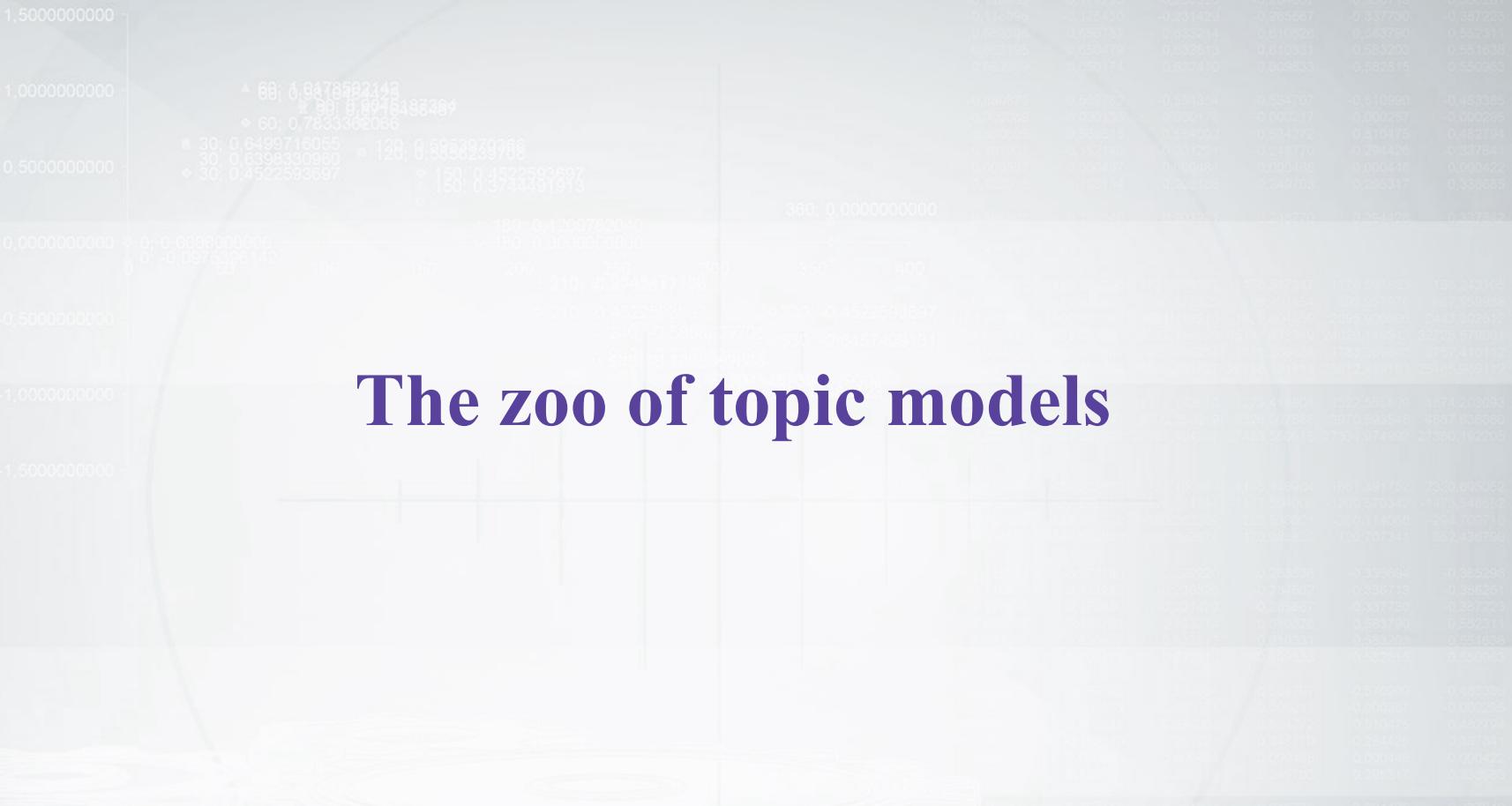


# The zoo of topic models



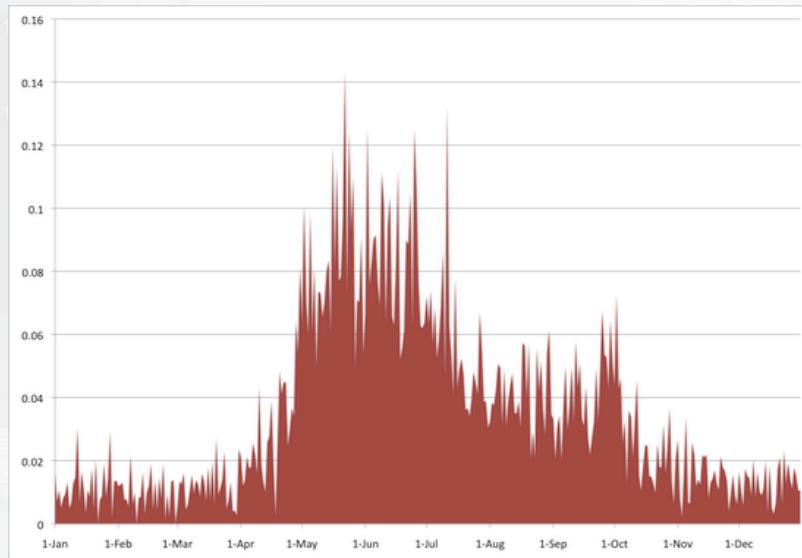
180 -0.417402  
EF 47.44  
G 50.10  
K 50.10  
Ks 50.000000  
Ye 57.133 YI 76.083319 RSetRotete \*chime\_A 50.000000 kx 0.0  
Z 50.10 ZI  
pi-5 0.099355  
VY(1-EF)\*cos(kx-m\*pi/2)  
Zp-Zc-EP\*pi\*cos(kx-m\*pi/2)

