Social Media & Network Analytics

Topic Modelling

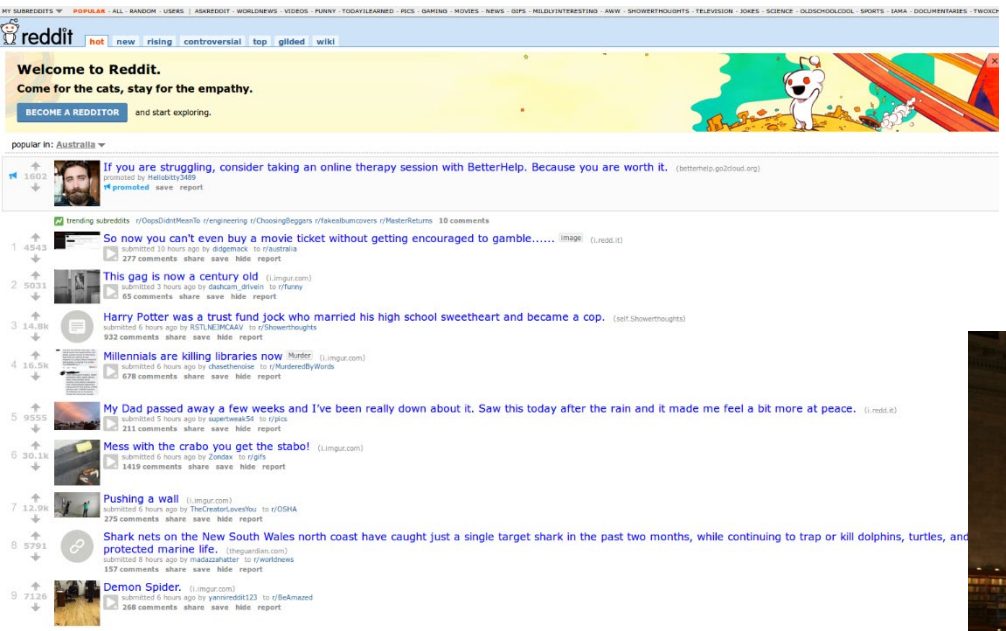
Acknowledgements

- David Blei's tutorial on Topic Modelling

Overview

- Introduction and motivation for topic modelling
- Probability and statistics
- High level explanation of topic modelling
- More detailed explanation of topic modelling
  - Model
  - How to fit models (briefly)

Document Corpus



Topic Models

- Topic models
  - Discover hidden themes/topics that are prevalent within the corpus
  - Annotate the documents according to these topics

- Topic models can help you automatically organise, summarise and understand large electronic corpus

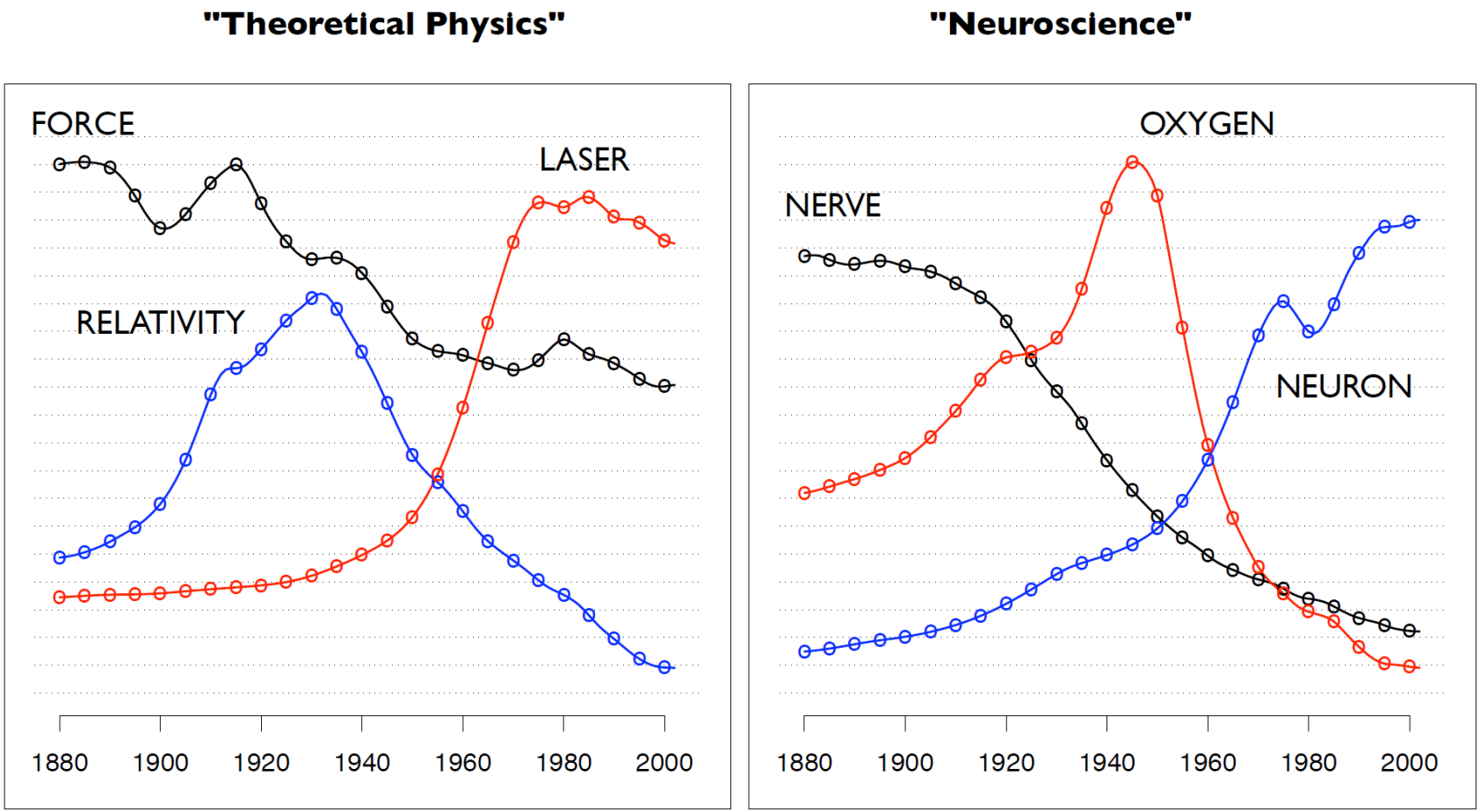
- Topic models is a form of **exploratory analysis**, unsupervised analysis

Discussion point: What is a topic?

