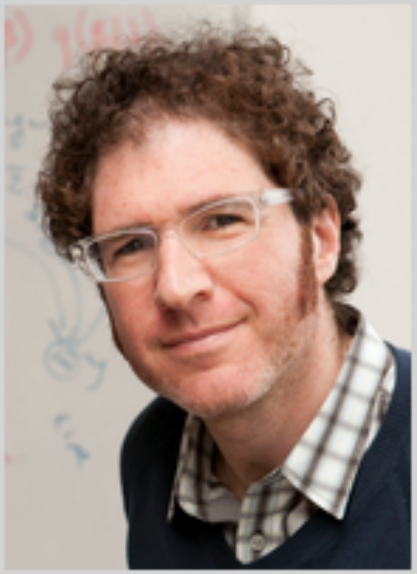
Statistical Topic Models

Chenghua Lin
Computing Science
University of Aberdeen

Outlines

- Graphical model
- Probability distributions
- Latent Dirichlet allocation (LDA)

Probabilistic topic models



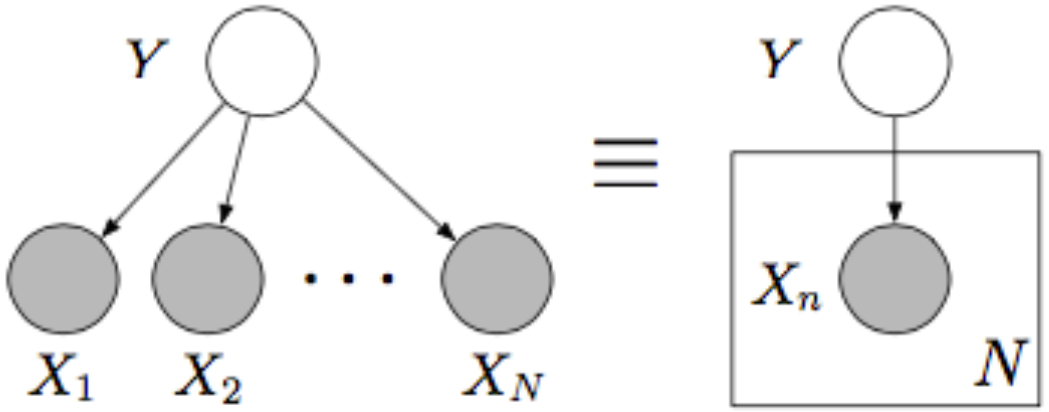
David M. Blei

Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

Graphical Model

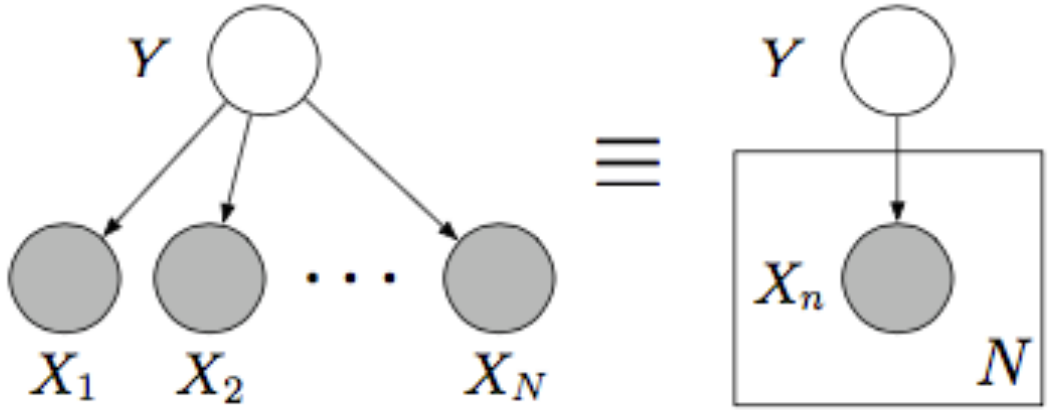
- What is graphical models
  - a probabilistic model for which a graph expresses the conditional dependence structure between random variables.
  - commonly used in probability theory, statistics, particularly Bayesian statistics, and machine learning.

Graphical Model



- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plates denote replicated structure

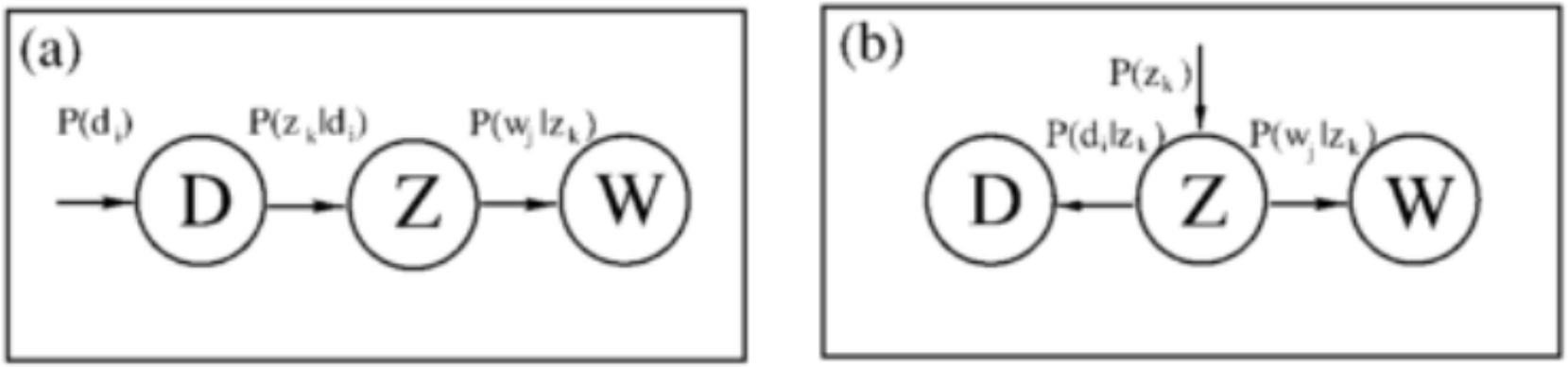
Graphical Model



- Structure of the graph denotes the pattern of conditional dependence between the resemble of random variables.
- E.g., this graph corresponds to

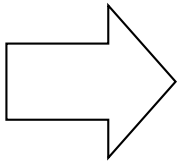
$$P(y, x_1, x_2 \ldots, x_N) = P(y) \prod_{i=1}^{N} P(x_n|y)$$

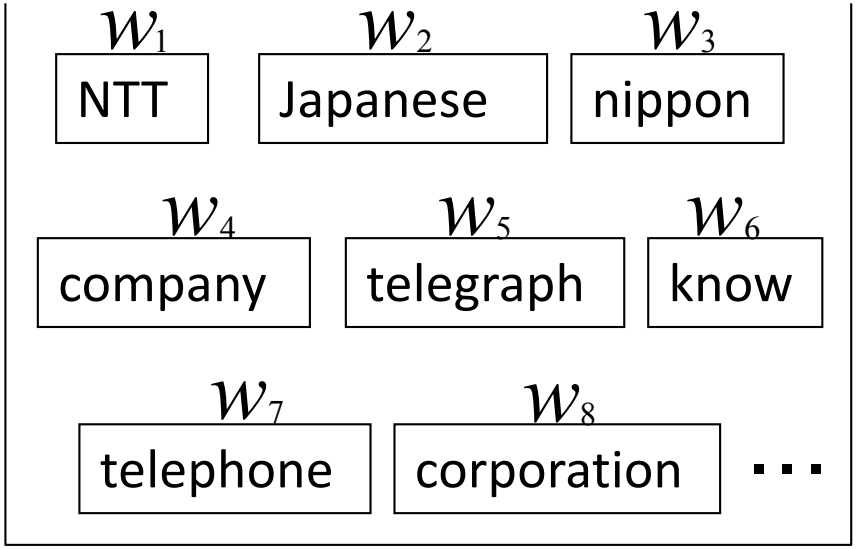
Graphical Model



Data: bag-of-words

Nippon Telegraph and Telephone Corporation, commonly known as NTT, is a Japanese telecommunications company headquartered …..





w

- a document is represented by a set of words (disregarding word order)

Bernoulli Distribution

- Consider a *single* binary random variable x:{0, 1}
- p(x=1|μ) = μ, where 0 ≤ μ ≤ 1.

$$\mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$$

or

Binomial Distribution

- Suppose we have N observations of the outcomes of a binary random variable x:{0, 1}, denoted as $D = \{x_1, x_2, \ldots, x_N\}$
- One wants to work out the distribution for the event that *m* out of the N observations have the outcome x = 1.

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

where $$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!}$$

- A Binomial distribution is a generalization of the Bernoulli Distribution when N = 1.

Multinomial Distribution

- A Multinomial variable can take one of the ***K*** possible mutual exclusive states, e.g. a dice where *k:* {1, 2, 3, 4, 5, 6}.
- Suppose you throw a dice N times, where the number of times you have the observation {*k = 1*} is denoted as m_1

$$\text{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

where $\binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1! m_2! \ldots m_K!}$

- A multinomial distribution is a generalization of the binomial distribution when ***K* = 2**

Lejeune Dirichlet
1805–1859

Johann Peter Gustav Lejeune Dirichlet was a modest and reserved mathematician who made contributions in number theory, mechanics, and astronomy, and who gave the first rigorous analysis of Fourier series. His family originated from Richelet in Belgium, and the name Lejeune Dirichlet comes from 'le jeune de Richelet' (the young person from Richelet). Dirichlet's first paper, which was published in 1825, brought him instant fame. It concerned Fermat's last theorem, which claims that there are no positive integer solutions to $x^n + y^n = z^n$ for $n > 2$. Dirichlet gave a partial proof for the case $n = 5$, which was sent to Legendre for review and who in turn completed the proof. Later, Dirichlet gave a complete proof for $n = 14$, although a full proof of Fermat's last theorem for arbitrary n had to wait until the work of Andrew Wiles in the closing years of the 20^{th} century.

Dirichlet Distribution

- A fair dice *k:* {1, 2, 3, 4, 5, 6}
  - $\boldsymbol{\mu} = \{\mu_1 = 1/6, \mu_2 = 1/6, \mu_3 = 1/6, \mu_4 = 1/6, \mu_5 = 1/6, \mu_6 = 1/6\}$
- An unfair dice *K*: {1, 2, 3, 4, 5, 6}
  - $\boldsymbol{\mu} = \{\mu_1 = 1/3, \mu_2 = 1/12, \mu_3 = 1/12, \mu_4 = 1/12, \mu_5 = 1/12, \mu_6 = 1/3\}$
- How do we model the probability of $\boldsymbol{\mu}$ in terms of some distributions?
  - That is, the distribution of a distribution.

