Quantitative text analysis: Topic Models

Blake Miller

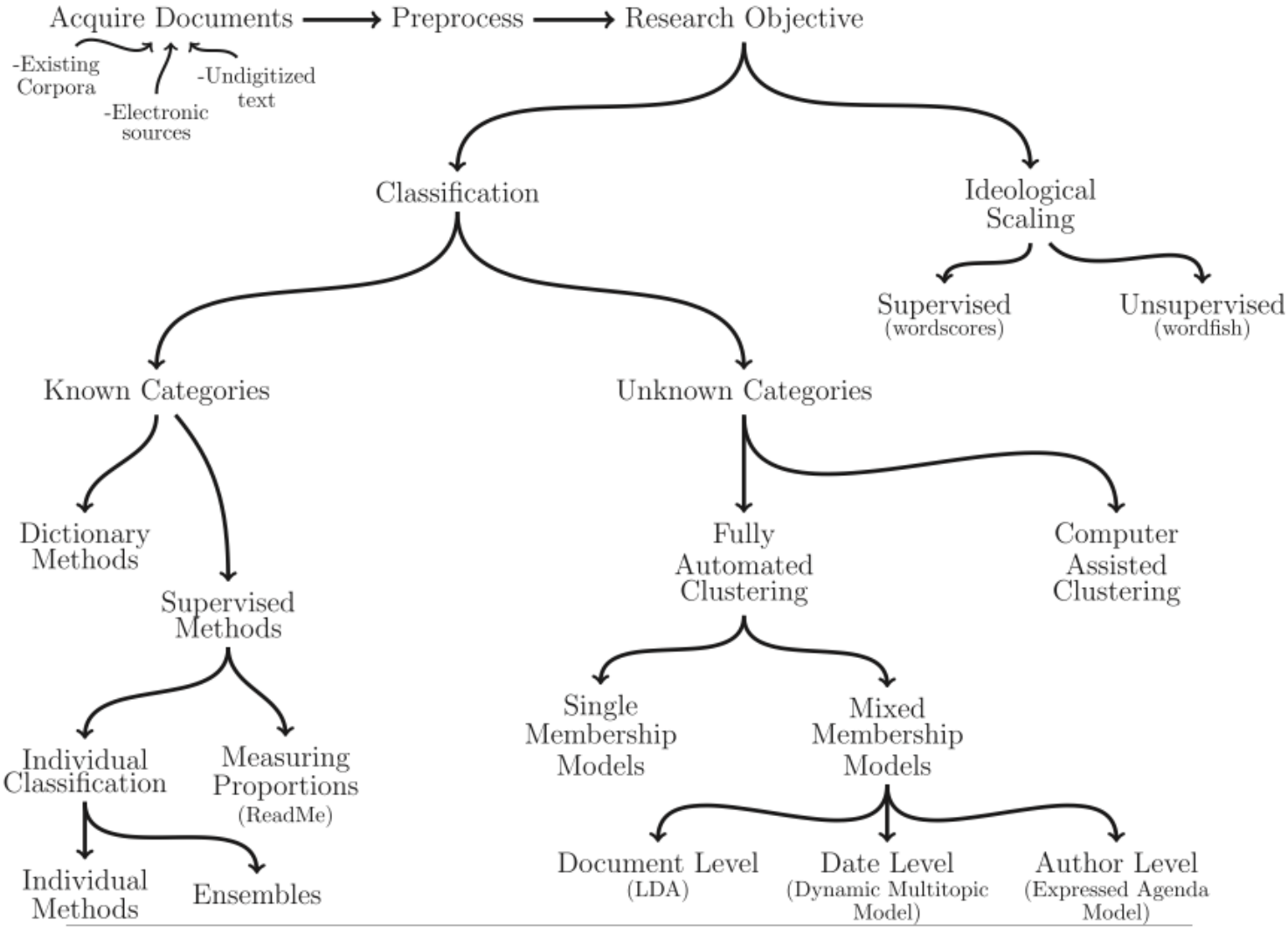
MY 459: Quantitative Text Analysis

March 5, 2023

Course website: lse-my459.github.io

1. Overview and Fundamentals
2. Descriptive Statistical Methods for Text Analysis
3. Automated Dictionary Methods
4. Machine Learning for Texts
5. Supervised Scaling Models for Texts
6. *Reading Week*
7. Unsupervised Models for Scaling Texts
8. Similarity and Clustering Methods
9. Topic models
10. Word embeddings/guided coding
11. Current topics in QTA/NLP (if no industrial action)

Overview of text as data methods



Outline

- Overview of topic models
- Latent Dirichlet Allocation (LDA)
- Examples
- Model Selection and Validation
- Extensions of LDA
- Implementations in R
- Guided coding

Topic Models

- Topic models are algorithms for discovering the main "themes" in an unstructured corpus
- Can be used to organize the collection according to the discovered themes
- Requires no prior information, training set, or human annotation – only a decision on K (number of topics)
- Most common: Latent Dirichlet Allocation (LDA) – Bayesian mixture model for discrete data where topics are assumed to be uncorrelated
- LDA provides a generative model that describes how the documents in a dataset were created
  - Each of the K *topics* is a distribution over a fixed vocabulary
  - Each document is a collection of words, generated according to a multinomial distribution, one for each of K topics

