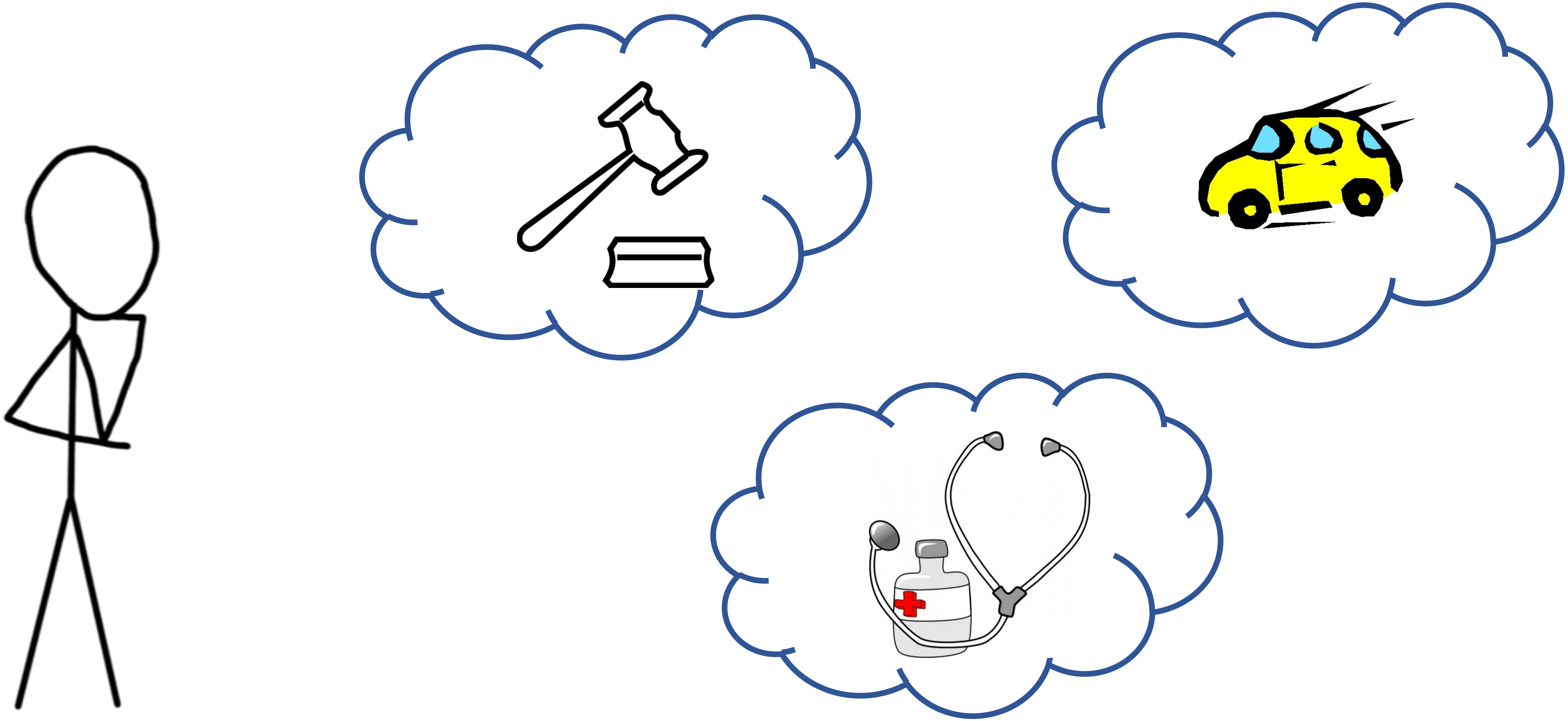
Human-in-the-Loop Interpretability Prior

Isaac Lage[1], Andrew Slavin Ross[1], Been Kim[2], Samuel J. Gershman[1] & Finale Doshi-Velez[1]

[1]Harvard University & [2]Google Brain

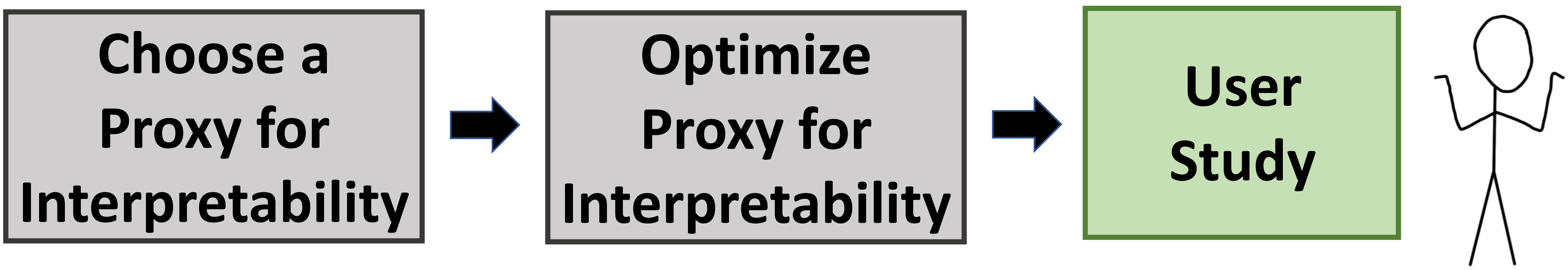
Poster: Today, 10:45 AM - 12:45 PM, Room 210 & 230 AB **#119**

