Text as Data in Social Science: Discovery, Measurement and Causal Inference

Brandon Stewart
Based on joint work with Molly Roberts, Justin Grimmer, Dustin Tingley, Rich Nielsen, Jacob Montgomery, Luwei Ying and others.

Princeton Sociology and Office of Population Research

December 6, 2019

Source Material

- *Text as Data: How to Make Large Scale Inferences from Language* — book manuscript in preparation
  (with Grimmer and Roberts)
- "stm: R package for Structural Topic Models" *Journal of Statistical Software*
  (with Roberts and Tingley)
- "Adjusting for Confounding with Text Matching" *American Journal of Political Science*
  (with Roberts and Nielsen)
- "Inferring Concepts from Topics: Towards Procedures for Validating Topics as Measures"
  (with Ying and Montgomery)
- and others...

Copies of papers at BrandonStewart.org

Big Data, Big Analytics

Big Data, Big Analytics

- Massive increase in unstructured text due to:

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new modes of social interaction (email, forums, Twitter)
  - new/improved data collection (wiki surveys, survey experiments)
  - digitization efforts (government documents, Google Books)

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new modes of social interaction (email, forums, Twitter)
  - new/improved data collection (wiki surveys, survey experiments)
  - digitization efforts (government documents, Google Books)
- Communities now leave digitized footprints we can measure

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new modes of social interaction (email, forums, Twitter)
  - new/improved data collection (wiki surveys, survey experiments)
  - digitization efforts (government documents, Google Books)
- Communities now leave digitized footprints we can measure
- Text analysis is a tool to augment (not replace!) human effort

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new modes of social interaction (email, forums, Twitter)
  - new/improved data collection (wiki surveys, survey experiments)
  - digitization efforts (government documents, Google Books)
- Communities now leave digitized footprints we can measure
- Text analysis is a tool to augment (not replace!) human effort
- Allows for flourishing of custom measurement strategies

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new modes of social interaction (email, forums, Twitter)
  - new/improved data collection (wiki surveys, survey experiments)
  - digitization efforts (government documents, Google Books)
- Communities now leave digitized footprints we can measure
- Text analysis is a tool to augment (not replace!) human effort
- Allows for flourishing of custom measurement strategies
- Combine insights from computer science and social science

Big Data, Big Analytics

- Massive increase in unstructured text due to:
  - new modes of social interaction (email, forums, Twitter)
  - new/improved data collection (wiki surveys, survey experiments)
  - digitization efforts (government documents, Google Books)
- Communities now leave digitized footprints we can measure
- Text analysis is a tool to augment (not replace!) human effort
- Allows for flourishing of custom measurement strategies
- Combine insights from computer science and social science
- This talk $\rightsquigarrow$ designing models for social science applications.

Three Tasks in Social Science Research Design

Three Tasks in Social Science Research Design

1) Discovery

Three Tasks in Social Science Research Design

1) Discovery
   - spark new concepts and explore the data

Three Tasks in Social Science Research Design

1) Discovery
   - spark new concepts and explore the data
   - once we acknowledge discovery as part of the research process, we can develop methods to improve it

Three Tasks in Social Science Research Design

1) Discovery
   - spark new concepts and explore the data
   - once we acknowledge discovery as part of the research process, we can develop methods to improve it
2) Measurement

Three Tasks in Social Science Research Design

1) Discovery
   - spark new concepts and explore the data
   - once we acknowledge discovery as part of the research process, we can develop methods to improve it
2) Measurement
   - operationalize a concept

Three Tasks in Social Science Research Design

1) Discovery
   - spark new concepts and explore the data
   - once we acknowledge discovery as part of the research process, we can develop methods to improve it
2) Measurement
   - operationalize a concept
   - can measure many phenomena at scale

Three Tasks in Social Science Research Design

1) Discovery
   - spark new concepts and explore the data
   - once we acknowledge discovery as part of the research process, we can develop methods to improve it
2) Measurement
   - operationalize a concept
   - can measure many phenomena at scale
3) Causal Inference

Three Tasks in Social Science Research Design

1) Discovery
   - spark new concepts and explore the data
   - once we acknowledge discovery as part of the research process, we can develop methods to improve it
2) Measurement
   - operationalize a concept
   - can measure many phenomena at scale
3) Causal Inference
   - assess the effect of a counterfactual intervention

Three Tasks in Social Science Research Design

1) Discovery
   - spark new concepts and explore the data
   - once we acknowledge discovery as part of the research process, we can develop methods to improve it
2) Measurement
   - operationalize a concept
   - can measure many phenomena at scale
3) Causal Inference
   - assess the effect of a counterfactual intervention
   - text as treatment, outcome and confounder

Roadmap

Roadmap

1) Trends in the literature
2) STM and Designing Methods of the Social Sciences
   - theory of use
   - customization
   - workflow
   - validation
3) Applications in Three Tasks
   - discovery
   - measurement
   - causal inference
4) Conclusions

Different Types of Methods for Different Goals

Different Types of Methods for Different Goals

- Supervised: pursuing a known goal

Different Types of Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents

Different Types of Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest

Different Types of Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest
  - highest effort before running the model

Different Types of Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest
  - highest effort before running the model
- Unsupervised: goal is to learn the goal

Different Types of Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest
  - highest effort before running the model
- Unsupervised: goal is to learn the goal
  - algorithm discovers themes/patterns in the texts

Different Types of Methods for Different Goals

- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest
  - highest effort before running the model
- Unsupervised: goal is to learn the goal
  - algorithm discovers themes/patterns in the texts
  - human interprets the results

Different Types of Methods for Different Goals

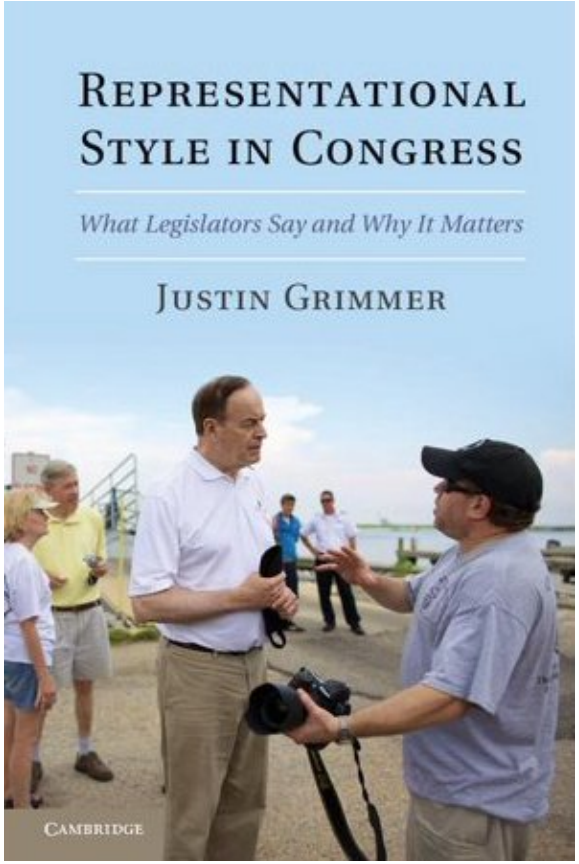
- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest
  - highest effort before running the model
- Unsupervised: goal is to learn the goal
  - algorithm discovers themes/patterns in the texts
  - human interprets the results
  - highest effort after running the model

Different Types of Methods for Different Goals

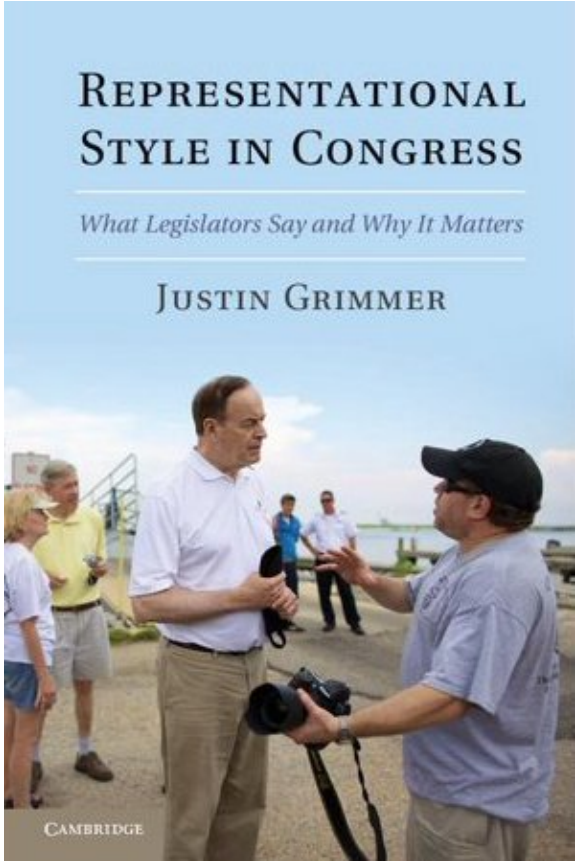
- Supervised: pursuing a known goal
  - human annotates a subset of documents
  - algorithm annotates the rest
  - highest effort before running the model
- Unsupervised: goal is to learn the goal
  - algorithm discovers themes/patterns in the texts
  - human interprets the results
  - highest effort after running the model
- Both strategies amplify human effort, each in different ways.

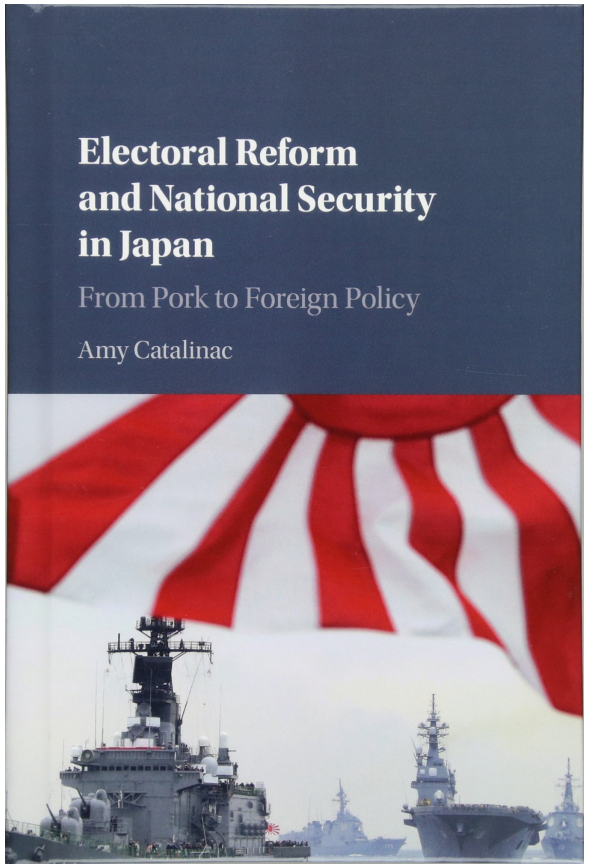
Books/Papers Embrace the Unsupervised Strategy

Books/Papers Embrace the Unsupervised Strategy

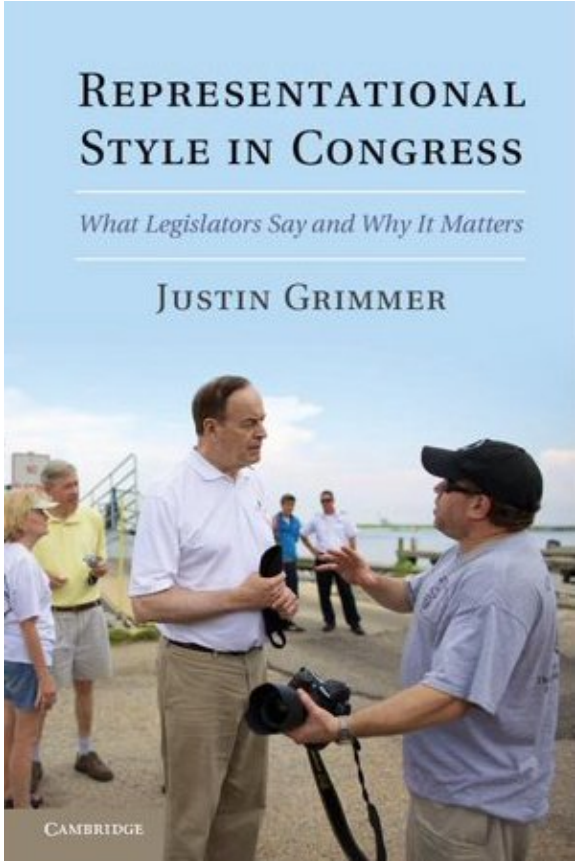


Books/Papers Embrace the Unsupervised Strategy





Books/Papers Embrace the Unsupervised Strategy



Making Austerity Popular: The Media and Mass Attitudes toward Fiscal Policy

Lucy Barnes University College London
Timothy Hicks University College London

Is the Left-Right Scale a Valid Measure of Ideology?

Individual-Level Variation in Associations with "Left" and "Right" and Left-Right Self-Placement

Paul C. Bauer[1] · Pablo Barberá[2] · Kathrin Ackermann[3] · Aaron Venetz[4]

From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections

Amy Catalinac, New York University

Islamophobia and Media Portrayals of Muslim Women: A Computational Text Analysis of US News Coverage

Rochelle Terman

Latent Dirichlet Allocation (Blei et al 2003)

- Core idea: proportional membership in topics.

