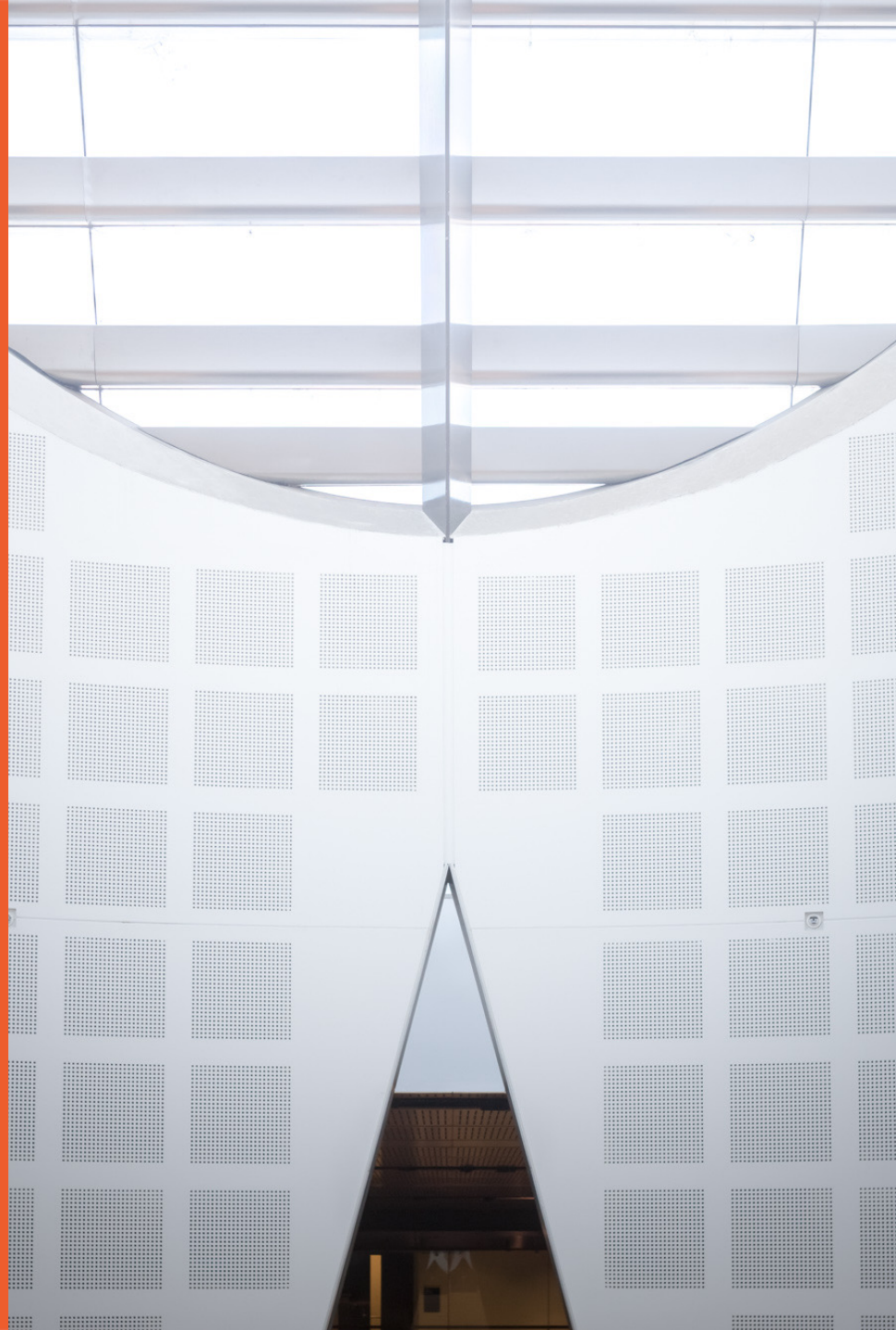
Social Media Analytics for Healthcare

Reference: Healthcare Data Analytics, Chapter 9



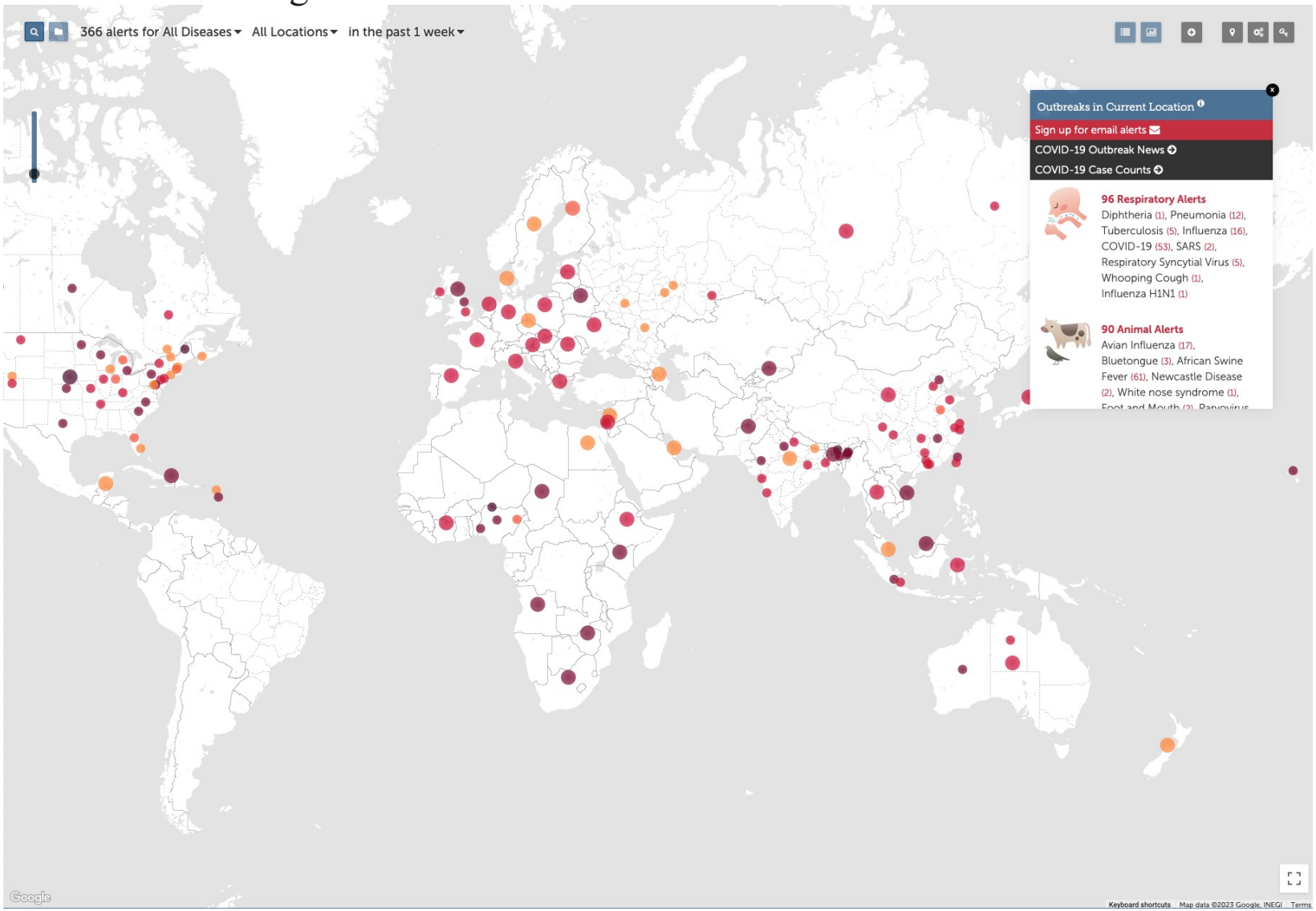
THE UNIVERSITY OF SYDNEY

Introduction

The emergence of social media resources in the form of social networking sites, blogs/microblogs, forums, question answering services, online communities and encyclopedias, designated a move **from passive consumption to active creation of diverse types of content by Internet users**.

