AI in Clinical Care

Stanford EMED Faculty Development

Erkin Ötleş
April 2024

Hello, World!

Medical Scientist Training Program Fellow

MD: x2024, Engineering PhD: 2022

ML Dev & Implementation Lead

Previously:

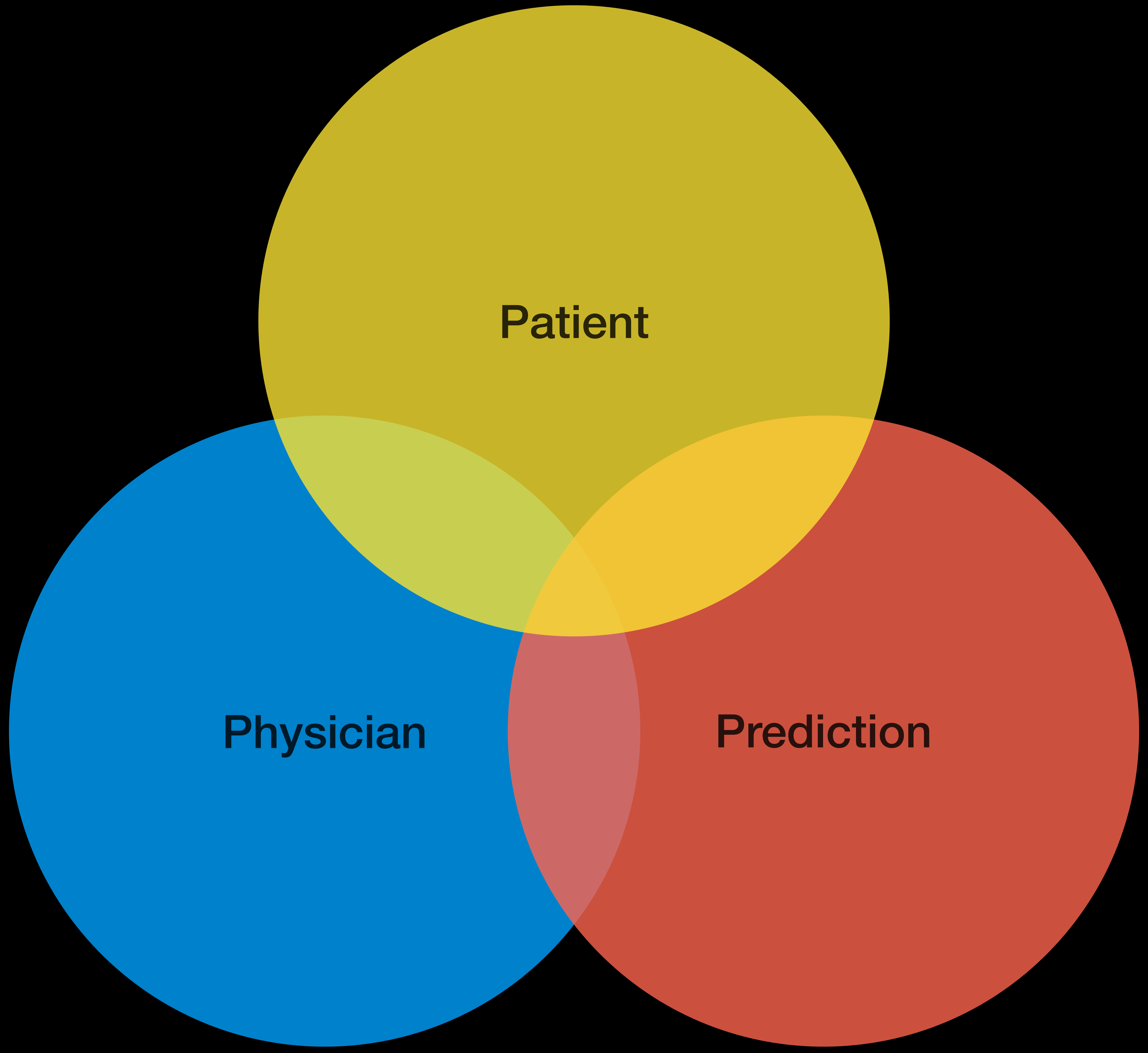
Healthcare Data Science Manager

Epic Engineer

Hello, World!

Dissertation: *Machine Learning for Healthcare: Model Development and Implementation in Longitudinal Settings*

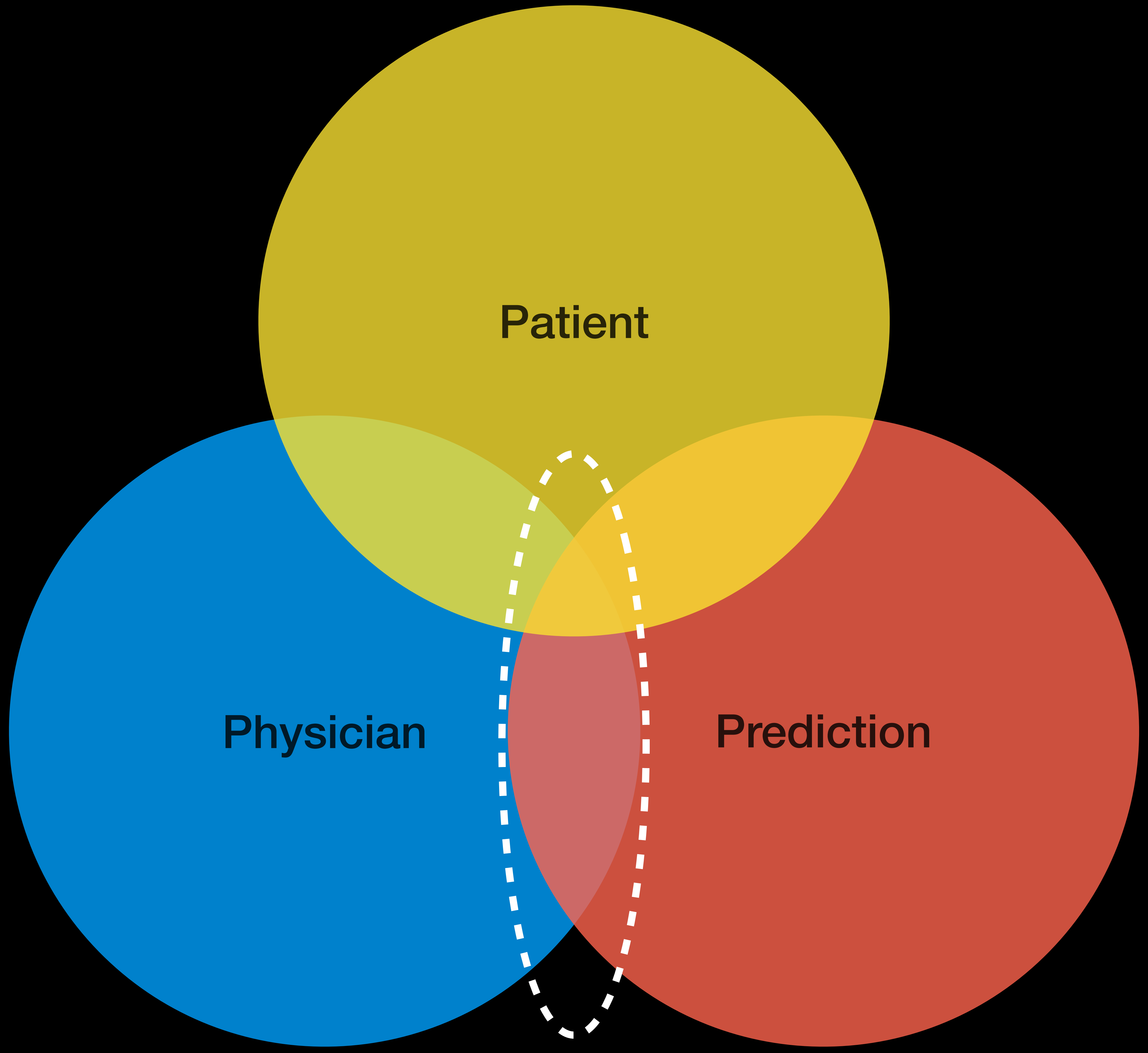
Co-advised: Jenna Wiens (CS) & Brian Denton (IOE)



Hello, World!

Dissertation: *Machine Learning for Healthcare: Model Development and Implementation in Longitudinal Settings*

Interested in computational approaches to make AI tools more useful for physicians and patients.



Potential Conflicts of Interest

Advise startups: clinical summarization using LLMs

Patent pending: AI prediction of health outcomes in patients with occupational injuries.

Small amount of stock in various technology & healthcare companies.

Agenda

Background

- Definitions
- Connections between generative & predictive AI

Clinical AI

- Repurposing predictive AI to do generative tasks
- Constant evaluation is fundamental

Upcoming Generative Clinical AI

- AI Scribes & Summarization
- Medical foundation models

Discussion

Background

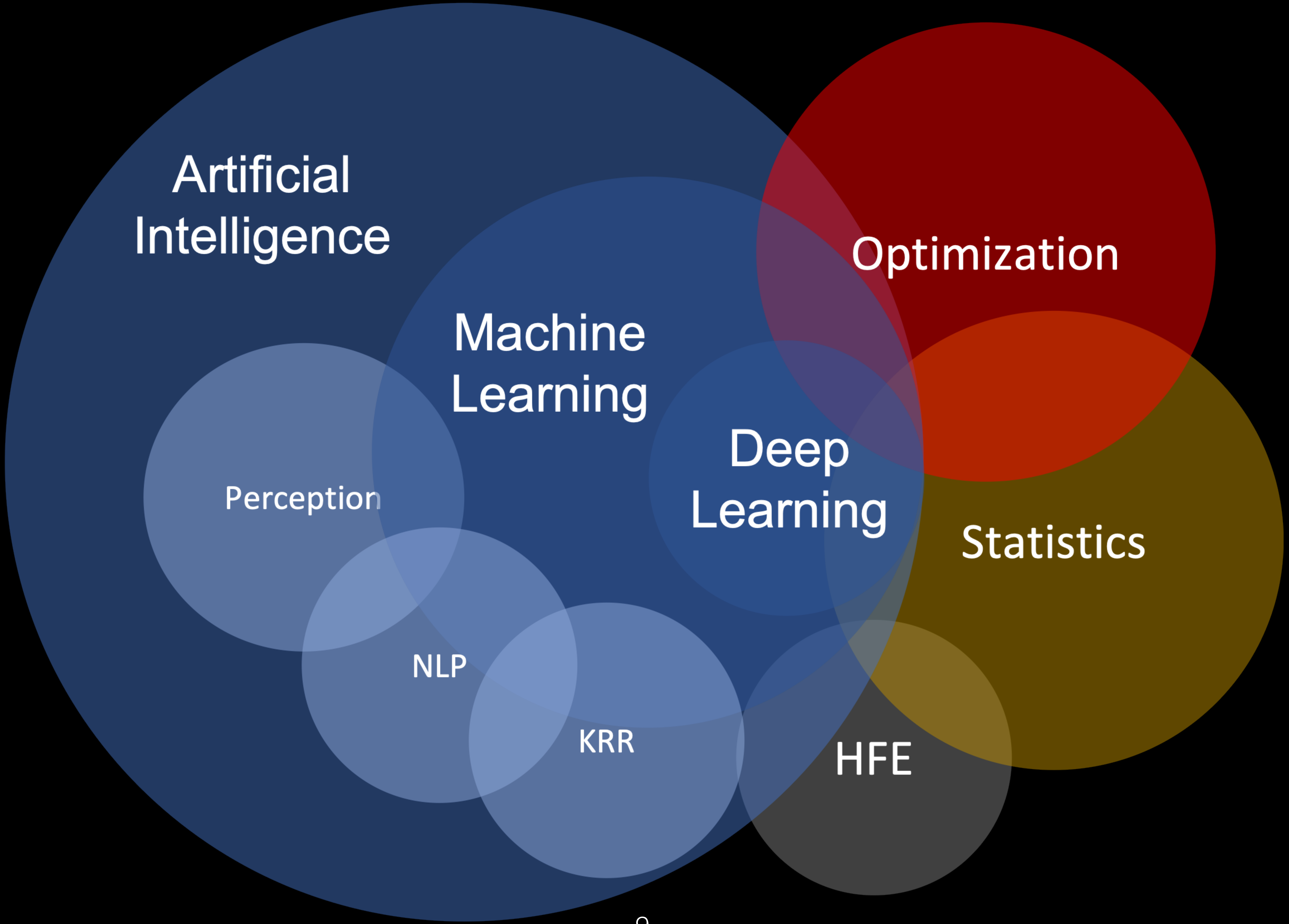
First, some definitions

Artificial Intelligence (AI): *intelligence* (perceiving, synthesizing, and inferring information) demonstrated by machines

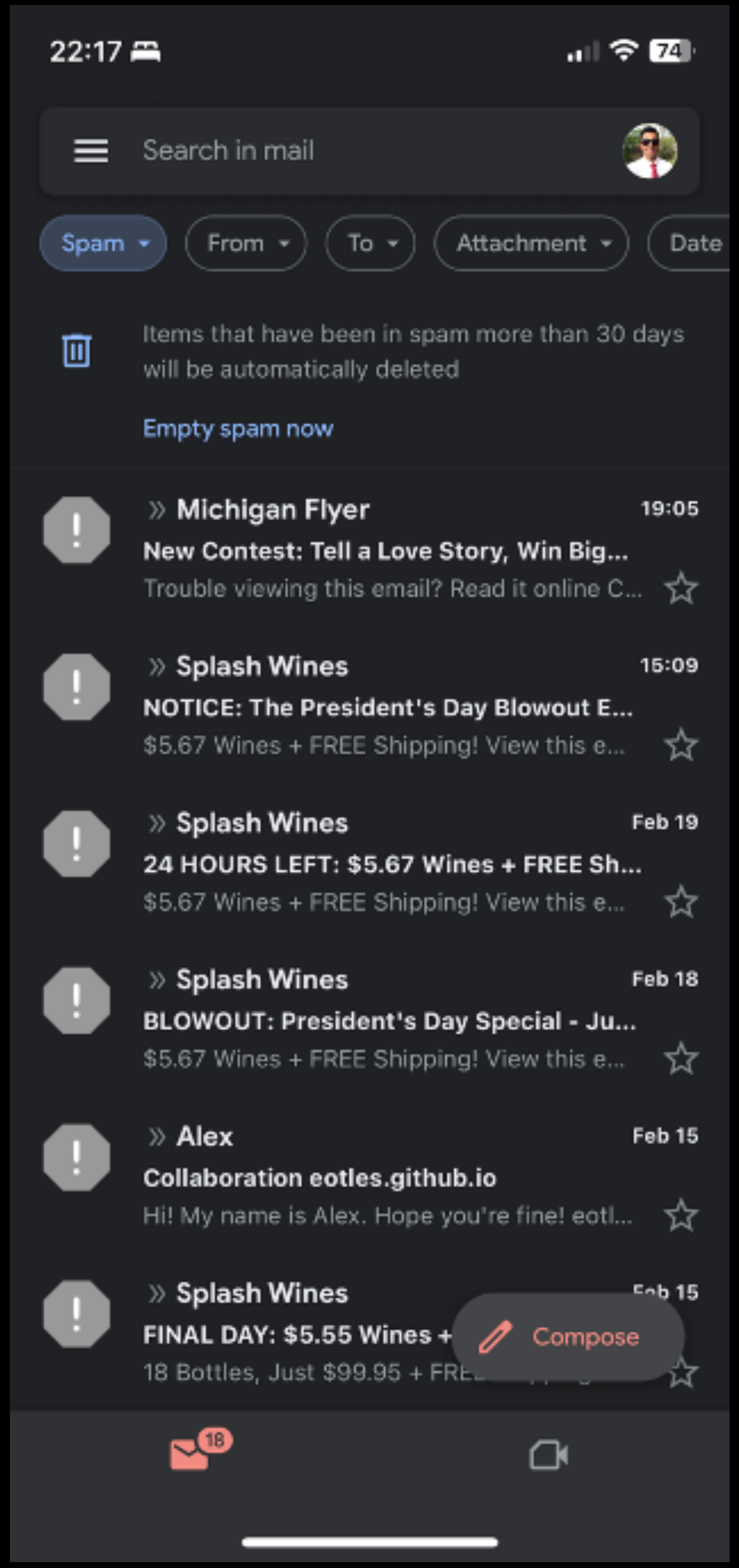
Machine Learning (ML): field of inquiry devoted to understanding and building methods that *learn* (use data to improve performance on a task).