Latent Dirichlet Allocation

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

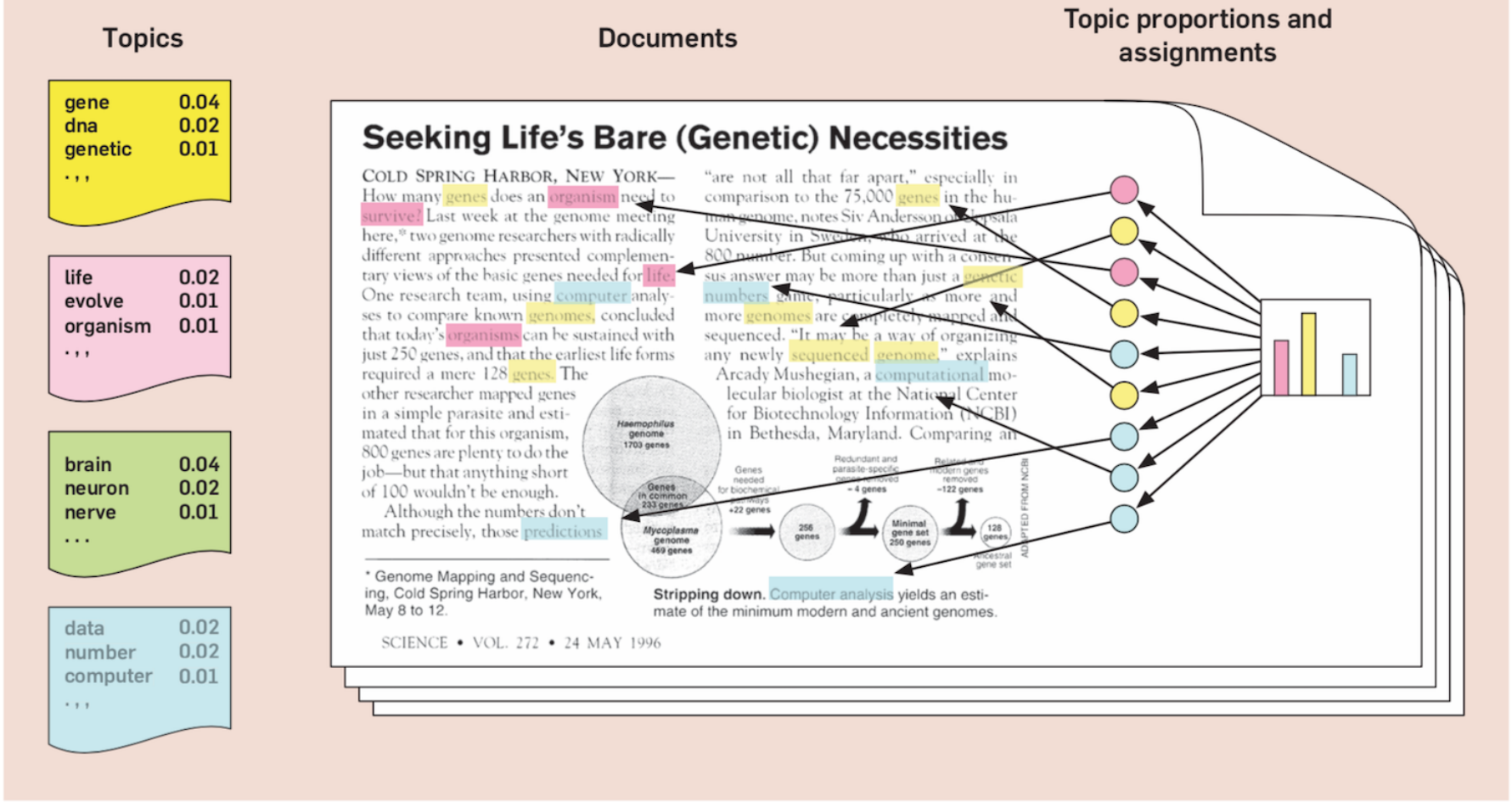


Illustration of the LDA generative process

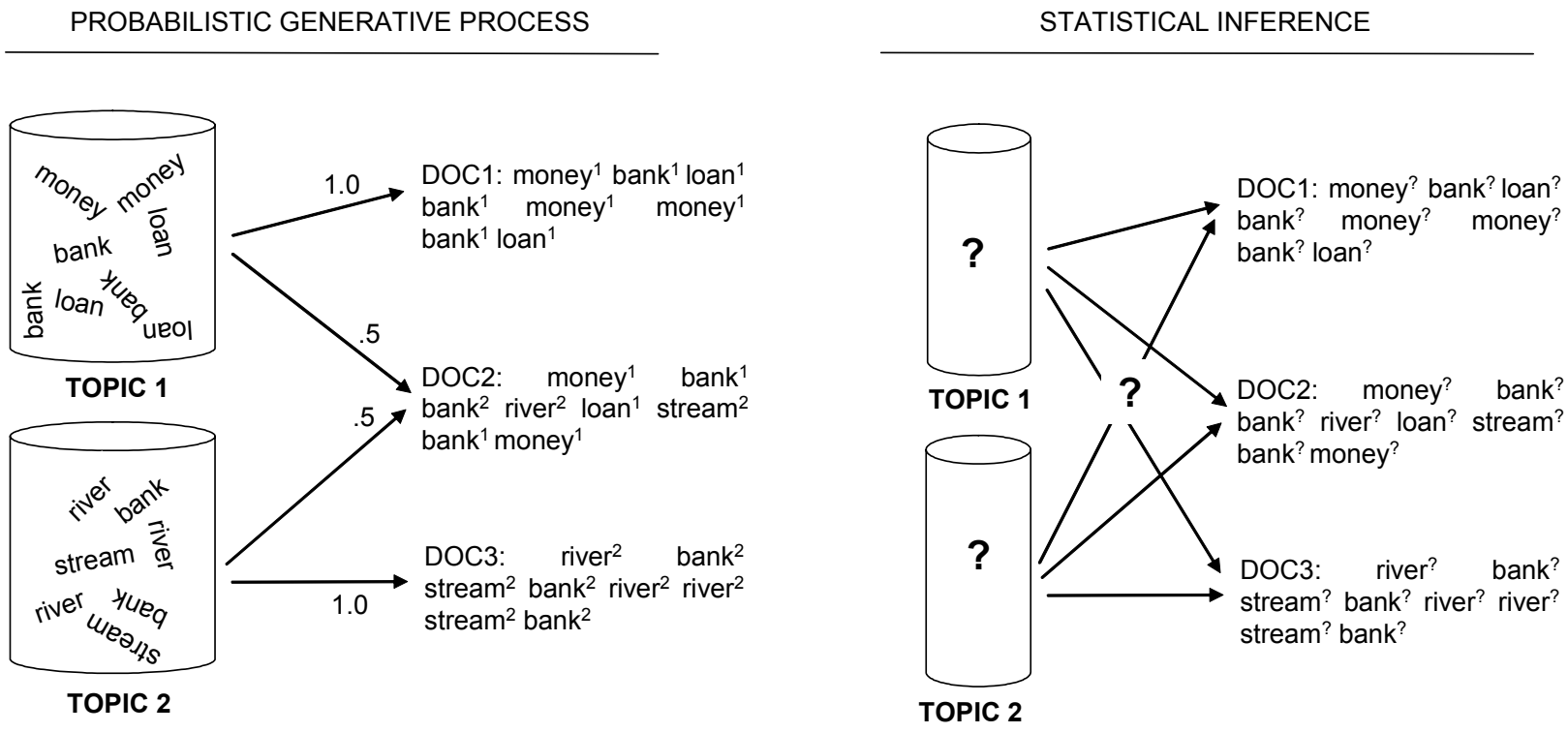


Figure 2. Illustration of the generative process and the problem of statistical inference underlying topic models

(from Steyvers and Griffiths 2007)

Topics example

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Figure 1. An illustration of four (out of 300) topics extracted from the TASA corpus.

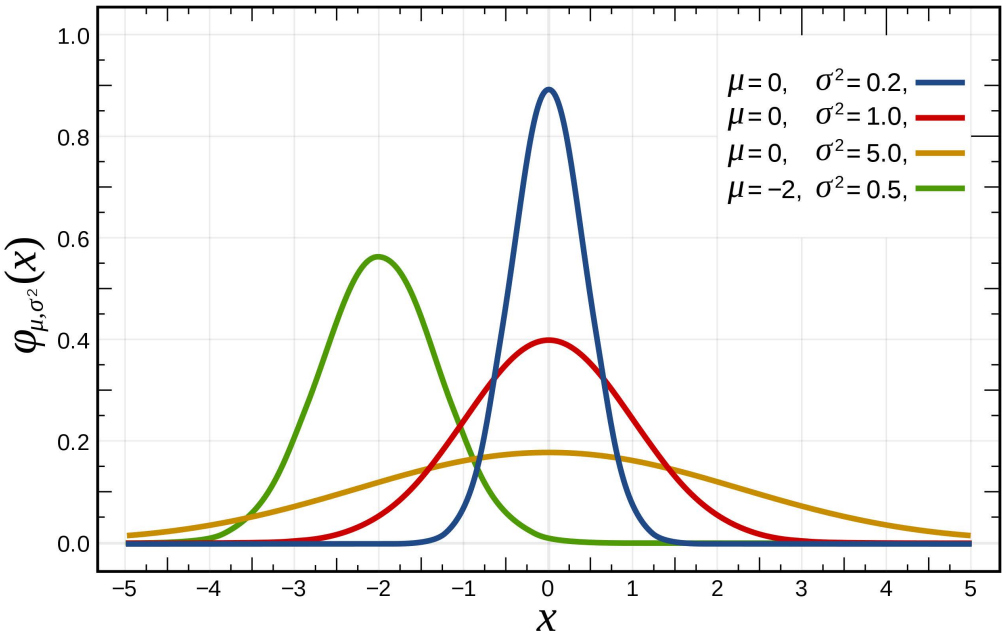
(from Steyvers and Griffiths 2007)

Often K is quite large!

Outline

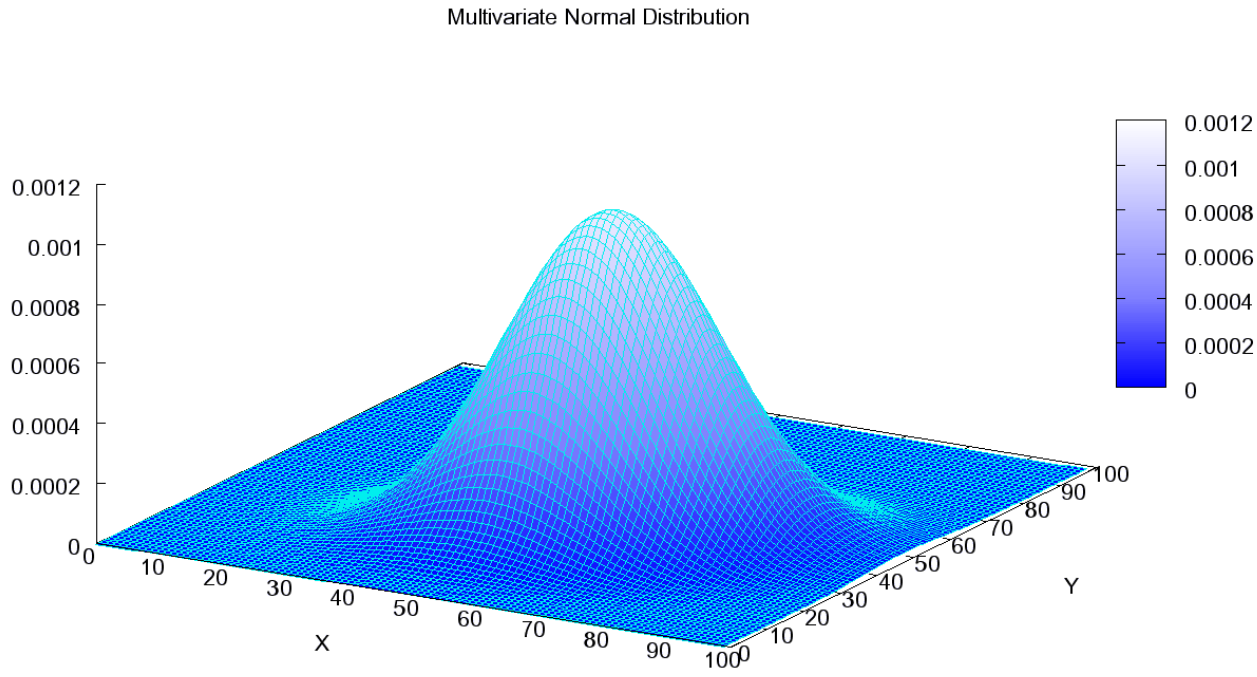
- Overview of topic models
- Latent Dirichlet Allocation (LDA)
- Examples
- Model Selection and Validation
- Extensions of LDA
- Implementations in R
- Guided coding

