K-Means for Document Clustering

School of Information Studies
Syracuse University

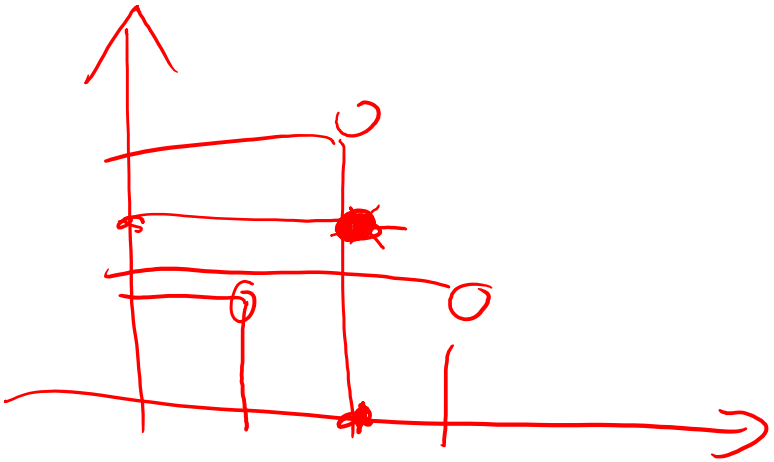
Similarity/Distance Measures

Euclidean distance

Cosine similarity measure

Centroid of a Cluster

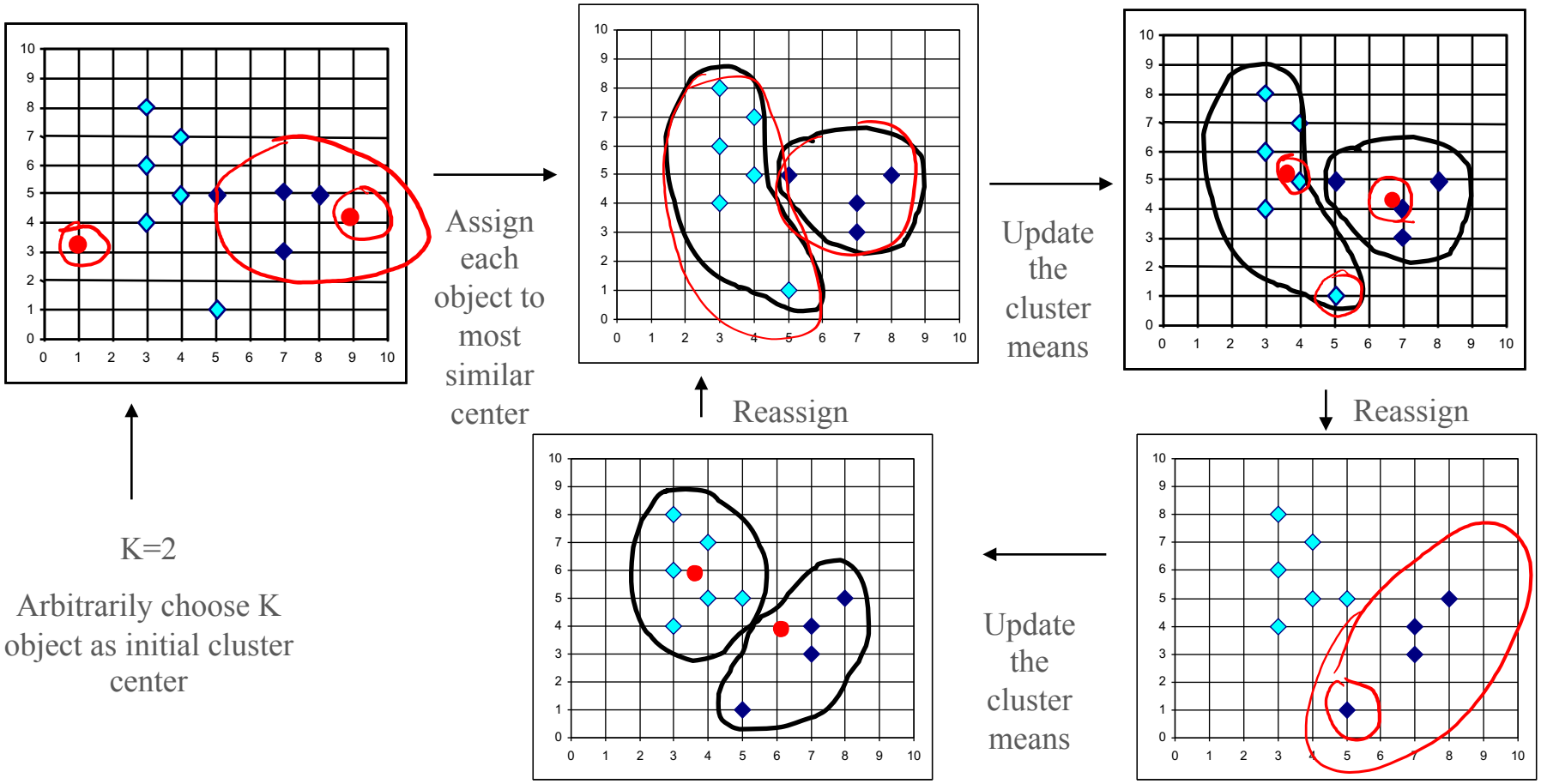
Centroid: The "center" of a cluster is a (pseudo) instance of data in which each attribute is the "mean" of all the attribute values in the cluster.