Latent Dirichlet Allocation (Blei et al 2003)

- Core idea: proportional membership in topics.
- We can learn the topics because:

Latent Dirichlet Allocation (Blei et al 2003)

- Core idea: proportional membership in topics.
- We can learn the topics because:
  - the model rewards a topic which contains as few words as possible

Latent Dirichlet Allocation (Blei et al 2003)

- Core idea: proportional membership in topics.
- We can learn the topics because:
  - the model rewards a topic which contains as few words as possible
  - but a document which contains as few topics as possible.

Latent Dirichlet Allocation (Blei et al 2003)

- Core idea: proportional membership in topics.
- We can learn the topics because:
  - the model rewards a topic which contains as few words as possible
  - but a document which contains as few topics as possible.
- This tension is what encourages the model to develop semantically cohesive sets of topics.

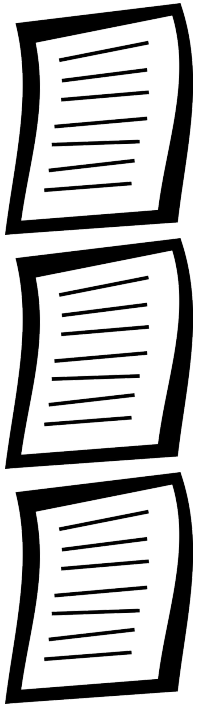
What this means in pictures

Say you have
a lot of people.

What this means in pictures

Say you have
a lot of people.

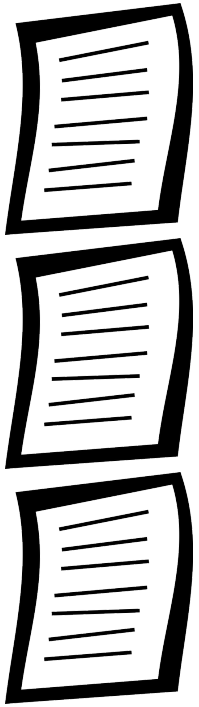
Each writes
some texts



What this means in pictures

Say you have
a lot of people.

Each writes
some texts

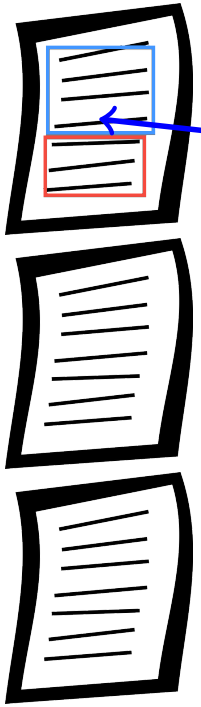


that discuss a few
different topics

What this means in pictures

Say you have
a lot of people.

Each writes
some texts



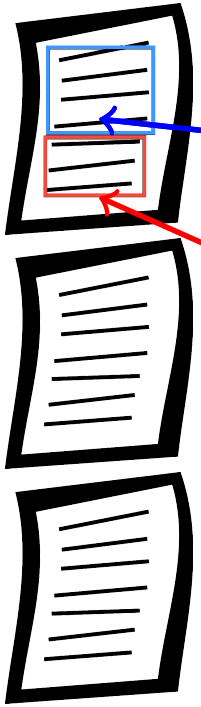
that discuss a few
different topics

Topic 1

What this means in pictures

Say you have
a lot of people.

Each writes
some texts



that discuss a few
different topics

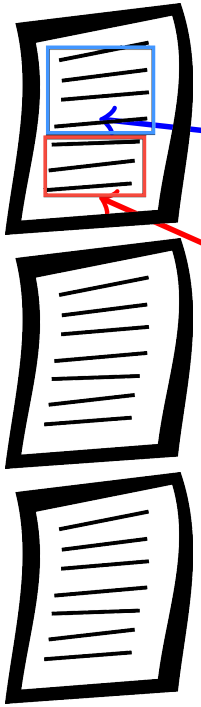
Topic 1

Topic 2

What this means in pictures

Say you have
a lot of people.

Each writes
some texts



that discuss a few
different topics

Topic 1

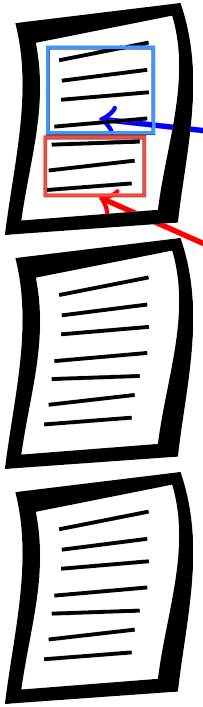
Topic 2

The Latent Dirichlet Allocation estimates:

What this means in pictures

Say you have
a lot of people.

Each writes
some texts



that discuss a few
different topics

Topic 1

Topic 2

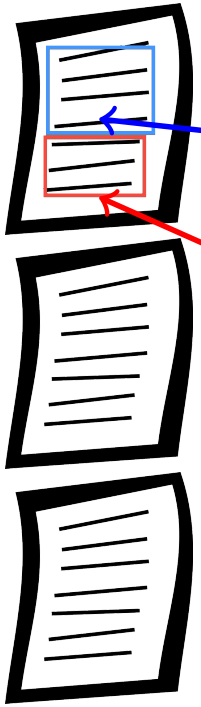
The Latent Dirichlet Allocation estimates:

1. The topics- each is a distribution over words

What this means in pictures

Say you have
a lot of people.

Each writes
some texts



that discuss a few
different topics

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The Latent Dirichlet Allocation estimates:

1. The topics- each is a distribution over words

What this means in pictures

Say you have
a lot of people.

Each writes
some texts

that discuss a few
different topics

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

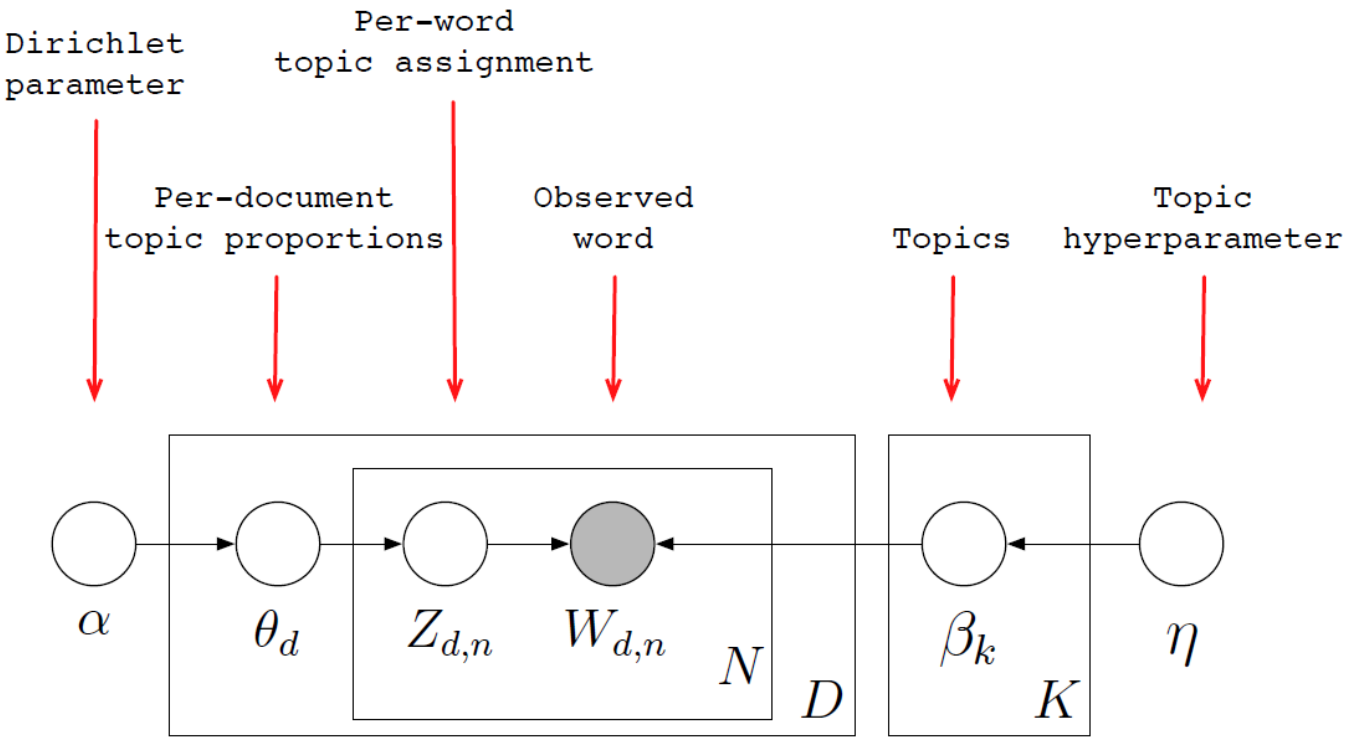
estimator, data, analysis, variance, model, inference

The Latent Dirichlet Allocation estimates:

1. The topics- each is a distribution over words
2. The proportion of each document in each topic

This is a Bayesian Model

Figure: Plate Notation of Latent Dirichlet Allocation



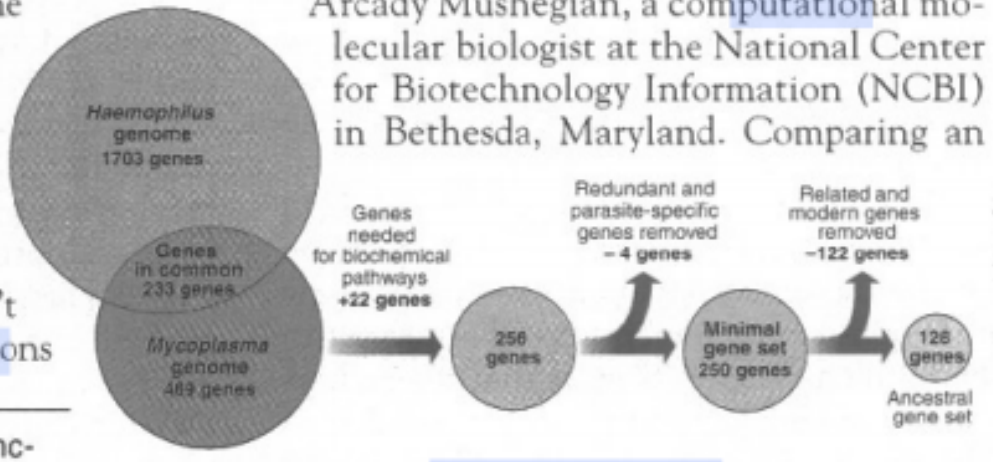
Graphic from Dave Blei

A Statistical Highlighter (With Many Colors)

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

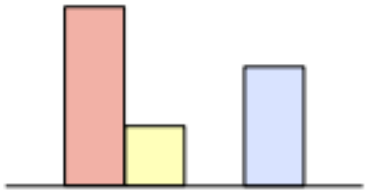


Image from Hanna Wallach

Leveraging Information Within and About Texts

Leveraging Information Within and About Texts

- Previous methods leverage the information within documents

Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
- But, we also have information about documents

Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
- But, we also have information about documents
- Use in Practice

Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
- But, we also have information about documents
- Use in Practice
  - social scientists want to measure variation in topics and explore how observed covariates drive trends in language

Leveraging Information Within and About Texts

- Previous methods leverage the information within documents
- But, we also have information about documents
- Use in Practice
  - social scientists want to measure variation in topics and explore how observed covariates drive trends in language
  - we want to tell a story not just about what, but how and why

Designing Methods in Social Science

Designing Methods in Social Science

- A series of special case models beginning to be proposed to handle structure of time (Quinn et al 2010), author (Grimmer 2010) and other settings.

Designing Methods in Social Science

- A series of special case models beginning to be proposed to handle structure of time (Quinn et al 2010), author (Grimmer 2010) and other settings.
- Not all social scientists have the time and background to design their own model. Can we build something more modular?

Designing Methods in Social Science

- A series of special case models beginning to be proposed to handle structure of time (Quinn et al 2010), author (Grimmer 2010) and other settings.
- Not all social scientists have the time and background to design their own model. Can we build something more modular?
- Four components to designing methods:

Designing Methods in Social Science

- A series of special case models beginning to be proposed to handle structure of time (Quinn et al 2010), author (Grimmer 2010) and other settings.
- Not all social scientists have the time and background to design their own model. Can we build something more modular?
- Four components to designing methods:
  1) Propose a Theory of Use

Designing Methods in Social Science

- A series of special case models beginning to be proposed to handle structure of time (Quinn et al 2010), author (Grimmer 2010) and other settings.
- Not all social scientists have the time and background to design their own model. Can we build something more modular?
- Four components to designing methods:
  1) Propose a Theory of Use
  2) Customize the Method

Designing Methods in Social Science

- A series of special case models beginning to be proposed to handle structure of time (Quinn et al 2010), author (Grimmer 2010) and other settings.
- Not all social scientists have the time and background to design their own model. Can we build something more modular?
- Four components to designing methods:
  1) Propose a Theory of Use
  2) Customize the Method
  3) Design the Workflow

Designing Methods in Social Science

- A series of special case models beginning to be proposed to handle structure of time (Quinn et al 2010), author (Grimmer 2010) and other settings.
- Not all social scientists have the time and background to design their own model. Can we build something more modular?
- Four components to designing methods:
  1) Propose a Theory of Use
  2) Customize the Method
  3) Design the Workflow
  4) Create Validations

Designing Methods in Social Science

- A series of special case models beginning to be proposed to handle structure of time (Quinn et al 2010), author (Grimmer 2010) and other settings.
- Not all social scientists have the time and background to design their own model. Can we build something more modular?
- Four components to designing methods:
  1) Propose a Theory of Use
  2) Customize the Method
  3) Design the Workflow
  4) Create Validations
- Illustration through the Structural Topic Model

1) Propose a Theory of Use

- Social scientists use text as data methods for different purposes than the fields where the ideas are originally proposed.

