Topic Models: Introduction

Pawan Goyal

CSE, IIT Kharagpur

Week 9, Lecture 1

Why Topic Modeling?

Information Overload

As more information becomes available, it becomes more difficult to find and discover what we need.

Why Topic Modeling?

Information Overload

As more information becomes available, it becomes more difficult to find and discover what we need.

Main Tools: Search and Links

- We type keywords into a search engine and find a set of related documents
- We look at these documents and possibly navigate to other documents

Why Topic Modeling?

Search Based-on themes

- Imagine searching and exploring documents based on themes that run through them.
- We might “zoom-in” or “zoom-out” to find specific or broader themes
- We might look at how themes change through time, how they are connected to each other
- Find the theme first and then examine the documents pertaining to that theme

Why Topic Modeling?

Topic Modeling

Provides methods for automatically organizing, understanding, searching and summarizing large electronic archives without any prior annotation or labeling

- Discover the hidden themes that pervade the collection
- Annotate the documents according to those themes
- Use annotations to organize, summarize, and search the texts

Applications: Discover Topics from a corpus

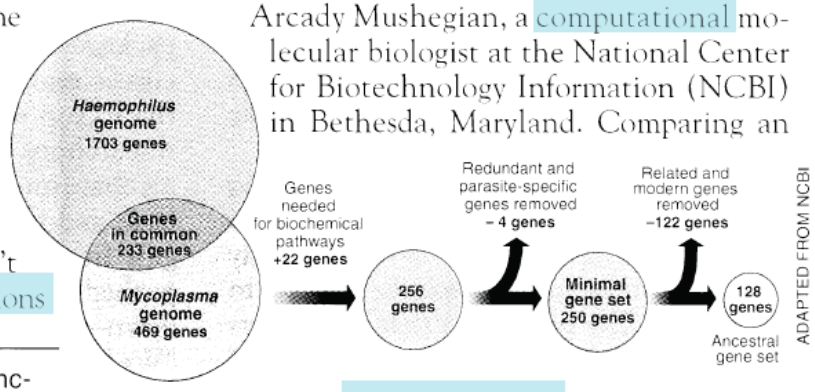
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living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

Intuition: Documents exhibit multiple topics

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

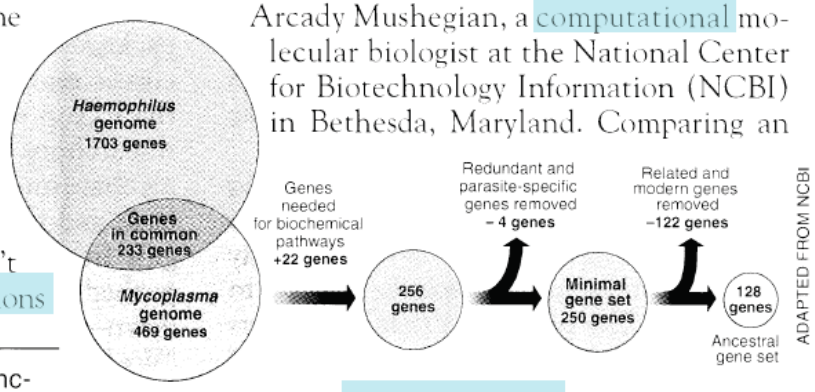
This articles is about using data analysis to determine the number of genes an organism needs to survive

Intuition: Documents exhibit multiple topics

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

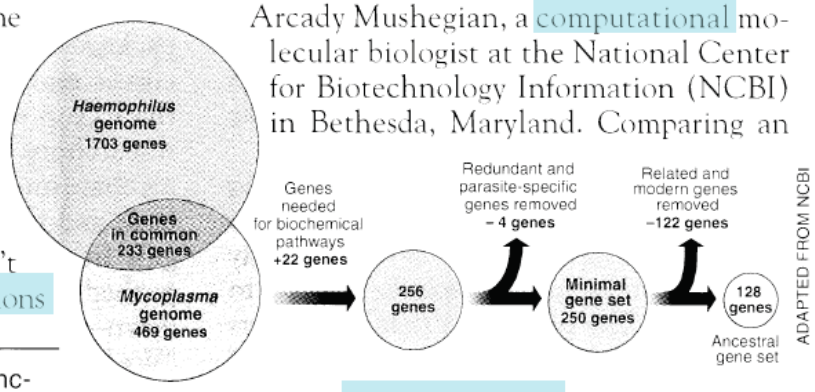
Highlighted words: 'blue': data analysis, 'pink': evolutionary biology, 'yellow': genetics