Social media goes beyond stating facts and describing events and provides **a wealth of information about public opinion on virtually any topic, including healthcare.**

HealthMap is a freely accessible, automated electronic information system for monitoring, organizing, and visualizing reports of global disease outbreaks according to geography, time, and infectious disease agent.



https://www.healthmap.org/

Introduction

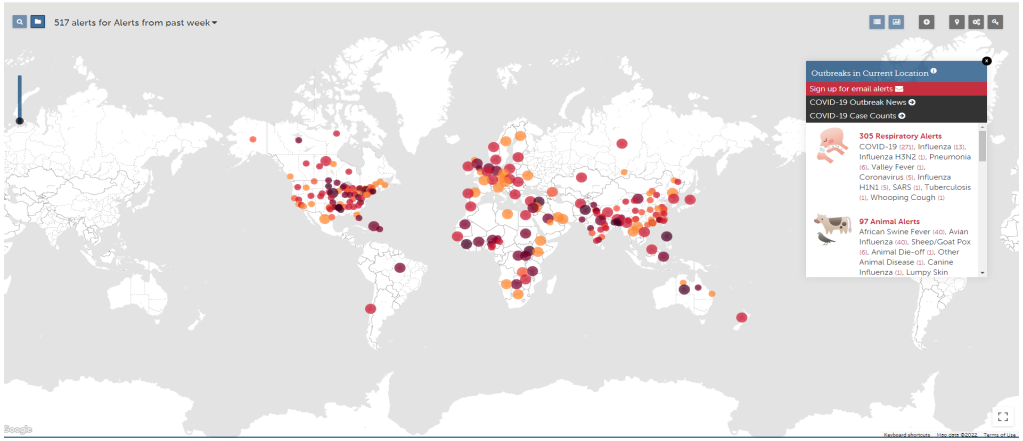


FIGURE: User interface of HealthMap

Recent studies report that 61% of American adults seek health information online and 37% have accessed or posted health information online. In addition to that, 72% of online adults in the United States are using social media. Of adult social media users, 23% follow their friends' personal health experiences or updates, 17% use social media to remember and memorialize people with a specific health condition and 15% obtain health information from social media sites.

Introduction

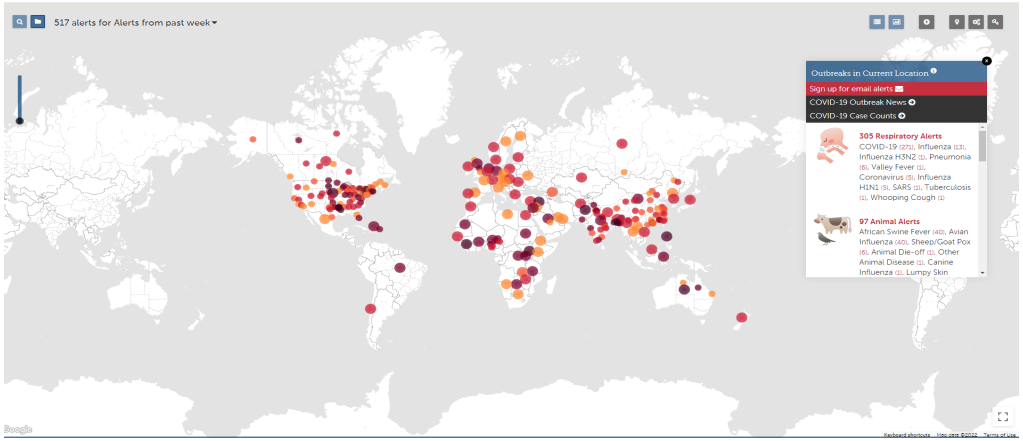


FIGURE: User interface of HealthMap

Web 2.0 services and platforms have been designed to encourage frequent **expression of people's thoughts and opinions** on a variety of issues as well as random details of their lives.

They have also made measurable what was previously unmeasurable and shed additional light on important questions in public health that have been either too expensive or outright impossible to answer, **such as distribution of health information in a population, tracking health information trends over time and identifying gaps between health information supply and demand.**

Introduction

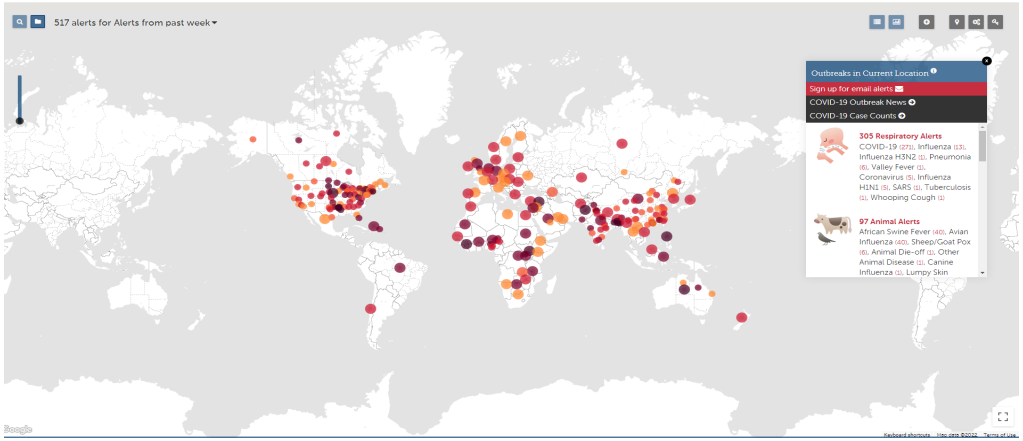


FIGURE: User interface of HealthMap

The fine granularity and pervasiveness of social media data models phenomena that were previously out of reach, including the **probability of a given individual to get sick with a disease.**

Although most individual social media posts and messages contain little informational value, **aggregation of millions of such messages can generate important knowledge.**

For example, knowing that a certain individual has contracted a flu based on his or her messages on social networking sites may not be an interesting fact by itself, but **millions of such messages can be used to track influenza rate in a state or a country.**

Introduction

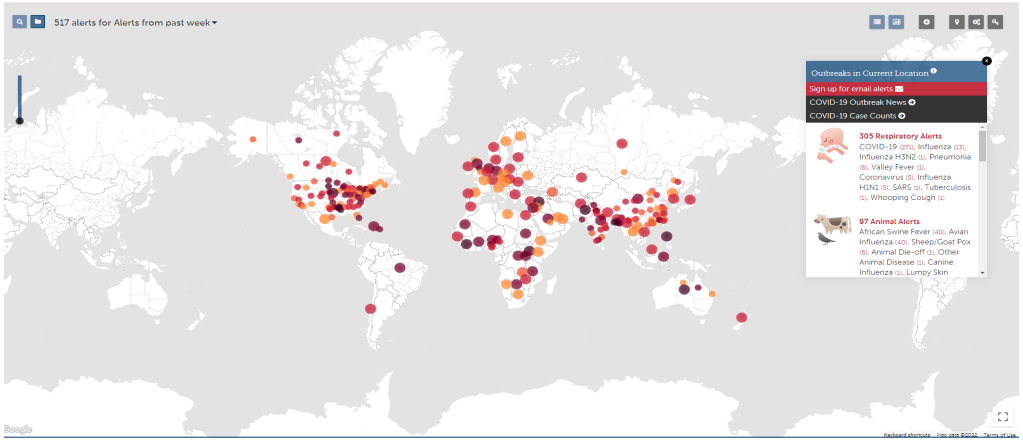


FIGURE: User interface of HealthMap

Social media data can be mined for patterns and knowledge that can be leveraged in descriptive as well as predictive models of population health. It can also improve the **overall effectiveness of public health monitoring and analysis and significantly reduce its latency.**

Introduction

Previous research work on social media analytics for healthcare has focused on the following three broad areas:

1. Methods for capturing aggregate health trends from social media data, such as outbreaks of infectious diseases, and analyzing the mechanisms underlying the spread of infectious diseases;

2. Methods for fine-grained analysis and processing of social media data, such as methods to detect reports of adverse drug interactions and medical events and to model the health status and well-being of individuals;

3. Studying how social media can be effectively used as a communication medium between patients, between patients and doctors and how to effectively leverage social media in interventions and health education campaigns.