1) Propose a Theory of Use

- Social scientists use text as data methods for different purposes than the fields where the ideas are originally proposed.
- Articulating the principles that guide usage of the model, helps us build more effective models.

1) Propose a Theory of Use

- Social scientists use text as data methods for different purposes than the fields where the ideas are originally proposed.
- Articulating the principles that guide usage of the model, helps us build more effective models.
- Our understanding: social scientists using document-level latent variables to tell stories about how and why topics allocated attention over time.

2) Customization

STM = LDA + Contextual Information

2) Customization

STM = LDA + Contextual Information

STM provides two ways to include contextual information

- Topic prevalence can vary by metadata
  - e.g. Democrats talk more about education than Republicans

2) Customization

STM = LDA + Contextual Information

STM provides two ways to include contextual information

- Topic prevalence can vary by metadata
  - e.g. Democrats talk more about education than Republicans
- Topic content can vary by metadata
  - e.g. Democrats are less likely to use the word "life" when talking about abortion than Republicans

STM: What this means in pictures

Say you have
a lot of people.

Each writes
some text

that discuss a few
different topics

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

STM: What this means in pictures

Say you have a lot of people.

Each writes some text

that discuss a few different topics

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

1. The words in each topic to vary by gender

STM: What this means in pictures

Say you have
a lot of people.

Each writes
some text

that discuss a few
different topics

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

1. The words in each topic to vary by gender

STM: What this means in pictures

Say you have
a lot of people.

Each writes
some text

that discuss a few
different topics

Group A

Group B

Group A

Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

1. The words in each topic to vary by gender
2. The topic proportions to vary by group

STM: What this means in pictures

Say you have
a lot of people.

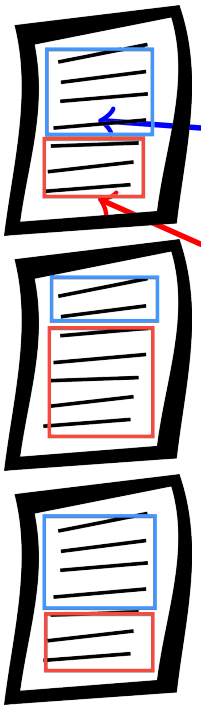
Each writes
some text

that discuss a few
different topics

Group A

Group B

Group A



Politics

congress, nations, power, votes, agreement, bargaining

Statistics

estimator, data, analysis, variance, model, inference

The STM Allows for:

1. The words in each topic to vary by gender
2. The topic proportions to vary by group

Mixed-Membership Topic Models

More formal terminology:

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $\mathbf{X}$

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $\mathbf{X}$
  - Topical content groups: D length vector $\mathbf{Y}$

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $\mathbf{X}$
  - Topical content groups: D length vector $\mathbf{Y}$
- Latent variables

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $\mathbf{X}$
  - Topical content groups: D length vector $\mathbf{Y}$
- Latent variables
  - $D \times K$ matrix θ: proportion of document on each topic.

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $\mathbf{X}$
  - Topical content groups: D length vector $\mathbf{Y}$
- Latent variables
  - $D \times K$ matrix θ: proportion of document on each topic.
  - $K \times V$ matrix β: probability of drawing a word conditional on topic.

Mixed-Membership Topic Models

More formal terminology:

- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional data for STM
  - Topic prevalence covariates: $D \times P$ matrix $\mathbf{X}$
  - Topical content groups: D length vector $\mathbf{Y}$
- Latent variables
  - $D \times K$ matrix θ: proportion of document on each topic.
  - $K \times V$ matrix β: probability of drawing a word conditional on topic.
  - Low rank approximation to expected counts: $\underset{D \times V}{\tilde{W}} \approx \underset{D \times K}{\theta} \underset{K \times V}{\beta}$

The Structural Topic Model

- Low rank approximation to expected counts: $\underset{D\times V}{\tilde{W}} \approx \underset{D\times K}{\theta} \underset{K\times V}{\beta}$
- θ, $D \times K$ document-topic matrix
- β, $K \times V$ topic-word matrix
- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{d,n}$ from Discrete(θ_d)
  - Draw observed word $w_{d,n}$ from Discrete($\beta_{k=z,}$)

The Structural Topic Model

- Low rank approximation to expected counts: $\underset{D\times V}{\tilde{W}} \approx \underset{D\times K}{\theta} \underset{K\times V}{\beta}$
- θ, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates
- β, $K \times V$ topic-word matrix
- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{d,n}$ from Discrete(θ_d)
  - Draw observed word $w_{d,n}$ from Discrete($\beta_{k=z,}$)

The Structural Topic Model

- Low rank approximation to expected counts: $\underset{D\times V}{\tilde{W}} \approx \underset{D\times K}{\theta} \underset{K\times V}{\beta}$
- θ, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{d,\cdot} \sim \text{LogisticNormal}(X_d\gamma, \Sigma)$
- β, $K \times V$ topic-word matrix

- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{d,n}$ from Discrete(θ_d)
  - Draw observed word $w_{d,n}$ from Discrete($\beta_{k=z,}$)

The Structural Topic Model

- Low rank approximation to expected counts: $\underset{D\times V}{\tilde{W}} \approx \underset{D\times K}{\theta} \underset{K\times V}{\beta}$
- θ, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{d,\cdot} \sim \text{LogisticNormal}(X_d\gamma, \Sigma)$
- β, $K \times V$ topic-word matrix $\Leftarrow$ multinomial logit with covariates

- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{d,n}$ from Discrete(θ_d)
  - Draw observed word $w_{d,n}$ from Discrete($\beta_{k=z,}$)

The Structural Topic Model

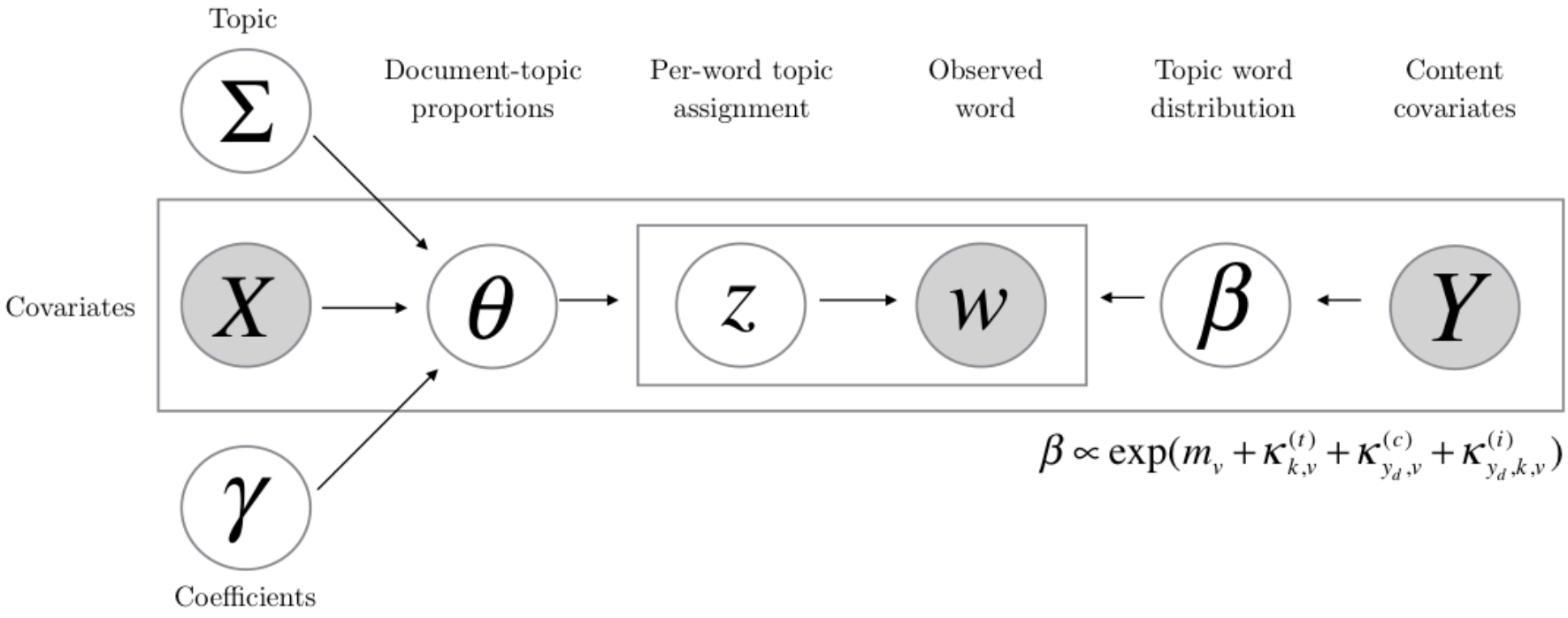
- Low rank approximation to expected counts: $\underset{D\times V}{\tilde{W}} \approx \underset{D\times K}{\theta}\underset{K\times V}{\beta}$
- θ, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{d,\cdot} \sim \text{LogisticNormal}(X_d\gamma, \Sigma)$
- β, $K \times V$ topic-word matrix $\Leftarrow$ multinomial logit with covariates
  - Each topic is now a sparse, covariate-specific deviation from a baseline distribution.
  - $\vec{\beta}_{k,\cdot} \propto \exp\left(m + \kappa^{(\text{topic})} + \kappa^{(\text{cov})} + \kappa^{(\text{int})}\right)$
  - Thee parts: topic, covariate, topic-covariate interaction

- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{d,n}$ from Discrete(θ_d)
  - Draw observed word $w_{d,n}$ from Discrete($\beta_{k=z,}$)

The Structural Topic Model

- Low rank approximation to expected counts: $\underset{D\times V}{\tilde{W}} \approx \underset{D\times K}{\theta}\underset{K\times V}{\beta}$
- θ, $D \times K$ document-topic matrix $\Leftarrow$ logistic normal glm with covariates
  - Covariate-specific prior with global topic covariance
  - $\theta_{d,\cdot} \sim \text{LogisticNormal}(X_d\gamma, \Sigma)$
- β, $K \times V$ topic-word matrix $\Leftarrow$ multinomial logit with covariates
  - Each topic is now a sparse, covariate-specific deviation from a baseline distribution.
  - $\vec{\beta}_{k,\cdot} \propto \exp\left(m + \kappa^{(\text{topic})} + \kappa^{(\text{cov})} + \kappa^{(\text{int})}\right)$
  - Thee parts: topic, covariate, topic-covariate interaction
  - β may instead by point-estimated
- Each token has a topic drawn from the document mixture
  - Draw token topic $z_{d,n}$ from Discrete(θ_d)
  - Draw observed word $w_{d,n}$ from Discrete($\beta_{k=z,}$)

Structural Topic Model



$$\beta \propto \exp(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v})$$

Problem Solving for Practical Use

Problem Solving for Practical Use

- Fast Estimation (Roberts, Stewart and Airoldi 2016)
  bayesian estimation using partially-collapsed non-conjugate variational inference

Problem Solving for Practical Use

- Fast Estimation (Roberts, Stewart and Airoldi 2016)
  bayesian estimation using partially-collapsed non-conjugate variational inference
- Uncertainty (Roberts et al 2014)
  method of composition to include uncertainty when estimating covariate effects.

Problem Solving for Practical Use

- Fast Estimation (Roberts, Stewart and Airoldi 2016)
  bayesian estimation using partially-collapsed non-conjugate variational inference
- Uncertainty (Roberts et al 2014)
  method of composition to include uncertainty when estimating covariate effects.
- Multimodality (Roberts et al 2016)
  initialization using a spectral method of moments estimator from Arora et al 2013

Problem Solving for Practical Use

- Fast Estimation (Roberts, Stewart and Airoldi 2016)
  bayesian estimation using partially-collapsed non-conjugate variational inference
- Uncertainty (Roberts et al 2014)
  method of composition to include uncertainty when estimating covariate effects.
- Multimodality (Roberts et al 2016)
  initialization using a spectral method of moments estimator from Arora et al 2013
- Multilingual Text (Lucas et al 2015)
  modeling of a translated term-document matrix with modeled language difference

3) Designing the Workflow

- Workflows more important than models

3) Designing the Workflow

- Workflows more important than models
- People use good practice when it is easy and they don't generally change defaults.

