Bayesian Machine Learning for Social Data Science

P. M. KRAFFT

UNIVERSITY of WASHINGTON

This Talk

BAYESIAN MACHINE LEARNING FOR SOCIAL DATA SCIENCE

1) Overview of my Work

2) Intro to Polarization Model

3) Brief Bayesian Inference Tutorial

4) Polarization Model

5) Ongoing and Future Work

SOCIAL SCIENCE

Computational Social Science

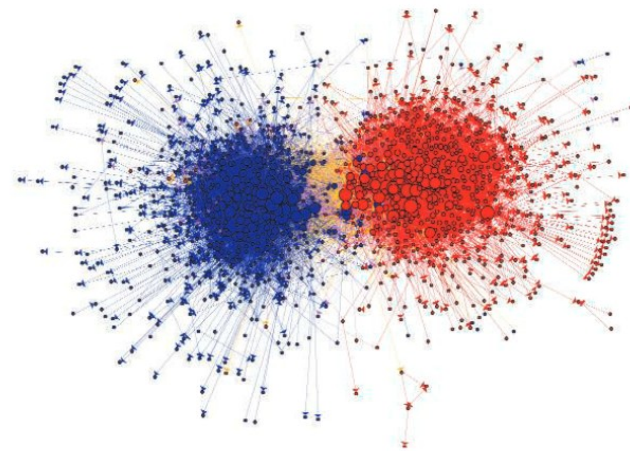
A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the



Data from the blogosphere. Shown is a link structure within a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it. [Reproduced from (8) with permission from the Association for Computing Machinery]

[1]Harvard University, Cambridge, MA, USA. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]University of Michigan, Ann Arbor, MI, USA. [4]New York University, New York, NY, USA. [5]Northeastern University, Boston, MA, USA. [6]Interdisciplinary Scientific Research, Seattle, WA, USA. [7]Northwestern University, Evanston, IL, USA. [8]University of California–San Diego, La Jolla, CA, USA. [9]Columbia University, New York, NY, USA [10]Cornell University, Ithaca, NY, USA. [11]Boston University, Boston, MA, USA. E-mail: david_lazer@harvard.edu. Complete affiliations are listed in the supporting online material.

www.sciencemag.org SCIENCE VOL 323 6 FEBRUARY 2009

IC²S²

SOCIAL SCIENCE

Computational Social Science

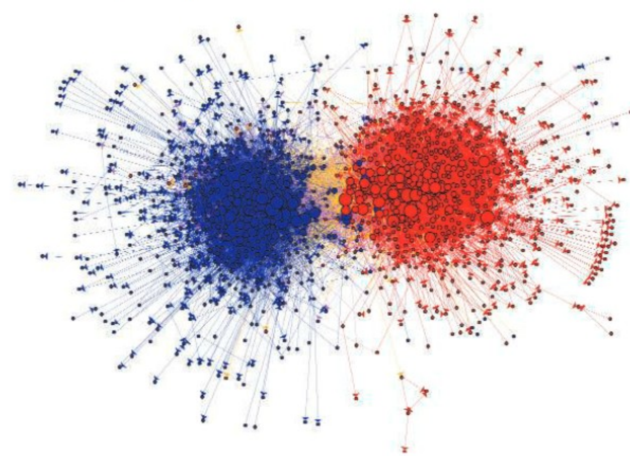
David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the



Data from the blogosphere. Shown is a link structure within a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it. [Reproduced from (8) with permission from the Association for Computing Machinery]

[1]Harvard University, Cambridge, MA, USA. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]University of Michigan, Ann Arbor, MI, USA. [4]New York University, New York, NY, USA. [5]Northeastern University, Boston, MA, USA. [6]Interdisciplinary Scientific Research, Seattle, WA, USA. [7]Northwestern University, Evanston, IL, USA. [8]University of California–San Diego, La Jolla, CA, USA. [9]Columbia University, New York, NY, USA [10]Cornell University, Ithaca, NY, USA. [11]Boston University, Boston, MA, USA. E-mail: david_lazer@harvard.edu. Complete affiliations are listed in the supporting online material.

www.sciencemag.org SCIENCE VOL 323 6 FEBRUARY 2009

Methodology

SOCIAL SCIENCE

Computational Social Science

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

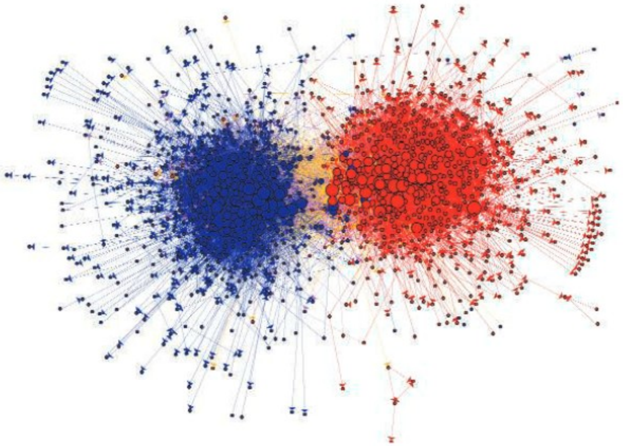
David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the

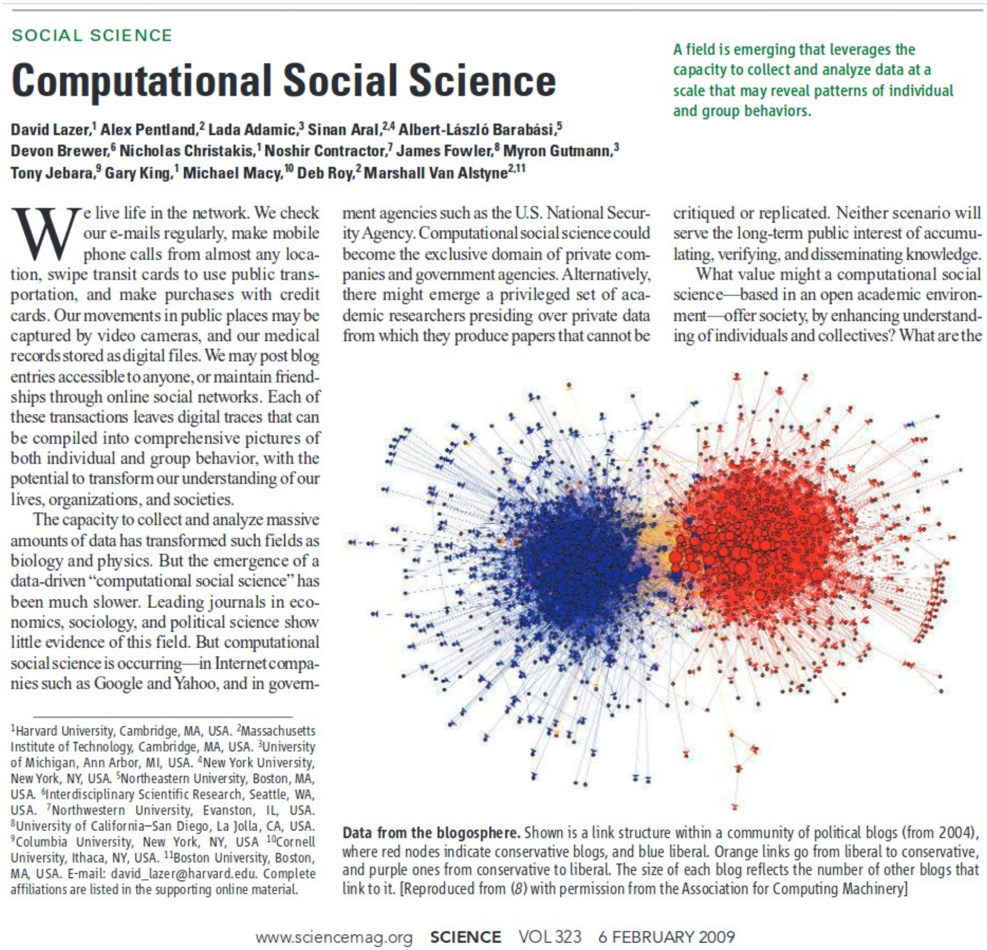
[1]Harvard University, Cambridge, MA, USA. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]University of Michigan, Ann Arbor, MI, USA. [4]New York University, New York, NY, USA. [5]Northeastern University, Boston, MA, USA. [6]Interdisciplinary Scientific Research, Seattle, WA, USA. [7]Northwestern University, Evanston, IL, USA. [8]University of California–San Diego, La Jolla, CA, USA. [9]Columbia University, New York, NY, USA [10]Cornell University, Ithaca, NY, USA. [11]Boston University, Boston, MA, USA. E-mail: david_lazer@harvard.edu. Complete affiliations are listed in the supporting online material.



Data from the blogosphere. Shown is a link structure within a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it. [Reproduced from (8) with permission from the Association for Computing Machinery]

www.sciencemag.org SCIENCE VOL 323 6 FEBRUARY 2009

Methodology



SOCIAL SCIENCE

Computational Social Science

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

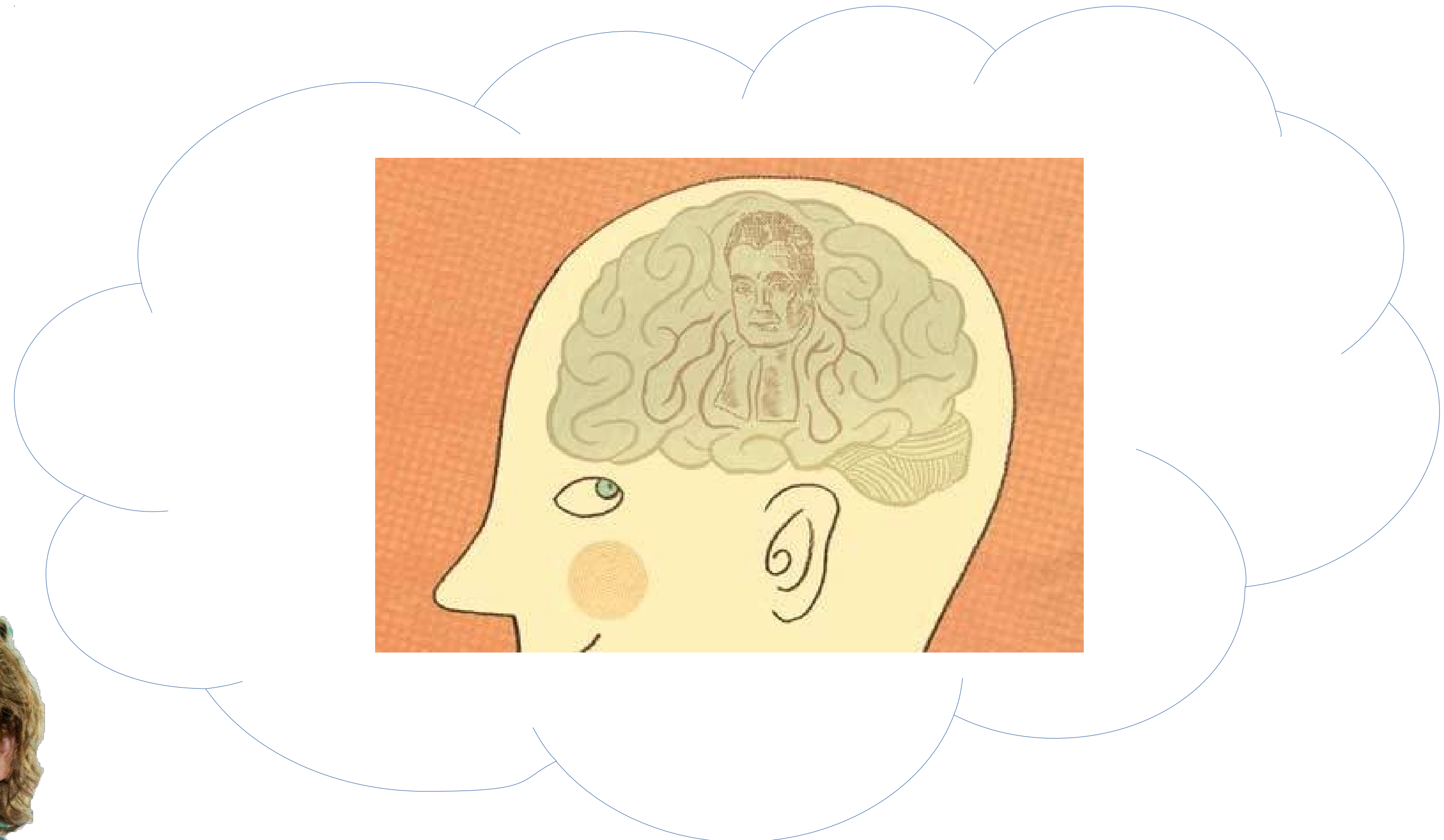
What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the

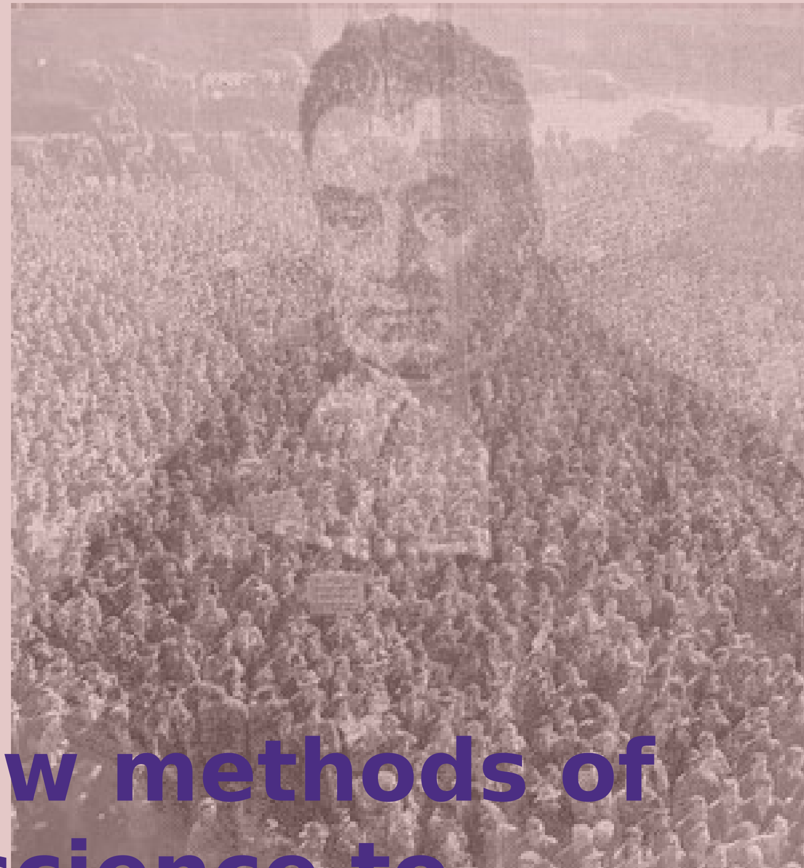
[1]Harvard University, Cambridge, MA, USA. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]University of Michigan, Ann Arbor, MI, USA. [4]New York University, New York, NY, USA. [5]Northeastern University, Boston, MA, USA. [6]Interdisciplinary Scientific Research, Seattle, WA, USA. [7]Northwestern University, Evanston, IL, USA. [8]University of California–San Diego, La Jolla, CA, USA. [9]Columbia University, New York, NY, USA [10]Cornell University, Ithaca, NY, USA. [11]Boston University, Boston, MA, USA. E-mail: david_lazer@harvard.edu. Complete affiliations are listed in the supporting online material.

Data from the blogosphere. Shown is a link structure within a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it. [Reproduced from (*8*) with permission from the Association for Computing Machinery]

www.sciencemag.org SCIENCE VOL 323 6 FEBRUARY 2009

Data





I develop new methods of social data science to understand rumors, fads, conspiracies, disinformation, public opinion, and information flow

This Talk

BAYESIAN MACHINE LEARNING FOR SOCIAL DATA SCIENCE

Methodology

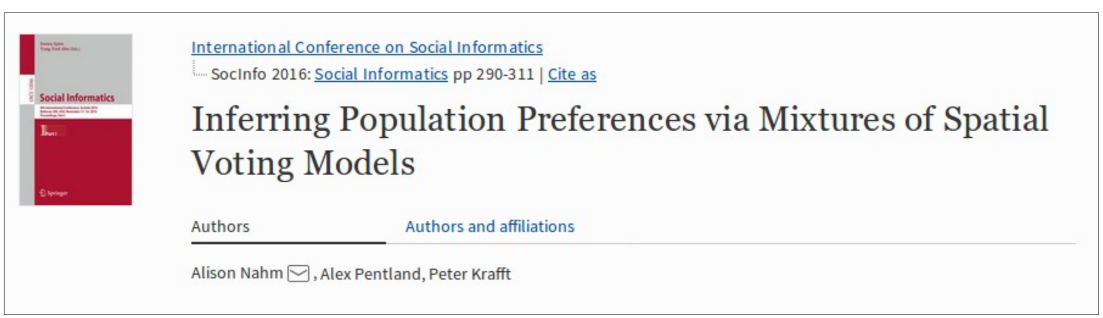
Epistemology

This Talk

BAYESIAN MACHINE LEARNING FOR SOCIAL DATA SCIENCE

Methodology

Epistemology



International Conference on Social Informatics

SocInfo 2016: Social Informatics pp 290-311 | Cite as

Inferring Population Preferences via Mixtures of Spatial Voting Models

Authors | Authors and affiliations

Alison Nahm, Alex Pentland, Peter Krafft

This Talk

BAYESIAN MACHINE LEARNING FOR SOCIAL DATA SCIENCE

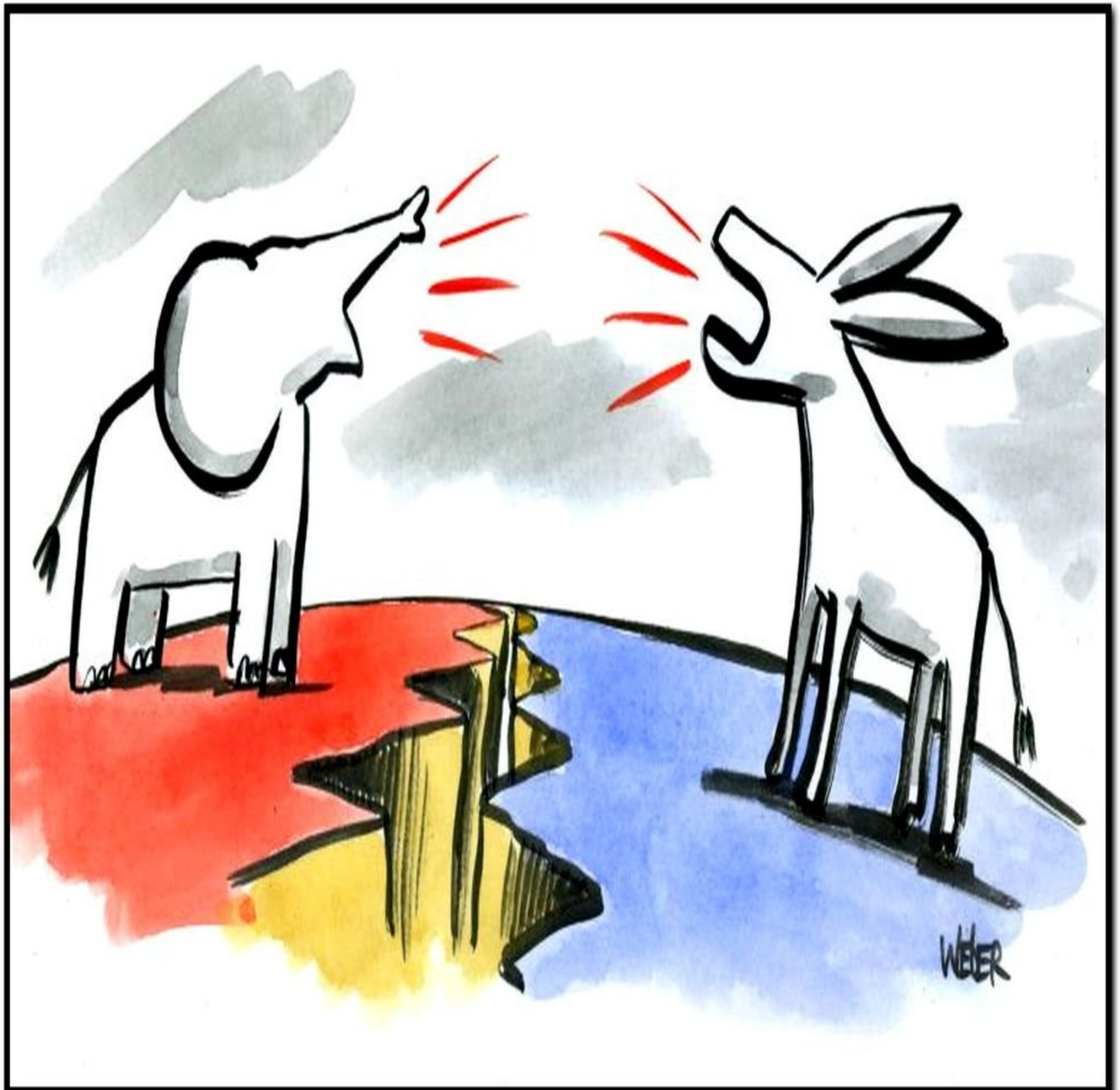
1) Overview of my Work

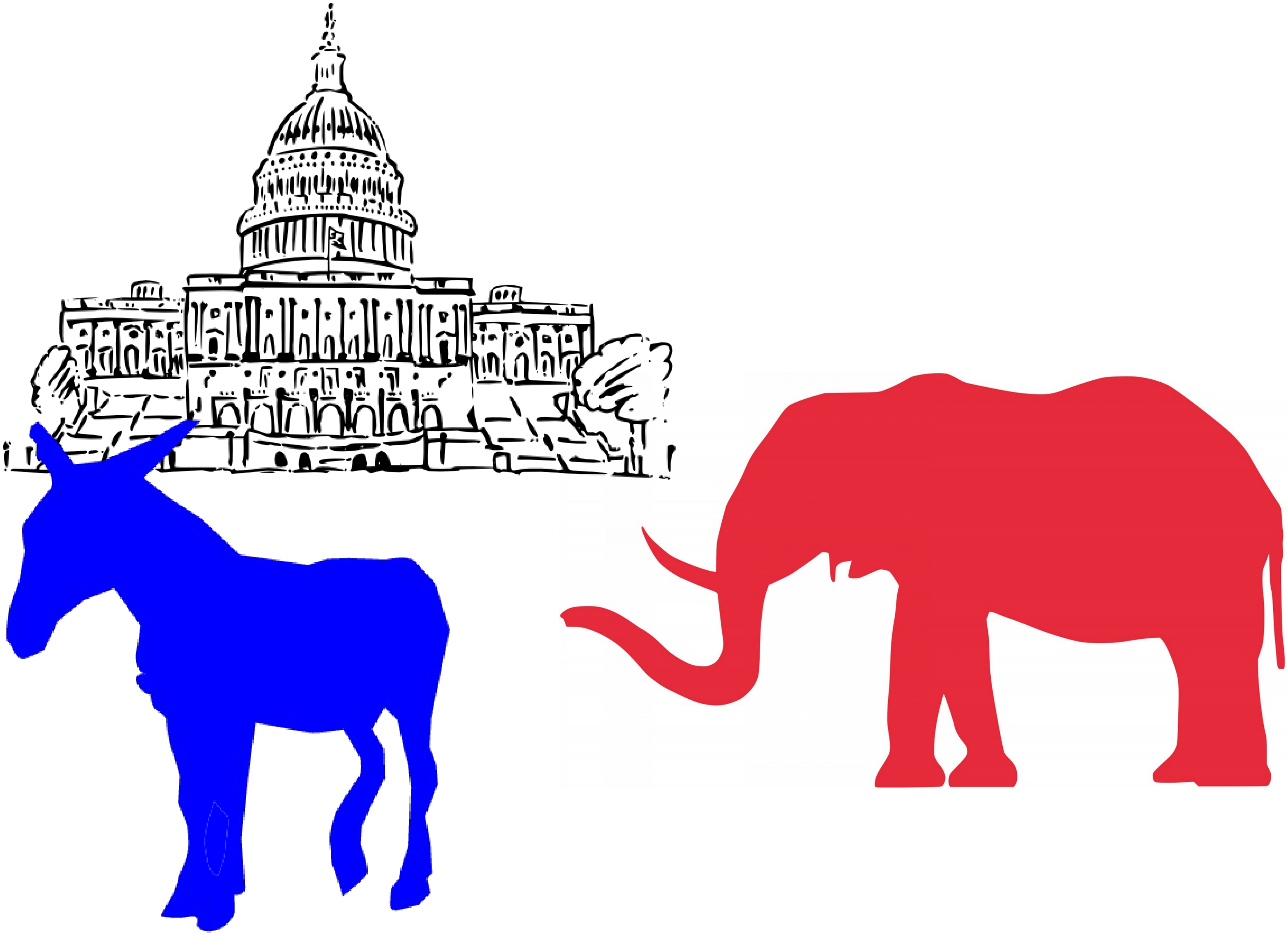
2) **Intro to Polarization Model**

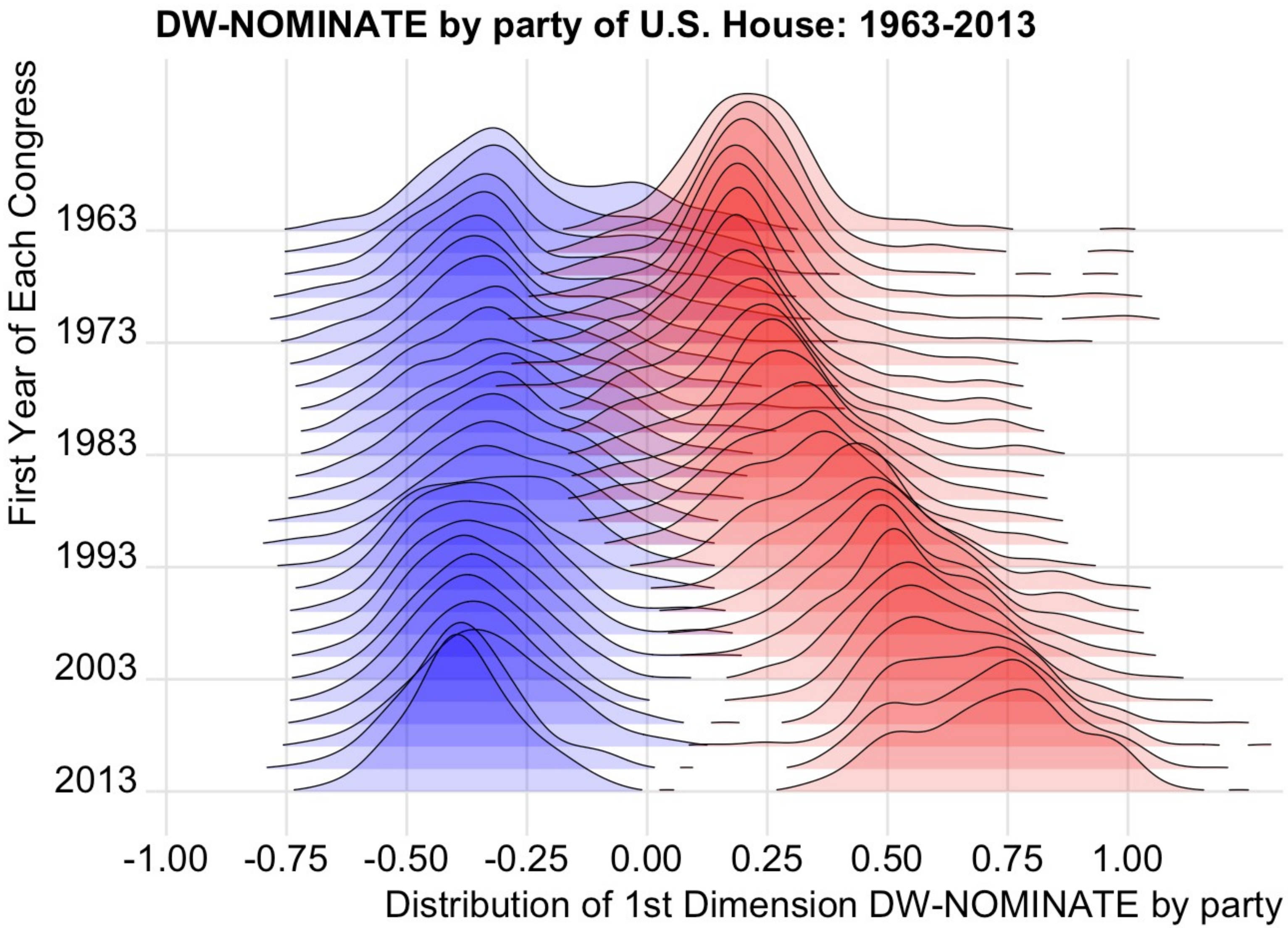
3) Brief Bayesian Inference Tutorial

4) Polarization Model

5) Ongoing and Future Work

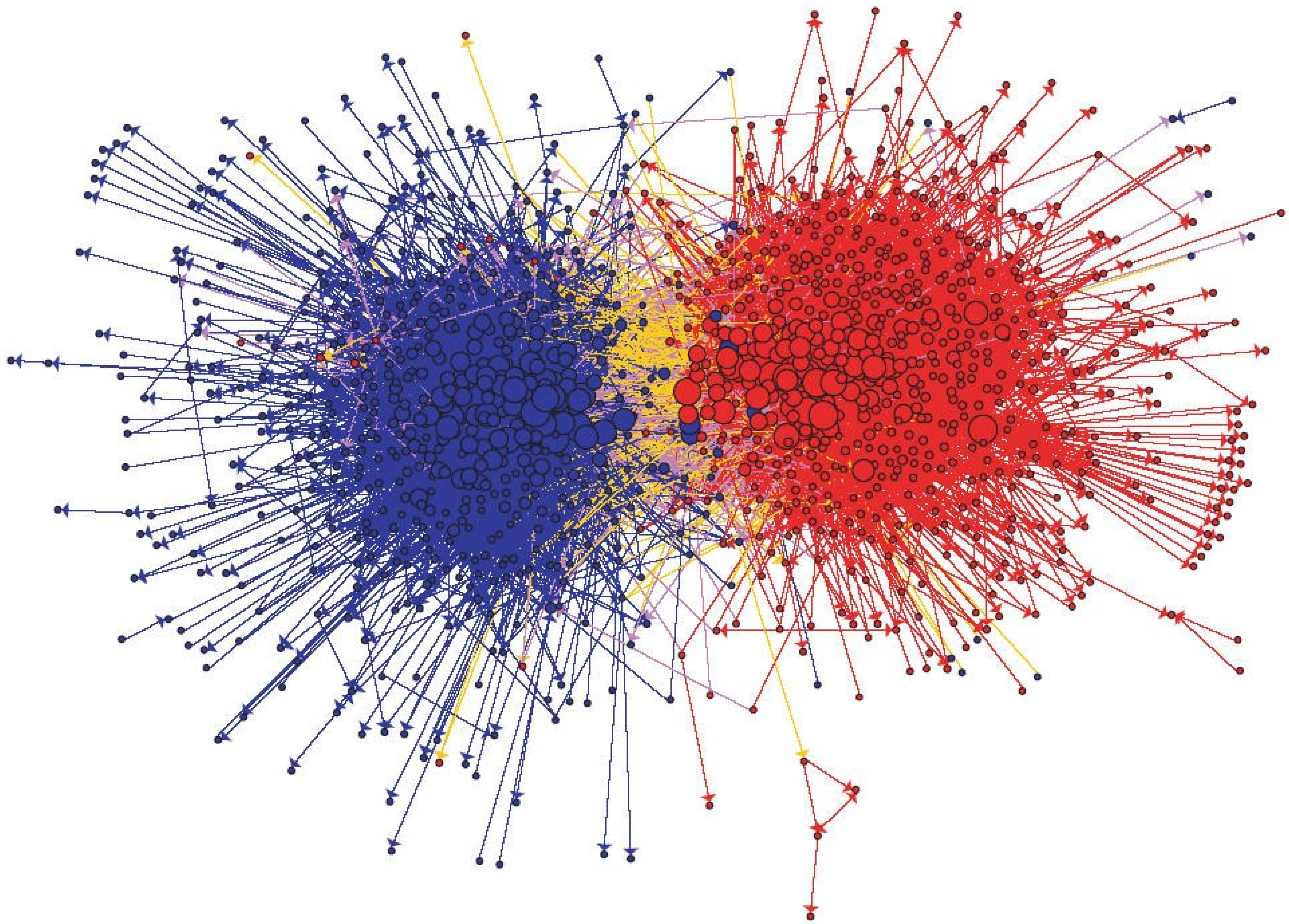

WEBER





(McDonald, 2017)

What about the American public?