Syndromic Surveillance Systems Based on Social Media

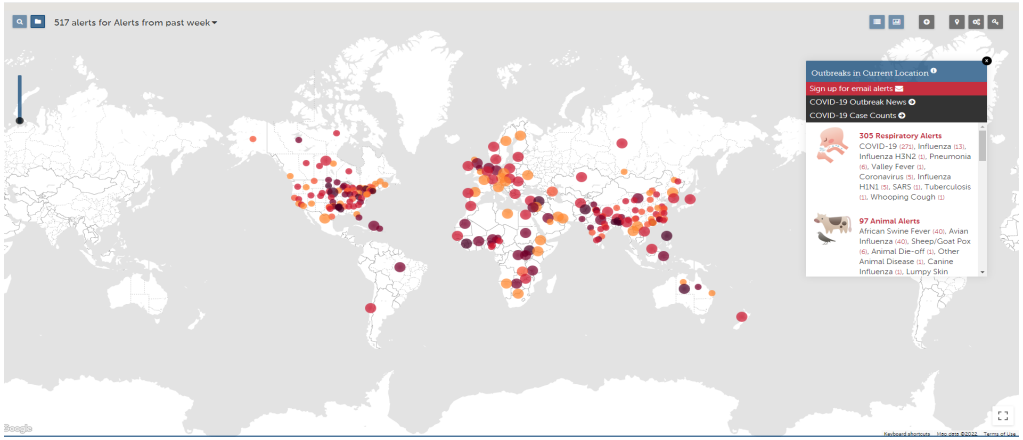


FIGURE: User interface of HealthMap

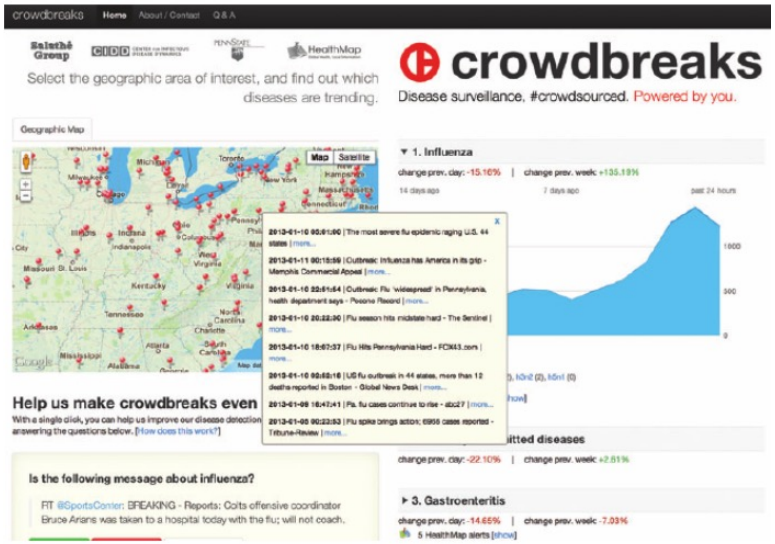
HealthMap is a system that monitors global media sources such as news wires and Web sites to provide a comprehensive view of ongoing disease activity around the world. It combines automated, around-the-clock data collection and processing with expert review and analysis. Visitors to the site could filter reports according to the suspected or confirmed cases of deaths from a disease and select a time interval to show its spread.

Syndromic Surveillance Systems Based on Social Media

User interface of FluNearYou

FluNearYou is an online system that integrates different types of data (weekly surveys completed by volunteers, CDC Flu Activity data and Google Flu Trends ILI data) to visualize the current and retrospective flu activity in the United States and Canada. It is a joint project between HealthMap, the American Public Health Association, Skoll Global Threats Fund and Boston Children's Hospital.

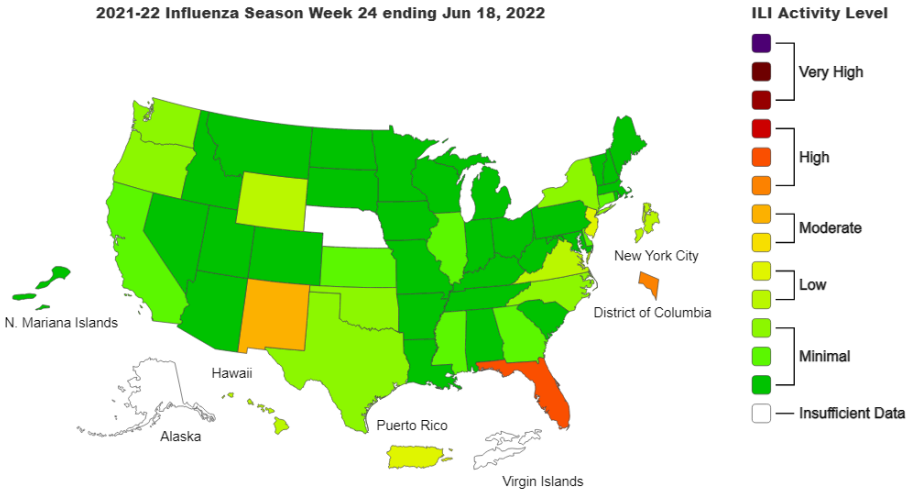
Syndromic Surveillance Systems Based on Social Media



User interface of Crowdbreaks

Crowdbreaks is a surveillance system that automatically collects the disease-related tweets, determines their location and visualizes them on a map (Figure 9.4). It employs a machine learning algorithm to assess whether a given tweet contains a reported case of a disease.

Syndromic Surveillance Systems Based on Social Media



Weekly U.S. Influenza Surveillance Report | CDC

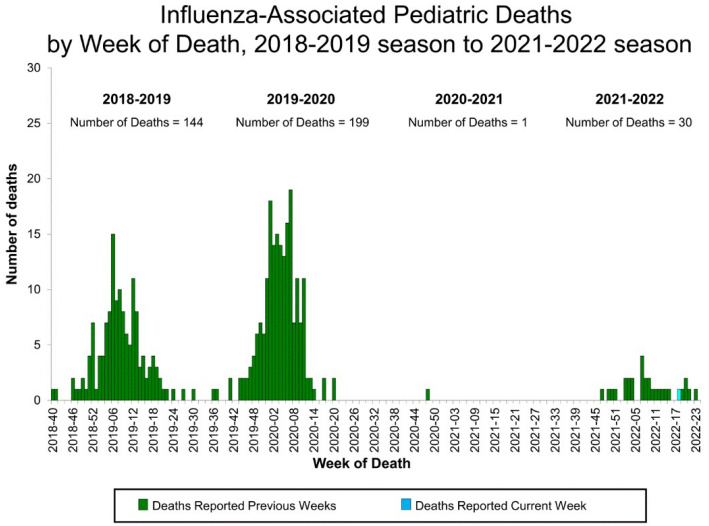
Epidemics of infectious diseases, such as influenza and cholera, are a major public health concern that is difficult to anticipate and model. **Seasonal influenza epidemics result in about three to five million cases of severe illnesses and about 250,000 to 500,000 deaths worldwide each year.**

Health organizations require accurate and timely disease surveillance techniques in order to respond to the emerging epidemics by better planning for surges in patient visits, therapeutic supplies and public health information dissemination campaigns.

Syndromic Surveillance Systems Based on Social Media

Public health monitoring has traditionally relied **on surveys and aggregating primary data from healthcare providers and pharmacists** (e.g., clinical encounters with healthcare professionals, sickleave and drug prescriptions). Syndromic surveillance, the monitoring of clinical syndromes that have significant impact on public health, is particularly required for episodic and widespread infections, such as seasonal influenza.

Many infectious **disease surveillance systems**, including those employed by Centers for Disease Control and Prevention (CDC) in the United States, Public Health Agency of Canada, etc. Although **survey-based surveillance** systems are effective tools in discovering disease outbreaks, they typically incur high operational costs and temporal lags in reporting the outbreaks, since an infectious disease case is recorded only after a patient visits a doctor's office and the information about it is sent to the appropriate public health agency. **During the deadly infectious disease outbreaks**, such as cholera, **this delay can hinder early epidemiological assessment and result in a greater number of fatalities.**



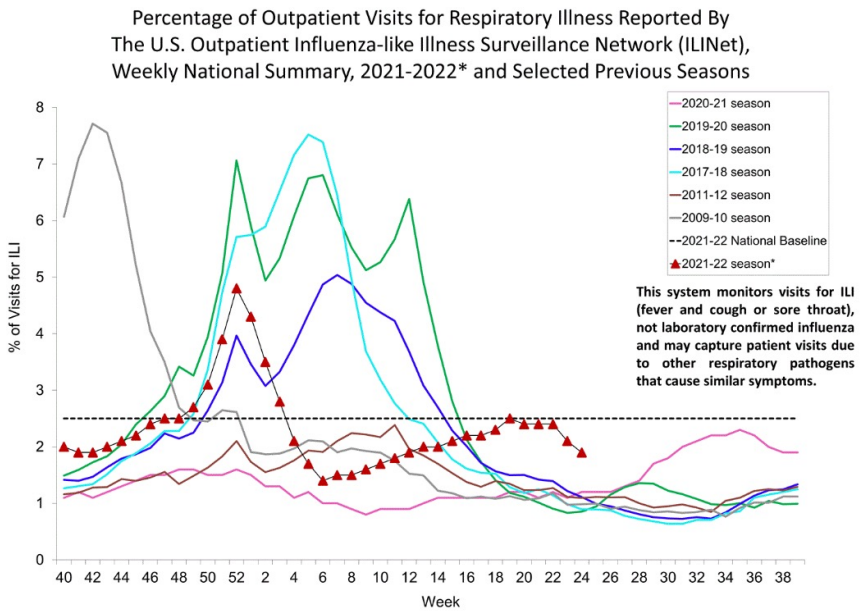
Weekly U.S. Influenza Surveillance Report | CDC

Social Media Analysis for Detection and Tracking of Infectious Disease Outbreaks

By contrast, social media data are available in near real time and therefore can provide much **earlier estimates of the magnitude and dynamics of an epidemic**. Social media platforms, such as Twitter, offer virtually unlimited volumes of publicly available data and population sample sizes that exceed those of paper surveys by several orders of magnitude.

