Statistical Methods for the Forensic Analysis of Geolocated Event Data

By: **Christopher Galbraith** (Department of Statistics, University of California, Irvine), Padhraic Smyth (Department of Computer Science, University of California, Irvine), and Hal S. Stern (Department of Statistics, University of California, Irvine)

From the proceedings of

The Digital Forensic Research Conference

DFRWS 2020 USA

Virtual -- July 20-24

Statistical Methods for the Forensic Analysis of Geolocated Event Data

Christopher Galbraith
Padhraic Smyth
Hal S. Stern

DFRWS USA
7.22.20

UCIRVINE
UNIVERSITY of CALIFORNIA · IRVINE

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.

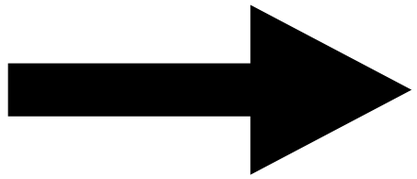
Outline

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

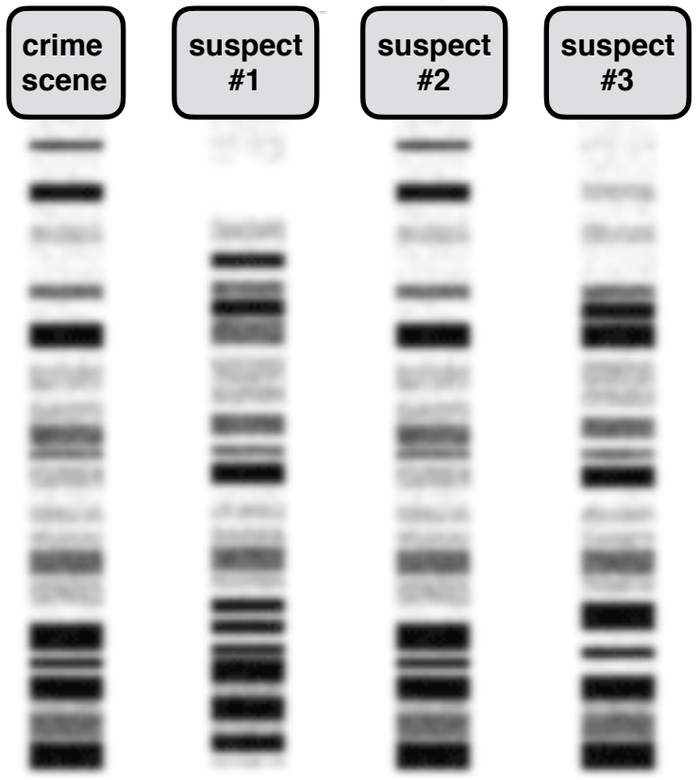

APPS
WIDGETS
12:30 AM
Browser
Calculator
Calendar
Camera
Chrome
Clock
Downloads
Drive
Email
File Explorer
Flashlight
FM Radio
Gallery
Gmail
Google
Google Settings
Google+
Hangouts
Maps
Messaging

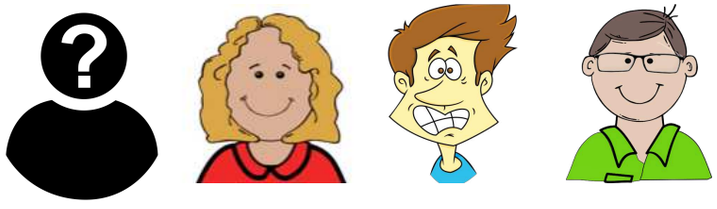

POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS



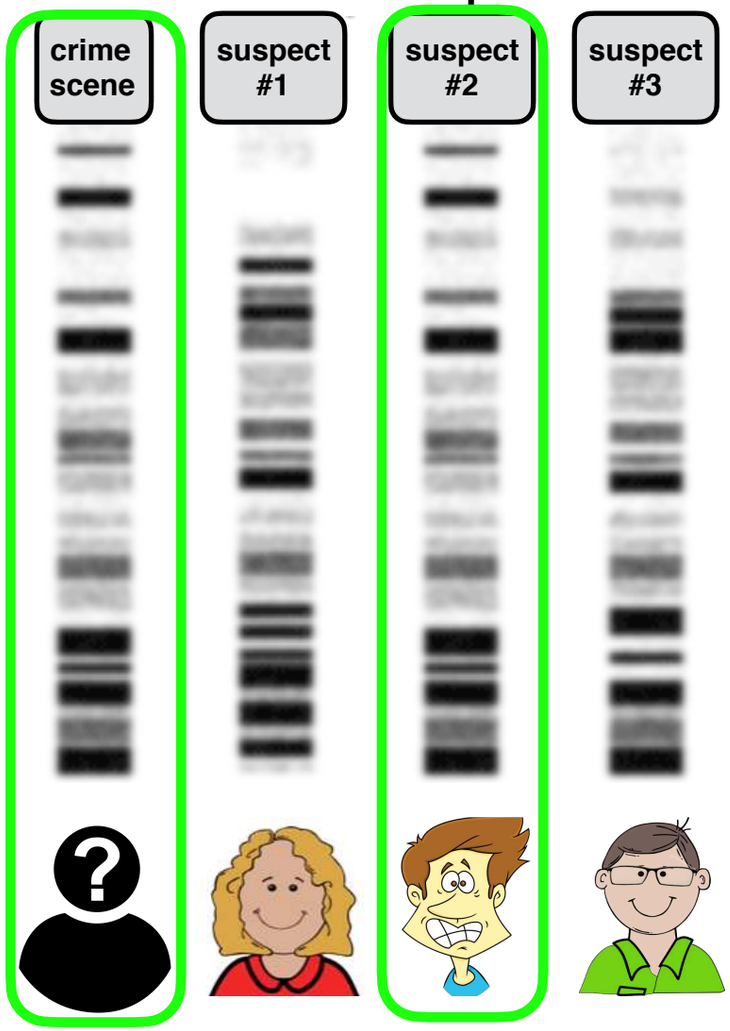

POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS
POLICE LINE DO NOT CROSS

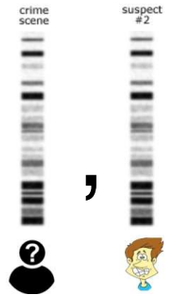
DNA Samples





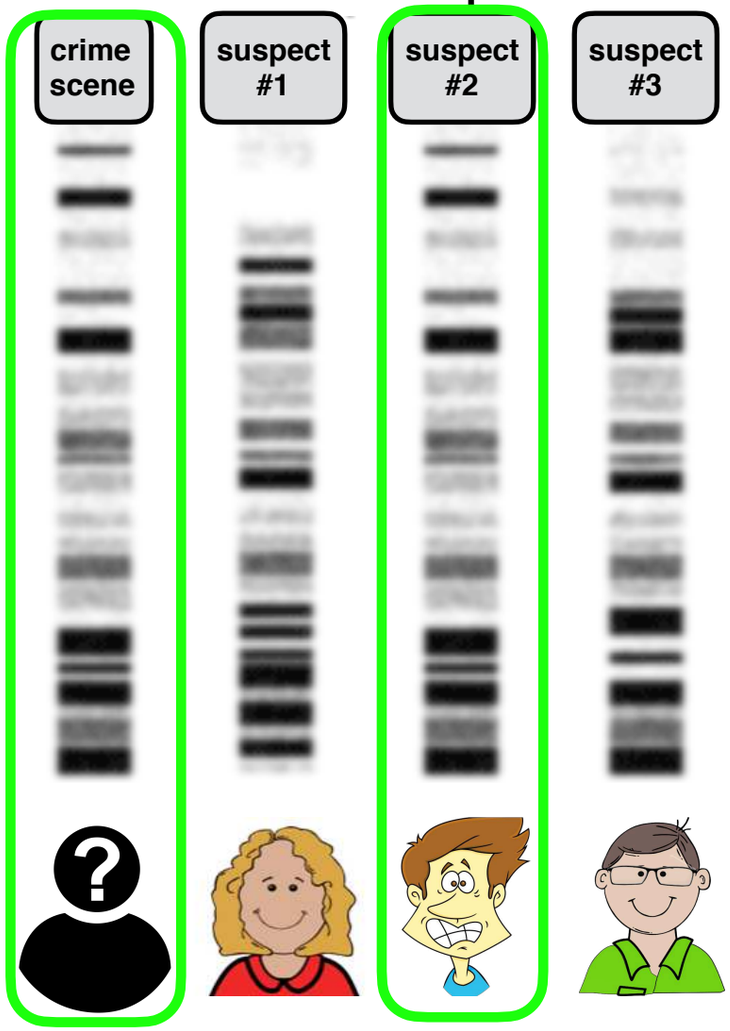
DNA Samples



i $\Pr($  $,$ $\mid$  $=$  $)$

[Steele & Balding, 2014]

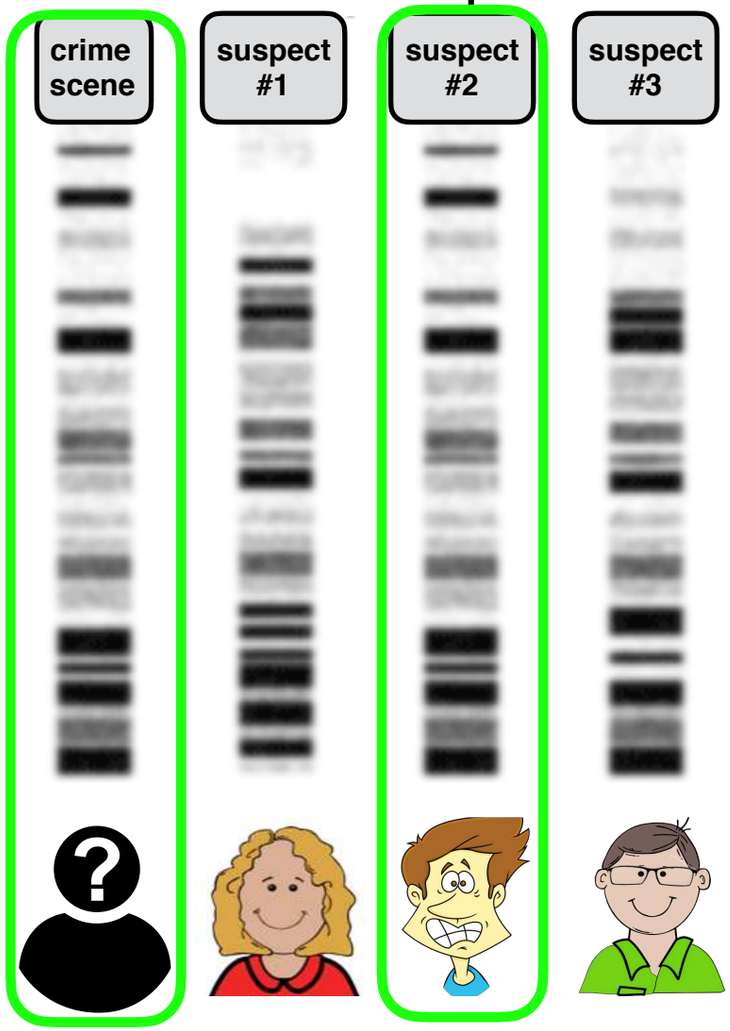
DNA Samples



i $\Pr(\;\text{crime scene}\;,\;\text{suspect \#2}\;|\;\text{?} = \text{suspect \#2}\;) = 1$

[Steele & Balding, 2014]

DNA Samples

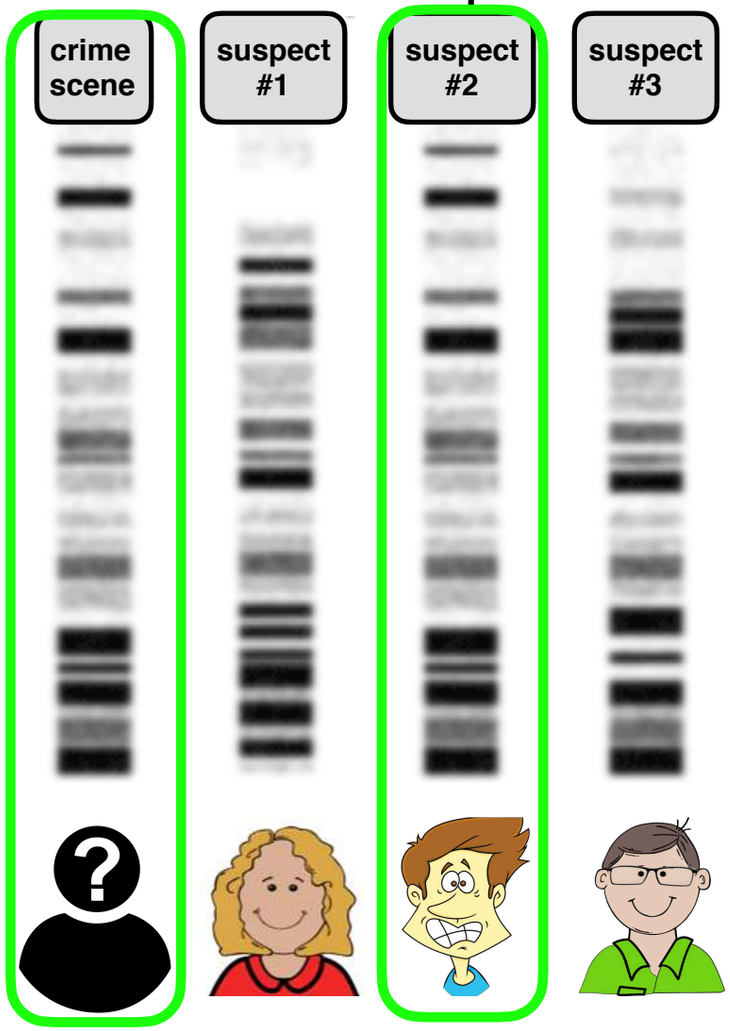


i $\Pr(\text{crime scene}, \text{suspect \#2} \mid \text{?} = \text{suspect \#2}) = 1$

ii $\Pr(\text{crime scene}, \text{suspect \#2} \mid \text{?} \neq \text{suspect \#2})$

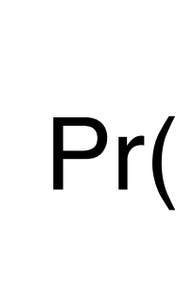
[Steele & Balding, 2014]

DNA Samples



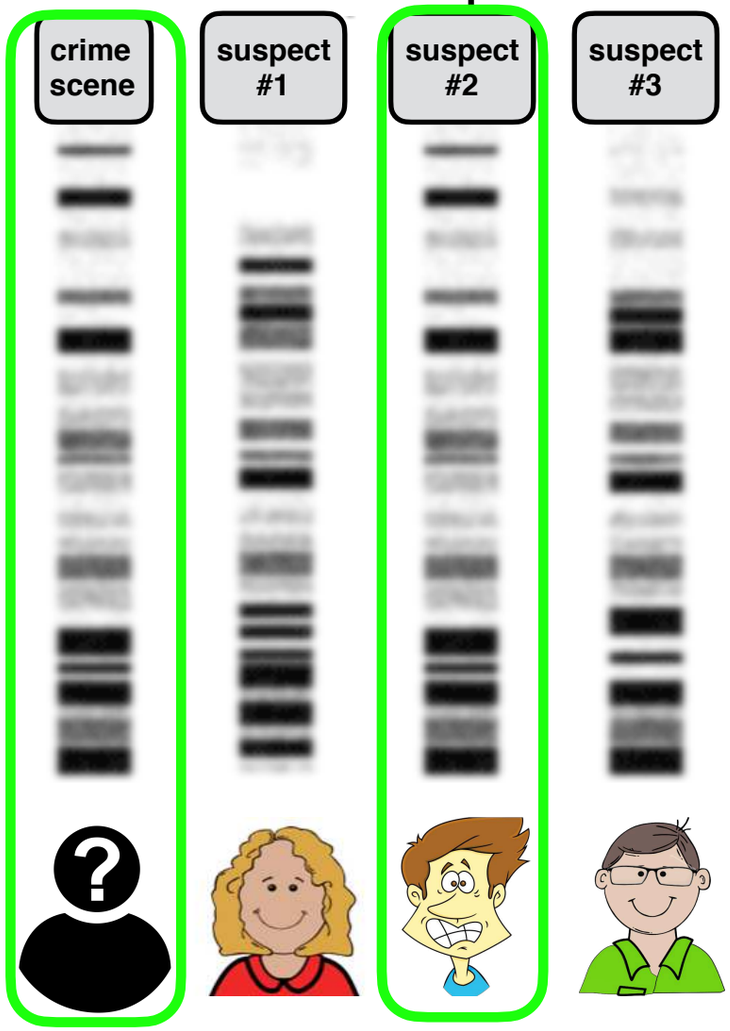
i $\Pr(\text{crime scene}, \text{suspect \#2} \mid ? = \text{suspect \#2}) = 1$

ii $\Pr(\text{crime scene}, \text{suspect \#2} \mid ? \neq \text{suspect \#2})$



[Steele & Balding, 2014]

DNA Samples



i $\Pr(\text{crime scene}, \text{suspect \#2} \mid \text{?} = \text{suspect \#2}) = 1$

