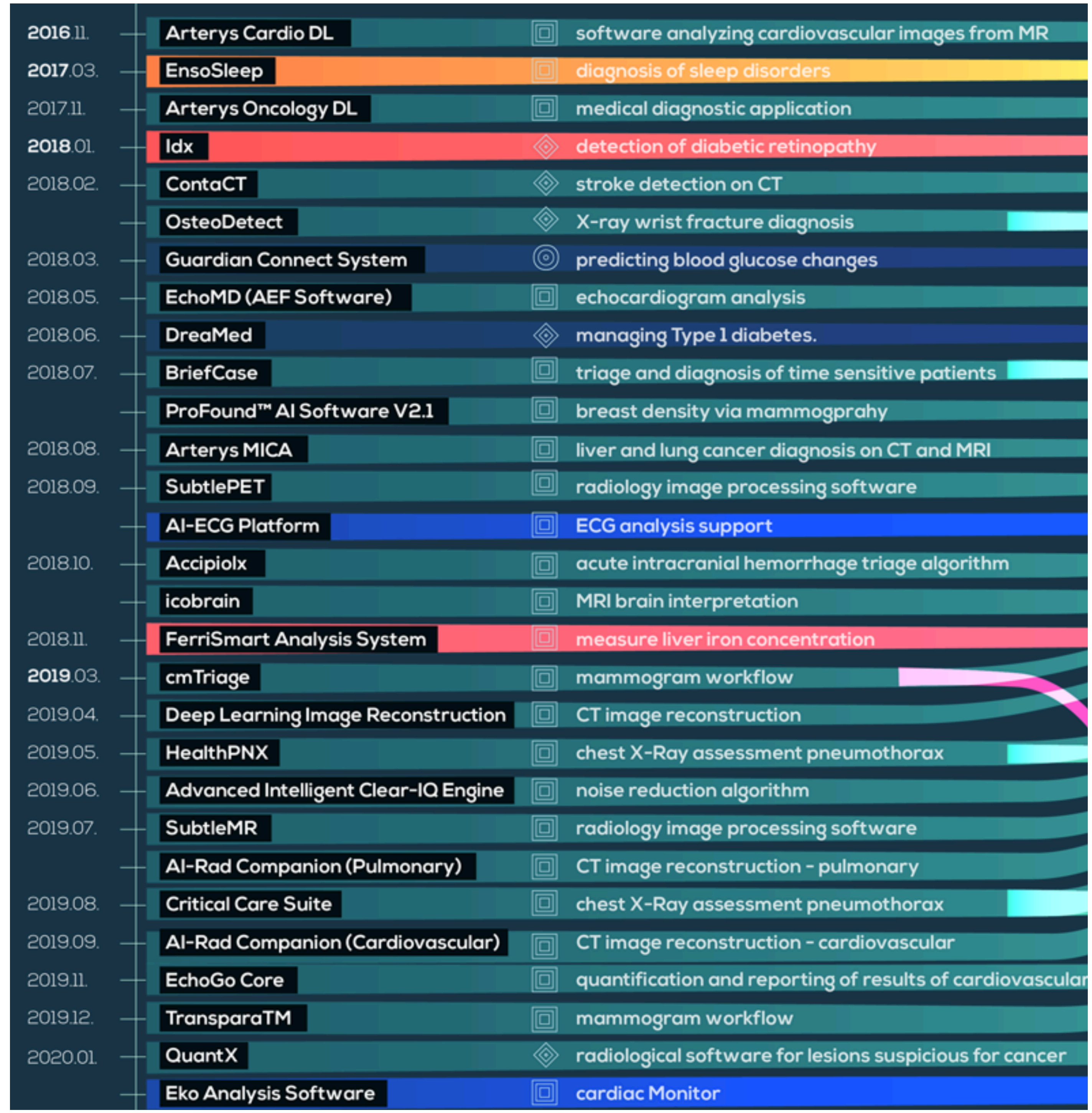
Approval policies for modifications to machine learning-based software as a medical device: A study of bio-creep

Jean Feng, Scott Emerson, Noah Simon
Biometrics 2021

Journal Club: April 28, 2022

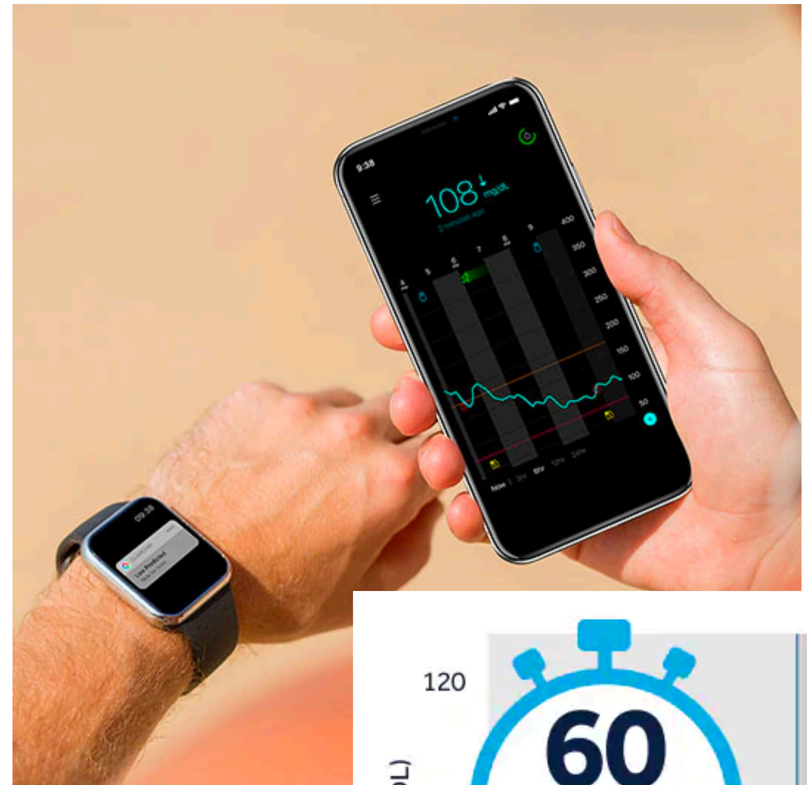
FDA Approvals for Artificial Intelligence/ Machine Learning-based Software-as-a-Medical-Device (SaMD)



