# Quantitative Content Analysis: Lecture 11

Matthias Haber

26 April 2017

# Today's outline

- Topic Models
- Wrapping up

# Topic models: basic idea

We often have collections of documents that we'd like to divide into natural groups so that we can understand them separately. Topic modeling is a method for unsupervised classification of such documents, which finds natural groups of items even when we're not sure what we're looking for.

- Topic models are exploratory probability models that
  - weaken the contraints required in dictionary based content analysis
  - have been intensively studied in the computer science literature
- Topic models work best with large amounts of text with a thematic structure

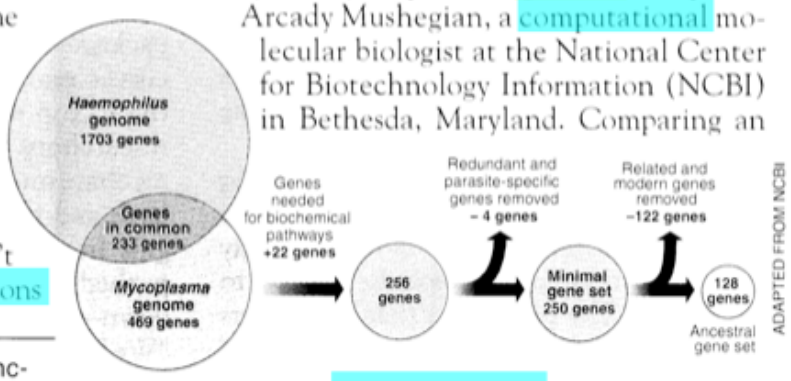
# Topic models: LDA

Latent Dirichlet allocation (LDA) is a popular method for fitting a topic model. It treats each document as a mixture of topics, and each topic as a mixture of words.