3) Designing the Workflow

- Workflows more important than models
- People use good practice when it is easy and they don't generally change defaults.
- `stm` Package in `R`
  - complete workflow: raw texts → figures
  - simple regression style syntax using formulas

```
mod.out <- stm(documents,vocab, K=10,
               prevalence= ~paper + s(time),
               data=metadata, init.type="Spectral")
```

  - many functions for summarization, visualization and checking

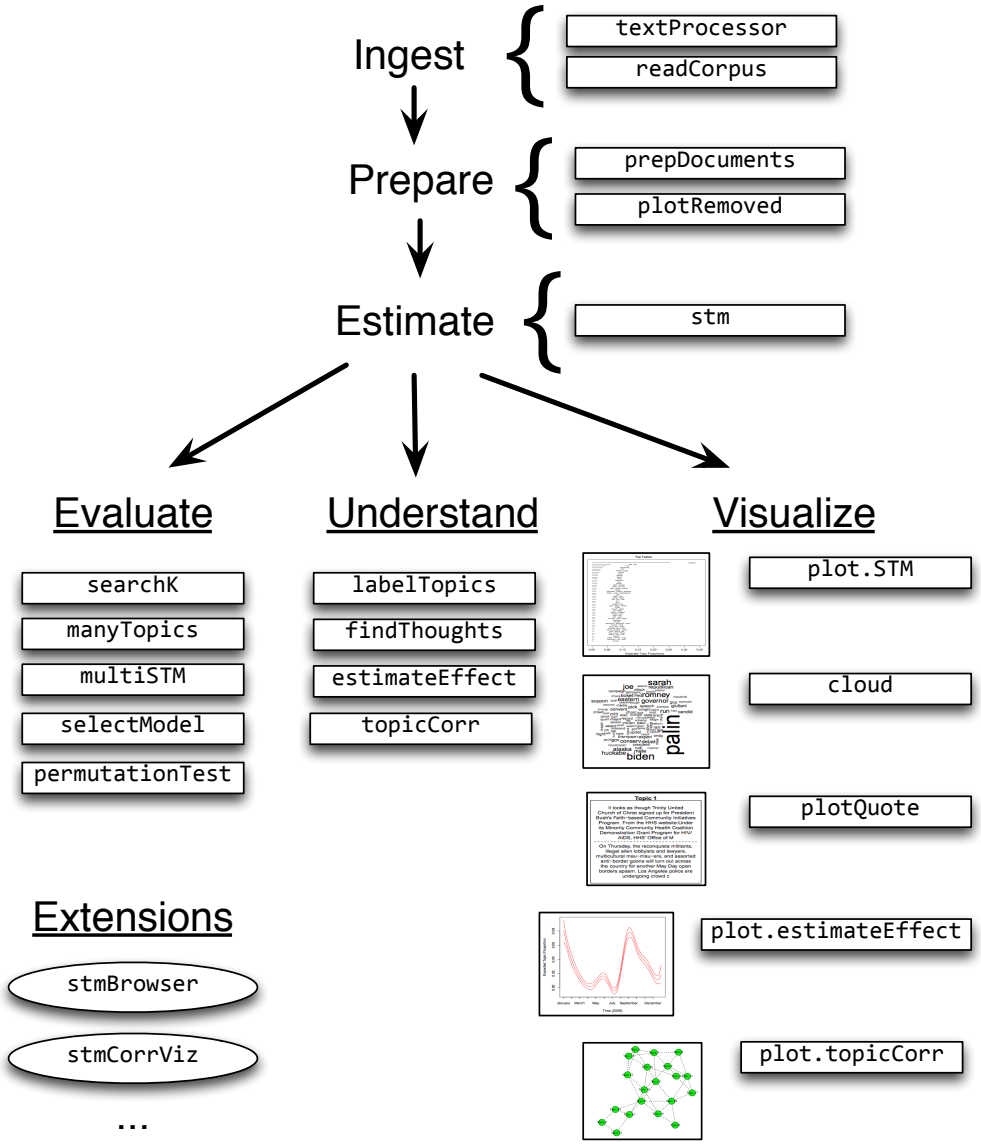
3) Designing the Workflow

- Workflows more important than models
- People use good practice when it is easy and they don't generally change defaults.
- `stm` Package in `R`
  - complete workflow: raw texts → figures
  - simple regression style syntax using formulas

```
mod.out <- stm(documents,vocab, K=10,
                    prevalence= ~paper + s(time),
                    data=metadata, init.type="Spectral")
```

  - many functions for summarization, visualization and checking
- Complete vignette online with examples

Workflow

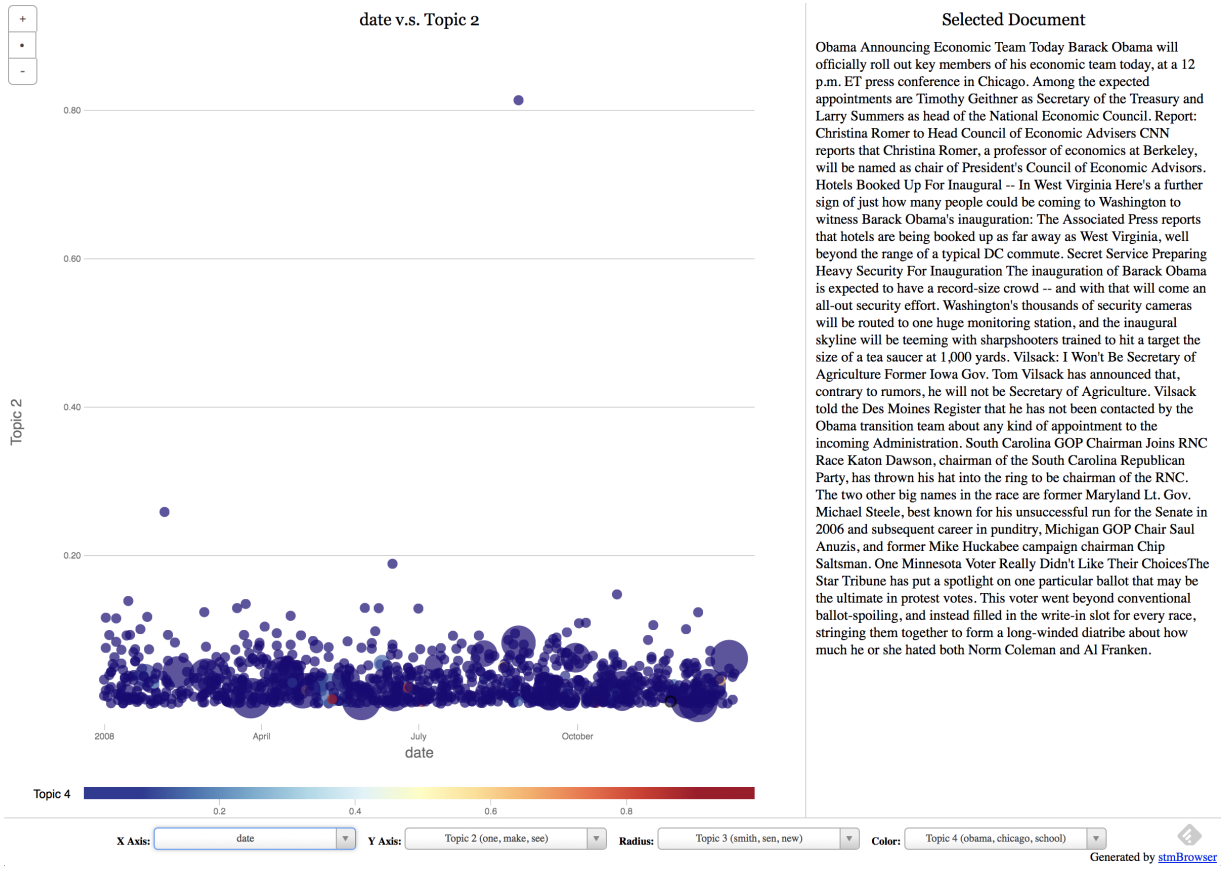

Ingest
textProcessor
readCorpus
Prepare
prepDocuments
plotRemoved
Estimate
stm
Evaluate
searchK
manyTopics
multiSTM
selectModel
permutationTest
Understand
labelTopics
findThoughts
estimateEffect
topicCorr
Visualize
plot.STM
cloud
plotQuote
plot.estimateEffect
plot.topicCorr
Extensions
stmBrowser
stmCorrViz
...

Auxiliary Visualization Packages

Auxiliary Visualization Packages

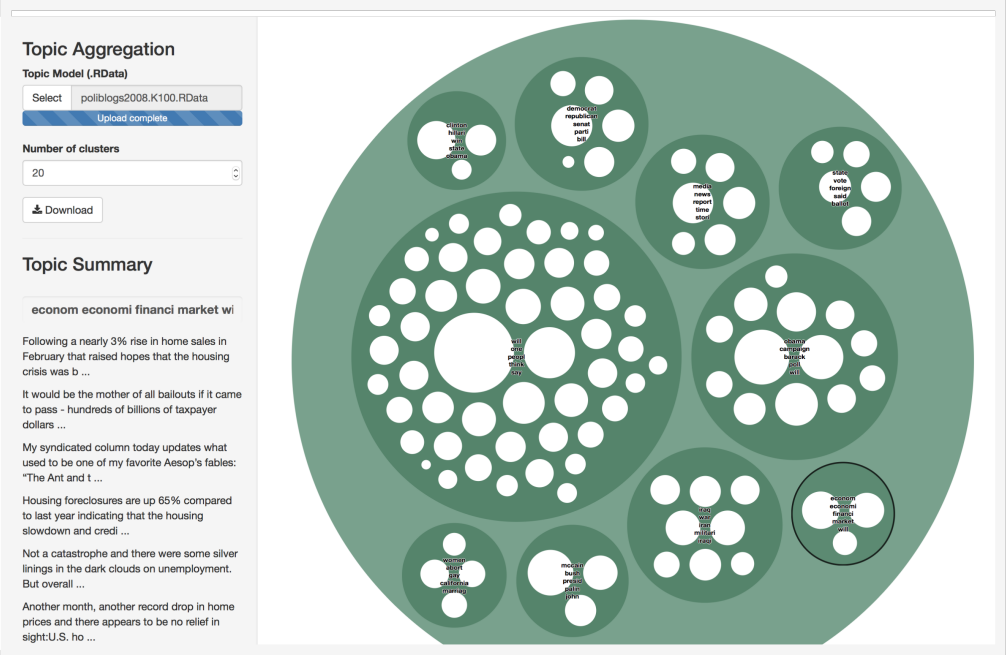
`stmBrowser`

(with Freeman, Chuang, Roberts and Tingley)



Auxiliary Visualization Packages

Trellis
(with Chaney and Schaffner)



Creating Validations

Creating Validations

- How do we know we are getting it right?

Creating Validations

- How do we know we are getting it right?
- Two related goals:
  1) validate measure in our application

Creating Validations

- How do we know we are getting it right?
- Two related goals:
  1) validate measure in our application
  2) evaluate new methods for producing measurements

Creating Validations

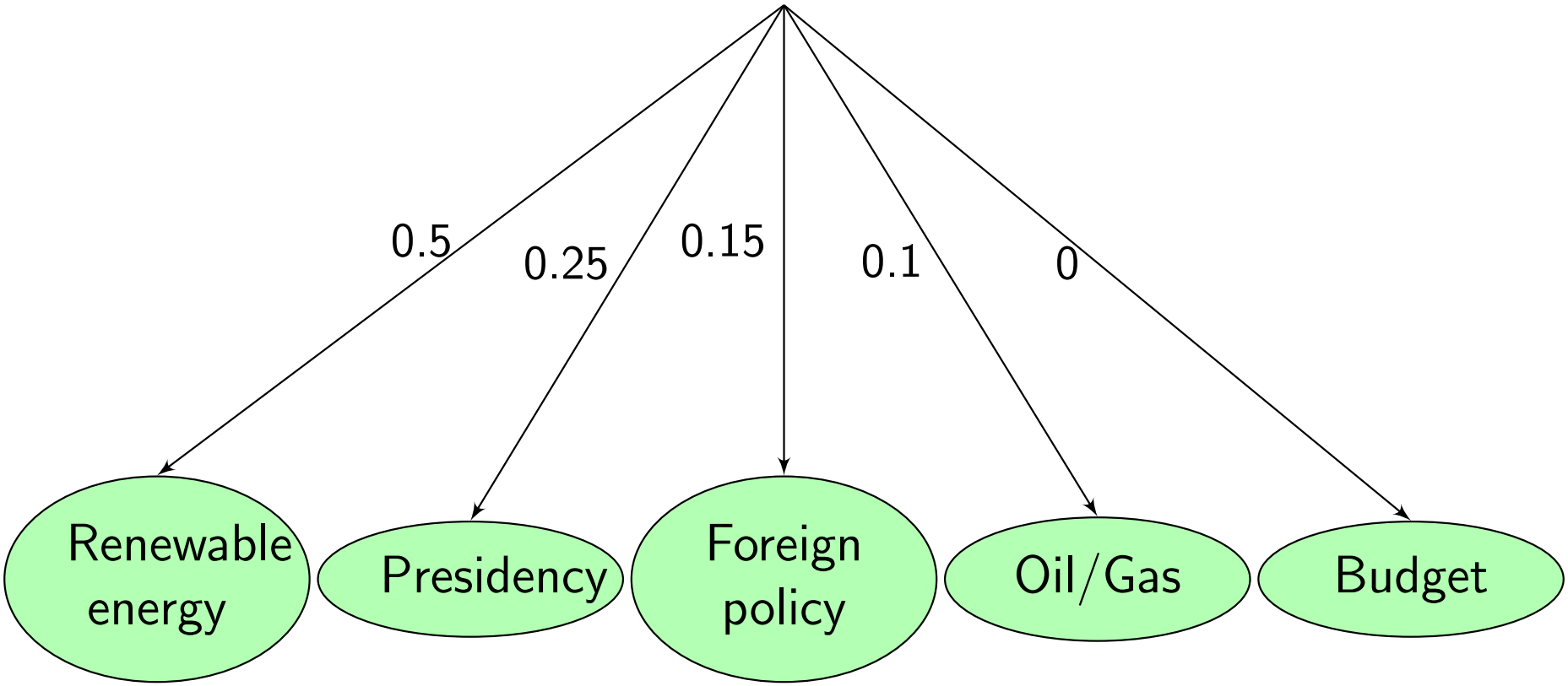
- How do we know we are getting it right?
- Two related goals:
  1) validate measure in our application
  2) evaluate new methods for producing measurements
- CS has advanced through clear, shared benchmarks
  shared task framework, leaderboards, and 'state-of-the-art' performance

Creating Validations

- How do we know we are getting it right?
- Two related goals:
  1) validate measure in our application
  2) evaluate new methods for producing measurements
- CS has advanced through clear, shared benchmarks
  shared task framework, leaderboards, and 'state-of-the-art' performance
- Difficult in social science — evaluation depends on the goal
  do custom measurements imply custom validations?

