

# Statistics in Practice

**Quantifying dependencies in fingerprint evidence  
(and how you can get involved in the statistics community)**

FOCUS Cluster Dinner Series

October 28, 2019



[bit.ly/focus-oct2019](https://bit.ly/focus-oct2019)

Maria Tackett  
Duke University



A little about me...





B.S. in Mathematics

✗ Minor in Computer Science

✗ Minor in Economics

✗ Minor in German

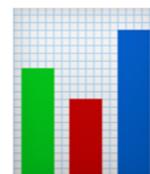


B.S. in Mathematics

✗ Minor in Computer Science

✗ Minor in Economics

✗ Minor in German



First statistics class junior year

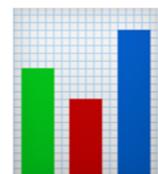


B.S. in Mathematics

✗ Minor in Computer Science

✗ Minor in Economics

✗ Minor in German



First statistics class junior year



M.S. in Statistics

# Statistician @ Capital One



Design of Experiments

Regression Modeling



Ph.D. in Statistics

**Assistant  
Professor of @  
the Practice**

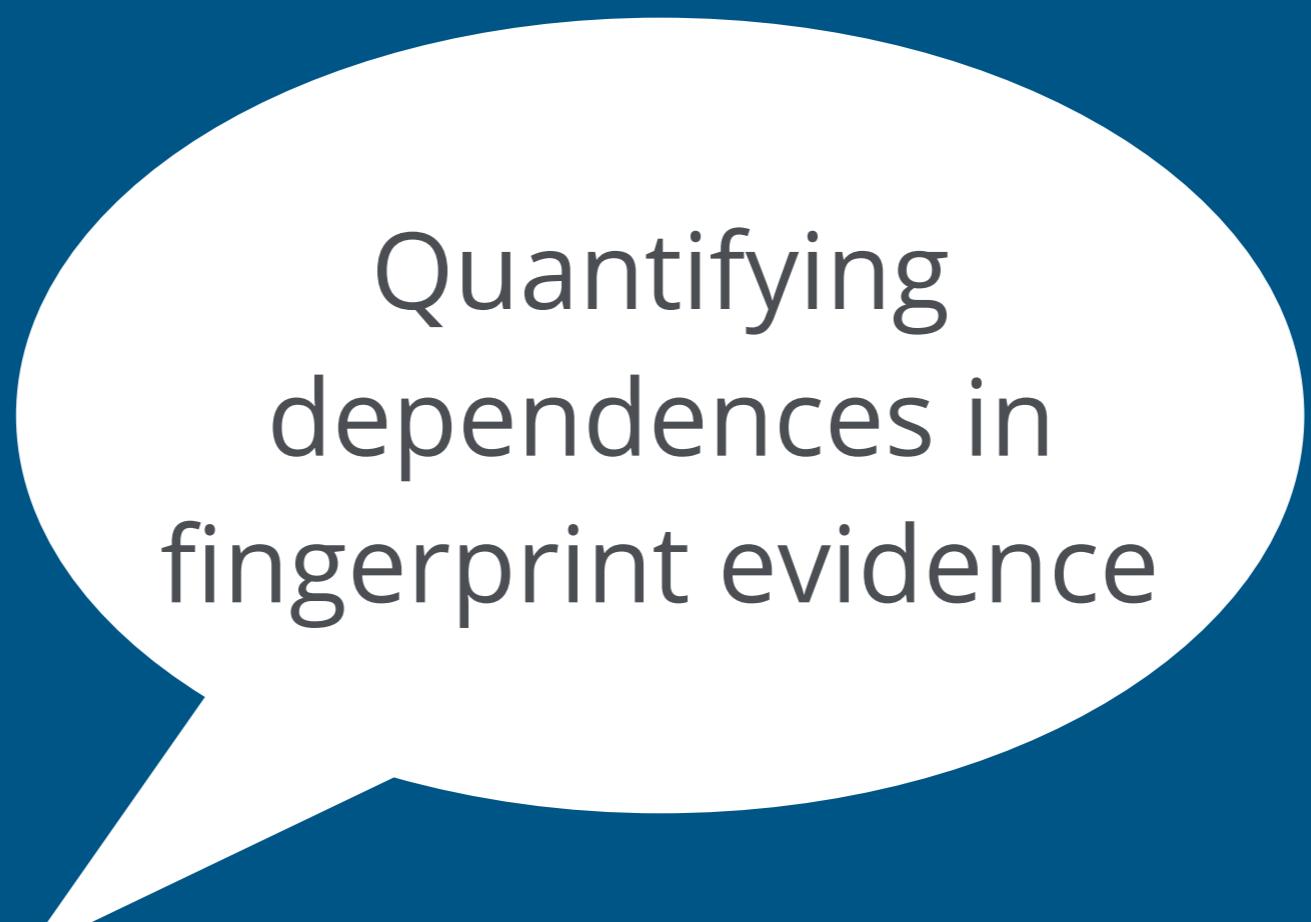


**Research**

Statistics Education  
  
Statistics in Criminal  
Justice and Forensic  
Science

**Teaching**

Intro to Data Science  
  
Regression Analysis



Quantifying  
dependences in  
fingerprint evidence

# Dissertation Research

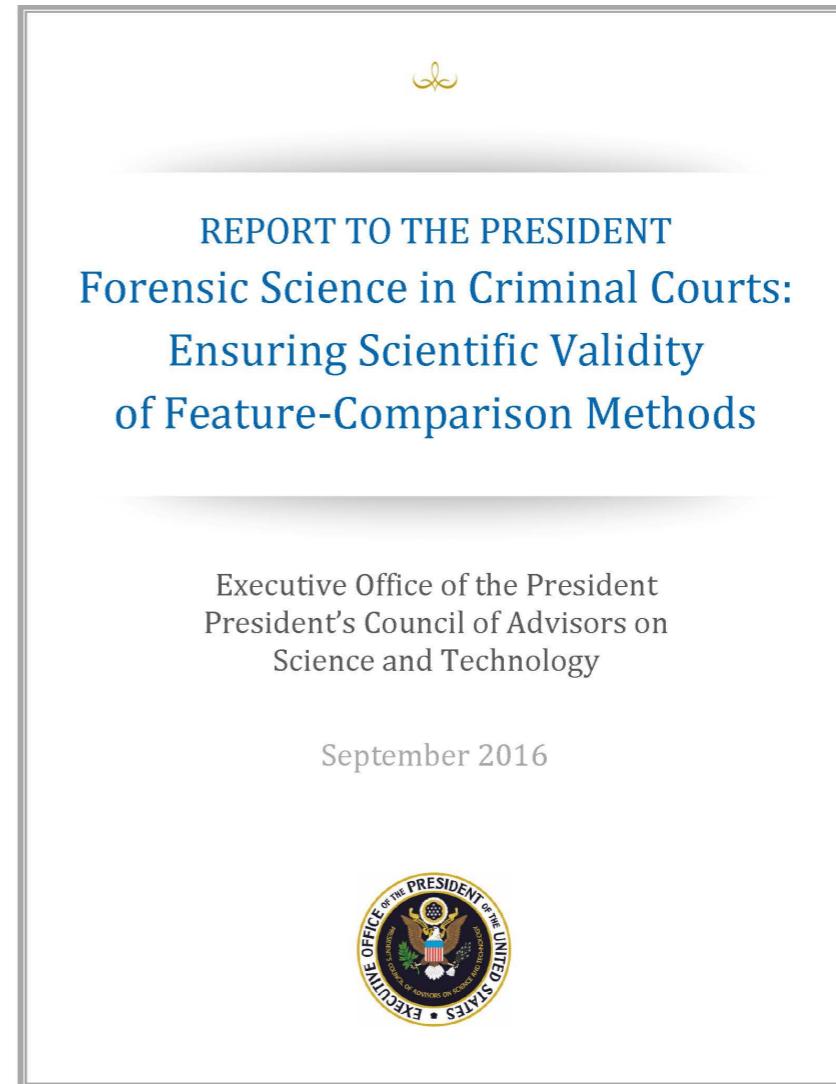
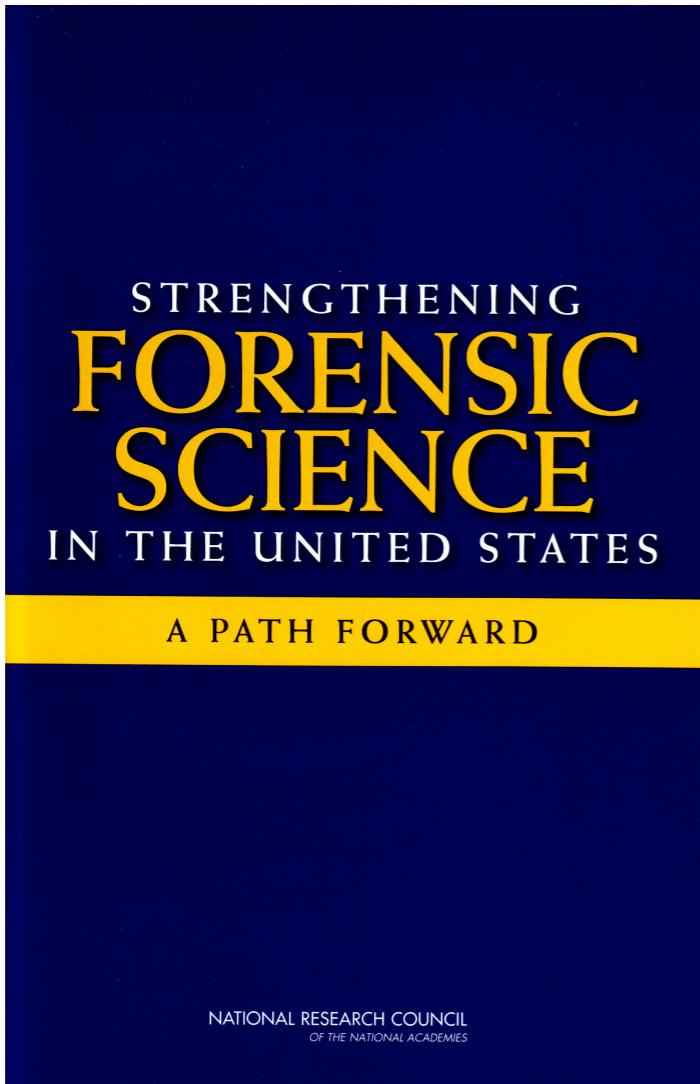
Used Bayesian methods to...

... decompose the sources of variability in two fingerprints

