no conversation swap regret wrt $\mathcal{H}_h$!

So that means **they will agree** with each other!
So must have no swap regret on both classes!

Goal: get swap regret wrt $\mathcal{H}_m \cup \mathcal{H}_h$

This is nice, but I thought we wanted to be competitive with $\mathcal{H}_J$!!

Remember $\mathcal{H}_J$?

$\mathcal{H}_J$ **:** The magical hypothesis class in the sky which has access to everyone's features.

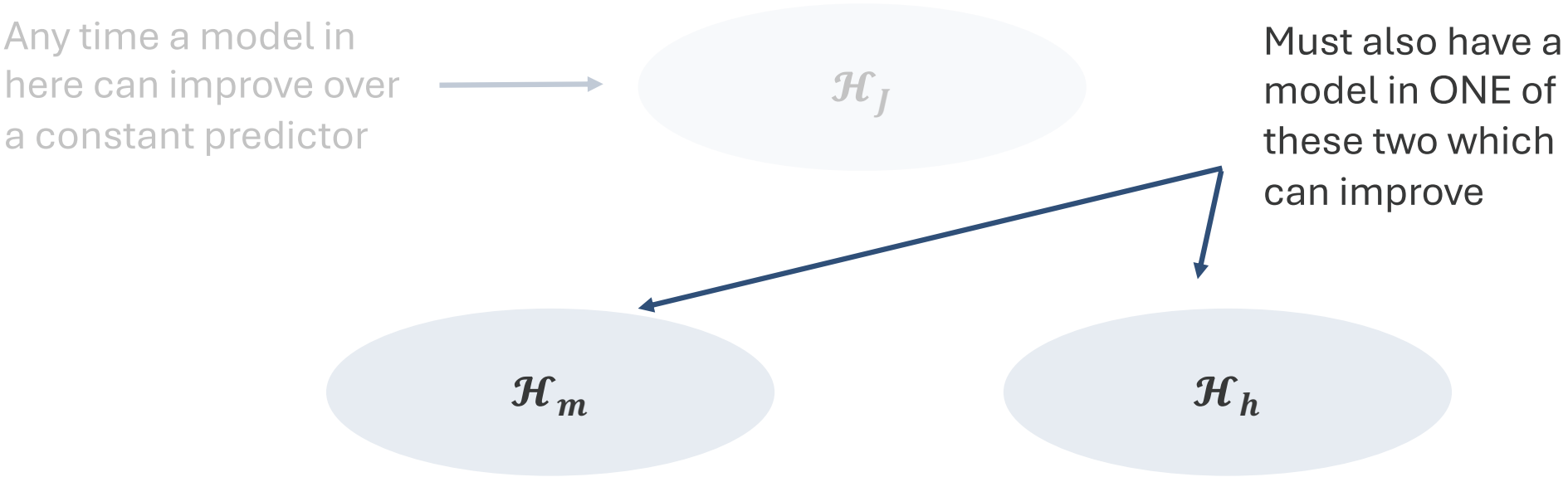
Ideal world

$\mathcal{H}_J$

$\mathcal{X}_m \cup \mathcal{X}_h \to [0,1]$

When is it possible for information aggregation to compete?

Final result: a **weak learning condition** on $\mathcal{H}_m$ and $\mathcal{H}_h$ such that **low conversation swap regret** implies you are competitive with $\mathcal{H}_J$.



Final result: a **weak learning condition** on $\mathcal{H}_m$ and $\mathcal{H}_h$ such that **low conversation swap regret** implies you are competitive with $\mathcal{H}_J$.

If your hypothesis classes satisfy this, enforcing low conversation swap regret suffices for information aggregation!

So now we know to reach agreement, and when it will make us better off!

We can agree quickly, and know when this means we agree on something good!

And we saw the power of **clever debiasing**!

Clever debiasing

Make predictions unbiased on the right subsequences

More Beyond

A glimpse at what we didn’t cover.

