Updating Clinical Risk Stratification Models Using Rank-Based Compatibility

Evaluating & Optimizing Clinician-Model Team Performance

INFORMS Healthcare 2023
Erkin Ötleş, Jenna Wiens, Brian T. Denton
July 2022

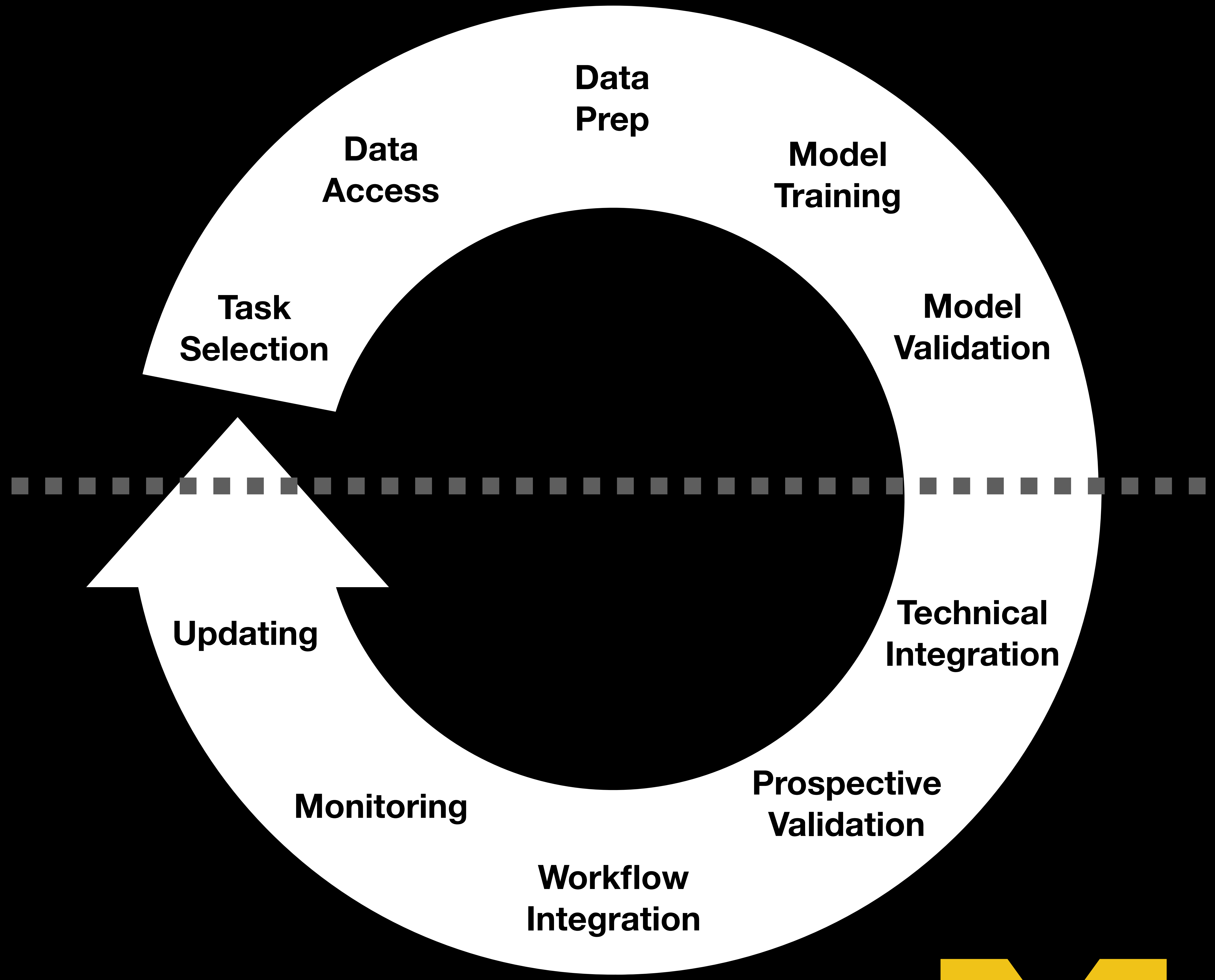
Hello, INFORMS!

Medical Scientist Training Program Fellow

- MD: x2023
- Engineering PhD: 2022
- Healthcare ML Dev & Implementation

Previously:

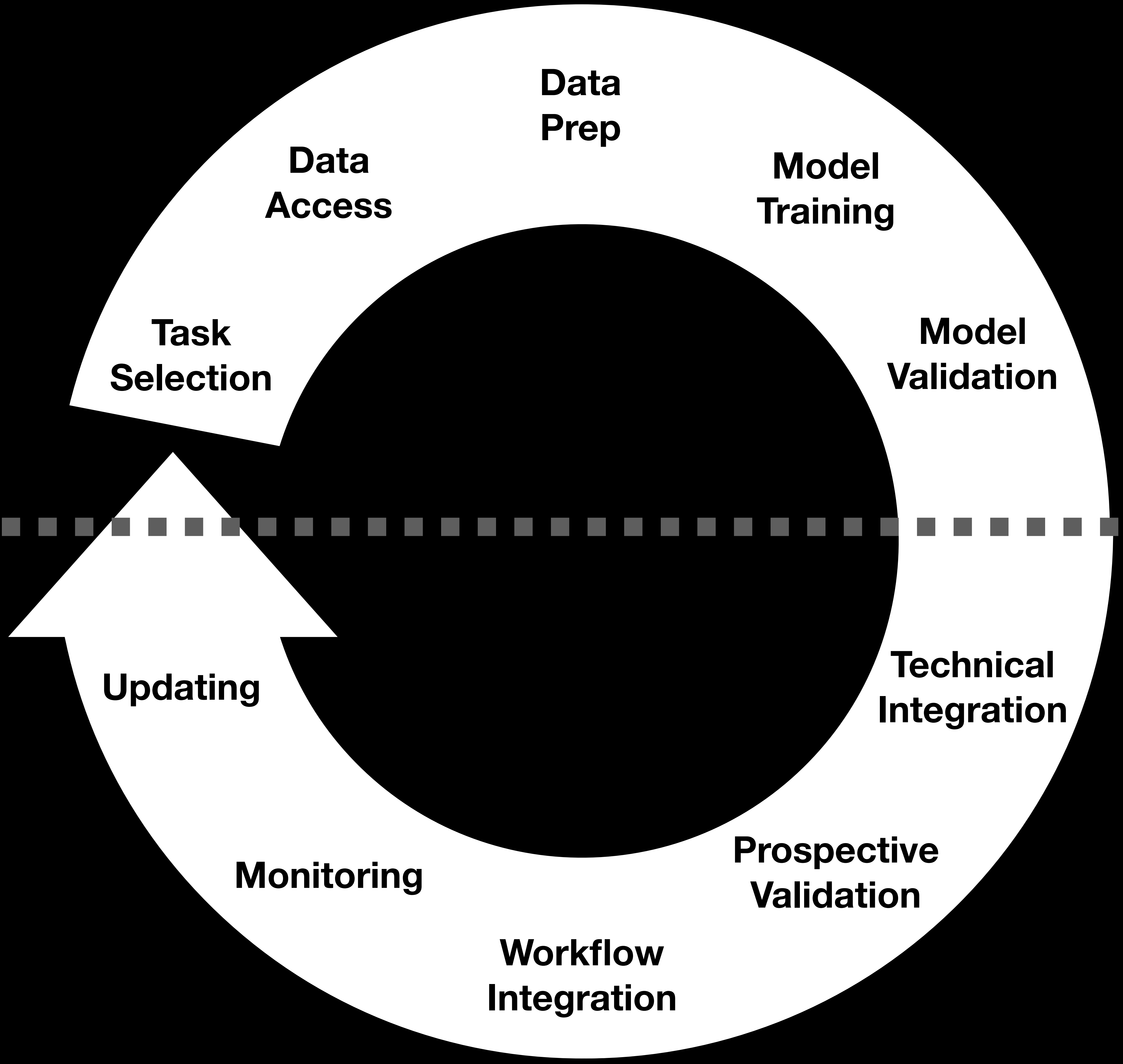
- Healthcare Data & Decision Science Manager
- Epic Ambulatory Solutions Engineer



Development & Implementation Experience Grounded in Clinical & Technical Knowledge.

Development

Creation of models



Integration into care

Implementation

Creation & Validation of Models Addressing Clinical Needs.

Development

Task Selection

Use clinical needs to drive technical foci

Data Access

Deep understanding of EHR & Claims data

Data Prep

Developed tools to automate data transformation

Model Training

Extensive experience building ML & AI models

Model Validation

Widely recognized validation studies

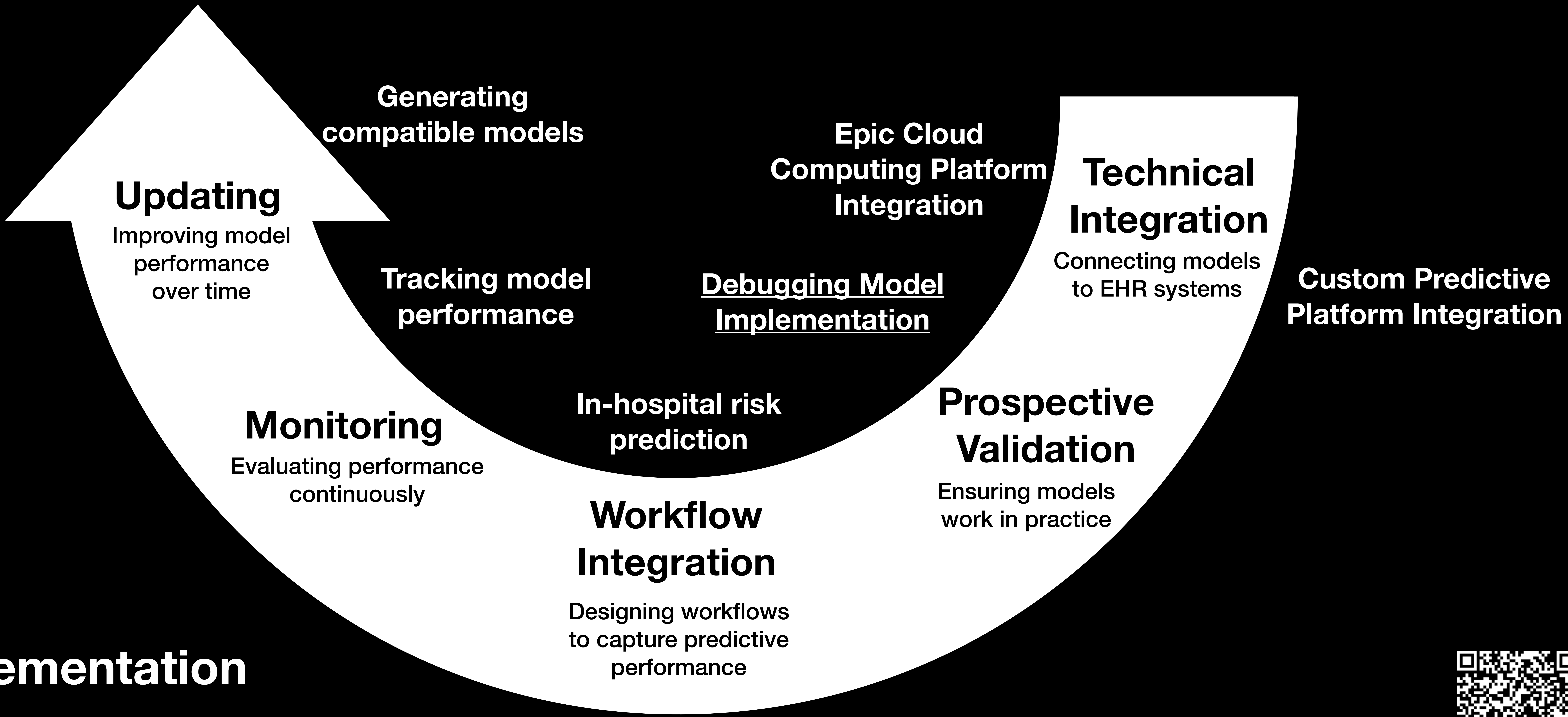
AI Health-State Prediction Patent

Data Preparation Tools

Covid Model BMJ

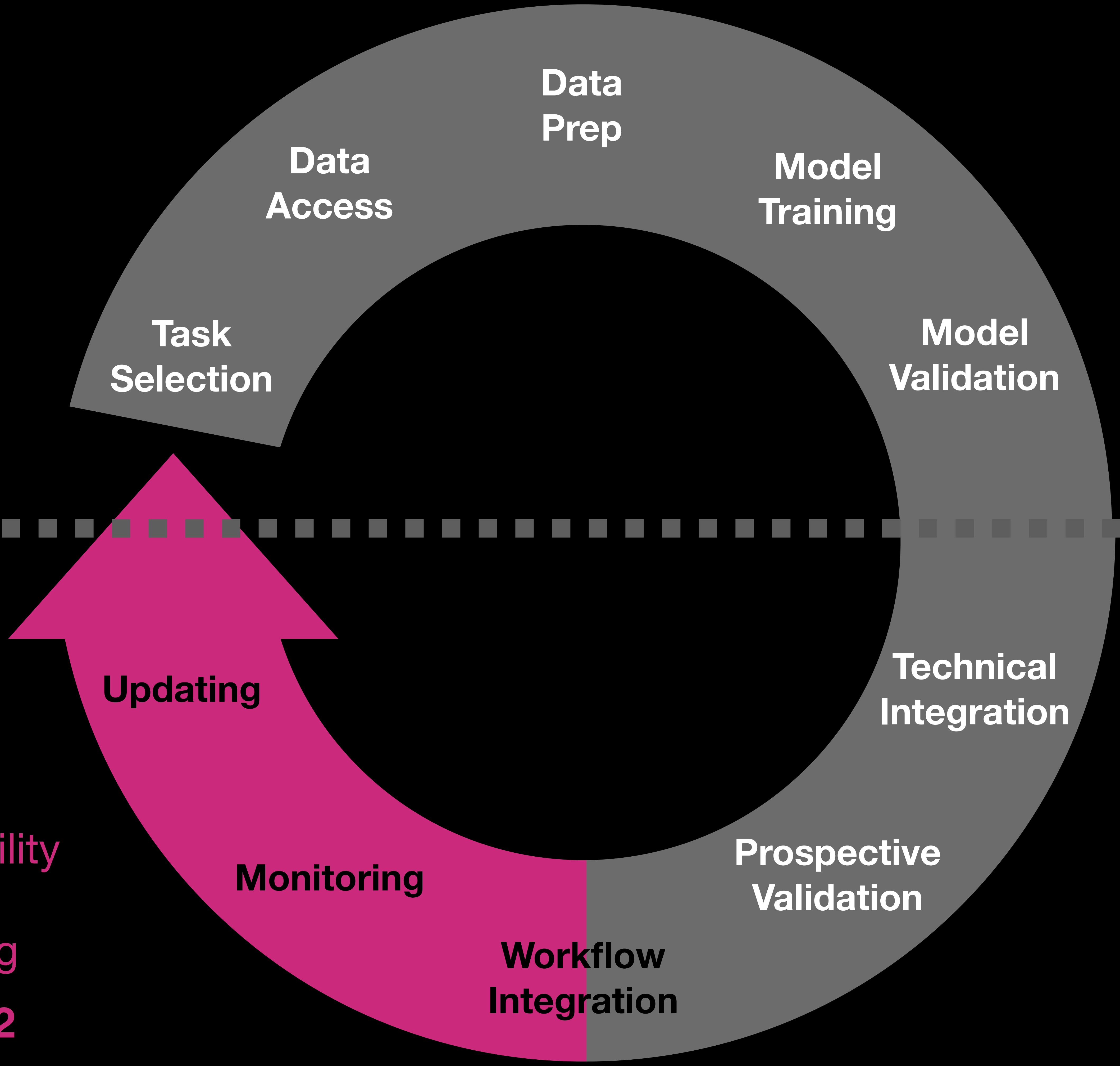
Epic Sepsis Model Validation

Implementation Focused on Bridging Workflow & Technical Considerations.



Implementation

Last but not least.



Chapter 2

Assess value of
longitudinal observations
for return to work prediction

Ötleş *et al.*, JAMIA '22

Chapter 3

Characterize changes
to prospective performance
after model implementation

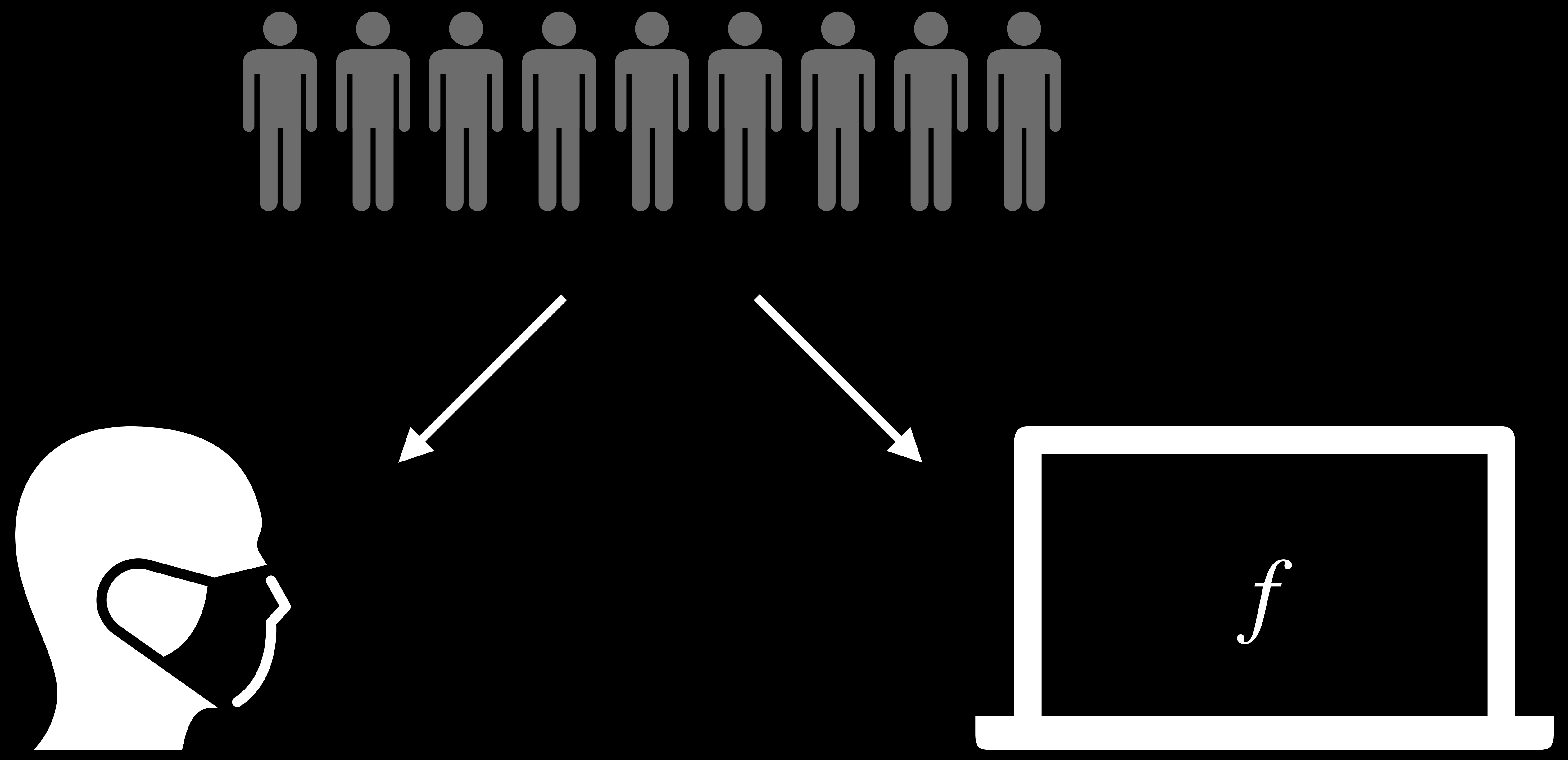
Ötleş *et al.*, MLHC '21

Chapter 4

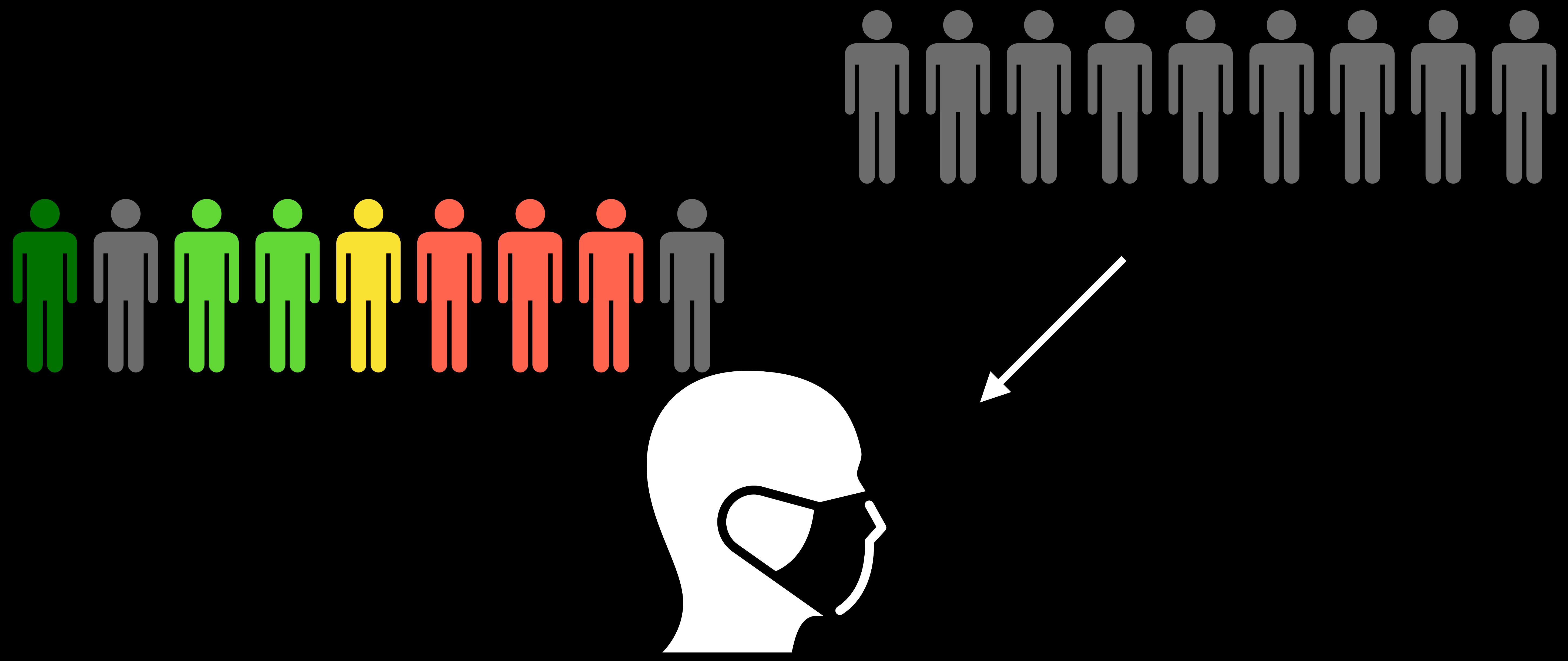
Develop rank-based compatibility
measurement & optimization
approaches for model updating

