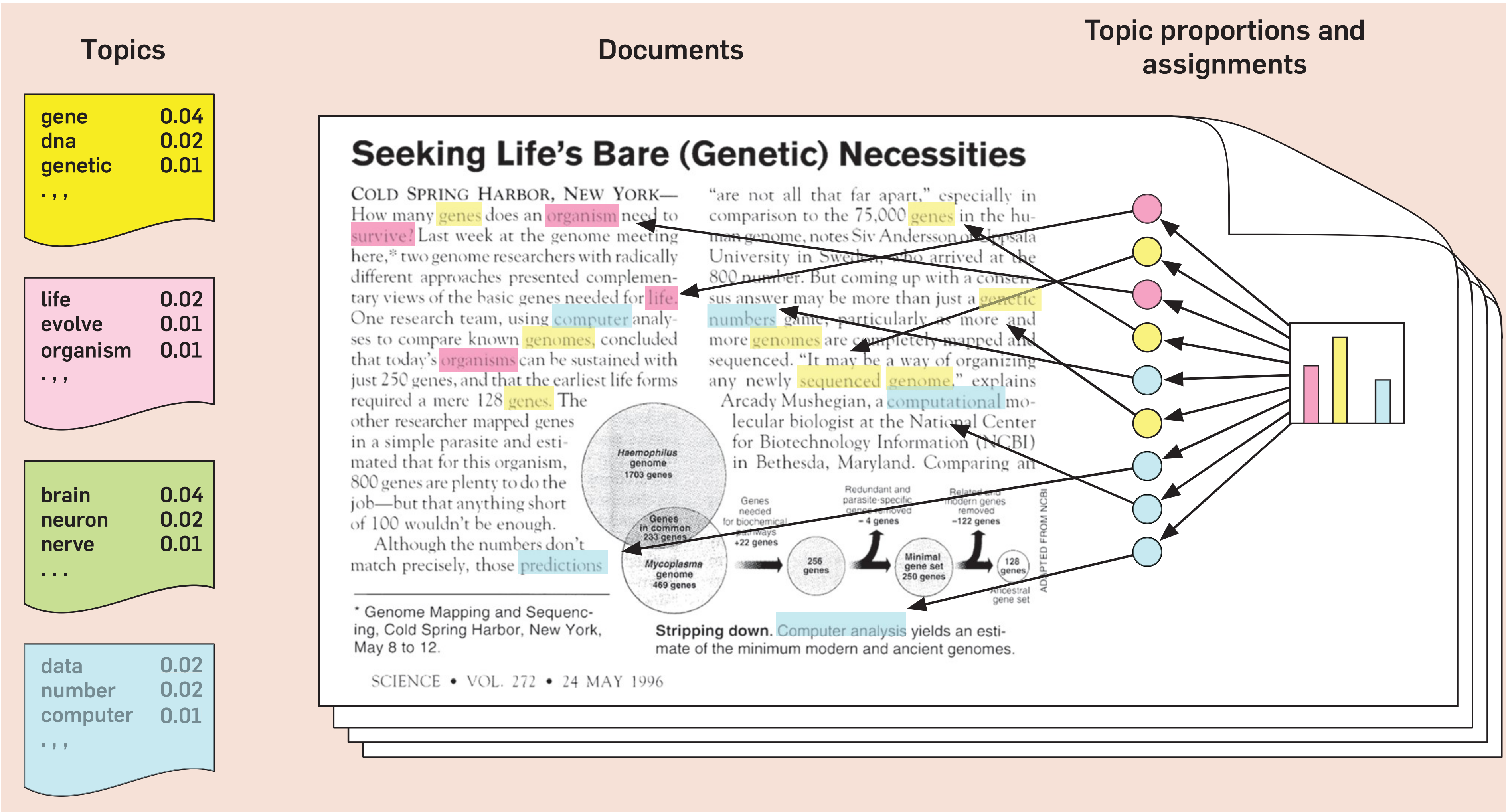
Last time. . .

Last time. . .



Topics
Documents
Topic proportions and assignments
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
. , ,
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
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

Extensions and Special Cases

Extensions and Special Cases

'vanilla' LDA is a popular model.

Extensions and Special Cases

‘vanilla’ LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts. . .

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts...

Assn each document is mix of multiple topics.

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts...

Assn each document is mix of multiple topics.

But each document is one topic. [EAM]

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts. . .

Assn each document is mix of multiple topics.

But each document is one topic. [EAM]

Assn topics in documents are uncorrelated.

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts. . .

Assn each document is mix of multiple topics.

But each document is one topic. [EAM]

Assn topics in documents are uncorrelated.

But given topic A, we are more likely to see topic B than C. [CTM]

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts...

Assn each document is mix of multiple topics.

But each document is one topic. [EAM]

Assn topics in documents are uncorrelated.

But given topic A, we are more likely to see topic B than C. [CTM]

Assn documents are exchangeable (over time).

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts...

Assn each document is mix of multiple topics.
But each document is one topic. [EAM]

Assn topics in documents are uncorrelated.
But given topic A, we are more likely to see topic B than C. [CTM]

Assn documents are exchangeable (over time).
But time matters for what topic 'means' [DTM]

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts. . .

Assn each document is mix of multiple topics.
But each document is one topic. [EAM]

Assn topics in documents are uncorrelated.
But given topic A, we are more likely to see topic B than C. [CTM]

Assn documents are exchangeable (over time).
But time matters for what topic 'means' [DTM]

Assn no covariates

Extensions and Special Cases

'vanilla' LDA is a popular model. But assumptions it makes are restrictive/inappropriate for some contexts...

Assn each document is mix of multiple topics.
But each document is one topic. [EAM]

Assn topics in documents are uncorrelated.
But given topic A, we are more likely to see topic B than C. [CTM]

Assn documents are exchangeable (over time).
But time matters for what topic 'means' [DTM]

Assn no covariates
But topic prevalence and topic content are $f(X)$ [STM]

Lots of other ideas!

Lots of other ideas!

hierarchical LDA, pachinko allocation, nonparametric pachinko allocation,factorial LDA, gamma-poisson factorization, shared component topic models, dirichlet multinomial regression topic models, nested hierarchical dirichlet process topic model, focused topic model, inverse regression topic model, ideal point topic model, discrete innite logistic normal topic model multilingual topic model, markov topic model, relational topic model, syntactic topic model, supervised latent dirichlet allocation

Expressed Agenda Model (Grimmer, 2010)

Expressed Agenda Model (Grimmer, 2010)

Suppose each document is assigned to one topic.

Expressed Agenda Model (Grimmer, 2010)

Suppose each document is assigned to one topic.

Each author allocates a latent (to us) proportion of time to each topic.

Expressed Agenda Model (Grimmer, 2010)

Suppose each document is assigned to one topic.

Each author allocates a latent (to us) proportion of time to each topic.

$\rightarrow$ special case of LDA, and used for measuring way Senators 'express' themselves to constituents via press releases.

Notice that set of topics is same across Senators,

Expressed Agenda Model (Grimmer, 2010)

Suppose each document is assigned to one topic.