Interpretability



Optimizing for Interpretability

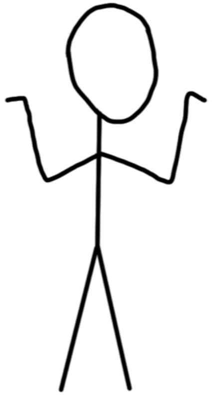
Previous Work



Optimizing for Interpretability

Previous Work

Choose a Proxy for Interpretability → **Optimize Proxy for Interpretability** → **User Study**

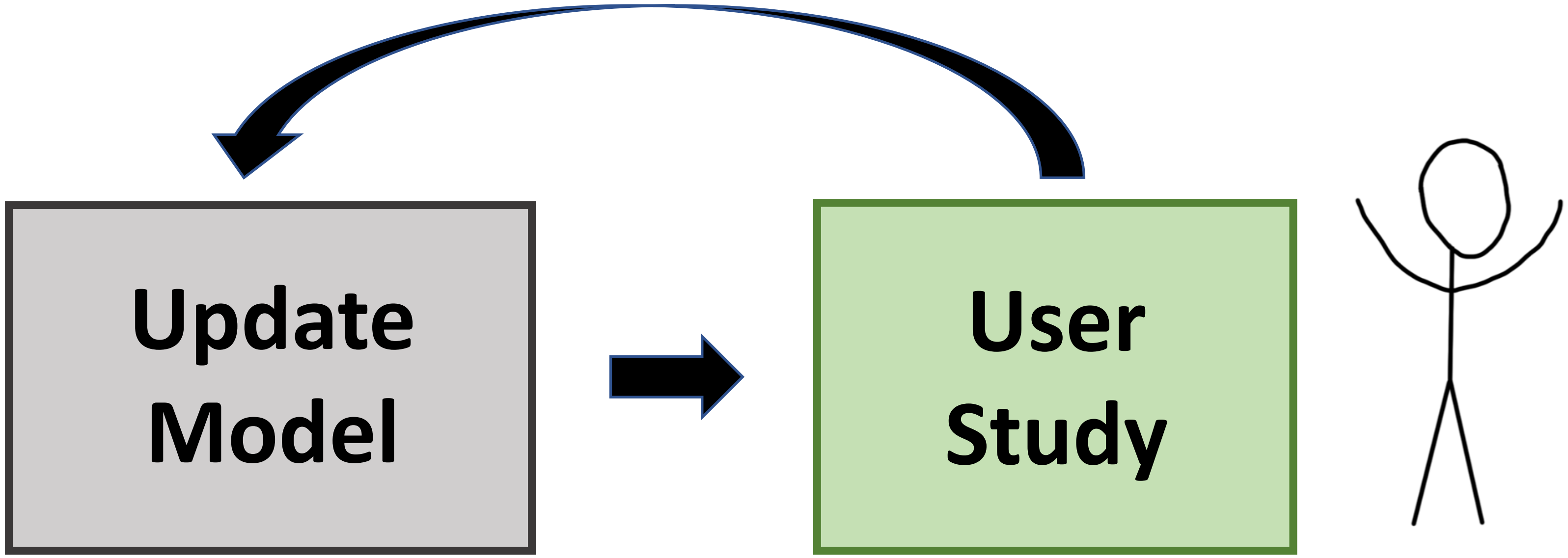


Which proxy?

How to use results to choose a better proxy?

Optimizing for Interpretability

Human-in-the-Loop Interpretability



No proxy!

Update model directly with results!