Review: Univariate probability density function



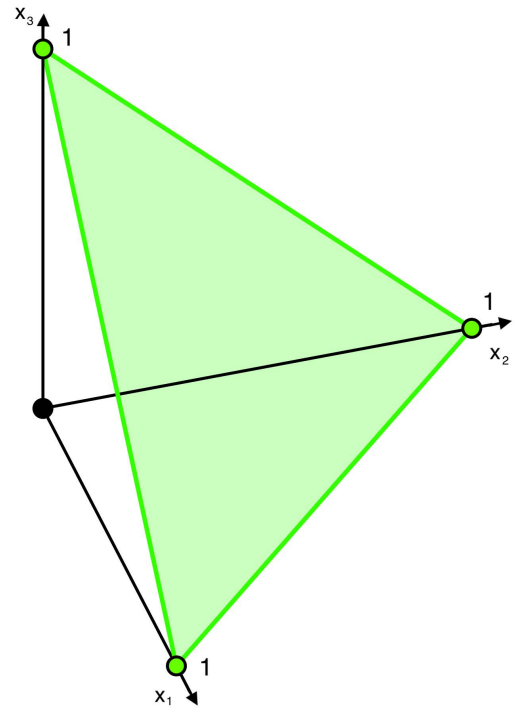
- Different parameter values (e.g. μ and σ for the normal distribution) change the distributions' shape
- The notation "$x \sim N(0, 1)$" denotes to sample or draw "x" from a standard normal distribution. This draw could e.g. return $x = -1.124$

Review: Multivariate probability density function



A draw $a \sim N(\mu, \Sigma)$ from this multivariate normal distribution could e.g. return $a = (-0.12, 1.2)$

Review: Graphical intuition of a standard simplex

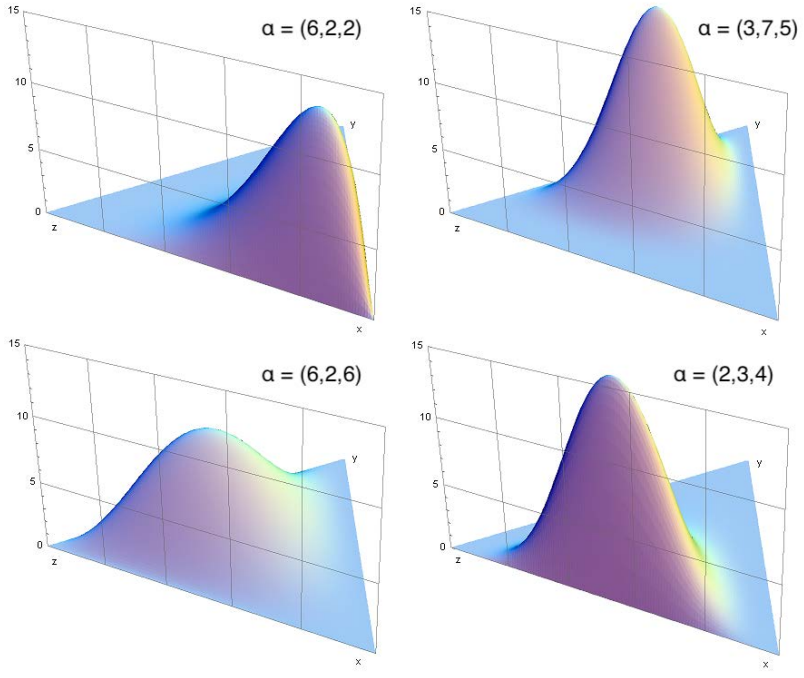


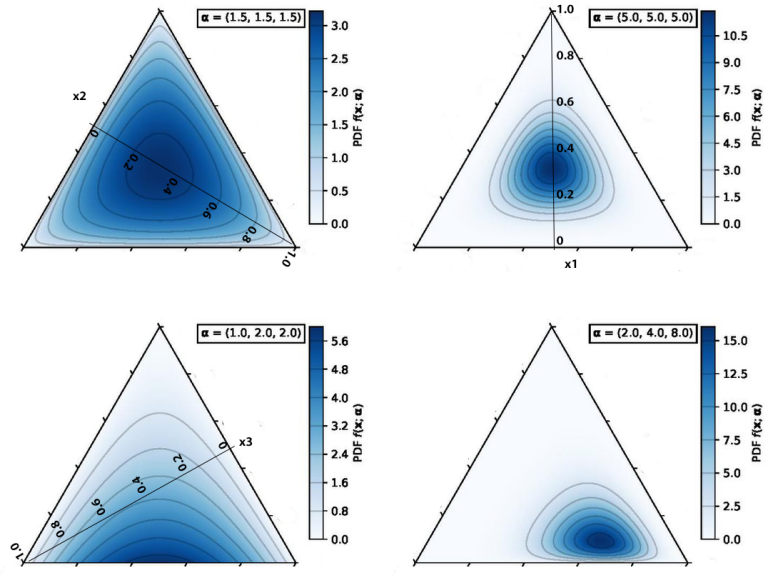
A point on the triangle is x_1, x_2, x_3, with $x_1, x_2, x_3 \in [0, 1]$ and $x_1 + x_2 + x_3 = 1$. For example, (0.05, 0.8, 0.15)

Generalises to higher dimensions with $x_1, ..., x_K \in [0, 1]$ and $\sum x_k = 1$

Key distribution 1: The Dirichlet distribution

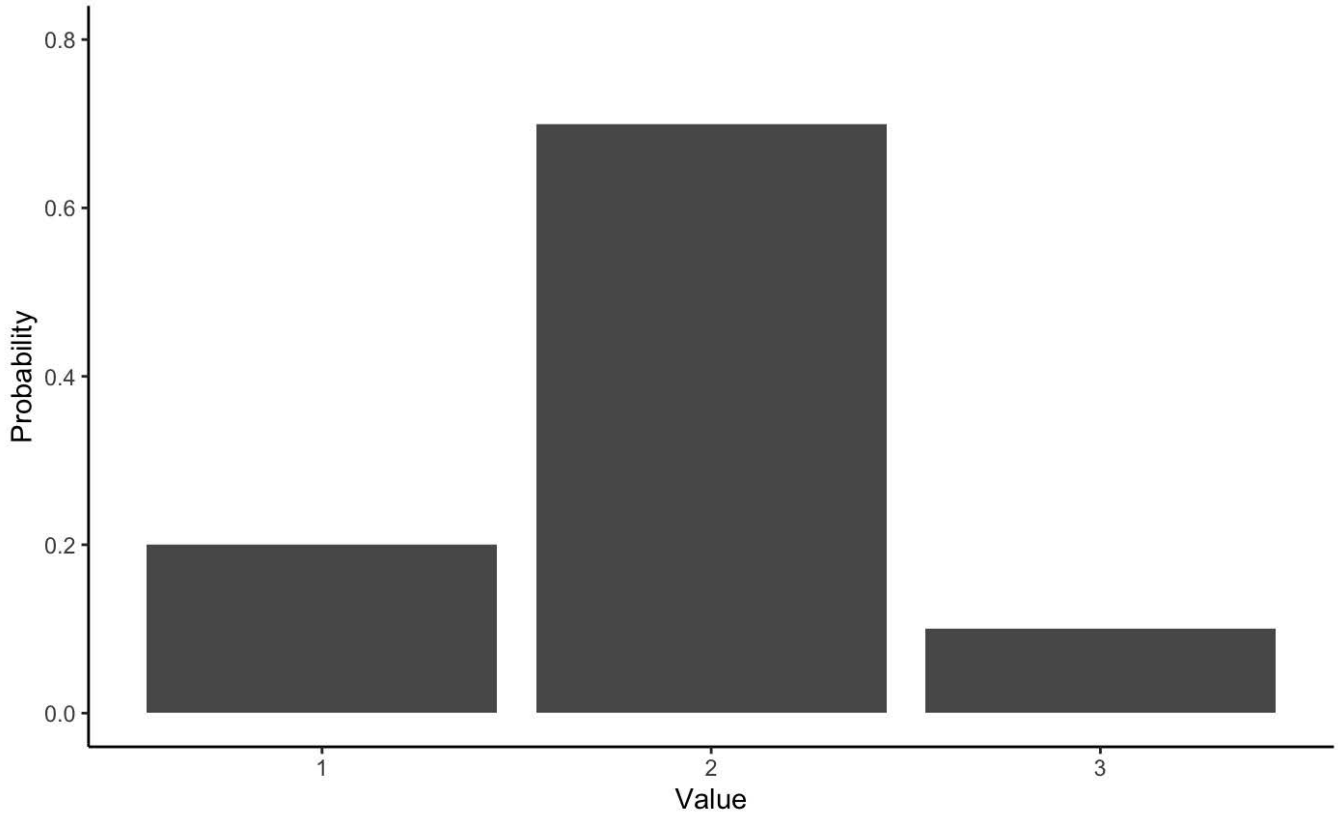
Dirichlet distribution: Probability distribution over a simplex