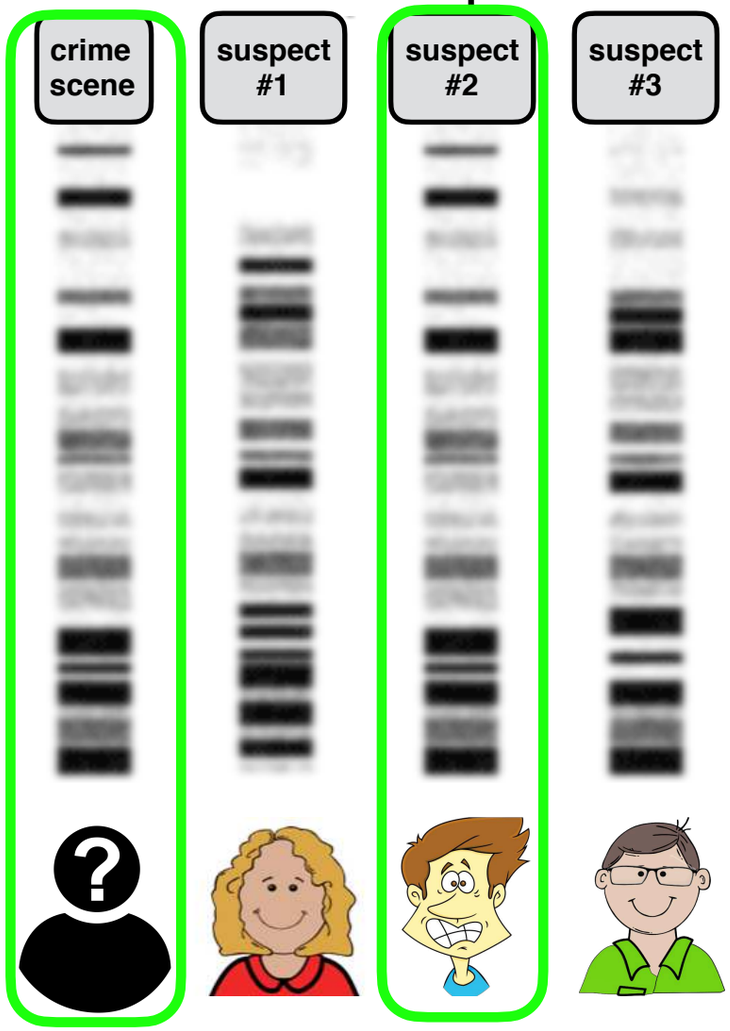
ii $\Pr(\text{crime scene}, \text{suspect \#2} \mid \text{?} \neq \text{suspect \#2})$

CODIS

Random Match Probability

[Steele & Balding, 2014]

DNA Samples



i $\Pr(\text{crime scene}, \text{suspect \#2} \mid \text{?} = \text{suspect \#2}) = 1$

ii $\Pr(\text{crime scene}, \text{suspect \#2} \mid \text{?} \neq \text{suspect \#2})$

CODIS

Random Match Probability

Likelihood Ratio

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

< 1	Samples from different sources
= 1	Inconclusive
> 1	Samples from same source

[Steele & Balding, 2014]

DNA Samples

i $\Pr(\text{crime scene}, \text{suspect \#2} \mid \text{?} = \text{suspect \#1}) = 1$

$\Pr(\text{crime scene}, \text{suspect \#2} \mid \text{?} \neq \text{suspect \#1})$

ch

sources

ource

Extraction

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

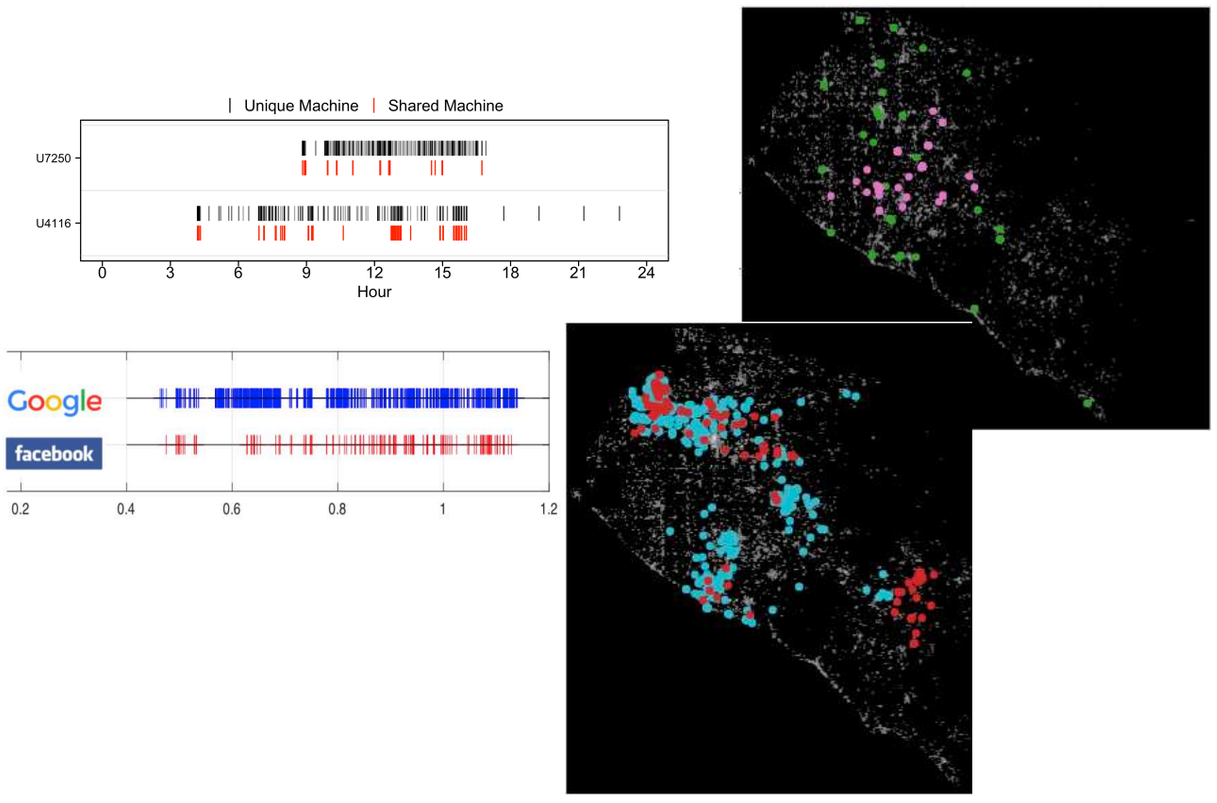
- Browser requests
- Web searches
- Email activity
- Phone/SMS
- Social media activity
- GPS locations
- File access
- Network activity
- Exercise/movement
- …

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]

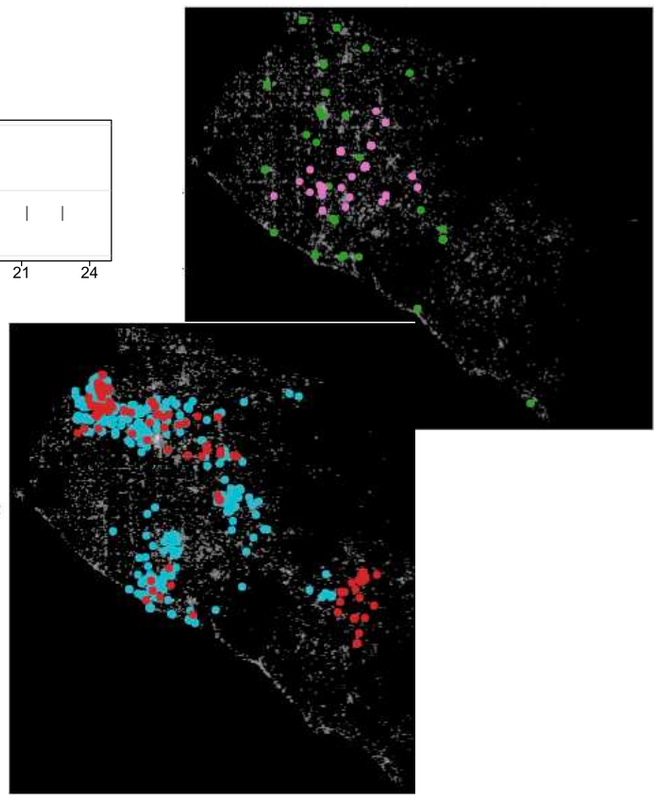
Extraction

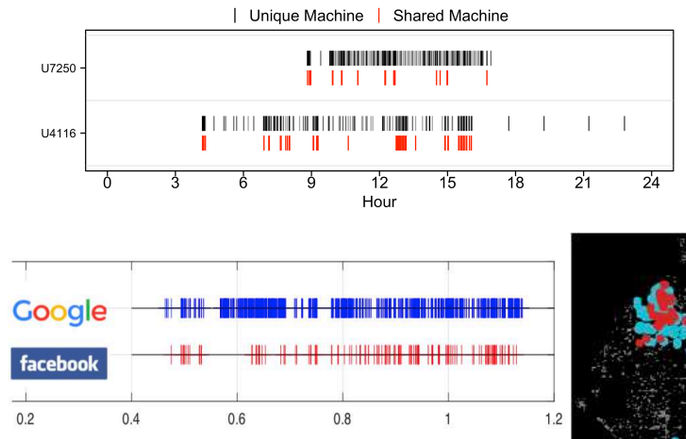
[SWDGE, 2019;
Roussev, 2016;
Casey, 2011]

- 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

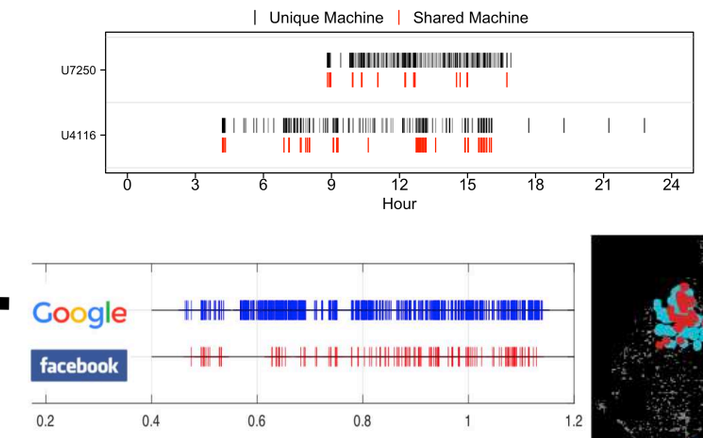
[SWDGE, 2019;
Roussev, 2016;
Casey, 2011]

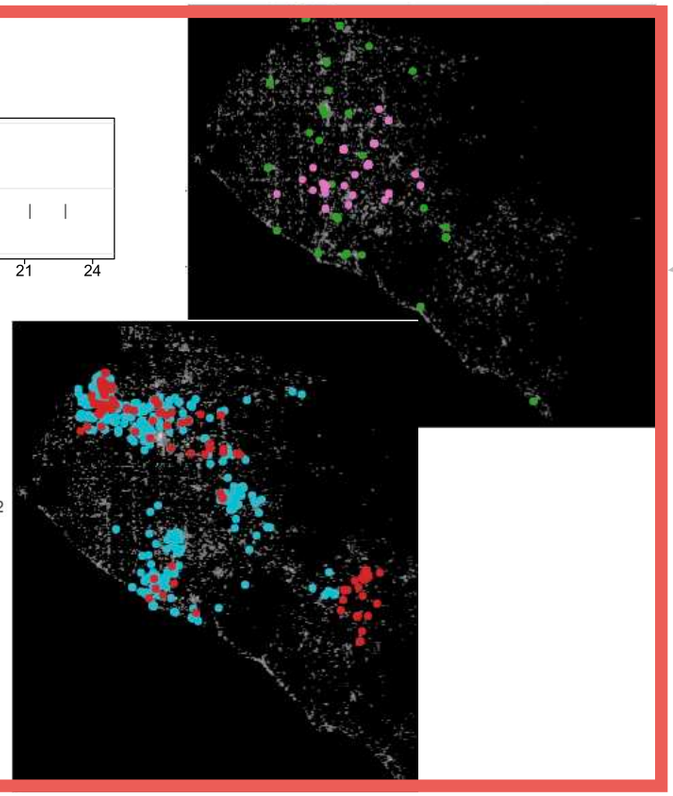
- Browser requests
- Web searches
- Email activity
- Phone/SMS
- Social media activity
- GPS locations
- File access
- Network activity
- Exercise/movement
- …

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



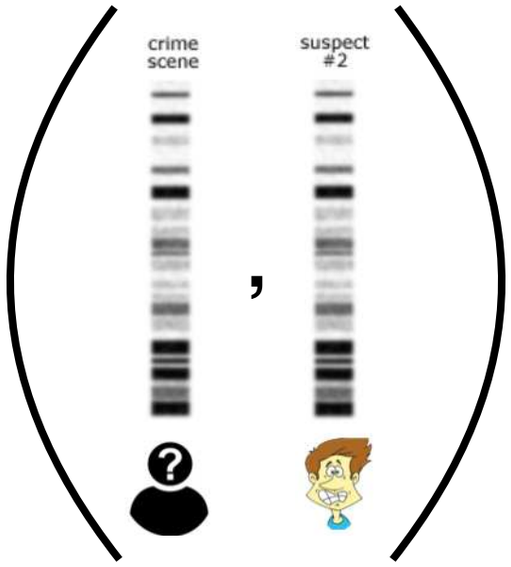


TOPIC OF
DFRWS PAPER

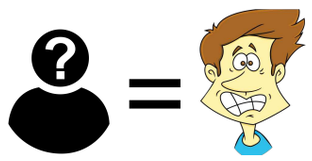
Statistical Approaches for Evaluating Forensic Evidence

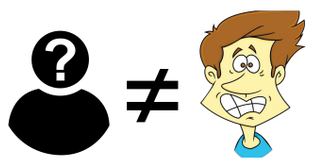
Goal

Assess the likelihood of observing



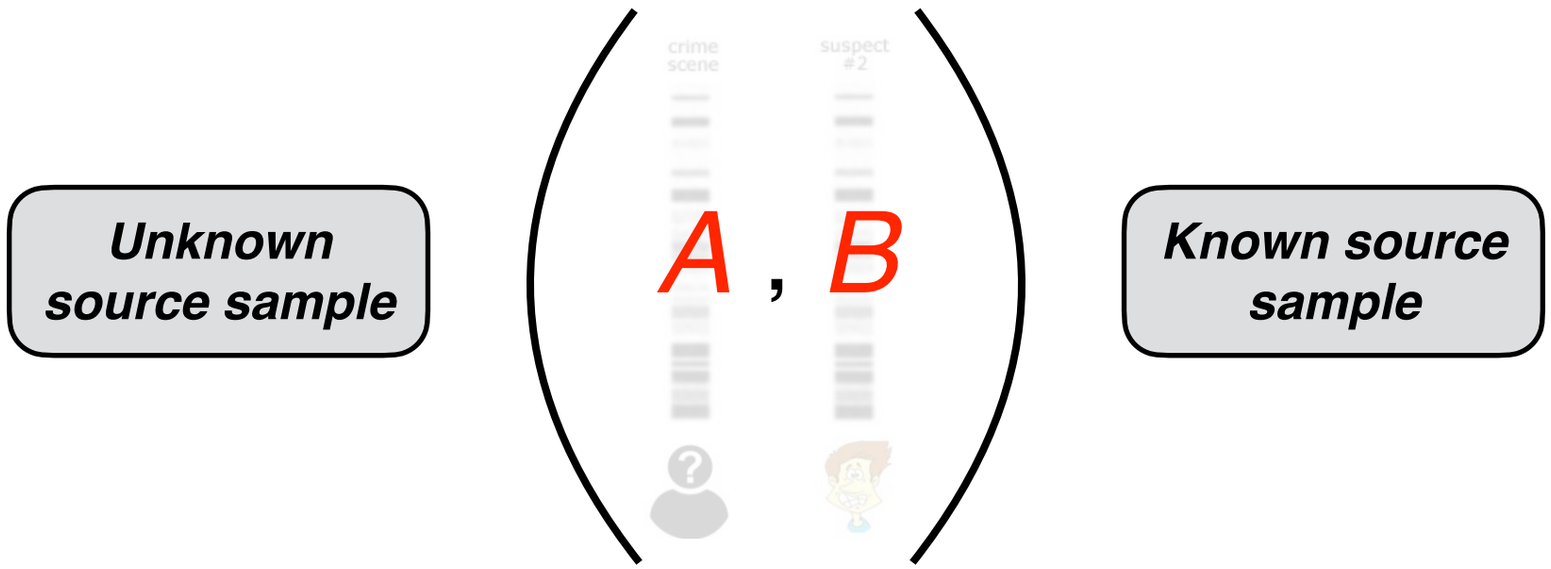
Under two competing hypotheses





Goal

Assess the likelihood of observing



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?