# Martha Ballard's diary

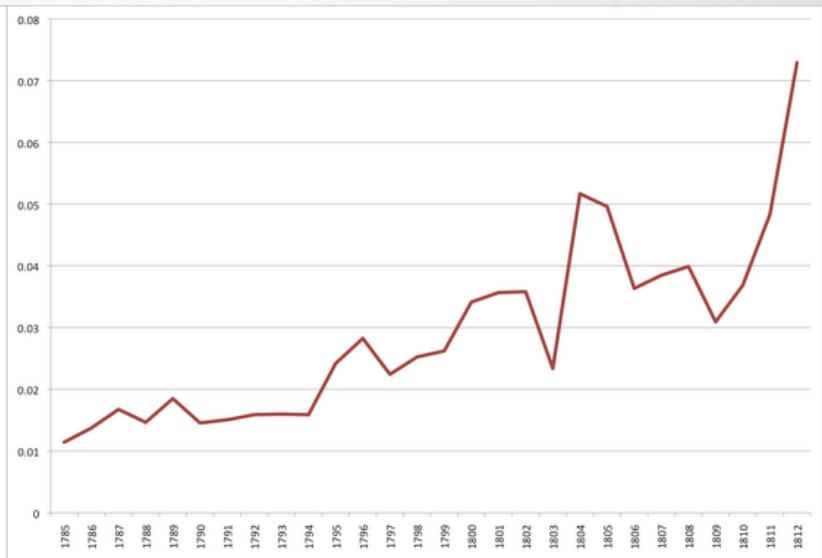
- Diary had daily entries over the course of 27 years
- Topic modeling helps to analyze it
- Revealed topics (the most probable words):
  - **GARDENING:** *garden worked clear beans corn warm planted matters cucumbers potatoes plants*
  - **CHURCH:** *meeting attended afternoon reverend worship foren mr famely st lecture discoarst administered*
  - **DEATH:** *day yesterday informed morn years death ye hear expired expired weak dead*
  - **SHOPPING:** *butter sugar carried candles wheat store flower*

# Martha Ballard's diary

- Diary had daily entries over the course of 27 years
- Topic modeling helps to analyze it
- How topics are developing through time:



Gardening (average year)



Emotions (1785-1812)

<http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/>

# Latent Dirichlet Allocation

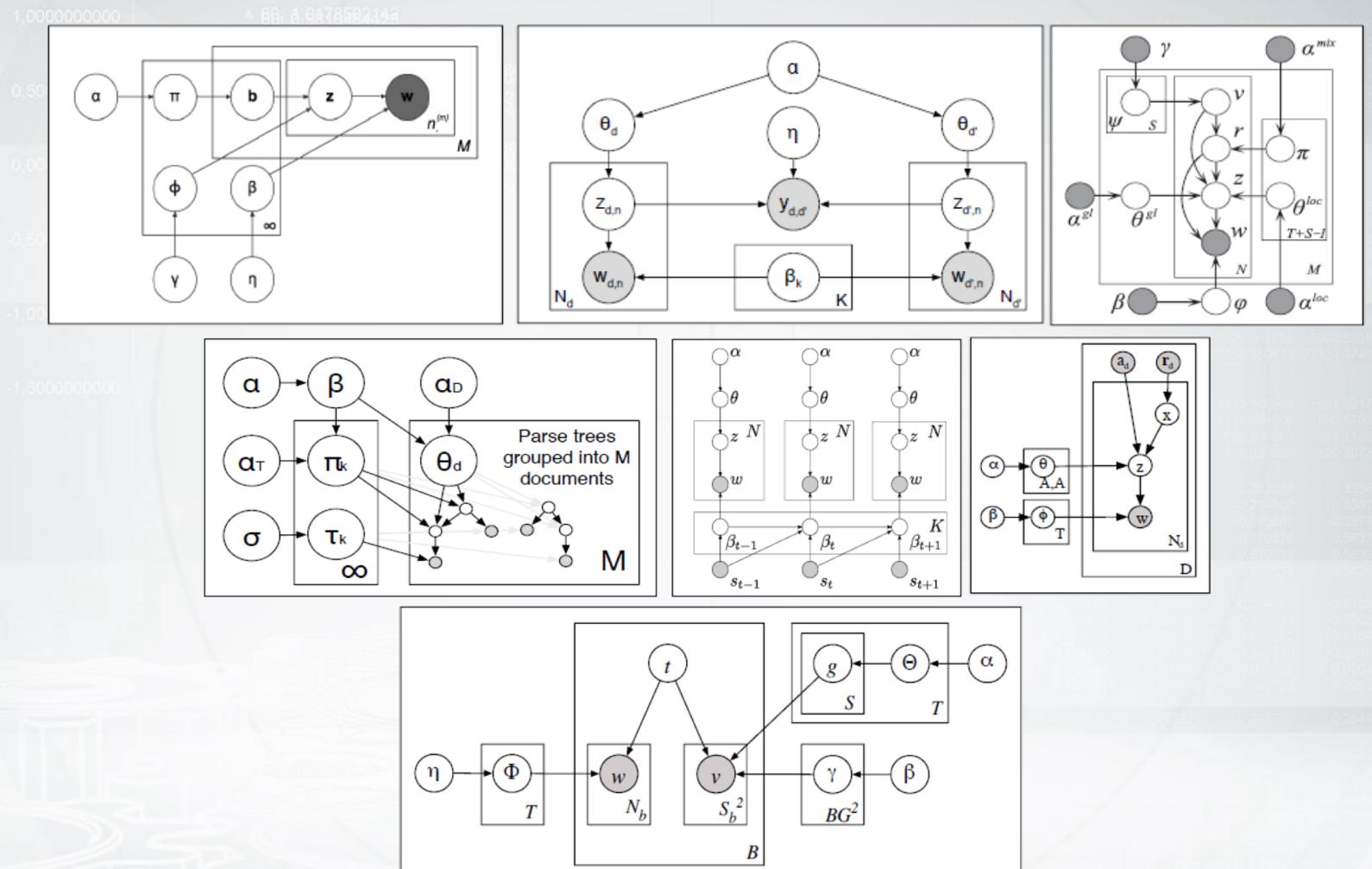
Dirichlet priors for  $\phi_t = (\phi_{wt})_{w \in W}$  and  $\theta_d = (\theta_{td})_{t \in T}$ :

