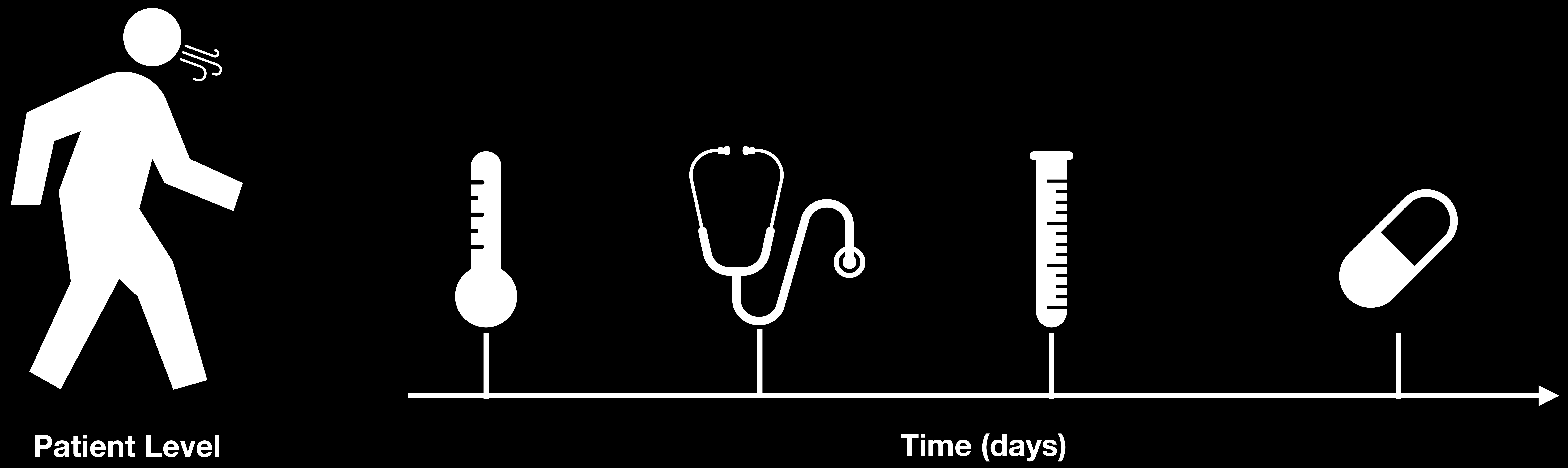
Machine Learning for Healthcare

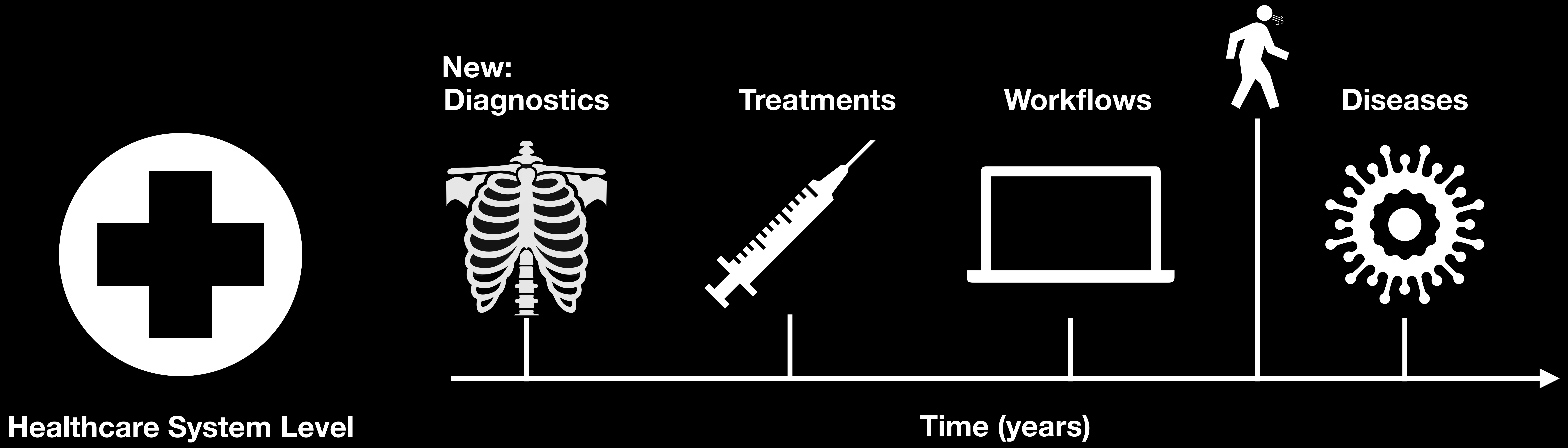
Lessons From Across The Healthcare ML Lifecycle

WPI Business Week
Erkin Ötleş
October 2023

Time plays an important role in medicine.



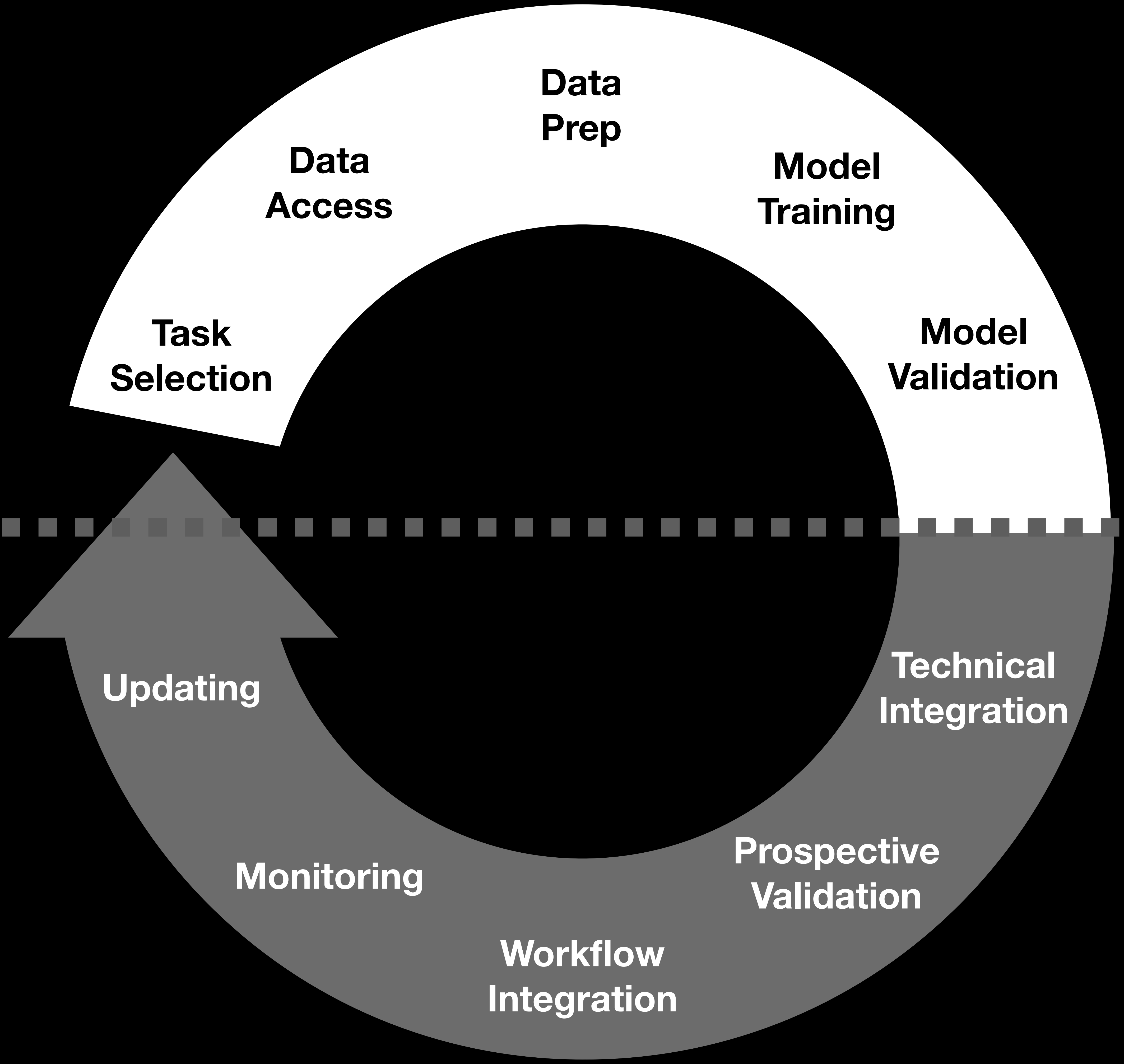
Time plays an important role in medicine.



Time complicates ML model development & implementation.

Development

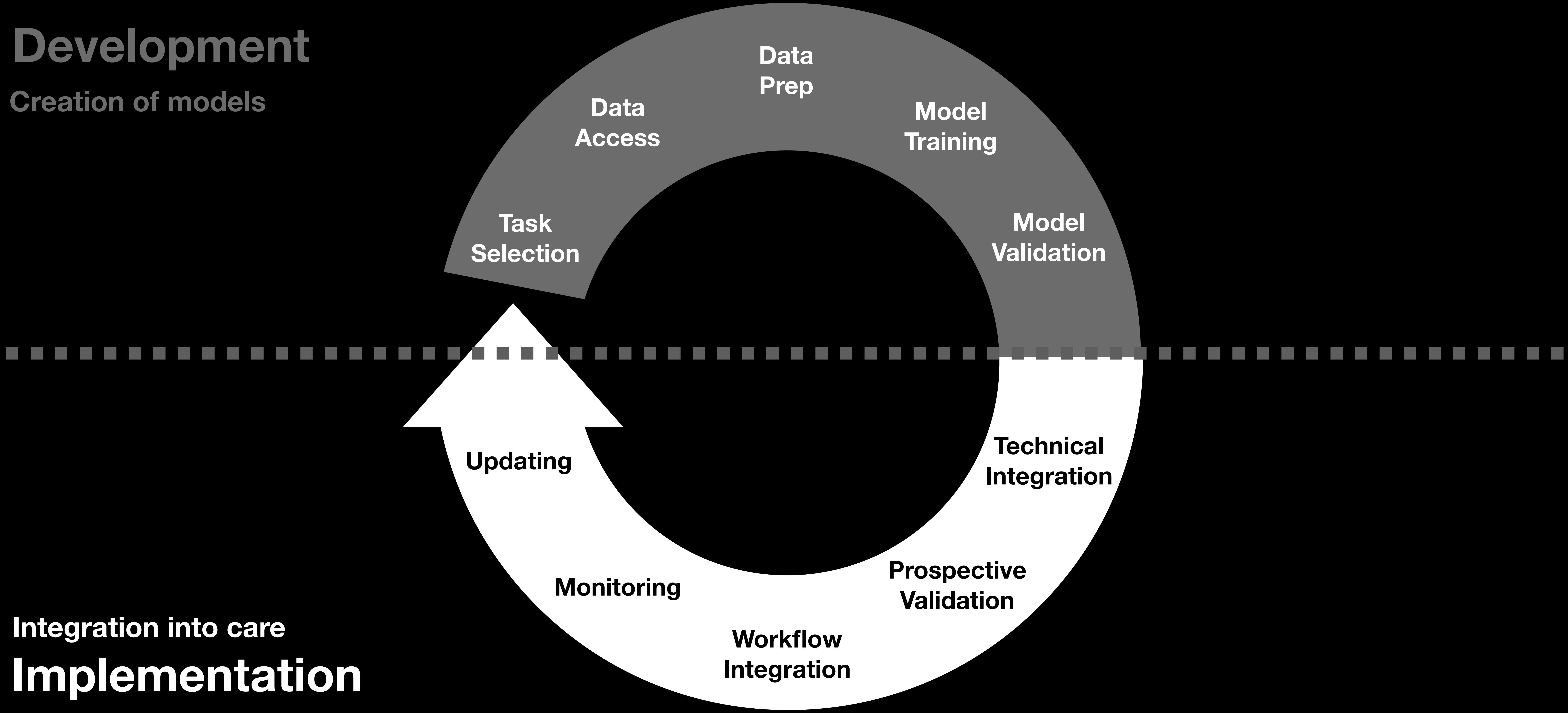
Creation of models



Integration into care

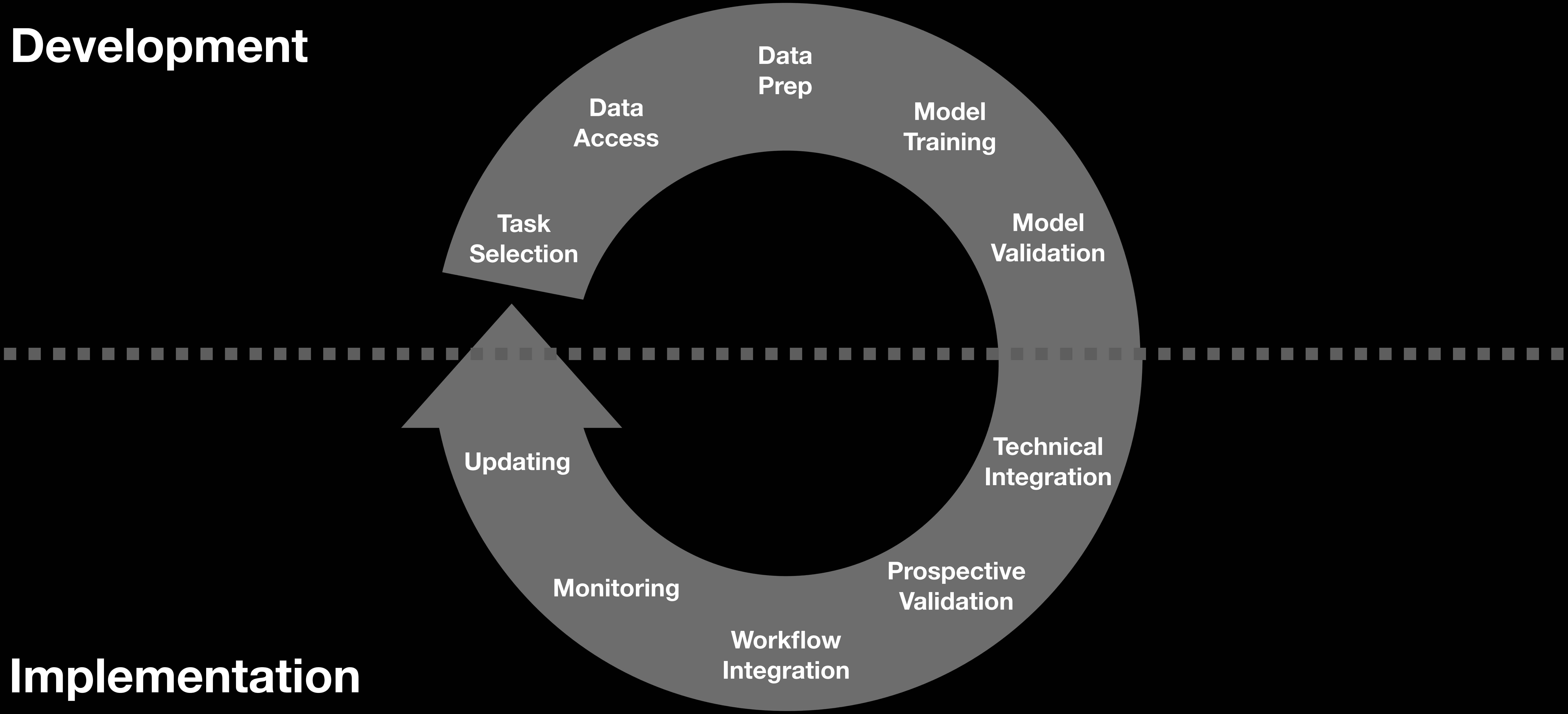
Implementation

Time complicates ML model development & implementation.

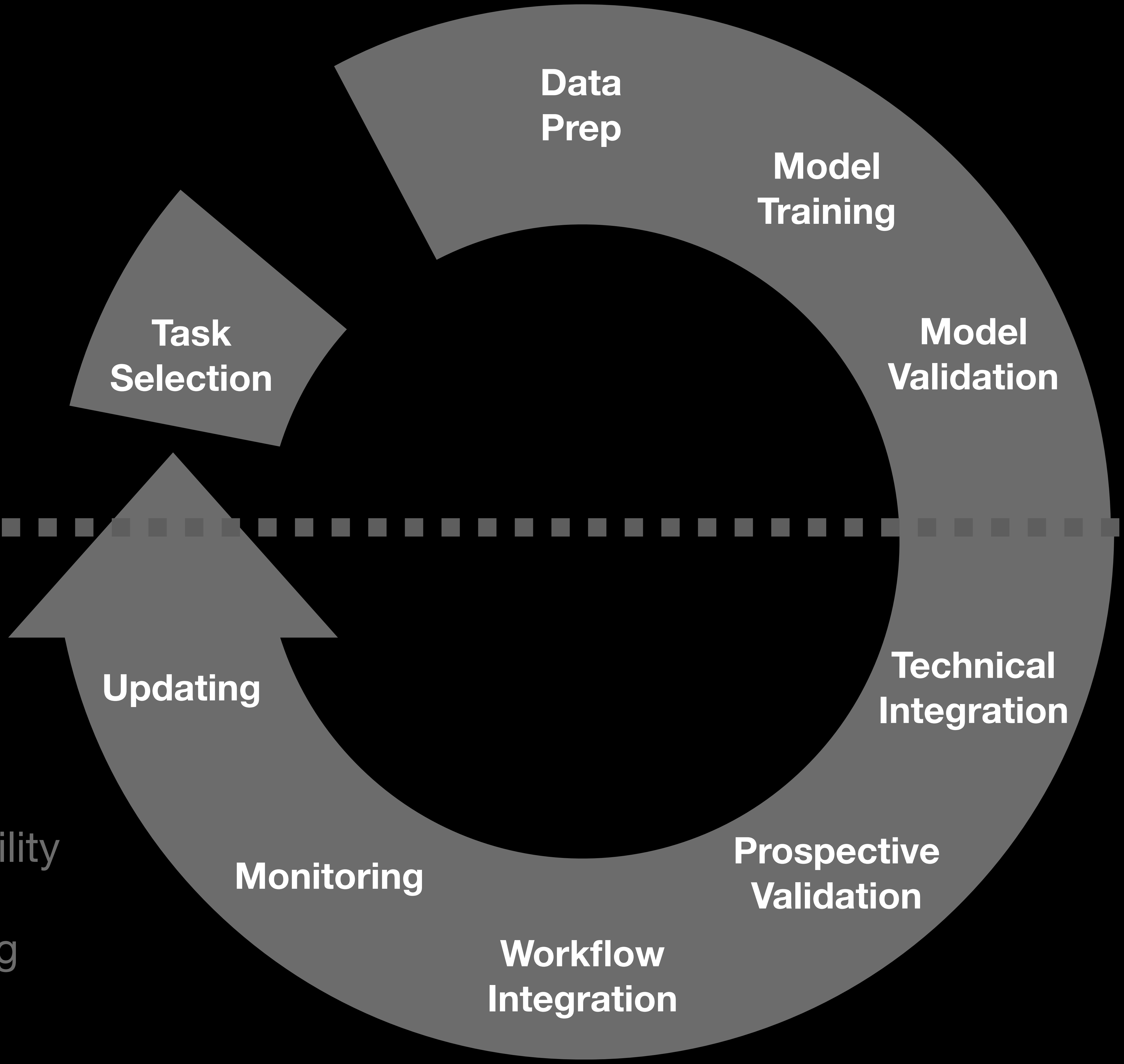


Thesis: by focusing on the longitudinal nature of healthcare, we can improve the performance of predictive models during development & implementation.

The ML Lifecycle



The ML Lifecycle



RTW Prediction

Assess value of longitudinal observations for return to work prediction

Ötleş *et al.*, JAMIA ‘22

Updating Models

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

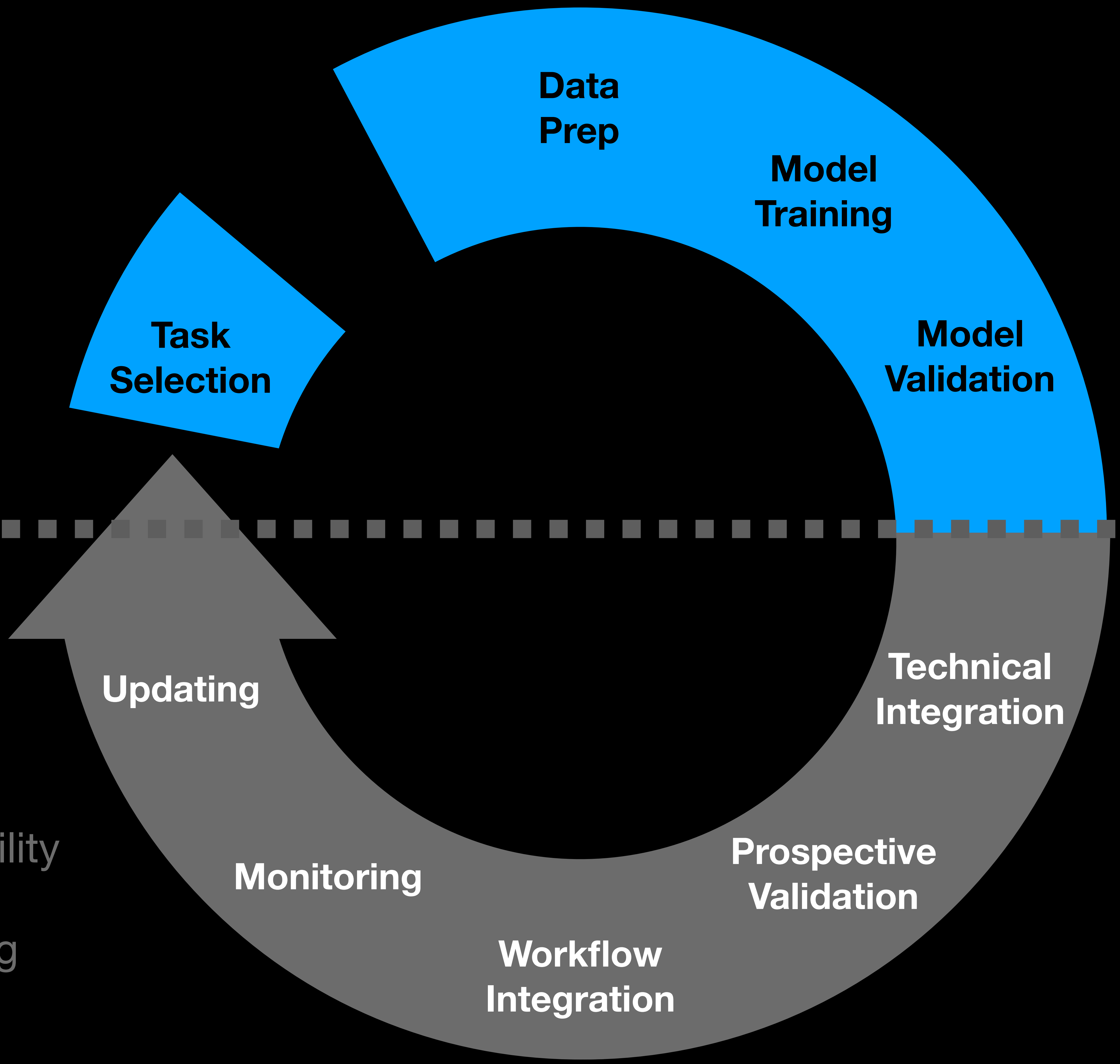
Ötleş *et al.*, MLHC '23

Performance Gap

Characterize changes to prospective performance after model implementation

Ötleş *et al.*, MLHC ‘21

The ML Lifecycle



RTW Prediction

Assess value of
longitudinal observations
for return to work prediction

Ötleş *et al.*, JAMIA ‘22

Updating Models

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

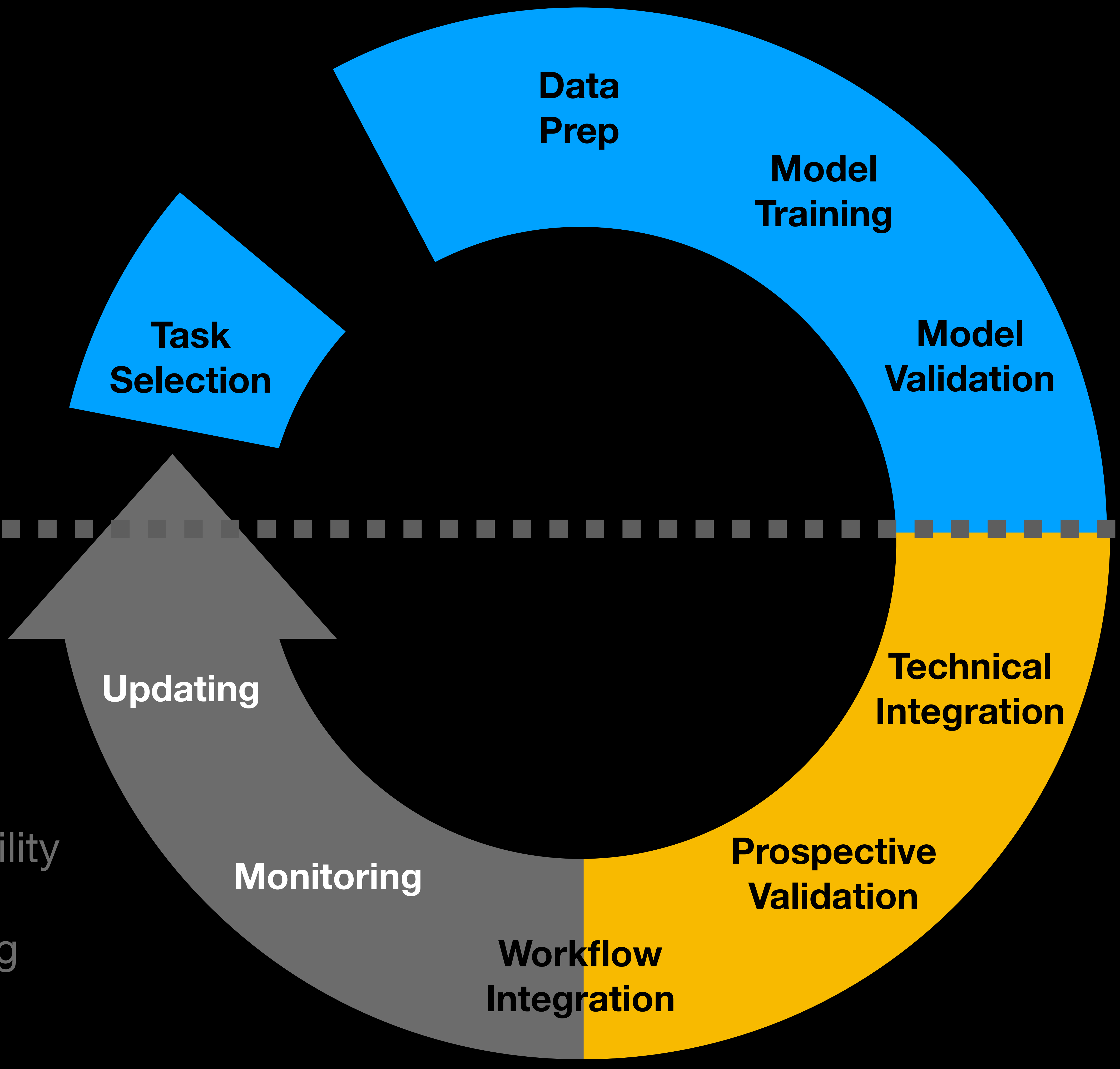
Ötleş *et al.*, MLHC '23

Performance Gap

Characterize changes
to prospective performance
after model implementation

Ötleş *et al.*, MLHC ‘21

The ML Lifecycle



RTW Prediction

Assess value of longitudinal observations for return to work prediction

Ötleş *et al.*, JAMIA '22

Updating Models

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

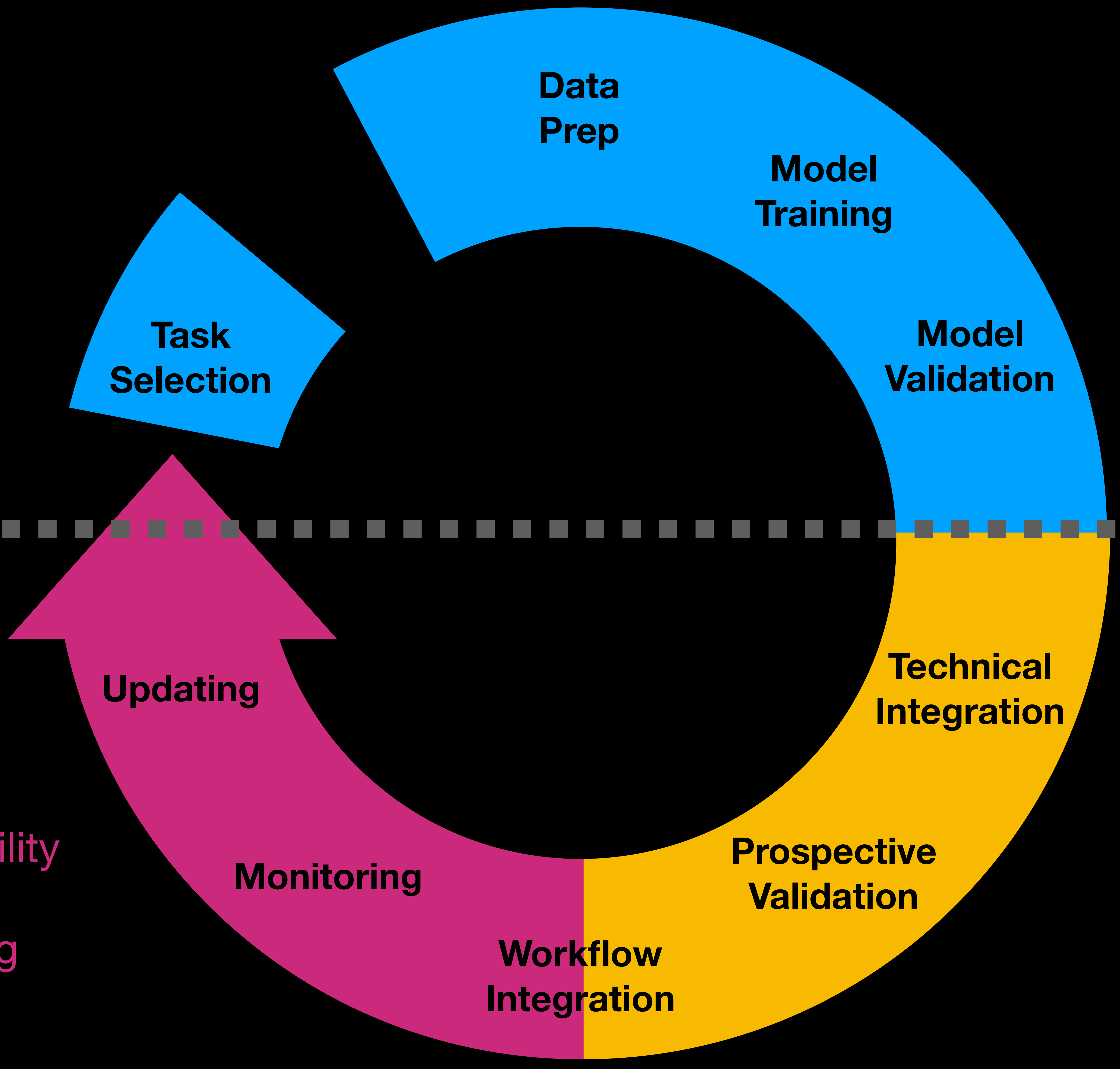
Ötleş *et al.*, MLHC '23

Performance Gap

Characterize changes to prospective performance after model implementation

Ötleş *et al.*, MLHC '21

The ML Lifecycle



RTW Prediction

Assess value of longitudinal observations for return to work prediction

Ötleş *et al.*, JAMIA '22

Updating Models

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

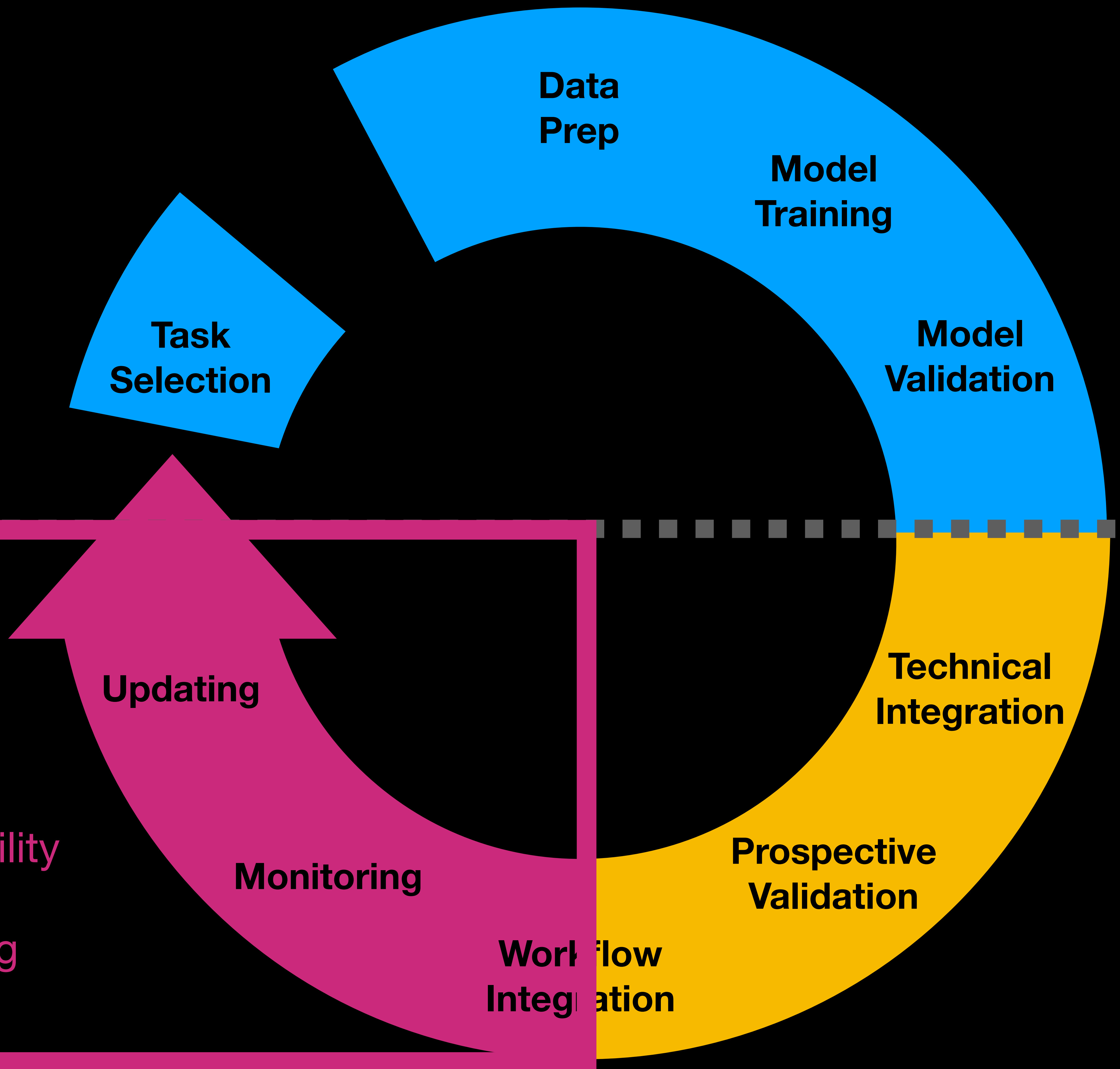
Ötleş *et al.*, MLHC '23

Performance Gap

Characterize changes to prospective performance after model implementation

Ötleş *et al.*, MLHC '21

What are we going to talk about?



