Topic Models

Dr Joemon M Jose & Dr Harry Nguyen
School of Computing Science
University of Glasgow

Why?

- As we have huge archives of news, blogs, web pages, scientific articles, books, images, sound, video, and social networks
  - It has become difficult to find and discover what we are looking for
- Current approach
  - Search
- However, imagine exploring a collection based on the themes run through them
  - Zoom in, zoom out

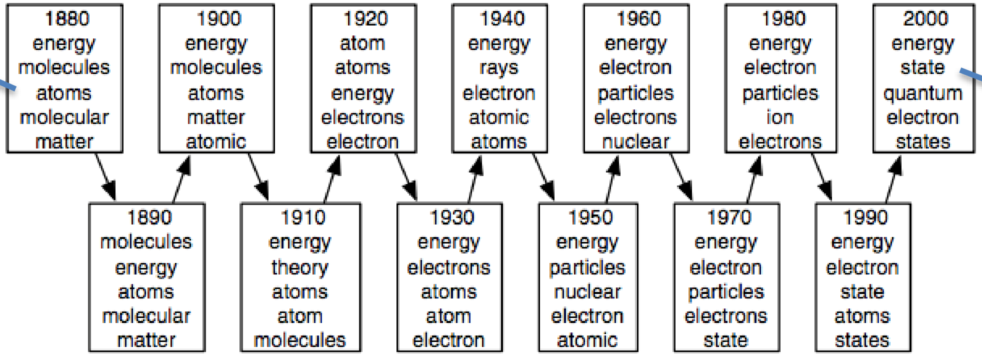
Discuss why searching is limited when exploring a collection

Exploring New York Times

- Themes to explore complete history of NYT
- Sections of the paper
  - Foreign Policy, national affairs, sports, …
- Zoom in on Foreign policy and various aspects of it
  - Chinese foreign policy, ..
- Throughout the exploration, original articles are accessible
- A new way to explore and digest the collection BUT
  - Very time consuming

Science from 1880 - 2002

Energy
Molecules
Atoms
Molecular
matter



1880
energy
molecules
atoms
molecular
matter

1890
molecules
energy
atoms
molecular
matter

1900
energy
molecules
atoms
matter
atomic

1910
energy
theory
atoms
atom
molecules

1920
atom
atoms
energy
electrons
electron

1930
energy
electrons
atoms
atom
electron

1940
energy
rays
electron
atomic
atoms

1950
energy
particles
nuclear
electron
atomic

1960
energy
electron
particles
electrons
nuclear

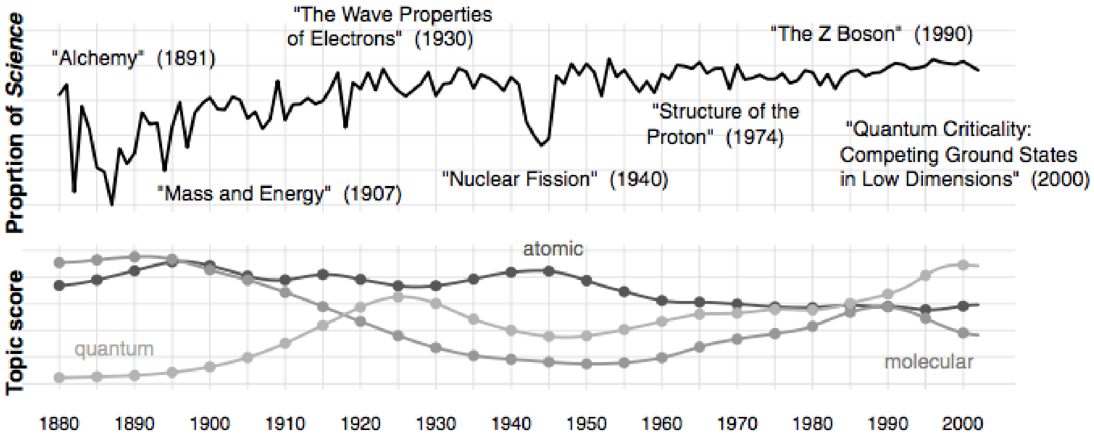
1970
energy
electron
particles
electrons
state

1980
energy
electron
particles
ion
electrons

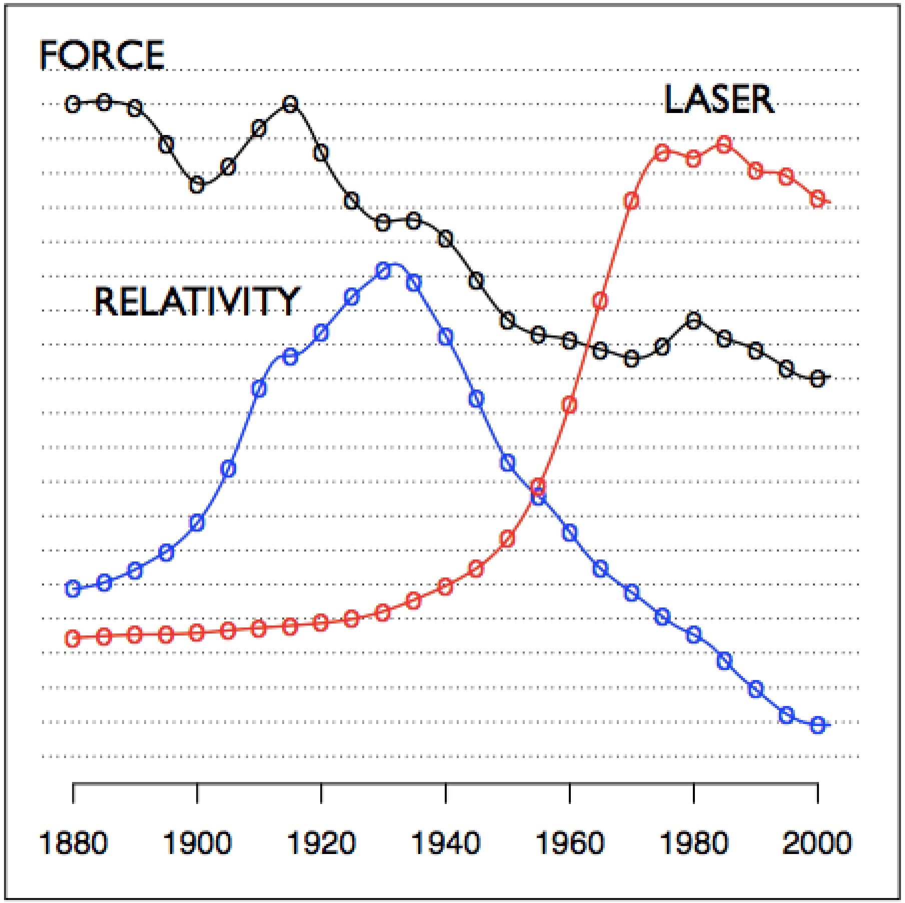
1990
energy
electron
state
atoms
states

2000
energy
state
quantum
electron
states

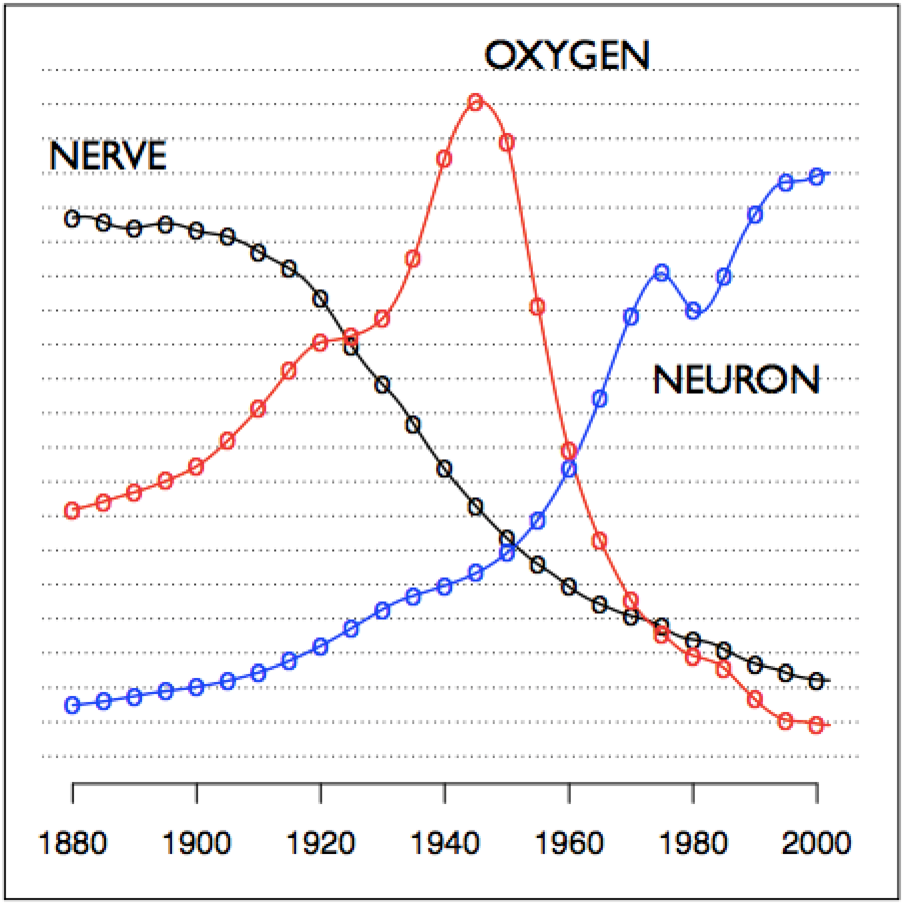
Energy
State
Quantum
Electron
states



"Theoretical Physics"



"Neuroscience"



Problems of Interest

Machine Learning/ Data mining

- What topics does this text collection "span"?
- Which documents are about a particular topic?
- Who writes about a particular topic?
- How have topics changed over time?

- How to represent the "gist" of a list of words?
- How to model associations between words?