... develop a method to quantify certainty in fingerprint examiner's decision, adjusting for the number of fingerprints examined



# Statistics in Forensic Science



# General Framework: Lindley (1977)

Lindley (1977) introduced a Bayesian approach to interpreting forensic evidence. Suppose  $Y$  is a latent print recovered from a crime scene and  $X$  is a print from a person of interest. We want to test the hypotheses

$H_0$ :  $X$  and  $Y$  were produced by the same source

$H_1$ :  $X$  and  $Y$  were produced by different sources

$$\frac{P(H_0 | E)}{P(H_1 | E)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(E | H_0)}{P(E | H_1)}$$

# General Framework: Lindley (1977)

Lindley (1977) introduced a Bayesian approach to interpreting forensic evidence. Suppose  $Y$  is a latent print recovered from a crime scene and  $X$  is a print from a person of interest. We want to test the hypotheses

$H_0$ :  $X$  and  $Y$  were produced by the same source

$H_1$ :  $X$  and  $Y$  were produced by different sources

Posterior odds of  
a match

$$\frac{P(H_0 | E)}{P(H_1 | E)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(E | H_0)}{P(E | H_1)}$$

# General Framework: Lindley (1977)

Lindley (1977) introduced a Bayesian approach to interpreting forensic evidence. Suppose  $Y$  is a latent print recovered from a crime scene and  $X$  is a print from a person of interest. We want to test the hypotheses

