Extensions and Adaptations of LDA

Lisa Posch

Topic Model Tutorial
Hannover, 2016

Why do we want to extend LDA?

We want to extend LDA so that we can

- include different characteristics
- explore different aspects

of our dataset.

Outline

- Quick Recap of LDA
- Extensions and Adaptations

Outline

- Quick Recap of LDA
- Extensions and Adaptations:
  - Labeled LDA
  - Polylingual Topic Model
  - Author-Topic Model
  - Topics over Time
  - Citation Influence Model

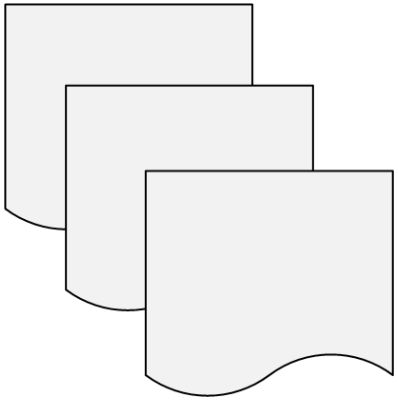
Goal of this Session

- You know that there are different topic models that are based on LDA.
- You have seen some specific adaptations of LDA
- and you know what they are used for.

LDA (recap)

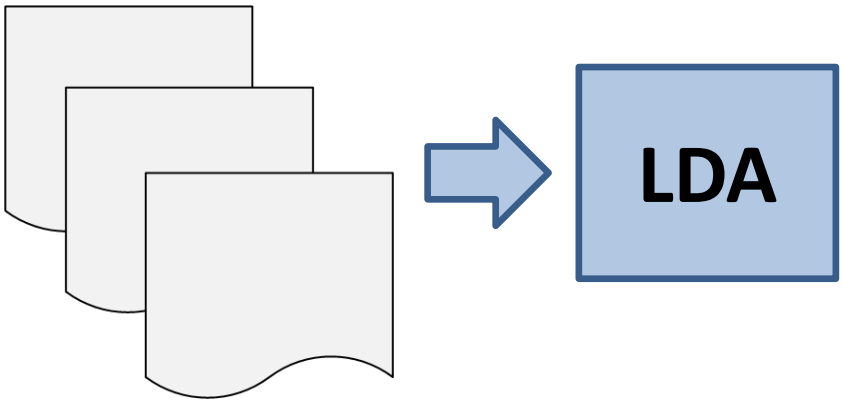
LDA

Collection of Documents



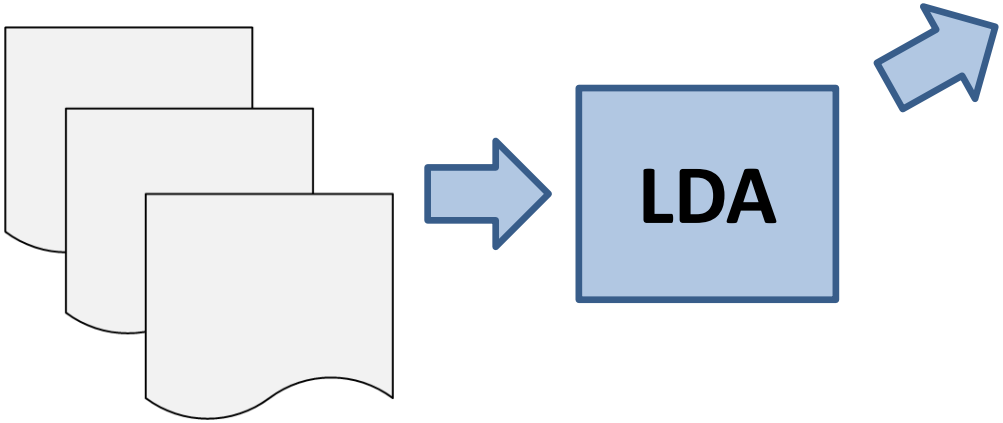
LDA

Collection of Documents



LDA

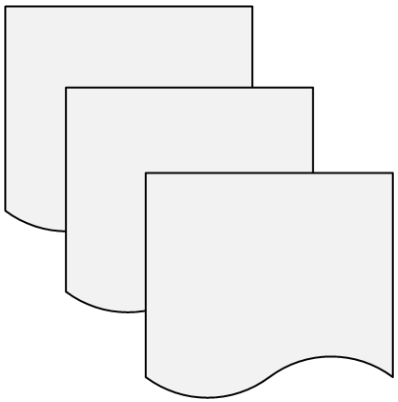
Collection of Documents

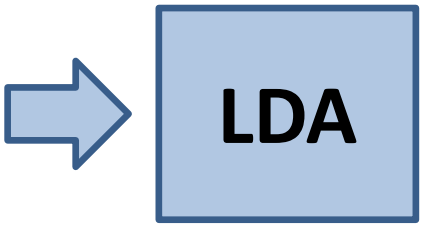


Topic 1	Topic 2	Topic 3
websci	food	sociology
conference	restaurant	social
germany	pizza	society
hannover	eat	behavior
social	steak	relationships
computer	cafe	quantitative

