Training Session: Topic models

Pablo Barberá
Networked Democracy Lab
University of Southern California

May 2, 2017

materials: github.com/NetDem-USC/training

Overview of text as data methods

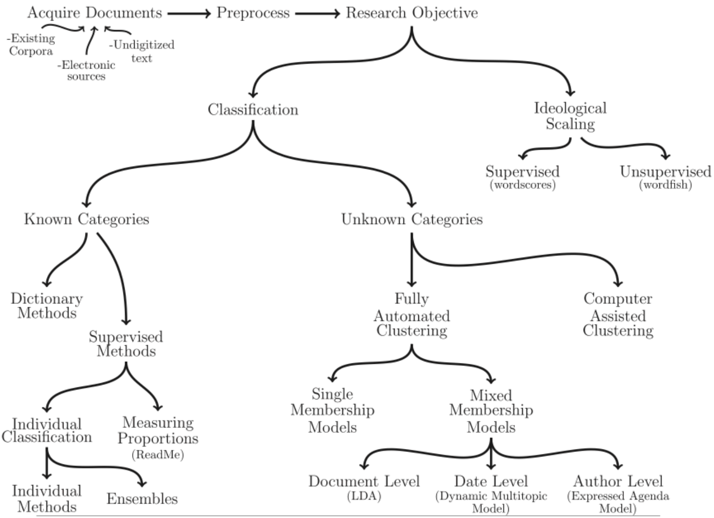


Fig. 1 in Grimmer and Stewart (2013)

Overview of text as data methods

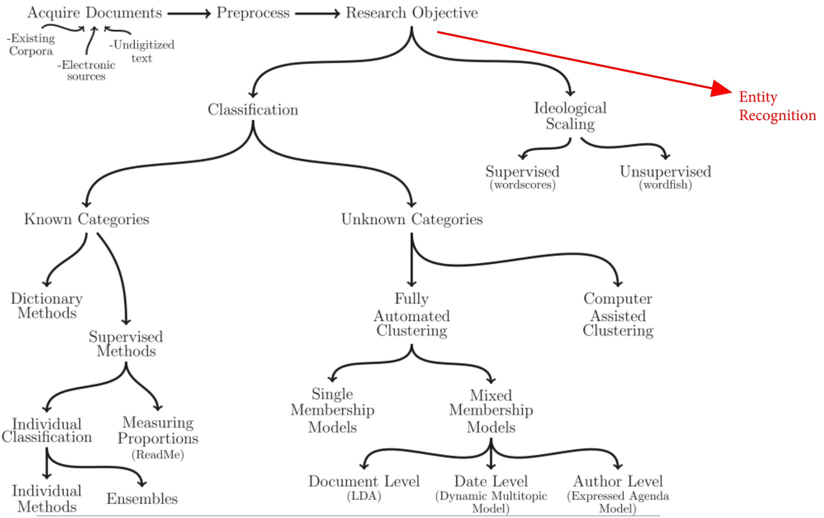


Fig. 1 in Grimmer and Stewart (2013)

Overview of text as data methods

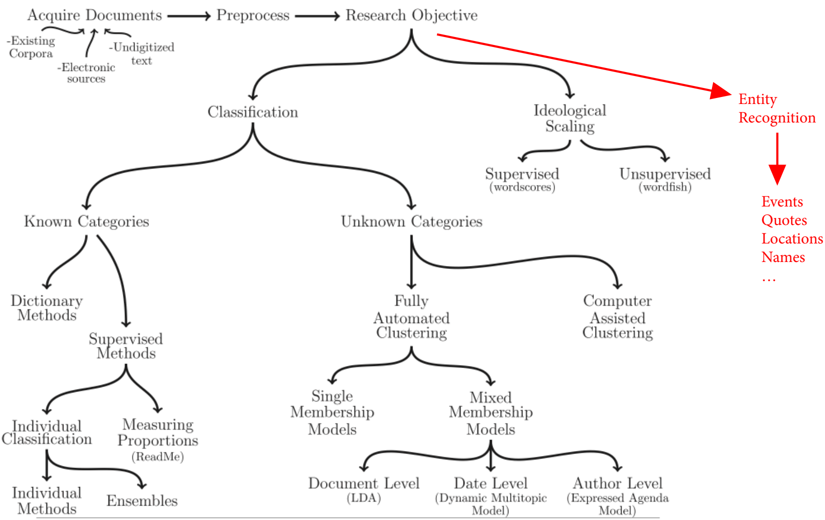


Fig. 1 in Grimmer and Stewart (2013)