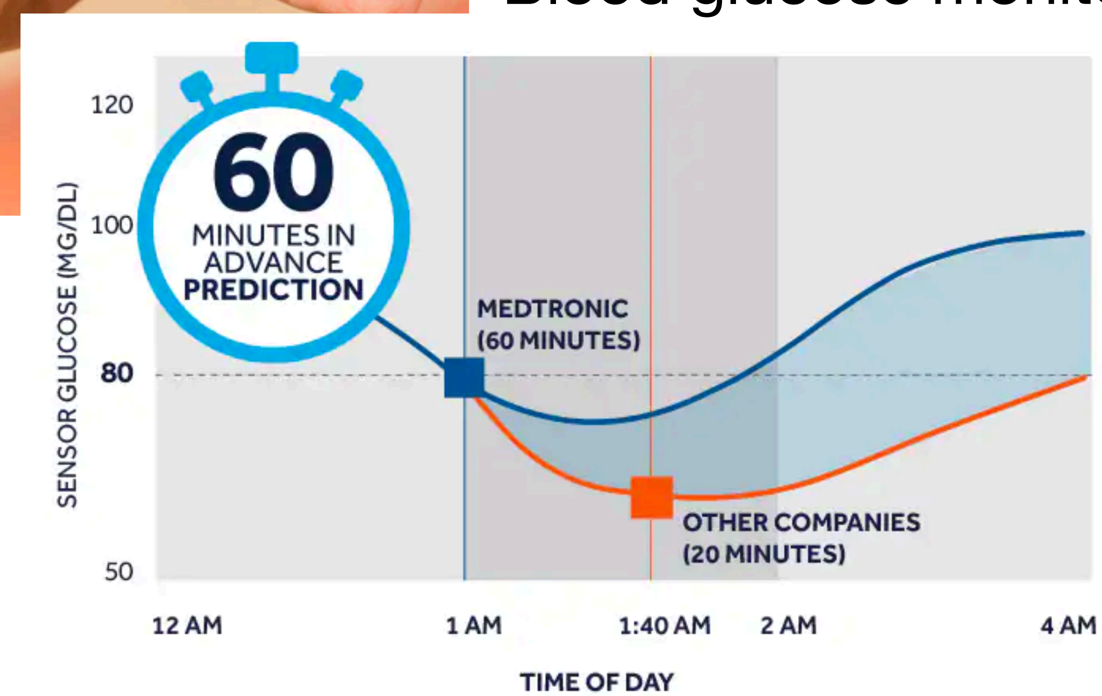
Benjamens, et. al. 2020

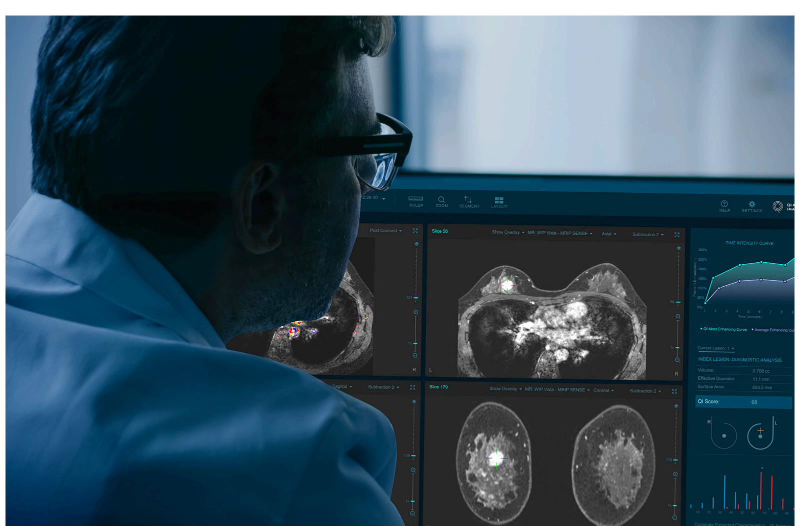
Examples

IDx-DR:
Diabetic retinopathy and macular edema



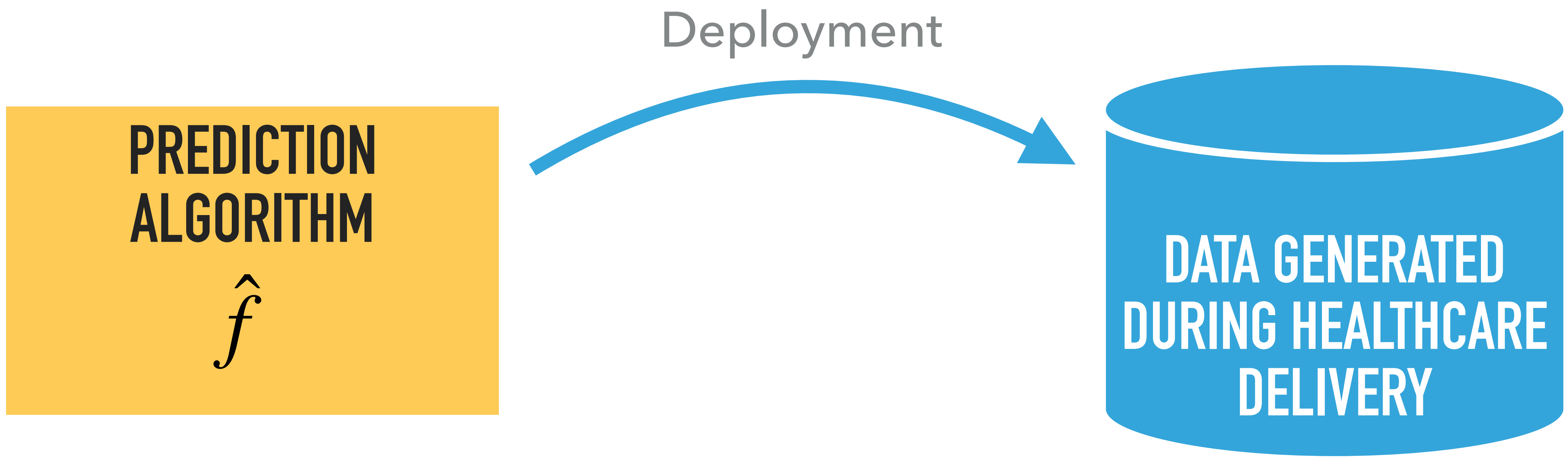
The Guardian Connect System, Medtronic:
Blood glucose monitor



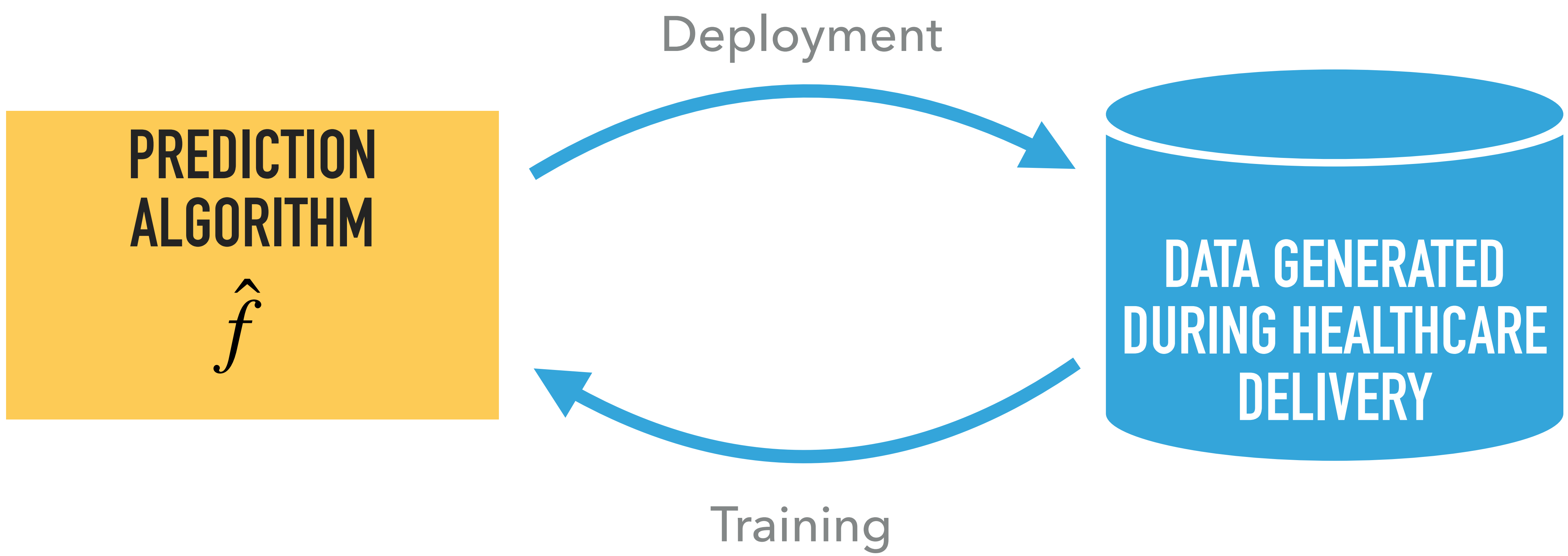


QuantX: Diagnose breast abnormalities

Machine learning in healthcare



Online machine learning in healthcare

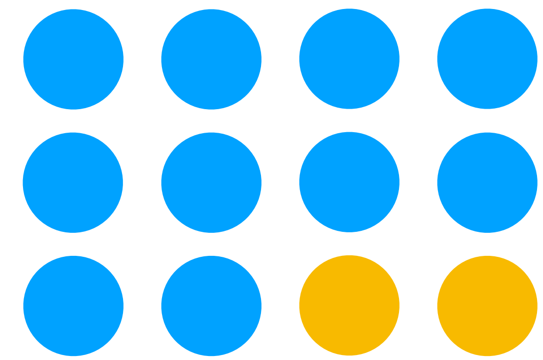


Iteration cycle in…

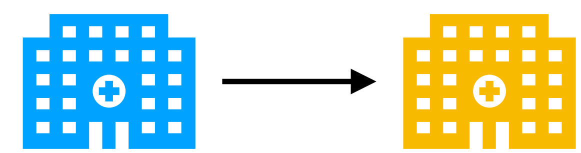
Drug development	ML algorithm development
Years	Days or Weeks

