Beyond Word Clouds

Combining Entities and Topics for Fine-Grained Analyses of Historical (and Political) Texts

Federico Nanni, Hiram Kümper and Simone Paolo Ponzetto
University of Mannheim

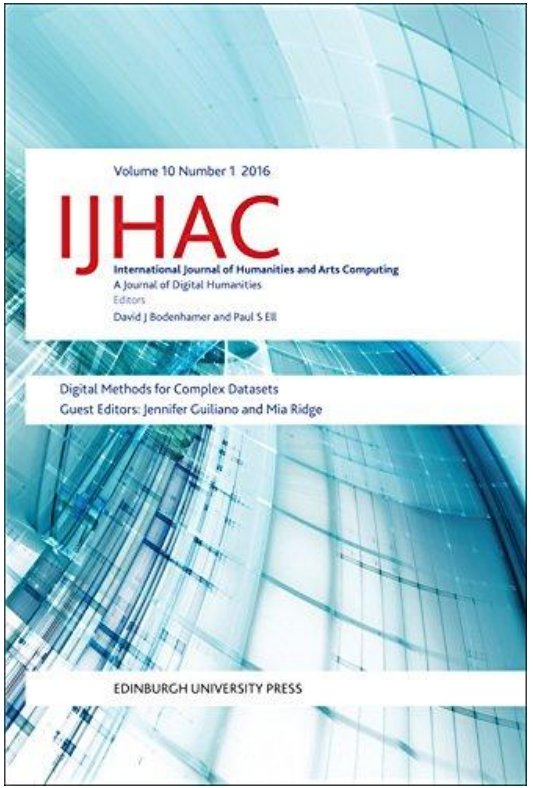
Background

- Bachelor and Master in **Contemporary History**
- Final year PhD student in **Digital Humanities** at UniBo
  - Thesis: “The Web as a Historical Corpus”
- Researcher at **Data and Web Science Group,** UniMannheim

Cfp: special issue of IJHAC

The **future of digital methods** for complex datasets.

Guest editors: Jennifer Giuliano and Mia Ridge



The future of digital history

Historians are dealing with large collections:

The future of digital history

Historians are dealing with large collections:

The future of digital history

Historians are dealing with large collections:

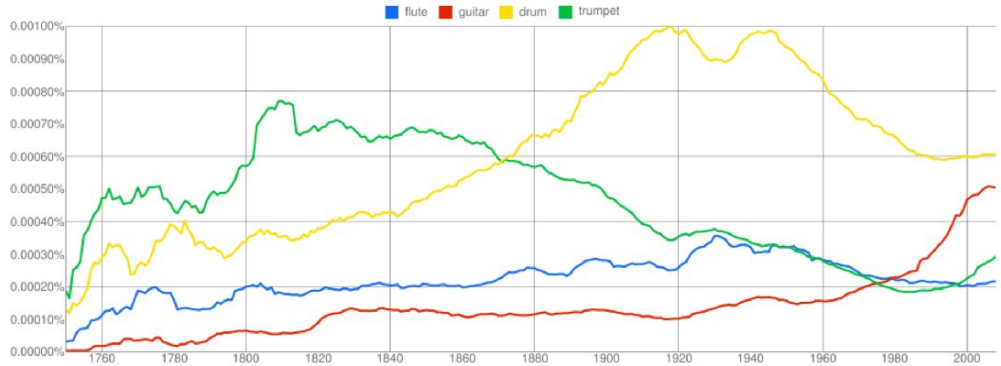
- Text exploration
- Information Retrieval
- Quantification

NLP and digital history

NLP and digital history

Google labs Books Ngram Viewer

Graph these case-sensitive comma-separated phrases: flute,guitar,trumpet,drum
between 1750 and 2008 from the corpus English with smoothing of 5 .
Search lots of books



Search in Google Books:

1750 - 1793	1794 - 1923	1924 - 1940	1941 - 1981	1982 - 2008	flute
1750 - 1825	1826 - 1979	1980 - 1992	1993 - 2001	2002 - 2008	guitar
1750 - 1777	1778 - 1801	1802 - 1817	1818 - 1952	1953 - 2008	trumpet
1750 - 1801	1802 - 1911	1912 - 1927	1928 - 1983	1984 - 2008	drum

Run your own experiment! Raw data is available for download here.

NLP and digital history

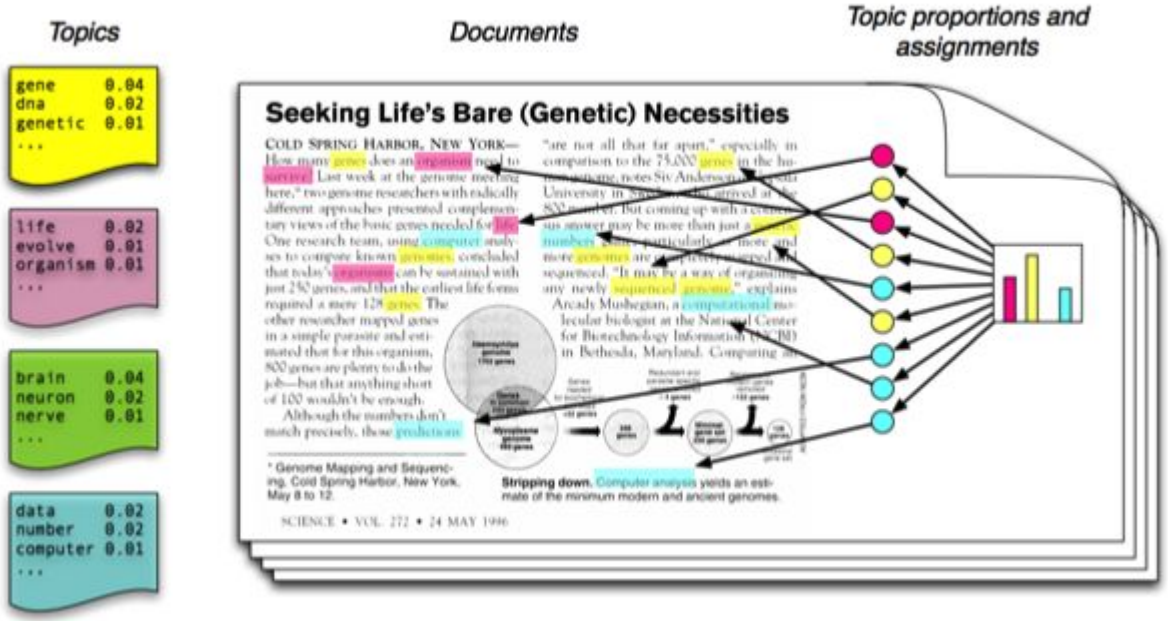
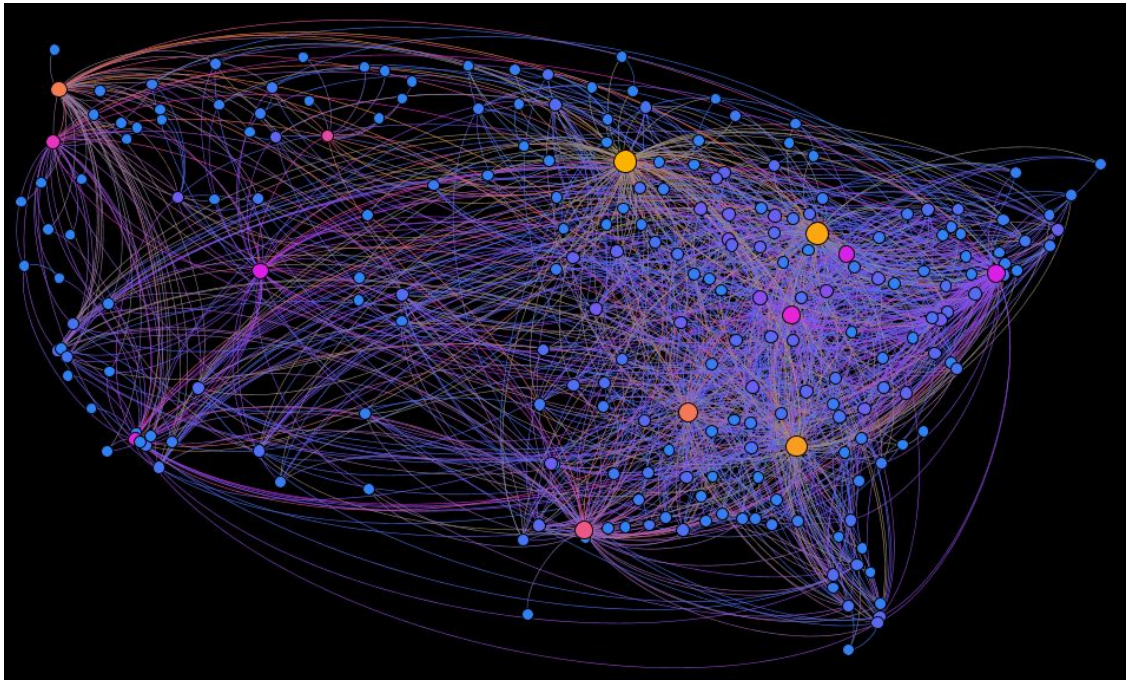
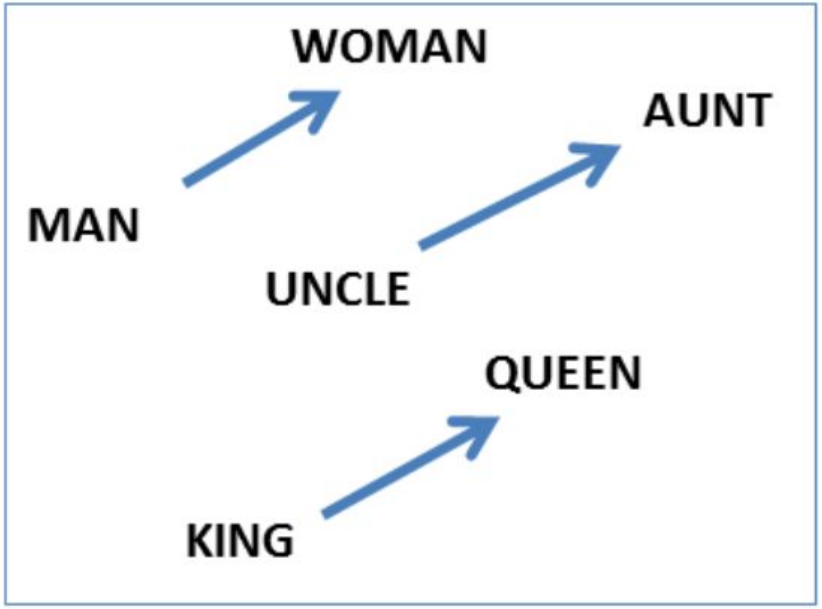


Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

NLP and digital history



NLP and digital history



Current problems

It’s very difficult to:

- Find the **right approach** for a specific question
- Establish its **reliability**
- Move **beyond** text exploration
- **Answer** a humanities research question

Our goal at Data and Web Science Group

Sustain **hypothesis-testing** analyses

Offer both tool implementations and **evaluation platforms**

Train students with experience both in **data science and humanities / social sciences**

Nanni, Kümper and Ponzetto, “Semi-supervised Textual Analysis and Historical Research Helping Each Other”, IJHAC, 2016.

Today's talk

Overview of three researches we recently conducted at DWS:

1) Entities as Topic Labels
2) Building Entity-based Collections of Global Events
3) Topic-Based Analysis of Political Positions

Entities as Topic Labels

Anne Lauscher[1], **Federico Nanni**[1], Pablo Ruiz Fabo[2] and Simone Paolo Ponzetto[1]

[1]Data and Web Science Group, University of Mannheim
[2]LATTICE Lab, École Normale Supérieure

Why topic models are awesome

They are able to identify the most important topics in a collection of documents.

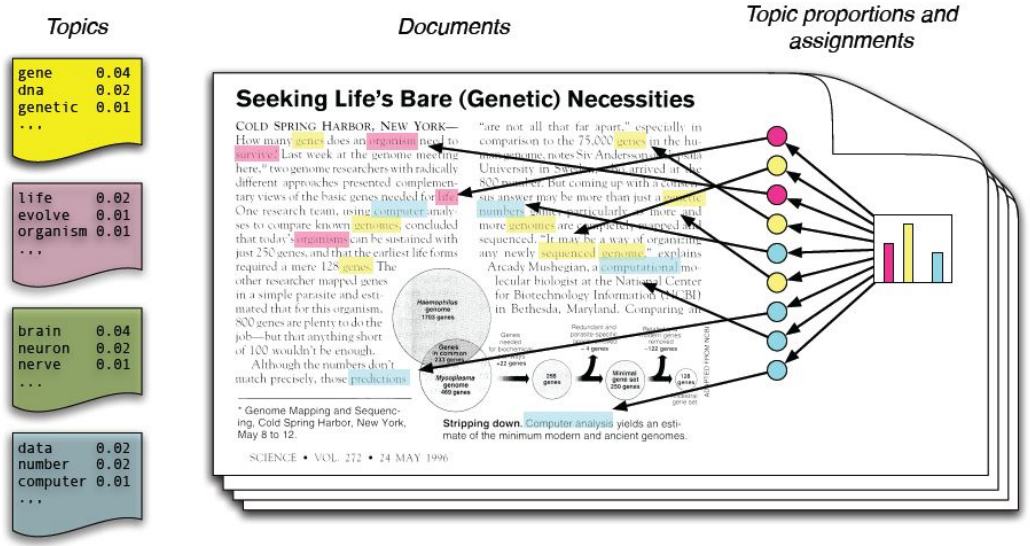


Figure from Blei (2012)

Why topic models are NOT awesome

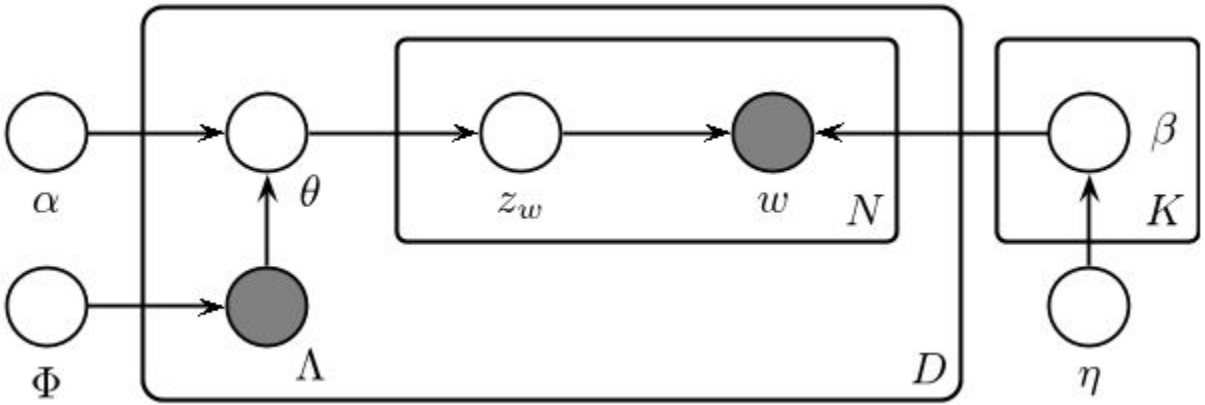
The topics obtained are difficult to interpret and evaluate.

Why topic models are NOT awesome

The topics obtained are difficult to interpret and evaluate.

Labeled LDA

Each **document** is described with one or more **labels**, each label is associated with a specific **topic** (Ramage et al., 2009).

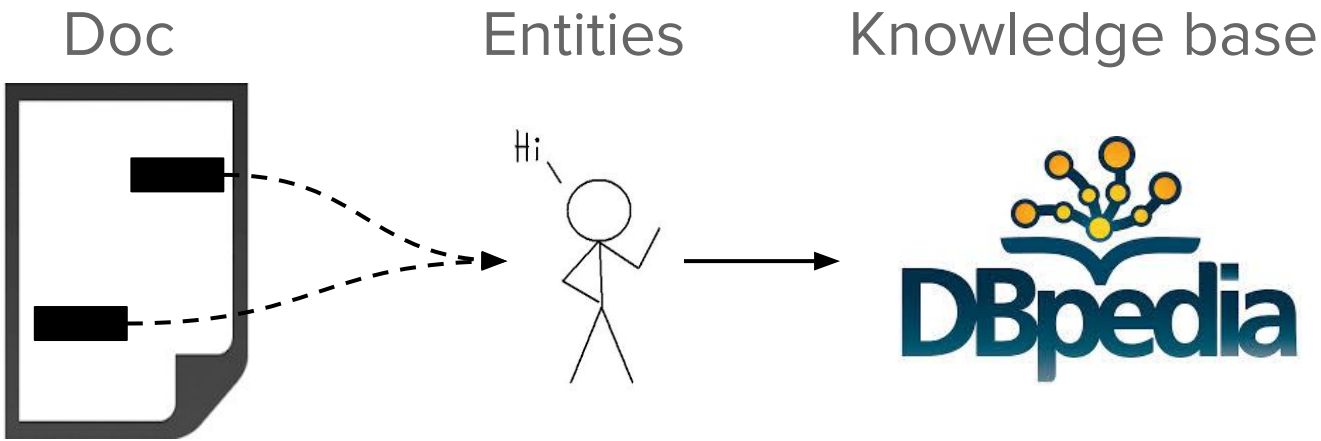


How to automatically obtain labels?