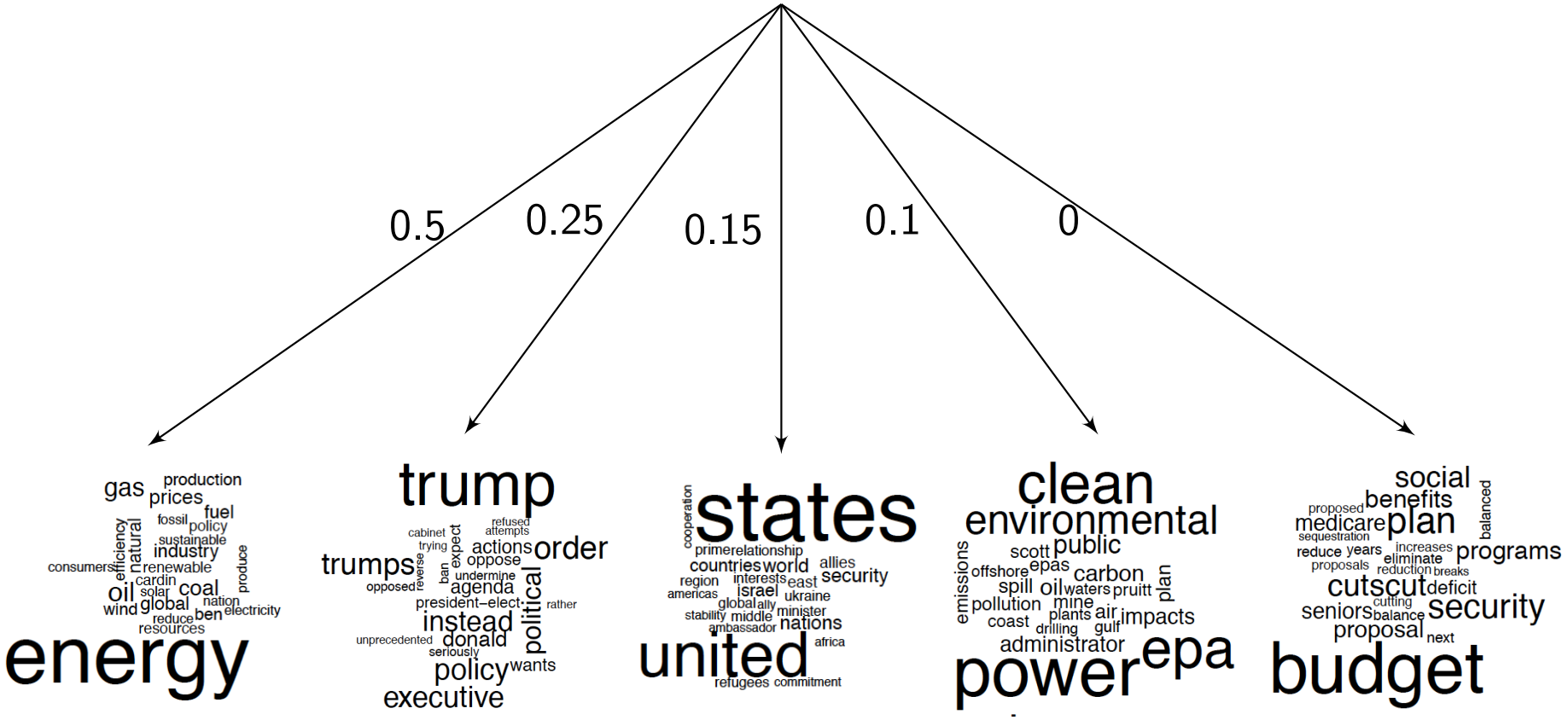
Topic Models as Presented (Ying et al 2019)

Why is Obama tapping oil reserves without adopting a national energy policy that will lessen America's dependence on foreign oil?

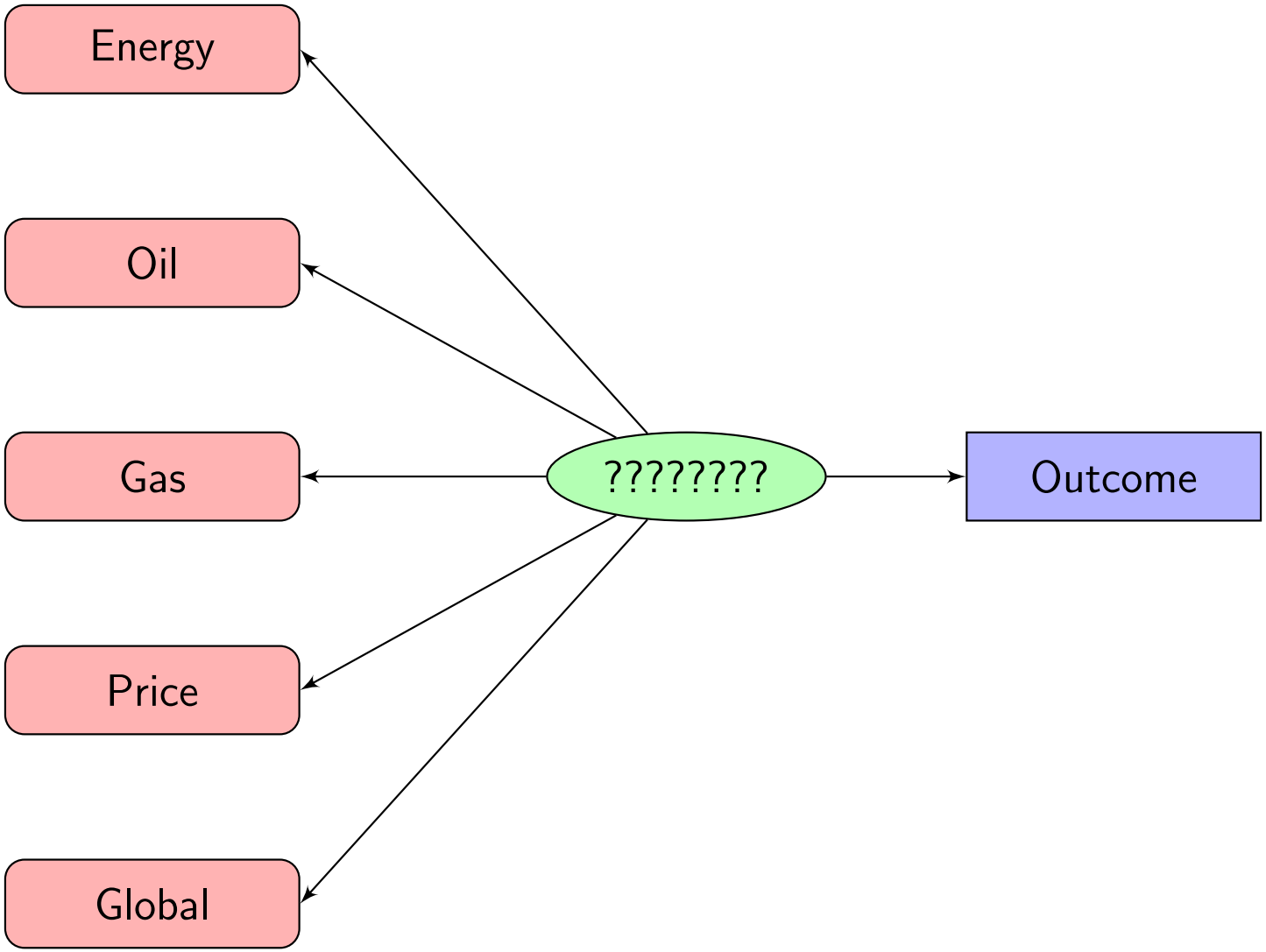


Topic Models Under the Hood

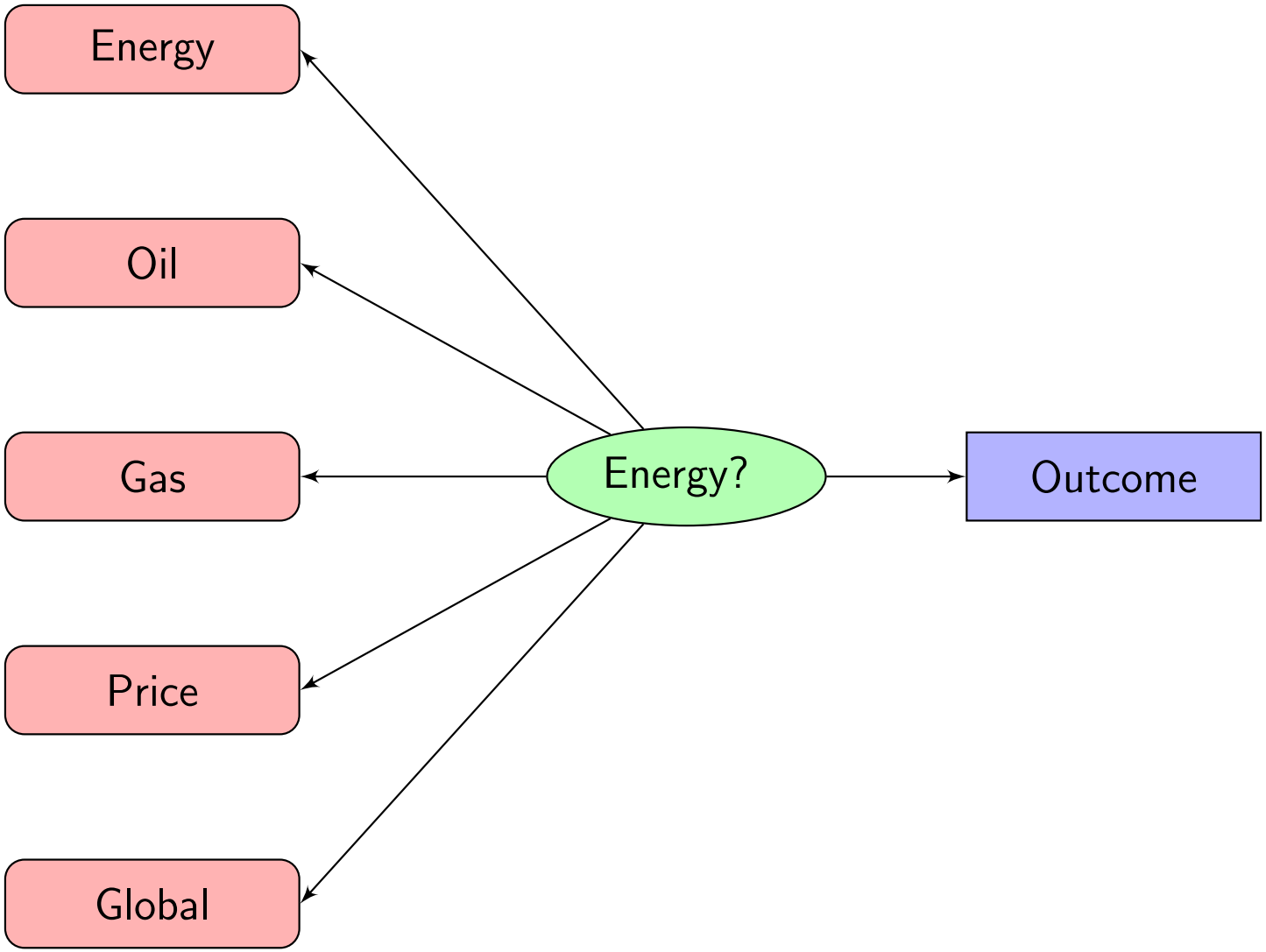
Why is Obama tapping oil reserves without adopting a national energy policy that will lessen America's dependence on foreign oil?



