

Winter Institute: Computer-Assisted Content Analysis

Will Lowe
University of Mannheim

Practicalities: Materials

We have a website:

<http://conjugateprior.org/teaching/texas/>

Password: quantext

Practicalities: Labs



Worked examples (usually in the tools themselves and also in an R wrapper)

Menu

Session 0: How could this possibly work?

Session 1: Dictionary-based ‘classical’ content analysis
and 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

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

Focus

Assumptions

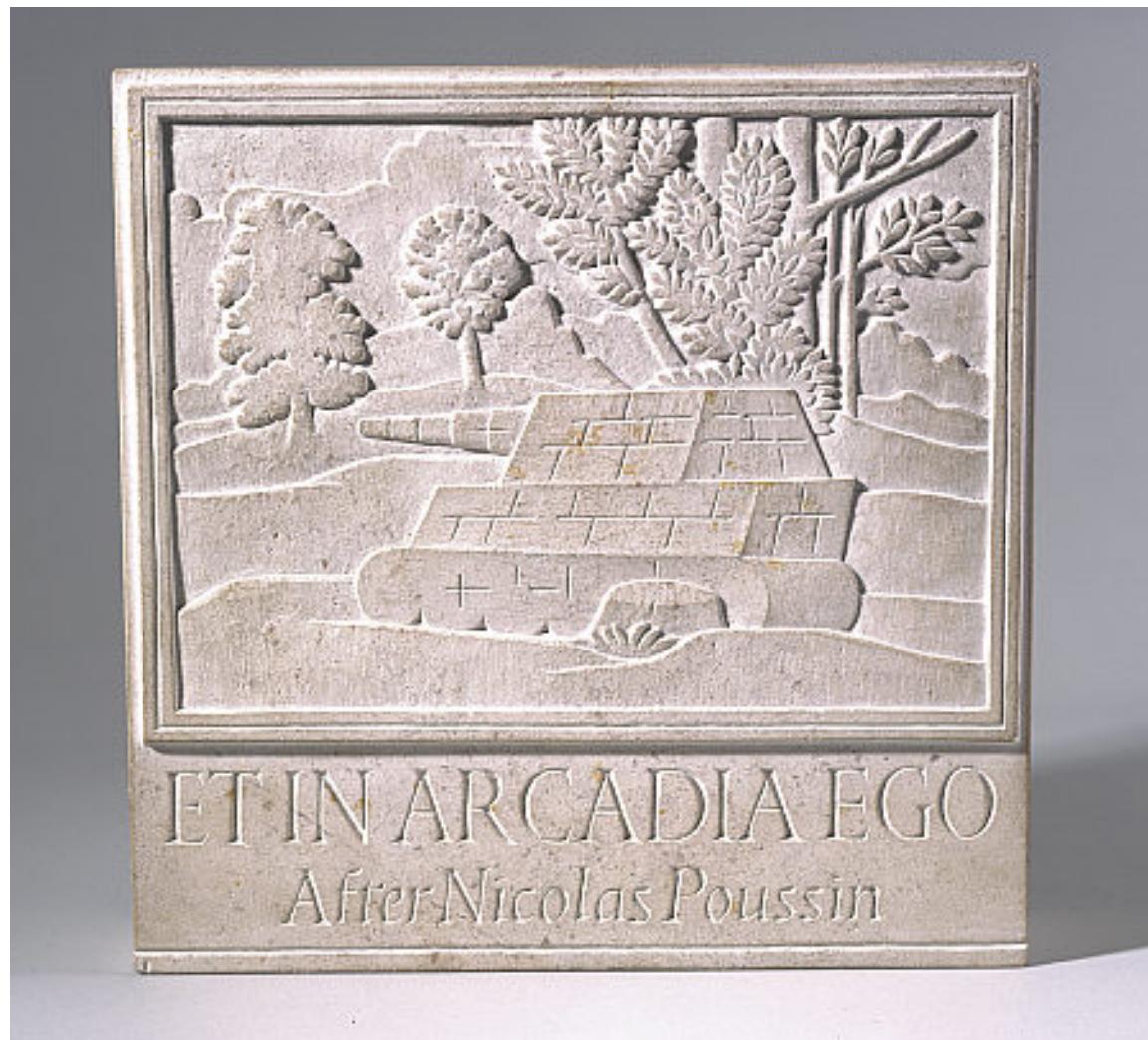
Mechanics

Interpretation

Pitfalls

Topics

How to learn about
party platforms
legislative agendas
parliamentary debates
bloggers
presidents
international terrorists
by counting (lots of) words...



Texas A&M, January 2015

The Transcendental Question

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

a.k.a how could this possibly work?

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.

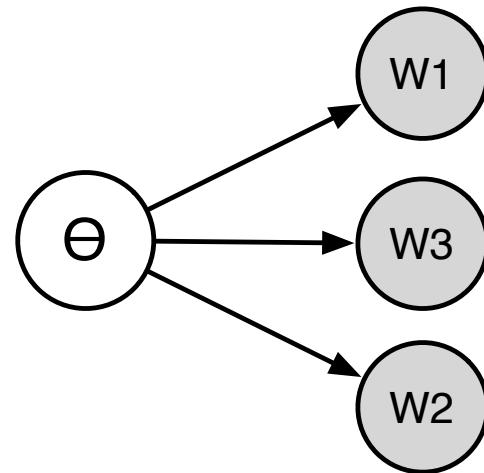
and *behaviour*, including *linguistic behaviour*, e.g.

yelling, muttering, cursing, lecturing
which *can* be directly observed.

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

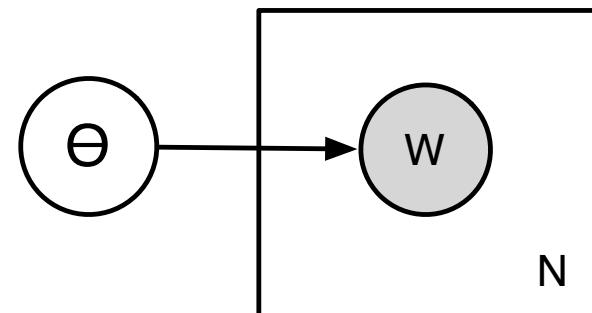
Communication

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



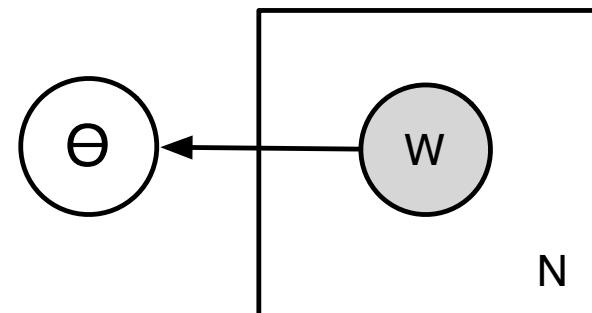
Communication

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*.

Words as Data

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...

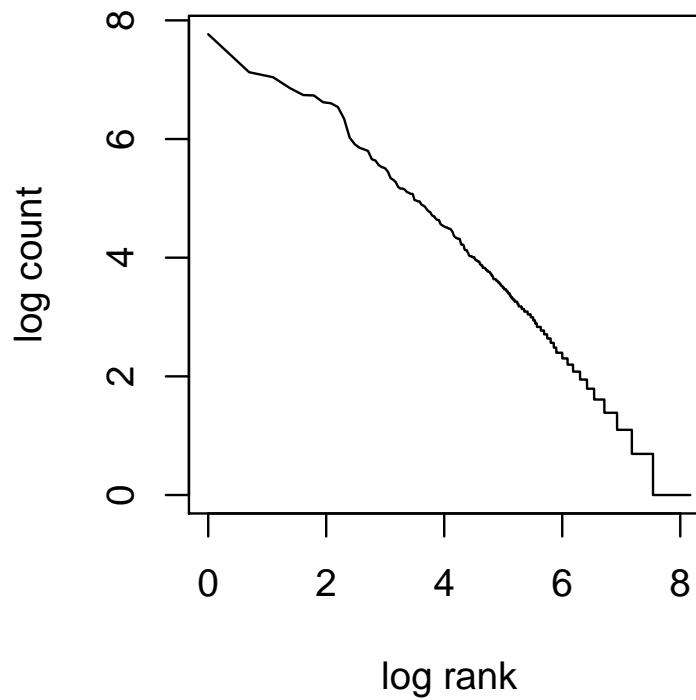
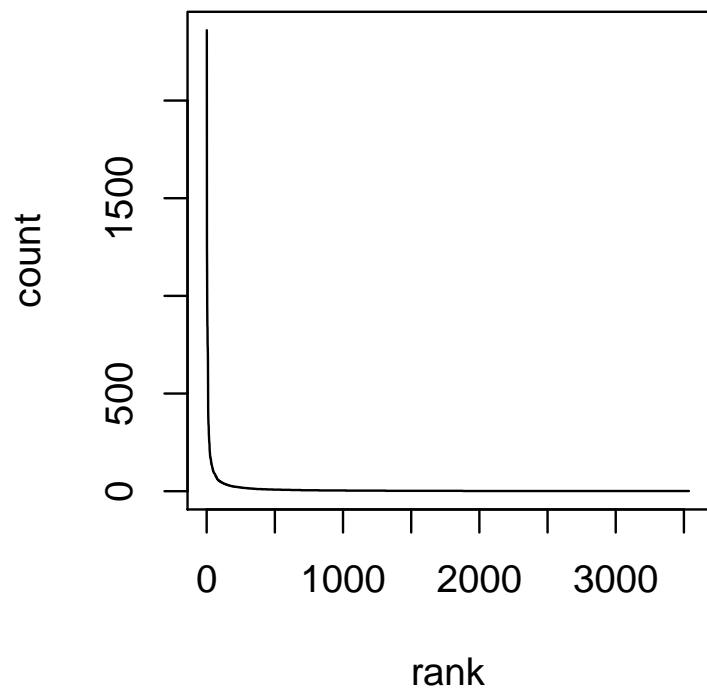
Zipf-Mandelbrot law (a pareto distribution in disguise)

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

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

Very fat tailed...

Zipf's Law



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	
		'doc' 1	'doc' 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
doc 1	1	1	1	2	0	1	$\dots \theta_1$
doc 2	0	0	0	1	1	0	$\dots \theta_2$

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



Texas A&M, January 2015

Statistical Models of Words: Poisson

These models are going to think of content as anything that systematically alters word and topic rates

Statistical Models of Words: Multinomial

For fixed document length counts are conditionally Multinomial:

$$P(W_1 \dots W_V) = \text{Multinomial}(W_1 \dots W_V; \pi_1 \dots \pi_V, N)$$
$$\frac{N!}{W_1! \dots W_V!} \prod_j \pi_j^{W_j}$$