Other Calibration Metrics

Smooth Calibration, Leaky Forecasts, Finite Recall, and Nash Dynamics

Authors: Dean Foster, Sergiu Hart. **Publication:** Games and Economic Behavior, 2018.

Introduces smooth calibration, showing it can be achieved deterministically with leaked finite-memory forecasts and yields approximate Nash play in repeated games.

Low-Degree Multicalibration

Authors: Parikshit Gopalan, Michael Kim, Mihir Singhal, Shengjia Zhao. **Publication:** COLT 2022.

Defines low-degree multicalibration as an efficient hierarchy between multiaccuracy and full multicalibration that retains key fairness and accuracy guarantees.

How Global Calibration Strengthens Multiaccuracy

Authors: Silvia Casacuberta, Parikshit Gopalan, Varun Kanade, Omer Reingold. **Publication:** FOCS 2025.

Clarifies the hierarchy between multiaccuracy, calibrated multiaccuracy, and multicalibration by showing that global calibration makes multiaccuracy much more powerful.

A Unifying Theory of Distance from Calibration

Authors: Jaroslaw Blasiok, Parikshit Gopalan, Lunjia Hu, Preetum Nakkiran. **Publication:** STOC 2023.

Defines distance to the nearest calibrated predictor as a ground-truth calibration metric and identifies efficiently estimable calibration measures consistent with it.

Calibration Error for Decision Making

Authors: Lunjia Hu, Yifan Wu. **Publication:** FOCS 2024.

Introduces Calibration Decision Loss as the worst-case downstream decision-payoff improvement from recalibrating forecasts and gives an efficient online algorithm with near-optimal expected error.

An Elementary Predictor Obtaining 2 sqrt(T) + 1 Distance to Calibration

Authors: Eshwar Ram Arunachaleswaran, Natalie Collina, Aaron Roth, Mirah Shi. **Publication:** SODA 2025.

Gives a simple deterministic online predictor achieving distance to calibration at most 2 sqrt(T) + 1 in the adversarial setting.

Truthful Calibration Metrics

Truthfulness of Calibration Measures

Authors: Nika Haghtalab, Mingda Qiao, Kunhe Yang, Eric Zhao. **Publication:** NeurIPS 2024.

Shows that many existing calibration measures are not truthful and introduces subsampled smooth calibration error as a more incentive-compatible alternative.

Truthfulness of Decision-Theoretic Calibration Measures

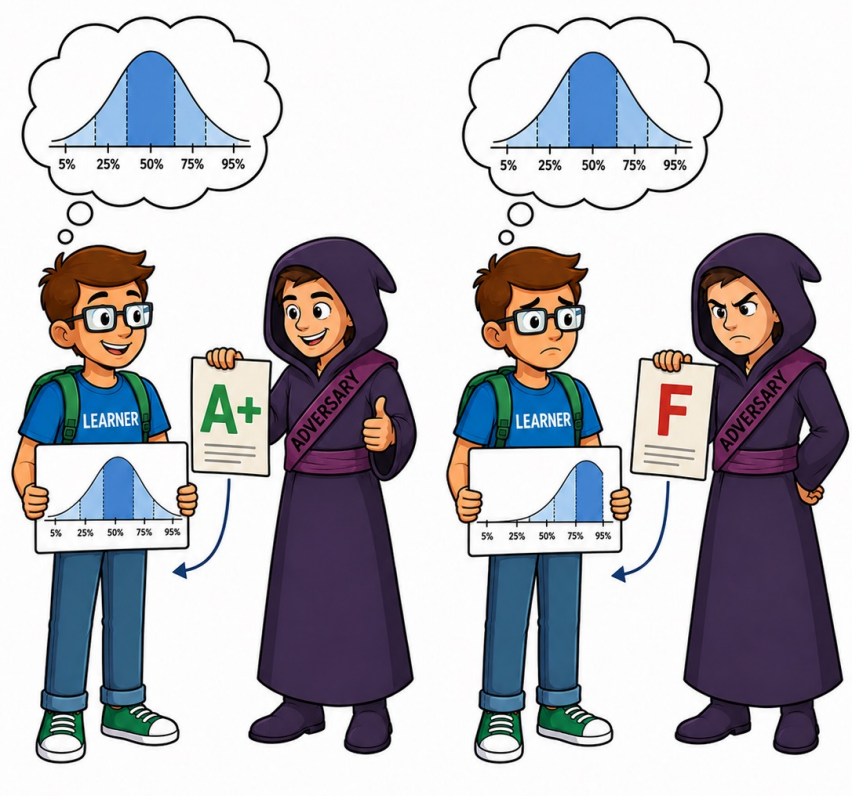
Authors: Mingda Qiao, Eric Zhao. **Publication:** COLT 2025.