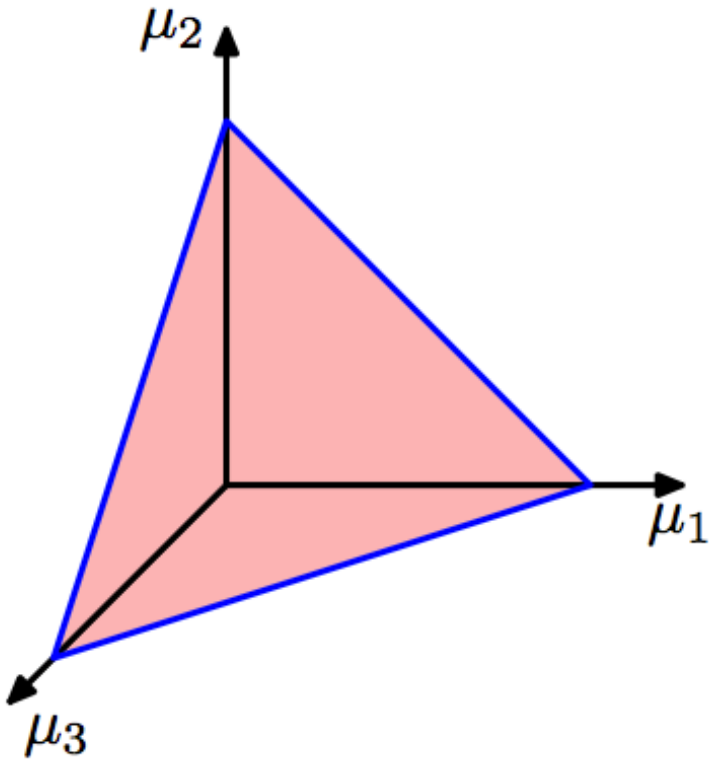
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

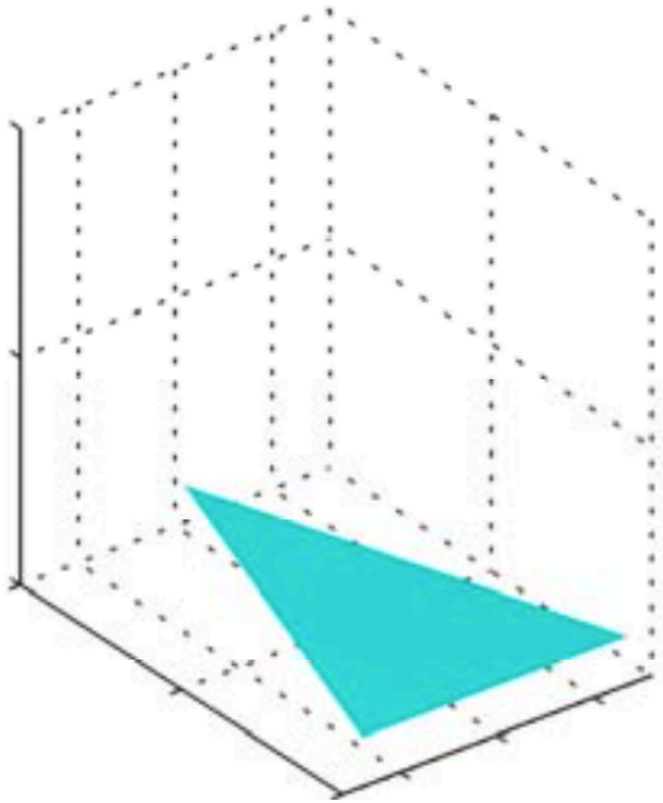
Dirichlet Distribution

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

- The Dirichlet distribution is an exponential family distribution over the simplex, i.e. positive vectors that sum to one.
  - $0 \leqslant \mu_k \leqslant 1$ and $\sum_k \mu_k = 1$
- The Dirichlet parameter α controls the mean shape and sparsity of **μ**.
- It is **conjugate** to the multinomial. Given a multinomial observation, the posterior distribution of **μ** is a Dirichlet.

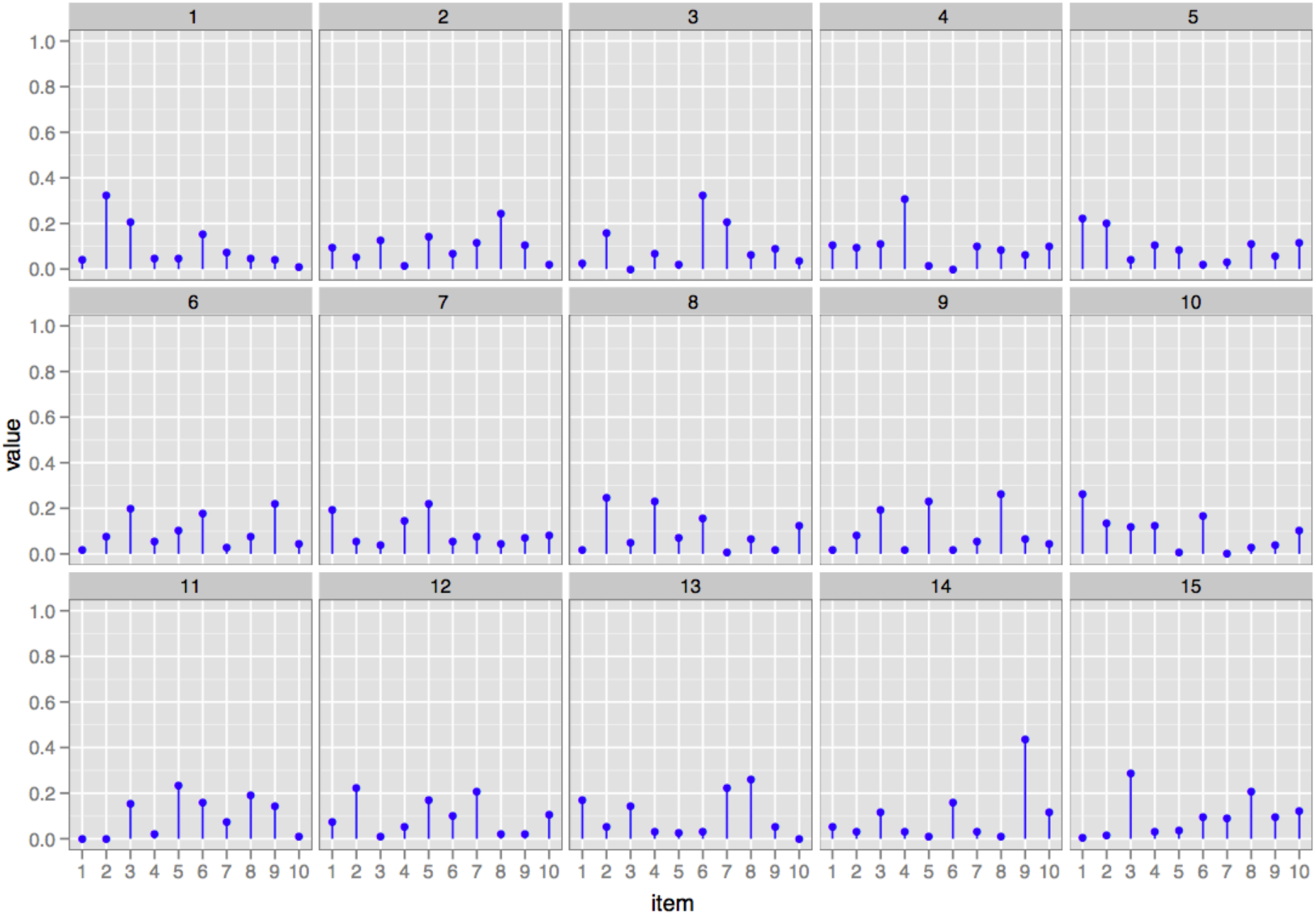
Dirichelet Distribution: α = 1



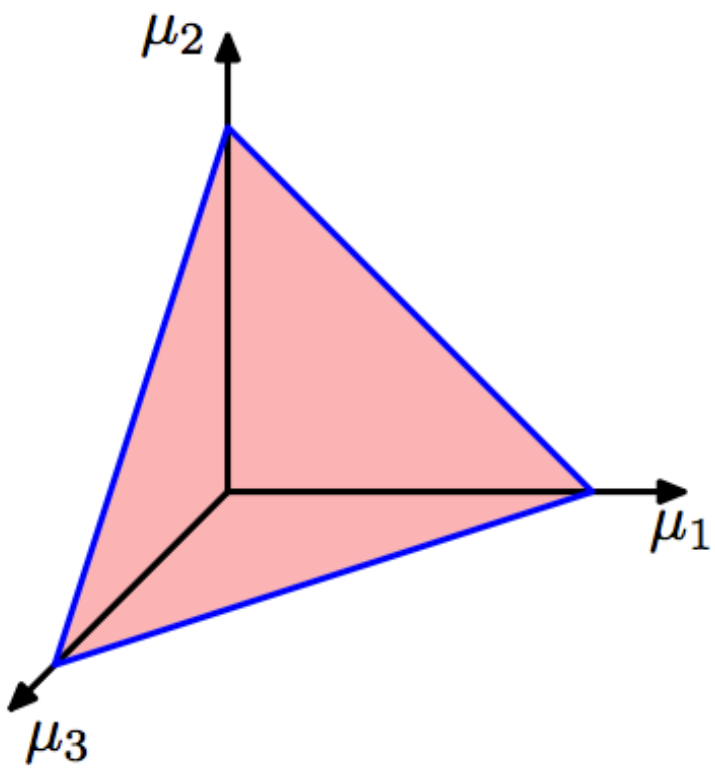


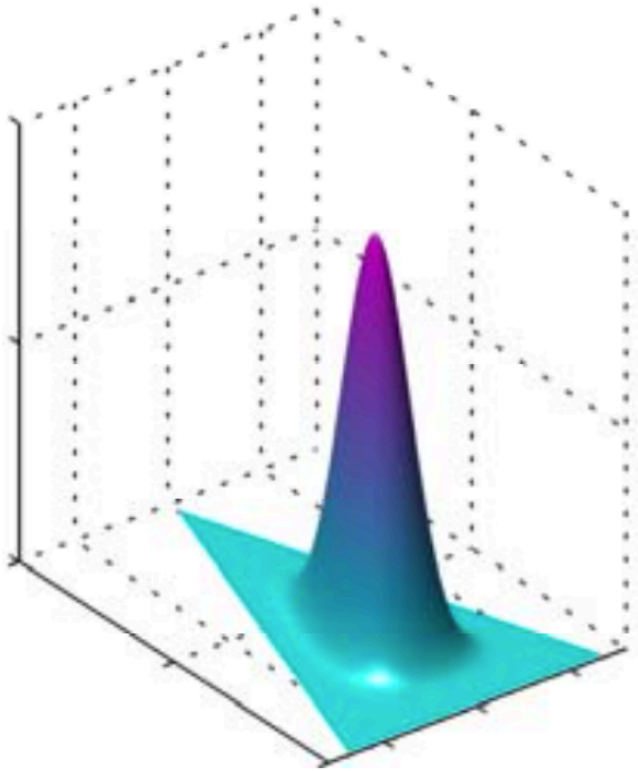
- $\boldsymbol{\mu} = \{\mu_1, \mu_2, \mu_3\}$
- Simplex dimensionality: K-1 = 3-1 = 2

