Statistical Methods for the Forensic Analysis of User-Event Data

Christopher Galbraith

PhD Defense
5.28.20

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

csafe
Center for Statistics and Applications in Forensic Evidence

The material presented here is based upon work supported by the National Institute of Science and Technology under Award No. 70NANB15H176. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Institute of Science and Technology, nor of the Center for Statistics and Applications in Forensic Evidence.

Publications

i **C. Galbraith,** P. Smyth, H. S. Stern. ***Statistical methods for the forensic analysis of geolocated event data.*** Digital Investigation 2020.

ii **C. Galbraith,** P. Smyth, H. S. Stern. ***Quantifying the association between discrete event time series with applications to digital forensics.*** J R Stat Soc Series A 2020.

iii **C. Galbraith,** P. Smyth. ***Analyzing user-event data using score-based likelihood ratios with marked point processes.*** Digital Investigation 2017.

Publications

i **C. Galbraith,** P. Smyth, H. S. Stern. ***Statistical methods for the forensic analysis of geolocated event data.*** Digital Investigation 2020.

ii **C. Galbraith,** P. Smyth, H. S. Stern. ***Quantifying the association between discrete event time series with applications to digital forensics.*** J R Stat Soc Series A 2020.

iii **C. Galbraith,** P. Smyth. ***Analyzing user-event data using score-based likelihood ratios with marked point processes.*** Digital Investigation 2017.

Dissertation

1 **Introduction**

2 **Computing Strength of Evidence with the Likelihood Ratio**

3 **Score-based Approaches for Computing Strength of Evidence** ii

4 **Spatial Event Data** i

5 **Temporal Event Data** ii iii

6 **Discussion on Future Directions**

Publications

i **C. Galbraith,** P. Smyth, H. S. Stern. ***Statistical methods for the forensic analysis of geolocated event data.*** Digital Investigation 2020.

ii **C. Galbraith,** P. Smyth, H. S. Stern. ***Quantifying the association between discrete event time series with applications to digital forensics.*** J R Stat Soc Series A 2020.

iii **C. Galbraith,** P. Smyth. ***Analyzing user-event data using score-based likelihood ratios with marked point processes.*** Digital Investigation 2017.

Dissertation

1. **Introduction**
2. **Computing Strength of Evidence with the Likelihood Ratio**
3. **Score-based Approaches for Computing Strength of Evidence** ii
4. **Spatial Event Data** i
5. **Temporal Event Data** ii iii
6. **Discussion on Future Directions**

Outline

1. Motivation
2. Quantifying Strength of Evidence
3. Empirical Evaluation Techniques
4. Application to Geolocated Event Data
5. Future Directions and Conclusions

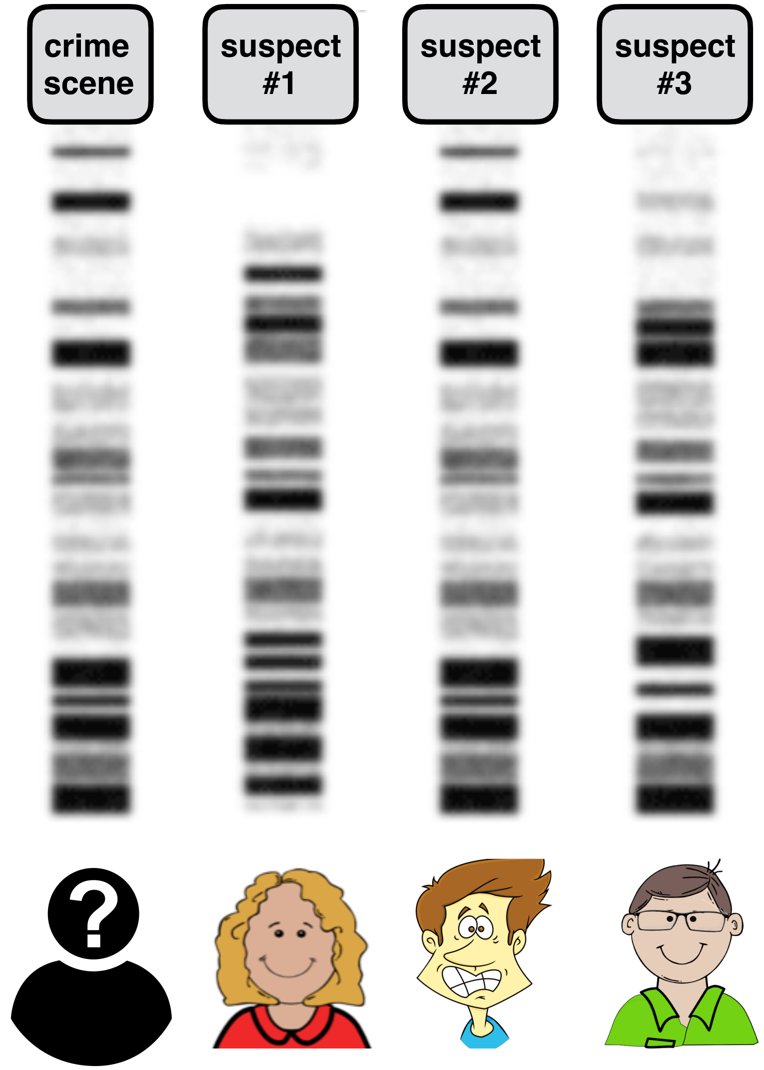
Motivation


POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS

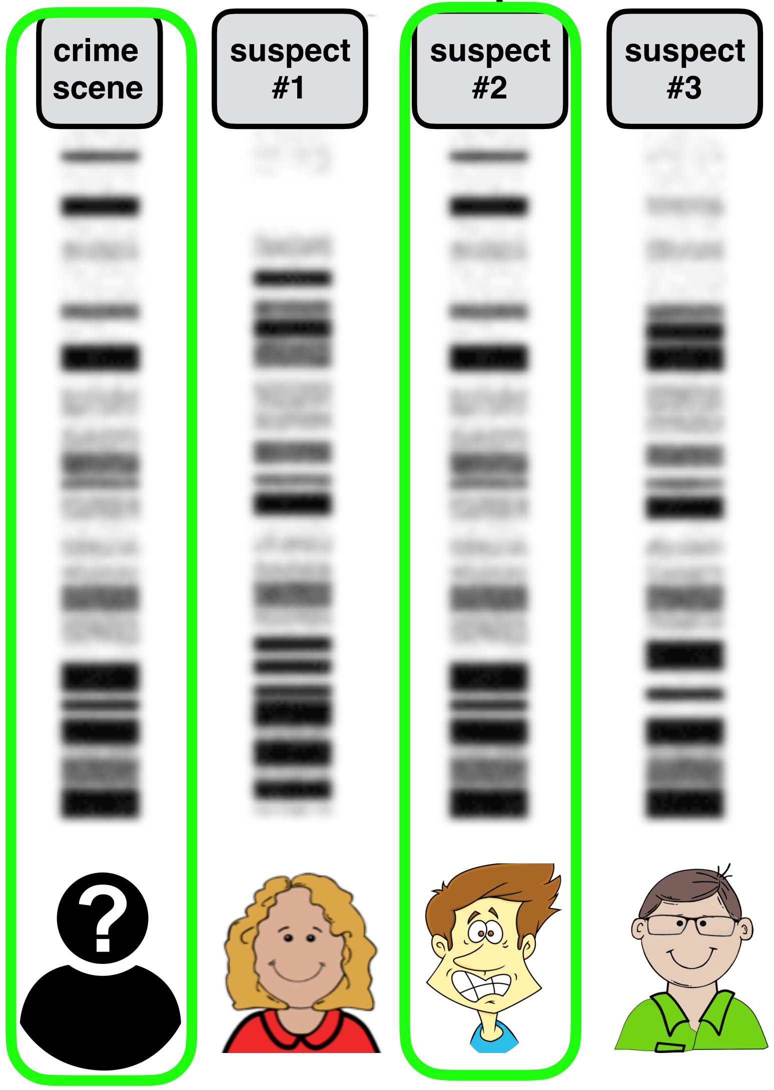

POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS

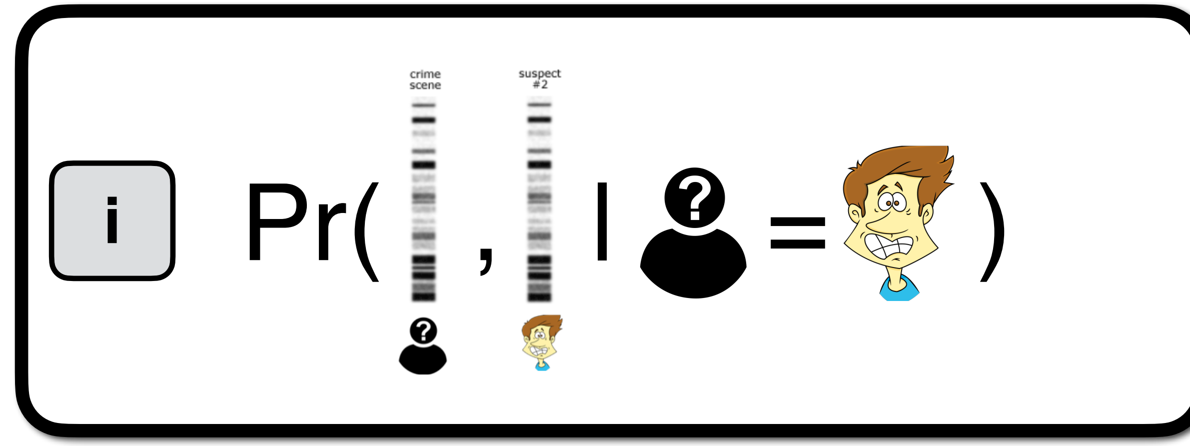

POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS

DNA Samples



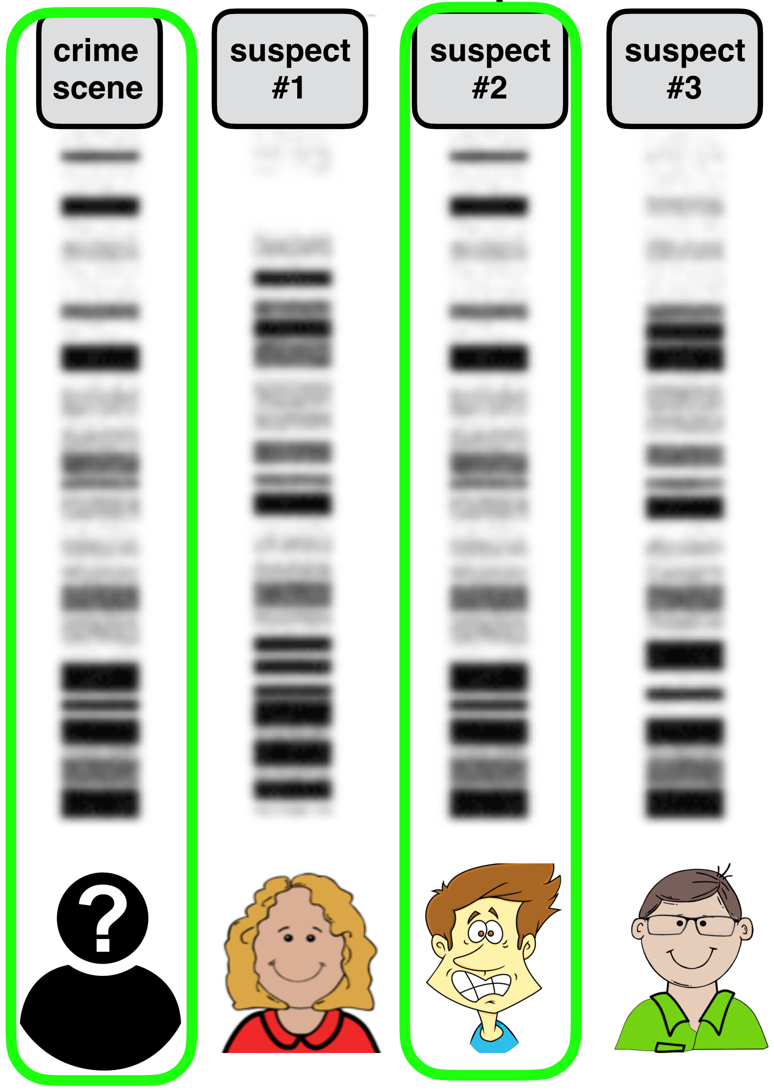
DNA Samples

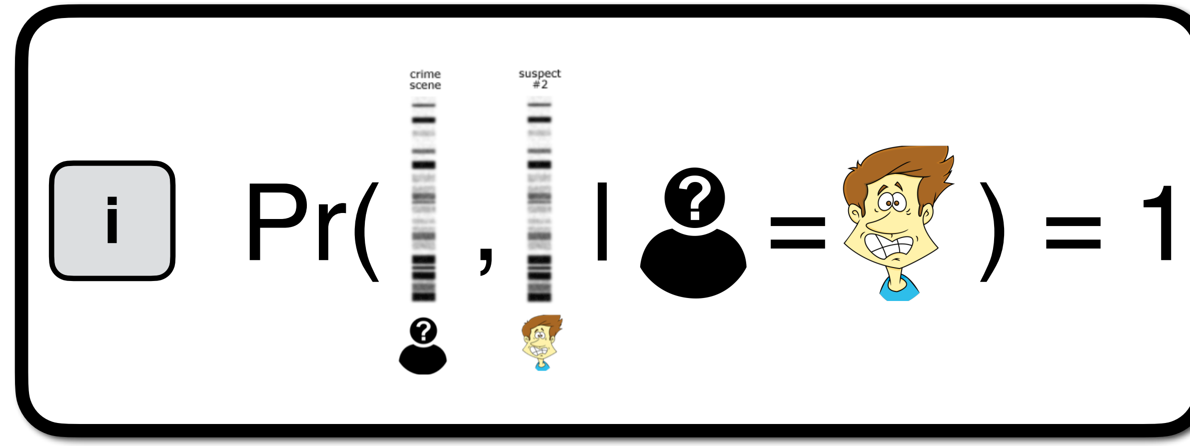




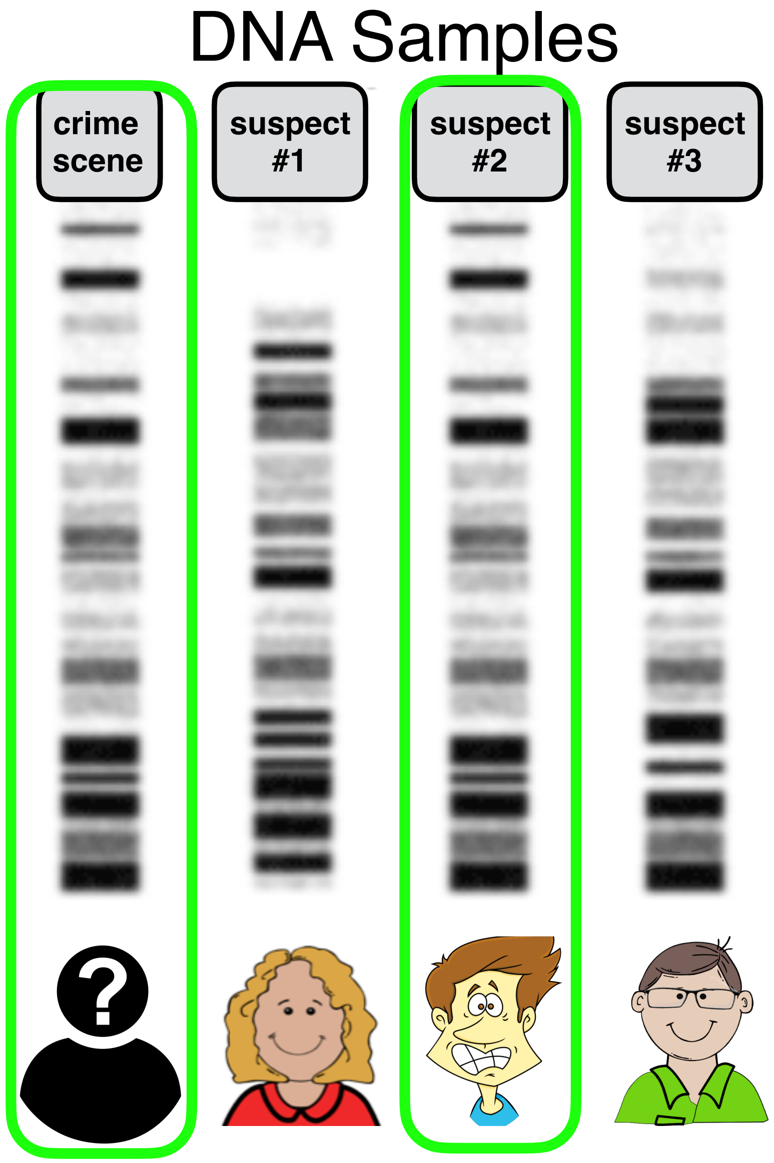
[Steele & Balding, 2014]

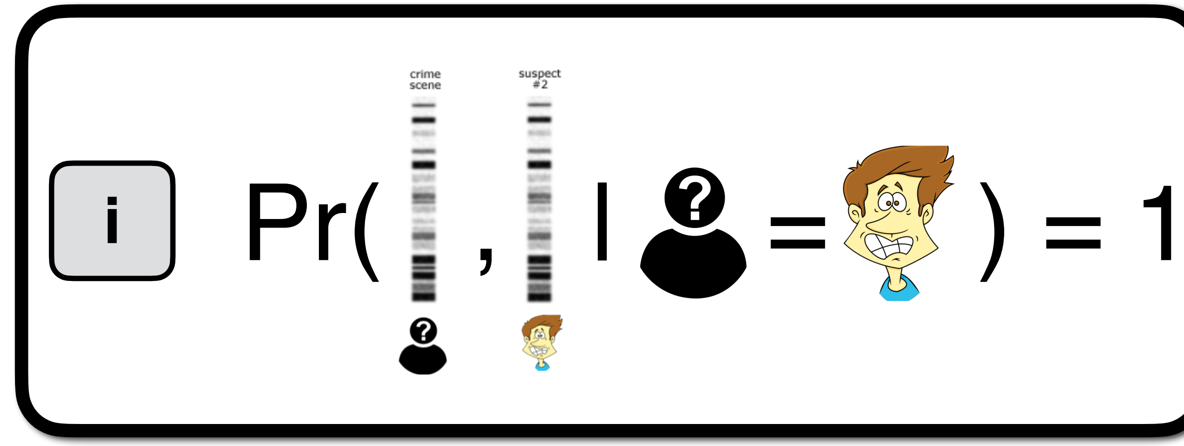
DNA Samples

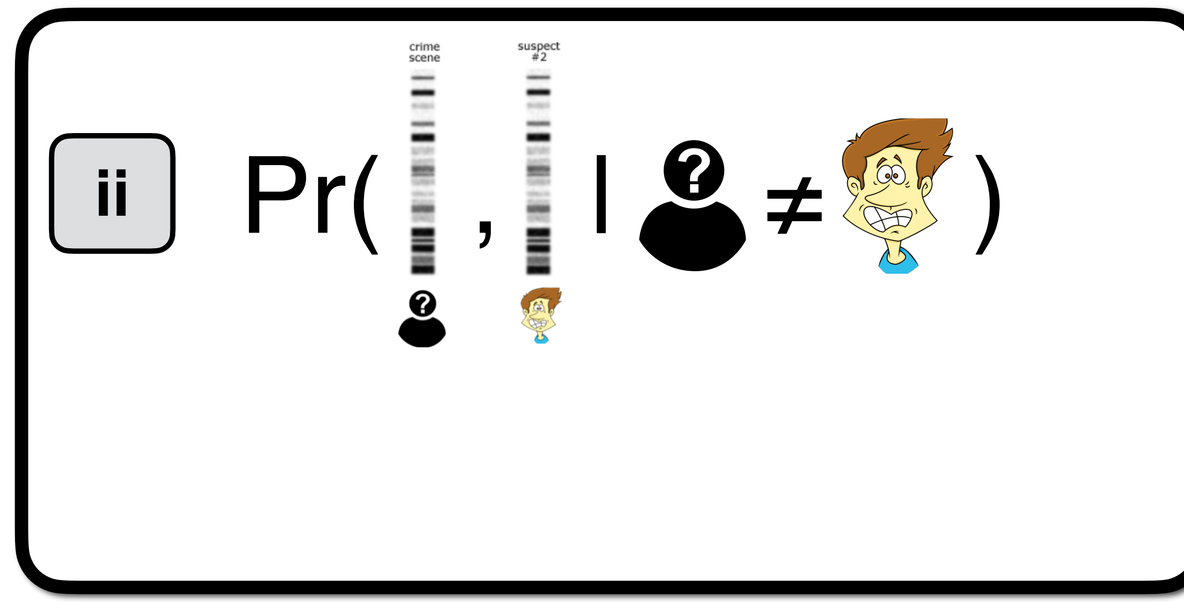




[Steele & Balding, 2014]

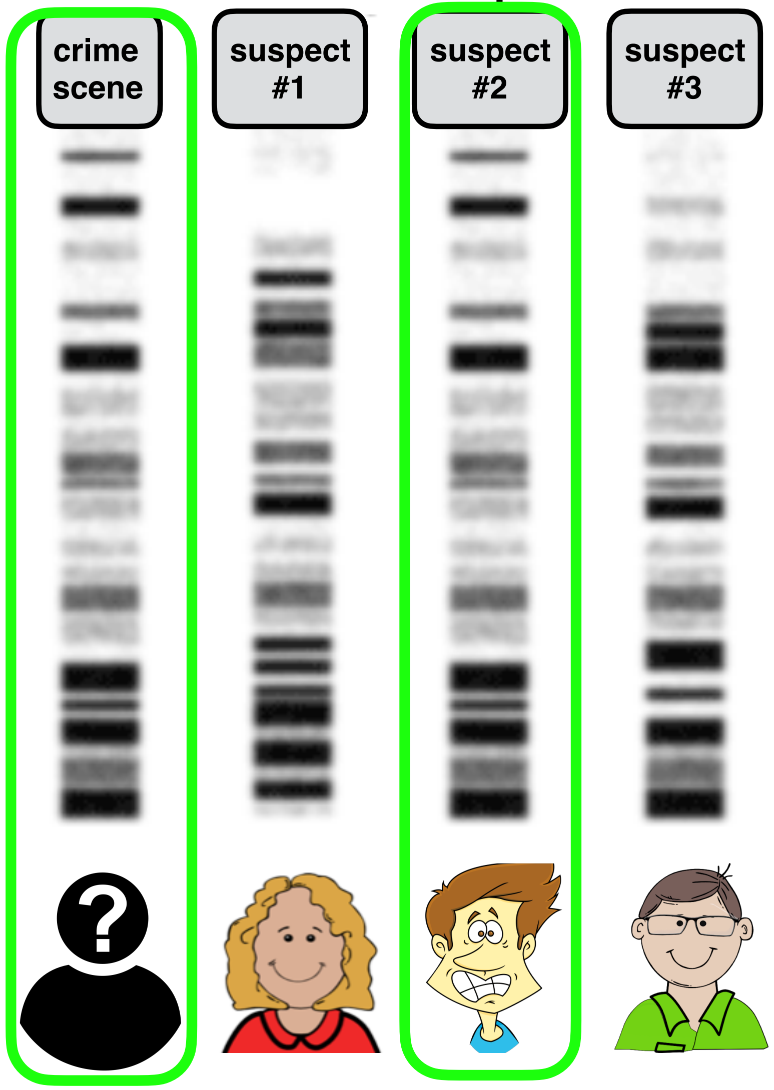


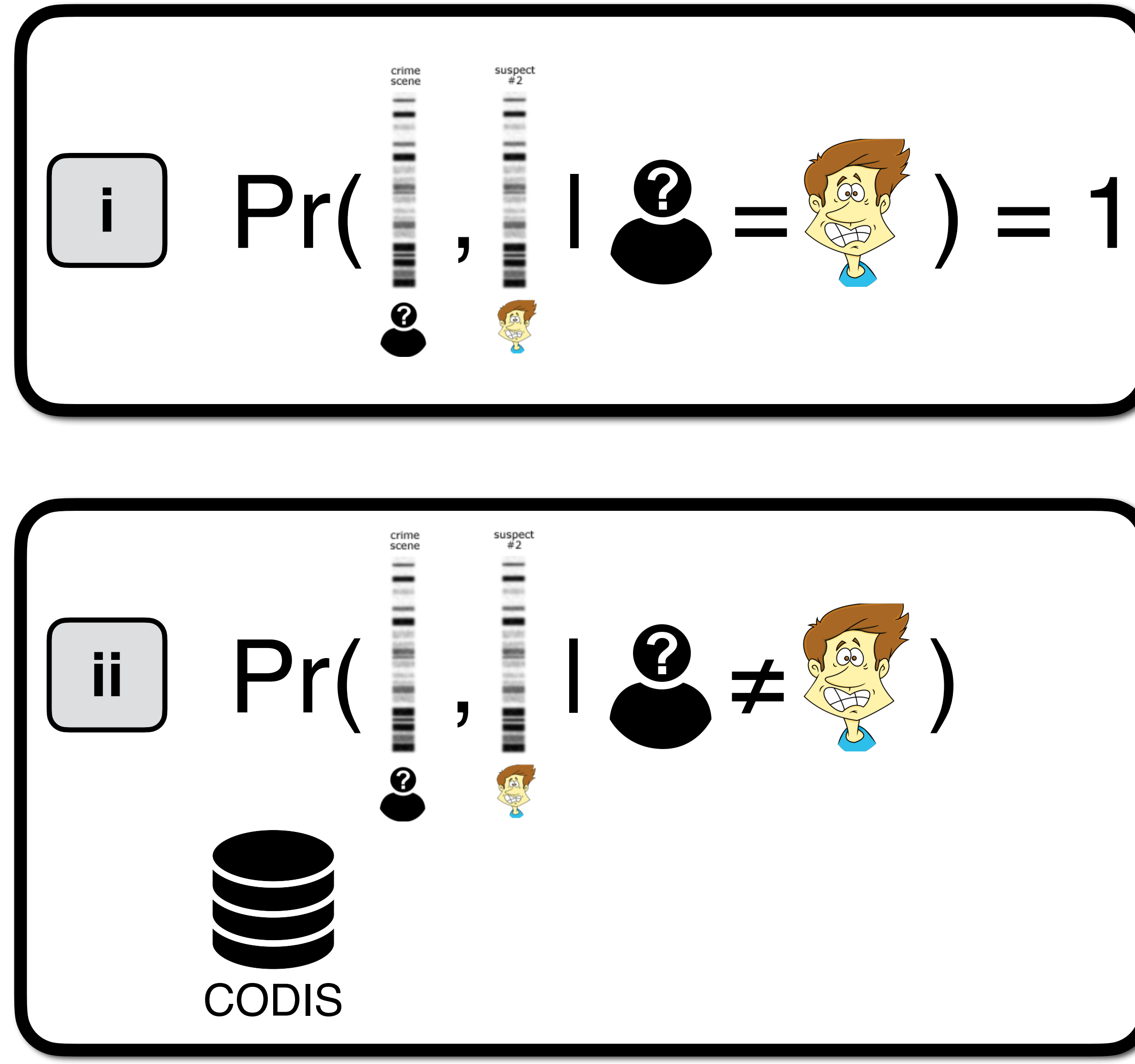




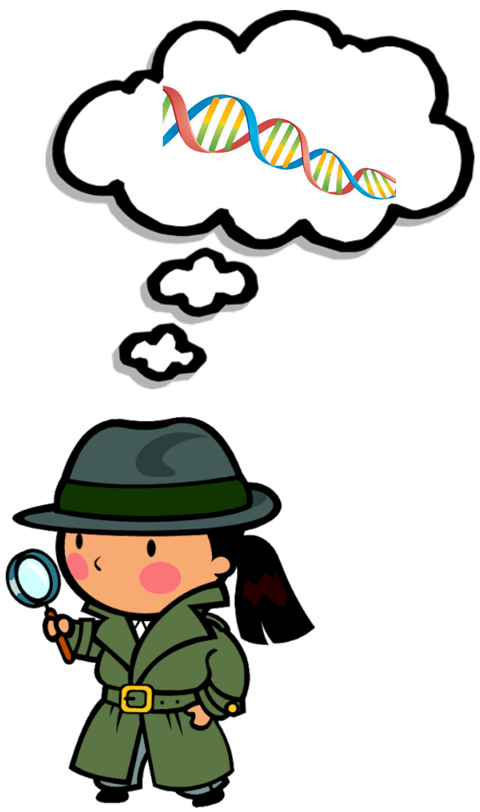
[Steele & Balding, 2014]

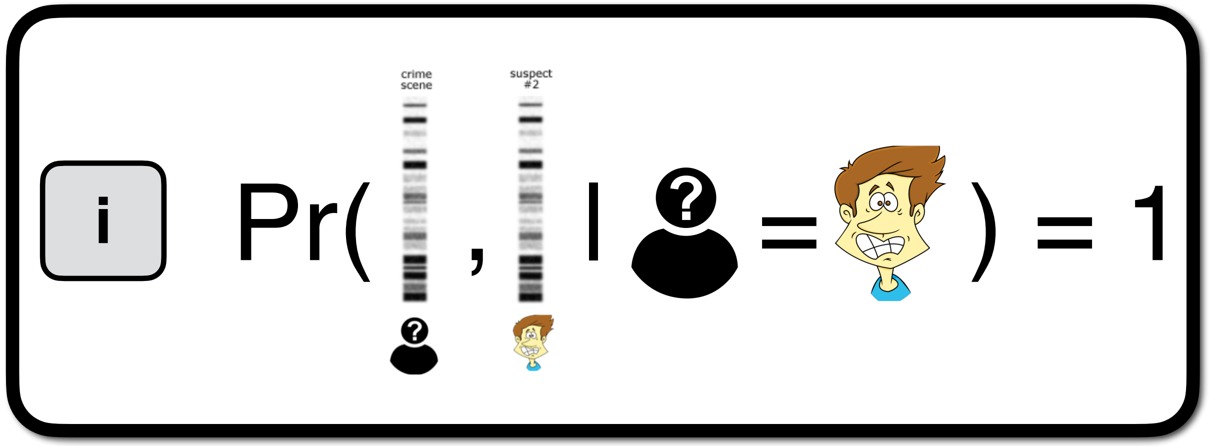
DNA Samples

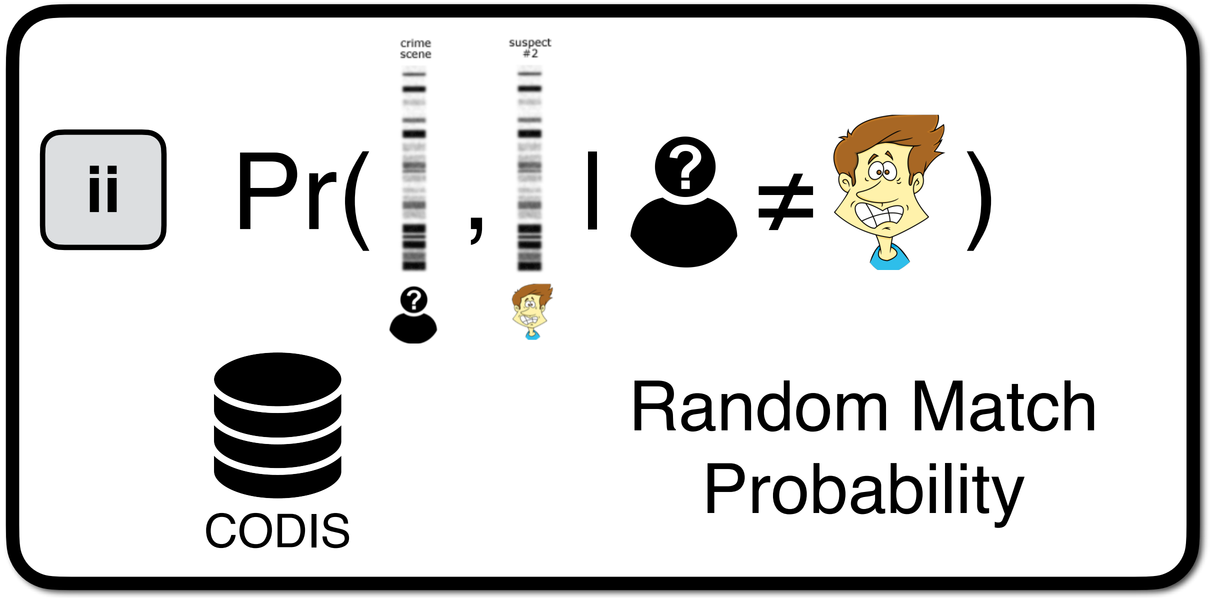




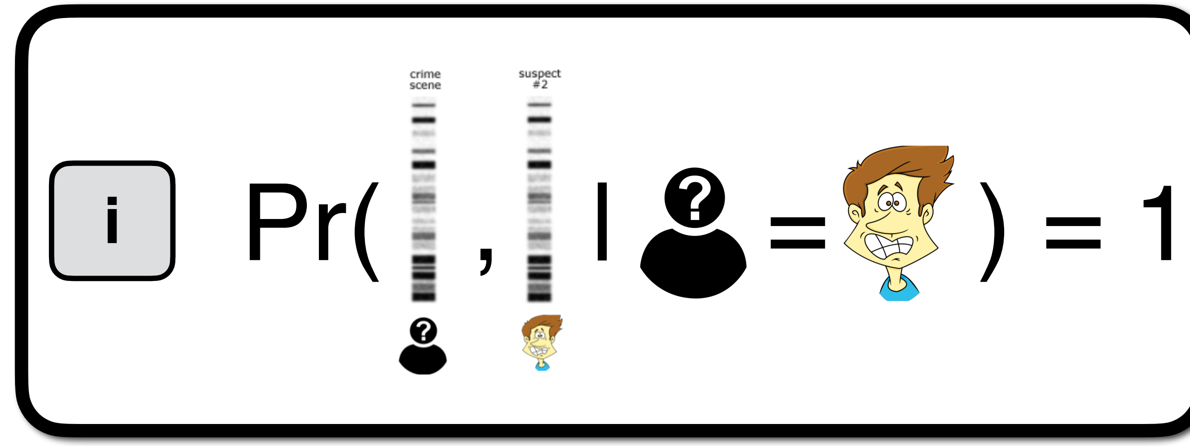
[Steele & Balding, 2014]

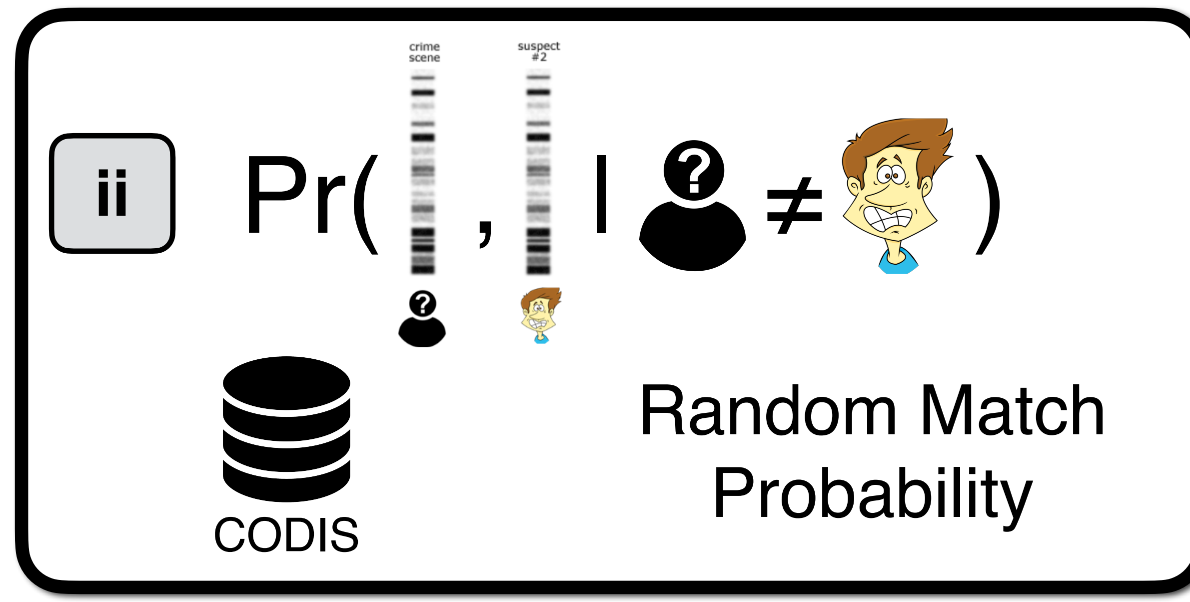






[Steele & Balding, 2014]