Finding the key symptomatic individuals along with other people, who may have already contracted the disease, can also be done **more effectively and in a timely manner by leveraging online social network data.**

Geographical metadata in the form of the coordinates associated with some of the **social media posts can play an important role in monitoring the impact and the geographical spread of an epidemic.**

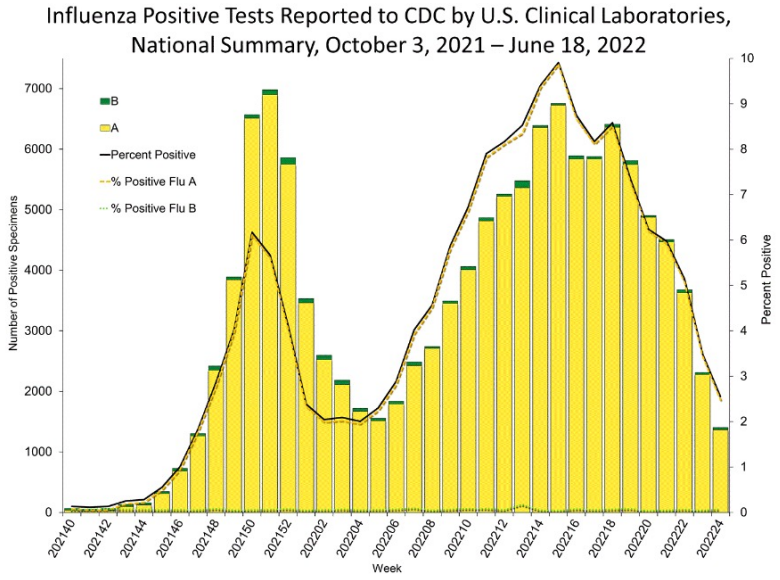


Weekly U.S. Influenza Surveillance Report | CDC

Using Search Query and Website Access Logs

An increasing number of people around the world are using the Internet to seek and disseminate health-related information. People search for health information for a variety of reasons: concerns about themselves, their family or friends.

According to the National Library of Medicine**, an estimated 113 million people in the United States use the Internet to find health-related information** with up to 8 million people searching for health-related information on a typical day.



Weekly U.S. Influenza Surveillance Report | CDC

Using Search Query and Website Access Logs

The general idea behind the proposed methods for monitoring public health based on the analysis of query logs of search engines is that the **interest of a general public in a certain public health topic can be approximated by the search query activity related to this topic**.

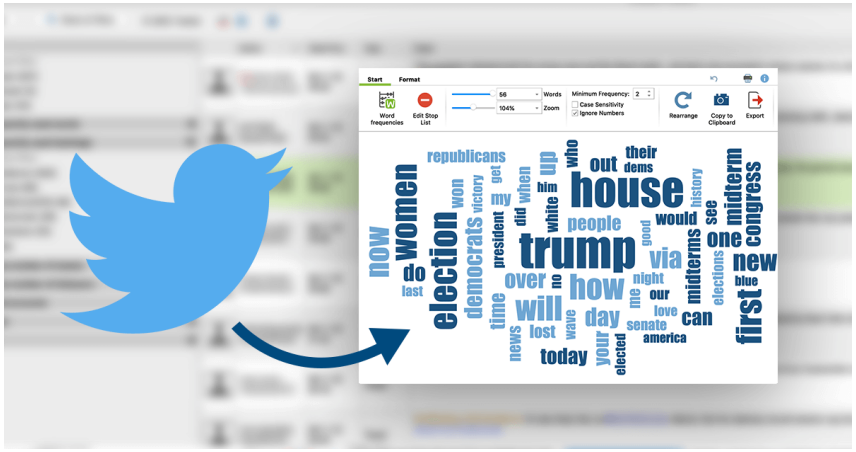
A joint study by CDC and Yahoo! suggested that Internet searches for specific cancers correlate with their estimated incidence, mortality and the volume of related news coverage. **They concluded that media coverage appears to play a powerful role in prompting online searches for cancer information.**

	Week 24	Data Cumulative since October 3, 2021 (Week 40)
No. of specimens tested	6,558	880,606
No. of positive specimens	60	24,197
Positive specimens by type/subtype		
Influenza A	57 (95.0%)	24,068 (99.5%)
(H1N1)pdm09	0	28 (0.1%)
H3N2	35 (100%)	18,806 (99.9%)
H3N2v	0	1 (<0.1%)
Subtyping not performed	22	5,233
Influenza B	3 (5.0%)	129 (0.5%)
Yamagata lineage	0	1 (2.4%)
Victoria lineage	0	40 (97.6%)
Lineage not performed	3	88

Weekly U.S. Influenza Surveillance Report | CDC

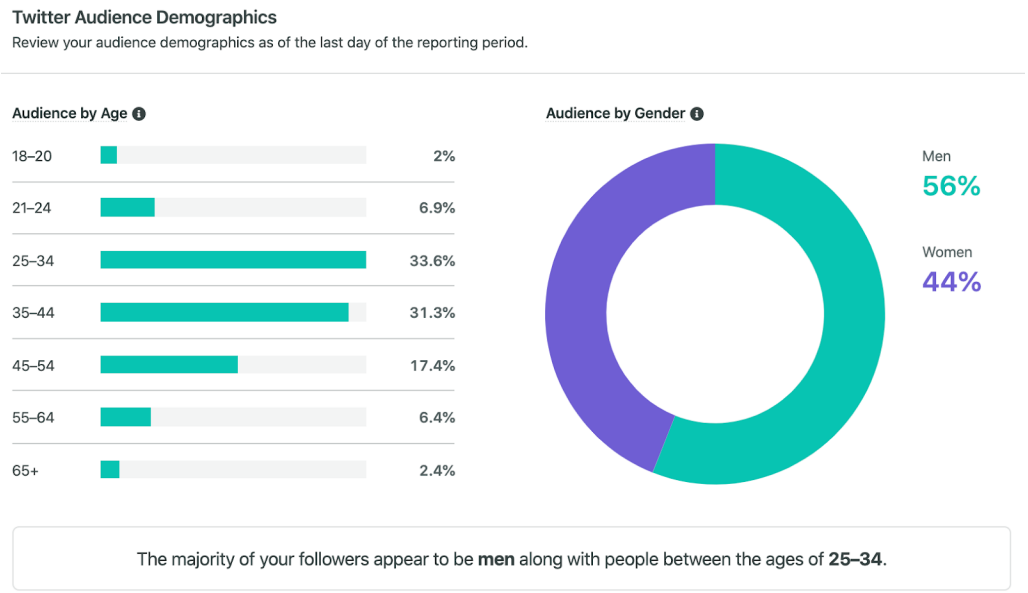
Using Twitter and Blogs

The emergence and rapid increase in popularity of **Twitter opened up a new research direction in Internet-based disease surveillance.** Twitter is a social networking and microblogging platform that enables users to create the posts limited to 140 characters and share them either with the general public or only with a specific group of people designated as “followers.”



The advantages of Twitter-based approaches for disease outbreak detection over the ones that are based on search query and access logs are twofold. First, although Twitter messages are **fairly short**, they are still more descriptive and provide more contextual information than search engine queries. Second, Twitter profiles often contain **rich meta-data associated with the users.** Twitter also has an advantage over other social media services in that **it offers a larger volume of mostly publicly available messages.**

Using Twitter and Blogs



The work of Ritterman et al. was one of the first to use **Twitter for infectious disease surveillance.** In particular, they used the dataset consisting of 48 million tweets collected over a period of two months, which covers the timespan between the first time when the news about H1N1 (or Swine Flu) virus first broke out and until the H1N1 pandemic was declared by theWorld Health Organization onMay 11, 2009.

In a recent work, Li and Cardie focused on the early detection of the flu pandemic and introduced a Bayesian approach based on the **spatio-temporal Markov Network** (which they call Flu Markov Network), which takes into account both the spatial information and the daily fluctuations in the number of posted tweets, for early stage unsupervised detection of flu.

What Is Text Mining?