α=1

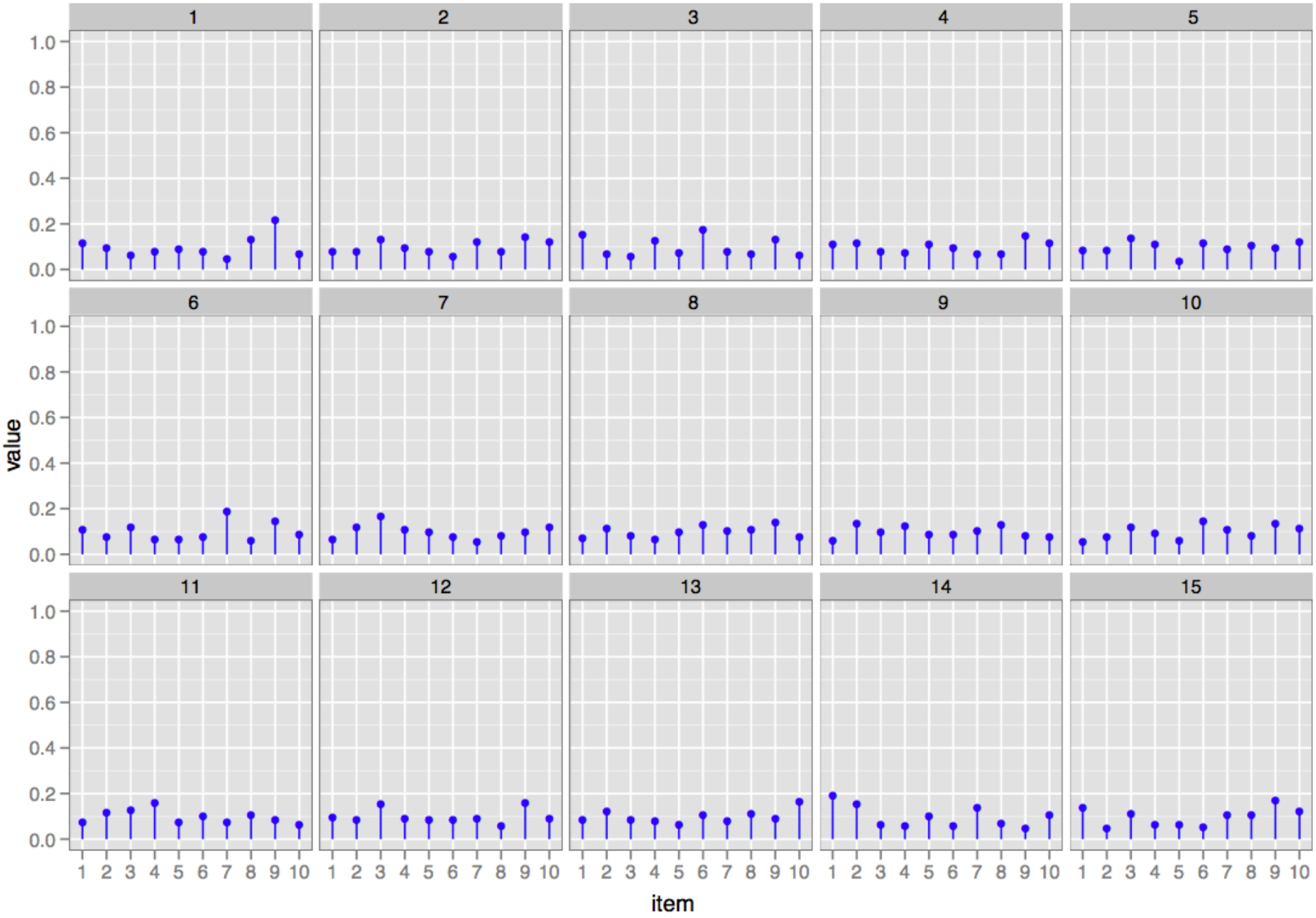

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
item
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5 6 7 8 9 10

Dirichelet Distribution: α > 1

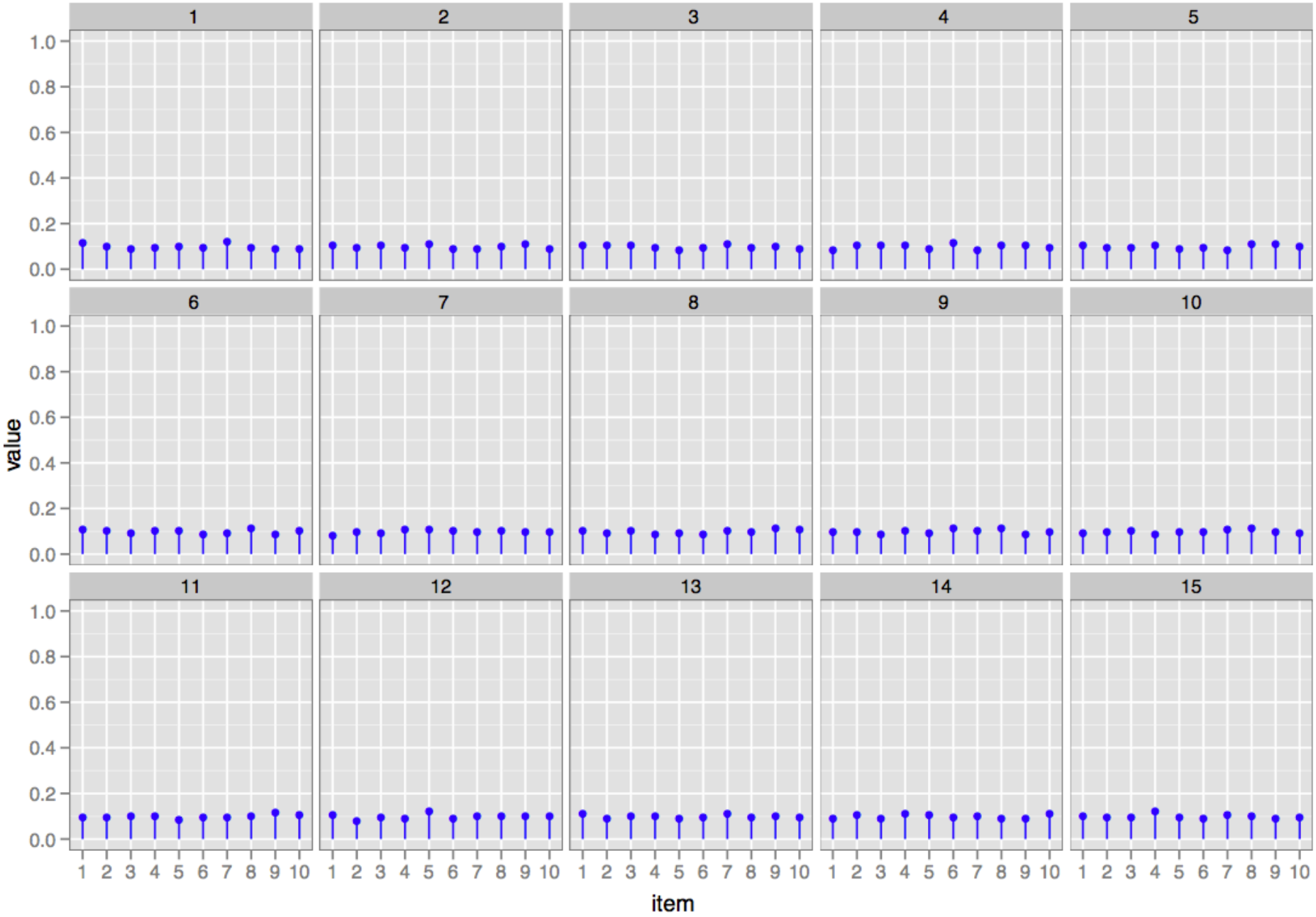




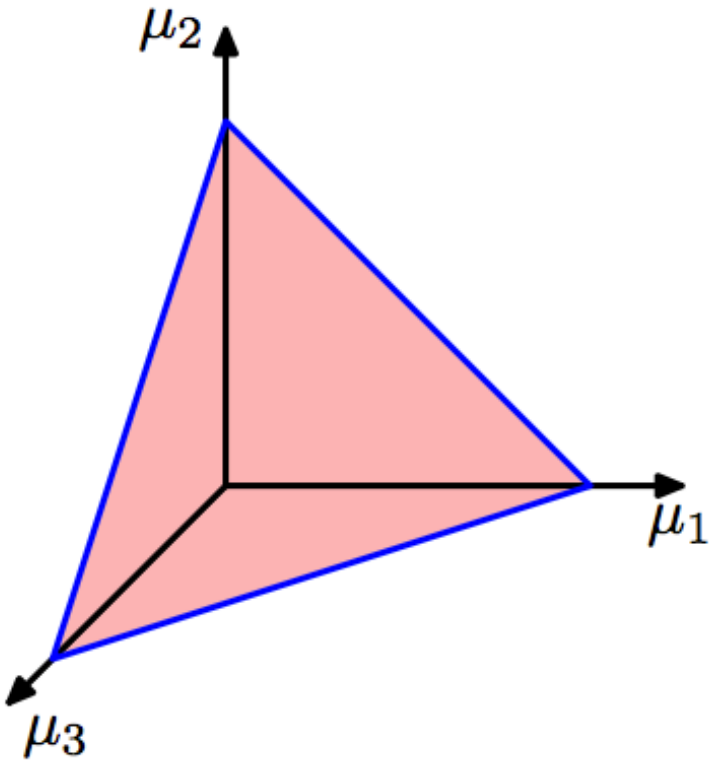
α=10

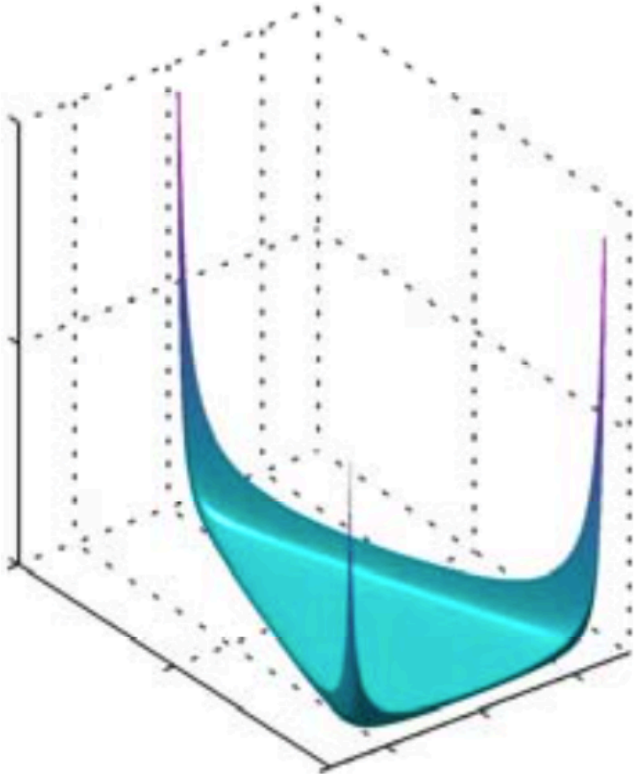

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
item
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5 6 7 8 9 10