$$Dir(\phi_t | \beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1} \quad \beta_0 = \sum_w \beta_w, \beta_t > 0$$

- Inference:
  - Variational Bayes
  - Gibbs Sampling
- Output:
  - Posterior probabilities for parameters (also Dirichlet!).

Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models, 2009.

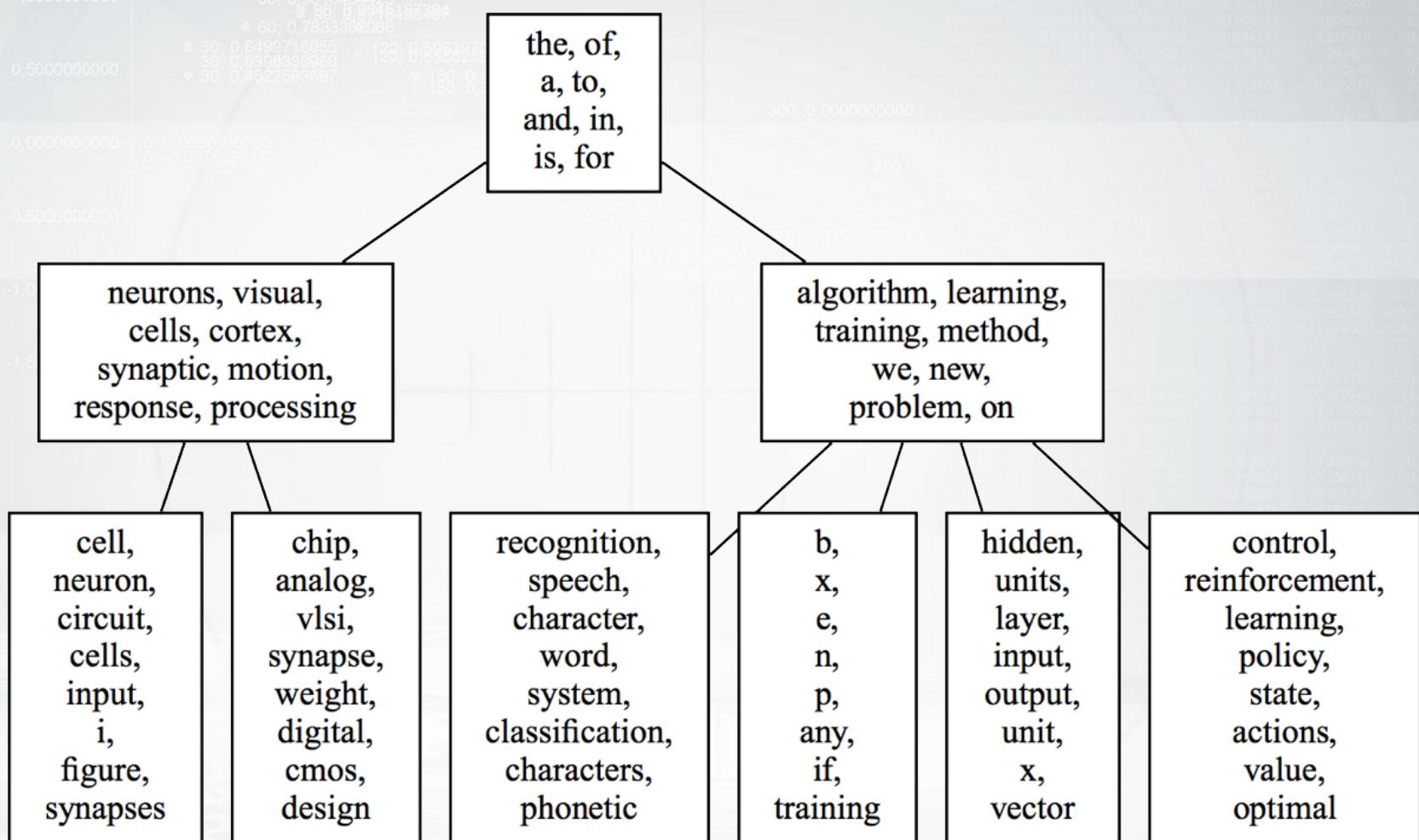
# Bayesian methods and graphical models



Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad.

