Building a Robot Judge: Data Science for the Law

7. Topic Models

Elliott Ash

Different Goals, Different Methods

- Supervised: Pursuing a known goal.
  - e.g., predicting whether a defendant will win his case.
  - machine learns to replicate human decision process.
- Unsupervised:
  - algorithm discovers themes/patterns in data (e.g. text)
    - e.g., k-means clustering of similar documents.
  - human interprets the results (e.g. clusters)
- Both strategies amplify human effort, each in different ways.
- Also: supervised learning models can be used to discover themes/patterns, and unsupervised learning models can be used in service of prediction or known goals.

- Today, focus on **unsupervised learning** using **topic models**.

Topic Models in Social Science

- Core methods for topic models were developed in computer science and statistics
  - summarize unstructured text
  - use words within document to infer subject
  - useful for dimension reduction
- Social scientists wanted to use topics as a form of measurement
  - how observed covariates drive trends in language
  - tell a story not just about what, but how and why
  - **topic models are more interpretable** than other methods, e.g. principal components analysis.

Some example questions

- How do U.S. politicians present their work to the public? What explains variation in representational style? (Grimmer 2013)

- Did electoral reform change the portfolio of issues addressed by politicians in Japan? (Catalinac 2016)

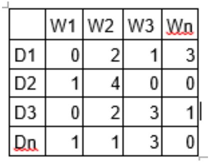
- What are the propaganda strategies of the Chinese government? (Roberts and Stewart 2016)

- How do central bankers respond to an increase in transparency over their discussions? (Hansen, McMahon, and Pray 2015)

Latent Dirichlet Allocation (LDA)

- Idea: documents exhibit each topic in some proportion.
  - Each document is a distribution over topics.
  - Each topic is a distribution over words.
- Latent Dirichlet Allocation (e.g. Blei 2012) estimates:
  - The distribution over words for each topic.
  - The proportion of a document in each topic, for each document.
- Maintained assumptions: Bag of words/phrases, and fix number of topics ex ante.

Document-term Matrix X

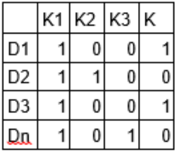


	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

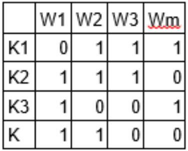
- A corpus of N documents $D_1, D_2, D_3 \ldots D_n$
- Vocabulary of M words $W_1, W_2 ... W_m$.
- The value of i, j cell gives the frequency count of word W_j in Document D_i.

Matrix factoring

- LDA factors the document-term matrix into two lower-dimensional matrices, M_1 and M_2:



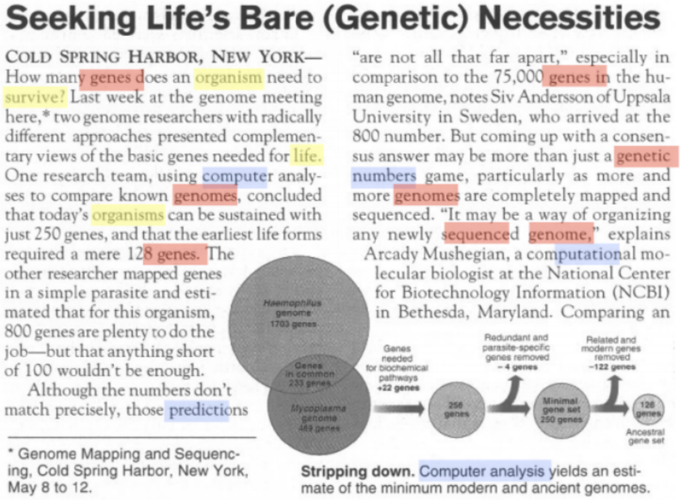
	K1	K2	K3	K
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
Dn	1	0	1	0



	W1	W2	W3	Wm
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

- M_1 is a $N \times K$document-topic matrix
- M_2 is a $K \times M$ topic-term matrix.

