Topic Modeling

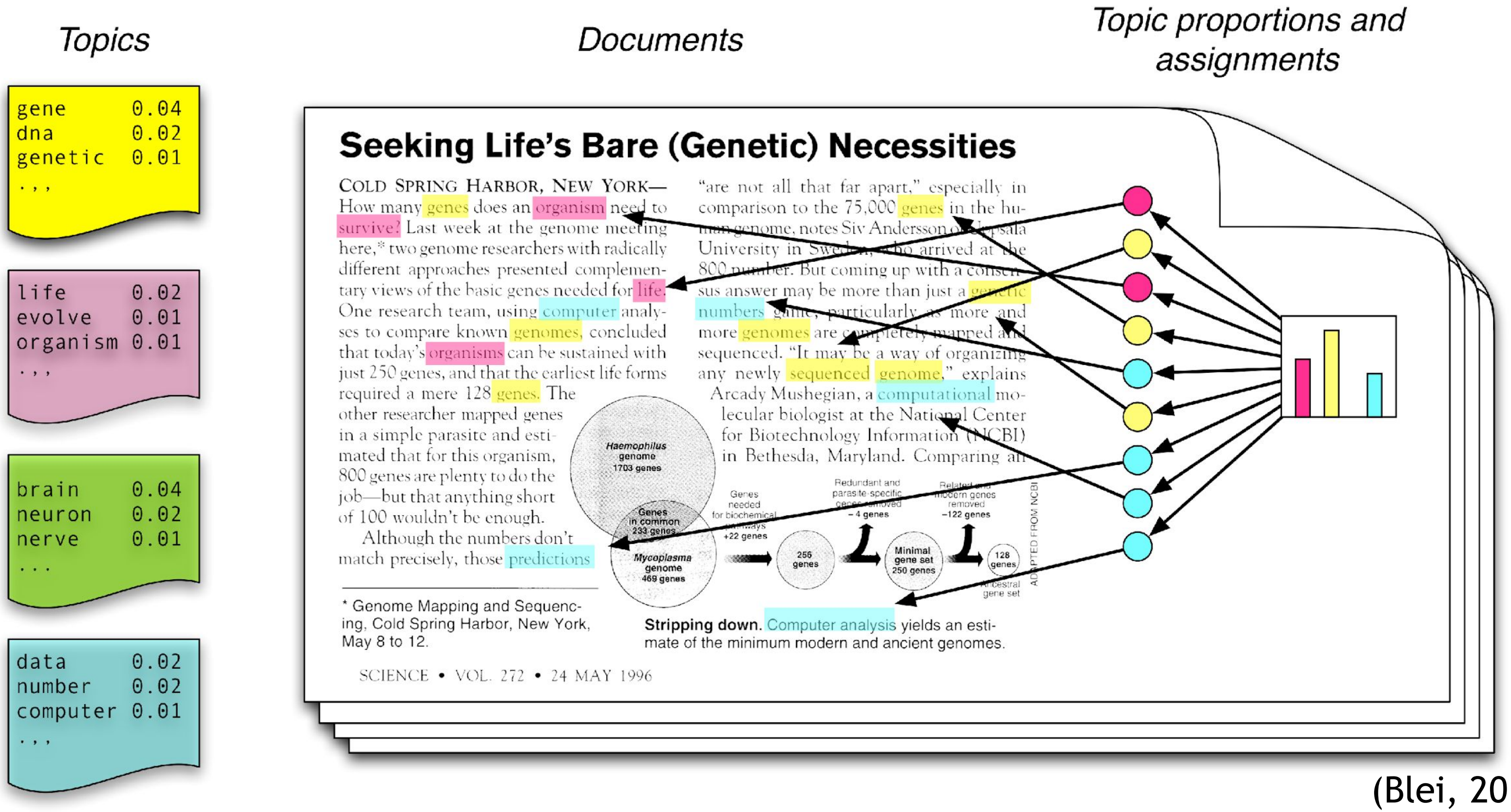
Topic Models

According to David Blei:

“Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections. These algorithms help us develop new ways to search, browse and summarize large archives of texts”

(http://www.cs.columbia.edu/~blei/topicmodeling.html)

Topic Models


Topics
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
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

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
...	

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.