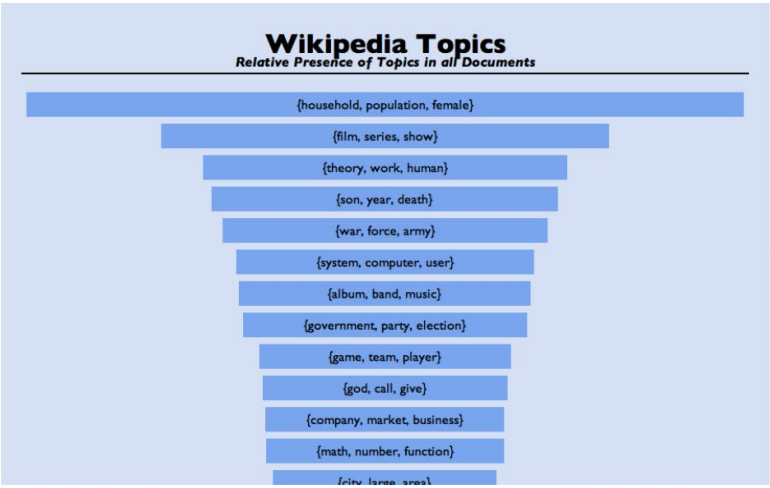
Topic Models

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 Models

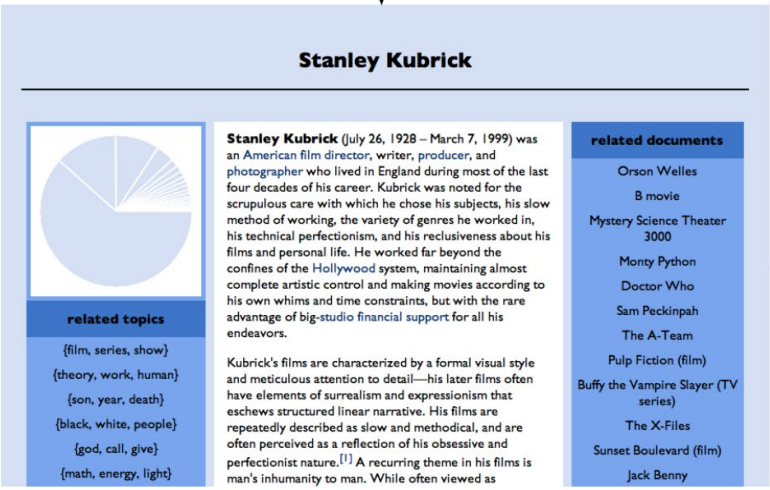


Topic Models



{film, series, show}

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



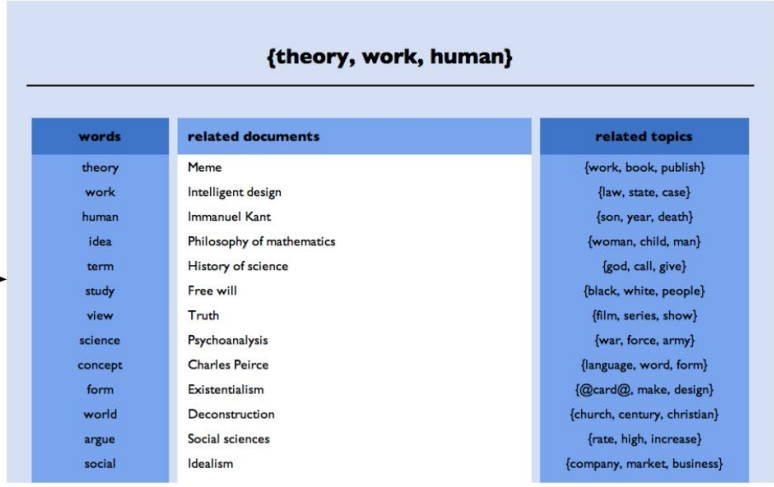
Stanley Kubrick

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 man's inhumanity to man. While often viewed as

related topics
{film, series, show}
{theory, work, human}
{son, year, death}
{black, white, people}
{god, call, give}
{math, energy, light}

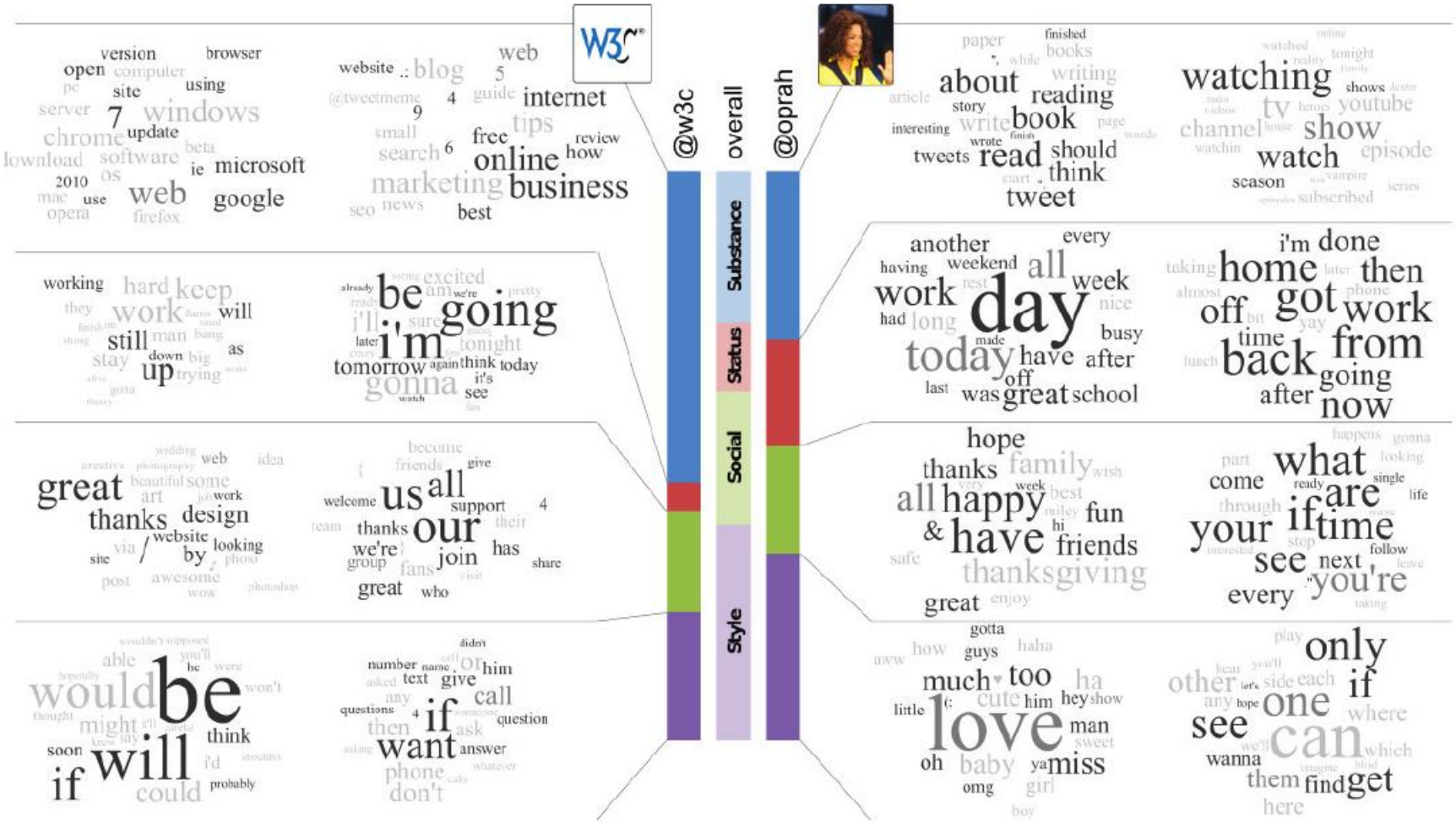
related documents
Orson Welles
B movie
Mystery Science Theater 3000
Monty Python
Doctor Who
Sam Peckinpah
The A-Team
Pulp Fiction (film)
Buffy the Vampire Slayer (TV series)
The X-Files
Sunset Boulevard (film)
Jack Benny



{theory, work, human}

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

Twitter topics



Daniel Ramage, Susan Dumais, Dan Liebling, ICWSM 2010

Overview

- Introduction and motivation for topic modelling
- Probability and statistics
- High level explanation of topic modelling
- More detailed explanation of topic modelling
  - Model
  - How to fit models (briefly)

Sample Space and Events

- Sample space Ω: all possible outcomes of an experiment
  - E.g., roll dice experiments: $\Omega = \{1,2,3,4,5,6\}$
  - For each $w \in \Omega$ we have a probability P(w) with:
    - $0 \leq P(w) \leq 1$ and $\sum_w P(w) = 1$
- Event E: a set of outcomes with certain probability P(E)
  - Any subset of Ω is a possible event
  - E.g., E = {1,2,3,4} is the event that dice roll comes < 5
  - Calculate probability of an event: $P(E) = \sum_{w \in E} P(w)$
  - P({1,2,3,4})=P(1)+P(2)+P(3)+P(4)=4/6=0.67

Random Variables

- Not convenient for large event spaces
- Random variable is a variable that helps to describe outcomes and events:
  - Discrete: possible values from a countable domain.
    - If X is the outcome of a dice throw, then $X \in \{1,2,3,4,5,6\}$
  - Boolean random variable $X \in \{True, False\}$
    - X is whether a dice will come less than 5
    - X is whether the Australian PM in 2100 will be a robot
    - X is whether a patient have Ebola
  - Continuous random variable: possible values from a continuous (infinite) domain
    - X is the height of the class
    - X is how far a car travels with 50L of petrol

How to Summarise the Probabilities of P(X)?

- Prior probabilities can be described by **probability distributions**.
- Weather is one of Sunny, Rain, Cloudy, Snow

Weather	Sunny	Rain	Cloudy	Snow
Probabilty	0.6	0.1	0.29	0.01

- P(Weather) = < 0.6, 0.1, 0.29, 0.01
- P(Weather) follows multinomial d

Need to sump up to 1!