A statistical highlighter



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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set

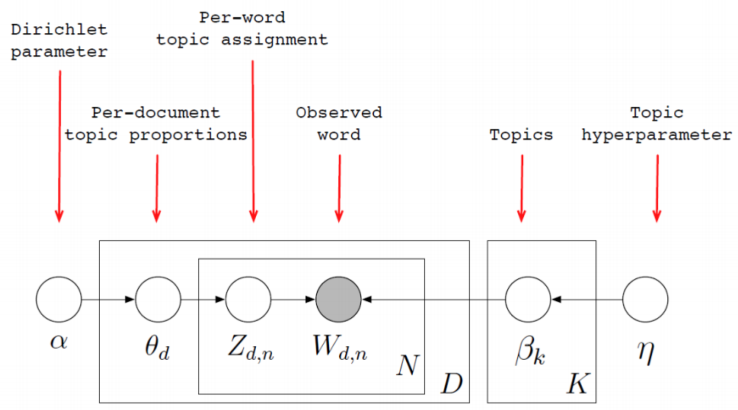
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

Image from Hanna Wallach

A Bayesian Model

Figure: Plate Notation of Latent Dirichlet Allocation



Source: Brandon Stewart Topic Models Slides.

LDA Parameters

- α: document-topic density
  - higher α means documents contain more topics, lower α means documents contain fewer topics
- β: topic-word density
  - higher β means topics have more words, while lower β means topics have fewer words
- Number of topics:
  - this is specified in advance, or can be chosen to optimize model fit.
  - the "statistically optimal" topic count is usually too high for the topics to be interpretable/useful.

Why does this work? Co-occurrence

- Where is the information for each word's topic?
  - We are learning the pattern of what words occur together.
- The model wants a topic to contain as few words as possible, but a document to contain as few topics as possible.
  - This tension is what makes the model work

LDA in Python

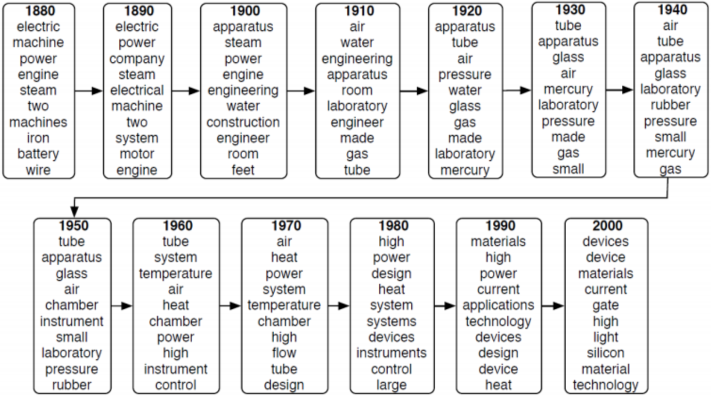
- gensim provides the best impementation of LDA in Python (gensim.models.LdaModel)
    - streams from disk (so works on arbitrarily large corpora), intuitive, fast (cython, parallelized)
    - See accompanying Jupyter notebooks.
- "passes" is number of times to go through the corpus
    - probably doesn't matter for large corpora
    - if your topics differ significantly across runs, you need more passes.
- Once trained, can easily get topic proportions for a document.

Extensions

- There are a ton of extensions/variants of LDA.
  - But almost all of them are very context-specific.
  - LDA is great because it works so well across different domains.

Dynamic Topic Model

Figure: Topic Evolution over Time



1880
electric
machine
power
engine
steam
two
machines
iron
battery
wire

1890
electric
power
company
steam
electrical
machine
two
system
motor
engine

1900
apparatus
steam
power
engine
engineering
water
construction
engineer
room
feet

1910
air
water
engineering
apparatus
room
laboratory
engineer
made
gas
tube

1920
apparatus
tube
air
pressure
water
glass
gas
made
laboratory
mercury

1930
tube
apparatus
glass
air
mercury
laboratory
pressure
made
gas
small

1940
air
tube
apparatus
glass
laboratory
rubber
pressure
small
mercury
gas

1950
tube
apparatus
glass
air
chamber
instrument
small
laboratory
pressure
rubber

1960
tube
system
temperature
air
heat
chamber
power
high
instrument
control

1970
air
heat
power
system
temperature
chamber
high
flow
tube
design

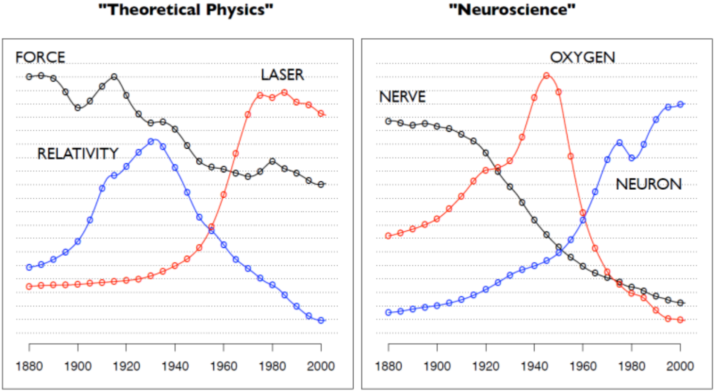
1980
high
power
design
heat
system
systems
devices
instruments
control
large