LDA

Collection of Documents





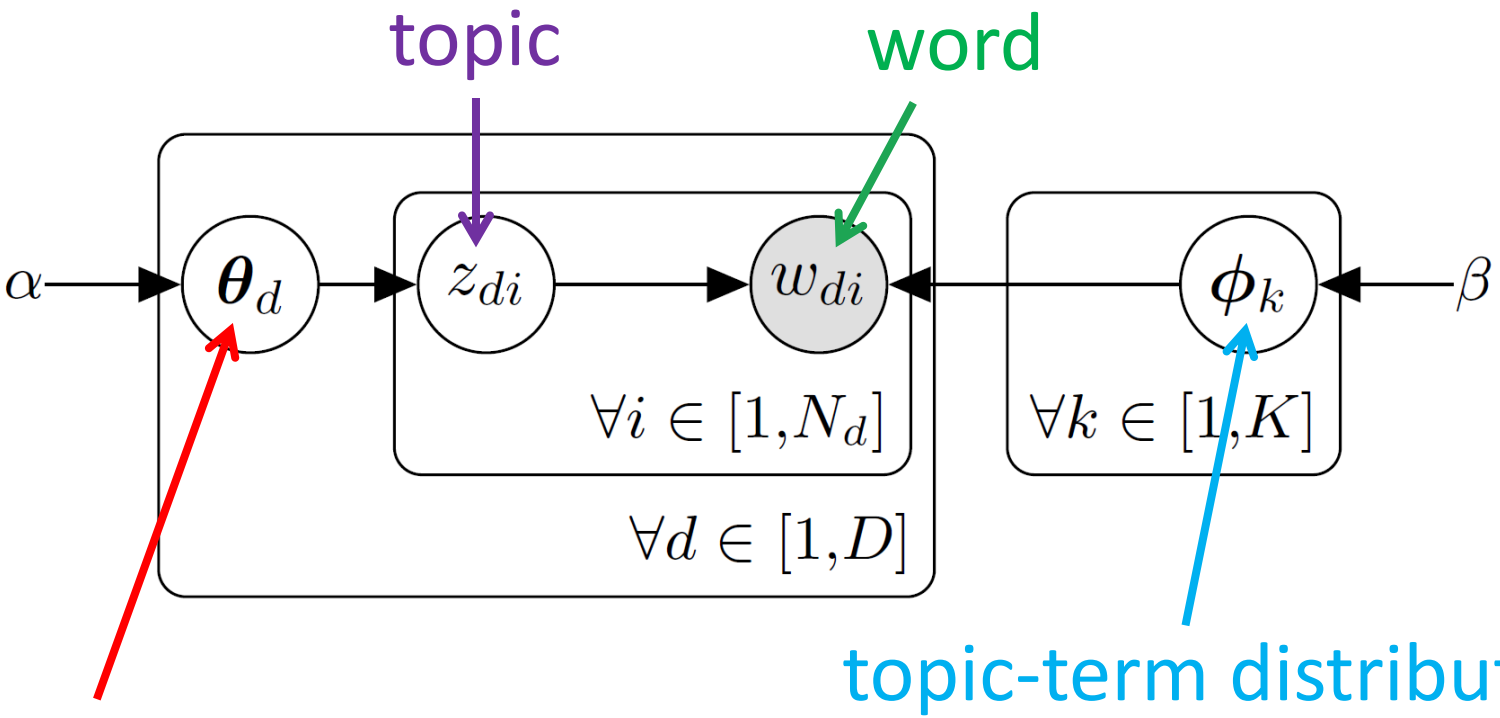
Topic 1	Topic 2	Topic 3
websci	food	sociology
conference	restaurant	social
germany	pizza	society
hannover	eat	behavior
social	steak	relationships
computer	cafe	quantitative

	Topic 1	Topic 2	Topic 3
Doc 1	40%	40%	20%
Doc 2	50%	10%	40%
Doc 3	5%	45%	50%

LDA – Generative Storyline

1. For each document, draw a distribution over topics
2. For each topic, draw a distribution over the vocabulary
3. For each word in each document:
   - Draw a topic
   - Draw a word from this topic

LDA – Plate Notation



Labeled LDA

Labeled LDA

L-LDA is supervised variant of LDA which takes labeled documents as input and creates a topic for each label.

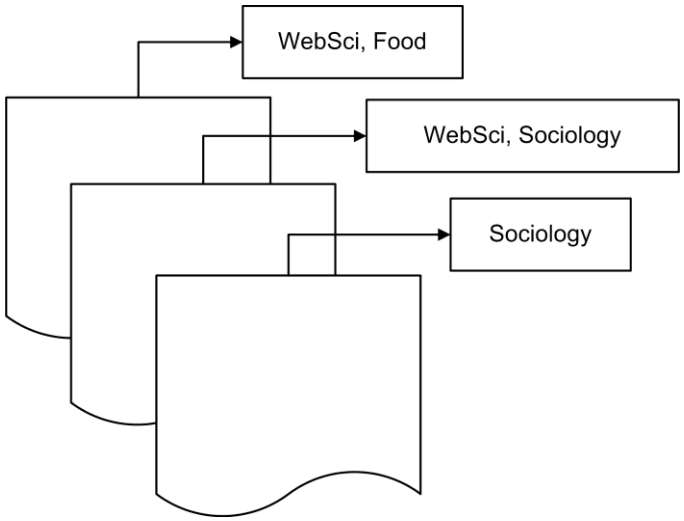
Ramage, Daniel, et al. "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*

Labeled LDA

L-LDA is supervised variant of LDA which takes labeled documents as input and creates a topic for each label.

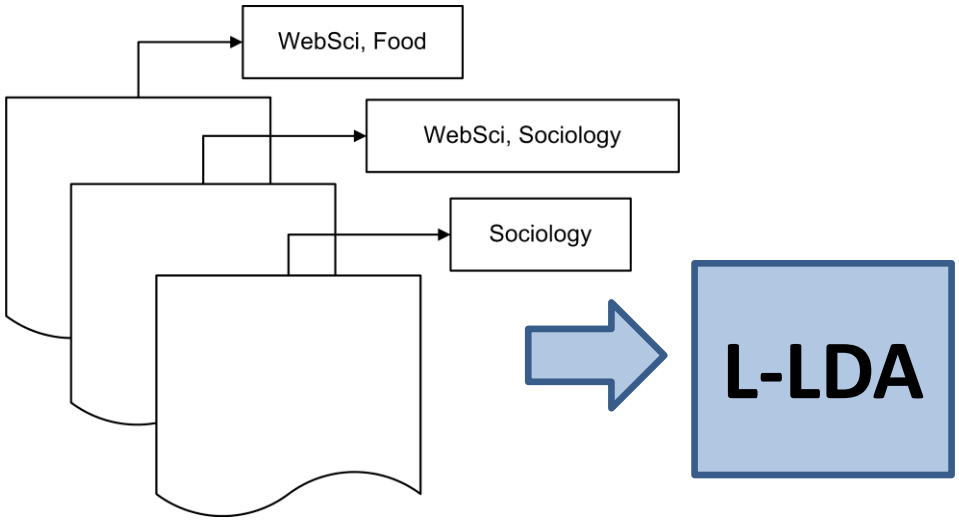
Dataset: text documents with multiple labels