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

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

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

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

[Stern, 2017; Thompson, 2017]

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

[Stern, 2017; Thompson, 2017]

“Strength of Evidence”

“Weight of Evidence”

[Pierce, 1878]

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

[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]

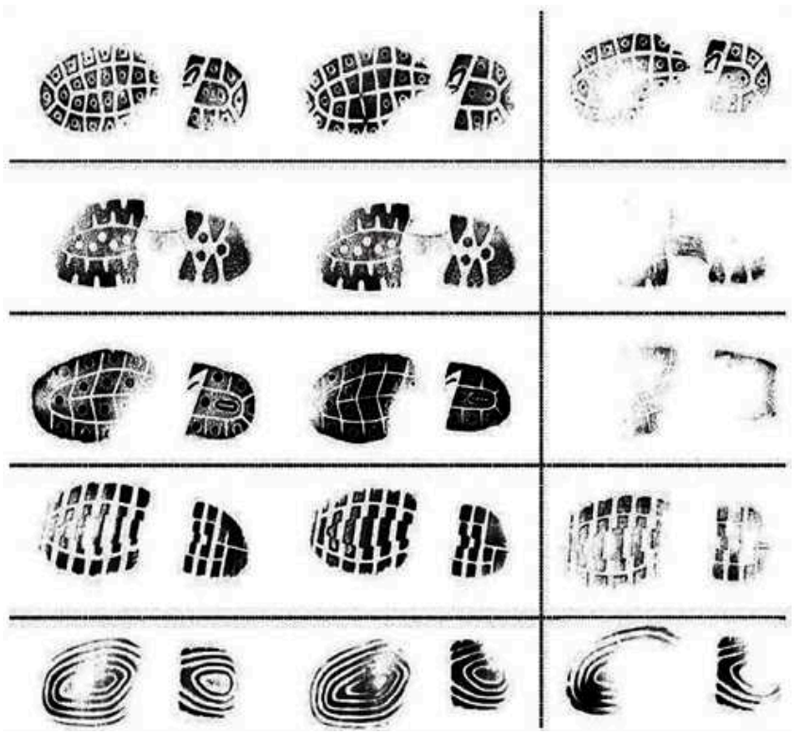
- Has been applied in a variety of forensic disciplines
  - → DNA [Aitken & Stoney, 1991; Evett & Weir, 1998; Steele & Balding, 2014]
  - → Fingerprints [Champod & Evett, 2001]
  - → Handwriting [Bozza et al., 2008]
  - → Speaker Recognition [Champod & Meuwly, 2000]

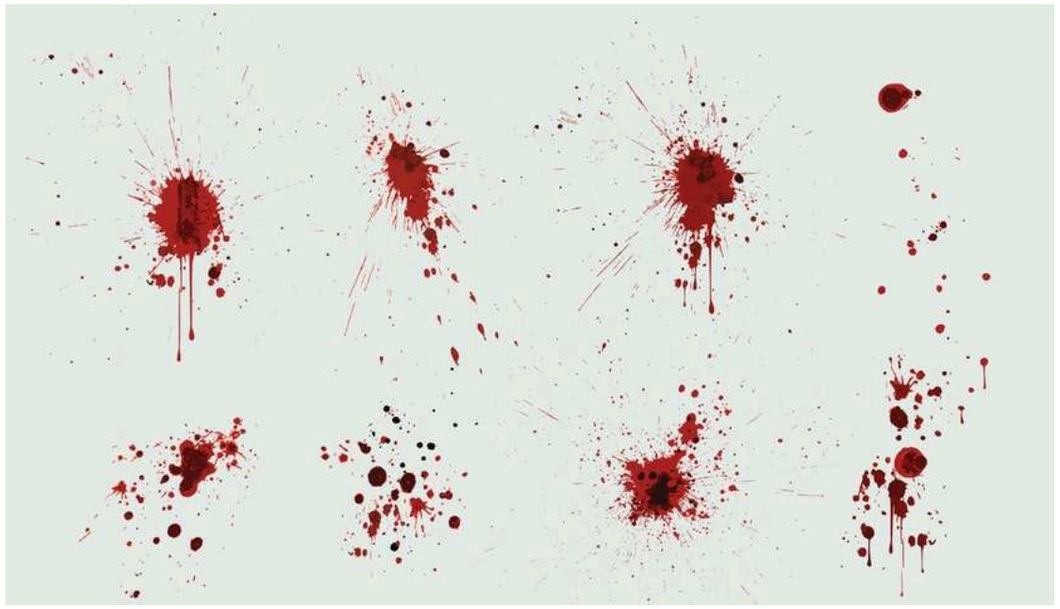
- Studies demonstrating its understanding
  - → Misconceptions [Martire et al., 2013, Thompson and Newman, 2015, Thompson et al., 2018]
  - → Verbal Equivalents [e.g., AFSP, 2009]

Why not always use the LR?

Why not always use the LR?

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



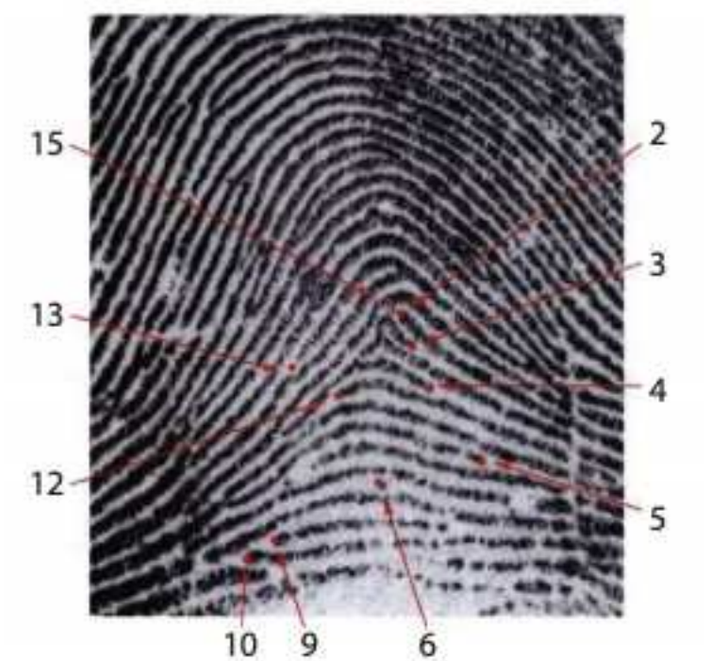


[Stern, 2017]

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*

$$\Delta(A, B)$$

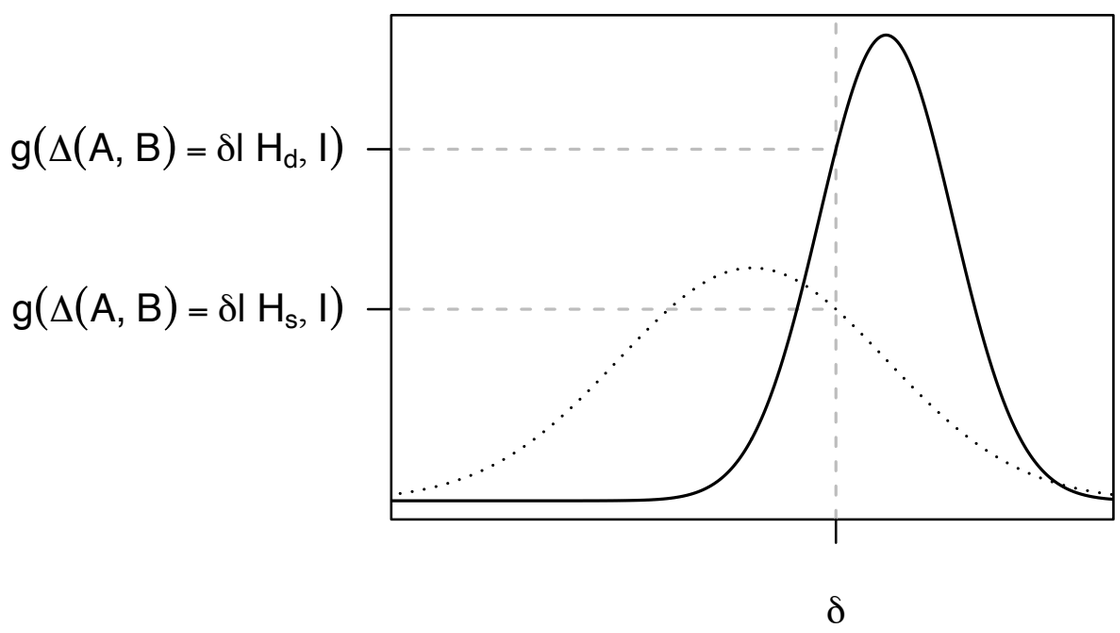
Score-based Approaches

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

$$\Delta(A, B)$$

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

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



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

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

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]

Evidence Evaluation Approaches

- **Likelihood Ratio:** Models evidence directly

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

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

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

Evidence Evaluation Approaches

- **Likelihood Ratio:** Models evidence directly

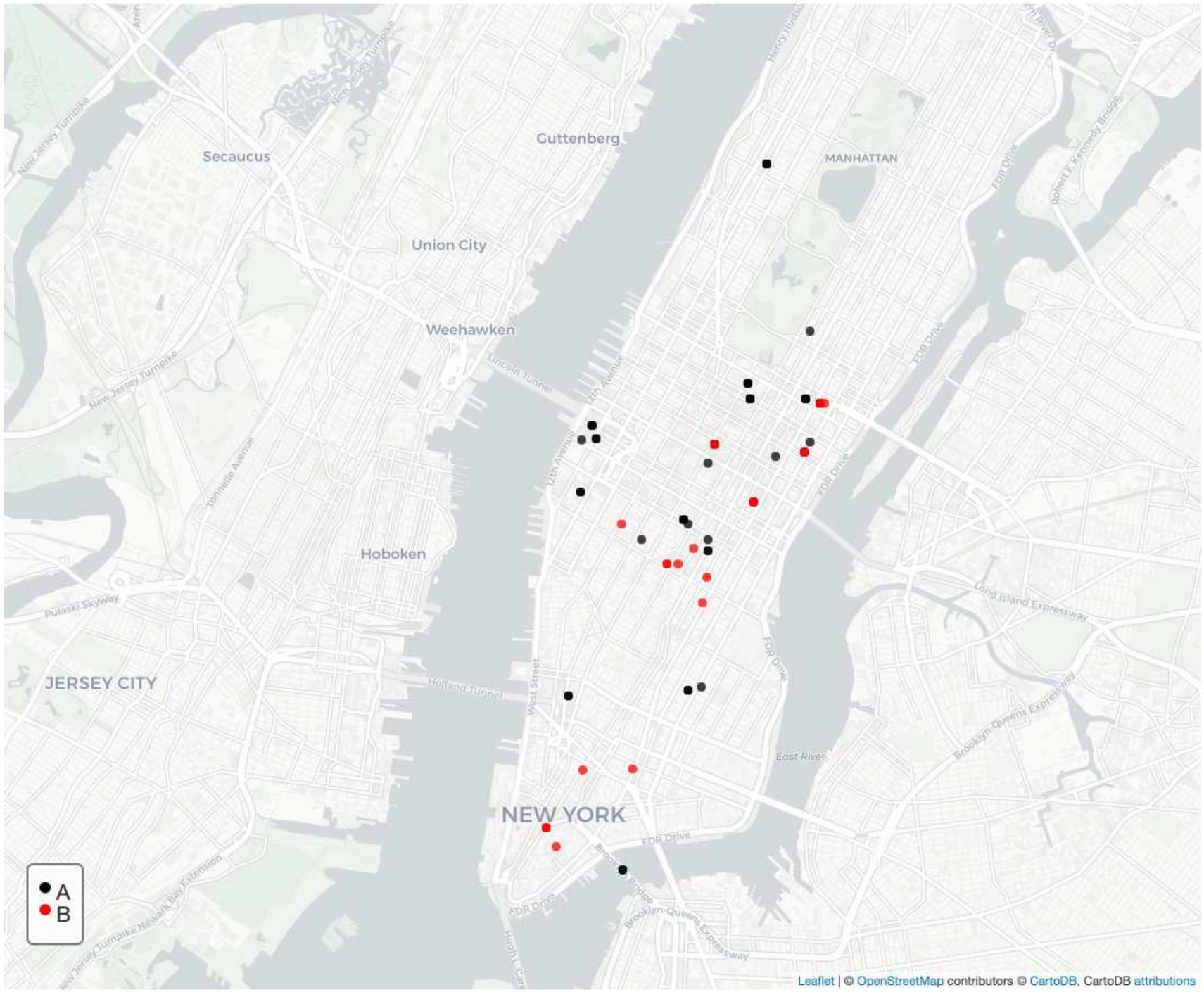
$$LR = \frac{Pr(A, B \mid H_s)}{Pr(A, B \mid H_d)}$$

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

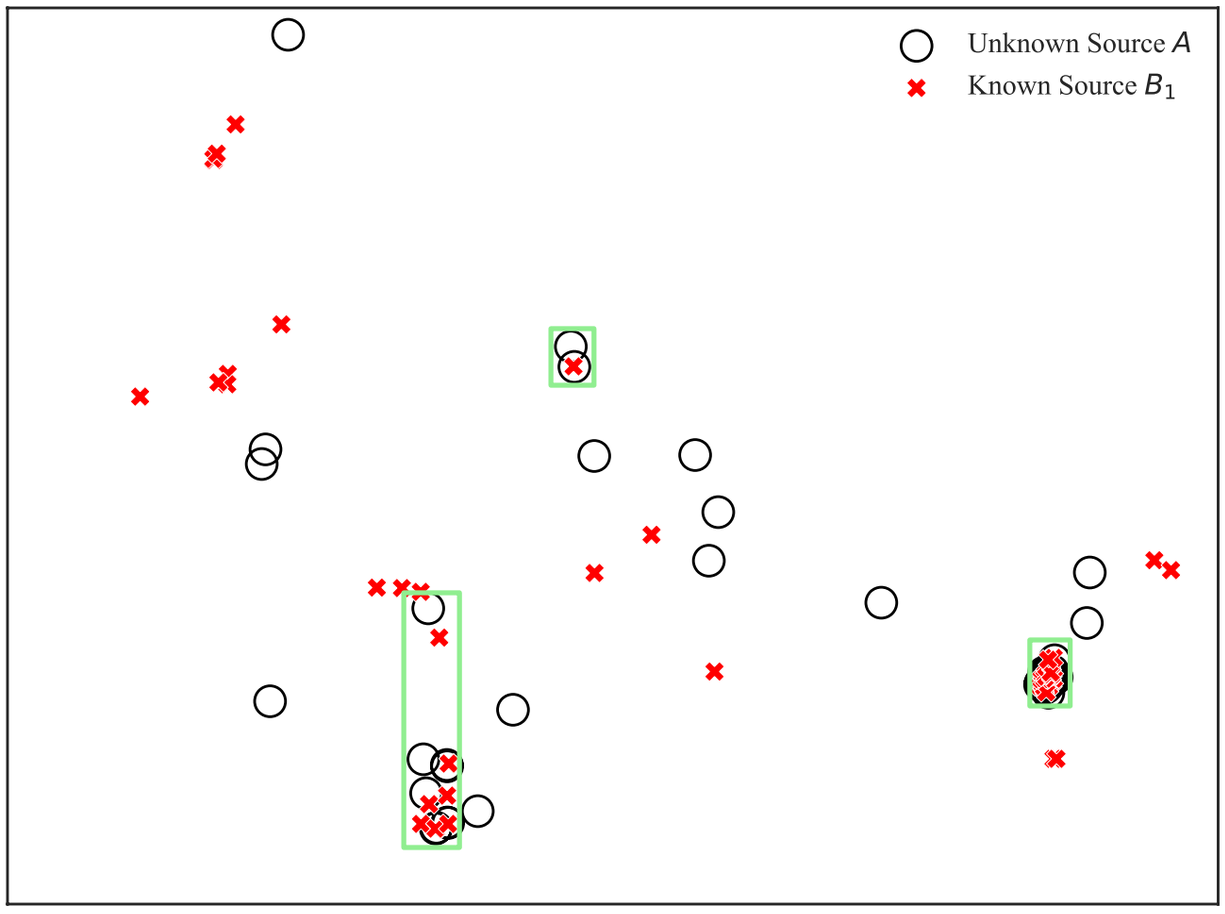
$$SLR_{\Delta} = \frac{g(\Delta(A, B) = \delta \mid H_s)}{g(\Delta(A, B) = \delta \mid H_d)}$$