Ötleş *et al.*, In submission '22

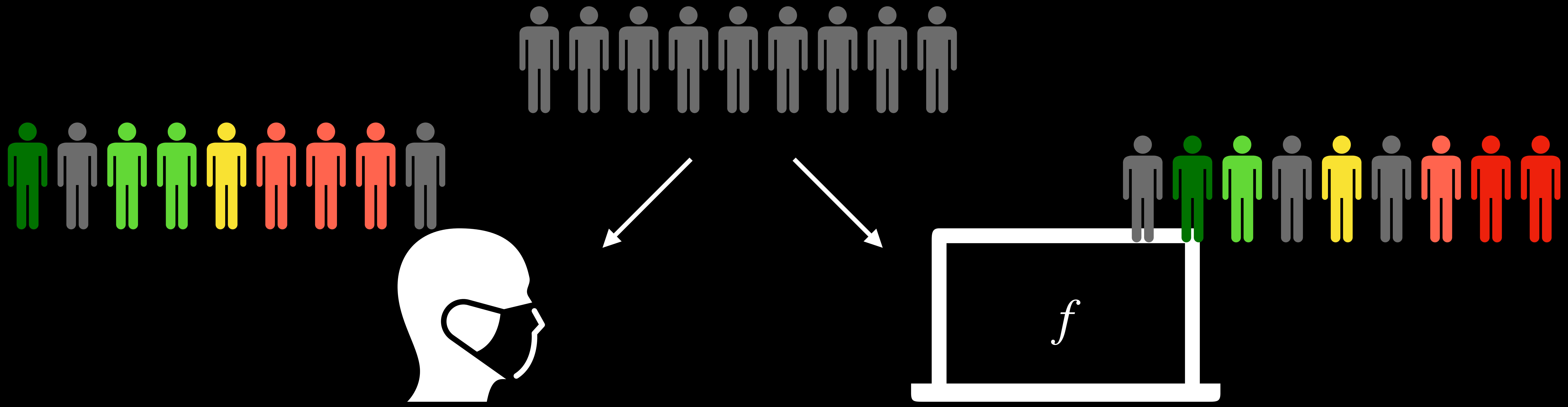
Physicians and models function as a team in healthcare settings.



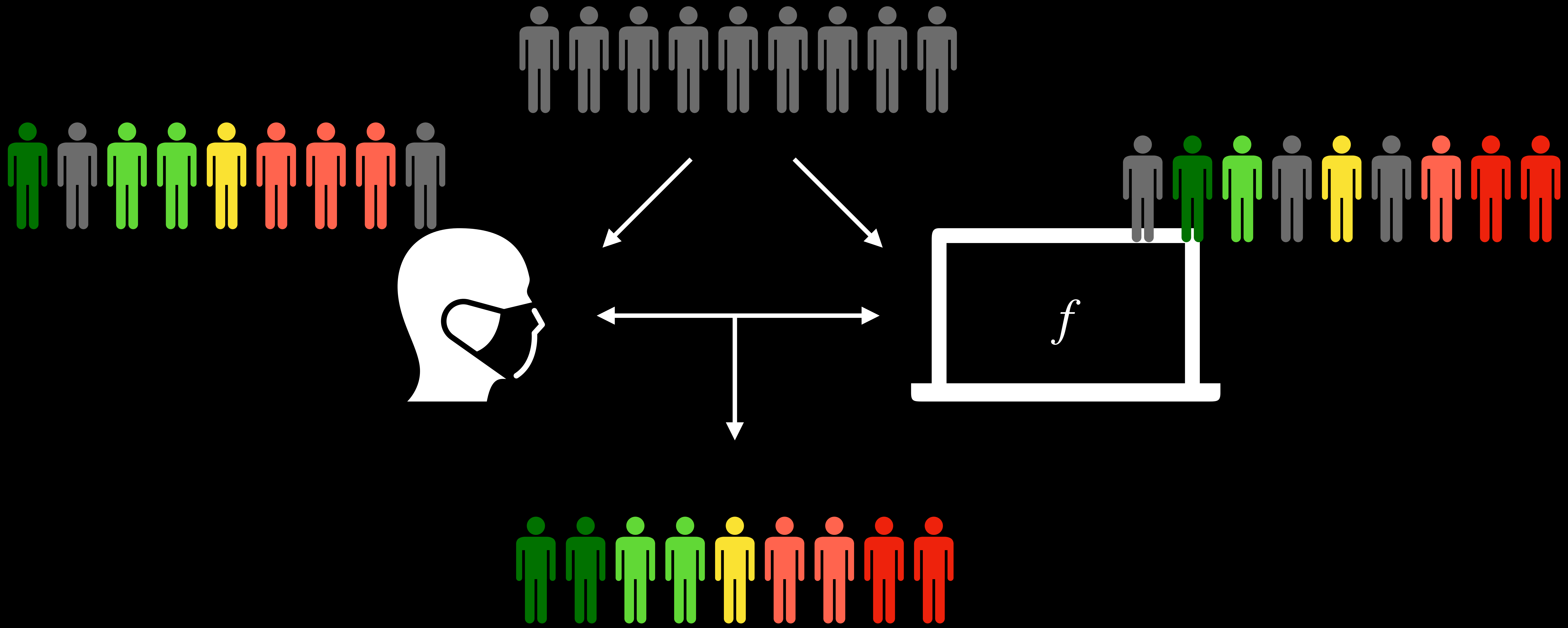
Physicians and models function as a team in healthcare settings.



Physicians and models function as a team in healthcare settings.



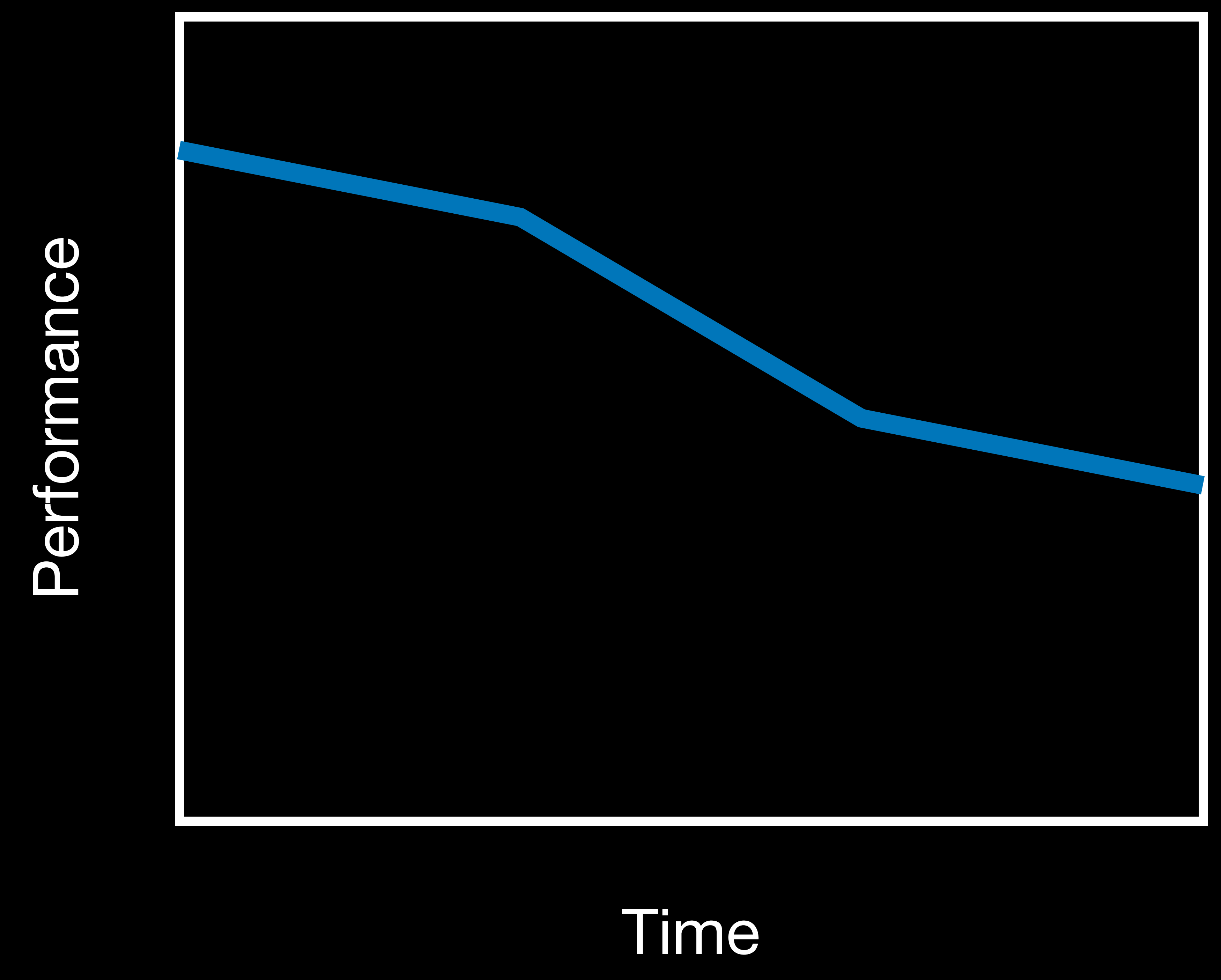
Physicians and models function as a team in healthcare settings.



This is complicated because we need to update models over time.

This is complicated because we need to update models over time.

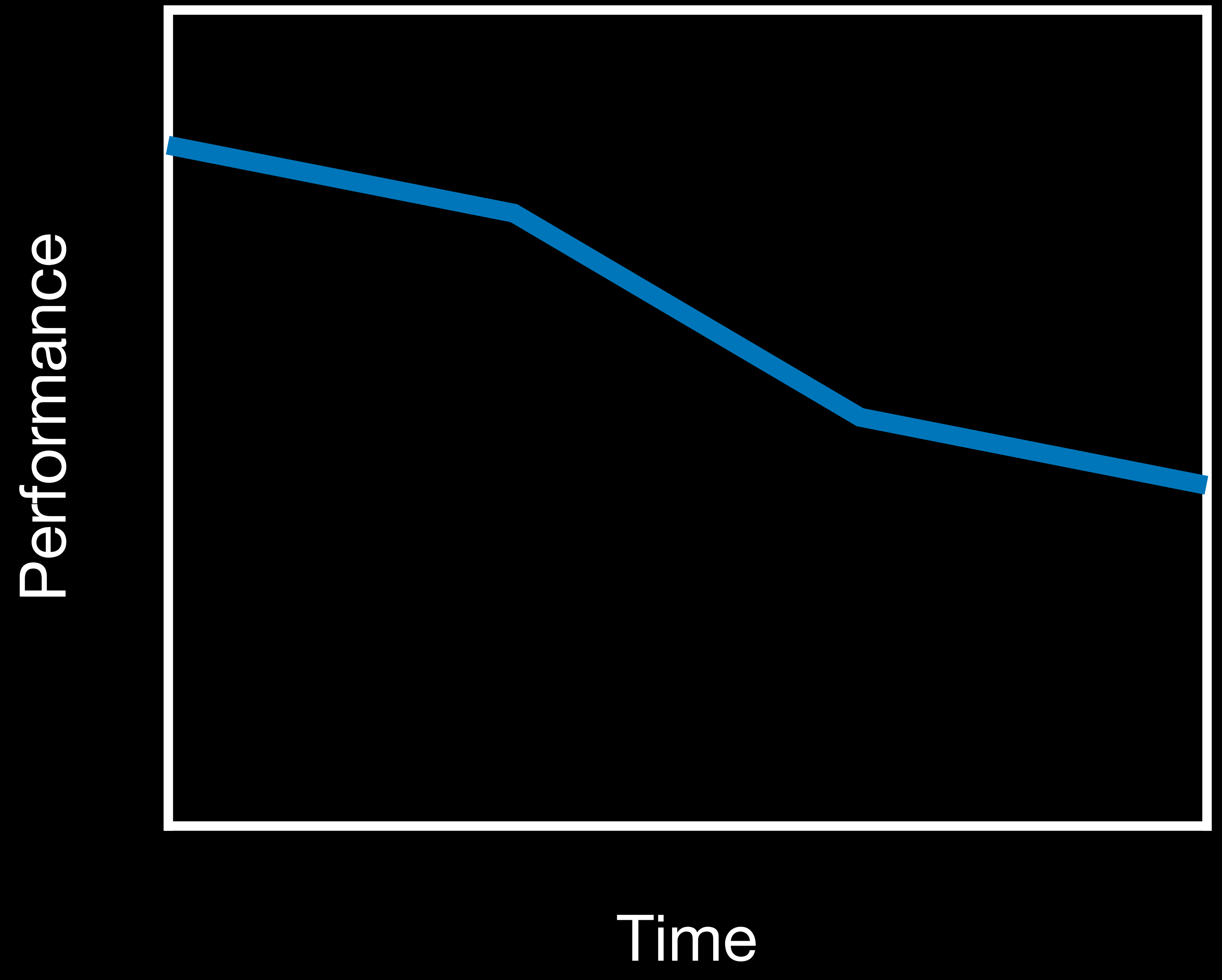
Counter performance degradation



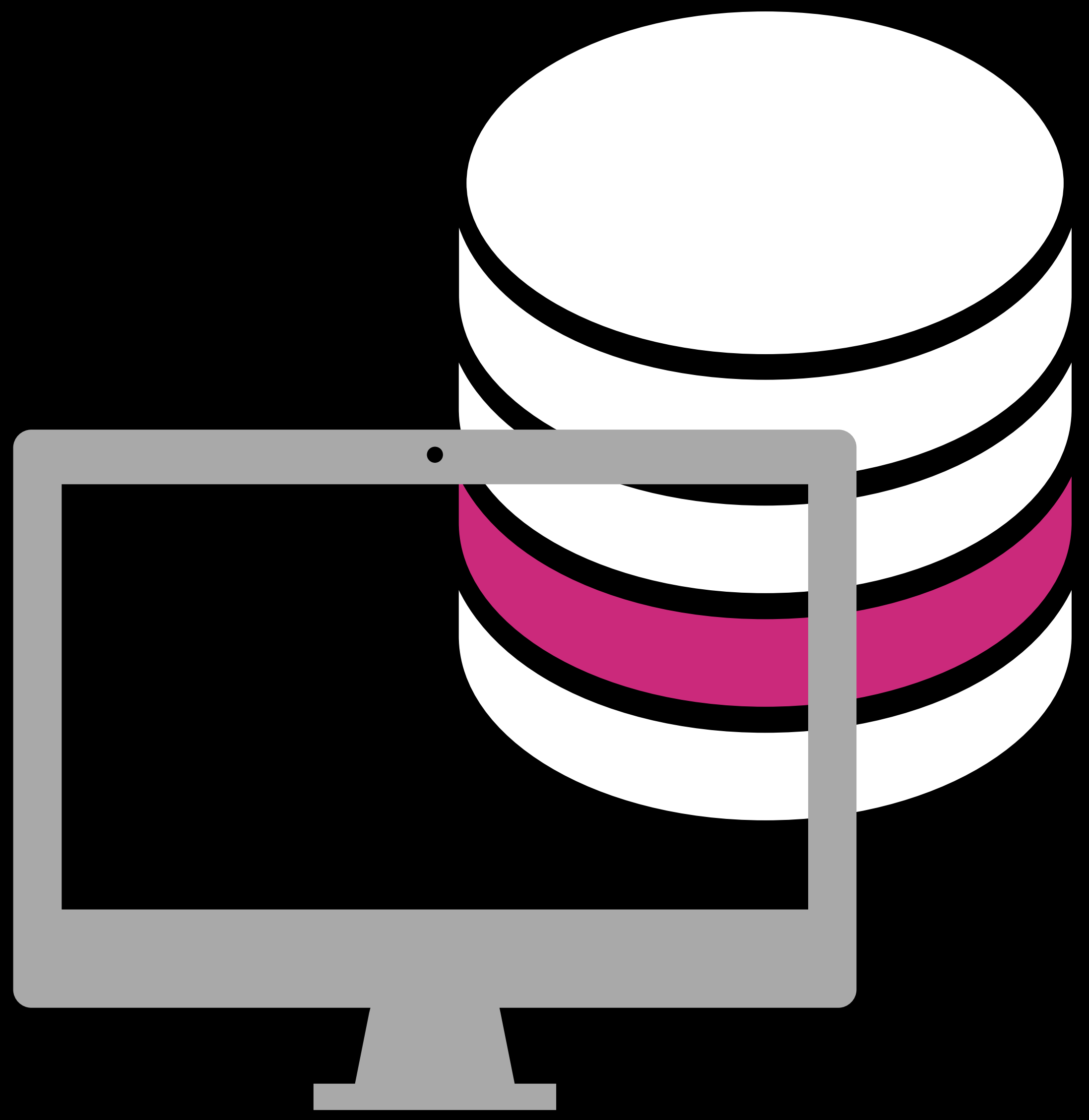
Finlayson 2021, David 2017, Hickey 2012, Minne 2012, Bansal 2019

This is complicated because we need to update models over time.