$H_0$ :  $X$  and  $Y$  were produced by the same source

$H_1$ :  $X$  and  $Y$  were produced by different sources

Prior odds  
(non-fingerprint  
evidence)

$$\frac{P(H_0 | E)}{P(H_1 | E)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(E | H_0)}{P(E | H_1)}$$

# General Framework: Lindley (1977)

Lindley (1977) introduced a Bayesian approach to interpreting forensic evidence. Suppose  $Y$  is a latent print recovered from a crime scene and  $X$  is a print from a person of interest. We want to test the hypotheses

$H_0$ :  $X$  and  $Y$  were produced by the same source

$H_1$ :  $X$  and  $Y$  were produced by different sources

Weight of  
fingerprint  
evidence

$$\frac{P(H_0 | E)}{P(H_1 | E)} = \frac{P(H_0)}{P(H_1)} \times \frac{P(E | H_0)}{P(E | H_1)}$$

**Issue:**

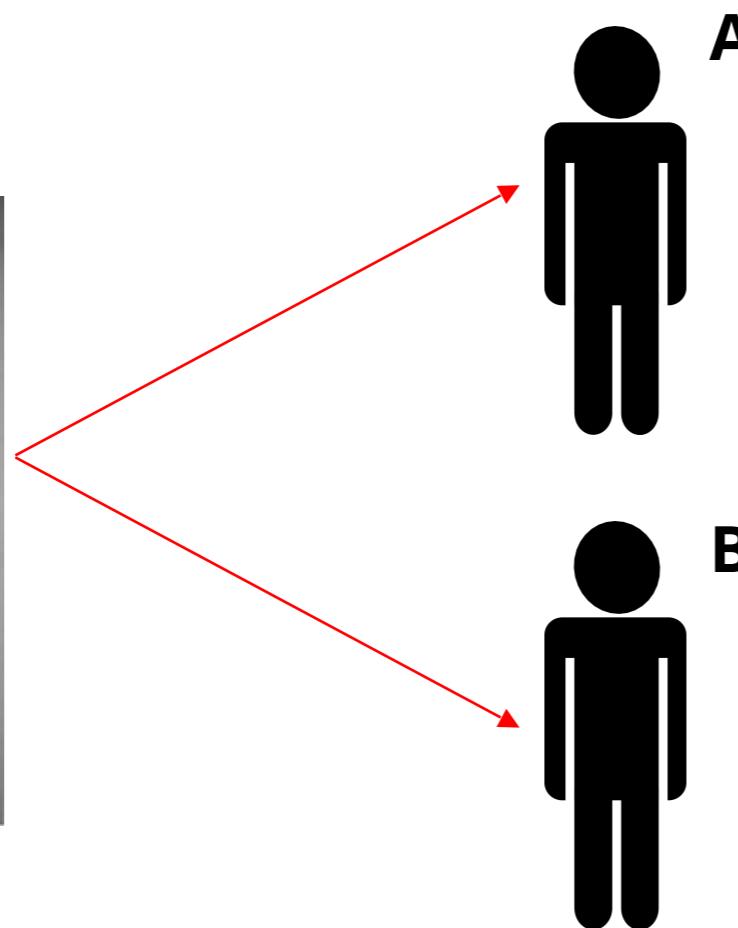
Examiners consider multiple candidates as a possible source of a latent fingerprint

**Proposed Solution:**

Quantify the weight of evidence in a way that accounts for the number of candidates examined