- A draw $b \sim Dir(\alpha)$ from this distribution could e.g. return $b = (0.2, 0.7, 0.1)$
- Hence, we can think of the draw from a Dirichlet distribution being itself a multinomial distribution (next slide)

Key distribution 2: Multinomial distribution



The multinomial distribution depicted has probabilities $[0.2, 0.7, 0.1]$. A draw $c \sim \text{Multinomial}([0.2, 0.7, 0.1])$ could e.g. return $c = 2$

Latent Dirichlet Allocation

- Document = random mixture over latent topics
- Topic = distribution over words

Probabilistic model with 3 steps:

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

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

Latent Dirichlet Allocation

Key parameters:

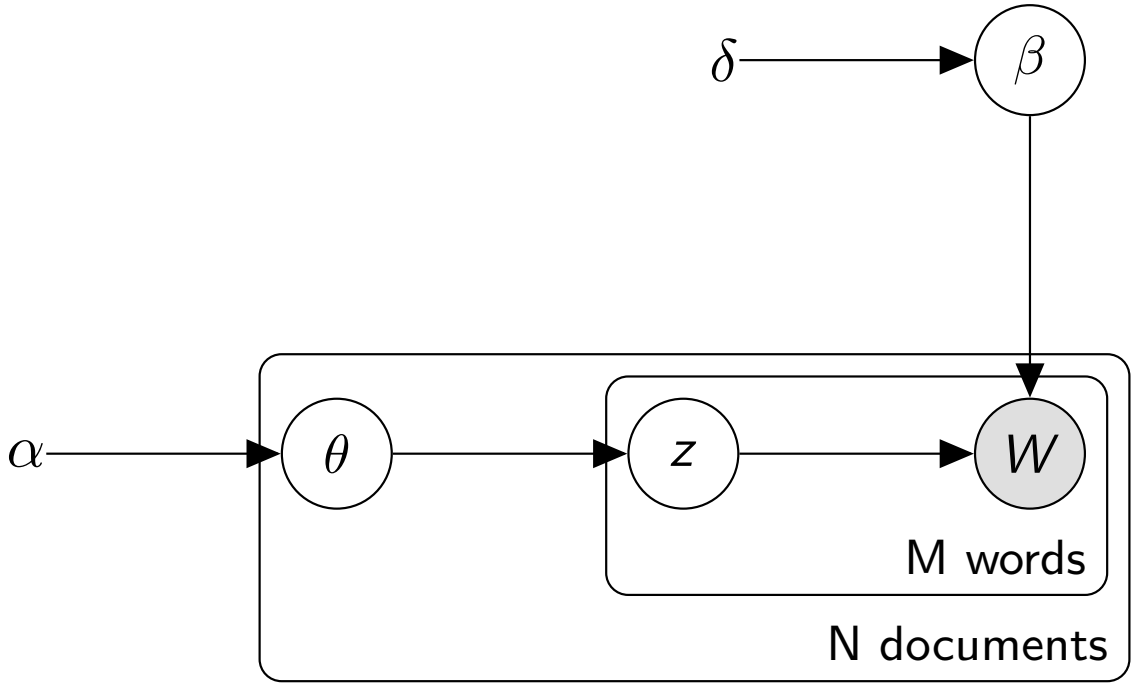
1. θ = matrix of dimensions N documents by K topics where θ_{ik} corresponds to the probability that document i belongs to topic k; i.e. assuming $K = 5$:

	T1	T2	T3	T4	T5
Document 1	0.15	0.15	0.05	0.10	0.55
Document 2	0.80	0.02	0.02	0.10	0.06
...					
Document N	0.01	0.01	0.96	0.01	0.01

2. β = matrix of dimensions K topics by M words where β_{km} corresponds to the probability that word m belongs to topic k; i.e. assuming $M = 6$:

	W1	W2	W3	W4	W5	W6
Topic 1	0.40	0.05	0.05	0.10	0.10	0.30
Topic 2	0.10	0.10	0.10	0.50	0.10	0.10
...						
Topic k	0.05	0.60	0.10	0.05	0.10	0.10

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

Outline

- Overview of topic models
- Latent Dirichlet Allocation (LDA)
- Examples
- Model Selection and Validation
- Extensions of LDA
- Implementations in R
- Guided coding

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)

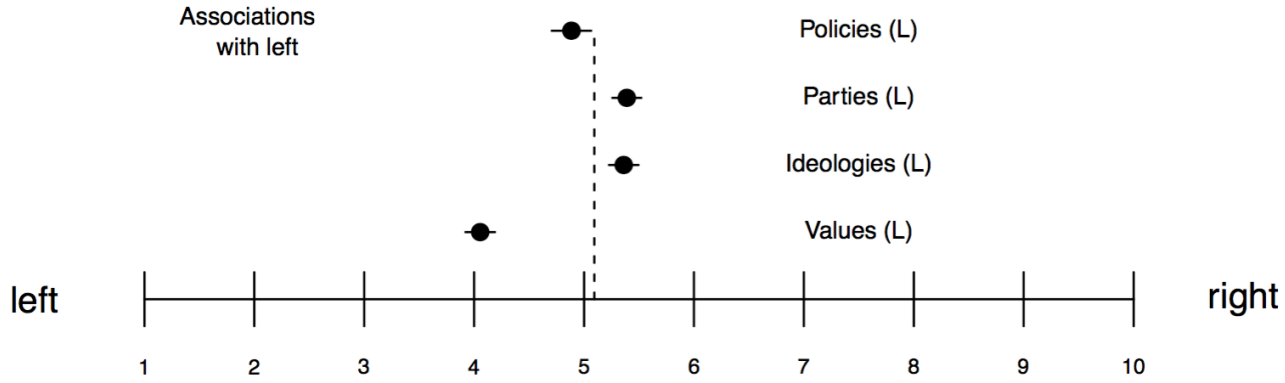
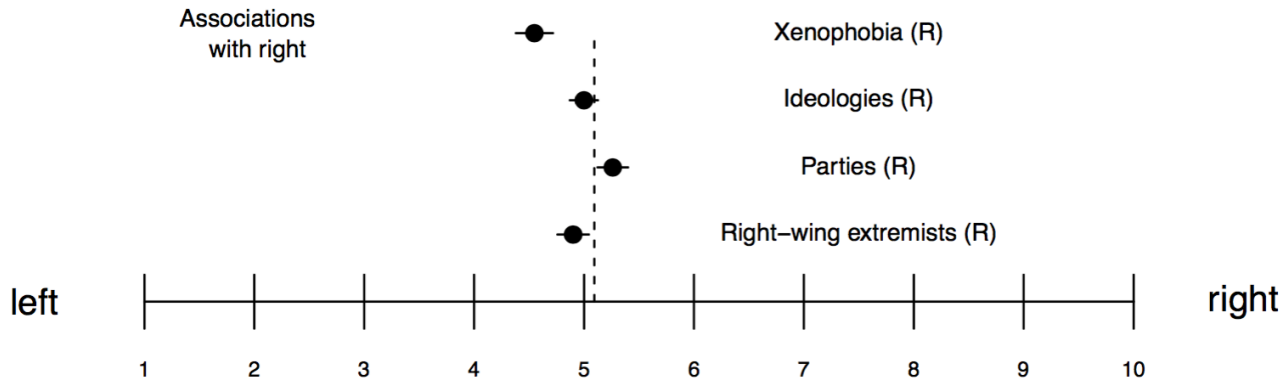