Likelihood Ratio

$$\frac{\text{i}}{\text{ii}}$$

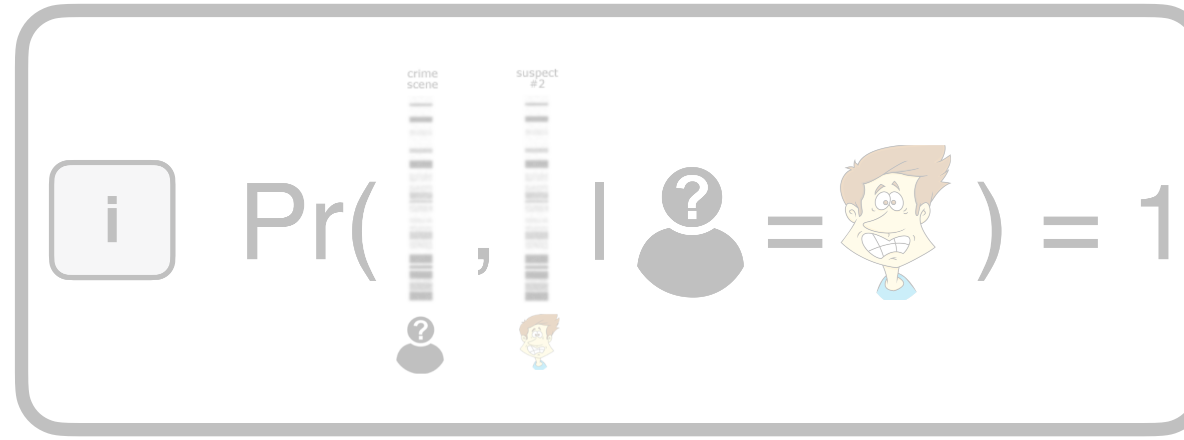
< 1 Samples from different sources

= 1 Inconclusive

> 1 Samples from same source

[Steele & Balding, 2014]

DNA Samples

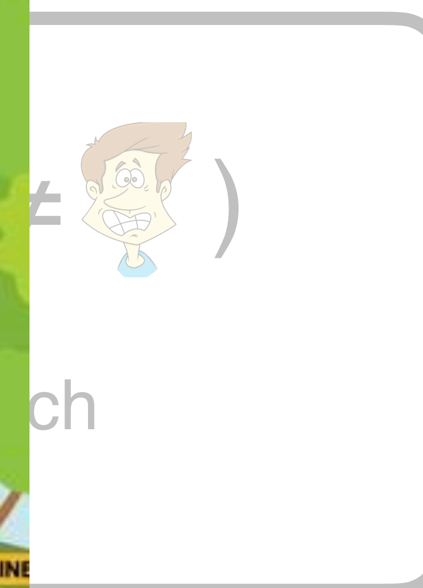


≠)

ch

sources

ource



Extraction

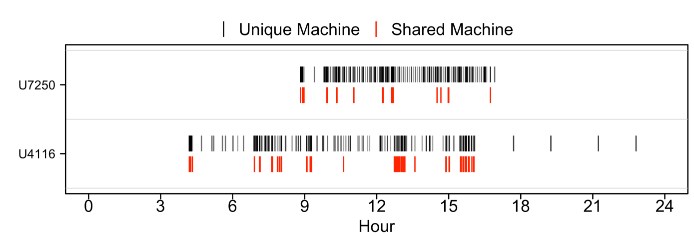
[SWDGE, 2019;
Roussev, 2016;
Casey, 2011]

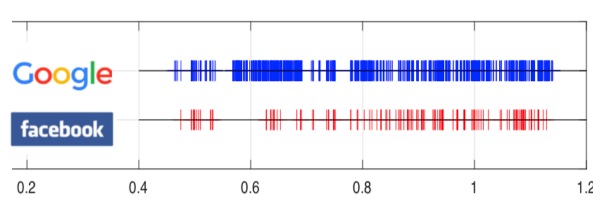
Browser requests
Web searches
Email activity
Phone/SMS
Social media activity
GPS locations
File access
Network activity
Exercise/movement
…

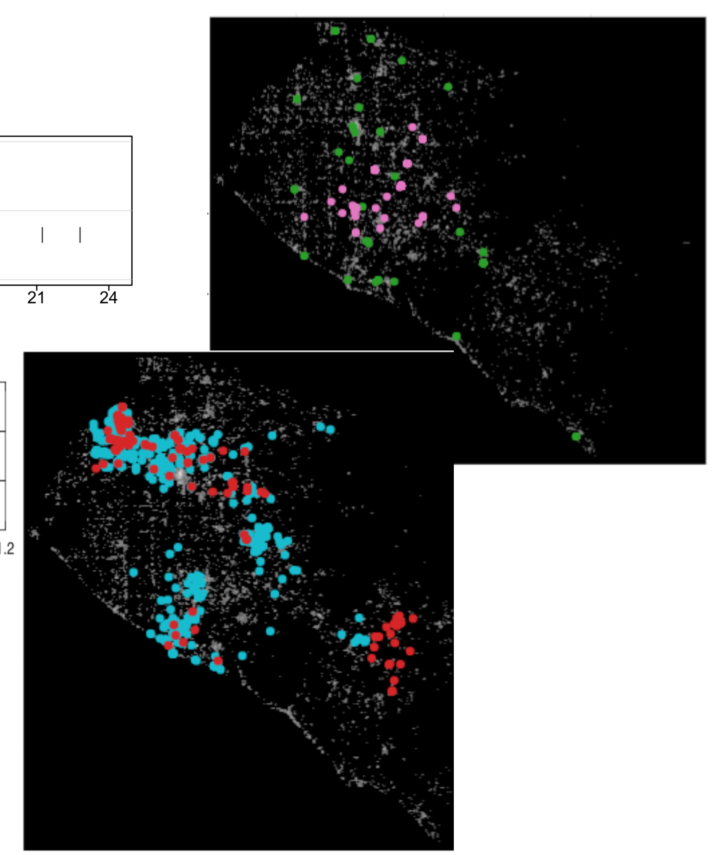
Extraction

[SWDGE, 2019;
Roussev, 2016;
Casey, 2011]

- Browser requests
- Web searches
- Email activity
- Phone/SMS
- Social media activity
- GPS locations
- File access
- Network activity
- Exercise/movement
- …







Analysis

&

Visualization

[Buchholz and Falk, 2005;
Grier, 2011;
Koven et al., 2016;
Gresty et al., 2016;
Kirchler et al., 2016]

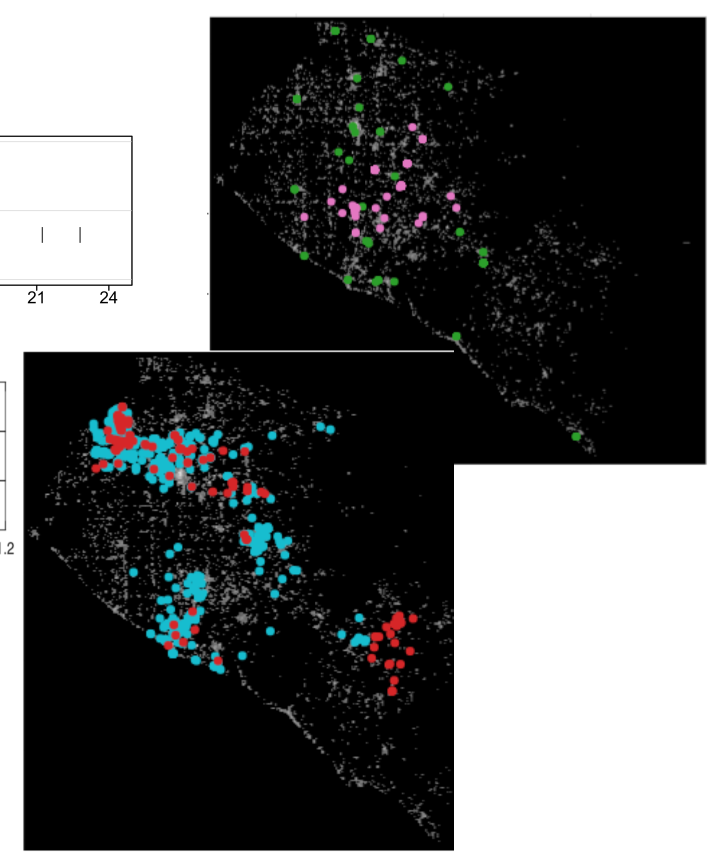
Extraction

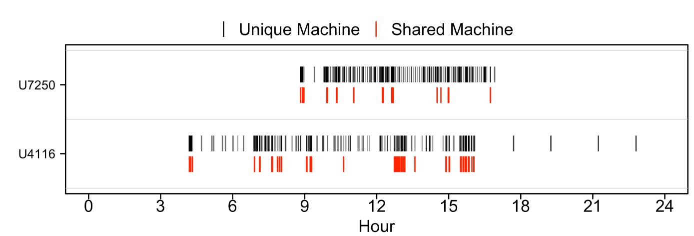
[SWDGE, 2019;
Roussev, 2016;
Casey, 2011]

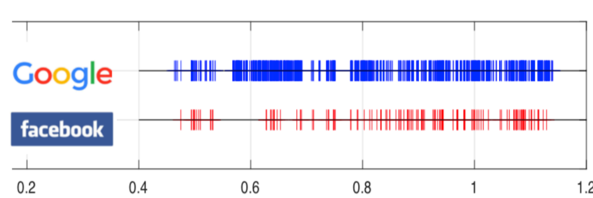
Browser requests
Web searches
Email activity
Phone/SMS
Social media activity
GPS locations
File access
Network activity
Exercise/movement
…

Analysis & Visualization

[Buchholz and Falk, 2005;
Grier, 2011;
Koven et al., 2016;
Gresty et al., 2016;
Kirchler et al., 2016]







?

Probabilistic conclusions regarding source, e.g., Likelihood Ratio

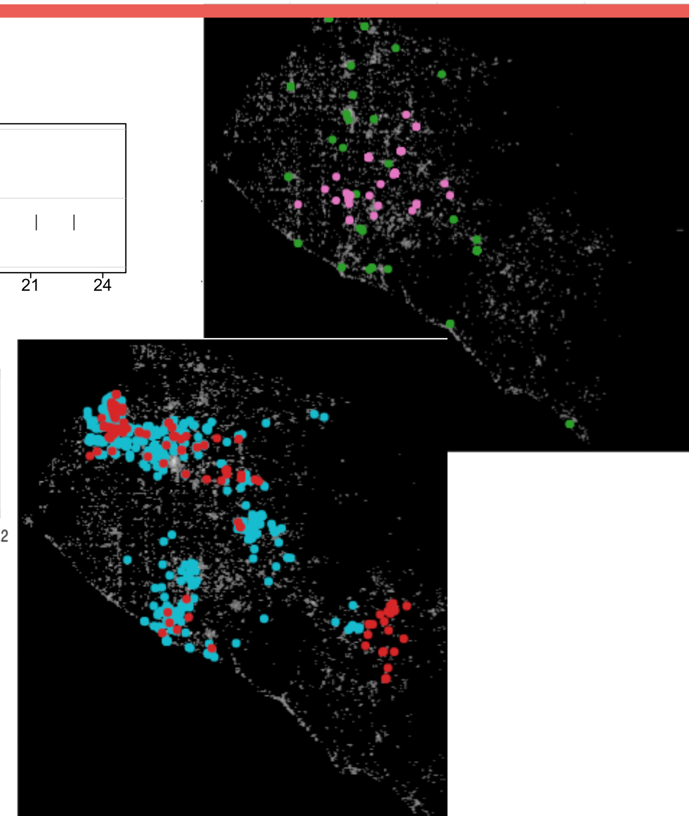
Extraction

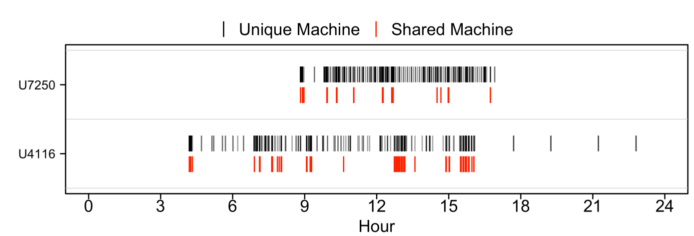
[SWDGE, 2019;
Roussev, 2016;
Casey, 2011]

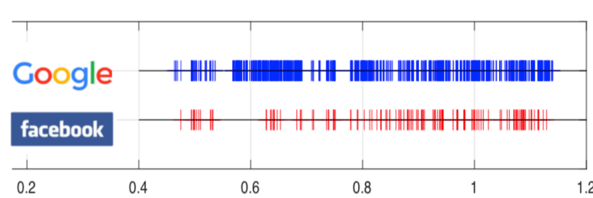
Browser requests
Web searches
Email activity
Phone/SMS
Social media activity
GPS locations
File access
Network activity
Exercise/movement
…

Analysis & Visualization

[Buchholz and Falk, 2005;
Grier, 2011;
Koven et al., 2016;
Gresty et al., 2016;
Kirchler et al., 2016]







Probabilistic conclusions regarding source, e.g., Likelihood Ratio

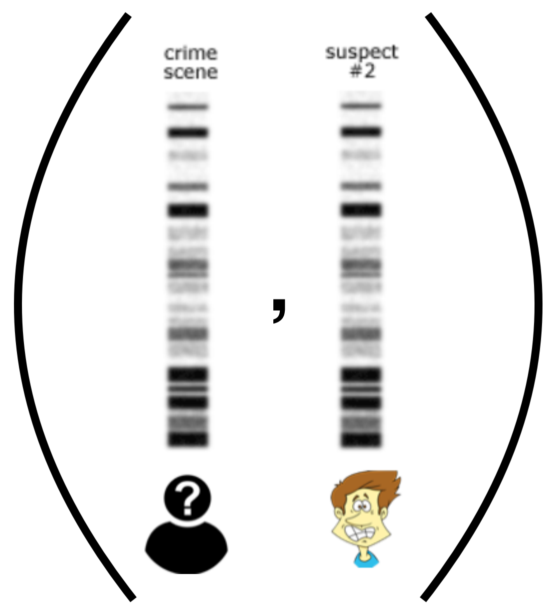
TOPIC OF DISSERTATION

BACKGROUND

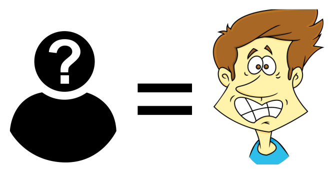
Statistical Approaches for Evaluating Forensic Evidence

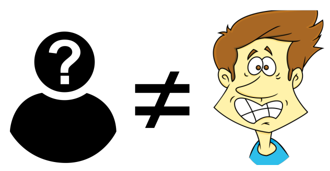
Goal

Assess the likelihood of observing



Under two competing hypotheses



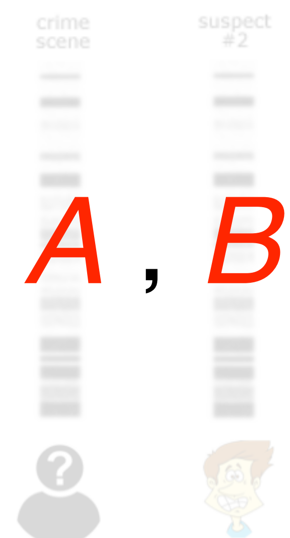


Goal

Assess the likelihood of observing

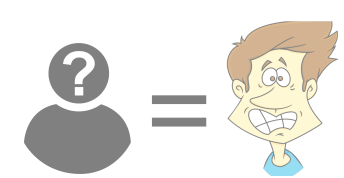
Unknown source sample

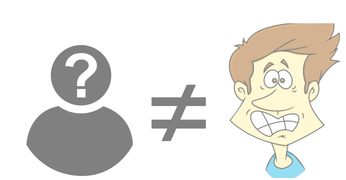
(*A* , *B*)



Known source sample

Under two competing hypotheses

H_s: (*A*, *B*) came from the same source 

H_d: (*A*, *B*) came from the different sources 

Wait…why aren't we interested in the probability of the source hypothesis *given the evidence*?

Wait…why aren't we interested in the probability of the source hypothesis given the evidence?

$$\underbrace{\frac{Pr(H_s \mid A, B, I)}{Pr(H_d \mid A, B, I)}}_{\text{posterior odds}}$$

Wait…why aren't we interested in the probability of the source hypothesis given the evidence?

$$\underbrace{\frac{Pr(H_s \mid A, B, I)}{Pr(H_d \mid A, B, I)}}_{\text{posterior odds}} = \overbrace{\frac{Pr(A, B \mid H_s, I)}{Pr(A, B \mid H_d, I)}}^{\text{likelihood ratio}} \cdot \underbrace{\frac{Pr(H_s \mid I)}{Pr(H_d \mid I)}}_{\text{prior odds}}$$

$$\underbrace{\frac{Pr(H_s\,|\,A,B,I)}{Pr(H_d\,|\,A,B,I)}}_{\text{posterior odds}} = \overbrace{\frac{Pr(A,B\,|\,H_s,I)}{Pr(A,B\,|\,H_d,I)}}^{\text{likelihood ratio}} \cdot \underbrace{\frac{Pr(H_s\,|\,I)}{Pr(H_d\,|\,I)}}_{\text{prior odds}}$$

[Stern, 2017; Thompson, 2017]

$$\underbrace{\frac{Pr(H_s \mid A, B, I)}{Pr(H_d \mid A, B, I)}}_{\text{posterior odds}} = \overbrace{\frac{Pr(A, B \mid H_s, I)}{Pr(A, B \mid H_d, I)}}^{\text{likelihood ratio}} \cdot \underbrace{\frac{Pr(H_s \mid I)}{Pr(H_d \mid I)}}_{\text{prior odds}}$$

[Stern, 2017; Thompson, 2017]

“Strength of Evidence”

“Weight of Evidence”

[Pierce, 1878]

$$\underbrace{\frac{Pr(H_s \mid A, B, I)}{Pr(H_d \mid A, B, I)}}_{\text{posterior odds}} = \overbrace{\frac{Pr(A, B \mid H_s, I)}{Pr(A, B \mid H_d, I)}}^{\text{likelihood ratio}} \cdot \underbrace{\frac{Pr(H_s \mid I)}{Pr(H_d \mid I)}}_{\text{prior odds}}$$

[Stern, 2017; Thompson, 2017]

The Likelihood Ratio

- Widely accepted as a “logically defensible way” to asses the strength of evidence [Willis et al., 2016]

The Likelihood Ratio

- Widely accepted as a “logically defensible way” to asses the strength of evidence [Willis et al., 2016]

- Has been applied in a variety of forensic disciplines

The Likelihood Ratio

- Widely accepted as a “logically defensible way” to asses the strength of evidence [Willis et al., 2016]

- Has been applied in a variety of forensic disciplines
  - DNA [Aitken & Stoney, 1991; Evett & Weir, 1998; Steele & Balding, 2014]
  - Fingerprints [Champod & Evett, 2001]
  - Handwriting [Bozza et al., 2008]
  - Speaker Recognition [Champod & Meuwly, 2000]

The Likelihood Ratio

- Widely accepted as a “logically defensible way” to asses the strength of evidence [Willis et al., 2016]
- Has been applied in a variety of forensic disciplines
  - DNA [Aitken & Stoney, 1991; Evett & Weir, 1998; Steele & Balding, 2014]
  - Fingerprints [Champod & Evett, 2001]
  - Handwriting [Bozza et al., 2008]
  - Speaker Recognition [Champod & Meuwly, 2000]
- Studies demonstrating its understanding

The Likelihood Ratio

- Widely accepted as a “logically defensible way” to asses the strength of evidence [Willis et al., 2016]

- Has been applied in a variety of forensic disciplines
  - DNA [Aitken & Stoney, 1991; Evett & Weir, 1998; Steele & Balding, 2014]
  - Fingerprints [Champod & Evett, 2001]
  - Handwriting [Bozza et al., 2008]
  - Speaker Recognition [Champod & Meuwly, 2000]

- Studies demonstrating its understanding
  - Misconceptions [Martire et al., 2013, Thompson and Newman, 2015, Thompson et al., 2018]

The Likelihood Ratio

- Widely accepted as a “logically defensible way” to asses the strength of evidence [Willis et al., 2016]

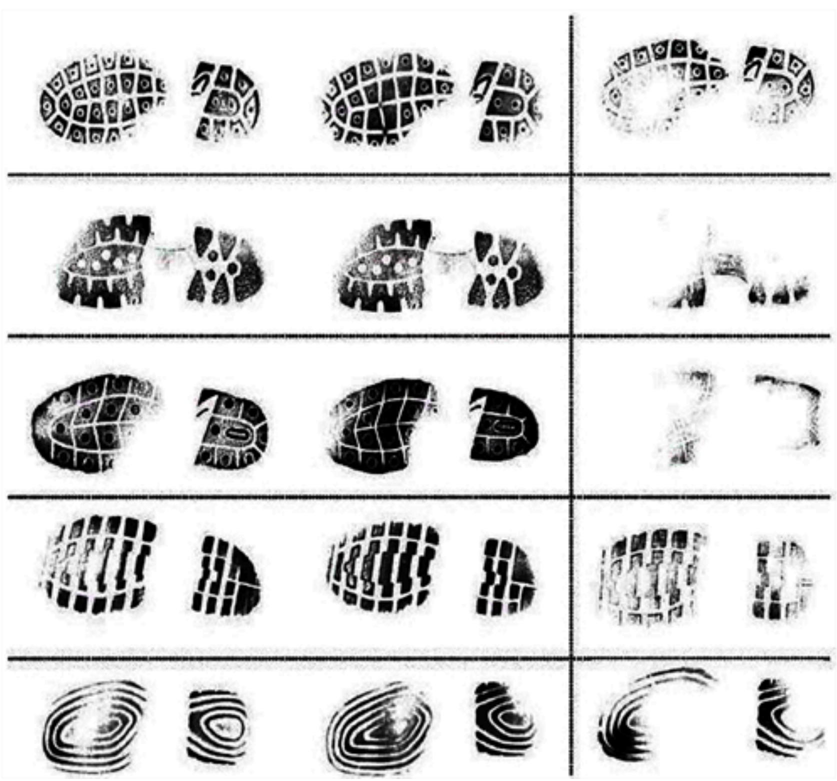
- Has been applied in a variety of forensic disciplines
  - → DNA [Aitken & Stoney, 1991; Evett & Weir, 1998; Steele & Balding, 2014]
  - → Fingerprints [Champod & Evett, 2001]
  - → Handwriting [Bozza et al., 2008]
  - → Speaker Recognition [Champod & Meuwly, 2000]

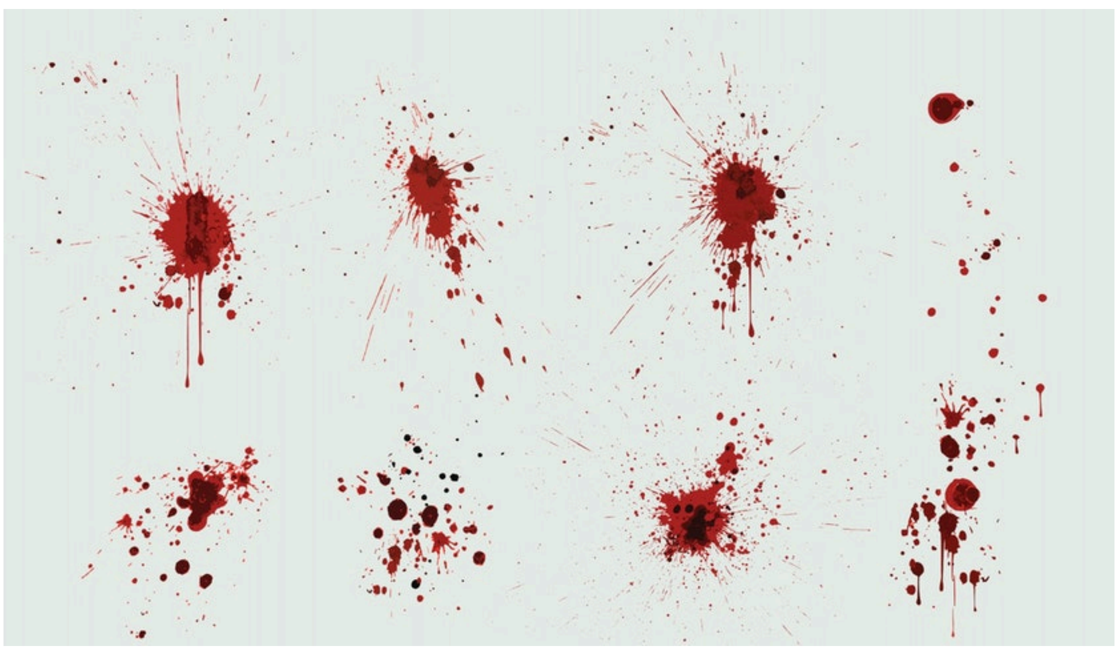
- Studies demonstrating its understanding
  - → Misconceptions [Martire et al., 2013, Thompson and Newman, 2015, Thompson et al., 2018]
  - → Verbal Equivalents [e.g., AFSP, 2009]

Why not always use the LR?

Why not always use the LR?

- **Complexity:** Evidence can be high-dimensional



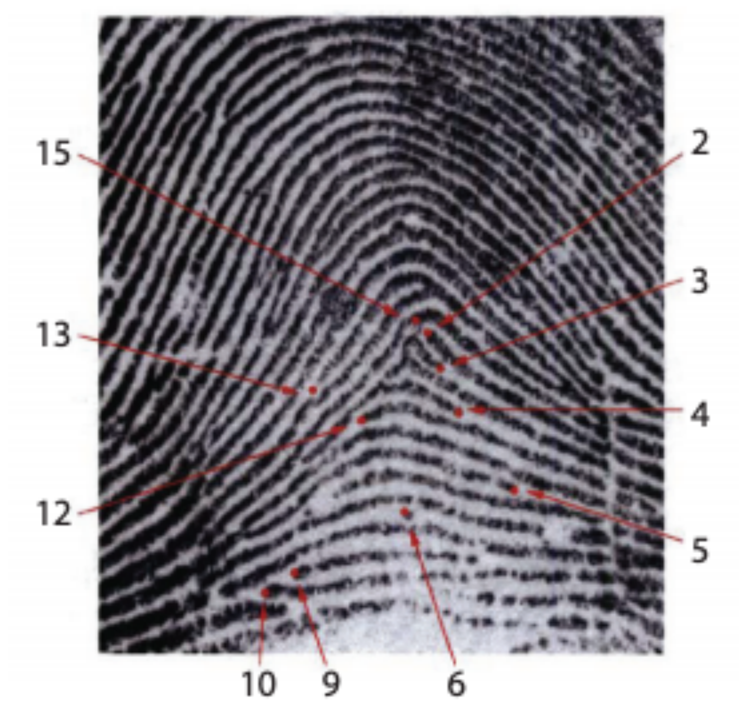


[Stern, 2017]

Why not always use the LR?

- **Complexity:** Evidence can be high-dimensional

- **Feature Selection:** Wide variety of features to consider



[Fine, 2016]

[Stern, 2017]

Why not always use the LR?

- **Complexity:** Evidence can be high-dimensional

- **Feature Selection:** Wide variety of features to consider

- **Appropriate Probability Models:** Must describe variation within a given source and between different sources

[Stern, 2017]

Why not always use the LR?

- **Complexity:** Evidence can be high-dimensional
- **Feature Selection:** Wide variety of features to consider
- **Appropriate Probability Models:** Must describe variation within a given source and between different sources
- **Reference Population:** Difficult to identify a *relevant* reference population to estimate model parameters & perform validation studies

[Stern, 2017]

Score-based Approaches

- Measure similarity between *A* and *B* via a *score function*

$$\Delta(A, B)$$

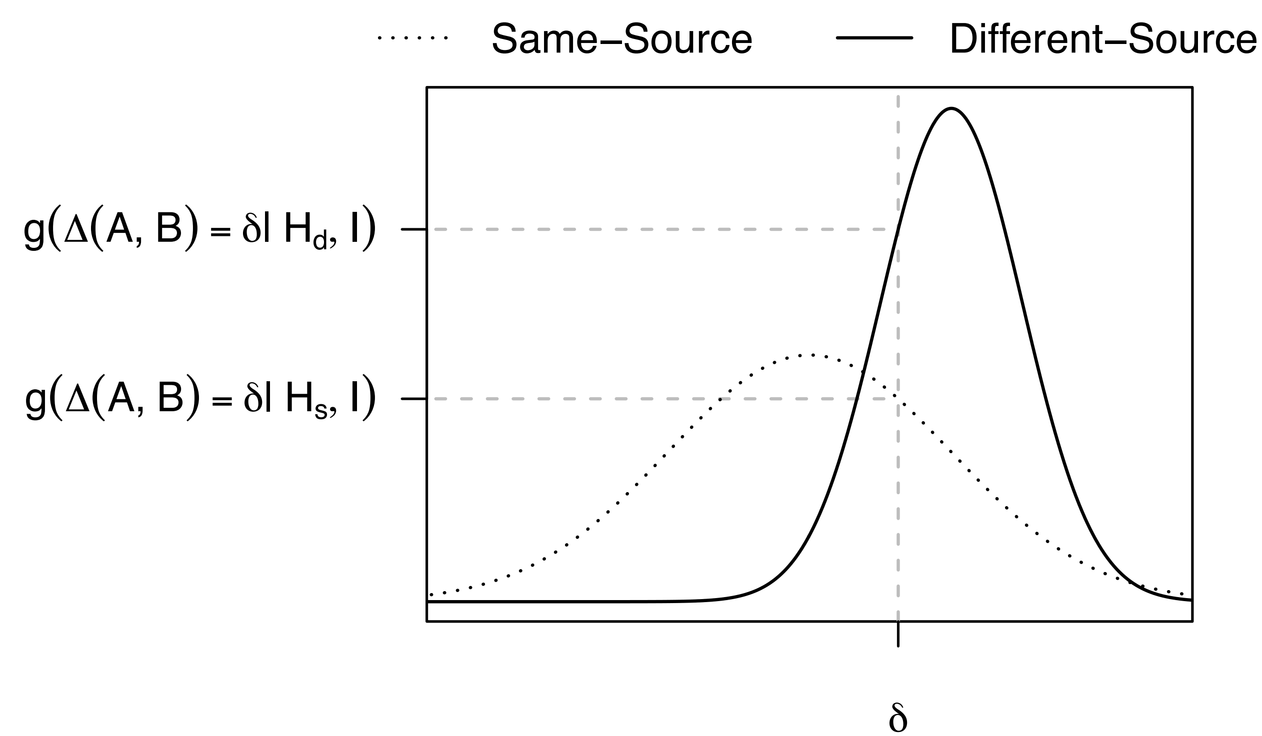
Score-based Approaches

- Measure similarity between *A* and *B* via a *score function*

$$\Delta(A, B)$$

- **Score-based Likelihood Ratio:** Compute a LR for the observed score

$$SLR_{\Delta} = \frac{g(\Delta(A,B) = \delta \,|\, H_s, I)}{g(\Delta(A,B) = \delta \,|\, H_d, I)}$$



Score-based Approaches

- Measure similarity between *A* and *B* via a *score function*

$$\Delta(A, B)$$

- **Score-based Likelihood Ratio:** Compute a LR for the observed score

$$SLR_{\Delta} = \frac{g(\Delta(A,B) = \delta \,|\, H_s, I)}{g(\Delta(A,B) = \delta \,|\, H_d, I)}$$

Gaining popularity in a variety of forensic disciplines

- Chemical Concentrations [Bolck et al., 2015]
- Speaker Recognition [Gonzalez-Rodriguez et al., 2007]
- Fingerprints [Alberink et al., 2013; Neumann et al., 2015]
- Handwriting [Hepler et al., 2012]

Score-based Approaches

- Measure similarity between *A* and *B* via a *score function*

$$\Delta(A, B)$$

- **Score-based Likelihood Ratio:** Compute a LR for the observed score

$$SLR_\Delta = \frac{g(\Delta(A,B) = \delta \,|\, H_s, I)}{g(\Delta(A,B) = \delta \,|\, H_d, I)}$$

CONTRIBUTION CHAPTER #3 [Galbraith, Smyth & Stern, JRSSA 2020]

- **Coincidental Match Probability:**

Score-based Approaches

- Measure similarity between *A* and *B* via a *score function*

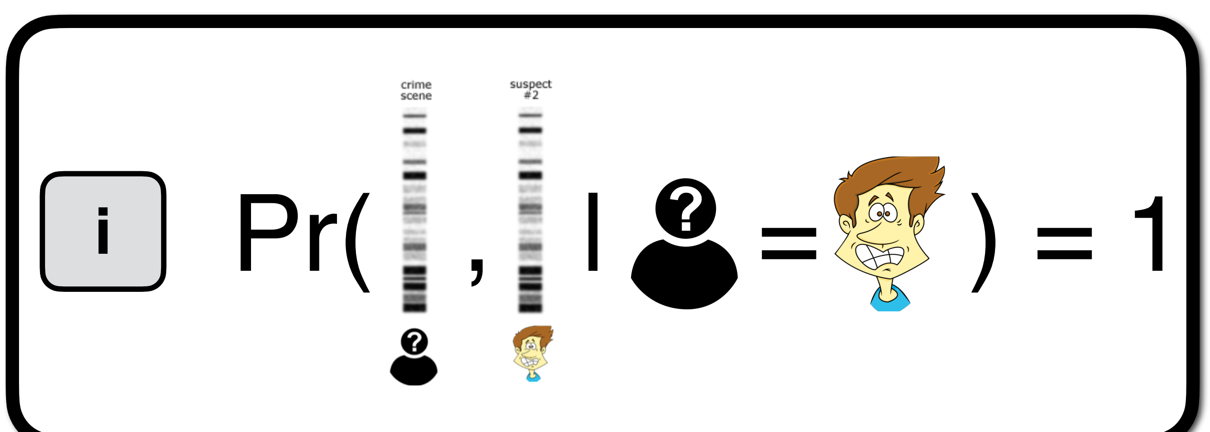
$$\Delta(A, B)$$

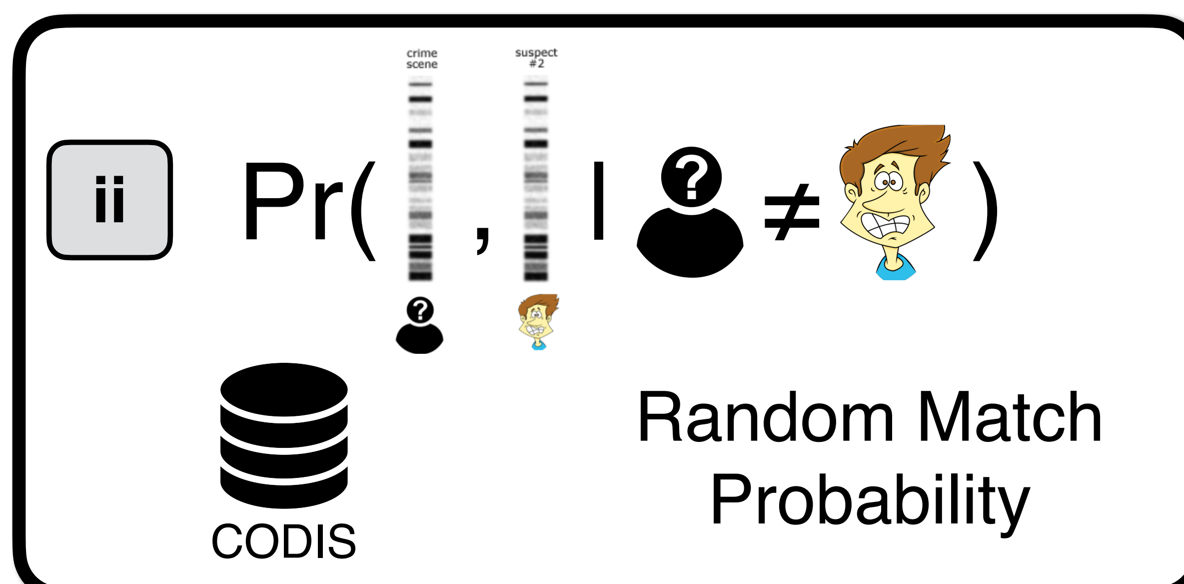
- **Score-based Likelihood Ratio:** Compute a LR for the observed score

$$SLR_\Delta = \frac{g(\Delta(A,B) = \delta \,|\, H_s, I)}{g(\Delta(A,B) = \delta \,|\, H_d, I)}$$