Interpretability Prior

Goal: Bias model to be **human interpretable**

$$\max_{M \in \mathcal{M}} p(X|M)p(M)$$

Bayesian Inference

Interpretability Prior

First: Formulate **Interpretability Encouraging Prior**

$$\max_{M \in \mathcal{M}} p(X|M)p(M)$$

Optimizing for Interpretability

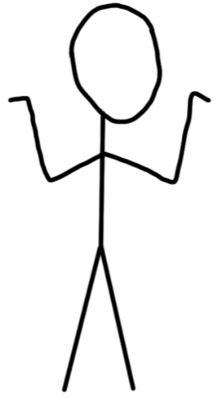
Can define a **prior**

Previous Work

Choose a Proxy for Interpretability

Optimize Proxy for Interpretability

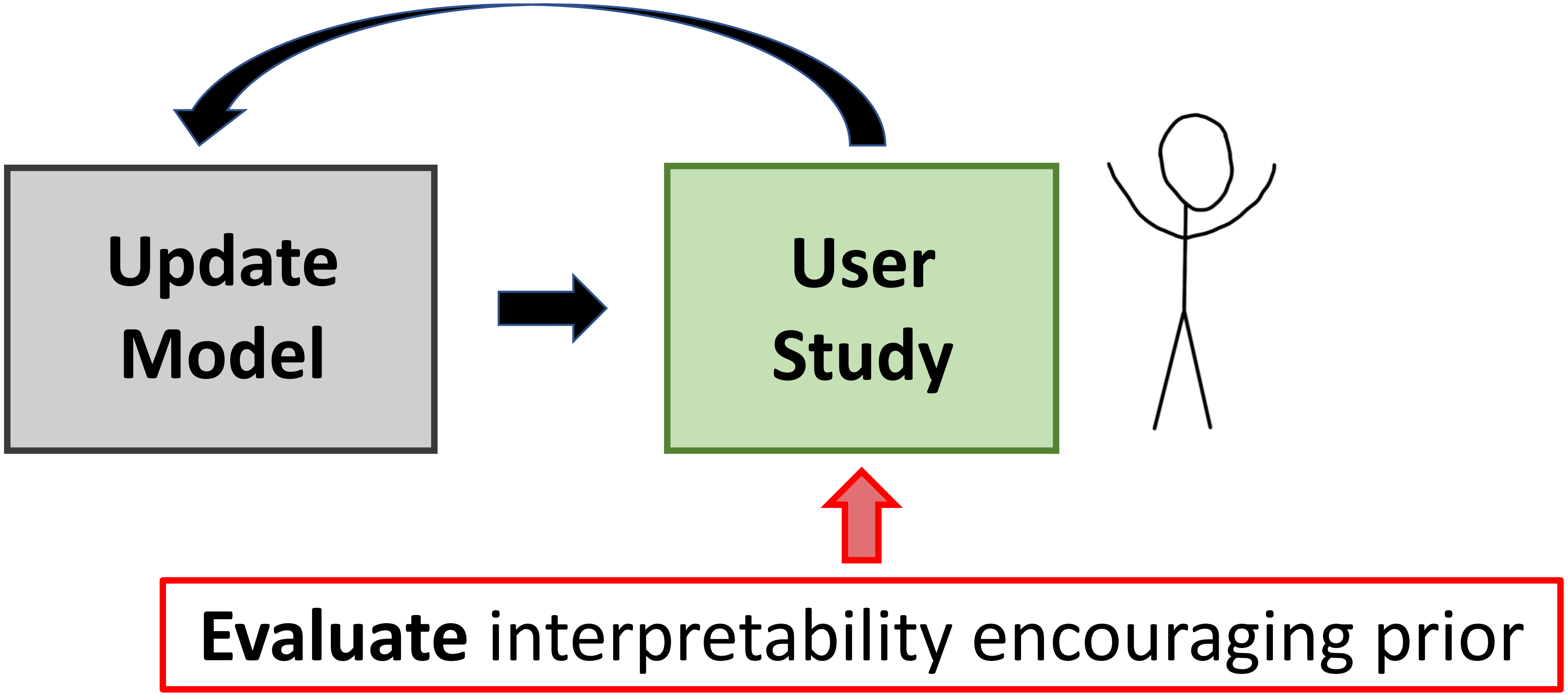
User Study



Which prior captures human interpretability?

Optimizing for Interpretability

Human-in-the-Loop Interpretability



Interpretability Prior

First: Formulate **Interpretability Encouraging Prior**

$$\max_{M \in \mathcal{M}} p(X|M)p(M)$$

Then: Identify **MAP** Solution

Interpretability Prior

$$\max_{M \in \mathcal{M}} p(X|M)p(M)$$

Likelihood: Easy

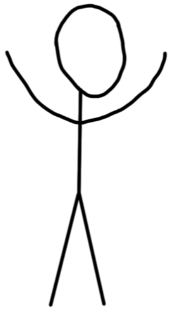
Evaluate computationally

No users!

Interpretability Prior

Prior: Hard

No closed form
Evaluate with
user studies!



$$\max_{M \in \mathcal{M}} p(X|M)p(M)$$

Likelihood: Easy

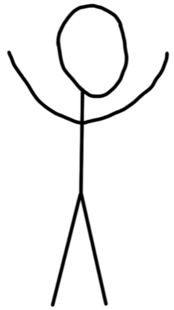
Evaluate computationally
No users!

Interpretability Prior

Prior: Hard

No closed form
Evaluate with
user studies!