# Scenario: Two primary suspects

**Latent Print**

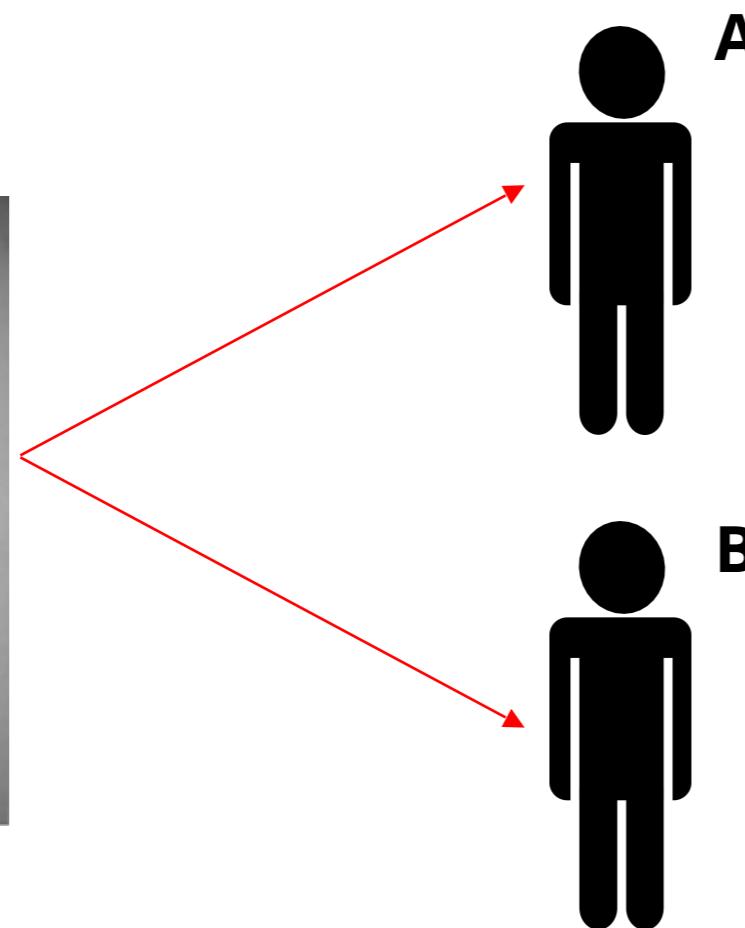


The latent print was created by

1. Candidate A
2. Candidate B
3. Neither candidate
4. Both candidates

# Scenario: Two Primary Suspects

**Latent Print**

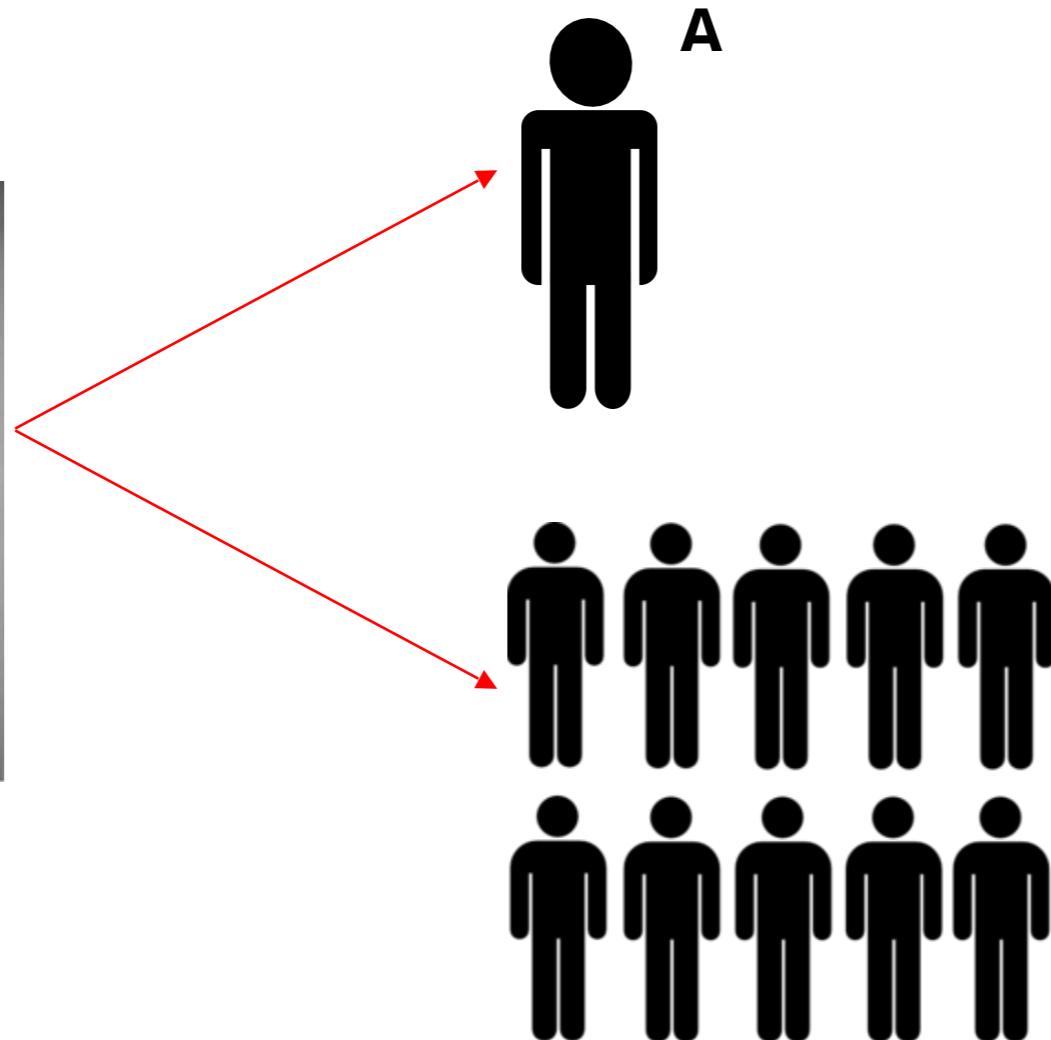


The latent print was created by

1. Candidate A
2. Candidate B
3. Neither candidate
4. Both candidates

# Scenario: Many suspects

**Latent Print**

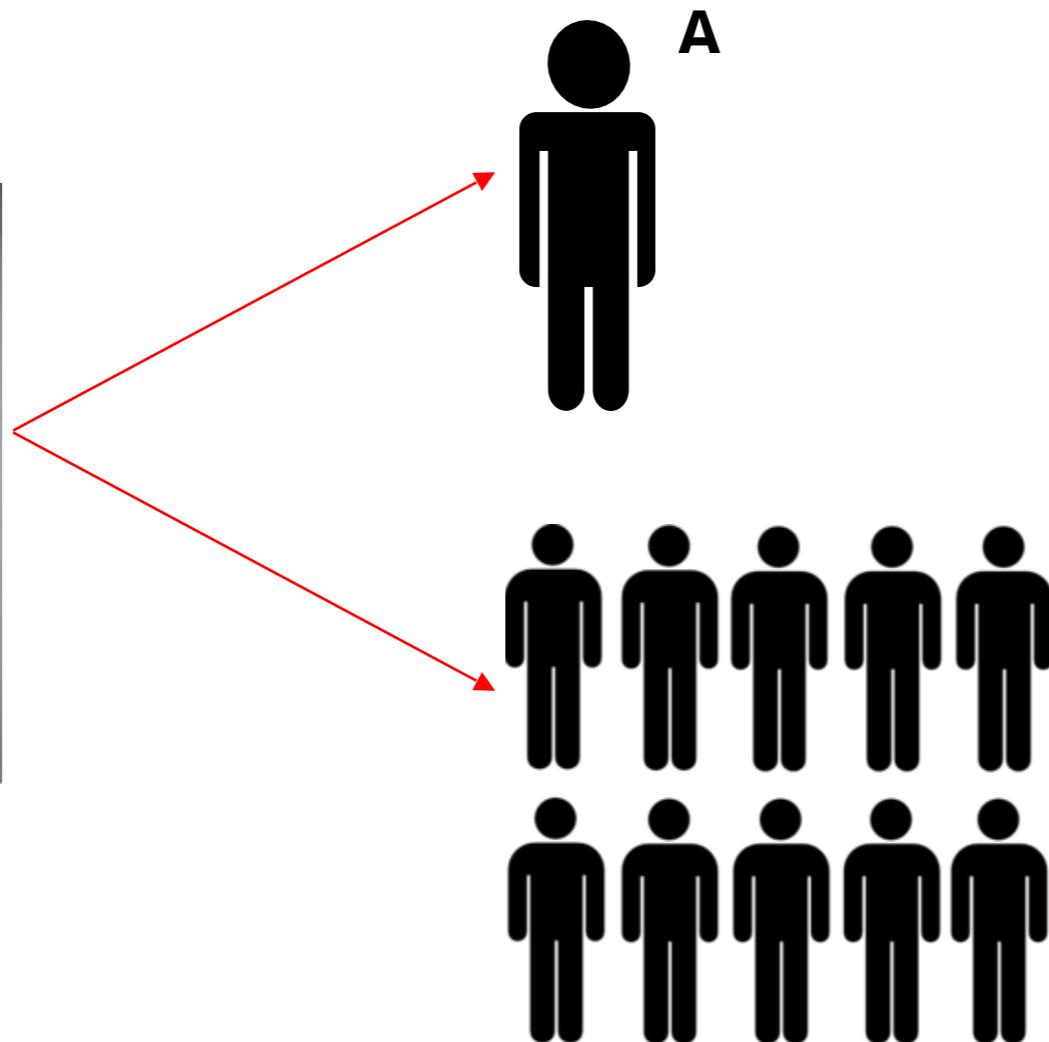


The latent print was created by

1. Candidate A
2. Candidate B
3. Candidate C
- ...
11. None of the candidates

# Scenario: Many suspects

**Latent Print**



The latent print was created by

1. Candidate A
2. Candidate B
3. Candidate C
- ...
11. None of the candidates

**Constraint:** At most one candidate could have created the latent print

# Set of possible models

**Constraint:** At most one candidate could have created the latent print

For  $i = 1, \dots, K$ ,

$$M_i : \begin{cases} \theta_{X_i} = \theta_Y \\ \theta_{X_j} \neq \theta_Y & j = 1, \dots, K \text{ such that } i \neq j \end{cases}$$

$$M_0: \theta_{X_1} \neq \theta_Y \quad \theta_{X_2} \neq \theta_Y \quad \dots \quad \theta_{X_K} \neq \theta_Y$$

$\theta_{X_i}$ : Measurements from candidate  $i$

$\theta_Y$ : Measurements from the latent print

# Weight of fingerprint evidence

By imposing this constraint on the model framework, we can measure the weight of evidence supporting the hypothesis that the latent print was created by Candidate  $i$  as

Weight of evidence  
that Candidate  $i$   
created the latent  
print

=

How closely the print  
from Candidate  $i$   
matches the latent  
print

X

Effect of evidence  
from other  
candidates

$BF_i$

$BF_{i0}$

$DF_i$

# Calculating $\text{BF}_i$

Let  $\mathbf{X} = X_1, \dots, X_K$ .

Suppose there are two models:  $M_i : \theta_{X_i} = \theta_Y$  and  $M_0 : \theta_{X_i} \neq \theta_Y$

Two Models: Weight of Evidence for  $M_i$

$$BF_{i0} = \frac{P(M_i|\mathbf{X})/P(M_0|\mathbf{X})}{P(M_i)/P(M_0)}$$

If we extend this to  $K + 1$  models,  $M_0, \dots, M_K$

$K + 1$  Models: Weight of Evidence for  $M_i$