1990
materials
high
power
current
applications
technology
devices
design
device
heat

2000
devices
device
materials
current
gate
high
light
silicon
material
technology

Dynamic Topic Model

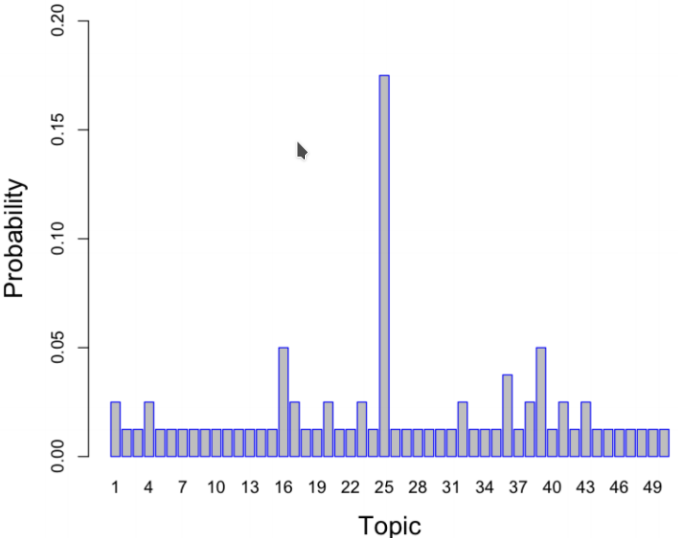
Figure: Word Use in Topics Over Time



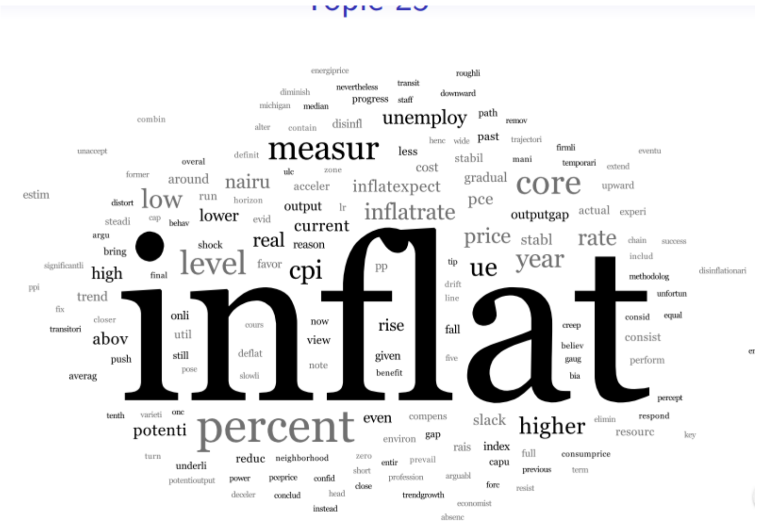
LDA on FOMC

- Hansen, McMahon, and Prat use LDA to analyze speech at the FOMC (Federal Open Market Committee).
    - private discussions among committee members at Federal Reserve (U.S. Central Bank)
    - transcripts: 150 meetings, 20 years, 46000 speeches
- They take the standard pre-processing steps and train LDA on the speeches.

Distribution of Attention



What is topic 25?



Overview of results

- They show that increasing transparency results in:
  - higher discipline / technocratic language (probably beneficial)
  - higher conformity (probably costly)
- Highlights tradeoffs from transparency in bureaucratic organizations.

Categorization of Union Contract Clauses

- Ash, MacLeod, and Naidu (2018) represent union contracts as a list of clauses:
  - $< Agent >< Obligation/Entitlement >< Action >$.
- the "action" segment of a clause includes connected pieces of the parse tree besides the subject (agent) and modal (obligation/entitlement).
  - How to encode actions as data?
- LDA Approach:
  - Classify each action clause by topic using Latent Dirichlet Allocation.
  - We got good results with 20 topics.

