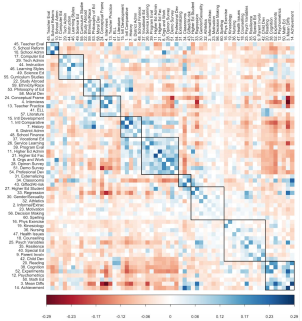
Topic Models

Sebastian Munoz-Najar

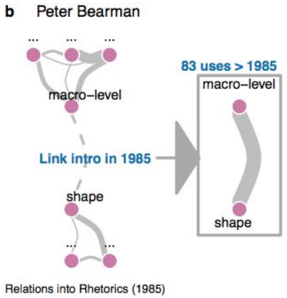
My research with topic models

Education Research 1980-2010
AERJ ‘19



How did education research discourse change over time? **TMs for research trends.**

Diversity Breeds Innovation
Arxiv ‘19



How are grad students rewarded for innovation? **TMs for phrase extraction.**

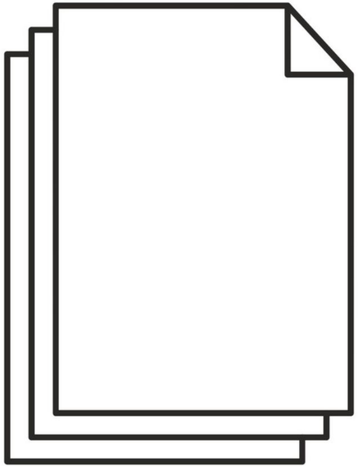
Today

- What is a topic model?
  - Intuition
  - Inputs and Outputs
  - Notation
- Kinds of topic models
- Validation
  - Exclusivity
  - Coherence
- Examples

What is a topic model?

The short answer

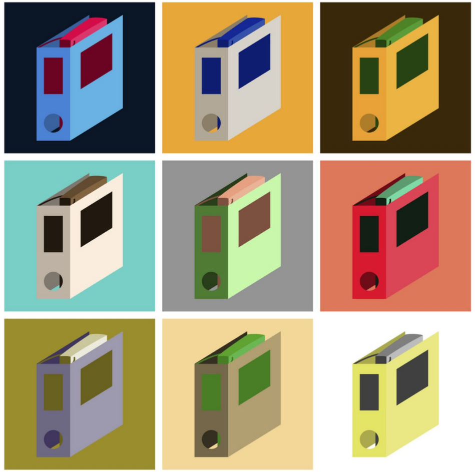
You have...



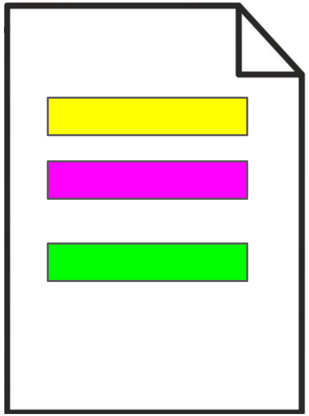
many texts

You want to know...

What the collection is about



What each text is about



Some involved answers

Technical Answer	**Translation**
Unsupervised Learning Algorithm	Classifies text without a predetermined classification scheme
Generative Probabilistic Model	Answers, “How is a text put together?” with probability distributions
Mixture Model	Represents texts as combinations of topics

Intuition

What is this text about? (i.e. Big themes)

This study used structural equation modeling to test a model of ethnic identity development among 513 Mexican-origin adolescents living in the United States.

(Umana-Taylor & Fine, 2004)

What did you need to know about language and the world?

What you know

- Language
  - Grammar
  - Vocabulary
  - Pragmatics
- Domain Knowledge
  - Intertextuality
  - Style
- Context Knowledge
  - Year
  - Author
  - Institution

What a TM knows (approximately)

- Language
  - ~~Grammar~~
  - **Vocabulary**
  - **~Pragmatics**
- Domain Knowledge
  - **~Intertextuality**
  - ~~Style~~
- Context Knowledge
  - **Year***
  - **Author***
  - **Institution***

