Computer-assisted content analysis

Will Lowe Princeton University
James Lo University of Southern California

May 20, 2016

IQMR 2016 Syracuse

Plan

http://conjugateprior.org/teaching/iqmr/materials.zip

or

http://aws-23fggg.com.org:8787

Practicalities: Labs

Two streams of labs. But you can switch…
(and we'll stop if you ask us)

Focus

Assumptions

Mechanics

Interpretation

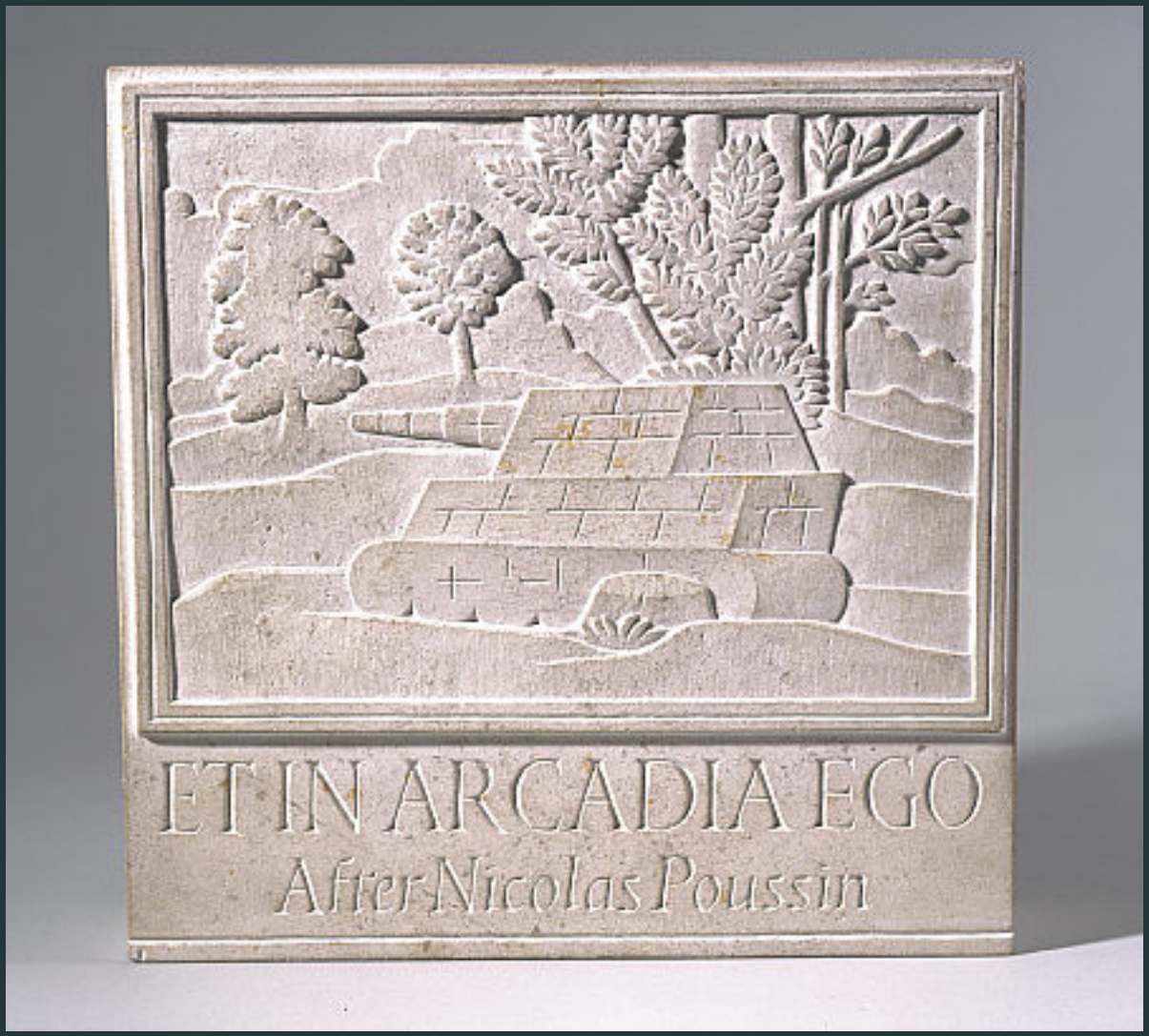
Pitfalls

Today

How to learn about

- party platforms
- legislative agendas
- parliamentary debates
- bloggers
- presidents
- international terrorists

by counting (lots of) words…


ET IN ARCADIA EGO
After Nicolas Poussin

First Section

The Transcendental Question

What are the *conditions for the possibility* of learning about these things by counting words?

Big Picture

There is a *message* or *content* that cannot be *directly* observed, e.g.

> The topic of my lecture, my position on a political issue, the importance of defence issues to a some political party.

and *behaviour*, including *linguistic behaviour*, e.g.

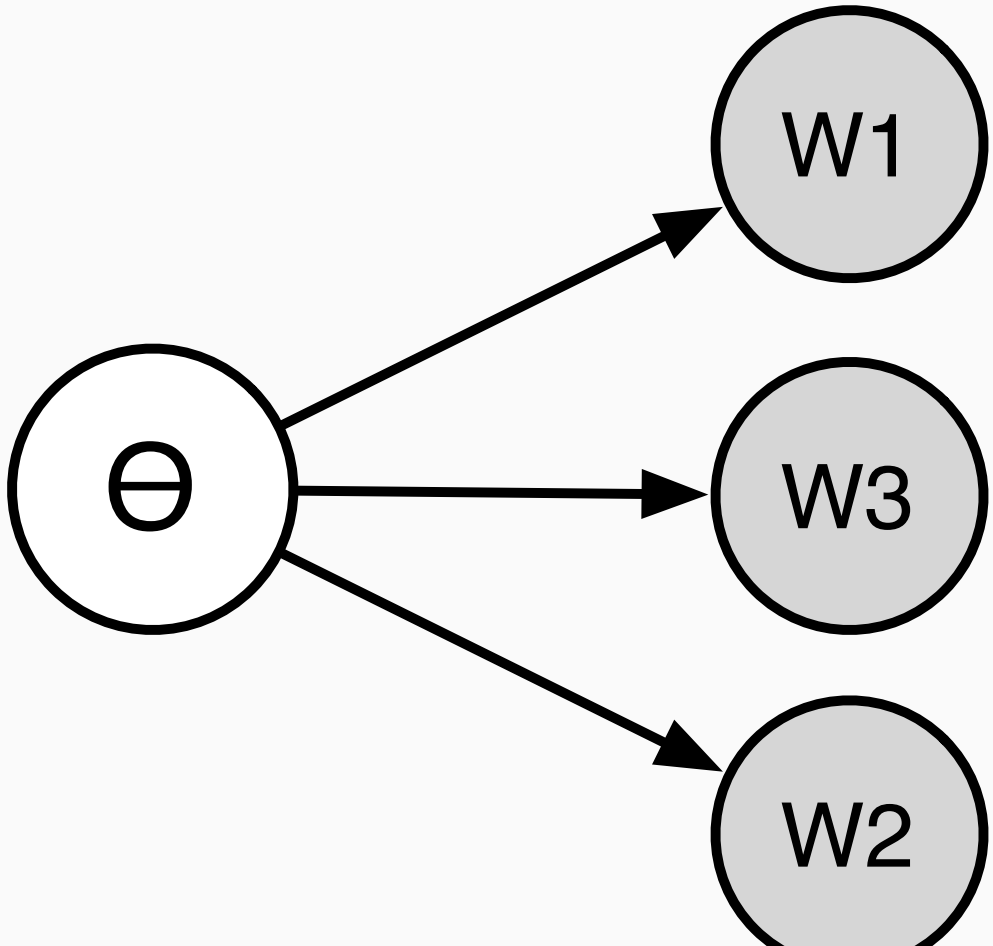
> yelling, muttering, cursing, lecturing

which *can* be directly observed.

Focus on the *expressed message* and the *words*…

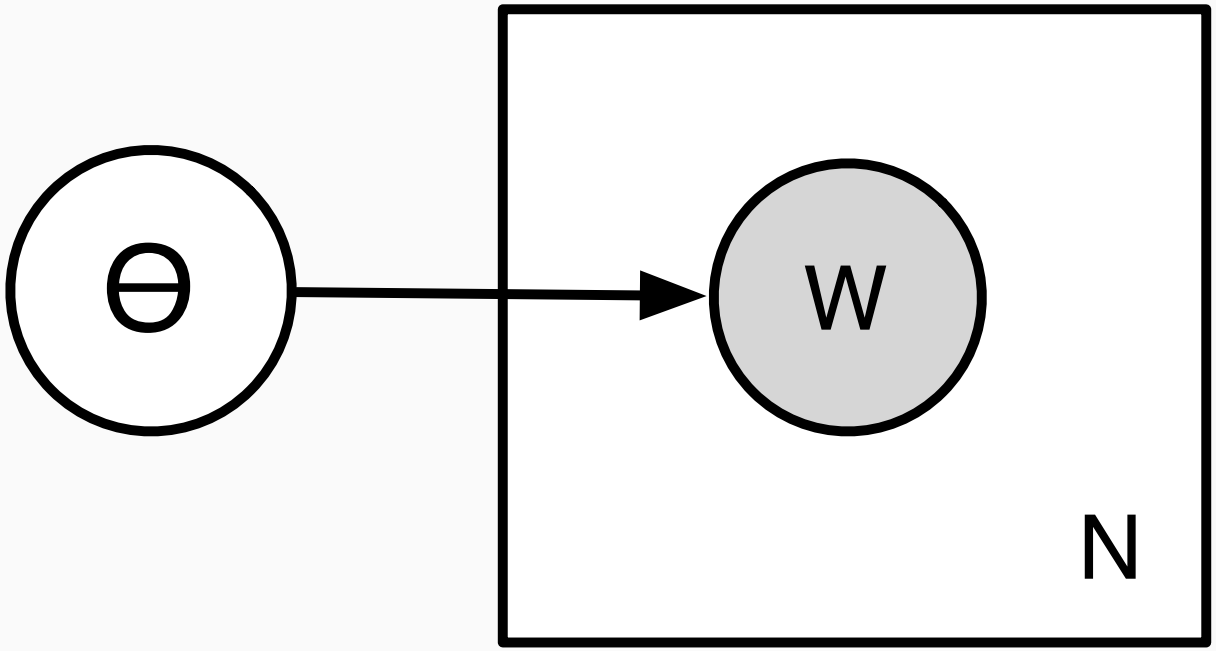
Communication

To *communicate* a message θ – to inform, persuade, demand, threaten, a producer (the speaker or writer) *generates* words of different kinds in different quantities



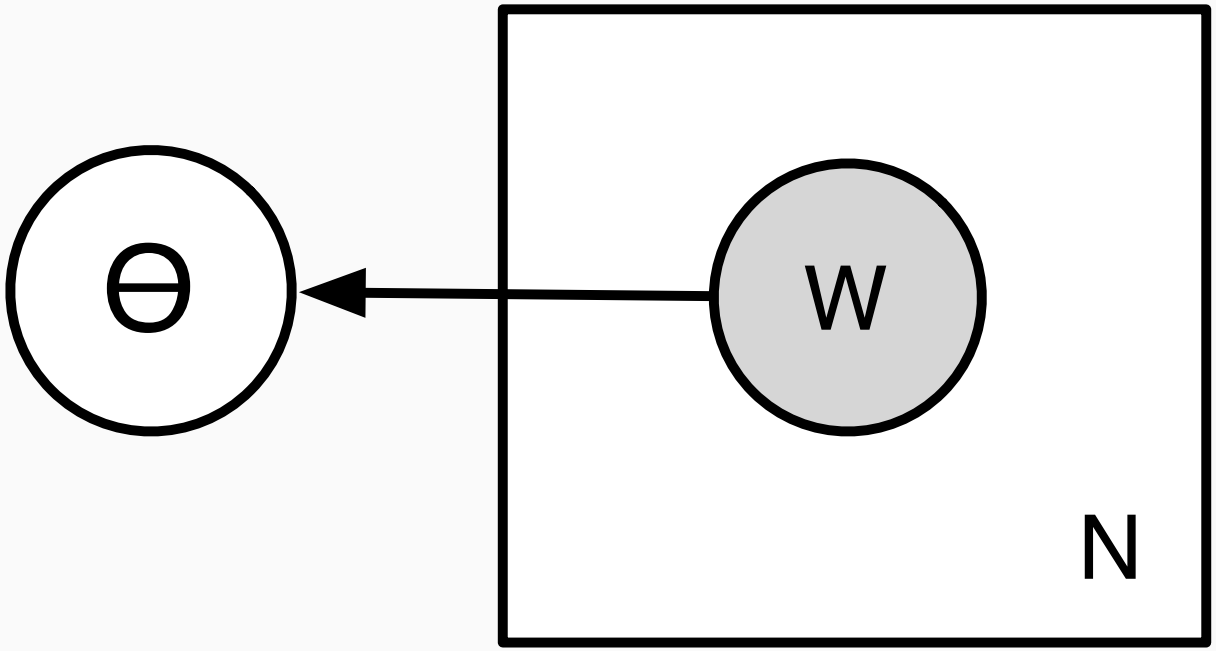
Communication

To *communicate* a message θ – to inform, persuade, demand, threaten, a producer (the speaker or writer) *generates* words of different kinds in different quantities



Communication

To *understand* a message the consumer (the hearer, reader, coder) uses those words to *reconstruct* the message



Communication

This is a stable (Searle, 1995) institutionally conventionalised (Lewis, 1969) but disruptable (Riker, 1996) communication process in which no finite set of words *uniquely* identifies any content (Quine, 1960; Davidson, 1977)

How to model this without having to solve the problems of linguistics (psychology, politics) first?

Rely on: instrumentality, conventionalisation, randomness and reflexivity

Instrumentality

Language use is as a *form of action* (Wittgenstein, 1953; Austin, 1975; Dawkins and Krebs, 1978)

Note the distinction between

> '*W* means *X*'
>
> versus
>
> '*W* is used to mean *X*'

Content analyses to work better when language usage is *stable* and *instrumental*…

Randomness

You know the content of your beliefs (probably) but others only infer them, on the basis of data

> c.f. 'How do I know what I think until I hear what I say?' (E. M. Forster)

The *primary data* are often the words you use

You almost never *say exactly the same words twice*, even when you haven't changed your mind. Hence words are the result of some kind of *sampling process*.