Overview of text as data methods

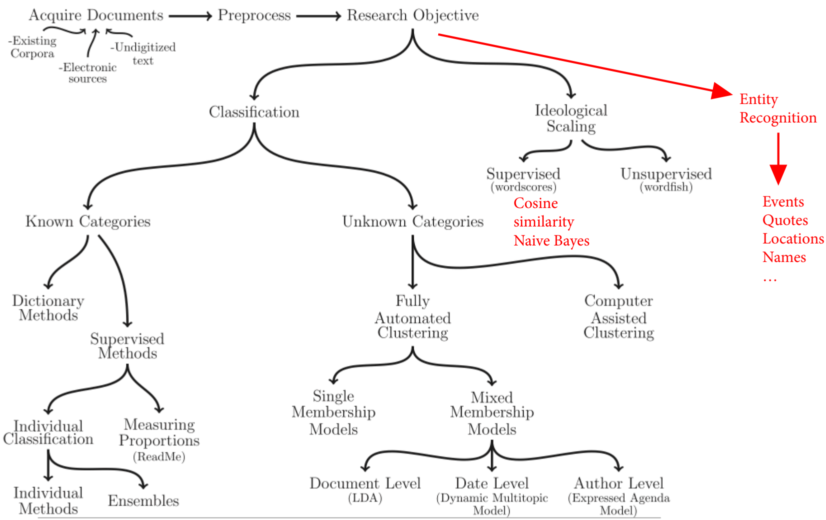


Fig. 1 in Grimmer and Stewart (2013)

Overview of text as data methods

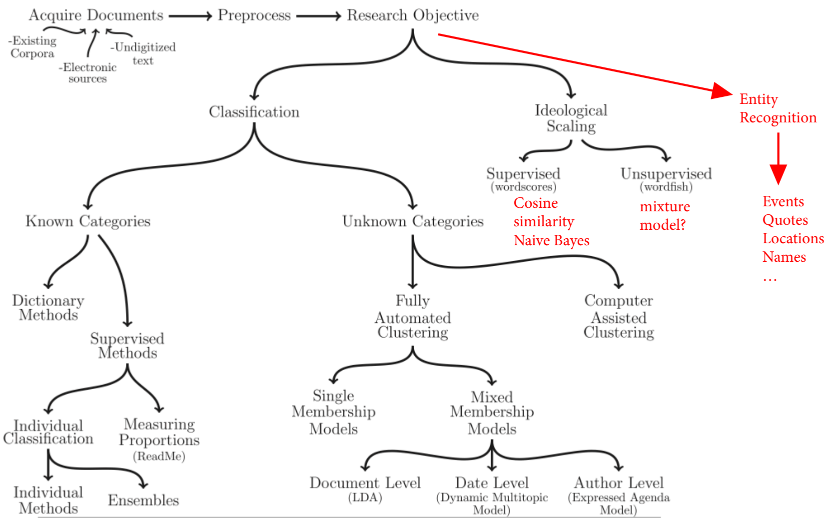


Fig. 1 in Grimmer and Stewart (2013)

Overview of text as data methods

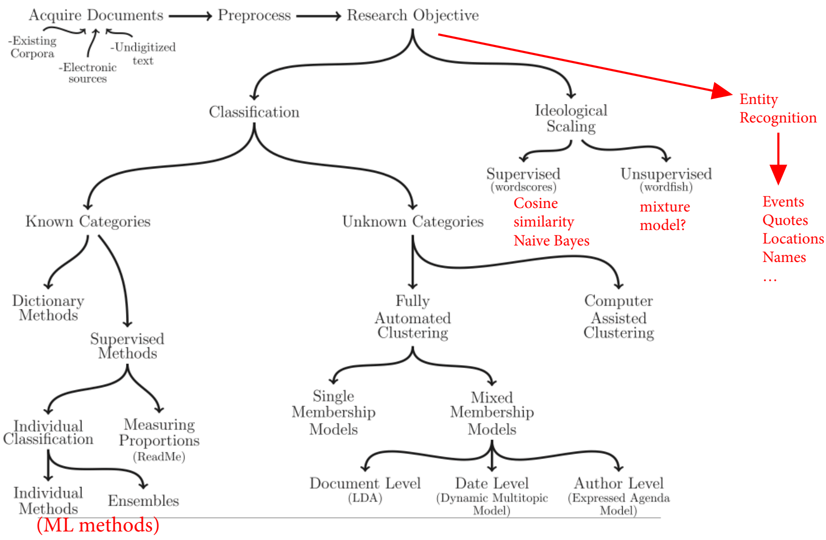


Fig. 1 in Grimmer and Stewart (2013)