Counter performance degradation

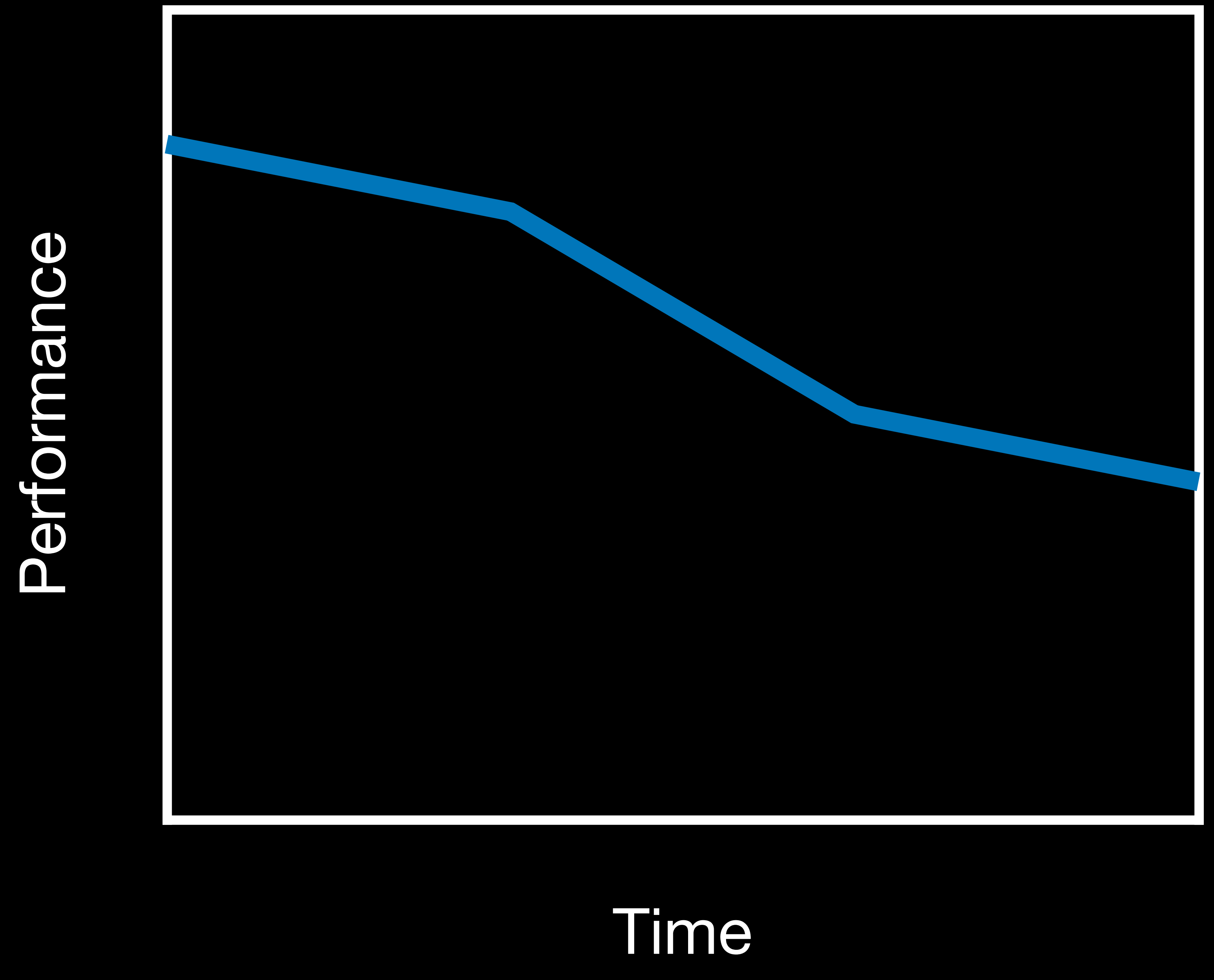


Adapt to infrastructure changes

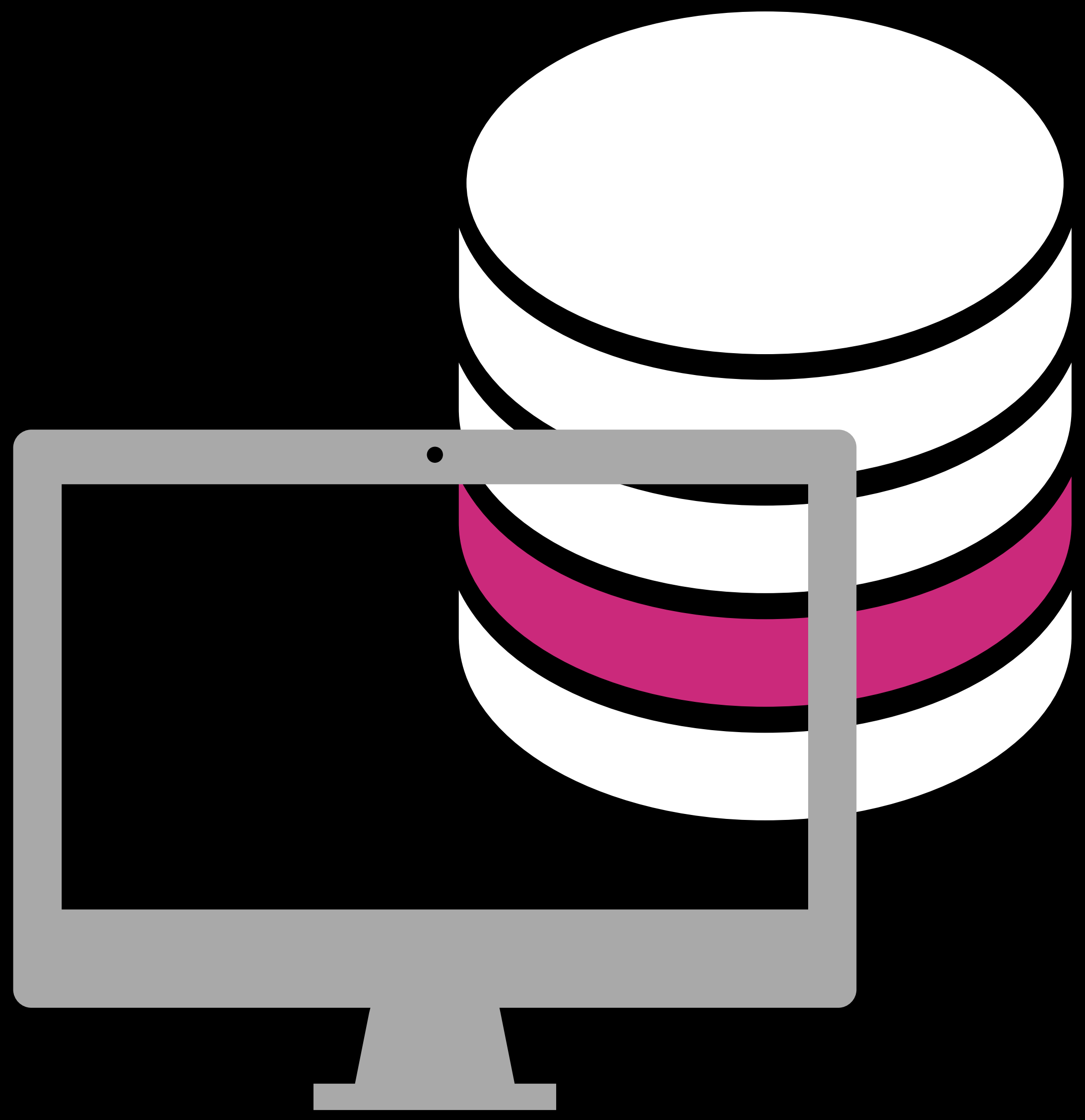


This is complicated because we need to update models over time.

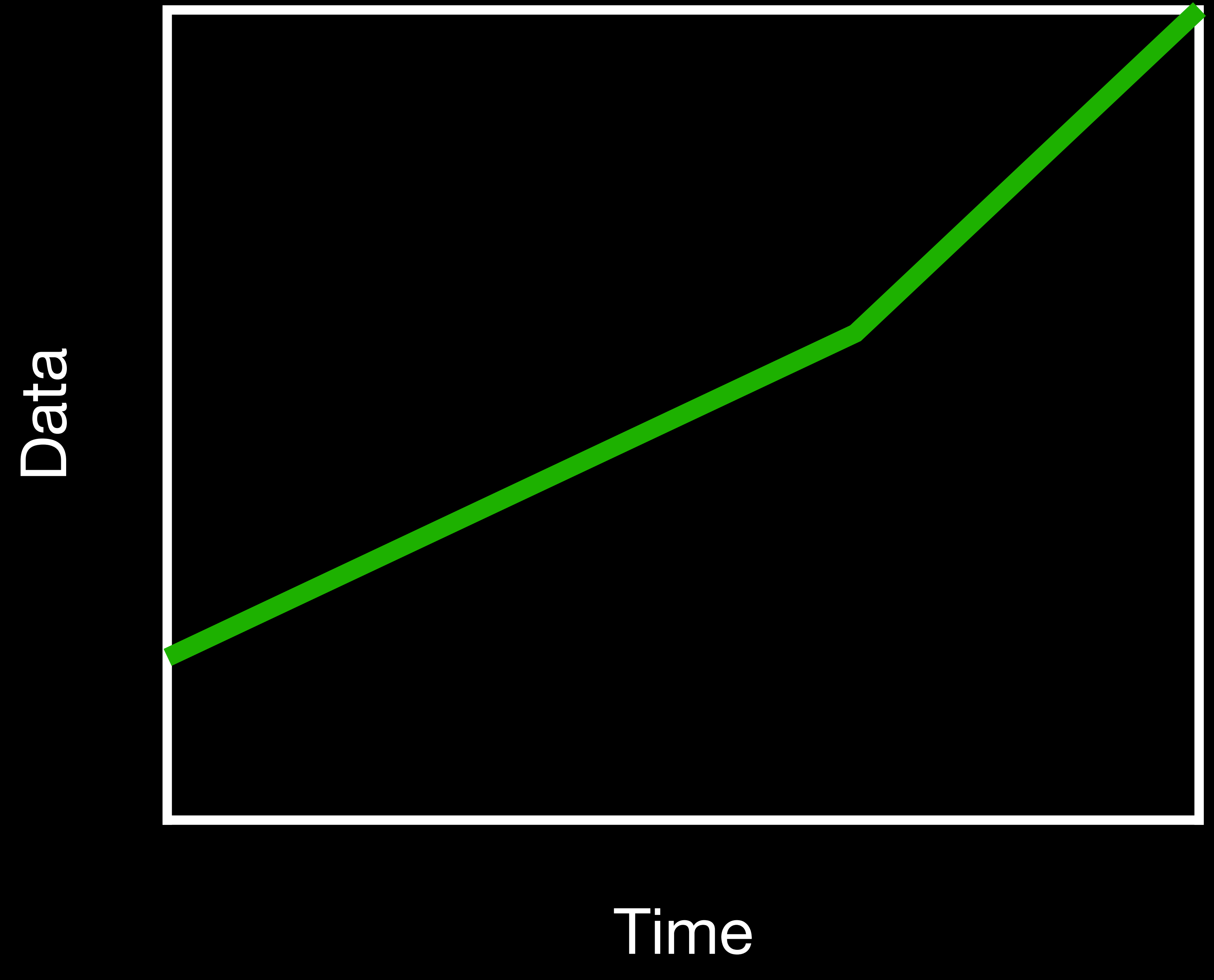
Counter performance degradation



Adapt to infrastructure changes

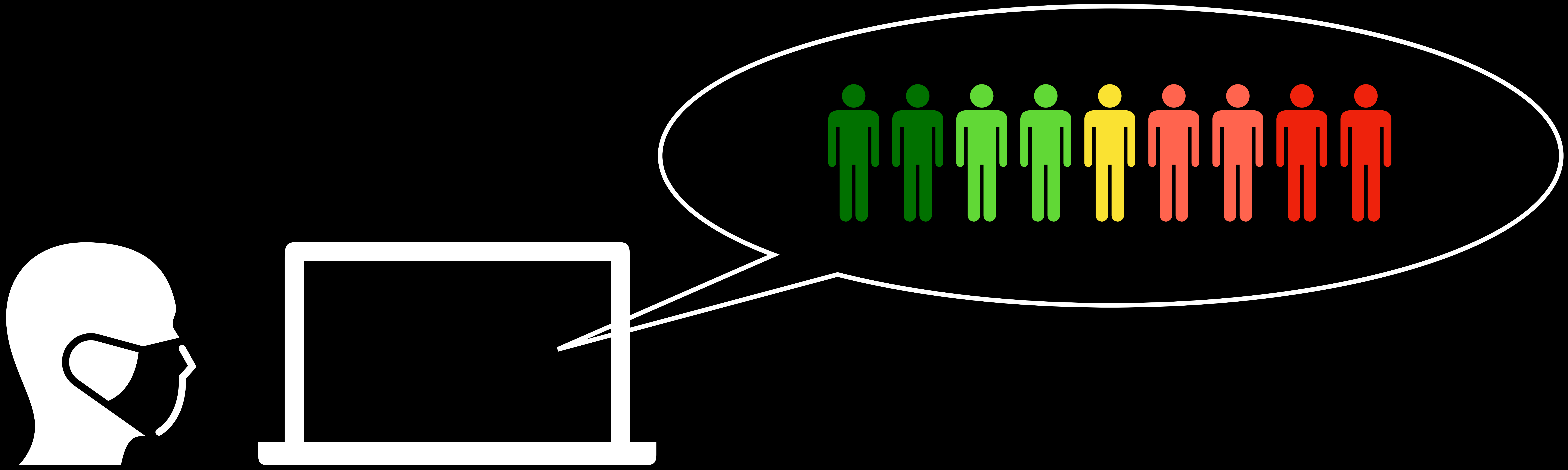


Incorporate new data

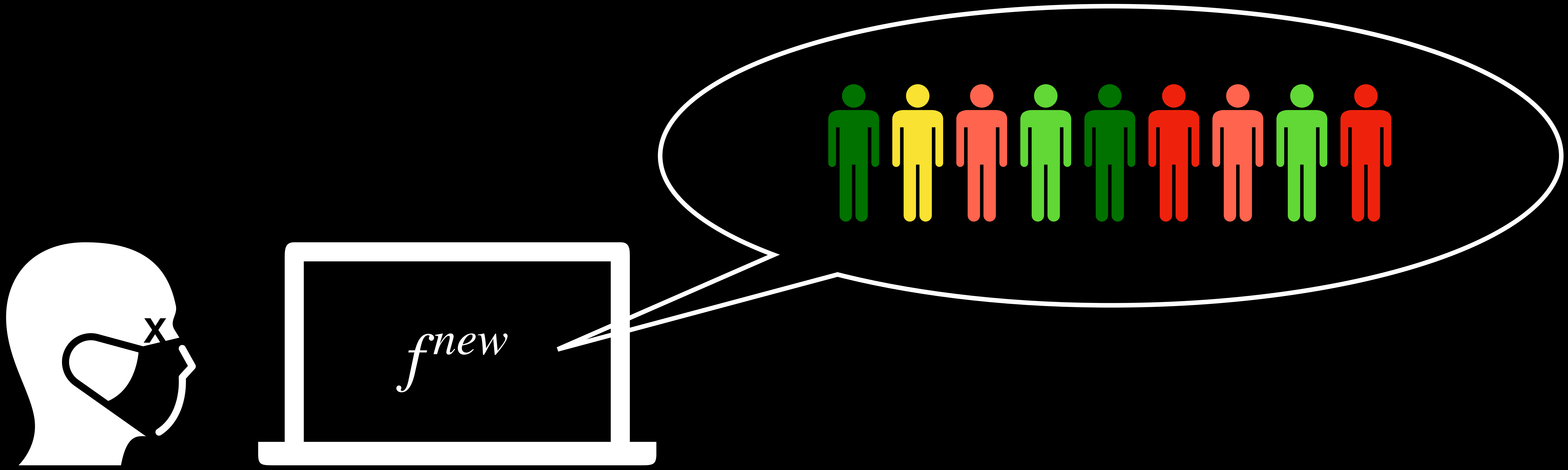


Finlayson 2021, David 2017, Hickey 2012, Minne 2012, Bansal 2019

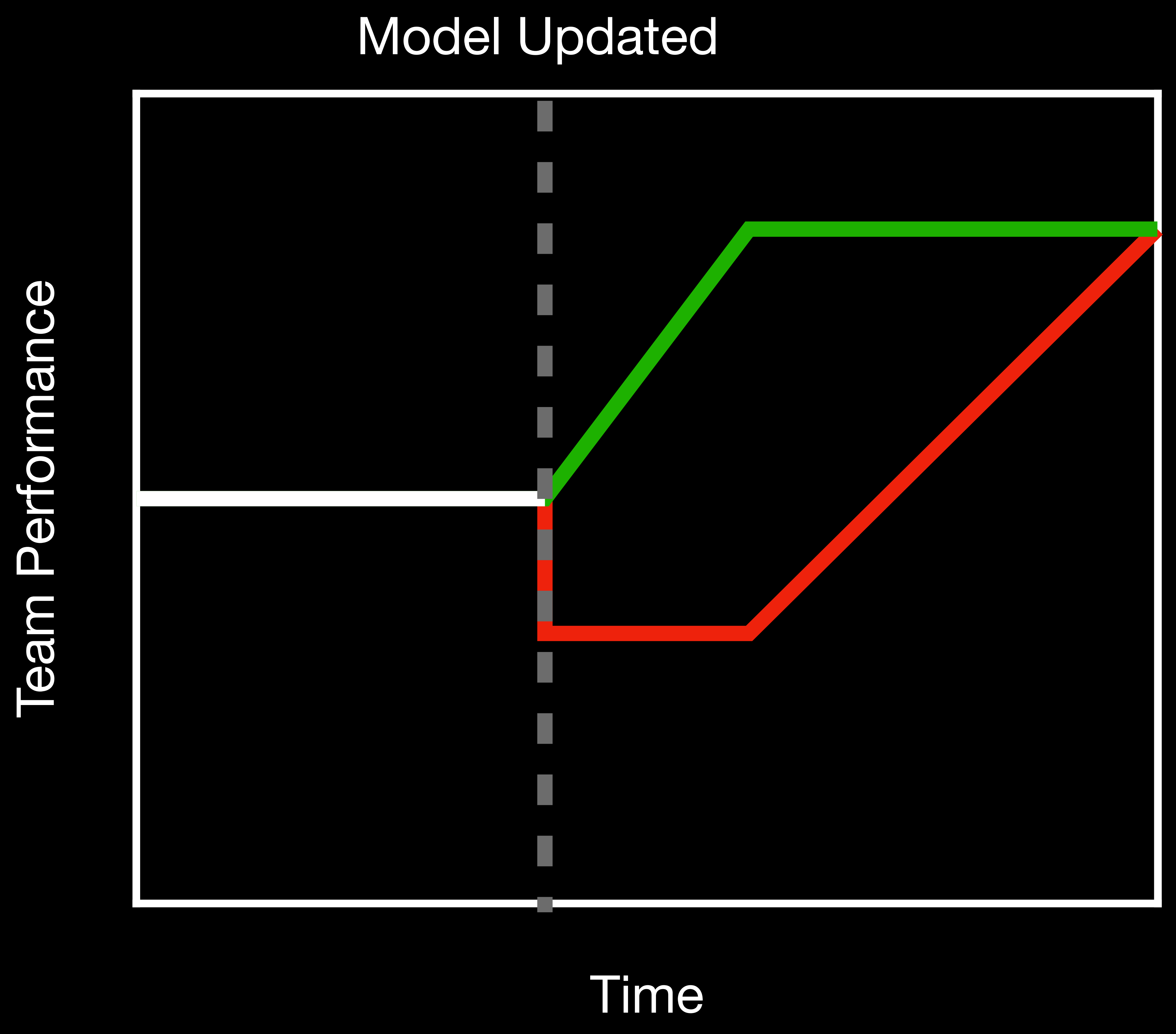
Updates can mess with user expectations.



Updates can mess with user expectations.



Team performance may suffer if models don't meet user expectations.



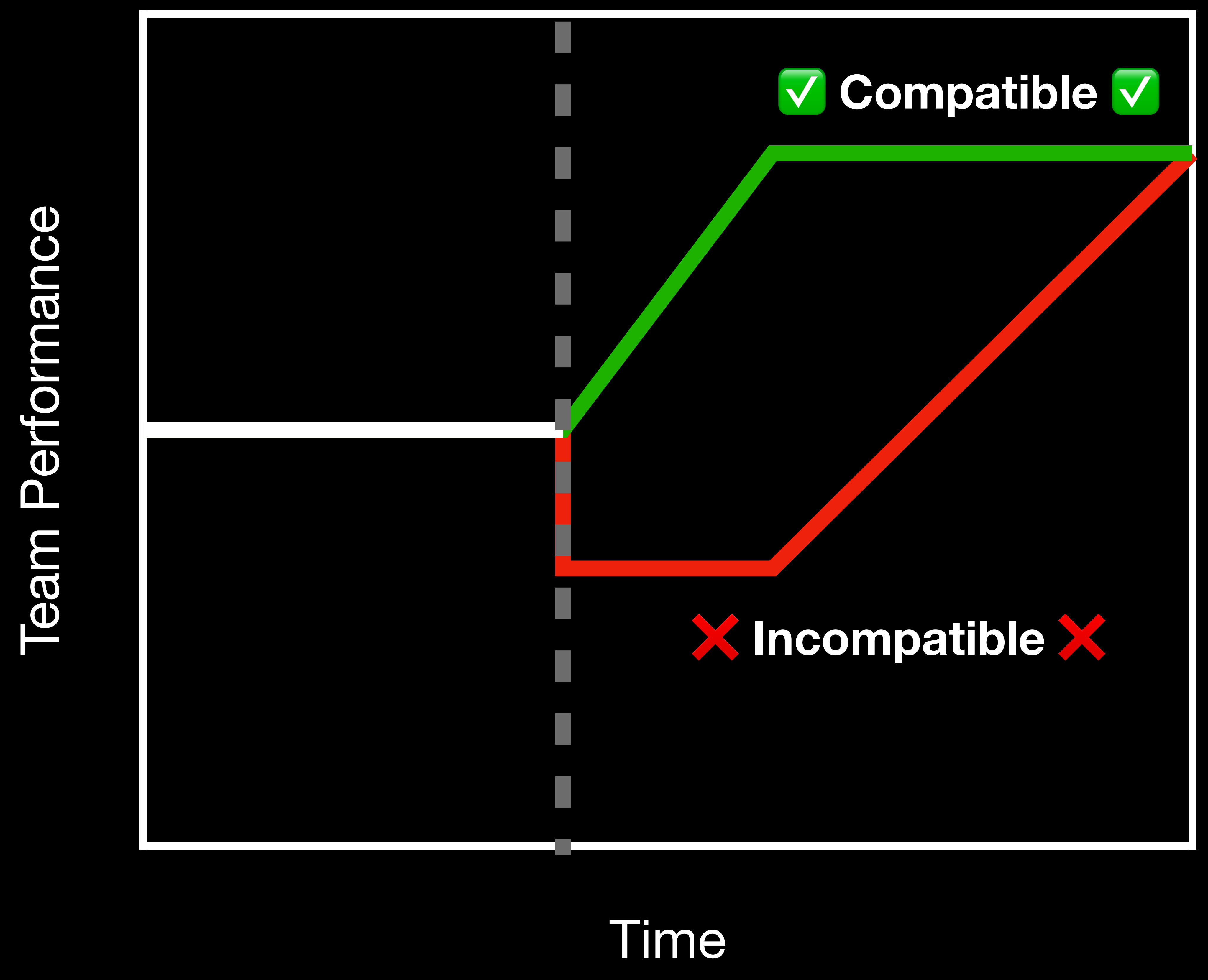
Adapted from Bansal 2019

Ideally updated models meet the expectations of users

Compatibility: the amount an updated model continues the correct behavior of an original model