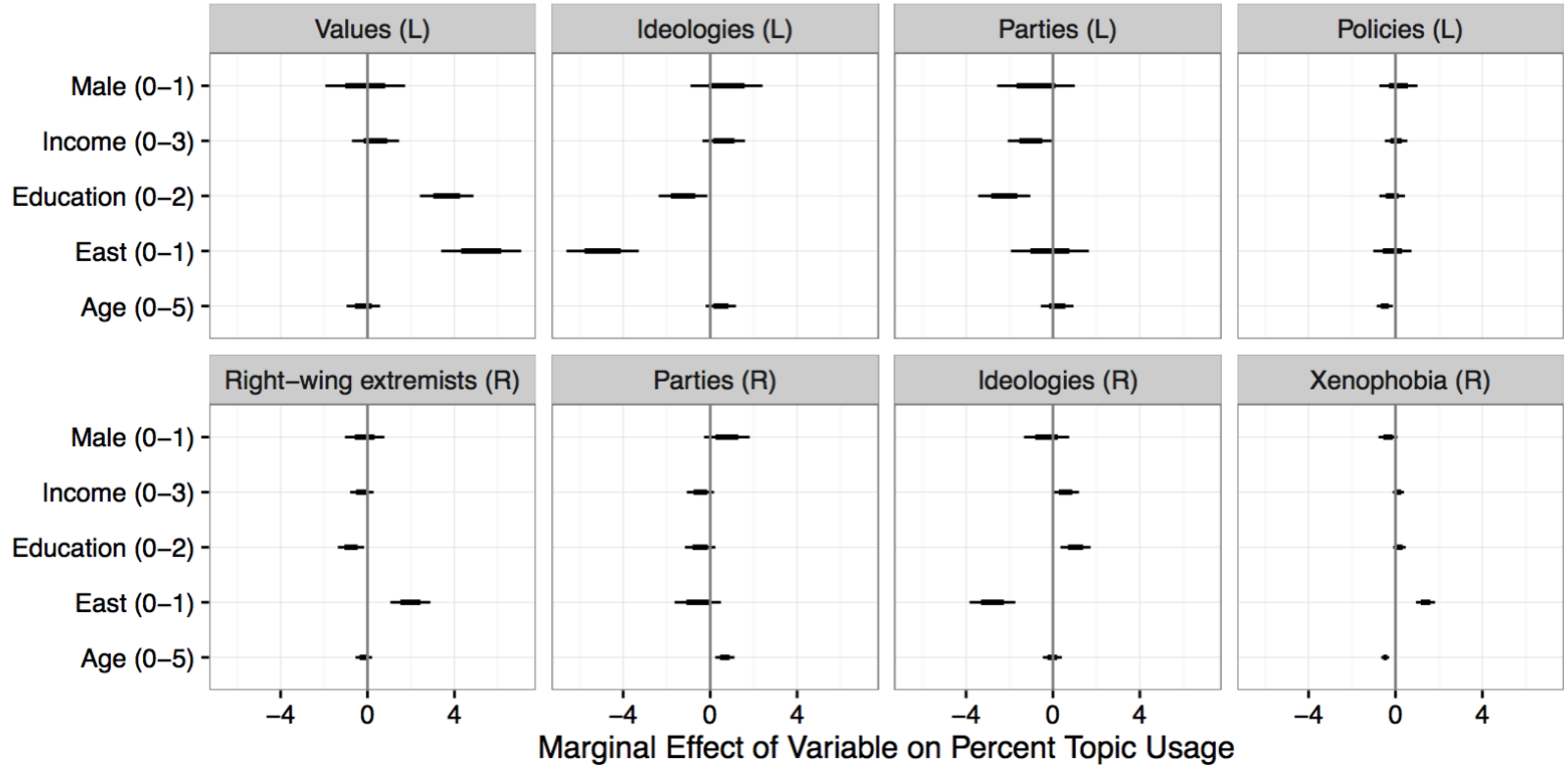
Fig. 7: Left-right scale means for different subsamples of associations with **right** (dashed = sample mean, bars = 95% Cis)



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

- A carefully documented project with very good average topic coherence is e.g. "Leaders or Followers? Measuring Political Responsiveness in the U.S. Congress Using Social Media Data" by Barberá et al. (2014)
- 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)
- Sidenote: For short texts, such as also e.g. survey responses, topic models such as sparse additive generative models (SAGE) might create more coherent topics than e.g. LDA (see e.g. Bauer et al, Political Behavior, 2017)
- $K = 100$ topics
- Validation: http://j.mp/lda-congress-demo

Outline

- Overview of topic models
- Latent Dirichlet Allocation (LDA)
- Examples
- Model Selection and Validation
- Extensions of LDA
- Implementations in R
- Guided coding

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?

Selecting K and beyond

- In particular selecting the amount of topics K, but also parameters, covariates, and potentially initialisations is a very challenging exercise without a single solution
- Researchers typically consult a combination of quantitative metrics and human judgement

Quantitative metrics

- Held-out likelihood or perplexity: For some held-out documents, how likely would the model have generated/predicted these documents
- Semantic coherence: For example, how likely do the most common words from a topic also co-occur in the same document?
- Exclusivity: Do words with high probability in one topic have low probabilities in others?
- Many automated metrics exist, see e.g. Grimmer and Stewart (2013), Mimno et al. (2011), Taddy (2012)

On held-out likelihood metrics

- Held-out likelihood and perplexity allow to judge the predictive ability of the model
- Yet, choosing K such that it achieves the highest held-out likelihood has serious limitations
- Rather than creating an auto-complete prediction model or something similar, the purpose of topic modelling most often is to obtain coherent topics that tell a story
- In their paper "Reading tea leaves" (2009) Chang et al. contrast likelihood based metrics with human judgements about topic coherence
- Also see this short video by one of the authors

Reading tea leaves

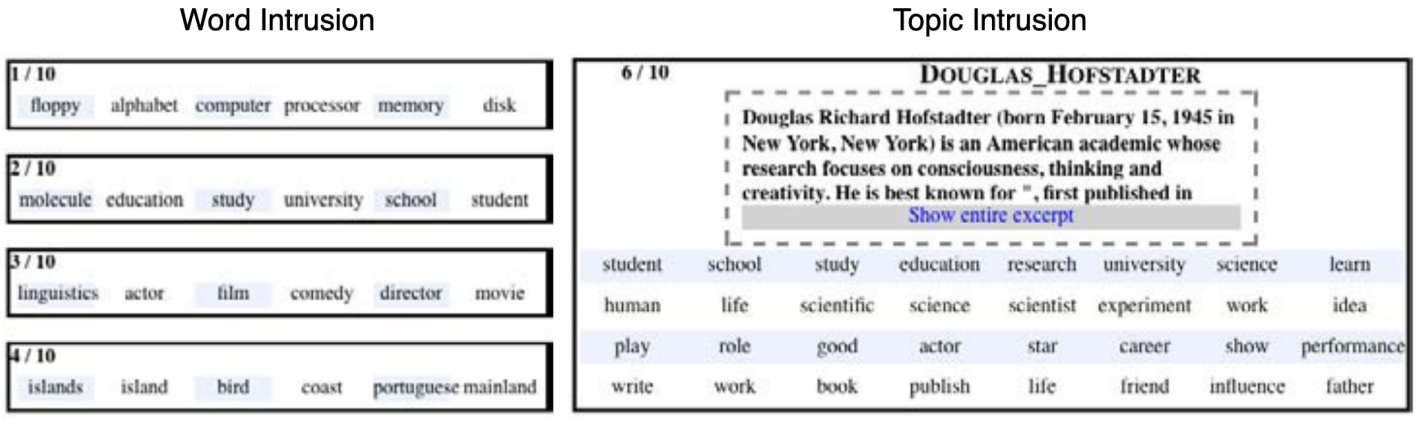


Figure 2: Screenshots of our two human tasks. In the word intrusion task (left), subjects are presented with a set of words and asked to select the word which does not belong with the others. In the *topic intrusion* task (right), users are given a document's title and the first few sentences of the document. The users must select which of the four groups of words does not belong.