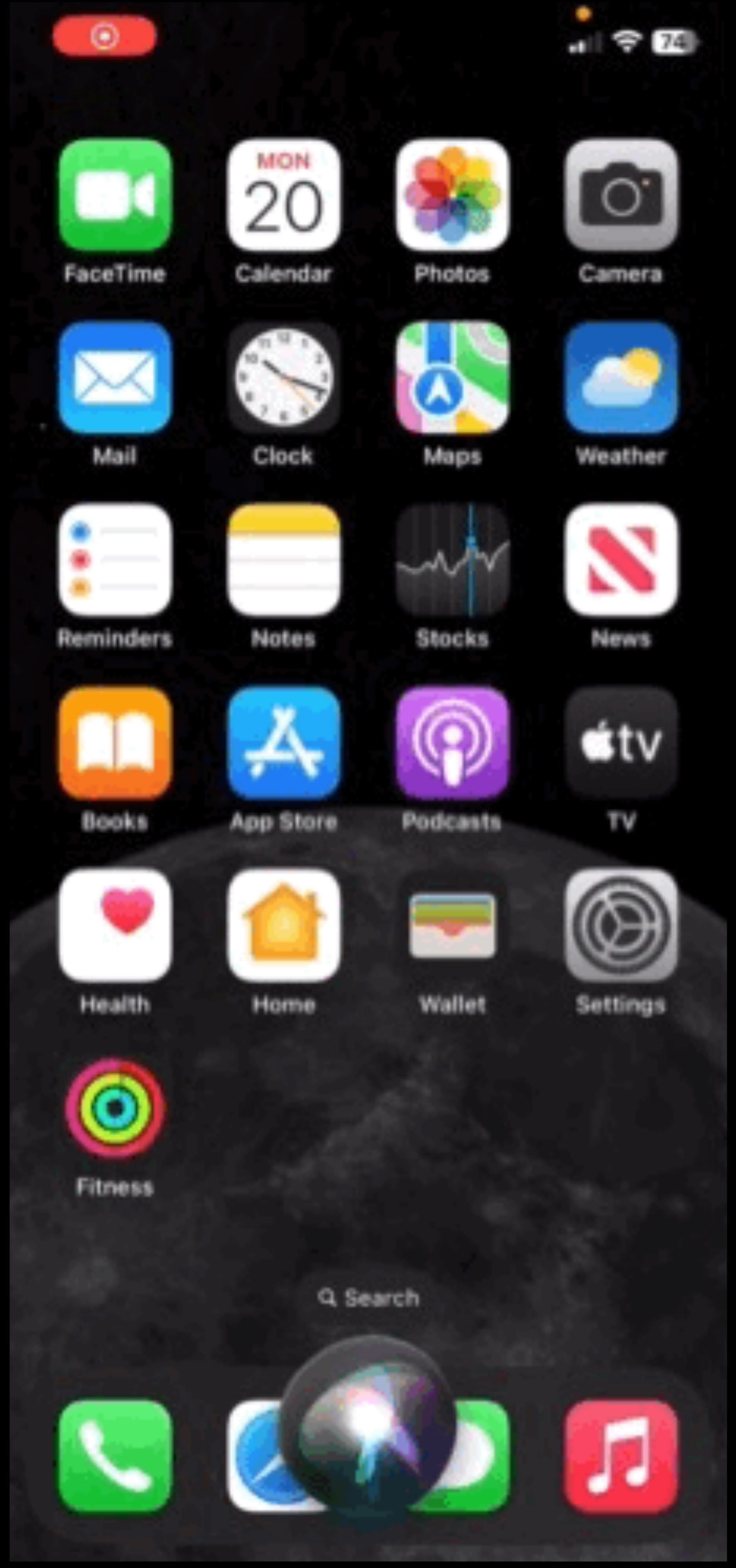
Adapted from wikipedia

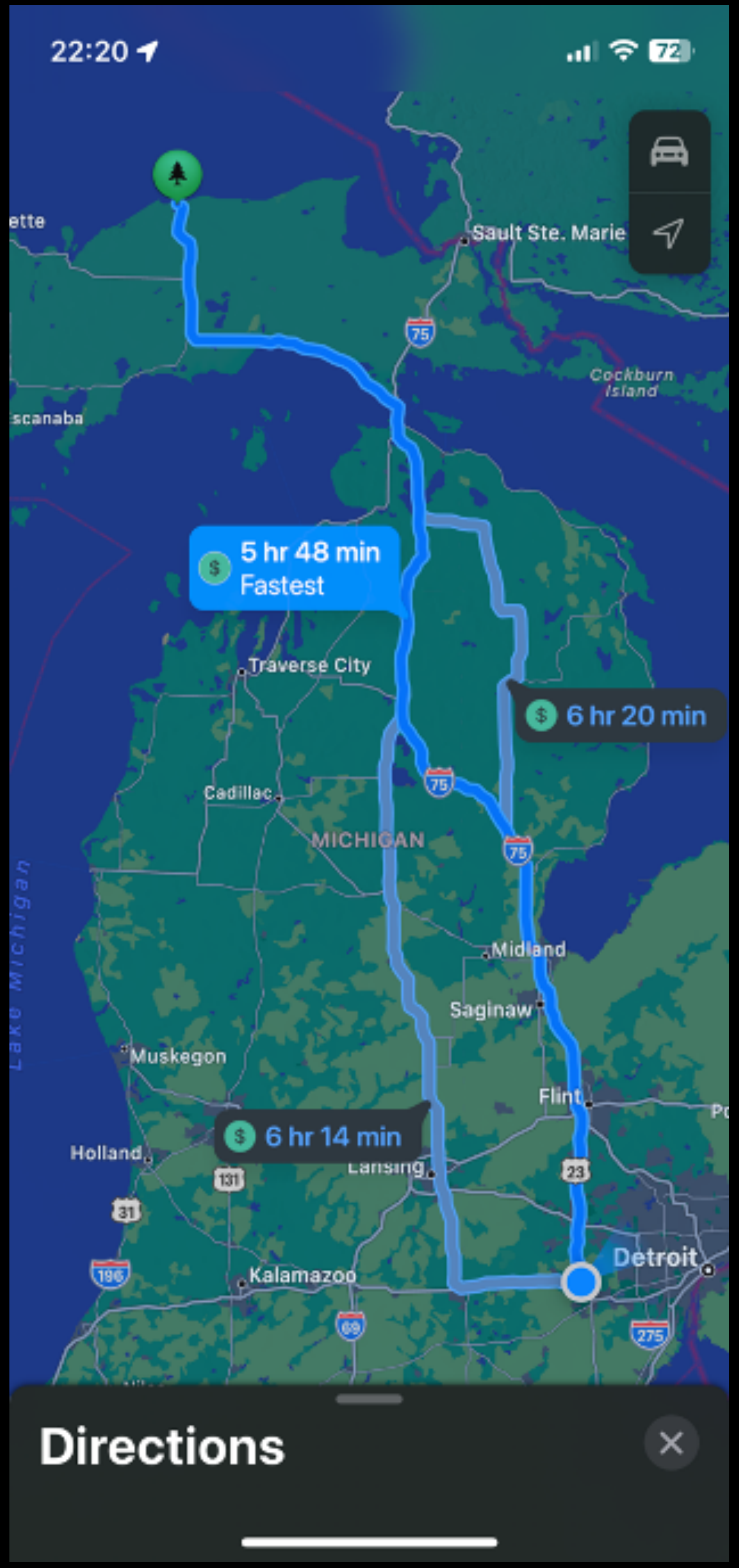
Nesting and overlapping concepts

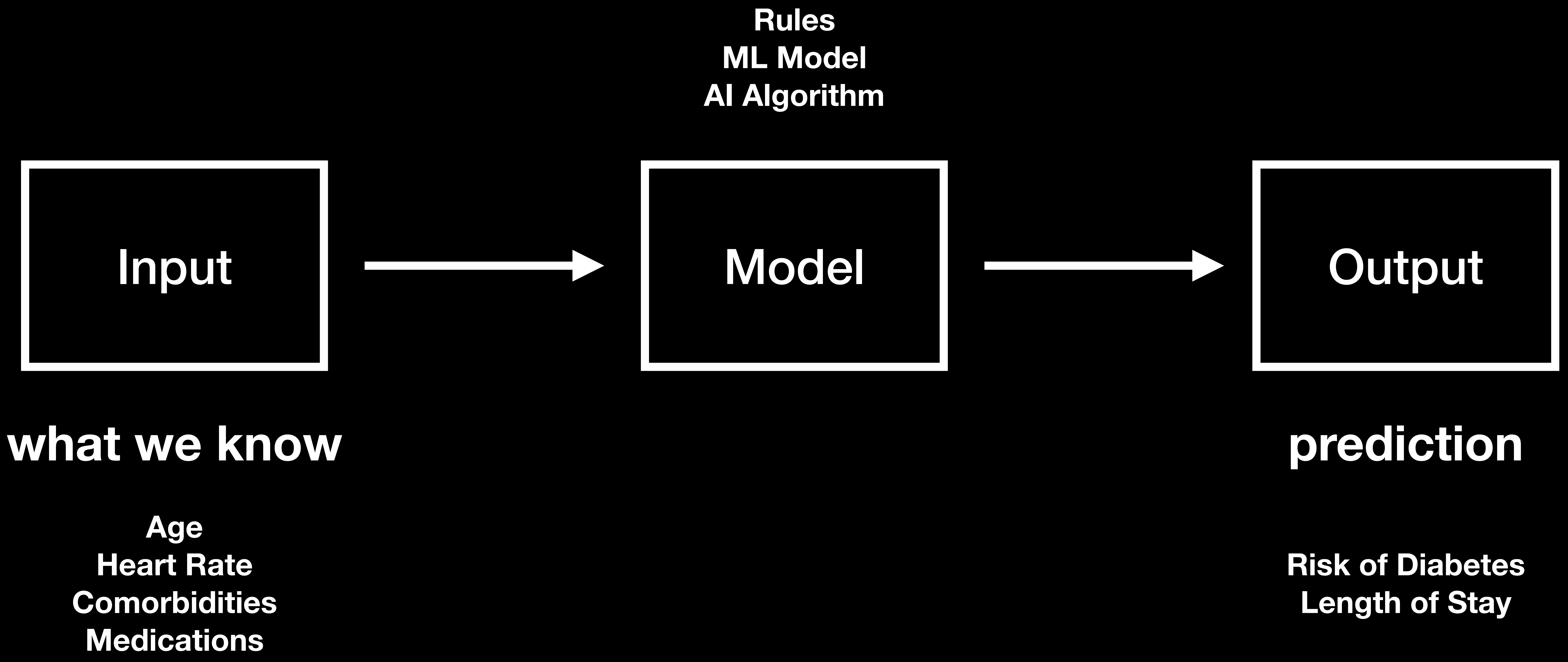


AI is ubiquitous in everyday life

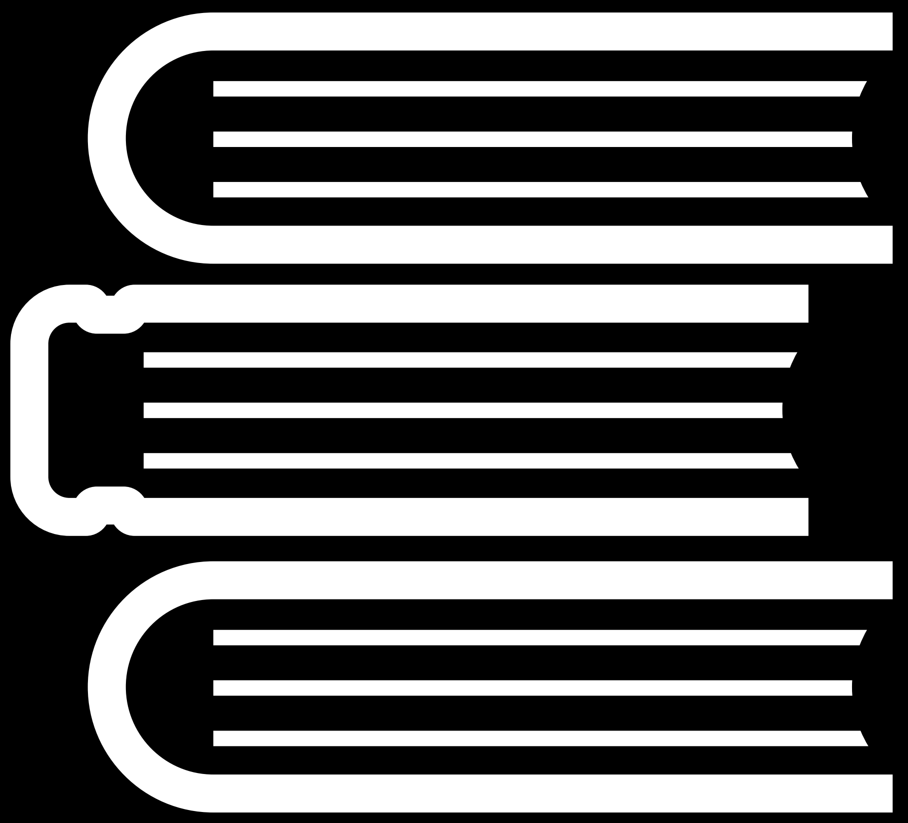


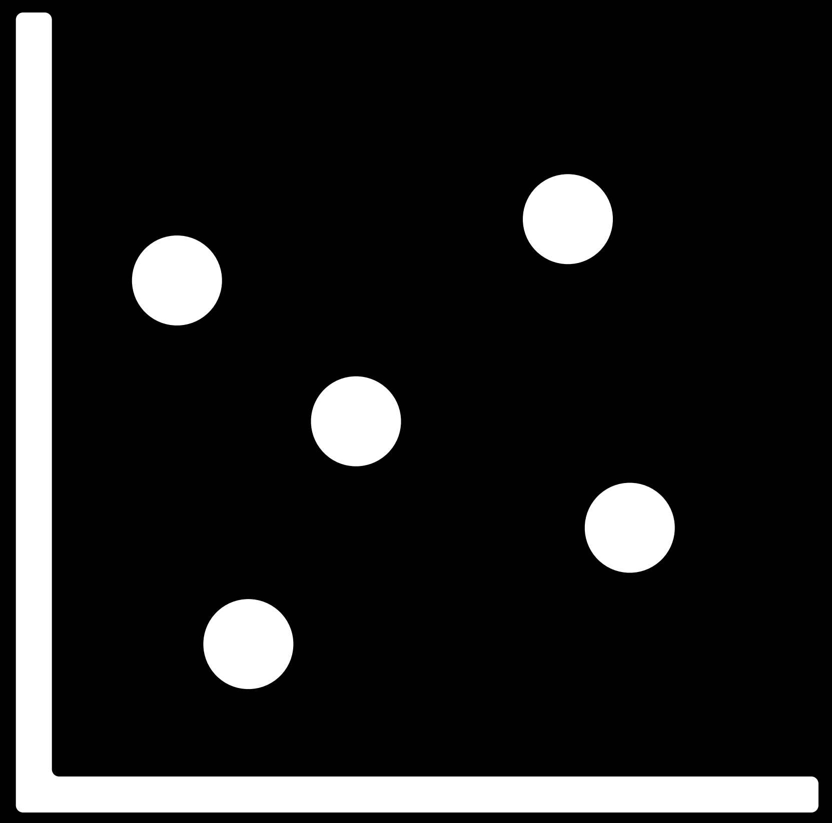





Rules
ML Model
AI Algorithm
Input
Model
Output
what we know
Age
Heart Rate
Comorbidities
Medications
prediction
Risk of Diabetes
Length of Stay

AI has the potential to advance medicine



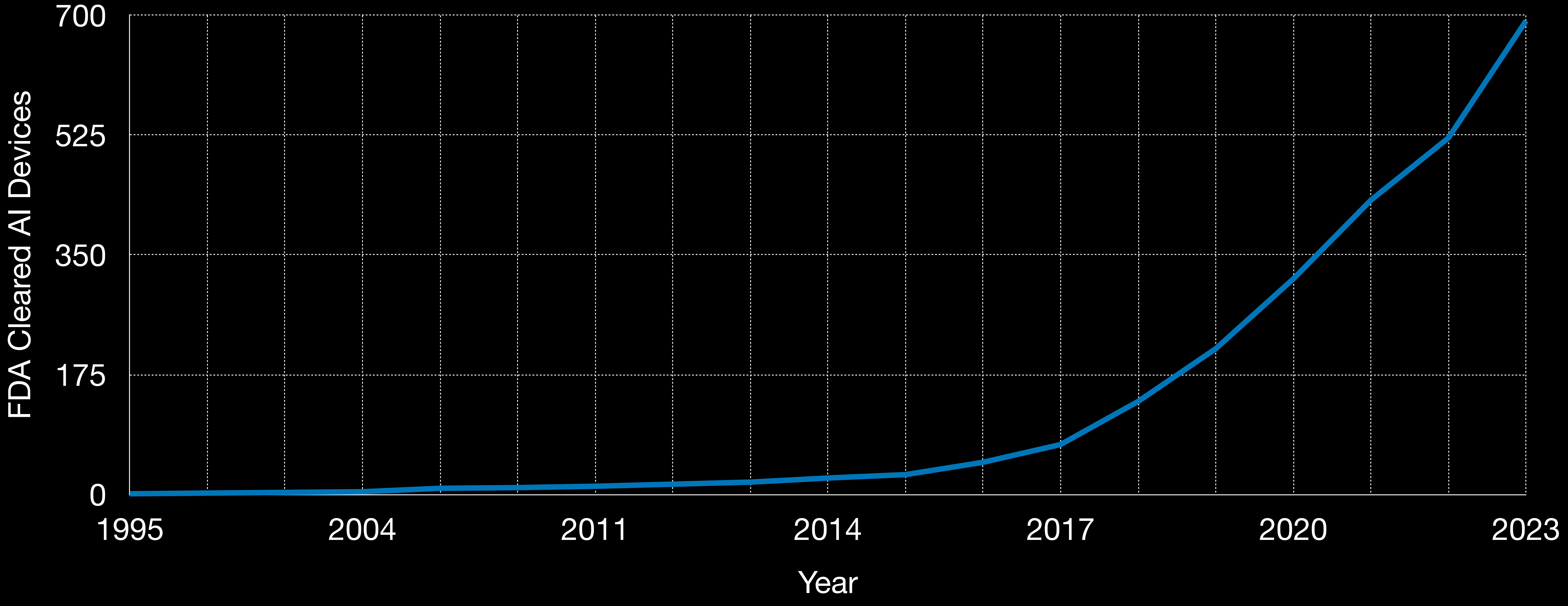


AI has techniques to rapidly **summarize** information, **predict** outcomes, and **learn** over time

Society has big expectations for AI in medicine

Increasing prevalence of medical AI

no FDA cleared generative AI tools as of 2023



Generative AI definitions

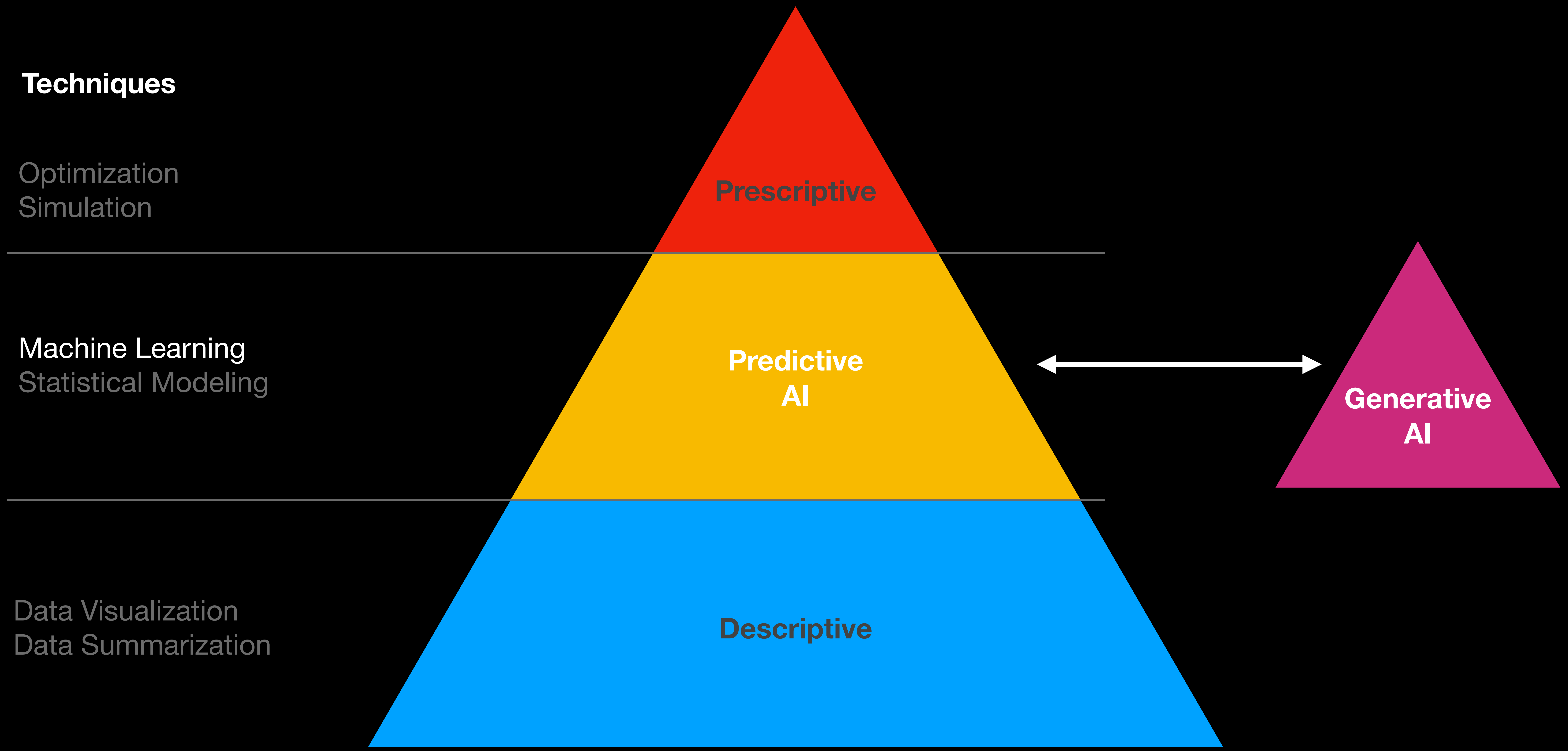
Generative AI: AI capable of *generating data* (text, images, etc.) using generative models, often in response to prompts.

Large Language Model (LLM): language model able to capable of *general-purpose language generation* and other language tasks.

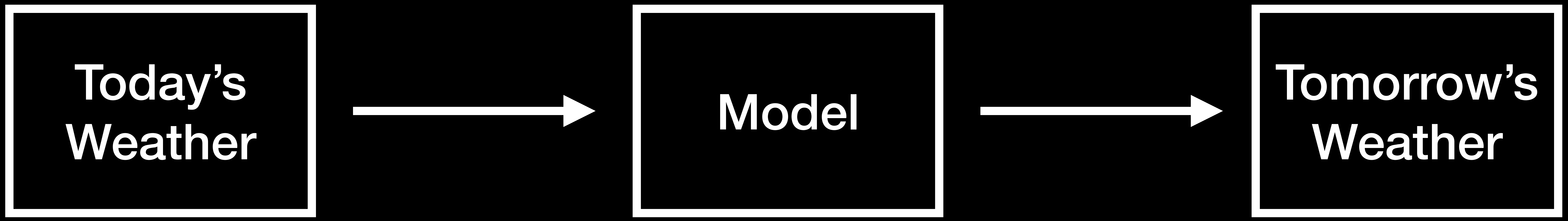
Foundation Model: a model that is trained on broad data such that it can be applied across a *wide range of use cases*.

Adapted from wikipedia

Connection between AI types



Will it rain tomorrow?



what we know

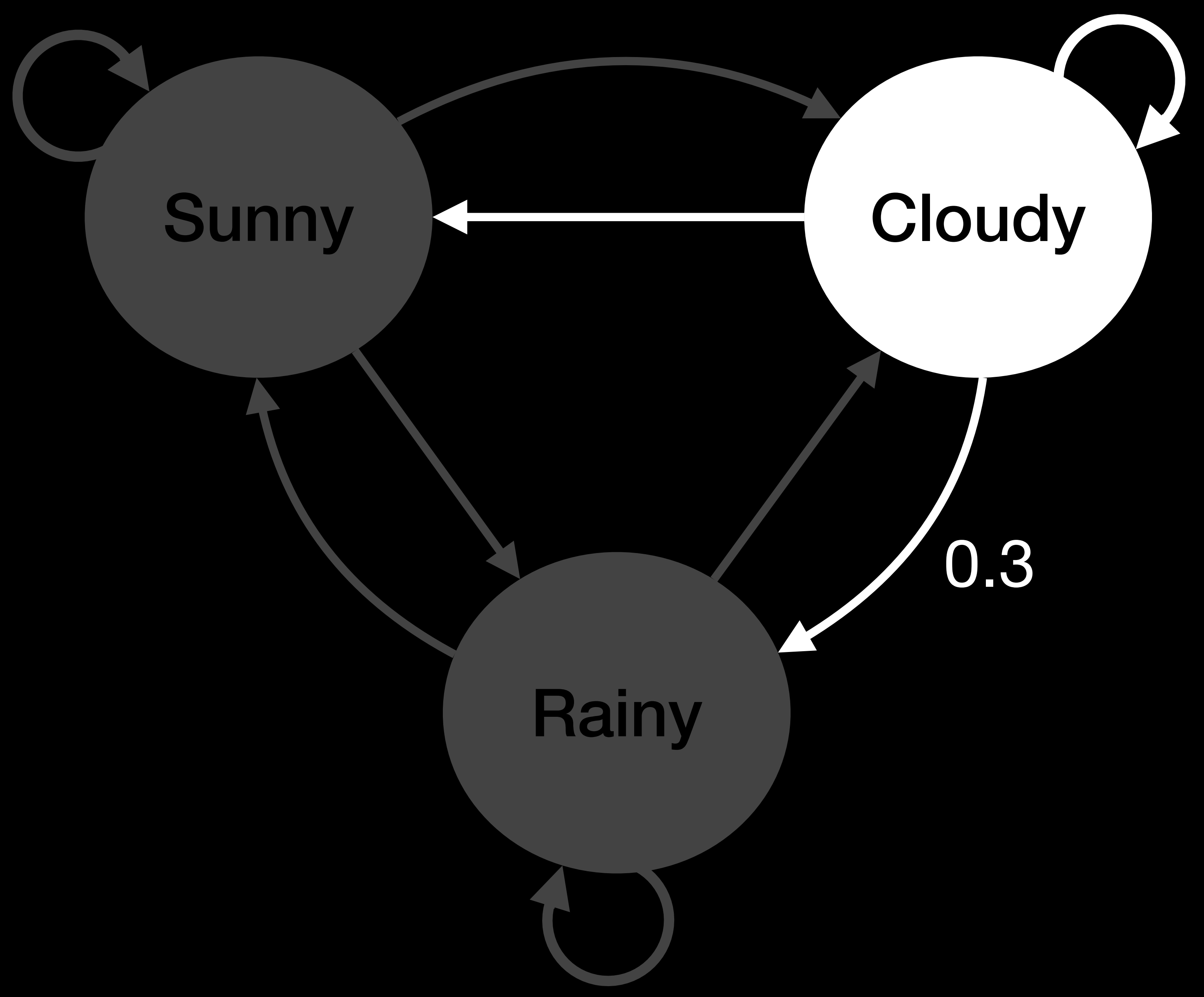
prediction

It is cloudy today.

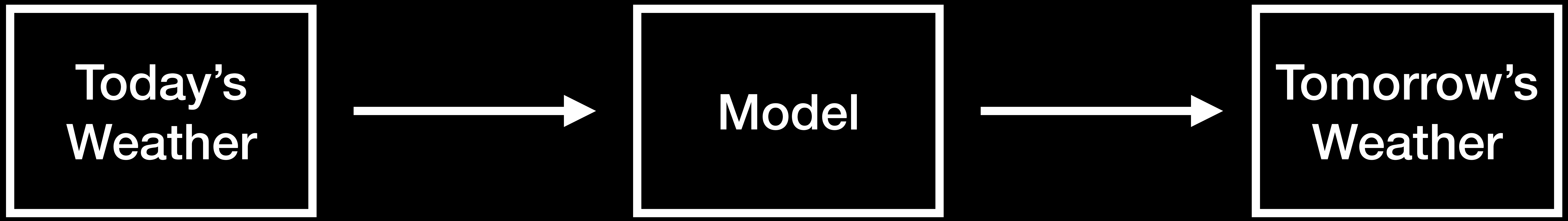
**What's the probability
it will rain tomorrow?**

Predictive

Predictive use of a weather model



What’s the weather next week look like?