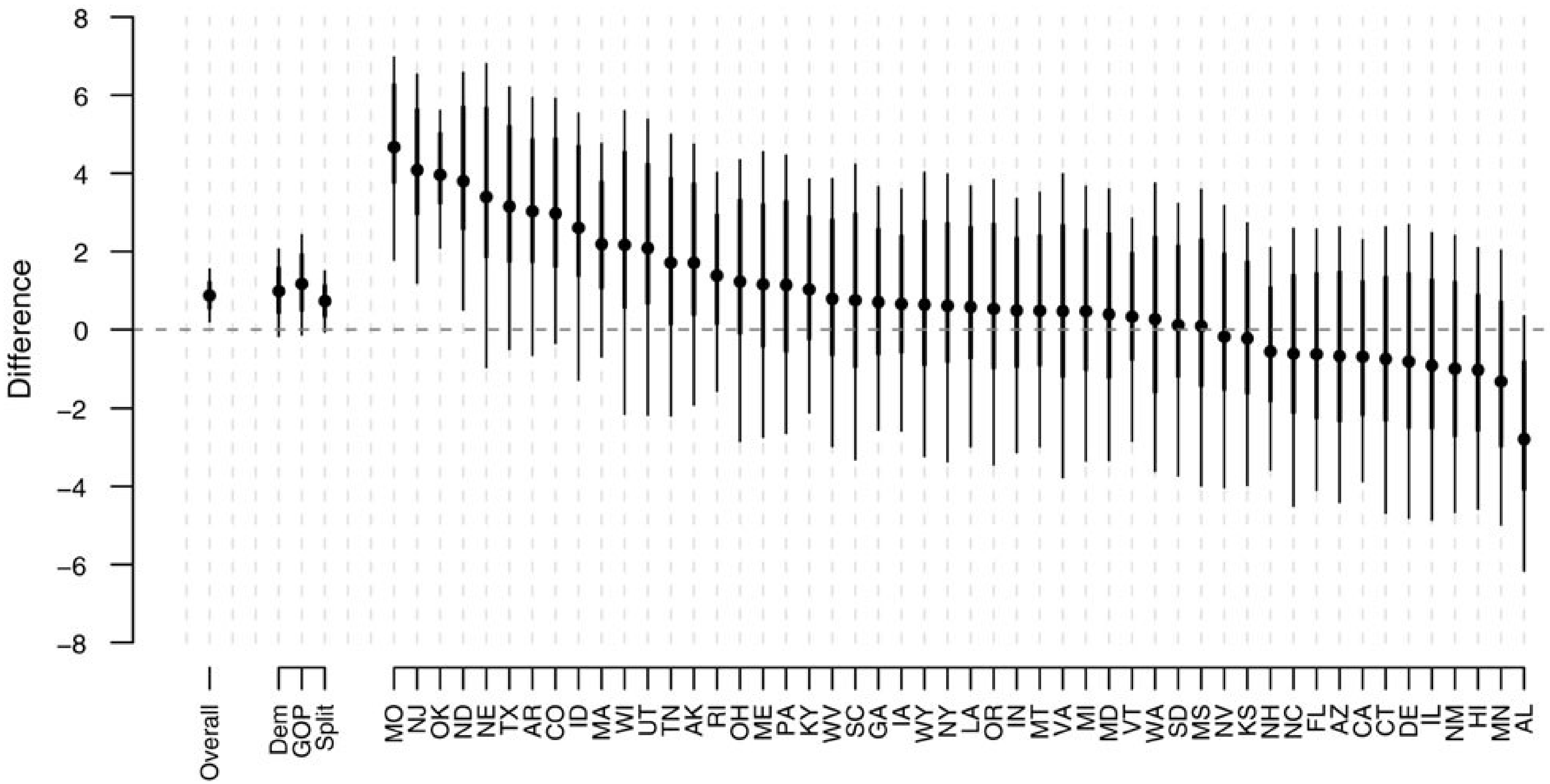
Each author allocates a latent (to us) proportion of time to each topic.

→ special case of LDA, and used for measuring way Senators 'express' themselves to constituents via press releases.

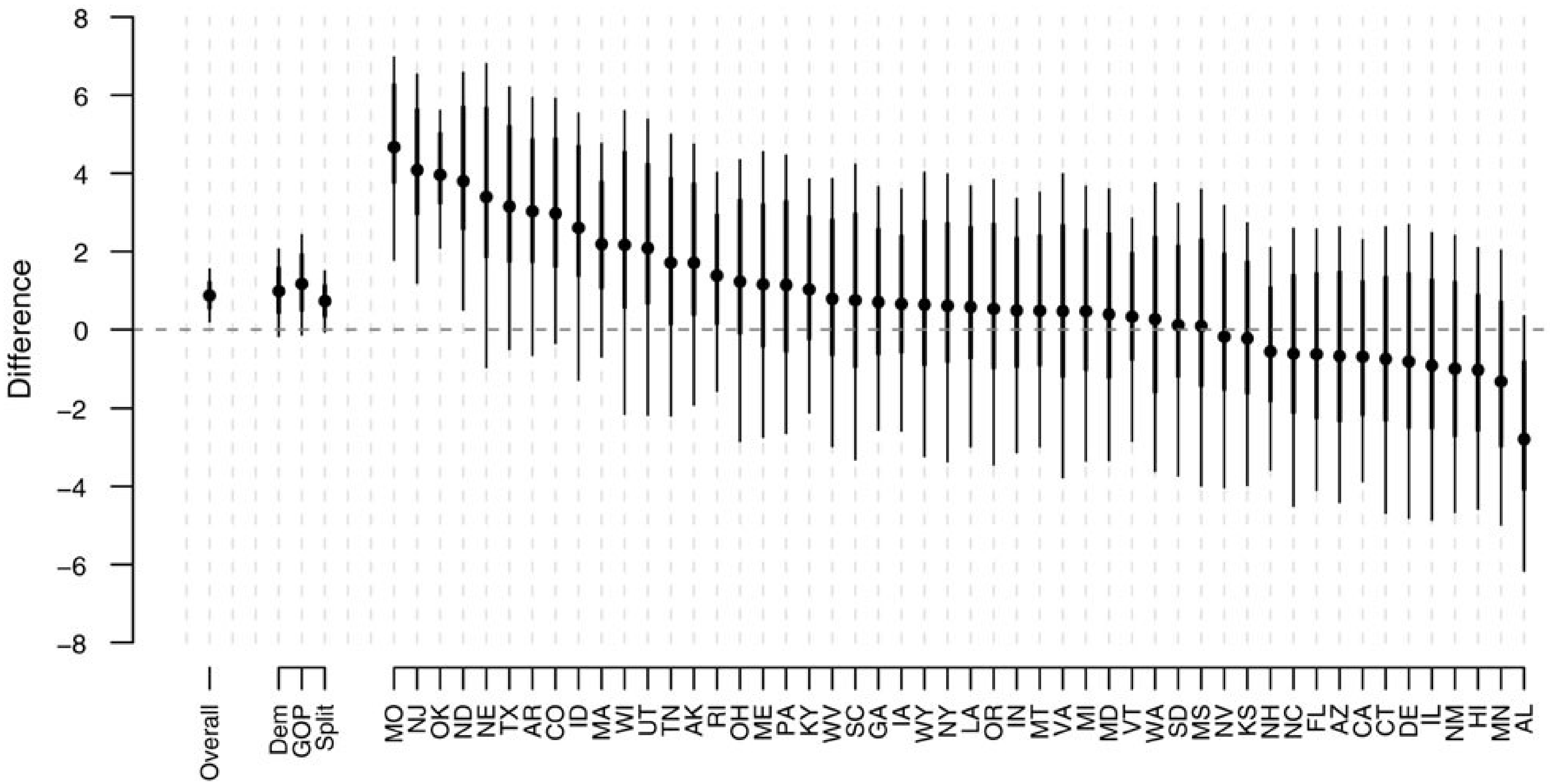
Notice that set of topics is same across Senators, but weights are allowed to vary across Senators.

Senators from same states have similar agendas

Senators from same states have similar agendas



Senators from same states have similar agendas



Senators from same states talk about more similar things than Senators from different states (generally).

Correlated Topic Model (Blei & Lafferty, 2007)

Correlated Topic Model (Blei & Lafferty, 2007)

If a news article talks about `finance` it's more likely to also talk about `law` than it is `baseball`.

Correlated Topic Model (Blei & Lafferty, 2007)

If a news article talks about `finance` it's more likely to also talk about `law` than it is `baseball`. But LDA doesn't allow this (consequence of Dirichlet prior on topic proportions).

Correlated Topic Model (Blei & Lafferty, 2007)

If a news article talks about `finance` it's more likely to also talk about `law` than it is `baseball`. But LDA doesn't allow this (consequence of Dirichlet prior on topic proportions).

The Correlated Topic Model allows for positive covariance between topics.

Correlated Topic Model (Blei & Lafferty, 2007)

If a news article talks about `finance` it's more likely to also talk about `law` than it is `baseball`. But LDA doesn't allow this (consequence of Dirichlet prior on topic proportions).

The Correlated Topic Model allows for positive covariance between topics. Does this by drawing topic proportions from a log normal.

Correlated Topic Model (Blei & Lafferty, 2007)

If a news article talks about `finance` it's more likely to also talk about `law` than it is `baseball`. But LDA doesn't allow this (consequence of Dirichlet prior on topic proportions).

The Correlated Topic Model allows for positive covariance between topics. Does this by drawing topic proportions from a log normal.

Shows improved model fit over LDA.

Correlated Topic Model (Blei & Lafferty, 2007)

If a news article talks about `finance` it's more likely to also talk about `law` than it is `baseball`. But LDA doesn't allow this (consequence of Dirichlet prior on topic proportions).

