The zoo of topic models

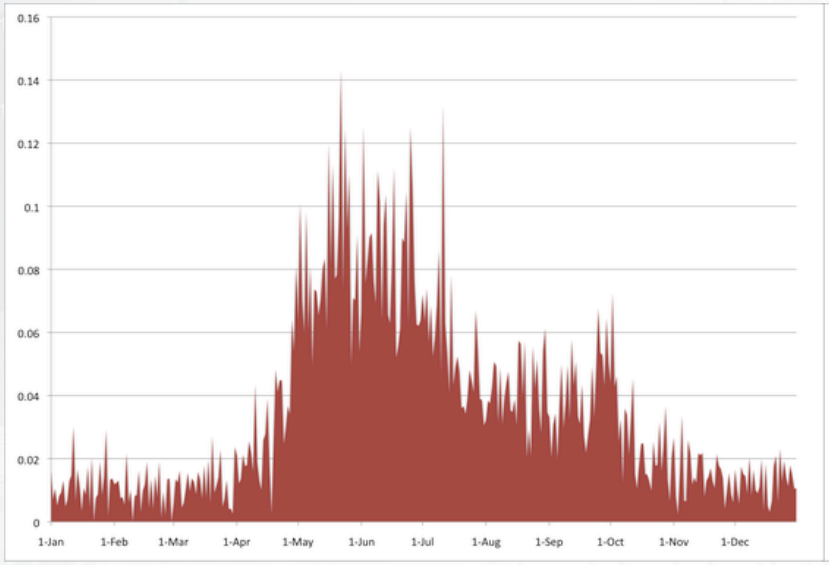
Martha Ballard's diary

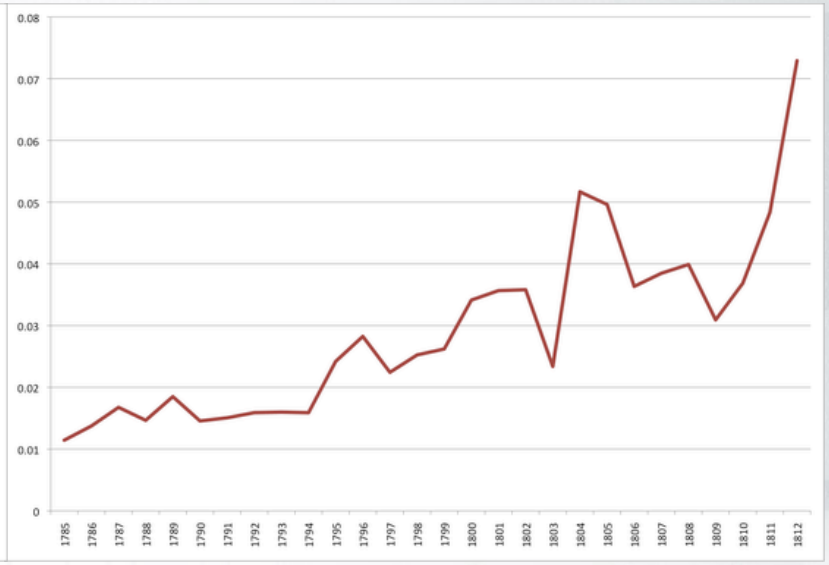
- Diary had daily entries over the course of 27 years
- Topic modeling helps to analyze it
- Revealed topics (the most probable words):

- ***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 $\theta_d = (\theta_{td})_{t \in T}$:

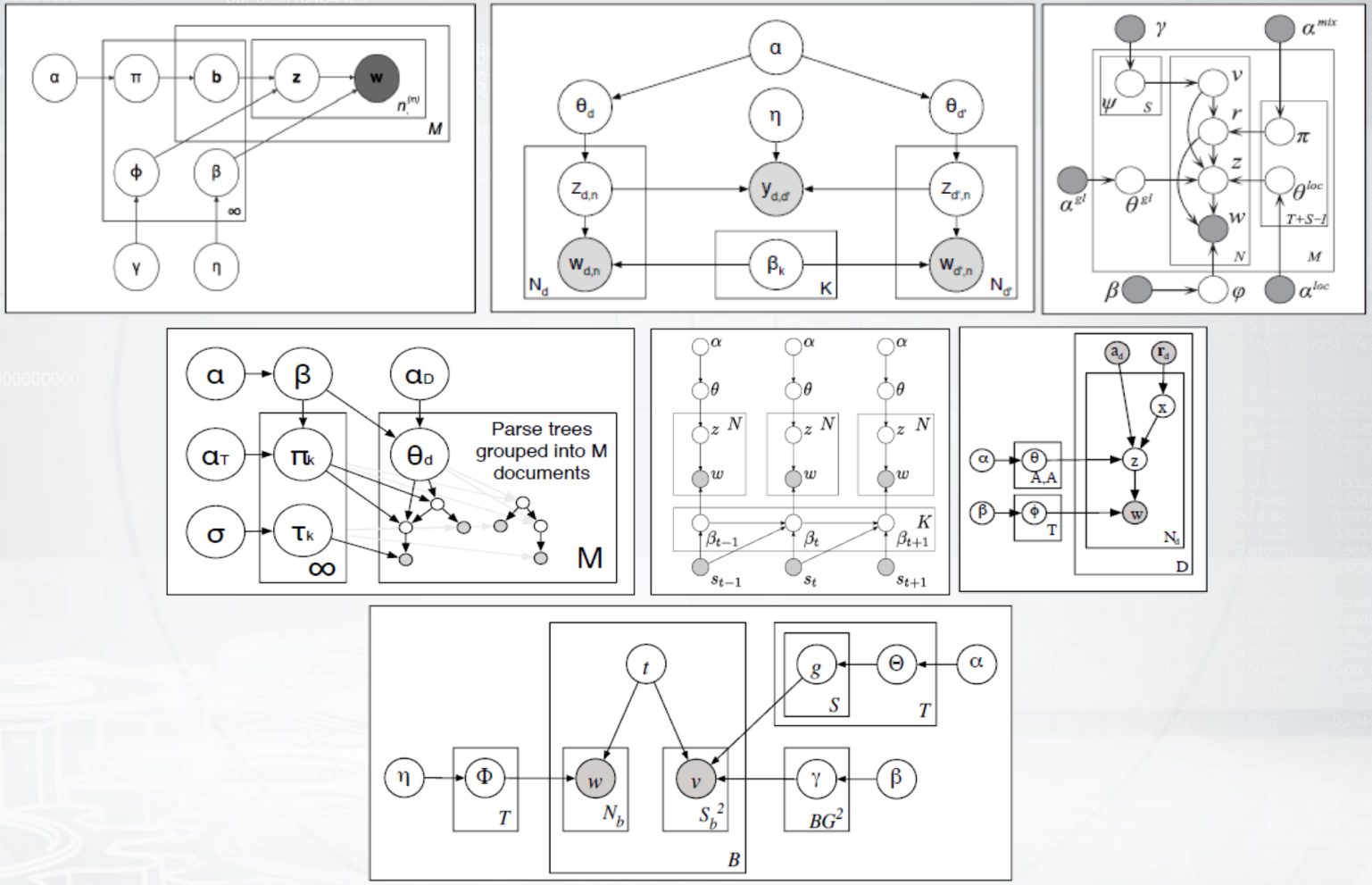
$$Dir(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1} \quad \beta_0 = \sum_w \beta_w, \beta_t > 0$$

- **Inference:**
  - Variational Bayes
  - Gibbs Sampling

- **Output:**
  - Posterior probabilities for parameters (also Dirichlet!).