We treat this process as *random* because we don't know or care about all the causes of variation

> (and because we're all secretly Bayesians)

Reflexivity

Politicians are often nice enough to talk as if they really communicate this way

> My theme here has, as it were, four heads. […] The first is articulated by the word “opportunity” […] the second is expressed by the word “choice” […] the third theme is summed up by the word “strength” [and] my fourth theme is expressed well by the word “renewal”
>
> (M. Thatcher, 1979)

[2, 7, 2, 8] in 4431 words

More Instrumentality

The secret of quantitative political text analysis:

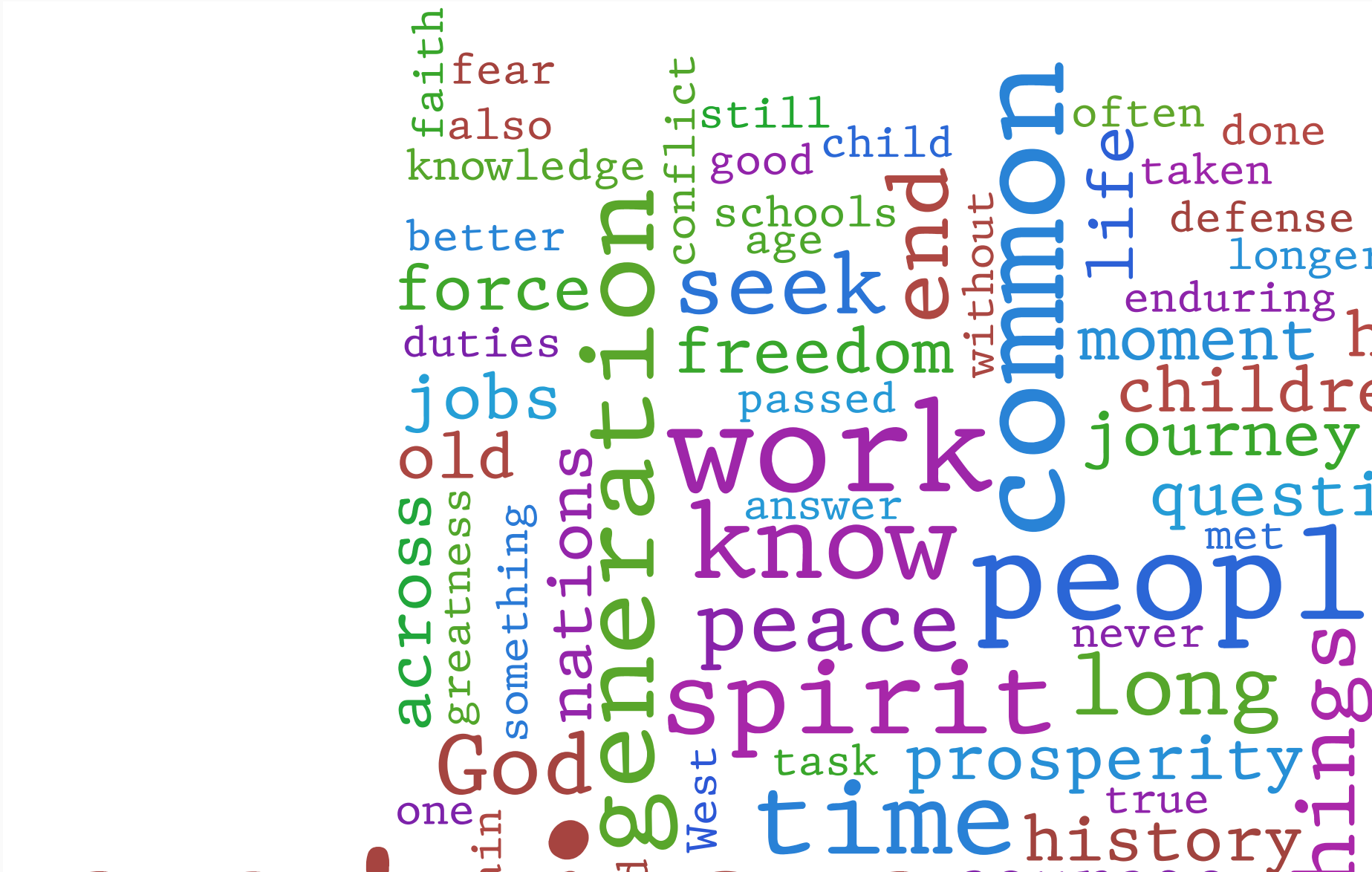
we aren't actually interested in words W

that's for linguists…

we aren't actually interested in what's in your head θ

that's for psychologists…

except as they help explain things we are interested in. They are *just data*.


faith
fear
also
knowledge
conflict
still
good
child
often
done
life
taken
schools
age
defense
better
seek
end
without
common
force
freedom
enduring
moment
duties
jobs
passed
old
work
journey
answer
across
greatness
nations
generation
know
met
something
peace
people
never
spirit
long
God
task
prosperity
West
time
true
history
one

Words as Data

What do we know about words *as data*?

They are *difficult*

- High dimensional
- Sparsely distributed (with skew)
- Not equally informative

Difficult Words

Example: Labour party (2010) manifesto compared to other parties in two elections

High D.	6343 word types in three manifestos
Sparse	Of these, Labour only uses 3675 (58%)
Skewed	Of these 1783 (49%) words appear exactly once, and 2854 (78%) appear <5 times

Average (non-academic) adult vocabulary contains about 10,000 words

Labour manifesto uses about 37% of commonly available types

Difficult Words

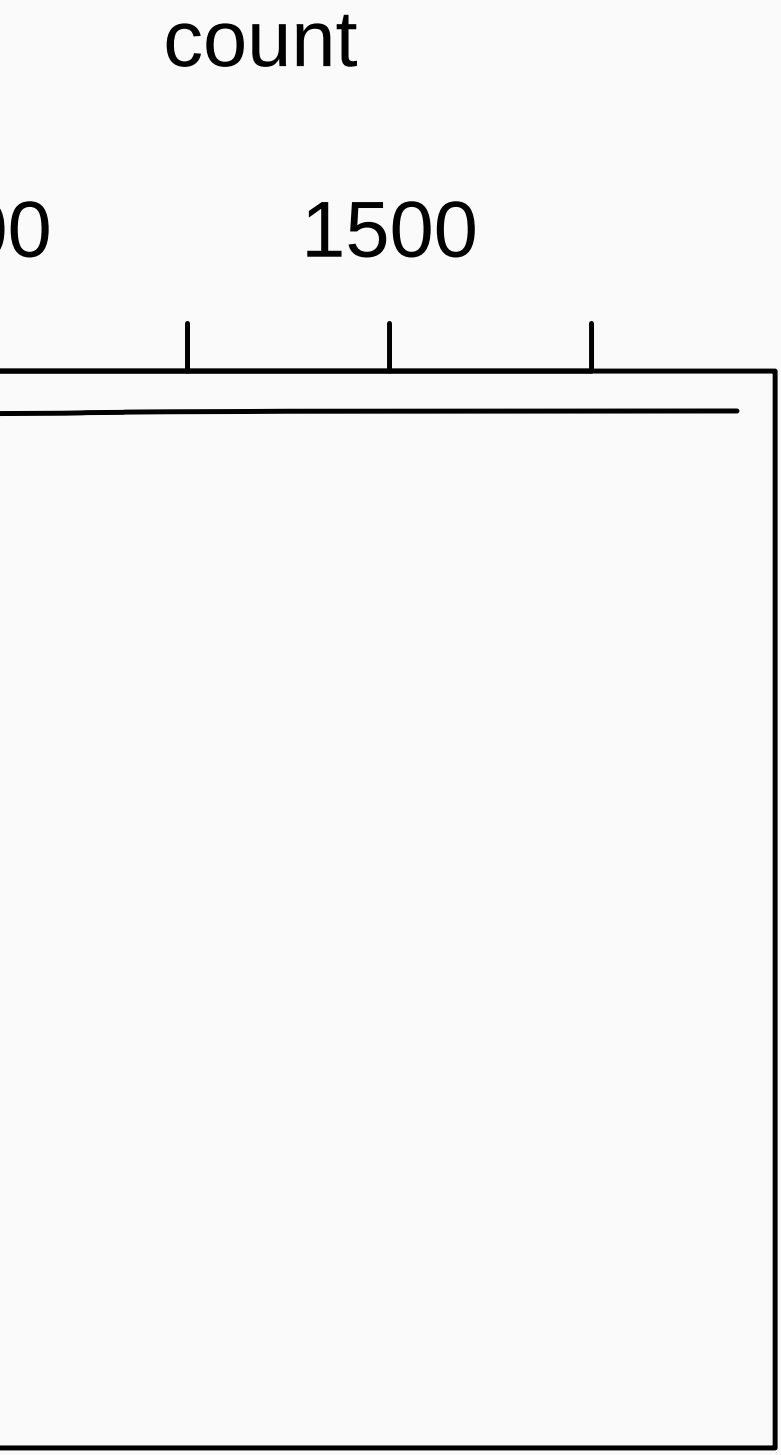
Words are not like your other data...

Zipf-Mandelbrot law (a pareto distribution in disguise)

$$P(w_i) \propto 1/r_i^a$$

where r_i is the frequency *rank* of word i and $a \approx 1$

Very fat tailed…



log count

Dealing with Difficult Words

For large amounts of text summaries are not enough. We need a *model*.

But we need to make some *exchangeability* assumptions or *conditional independence* assumptions before we can even do that…

These are typically called the ‘bag of words’. Specifically:

Text As You Might Read or Hear It

"As I look ahead I am filled with foreboding. Like the Roman I seem to see 'the river Tiber flowing with much blood'…"
(E. Powell, 1968)

Punctuation Invariance

"As I look ahead I am filled with foreboding. Like the Roman I seem to see 'the river Tiber flowing with much blood'…"
(E. Powell, 1968)

index	token
1	as
2	i
3	look
4	ahead
5	i
6	am
7	…

index	token
1	like
2	the
3	roman
4	i
5	seem
6	to
7	…

Lexical Univocality

type	count
as	1
i	2
look	1
ahead	1
am	1
…	…

token	count
like	1
the	1
roman	1
i	1
seem	1
to	1
…	…

Order invariance

		unit	
		sent.1	sent.2
type	ahead	1	0
	am	1	0
	as	1	0
	i	2	1
	like	0	1
	look	1	0
	roman	0	1
	seem	0	1
	the	0	1
	to	0	1
	…	…	…

Count Data

Yes, we have turned a corpus into a contingency table.

Everything you learned in your categorical data analysis course applies

except that some variables of interest: θ are *not observed*

Down To Business

	ahead	am	am	i	like	look		content
sent.1	1		1	1	2	0	1	…	θ_1
sent.2	0		0	0	1	1	0	…	θ_2

For each research problem involving content analysis we need to ask:

What *structure* θ has

What *modeling strategy* to take

What the *relationship* is between θ and the words (i.e. the model)

The Structure of θ

Categorical structure: topics of newspaper articles or speeches
Data type: nominal or ordinal (category labels)

Dimensional structure: In spatial politics well-ordered preferences imply distances in an ideological space
Data type: Real numbers

Network structure: citation analyses, social network analysis
Data type: nodes and arcs

Predicative structures: assertions about the beliefs of other political actors
Data type: ??

Statistical Models of Words: Poisson

Word counts/rates are conditionally Poisson:

$$\begin{aligned} P(W_j) &= \text{Poisson}(\lambda_j) \\ &= \frac{{\lambda_j}^{W_j}\, e^{-\lambda}}{W_j!} \end{aligned}$$

Expected W_j (and its variance) is λ_j

Models are mostly proportional because multiplicative

Conditional on what? Typically on θ

Statistical Models of Words: Poisson

Statistical Models of Words: Poisson

We are going to think of content as systematically altering word and topic rates

Statistical Models of Words: Multinomial

For fixed document length counts are conditionally Multinomial:

$$P(W_1 \ldots W_V) = \text{Multinomial}(W_1 \ldots W_V; \pi_1 \ldots \pi_V, N_i)$$

$$\frac{N!}{W_1! \ldots W_V!} \prod_j^V \pi_j^{W_j}$$

Expected W_i is $N\pi_i$

Covariance of W_i and W_j is $-N\pi_i\pi_j$ (budget constraint)

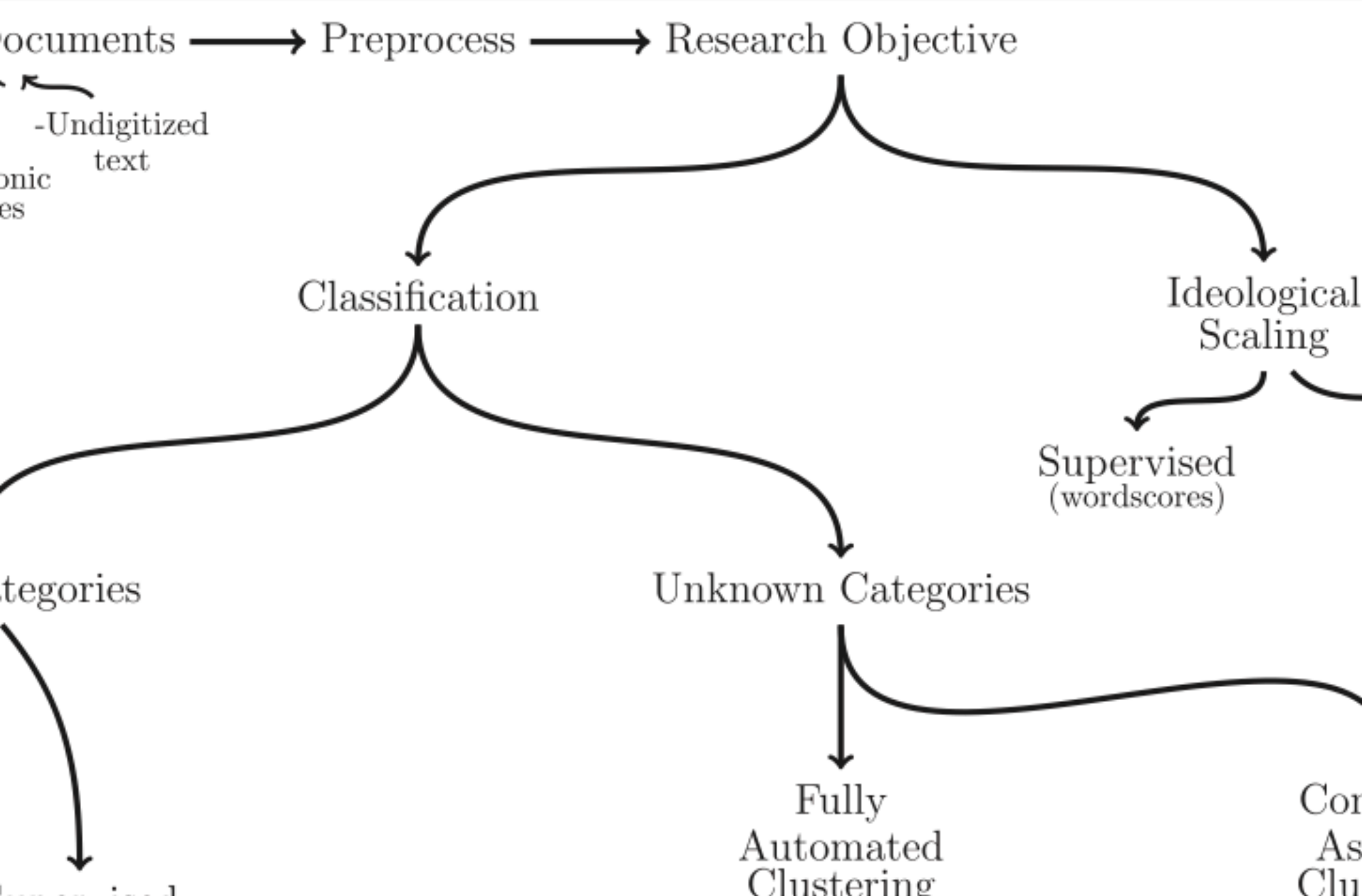
Statistical Models of Words: Multinomial

Statistical Models of Words: Multinomial

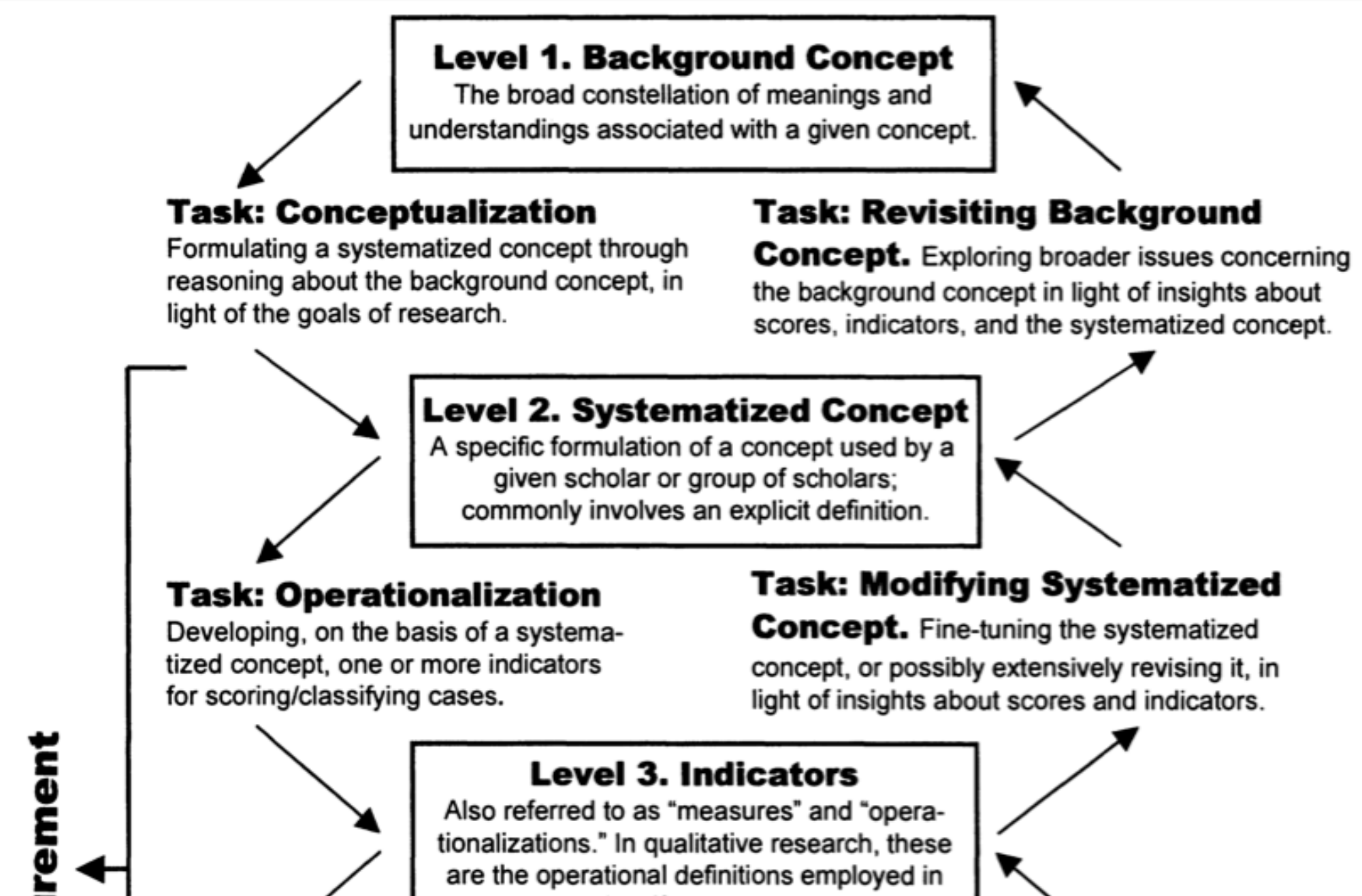
We are going to think about *content topics* as like this

We build measurement devices

(Almost) the big picture



Really the big picture



Modeling strategies

We can model the contents of a word frequency matrix in several ways

$\theta \longleftarrow$ words: Go for $P(\theta \mid \text{words})$ *directly*

Requires some *observed* θ, and lots of *careful* regression modeling, or manual coding

$\theta \longrightarrow$ words: Get $P(\theta \mid \text{words})$ *indirectly*

Model words as a function of θ, add a prior, and infer θ using Bayes theorem

$$P(\theta \mid \text{words}) = \frac{P(\text{words} \mid \theta)P(\theta)}{\sum_k^\theta P(\text{words} \mid \theta_k)P(\theta_k)}$$

Modeling Strategies

These strategies correspond to classification, classical content analysis and topic models, and scaling models

That is to say:

> Next session, this session, and the last session

Commitment issues

What are we committing to in this quantitative content analysis framework?