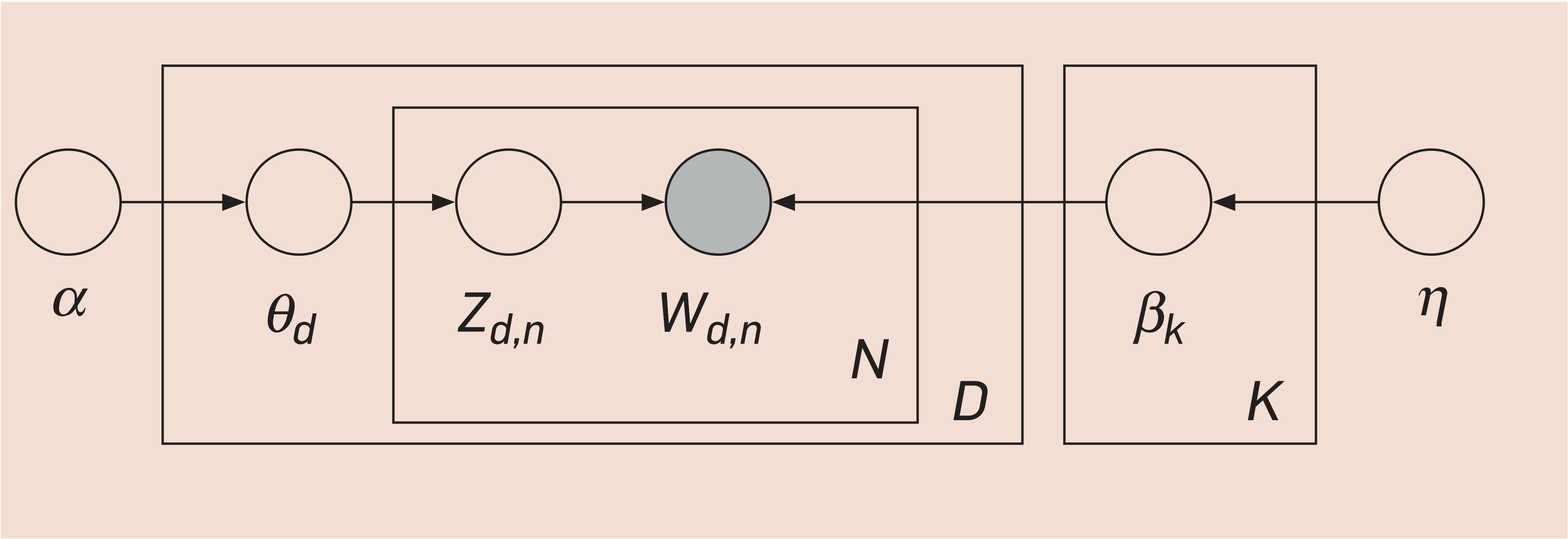
The Correlated Topic Model allows for positive covariance between topics. Does this by drawing topic proportions from a log normal.

Shows improved model fit over LDA. BTW, note that STM (below) reduces to CTM if no covariates are specified.

Dynamic Topic Model (Blei & Lafferty, 2006)

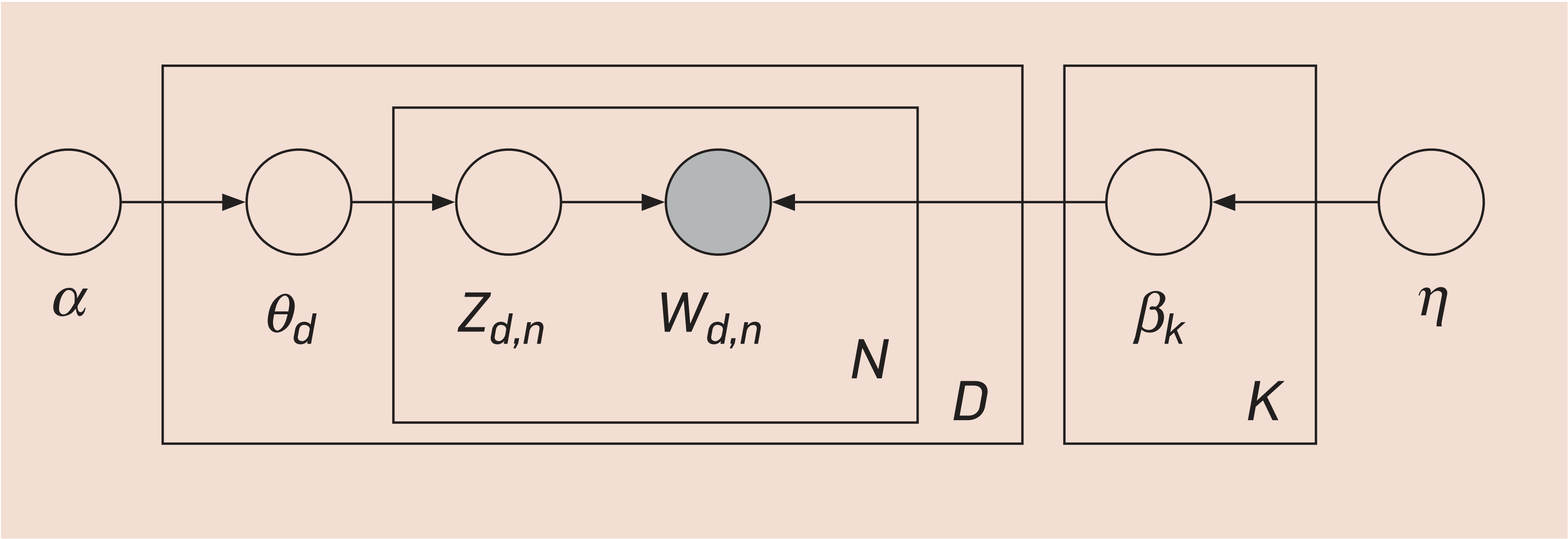
Dynamic Topic Model (Blei & Lafferty, 2006)

Recall LDA...



Dynamic Topic Model (Blei & Lafferty, 2006)

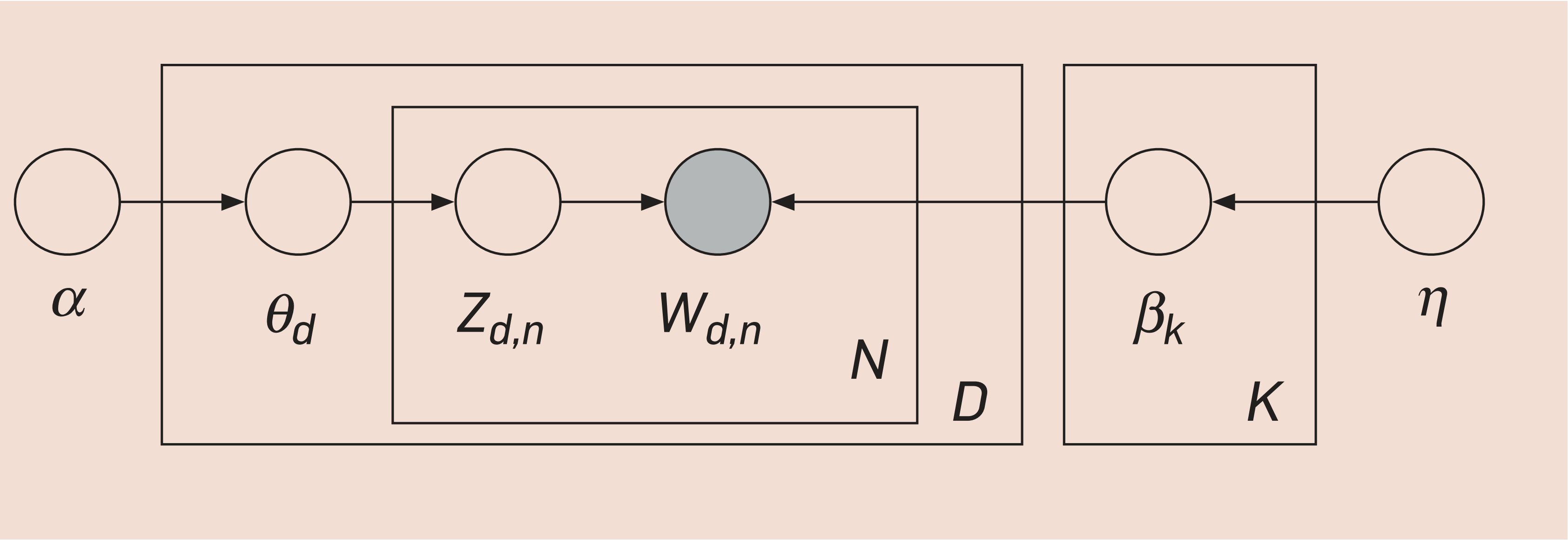
Recall LDA...