Labeled LDA



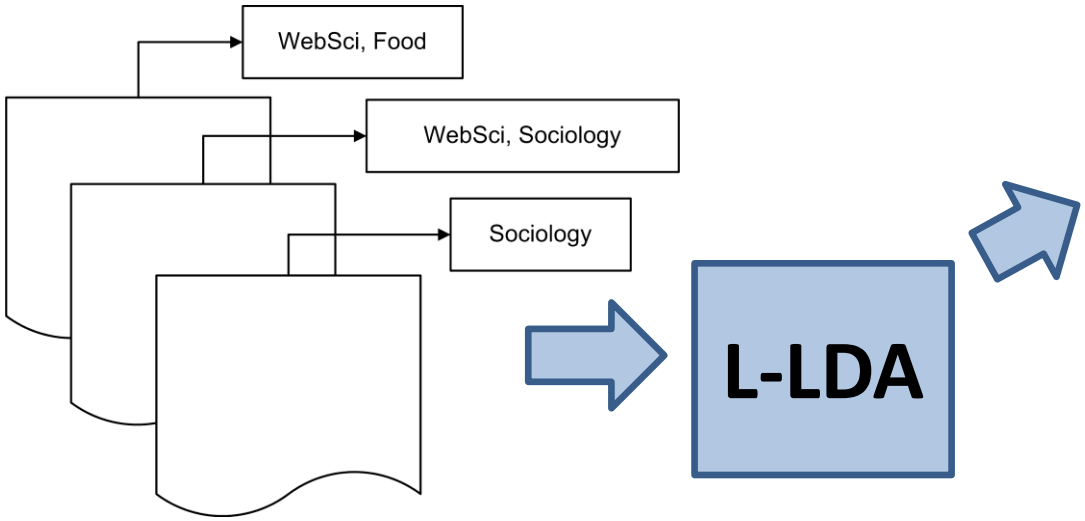
Collection of labeled Documents

Labeled LDA



Collection of labeled Documents

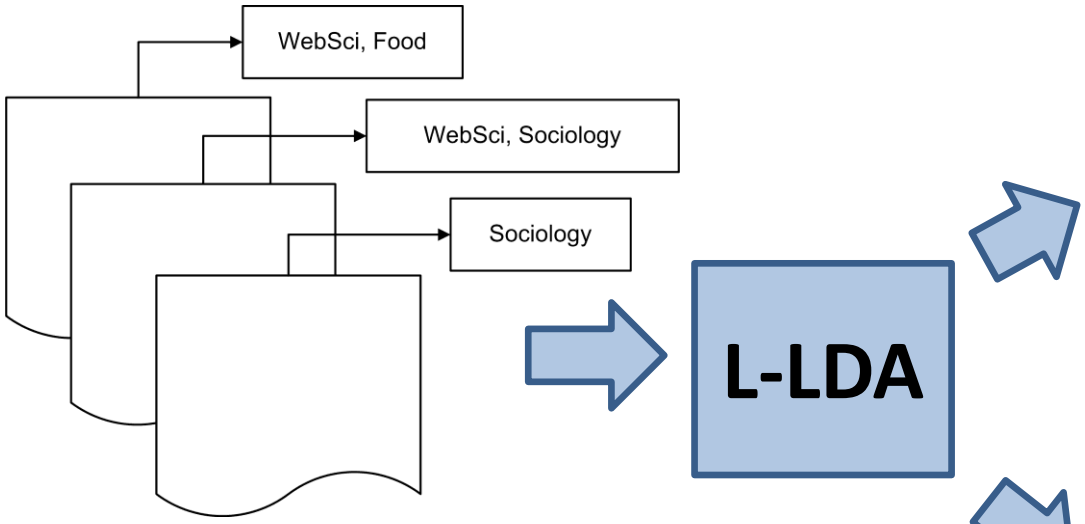
Labeled LDA



WebSci	Food	Sociology
websci	food	sociology
conference	restaurant	social
germany	pizza	society
hannover	eat	behavior
social	steak	relationships
computer	cafe	quantitative

Collection of labeled Documents

Labeled LDA



WebSci	Food	Sociology
websci	food	sociology
conference	restaurant	social
germany	pizza	society
hannover	eat	behavior
social	steak	relationships
computer	cafe	quantitative