Learning Objectives

- Understand topic models
- Discuss the need for topic models
- Look at LDA
- Semantic Topic Modelling

Topic Models

- A suite of algorithms that aim to discover and annotate large archives documents with thematic information
- Topic modelling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through tem,
  - How these themes are connected to each other, and how they change over time
- **They do not require any prior annotation**
  - In machine learning based approach, we often use an annotated collection (in supervised methods)

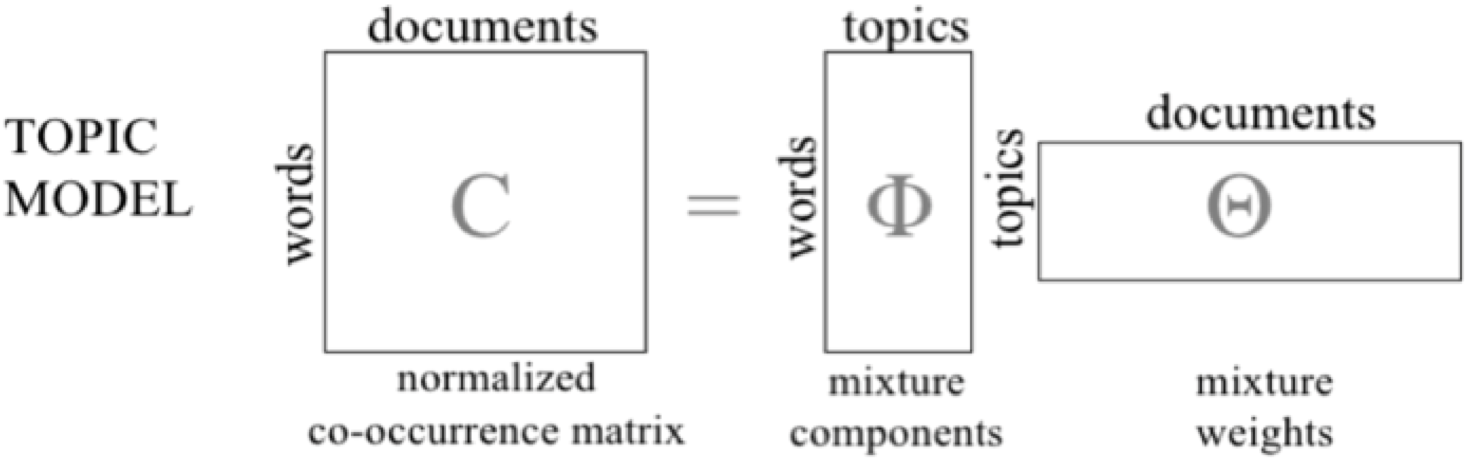
What is a topic model? Why do we need them?
Do we need a training set to generate topic models?

Topic & Intuition

- Topic
  - Not predefined but mined
  - **Defined as a probability distribution over the words/ fixed vocabulary**
    - Still alluding to more general meaning of a theme or subject of discourse
- Intuition
  - **Documents exhibit multiple topics**

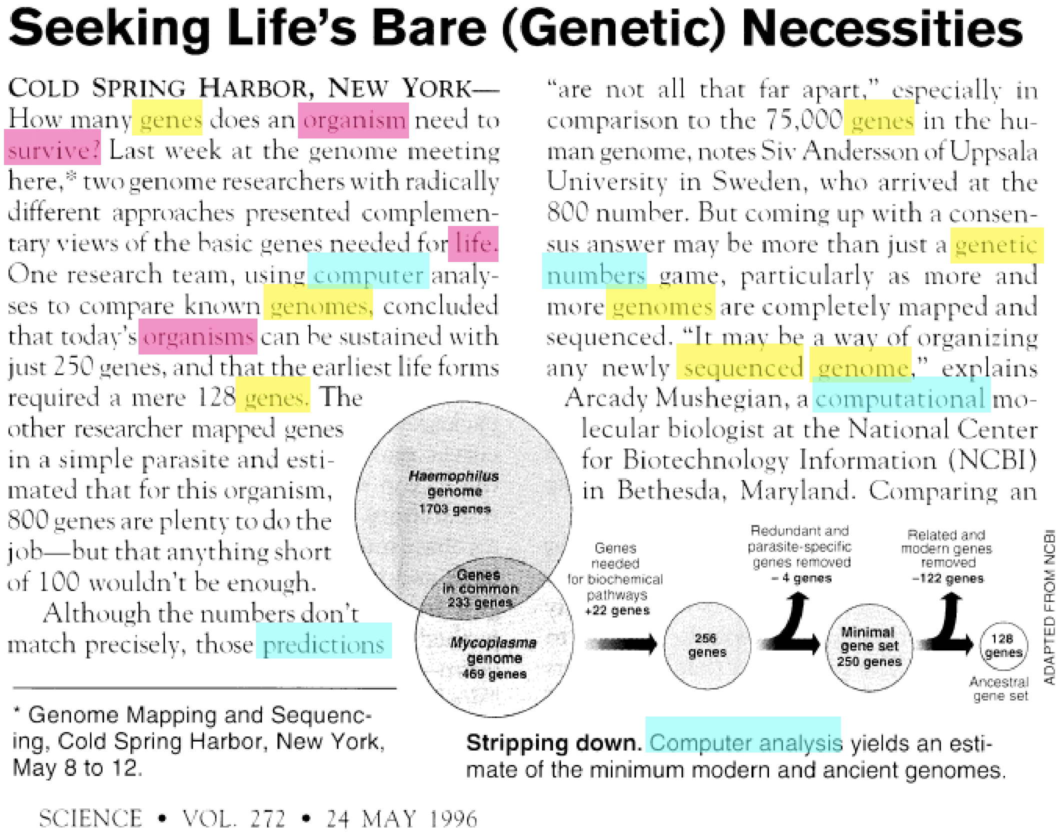
D Blei: Probabilistic Topic Models
doi:10.1145/2133806.2133826

Topic Model



(Kozareva 2013)

Document



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

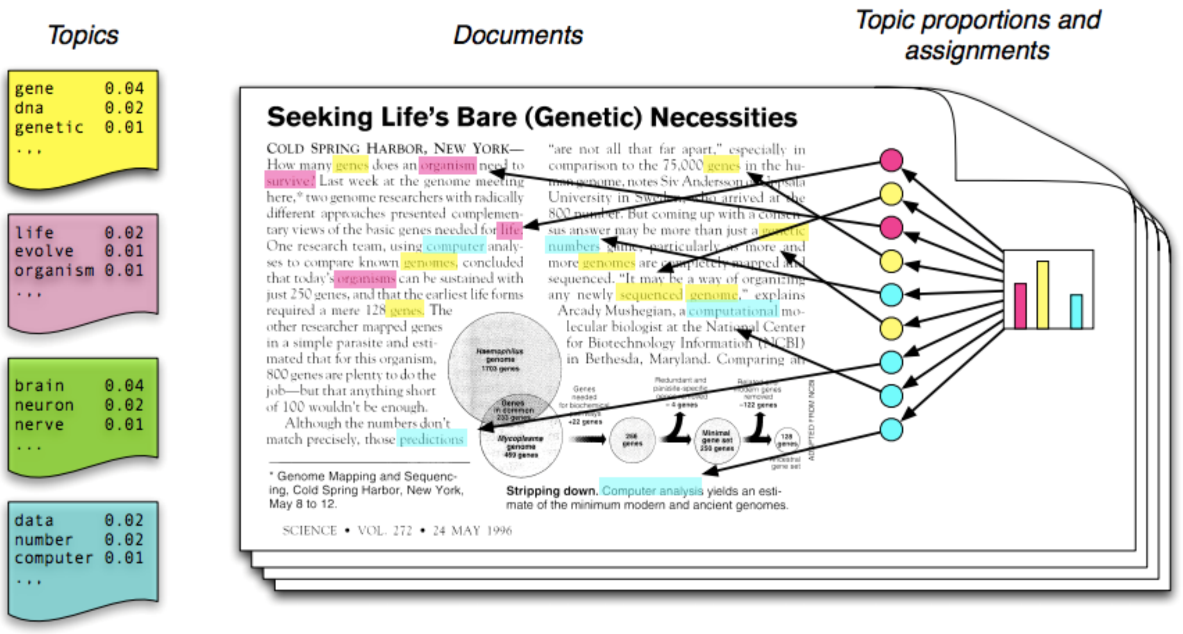
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Intuition behind topic models!

This article blends genetics, data analysis, and evolutionary biology in different proportions


Topics
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Documents
Topic proportions and assignments
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
268 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

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

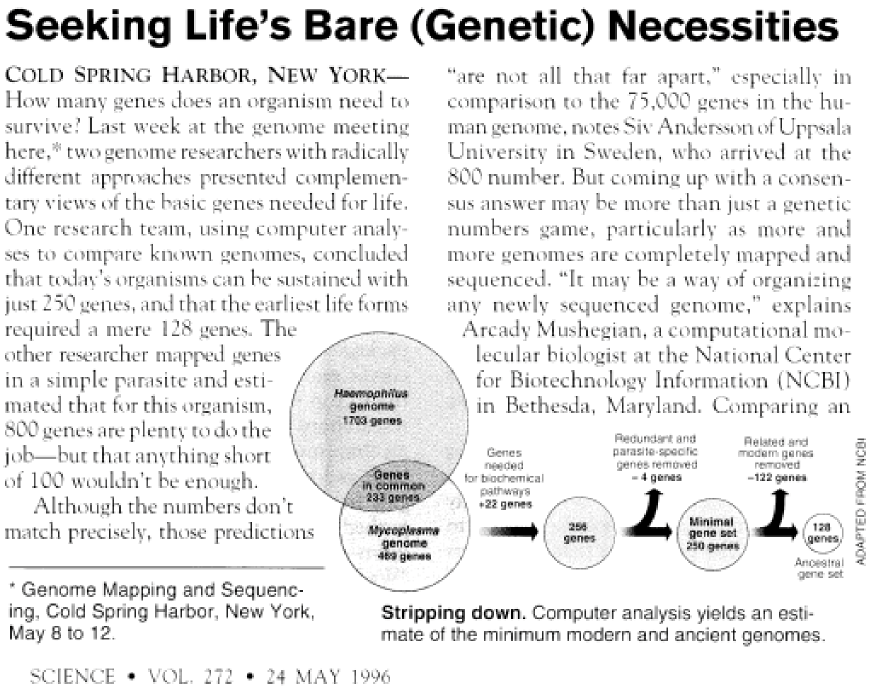
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

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

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.

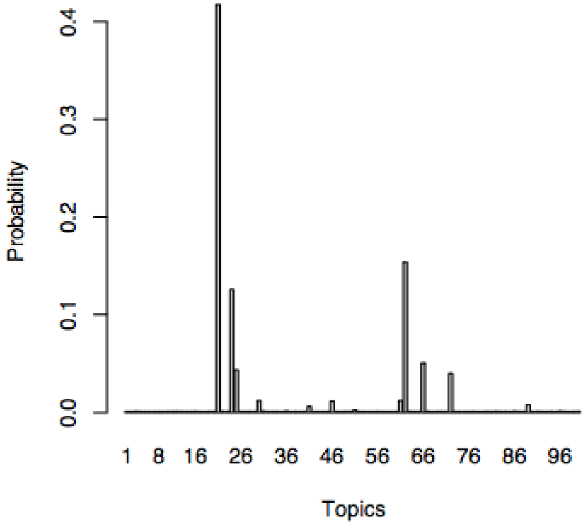
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

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.