RTW Prediction

Assess value of
longitudinal observations
for return to work prediction

Ötleş *et al.*, JAMIA '22

Performance Gap

Characterize changes
to prospective performance
after model implementation

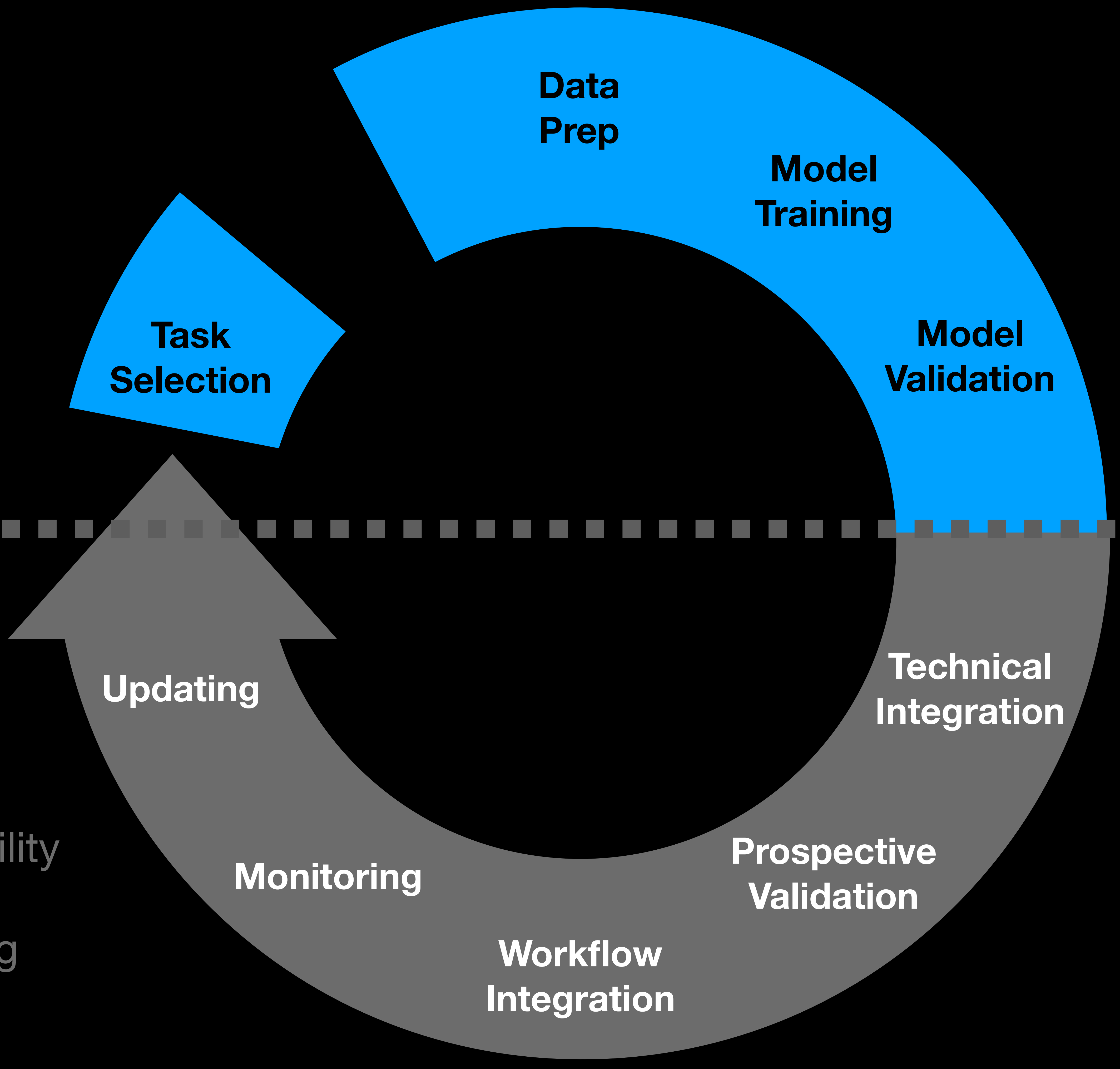
Ötleş *et al.*, MLHC '21

Updating Models

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

Ötleş *et al.*, MLHC '23

Let's dive in.



RTW Prediction

Assess value of longitudinal observations for return to work prediction

Ötleş *et al.*, JAMIA '22

Updating Models

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

Ötleş *et al.*, MLHC '23

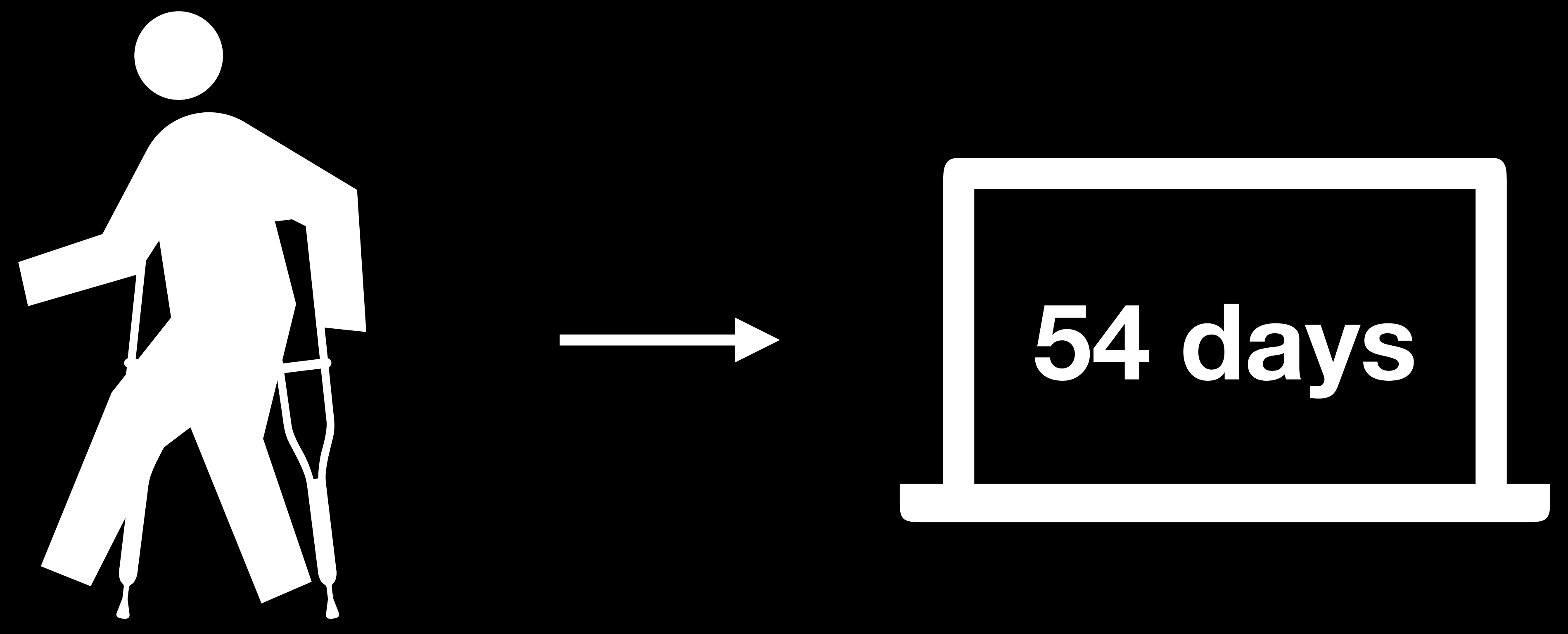
Performance Gap

Characterize changes to prospective performance after model implementation

Ötleş *et al.*, MLHC '21

Existing return to work models ignore longitudinal observations.

55 y.o. male
postal worker

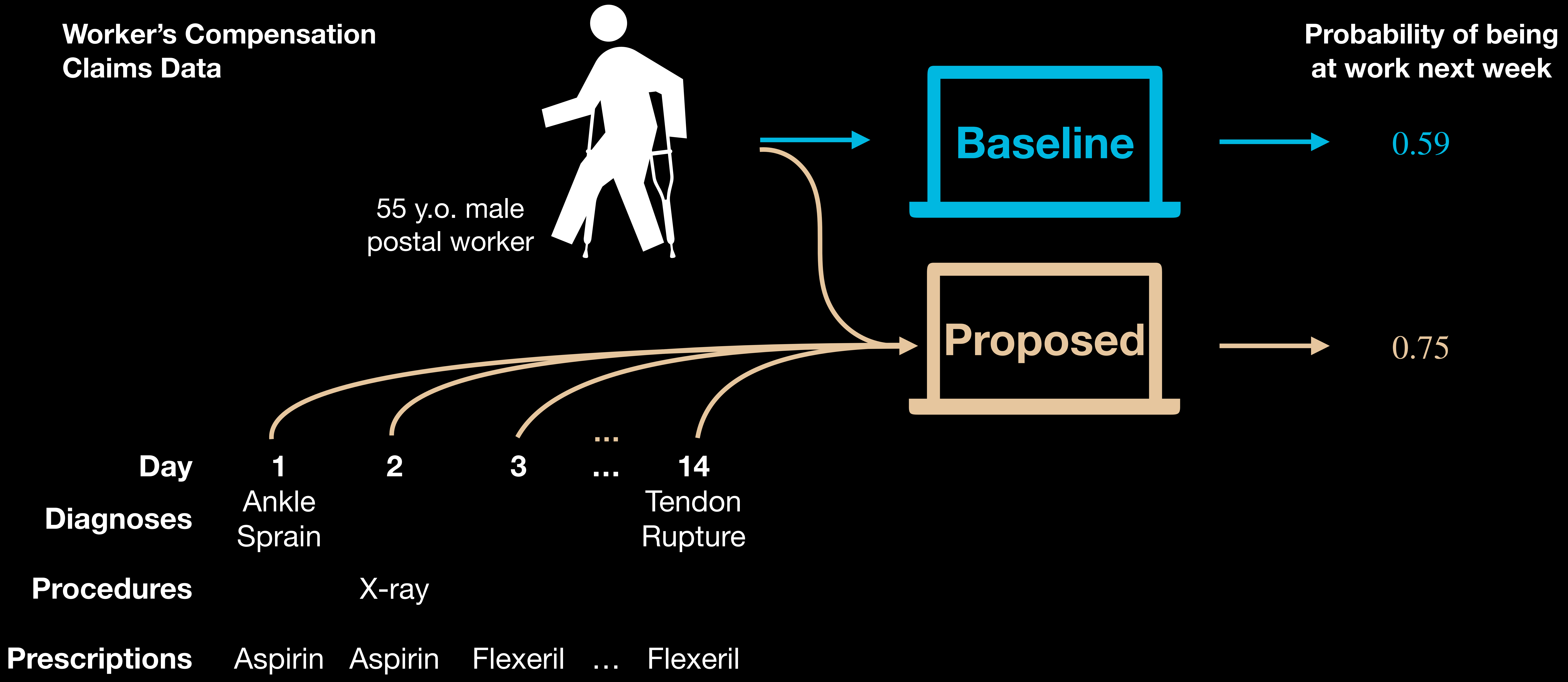


Q: *What is the value of longitudinal observations in return to work prediction?*

More specifically: when predicting future work-status do we observe a performance improvement when using longitudinal observations collected beyond the time of injury?

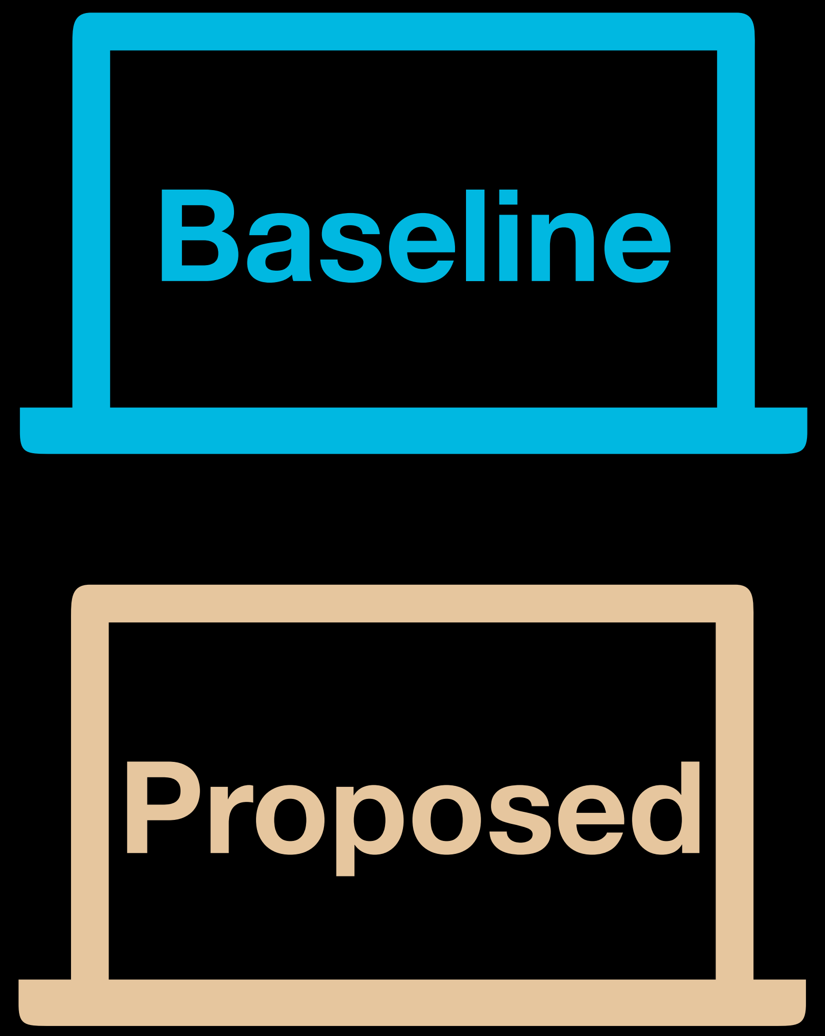
Hypothesis: Yes. Longitudinal observations improve predictions in other healthcare task.

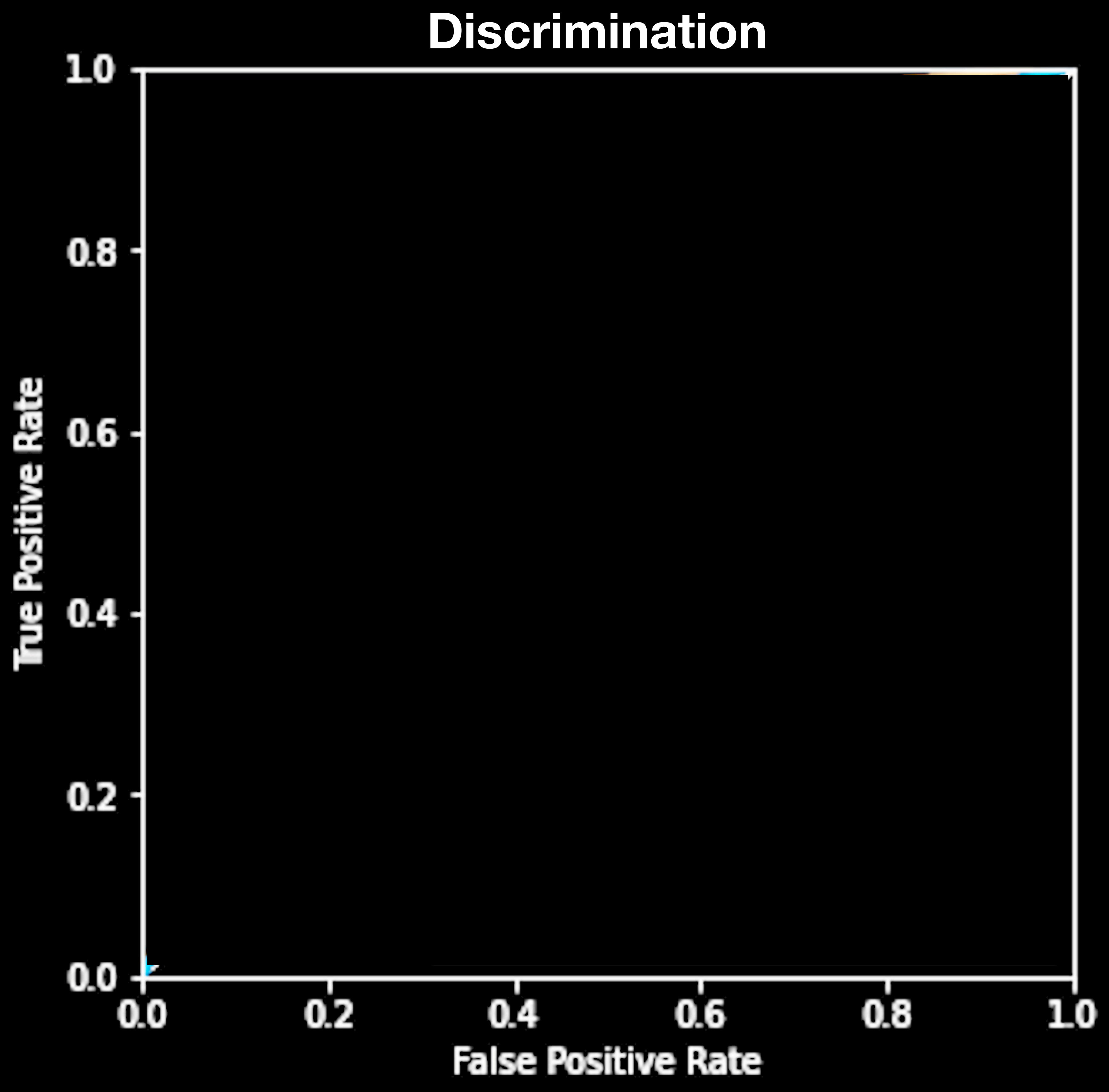
Experimental Setup

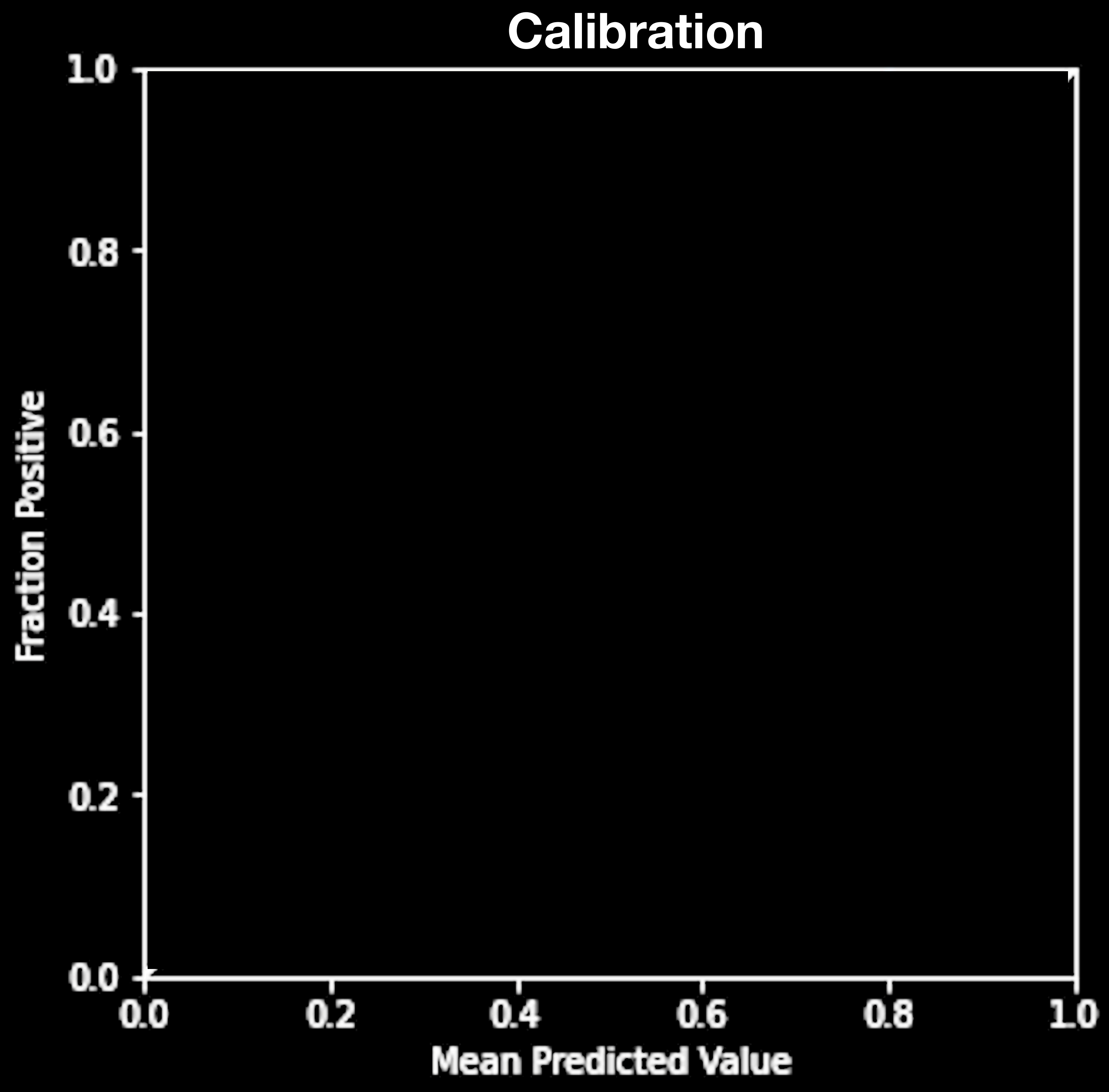


Ötleş, Seymour, Wang, Denton. Dynamic prediction of work status for workers with occupational injuries: assessing the value of longitudinal observations. *Journal of the American Medical Informatics Association.* 2022 [In Press]

Results

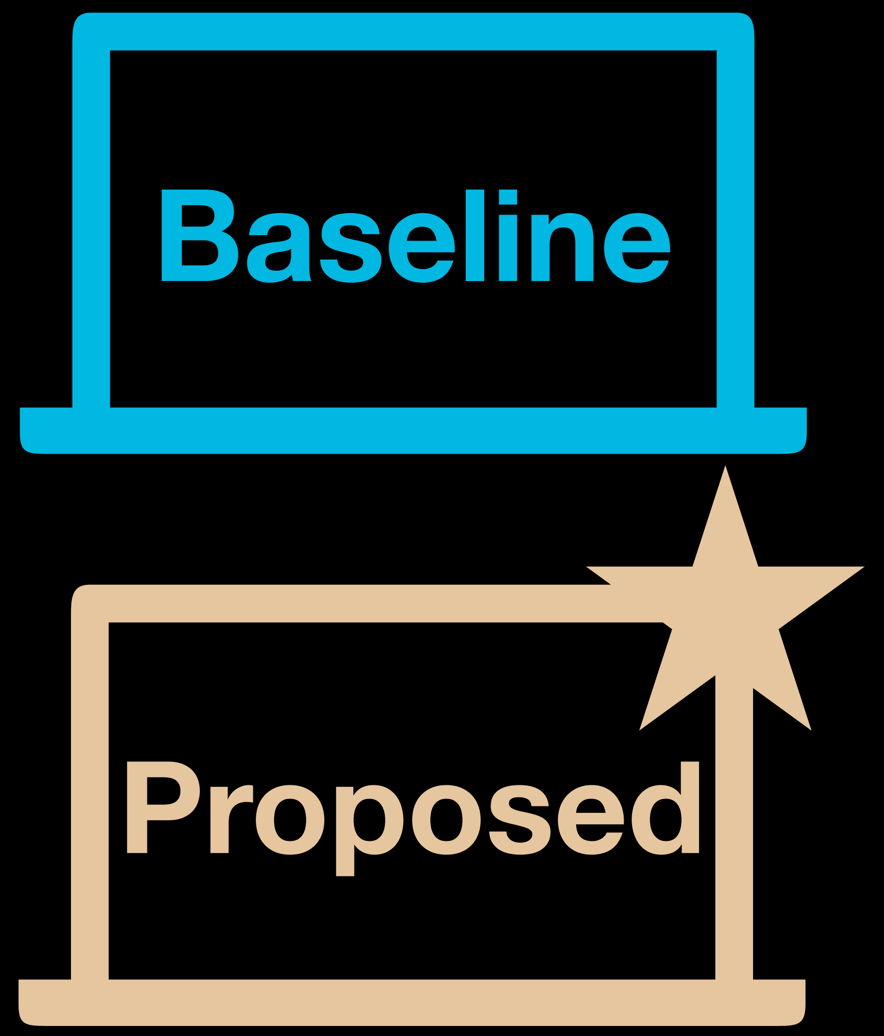


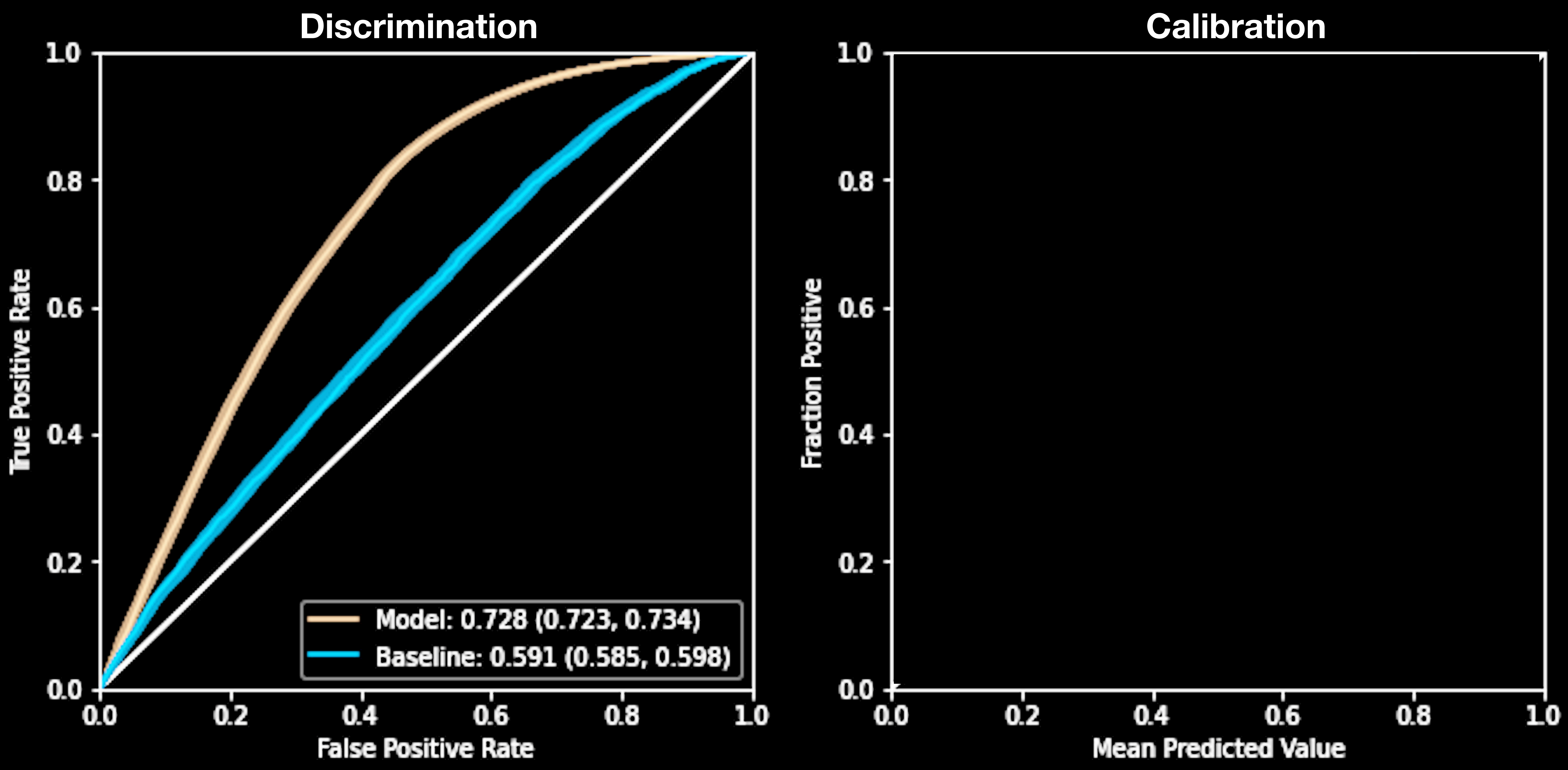




Ötleş, Seymour, Wang, Denton. Dynamic prediction of work status for workers with occupational injuries: assessing the value of longitudinal observations. *Journal of the American Medical Informatics Association.* 2022 [In Press]

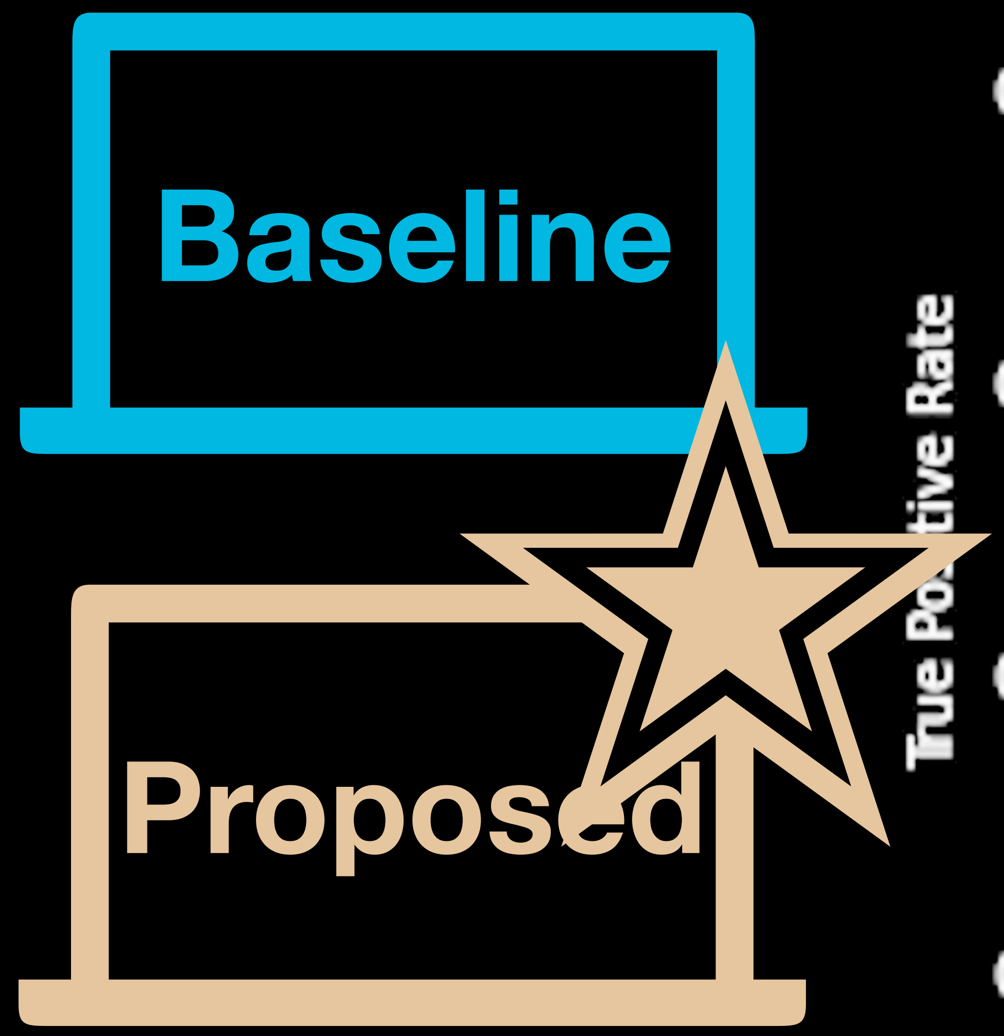
Results

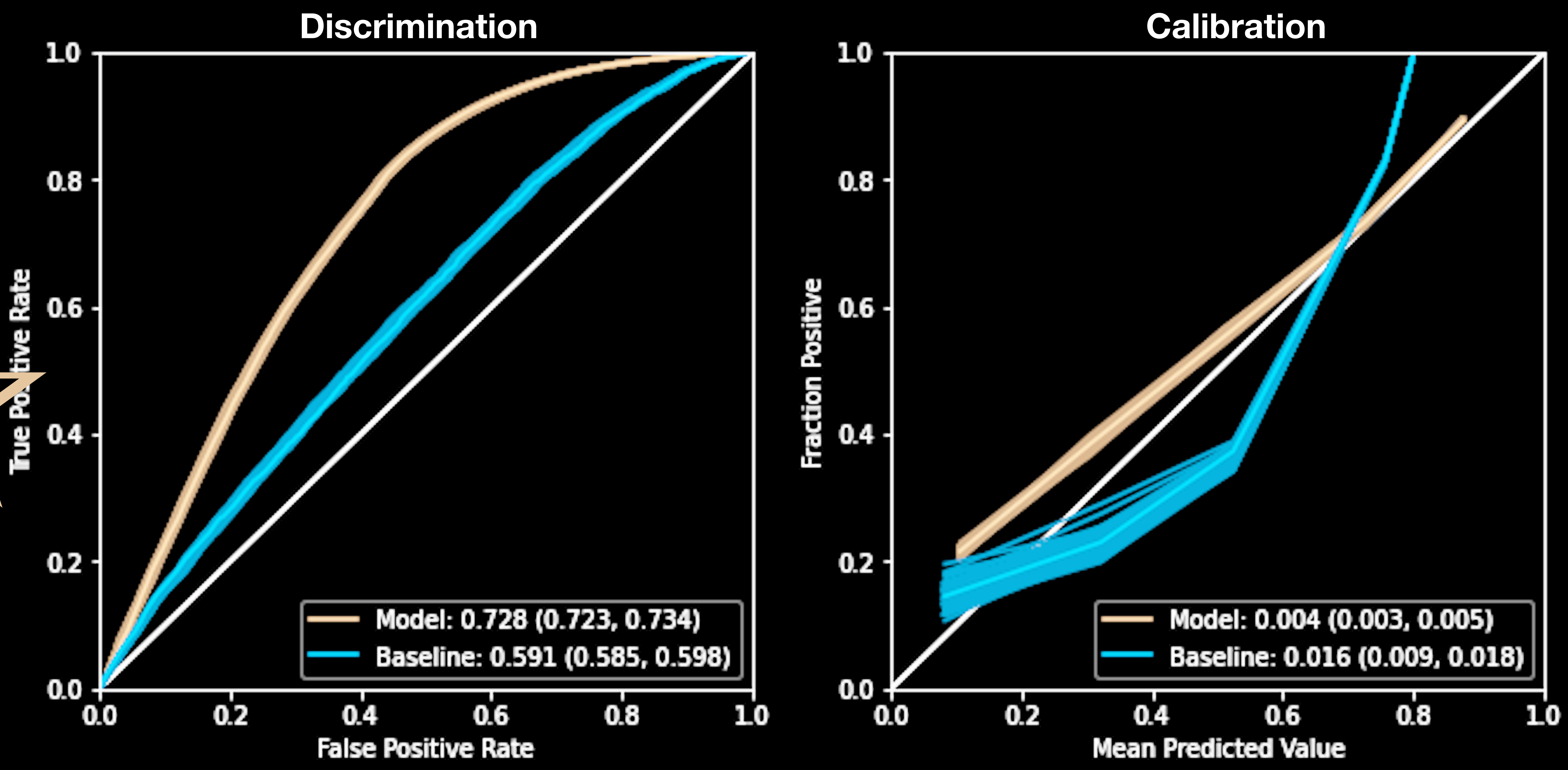




Ötleş, Seymour, Wang, Denton. Dynamic prediction of work status for workers with occupational injuries: assessing the value of longitudinal observations. *Journal of the American Medical Informatics Association.* 2022 [In Press]