CONTRIBUTION CHAPTER #3 [Galbraith, Smyth & Stern, JRSSA 2020]

- **Coincidental Match Probability:**





Score-based Approaches

- Measure similarity between *A* and *B* via a *score function*

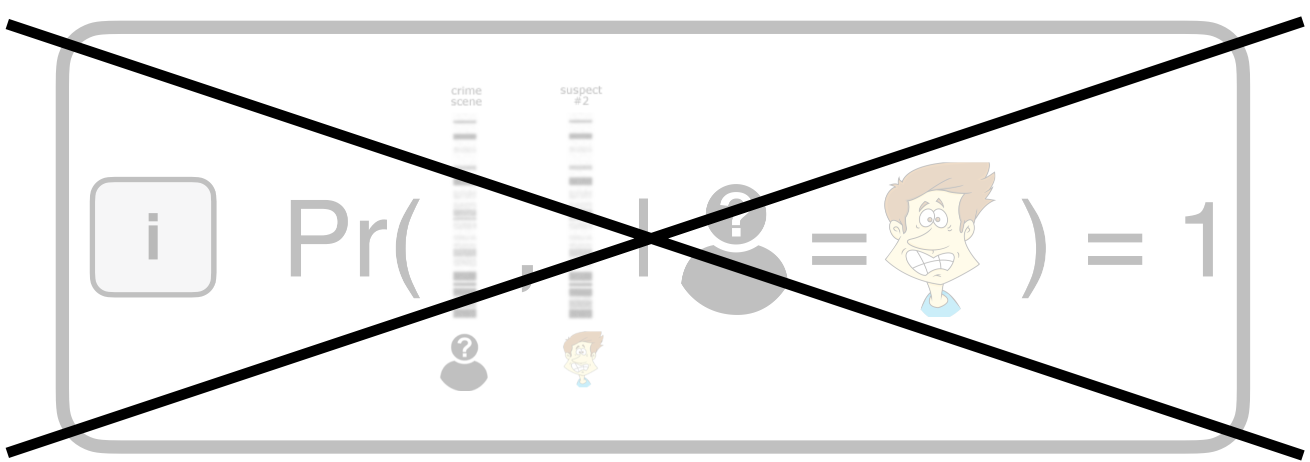
$$\Delta(A, B)$$

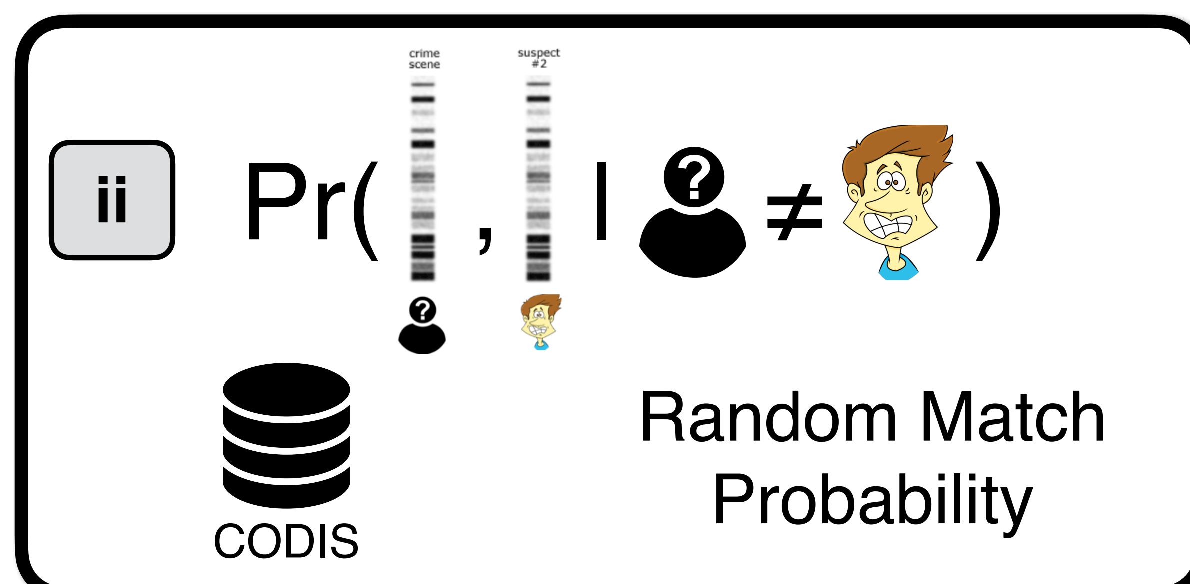
- **Score-based Likelihood Ratio:** Compute a LR for the observed score

$$SLR_{\Delta} = \frac{g(\Delta(A,B) = \delta \,|\, H_s, I)}{g(\Delta(A,B) = \delta \,|\, H_d, I)}$$

CONTRIBUTION CHAPTER #3 [Galbraith, Smyth & Stern, JRSSA 2020]

- **Coincidental Match Probability:**





Score-based Approaches

- Measure similarity between *A* and *B* via a *score function*

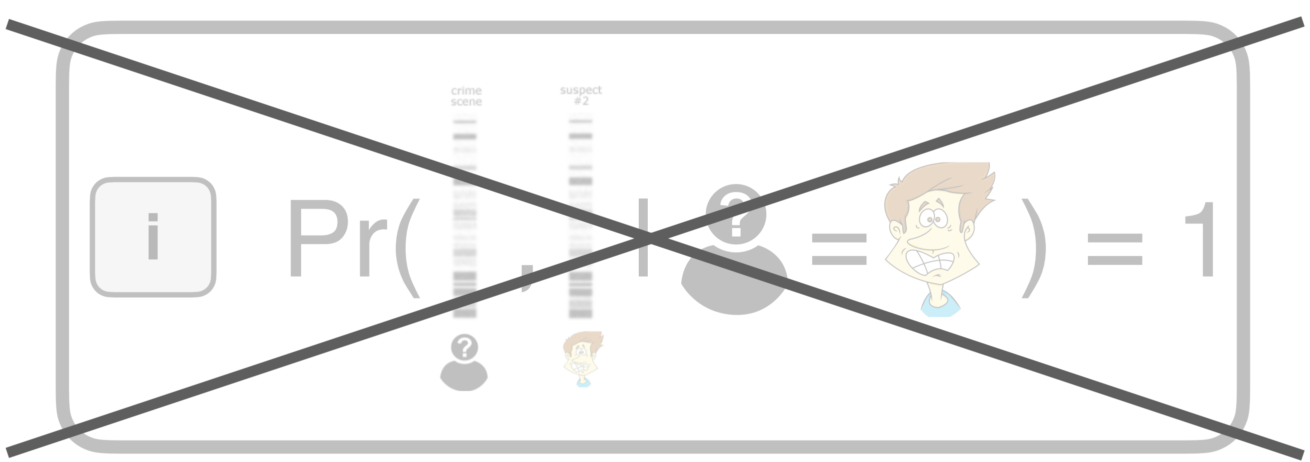
$$\Delta(A, B)$$

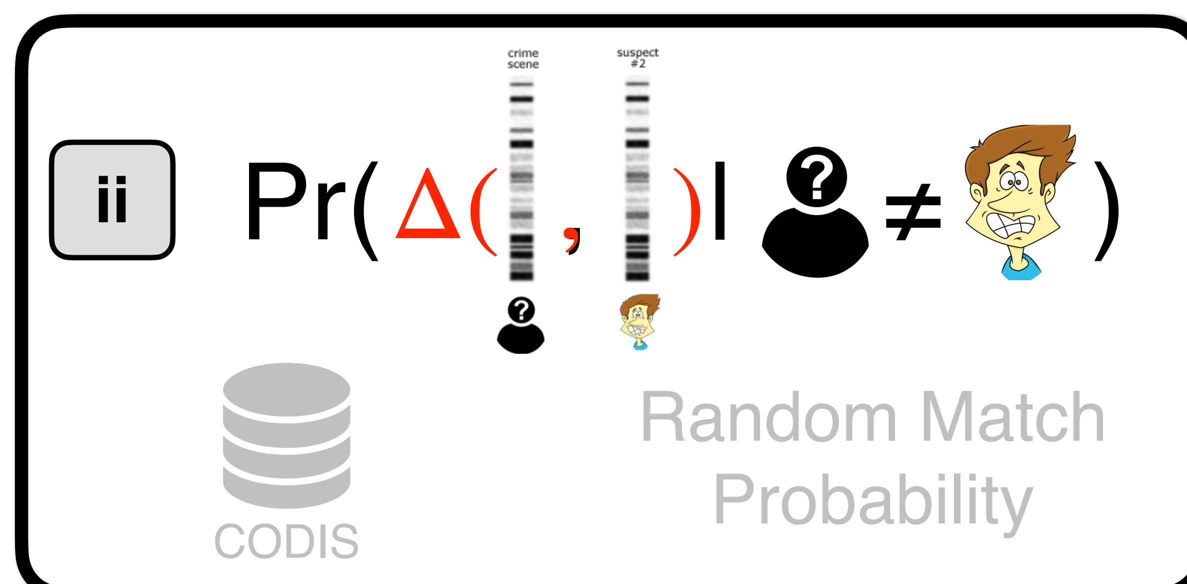
- **Score-based Likelihood Ratio:** Compute a LR for the observed score

$$SLR_{\Delta} = \frac{g(\Delta(A,B) = \delta \,|\, H_s, I)}{g(\Delta(A,B) = \delta \,|\, H_d, I)}$$

CONTRIBUTION CHAPTER #3 [Galbraith, Smyth & Stern, JRSSA 2020]

- **Coincidental Match Probability:**





Score-based Approaches

- Measure similarity between *A* and *B* via a *score function*

$$\Delta(A, B)$$

- **Score-based Likelihood Ratio:** Compute a LR for the observed score

$$SLR_{\Delta} = \frac{g(\Delta(A, B) = \delta \,|\, H_s, I)}{g(\Delta(A, B) = \delta \,|\, H_d, I)}$$

CONTRIBUTION CHAPTER #3 [Galbraith, Smyth & Stern, JRSSA 2020]

- **Coincidental Match Probability:** Probability that different-source evidence has a more extreme score than the observed score

Score-based Approaches

- Measure similarity between *A* and *B* via a *score function*

$$\Delta(A, B)$$

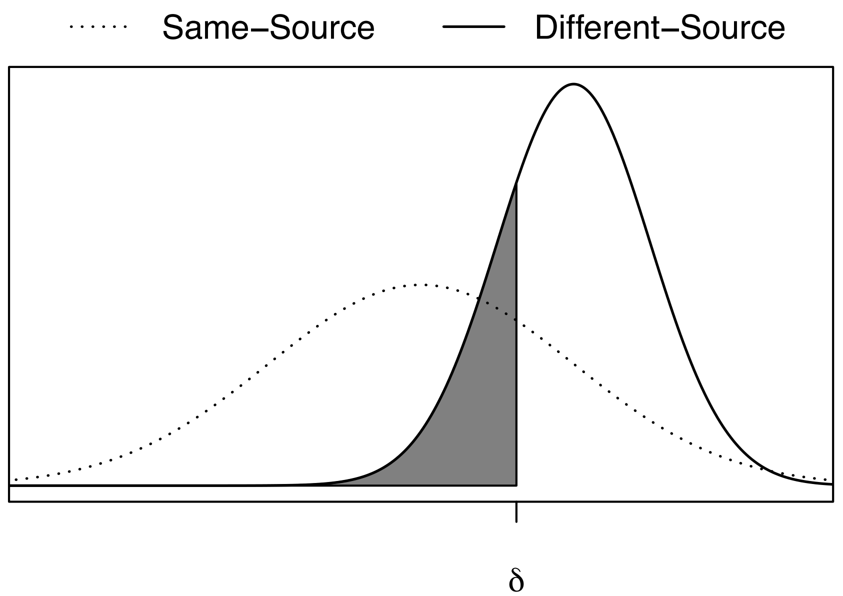
- **Score-based Likelihood Ratio:** Compute a LR for the observed score

$$SLR_\Delta = \frac{g(\Delta(A,B) = \delta \mid H_s, I)}{g(\Delta(A,B) = \delta \mid H_d, I)}$$

CONTRIBUTION CHAPTER #3 [Galbraith, Smyth & Stern, JRSSA 2020]

- **Coincidental Match Probability:** Probability that different-source evidence has a more extreme score than the observed score

$$CMP_\Delta = Pr(\Delta(A,B) < \delta \mid H_d, I)$$



Evidence Evaluation Approaches

- **Likelihood Ratio:** Models evidence directly

$$LR = \frac{Pr(A, B \mid H_s, I)}{Pr(A, B \mid H_d, I)}$$

- **Score-based Likelihood Ratio:** Models low-dimensional summary of the evidence, $\Delta(A, B)$

$$SLR_\Delta = \frac{g(\Delta(A, B) = \delta \mid H_s, I)}{g(\Delta(A, B) = \delta \mid H_d, I)}$$

CONTRIBUTION CHAPTER #3 [Galbraith, Smyth & Stern, JRSSA 2020]

- **Coincidental Match Probability:** Focus on different-source score distribution; similar to RMP, but we don't determine a "match" first

$$CMP_\Delta = Pr(\Delta(A, B) < \delta \mid H_d, I)$$

BACKGROUND

Empirical Evaluation Techniques

Empirical Evaluation Techniques

- **Validation Data:** Sample from relevant reference population
  - → $\mathscr{D}_s^*$ known same-source evidence
  - → $\mathscr{D}_d^*$ known different-source evidence

Empirical Evaluation Techniques

- **Validation Data:** Sample from relevant reference population
  - → $\mathscr{D}_s^*$ known same-source evidence
  - → $\mathscr{D}_d^*$ known different-source evidence

- **Classification Performance:** TP/FP rates, AUC

Empirical Evaluation Techniques

- **Validation Data:** Sample from relevant reference population
  - → $\mathscr{D}_s^*$ known same-source evidence
  - → $\mathscr{D}_d^*$ known different-source evidence

- **Classification Performance:** TP/FP rates, AUC

- **Calibration:** Same-source evidence should have larger LR/SLR (or smaller CMP) values than different-source evidence, e.g.,
  - → $LR_s \in \mathscr{D}_s^*, LR_d \in \mathscr{D}_d^* \Rightarrow LR_d < LR_s$

Empirical Evaluation Techniques

- **Validation Data:** Sample from relevant reference population
  - → $\mathscr{D}_s^*$ known same-source evidence
  - → $\mathscr{D}_d^*$ known different-source evidence

- **Classification Performance:** TP/FP rates, AUC

- **Calibration:** Same-source evidence should have larger LR/SLR (or smaller CMP) values than different-source evidence

- **Information-theoretic Evaluation:** How much does the LR/SLR value reduce the uncertainty regarding the source hypotheses?
  - → Empirical cross-entropy [Brümmer & du Preez, 2006; Ramos, 2007]

Empirical Cross-Entropy

- **Cross-entropy:** $\mathscr{U}_{Q||P}(H_s \mid E) = -\,\mathbb{E}_{Q(E,H_s)} \log P(H_s \mid E)$

Empirical Cross-Entropy

- **Cross-entropy:** $\mathscr{U}_{Q||P}(H_s \mid E) = -\,\mathbb{E}_{Q(E,H_s)} \log P(H_s \mid E)$

$$Q(H_s \mid E) = \begin{cases} 1, & H_s \text{ true} \\ 0, & H_d \text{ true} \end{cases}$$

Target Posterior

Empirical Cross-Entropy

- **Cross-entropy:** $\mathscr{U}_{Q||P}(H_s | E) = -\mathbb{E}_{Q(E,H_s)} \log P(H_s | E)$

$$Q(H_s | E) = \begin{cases} 1, & H_s \text{ true} \\ 0, & H_d \text{ true} \end{cases}$$

Target Posterior

$$P(H_s | E) = \frac{LR \times \frac{P(H_s)}{P(H_d)}}{1 + LR \times \frac{P(H_s)}{P(H_d)}}$$

Posterior (from evidence evaluation)

Empirical Cross-Entropy

- **Cross-entropy:** $\mathscr{U}_{Q||P}(H_s \mid E) = -\,\mathbb{E}_{Q(E,H_s)} \log P(H_s \mid E)$

$$Q(H_s \mid E) = \begin{cases} 1, & H_s \text{ true} \\ 0, & H_d \text{ true} \end{cases}$$

Target Posterior

$$P(H_s \mid E) = \frac{LR \times \frac{P(H_s)}{P(H_d)}}{1 + LR \times \frac{P(H_s)}{P(H_d)}}$$

Posterior (from evidence evaluation)

→ Using above target posterior, $\mathscr{U}_{Q||P}(H_s \mid E) = D_{Q||P}(H_s \mid E)$

Empirical Cross-Entropy

- **Cross-entropy:** $\mathscr{U}_{Q||P}(H_s|E) = -\mathbb{E}_{Q(E,H_s)} \log P(H_s|E)$

$$Q(H_s|E) = \begin{cases} 1, & H_s \text{ true} \\ 0, & H_d \text{ true} \end{cases}$$

Target Posterior

$$P(H_s|E) = \frac{LR \times \frac{P(H_s)}{P(H_d)}}{1 + LR \times \frac{P(H_s)}{P(H_d)}}$$

Posterior (from evidence evaluation)

→ Using above target posterior, $\mathscr{U}_{Q||P}(H_s|E) = D_{Q||P}(H_s|E)$

→ Equivalent to Bayes risk for logarithmic loss [Proof in CHAPTER #2]

Empirical Cross-Entropy

- **Cross-entropy:** $\mathscr{U}_{Q||P}(H_s|E) = -\mathbb{E}_{Q(E,H_s)} \log P(H_s|E)$

$$Q(H_s|E) = \begin{cases} 1, & H_s \text{ true} \\ 0, & H_d \text{ true} \end{cases}$$

Target Posterior

$$P(H_s|E) = \frac{LR \times \frac{P(H_s)}{P(H_d)}}{1 + LR \times \frac{P(H_s)}{P(H_d)}}$$

Posterior (from evidence evaluation)

→ Using above target posterior, $\mathscr{U}_{Q||P}(H_s|E) = D_{Q||P}(H_s|E)$

→ Equivalent to Bayes risk for logarithmic loss [Proof in CHAPTER #2]

- **Empirical Cross-entropy:** Estimate $\mathscr{U}_{Q||P}(H_s|E)$ by averaging over validation data

Empirical Cross-Entropy

- **Cross-entropy:** $\mathcal{U}_{Q||P}(H_s|E) = -\mathbb{E}_{Q(E,H_s)} \log P(H_s|E)$

$$Q(H_s|E) = \begin{cases} 1, & H_s \text{ true} \\ 0, & H_d \text{ true} \end{cases}$$

Target Posterior

$$P(H_s|E) = \frac{LR \times \frac{P(H_s)}{P(H_d)}}{1 + LR \times \frac{P(H_s)}{P(H_d)}}$$

Posterior (from evidence evaluation)

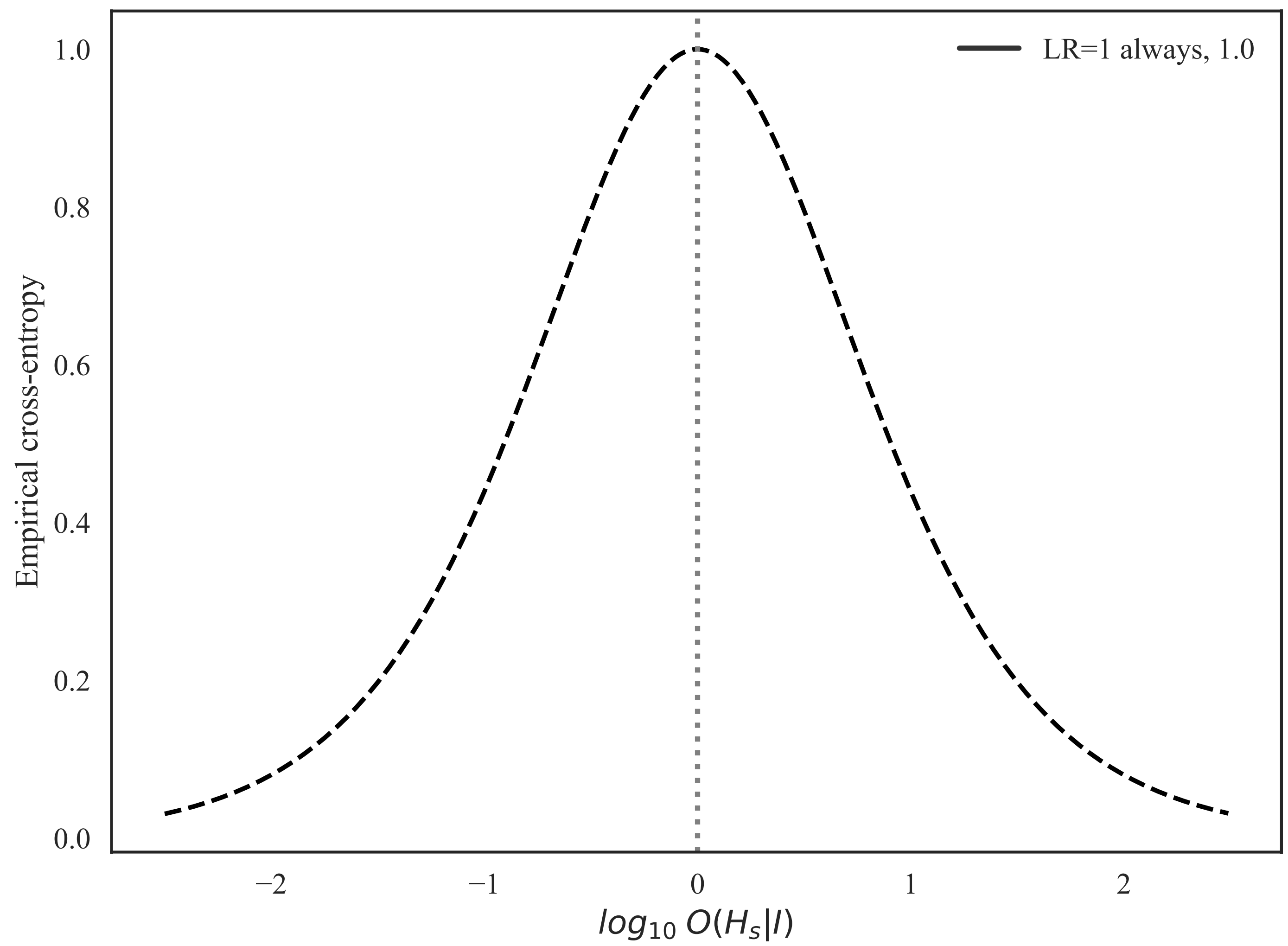
→ Using above target posterior, $\mathcal{U}_{Q||P}(H_s|E) = D_{Q||P}(H_s|E)$

→ Equivalent to Bayes risk for logarithmic loss [Proof in CHAPTER #2]

- **Empirical Cross-entropy:** Estimate $\mathcal{U}_{Q||P}(H_s|E)$ by averaging over validation data

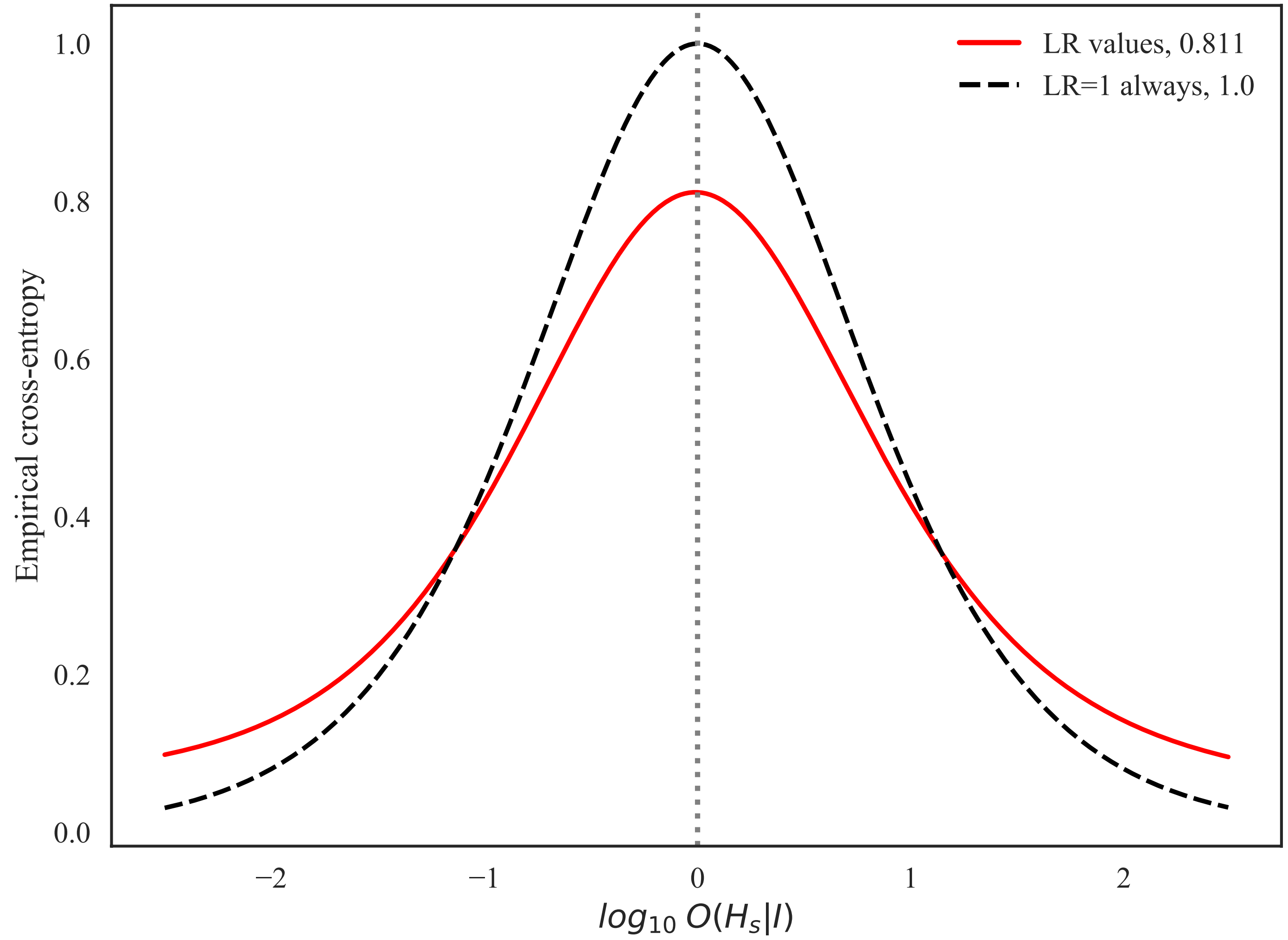
→ Repeat over a range of priors $P(H_s)$

ECE Plot



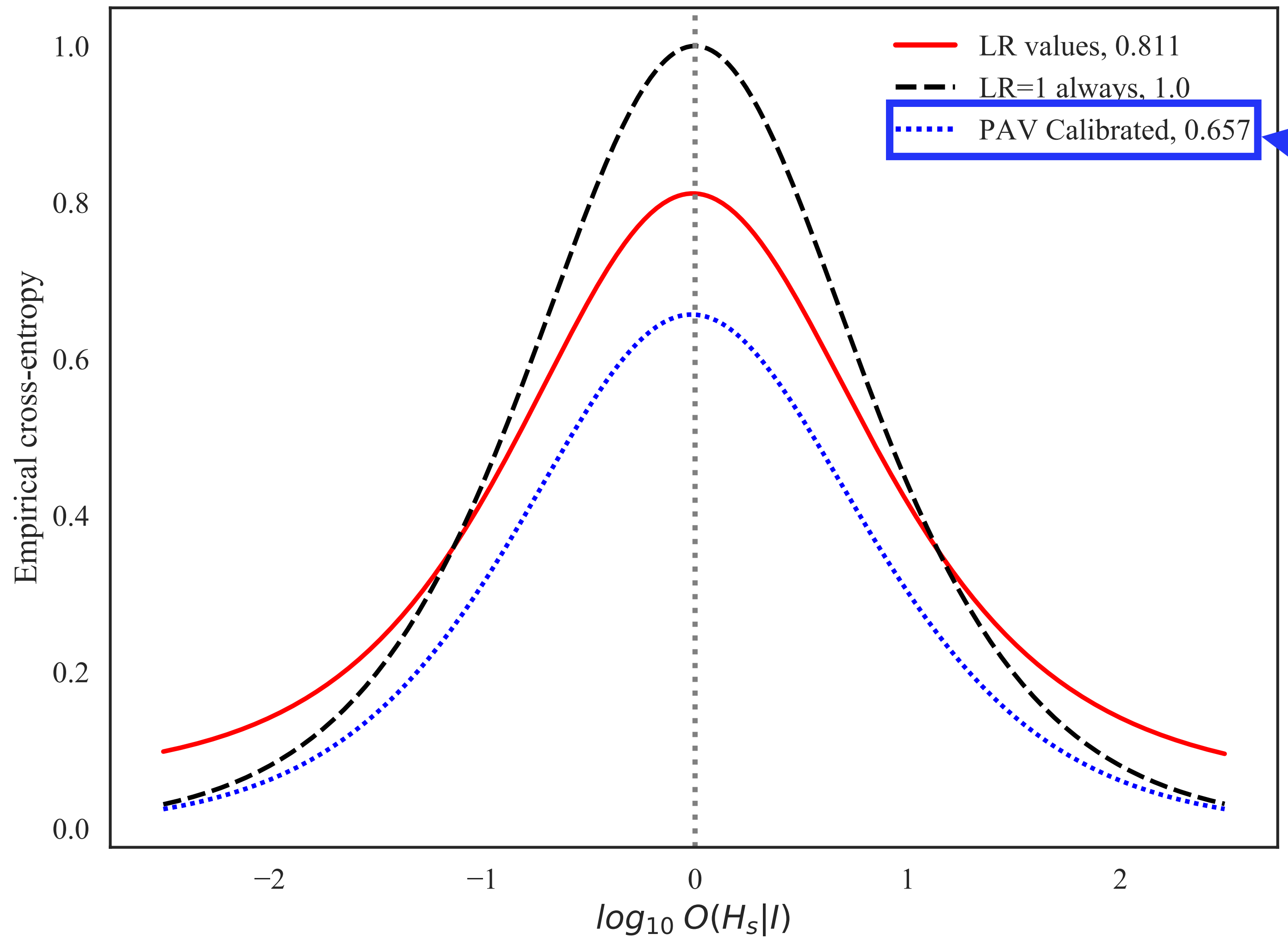
LR=1 always, 1.0
Empirical cross-entropy
1.0
0.8
0.6
0.4
0.2
0.0
−2
−1
0
1
2
$log_{10}\, O(H_s|I)$

ECE Plot



LR values, 0.811
LR=1 always, 1.0
Empirical cross-entropy
1.0
0.8
0.6
0.4
0.2
0.0
−2
−1
0
1
2
$log_{10}\, O(H_s|I)$