Intuition: Documents exhibit multiple topics

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

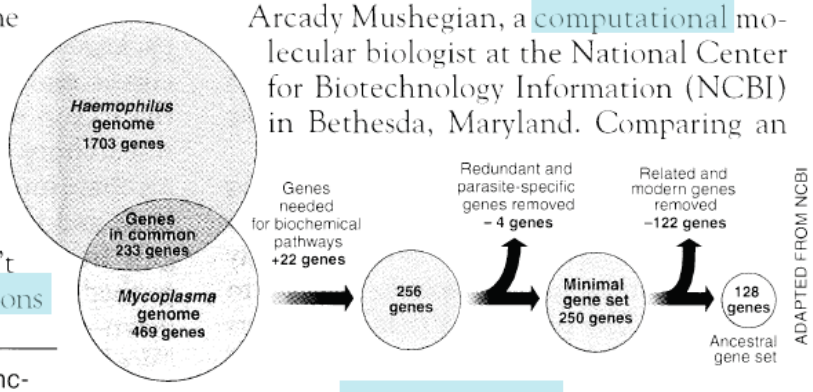
The article blends genetics, data analysis and evolutionary biology in different proportions

Intuition: Documents exhibit multiple topics

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

Knowing that this article blends those topics would help situate it in a collection of scientific articles

Topic Model: Basic Idea

A generative statistical model that captures this intuition.

Generative Model

Documents are mixture of topics, where a topic is a probability distribution over words.

Topic Model: Basic Idea

A generative statistical model that captures this intuition.

Generative Model

Documents are mixture of topics, where a topic is a probability distribution over words.
genetics topic has words about genetics with high probability and the *evolutionary biology* topic has words about evolutionary biology with high probability.

