Mining Text Data

(Lecture for SI671 – Data Mining: Methods and Applications)

October 20th, 2016

Instructor: V.G. Vinod Vydiswaran

vgvinodv@umich.edu

Department of Learning Health Sciences

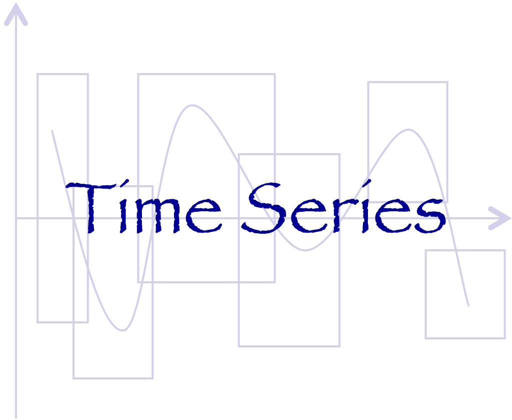
University of Michigan Medical School

Some slides are adopted from Prof. Qiaozhu Mei's lectures

Views of Data Formulation

Itemsets

Matrix



Time Series

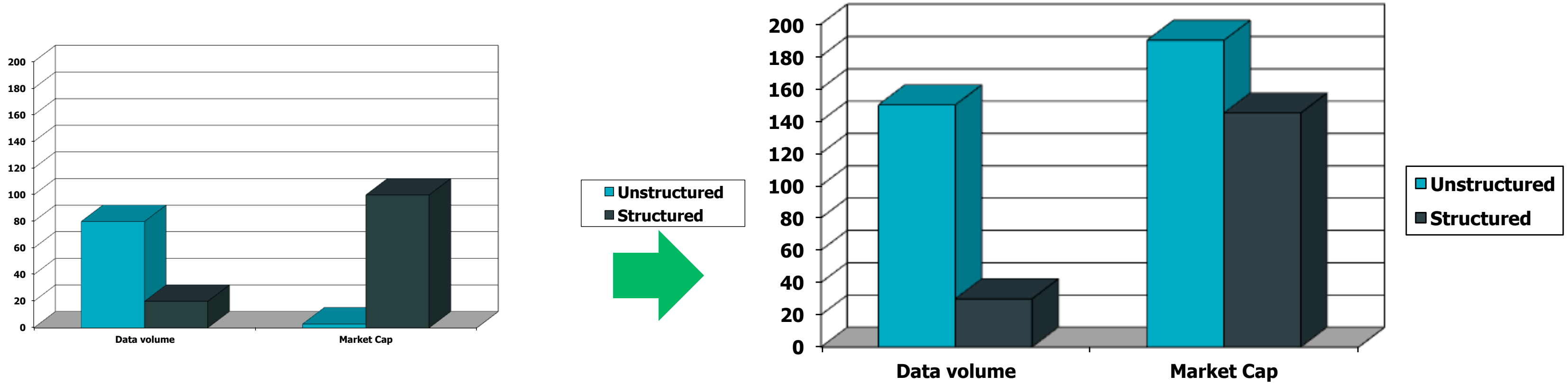
SYNTENIC ASSEMBLIES FOR CG15386

Sequences

Networks

Stream

The Power of Text Data



1996

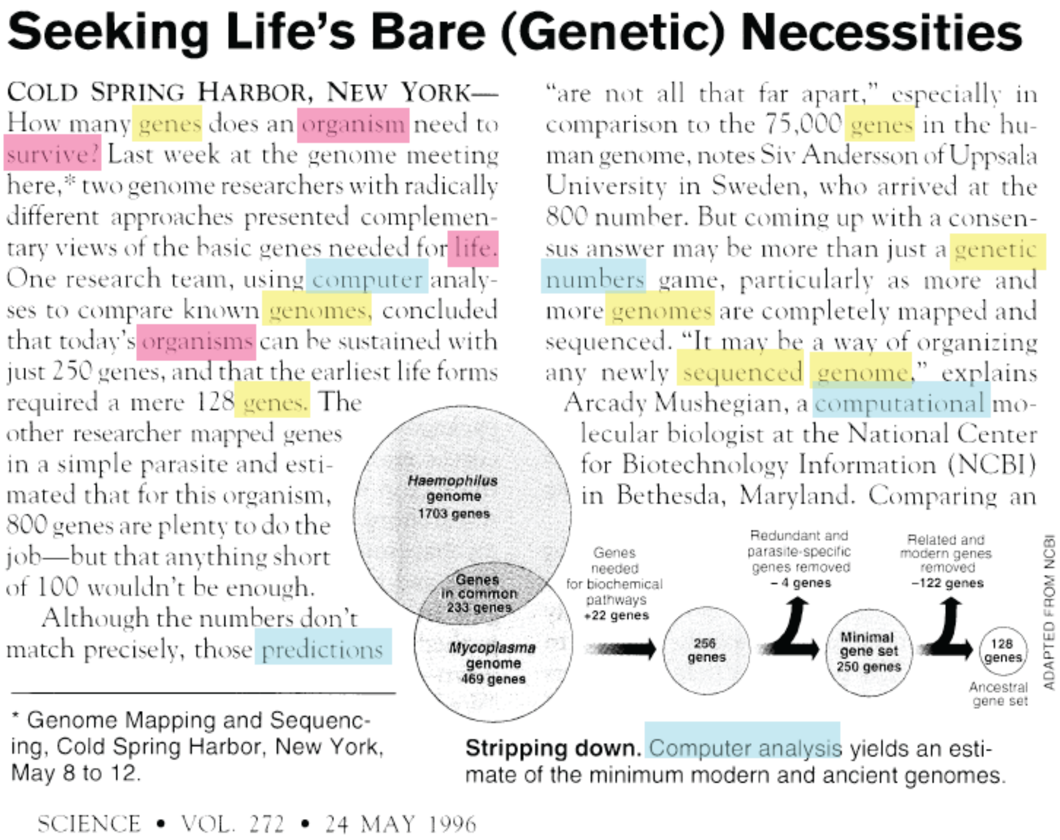
2009

- Figure from Chris Manning

Research Problems in Text Data Mining

- Similarity
- Text classification
- Text clustering
- Labeling, e.g., entity extraction
- Ranking, e.g., search engines
- Topic modeling
- Sentiment analysis
- Contextual analysis
- …

Probabilistic Topic Modeling



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Latent Dirichlet Allocation
(Blei et al. 03)

Topic 1: Genetics
gene, sequence, genome, molecule, …

Topic 2: Computation
computer, predictions, analysis, number, …

Topic 3: Life science
life, organism, survive, biology, …

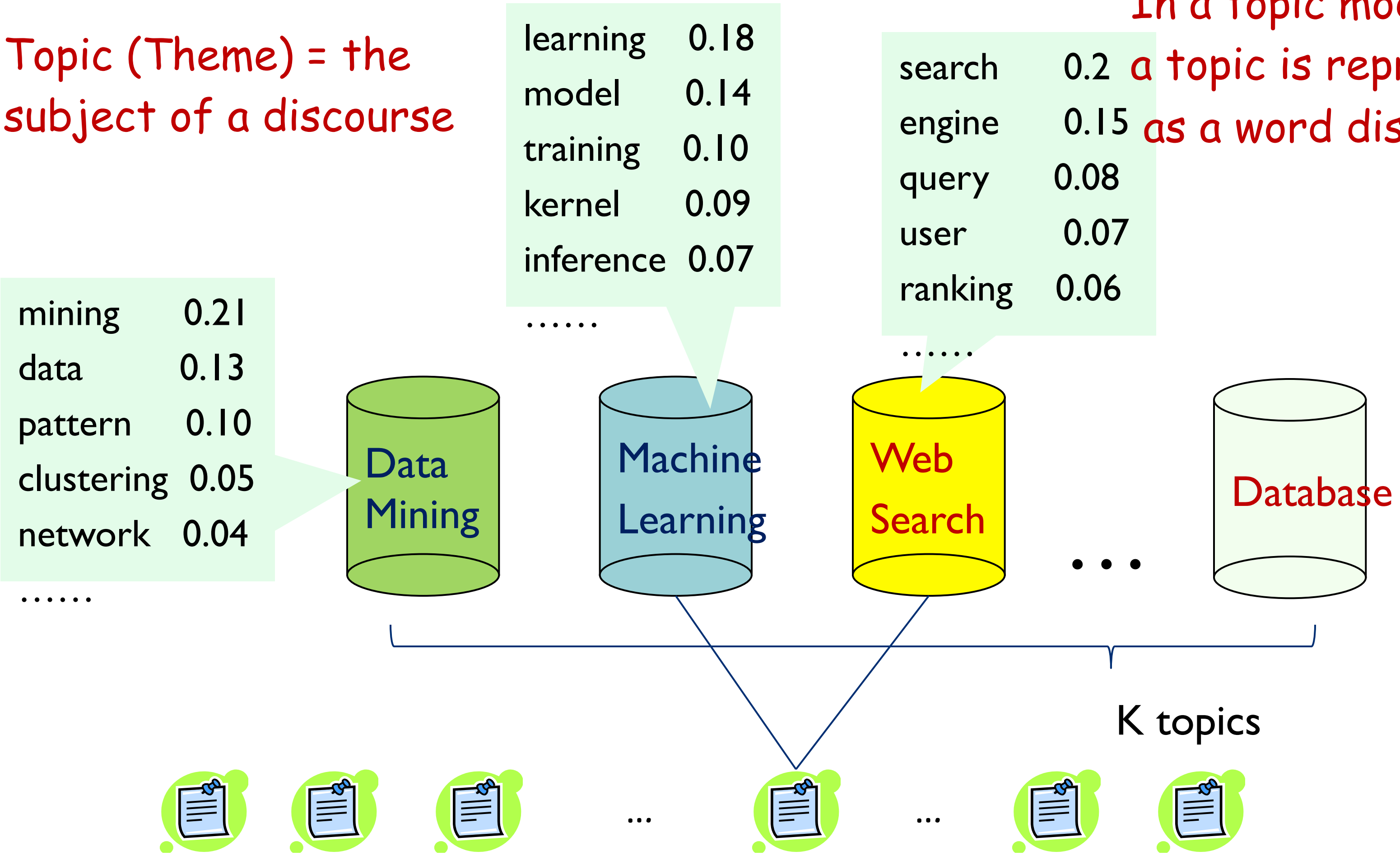
- Figure from David Blei's slide

Simple intuition: Documents exhibit multiple topics.