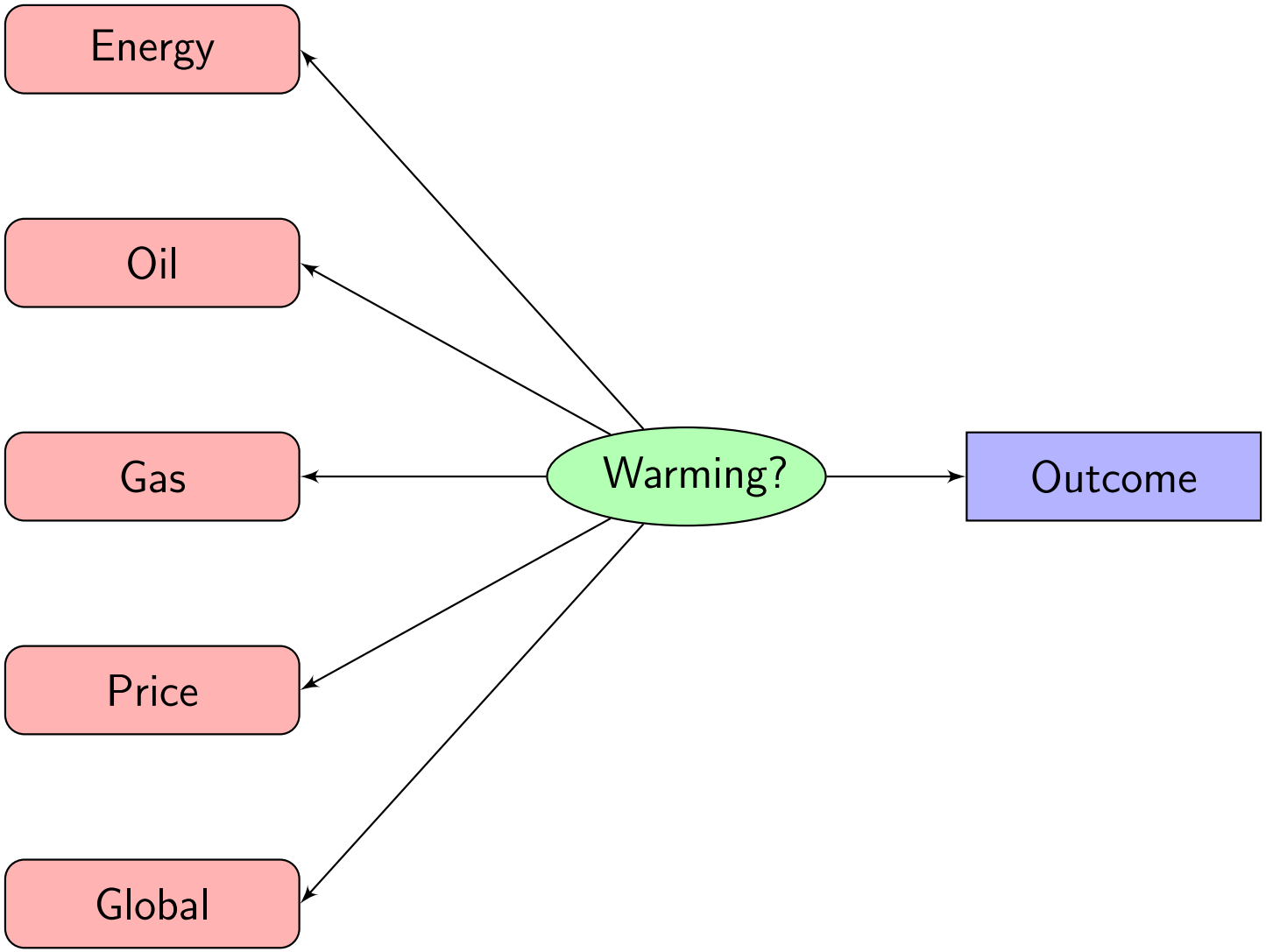
What's the concept in this cloud?: It matters



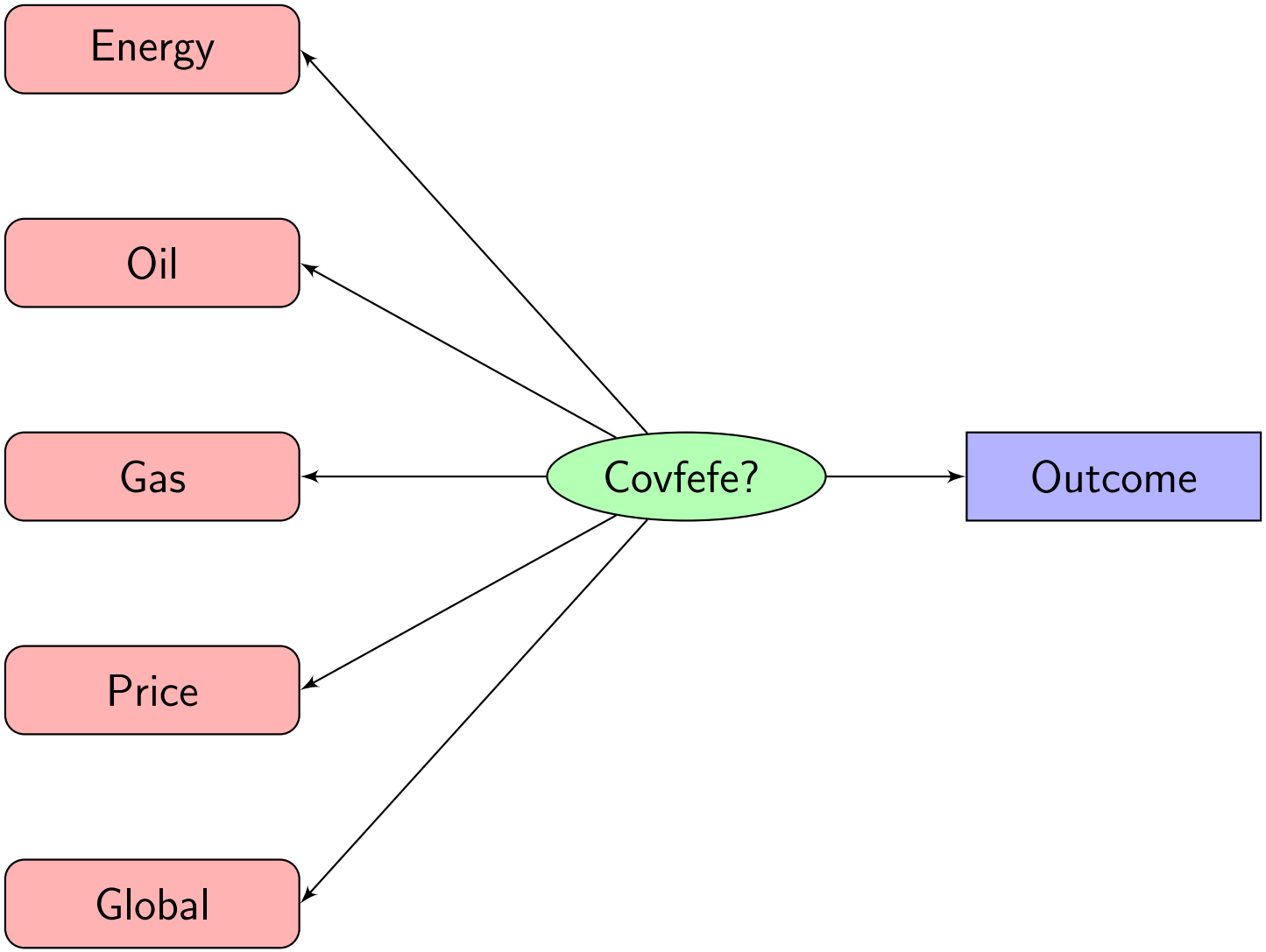
Energy?



Global Warming?



Are we just modeling a mashup?



Existing Approaches to Validation

Existing Approaches to Validation

1) Qualitative Validation
   careful reading
   private, application-specific, not-scalable

Existing Approaches to Validation

1) Qualitative Validation
   careful reading
   private, application-specific, not-scalable
2) Model Fit
   held-out log likelihood, residual dispersion
   public, assumes the model, scalable

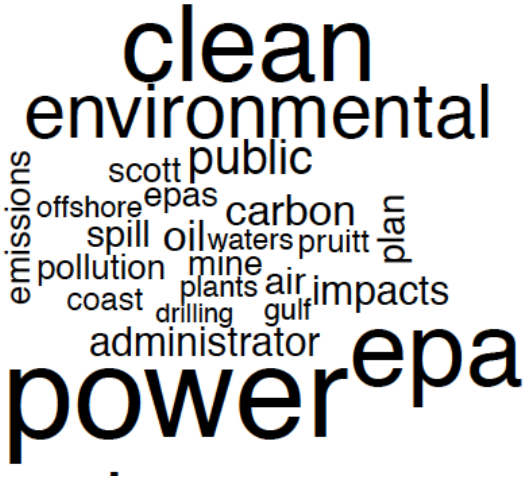
Existing Approaches to Validation

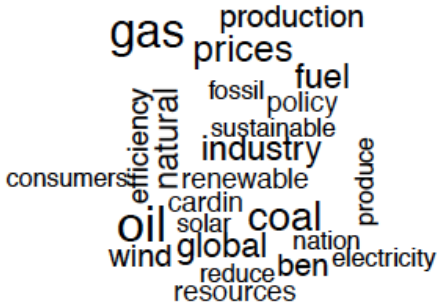
1) Qualitative Validation
   careful reading
   private, application-specific, not-scalable
2) Model Fit
   held-out log likelihood, residual dispersion
   public, assumes the model, scalable
3) Quality-Surrogate Function
   semantic coherence, exclusivity, crowd-sourced evaluations
   public, assumes the surrogate, scalable

Existing Approaches to Validation

1) Qualitative Validation
   careful reading
   private, application-specific, not-scalable
2) Model Fit
   held-out log likelihood, residual dispersion
   public, assumes the model, scalable
3) Quality-Surrogate Function
   semantic coherence, exclusivity, crowd-sourced evaluations
   public, assumes the surrogate, scalable
4) Bespoke Evaluation
   alternate measures, external events, supervised checks
   public-ish, application-specific, maybe scalable?

Validations as a Reader: Wordclouds





When you explain qualitative validation

Wisdom of crowds (Ying et al 2019)

- Topics validation as a set of little games

[PDF] **Reading tea leaves**: How humans interpret topic models

J Chang, S Gerrish, C Wang… - Advances in neural …, 2009 - papers.nips.cc

Probabilistic topic models are a popular tool for the unsupervised analysis of text, providing both a predictive model of future text and a latent topic representation of the corpus. Practitioners typically assume that the latent space is semantically meaningful. It is used to …

☆ ❝❝ Cited by 1413 Related articles All 24 versions »

Wisdom of crowds (Ying et al 2019)

- Topics validation as a set of little games

  [PDF] **Reading tea leaves**: How humans interpret topic models

  J Chang, S Gerrish, C Wang… - Advances in neural …, 2009 - papers.nips.cc

  Probabilistic topic models are a popular tool for the unsupervised analysis of text, providing both a predictive model of future text and a latent topic representation of the corpus. Practitioners typically assume that the latent space is semantically meaningful. It is used to …

  ☆ ❞❞ Cited by 1413 Related articles All 24 versions »

- Less convincing than customized validations, but ...
  - More generalizable
  - Almost always available

Wisdom of crowds (Ying et al 2019)

- Topics validation as a set of little games

  [PDF] **Reading tea leaves**: How humans interpret topic models
  J Chang, S Gerrish, C Wang… - Advances in neural …, 2009 - papers.nips.cc
  Probabilistic topic models are a popular tool for the unsupervised analysis of text, providing both a predictive model of future text and a latent topic representation of the corpus. Practitioners typically assume that the latent space is semantically meaningful. It is used to …
  ☆ ❞❞ Cited by 1413 Related articles All 24 versions »

- Less convincing than customized validations, but ...
  - More generalizable
  - Almost always available
- Our aim:
  - Tweak existing method
  - Add new "games" for validating topic labels

Wisdom of crowds (Ying et al 2019)

- Topics validation as a set of little games

  [PDF] **Reading tea leaves**: How humans interpret topic models
  J Chang, S Gerrish, C Wang… - Advances in neural …, 2009 - papers.nips.cc
  Probabilistic topic models are a popular tool for the unsupervised analysis of text, providing both a predictive model of future text and a latent topic representation of the corpus. Practitioners typically assume that the latent space is semantically meaningful. It is used to …
  ☆ ❞❞ Cited by 1413 Related articles All 24 versions »

- Less convincing than customized validations, but ...
  - More generalizable
  - Almost always available
- Our aim:
  - Tweak existing method
  - Add new "games" for validating topic labels