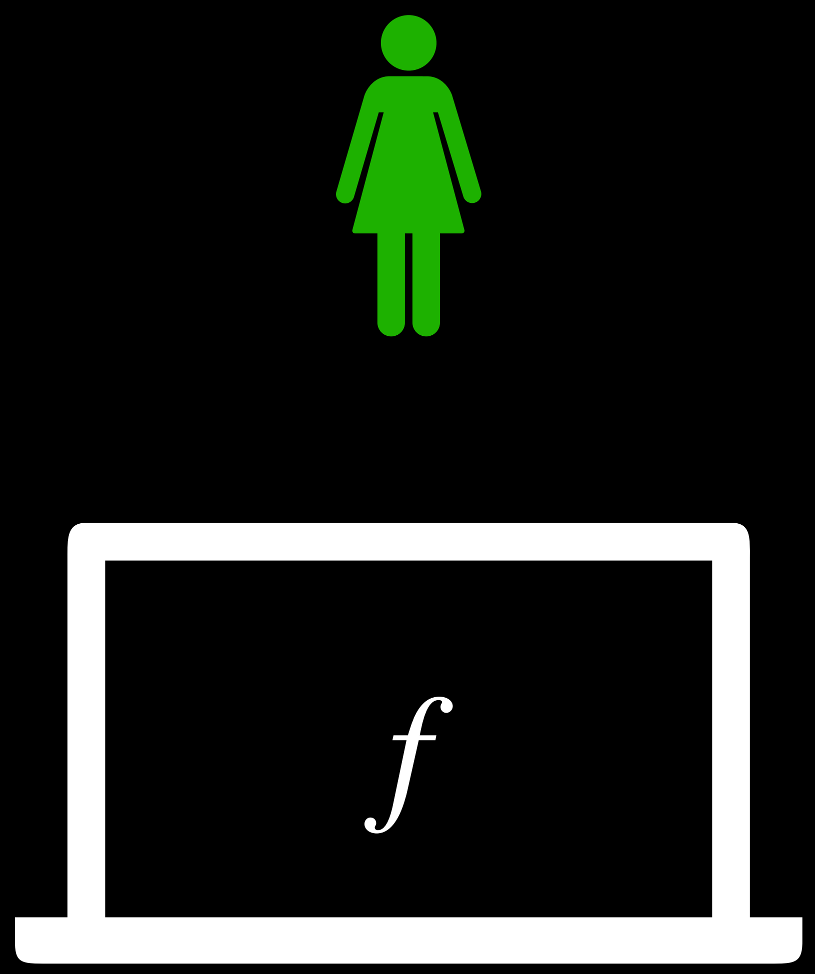
Way to measure user expectations

Goal: updated models should have high compatibility

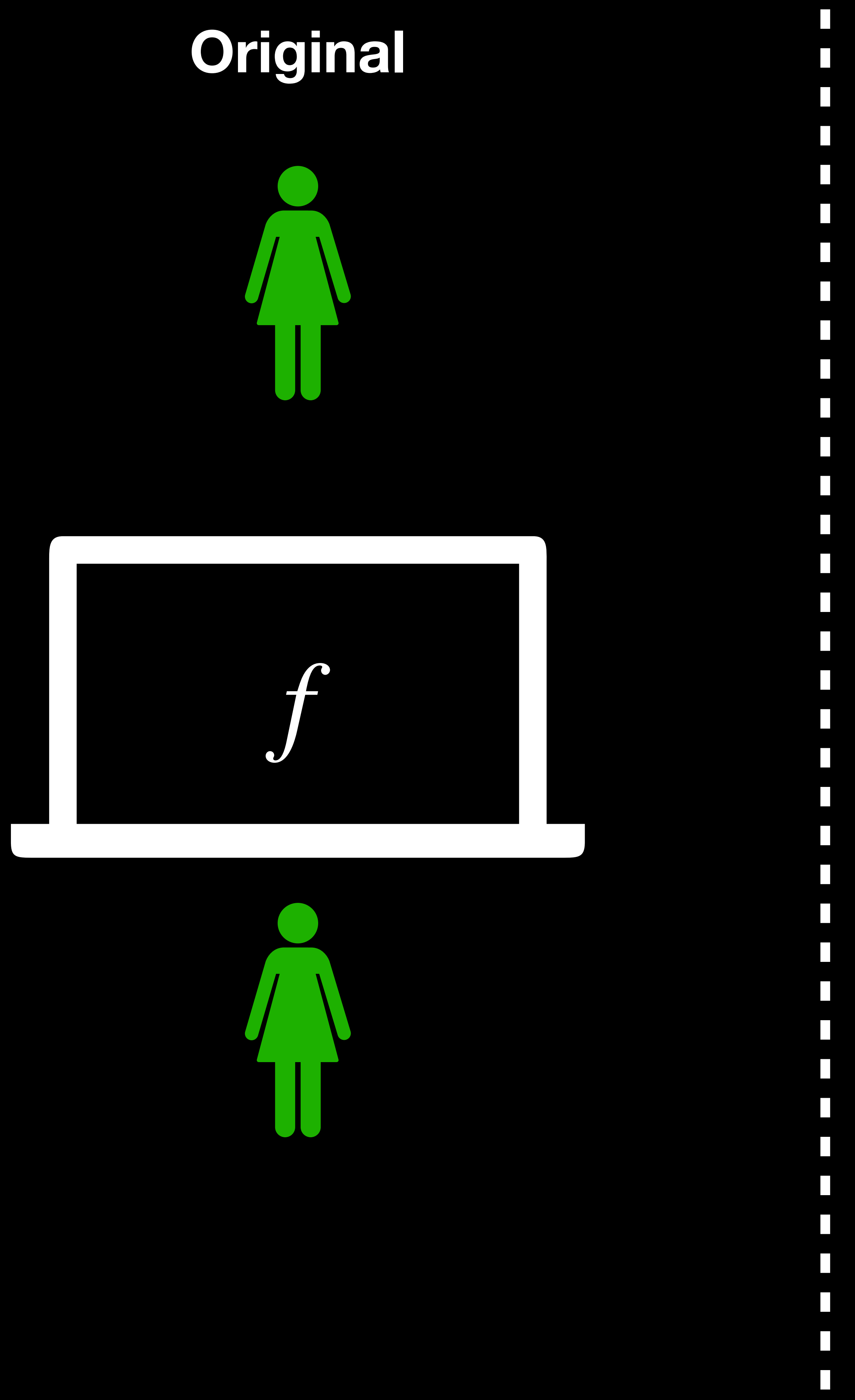


Compatibility can be assessed by using the original and updated models for the same predictive task.

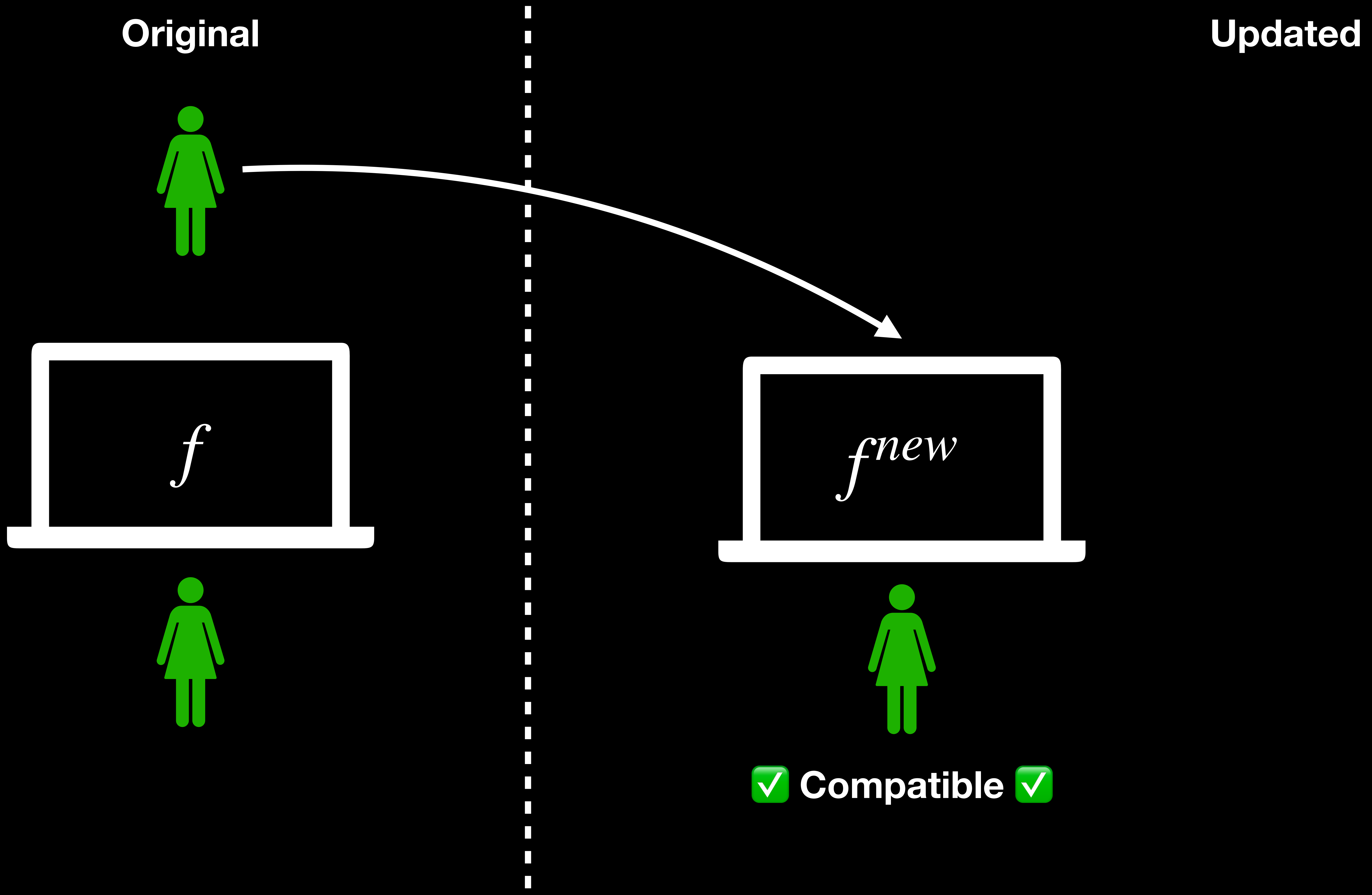
Original



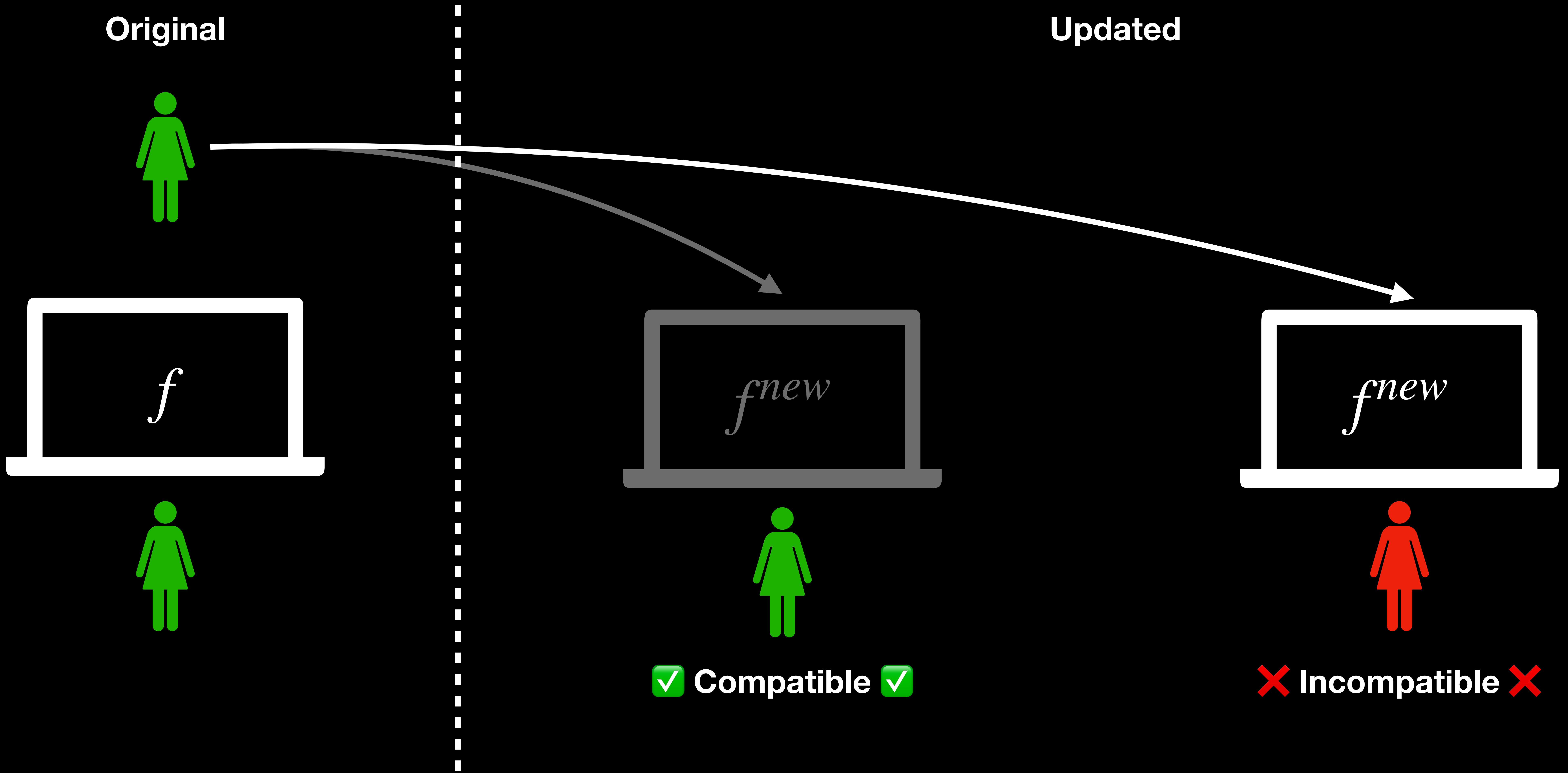
Compatibility can be assessed by using the original and updated models for the same predictive task.



Compatibility can be assessed by using the original and updated models for the same predictive task.



Compatibility can be assessed by using the original and updated models for the same predictive task.



Backwards Trust Compatibility C^{BT}

The chance that the updated model's labels are correct, given that the original model's labels were correct.

$$C^{BT}(f^o, f^u) = \frac{\text{\# patients both models label correctly}}{\text{\# patients original model labels correctly}}$$

Bansal 2019

Backwards Trust Compatibility C^{BT}

The chance that the updated model's labels are correct, given that the original model's labels were correct.

$$C^{BT}(f^o, f^u) = \frac{\text{\# patients both models label correctly}}{\text{\# patients original model labels correctly}}$$

original model **updated model**

Backwards Trust Compatibility C^{BT}

The chance that the updated model's labels are correct, given that the original model's labels were correct.

$$C^{BT}(f^o, f^u) = \frac{\text{\# patients both models label correctly}}{\text{\# patients original model labels correctly}}$$

Backwards Trust Compatibility C^{BT}

The chance that the updated model's labels are correct, given that the original model's labels were correct.

$$C^{BT}(f^o, f^u) = \frac{\text{\# patients both models label correctly}}{\text{\# patients original model labels correctly}}$$

1 → perfect compatibility, 0 → perfect incompatibility

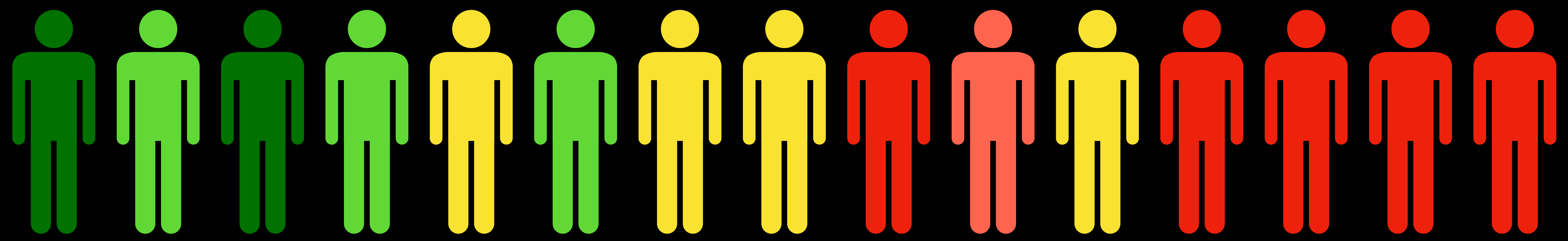
Bansal 2019

Problems with existing compatibility measures.

Existing measure depends on equality comparison

Problematic for use in risk stratification model & healthcare settings

- Depends on setting a single decision threshold

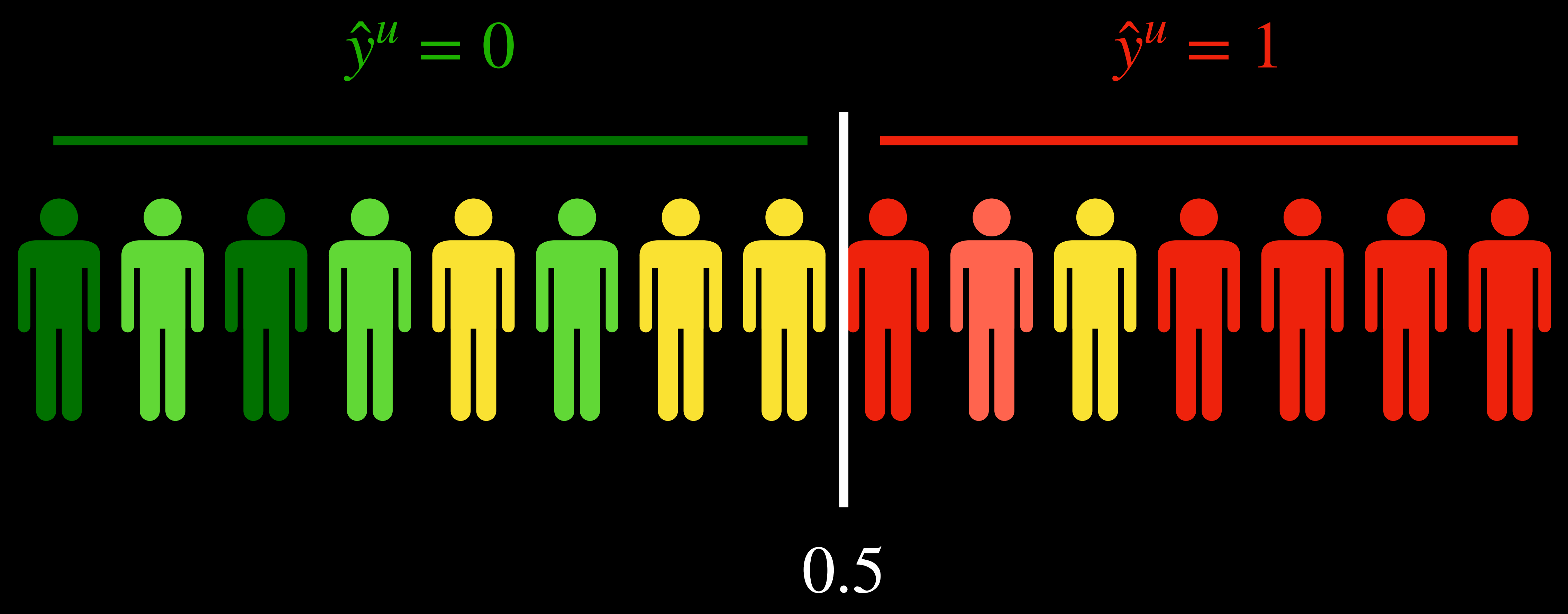


Problems with existing compatibility measures.

Existing measure depends on equality comparison

Problematic for use in risk stratification model & healthcare settings

Depends on setting a single decision threshold

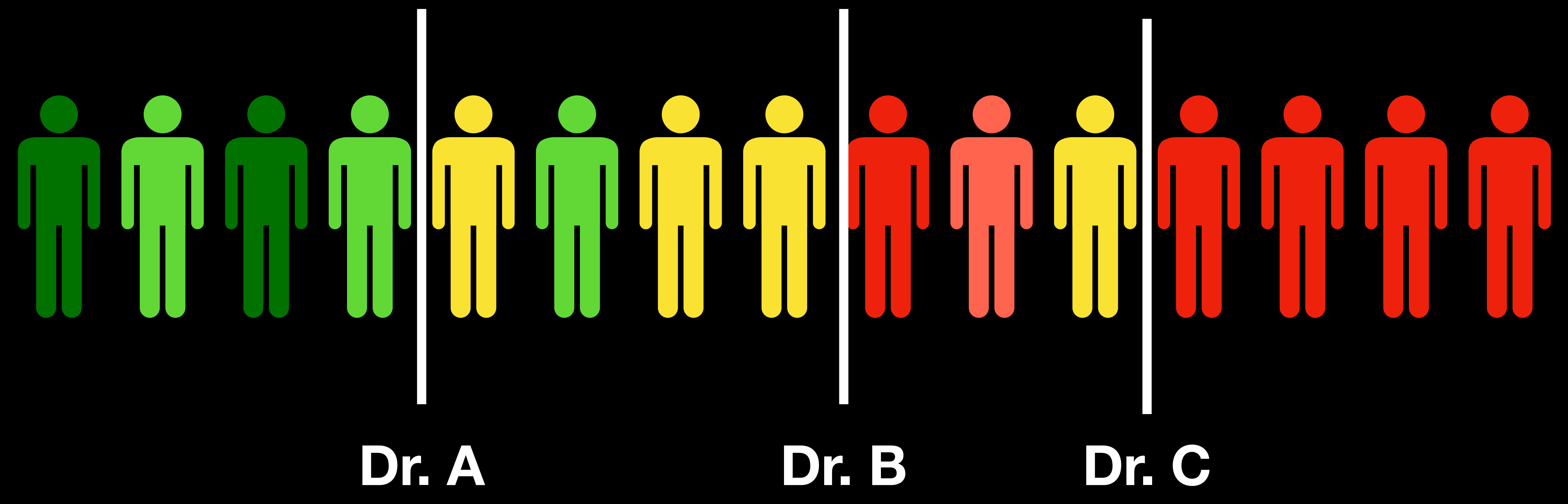


Problems with existing compatibility measures.