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



Texas A&M, January 2015

Statistical Models of Words: Multinomial

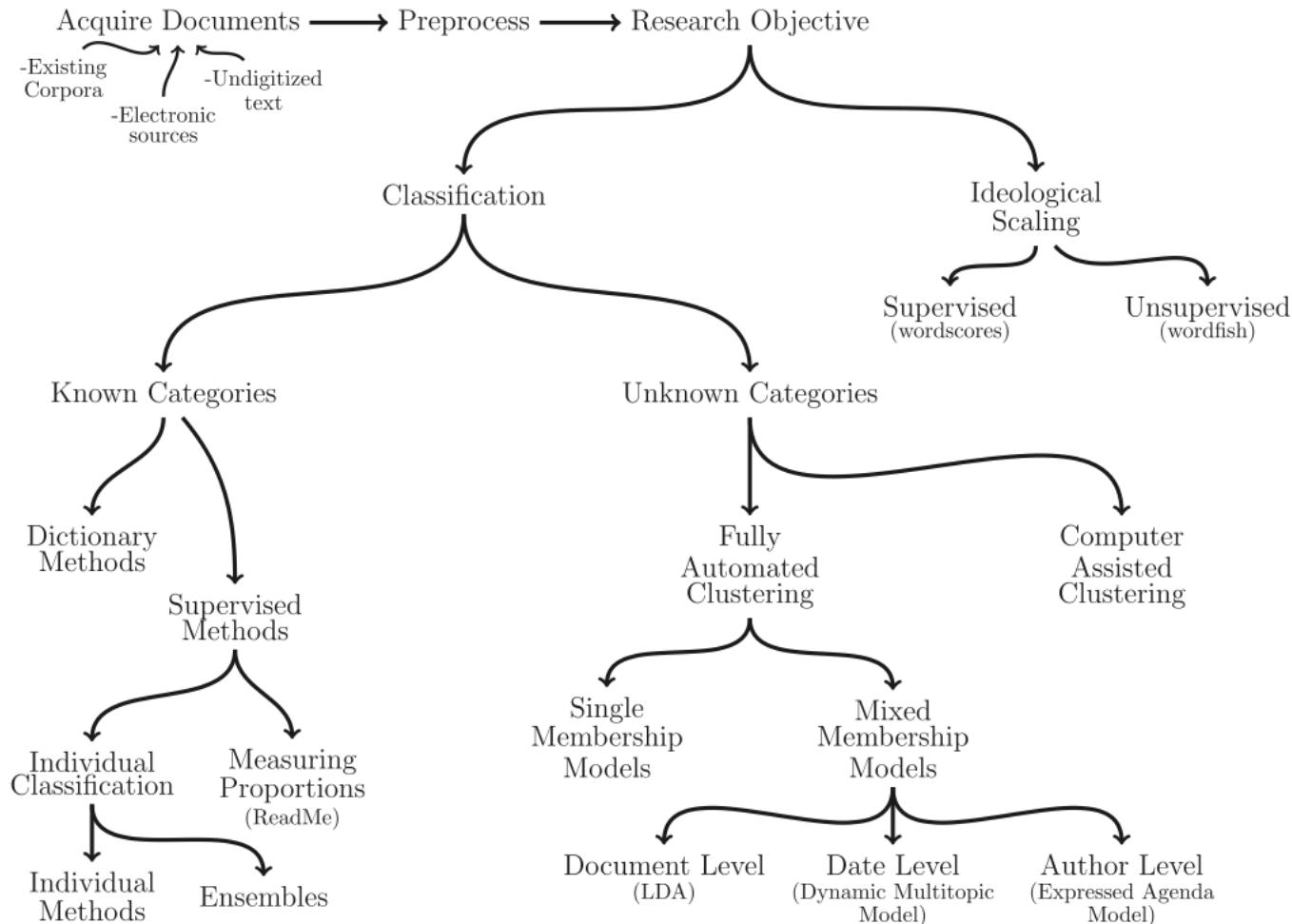
We are going to think about *content topics* as like this

We build measurement devices

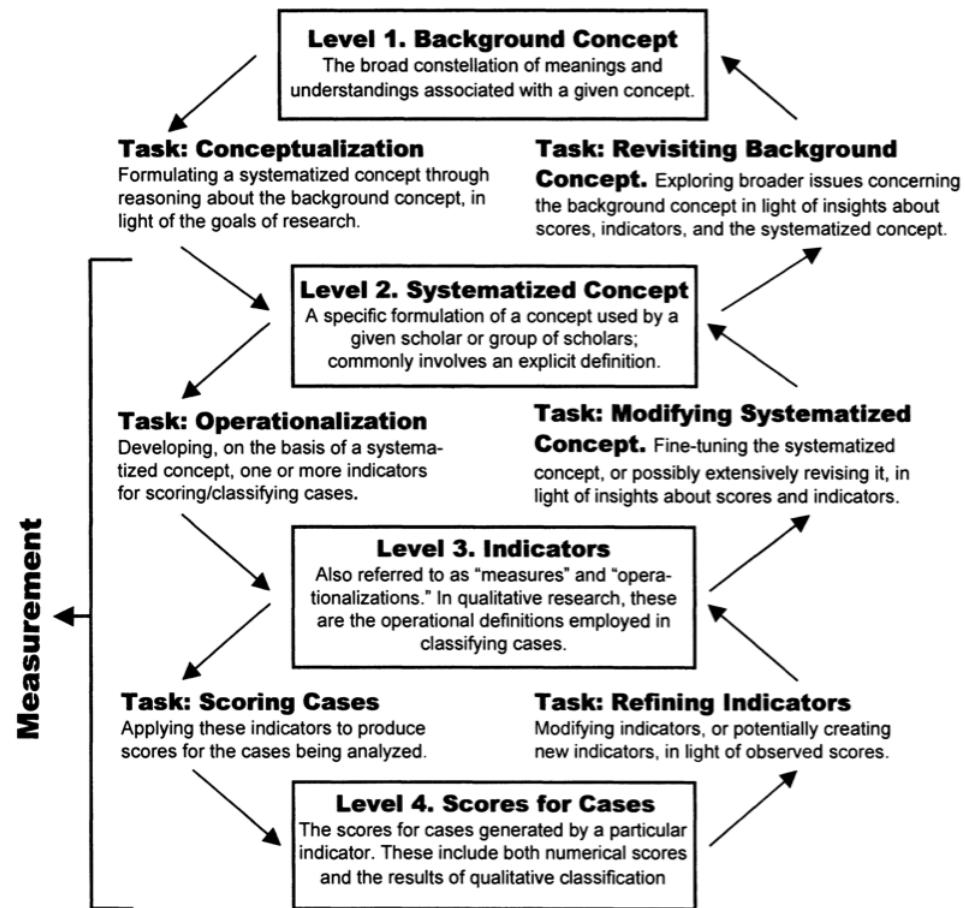


Texas A&M, January 2015

(Almost) the big picture



Really the big picture



Modeling strategies

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

$\theta \leftarrow$ words: Go for $P(\theta | \text{words})$ *directly*

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

$\theta \rightarrow$ words: Get $P(\theta | \text{words})$ *indirectly*

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

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

Modeling Strategies

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

That is to say

the next three days

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)



Texas A&M, January 2015

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 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.

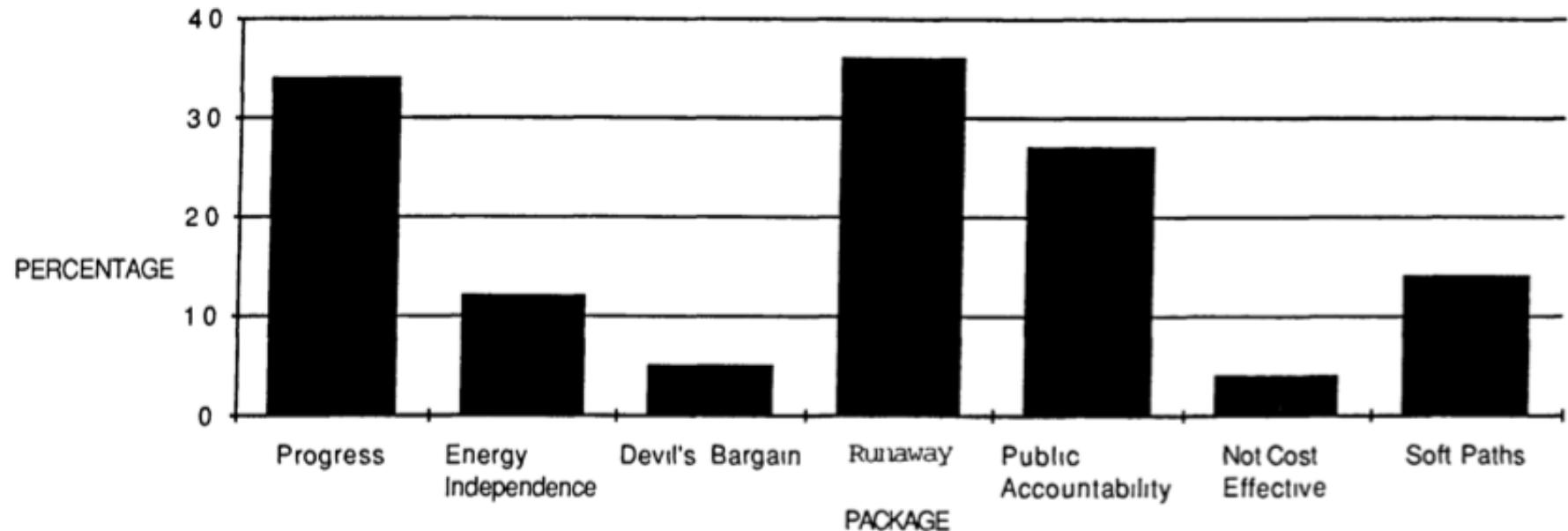
pro-welfare state vs. anti-welfare state, lots of CMP categories

But they are invariably treated as *nominal level* variables

We are typically interested in them for

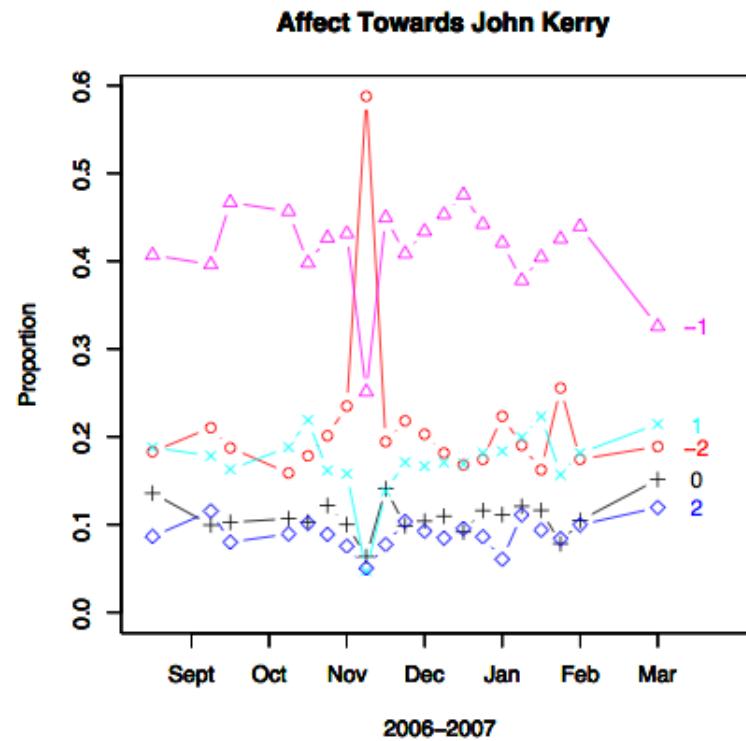
simple descriptions, making comparisons, tracing temporal dynamics

Talking Like a Newspaper



Gamson and Modigliani (1989)

Talking like a Candidate



Hopkins and King (2010)

Texas A&M, January 2015

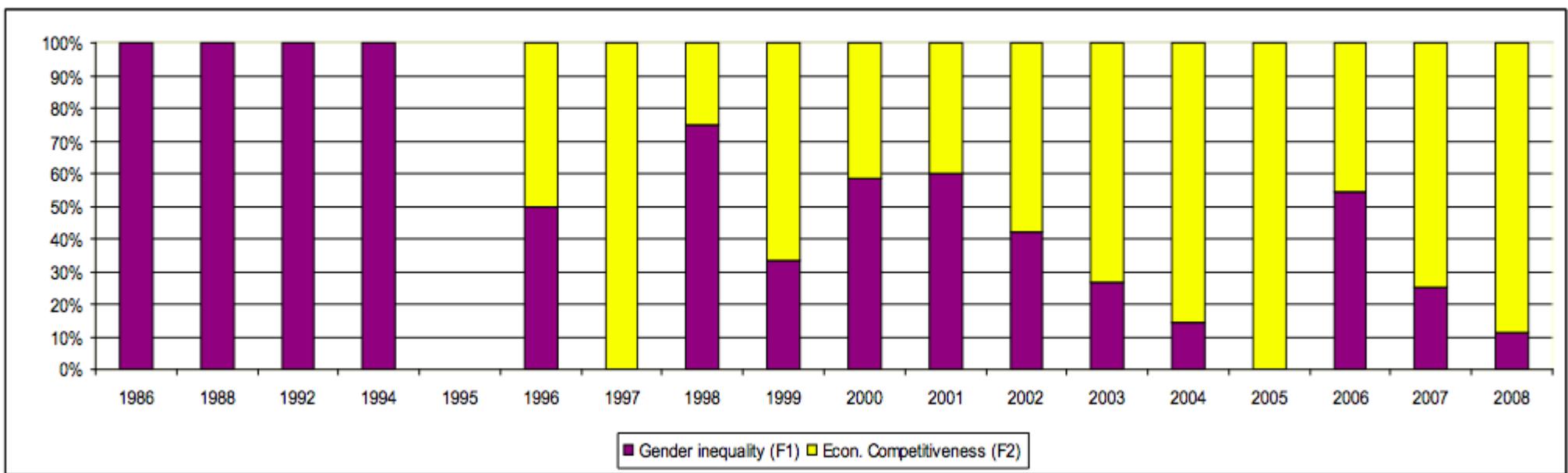
Talking like a Terrorist

	Bin Laden (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 Laden, Zawahiri, and Controls. Documents whose source indicates "Both" (n=3) or "Unknown" (n=2) were excluded due to their small sample sizes.

