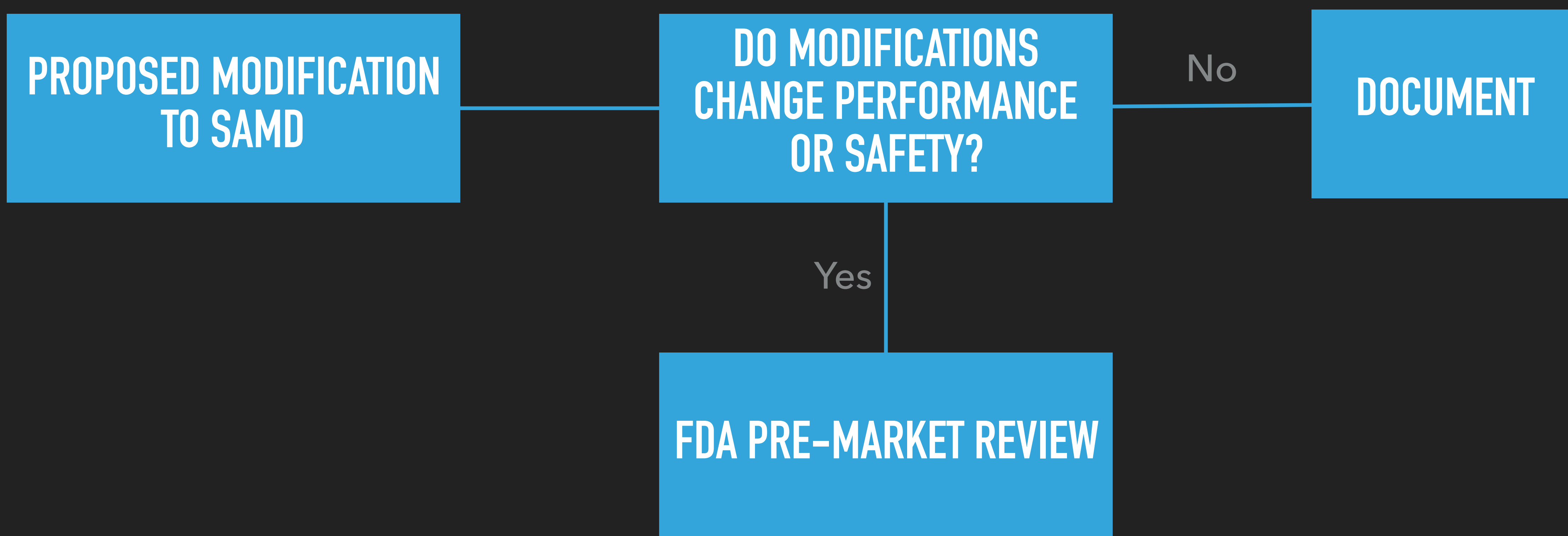
JEAN FENG

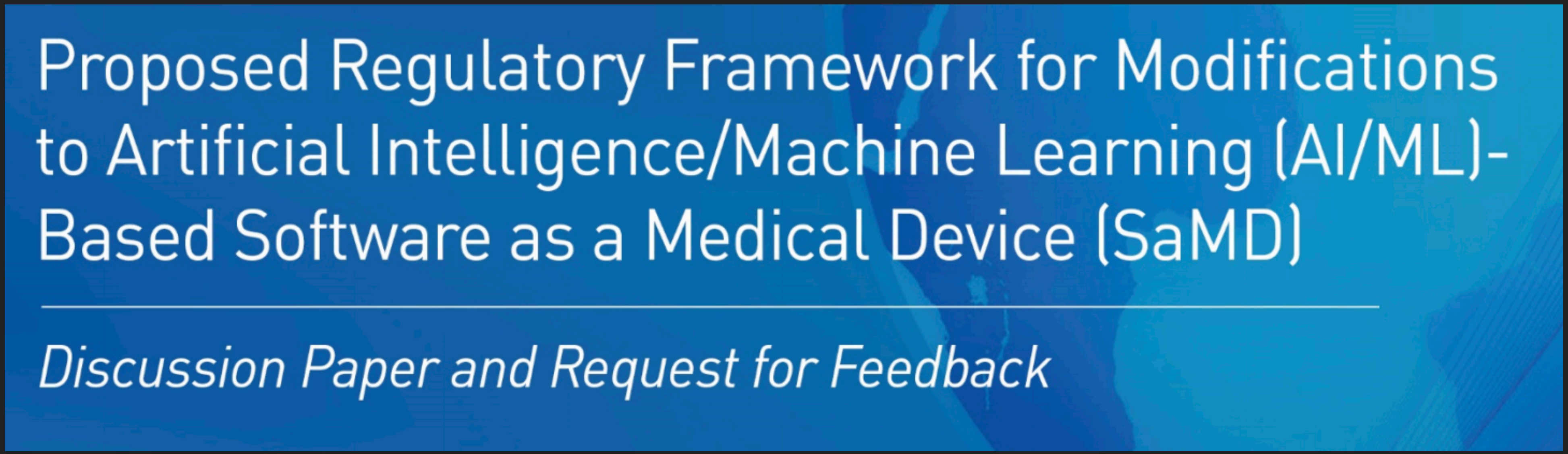
UNIVERSITY OF WASHINGTON

APPROVAL POLICIES FOR MODIFICATIONS TO MACHINE LEARNING-BASED SOFTWARE AS A MEDICAL DEVICE: A STUDY OF BIO-CREEP

CURRENT FDA POLICY FOR SOFTWARE AS A MEDICAL DEVICE (SAMD)