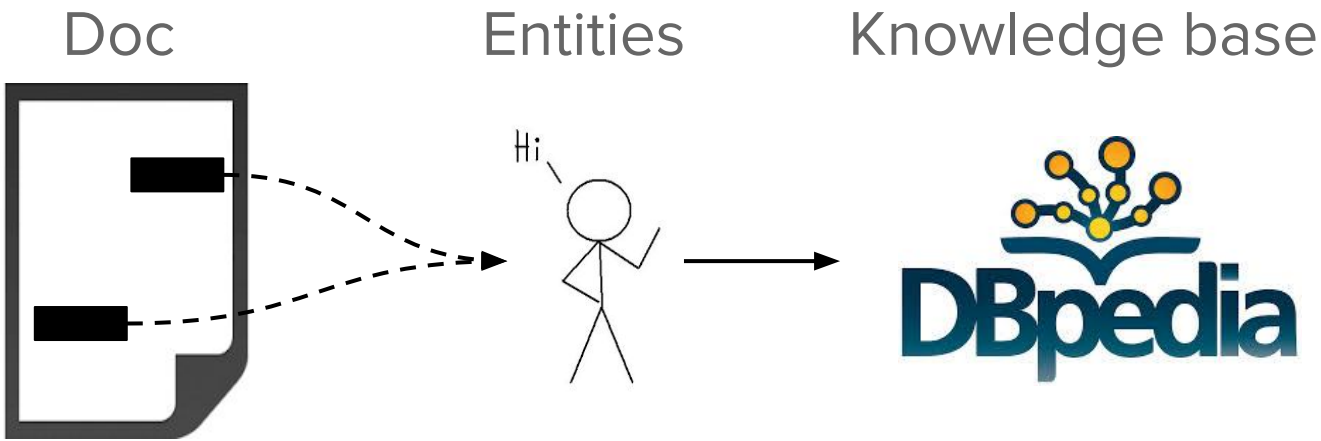
Different approaches:

- Keyphrase digger (FBK - Trento)
- Labeling the obtained topics (Hulpus et al., 2014)

Our approach: entities



Our approach: entities



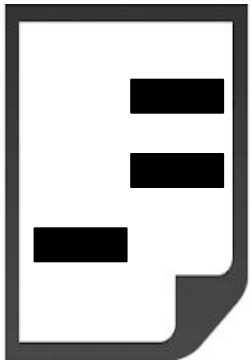
Babelfy "Word sense disambiguation and entity linking together!"

TAG·ME

DBpedia Spotlight

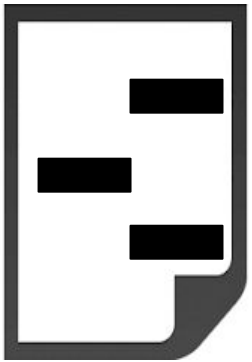
Our approach: entity ranking

Doc1



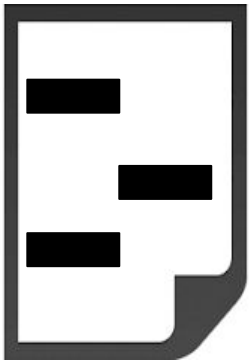
Entity1
Entity2
Entity3

Doc2



Entity1
Entity2
Entity3

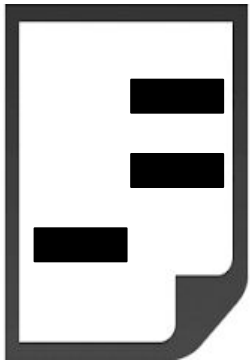
Doc3



Entity1
Entity2
Entity3

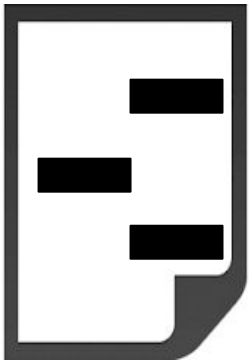
Our approach: entities as topic labels

Doc1



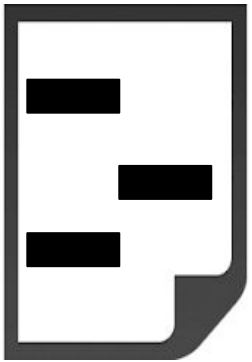
Entity1
Entity2
Entity3

Doc2



Entity1
Entity2
Entity3

Doc3



Entity1
Entity2
Entity3

Labeled LDA!

Different case-study

Transcripts from **European Parliament**'s fifth term (1999-2004).

Threads in the **Enron Corpus** (600.000 emails, 158 employees).

Discussions in the **Hillary Clinton Email Dataset**, a collection of redacted versions of emails (available on Kaggle).

Europarl corpus

Examined most relevant topics addressed by each party in the European Parliament's fifth term (1999-2004).

Les Verts (France)

	Label: Consumer (47%)	**Label: GMO** (34%)
Topic words	product directive consumer safety law market	human health agreement food measure sustainable

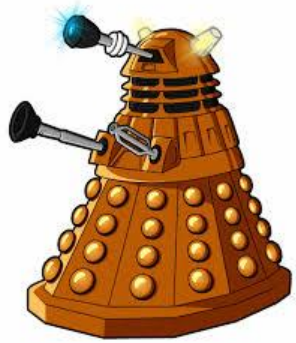
Europarl corpus

Examined most relevant topics addressed by each party in the European Parliament's fifth term (1999-2004).

Les Verts (France)

	Label: Consumer (47%)	**Label: GMO** (34%)
Topic words	product directive consumer safety law market	human health agreement food measure sustainable

EVALUATE!
EVALUATE!
EVALUATE!



How to evaluate it?

1. Label selection

Les Verts (France)

Consumer **GMO**

How to evaluate it?

1. Label selection
2. Label ranking

Les Verts (France)

Consumer (**47%**) GMO (**34%**)

How to evaluate it?

1. Label selection
2. Label ranking
3. Label-topic relation

Les Verts (France)

Consumer (47%)	GMO (34%)
product	human
directive	health
consumer	agreement
safety	food
law	measure
market	sustainable

How to evaluate it?

1. Label selection
2. Label ranking
3. Label-topic relation

Following slides introduce
the ongoing master thesis work
of **Anne Lauscher** (Univ. Mannheim)

Les Verts (France)

Consumer (47%)	GMO (34%)
product	human
directive	health
consumer	agreement
safety	food
law	measure
market	sustainable

Label selection and ranking

Country Weather
Limiting magnitude
Marginal cost
Economic growth
Brand management
Barriers to entry
Fax
Three Mile Island accident
George W. Bush
Cambridge Energy Research Associates
Macroeconomics
Customer relationship management
Loving Every Minute (album)
Wave function
Drought

Is the country facing sustained tight markets?:
In general, capacity margins have been falling over the last few years as utilities have refraind from building baseload capacity, and others have focused on developing primarily peaking capacity. The last large building boom of coal plants ended in early 1970's and nuclear boom dropped sharply after Three Mile Island incident in 1979. During 1990's, projected capacity margins have fallen from the 15 - 20 % toward 10% and below. Outside of the California situation, NYC also poses a potential risk for this summer.
If general, however, by 2002 - 2003, the amount of capacity proposed in each region more than covers normal load growth for meeting peak hour demand. Remaining question relates to performance of existing coal and nuclear stacks, also what happens during periods of persistent drought.

Will voaltility and prices remain high or lessen?:
To the extent that more capacity becomes merchant oriented focusing in marginal cost economics, and transmission congestion persists, increased volatility will continue, especially in ISO/pool type environments. Prices will reflect primary fuel dynamics, especially the interplay between gas and oil. The case for lower prices will reflect an overbuild scenario beyond this year, coupled with slowing economic activity and low incidence of extreme weather events.

How fast can generation be added and what returns should be expected?
The variability in development of greenfield capacity depends on time to permit at state or local levels. Construction time is fairly constant. In general 18 months is reasonable time frame from concept to first fire. To the extent that power plants are project financed with minimum 30 equity, returns should be consistent with other comparable project financed opportunities available to fund managers. We do not expect any more fully debt financed facilities in the near term.

Can companies make major profits owning generation long term?

Label selection and ranking

Country Weather
Limiting magnitude
Marginal cost
Economic growth
Brand management
Barriers to entry
Fax
Three Mile Island accident
George W. Bush
Cambridge Energy Research Associates
Macroeconomics
Customer relationship management
Loving Every Minute (album)
Wave function
Drought

Is the country facing sustained tight markets?:
In general, capacity margins have been falling over the last few years as utilities have refraind from building baseload capacity, and others have focused on developing primarily peaking capacity. The last large building boom of coal plants ended in early 1970's and nuclear boom dropped sharply after Three Mile Island incident in 1979. During 1990's, projected capacity margins have fallen from the 15 - 20 % toward 10% and below. Outside of the California situation, NYC also poses a potential risk for this summer.
If general, however, by 2002 - 2003, the amount of capacity proposed in each region more than covers normal load growth for meeting peak hour demand. Remaining question relates to performance of existing coal and nuclear stacks, also what happens during periods of persistent drought.