Probably less than you think…

Assumptions:

- θ is socially / institutionally constructed: only linguists care about the 'real' thing
  - Note: this does not rule out being objectively correct – there is a fact about the worth of money in my pocket
- There are no differences in θ that make no verbal difference (basically Pragmatism)

Absence is an observation

Don't be fooled…

Statistical models of text deal with absence as well as presence: zeros count

Absence is informative *to the extent it is surprising*

Surprise implies expectations; expectations imply a model (Kant *and* contemporary neuroscience)

Theory measurement separation

Discourse analytic approaches tend to *tightly couple* theory and 'measurement' components

(This is contingent…)

We will try as far as possible to separate them…

Our concerns: validity, stability

Can rely on: transparency, reliability, replicability

Menu

Session 1:

- Dictionary-Based 'Classical' Content Analysis
  - Content analysis as a model
  - Applications
  - Measurement error

Session 2: Topic Models and Classification

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

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

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.

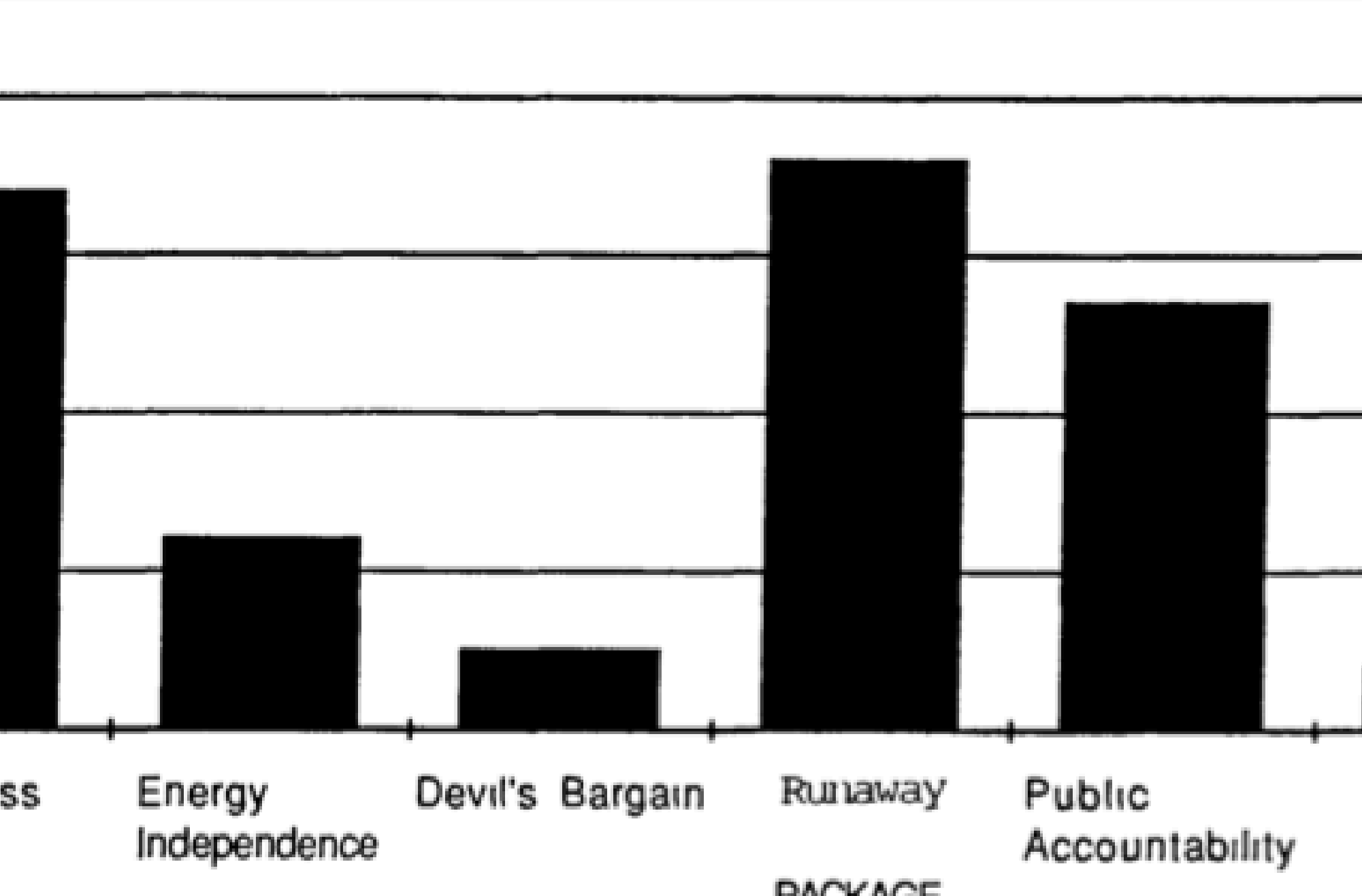
> pro-welfare state vs. anti-welfare state, lots of CMP categories