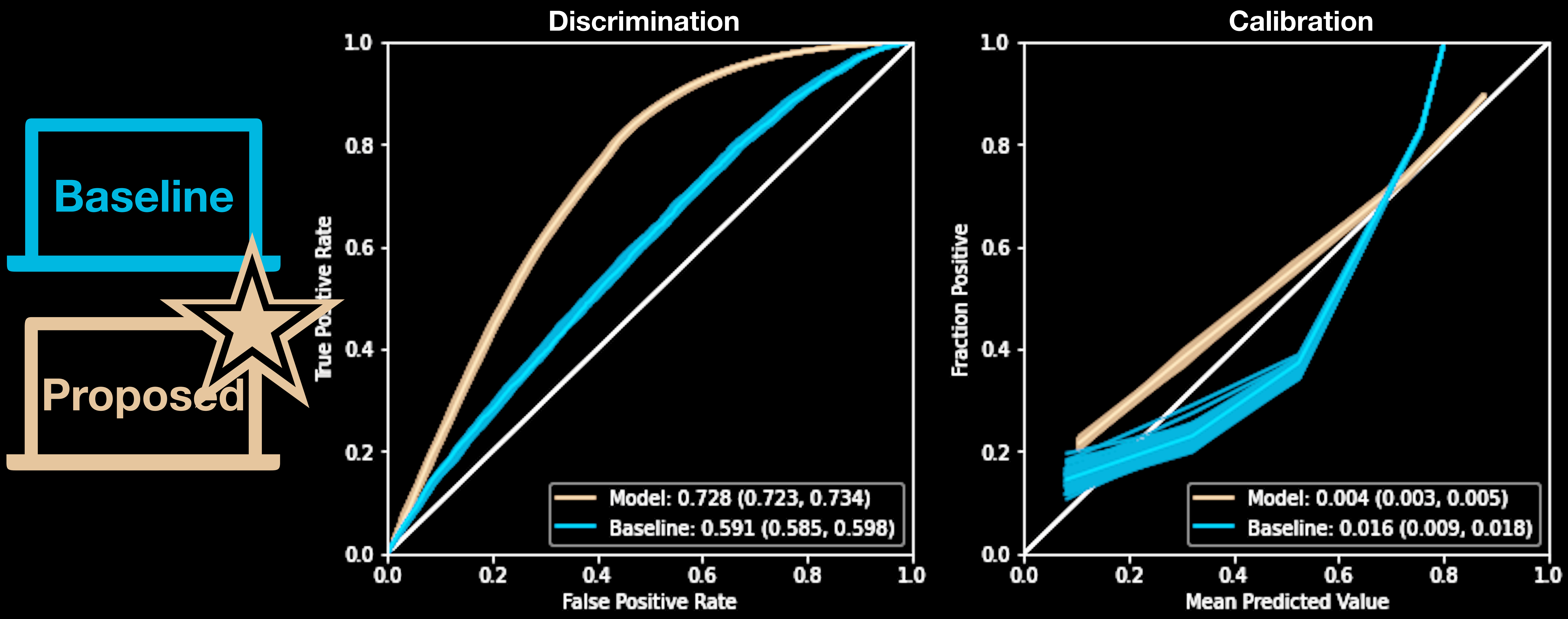
Results





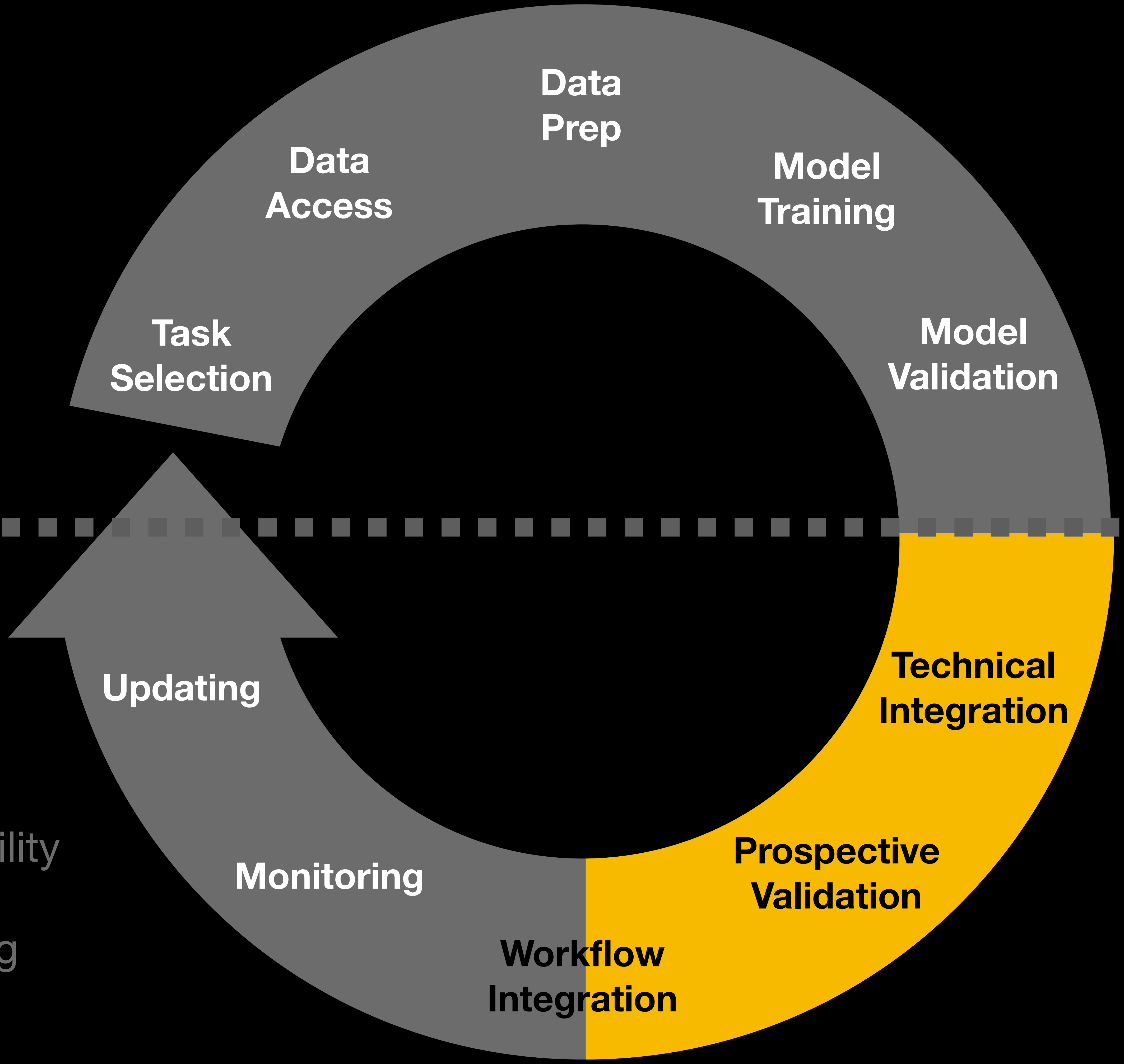
Ötleş, Seymour, Wang, Denton. Dynamic prediction of work status for workers with occupational injuries: assessing the value of longitudinal observations. *Journal of the American Medical Informatics Association.* 2022 [In Press]

Main Takeaway from RTW Prediction.



Return to work prediction benefits from longitudinal observations.

On to the next one.



RTW Prediction

Assess value of longitudinal observations for return to work prediction

Ötleş *et al.*, JAMIA ‘22

Updating Models

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

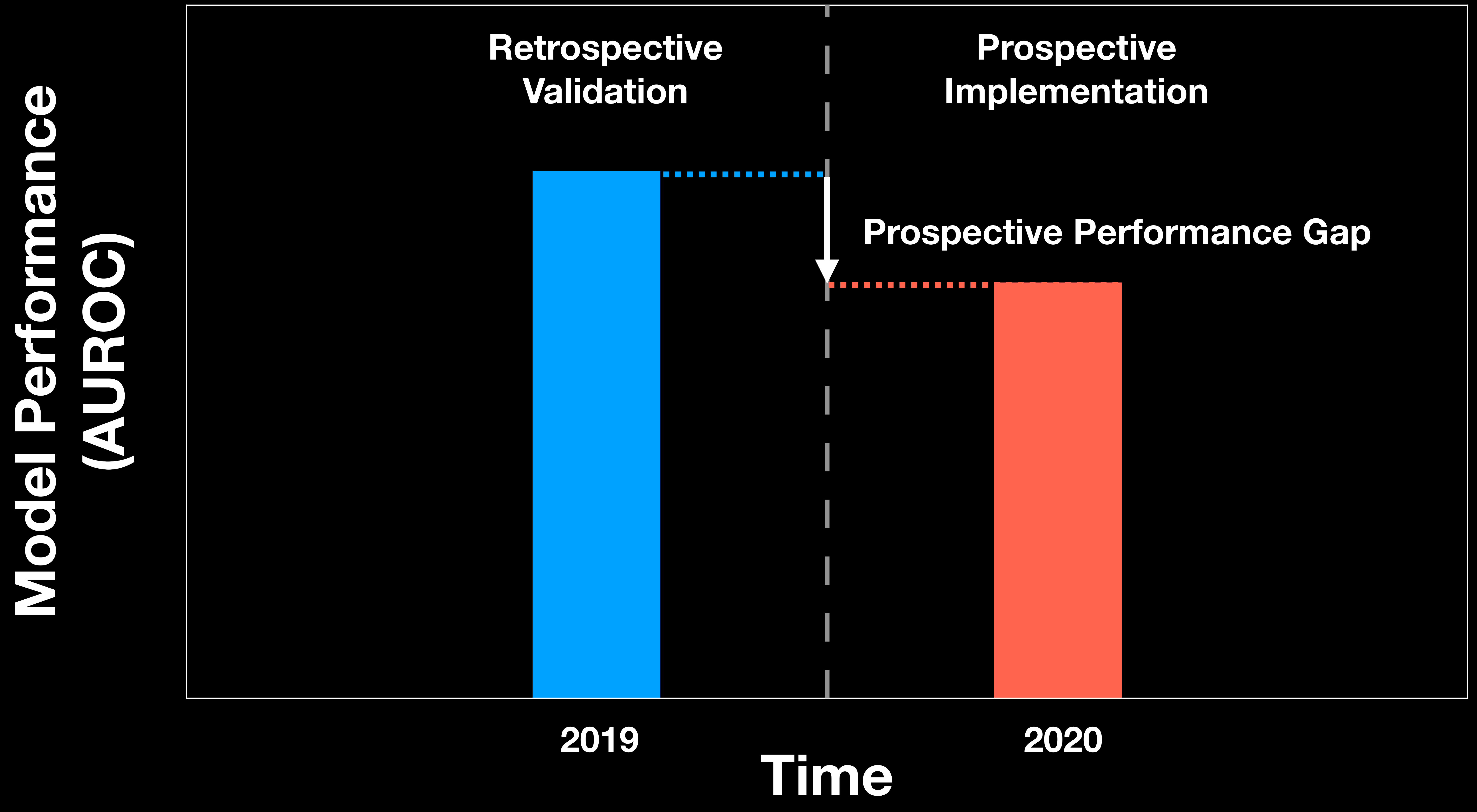
Ötleş *et al.*, MLHC '23

Performance Gap

Characterize changes to prospective performance after model implementation

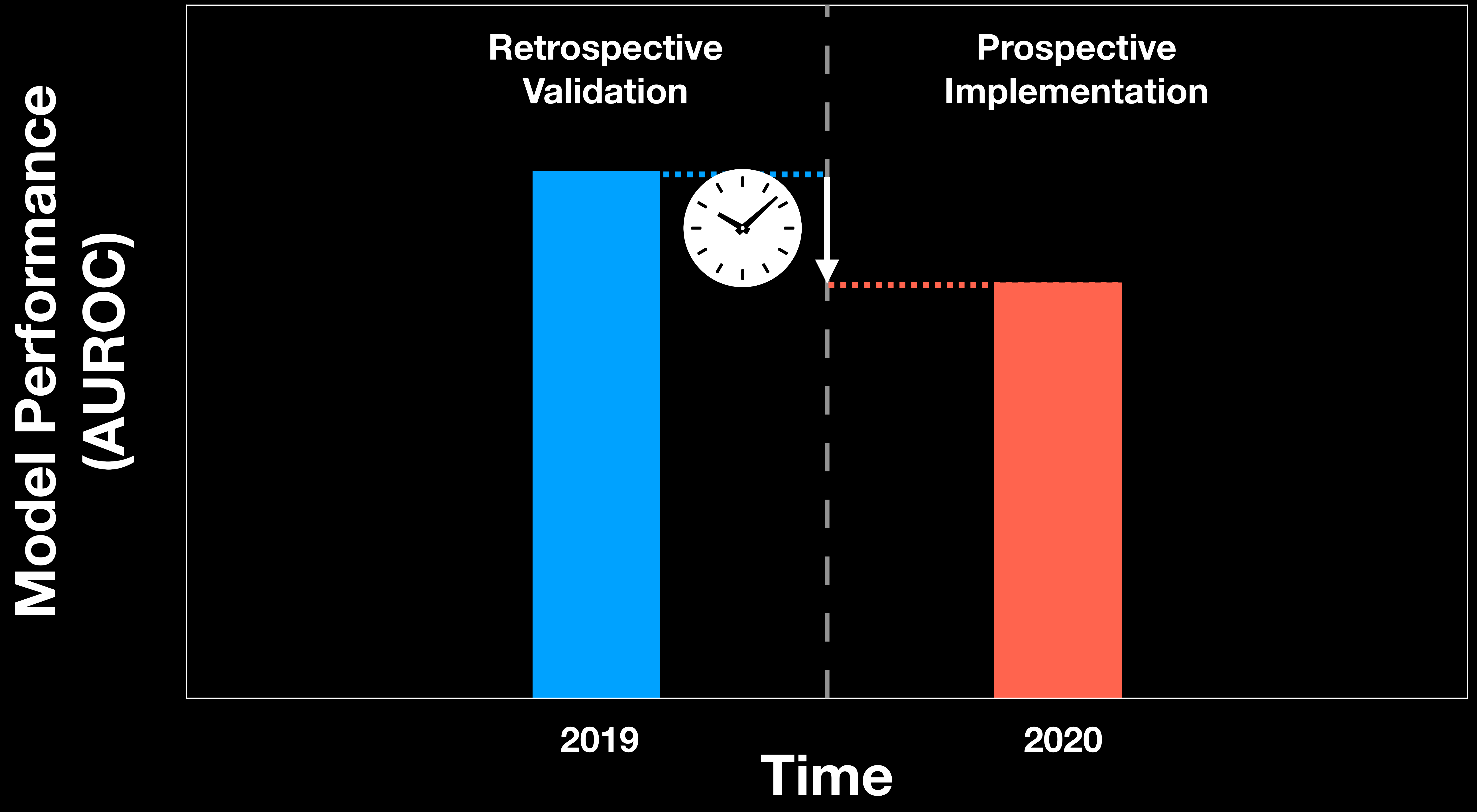
Ötleş *et al.*, MLHC ‘21

Model performance may degrade after implementation.



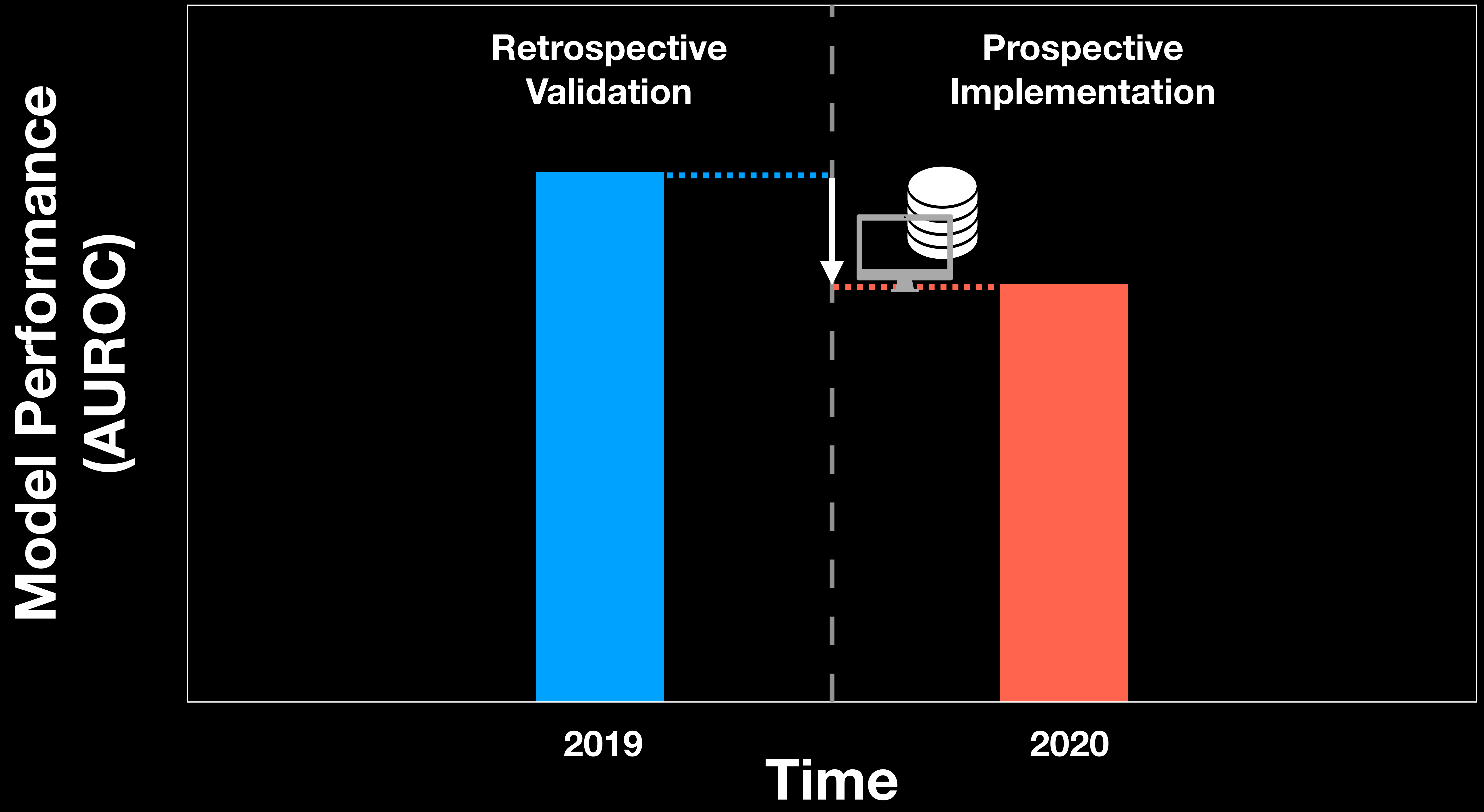
Davis 2017, Hickey 2013, Minne 2012

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

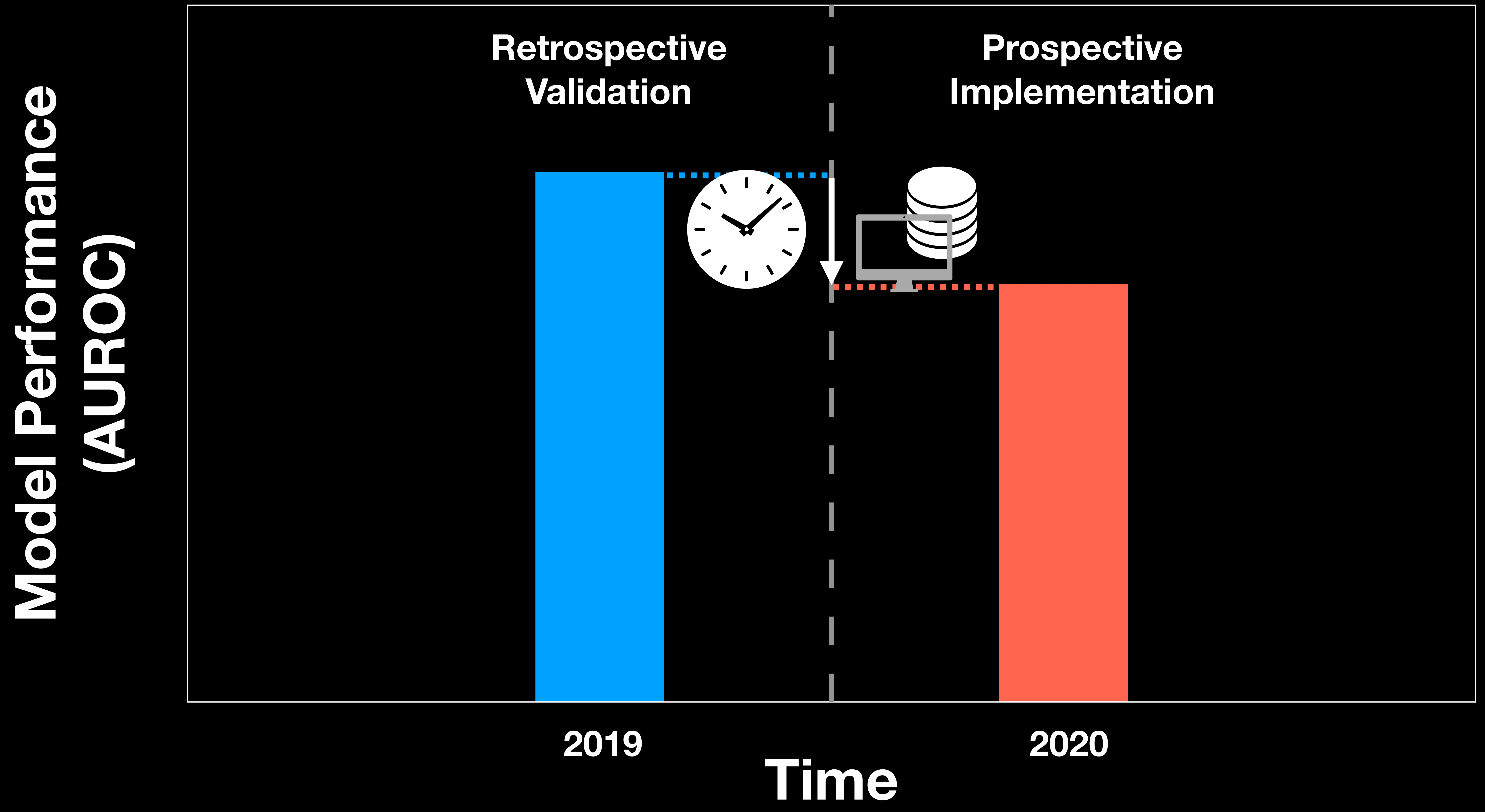


Davis 2017, Hickey 2013, Minne 2012

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



Degradation due to temporal & infrastructure shift.

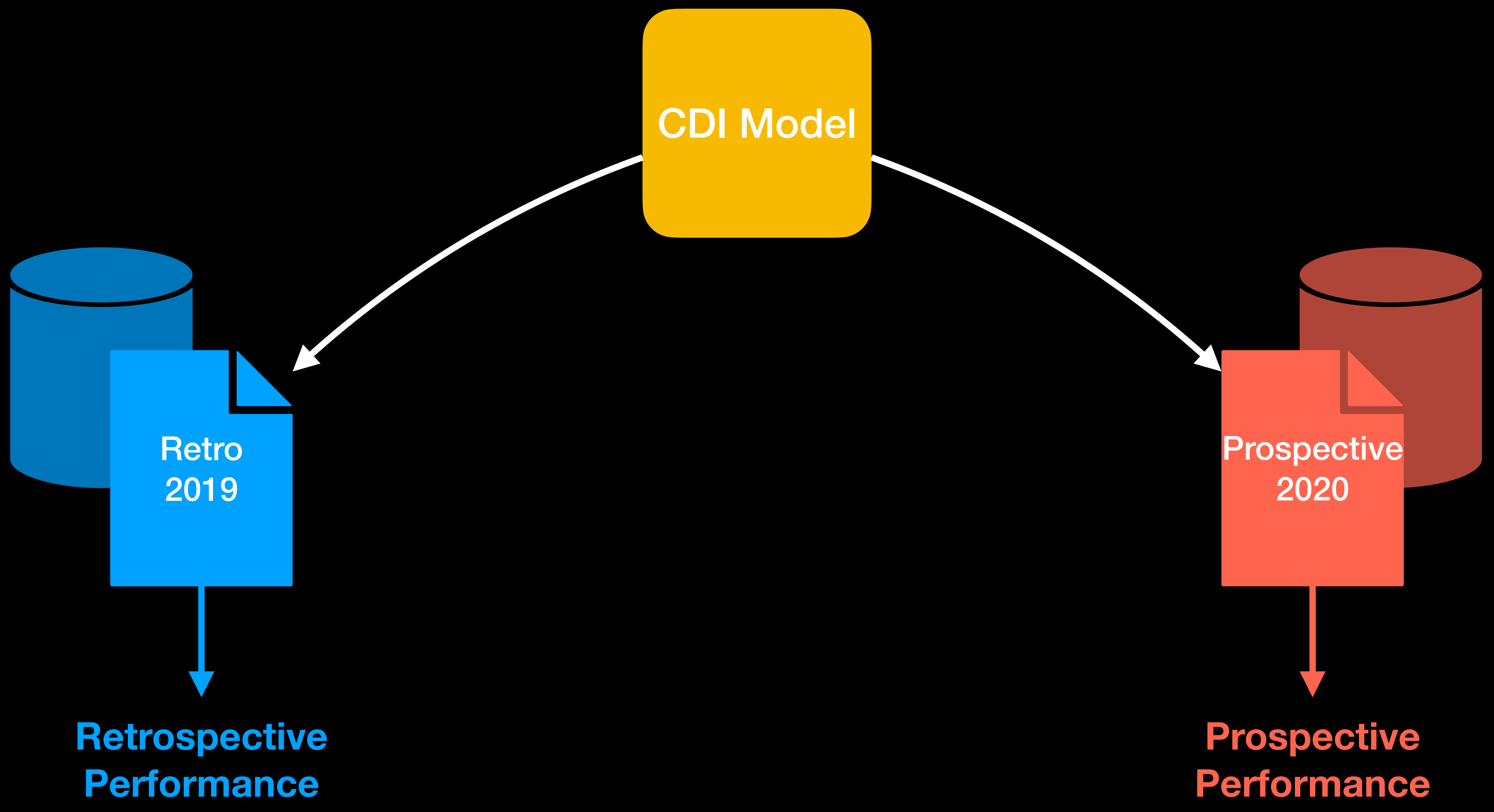


Q: *What are the causes of the prospective performance gap?*

More specifically: what are the temporal and infrastructure shift contributions to the prospective performance gap of a prospectively implemented model?

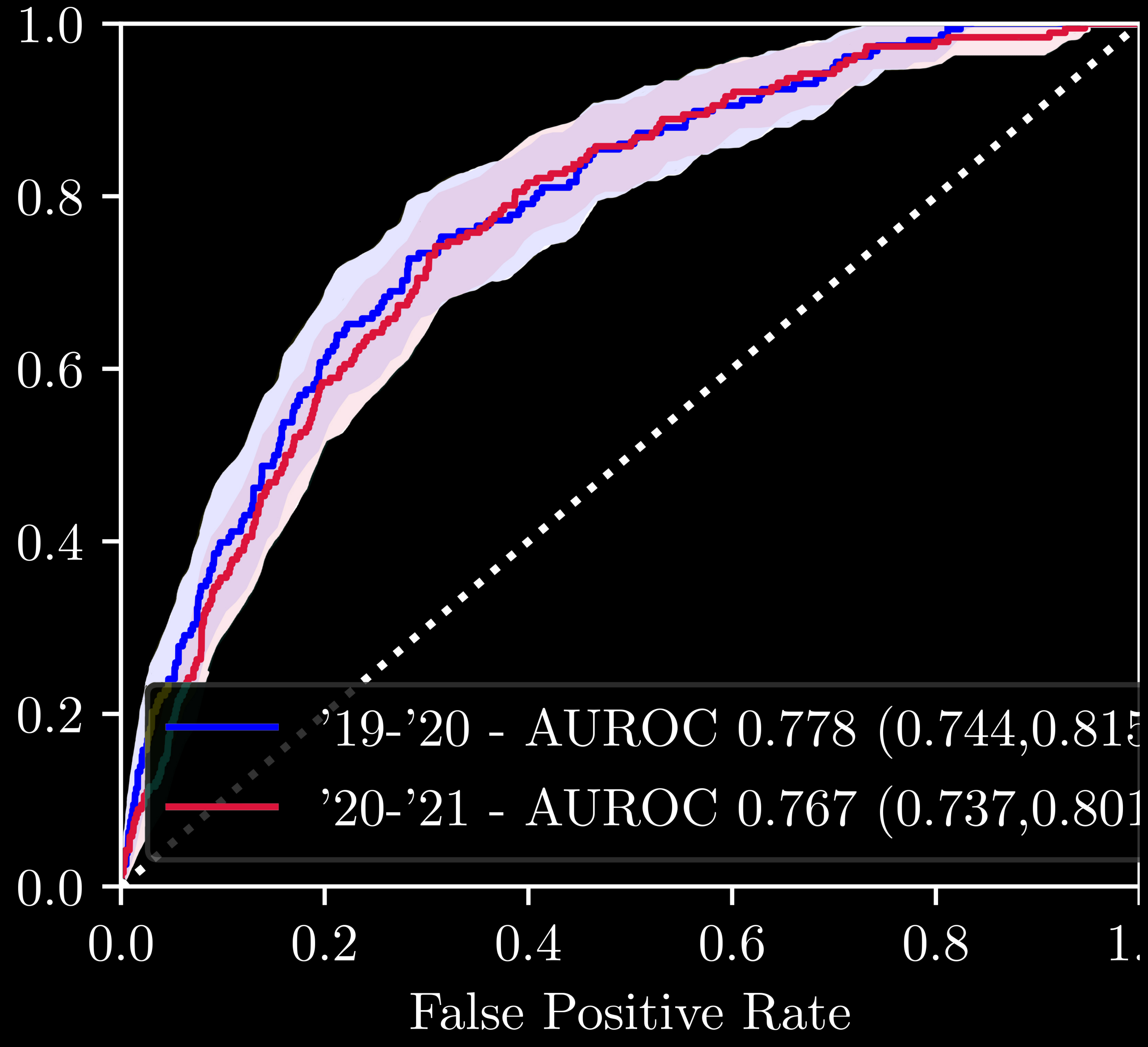
Hypothesis: Infrastructure will be an important component of the prospective performance gap.

Experimental Setup



Ötleş, Oh, Li, Bochinski, Joo, Ortwine, Shenoy, Washer, Young, Rao, Wiens.
Mind the performance gap: examining dataset shift during prospective validation.
Machine Learning for Healthcare Conference. 2021.

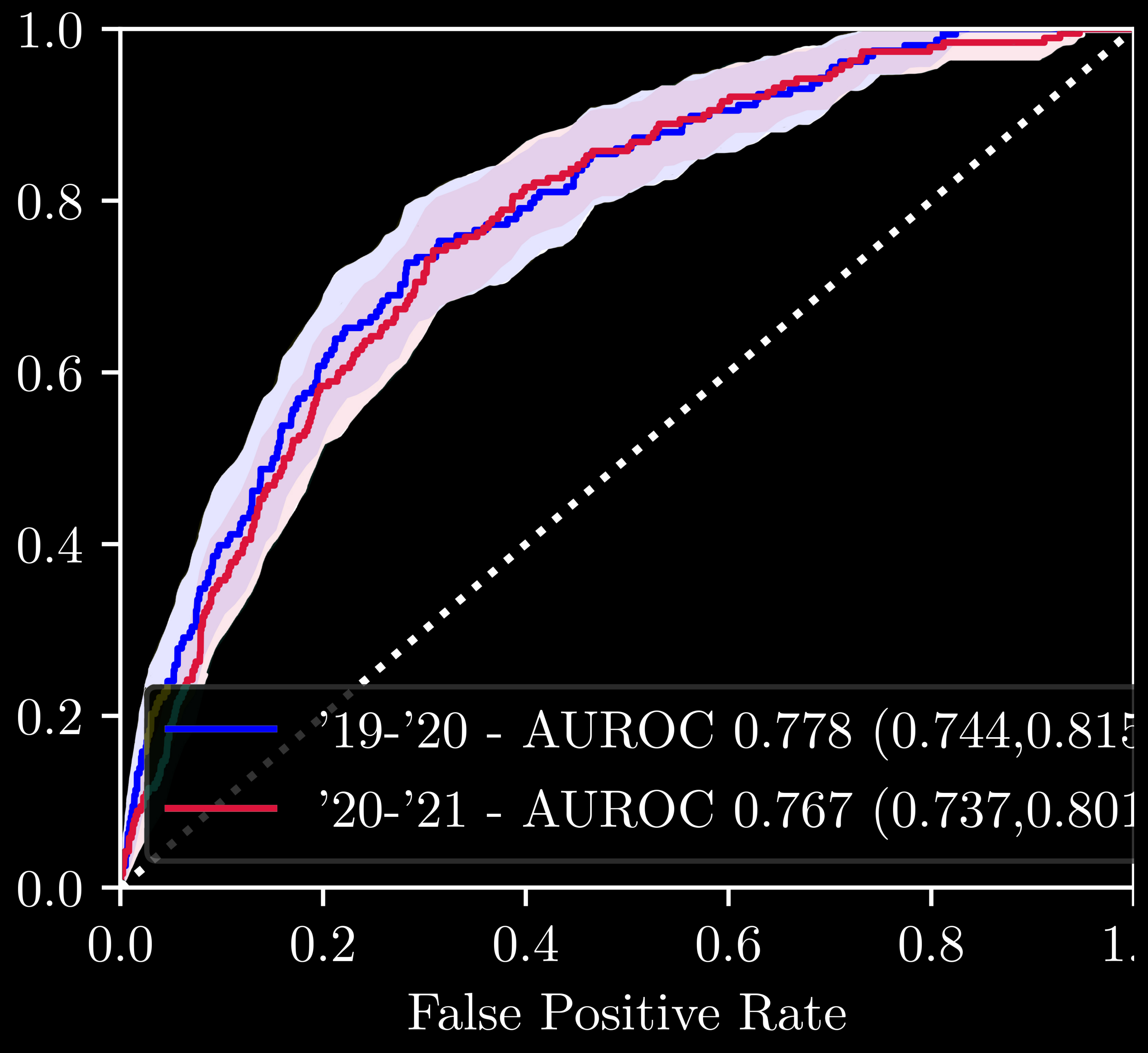
Results



Ötleş, Oh, Li, Bochinski, Joo, Ortwine, Shenoy, Washer, Young, Rao, Wiens.
Mind the performance gap: examining dataset shift during prospective validation.
Machine Learning for Healthcare Conference. 2021.

