Semantic Analysis: Document Clustering

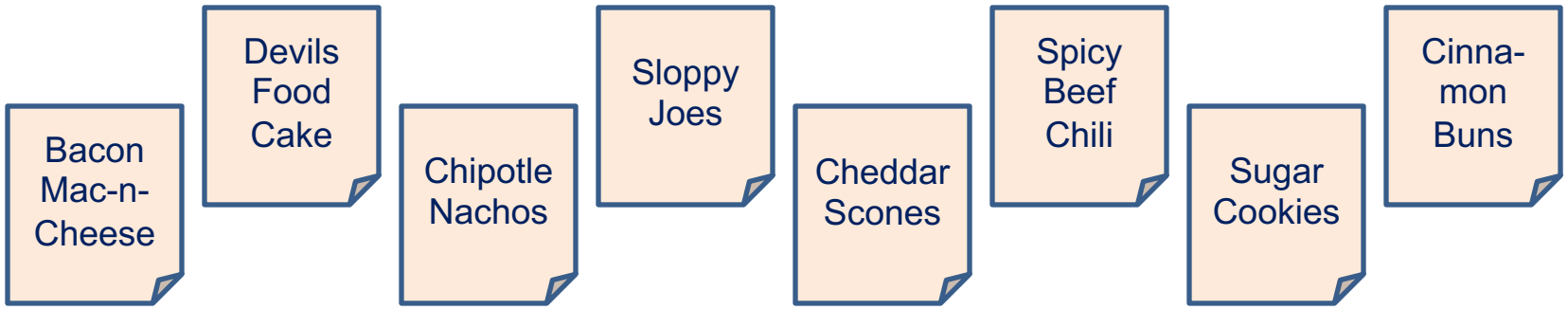
What Is Document Clustering?

In document clustering, we organize a set of documents into groups having similar characteristics.

An easy example to think about is a collection of recipes…

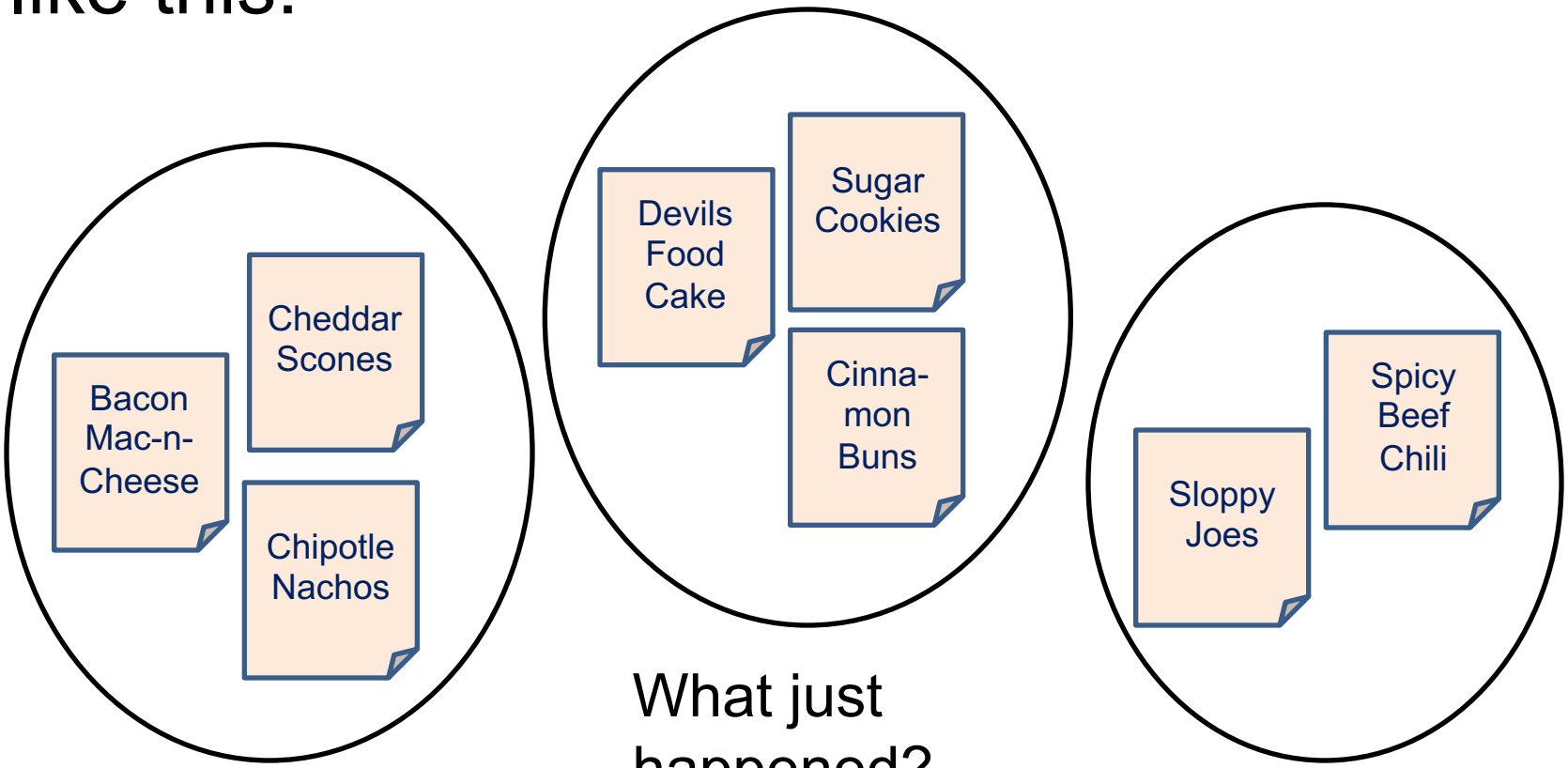
What Is Document Clustering?

When clustering documents, we start with something that looks like this:



What Is Document Clustering?

And we end up with something that looks like this:

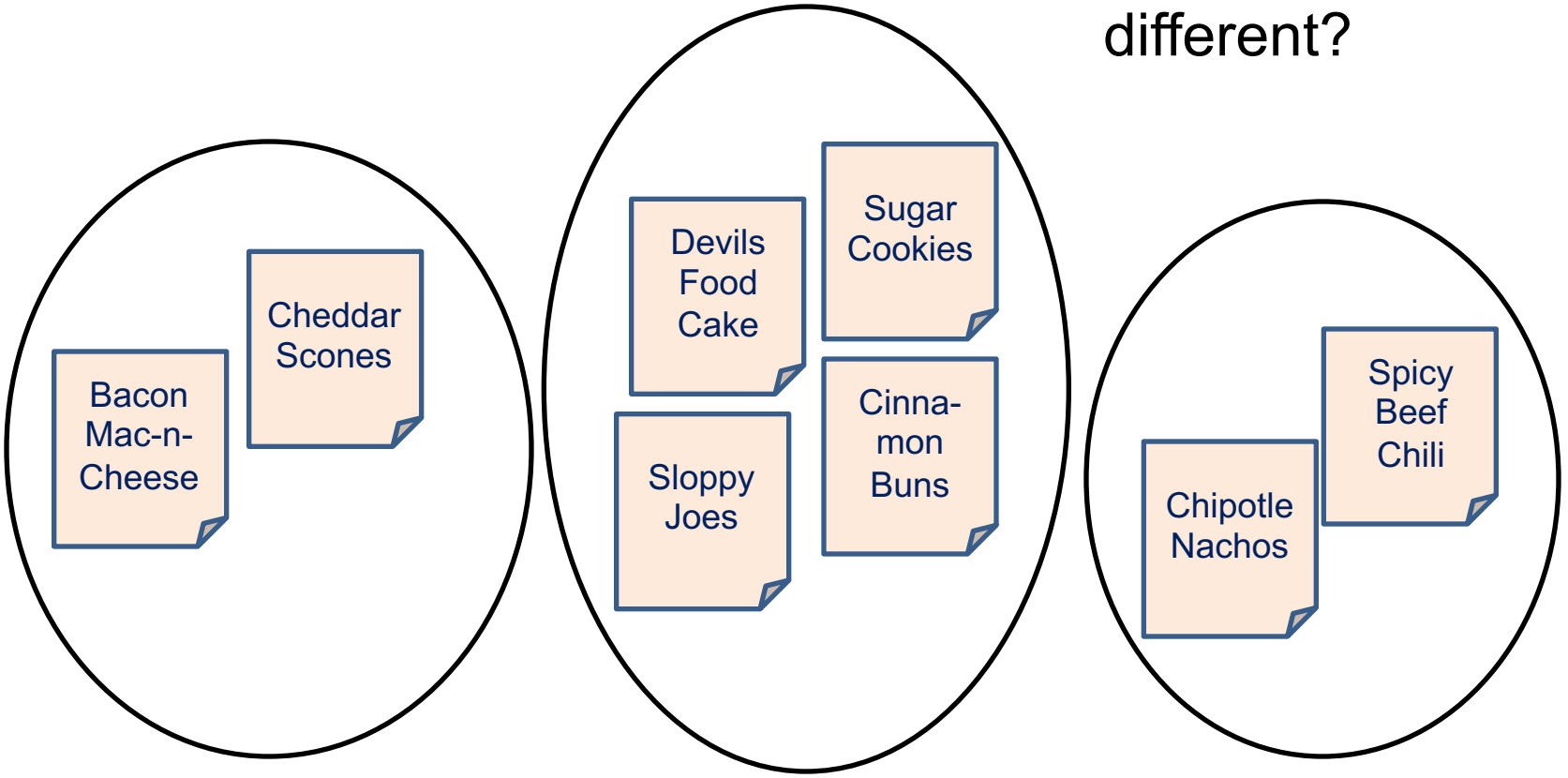


What just happened?