Text as a Mixture of Topics

Topic (Theme) = the subject of a discourse

In a topic model, a topic is represented as a word distribution



mining	0.21
data	0.13
pattern	0.10
clustering	0.05
network	0.04
……	

learning	0.18
model	0.14
training	0.10
kernel	0.09
inference	0.07
……	

search	0.2
engine	0.15
query	0.08
user	0.07
ranking	0.06
……	

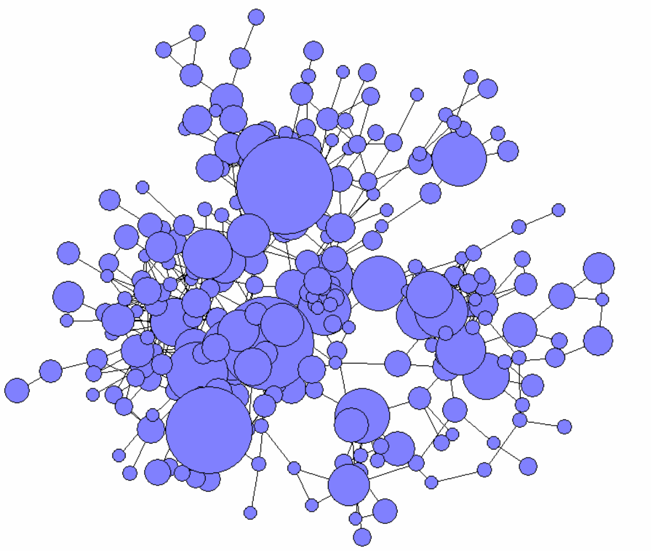
K topics

Using machine learning for web search

Example of Topic Modeling

Topic 1		Topic 2		Topic 3		Topic 4	
term	0.02	peer	0.02	visual	0.02	interface	0.02
question	0.02	patterns	0.01	analog	0.02	towards	0.02
protein	0.01	mining	0.01	neurons	0.02	browsing	0.02
training	0.01	clusters	0.01	vlsi	0.01	xml	0.01

Can we recover the four topics and assign papers/ authors to topics?



Papers of authors collected from
4 Conferences:
SIGIR, KDD, NIPS, WWW

Results of Topic Modeling

Topic 1		Topic 2		Topic 3		Topic 4	
retrieval	0.13	mining	0.11	neural	0.06	web	0.05
information	0.05	data	0.06	learning	0.02	service	0.03
document	0.03	discovery	0.03	networks	0.02	semantic	0.03
query	0.03	databases	0.02	recognition	0.02	services	0.03
text	0.03	rules	0.02	analog	0.01	peer	0.02
search	0.03	association	0.02	vlsi	0.01	ontologies	0.02
evaluation	0.02	patterns	0.02	neurons	0.01	rdf	0.02
user	0.02	frequent	0.01	gaussian	0.01	management	0.01
relevance	0.02	streams	0.01	network	0.01	ontology	0.01

Word distributions present 4 topics

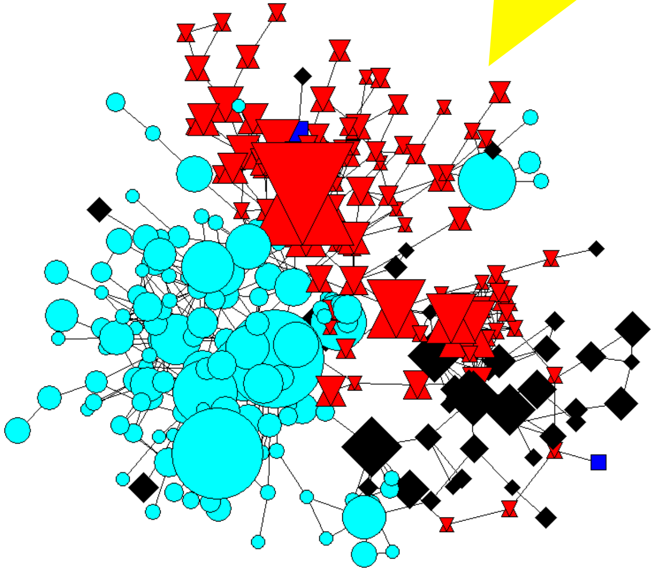
Web

Information Retrieval

Data mining

Machine learning

Documents assigned to topics



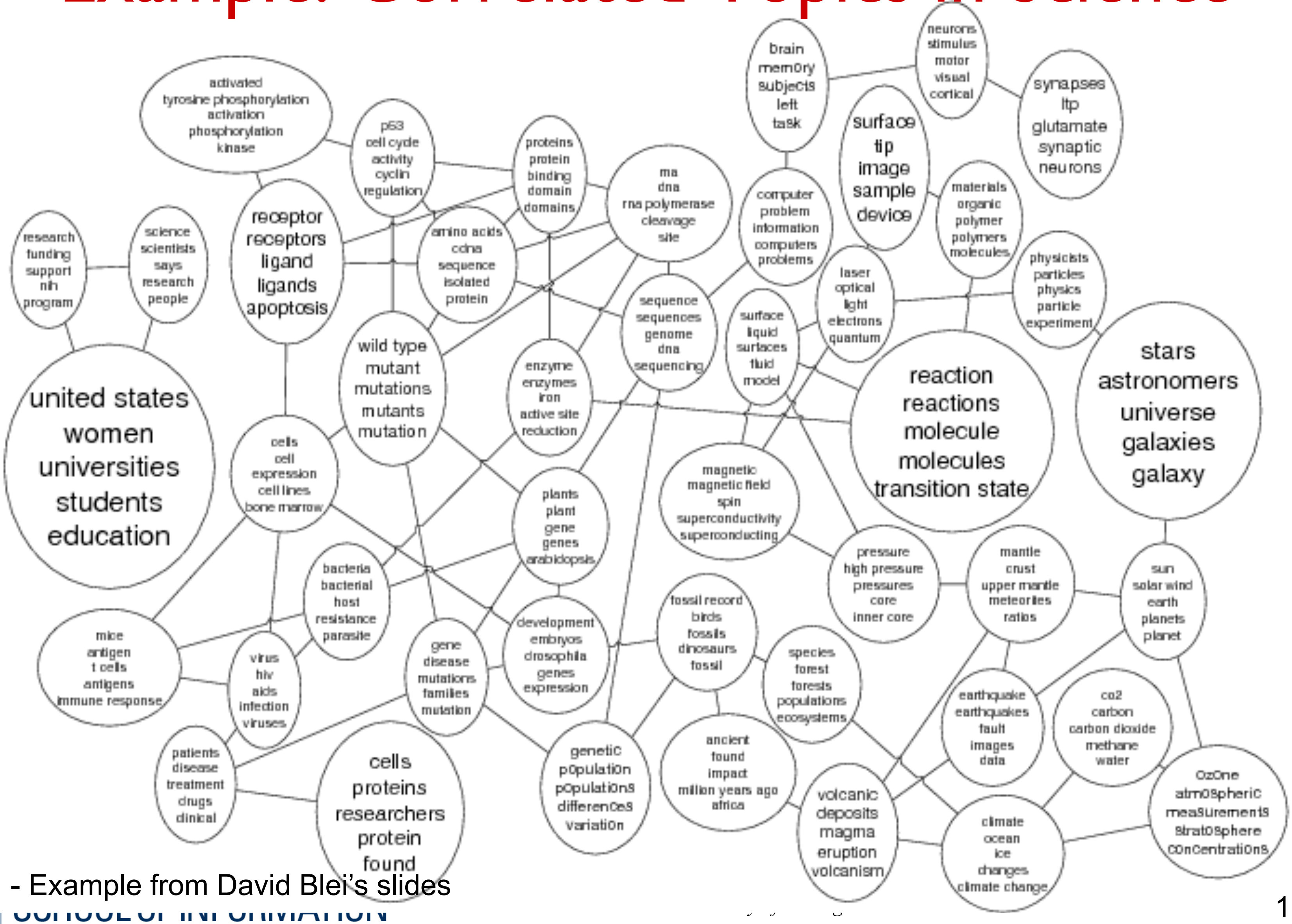
- Mei et al., 2008. Topic modeling with network regularization