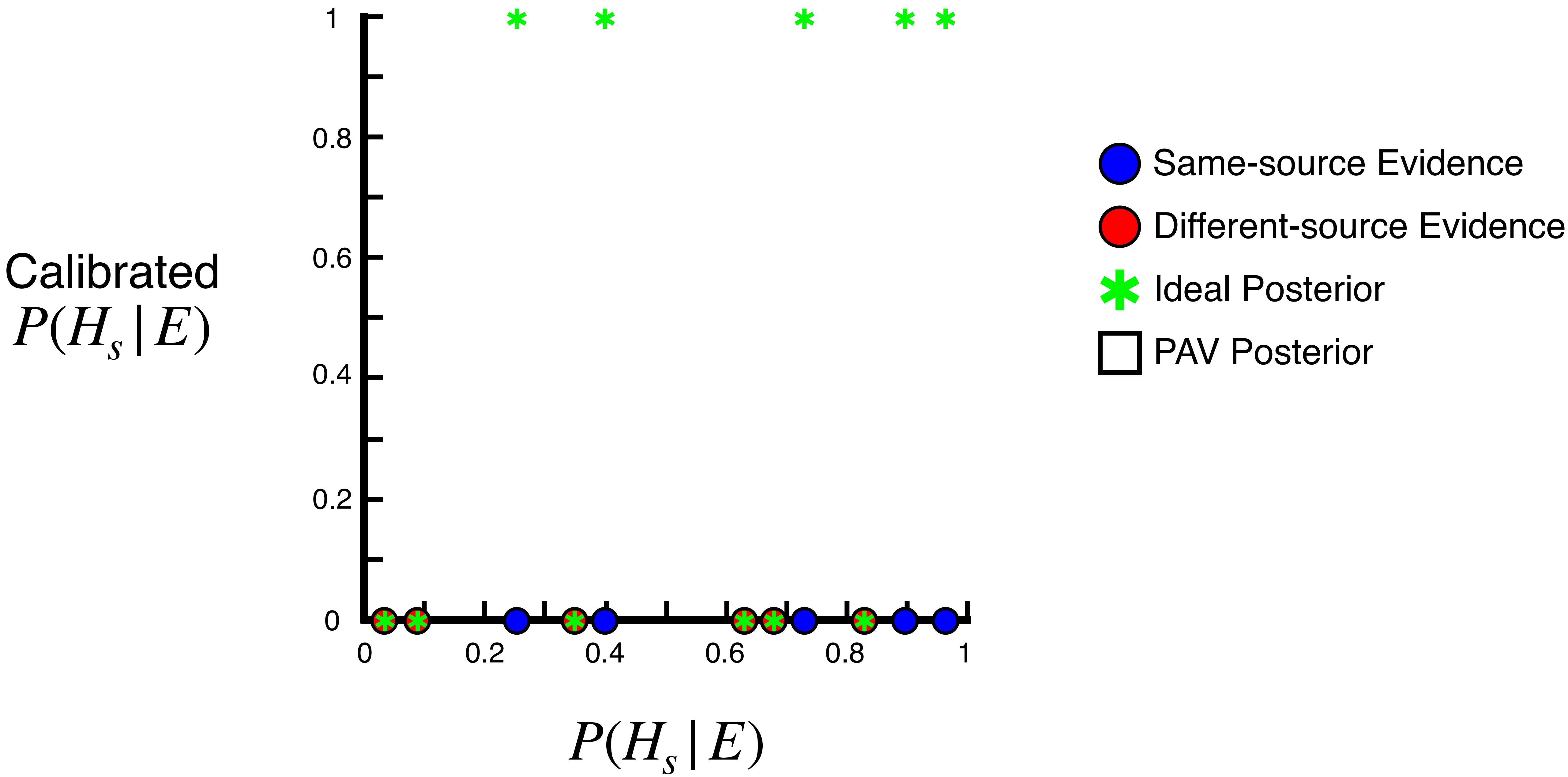
ECE Plot



Isotonic Regression

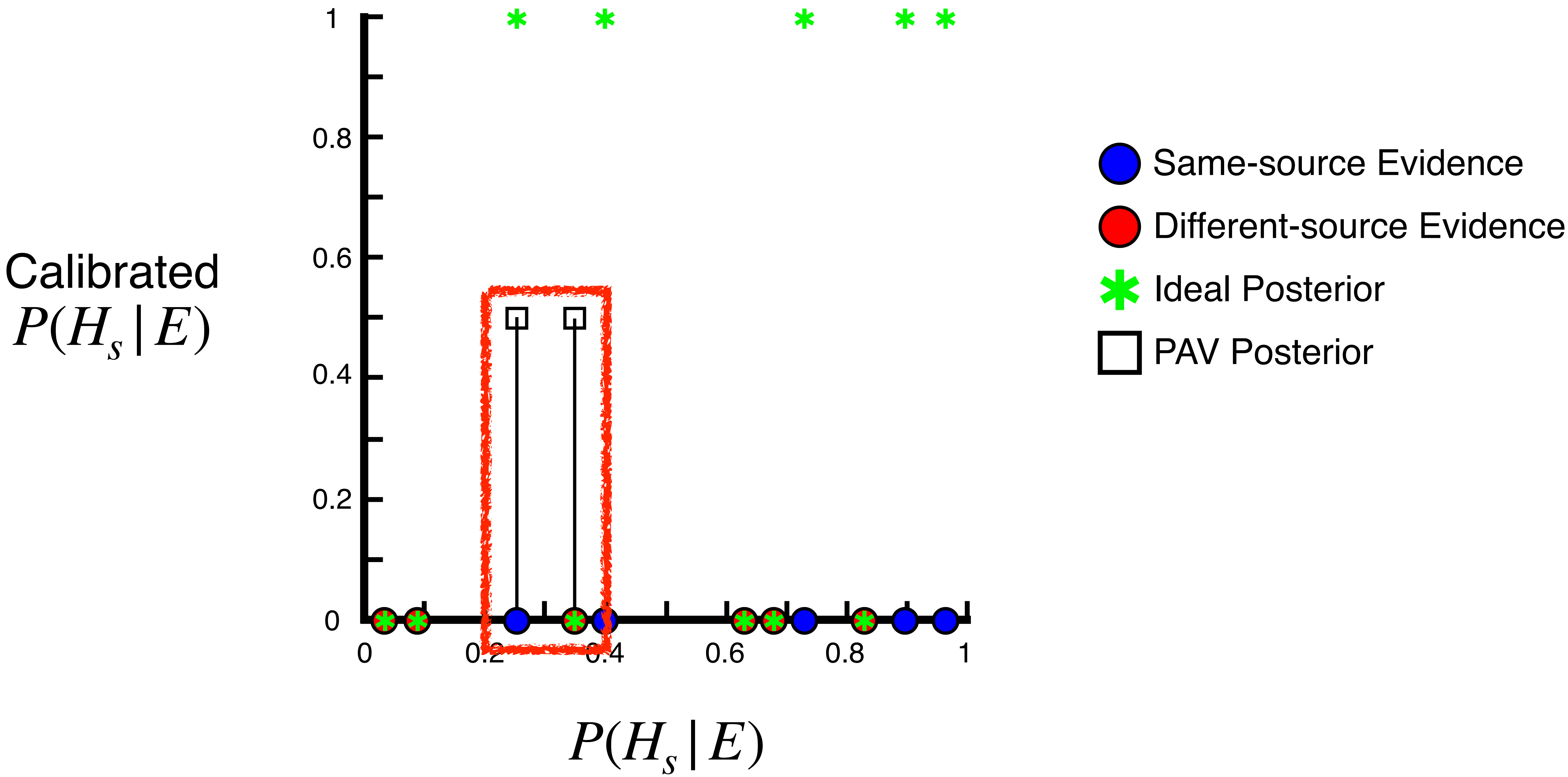
[Zadrozny & Elkan, 2002]

Aside: Calibration via Isotonic Regression



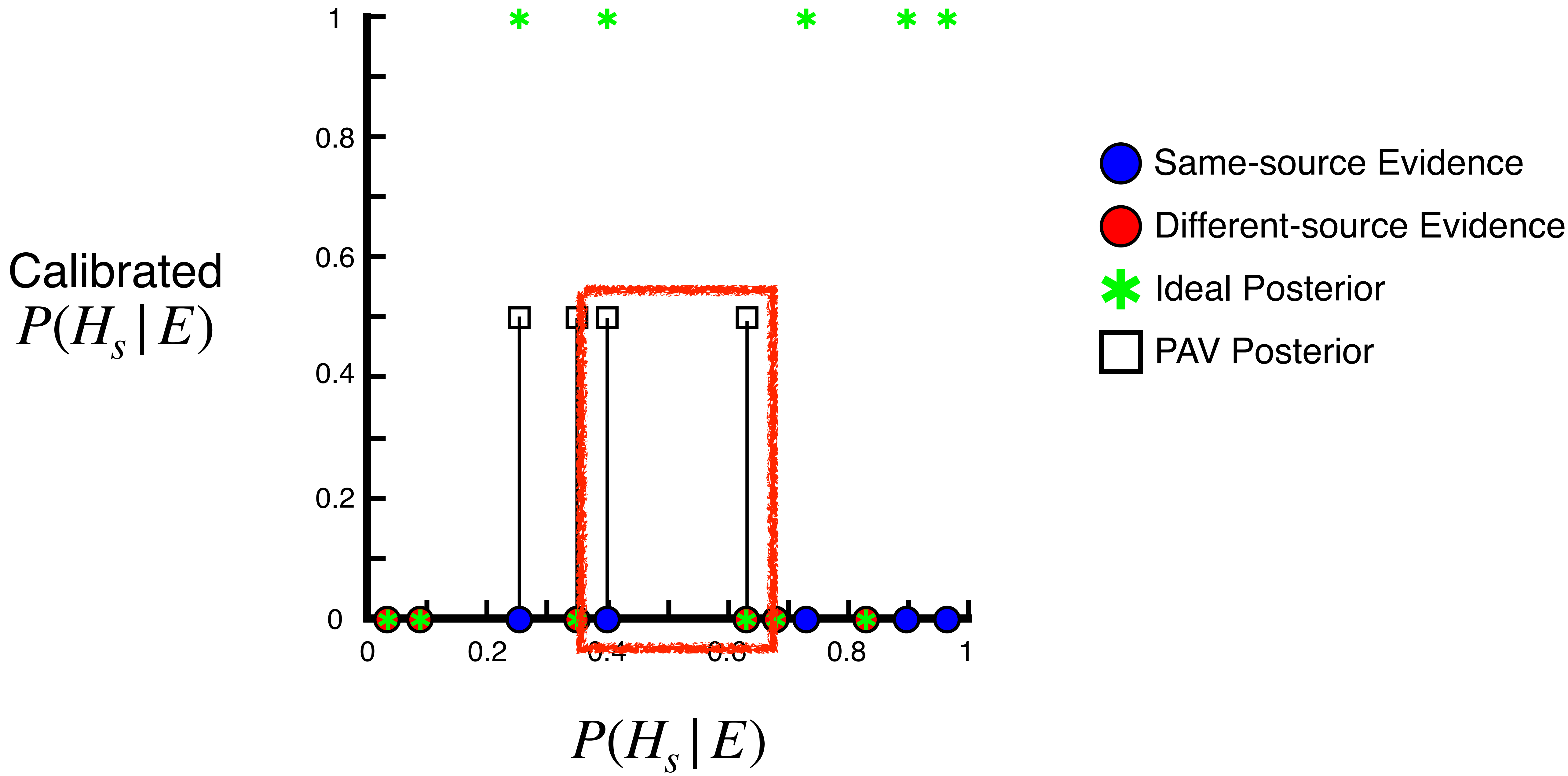
[Barlow, 1972]

Aside: Calibration via Isotonic Regression



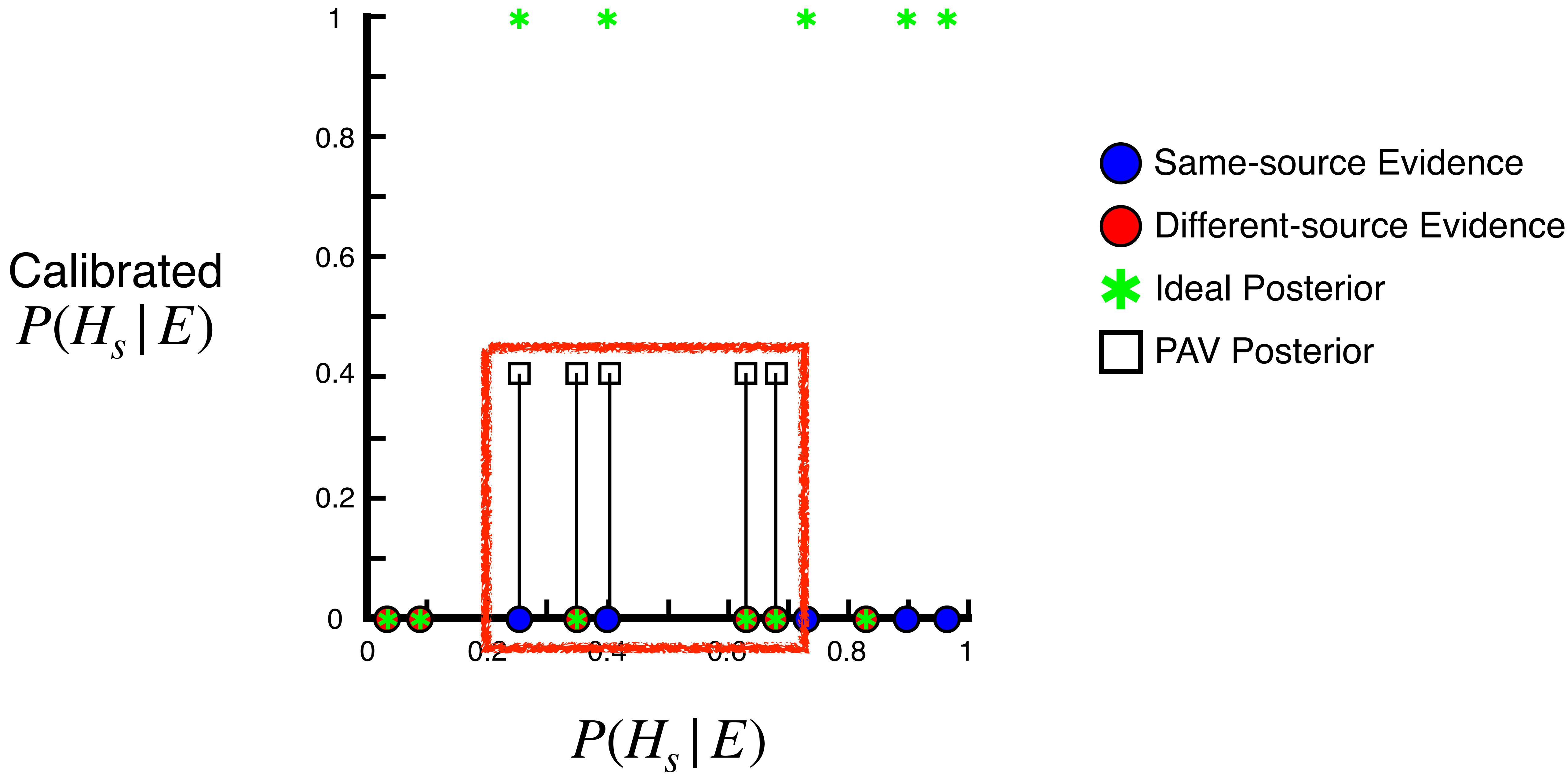
[Barlow, 1972]

Aside: Calibration via Isotonic Regression



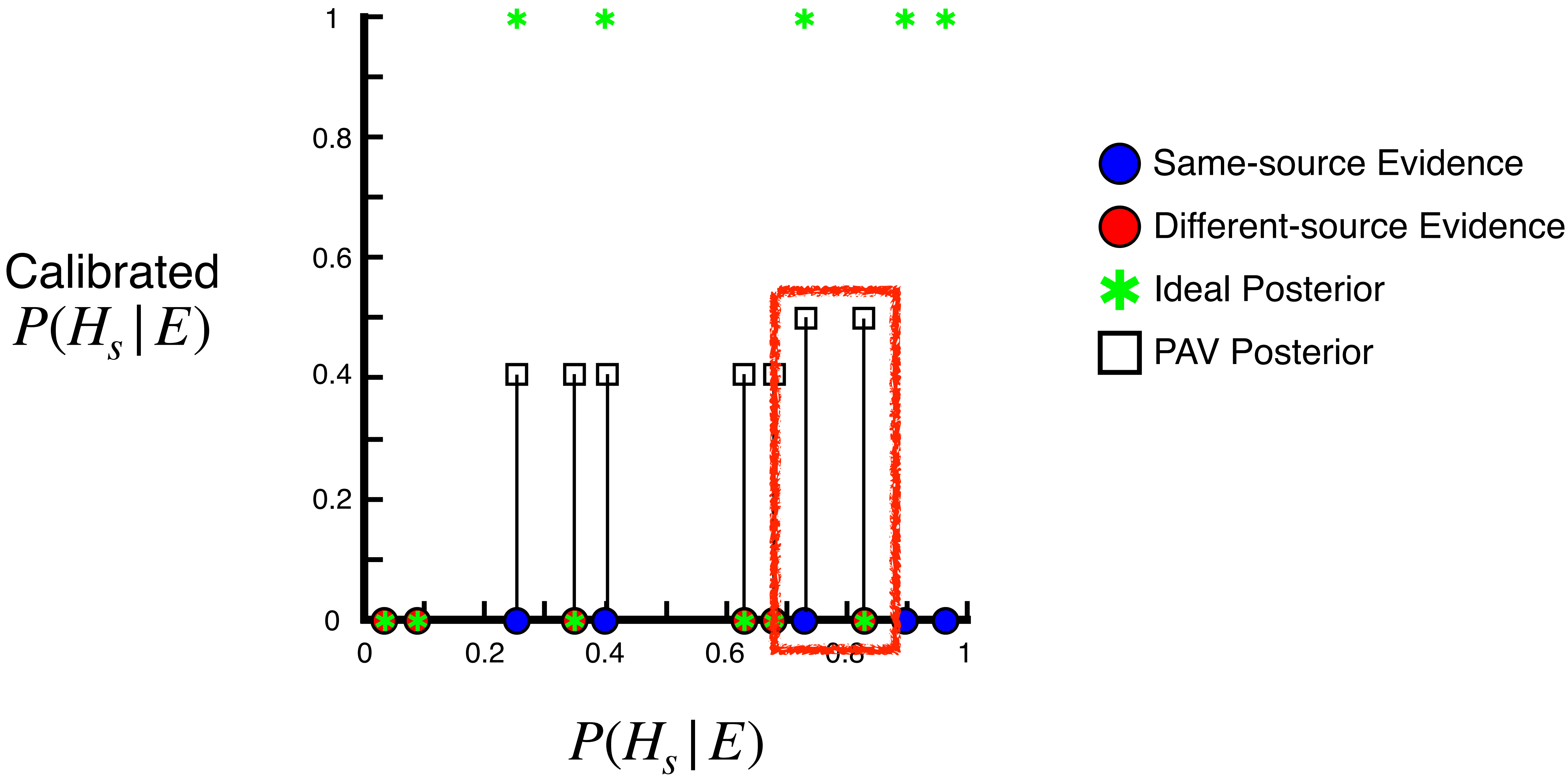
[Barlow, 1972]

Aside: Calibration via Isotonic Regression



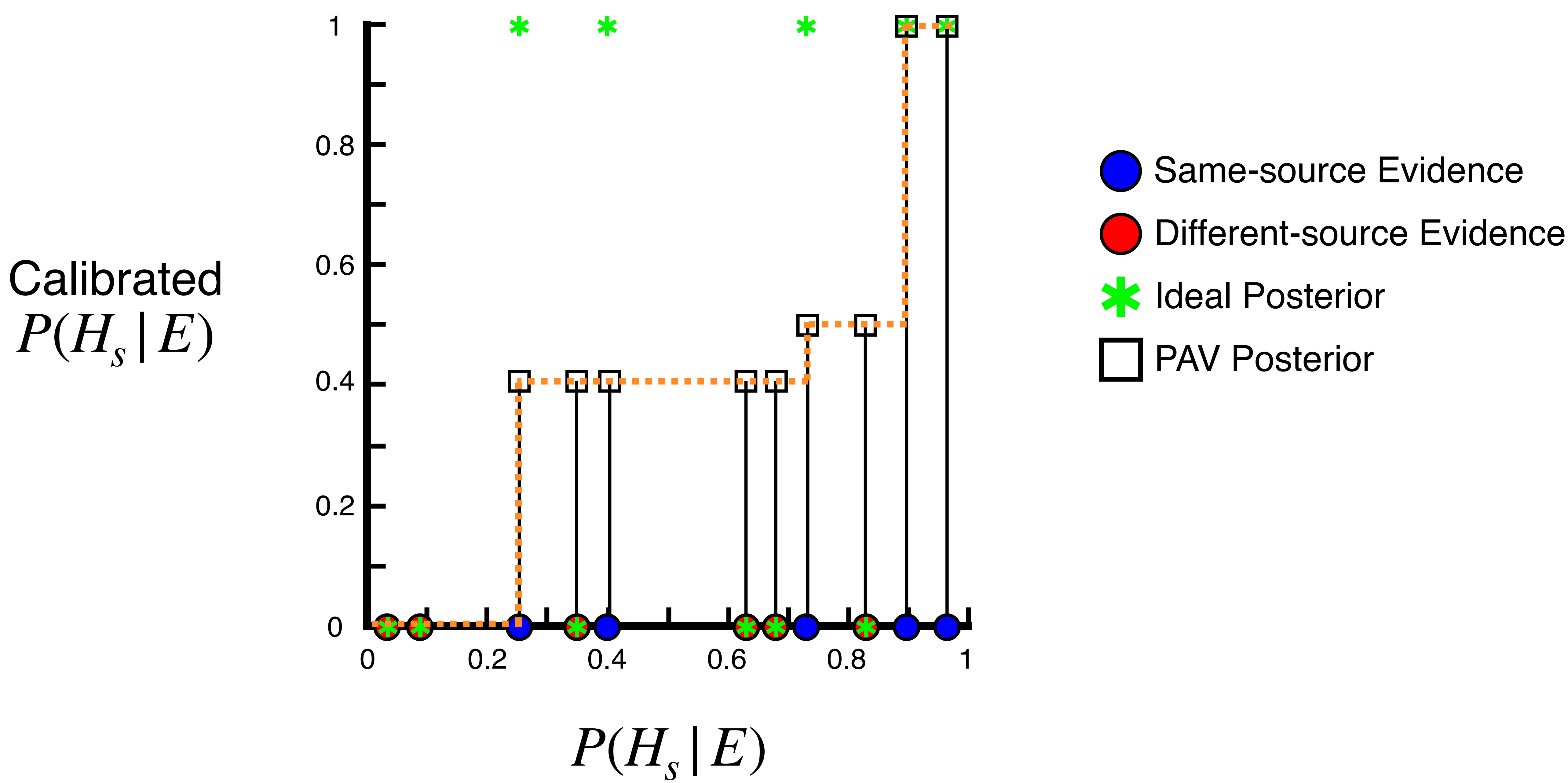
[Barlow, 1972]

Aside: Calibration via Isotonic Regression



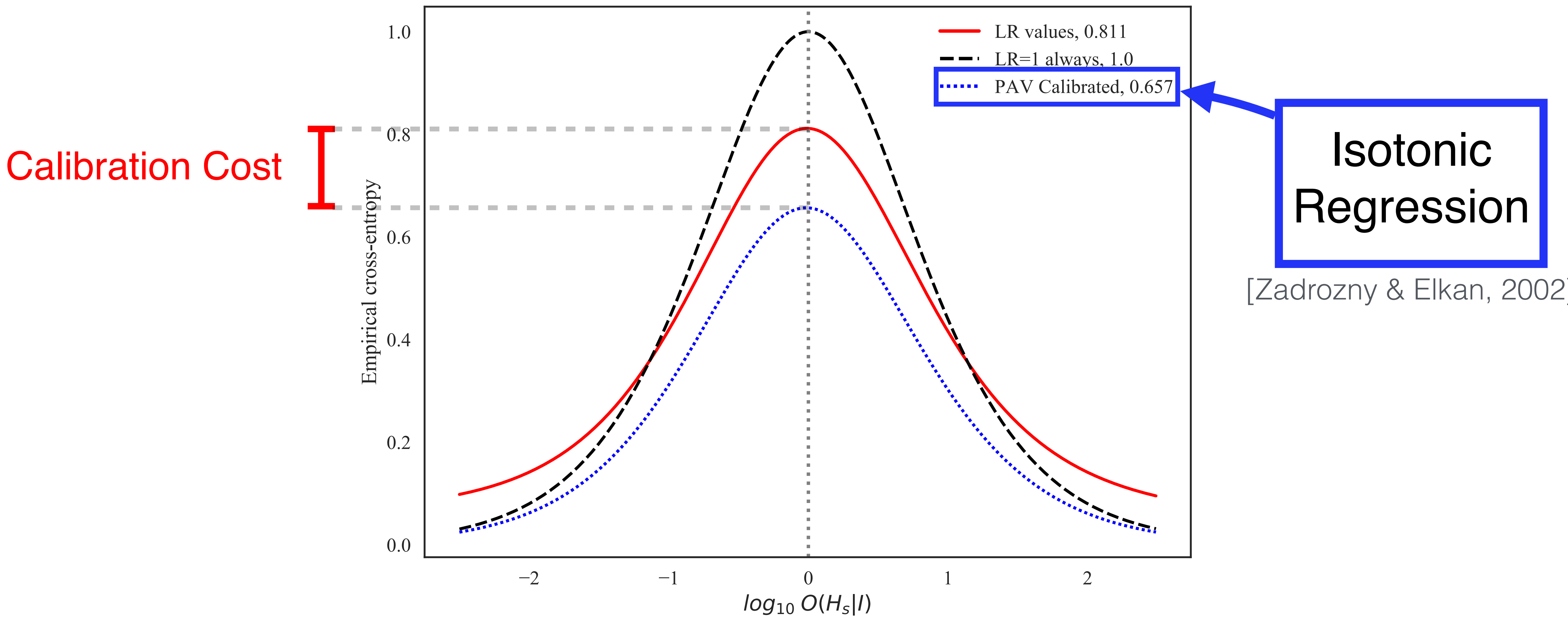
[Barlow, 1972]

Aside: Calibration via Isotonic Regression

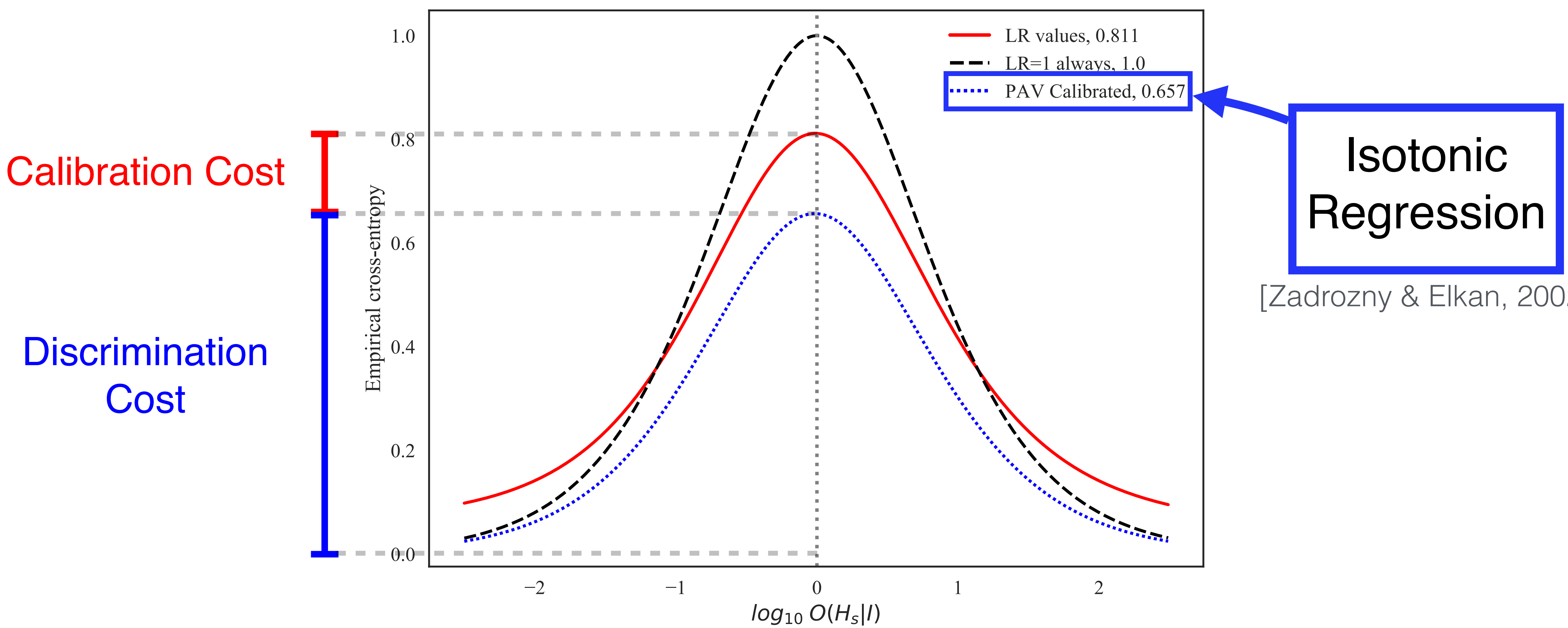


[Barlow, 1972]

ECE Plot


Calibration Cost
1.0
0.8
0.6
0.4
0.2
0.0
Empirical cross-entropy
LR values, 0.811
LR=1 always, 1.0
PAV Calibrated, 0.657
−2
−1
0
1
2
$log_{10}\, O(H_s|I)$
Isotonic Regression
[Zadrozny & Elkan, 2002]

ECE Plot

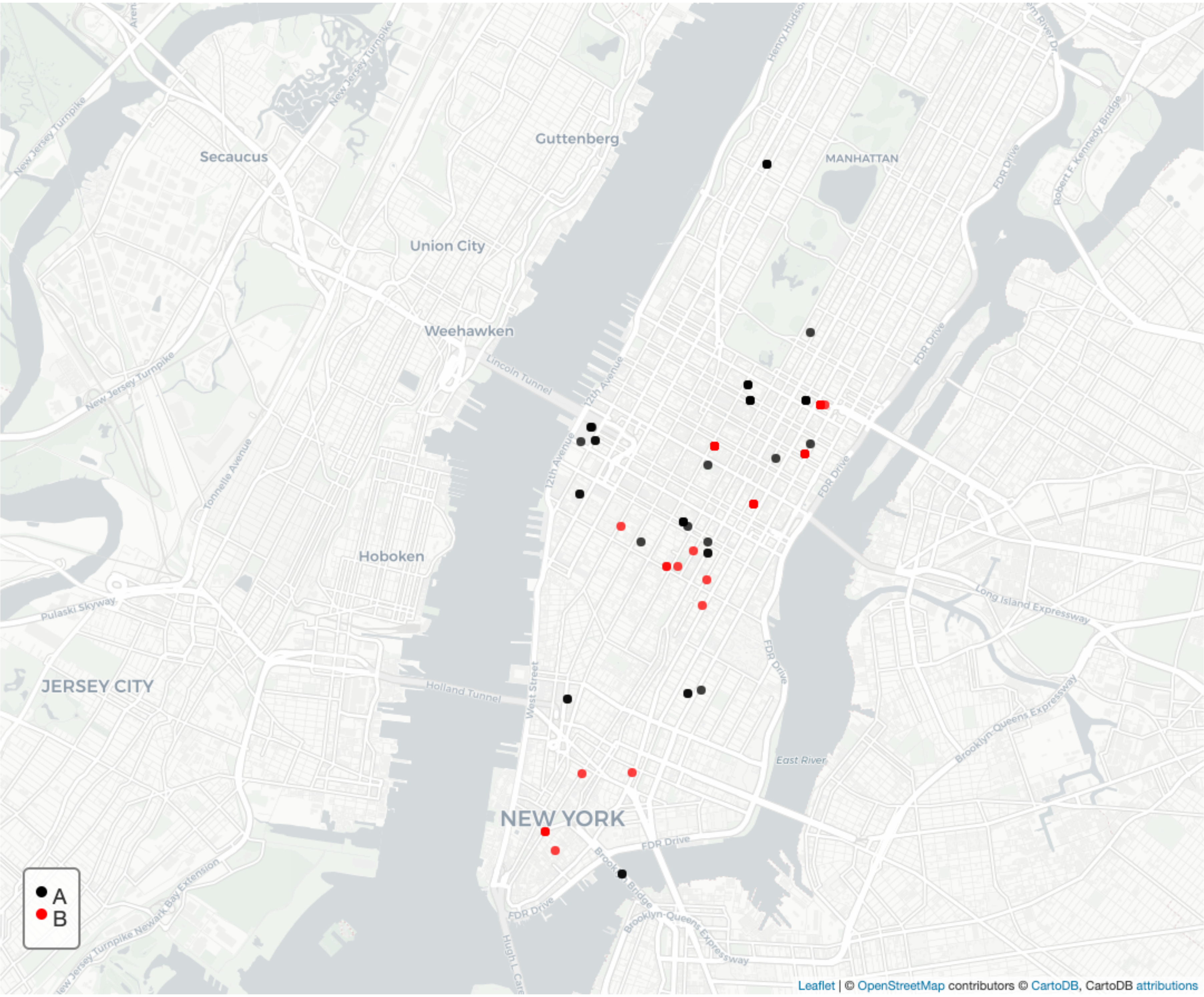


[Zadrozny & Elkan, 2002]

CONTRIBUTION

Quantifying the Strength of Geolocated Event Evidence

CHAPTER #4 [Galbraith, Smyth & Stern, Digital Investigation 2020]



Secaucus
Guttenberg
MANHATTAN
Union City
Weehawken
Hoboken
JERSEY CITY
NEW YORK
East River
A
B
Leaflet | © OpenStreetMap contributors © CartoDB, CartoDB attributions

Revisiting the LR

$$LR = \frac{Pr(A, B \mid H_s)}{Pr(A, B \mid H_d)} = \frac{Pr(B \mid A, H_s)}{Pr(B \mid A, H_d)} \cdot \frac{Pr(A \mid H_s)}{Pr(A \mid H_d)}$$

Revisiting the LR

$$Pr(A \mid H_s) = Pr(A \mid H_d) = Pr(A)$$

$$LR = \frac{Pr(A, B \mid H_s)}{Pr(A, B \mid H_d)} = \frac{Pr(B \mid A, H_s)}{Pr(B \mid A, H_d)} \boxed{\frac{Pr(A \mid H_s)}{Pr(A \mid H_d)} = 1}$$

[Stern, 2017]

Revisiting the LR

$$Pr(A \mid H_s) = Pr(A \mid H_d) = Pr(A)$$

$$LR = \frac{Pr(A, B \mid H_s)}{Pr(A, B \mid H_d)} = \frac{Pr(B \mid A, H_s)}{Pr(B \mid A, H_d)} \cdot \frac{Pr(A \mid H_s)}{Pr(A \mid H_d)}$$

$$\frac{Pr(A \mid H_s)}{Pr(A \mid H_d)} = 1$$

$$Pr(B \mid A, H_d) = Pr(B \mid H_d)$$

[Stern, 2017]

Revisiting the LR

$$Pr(A \mid H_s) = Pr(A \mid H_d) = Pr(A)$$

$$LR = \frac{Pr(A, B \mid H_s)}{Pr(A, B \mid H_d)} = \frac{Pr(B \mid A, H_s)}{Pr(B \mid A, H_d)} \cdot \frac{Pr(A \mid H_s)}{Pr(A \mid H_d)}$$

1

$$Pr(B \mid A, H_d) = Pr(B \mid H_d)$$

$$\Rightarrow LR = \frac{f(B \mid A, H_s)}{f(B \mid H_d)}$$

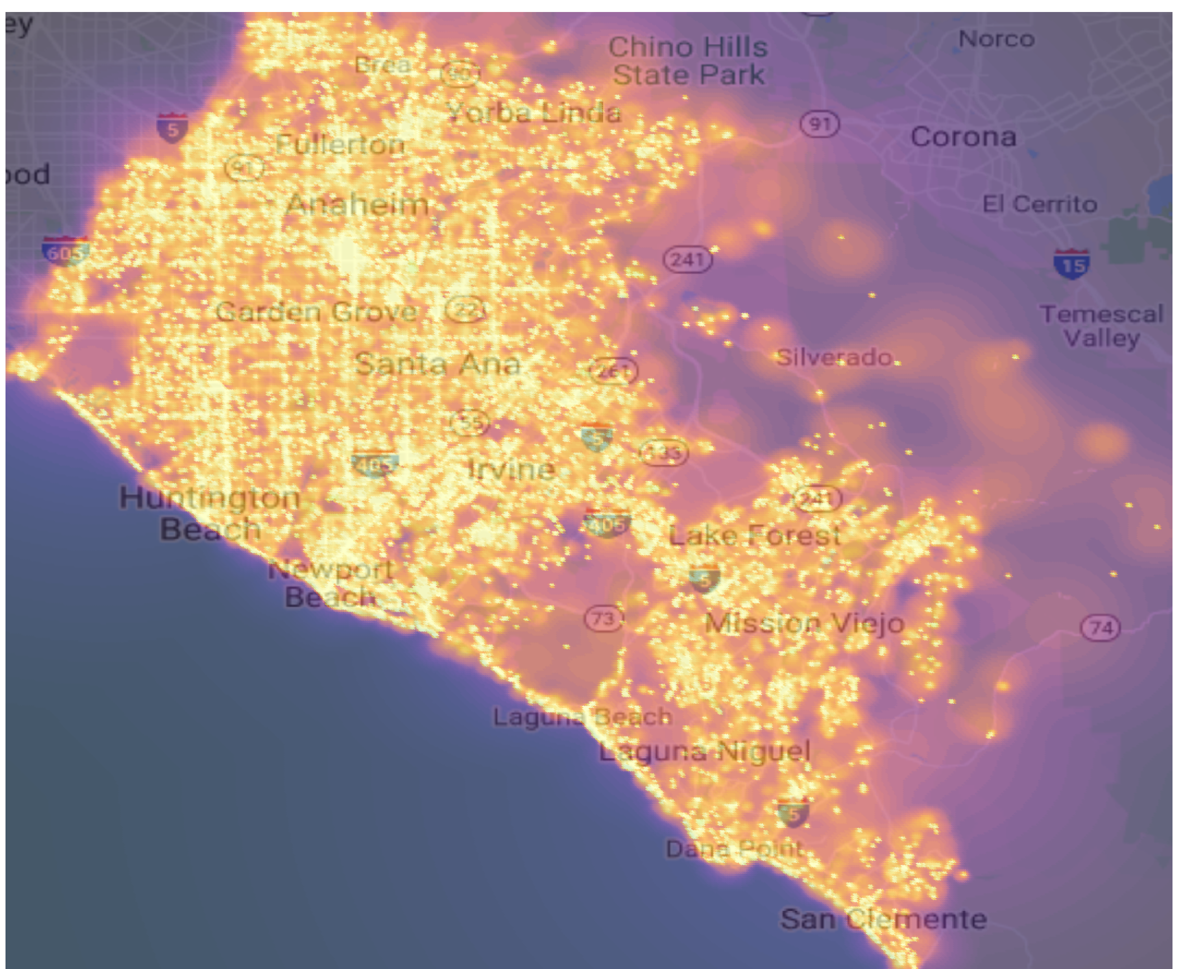
[Stern, 2017]

