IS590Labtime

topic: Expectation-maximization (EM) algorithm and Latent Dirichlet Allocation (LDA) algorithm

Date: 10/18

Time: 4:40-5:30pm

Instructor: Yingjun Guan

A big picture of Machine Learning.

Super VIP Cheatsheet: Machine Learning

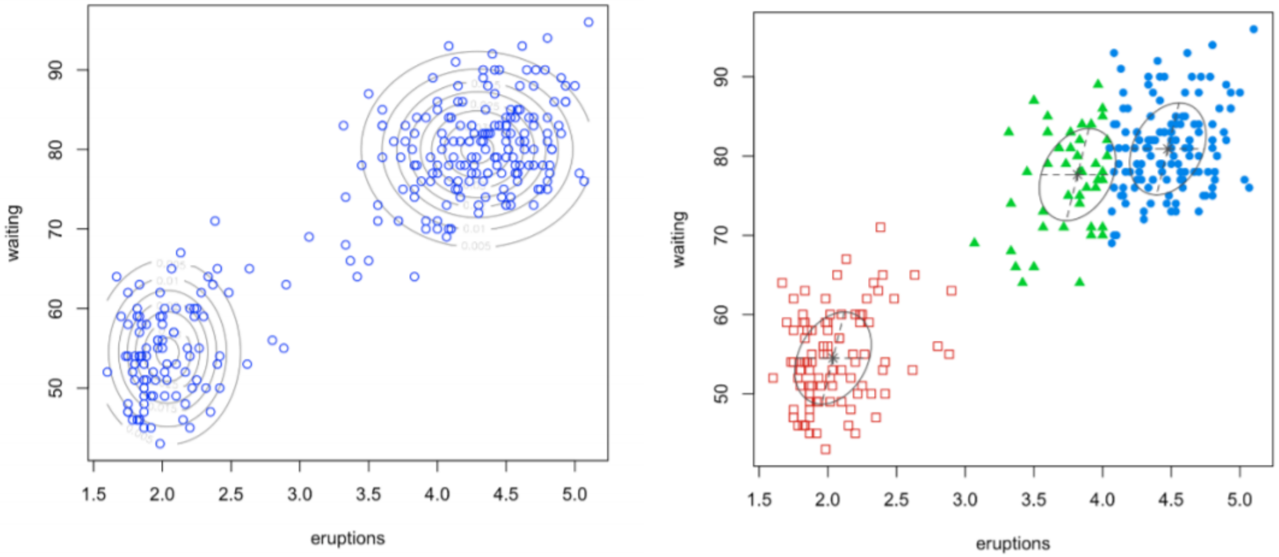
Afshine Amidi and Shervine Amidi

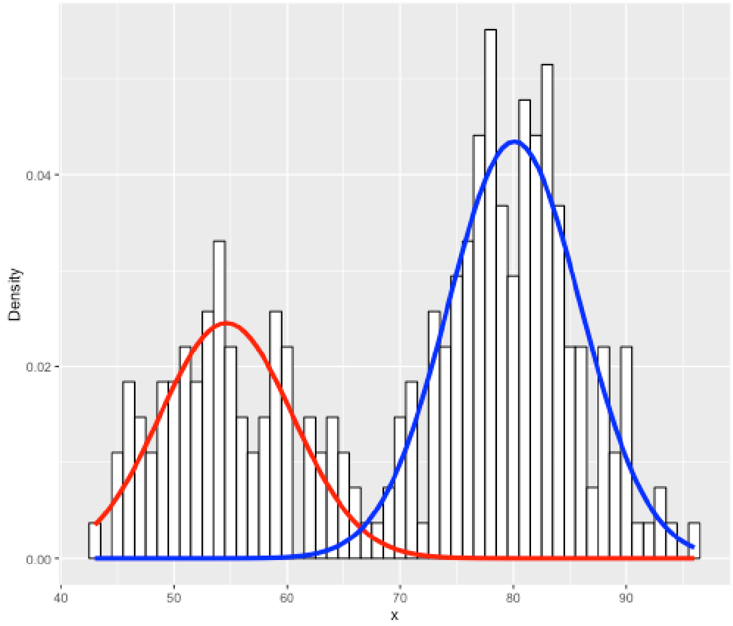
October 6, 2018

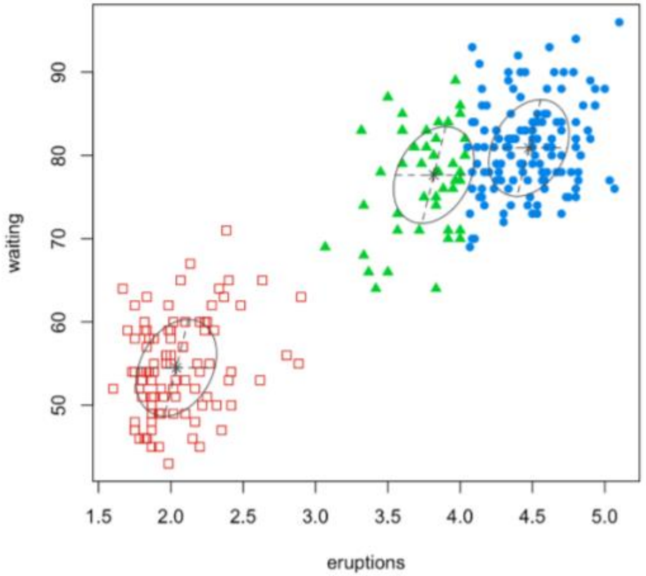
Contents

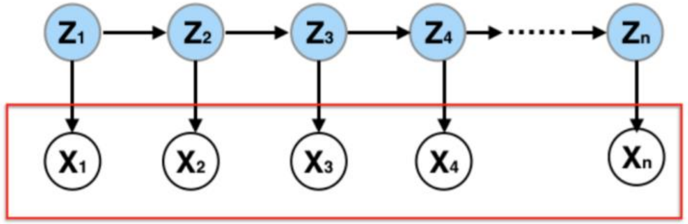
Today, we're talking about model-based clustering.

Model-based clustering refers to clustering a set of data points $(x_1, \ldots, x_n)$ by fitting a mixture model on this data set, where each cluster corresponds to a component of the mixture model.









