# Menu

Session 0: How could this possibly work?

Session 1: Dictionary-based 'classical' content analysis and topic models

- Content analysis as a model
- Applications
- Measurement error and how to avoid it
- Fun (and trouble) with topic models

Session 2: Classification and evaluation

Session 3: Scaling Models

# Goals

For each class of methods

- You have a good idea of what can, can't and might be able to do with each method
- You know what can go wrong, how to spot it, and how to work around it

# Goals

For each class of methods

- You have a good idea of what can, can't and might be able to do with each method
- You know what can go wrong, how to spot it, and how to work around it

And we'll also have time to try some of them out. . .

Texas A&M, January 2015

# Classical Content Analysis

*Content* is, or is constructed from, *categories* e.g.

human rights, welfare state, national security

Substantively these often have *valence*, e.g.

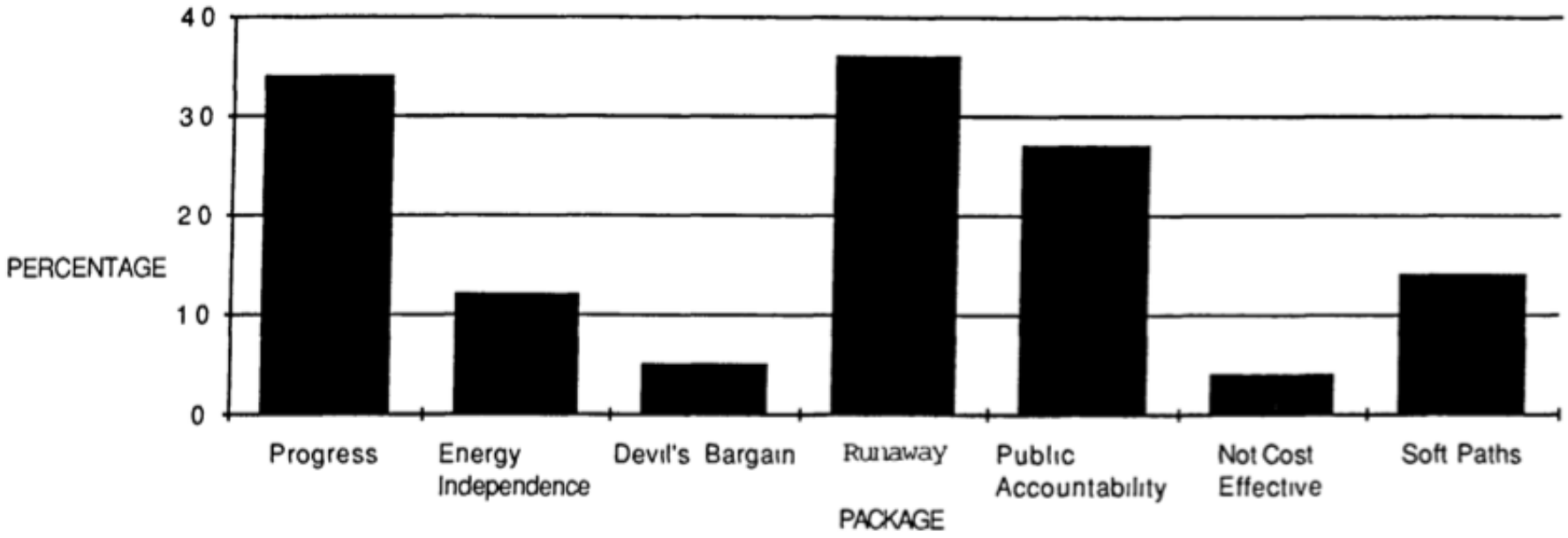
pro-welfare state vs. anti-welfare state, lots of CMP categories

But they are invariably treated as *nominal level* variables

We are typically interested in them for

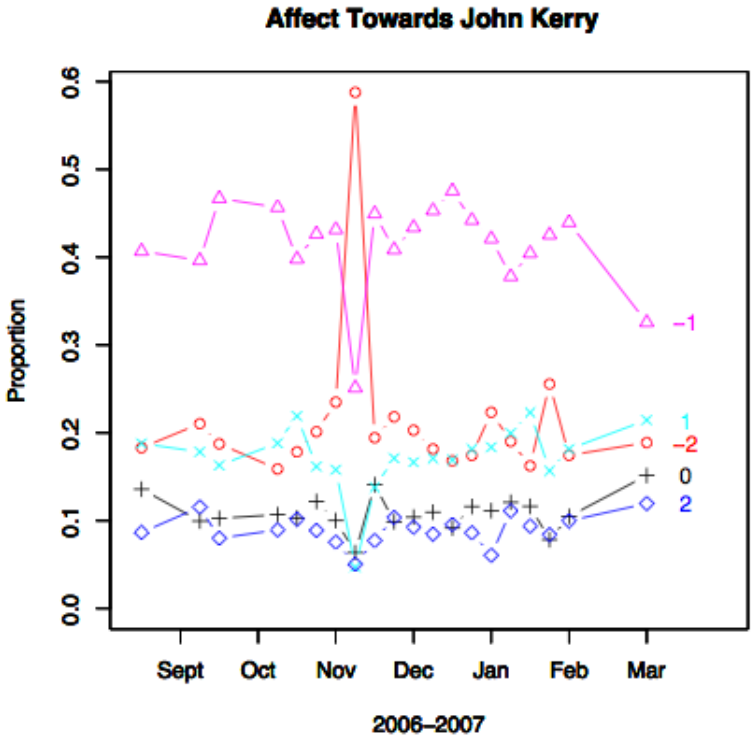
simple descriptions, making comparisons, tracing temporal dynamics

# Talking Like a Newspaper




Gamson and Modigliani (1989)

# Talking like a Candidate




Hopkins and King (2010)

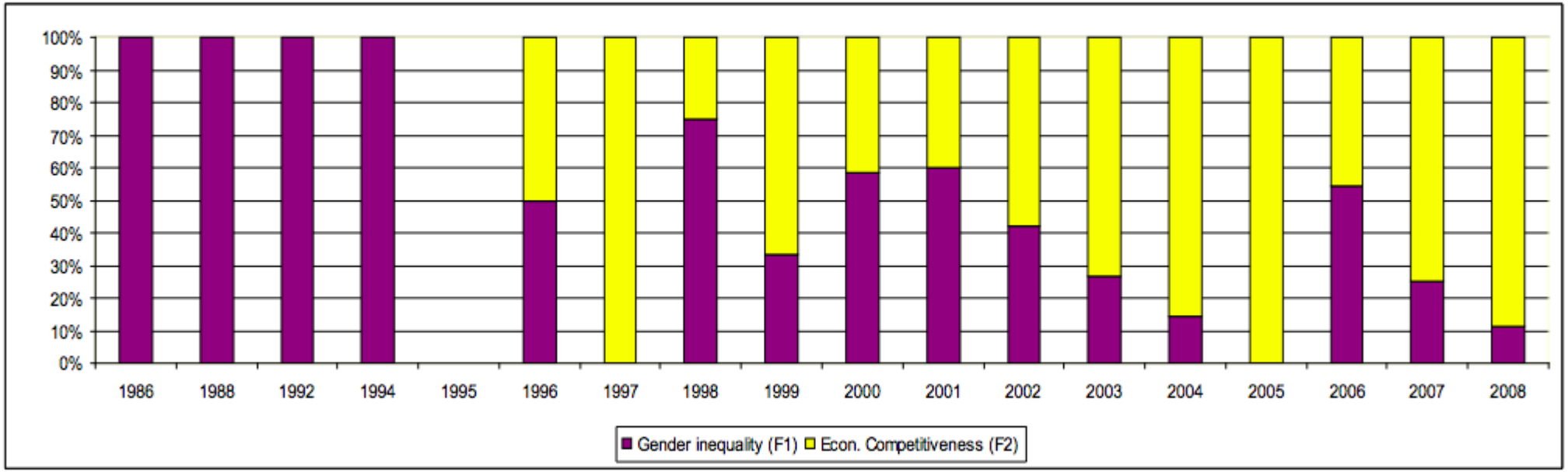
# Talking like a Terrorist

| | Bin Ladin (1988 to 2006) N = 28 | Zawahiri (2003 to 2006) N = 15 | Controls N = 17 | p (two-tailed) |
|---|---|---|---|---|
| Word Count | 2511.5 | 1996.4 | 4767.5 | |
| Big words (greater than 6 letters) | 21.2a | 23.6b | 21.1a | .05 |
| Pronouns | 9.15ab | 9.83b | 8.16a | .09 |
| I (e.g. I, me, my) | 0.61 | 0.90 | 0.83 | |
| We (e.g. we, our, us) | 1.94 | 1.79 | 1.95 | |
| You (e.g. you, your, yours) | 1.73 | 1.69 | 0.87 | |
| He/she (e.g. he, hers, they) | 1.42 | 1.42 | 1.37 | |
| They (e.g., they, them) | 2.17a | 2.29a | 1.43b | .03 |
| Prepositions | 14.8 | 14.7 | 15.0 | |
| Articles (e.g. a, an, the) | 9.07 | 8.53 | 9.19 | |
| Exclusive Words (but, exclude) | 2.72 | 2.62 | 3.17 | |
| Affect | 5.13a | 5.12a | 3.91b | .01 |
| Positive emotion (happy, joy, love) | 2.57a | 2.83a | 2.03b | .01 |
| Negative emotion (awful, cry, hate) | 2.52a | 2.28ab | 1.87b | .03 |
| Anger words (hate, kill) | 1.49a | 1.32a | 0.89b | .01 |
| Cognitive Mechanisms | 4.43 | 4.56 | 4.86 | |
| Time (clock, hour) | 2.40b | 1.89a | 2.69b | .01 |
| Past tense verbs | 2.21a | 1.63a | 2.94b | .01 |
| Social Processes | 11.4a | 10.7ab | 9.29b | .04 |
| Humans (e.g. child, people, selves) | 0.95ab | 0.52a | 1.12b | .05 |
| Family (mother, father) | 0.46ab | 0.52a | 0.25b | .08 |
| Content | | | | |
| Death (e.g. dead, killing, murder) | 0.55 | 0.47 | 0.64 | |
| Achievement | 0.94 | 0.89 | 0.81 | |
| Money (e.g. buy, economy, wealth) | 0.34 | 0.38 | 0.58 | |
| Religion (e.g. faith, Jew, sacred) | 2.41 | 1.84 | 1.89 | |

Note. Numbers are mean percentages of total words per text file. Statistical tests are between Bin Ladin, Zawahiri, and Controls. Documents whose source indicates "Both" (n=3) or "Unknown" (n=2) were excluded due to their small sample sizes.