The *K-Means* Clustering Method

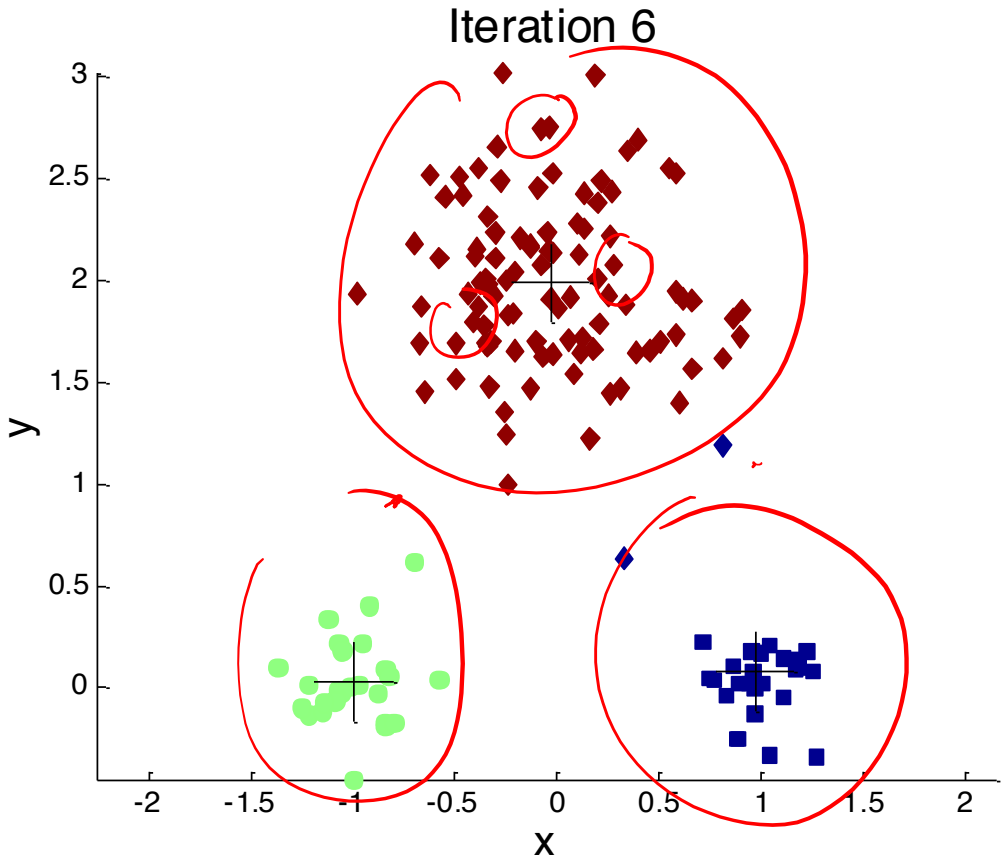
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

The *K-Means* Clustering Method: Example



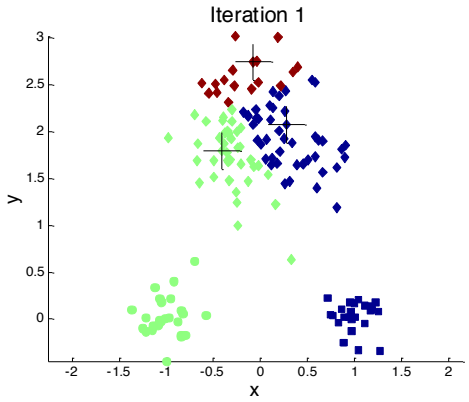
School of Information Studies
Syracuse University

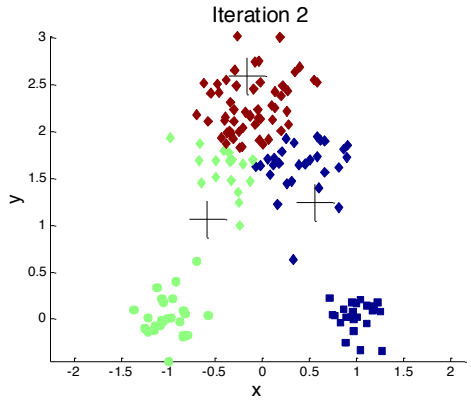
Importance of Choosing Initial Centroids