Figure: From: "Reading Tea Leaves: How Humans Interpret Topic Models" by Chang et al., 2009

- Likelihood based metrics were actually negatively correlated with human metrics about topic coherence

Quantitative metrics

- For a discussion of model selection for (structural) topic models (e.g. different choices of K, initialisations, and covariates) and evaluation, see e.g. Section "3.4. Evaluate: Model selection and search" in the package vignette or the Section *Model Specification and Selection* in "Structural Topic Models for Open-Ended Survey Responses" by Roberts et al. (2014)

- In Roberts et al. (2014), the authors e.g. argue that "a semantically interpretable topic has two qualities: (1) it is cohesive in the sense that high-probability words for the topic tend to co-occur within documents, and (2) it is exclusive in the sense that the top words for that topic are unlikely to appear with in top words of other topics."

- Semantic coherence and exclusivity, with many other quantitative metrics, are outputs of the function 'searchK' in the 'stm' package

Takeways

- No quantitative metric can replace human judgement when selecting K or other model parameters, and evaluating the fit of a particular topic model more generally
- "The most effective method for assessing model fit is to carefully read documents that are closely associated with particular topics to verify that the semantic concept covered by the topic is reflected in the text." from "A Model of Text for Experimentation in the Social Sciences" by Roberts et al. (2016)
- The 'plotQuote' function in 'stm' allows to plot documents highly associated with certain topics
- Key consideration: What is the goal of the current model?
- To generate coherent topics which describe themes? Combine careful human reading with quantitative metrics
- To use topic models e.g. for document embeddings to find similar documents? As input in a predictive model? Try to select K and other parameters such that they maximise whatever the objective function

Outline

- Overview of topic models
- Latent Dirichlet Allocation (LDA)
- Examples
- Model Selection and Validation
- Extensions of LDA
- Implementations in R
- Guided coding

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)

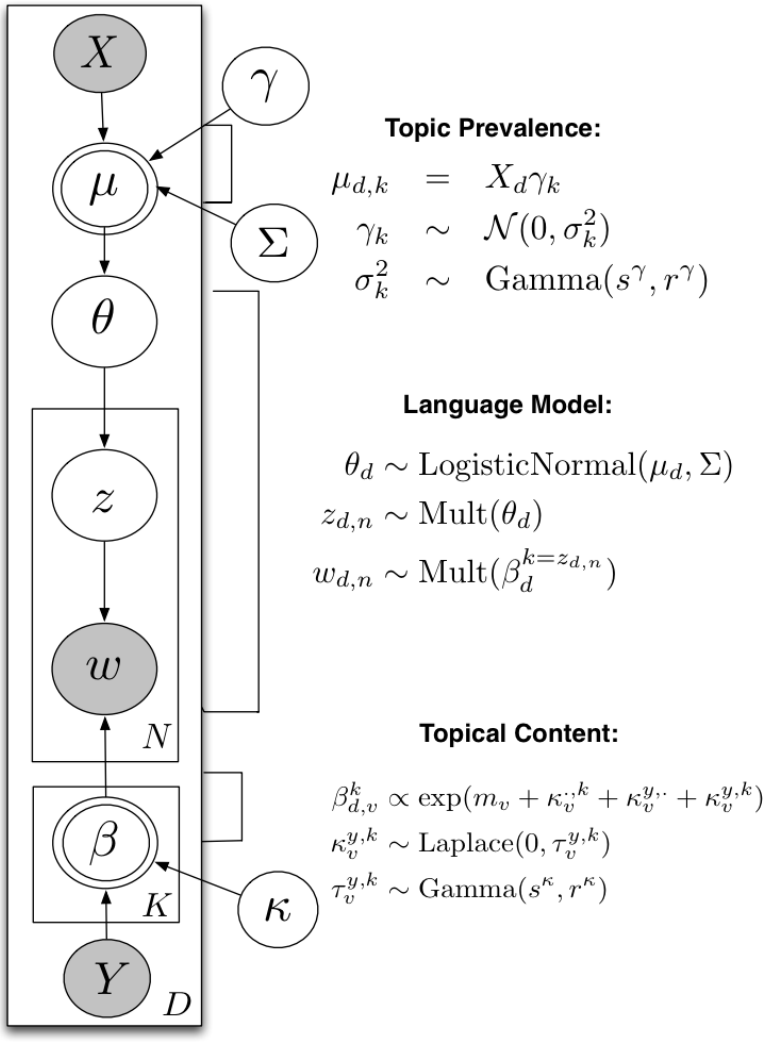
Why?

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

Structural topic model (STM)

- Basic idea: 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)
- Including context improves the model:
  - more accurate estimation
  - better qualitative interpretability

Structural topic model (STM)



Topic Prevalence:

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

Language Model:

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

Topical Content:

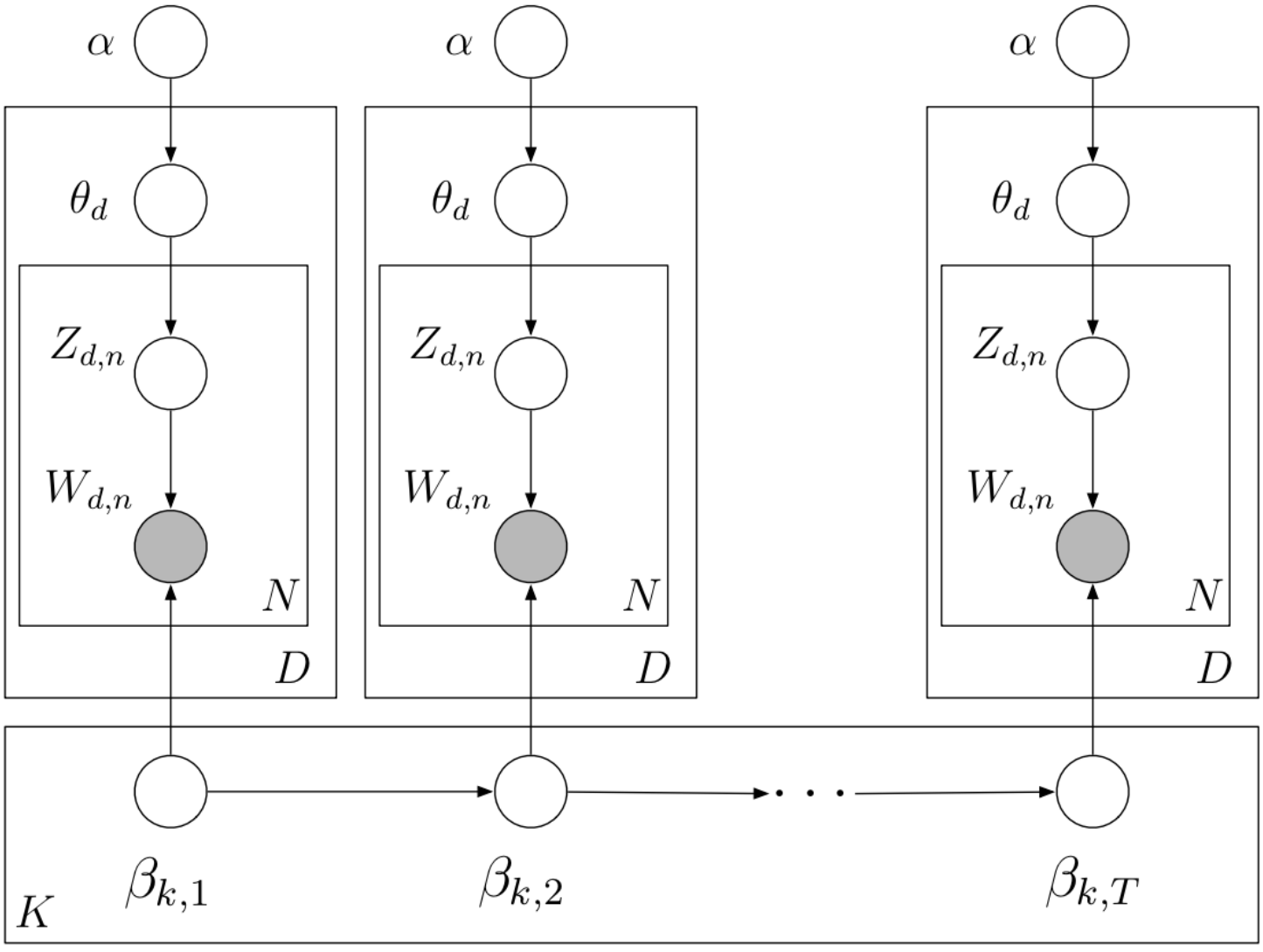
$$\begin{aligned} \beta_{d,v}^k &\propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k}) \\ \kappa_v^{y,k} &\sim \text{Laplace}(0, \tau_v^{y,k}) \\ \tau_v^{y,k} &\sim \text{Gamma}(s^\kappa, r^\kappa) \end{aligned}$$