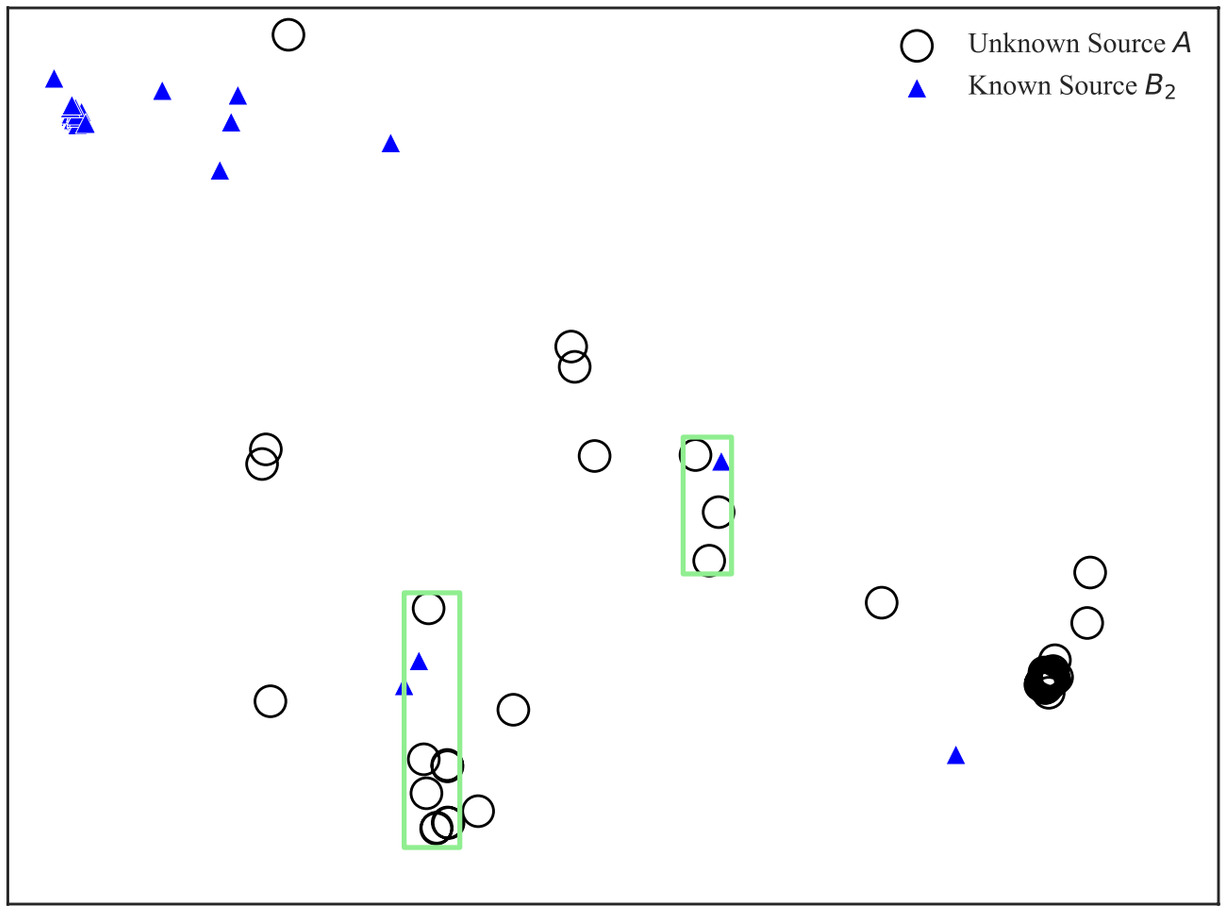
CONTRIBUTION

Quantifying the Strength of Geolocated Event Evidence


Secaucus
Guttenberg
MANHATTAN
Union City
Weehawken
Lincoln Tunnel
Hoboken
JERSEY CITY
Holland Tunnel
NEW YORK
East River
FDR Drive
Robert F. Kennedy Bridge
Brooklyn-Queens Expressway
Long Island Expressway
Pulaski Skyway
New Jersey Turnpike
Tonnelle Avenue
12th Avenue
West Street
A
B
Leaflet | © OpenStreetMap contributors © CartoDB, CartoDB attributions

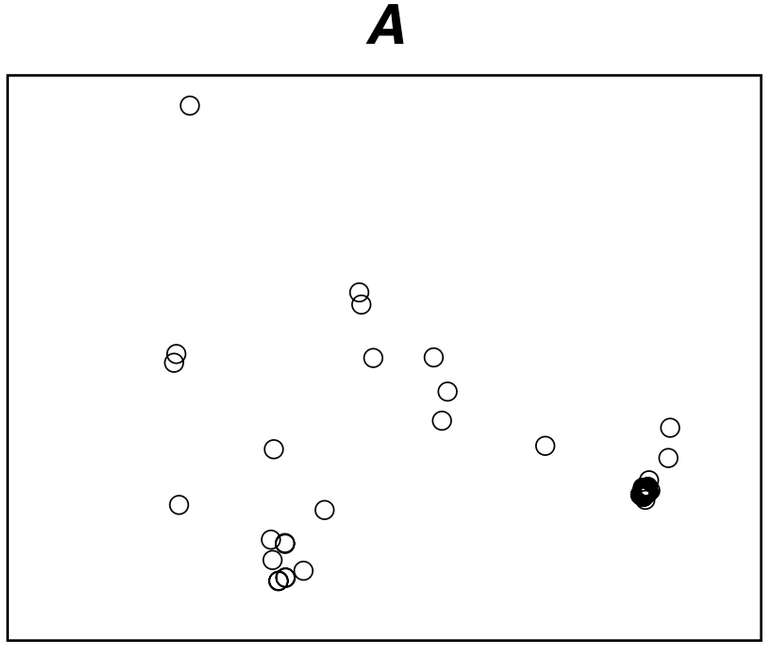
Geofence Warrants



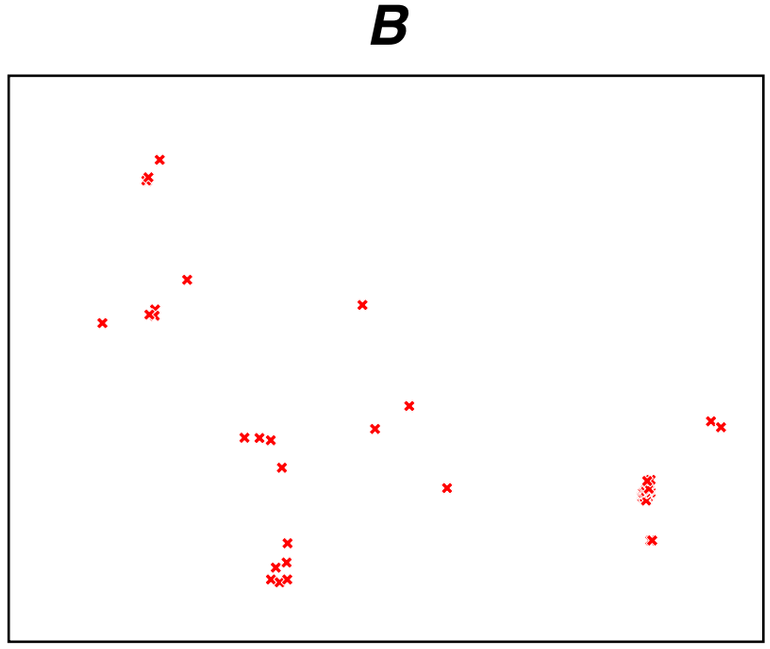


Revisiting the LR

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



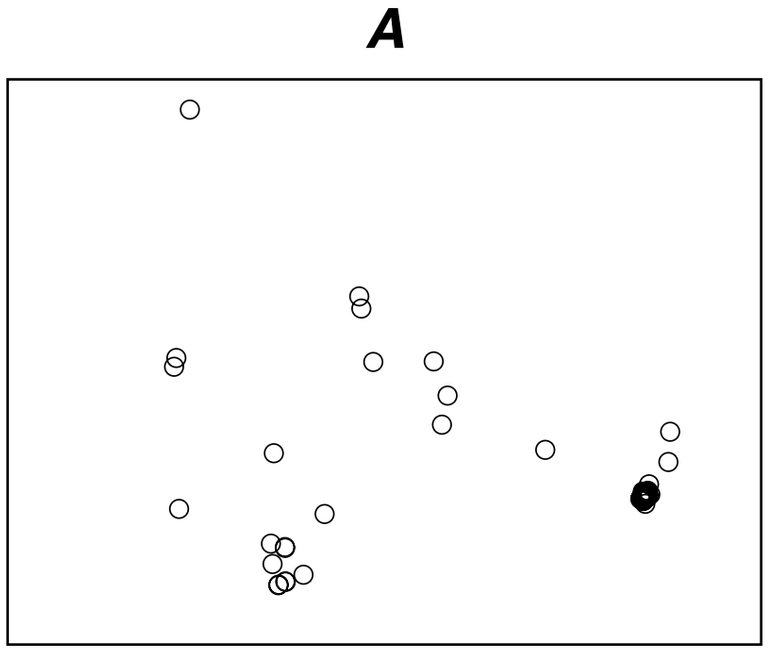
Unknown source events



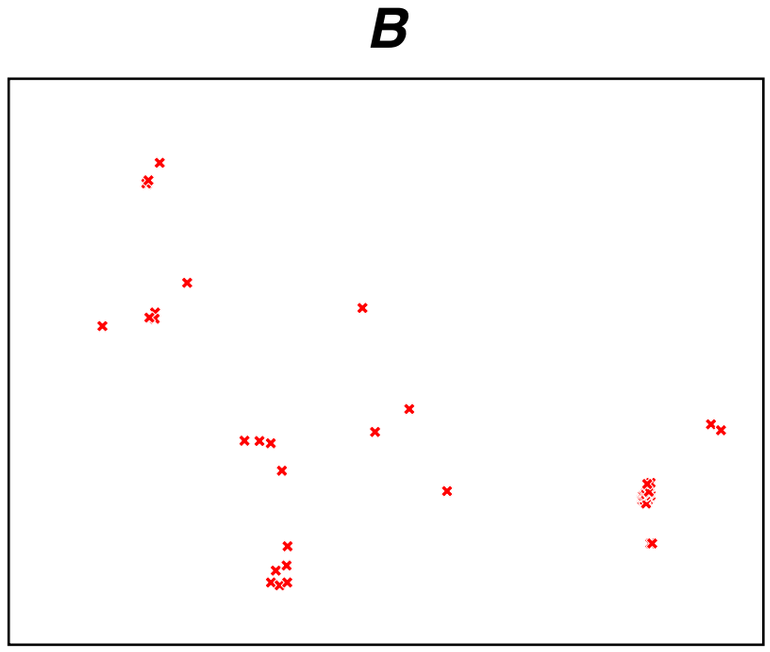
Known source events

Revisiting the LR

$$LR = \frac{Pr(A, B \mid H_s)}{Pr(A, B \mid H_d)} \Rightarrow \quad \underset{\textbf{Maths}}{\cdots} \quad \Rightarrow LR = \frac{f(B \mid A, H_s)}{f(B \mid H_d)}$$



Unknown source events

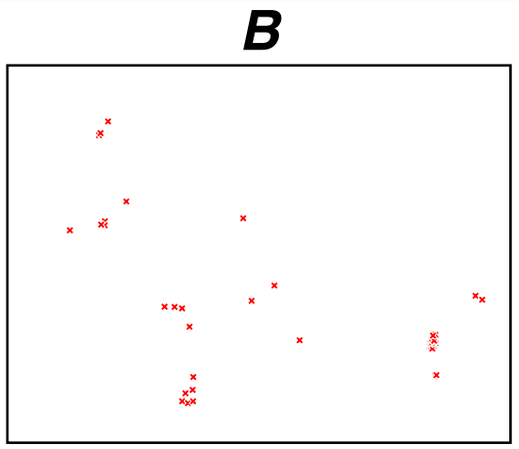


Known source events

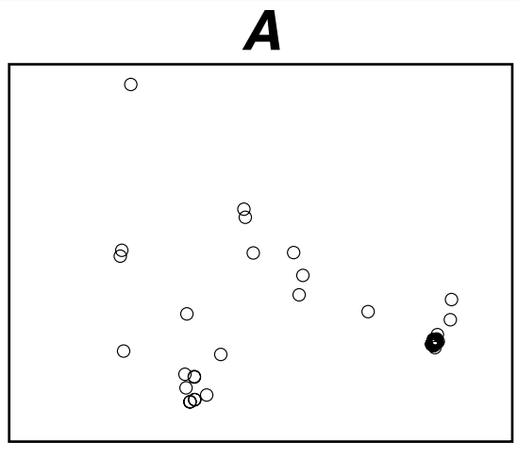
Revisiting the LR

$$LR = \frac{Pr(A, B \mid H_s)}{Pr(A, B \mid H_d)} \Rightarrow \quad \underset{\text{Maths}}{\cdots} \quad \Rightarrow LR = \frac{f(B \mid A, H_s)}{f(B \mid H_d)}$$

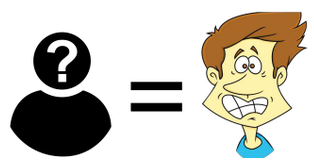
$$LR = \frac{\text{Prob of } \boldsymbol{B} \text{ given } \boldsymbol{A} \text{ \& } \text{?} = \text{}}{\text{Prob of } \boldsymbol{B} \text{ given } \text{?} \neq \text{}}$$