## 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

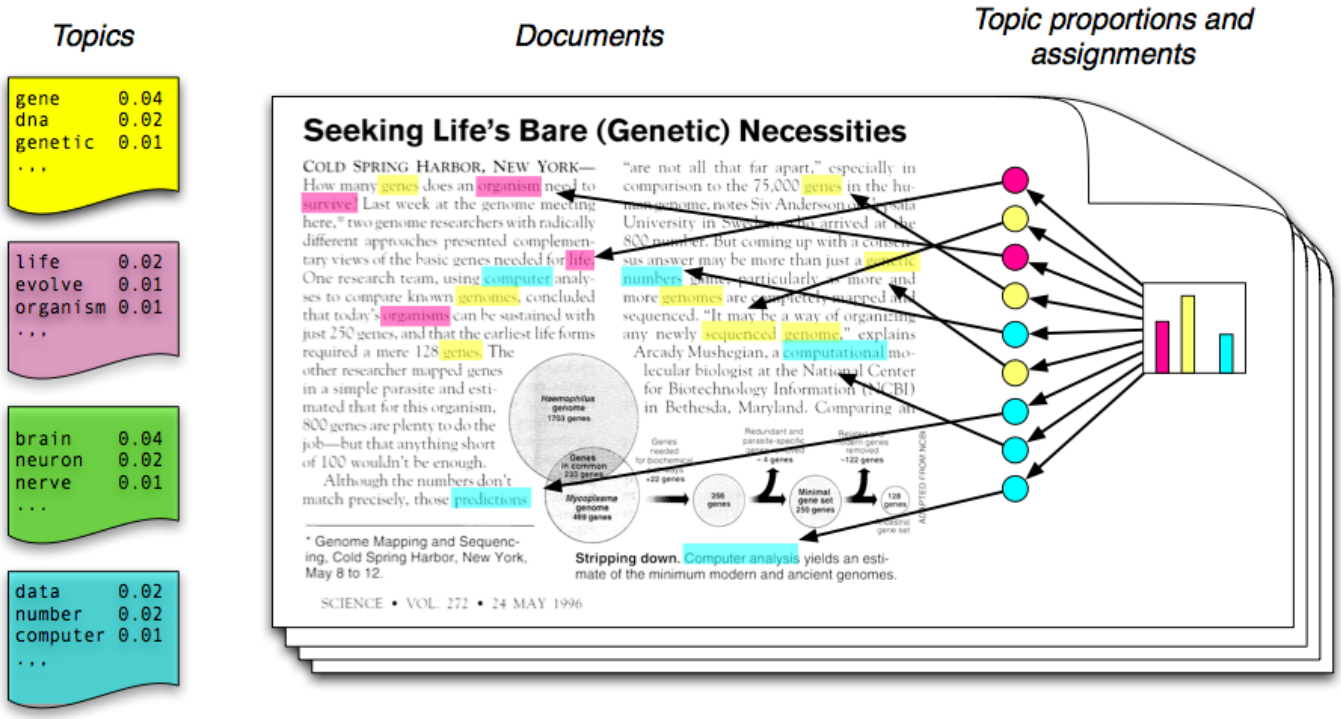



* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

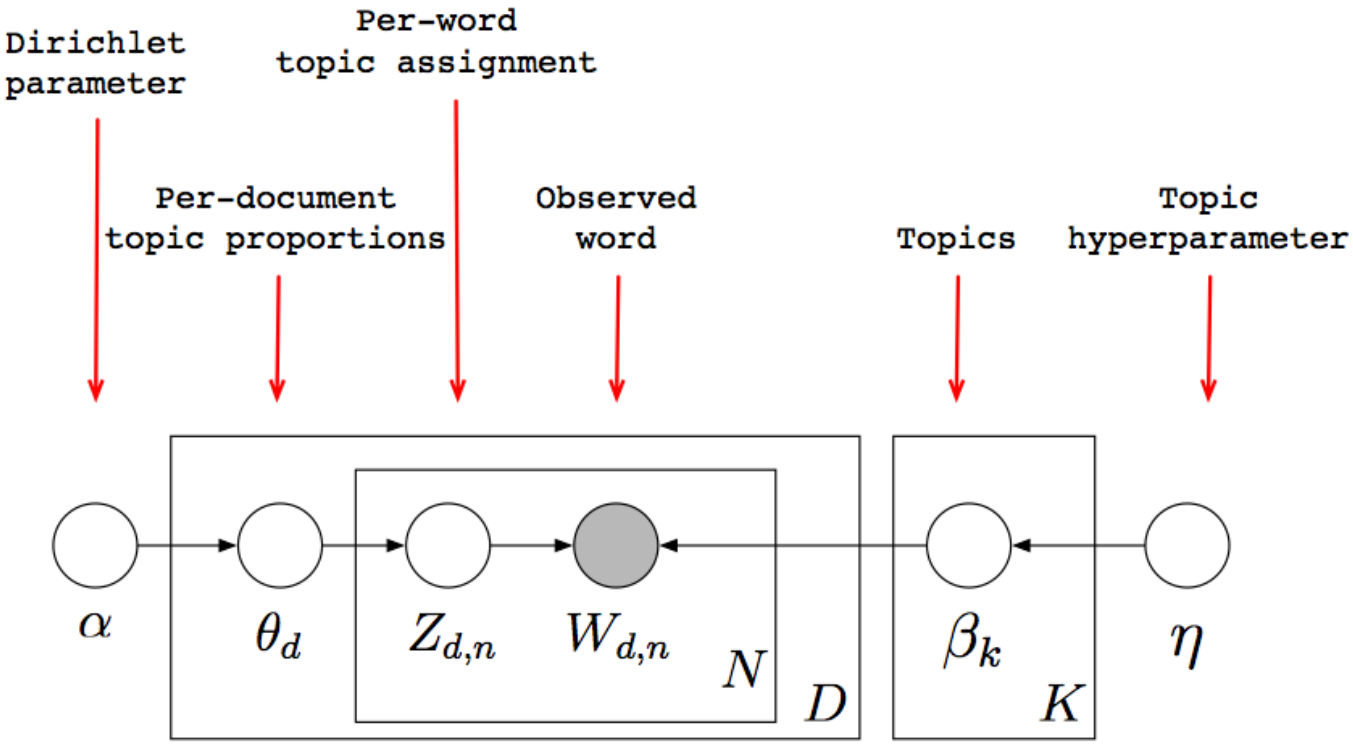
**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