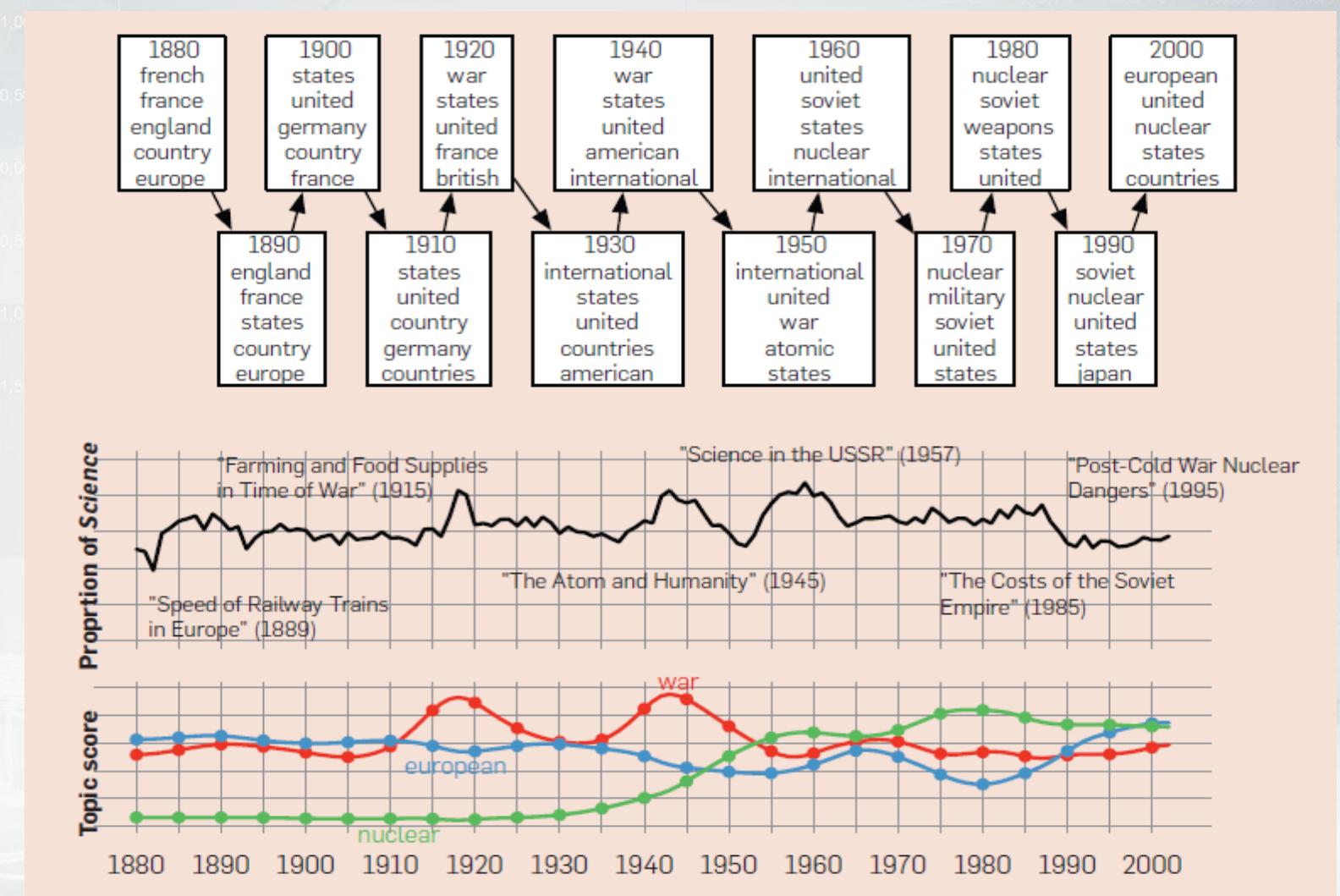
Knowledge discovery through directed probabilistic topic models: a survey, 2010.