"The objective of Text Mining is to exploit information contained in textual documents in various ways, including …discovery of patterns and trends in data, associations among entities, predictive rules, etc." (Grobelnik et al., 2001)

"Another way to view text data mining is as a process of exploratory data analysis that leads to heretofore unknown information, or to answers for questions for which the answer is not currently known." (Hearst, 1999)

Two Different Views of Text Mining

- Data Mining View: Explore patterns in textual data
  - Find latent topics
  - Find topical trends
  - Find outliers and other hidden patterns

Shallow mining

- Natural Language Processing View: Make inferences based on partial understanding natural language text
  - Information extraction
  - Question answering

Deep mining

Applications of Text Mining

- Direct applications: Go beyond search to find knowledge
  - Question-driven (Bioinformatics, Business Intelligence, etc): We have specific questions; how can we exploit data mining to answer the questions?
  - Data-driven (WWW, literature, email, customer reviews, etc): We have a lot of data; what can we do with it?
- Indirect applications
  - Assist information access (e.g., discover latent topics to better summarize search results)
  - Assist information organization (e.g., discover hidden structures)

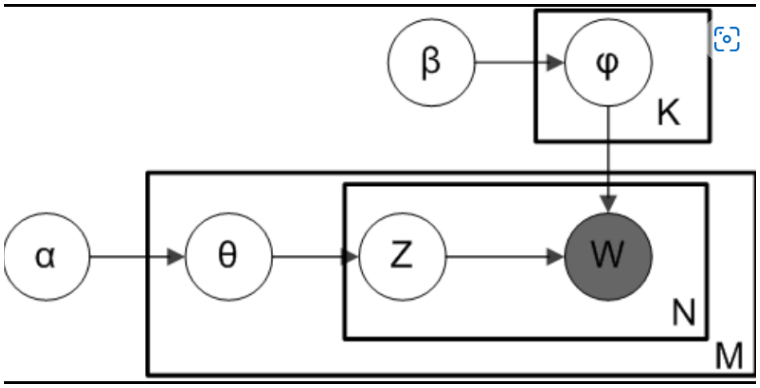
Text Mining Methods

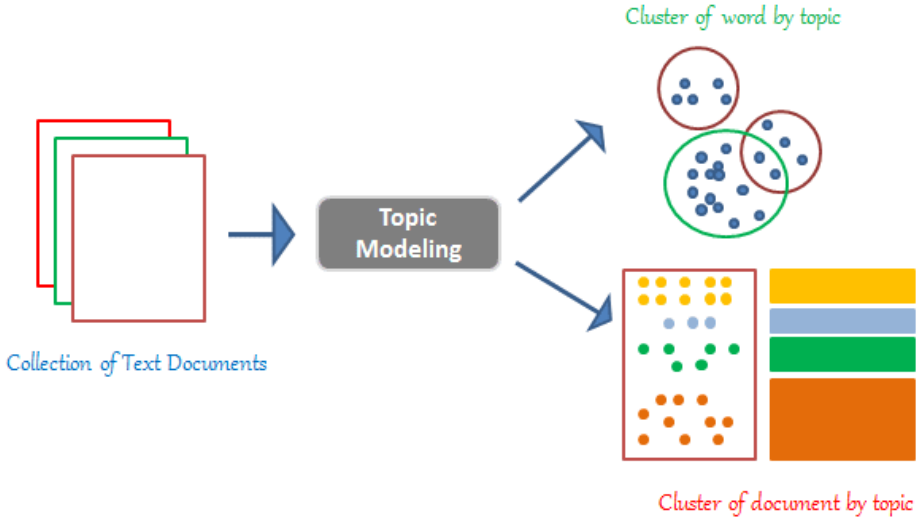
- Data Mining Style: View text as high dimensional data
  - Frequent pattern finding
  - Association analysis
  - Outlier detection
- Information Retrieval Style: Fine granularity topical analysis
  - Topic extraction
  - Exploit term weighting and text similarity measures
- Natural Language Processing Style: Information Extraction
  - Entity extraction
  - Relation extraction
  - Sentiment analysis
  - Question answering
- Machine Learning Style: Unsupervised or semi-supervised learning
  - Mixture models
  - Dimension reduction

Topic Models for Analyzing Health-Related Content

Methods capable of aggregating healthcare-related content created by millions of social media users can provide extensive near real-time information about population health and different population characteristics, which is invaluable to public health researchers.

Topic models, such as Latent Dirichlet Allocation (LDA), are probabilistic latent variable generative models, which associate hidden variables controlling topical assignments with terms in document collections. They were designed to summarize information in large textual corpora by revealing their latent thematic structure in the form of clusters of semantically related words.





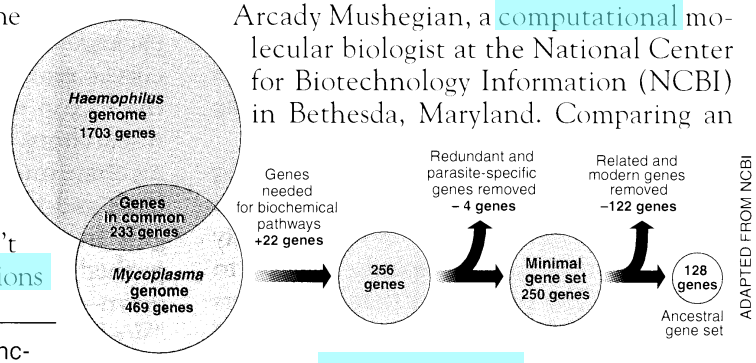
Basic Topic Model: LDA

* Slides borrowed from David Blei

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.

Documents exhibit multiple topics.

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

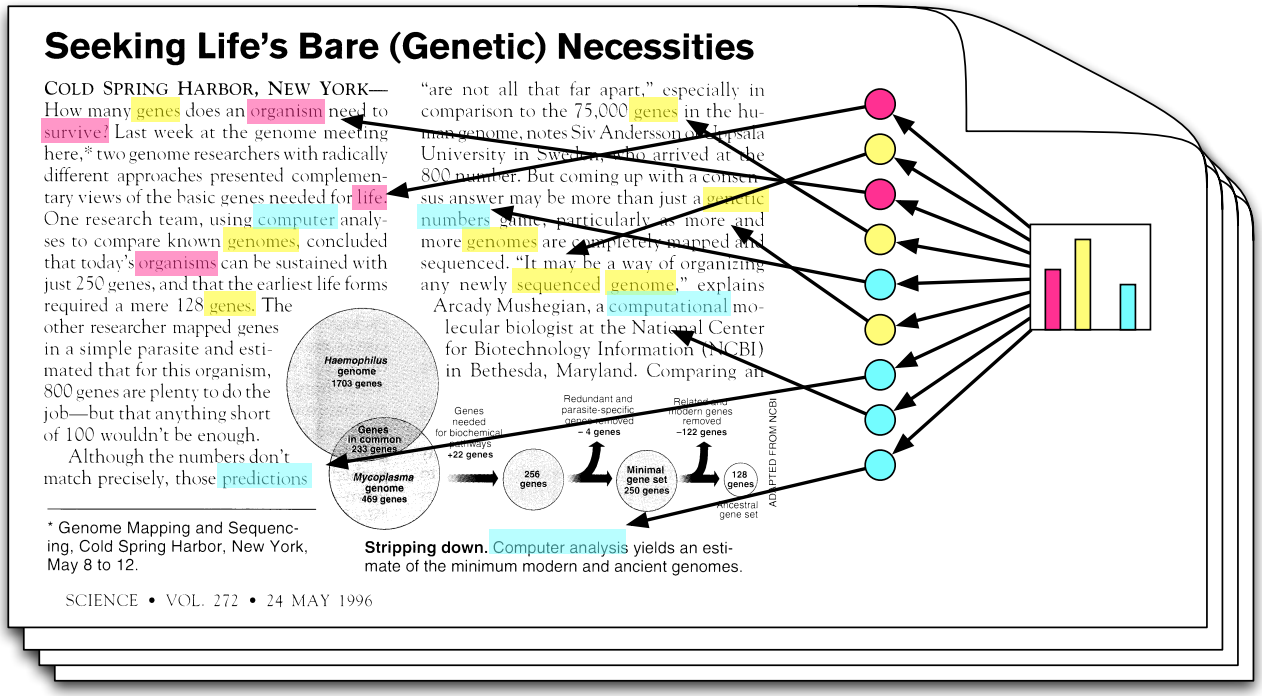
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



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.

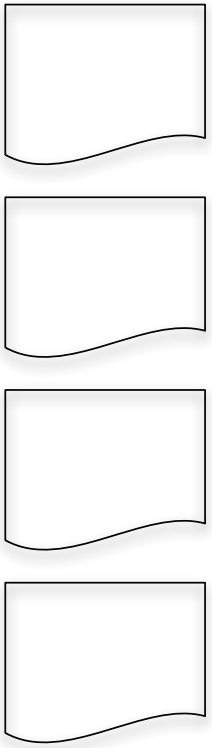
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