Overview of text as data methods

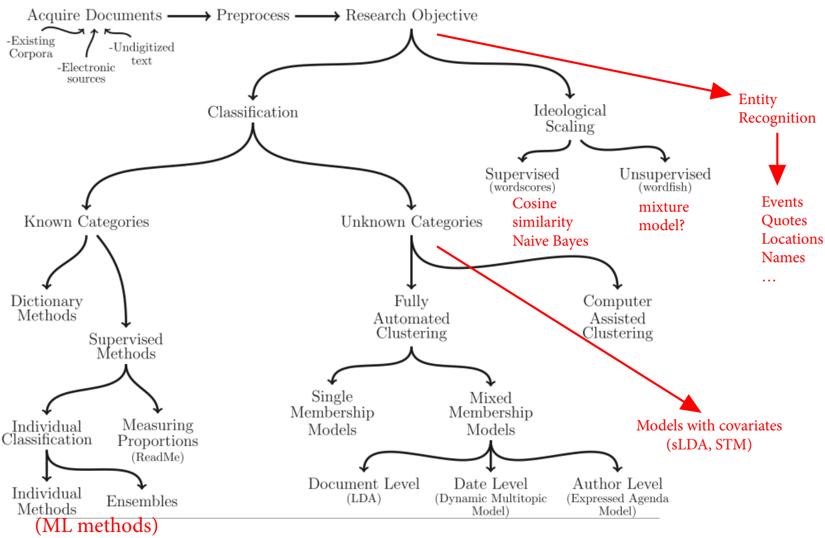
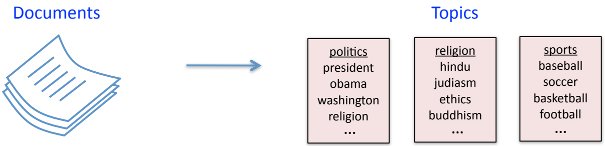


Fig. 1 in Grimmer and Stewart (2013)

Latent Dirichlet allocation (LDA)

- **Topic models** are powerful tools for exploring large data sets and for making inferences about the content of documents



Latent Dirichlet allocation (LDA)

- **Topic models** are powerful tools for exploring large data sets and for making inferences about the content of documents

Documents

Topics

politics	religion	sports
president	hindu	baseball
obama	judiasm	soccer
washington	ethics	basketball
religion	buddhism	football
...	...	...

- Many applications in information retrieval, document summarization, and classification

New document

Words $w_1, \ldots, w_N$

What is this document about?

weather	.50
finance	.49
sports	.01

Distribution of topics θ

Latent Dirichlet allocation (LDA)

- **Topic models** are powerful tools for exploring large data sets and for making inferences about the content of documents

Documents

Topics

politics	religion	sports
president	hindu	baseball
obama	judiasm	soccer
washington	ethics	basketball
religion	buddhism	football
…	…	…

- Many applications in information retrieval, document summarization, and classification

New document

Words $w_1, \ldots, w_N$

What is this document about?

weather	.50
finance	.49
sports	.01

Distribution of topics θ

- LDA is one of the simplest and most widely used topic models

Latent Dirichlet Allocation

- Document = random mixture over latent topics
- Topic = distribution over n-grams

Probabilistic model with 3 steps:

1. Choose $\theta_i \sim \text{Dirichlet}(\alpha)$
2. Choose $\beta_k \sim \text{Dirichlet}(\delta)$
3. For each word in document i:
    - Choose a topic $z_m \sim \text{Multinomial}(\theta_i)$
    - Choose a word $w_{im} \sim \text{Multinomial}(\beta_{i,k=z_m})$

where:
α=parameter of Dirichlet prior on distribution of topics over docs.
θ_i=topic distribution for document i
δ=parameter of Dirichlet prior on distribution of words over topics
β_k=word distribution for topic k

Plate notation

Plate notation

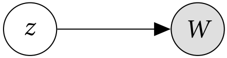
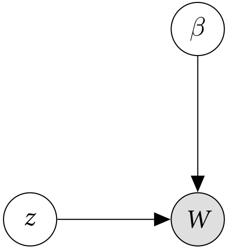
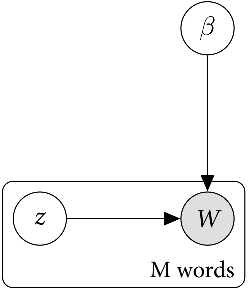


Plate notation



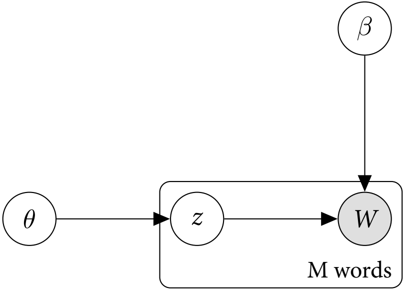
$\beta = M \times K$ matrix where β_{im} indicates prob(topic=k) for word m