$$BF_i = \frac{P(M_i|\mathbf{X})/(1 - P(M_i|\mathbf{X}))}{P(M_i)/(1 - P(M_i))}$$

# Calculating $\text{BF}_i$

$$\text{BF}_i = \frac{P(M_i|\mathbf{X})/(1 - P(M_i|\mathbf{X}))}{P(M_i)/(1 - P(M_i))}$$

= ...

$$= \text{BF}_{i0} \left[ 1 + \sum_{\substack{k=1 \\ k \neq i}}^K \frac{P(M_k)}{P(M_0)} \right] \Big/ \left[ 1 + \sum_{\substack{k=1 \\ k \neq i}}^K \frac{P(M_k)}{P(M_0)} \text{BF}_{k0} \right]$$

$$= \text{BF}_{i0} \times \text{DF}_i$$

$$= \text{Bayes Factor} \times \text{Dependency Factor}$$

# Interpreting $\text{BF}_i$

$$\text{BF}_i = \text{BF}_{i0} \times \text{DF}_i$$

## Bayes Factor, $\text{BF}_{i0}$

- Indicates how closely the fingerprint from Candidate  $i$  matches the latent print
- Interpreted independent from the other candidates examined

## Dependency Factor, $\text{DF}_i$

- Indicates the effect of evidence from other candidates
- Interpretation:
  - $\text{DF}_i > 1$  : Additional candidates show weak evidence
  - $\text{DF}_i < 1$  : Additional candidates show strong evidence

# Implications for fingerprint examiners

$$2 \log(BF_i) = 2 \log(BF_{i0}) + 2 \log(DF_i)$$

- ✓ Provides a probative value to express level of uncertainty  
(Dror and Mnookin, 2010)
- ✓ Interpreted using a common scale for assessing the weight of evidence (Kass and Raftery, 1995)
- ✓ Eliminates the need for examiners to adjust their decision criteria based on the number of candidates examined

# Example: Three Candidates Examined

Current Approach			Proposed Approach		
Candidate A	Candidate B	Candidate C	Candidate A	Candidate B	Candidate C
6.000	-5.889	-5.889	5.800	-5.889	-5.889
6.000	-2.197	-5.889	5.347	-2.197	-5.889
6.000	-2.197	-2.197	4.978	-2.197	-2.197
6.000	0.000	-5.889	4.562	0.000	-5.889
6.000	0.000	-2.197	4.305	0.000	-2.197
6.000	0.000	0.000	3.803	0.000	0.000
6.000	0.000	2.773	2.416	0.000	2.773
6.000	2.773	-5.889	2.760	2.773	-5.889
6.000	2.773	-2.197	2.652	2.773	-2.197
6.000	2.773	2.773	1.605	2.773	2.773
6.000	5.889	-5.889	0.003	5.889	-5.889
6.000	5.889	-2.197	-0.025	5.889	-2.197
6.000	5.889	0.000	-0.089	5.889	0.000
6.000	5.889	2.773	-0.356	5.889	2.773
6.000	5.889	5.889	-1.327	5.889	5.889