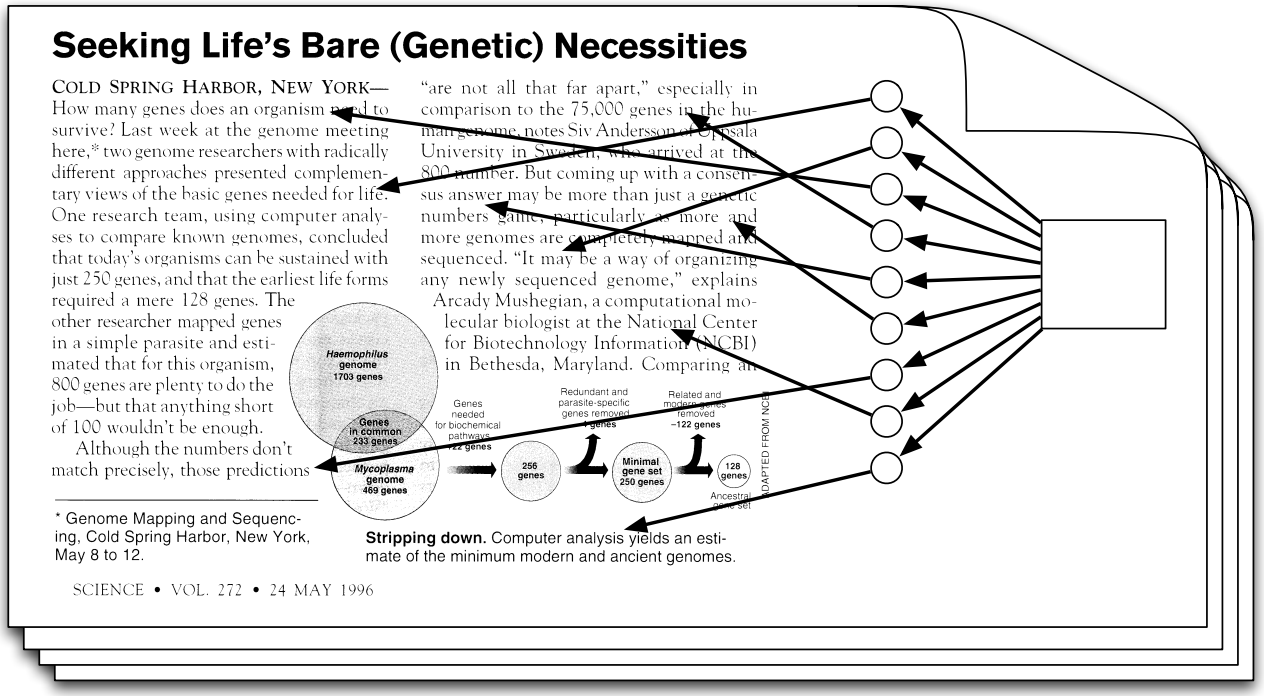
Latent Dirichlet Allocation

Topics

Documents

Topic proportions and assignments





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.

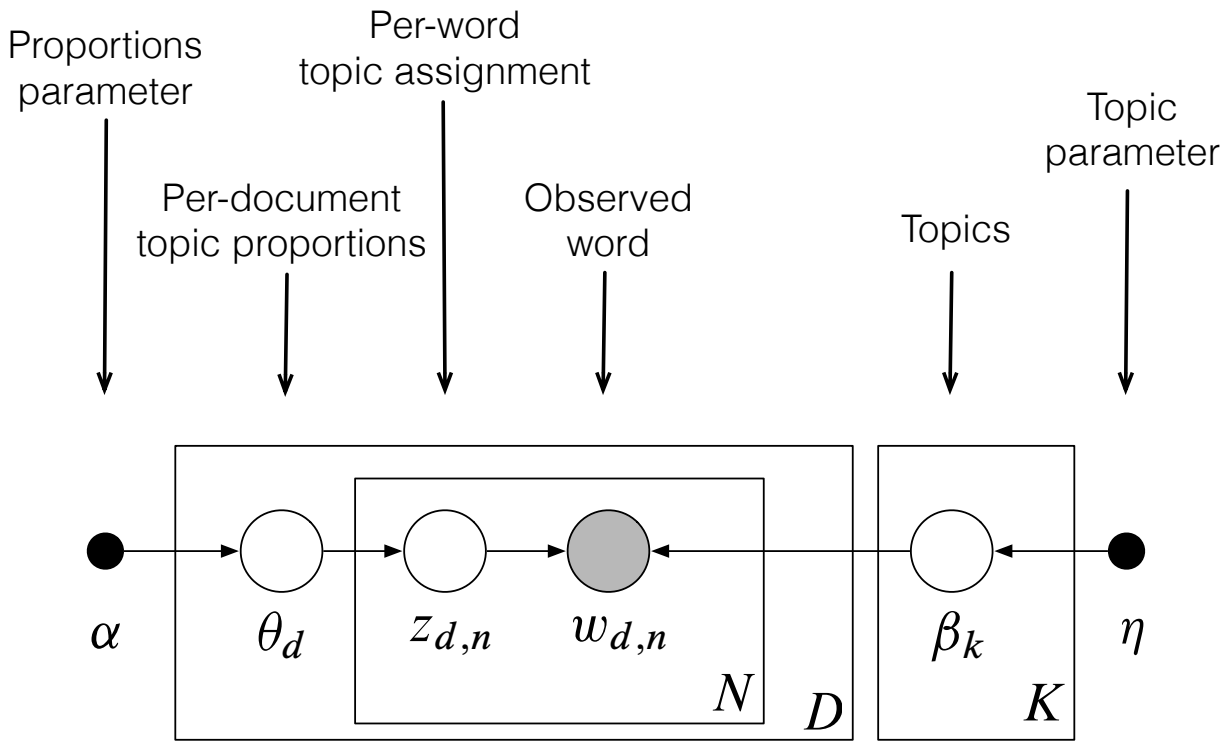
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