Latent variables are hidden/unobserved data help to estimate the clustering, thus could be described in different ways.

What is EM?

- The Expectation-Maximization (EM) algorithm is an iterative method that finds the MLE by enlarging the sample with unobserved latent data.

Parameter → Probability [Expectation]

- E-step: Evaluate the posterior probability $Q_i(z^{(i)})$ that each data point $x^{(i)}$ came from a particular cluster $z^{(i)}$ as follows:

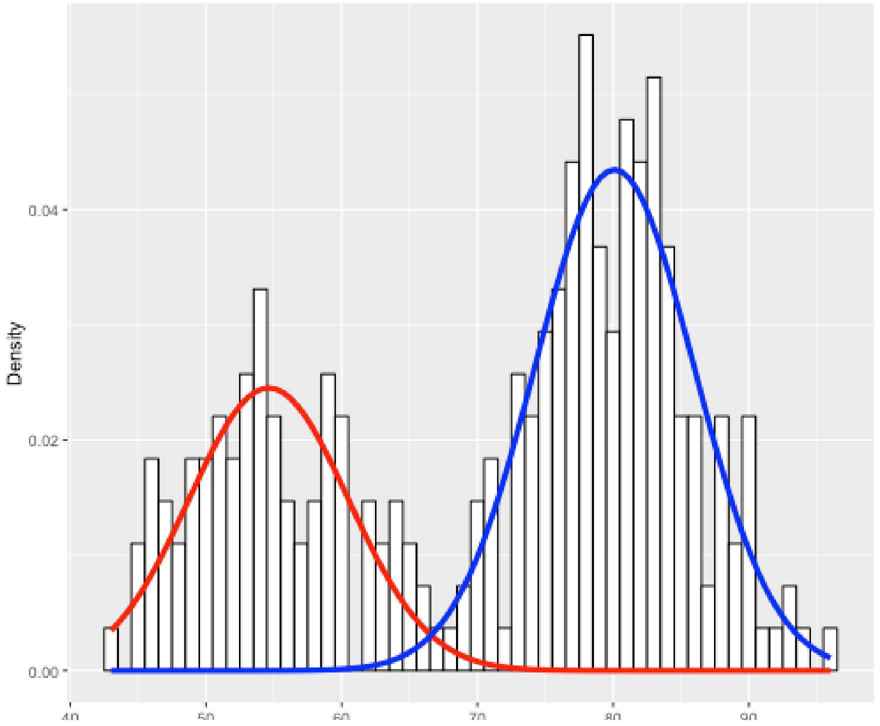
$$\boxed{Q_i(z^{(i)}) = P(z^{(i)}|x^{(i)};\theta)}$$