α=100

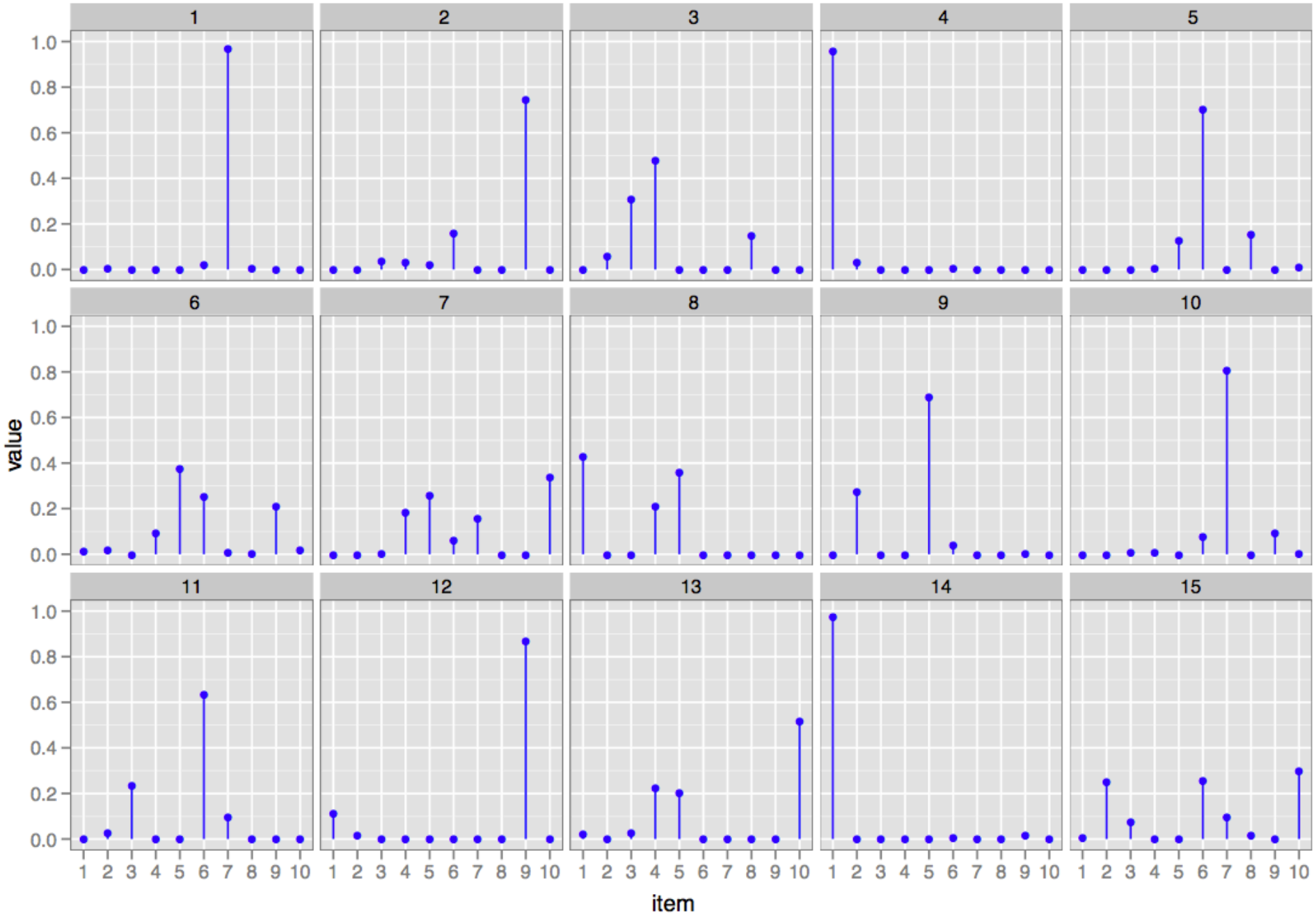

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
item
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5 6 7 8 9 10

Dirichelet Distribution: α < 1

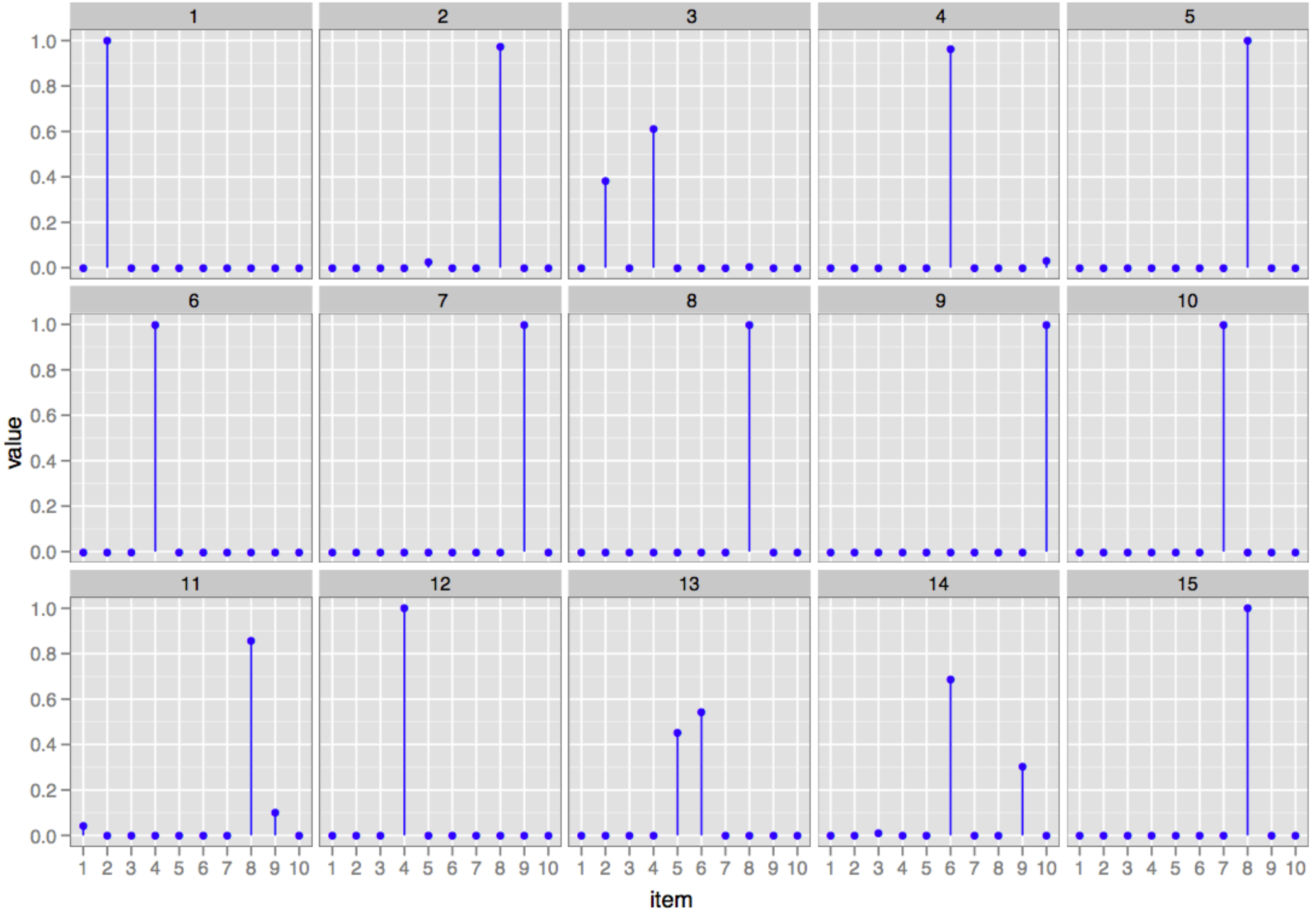




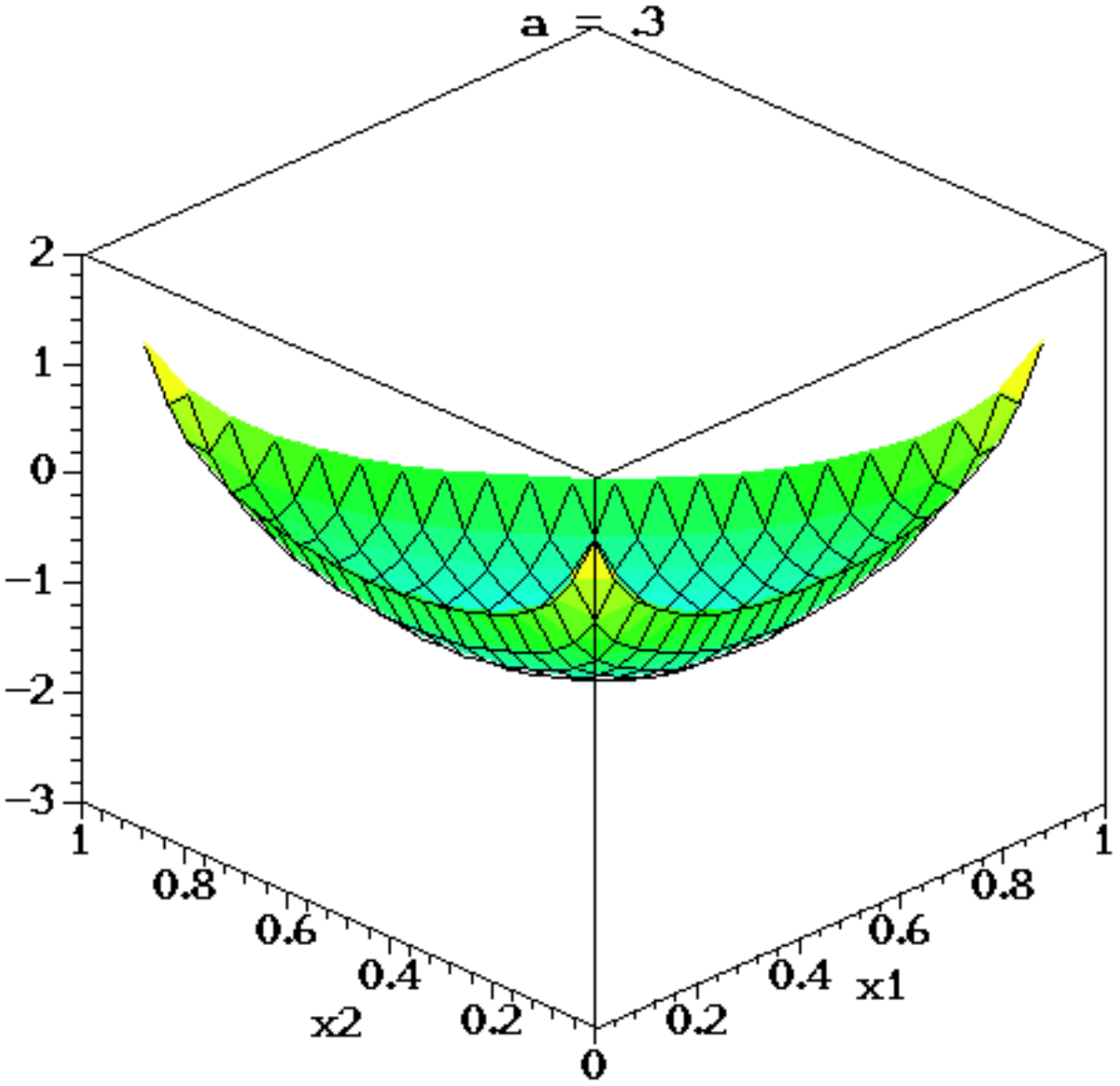
α=0.1


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
item
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5 6 7 8 9 10