Plate notation



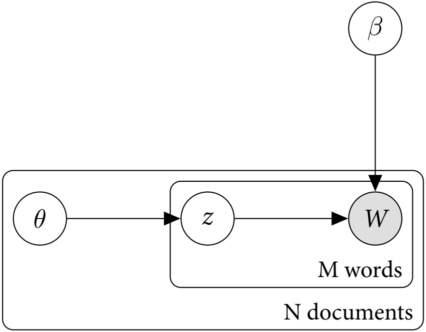
$\beta = M \times K$ matrix where β_{im} indicates prob(topic=k) for word m

Plate notation



$\beta = M \times K$ matrix where β_{im} indicates prob(topic=k) for word m
$\theta = N \times K$ matrix where θ_{ik} indicates prob(topic=k) for document i

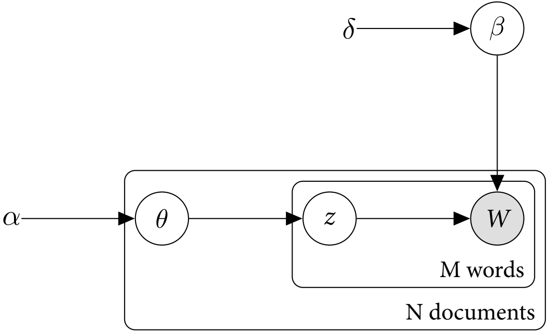
Plate notation



$\beta = M \times K$ matrix where β_{im} indicates prob(topic=k) for word m

$\theta = N \times K$ matrix where θ_{ik} indicates prob(topic=k) for document i

Plate notation



$\beta = M \times K$ matrix where β_{im} indicates prob(topic=k) for word m

$\theta = N \times K$ matrix where θ_{ik} indicates prob(topic=k) for document i

Validation

From Quinn et al, AJPS, 2010:

1. Semantic validity

Validation

From Quinn et al, AJPS, 2010:

1. Semantic validity
   - Do the topics identify coherent groups of tweets that are internally homogenous, and are related to each other in a meaningful way?

Validation

From Quinn et al, AJPS, 2010:

1. Semantic validity
   - Do the topics identify coherent groups of tweets that are internally homogenous, and are related to each other in a meaningful way?
2. Convergent/discriminant construct validity

Validation

From Quinn et al, AJPS, 2010:

1. Semantic validity
   - Do the topics identify coherent groups of tweets that are internally homogenous, and are related to each other in a meaningful way?
2. Convergent/discriminant construct validity
   - Do the topics match existing measures where they should match?

Validation

From Quinn et al, AJPS, 2010:

1. Semantic validity
   - Do the topics identify coherent groups of tweets that are internally homogenous, and are related to each other in a meaningful way?
2. Convergent/discriminant construct validity
   - Do the topics match existing measures where they should match?
   - Do they depart from existing measures where they should depart?

Validation

From Quinn et al, AJPS, 2010:

1. Semantic validity
   - Do the topics identify coherent groups of tweets that are internally homogenous, and are related to each other in a meaningful way?
2. Convergent/discriminant construct validity
   - Do the topics match existing measures where they should match?
   - Do they depart from existing measures where they should depart?
3. Predictive validity

Validation

From Quinn et al, AJPS, 2010:

1. Semantic validity
   - Do the topics identify coherent groups of tweets that are internally homogenous, and are related to each other in a meaningful way?
2. Convergent/discriminant construct validity
   - Do the topics match existing measures where they should match?
   - Do they depart from existing measures where they should depart?
3. Predictive validity
   - Does variation in topic usage correspond with expected events?

Validation

From Quinn et al, AJPS, 2010:

1. Semantic validity
   - Do the topics identify coherent groups of tweets that are internally homogenous, and are related to each other in a meaningful way?
2. Convergent/discriminant construct validity
   - Do the topics match existing measures where they should match?
   - Do they depart from existing measures where they should depart?
3. Predictive validity
   - Does variation in topic usage correspond with expected events?
4. Hypothesis validity

Validation

From Quinn et al, AJPS, 2010:

1. Semantic validity
   - Do the topics identify coherent groups of tweets that are internally homogenous, and are related to each other in a meaningful way?
2. Convergent/discriminant construct validity
   - Do the topics match existing measures where they should match?
   - Do they depart from existing measures where they should depart?
3. Predictive validity
   - Does variation in topic usage correspond with expected events?
4. Hypothesis validity
   - Can topic variation be used effectively to test substantive hypotheses?

Example: open-ended survey responses

Bauer, Barberá *et al*, *Political Behavior*, 2016.