Joint Distribution

- Instead of one random variable, the world can be described by two (or more) random variables:
  - Roll two dice, X = roll of first dice, Y = roll of second dice
  - X = sunny, Y = cold, Z = headache
- Joint probability distribution
  - Specification of probabilities for all combination of events.

sunny

	cold	~cold
headache	0.108	0.012
~headache	0.016	0.064

~sunny

	cold	~cold
headache	0.072	0.008
~headache	0.144	0.576

P(headache ∧ sunny ∧ cold) = 0.108

P(headache∧~sunny∧~cold) = 0.008

Joint Distribution

- Given two random variables A and B:

Joint Distribution:

$$\Pr(A = a \wedge B = b)$$
$$\Pr(a, b)$$

Marginisation (sumout rule):

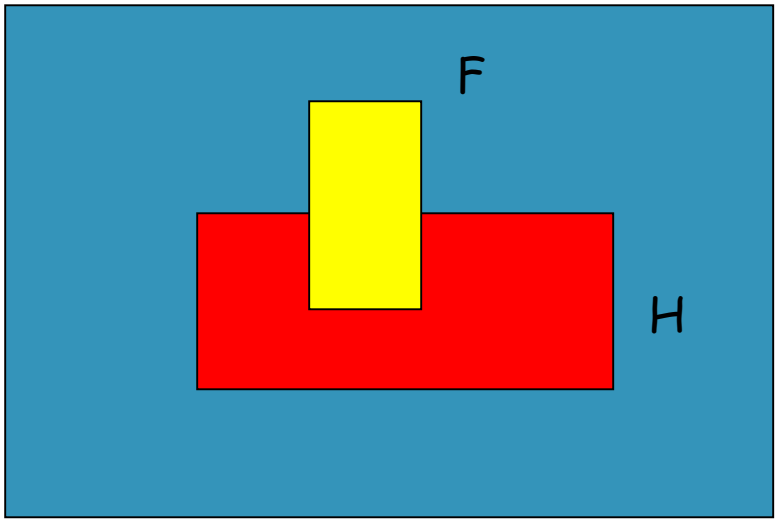
$$\Pr(A = a) = \sum_b \Pr(A = a \wedge B = b)$$
$$\Pr(B = b) = \sum_a \Pr(A = a \wedge B = b)$$

Conditional Probability

- $\Pr(A \mid B)$= probability of A being true given that we know B
- For example: $\Pr(cavity \mid toothache \wedge headache)$
- More general: $\Pr(illness \mid symptoms)$
- $\Pr(A \mid B)$ fraction of worlds in which B is true that also have A true.

H="Have headache"
F="Have Flu"

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2



Headaches are rare and flu is rarer, but *if you have the flu, then there is a 50-50 chance you will have a headache.*

Bayes Rule

- Motivation: when calculating $\Pr(B \mid A)$ we often know distribution $\Pr(A \mid B)$
- For example:
  - A represents symptoms evidences
  - B represents illness or condition
- $P(A|B)P(B) = P(A \wedge B) = P(B \wedge A) = P(B|A)P(A)$
- Bayes rule:
  - $P(\mathrm{B}|\mathrm{A}) = \frac{P(A|B)P(B)}{P(A)}$

Using Bayes Rule for Inference

- Often we want to form a hypothesis about the world based on what we have observed.
- Bayes rule: allows us to state the belief given to hypothesis H, given evidence e.

Likelihood

Prior probability

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)}$$

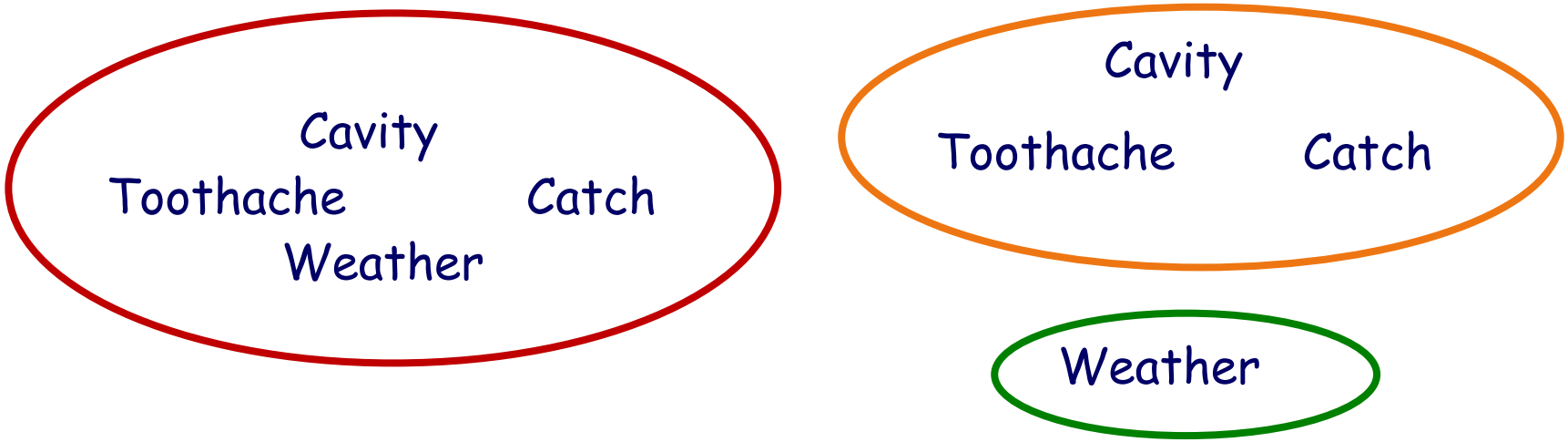
Posterior probability

Normalizing constant

Independence

- A and B are independent iff

  $P(A|B) = P(A)$ or $P(B|A)=P(B)$ or $P(A,B)=P(A)P(B)$



P(Toothache, Catch, Cavity | Weather) = P(Toothache, Catch, Cavity)

P(Toothache, Catch, Cavity, Weather) = P(Toothache, Catch, Cavity) P(Weather)

What good is independence?

- Suppose (say, boolean) variables $X_1, X_2, \ldots, X_n$ are mutually independent
  - We can specify full joint distribution using only n parameters (linear) instead of (exponential)
- How? Simply specify $P(X_1), \ldots, P(X_n)$
  - $P(X_1, \ldots, X_n) = P(X_1)P(X_2) \ldots P(X_n) = \prod_{i=1}^{n} P(X_i)$

Conditional Independence

- Independence is very strong and hence rare …
- Conditional independence is more common

  Consider P(catch | toothache, cavity)

depends on

but not on

P(catch | toothache, cavity) = P(catch| cavity)

P(catch | toothache, ¬ cavity) = P(catch| ¬ cavity)

P(Catch | Toothache, Cavity) = P(Catch| Cavity)

Conditional Independence

- x and y are conditionally independent given z *iff*
  - $P(x|y,z) = P(x|z)$ or
  - $P(y|z)=P(y|x,z)$ or
  - $P(x,y|z)=P(x|z)P(y|z)$
- e.g., learning someone's mark on SMAN exam can influence the probability you assign a specific GPA to them; but if you already knew the final grade, learning the exam mark would not influence your GPA assessment

Overview

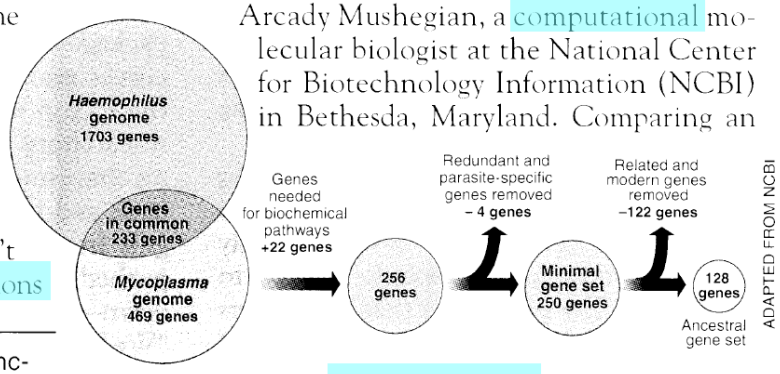
- Introduction and motivation for topic modelling
- Probability and statistics
- High level explanation of topic modelling
- More detailed explanation of topic modelling
  - Model
  - How to fit models (briefly)

Latent Dirichlet Allocation (LDA)