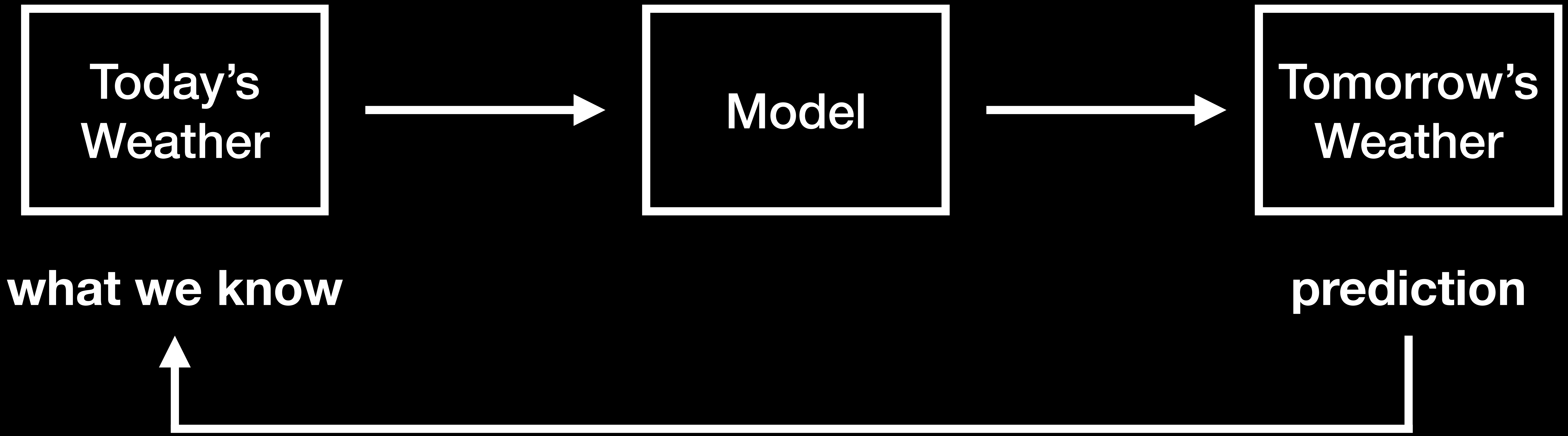


what we know

prediction

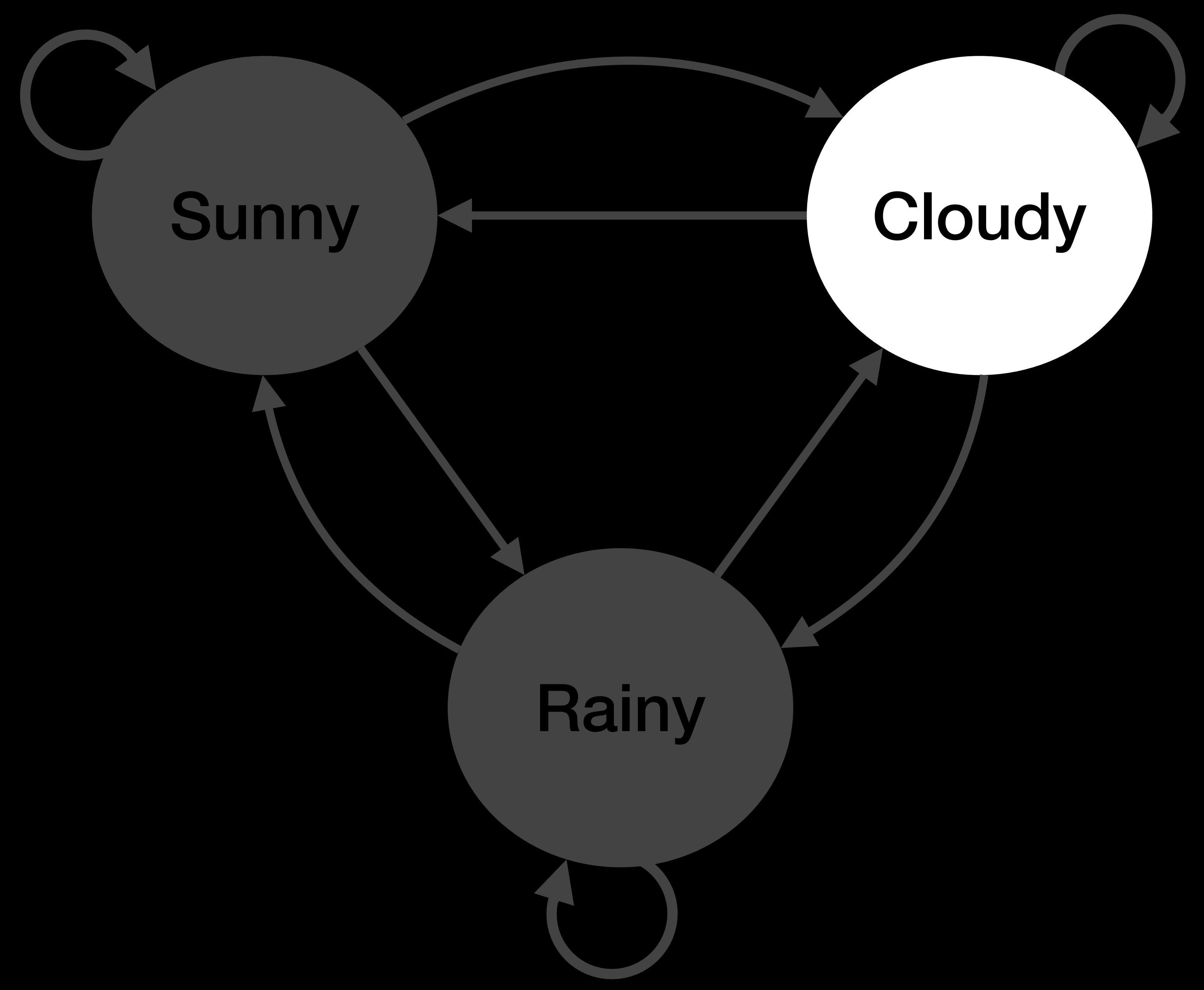
Generative

What’s the weather next week look like?



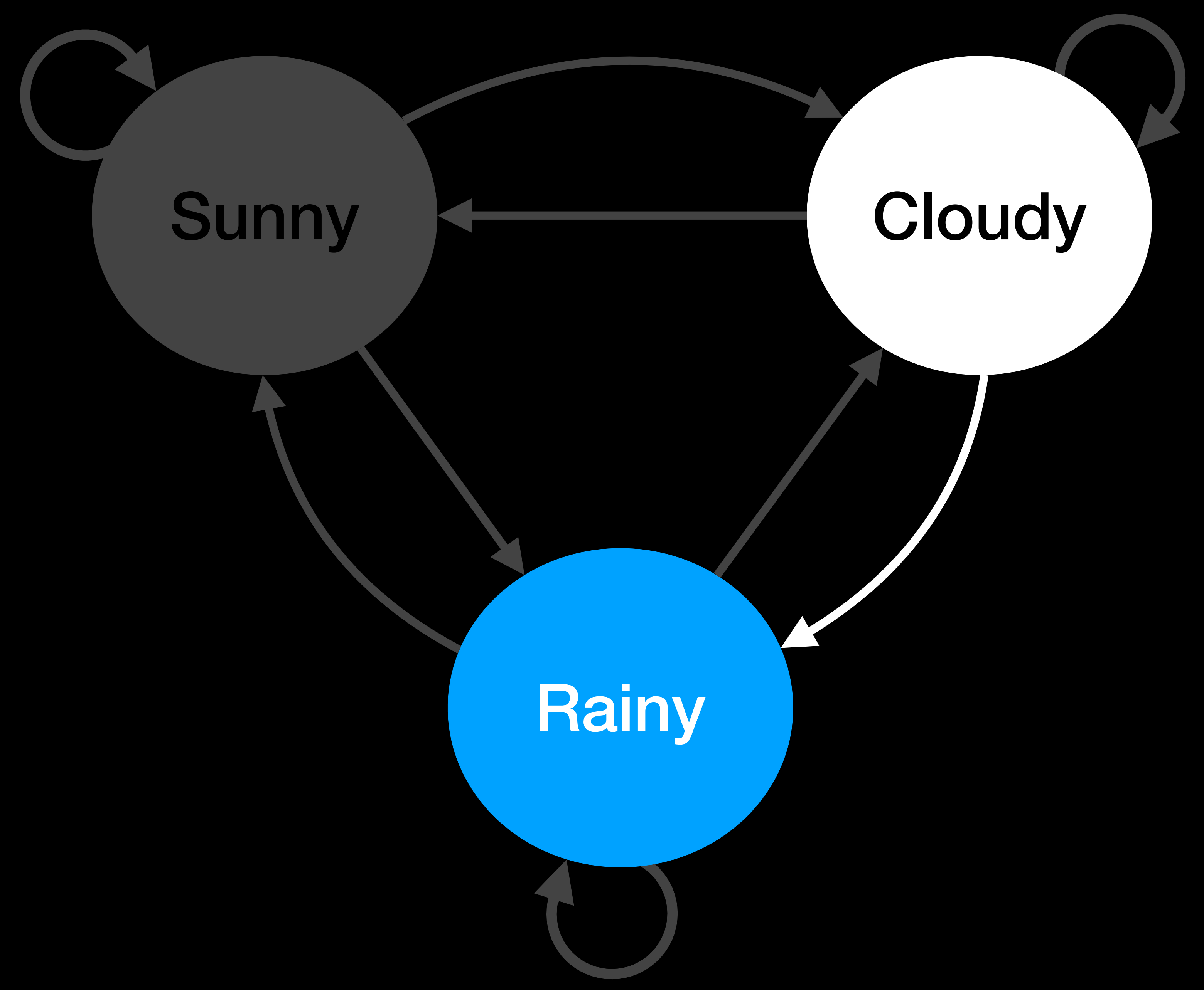
Generative

Generative use of a weather model



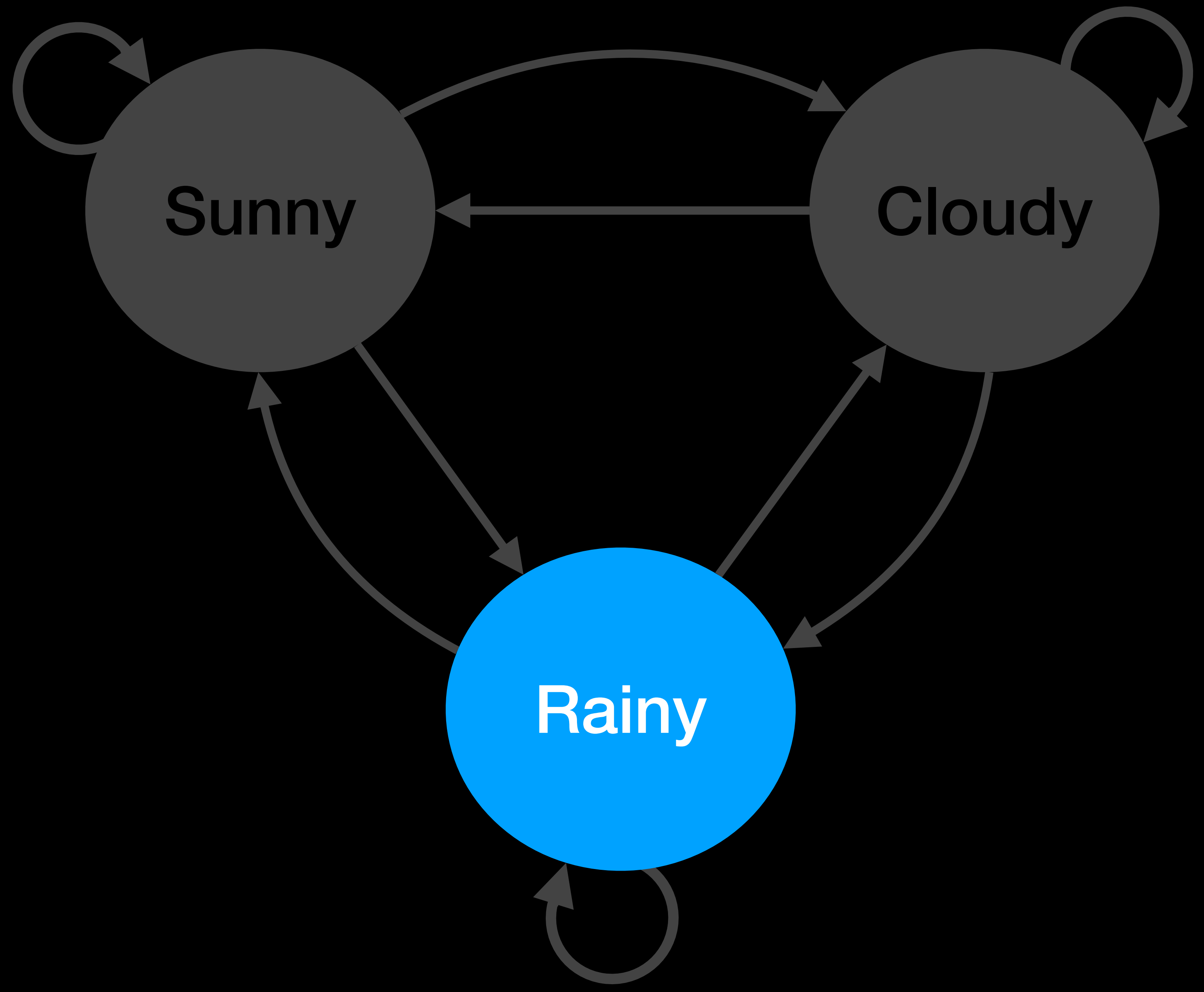
Cloudy

Generative use of a weather model



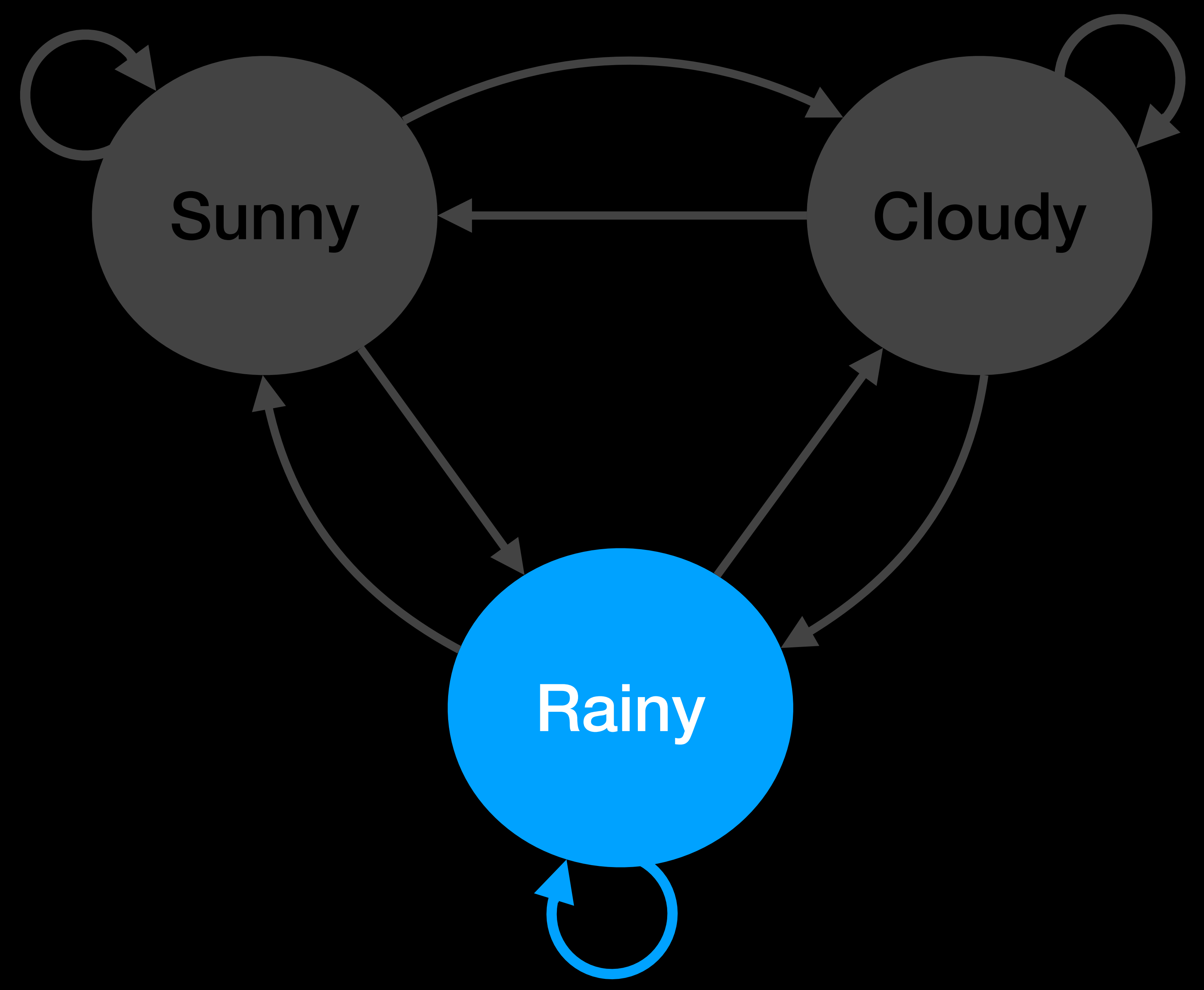
Cloudy

Generative use of a weather model



Cloudy, Rainy

Generative use of a weather model



Cloudy, Rainy, Rainy

ChatGPT = Chatbot + GPT3

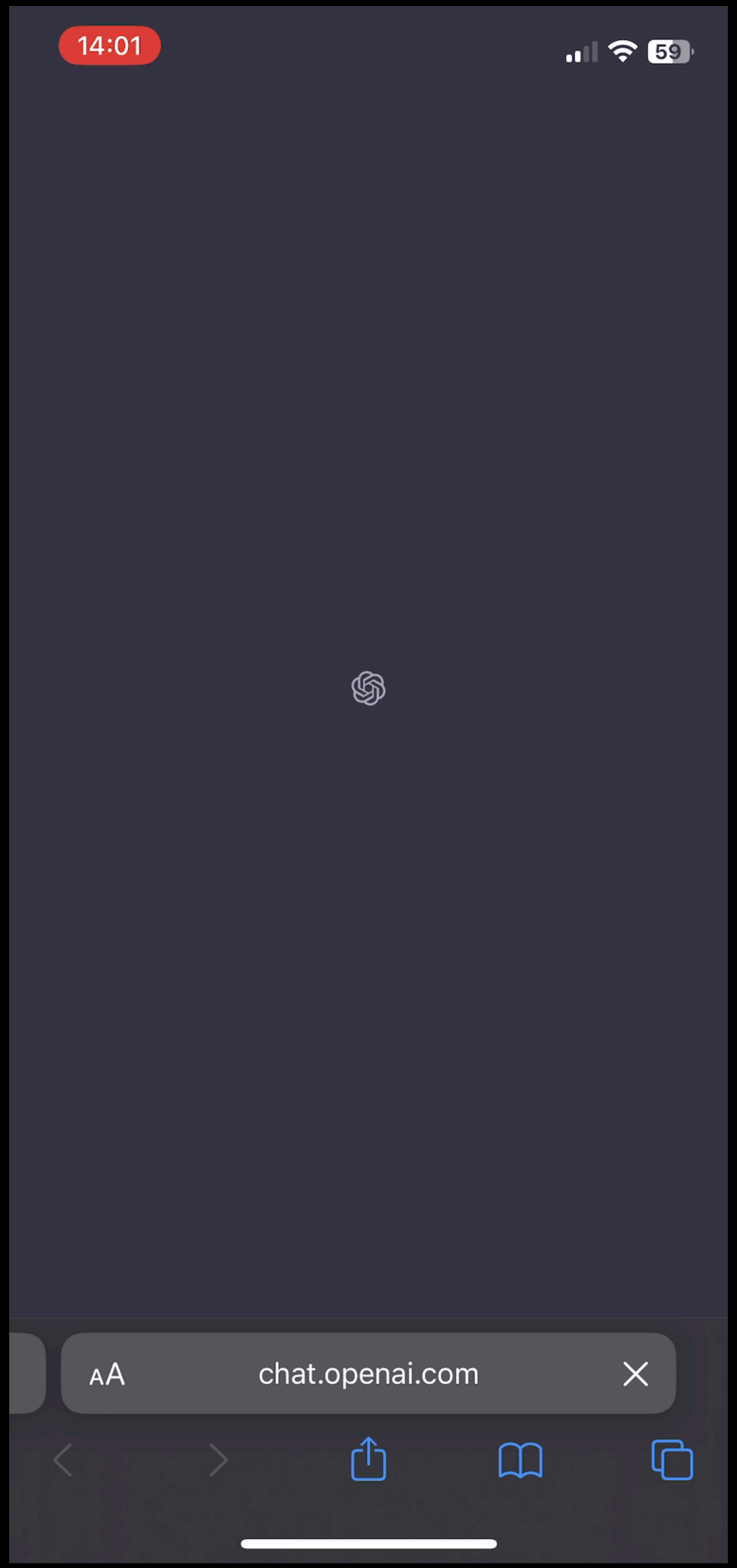
Chatbot: developed by OpenAI
mix of supervised & reinforcement learning

GPT3: Generative Pre-trained Transformer 3
type of **large language model** (fancy predictive text)

“The quick brown fox jumps over the _____”

Lazy 95%
Slow 2%
Fun 1%
…
Zyzzyva 0%

Trained on all available text on the internet



Major issues with large language models

Based on what ever data it was trained on

- May not be relevant, accurate, or pleasant

Generative process is inherently stochastic

- Response choices and sentence construction depend on sampling distributions randomly

Hard to evaluate and verify

- How often will it be right? What is right?

Clinical AI

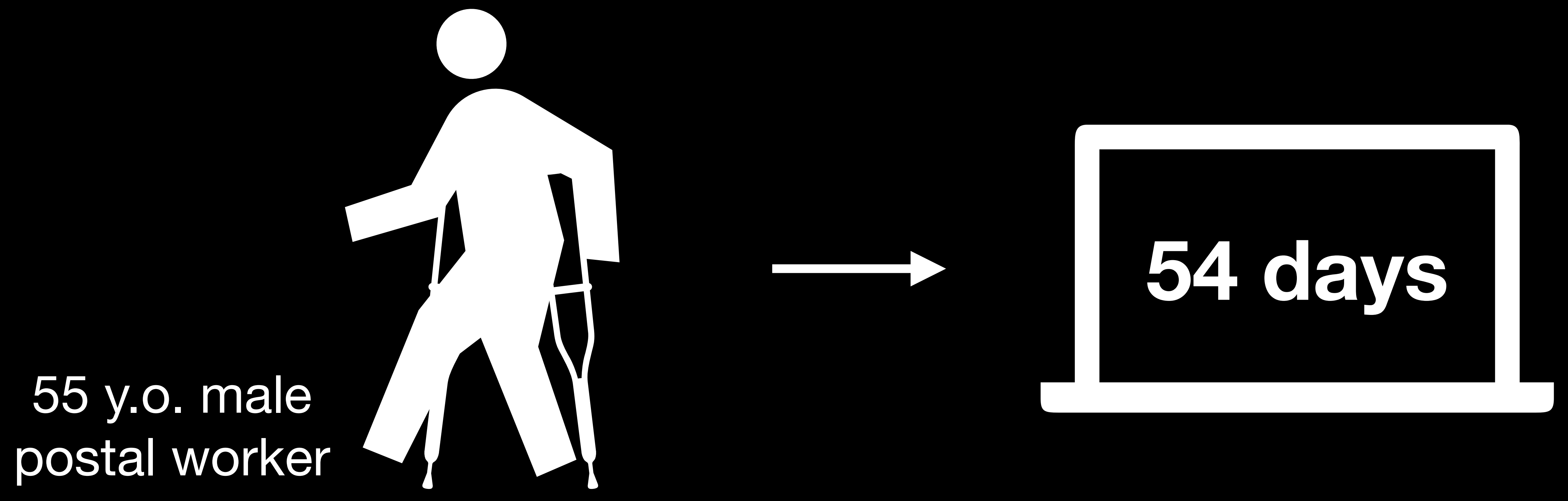
Repurposing Predictive Models

Return to Work

Our OI Goal

Predict the work-status over the course of a patient’s recovery.