Results

Performance Gap: 0.011

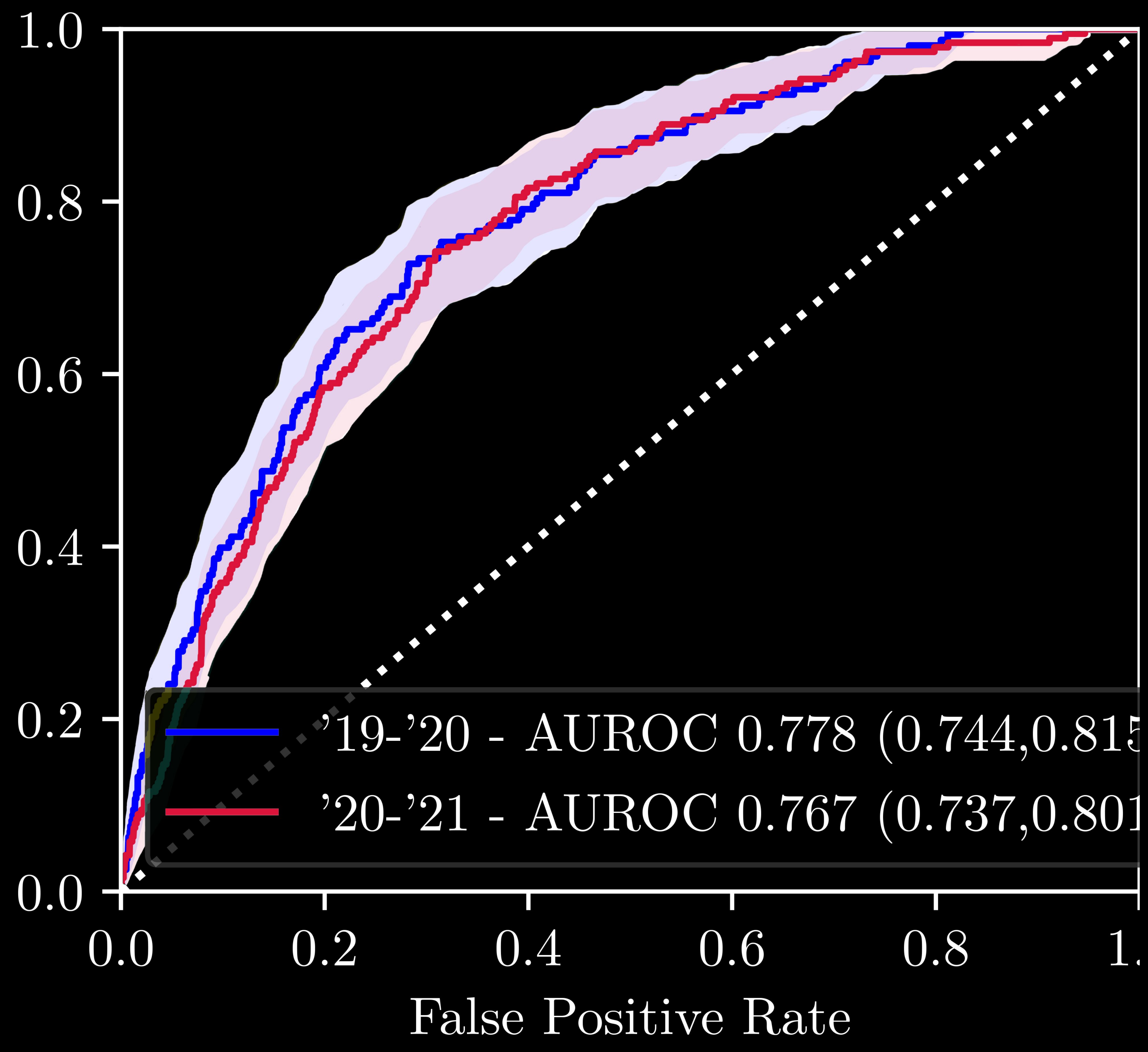


Ötleş, Oh, Li, Bochinski, Joo, Ortwine, Shenoy, Washer, Young, Rao, Wiens.
Mind the performance gap: examining dataset shift during prospective validation.
Machine Learning for Healthcare Conference. 2021.

Results

Performance Gap: 0.011

Temporal Shift: -0.005



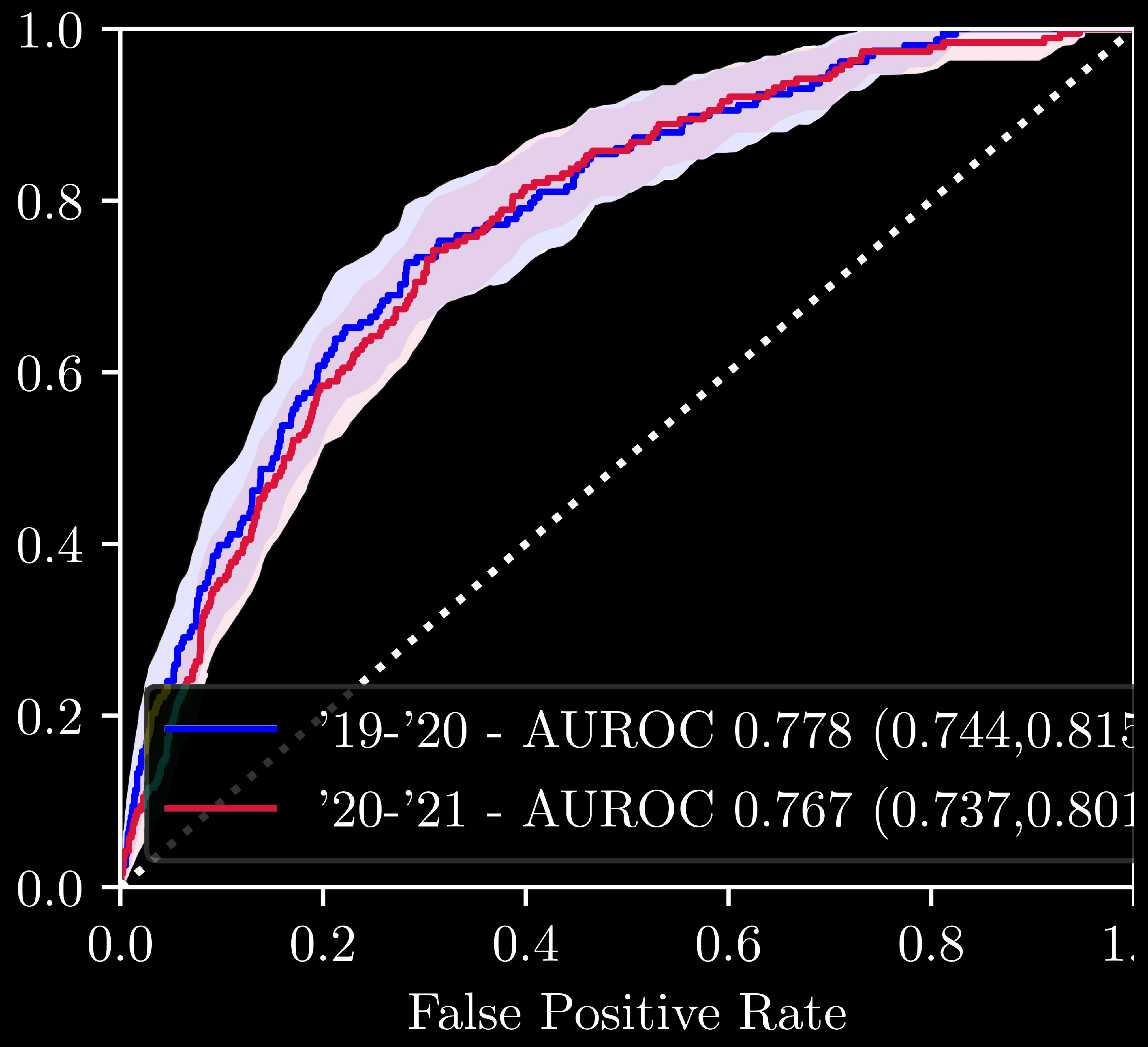
Ötleş, Oh, Li, Bochinski, Joo, Ortwine, Shenoy, Washer, Young, Rao, Wiens.
Mind the performance gap: examining dataset shift during prospective validation.
Machine Learning for Healthcare Conference. 2021.

Results

Performance Gap: 0.011

Temporal Shift: -0.005

Infrastructure Shift: 0.016



Ötleş, Oh, Li, Bochinski, Joo, Ortwine, Shenoy, Washer, Young, Rao, Wiens.
Mind the performance gap: examining dataset shift during prospective validation.
Machine Learning for Healthcare Conference. 2021.

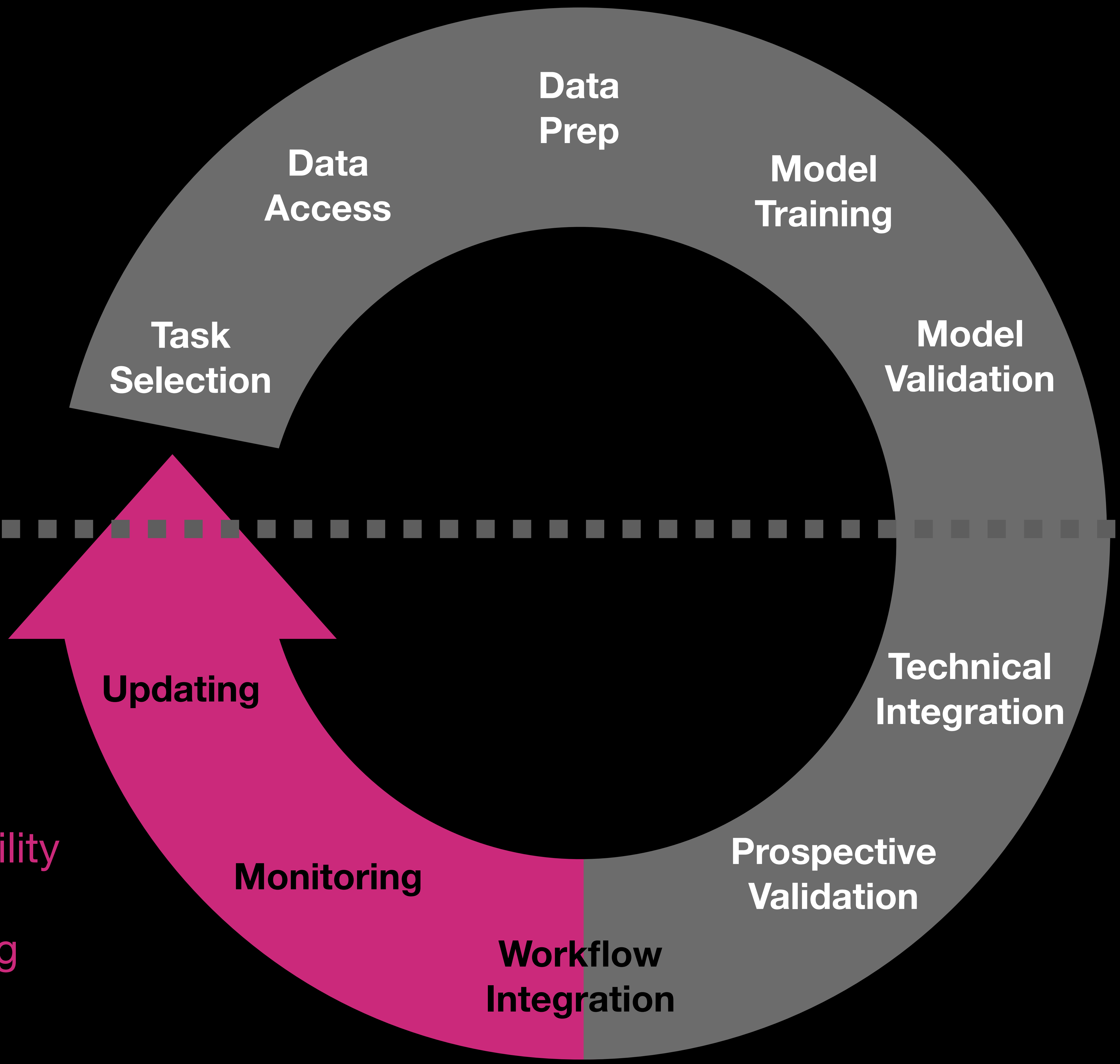
Infrastructure Matters.

Observed prospective performance gap primarily driven by infrastructure shift:

- source & processing differ, requires infrastructure updates
- timing of access differs, *when* you access the data affects *what* you access

Infrastructure shift can be mitigated, but first it must be recognized!

Last but not least.



RTW Prediction

Assess value of longitudinal observations for return to work prediction

Ötleş *et al.*, JAMIA '22

Updating Models

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

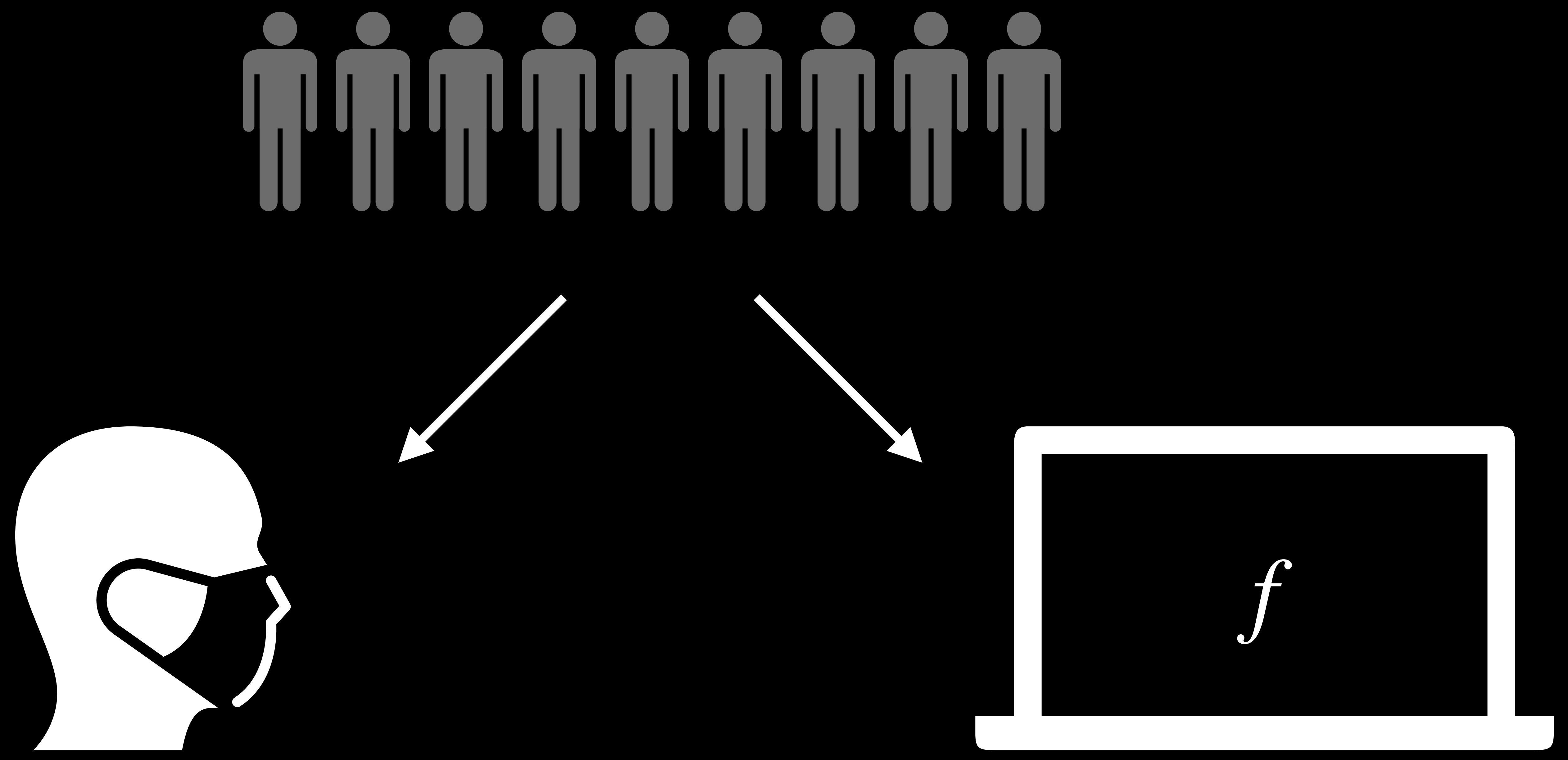
Ötleş *et al.*, MLHC '23

Performance Gap

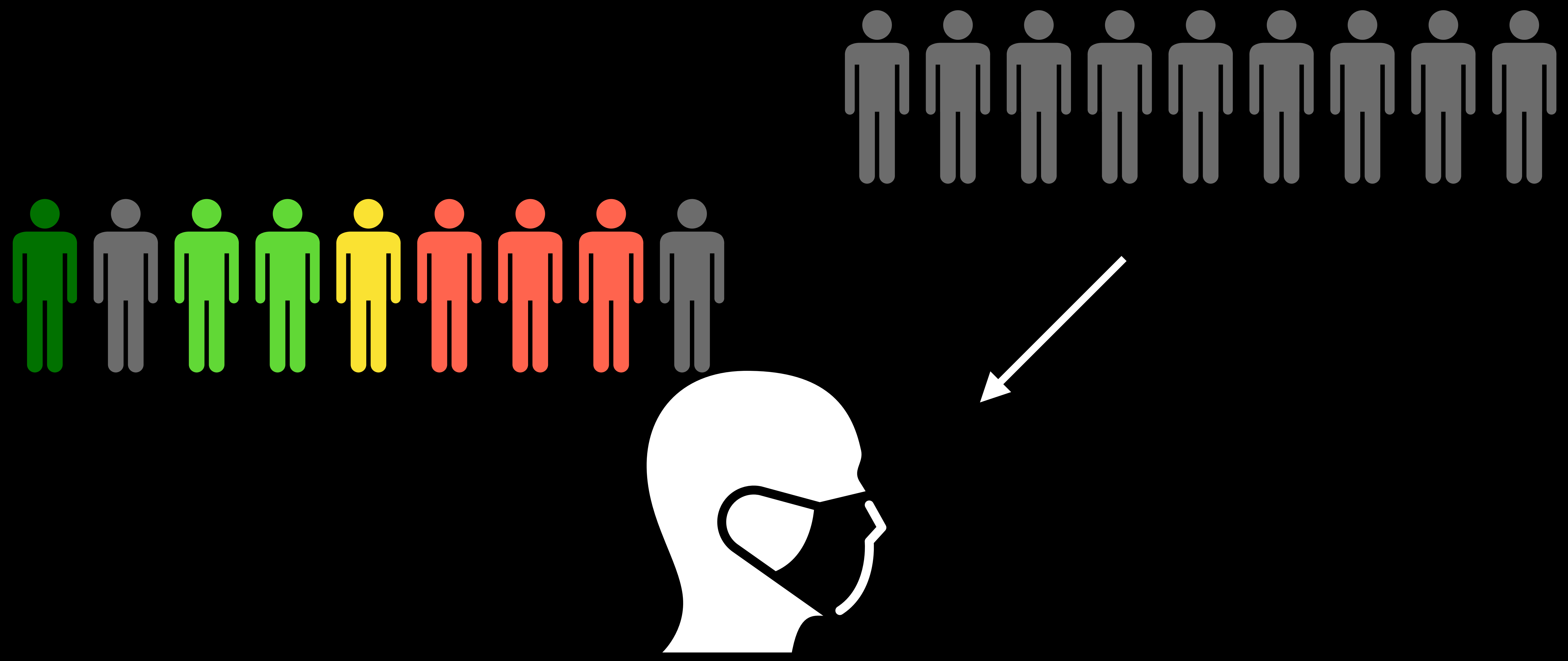
Characterize changes to prospective performance after model implementation

Ötleş *et al.*, MLHC '21

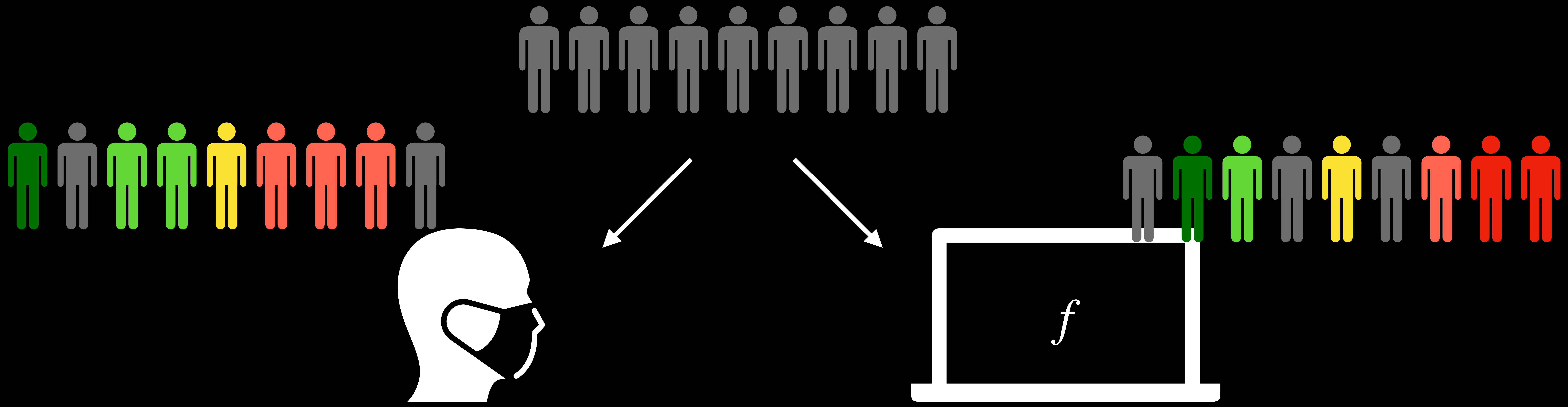
Physicians and models function as a team in healthcare settings.



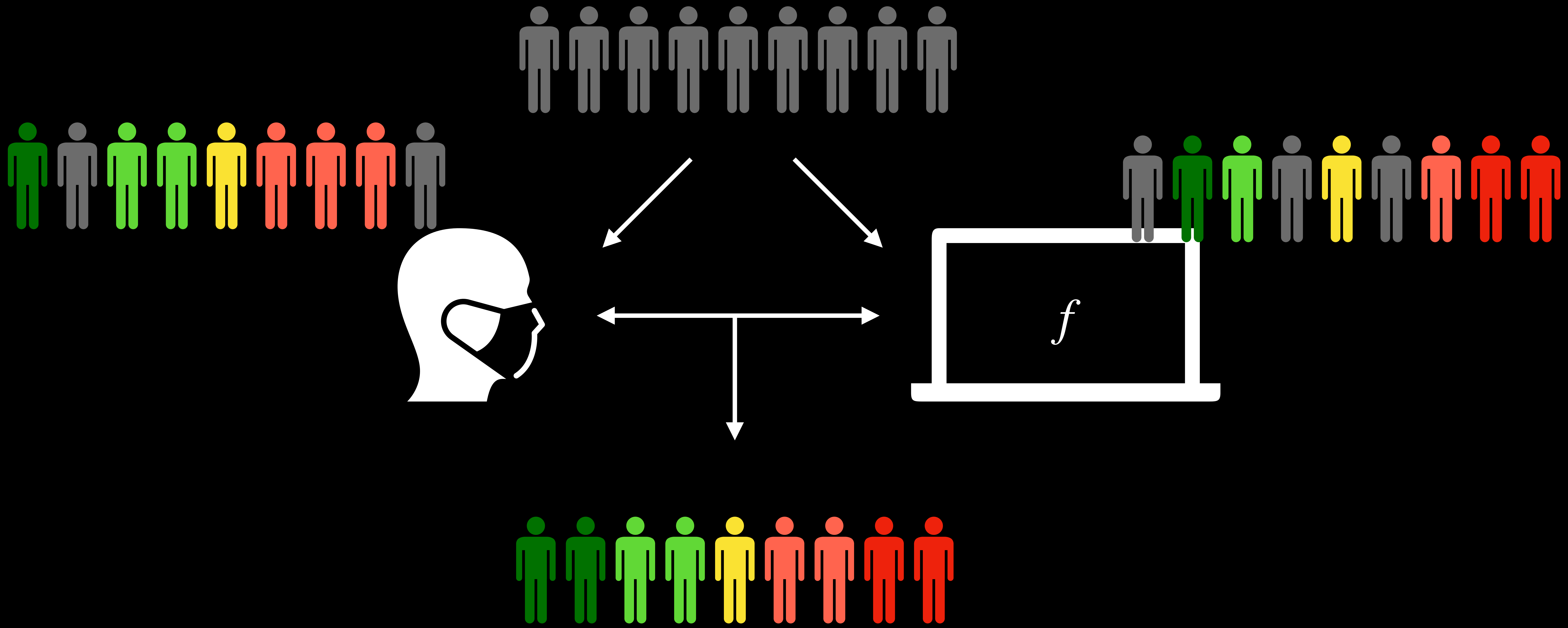
Physicians and models function as a team in healthcare settings.



Physicians and models function as a team in healthcare settings.



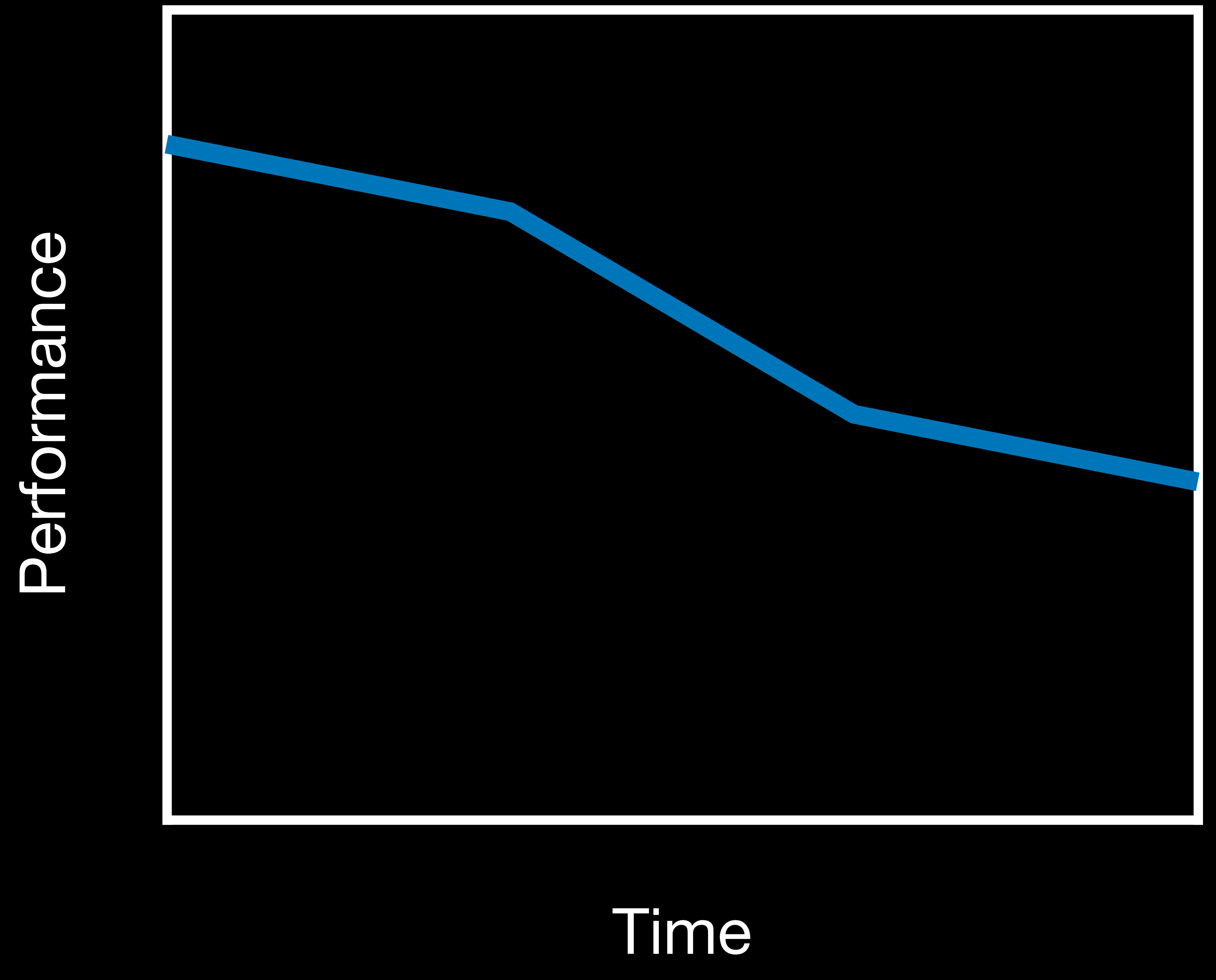
Physicians and models function as a team in healthcare settings.



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

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

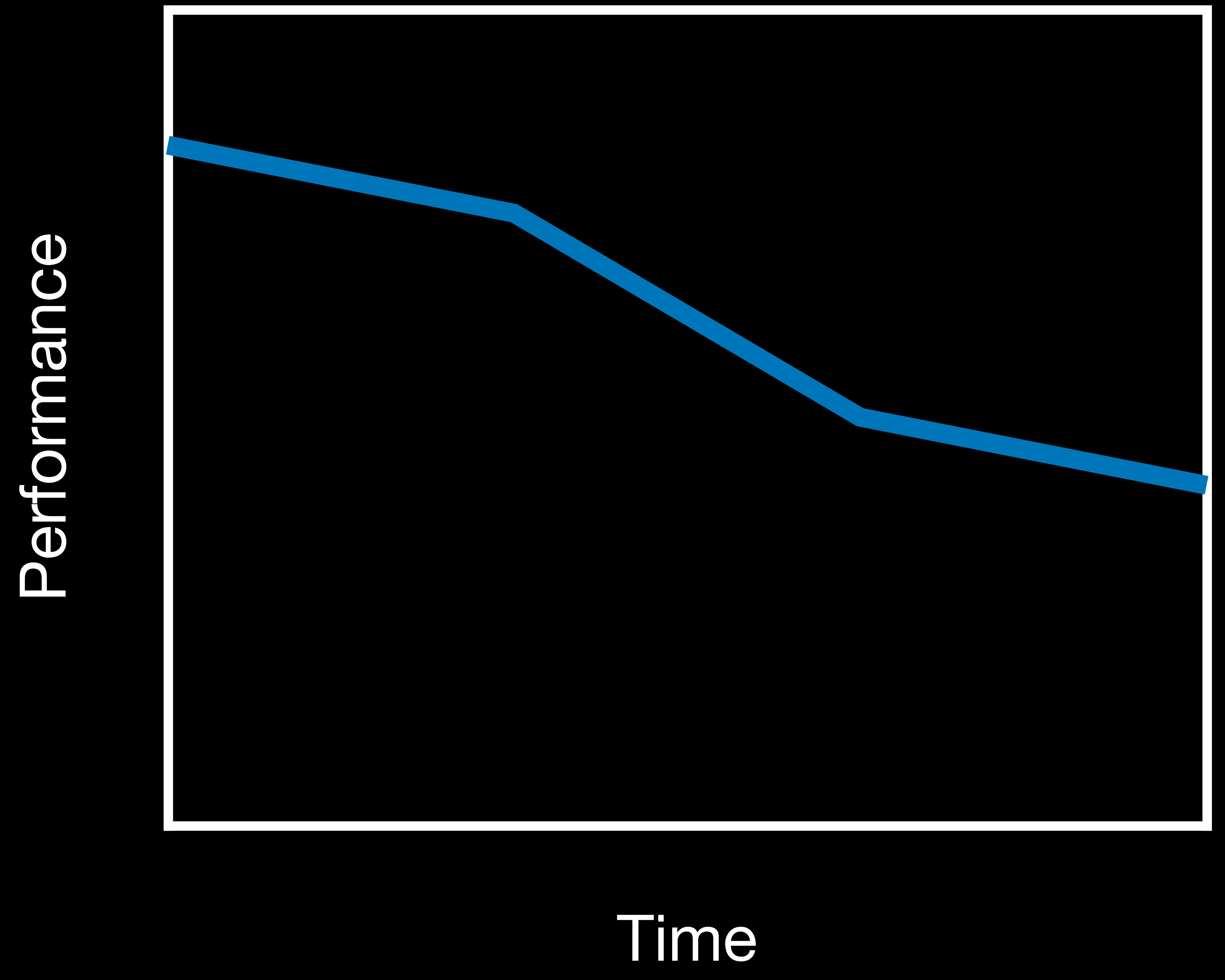
Counter performance degradation



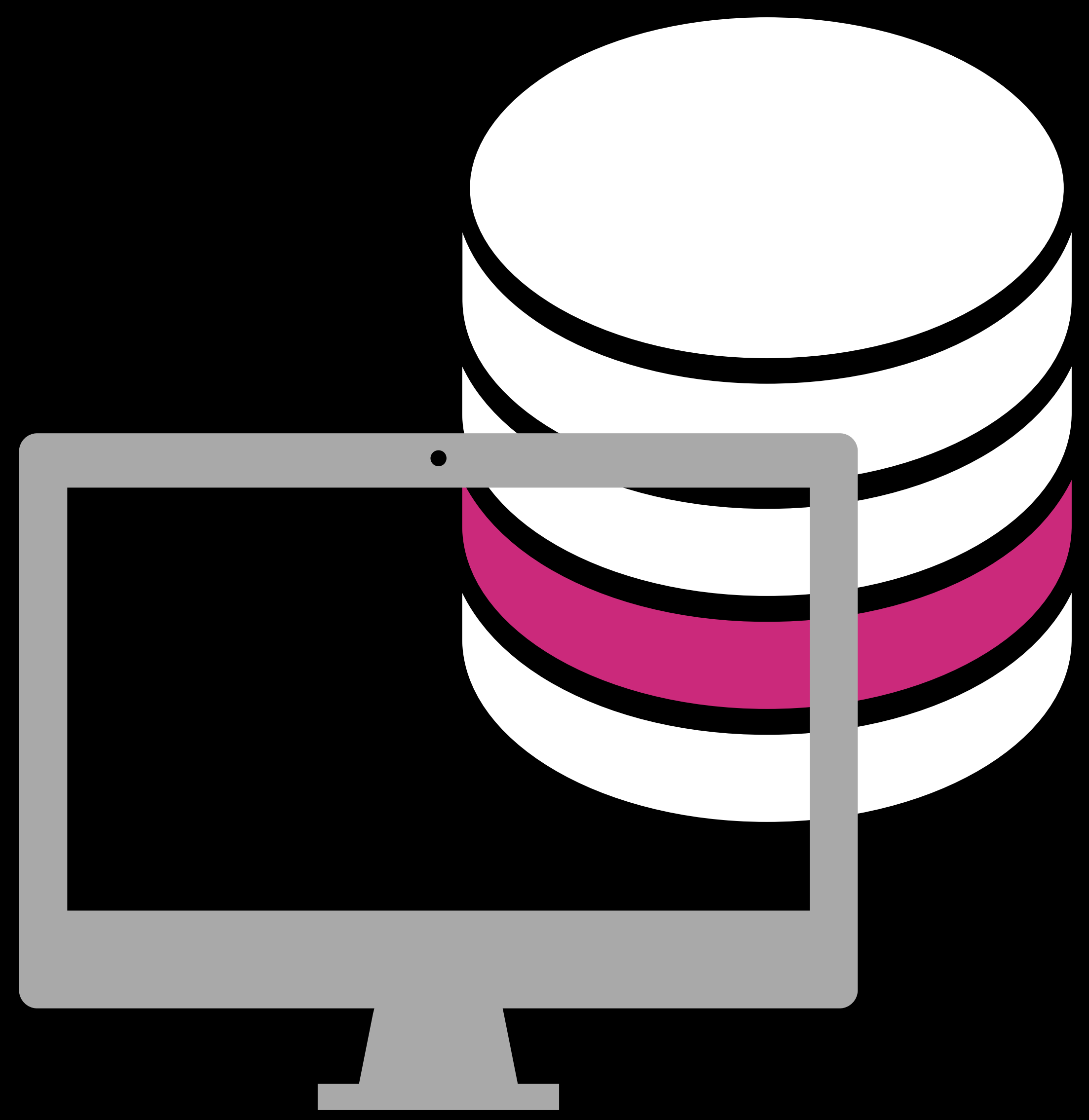
Finlayson 2021, David 2017, Hickey 2012, Minne 2012, Bansal 2019