But they are invariably treated as *nominal level* variables

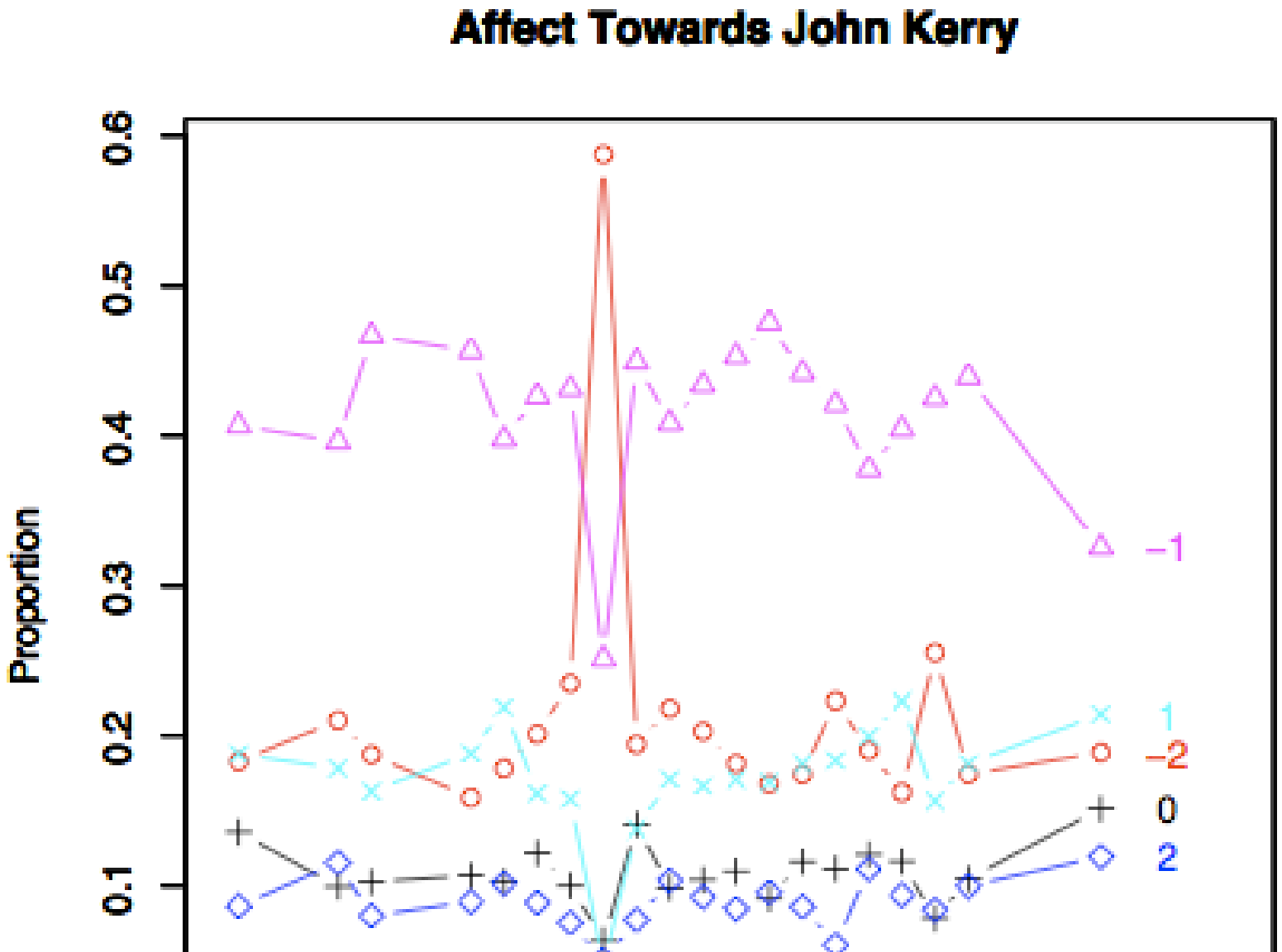
We are typically interested in them for

> simple descriptions, making comparisons, tracing temporal dynamics

Talking Like a Newspaper



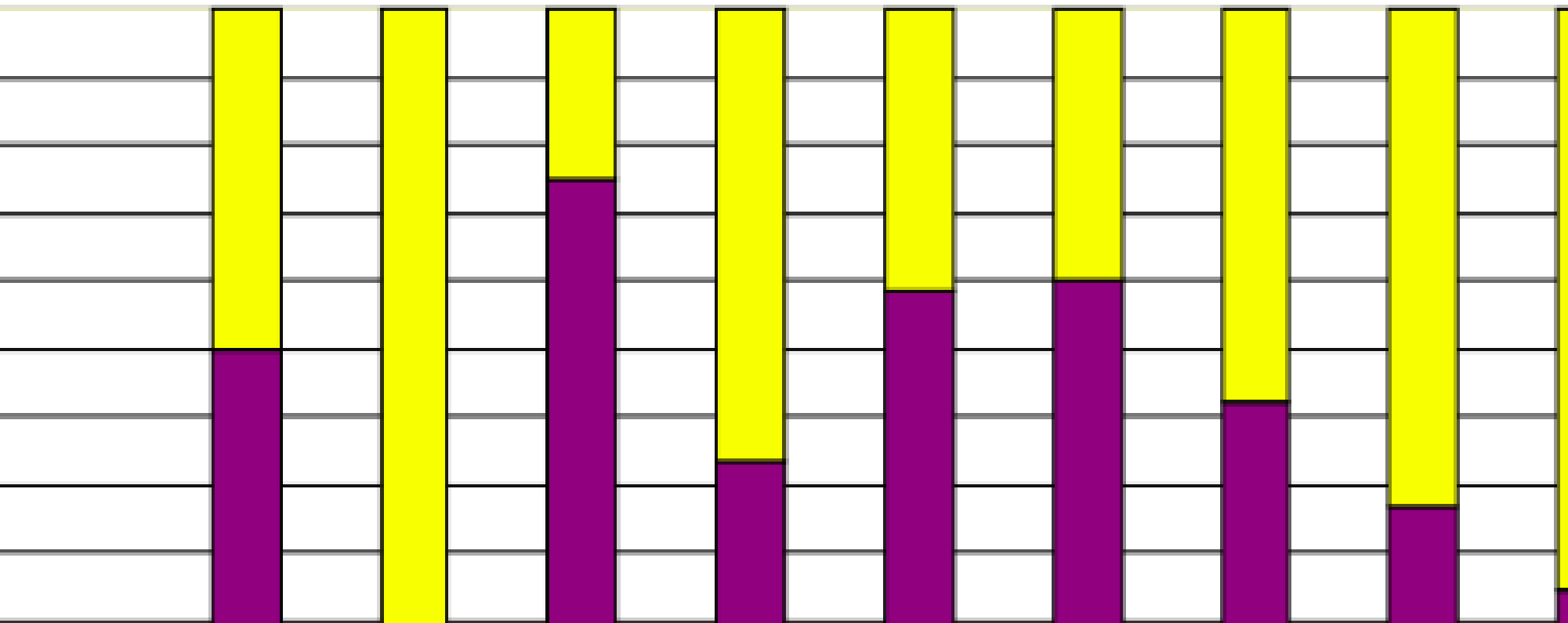
Talking like a Presidential Candidate



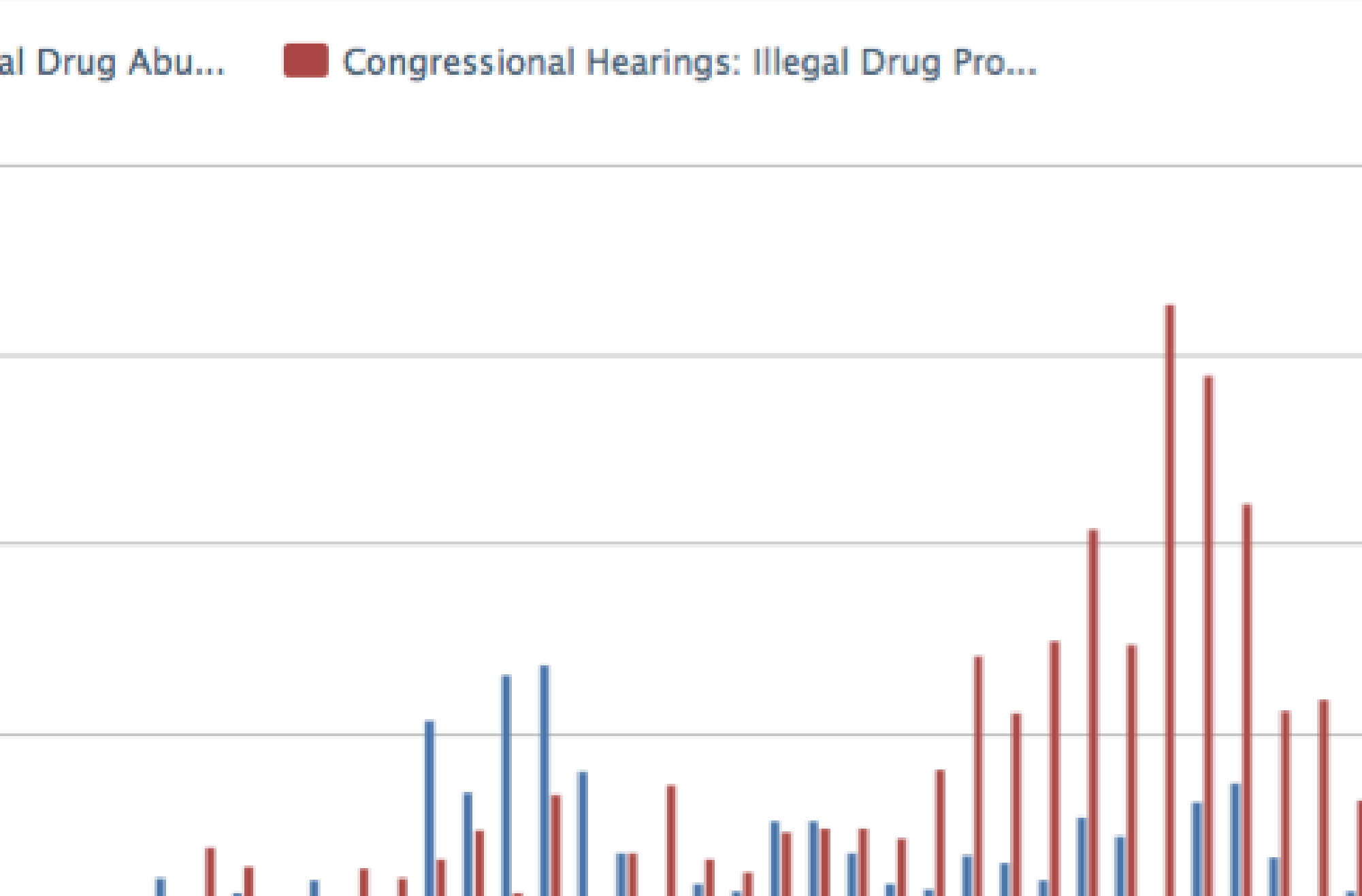
Talking like a Terrorist

	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

lative proportions of policy frames F1 and F2 in secondary



Talking About Drugs

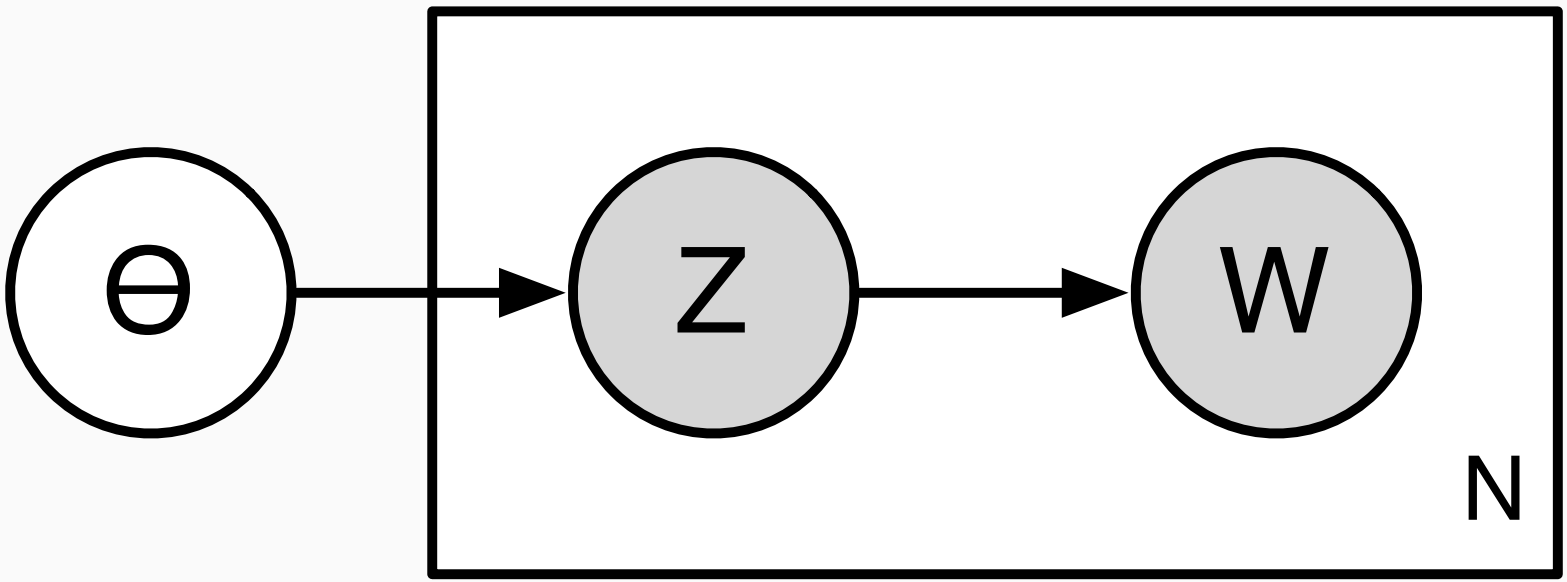


Classical Content Analysis

Categories are

equivalence classes over words

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



Classical Content Analysis

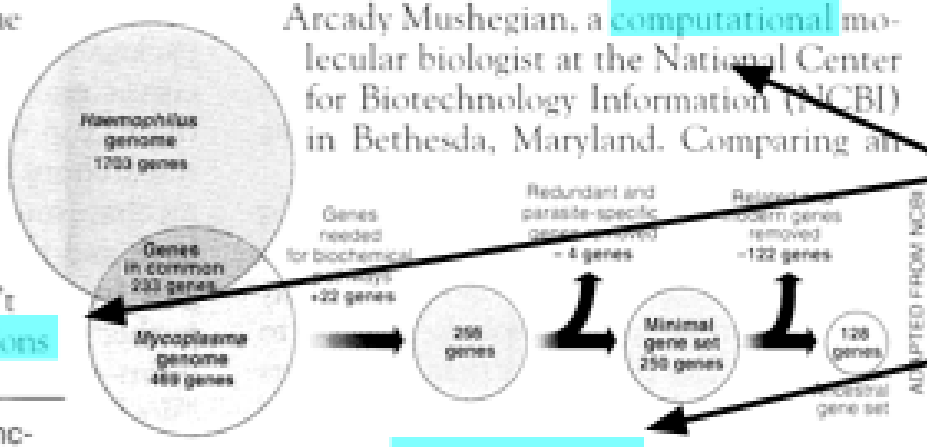
Documents

Topic propor assignn

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



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.

Content Analysis Dictionary

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

from Laver and Garry's (2000) dictionary

P(Z | W)

Translation: For each word:

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

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: $P(k) = 1/K$

Using a dictionary

Define the category *counts*

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

and estimate category relative *proportions* using

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

When θ is a set of multinomial parameters, *and the model assumptions are correct*, this is a reasonable estimator

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

$$\hat{\theta}_{\textit{state reg}} = 0.029 \ [0.027, 0.030]$$

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

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

Rates are more intuitive

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]

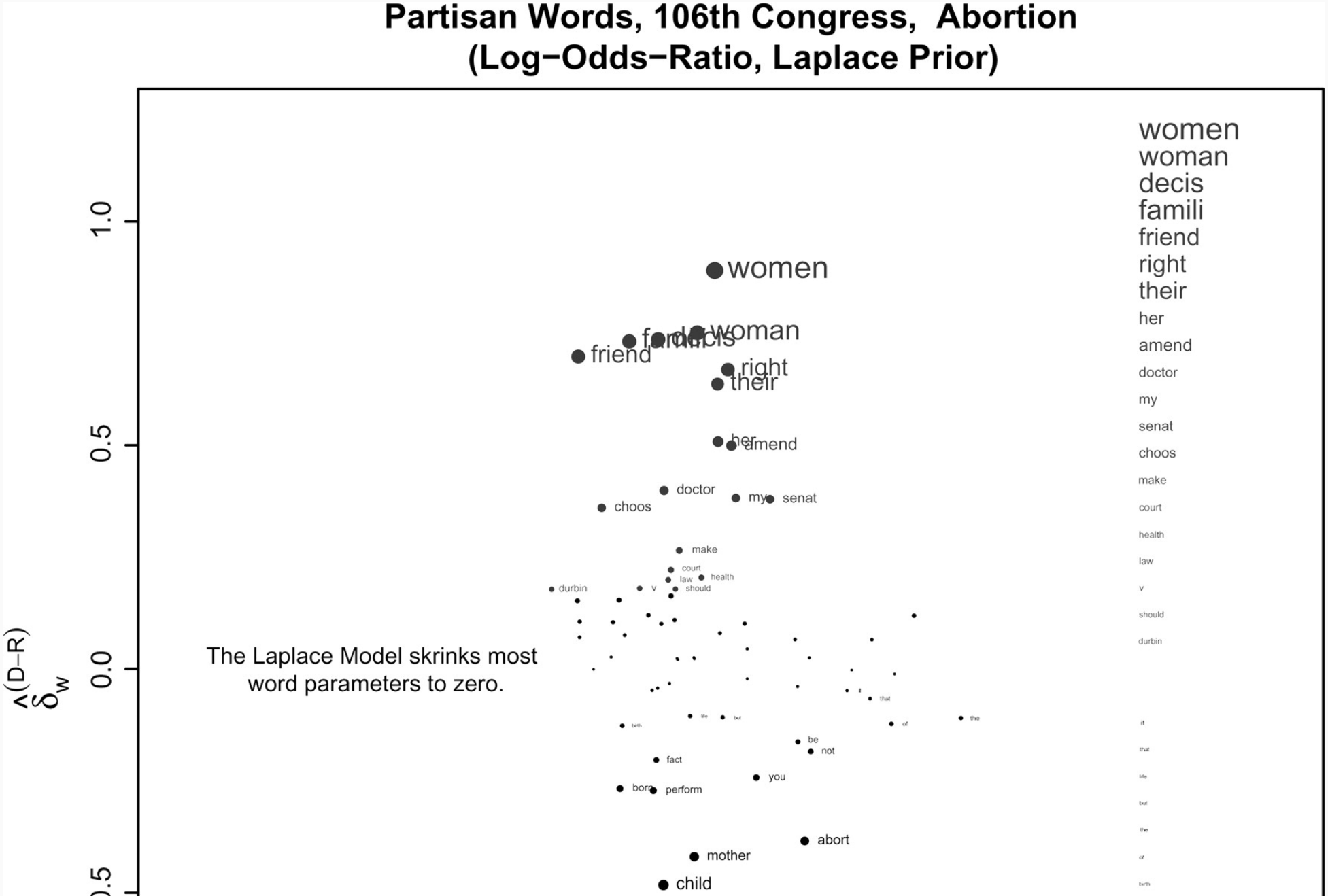
Conservative manifesto generates *market econ* words 45% more often

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

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

Equivalently Labour generates them about *twice* as often

Extensions: Regularised log odds



What to report?

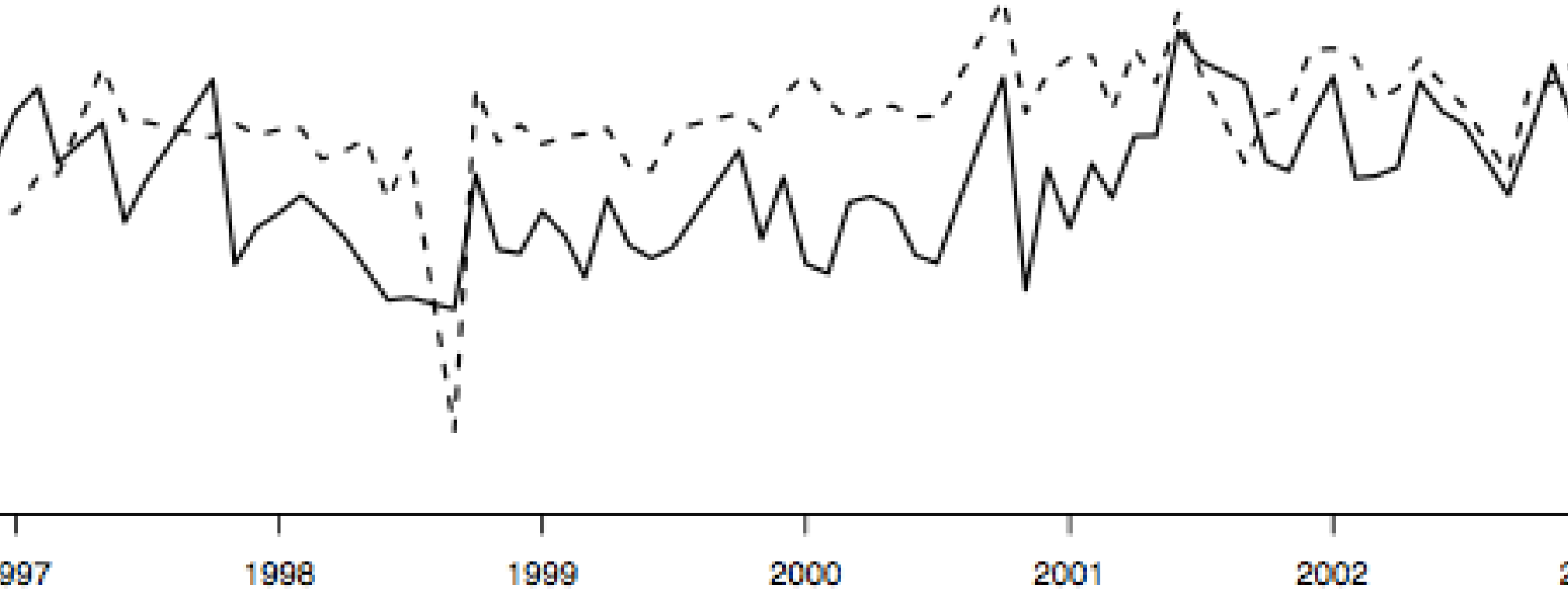
Not all choices are constant or comparable…

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!

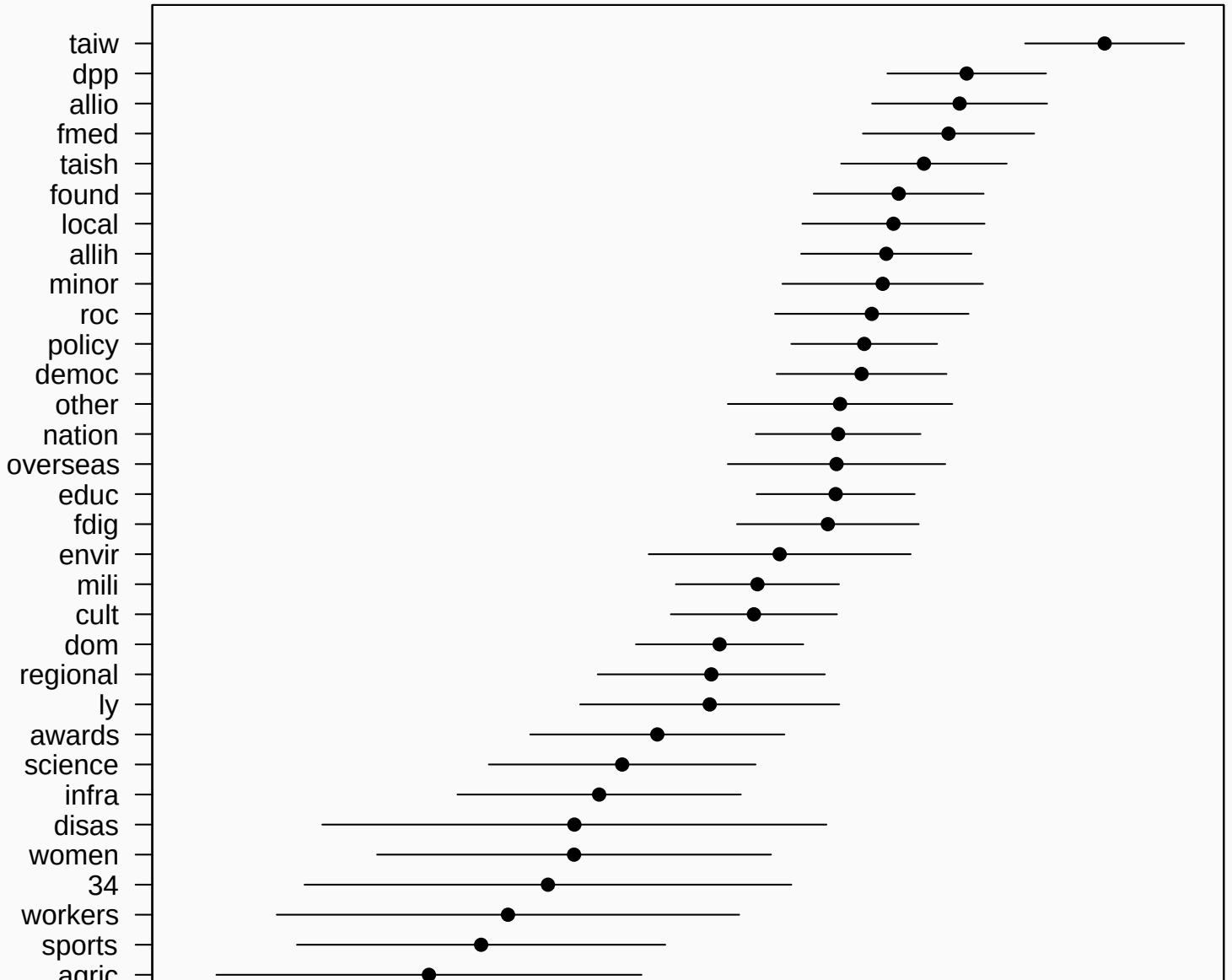
(if length is uninformative & category probability constant)

Content as something to explain

Example: district vs party focus



Content as something to explain (Sullivan and Lowe 2007)



OK, how do I do this myself?