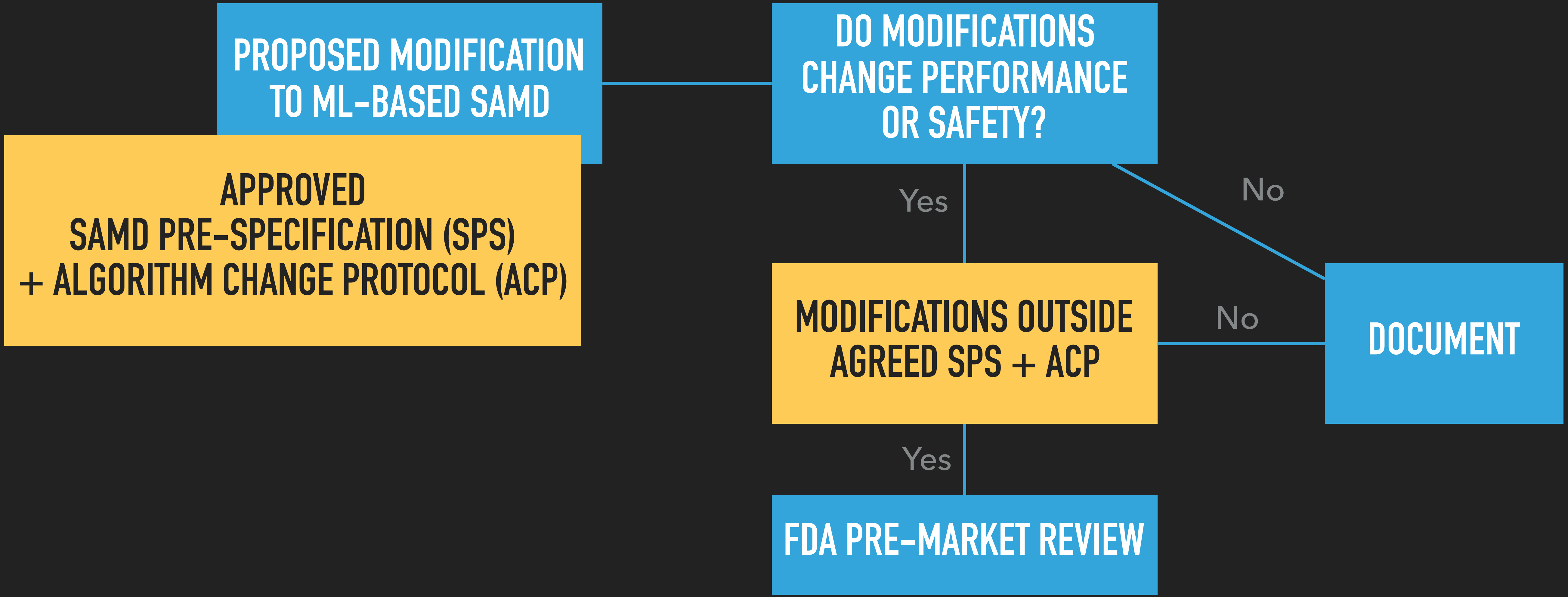


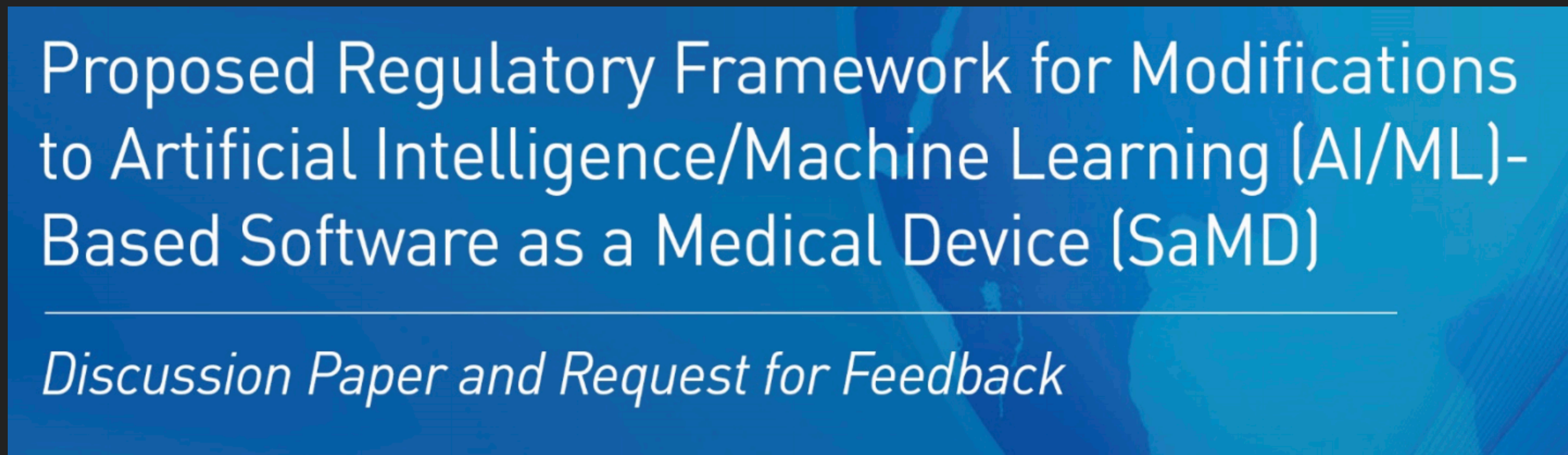
REGULATING ML



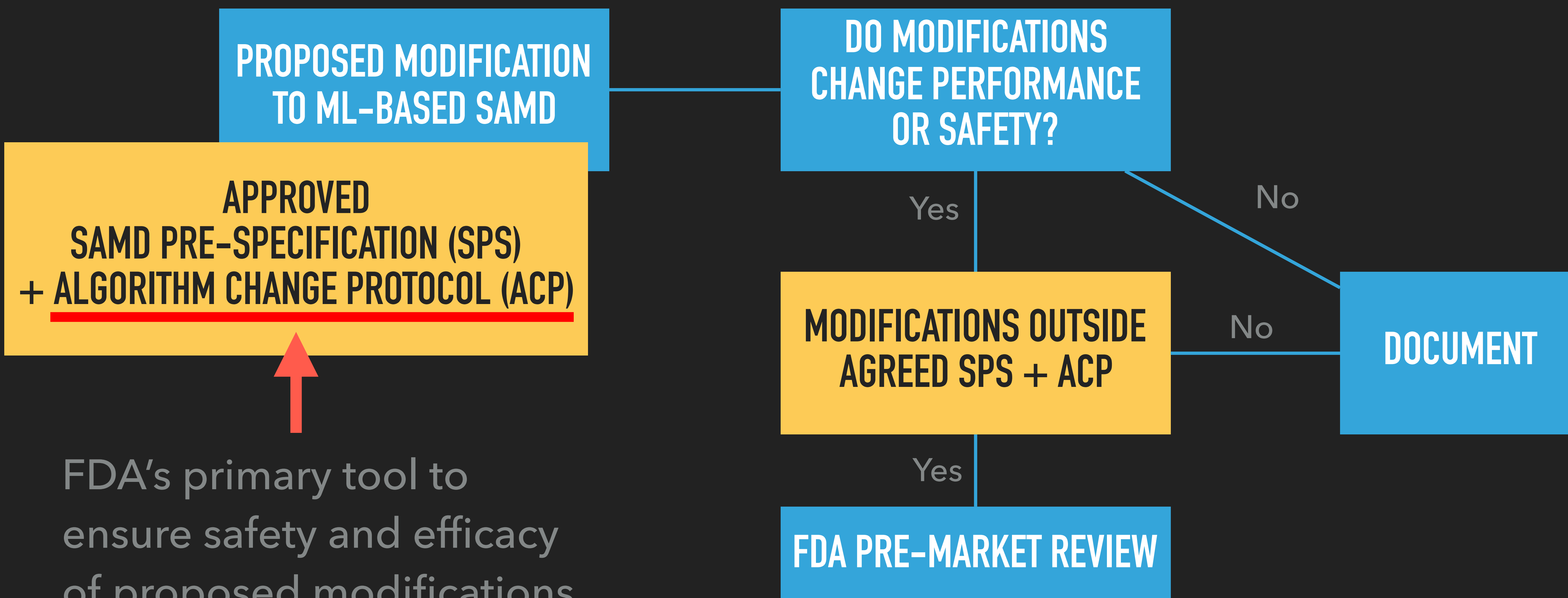
Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)