...there are multiple documents, but we don't care about their order.

Dynamic Topic Model (Blei & Lafferty, 2006)

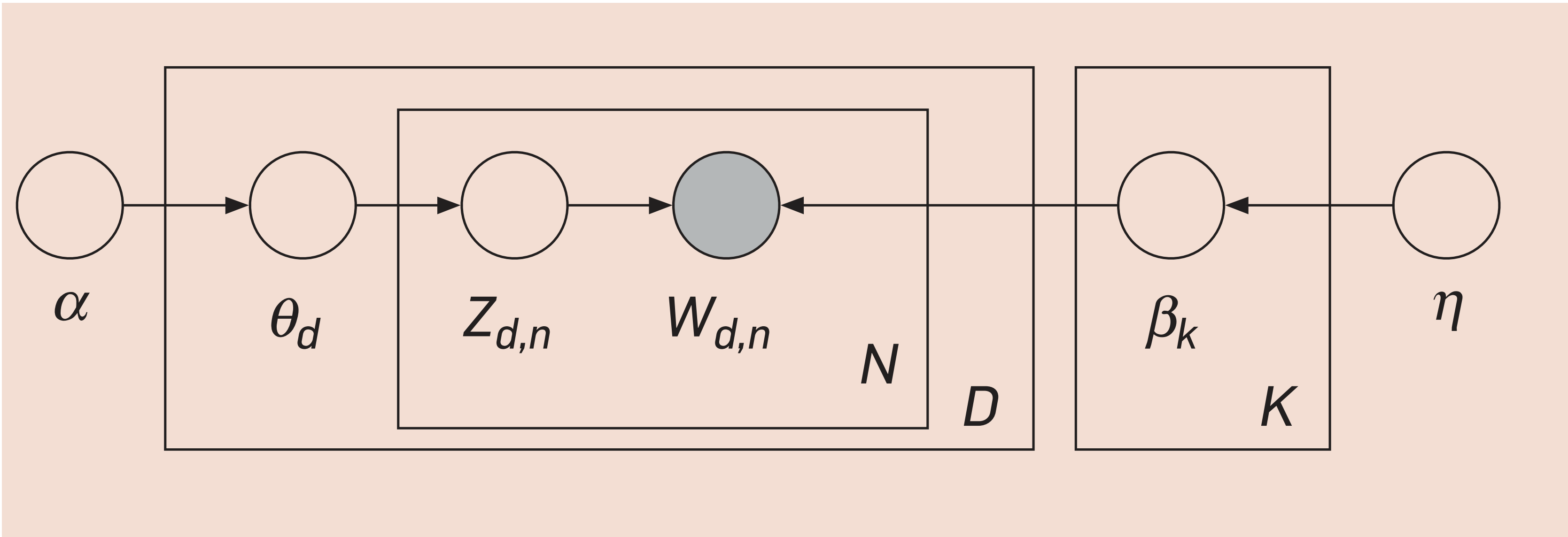
Recall LDA...



...there are multiple documents, but we don't care about their order. Our results are the 'same' regardless of how we reorder the documents and feed them to the model.

Dynamic Topic Model (Blei & Lafferty, 2006)

Recall LDA. . .

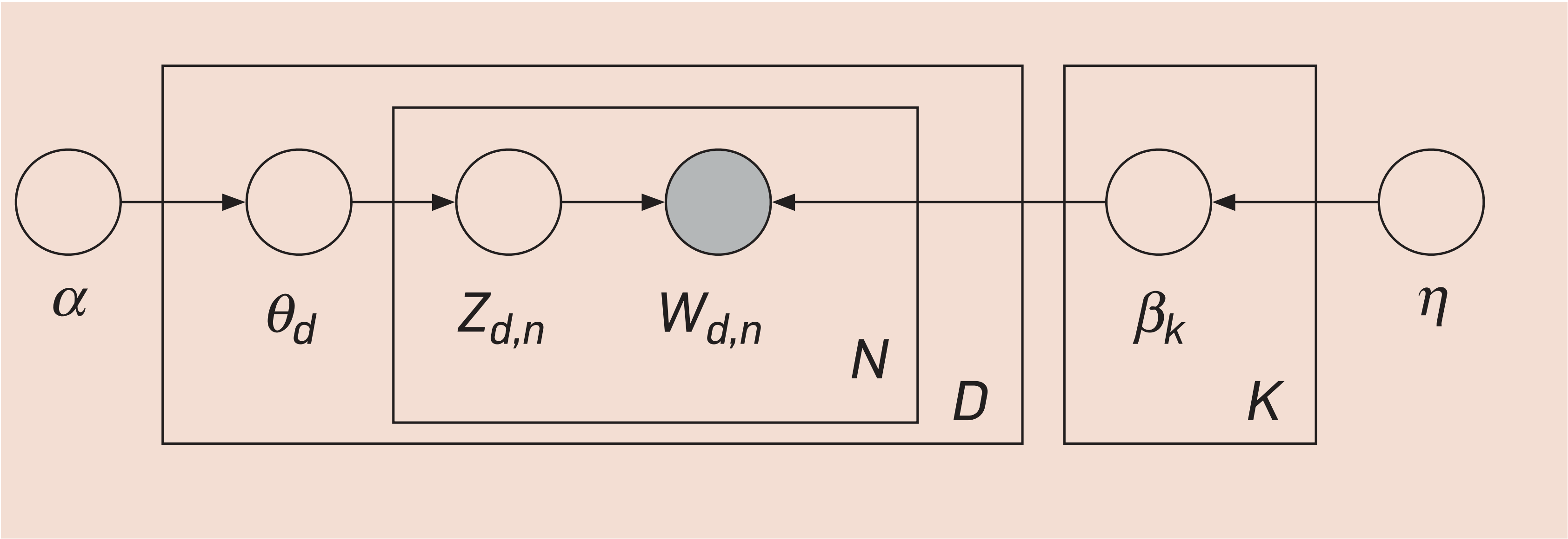


. . . there are multiple documents, but we don't care about their order. Our results are the 'same' regardless of how we reorder the documents and feed them to the model.

Dynamic Topic Model has a different model for each time period,

Dynamic Topic Model (Blei & Lafferty, 2006)

Recall LDA. . .