Maximise measurement validity

Minimise *measurement error*

(Sell high, buy low)

Measurement Error

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

	$P(Z = \textit{state reg} \mid \text{W})$	$P(Z = \textit{market econ} \mid \text{W})$
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

	$P(Z = state\ reg \mid W)$	$P(Z = market\ econ \mid W)$
benefit	1	0
assets	0	1

Measurement Error

implies the **confusion matrix**

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

But what if P(W='asset' | Z=*state reg*) > 0?

Then P(Z=*state reg* | W='asset') < 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

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*)

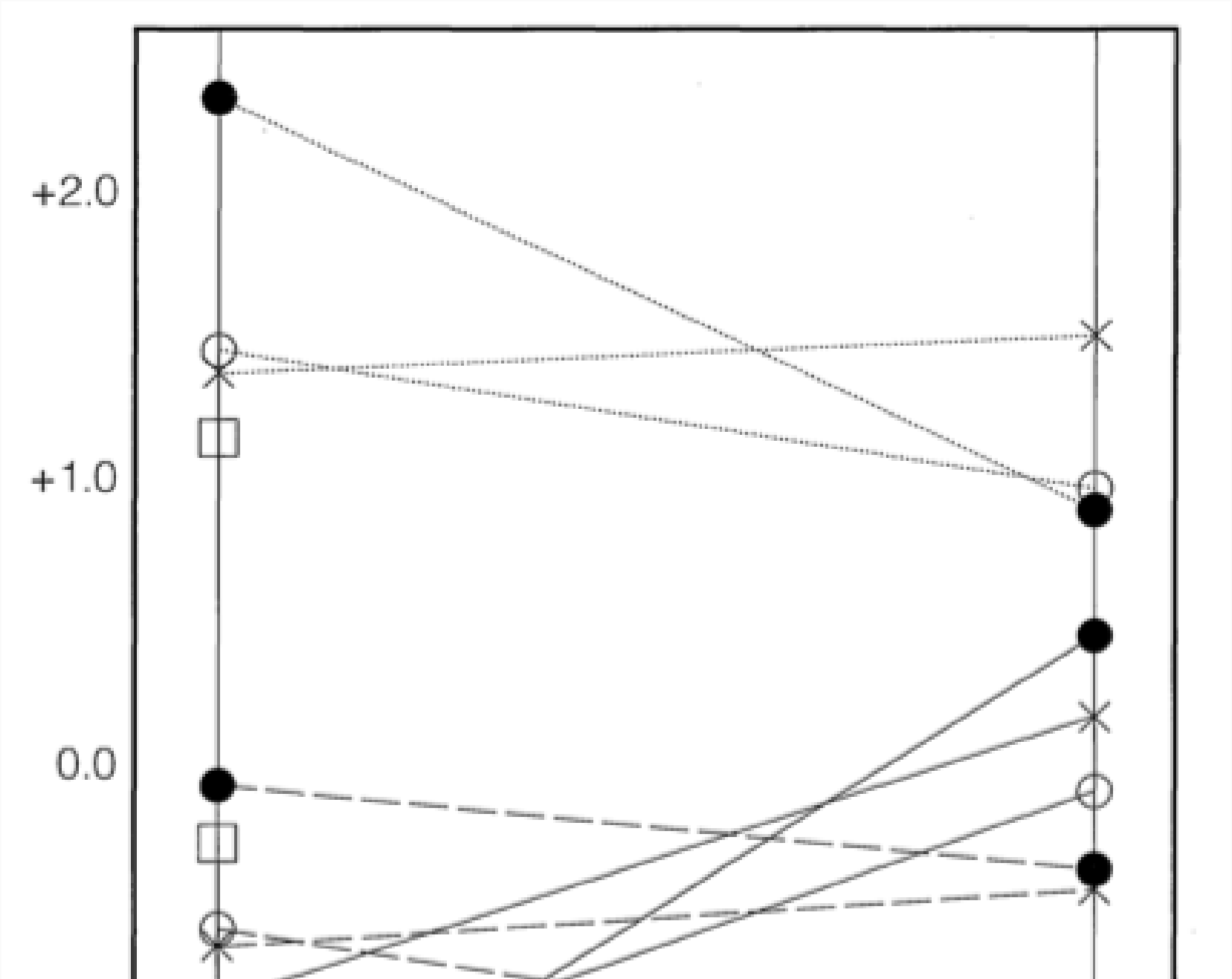
Measurement Error

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)

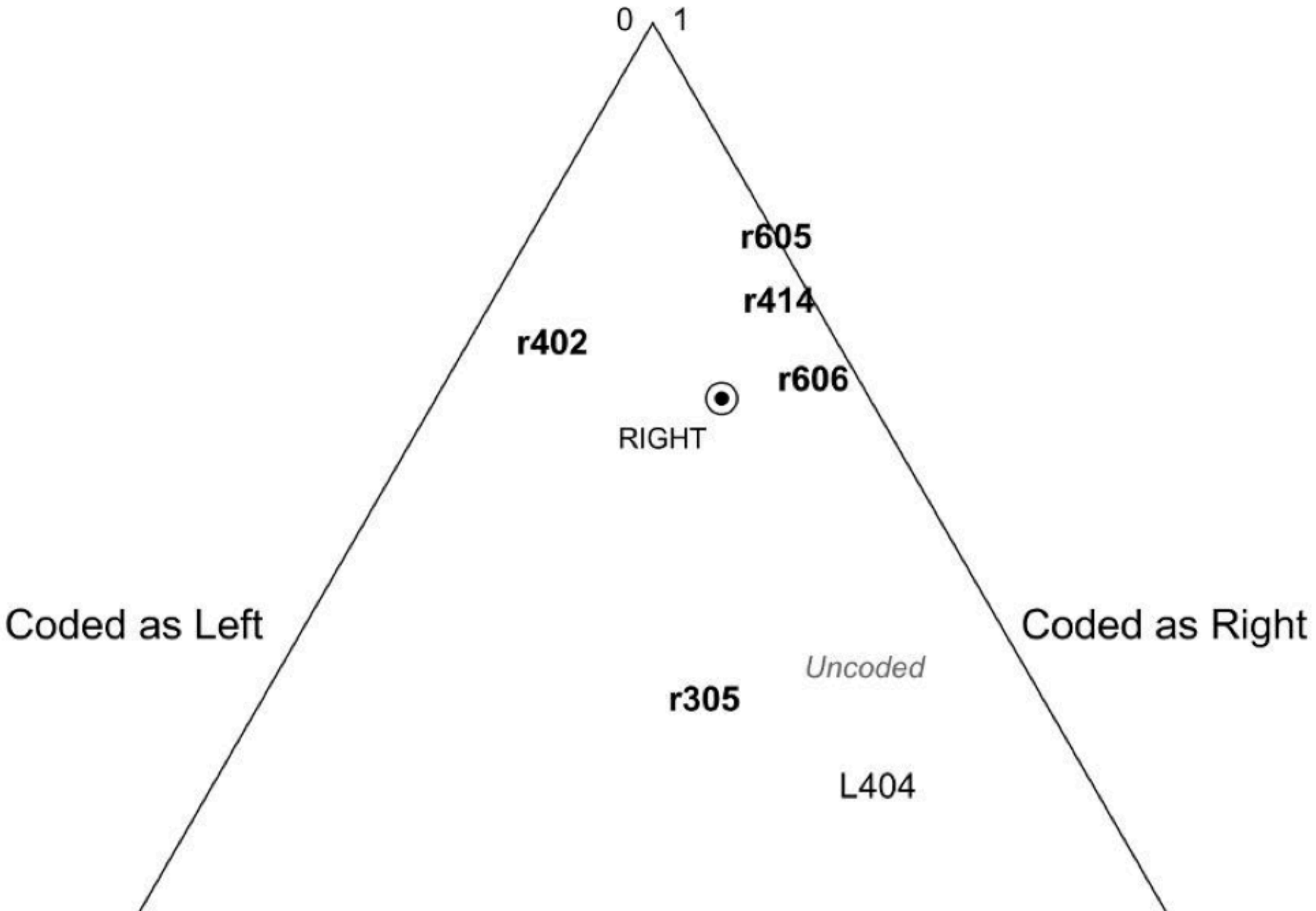


In Action (Mikhaylov et al. 2012)

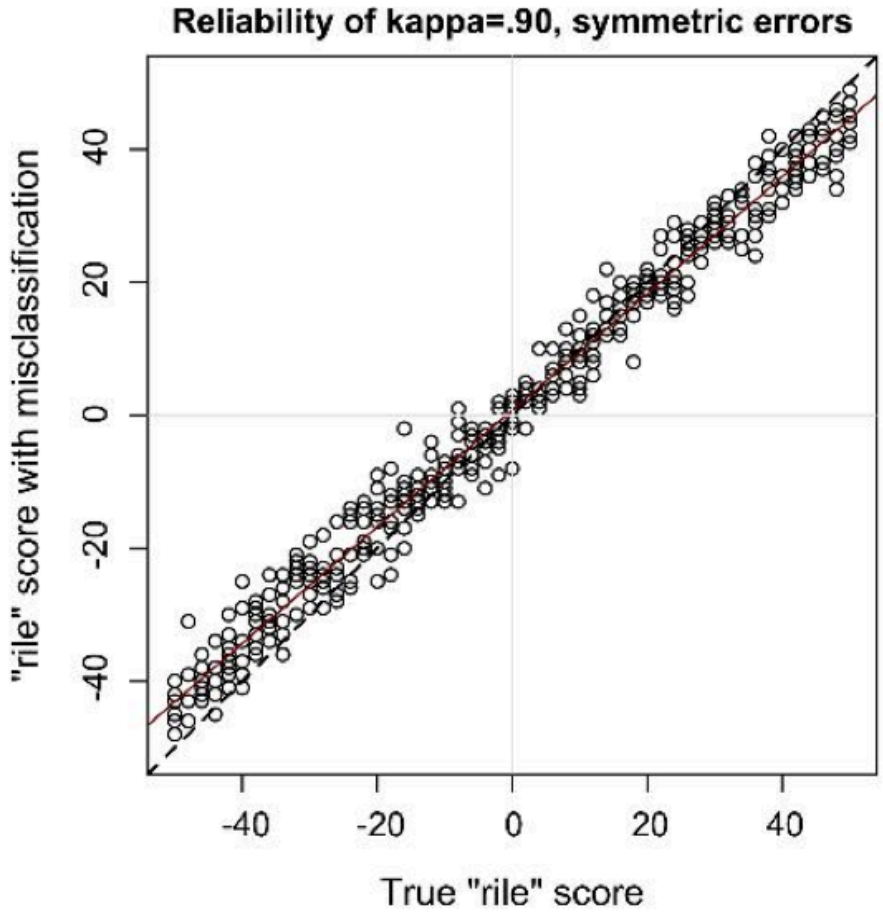
Table 3 Misclassification matrix for true versus observed Rile

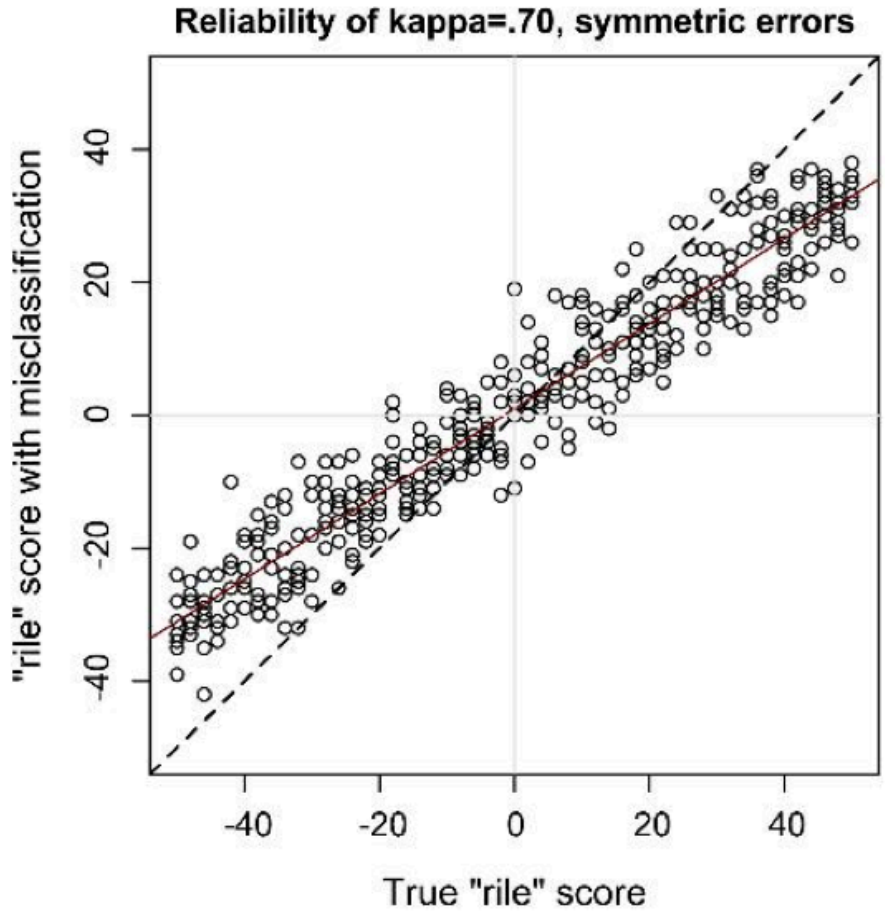
	True Rile category		
	Left	*None*	*Rig*
Left	430	188	10
	0.59	0.19	0.1
None	254	712	19
	0.35	**0.70**	0.2
Right	41	115	65
	0.06	0.11	**0.6**
Total	725	1015	94
lse negative rate	0.41	0.30	0.3

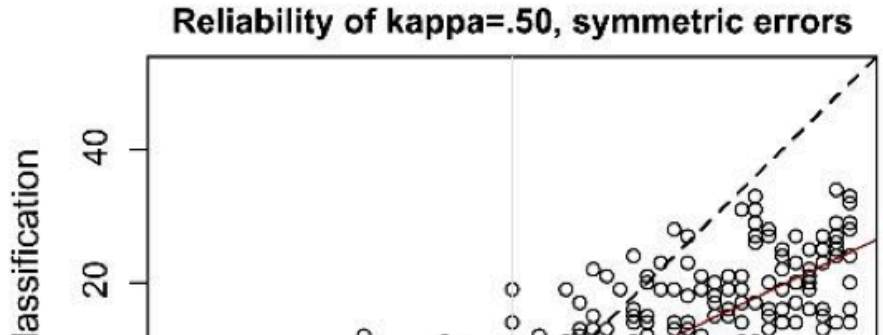
In Action (Mikhaylov et al. 2012)

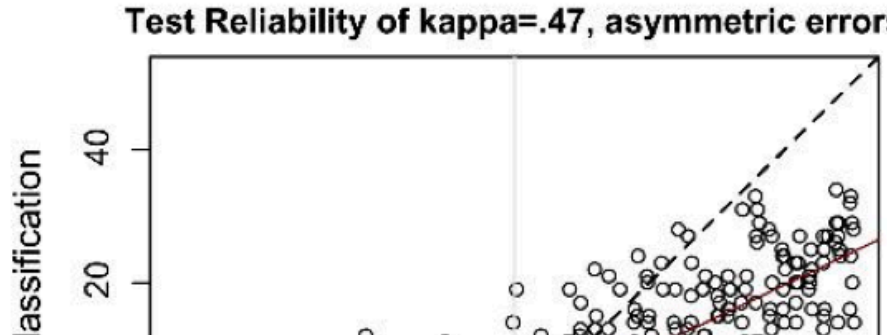


In Action (Mikhaylov et al. 2012)