# Topic models: LDA (II)

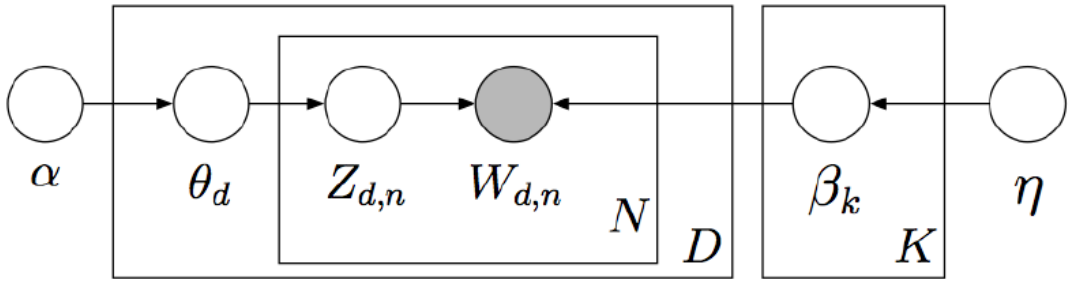


# Topic models: LDA (III)




Each piece of the structure is a random variable.

# Topic models: LDA (IV)




1. Draw each topic $\beta_k \sim \text{Dir}(\eta)$, for $k \in \{1, \ldots, K\}$.
2. For each document:
   1. Draw topic proportions $\theta_d \sim \text{Dir}(\alpha)$.
   2. For each word:
      1. Draw $Z_{d,n} \sim \text{Mult}(\theta_d)$.
      2. Draw $W_{d,n} \sim \text{Mult}(\beta_{Z_{d,n}})$.

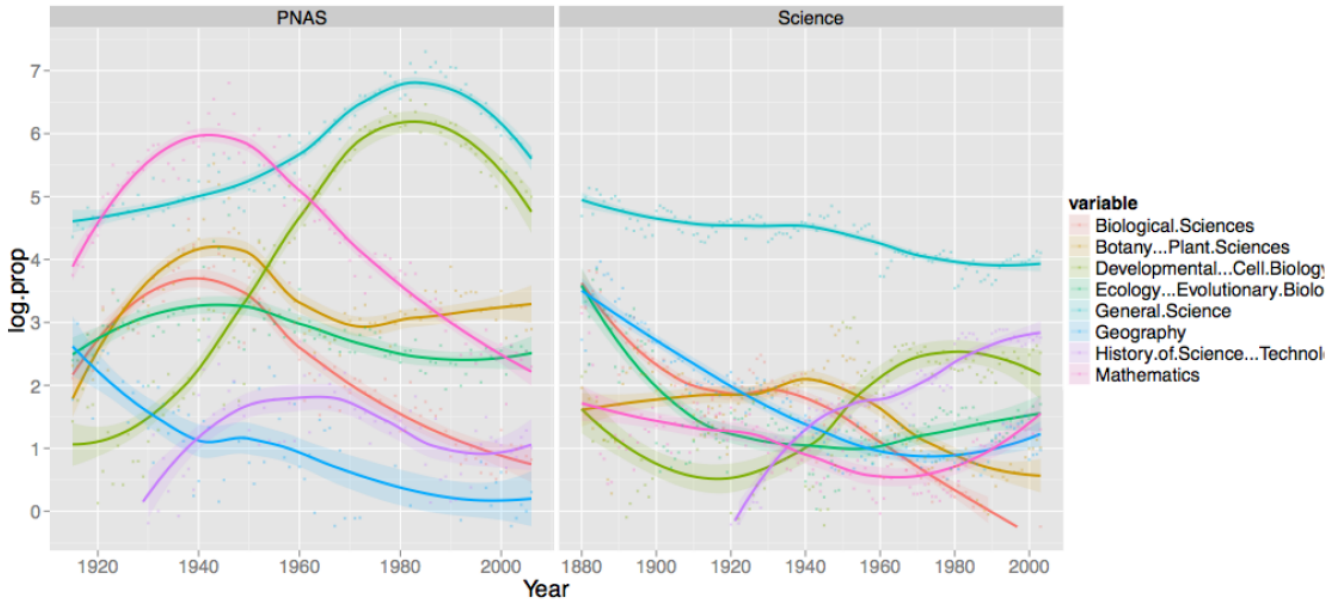
# Topic model: LDA (V)