# Hierarchical topic models



D. Blei et. al. Hierarchical Topic Models and the Nested Chinese Restaurant Process, NIPS-2003.

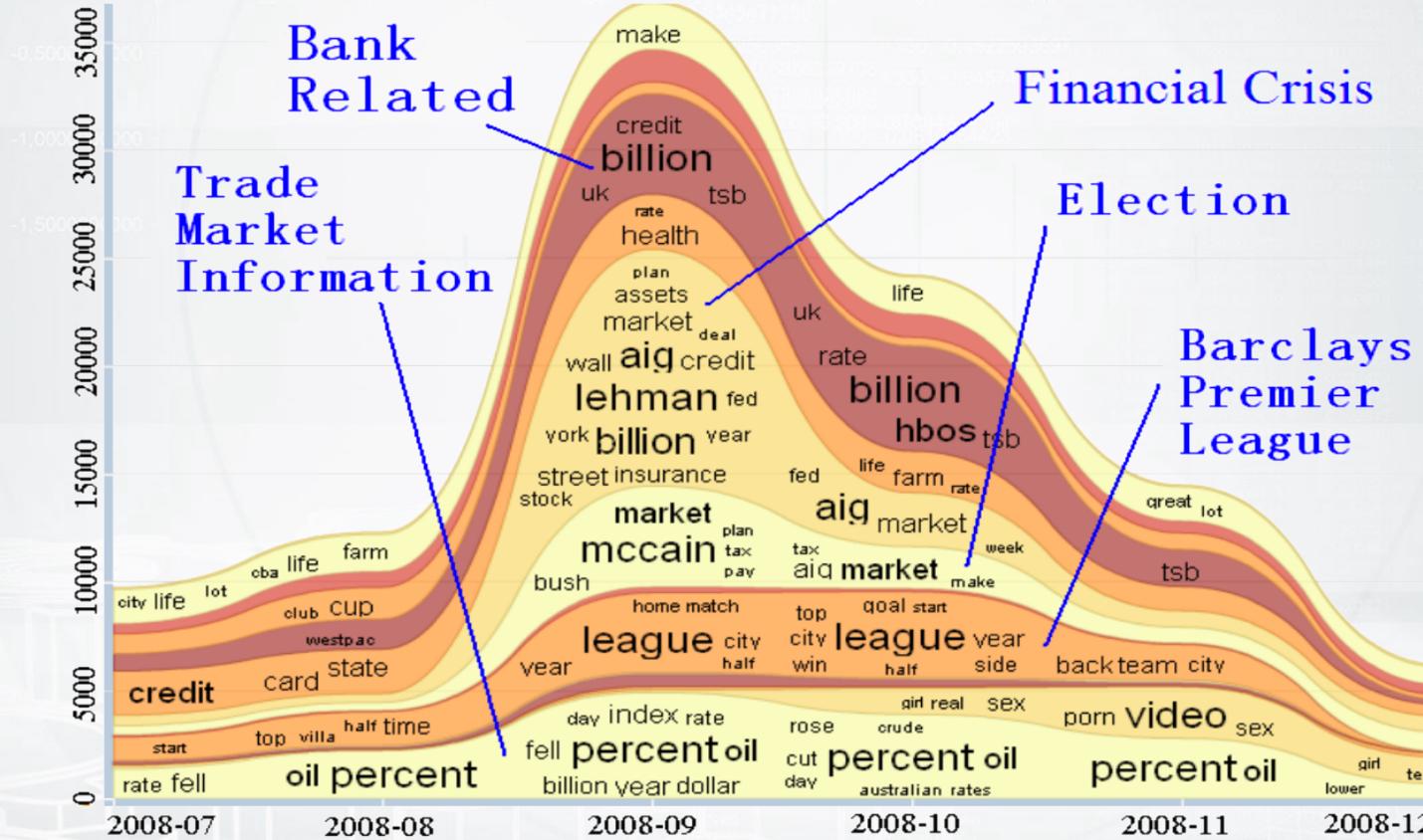
# Dynamic topic models



David Blei, Probabilistic Topic Models, 2012.

# Dynamic topic models

Topic detection and analysis of news flows:

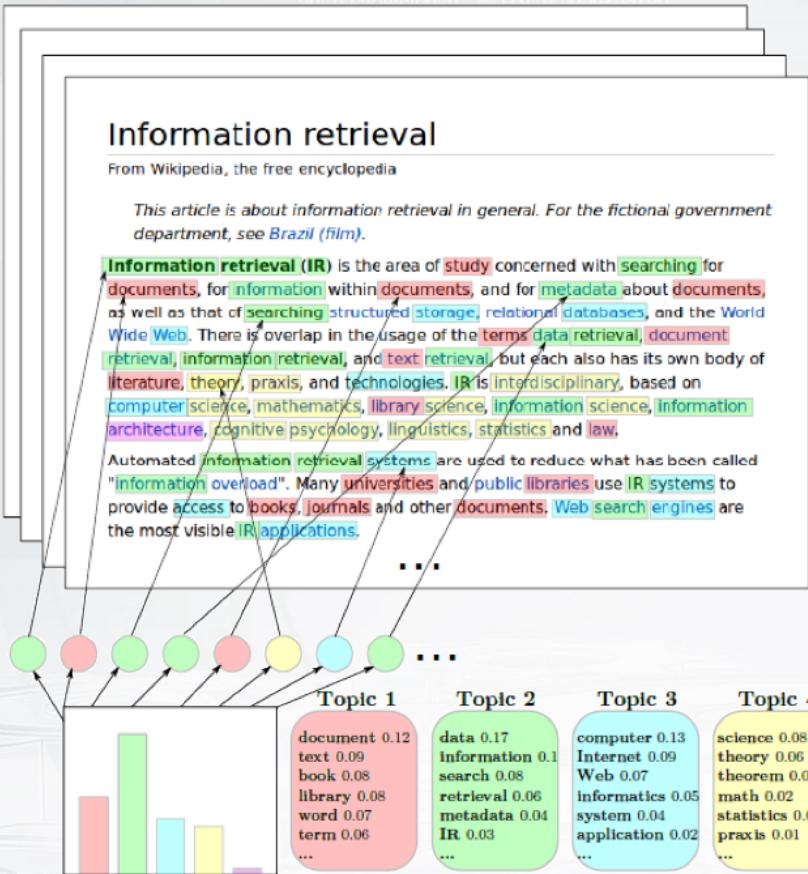


Jianwen Zhang, Yangqiu Song, Changshui Zhang, Shixia Liu Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying, KDD-2010.