What Is Document Clustering?

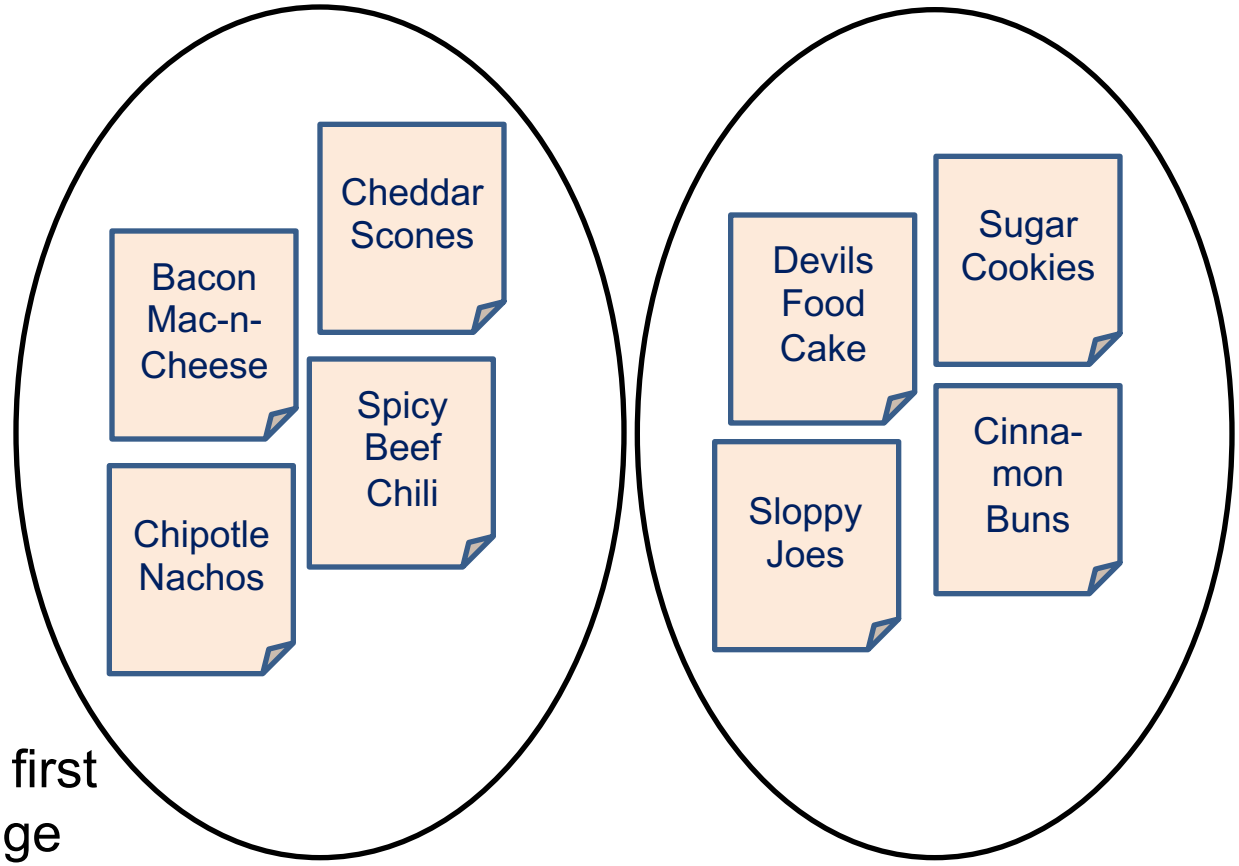
Or this:

How is this different?



What Is Document Clustering?

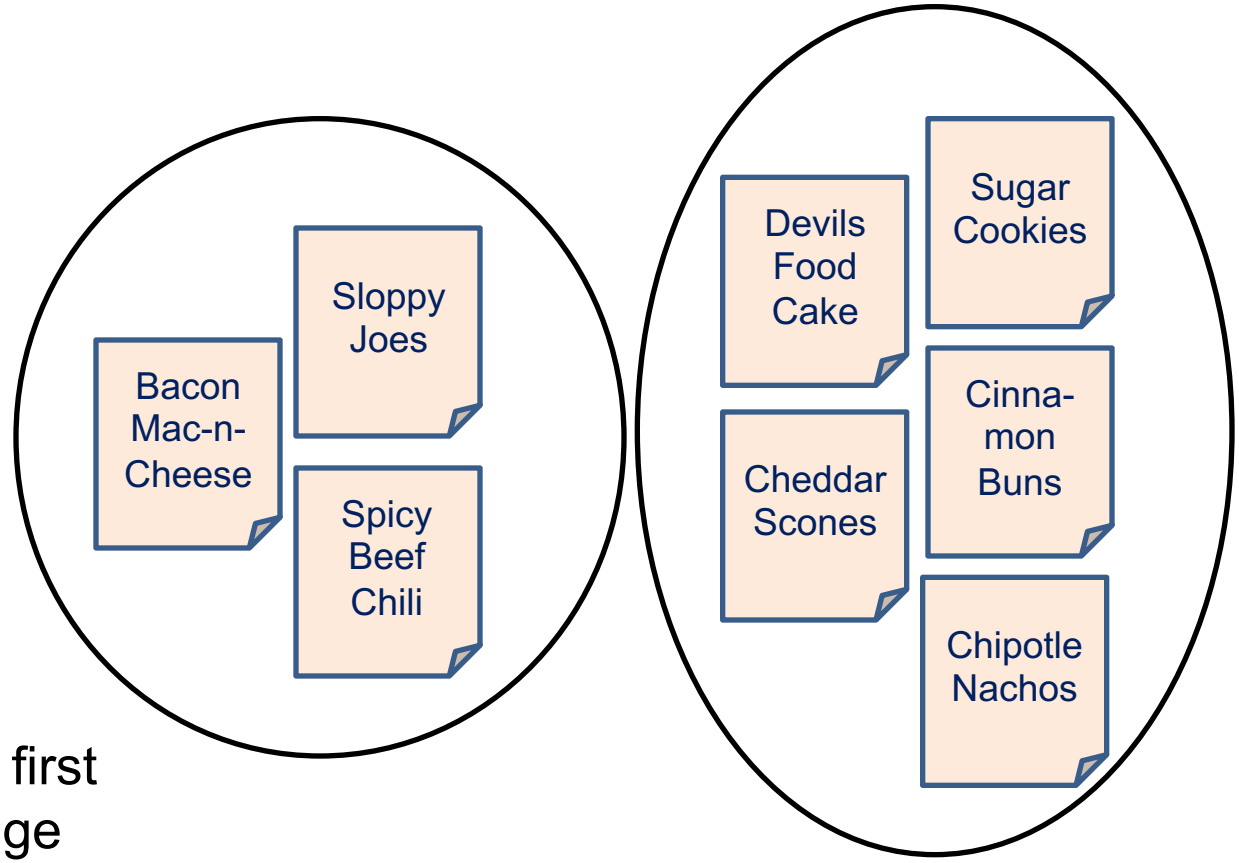
Or this:



How did the first cluster change characteristics?

What Is Document Clustering?

Or this:



How did the first cluster change again?

Clustering vs. Classification

Clustering is a very different approach to grouping documents, from classification.