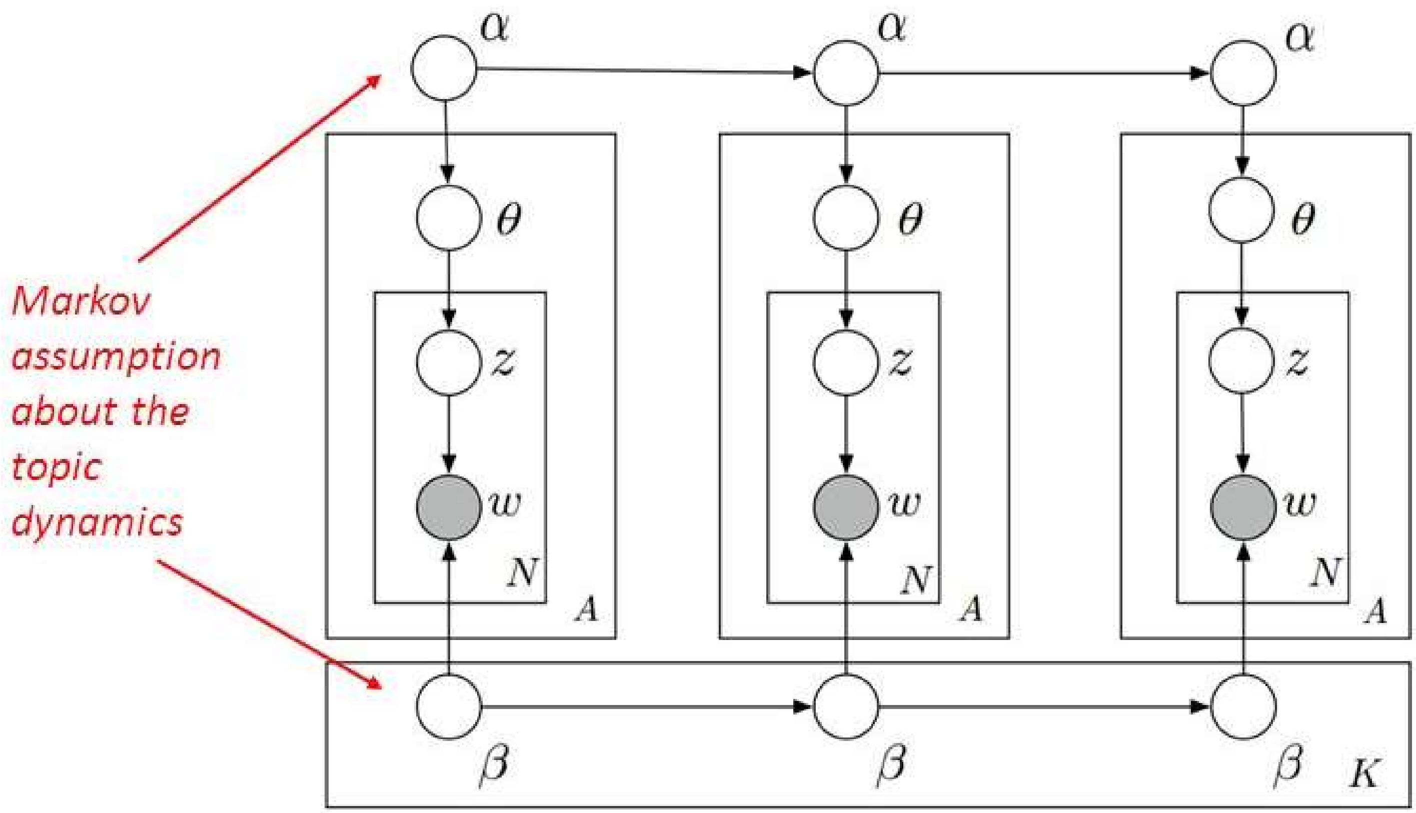


. . . there are multiple documents, but we don't care about their order. Our results are the 'same' regardless of how we reorder the documents and feed them to the model.

Dynamic Topic Model has a different model for each time period, with topics allowed to evolve over time. . .

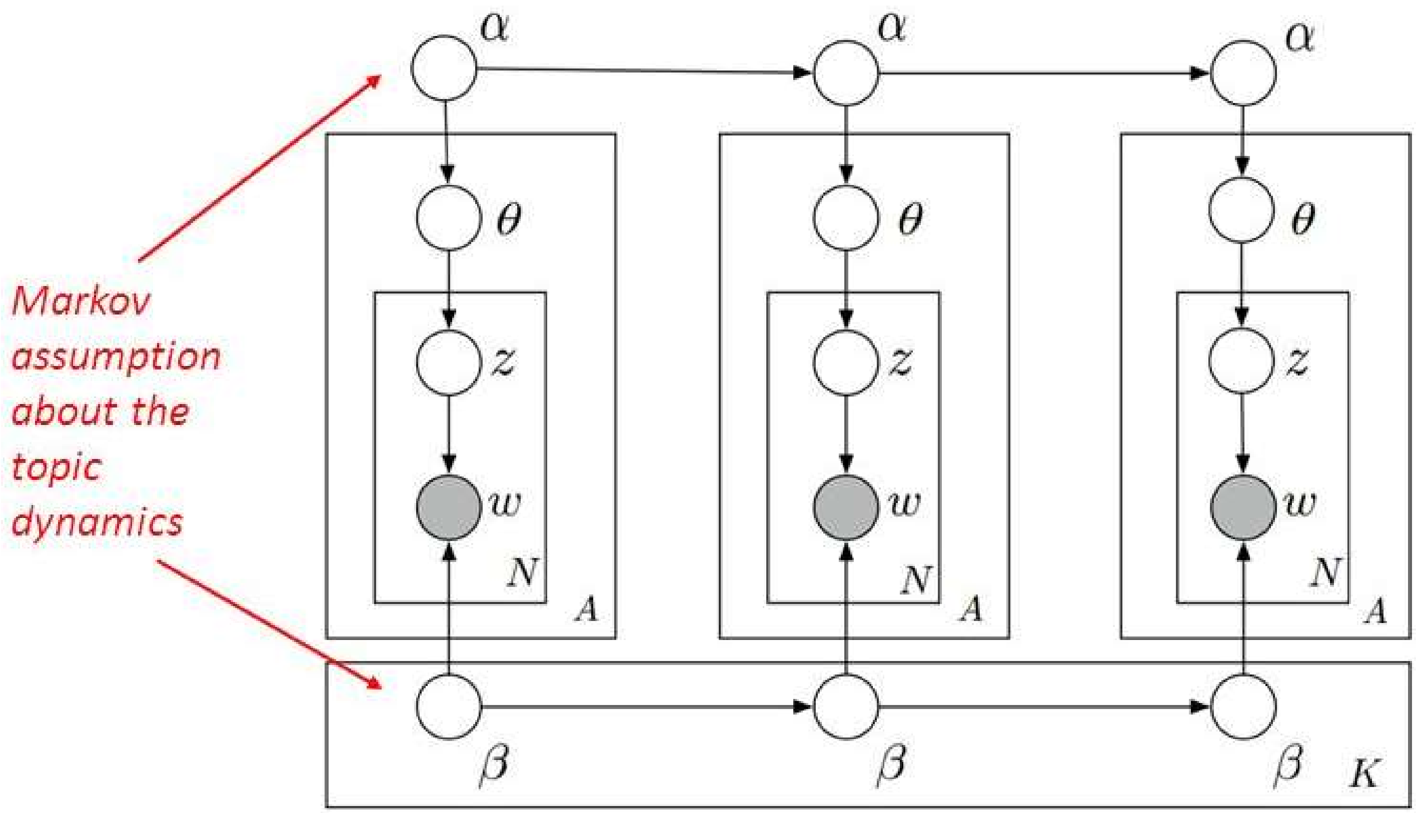
So. . .

So. . .



Markov
assumption
about the
topic
dynamics
α
α
α
θ
θ
θ
z
z
z
w
w
w
N
N
N
A
A
A
β
β
β
K

So. . .



Now, mean parameters for the topic proportions (αs) and the what's in the topics (in terms of words, βs) are connected over time via a simple evolutionary process (West & Harrison, 1997).

How to Analyze Political Attention with Minimal Assumptions and Costs (Quinn et al, 2010)

How to Analyze Political Attention with Minimal Assumptions and Costs (Quinn et al, 2010)

What is the agenda of the Senate?

How to Analyze Political Attention with Minimal Assumptions and Costs (Quinn et al, 2010)

What is the agenda of the Senate? $\rightarrow$ dynamic topic model of over $100,000$ speeches.

How to Analyze Political Attention with Minimal Assumptions and Costs (Quinn et al, 2010)

What is the agenda of the Senate? $\rightarrow$ dynamic topic model of over $100,000$ speeches.

Assume slightly different evolution process (Dynamic Linear Model),

How to Analyze Political Attention with Minimal Assumptions and Costs (Quinn et al, 2010)

What is the agenda of the Senate? $\rightarrow$ dynamic topic model of over $100,000$ speeches.

Assume slightly different evolution process (Dynamic Linear Model), and only one topic per speech (like Grimmer).

How to Analyze Political Attention with Minimal Assumptions and Costs (Quinn et al, 2010)

What is the agenda of the Senate? $\rightarrow$ dynamic topic model of over $100,000$ speeches.

Assume slightly different evolution process (Dynamic Linear Model), and only one topic per speech (like Grimmer). Model is fit differently too.