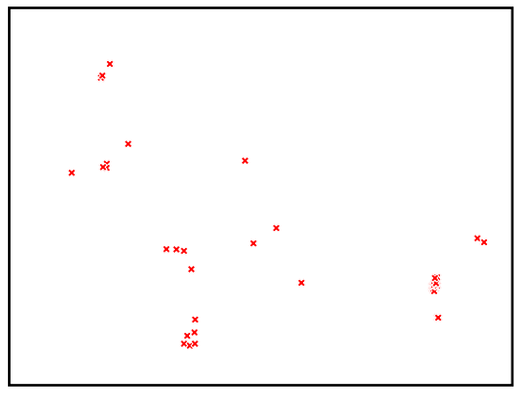


Known source events

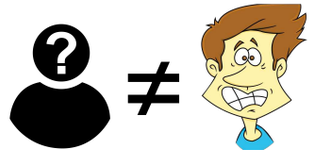


Unknown source events



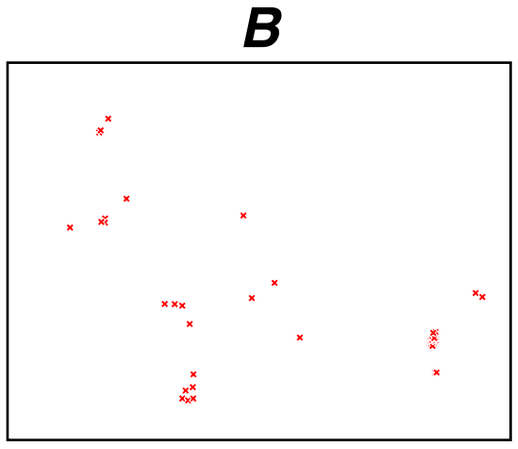


Known source events



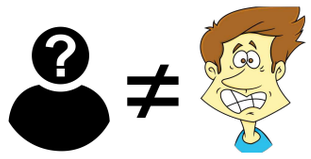
Random match probability

Prob of



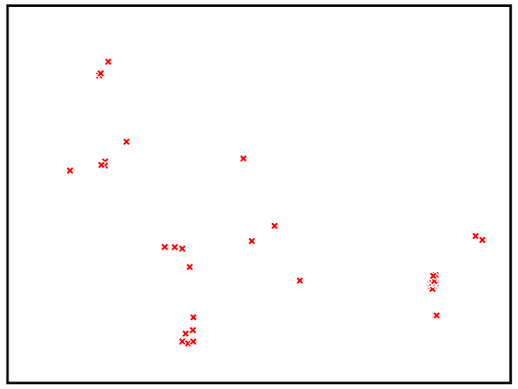
Known source events

given



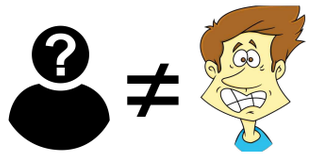
Random match probability

Prob of

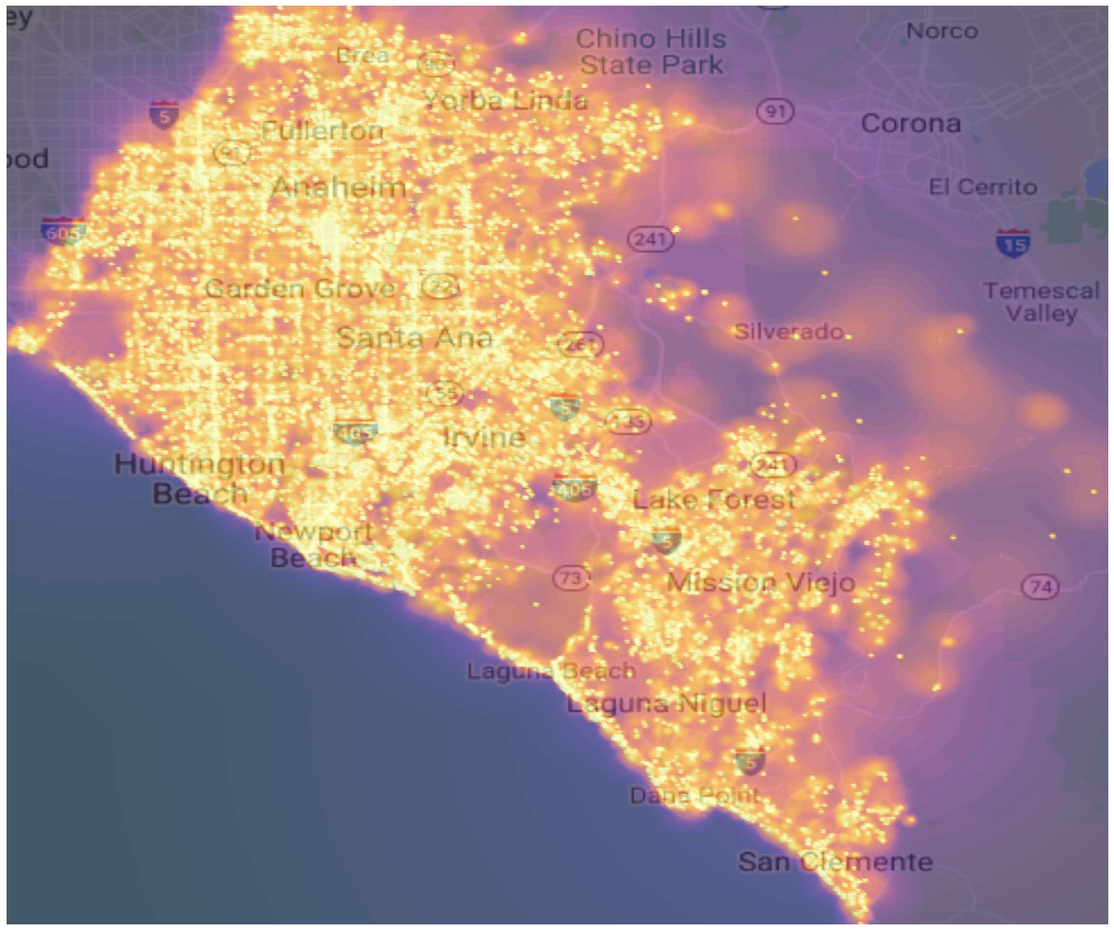


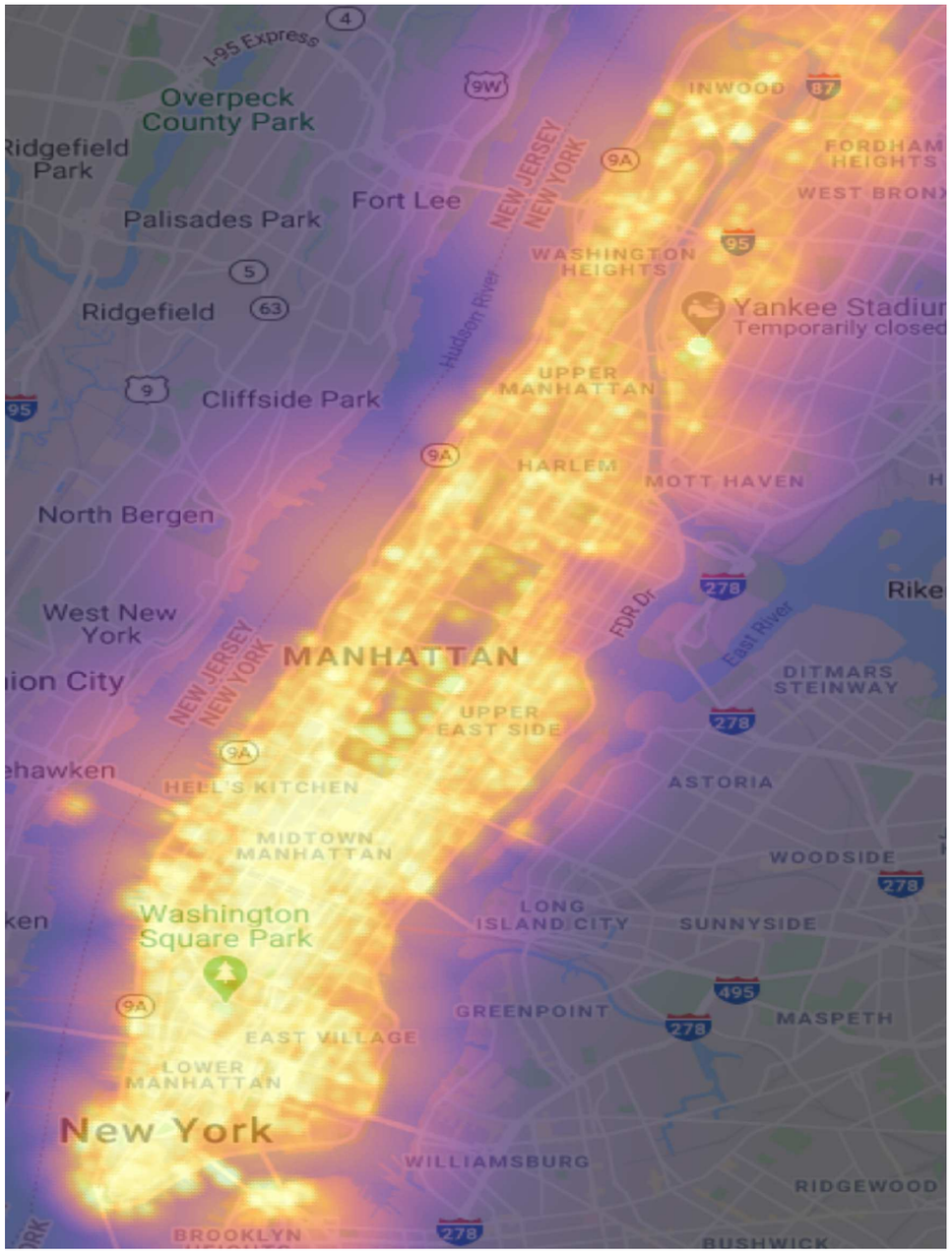
Known source events

given



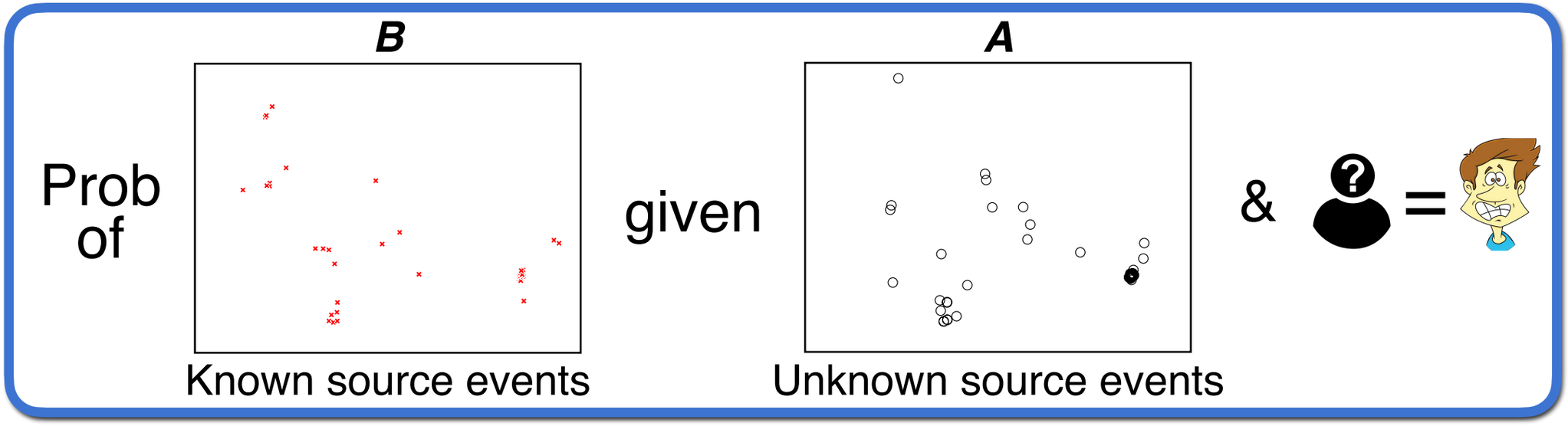
Random match probability



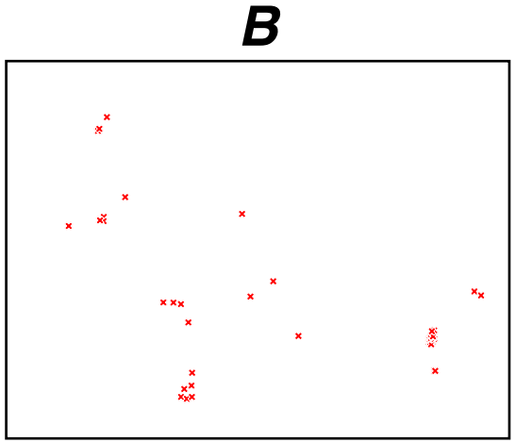


Adaptive Bandwidth Kernel Density Estimators

[Breiman et al., 1977]

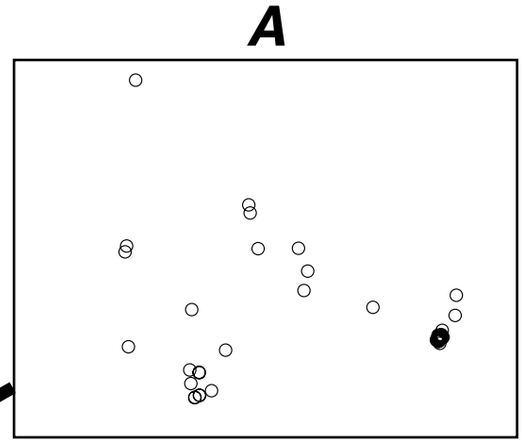

Prob of
B
Known source events
given
A
Unknown source events
&
=

Prob of



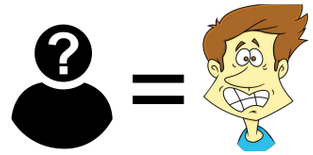
Known source events

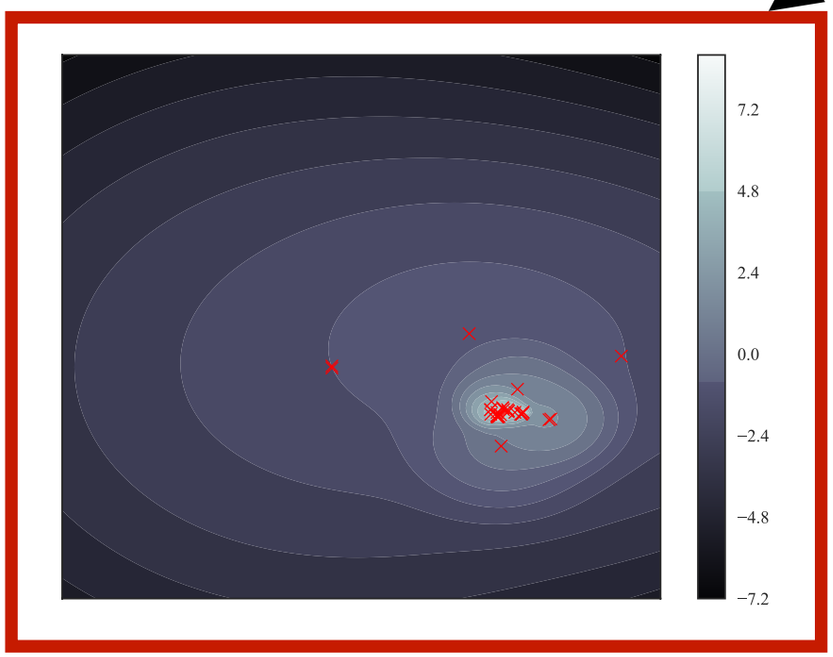
given



Unknown source events

&

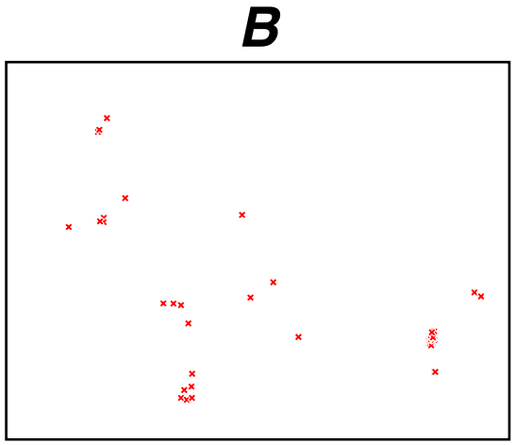




Individual Component

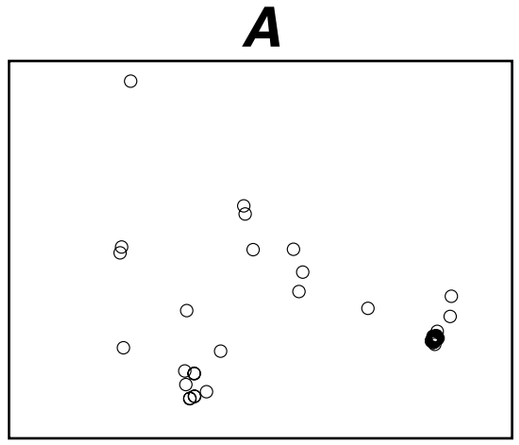
[Lichman & Smyth, 2014]

Prob of



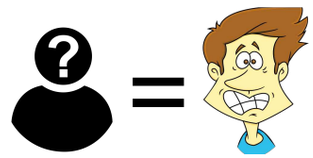
Known source events

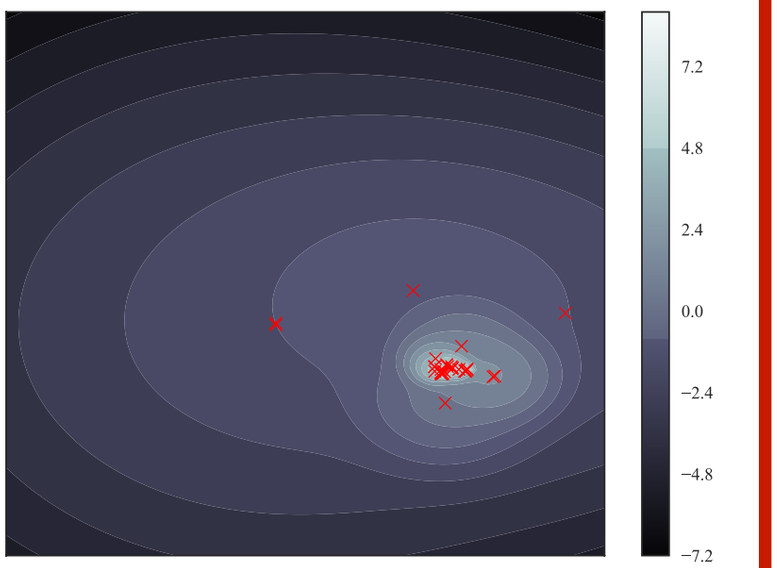
given



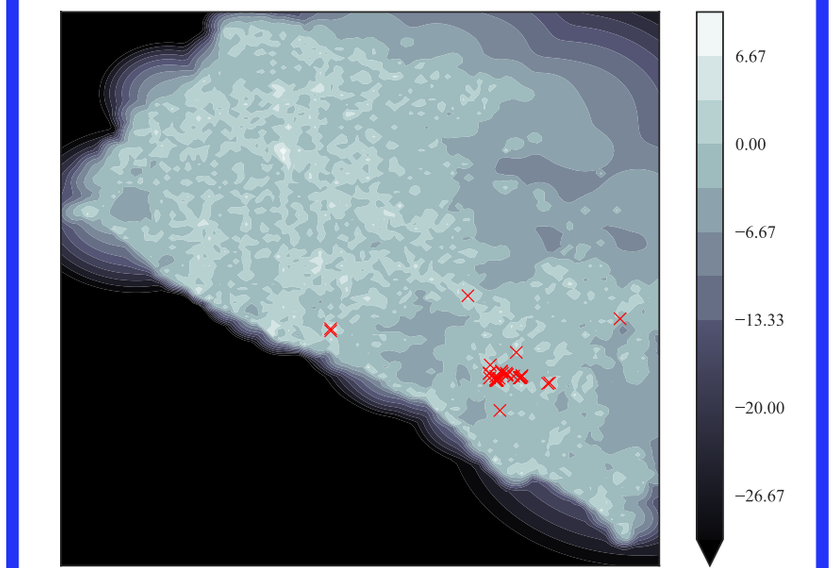
Unknown source events

&





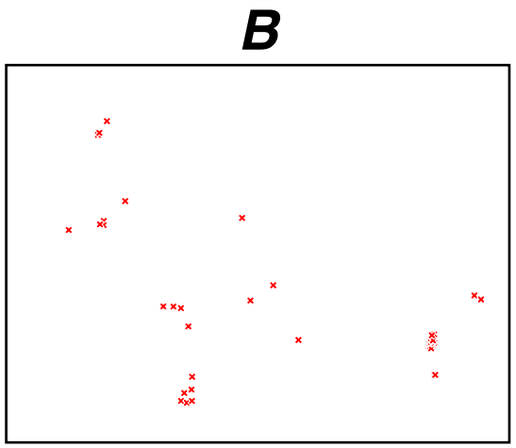
Individual Component



Population Component

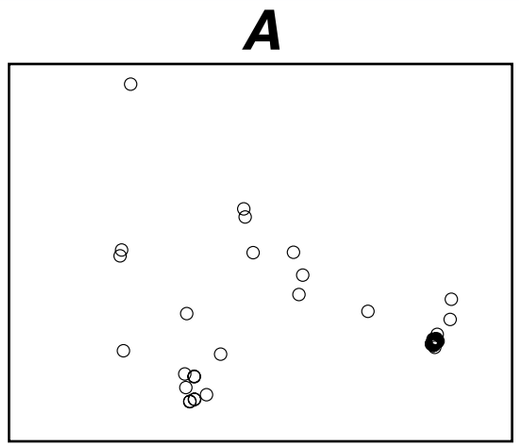
[Lichman & Smyth, 2014]