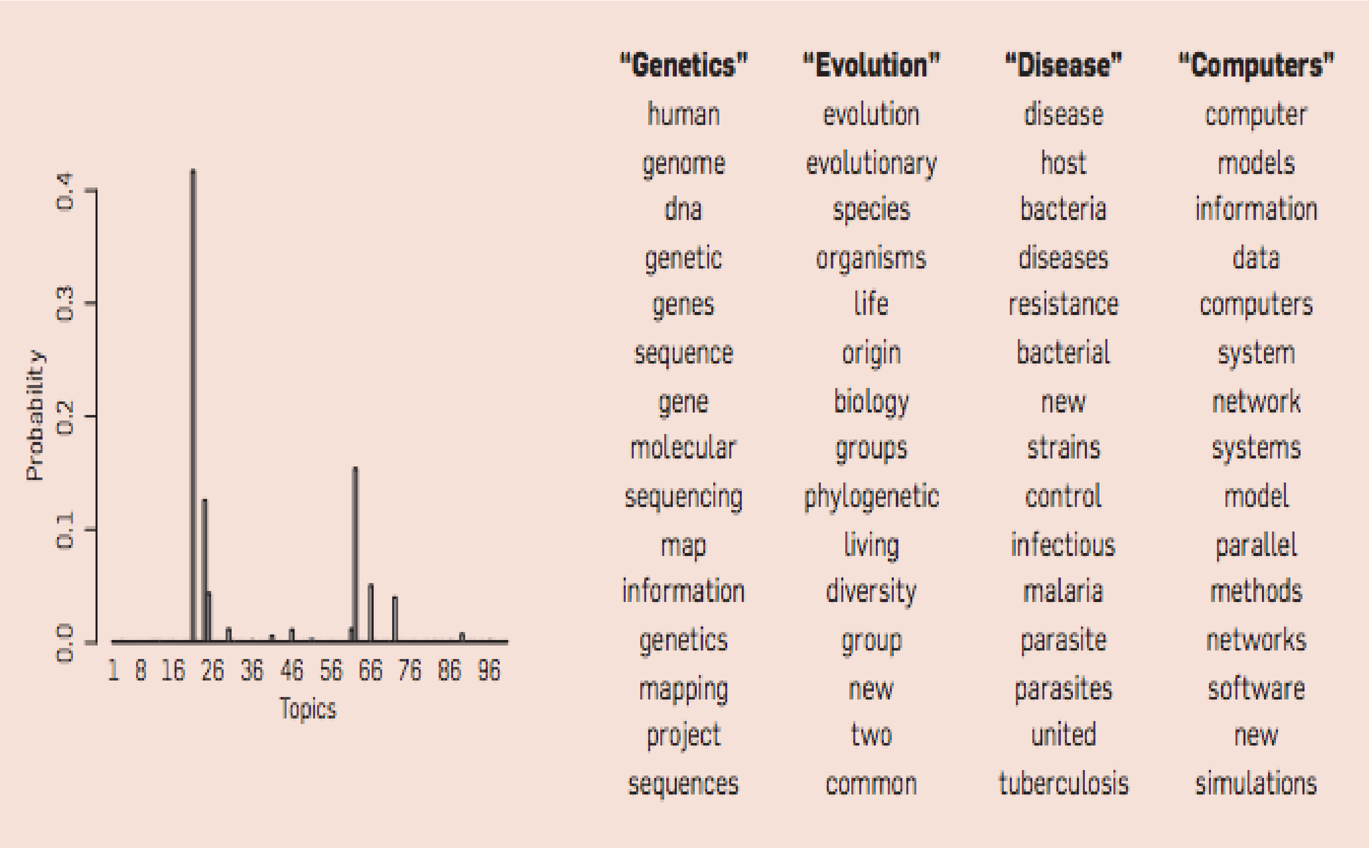
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996



Topic modeling algorithm to explore 17,000 articles; 100 topics assumed

Most probable terms for each of the topics



"Genetics"	"Evolution"	"Disease"	"Computers"
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

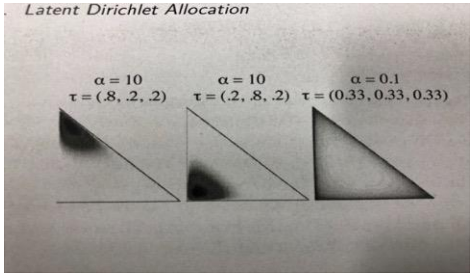
Latent Dirichlet Allocation (LDA)

- Is a statistical model of document collections that tries to capture *this intuition*
- Topic
  - Defined as a distribution over the words/ fixed vocabulary
    - E.g., genetic topic has words about genetics (sequenced, genes) with high probability
    - Evolutionary biology has words like life, organism with high probability

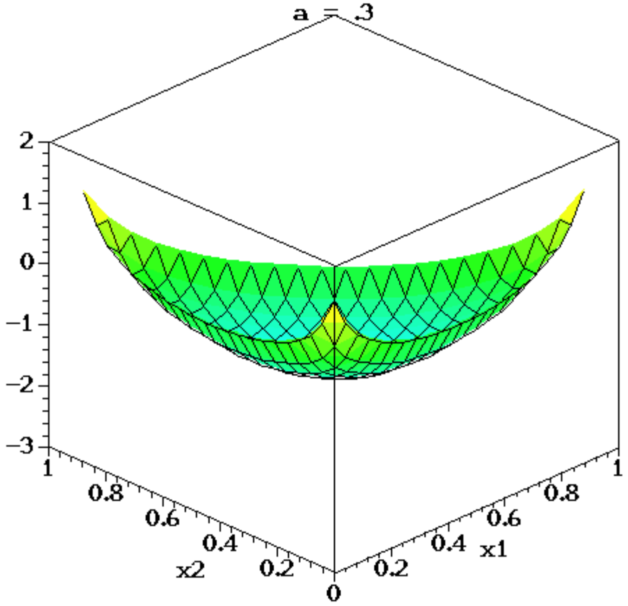
What is an LDA model? What is the underlying intuition ?

Dirichlet

- Dirichlet distributions produce probability vectors that can be used as parameters of discrete distributions
  - Mean – base measure
    - τ – a vector
    - Values you get if you averaged many draws from the Dirichlet
  - A concentration parameter a
    - Controls how far away individuals draws are from the base measure
    - $a_k=a_0\tau_k$



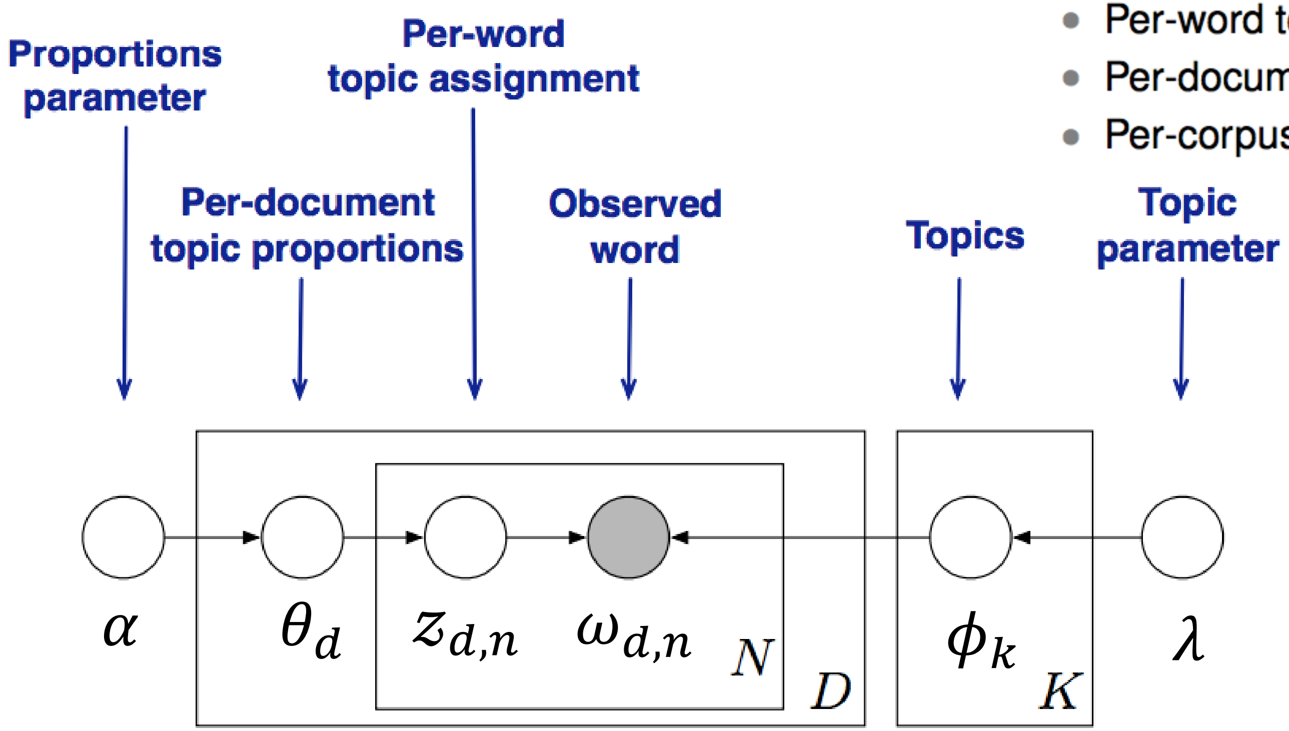
Dirichlet



Graphical Model

From a collection of documents, infer

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



Per-Corpus Topic Distributions

- The user specifies that there are K distinct topics
  - Each of the K topics is drawn from a Dirchlet distribution with a
    - Uniform base distribution (u) and concentration parameter λ

$$f_k \sim Dir(l u)$$

Document allocations

- Distributions over topics of each document

$$q_d \sim Dir(au)$$

LDA Process

- Step #1: Randomly choose a distribution over topics
- Step #2: For each word in the document
  - (#a) Randomly choose a topic from the distribution over topics in step #1
  - (#b) Randomly choose a word from the corresponding distribution in vocabulary

Describe the LDA process

Topic Modelling Approaches

- Number of possible topic structures is exponentially large
- Approximate the posterior distribution
- Topic modelling algorithms form an approximation of equation, by adapting an alternative distribution over latent topic structure to be close to the true posterior

Two approaches:

1. **Sampling based!**
    - Attempt to collect samples from the posterior to approximate it with an empirical distribution – Gibbs sampling!
2. **Variational methods!**
    - Deterministic alternative to sampling based methods
    - Posit a parametrised family of distributions over the hidden structure and then find the member of that family that is closest to the posterior

Gibbs Sampling

- Start with random assignments of words to topics

- Repeat M iterations
  - Repeat for all words *i*
    - Sample a new topic assignment for word *i* conditioned on all other topic assignments

16 Artificial Documents

documents

		●●●●	●●●●●●	●●●●●●
		●●●●●	●●●●●●●	●●●●
		●●●●●●●	●●●●●	●●●●
		●●●●●●●	●●●●●●	●●●
		●●●●●●●	●●	●●●●●●●
		●●●●●●●●●	●●●	●●●●
●		●●●●	●●●●●●	●●●●●
●	●●	●●●●●●	●●●●	●●●
●	●●●	●●●●●●	●●●●	●●
●●	●●●	●●●●●●	●	●●●●
●●	●●●	●●●●●●●	●●●	●
●●●	●●●●●●	●●●●●●	●	
●●●●●●	●●●	●●●●●●		●
●●	●●●●●●●●	●●●●●●		
●●●●	●●●●●●●	●●●●●		
●●●●●	●●●●●●●	●●●●		

Can we recover the original topics and topic mixtures from this data?