Discussion Paper and Request for Feedback




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



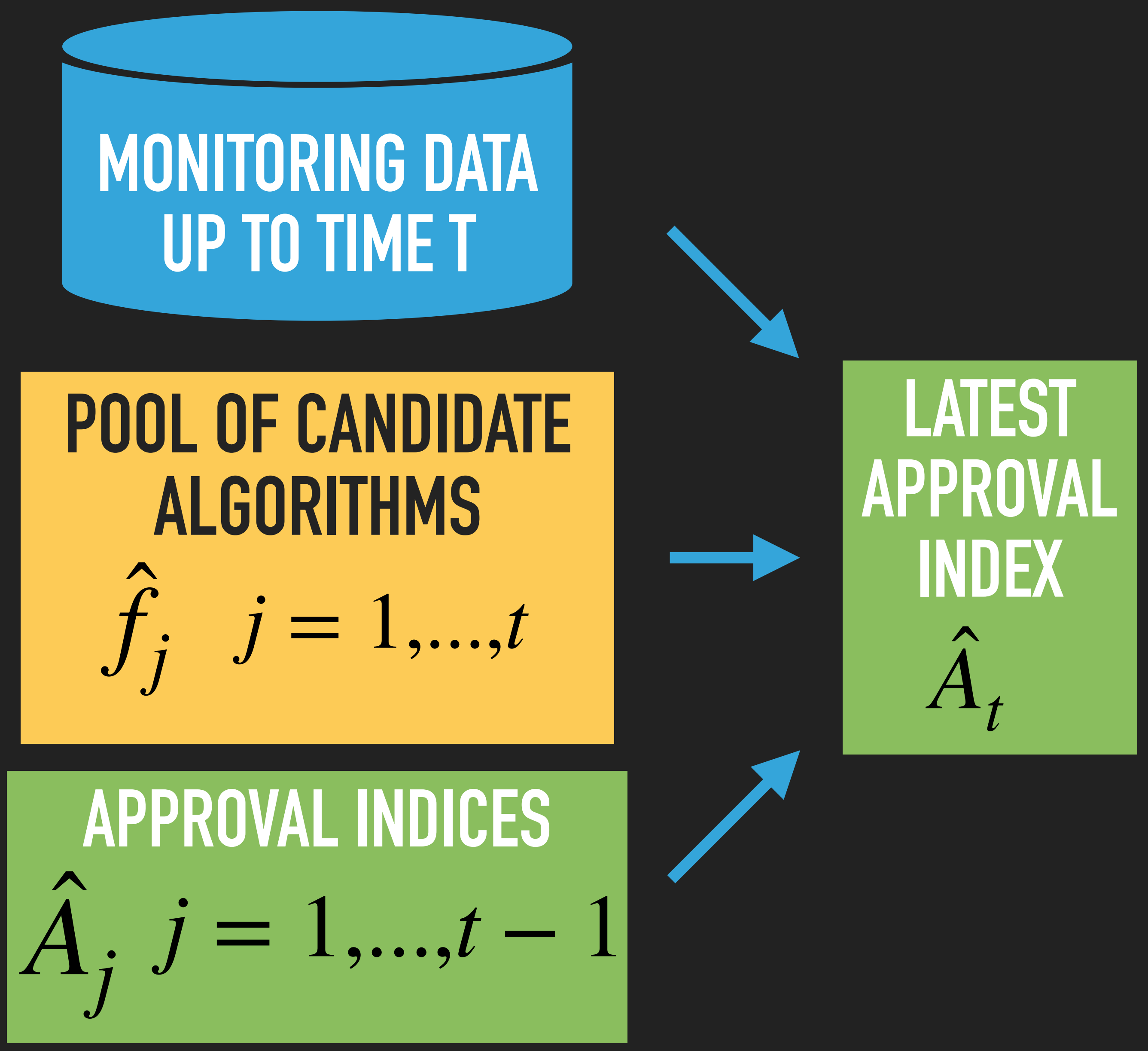
FDA's primary tool to ensure safety and efficacy of proposed modifications