XXXVI

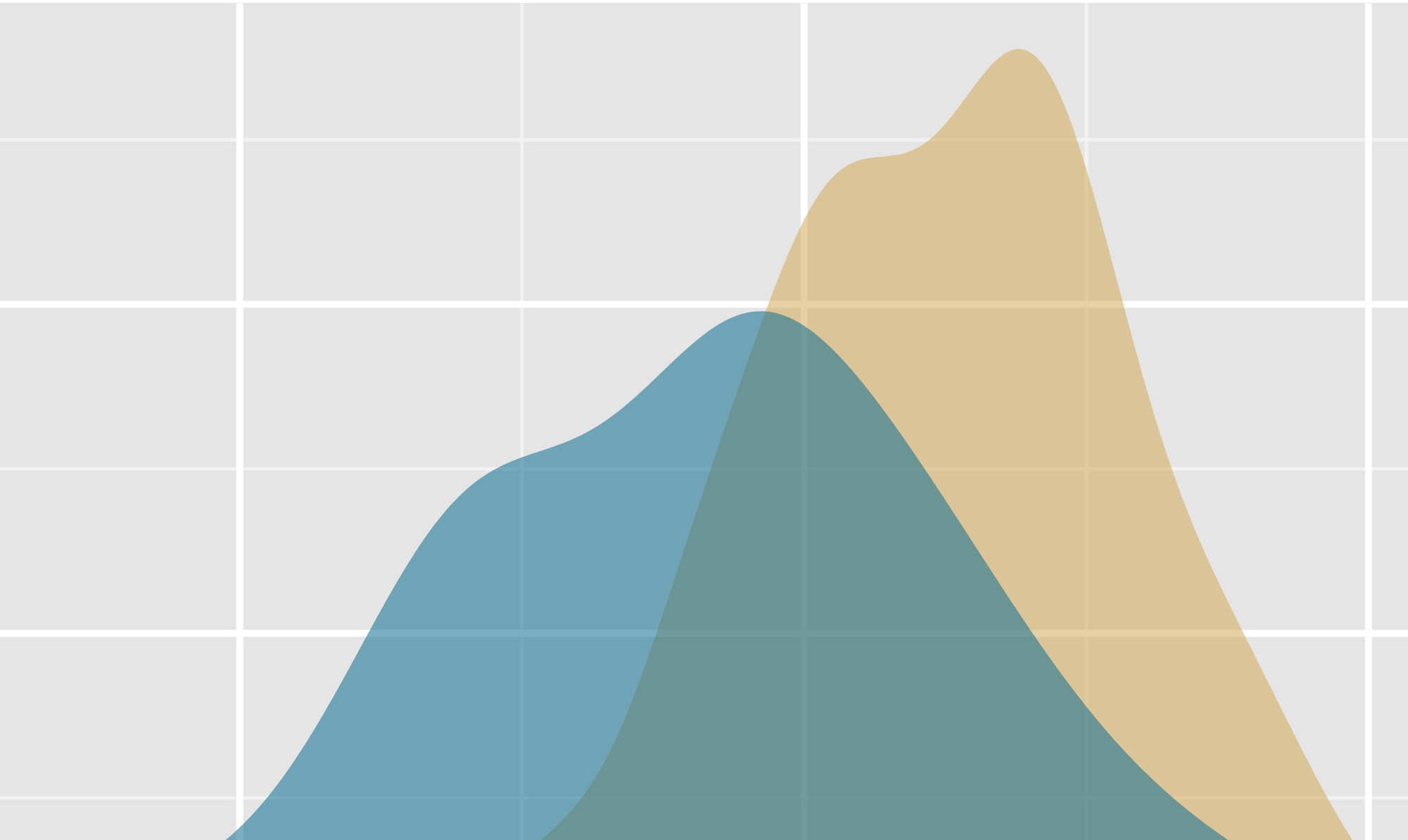
Solutions: Try to Avoid it

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



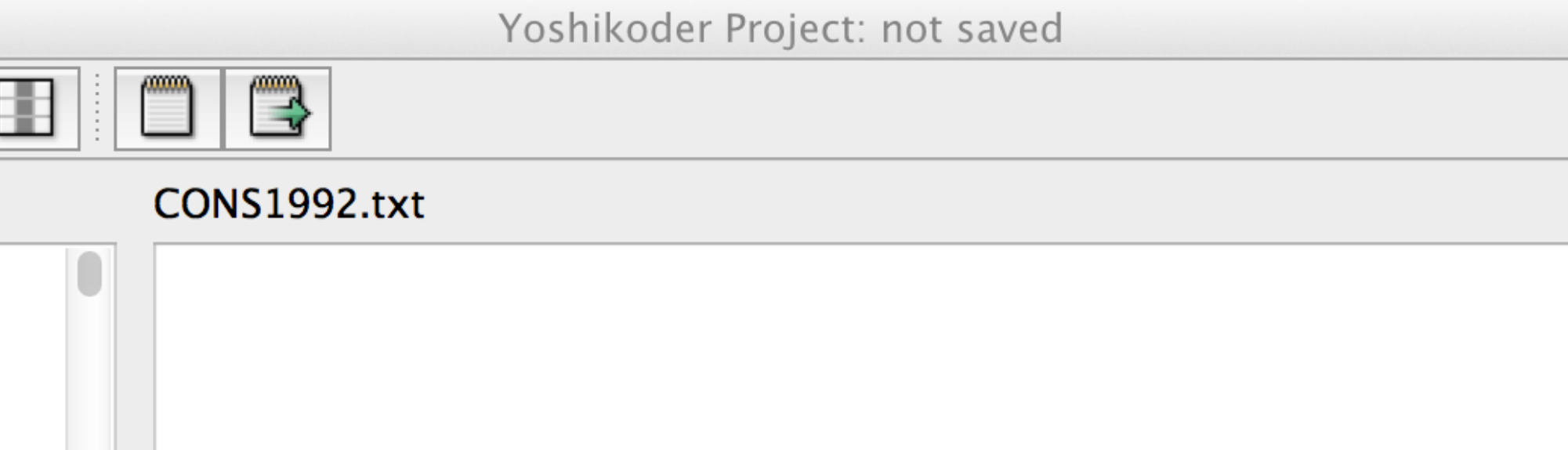
Solutions: Try to Avoid it

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?



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)

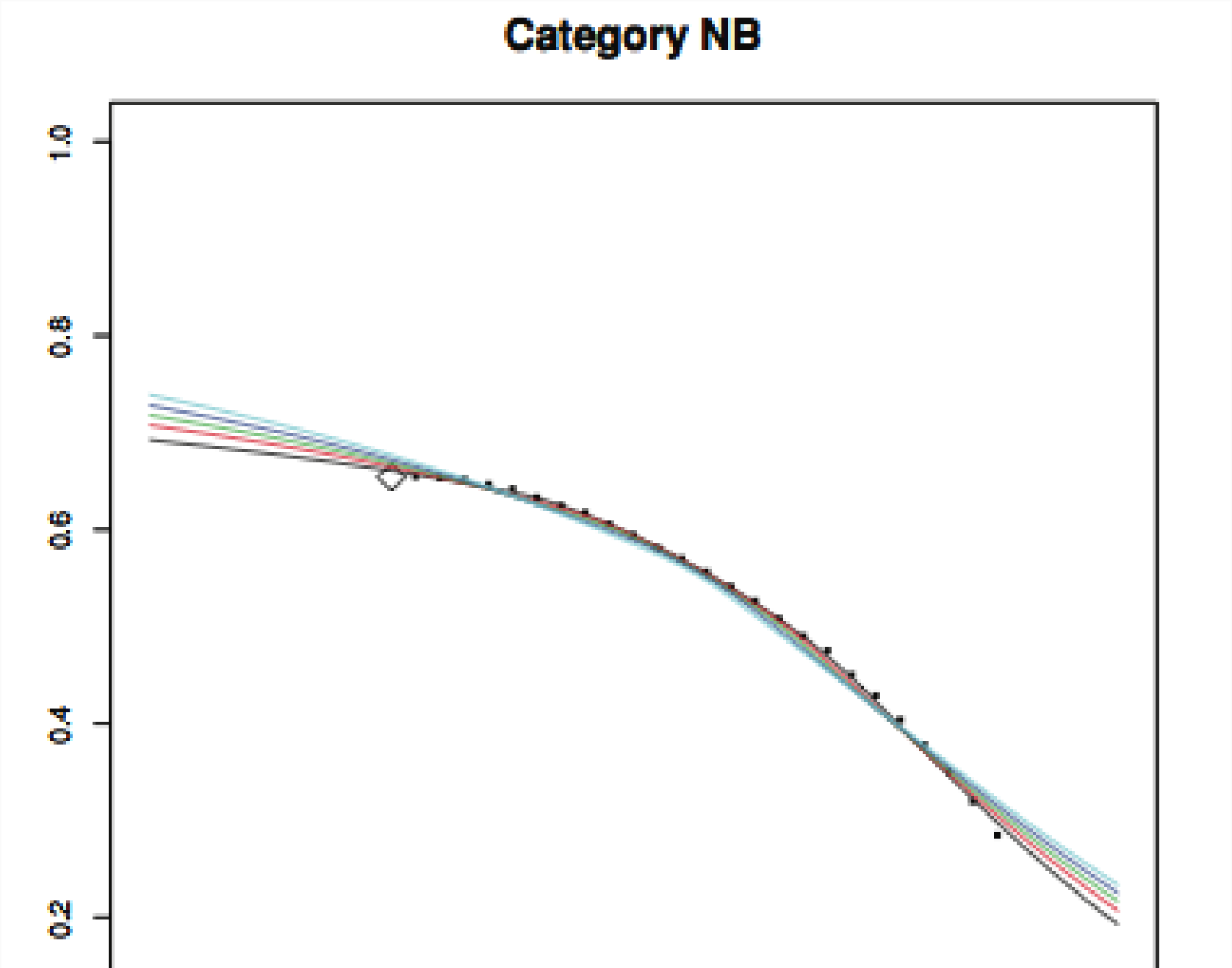
Coder training provides information about how a category proportion estimate changes as coders get better

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 (3)

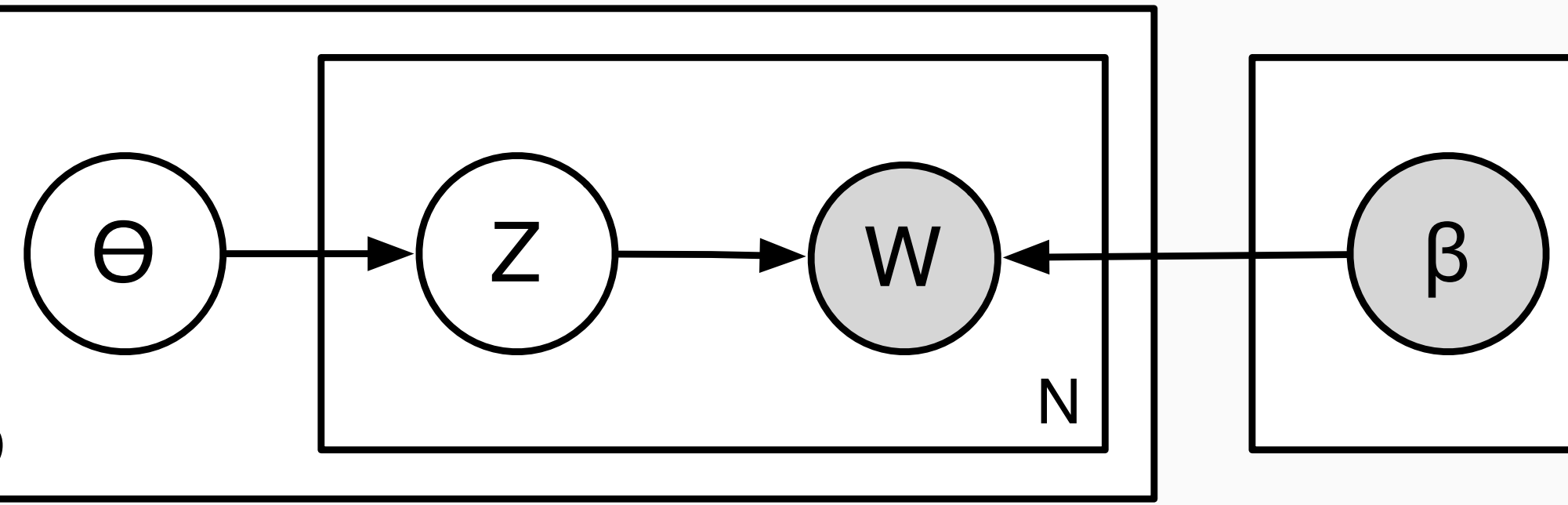
Topic models, e.g. Latent Dirichlet Allocation (Blei et al.) **add**:

A *probabilistic* view of the relationship between W, Z and θ

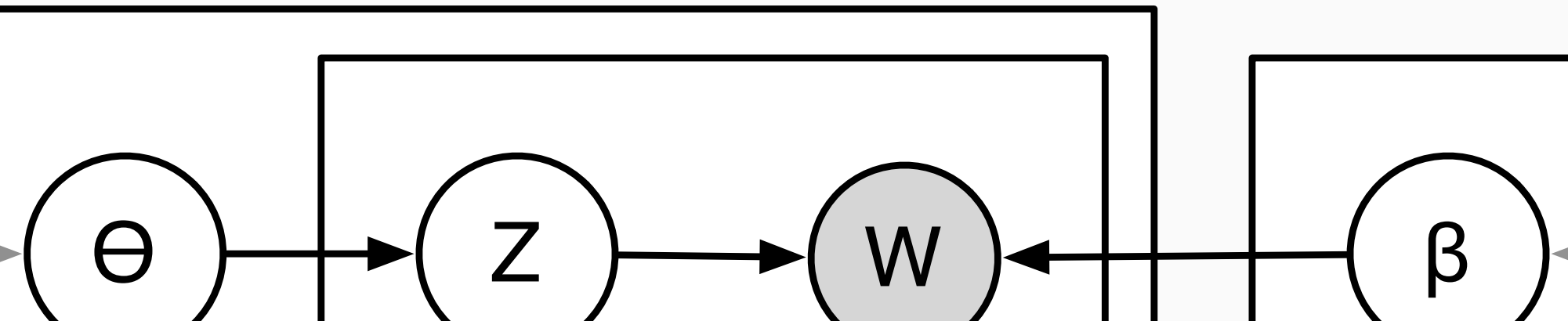
Full statistical framework for learning most aspects of the relationship

Topic Models as Classical Content Analysis

From



(via pLSI) to



Generative probability model

Assumptions:

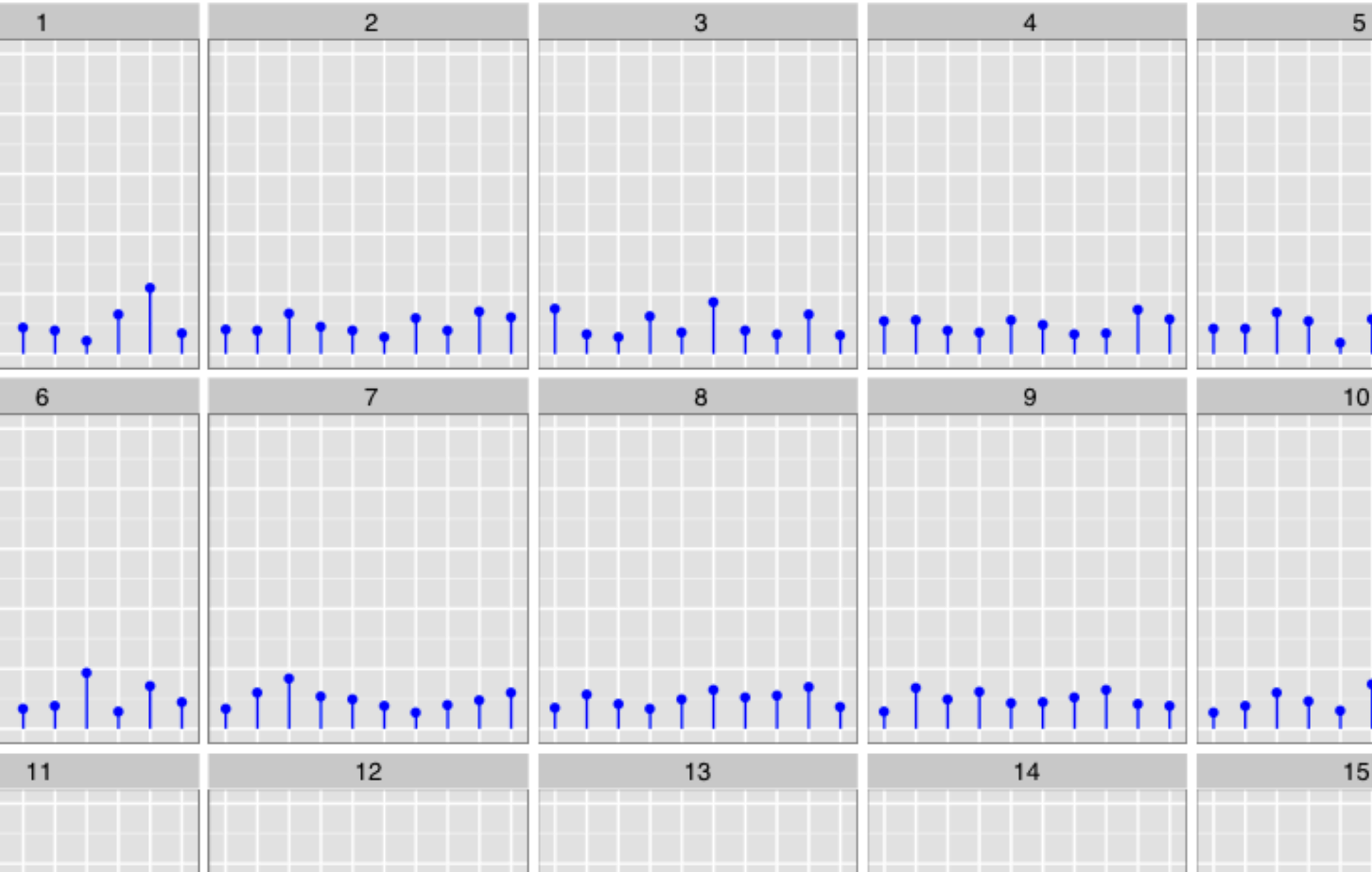
W and Z are multinomial.