Starting the Gibbs Sampling

- Assign word tokens randomly to topics (●=topic 1; ○=topic 2)

		○ ○ ○ ○	● ○ ○ ○ ● ○	● ● ○ ● ○ ○
		○ ○ ● ○ ○	● ● ● ● ● ● ○	● ○ ○ ●
		○ ○ ○ ● ○ ○ ○	○ ● ○ ● ○	● ○ ○ ○
		● ● ● ○ ● ○ ○	○ ● ● ○ ○ ○	○ ○ ○
		● ● ○ ● ○ ● ○	● ○	○ ● ○ ○ ○ ○ ○
		○ ● ● ○ ● ● ● ● ●	○ ● ○	○ ○ ● ●
○		○ ● ● ●	● ● ○ ○ ● ○	○ ● ● ● ○
●	○ ●	○ ○ ● ● ● ●	○ ● ● ○	● ● ○
●	○ ○ ●	○ ○ ○ ○ ○ ●	● ○ ● ●	○ ●
● ○	● ● ○	● ○ ○ ○ ○ ○	●	● ○ ○ ●
○ ●	○ ● ●	○ ○ ○ ● ● ○ ○	● ● ●	●
○ ○ ○	○ ○ ○ ○ ● ○	● ○ ● ● ○ ●	○	
○ ○ ○ ● ● ●	○ ● ○	● ○ ○ ○ ● ●		○
○ ○	● ● ○ ○ ○ ● ● ●	● ● ○ ● ○ ○		
○ ● ● ●	● ● ● ○ ○ ● ○	● ○ ● ○ ●		
● ○ ● ● ○	● ● ○ ○ ○ ○ ●	● ● ● ○		

After 1 iteration

		● ● ○ ○	○ ○ ○ ○ ○ ●	● ○ ○ ○ ○ ○
		● ○ ○ ○ ○	○ ● ● ● ● ● ○	○ ○ ○ ●
		○ ○ ○ ○ ○ ○ ●	○ ○ ○ ○ ●	○ ○ ● ○
		○ ○ ○ ○ ○ ○ ○	● ○ ○ ○ ○ ○	○ ○ ○
		● ● ● ● ● ● ○	● ●	● ○ ● ○ ● ● ●
		● ○ ● ○ ● ● ● ○ ●	● ● ●	● ○ ● ●
●		● ● ● ●	● ● ● ● ● ●	● ● ● ● ●
○	● ○	● ● ● ● ● ●	● ● ● ●	● ● ●
●	○ ● ●	○ ○ ○ ● ○ ○	○ ● ● ●	● ●
○ ●	○ ○ ○	○ ○ ○ ○ ● ○	○	● ○ ○ ○
○ ●	● ● ○	● ● ● ● ● ● ○	○ ○ ●	○
○ ● ●	○ ○ ● ○ ○ ●	○ ○ ○ ○ ○ ○	○	
● ● ● ● ● ○	○ ● ○	○ ○ ○ ● ○ ○		●
● ●	○ ○ ○ ● ○ ○ ○ ○	○ ● ● ● ○ ●		
● ○ ○ ○	○ ○ ○ ○ ● ○ ○	○ ○ ○ ● ○		
● ● ● ● ●	○ ● ○ ● ○ ● ●	● ○ ● ●		

After 4 iterations

		● ● ● ●	● ● ● ● ● ●	● ● ● ● ● ●
		● ○ ○ ● ○	● ○ ● ● ● ● ●	● ● ● ●
		○ ○ ● ○ ● ● ●	● ● ● ○ ●	○ ○ ○ ●
		○ ○ ○ ○ ○ ○ ○	○ ● ● ● ● ○	○ ● ○
		● ○ ● ● ● ● ●	● ●	● ● ● ● ● ● ●
		● ● ● ● ● ● ● ● ●	● ● ●	● ● ● ●
●		● ● ● ●	● ● ● ● ● ●	● ● ● ● ●
○	○ ○	○ ● ○ ○ ● ○	● ○ ● ●	● ● ●
●	● ○ ●	● ● ● ● ● ○	● ● ● ●	● ●
● ○	○ ○ ○	○ ● ● ○ ○ ●	○	● ● ○ ○
○ ○	○ ○ ○	○ ● ○ ● ○ ○ ○	● ○ ○	○
○ ○ ○	○ ○ ○ ○ ○ ○	● ○ ○ ○ ○ ○	○	
○ ○ ○ ○ ○ ○	○ ○ ○	○ ● ● ○ ○ ○		●
○ ○	○ ○ ○ ○ ○ ○ ○ ○	○ ○ ○ ○ ○ ○		
○ ○ ○ ○	○ ○ ○ ○ ○ ○ ○	○ ○ ○ ○ ○		
○ ○ ○ ○ ○	○ ○ ○ ○ ○ ○ ○	○ ○ ○ ○		

After 32 iterations

●

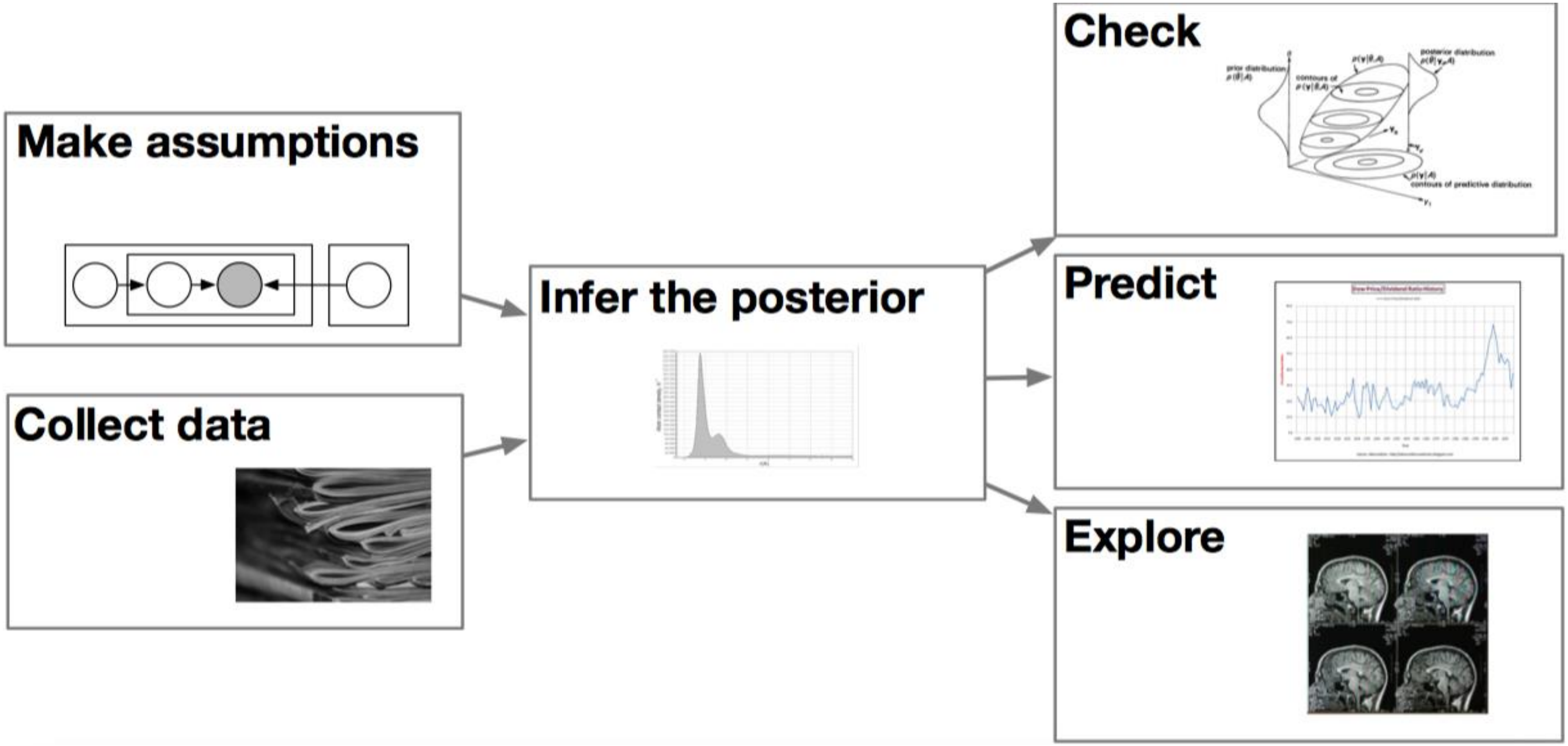
topic 1	
stream	.40
bank	.35
river	.25

○

topic 2	
bank	.39
money	.32
loan	.29

		●●●●	●●●●●●	●●●●●●
		●●●○●	●●●●●●●	●●●●
		●●●●●●●	●●●●●	●●●●
		●●●●●●●	●●●●●●	●●●
		●●●○●●●	●●	●●●●●●●
		●●●●●●●●●	●●●	●●●●
○		●●●○	●●●●●●	●●●●●
○	○○	●○○○●○	●●●●	●●●
○	○○○	●○●●●●	●●●●	●●
○○	○○○	○●●●●	●	●●●●
○○	○○○	○○○●○○○	●●●	●
○○○	○○○○○○	○●○●●●	●	
○○○○○○	○○○	●○○○○○		●
○○	○○○○○○○○	○○○○○○		
○○○○	○○○○○○○	○○○○○		
○○○○○	○○○○○○○	○○○○		