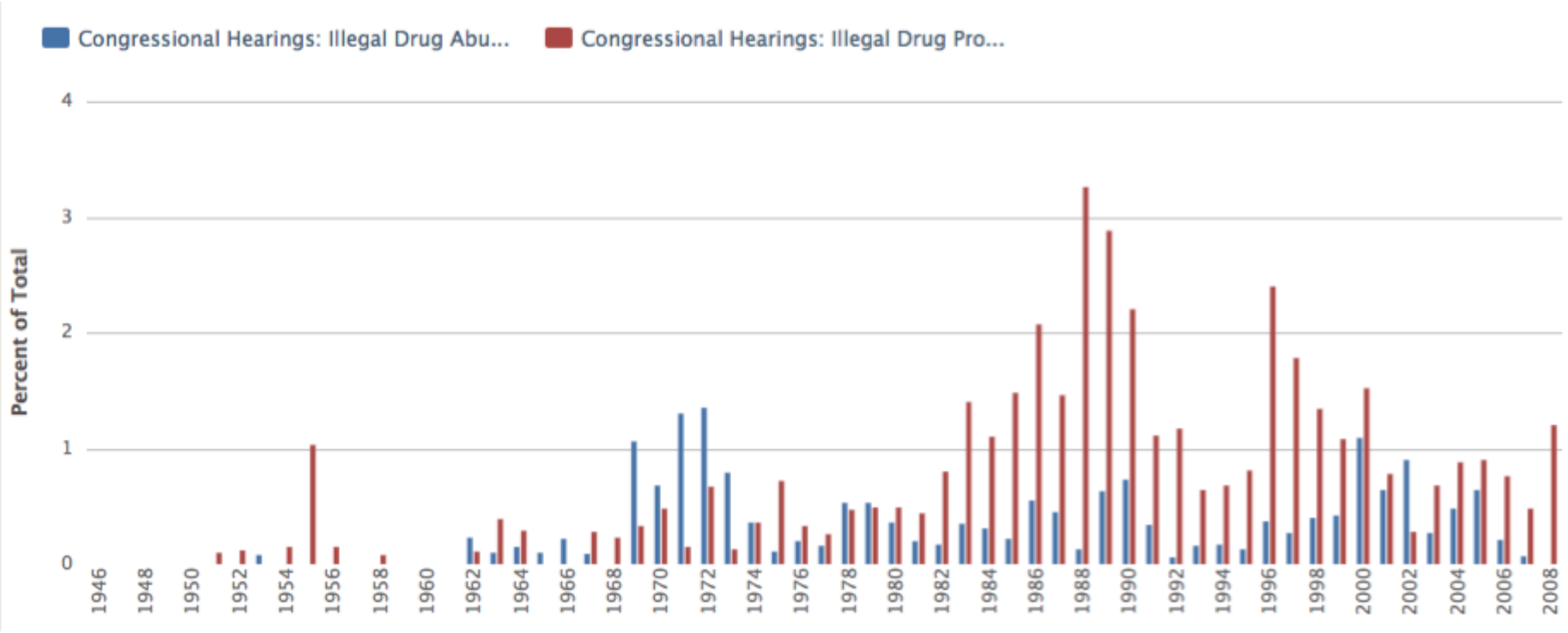
# Talking Like the Commission

*Figure 4.2-2 Relative proportions of policy frames F1 and F2 in secondary EU legislation*




Source: Radulova (2009)

# Talking About Drugs



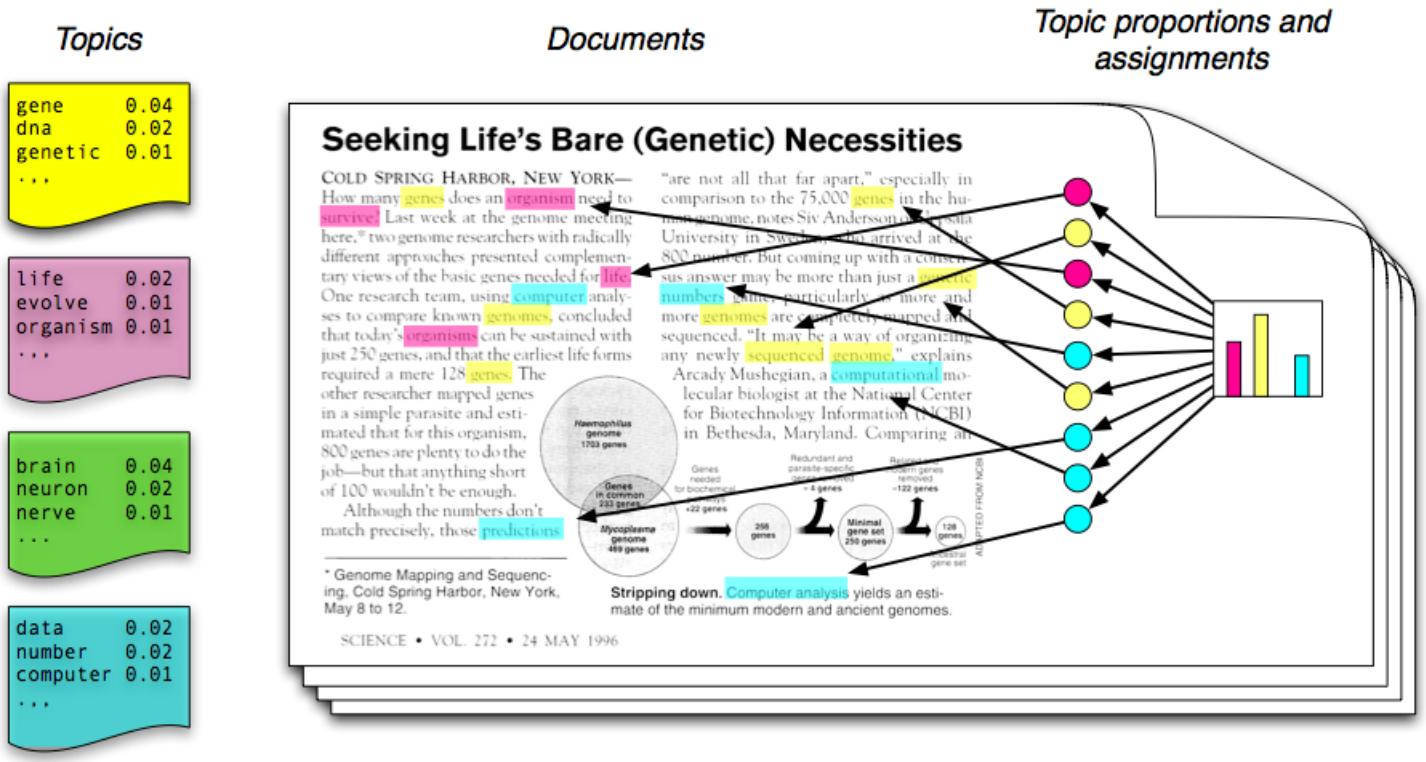

The Congressional Bills Project website (retrieved 2010)

# Classical Content Analysis

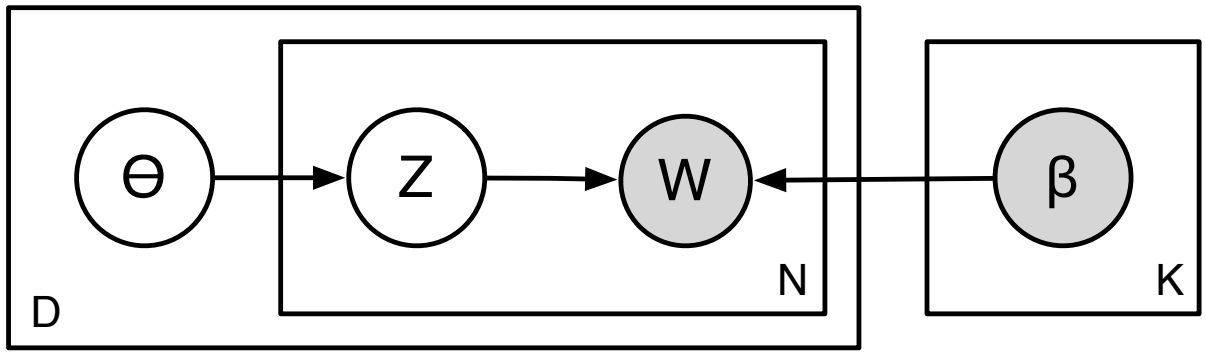
Categories are

equivalence classes over words

representable as assignments of a K-valued category membership variable $Z$ to each word


Topics
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Documents
Topic proportions and assignments
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 basics +22 genes
256 genes
Redundant and parasite-specific genes removed ~6 genes
Minimal gene set 250 genes
Biosynthetic genes removed ~122 genes
128 genes
Minimal gene set
ADAPTED FROM NCBI
* 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

# Classical Content Analysis



# Content Analysis Dictionary

| | |
|---|---|
| ECONOMY | +STATE |
| | accommodation |
| | age |
| | ambulance |
| | assist |
| | . . . |
| | -STATE |
| | choice* |
| | compet* |
| | constrain* |
| | . . . |

from Laver and Garry's (2000) dictionary

# As a posterior: P(Z | W)

Translation: For each word:

| | $P(Z = \textit{state reg} \mid \mathrm{W})$ | $P(Z = \textit{market econ} \mid \mathrm{W})$ |
|---|---|---|
| age | 1 | 0 |
| benefit | 1 | 0 |
| ... | ... | ... |
| assets | 0 | 1 |
| bid | 0 | 1 |
| ... | ... | ... |

# . . . from a underspecified likelihood: P(W | Z)

Note that the *only* way this could be true is if

| | *state reg* | *market econ* |
|---|---|---|
| $P$(age \| $Z$) | a | **0** |
| $P$(benefit \| $Z$) | b | **0** |
| . . . | . . . | . . . |
| $P$(assets \| $Z$) | **0** | c |
| $P$(bid \| $Z$) | **0** | d |
| . . . | . . . | . . . |

where a, b, c, and d $> 0$ and all categories are equally likely: $P(k) = 1/K$

# . . . leading to a posterior over content

Define the category *counts*

$$Z_k = \sum_i^N P(Z = k \mid W_i)$$

and estimate category relative *proportions* using

$$\hat{\theta}_k = \frac{Z_k}{\Sigma_j^K Z_j}$$

When $\theta$ is a set of multinomial parameters, *and the model assumptions are correct*, this could be a reasonable estimator

# Reconstruction

Dictionary-based content analysis was *not* developed this way

> Originally (e.g. Stone 1966) there was no probability model at all

# Connecting CCA content to politics

We're usually interested in category proportions per unit (usually document), e.g.

*How much* of this document is about national defense?

What is the *difference* of aggregated left and aggregated right categories (RILE)

How does the *balance* of human rights and national defense change over time?

# Inference About Content

Statistically speaking, the three types of measures are

- a proportion
- a difference of proportions
- a ratio of proportions

Under certain sampling assumptions we can make inferences about a population

# Inference About Proportions

Example: in the 2001 Labour manifesto there are 872 matches to Laver and Garry's *state reg* category

- 0.029 (nearly 3%) of the document's words
- 0.066 (about 6%) of words that matched *any* categories

The document has 30157 words, so the *first* proportion is estimated as

$$\widehat{\theta}_{state\ reg} = 0.029 \quad [0.027, 0.030]$$

What does this mean?

# Inference About Proportions

Think of the party headquarters repeatedly *drafting* this manifesto

The true proportion – the one suitable to the party's policies – is fixed but every draft is slightly different

The confidence interval reflects the fact that we expect long manifestos to have more precise information about policy

This interval is computed as if every word was a new (conditionally) independent piece of of information

# Reporting: Rates

Don't report proportions if you don't need to.

*Rates/ratios* are more intuitive

e.g. the rate of dictionary matches per $B$ words is

$$\lambda_B = \theta B$$

which is a more interpretable proportion, e.g.

29 times per 1000 words

Different measures correspond to different choices of $B$.

# Ratios: How 'New' was New Labour?

Was the Conservative party in 1992 more or less for state intervention than 'New' Labour in 1997?

Compare instances of *state reg* and *market econ* in the manifestos

| Party | Counts | |
|---|---|---|
| | *state reg* | *market econ* |
| Conservative | 320 | 643 |
| Labour | 396 | 268 |

# Risk Ratios

Compute two *risk ratios*:

$$RR_{\textit{state reg}} = \frac{P(\textit{state reg} \mid \text{cons})}{P(\textit{state reg} \mid \text{lab})}$$

$$RR_{\textit{market econ}} = \frac{P(\textit{market econ} \mid \text{cons})}{P(\textit{market econ} \mid \text{lab})}$$

and 95% confidence intervals

# Interpreting Risk Ratios

If $RR = 1$ then the category occurs at the same rate in labour and conservative manifestos

If $RR = 2$ then the conservative manifesto contains *twice* as much *state reg* language as the labour manifesto

If $RR = .5$ then the conservative manifesto contains *half* as much *state reg* language as the labour manifesto

If the confidence interval for $RR$ contains 1 then we *no evidence* that *state reg* and *market econ* occur at different rates

# Risk Ratios

| | Risk Ratio |
|---|---|
| *market econ* | 1.45 [1.26, 1.67] |
| *state reg* | 0.49 [0.42, 0.57] |