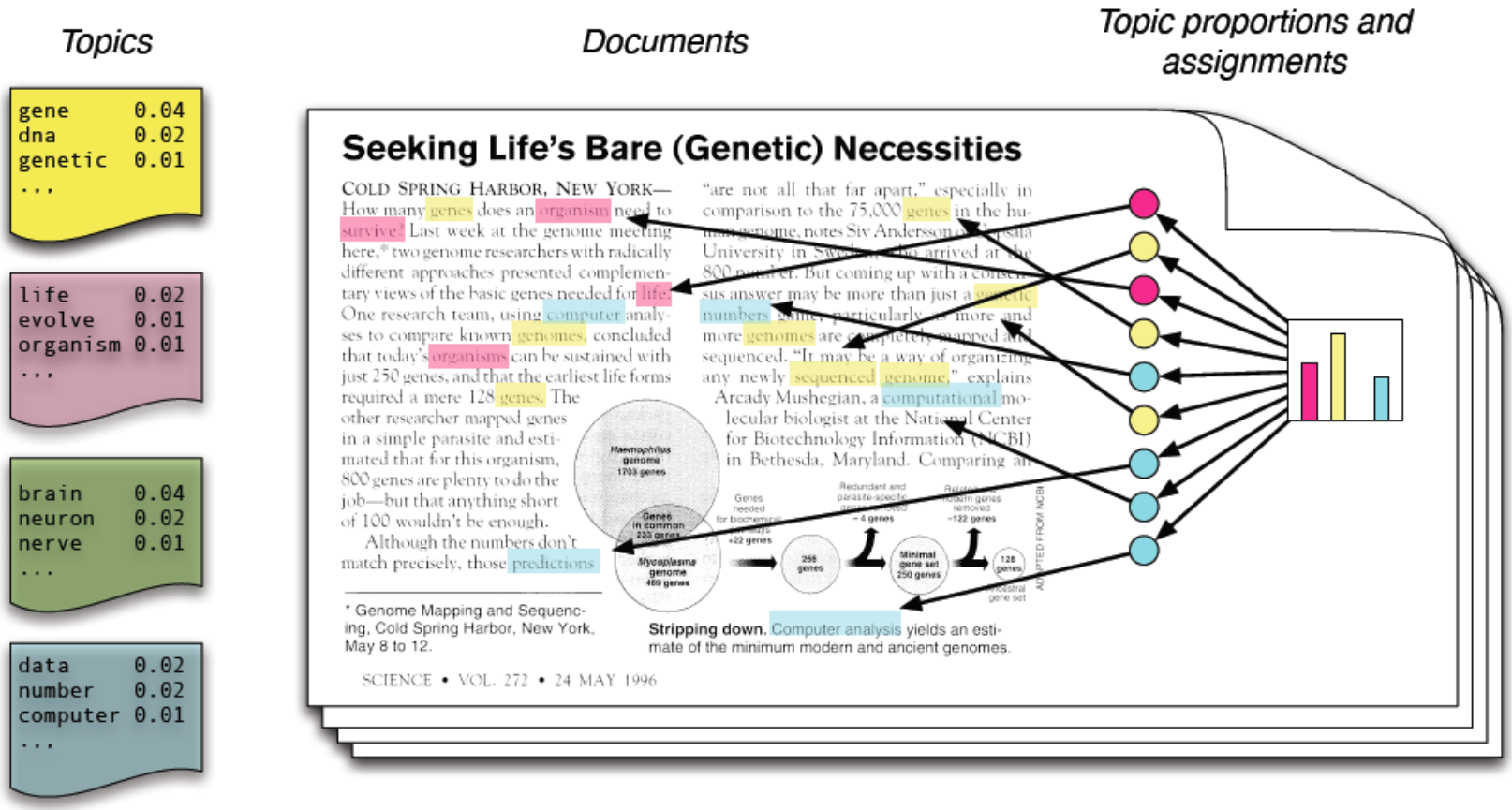
Topic Model: Basic Idea

A generative statistical model that captures this intuition.