α=0.01


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
value
item
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5 6 7 8 9 10

- Let’ e
  den
  cha o
  (2.0
  equ



Latent Dirichlet Allocation

Topic Models

A topic model is a generative probabilistic model for discrete data with latent structure

- Generative model for capturing semantic properties of text documents
- Can be applied to a wide variety of data
  - images, purchase logs, social network, music
- Easy to extend and implement

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

Topic Extraction

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Input: Documents

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

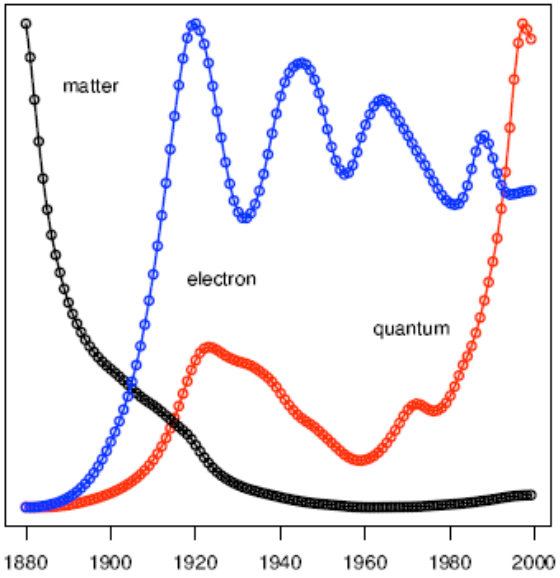
Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

D. Blei, A. Ng, M. Jordan, Latent Dirichlet Allocation, JMLR2003

Dynamic Topic Analysis

1881	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
force	motion	magnet	force	atom	ray	energy	energy	radiat	electron	electron	electron	state
energy	force	electric	magnet	theory	measure	measure	radiat	energy	energy	energy	atom	energy
motion	magnet	measure	theory	electron	energy	electron	ray	electron	atom	particle	energy	electron
differ	energy	force	electric	energy	theory	light	electron	measure	measure	field	structur	magnet
light	measure	theory	atom	measure	light	atom	measure	ray	radiat	radiat	field	field
measure	differ	system	system	ray	wave	particle	atom	atom	field	model	model	atom
magnet	direct	motion	measure	electr	radiat	ray	particle	field	ray	atom	state	system
direct	line	line	line	line	atom	radiat	two	two	model	two	two	two
matter	result	point	energy	force	electric	point	light	particle	particle	ray	magnet	quantum
result	light	differ	body	value	value	theory	absorpt	observe	magnet	measure	ray	physic

"Atomic Physics"

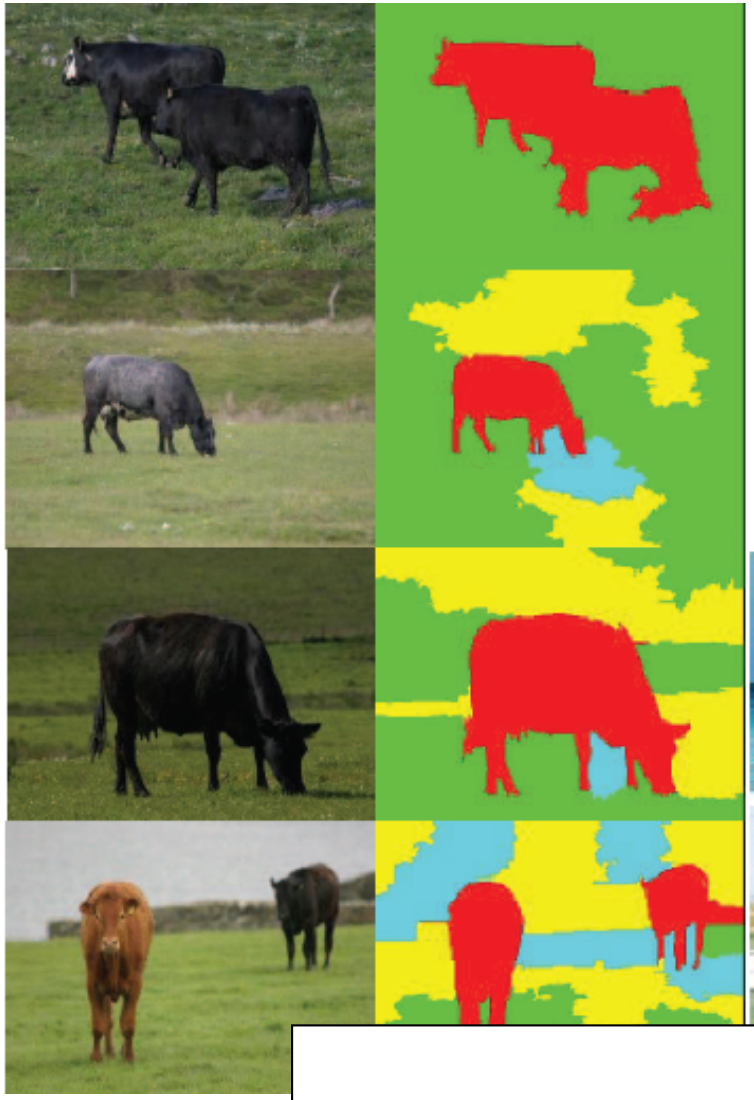


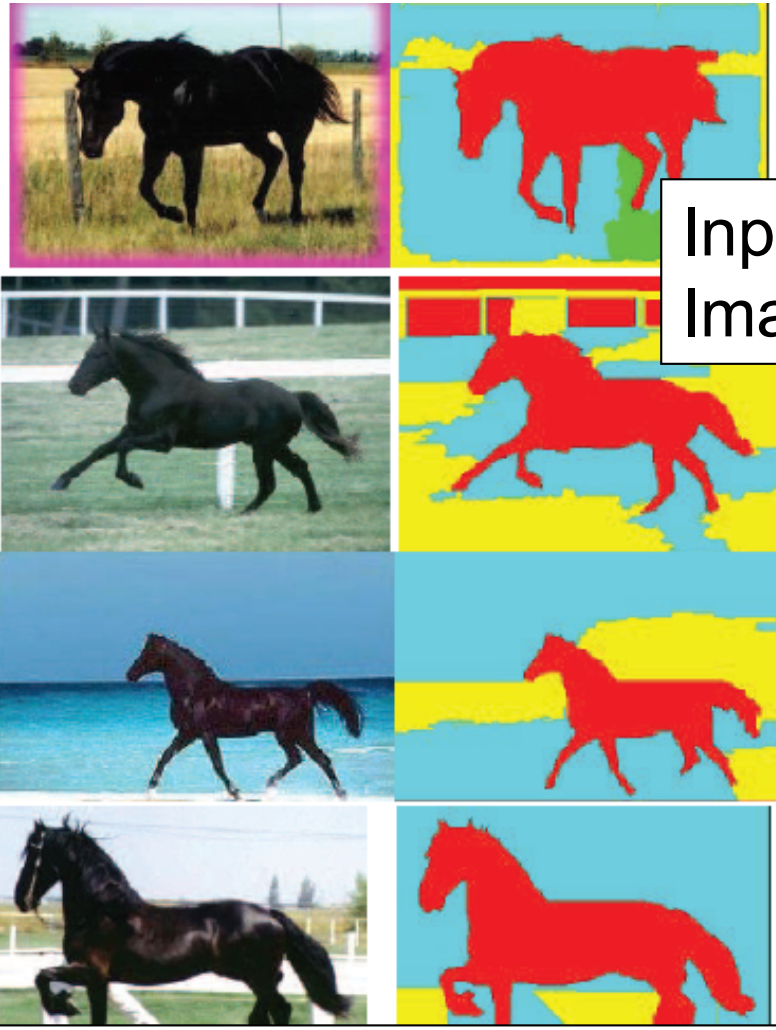
1881 On Matter as a form of Energy
1892 Non-Euclidean Geometry
1900 On Kathode Rays and Some Related Phenomena
1917 ``Keep Your Eye on the Ball"
1920 The Arrangement of Atoms in Some Common Metals
1933 Studies in Nuclear Physics
1943 Aristotle, Newton, Einstein. II
1950 Instrumentation for Radioactivity
1965 Lasers
1975 Particle Physics: Evidence for Magnetic Monopole Obtained
1985 Fermilab Tests its Antiproton Factory
1999 Quantum Computing with Electrons Floating on Liquid Helium