Introduces a decision-theoretic calibration measure that is both truthful and useful for downstream no-regret decision-making, while proving limits for such measures.

A Perfectly Truthful Calibration Measure

Authors: Jason Hartline, Lunjia Hu, Yifan Wu. **Publication:** arXiv, 2025.

Introduces averaged two-bin calibration error, a perfectly truthful batch calibration measure that is sound, complete, continuous, and efficiently computable.



High Dimensional Prediction

High-Dimensional Prediction for Sequential Decision Making

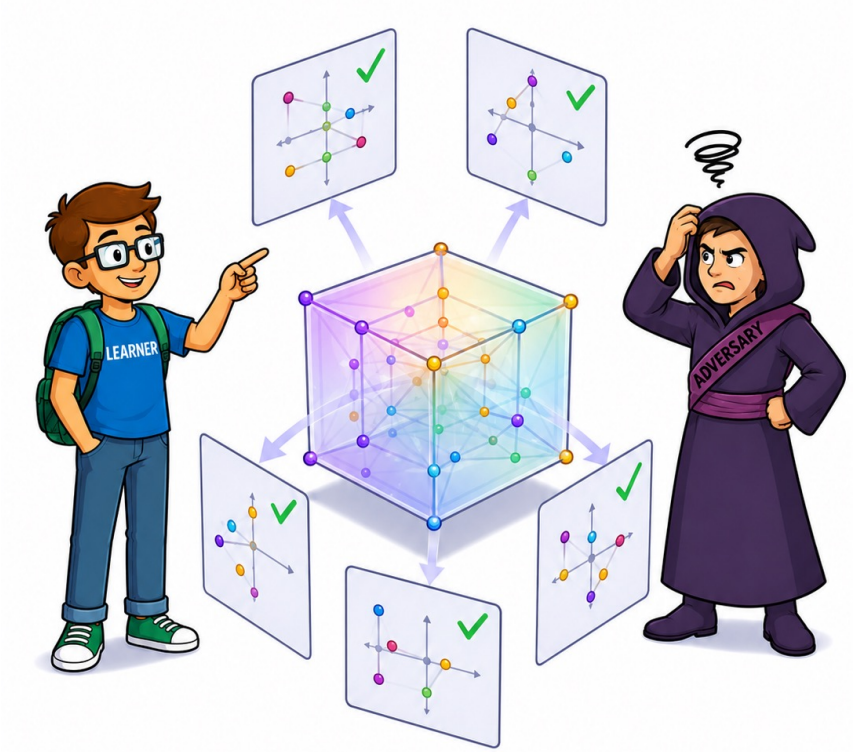
Authors: Georgy Noarov, Ramya Ramalingam, Aaron Roth, Stephan Xie. **Publication:** ICML 2025.

Gives efficient high-dimensional event-unbiased forecasts that support downstream decision-makers, conditional regret guarantees, and online multicalibration rates.

An Efficient Black-Box Reduction from Online Learning to Multicalibration, and a New Route to Phi-Regret Minimization

Authors: Gabriele Farina, Juan Carlos Perdomo. **Publication:** arXiv, 2026.