Existing return to work models ignore longitudinal observations.



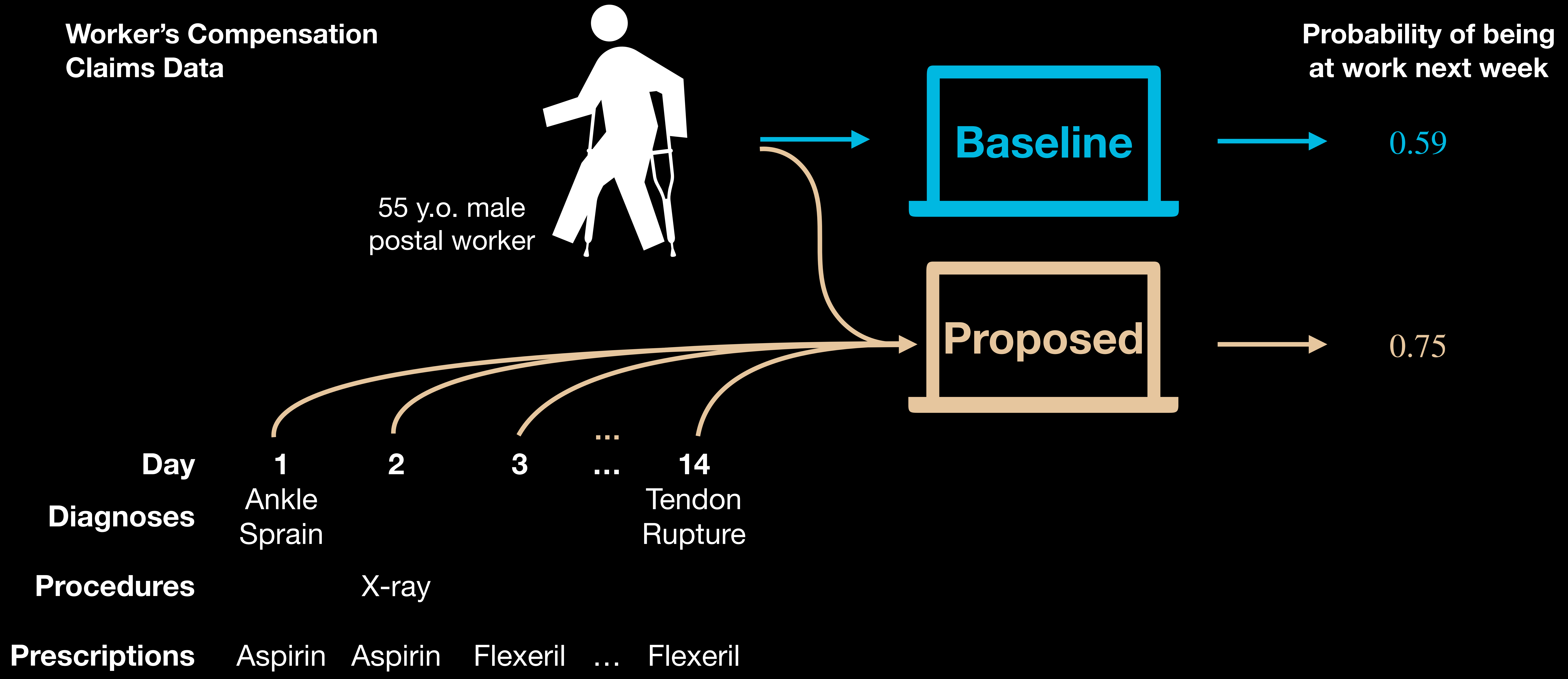
Gaspar 2017, MCG 2022, MDG 2022

What is the value of longitudinal observations in return to work prediction?

Do we observe a performance improvement when using longitudinal observations collected beyond the time of injury?

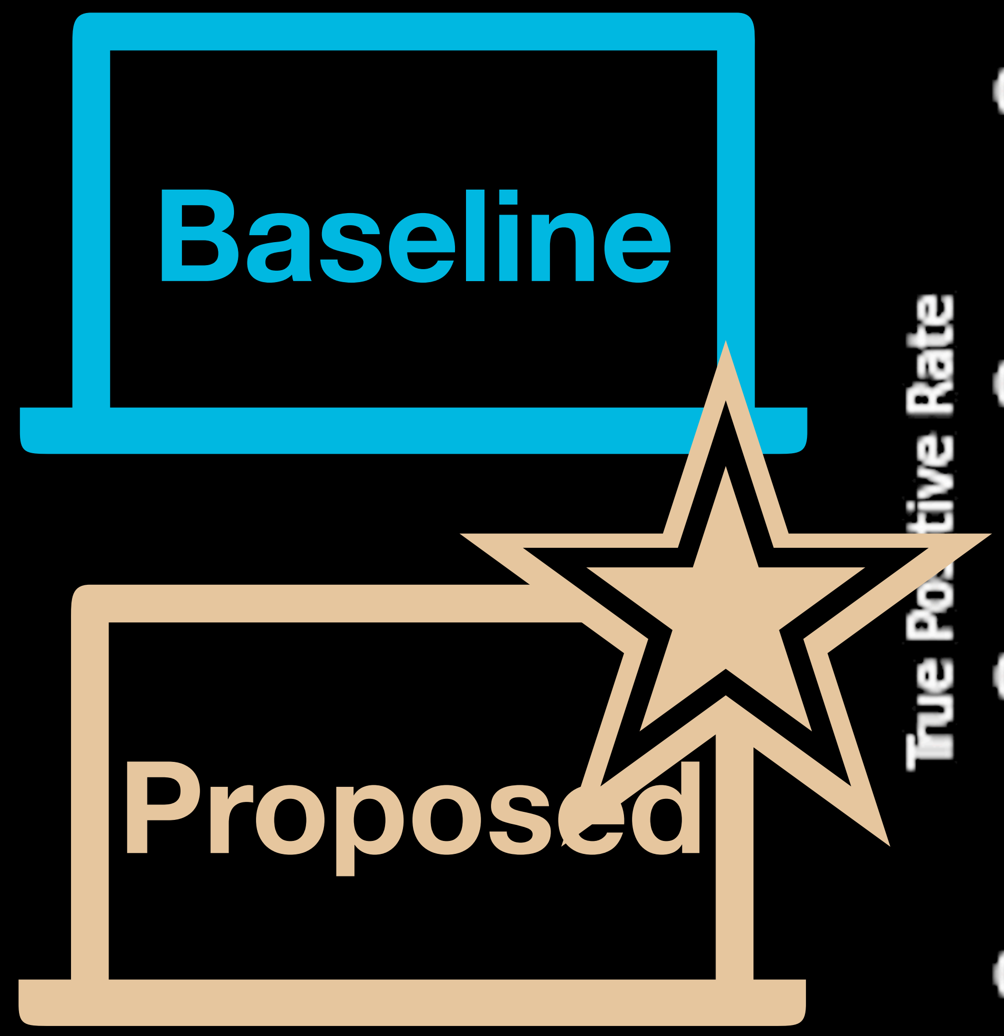
Presume longitudinal observations improve predictions in other healthcare task.

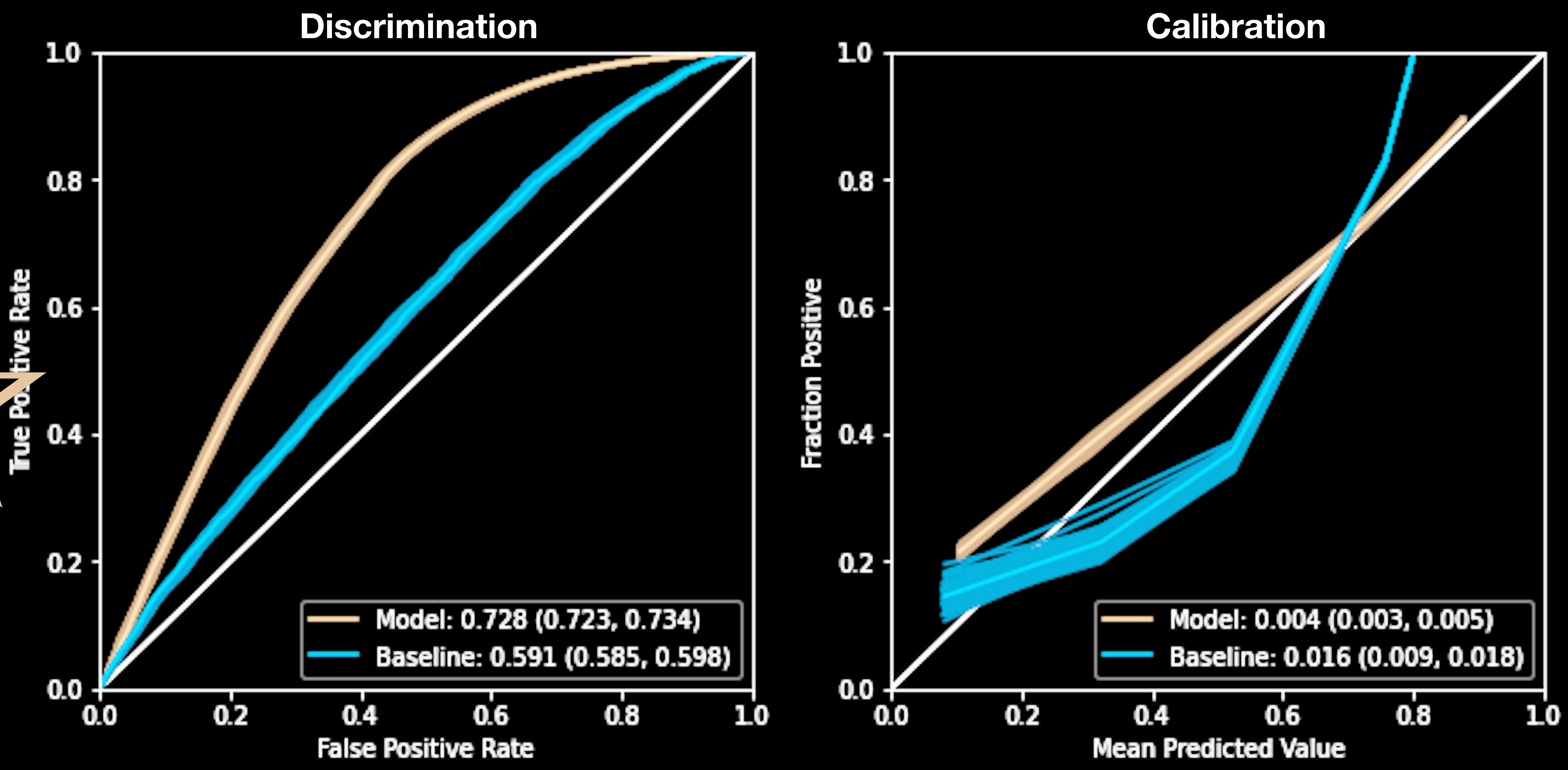
Experimental setup



Ötleş et al. 2022

Results

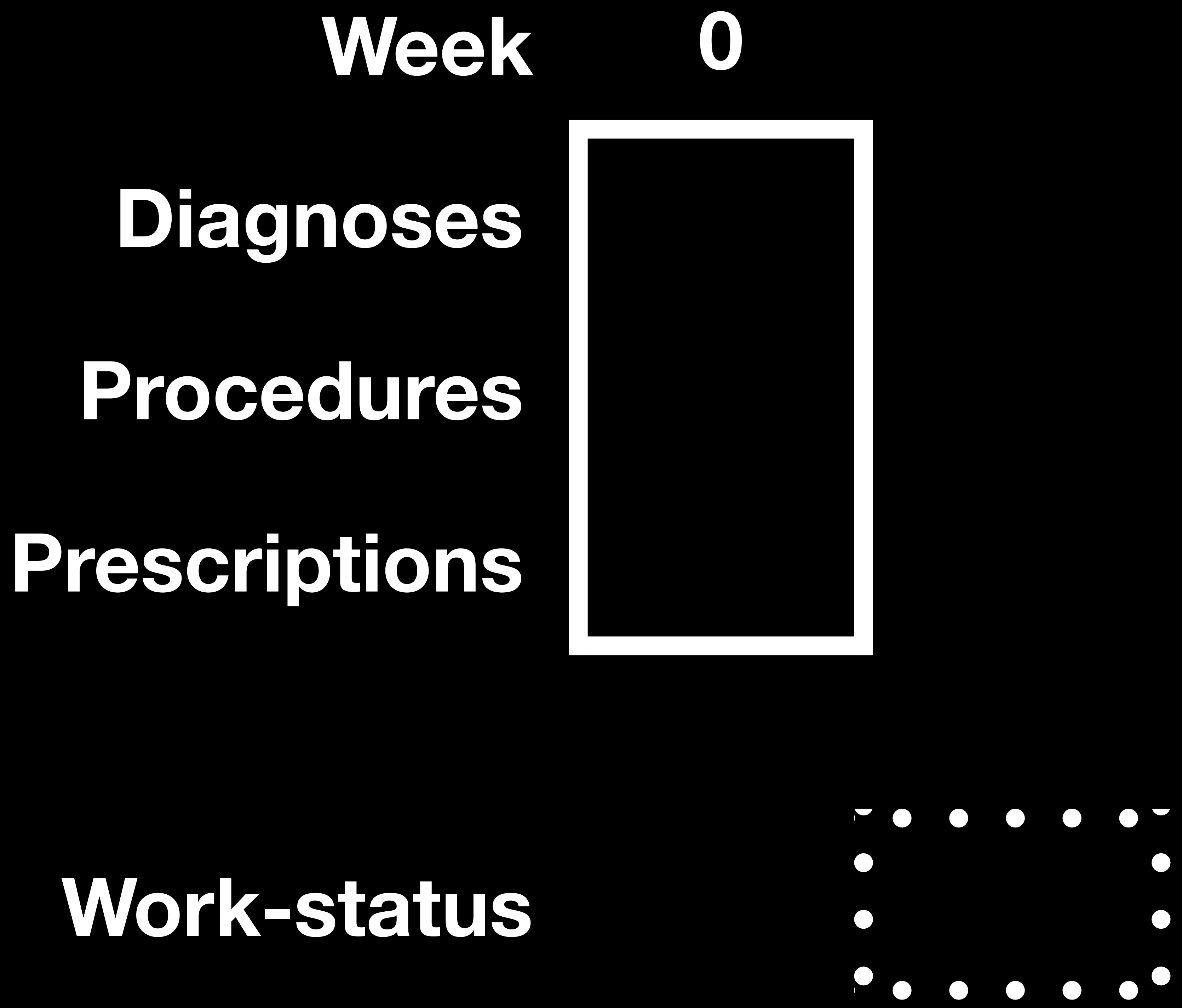




Ötleş et al. 2022

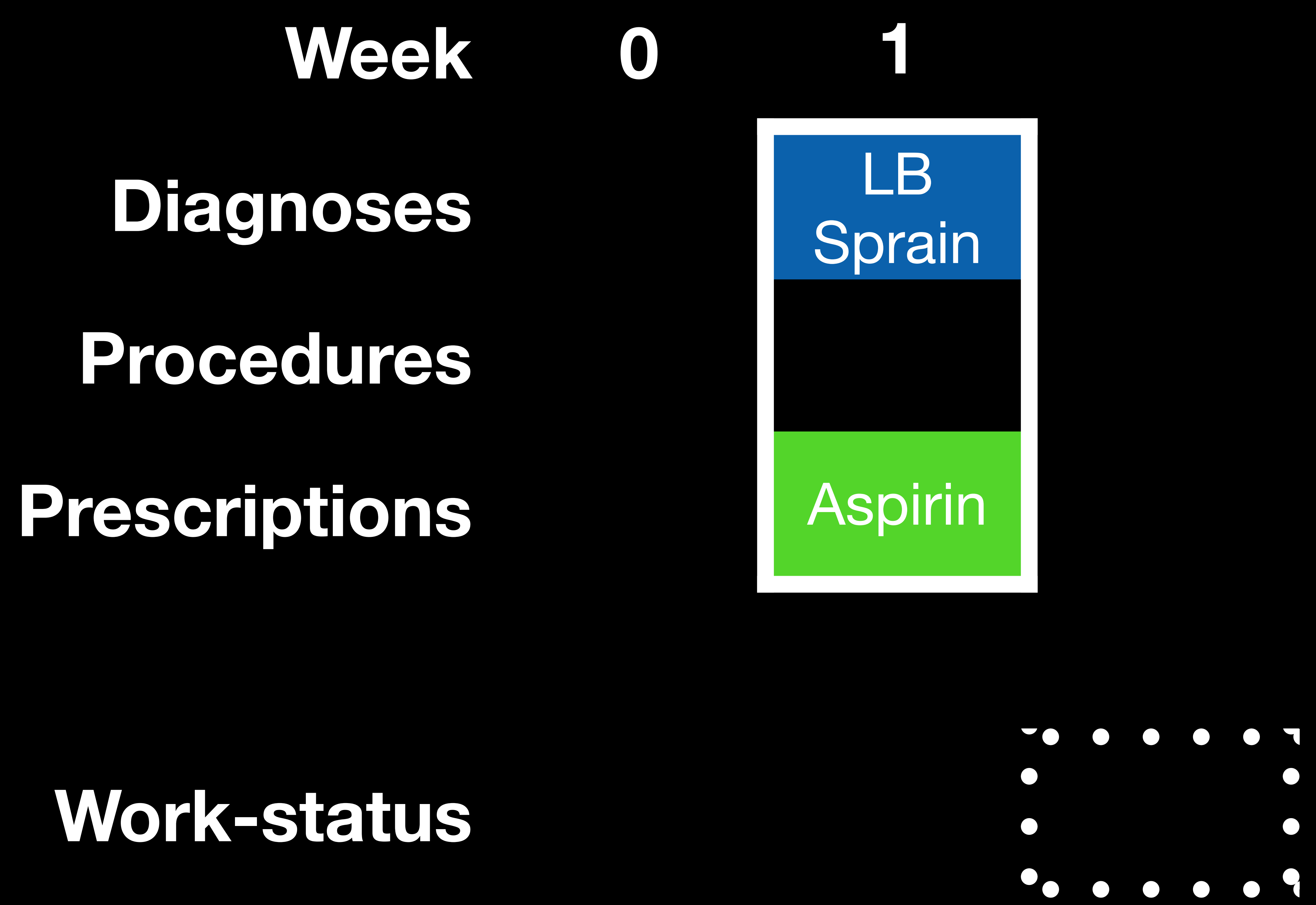
Underlying predictive model can be converted to a generative model.

Sequential prediction of recovery



25y Male Dairy Farmer

Sequential prediction of recovery



25y Male Dairy Farmer

Sequential prediction of recovery

Week	0	1	2
Diagnoses		LB Sprain	
Procedures			Chiro.
Prescriptions		Aspirin	
Work-status			

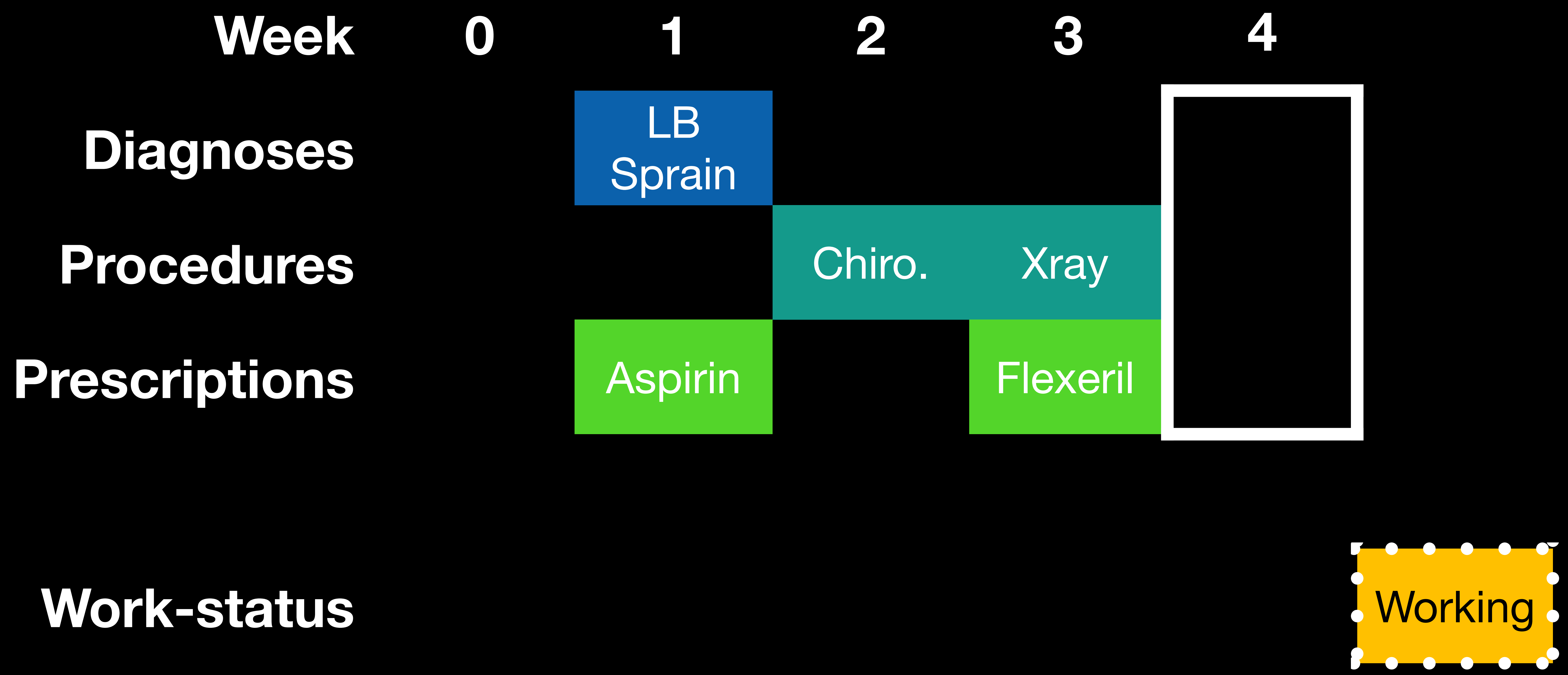
25y Male Dairy Farmer

Sequential prediction of recovery

Week	0	1	2	3
Diagnoses		LB Sprain		
Procedures			Chiro.	Xray
Prescriptions		Aspirin		Flexeril
Work-status				