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



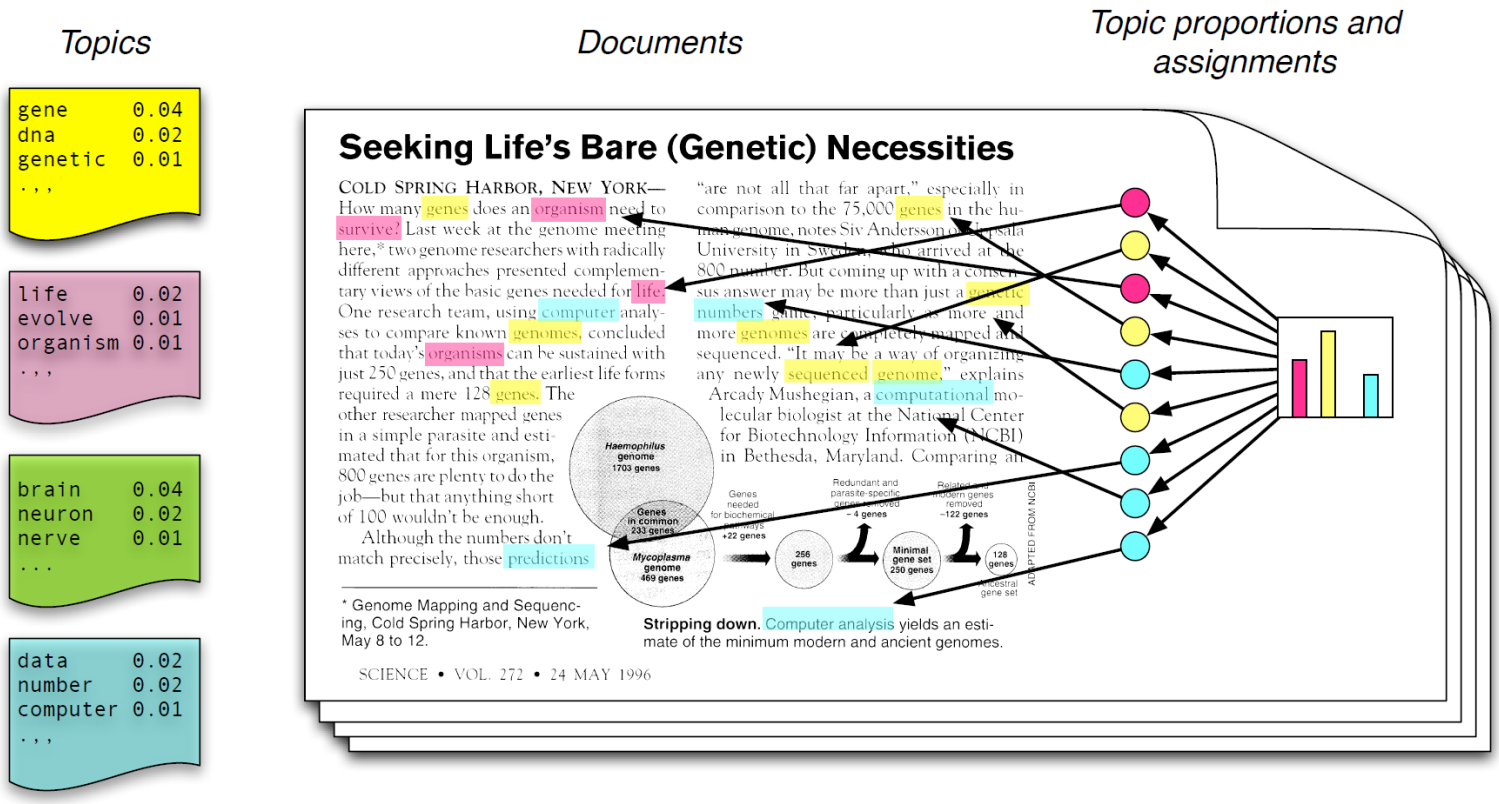
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

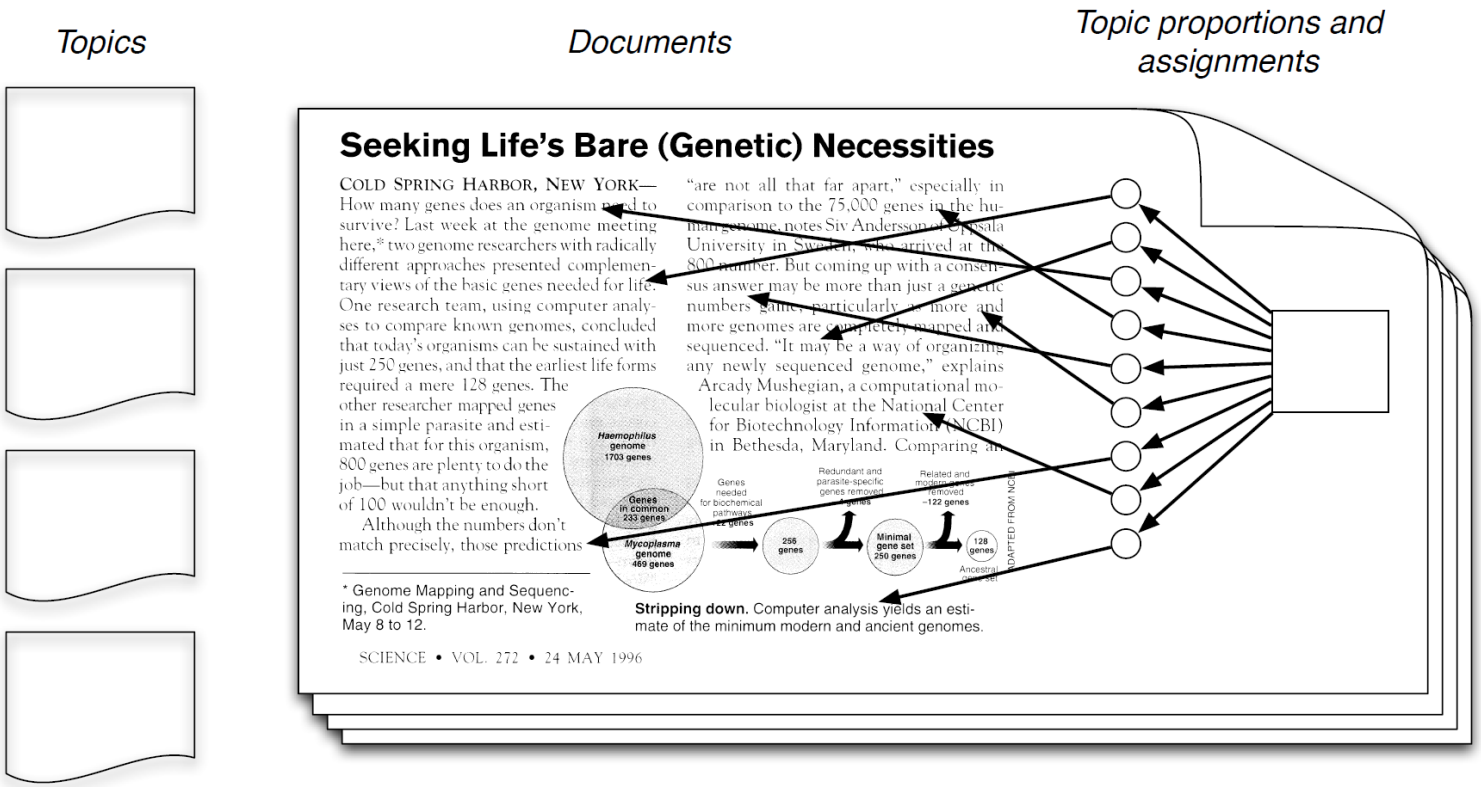
Simple Intuition: Documents consists of multiple topics

Latent Dirichlet Allocation (LDA)



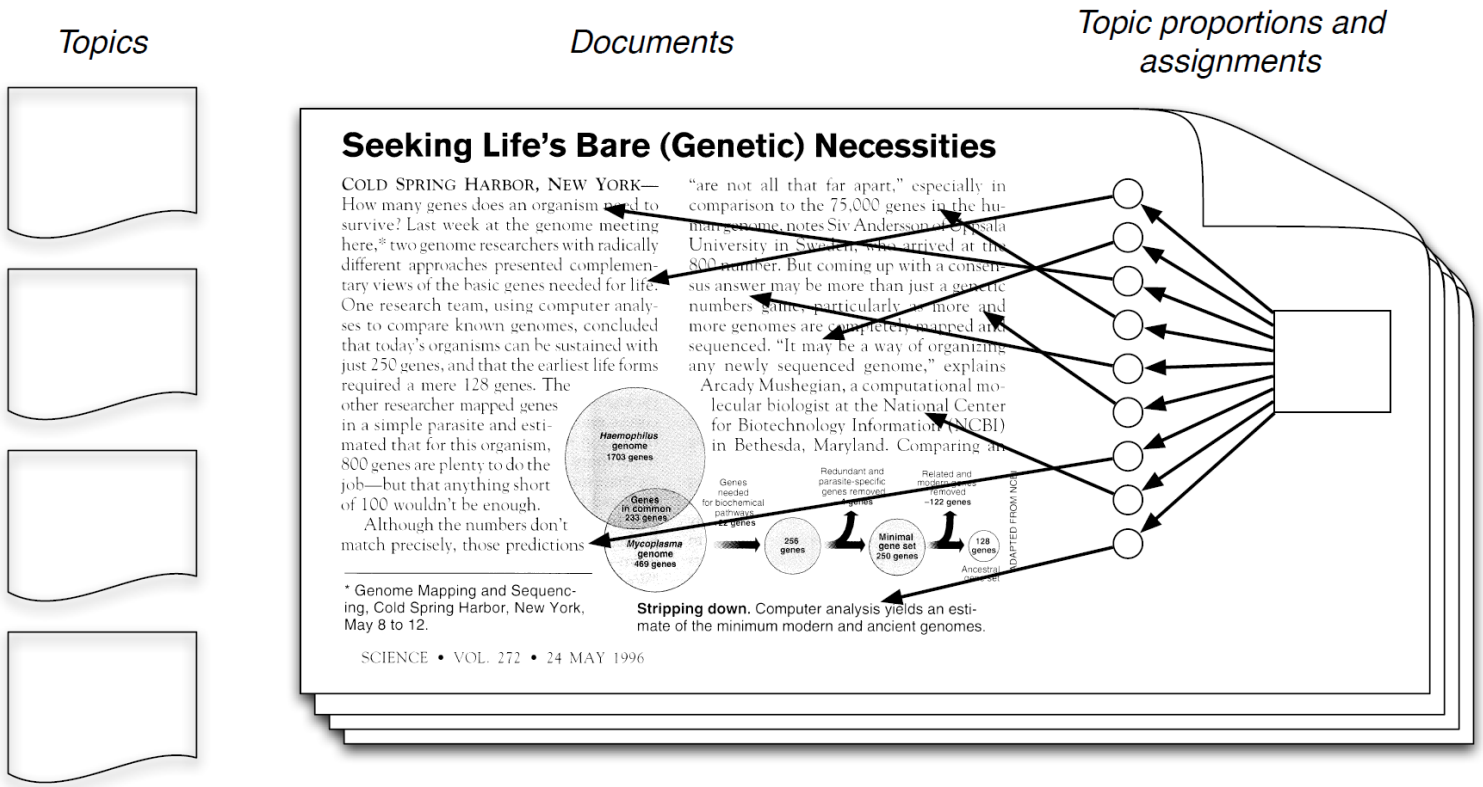
- Each **topic** is a distribution of words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Latent Dirichlet Allocation (LDA)