Gives a GGM-style black-box reduction showing that online multicalibration can be achieved by combining a no-regret learner over test functions with an expected variational inequality solver, yielding oracle-efficient sqrt(T)-type guarantees.



Calibration Beyond the Worst Case

Instance-Adaptive Online Multicalibration

Authors: Zhiming Huang, Jamie Morgenstern, Aaron Roth, Claire Jie Zhang. **Publication:** arXiv, 2026.

Gives an efficient online multicalibration algorithm that recovers worst-case-optimal rates while adapting to easier stochastic and piecewise-stationary instances.

Adaptive Calibration in Non-Stationary Environments

Authors: Junyan Liu, Haipeng Luo, Lillian J. Ratliff. **Publication:** arXiv, 2026.

Develops online calibration algorithms whose guarantees adapt to the degree of non-stationarity, interpolating between stationary, piecewise-stationary, and adversarial regimes under multiple calibration measures.

Multicalibration As Robustness to Distribution Shift

Universal Adaptability: Target-Independent Inference that Competes with Propensity Scoring

Authors: Michael P. Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, Omer Reingold. **Publication:** PNAS, 2022.

Uses multicalibration to build a single source-population estimator whose inferences remain valid across many downstream target populations.

Multi-CATE: Multi-Accurate Conditional Average Treatment Effect Estimation Robust to Unknown Covariate Shifts

Authors: Christoph Kern, Michael P. Kim, Angela Zhou. **Publication:** arXiv, 2024.

Uses multiaccurate post-processing to make conditional average treatment-effect estimates robust to unknown covariate shifts at deployment time.

Bridging Multicalibration and Out-of-Distribution Generalization Beyond Covariate Shift

Authors: Jiayun Wu, Jiashuo Liu, Peng Cui, Zhiwei Steven Wu. **Publication:** NeurIPS 2024.

Extends multicalibration to label-dependent grouping functions, linking it to invariance and robust out-of-distribution generalization beyond covariate shift.

Boosting and Hard Core Sets

Multicalibration as Boosting for Regression

Authors: Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, Jessica Sorrell. **Publication:** ICML 2023.

Characterizes multicalibration through a squared-error swap-regret condition and gives a regression-oracle boosting algorithm with agnostic guarantees.

From Pseudorandomness to Multi-Group Fairness and Back

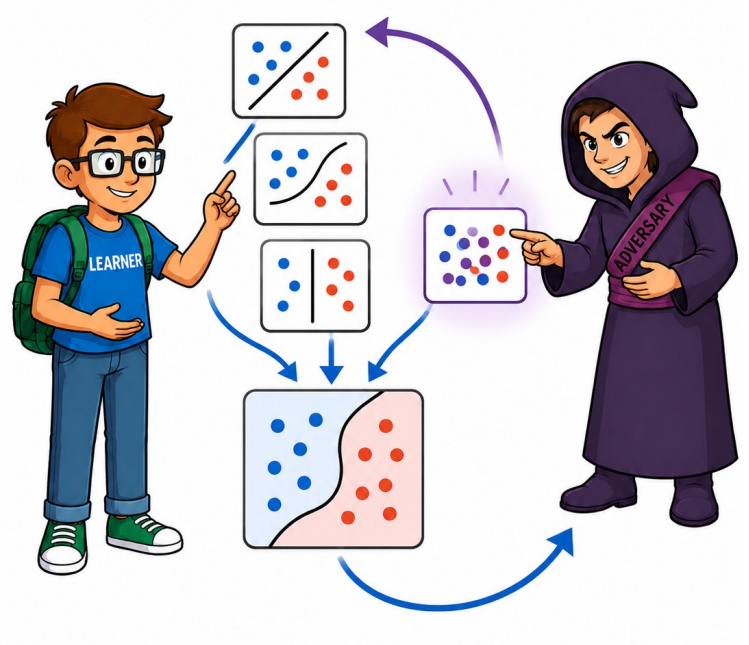
Authors: Cynthia Dwork, Daniel Lee, Huijia Lin, Pranay Tankala. **Publication:** COLT 2023.