Input:
Documents
with time info

D. Blei, J. Lafferty, Dynamic Topic Models, ICML2006

Image Recognition



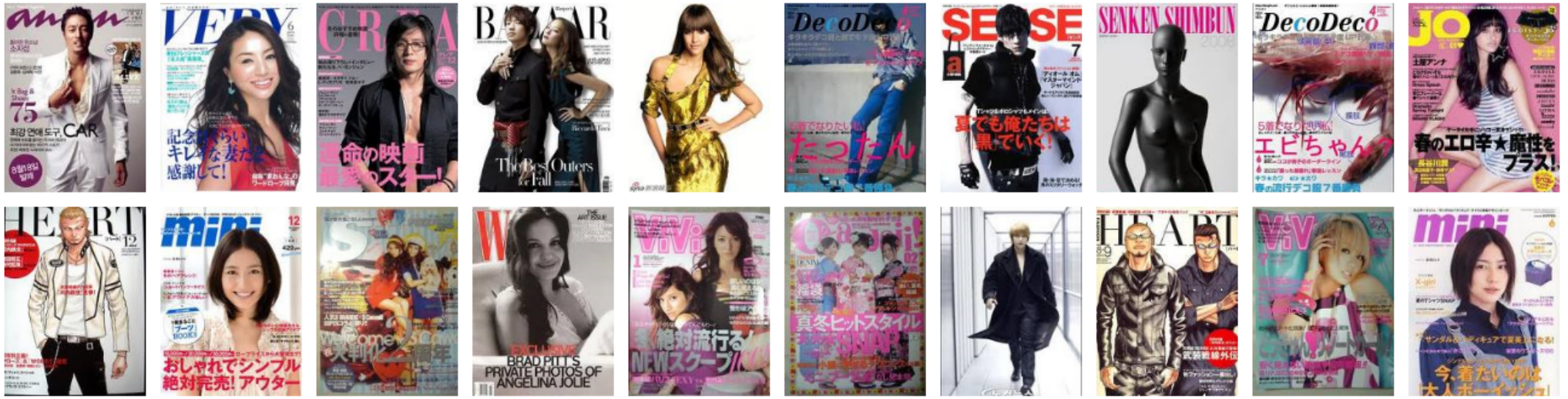


Input: Images

[L. Cao and L. Fei-Fei. **Spatially coherent latent topic model for concurrent object segmentation and classification . ICCV2007]**

Fashion Recommendation

- Fashion magazines contain photos of fashion models
  – their clothing coordinates serve as useful references
- We propose a recommender system for coordinates by using the photos



Task

Input

Photo of
tops (bottoms)

Output

Photo of
bottoms (tops)

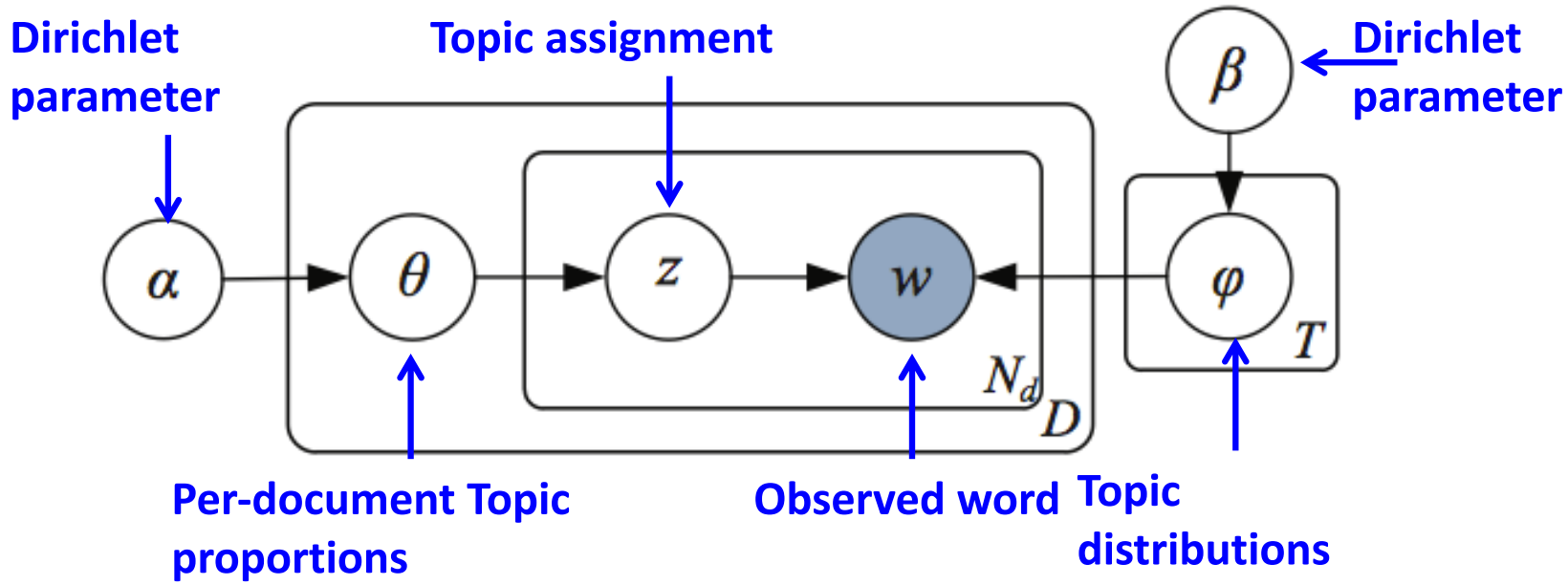
- Learn coordinates from photos
- Helpful when
  - selecting clothes from her wardrobe
  - buying clothes that match with her own clothes

Latent Dirichlet Allocation (LDA)

- LDA (Blei, Ng & Jordan, 2003)
  - A fully unsupervised Bayesian model
  - Assumes that documents exhibit multiple topics as "**theme**" or "**gist**")
  - Each topic is a distribution over words which ha semantic relation with one another

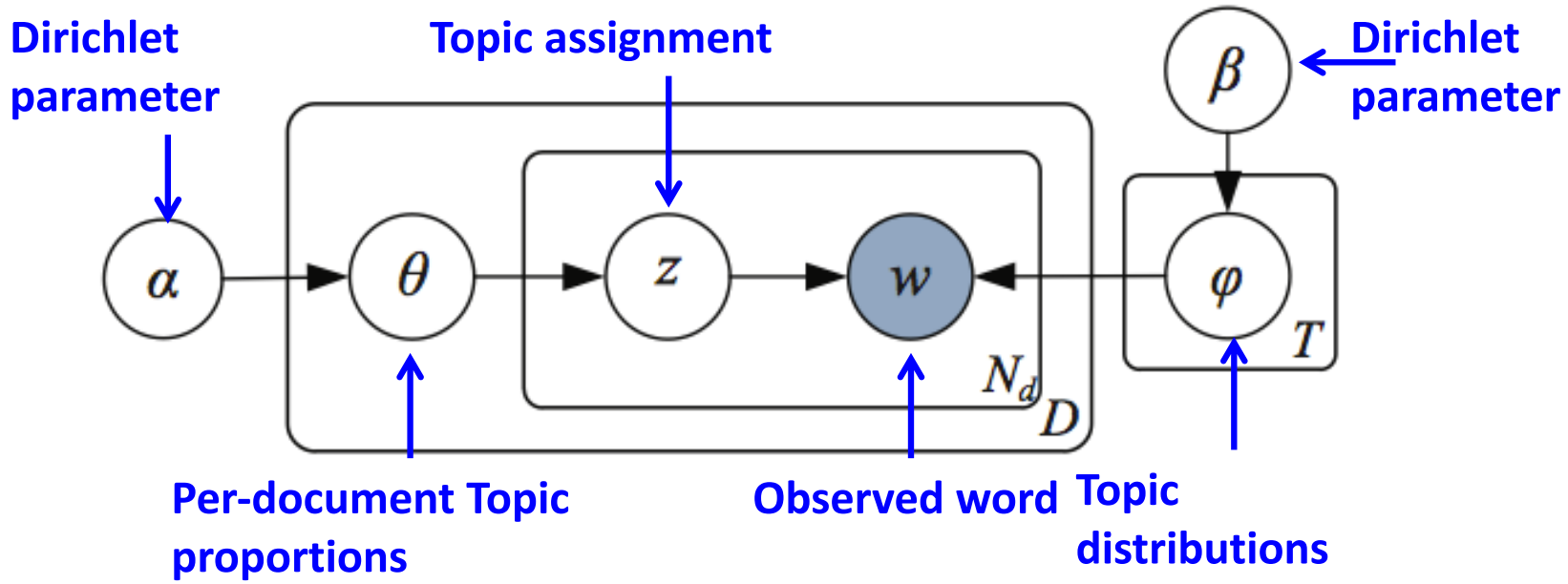
gene
dna
cell
sequence
genetics
mapping
human
molecular
…

LDA Model



- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; unshaded nodes are hidden.
- Plates indicate replicated variables.

Model Parameters



- θ: per-document topic proportion
- φ: per-corpus topic word distribution
- Z: per-word topic assignment

LDA Model

Dirichl
param

hlet
meter

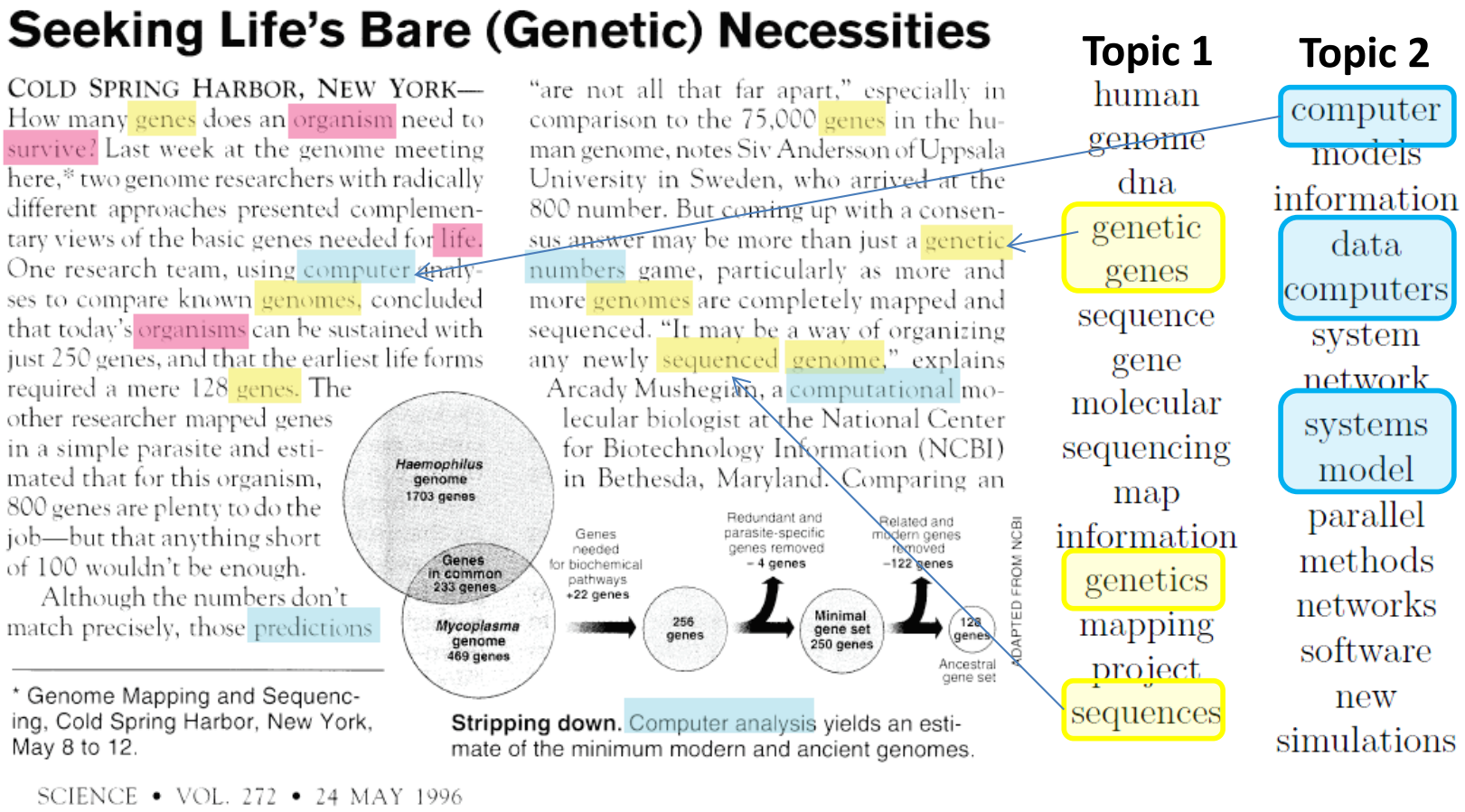
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