θ and β are Dirichlet

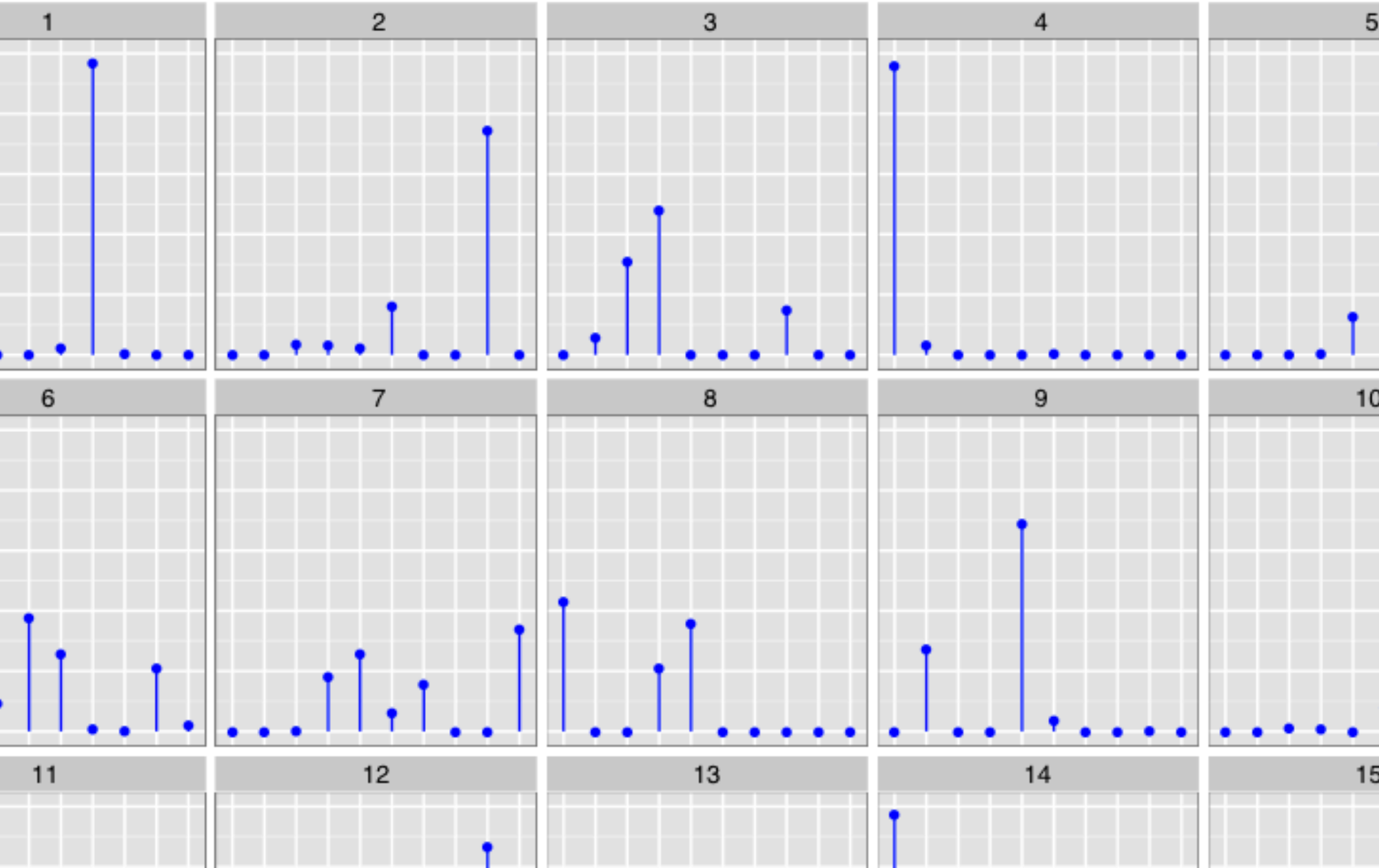
α and η are hyperparameters for Dirichlet parameters

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

$a = 10$



$a = .1$



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*

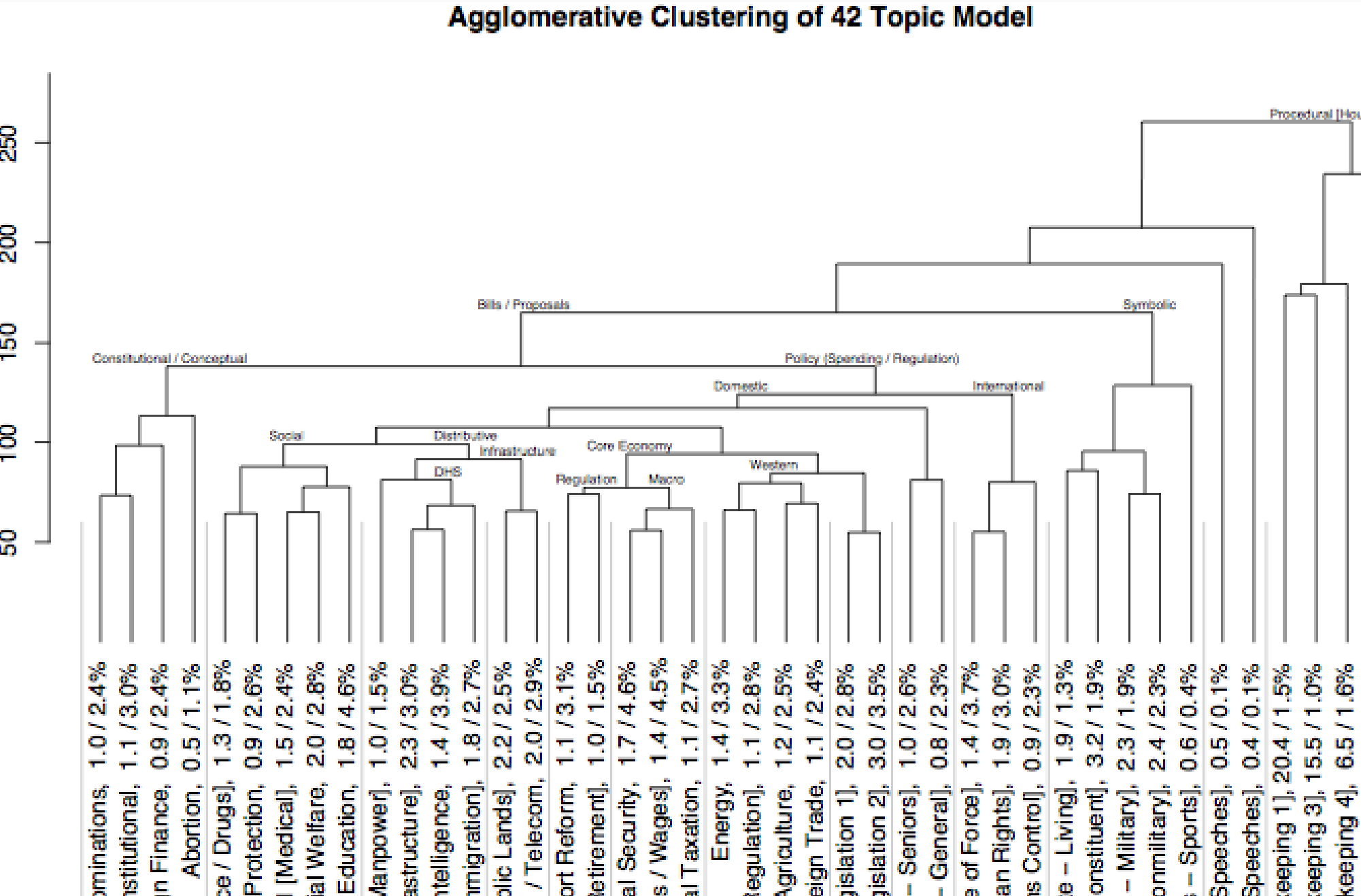
Upshot:

> *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?

	Keys
ons	*nomine, confirm, nomin, circuit, hear, court, judg,*
	case, court, attornei, supreme, justic, nomin, judg,
e	*campaign, candid, elect, monei, contribut, polit, soft,*
	procedur, abort, babi, thi, life, doctor, human, ban,
	enforc, act, crime, gun, law, victim, violenc, abus,
	gun, tobacco, smoke, kid, show, firearm, crime, kill,
]	*diseas, cancer, research, health, prevent, patient,*
	care, health, act, home, hospit, support, children,
	school, teacher, educ, student, children, test, local,
power]	*veteran, va, forc, militari, care, reserv, serv, men,*
structure]	*appropri, defens, forc, report, request, confer, guard,*
	intellig, homeland, commiss, depart, agenc, director,
]	*act, inform, enforc, record, law, court, section,*

How is it related to others topics?



and other schemes?

y Classification Systems

lusters

onceptual (Partisan conflict)

stributive (Common good / State v state)

urban

v them)

pinksi (2006)

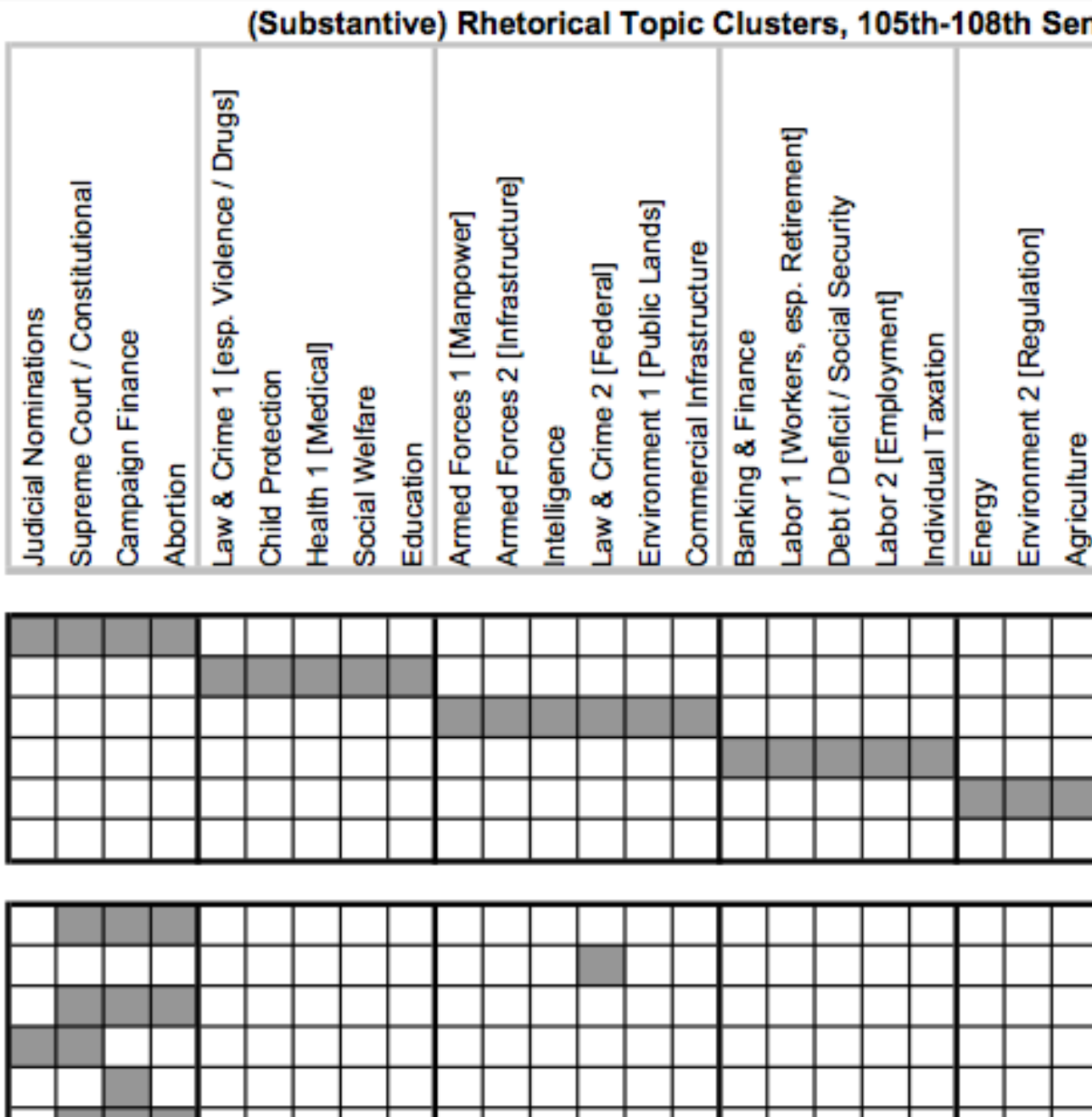
rty]

nbership / nation]

rights]

pe [Governmental org.]

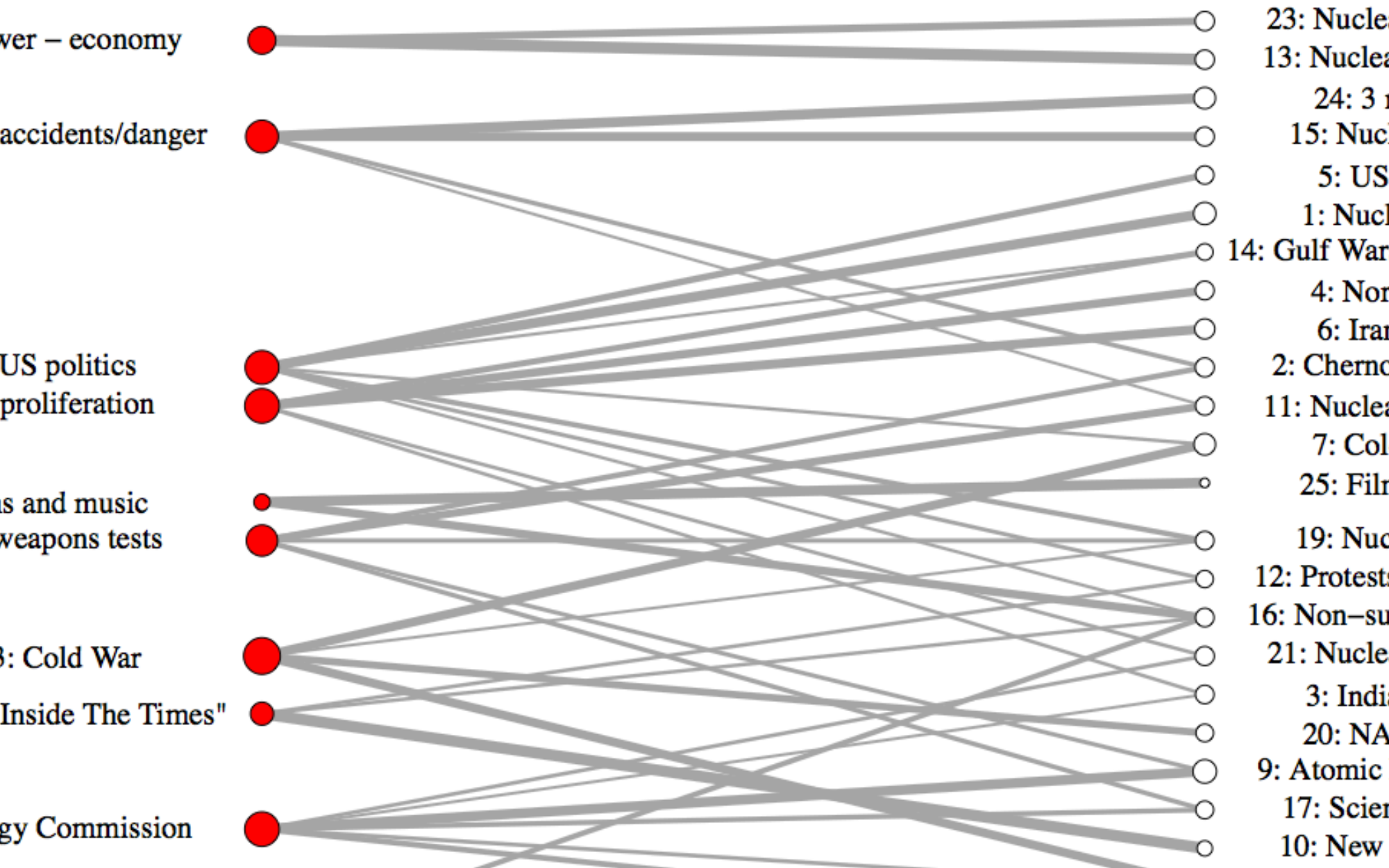
pe [Representation]



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?