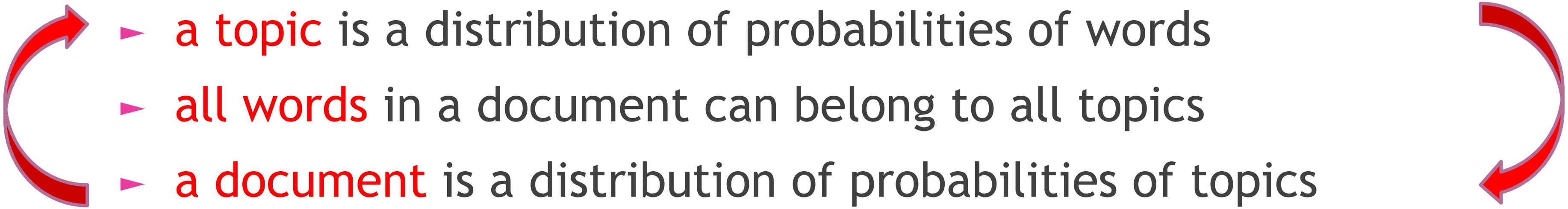
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

(Blei, 2012)

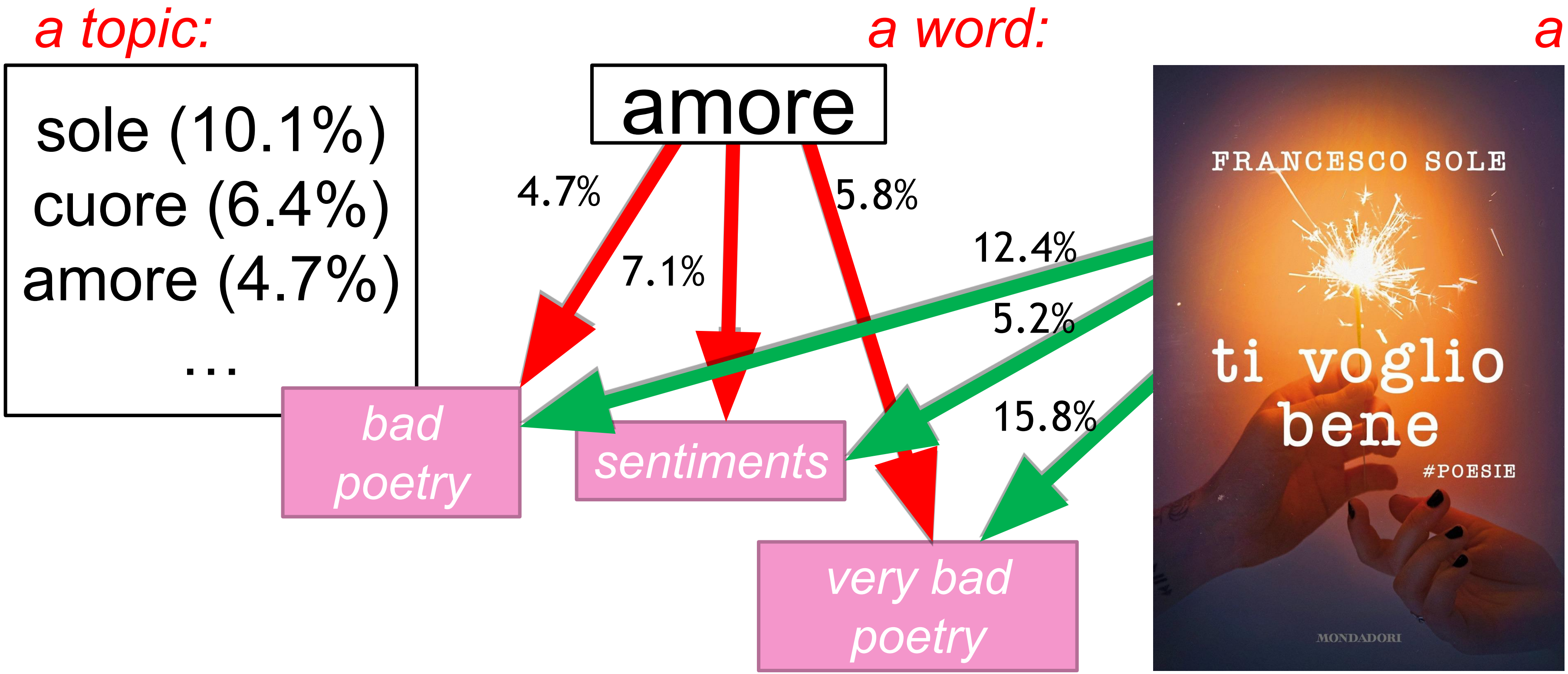
LDA Topic Models

LDA = Latent Dirichlet Allocation



- a topic is a distribution of probabilities of words
- all words in a document can belong to all topics
- a document is a distribution of probabilities of topics