	Characteristics of groupings	ML approach	Labelling of groupings
Classification	**Present in advance**	**Supervised**	**Prelabelled**
Clustering	**Not present in advance**	**Unsupervised**	**Not readily labelled**

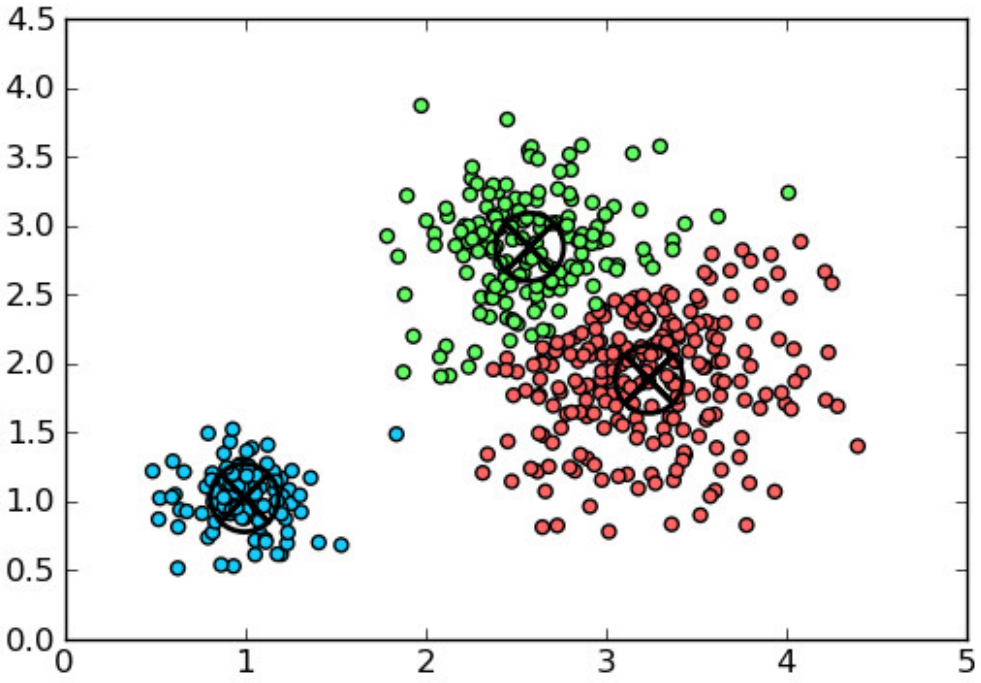
Methods of Clustering Documents

There are many types of clustering methods, but two of the most common are:

- Centroid-based methods
  - Each cluster has a central representative member
  - There's no hierarchy among the clusters
- Hierarchical methods
  - Smaller clusters are members of larger clusters, forming a hierarchy
  - Does not require having central members

Centroid Clustering

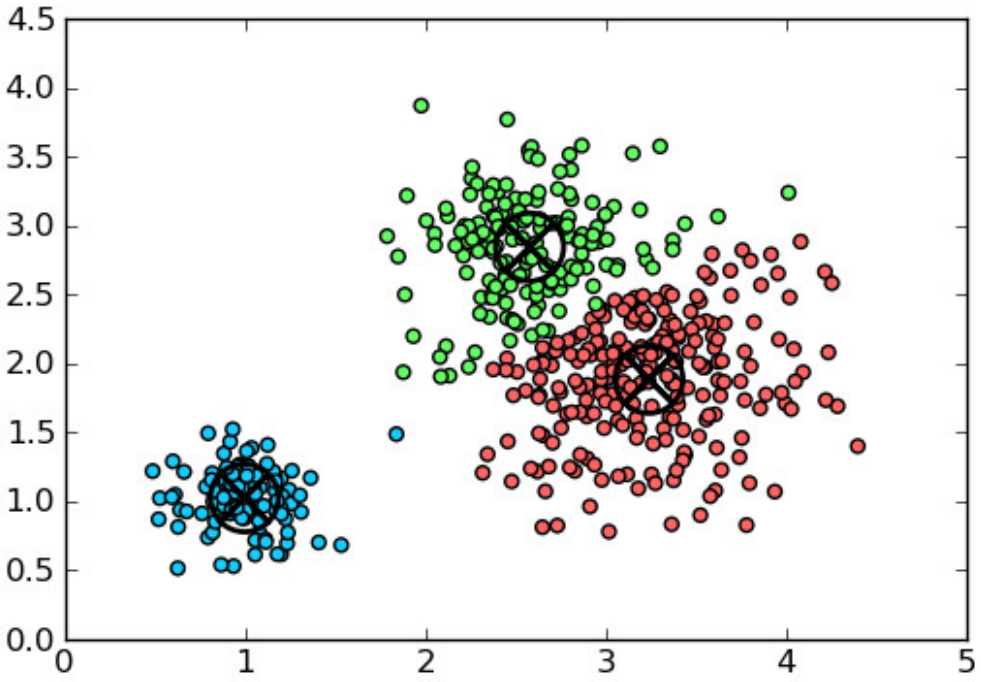
If the following represents a vector space for numerous documents, then *centroid-based clustering* is the grouping of documents as indicated by the colors.



Graph from Pacula, 2011

Centroid Clustering

Each ⊗ is a *centroid*, i.e., the point within each cluster that is considered the most representative of that group.



Centroids are typically at the mean or median value of the data in a cluster.

DataScience@SMU

Semantic Analysis:
Working with Clusters

Working with Clusters

- So we've generated some clusters. Now what? How do we know if they're good? How can we understand what they "mean"?
- We need to be able to visualize and label clusters, and leverage them in applications.

Visualizing Clusters

- In our examples we used simplified, two-dimensional graphs to understand the basics of clustering.
- But in a real project, there is a many-dimensional vector space—so visualization is nontrivial.
- We first have to perform *dimensionality reduction*, so that we can do a second plot.

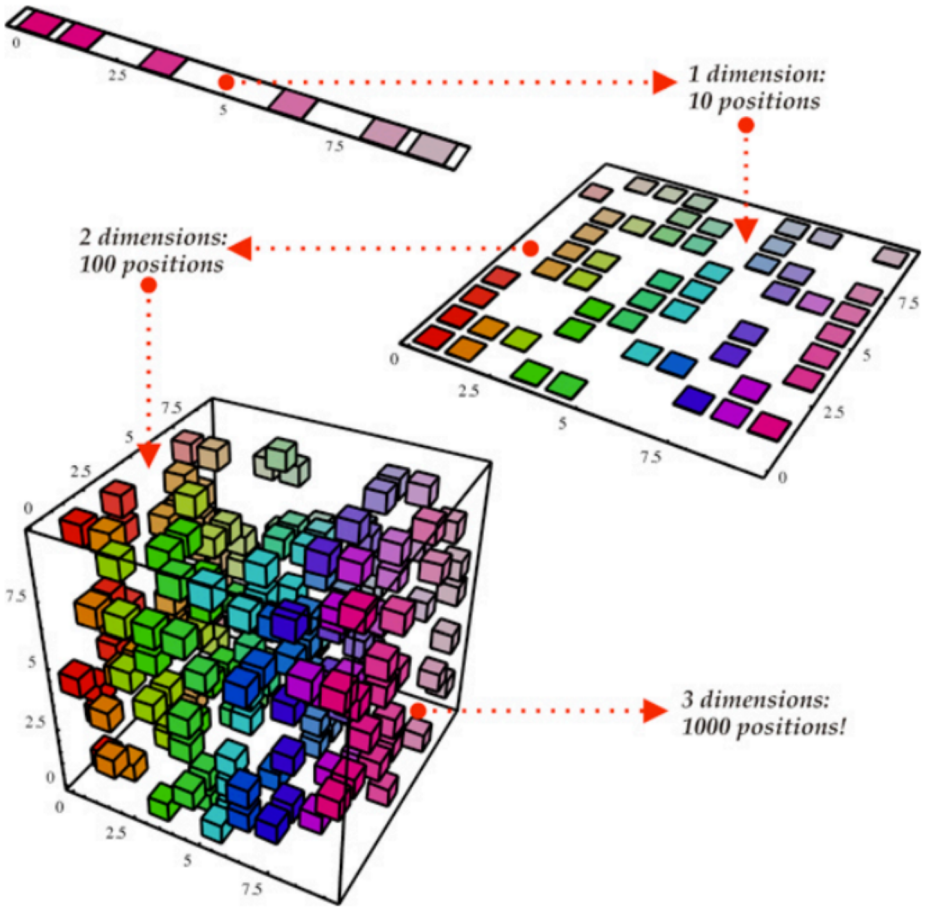
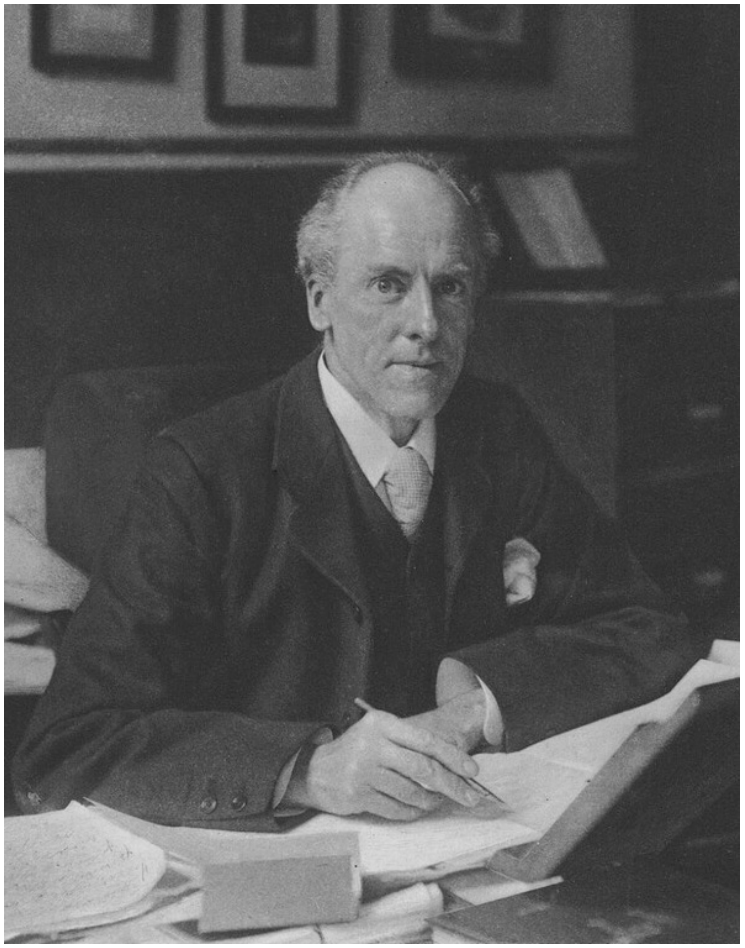