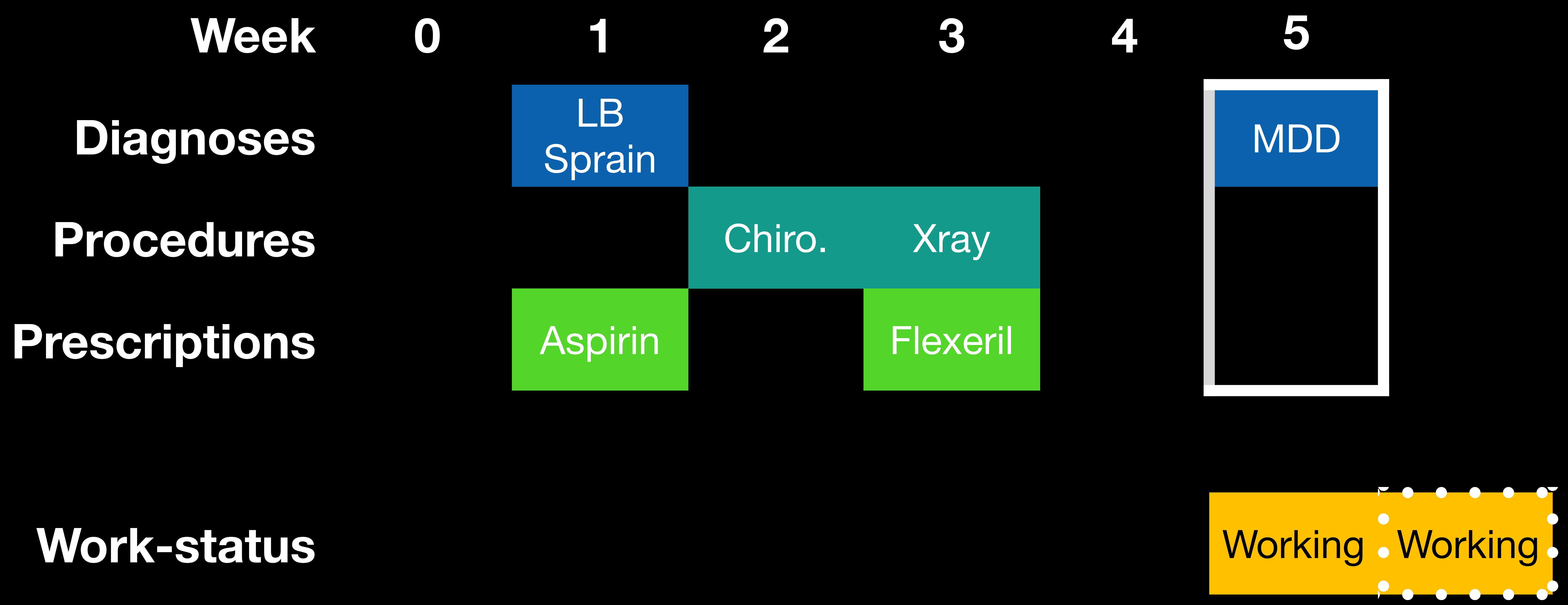
25y Male Dairy Farmer

Sequential prediction of recovery



25y Male Dairy Farmer

Sequential prediction of recovery



25y Male Dairy Farmer

Sequential prediction of recovery

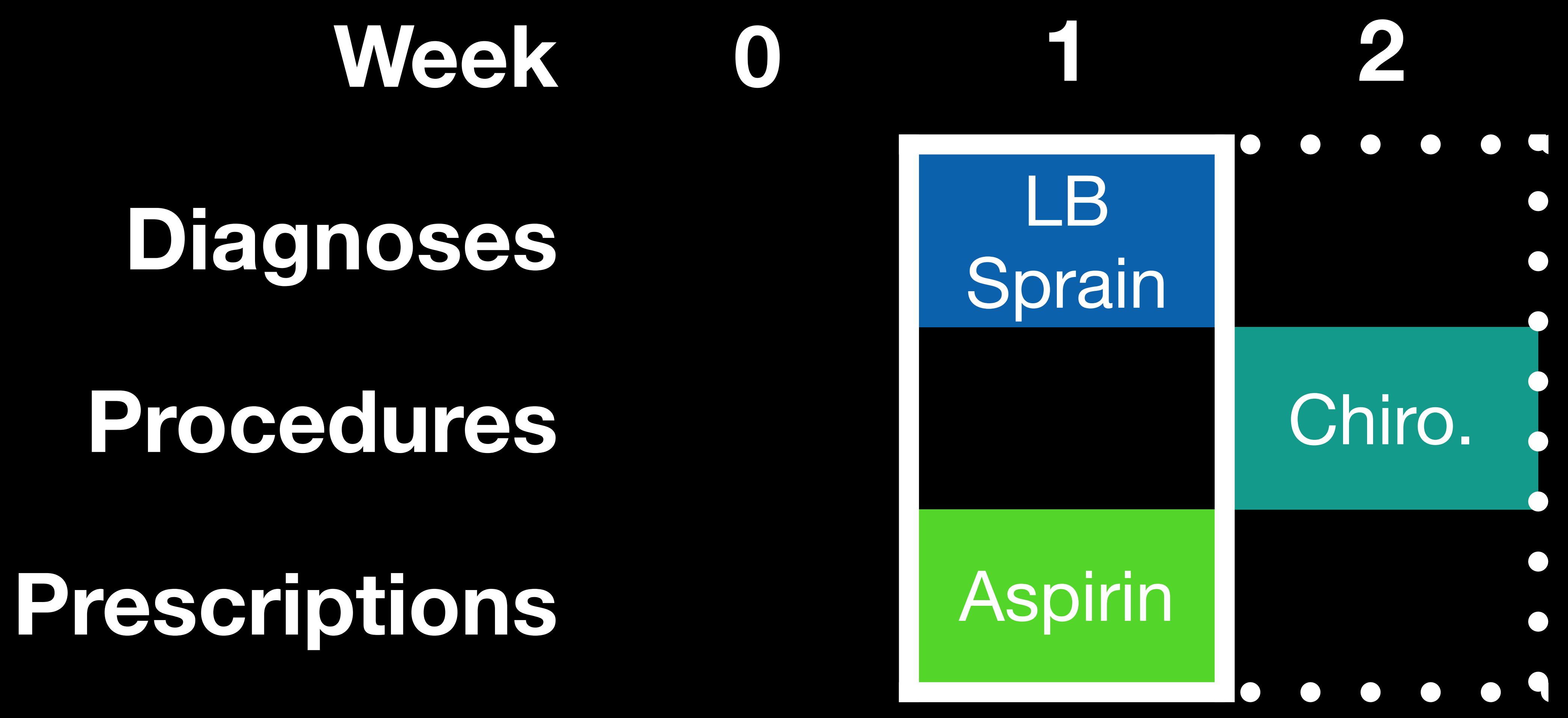
Week	0	1	2	3	4	5	6
Diagnoses		LB Sprain				MDD	
Procedures			Chiro.	Xray			MRI
Prescriptions		Aspirin		Flexeril			
Work-status						Working	Working

25y Male Dairy Farmer

Generation of recovery trajectory

25y Male Dairy Farmer

Generation of recovery trajectory



25y Male Dairy Farmer

Generation of recovery trajectory

Week	0	1	2	3
Diagnoses		LB Sprain		
Procedures			Chiro.	Xray
Prescriptions		Aspirin		Flexeril

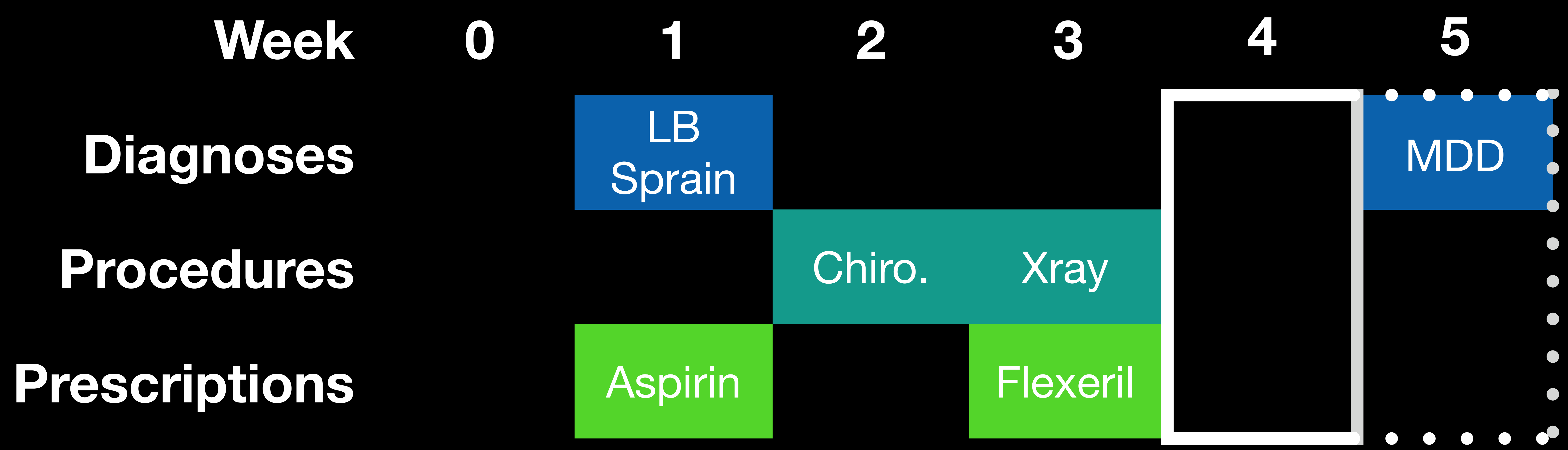
25y Male Dairy Farmer

Generation of recovery trajectory

Week	0	1	2	3	4
Diagnoses		LB Sprain			
Procedures			Chiro.	Xray	
Prescriptions		Aspirin		Flexeril	

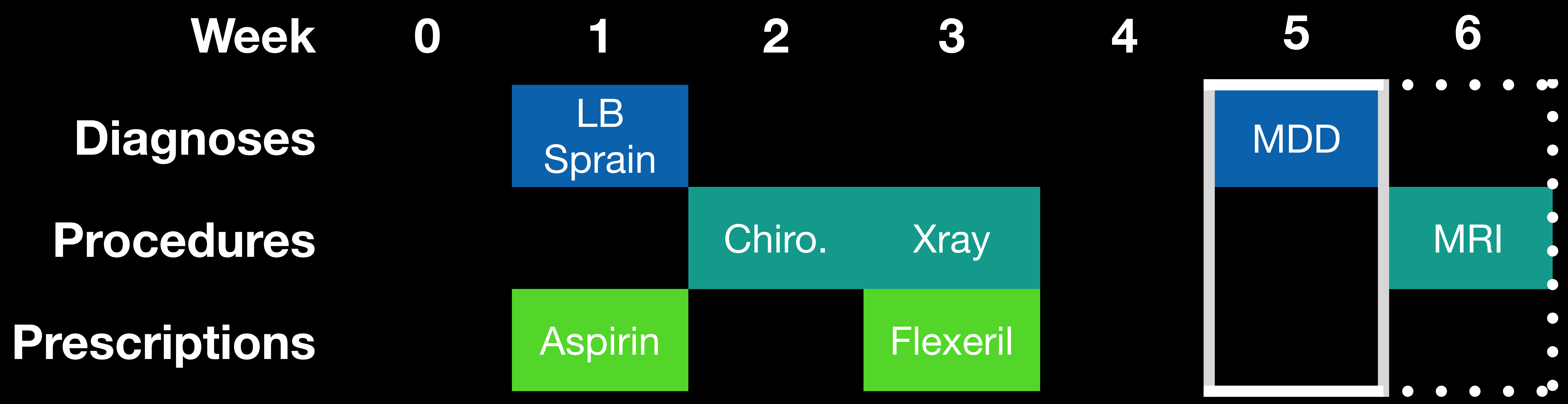
25y Male Dairy Farmer

Generation of recovery trajectory



25y Male Dairy Farmer

Generation of recovery trajectory

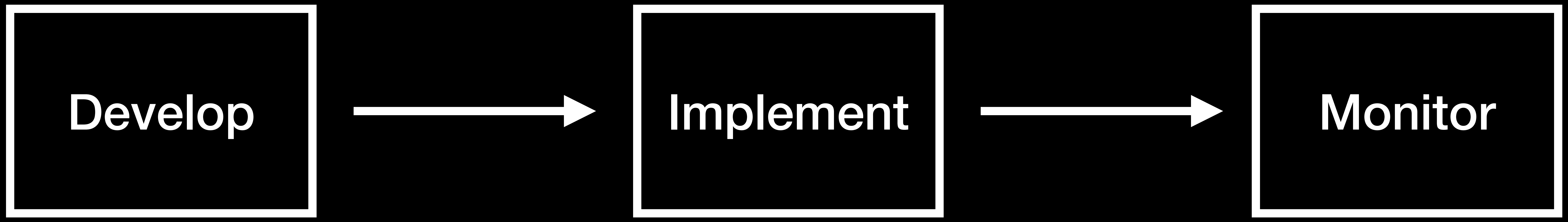


25y Male Dairy Farmer

Constant evaluation is fundamental

Prostate cancer
C. difficile infection risk
Sepsis

Simplified model lifecycle

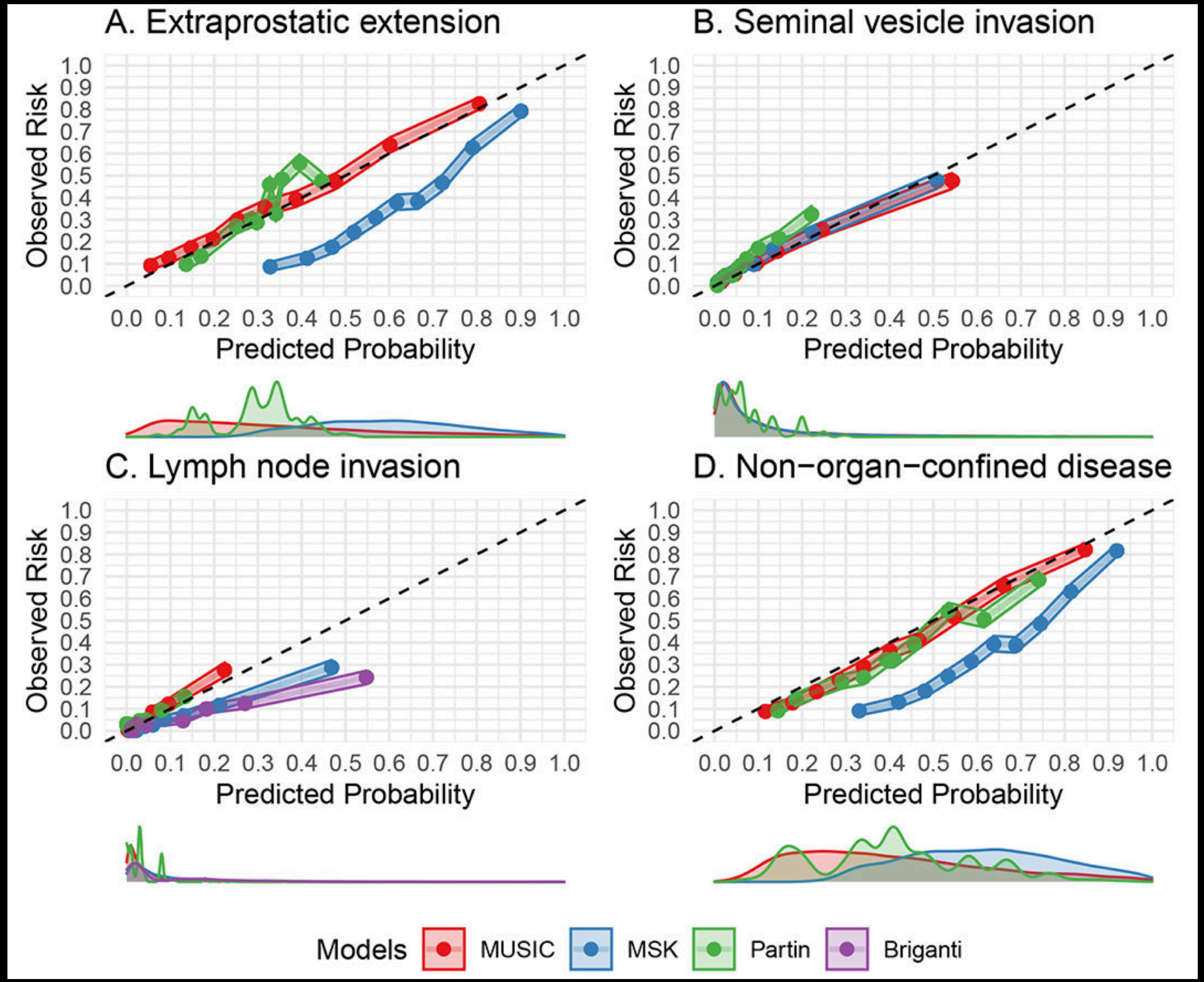


Simplified model lifecycle

Evaluation of prostate cancer pathological outcomes prediction

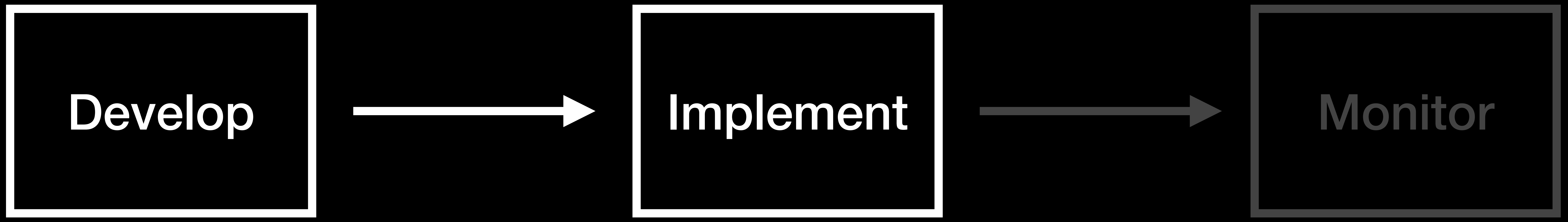
Internal vs. External Validation

Models developed by other institutions under perform compared to training institution specific models



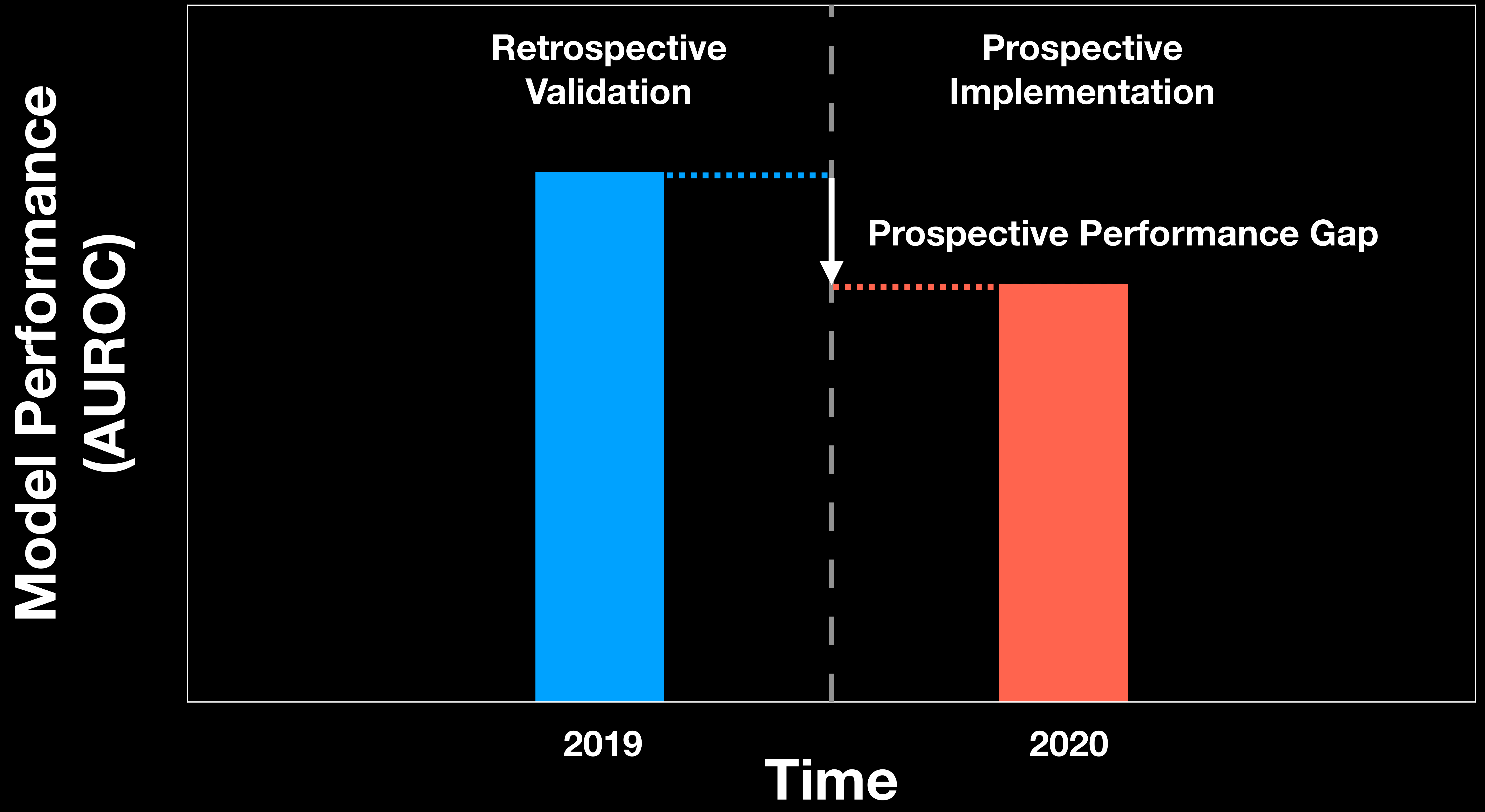
Ötleş et al. 2022

Simplified model lifecycle



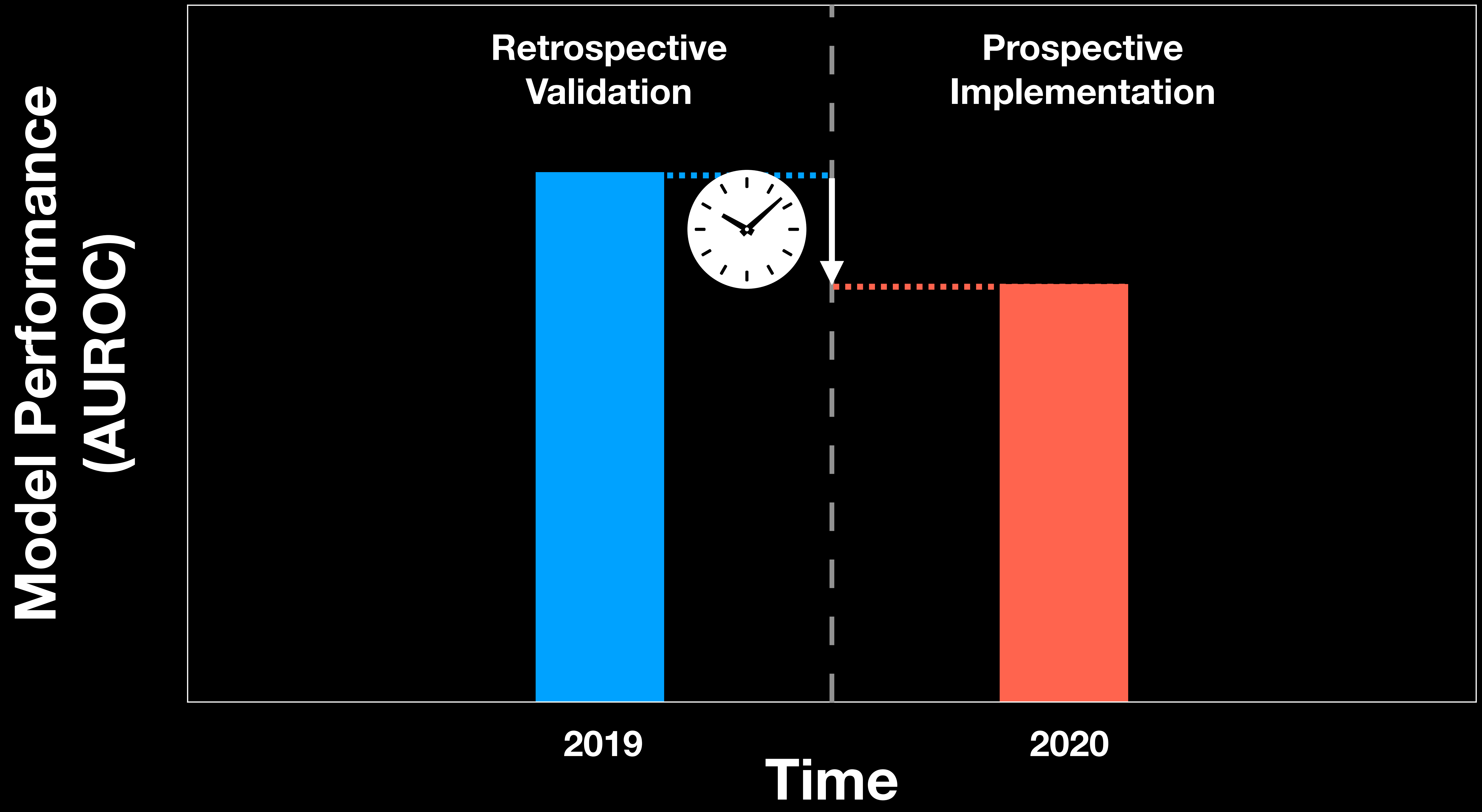
Prospective evaluation of inpatient *C. difficile* infection risk prediction

Model performance may degrade after implementation.