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

Counter performance degradation



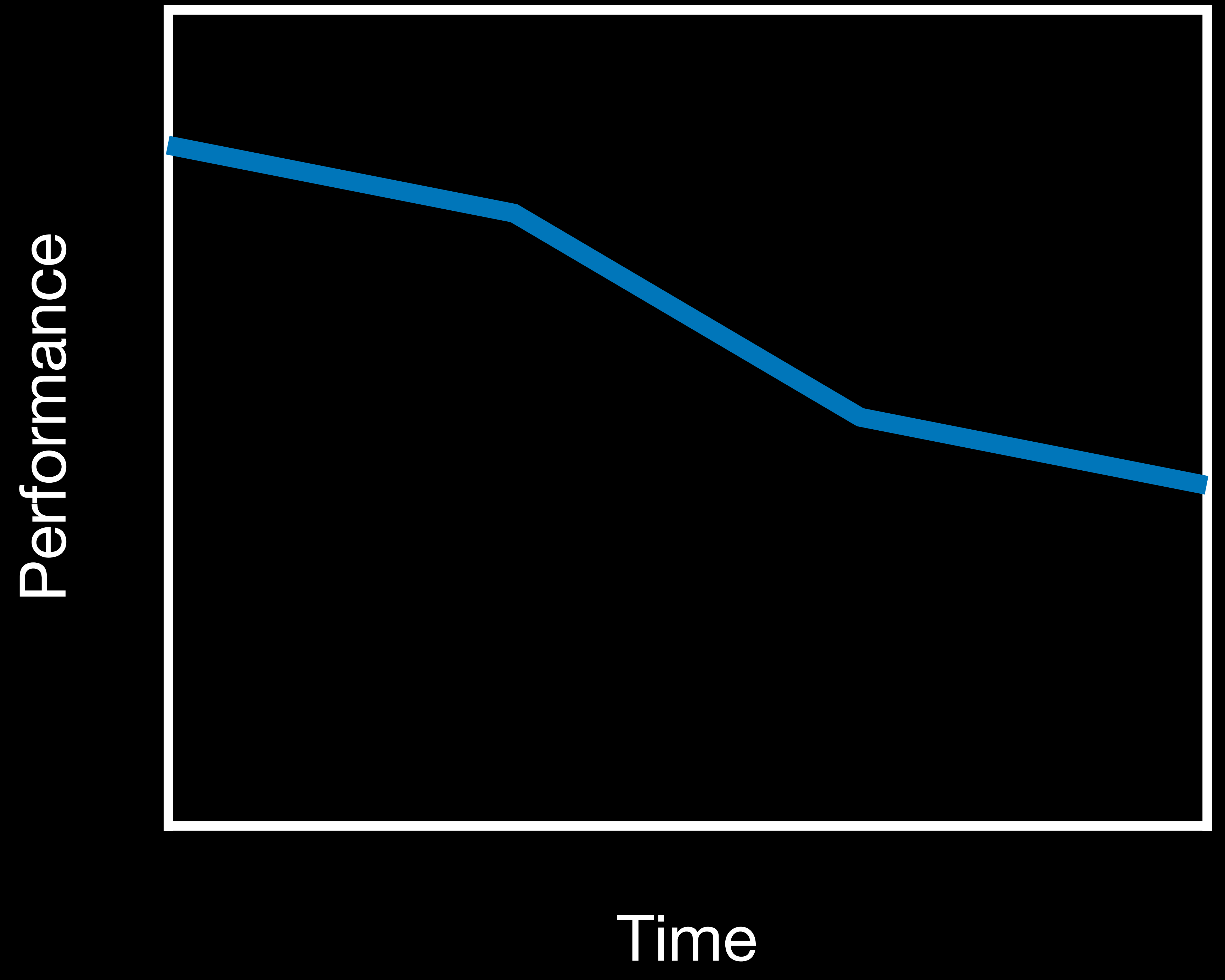
Adapt to infrastructure changes



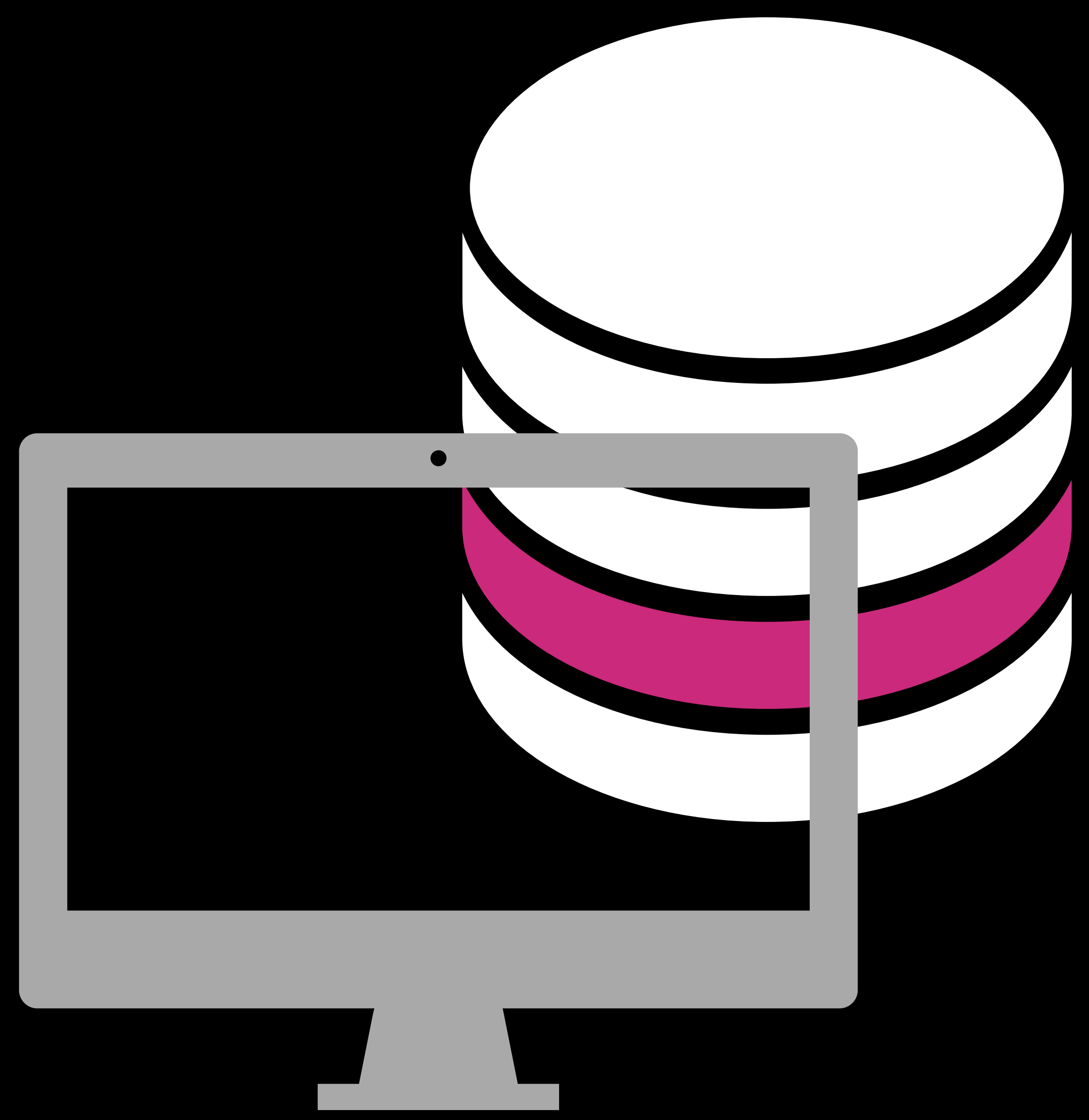
Finlayson 2021, David 2017, Hickey 2012, Minne 2012, Bansal 2019

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

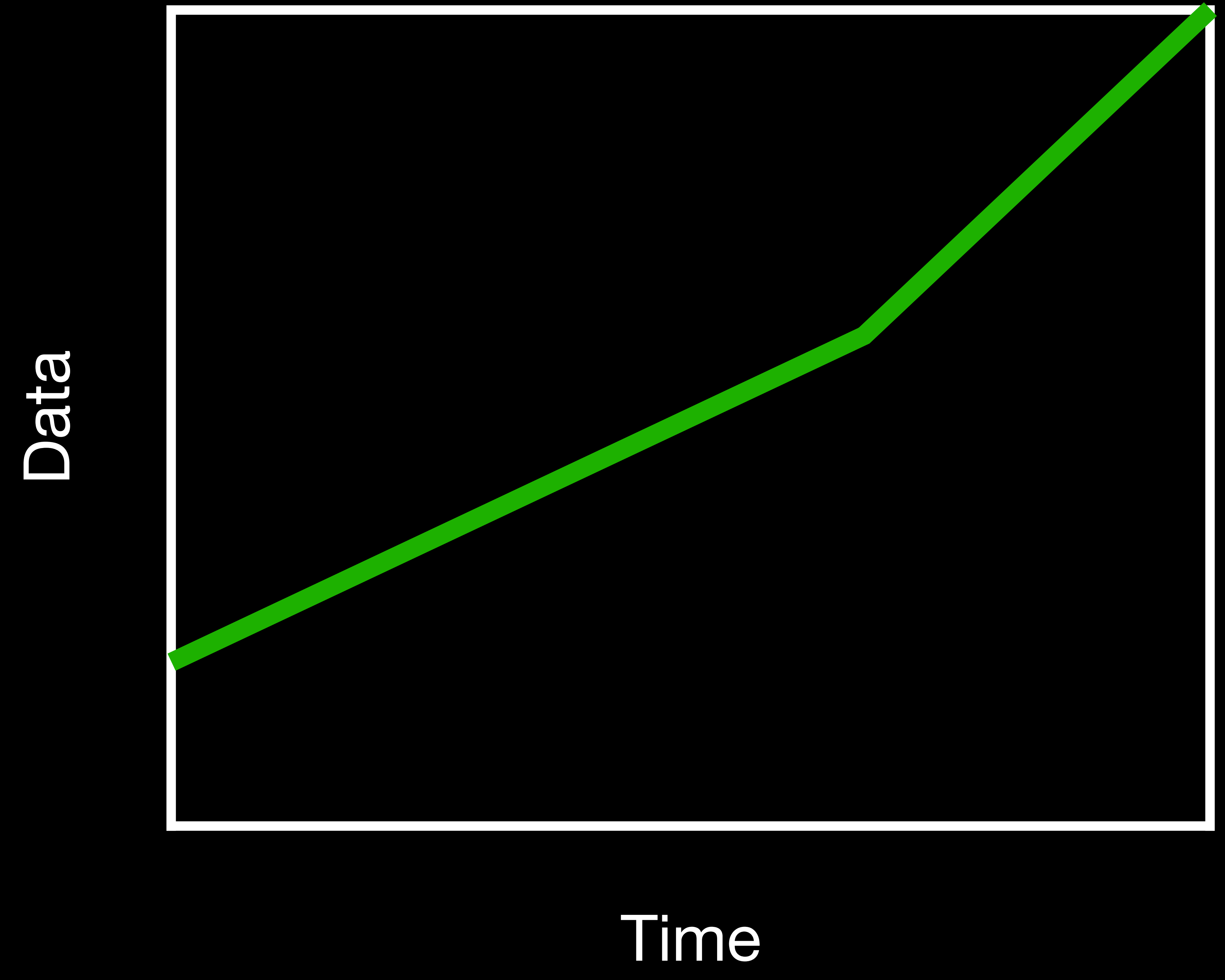
Counter performance degradation



Adapt to infrastructure changes

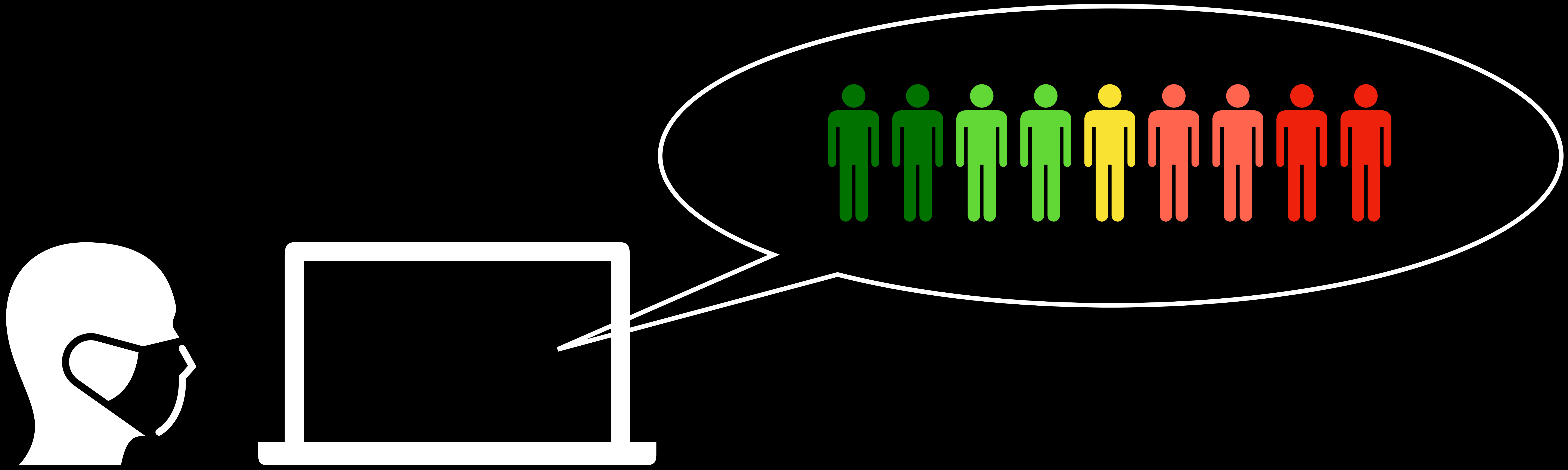


Incorporate new data

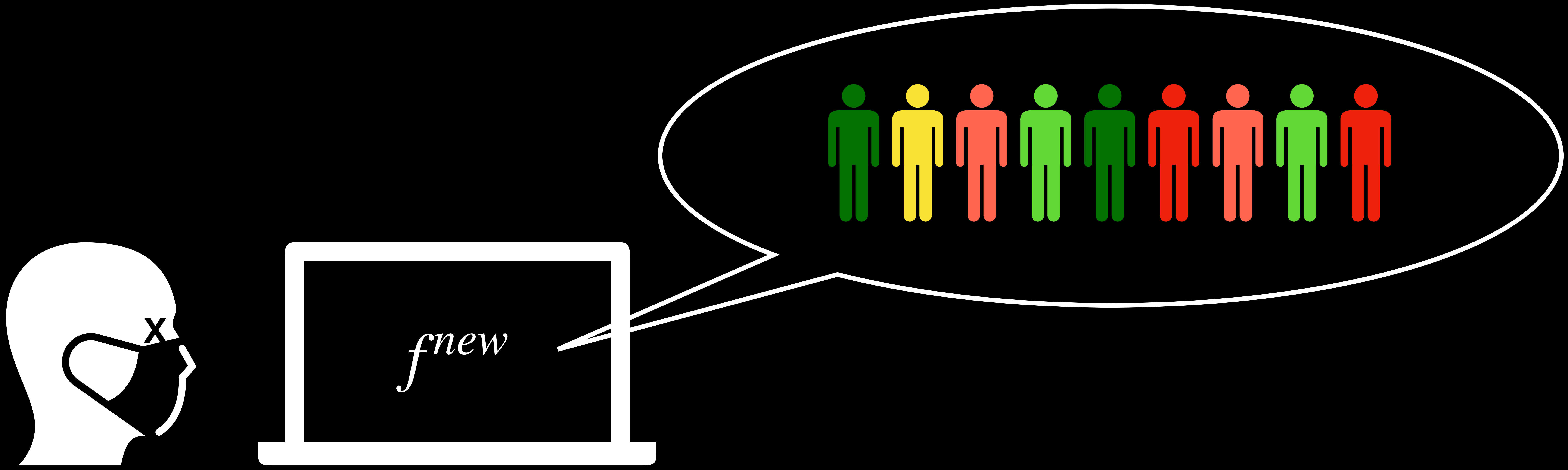


Finlayson 2021, David 2017, Hickey 2012, Minne 2012, Bansal 2019

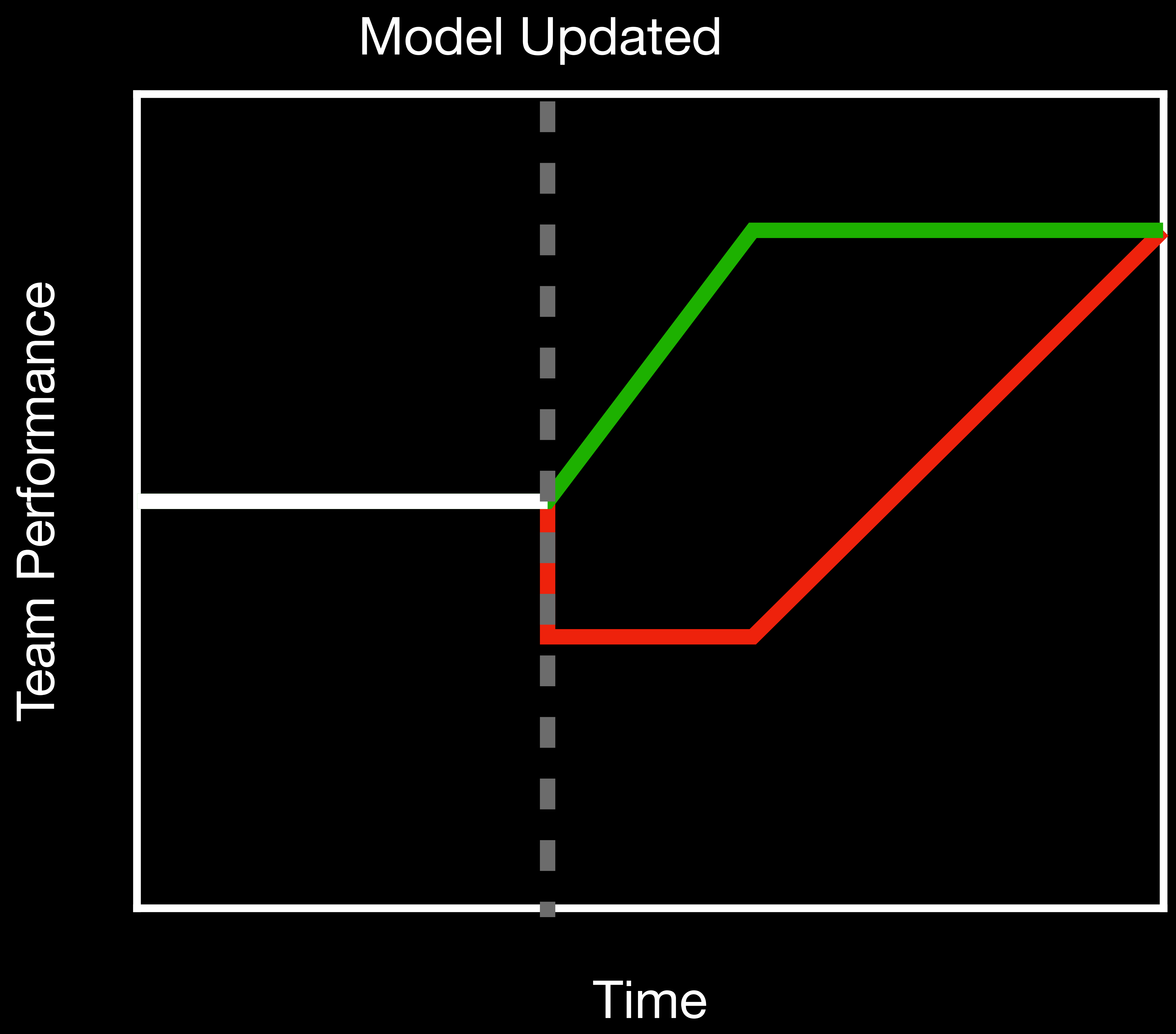
Updates can mess with user expectations.



Updates can mess with user expectations.



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



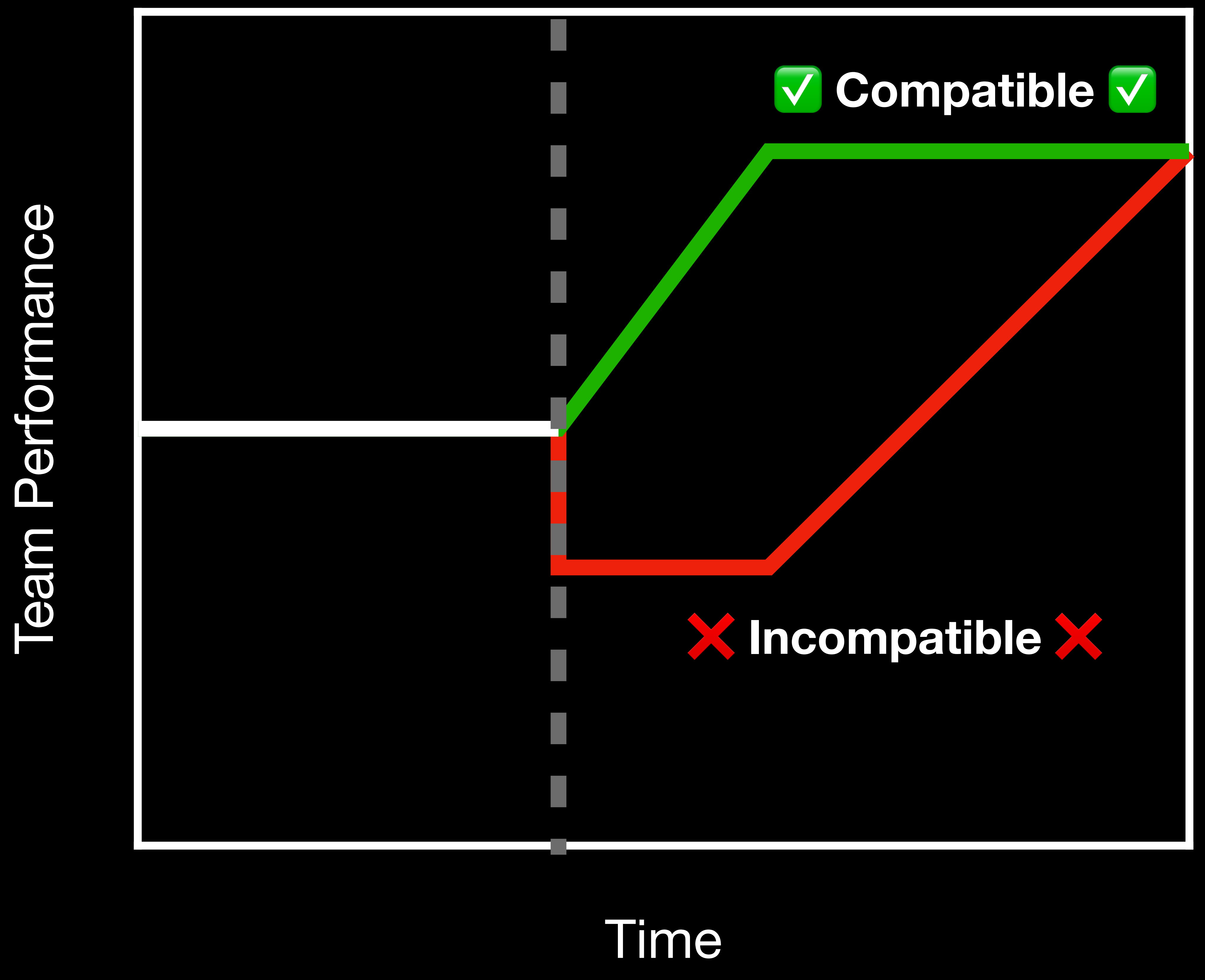
Adapted from Bansal 2019

Ideally updated models meet the expectations of users

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

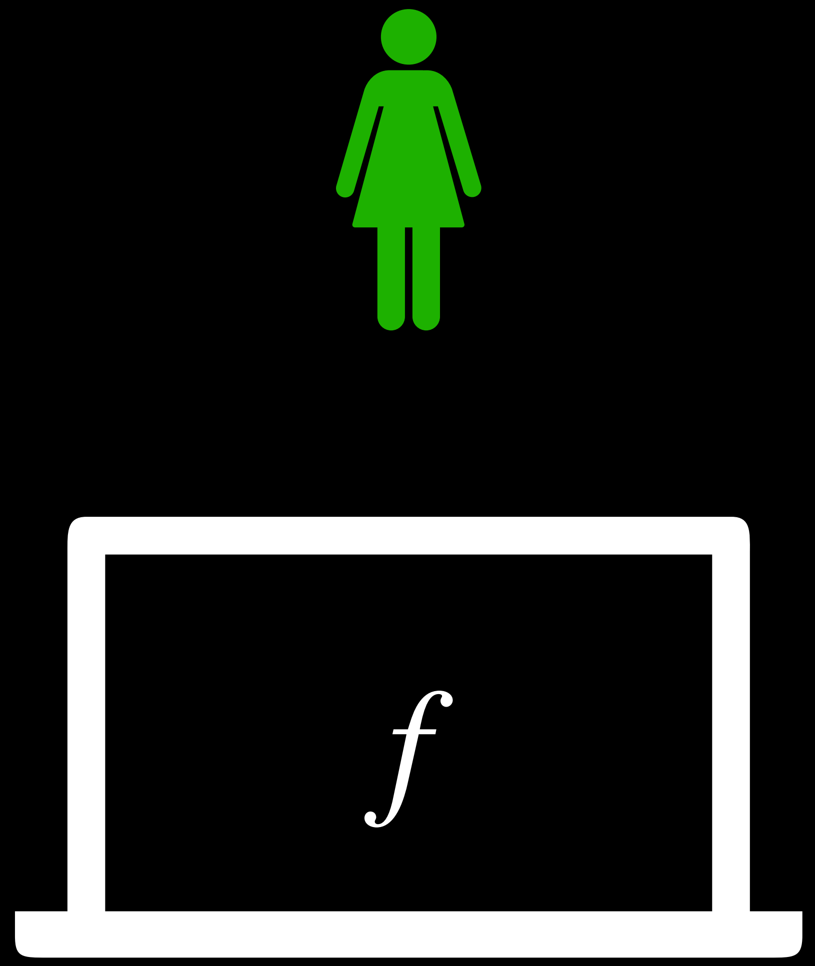
Way to measure user expectations

Goal: updated models should have high compatibility

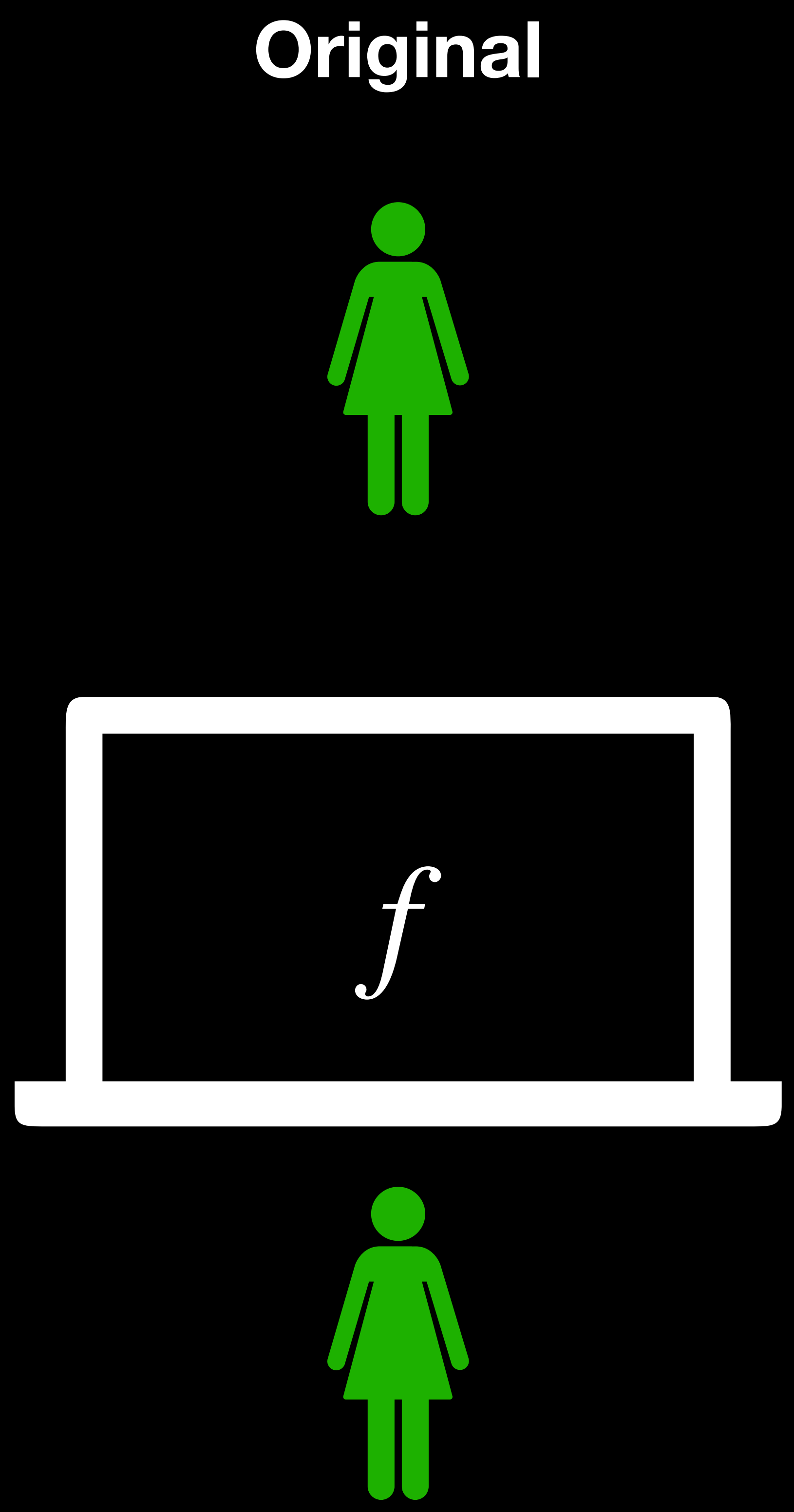


Compatibility can be assessed by using the original and updated models for the same predictive task.

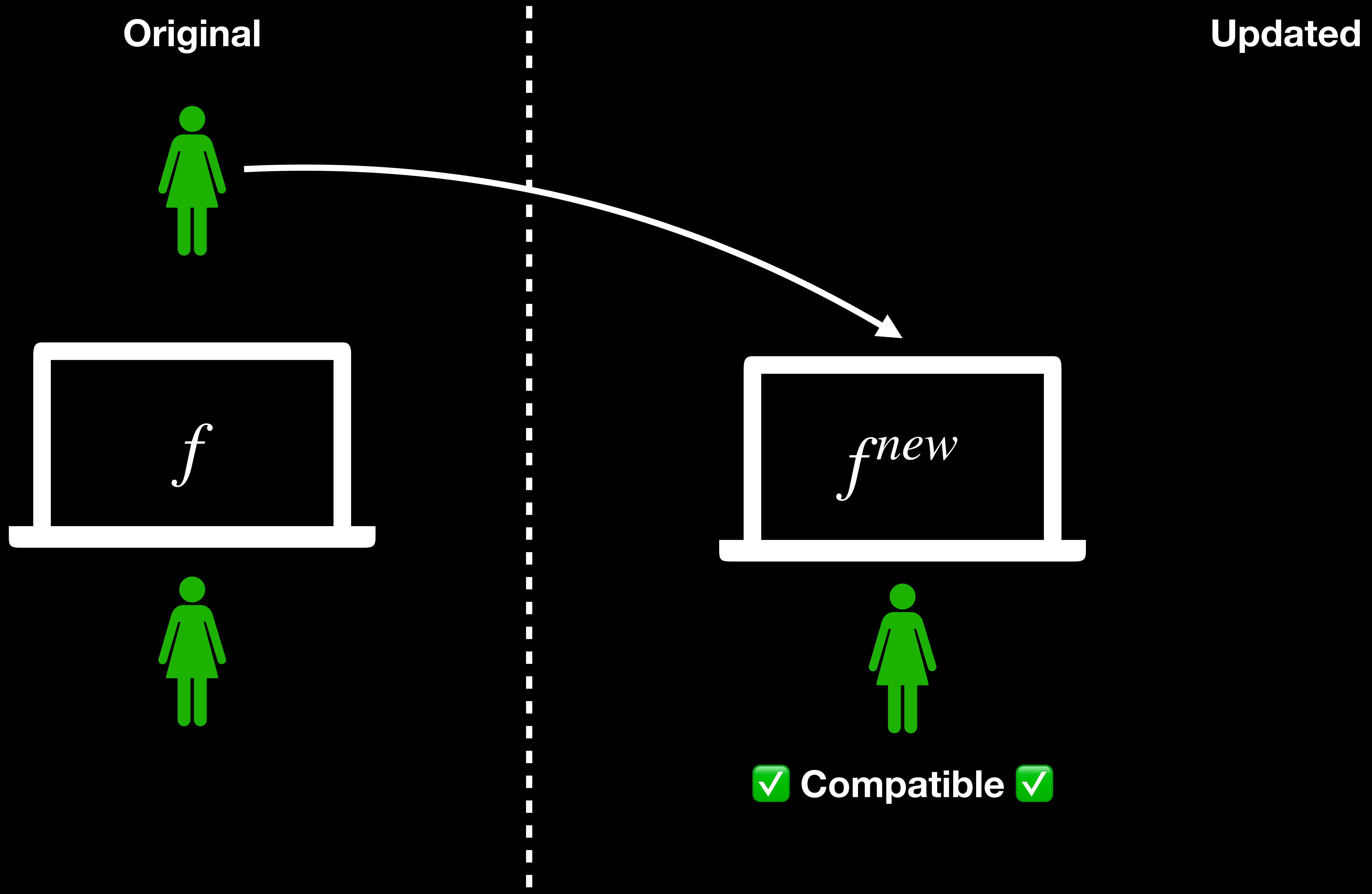
Original



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

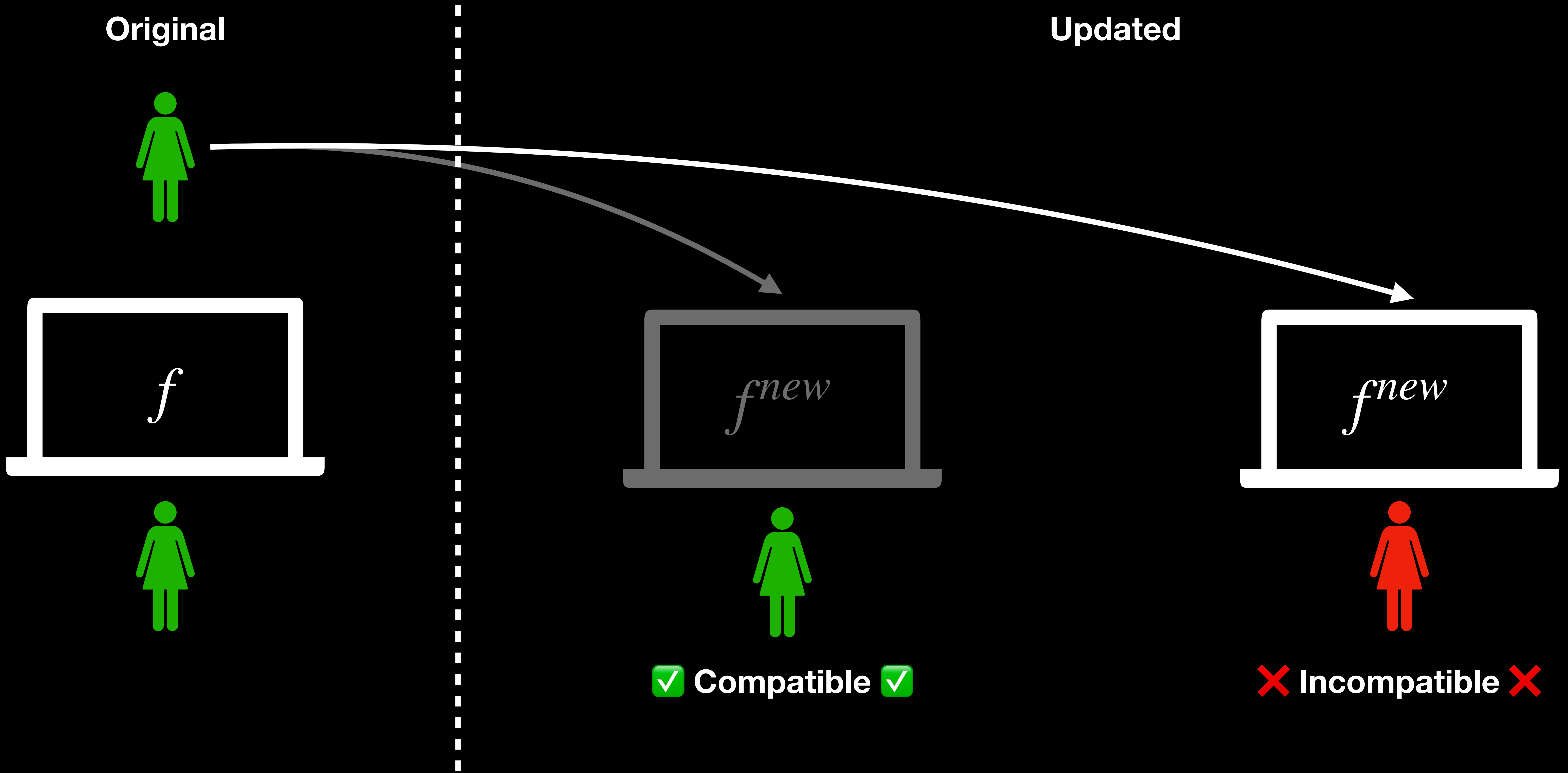


Compatibility can be assessed by using the original and updated models for the same predictive task.