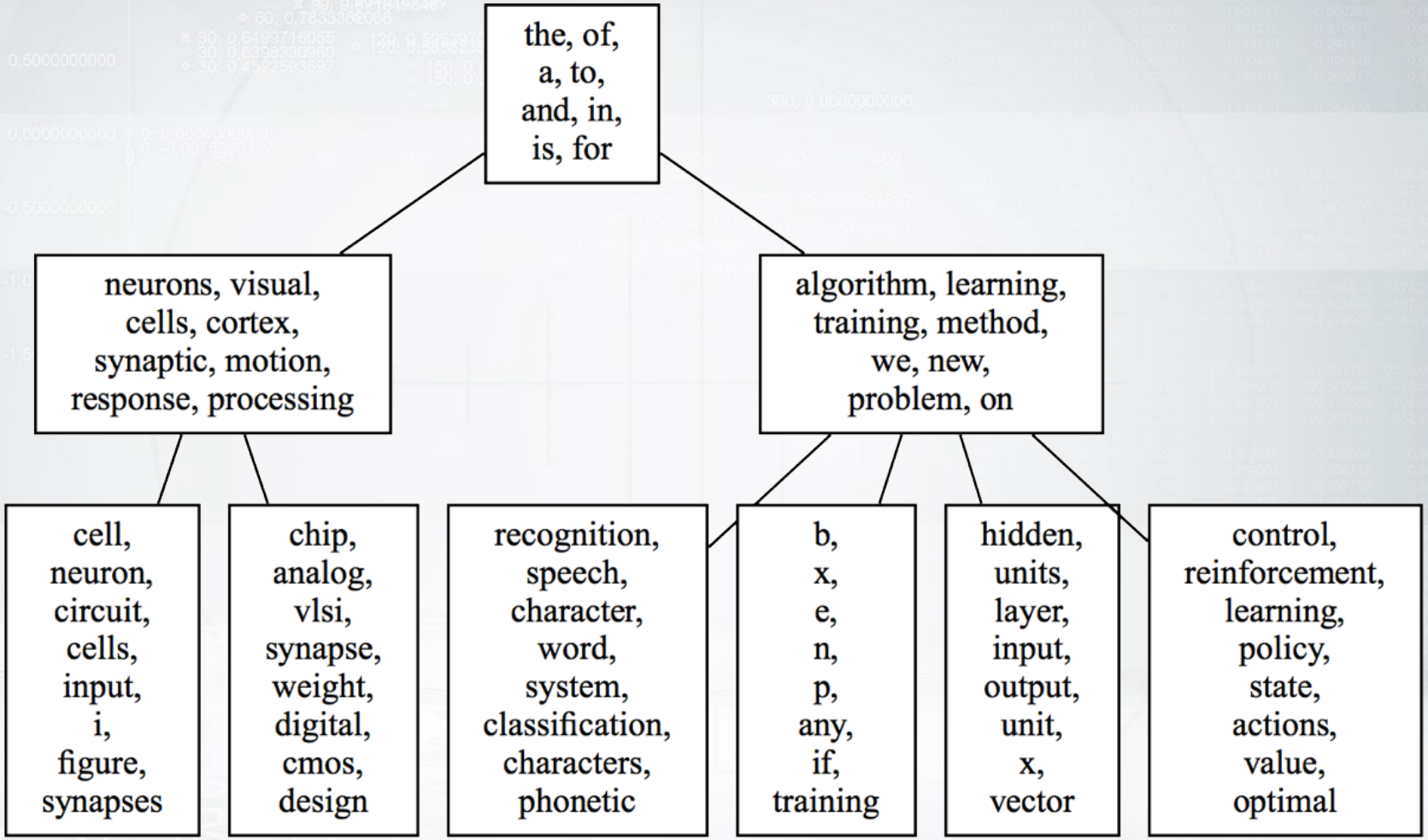
Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models, 2009.