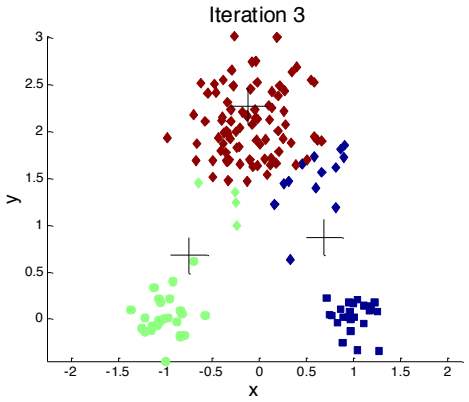


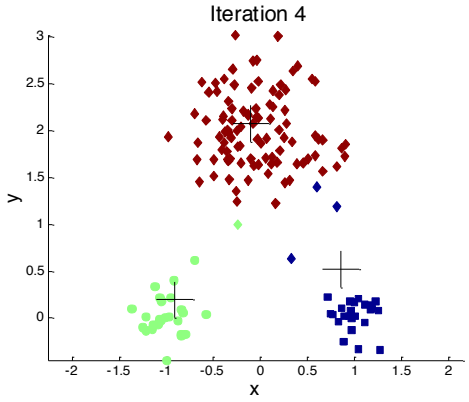
A good clustering result

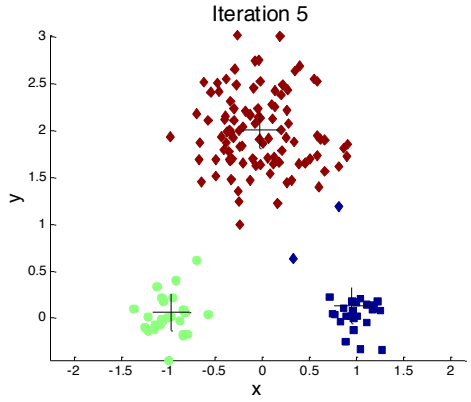
Importance of Choosing Initial Centroids