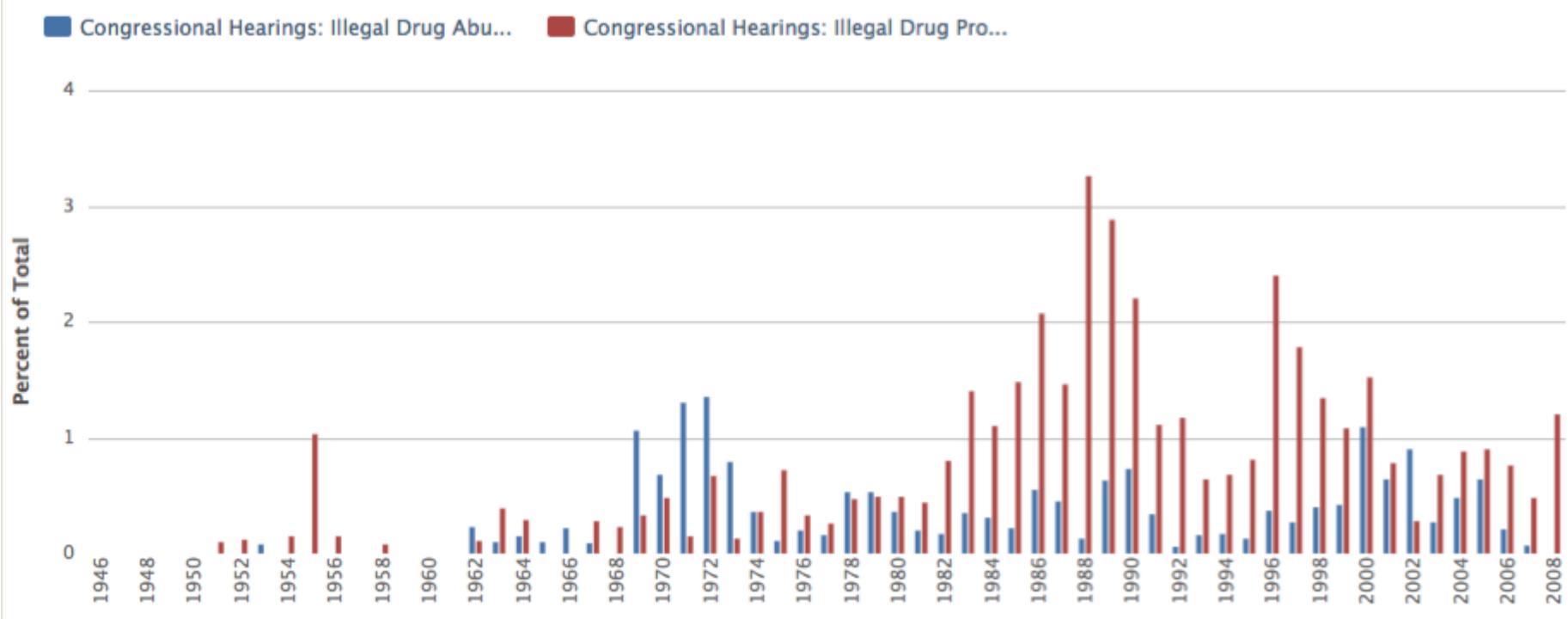
Talking Like the Commission

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

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.

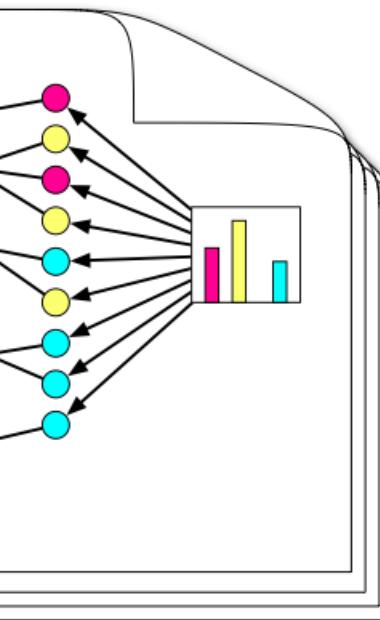
Although the numbers don't match precisely, those predictions

"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 all

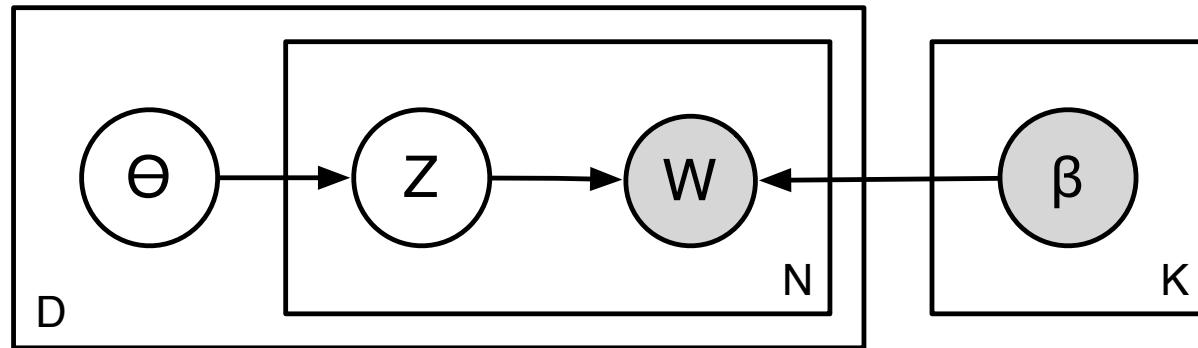
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



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:

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

	<i>state reg</i>	<i>market econ</i>
$P(\text{age} Z)$	a	0
$P(\text{benefit} Z)$	b	0
...
$P(\text{assets} Z)$	0	c
$P(\text{bid} Z)$	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

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

When θ 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

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

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 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	
	<i>state reg</i>	<i>market econ</i>
Conservative	320	643
Labour	396	268

Risk Ratios

Compute two *risk ratios*:

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

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

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
<i>market econ</i>	1.45 [1.26, 1.67]
<i>state reg</i>	0.49 [0.42, 0.57]

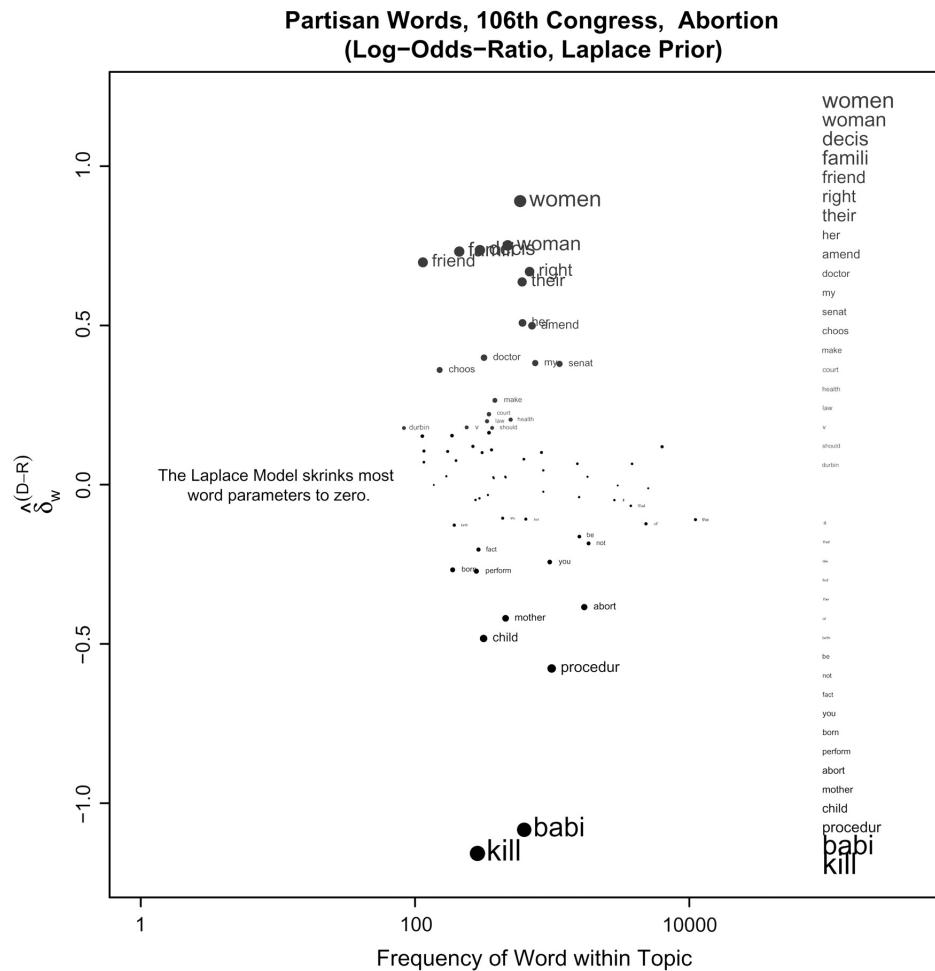
Conservative manifesto generates *market econ* words 45% more often

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

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

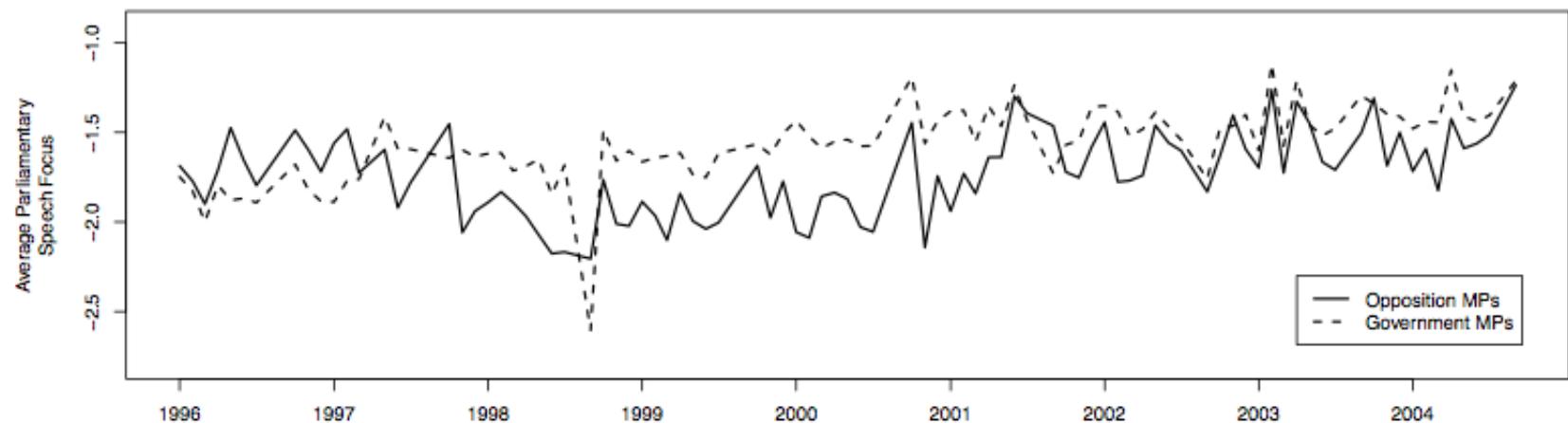
Equivalently Labour generates them about *twice* as often