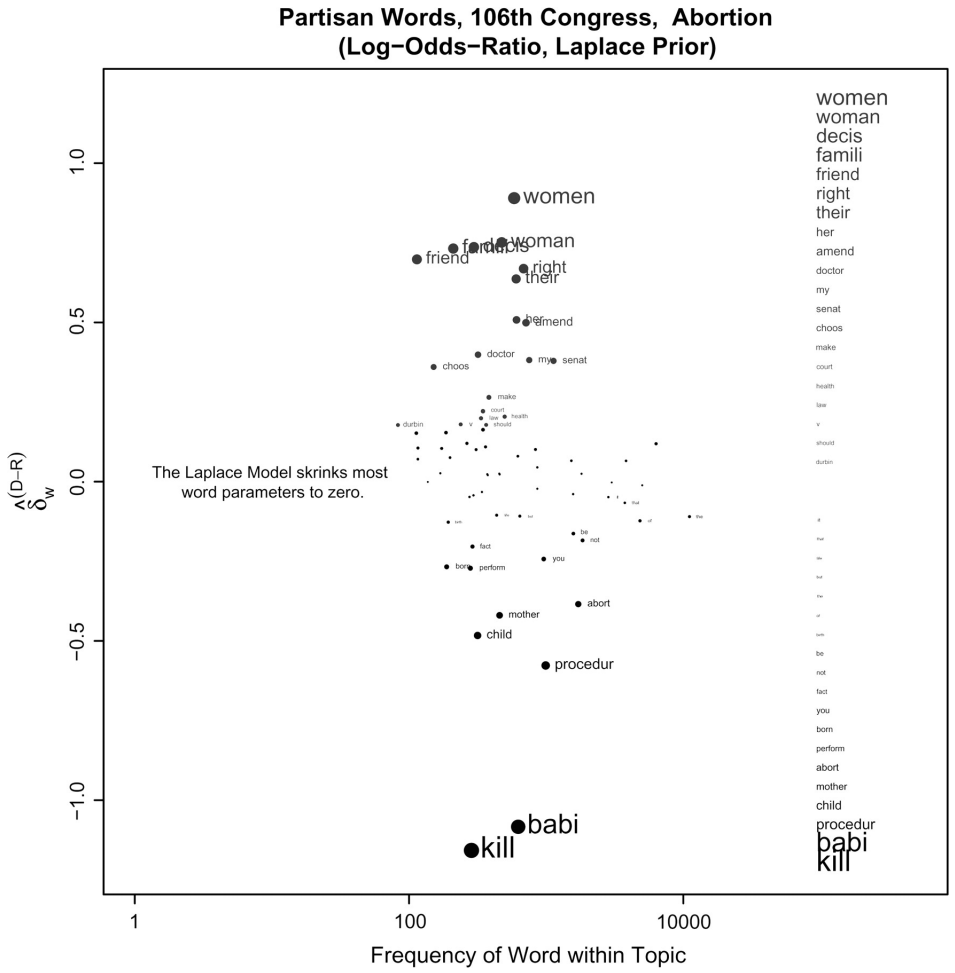
Conservative manifesto generates *market econ* words 45% more often

$$45\% = 100(1.45 - 1)\%$$

Conservative manifesto only generates 49% as many *state reg* words as Labour.

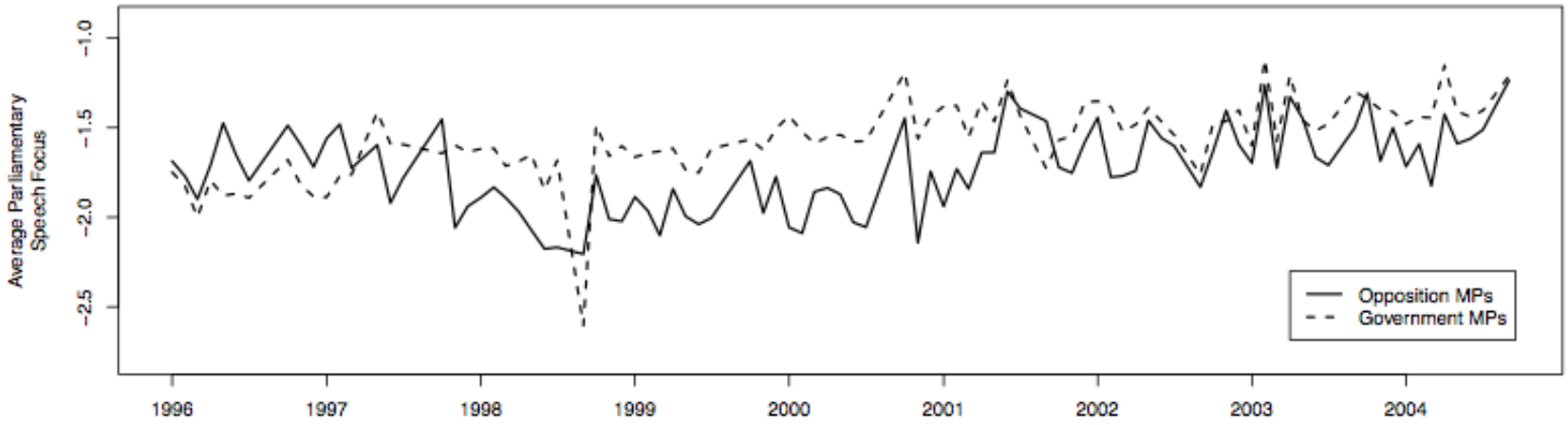
Equivalently Labour generates them about *twice* as often

# Extensions: Regularised log odds



# . . . as dependent variable

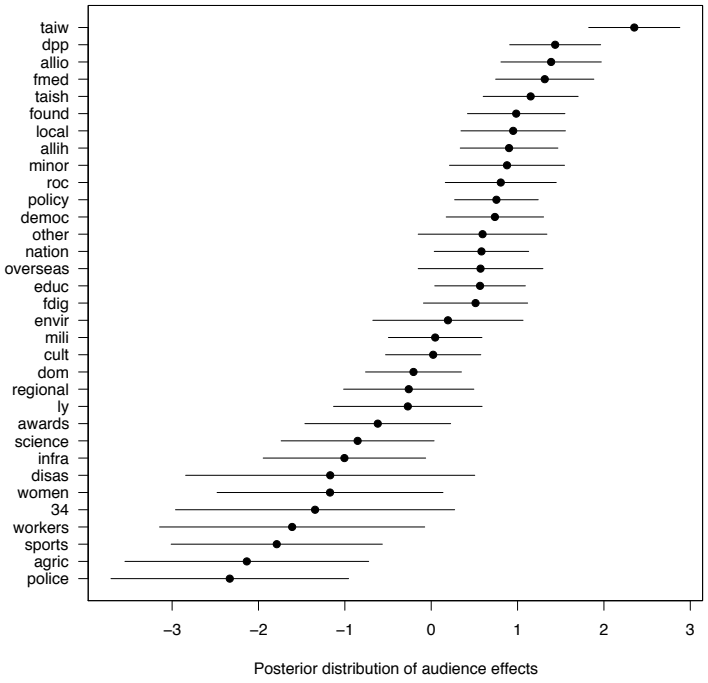
Example: district vs party focus




Data: [*district words*, *party words*] (Kellerman & Proksch, MS)

Here, a *logged ratio* of two categories

# Content as something to explain (Sullivan and Lowe 2007)




*independence words* ← audience + offset(doc length)

# What to report?

Not all choices are constant or comparable. . .

| Quantity | Comparability |
|---|---|
| Count | No |
| Proportion of words | Yes |
| Proportion of category matches | Yes |
| Rate per B words | Yes |
| Rate per sentence / paragraph | No |
| Difference of proportions | Yes |
| (Logged) ratio of counts | Yes |
| Significance of tests | No! |

(if length is uninformative & category probability constant)

# OK, how do I make such an instrument?

Maximise measurement validity

Minimise *measurement error*

(Sell high, buy low)

Texas A&M, January 2015

# Measurement Error

Measurement error in classical content analysis is primarily failure of *this* assumption:

| | $P(Z = \textit{state reg} \mid \mathrm{W})$ | $P(Z = \textit{market econ} \mid \mathrm{W})$ |
|---|---|---|
| age | 1 | 0 |
| benefit | 1 | 0 |
| ... | ... | ... |
| assets | 0 | 1 |
| bid | 0 | 1 |
| ... | ... | ... |

# Measurement Error

What are the effects of measurement error in category counts?

Being directly wrong, e.g.

My estimated rates are too *low* (bias)
Some of my estimates are more biased than others

Being indirectly wrong, e.g.

My carefully constructed left-right measure is too *centrist*
My effect sizes appear to be much too small

# Measurement Error

Assume

Vocabulary of only two words 'benefit' and 'assets'

a *subtractive* measure of position:
$Z_{market\ econ} - Z_{state\ reg}$

Then

| | $P(Z = state\ reg \mid \mathrm{W})$ | $P(Z = market\ econ \mid \mathrm{W})$ |
|---|---|---|
| benefit | 1 | 0 |
| assets | 0 | 1 |

# Measurement Error

implies the **confusion matrix**

| | | True | |
|---|---|---|---|
| | | *state reg* | *market econ* |
| Observed | *state reg* | 1 | 0 |
| | *market econ* | 0 | 1 |

# Measurement Error

Example: What if $P(W \mid Z)$ slips to

| | *state reg* | *market econ* |
|---|---|---|
| $P$(benefit \| Z) | 0.7 | 0.2 |
| $P$(assets \| Z) | 0.3 | 0.8 |

P(W=‘asset’ | Z=*state reg*) $>$ 0 so

P(Z=*state reg* | W=‘asset’) $<$ 1

# Measurement Error

When $Z_{market\ econ} = 10$ and $Z_{state\ reg} = 20$ the true difference is

(10-20)/(10+20) = -0.33

Under the perfect measurement model this would be realised (on average) as

20 'benefit's and 10 'assets'

# Measurement Error

Under our *imperfect* measurement it is realised (on average) as

16 'benefit's (14 from *state reg* but 2 from *market econ*) and

14 'assets' (8 from *market econ* but 6 from *state reg*)