Example: More Topics

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

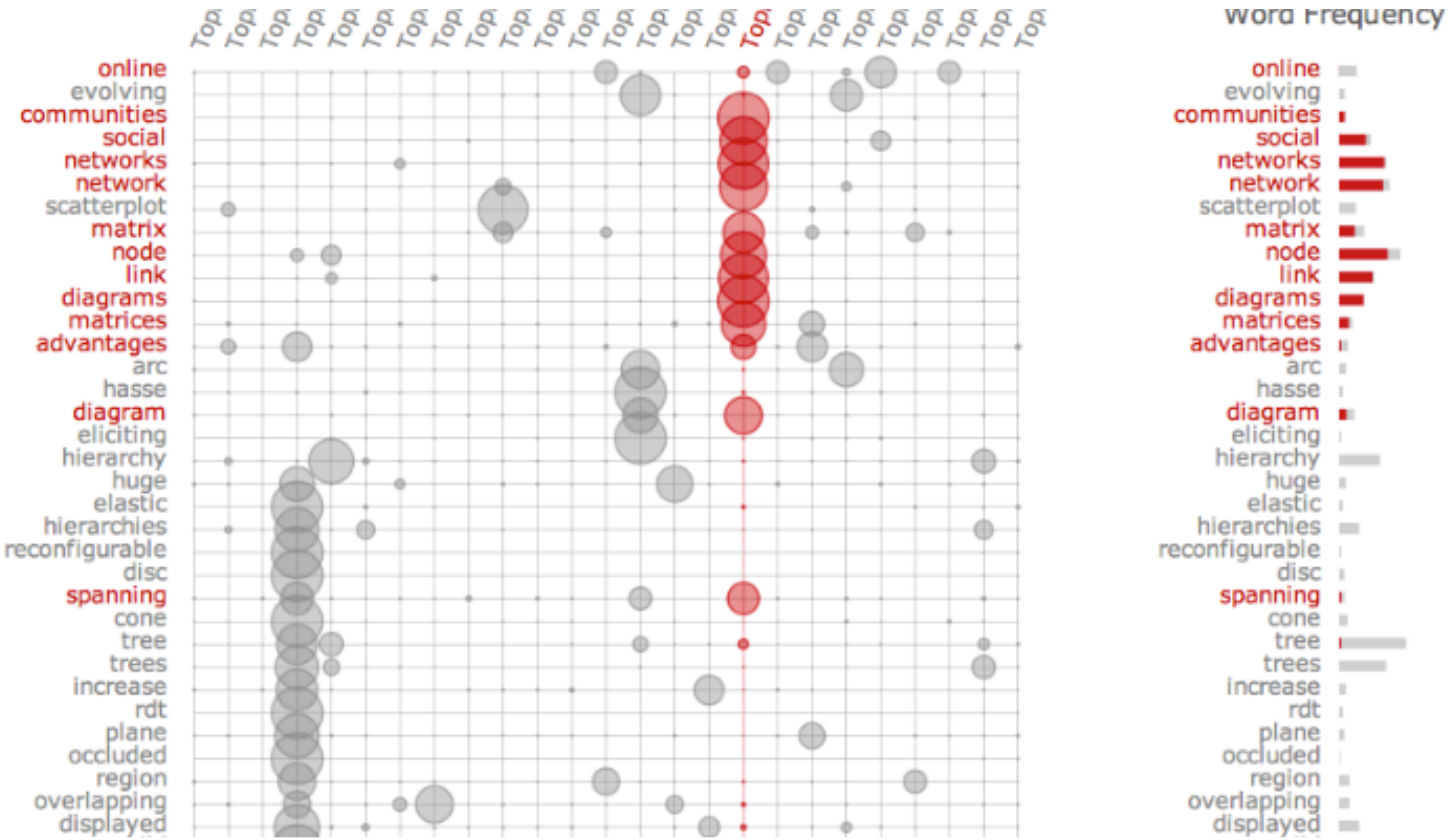
- Example from David Blei's slides

Example: Correlated Topics in Science


activated tyrosine phosphorylation activation phosphorylation kinase
p53 cell cycle activity cyclin regulation
proteins protein binding domain domains
rna dna rna polymerase cleavage site
brain memory subjects left task
neurons stimulus motor visual cortical
synapses ltp glutamate synaptic neurons
surface tip image sample device
computer problem information computers problems
materials organic polymer polymers molecules
research funding support nih program
science scientists says research people
receptor receptors ligand ligands apoptosis
amino acids cdna sequence isolated protein
sequence sequences genome dna sequencing
surface liquid surfaces fluid model
laser optical light electrons quantum
physicists particles physics particle experiment
united states women universities students education
wild type mutant mutations mutants mutation
enzyme enzymes iron active site reduction
reaction reactions molecule molecules transition state
stars astronomers universe galaxies galaxy
cells cell expression cell lines bone marrow
plants plant gene genes arabidopsis
magnetic magnetic field spin superconductivity superconducting
pressure high pressure pressures core inner core
mantle crust upper mantle meteorites ratios
sun solar wind earth planets planet
mice antigen t cells antigens immune response
bacteria bacterial host resistance parasite
virus hiv aids infection viruses
gene disease mutations families mutation
development embryos drosophila genes expression
fossil record birds fossils dinosaurs fossil
species forest forests populations ecosystems
earthquake earthquakes fault images data
co2 carbon carbon dioxide methane water
patients disease treatment drugs clinical
cells proteins researchers protein found
genetic population populations differences variation
ancient found impact million years ago africa
volcanic deposits magma eruption volcanism
climate ocean ice changes climate change
ozone atmospheric measurements stratosphere concentrations

- Example from David Blei's slides

Example: Visualization of Topics



Source- http://vis.stanford.edu/papers/termite

Why Bother?

- Exploratory analysis – a coarse-level analysis of what's in a given text collection
  - Essentially a text clustering problem
  - Cluster documents and words simultaneously
- Why not K-means or latent semantic indexing?
  - Relation to K-means: K clusters, word distribution of a topic is similar to the center of clusters
  - Problem: a document usually contain multiple topics
  - Relation to LSI: clustering documents/words simultaneously, K "soft" clusters.
  - Problem: no probabilistic interpretation

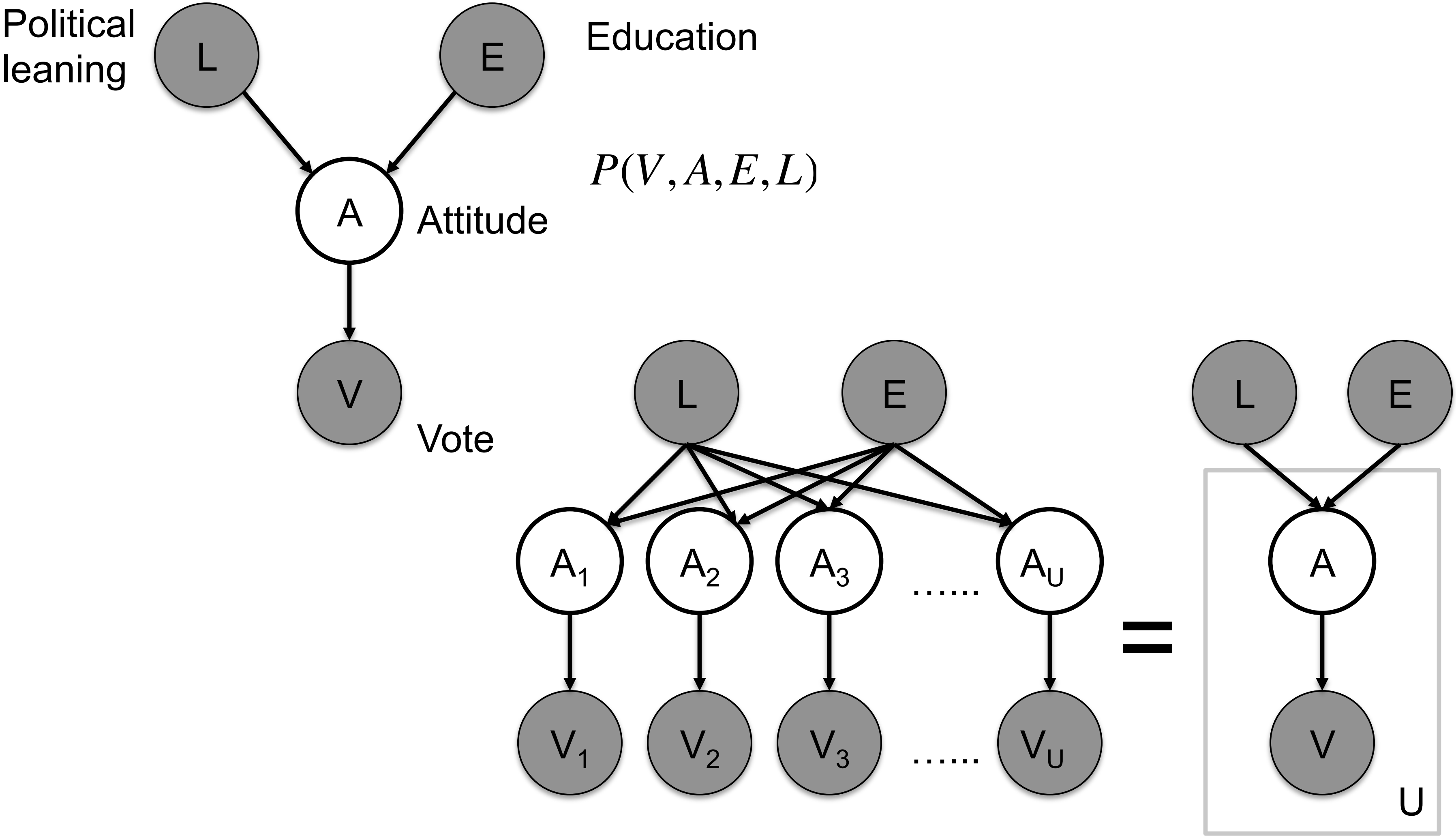
Terminology

- Latent variables
  - Random variables that are not directly observable, but can be inferred from other observable variables
  - E.g., market sentiment, user interest, topics, …
  - Latent variable model: probabilistic model with latent variables
- Generative model
  - A probabilistic model for the observable data, X.
  - Joint distribution of data and labels (e.g., P(X, Y))
  - In contrast to discriminative models (e.g., P(Y|X))
  - There is always a “generative process.”

Terminology

- Graphical models
  - a probabilistic model with a graph to denote the conditional independence structure between random variables
  - Directed: Bayesian Networks (acyclic)
  - Undirected: Markov random fields (Markov networks)
- Mixture models
  - a mixture distribution of the presence of sub-populations within an overall population