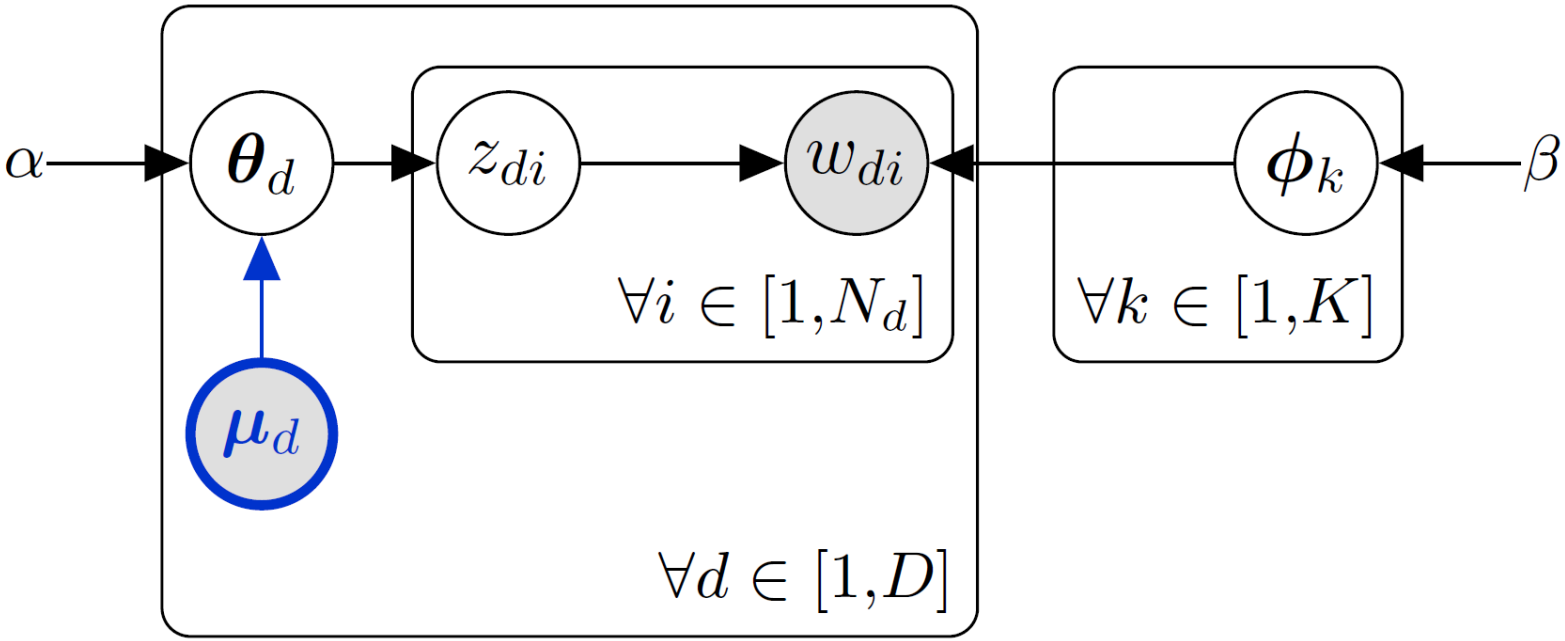
Collection of labeled Documents

	WebSci	Food	Sociology
Doc 1	70%	30%	0%
Doc 2	45%	0%	55%
Doc 3	0%	0%	100%

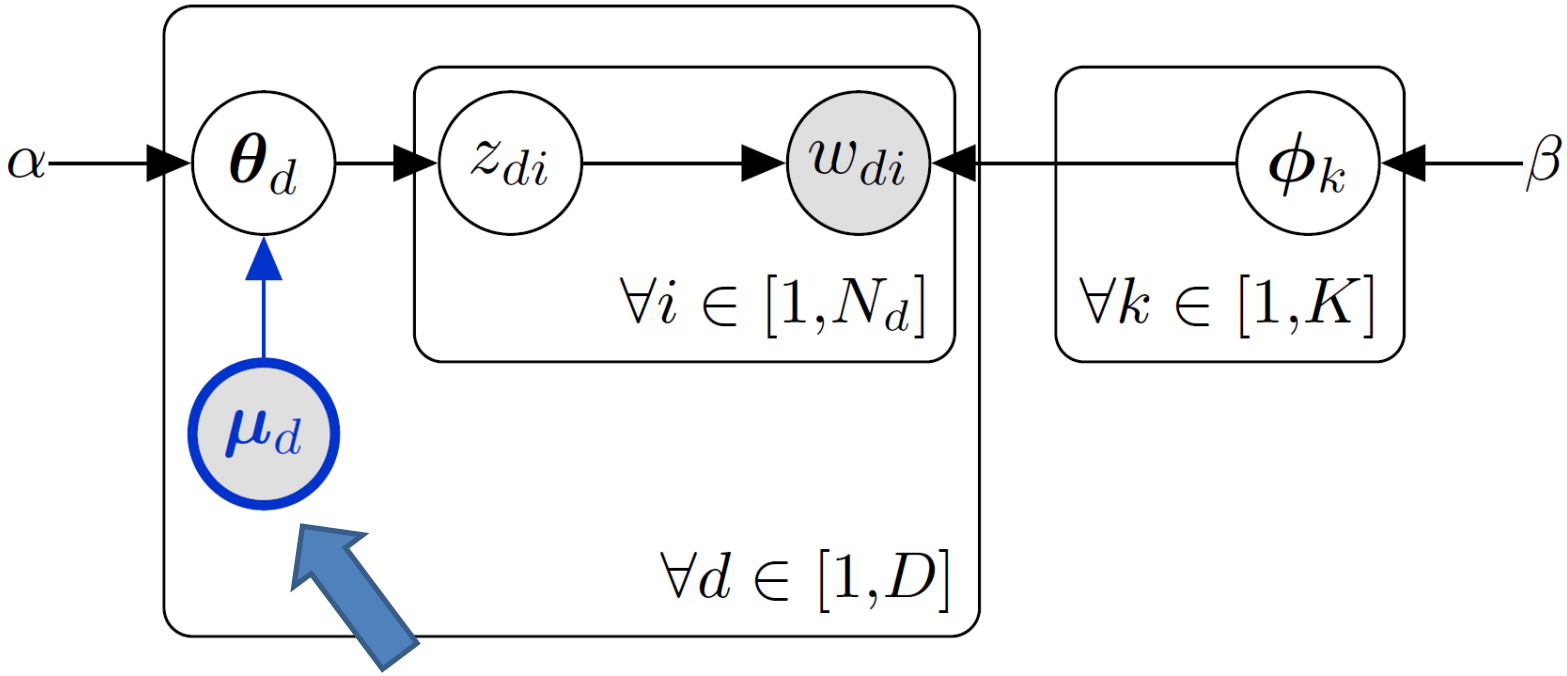
L-LDA – Generative Storyline

1. For each document, draw a distribution over topics, restricted to the document‘s labels
2. For each topic, draw a distribution over the vocabulary
3. For each word in each document:
   - Draw a topic, from the permitted topics
   - Draw a word from this topic

L-LDA – Plate Notation



L-LDA – Plate Notation



topic restriction

L-LDA – Examples

- Publications, labeled with a classification system
  - Create a topic for each class in the classification system
- Tagged Blog entries
  - Create a topic for each tag

Polylingual Topic Model

Polylingual Topic Model

PLTM is a topic model for corpora where the documents are available in several languages.

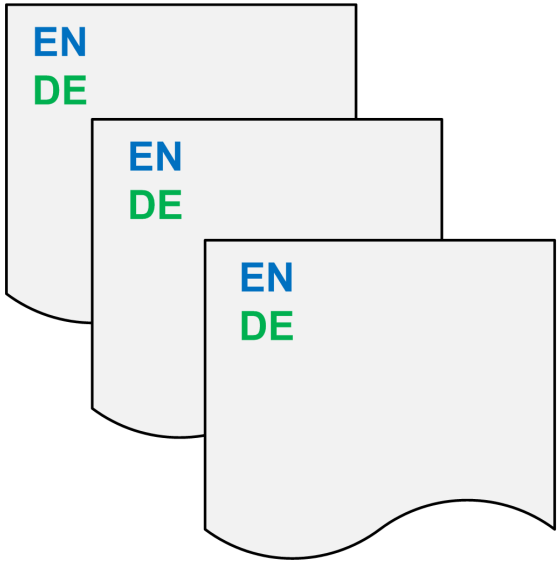
The sets of documents should be loosely equivalent to each other.

Mimno, David, et al. "Polylingual topic models."
Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009.