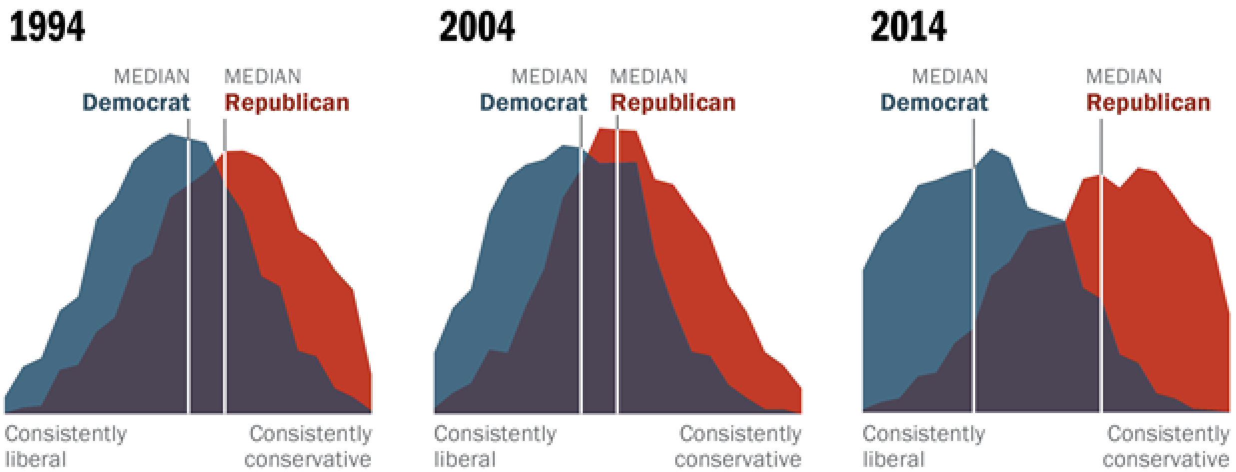
(Adamic and Glance, 2005)

Mass Polarization

RECENT POLLS

Democrats and Republicans More Ideologically Divided than in the Past

Distribution of Democrats and Republicans on a 10-item scale of political values



Source: 2014 Political Polarization in the American Public
Notes: Ideological consistency based on a scale of 10 political values questions (see Appendix A). The blue area in this chart represents the ideological distribution of Democrats; the red area of Republicans. The overlap of these two distributions is shaded purple. Republicans include Republican-leaning independents; Democrats include Democratic-leaning independents (see Appendix B).

PEW RESEARCH CENTER

Mass Polarization

POTENTIAL MECHANISMS

Exposure to opposing views on social media can increase political polarization

Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky

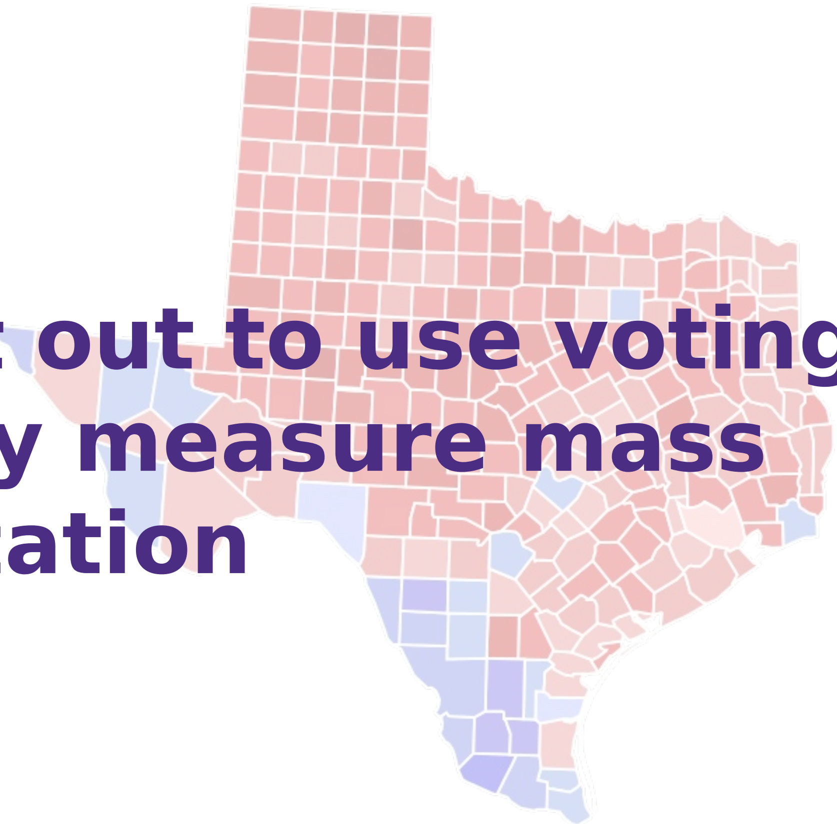
PNAS September 11, 2018 115 (37) 9216-9221; published ahead of print August 28, 2018
https://doi.org/10.1073/pnas.1804840115

Mass Polarization

EXISTING SCHOLARLY WORK

- **Increasing polarization in the electorate**
  (e.g., Abramowitz and Saunders, 2008)
- **Little gap in the center and comparatively moderate**
  (e.g., Fiorina and Abrams, 2008)

We set out to use voting data to directly measure mass polarization



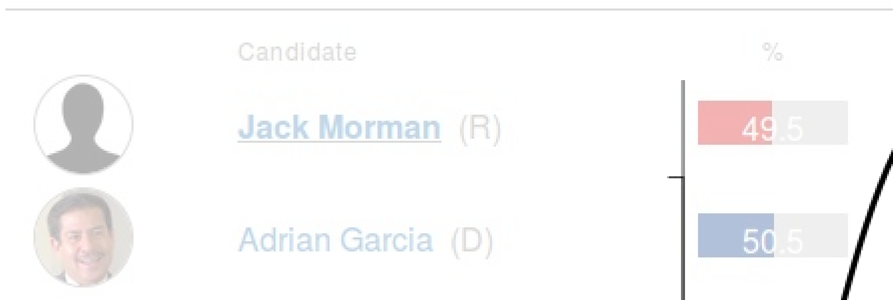
Precinct-Level Voting Data

What we have:

Candidate	%
Jack Morman (R)	49.5
Adrian Garcia (D)	50.5

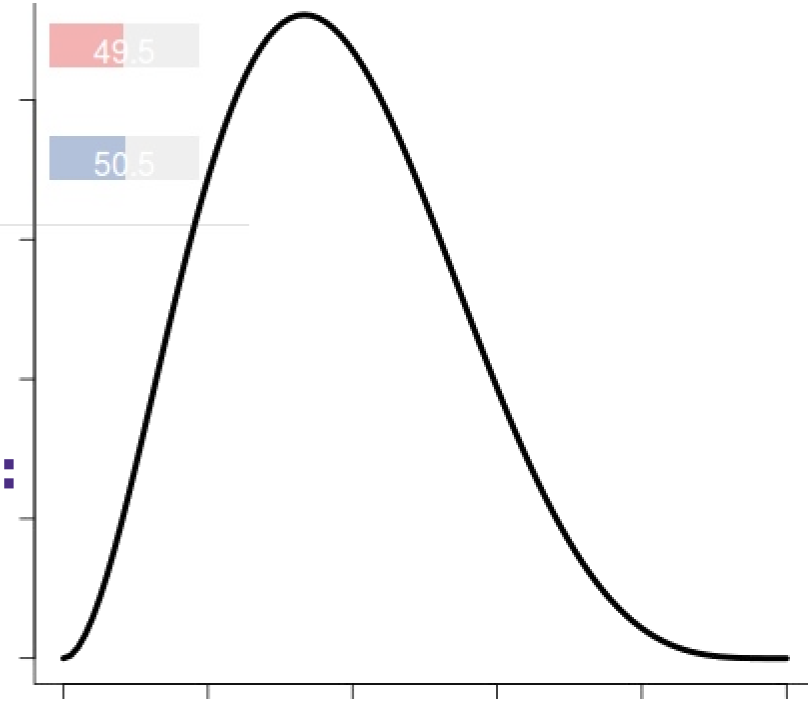
Precinct-Level Voting Data

What we have:



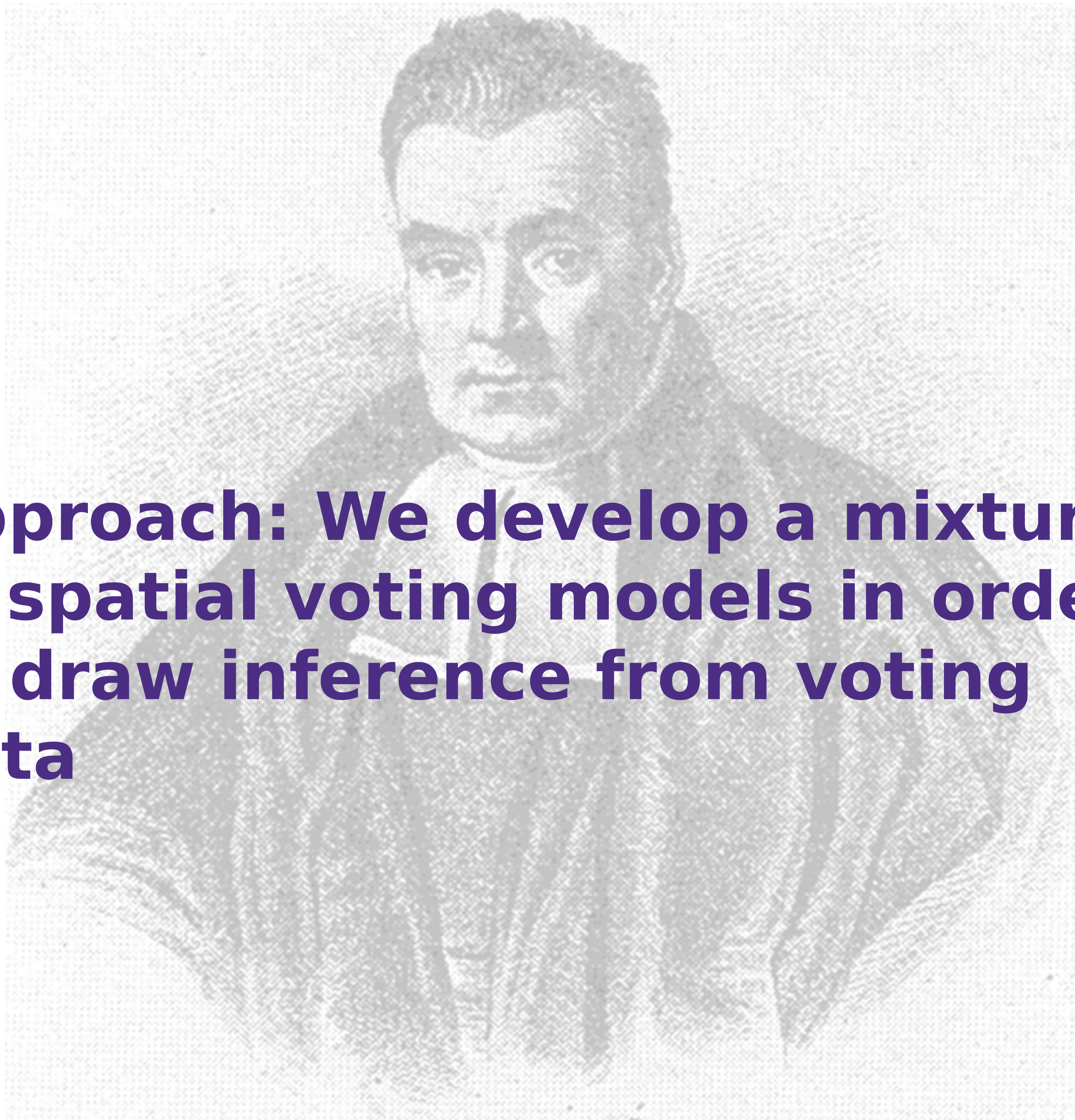
	Candidate	%
	Jack Morman (R)	49.5
	Adrian Garcia (D)	50.5

What we want:



Using Voter Data

- Challenge 1: Coarse candidate data
- Challenge 2: Censored voter data
- Challenge 3: Sparse data



Approach: We develop a mixture of spatial voting models in order to draw inference from voting data

This Talk

BAYESIAN MACHINE LEARNING FOR SOCIAL DATA SCIENCE

1) Overview of my Work

2) Intro to Polarization Model

3) **Brief Bayesian Inference Tutorial**

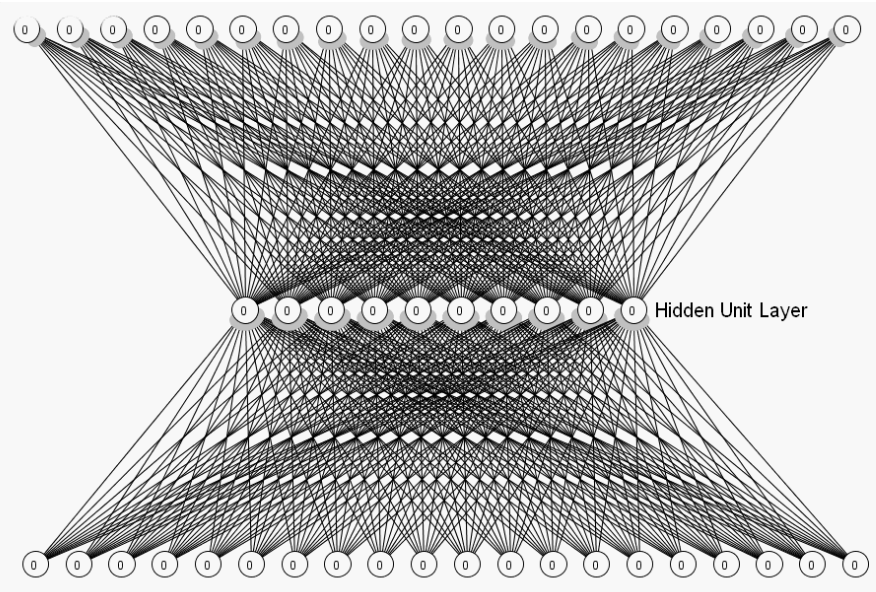
4) Polarization Model

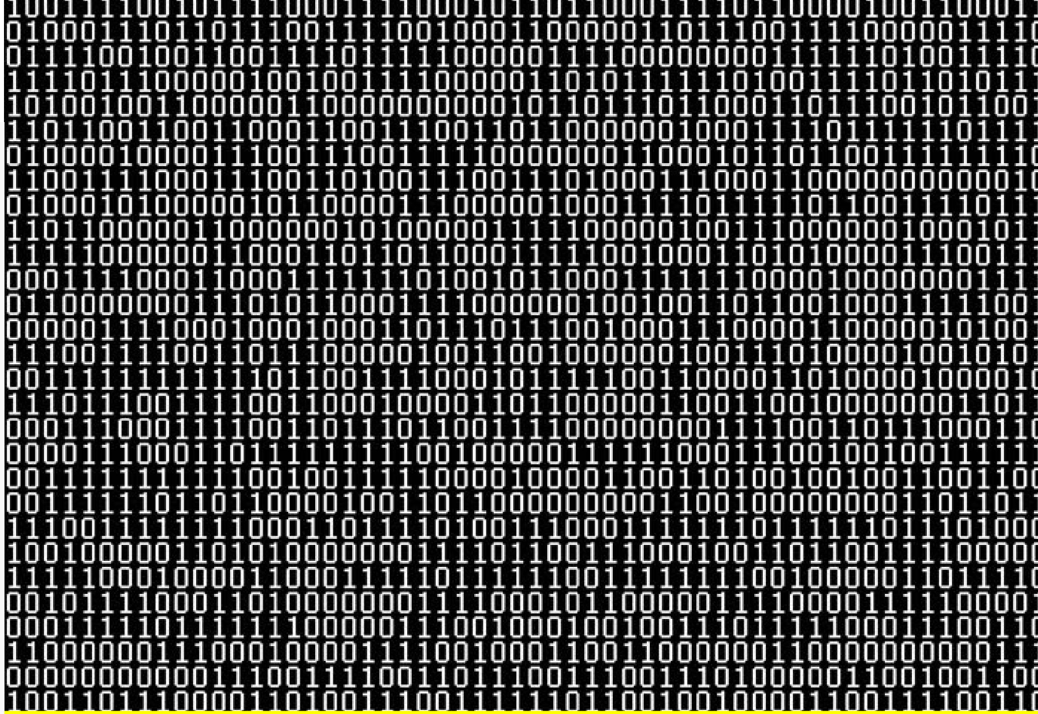
5) Ongoing and Future Work

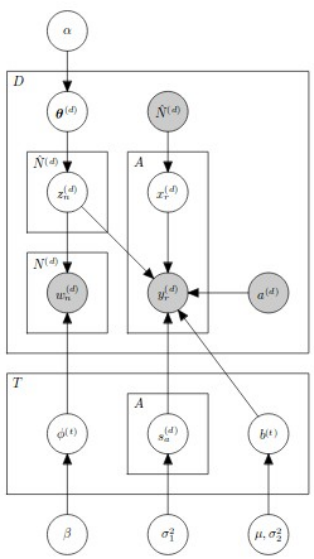
How many people here could write down Bayes' Rule from memory?

Why use Bayesian machine learning?

Neural Model



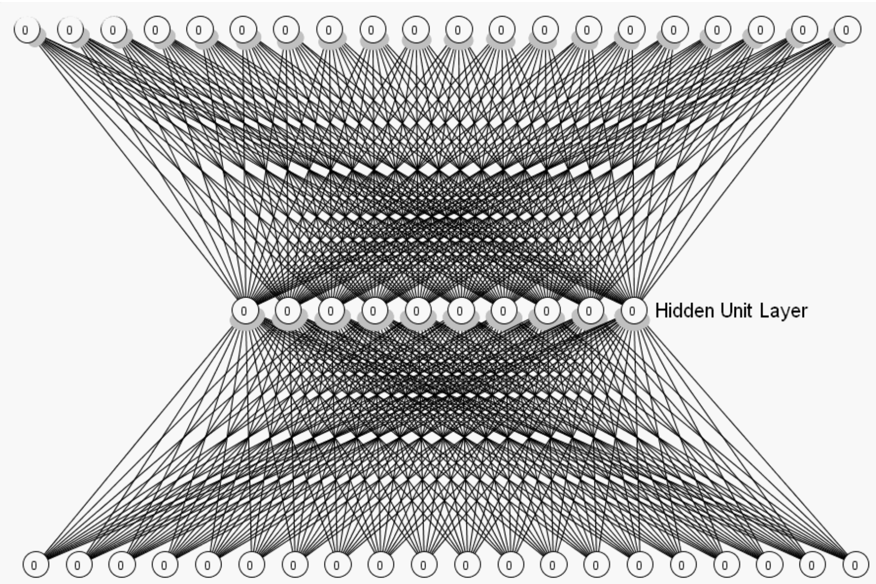


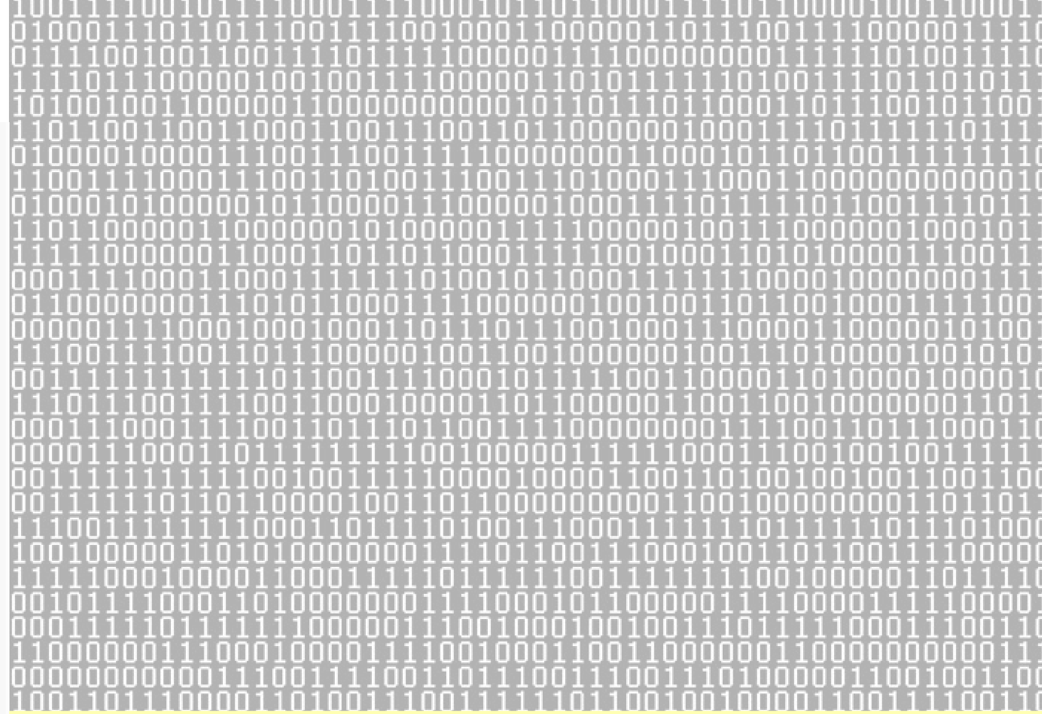


10 11101 011 10 001 1010 10

Bayesian Model

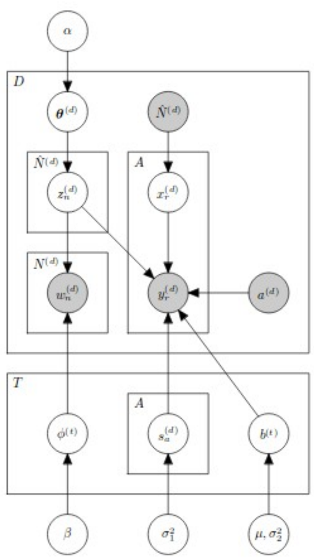
Neural Model





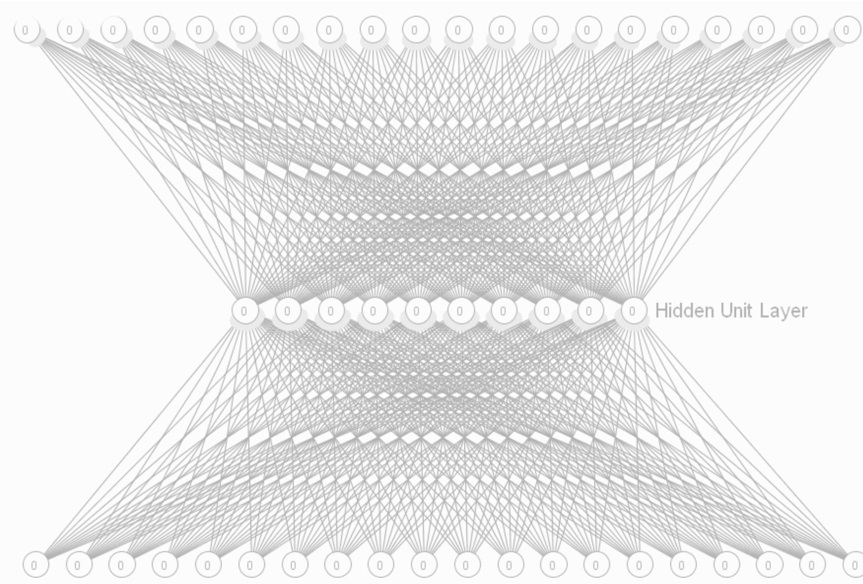
Stength #1:
Interpretable, structured models

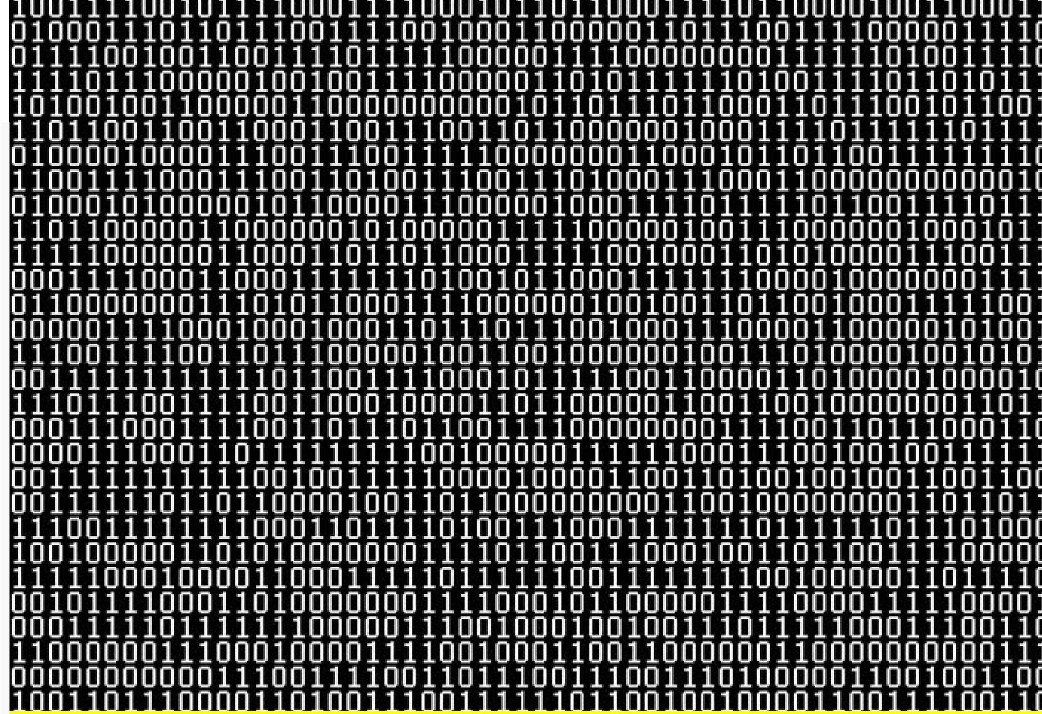
10 11101 011 10 001 1010 10



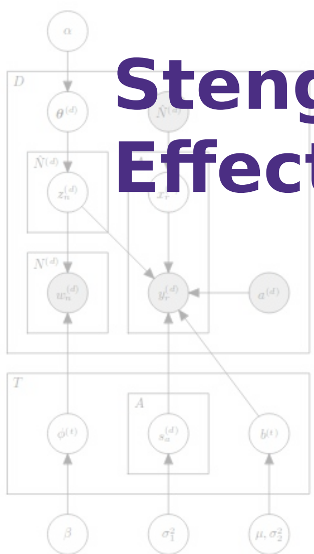
Bayesian Model

Neural Model





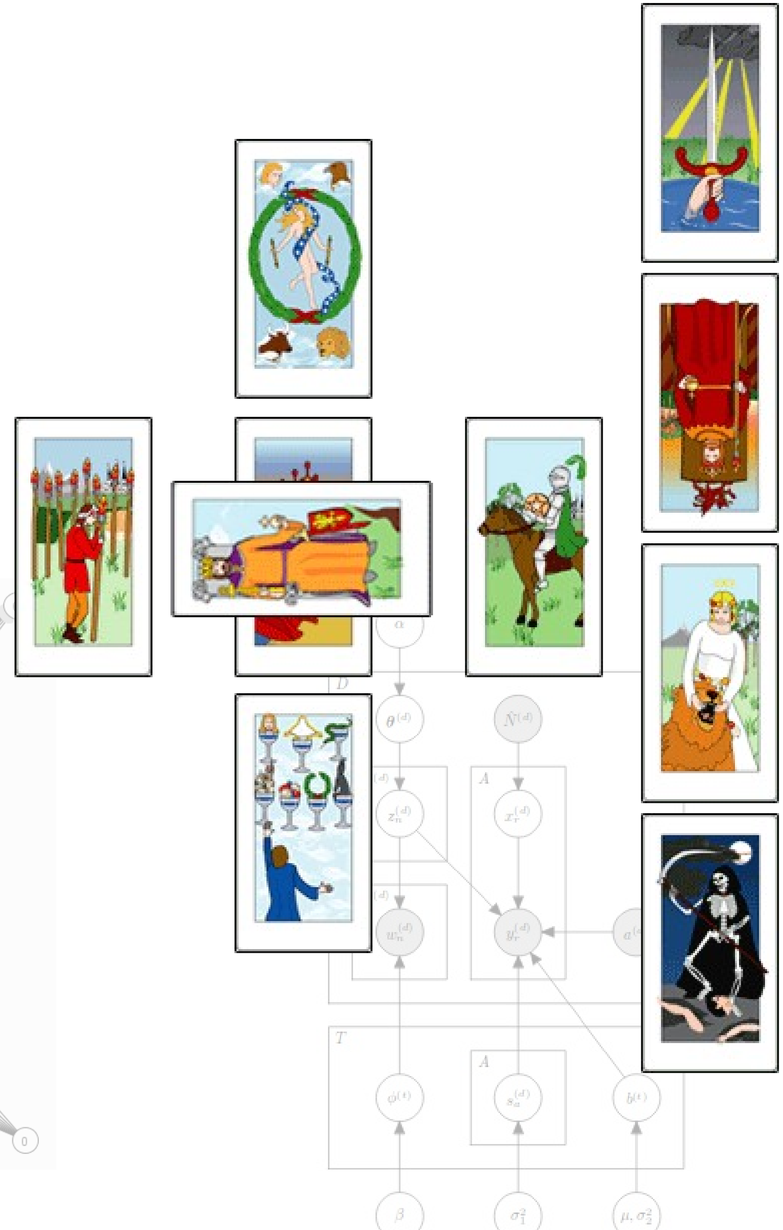
Stength #2:
Effective with small data



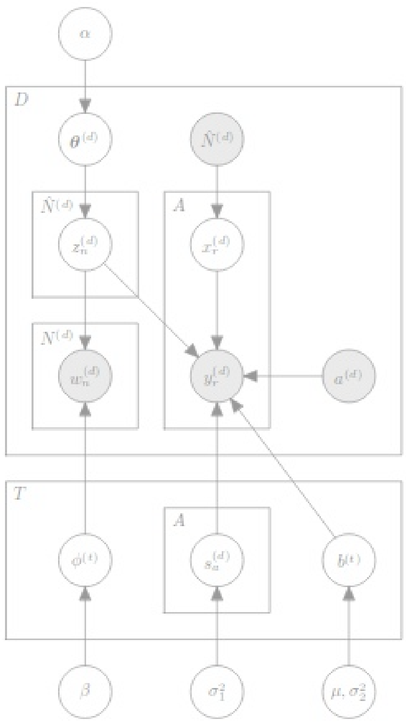
10 11101 011 10 001 1010 10

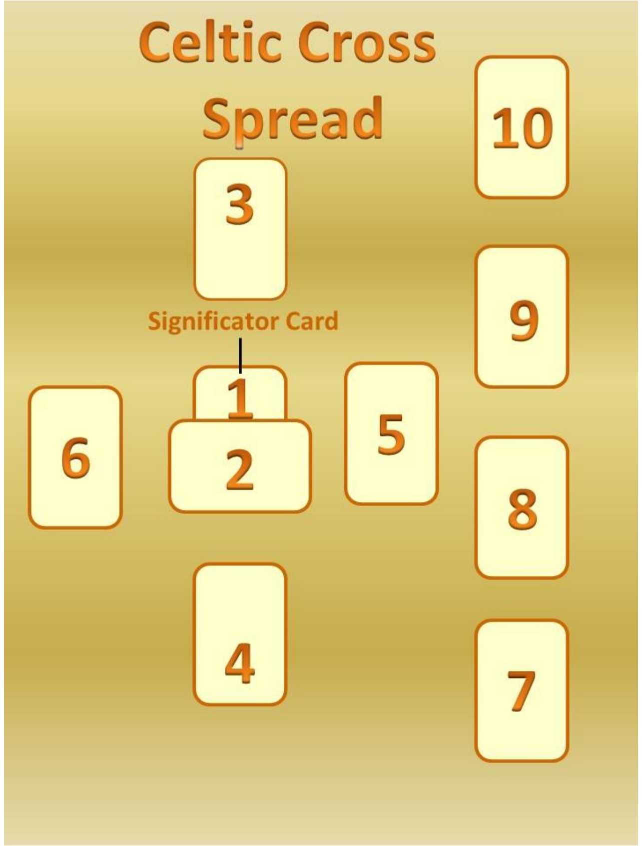
Bayesian Model


Hidden Unit Layer

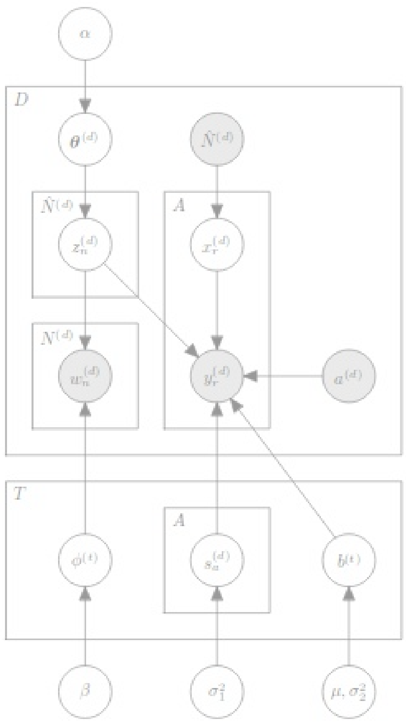


Bayesian Inference





Bayesian Inference



How many people here know the nitty-gritty of how Bayesian inference in mixture models works?