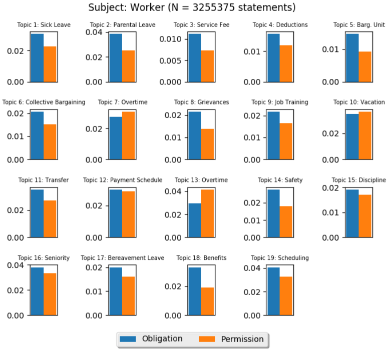
LDA Topics (1 of 2)

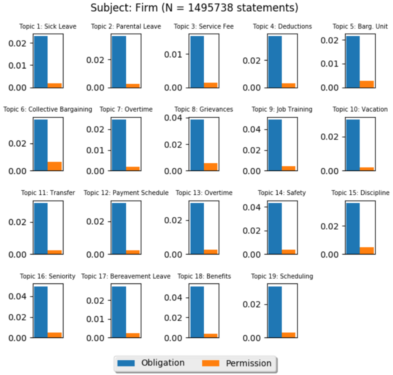
- 1 -- **"Sick Leave"** -- period month sick leave six probationary credit three complete employment twelve absent completion accumulate date exceed consecutive professional
- 2 -- **"Parental Leave"** -- leave absence pay request date grant prior week parental commencement pregnancy write maternity duty witness advance approve notice
- 4 -- **"Payroll"** -- change due result deduction amount status deduct monthly payroll reduction affect cheque technological fee employment orientation statement
- 5 -- **"Bargaining Unit"** -- unit bargaining person appointment appoint employ outside activity membership represent agent terminal sole select exercise ontario bargain behalf
- 7 -- **"Overtime"** -- hour shift work schedule overtime period call rest meal half minute start end break duty sunday weekend saturday two friday
- 8 -- **"Grievances"** -- grievance party procedure arbitration writing decision write step matter arbitrator committee complaint submit final dispute request name process
- 9 -- **"Job Training"** -- requirement operation training require equipment individual meet service responsibility provide program area manner performance" business duty operational
- 10 -- **"Vacation Leave"** -- year vacation service pay date employment week continuous effective two annual entitlement percent january salary earn termination period follow

LDA Topics (2 of 2)

- 14 – **"Medical Leave/Injuries"** medical reasonable illness reason certificate unable duty injury course require due provide information circumstance accident personal condition examination reasonably
- 15 -- **"Discipline/Firing"** -- school act safety committee health action discharge labour cause discipline disciplinary file application canada public relations suspension regulation authority accordance
- 16 -- **"Seniority"** -- seniority lay position list layoff vacancy recall transfer post temporary qualification permanent job hire fill date provide ability copy basis
- 17 -- **"Work-Related Deaths"** – article accordance law child spouse pursuant family death include immediate parent purpose require city office paragraph funeral
- 18 -- **"Insurance/Benefits"** -- benefit plan insurance payment cost premium eligible provide receive compensation disability pay coverage pension receipt term amount
- 19 -- **"Scheduling"** -- work hour day week schedule two return perform normal regular report normally excess regularly require notice eight teaching available emergency

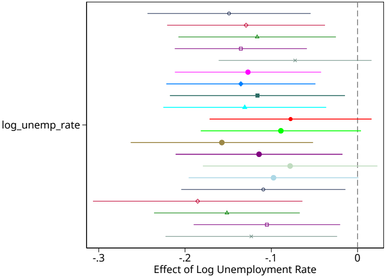
Workers Have More Entitlements Relative To Obligations





- Workers get more authority at work than employers, consistently across work areas.

Which contracts topics change in response to change in local unemployment?



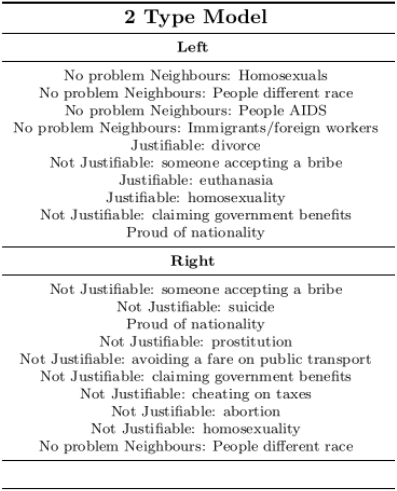
- Topic 7 (turquoise, square) is “Payment Schedule”; Topic 9 (bright red) is “Job Training”; Topic 10 (bright green) is “Vacations”; Topic 16 (dark red, diamond) is “Seniority”.