Example: Directed Graphical Model



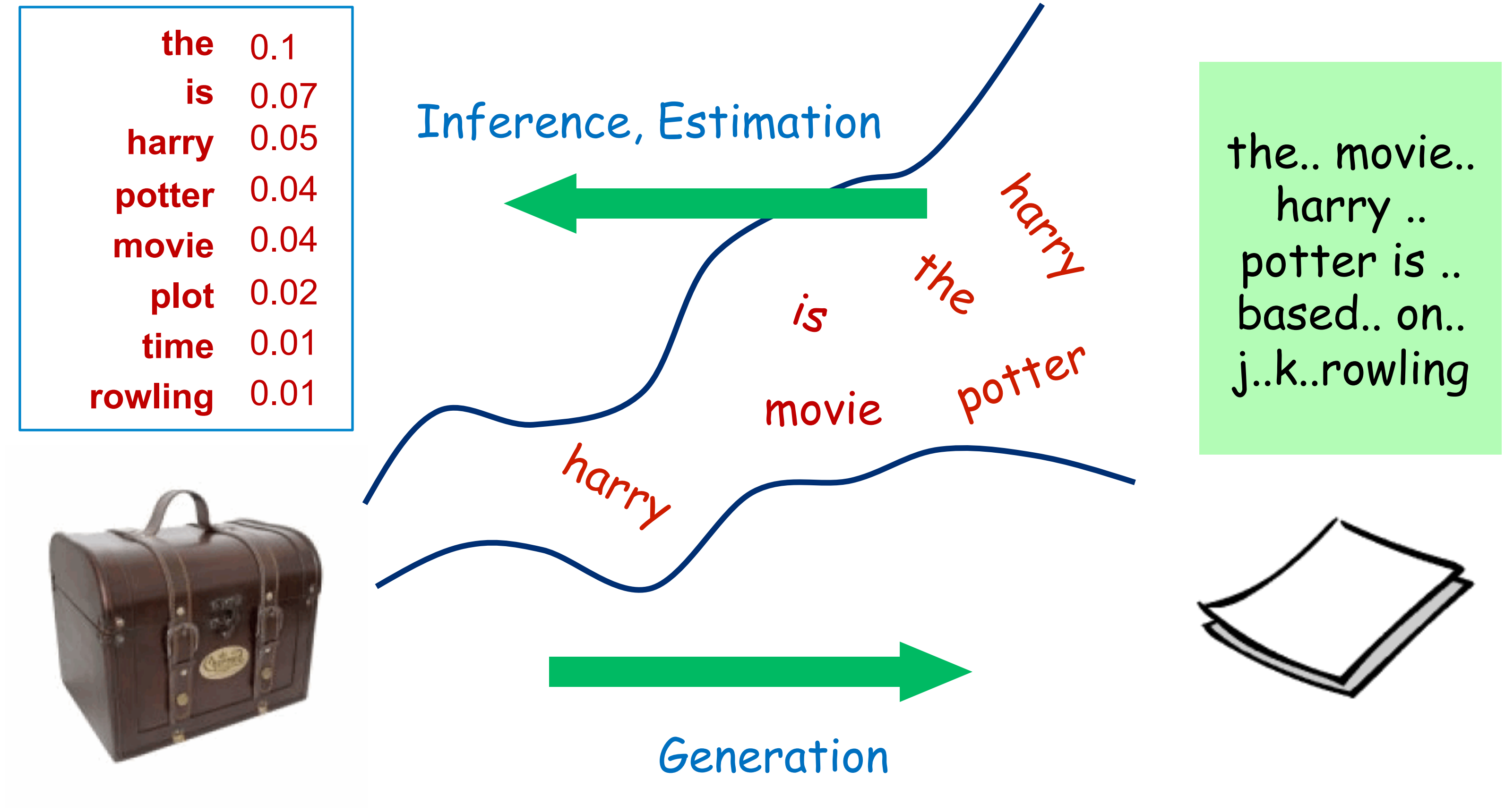
$P(V, A, E, L)$

Probabilistic Topic models

- Probabilistic topic model:
  - A special case of a directed graphical model
  - A generative model of text data, with latent variables corresponding to the hidden topics in text
  - An extension of mixture models (a document to be a “mixture” of different topics)
- Instances of probabilistic topic models:
  - Probabilistic latent semantic analysis (PLSA) [Hofmann, 1999]
  - Latent Dirichlet Allocation (LDA) [Blei et al., 2003]
  - Many, many others…

Generative Model of Text

$$P(Text \mid Model)$$



the	0.1
is	0.07
harry	0.05
potter	0.04
movie	0.04
plot	0.02
time	0.01
rowling	0.01

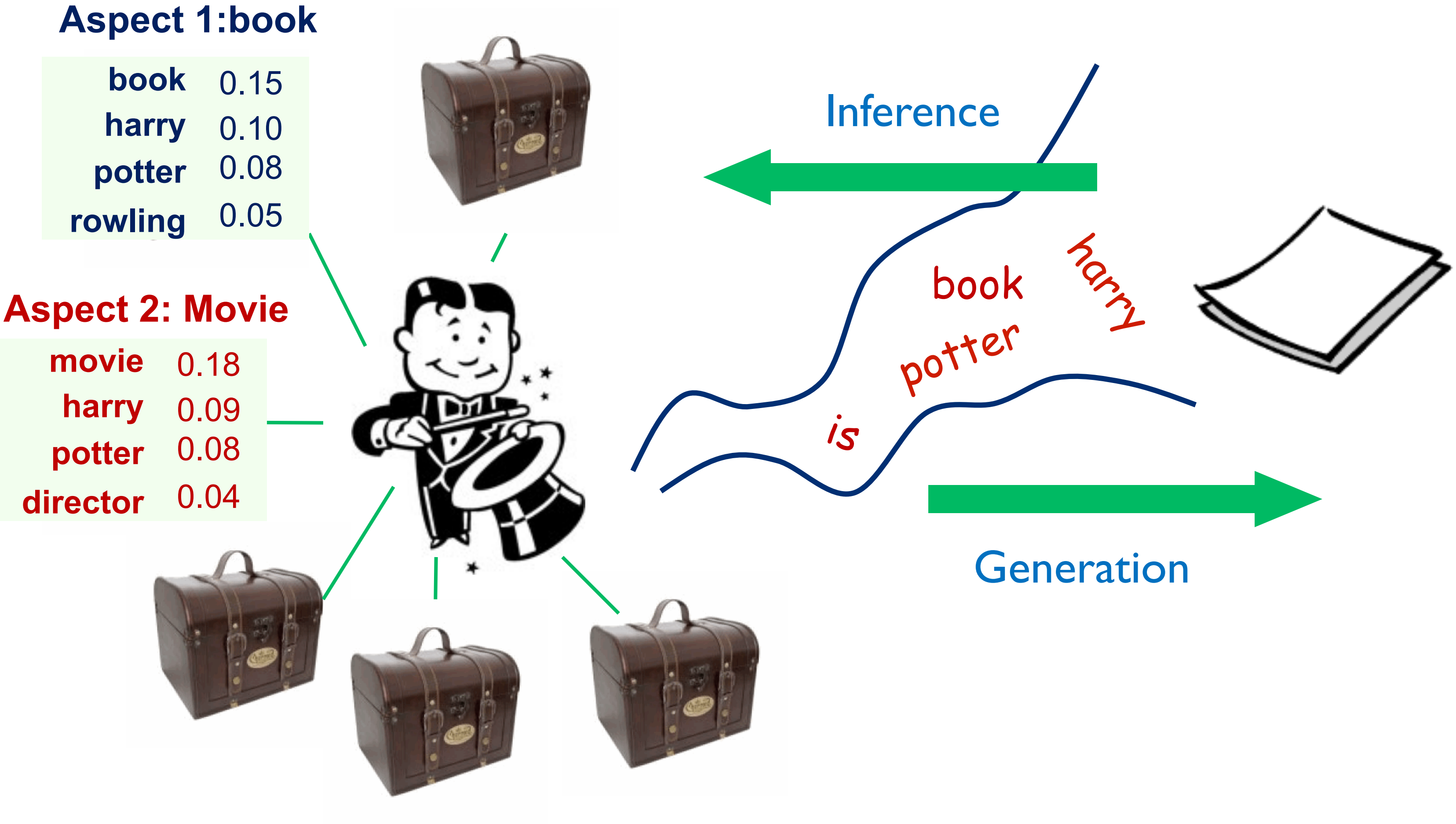
Inference, Estimation

the.. movie..
harry ..
potter is ..
based.. on..
j..k..rowling

Generation

A Generative Model can be Much More Complicated (e.g., a Mixture Model)

$P(\text{Text} \mid \text{Model})$



Aspect 1:book

book	0.15
harry	0.10
potter	0.08
rowling	0.05

Aspect 2: Movie

movie	0.18
harry	0.09
potter	0.08
director	0.04

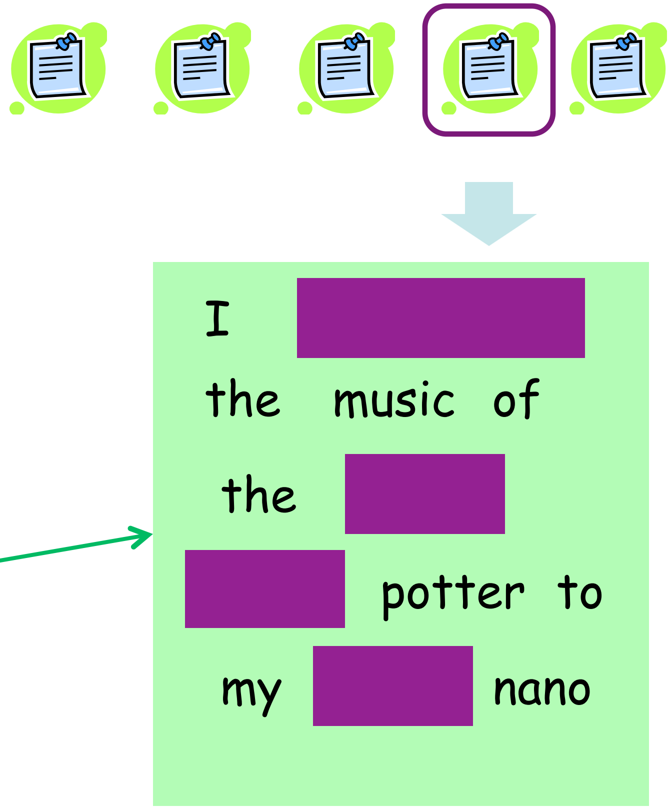
The Simplest Generative Model: Unigram Model

- Only one topic in the entire corpus, or one topic in each document
- Generative process:
- For each document d:
  - For each word token w_n in d:
  - Choose a word w according to the multinomial distribution P(w).

Simple Unigram – Generative Process

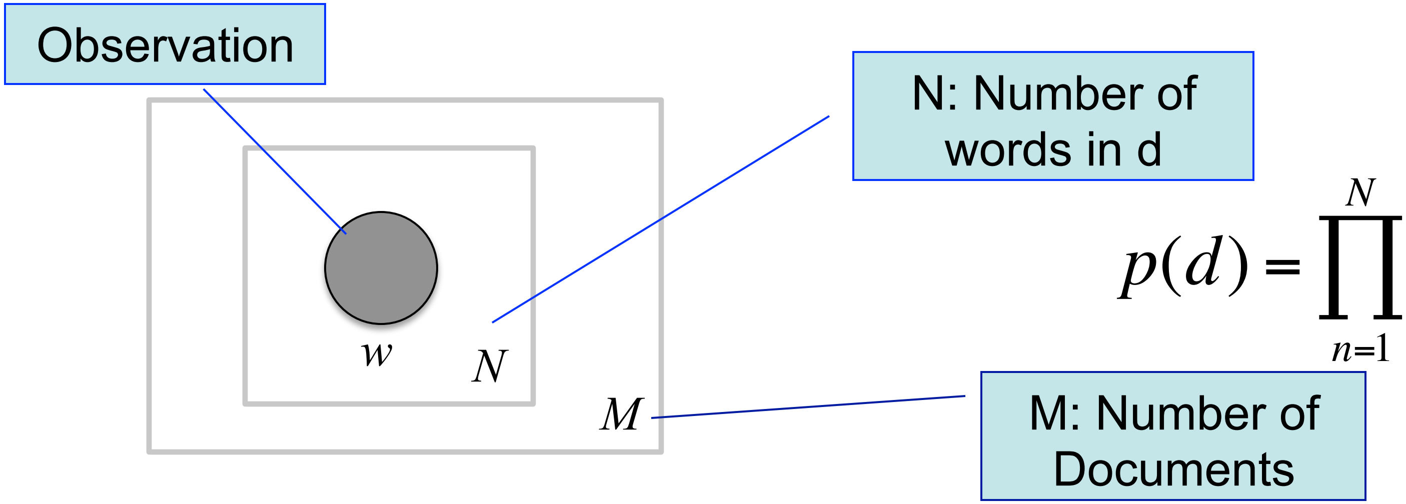
$$P(d) = \prod_{w \in d} P(w)$$

movie	0.10
harry	**0.09**
potter	0.05
ipod	**0.01**
music	0.02



I ▇
the music of
the ▇
▇ potter to
my ▇ nano

Graphical Structure of Unigram Models



Observation

N: Number of words in d

M: Number of Documents

$$p(d) = \prod_{n=1}^{N} p(w_n)$$

- No dependency among variables
- Parameter: p(w)
- Parameter estimation:
  - Maximum likelihood estimation (MLE)
- Applications:
  - language modeling based information retrieval