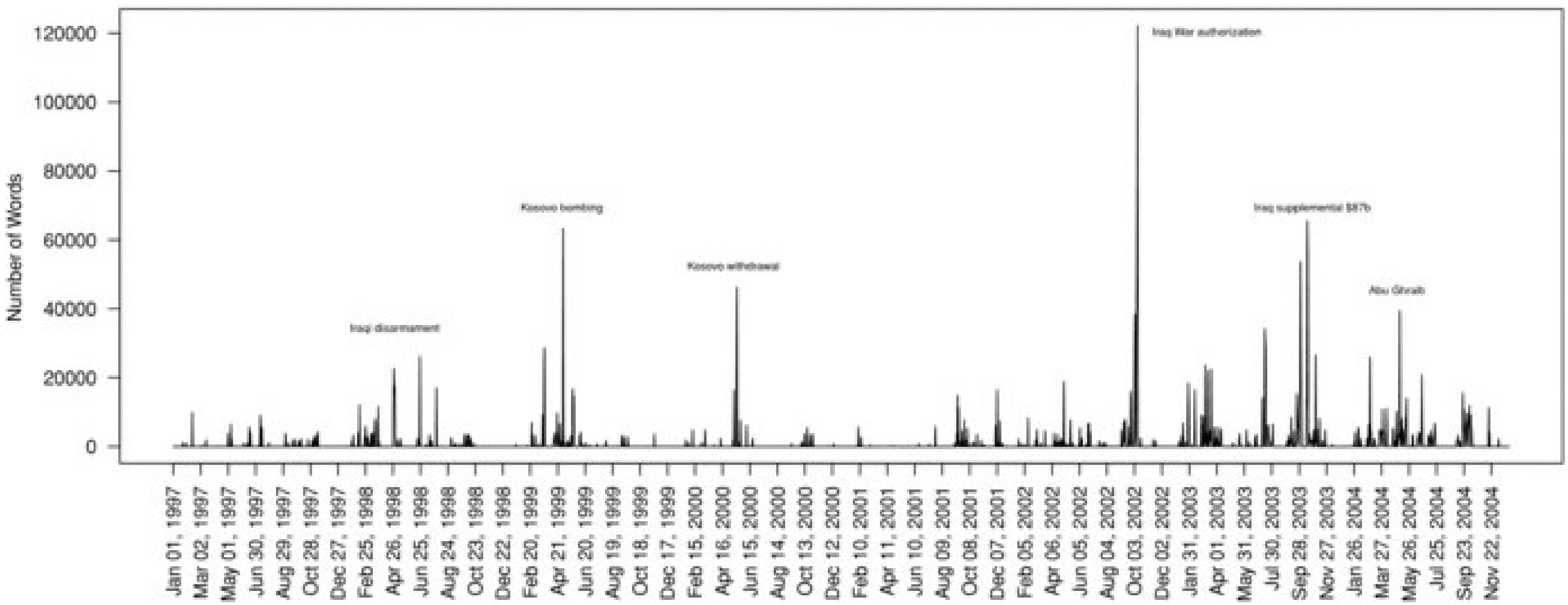
How to Analyze Political Attention with Minimal Assumptions and Costs (Quinn et al, 2010)

What is the agenda of the Senate? $\rightarrow$ dynamic topic model of over $100,000$ speeches.

Assume slightly different evolution process (Dynamic Linear Model), and only one topic per speech (like Grimmer). Model is fit differently too.

BTW, paper has a lot of validation!

Attention to Defense [Use of Force]



(b) *The Number of Words on the 'Defense [Use of Force]' Topic Per Day*

Structural Topic Model (Roberts et al.)

Structural Topic Model (Roberts et al.)

STM = LDA + contextual information.

Structural Topic Model (Roberts et al.)

STM = LDA + contextual information.

$\rightarrow$ topic prevalence varies by covariates

Structural Topic Model (Roberts et al.)

STM = LDA + contextual information.

$\rightarrow$ topic prevalence varies by covariates

e.g. women may report issues with depression more than men do.

Structural Topic Model (Roberts et al.)

STM = LDA + contextual information.

$\rightarrow$ topic prevalence varies by covariates

e.g. women may report issues with depression more than men do.

$\rightarrow$ topic content varies by covariates

Structural Topic Model (Roberts et al.)

STM = LDA + contextual information.

$\rightarrow$ topic prevalence varies by covariates

e.g. women may report issues with depression more than men do.

$\rightarrow$ topic content varies by covariates

e.g. young people ($X = 0$) may talk about depression differently to the way old people ($X = 1$) talk about it.

Structural Topic Model (Roberts et al.)

STM = LDA + contextual information.

$\rightarrow$ topic prevalence varies by covariates

e.g. women may report issues with depression more than men do.

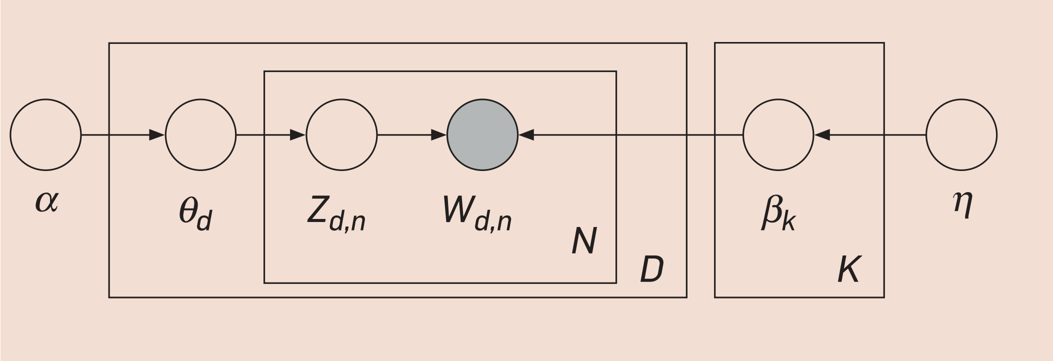
$\rightarrow$ topic content varies by covariates

e.g. young people ($X = 0$) may talk about depression differently to the way old people ($X = 1$) talk about it.

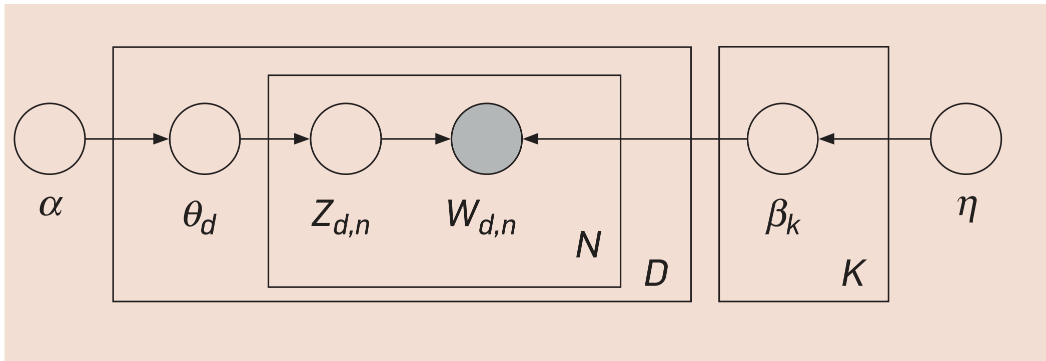
Including covariates allows for (a) more accurate estimation and (b) better interpretatability.

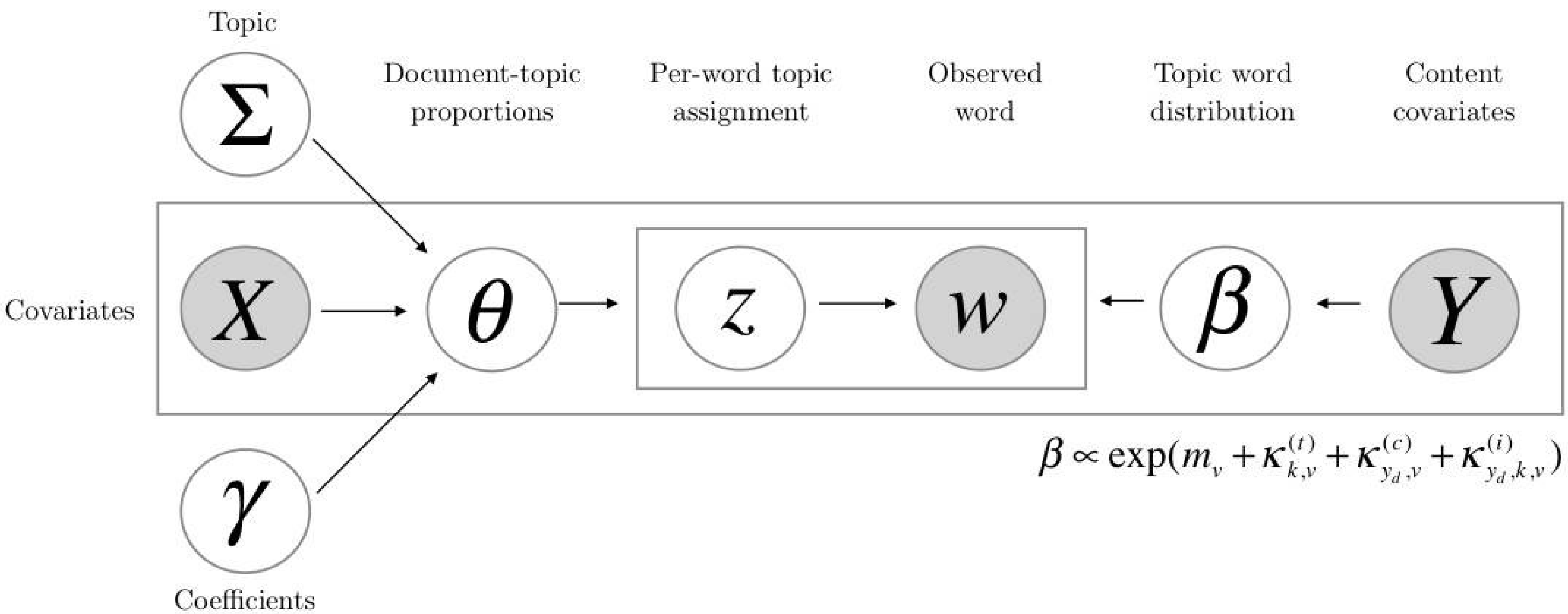
Compare: Plate Diagram

Compare: Plate Diagram



Compare: Plate Diagram





$$\beta \propto \exp(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_d,v} + \kappa^{(i)}_{y_d,k,v})$$