# Multilingual topic models

1.0000000000

## English corpus



### Information retrieval

From Wikipedia, the free encyclopedia

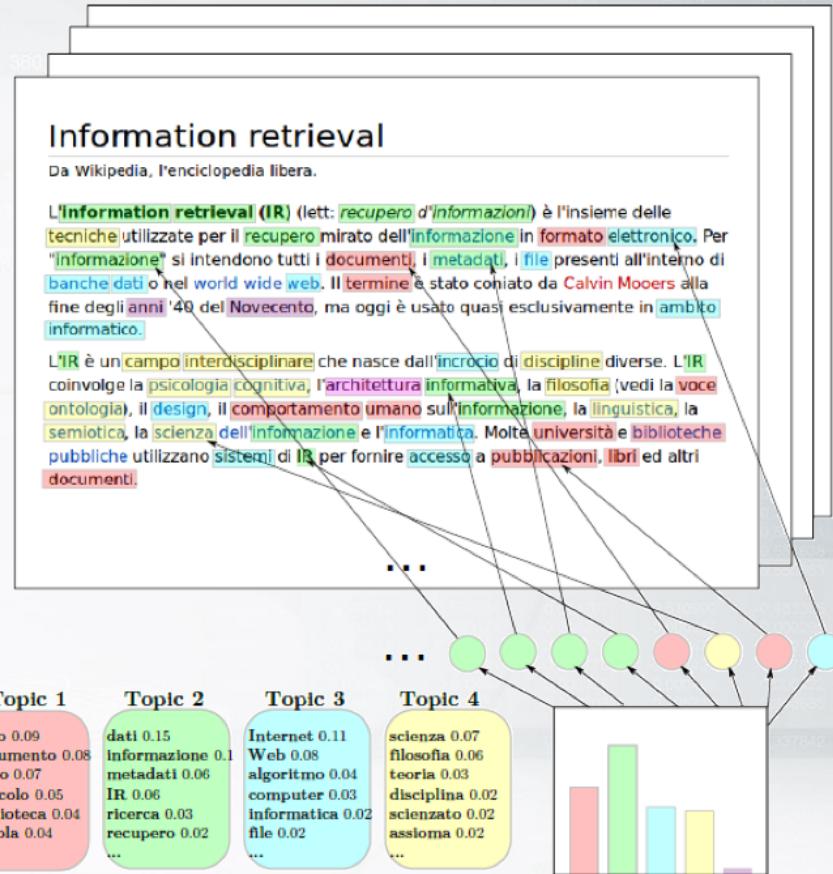
This article is about information retrieval in general. For the fictional government department, see *Brazil (film)*.

**Information retrieval (IR)** is the area of **study** concerned with **searching** for **documents**, for **information** within **documents**, and for **metadata** about **documents**, as well as that of **searching** **structured storage**, **relational databases**, and the **World Wide Web**. There is overlap in the usage of the terms **data retrieval**, **document retrieval**, **information retrieval**, and **text retrieval**, but each also has its own body of **literature**, **theory**, **praxis**, and **technologies**. IR is **interdisciplinary**, based on **computer science**, **mathematics**, **library science**, **information science**, **information architecture**, **cognitive psychology**, **linguistics**, **statistics** and **law**.

Automated **information retrieval systems** are used to reduce what has been called "**information overload**". Many **universities** and **public libraries** use **IR systems** to provide **access** to books, **journals** and other documents. **Web search engines** are the most visible **IR applications**.

-0.115781	0.173160	0.229220	0.263536	-0.335684	0.385298
-0.116939	0.174295	0.230325	0.264602	-0.336713	0.386261
-0.118095	0.177430	0.231423	0.265687	-0.337730	0.387223
-0.119254	0.180567	0.232524	0.266766	0.543203	0.552311
-0.120400	0.183694	0.233624	0.267843	0.544303	0.553215
-0.121549	0.186821	0.234721	0.268920	0.545390	0.554083
-0.122690	0.190948	0.235818	0.270000	-0.510990	-0.483385
-0.123823	0.194075	0.236915	0.271075	-0.000257	-0.000285
-0.124955	0.197202	0.238012	0.272150	0.510475	0.482284

## Italian corpus



### Information retrieval

Da Wikipedia, l'enciclopedia libera.

L'**information retrieval (IR)** (lett: *recupero d'informazioni*) è l'insieme delle tecniche utilizzate per il recupero mirato dell'informazione in formato elettronico. Per "informazione" si intendono tutti i documenti, i metadati, i file presenti all'interno di banche dati o nel world wide web. Il termine è stato coniato da *Calvin Mooers* alla fine degli anni '40 del Novecento, ma oggi è usato quasi esclusivamente in ambito informatico.

L'**IR** è un campo interdisciplinare che nasce dall'incrocio di discipline diverse. L'**IR** coinvolge la **psicologia cognitiva**, l'**architettura informativa**, la **filosofia** (vedi la voce **ontologia**), il **design**, il **comportamento umano sull'informazione**, la **linguistica**, la **semiotica**, la scienza dell'**informazione** e l'**informatica**. Molte **università** e **biblioteche pubbliche** utilizzano **sistemi** di **IR** per fornire **accesso** a **pubblicazioni**, **ibri** ed altri **documenti**.

I. Vulic, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: a short overview of its methodology with applications, NIPS-2012.