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

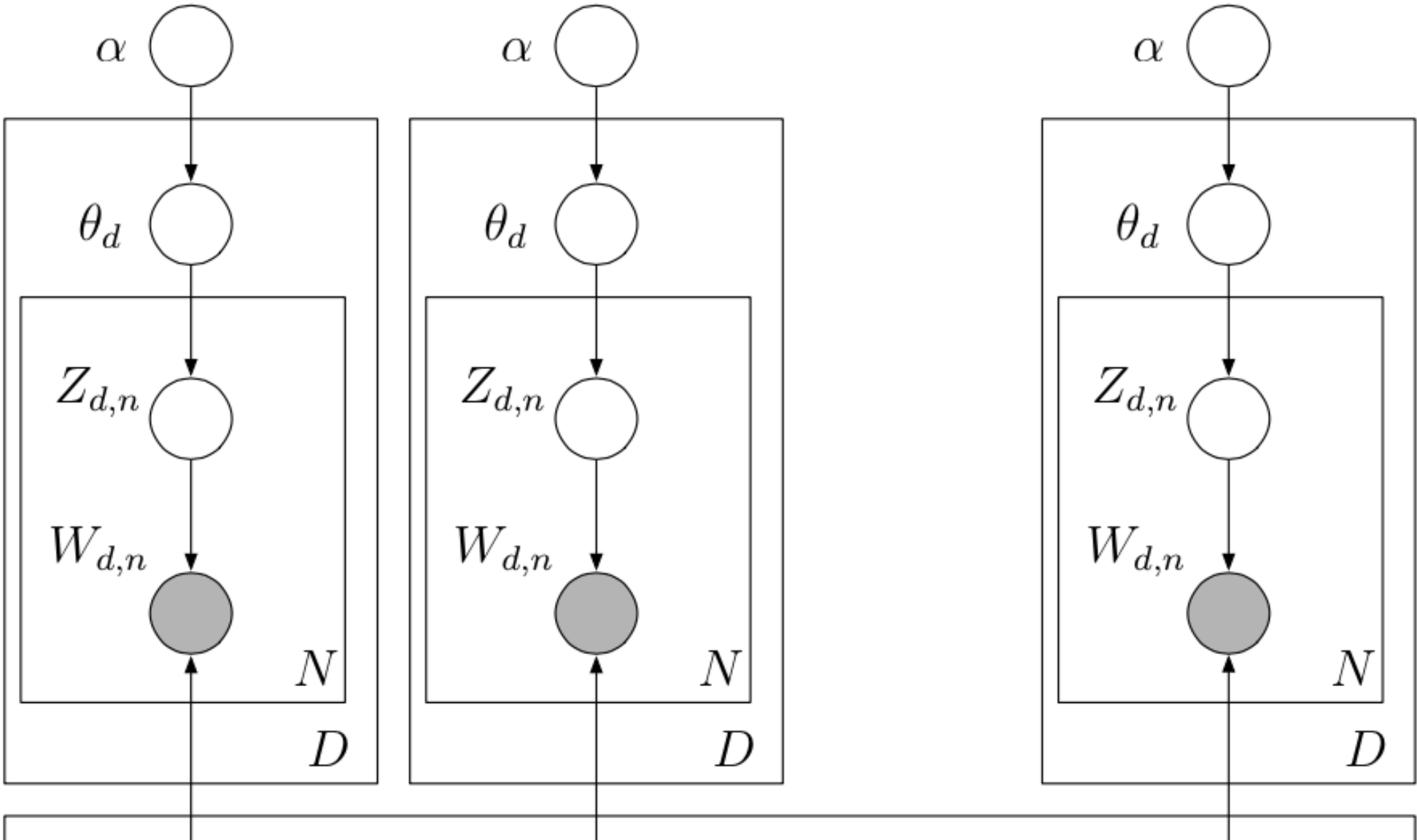
> Airoldi et al. 2011 Table 1 provide references for 33!

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

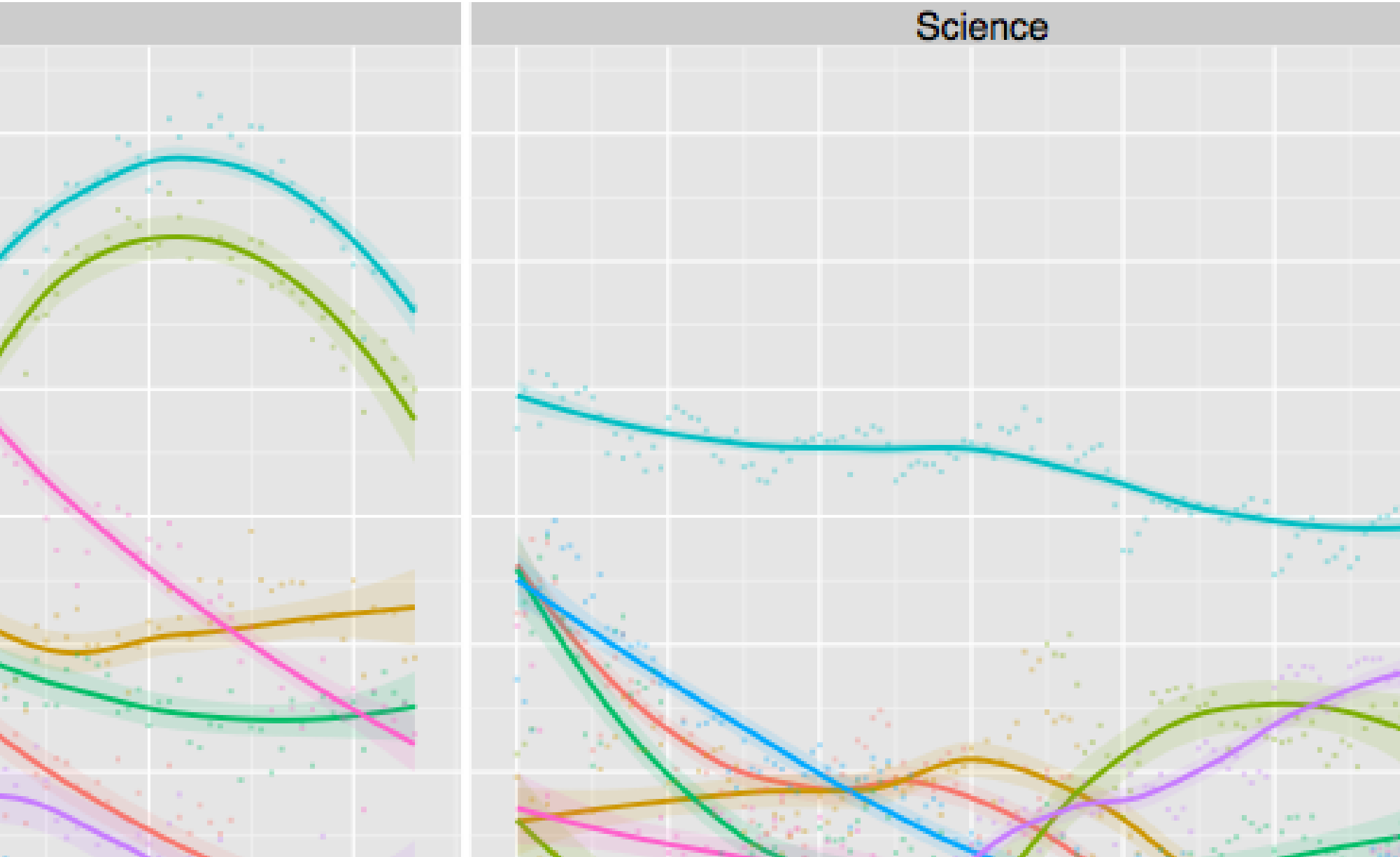
> Dynamic topic models
>
> Structural topic models (`http://structuraltopicmodel.com`)

Dynamic Topic Models (1)

Words ‘drift’ from topic to topic over time



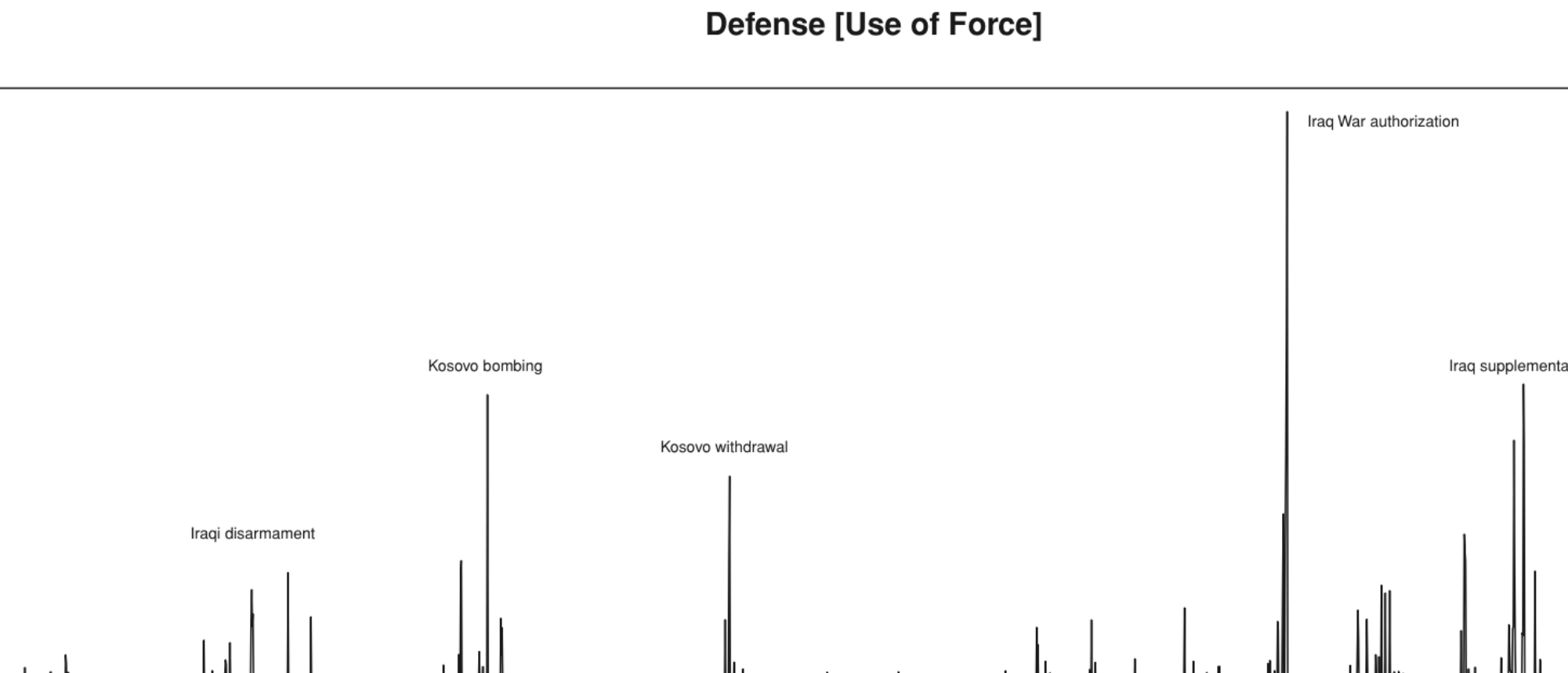
Dynamic Topic Models (1)



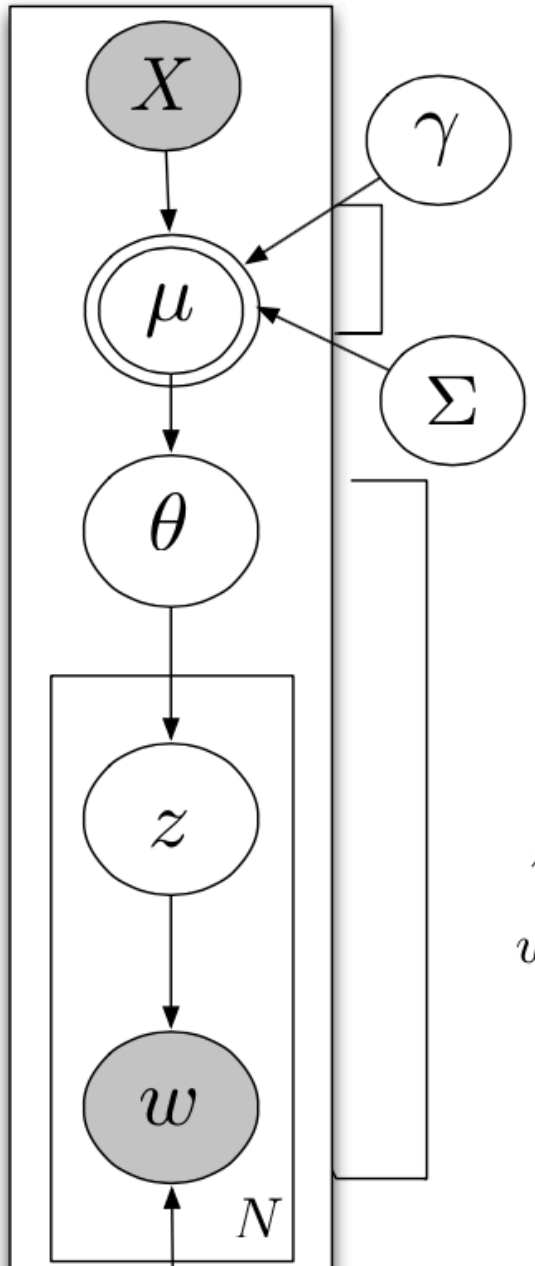
Dynamic Topic Models (2)

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

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



Structural Topic Models



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:

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

-1.25

F: Muslim, Jihad, Islam, fight, Jihadi fighters, pathway, almighty, that
FREX: jihad, fighting, jihadist fighters, pulpit, approves of us, annotated, to fight, vicinity
F: مسلم, جهاد, اسلام, قتل, مجاهد, سبيل, تعال, ذين FREX: جهاد, قتال, مجاهد, منبر, يوافقنا, مذيل, يقاتل, بجوار

F: person, life, soul/self, knowledge/science, society, work, image, material/physical
FREX: imagine, morals, develop, society, product,necessarily, environment, traditions, activity
F: انس, حيا, نفس, علم, مجتمع, عمل, صور, ماد FREX: تصور, اخلاق, تطور, مجتمع, انتاج, حتم, بيئ, تقاليد

F: Arab, Jews, country, Islam, A.D., year, West, Muslim
FREX: capitol, Asia, Iran, South, Washington, A.D., Russia, Turkey
F: عرب, يهود, دول, اسلام, م, سن, غرب, مسلم FREX: عاصمت, اسيا, ايرا, جنوب, اشنطن, م, روسيا, تركيا

F: said, prayers (be upon him), peace (be upon him), almighty, messenger, glory, prophet, that
FREX: almighty, almighty, glory, bless you, magic, punishment, hypocrisy, sins
F: قال, صل, سلم, تعال, رسول, سبح, نب, ذين FREX: وجل, عز, سبح, تبارك, سحر, عذاب, رياء, ذنوب

F: prayer, pray, son, prophet, sheikh, mosque, fatwas, group
FREX: prostration, prostrated, Abd al-Aziz, supplicant, Baz, prayer space, omission, prostration
F: صلا, صل, سلم, بن, نب, شيخ, مسجد, فتاو FREX: ركع, ركعت, عبدالعزيز, ماموم, باز, مصل, سهو, ركوع

F: day, fasting, Ashura, Ramadan, sheikh, group, fatwas, Uthaymeen
FREX: wash, one who fasts, fasting, fasting, to break fast, Ramadan, travel, dirty
F: يوم, صيام, عشر, رمض, شيخ, مجموع, فتاو, عثيم FREX: غسل, صائم, صيام, صوم, يفطر, رمض, مسافر, نجاس

Wrapping up

Lab Time