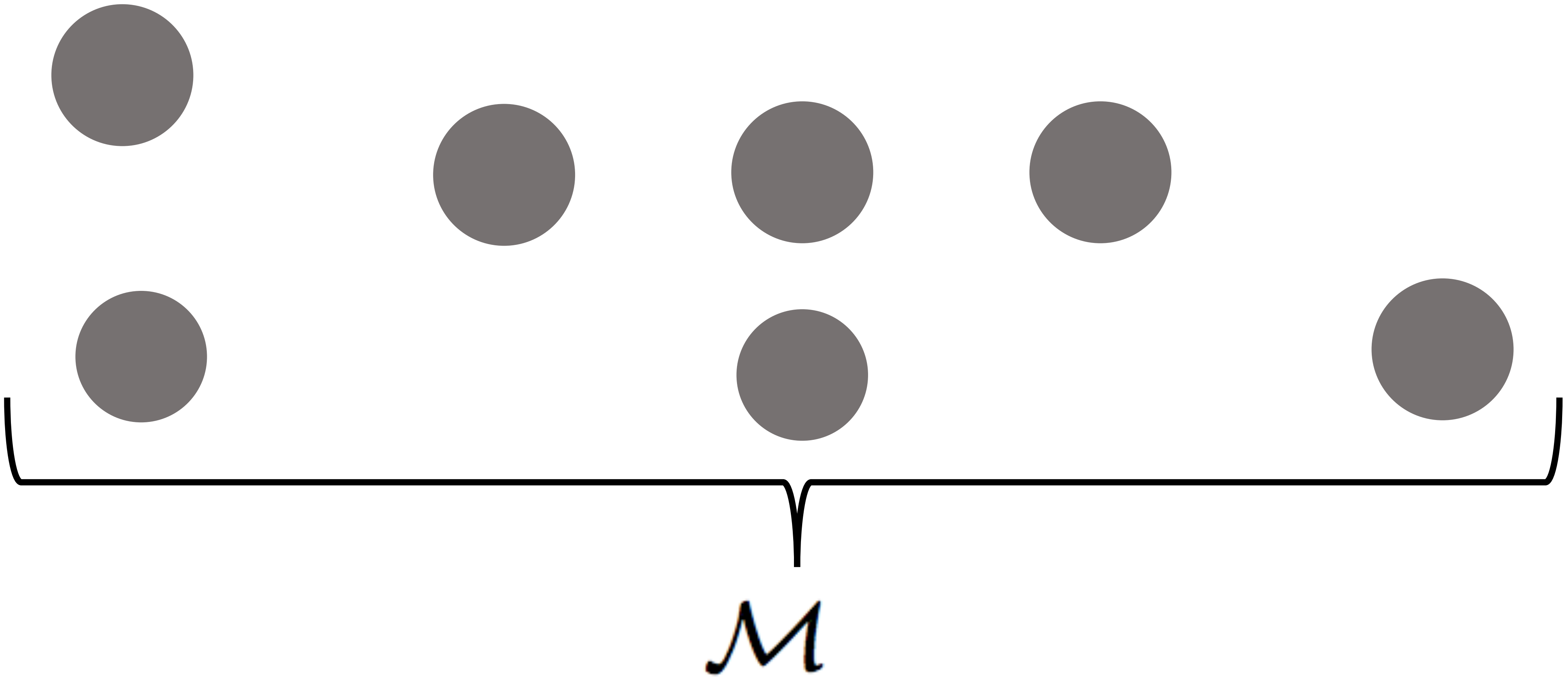
$$\max_{M \in \mathcal{M}} p(X|M)p(M)$$



Challenge: Approximate MAP with **few evaluations of prior**

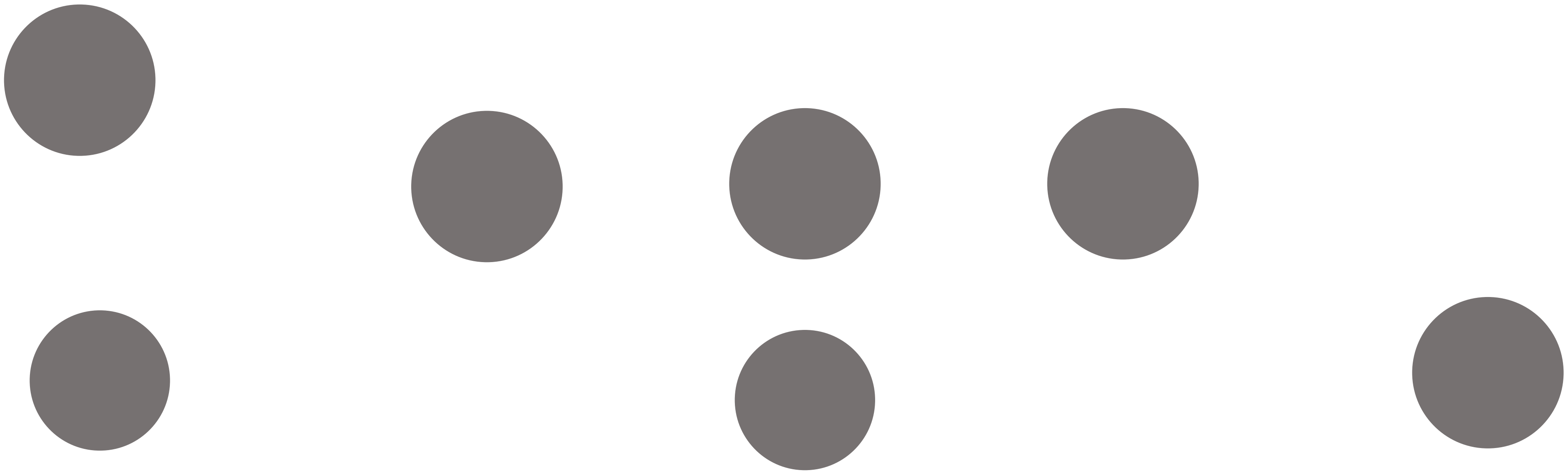
Simplified Cartoon of Our Approach

Step 1: Identify Diverse, High Likelihood Models



Simplified Cartoon of Our Approach

Step 1: Identify Diverse, High Likelihood Models