Intro to Bayesian Inference

BASIC MECHANICS

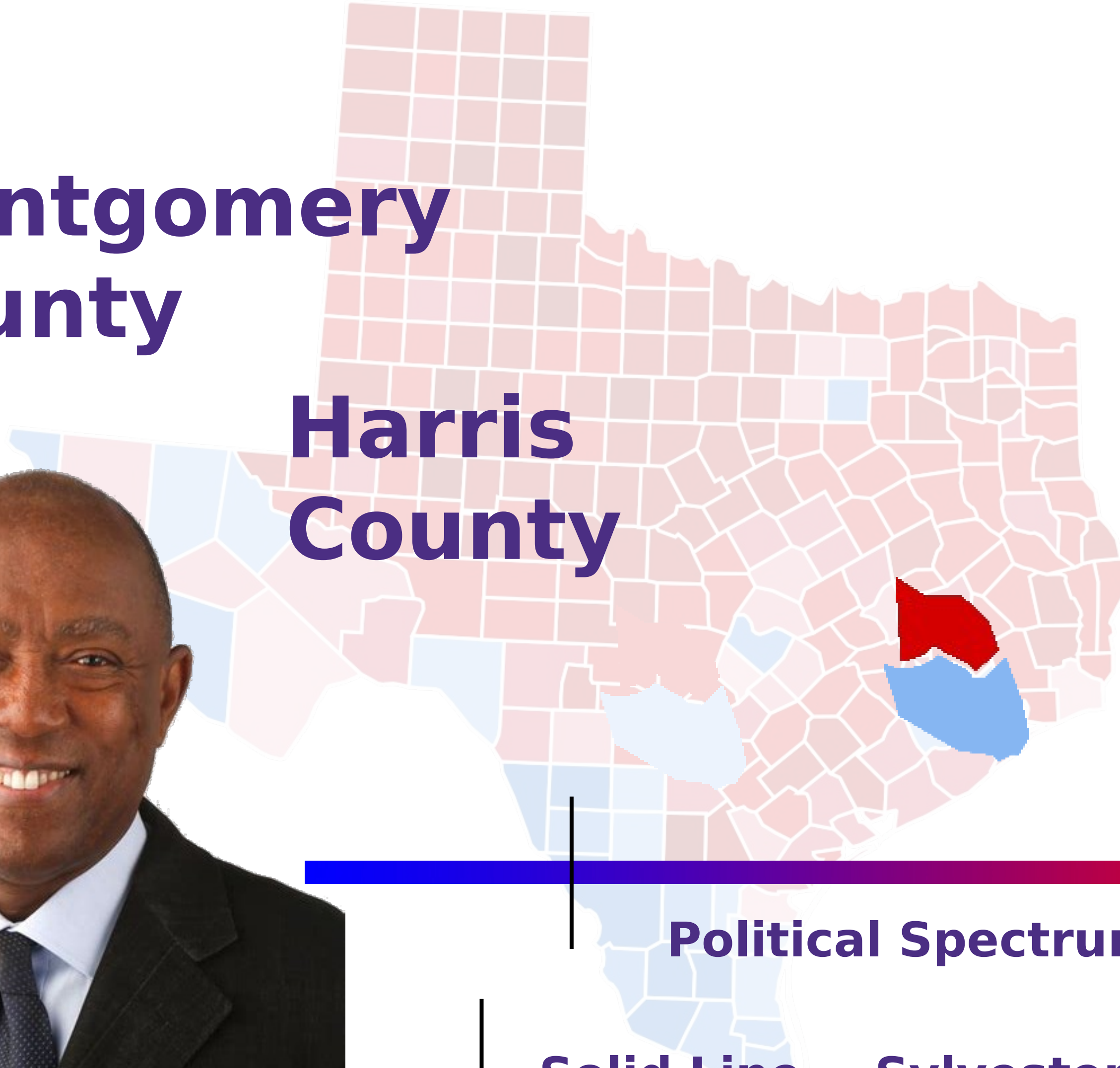
1) Identify data

2) Specify probabilistic generative model

3) Infer model parameters

Montgomery County

Harris County



Political Spectrum

Solid Line = Sylvester's Political Position

Intro to Bayesian Inference

BASIC MECHANICS

1) Identify data

2) Specify probabilistic generative model

3) Infer model parameters

STORY TELLING FOR EARTHLY SURVIVAL

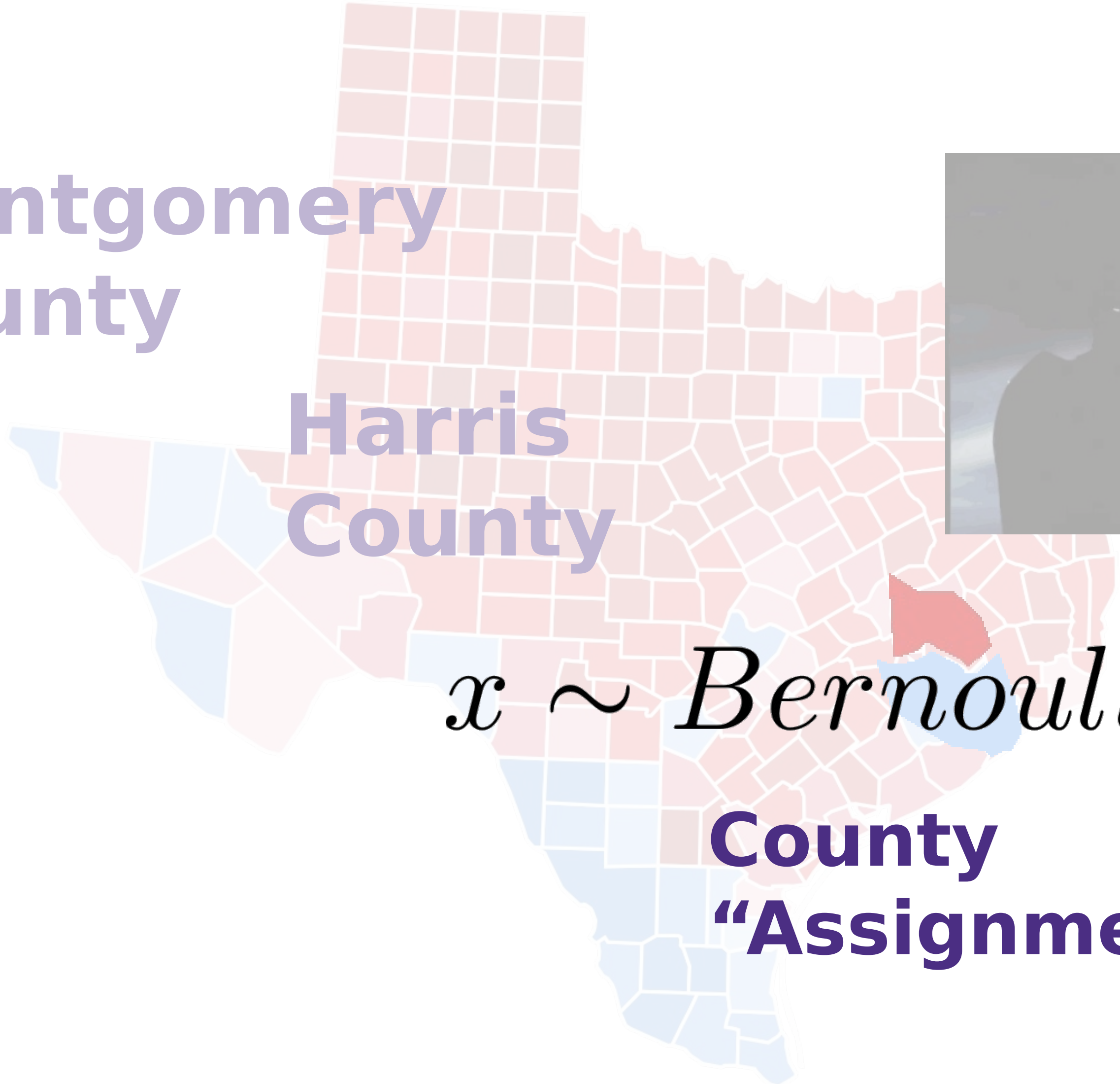
EXAMPLE GENERATIVE STORY

A CHOOSE-YOUR-OWN ADVENTURE, WRITTEN IN MATH

$$
\begin{aligned}
&x \sim Bernoulli(0.5) \\
&\textbf{if } x = 0\text{:} \\
&\qquad y \sim Normal(-1, 1) \\
&\textbf{if } x = 1\text{:} \\
&\qquad y \sim Normal(1, 1)
\end{aligned}
$$

Montgomery County

Harris County

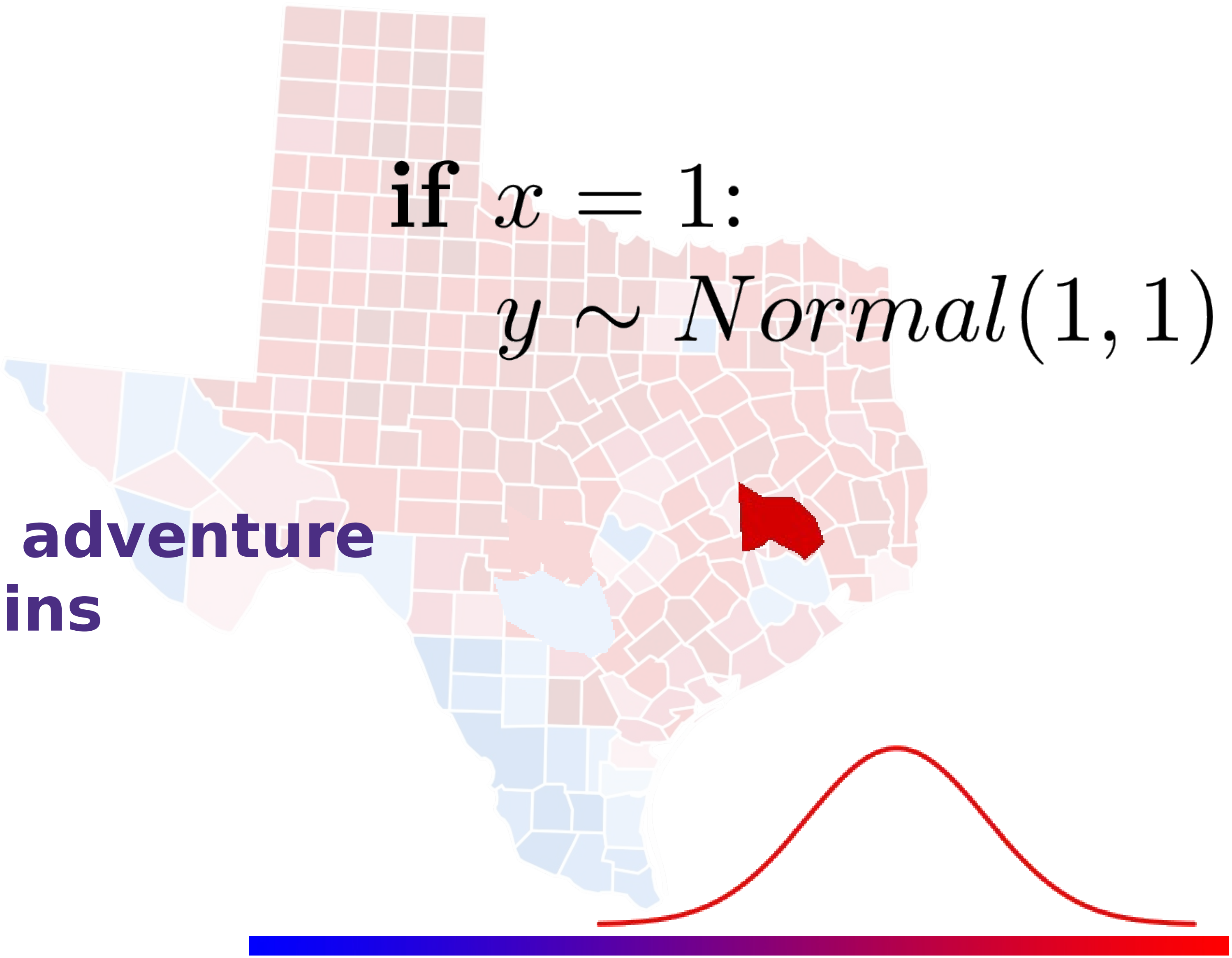


$$x \sim Bernoulli(0.5)$$

County "Assignment"

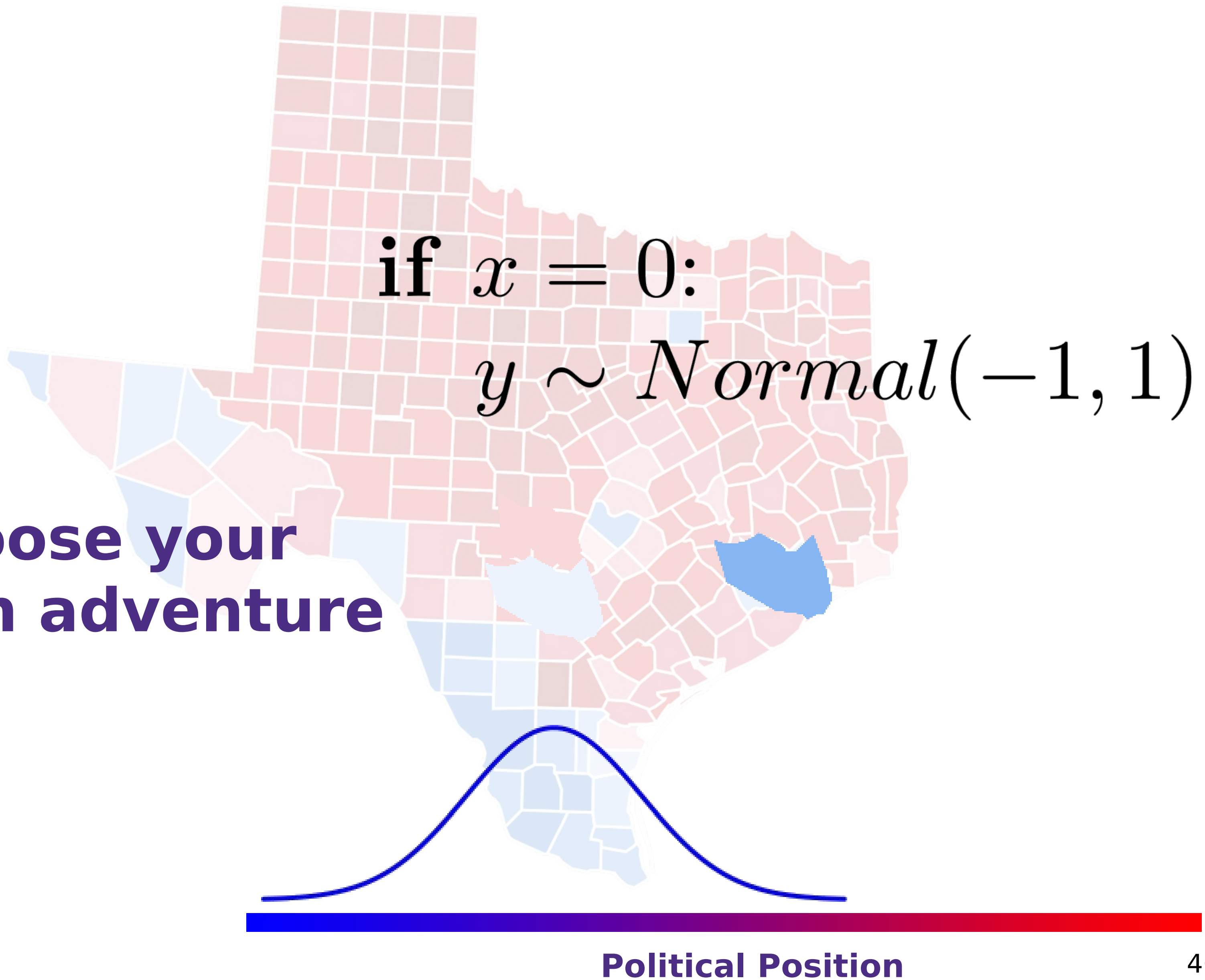
$$\textbf{if } x = 1:$$
$$y \sim Normal(1, 1)$$

The adventure begins

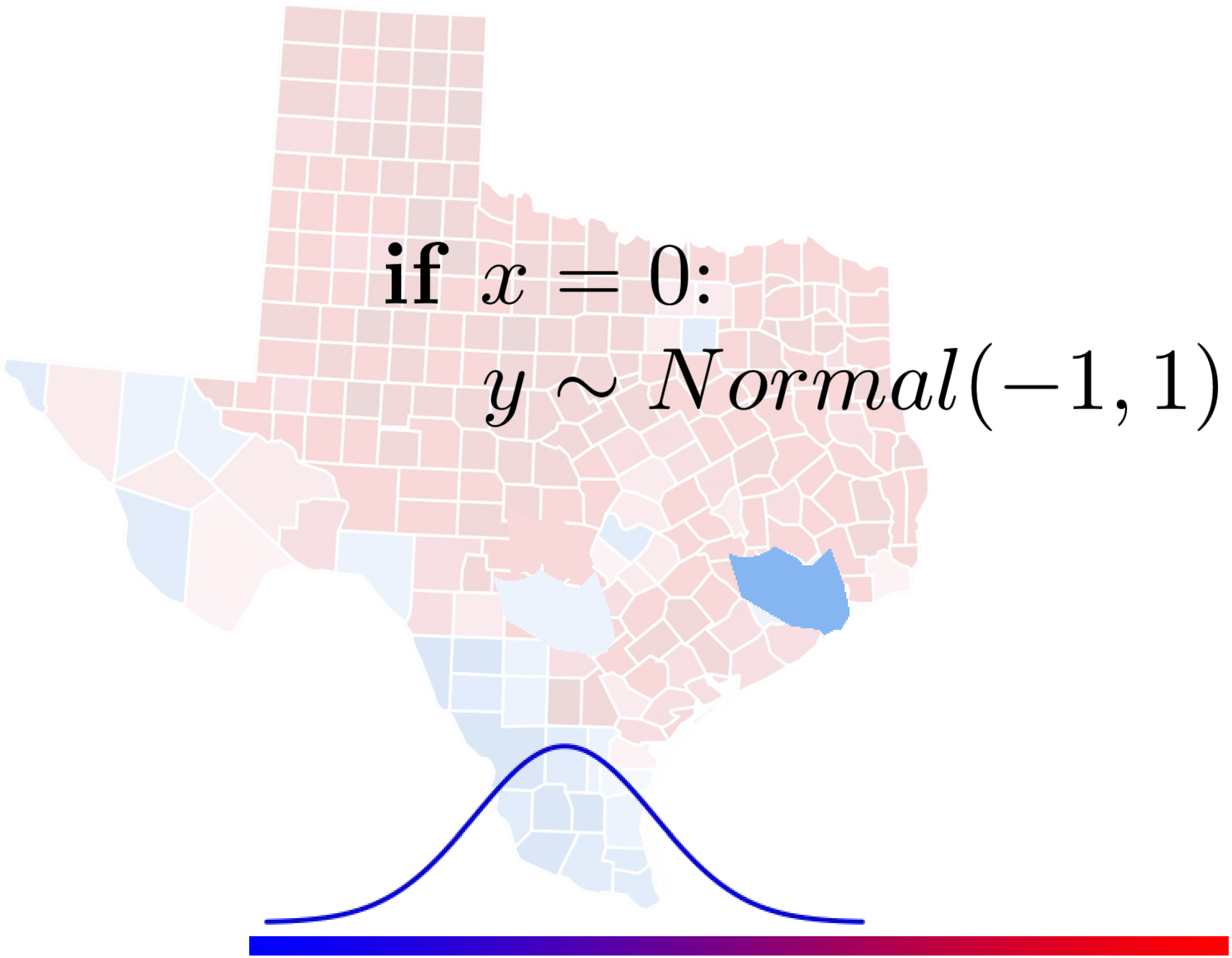


$$\textbf{if } x = 0\text{:}$$
$$y \sim Normal(-1, 1)$$

Choose your own adventure



$$\textbf{if } x = 0:\quad y \sim Normal(-1, 1)$$

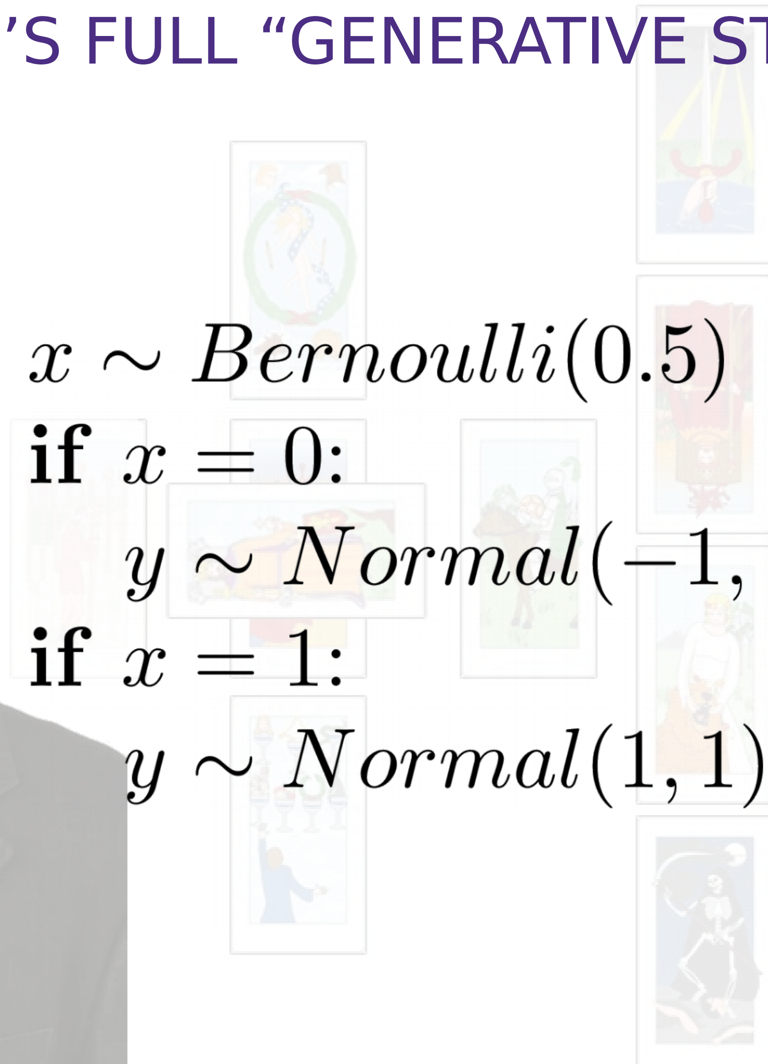


Political Position

A narrative structure that organizes information

Intro to Bayesian Inference

SYLVESTER'S FULL "GENERATIVE STORY"



$$
\begin{aligned}
&x \sim Bernoulli(0.5) \\
&\textbf{if } x = 0\text{:} \\
&\quad y \sim Normal(-1, 1) \\
&\textbf{if } x = 1\text{:} \\
&\quad y \sim Normal(1, 1)
\end{aligned}
$$

Intro to Bayesian Inference

BASIC MECHANICS

1) Identify data

2) Specify probabilistic generative model

3) Infer model parameters

Inverting the Generative Story

TAKING SYLVESTER'S STORY

$$x \sim Bernoulli(0.5)$$

$$\textbf{if } x = 0\text{:}$$

$$\quad y \sim Normal(-1, 1)$$

$$\textbf{if } x = 1\text{:}$$

$$\quad y \sim Normal(1, 1)$$

$$\textbf{output } y$$

Intro to Bayesian Inference

INFERENCE PROCESS

$x \sim Bernoulli(0.5)$
if $x = 0$:
  $y \sim Normal(-1, 1)$
if $x = 1$:
  $y \sim Normal(1, 1)$
output y

Observe!

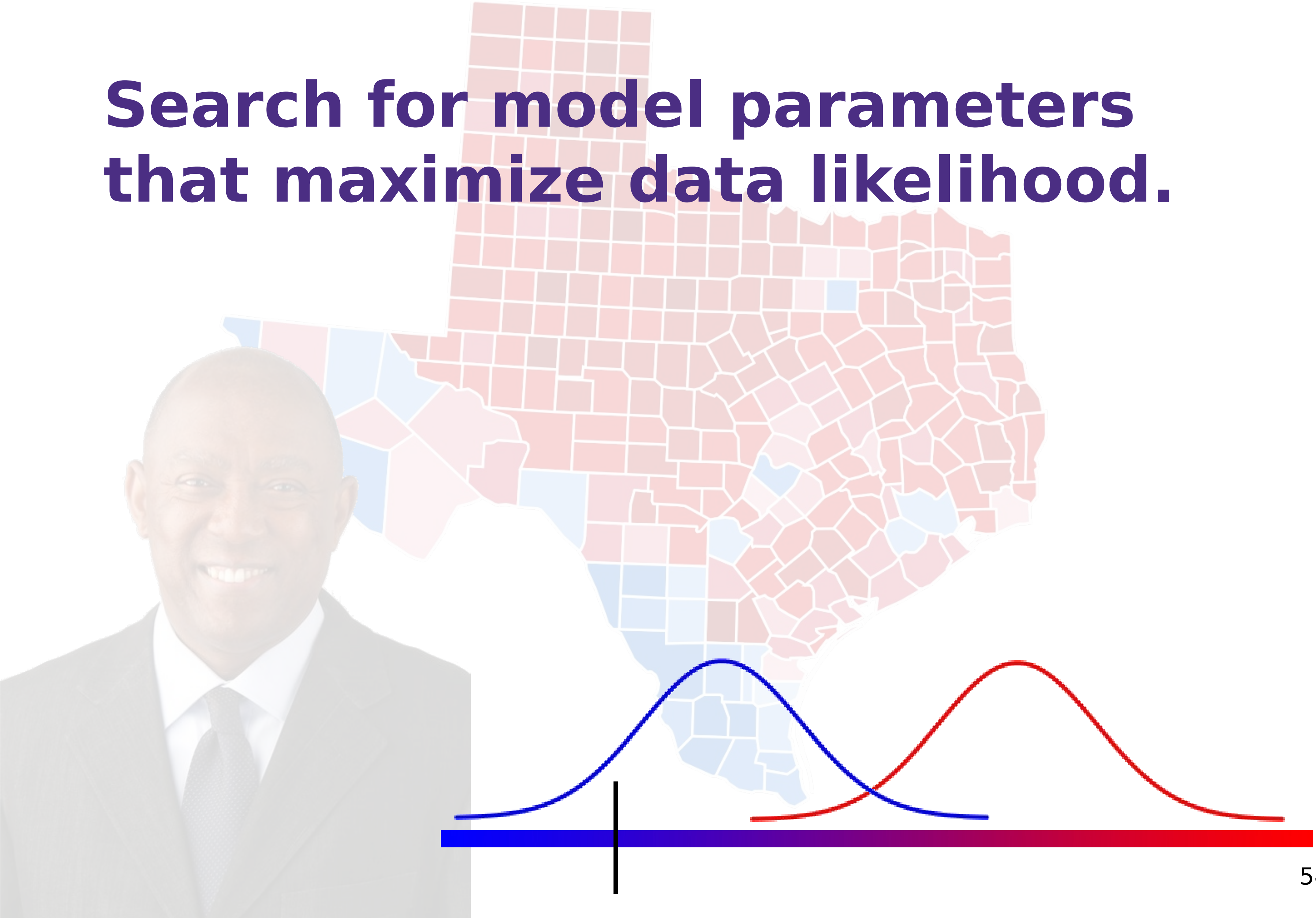
Intro to Bayesian Inference

INFERENCE PROCESS

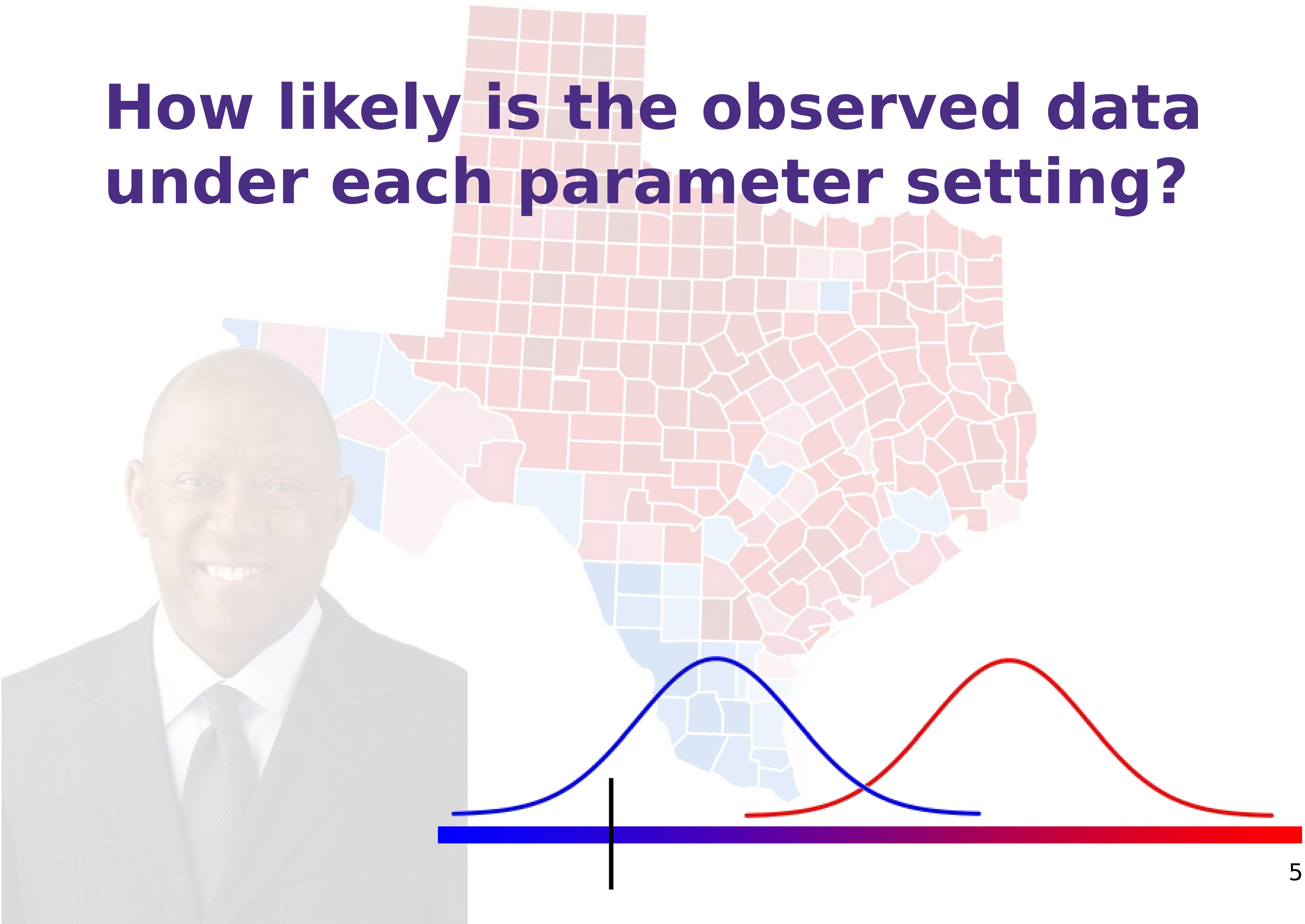
$$
\begin{aligned}
&x \sim Bernoulli(0.5)\\
&\textbf{if } x = 0\text{:}\\
&\quad y \sim Normal(-1, 1)\\
&\textbf{if } x = 1\text{:}\\
&\quad y \sim Normal(1, 1)\\
&\textbf{output } y
\end{aligned}
$$

Infer!