Generative Model

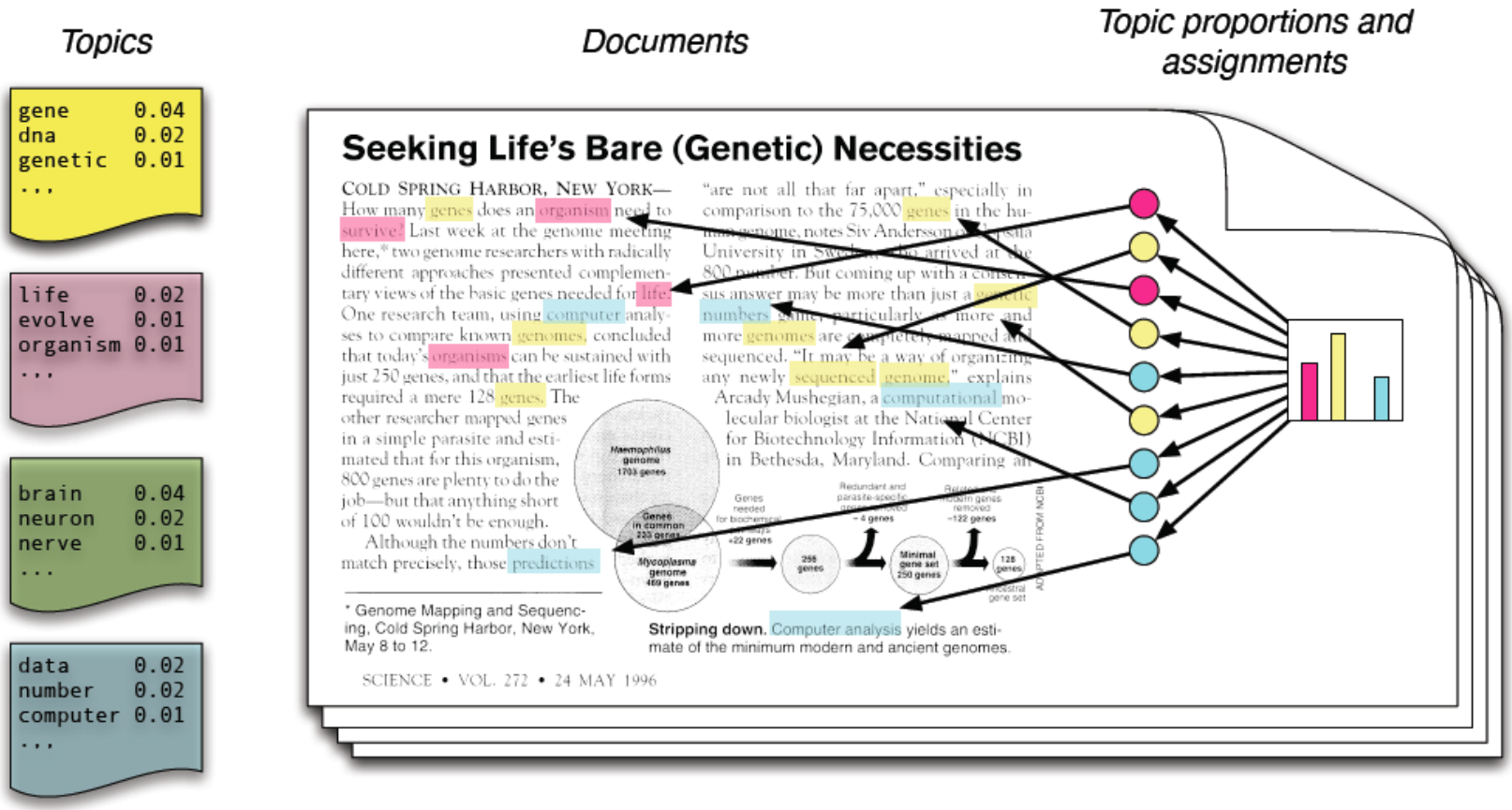
Documents are mixture of topics, where a topic is a probability distribution over words.
genetics topic has words about genetics with high probability and the *evolutionary biology* topic has words about evolutionary biology with high probability.
Technically, the generative model assumes that the topics are generated first, before the documents.