Evidence supporting Candidate  $i$  as source of latent print

- 6: Strong
- 2 to 6: Positive
- 0 to 2: Neutral

# Example: Three Candidates Examined

Current Approach			Proposed Approach		
Candidate A	Candidate B	Candidate C	Candidate A	Candidate B	Candidate C
6.000	-5.889	-5.889	5.800	-5.889	-5.889
6.000	-2.197	-5.889	5.347	-2.197	-5.889
6.000	-2.197	-2.197	4.978	-2.197	-2.197
6.000	0.000	-5.889	4.562	0.000	-5.889
6.000	0.000	-2.197	4.305	0.000	-2.197
6.000	0.000	0.000	3.803	0.000	0.000
6.000	0.000	2.773	2.416	0.000	2.773
6.000	2.773	-5.889	2.760	2.773	-5.889
6.000	2.773	-2.197	2.652	2.773	-2.197
6.000	2.773	2.773	1.605	2.773	2.773
6.000	5.889	-5.889	0.003	5.889	-5.889
6.000	5.889	-2.197	-0.025	5.889	-2.197
6.000	5.889	0.000	-0.089	5.889	0.000
6.000	5.889	2.773	-0.356	5.889	2.773
6.000	5.889	5.889	-1.327	5.889	5.889

Evidence supporting Candidate  $i$  as source of latent print

- 6: Strong
- 2 to 6: Positive
- 0 to 2: Neutral

# Example: Three Candidates Examined

Current Approach			Proposed Approach		
Candidate A	Candidate B	Candidate C	Candidate A	Candidate B	Candidate C
6.000	-5.889	-5.889	5.800	-5.889	-5.889
6.000	-2.197	-5.889	5.347	-2.197	-5.889
6.000	-2.197	-2.197	4.978	-2.197	-2.197
6.000	0.000	-5.889	4.562	0.000	-5.889
6.000	0.000	-2.197	4.305	0.000	-2.197
6.000	0.000	0.000	3.803	0.000	0.000
6.000	0.000	2.773	2.416	0.000	2.773
6.000	2.773	-5.889	2.760	2.773	-5.889
6.000	2.773	-2.197	2.652	2.773	-2.197
6.000	2.773	2.773	1.605	2.773	2.773
6.000	5.889	-5.889	0.003	5.889	-5.889
6.000	5.889	-2.197	-0.025	5.889	-2.197
6.000	5.889	0.000	-0.089	5.889	0.000
6.000	5.889	2.773	-0.356	5.889	2.773
6.000	5.889	5.889	-1.327	5.889	5.889

Evidence supporting Candidate  $i$  as source of latent print

- 6: Strong
- 2 to 6: Positive
- 0 to 2: Neutral

# Example: Three Candidates Examined

Current Approach			Proposed Approach		
Candidate A	Candidate B	Candidate C	Candidate A	Candidate B	Candidate C
6.000	-5.889	-5.889	5.800	-5.889	-5.889
6.000	-2.197	-5.889	5.347	-2.197	-5.889
6.000	-2.197	-2.197	4.978	-2.197	-2.197
6.000	0.000	-5.889	4.562	0.000	-5.889
6.000	0.000	-2.197	4.305	0.000	-2.197
6.000	0.000	0.000	3.803	0.000	0.000
6.000	0.000	2.773	2.416	0.000	2.773
6.000	2.773	-5.889	2.760	2.773	-5.889
6.000	2.773	-2.197	2.652	2.773	-2.197
6.000	2.773	2.773	1.605	2.773	2.773
6.000	5.889	-5.889	0.003	5.889	-5.889
6.000	5.889	-2.197	-0.025	5.889	-2.197
6.000	5.889	0.000	-0.089	5.889	0.000
6.000	5.889	2.773	-0.356	5.889	2.773
6.000	5.889	5.889	-1.327	5.889	5.889

Evidence supporting Candidate  $i$  as source of latent print

- 6: Strong
- 2 to 6: Positive
- 0 to 2: Neutral



How can you get involved?

# Statistical Science

- ✓ STA 210: Regression Analysis
- ✓ STA 240L: Probability for Statistics
- ✓ STA 250: Statistics
- ✓ STA 360: Bayesian Inference and Modern Statistical Methods
- ✓ STA 440: Case Studies in the Practice of Statistics
- + Electives!

On campus



**DEADLINE  
EXTENDED TO  
MONDAY AT  
11:59 P.M.!**

**Sign-up now at:  
[dukeml.org/register](http://dukeml.org/register)**

# On campus



# On campus

Join us for the ...

## Electronic Undergraduate Research Conference

**Friday, November 1      Sociology Psychology 238**

11:30 am - 12:20 pm	Virtual Video Presentation Session
12:20 pm - 12:30 pm	Opening Remarks
12:30 pm - 2:00 pm	Plenary Talks by USPROC Award Winners
2:00 pm - 3:00 pm	Keynote Address <i>Finding Your Balance: Foundations vs. Evolution in Data Science</i>
3:00 pm - 3:45 pm	Graduate School Information Session
3:45 pm - 4:30 pm	Panel Discussion About Careers in Industry and Government
4:30 pm - 4:40 pm	Closing Remarks