Mixture of Unigrams

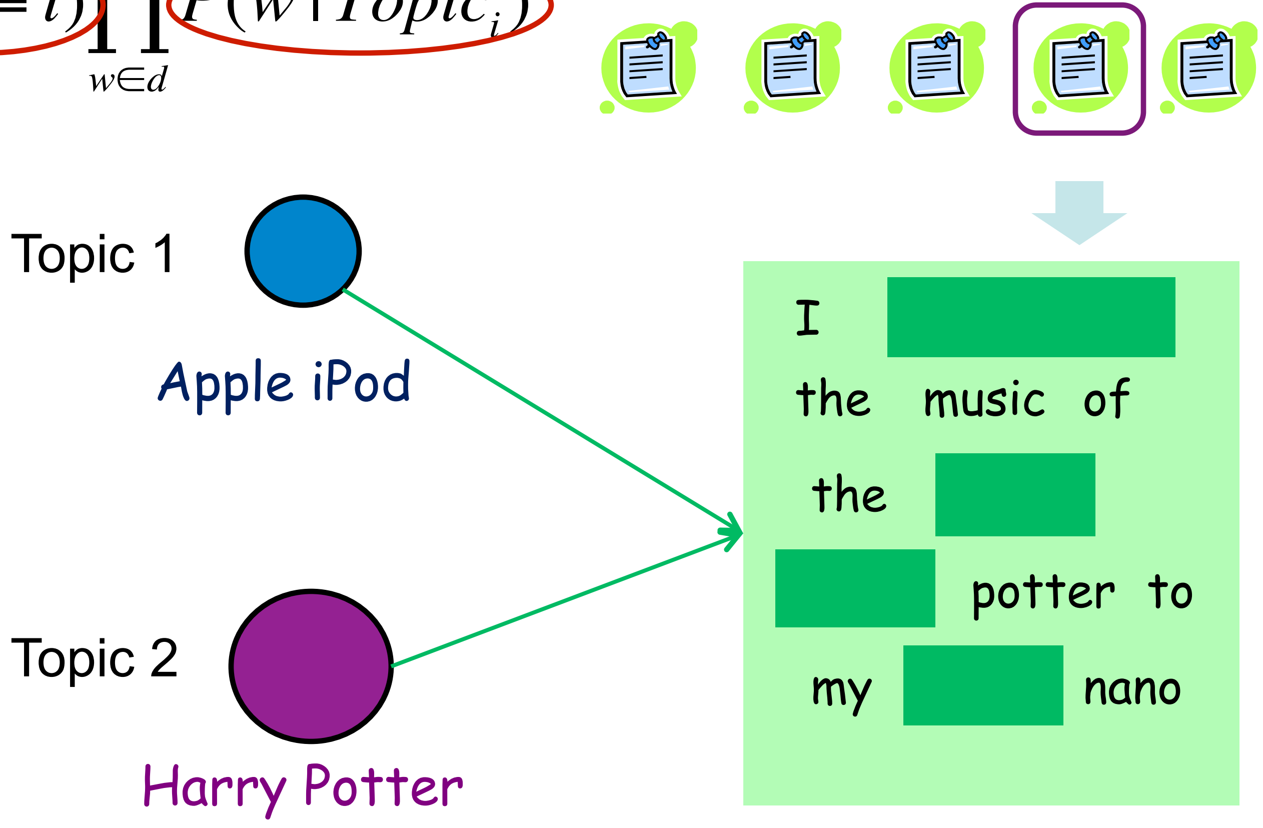
- Allow multiple topics in a corpus.
- The generative model of text becomes a mixture model
- Generative process:
  - For every document d
    - Choose a topic z according to a distribution P(z)
    - For every word token in d
      - Choose a word w_n according to word distribution P(w|z)

Mixture of Unigrams – Generative Process

$$P(d)=\sum_{i=1..K} P(z=i)\prod_{w\in d} P(w\,|\,Topic_i)$$

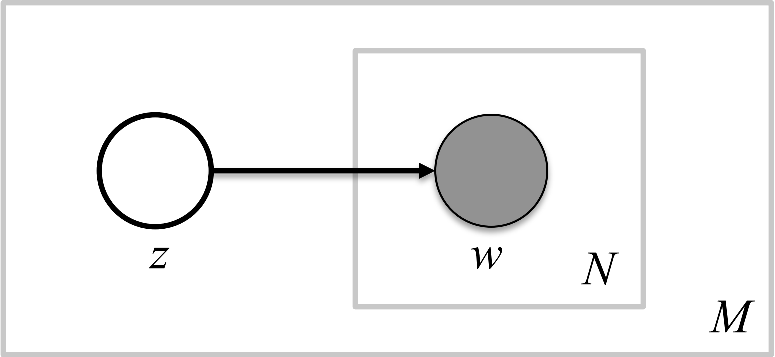
ipod	0.15
nano	0.08
music	0.05
download	0.02
apple	0.01

movie	*0.10*
harry	***0.09***
potter	*0.05*
ipod	***0.01***
music	*0.02*



Mixture of Unigrams – Graphical Model

- There are *k* topics in the collection, but each document only cover one topic



$$p(d) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n | z)$$

- Estimation: MLE and EM Algorithm
- Application: simple document clustering

Mixture of Unigram: Pros and Cons

- Allow each document to select among multiple identities (topics)
- But once decided, all words in the document share the identity of the document.
- Essentially a probabilistic version of K-means.
- Simple and easy to estimate
- Performs well on short documents, e.g., tweets
- Cons:
  - Usually different parts of a document present different topics
  - Desirable: words in a document to take different identities

Probabilistic Latent Semantic Analysis (Hofmann, 1999)

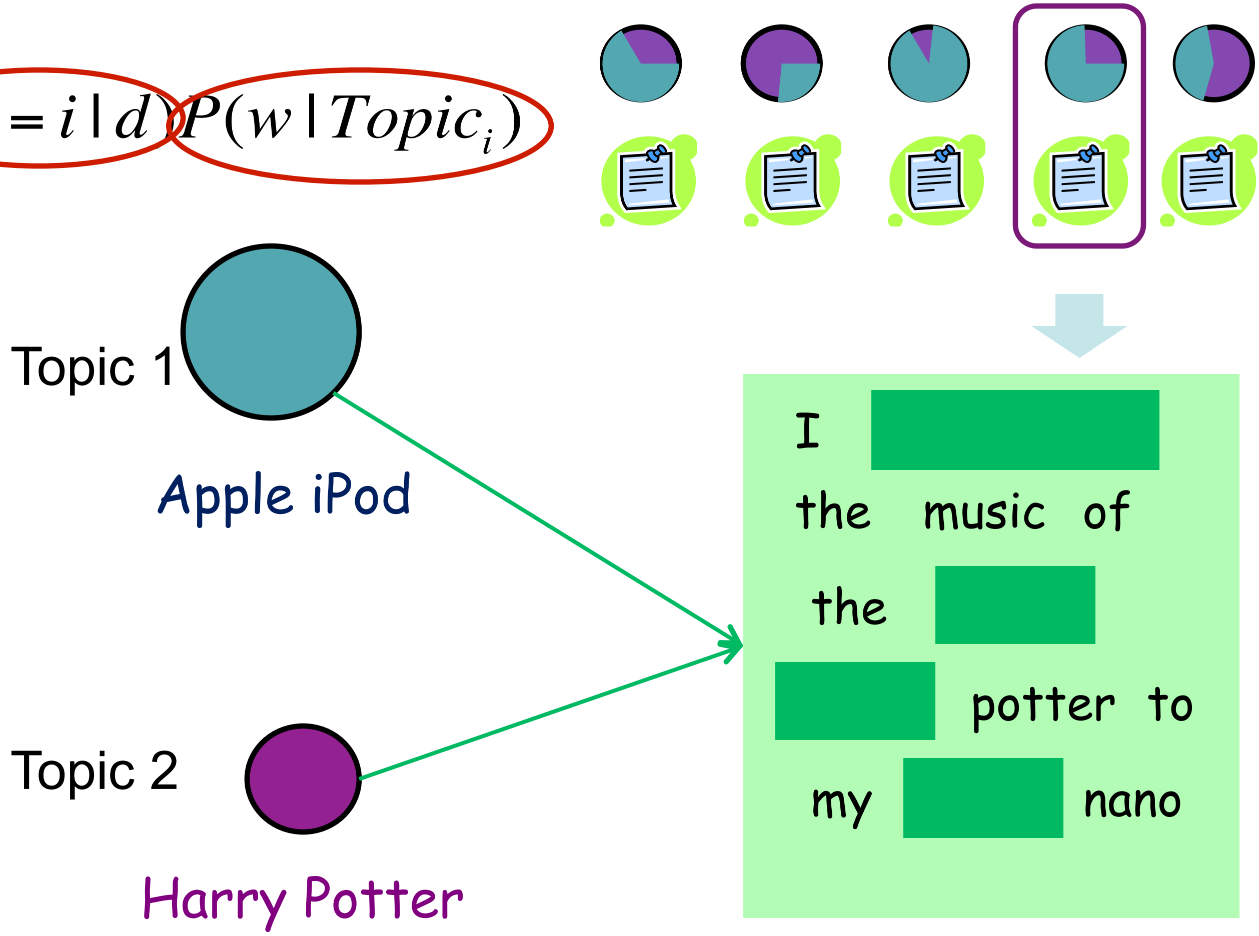
- Also known as probabilistic latent semantic indexing (PLSI)
- Generative process:
  - For every document d
    - For every word token in d
      - Choose a topic z_n according to P(z|d)
      - Choose a word w_n according to $P(w|z_n)$
  - Both P(z|d) and P(w|z) are fixed but unknown parameters – to be estimated