Prob of



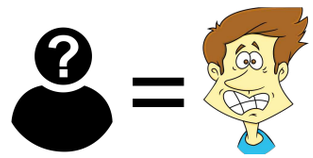
Known source events

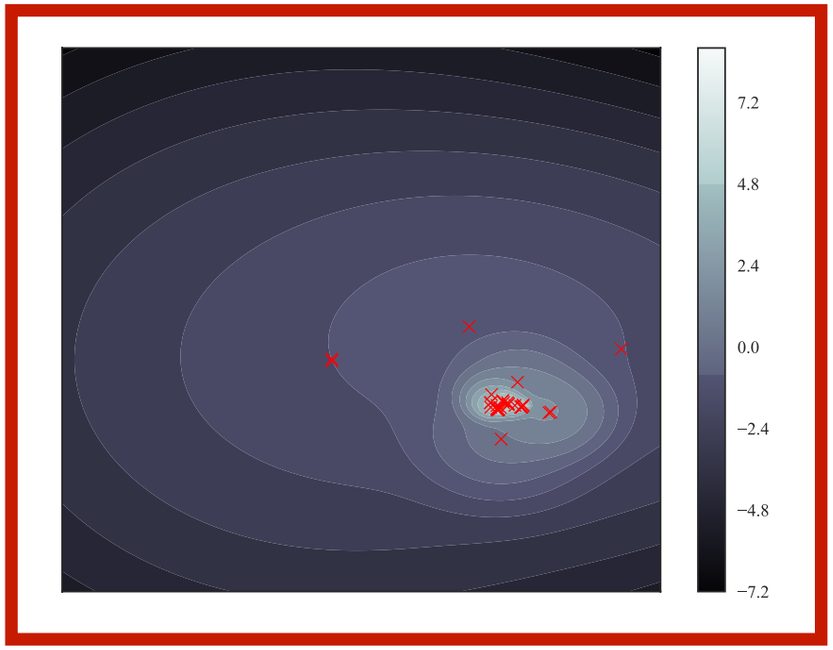
given



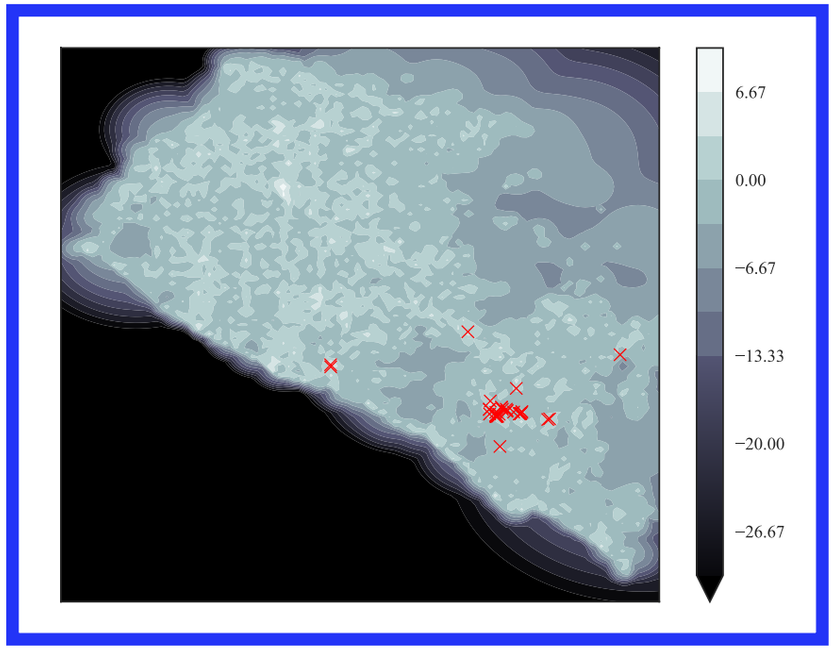
Unknown source events

&

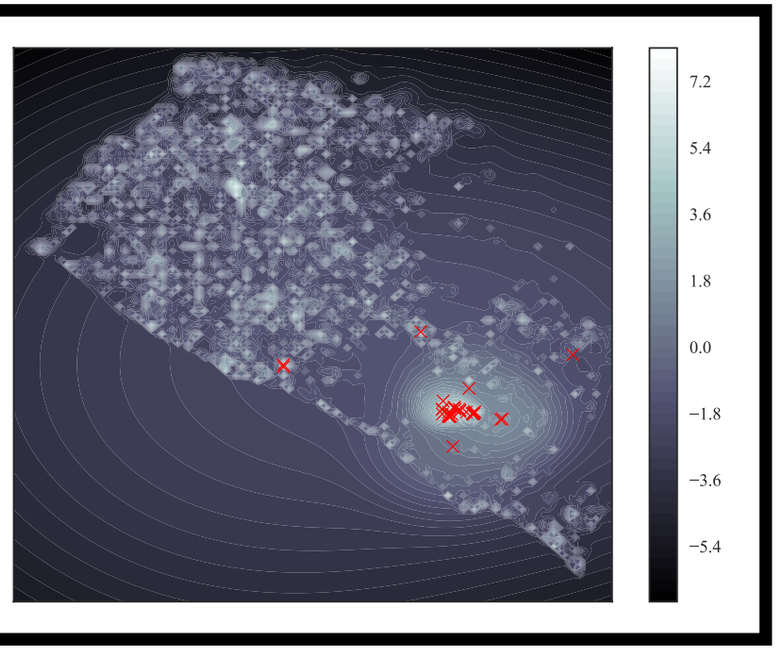




Individual Component



Population Component



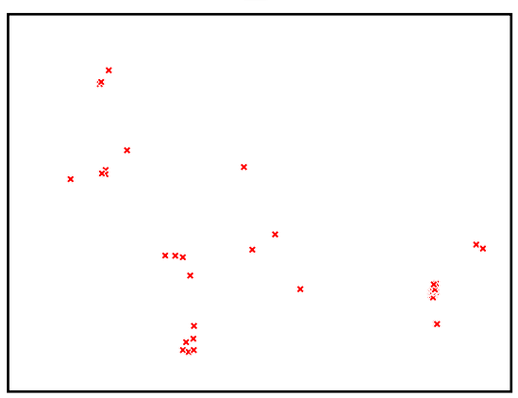
Mixture

80% individual,
20% population

[Lichman & Smyth, 2014]