Image from Bigsnarf blog 2013

Principal Component Analysis

- PCA was introduced by Pearson in 1901.
- It was modified over the years until it reached its now-prevalent form with Michel Loève in 1963.

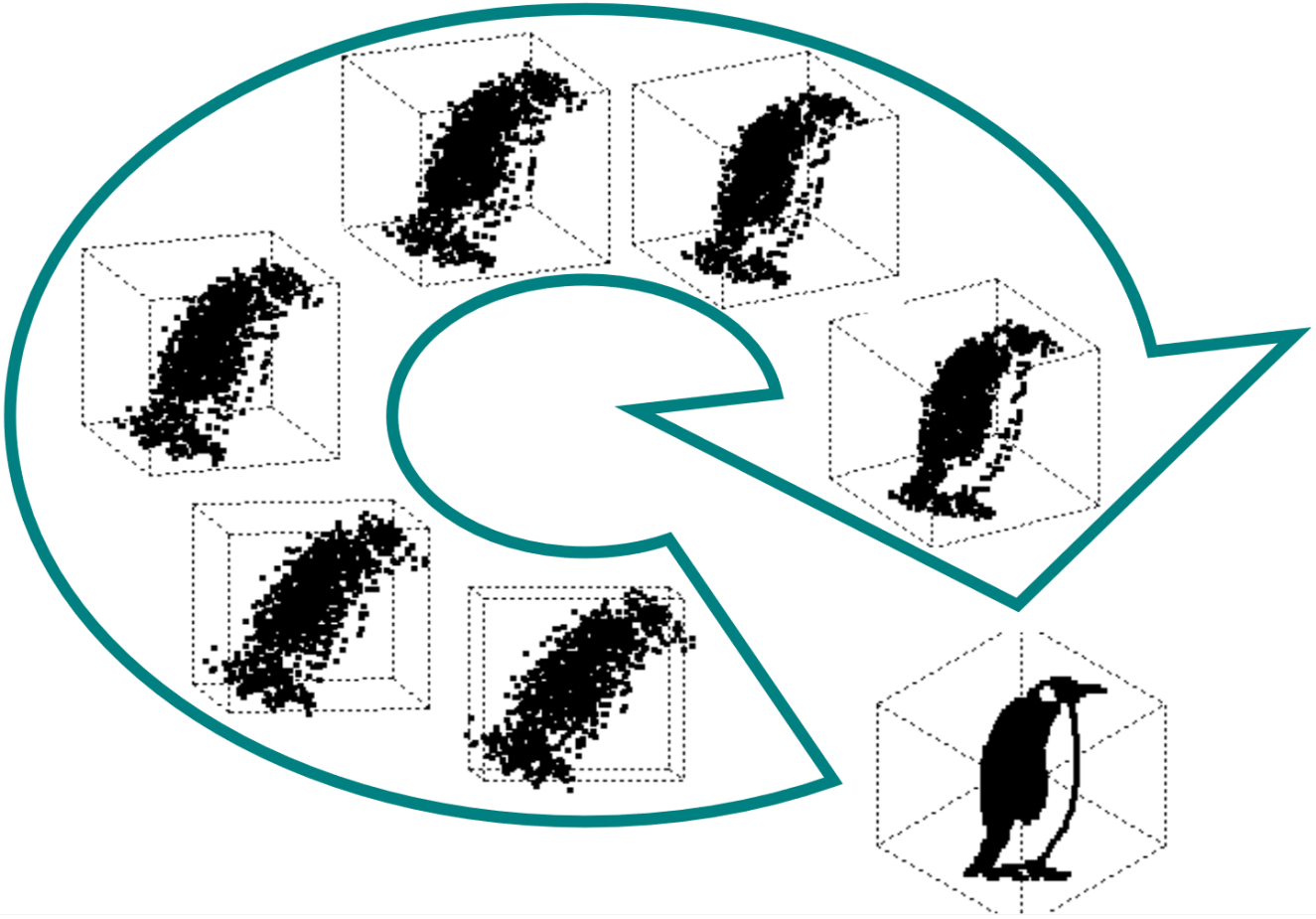
Michel Loève



Karl Pearson

PCA

See below how a 3D plot is rotated in various ways—notice that one of the rotations shows more variance (the dot patterns don't blur together as much as in the other rotations).