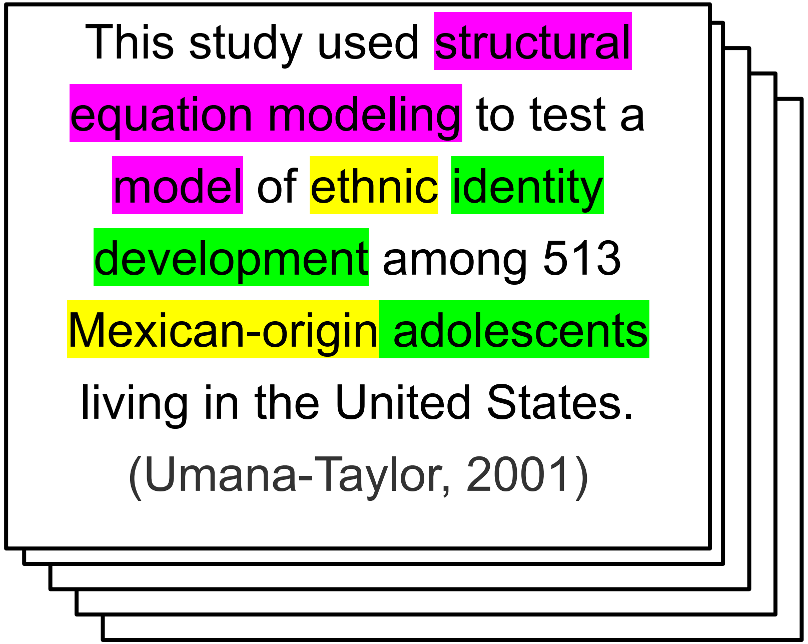
~ : approximates
* : depends on model

What is this text about? (i.e. Big themes)

model,
structural_equation_model,
factor_analysis,
….

adolescent,
identitiy_development,
socialization
….

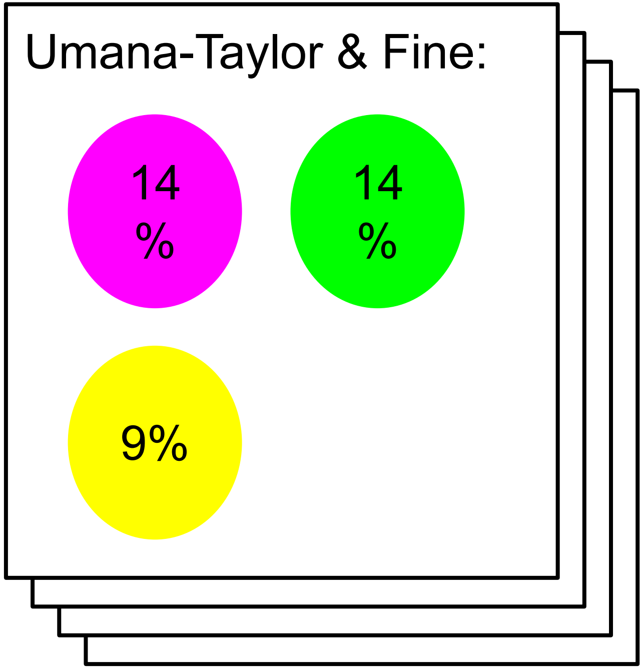
ethnic,
mexican-origin,
chicano
….



This study used structural equation modeling to test a model of ethnic identity development among 513 Mexican-origin adolescents living in the United States.

(Umana-Taylor, 2001)

What is this text about? (i.e. Big themes)



Umana-Taylor & Fine:

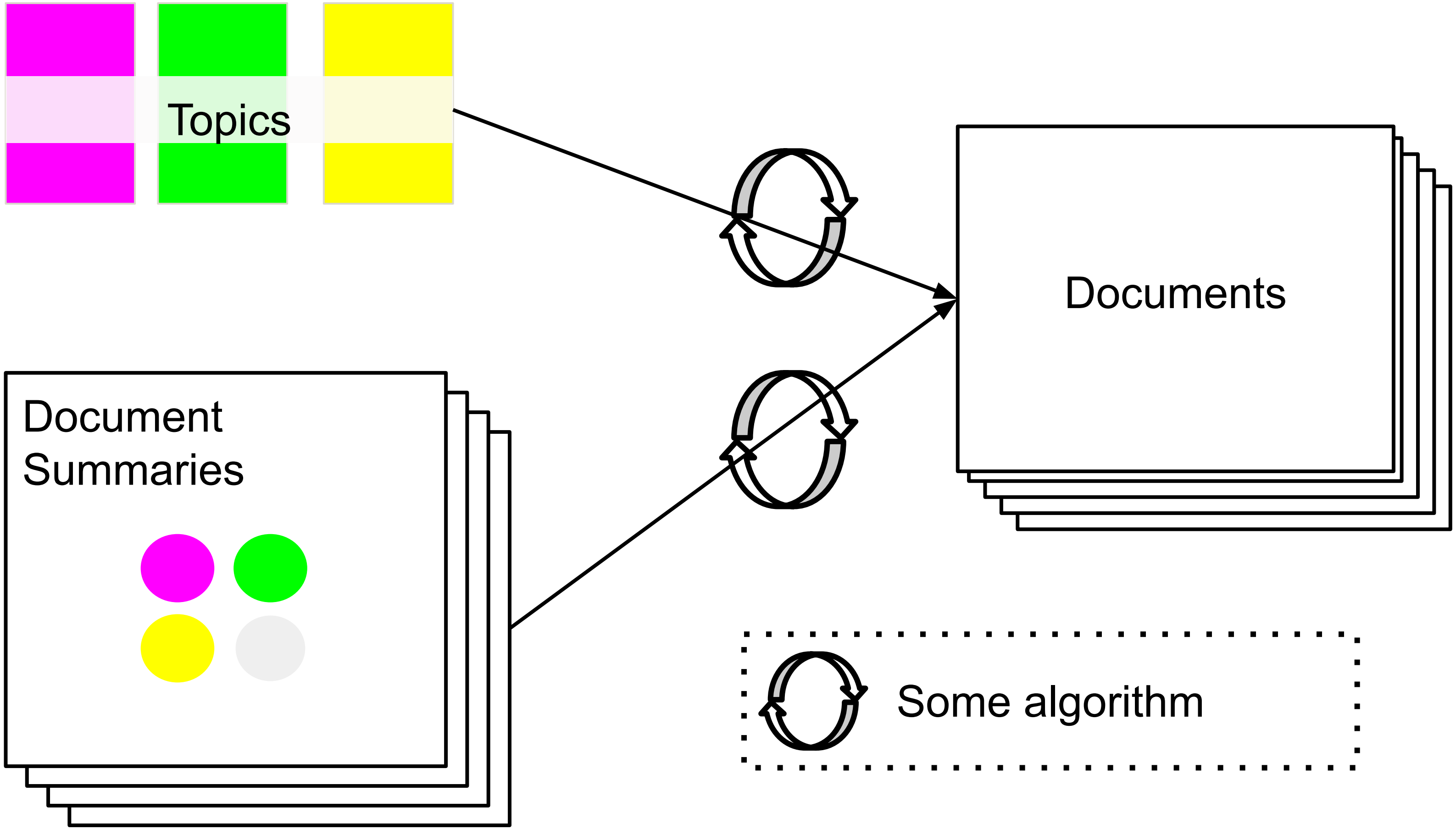
14 %

14 %

9%

This study used structural equation modeling to test a model of ethnic identity development among 513 Mexican-origin adolescents living in the United States.

(Umana-Taylor, 2001)


Topics
Documents
Document Summaries
Some algorithm

Inputs and outputs

Input

- Document-Term Matrix (MxV)
  - M documents
  - V words
- Number of Topics (K)
- Covariates (MxX)

ID	**educ**	**...**	**stud**	**V**
1	3	...	1	10
...	...	...	...	...
M	5	...	5	0

Output

- Topics (KxV)
  - Word Proportions
- Document Summaries (MxK)
  - Topic Proportions

ID	**Topic 1**	**...**	**Topic K**
1	.001	...	.2
...	...	...	...
M	.1	...	.00009

Topics	**educ**	**...**	**stud**	**V**
1	.2	...	.2	.001
…	...	...	...	...
K	.2	...	.001	.6

