Modeling Topics from Text

March 20, 2013

Topics in Text

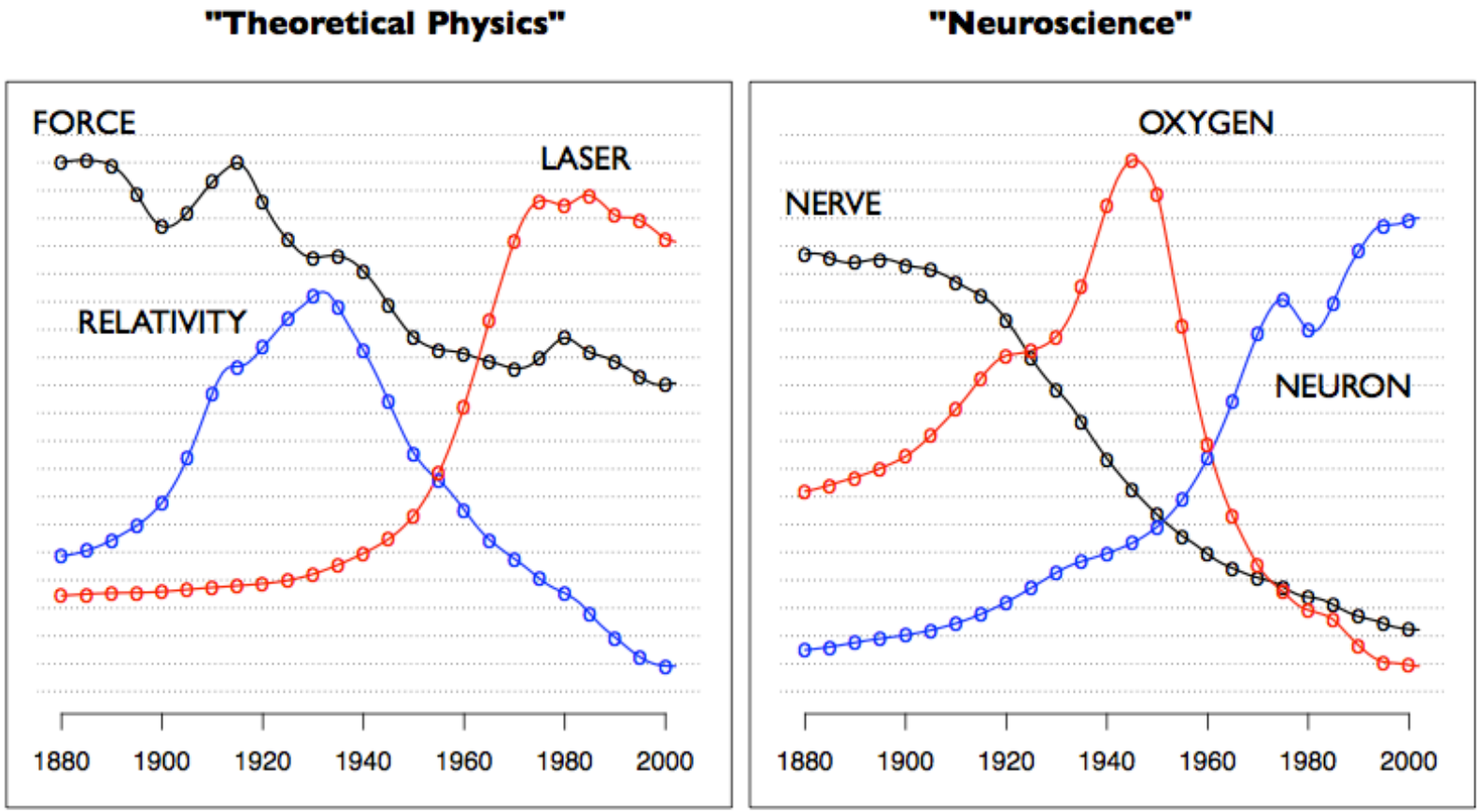
- Ideas, concepts, and meanings are (usually) organized thematically
  - Look at patterns of communication to understand these themes
  - Make simplifying assumptions about language and meaning to uncover the organizing features of documents
- Think about topics as ways people link ideas or concepts together across different political domains
  - E.g., Legislative debates, party platforms, advertisements, legal decisions, statutes, newspapers, magazines, academic journals, historical records...
  - Revealing behavior or attitudes through words
  - Topic models offer ways to organize, search and summarize information contained in strings of words

Topics in *Science*

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Source: Blei (2012)

Topics in *Science*



Source: Blei (2012)