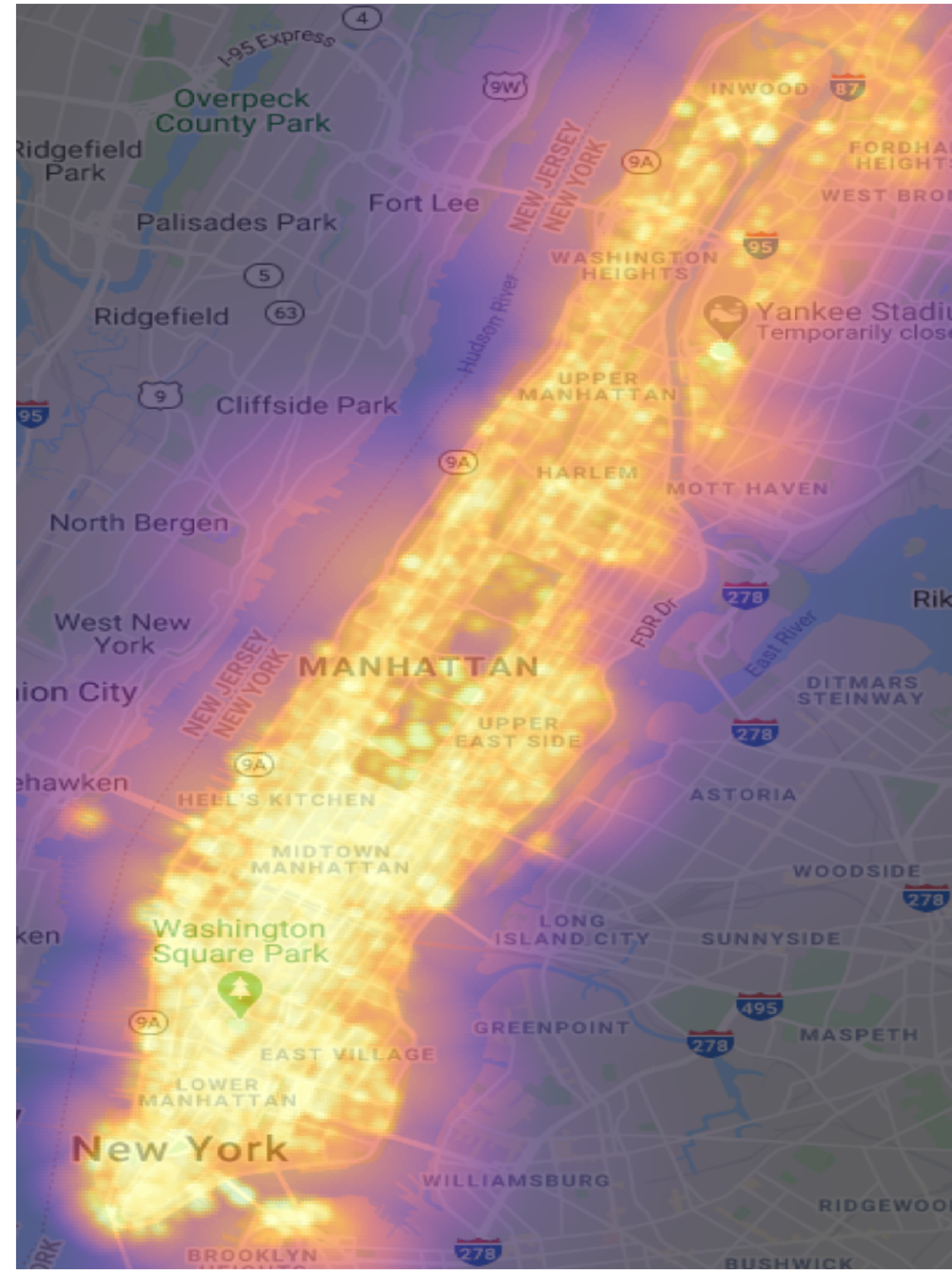
$$LR = \frac{f(B \mid A, H_s)}{f(B \mid H_d)}$$

$$\hat{f}(B \mid H_d) = \prod_{j=1}^{n_b} f_{KD}(s_j^b \mid \mathscr{D})$$

$$LR = \frac{f(B \mid A, H_s)}{f(B \mid H_d)}$$

$$\hat{f}(B \mid H_d) = \prod_{j=1}^{n_b} f_{KD}(s_j^b \mid \mathscr{D})$$





Adaptive Bandwidth
Kernel Density Estimators
[Breiman et al., 1977]

$$LR = \frac{f(B \mid A, H_s)}{f(B \mid H_d)}$$

$$\hat{f}(B \mid A, H_s) = \prod_{j=1}^{n_b} f_{MKD}(s_j^b \mid A, \mathscr{D}, \alpha)$$

[Lichman & Smyth, 2014]

$$\hat{f}(B \mid H_d) = \prod_{j=1}^{n_b} f_{KD}(s_j^b \mid \mathscr{D})$$

$$LR = \frac{f(B \mid A, H_s)}{f(B \mid H_d)}$$

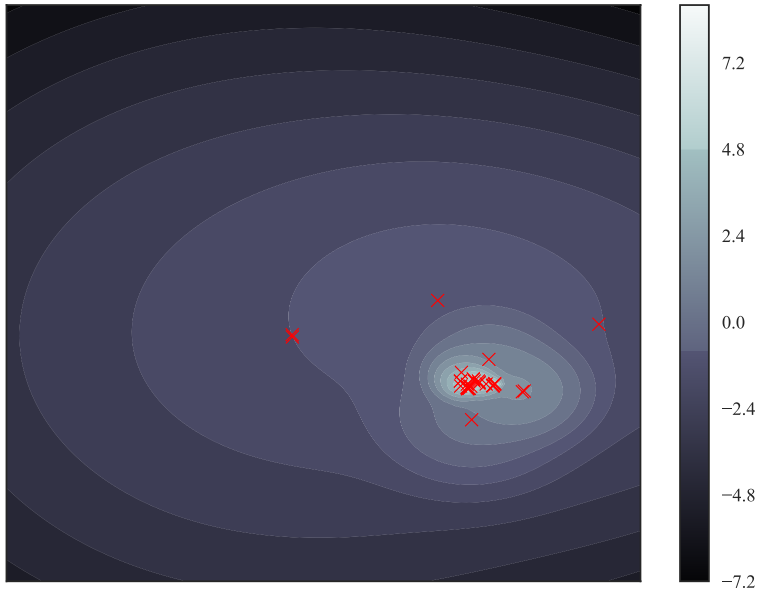
$$\hat{f}(B \mid A, H_s) = \prod_{j=1}^{n_b} f_{MKD}(s_j^b \mid A, \mathscr{D}, \alpha)$$

$$f_{MKD}(s_j^b \mid A, \mathscr{D}, \alpha) = \alpha f_{KD}(s_j^b \mid A)$$

Individual Component

[Lichman & Smyth, 2014]

$$\hat{f}(B \mid H_d) = \prod_{j=1}^{n_b} f_{KD}(s_j^b \mid \mathscr{D})$$



$$LR = \frac{f(B \mid A, H_s)}{f(B \mid H_d)}$$

$$\hat{f}(B \mid A, H_s) = \prod_{j=1}^{n_b} f_{MKD}(s_j^b \mid A, \mathcal{D}, \alpha)$$

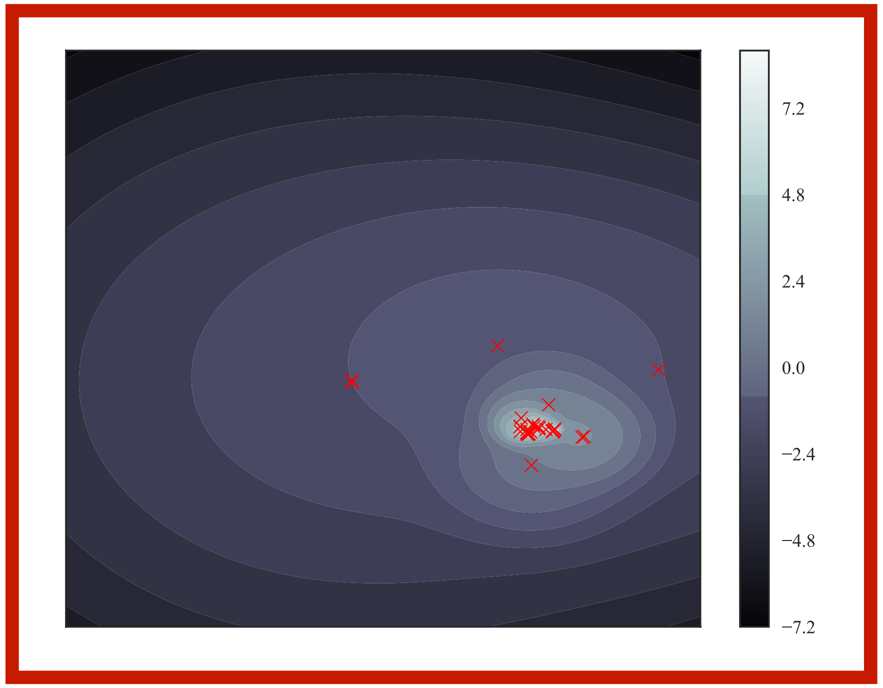
$$f_{MKD}(s_j^b \mid A, \mathcal{D}, \alpha) = \alpha f_{KD}(s_j^b \mid A) + (1-\alpha) f_{KD}(s_j^b \mid \mathcal{D})$$

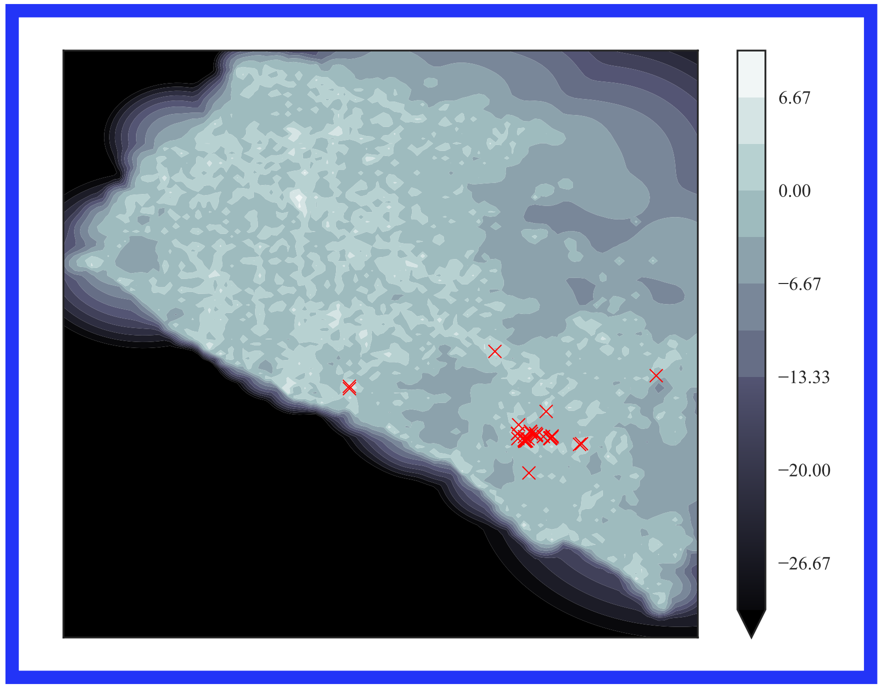
Individual Component

Population Component

[Lichman & Smyth, 2014]

$$\hat{f}(B \mid H_d) = \prod_{j=1}^{n_b} f_{KD}(s_j^b \mid \mathcal{D})$$





$$LR = \frac{f(B \mid A, H_s)}{f(B \mid H_d)}$$

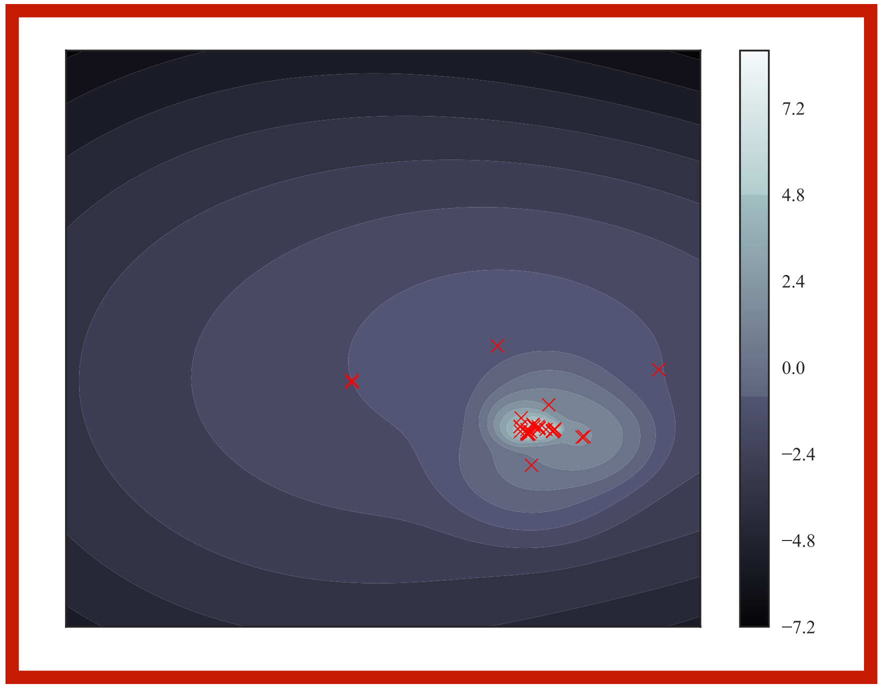
$$\hat{f}(B \mid A, H_s) = \prod_{j=1}^{n_b} f_{MKD}(s_j^b \mid A, \mathscr{D}, \alpha)$$

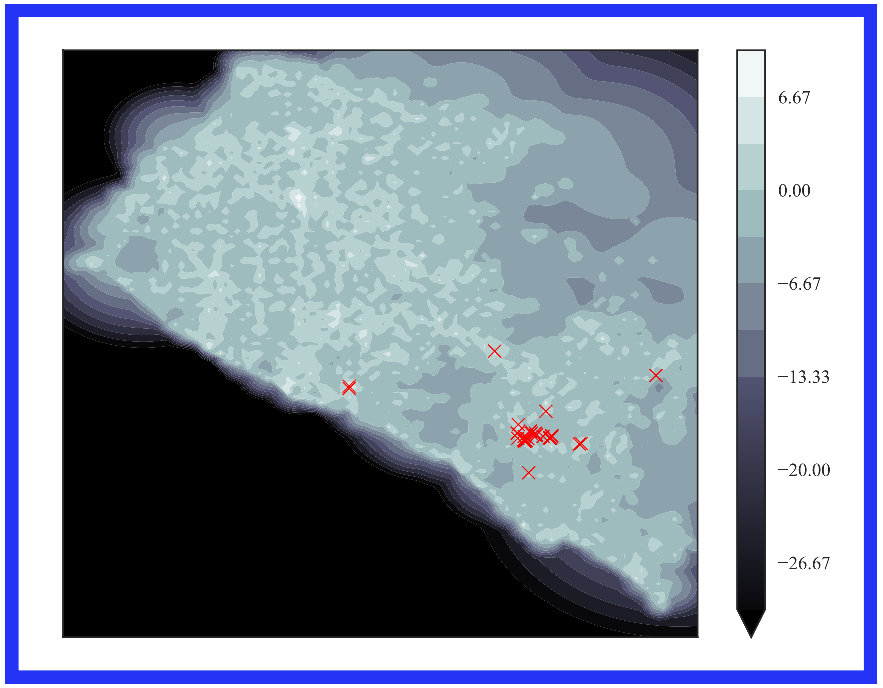
$$f_{MKD}(s_j^b \mid A, \mathscr{D}, \alpha) = \alpha f_{KD}(s_j^b \mid A) + (1-\alpha) f_{KD}(s_j^b \mid \mathscr{D})$$

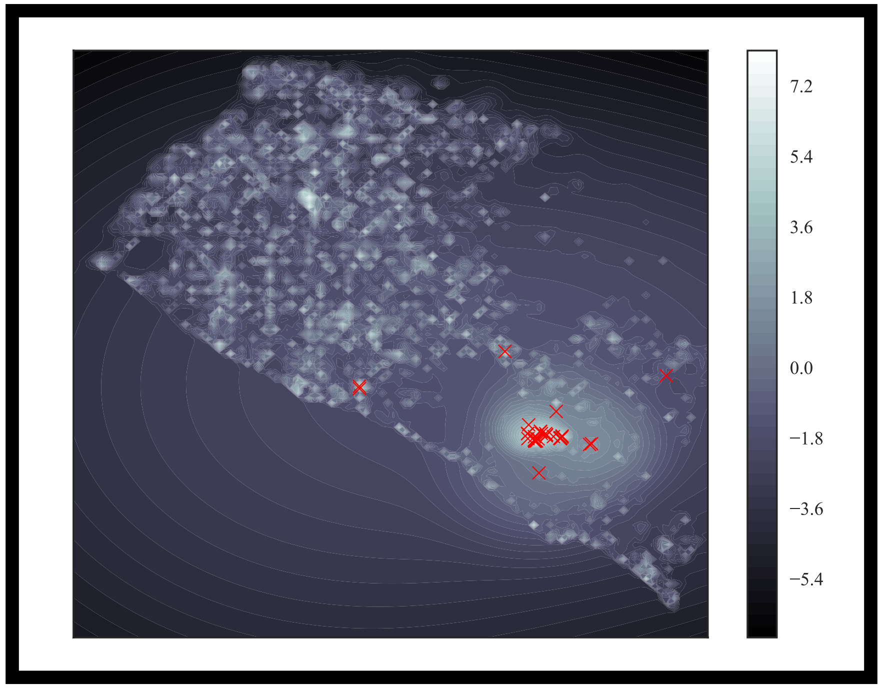
Mixing Weight — Individual Component — Population Component

[Lichman & Smyth, 2014]

$$\hat{f}(B \mid H_d) = \prod_{j=1}^{n_b} f_{KD}(s_j^b \mid \mathscr{D})$$







$\alpha = 0.8$

$$LR = \frac{f(B \mid A, H_s)}{f(B \mid H_d)}$$

$$\hat{f}(B \mid A, H_s) = \prod_{j=1}^{n_b} f_{MKD}(s_j^b \mid A, \mathscr{D}, \alpha)$$

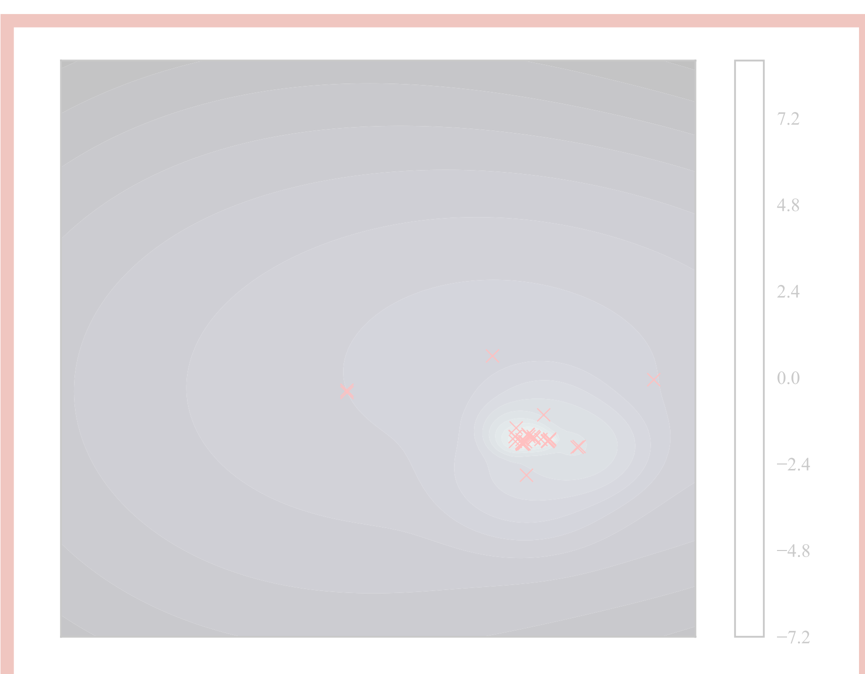
$$f_{MKD}(s_j^b \mid A, \mathscr{D}, \alpha) = \alpha f_{KD}(s_j^b \mid A) + (1-\alpha) f_{KD}(s_j^b \mid \mathscr{D})$$

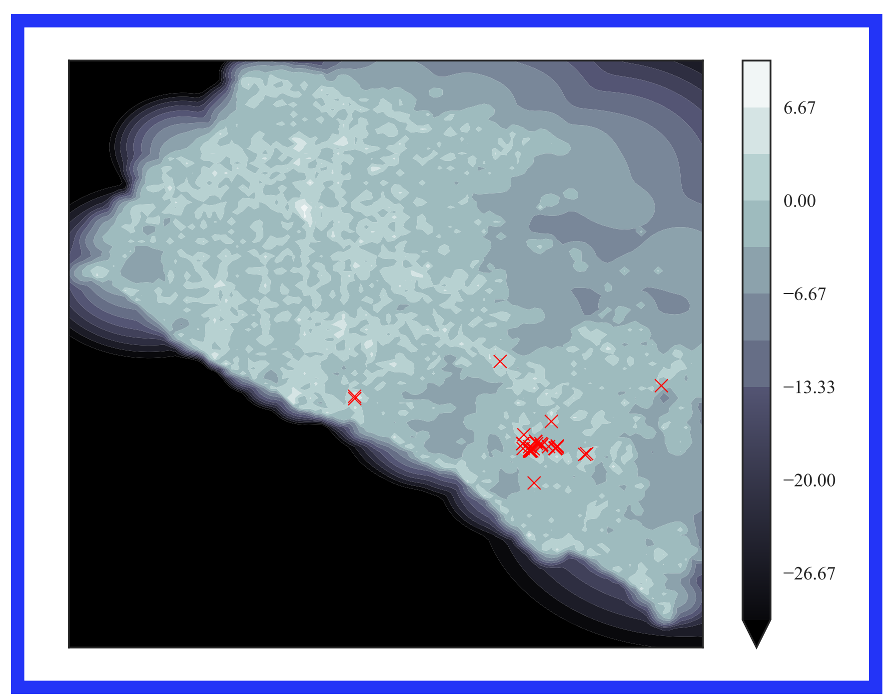
Mixing Weight | Individual Component | Population Component

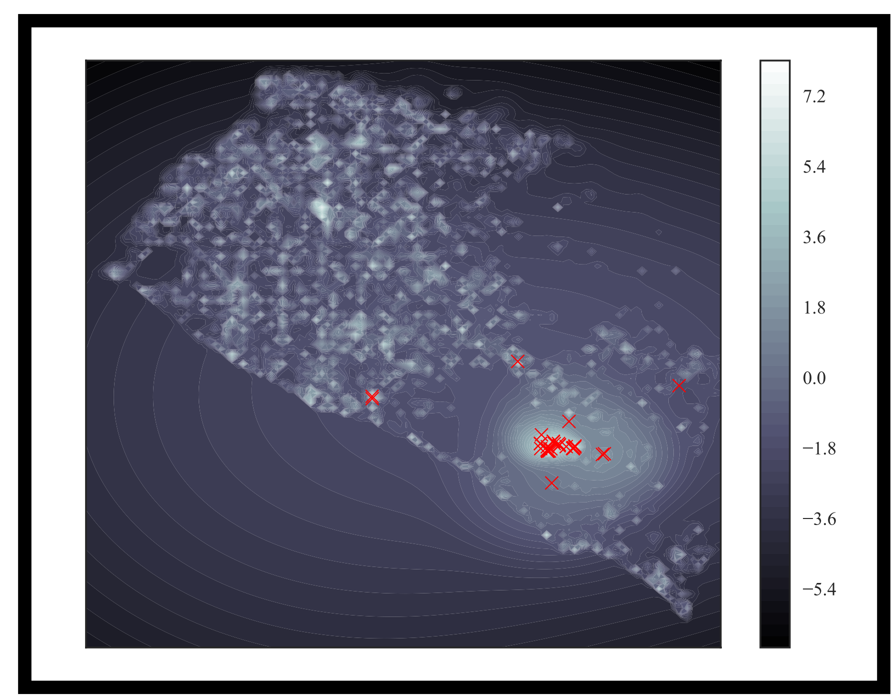
[Lichman & Smyth, 2014]

$$\hat{f}(B \mid H_d) = \prod_{j=1}^{n_b} f_{KD}(s_j^b \mid \mathscr{D})$$

$$\widehat{LR} = \frac{\hat{f}(B \mid A, H_s)}{\hat{f}(B \mid H_d)}$$







$\alpha = 0.8$

What about score-based approaches?

Score Functions

- Techniques to characterize spatial point patterns generally fall into two categories [Haggett, 1977]

Score Functions

- Techniques to characterize spatial point patterns generally fall into two categories [Haggett, 1977]
  - → Distance-based
  - → Area-based

Score Functions

- Techniques to characterize spatial point patterns generally fall into two categories [Haggett, 1977]
  - → Distance-based
  - → Area-based

- Use distance-based score functions $\Delta(A, B)$ to quantify the similarity of the points within the sets A and B

Score Functions

- Techniques to characterize spatial point patterns generally fall into two categories [Haggett, 1977]
  - Distance-based
  - Area-based

- Use distance-based score functions $\Delta(A, B)$ to quantify the similarity of the points within the sets A and B
  - Average nearest-neighbor distance $\overline{D}_{min}(B, A \mid \Omega^b)$
  - Earth-mover's distance $EMD(B, A \mid \Omega^b, \Omega^a)$

Score Functions

- Techniques to characterize spatial point patterns generally fall into two categories [Haggett, 1977]
  - → Distance-based
  - → Area-based

- Use distance-based score functions $\Delta(A, B)$ to quantify the similarity of the points within the sets A and B
  - → Average nearest-neighbor distance $\overline{D}_{min}(B, A \mid \Omega^b)$
  - → Earth-mover's distance $EMD(B, A \mid \Omega^b, \Omega^a)$

- Incorporate area-based information via weights Ω^a, Ω^b

Score Functions

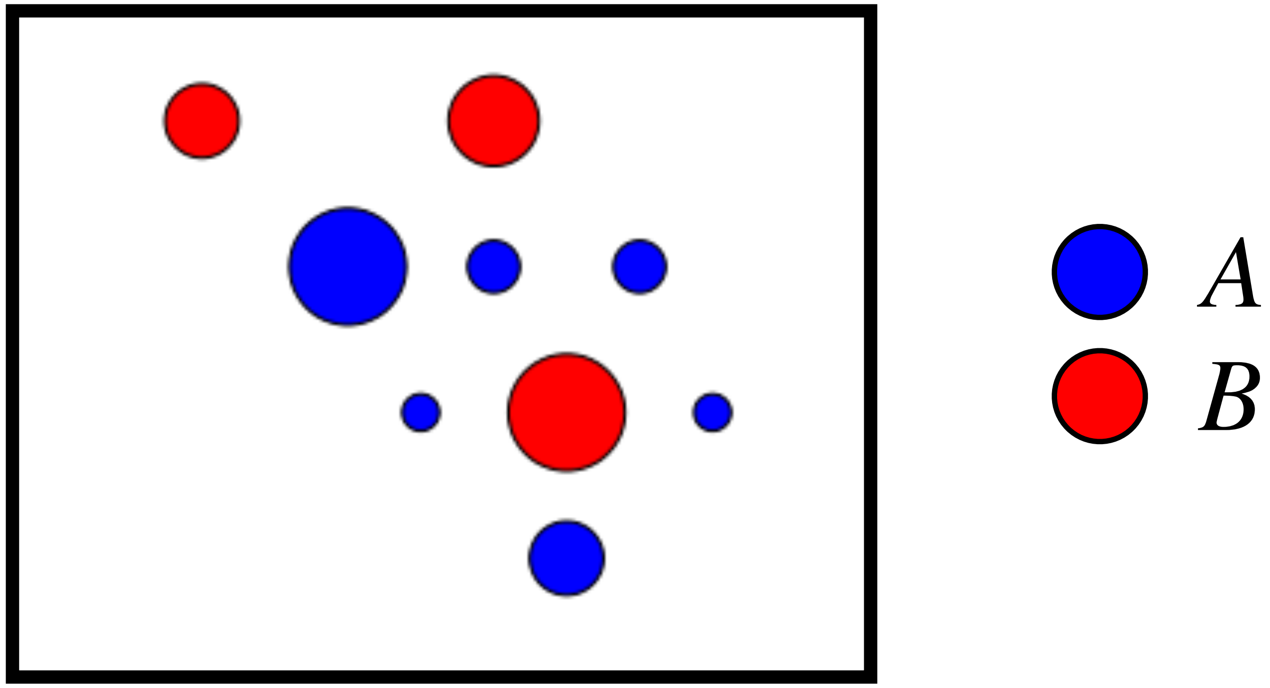
- Techniques to characterize spatial point patterns generally fall into two categories [Haggett, 1977]
  - Distance-based
  - Area-based

- Use distance-based score functions $\Delta(A, B)$ to quantify the similarity of the points within the sets A and B
  - Average nearest-neighbor distance $\overline{D}_{min}(B, A \mid \Omega^b)$
  - Earth-mover's distance $EMD(B, A \mid \Omega^b, \Omega^a)$

- Incorporate area-based information via weights Ω^a, Ω^b

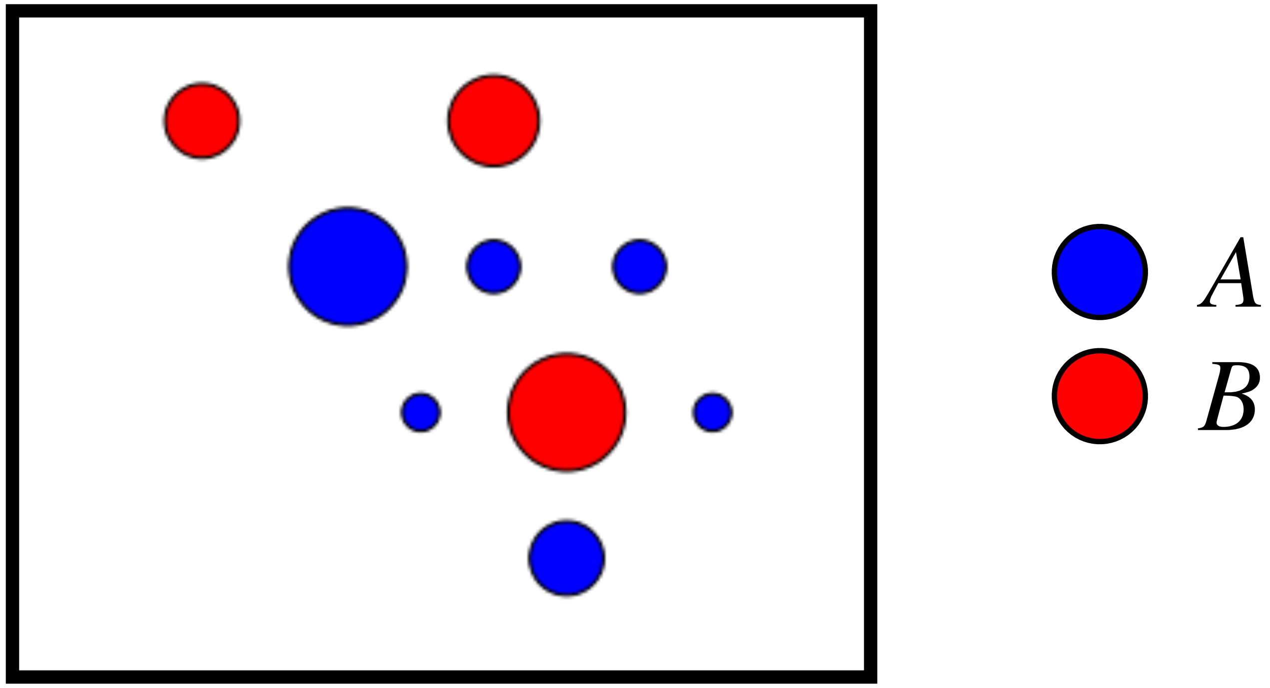
Earth-mover's distance

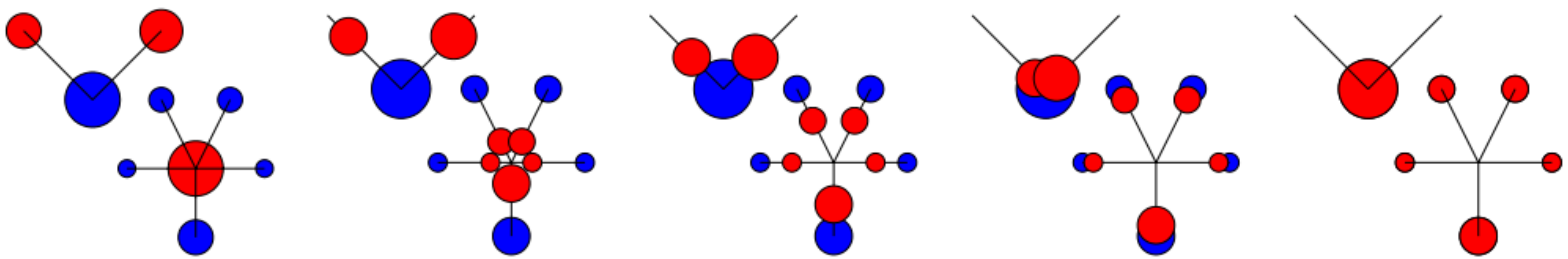
$EMD(B, A \mid \Omega^b, \Omega^a)$



[Cohen, 1999]