Notation

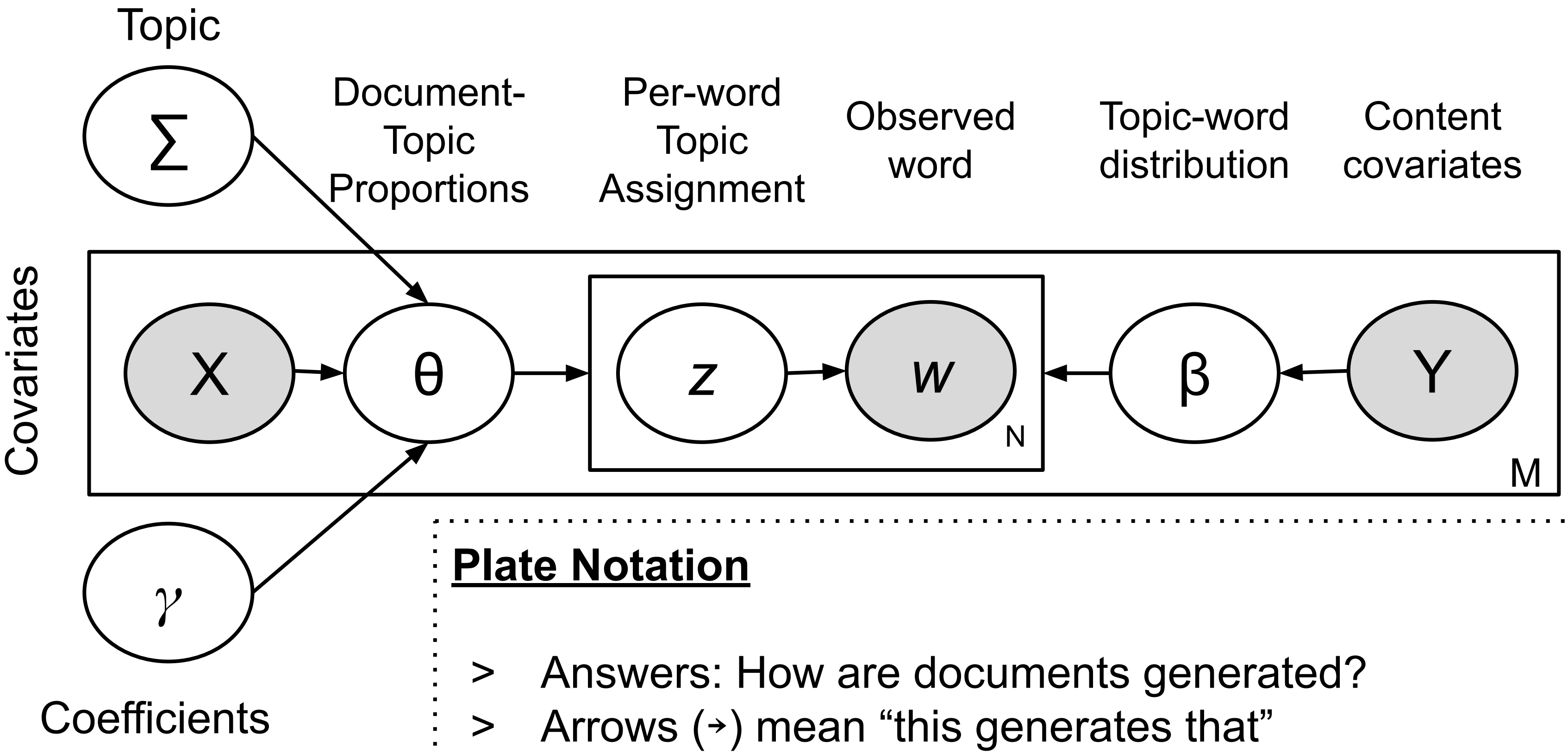


Plate Notation

> Answers: How are documents generated?

> Arrows (→) mean “this generates that”

> Plates say how many times you repeat a process:

◆ M documents, N terms per document

Kinds of TM

Name	**Gist**	**Learns**
LDA	Vanilla	Topics, Document Summaries
Correlational TM	Topics are correlated!	Topic Correlations
Dynamic TM	Words in topics change!	Dynamic Topics
Hierarchical TM	Topics have hierarchy!	Topic Hierarchy
Bayesian Hierarchical TM	Documents are nested in authors!	Author topical priorities
Labelled LDA	Documents have labels!	Labelled Topics
Relational TM, Topic-Link LDA	Documents have relations!	Links
Structural TM	Documents have all kinds of metadata!	Coefficients of covariates

Validation

Goal: Select Number of Topics (K)

What is wrong with these topics?

education,
model,
structural_equation_model,
factor_analysis,

education,
adolescent,
identitiy_development,
socialization

education
ethnic,
mexican-origin,
chicano

What is wrong with these topics?

education,
model,
structural_equation_model,
factor_analysis,

education,
adolescent,
identitiy_development,
socialization

education,
ethnic,
mexican-origin,
chicano

They lack exclusivity!