Search for model parameters that maximize data likelihood.

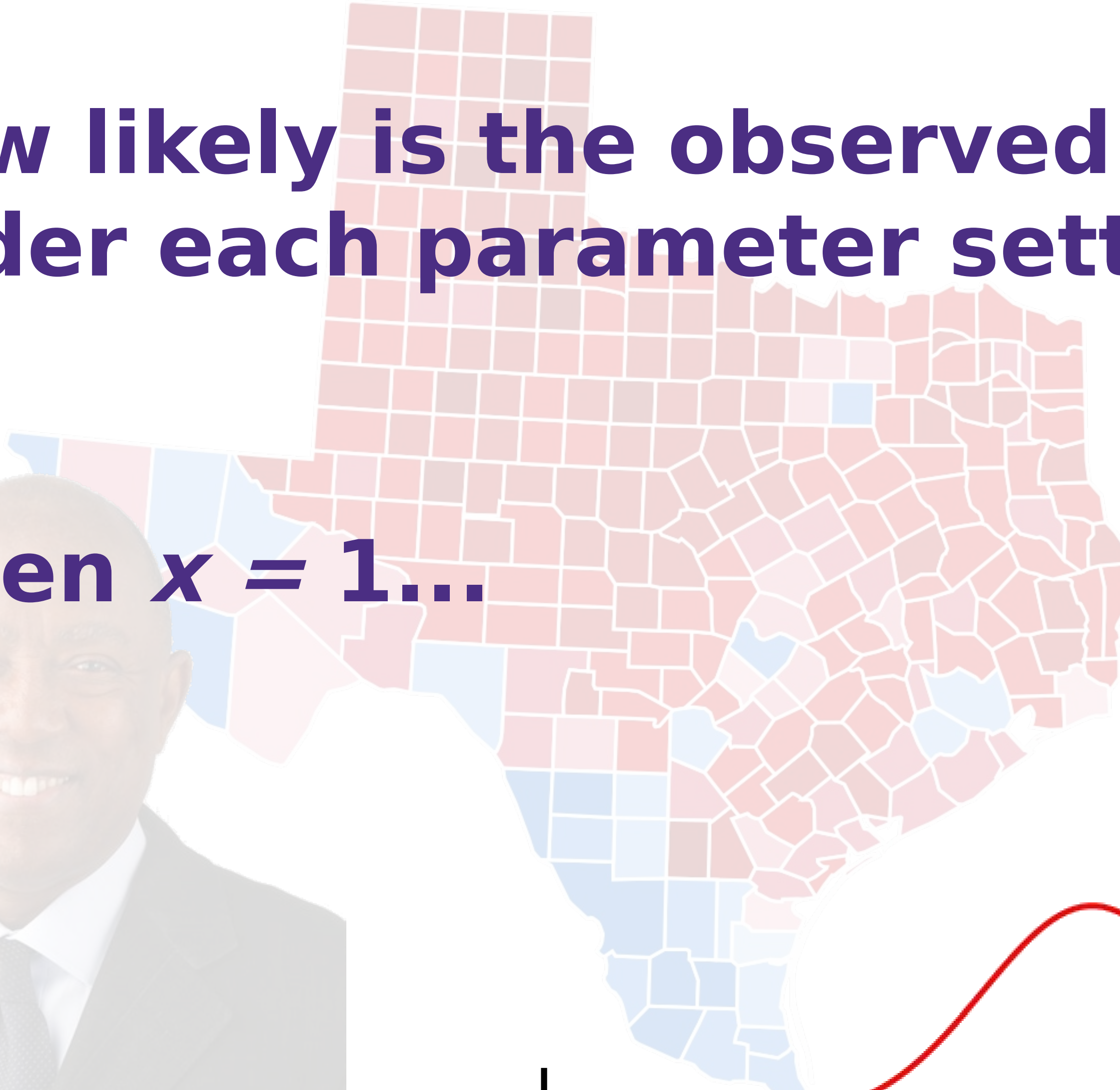


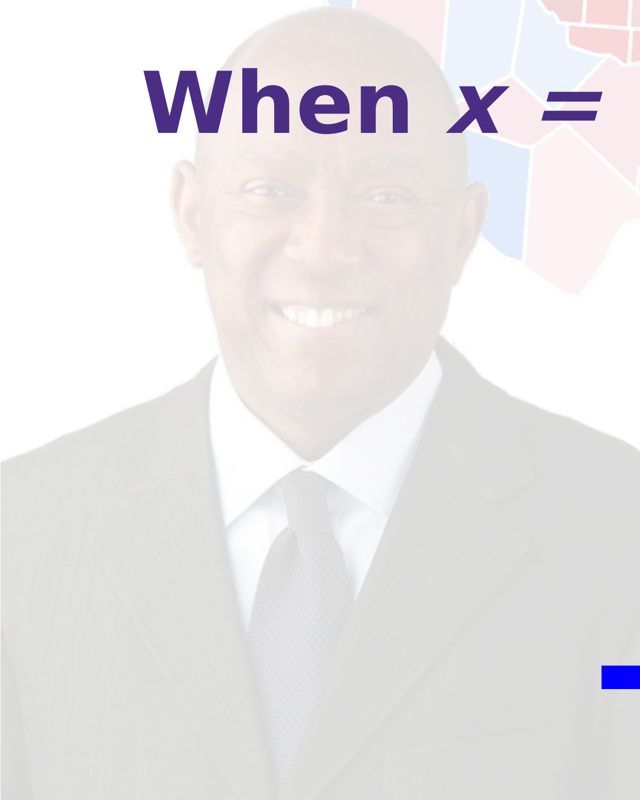
How likely is the observed data under each parameter setting?

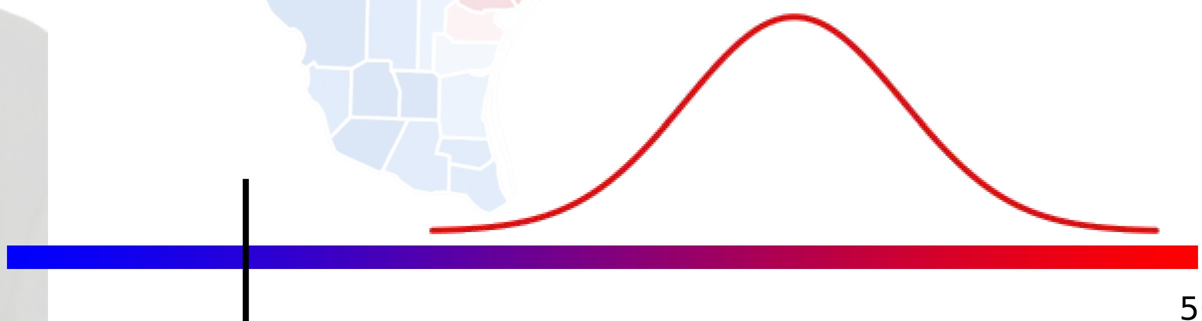


How likely is the observed data under each parameter setting?

When $x = 1$...

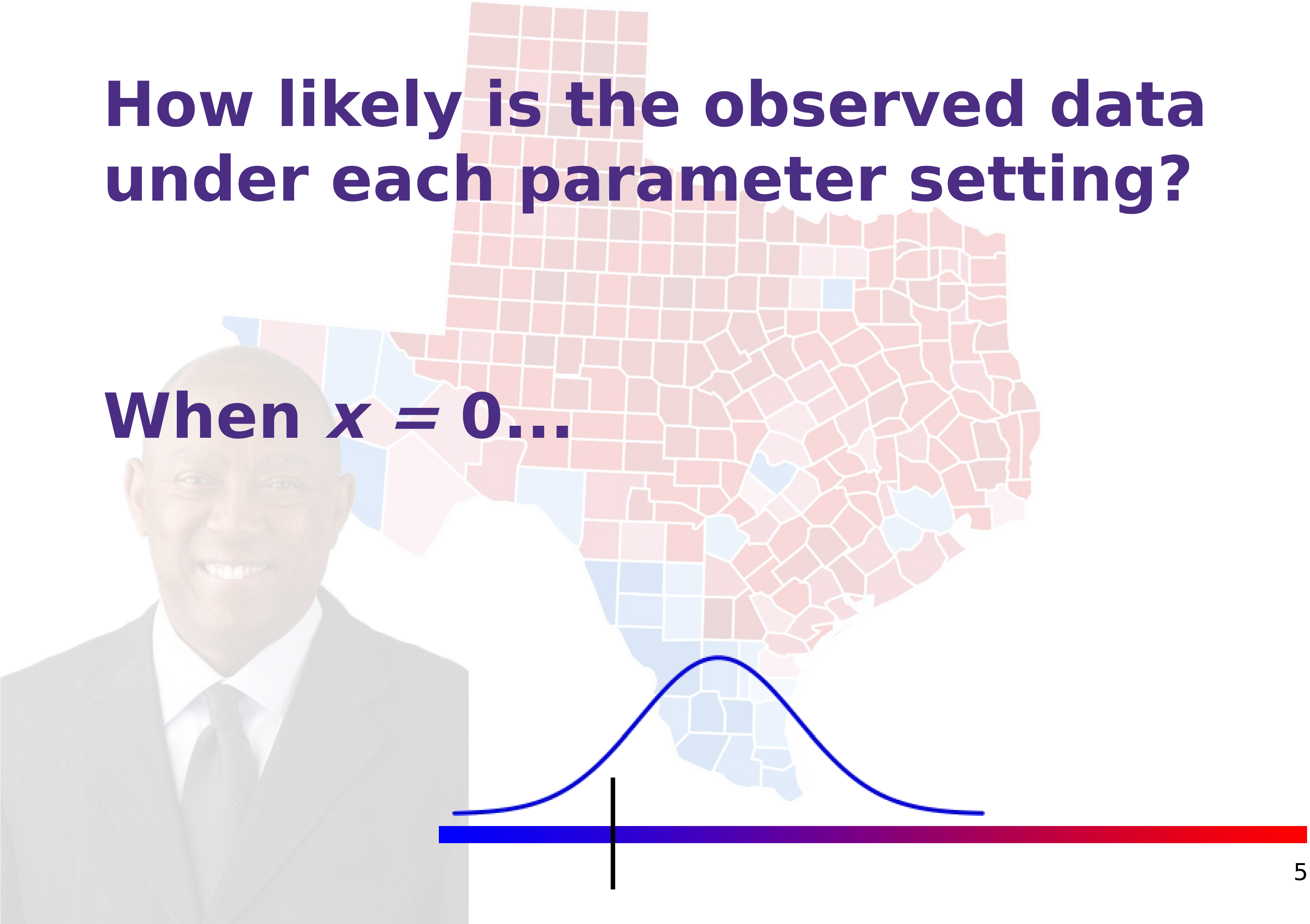






How likely is the observed data under each parameter setting?

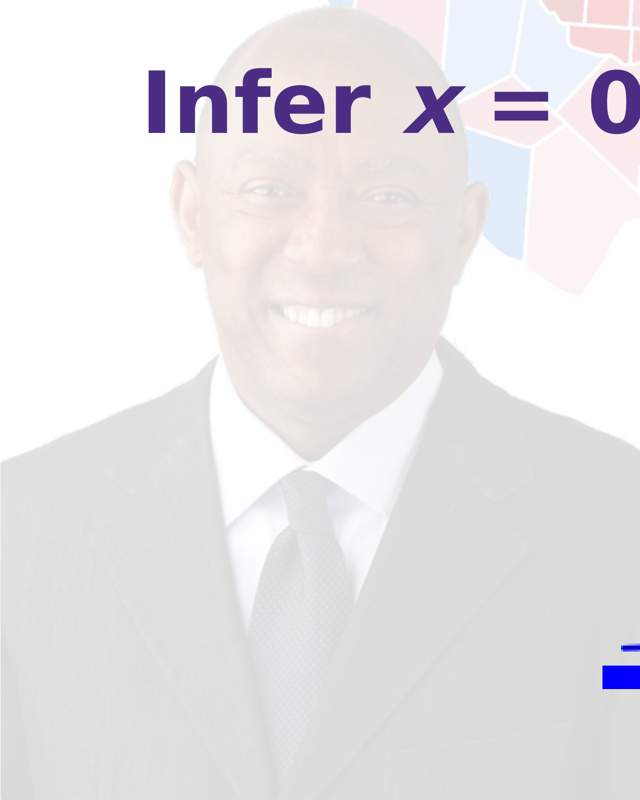
When $x = 0$…

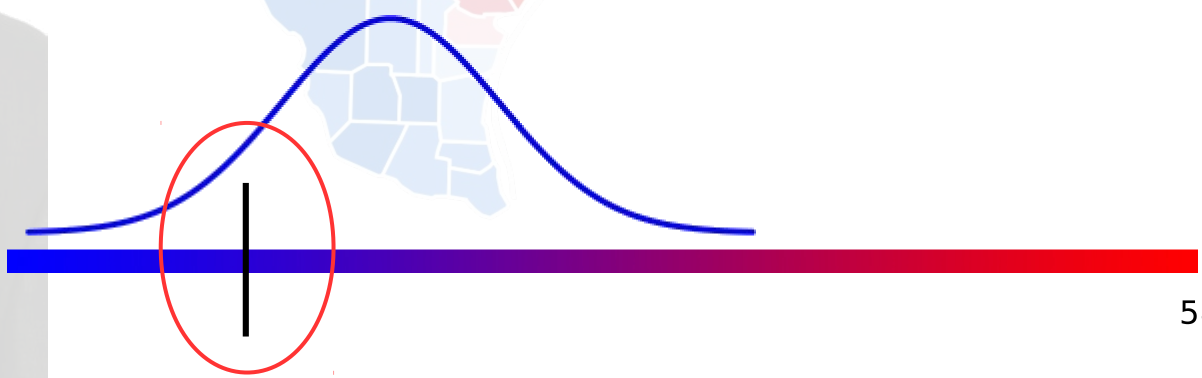


Which parameter setting best explains the observed data?

Infer $x = 0$!

Automated method





Intro to Bayesian Inference

DISCUSSION

- **Mixture of Normal distributions**
- **Same basic model structure as in “topic modeling”**
- **Same basic model structure as in the application in this talk**

This Talk

BAYESIAN MACHINE LEARNING FOR SOCIAL DATA SCIENCE

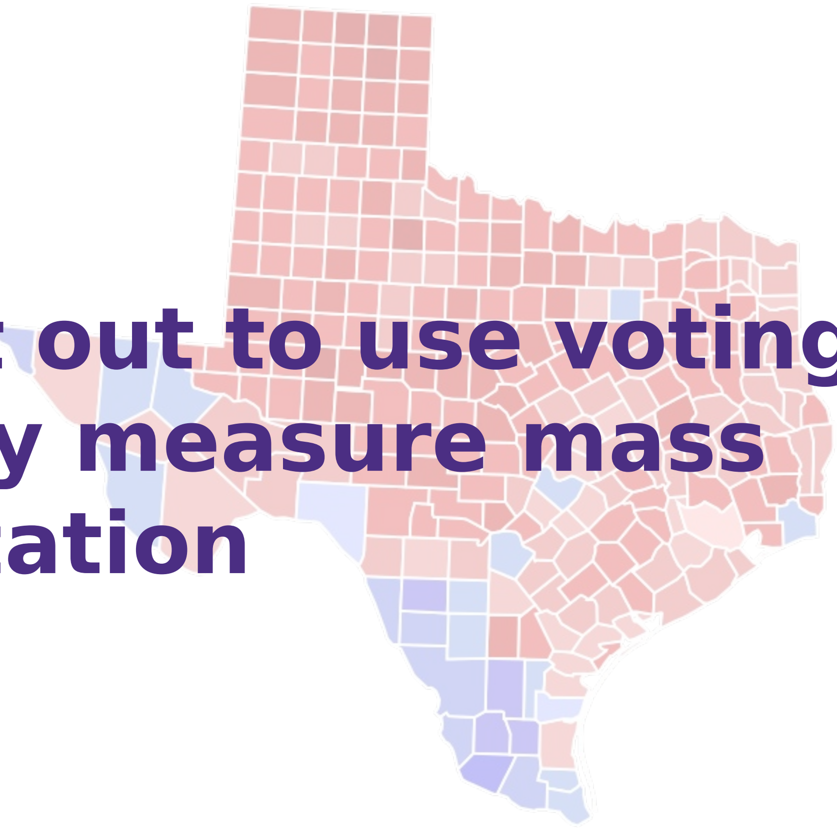
1) Overview of my Work

2) Intro to Polarization Model

3) Brief Bayesian Inference Tutorial

4) Polarization Model

5) Ongoing and Future Work



We set out to use voting data to directly measure mass polarization

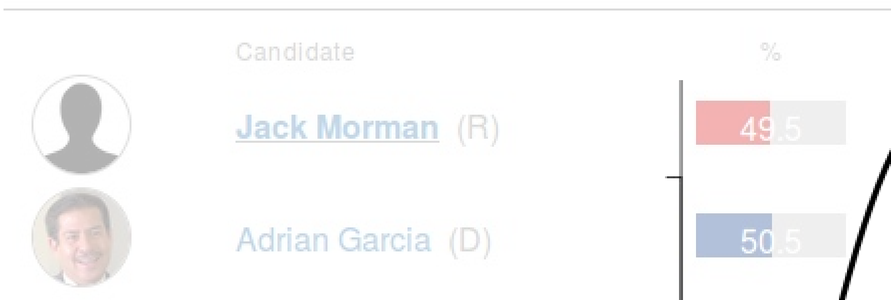
Precinct-Level Voting Data

What we have:

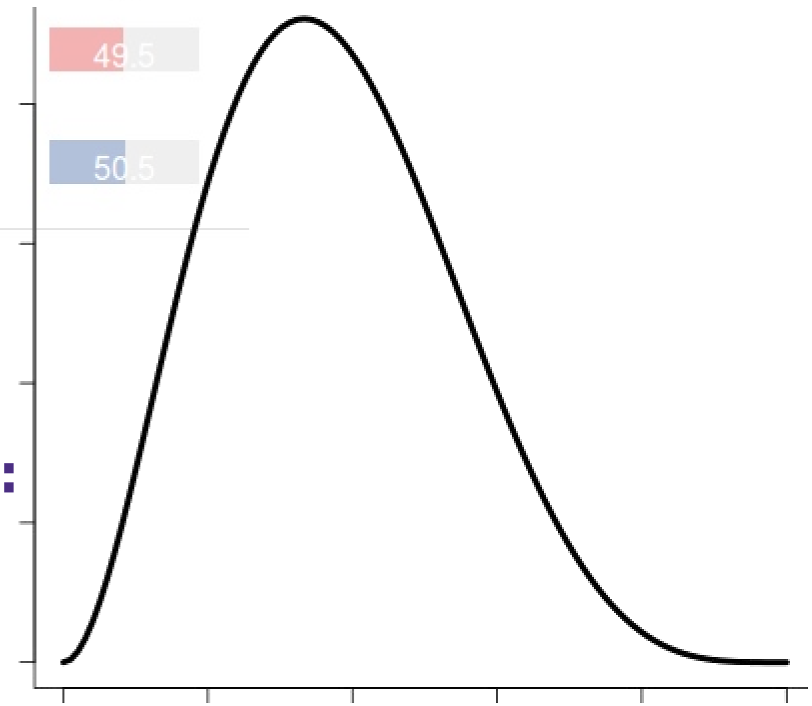
	Candidate	%
	Jack Morman (R)	49.5
	Adrian Garcia (D)	50.5

Precinct-Level Voting Data

What we have:



What we want:



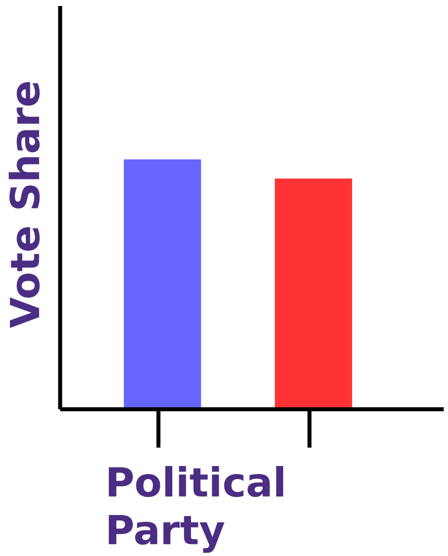
Using Voter Data

THREE CHALLENGES TO A SIMPLE STORY

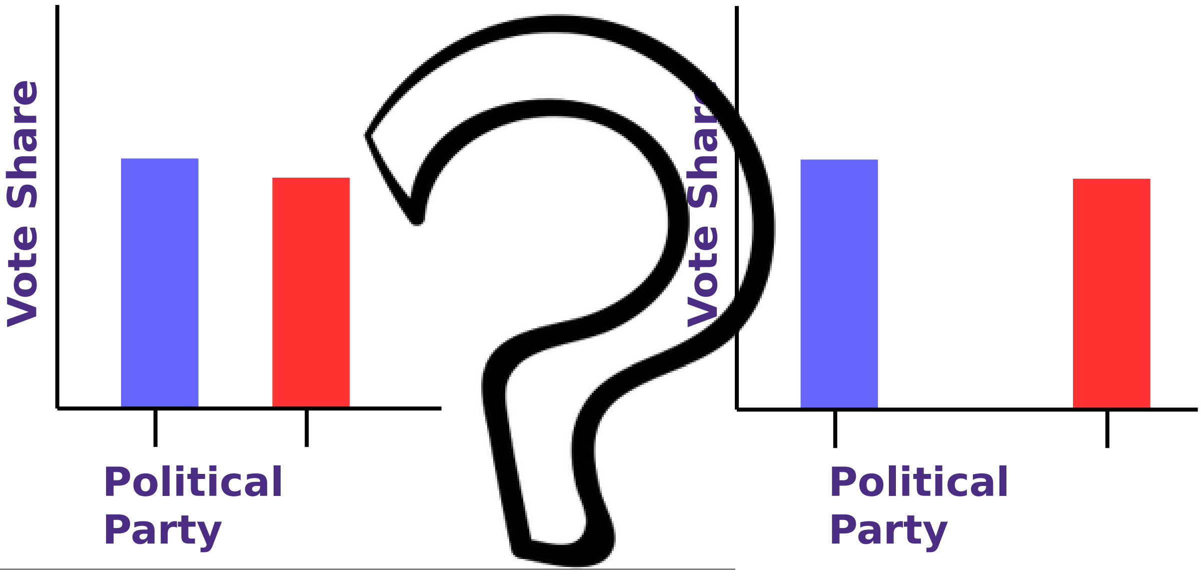
- **Challenge 1: Coarse candidate data**
- **Challenge 2: Censored voter data**
- **Challenge 3: Sparse data**

Challenge 1: Coarse Data

NO MEASUREMENT OF CANDIDATE POSITIONS



Challenge 1: Coarse Data



Assumption 1: CF-Scores

- Campaign finance (CF) scores
- Alternative to DW-NOMINATE
- Scores for all candidates

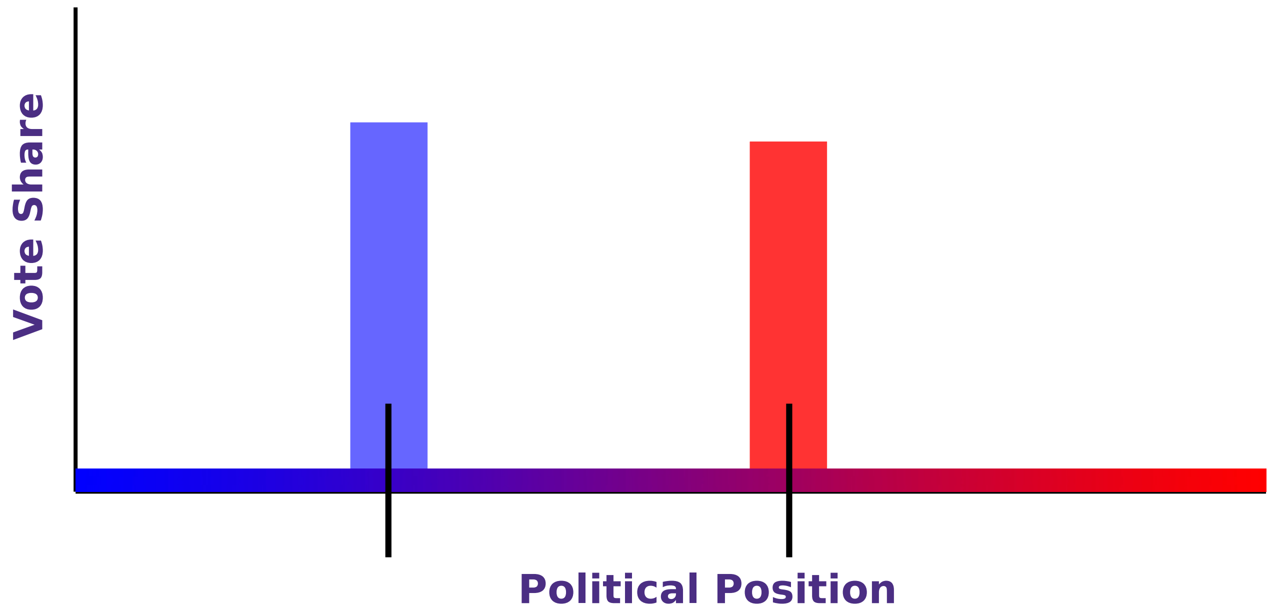
Political Position

Solid Line = Candidate Position

(Bonica, 2014)