- Data: General Social Survey (2008) in Germany
- Responses to questions: *Would you please tell me what you associate with the term "left"?and would you please tell me what you associate with the term "right"?*
- Open-ended questions minimize priming and potential interviewer effects
- Sparse Additive Generative model instead of LDA (more coherent topics for short text)
- $K = 4$ topics for each question

Example: open-ended survey responses

Table 1: Top scoring words associated with each topic, and English translations)

Left topic 1: **Parties** (proportion = .26, average lr-scale value = 5.38) linke, spd, partei, linken, pds, politik, kommunisten, parteien, grünen, punks *the left, spd, party, the left, pds, politics, communists, parties, greens, punks*
Left topic 2: **Ideologies** (proportion = .26, average lr-scale value = 5.36) kommunismus, links, sozialismus, lafontaine, rechts, aber, gysi, linkspartei, richtung, gleichmacherei *communism, left, socialism, lafontaine, right, but, gysi, left party, direction, levelling*
Left topic 3: **Values** (proportion = .24, average lr-scale value = 4.06) soziale, gerechtigkeit, demokratie, soziales, bürger, gleichheit, gleiche, freiheit, rechte, gleichberechtigung *social, justice, democracy, social, citizen, equality, equal, freedom, rights, equal rights*
Left topic 4: **Policies** (proportion = .24, average lr-scale value =4.89) sozial, menschen, leute, ddr, verbinde, kleinen, einstellung, umverteilung, sozialen, vertreten *social, humans, people, ddr, associate, the little, attitude, redistribution, social, represent*

Right topic 1: **Ideologies** (proportion = .27, average lr-scale value = 5.00) konservativ, nationalsozialismus, rechtsradikal, radikal, ordnung, politik, nazi, recht, menschen, konservative *conservative, national socialism, right-wing radicalism, radical, order, politics, nazi, right, people, conservatives*
Right topic 2: **Parties** (proportion = .25, average lr-scale value = 5.26) npd, rechts, cdu, csu, rechten, parteien, leute, aber, verbinde, rechtsradikalen *npd, right, cdu, csu, the right, parties, people, but, associate, right-wing radicalists*
Right topic 3: **Xenophobia** (proportion = .25, average lr-scale value = 4.55) ausländerfeindlichkeit, gewalt, ausländer, demokratie, nationalismus, rechtsradikalismus, diktatur, national, intoleranz, faschismus *xenophobia, violence, foreigners, democracy, nationalism, right-wing radicalism, dictatorship, national, intolerance, fascism*
Right topic 4: **Right-wing extremists** (proportion = .23, average lr-scale value = 4.90) nazis, neonazis, rechtsradikale, rechte, radikale, radikalismus, partei, ausländerfeindlich, reich, nationale *nazis, neonazis, right-wing radicalists, rightists, radicals, radicalism, party, xenophobia, rich, national*

Note: "proportion" indicates the average estimated probability that any given response is assigned to a topic. "average lr-scale value" is the mean position on the left-right scale (from 0 to 10) of individuals whose highest probability belongs to that particular topic.

Bauer, Barberá *et al*, *Political Behavior*, 2016.

Example: open-ended survey responses

Fig. 6: Left-right scale means for different subsamples of associations with **left** (dashed = sample mean, bars = 95% Cis)

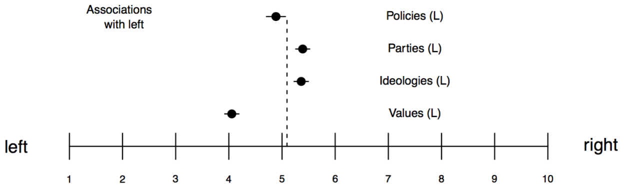
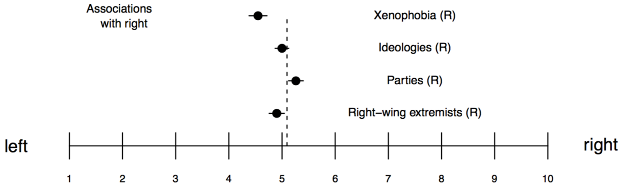


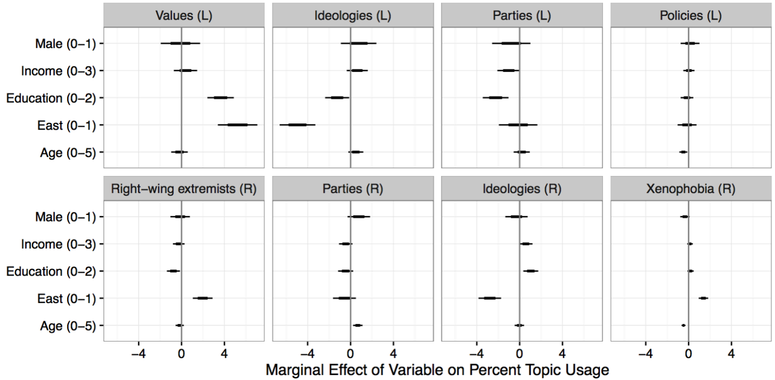
Fig. 7: Left-right scale means for different subsamples of associations with **right** (dashed = sample mean, bars = 95% Cis)



Bauer, Barberá *et al*, *Political Behavior*, 2016.