Polylingual Topic Model

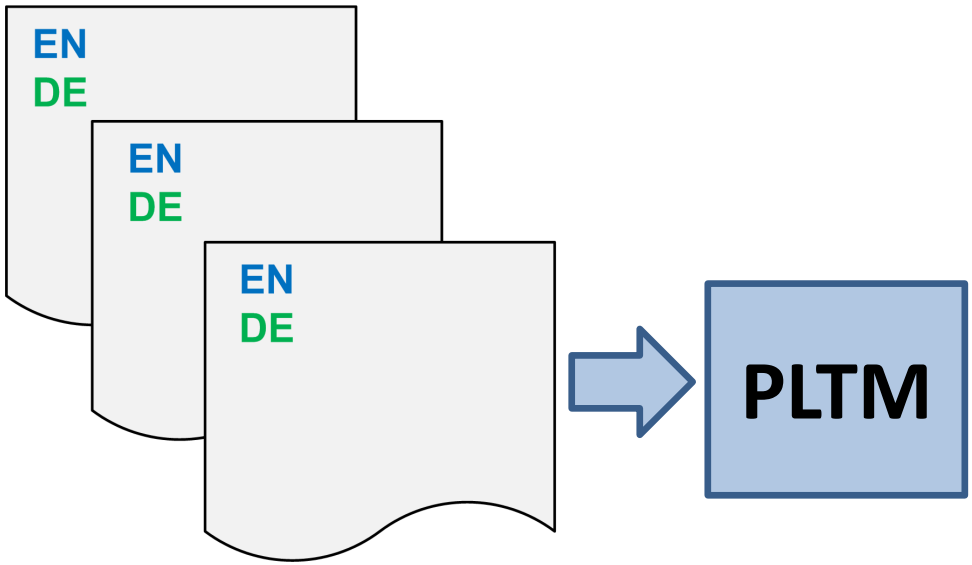
PLTM uses a separate vocabulary for each language, and each topic has a word distribution for each language.

Polylingual Topic Model



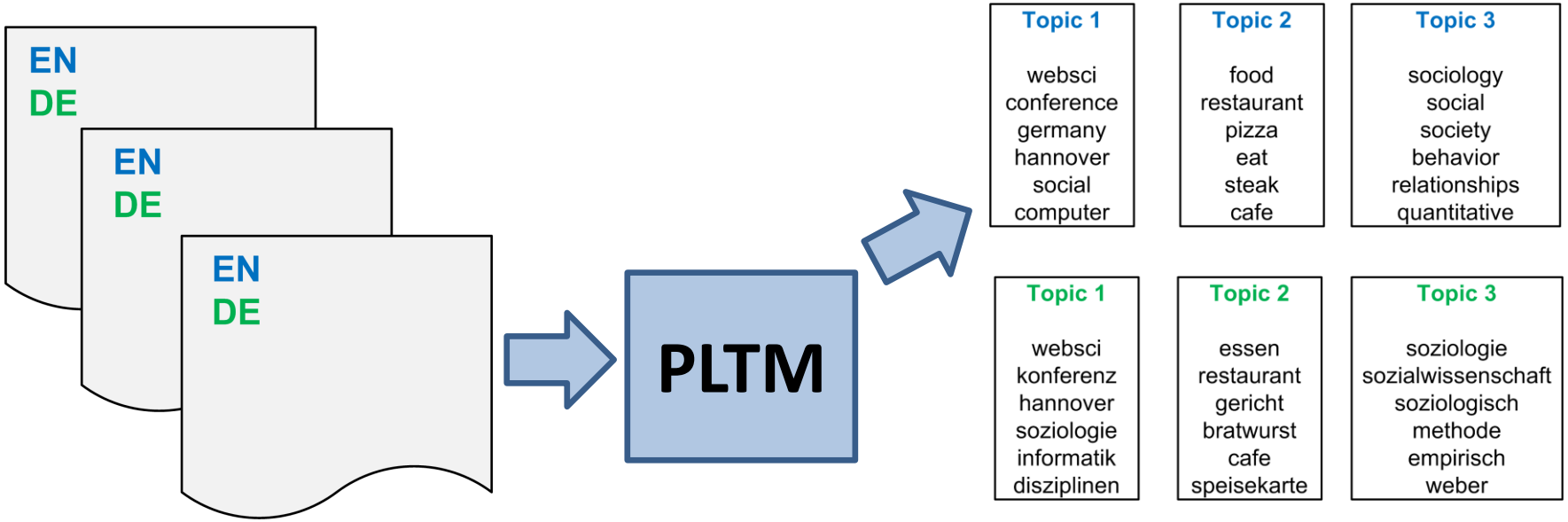
Collection of multilingual Documents

Polylingual Topic Model



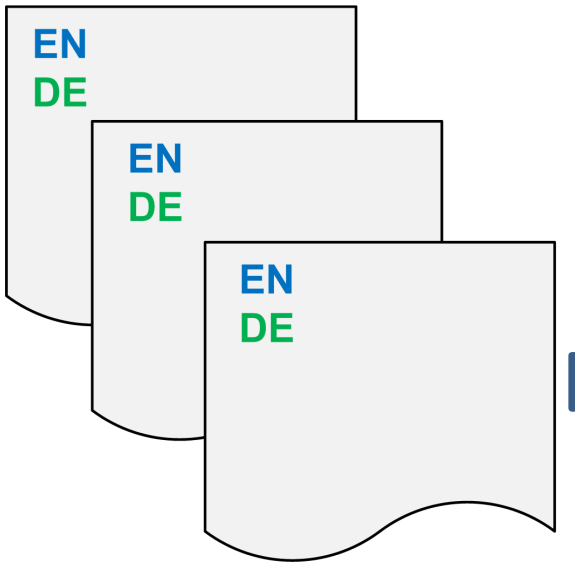
Collection of multilingual Documents

Polylingual Topic Model



Collection of multilingual Documents

Polylingual Topic Model



PLTM

Topic 1	Topic 2	Topic 3
websci	food	sociology
conference	restaurant	social
germany	pizza	society
hannover	eat	behavior
social	steak	relationships
computer	cafe	quantitative

Topic 1	Topic 2	Topic 3
websci	essen	soziologie
konferenz	restaurant	sozialwissenschaft
hannover	gericht	soziologisch
soziologie	bratwurst	methode
informatik	cafe	empirisch
disziplinen	speisekarte	weber