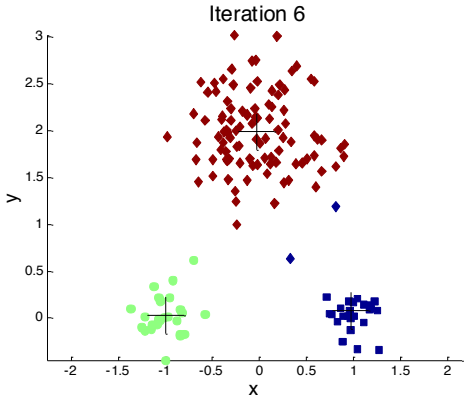




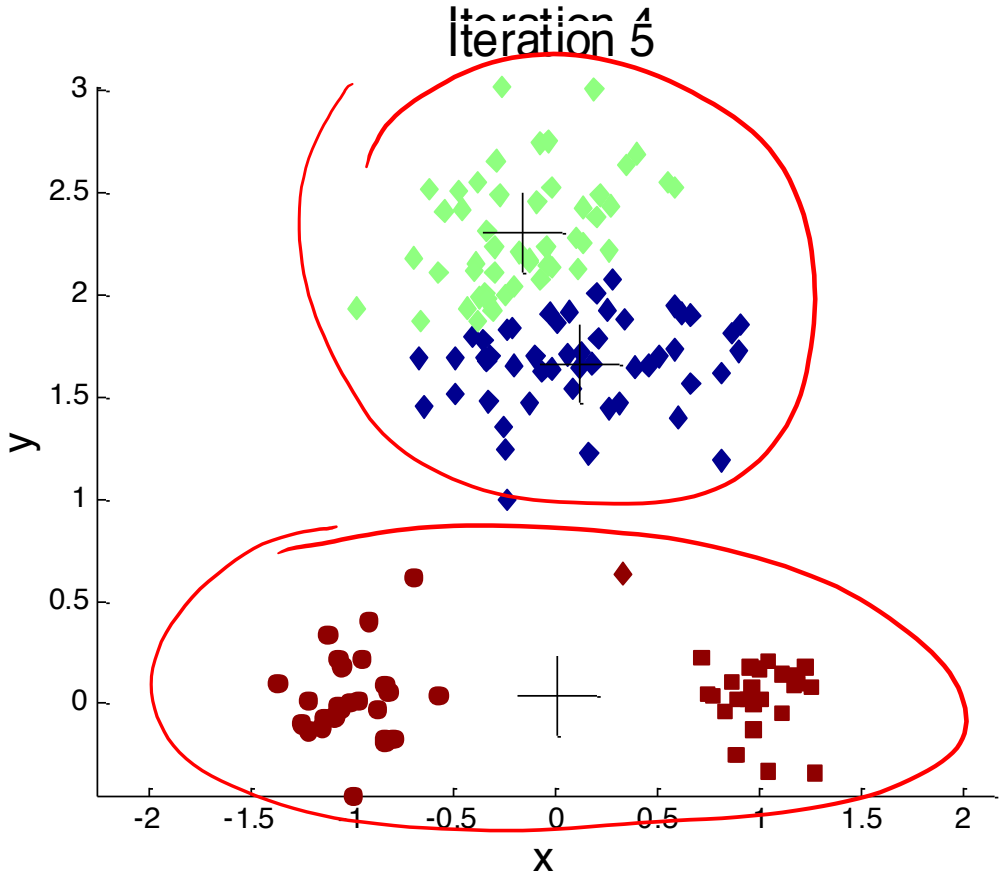






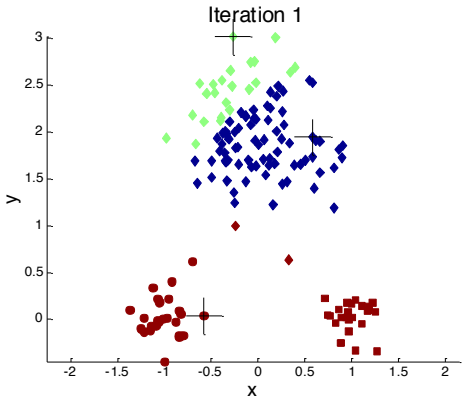


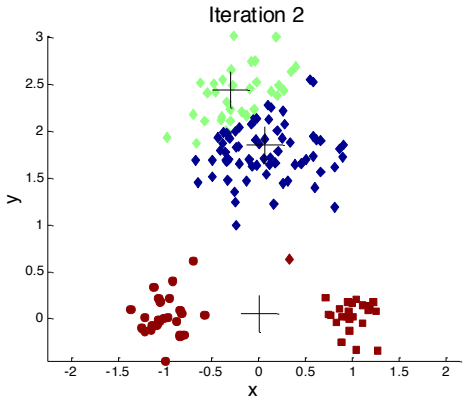
Importance of Choosing Initial Centroids

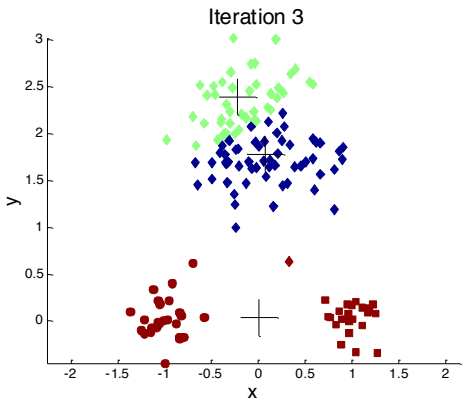


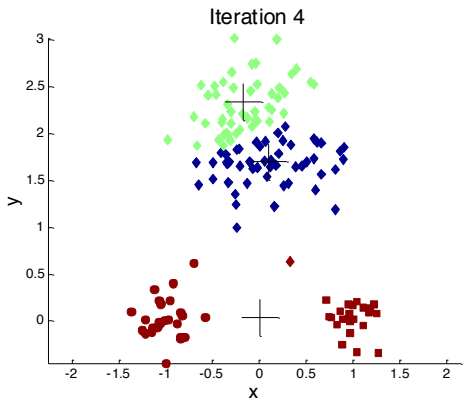
A less meaningful result

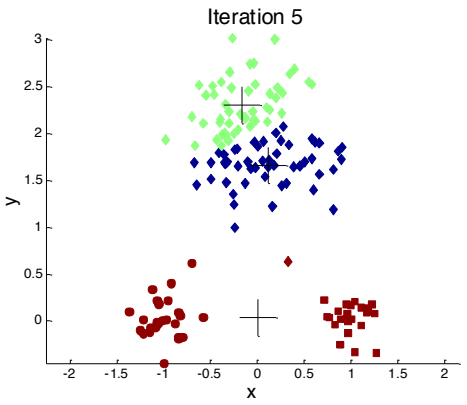
Importance of Choosing Initial Centroids











How to Choose Initial Centroids?

Multiple runs, changing random seeds every time

- Each random seed corresponds to one set of randomly chosen centroids.