Online learning: Benefits

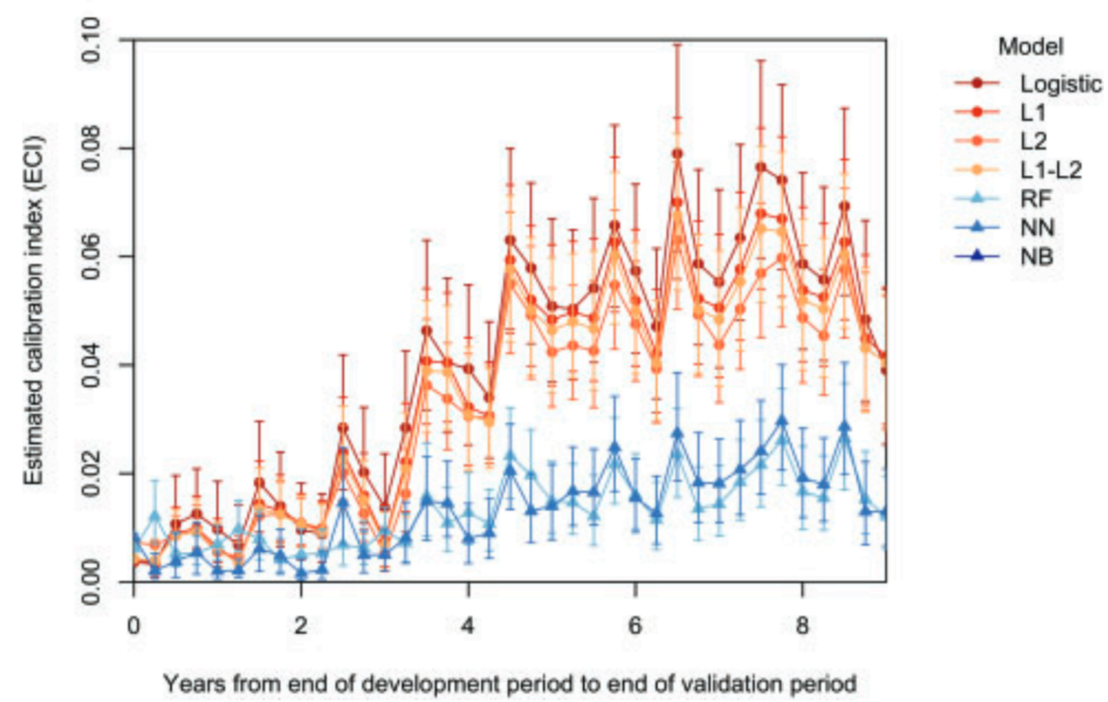
- Improve performance on average and/or within subpopulations



- Localize a model to a new medical site



- Adapt to distribution shifts



Davis 2017

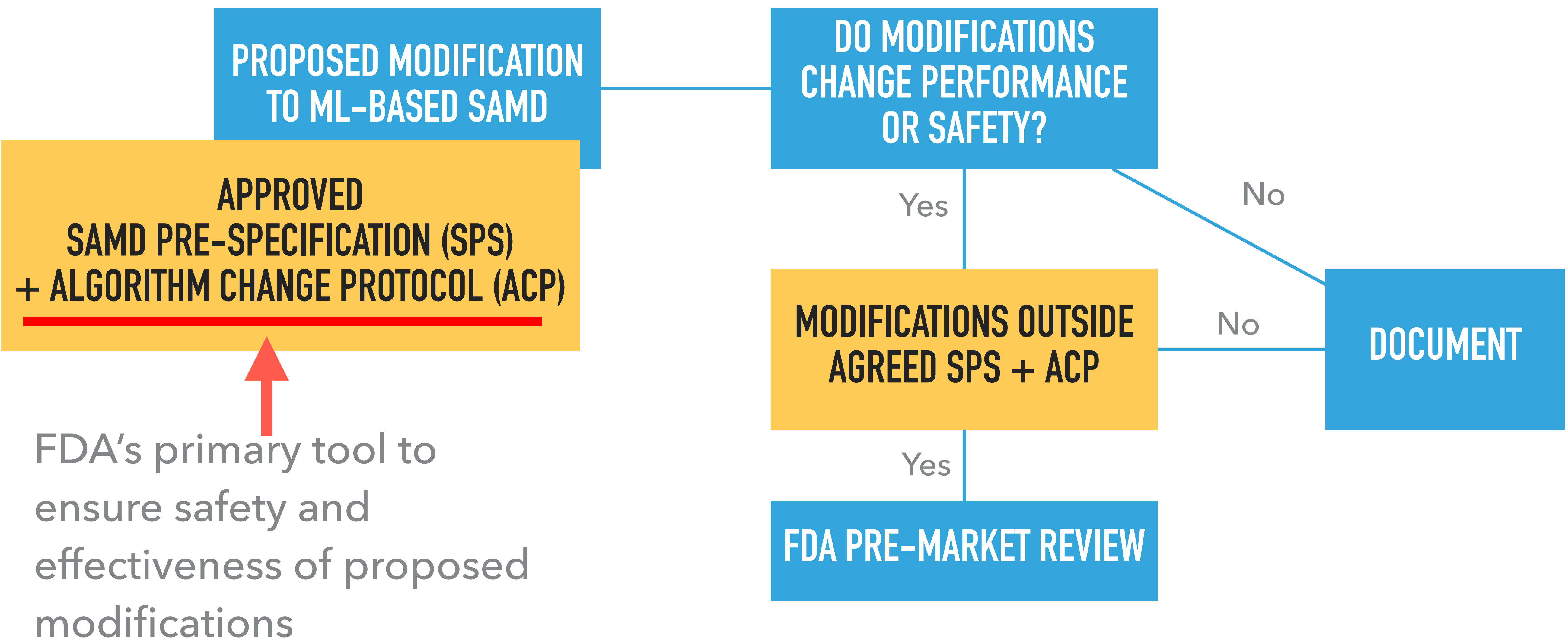
- …

Online learning: Risks

- Algorithmic modifications are not guaranteed to improve performance due to:
  - Over-updating
  - Catastrophic forgetting
  - Feedback cycles
  - Multiple hypothesis testing
  - Observational data and confounding
  - Machine-human interaction
  - Data quality
  - …

Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)

Discussion Paper and Request for Feedback



Algorithm change protocols with statistical guarantees

1. **Online hypothesis testing**
   - Feng, Jean, Scott Emerson, and Noah Simon. 2021. “Approval Policies for Modifications to Machine Learning-Based Software as a Medical Device: A Study of Bio-Creep.” Biometrics.

2. **Game-theoretic online learning**
   - Feng, Jean. 2021. “Learning to Safely Approve Updates to Machine Learning Algorithms.” *Proceedings of the Conference on Health, Inference, and Learning.*

3. **Bayesian inference**
   - Feng, Jean, Berkman Sahiner, Alexej Gossmann, and Romain Pirracchio. 2021. Bayesian logistic regression for online recalibration and revision of clinical prediction models with guarantees. Journal of the American Medical Informatics Association.

Problem statement

Design a performance evaluation component of the Algorithm Change Protocol (pACP) that approves good modifications quickly and controls the rate at which bad modifications are approved.

Steps:

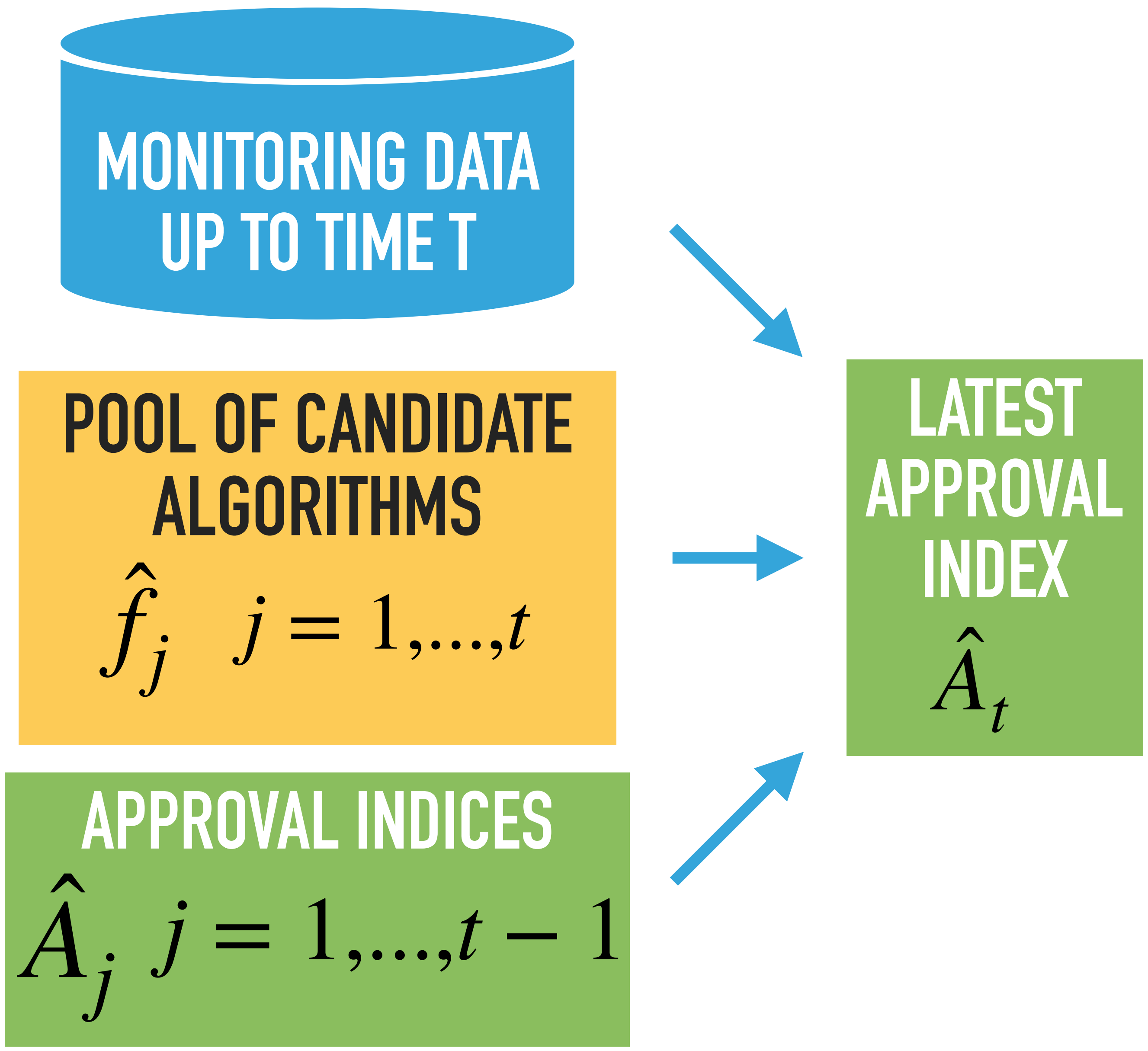
1) Define what an acceptable modification is.

2) Define a statistical framework for evaluating pACPs.

3) Design pACPs.

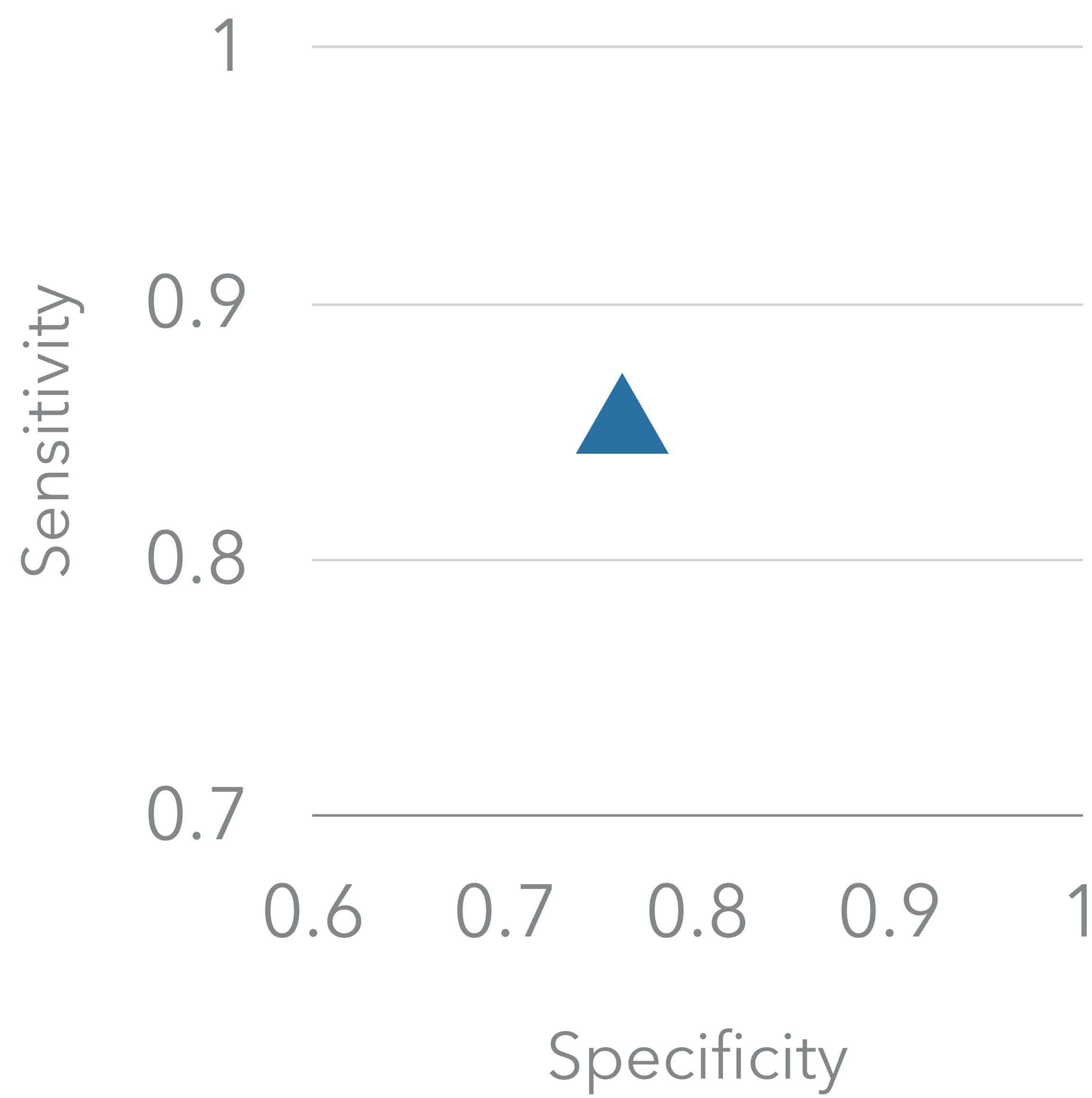
Problem Setup

- Let's start simple with IID data.
- At time points t = 1,2,…
  - Collect new batch of monitoring data $\{(x_{i,t}, y_{i,t}) : i = 1,\ldots,n\}$
  - Company proposes new candidate algorithm $\hat{f}_t$
  - The index of the most recently approved algorithm by the pACP is $\hat{A}_t$