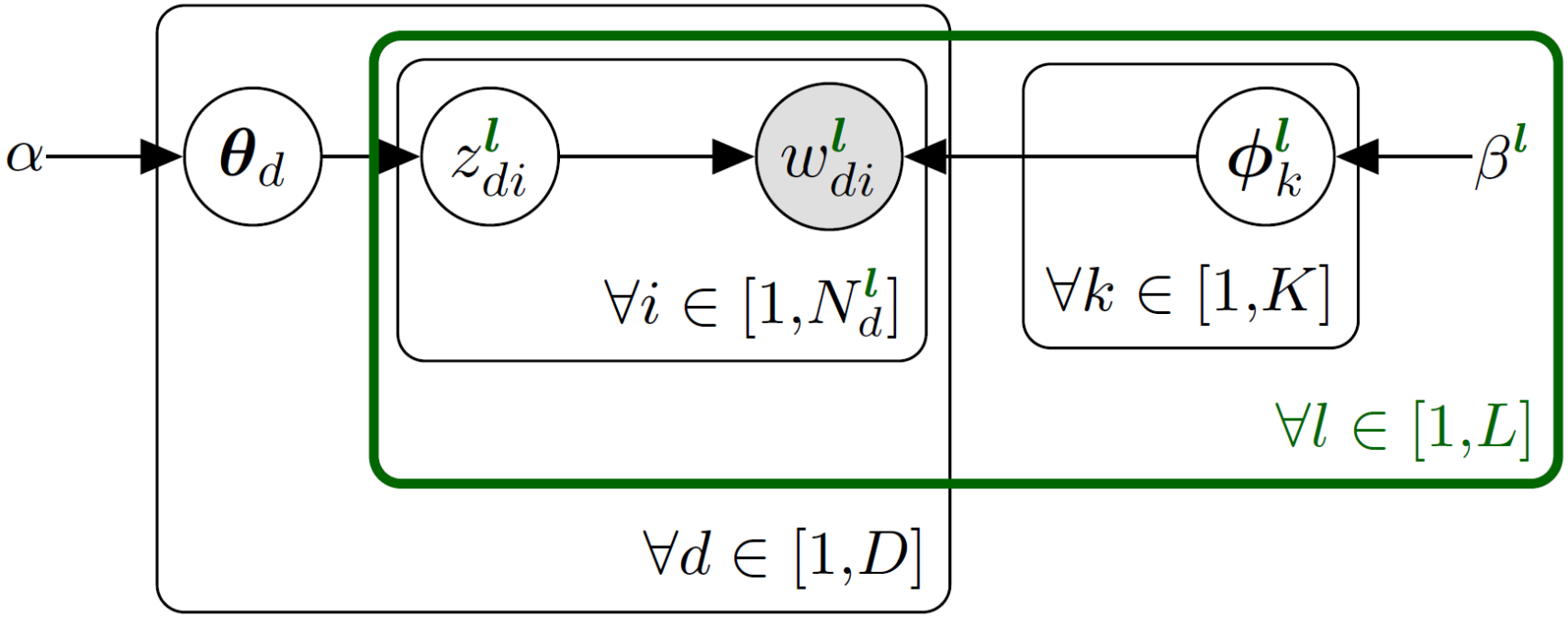
Collection of multilingual Documents

	Topic 1	Topic 2	Topic 3
Doc 1	40%	40%	20%
Doc 2	50%	10%	40%
Doc 3	5%	45%	50%

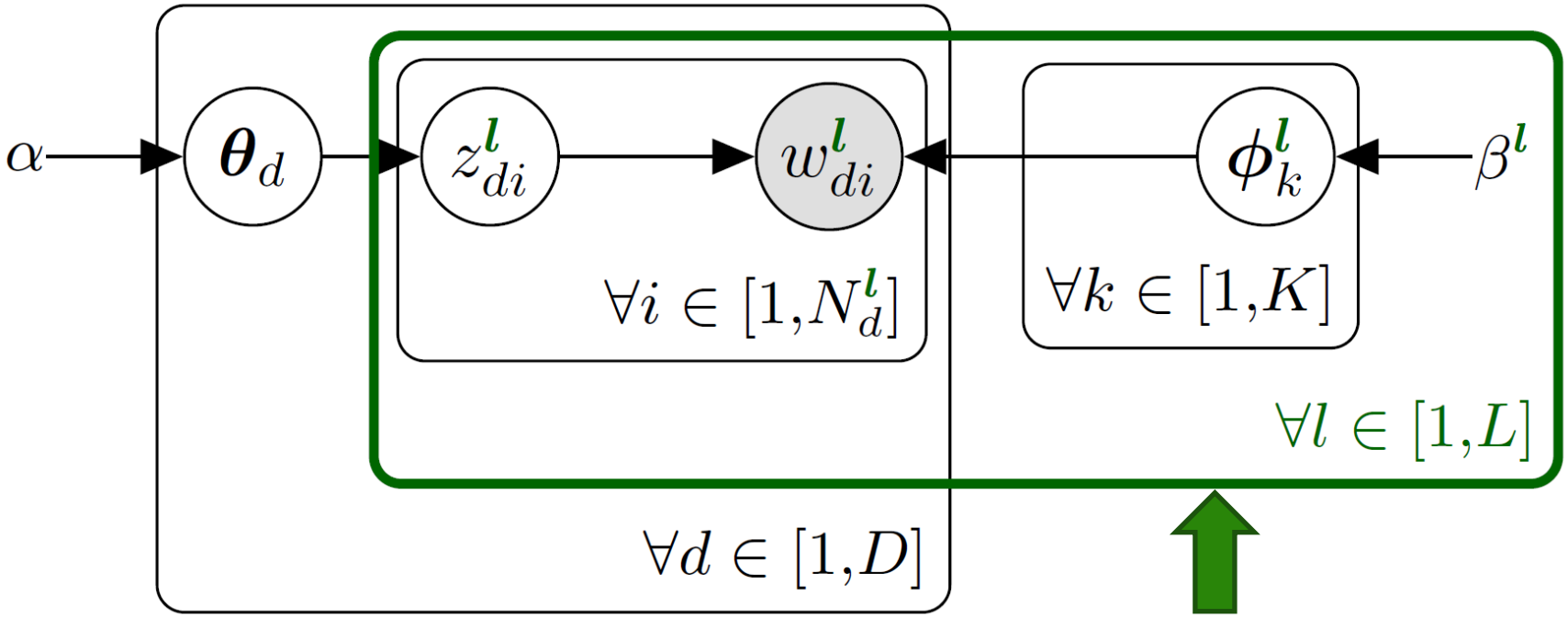
PLTM – Generative Storyline

1. For each document, draw a distribution over topics
2. For each topic, in each language, draw a distribution over the vocabulary of this language
3. For each word in each language in each document:
   - Draw a topic
   - Draw a word from this language-specific topic

PLTM – Plate Notation



PLTM – Plate Notation



multiple languages

PLTM - Examples

- Wikipedia articles in several languages
  - Create topics for each language
- Documents that are annotated with a controlled vocabulary
  - Create topics for both the natural language and the controlled vocabulary

Author-Topic Model

Author-Topic Model

- The Author-Topic model extends LDA to include authorship information.