- Topic models giveth:
  - a probabilistic view of the relationship between W, Z and $\theta$
  - a full statistical framework for learning most aspects of the relationship
- and taketh away:
  - substantive control: You do not get to assert what the topics mean (inevitable when the Z and $\theta$ are both unobserved)

# Topic model: LDA (VI)

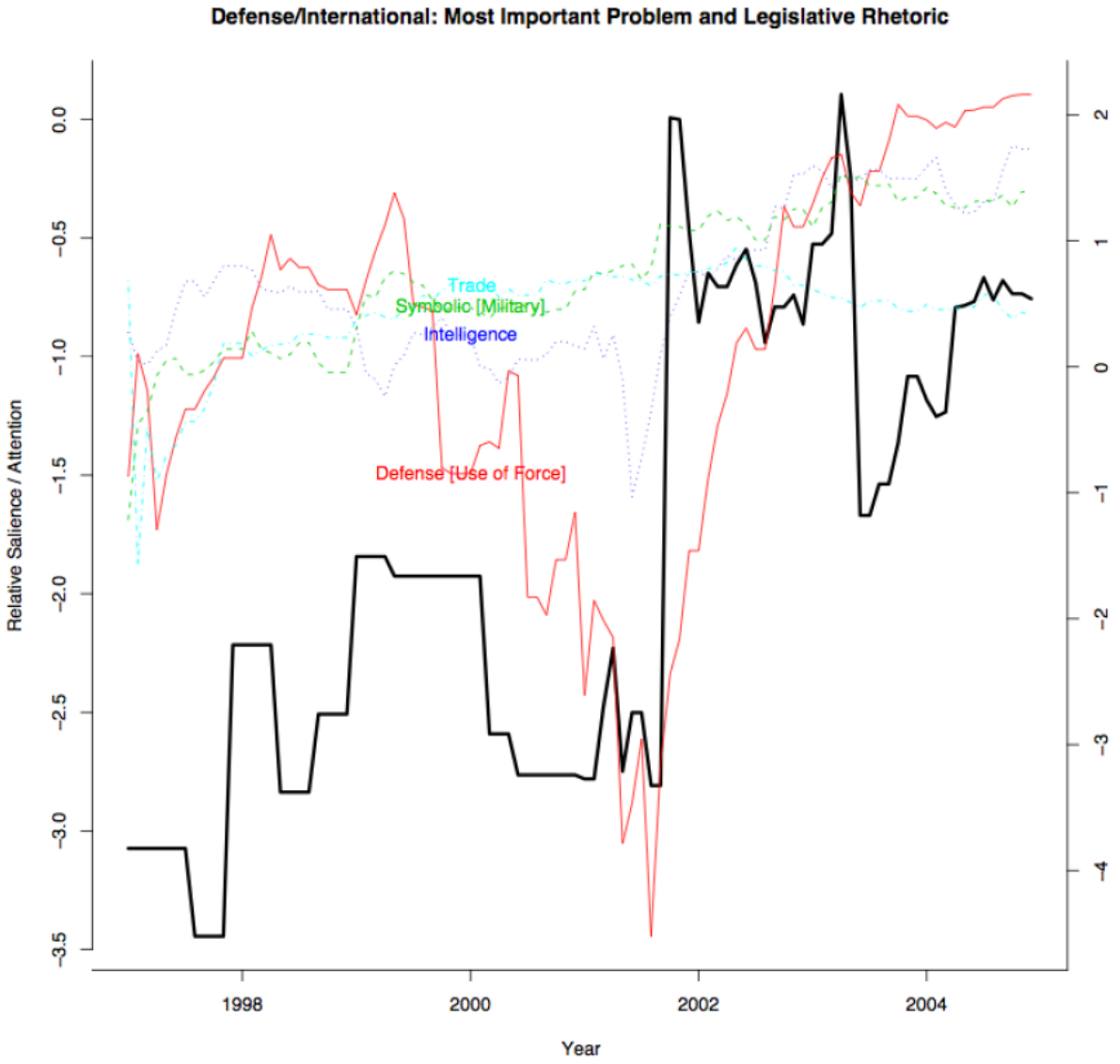
- What control is left?
  - $\alpha$ the parameter of a Dirichlet prior for the distributions (not number) of topics
  - $\eta$ the parameter of a Dirichlet prior over the distributions of words in topics
- Roughly, larger values allow more variation and less sparse representations
- Topic models are admixtures: mixtures of mixtures