Extensions: Regularised log odds



... as dependent variable

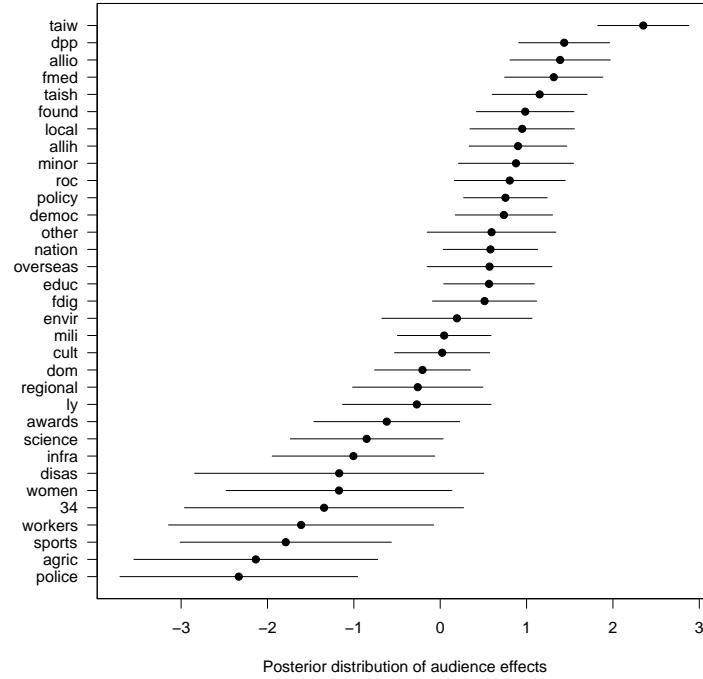
Example: district vs party focus



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 \leftarrow 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:

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

$$Z_{\text{market econ}} - Z_{\text{state reg}}$$

Then

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

Measurement Error

implies the **confusion matrix**

		True	
		<i>state reg</i>	<i>market econ</i>
Observed	<i>state reg</i>	1	0
	<i>market econ</i>	0	1

Measurement Error

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

	<i>state reg</i>	<i>market econ</i>
$P(\text{benefit} Z)$	0.7	0.2
$P(\text{assets} Z)$	0.3	0.8

$P(W=\text{'asset'} | Z=\text{state reg}) > 0$ so

$P(Z=\text{state reg} | W=\text{'asset'}) < 1$

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*)

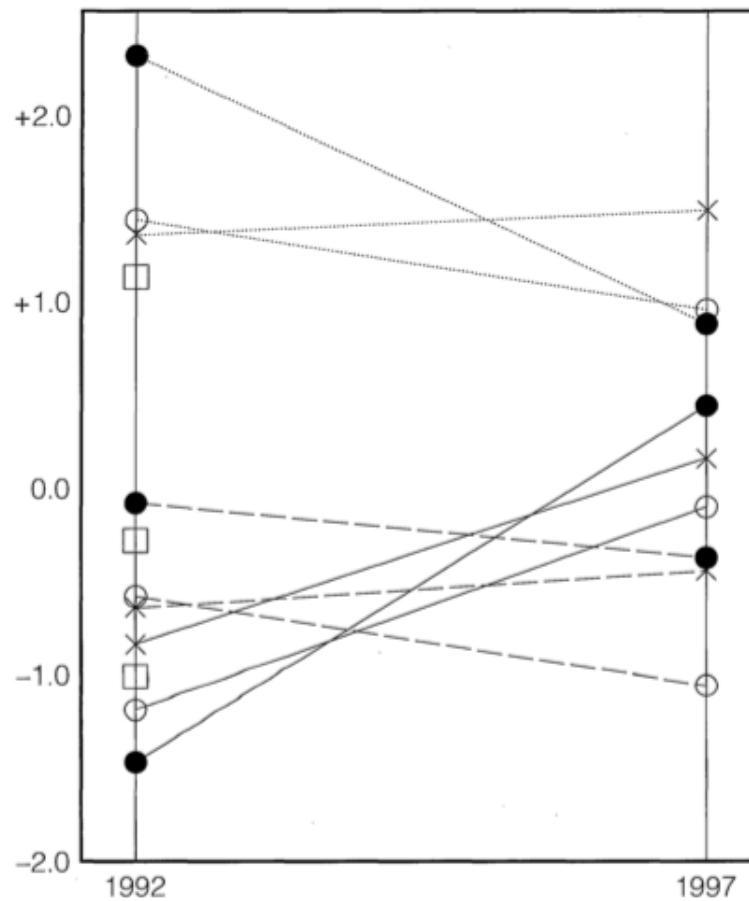
Measurement Error

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)



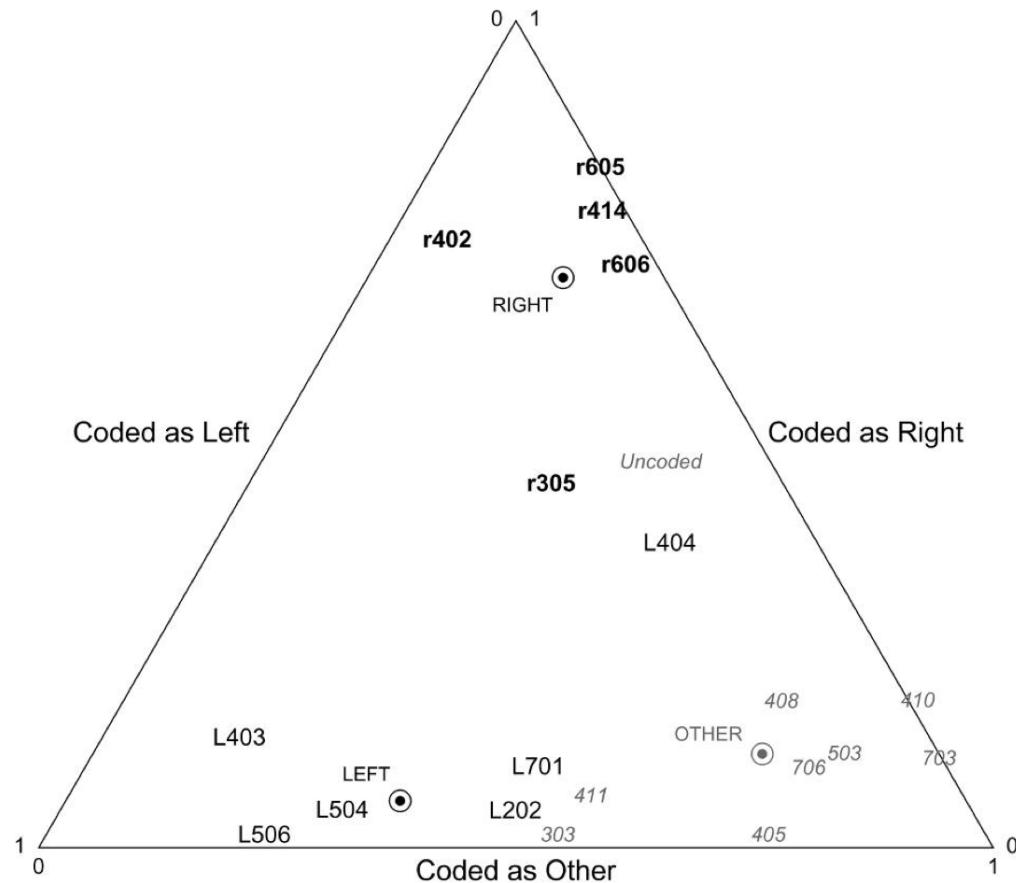
In Action (Mikhaylov et al. 2012)

Table 3 Misclassification matrix for true versus observed Rile

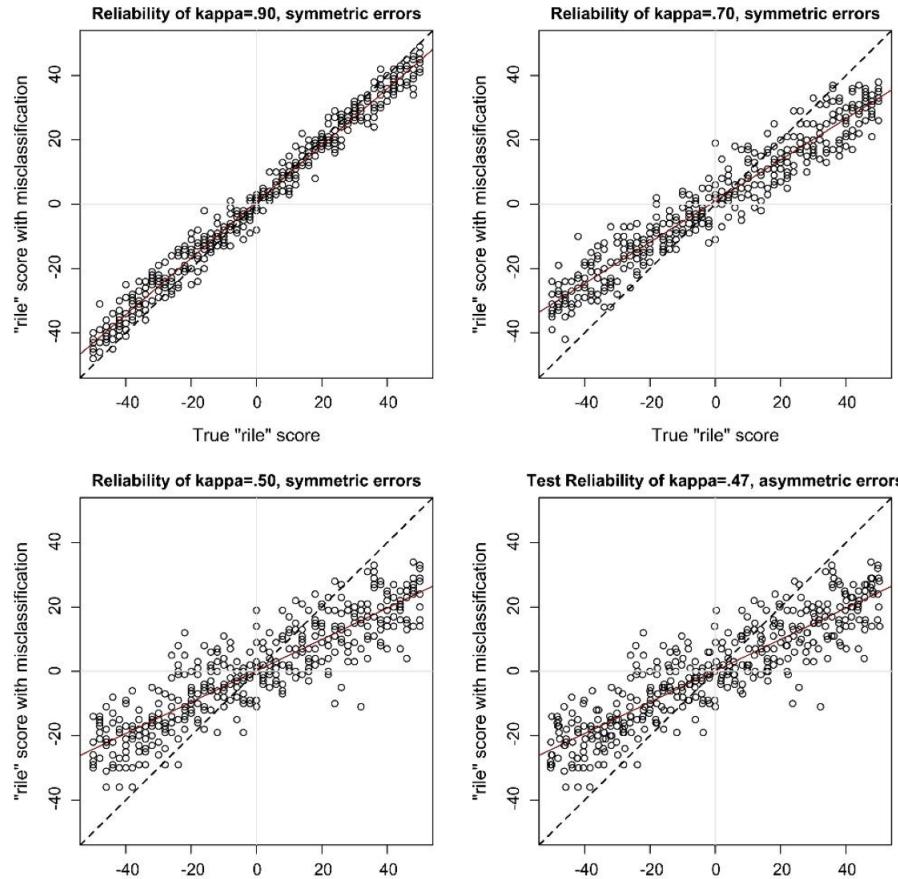
		<i>True Rile category</i>			
		<i>Left</i>	<i>None</i>	<i>Right</i>	<i>Total</i>
Coded Rile	Left	430	188	100	718
		0.59	0.19	0.11	
	None	254	712	193	1159
		0.35	0.70	0.20	
	Right	41	115	650	806
		0.06	0.11	0.69	
	Total	725	1015	943	1668
False negative rate		0.41	0.30	0.31	
False positive rate		0.15	0.27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