What is wrong with this topic?

model,
structural_equation_model,
factor_analysis, ...

All Documents:

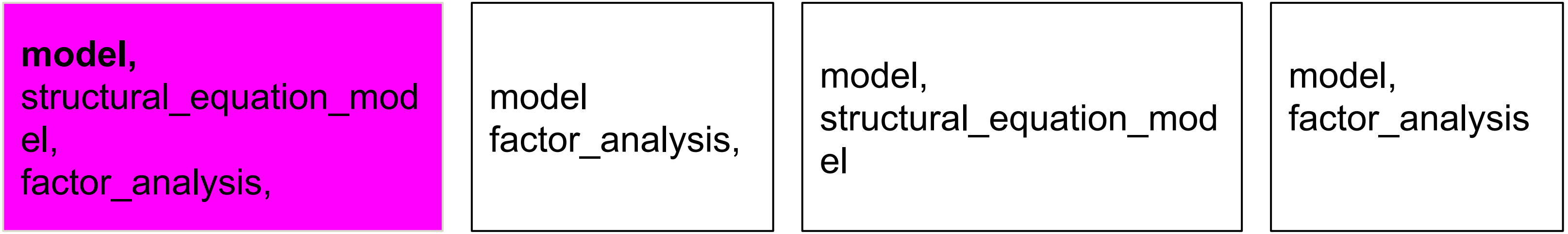
model,
factor_analysis,

model,
structural_equation_model

model,
factor_analysis

What is wrong with this topic?

All Documents:



model,
structural_equation_model,
factor_analysis,

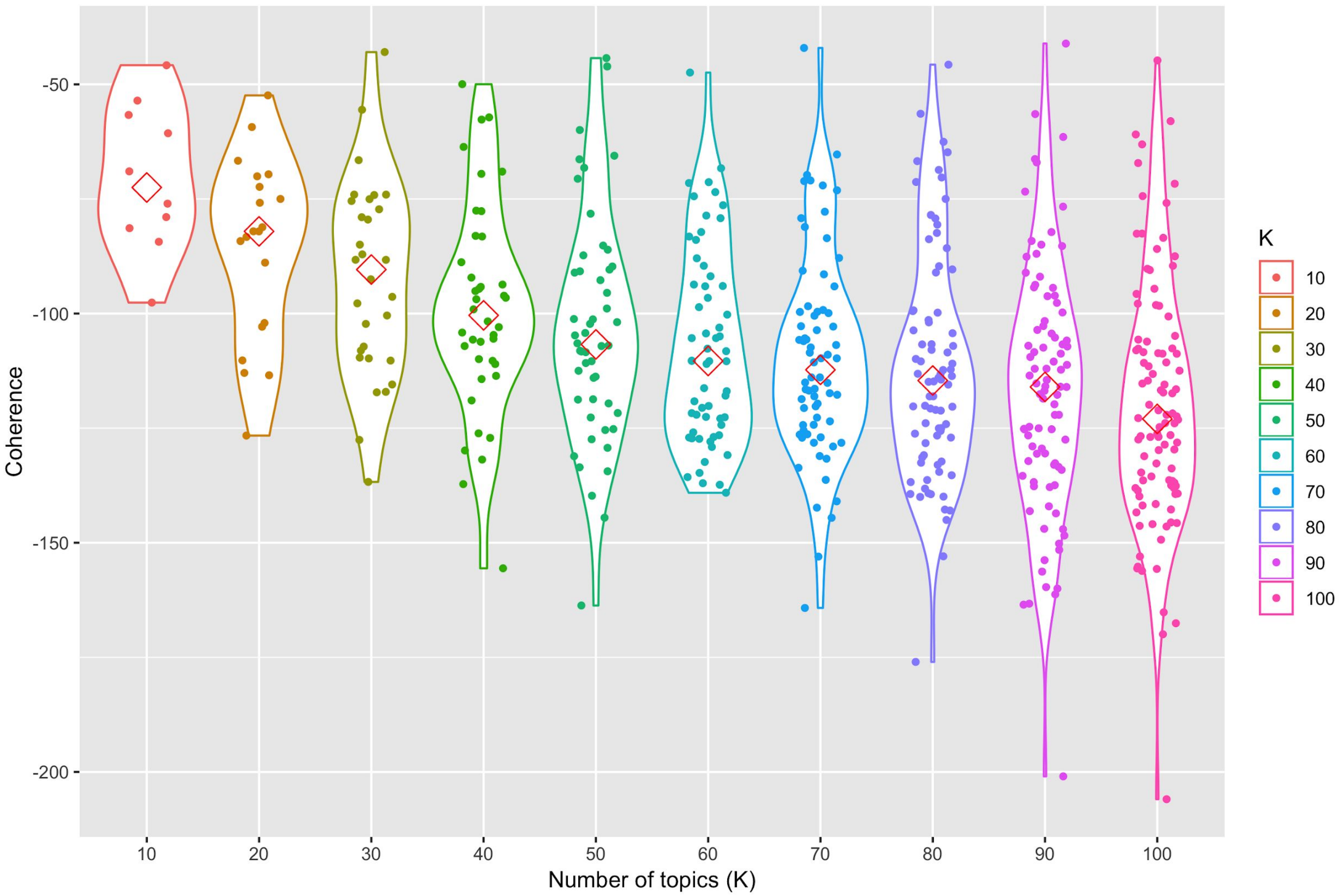
model
factor_analysis,

model,
structural_equation_model

model,
factor_analysis

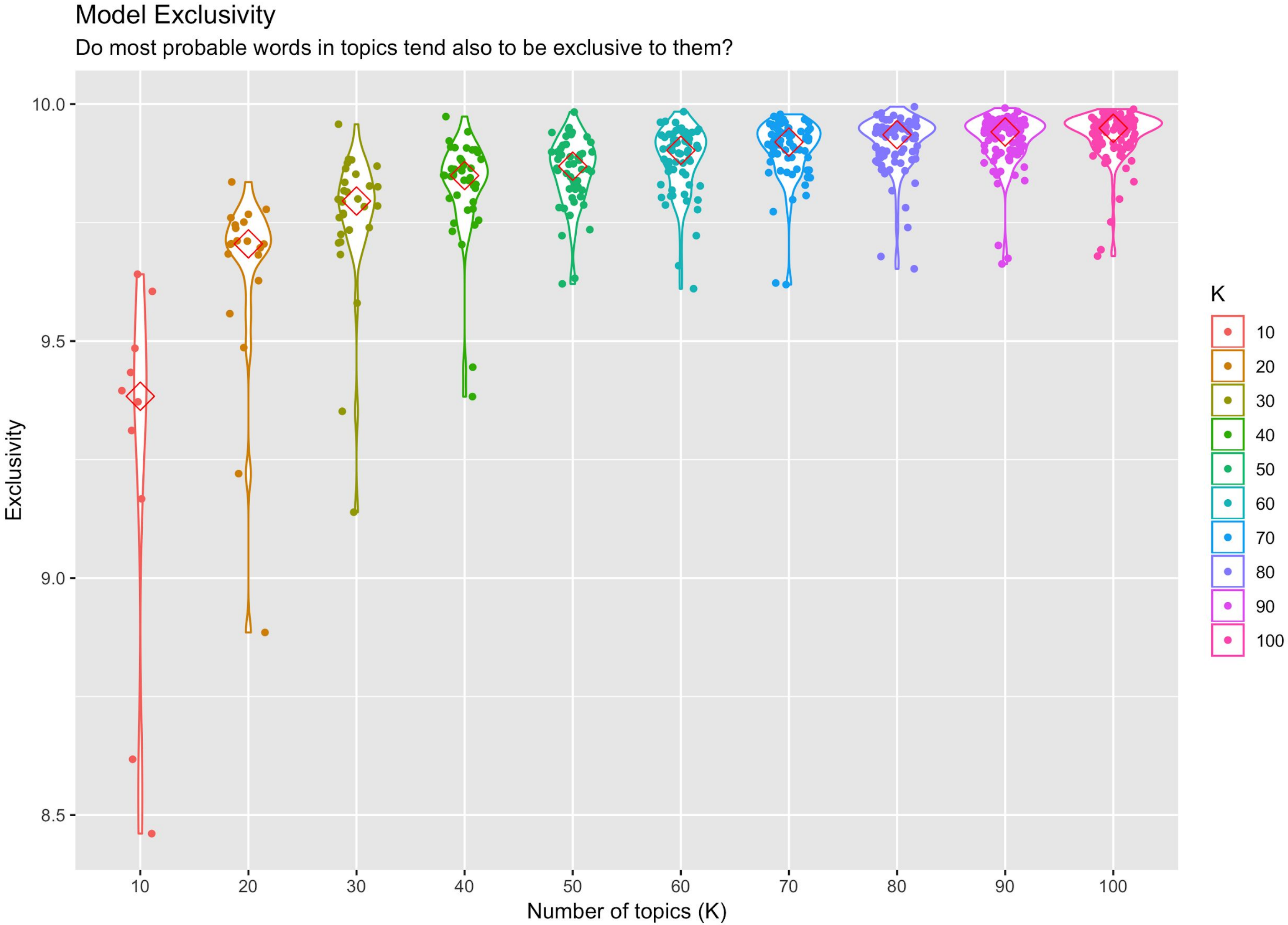
They lack coherence!