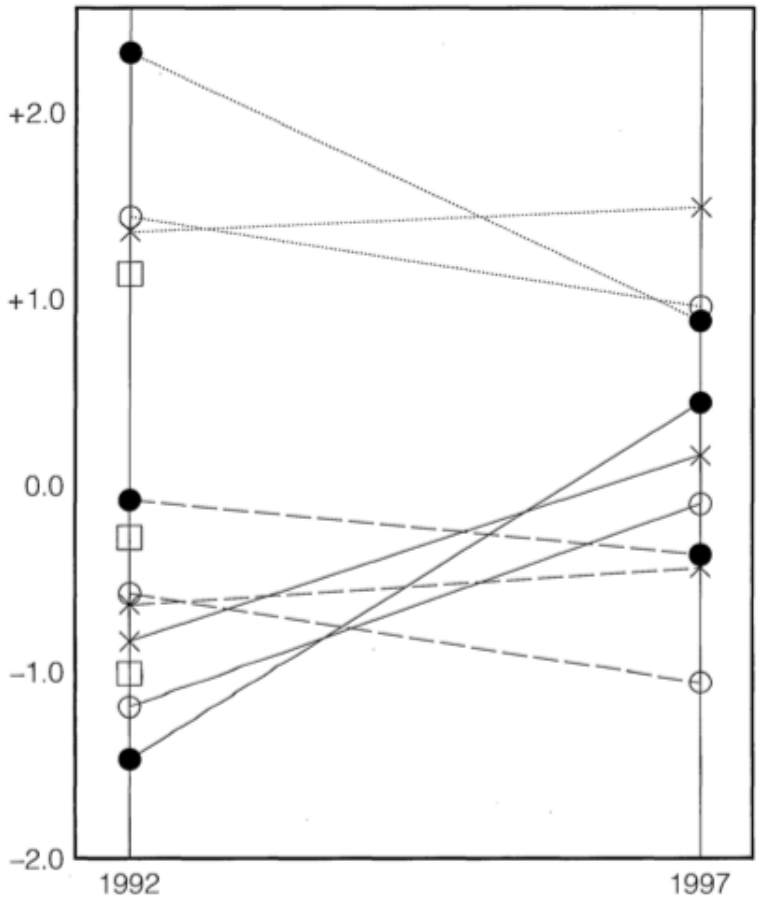
# Measurement Error

The proportional difference measure is now (14-16)/(14+16) = -0.07

Apparently much closer to the centre, but only because of measurement error

*All* relative measures will have this problem

# In Action (Laver and Garry 2000)



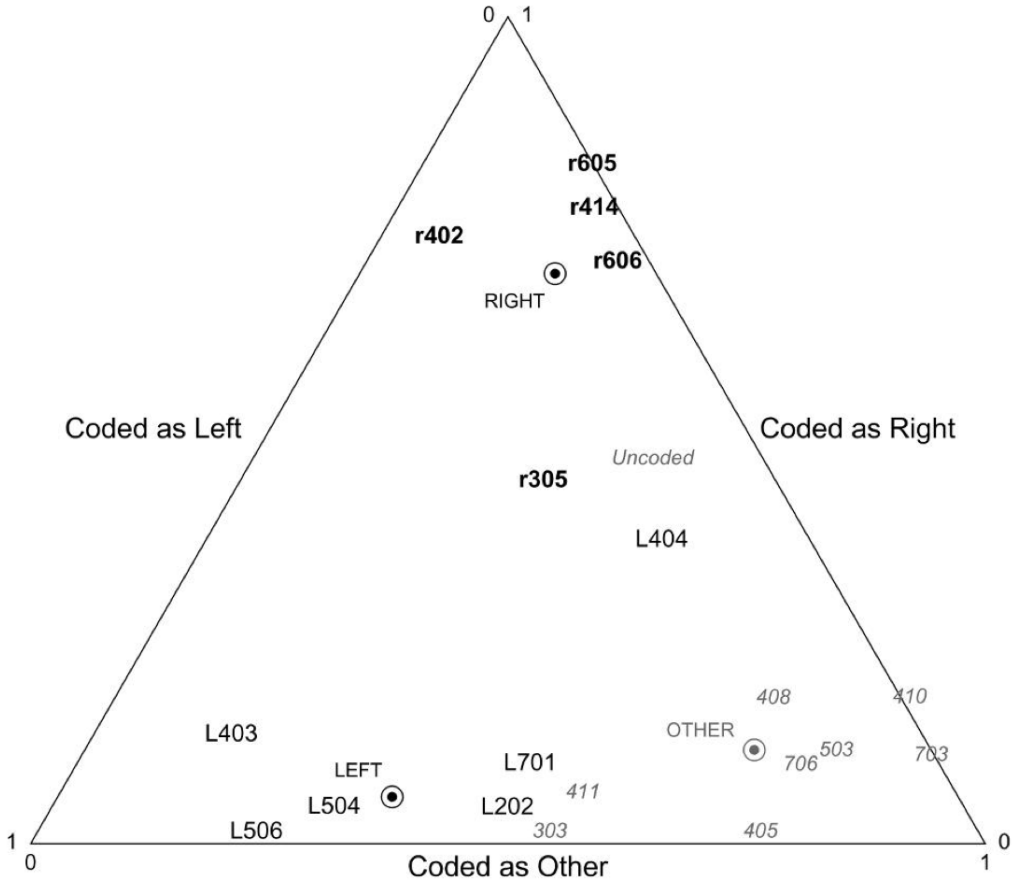
# In Action (Mikhaylov et al. 2012)

**Table 3** Misclassification matrix for true versus observed Rile

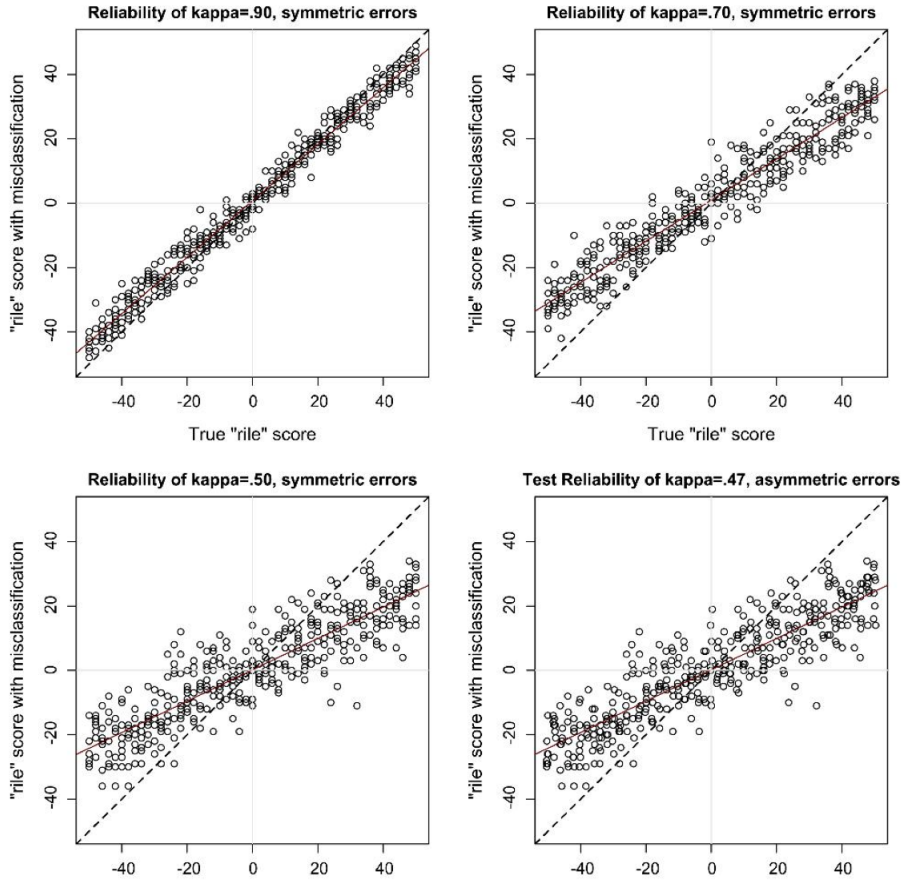
| | | *True Rile category* | | | |
|---|---|---|---|---|---|
| | | *Left* | *None* | *Right* | *Total* |
| | Left | 430 | 188 | 100 | 718 |
| | | **0.59** | 0.19 | 0.11 | |
| Coded | None | 254 | 712 | 193 | 1159 |
| Rile | | 0.35 | **0.70** | 0.20 | |
| | Right | 41 | 115 | 650 | 806 |
| | | 0.06 | 0.11 | **0.69** | |
| | Total | 725 | 1015 | 943 | 1668 |
| | False negative rate | 0.41 | 0.30 | 0.31 | |
| | False positive rate | 0.15 | 0.27 | 0.09 | |

*Note*. The top figure in each cell is the raw count; the bottom figure is the column proportion. The figures are empirically computed from combined British and New Zealand manifesto tests. The false negative rate is 1—sensitivity, whereas the false positive rate is 1—specificity.

# In Action (Mikhaylov et al. 2012)



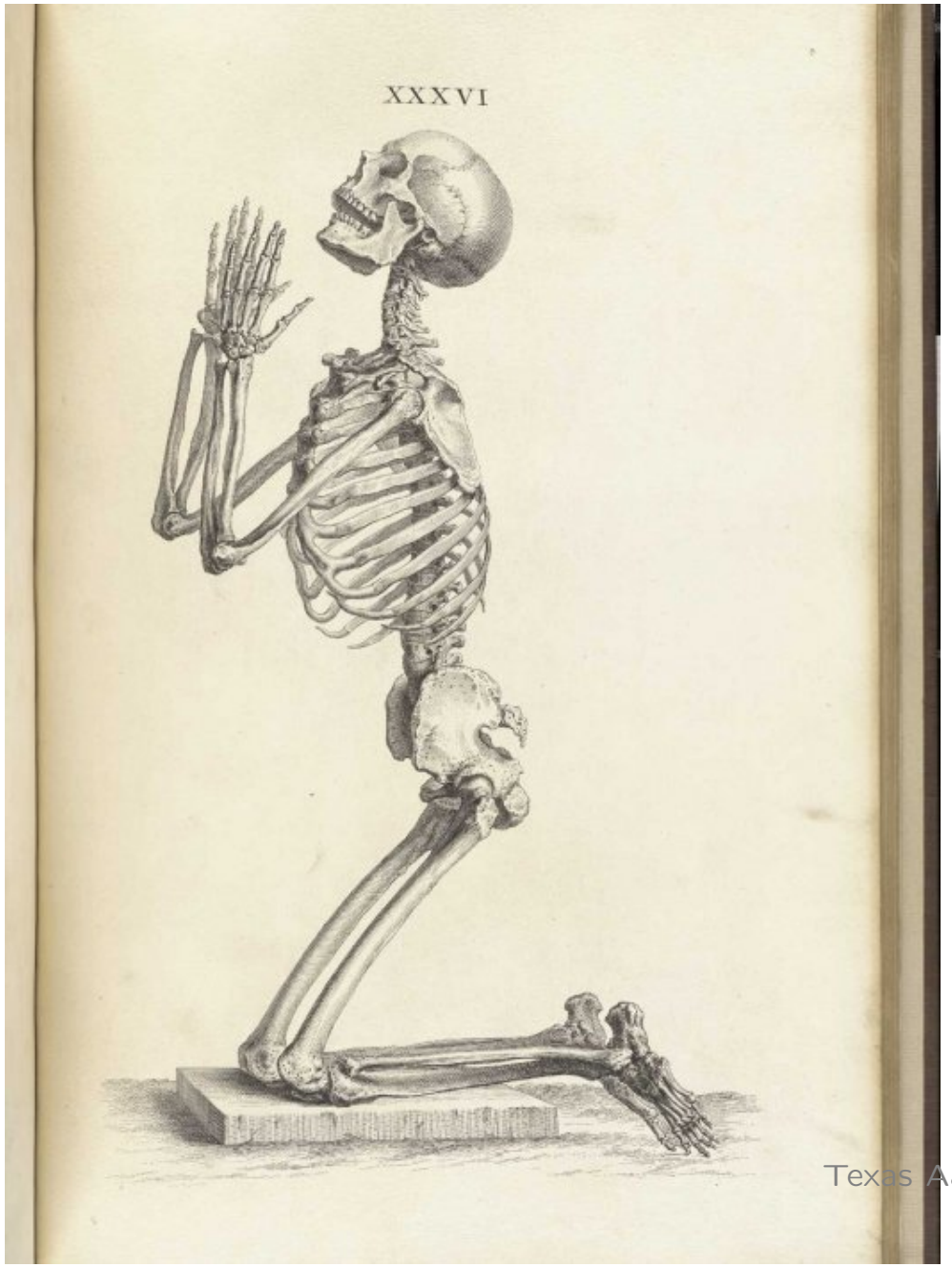
# In Action (Mikhaylov et al. 2012)



Texas A&M, January 2015

# Solutions

Texas A&M, January 2015


XXXVI

# Solutions: Try to Avoid it

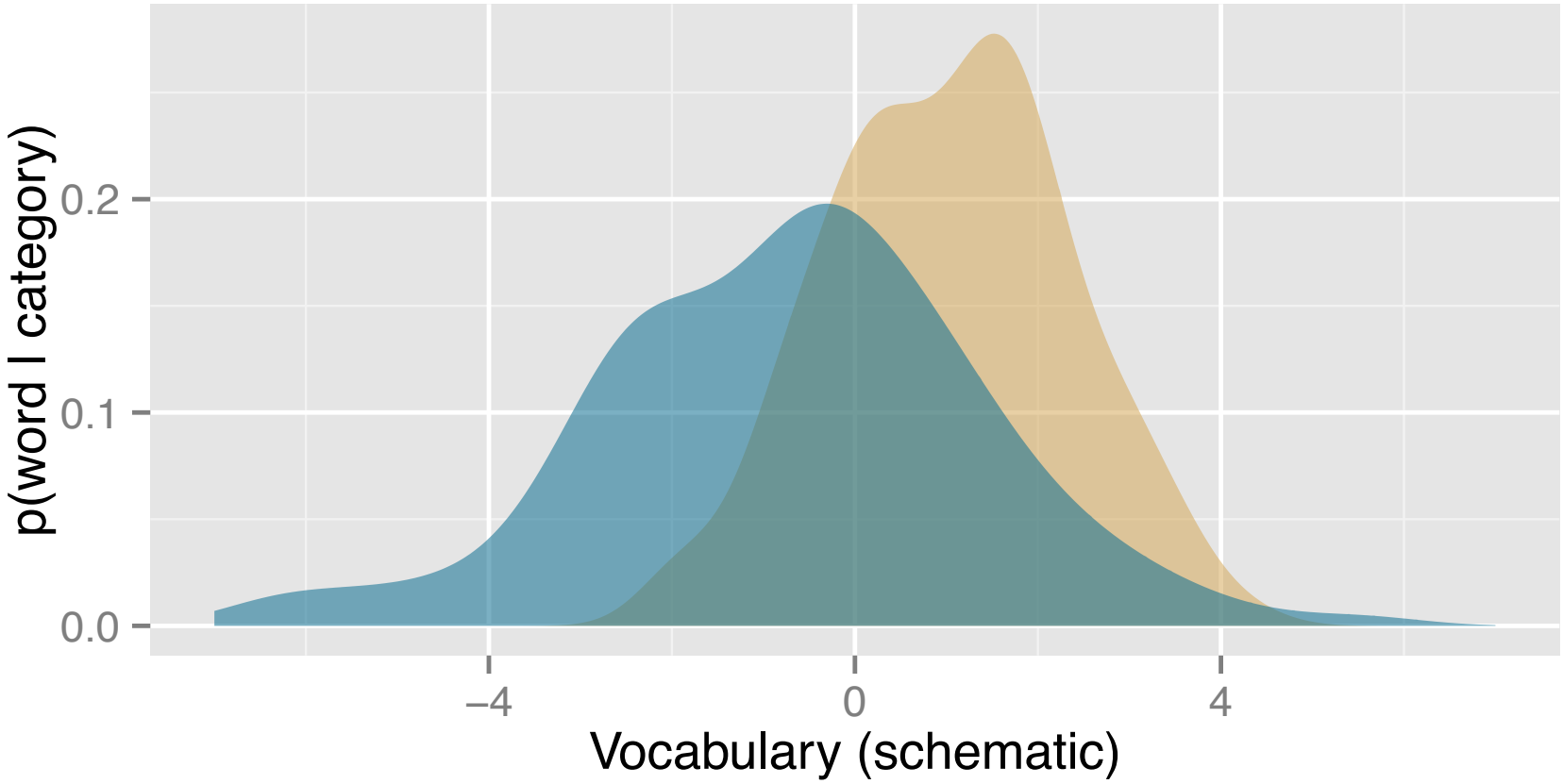
A non-intuitive fact about content dictionaries:

**Precision**: proportion of words used the way your dictionary assumes

**Recall**: proportion of words used that way that are in your dictionary

*always* trade-off. . .