LDA in one picture



Make assumptions

Collect data

Infer the posterior

Check

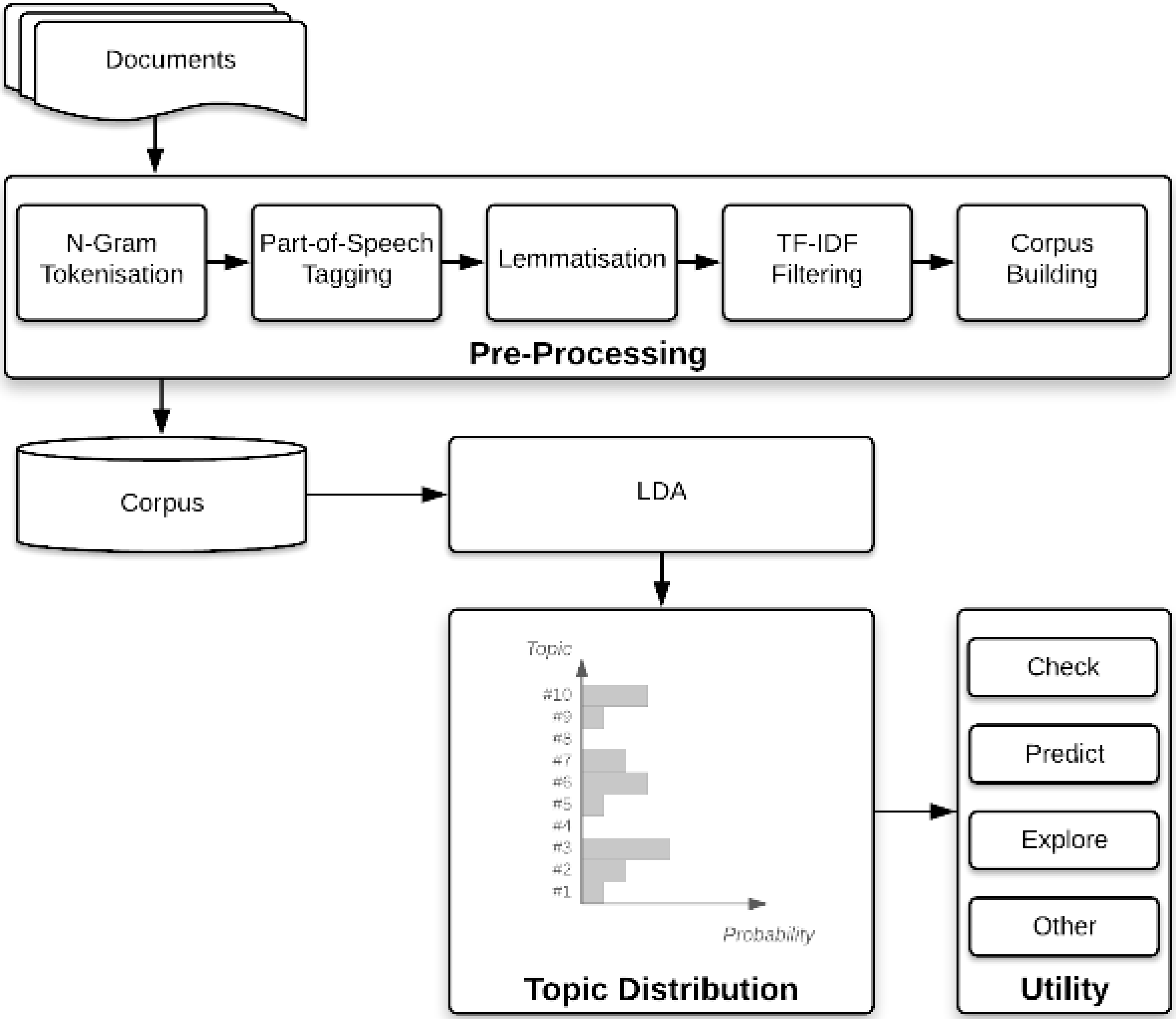
Predict

Explore

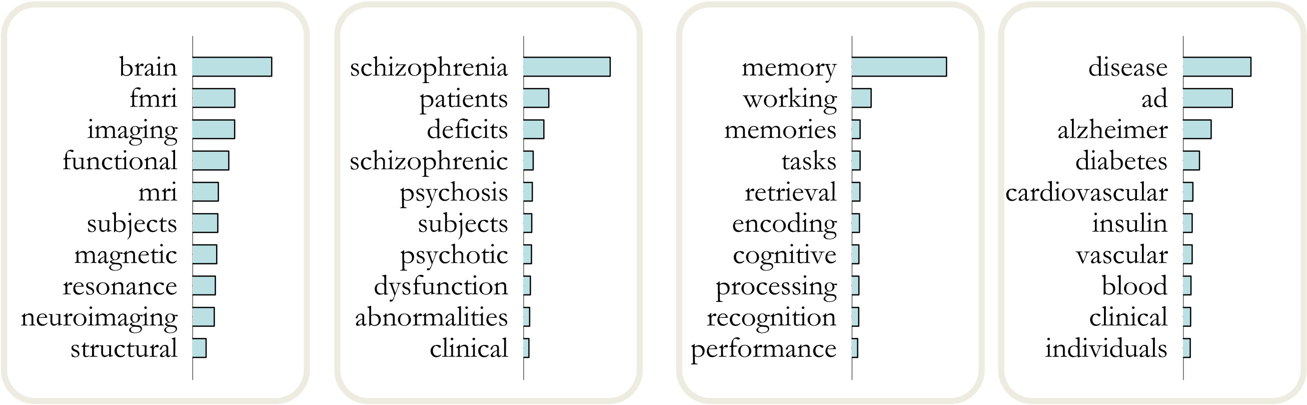
prior distribution $p(\theta|A)$
θ
contours of $p(y|\theta,A)$
$p(y|\theta,A)$
posterior distribution $p(\theta|y_o,A)$
y_o
y_r
$p(y|A)$ contours of predictive distribution
y_1

(Blei 2012)

Practical


Documents
N-Gram Tokenisation
Part-of-Speech Tagging
Lemmatisation
TF-IDF Filtering
Corpus Building
Pre-Processing
Corpus
LDA
Topic
#10
#9
#8
#7
#6
#5
#4
#3
#2
#1
Probability
Topic Distribution
Check
Predict
Explore
Other
Utility

Example Topics extracted from NIH/NSF grants



Important point: these distributions are learned in a completely automated "unsupervised" fashion from the data

Topics are like clusters of documents; however, they are distributed across the documents

Our goal in topic modelling

- The goal of topic modeling is to automatically discover the topics from a collection of documents
- Documents are observed
  - Topics, per-document, per-word topic assignments – hidden
  - Hence latent!
- **The central computation problem for topic modelling is to use the observed documents to infer hidden topic structure**
- Think it as reversing the generative process
  - What is the hidden structure that likely generated the observed collection?

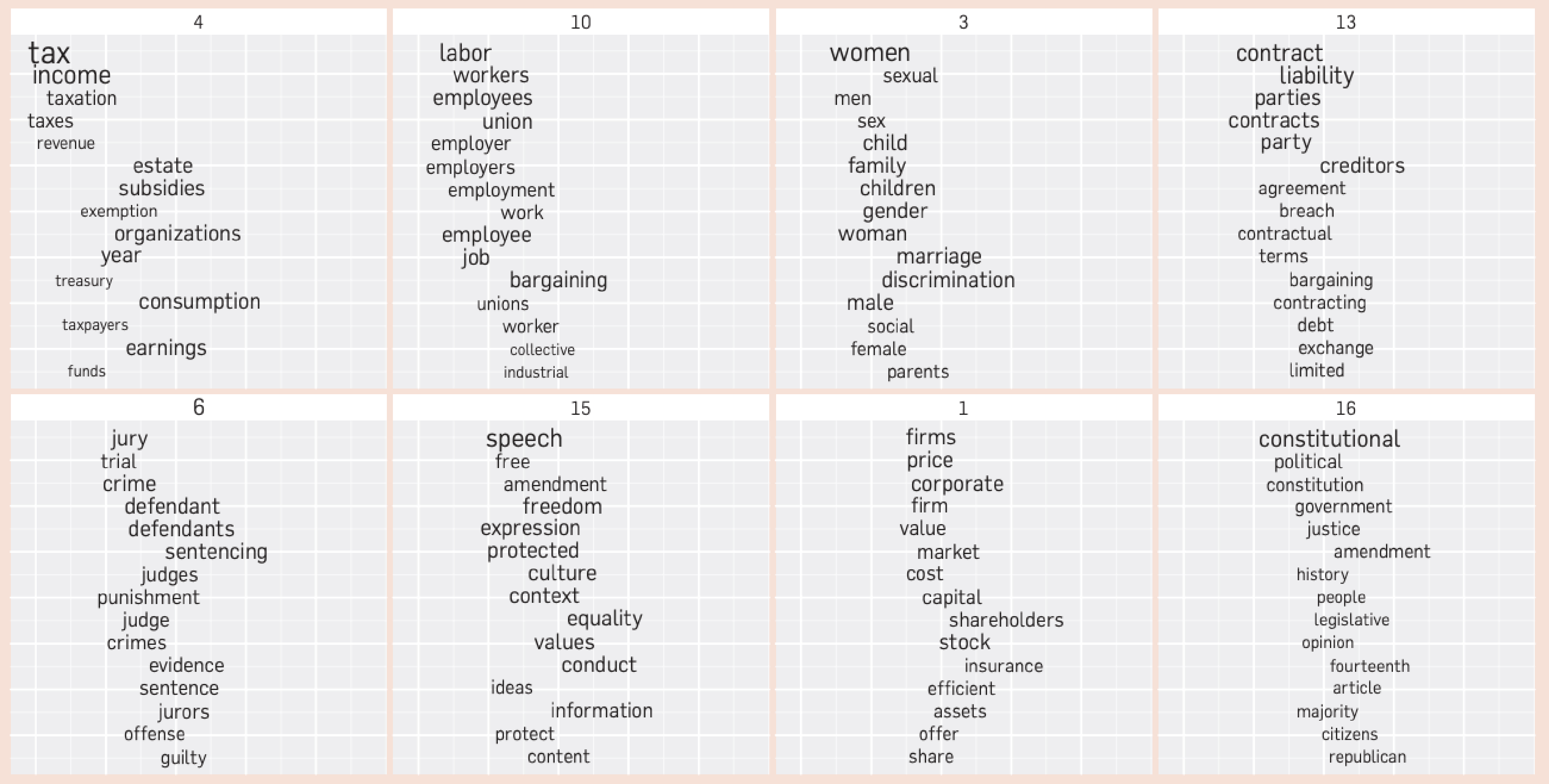
Discuss the central computation problem in topic modelling