Compare *SSE (Sum of Squared Errors)* for each run

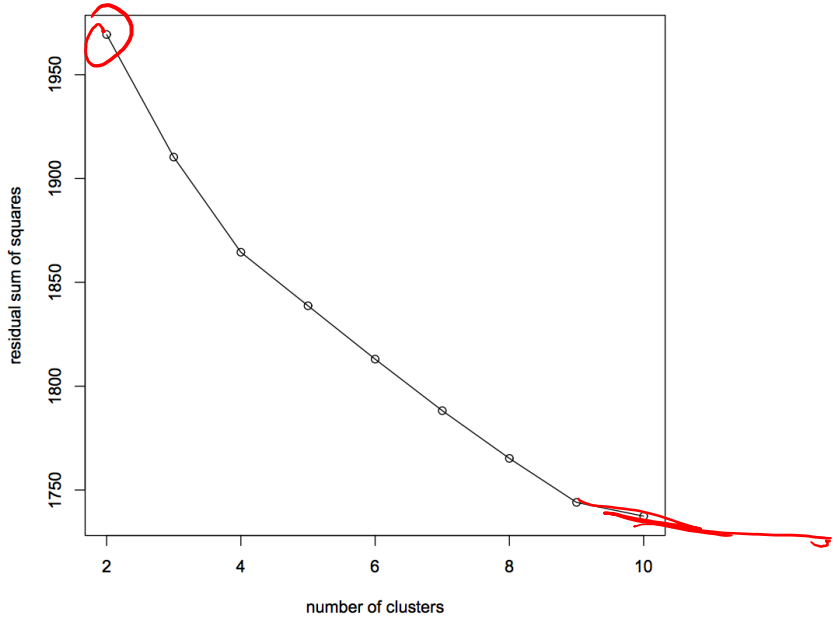
- x is a data point in cluster C_i and m_i is the centroid/medoid for cluster C_i.
- For each point, the error is the distance to the centroid/medoid.
- To get SSE, we square these errors and sum them:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- Compare the *SSE* for finding the best initial centroids when k (the number of clusters) is the same.

How to Choose K (Number of Clusters)?

The "elbow" method



► **Figure 16.8** Estimated minimal residual sum of squares as a function of the number of clusters in K-means. In this clustering of 1203 Reuters-RCV1 documents, there are two points where the $\widehat{\text{RSS}}_{\text{min}}$ curve flattens: at 4 clusters and at 9 clusters. The documents were selected from the categories *China*, *Germany*, *Russia* and *Sports*, so the $K = 4$ clustering is closest to the Reuters classification.

What If the Iteration Never Stops?

Set maximum number of iterations

Set minimum value of SSE change

Variations of the *K-Means* Method

One variation is the mixture models (soft clustering)

- Estimates clusters from probability distributions
- Includes the expectation maximization (EM) algorithm

Cluster Validity

For supervised classification we have a variety of measures to evaluate how good our model is.

- Accuracy, precision, recall

For cluster analysis, the analogous question is how to evaluate the "goodness" of the resulting clusters.

But "clusters are in the eye of the beholder"!

LDA in Theory and Applications

School of Information Studies
Syracuse University

Topic Modeling

Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes.

Why Topic Modeling?

Assume you have all the New York Times articles about the Middle East in the past 50 years. How can you find out what are the main concerns in Middle East? And how did the focus change over time?

A human expert would follow all articles and write a review, but it takes a lot of time.

A reader might just want a bird's-eye view of the major events that have been brought up in the past 50 years.

Topic modeling provides such a bird's-eye view.

Topic Modeling Algorithms

Latent Semantic Analysis (LSA)

- Landauer, T. K., & Dumais, S. (2008). Latent semantic analysis. *Scholarpedia*, *3*(11), 4356.