- In reality, we only observe the documents
- The other structures are **hidden variables**

Latent Dirichlet Allocation (LDA)



- Our goal is to **infer** the hidden variables
- Compute their distribution conditioned on the documents

  p(topics, proportions, assignments | documents)

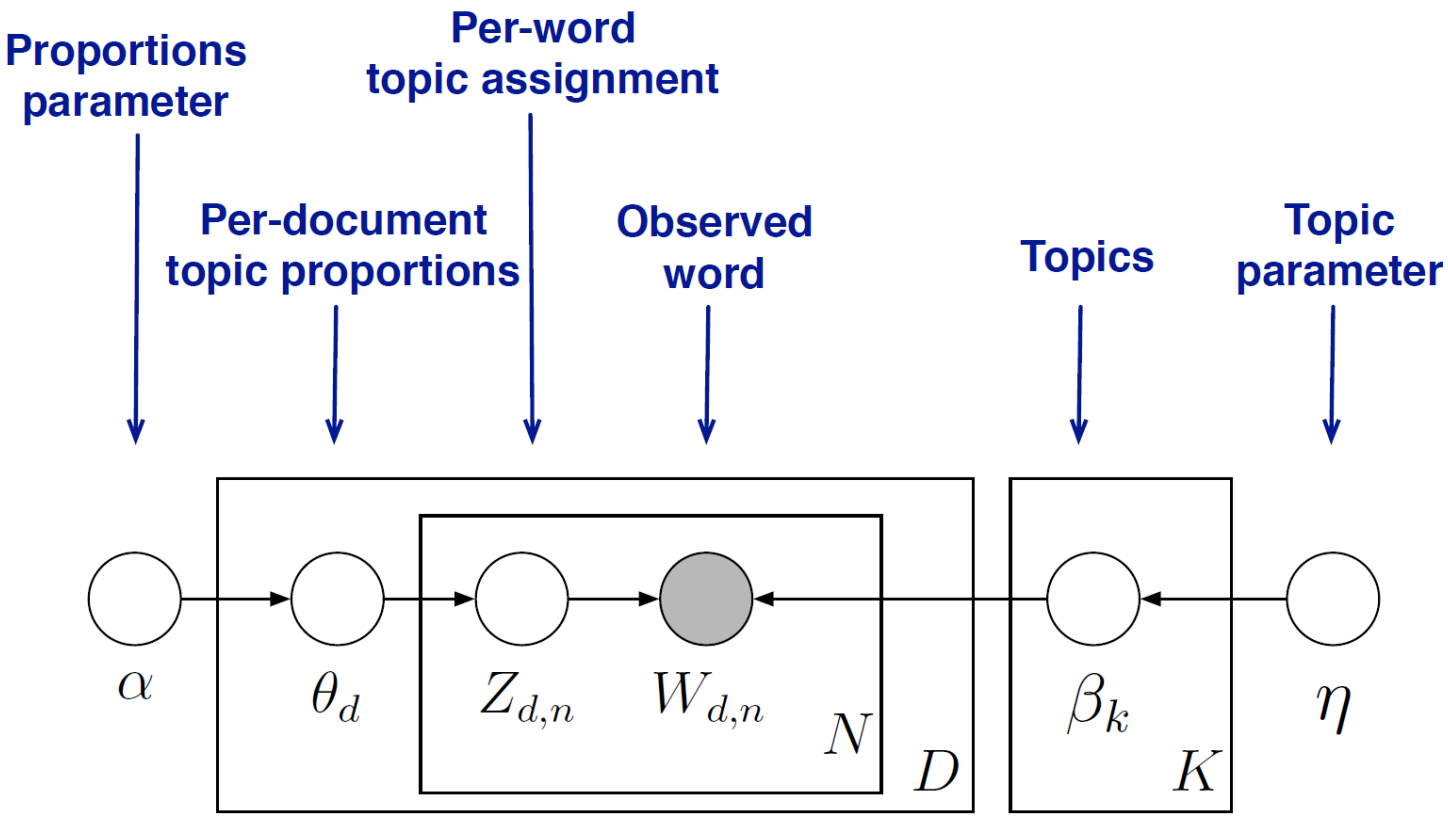
Overview

- Introduction and motivation for topic modelling
- High level explanation of topic modelling
- More detailed explanation of topic modelling
  - Model
  - How to fit models (briefly)

Latent Dirichlet Analysis (LDA) Model

- LDA is a mixture of topics, that outputs words with certain probabilities
- It assumes a document is produced/written as follows:
  - There is K topics that have a distribution of words,
    - e.g, genetics topic, we might generate the word “gene” with probability 4%, “dna” with probability 2% and “genetics” with probability 1% etc
  - Decide on the number of words N the document will have
  - Choose a topic mixture for the document
    - E.g., if genetics and evolution topics have 30% and 20% chance appearing, you might choose the document to be 30% about genetics and 20% about evolution.
  - Generate each word w_i in the document by:
    - First picking a topic (according to the distribution we sampled above, e.g., 30% probability been genetics and 20% been evolution)
    - Using the topic selected to generate the word itself, according to the topics distribution, e.g., if we selected genetics topic, we generate the word “dna”
- Using this model, LDA starts from the words and try to infer the topics

LDA model

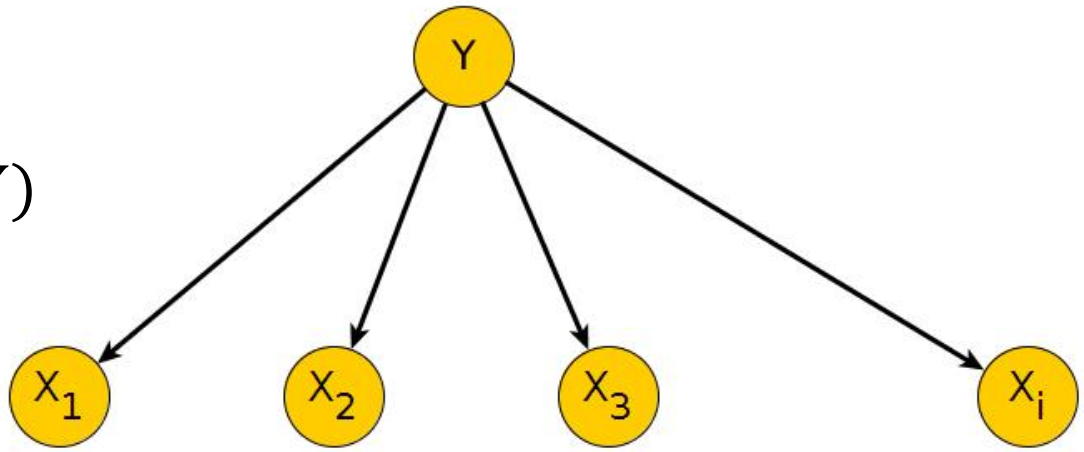


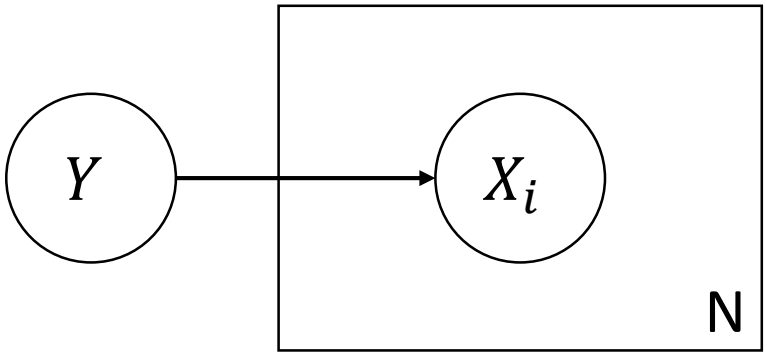
$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) =$$