- M-step: Use the posterior probabilities $Q_i(z^{(i)})$ as cluster specific weights on data points $x^{(i)}$ to separately re-estimate each cluster model as follows:

$$\boxed{\theta_i = \underset{\theta}{\operatorname{argmax}} \sum_i \int_{z^{(i)}} Q_i(z^{(i)}) \log\left(\frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}\right) dz^{(i)}}$$

Probability → Parameter [Maximization]

How to explain EM in equations (n-component gaussian mixture)



$$\log p(\mathbf{x}|\theta) = \sum_{i=1}^{n} \log \left[\pi\phi_{\mu_1,\sigma_1^2}(x_i) + (1-\pi)\phi_{\mu_2,\sigma_2^2}(x_i) \right].$$

- E-step: Let θ_0 denote the current value of θ. Find $p(\mathbf{Z}|\mathbf{x},\theta_0)$, the distribution of the latent variable $\mathbf{Z}$ given the data $\mathbf{x}$ and θ_0, and then calculate the following expectation

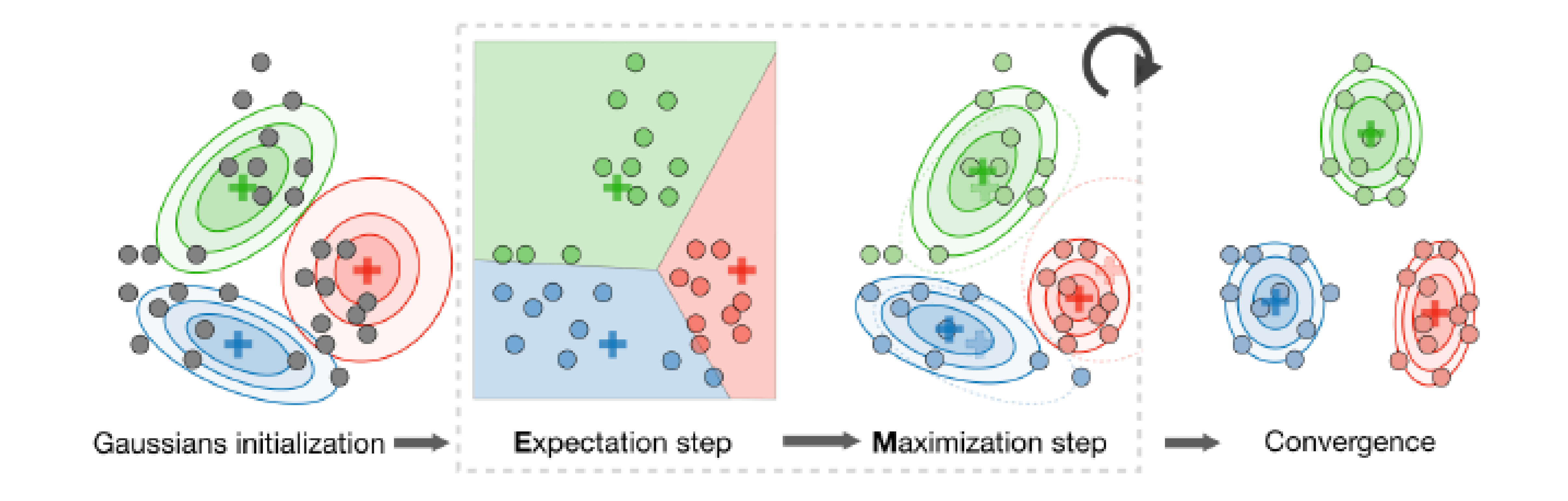
$$g(\theta) = \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta_0} \log p(\mathbf{x},\mathbf{Z}|\theta)$$