**Snacks provided!**

**Join when  
you can!**

More details available at  
<https://www.causeweb.org/usproc/eusrc/2019/program>

# Outside of Duke

## **Undergraduate Statistics Class Project Competition**



December 20



<https://www.causeweb.org/usproc/usclap>

## **Undergraduate Statistics Research Project Competition**

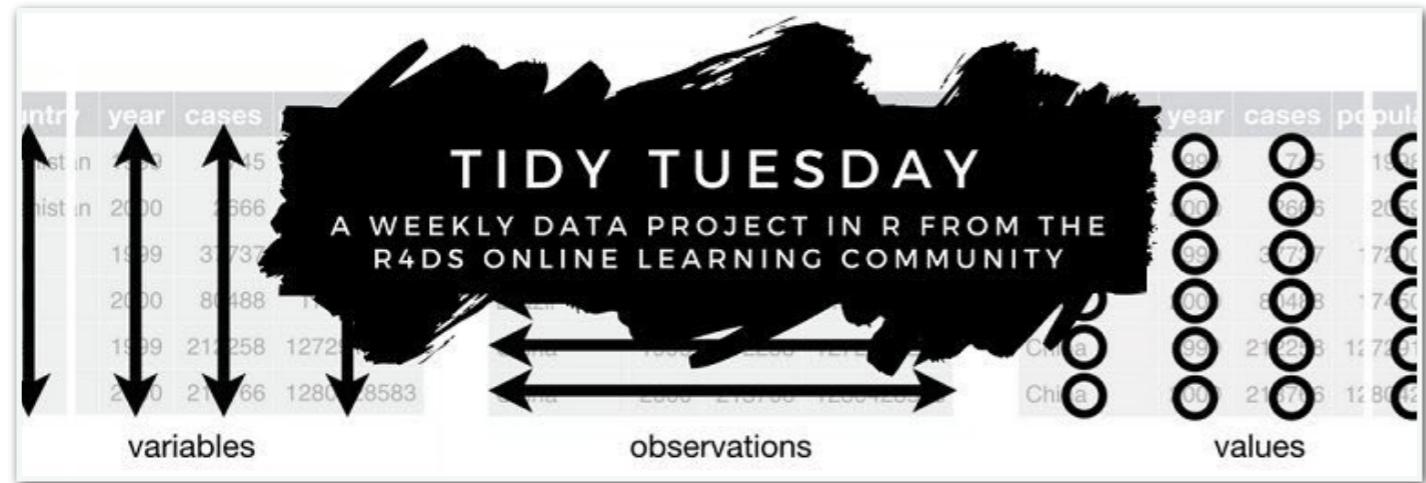


December 20

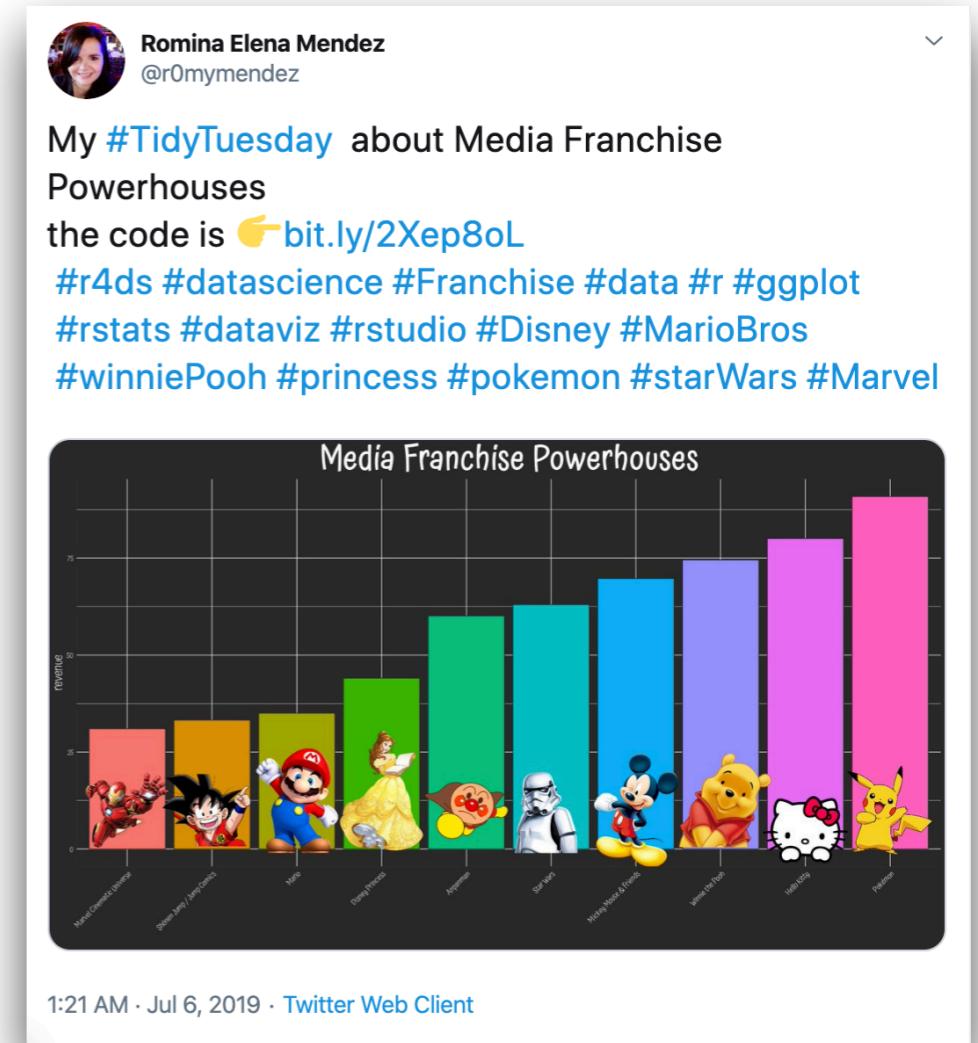
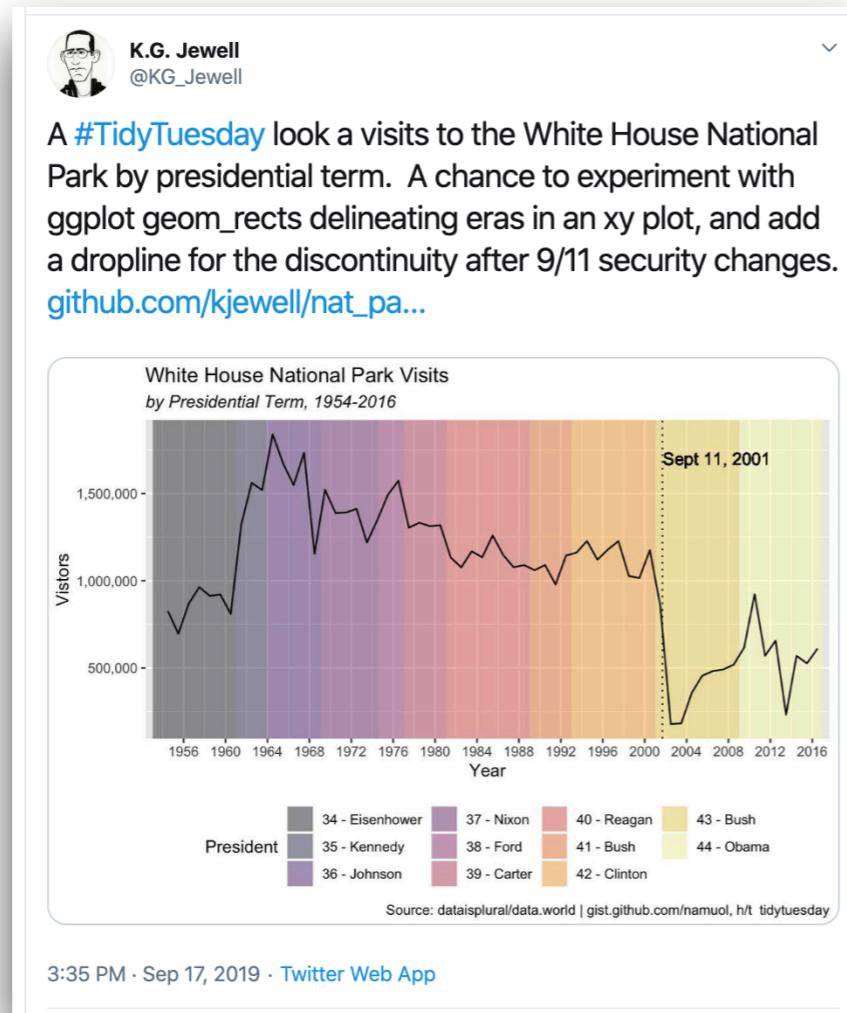