PROBLEM SETUP

- **Automatic Algorithm Change Protocol (aACP)**: an ACP executed without human intervention
- ML-based SAMD is a black-box prediction model f.

PROBLEM SETUP

- At time points t = 1,2,…
  - Collect new batch of monitoring data
  - Company proposes new candidate algorithm $\hat{f}_t$
  - Index of the most recently approved algorithm is $\hat{A}_t$



GOAL

Design automatic Algorithm Change Protocols that approve good modifications quickly and control the rate at which bad modifications are approved.

GOAL

1) Define what an acceptable modification is.

2) Define a statistical framework for evaluating automatic Algorithm Change Protocols.

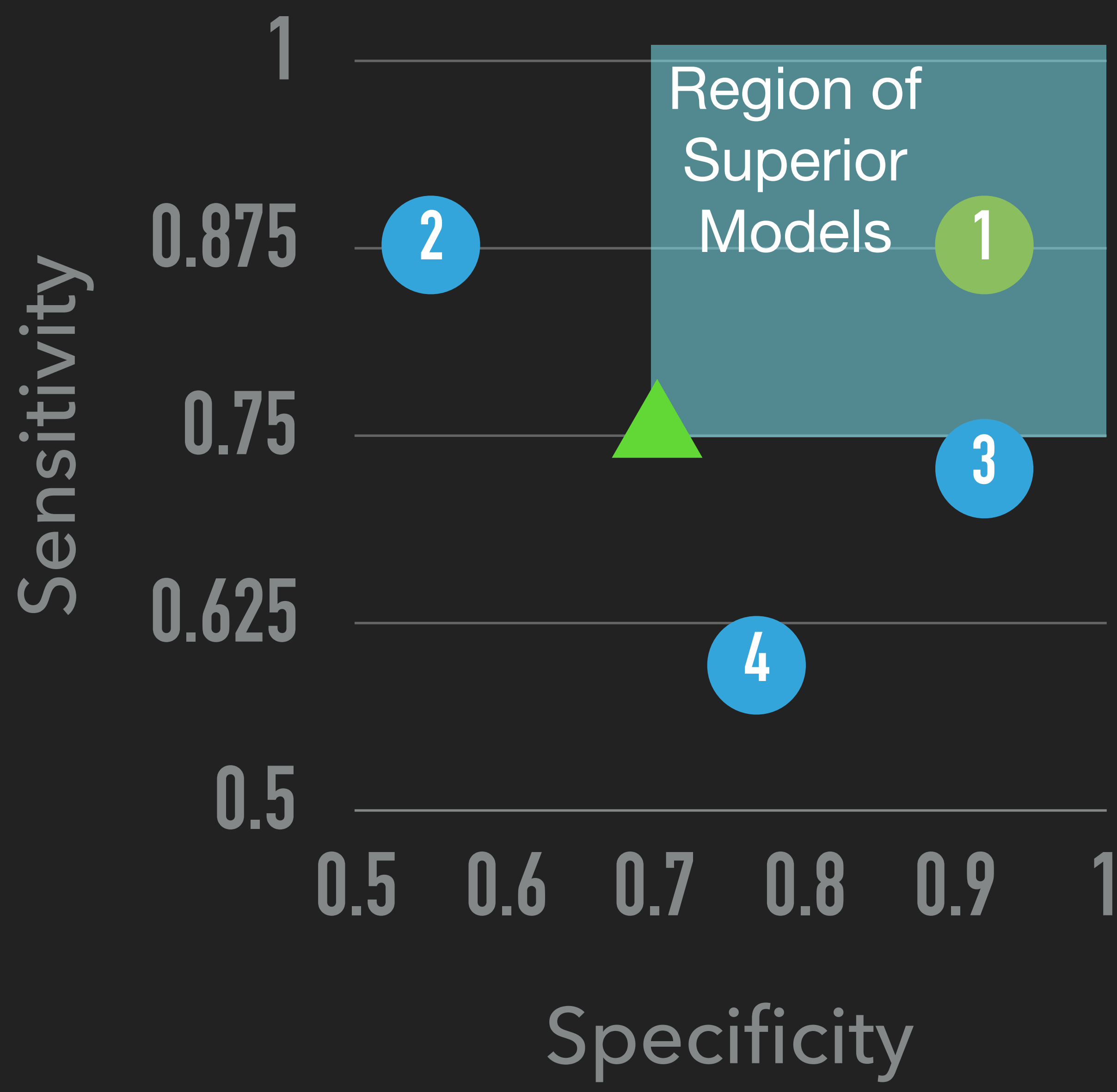
3) Design automatic Algorithm Change Protocols that approve good modifications quickly and control the rate at which bad modifications are approved.