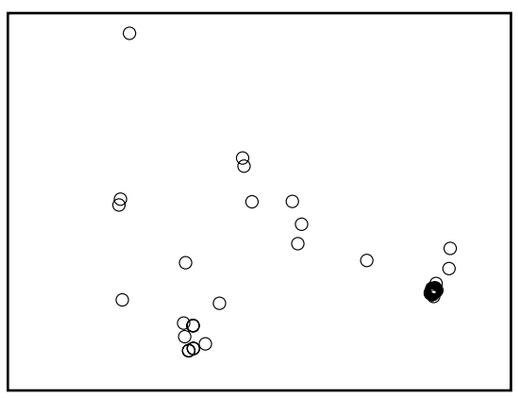
$$LR = \frac{}{}$$

Prob of



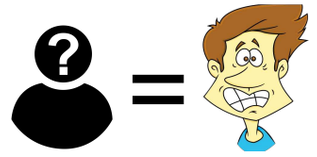
Known source events

given

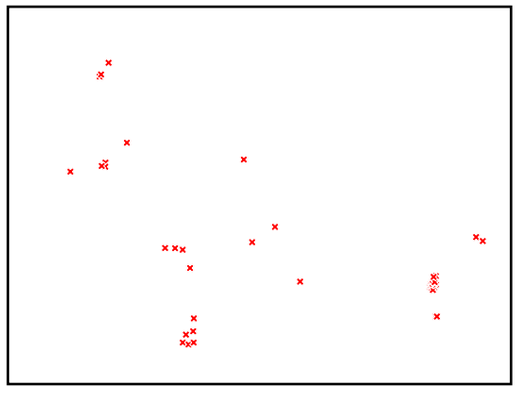


Unknown source events

&

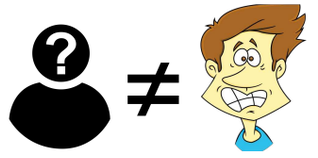


Prob of



Known source events

given

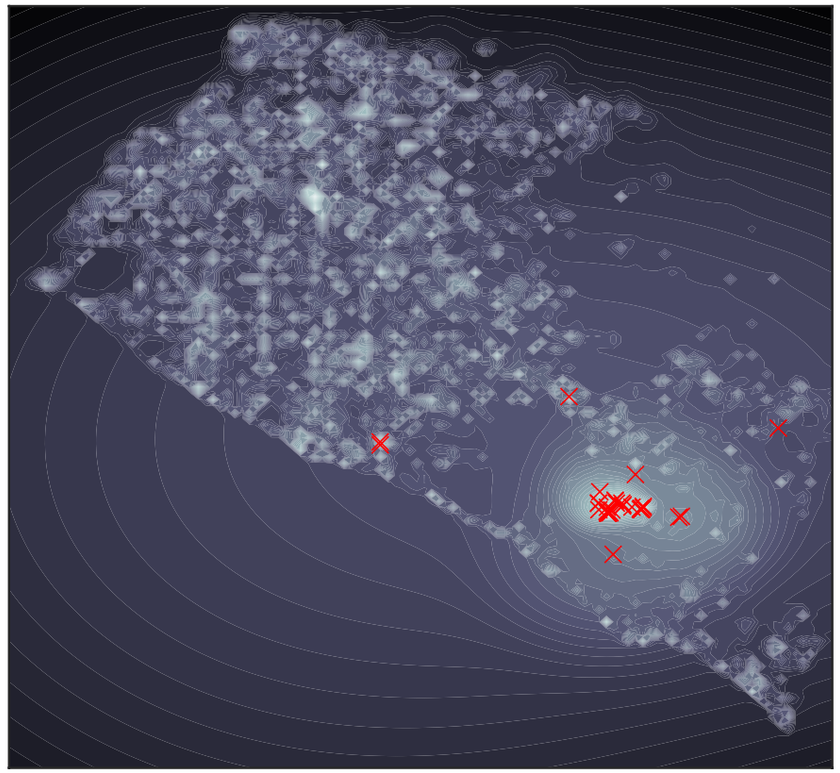


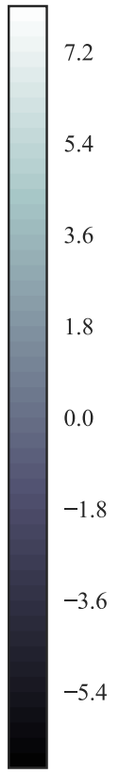
Random match probability

$$LR = \frac{}{}$$

Prob of

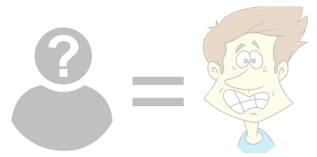
Known so



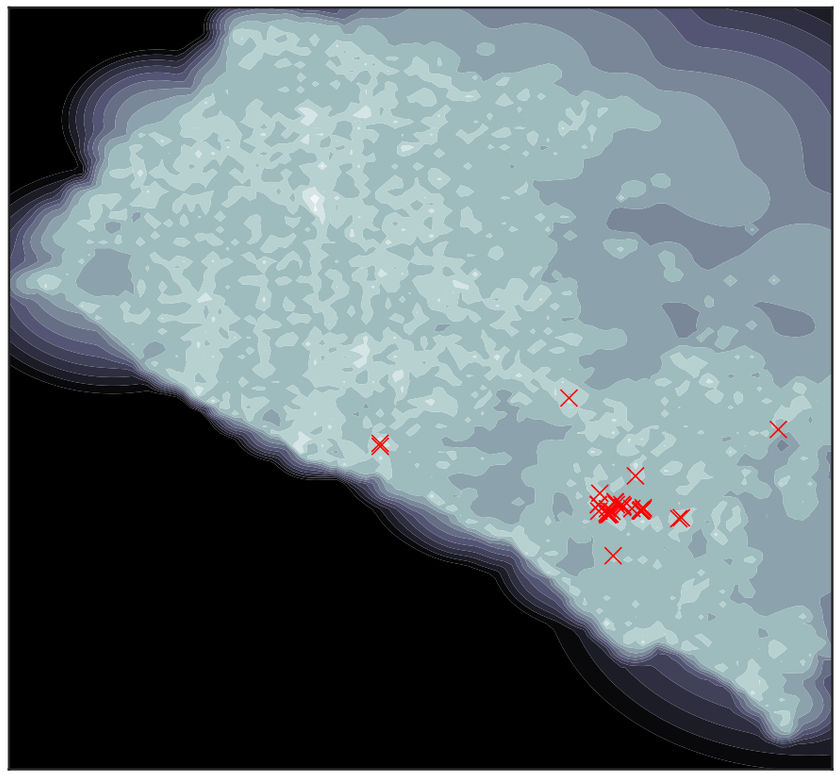


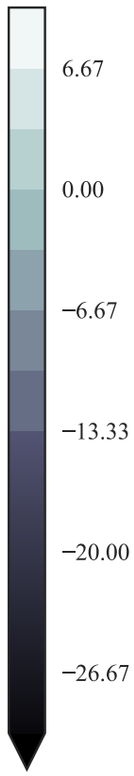
ts

&



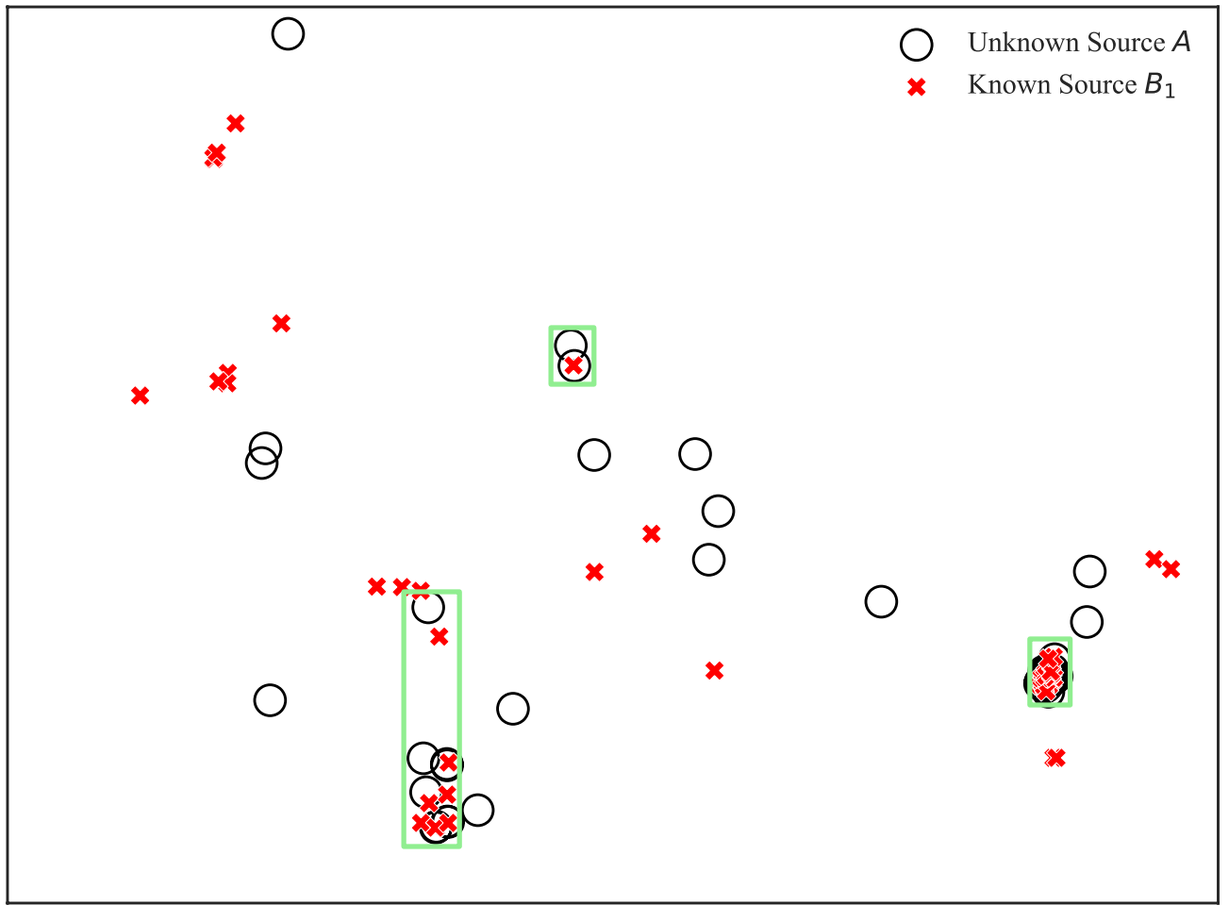
Prob of





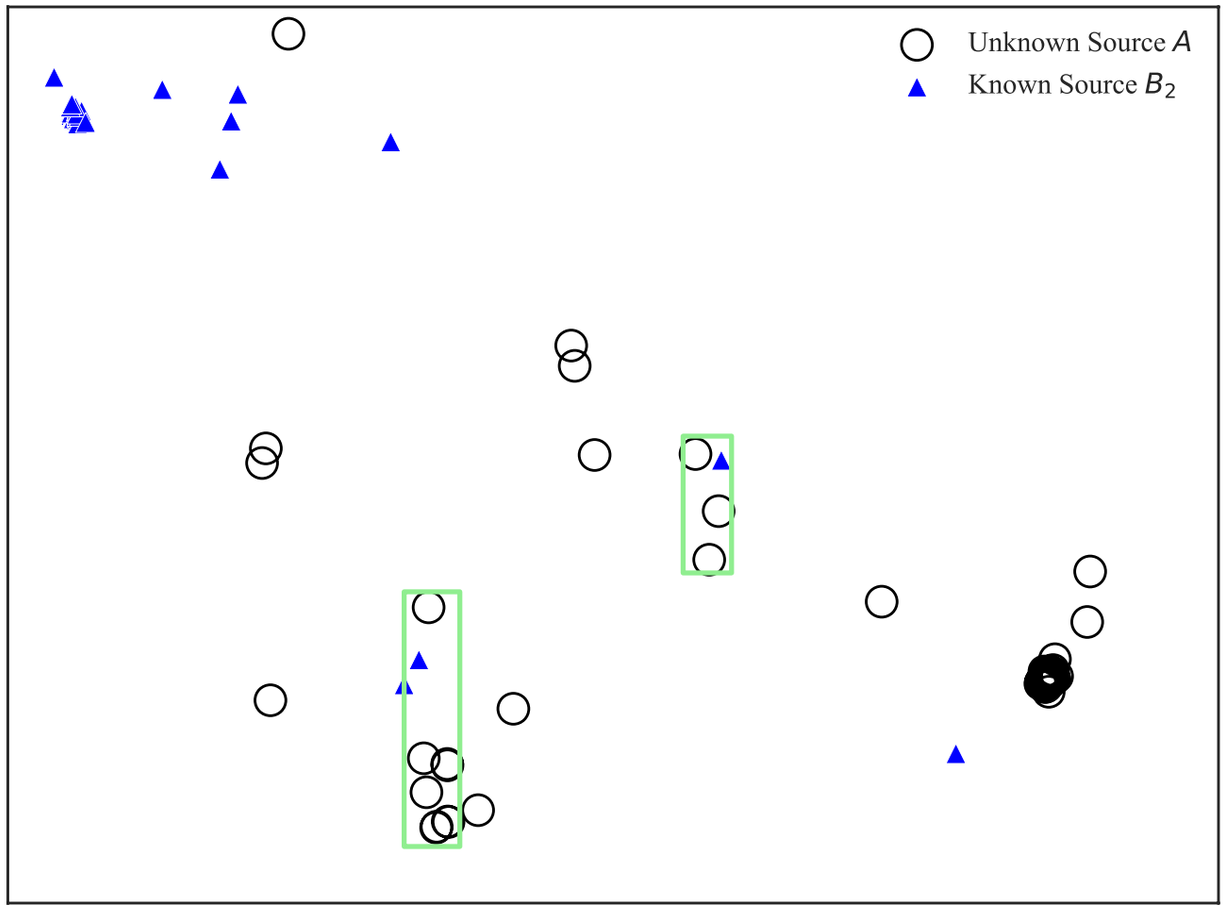
match
bility

Geofence Warrants Revisited



$$LR \approx 1137$$

Strong support for same-source hypothesis



$$LR \approx 2.8 \times 10^{-28}$$

Exclusion

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

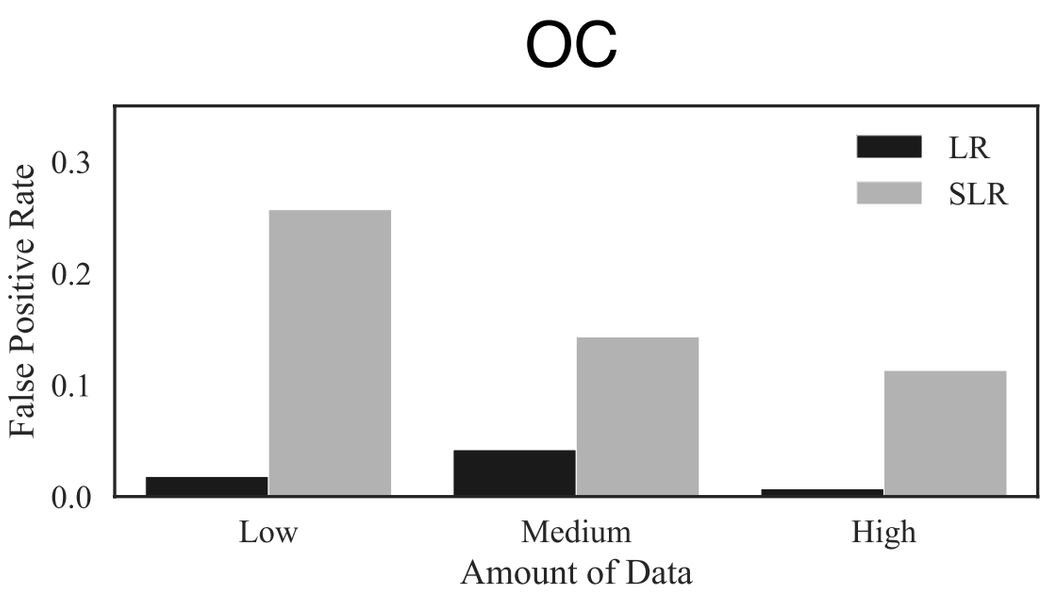
Region	Method[1]	TP Rate[2]	FP Rate[2]	AUC
OC	LR	0.380	**0.038**	**0.845**
	SLR	**0.614**	0.162	0.783
NY	LR	0.285	**0.089**	**0.768**
	SLR	**0.511**	0.235	0.685

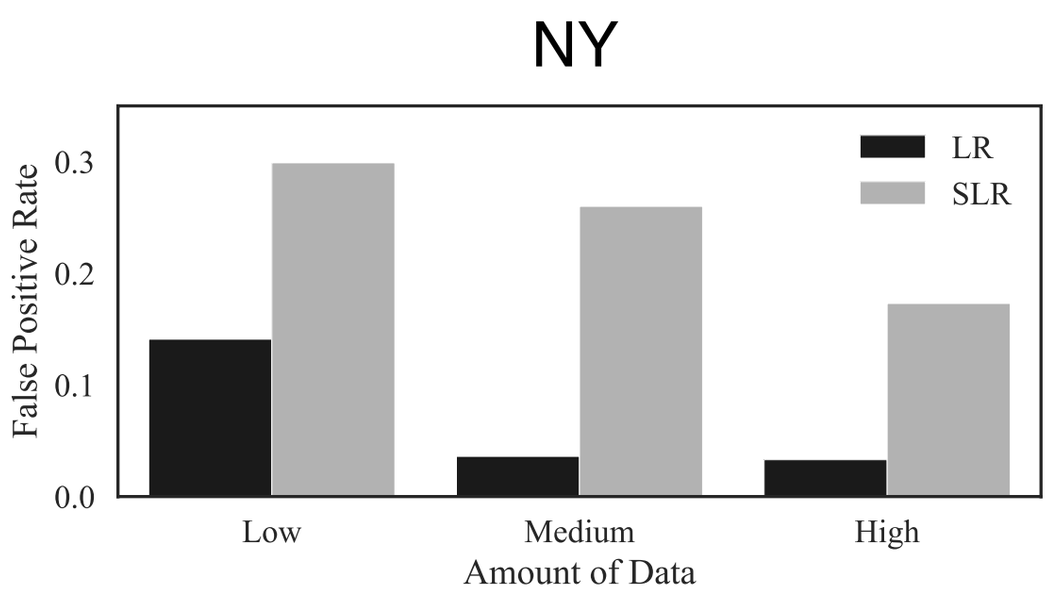
(1) LR with $\alpha(n_a)$ weights; SLR using earth mover's distance with account weights
(2) LR & SLR threshold is 1

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
NY	LR	0.285	**0.089**	**0.768**
	SLR_{EMD}	**0.511**	0.235	0.685

(1) LR with $\alpha(n_a)$ weights; SLR using earth mover's distance with account weights
(2) LR & SLR threshold is 1





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.

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

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.

- [ ] Presented a framework for estimating **likelihood ratios** for forensic applications with **geolocated event data.**

Questions

Appendix

Score Functions

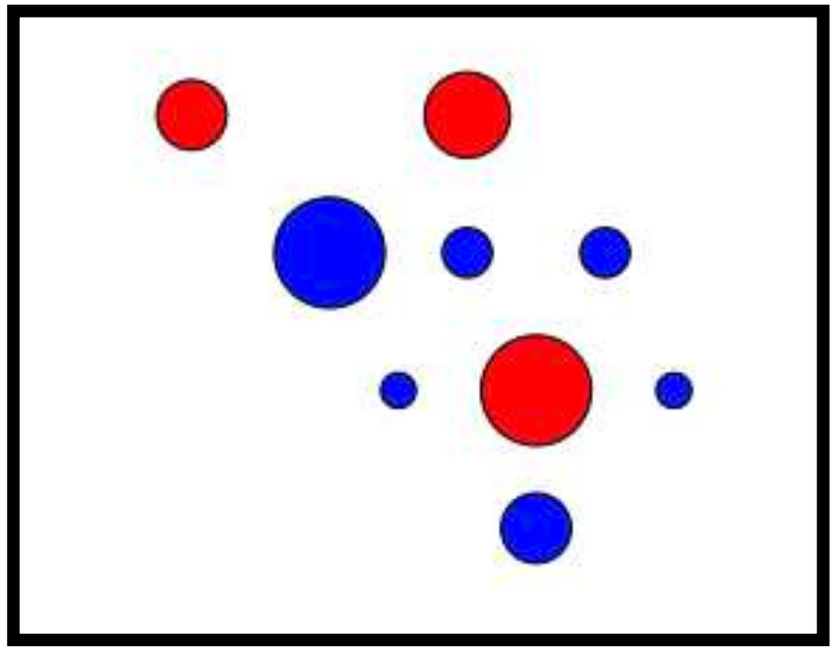
- Techniques to characterize spatial point patterns generally fall into two categories [Haggett, 1977]
  - Distance-based
  - Area-based

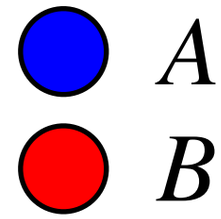
- Use distance-based score functions $\Delta(A, B)$ to quantify the similarity of the points within the sets A and B

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

Earth-mover's distance

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

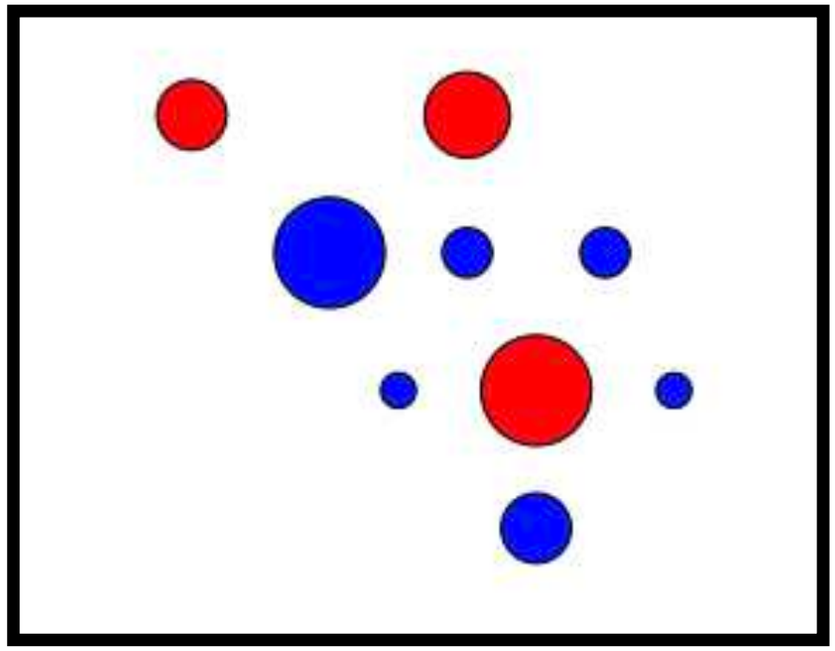


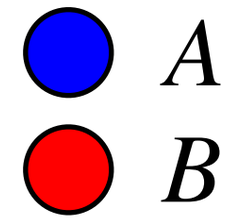


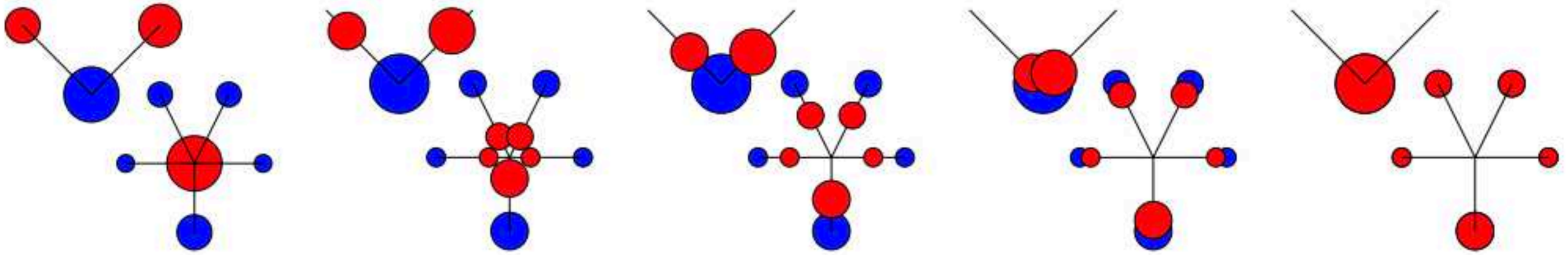
[Cohen, 1999]

Earth-mover's distance

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







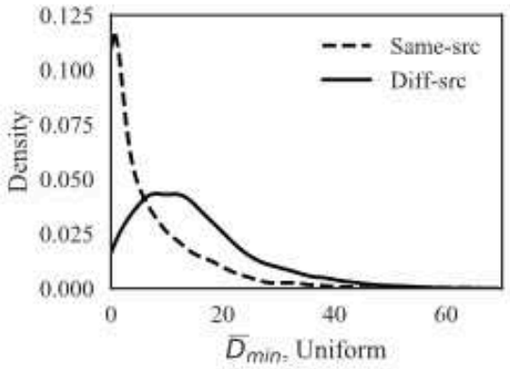
[Cohen, 1999]

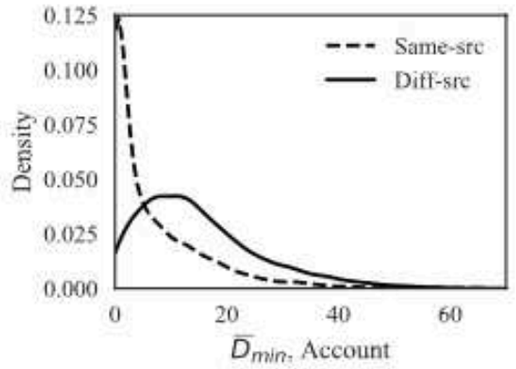
[Lichman, 2017]