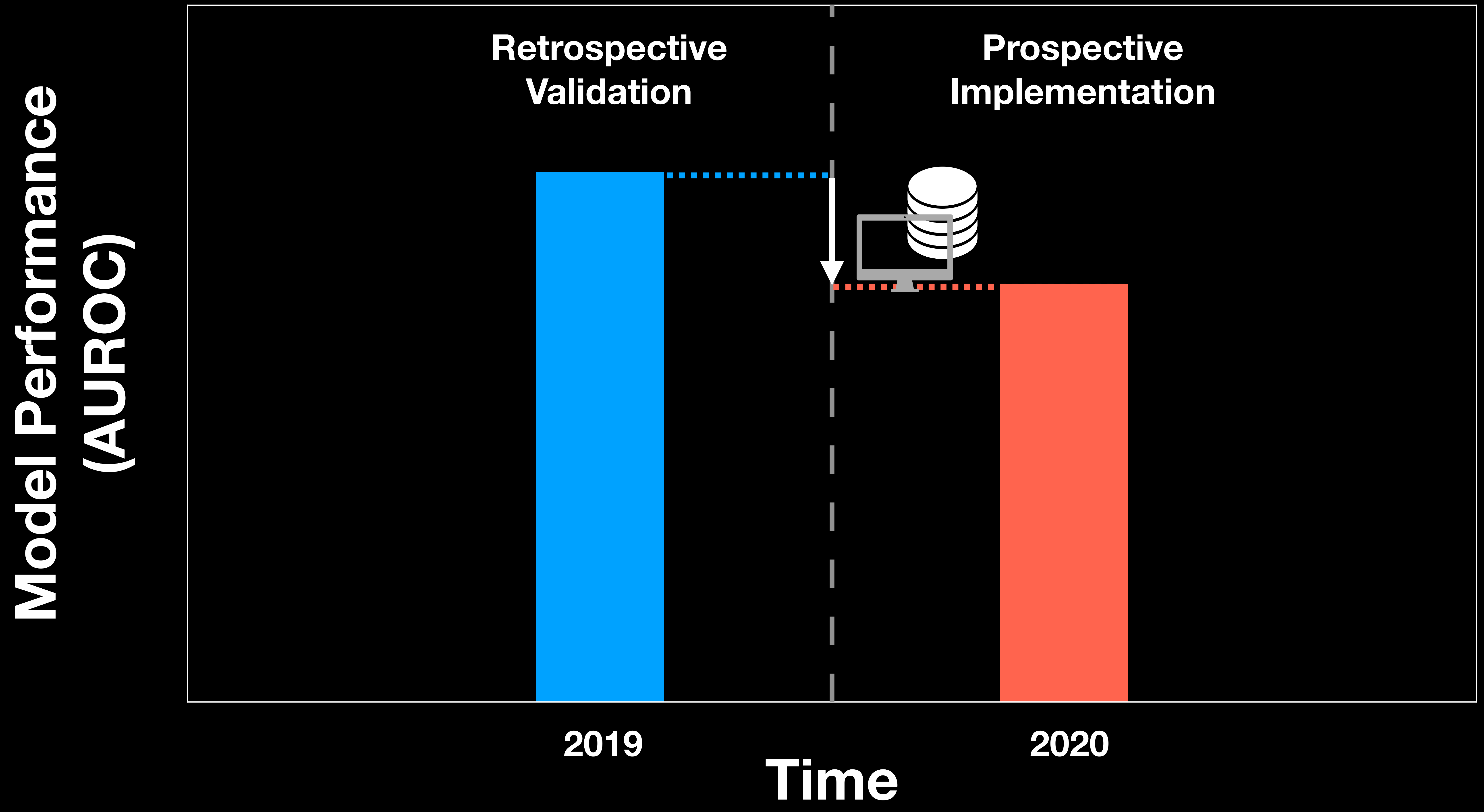


Davis 2017, Hickey 2013, Minne 2012

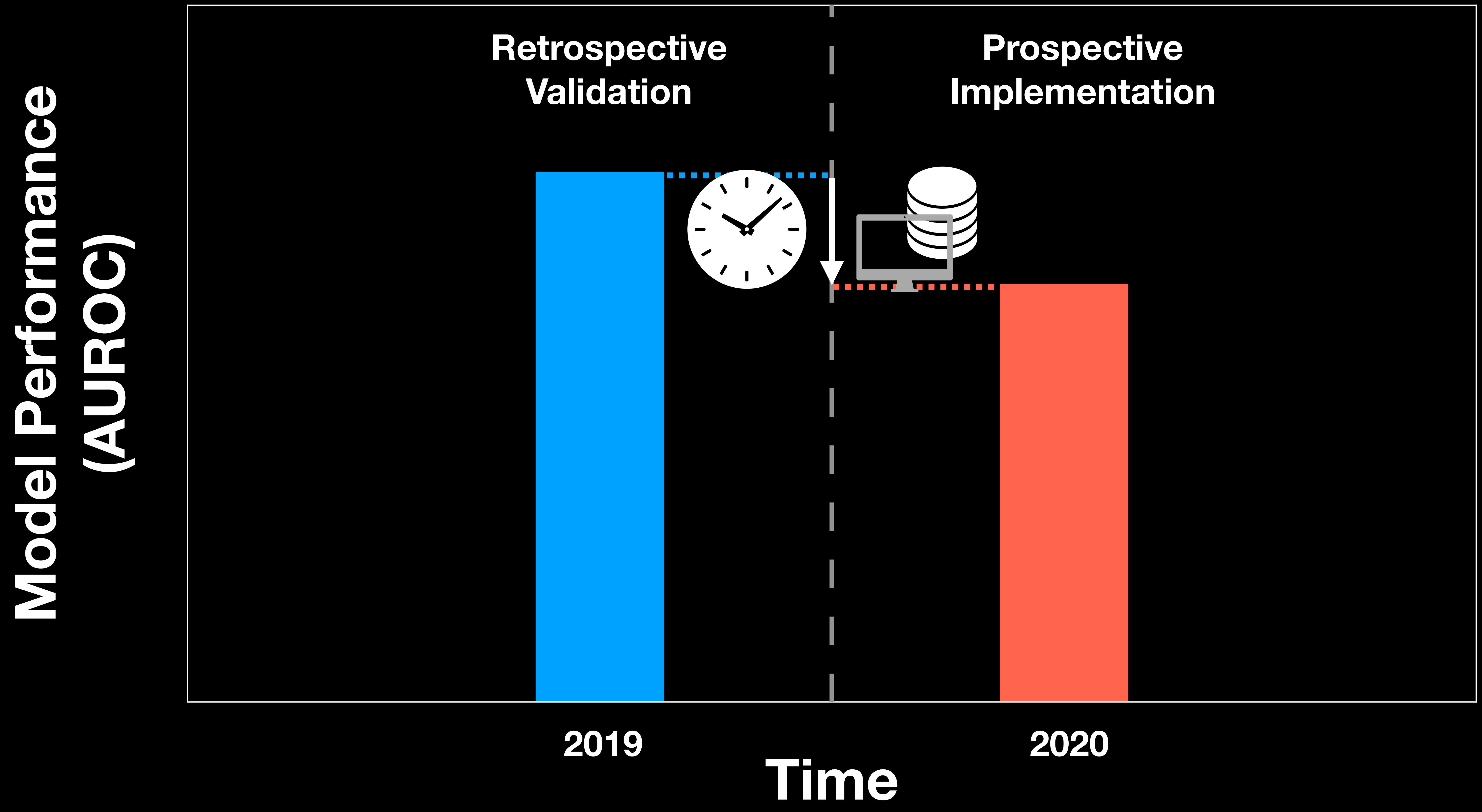
This degradation is often attributed to changes in populations & practice that occurs over time.



However, changes in IT infrastructure may also affect the prospective performance gap.

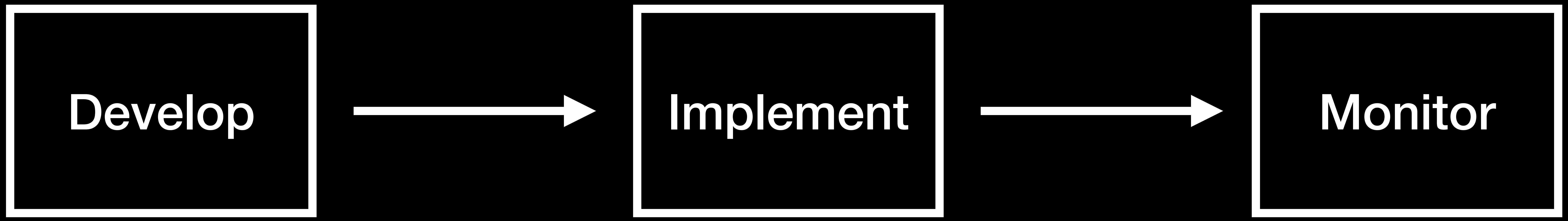


Degradation due to temporal & infrastructure shift.



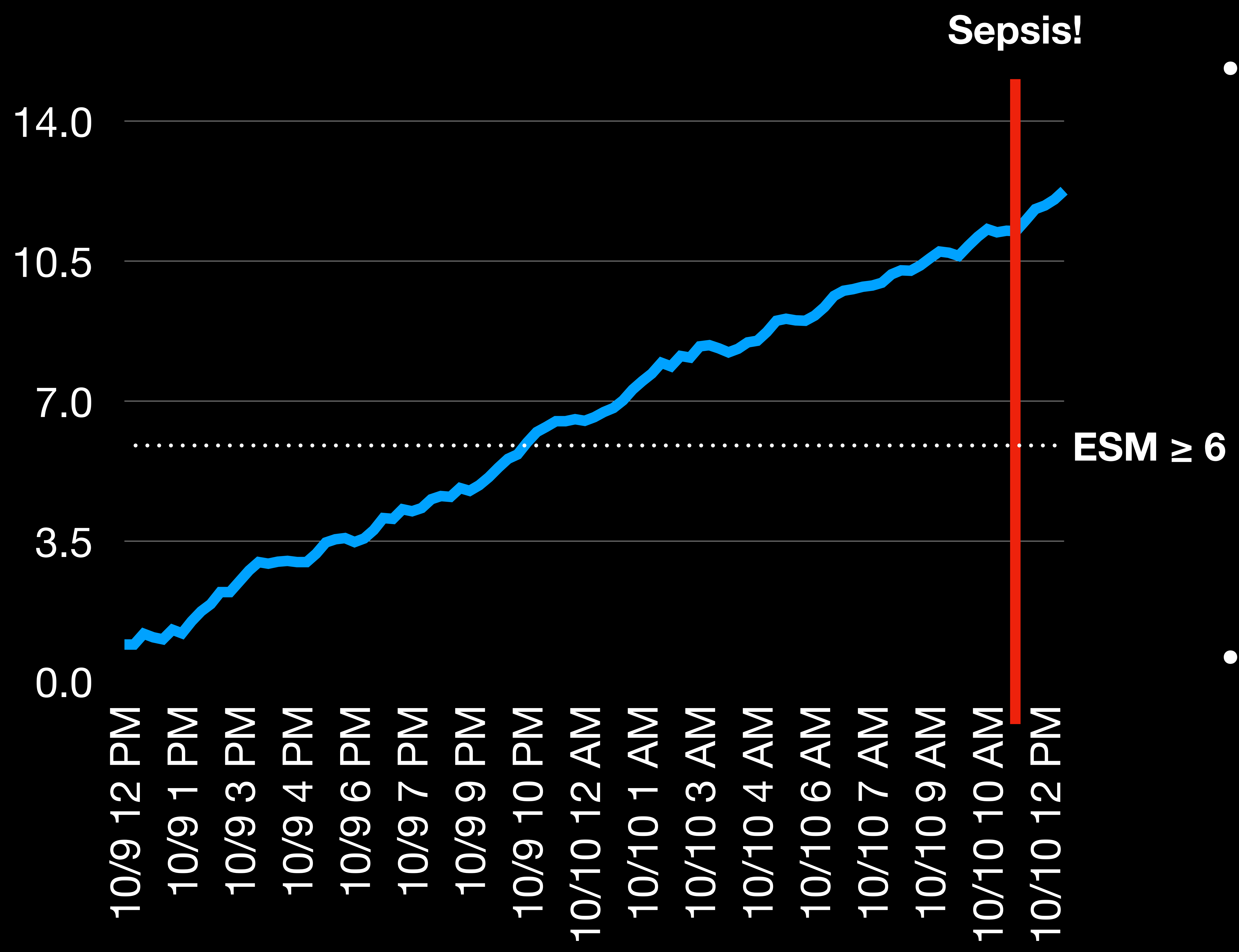
Ötleş & Oh et al. 2021

Simplified model lifecycle



Prospective evaluation of Epic sepsis model

Epic Sepsis Model



- Development
  - Inputs: vital signs, medication orders, lab values, comorbidities, and demographic information.
  - Outputs: ICD-9 code indicating diagnosis of sepsis - timing 6hrs prior to clinical intervention
- Implementation
  - Runs every 15 minutes on all patients in hospital
  - Expected AUROC performance ~ 0.8

Wong et al. 2021

Table 2

Table 2. ESM Performance

Model performance	Hospitalization	Time horizons 24 h	12 h	8 h	4 h
Outcome incidence, %	6.6	0.43	0.29	0.22	0.14
Area under the receiver operating characteristic curve (95% CI)	0.63 (0.62-0.64)	0.72 (0.72-0.72)	0.73 (0.73-0.74)	0.74 (0.74-0.75)	0.76 (0.75-0.76)
Positive predictive value (ESM score ≥6), %	12	2.4	1.7	1.4	0.92
No. needed to evaluate (ESM score ≥6)[a]	8	42	59	73	109

Abbreviation: ESM, Epic Sepsis Model.

[a] The number needed to evaluate makes different assumptions at the hospitalization and time horizon levels. At the hospitalization level, the number needed to evaluate assumes that each patient would be evaluated only the first time the ESM score is 6 or higher. For each time horizon, the number needed to evaluate assumes that each patient would be evaluated every time the ESM score is 6 or higher.

Wong et al. 2021

Confusion Matrix Math

	Sepsis	No Sepsis	
ESM ≥ 6	843	5,948	**6,791**
ESM < 6	1,709	29,955	**31,664**
	2,552	**35,903**	**38,445**

Wong et al. 2021

Confusion Matrix Math

	Sepsis	No Sepsis	
ESM ≥ 6	843	5,948	**6,791**
ESM < 6	1,709	29,955	**31,664**
	2,552	**35,903**	**38,445**

$$PPV = \frac{TP}{TP + FP} = \frac{843}{6791} \approx 12\,\%$$

Wong et al. 2021

Confusion Matrix Math

	Sepsis	No Sepsis	
ESM ≥ 6	843	5,948	**6,791**
ESM < 6	1,709	29,955	**31,664**
	2,552	**35,903**	**38,445**

$$NNE = \frac{1}{PPV} = \frac{6791}{843} \approx 8$$

Wong et al. 2021

Confusion Matrix Math

	Sepsis (No Abx)	Sepsis (Abx)	
ESM ≥ 6	183	660	**843**
ESM < 6	679	1,030	**1,709**
	862	**1,690**	**2,552**

Wong et al. 2021

Confusion Matrix Math

	Sepsis (No Abx)	Sepsis (Abx)	
ESM ≥ 6	183	660	**843**
ESM < 6	679	1,030	**1,709**
	862	**1,690**	**2,552**

$$P(Useful \mid Correct) = \frac{P(Useful \cap Correct)}{P(Correct)} = \frac{183}{843} \approx 22\,\%$$

Wong et al. 2021

Why such a big difference between expected & observed performance?

Subtle choice of outcome definition

Development: ICD-9 code indicating diagnosis of sepsis

Our outcome: Health catalyst operational sepsis outcome

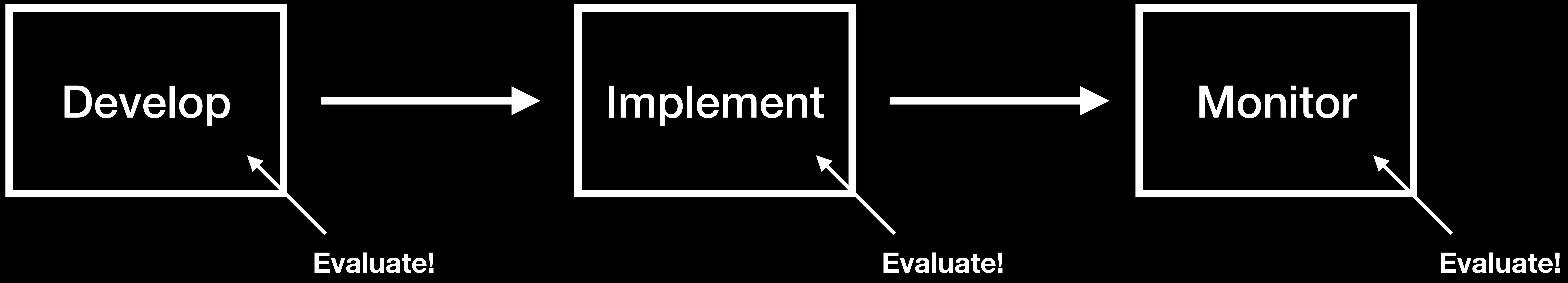
Billing lags behind actual clinical care

Sensitivity Analysis

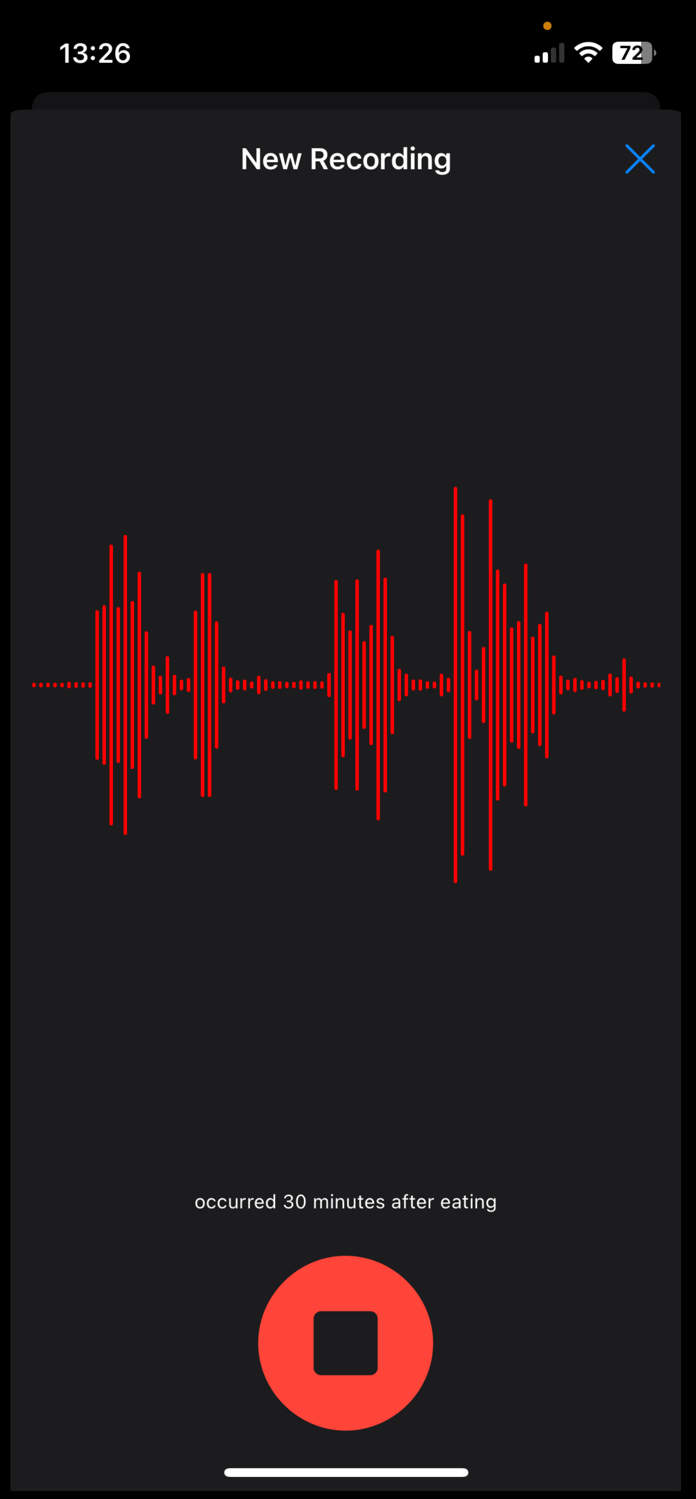
When ESM scores up to 3 hours after the onset of sepsis were included, the hospitalization-level AUC improved to 0.80 (95% CI, 0.79-0.81).