EVALUATION METRICS

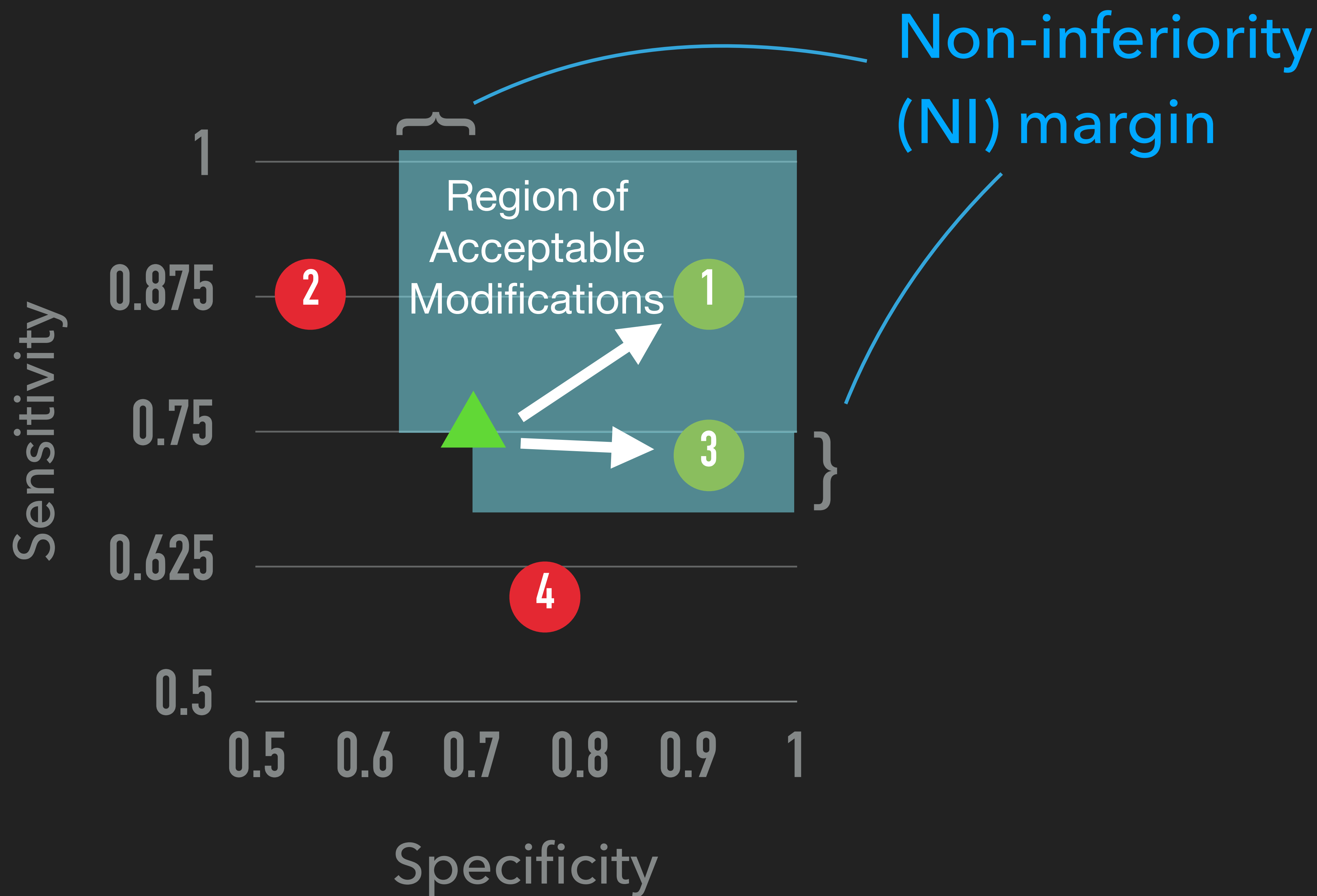
- Evaluate ML-based SaMD according to metrics