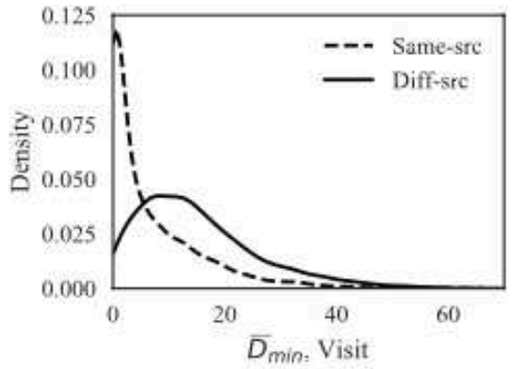
Number of Points ↑

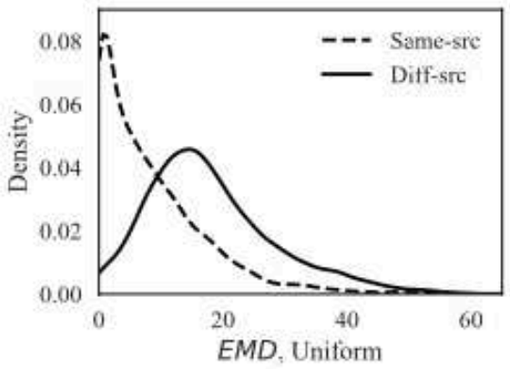
Weight ↓

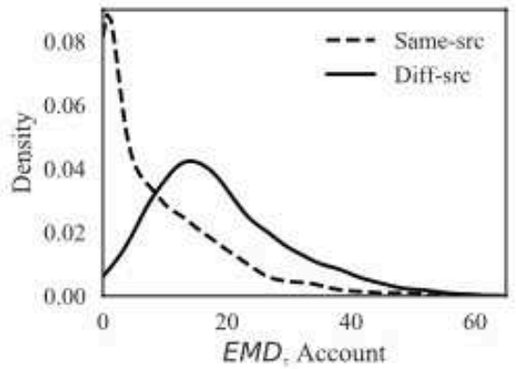
OC

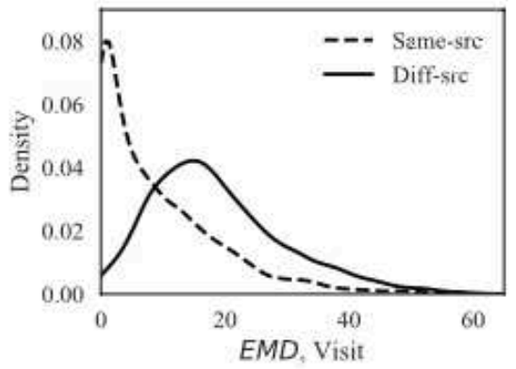

Same-src
Diff-src
Density
$\overline{D}_{min}$, Uniform


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

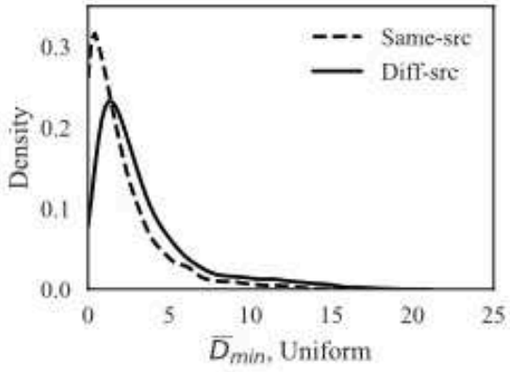

Same-src
Diff-src
Density
$\overline{D}_{min}$, Visit

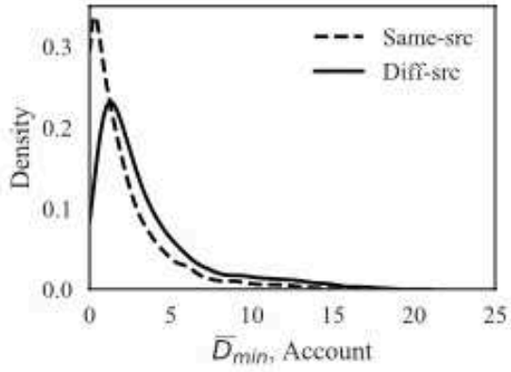

Same-src
Diff-src
Density
EMD, Uniform

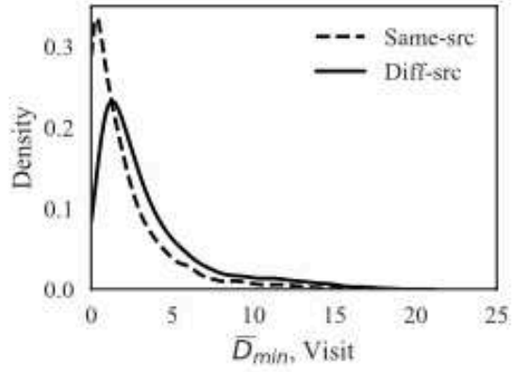

Same-src
Diff-src
Density
EMD, Account

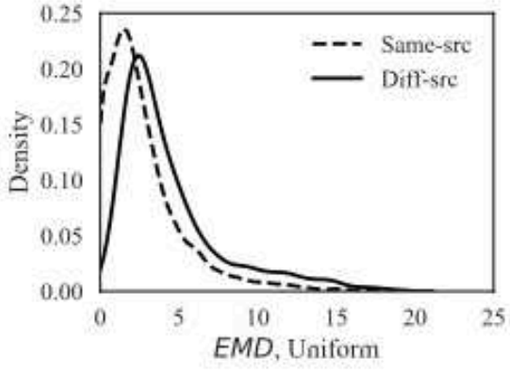

Same-src
Diff-src
Density
EMD, Visit

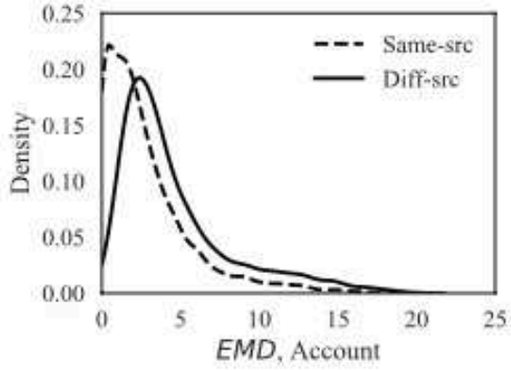
NY

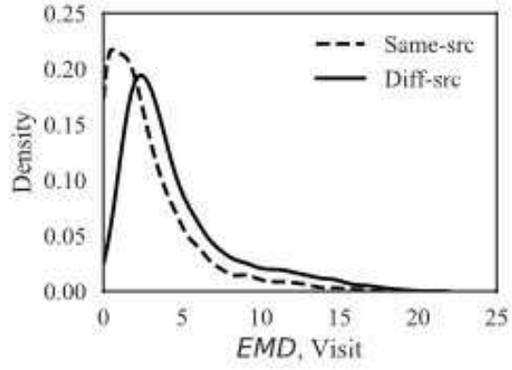

Same-src
Diff-src
Density
$\overline{D}_{min}$, Uniform


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


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


Same-src
Diff-src
Density
EMD, Uniform


Same-src
Diff-src
Density
EMD, Account

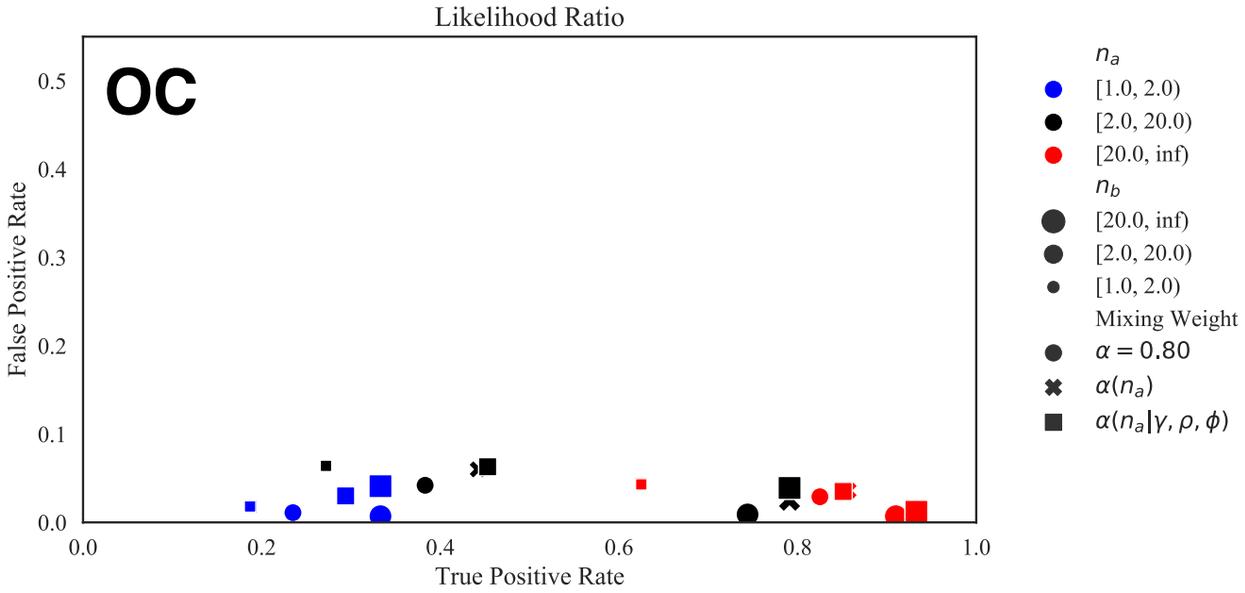

Same-src
Diff-src
Density
EMD, Visit

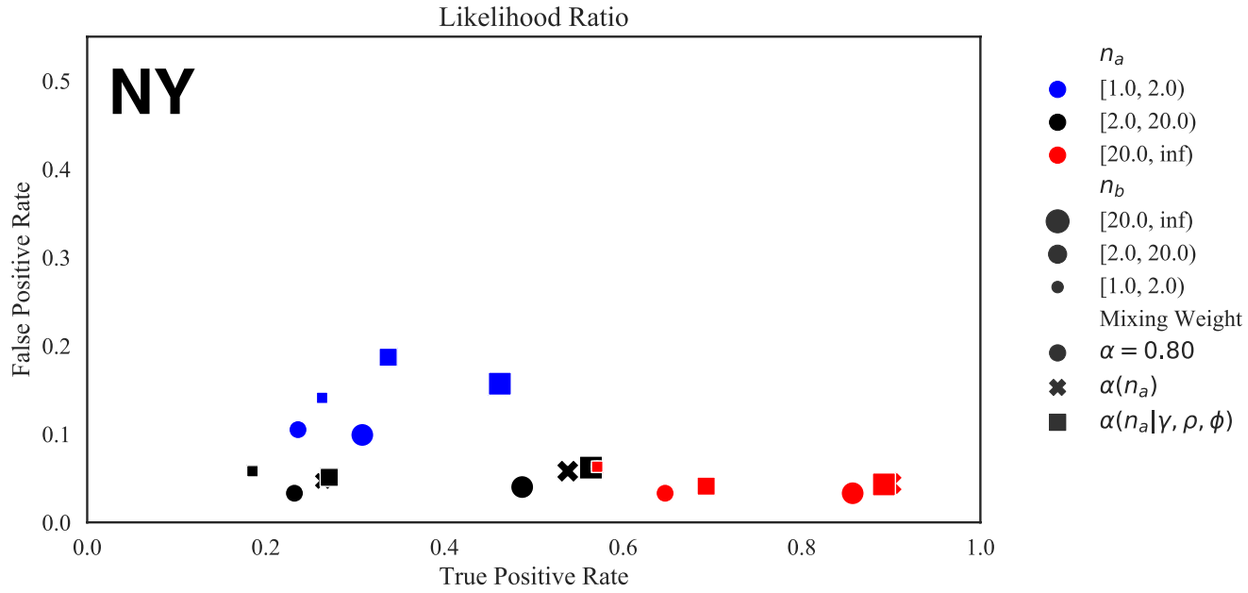
Region	Weight	TP@1	FP@1	AUC
OC	0.80	0.340	**0.026**	0.787
	$\alpha(n_a)$	**0.380**	0.038	**0.845**
	$\alpha(n_a \vert \gamma, \rho, \phi)$	0.375	0.037	0.817
NY	0.80	0.251	**0.067**	0.711
	$\alpha(n_a)$	**0.285**	0.089	**0.768**
	$\alpha(n_a \vert \gamma, \rho, \phi)$	0.282	0.088	0.734

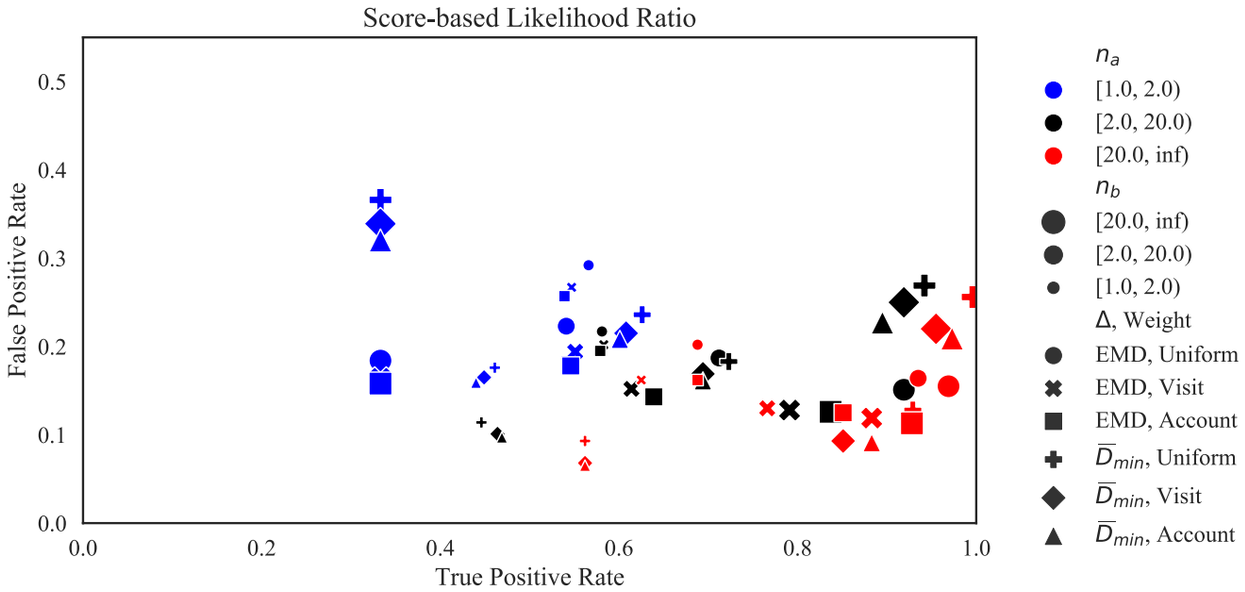
LR

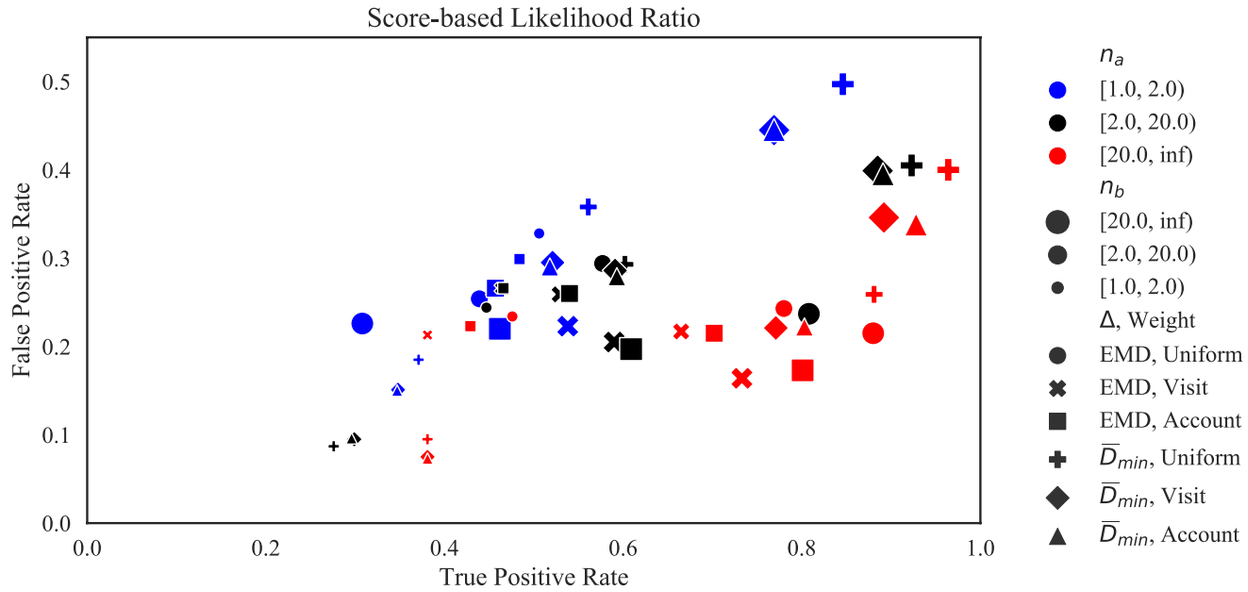
Region	Δ	Weights	TP@1	FP@1	AUC
OC	$\overline{D}_{min}$	Uniform	0.628	0.202	0.768
	$\overline{D}_{min}$	Account	0.610	0.171	0.774
	$\overline{D}_{min}$	Visit	0.611	0.180	0.768
	EMD	Uniform	**0.654**	0.197	**0.790**
	EMD	Account	0.614	**0.162**	0.783
	EMD	Visit	0.602	0.169	0.774
NY	$\overline{D}_{min}$	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


Likelihood Ratio
OC
False Positive Rate
True Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
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
$\alpha = 0.80$
$\alpha(n_a)$
$\alpha(n_a|\gamma, \rho, \phi)$


Likelihood Ratio
NY
False Positive Rate
True Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
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
$\alpha = 0.80$
$\alpha(n_a)$
$\alpha(n_a|\gamma, \rho, \phi)$


Score-based Likelihood Ratio
False Positive Rate
True Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
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)
Δ, Weight
EMD, Uniform
EMD, Visit
EMD, Account
$\overline{D}_{min}$, Uniform
$\overline{D}_{min}$, Visit
$\overline{D}_{min}$, Account


Score-based Likelihood Ratio
False Positive Rate
True Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
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)
Δ, Weight
EMD, Uniform
EMD, Visit
EMD, Account
$\overline{D}_{min}$, Uniform
$\overline{D}_{min}$, Visit
$\overline{D}_{min}$, Account