# Additive Regularization for Topic Models

How to combine all those extensions in one model?

$$\text{PLSA: } \mathcal{L} = \sum_{d \in D} \sum_{w \in W} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

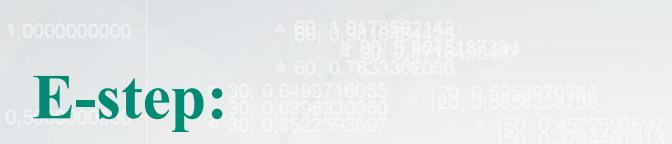
$$\text{ARTM: } \mathcal{L} + \sum_{i=1}^n \tau_i R_i(\Phi, \theta) \rightarrow \max_{\Phi, \Theta}$$

Example of a regularizer – diversity of topics:

$$R_i(\Phi) = - \sum_{t \neq s} \sum_w \phi_{wt} \phi_{ws}$$

K. Vorontsov, A. Potapenko Additive Regularization of Topic Models, 2015.

# Regularized EM-algorithm



**E-step:**

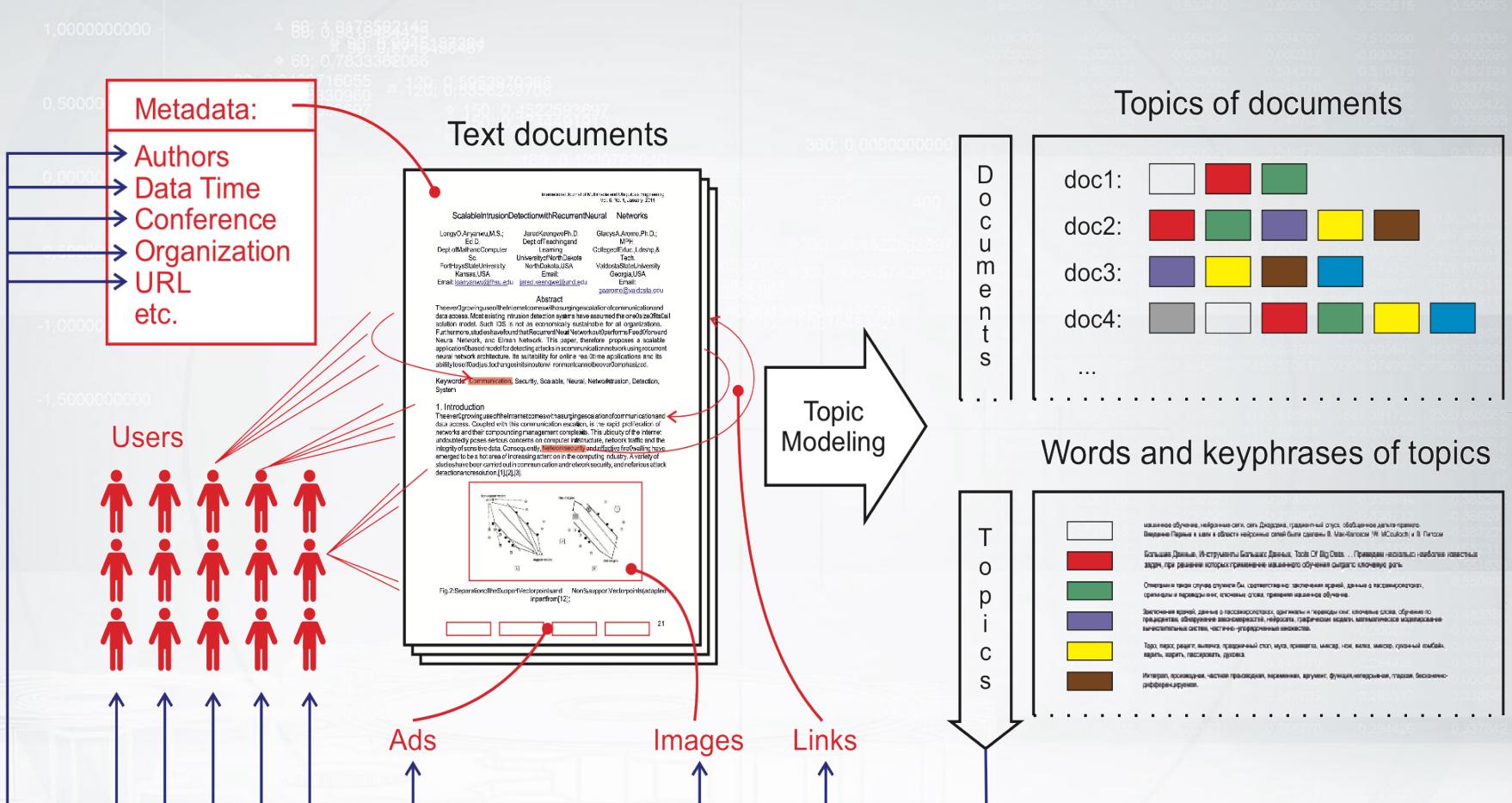
$$p(t|d, w) = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}$$

**M-step:**

$$\phi_{wt} = \text{norm}_{w \in W} \left( \sum_{d \in D} n_{dw} p(t|d, w) + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$$

$$\theta_{td} = \text{norm}_{t \in T} \left( \sum_{w \in d} n_{dw} p(t|d, w) + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$$

# Multimodal topic models



K. Vorontsov et. al. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections, 2015.

# Multi-ARTM

How to incorporate tokens of additional modalities?