Example: open-ended survey responses

Fig. 9: Systematic relationship between associations with "left" and "right" and characteristics of respondents



Note: Each line indicates a 95% confidence interval (and 66% confidence interval in darker color) for the coefficient of eight different regressions of topic usage (in a scale from 0 to 100) at the respondent level on seven individual-level characteristics. The line on the bottom right corner (second row, second plot), for example, shows that individual a one-category change in age is associated with around one percentage point increase in the probability that the individual associated "right" with political parties.

Bauer, Barberá *et al*, *Political Behavior*, 2016.

Example: topics in US legislators' tweets

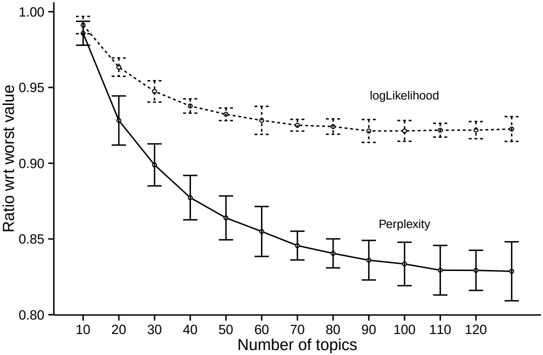
- Data: 651,116 tweets sent by US legislators from January 2013 to December 2014.
- 2,920 documents = 730 days $\times$ 2 chambers $\times$ 2 parties
- Why aggregating? Applications that aggregate by author or day outperform tweet-level analyses (Hong and Davidson, 2010)
- $K = 100$ topics (more on this later)
- Validation: http://j.mp/lda-congress-demo

Choosing the number of topics

- Choosing K is "one of the most difficult questions in unsupervised learning" (Grimmer and Stewart, 2013, p.19)

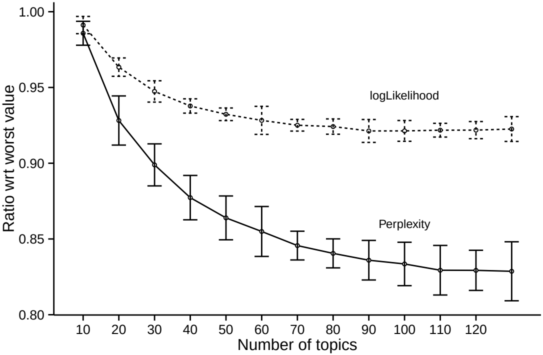
Choosing the number of topics

- Choosing K is "one of the most difficult questions in unsupervised learning" (Grimmer and Stewart, 2013, p.19)
- We chose $K = 100$ based on cross-validated model fit.



Choosing the number of topics

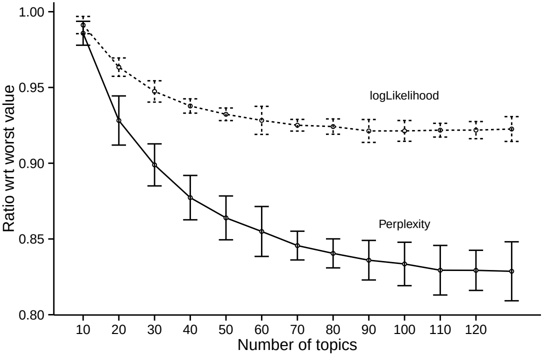
- Choosing K is “one of the most difficult questions in unsupervised learning” (Grimmer and Stewart, 2013, p.19)
- We chose $K = 100$ based on cross-validated model fit.



- **BUT**: “there is often a negative relationship between the best-fitting model and the substantive information provided”.

Choosing the number of topics

- Choosing K is "one of the most difficult questions in unsupervised learning" (Grimmer and Stewart, 2013, p.19)
- We chose $K = 100$ based on cross-validated model fit.



- **BUT**: "there is often a negative relationship between the best-fitting model and the substantive information provided".
- GS propose to choose K based on "substantive fit."