Bayesian methods and graphical models



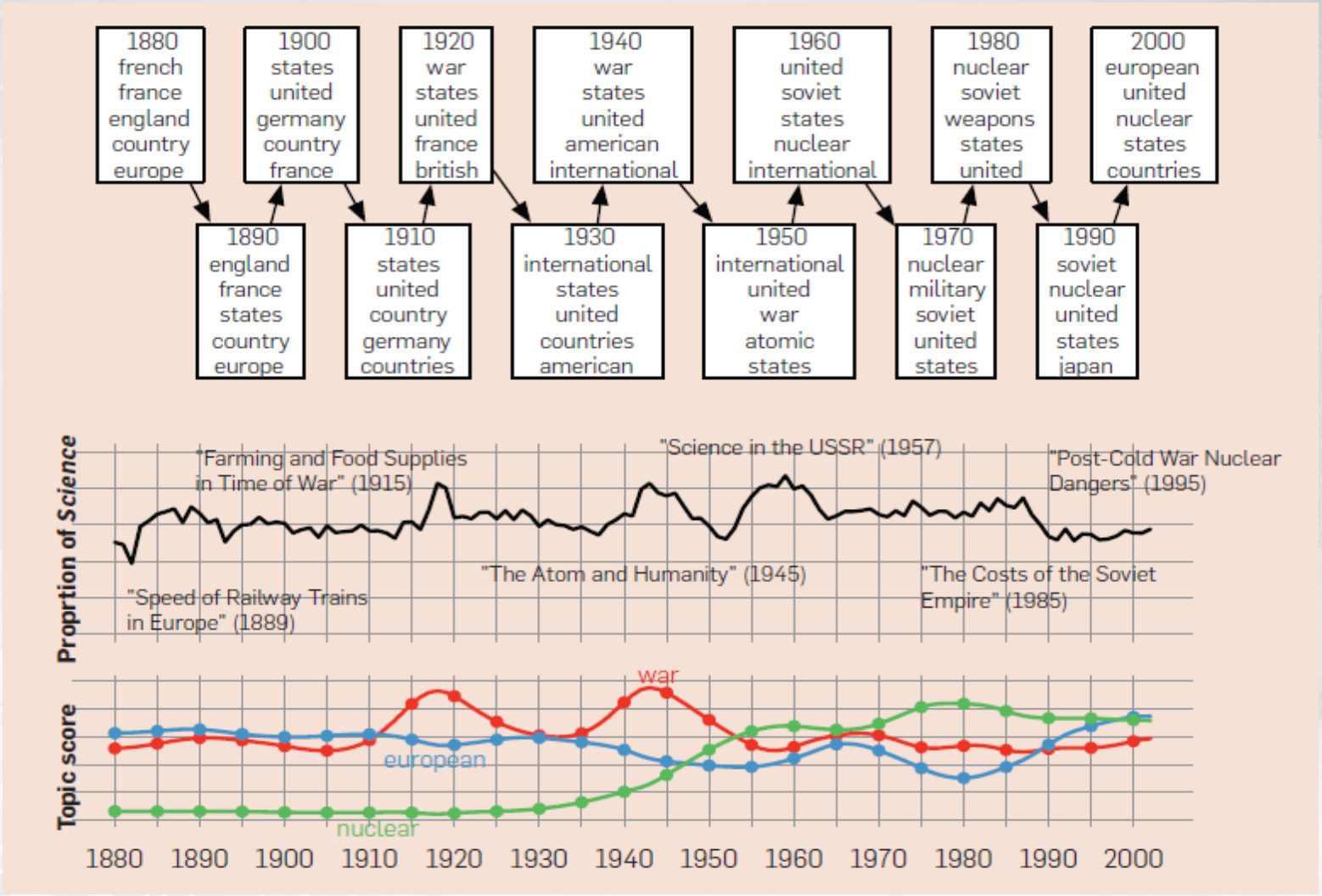
Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad.
Knowledge discovery through directed probabilistic topic models: a survey, 2010.