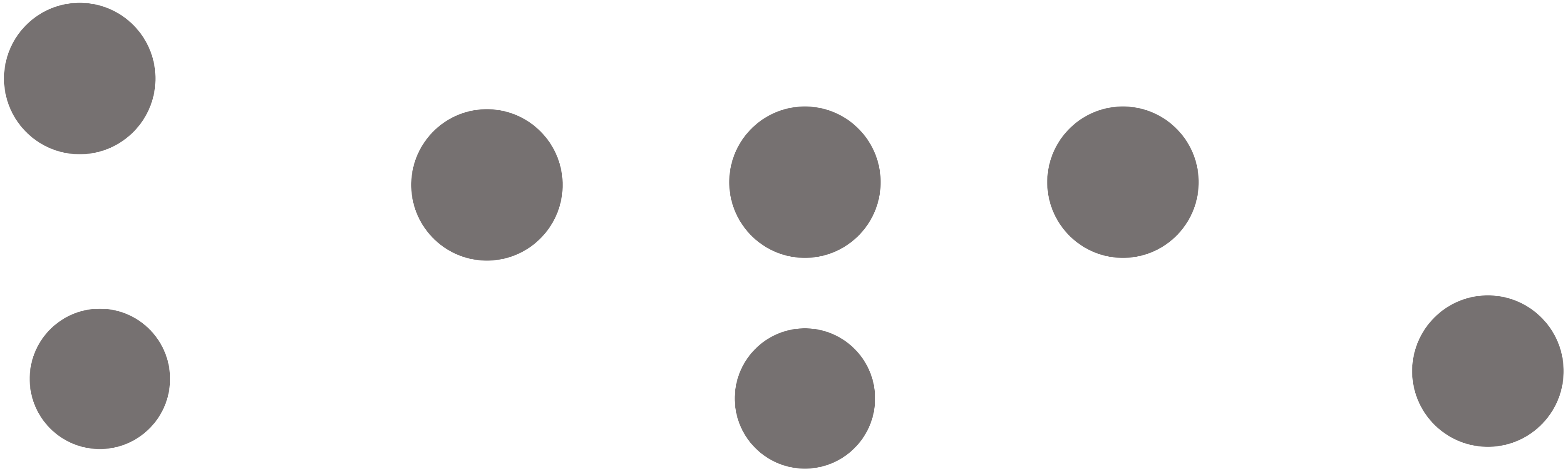
Candidate MAP 1:
Likelihood = **HIGH**

Candidate MAP 2:
Likelihood = **HIGH**

Candidate MAP 3:
Likelihood = **HIGH**

Simplified Cartoon of Our Approach

Step 1: Identify Diverse, High Likelihood Models



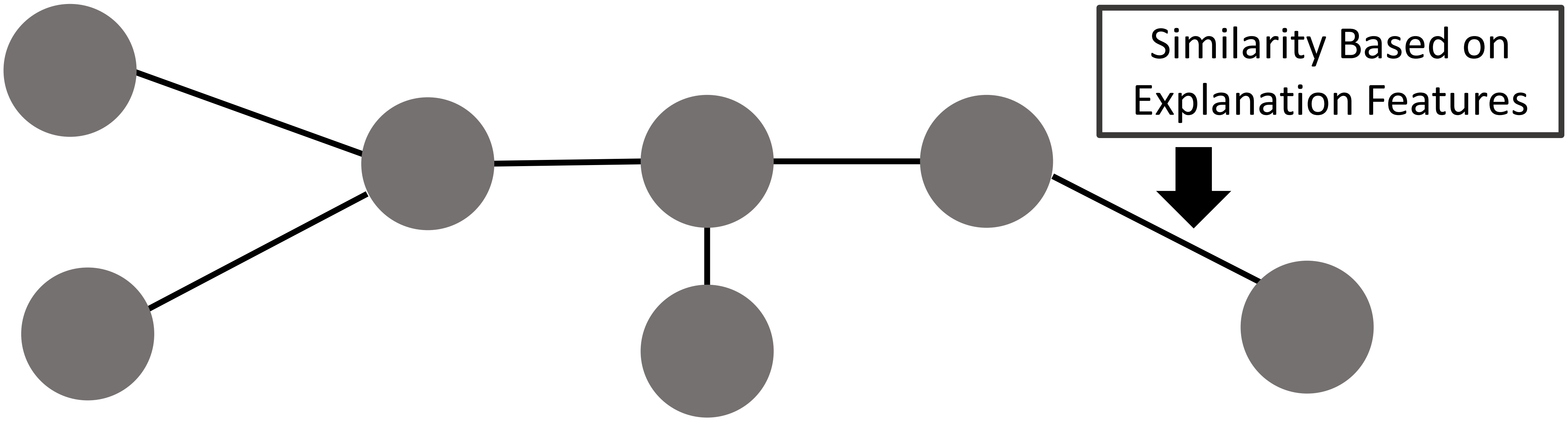
Candidate MAP 1:
Likelihood = **HIGH**
Prior = **?**