LDA on CEO task data

- LDA can be used on any type of co-occurence data.
  - Bandiera, Hansen, Prat, and Sadun (2017): use LDA to do dimension reduction on CEO tasks using time allocation data.
    - model endogenously identifies a “leader” CEO topic, and a “micro-manager” CEO topic

LDA for ideology

- Draca and Schwarz (2018): use LDA to dimension-reduce the features of political attitudes survey responses.



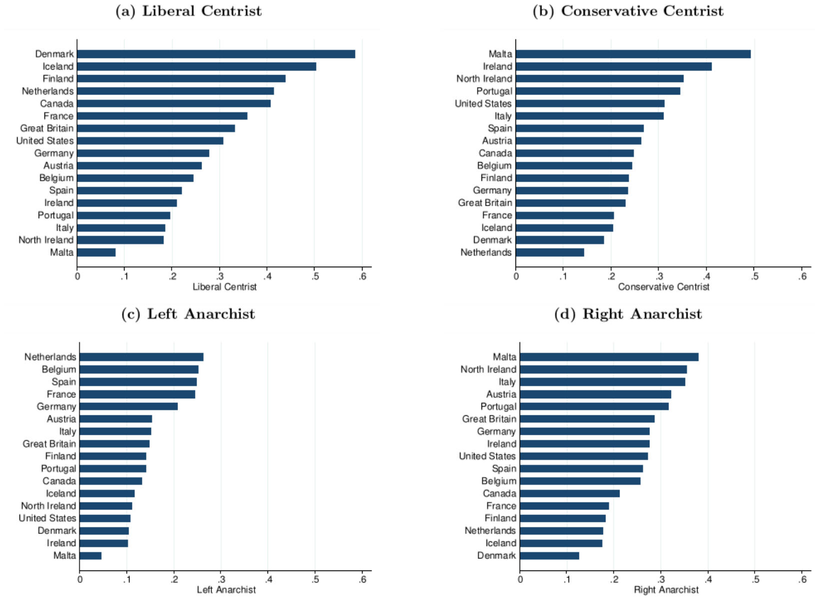
2 Type Model
Left
No problem Neighbours: Homosexuals
No problem Neighbours: People different race
No problem Neighbours: People AIDS
No problem Neighbours: Immigrants/foreign workers
Justifiable: divorce
Not Justifiable: someone accepting a bribe
Justifiable: euthanasia
Justifiable: homosexuality
Not Justifiable: claiming government benefits
Proud of nationality
Right
Not Justifiable: someone accepting a bribe
Not Justifiable: suicide
Proud of nationality
Not Justifiable: prostitution
Not Justifiable: avoiding a fare on public transport
Not Justifiable: claiming government benefits
Not Justifiable: cheating on taxes
Not Justifiable: abortion
Not Justifiable: homosexuality
No problem Neighbours: People different race

- running LDA with two topics gives intuitive liberal-conservative categories.

3 Type Model
Liberal Centrist
Confidence: Police
Confidence: Justice System/Courts
No problem Neighbours: Homosexuals
No problem Neighbours: People different race
No problem Neighbours: People AIDS
No problem Neighbours: Immigrants/foreign workers
Proud of nationality
Not Justifiable: someone accepting a bribe
Confidence: The Civil Services
Not Justifiable: cheating on taxes
Conservative Centrist
Not Justifiable: homosexuality
Not Justifiable: abortion
Not Justifiable: suicide
Not Justifiable: prostitution
Proud of nationality
Not Justifiable: someone accepting a bribe
Not Justifiable: avoiding a fare on public transport
Not Justifiable: claiming government benefits
Not Justifiable: cheating on taxes
Not Justifiable: euthanasia
Anarchist
No Confidence: Civil Services
No Confidence: Parliament
No Confidence: Churches
No Confidence: Major Companies
No Confidence: Justice System/Courts
No Confidence: The Press
No problem Neighbours: Homosexuals
No problem Neighbours: People different race
No problem Neighbours: People AIDS
No Confidence: Labour Unions