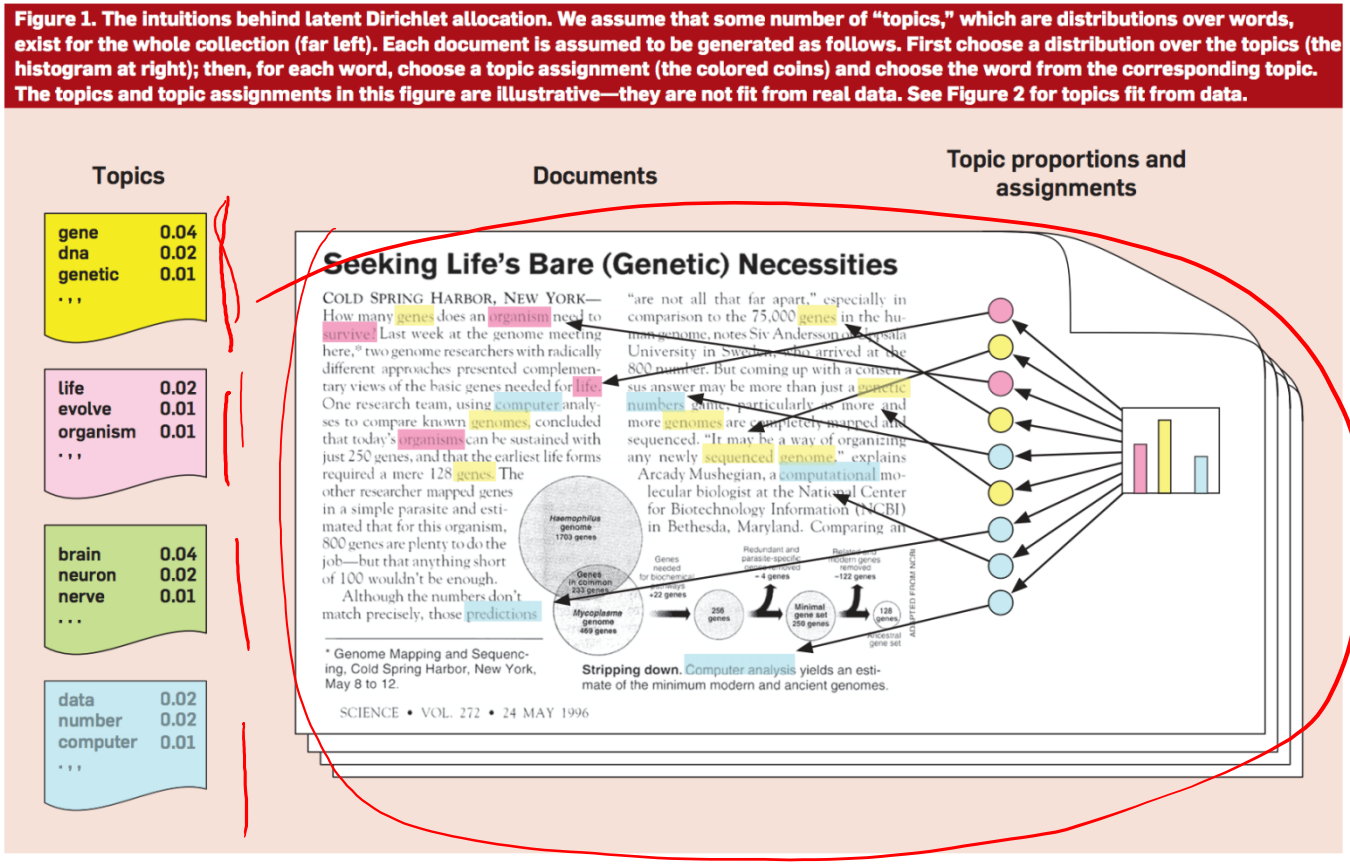
Probabilistic Latent Semantic Indexing (PLSI)

- Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 289–296)

Latent Dirichlit Allocation (LDA)

- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

The Intuition Behind LDA



Documents exhibit multiple topics

What Is a Topic?

A topic is defined as a distribution over a vocabulary.

For example, the genetics topic has words about genetics with high probability, and the evolutionary biology topic has words about evolutionary biology with high probability.

Topics as Distributions of Vocabulary

Each word has a probabilistic relevance to each topic.

	Vocabulary								
Topics	Gene	DNA	Genetic	Life	Evolve	Organism	Brain	Neural	Nerve
1	**0.04**	**0.02**	**0.01**	0.005	0.001	0.0001	0.000001	0.000001	0.00000001
2	0.001	0.001	0.00001	**0.02**	**0.01**	**0.01**	0.00001	0.00001	0.00001
3	0.001	0.001	0.0001	0.000001	0.000001	0.00001	**0.04**	**0.02**	**0.01**
…									

Fitting the Topic Model

Standard statistical techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of documents.

- The text collection that you are analyzing is the "result" of this generative process.
- Bayesian rule and other math tools are used to invert this generative process to find out the "hidden topics" that generated this text collection.

Topic Modeling

Input: a text collection

Assumption:

- A text collection is "generated" by N topics.
- Each doc is a mixture of the topics.
- Each topic is a distribution of word weights.

Method: a generative model like Latent Dirichlet Allocation (LDA)

The Generative Process

Step 1: Randomly choose a distribution over topics.

Step 2: For each word in the document:

a. Randomly choose a topic from the distribution over topics in Step 1.

b. Randomly choose a word from the corresponding distribution over the vocabulary.

An Example

To make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic.

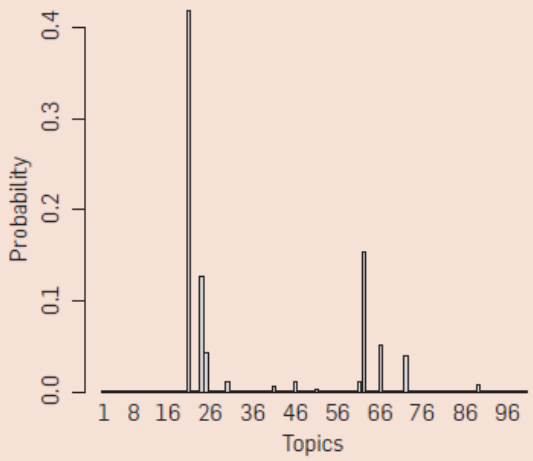
Example: Assume three topics: sports, health, university policy.