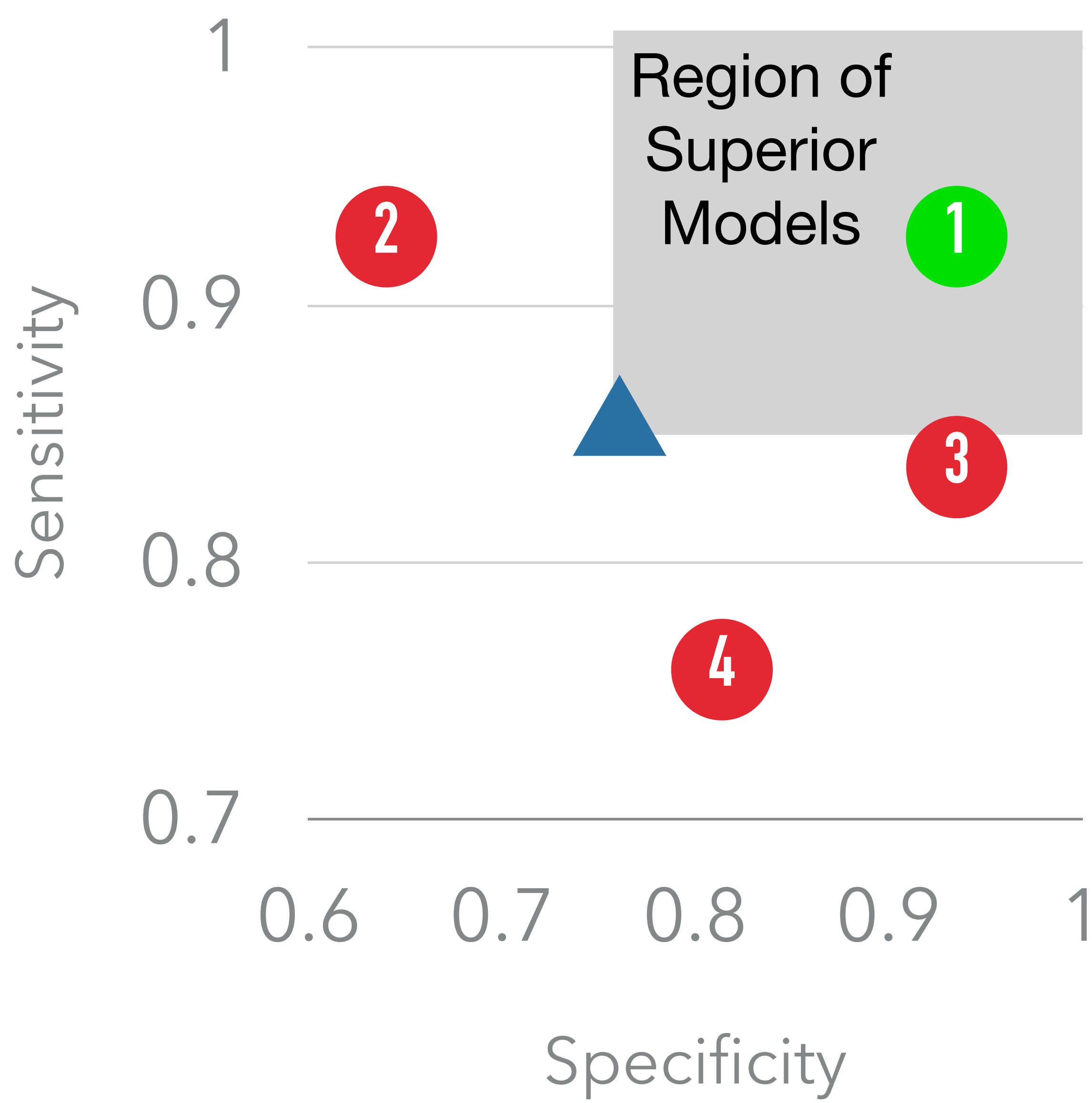


Performance evaluation

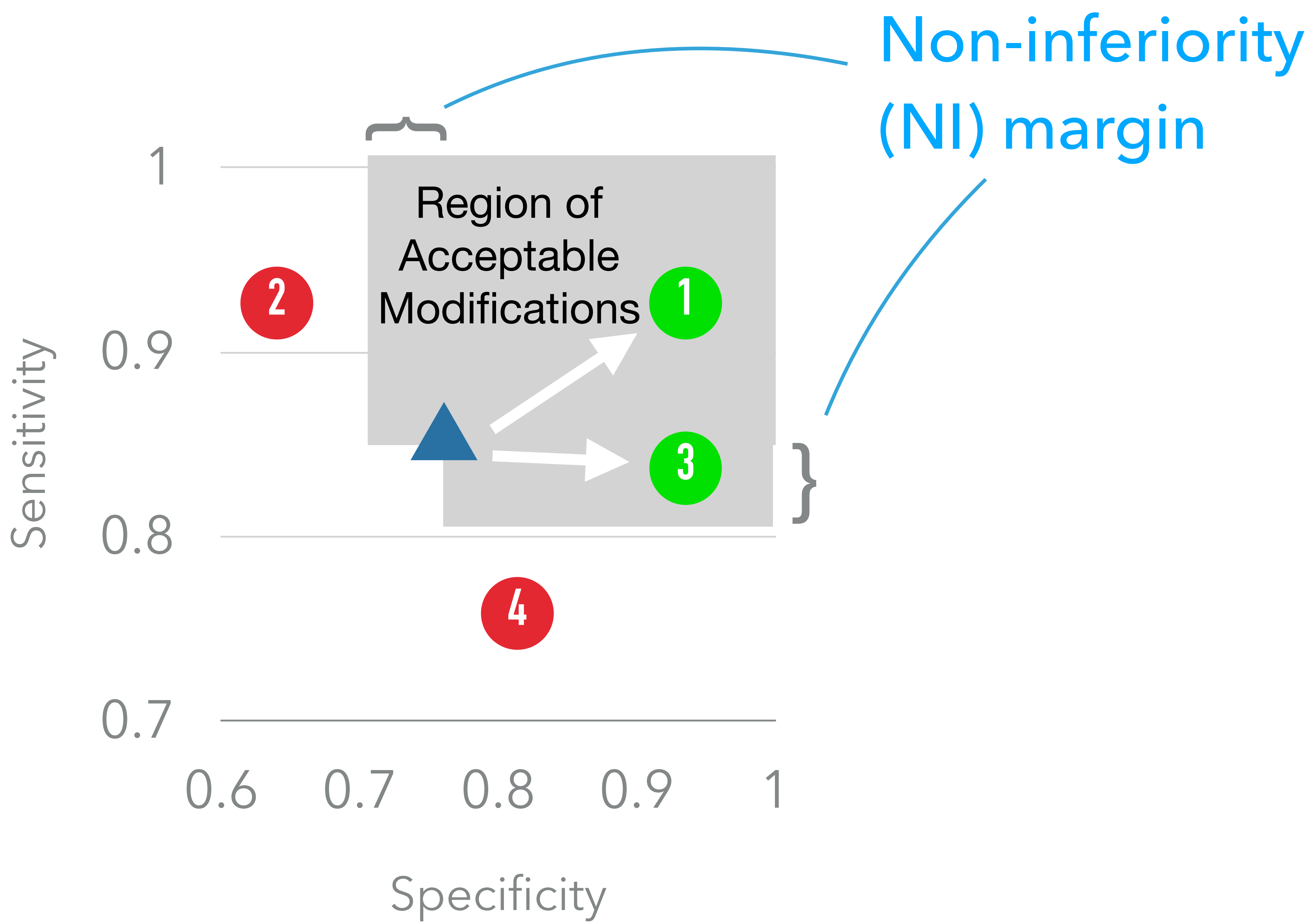
In practice, a model is evaluated using multiple performance metrics.



What is an acceptable modification?



What is an acceptable modification?



Acceptable modifications

<u>Definition</u>: A modification from algorithm f to f' is acceptable for non-inferiority margin ϵ, $f \rightarrow_{\epsilon} f'$, if it is:

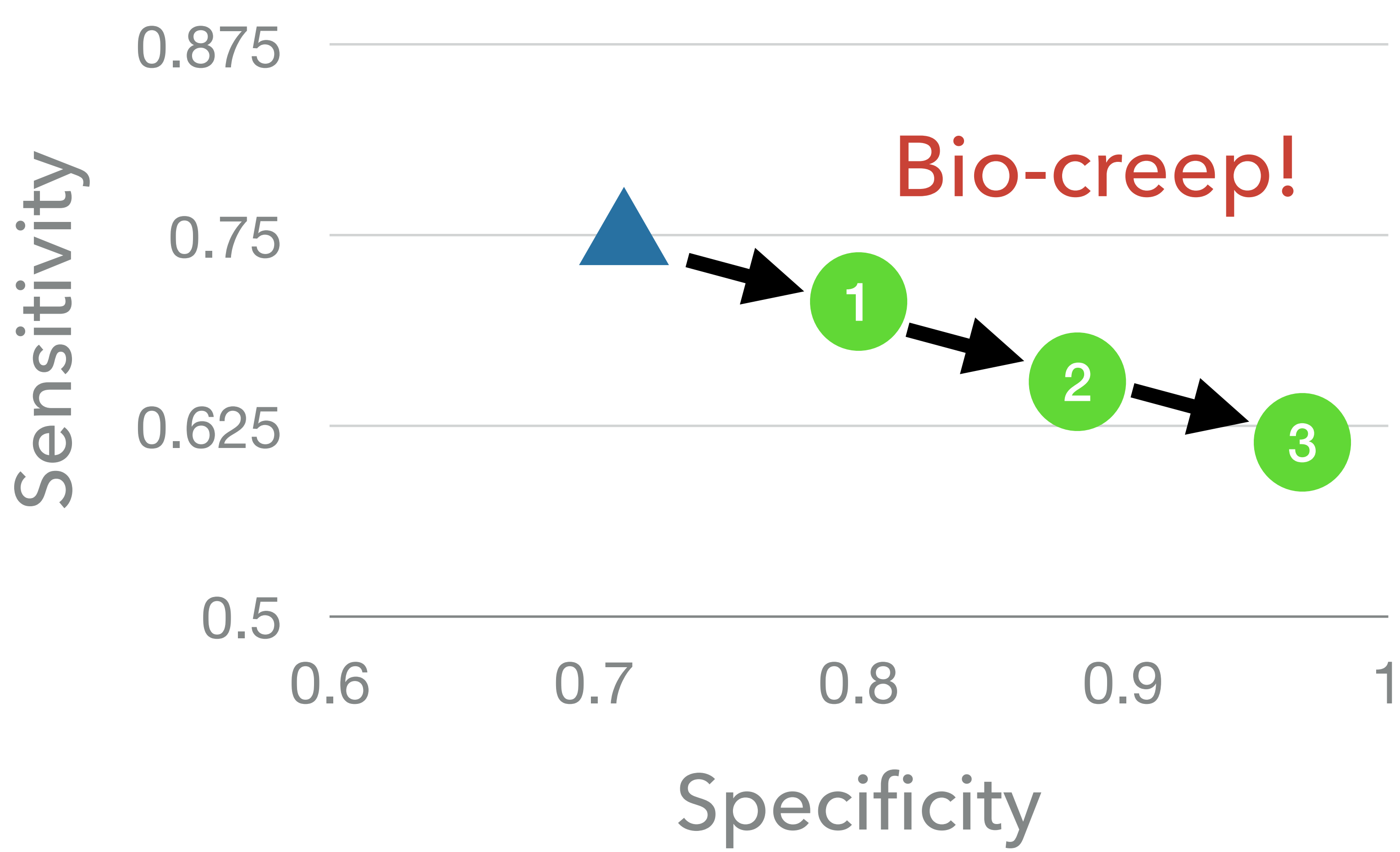
- Non-inferior with respect to all metrics
  $m_k(f) - \epsilon \leq m_k(f') \quad \forall k = 1,...,K$

- Superior in at least one metric
  $m_k(f) < m_k(f') \quad \exists k \in \{1,...,K\}$

Online error for a pACP

- **Definition:** The expected bad approval count at time T

$$\mathrm{BAC}(T) = \mathbb{E}\left[\sum_{t=1}^{T} 1\left\{\text{Approved unacceptable modification at time } t\right\}\right]$$



Online error for a pACP

- **Definition:** The expected bad approval count at time T

$$\mathrm{BAC}(T) = \mathbb{E}\left[\sum_{t=1}^{T} 1\left\{\exists t' = 1,\ldots,t-1 \text{ s.t. } \hat{f}_{\hat{A}_{t'}} \not\rightarrow_{\epsilon} \hat{f}_{\hat{A}_t}\right\}\right]$$

"FWER"

- **Definition:** The expected bad approval and benchmark ratio at time T

$$\mathrm{BABR}(T) = \mathbb{E}\left[\frac{\sum_{t=1}^{T} 1\left\{\exists t' = 1,\ldots,t-1 \text{ s.t. } \hat{f}_{\hat{A}_{t'}} \not\rightarrow_{\epsilon} \hat{f}_{\hat{A}_t}\right\}}{1 + \sum_{t=1}^{T} 1\left\{\hat{B}_t \neq \hat{B}_{t-1}\right\}}\right]$$

"FDR"

A zoo of pACPs

- **Without error rate control:**
  - **pACP-Blind**: Approve everything
  - **pACP-Reset**: Compare to the latest approval with fixed p-value threshold
- **With error rate control:**
  - **pACP-Locked**: Do not approve anything
  - **pACP-BAC**: Controls expected Bad Approval Count using alpha-spending, group-sequential, and gate-keeping methods
  - **pACP-BABR**: Controls expected Bad Approval and Benchmark Ratios using alpha-investing, group-sequential, and gate-keeping methods

A simple protocol with no error control

pACP-Reset

Select fixed level α. At time t = 1,2,…

- For each candidate modification $\hat{f}_{t'}$, test if it is acceptable to the currently approved model $\hat{f}_{\hat{A}_t}$ ($H^0 : \hat{f}_{\hat{A}_t} \not\to_{\epsilon} \hat{f}_{t'}$) using prospectively-collected monitoring data.
- Approve the latest modification with p-value smaller than α

Controlling BAC

pACP-BAC

At time t = 1,2,…

- Pre-specify testing procedure for new candidate $\hat{f}_t$: Test the following sequence of null hypotheses using significant thresholds selected using **alpha-spending** and **group-sequential methods**.

  - $H_1^0 : \hat{f}_{\hat{A}_1} \nrightarrow_\epsilon \hat{f}_t$
  - $H_2^0 : \hat{f}_{\hat{A}_2} \nrightarrow_\epsilon \hat{f}_t$
  - …
  - $H_t^0 : \hat{f}_{\hat{A}_t} \nrightarrow_\epsilon \hat{f}_t$

  Gate-keeping

- Evaluate all candidate algorithms using pre-specified procedure.
- Approve the latest modification that rejects all hypotheses.

A zoo of pACPs

- **Without error rate control:**
  - **pACP-Blind**: Approve everything
  - **pACP-Reset**: Compare to the latest approval with fixed p-value threshold
- **With error rate control:**
  - **pACP-Locked**: Do not approve anything
  - **pACP-BAC**: Controls expected Bad Approval Count using alpha-spending, group-sequential, and gate-keeping methods
  - **pACP-BABR**: Controls expected Bad Approval and Benchmark Ratios using alpha-investing, group-sequential, and gate-keeping methods