Generative Model for LDA



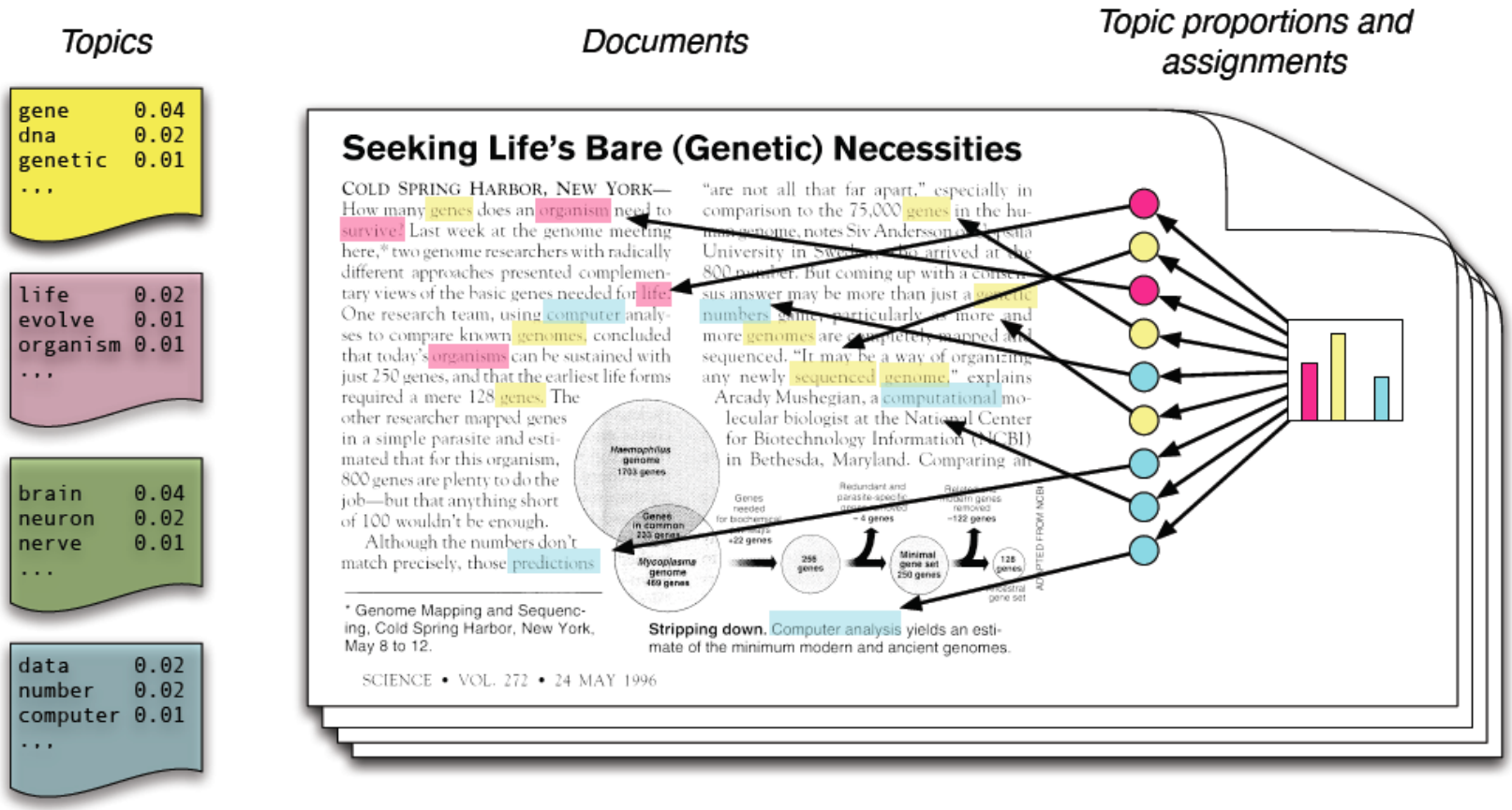
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Generative Model for LDA



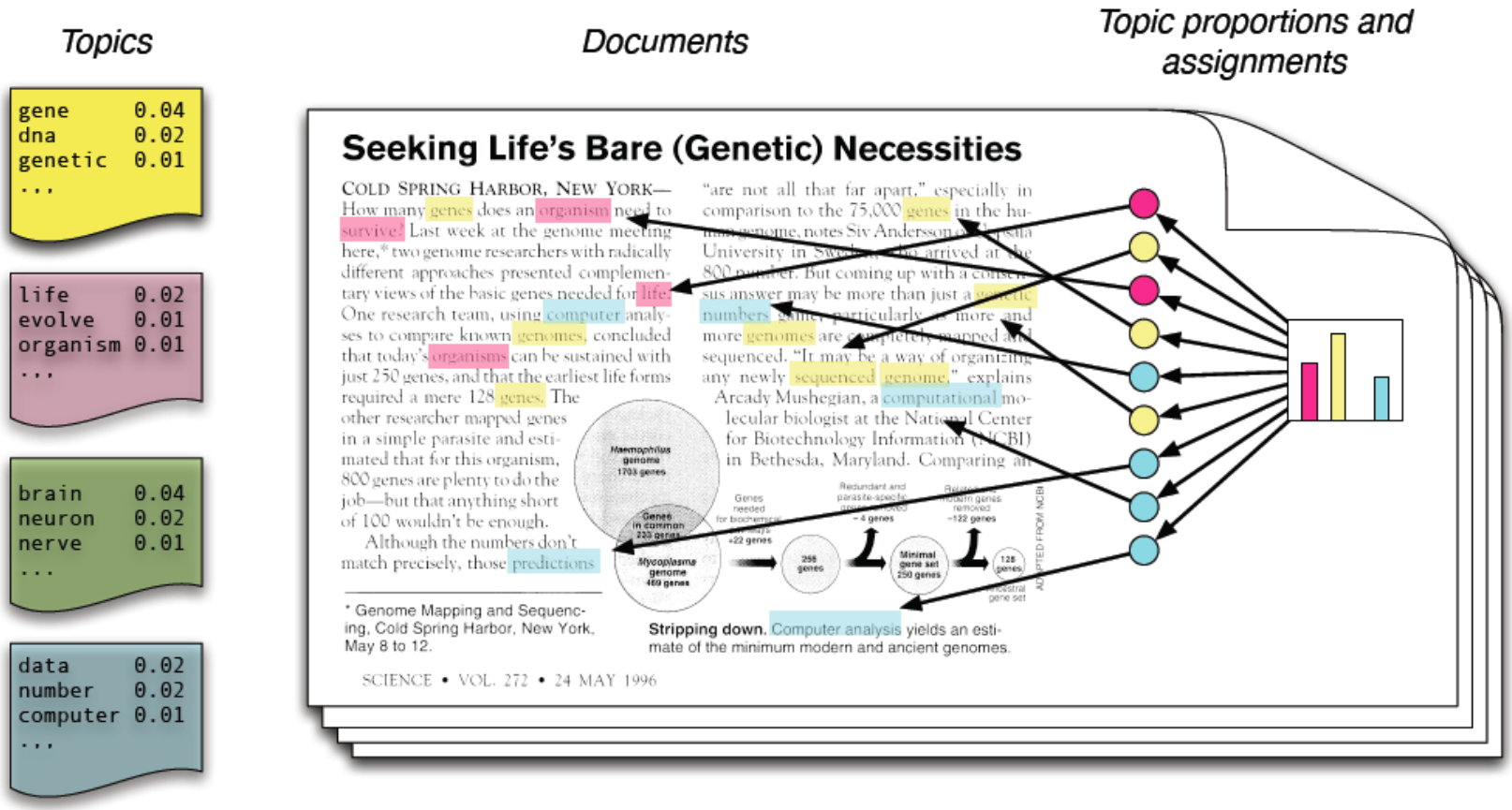
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Generative Model for LDA