- Int
  - –
  - –
  - –

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Discover topics from documents



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

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

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

SCIENCE • VOL. 272 • 24 MAY 1996

Topic 1
human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
information
genetics
mapping
project
sequences

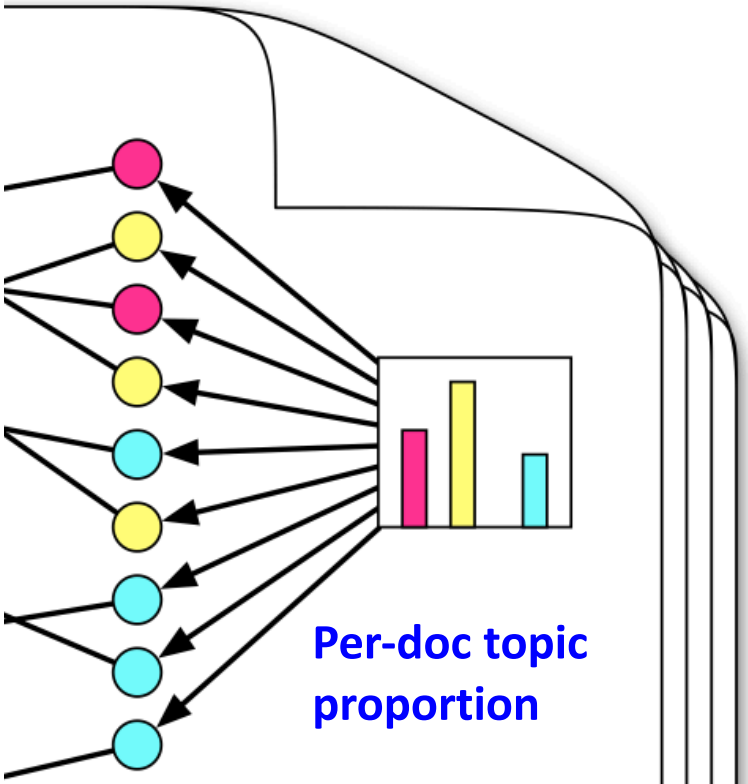
Topic 2
computer
models
information
data
computers
system
network
systems
model
parallel
methods
networks
software
new
simulations

From David Blei

- The co-occuring words in each topic tend to have strong semantic association with each other.

w1 w2 w3 w4 ?? …

Generate a document with a bulk of words …



Topics:

gene	0.04
dna	0.02
cell	0.01
…	

life	0.02
evolve	0.01
organism	0.01
…	

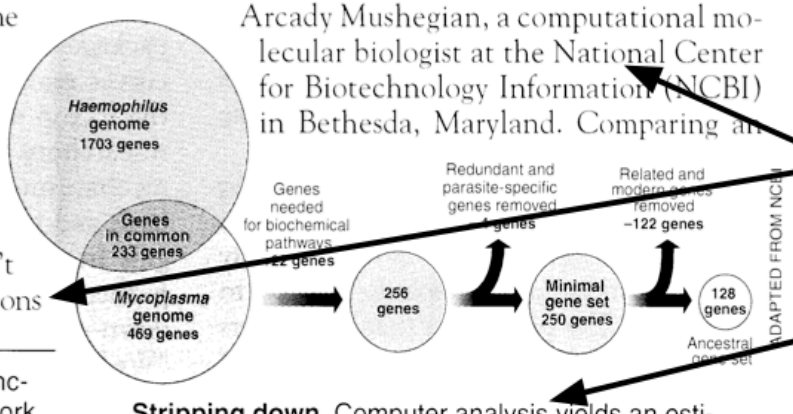
brain	0.04
neuron	0.02
nerve	0.01
…	

data	0.04
number	0.04
computer	0.04
…	

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



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.

Topics:

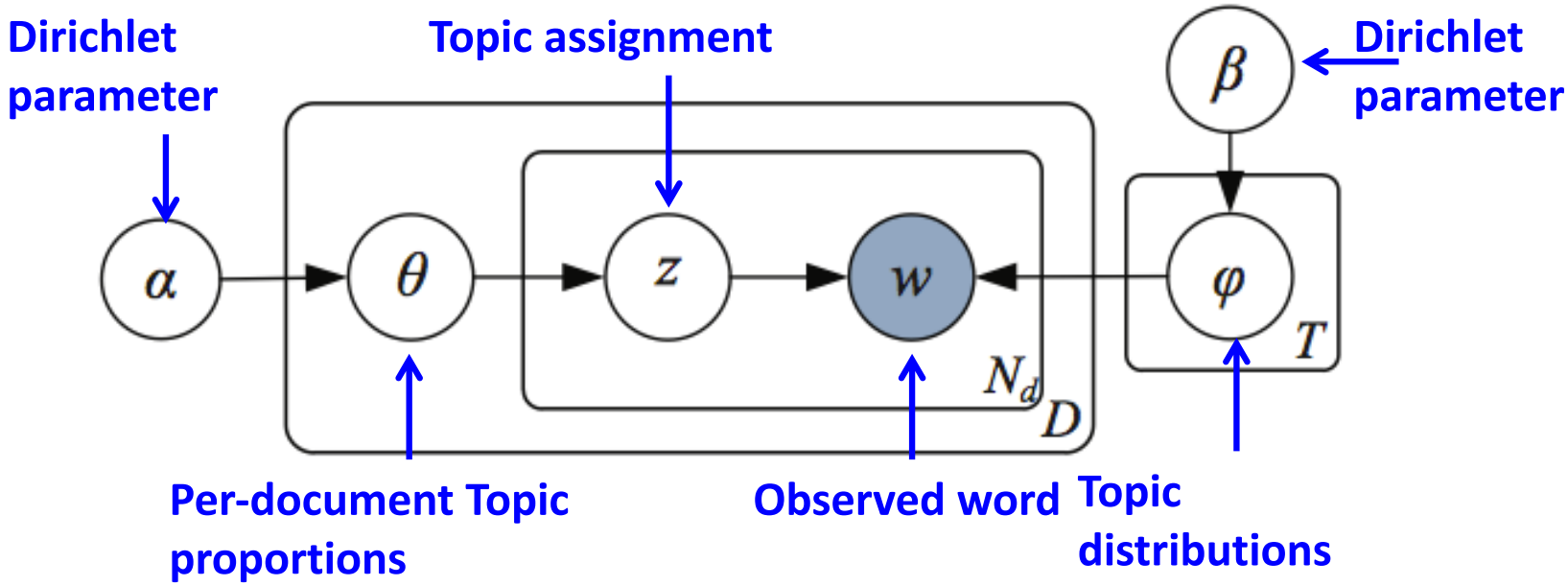
?
...

?

?

?

LDA Model



From a collection of documents, we need to infer

- Per-word topic assignment $z_{d,n}$
- Per-document topic proportions θ_d
- Per-corpus topic distributions φ_k

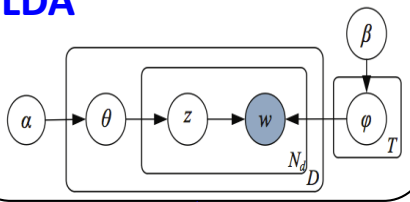
Estimate a posterior, $p(\theta, z, \varphi | w)$.

- Gibbs sampling
- Variational Bayes inference

Applications

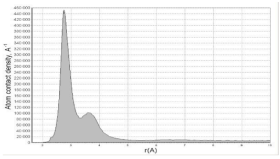
$$p(\beta,\theta,\mathbf{z},\mathbf{w}) = \left(\prod_{i=1}^{K} p(\beta_i|\eta)\right)\left(\prod_{d=1}^{D} p(\theta_d|\alpha)\prod_{n=1}^{N} p(z_{d,n}|\theta_d)p(w_{d,n}|\beta_{1:K},z_{d,n})\right)$$

LDA

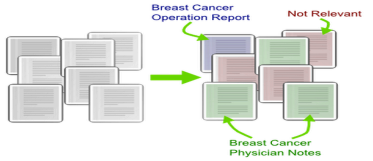


Data

Posterior inference



Classification

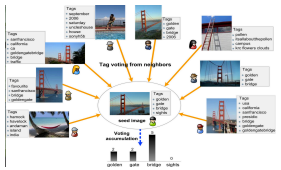


Information retrieval

Google

Recommendation

Data exploration



Posterior

- Computing the posterior distribution of the hidden variables given a document

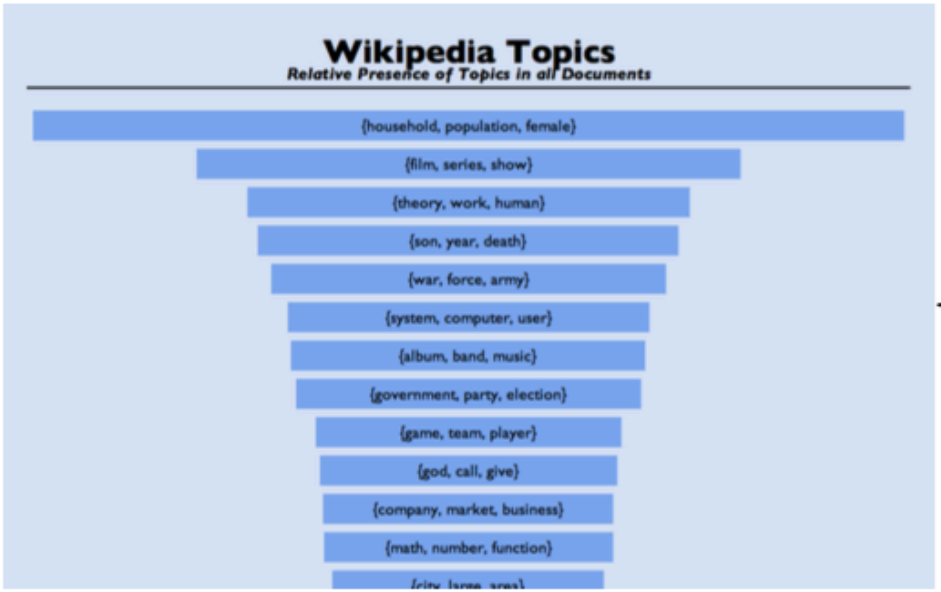
$$P(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{P(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \boldsymbol{\alpha}, \boldsymbol{\beta})}{P(\mathbf{w} | \boldsymbol{\alpha}, \boldsymbol{\beta})}$$

$$p(\boldsymbol{w}|\alpha, \beta) = \iint p(\theta_d|\alpha) \cdot p(\varphi|\beta) \cdot \prod_{n=1}^{N_d} p\left(w_{d,n}|\theta_d, \varphi\right) d\varphi d\theta_d$$