Hierarchical topic models



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

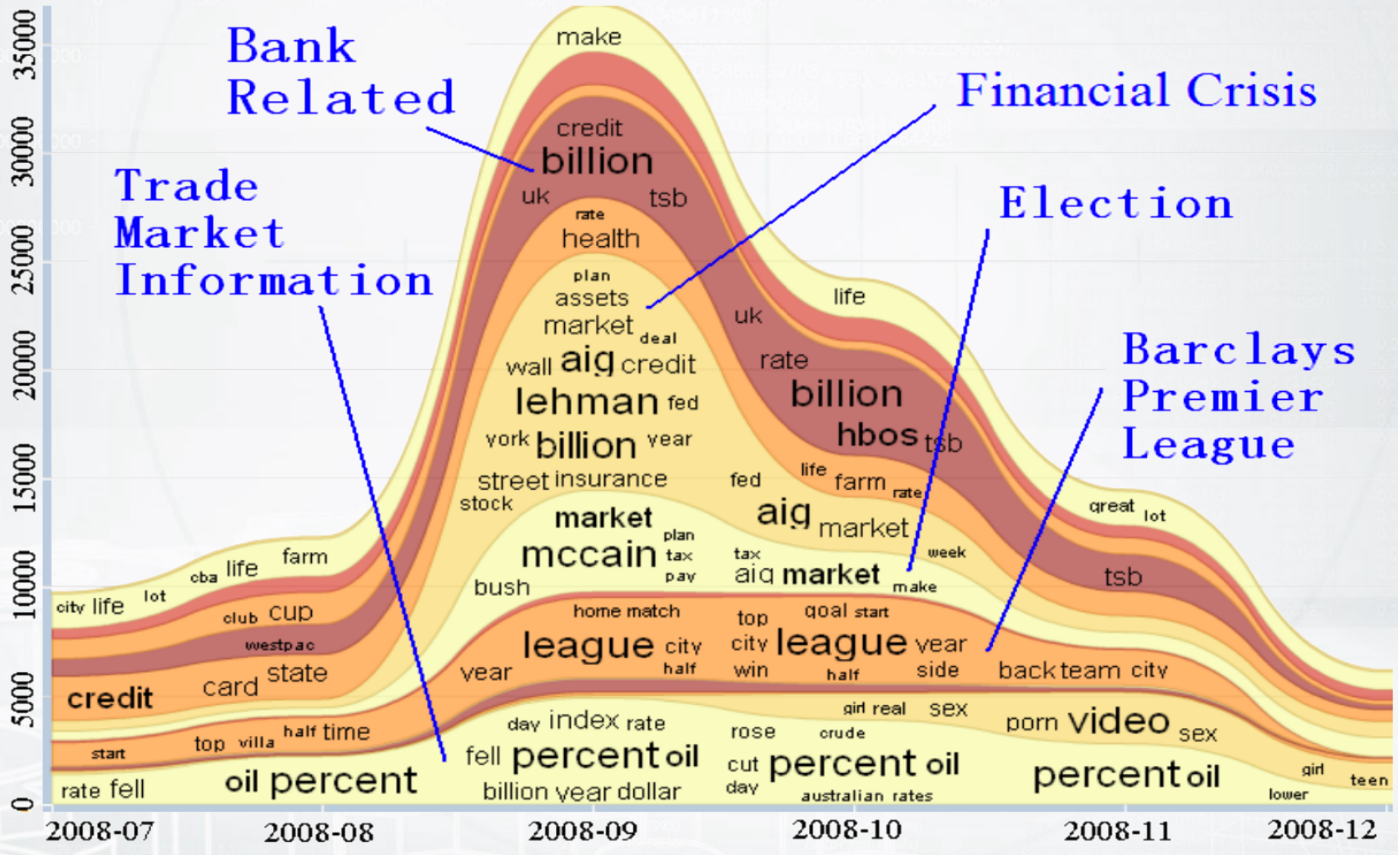
Dynamic topic models



David Blei, Probabilistic Topic Models, 2012.