Existing measure depends on equality comparison

Problematic for use in risk stratification model & healthcare settings

- Depends on setting a single decision threshold

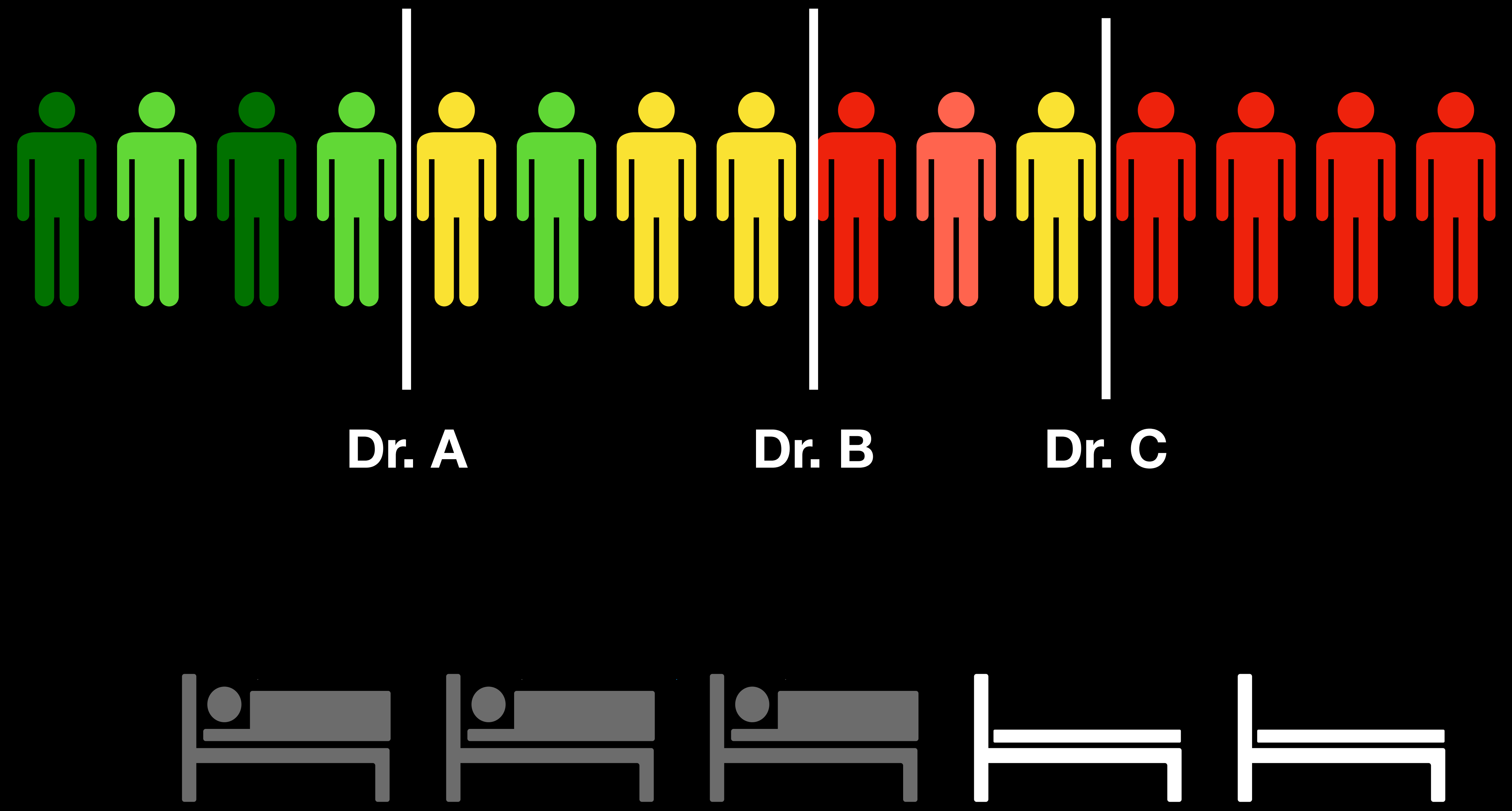


Problems with existing compatibility measures.

Existing measure depends on equality comparison

Problematic for use in risk stratification model & healthcare settings

- Depends on setting a single decision threshold



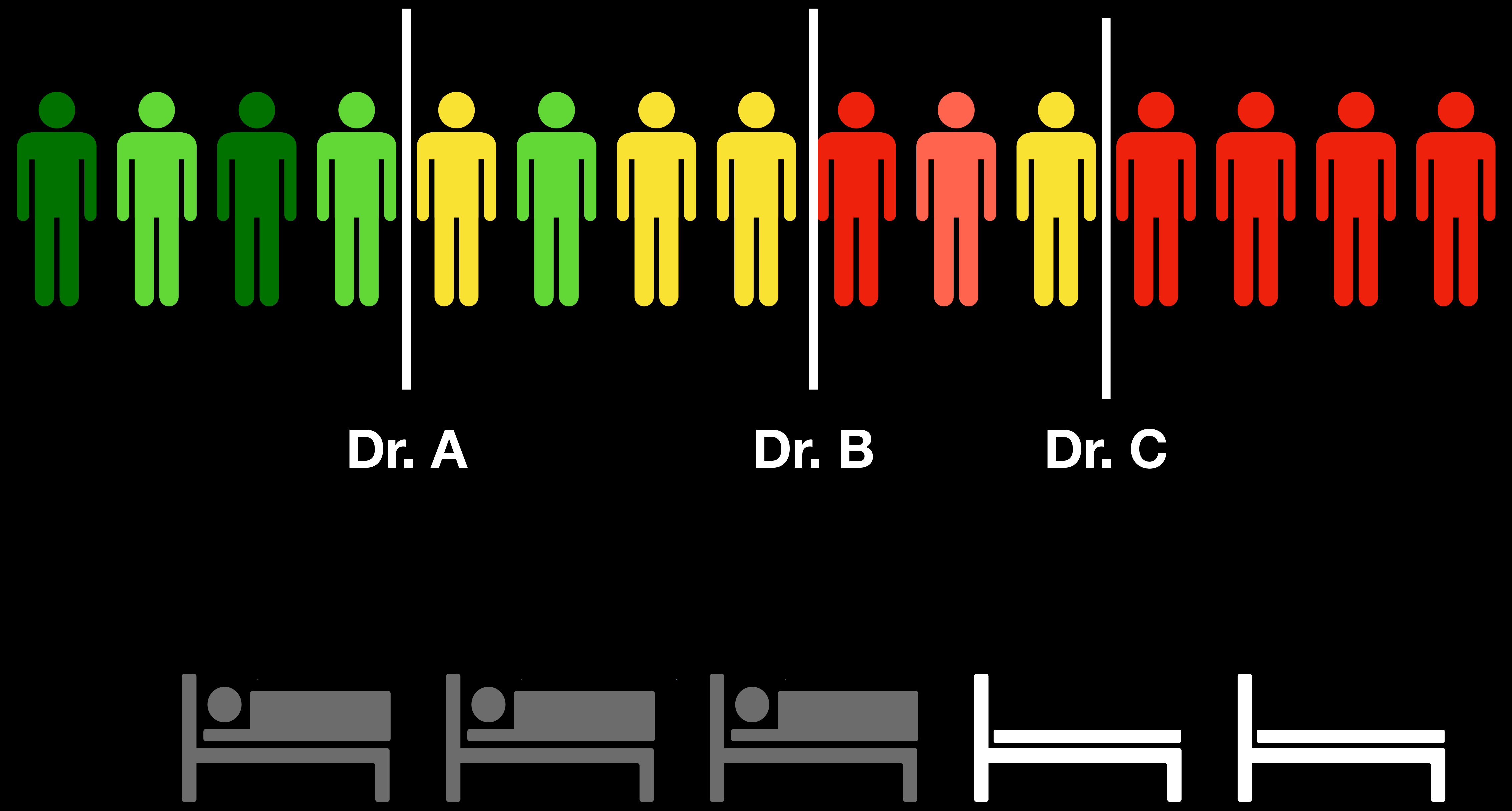
Problems with existing compatibility measures.

Existing measure depends on equality comparison

Problematic for use in risk stratification model & healthcare settings

- Depends on setting a single decision threshold

- No direct relationship with AUROC



Our contributions

Define a new rank-based compatibility measure (C^R)

Characterize C^R and its relationship with AUROC

Custom loss function to engineer model updates with improved C^R

Intuition: C^R should inherit from both C^{BT} & AUROC

$$C^{BT}(f^o, f^u) = \frac{\text{\# patients both models label correctly}}{\text{\# patients original model labels correctly}}$$

$$AUROC(f^o) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o)}{m}$$

Evaluate correct behavior of both models
Normalized based on original model's behavior

"Correctness" based on risk estimate ordering

Rank-based compatibility C^R

Agreement of risk estimate rankings produced by original & updated models given original ranked correctly:

$$C^R(f^o, f^u) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o) \cdot \mathbf{1}(\hat{p}_i^u < \hat{p}_j^u)}{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o)}$$

Rank-based compatibility C^R

Agreement of risk estimate rankings produced by original & updated models given original ranked correctly:

$$C^R(\boxed{f^o}, \boxed{f^u}) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o) \cdot \mathbf{1}(\hat{p}_i^u < \hat{p}_j^u)}{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o)}$$

original model **updated model**

Rank-based compatibility C^R

Agreement of risk estimate rankings produced by original & updated models given original ranked correctly:

$$C^R(f^o, f^u) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o) \cdot \mathbf{1}(\hat{p}_i^u < \hat{p}_j^u)}{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o)}$$

original model ranks correctly

Rank-based compatibility C^R

Agreement of risk estimate rankings produced by original & updated models given original ranked correctly:

$$C^R(f^o, f^u) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o) \cdot \mathbf{1}(\hat{p}_i^u < \hat{p}_j^u)}{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o)}$$

updated model ranks correctly

original model ranks correctly

Rank-based compatibility C^R

Agreement of risk estimate rankings produced by original & updated models given original ranked correctly:

$$C^R(f^o, f^u) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o) \cdot \mathbf{1}(\hat{p}_i^u < \hat{p}_j^u)}{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o)}$$

both models rank correctly

original model ranks correctly

Rank-based compatibility C^R

Agreement of risk estimate rankings produced by original & updated models given original ranked correctly:

$$C^R(f^o, f^u) = \frac{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o) \cdot \mathbf{1}(\hat{p}_i^u < \hat{p}_j^u)}{\sum_{i \in I^0} \sum_{j \in I^1} \mathbf{1}(\hat{p}_i^o < \hat{p}_j^o)}$$

1 → perfect compatibility, 0 → perfect incompatibility

C^R is a new compatibility measure inspired by AUROC

Not threshold dependent.

Has a direct relationship with AUROC which we can use to assess trade-offs.

Lower bound of C^R expected to increase as model performance increases.

Q1: *Do we get C^R for free when we make updated models targeting AUROC?*

More specifically: do we observe $C^R = 1$ (or very close) when we train updated models using binary cross entropy loss?

Hypothesis: No, analytically we'd expect that C^R is centered at a region away from the upper and lower bounds.

Q1: Experimental Setup

Original Model
$n = 1{,}000$

Original Model Development
$n = 500$

Original Model Validation
$n = 500$

Updated Model Dataset
$n = 5{,}000$

Updated Model Development
$n = 2{,}500$

Updated Model Validation
$n = 2{,}500$

Evaluation Dataset
$n = 2{,}577$

MIMIC-III Mortality Dataset

Alistair 2016, Bansal 2019, Tang 2020

Q1: Experimental Setup

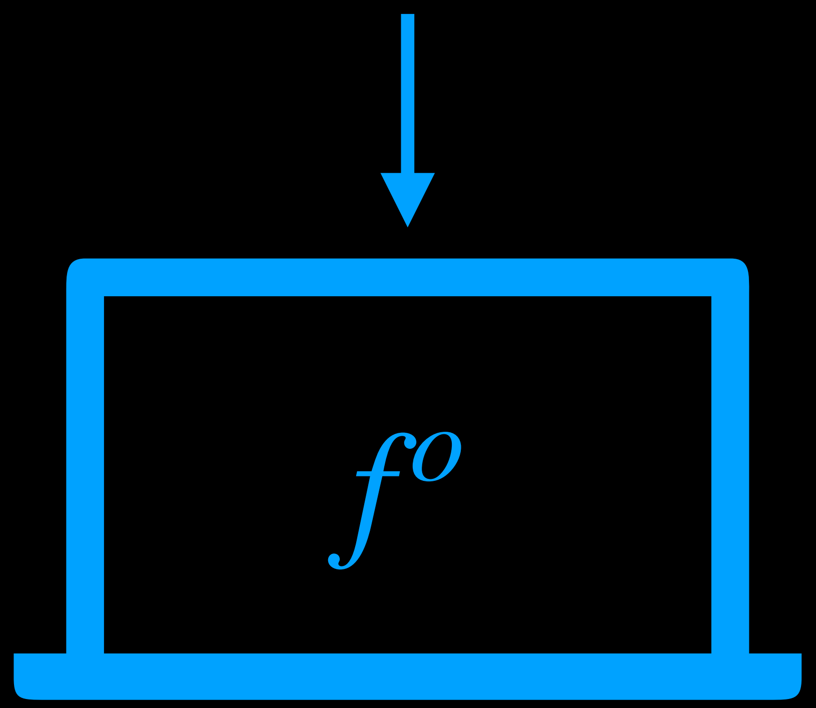
Original Model
$n = 1{,}000$

- **Original Model Development**
  $n = 500$
- **Original Model Validation**
  $n = 500$

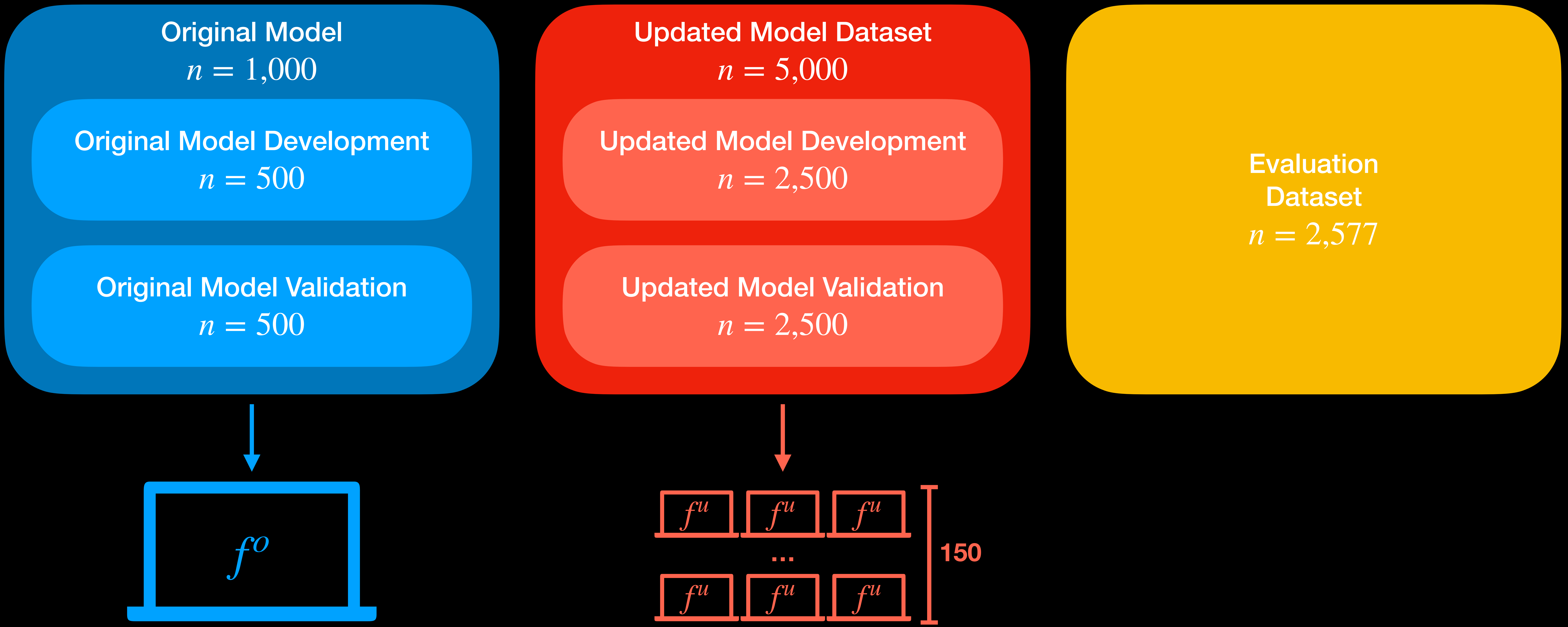
Updated Model Dataset
$n = 5{,}000$

- **Updated Model Development**
  $n = 2{,}500$
- **Updated Model Validation**
  $n = 2{,}500$