- Each author has a distribution over topics.

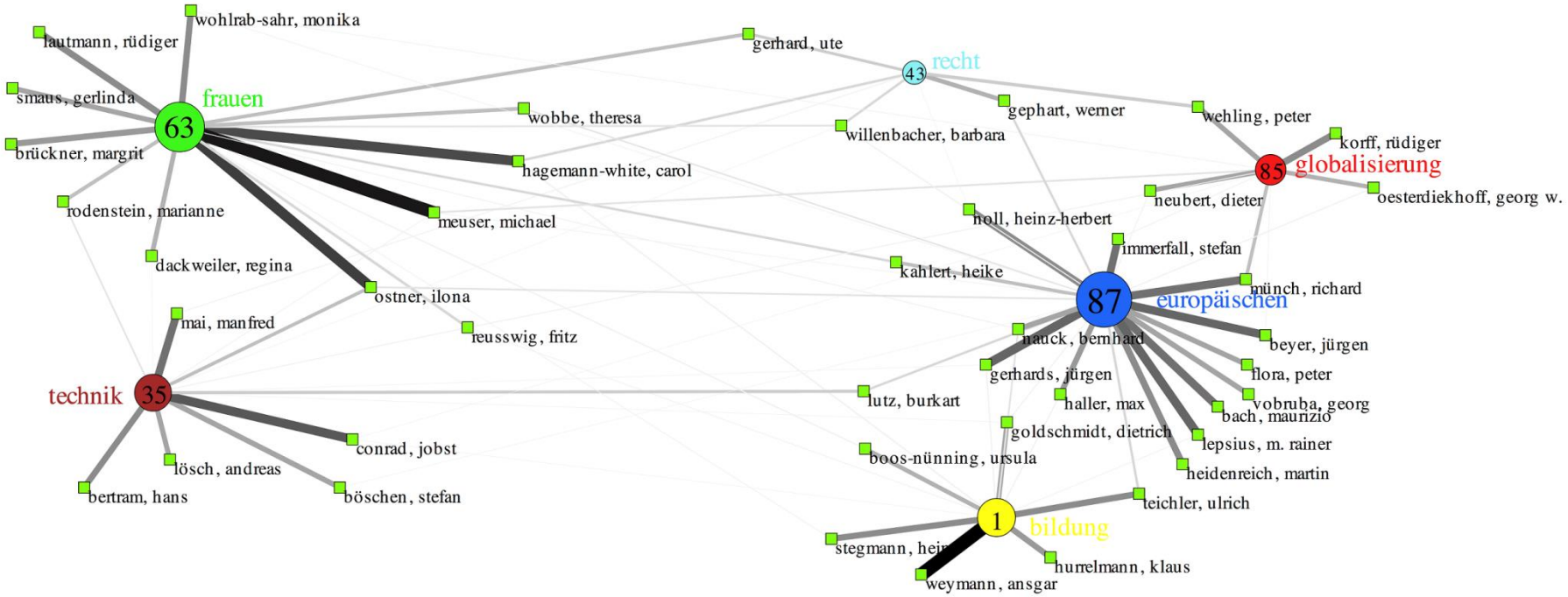
Rosen-Zvi, Michal, et al. "The author-topic model for authors and documents."
Proceedings of the 20th conference on Uncertainty in artificial intelligence. AUAI Press, 2004.

Author-Topic Model

- For each word in a document
  - choose an author,
  - then choose a topic from that author‘s topic distribution and
  - generate a word from that topic.

Rosen-Zvi, Michal, et al. "The author-topic model for authors and documents." *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, 2004.

Author-Topic Model



Bleier, Arnim, and Andreas Strotmann. "Towards an Author-Topic-Term-Model Visualization of 100 Years of German Sociological Society Proceedings."

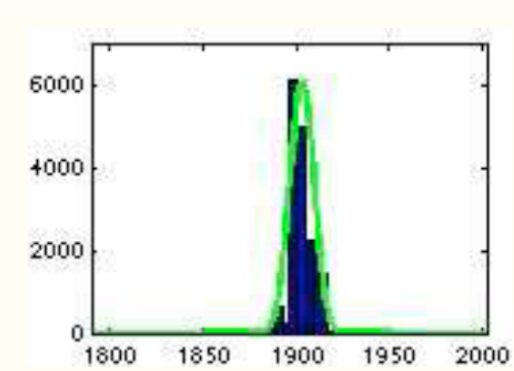
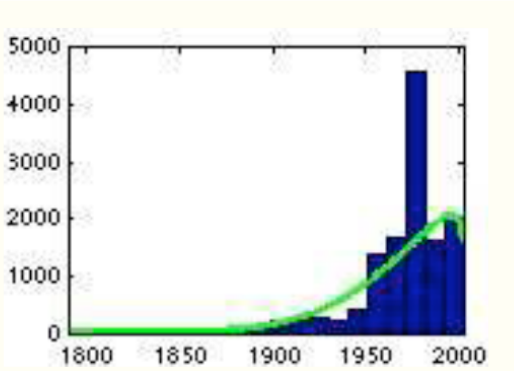
Topics over Time

Topics over Time

- Topics generate both words and observed timestamps.

- Jointly models word co-occurrences and localization in time.

Topics over Time

Panama Canal		Cold War	
			
government	0.02928	world	0.01875
united	0.02132	states	0.01717
states	0.02067	security	0.01710
islands	0.01167	soviet	0.01664
canal	0.01014	united	0.01491
american	0.00872	nuclear	0.01454
cuba	0.00834	peace	0.01408

Wang, Xuerui, and Andrew McCallum. "Topics over time: a non-Markov continuous-time model of topical trends." *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006.