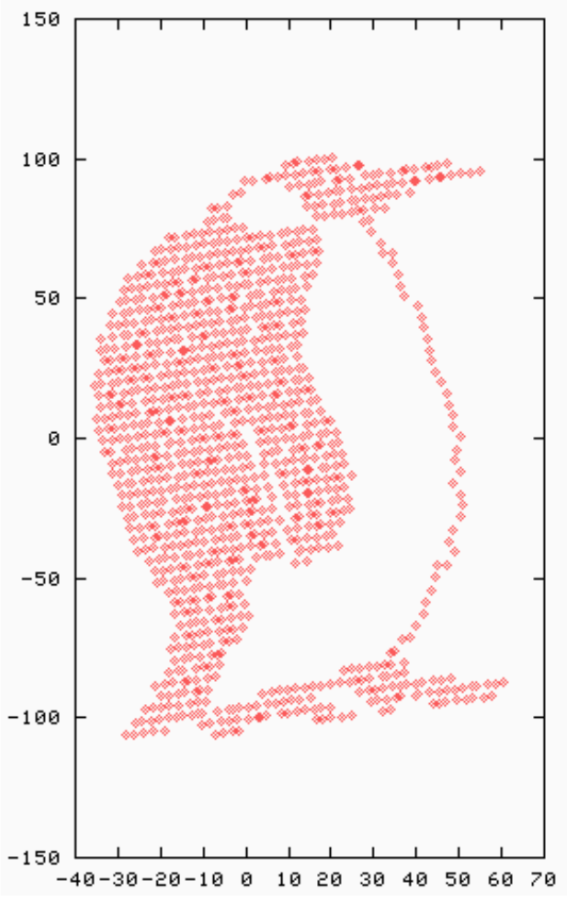
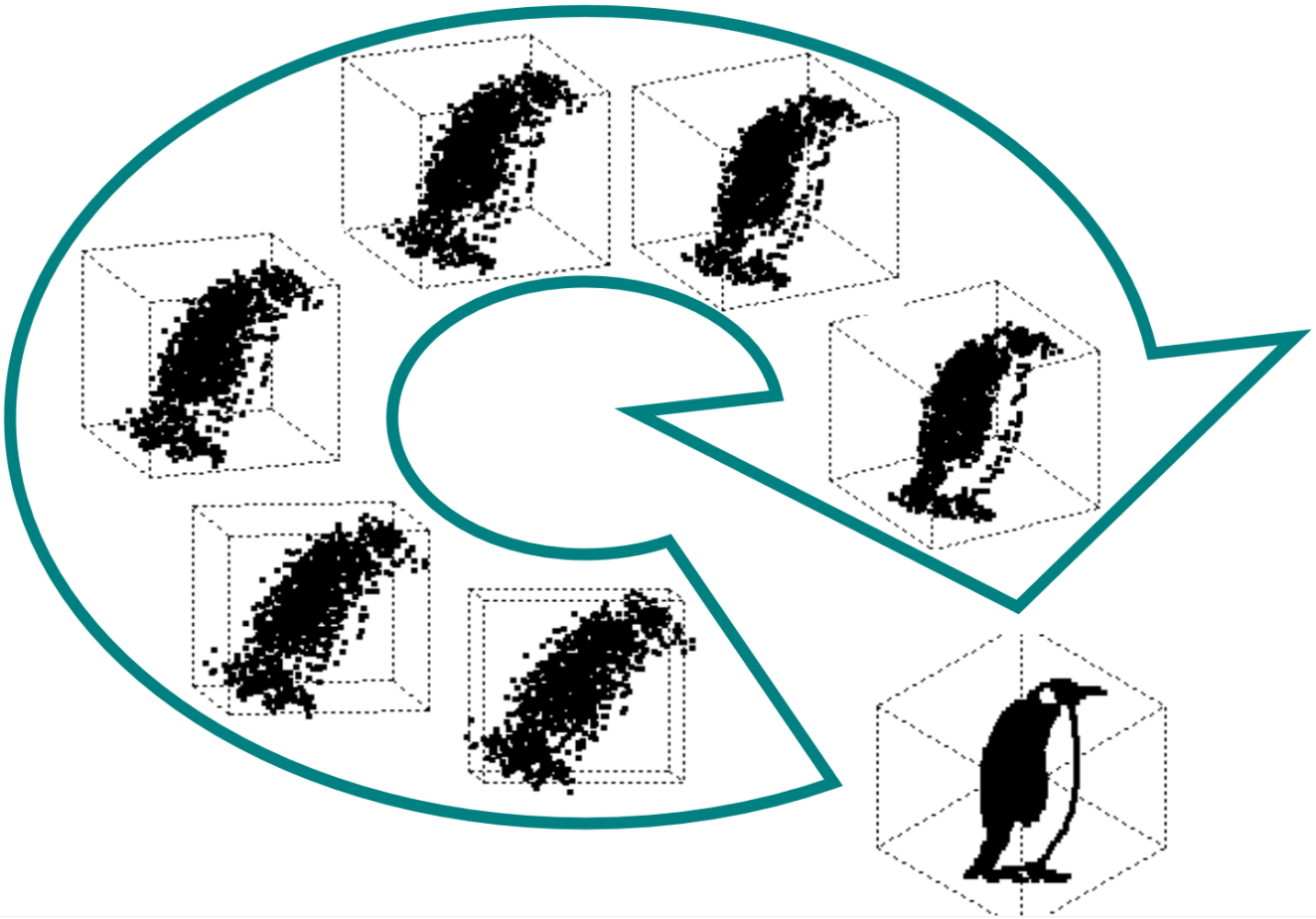
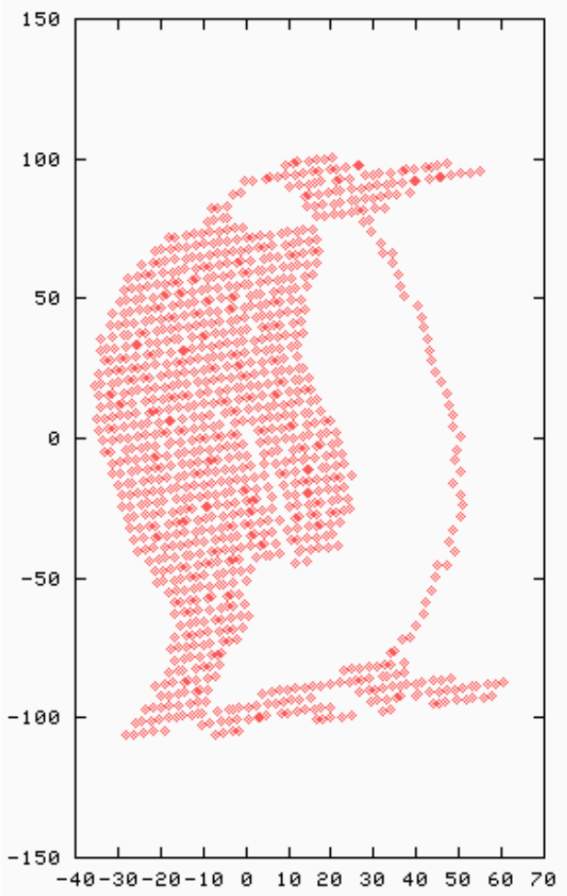


Image from Gutierrez-Osuna 2002

PCA

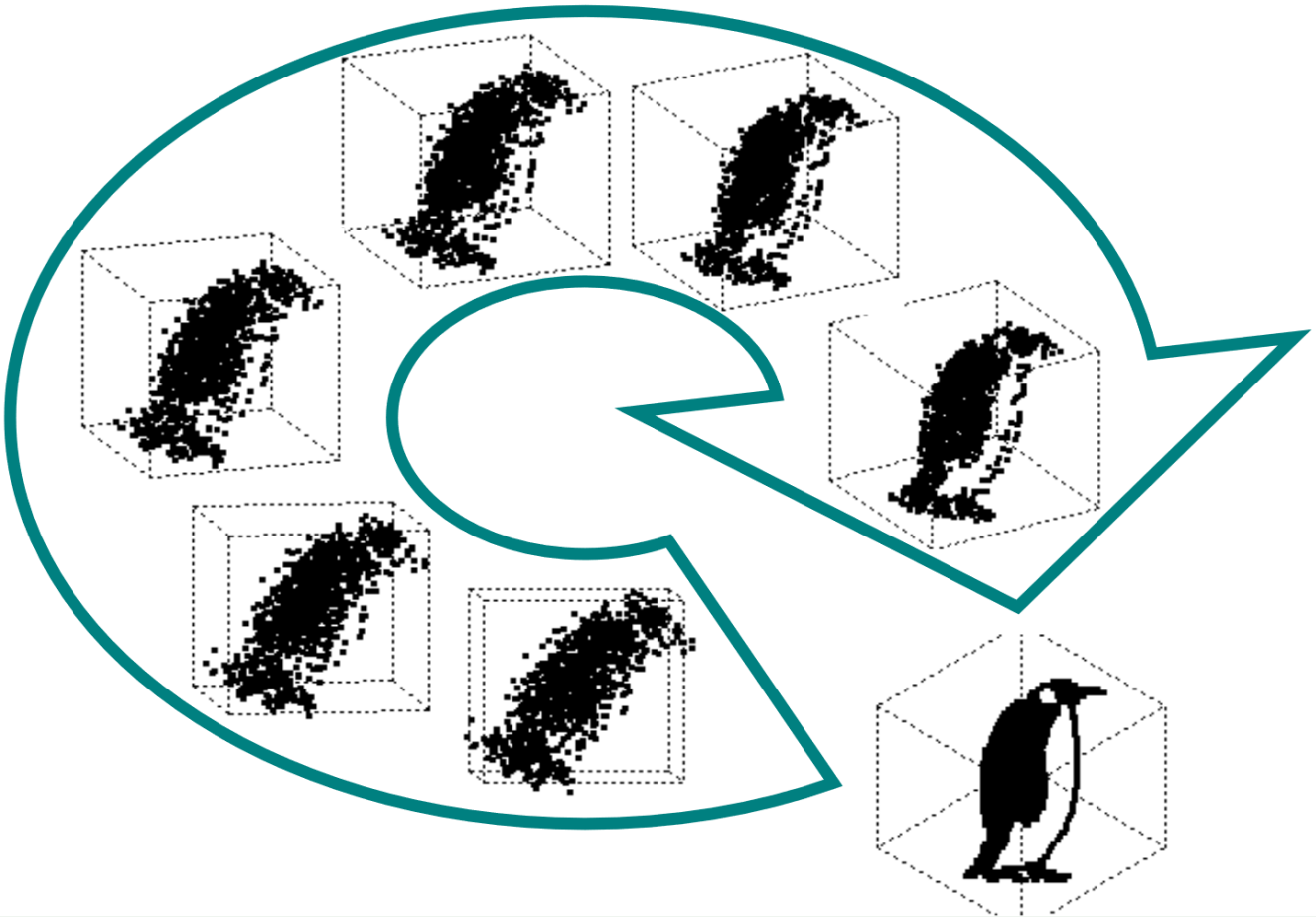
PCA finds *the rotation that is best* for projecting the 3D plot onto a merely 2D plot.