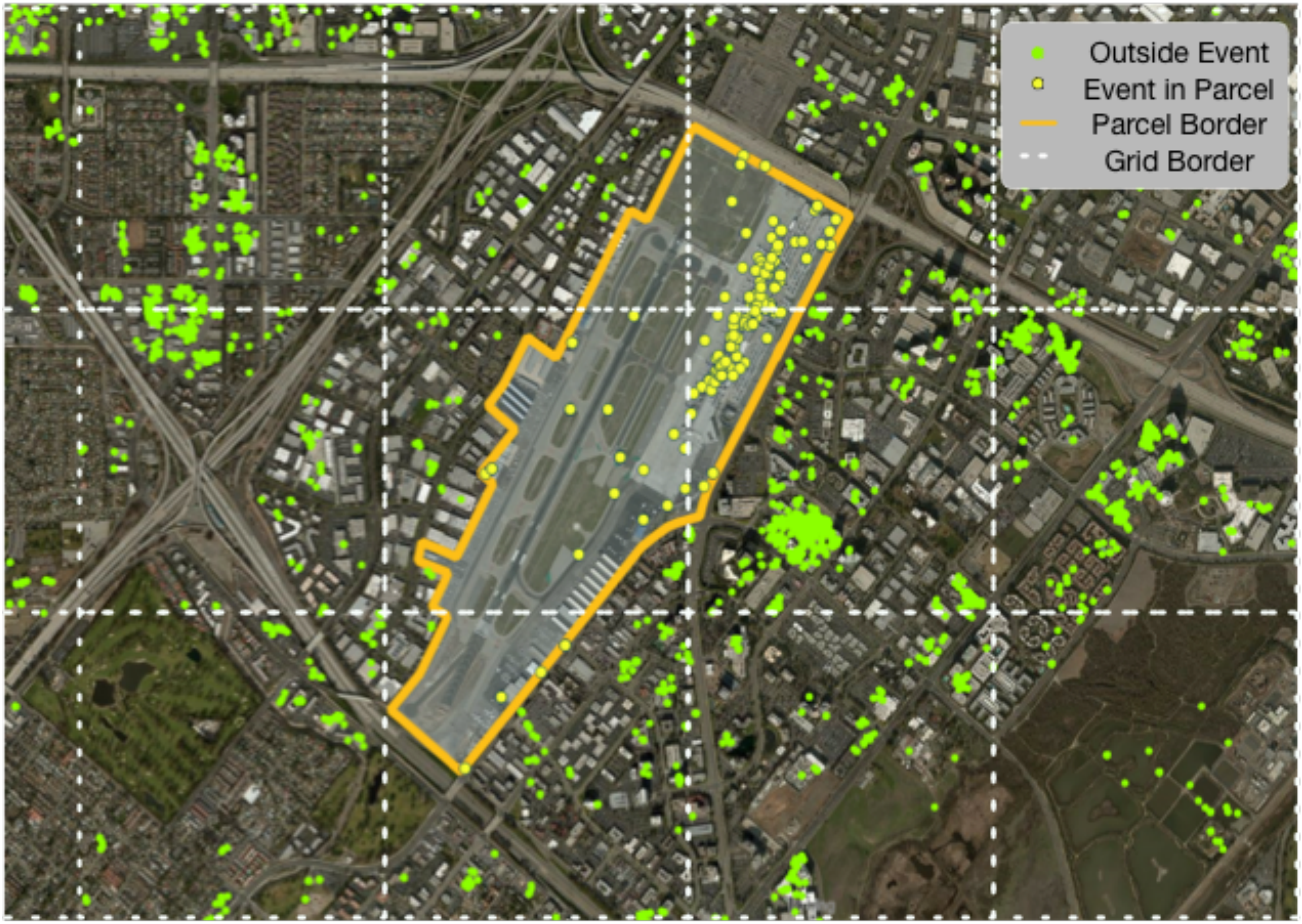
Earth-mover's distance
$EMD(B, A \mid \Omega^b, \Omega^a)$





[Cohen, 1999]

Weights Ω^a, Ω^b



[Lichman, 2017]

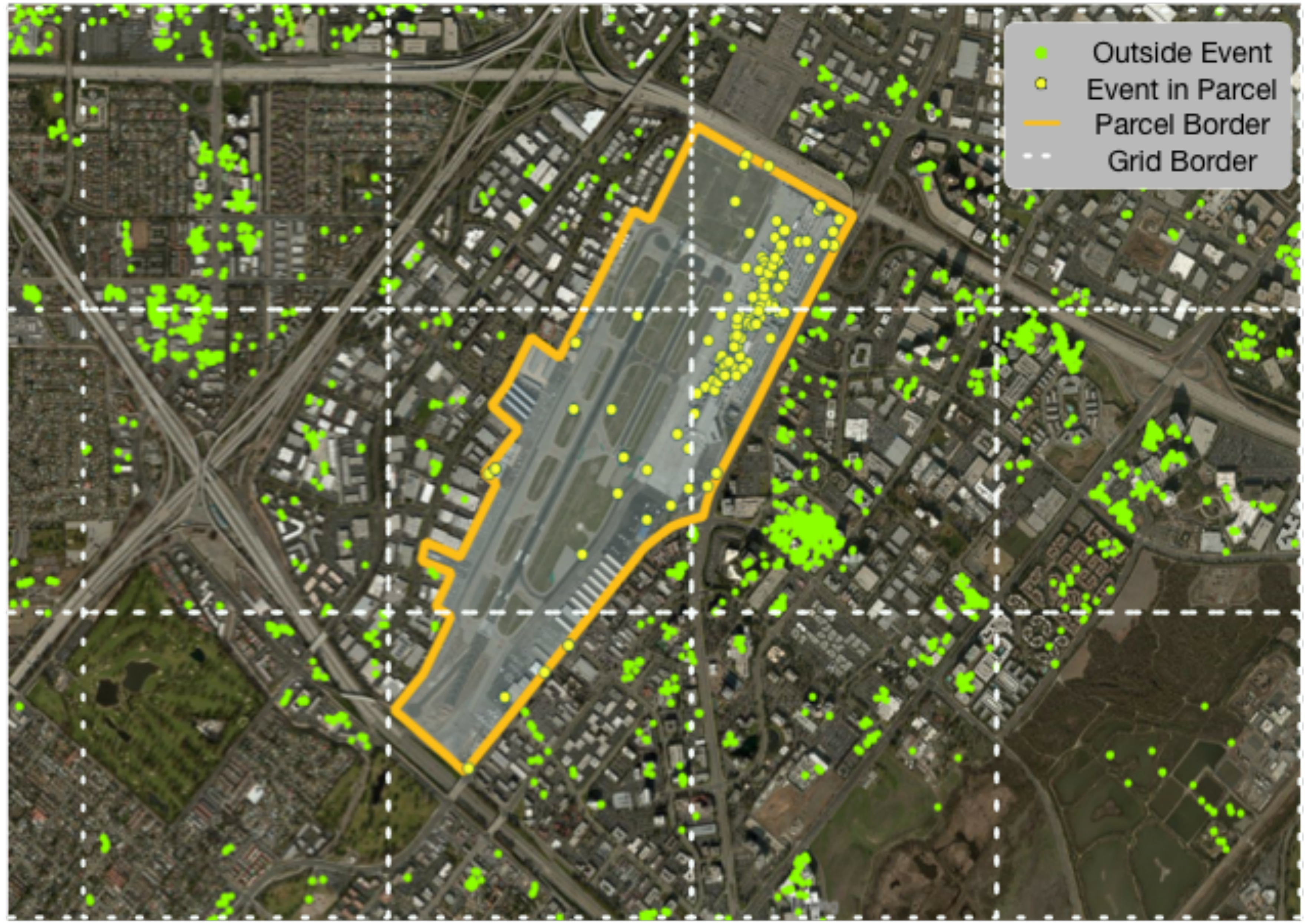
Weights Ω^a, Ω^b

- ☐ Uniform
- ☐ Visits

  $\omega_j \propto [n_{vis}(\ell(s_j))]^{-1}$

- ☐ Accounts

  $\omega_j \propto [n_{acc}(\ell(s_j))]^{-1}$



[Lichman, 2017]

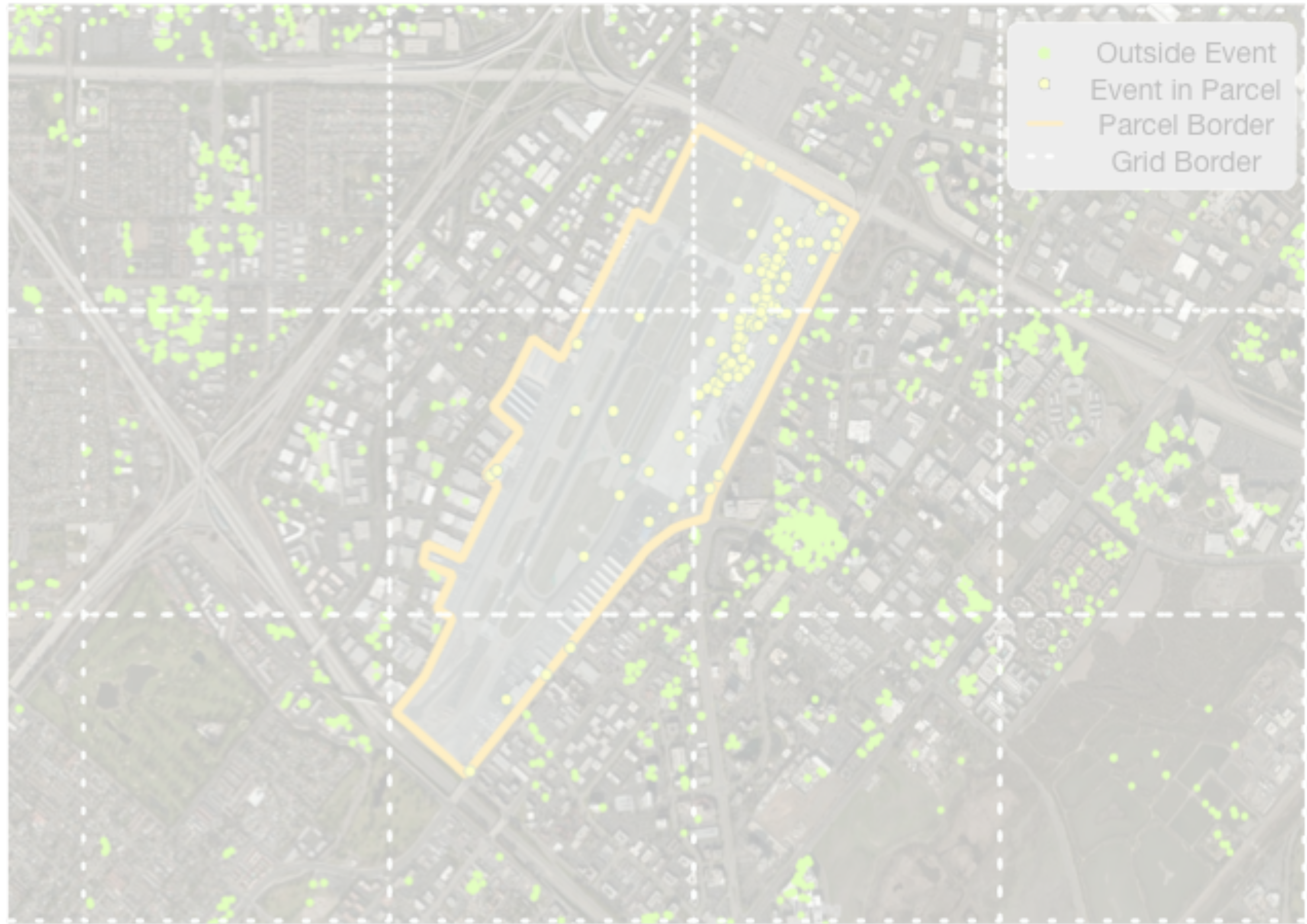
Weights Ω^a, Ω^b

- Uniform
- Visits

  $\omega_j \propto [n_{vis}(\ell(s_j))]^{-1}$

- Accounts

  $\omega_j \propto [n_{acc}(\ell(s_j))]^{-1}$



[Lichman, 2017]

Case Study

- Collected Twitter data from May 2015 to Feb 2016
  - Orange County, CA
  - Manhattan, New York, NY
- A and B are consecutive months from the same account

Region	Accounts	Visits in A	Visits in B
OC	6,714	44,310 (6.6)	38,697 (5.8)
NY	13,523	72,799 (5.4)	65,852 (4.9)

Case Study

- Collected Twitter data from May 2015 to Feb 2016
  - Orange County, CA
  - Manhattan, New York, NY

- A and B are consecutive months from the same account

Region	Accounts	Visits in A	Visits in B
OC	6,714	44,310 (6.6)	38,697 (5.8)
NY	13,523	72,799 (5.4)	65,852 (4.9)

- Results based on stratified sample based on n_a and n_b for different-source evidence

Results

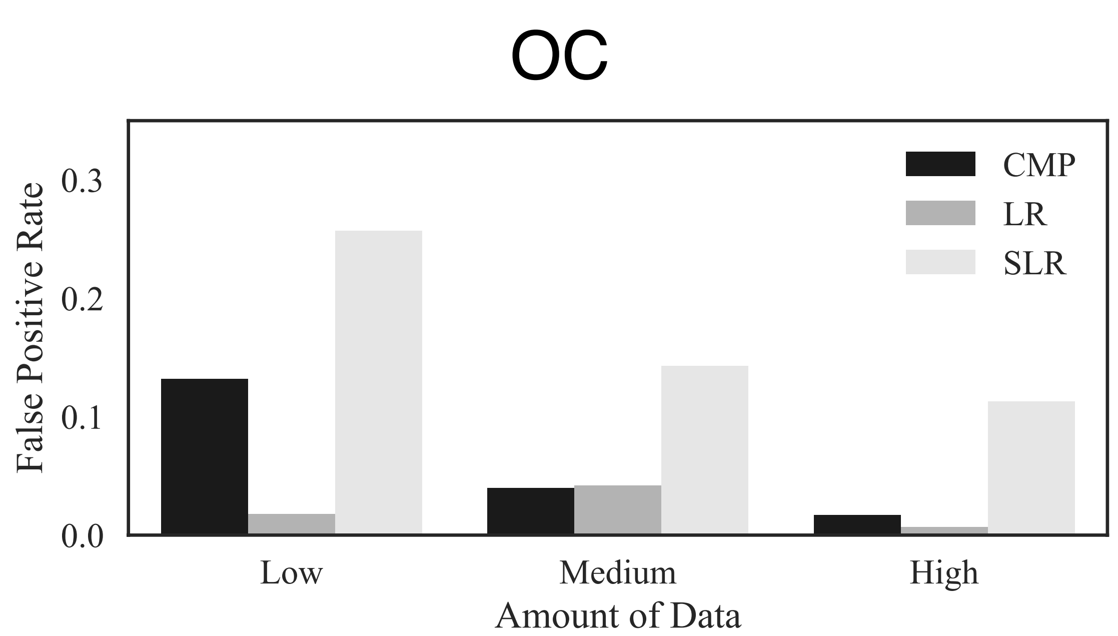
Region	Method[1]	TP Rate[2]	FP Rate[2]	AUC
OC	LR	0.380	**0.038**	**0.845**
	SLR_{EMD}	**0.614**	0.162	0.783
	CMP_{EMD}	0.448	0.208	0.784
NY	LR	0.285	**0.089**	**0.768**
	SLR_{EMD}	**0.511**	0.235	0.685
	CMP_{EMD}	0.283	0.161	0.686

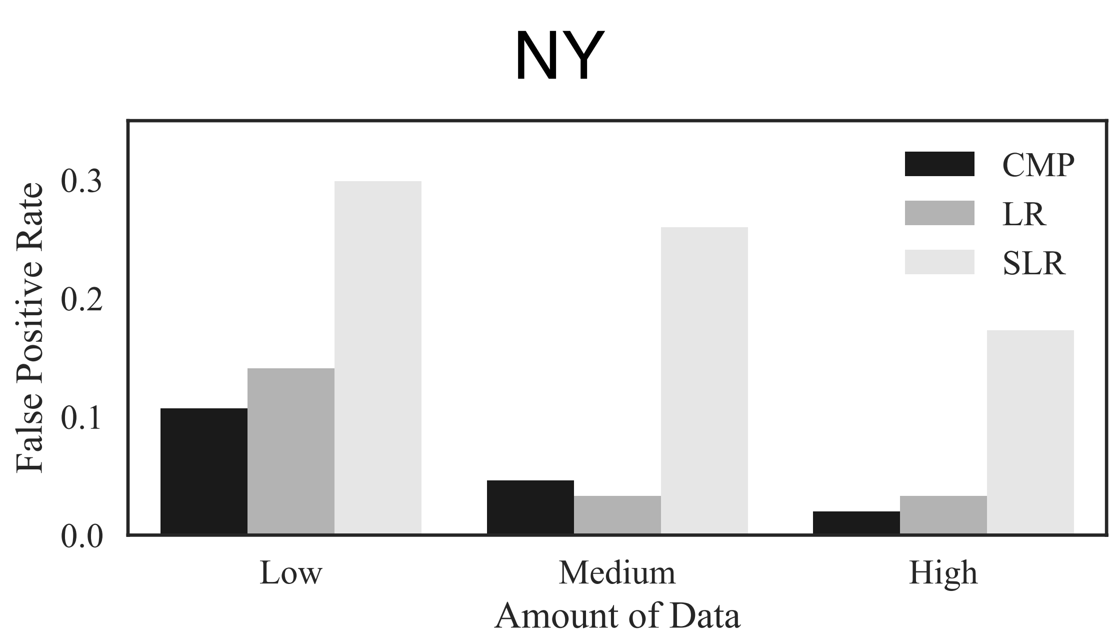
(1) LR with $\alpha(n_a)$ weights; SLR_{EMD} & CMP_{EMD} with account weights
(2) LR & SLR threshold is 1; CMP threshold is 0.05

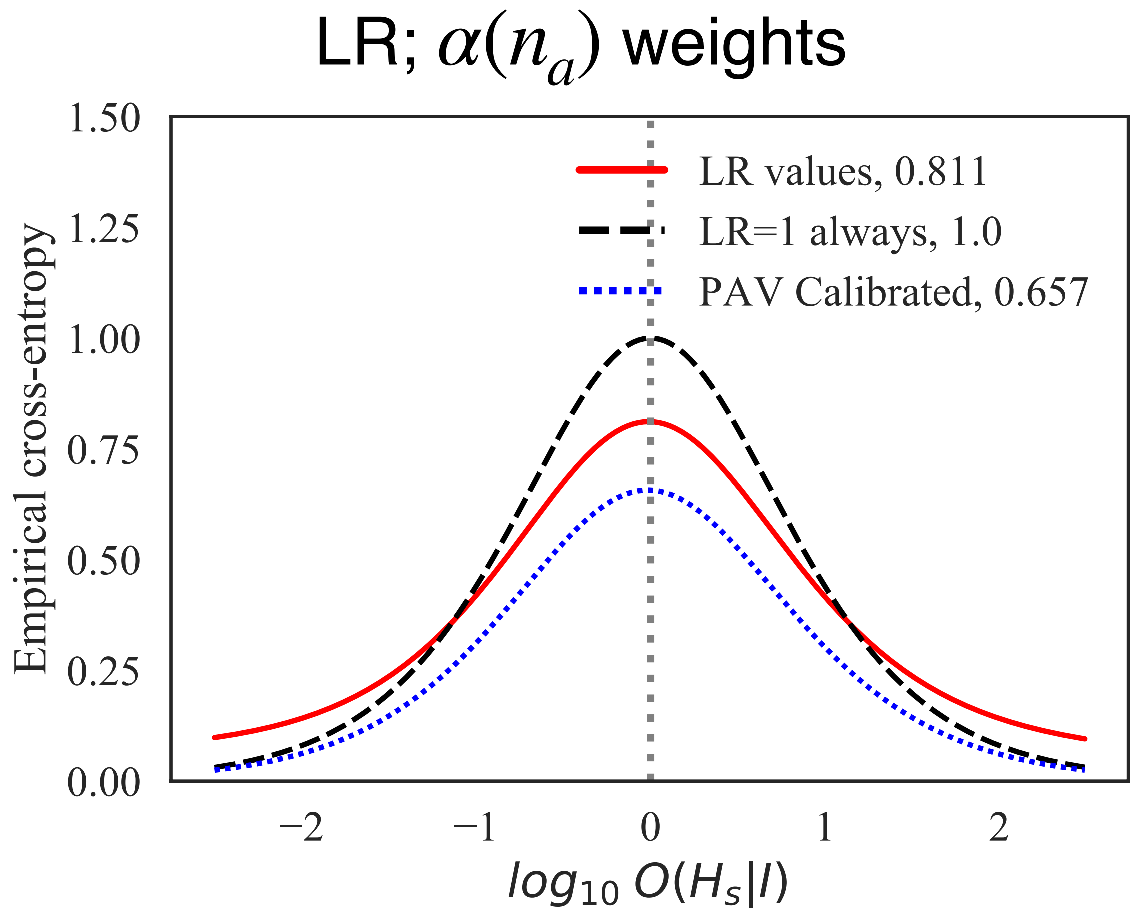
Results

Region	Method[1]	TP Rate[2]	FP Rate[2]	AUC
OC	LR	0.380	**0.038**	**0.845**
	SLR_{EMD}	**0.614**	0.162	0.783
	CMP_{EMD}	0.448	0.208	0.784
NY	LR	0.285	**0.089**	**0.768**
	SLR_{EMD}	**0.511**	0.235	0.685
	CMP_{EMD}	0.283	0.161	0.686

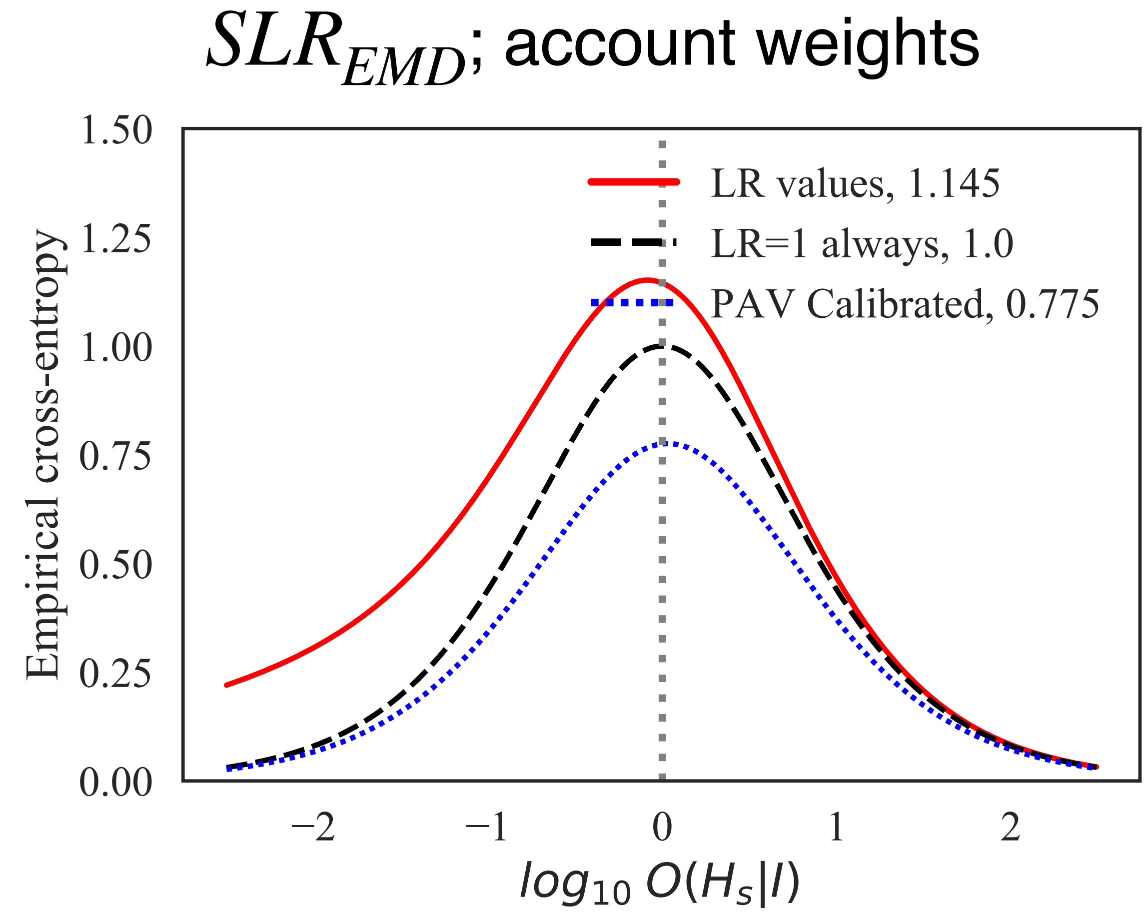
(1) LR with $\alpha(n_a)$ weights; SLR_{EMD} & CMP_{EMD} with account weights
(2) LR & SLR threshold is 1; CMP threshold is 0.05





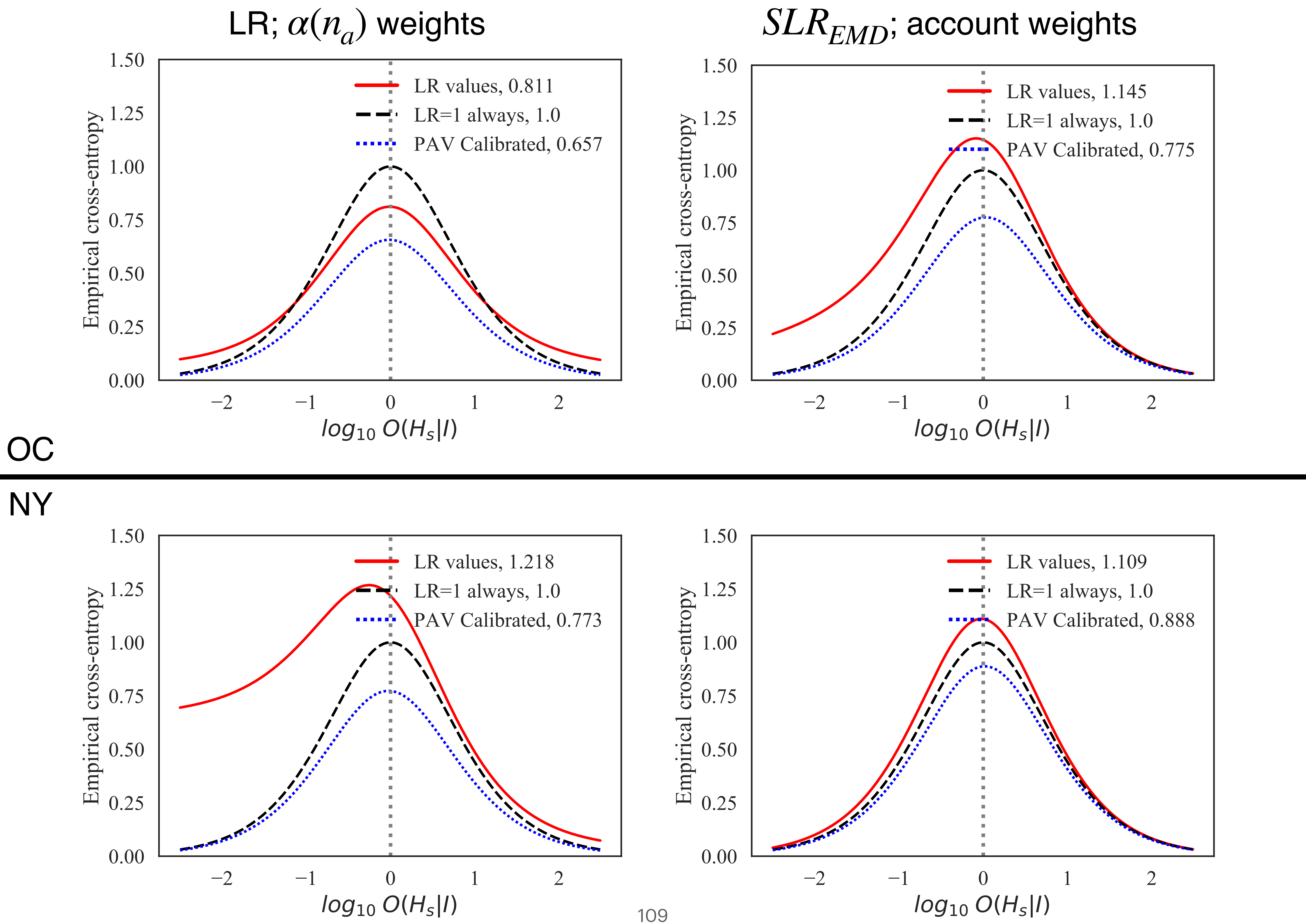


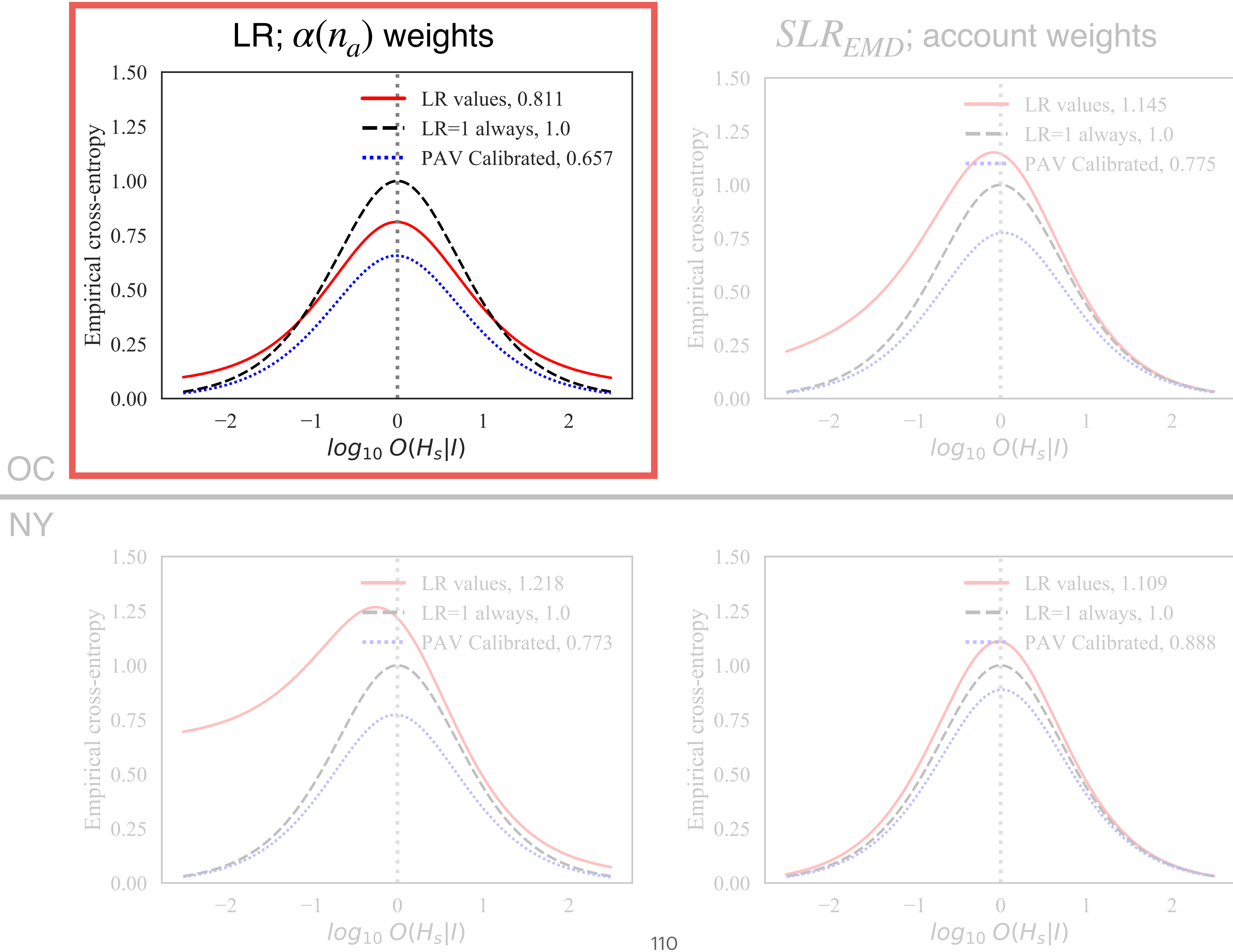
LR; $\alpha(n_a)$ weights
LR values, 0.811
LR=1 always, 1.0
PAV Calibrated, 0.657
Empirical cross-entropy
$log_{10}\ O(H_s|I)$



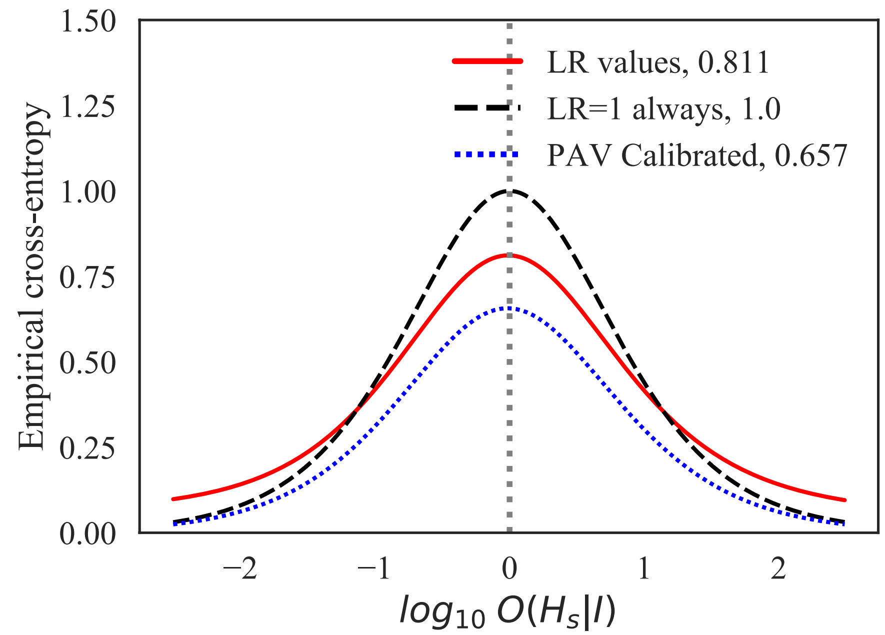
SLR_{EMD}; account weights
LR values, 1.145
LR=1 always, 1.0
PAV Calibrated, 0.775
Empirical cross-entropy
$log_{10}\ O(H_s|I)$

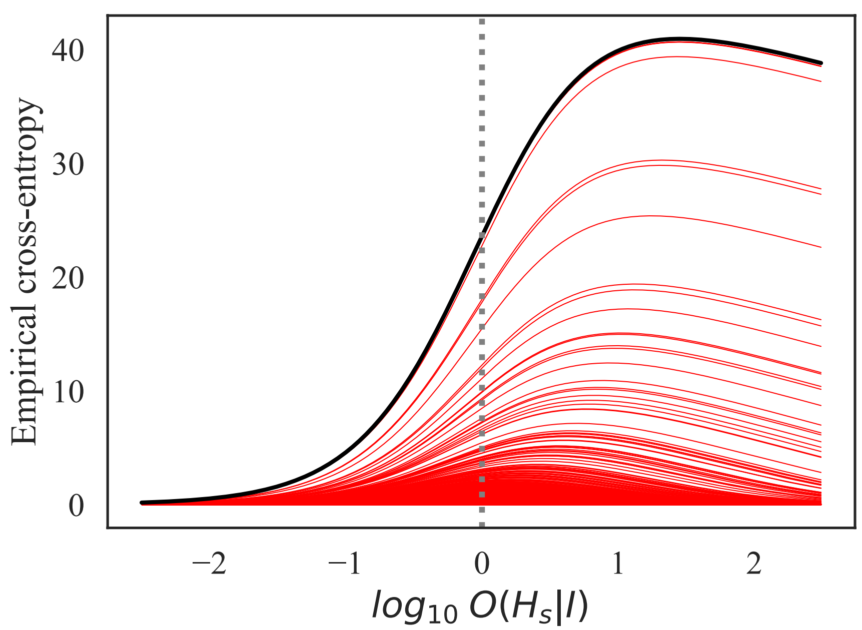
OC


LR; $\alpha(n_a)$ weights
SLR_{EMD}; account weights
OC
NY
Empirical cross-entropy
$log_{10}\ O(H_s|I)$
LR values, 0.811
LR=1 always, 1.0
PAV Calibrated, 0.657
LR values, 1.145
LR=1 always, 1.0
PAV Calibrated, 0.775
LR values, 1.218
LR=1 always, 1.0
PAV Calibrated, 0.773
LR values, 1.109
LR=1 always, 1.0
PAV Calibrated, 0.888