Dynamic topic models

Topic detection and analysis of news flows:

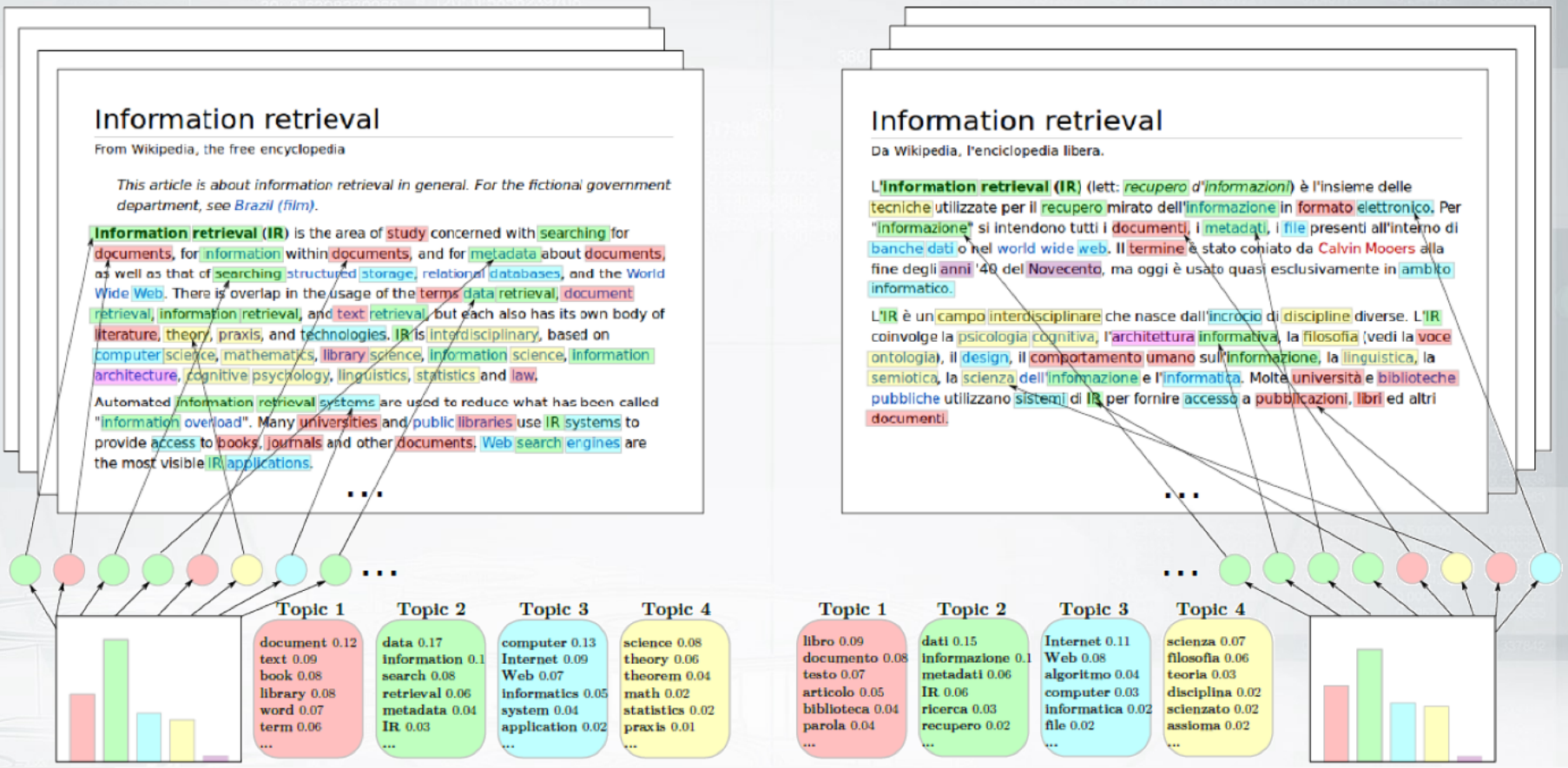


Jianwen Zhang, Yangqiu Song, Changshui Zhang, Shixia Liu Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying, KDD-2010.

Multilingual topic models

English corpus

Italian corpus



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?