Challenge 2: Censored Data

NO DIRECT OBSERVATION OF VOTER POSITIONS

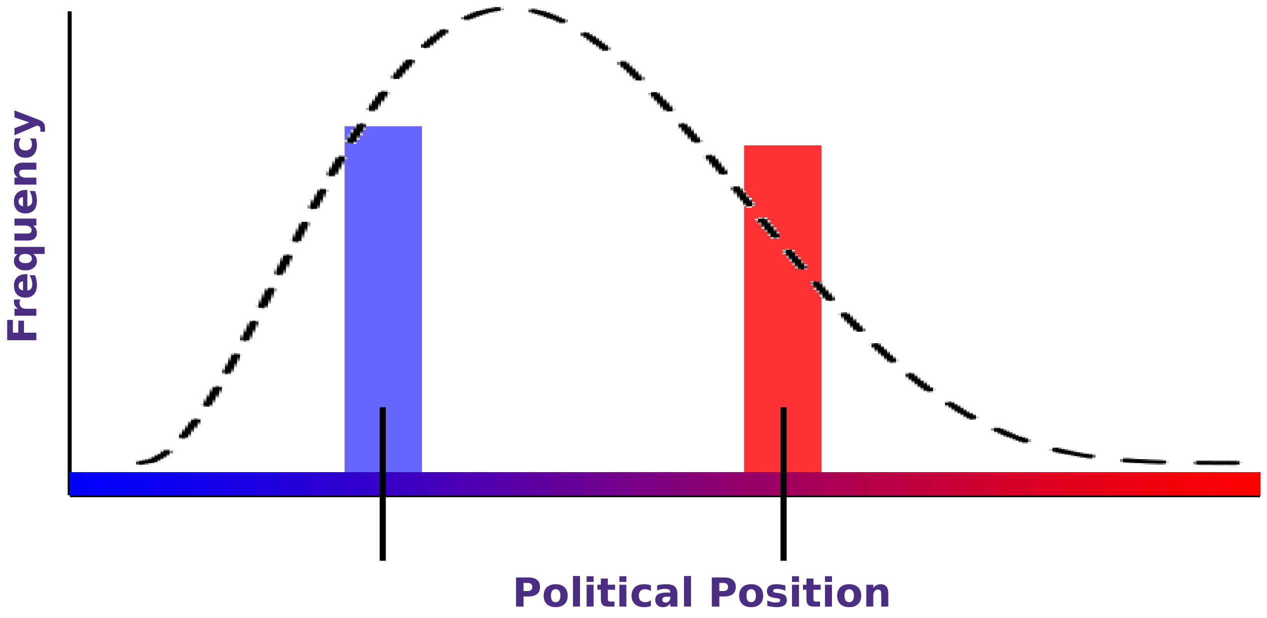


	Candidate	%
silhouette	Jack Morman (R)	49.5
	Adrian Garcia (D)	50.5

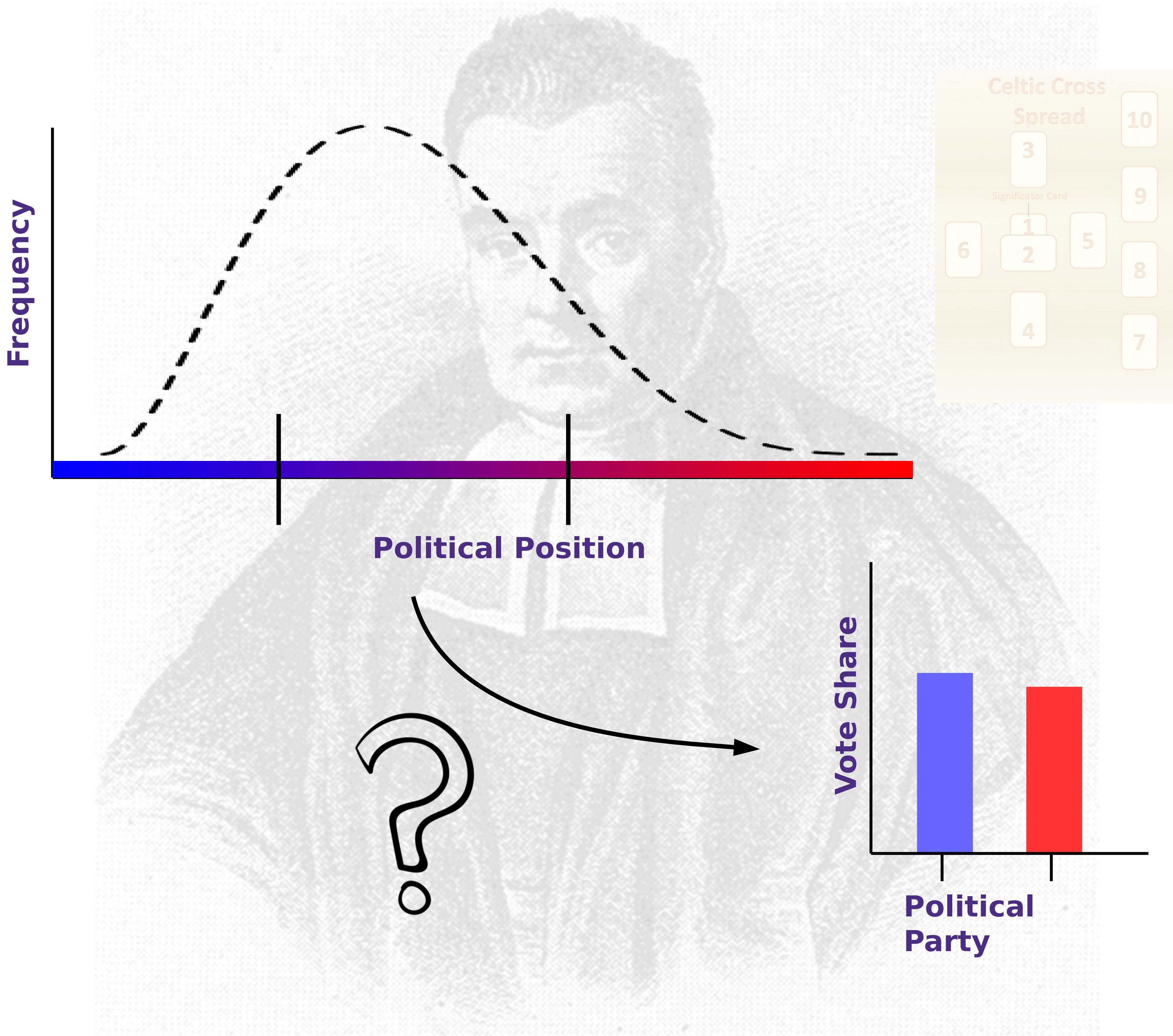
Solid Line = Candidate Position

Challenge 2: Censored Data

VOTER POSITIONS MAY DIFFER FROM CANDIDATES’



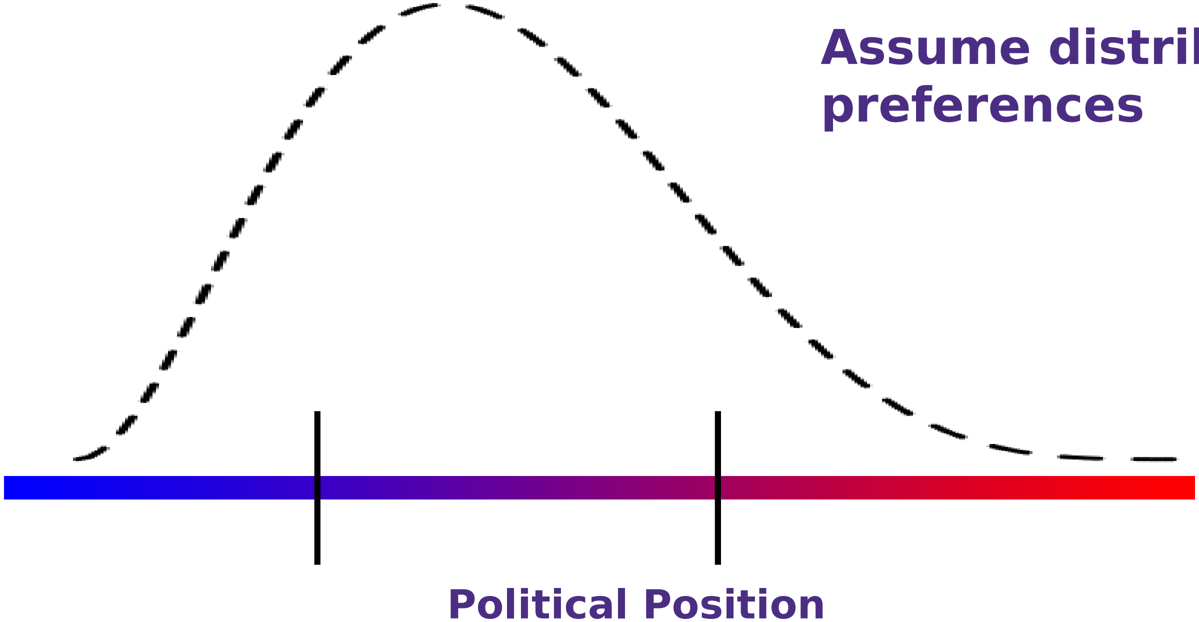
Dotted Curve = Voter Distribution (Unknown)
Solid Line = Candidate Position


Frequency
Political Position
Vote Share
Political Party
Celtic Cross Spread
Significator Card
1
2
3
4
5
6
7
8
9
10

Assumption 2: Spatial Voting

THE FIRST CHAPTER OF OUR "GENERATIVE STORY"

Assume distribution of voter preferences



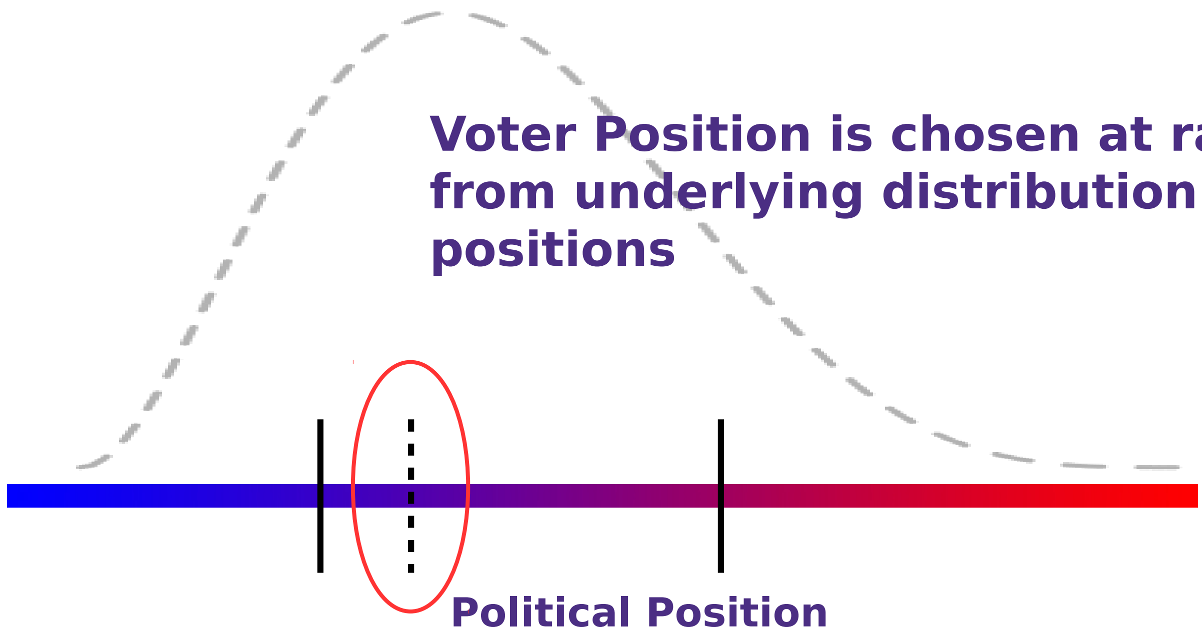
Solid Line = Candidate Position (Known)

(Downs, 1957)

Assumption 2: Spatial Voting

THE FIRST CHAPTER OF OUR “GENERATIVE STORY”

Voter Position is chosen at random from underlying distribution of political positions



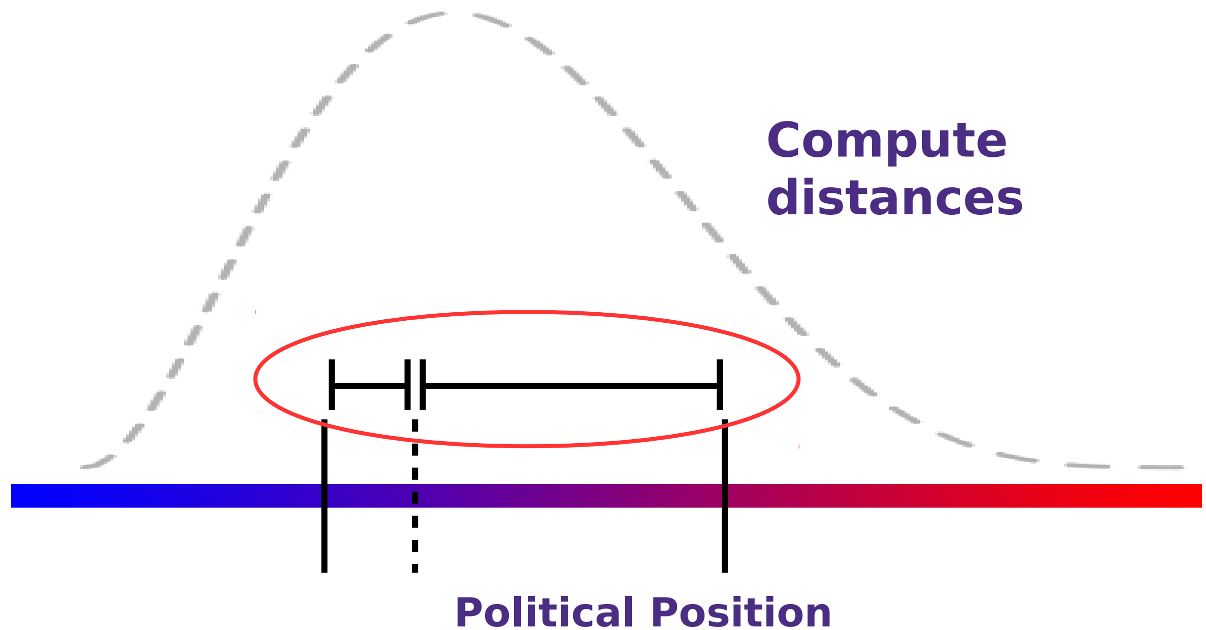
Solid Line = Candidate Position (Known)

Dotted Line = Voter Position (Unknown)

(Downs, 1957)

Assumption 2: Spatial Voting

THE FIRST CHAPTER OF OUR “GENERATIVE STORY”



Solid Line = Candidate Position (Known)

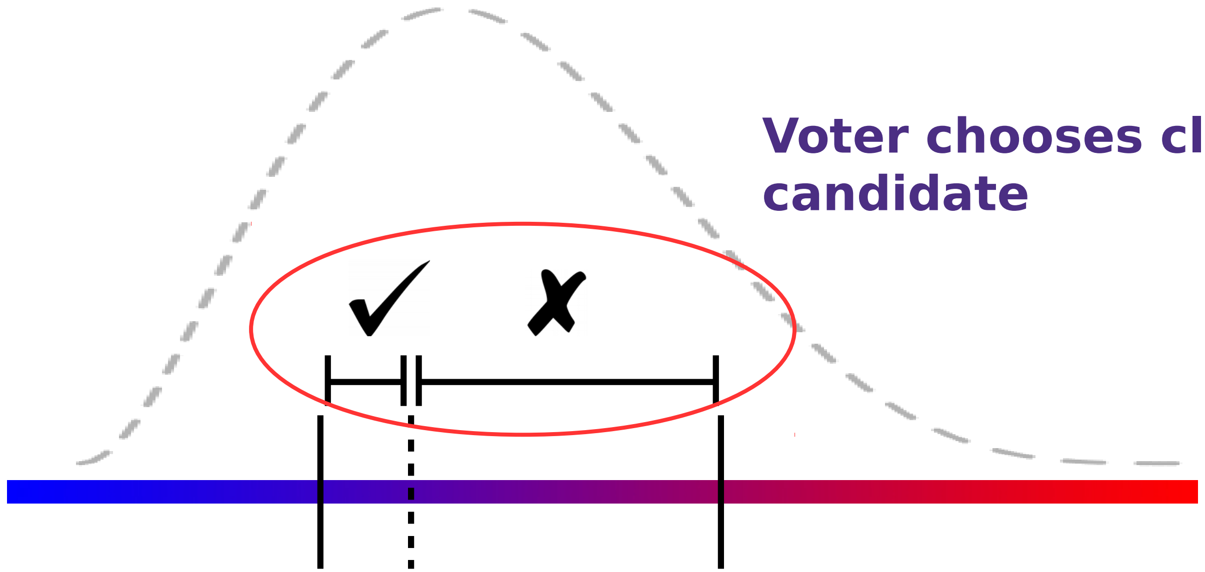
Dotted Line = Voter Position (Unknown)

(Downs, 1957)

Assumption 2: Spatial Voting

THE FIRST CHAPTER OF OUR “GENERATIVE STORY”

Voter chooses closest candidate



Political Position

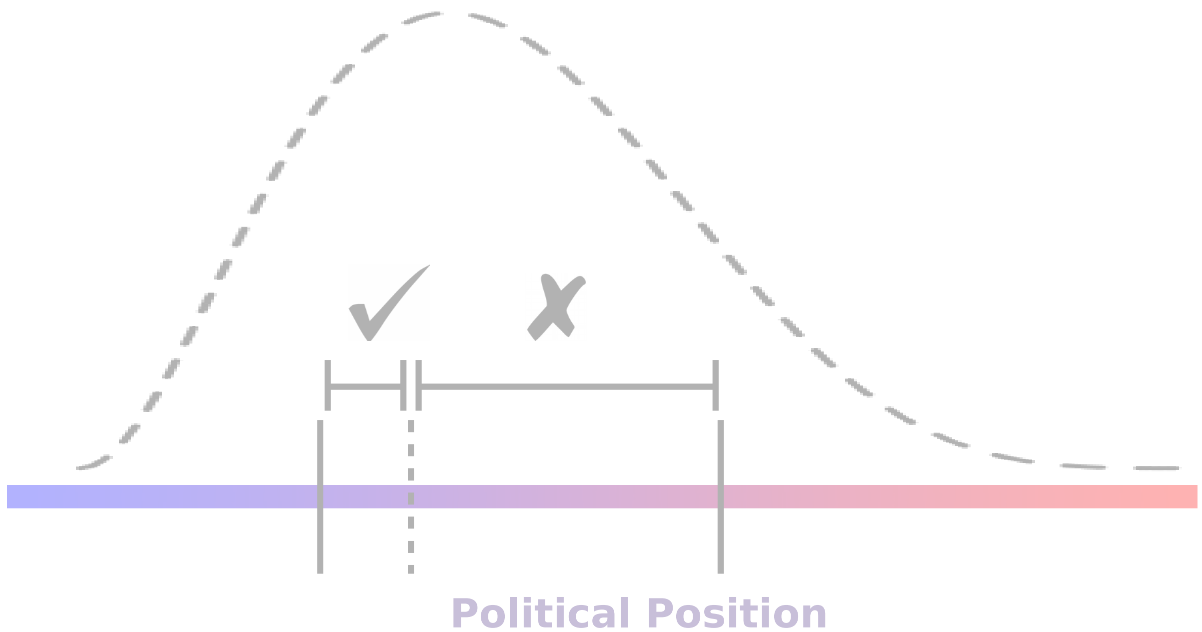
Solid Line = Candidate Position (Known)

Dotted Line = Voter Position (Unknown)

(Downs, 1957)

Assumption 2: Spatial Voting

THE FIRST CHAPTER OF OUR "GENERATIVE STORY"



$$\boldsymbol{y}_i \sim Normal(\mu_{x_i}, \sigma_{x_i}) \text{ \#Voter positions}$$

$$n_i = \sum_{j=t}^{N_i} \mathbb{1}(|y_{ji} - c_{0i}| < |y_{ji} - c_{1i}|) \text{ \#Votes cast}$$

(Downs, 1957)

Assumption 2: Spatial Voting

BAYESIAN INFERENCE

- **“Inverting” model allows us to infer distribution of voter positions**
- **What distribution of voter positions makes the observed votes most likely?**

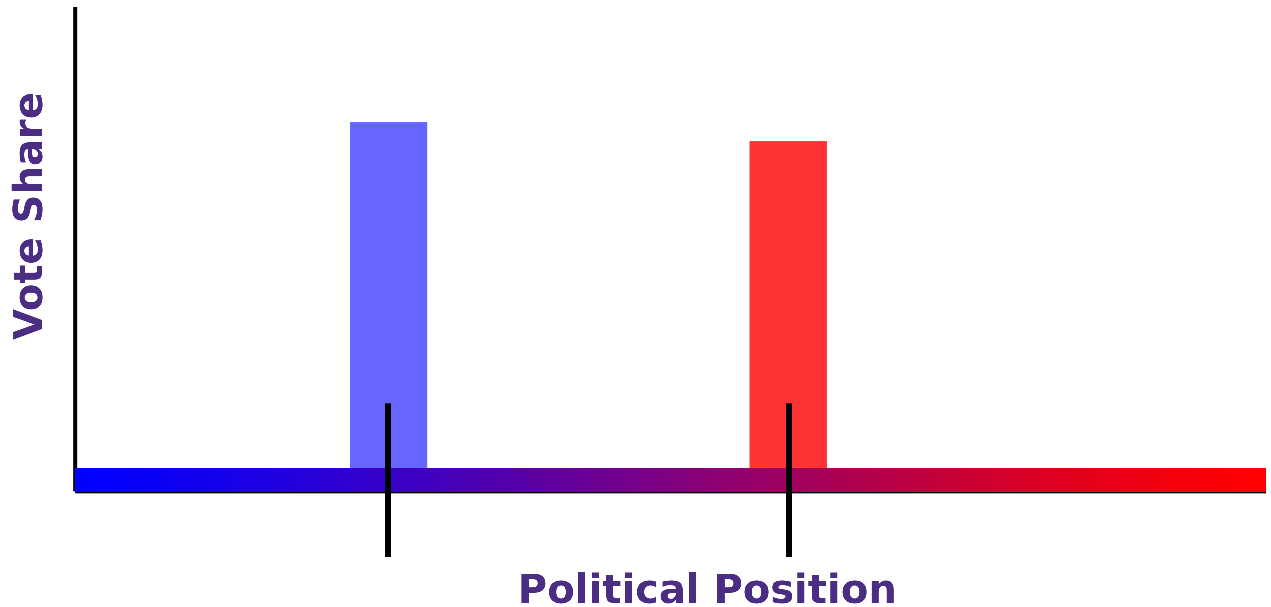
Assumption 2: Spatial Voting

BAYESIAN INFERENCE

- **Story doesn’t have to be literally true**
- **Vote by party versus position**
- **Other factors ignored**
- **“Revealed preferences” model**

Challenge 3: Sparse Data

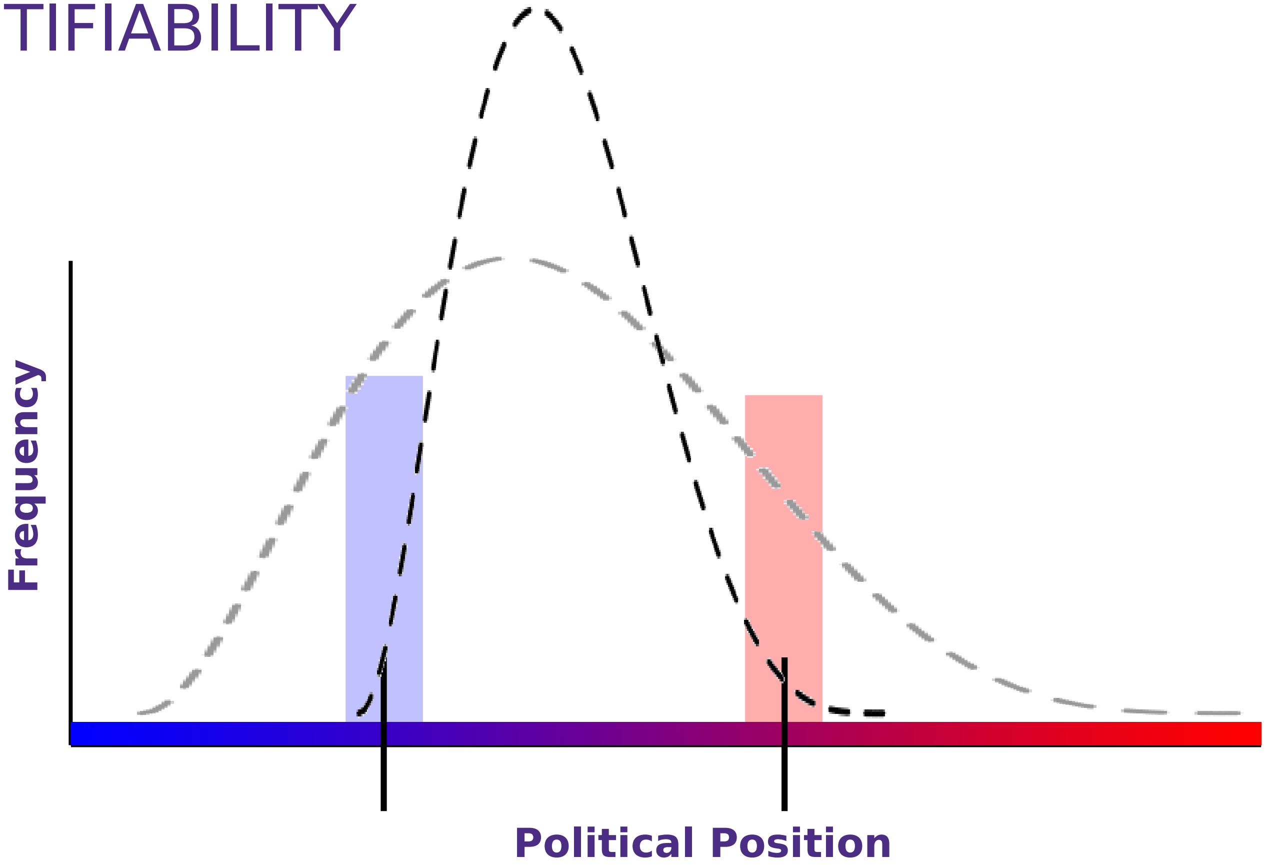
ONLY TWO DATA POINTS PER PRECINCT!



Solid Line = Candidate Position

Challenge 3: Sparse Data

IDENTIFIABILITY

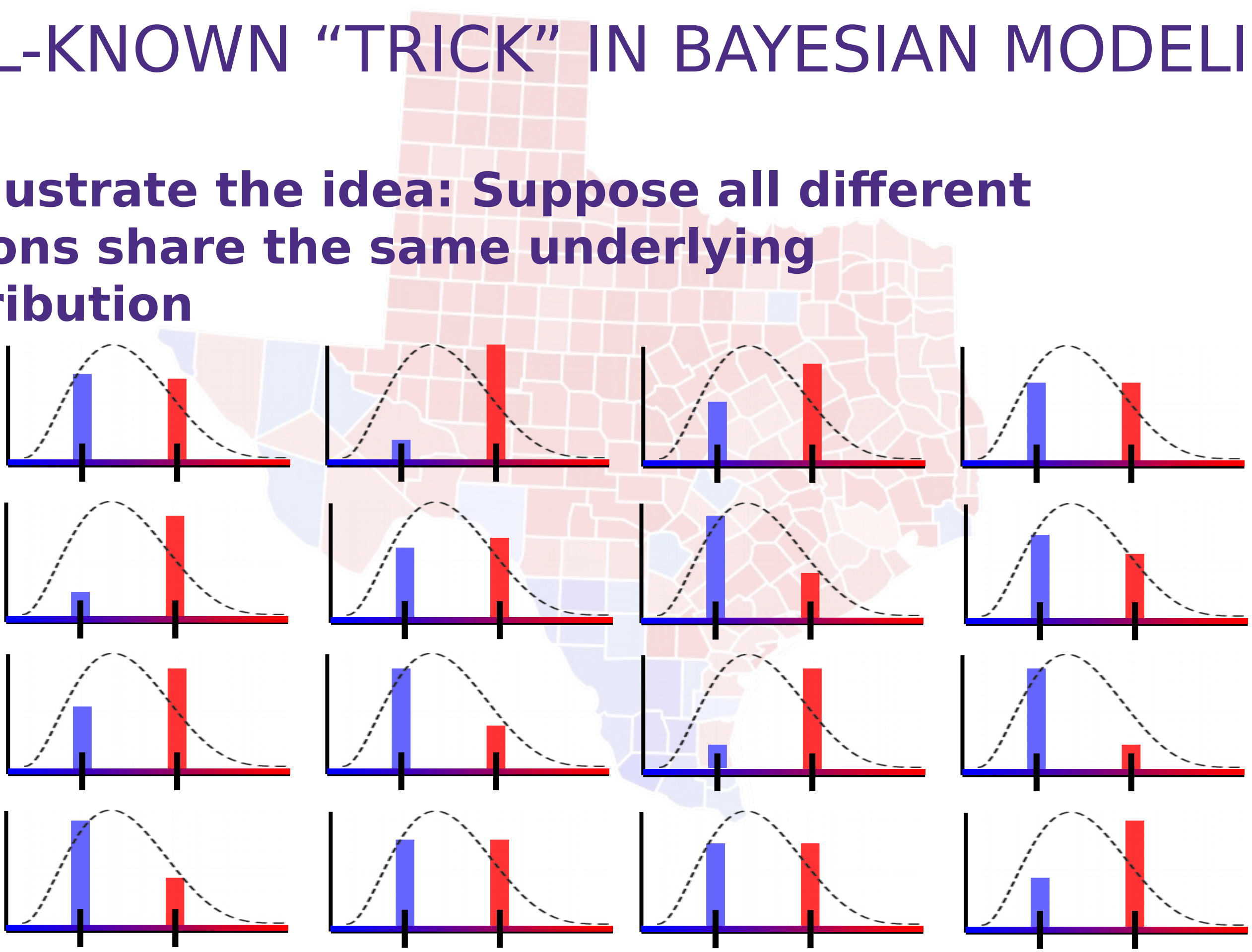


Dotted Curve = Voter Distribution (Unknown)
Solid Line = Candidate Position

Assumption 3: Statistical Pooling

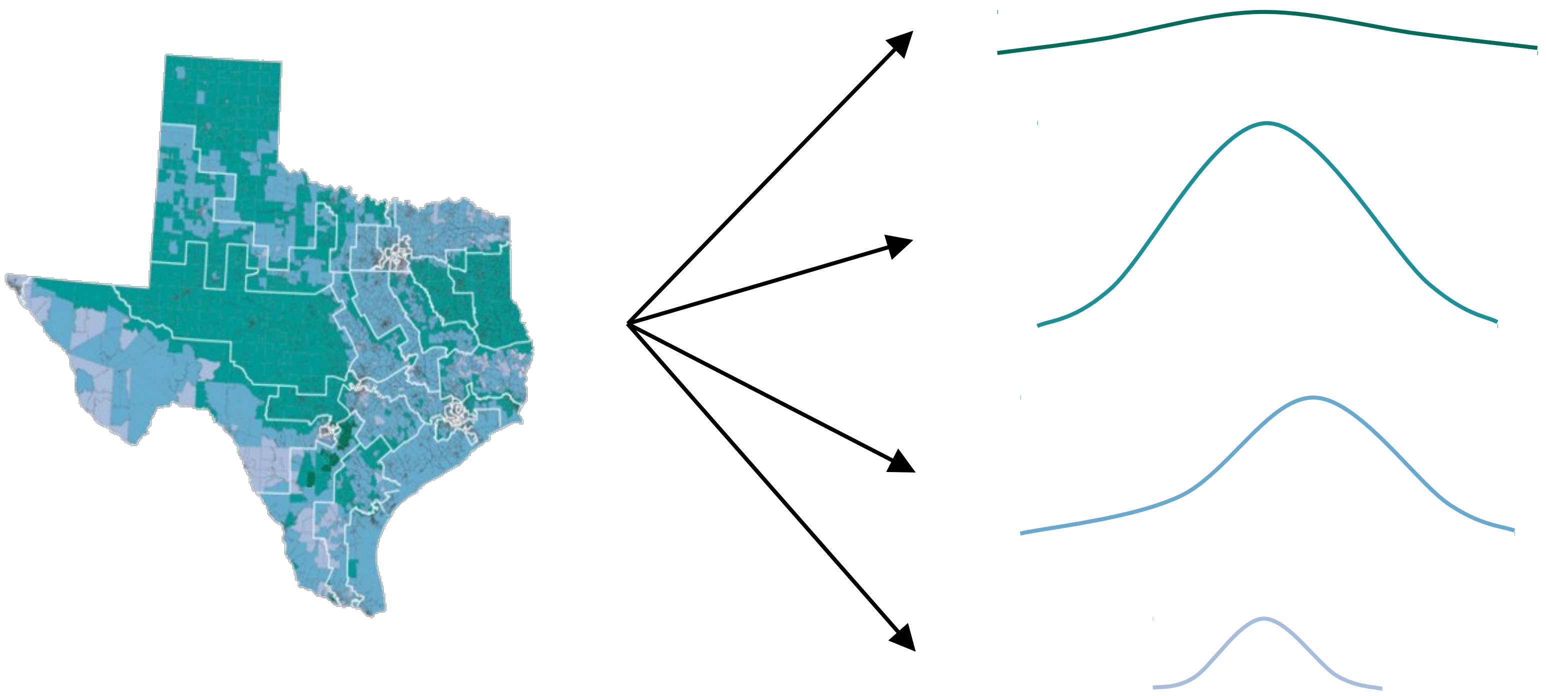
WELL-KNOWN "TRICK" IN BAYESIAN MODELING

To illustrate the idea: Suppose all different regions share the same underlying distribution