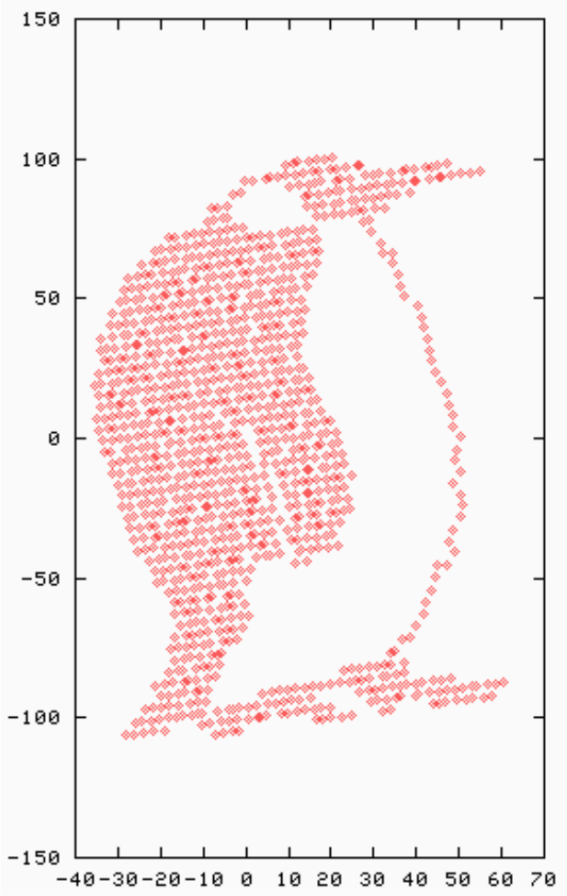




PCA

PCA can do this even if there are 4, 5, 6, etc. dimensions—lucky for us!





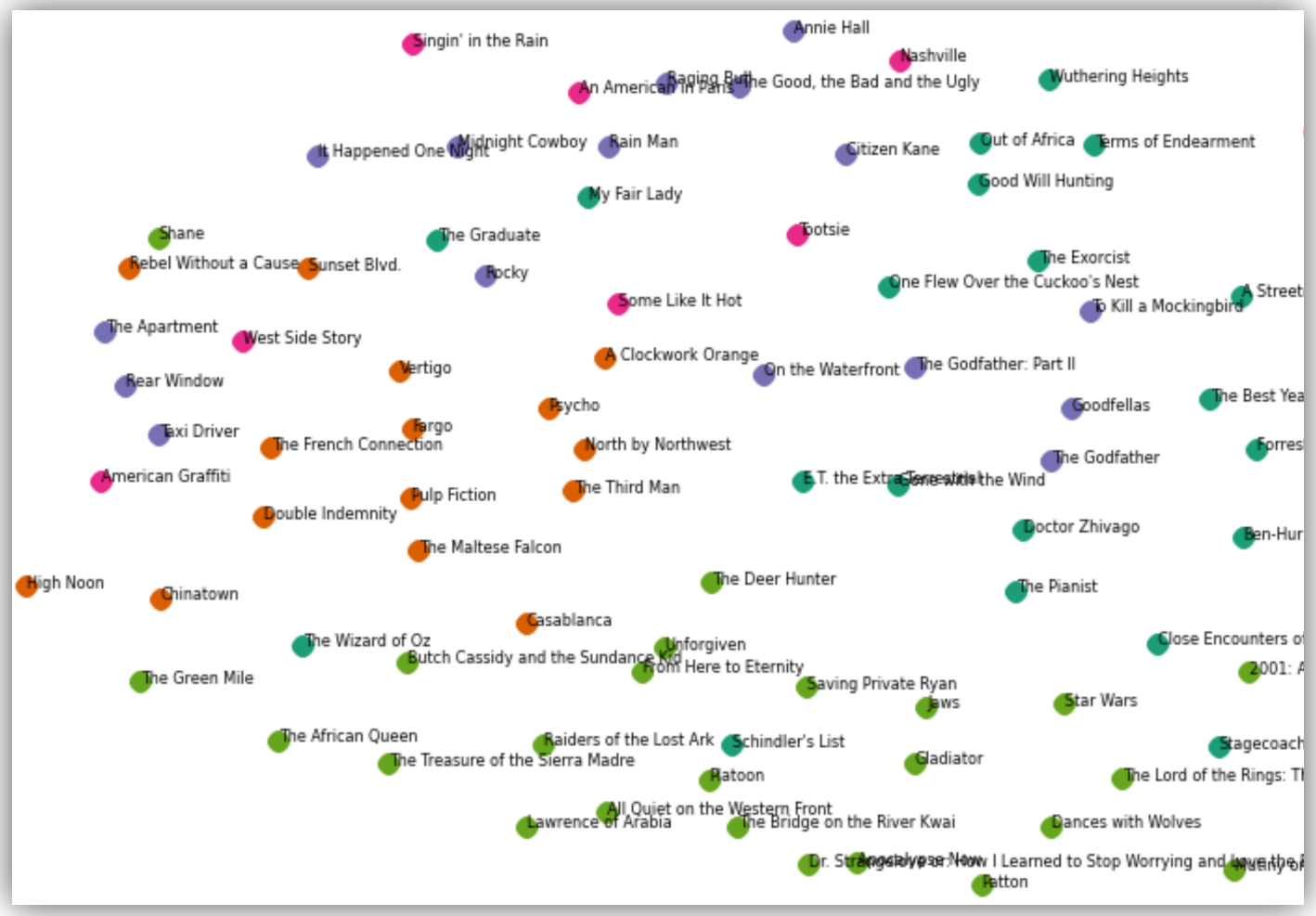
Visualizing Clusters

- It's not hard to run PCA on our data from within Python:

```
1 from sklearn.decomposition import PCA
2 pca = PCA(n_components=2).fit(myclusters.data)
3 pca_2d = pca.transform(myclusters.data)
```