Simulation studies

- Desired properties

  1. Low rate of bad approvals

  2. High rate of good approvals

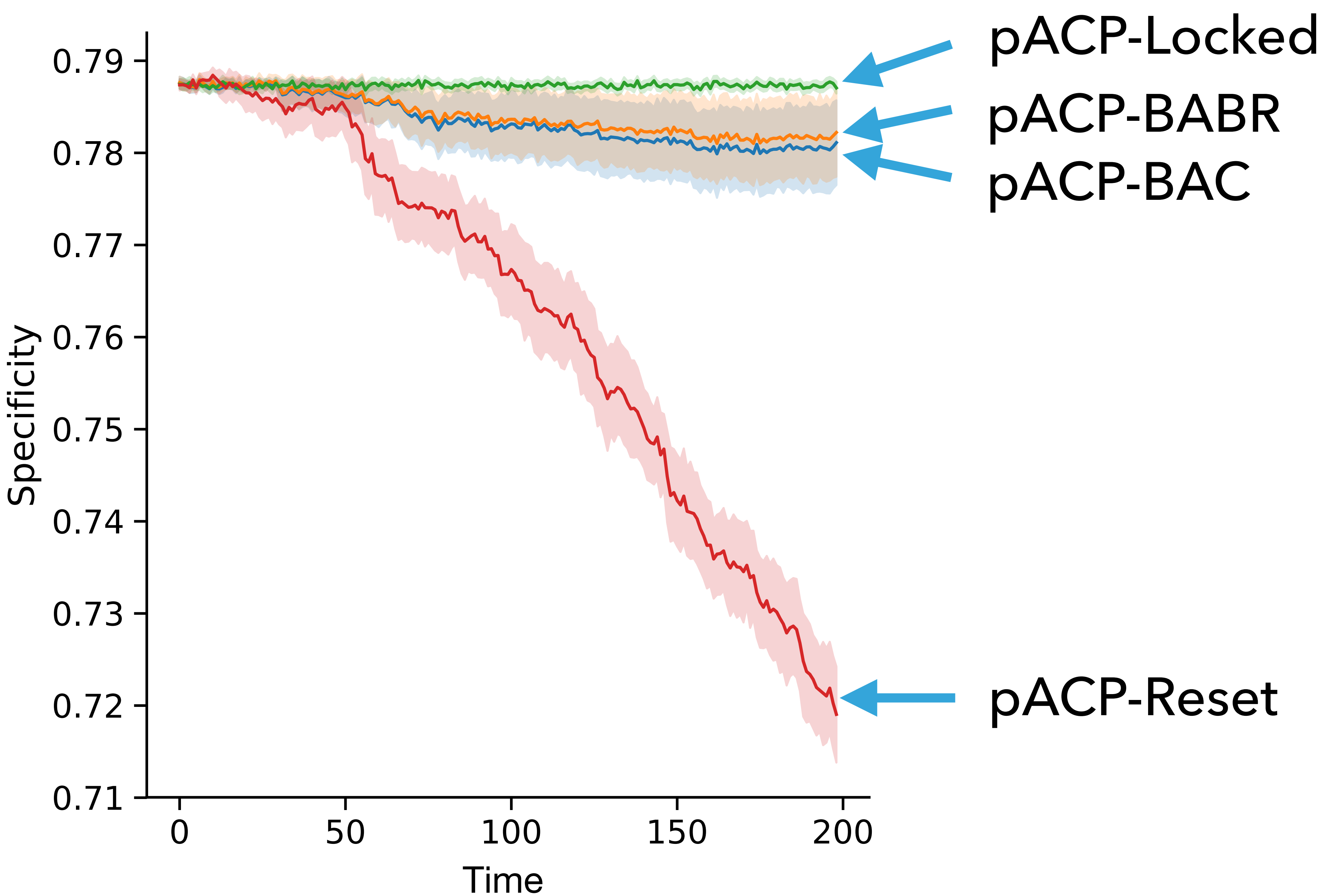
- Setup

  - Monitoring data is IID at each time point and across time points

  - Binary prediction problem

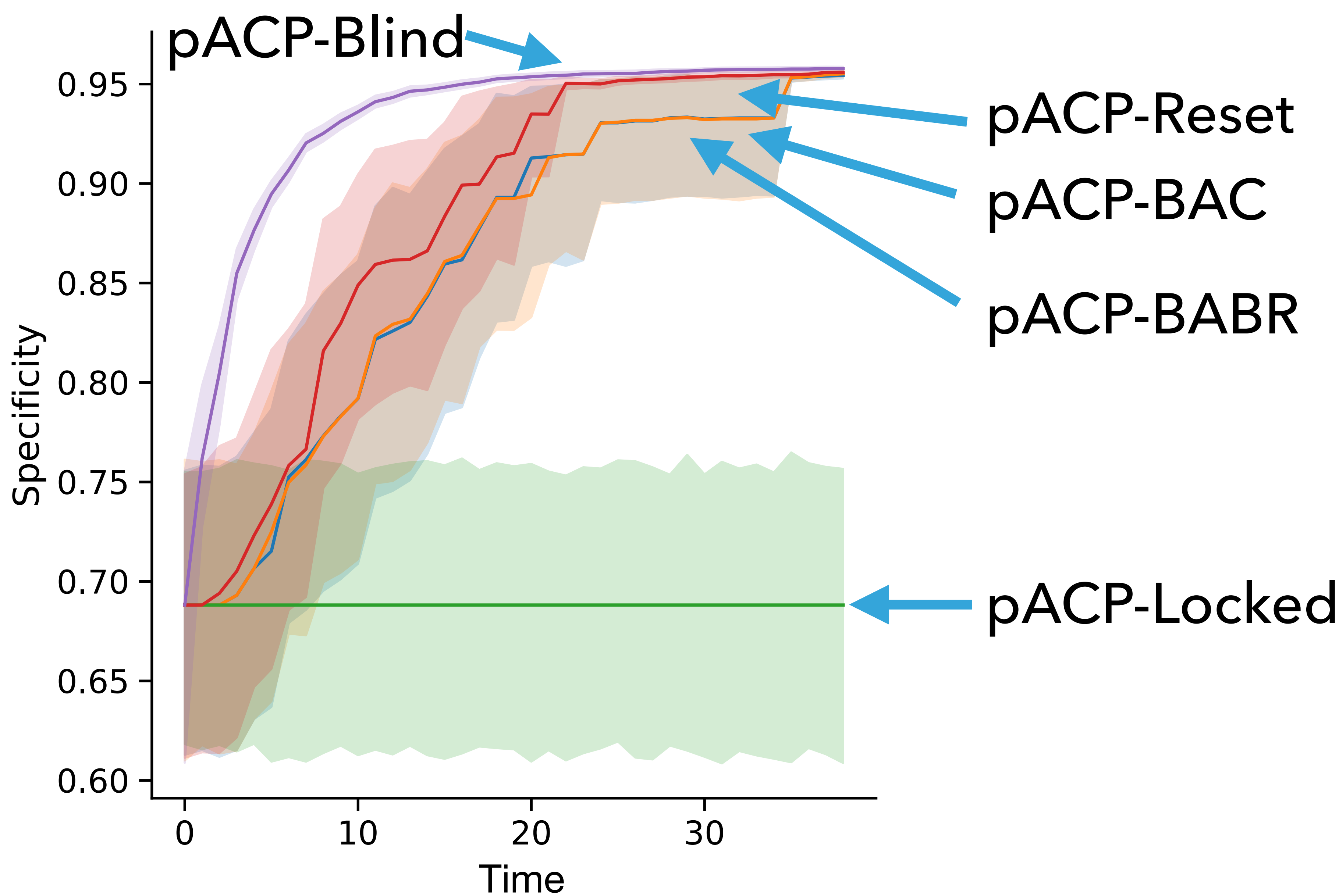
Simulation: mostly deleterious modifications

Proposed modifications deteriorate over time



Simulation: mostly beneficial modifications

Train new models using the accumulating monitoring data



Summary

- Bio-creep is a concern, even in this idealized scenario with IID data. *Designing a pACP cannot be taken lightly!*

- If we carefully design pACPs, we can approve good modifications quickly while protecting against bad modifications.

Algorithm change protocols with statistical guarantees

1. Online hypothesis testing

- Feng, Jean, Scott Emerson, and Noah Simon. 2021. "Approval Policies for Modifications to Machine Learning-Based Software as a Medical Device: A Study of Bio-Creep." Biometrics.

- *Black-box modifications*
- *Stationary data*

2. Game-theoretic online learning

- Feng, Jean. 2021. "Learning to Safely Approve Updates to Machine Learning Algorithms." *Proceedings of the Conference on Health, Inference, and Learning.*

- *Black-box modifications*
- *Nonstationary data*
- *Faster approval*

3. Bayesian inference

- Feng, Jean, Berkman Sahiner, Alexej Gossmann, and Romain Pirracchio. 2021. Bayesian logistic regression for online recalibration and revision of clinical prediction models with guarantees. Journal of the American Medical Informatics Association.

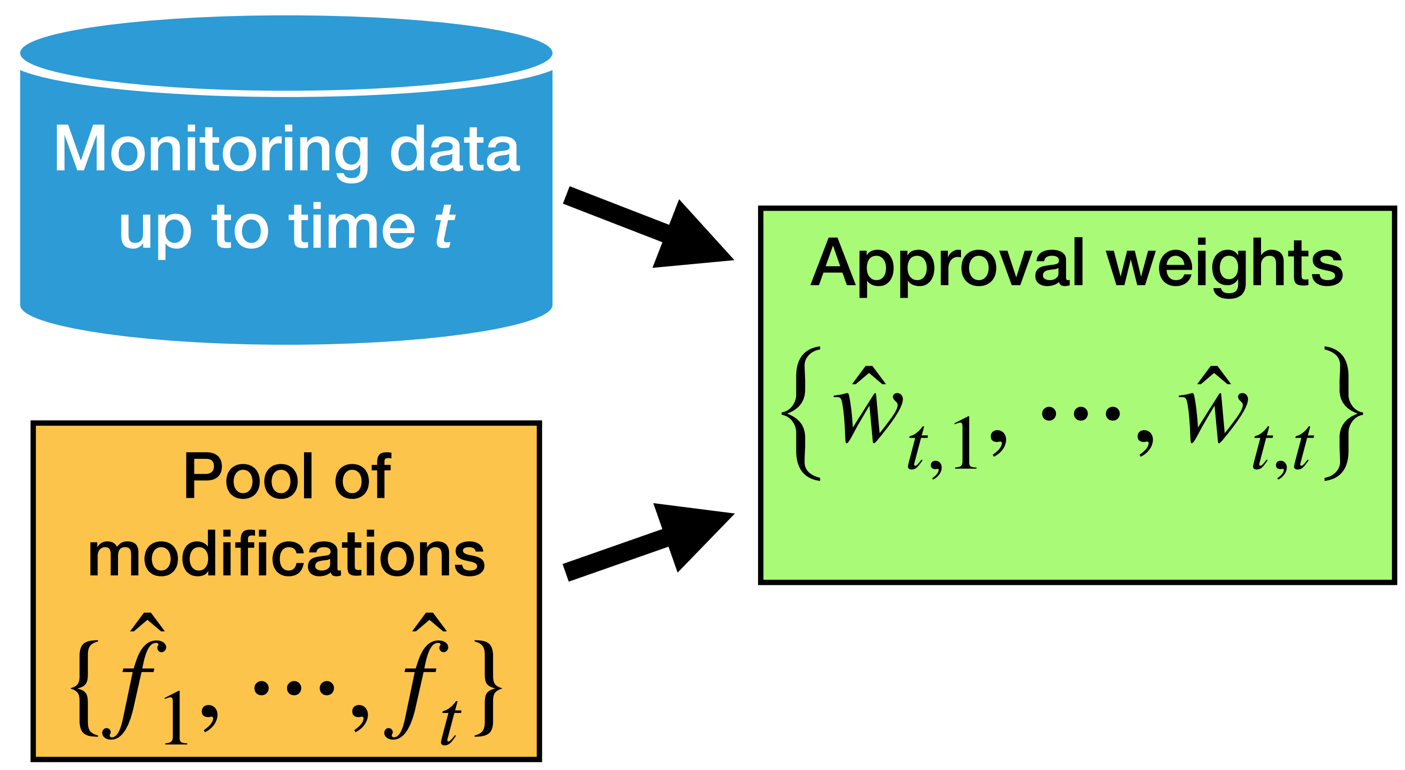
Approach 2: Game-theoretic online learning