# Variations: dynamic topic models (II)




- Quinn et al. analyze 118,065 such congressional speeches from 1997-2004.
- $\theta$ has Markovian dynamics for smooth movement in topic proportions.
- Note: This does not allow variation in the way topics are expressed in words

# Application: policy agenda

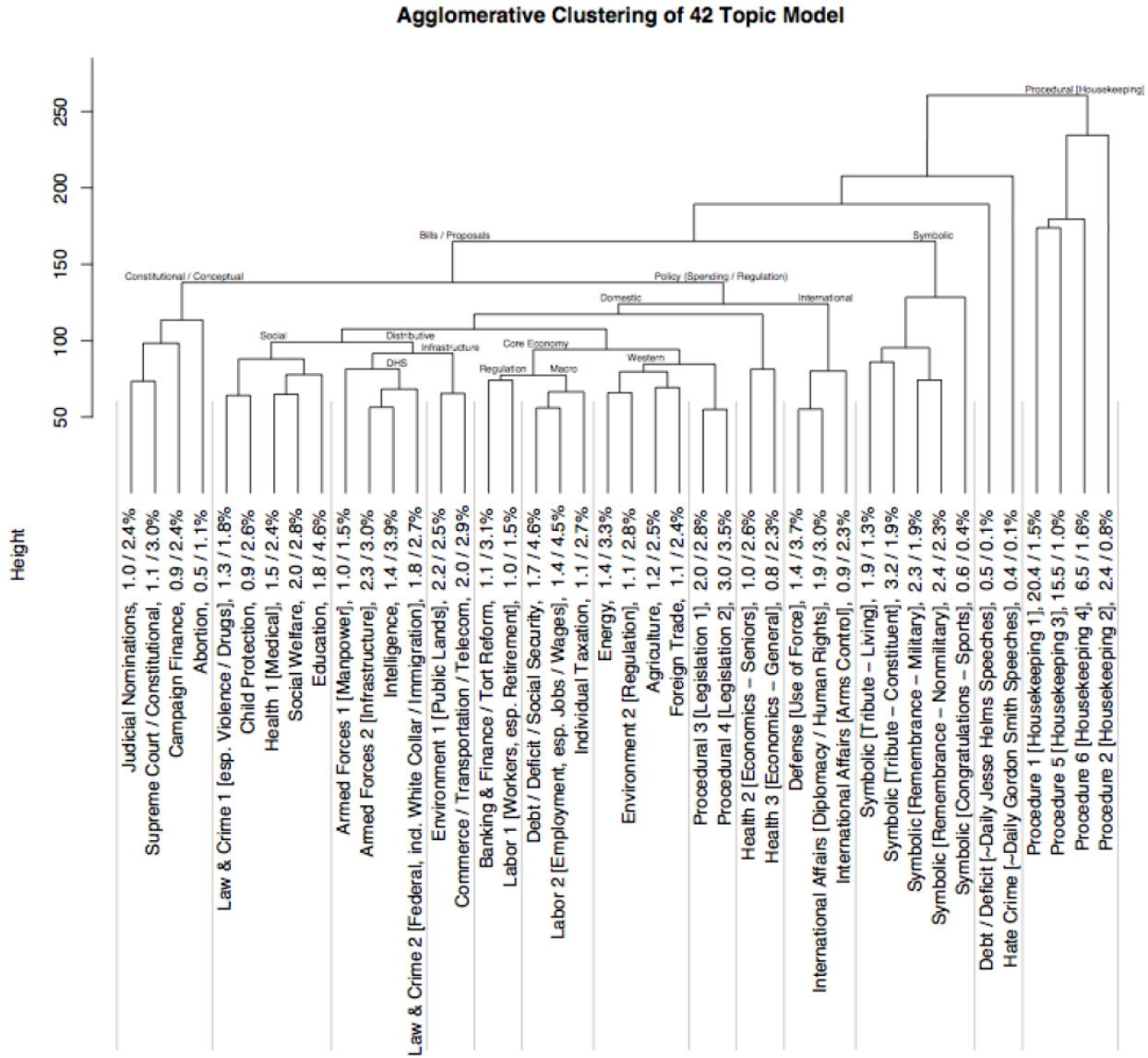


# Defining topics

## Defining Topics

| Topic (Short Label) | Keys |
|---|---|
| 1. Judicial Nominations | *nomine, confirm, nomin, circuit, hear, court, judg, judici, case, vacanc* |
| 2. Constitutional | *case, court, attornei, supreme, justic, nomin, judg, m, decis, constitut* |
| 3. Campaign Finance | *campaign, candid, elect, monei, contribut, polit, soft, ad, parti, limit* |
| 4. Abortion | *procedur, abort, babi, thi, life, doctor, human, ban, decis, or* |
| 5. Crime 1 [Violent] | *enforc, act, crime, gun, law, victim, violenc, abus, prevent, juvenil* |
| 6. Child Protection | *gun, tobacco, smoke, kid, show, firearm, crime, kill, law, school* |
| 7. Health 1 [Medical] | *diseas, cancer, research, health, prevent, patient, treatment, devic, food* |