$$m_k : \mathscr{F} \mapsto \mathbb{R} \quad k = 1,..,K$$

ACCEPTABLE MODIFICATIONS



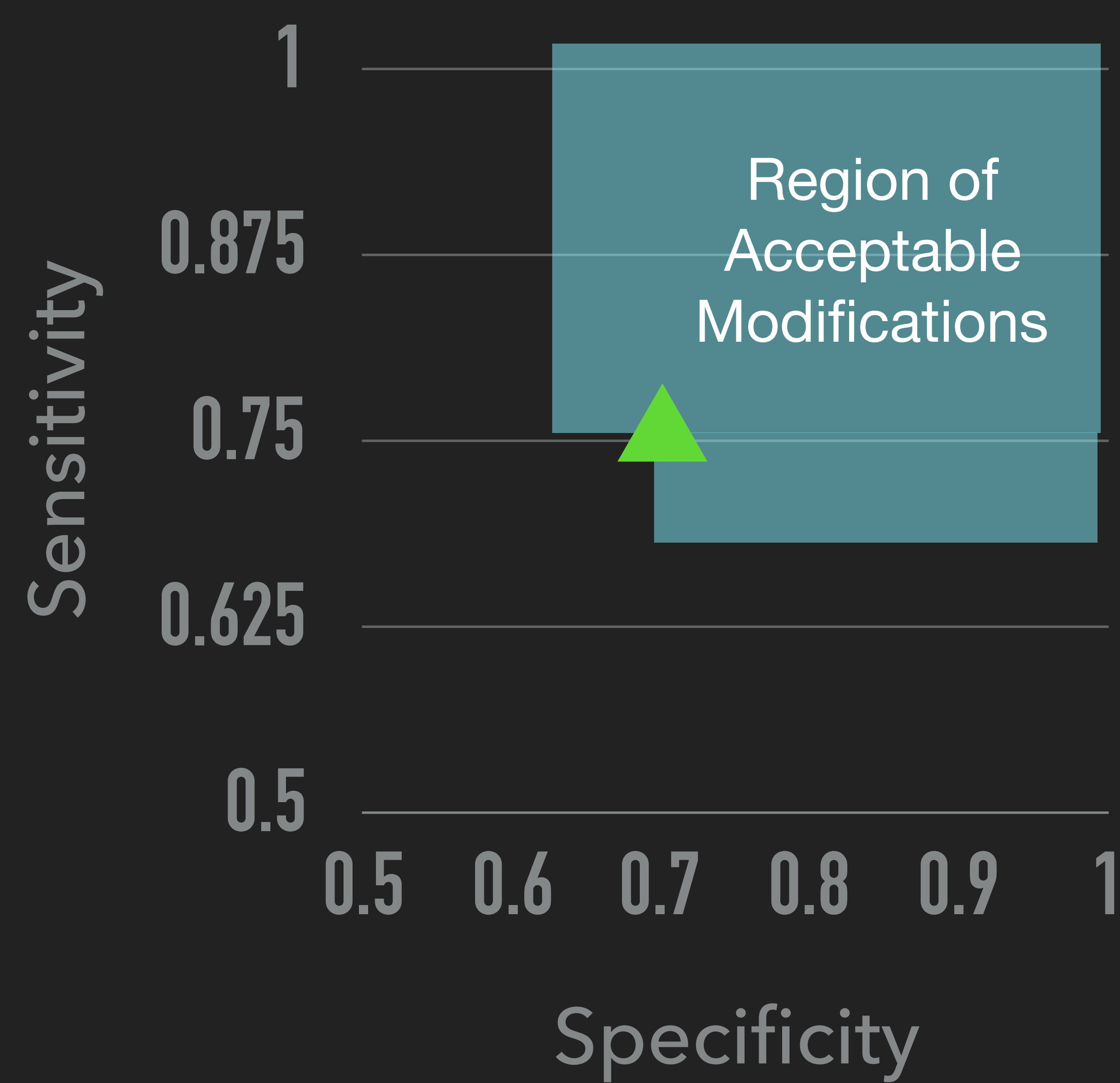
ACCEPTABLE MODIFICATIONS



ACCEPTABLE MODIFICATIONS AND ACCEPTABILITY GRAPHS

Definition: 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
  $m_k(f) - \epsilon \leq m_k(f') \quad \forall k = 1,\ldots,K$
- Superior in at least one metric
  $m_k(f) \leq m_k(f') \quad \exists k \in \{1,\ldots,K\}$



EVALUATING AUTOMATIC ACPS

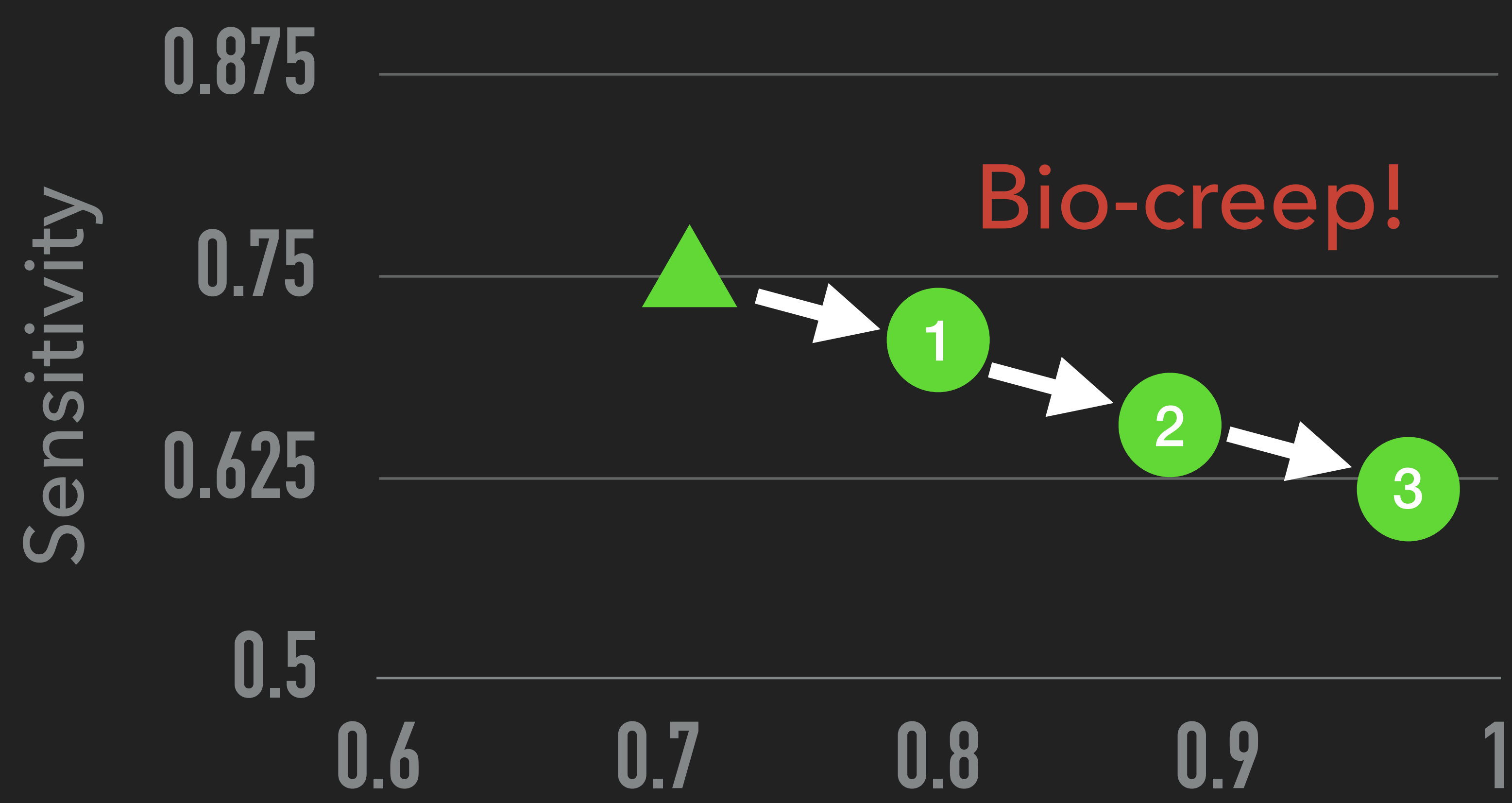
- **Definition:** The expected bad approval count at time T