Assumption 3: Statistical Pooling

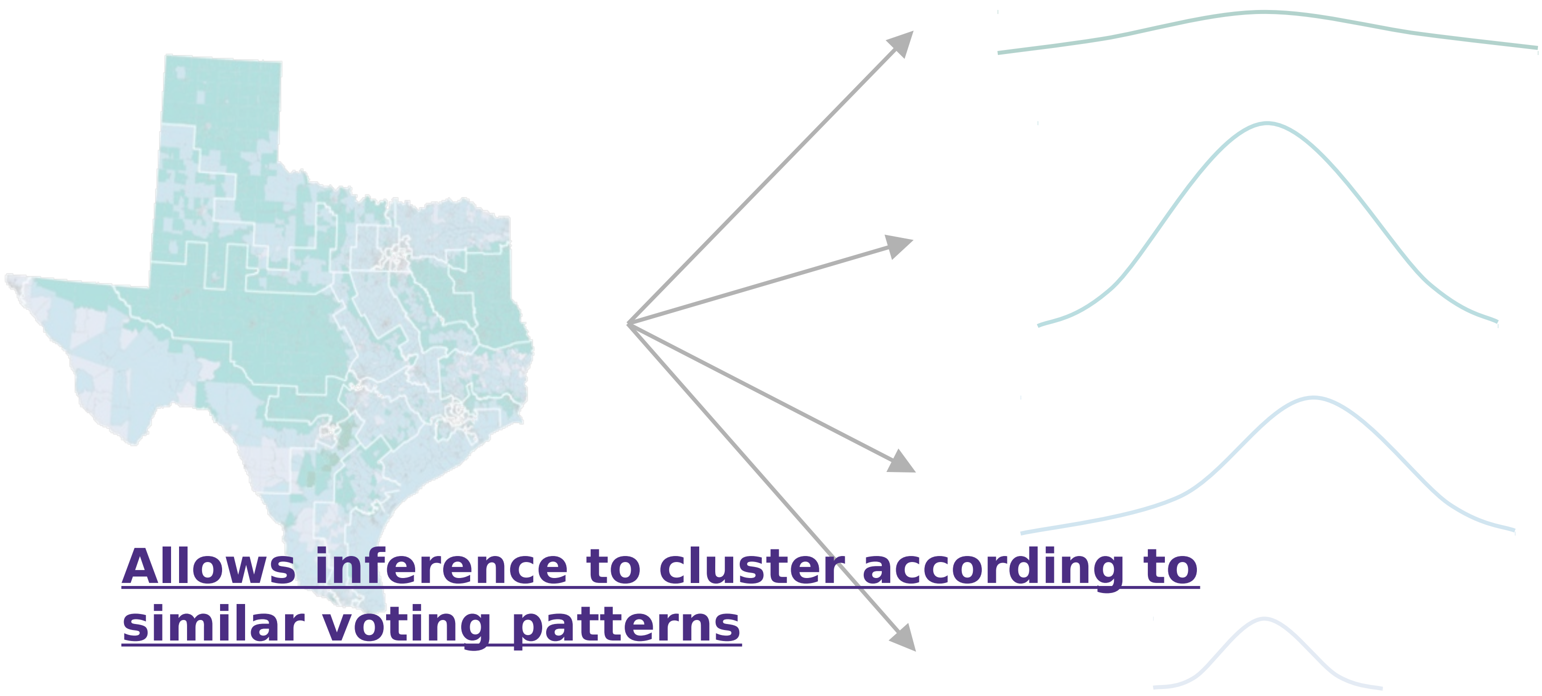
THE SECOND CHAPTER IN OUR "GENERATIVE STORY"



Suppose some regions share the same underlying distribution

Assumption 3: Statistical Pooling

THE SECOND CHAPTER IN OUR "GENERATIVE STORY"

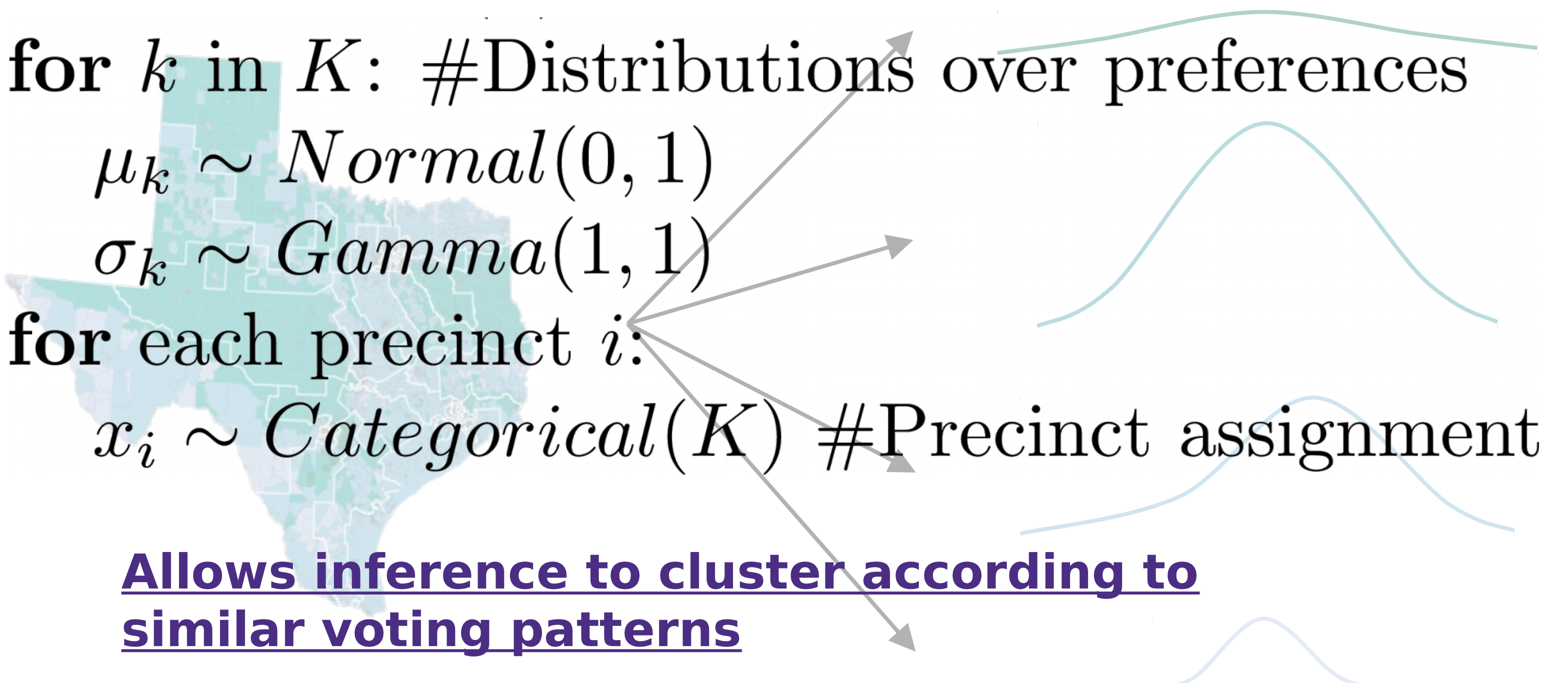


Allows inference to cluster according to similar voting patterns

Suppose some regions share the same underlying distribution

Assumption 3: Statistical Pooling

THE SECOND CHAPTER IN OUR "GENERATIVE STORY"



$$
\begin{aligned}
&\textbf{for } k \text{ in } K\text{: } \#\text{Distributions over preferences} \\
&\quad \mu_k \sim Normal(0, 1) \\
&\quad \sigma_k \sim Gamma(1, 1) \\
&\textbf{for } \text{each precinct } i\text{:} \\
&\quad x_i \sim Categorical(K) \ \#\text{Precinct assignment}
\end{aligned}
$$

Allows inference to cluster according to similar voting patterns

Suppose some regions share the same underlying distribution

Mixture of Spatial Voting Models

BAYESIAN INFERENCE

What region’s voting patterns are best explained by the same distributions of political positions?

What distributions are needed to best explain observed votes?

Mixture of Spatial Voting Models

Given $\boldsymbol{c}$ #Candidate positions
Given K #Number of clusters
$\theta \sim Dirichlet(\mathbf{1})$
for k in K: #Distributions over preferences
$\quad \mu_k \sim Normal(0, 1)$
$\quad \sigma_k \sim Gamma(1, 1)$
for each precinct i:
$\quad x_i \sim Categorical(K)$ #Precinct assignment
$\quad \boldsymbol{y}_i \sim Normal(\mu_{x_i}, \sigma_{x_i})$ #Voter positions
$\quad n_i = \sum_{j=t}^{N_i} \mathbb{1}(|y_{ji} - c_{0i}| < |y_{ji} - c_{1i}|)$ #Votes cast

Mixture of Spatial Voting Models

Given $\boldsymbol{c}$ #Candidate positions
Given K #Number of clusters
$\theta \sim Dirichlet(\mathbf{1})$
for k in K: #Distributions over preferences
$\quad \mu_k \sim Normal(0, 1)$
$\quad \sigma_k \sim Gamma(1, 1)$
for each precinct i:
$\quad x_i \sim Categorical(K)$ #Precinct assignment
$\quad \boldsymbol{y}_i \sim Normal(\mu_{x_i}, \sigma_{x_i})$ #Voter positions
$\quad n_i = \sum_{j=t}^{N_i} \mathbb{1}(|y_{ji} - c_{0i}| < |y_{ji} - c_{1i}|)$ #Votes cast

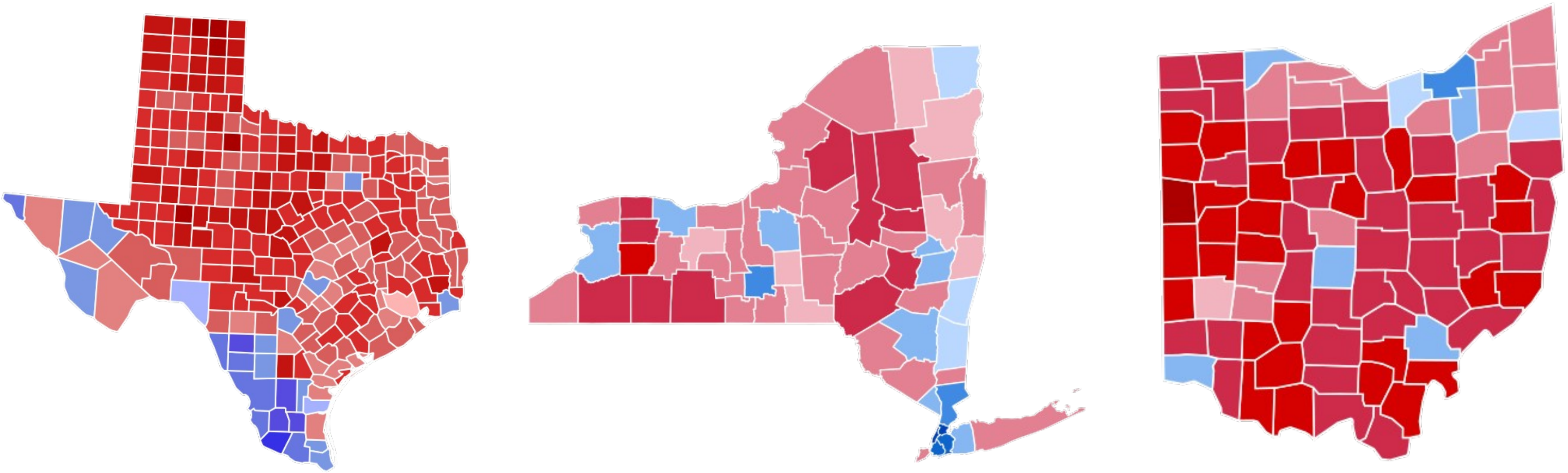
Observe!

Mixture of Spatial Voting Models

Given $\boldsymbol{c}$ #Candidate positions
Given K #Number of clusters
$\theta \sim Dirichlet(\mathbf{1})$
for k in K: #Distributions over preferences
$\quad \mu_k \sim Normal(0, 1)$
$\quad \sigma_k \sim Gamma(1, 1)$
for each precinct i:
$\quad x_i \sim Categorical(K)$ #Precinct assignment
$\quad \boldsymbol{y}_i \sim Normal(\mu_{x_i}, \sigma_{x_i})$ #Voter positions
$\quad n_i = \sum_{j=t}^{N_i} \mathbb{1}(|y_{ji} - c_{0i}| < |y_{ji} - c_{1i}|)$ #Votes cast

Infer!

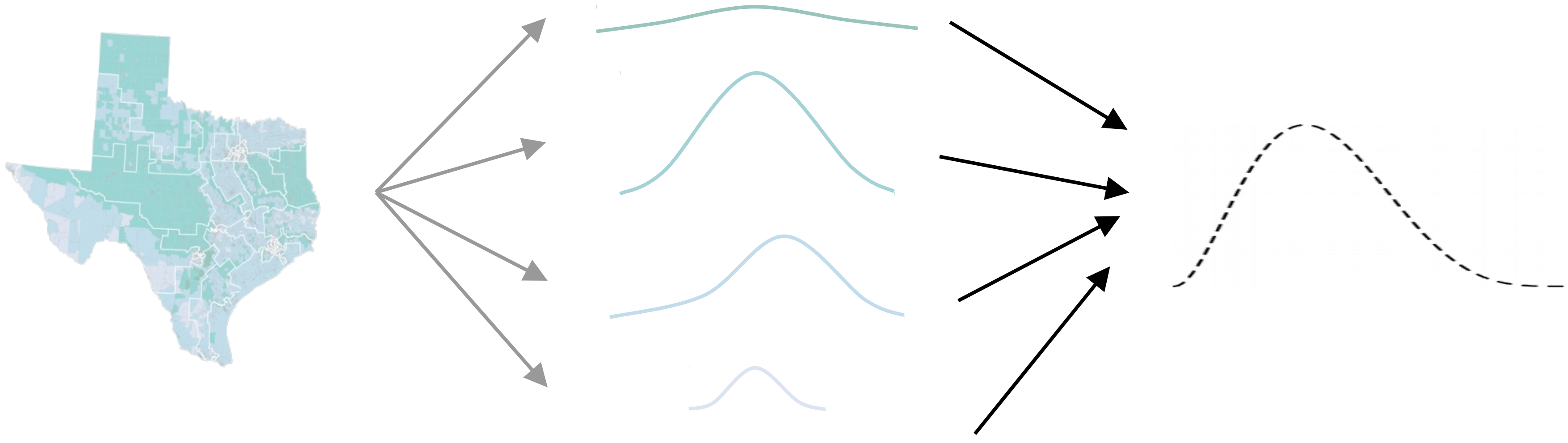
Data



2006, 2008, 2010

Aggregation Method

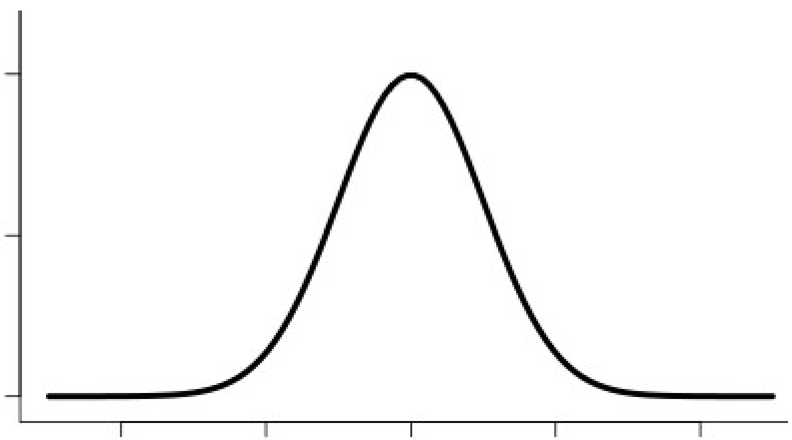
COMBINE PRECINCTS TO STATE / COUNTY LEVEL



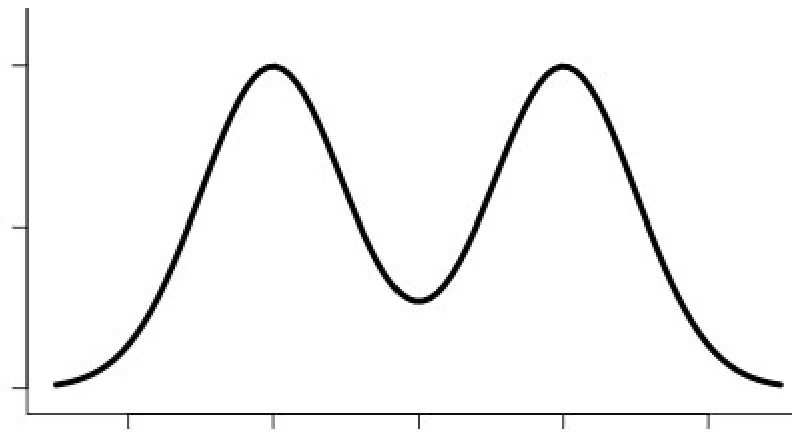
Mass Polarization

DIFFERENT FORMS OF POLARIZATION

Bimodality



Political Position



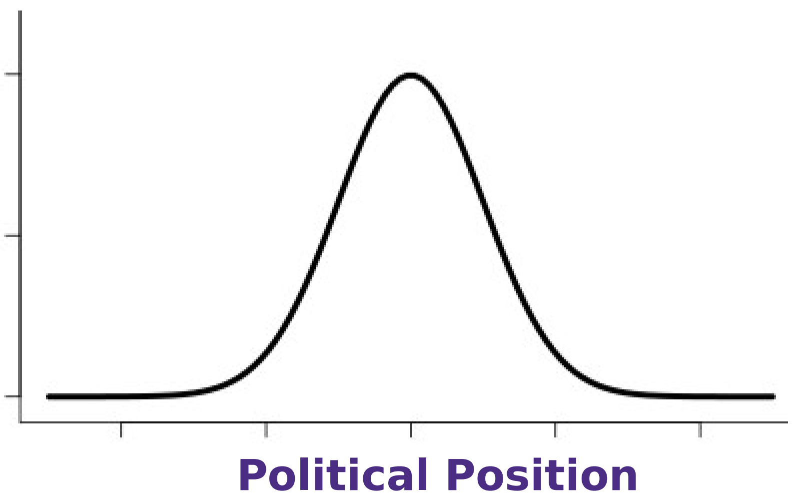
Political Position

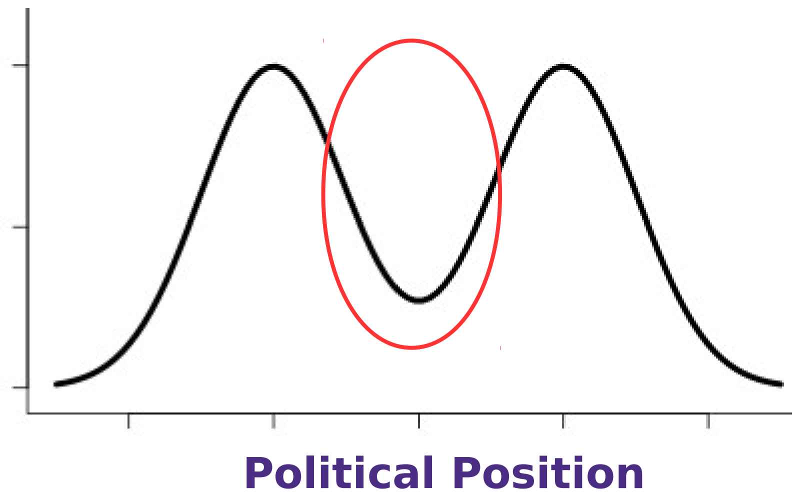
(DiMaggio et al., 1996)

Mass Polarization

DIFFERENT FORMS OF POLARIZATION

Bimodality



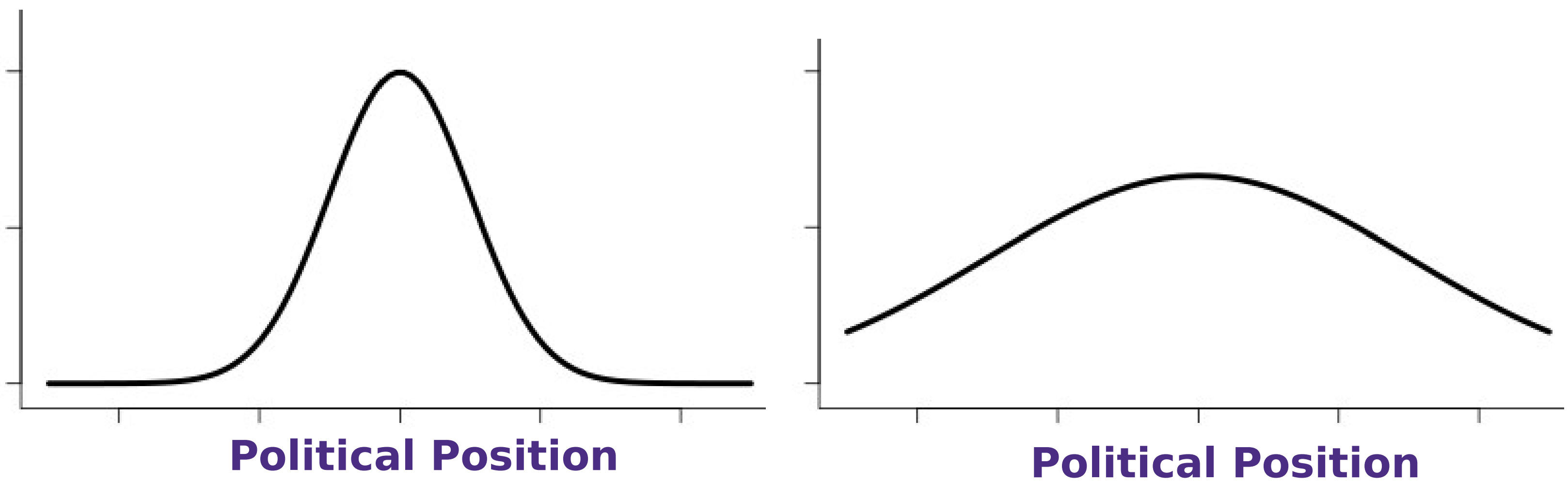


(DiMaggio et al., 1996)

Mass Polarization

DIFFERENT FORMS OF POLARIZATION

Dispersion

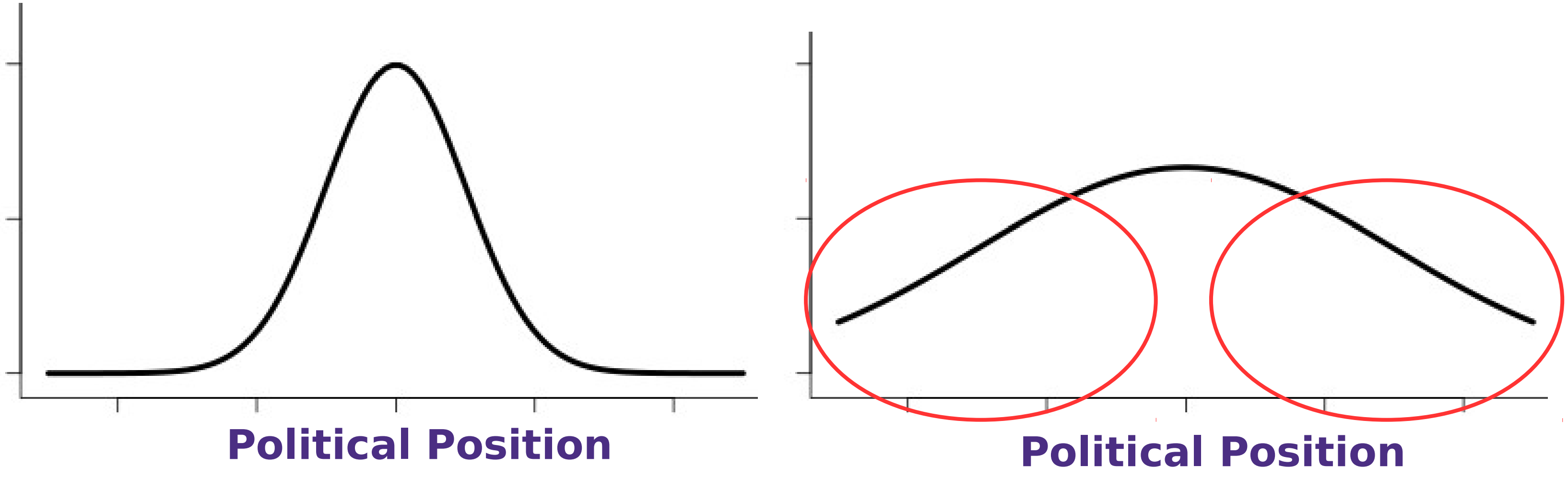


(DiMaggio et al., 1996)

Mass Polarization

DIFFERENT FORMS OF POLARIZATION

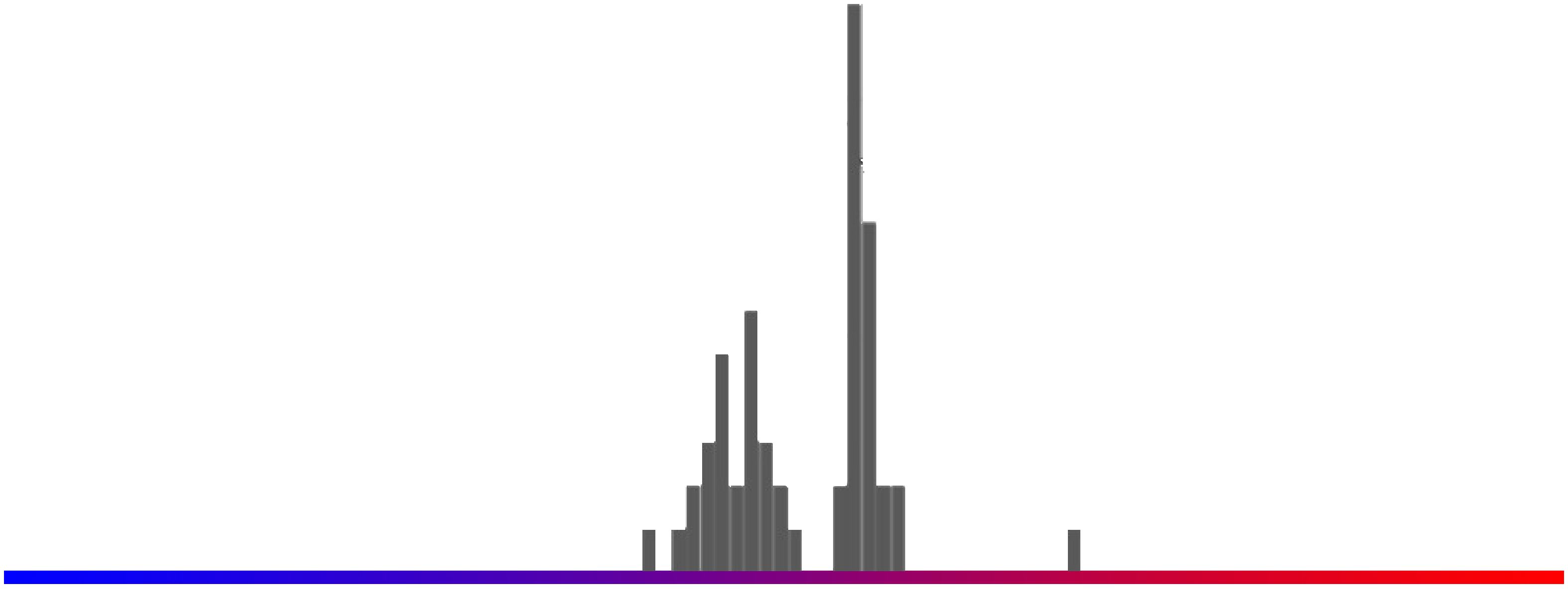
Dispersion



(DiMaggio et al., 1996)

Results

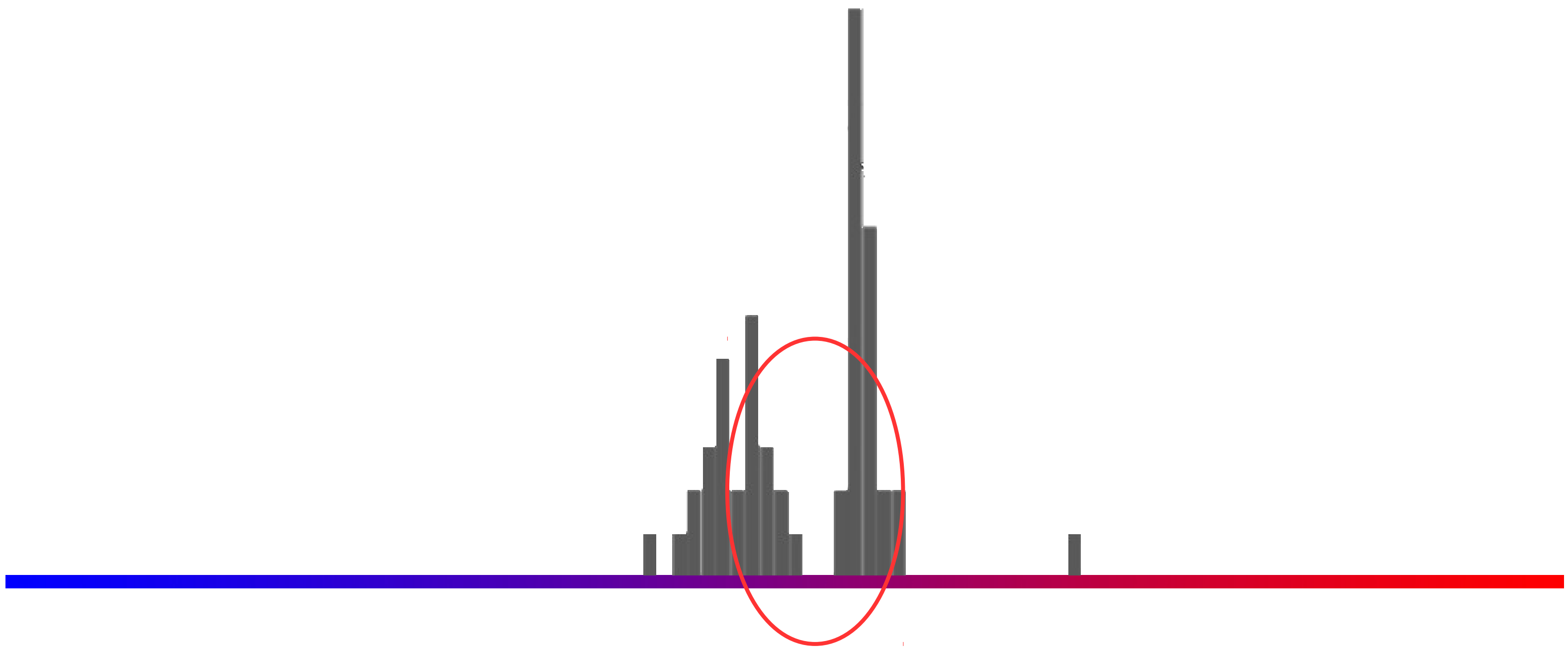
TEXAS 2008



Histogram = Candidates' Political Positions (Known from CF-Scores)