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

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

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:

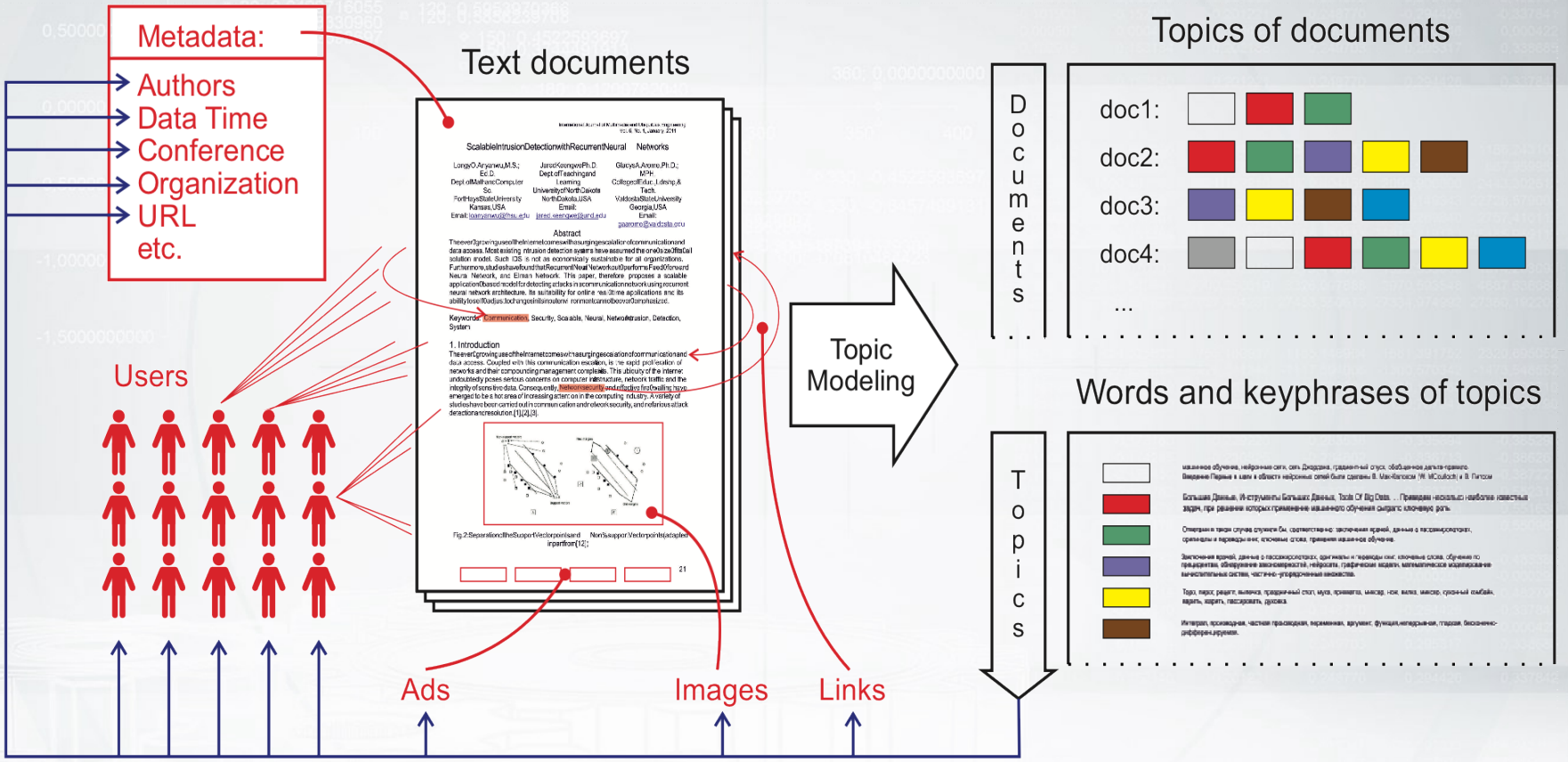
$$p(t|d,w) = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}}$$

M-step:

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

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

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

Multi-ARTM:

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

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

Inter-modality similarities

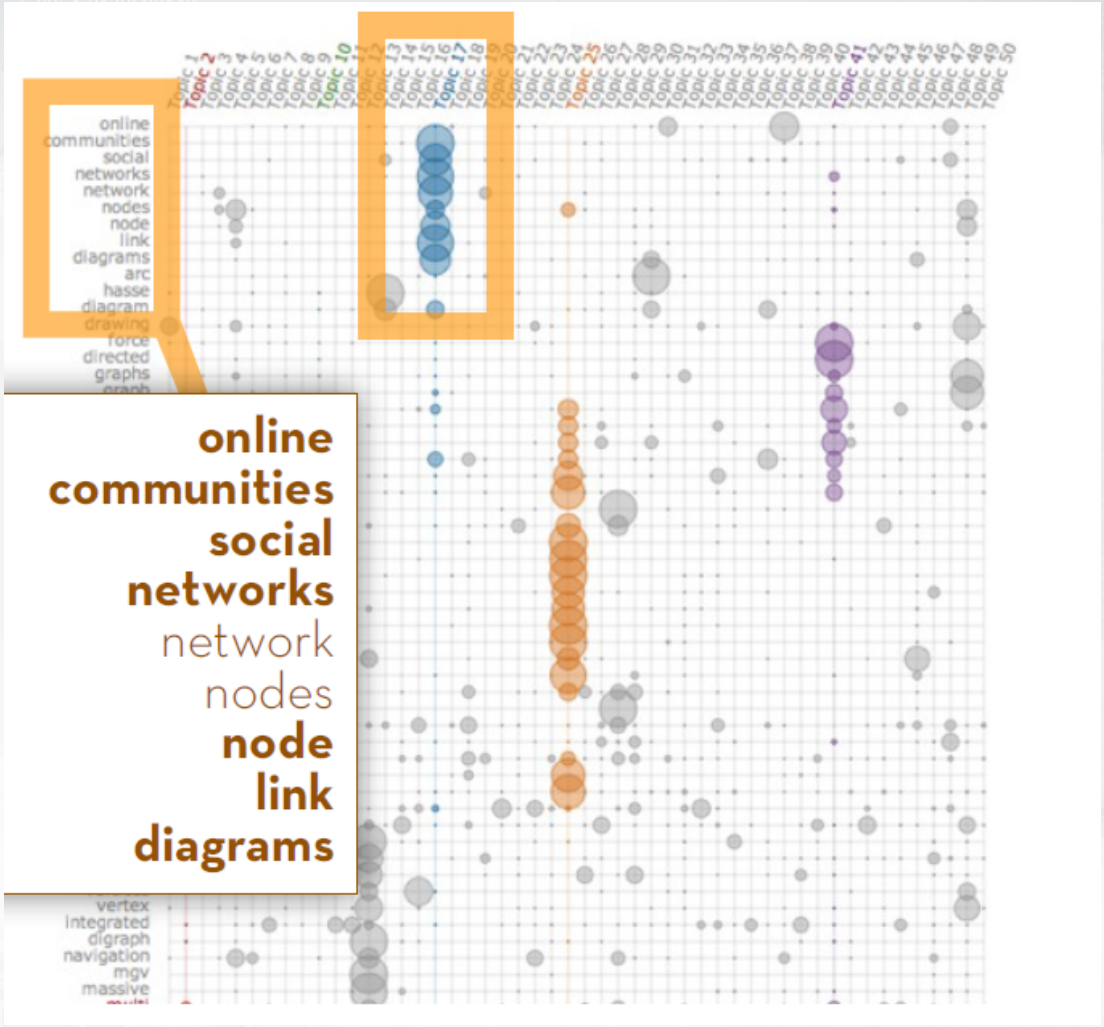
2015-12-18 Star Wars Release	2016-02-29 The Oscars	2015-05-09 Victory Day
jedi	statuette	great
sith	award	anniversary
fett	nomination	normandy
anakin	linklater	parade
chewbacca	oscar	demonstration
film series	birdman	vladimir
hamill	win	celebration
prequel	criticism	concentration
awaken	director	auschwitz
boyega	lubezki	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
- **Gensim** is a library of text analysis for Python, radimrehurek.com/gensim
- **MALLET** is a library of text analysis for Java mallet.cs.umass.edu
- **Vowpal Wabbit** has a fast implementation of online LDA 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