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

EVALUATING AUTOMATIC ACPS

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

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



EVALUATING AUTOMATIC ACPS

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

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

EVALUATING AUTOMATIC ACPS

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

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

"FWER"

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

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

"FDR"

AUTOMATIC ALGORITHM CHANGE PROTOCOLS

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

aACP-Reset (no error control)

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 $\hat{f}_{\hat{A}_t}$ ($H^0 : \hat{f}_{\hat{A}_t} \not\rightarrow_\epsilon \hat{f}_{t'}$) using prospectively-collected monitoring data.
- Approve the latest modification with p-value smaller than α

aACP-BAC (controls BAC)

At time t = 1,2,…

- For each candidate modification $\hat{f}_{t'}$, test the null hypotheses following a **gate-keeping** procedure at alpha levels chosen using **group-sequential** and **alpha-spending** procedures:
  - $H_1^0 : \hat{f}_{\hat{A}_1} \not\rightarrow_\epsilon \hat{f}_{t'}$
  - $H_2^0 : \hat{f}_{\hat{A}_2} \not\rightarrow_\epsilon \hat{f}_{t'}$
  - …
  - $H_t^0 : \hat{f}_{\hat{A}_t} \not\rightarrow_\epsilon \hat{f}_{t'}$

  Gate-keeping
- Approve the latest modification that rejects all hypotheses

AUTOMATIC ALGORITHM CHANGE PROTOCOLS

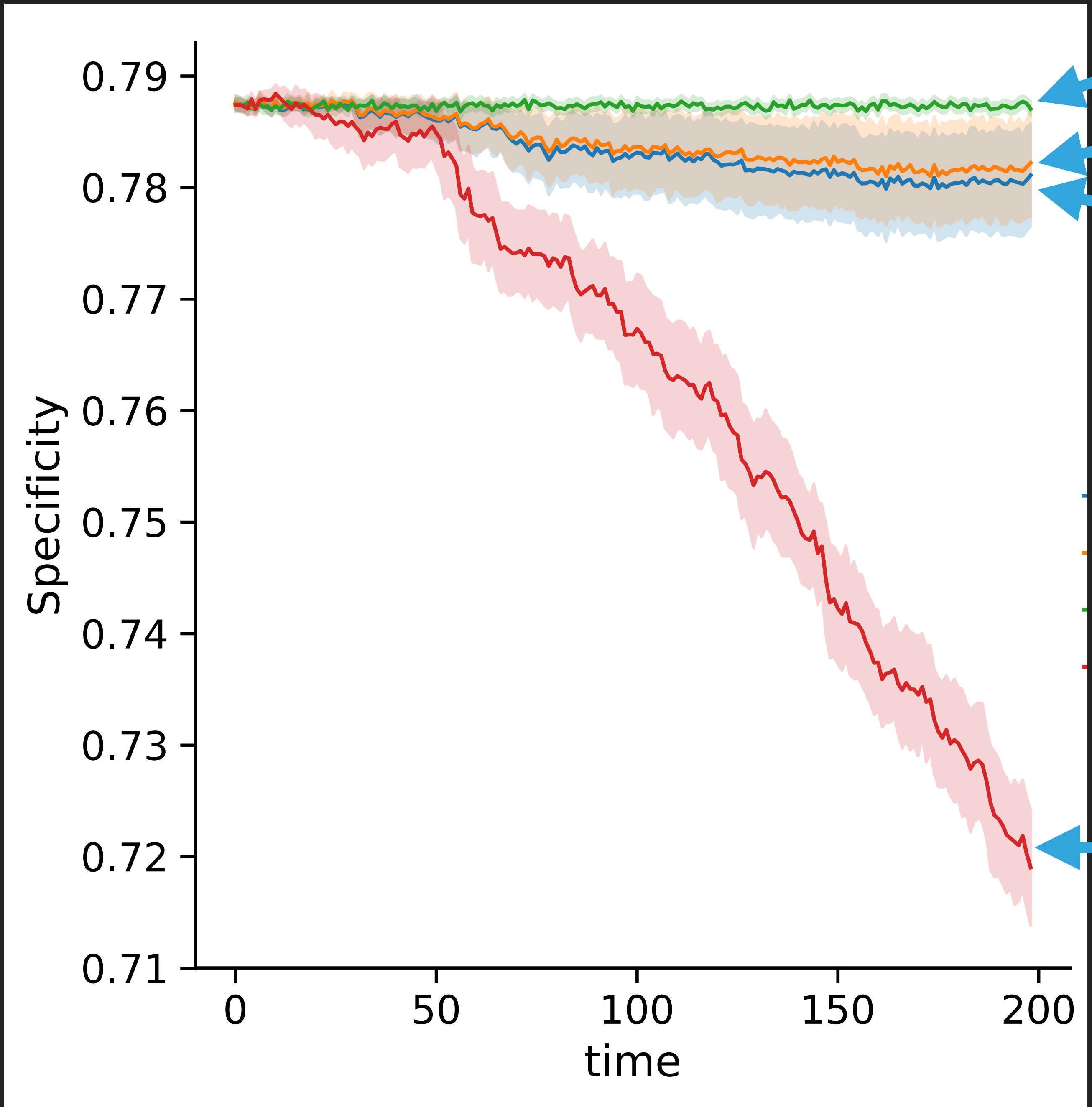
- Without error rate control:
  - **aACP-Blind**: Approve everything
  - **aACP-Reset**: Compare to the latest approval with fixed p-value threshold
- With error rate control:
  - **aACP-BAC**: Controls expected Bad Approval Count using alpha-spending, group-sequential, and gate-keeping methods
  - **aACP-BABR**: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