Bansal 2019

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.

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

Backwards Trust Compatibility C^{BT}

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

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

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

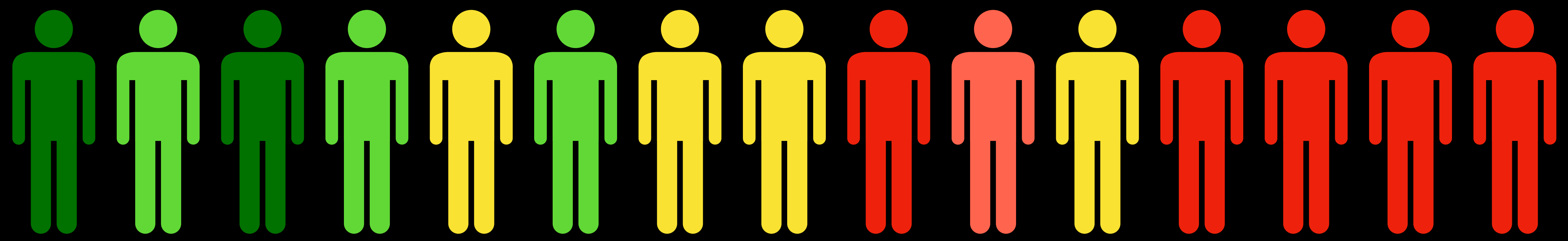
Bansal 2019

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

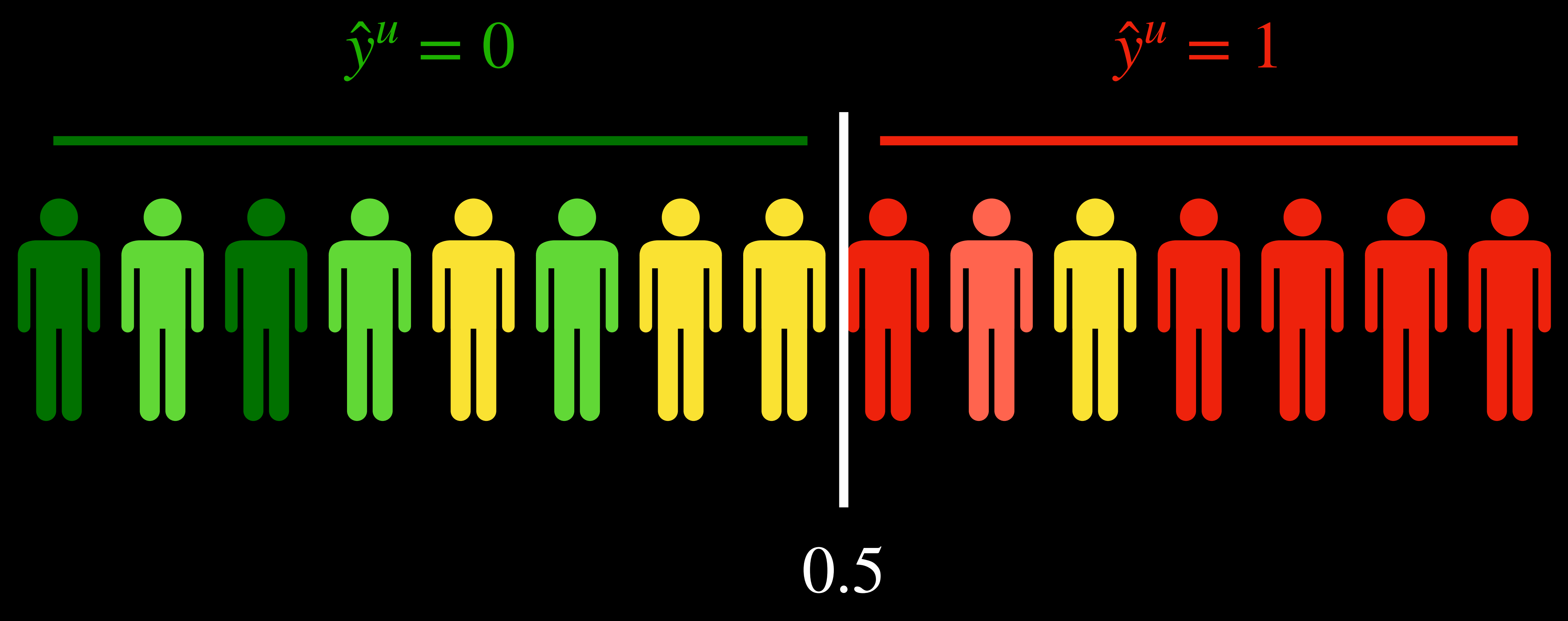


Problems with existing compatibility measures.

Existing measure depends on equality comparison

Problematic for use in risk stratification model & healthcare settings

- Depends on setting a single decision threshold

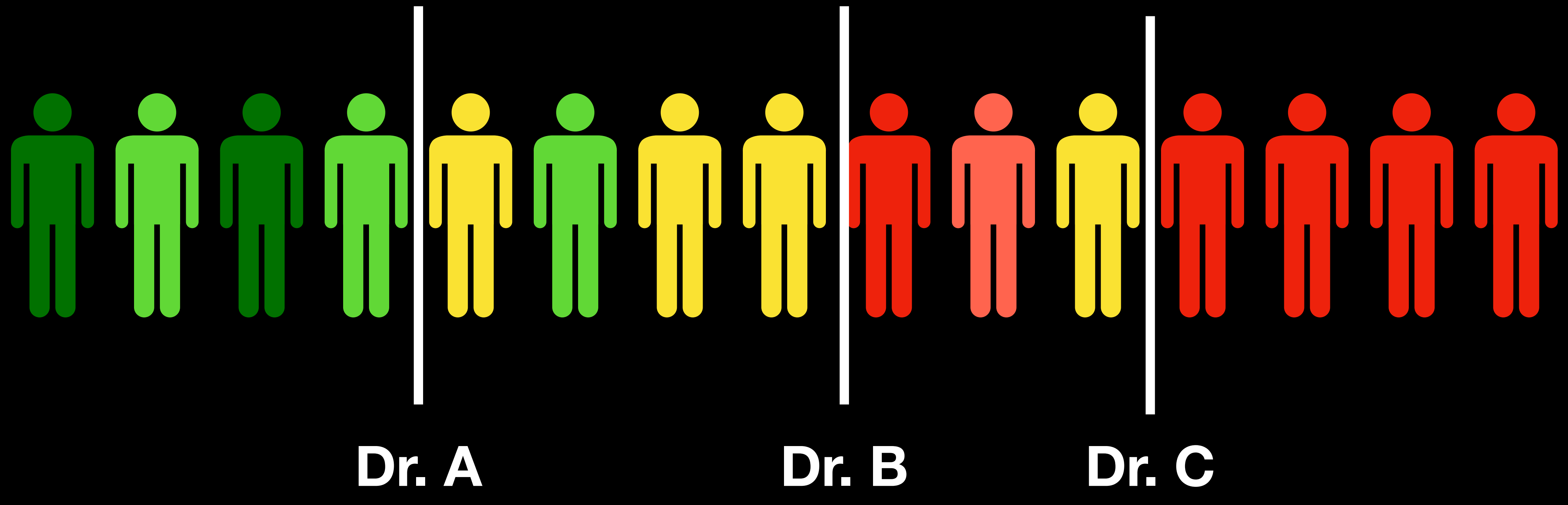


Problems with existing compatibility measures.

Existing measure depends on equality comparison

Problematic for use in risk stratification model & healthcare settings

- Depends on setting a single decision threshold

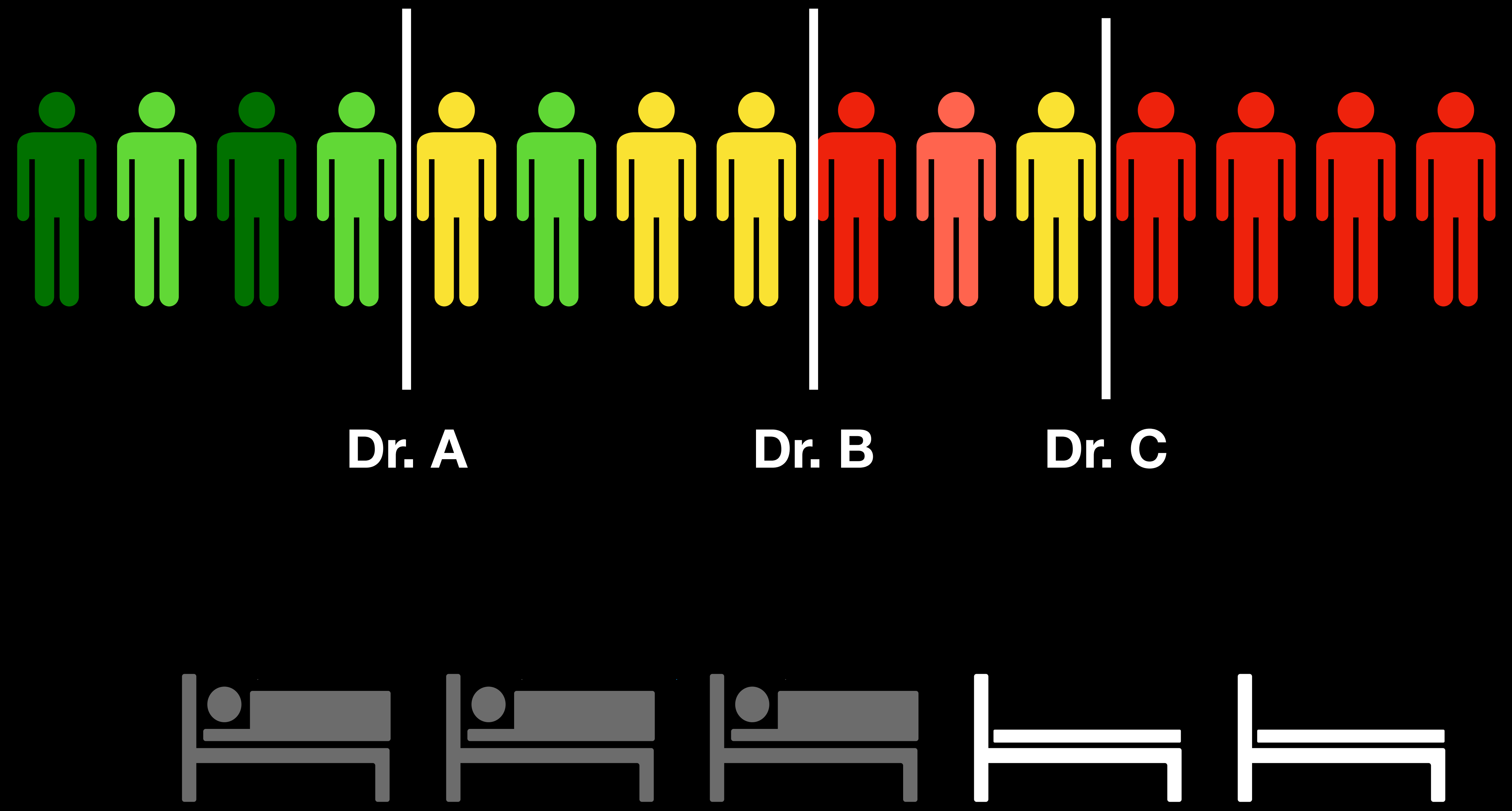


Problems with existing compatibility measures.

Existing measure depends on equality comparison

Problematic for use in risk stratification model & healthcare settings

- Depends on setting a single decision threshold



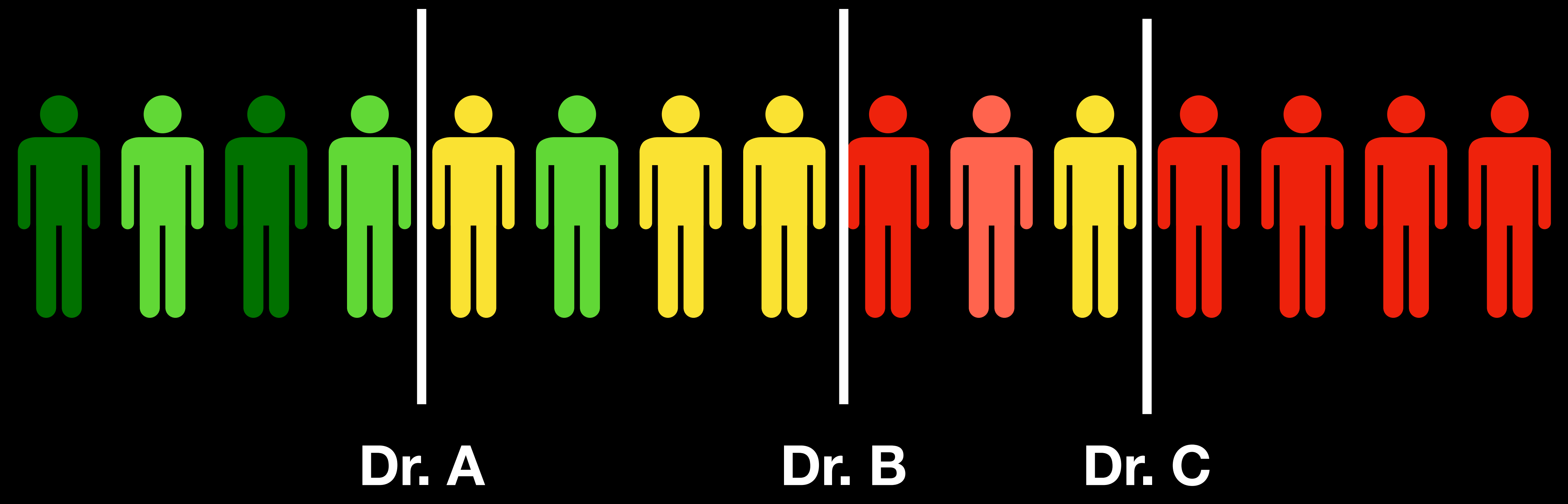
Problems with existing compatibility measures.

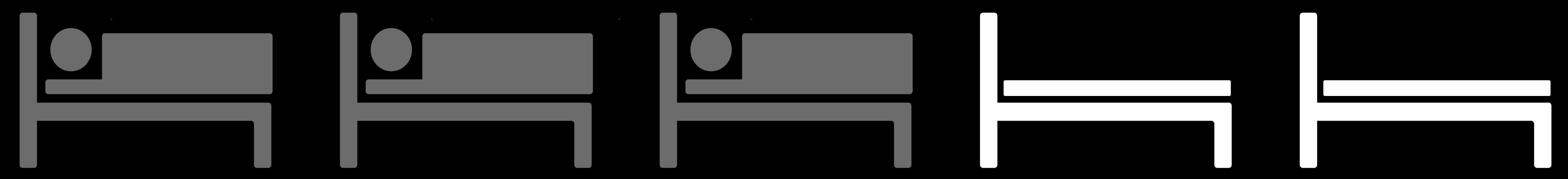
Existing measure depends on equality comparison

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:

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

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
$n = 500$

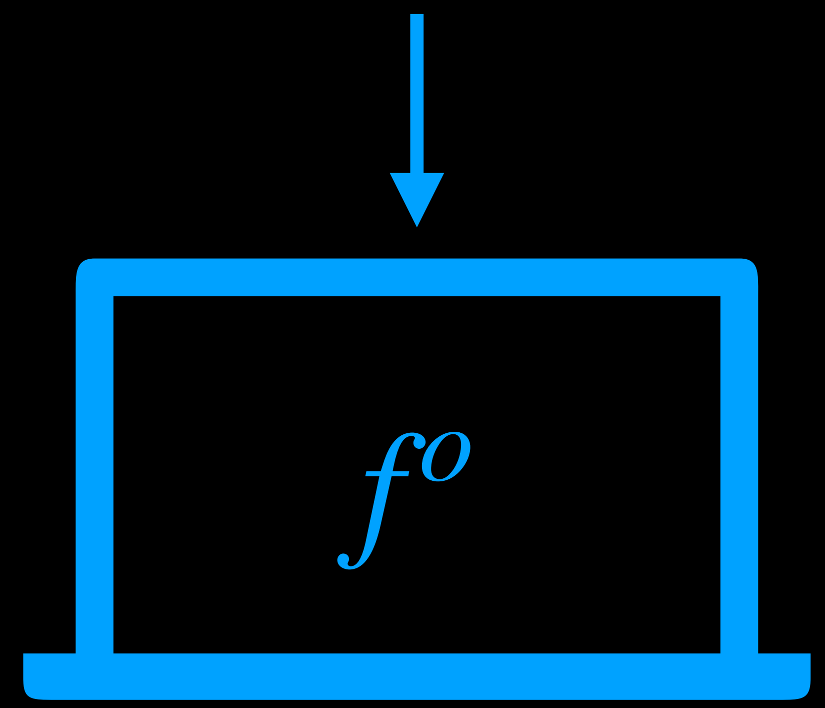
Original Model Validation
$n = 500$

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

Original Model
$n = 1{,}000$

Original Model Development
$n = 500$

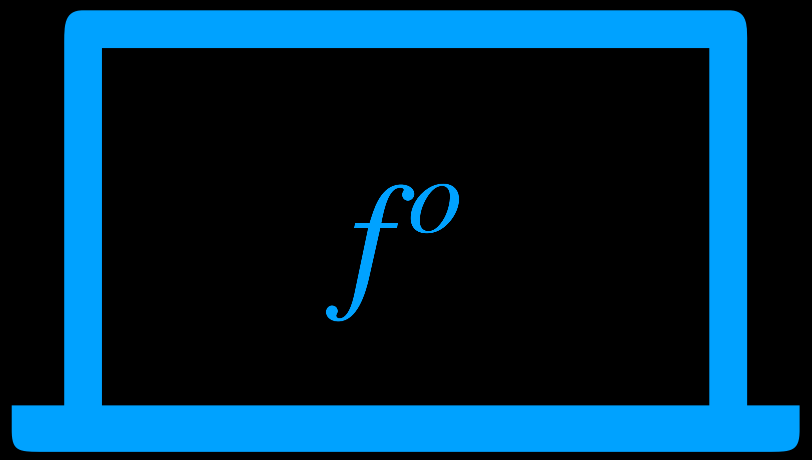
Original Model Validation
$n = 500$

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

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

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

Evaluation Dataset
$n = 2{,}577$



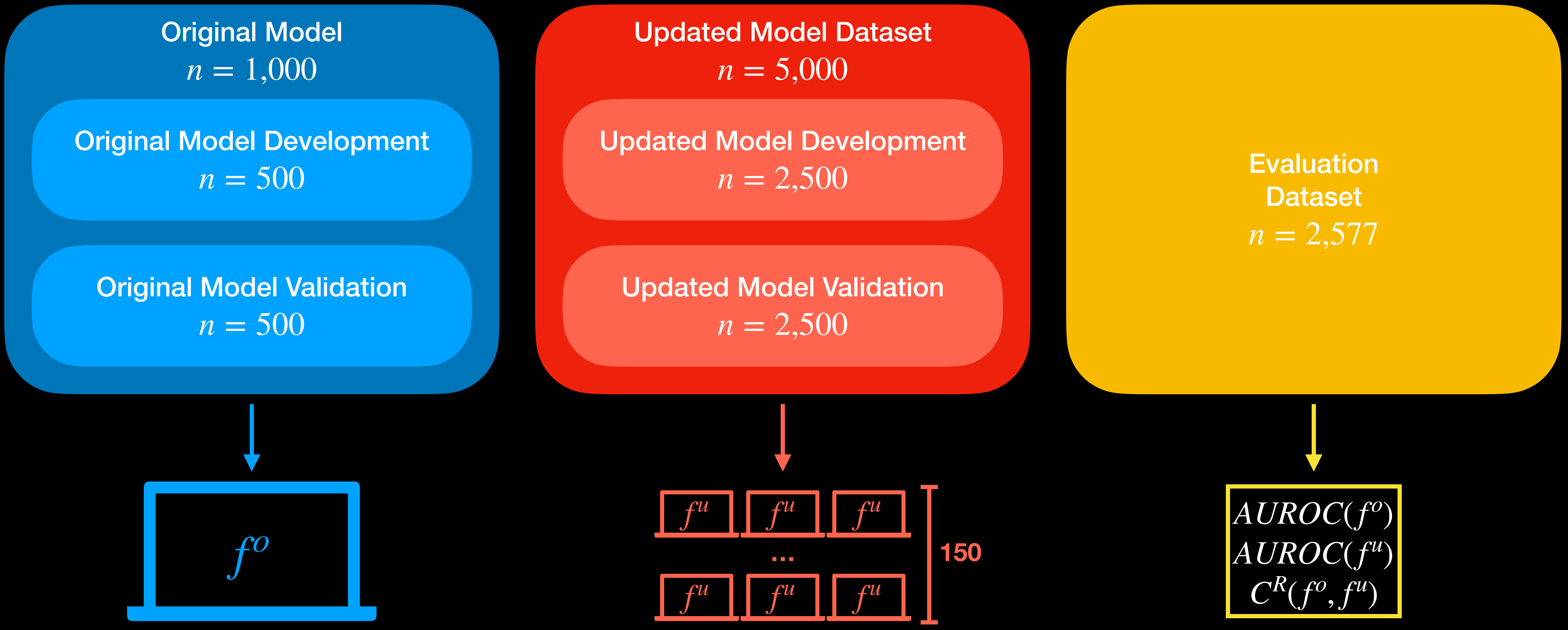
f^u f^u f^u

...

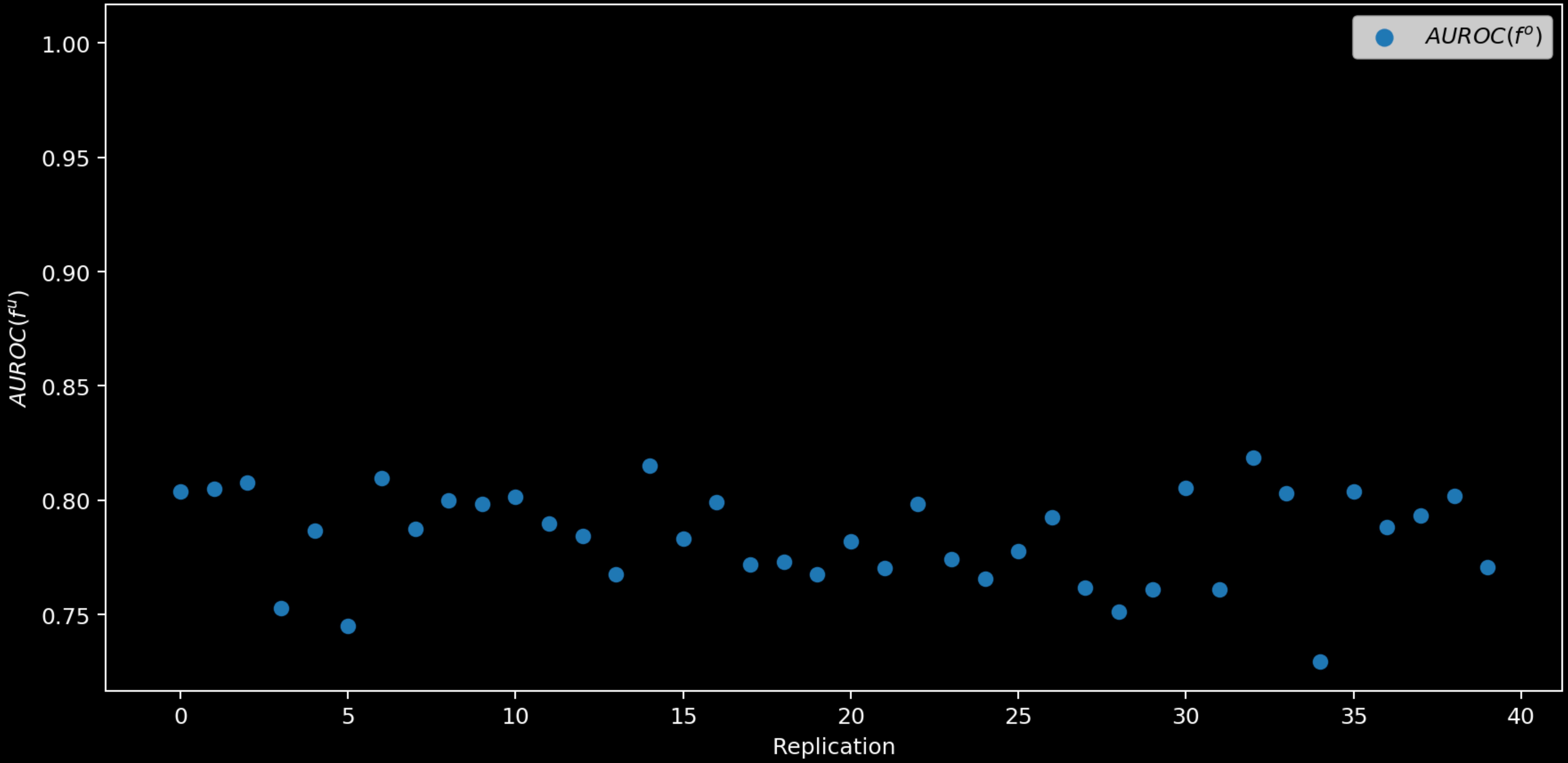
f^u f^u f^u

150

Q1: Experimental Setup

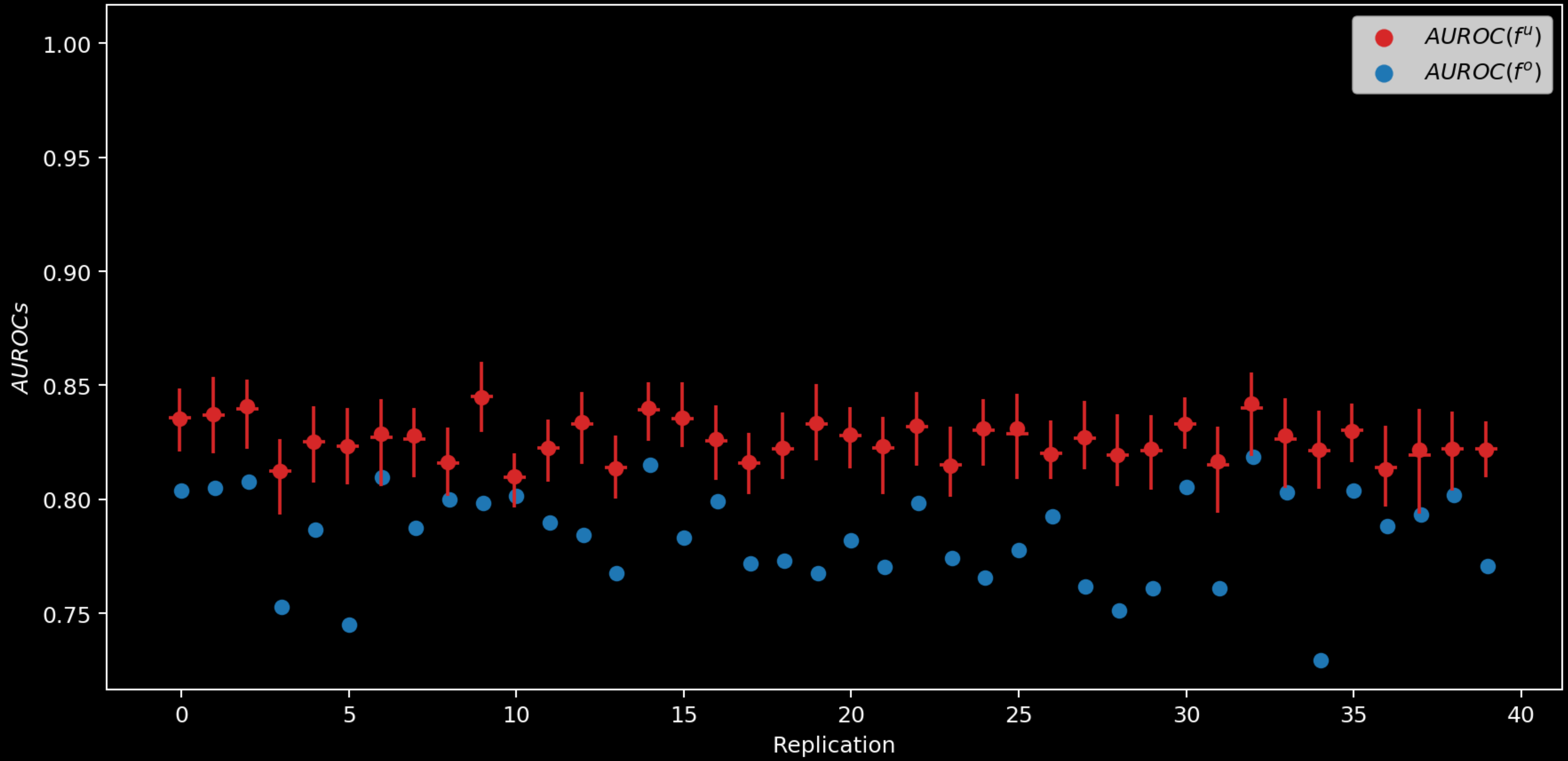


Results



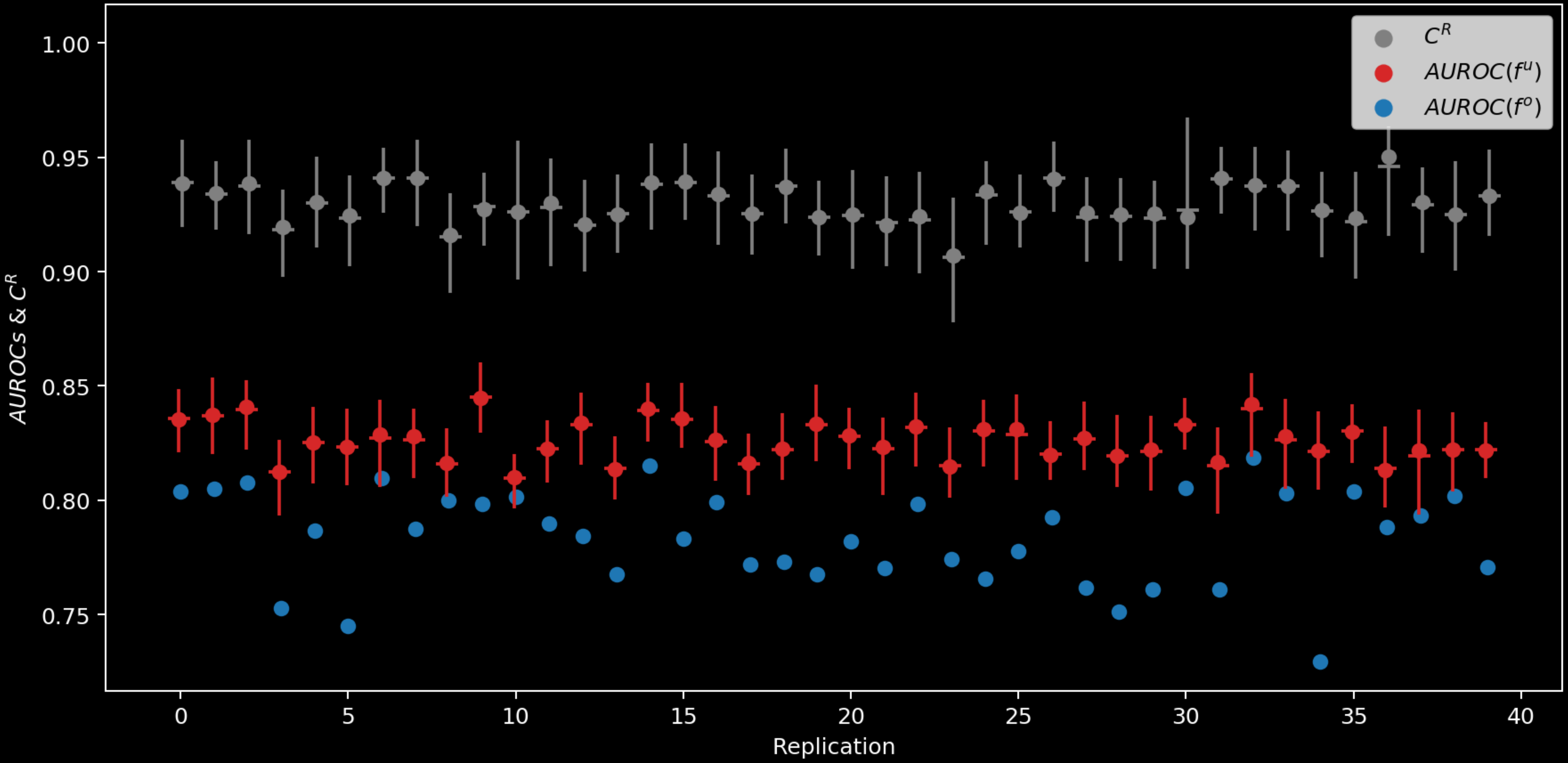
$AUROC(f^o)$
$AUROC(f^u)$
1.00
0.95
0.90
0.85
0.80
0.75
0
5
10
15
20
25
30
35
40
Replication