# Intuition: Precision and Recall



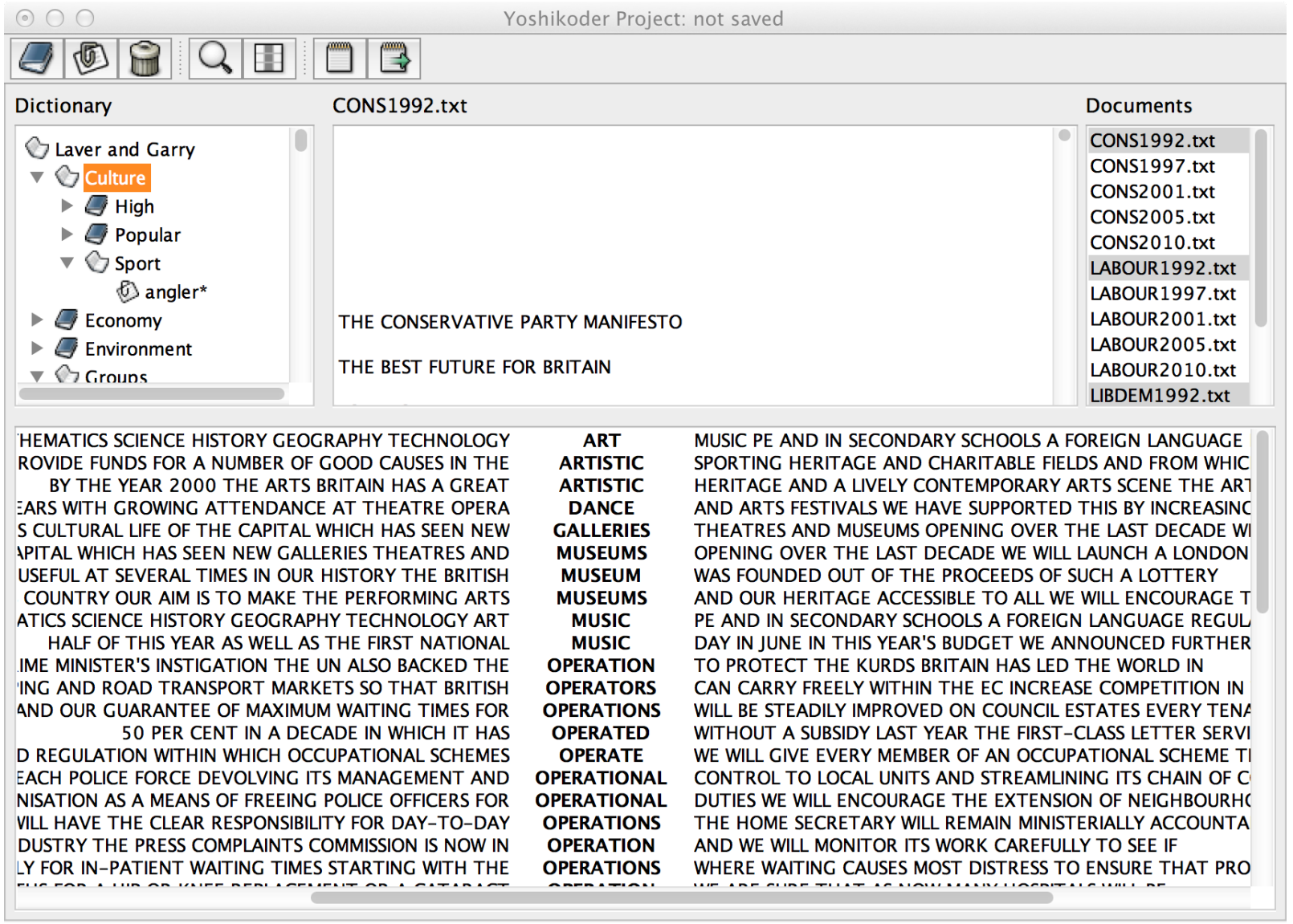
# Solutions: Try to Avoid it

Keyword in context analyses allow you to scan all contexts of a word

- How many of them are the sense or usage you want?

Compromise: KWIC analyses allow you to scan all contexts of a word

- What proportion of tokens are the sense or usage you want?


Yoshikoder Project: not saved
Dictionary
CONS1992.txt
Documents
Laver and Garry
Culture
High
Popular
Sport
angler*
Economy
Environment
Groups
THE CONSERVATIVE PARTY MANIFESTO
THE BEST FUTURE FOR BRITAIN
CONS1992.txt
CONS1997.txt
CONS2001.txt
CONS2005.txt
CONS2010.txt
LABOUR1992.txt
LABOUR1997.txt
LABOUR2001.txt
LABOUR2005.txt
LABOUR2010.txt
LIBDEM1992.txt
HEMATICS SCIENCE HISTORY GEOGRAPHY TECHNOLOGY ART MUSIC PE AND IN SECONDARY SCHOOLS A FOREIGN LANGUAGE
ROVIDE FUNDS FOR A NUMBER OF GOOD CAUSES IN THE ARTISTIC SPORTING HERITAGE AND CHARITABLE FIELDS AND FROM WHIC
BY THE YEAR 2000 THE ARTS BRITAIN HAS A GREAT ARTISTIC HERITAGE AND A LIVELY CONTEMPORARY ARTS SCENE THE ART
EARS WITH GROWING ATTENDANCE AT THEATRE OPERA DANCE AND ARTS FESTIVALS WE HAVE SUPPORTED THIS BY INCREASINC
S CULTURAL LIFE OF THE CAPITAL WHICH HAS SEEN NEW GALLERIES THEATRES AND MUSEUMS OPENING OVER THE LAST DECADE WI
APITAL WHICH HAS SEEN NEW GALLERIES THEATRES AND MUSEUMS OPENING OVER THE LAST DECADE WE WILL LAUNCH A LONDON
USEFUL AT SEVERAL TIMES IN OUR HISTORY THE BRITISH MUSEUM WAS FOUNDED OUT OF THE PROCEEDS OF SUCH A LOTTERY
COUNTRY OUR AIM IS TO MAKE THE PERFORMING ARTS MUSEUMS AND OUR HERITAGE ACCESSIBLE TO ALL WE WILL ENCOURAGE T
ATICS SCIENCE HISTORY GEOGRAPHY TECHNOLOGY ART MUSIC PE AND IN SECONDARY SCHOOLS A FOREIGN LANGUAGE REGUL
HALF OF THIS YEAR AS WELL AS THE FIRST NATIONAL MUSIC DAY IN JUNE IN THIS YEAR'S BUDGET WE ANNOUNCED FURTHER
IME MINISTER'S INSTIGATION THE UN ALSO BACKED THE OPERATION TO PROTECT THE KURDS BRITAIN HAS LED THE WORLD IN
ING AND ROAD TRANSPORT MARKETS SO THAT BRITISH OPERATORS CAN CARRY FREELY WITHIN THE EC INCREASE COMPETITION IN
AND OUR GUARANTEE OF MAXIMUM WAITING TIMES FOR OPERATIONS WILL BE STEADILY IMPROVED ON COUNCIL ESTATES EVERY TENA
50 PER CENT IN A DECADE IN WHICH IT HAS OPERATED WITHOUT A SUBSIDY LAST YEAR THE FIRST-CLASS LETTER SERVI
D REGULATION WITHIN WHICH OCCUPATIONAL SCHEMES OPERATE WE WILL GIVE EVERY MEMBER OF AN OCCUPATIONAL SCHEME TI
EACH POLICE FORCE DEVOLVING ITS MANAGEMENT AND OPERATIONAL CONTROL TO LOCAL UNITS AND STREAMLINING ITS CHAIN OF C
NISATION AS A MEANS OF FREEING POLICE OFFICERS FOR OPERATIONAL DUTIES WE WILL ENCOURAGE THE EXTENSION OF NEIGHBOURHC
VILL HAVE THE CLEAR RESPONSIBILITY FOR DAY-TO-DAY OPERATIONS THE HOME SECRETARY WILL REMAIN MINISTERIALLY ACCOUNTA
DUSTRY THE PRESS COMPLAINTS COMMISSION IS NOW IN OPERATION AND WE WILL MONITOR ITS WORK CAREFULLY TO SEE IF
LY FOR IN-PATIENT WAITING TIMES STARTING WITH THE OPERATIONS WHERE WAITING CAUSES MOST DISTRESS TO ENSURE THAT PRO

# Recovering from Measurement Error: Model it (1)

We can recover from measurement error if we know enough about it (Hopkins and King 2010, King and Liu 2007)

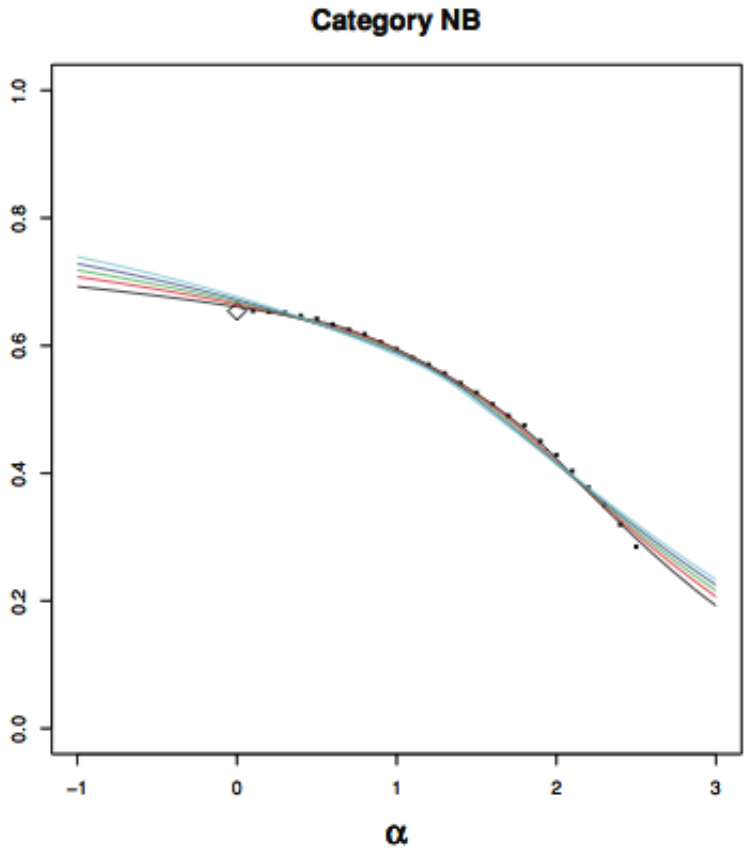
Coder training provides information about how a category proportion estimate changes as coders get better

Intuition:

> Model this trajectory, then extrapolate it forwards. . .

In regression contexts this is called SIMEX (Cook and Stefanski, 1995).