SIMULATIONS

- Setup
  - Monitoring data is IID at each time point and across time points
  - Binary prediction problem
- Desired properties
  1. Low rate of bad approvals
  2. High rate of good approvals

RESULT: AACP-BAC AND -BABR PROTECT AGAINST BIO-CREEP

- Proposed modifications deteriorate over time



aACP-Fixed

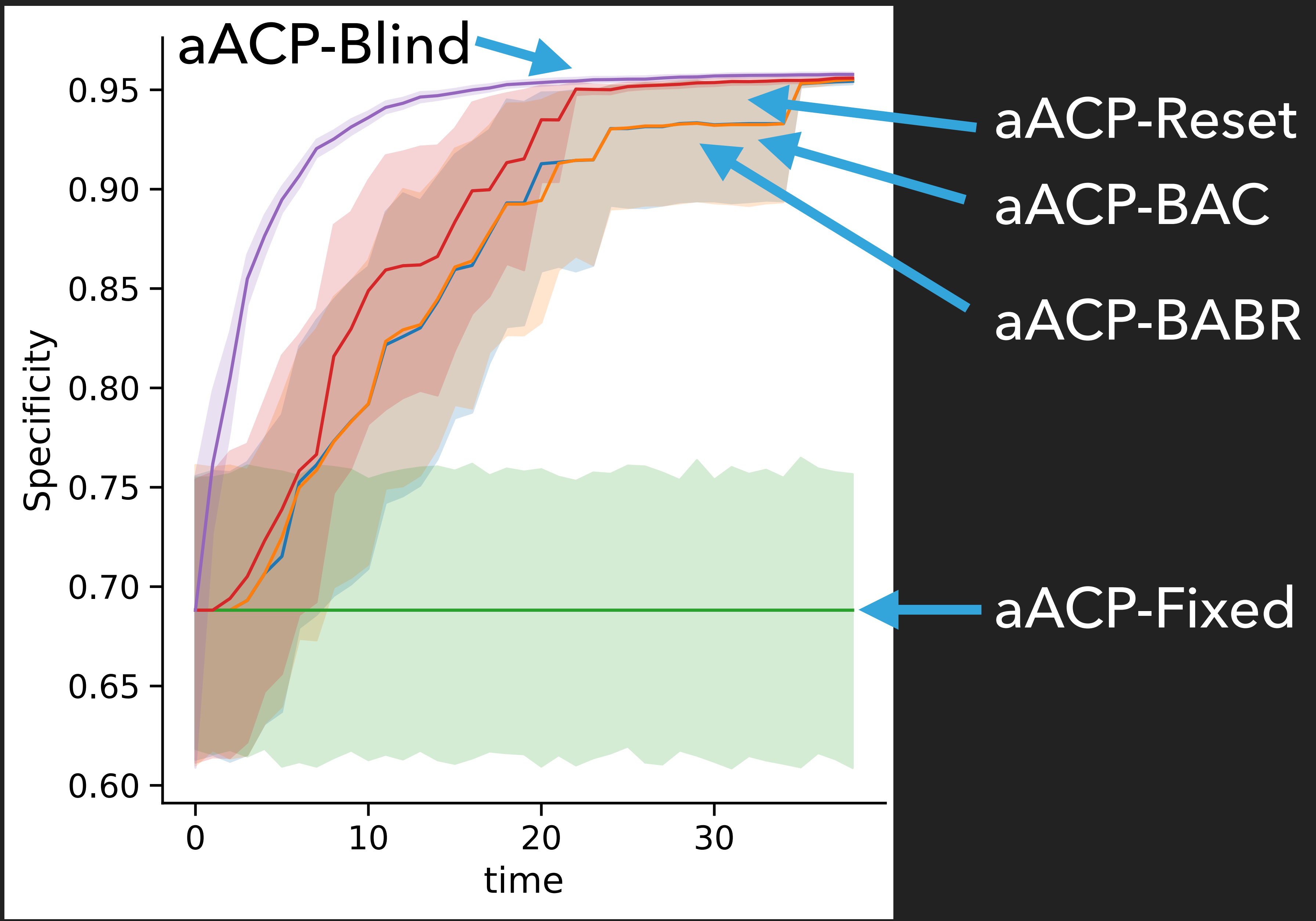
aACP-BABR

aACP-BAC

aACP-Reset

RESULT: MODELS IMPROVE AT SIMILAR RATES USING AACP-BAC, AACP-BABR, AND AACP-RESET

- Train new models using the accumulating monitoring data



THANKS!

Jean Feng, Scott Emerson, Noah Simon

https://arxiv.org/abs/1912.12413