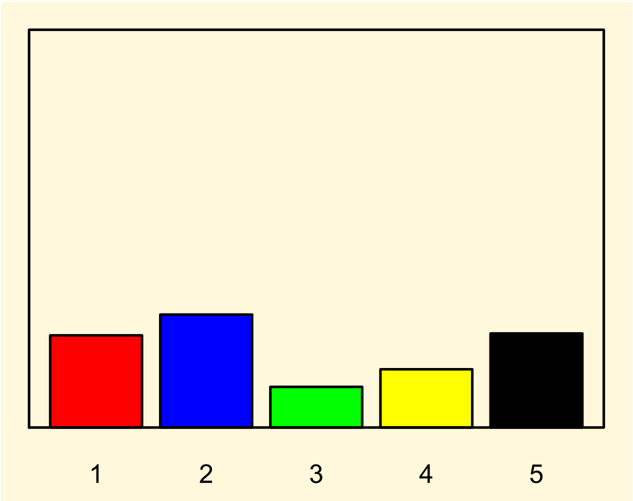
Compare: Per Document Topic Distribution (θ)

Compare: Per Document Topic Distribution (θ)

LDA: each document has some topic distribution.

Compare: Per Document Topic Distribution (θ)

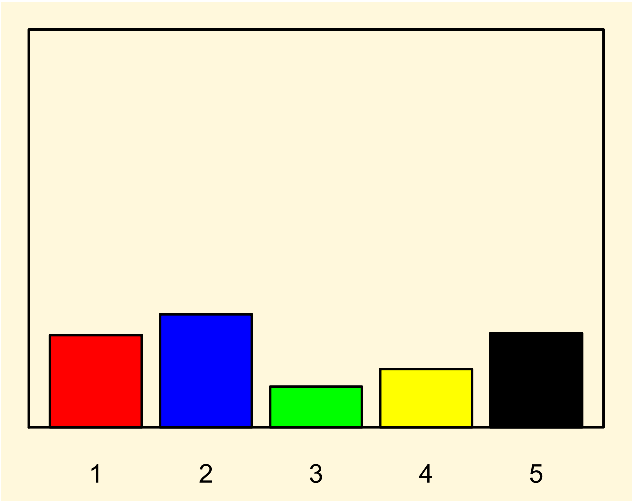
LDA: each document has some topic distribution.



Compare: Per Document Topic Distribution (θ)

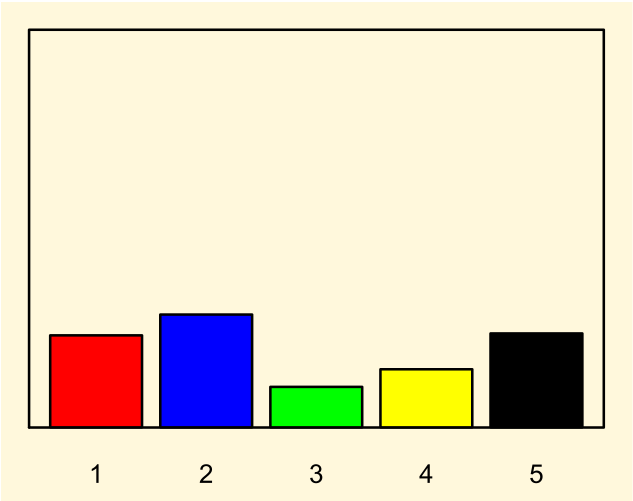
LDA: each document has some topic distribution.

STM, that topic distribution ('prevalence') is a function of the document metadata.



Compare: Per Document Topic Distribution (θ)

LDA: each document has some topic distribution.

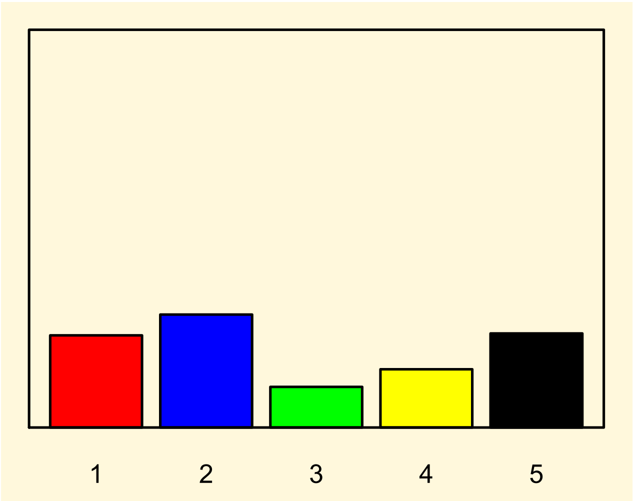


STM, that topic distribution ('prevalence') is a function of the document metadata.

e.g. perhaps male author ($X = 0$) documents have different topics relative to female ($X = 1$) author docs.

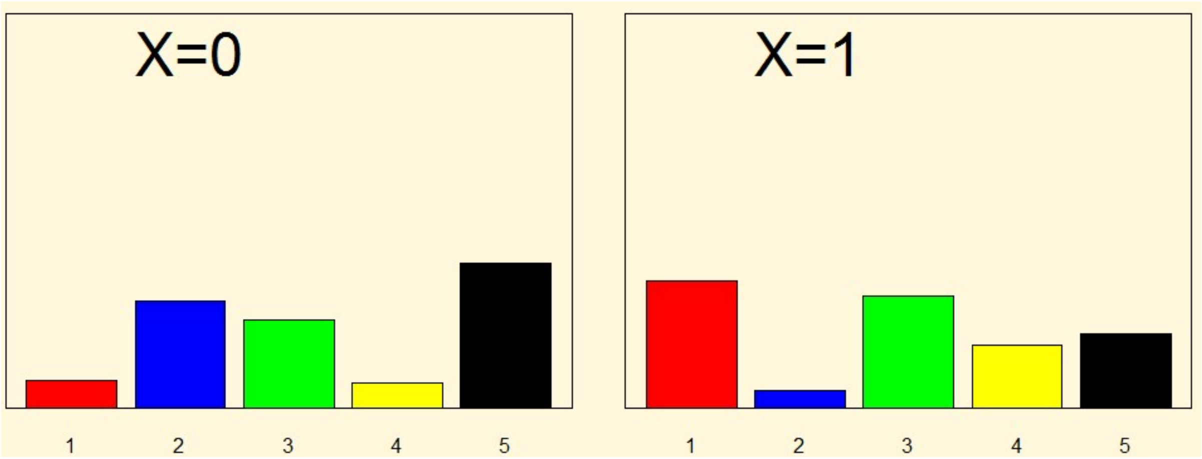
Compare: Per Document Topic Distribution (θ)

LDA: each document has some topic distribution.



STM, that topic distribution ('prevalence') is a function of the document metadata.

e.g. perhaps male author ($X = 0$) documents have different topics relative to female ($X = 1$) author docs.