# Recovering from Measurement Error: Model It (1)



# Recovering from Measurement Error: Model It (2)

Under measurement error

- A observed category proportions are generated by a *mixture* of categories
- The weights for this mixture are the true category proportions

Given the confusion matrix (or the true generation probabilities), we can *infer* the true proportions

# Recovering from Measurement Error: Model It (2)

Intuition

$$P(W) = \sum_{k}^{K} P(W \mid Z = k) P(Z = k)$$

has the form

$$Y = X\theta$$

$$\begin{bmatrix} 0.53 \\ 0.46 \end{bmatrix} = \begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

is solved as [0.67, 0.33]

# Recovering from Measurement Error: Model It (2)

Applied to Mikhaylov error data:

If $(P, T)$ is

| | L | N | R |
|---|---|---|---|
| L | 430.00 | 188.00 | 100.00 |
| N | 254.00 | 712.00 | 193.00 |
| R | 41.00 | 115.00 | 650.00 |

So $P(C \mid T)$ is

| | L | N | R |
|---|---|---|---|
| L | 0.59 | 0.19 | 0.11 |
| N | 0.35 | 0.70 | 0.20 |
| R | 0.06 | 0.11 | 0.69 |

Implication:

If [L, N, R] = [20, 0, 10]

we would expect to see about [13, 9, 8]

Invert $P(C \mid T)$:

| | L | N | R |
|---|---|---|---|
| L | 2.00 | -0.50 | -0.16 |
| N | -1.00 | 1.75 | -0.37 |
| R | 0.00 | -0.25 | 1.52 |

and multiply to get an estimate of the true counts. . .

Example:

Given counts [13, 9, 8] we get

[L 20.19, -0.16, 9.98] $\approx$ [20, 0, 10]

# Notes:

This is all *in expectation*.

We are ignoring measurement error *in the error matrix*

This method may violate sensible prior information

Works for anything that makes errors (human or machine)

# Recovering from Measurement Error: Model It (3)

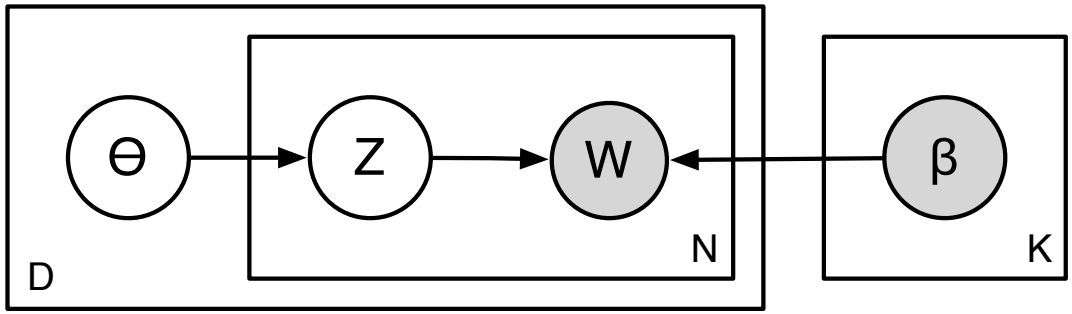
Topic models, e.g. Latent Dirichlet Allocation (Blei et al.) **add**:

A *probabilistic* view of the relationship between W, Z and $\theta$

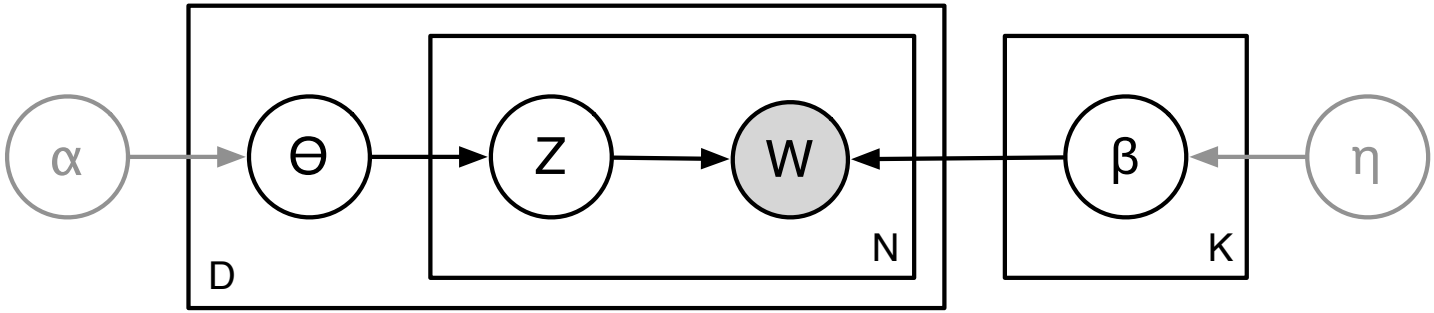
Full statistical framework for learning most aspects of the relationship

# Topic Models as Classical Content Analysis

From




(via pLSI) to



# Generative probability model

Assumptions:

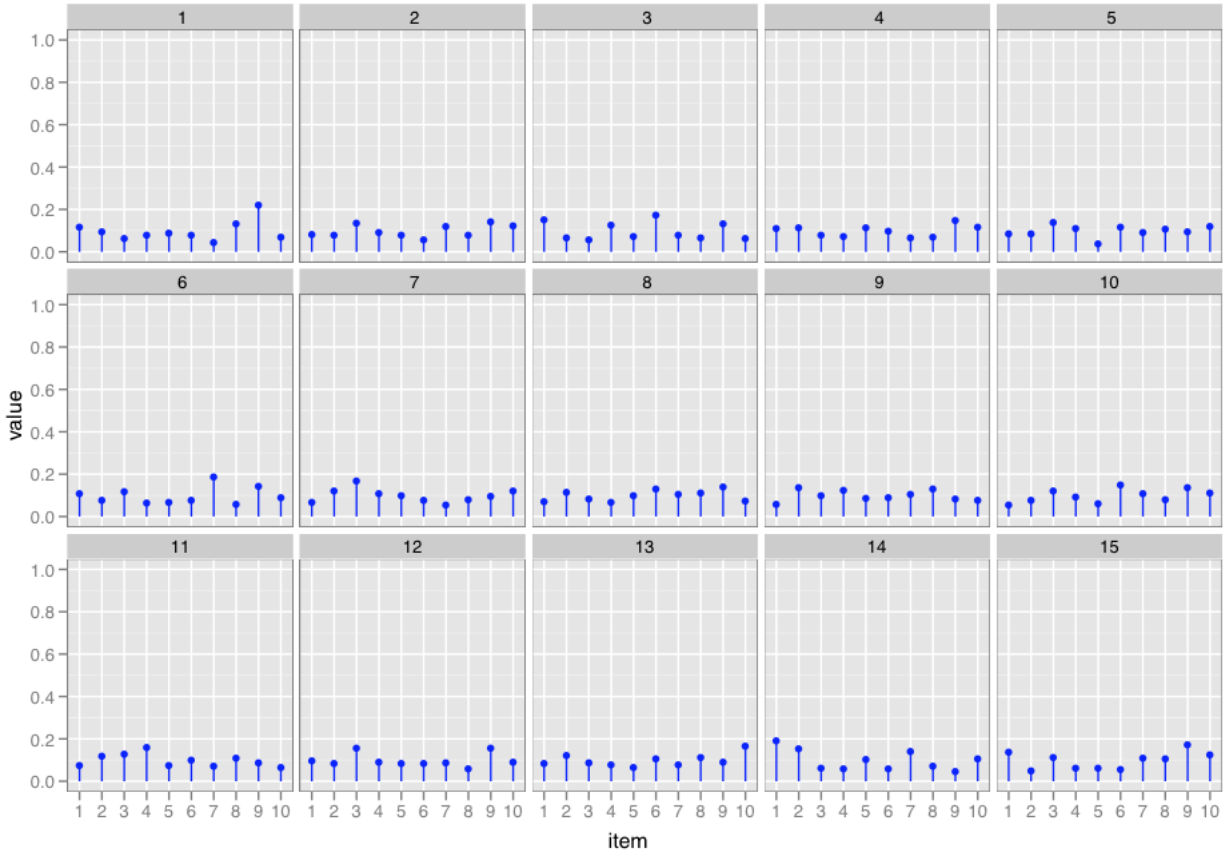
W and Z are multinomial.

$\theta$ and $\beta$ are Dirichlet

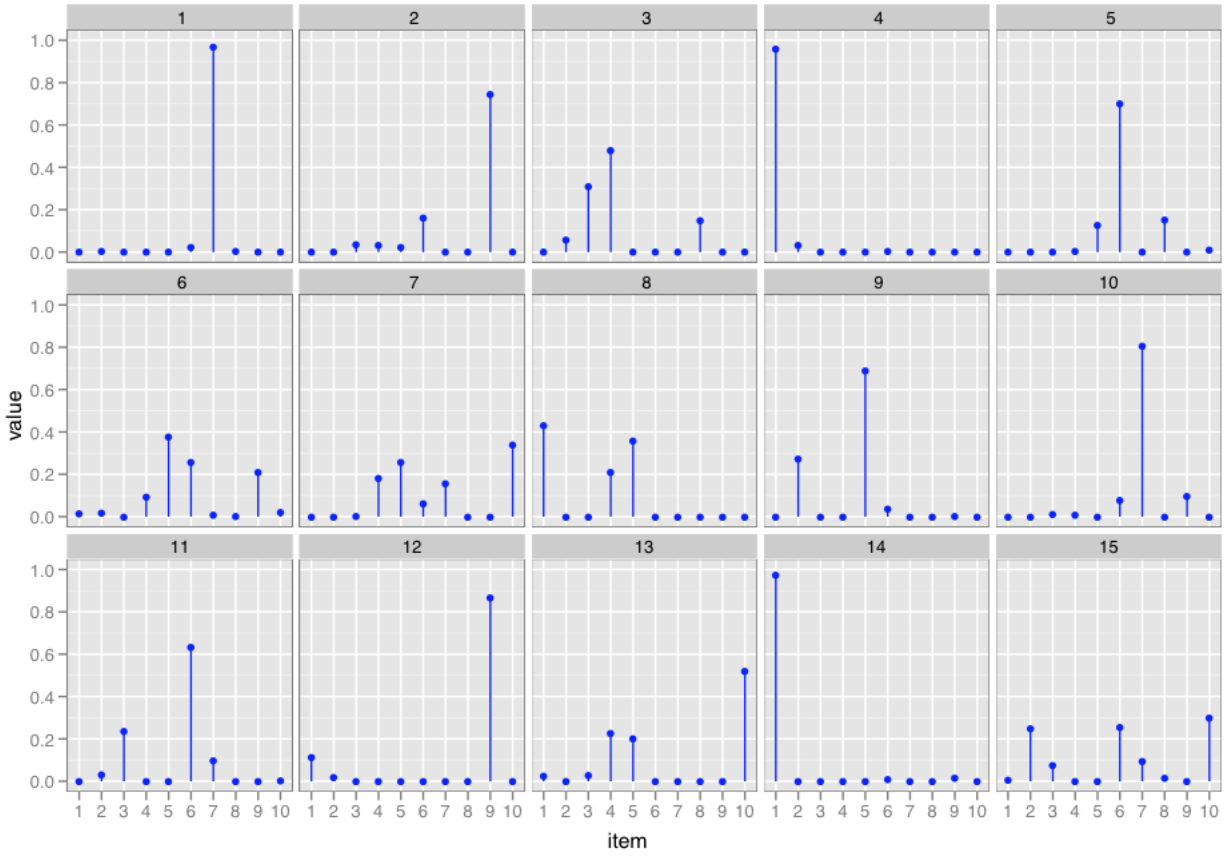
$\alpha$ and $\eta$ are hyperparameters for Dirichlet parameters

- Magnitude controls expected sparsity of words per topic and topics realised per document
- Often optimised directly (a la 'Empirical Bayes') rather than integrated out

# $\alpha = 10$


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
item
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5 6 7 8 9 10

# $\alpha = .1$


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
item
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5 6 7 8 9 10

# Interpretation

Topic assignment measurement process is modeled (yay!)

Topic meaning is no longer under your control (boo!)