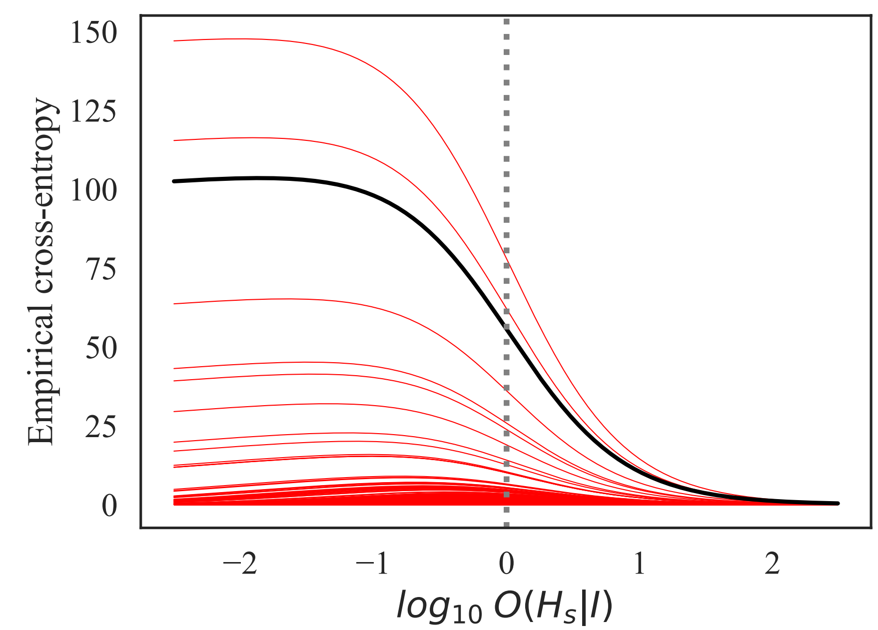

LR; $\alpha(n_a)$ weights
Empirical cross-entropy
1.50
1.25
1.00
0.75
0.50
0.25
0.00
LR values, 0.811
LR=1 always, 1.0
PAV Calibrated, 0.657
−2
−1
0
1
2
$log_{10}\, O(H_s|I)$
SLR_{EMD}; account weights
LR values, 1.145
LR=1 always, 1.0
PAV Calibrated, 0.775
OC
NY
LR values, 1.218
LR=1 always, 1.0
PAV Calibrated, 0.773
LR values, 1.109
LR=1 always, 1.0
PAV Calibrated, 0.888