Utility of topic models

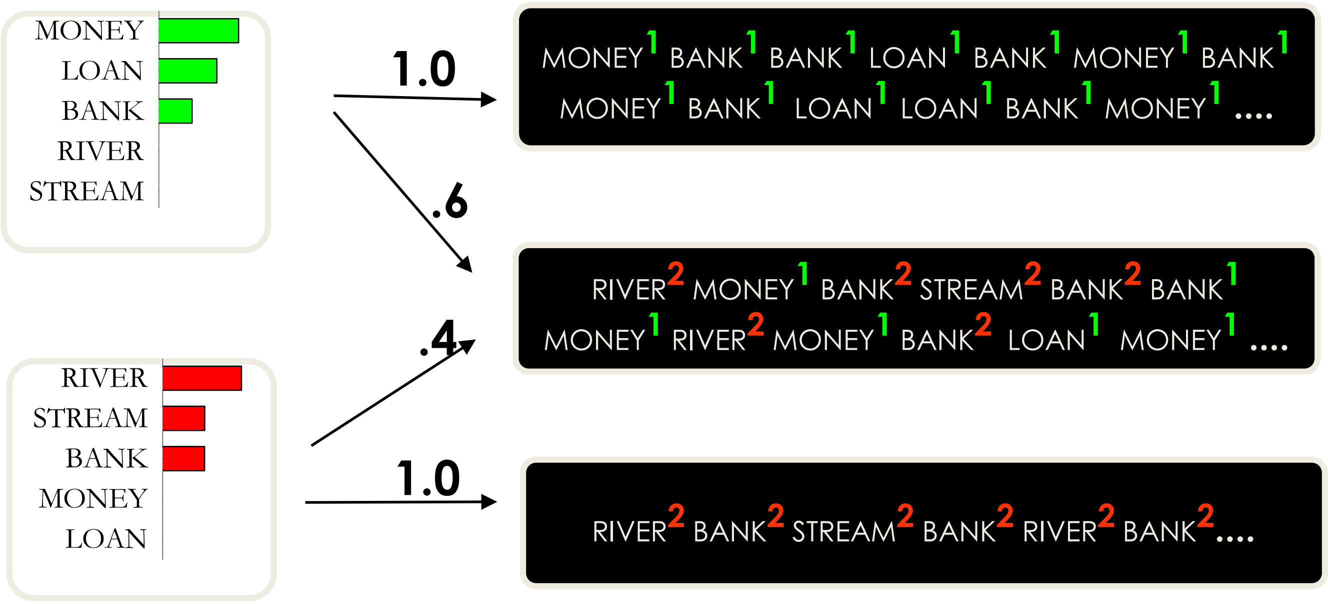
- The utility of topic models stem from the fact that the inferred hidden structure resembles the thematic structure of the collection
- inferred hidden structure
  - Annotates each document in the collection
  - Which can be used for information retrieval, classification etc.
- Topic models provide an algorithmic solution to manage, organize and annotate the large archive texts

Discuss the utility of topic models in exploring a textual collection

Yale Law Journal



Example of generating words



MONEY^{1} BANK^{1} BANK^{1} LOAN^{1} BANK^{1} MONEY^{1} BANK^{1}
MONEY^{1} BANK^{1} LOAN^{1} LOAN^{1} BANK^{1} MONEY^{1} ....

RIVER^{2} MONEY^{1} BANK^{2} STREAM^{2} BANK^{2} BANK^{1}
MONEY^{1} RIVER^{2} MONEY^{1} BANK^{2} LOAN^{1} MONEY^{1} ....

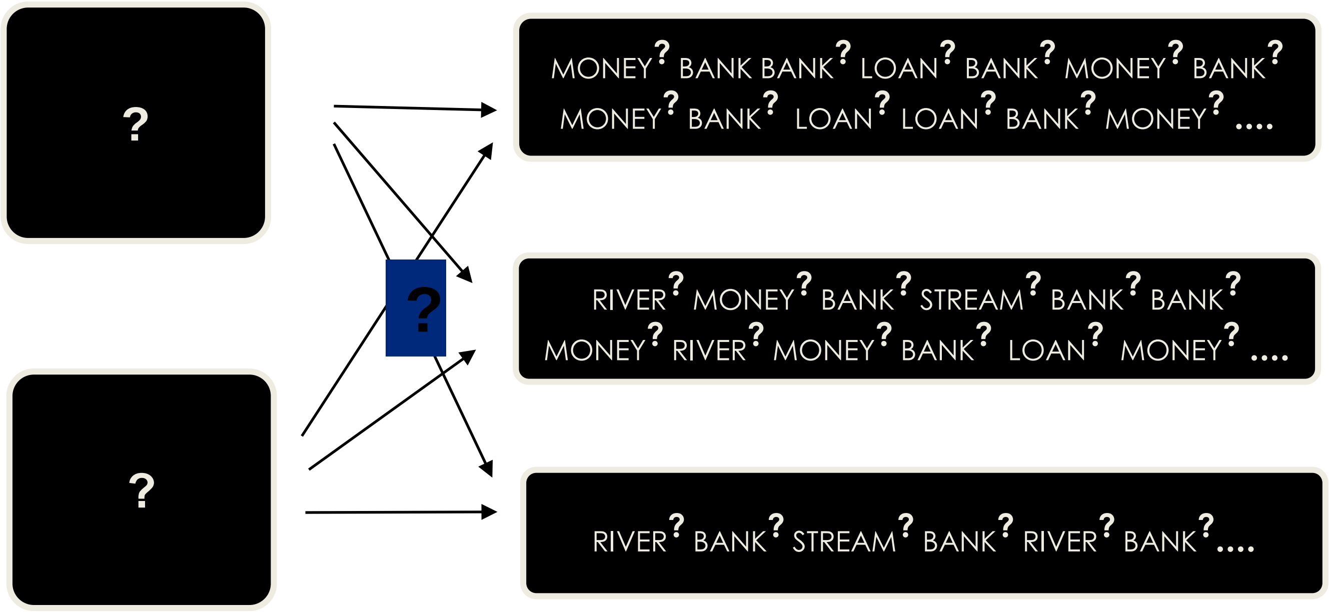
RIVER^{2} BANK^{2} STREAM^{2} BANK^{2} RIVER^{2} BANK^{2}....

Documents and topic assignments

Mixtures θ

Topic
s
ϕ

Inference



?

?

?

MONEY? BANK BANK? LOAN? BANK? MONEY? BANK?
MONEY? BANK? LOAN? LOAN? BANK? MONEY? ....

RIVER? MONEY? BANK? STREAM? BANK? BANK?
MONEY? RIVER? MONEY? BANK? LOAN? MONEY? ....

RIVER? BANK? STREAM? BANK? RIVER? BANK?....

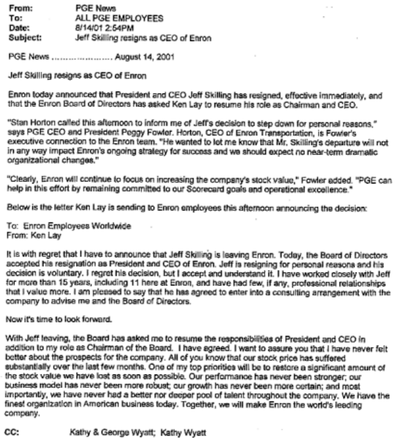
Documents and topic assignments

Mixtures
θ

Topic
s
ϕ

Extracting Topics from Email Conversation

20 News Groups



From: PGE News
To: ALL PGE EMPLOYEES
Date: 8/14/01 2:54PM
Subject: Jeff Skilling resigns as CEO of Enron

PGE News .... August 14, 2001

Jeff Skilling resigns as CEO of Enron

Enron today announced that President and CEO Jeff Skilling has resigned, effective immediately, and that the Enron Board of Directors has asked Ken Lay to resume his role as Chairman and CEO.

"Stan Horton called this afternoon to inform me of Jeff's decision to step down for personal reasons," says PGE CEO and President Peggy Fowler. Horton, CEO of Enron Transportation, is Fowler's executive connection to the Enron team. "He wanted to let me know that Mr. Skilling's departure will not in any way impact Enron's ongoing strategy for success and we should expect no near-term dramatic organizational changes."

"Clearly, Enron will continue to focus on increasing the company's stock value," Fowler added. "PGE can help in this effort by remaining committed to our Scorecard goals and operational excellence."

Below is the letter Ken Lay is sending to Enron employees this afternoon announcing the decision:

To: Enron Employees Worldwide
From: Ken Lay

It is with regret that I have to announce that Jeff Skilling is leaving Enron. Today, the Board of Directors accepted his resignation as President and CEO of Enron. Jeff is resigning for personal reasons and his decision is voluntary. I regret his decision, but I accept and understand it. I have worked closely with Jeff for more than 15 years, including 11 here at Enron, and have had few, if any, professional relationships that I value more. I am pleased to say that he has agreed to enter into a consulting arrangement with the company to advise me and the Board of Directors.

Now it's time to look forward.

With Jeff leaving, the Board has asked me to resume the responsibilities of President and CEO in addition to my role as Chairman of the Board. I have agreed. I want to assure you that I have never felt better about the prospects for the company. All of you know that our stock price has suffered substantially over the last few months. One of my top priorities will be to restore a significant amount of the stock value we have lost as soon as possible. Our performance has never been stronger; our business model has never been more robust; our growth has never been more certain; and most importantly, we have never had a better nor deeper pool of talent throughout the company. We have the finest organization in American business today. Together, we will make Enron the world's leading company.

CC: Kathy & George Wyatt; Kathy Wyatt

20,000 emails

1999-2002

TEXANS
WIN
FOOTBALL
FANTASY
SPORTSLINE
PLAY
TEAM
GAME
SPORTS
GAMES

GOD
LIFE
MAN
PEOPLE
CHRIST
FAITH
LORD
JESUS
SPIRITUAL
VISIT

TRAVEL
ROUNDTRIP
SAVE
DEALS
HOTEL
BOOK
SALE
FARES
TRIP
CITIES

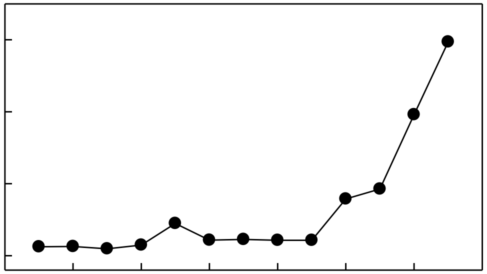
FERC
MARKET
ISO
COMMISSION
ORDER
FILING
COMMENTS
PRICE
CALIFORNIA
FILED

POWER
CALIFORNIA
ELECTRICITY
UTILITIES
PRICES
MARKET
PRICE
UTILITY
CUSTOMERS
ELECTRIC

STATE
PLAN
CALIFORNIA
DAVIS
RATE
BANKRUPTCY
SOCAL
POWER
BONDS
MOU

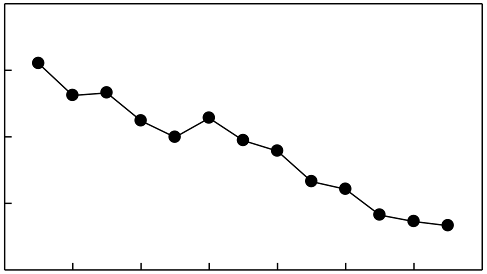
Topic trends in NIPS conference

... NN’s become more popular.