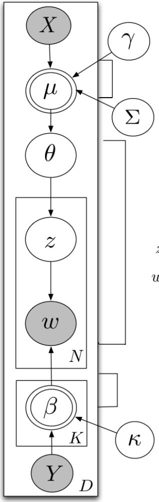
Extensions of LDA

1. Structural topic model (Roberts et al, 2014, AJPS)
2. Dynamic topic model (Blei and Lafferty, 2006, ICML; Quinn et al, 2010, AJPS)
3. Hierarchical topic model (Griffiths and Tenembaun, 2004, NIPS; Grimmer, 2010, PA)

Why?

- Substantive reasons: incorporate specific elements of DGP into estimation
- Statistical reasons: structure can lead to better topics.

Structural topic model



Topic Prevalence:

$$\mu_{d,k} = X_d \gamma_k$$
$$\gamma_k \sim \mathcal{N}(0, \sigma_k^2)$$
$$\sigma_k^2 \sim \text{Gamma}(s^\gamma, r^\gamma)$$

Language Model:

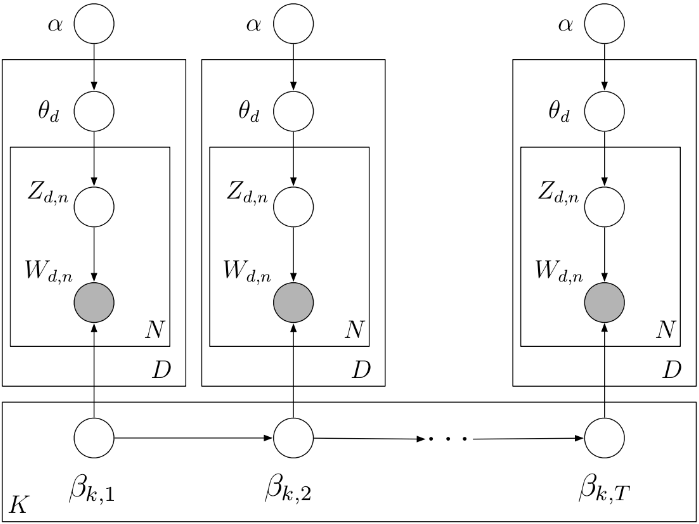
$$\theta_d \sim \text{LogisticNormal}(\mu_d, \Sigma)$$
$$z_{d,n} \sim \text{Mult}(\theta_d)$$
$$w_{d,n} \sim \text{Mult}(\beta_d^{k=z_{d,n}})$$

Topical Content:

$$\beta_{d,v}^k \propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k})$$
$$\kappa_v^{y,k} \sim \text{Laplace}(0, \tau_v^{y,k})$$
$$\tau_v^{y,k} \sim \text{Gamma}(s^\kappa, r^\kappa)$$

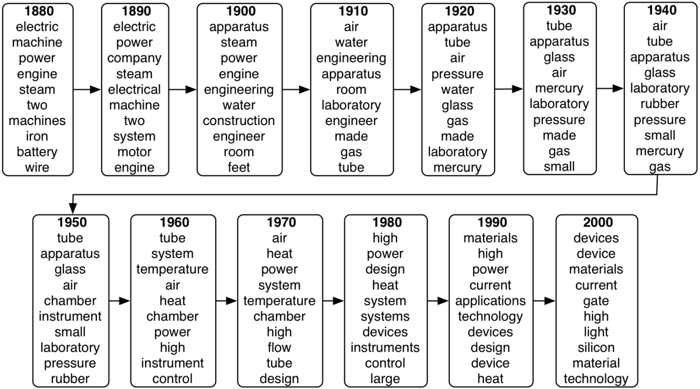
- **Prevalence**: Prior on the mixture over topics is now document-specific, and can be a function of covariates (documents with similar covariates will tend to be about the same topics)
- **Content**: distribution over words is now document-specific and can be a function of covariates (documents with similar covariates will tend to use similar words to refer to the same topic)

Dynamic topic model



Source: Blei, "Modeling Science"

Dynamic topic model



Source: Blei, "Modeling Science"

Wordfish (Slapin and Proksch, 2008, AJPS)

- Goal: unsupervised scaling of ideological positions

Wordfish (Slapin and Proksch, 2008, AJPS)

- Goal: unsupervised scaling of ideological positions
- Ideology of politician i, θ_i is a position in a latent scale.