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

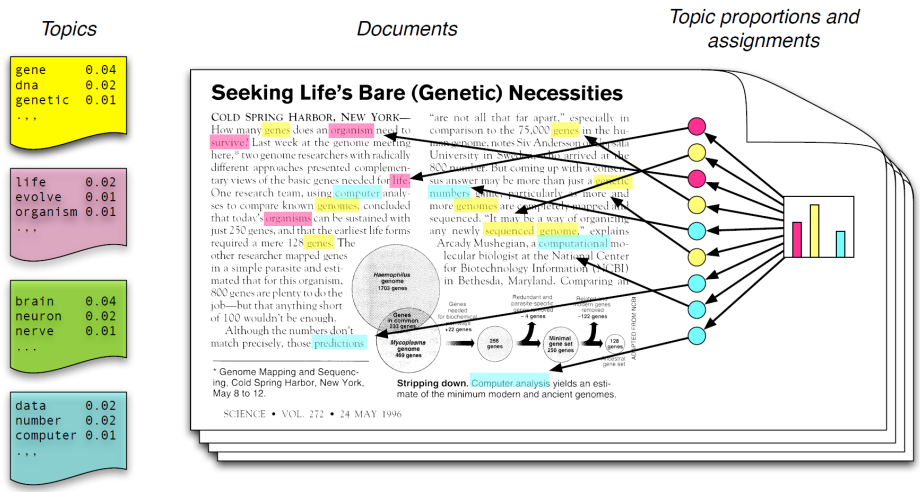
Plate Notation

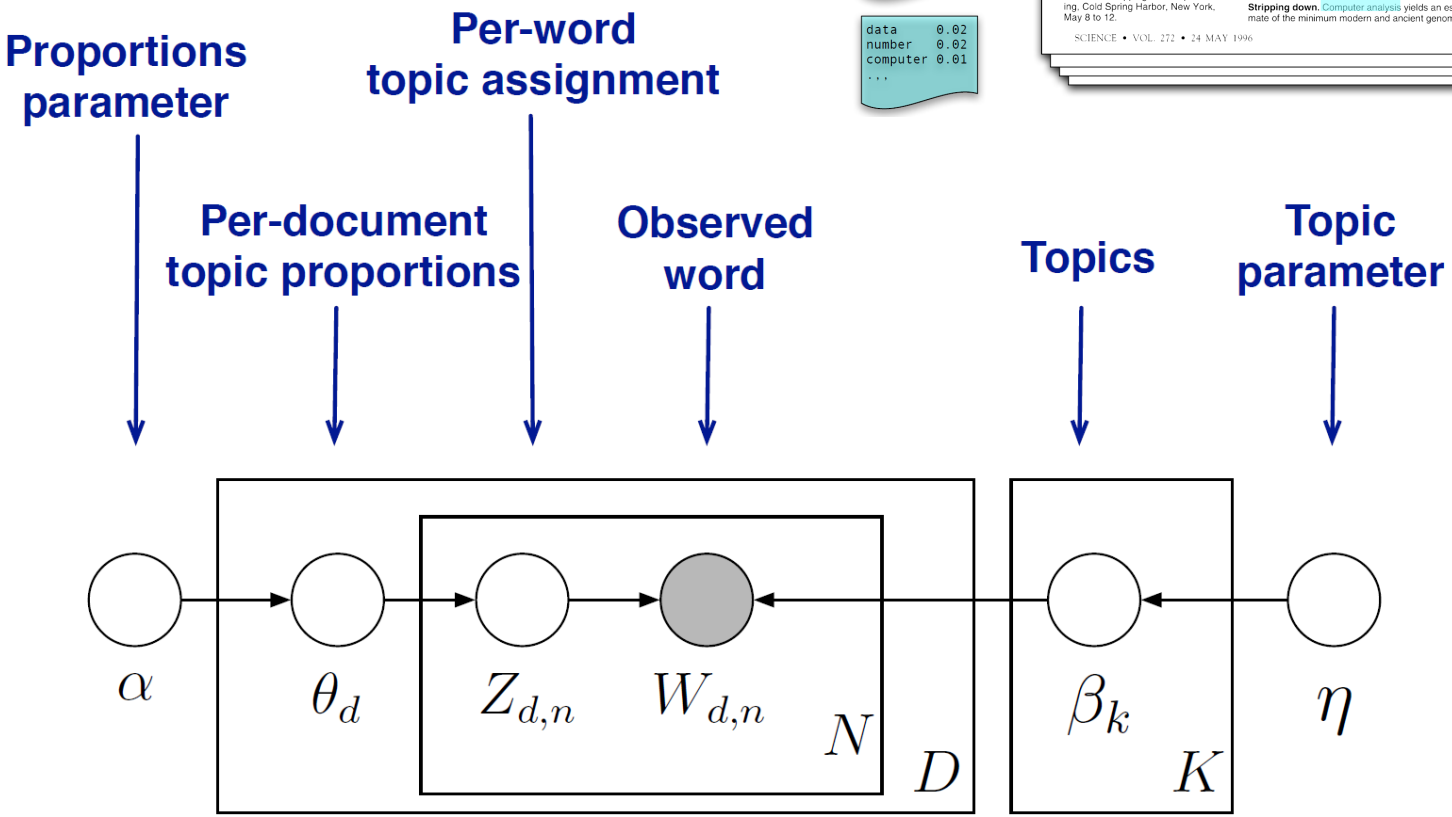
$$P(Y, \boldsymbol{X}) = P(Y) \prod_{i}^{N} P(X_i | Y)$$



