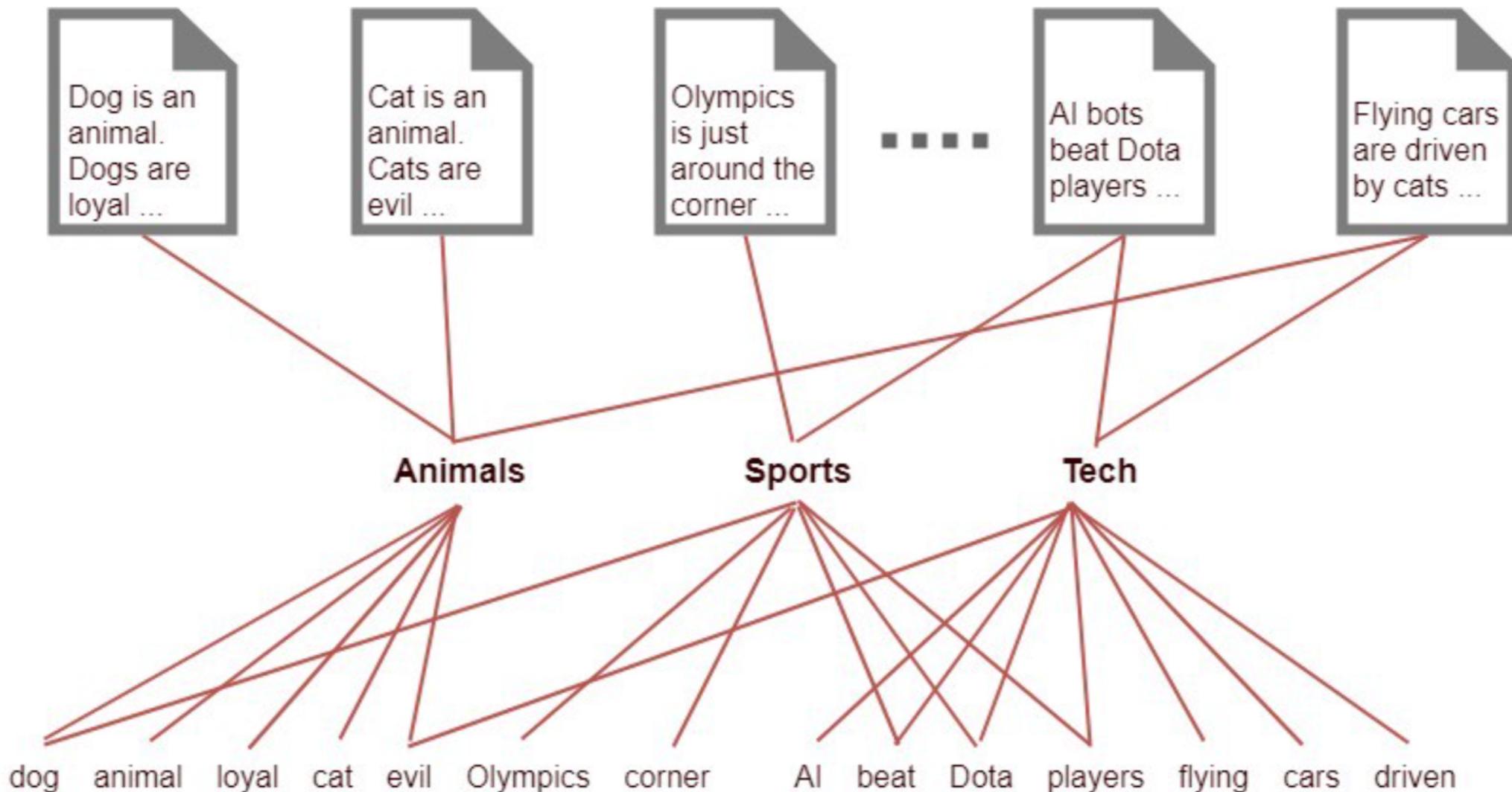


Modeling documents just with words. You can see that we can't really infer any useful information due to the large amount of connections