Example: Random 4 Word Set Intrusion

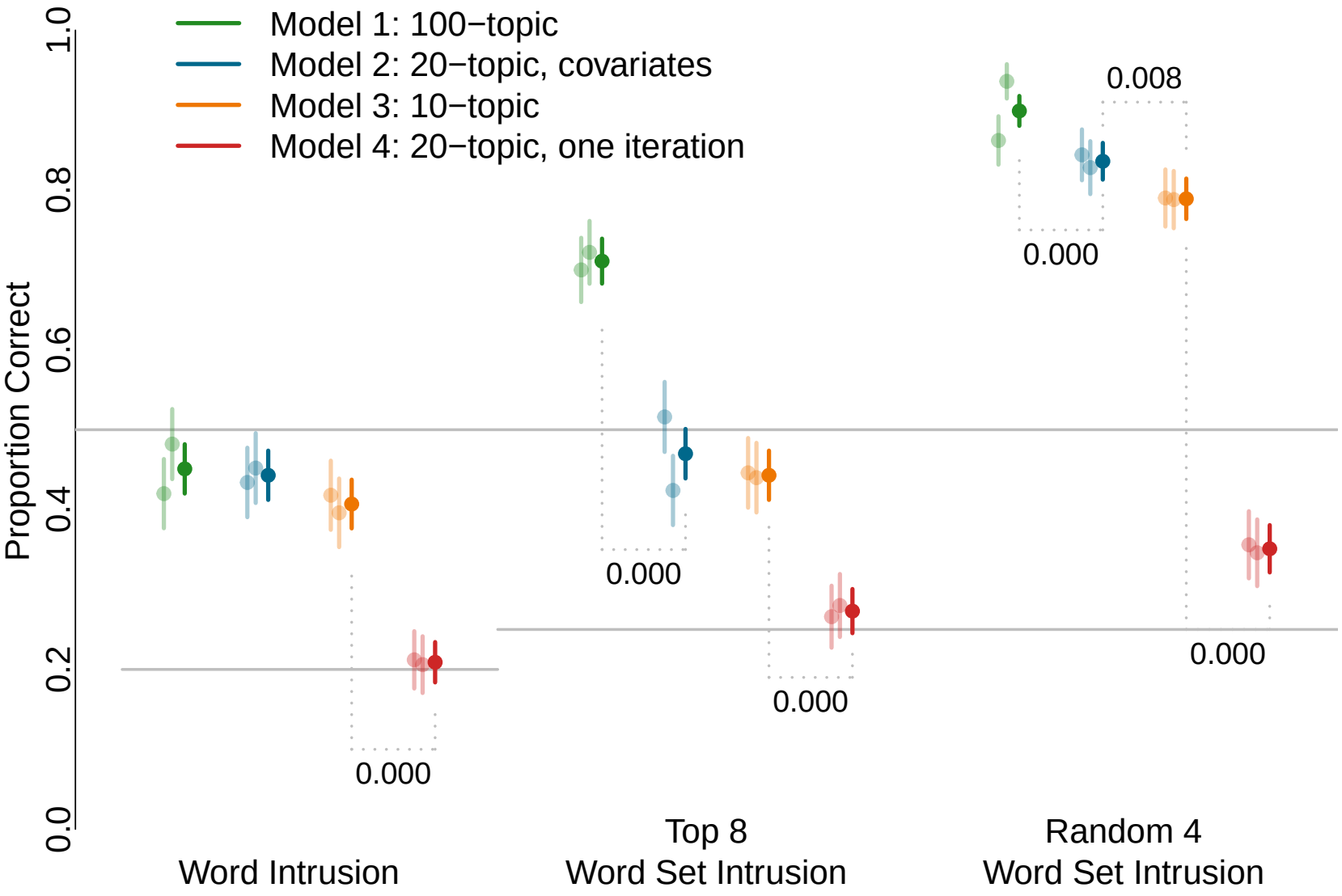
When it comes to medical treatment, children are not small adults, and the EMSC program recognizes the special care and attention they deserve. This low-cost program has saved the lives of countless children and adolescents in the past 30 years, and I urge my colleagues to support this critically important program.

After reading the above passage, please click on the set of words below that is MOST UNRELATED to the passage.

- health, human, hospital, planned
- visiting, pictured, tour, visit
- care, access, serves, women
- human, centers, medical, improve

Submit

New Games Discriminate Better



Example: Label Intrusion (LI)

Great to visit a Kansas energy producer today. They are creating jobs and energy for Kansas and the nation. Keep up the good work!

After reading the above passage, please click on the label below that is MOST UNRELATED to the passage.

- Renewable Energy
- Jobs & Keystone Pipeline
- Raising the Debt Ceiling
- Economic & Job Growth

Submit

Adoption

stm: An R Package for the Structural Topic Model

Website: www.structuraltopicmodel.com

Vignette: Here

Authors: Molly Roberts, Brandon Stewart and Dustin Tingley

Please email all comments/questions to bms4 [AT] princeton.edu

CRAN 1.3.4 | build passing | downloads 5682/month | downloads 113K

Adoption

View on GitHub

stm

An R Package for the Structural Topic Model

tar.gz .zip

Download the Vignette

Authors: Molly Roberts, Brandon Stewart and Dustin Tingley

Please email all comments/questions to bms4 [AT] princeton.edu

News

July 31, 2018

1. We were honored to win the Political Methodology Society's Statistical Software Award for 2018.
2. Added Mikael Johannesson's stmprinter package to Supporting Packages.

Adoption

Published Applications

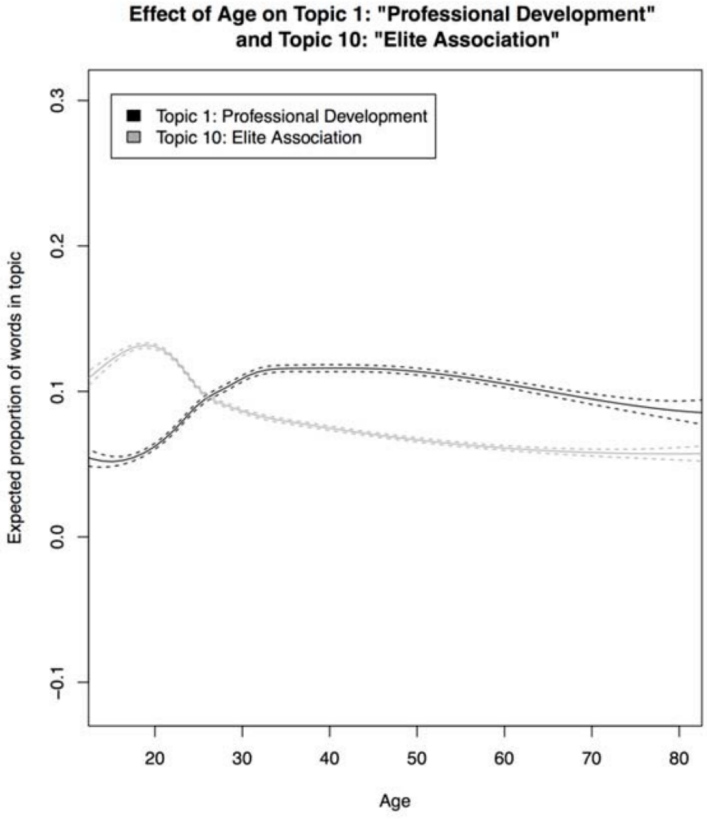
If you have published a paper using stm that you would like to see included here please email us.

1. Schwemmer and Ziewiecki. "Social media Sellout: The Increasing Role of Product Promotion on YouTube." *Social Media + Society.* 2018.
2. Dybowski and Adämmer. "The economic effects of U.S. presidential tax communication: Evidence from a correlated topic model" *European Journal of Political Economy* 2018.
3. Rothschild, Howat, Shafranek, Busby. "Pigeonholing Partisans: Stereotypes of Party Supporters and Partisan Polarization." *Political Behavior* 2018.
4. Chandelier, Steuckardt, Mathevet, Diwersy, Gimenez. "Content analysis of newspaper coverage of wolf recolonization in France using structural topic modeling." *Biological Conservation* 2018.
5. Cerchiello and Nicola. "Assessing News Contagion in Finance" *Econometrics* 2018.
6. Nelson, Laura K. "Computational Grounded Theory: A Methodological Framework" *Sociological methods & Research* 2018.
7. Bohr and Dunlap. "Key Topics in environmental sociology, 1990–2014: results from a computational text analysis"*Environmental Sociology* 2018.
8. Banks, Woznyj, Wesslen and Ross. "A Review of Best Practice Recommendations for Text Analysis in R (and a User-Friendly App)"*Journal of Business and Psychology* 2018.
9. Hagen, Harrison and Dumas. "Data Analytics for Policy Informatics: The Case of E-Petitioning" *Policy Analytics, Modelling, and Informatics* 2018.
10. Kuhn, Kenneth D. "Using structural topic modeling to identify latent topics and trends in aviation incident reports" *Transportation Research Part C: Emerging Technologies* 2018.
11. Tvinnereim, Flottum, Gjerstad, Johannesson and Nordo. "Citizens' preferences for tackling climate change. Quantitative and qualitative analyses of their freely formulated solutions "*Global Environmental Change* 2017.

Applications: Discovery

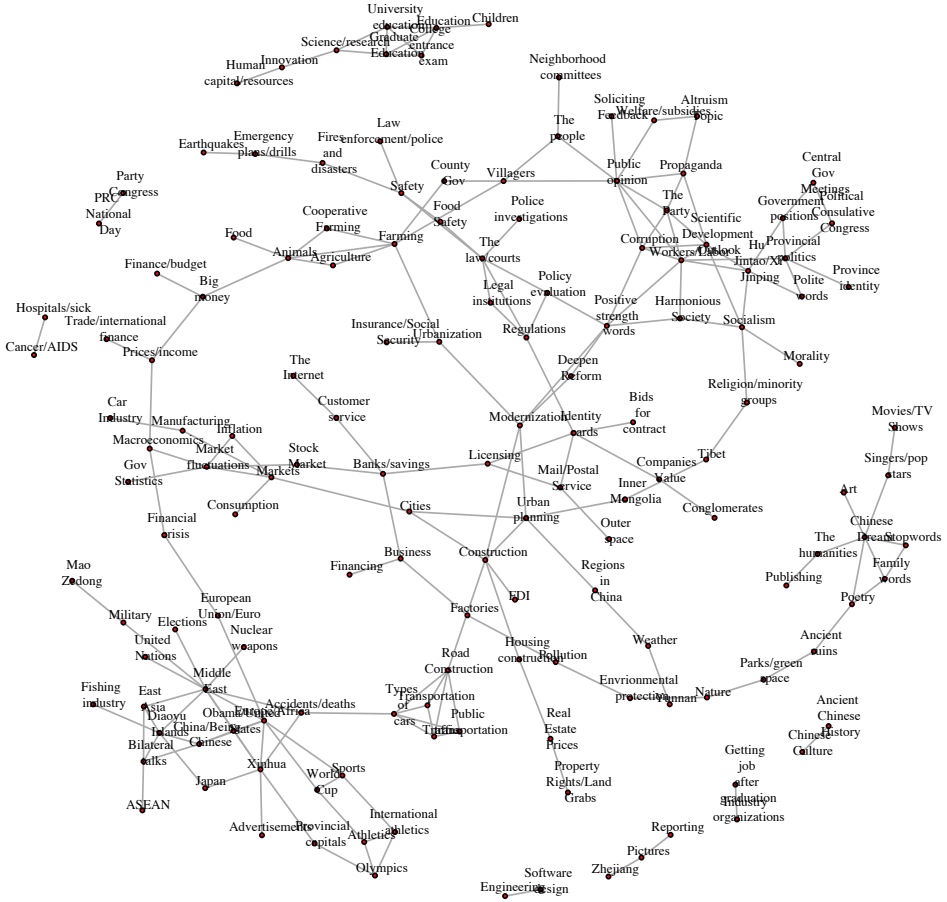
Applications: Discovery

Computer-Assisted Reading and Discovery for Student-Generated Text in Massive Open Online Courses
(with Reich et al in *Journal of Learning Analytics*)



Applications: Discovery

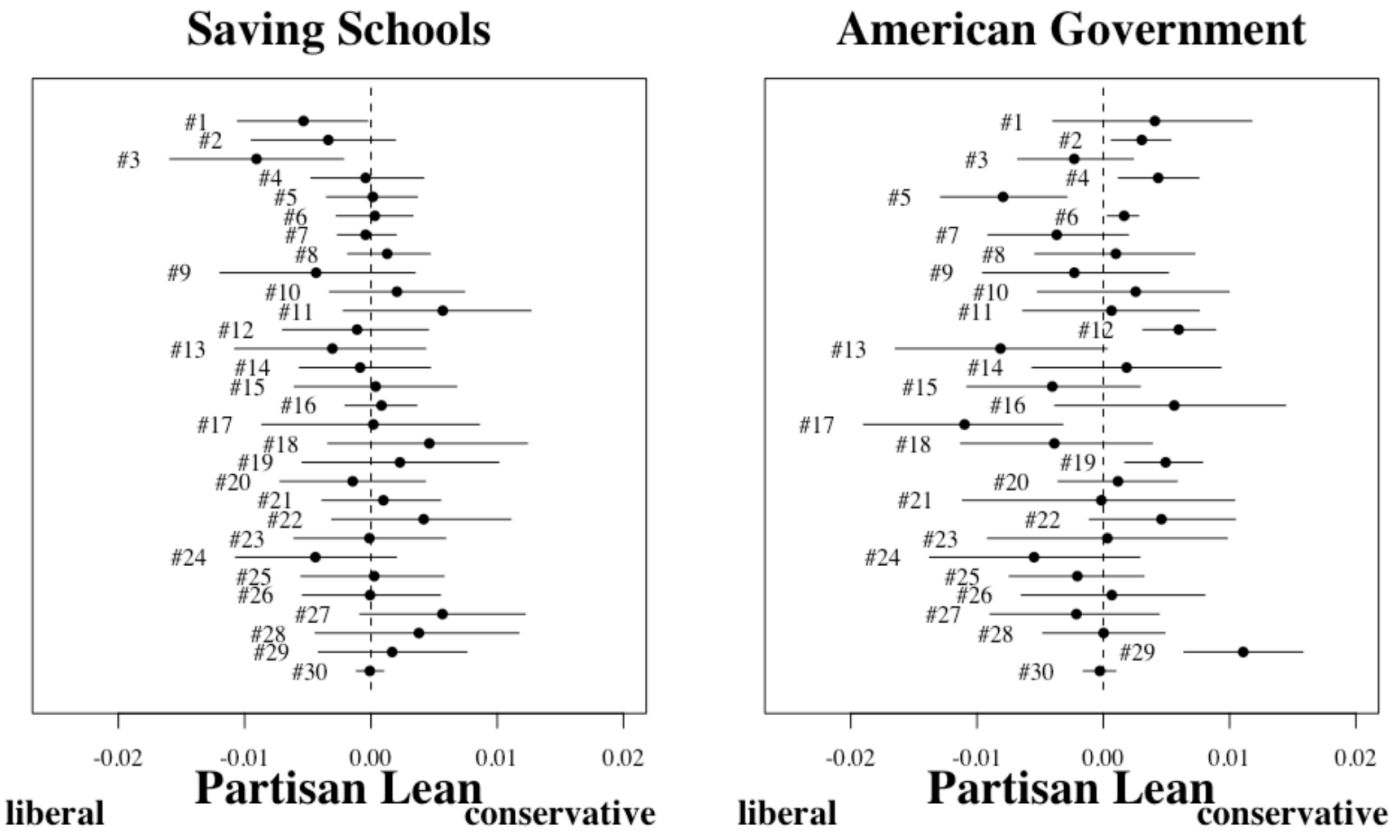
Propaganda in China
(with Roberts)