LDA Topic Models



a topic:

sole (10.1%)
cuore (6.4%)
amore (4.7%)
...

a word:

amore

4.7%

7.1%

5.8%

a

12.4%

5.2%

15.8%

bad poetry

sentiments

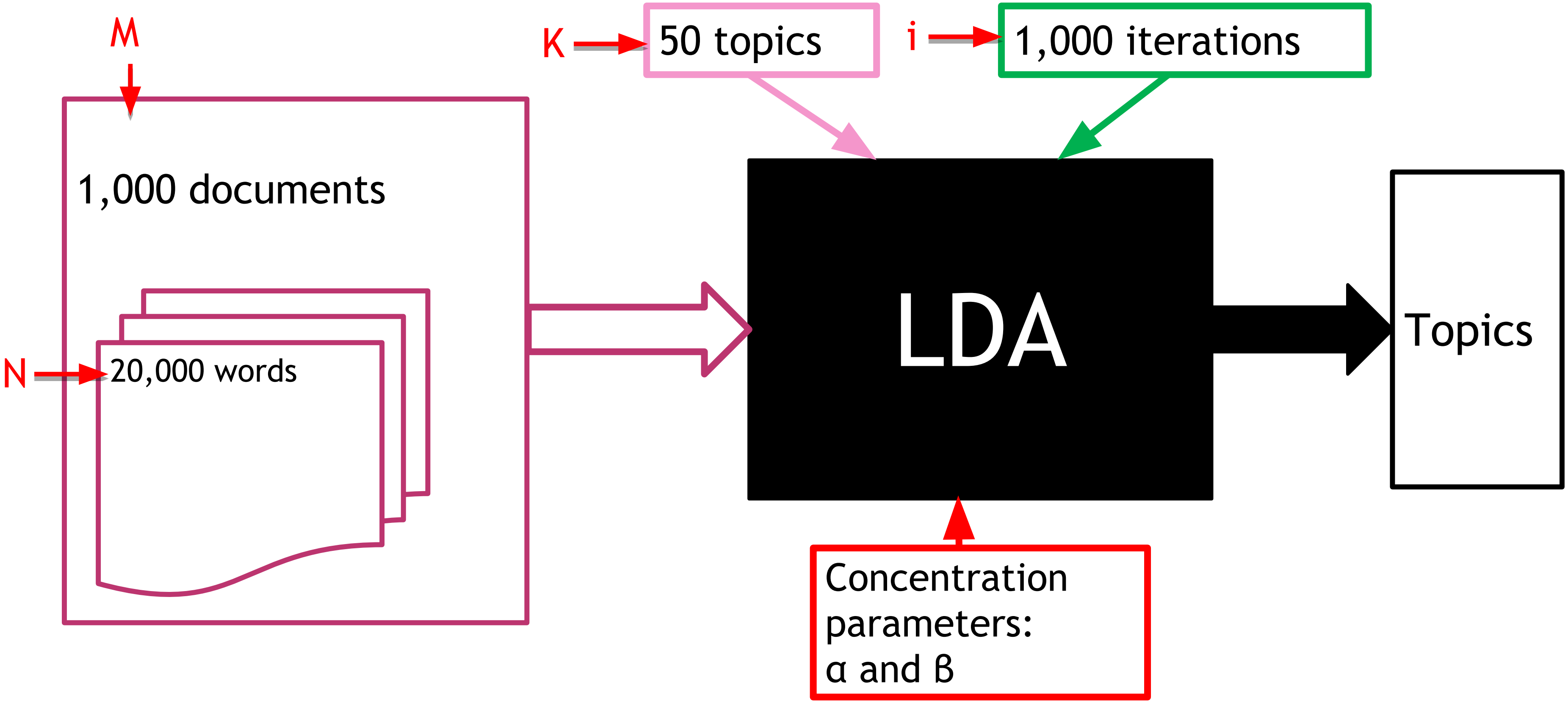
very bad poetry

LDA: How Does it Work?