Results

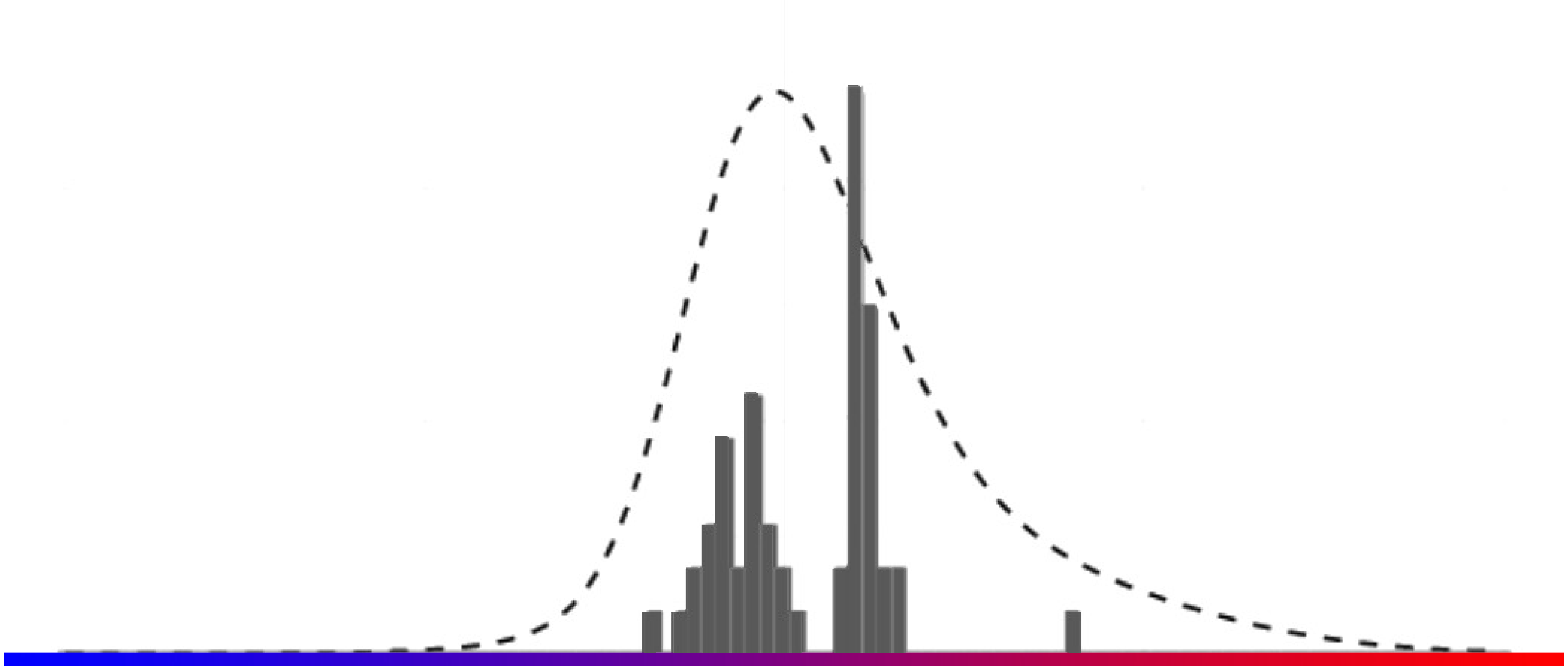
TEXAS 2008



Histogram = Candidates' Political Positions (Known from CF-Scores)

Results

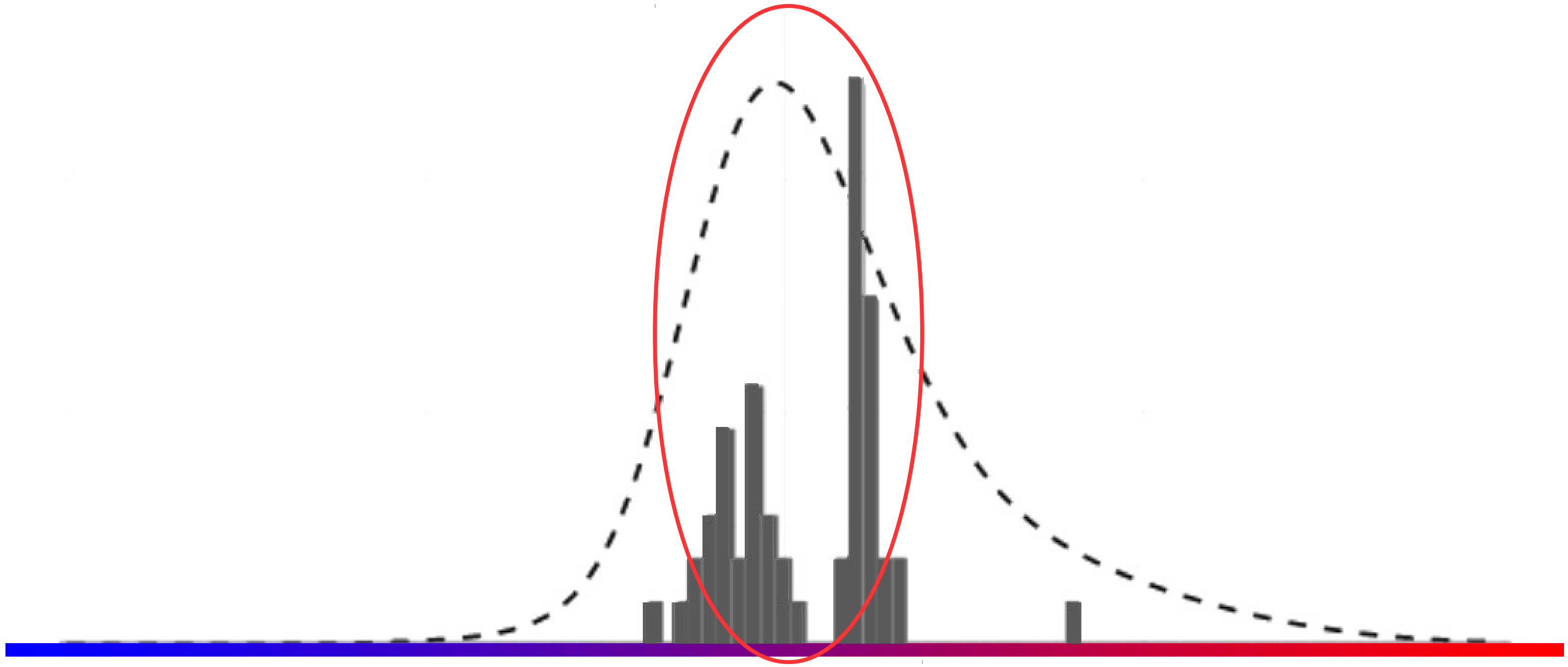
TEXAS 2008



Histogram = Candidates' Political Positions
Dotten Curve = <u>Inferred</u> Distribution of Voter Positions

Results

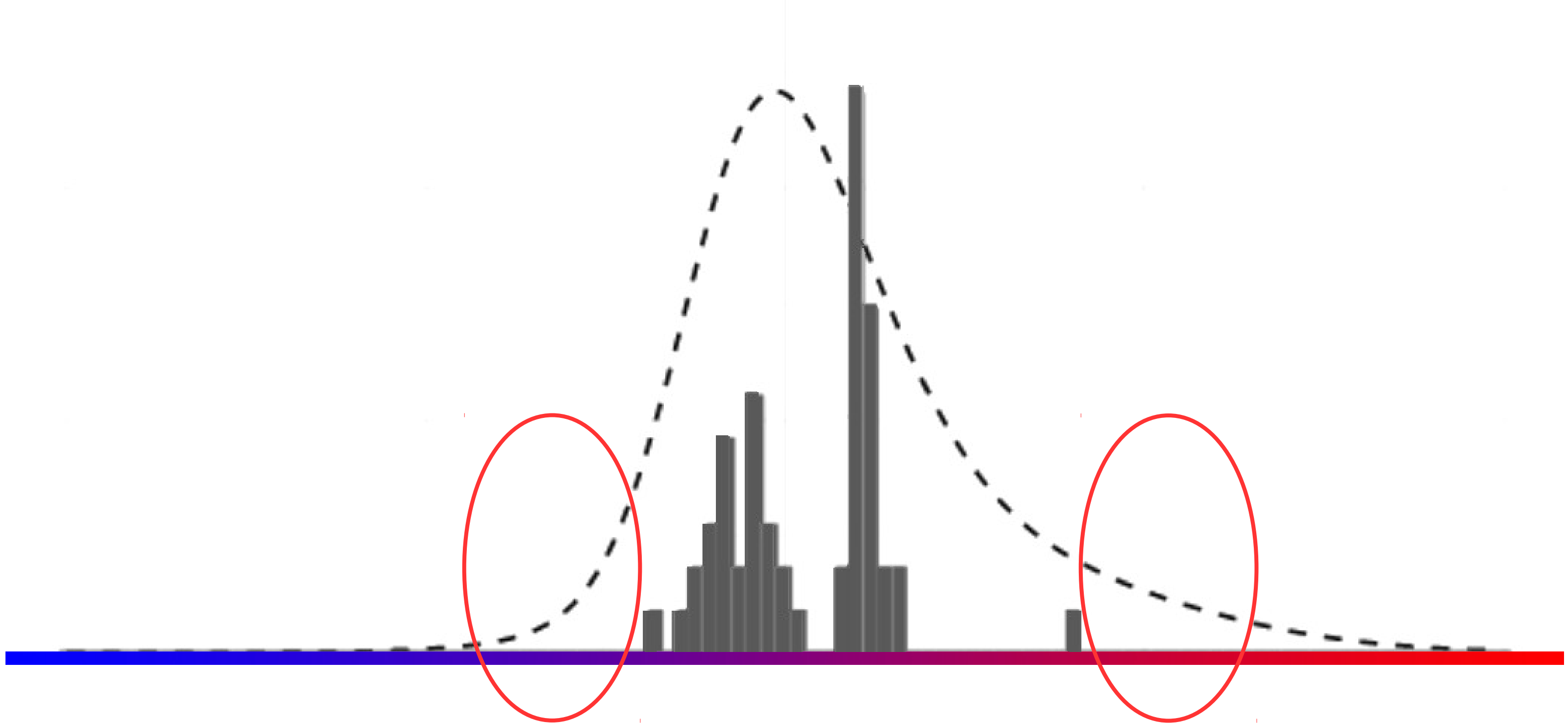
TEXAS 2008



Histogram = Candidates' Political Positions
Dotten Curve = <u>Inferred</u> Distribution of Voter Positions

Results

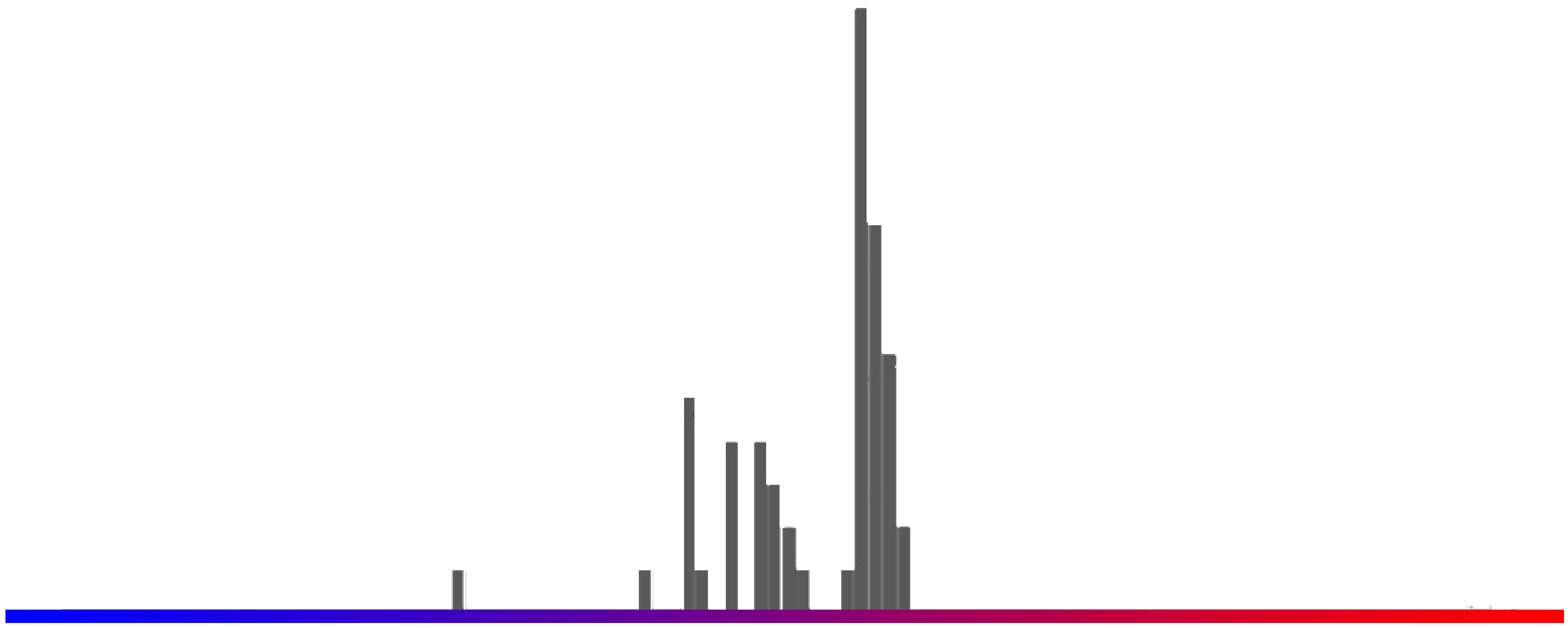
TEXAS 2008



Histogram = Candidates' Political Positions
Dotten Curve = <u>Inferred</u> Distribution of Voter Positions

Results

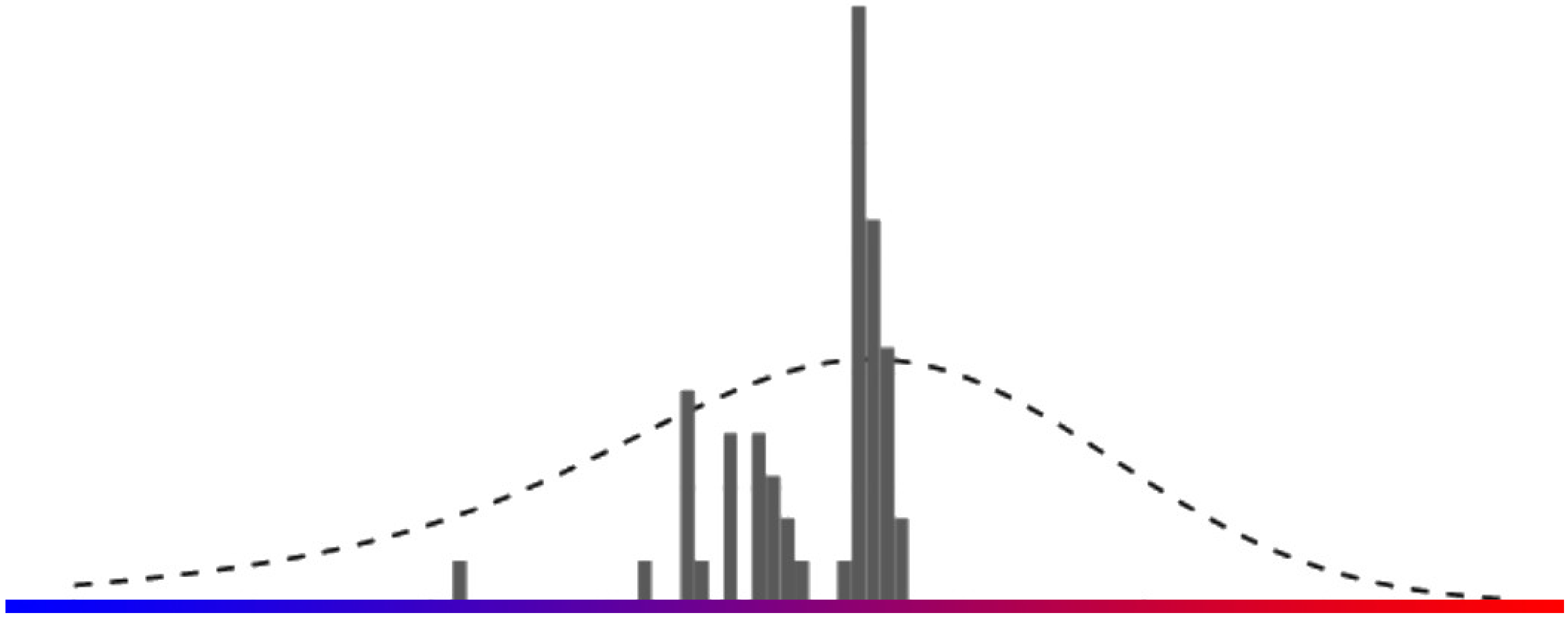
TEXAS 2010



Histogram = Candidates' Political Positions

Results

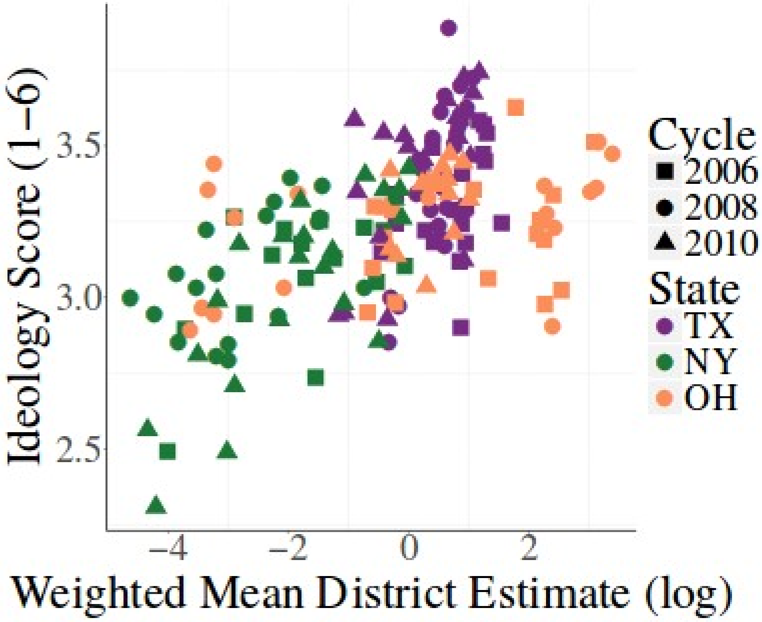
TEXAS 2010



Histogram = Candidates' Political Positions
Dotten Curve = <u>Inferred</u> Distribution of Voter Positions

Construct Validity

OUR ESTIMATES CORRELATE WITH EXISTING METRICS



CCES Data

(Also checked against several other metrics)

International Conference on Social Informatics

SocInfo 2016: Social Informatics pp 290-311 | Cite as

Inferring Population Preferences via Mixtures of Spatial Voting Models

Authors Authors and affiliations

Alison Nahm, Alex Pentland, Peter Krafft

We set out to use voting data to directly measure mass polarization

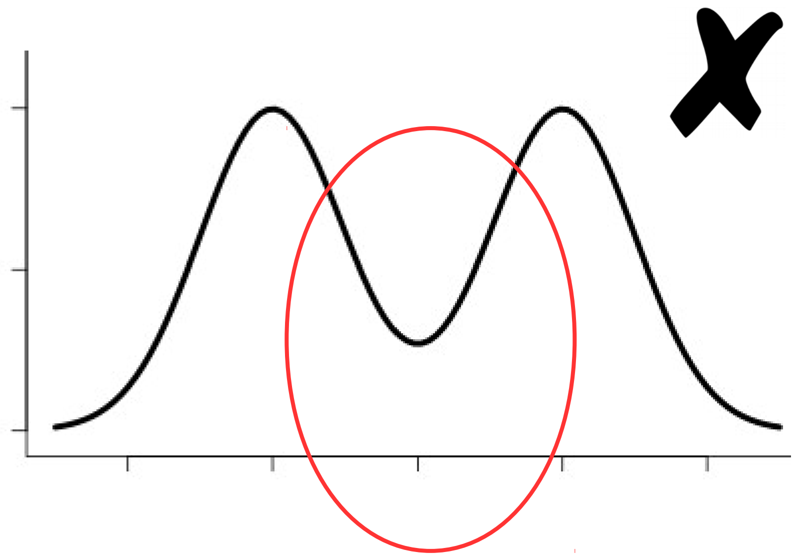
International Conference on Social Informatics

SocInfo 2016: Social Informatics pp 290-311 | Cite as

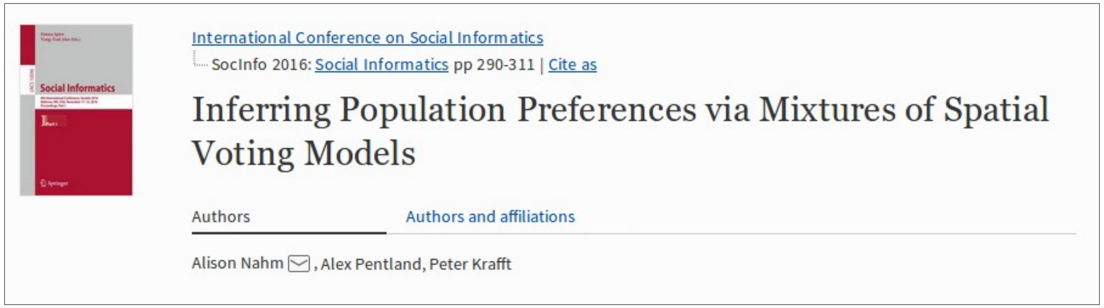
Inferring Population Preferences via Mixtures of Spatial Voting Models

Authors   Authors and affiliations

Alison Nahm, Alex Pentland, Peter Krafft



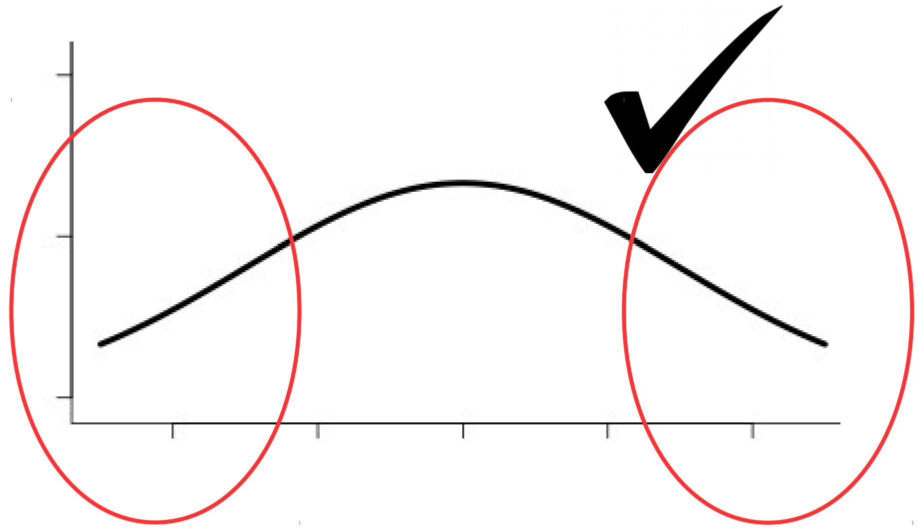
We find reliably <u>lower levels of bimodality</u> than in distributions of candidate positions


International Conference on Social Informatics
SocInfo 2016: Social Informatics pp 290-311 | Cite as

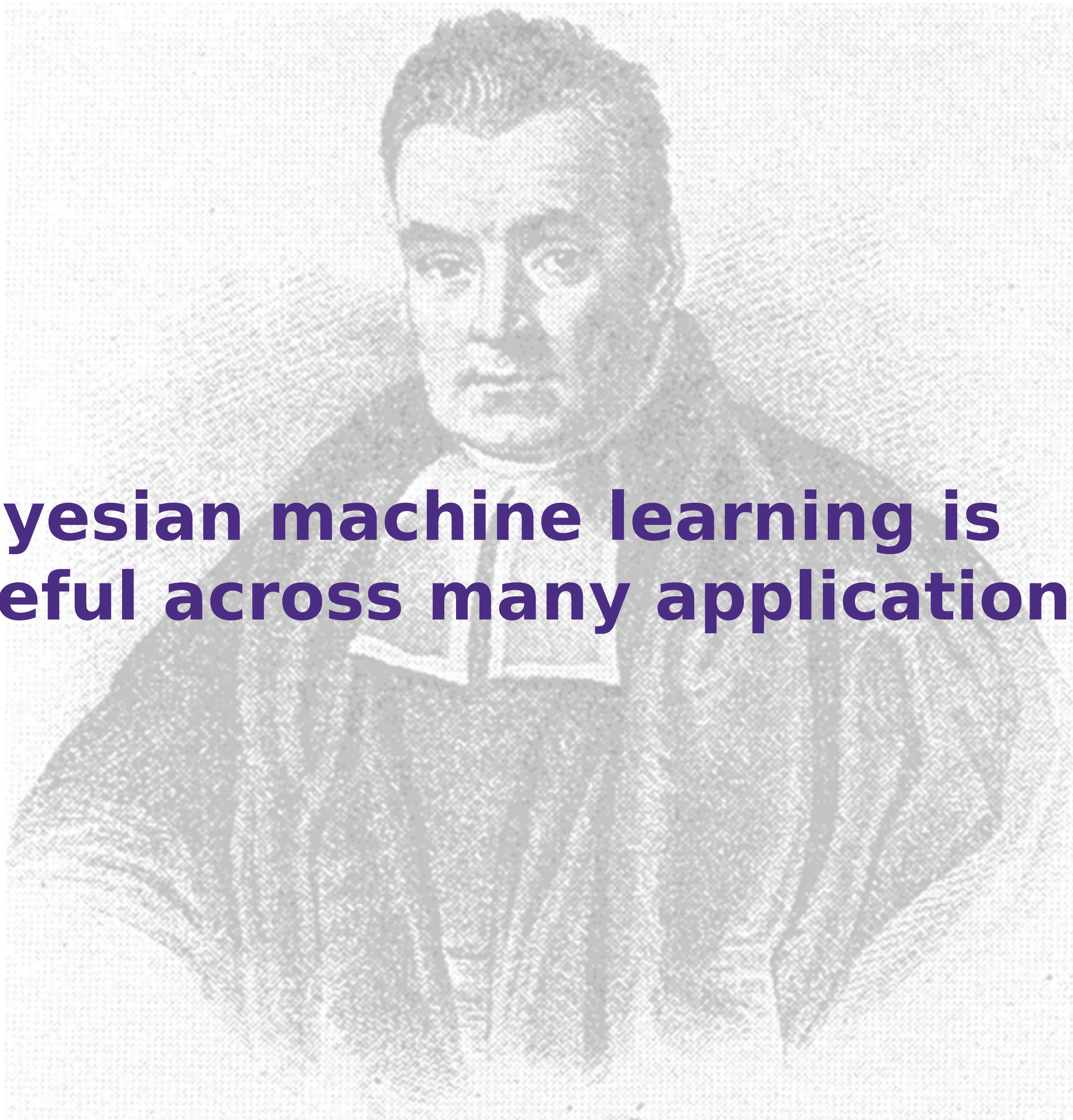
Inferring Population Preferences via Mixtures of Spatial Voting Models