Candidate MAP 2:
Likelihood = **HIGH**
Prior = **?**

Candidate MAP 3:
Likelihood = **HIGH**
Prior = **?**

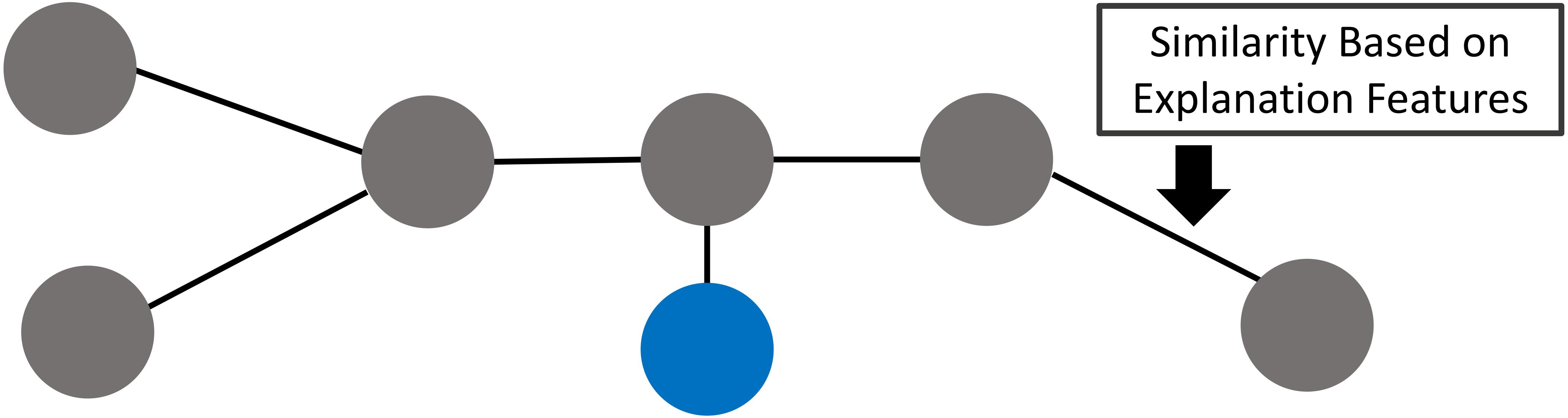
Simplified Cartoon of Our Approach

Step 2: Bayesian Optimization with User Studies



Simplified Cartoon of Our Approach

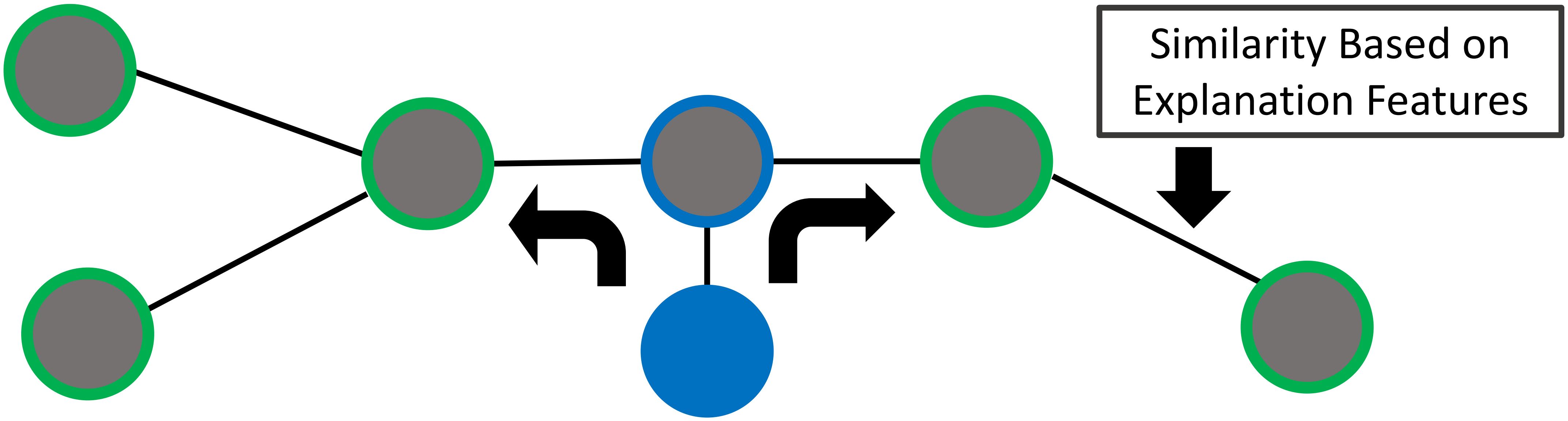
Step 2: Bayesian Optimization with User Studies