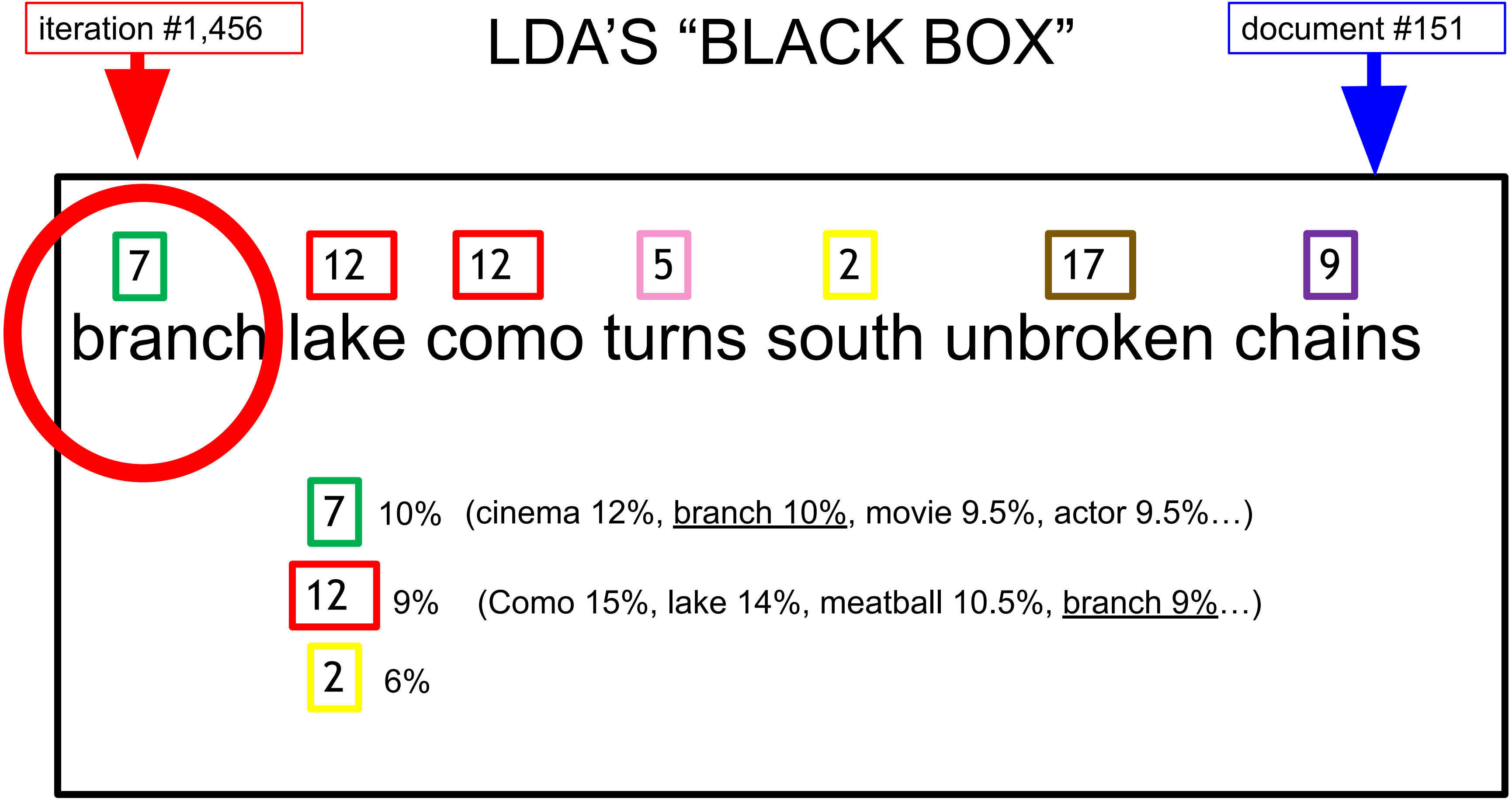
- Initialize topic assignments randomly
- For each word in each document:
  - re-sample topic for word,

    given all other words and their current topic assignments
- Iterate *n* times!

LDA: How Does it Work?