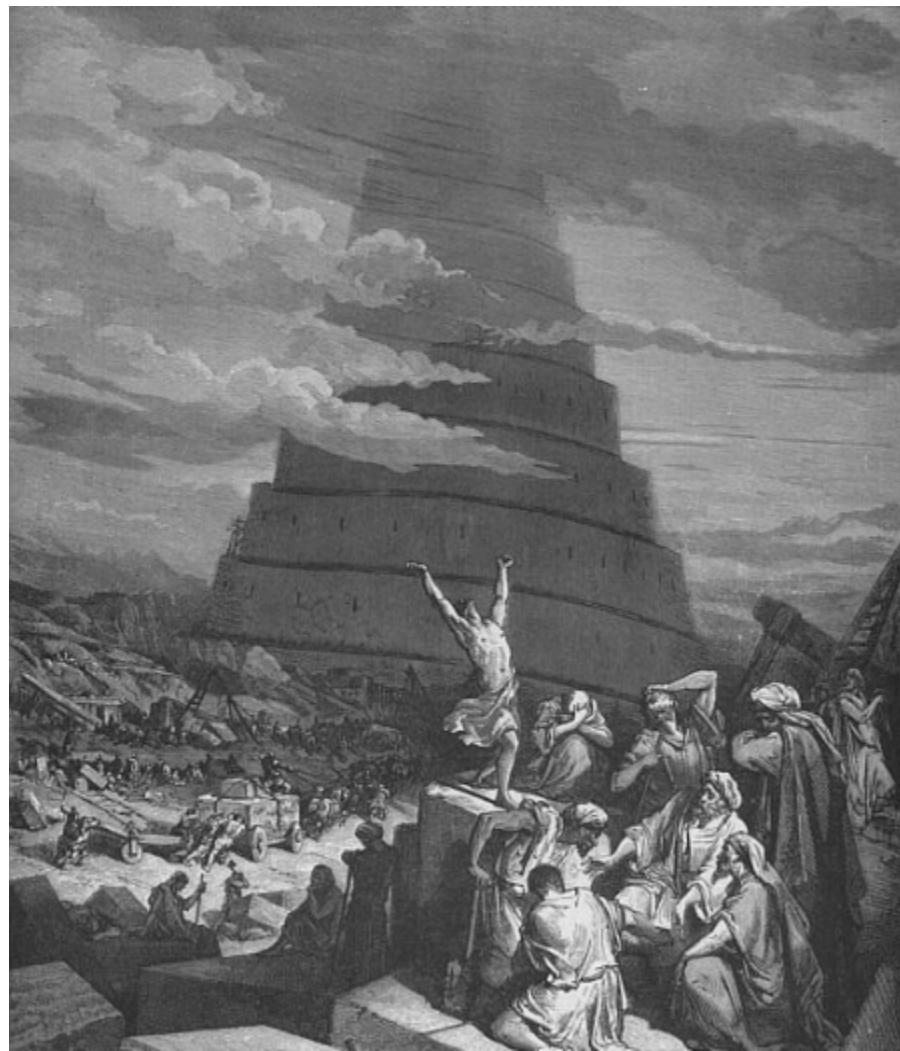
Note. The top figure in each cell is the raw count; the bottom figure is the column proportion. The figures are empirically computed from combined British and New Zealand manifesto tests. The false negative rate is 1—sensitivity, whereas the false positive rate is 1—specificity.

In Action (Mikhaylov et al. 2012)



In Action (Mikhaylov et al. 2012)



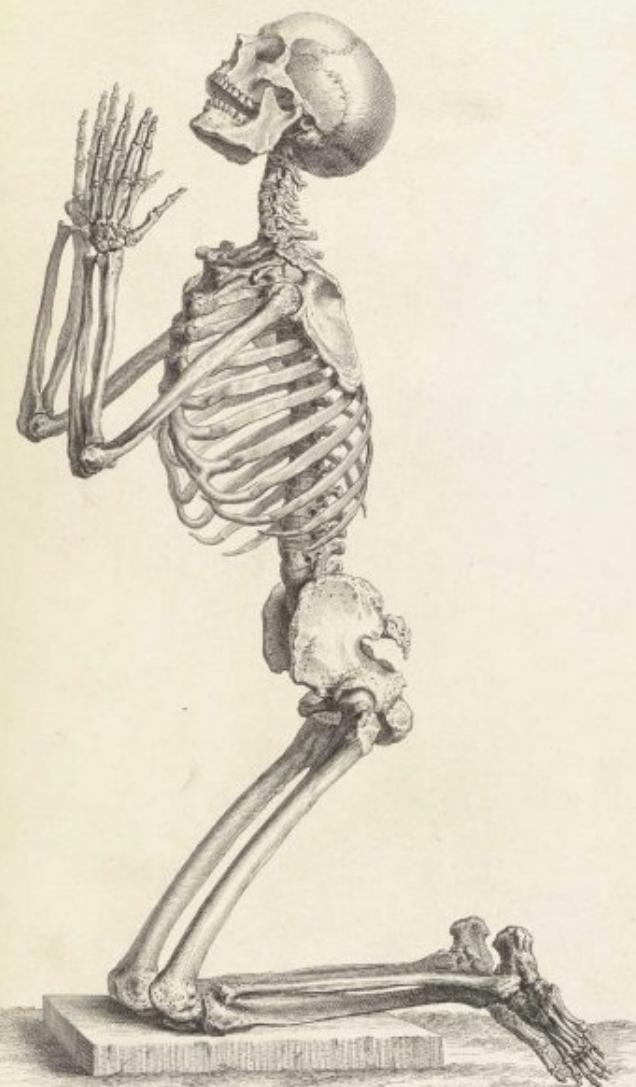


Texas A&M, January 2015

Solutions

Texas A&M, January 2015

XXXVI



Texas A&M, January 2015

Solutions: Try to Avoid it

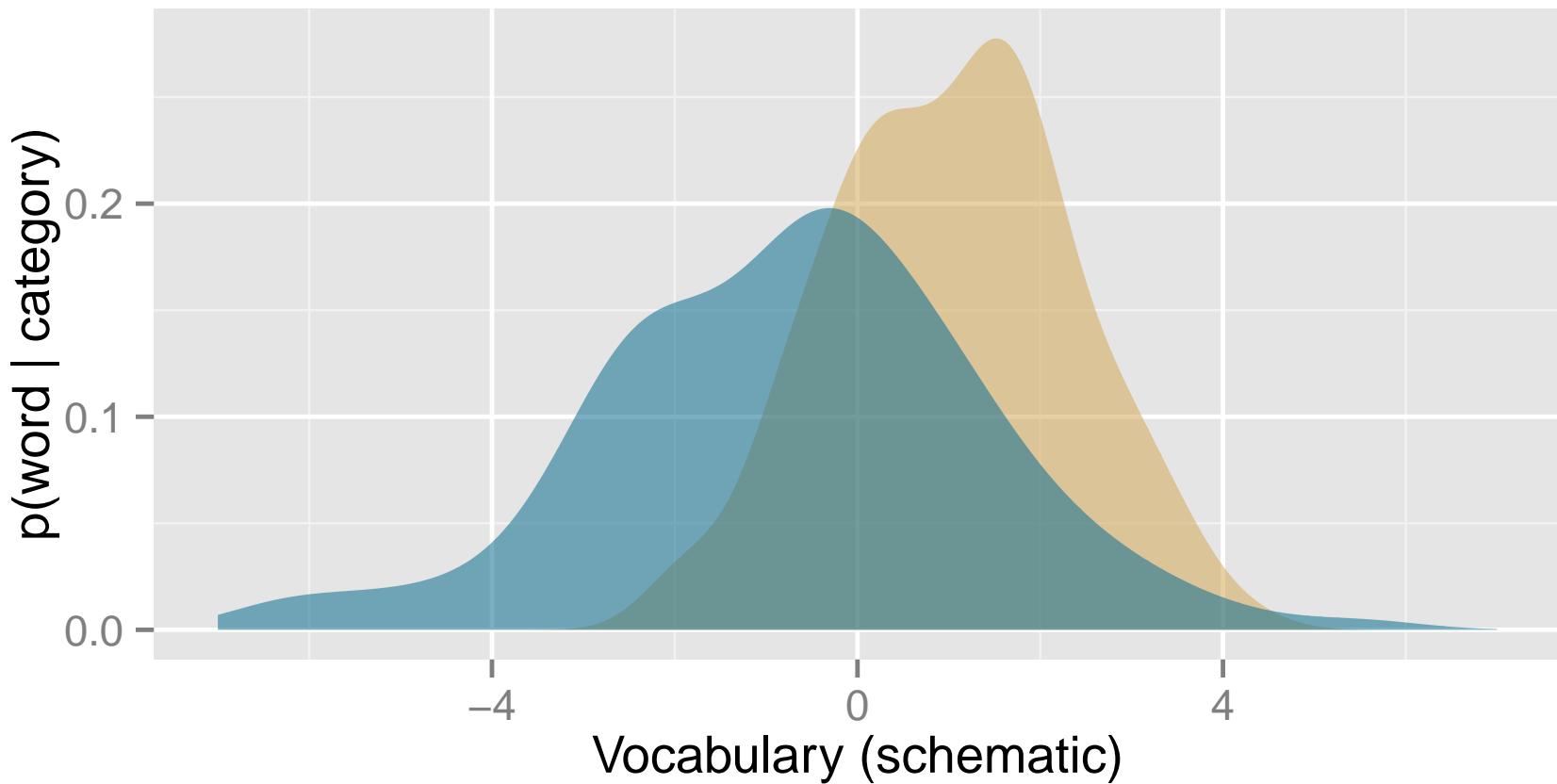
A non-intuitive fact about content dictionaries:

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

The screenshot shows the Yoshikoder application window. The title bar says "Yoshikoder Project: not saved". The menu bar has icons for file operations like Open, Save, Print, and Search. The left sidebar is titled "Dictionary" and contains a tree view of topics: Laver and Garry, Culture (selected), High, Popular, Sport, angler*, Economy, Environment, and Groups. The main content area is titled "CONS1992.txt" and displays two sections: "THE CONSERVATIVE PARTY MANIFESTO" and "THE BEST FUTURE FOR BRITAIN". The right sidebar is titled "Documents" and lists several files: CONS1992.txt, CONS1997.txt, CONS2001.txt, CONS2005.txt, CONS2010.txt, LABOUR1992.txt (selected), LABOUR1997.txt, LABOUR2001.txt, LABOUR2005.txt, LABOUR2010.txt, and LIBDEM1992.txt.

HEMATICS SCIENCE HISTORY GEOGRAPHY TECHNOLOGY ROVIDE FUNDS FOR A NUMBER OF GOOD CAUSES IN THE BY THE YEAR 2000 THE ARTS BRITAIN HAS A GREAT EARS WITH GROWING ATTENDANCE AT THEATRE OPERA S CULTURAL LIFE OF THE CAPITAL WHICH HAS SEEN NEW APITAL WHICH HAS SEEN NEW GALLERIES THEATRES AND USEFUL AT SEVERAL TIMES IN OUR HISTORY THE BRITISH COUNTRY OUR AIM IS TO MAKE THE PERFORMING ARTS ATICS SCIENCE HISTORY GEOGRAPHY TECHNOLOGY ART	ART ARTISTIC ARTISTIC DANCE GALLERIES MUSEUMS MUSEUM MUSEUMS MUSIC MUSIC OPERATION OPERATORS OPERATIONS OPERATED OPERATE OPERATIONAL OPERATIONAL OPERATIONS OPERATION OPERATIONS OPERATION	MUSIC PE AND IN SECONDARY SCHOOLS A FOREIGN LANGUAGE SPORTING HERITAGE AND CHARITABLE FIELDS AND FROM WHIC HERITAGE AND A LIVELY CONTEMPORARY ARTS SCENE THE ART AND ARTS FESTIVALS WE HAVE SUPPORTED THIS BY INCREASINC THEATRES AND MUSEUMS OPENING OVER THE LAST DECADE WI OPENING OVER THE LAST DECADE WE WILL LAUNCH A LONDON WAS FOUNDED OUT OF THE PROCEEDS OF SUCH A LOTTERY AND OUR HERITAGE ACCESSIBLE TO ALL WE WILL ENCOURAGE T PE AND IN SECONDARY SCHOOLS A FOREIGN LANGUAGE REGUL DAY IN JUNE IN THIS YEAR'S BUDGET WE ANNOUNCED FURTHER TO PROTECT THE KURDS BRITAIN HAS LED THE WORLD IN CAN CARRY FREELY WITHIN THE EC INCREASE COMPETITION IN WILL BE STEADILY IMPROVED ON COUNCIL ESTATES EVERY TENA WITHOUT A SUBSIDY LAST YEAR THE FIRST-CLASS LETTER SERVI WE WILL GIVE EVERY MEMBER OF AN OCCUPATIONAL SCHEME TI CONTROL TO LOCAL UNITS AND STREAMLINING ITS CHAIN OF C DUTIES WE WILL ENCOURAGE THE EXTENSION OF NEIGHBOURHO THE HOME SECRETARY WILL REMAIN MINISTERIALLY ACCOUNTA AND WE WILL MONITOR ITS WORK CAREFULLY TO SEE IF WHERE WAITING CAUSES MOST DISTRESS TO ENSURE THAT PRO WE ARE SURE THAT AS NOW MANY HOSPITALS WILL BE
---	--	---