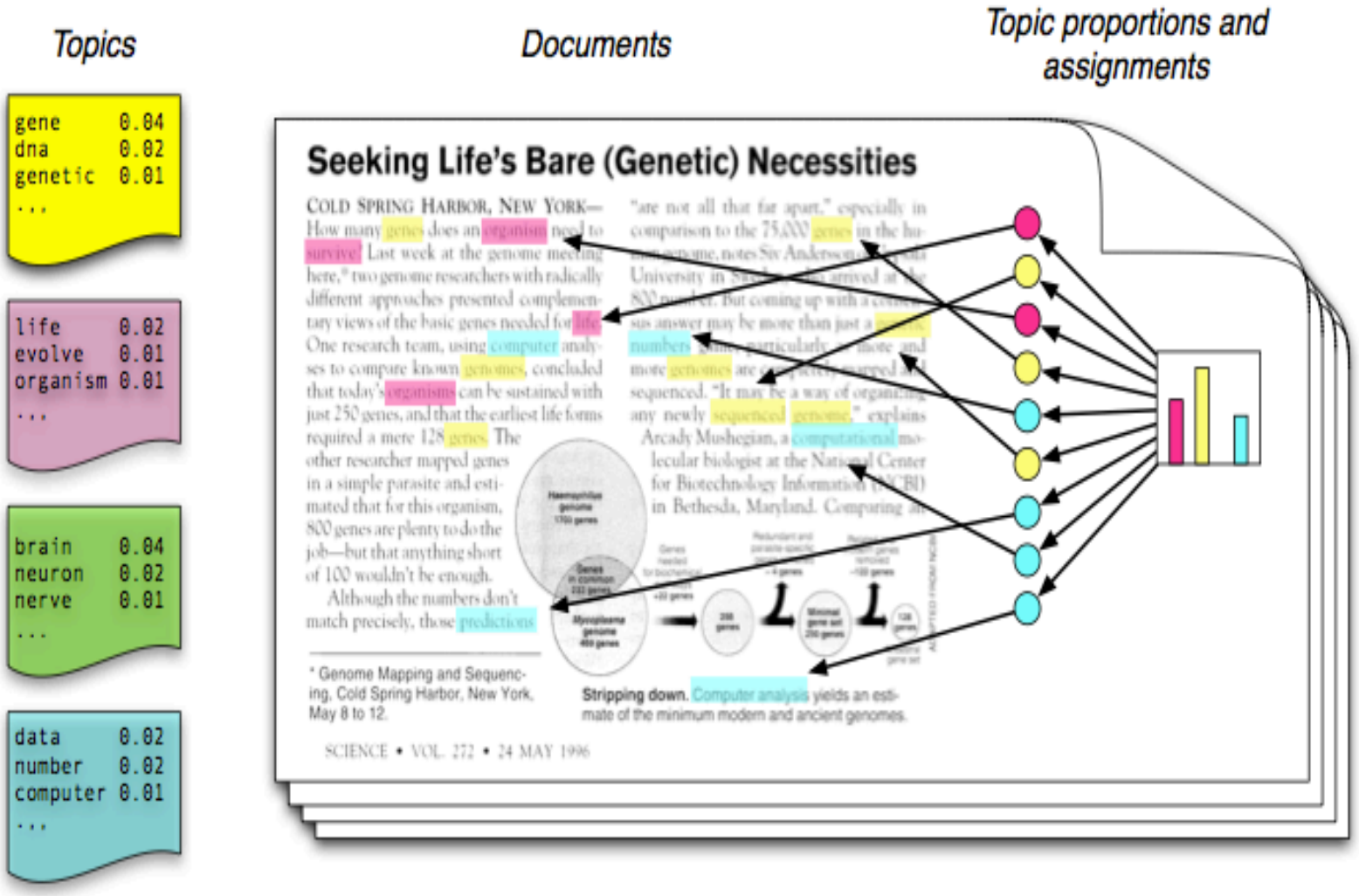
Partisan Topics in Campaigns

- 'Republican' Topics in Ads
  - cut taxes sales income death
  - health care special interests kids
  - republican reagan contra budget spending
  - welfare cut work reform contract
  - term limits perks reform raises
- 'Democratic' Topics in Ads
  - school education teachers class kids
  - social security gap seniors retire
  - care health doctors patients universal
  - drug crimes death criminals penalty
  - environment protecting oil economy jobs