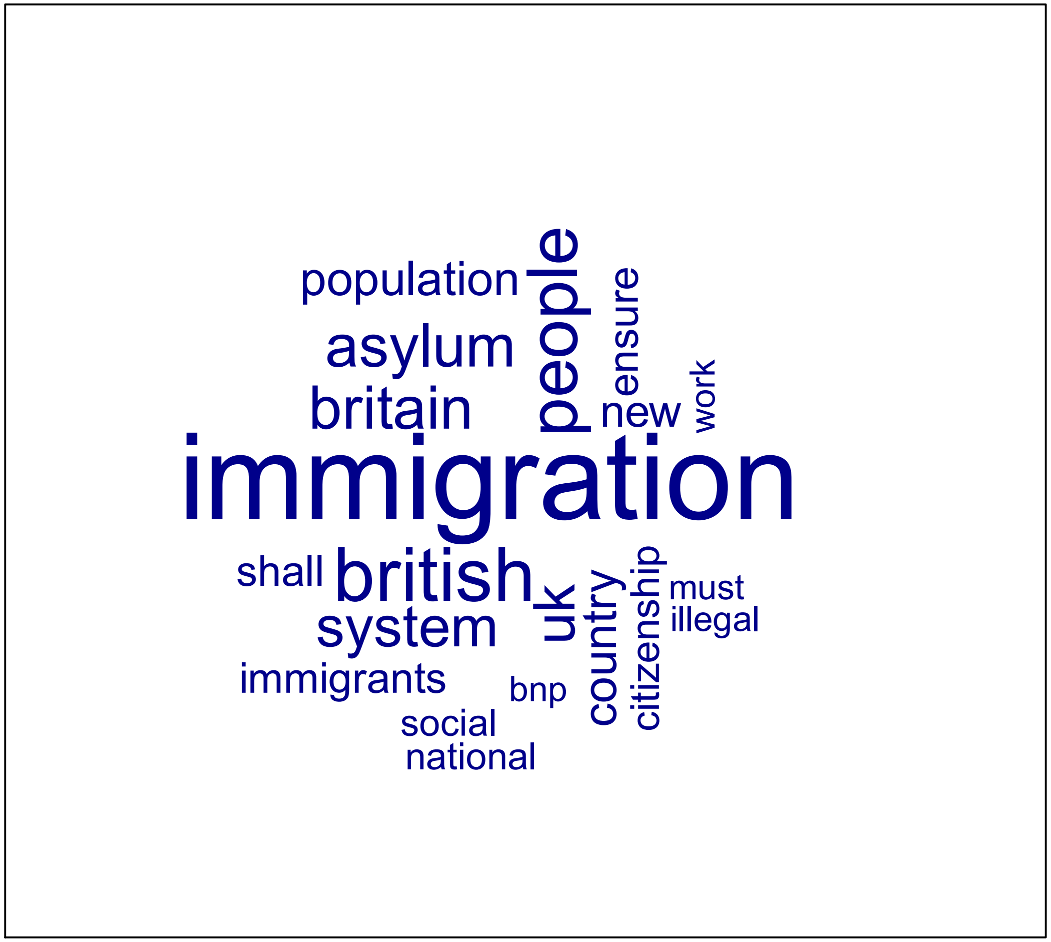
Compare: Per Topic Word Distribution (β)

Compare: Per Topic Word Distribution (β)

LDA: topic ('immigration') has a given distribution over words.

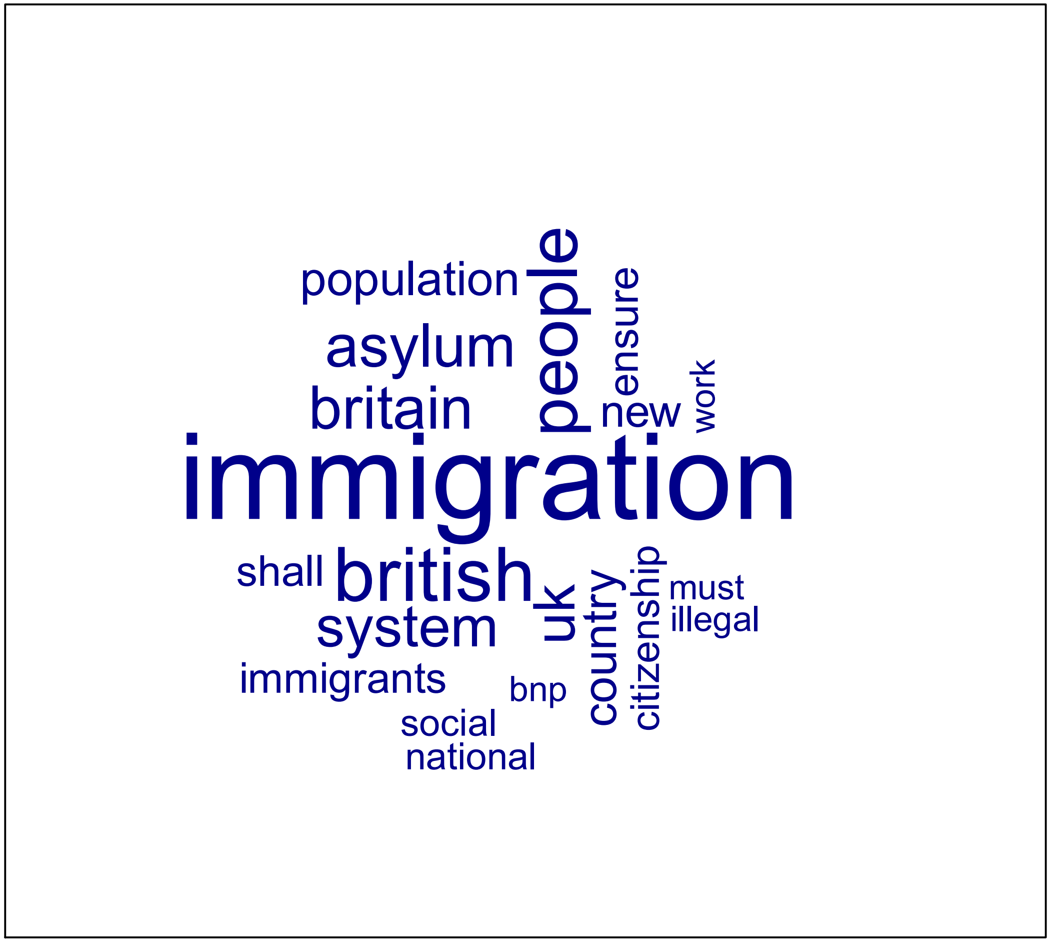
Compare: Per Topic Word Distribution (β)

LDA: topic ('immigration') has a given distribution over words.



Compare: Per Topic Word Distribution (β)

LDA: topic ('immigration') has a given distribution over words.



STM: that word distribution ('content') is a function of the document metadata.

STM: that word distribution ('content') is a function of the document metadata.

e.g. perhaps right parties ($Y = 0$) talk about a given topic differently to left ($Y = 1$) parties.

STM: that word distribution ('content') is a function of the document metadata.

e.g. perhaps right parties ($Y = 0$) talk about a given topic differently to left ($Y = 1$) parties.

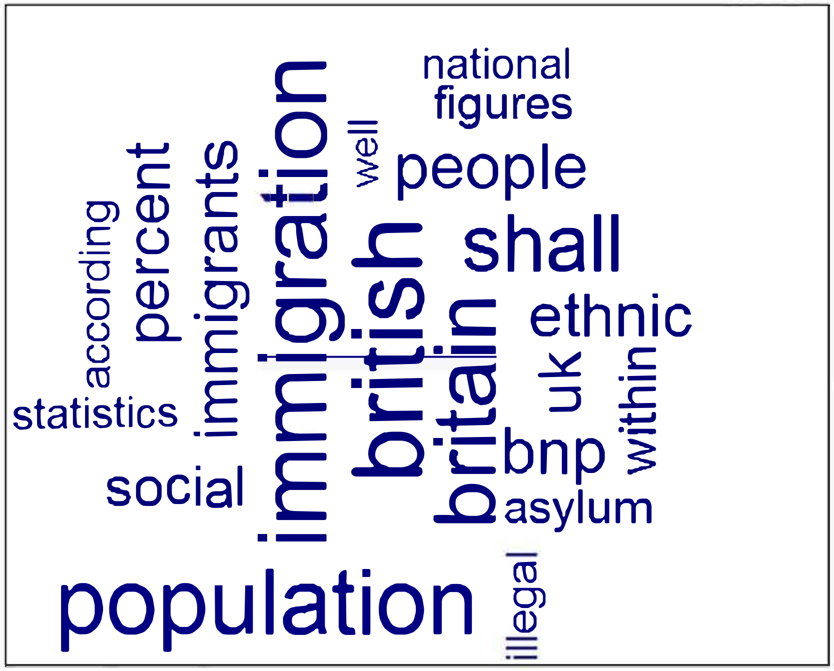
In practice, `content` needs to a single discrete variable.

STM: that word distribution ('content') is a function of the document metadata.

e.g. perhaps right parties ($Y = 0$) talk about a given topic differently to left ($Y = 1$) parties.

In practice, `content` needs to a single discrete variable.

Y=0



Y=1