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)

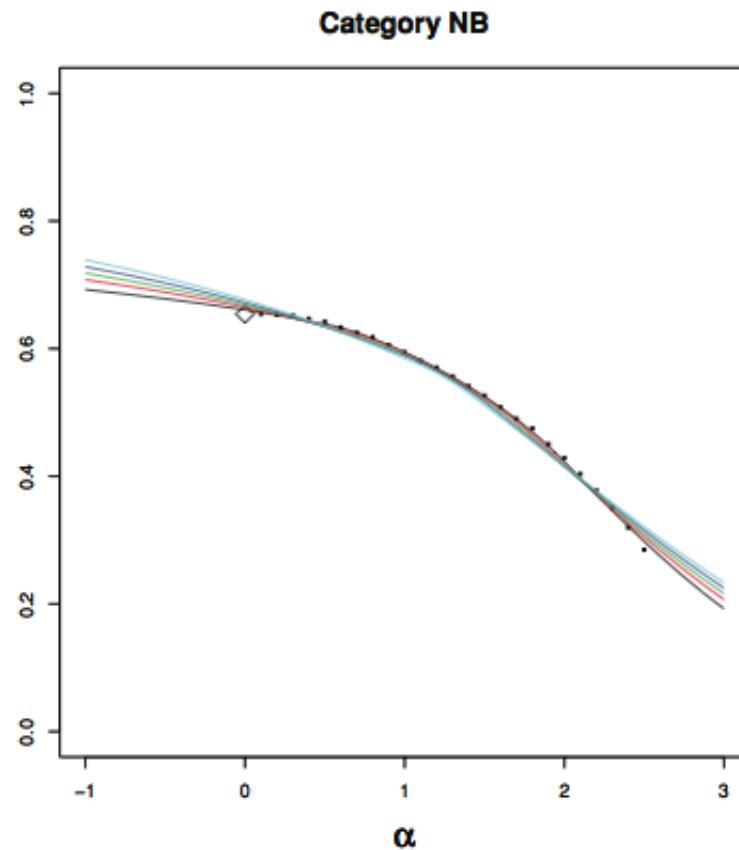
Coder training provides information about how a category proportion estimate changes as coders get better

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

$$P(W) = \sum_k^K P(W | Z = k)P(Z = k)$$

has the form

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

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 $P(C | T)$ 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)

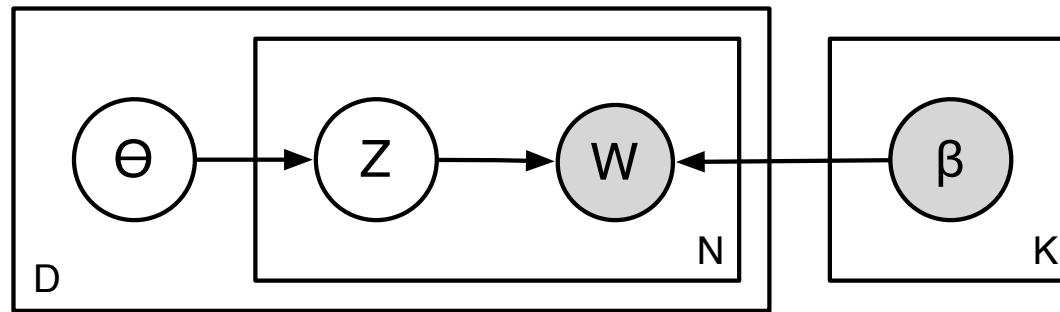
Topic models, e.g. Latent Dirichlet Allocation (Blei et al.) **add**:

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

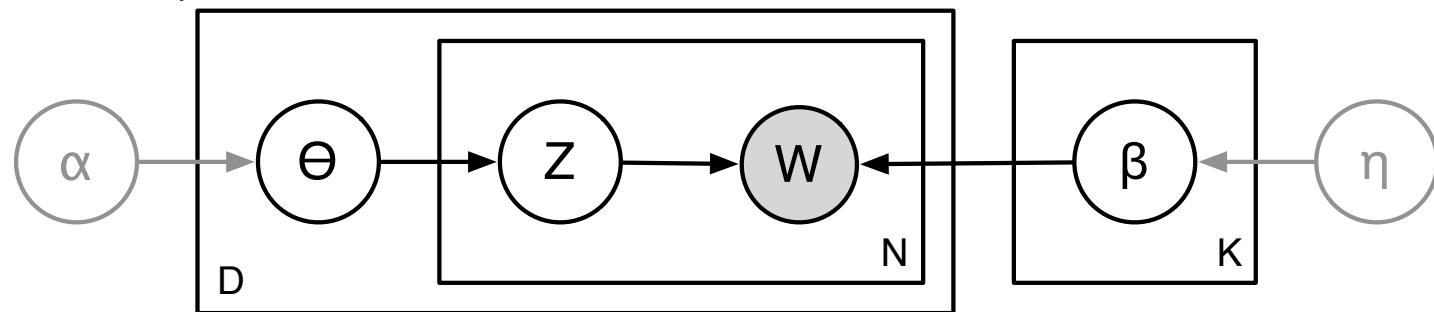
Full statistical framework for learning most aspects of the relationship

Topic Models as Classical Content Analysis

From



(via pLSI) to



Generative probability model

Assumptions:

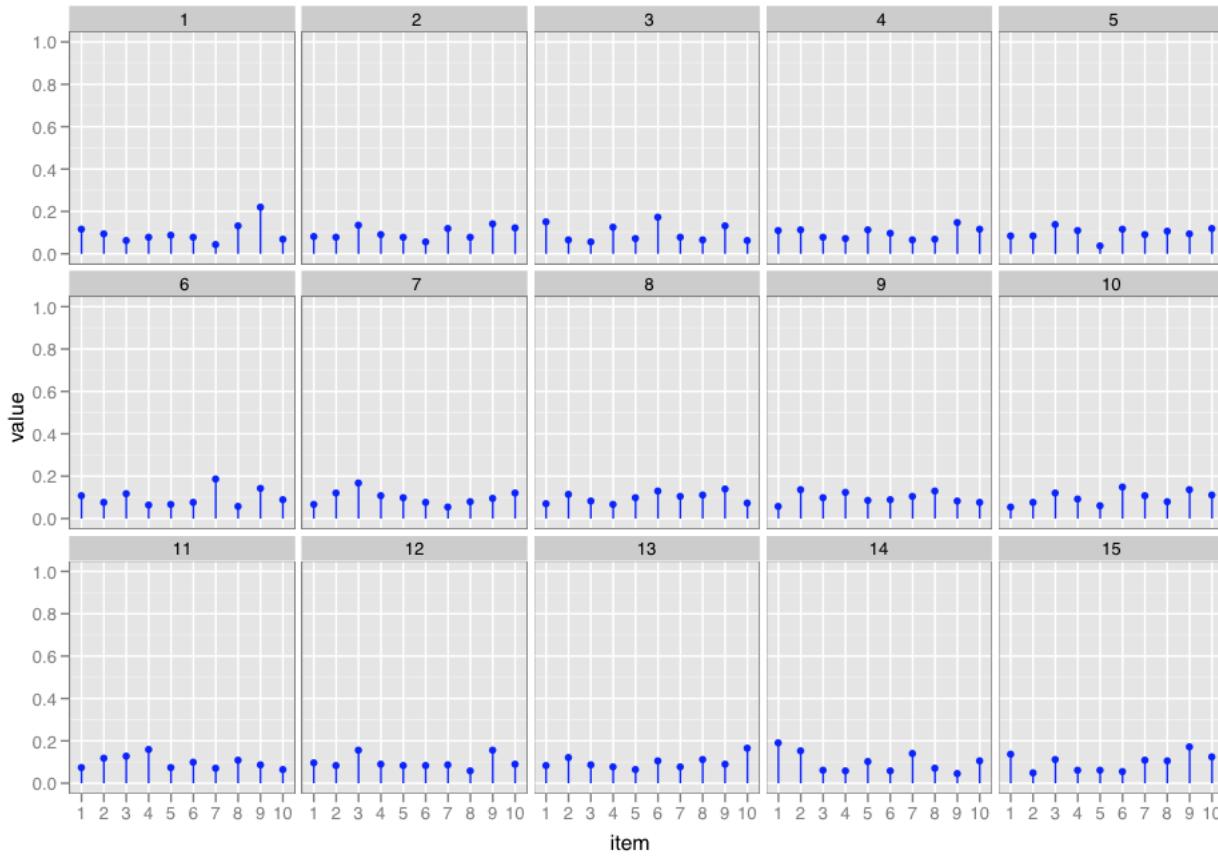
W and Z are multinomial.

θ and β are Dirichlet

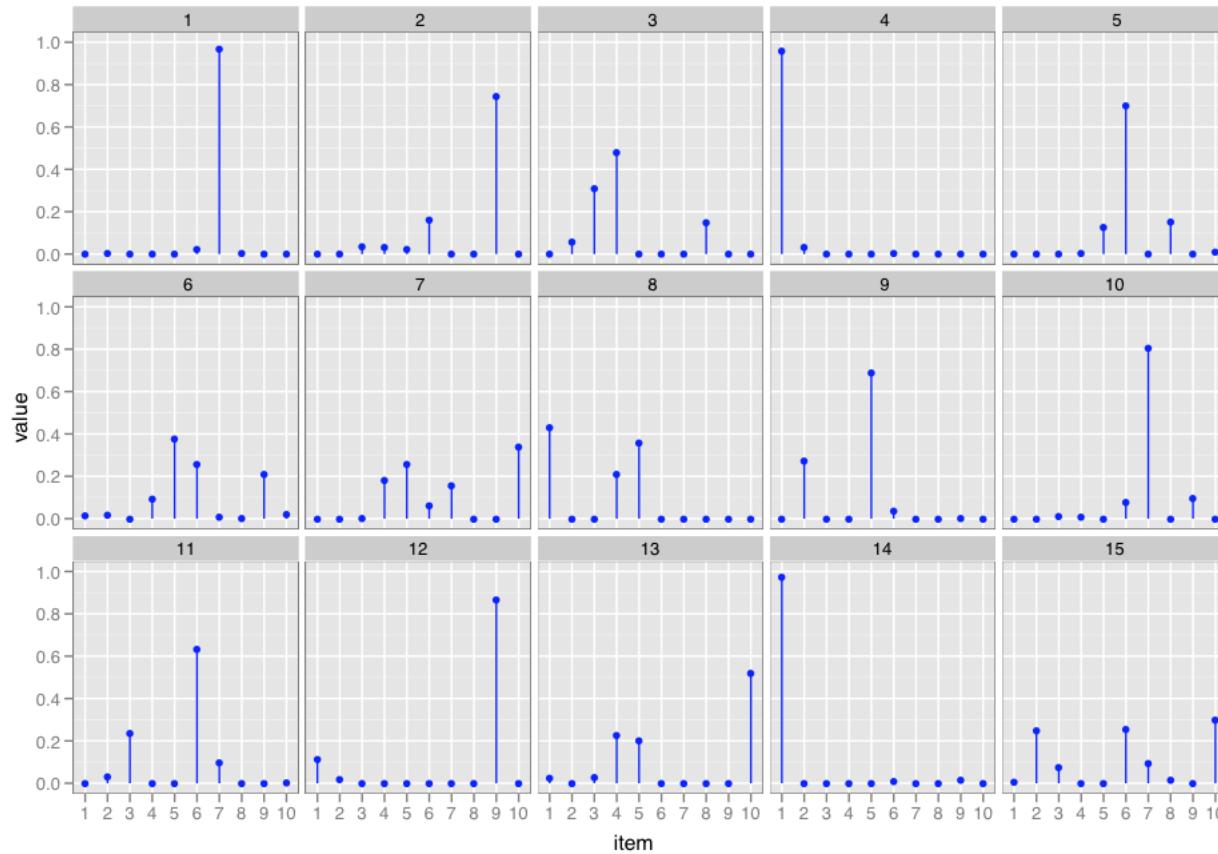
α and η are hyperparameters for Dirichlet parameters

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

$$\alpha = 10$$



$$\alpha = .1$$



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.