Connects multi-group fairness to pseudorandomness through statistical-distance variants of multicalibration, yielding algorithms and a real-valued hardcore lemma.

Complexity-Theoretic Implications of Multicalibration

Authors: Silvia Casacuberta, Cynthia Dwork, Salil Vadhan. **Publication:** STOC 2024.

Uses multicalibration to strengthen regularity-lemma applications including hardcore lemmas, dense-model theorems, and conditional pseudo-min-entropy equivalences.



Calibration Beyond Means and Conformal Prediction

Moment Multicalibration for Uncertainty Estimation

Authors: Christopher Jung, Changhwa Lee, Mallesh M. Pai, Aaron Roth, Rakesh Vohra. **Publication:** COLT 2021.

Extends multicalibration from means to higher moments, enabling calibrated uncertainty estimates such as variances.

Practical Adversarial Multivalid Conformal Prediction

Authors: Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, Aaron Roth. **Publication:** NeurIPS 2022.

Gives a lightweight sequential conformal method with threshold-conditional and subgroup-conditional empirical coverage guarantees without a held-out calibration set.

Batch Multivalid Conformal Prediction

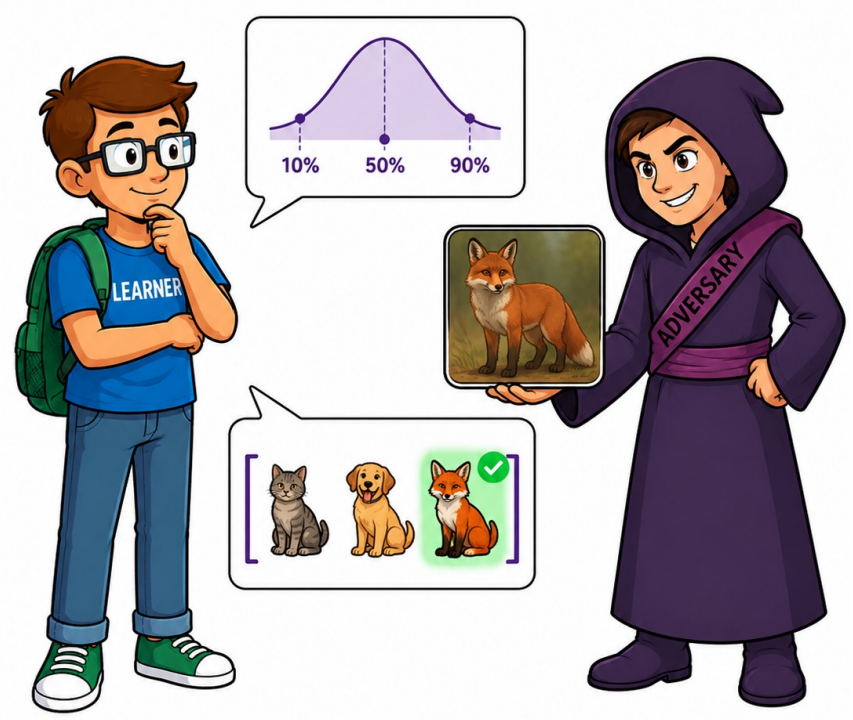
Authors: Christopher Jung, Georgy Noarov, Ramya Ramalingam, Aaron Roth. **Publication:** ICLR 2023.

Develops batch conformal algorithms whose coverage guarantees hold simultaneously conditional on group membership and nonconformity threshold.

The Statistical Scope of Multicalibration

Authors: Georgy Noarov, Aaron Roth. **Publication:** ICML 2023.

Characterizes which continuous scalar distributional properties can be multicalibrated by showing that multicalibration is possible exactly for elicitable properties.



Thank you --- and Questions Please!

For an annotated bibliography, see:
https://calibration-tutorial.github.io/