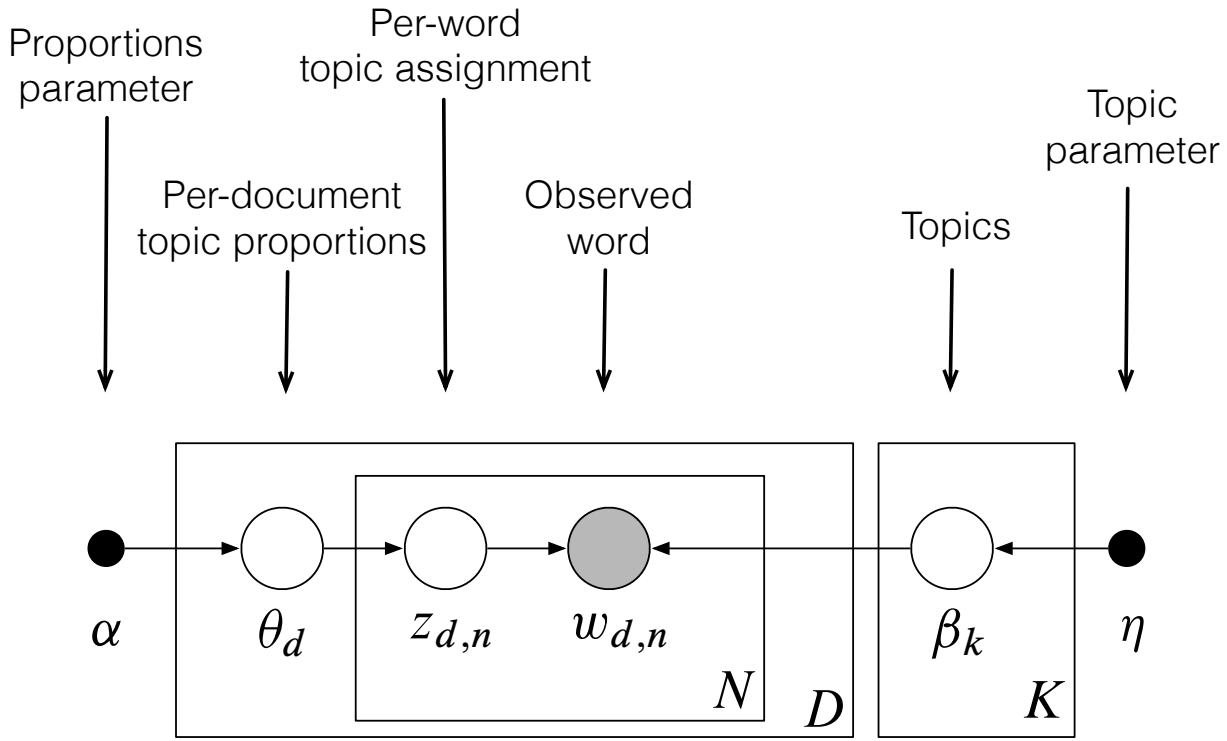
SCIENCE • VOL. 272 • 24 MAY 1996

Latent Dirichlet Allocation



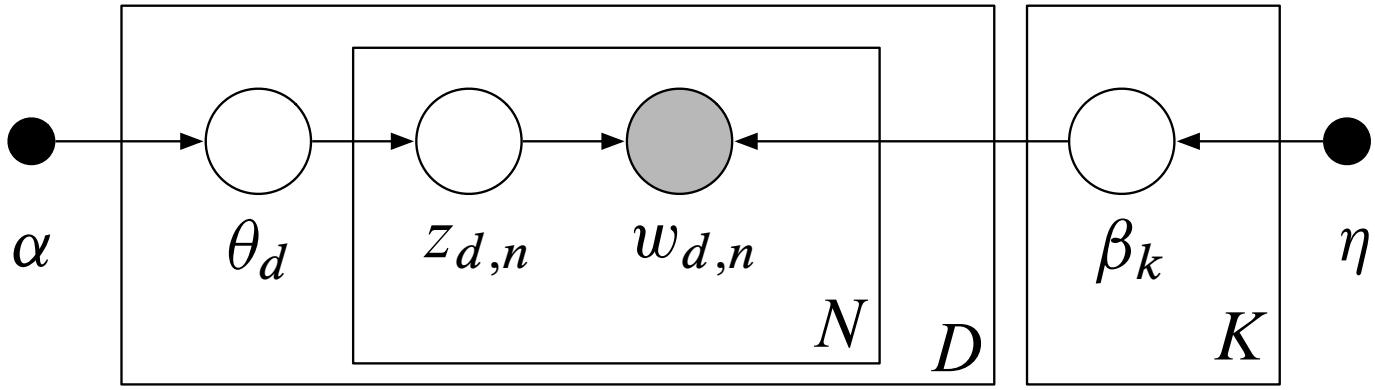
LDA as a graphical model

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

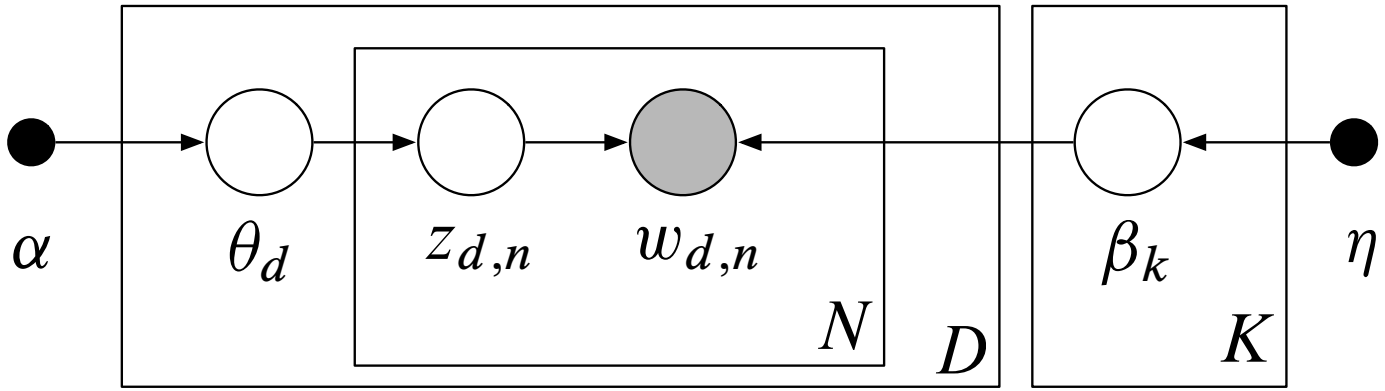


LDA as a graphical model

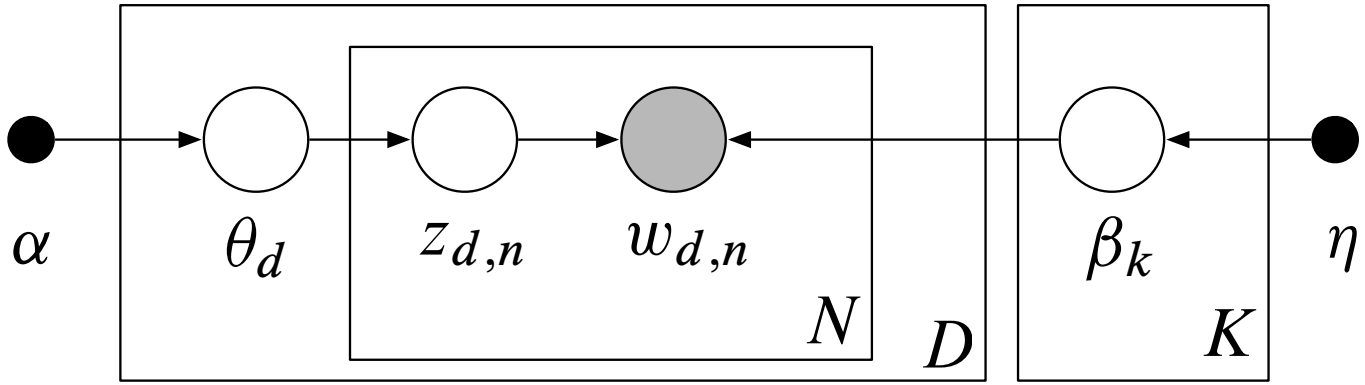
- Defines a factorization of the joint probability distribution
- Encodes independence assumptions about the variables
- Connects to algorithms for computing with data



- The joint defines a posterior, $p(\theta, z, \beta \mid w)$.
- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Then use posterior expectations to perform the task at hand: information retrieval, document similarity, exploration, and others.



- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Distributed sampling (Newman et al., 2008; Ahmed et al., 2012)
- Collapsed variational inference (Teh et al., 2006)
- Stochastic inference (Hoffman et al., 2010, 2013; Mimno et al., 2012)
- Factorization inference (Arora et al., 2012; Anandkumar et al., 2012)
- Amortized inference (Srivastava and Sutton, 2016)



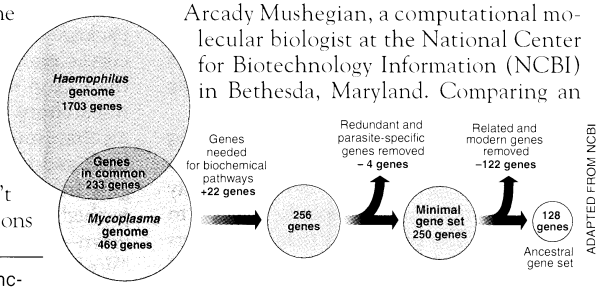
- LDA in R [https://cran.r-project.org/web/packages/lda/]
- GenSim [https://radimrehurek.com/gensim]
- Mallet [http://mallet.cs.umass.edu]
- Vowpal Wabbit [http://hunch.net/∼vw/]
- Apache Spark [http://spark.apache.org/]
- SciKit Learn [http://scikit-learn.org/]

- **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.

Seeking Life's Bare (Genetic) Necessities

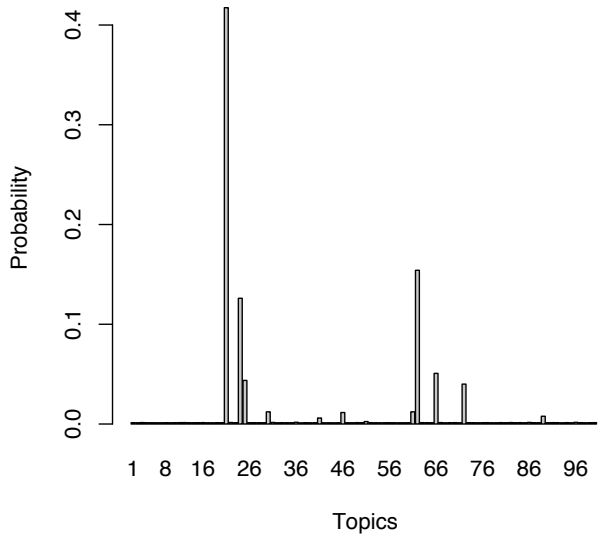
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known 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.



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

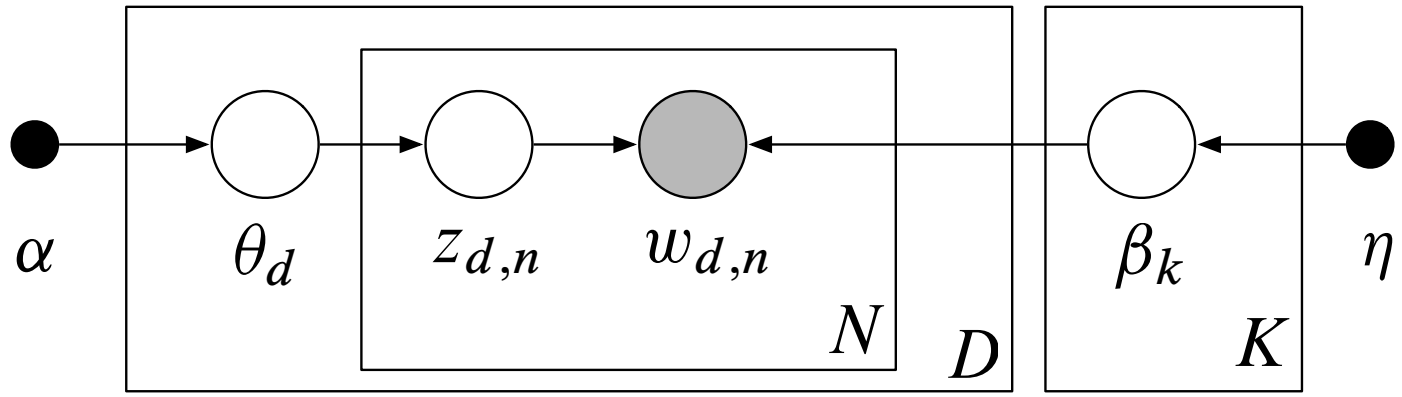
1	2	3	4	5
Game	Life	Film	Book	Wine
Season	Know	Movie	Life	Street
Team	School	Show	Books	Hotel
Coach	Street	Life	Novel	House
Play	Man	Television	Story	Room
Points	Family	Films	Man	Night
Games	Says	Director	Author	Place
Giants	House	Man	House	Restaurant
Second	Children	Story	War	Park
Players	Night	Says	Children	Garden