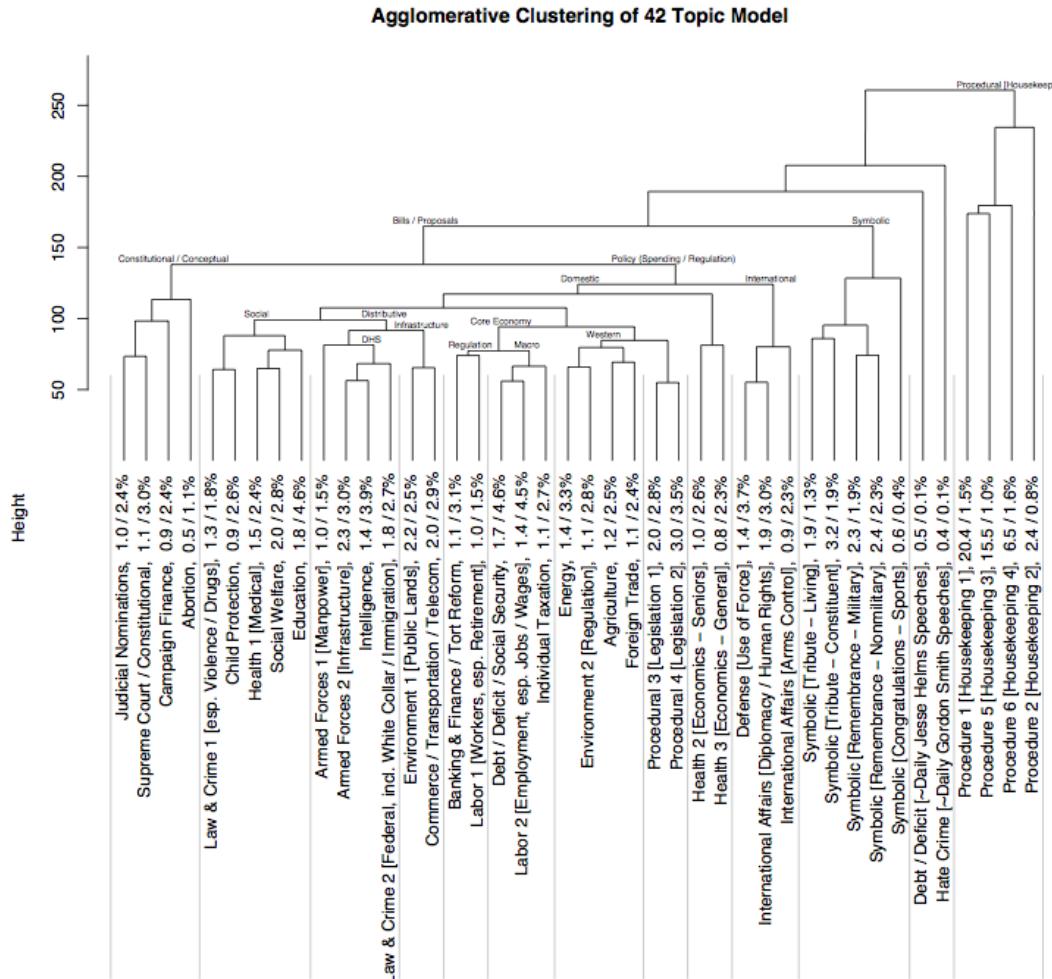
Topic *number* is now an open question

What is this topic anyway?

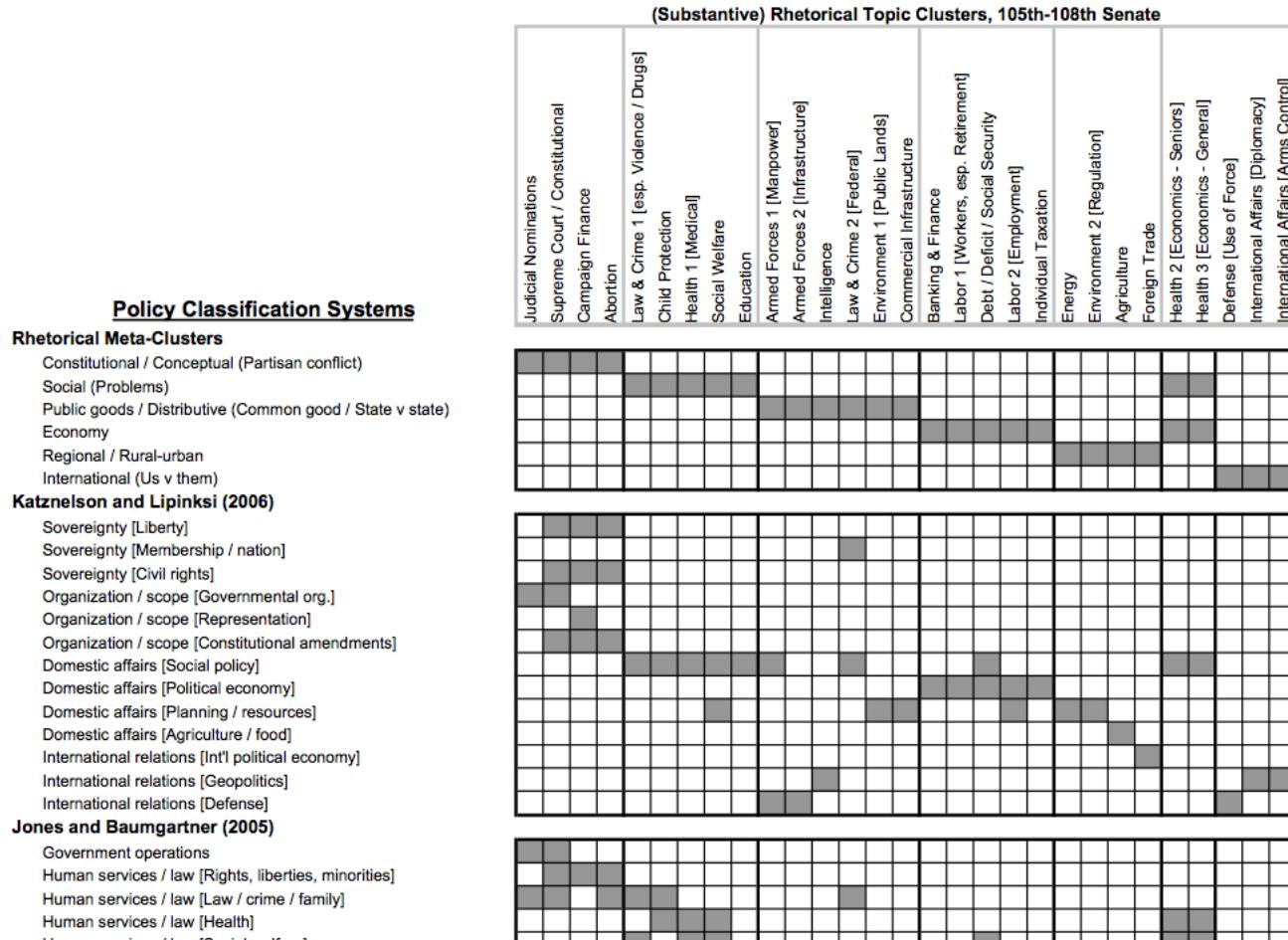
Topic (Short Label)	Keys
1. Judicial Nominations	<i>nomine, confirm, nomin, circuit, hear, court, judg, judici, case, vacanc</i>
2. Constitutional	<i>case, court, attornei, supreme, justic, nomin, judg, m, decis, constitut</i>
3. Campaign Finance	<i>campaign, candid, elect, monei, contribut, polit, soft,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

(Quinn et al. 2010)

How is it related to others topics?



and other schemes?

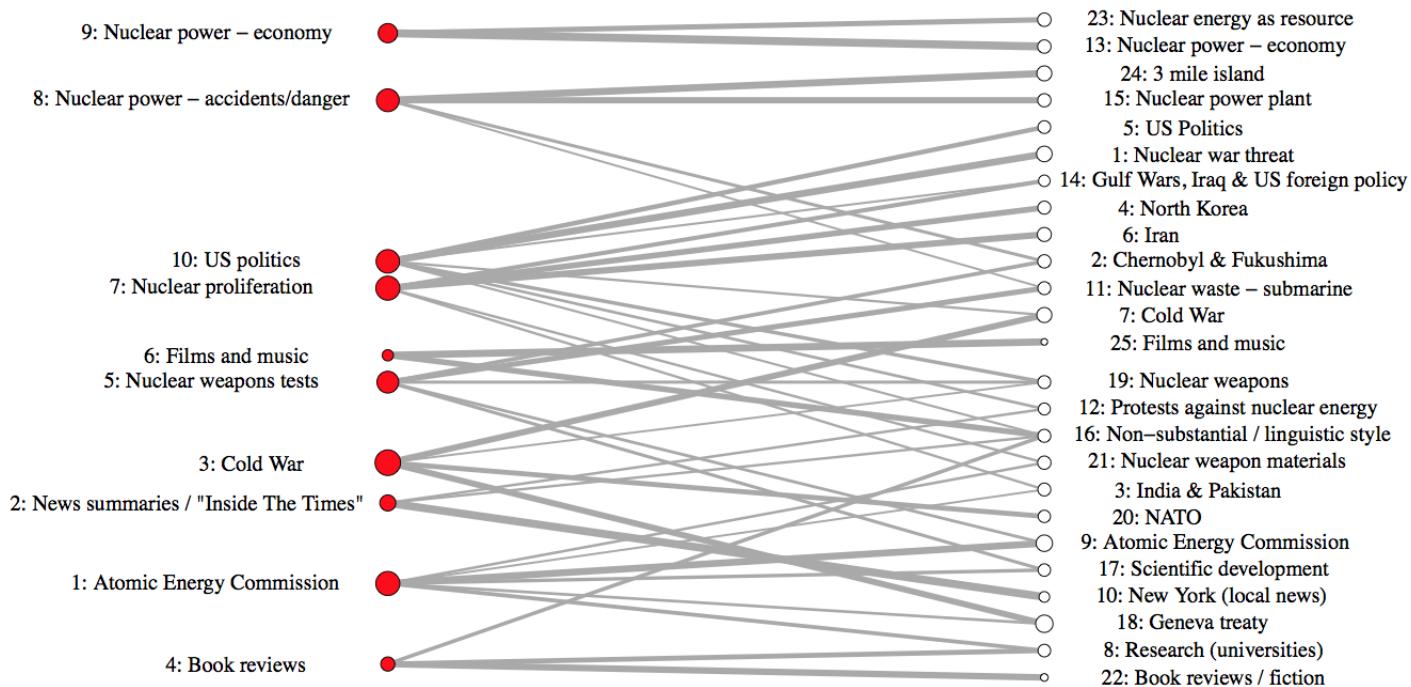


Topic Models: How *many* topics?