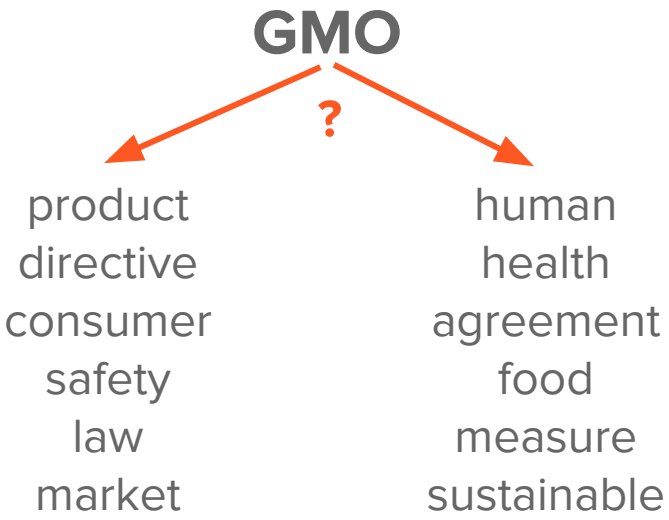
Will voaltility and prices remain high or lessen?:
To the extent that more capacity becomes merchant oriented focusing in marginal cost economics, and transmission congestion persists, increased volatility will continue, especially in ISO/pool type environments. Prices will reflect primary fuel dynamics, especially the interplay between gas and oil. The case for lower prices will reflect an overbuild scenario beyond this year, coupled with slowing economic activity and low incidence of extreme weather events.

How fast can generation be added and what returns should be expected?
The variability in development of greenfield capacity depends on time to permit at state or local levels. Construction time is fairly constant. In general 18 months is reasonable time frame from concept to first fire. To the extent that power plants are project financed with minimum 30 equity, returns should be consistent with other comparable project financed opportunities available to fund managers. We do not expect any more fully debt financed facilities in the near term.

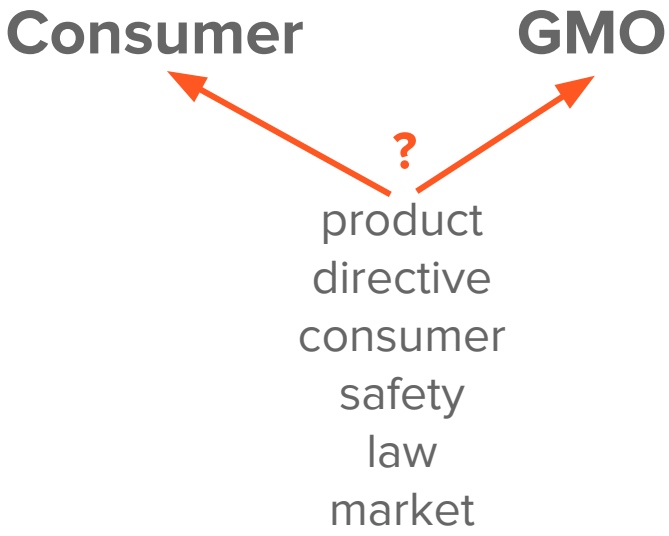
Can companies make major profits owning generation long term?

Mean Average Precision

Topic-label relation



Topic-label relation



First results - 50 docs labeled

	Avg number of label selected	Recall on user selection	Docs with 0 annotations
EuroParl	4	0.88	2
EnronCorpus	4	0.91	7
ClintonCorpus	4	0.95	1

First results - ranking

	MAP TF-IDF Ranking	MAP LLDA Ranking
RandomBaseline	0.30	0.30
EuroParl	0.51	**0.54**
EnronCorpus	0.40	**0.41**
ClintonCorpus	0.48	**0.52**

Conclusion

Entity-labels could **drastically improve** topic interpretability.

However it is necessary to **always evaluate** them.

We will release:

- The pipeline for labeling topics with entities
- A tool for evaluating each step of the process

UNIVERSITÄT MANNHEIM

Building Entity-based Collections of Global Events

Federico Nanni, Simone Paolo Ponzetto and Laura Dietz

Data and Web Science Group
University of Mannheim

{federico,simone,dietz}@informatik.uni-mannheim.de

Global events


Santa Ana Register

FINAL EDITION

BILLIONS LOST AS STOCKS CRASH

Pantages In Jail Awaits Sentence To State Prison

GUILTY SAYS JURY AFTER MANY HOURS

SENATE "FARM BLOC" STARTS BATTLE FOR HIGHER TARIFFS

STATE PLANS TO ROUND UP TAX DODGERS

BINGHAM TAKES EXCEPTION TO SENATE LOBBY

Alliance Of Two Factions May Win Out

JURY BELIEVES HER STORY

MAINSTAYS OF MARKET IN PLUNGE

THREE U. S. STUDENTS ARE CRASH VICTIMS

CHRISTIANITY MUST DESTROY WAR OR WAR WILL OVERTHROW CHURCH, SAYS PASTOR SCHROCK

Global events


Santa Ana Register

FINAL EDITION

BILLIONS LOST AS STOCKS CRASH

Pantages In Jail Awaits Sentence To State Prison

GUILTY SAYS JURY AFTER MANY HOURS

SENATE "FARM BLOC" STARTS BATTLE FOR HIGHER TARIFFS

STATE PLANS TO ROUND UP TAX DODGERS

BINGHAM TAKES EXCEPTION TO SENATE LOBBY

Alliance Of Two Factions May Win Out

JURY BELIEVES HER STORY

MAINSTAYS OF MARKET IN PLUNGE

THREE U. S. C. STUDENTS ARE CRASH VICTIMS

CHRISTIANITY MUST DESTROY WAR OR WAR WILL OVERTHROW CHURCH, SAYS PASTOR SCHROCK

Identified by a common name: the **Wall Street Crash of 1929**.

Also known as: the **Black Tuesday**, the **Great Crash**, or the **Stock Market Crash**.

Global events

Santa Ana Register

FINAL EDITION

BILLIONS LOST AS STOCKS CRASH

Pantages In Jail Awaits Sentence To State Prison

GUILTY SAYS JURY AFTER MANY HOURS

SENATE "FARM BLOC" STARTS BATTLE FOR HIGHER TARIFFS

JURY BELIEVES HER STORY

MAINSTAYS OF MARKET IN PLUNGE

STATE PLANS TO ROUND UP TAX DODGERS

BINGHAM TAKES EXCEPTION TO SENATE LOBBY

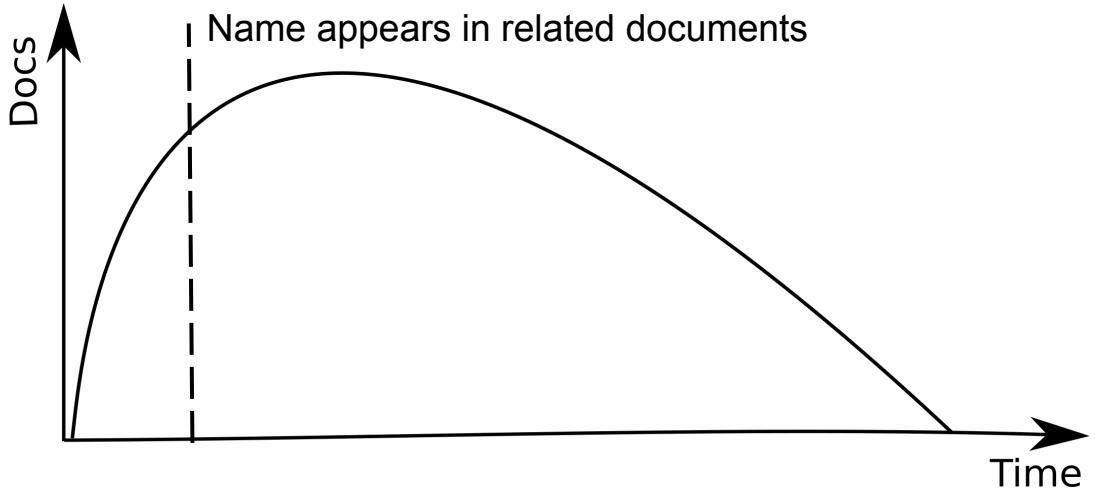
Alliance Of Two Factions May Win Out

THREE U. S. C. STUDENTS ARE CRASH VICTIMS

CHRISTIANITY MUST DESTROY WAR OR WAR WILL OVERTHROW CHURCH, SAYS PASTOR SCHROCK

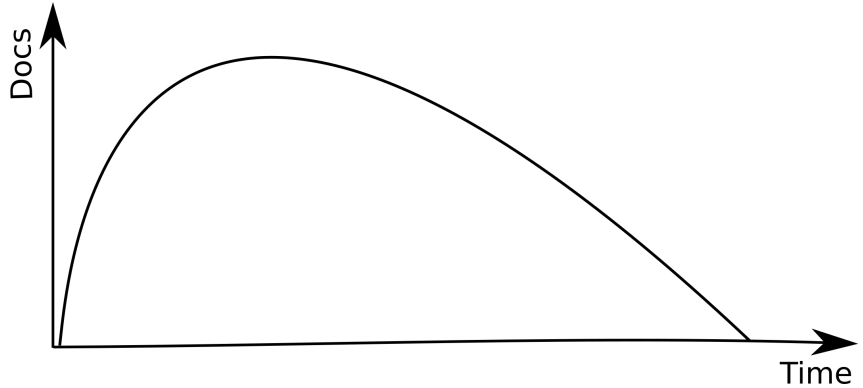
Identified by a common name: the **Wall Street Crash of 1929**.