<https://www.causeweb.org/usproc/usresp>

# Online



<https://github.com/rfordatascience/tidytuesday>

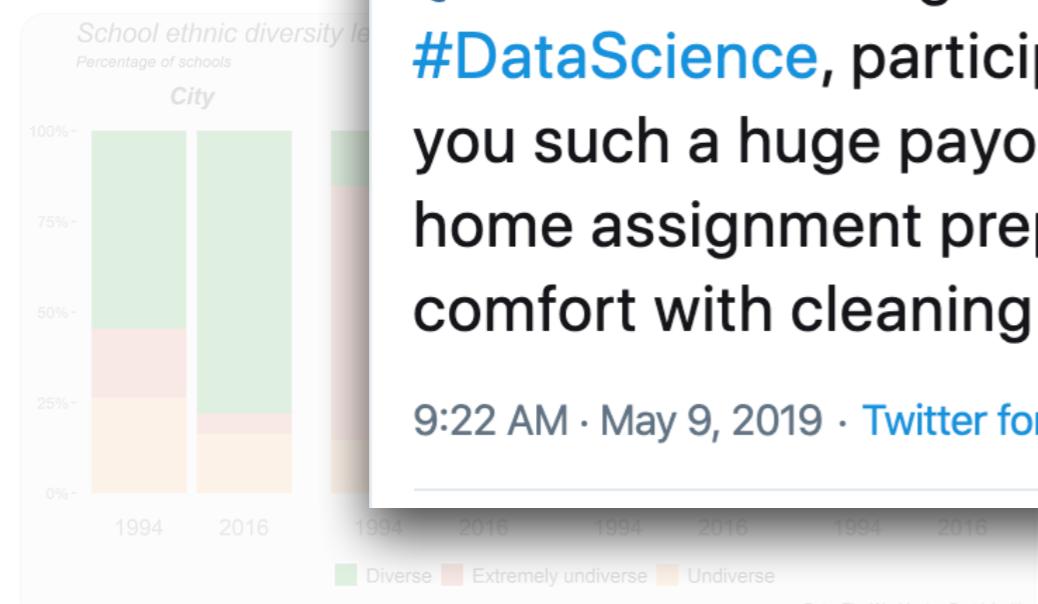




Amit Levinson  
@Amit\_Levinson

#TidyTuesday plot of school ethnic diversity levels across residency types for 1994 & 2016

I love seeing how it thoughts on minority diverse in cities > schools code: [bit.ly/2nuiv5](https://bit.ly/2nuiv5)



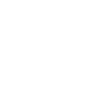
Amit Levinson @Amit\_Levinson · Sep 26

Replying to @Amit\_Levinson

I learned a lot this week!

Analysis: using stringr package instead of base r to change strings.

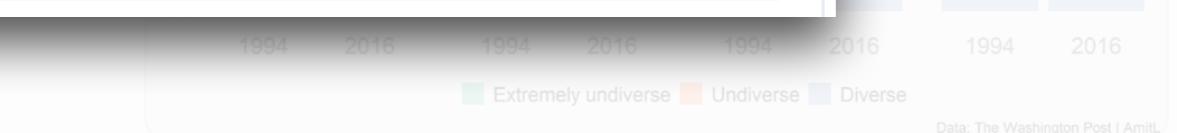
Visualization: used a stacked bar with 3 variables and various levels. I initially plot, but i think stacked



Rika Gorn 🎃  
@RikaGorn

🗣 Friends looking to get into analytics and #DataScience, participating in #TidyTuesday will give you such a huge payoff including: a portfolio, take home assignment prep, #rstats & #dataviz learning, comfort with cleaning messy datasets, a community!

9:22 AM · May 9, 2019 · Twitter for Android



3:44 PM · Sep 26, 2019 · Twitter Web App

1

# Thank You!

[bit.ly/focus-oct2019](https://bit.ly/focus-oct2019)

Maria Tackett  
Duke University

# References

- Dror, I. and Mnookin, J. 2010. The Use of Technology in Human Expert Domains: Challenges and Risks Arising from the Use of Automated Fingerprint Identification Systems in Forensic Science, *Law, Probability & Risk*, 9: 47 - 67.
- Kass, R. and Raftery, A. 1995. Bayes Factors, *Journal of the American Statistical Association*, 90: 773 - 795.