PLSA – Generative Process

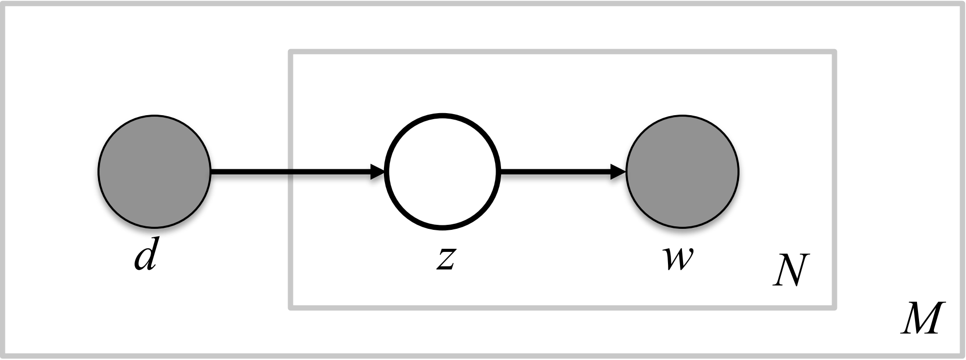
$$P(d) = \prod_{w \in d} \sum_{i=1..K} P(z = i \mid d) P(w \mid Topic_i)$$

ipod	**0.15**
nano	0.08
music	0.05
download	0.02
apple	0.01

movie	*0.10*
harry	***0.09***
potter	*0.05*
ipod	*0.01*
music	*0.02*



Graphical Model of PLSA



$$p(w_n, d) = p(d) \sum_{z} p(w_n \mid z) p(z \mid d)$$

- Assume uniform distribution of p(d)
- Parameters: P(w|z), P(z|d)
- Parameter estimation: MLE, Expectation-Maximization

PLSA: Pros and Cons

- One of the most effective topic models
- Allow a document to present multiple identities (topics)
- Allow words in the same document to take different identities. Word identities are affected, but not forced to be identical to the document's identity.
- A "soft" version of K-means
- Cons:
  - Not a complete Bayesian model: hard to interpret unseen documents
  - Overfits the data (because number of parameters grows linearly with the number of documents).

Latent Dirichlet Allocation
(Blei&Ng&Jordan, 2003)

- The Bayesian version of PLSA
- Treats the document-topic mixture weights as a k-parameter hidden random variable (a multinomial).
- Places a Dirichlet prior on all the multinomial mixing weights. Sample the topic mixture weights once per document.
- The weights for word multinomial distributions are still considered as fixed parameters to be estimated.
- For a fuller Bayesian approach, can place a Dirichlet prior to these word multinomial distributions to smooth the probabilities.

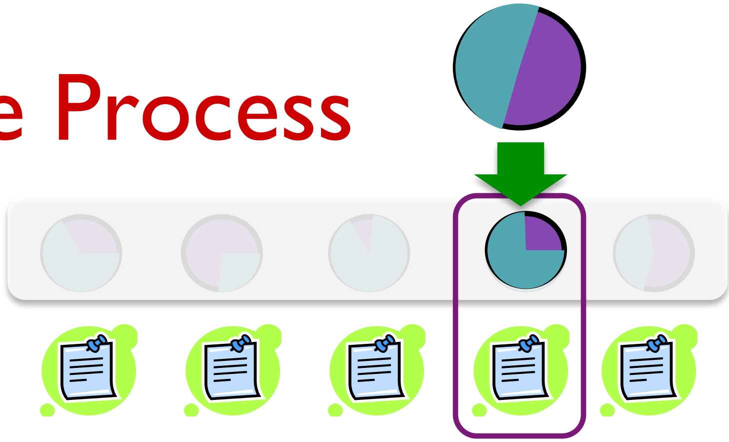
LDA – Generative Process

- For each document d:
  - Choose document length $N \sim Poisson(\xi)$
  - Choose a topic mixture distribution $\theta \sim Dir(\alpha)$
  - For each word token in d:
    - Choose a topic $z_n \sim multinomial(\theta)$
    - Choose a word w_n from a multinomial distribution conditioned on the topic z_n.

β is a k by V matrix parameterized with the word probabilities.

$$[\beta]_{k \times V} \quad \beta_{ij} = p(w_j | z = i)$$

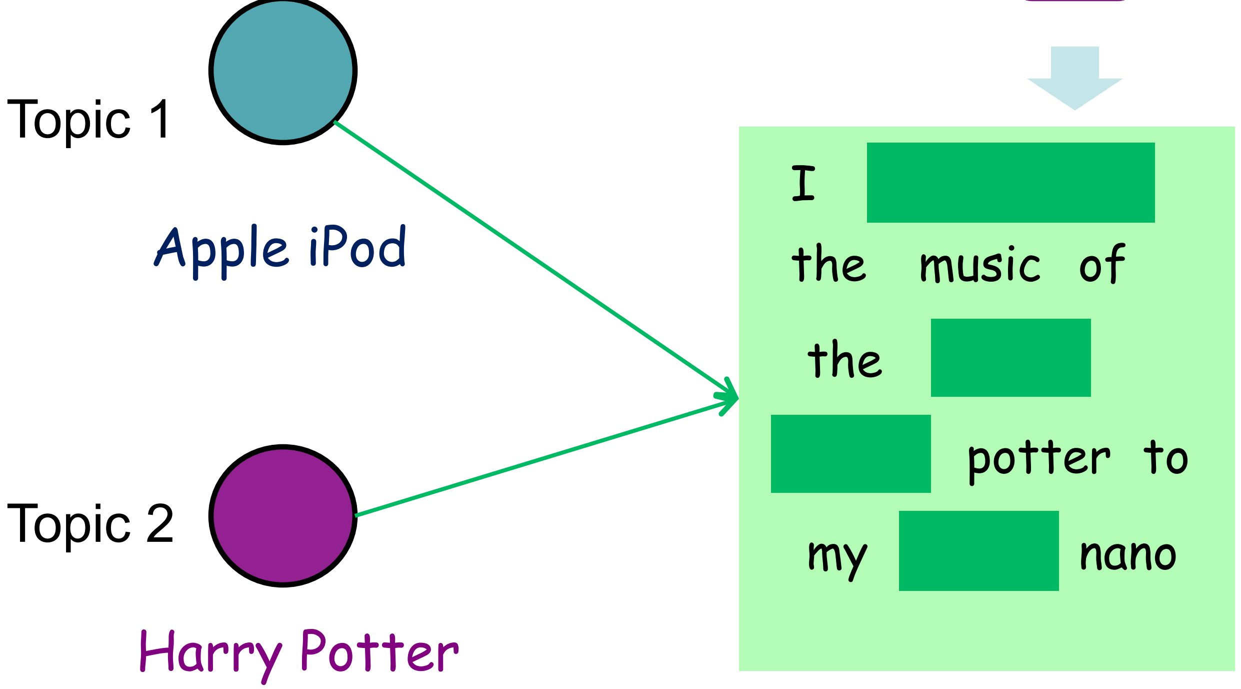
LDA – Generative Process



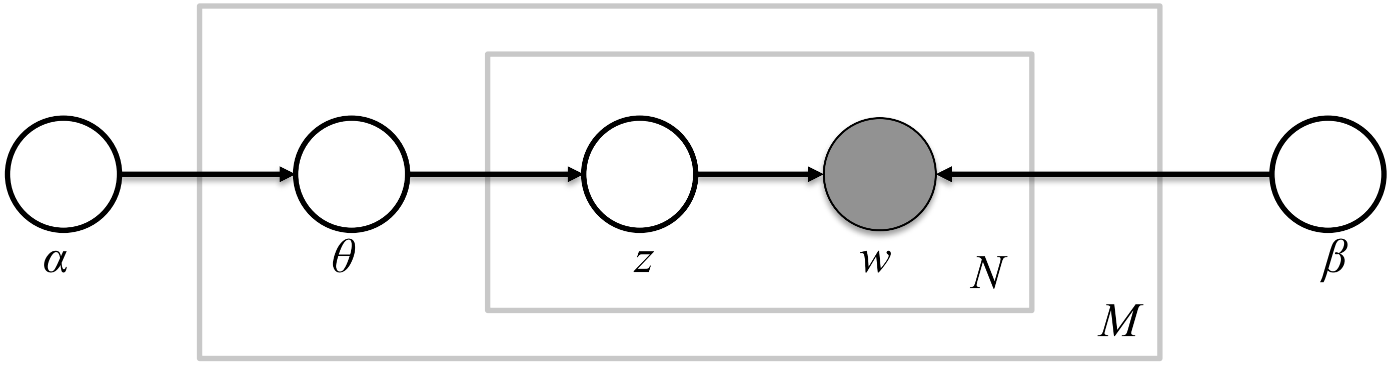
$$P(d)=\prod_{w\in d}\sum_{i=1..K} P(z=i\,|\,d)\,P(w\,|\,Topic_i)$$

ipod	**0.15**
nano	0.08
music	0.05
download	0.02
apple	0.01

movie	*0.10*
harry	***0.09***
potter	*0.05*
ipod	*0.01*
music	*0.02*



Graphical Model of LDA



Complete likelihood: $p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) =$

(if we observe all latent variables)

Data likelihood:

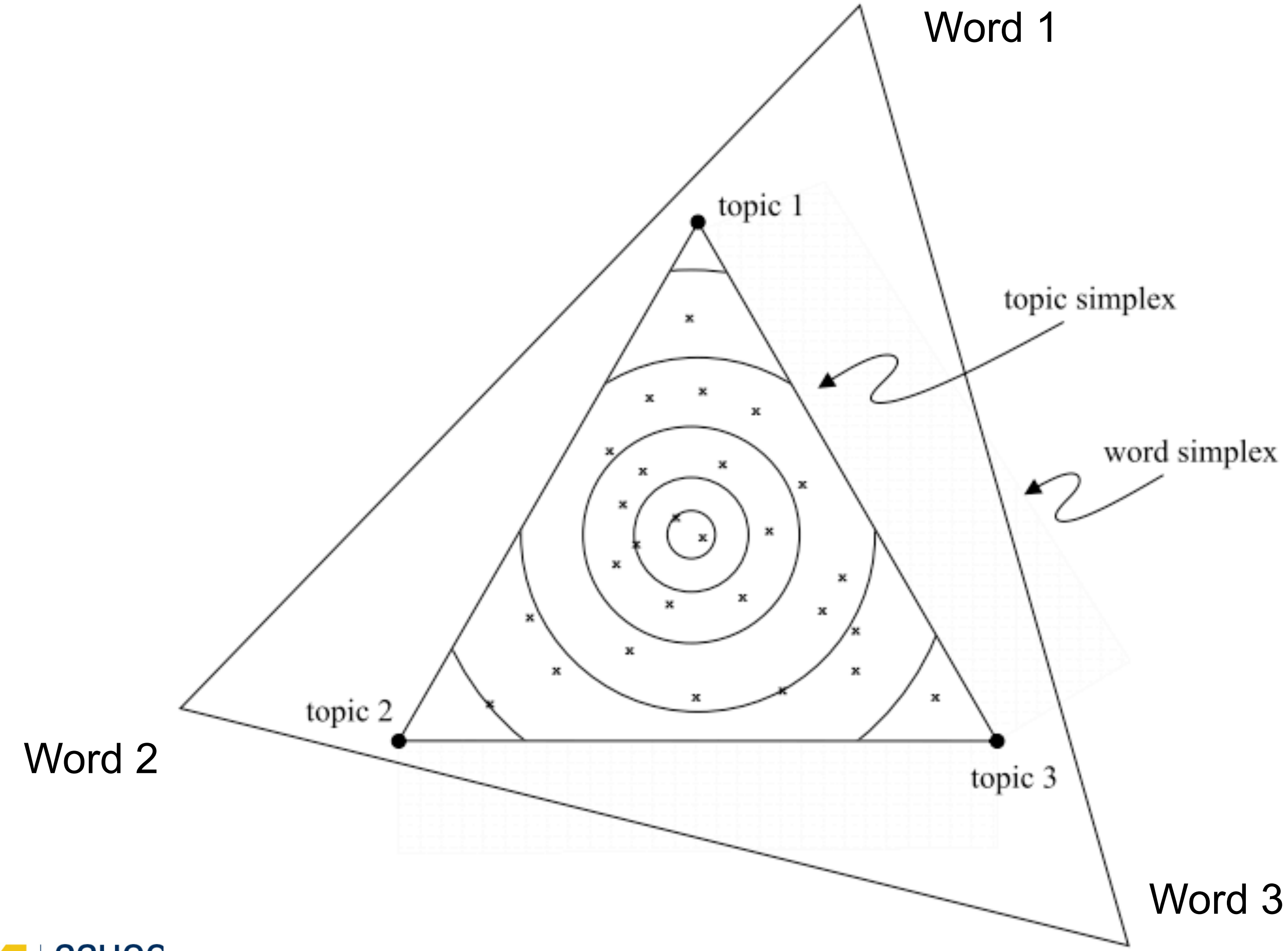
- P(z|d) are no longer fixed parameters to be estimated...
- But inference becomes much more complicated

Dirichlet Prior of Multinomial

- Bayesian Theory: Nothing is fixed, everything is random.
- Choose prior for the multinomial distribution of the topic mixture weights: Dirichlet distribution is conjugate to the multinomial distribution, which is natural to choose.
- A *k*-dimensional Dirichlet random variable θ can take values in the (k-1)-simplex, and has the following probability density on this simplex:

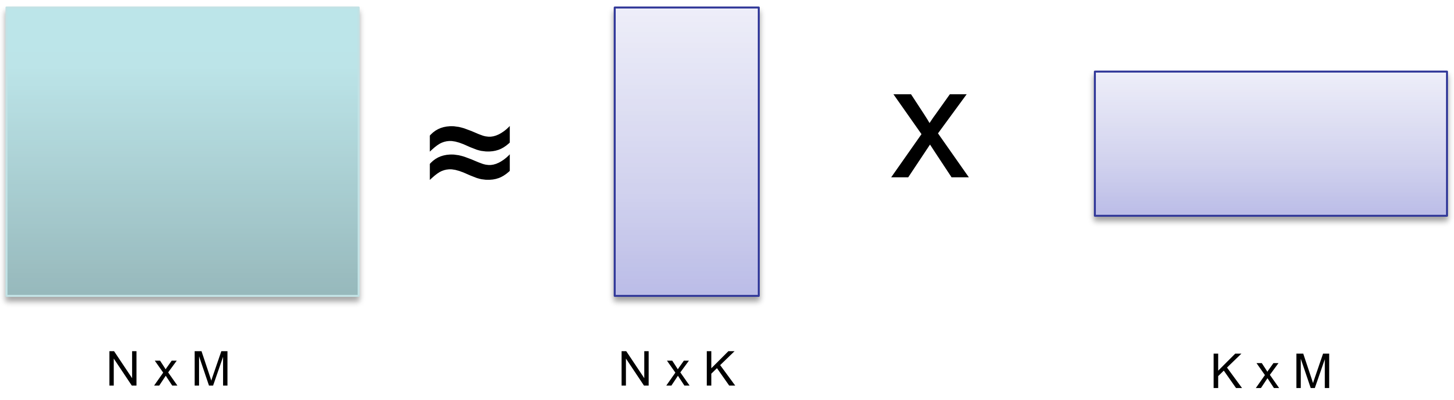
$$(1)\ p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}$$

The Simplex Interpretation



Wait…

- Wasn't this similar to matrix decomposition?
- Indeed… PLSI is the probabilistic version of LSI (i.e., dimension reduction using SVD)
- PLSA is also proved to be equivalent to non-negative matrix factorization (with a specific loss function)



N words, M documents, K topics

Inference & Parameter Estimation

- PLSA:
  - Parameters: $\Lambda = \left\{ P(z \mid d), P(w \mid z) \right\}_{\forall z,d,w}$
  - Maximum Likelihood Estimation (MLE):

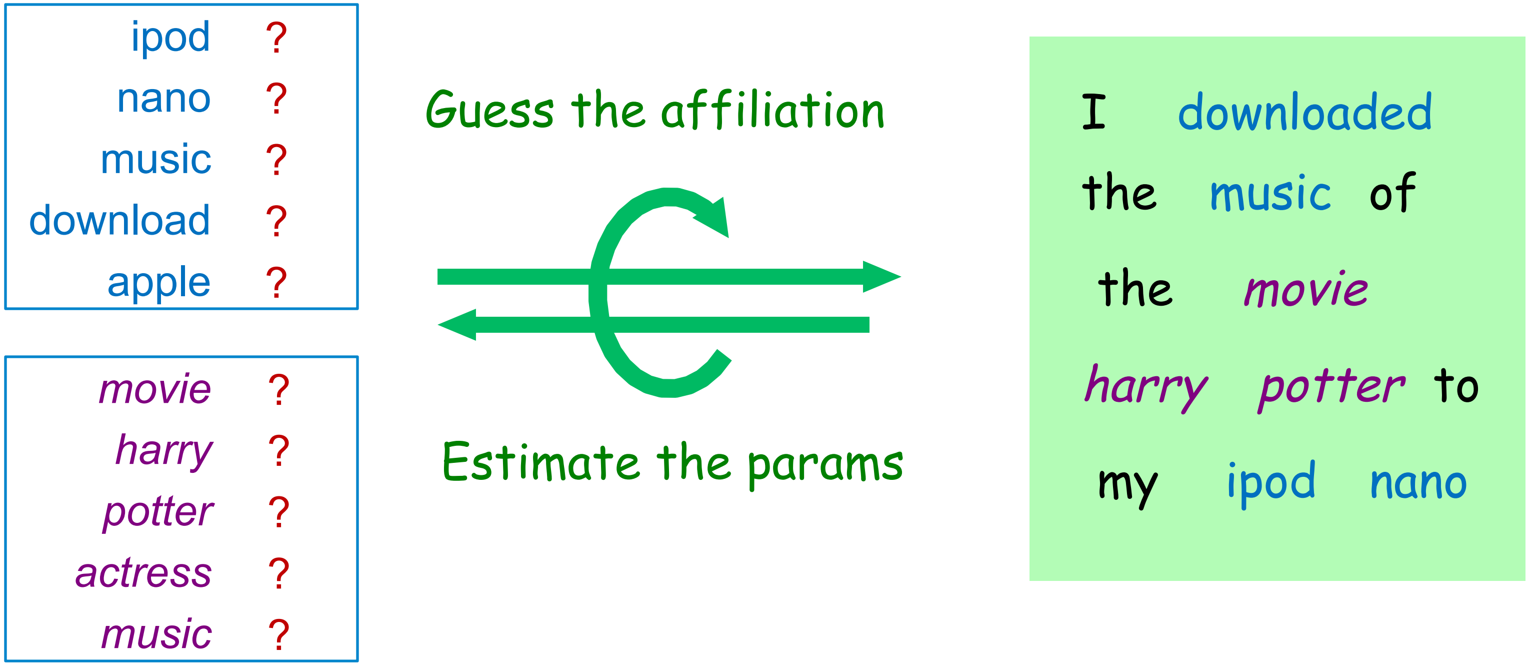
$$\Lambda^* = \arg\max_\Lambda P(D \mid \Lambda)$$

- LDA:
  - Corpus level parameters: *α*, *β*: sampled once in the corpus creating generative model.
  - Document level parameters: θ, z: sampled once to generate a document

$$(\theta, z)^* = \operatorname{argmax}_{(\theta,z)} P(\theta, z \mid D, \alpha, \beta) = \operatorname{argmax}_{(\theta,z)} \frac{P(D \mid \theta, z, \alpha, \beta) P(\theta, z \mid \alpha, \beta)}{P(D \mid \alpha, \beta)}$$

Inference Methods

- Parameter estimation/inference methods:
  - Expectation Maximization, Variational inference, Gibbs sampling, …



ipod	?
nano	?
music	?
download	?
apple	?

movie	?
harry	?
potter	?
actress	?
music	?

Guess the affiliation

Estimate the params

I downloaded
the music of
the *movie*
harry potter to
my ipod nano

Topic Modeling in Practice

- Number of topics have to be predetermined
  - Finding the right number of topics can be hard
  - Bayesian model selection – enumerate all K
  - Non-parametric topic models – more complicated
- Topics need to be interpreted
  - Reading the word distribution can be hard
  - Methods to generate labels for topics (Mei et al., 2007)
- Evaluation is subjective
  - Likelihood of held-out data, quality of topics, quality of document clusters, human judgments, …

Topic Modeling in Practice (Cont.)