Words are modeled by a set of topics and documents are modeled by a set of topics. The relationships are clearer than the first example because there's a fewer connections than the first example.

|        | Word1 | word2 | word3 | word4 | ..... |
|--------|-------|-------|-------|-------|-------|
| Topic1 | 0.01  | 0.23  | 0.19  | 0.03  |       |
| Topic2 | 0.21  | 0.07  | 0.48  | 0.02  |       |
| Topic3 | 0.53  | 0.01  | 0.17  | 0.04  |       |

Each topic contains a score for all the words in the corpus.

## A layman's example

Suppose you have various photographs(*documents*) with captions(*words*).

You want to display them in a gallery so you decide to categorize the photographs on various themes(*topics*) based on which you will create different sections in your gallery.



Photo by [Soragrit Wongsa](#) on [Unsplash](#)

|        | Word1 | word2 | word3 | word4 | ..... |
|--------|-------|-------|-------|-------|-------|
| Topic1 | 0.01  | 0.23  | 0.19  | 0.03  |       |
| Topic2 | 0.21  | 0.07  | 0.48  | 0.02  |       |
| Topic3 | 0.53  | 0.01  | 0.17  | 0.04  |       |

Each topic contains a score for all the words in the corpus.

**Figure 1. The intuitions behind latent Dirichlet allocation.** We assume that some number of “topics,” which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.

### 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

## Seeking Life’s Bare (Genetic) Necessities

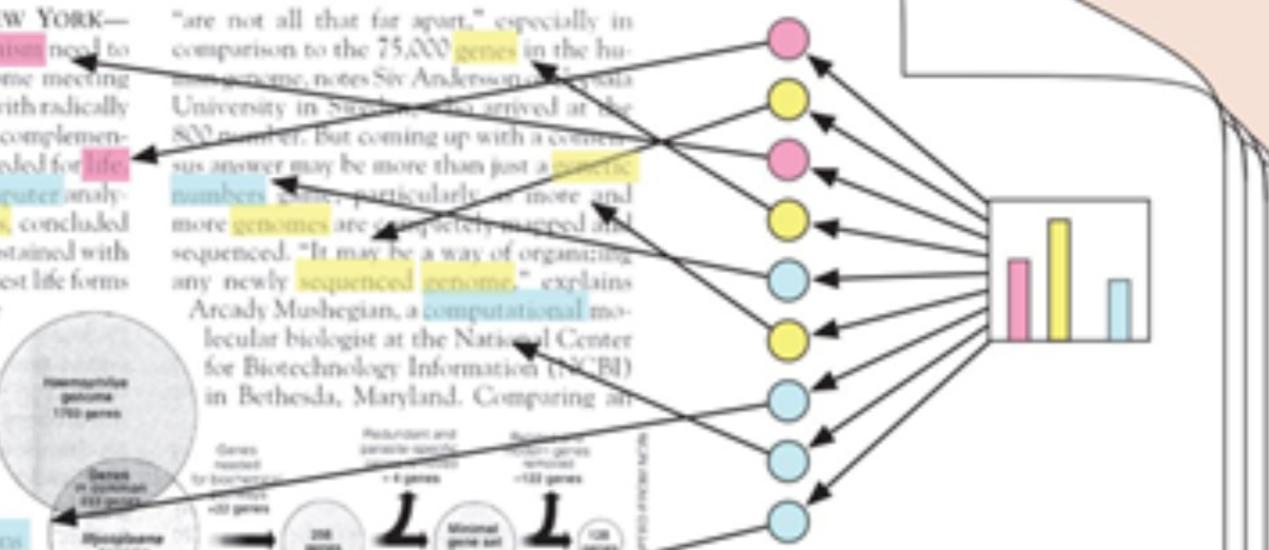
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known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

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

### Topic proportions and assignments



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