Also known as: the **Black Tuesday**, the **Great Crash**, or the **Stock Market Crash**.



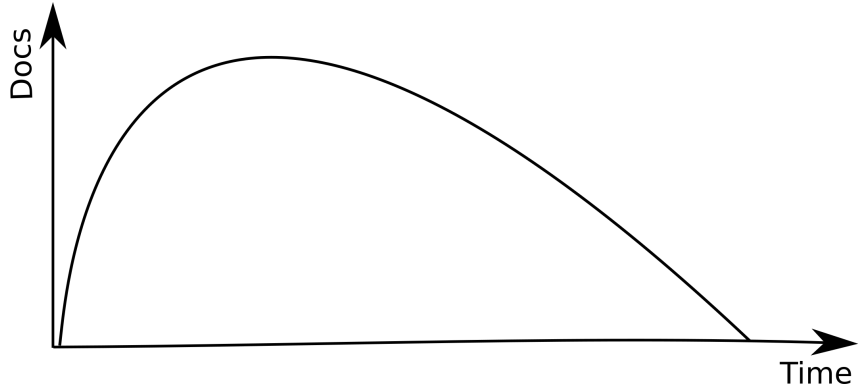
Retrospective analyses

Studying how **events are perceived by society** is a fundamental research task for humanists and social scientists.



Retrospective analyses

Studying how **events are perceived by society** is a fundamental research task for humanists and social scientists.



HATHI TRUST

europeana newspapers

INTERNET ARCHIVE

Solutions: doc filtering, focused crawling

Collect the documents that contain **all query words** (e.g. Kedzie et al., 2014).

E.g. query: “wall street crash tuesday stock market 1929”

The FIRST with the LATEST
Full United Press Leased Wire

People's Paper — Orange County Daily Evening — Santa Ana Register

FINAL EDITION

VOL. XXIV. NO. 285. SANTA ANA, CALIFORNIA, MONDAY, OCTOBER 28, 1929 20 PAGES 3c Per Copy. 65c Per Month

BILLIONS LOST AS STOCKS CRASH

Pantages In Jail Awaits Sentence To State Prison

GUILTY SAYS JURY AFTER MANY HOURS

'Theater Man Declares He 'Got Raw Deal' and Did Not Even Have Chance

MISS PRINGLE GLAD

In Statement Says Not Sure Whether She Will Continue Her Stage Career

LOS ANGELES, Oct. 28.—(UP)—

SENATE "FARM BLOC" STARTS BATTLE FOR HIGHER TARIFFS

STATE PLANS TO ROUND UP TAX DODGERS

Banks and Corporations to Be Object of Campaign Says Commissioner

SACRAMENTO, Oct. 28.—(UP)

Friend Brings Man To Jail On Rum Charge

BINGHAM TAKES EXCEPTION TO SENATE LOBBY

Charges Committee With Unfair Political Tactics During Probe

WASHINGTON, Oct. 28.—(UP)—

Alliance Of Two Factions May Win Out

Western Republicans Group With Southern Democrats to Increase Rates

WASHINGTON, Oct. 28.—

THREE U. S. C STUDENTS ARE CRASH VICTIMS

Auto Plunges Into Engine As Rooters Return from Big Football Game

TULARE, Calif., Oct. 26.—(UP)—

JURY BELIEVES HER STORY

CHRISTIANITY MUST DESTROY WAR OR WAR WILL OVERTHROW CHURCH, SAYS PASTOR SCHROCK

"WAR is a problem in arithmetic," said the Rev. Perry Frederick Schrock, in his sermon at the First Congregational church last night. His subject grew out of a

MAINSTAYS OF MARKET IN PLUNGE

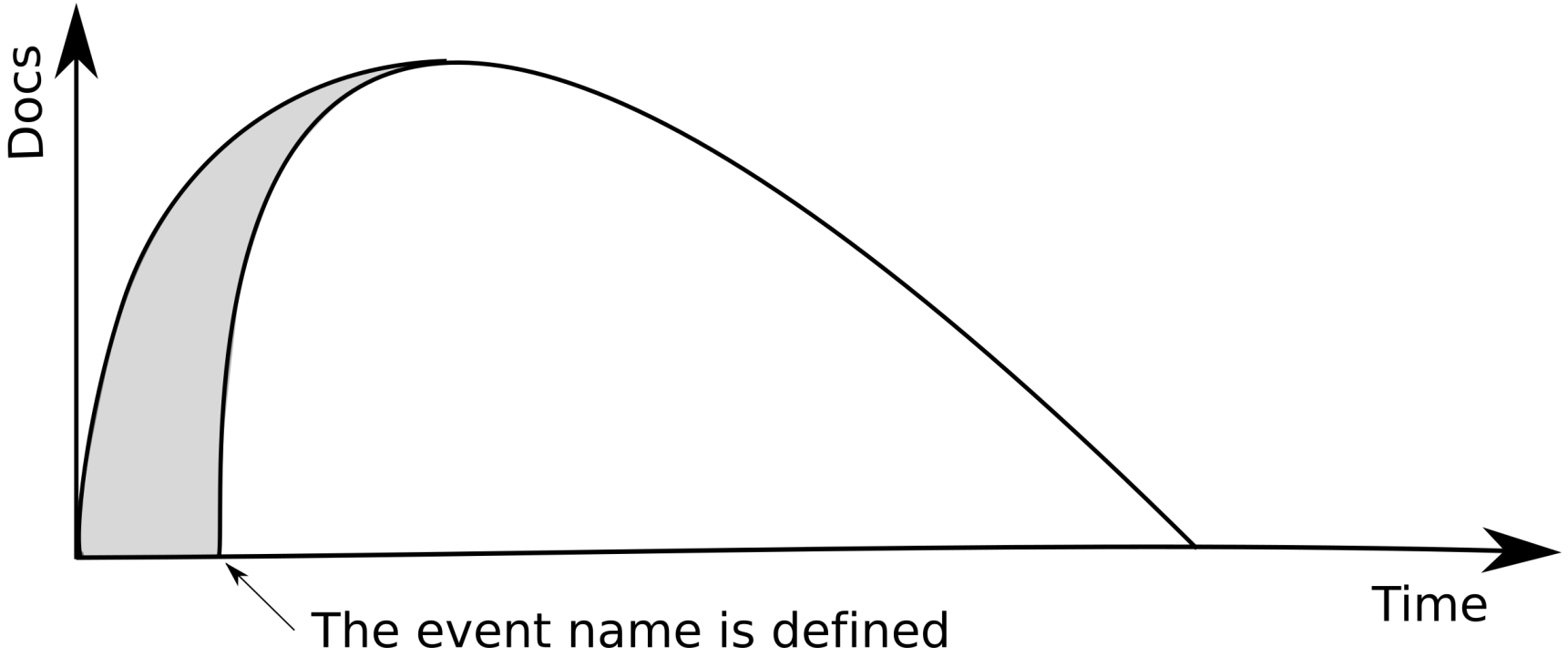
Banking Support Unable to Prevent Break Accompanied by Wild Trading