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
$n = 500$

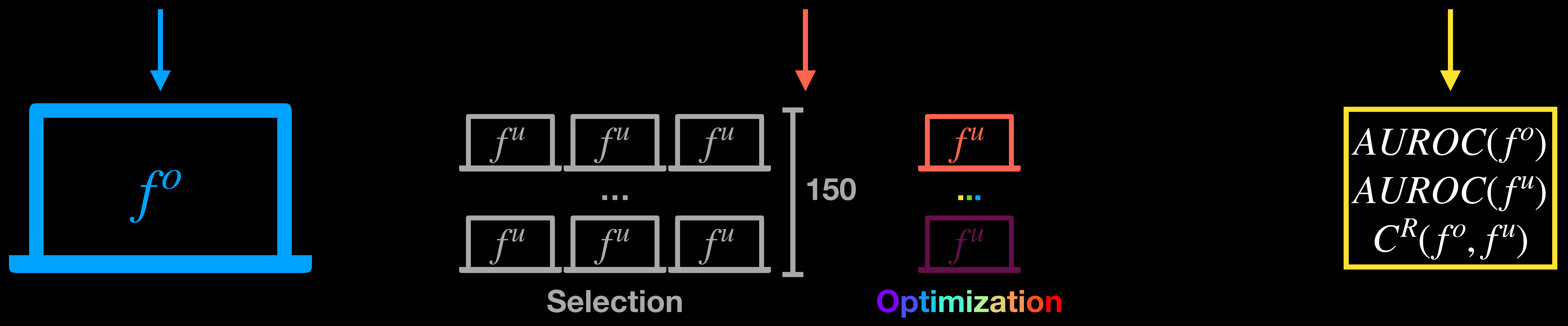
Original Model Validation
$n = 500$

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

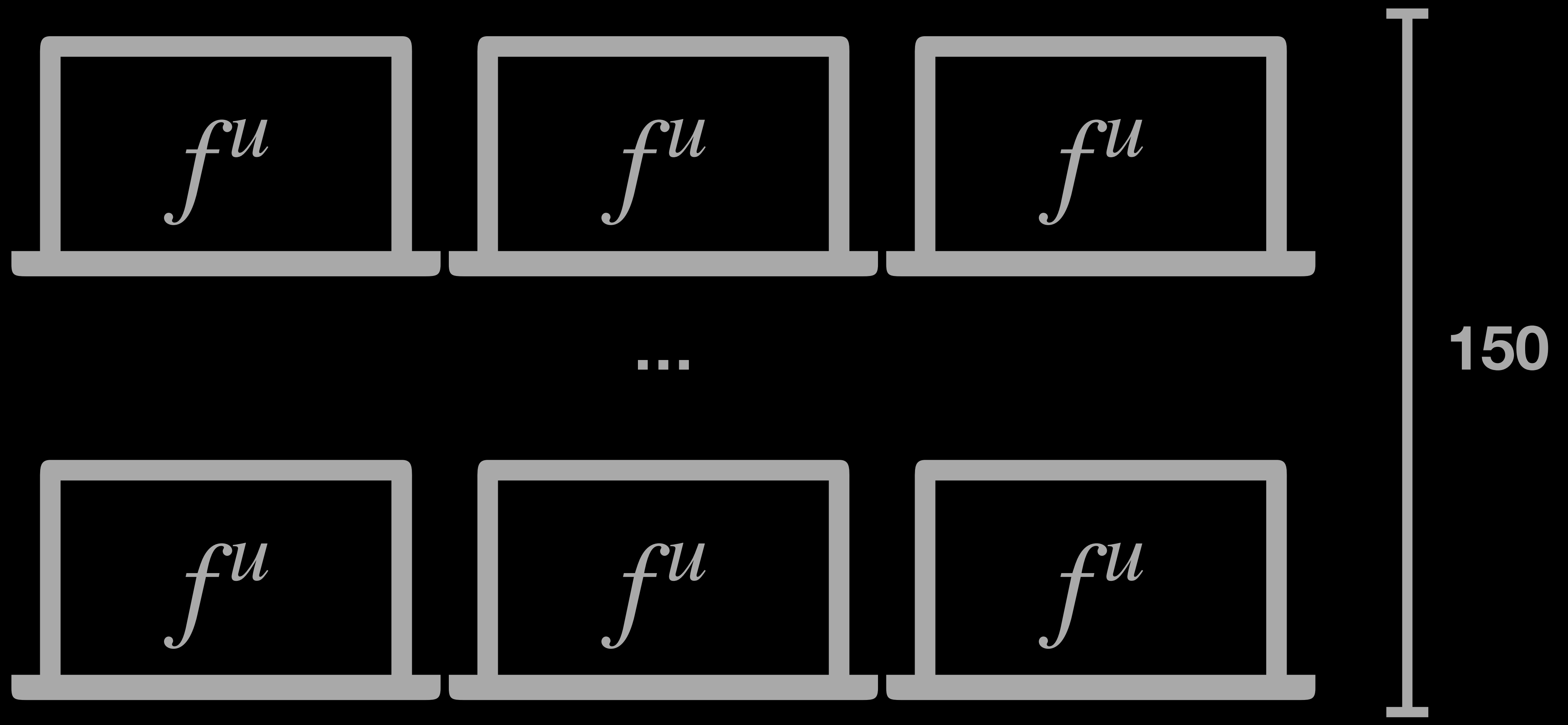
Updated Model Development
$n = 2{,}500$

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

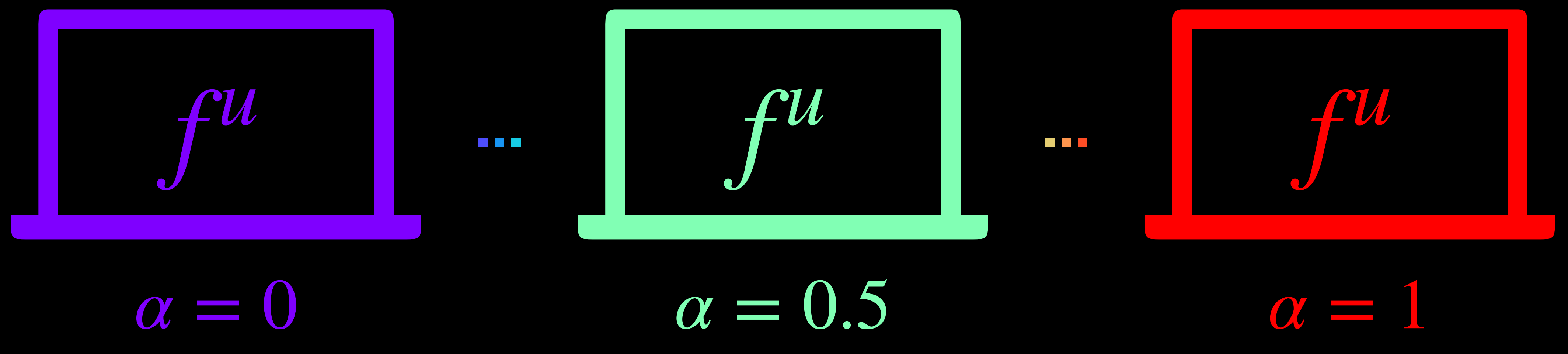
Evaluation Dataset
$n = 2{,}577$



Q2: Updated Models Selection vs. Optimization



Selection

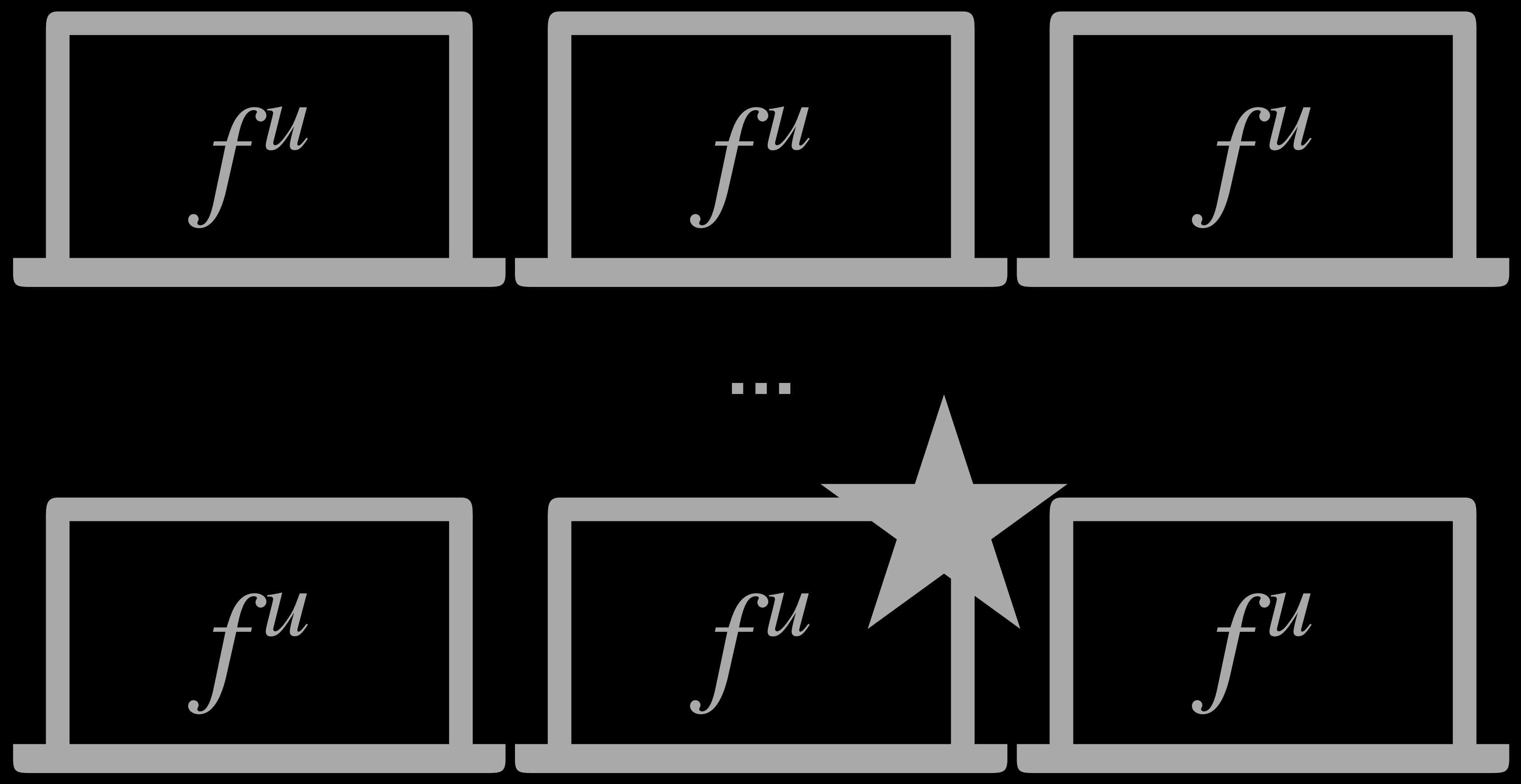


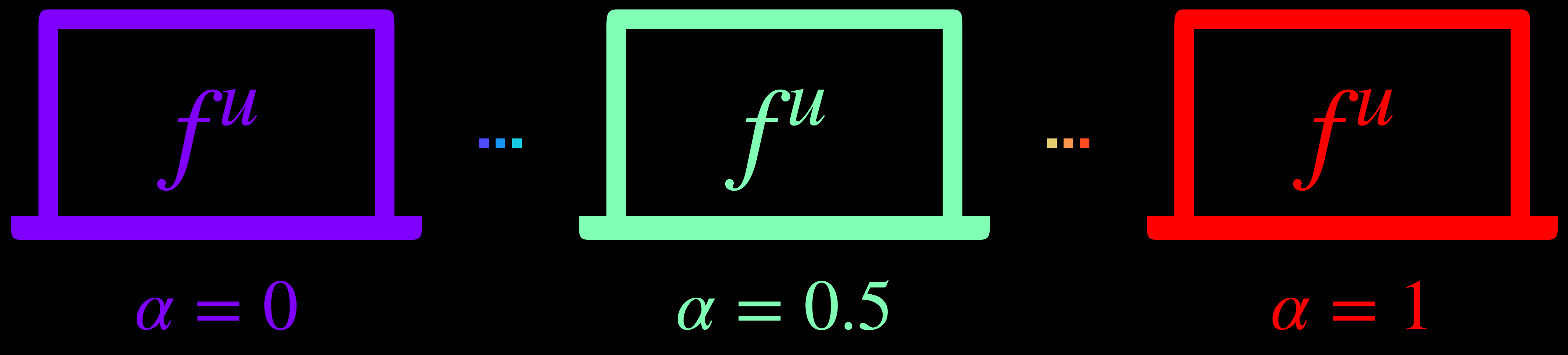
Optimization

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

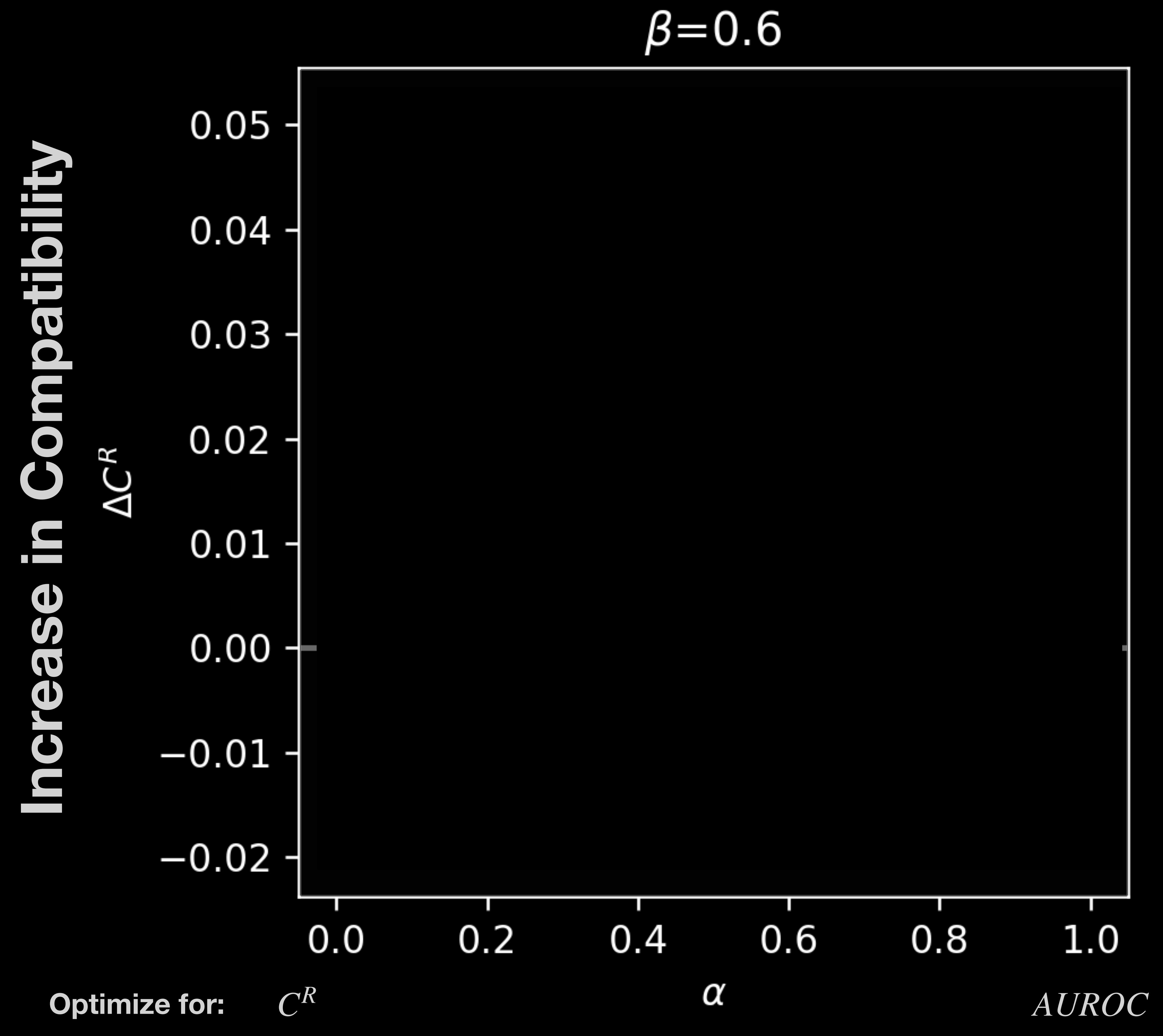
Optimization

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

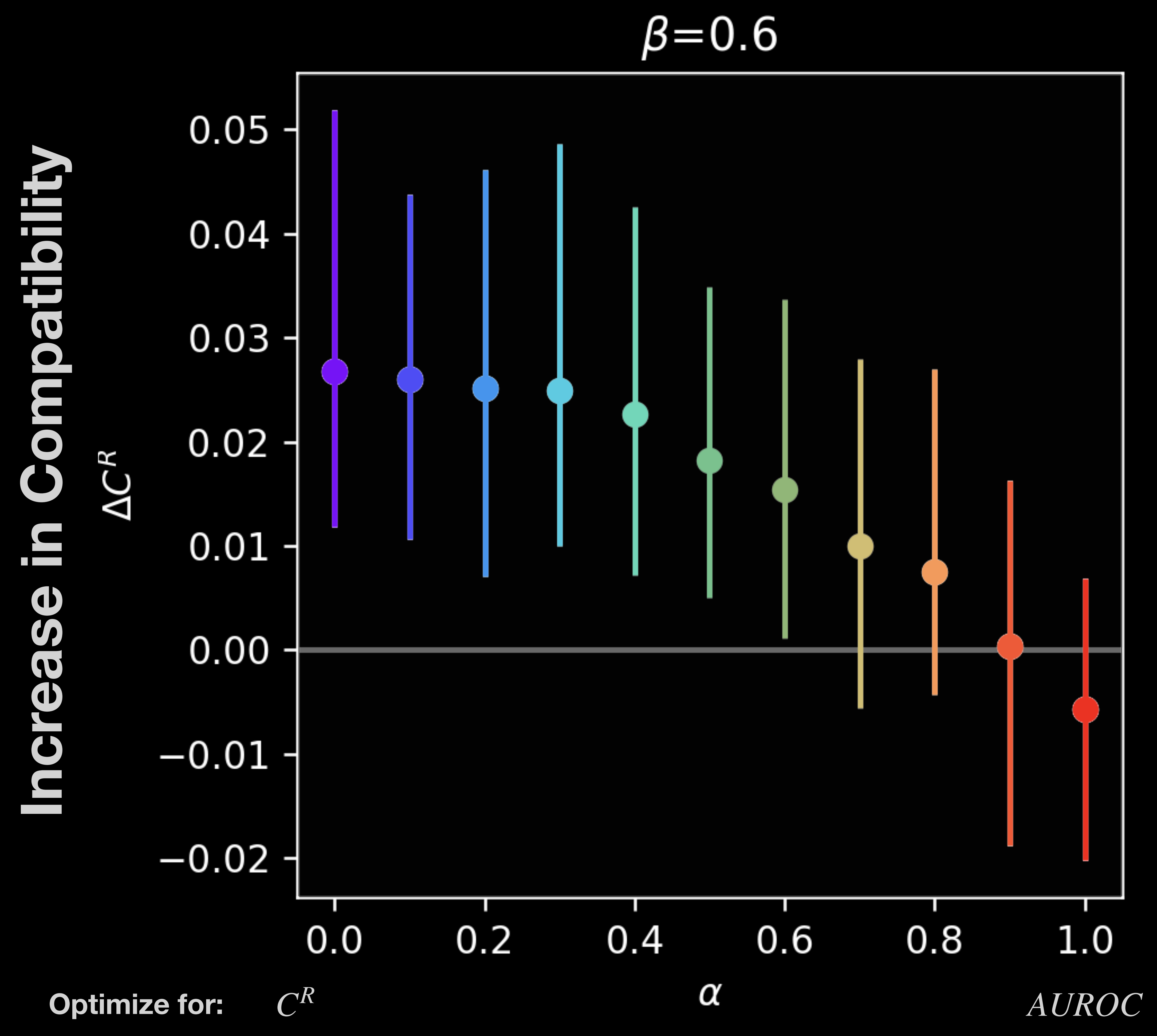
For example model with best validation AUROC

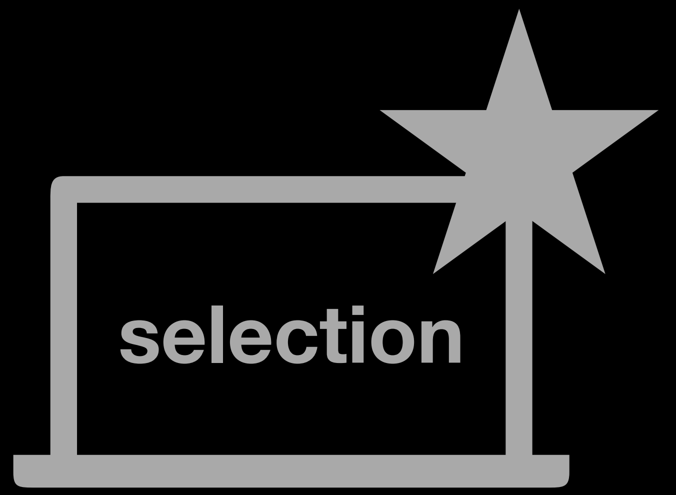
Examine difference in held out evaluation C^R and AUROC

Q2: C^R performance results

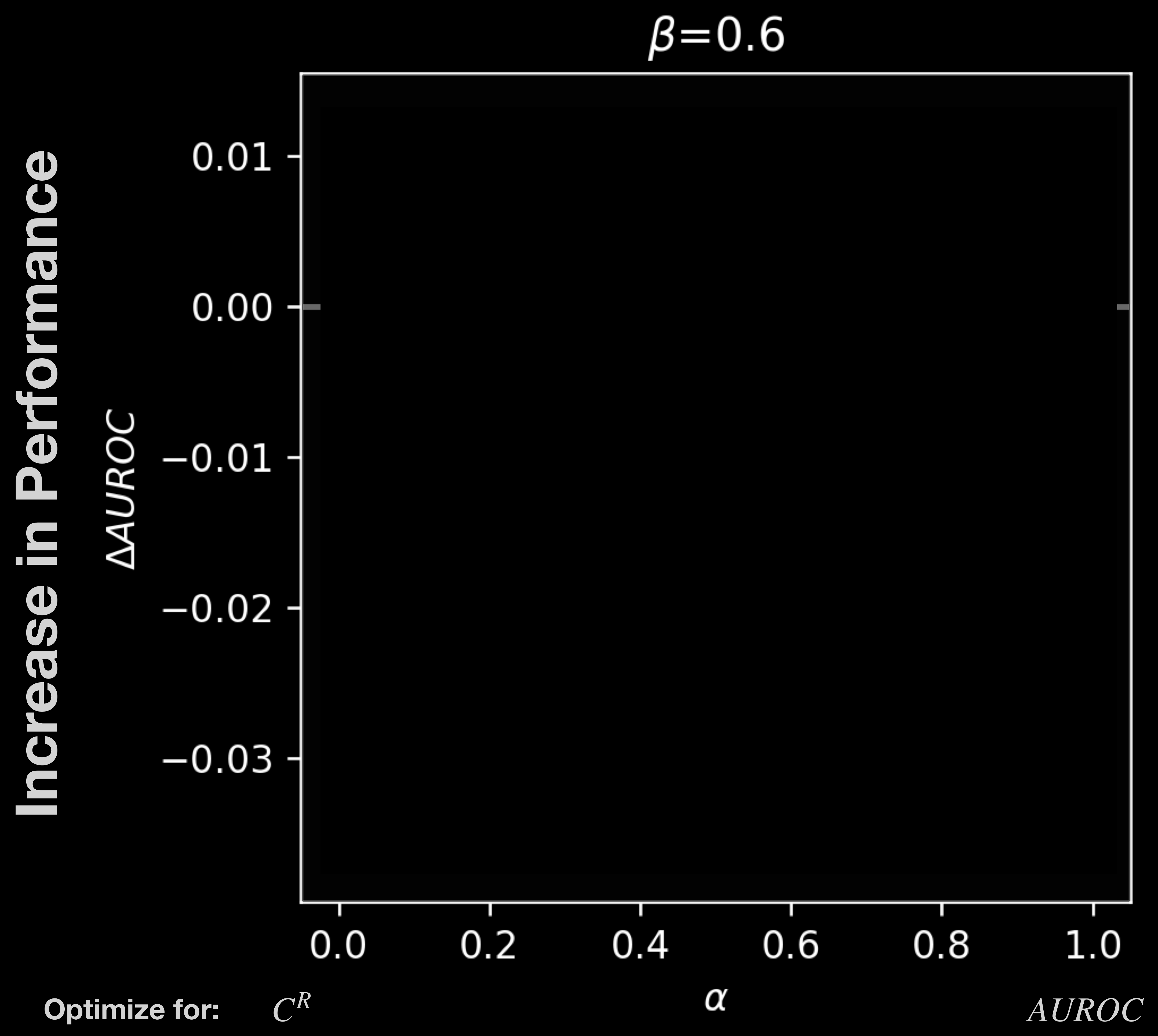


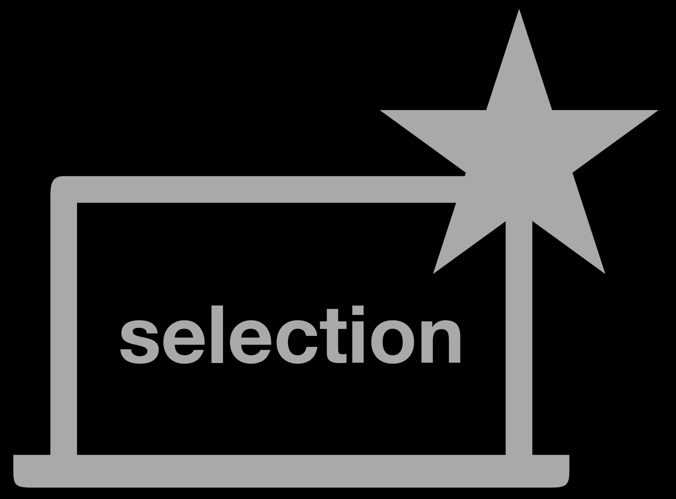
Q2: C^R performance results



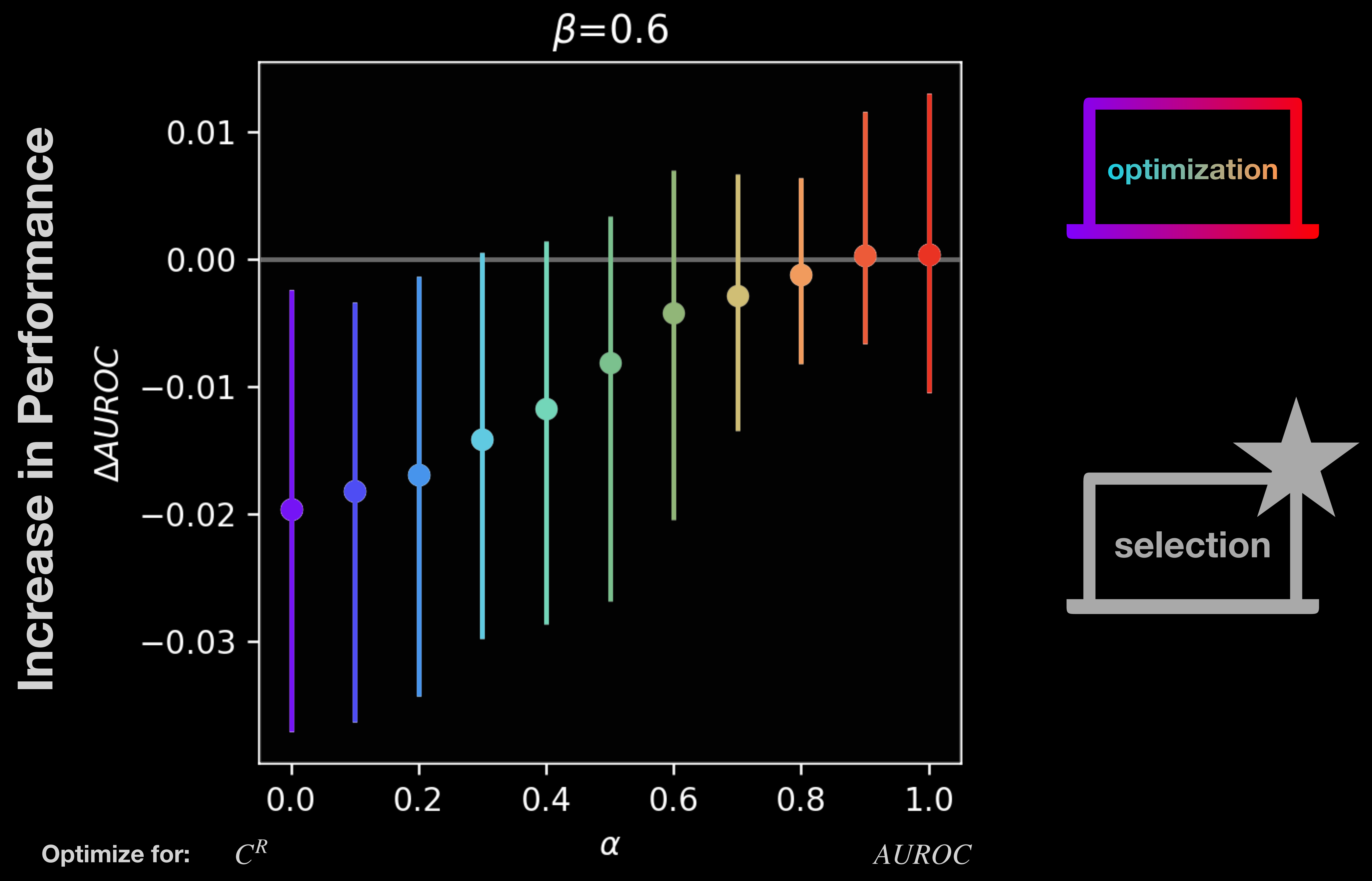


Q2: $AUROC$ performance results

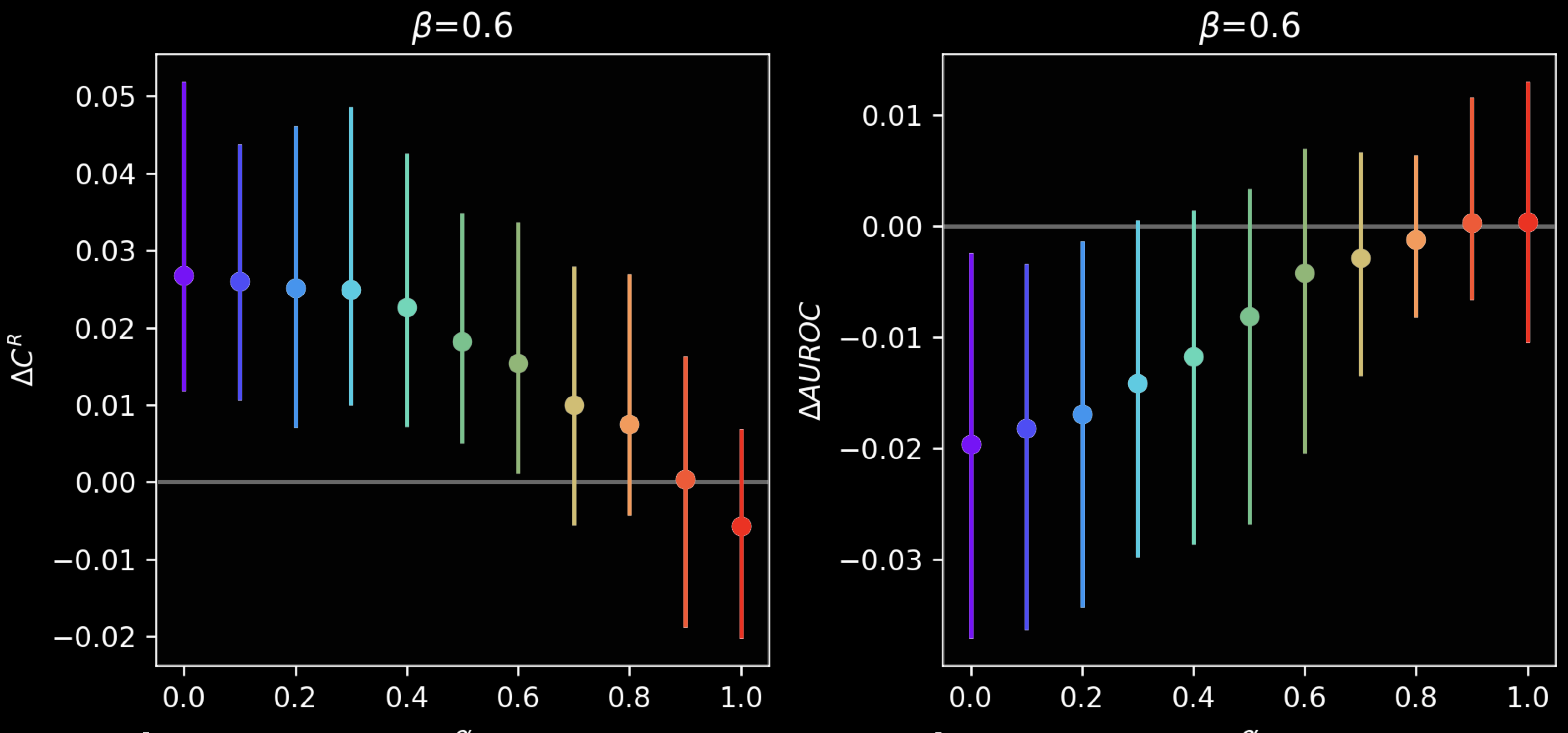




Q2: $AUROC$ performance results

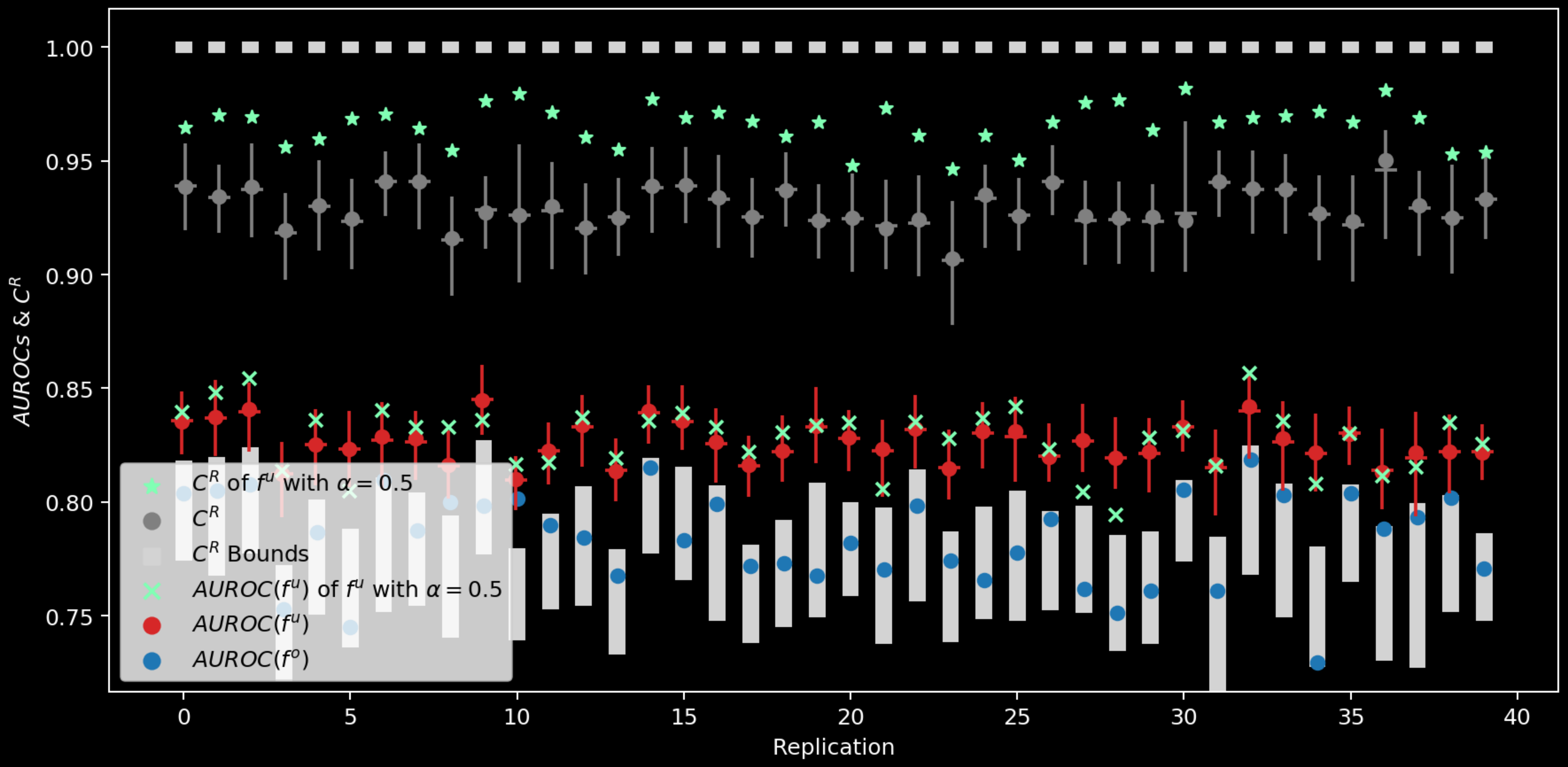


Q2: Performance results



Optimize for: C^R … $AUROC$ (left chart); C^R … $AUROC$ (right chart)

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?