Can we trivially maximize these metrics?



Model Coherence

Do the pairs of most probable words in each topic have high co-document frequencies?



Model Exclusivity

Do most probable words in topics tend also to be exclusive to them?

Examples

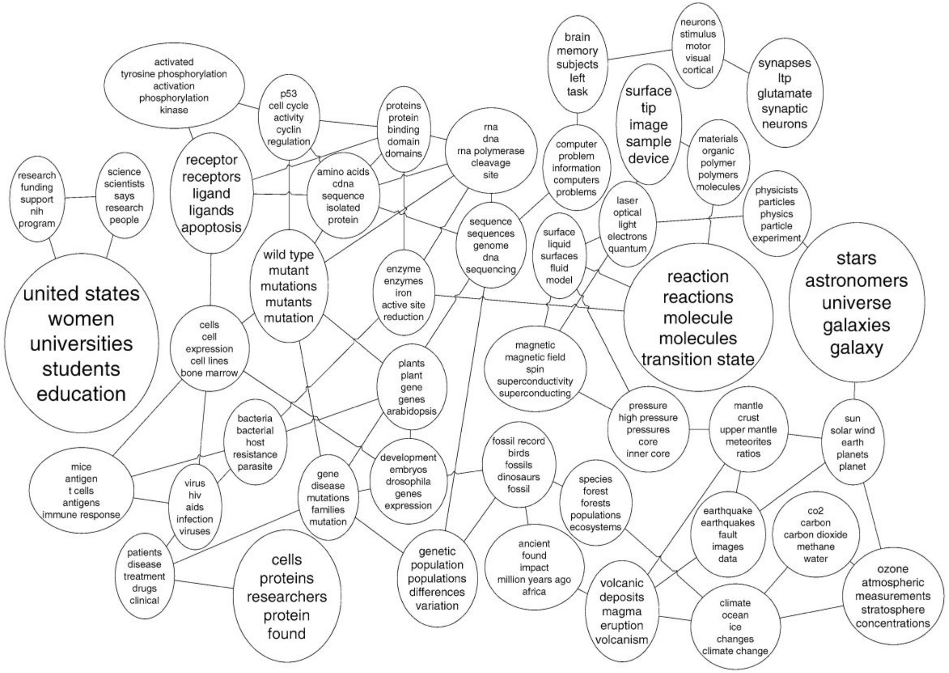
David Blei and John Lafferty (2007), A Correlated Topic Model of Science

Question: Can we model topic correlations?

Model: Correlational Topic Model

Data: articles from Science published from 1990–1999

Result: Yes



Derek Greene and James P., (2017) Cross Exploring the Political Agenda of the European Parliament

Question: How has the political agenda of the European Parliament (EP) plenary evolved over time?