To "generate" a new document about university policy on student athletes' health, assume the content of the document to be 1/4 about sports, 1/4 about health, and 1/2 about policy.

To "generate" a word in the document, randomly choose a topic based on the above distribution, and then draw a word from that topic.

What Does a Topic Model Look Like?

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.

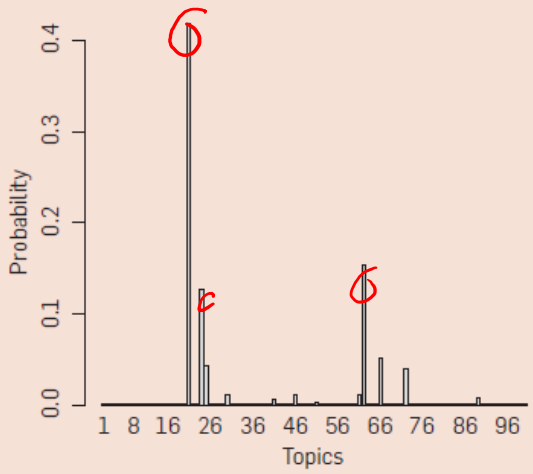


genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information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

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84.

What Does a Topic Model Look Like?

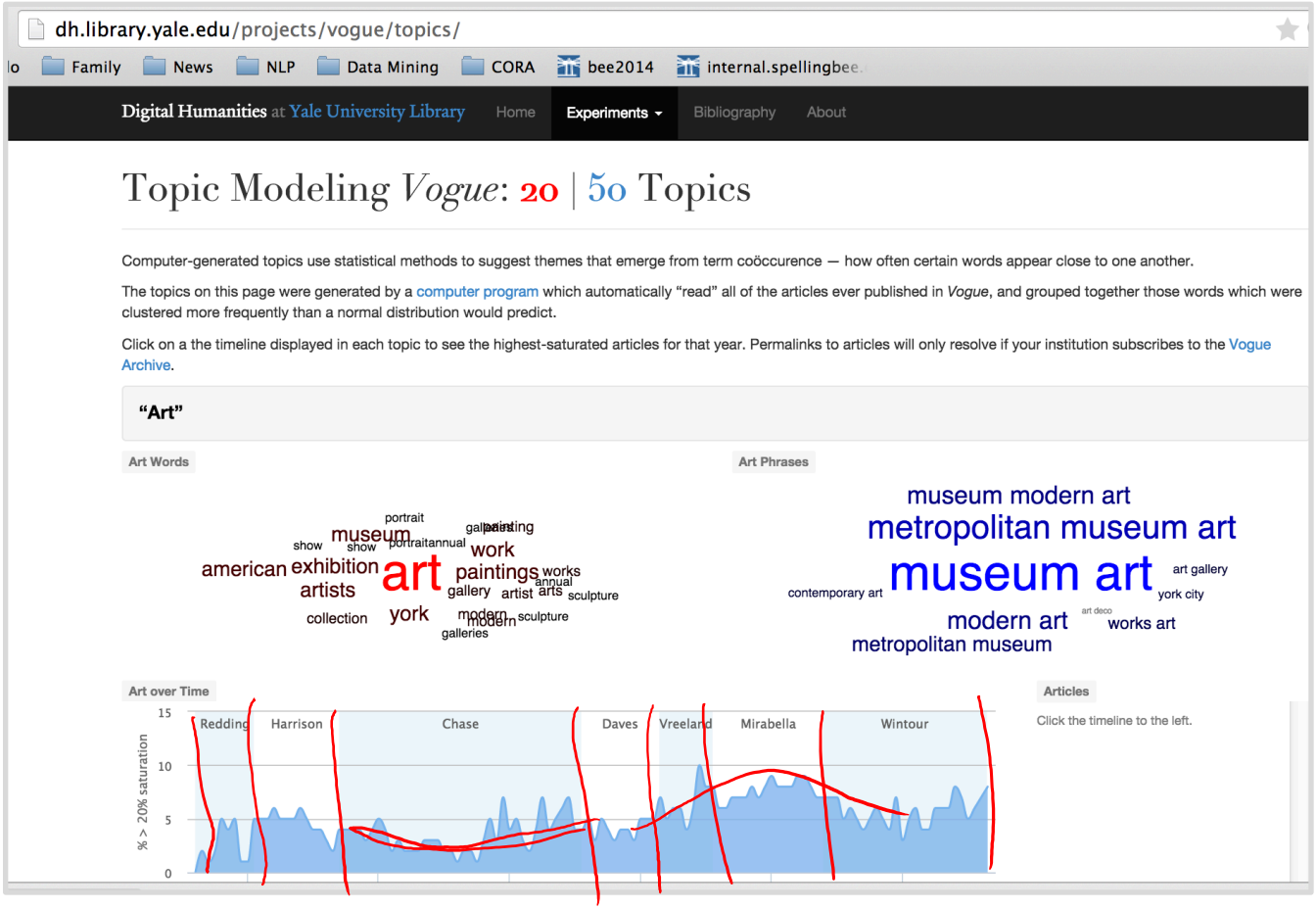
Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