- **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)

Structural topic model (STM)

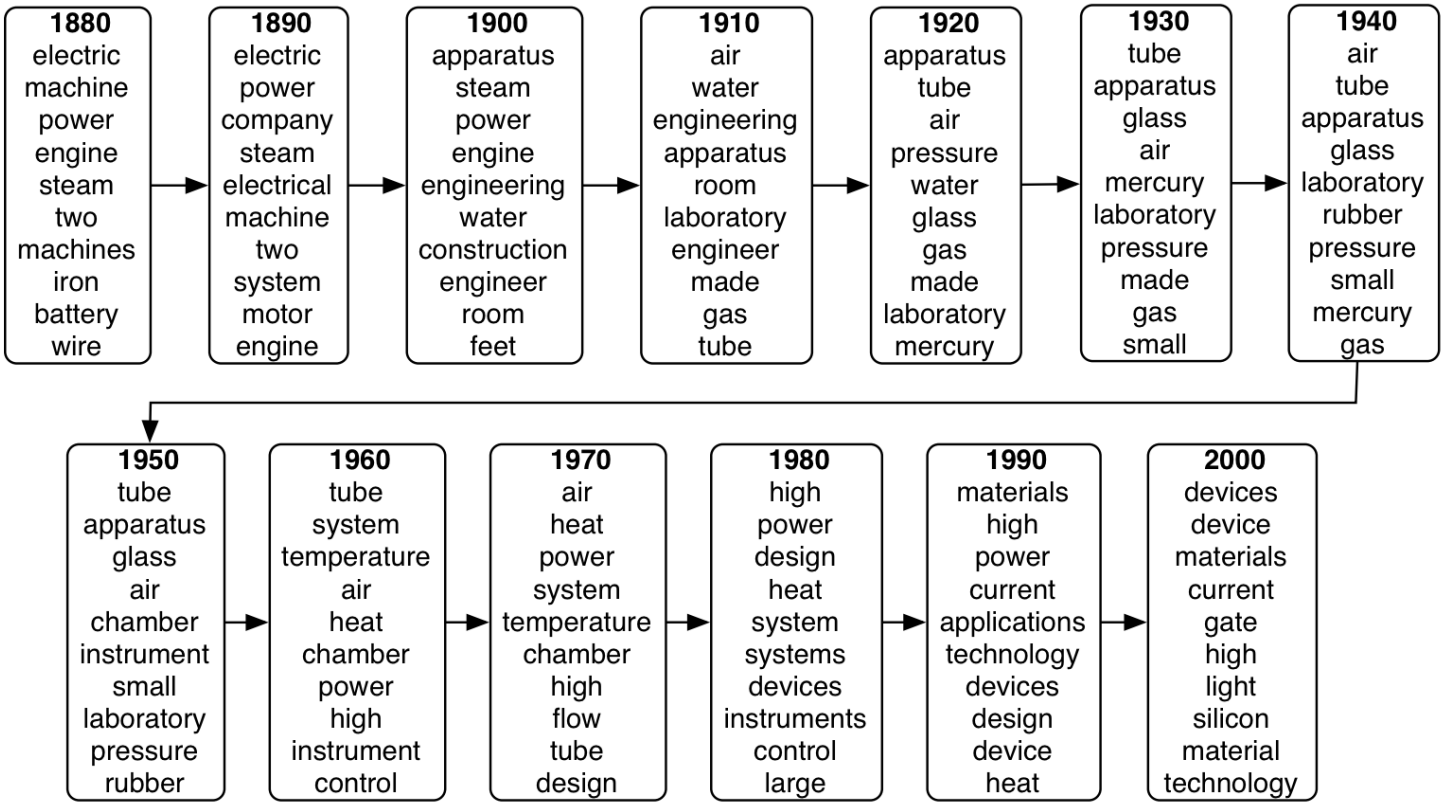
- User specifies the number of topics: K
- Observed data for standard topic models
  - Each document ($d \in 1 \ldots D$) is a collection of N_d tokens
  - Each token is a particular word from a dictionary of V entries
  - Data summarized in a single matrix $D \times V$ matrix $\mathbf{W}$
- Additional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

Dynamic topic model



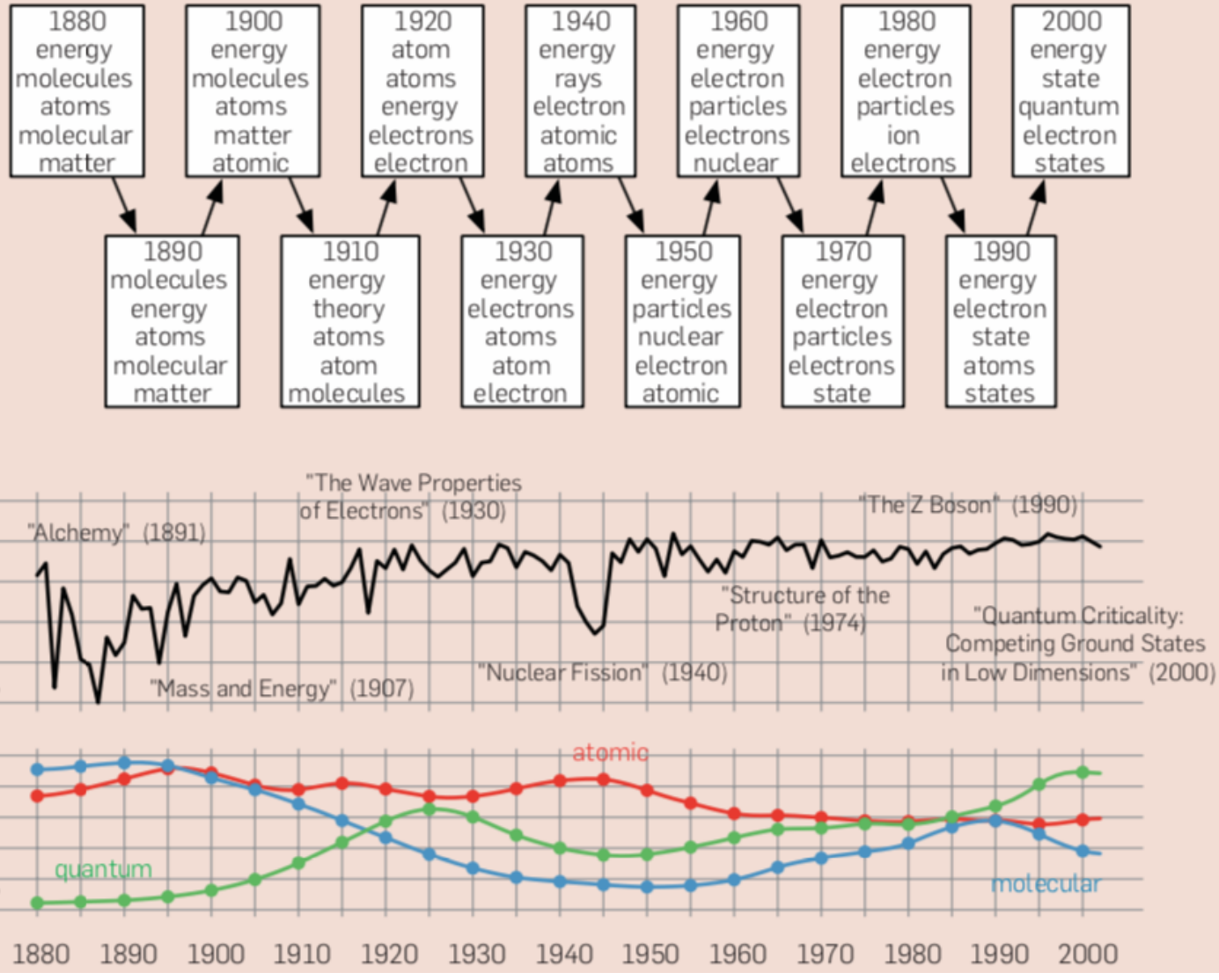
Source: Blei, "Modeling Science"