Model: Dynamic Topic Model

Data: all English language legislative speeches in the EP plenary from the period 1999 to 2014

Result: Agenda responds to external shocks!

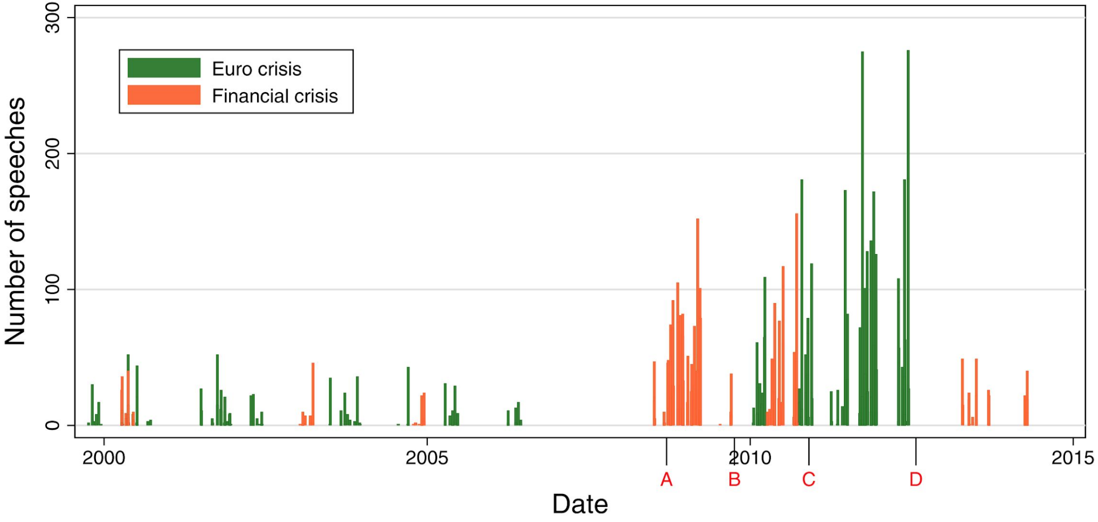


Figure 7. “Financial & Euro crises” dynamic topics.

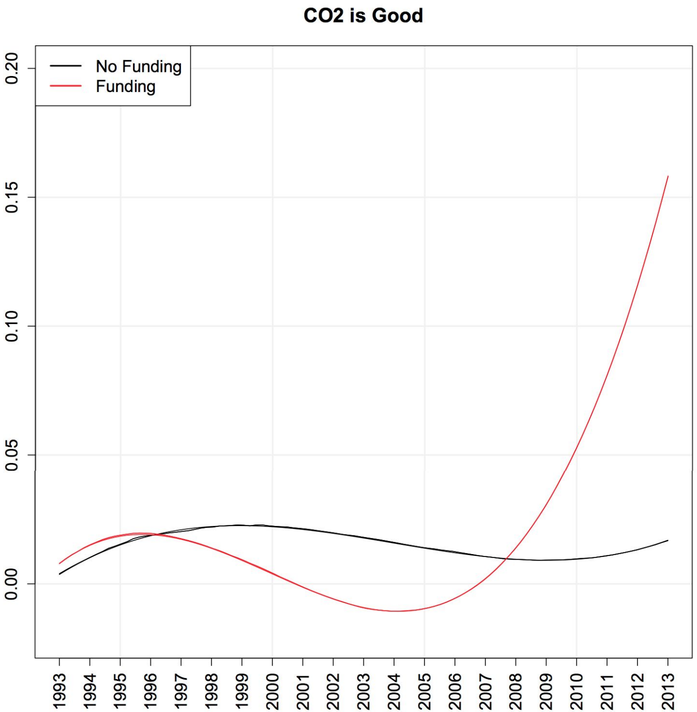
Justin Farrell (2016), Corporate funding and ideological polarization about climate change

Question: How does corporate funding influence the production and content of polarization efforts in climate change debates?

Model: Structural Topic Model

Data: every text about climate change produced by 164 organizations between 1993 and 2013

Result: organizations that received corporate funding were more likely to have written and disseminated contrarian texts.


CO2 is Good
No Funding
Funding