- In general, it is a good tool for exploratory analysis
  - Understanding qualitatively what's in the corpus.
- Not the first choice if your have a particular task in mind
  - Classification: try SVM first
  - Labeling: try CRF
  - Clustering: try k-means, spectral clustering, etc.
- But can provide "deep" features for those tasks
- Works well when few training examples are available
- Extremely powerful when contextual variables are observed and concerned with
  - Very easy to extend, by adding variables into the Bayesian network (graphical structure)

Things You Don't Know About LDA

- There are latent requirements of your data…
- Do not apply LDA directly if
  - You have too few documents
  - Your documents are too short
  - Your goal is to find smallish, tight clusters
  - The co-occurrence assumption doesn't hold in your data
- Try to leverage meta-data or context information to improve the performance of topic modeling
  - Authorship
  - Time, geographic location
  - Networks of documents and words
  - User-guidance, …

Resources of Topic Modeling

- Topic Modeling Bibliography
  - Collected by David Mimno: http://www.cs.princeton.edu/~mimno/topics.html
- Toolkits:
  - LDA++: Easy-to-use C++ implementation
  - MALLET: Java package, with good support of text mining in general
  - Mahout: Machine learning toolkit adaptable on Hadoop
- Topic modeling in digital humanities
  - http://mith.umd.edu/topic-modeling-in-the-humanities-an-overview/

Sentiment and Opinion Analysis

Sentiment and Opinion Analysis

- Computational study of subjectivity in text
  - opinions, sentiments, subjectivity, evaluations, attitudes, appraisal, affects, views, emotions, etc.
  - Reviews, blogs, tweets, discussions, news, comments, feedback, or any other documents
- Terminology:
  - Sentiment analysis is more widely used in industry.
  - Both are widely used in academia
  - Sometimes used interchangeably.

- Slide partially from Prof. Bing Liu's tutorial

Why Bother?

- "Opinions" are key influencers of our behaviors.
- Our beliefs and perceptions of reality are conditioned on how others see the world.
- Whenever we need to make a decision, we often seek out the opinions of others. In the past,
  - Individuals: seek opinions from friends and family
  - Organizations: use surveys, focus groups, opinion polls, consultants.

- Slide partially from Prof. Bing Liu's tutorial

Applications of Sentiment Analysis

- Businesses and organizations
  - Benchmark products and services; market intelligence.
  - Businesses spend a huge amount of money to find consumer opinions using consultants, surveys and focus groups, etc.
- Individuals
  - Make decisions to buy products or to use services
  - Find public opinions about political candidates and issues
- Search engines
  - Provide opinion polarized summary of products
- Social science, political science, health informatics, …
  - Understanding public opinions, concerns, …
- Finance, …

Different Levels of Sentiment Analysis

- Sentiment classification
  - Whether a document/sentence conveys a sentiment
- Sentiment polarities
  - Positive, negative, neutral sentiments
- Sentiment in real scales
- Mood, emotion
  - Multi-class classification
- Attitude
  - Sentiment towards particular subjects
- Topic-sentiment mixture
  - Latent aspects of sentiments/opinions
- …

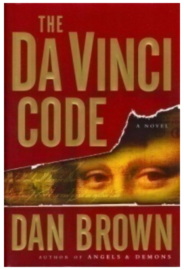
Sentiment Classification

- Yet another text classification problem
- But much more challenging than topic-based classification!
- In general, it touches almost every aspect of NLP
- Human agreement is around 70% or lower
- For human agreed examples, performance is usually between 80% to 90%.
- Evaluation needs to be done on real applications

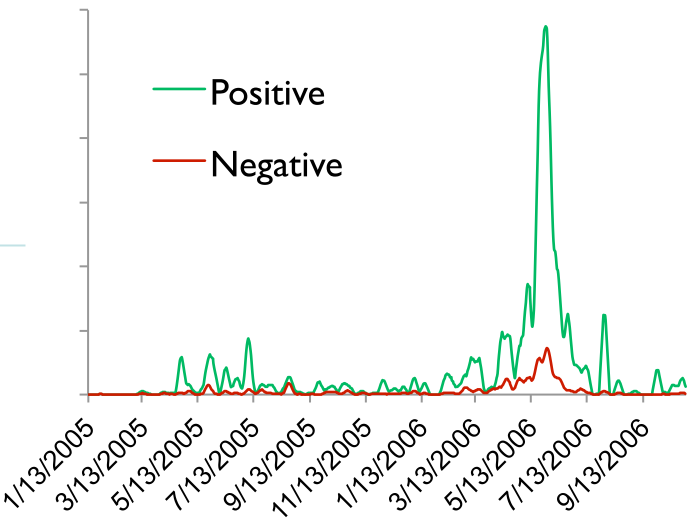
Features Matter

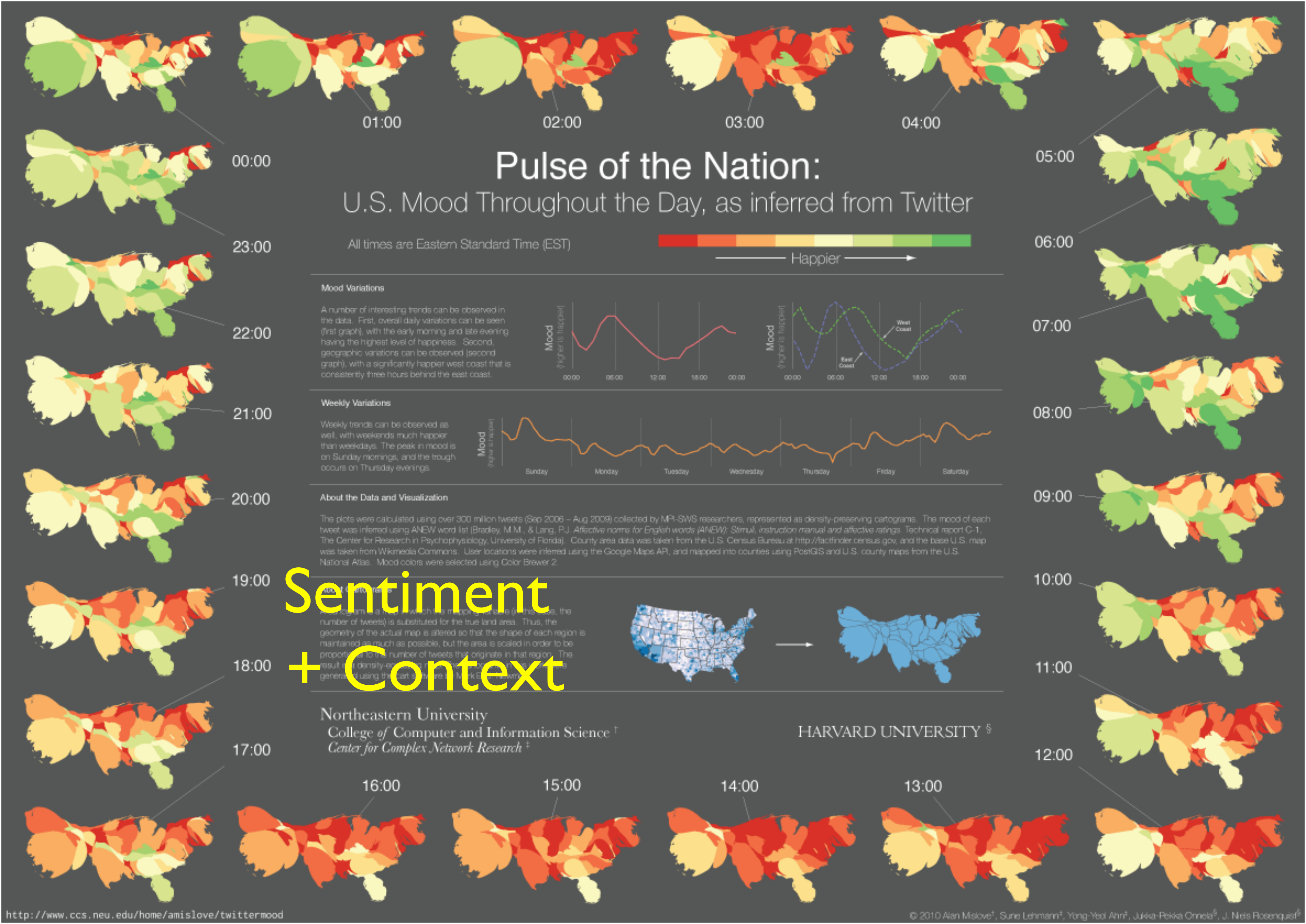
- Keyword features
  - Most useful features in topic classification
  - Presence vs. frequency
  - Dictionary match performs reasonably well.
- Part-of-speech
  - Very effective in sentiment classification
- Syntax vs. word proximity
- Negation tends to be important
- Topic-related features
- Specific features – e.g., emoticons
- Co-reference, sarcasm, metaphor, …

Topics + Sentiment = Faceted Opinion Summary

	👍	👎
	Tom Hanks, who is my favorite movie star act the leading role.	protesting... will lose your faith by watching the movie.
 THE DA VINCI CODE A NOVEL DAN BROWN AUTHOR OF ANGELS & DEMONS	a good book to past time.	... so sick of people making such a big deal about a fiction **book**

The Da Vinci Code





Pulse of the Nation:

U.S. Mood Throughout the Day, as inferred from Twitter

All times are Eastern Standard Time (EST)

Mood Variations

A number of interesting trends can be observed in the data. First, overall daily variations can be seen (first graph), with the early morning and late evening having the highest level of happiness. Second, geographic variations can be observed (second graph), with a significantly happier west coast that is consistently three hours behind the east coast.

Weekly Variations

Weekly trends can be observed as well, with weekends much happier than weekdays. The peak in mood is on Sunday mornings, and the trough occurs on Thursday evenings.

About the Data and Visualization

The plots were calculated using over 300 million tweets (Sep 2006 – Aug 2009) collected by MPI-SWS researchers, represented as density-preserving cartograms. The mood of each tweet was inferred using ANEW word list (Bradley, M.M., & Lang, P.J. *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings.* Technical report C-1, The Center for Research in Psychophysiology, University of Florida). County area data was taken from the U.S. Census Bureau at http://factfinder.census.gov, and the base U.S. map was taken from Wikimedia Commons. User locations were inferred using the Google Maps API, and mapped into counties using PostGIS and U.S. county maps from the U.S. National Atlas. Mood colors were selected using Color Brewer 2.

About Cartograms

A cartogram is a map in which the mapping variable (in this case, the number of tweets) is substituted for the true land area. Thus, the geometry of the actual map is altered so that the shape of each region is maintained as much as possible, but the area is scaled in order to be proportional to the number of tweets that originate in that region. The result is a density-equalizing map. The cartograms in this work were generated using the cart software by Mark E.J. Newman.

Northeastern University
College *of* Computer and Information Science †
Center for Complex Network Research ‡

HARVARD UNIVERSITY §

http://www.ccs.neu.edu/home/amislove/twittermood

© 2010 Alan Mislove†, Sune Lehmann‡, Yong-Yeol Ahn‡, Jukka-Pekka Onnela§, J. Niels Rosenquist§

Sentiment + Context

What You Should Know

- Major research issues of text mining
- Topic modeling and sentiment analysis are two trending topics
- Topic modeling as an exploratory tool
  - LDA and its extensions can be effective when data satisfies particular properties
- Sentiment classification is more challenging than topic classification, and should be treated differently
- Yet many problems to be solved