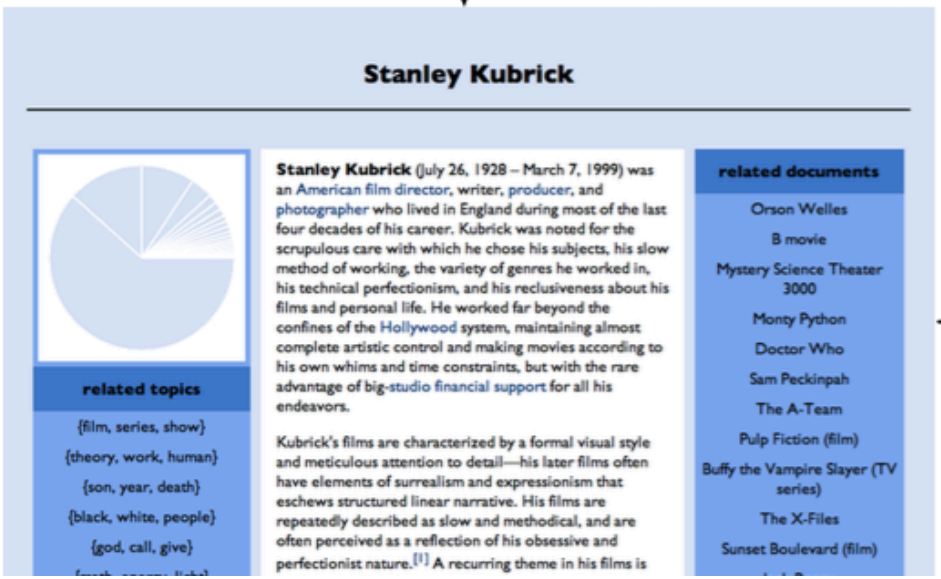
Intractable Posterior

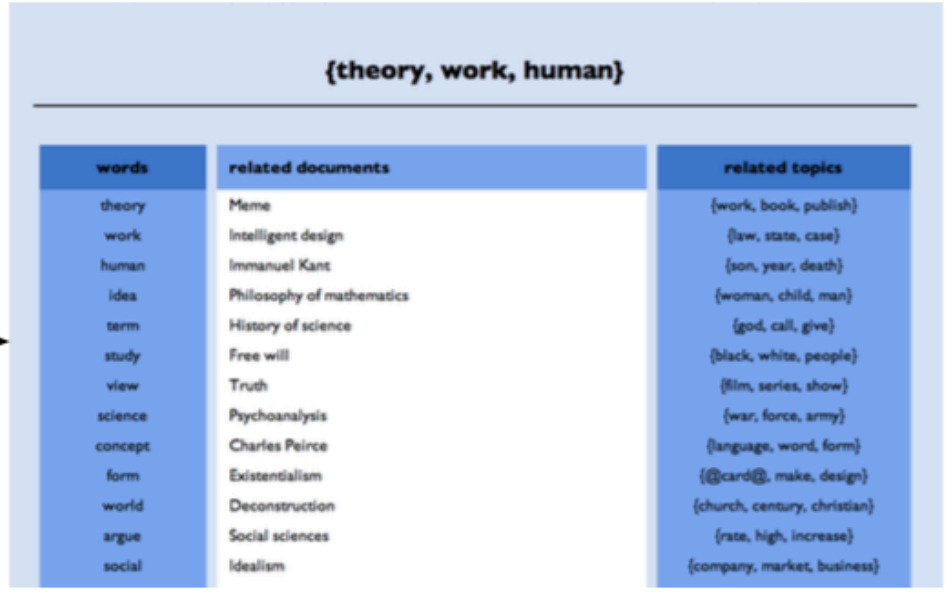
$$p(\boldsymbol{w}|\alpha,\beta) = \iint p(\theta_d|\alpha)\cdot p(\varphi|\beta)\cdot \prod_{n=1}^{N_d} p\left(w_{d,n}\middle|\theta_d,\varphi\right) d\varphi d\theta_d$$

- The integral in this expression is intractable due to the coupled parameters θ and φ, and is thus usually estimated by using
  - MCMC approaches, e.g. Gibbs Sampling
  - Variational Bayes
  - Expectation propagation

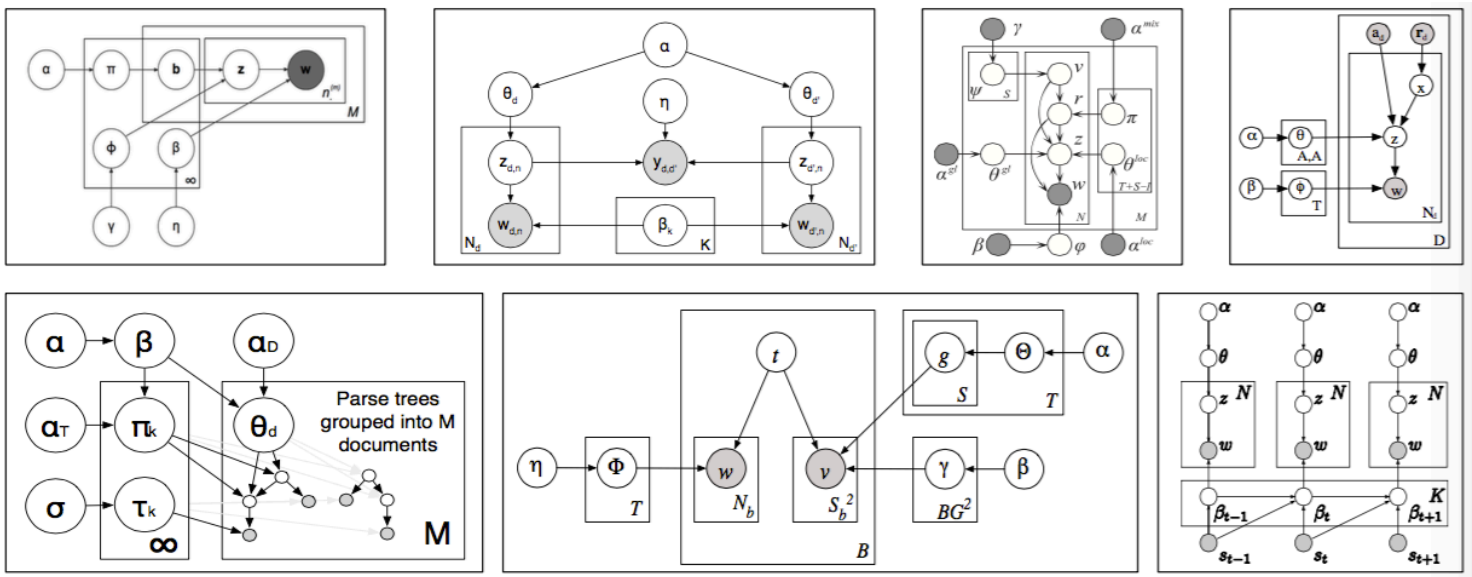

Wikipedia Topics
Relative Presence of Topics in all Documents
{household, population, female}
{film, series, show}
{theory, work, human}
{son, year, death}
{war, force, army}
{system, computer, user}
{album, band, music}
{government, party, election}
{game, team, player}
{god, call, give}
{company, market, business}
{math, number, function}


{film, series, show}
words
related documents
related topics
film
series
show
character
play
make
episode
movie
good
release
feature
television
star
The X-Files
Orson Welles
Stanley Kubrick
B movie
Mystery Science Theater 3000
Monty Python
Doctor Who
Sam Peckinpah
Married... with Children
History of film
The A-Team
Pulp Fiction (film)
Mad (magazine)
{son, year, death}
{work, book, publish}
{album, band, music}
{woman, child, man}
{law, state, case}
{black, white, people}
{theory, work, human}
{@card@, make, design}
{war, force, army}
{god, call, give}
{game, team, player}
{day, year, event}
{company, market, business}


Stanley Kubrick
related topics
{film, series, show}
{theory, work, human}
{son, year, death}
{black, white, people}
{god, call, give}
Stanley Kubrick (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow method of working, the variety of genres he worked in, his technical perfectionism, and his reclusiveness about his films and personal life. He worked far beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his own whims and time constraints, but with the rare advantage of big-studio financial support for all his endeavors.
Kubrick's films are characterized by a formal visual style and meticulous attention to detail—his later films often have elements of surrealism and expressionism that eschews structured linear narrative. His films are repeatedly described as slow and methodical, and are often perceived as a reflection of his obsessive and perfectionist nature.[1] A recurring theme in his films is
related documents
Orson Welles
B movie
Mystery Science Theater 3000
Monty Python
Doctor Who
Sam Peckinpah
The A-Team
Pulp Fiction (film)
Buffy the Vampire Slayer (TV series)
The X-Files
Sunset Boulevard (film)


{theory, work, human}
words
related documents
related topics
theory
work
human
idea
term
study
view
science
concept
form
world
argue
social
Meme
Intelligent design
Immanuel Kant
Philosophy of mathematics
History of science
Free will
Truth
Psychoanalysis
Charles Peirce
Existentialism
Deconstruction
Social sciences
Idealism
{work, book, publish}
{law, state, case}
{son, year, death}
{woman, child, man}
{god, call, give}
{black, white, people}
{film, series, show}
{war, force, army}
{language, word, form}
{@card@, make, design}
{church, century, christian}
{rate, high, increase}
{company, market, business}

Other LDA Model Variants



- In order to design a new model, you need to know how to derive the model posterior, where the mainstream approaches are Gibbs sampling and Variational Bayes.

Summary

- LDA is a probabilistic model of text. It casts the problem of discovering themes in large document collections as a clustering problem.
- It lets us visualize the hidden thematic structure in large collections, and generalize new data to fit into that structure.
- LDA is easy to extend for other applications, e.g. sentiment analysis, etc.
- What you should know
  - Graphical model
  - What are topic models?
  - What kinds of things topic models can do?
  - The meaning of LDA model parameters