LDA Topics Models

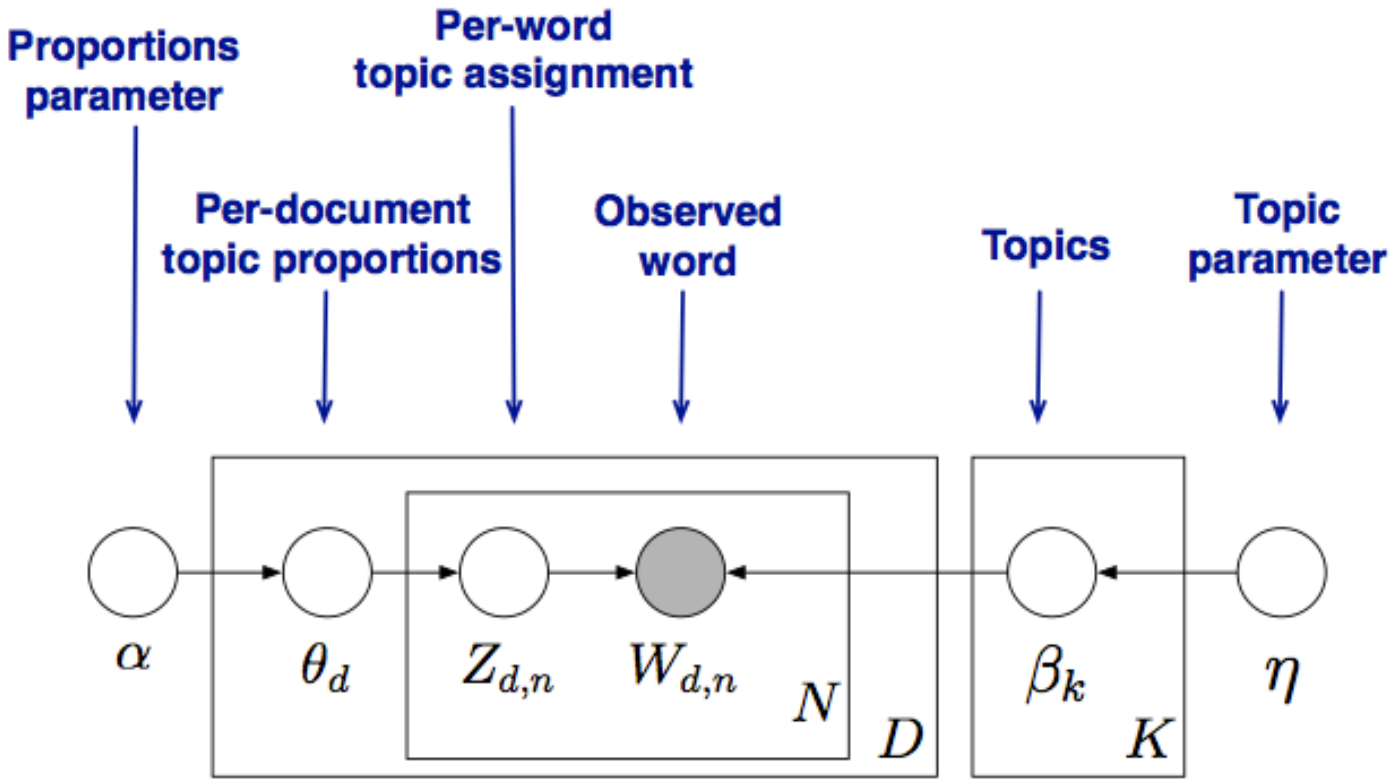
- Latent Dirichlet Allocation (LDA)
  - Each document contains a number of topics
  - Utilize the strings of words and phrases in each document to identify the K topics obtained across the entire corpus
- The LDA Model
  - A topic is a distribution over words (i.e., composed of words)
  - Each document is a mixture of topics
  - Each word is drawn from one of the topics

Topic Mixtures



Source: Blei (2012)

Graphical LDA



$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

Source: Blei (2012)

LDA Estimation

- The above model defines the joint posterior distribution
  - Estimation is fully (hierarchical) Bayesian
  - Mean field variational methods (Blei et al. 2001; Grimmer 2010)
- For D documents
  - Each word is assigned a vector for topic membership $z_{d,n}$
  - Each document d is assigned topic proportions θ_d (given words used)
  - Entire corpus of documents is assigned topic distributions β_k over K topics (given words and documents)

An Intuition for Why LDA Works

- Classifying commonly occuring words
  - Goal is to place words into the same 'bins' proportional to their co-occurance
  - Do this by maximizing the probability of observing a word w in a document given all the other words in d
- Word probabilities are maximized when dividing the words (widely) amongst the topics
  - More words means more density is spread across the topics
  - Allowing mixtures of topics for each document allows us to observe (maximize) a word-document probability
- Dirichlet priors for the multinomial topic mixtures 'encourage' sparsity
  - Documents are penalized for using too many topics
  - In probability, upweights tightly co-occuring words

References and Resources

- Blei, David. 2011. 'Probabilistic Topic Models': `www.cs.princeton.edu/~blei/kdd-tutorial.pdf`
- Blei, David and John Lafferty. 2009. 'Topic Models': `www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf`
- Gerrish, Sean and David Blei. 2010. 'The Ideal Point Topic Model': `http://people.cs.umass.edu/~wallach/workshops/nips2010css/papers/gerrish.pdf`
- Grimmer, Justin. 2010. 'An Introduction to Bayesian Inference via Variational Approximations': `www.stanford.edu/~jgrimmer/VariationalFinal.pdf`
- Some software: `www.cs.princeton.edu/~blei/topicmodeling.html`