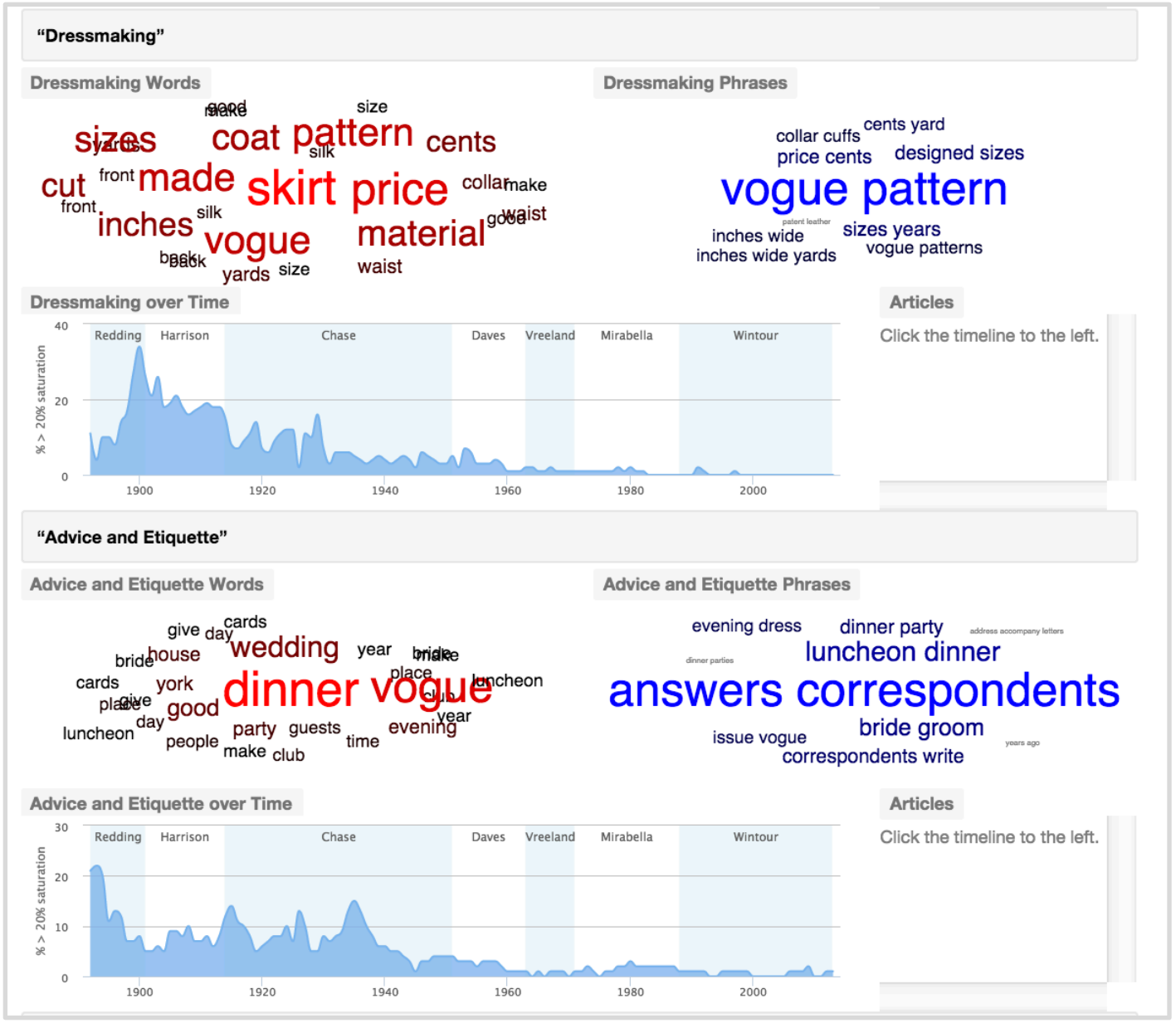
"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	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

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84.

Topic Trend Analysis



School of Information Studies
Syracuse University



School of Information Studies
Syracuse University