LDA'S "BLACK BOX"



iteration #1,456

document #151

7 12 12 5 2 17 9

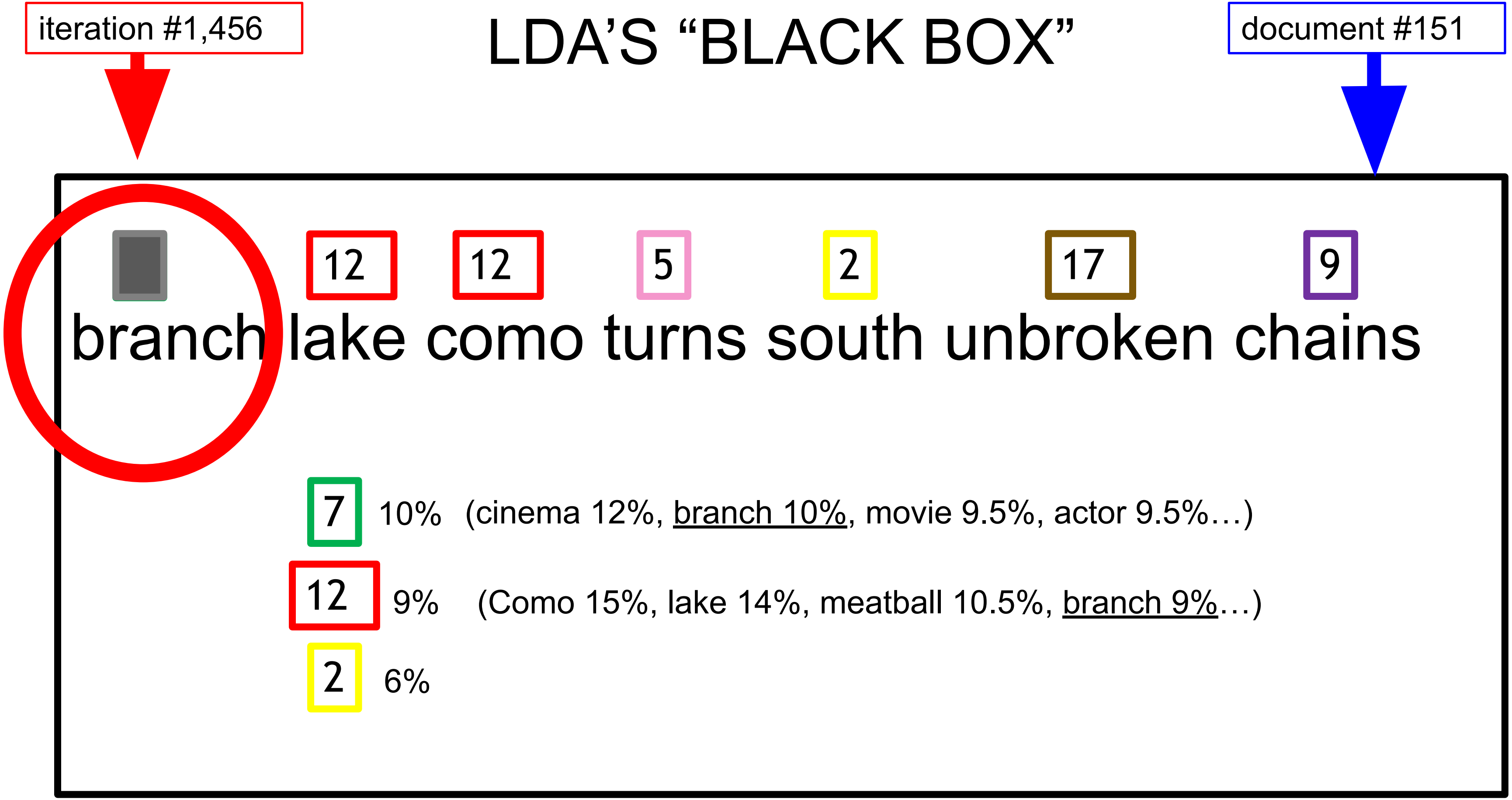
branch lake como turns south unbroken chains

7 10% (cinema 12%, branch 10%, movie 9.5%, actor 9.5%…)

12 9% (Como 15%, lake 14%, meatball 10.5%, branch 9%…)