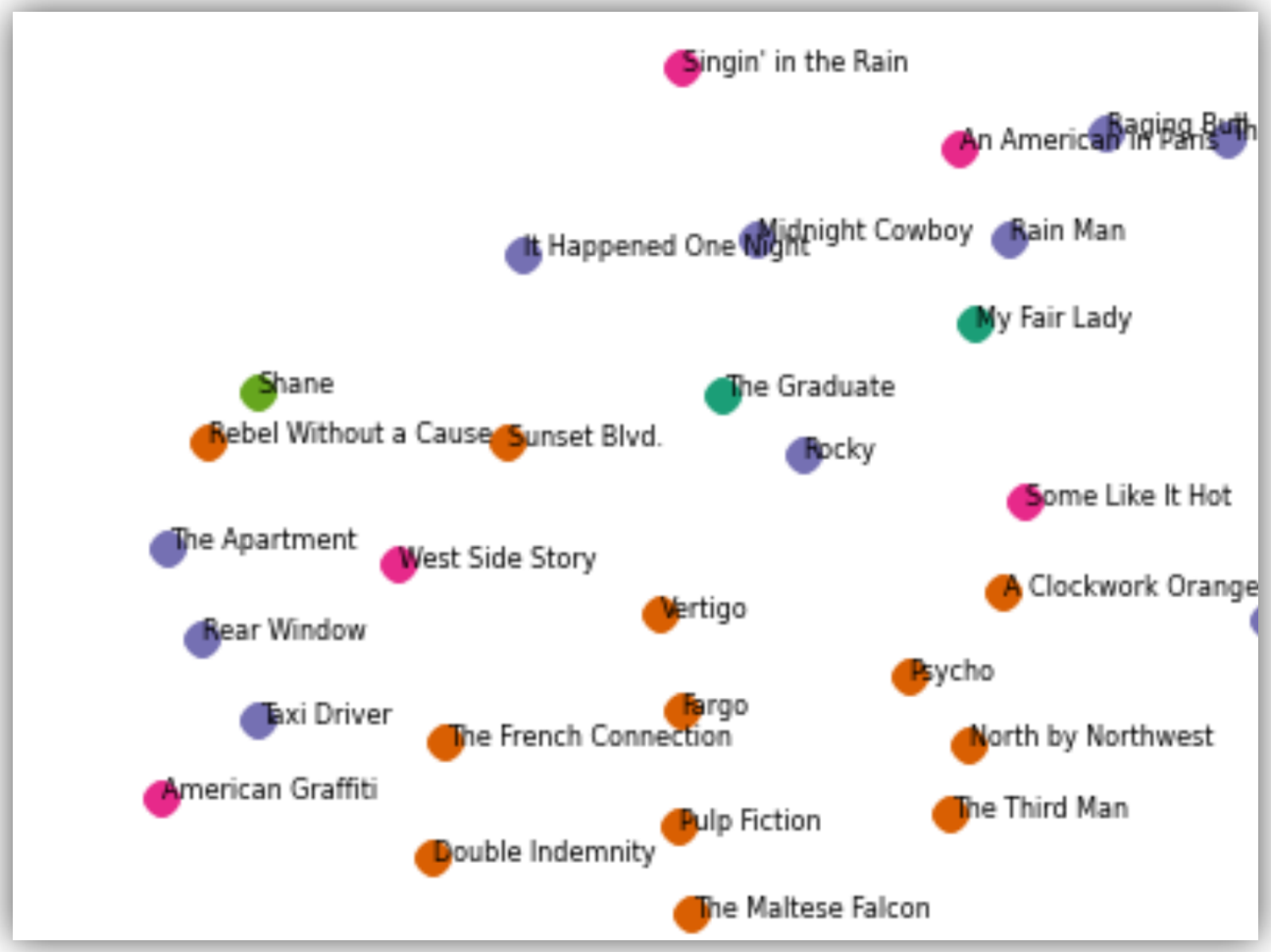
2D Plots of Clusters

Once we've reduced dimensions down to two, we can create plots like this:



2D Plots of Clusters

Each color = movie descriptions are in same cluster.



DataScience@SMU

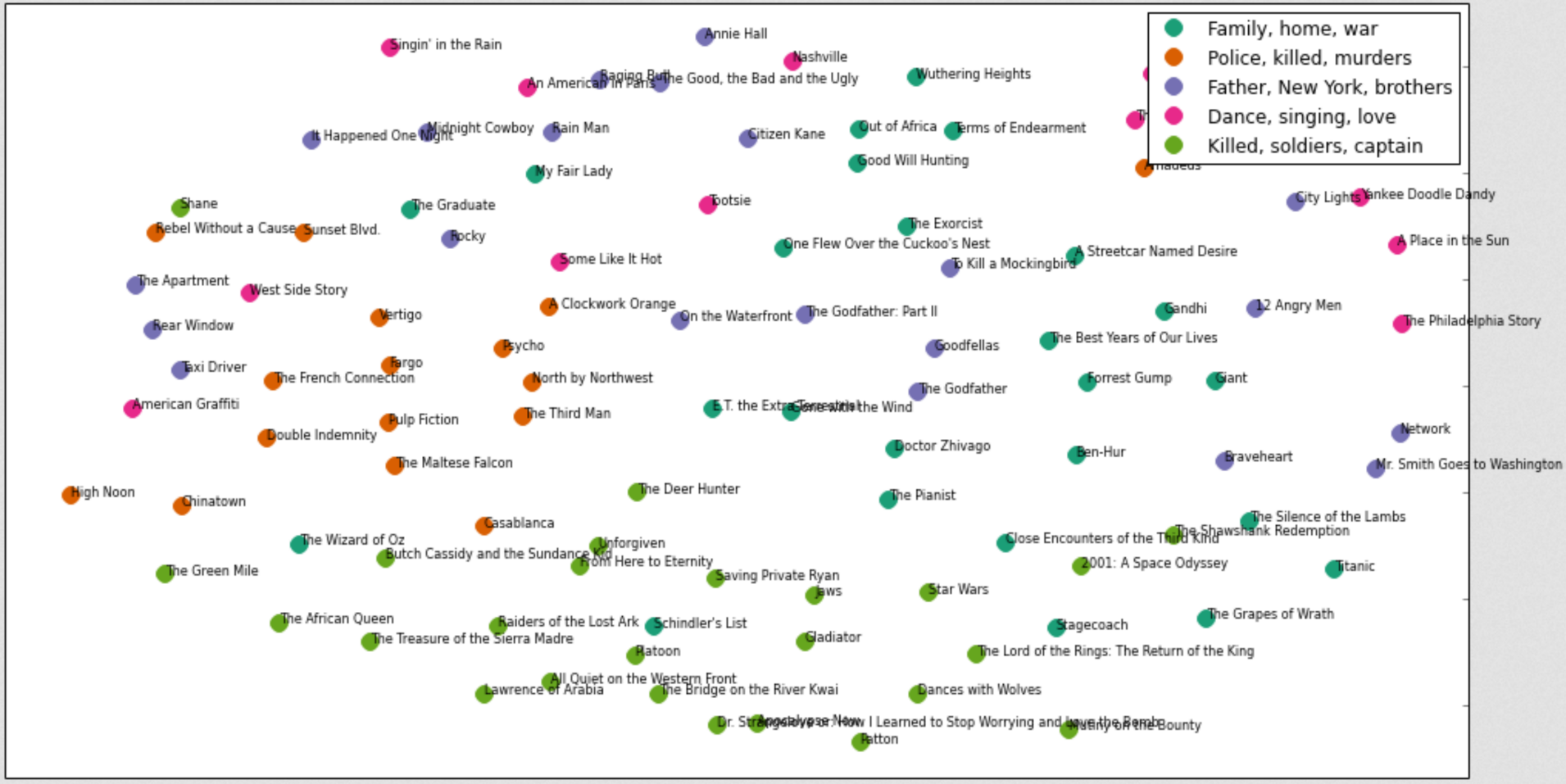
Semantic Analysis:
Working with Clusters

Labelling Clusters

- Clusters don’t come with ready-made names or titles.
- The clusters were made from long vectors, wherein any number of features could characterize a vector.
- The most straightforward approach to labelling clusters is feature extraction—taking the *top n features.*

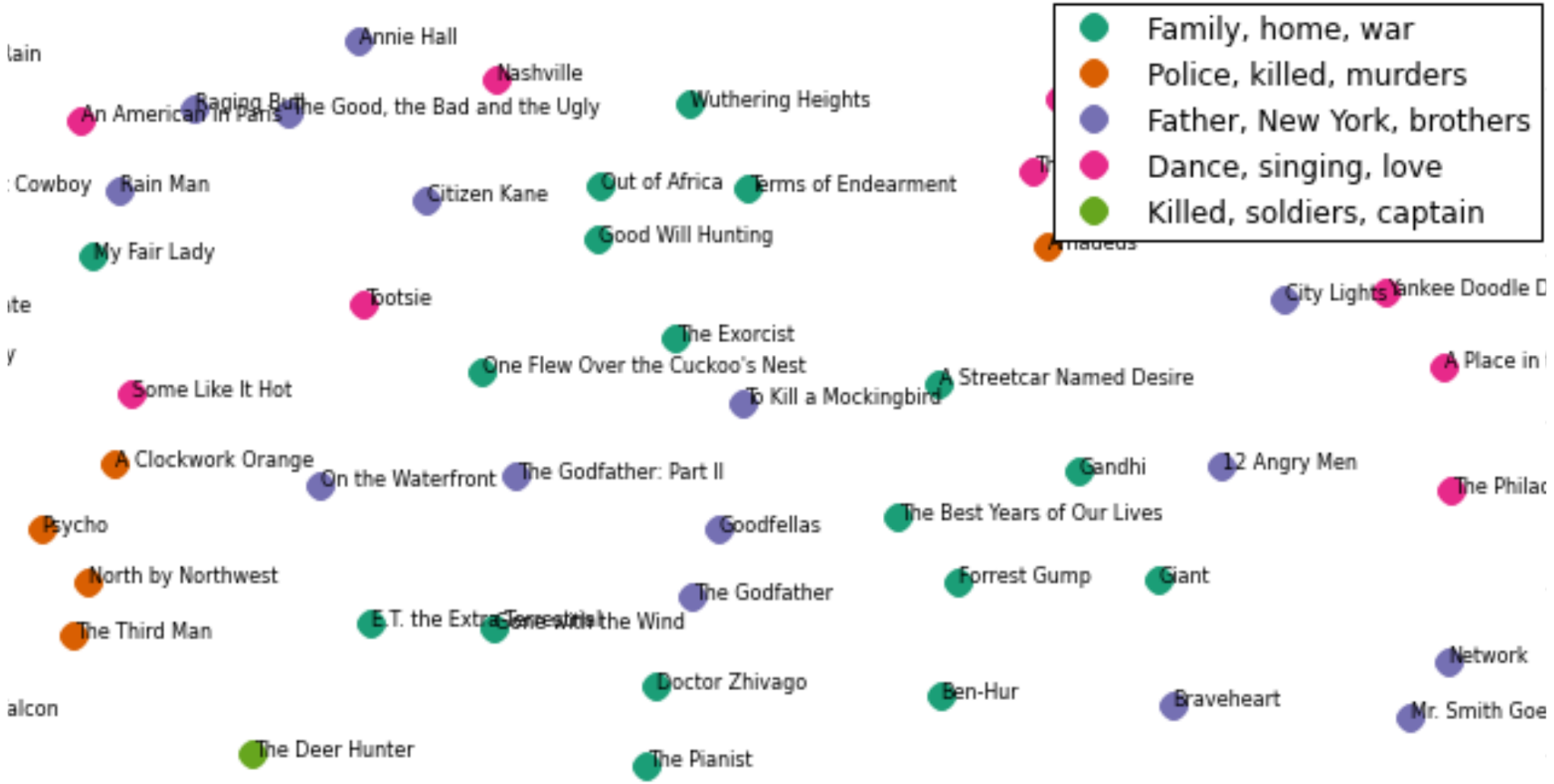
Labelling Clusters

- That turns it into this:



Labelling Clusters

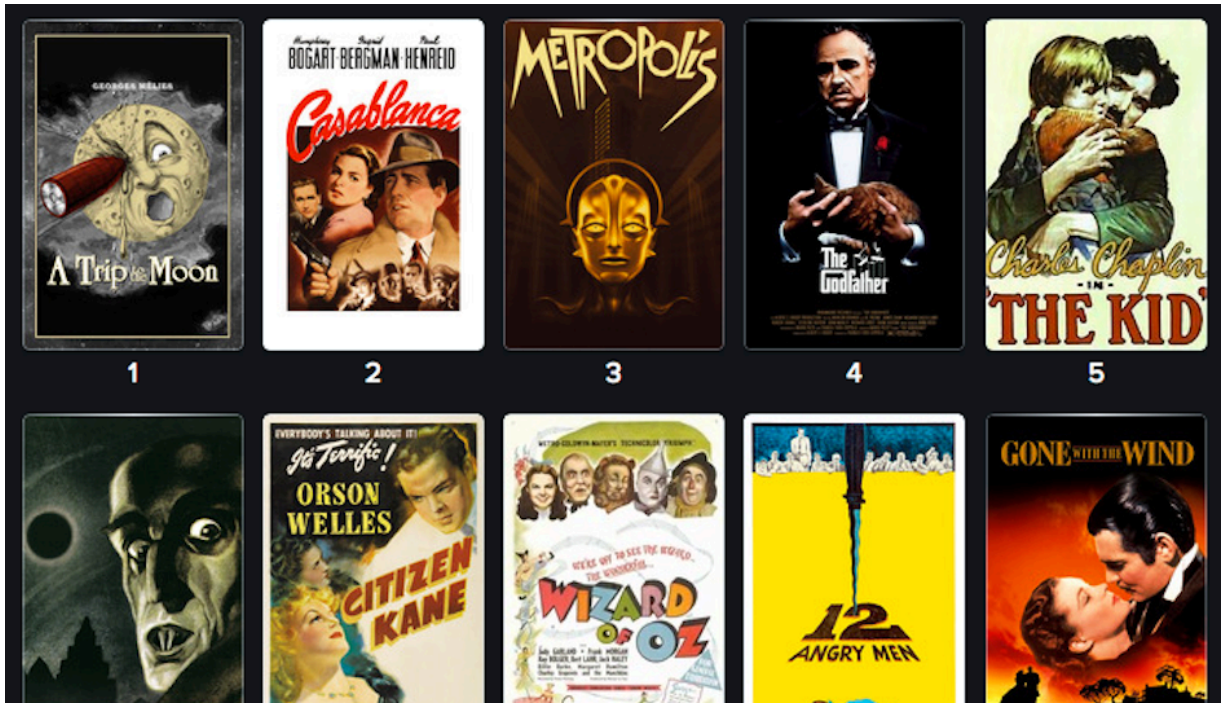
- The labels are actually helpful!



Applications of Clustering

One popular application is cohort profiling, i.e., explain how the top (or bottom) *n* [cell phones, TV shows, news articles] fall into clusters.

- We just saw this done with a list of the “best” movies.



Applications of Clustering

Another application is work-flow management—how to route *m* items to *n* reviewers, where similarity of items will be conducive to greater efficiency by a reviewer.

- Interviewing job candidates or grant applications

Applications of Clustering in IR ("Information Retrieval")

- Search result clustering
  - More effective search engine results page
- Collection clustering
  - Alternative to search—browse/explore
- Cluster-based retrieval
  - Exemplar based:
    "more like this" or "more like these"

Search Result Clustering

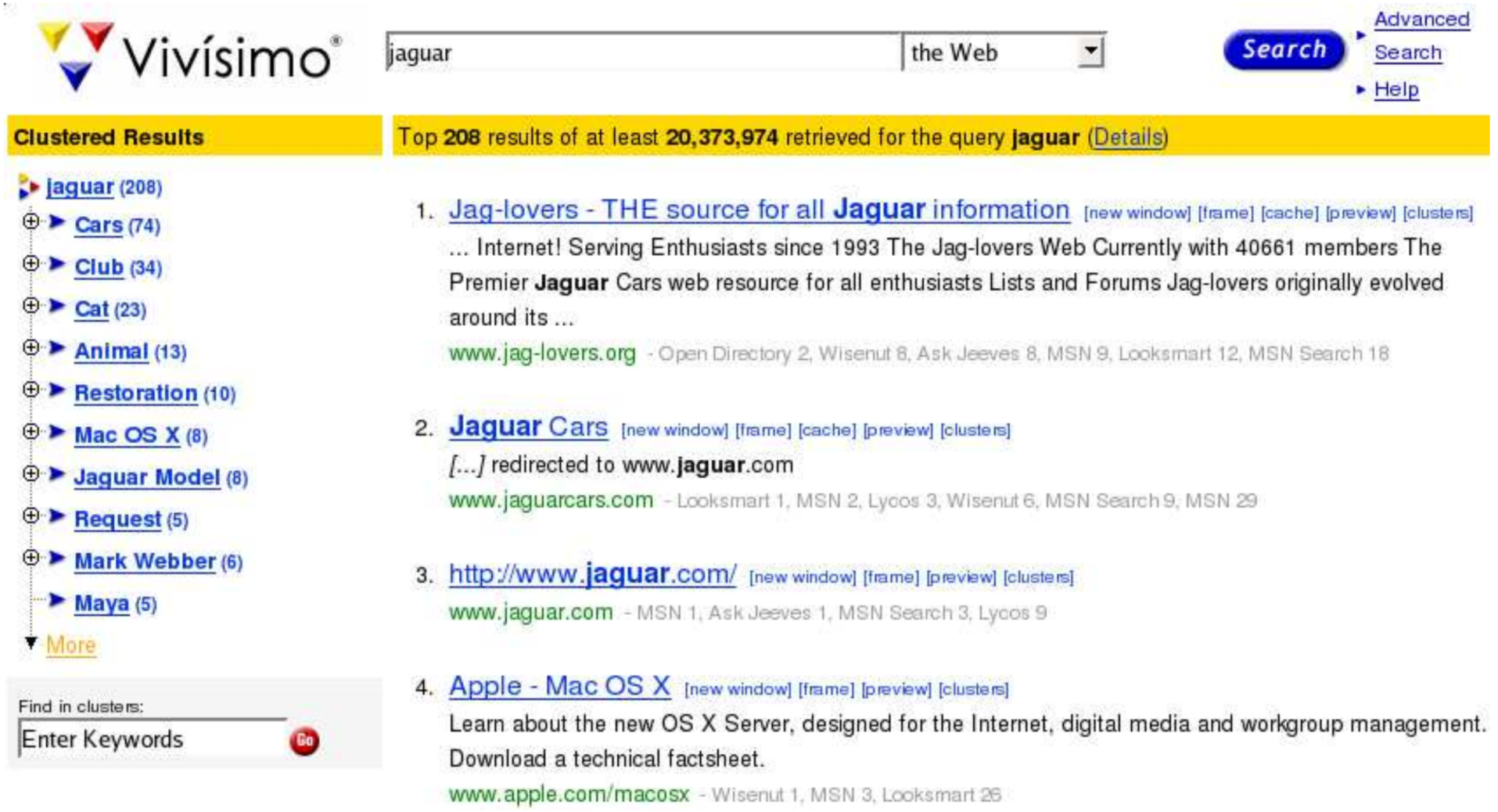


Image from Manning, Raghavan, Schuetze 2009. Vivisimo was acquired by IBM in 2012.

DataScience@SMU