LAYER
NET
NEURAL
LAYERS
NETS
ARCHITECTURE
NUMBER
FEEDFORWARD
SINGLE

SVM on the decline ...



KERNEL
SUPPORT
VECTOR
MARGIN
SVM
KERNELS
SPACE
DATA
MACHINES

What is Heterogeneous Topic Modelling?

- Discover the abstract "**topics**" that occur in a heterogeneous collection of documents.
  - Twitter
  - News
  - Blogs
- Mining common topics from disparate sources
  - unbiased and comprehensive topics

Challenges

- Lexical gap
- Time gap
- Inconsistent signals

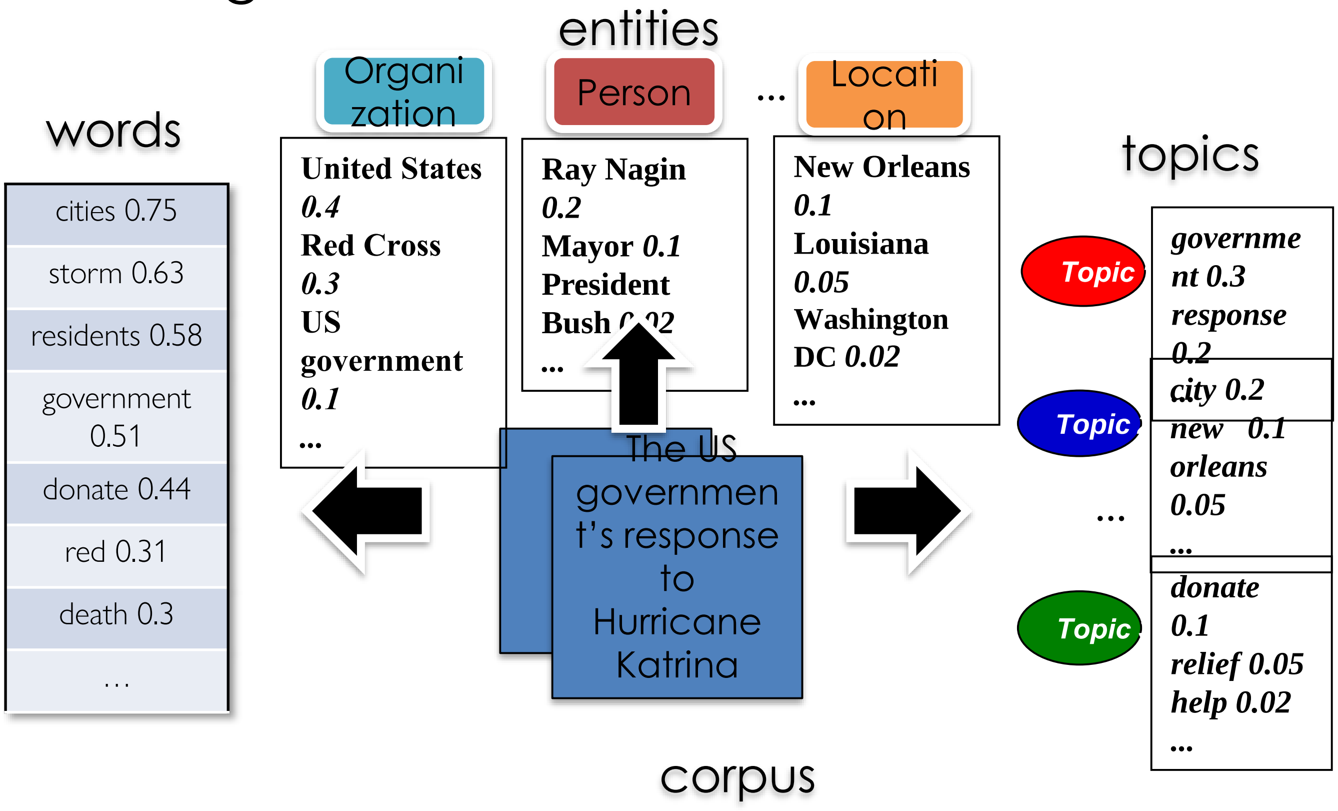
What are the challenges in discovering topics from heterogeneous streams of data?

Semantic Graph in Topic Modelling (SGMM)

- Many sources are useful
  - E.g., blogs; news; twitter;
  - How can we combine them? (semantic graph analysis in text mining for linking multiple text streams)
- Not all entities are equally important
  - E.g. person’s name v.s. locations
  - How can we know which entities are more useful? (entities weighting in semantic graph)
- General methodology to model context in text
  - A unified framework for mining topics from multiple streams (Similar timestamps for similar semantic graphs)
- Many applications (search engine, information browsing)

Motivation

- Making sense of documents collection



entities

Organization

Person

…

Location

words

cities 0.75

storm 0.63

residents 0.58

government 0.51

donate 0.44

red 0.31

death 0.3

…

United States *0.4*
Red Cross *0.3*
US government *0.1*
…

Ray Nagin *0.2*
Mayor *0.1*
President Bush *0.02*
…

New Orleans *0.1*
Louisiana *0.05*
Washington DC *0.02*
…

topics

Topic

government 0.3
response 0.2
…

Topic

city 0.2
new 0.1
orleans 0.05
…

…

Topic

donate 0.1
relief 0.05
help 0.02
…

The US government's response to Hurricane Katrina

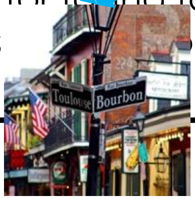
corpus

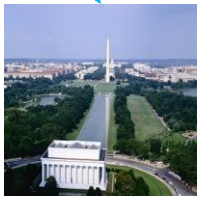
Example: Linking Entities to Knowledge Base

Doc 1: The criticism consisted primarily of condemnations of mismanagement in response to Hurricane Katrina. Specifically, there was a delayed response to the flooding of New Orleans, Louisiana. New Or Mayor Ray Nagin was als ized for f ing to im nt his n plc

Doc 2: Bush was criticized for not returning to Washington, D.C. from his vacation in Texas until after Wednesday afternoon. On the morning of August 28, the president telephoned Mayor Nagin to "plead" for a mandatory evacuation of New Orleans, and Nagin and Gov. Blanco decided to ev the city in response to eq

WIKIPEDIA
The Free Encyclopedia





"Entities" are what a large part of our knowledge is about

What Is Entity Recognition and Typing (ER)

- **Identify token spans of entity mentions in text, and classify them into predefined set of types of interest**

  [Barack Obama] arrived this afternoon in [Washington, D.C.]. [President Obama]'s wife [Michelle] accompanied him

  [TNF alpha] is produced chiefly by activated [macrophages]

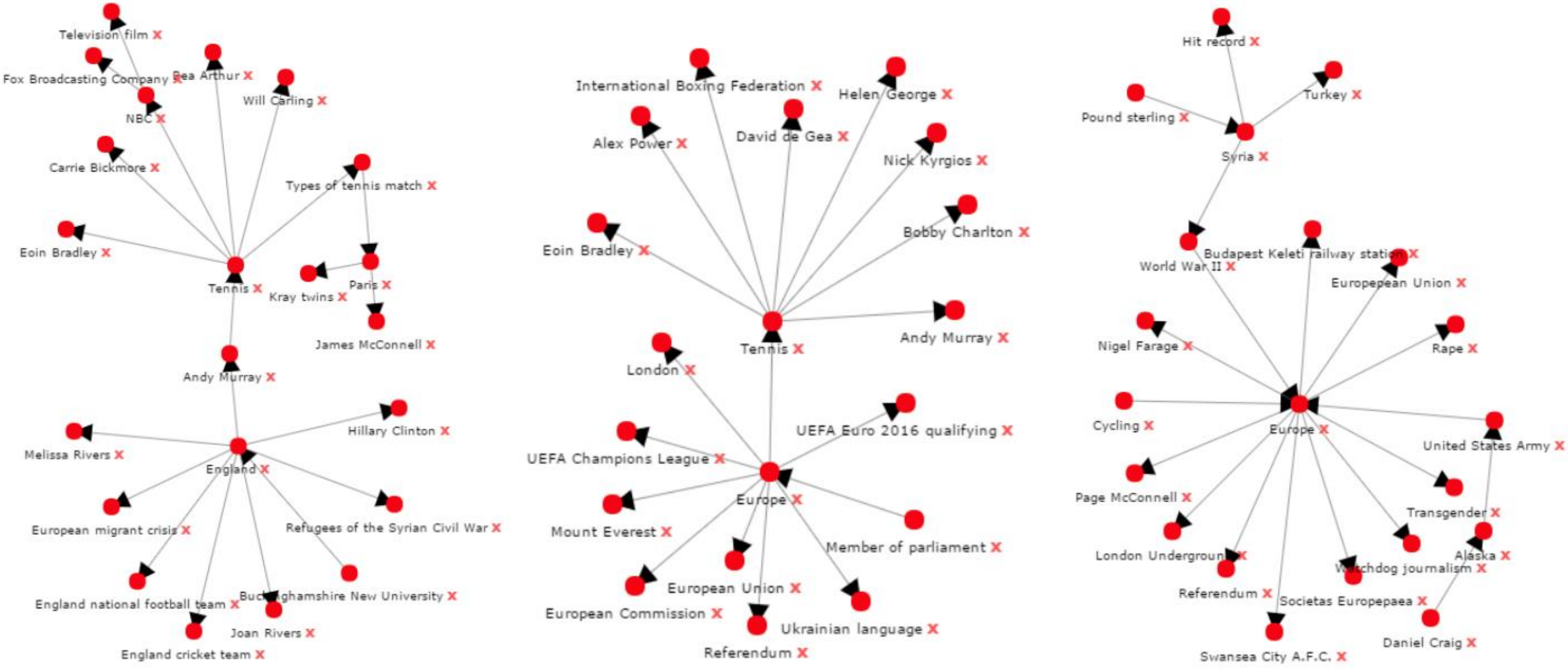
PERSON
LOCATION

PROTEIN
CELL

Semantic Graph Construction

- Apply Named Entity recognition tool DBpedia Spotlight.
  - 1. Remove the isolated entities
  - 2. Remove the infrequent entities (document frequency)
- Search a sub-graph of DBpedia with the entities already identified
  - put intermediate entities found along the paths into the graph.