| 8. Social Welfare | *care, health, act, home, hospit, support, children, educ, student, nurs* |
| 9. Education | *school, teacher, educ, student, children, test, local, learn, district, class* |
| 10. Military 1 [Manpower] | *veteran, va, forc, militari, care, reserv, serv, men, guard, member* |
| 11. Military 2 [Infrastructure] | *appropri, defens, forc, report, request, confer, guard, depart, fund, project* |
| 12. Intelligence | *intellig, homeland, commiss, depart, agenc, director, secur, base, defens* |
| 13. Crime 2 [Federal] | *act, inform, enforc, record, law, court, section, crimin, internet, investig* |
| 14. Environment 1 [Public Lands] | *land, water, park, act, river, natur, wildlif, area, conserv, forest* |
| 15. Commercial Infrastructure | *small, busi, act, highwai, transport, internet, loan, credit, local , capit* |
| 16. Banking / Finance | *bankruptci, bank, credit, case, ir, compani, file, card, financi, lawyer* |
| 17. Labor 1 [Workers] | *worker, social, retir, benefit, plan, act, employ, pension, small, employe* |

# Defining topics (II)



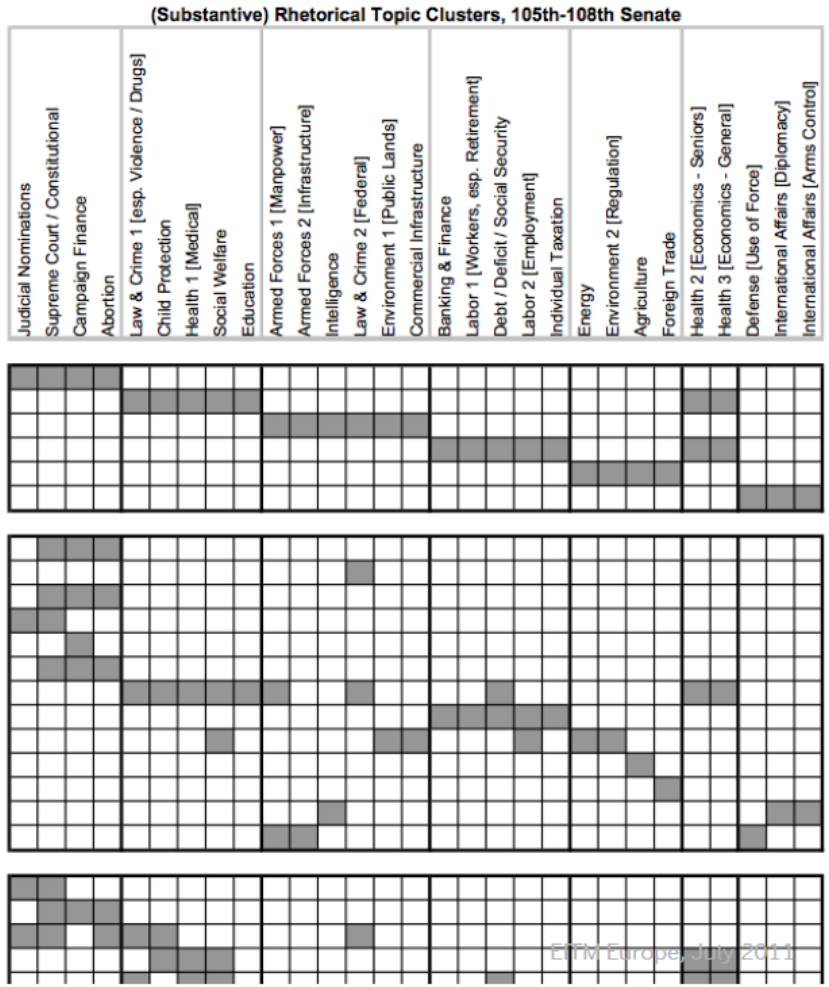
# Defining topics (III)