TICKERS FAR BEHIND

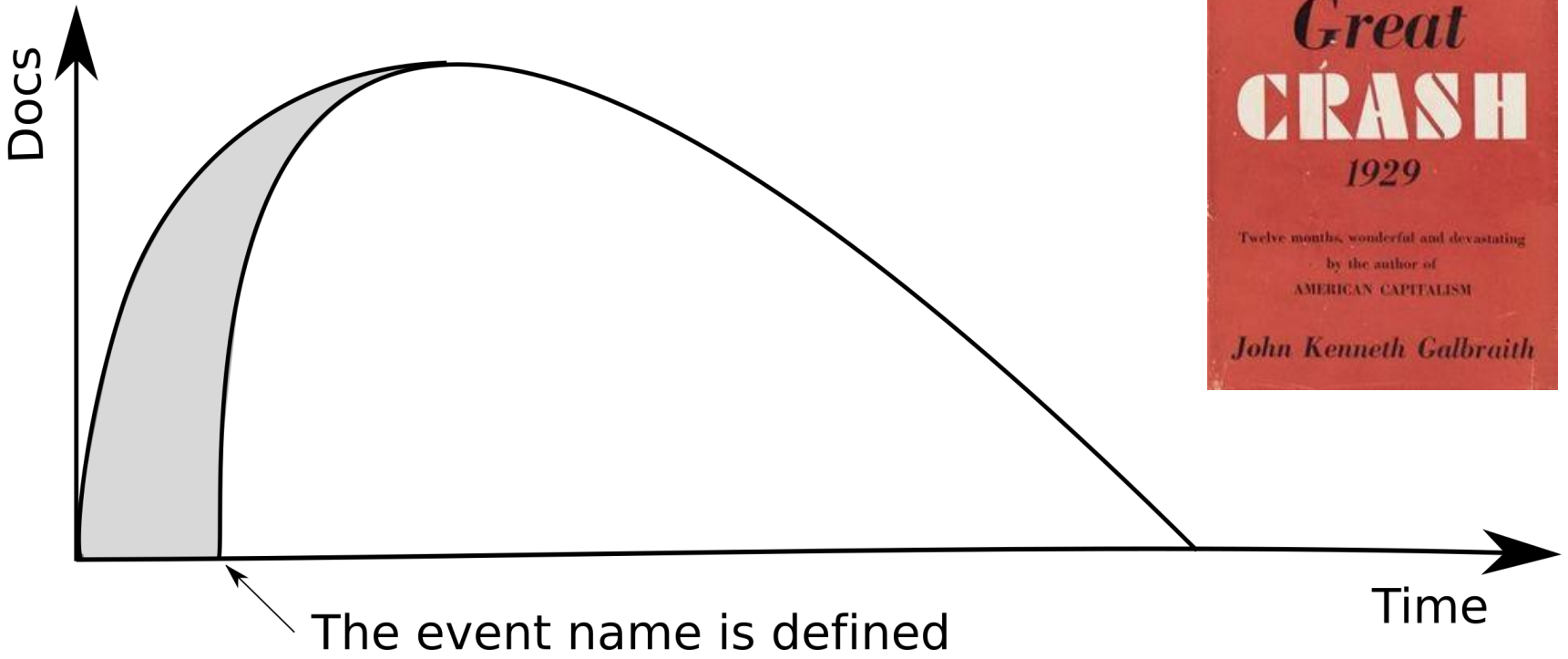
General Electric, U. S. Steel, Other Leading Issues Caught In Crash

NEW YORK, Oct. 28.—(UP)—

Missing the early stages



Missing the early stages



A contemporary example

@ReallyVirtual
Sohaib Athar

Helicopter hovering above Abbottabad at 1AM (is a rare event).

18 hours ago via TweetDeck

A contemporary example

@ReallyVirtual
Sohaib Athar

Helicopter hovering above Abbottabad at 1AM (is a rare event).

18 hours ago via TweetDeck

Brief idea for finding early stories

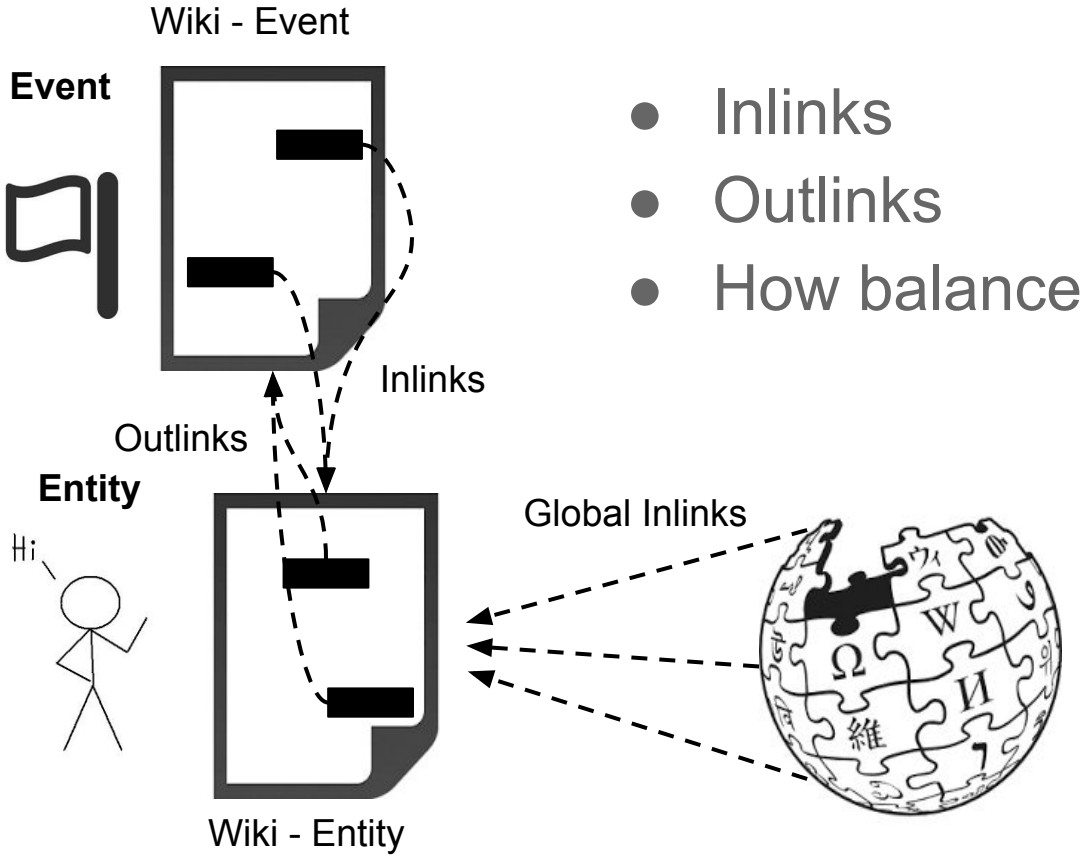
Given a named event:

1. Identify **related entities**
2. Retrieving **text passages** with entity in context
3. Building a **language model** for each entity

=> Entity event-query expansion!

How do we identify related entities?

Simple graph-based entity relatedness



- Inlinks
- Outlinks
- How balanced they are

Identifying related entities

System
Stics
Wiki2Vec
WikipediaRanking
Eventipedia (our)

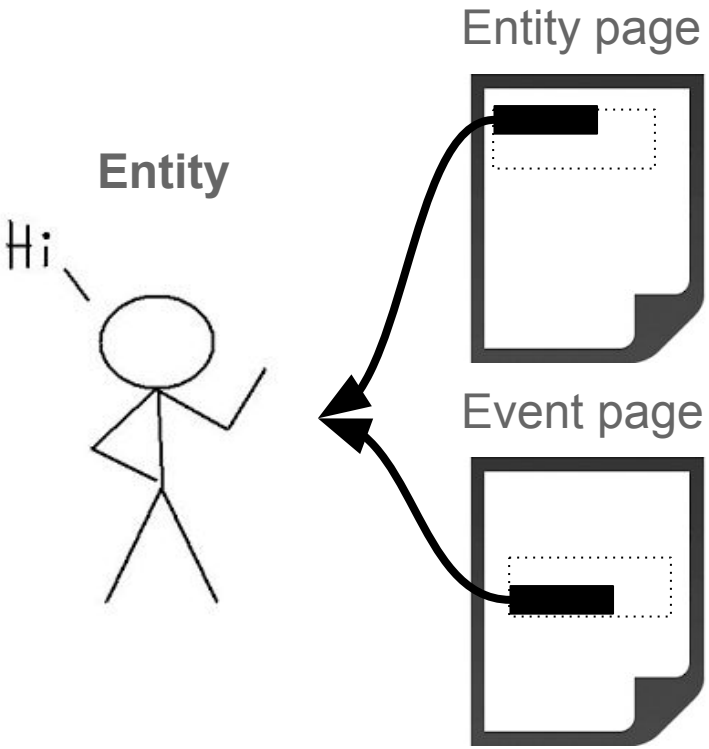
Gold standard: 10 global events (between 2012 and 2014). Human annotators assess the relevance of retrieved entities on a binary scale.