Topic modeling is (mostly) *exploratory* rather than *confirmatory*

Upshot:

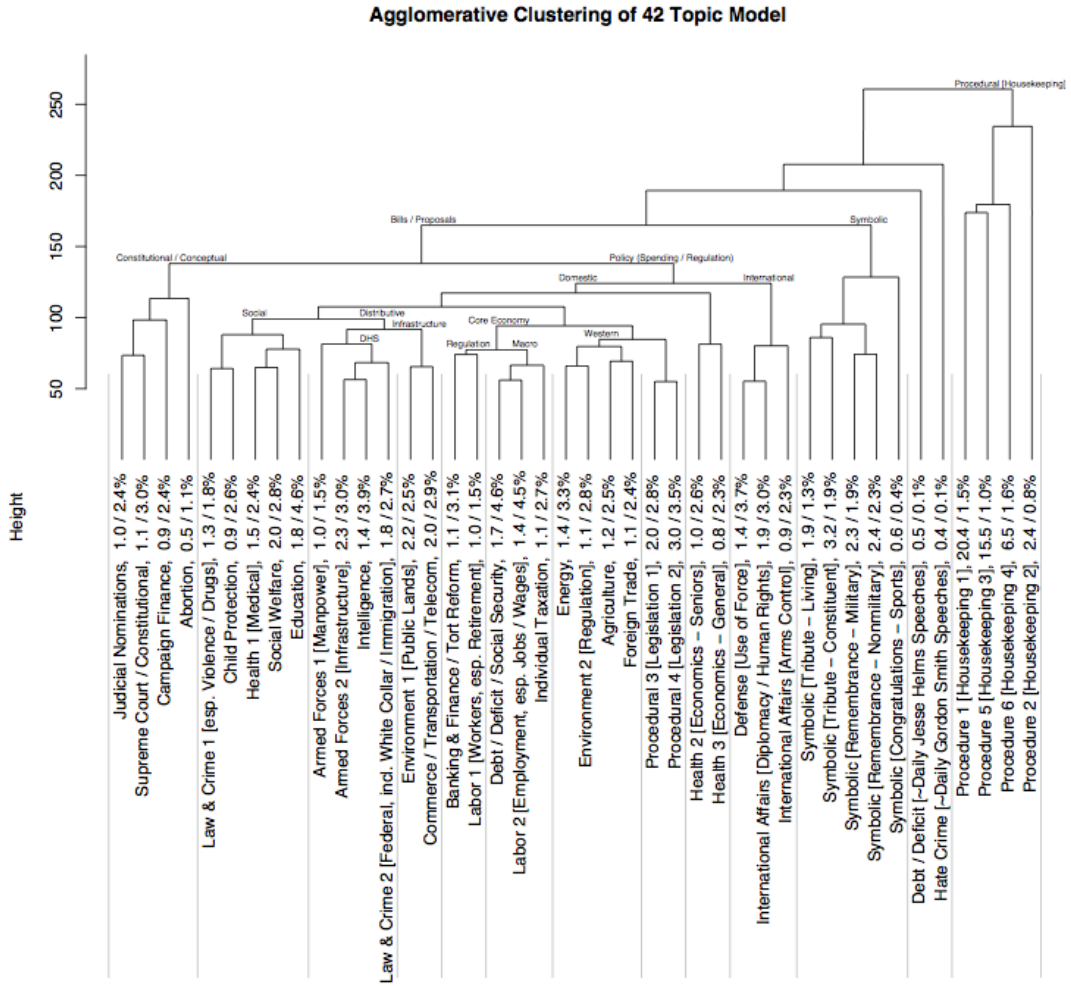
> *You* have to make the case that your topic model is capturing the topic you want to be capturing and not others.
>
> Topic *number* is now an open question

# What is this topic anyway?

| 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* |

(Quinn et al. 2010)

# How is it related to others topics?



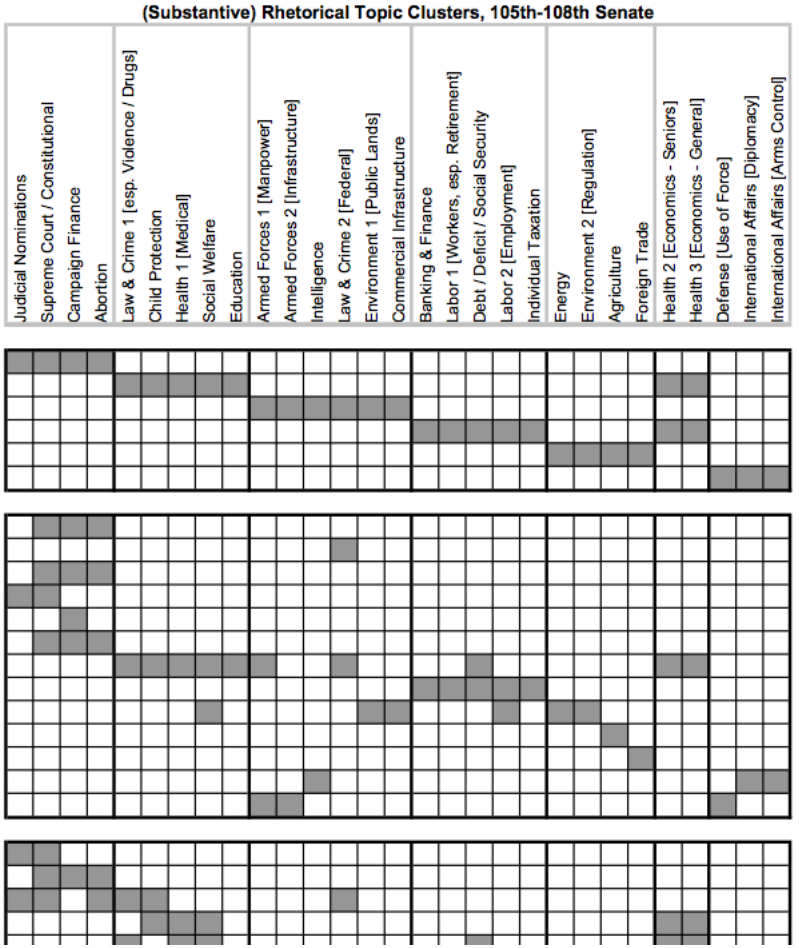
# and other schemes?

**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]

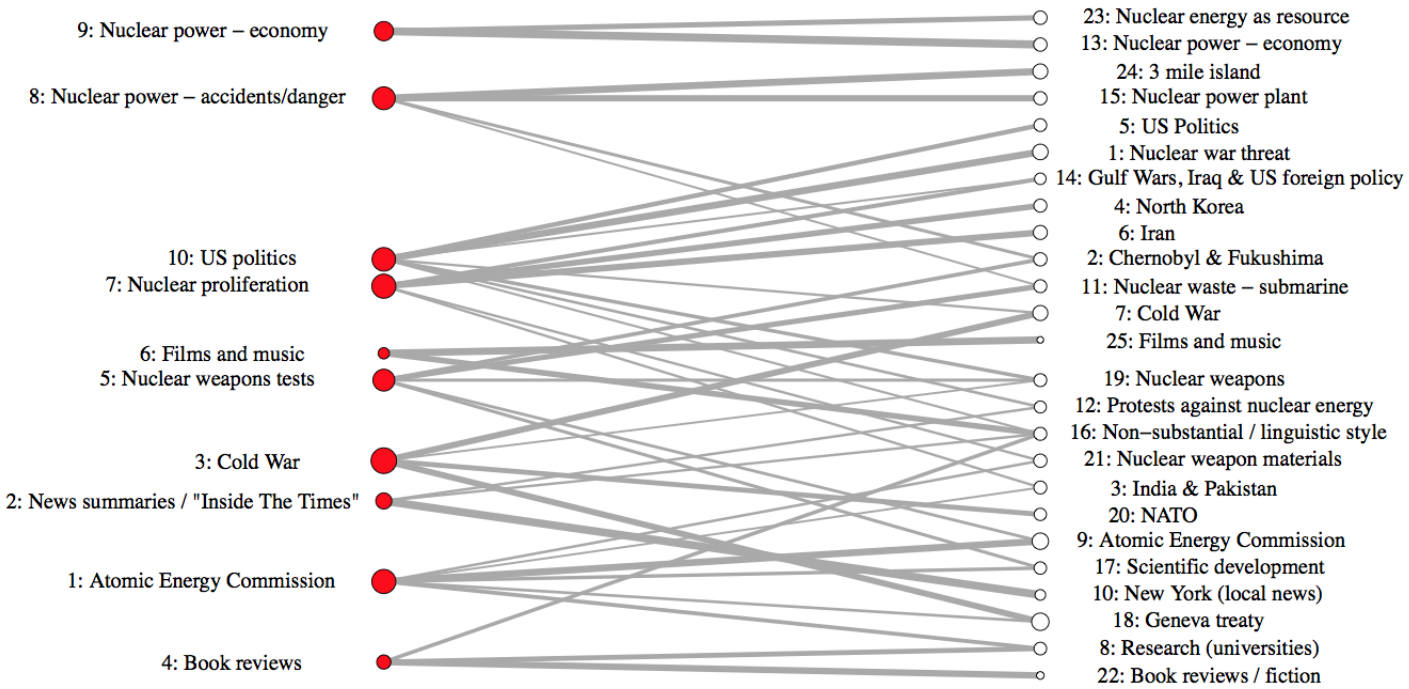
**Jones and Baumgartner (2005)**
- Government operations
- Human services / law [Rights, liberties, minorities]
- Human services / law [Law / crime / family]
- Human services / law [Health]



# Topic Models: How *many* topics?

The results presented in this paper [. . . ] assume there are 43 topics present in the data. I varied the number of assumed topics from only five topics, up to 85 different topics. Assuming too few topics resulted in distinct issues being lumped together, whereas too many topics results in several clusters referring to the same issues. *During my tests, 43 issues represented a decent middle ground.* (Grimmer 2010, p.12)

# Nested topics?




from Jacobi et al (2014) – Hierarchical Dirichlet Process models models *do* enforce nested topics

# A troubling tension

There are two natural metrics for evaluating topic models (or any other kind of computer assisted text analysis)

- Statistical performance, e.g. held-out likelihood
- Substantive usefulness, e.g. topic coherence

Experiments by Chang et al. (2009), using

- word intrusion (which word does not belong?)
- topic intrusion (which topic does not belong?)

suggest that these are robustly *negatively* correlated.

# Variations

There have been a huge number of variations on and extensions of the basic topic model

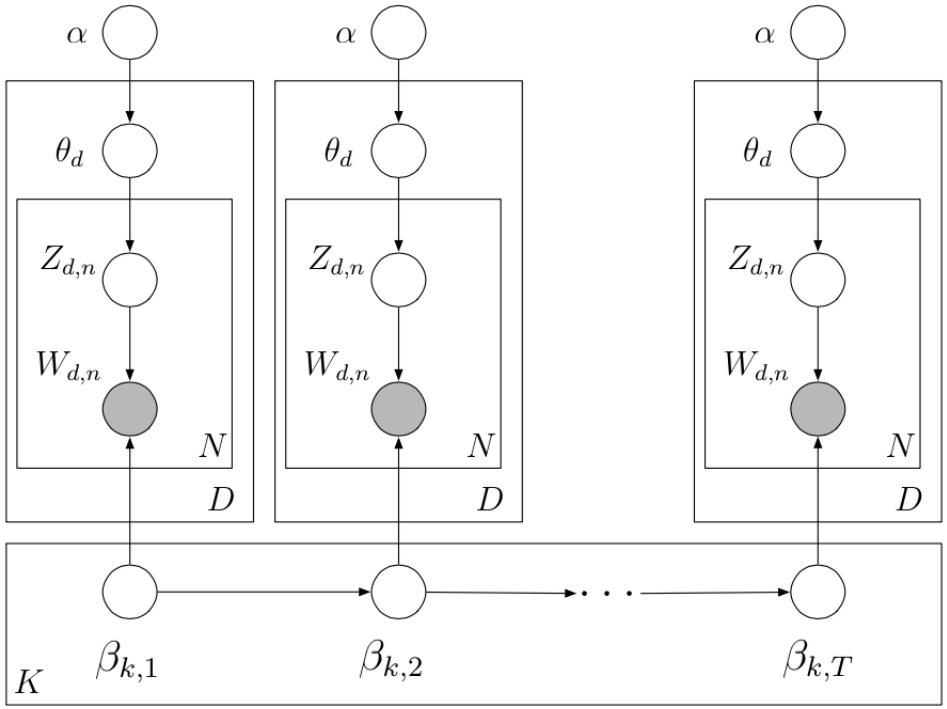
- Airoldi et al. 2011 Table 1 provide references for 33!