6	7	8	9	10
Bush	Building	Won	Yankees	Government
Campaign	Street	Team	Game	War
Clinton	Square	Second	Mets	Military
Republican	Housing	Race	Season	Officials
House	House	Round	Run	Iraq
Party	Buildings	Cup	League	Forces
Democratic	Development	Open	Baseball	Iraqi
Political	Space	Game	Team	Army
Democrats	Percent	Play	Games	Troops
Senator	Real	Win	Hit	Soldiers

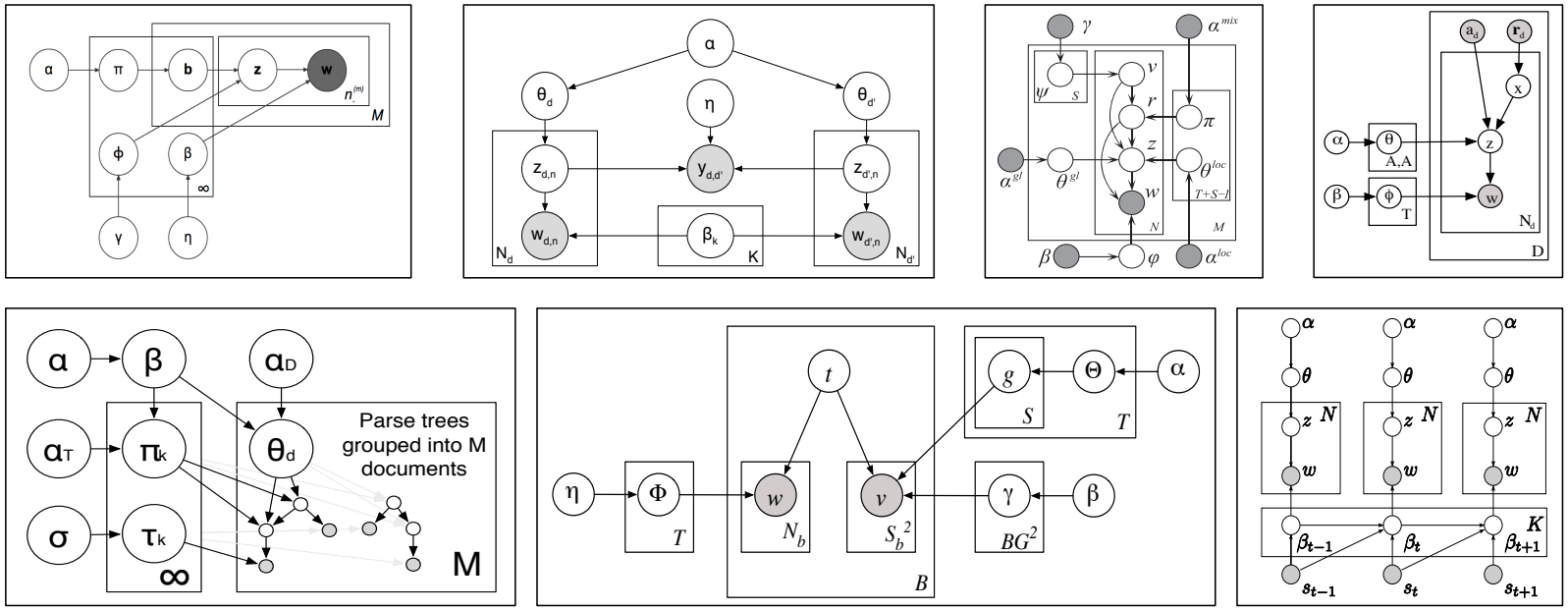
11	12	13	14	15
Children	Stock	Church	Art	Police
School	Percent	War	Museum	Yesterday
Women	Companies	Women	Show	Man
Family	Fund	Life	Gallery	Officer
Parents	Market	Black	Works	Officers
Child	Bank	Political	Artists	Case
Life	Investors	Catholic	Street	Found
Says	Funds	Government	Artist	Charged
Help	Financial	Jewish	Paintings	Street
Mother	Business	Pope	Exhibition	Shot

How does LDA "work"?

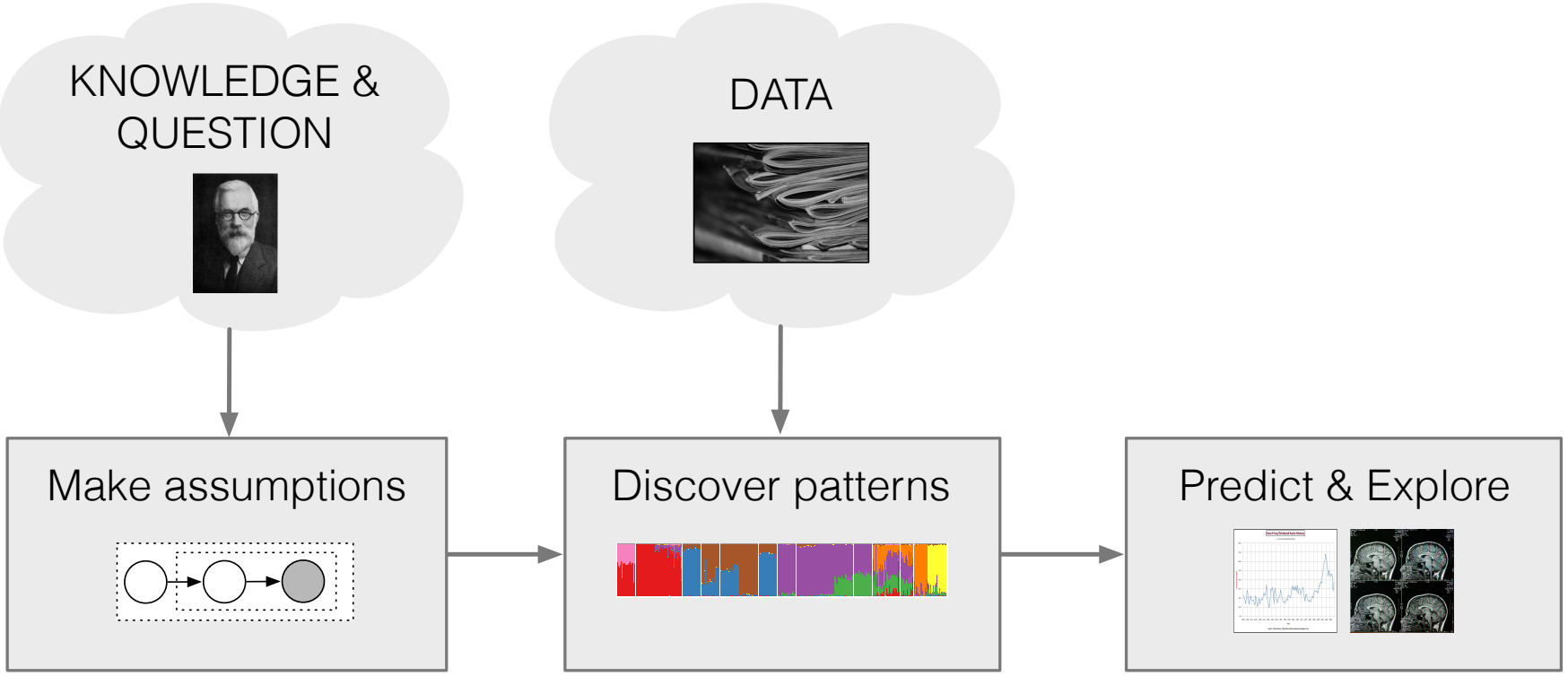
- LDA trades off two goals.
  1. In each **document**, allocate its words to **few topics**.
  2. In each **topic**, assign high probability to **few terms**.
- These goals are at odds.
  - Putting a document in a single topic makes #2 hard:
    All of its words must have probability under that topic.
  - Putting very few words in each topic makes #1 hard:
    To cover a document's words, it must assign many topics to it.
- Trading off these goals finds groups of tightly co-occurring words.



- Summary: LDA discovers themes through posterior inference.
- Other perspectives
  - Latent semantic analysis [Deerwester et al., 1990; Hofmann, 1999]
  - A mixed-membership model [Erosheva, 2004]
  - PCA and matrix factorization [Jakulin and Buntine, 2002]
  - Was independently invented for genetics [Pritchard et al., 2000]



- LDA has become a building block that enables many applications.
- Algorithmic improvements let us fit models to massive data.
  (See VW, Gensim, Mallet, others.)
- Organizing and finding patterns in text is important
  in the sciences, humanities, industry, and culture.



- Case study in **text analysis with probability models**
- Topic modeling research
  - develops new models.
  - develops new inference algorithms.
  - develops new applications, visualizations, tools.