User study 1:
Prior = MEDIUM

Simplified Cartoon of Our Approach

Step 2: Bayesian Optimization with User Studies

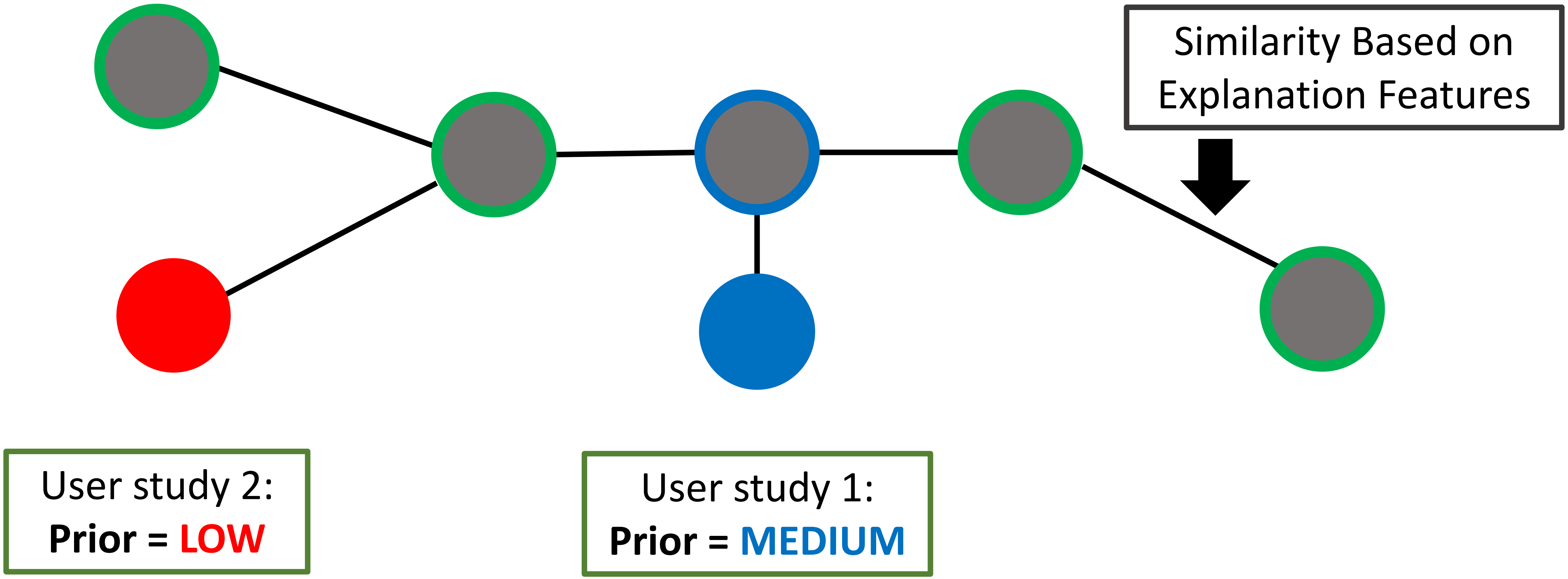


Prior Estimate:
Prior = HIGH?