Applications: Measurement

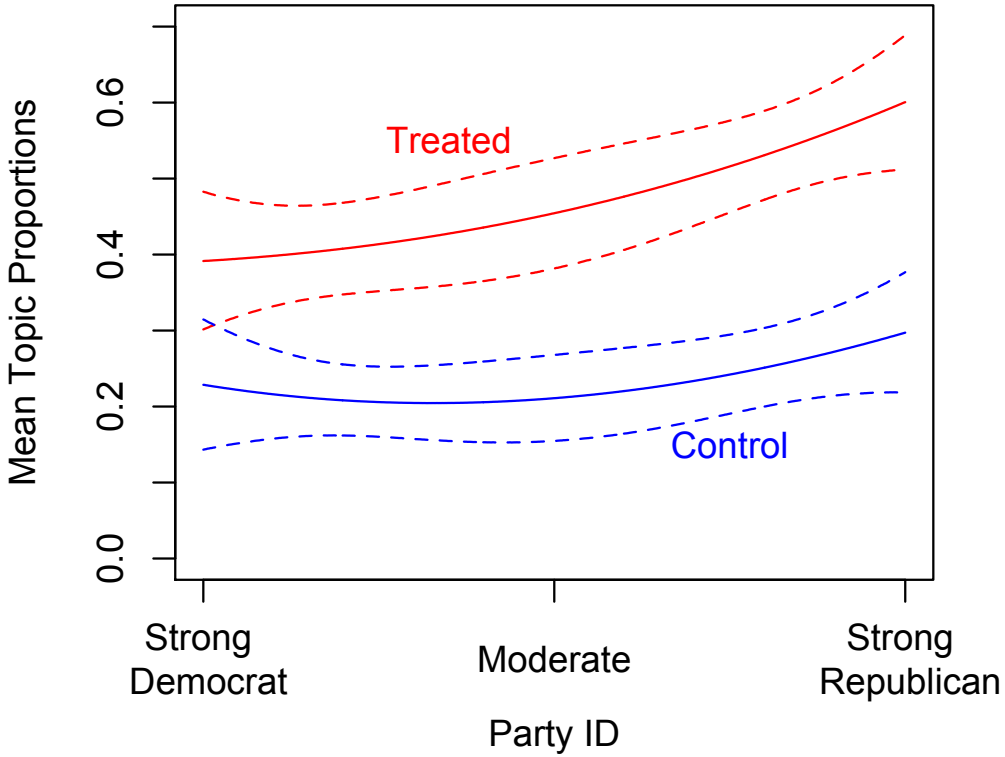
Applications: Measurement

The Civic Mission of MOOCs: Computational Measures of Engagement Across Differences in Online Courses.
(with Yeomans et al in *IJAED*)



Applications: Measurement

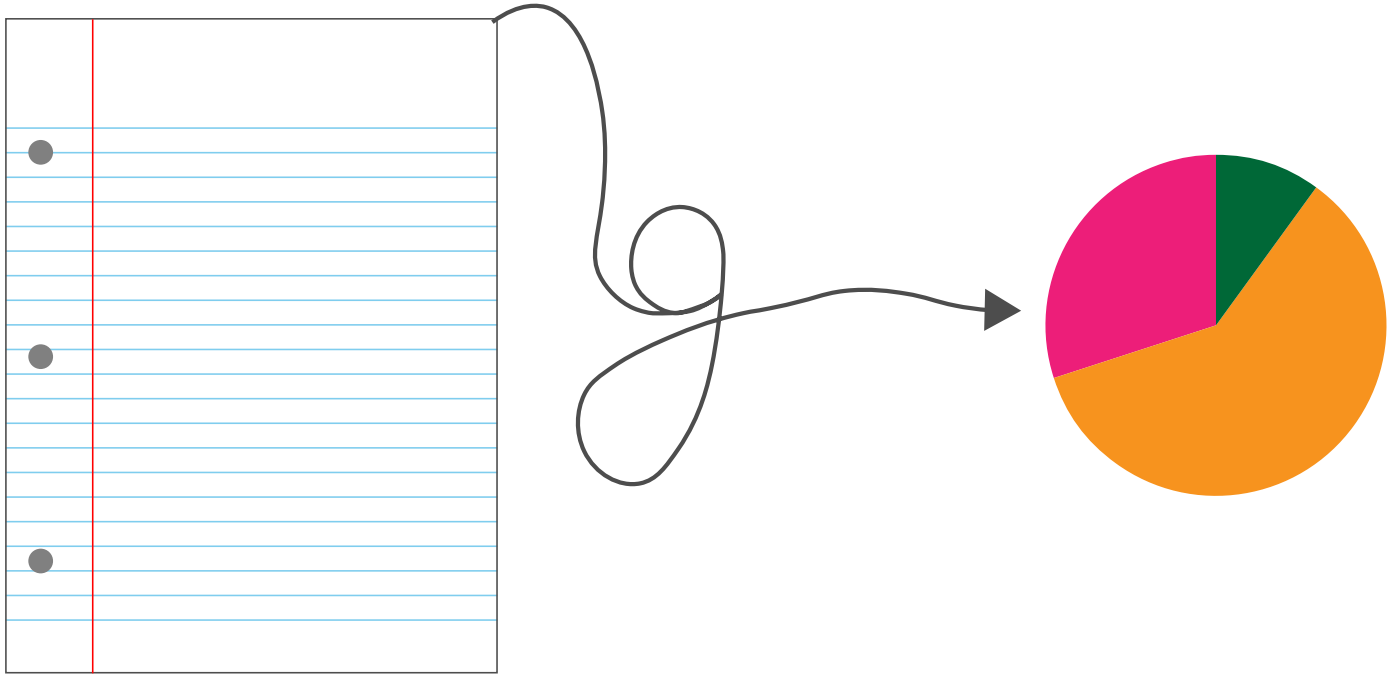
Structural topic models for open-ended survey responses
(with Roberts et al in *American Journal of Political Science*)



Applications: Causal Inference

Applications: Causal Inference

How to Make Causal Inferences Using Text
(with Egami, Fong, Grimmer and Roberts)



Application: "Adjusting for Confounding with Text Matching" (Roberts, Stewart and Nielsen)

Application: "Adjusting for Confounding with Text Matching" (Roberts, Stewart and Nielsen)

- Repression of speech in China $\rightsquigarrow$ internet censorship

Application: "Adjusting for Confounding with Text Matching" (Roberts, Stewart and Nielsen)

- Repression of speech in China $\rightsquigarrow$ internet censorship
- But, censoring the whole internet is hard
  (# of bloggers $\gg$ # of censors)

Application: "Adjusting for Confounding with Text Matching" (Roberts, Stewart and Nielsen)

- Repression of speech in China $\rightsquigarrow$ internet censorship
- But, censoring the whole internet is hard
  (# of bloggers $\gg$ # of censors)
- Limited external enforcement $\rightsquigarrow$ self-policing & targeting

Application: "Adjusting for Confounding with Text Matching" (Roberts, Stewart and Nielsen)

- Repression of speech in China $\rightsquigarrow$ internet censorship
- But, censoring the whole internet is hard
  (# of bloggers $\gg$ # of censors)
- Limited external enforcement $\rightsquigarrow$ self-policing & targeting
- People aren't censored at random $\rightsquigarrow$ causality is tricky

The Perfect Experiment

The Perfect Experiment

The Perfect Experiment

1. Be the Chinese government

The Perfect Experiment

1. Be the Chinese government
2. Randomly assign censorship

The Perfect Experiment

1. Be the Chinese government
2. Randomly assign censorship
3. See what bloggers write after censorship

The Perfect Experiment

1. Be the Chinese government
2. Randomly assign censorship
3. See what bloggers write after censorship

Problem 1: unethical

The Perfect Experiment

1. Be the Chinese government
2. Randomly assign censorship
3. See what bloggers write after censorship

Problem 1: unethical
Problem 2: we aren't the Chinese government

Approximating the Perfect Experiment

An approximation:

Find two bloggers

Approximating the Perfect Experiment

An approximation:

Find two bloggers

✓similar users,

Approximating the Perfect Experiment

An approximation:

Find two bloggers

✓similar users,
✓similar censorship histories,

Approximating the Perfect Experiment

An approximation:

Find two bloggers

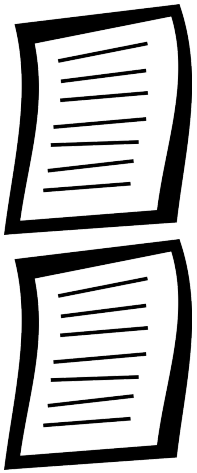
✓similar users,
✓similar censorship histories,
✓similar numbers of posts

Approximating the Perfect Experiment

An approximation:

Find two bloggers with similar posts

✓similar users,
✓similar censorship histories,
✓similar numbers of posts

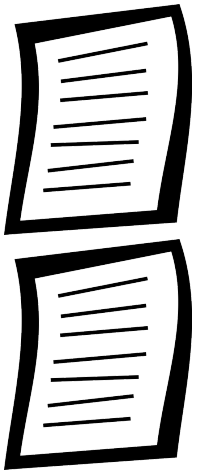


Approximating the Perfect Experiment

An approximation:

Find two bloggers with similar posts

✓similar users,
✓similar censorship histories,
✓similar numbers of posts



written on the same day

Approximating the Perfect Experiment

An approximation:

✓similar users,
✓similar censorship histories,
✓similar numbers of posts

Find two bloggers

Only one censored

Approximating the Perfect Experiment

An approximation:

✓similar users,
✓similar censorship histories,
✓similar numbers of posts

Find two bloggers

Only one censored

Censorship 'Mistake'

Approximating the Perfect Experiment

An approximation:

Find two bloggers

Only one censored

✓similar users,
✓similar censorship histories,
✓similar numbers of posts

Censorship 'Mistake'

Does the censored blogger's behavior change?

Text Matching

Text Matching

- Text as pre-treatment confounder

Text Matching

- Text as pre-treatment confounder $\rightsquigarrow$ a frequent problem

Text Matching

- Text as pre-treatment confounder $\rightsquigarrow$ a frequent problem
- Applications
  - Does censorship change a bloggers behavior?
  - Do targeted killings of extremists create interest in their work?
  - In Int. Relations, are women cited less frequently than men?

Text Matching

- Text as pre-treatment confounder $\rightsquigarrow$ a frequent problem
- Applications
  - Does censorship change a bloggers behavior?
  - Do targeted killings of extremists create interest in their work?
  - In Int. Relations, are women cited less frequently than men?
- We propose density estimate (from topic model) and propensity score (probability of censorship).

Text Matching

- Text as pre-treatment confounder $\rightsquigarrow$ a frequent problem
- Applications
  - Does censorship change a bloggers behavior?
  - Do targeted killings of extremists create interest in their work?
  - In Int. Relations, are women cited less frequently than men?
- We propose density estimate (from topic model) and propensity score (probability of censorship).
- Retain interpretability of matches for human evaluation.

How do bloggers react to censorship?

How do bloggers react to censorship?

- Data: 4,685 social media users on Weibo (Fu, Chan and Chau, 2013)

How do bloggers react to censorship?

- Data: 4,685 social media users on Weibo (Fu, Chan and Chau, 2013)
- Match on density estimate, propensity score and also: date, previous post rate, prior censorship experience

How do bloggers react to censorship?

- Data: 4,685 social media users on Weibo (Fu, Chan and Chau, 2013)
- Match on density estimate, propensity score and also: date, previous post rate, prior censorship experience
- What is the effect of censorship?

How do bloggers react to censorship?

- Data: 4,685 social media users on Weibo (Fu, Chan and Chau, 2013)
- Match on density estimate, propensity score and also: date, previous post rate, prior censorship experience
- What is the effect of censorship?

Censored users have almost double the censorship rate but still post at the same rate.
Consistent with algorithmic targeting or increased posting of sensitive content.

Conclusions

Conclusions

- Focus on tasks can be helpful in building applications.

Conclusions

- Focus on tasks can be helpful in building applications.
- Three core tasks: discovery, measurement and causal inference.

Conclusions

- Focus on tasks can be helpful in building applications.
- Three core tasks: discovery, measurement and causal inference.
- Four steps in building a method:

Conclusions

- Focus on tasks can be helpful in building applications.
- Three core tasks: discovery, measurement and causal inference.
- Four steps in building a method:
  - propose a theory of use
  - customize the method
  - design the workflow
  - create validations

Conclusions

- Focus on tasks can be helpful in building applications.
- Three core tasks: discovery, measurement and causal inference.
- Four steps in building a method:
  - propose a theory of use
  - customize the method
  - design the workflow
  - create validations
- Future directions in new kinds of data and new forms of supervision

For more information

brandonstewart.org

structuraltopicmodel.com