which is

$$\sum_{\mathbf{z}} p(\mathbf{Z}=\mathbf{z}|\mathbf{x},\theta_0) \log p(\mathbf{x},\mathbf{z}|\theta), \quad \text{or} \quad \int p(\mathbf{z}|\mathbf{x},\theta_0) \log p(\mathbf{x},\mathbf{z}|\theta) d\mathbf{z}.$$

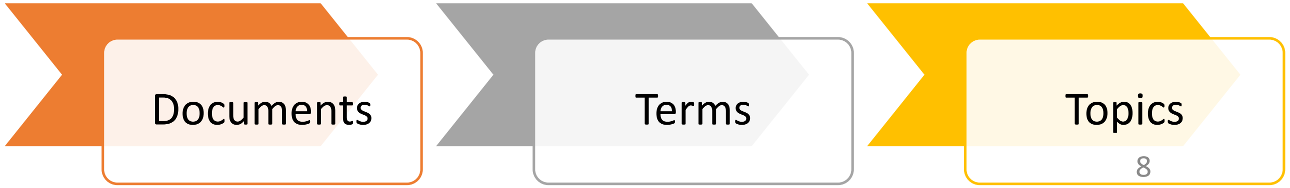
- M-step: Find θ_1 that maximizes $g(\theta)$.
- Replace θ_0 by θ_1 and repeat the above E and M steps until convergence.

How to explain EM in figures. (n-component gaussian mixture)



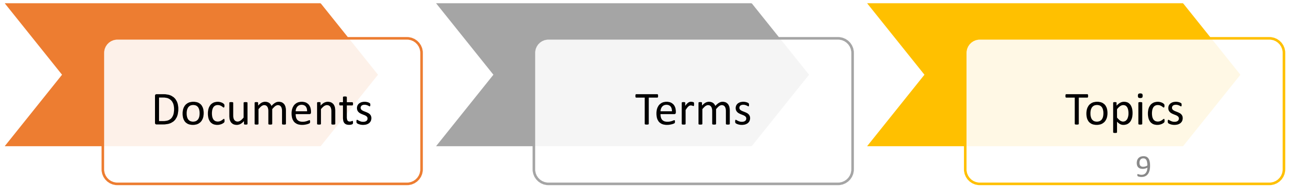
Hypotheses and targets of LDA (Latent Dirichlet Allocation)

- We want to find themes (or topics) in documents
  - useful for e.g. search or browsing
- We don't want to do supervised topic classification
  - rather not fix topics in advance nor do manual annotation
- Need an approach which automatically teases out the topics
- This is essentially a clustering problem
  - can think of both words and documents as being clustered



Key Assumptions behind the LDA topic model

- Documents exhibit multiple topics (but typically not many)
- LDA is a probabilistic model with a corresponding generative process
  - each document is assumed to be generated by this (simple) process
- A topic is a distribution over a fixed vocabulary
  - these topics are assumed to be generated first, before the documents
- Only the number of topics is specified in advance



Now, let's look at the generative process of LDA.