Identifying related entities

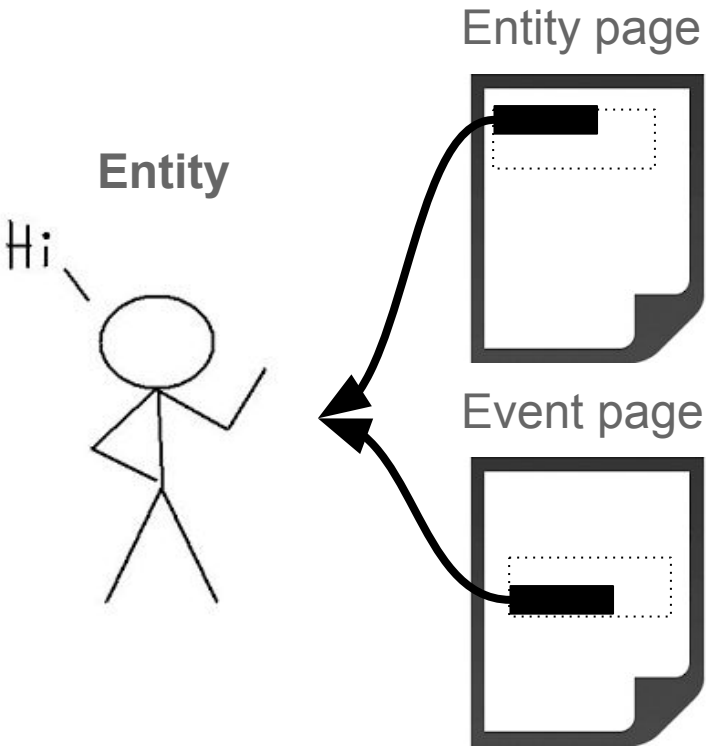
System	MAP@10	Micro-Prec@10
Stics	0.54 ± 0.07	0.59 ± 0.05
Wiki2Vec	0.59 ± 0.11	0.64 ± 0.04
WikipediaRanking	0.66 ± 0.09	0.71 ± 0.05
Eventipedia (our)	$\mathbf{0.74 \pm 0.05}$	$\mathbf{0.81 \pm 0.04}$

Gold standard: 10 global events (between 2012 and 2014). Human annotators assess the relevance of retrieved entities on a binary scale.

Retrieving explanatory passages

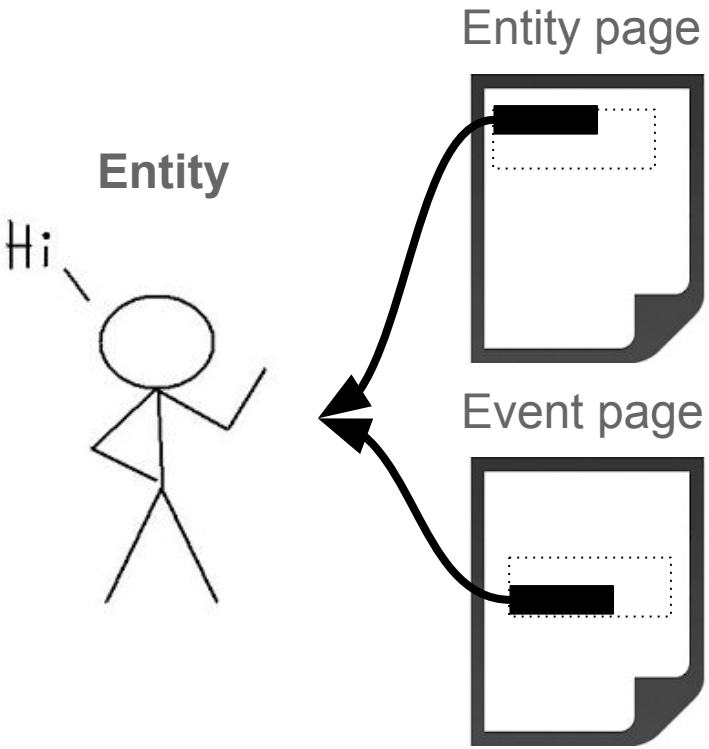


Retrieving explanatory passages



“**Russia**, also officially known as the Russian Federation, is a sovereign state in northern Eurasia. It is a federal semi-presidential republic.”

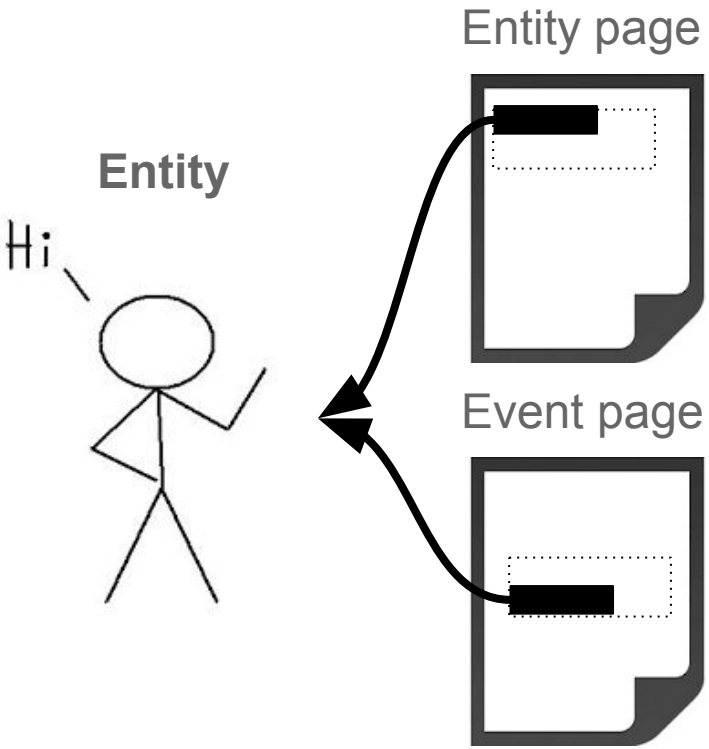
Retrieving explanatory passages



"**Russia**, also officially known as the Russian Federation, is a sovereign state in northern Eurasia. It is a federal semi-presidential republic."

"From the early stages, the Syrian government received technical, financial, military and political support from **Russia**, Iran and Iraq."

Retrieving explanatory passages



	% Good
Entity page snippet	45%
Event page snippet	**68%**

Conclusions

A simple **entity relatedness approach** is useful for identifying entities related to an event.

Entity needs to be considered **in the context of the event** for building the language model.

Next step: evaluate “early stories detection”.

Nanni, Federico, Simone Paolo Ponzetto, and Laura Dietz. "Entity Relatedness for Retrospective Analyses of Global Events.", NLP+CSS Workshop at WebSci16, 2016.

UNIVERSITÄT MANNHEIM

Topic-Based Analysis of Political Positions in US Electoral Campaigns

Federico Nanni, Caecilia Zirn, Goran Glavas
Jason Eichorst and Simone Paolo Ponzetto

Data and Web Science Group
Collaborative Research Center SFB 884
University of Mannheim

{federico,caecilia, goran,simone}@informatik.uni-mannheim.de

Introduction

- Campaigns designed to **inform voters** on candidates ideas

Introduction

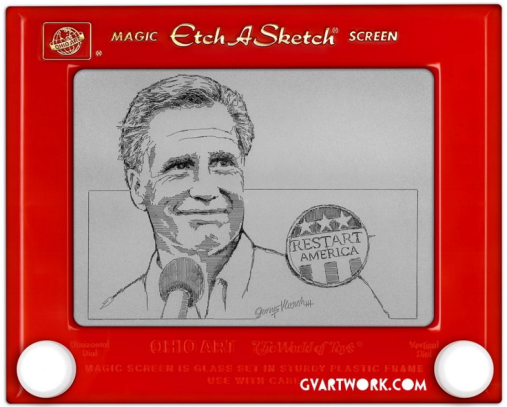
- Campaigns designed to **inform voters** on candidates ideas
- **Converging on a position** is an interesting process because candidates must satisfy
    - Interests of party members and groups during primaries
    - Interests of voters in general elections

Introduction

- Campaigns designed to **inform voters** on candidates ideas
- **Converging on a position** is an interesting process because candidates must satisfy
    - Interests of party members and groups during primaries
    - Interests of voters in general elections
- Often, there is a **notable shift** in candidate positions between primaries and general elections

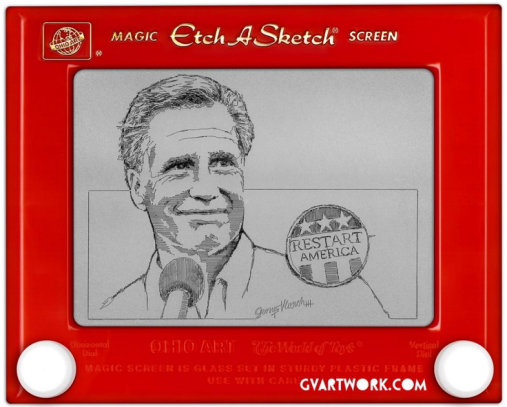
Mitt Romney's Etch-a-Sketch

"Everything changes. It's **almost like an Etch-a-Sketch**. You can kind of shake it up and restart all over again."