**Policy Classification Systems**

**Rhetorical Meta-Clusters**
- Constitutional / Conceptual (Partisan conflict)
- Social (Problems)
- Public goods / Distributive (Common good / State v state)
- Economy
- Regional / Rural-urban
- International (Us v them)

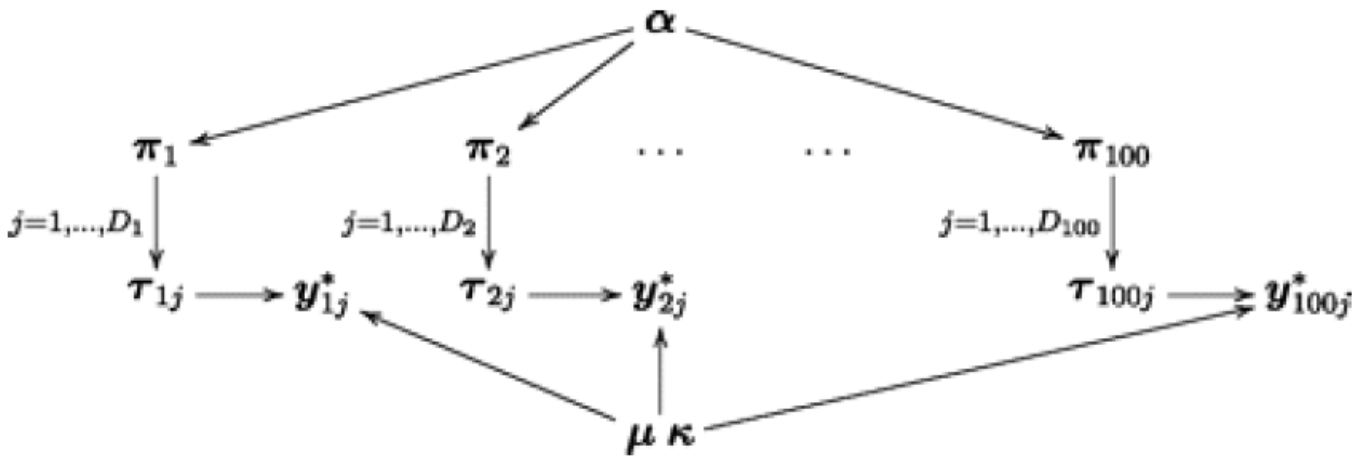
**Katznelson and Lipinksi (2006)**
- Sovereignty [Liberty]
- Sovereignty [Membership / nation]
- Sovereignty [Civil rights]
- Organization / scope [Governmental org.]
- Organization / scope [Representation]
- Organization / scope [Constitutional amendments]
- Domestic affairs [Social policy]
- Domestic affairs [Political economy]
- Domestic affairs [Planning / resources]
- Domestic affairs [Agriculture / food]
- International relations [Int'l political economy]
- International relations [Geopolitics]
- International relations [Defense]

**Jones and Baumgartner (2005)**
- Government operations
- Human services / law [Rights, liberties, minorities]
- Human services / law [Law / crime / family]
- Human services / law [Health]



# Variations: expressed agenda model

In a simpler variation on LDA, Grimmer (2009) defines an expressed agenda model as




- Here there are not multiple topics per press release, but there are observed authors drawn from a population

# Variations: correlated topic models

- The Dirichlet multinomial assumptions hide a constraint about topic covariation
  - LDA cannot represent free covariation of topic proportions
  - The correlated topic model can
- Replace the Dirichlet with a Logistic Normal structure (Aitchison, 1986) with arbitrary covariance matrix

# Topic models in R

- `topicmodels`
  - LDA & CTM
-

# Text analysis workflow