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 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:

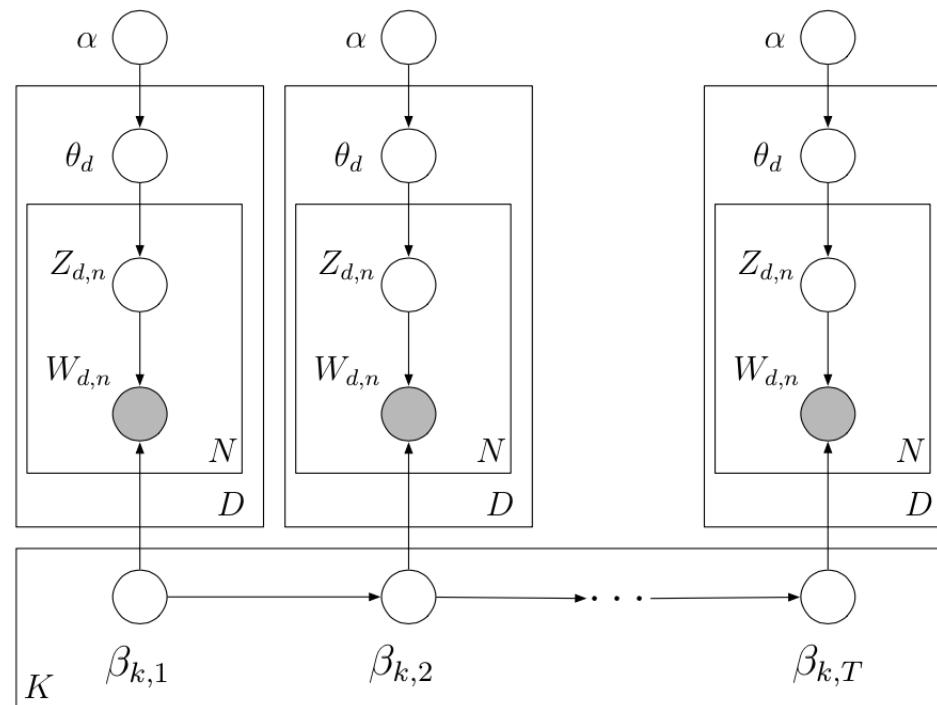
Dynamic topic models

Structural topic models

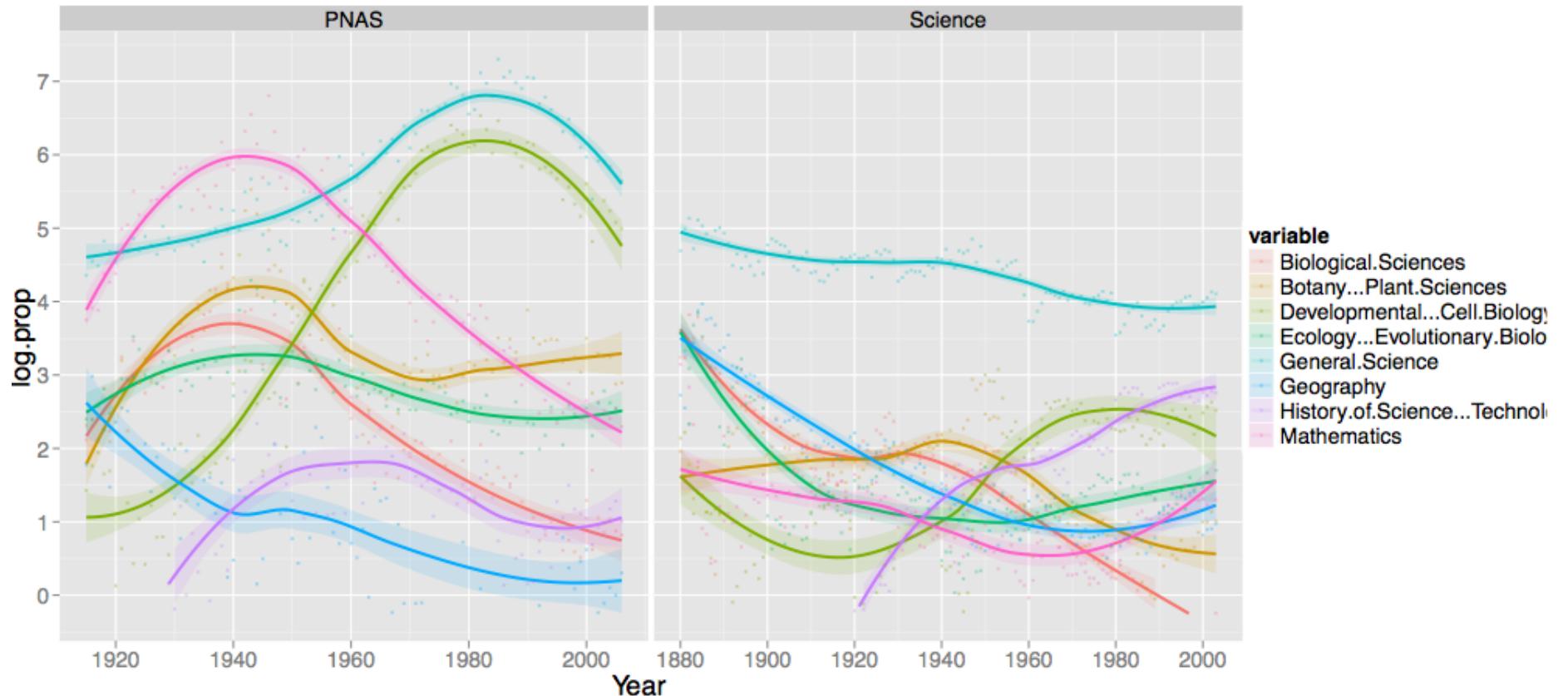
(<http://structuraltopicmodel.com>)

Dynamic Topic Models (1)

Words ‘drift’ from topic to topic over time



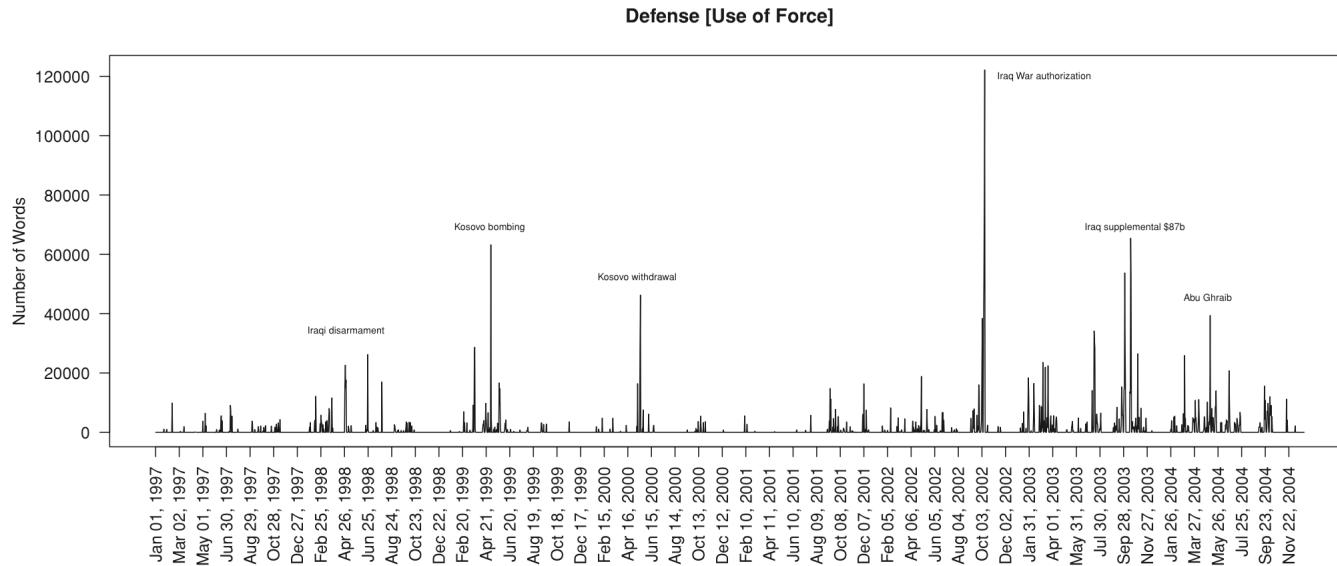
Dynamic Topic Models (1)



Dynamic Topic Models (2)

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

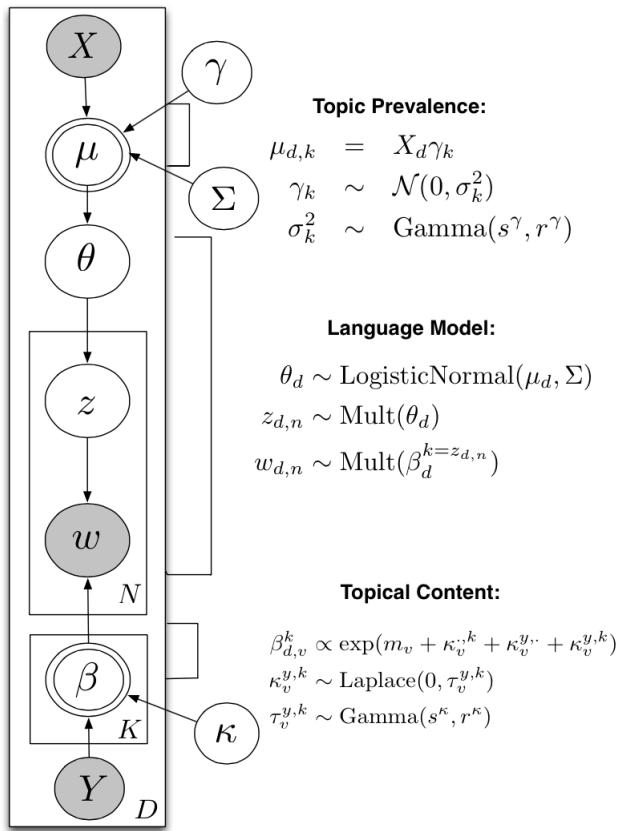
To track a smoothly moving policy agenda
(congressional speeches from 1997-2004)



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

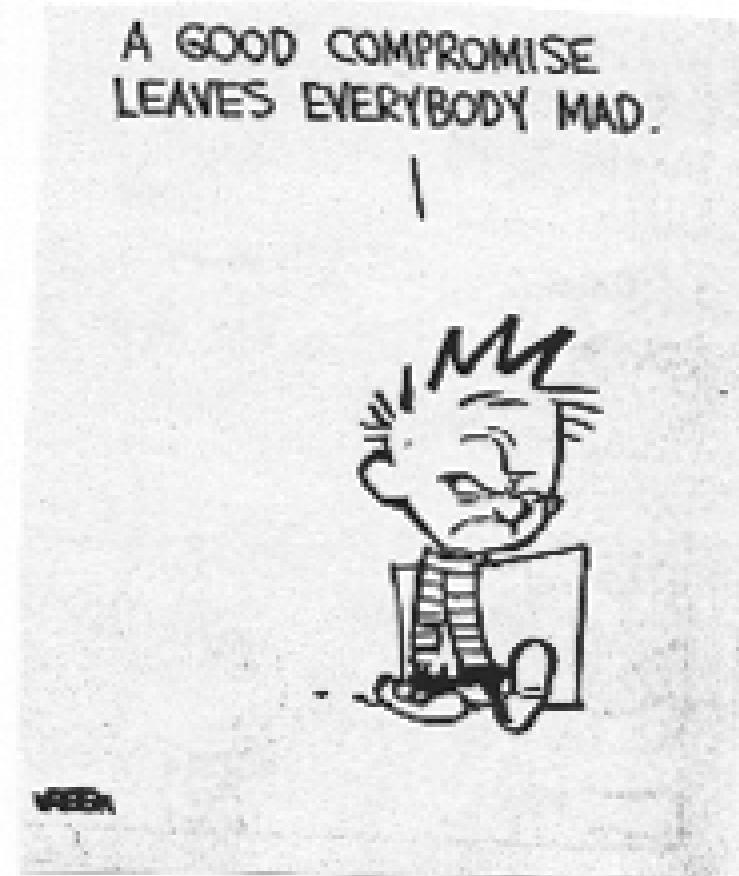
STM ‘Structural Topic Model’: Topic proportion correlations are regressions on covariates

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

Structural Topic Models



Wrapping up



Texas A&M, January 2015

Texas A&M, January 2015

Lab Time



Texas A&M, January 2015