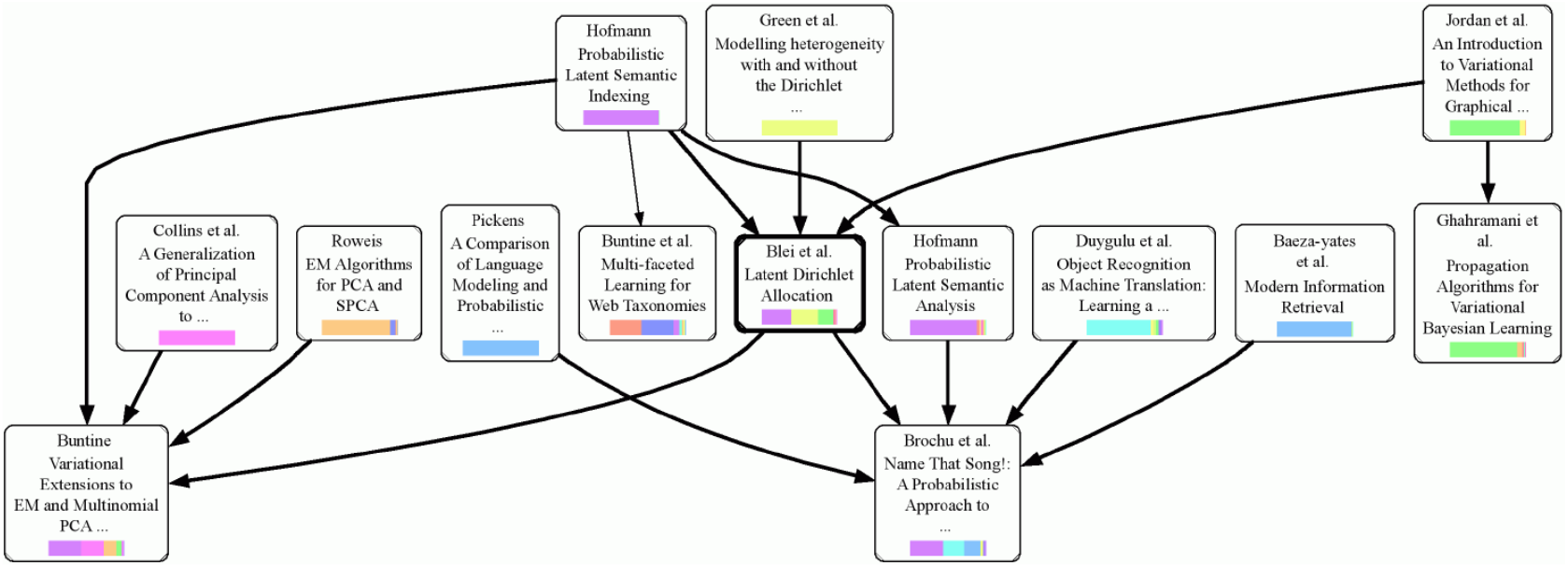
Citation Influence Model

Citation Influence Model

- Estimates the weight of edges in a citation graph, i.e. the strength of influence one publication has on another.
- Incorporates the aspects of topical innovation and topical inheritance via citations.

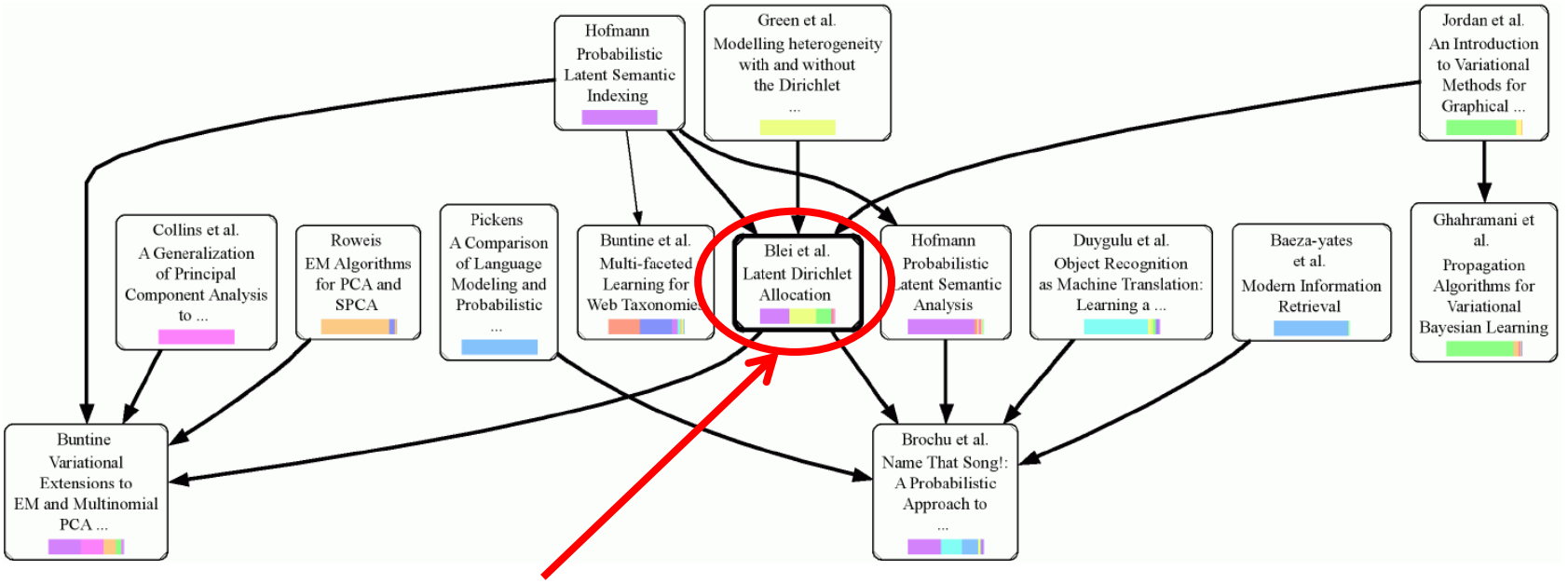
Dietz, Laura, Steffen Bickel, and Tobias Scheffer. "Unsupervised prediction of citation influences." *Proceedings of the 24th international conference on Machine learning*. ACM, 2007.

Citation Influence Model



Dietz, Laura, Steffen Bickel, and Tobias Scheffer. "Unsupervised prediction of citation influences." *Proceedings of the 24th international conference on Machine learning*. ACM, 2007.

Citation Influence Model



original LDA Paper

Summary

- Labeled LDA: for labeled documents
- Polylingual Topic Model: multilingual documents
- Author-Topic Model: authors‘ interests
- Topics over Time: topics‘ localization in time
- Citation Influence Model: strength of influence

Which topic model you want to use depends on your data and on which questions you want to answer.

Questions,

but first….

How could topic models be useful for your research?

Which aspects of datasets would you want to explore with topic models?

… discuss with your neighbor!

What questions do you have? Which things are still unclear to you?

etc. etc. etc

What would you like to know more about?

Thank you!