Error Analysis



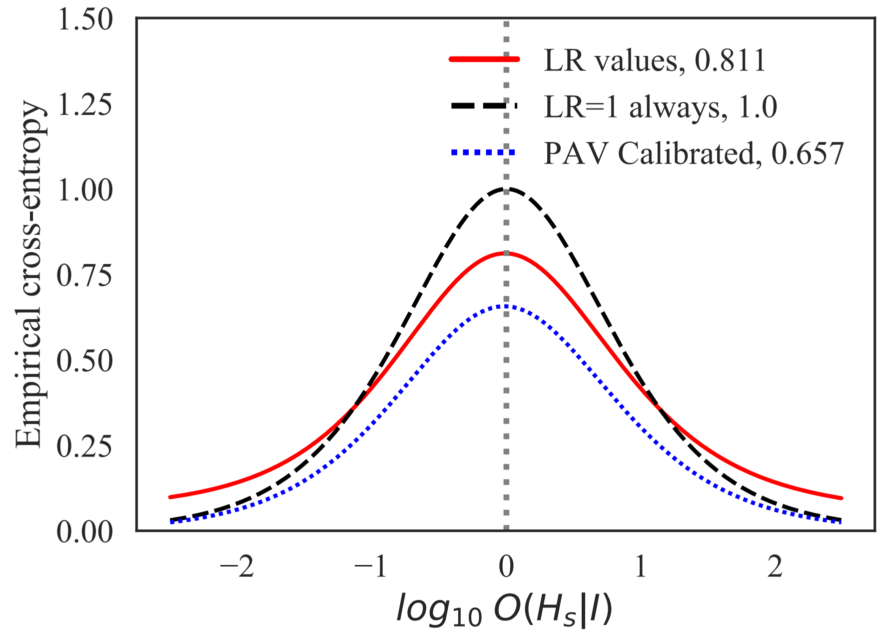


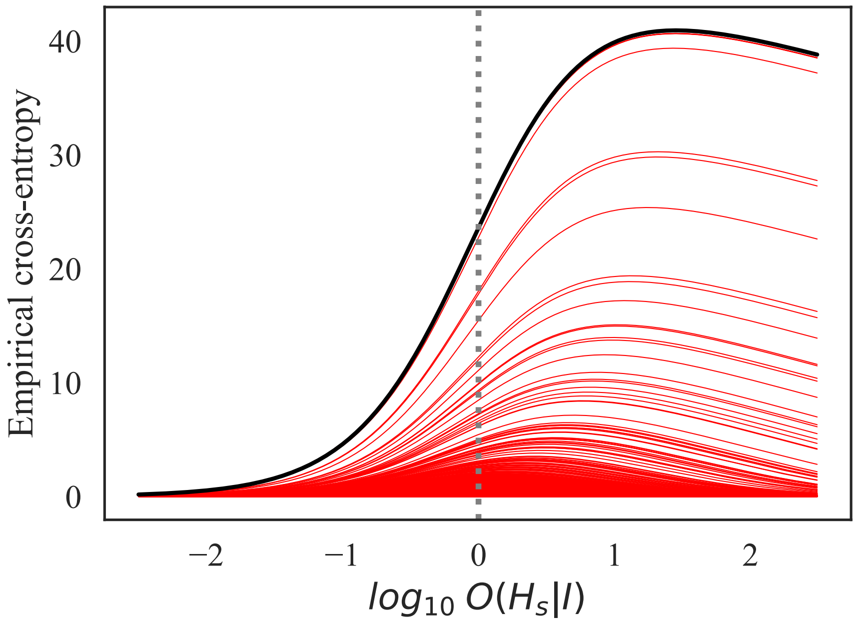
H_s true

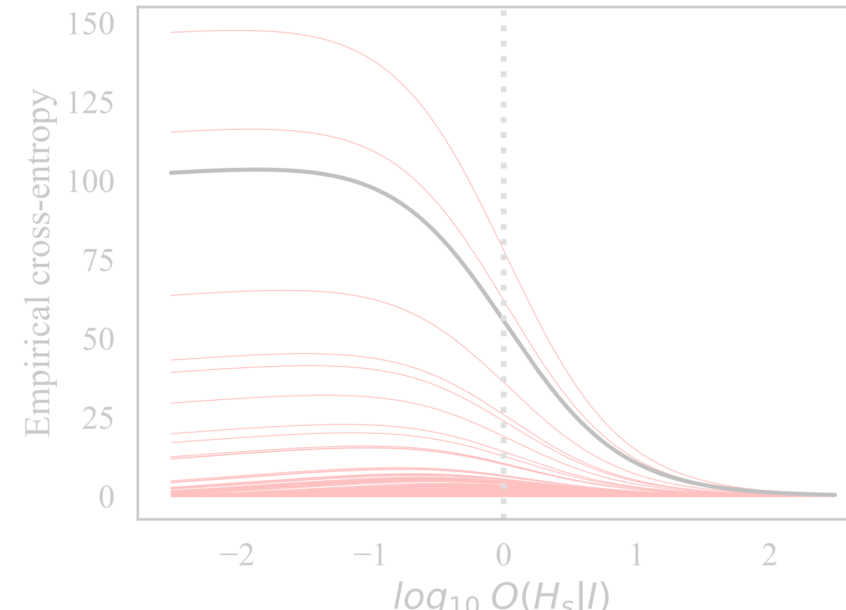


H_d true

*OC data; LR with $\alpha(n_a)$ weights

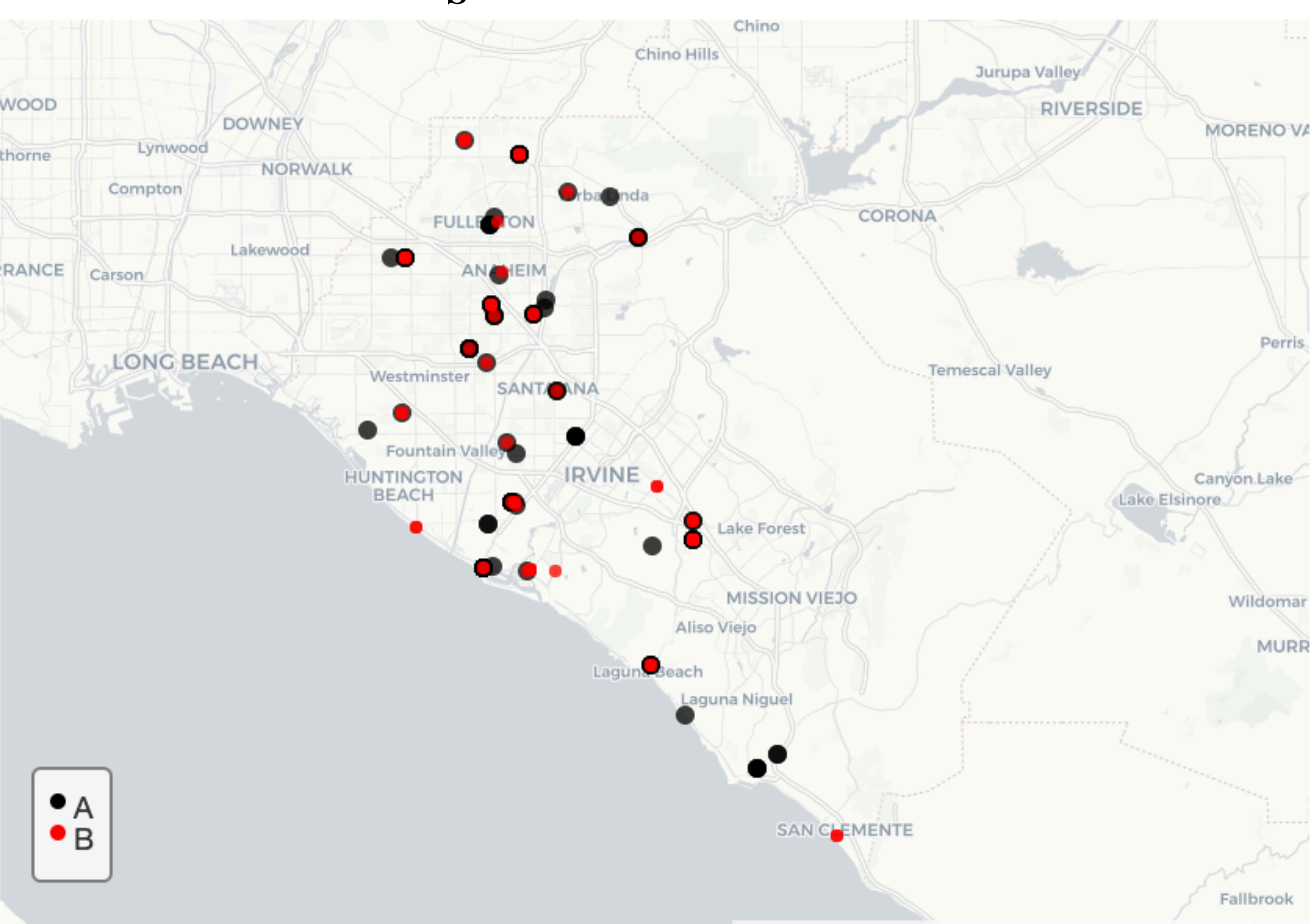


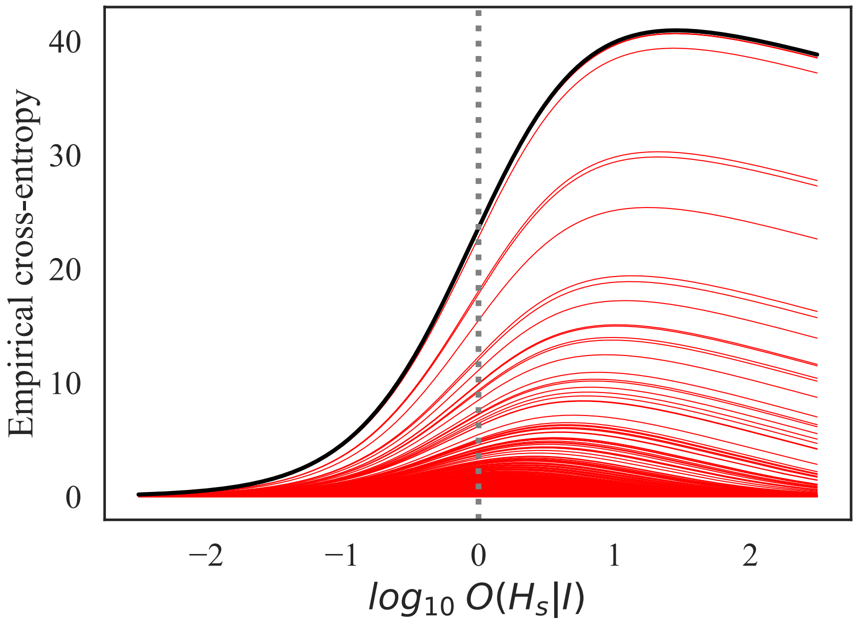


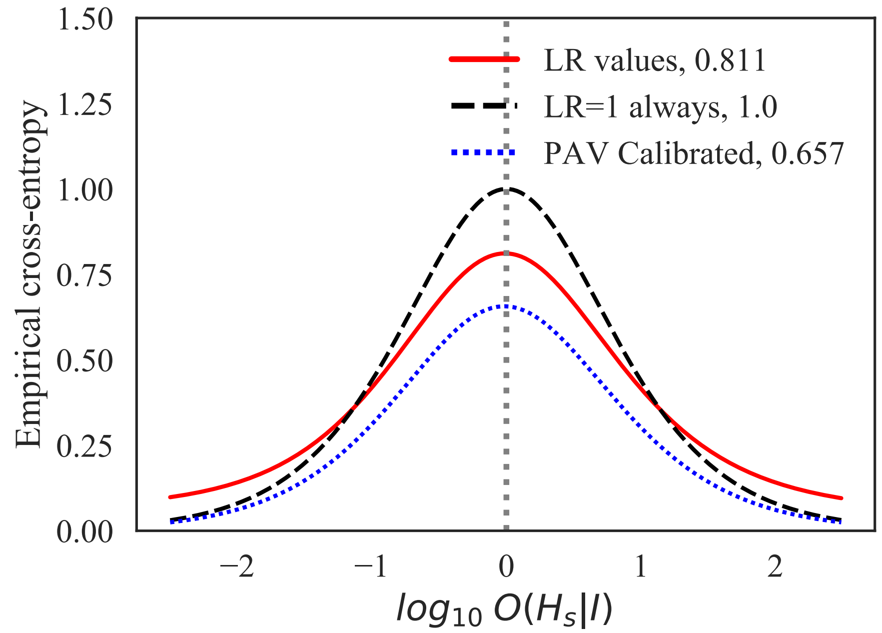


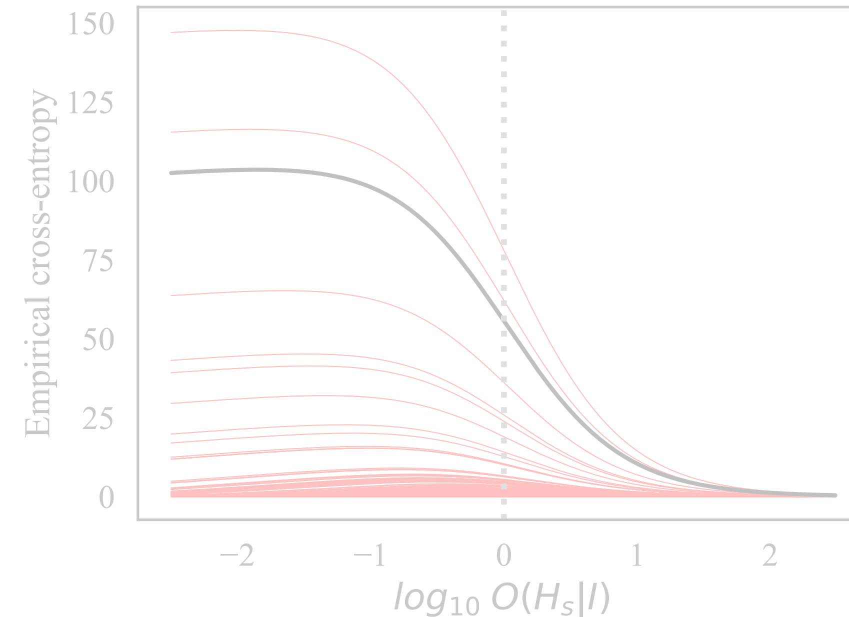
H_s true

H_d true



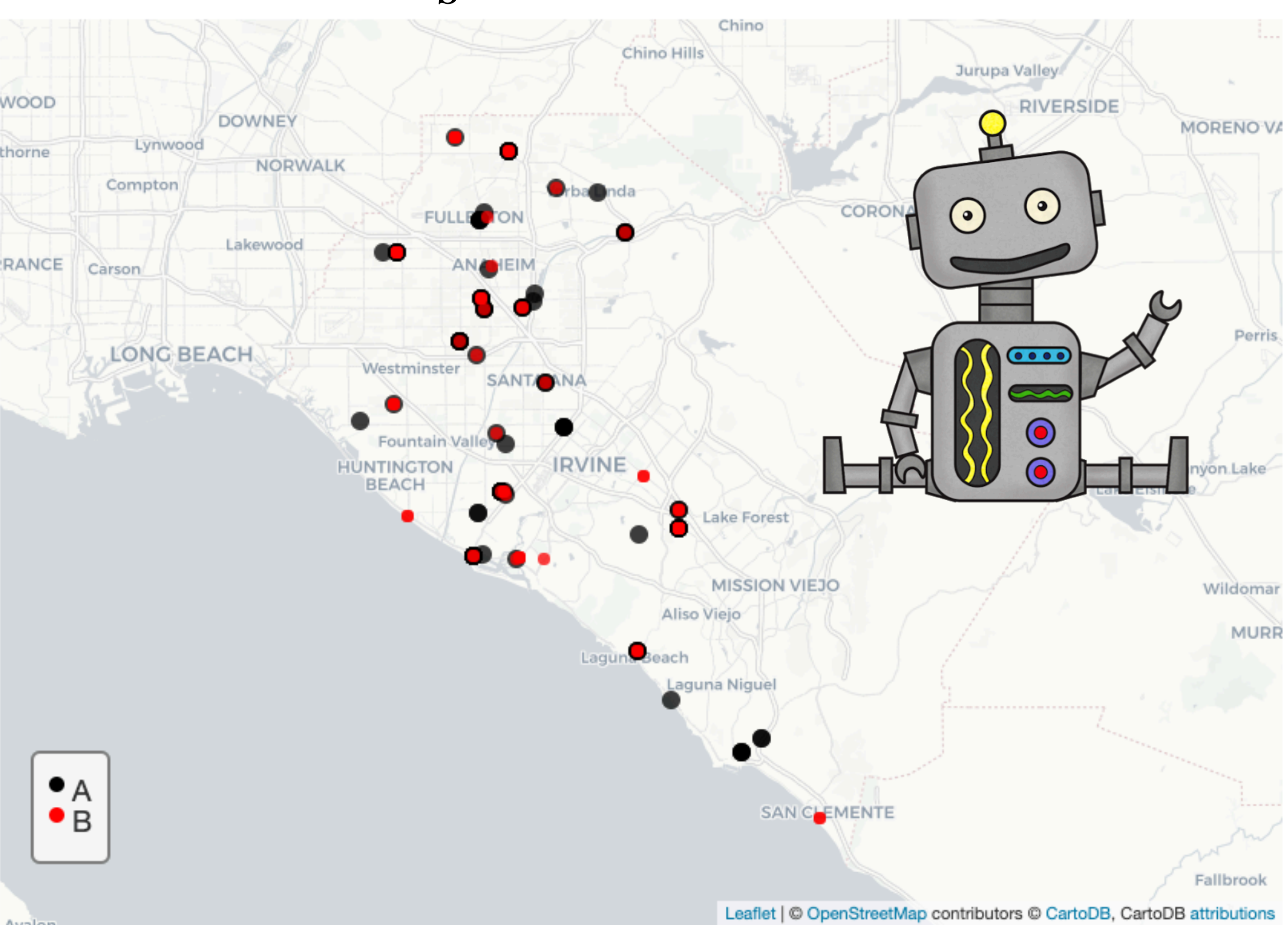


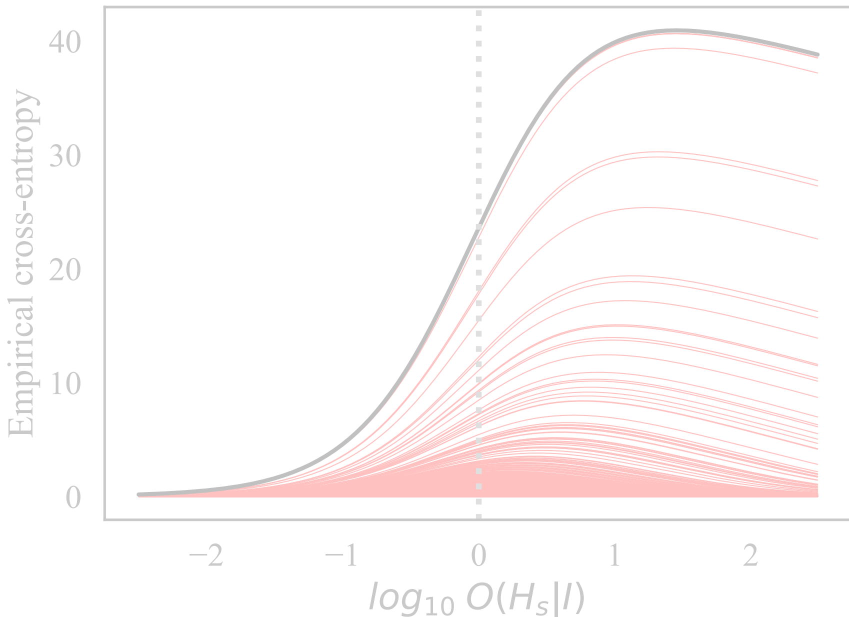


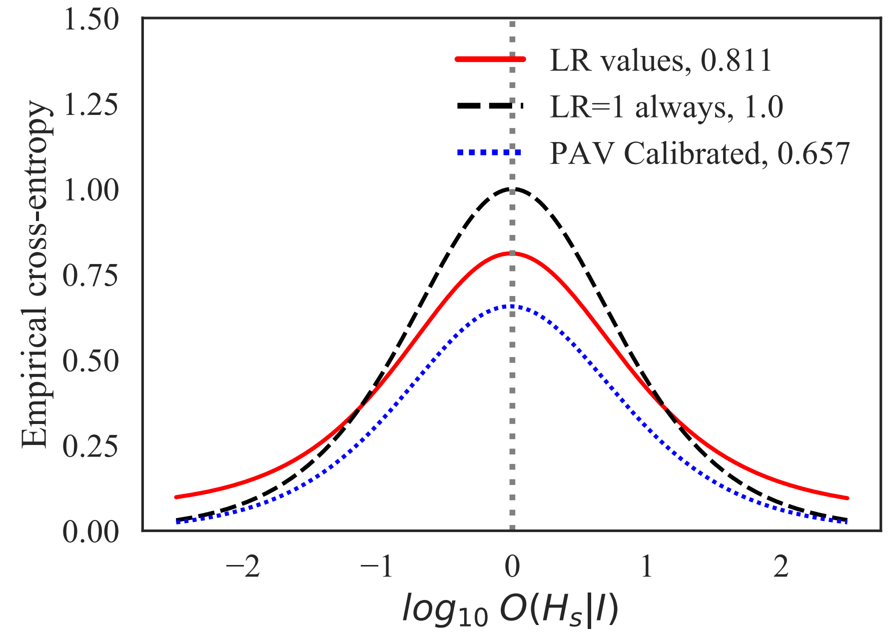


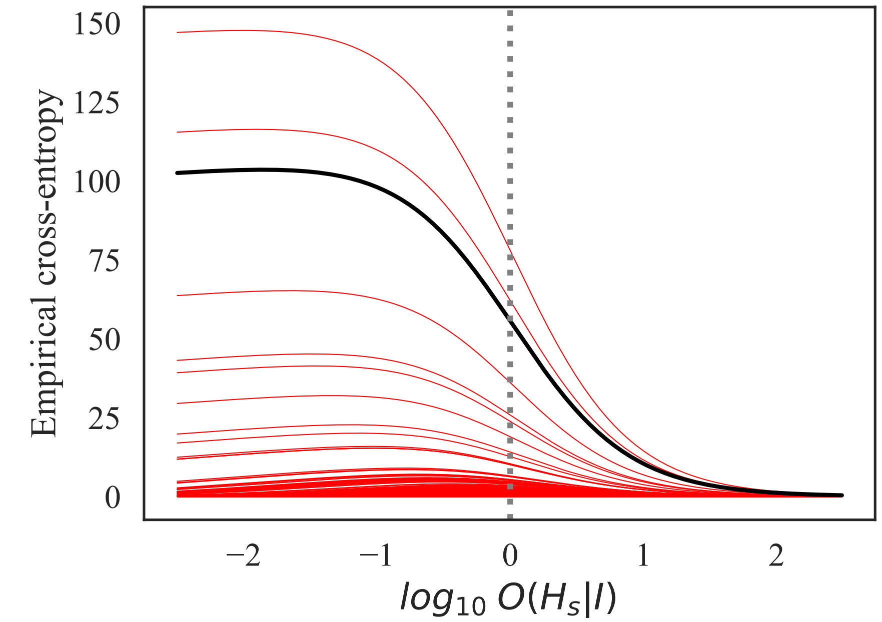
H_s true

H_d true



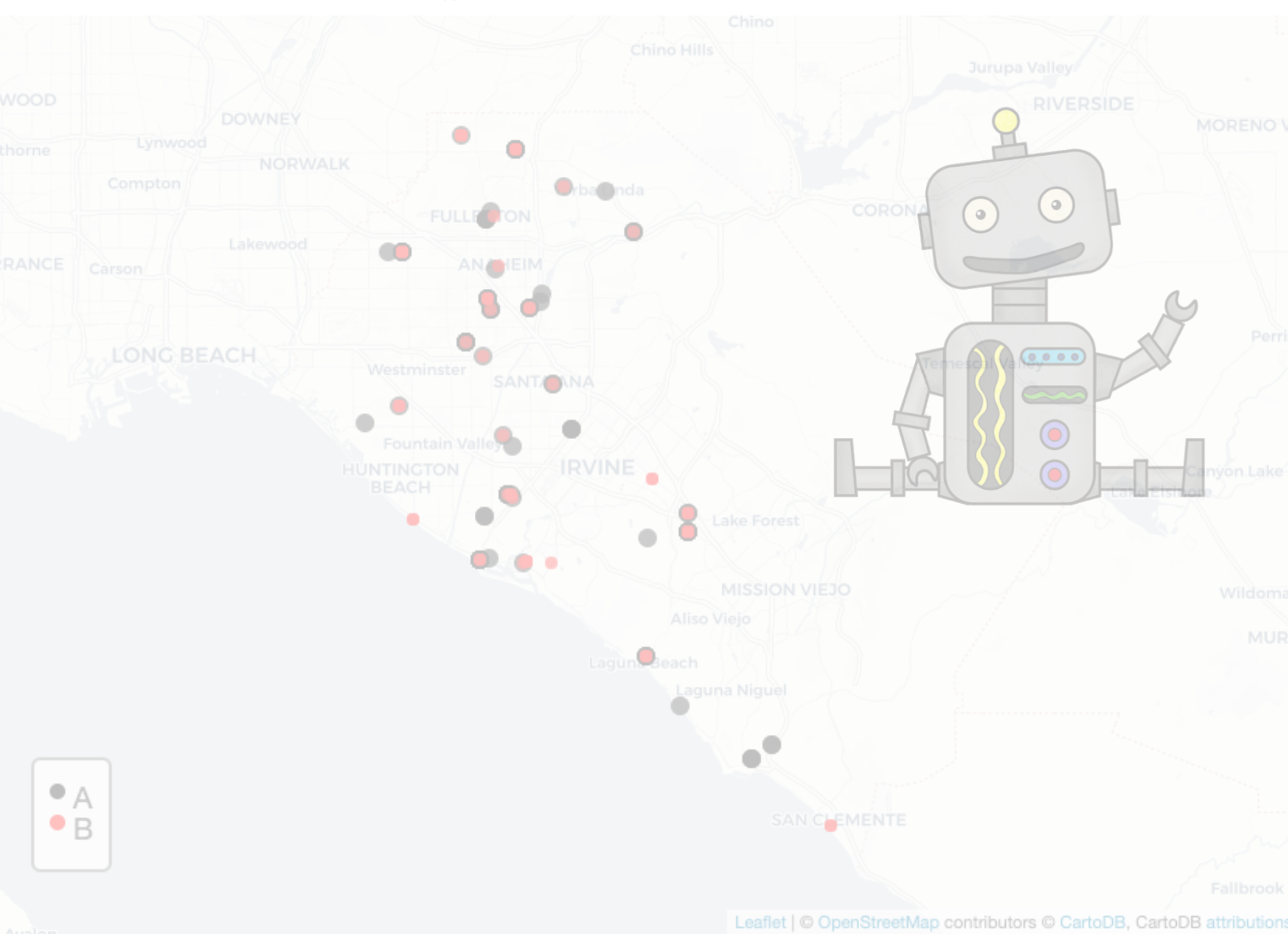


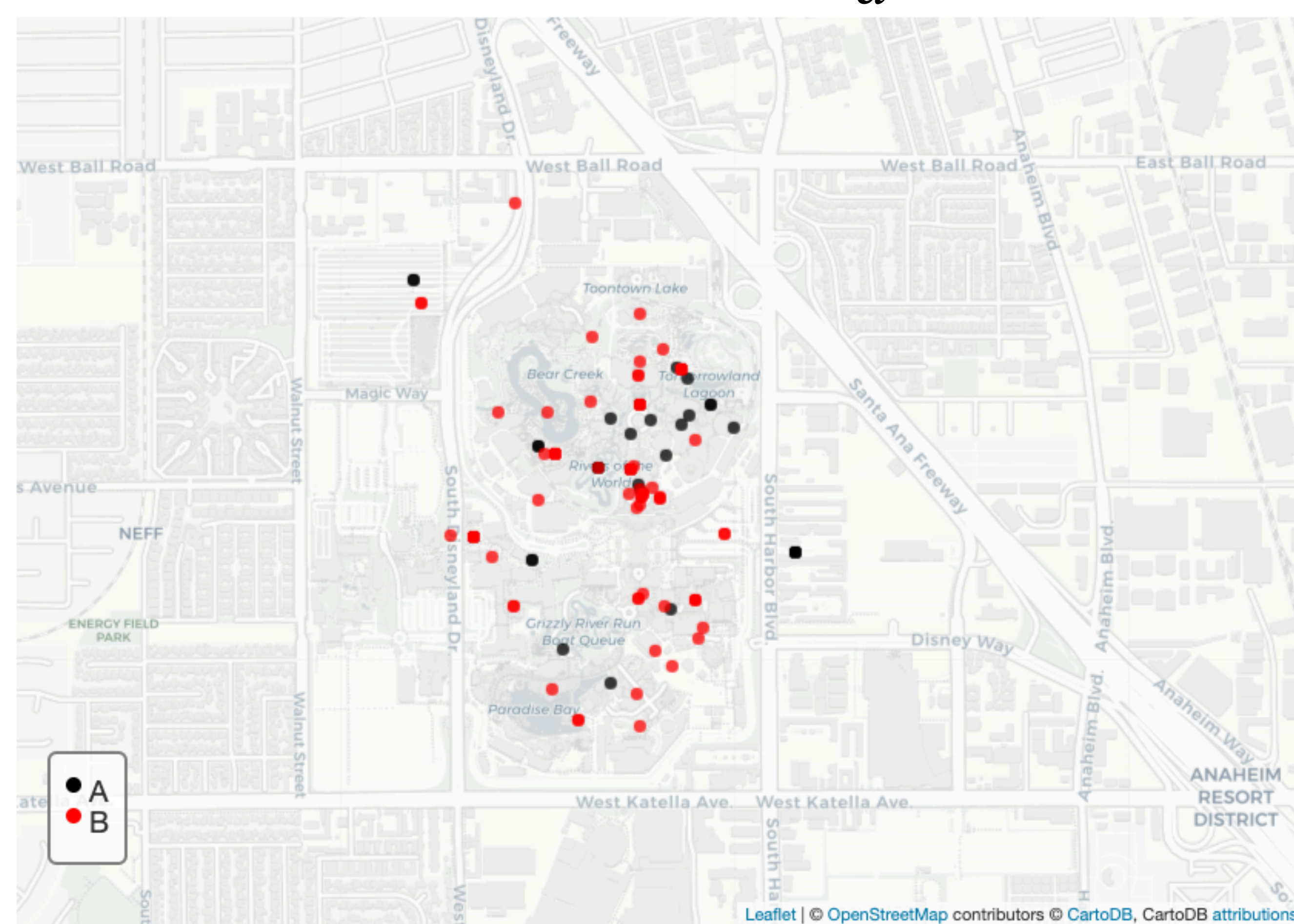


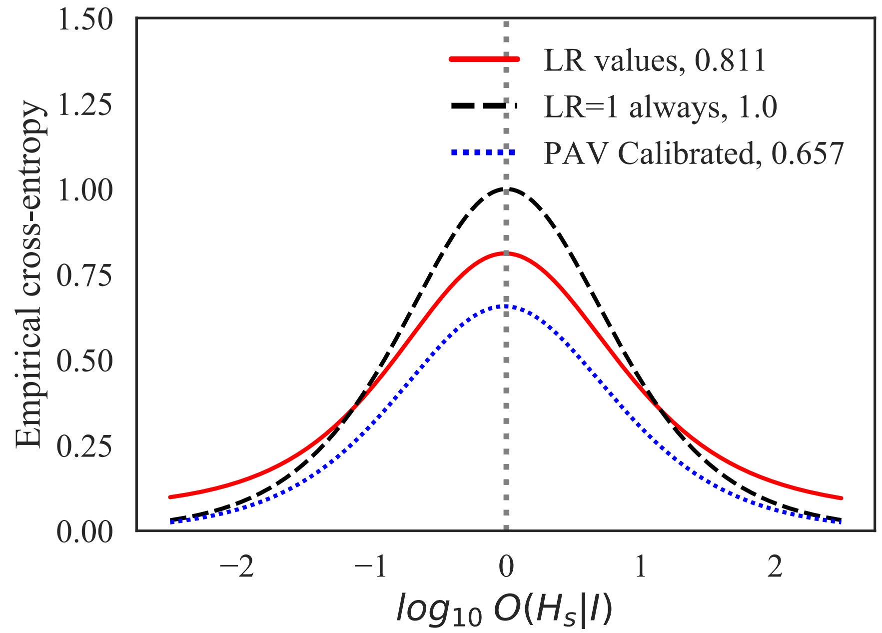


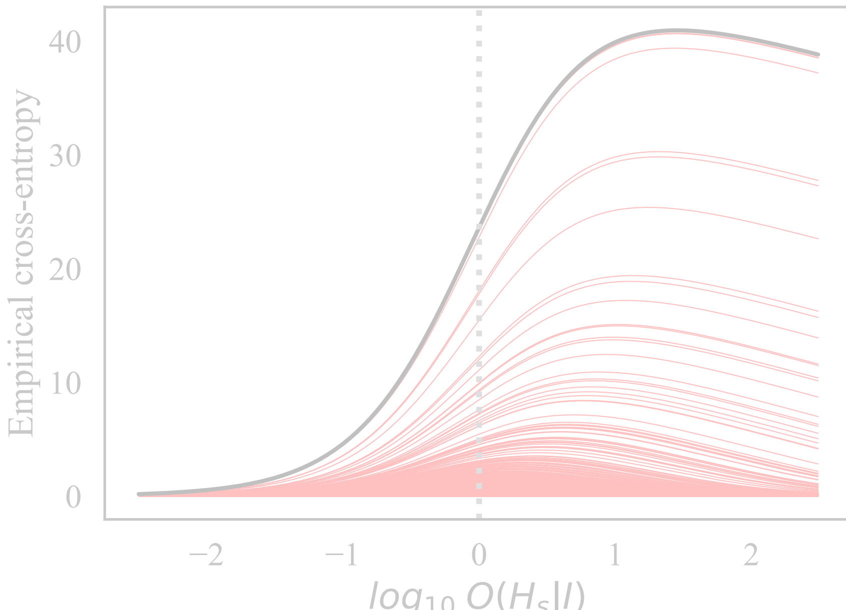
H_s true

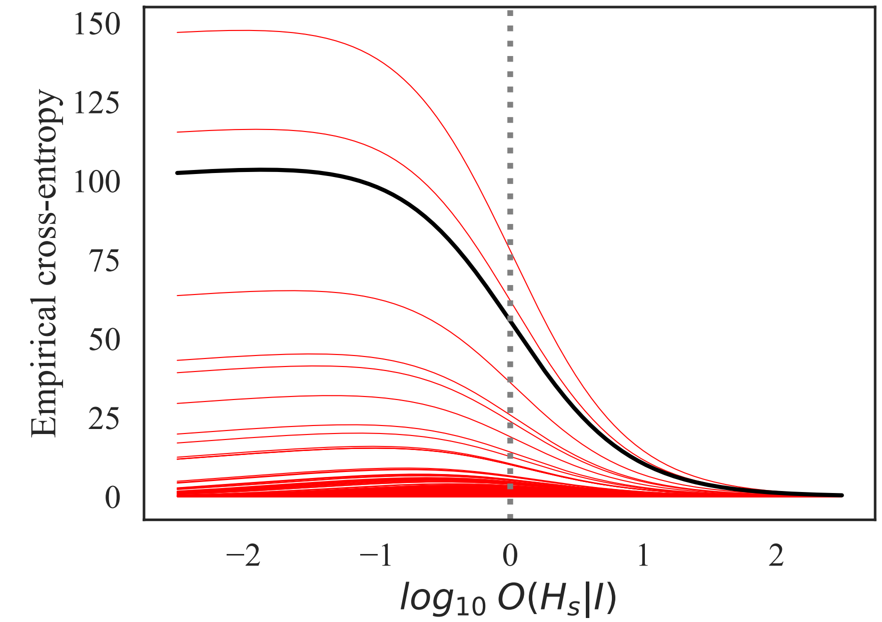
H_d true





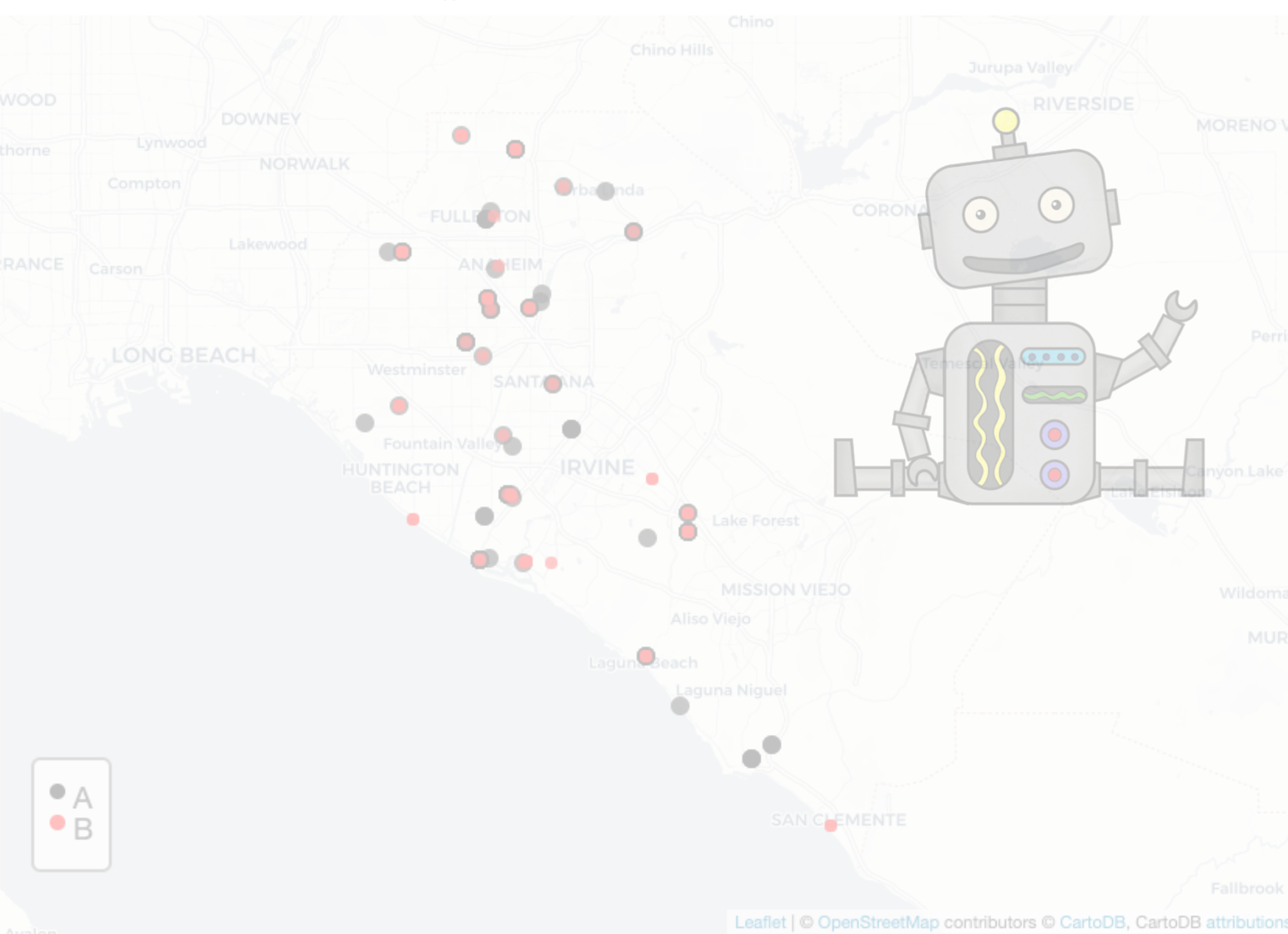


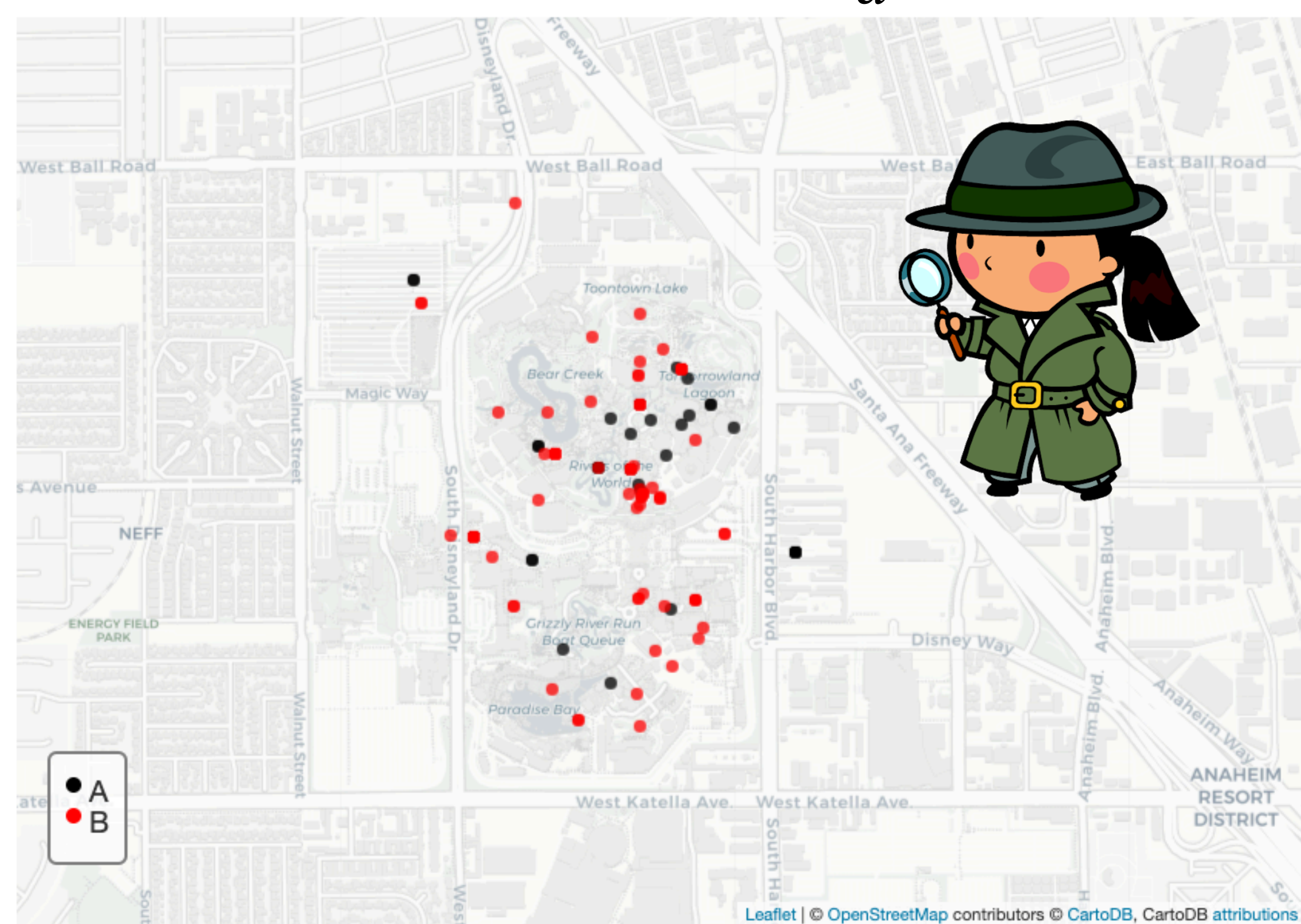




H_s true

H_d true





Future Directions and Summary

Future Directions

- **Reference Data:** Collect & share relevant digital data amongst law enforcement & researchers, e.g., start to build CODIS-like databases.

- **Assessment Techniques:** Classification performance & calibration are good ways to assess a method, but “misclassified” evidence complicates things…is there a systematic way to handle this?

- **Discovery:** Finding the most likely known source in a database given an unknown source sample…quickly.

- **Model Extensions:**
  - → Spatio-temporal models
  - → Incorporating event metadata

Summary

- **Statistical approaches** play a key role in the **forensic analysis** of a wide variety of evidence.

- **Digital evidence** is lagging behind other forensic disciplines.

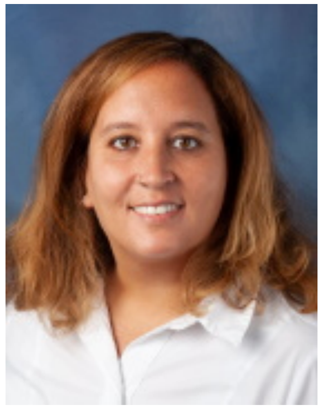
- **Contributions presented:**
  - *Coincidental Match Probability:* Novel technique for quantifying strength of evidence
  - *Geolocated Event Data:* Framework for estimating LRs and investigation of appropriate score functions

Many Thanks to…

MY ADVISOR

Many Thanks to…

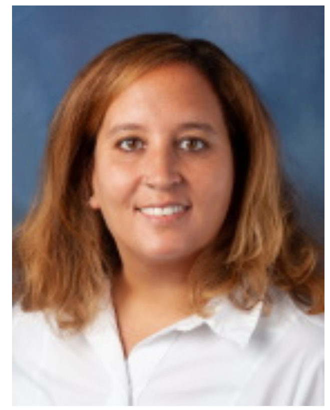
MY COMMITTEE



Many Thanks to…

MY COMMITTEE



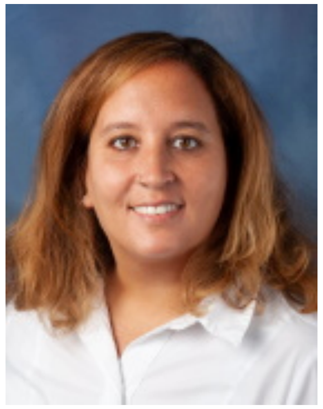


MY RESEARCH GROUP

UCI DATALAB

Many Thanks to…

MY COMMITTEE



MY RESEARCH GROUP

UCI DATALAB

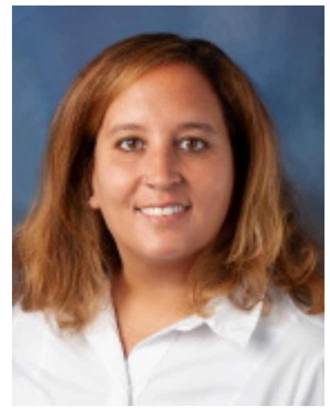
BREN:ICS

INFORMATION AND COMPUTER SCIENCES

UCIRVINE | UNIVERSITY of CALIFORNIA

Many Thanks to...

MY COMMITTEE



MY RESEARCH GROUP

UCI
DATALAB

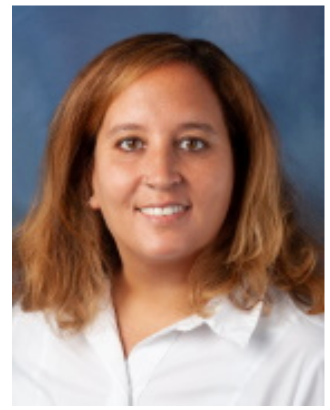
NIST
National Institute of
Standards and Technology
U.S. Department of Commerce

csafe
Center for Statistics and
Applications in Forensic Evidence

BREN:ICS
INFORMATION AND COMPUTER SCIENCES
UCIRVINE | UNIVERSITY of CALIFORNIA

Many Thanks to…

MY COMMITTEE



OBSIDIAN

CYLANCE

SOUTH DAKOTA STATE UNIVERSITY

MY RESEARCH GROUP

UCI DATALAB

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

csafe
Center for Statistics and Applications in Forensic Evidence

BREN:ICS
INFORMATION AND COMPUTER SCIENCES
UCIRVINE | UNIVERSITY of CALIFORNIA

Questions

Appendix

Cross-Entropy

$$\mathcal{U}_{Q||P}(H_s | E) = -\mathbb{E}_{Q(E,H_s)} \log P(H_s | E)$$
$$= -\sum_{k=0}^{1} Q(H_s = k) \int q(e | H_s = k) \log P(H_s = k | e) de \,.$$

$$\mathcal{U}_{Q||P}(H_s | E) = \mathcal{U}_{Q}(H_s | E) + D_{Q||P}(H_s | E)$$

Log Loss

$$L\left[Q(H_s | e), P(H_s | e)\right] = -Q(H_s | e)\log P(H_s | e) - (1 - Q(H_s | e))\log(1 - P(H_s | e))$$

Risk

$$R(Q, P) = \mathbb{E}_{q(E|H_s)} L\left[Q(H_s | E), P(H_s | E)\right] = \int q(e | H_s) L\left[Q(H_s | e), P(H_s | e)\right] de$$

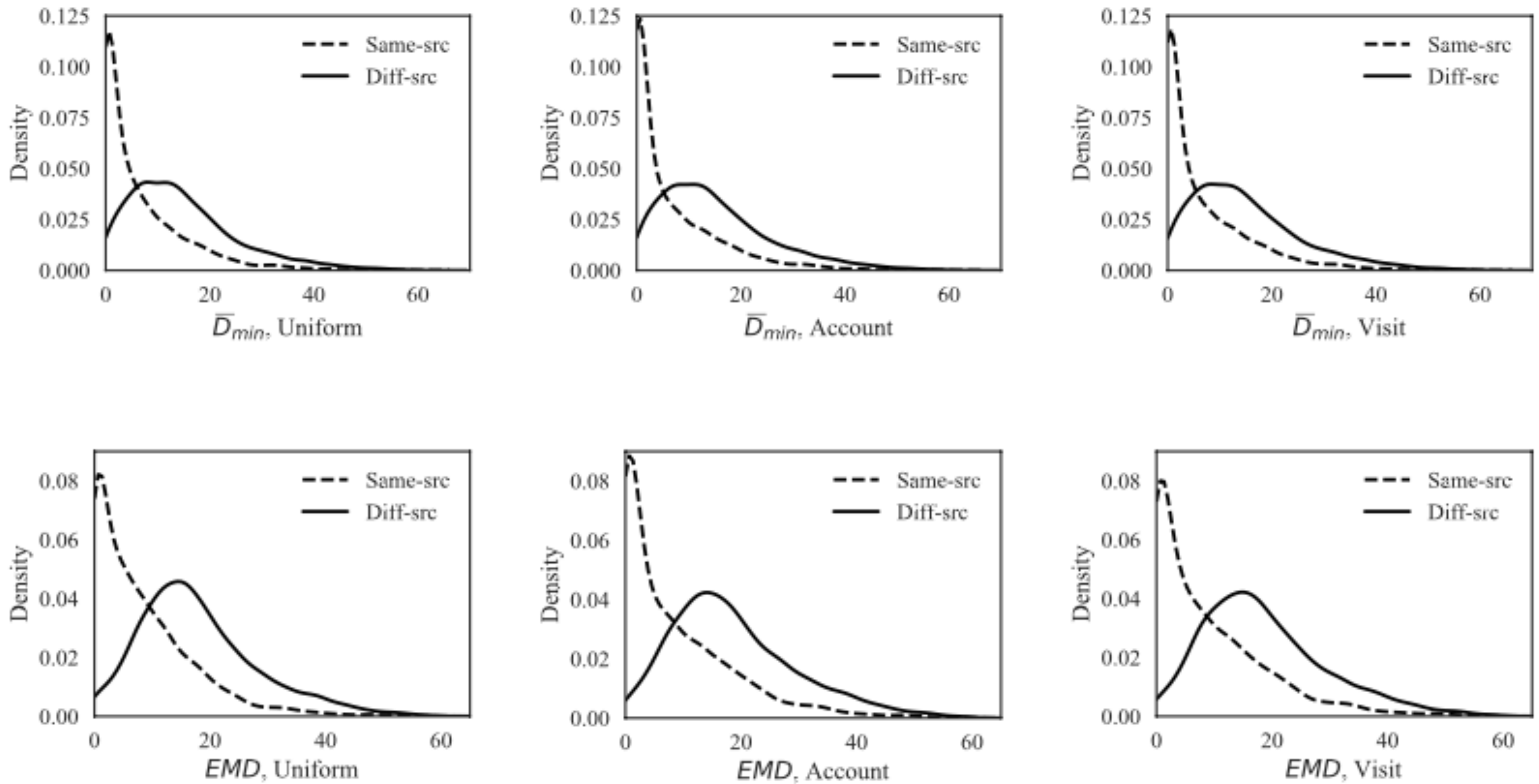
Bayes Risk

$$R_B(P) = \sum_{k=0}^{1} Q(H_s = k) \mathbb{E}_{q(E|H_s=k)} L\left[Q(H_s | e), P(H_s | e)\right]$$
$$= -\sum_{k=0}^{1} Q(H_s = k) \int q(e | H_s = k) \log P(H_s = k | e) de$$
$$= \mathcal{U}_{Q||P}(\theta | E)$$

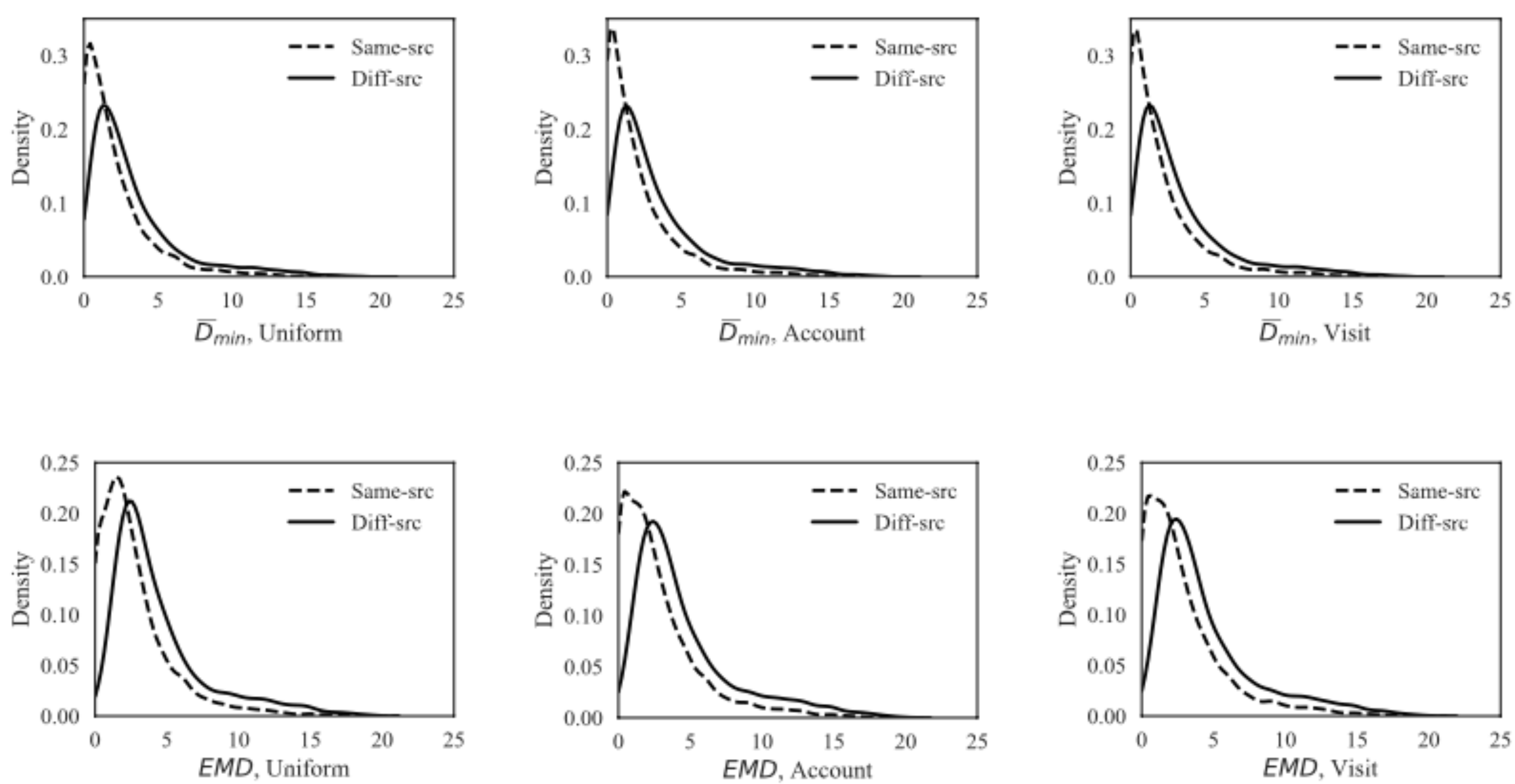
Empirical Cross-Entropy

$$ECE = -\frac{P(H_s)}{N_s^*} \sum_{i \in \mathcal{D}_s^*} \log P(H_s | e_i) - \frac{1 - P(H_s)}{N_d^*} \sum_{j \in \mathcal{D}_d^*} \log(1 - P(H_s | e_j))$$

OC


Same-src
Diff-src
Density
$\overline{D}_{min}$, Uniform
$\overline{D}_{min}$, Account
$\overline{D}_{min}$, Visit
EMD, Uniform
EMD, Account
EMD, Visit

NY


Same-src
Diff-src
Density
$\overline{D}_{min}$, Uniform
$\overline{D}_{min}$, Account
$\overline{D}_{min}$, Visit
EMD, Uniform
EMD, Account
EMD, Visit

Region	Weight	TP@1	FP@1	AUC
OC	0.80	0.340	**0.026**	0.787
	$\alpha(n_a)$	**0.380**	0.038	**0.845**
	$\alpha(n_a \vert \gamma, \rho, \phi)$	0.375	0.037	0.817
NY	0.80	0.251	**0.067**	0.711
	$\alpha(n_a)$	**0.285**	0.089	**0.768**
	$\alpha(n_a \vert \gamma, \rho, \phi)$	0.282	0.088	0.734

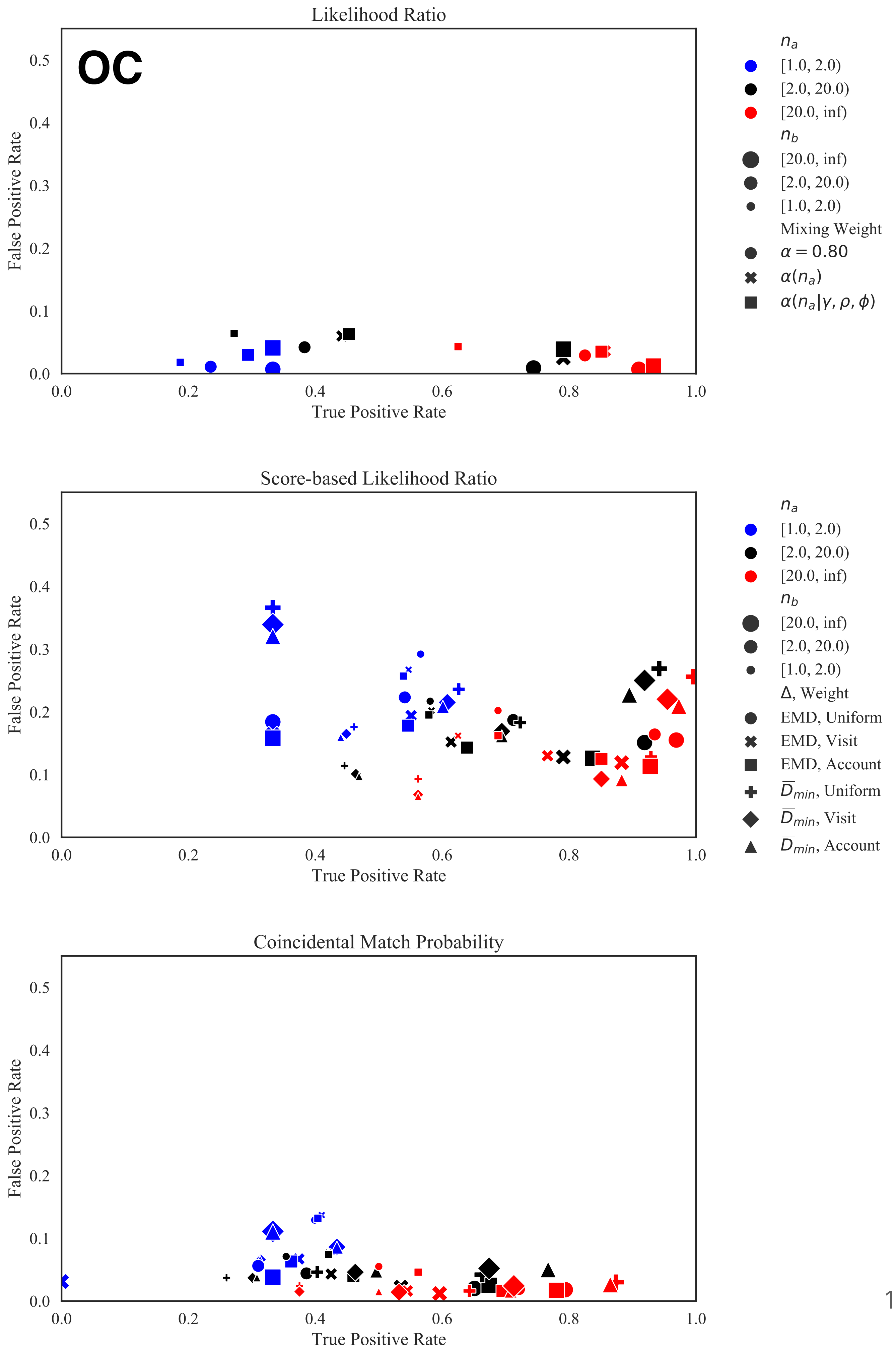
LR

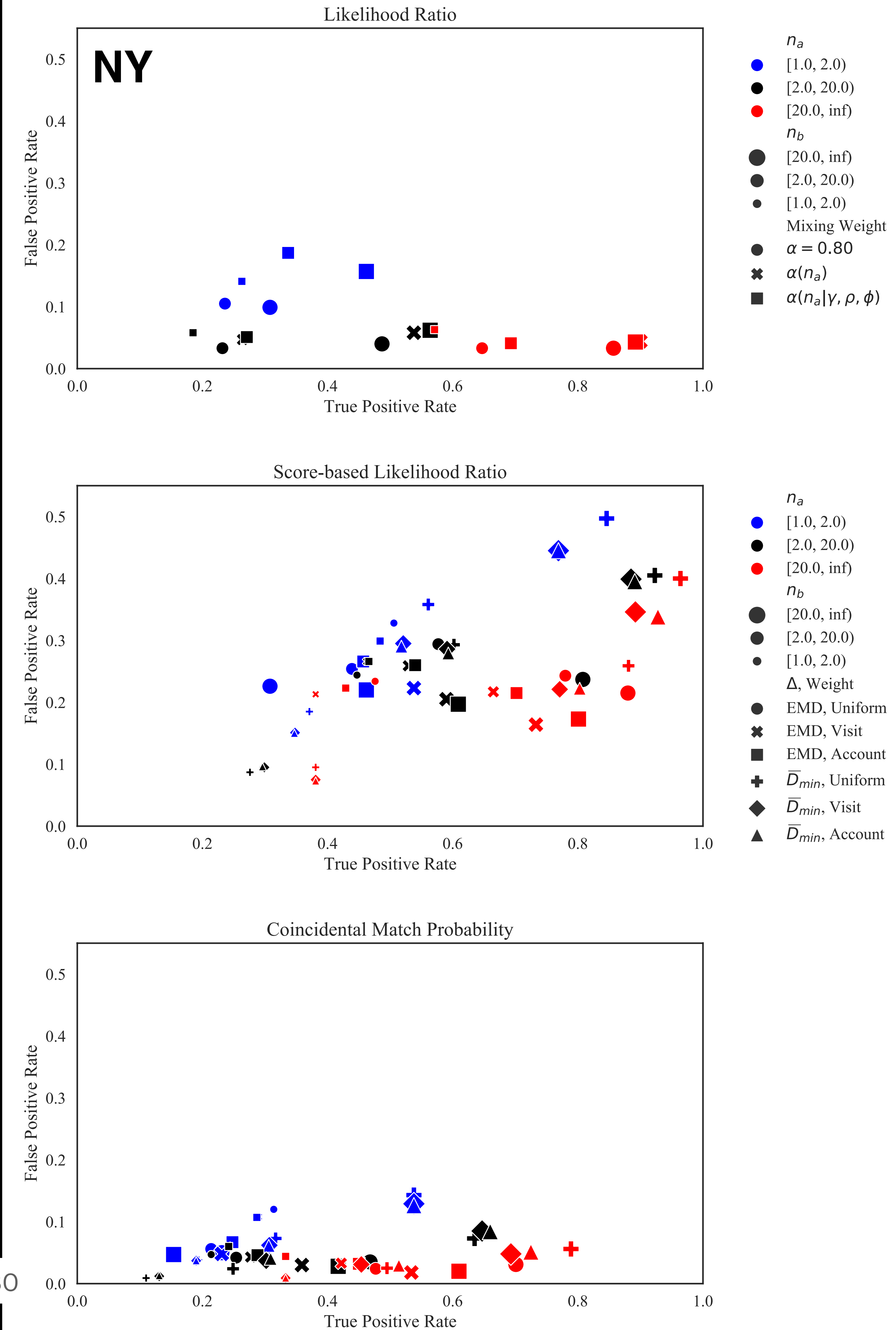
Region	Δ	Weights	TP@1	FP@1	AUC
OC	$\overline{D}_{min}$	Uniform	0.628	0.202	0.768
	$\overline{D}_{min}$	Account	0.610	0.171	0.774
	$\overline{D}_{min}$	Visit	0.611	0.180	0.768
	EMD	Uniform	**0.654**	0.197	**0.790**
	EMD	Account	0.614	**0.162**	0.783
	EMD	Visit	0.602	0.169	0.774
NY	$\overline{D}_{min}$	Uniform	0.508	0.287	0.656
	$\overline{D}_{min}$	Account	0.494	0.254	0.666
	$\overline{D}_{min}$	Visit	0.493	0.257	0.663
	EMD	Uniform	**0.530**	0.253	**0.686**
	EMD	Account	0.511	0.235	0.685
	EMD	Visit	0.504	**0.234**	0.679

SLR

Region	Δ	Weights	TP@0.05	TP@0.01	AUC
OC	$\overline{D}_{min}$	Uniform	0.389	0.187	0.771
	$\overline{D}_{min}$	Account	0.441	**0.236**	0.776
	$\overline{D}_{min}$	Visit	0.415	0.209	0.771
	EMD	Uniform	0.397	0.154	**0.791**
	EMD	Account	**0.448**	0.208	0.784
	EMD	Visit	0.425	0.182	0.775
NY	$\overline{D}_{min}$	Uniform	0.242	0.153	0.656
	$\overline{D}_{min}$	Account	0.269	**0.186**	0.667
	$\overline{D}_{min}$	Visit	0.264	0.179	0.665
	EMD	Uniform	0.265	0.139	**0.687**
	EMD	Account	**0.283**	0.161	0.686
	EMD	Visit	0.276	0.156	0.681

CMP


OC
Likelihood Ratio
Score-based Likelihood Ratio
Coincidental Match Probability
False Positive Rate
True Positive Rate
n_a
[1.0, 2.0)
[2.0, 20.0)
[20.0, inf)
n_b
[20.0, inf)
[2.0, 20.0)
[1.0, 2.0)
Mixing Weight
$\alpha = 0.80$
$\alpha(n_a)$
$\alpha(n_a|\gamma, \rho, \phi)$
Δ, Weight
EMD, Uniform
EMD, Visit
EMD, Account
$\overline{D}_{min}$, Uniform
$\overline{D}_{min}$, Visit
$\overline{D}_{min}$, Account


NY
Likelihood Ratio
Score-based Likelihood Ratio
Coincidental Match Probability
False Positive Rate
True Positive Rate
n_a
[1.0, 2.0)
[2.0, 20.0)
[20.0, inf)
n_b
[20.0, inf)
[2.0, 20.0)
[1.0, 2.0)
Mixing Weight
$\alpha = 0.80$
$\alpha(n_a)$
$\alpha(n_a|\gamma, \rho, \phi)$
Δ, Weight
EMD, Uniform
EMD, Visit
EMD, Account
$\overline{D}_{min}$, Uniform
$\overline{D}_{min}$, Visit
$\overline{D}_{min}$, Account