Evaluation Dataset
$n = 2{,}577$



Q1: Experimental Setup



Q1: Experimental Setup

Original Model
$n = 1{,}000$

Original Model Development
$n = 500$

Original Model Validation
$n = 500$

Updated Model Dataset
$n = 5{,}000$

Updated Model Development
$n = 2{,}500$

Updated Model Validation
$n = 2{,}500$

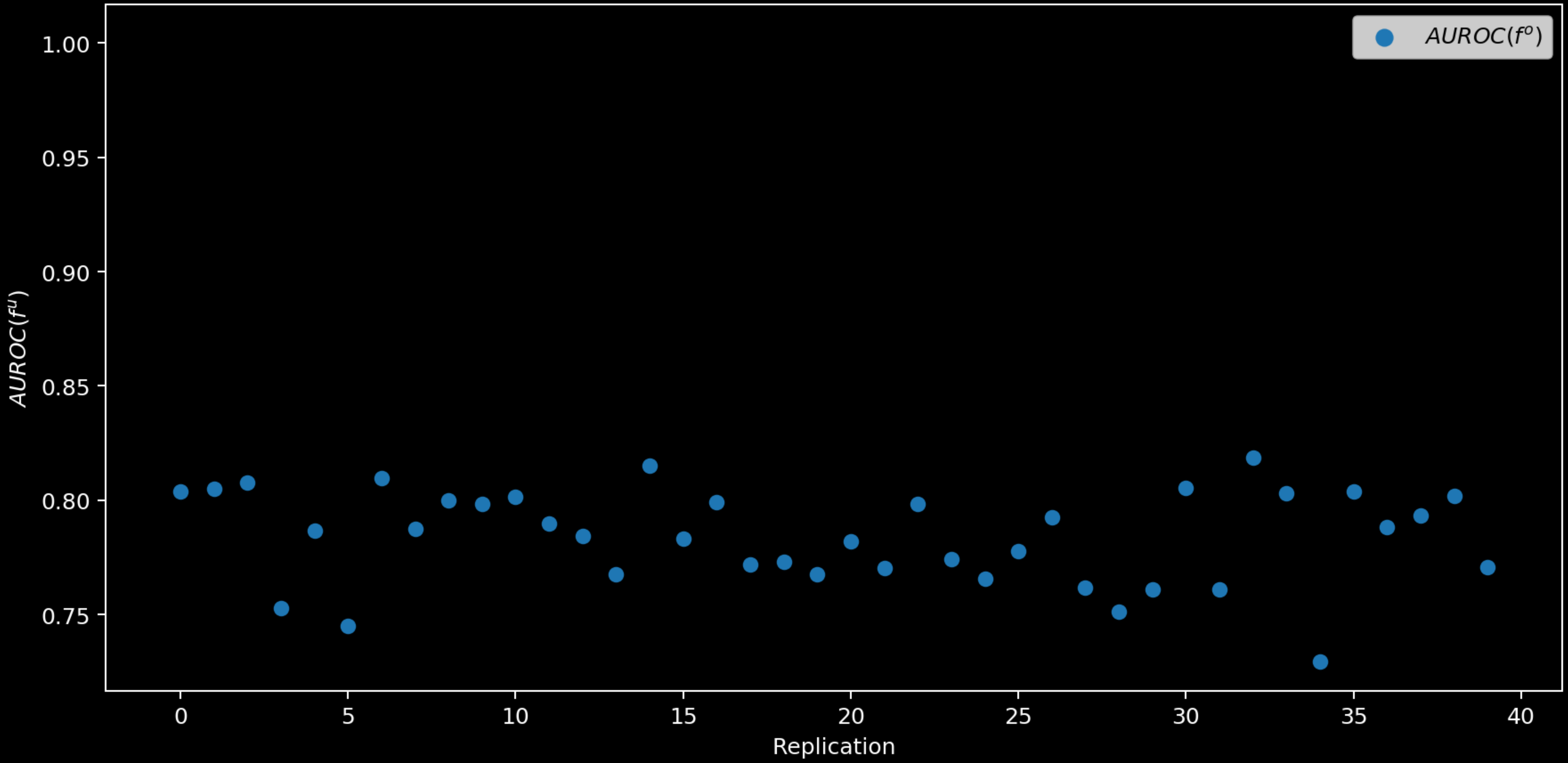
Evaluation Dataset
$n = 2{,}577$

f^o

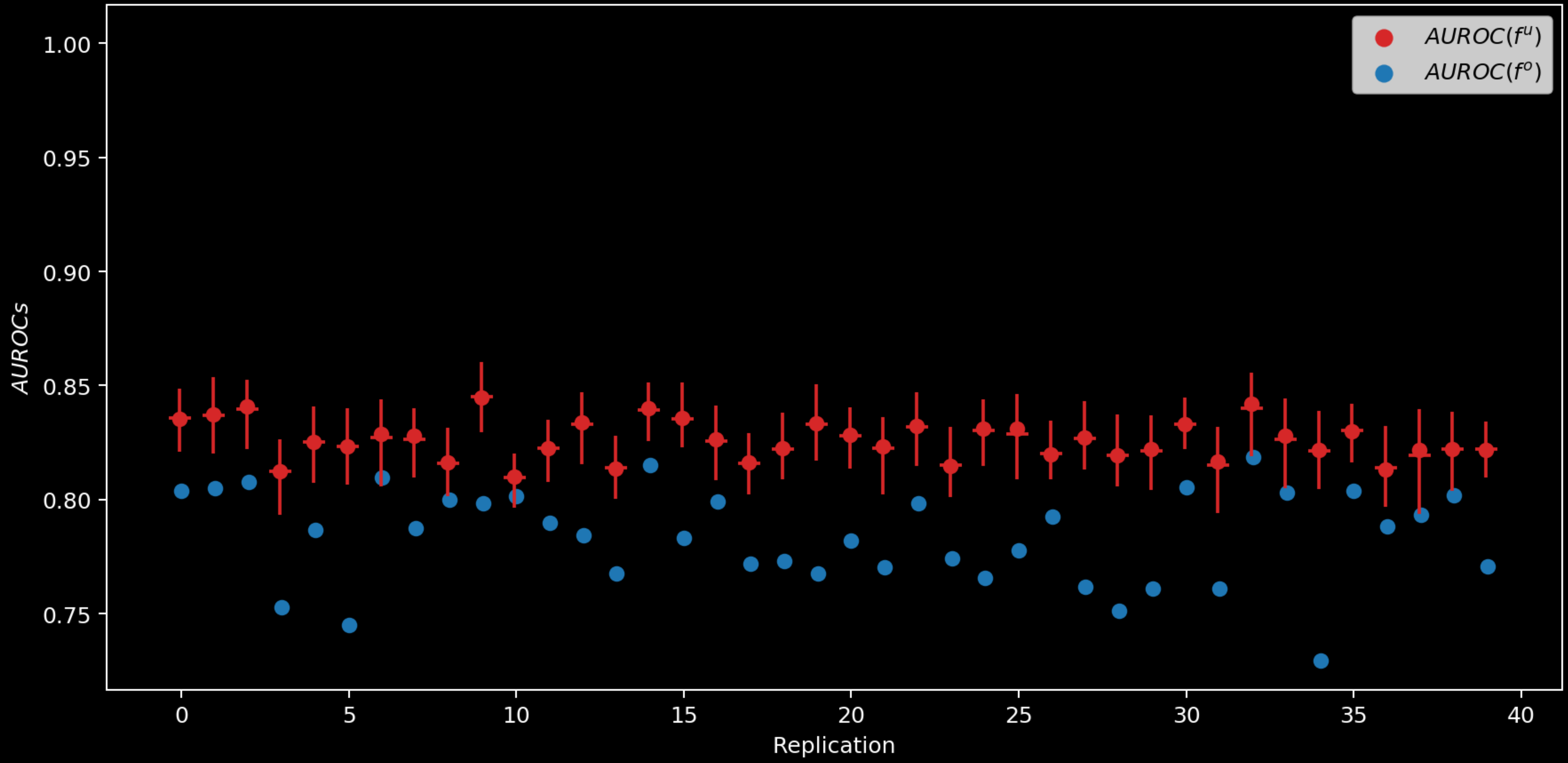
f^u f^u f^u
...
f^u f^u f^u
150

$AUROC(f^o)$
$AUROC(f^u)$
$C^R(f^o, f^u)$

Results

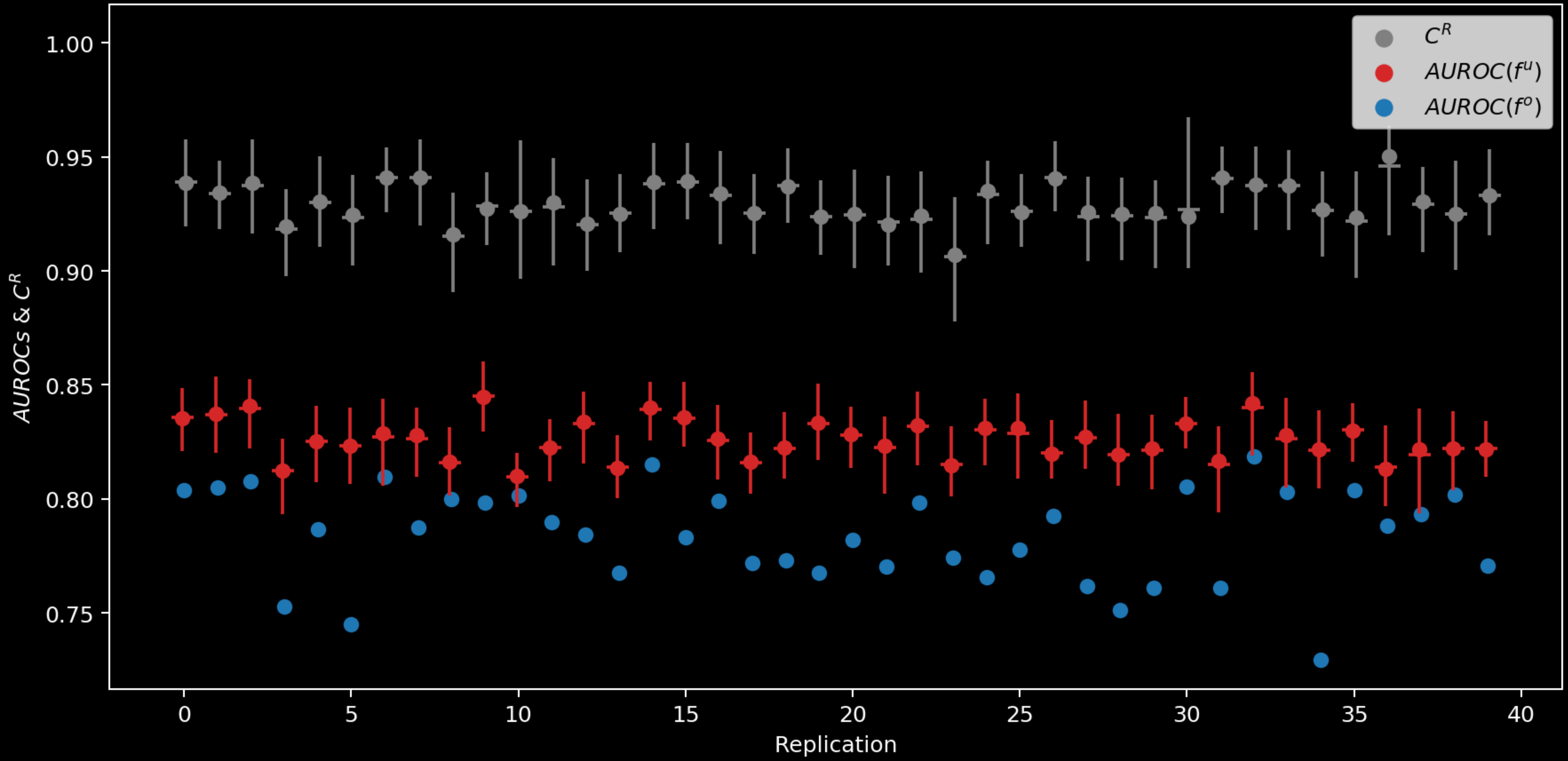

$AUROC(f^o)$
$AUROC(f^u)$
1.00
0.95
0.90
0.85
0.80
0.75
0 5 10 15 20 25 30 35 40
Replication

Results



$AUROC(f^u)$
$AUROC(f^o)$
1.00
0.95
0.90
0.85
0.80
0.75
$AUROCs$
0
5
10
15
20
25
30
35
40
Replication

Results



C^R
$AUROC(f^u)$
$AUROC(f^o)$
1.00
0.95
0.90
0.85
0.80
0.75
$AUROCs$ & C^R
0
5
10
15
20
25
30
35
40
Replication

Q1: ***Do we get C^R for free when we make updated models targeting AUROC?***

No.

We observe updated models have a limited range in C^R.

Analytical results suggest that there’s a large search space .

Motivates techniques to search for updated models that have higher C^R.

Risk stratification models are usually trained with binary cross entropy loss.

Binary cross entropy loss:

$$\mathscr{L}^{BCE}(f) = -\sum_{i \in I^0} \log(1 - \hat{p}_i) - \sum_{j \in I^1} \log(\hat{p}_j)$$

Minimization of $\mathscr{L}^{BCE}$ leads to higher AUROC because risk estimates tend to align with labels.

Risk stratification models are usually trained with binary cross entropy loss.

Binary cross entropy loss:

Make 0-labeled patients have low risk estimates

$$\mathscr{L}^{BCE}(f) = \boxed{-\sum_{i \in I^0} \log(1-\hat{p}_i)} - \sum_{j \in I^1} \log(\hat{p}_j)$$

Minimization of $\mathscr{L}^{BCE}$ leads to higher AUROC because risk estimates tend to align with labels.

Risk stratification models are usually trained with binary cross entropy loss.

Binary cross entropy loss:

Make 0-labeled patients have low risk estimates

Make 1-labeled patients have high risk estimates

$$\mathscr{L}^{BCE}(f) = \boxed{-\sum_{i \in I^0} \log(1 - \hat{p}_i)} \boxed{-\sum_{j \in I^1} \log(\hat{p}_j)}$$

Minimization of $\mathscr{L}^{BCE}$ leads to higher AUROC because risk estimates tend to align with labels.

Risk stratification models are usually trained with binary cross entropy loss.

Binary cross entropy loss:

Make 0-labeled patients have low risk estimates

Make 1-labeled patients have high risk estimates

$$\mathscr{L}^{BCE}(f) = \boxed{-\sum_{i \in I^0} \log(1-\hat{p}_i)} \boxed{-\sum_{j \in I^1} \log(\hat{p}_j)}$$

Minimization of $\mathscr{L}^{BCE}$ leads to higher AUROC because risk estimates tend to align with labels.

No focus on compatibility between the updated and original model.

We introduce rank-based incompatibility loss.

Rank-based incompatibility loss:

$$\mathscr{L}^R(f^o, f^u) = 1 - C^R(f^o, f^u)$$

Minimization of $\mathscr{L}^R$ will lead to higher levels of C^R.

Differentiable approximation $\widetilde{\mathscr{L}^R}$ for SGD.

Weighted loss trades-off between binary cross entropy and compatibility.

Weighted loss function:

$$\alpha \mathscr{L}^{BCE}(f^u) + (1-\alpha)\widetilde{\mathscr{L}^R}(f^o, f^u)$$

where $\alpha \in [0,1]$

When:

$\alpha = 1$ then only minimize $\mathscr{L}^{BCE}$, ↑ AUROC

$\alpha = 0$ then only minimize $\widetilde{\mathscr{L}^R}$, ↑ C^R

$\alpha = 0.5$ then balance $\mathscr{L}^{BCE}$ and $\widetilde{\mathscr{L}^R}$

Q2: *Can we make updated models with higher levels of C^R?*

Specifically: Compared to standard update model generation and selection approaches, can we use the weighted loss function to generate updates with better C^R?

Hypothesis: using weighted loss function will produce models with better C^R.

Also, can this be accomplished without a loss of AUROC?

Q2: Extends Previous Experimental Setup

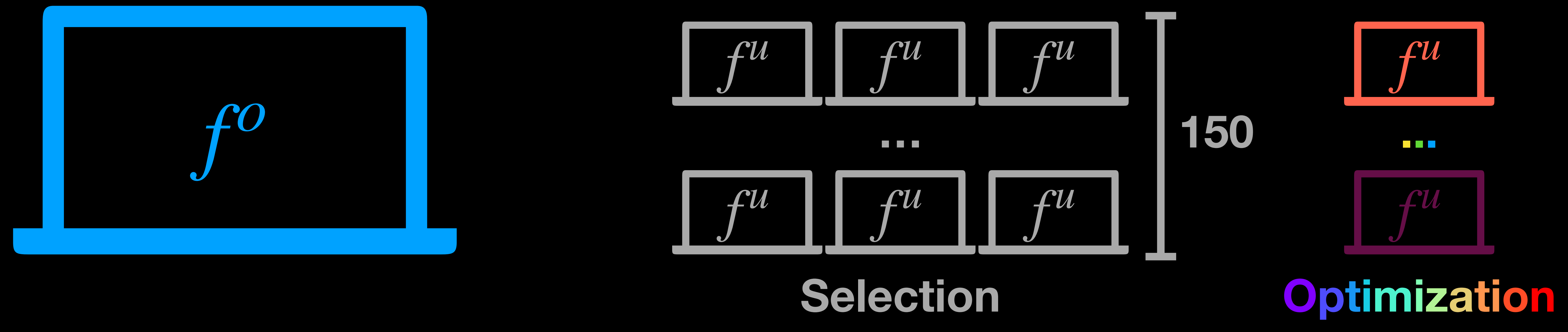
Original Model
$n = 1{,}000$

- Original Model Development $n = 500$
- Original Model Validation $n = 500$

Updated Model Dataset
$n = 5{,}000$

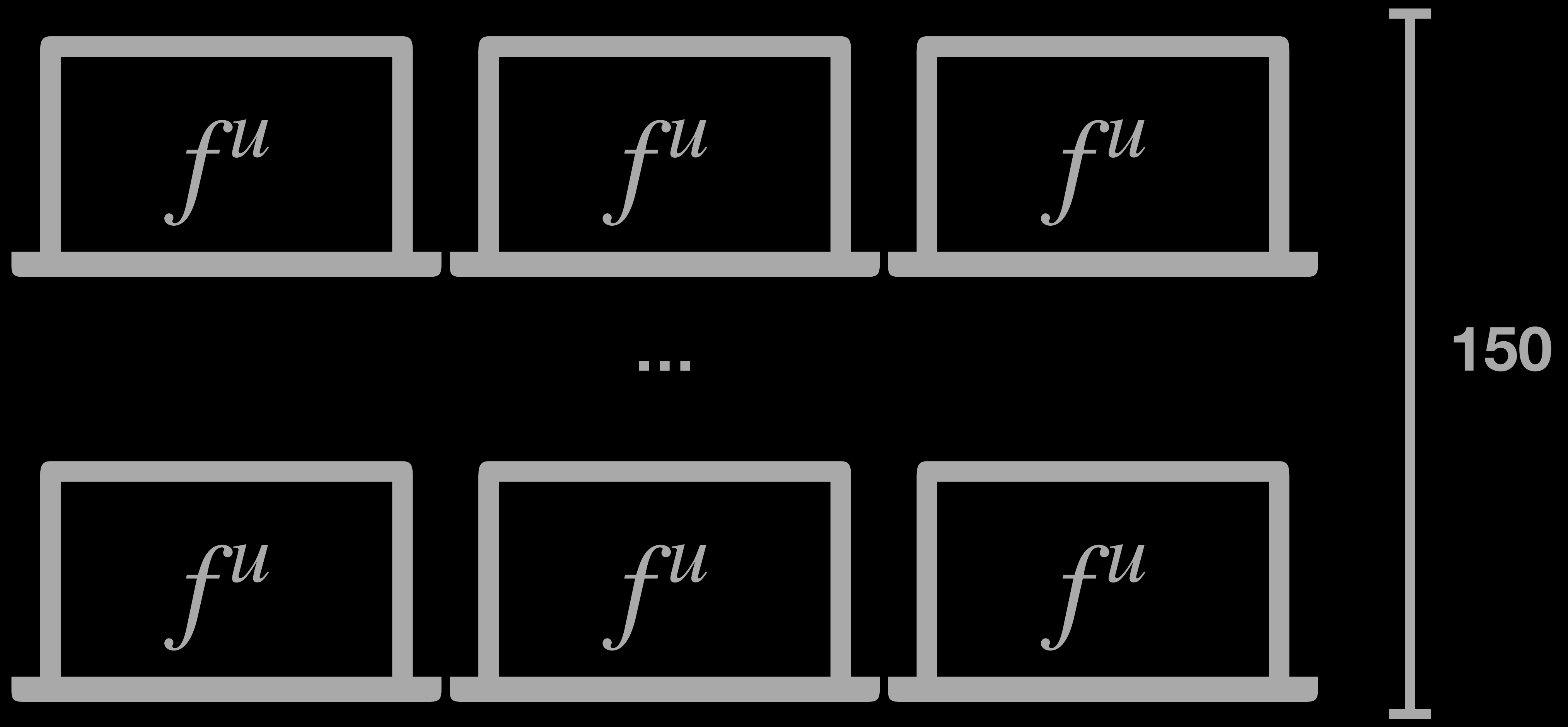
- Updated Model Development $n = 2{,}500$
- Updated Model Validation $n = 2{,}500$

Evaluation Dataset
$n = 2{,}577$

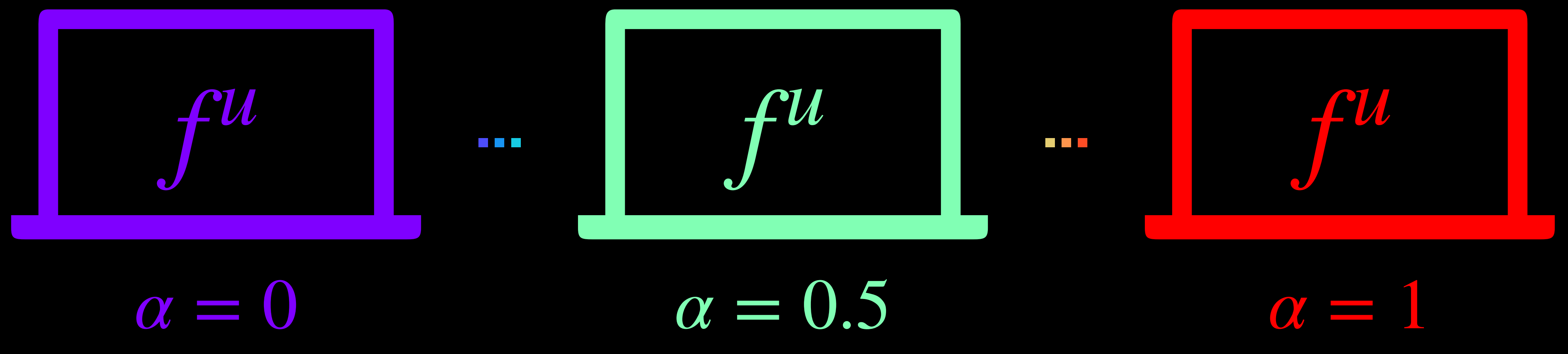


$AUROC(f^o)$
$AUROC(f^u)$
$C^R(f^o, f^u)$

Q2: Updated Models Selection vs. Optimization



Selection

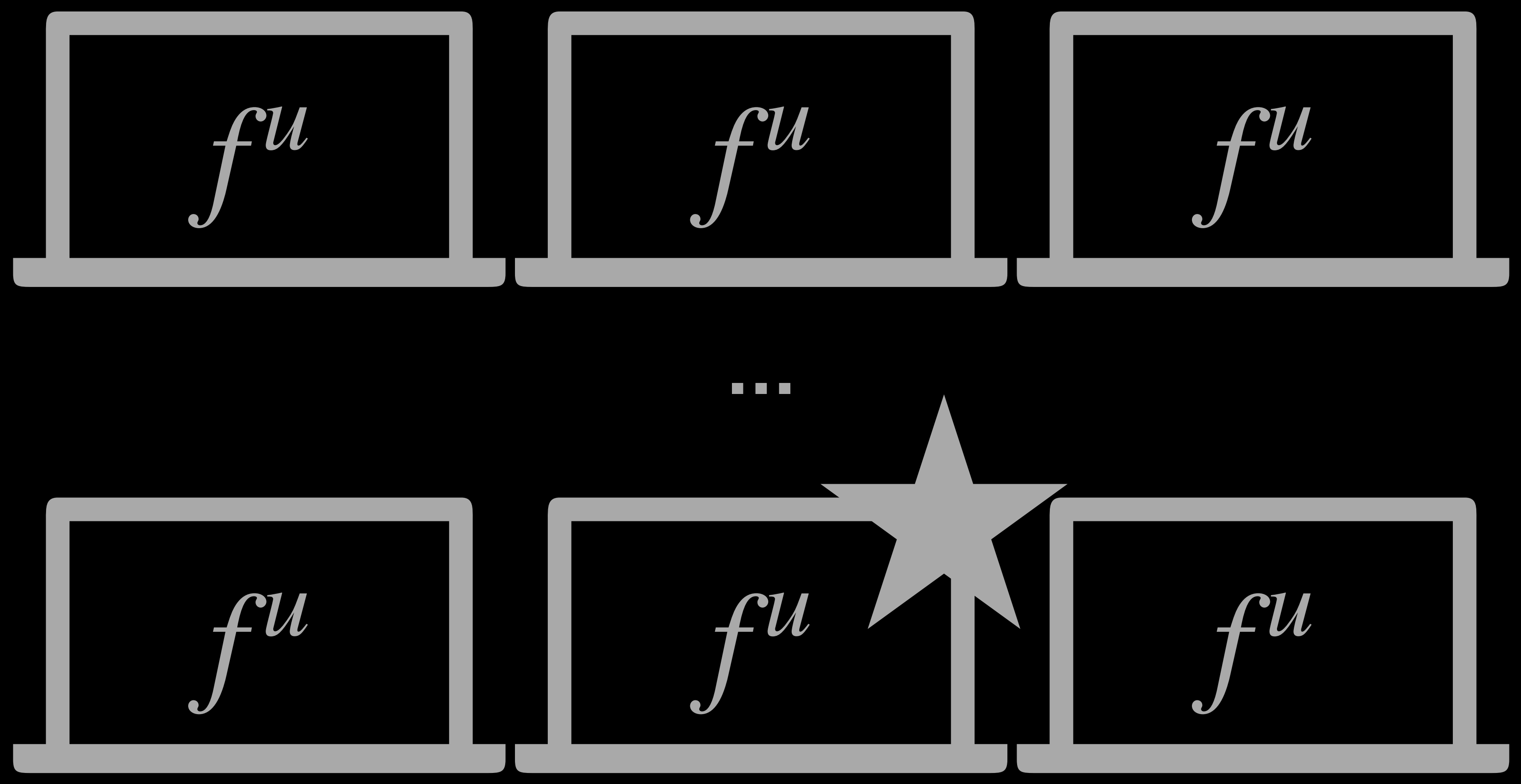


Optimization

150 "selection" models created through training with $\mathscr{L}^{BCE}$ and randomly resampling the development dataset