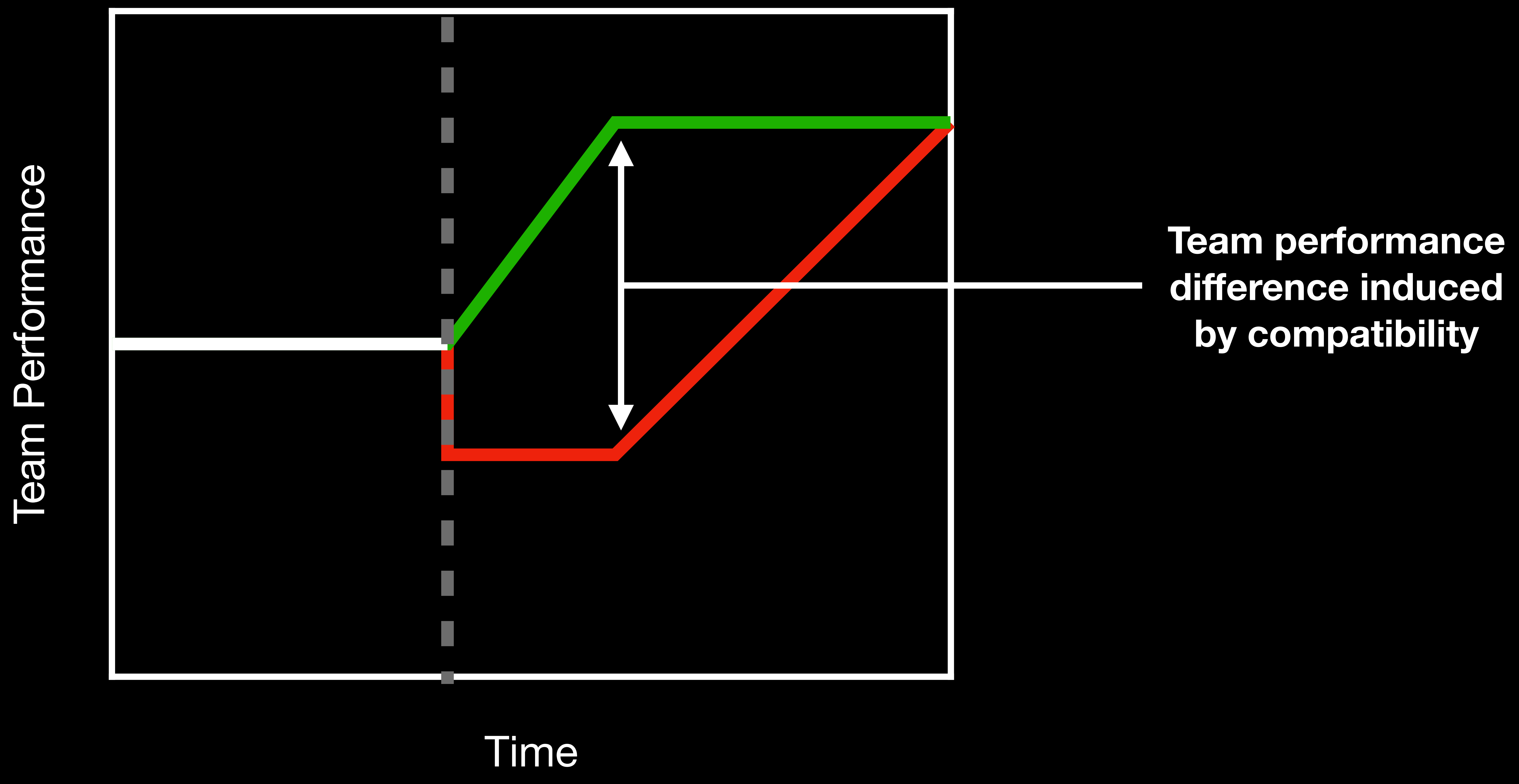
Yes, using our weighted loss function.

Does that come at a cost in terms of AUROC?

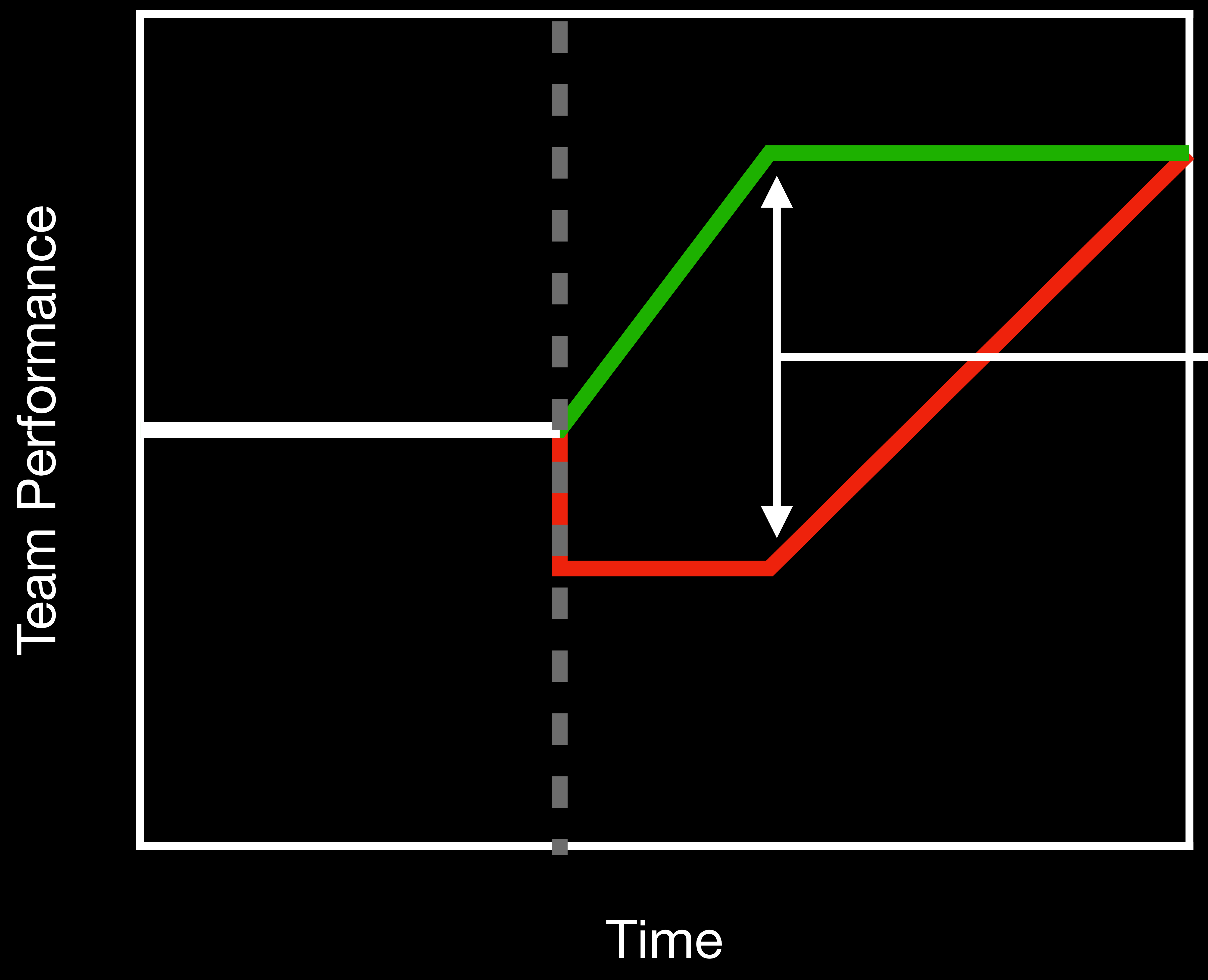
Sort of…

Why trade-off AUROC for C^R?

Why trade-off AUROC for C^R?



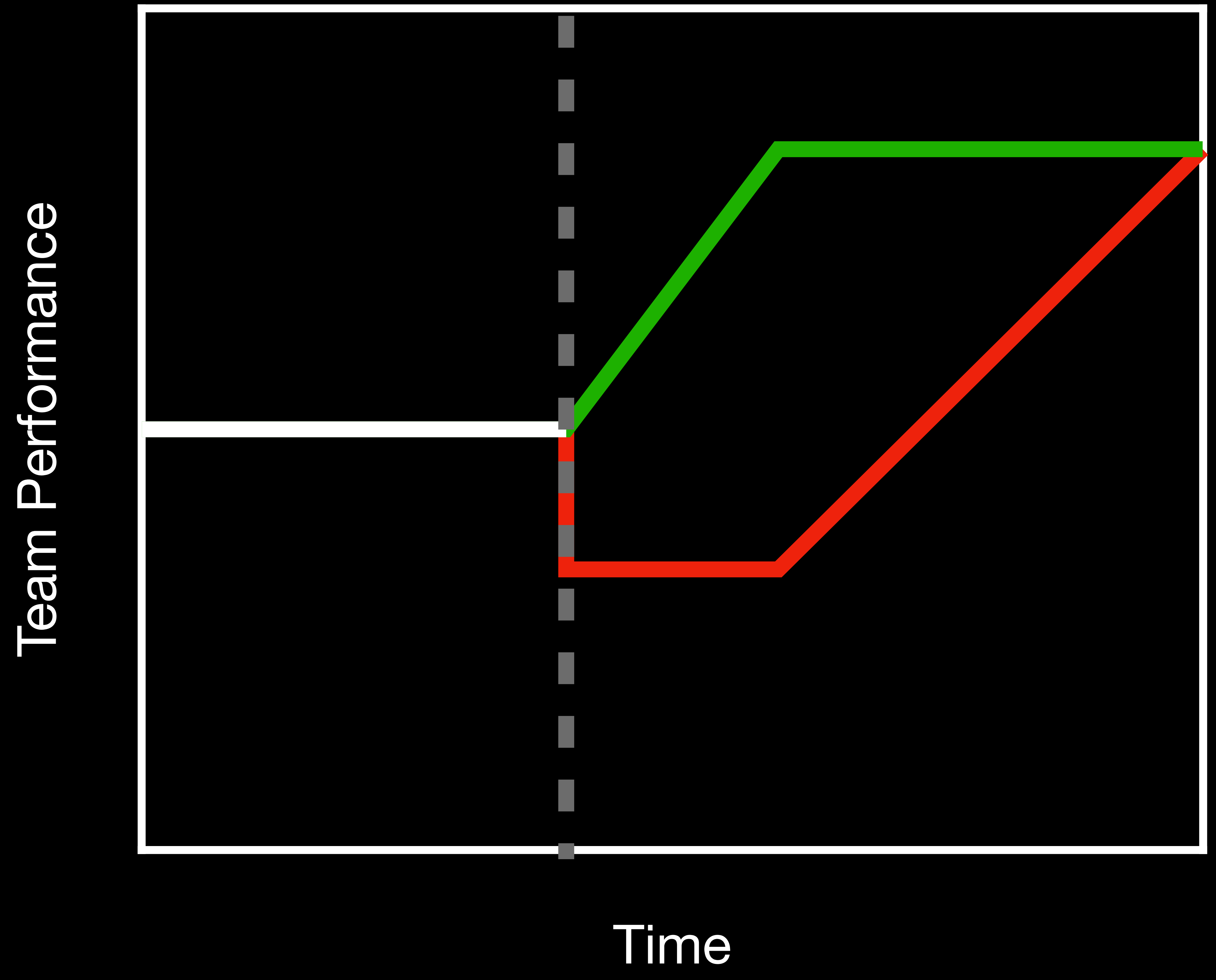
Why trade-off AUROC for C^R?



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

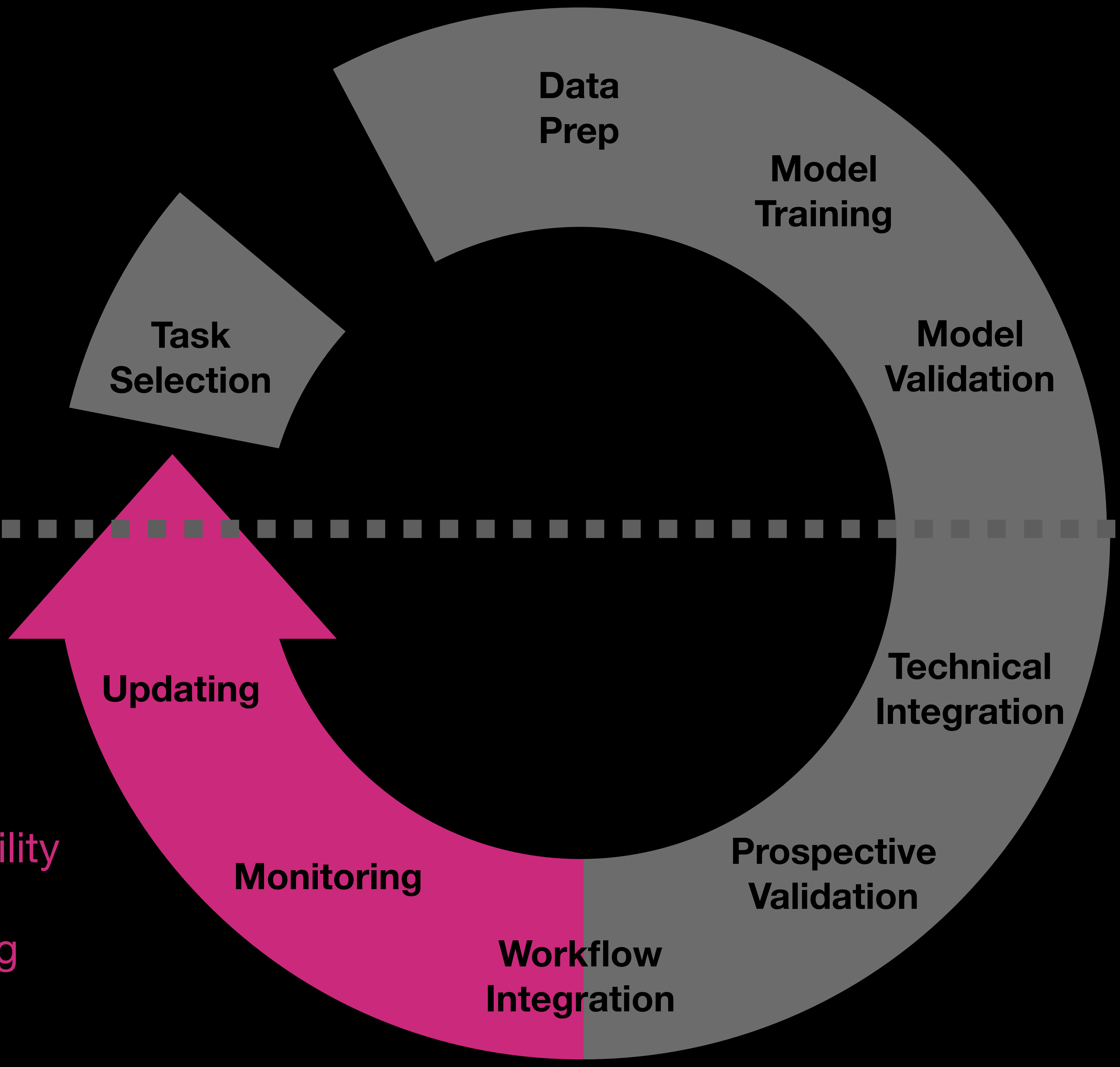
My work enables model developers to find these desirable updated risk stratification models.

Updated models may be assessed with our new rank-based compatibility measure.

Model developers may use our weighted loss function to create model updates with improved C^R.

Additionally, we present a real-world model updating case-study.

The ML Lifecycle



RTW Prediction

Assess value of longitudinal observations for return to work prediction

Ötleş *et al.*, JAMIA '22

Performance Gap

Characterize changes to prospective performance after model implementation

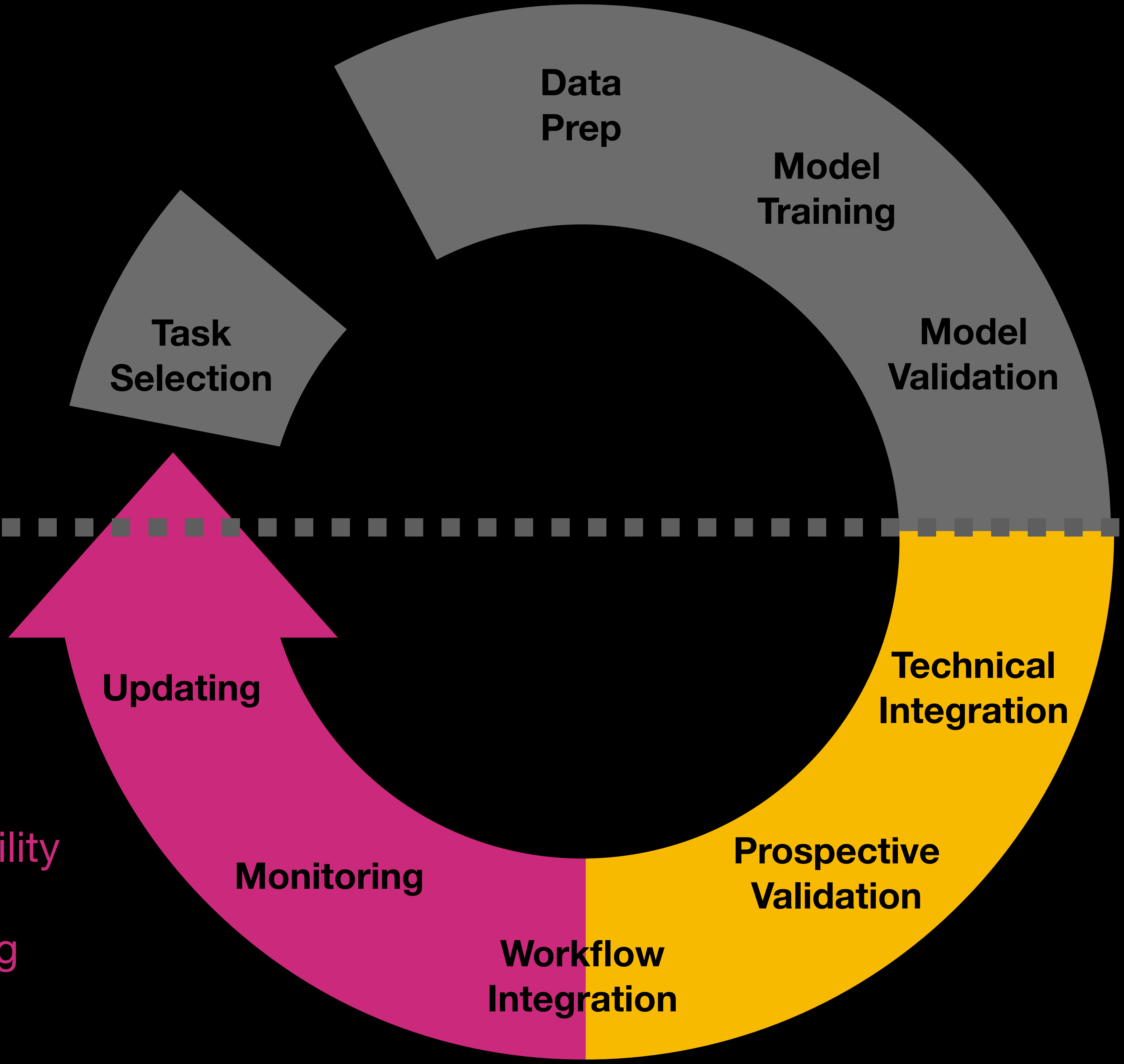
Ötleş *et al.*, MLHC '21

Updating Models

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

Ötleş *et al.*, MLHC '23

The ML Lifecycle



RTW Prediction

Assess value of longitudinal observations for return to work prediction

Ötleş *et al.*, JAMIA '22

Updating Models

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

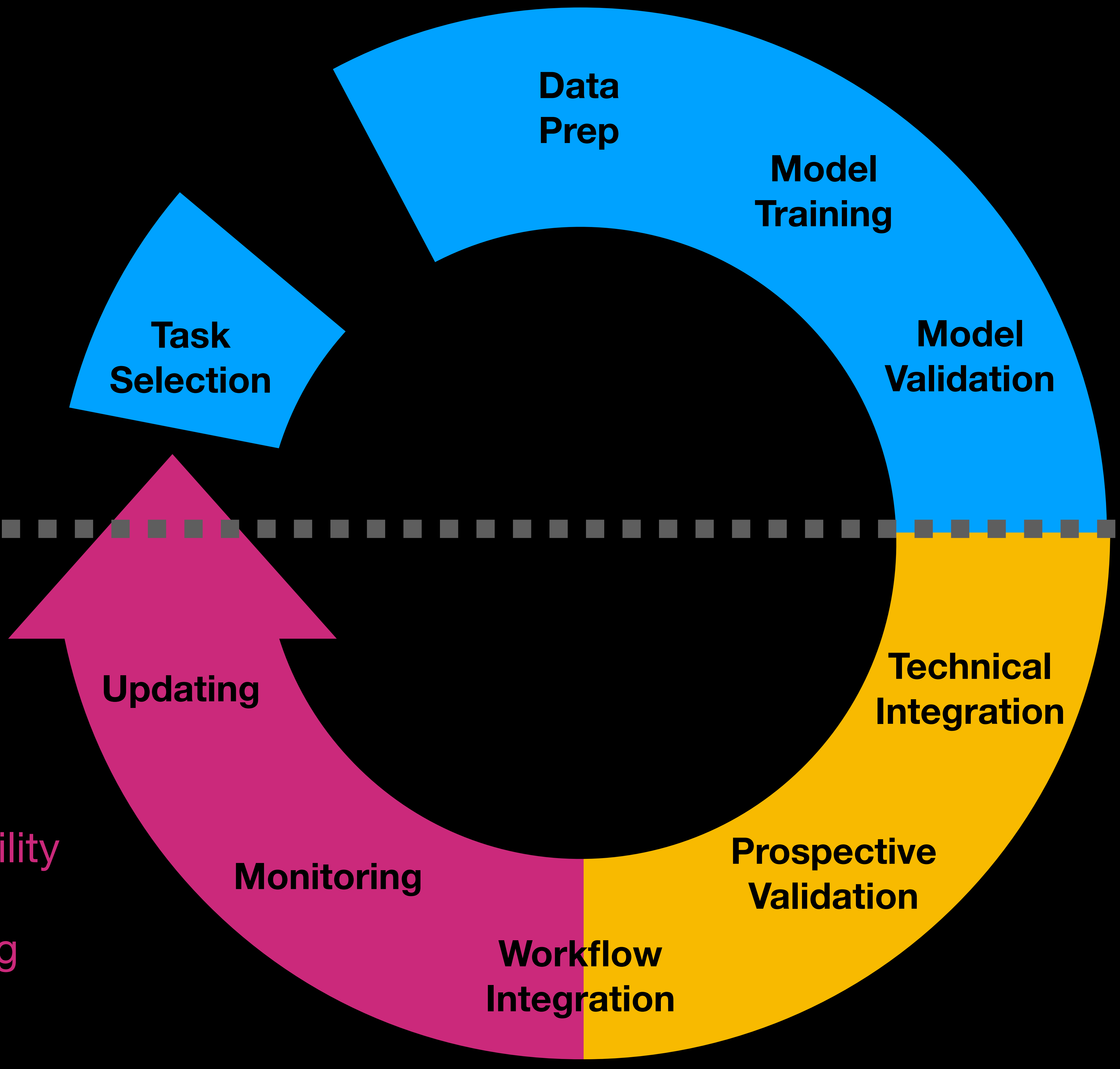
Ötleş *et al.*, MLHC '23

Performance Gap

Characterize changes to prospective performance after model implementation

Ötleş *et al.*, MLHC '21

The ML Lifecycle



RTW Prediction

Assess value of
longitudinal observations
for return to work prediction

Ötleş *et al.*, JAMIA '22

Updating Models

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

Ötleş *et al.*, MLHC '23

Performance Gap

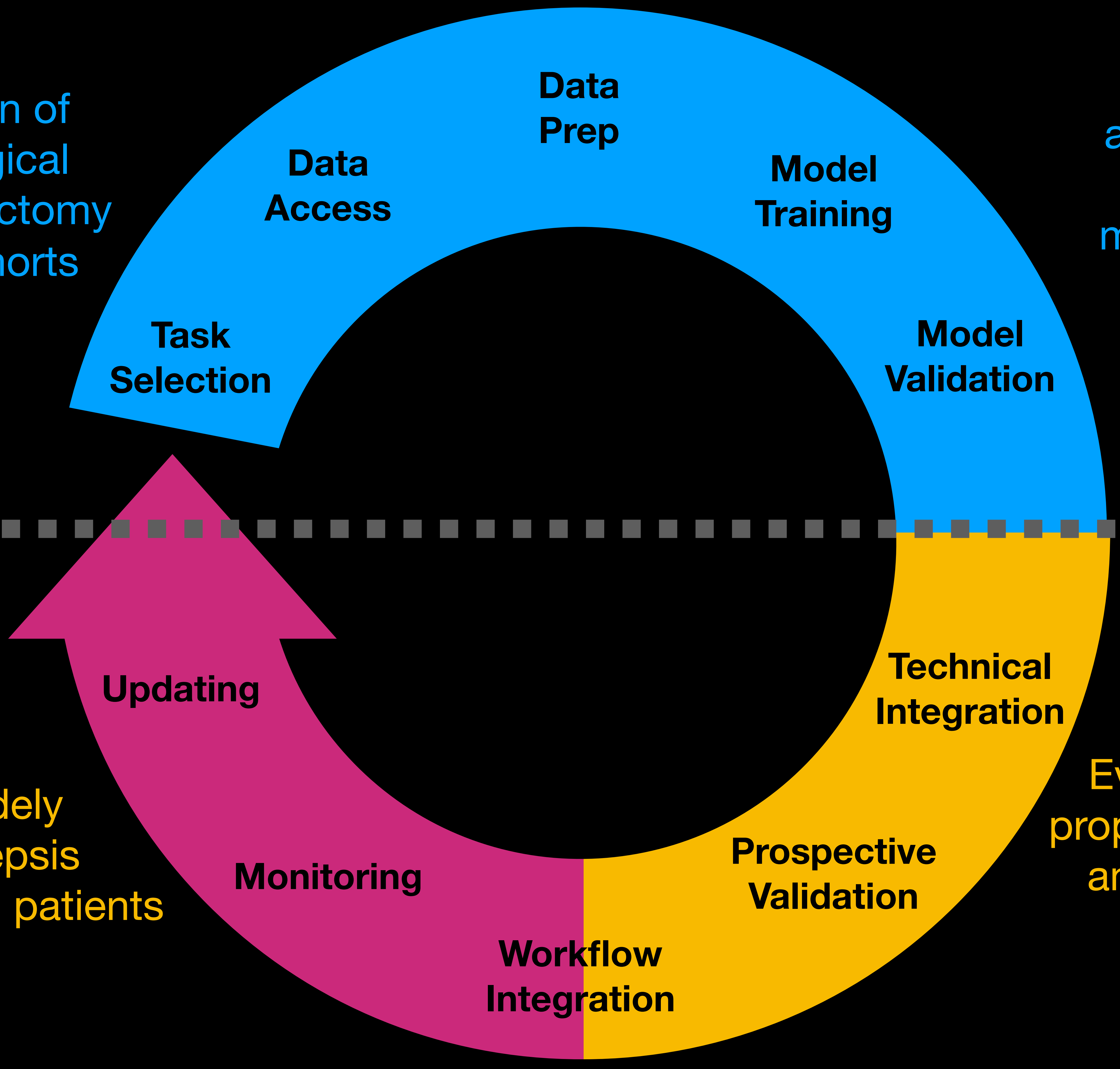
Characterize changes
to prospective performance
after model implementation

Ötleş *et al.*, MLHC '21

Other projects

Development and validation of models to predict pathological outcomes of radical prostatectomy in regional and national cohorts
Ötleş *et al.*, 2022

Early identification of patients admitted to hospital for covid-19 at risk of clinical deterioration: model development and multisite external validation study
Kamran *et al.*, 2022

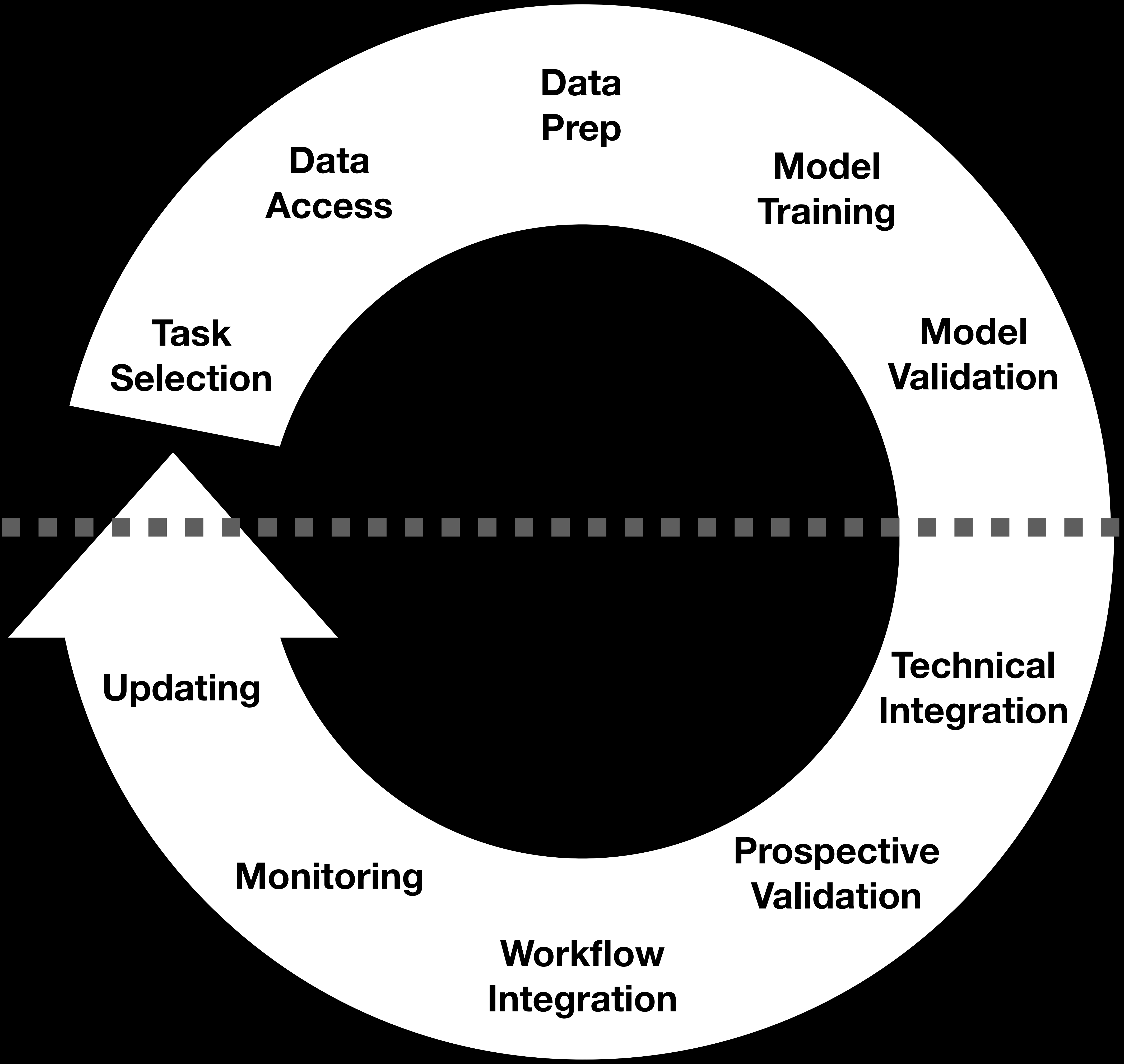


External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients
Wong *et al.*, 2022

Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19
Singh *et al.*, 2021

Time plays a crucial role in the development and implementation of ML models used in healthcare.

Development



Implementation

Future directions

Developing models less sensitive to infrastructure shifts

Updating

- Personalized updating
- Understanding user preferences for model updating
- Updating models over time to ensure best team performance

Thesis: by focusing on the longitudinal nature of healthcare, we can improve the performance of predictive models during development & implementation.