User study 1:
Prior = MEDIUM

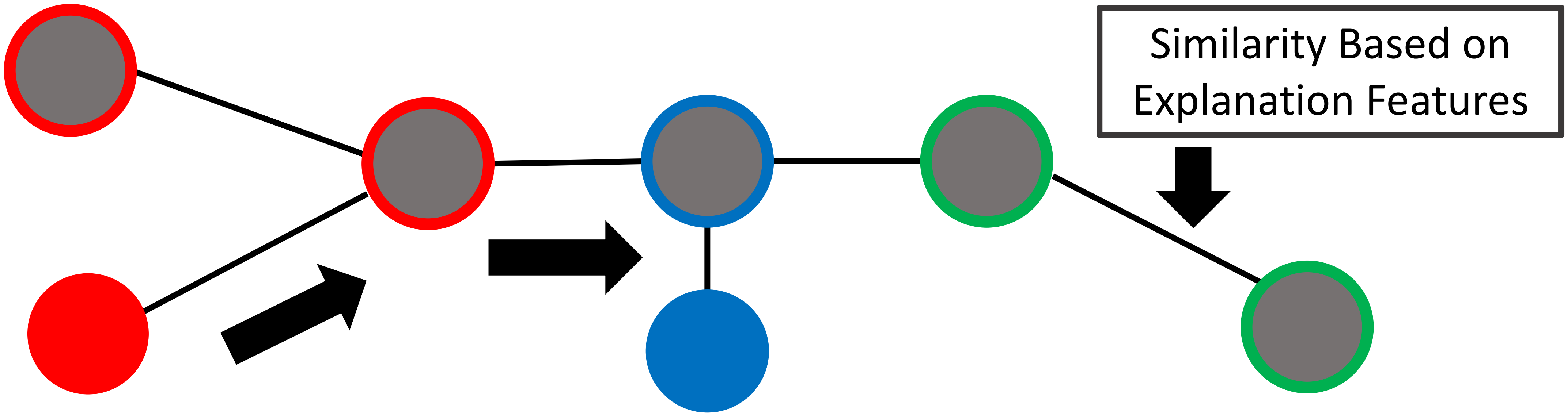
Simplified Cartoon of Our Approach

Step 2: Bayesian Optimization with User Studies



Simplified Cartoon of Our Approach

Step 2: Bayesian Optimization with User Studies



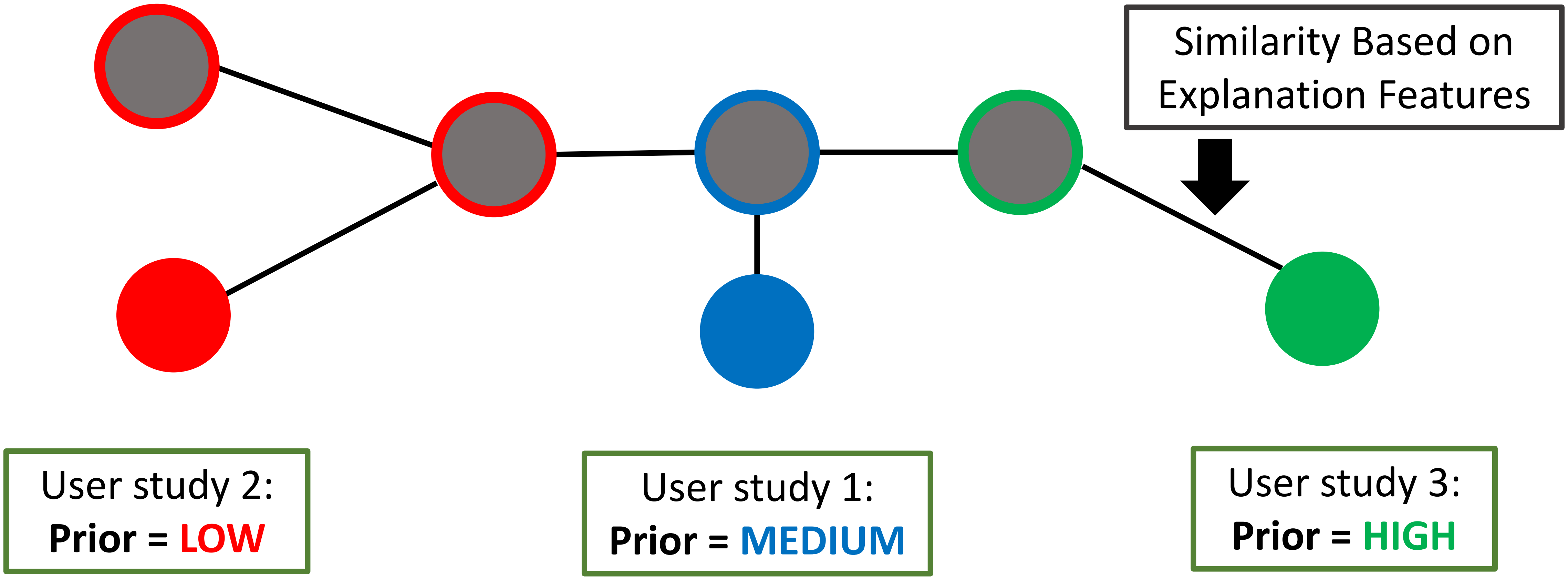
User study 2:
Prior = LOW

User study 1:
Prior = MEDIUM

Prior Estimate:
Prior = HIGH?

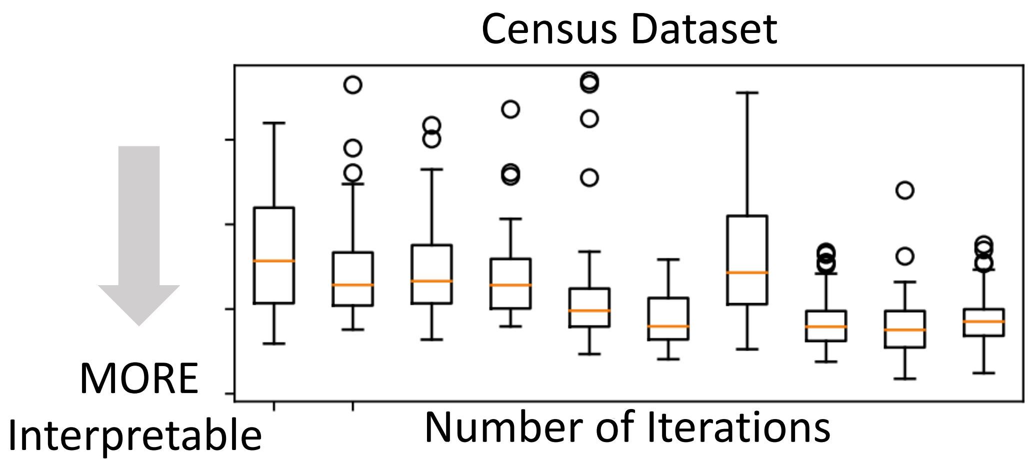
Simplified Cartoon of Our Approach

Step 2: Bayesian Optimization with User Studies



Main Takeaways

- We optimize for interpretability directly with human feedback
- Our approach efficiently identifies human-interpretable and predictive models
- MAP approximations correspond to different interpretability proxies on different datasets



Poster: Today, 10:45 AM - 12:45 PM, Room 210 & 230 AB **#119**