- 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

Generative Model for LDA



- 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

What does the statistical model reflect?

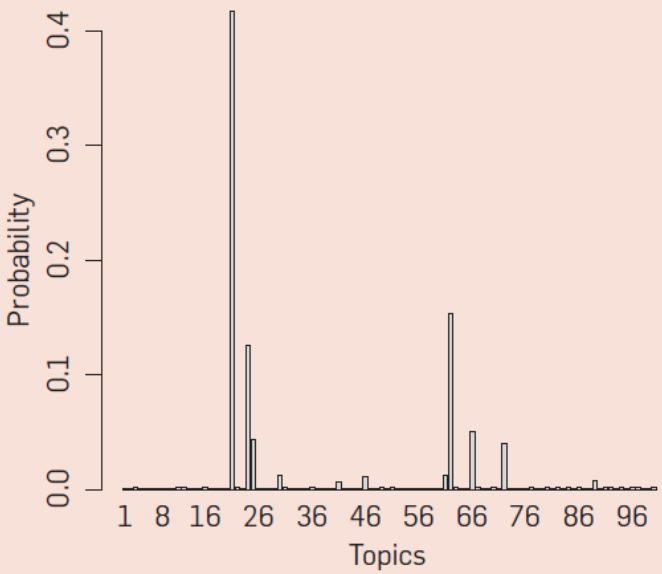
- All the document in the collection share the same set of topics, but each document exhibits those topics in different proportions
- Each word in each document is drawn from one of the topics, where the selected topic is chosen from the per-document distribution over topics

What does the statistical model reflect?

- All the document in the collection share the same set of topics, but each document exhibits those topics in different proportions
- Each word in each document is drawn from one of the topics, where the selected topic is chosen from the per-document distribution over topics

In the example article, the distribution over topics would place probability on *genetics*, *data analytics* and *evolutionary biology*, and each word is drawn from one of those three topics.

Real Inference with LDA for the example article



"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

Central Problem of LDA

- The documents themselves are observed, while the topic structure - the topics, per-document topic distributions, and the per-document per-word topic assignments - is *hidden structure*.
- The central computational problem is to use the observed documents to infer the hidden topic structure, i.e. *reversing* the generative process.