Wordfish (Slapin and Proksch, 2008, AJPS)

- Goal: unsupervised scaling of ideological positions
- Ideology of politician i, θ_i is a position in a latent scale.
- Word usage is drawn from a Poisson-IRT model:

$$
\begin{aligned}
W_{im} &\sim \text{Poisson}(\lambda_{im}) \\
\lambda_{im} &= exp(\alpha_i + \psi_m + \beta_m \times \theta_i)
\end{aligned}
$$

Wordfish (Slapin and Proksch, 2008, AJPS)

- Goal: unsupervised scaling of ideological positions
- Ideology of politician i, θ_i is a position in a latent scale.
- Word usage is drawn from a Poisson-IRT model:

$$W_{im} \sim \text{Poisson}(\lambda_{im})$$
$$\lambda_{im} = exp(\alpha_i + \psi_m + \beta_m \times \theta_i)$$

- where:

Wordfish (Slapin and Proksch, 2008, AJPS)

- Goal: unsupervised scaling of ideological positions
- Ideology of politician i, θ_i is a position in a latent scale.
- Word usage is drawn from a Poisson-IRT model:

$$
\begin{aligned}
W_{im} &\sim \text{Poisson}(\lambda_{im}) \\
\lambda_{im} &= exp(\alpha_i + \psi_m + \beta_m \times \theta_i)
\end{aligned}
$$

- where:

  α_i is "loquaciousness" of politician i

Wordfish (Slapin and Proksch, 2008, AJPS)

- Goal: unsupervised scaling of ideological positions
- Ideology of politician i, θ_i is a position in a latent scale.
- Word usage is drawn from a Poisson-IRT model:

$$
\begin{aligned}
W_{im} &\sim \text{Poisson}(\lambda_{im}) \\
\lambda_{im} &= exp(\alpha_i + \psi_m + \beta_m \times \theta_i)
\end{aligned}
$$

- where:

  α_i is "loquaciousness" of politician i
  ψ_m is frequency of word m

Wordfish (Slapin and Proksch, 2008, AJPS)

- Goal: unsupervised scaling of ideological positions
- Ideology of politician i, θ_i is a position in a latent scale.
- Word usage is drawn from a Poisson-IRT model:

$$W_{im} \sim \text{Poisson}(\lambda_{im})$$
$$\lambda_{im} = exp(\alpha_i + \psi_m + \beta_m \times \theta_i)$$

- where:

  α_i is "loquaciousness" of politician i
  ψ_m is frequency of word m
  β_m is discrimination parameter of word m

Wordfish (Slapin and Proksch, 2008, AJPS)

- Goal: unsupervised scaling of ideological positions
- Ideology of politician i, θ_i is a position in a latent scale.
- Word usage is drawn from a Poisson-IRT model:

$$
\begin{aligned}
W_{im} &\sim \text{Poisson}(\lambda_{im}) \\
\lambda_{im} &= exp(\alpha_i + \psi_m + \beta_m \times \theta_i)
\end{aligned}
$$

- where:

  α_i is "loquaciousness" of politician i
  ψ_m is frequency of word m
  β_m is discrimination parameter of word m

- Estimation using EM algorithm.

Wordfish (Slapin and Proksch, 2008, AJPS)

- Goal: unsupervised scaling of ideological positions
- Ideology of politician i, θ_i is a position in a latent scale.
- Word usage is drawn from a Poisson-IRT model:

$$W_{im} \sim \text{Poisson}(\lambda_{im})$$
$$\lambda_{im} = exp(\alpha_i + \psi_m + \beta_m \times \theta_i)$$

- where:

  α_i is "loquaciousness" of politician i
  ψ_m is frequency of word m
  β_m is discrimination parameter of word m

- Estimation using EM algorithm.
- Identification:

Wordfish (Slapin and Proksch, 2008, AJPS)

- Goal: unsupervised scaling of ideological positions
- Ideology of politician i, θ_i is a position in a latent scale.
- Word usage is drawn from a Poisson-IRT model:

$$W_{im} \sim \text{Poisson}(\lambda_{im})$$
$$\lambda_{im} = exp(\alpha_i + \psi_m + \beta_m \times \theta_i)$$

- where:

  α_i is "loquaciousness" of politician i
  ψ_m is frequency of word m
  β_m is discrimination parameter of word m

- Estimation using EM algorithm.
- Identification:
  - Unit variance restriction for θ_i

Wordfish (Slapin and Proksch, 2008, AJPS)

- Goal: unsupervised scaling of ideological positions
- Ideology of politician i, θ_i is a position in a latent scale.
- Word usage is drawn from a Poisson-IRT model:

$$W_{im} \sim \text{Poisson}(\lambda_{im})$$
$$\lambda_{im} = exp(\alpha_i + \psi_m + \beta_m \times \theta_i)$$

- where:

  α_i is "loquaciousness" of politician i
  ψ_m is frequency of word m
  β_m is discrimination parameter of word m

- Estimation using EM algorithm.
- Identification:
  - Unit variance restriction for θ_i
  - Choose a and b such that $\theta_a > \theta_b$