11 "optimization" models created through training with weighted loss with $\alpha = \{0, 0.1, \ldots, 0.9, 1\}$

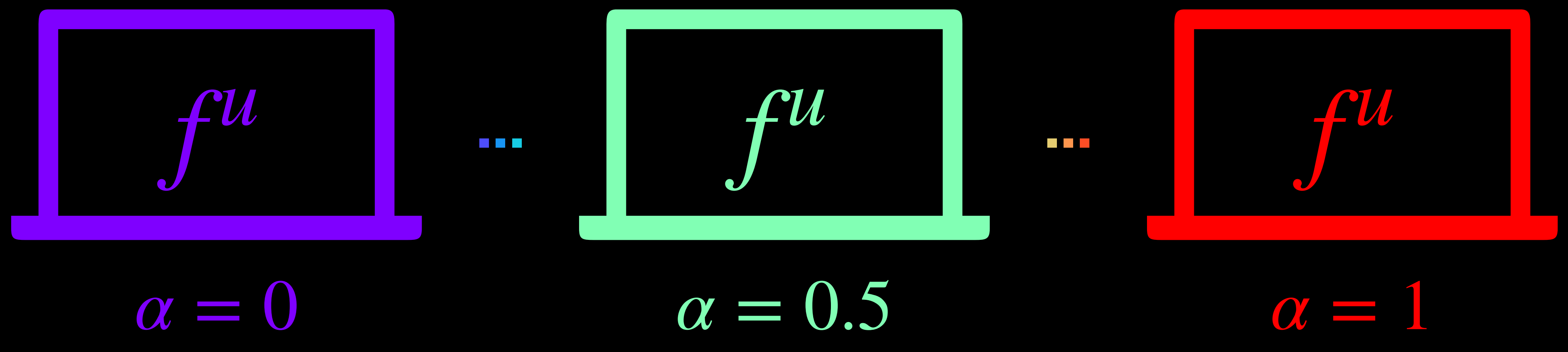
Q2: Updated Models Selection vs. Optimization



Selection

Use a selection procedure to pick an updated model to use as a baseline

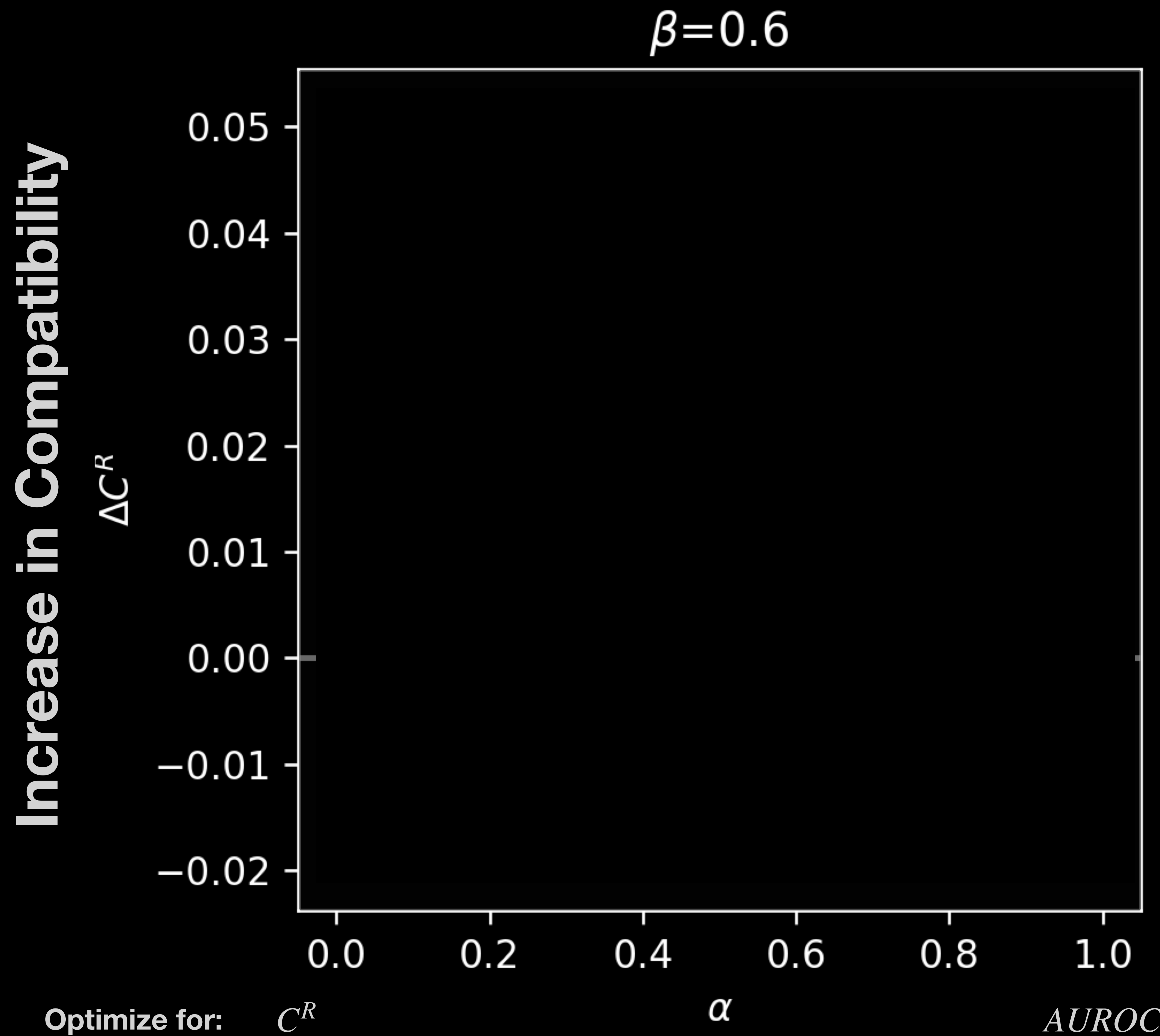
For example model with best validation AUROC



Optimization

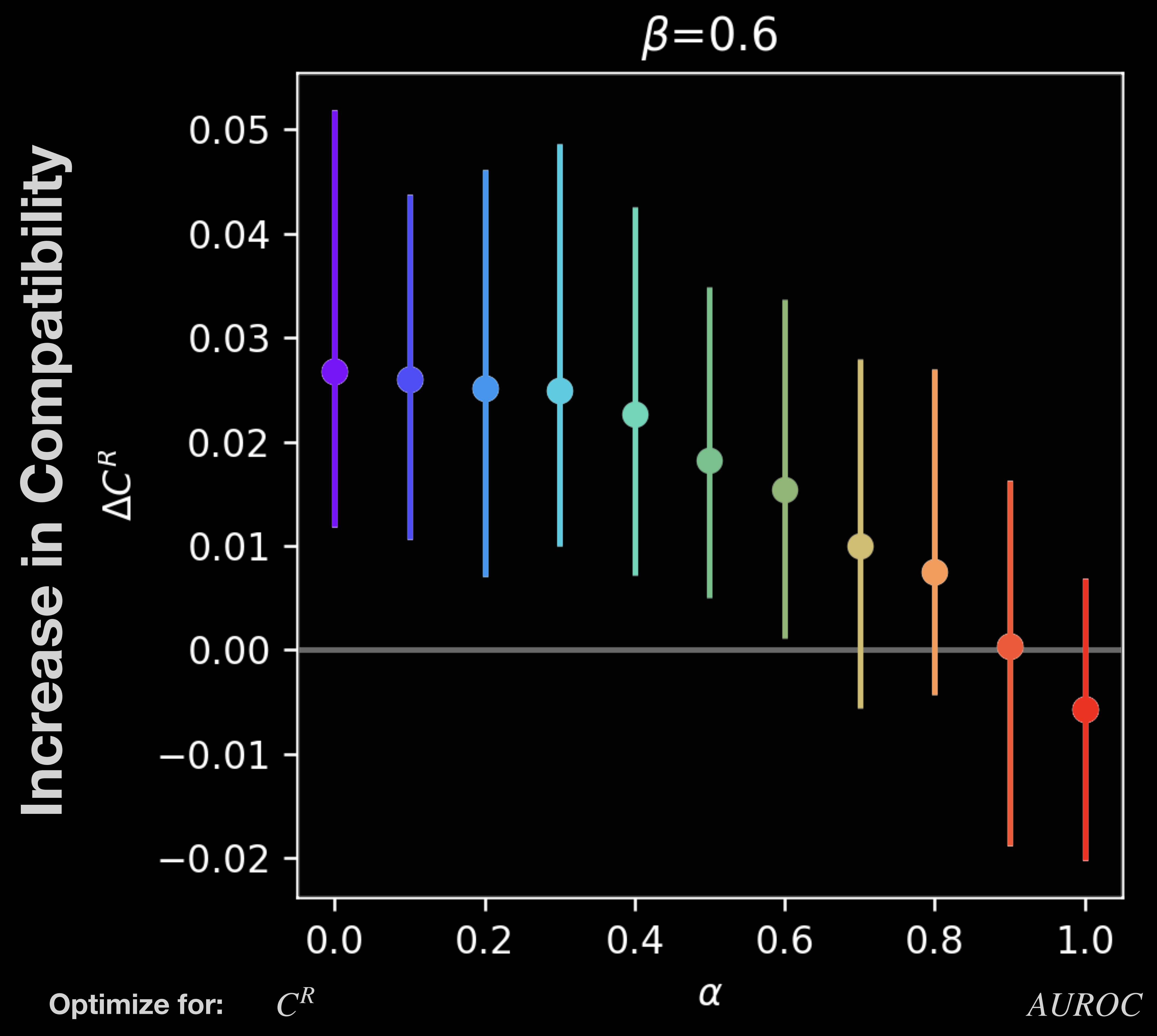
Examine difference in held out evaluation C^R and AUROC