I'll briefly mention two variations of interest to political science:

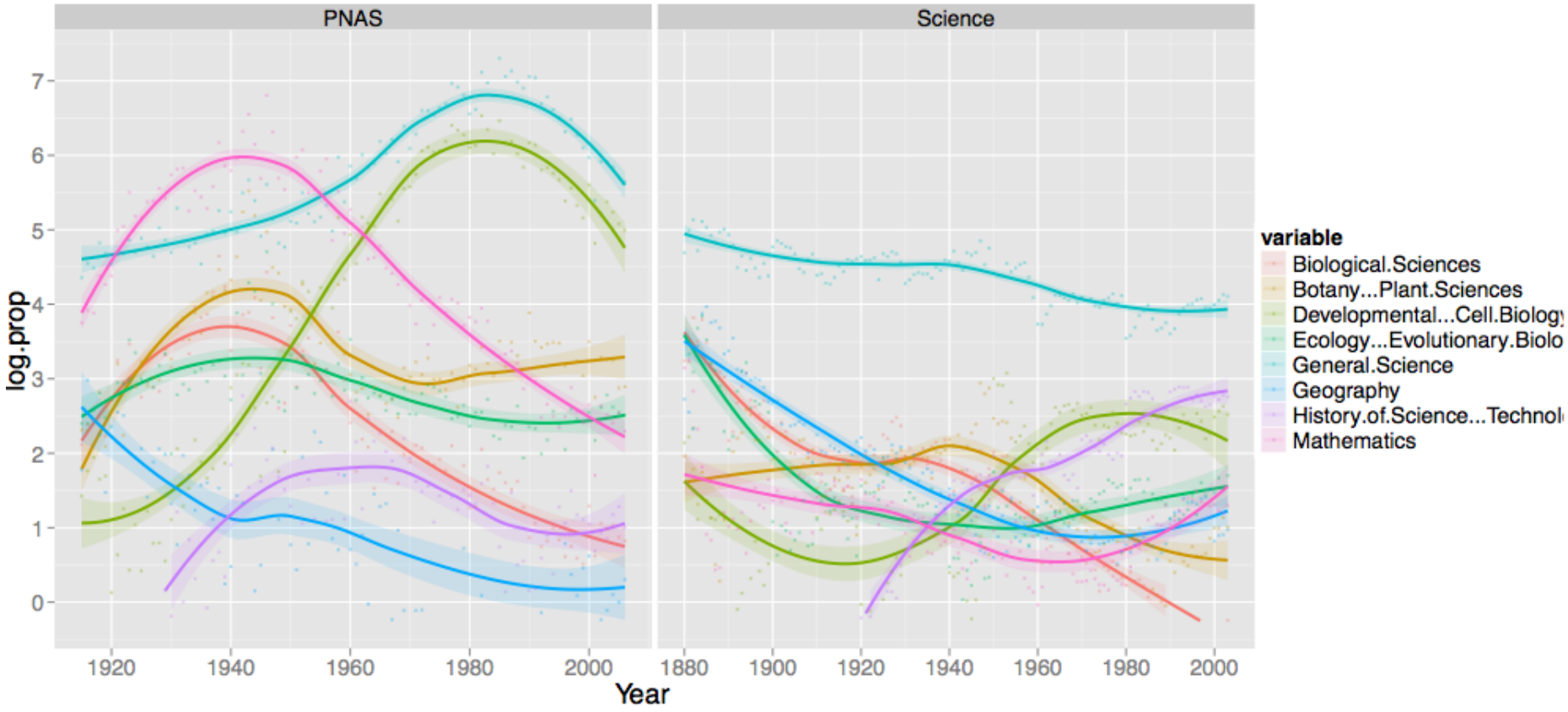
- Dynamic topic models
- Structural topic models (`http://structuraltopicmodel.com`)

# Dynamic Topic Models (1)

Words 'drift' from topic to topic over time



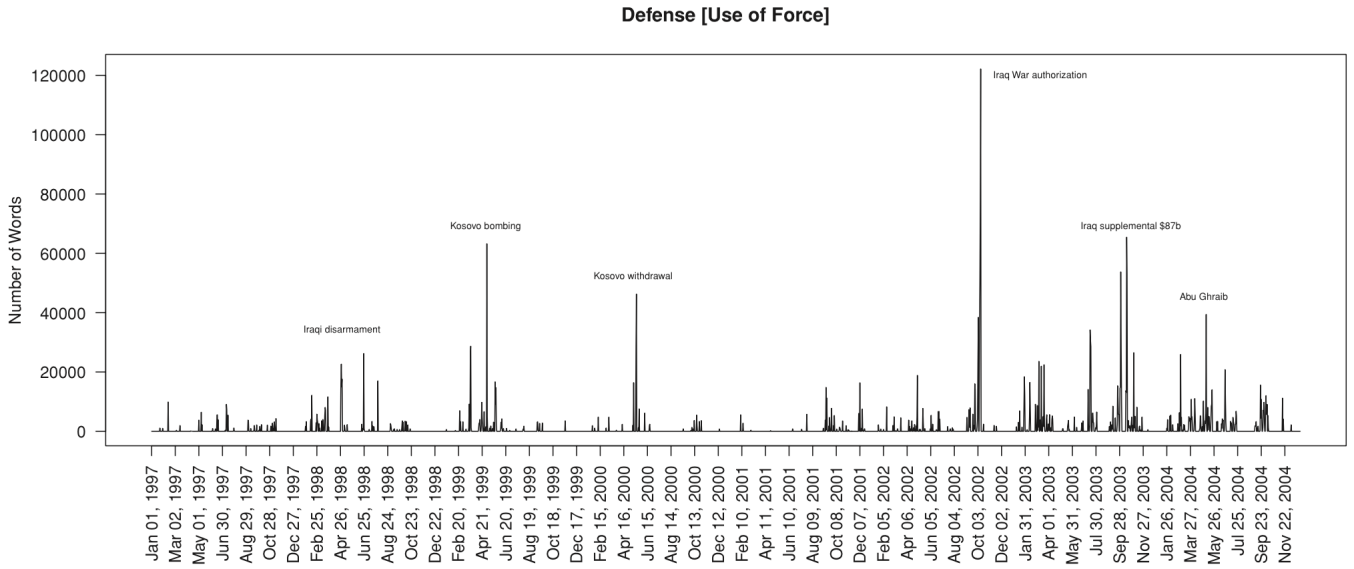
# Dynamic Topic Models (1)



# Dynamic Topic Models (2)

Alternatively (Quinn et al 2010) document topic proportions 'drift' over time

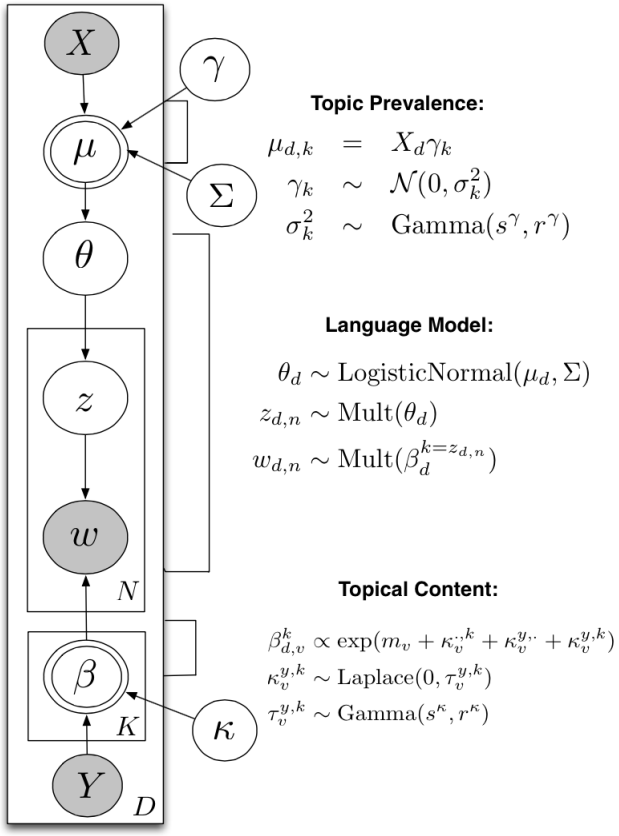
To track a smoothly moving policy agenda (congressional speeches from 1997-2004)



Note:

It's not clear you can have both types of dynamics!

# Structural Topic Models




Covariates for topics proportions and on topic words

# Structural Topic Models

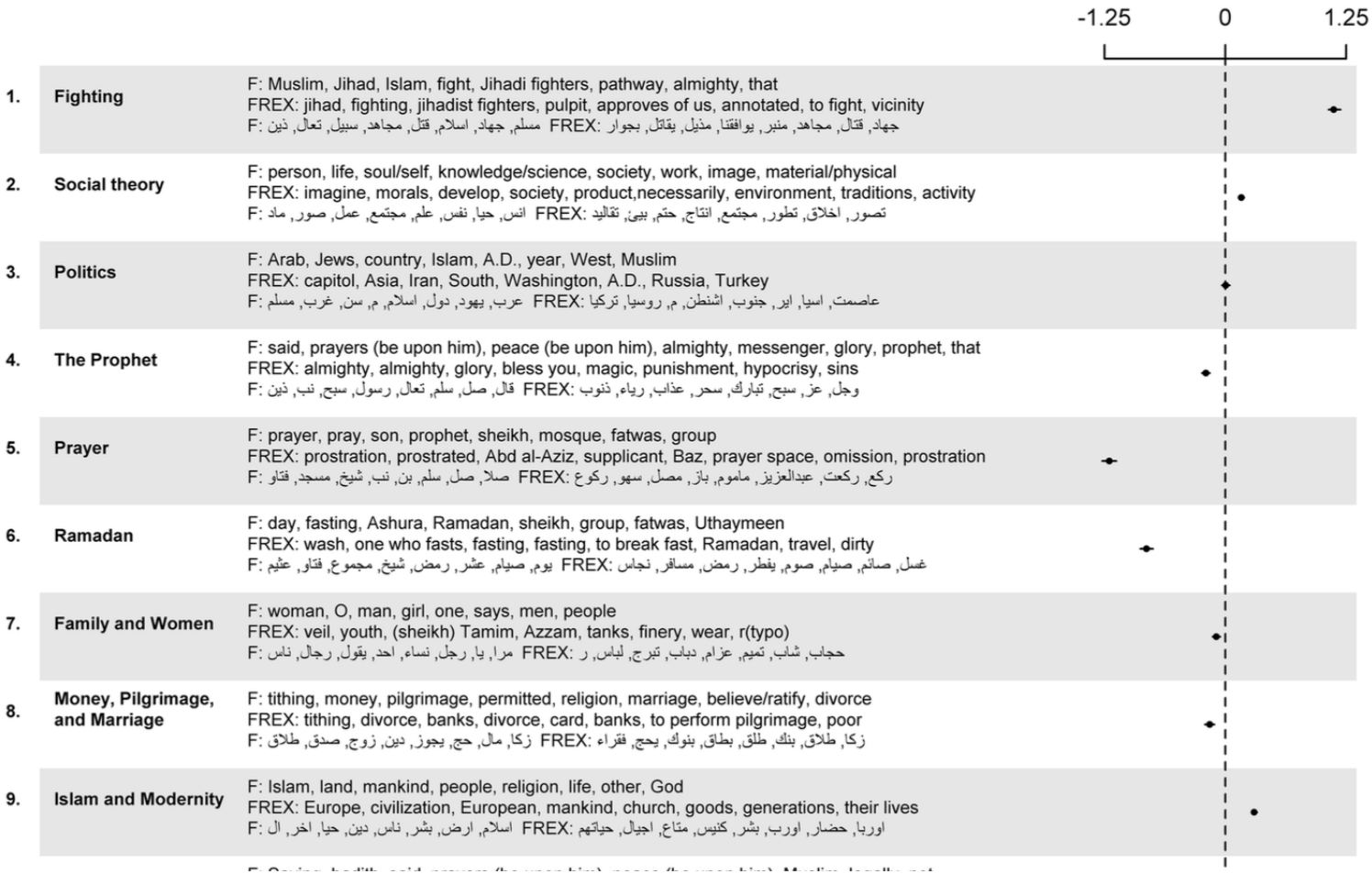
Think SEM/Multilevel measurement models for text...

Some history of modeling topic proportions:

- LDA: Topic proportions are *nearly* a priori independent
- CTM 'Correlated Topic Model': Topic proportions can be correlated via a Normal distribution in the space of topic proportion logits
- STM 'Structural Topic Model': Topic proportion correlations are regressions on covariates

Covariates are intended to facilitate interpretation/explanation (also provide measurement bias)

# Structural Topic Models



# Wrapping up