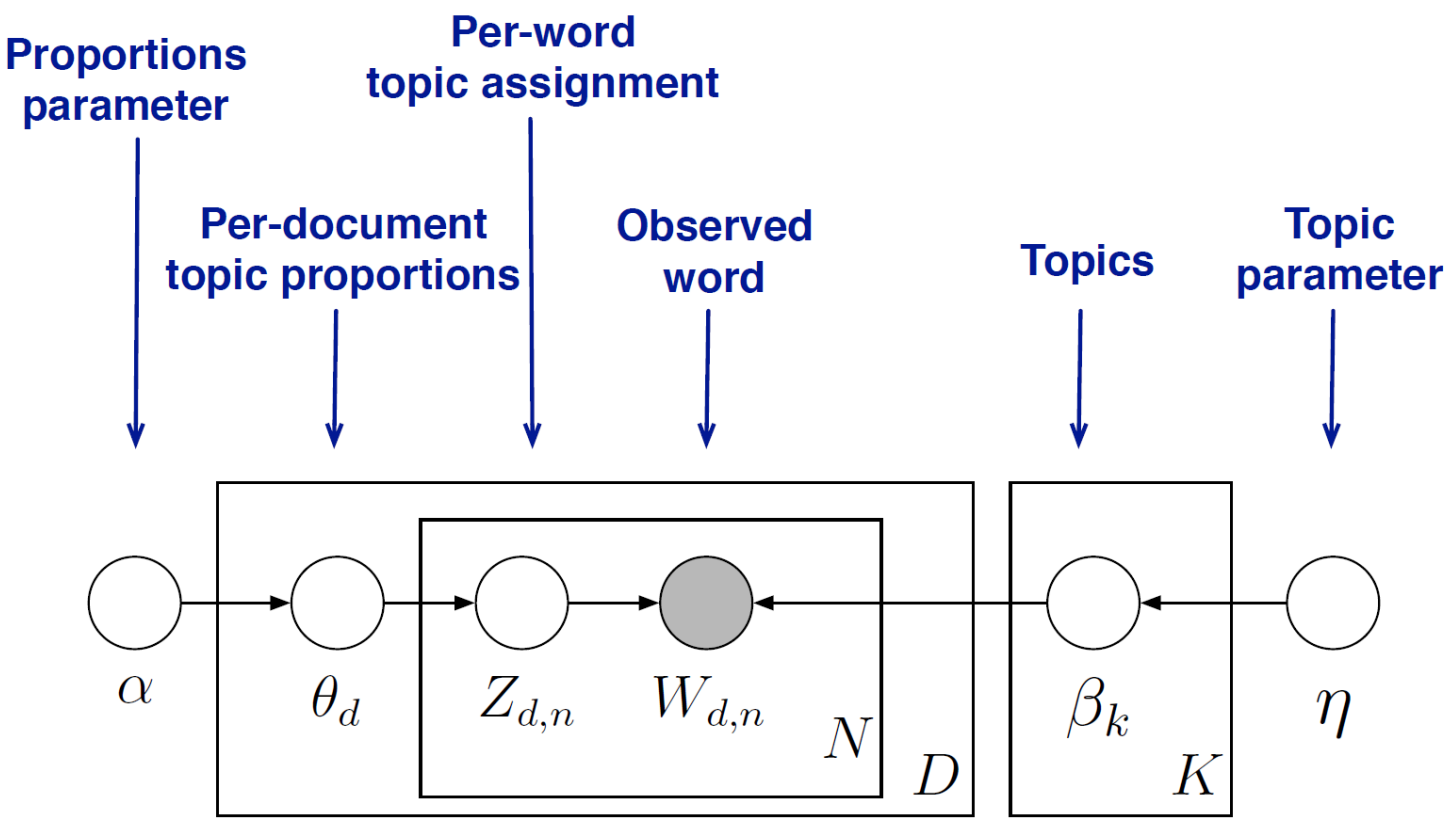


LDA as a graphical model



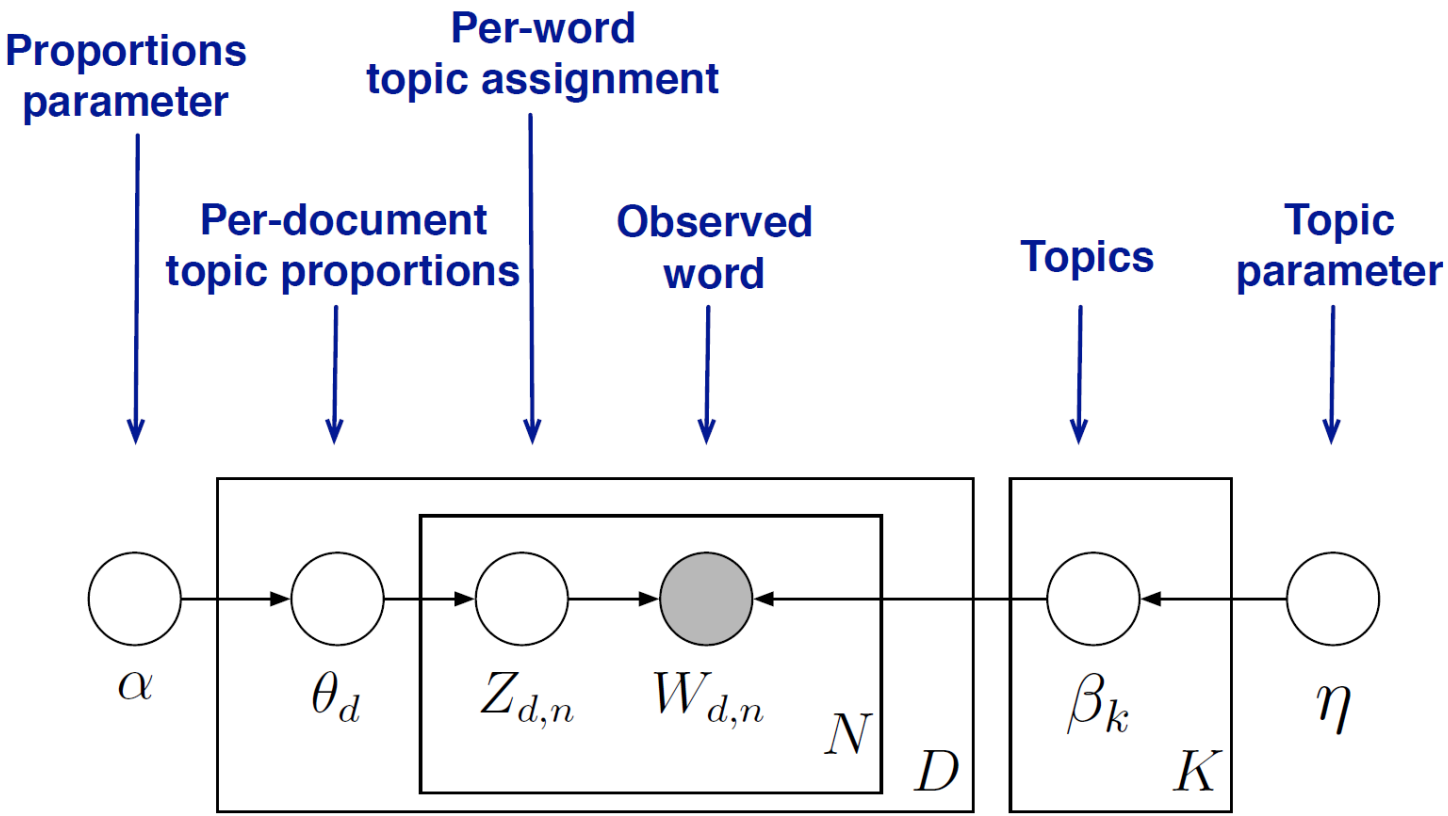


LDA as a graphical model



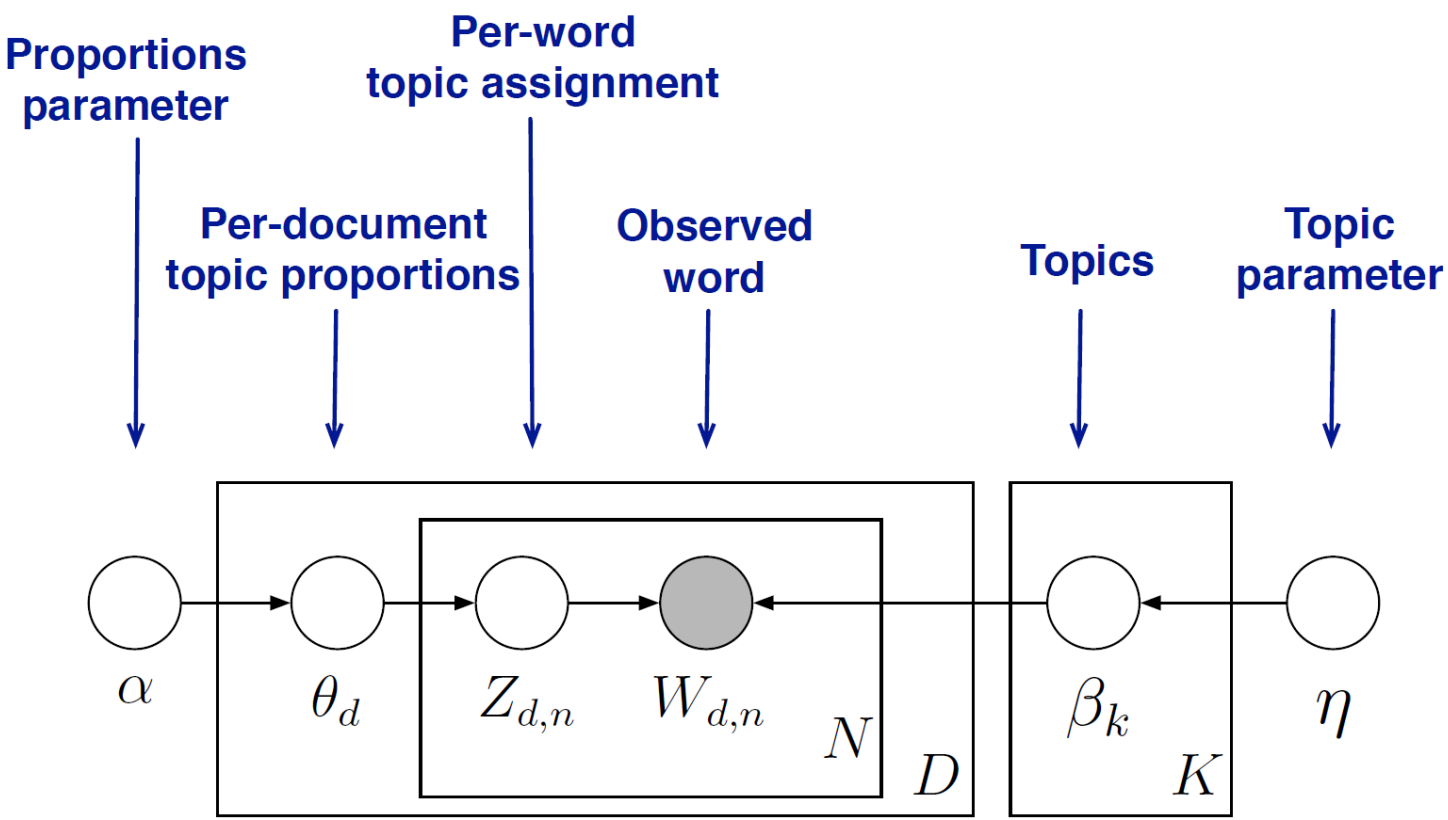
- Nodes are random variables, edges indicate dependence
- Shaded nodes are observed
- Plates indicate replicated variables

LDA as a graphical model



- Encodes model **assumptions**
- Defines a **factorisation** over the joint distribution
- Can use the large range of **algorithmic** approaches for model fitting

LDA as a graphical model



$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) =$$

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

Example Inference

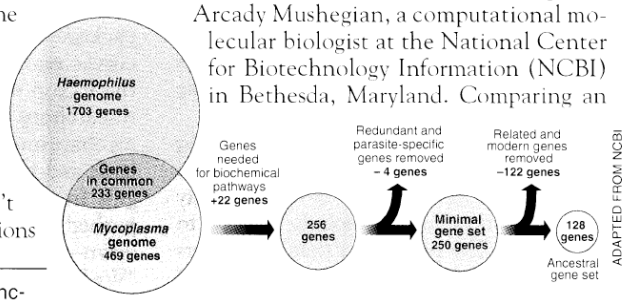
- Data: The OCR'ed collection of Science from 1990-2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- Model 100-topic LDA model using variational inference

Example Inference

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.