Semantic Graphs



(a) NEWSIR on 01/09/2015

(b) NEWSIR on 02/09/2015

(c) NEWSIR on 03/09/2015

Local Semantic Graph

- A semantic graph is built for each timestamp
  - Day 1, 2, 3, …
- Alleviate Asynchronous communication.

How does local semantic graph address the asynchronism between channels of data?

Global Semantic Graph

- A semantic graph is built over the entire corpus
- Bridge Lexical gap

Discuss the role of global semantic graph in topic modelling

Semantic Graph in Topic Modelling (SGMM)

- Biased propagation
  - Textual information
  - Semantic information
- Focus on entities
  - Topic distribution of an entity is computed
    - By average topic distribution of connected documents
    - Connected entities of the semantic graph
- then topic distribution of a document is then biased propagation of topic distribution its content and those of the entity based topic distribution

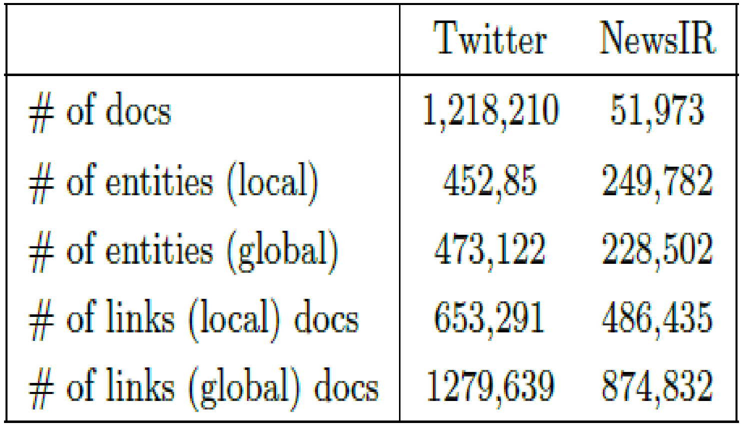
Baselines

- SMM: simple mixture model
  - The baseline approach that simply merges multiple streams and then apply topic model
- CCMM: cross collection mixture model
  - The state-of-the-art approach that distinguish common topics from local topics and structure asynchronous streams with a background language model
  - Drawbacks:
    - It assumes a shared time distribution
    - Word-level analysis
- SGMM: The semantic graph based mixture model

Long Chen, Joemon M. Jose, Haitao Yu, Fajie Yuan:
A Semantic Graph-Based Approach for Mining Common Topics from
Multiple Asynchronous Text Streams. WWW 2017: 1201-1209

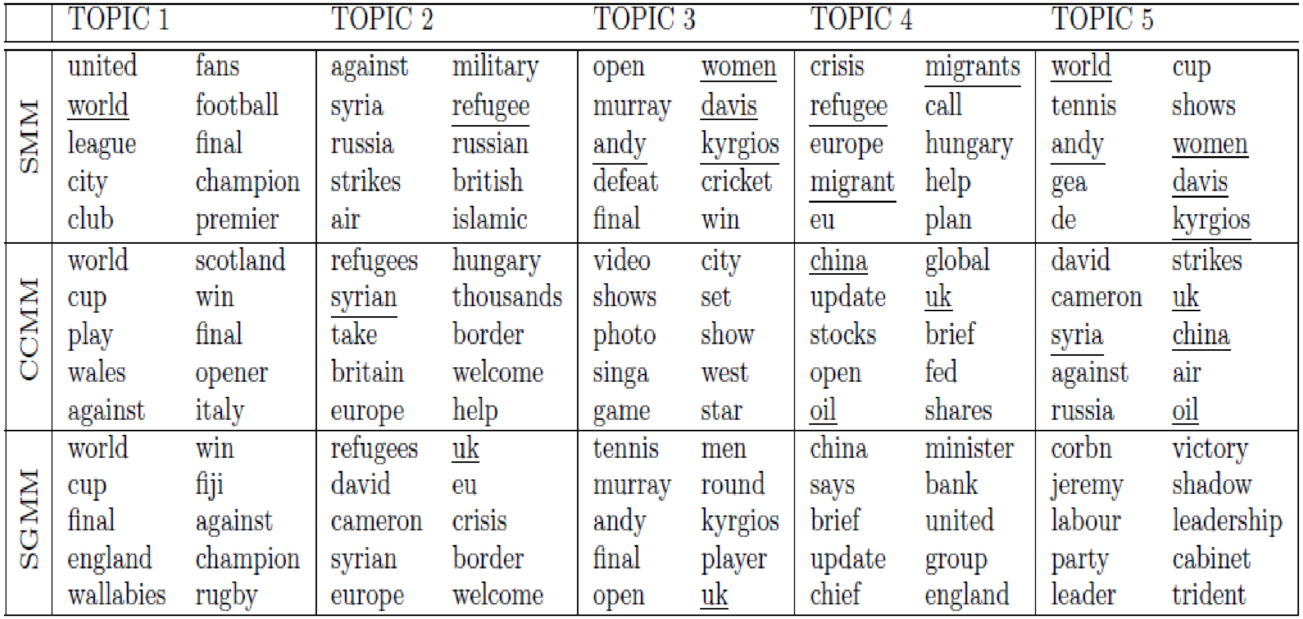
Dataset

- Experiments conducted on two real-world datasets:



	Twitter	NewsIR
# of docs	1,218,210	51,973
# of entities (local)	452,85	249,782
# of entities (global)	473,122	228,502
# of links (local) docs	653,291	486,435
# of links (global) docs	1279,639	874,832

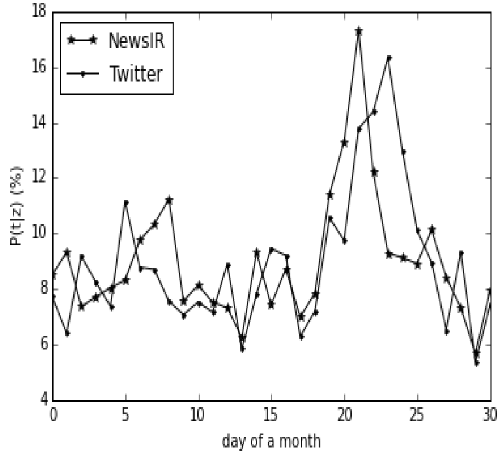
Experimental Results



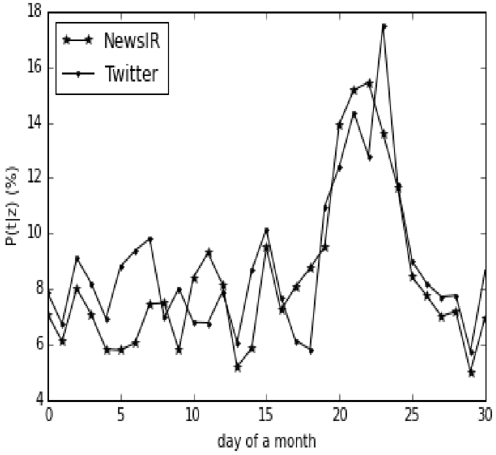
	TOPIC 1		TOPIC 2		TOPIC 3		TOPIC 4		TOPIC 5	
SMM	united	fans	against	military	open	women	crisis	migrants	world	cup
	world	football	syria	refugee	murray	davis	refugee	call	tennis	shows
	league	final	russia	russian	andy	kyrgios	europe	hungary	andy	women
	city	champion	strikes	british	defeat	cricket	migrant	help	gea	davis
	club	premier	air	islamic	final	win	eu	plan	de	kyrgios
CCMM	world	scotland	refugees	hungary	video	city	china	global	david	strikes
	cup	win	syrian	thousands	shows	set	update	uk	cameron	uk
	play	final	take	border	photo	show	stocks	brief	syria	china
	wales	opener	britain	welcome	singa	west	open	fed	against	air
	against	italy	europe	help	game	star	oil	shares	russia	oil
SGMM	world	win	refugees	uk	tennis	men	china	minister	corbn	victory
	cup	fiji	david	eu	murray	round	says	bank	jeremy	shadow
	final	against	cameron	crisis	andy	kyrgios	brief	united	labour	leadership
	england	champion	syrian	border	final	player	update	group	party	cabinet
	wallabies	rugby	europe	welcome	open	uk	chief	england	leader	trident

Experimental Results

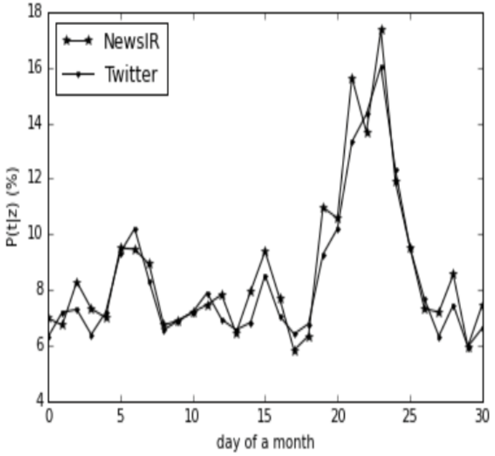
SMM



CCM



SGMM



Summary

- Discussed
  - Topic modelling/LDA
  - a novel semantic graph based topic model (SGMM)
- It supersedes the existing ones since:
  1. homogeneous networks (i.e., entity to entity relations)
  2. heterogeneous networks (i.e., entity to document relations)
  3. both local and global representation of documents

Software

- Entity Recognition
  - Tagme: https://github.com/shangjingbo1226/SegPhrase
  - Dbpedia Spotlight: https://github.com/dbpedia-spotlight/dbpedia-spotlight
- Dbpedia Dataset: http://oldwiki.dbpedia.org/Downloads2014/
- NewsIR: https://webscope.sandbox.yahoo.com/
- SGMM: https://github.com/long4glasgow/Semantic-Mixture-Model

Summary

- Data streaming systems
  - Twitter and social aspects
  - Technology
  - Event detection
- Making Sense
  - Crowd sourcing
  - Event detection evaluation
- Core Science
  - Emotion
  - Knowledge graph
  - Topic modelling
- Exploitation
  - Digital Marking

Project Presentation

- **Tuesday 26th November 2019**

- Group Presentation:
  - 5-Minute Presentation (Strict)
    - Design Architecture for Twitter Crawling
    - Basic Analytics of Crawled Tweets
    - Advanced Analytics of Crawled Tweets
      - Solution Design
      - Results
      - Discussion and Findings
  - 3-Minute Q&A