Mitt Romney's Etch-a-Sketch

"Everything changes. It's **almost like an Etch-a-Sketch**. You can kind of shake it up and restart all over again."



Starting Hypothesis: For certain **topics** these changes might be **more prominent** than for others

Topics

Coarse-grained topics from **Comparative Manifesto Project**:

1. External relations
2. Freedom and Democracy
3. Political System
4. Economy
5. Welfare and Quality of Life
6. Fabric of Society
7. Social Groups

Speech dataset

Speeches **manually labeled** with topics at paragraph level. Two annotators, moderate IAA of 0.55 (Cohen's kappa).

Contains altogether **9 speeches** from 2008, 2012, and 2016 elections (around 1k paragraphs).

Topical classification

Topic classifier: SVM with **lexical** and **semantic** features.

Two experimental settings:

1. **Domain transfer setting**: training the model on manifestos, testing on speeches
2. **Pure speeches setting** – Folded cross-validation on speeches

Baseline model: SVM with bag-of-words features.

Topical classification

Classification results (in terms of F1 score):

- Domain transfer setting (training on manifestos): 36.2%
- Baseline model (BoW SVM, CV on speeches): 71.2%
- Pure speech setting (CV on speeches): **78.6%**

Conclusion: **Transfer learning doesn't work** between different domains of political text.

Political scaling

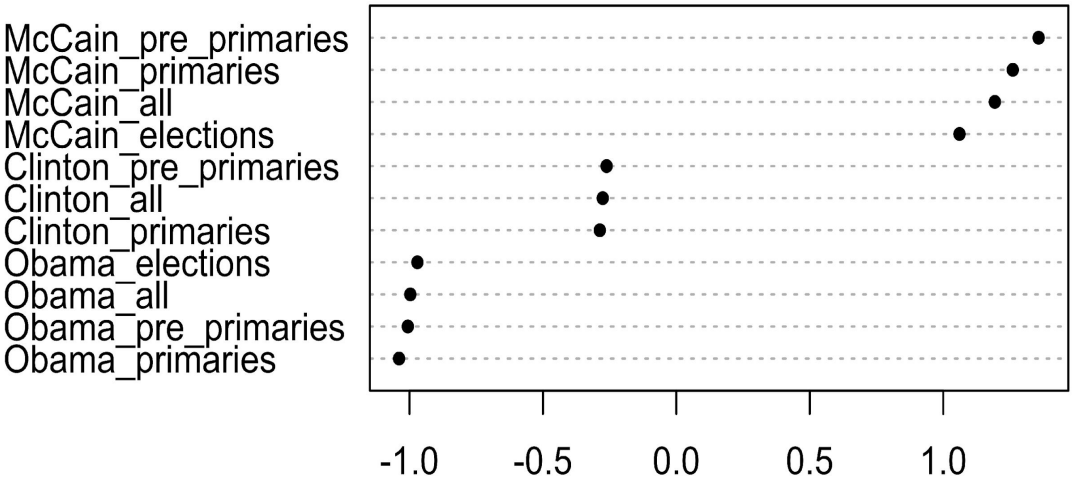
Best performing model made **topic predictions** on the speeches.

For each candidate, we concatenate all paragraphs with **same topic and same phase** (pre-primaries, primaries, elections).

Finally, we feed each phase-topic slice to the **Wordfish** tool.

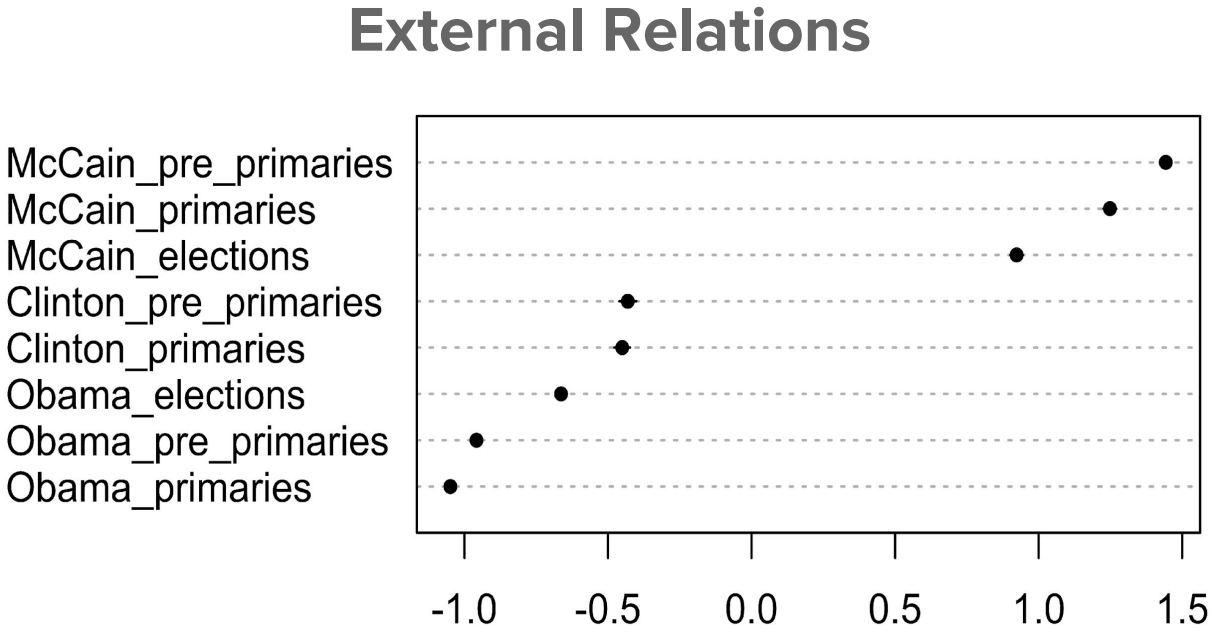
Coarse-grained analysis (2008)

First we analyzed the **general positions** (Wordfish on entire speeches), as a baseline for analysis.



Fine-grained analysis (2008)

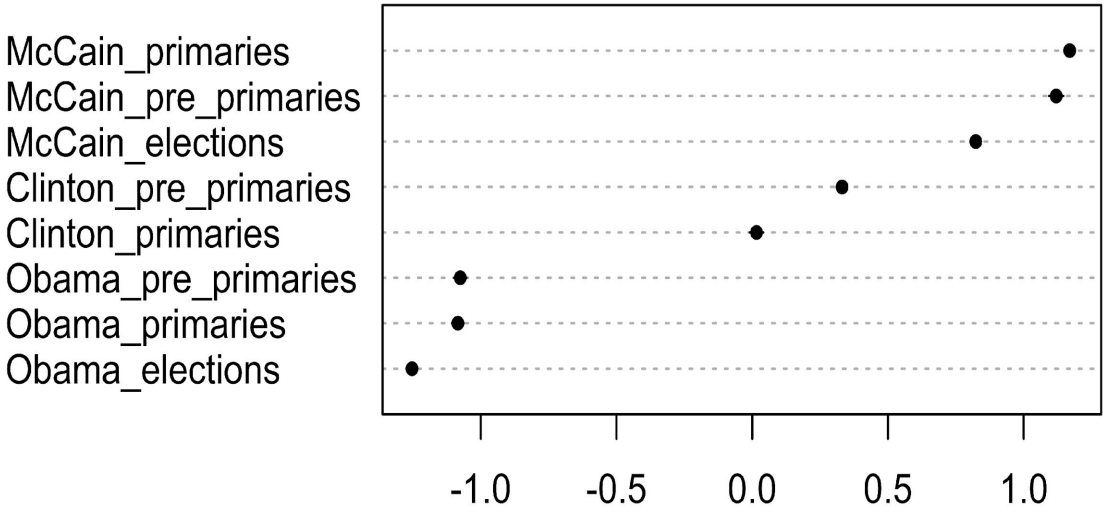
Next, we analyzed topic-specific positions:



Fine-grained analysis (2008)

Next, we analyzed topic-specific positions:

Welfare and Quality of Life



Conclusion

Topic-based position analysis could offer **fine-grained perspectives** on political campaigns.

Important **evaluating** approaches for the task.

We are working on a **Python** implementation of the pipeline!

Final wrap-up

Our approach to DH has a **strong NLP approach**.

Focus on:

1. Hypothesis-testing analyses
2. Tool evaluation
3. Train researchers with highly interdisciplinary profiles

UNIVERSITÄT MANNHEIM

Thanks!

Data and Web Science Group

http://dws.informatik.uni-mannheim.de

{federico,simone}@informatik.uni-mannheim.de

Historisches Institut

http://www.geschichte.uni-mannheim.de

hiram.kuemper@uni-mannheim.de