Authors | Authors and affiliations

Alison Nahm, Alex Pentland, Peter Krafft



BUT reliably <u>higher levels of dispersion</u>



Bayesian machine learning is useful across many applications

This Talk

BAYESIAN MACHINE LEARNING FOR SOCIAL DATA SCIENCE

1) Overview of my Work

2) Intro to Polarization Model

3) Brief Bayesian Inference Tutorial

4) Polarization Model

5) Ongoing and Future Wor

Methodology

Epistemology

My Work in the Broader Web of OII

Popularity Dynamics

Media Manipulation

Social Learning

Rumors

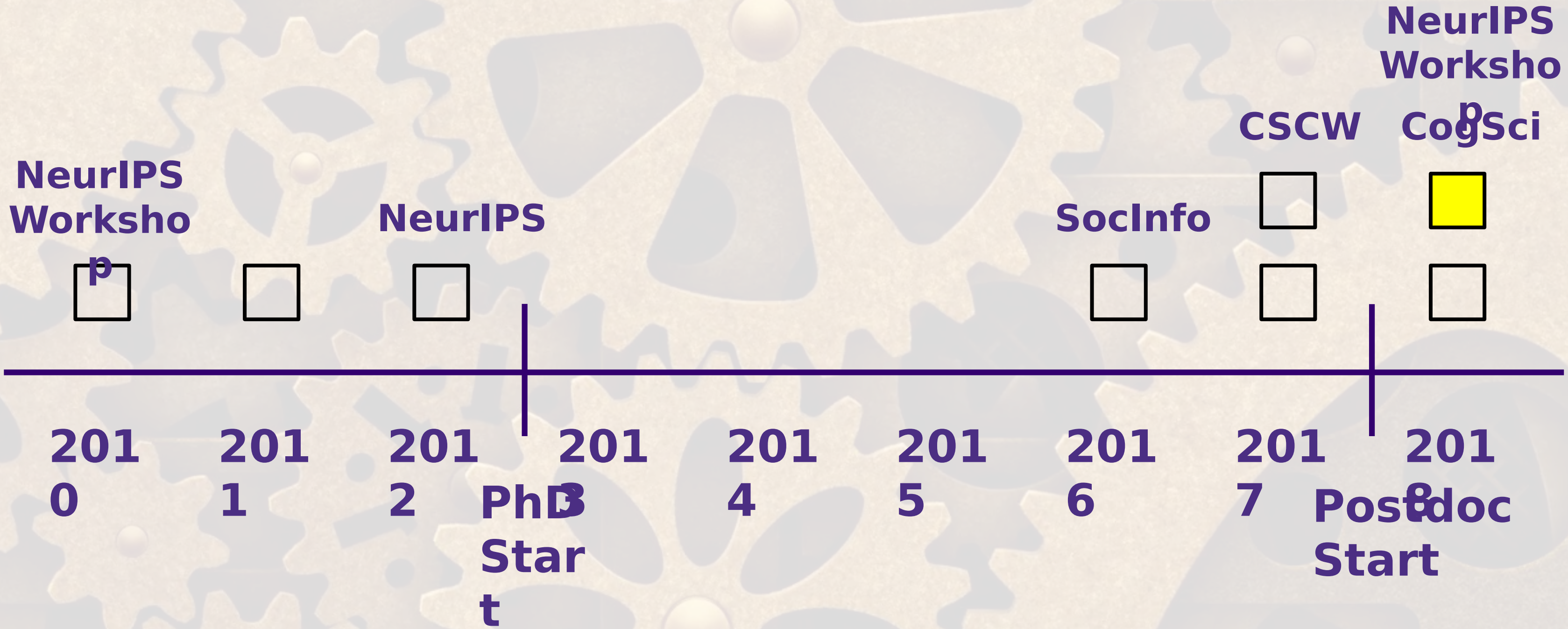
Digital Experiments

Alt-Right

Digital Ethnography

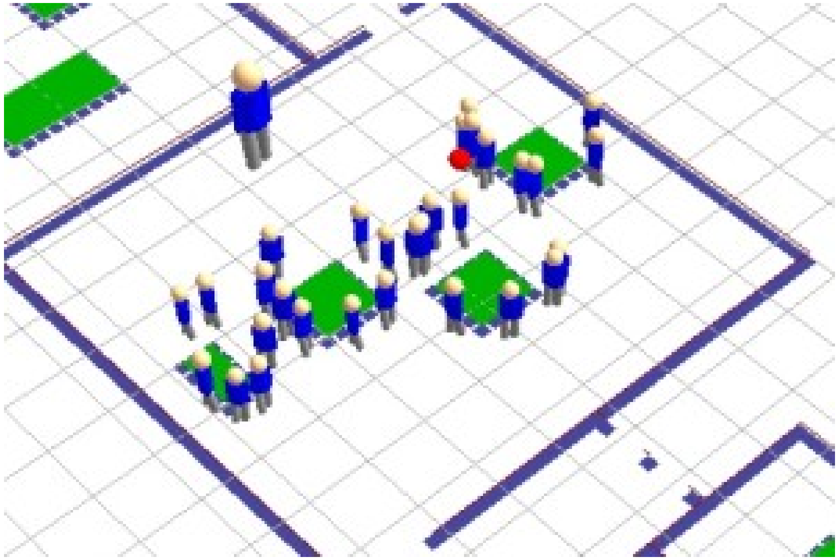
Cryptocoins

Machine Learning

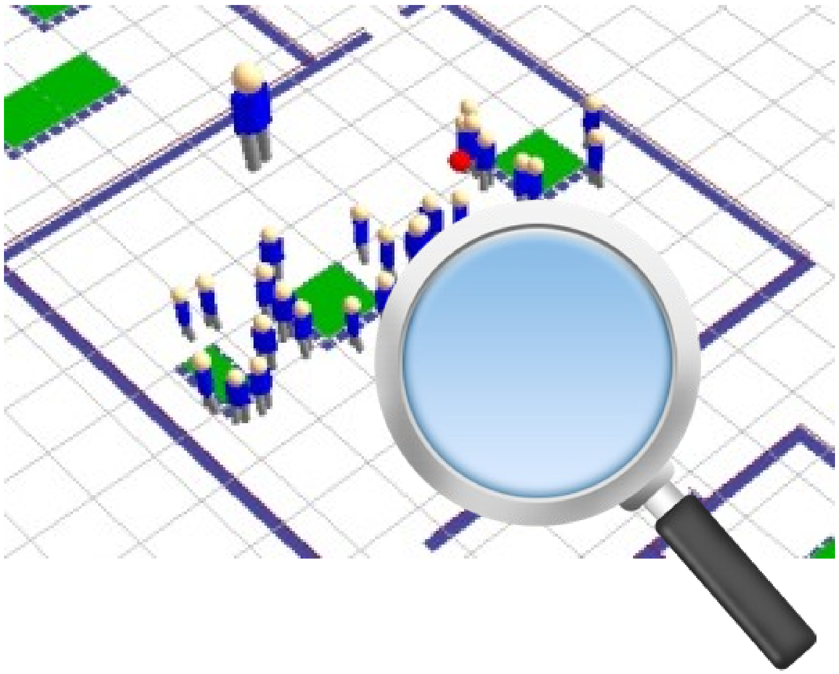


Methodology

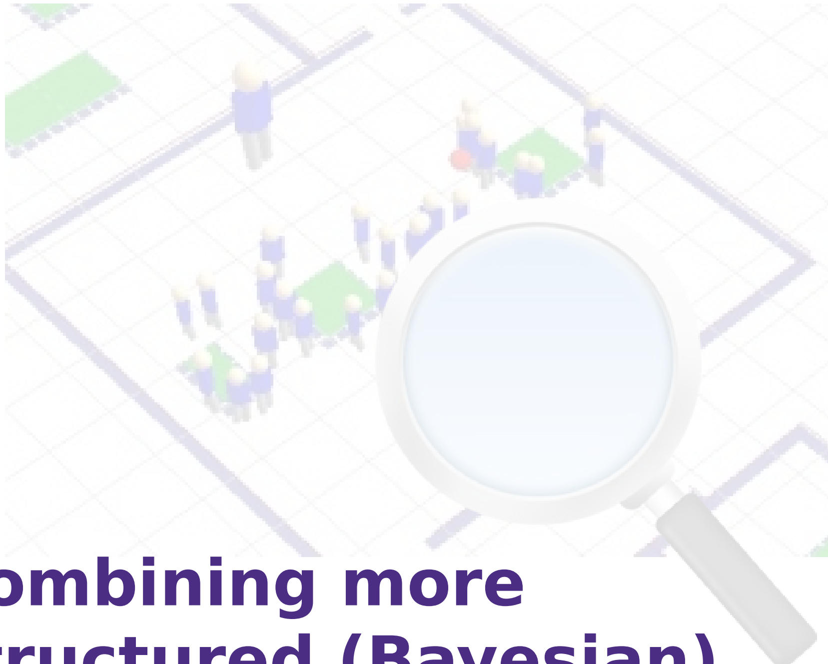
Epistemology

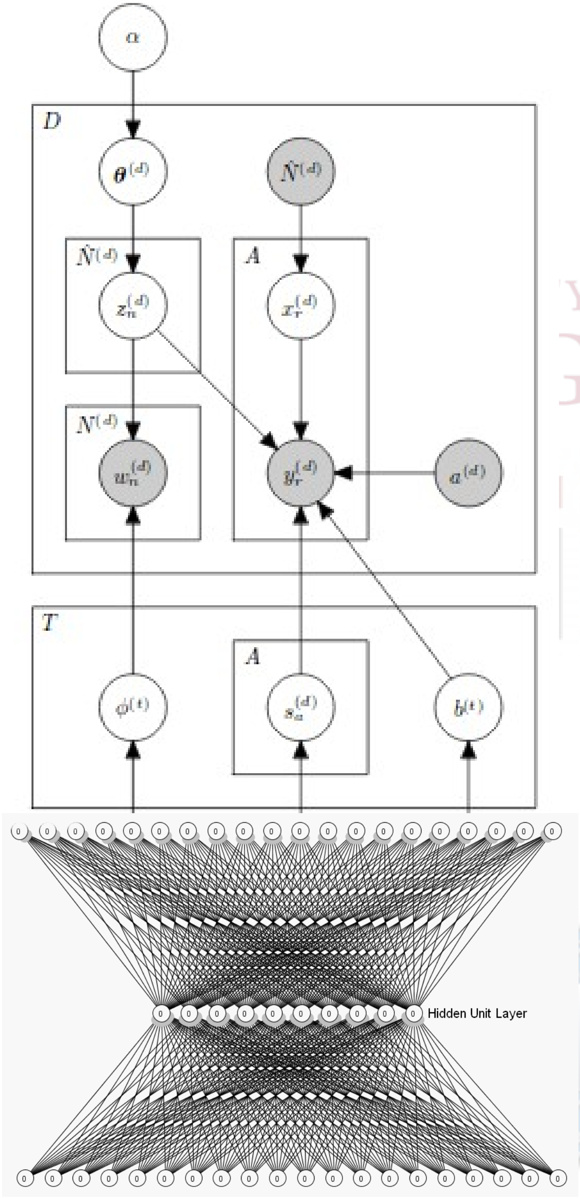


Senior Personnel
$2,000,000 2-year grant



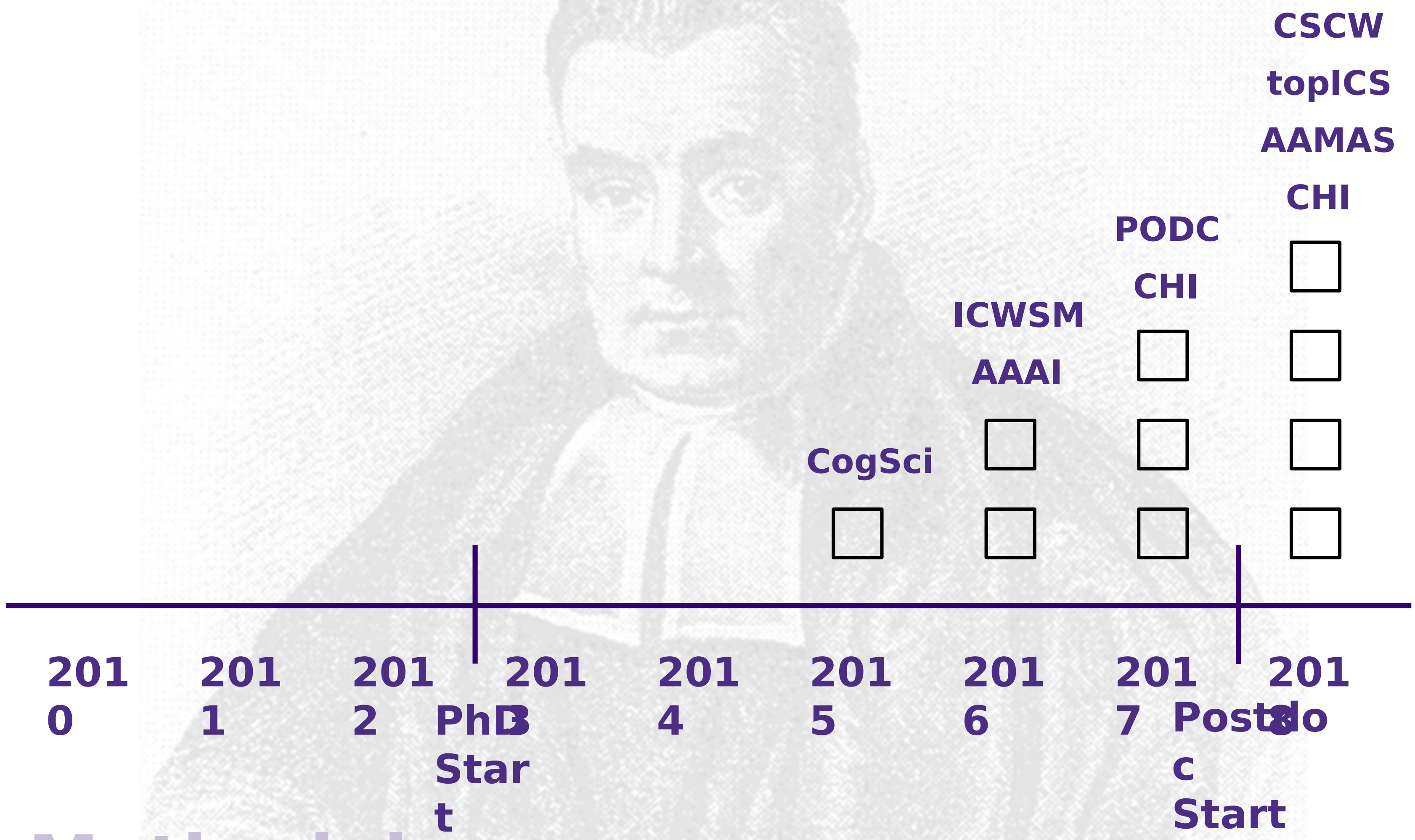
Senior Personnel $2,000,000 2-year grant





Combining more structured (Bayesian) and less structured (Neural) models

Senior Personnel
$2,000,000 2-year grant



Methodology

Epistemology

Collective Intelligence

Fads and Rumors

CSCW
topICS
AAMAS
CHI

PODC
CHI

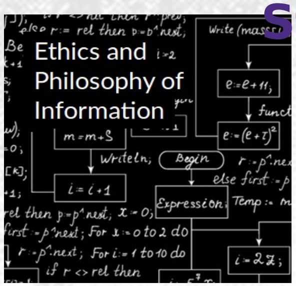
ICWSM
AAAI

CogSci

2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018

PhD Start

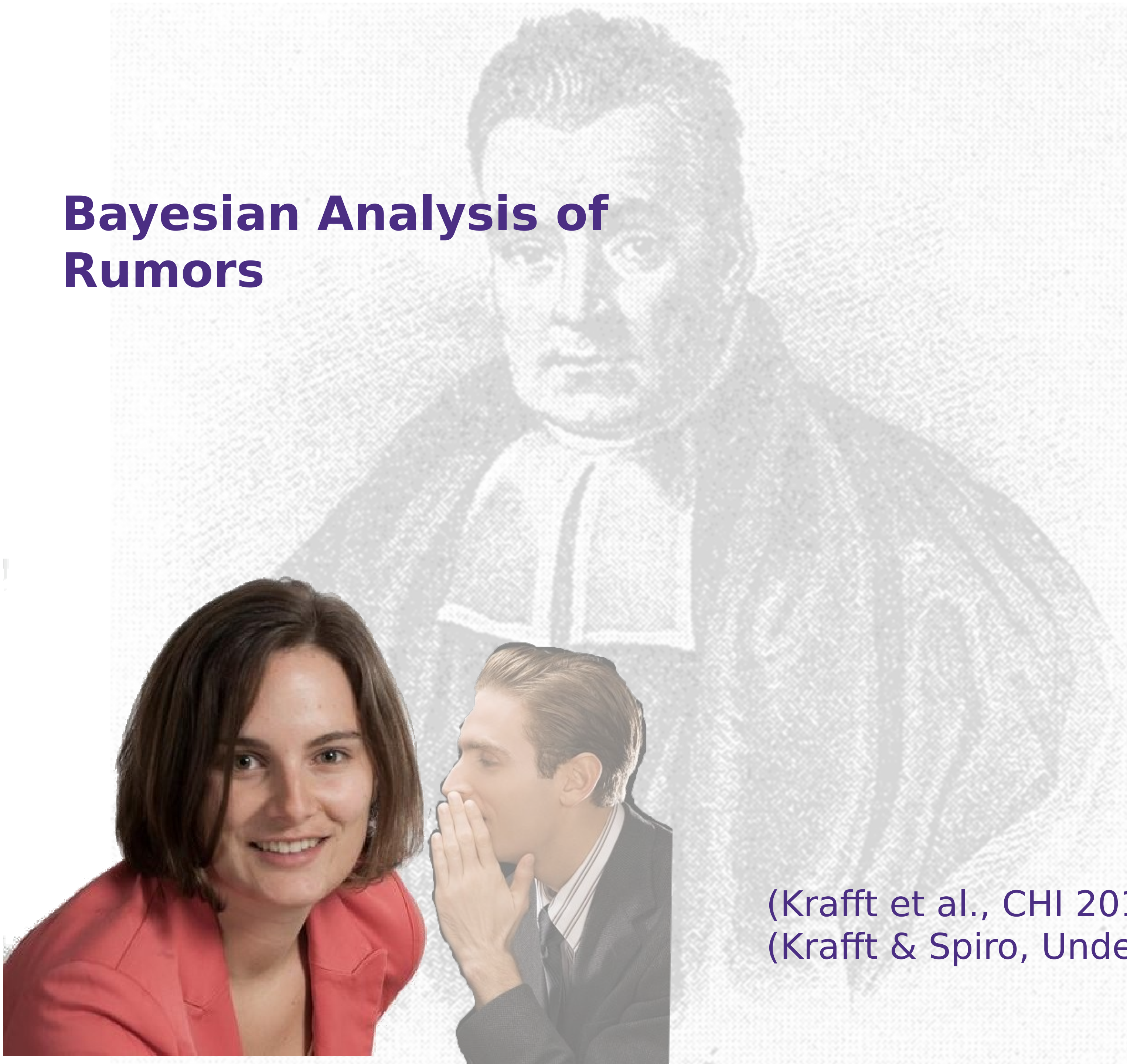
Postdoc Start



Methodology

Epistemology

Bayesian Analysis of Rumors



(Krafft et al., CHI 2017)
(Krafft & Spiro, Under Review)

Bayesian Analysis of Rumors

Centre for Experimental Social Sciences
Nuffield

A bombing at the Boston Marathon has occurred.

Please contribute to the ongoing discussion about the event by typing a short message (140 characters or fewer):

Submit

Relieved to hear that the girl and her family that I know that was running at the Boston Marathon is safe. #PrayersForBoston

Both girls I knew running the marathon are safe & sound!! God is good!

An eight year old girl who was doing an amazing thing running a marathon, was killed. I cant stand our world anymore

Work that touches on both

Methodology

Epistemology

Consider the following painting:

Is it a better example of Photorealism (A) or Realism (B)?

A B

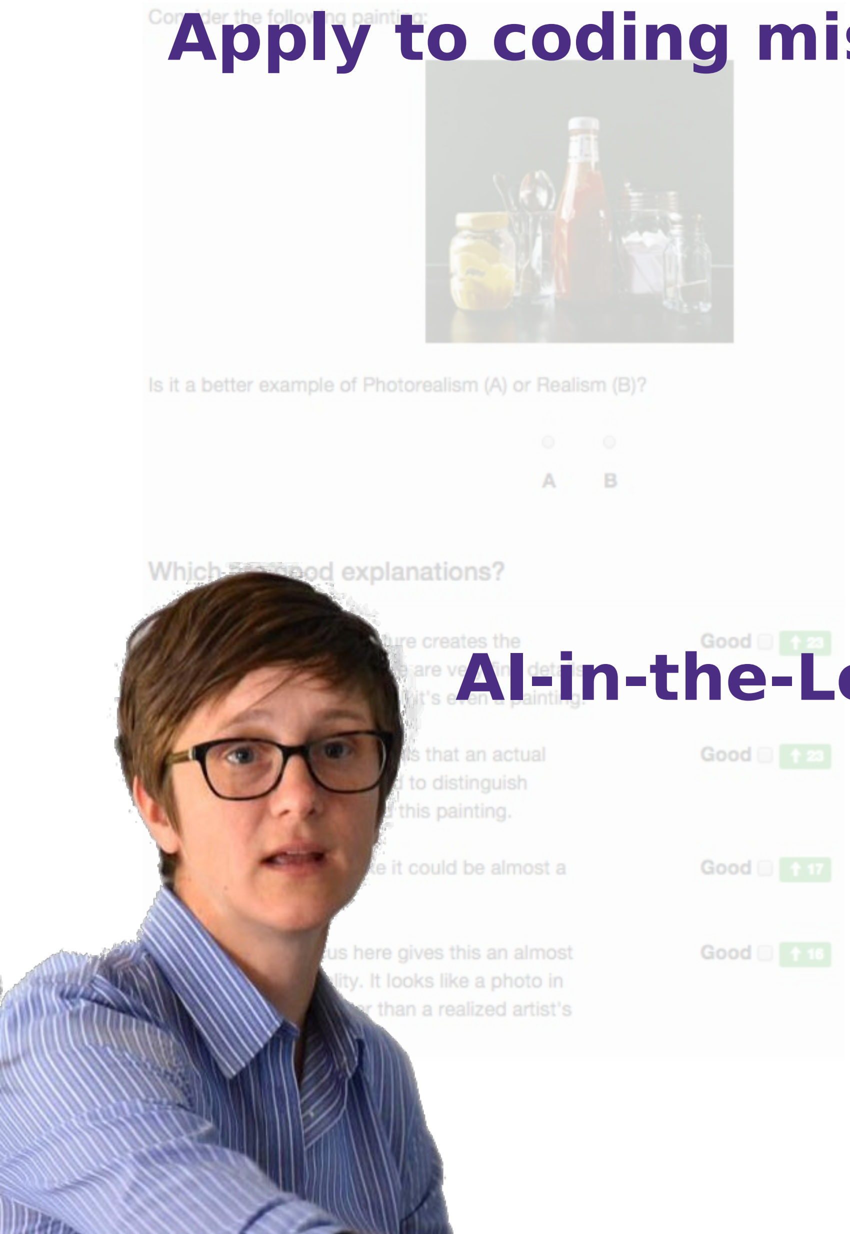
Which are good explanations?

The detail in this picture creates the illusion of reality. There are very fine details that make it hard to tell it's even a painting.	Good	↑ 23
It provides all the details that an actual photo would. It is hard to distinguish between a photo and this painting.	Good	↑ 23
Because it looks like it could be almost a photograph	Good	↑ 17
The sharp focus here gives this an almost hyper-real quality. It looks like a photo in sharp focus, rather than a realized artist's perception of reality.	Good	↑ 16

Collective Intelligence

(Celis, Krafft, Kobe, ICWSM 2016)

Apply to coding misinformation?



Collective Intelligence

AI-in-the-Loop Crowdsourcing

(Celis, Krafft

Apply to coding misinformation?

Collective Intelligence

AI-in-the-Loop Crowdsourcing

EPSRC
Pioneering research and skills

Celis, Krafft

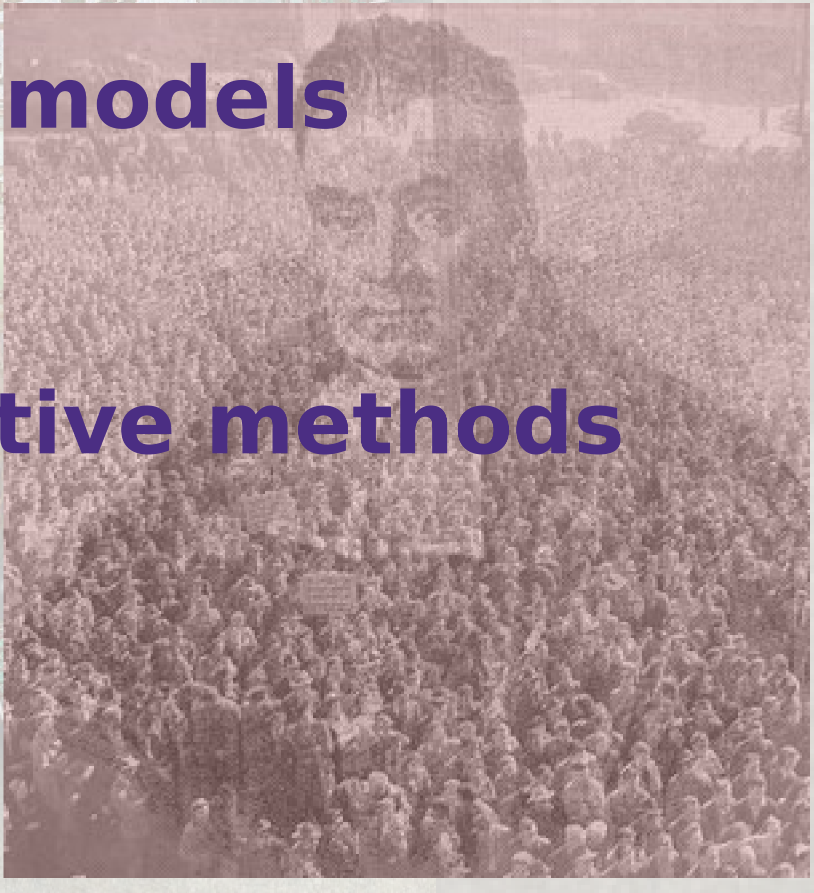
3-5 Year Plan:

Bridging the Cyber and the Social

1) Semi-structured models

2) AI-in-the-Loop

3) Bayesian qualitative methods



Methodology

Epistemology

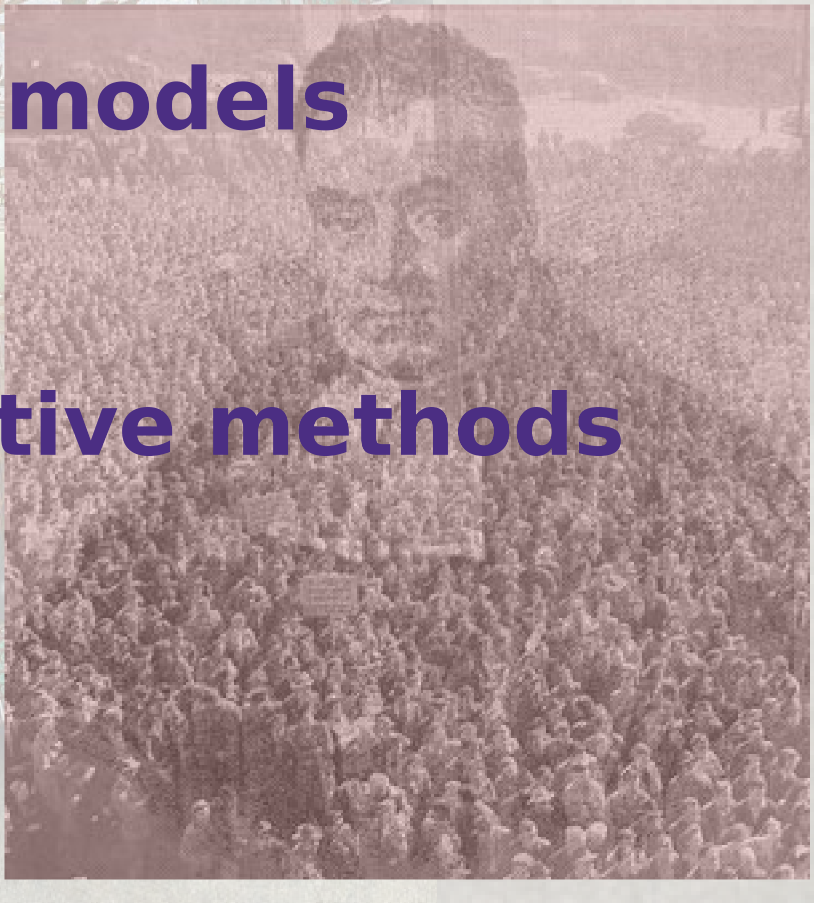
3-5 Year Plan:

<u>**Bridging the Cyber and the Social**</u>

1) Semi-structured models

2) AI-in-the-Loop

3) Bayesian qualitative methods



Methodology

Epistemology

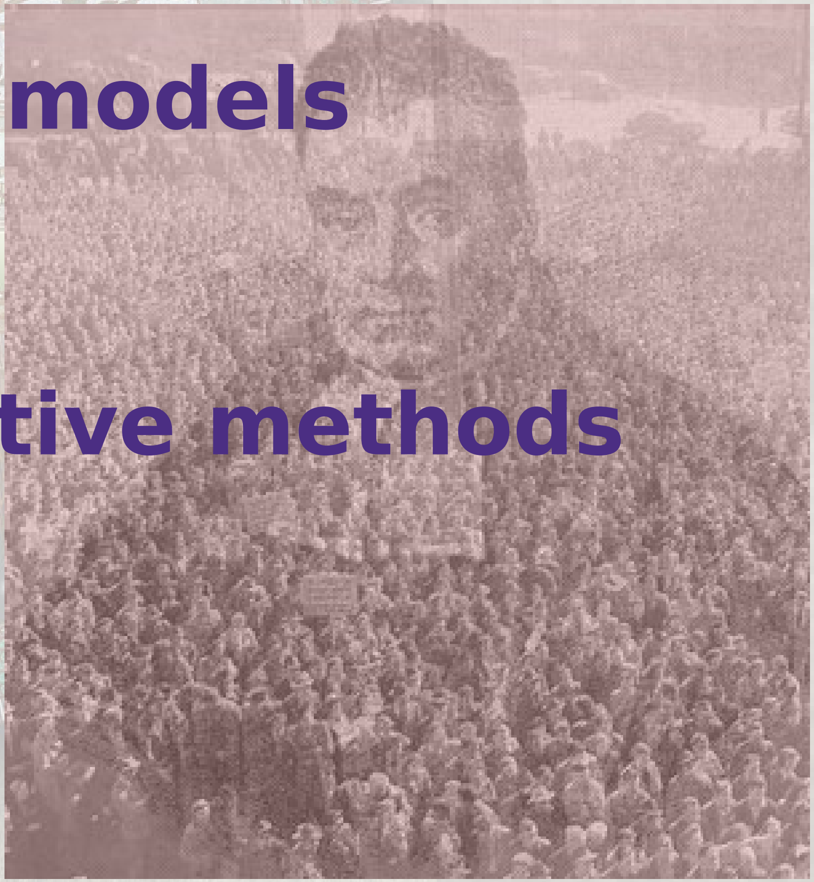
3-5 Year Plan:

<u>Bridging the Cyber and the Social</u>

1) Semi-structured models

2) AI-in-the-Loop

3) Bayesian qualitative methods



Methodology

Epistemology

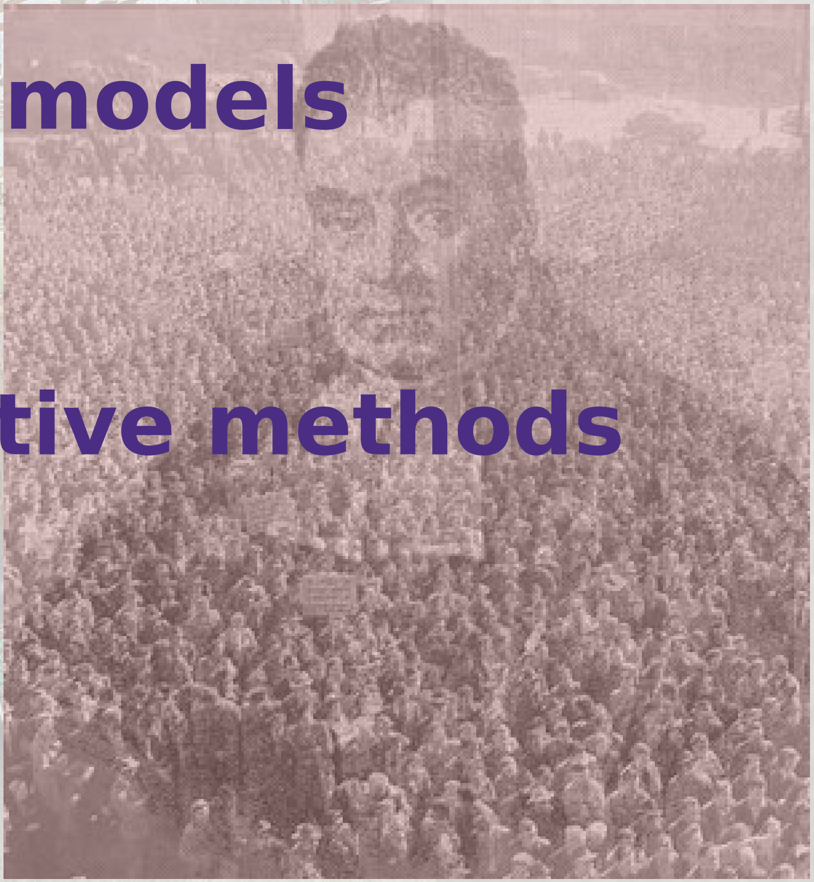
3-5 Year Plan:

<u>Bridging the Cyber and the Social</u>

1) Semi-structured models

2) AI-in-the-Loop

3) Bayesian qualitative methods



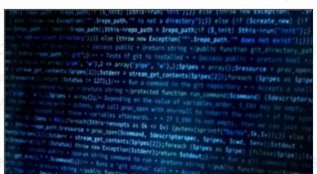
Methodology

Epistemology

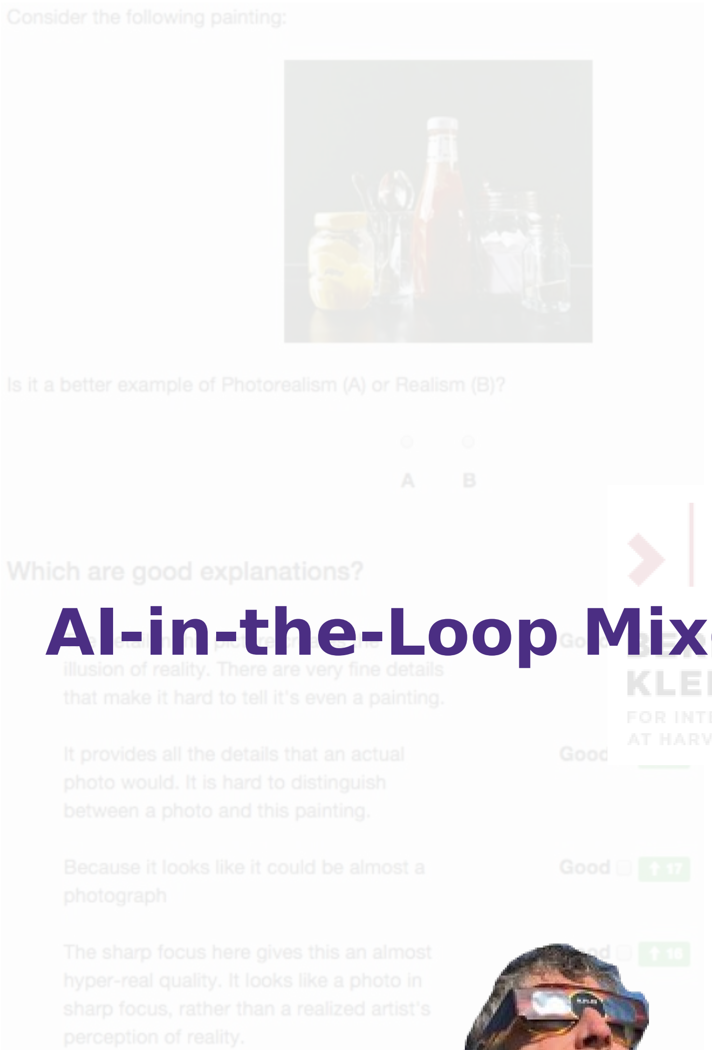
Relation to Fairness, Bias, and Inclusion

Collective Intelligence

Participatory Machine Learning



Ethical auditing for automated decision-making



CRITICAL PLATFORM STUDIES GROUP

At the University of Washington

AI-in-the-Loop Mixed Methods