**PLSA:**  $\mathcal{L} = \sum_{d \in D} \sum_{w \in W} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$

**Multi-ARTM:**

$$\sum_{m \in M} \lambda_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

- Each topic is characterized by several probability distributions
- More parameters, still trained with EM-algorithm

# Inter-modality similarities

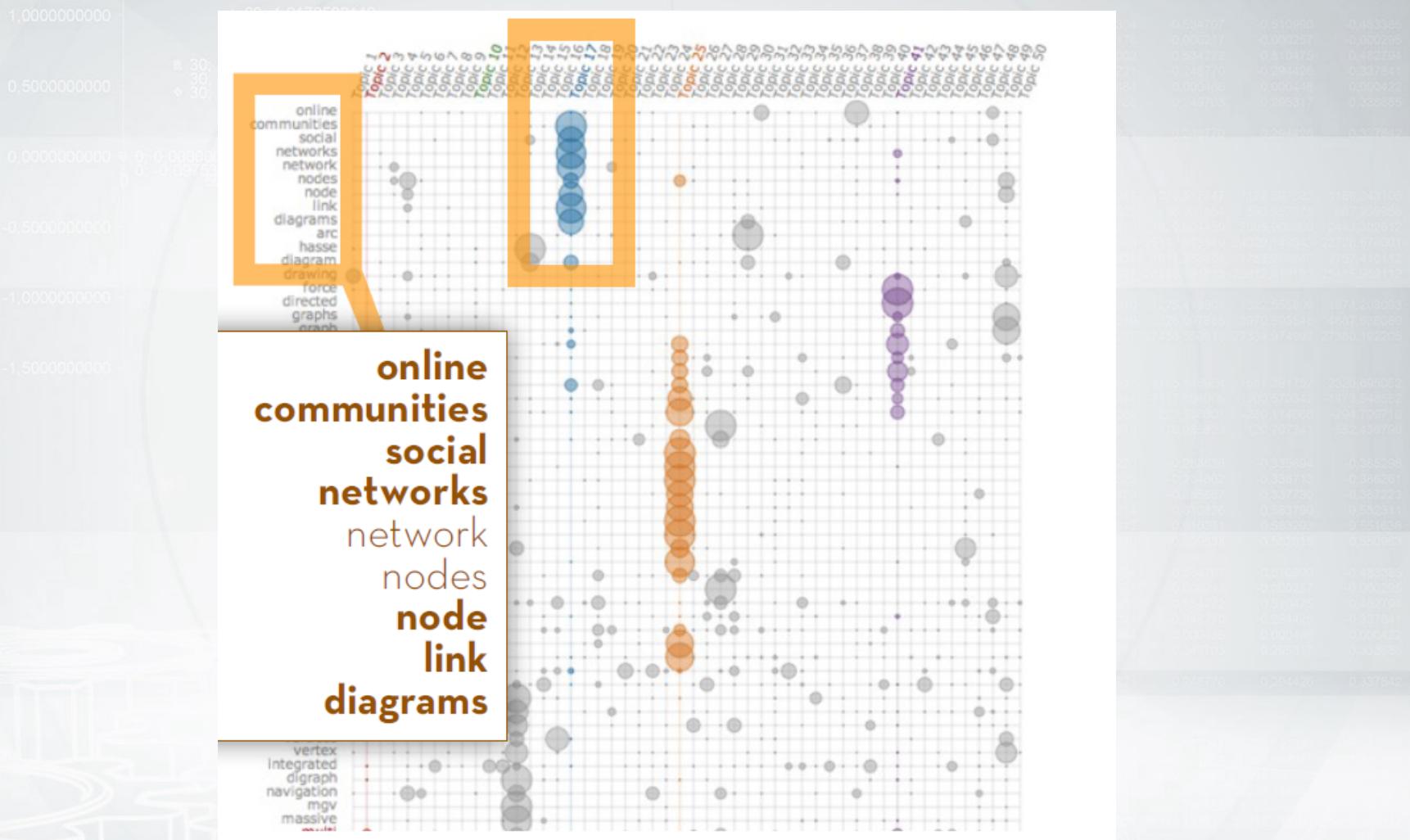
2015-12-18 Star Wars Release	2016-02-29 The Oscars	2015-05-09 Victory Day
jedi sith fett anakin chewbacca film series hamill prequel awaken boyega	statuette award nomination linklater oscar birdman win criticism director lubezki	great anniversary normandy parade demonstration vladimir celebration concentration auschwitz photograph

Potapenko, Popov, Vorontsov: Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks, 2017.

# Libraries for topic modeling

- **BigARTM** is an open-source library for Additive Regularization of Topic Models, [bigartm.org](http://bigartm.org)
- **Gensim** is a library of text analysis for Python, [radimrehurek.com/gensim](http://radimrehurek.com/gensim)
- **MALLET** is a library of text analysis for Java [mallet.cs.umass.edu](http://mallet.cs.umass.edu)
- **Vowpal Wabbit** has a fast implementation of online LDA [hunch.net/~vw/](http://hunch.net/~vw/)

# A few words about visualization



J. Chuang, C. D. Manning, J. Heer – Termite: Visualization Techniques For Assessing Textual Topic Models, 2012

# 380 ways to visualize: textvis.lnu.se



0.173160	0.229220	0.263536	-0.335684	0.385298
0.174295	-0.290325	-0.284602	-0.336713	0.386261
0.175430	-0.231423	0.265687	-0.337730	-0.387223
0.176565	0.155214	0.1610526	0.1543780	0.1552311
0.177700	0.033813	0.010331	0.0381003	0.0381003
0.178835	0.020410	0.009833	0.020515	0.020515