makes a big difference

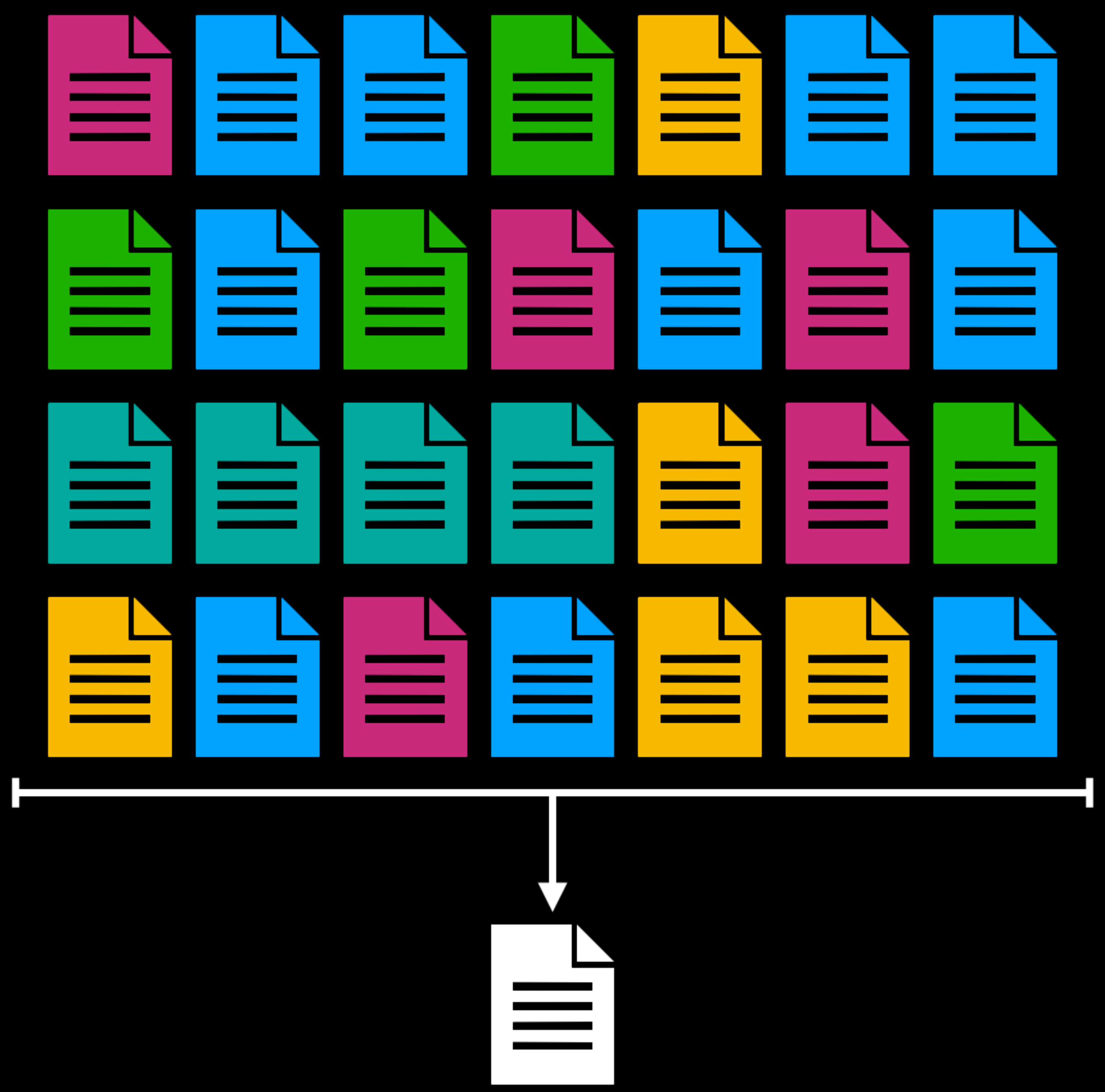
All models are wrong, but some are useful…



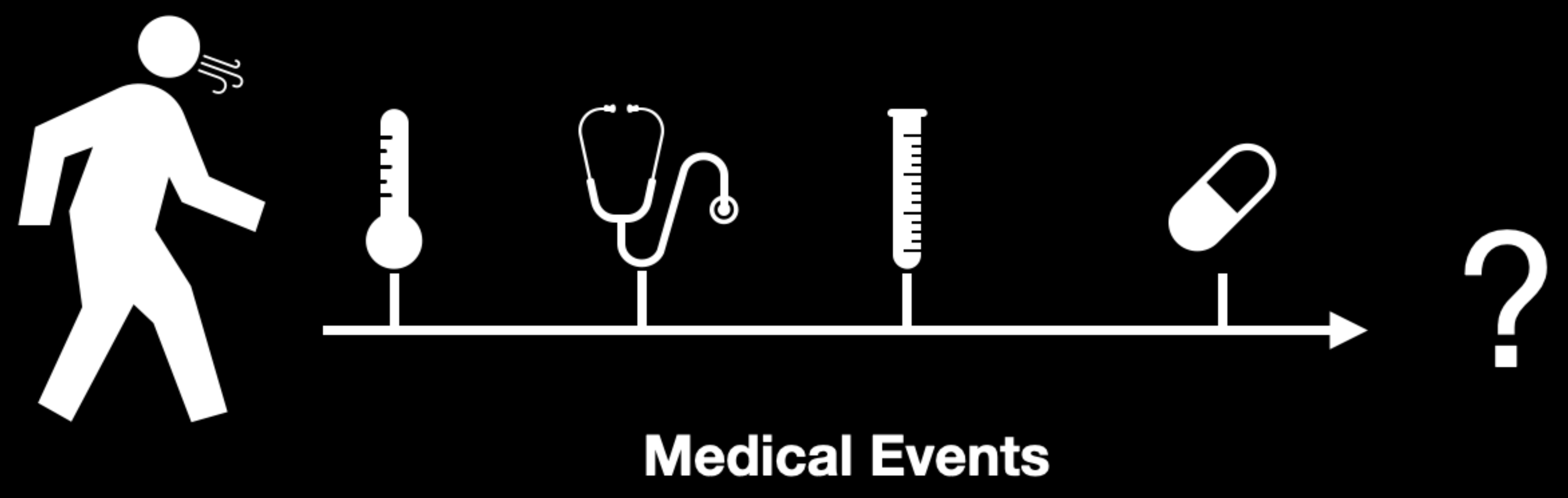
Generative AI Tools Coming Down the Pike



AI Scribe



AI Chart Summarization



Medical Foundation Models

AI Scribes

Goals:

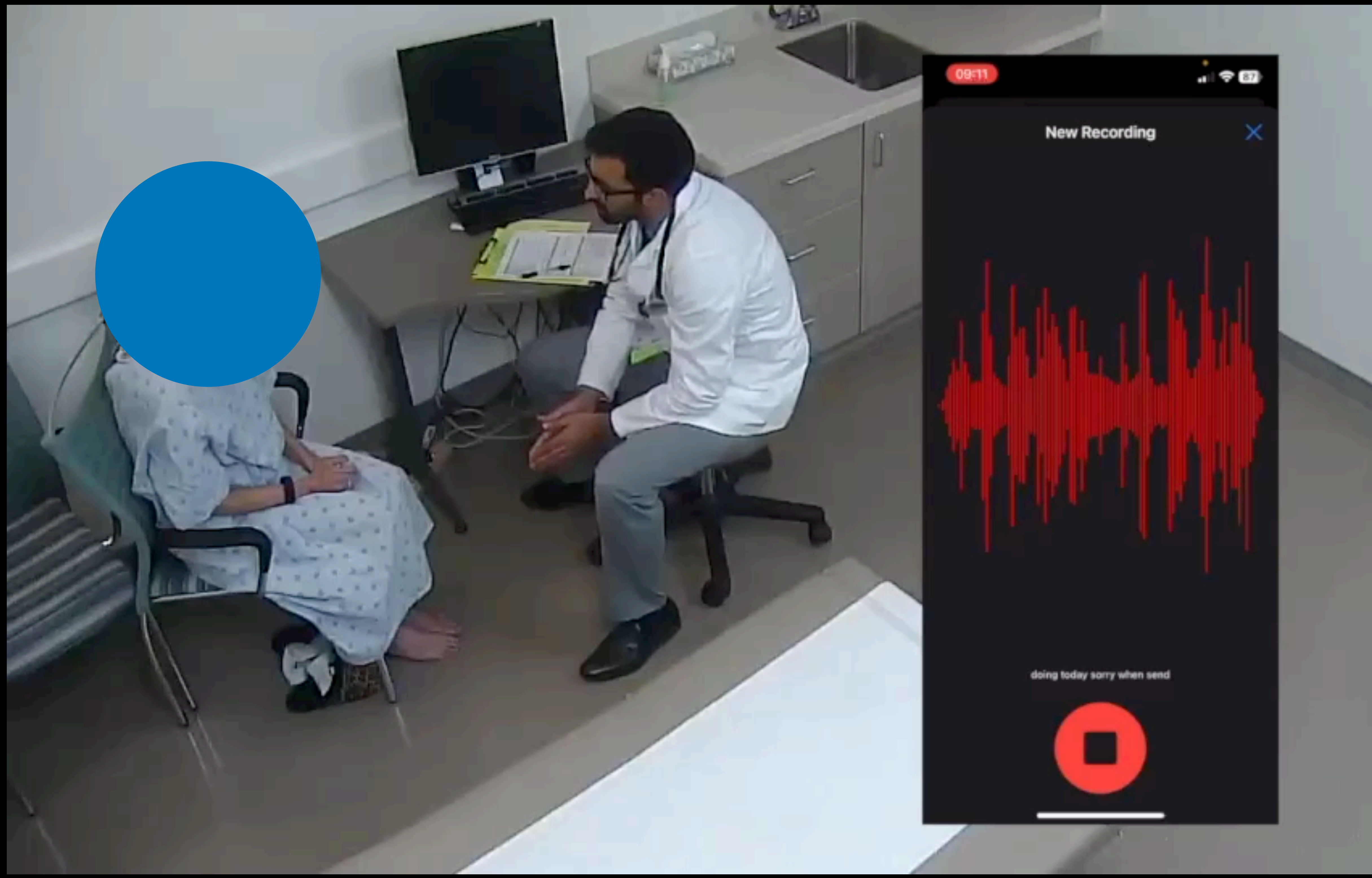
- Reduce burden of note creation
- Facilitate more face-to-face time

Technology:

- App records encounter
- Recording transcribed & converted to a note via LLM

Landscape:

- 2y ago very hard to build
- Now extremely easy to build - hard to validate



AI Scribes

Goals:

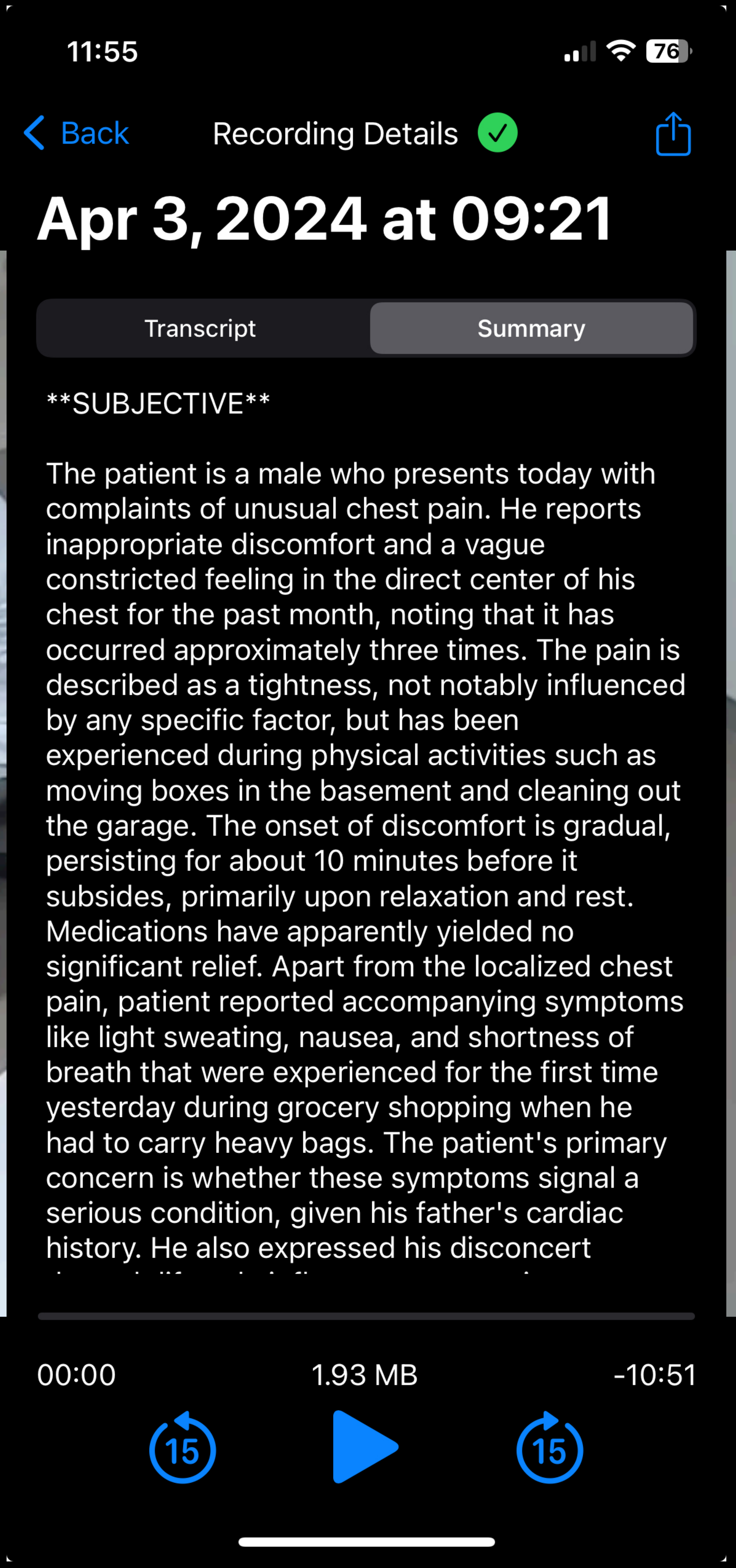
- Reduce burden of note creation
- Facilitate more face-to-face time

Technology:

- App records encounter
- Recording transcribed & converted to a note via LLM

Landscape:

- 2y ago very hard to build
- Now extremely easy to build - hard to validate



Do AI scribes actually benefit physicians?

- TPMG studied an AI scribe
    - From unnamed vendor
    - Accessible to a wide range of physicians
- As of publication time
    - 3k physicians, 303k encounters
- Studied
    - PJ time
    - Time in notes
    - Note quality

Tierney et al. 2024

Do AI scribes actually benefit physicians?

TPMG studied an AI scribe

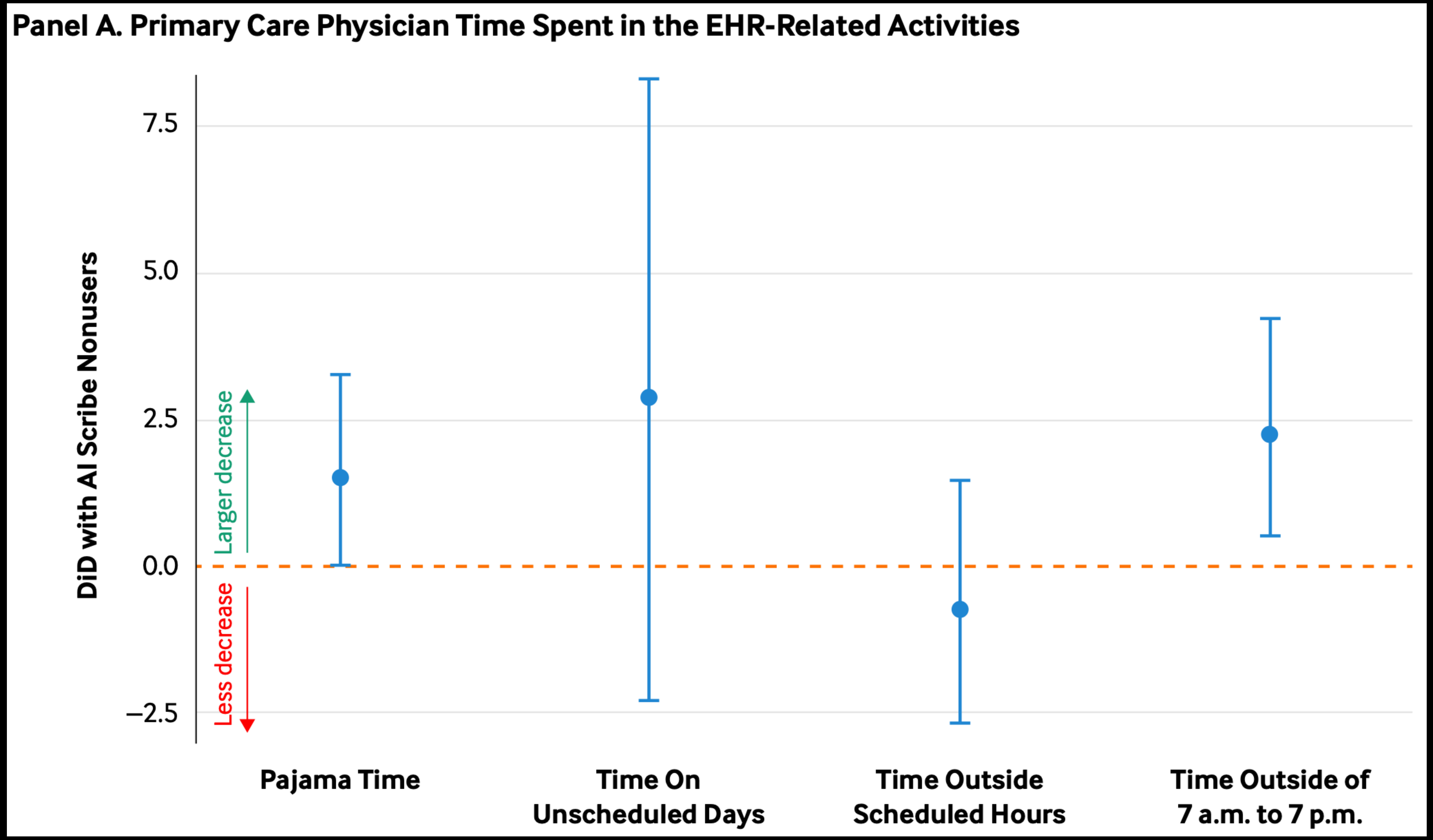
- From unnamed vendor
- Accessible to a wide range of physicians

As of publication time

- 3k physicians, 303k encounters

Studied

- PJ time ↓
- Time in notes
- Note quality



Tierney et al. 2024

Do AI scribes actually benefit physicians?

TPMG studied an AI scribe

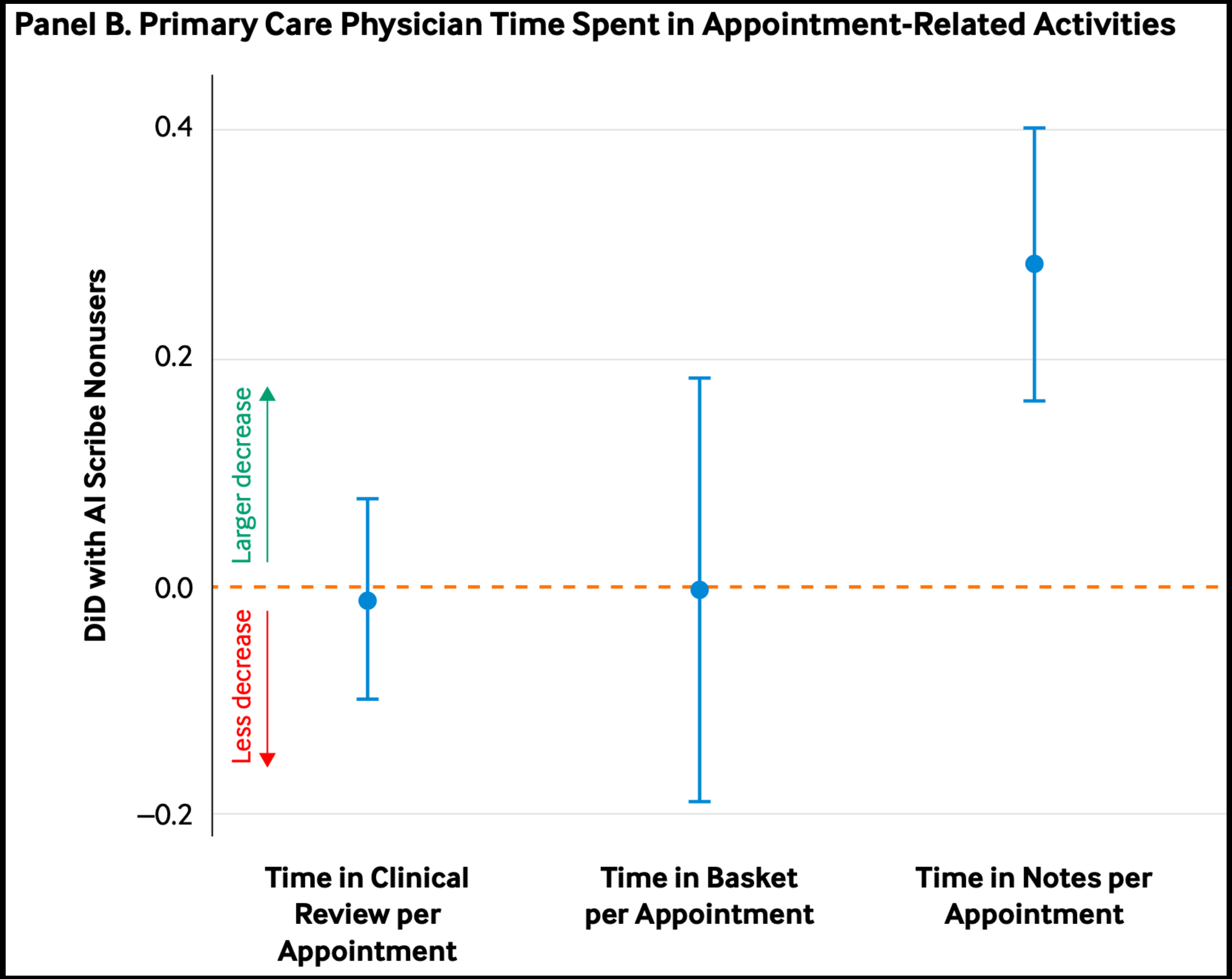
- From unnamed vendor
- Accessible to a wide range of physicians

As of publication time

- 3k physicians, 303k encounters

Studied

- PJ time ↓
- Time in notes ↓
- Note quality



Tierney et al. 2024

Do AI scribes actually benefit physicians?