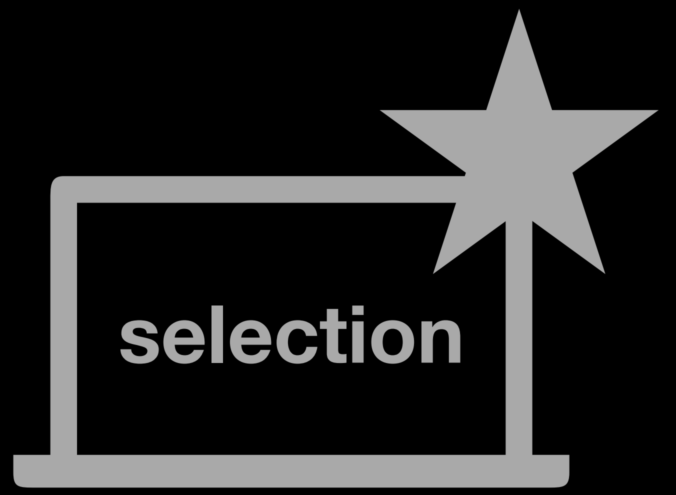
Q2: C^R performance results



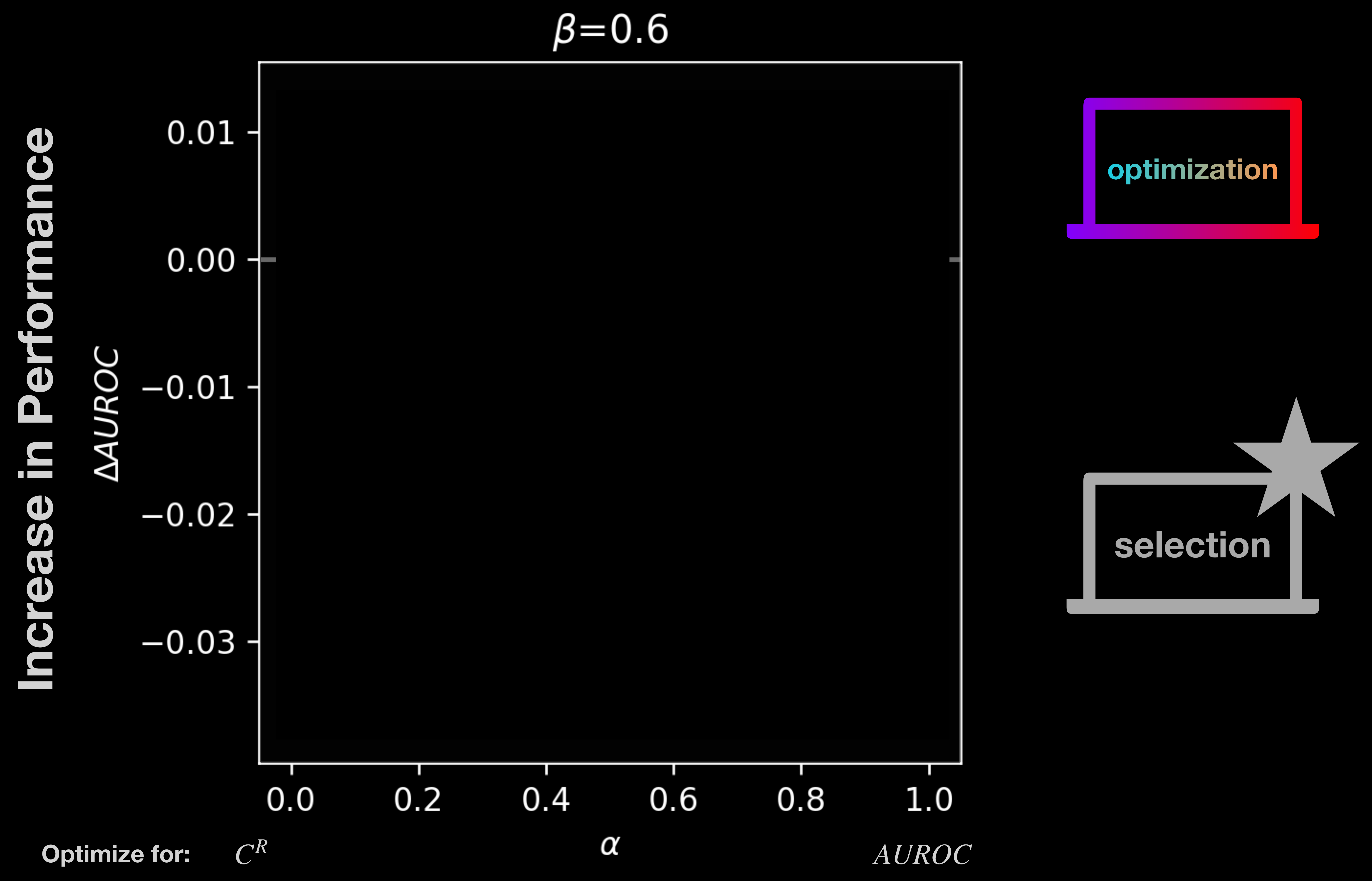
Optimize for: C^R $AUROC$

Q2: C^R performance results

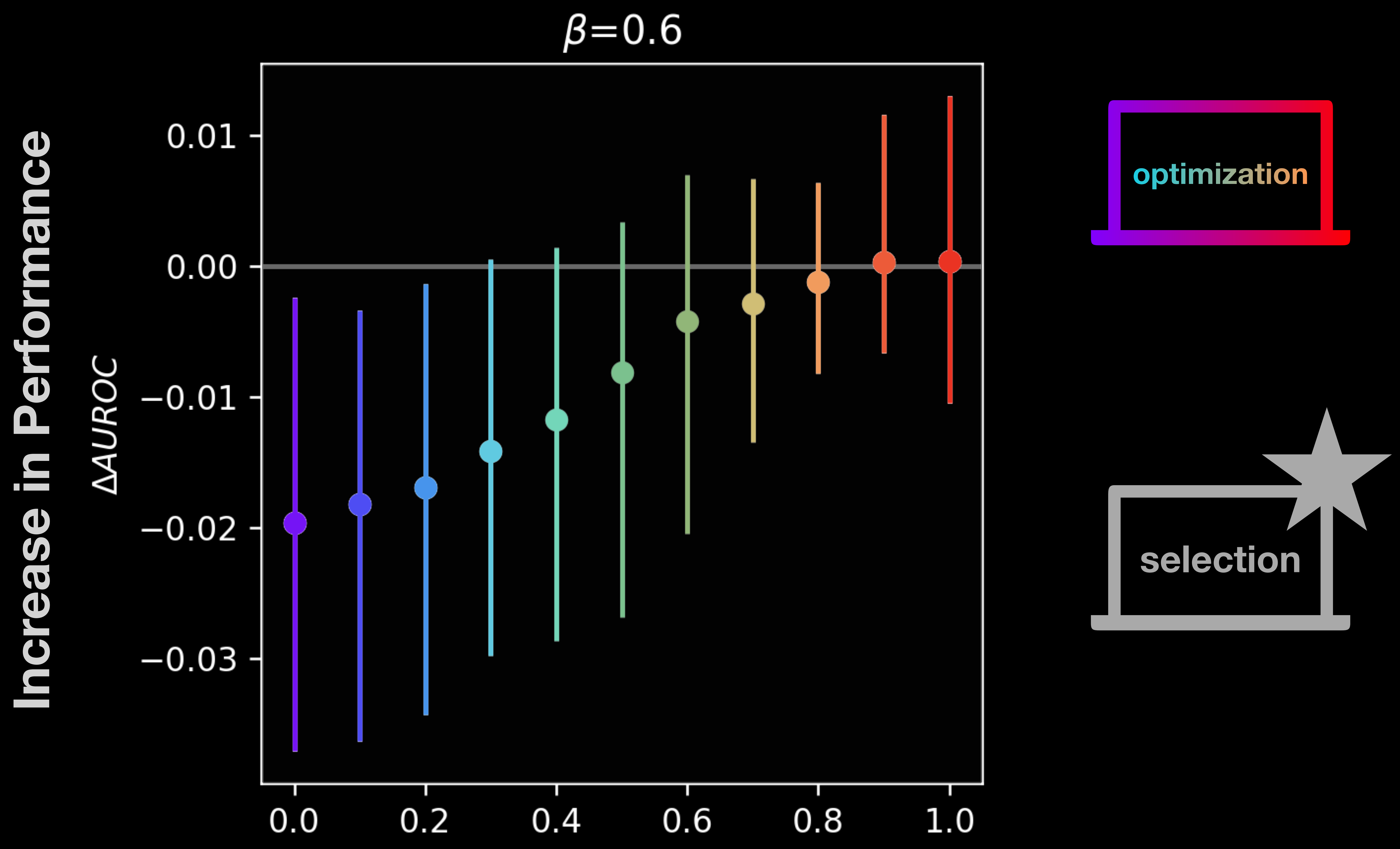




Q2: $AUROC$ performance results

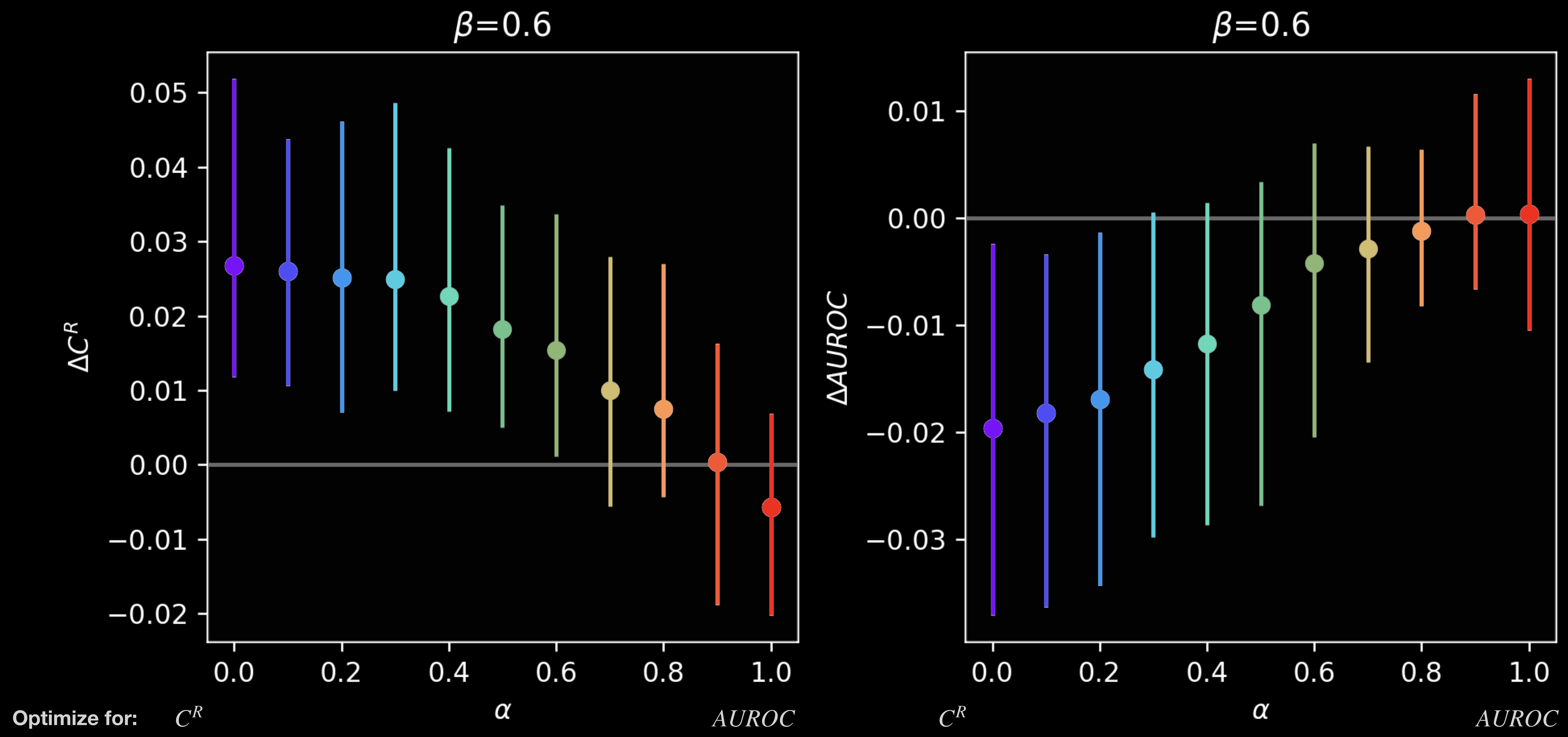


Q2: $AUROC$ performance results

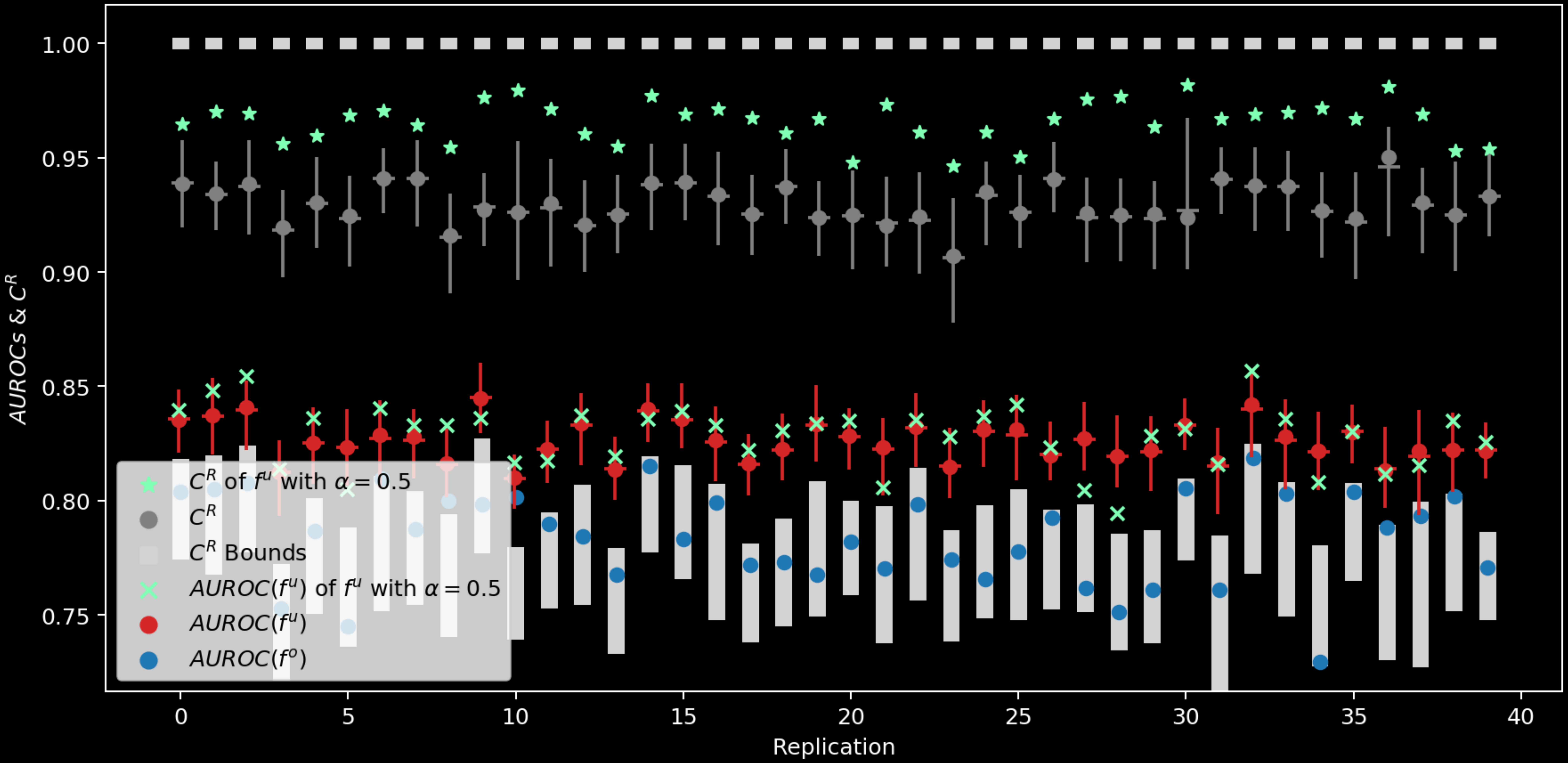


Optimize for: C^R $AUROC$

Q2: Performance results



Q2: α=0.6 yields promising updated models



Summary of experiments

Do we get C^R *for free when we make updated models targeting AUROC?*

No.

Can we make updated models with higher levels of C^R*?*

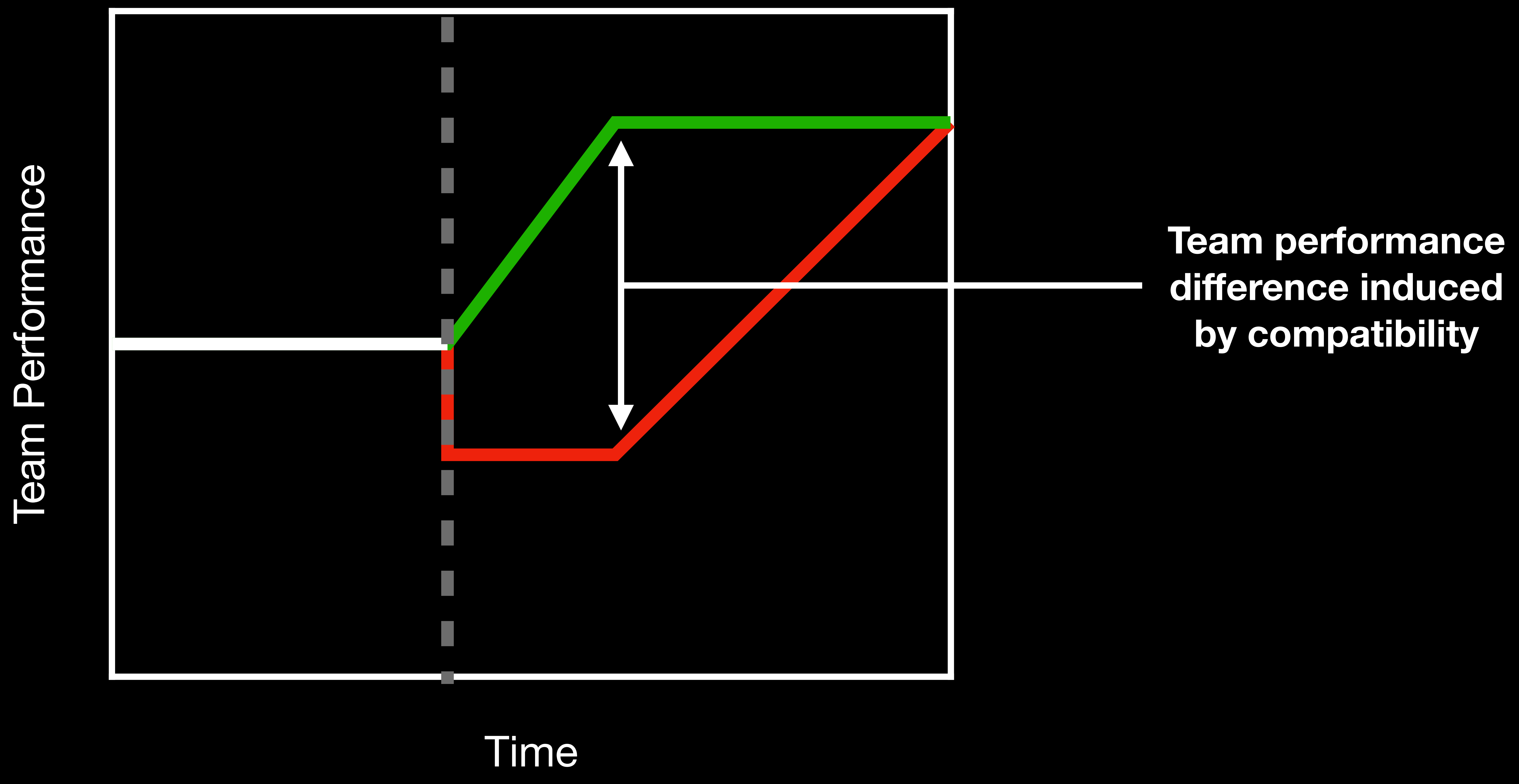
Yes, using our weighted loss function.

Does that come at a cost in terms of AUROC?

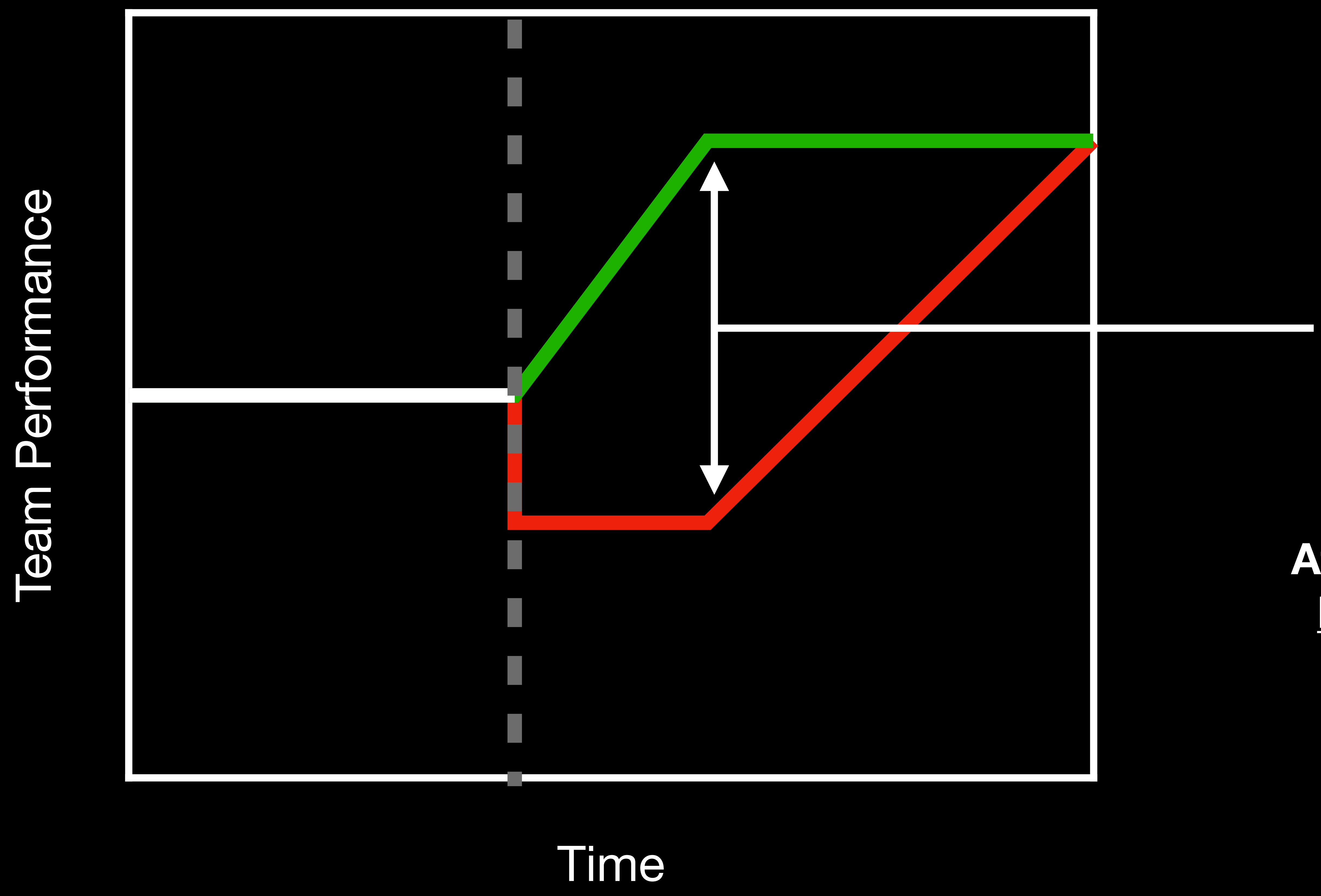
Sort of…

Why trade-off AUROC for C^R?

Why trade-off AUROC for C^R?



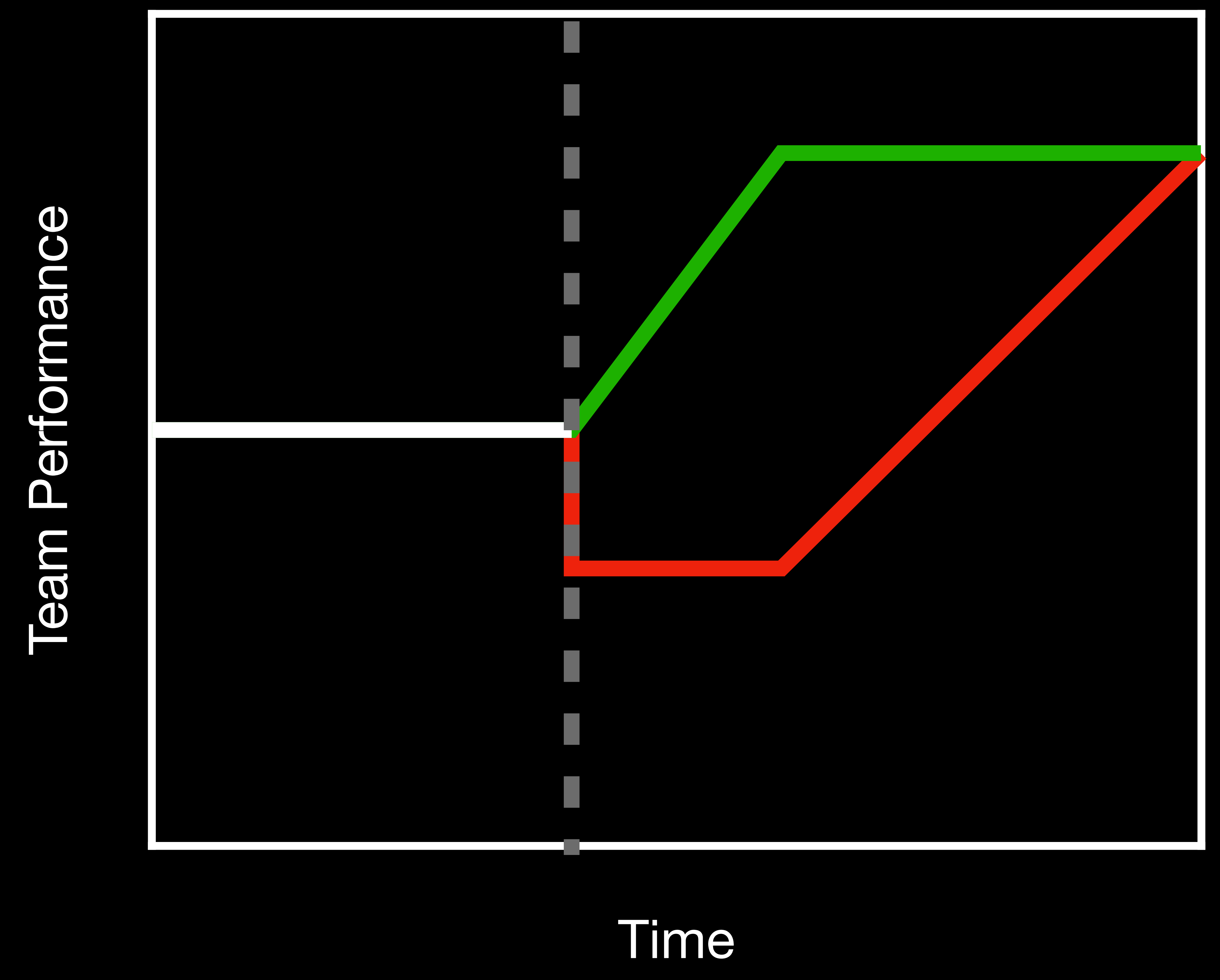
Why trade-off AUROC for C^R?



Team performance difference induced by compatibility

Affects physician users.
<u>Impacts patient lives.</u>

Why trade-off AUROC for C^R?



We should optimize for good team performance.

C^R is a new compatibility measure inspired by AUROC

Not threshold dependent: ↑ clinical utility

Has direct relationship with AUROC

Can balance AUROC and C^R

Using $\widetilde{\mathscr{L}^R}$ → ↑C^R & ↑AUROC

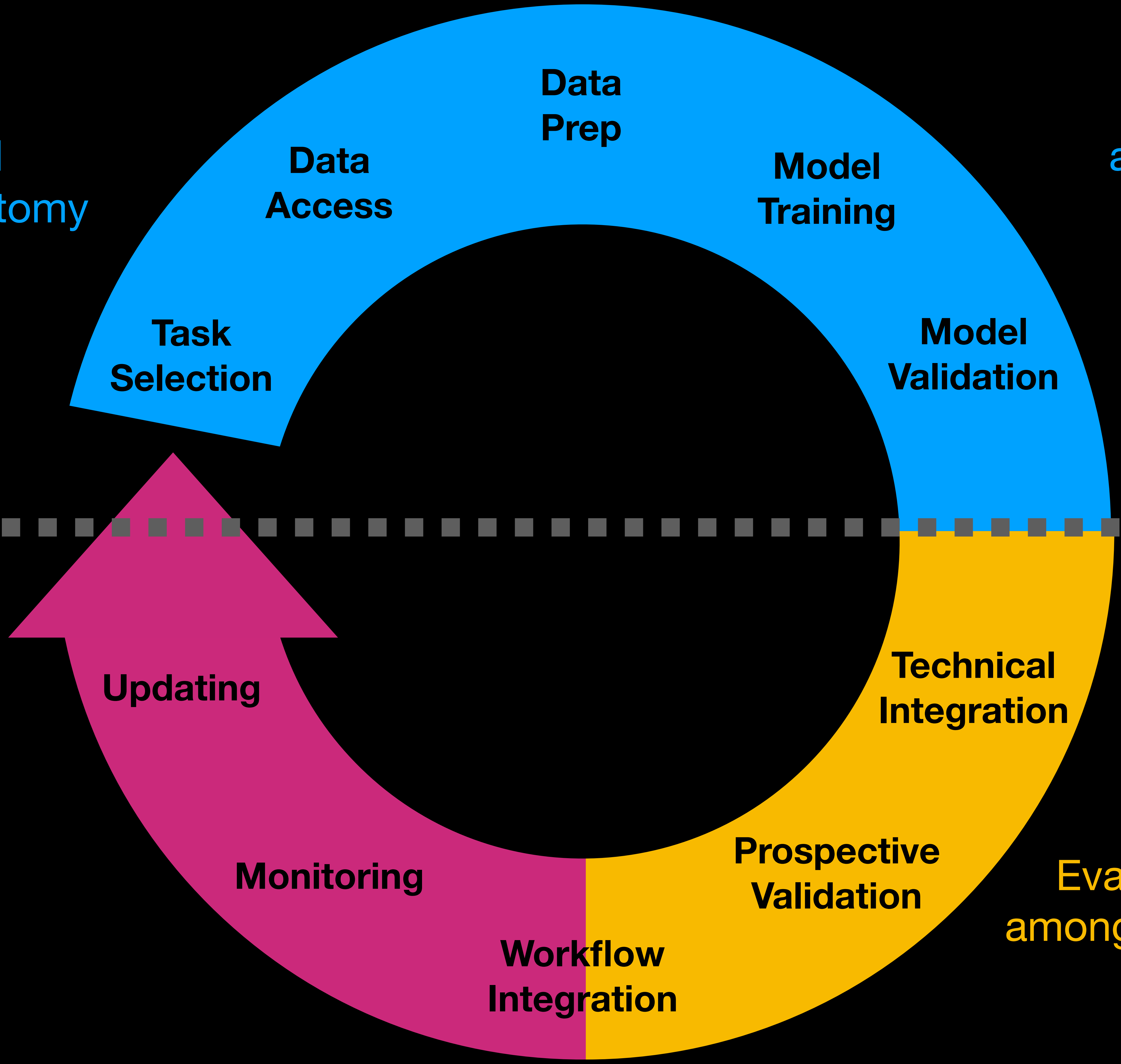
Real-world model updating case-study

eotles@umich.edu
@eotles

Rank-based compatibility improves the whole life-cycle

Development & validation of models to predict pathological outcomes of radical prostatectomy
<u>**Ötleş et al., 2022**</u>

Early identification of patients admitted to hospital for covid-19 at risk of clinical deterioration
<u>**Kamran et al., 2022**</u>



Rank-based compatibility measurement & optimization for model updating

Validation of the Epic sepsis model
<u>**Wong et al., 2022**</u>

Evaluation of the Epic deterioration index among hospitalized patients with COVID-19
<u>Singh et al., 2021</u>