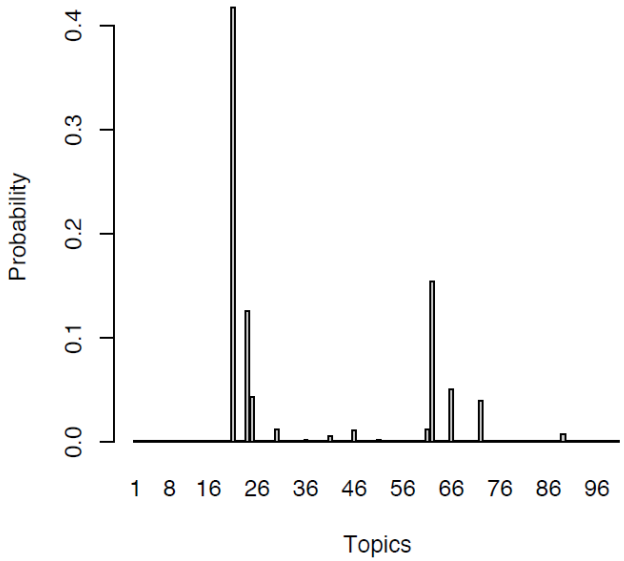
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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.

SCIENCE • VOL. 272 • 24 MAY 1996



Example Inference

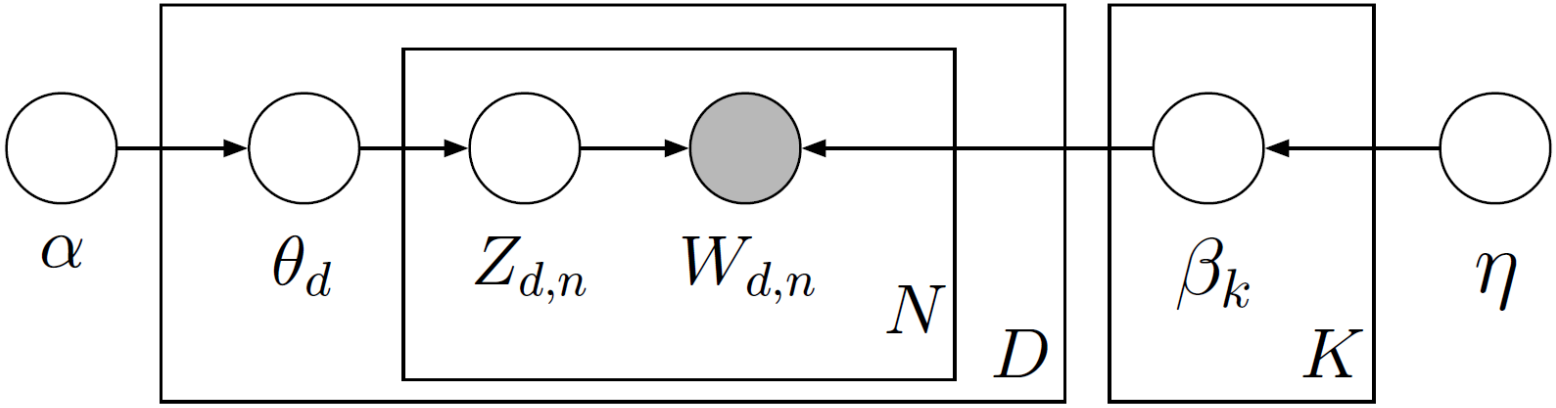
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living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

Example Inference

Why does LDA “work”?

- LDA encourages topics to have few prominent words, and documents to have few prominent topics
- It is a mixture, this find clusters (topics) of co-occurring words
- It is flexible in modelling documents of singleton topics to many topics

LDA Summary

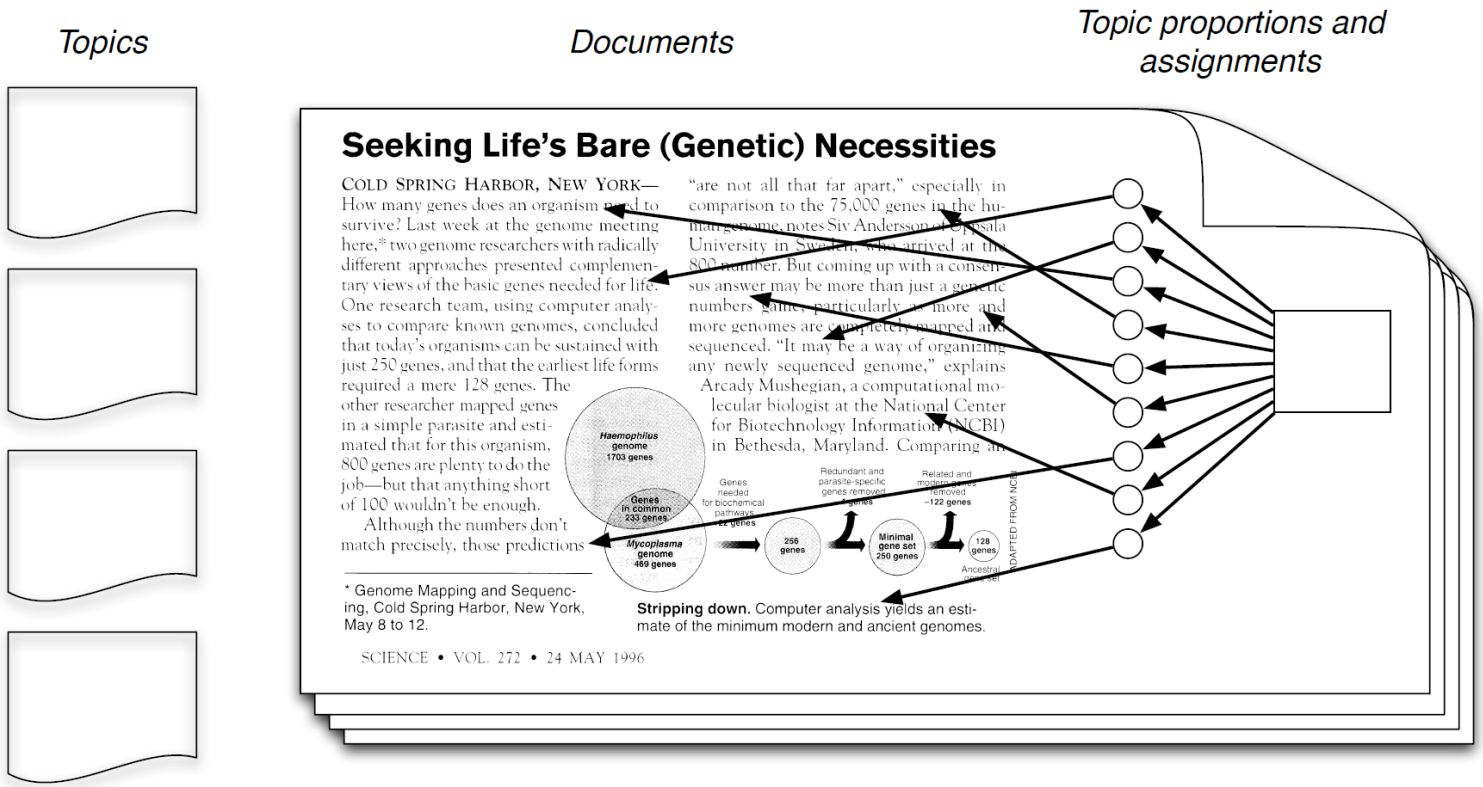


- LDA is a probabilistic model of text. It casts the problem of discovering themes/topics in large collections of documents as a posterior inference problem
- It is a mixture model
- Many more complex models built on top of LDA

Overview

- Introduction and motivation for topic modelling
- High level explanation of topic modelling
- More detailed explanation of topic modelling
  - Model
  - How to fit models (briefly)

Posterior Inference



- Our goal is to compute the distribution of the hidden variables conditioned on the documents

P(topics, proportions, assignments | documents)

Inference Approaches

- Essentially alternate optimising the local and global variables
- Some examples
  - Mean field variational Methods (Blei et al, 2001, 2003)
  - Expectation propagation (Minka et al, 2002)
  - Collapsed Gibbs sampling (Griffith et al, 2002)
  - Collapsed variational inference (Teh et al, 2006)
  - Online variational inference (Hoffman et al, 2010)
  - …

Summary

- Introduced topic modelling
  - Exploratory approach
  - Find topics, particularly in large document sets
- Probabilistic model
  - Intuition and more detailed model