To generate a document:

1. Randomly choose a distribution over topics

2. For each word in the document

   a. randomly choose a topic from the distribution over topics
   b. randomly choose a word from the corresponding topic (distribution over the vocabulary)

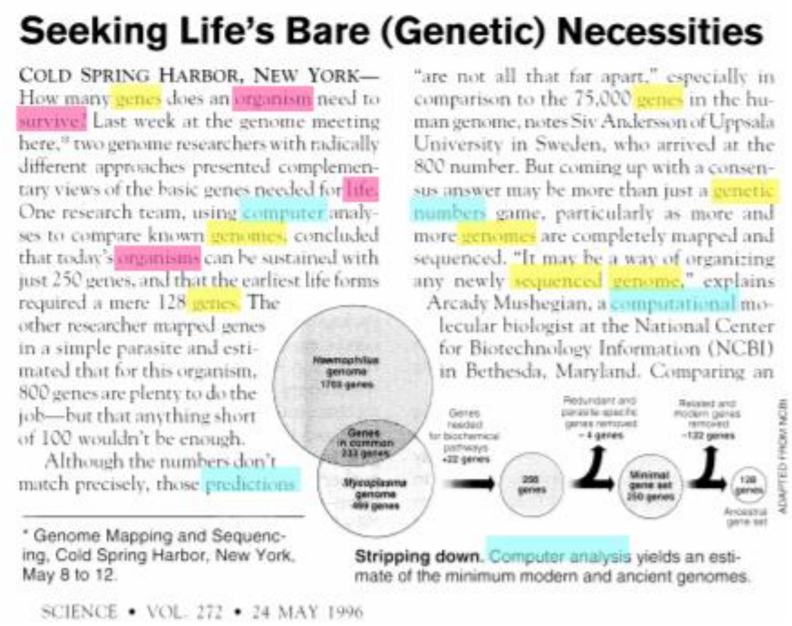
- Note that we need a distribution over a distribution (for step 1)

- Note that words are generated independently of other words (unigram bag-of-words model)

Documents

Terms

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.

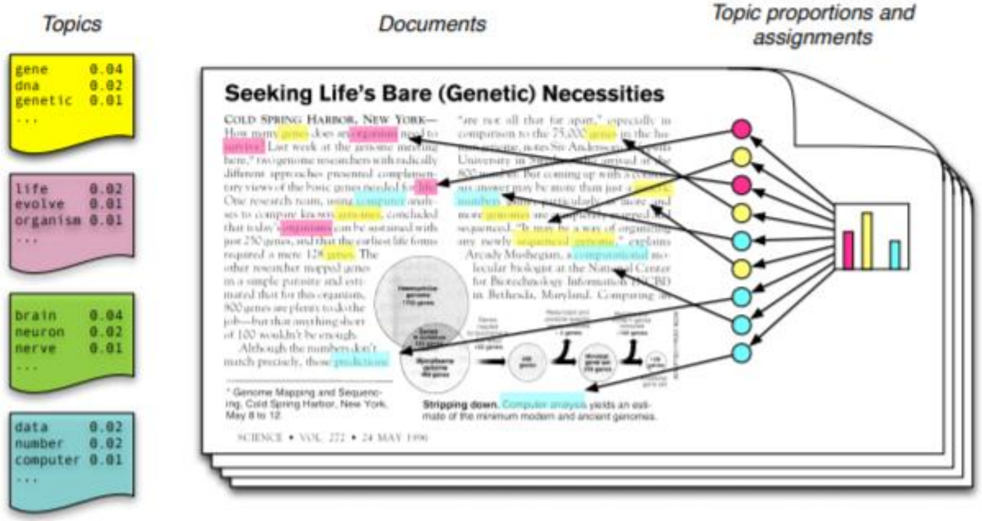
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

Simple intuition: Documents exhibit multiple topics.