2 6%

LDA’S “BLACK BOX”



iteration #1,456

document #151

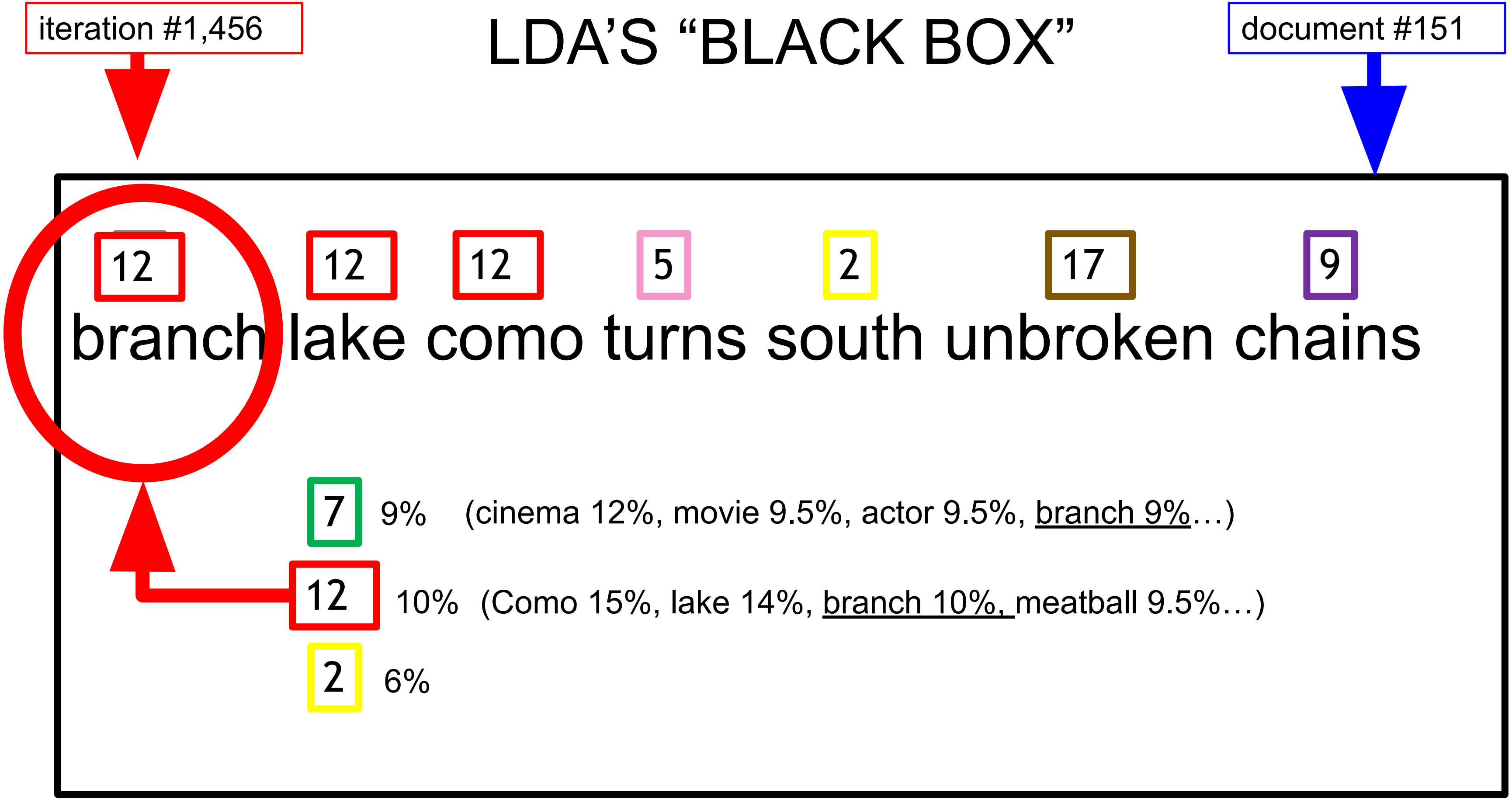
branch lake como turns south unbroken chains

7 10% (cinema 12%, branch 10%, movie 9.5%, actor 9.5%…)

12 9% (Como 15%, lake 14%, meatball 10.5%, branch 9%…)

2 6%

LDA'S "BLACK BOX"



iteration #1,456

document #151

12	12	12	5	2	17	9
branch	lake	como	turns	south	unbroken	chains

7 9% (cinema 12%, movie 9.5%, actor 9.5%, branch 9%…)

12 10% (Como 15%, lake 14%, branch 10%, meatball 9.5%…)

2 6%