4 Type Model
Liberal Centrist
Confidence: Police
No problem Neighbours: Homosexuals
No problem Neighbours: People AIDS
No problem Neighbours: People different race
No problem Neighbours: Immigrants/foreign workers
Not Justifiable: someone accepting a bribe
Proud of nationality
Not Justifiable: claiming government benefits
Justifiable: divorce
Not Justifiable: cheating on taxes
Conservative Centrist
Confidence: Police
Confidence: Churches
Not Justifiable: suicide
Proud of nationality
Confidence: Armed Forces
Not Justifiable: prostitution
Not Justifiable: abortion
Not Justifiable: someone accepting a bribe
Confidence: Justice System/Courts
Confidence: The Civil Services
Left Anarchist
Justifiable: divorce
No Confidence: Churches
Justifiable: euthanasia
No Confidence: Armed Forces
No Confidence: Parliament
No Confidence: Civil Services
No problem Neighbours: Homosexuals
Justifiable: homosexuality
Justifiable: abortion
No problem Neighbours: People different race
Right Anarchist
No Confidence: Parliament
No Confidence: Civil Services
No Confidence: Labour Unions
No Confidence: The Press
No Confidence: Justice System/Courts
Not Justifiable: someone accepting a bribe
Not Justifiable: avoiding a fare on public transport
Not Justifiable: claiming government benefits
Not Justifiable: suicide
Not Justifiable: cheating on taxes

Figure 4: Country-Level Type Shares


(a) Liberal Centrist

(b) Conservative Centrist

(c) Left Anarchist

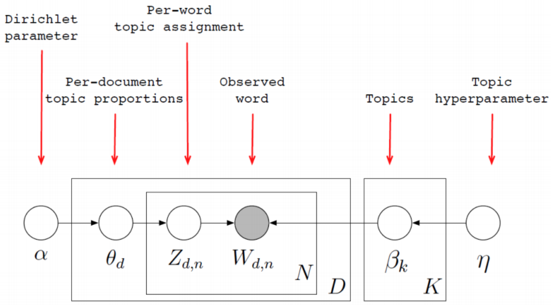
(d) Right Anarchist

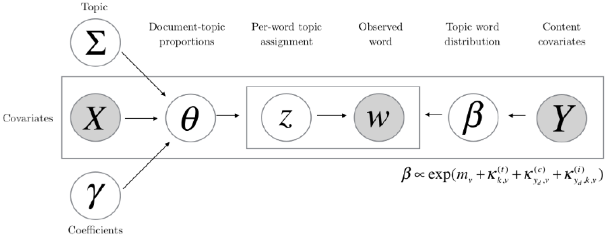
Notes: This figure shows the mean country-level θ type shares aggregated over individuals. Country means calculated using WVS sample weights.

Structural Topic Model = LDA + Metadata

- STM provides two ways to include contextual information:
    - Topic prevalence can vary by metadata
        - e.g. Republicans talk about military issues more then Democrats
    - Topic content can vary by metadata
        - e.g. Republicans talk about military issues differently from Democrats.

LDA vs. STM – Illustration




Topic
Σ
Document-topic proportions
Per-word topic assignment
Observed word
Topic word distribution
Content covariates
Covariates
X
θ
z
w
β
Y
γ
Coefficients
$\beta \propto \exp(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v})$

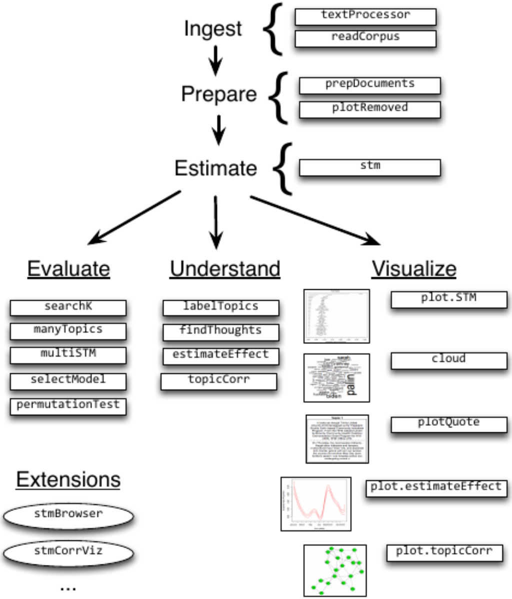
stm Package in R

- 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

stm has great functions/features



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

...

Caveats

- Structural topic model is not a prediction model:
  - it will tell you which topics or features correlate with an outcome, but it will not provide an in-sample or out-of-sample prediction for an outcome
- STM does not work with streaming data (yet)
  - have to load the whole corpus into memory