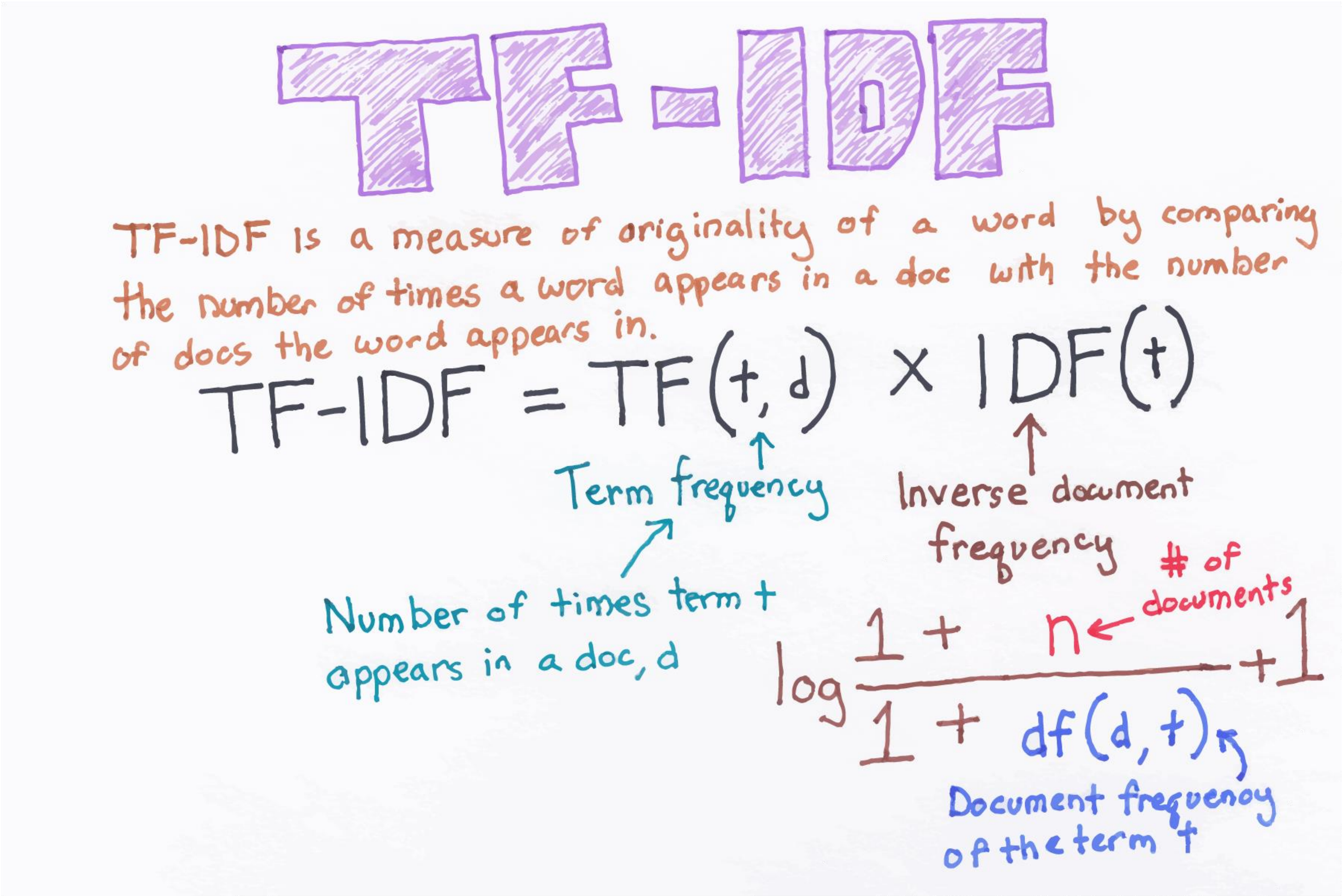
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Documents → terms → topics

R-implementation

- Are you with me so far?
- Any questions are welcome.
- Thank you for the feedback last week.
  - Time
  - Hands-on practice
  - Topics
  - Pace/background

TF-IDF: one way of textual feature selection



Reference:

- Online tutorial. https://www.cl.cam.ac.uk/teaching/1213/L101/clark_lectures/lect7.pdf
- David Blei's webpage is a good place to start
- Intro to EM on Youtube https://www.youtube.com/watch?v=REypj2sy_5U
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.
- https://www.tidytextmining.com/topicmodeling.html

Reference: (cont'd)

- Repository for today's lab:
  - Lecture slides
  - R file.
  - Other references.
- Online Access to today's lab.
  - 590DT student: log in through moodle directly.
  - Guest will be available through the following link: https://us.bbcollab.com/guest/3DA671955178CE4AF10F31B05C983E27
- Feedback for today's lab.
  - https://goo.gl/forms/qpDmxEsYzUwbK7zx2