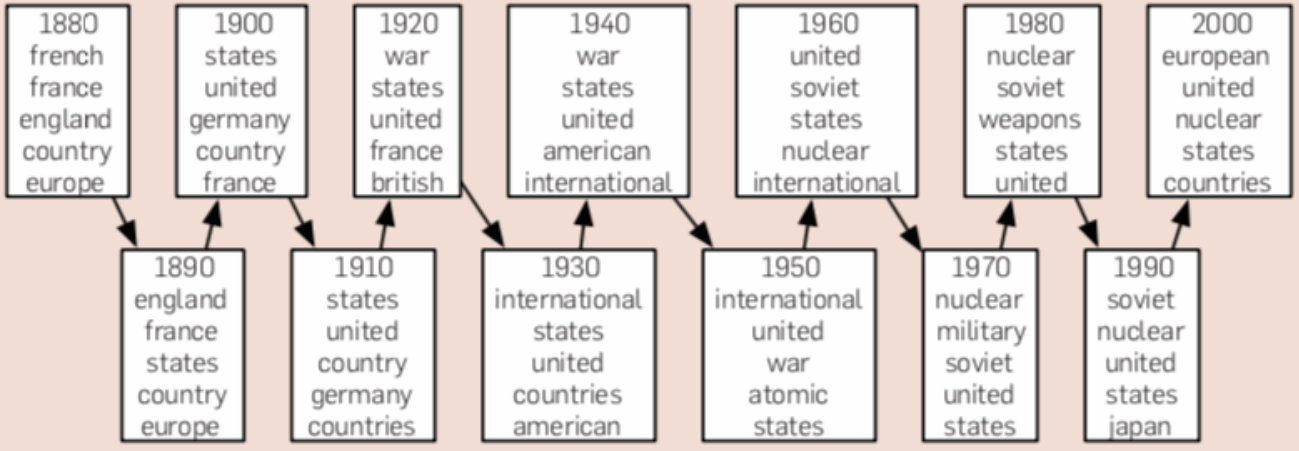
Dynamic topic model

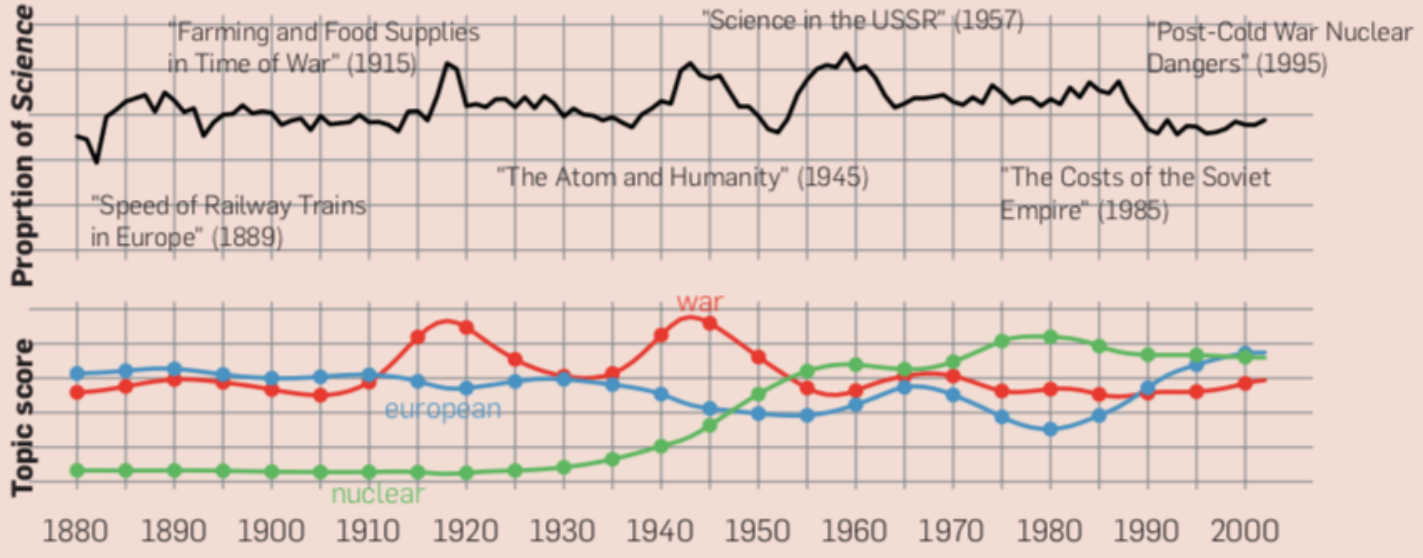


Source: Blei, "Modeling Science"

Figure 5. Two topics from a dynamic topic model. This model was fit to *Science* from 1880 to 2002. We have illustrated the top words at each decade.




1880
french
france
england
country
europe
1890
england
france
states
country
europe
1900
states
united
germany
country
france
1910
states
united
country
germany
countries
1920
war
states
united
france
british
1930
international
states
united
countries
american
1940
war
states
united
american
international
1950
international
united
war
atomic
states
1960
united
soviet
states
nuclear
international
1970
nuclear
military
soviet
united
states
1980
nuclear
soviet
weapons
states
united
1990
soviet
nuclear
united
states
japan
2000
european
united
nuclear
states
countries


Proprtion of Science
"Farming and Food Supplies in Time of War" (1915)
"Science in the USSR" (1957)
"Post-Cold War Nuclear Dangers" (1995)
"Speed of Railway Trains in Europe" (1889)
"The Atom and Humanity" (1945)
"The Costs of the Soviet Empire" (1985)
Topic score
war
european
nuclear
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000

Outline

- Overview of topic models
- Latent Dirichlet Allocation (LDA)
- Examples
- Model Selection and Validation
- Extensions of LDA
- Implementations in R

Implementations in R

- For an implementation of the LDA and CTM, see e.g. the package ‘topicmodels‘
- We will focus on the ‘stm‘ package which offers a range of helpful functionalities
- Without added covariates, the ‘stm‘ function also estimates a standard CTM, with covariates a structural topic model
- Further helpful package to visualise topic models are ‘LDAvis‘ or, the stm-specific, ‘stminsights‘

Functionalities ‘stm‘ package

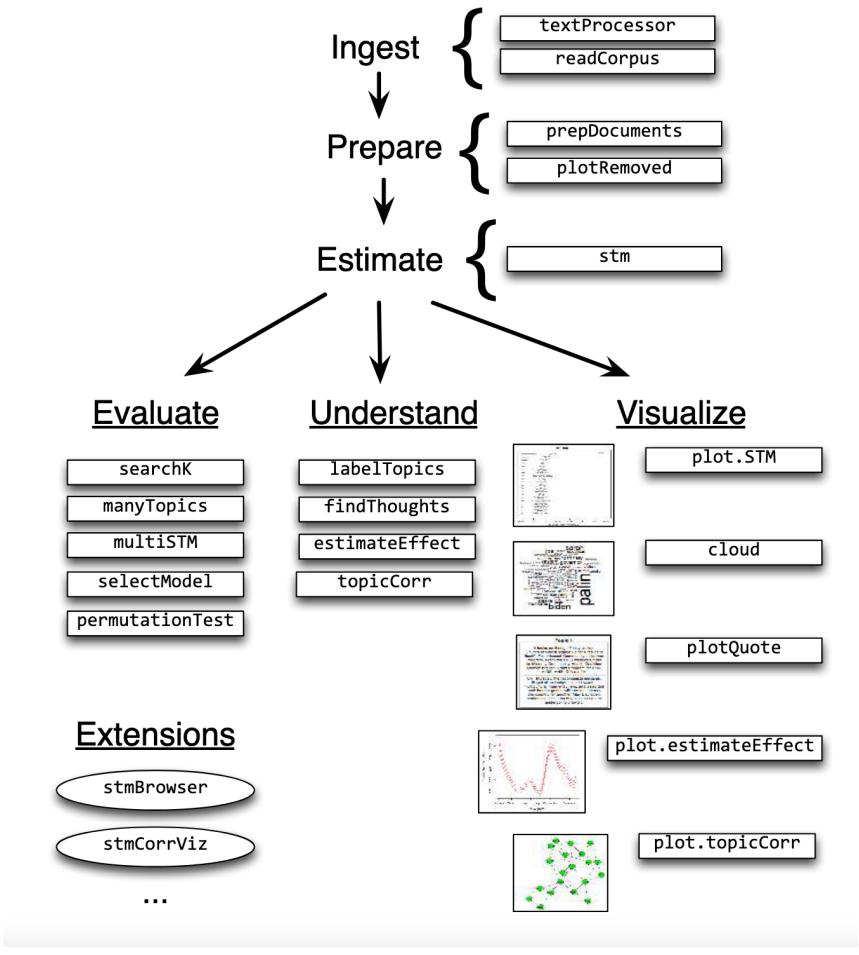


Figure: From the vignette “stm: RPackage for Structural Topic Models” by Roberts, Stewart, and Tingley