TPMG studied an AI scribe

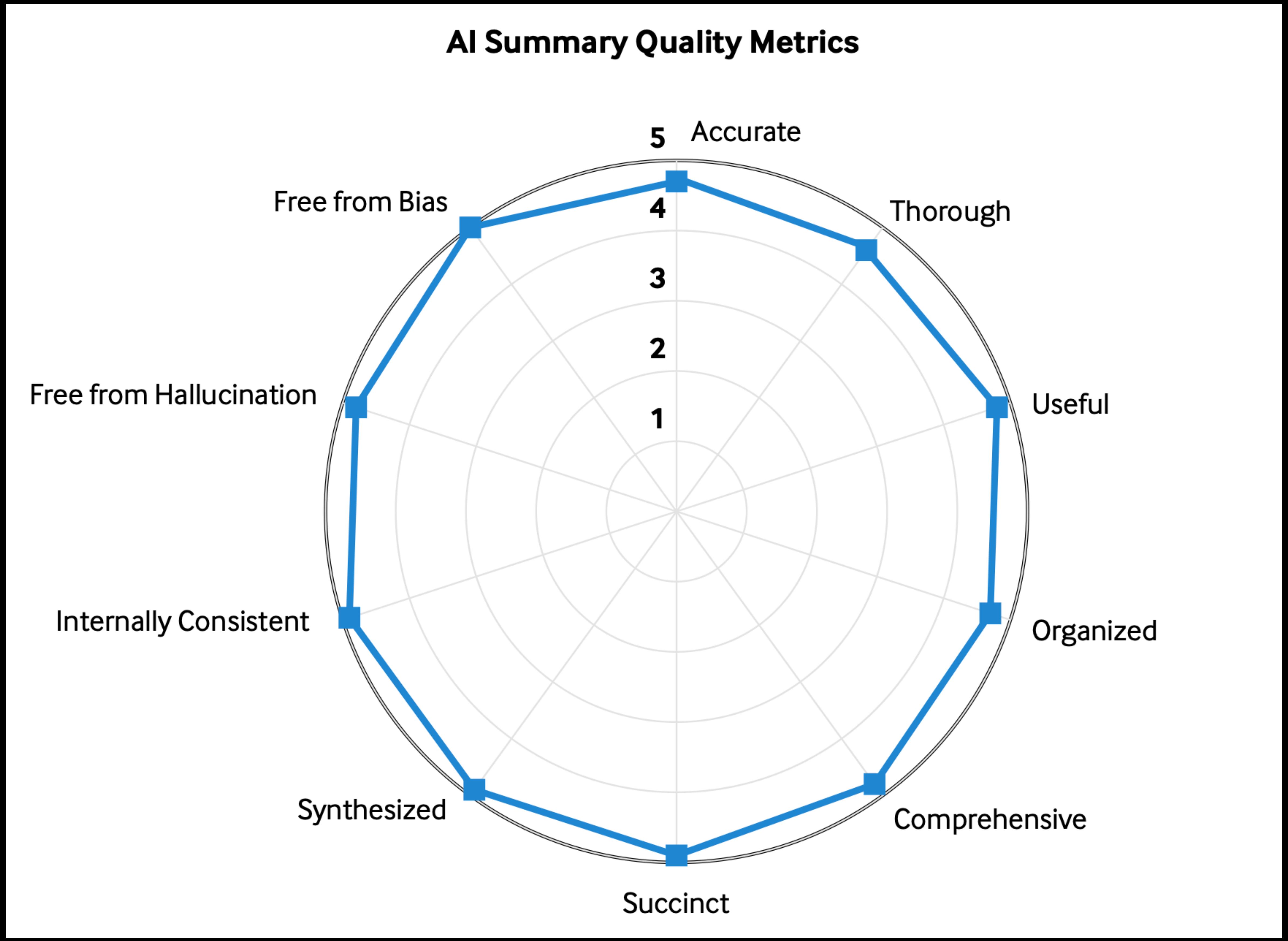
- From unnamed vendor
- Accessible to a wide range of physicians

As of publication time

- 3k physicians, 303k encounters

Studied

- PJ time ↓
- Time in notes ↓
- Note quality ~



Tierney et al. 2024

AI Chart Summarization

Goals:

Reduce burden of chart review

Help highlight relevant info

Reduce irrelevant info

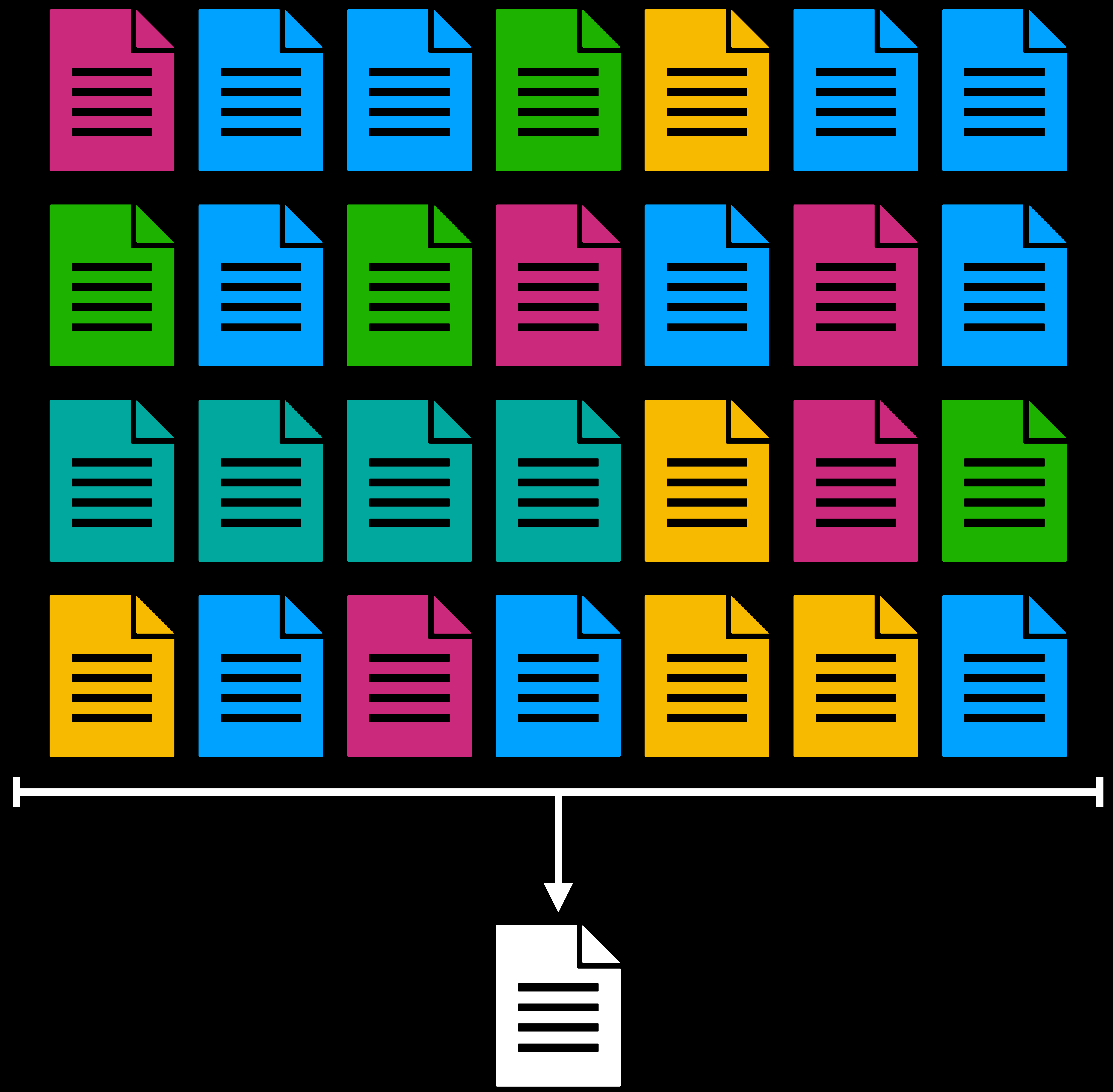
Technology:

LLM operating over all available notes

Some discussion of clinician-in-the-loop

Landscape:

Easy to build very hard to validate



Example Chart Summarization

ChatGPT

J. Doe, a 58-year-old male with a complex medical history, presented to the ED on 10/02/2025 with severe right upper quadrant (RUQ) pain that started 30 minutes after dinner. His medical background includes:

- **Type 2 Diabetes Mellitus**: Managed with a regimen that includes Metformin and a GLP-1 receptor agonist. His glycemic control has been generally good, with a recent HbA1c of 6.9%.
- **Hypertension**: Well-controlled on a regimen including Lisinopril and Metoprolol succinate. His blood pressure was 132/82 mmHg at the last primary care follow-up.
- **Hyperlipidemia**: Managed with Atorvastatin 80 mg nightly, achieving LDL levels within the target range.
- **Coronary Artery Disease**: History of NSTEMI managed with percutaneous coronary intervention (PCI) and stent placement, currently on dual antiplatelet therapy with Aspirin and Clopidogrel.
- **Chronic Kidney Disease (CKD) Stage 3**: Diagnosed around March 2025, likely secondary to his long-standing hypertension and diabetes. His management includes close monitoring of renal function and adjustments to medications to minimize nephrotoxicity. Recent serum creatinine stabilized at 1.7 mg/dL, with an eGFR of approximately 46 mL/min/1.73m^2.
- **Heart Failure with Preserved Ejection Fraction (HFpEF)**: Diagnosed in June 2025 following an admission for symptoms suggestive of volume overload, including shortness of breath and bilateral ankle swelling. His heart failure is managed with Furosemide, Lisinopril, and Metoprolol succinate, emphasizing sodium and fluid restriction.
- **Orthostatic Hypotension**: Episodes noted, particularly related to antihypertensive therapy adjustments.

In recent months, Mr. Doe has generally reported improvement in his symptoms, attributing mild fatigue to increased activity levels. He's been adhering to his medication regimen, dietary restrictions, and fluid intake recommendations, completing a cardiac rehabilitation program, and maintaining regular physical activity.

Given his RUQ pain, considerations might include gallbladder issues (e.g., cholelithiasis, cholecystitis) given the postprandial onset of symptoms, as well as possible hepatic or pancreatic involvement. His history of CKD and HFpEF, coupled

Clinical Foundation Models

Goals:

Build general models that can help us answer many different clinical questions

Operate over physiologic values & medical events instead of words

Technology:

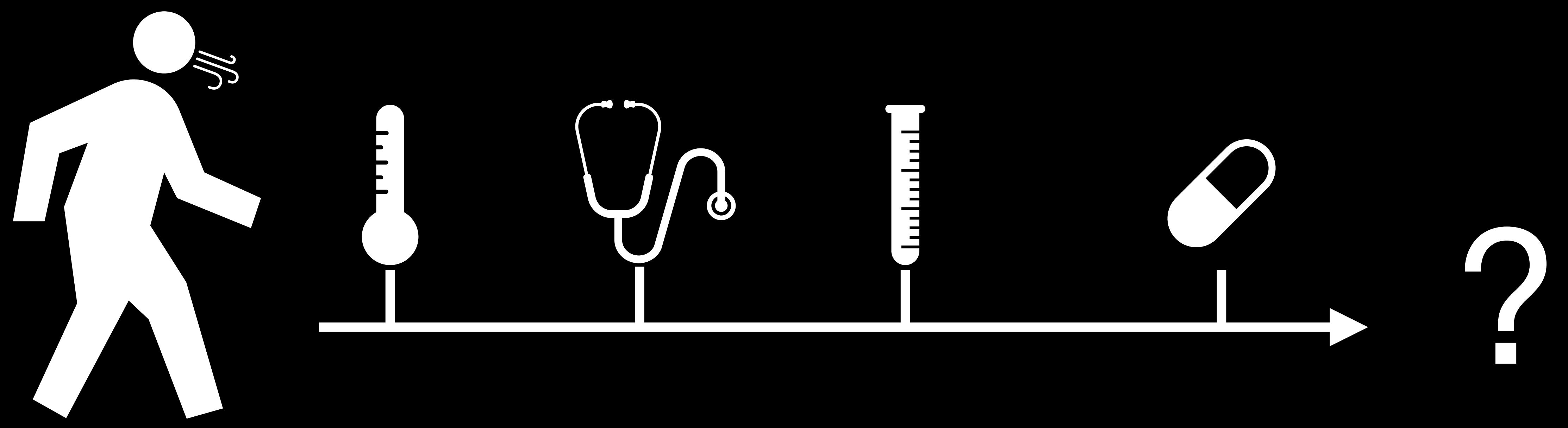
Transformer operating over all available EMR data

Landscape:

Hard to build hard to validate

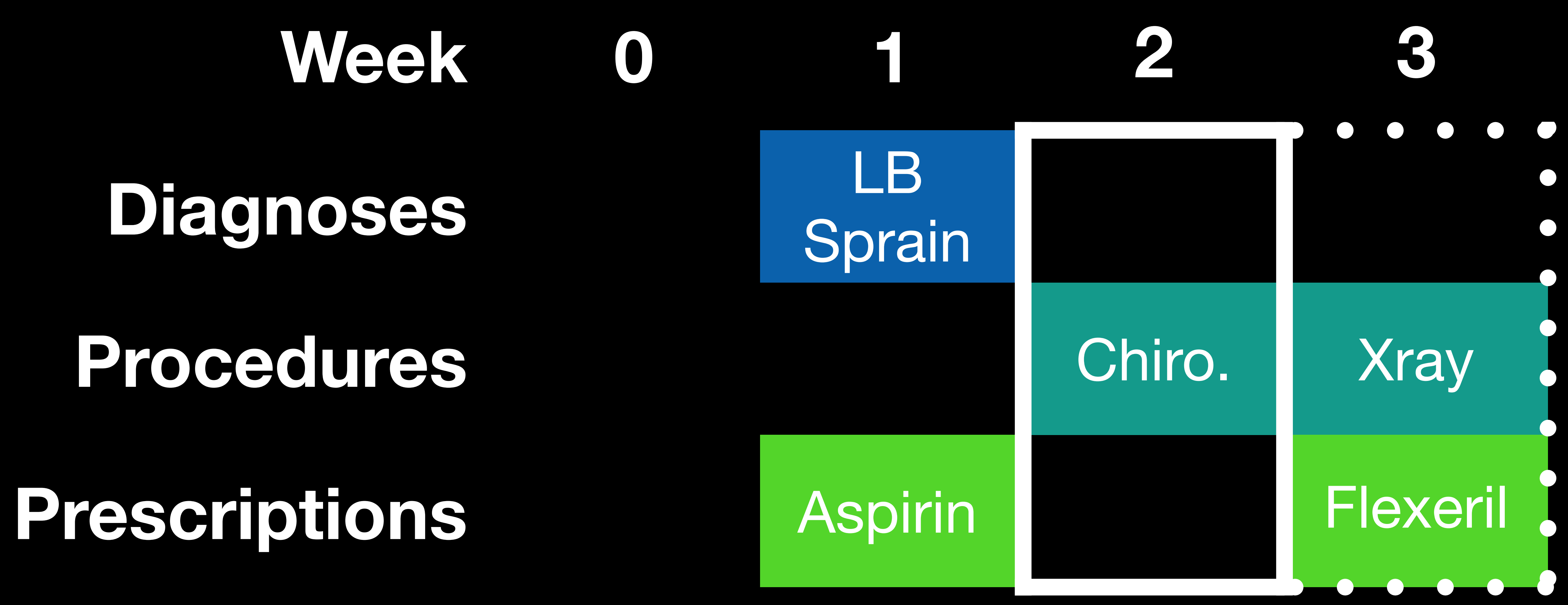
Natural Langauge

The quick brown fox jumps over … ?



Medical Events

Similar to Recovery Trajectory Generation



Doesn't involve repurposing another AI model. Need to make special medical AI models.

Resource intensive

- Computationally expensive
- Massive data needs
- Specialized engineering skills

Need to keep in mind

- Privacy
- Bias
- Interprebility
- Maintenance

Takeaways

Generative AI is special case of AI

- Having a general understanding of AI aids in understanding generative AI
- Models can be used in both a generative and predictive sense

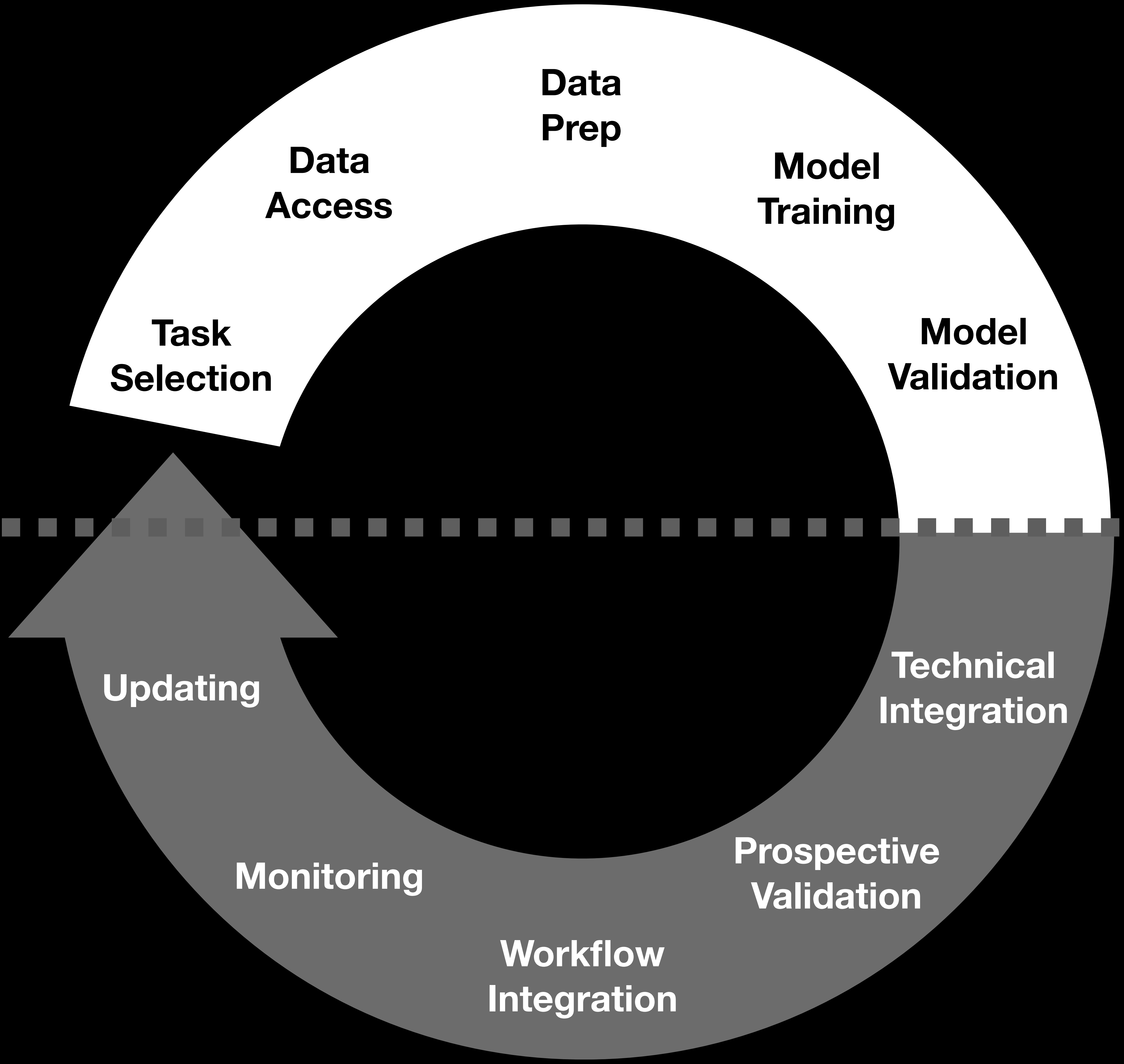
Evaluation is critical in AI

- Generative AI is harder to evaluate because the larger amount of use cases
- Population biases may be harder to detect

Most of these tools are still in development

Development

Creation of models



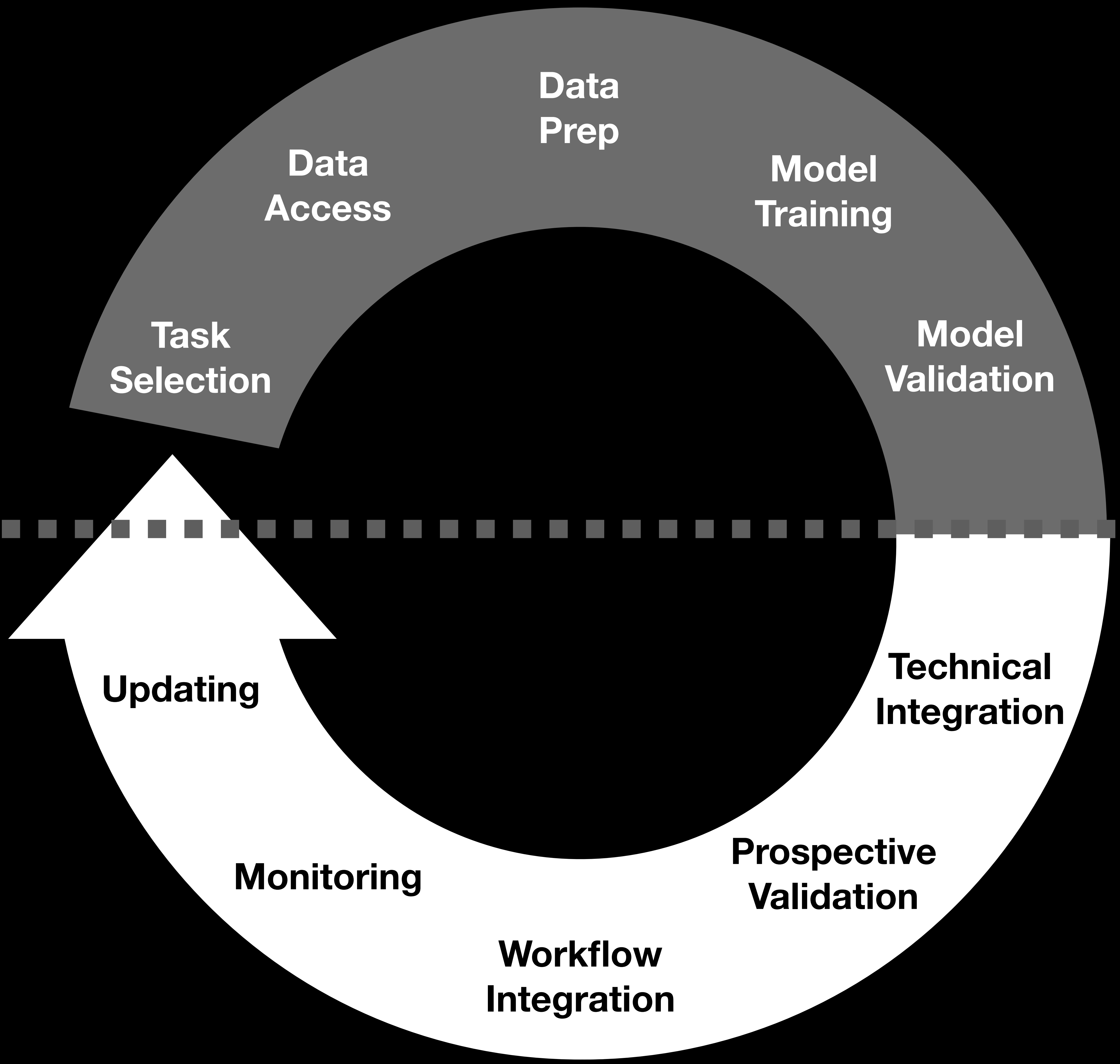
Integration into care

Implementation

Physicians need to drive the implementation

Development

Creation of models



Integration into care

Implementation

Questions?

Comments? Concerns? Discussion.

Erkin Ötleş
Twitter: @eotles
eotles.com
eotles@umich.edu