- Game-theoretic online learning procedures provide performance guarantees under ***arbitrary distribution shifts*** in terms of regret bounds.

- These guarantees are weak when sample sizes are small, which is common in medical settings.

- We developed a new algorithm called "Learning to approve" (L2A), which ***dynamically weights black-box modifications*** based on their past performance.

  ➜ Faster approval



Algorithm change protocols with statistical guarantees

1. Online hypothesis testing

- Feng, Jean, Scott Emerson, and Noah Simon. 2021. “Approval Policies for Modifications to Machine Learning-Based Software as a Medical Device: A Study of Bio-Creep.” Biometrics.

- *Black-box modifications*
- *Stationary data*

2. Game-theoretic online learning

- Feng, Jean. 2021. “Learning to Safely Approve Updates to Machine Learning Algorithms.” *Proceedings of the Conference on Health, Inference, and Learning.*

- *Black-box modifications*
- *Nonstationary data*
- *Faster approval*

3. Bayesian inference

- Feng, Jean, Berkman Sahiner, Alexej Gossmann, and Romain Pirracchio. 2021. Bayesian logistic regression for online recalibration and revision of clinical prediction models with guarantees. Journal of the American Medical Informatics Association.

- *Parametric modifications*
- *Nonstationary data*
- *Fastest approval rates*

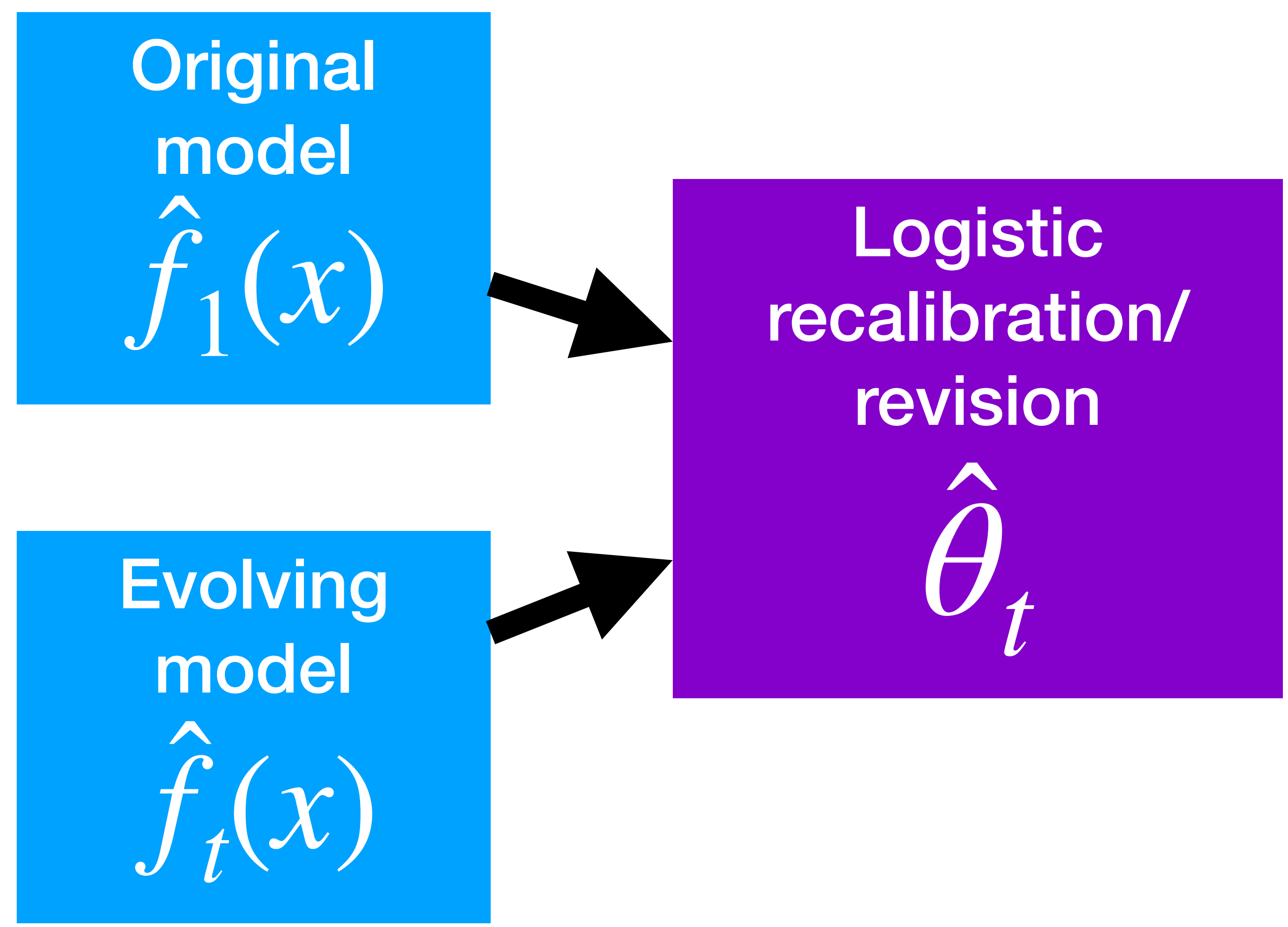
Approach 3: Bayesian inference

- In practice, the most common modification applied to ML algorithms is ***logistic recalibration or revision***.

- We can continually update the parameters of a logistic recalibration/revision model using Bayesian inference.

  ➜ ***Even faster approval***

- We derive regret bounds for Bayesian logistic recalibration/revision that hold under ***arbitrary distribution shifts***.



Algorithm change protocols with statistical guarantees

1. Online hypothesis testing

- Feng, Jean, Scott Emerson, and Noah Simon. 2021. “Approval Policies for Modifications to Machine Learning-Based Software as a Medical Device: A Study of Bio-Creep.” Biometrics.

- *Black-box modifications*
- *Stationary data*

2. Game-theoretic online learning

- Feng, Jean. 2021. “Learning to Safely Approve Updates to Machine Learning Algorithms.” *Proceedings of the Conference on Health, Inference, and Learning.*

- *Black-box modifications*
- *Nonstationary data*
- *Faster approval*

3. Bayesian inference

- Feng, Jean, Berkman Sahiner, Alexej Gossmann, and Romain Pirracchio. 2021. Bayesian logistic regression for online recalibration and revision of clinical prediction models with guarantees. Journal of the American Medical Informatics Association.

- *Parametric modifications*
- *Nonstationary data*
- *Fastest approval rates*

4. Others?

Acknowledgments

- Our team working on ML regulation
  - Scott Emerson (University of Washington)
  - Noah Simon (University of Washington)
  - Romain Pirracchio (UCSF)
  - Alexej Gossmann (FDA)
  - Berkman Sahiner (FDA)

- Support from the UCSF-Stanford CERSI program

  (Disclaimer: The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by FDA/HHS, or the U.S. Government.)