Exploring EDA, Clustering and Data Preprocessing
*Lecture **2***

Taking Raw Data Towards Analysis

Vincent Croft

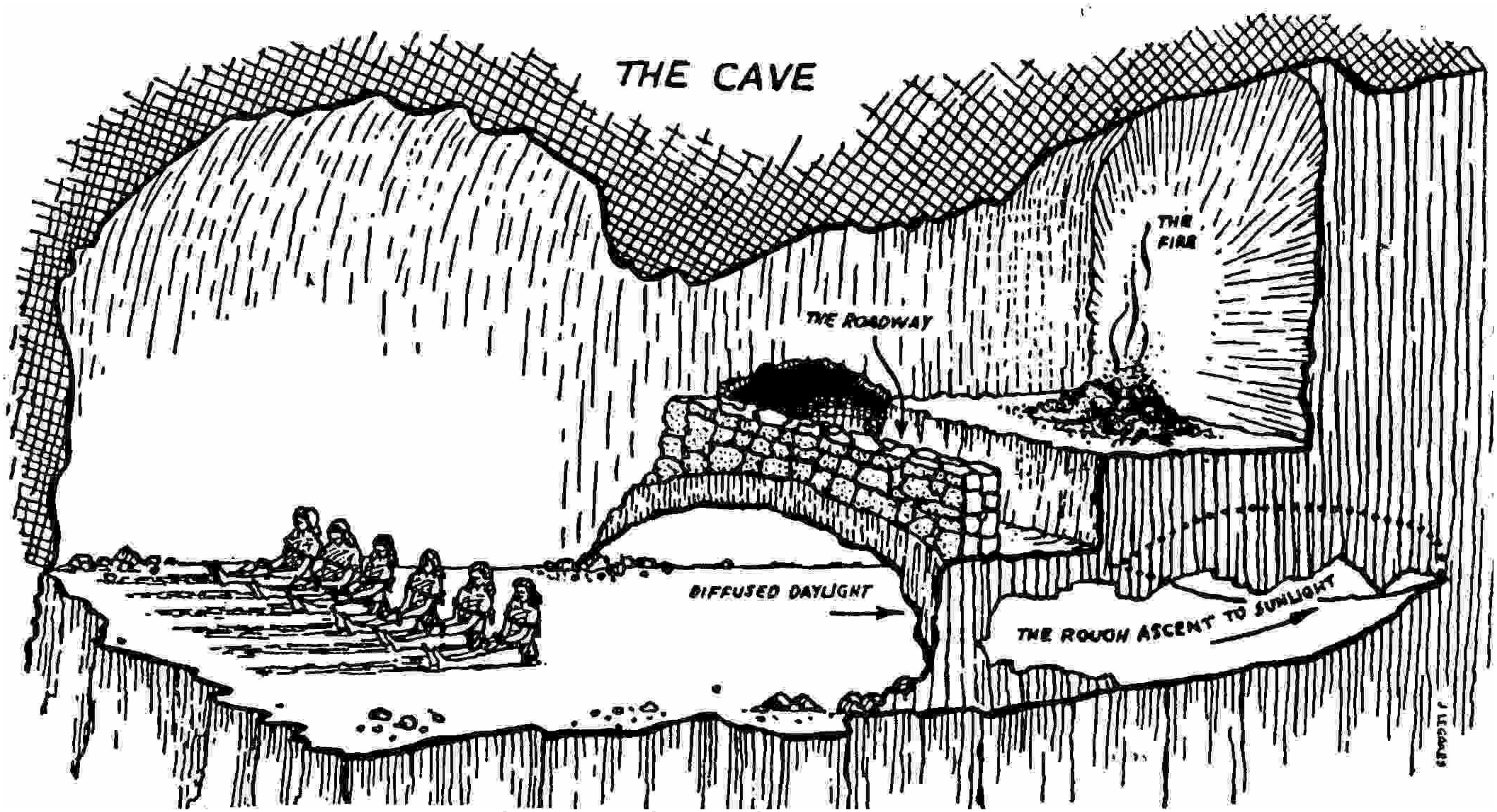
NIKHEF - Nijmegen

Inverted CERN School of Computing, 23-24 February 2015

CERN School of Computing

The path towards the sunlight…

- Our eyes see hundreds of colours, our ears hear thousands of frequencies, our user logs thousands of alphanumeric values… How do we keep ourselves from being overwhelmed.



Outline

- **Mapping**
- **Clustering**
- **Data Reduction**

- **Higher focus on examples**
- **Using real data from internet**
- **Brief introduction to scalable data analysis on big data**

Worked Examples

- All examples will be available online
- If you are not here in person or want to see the examples presented for yourself please see the support documentation on my institute web page.

Vince Croft - Home Page

http://www.nikhef.nl/~vcroft/

http://www.nikhef.nl/~vcroft/exploringEDA.pdf

http://www.nikhef.nl/~vcroft/takingRawDataTowardsAnalysis.pdf

CERN School of Computing

Mapping – Heat Maps

- **One last page in R**

Fund Category	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013 YTD
Total Equity Funds	4%	3%	3%	1%	-3%	2%	2%	-1%	1%	3.4%
Total Developed Market Equity Funds	4%	2%	2%	-1%	-3%	-1%	0%	0%	0%	3.8%
International Equity Funds	8%	6%	7%	6%	-4%	4%	1%	1%	1%	3.8%
US Equity Funds	1%	-1%	-1%	0%	0%	-4%	0%	0%	-1%	3.5%
Western Europe Equity Funds	1%	-1%	7%	-13%	-12%	1%	-3%	-2%	-2%	0.4%
Japan Equity Funds	52%	44%	0%	-27%	-18%	-19%	-3%	5%	10%	24.7%
Pacific Equity Funds	7%	-3%	12%	-1%	-16%	17%	8%	-8%	1%	7.9%
Total Emerging Market Equity Funds	3%	16%	11%	12%	-7%	27%	16%	-5%	7%	0.4%
Global Emerging Market Equity Funds	-10%	3%	4%	10%	-4%	32%	23%	-1%	12%	2.5%
EMEA Equity funds	27%	40%	-6%	-2%	-8%	11%	20%	-11%	-4%	-7.4%
Latin America Equity Funds	10%	81%	27%	46%	-12%	48%	4%	-12%	-1%	-8.5%
Asia Pacific Ex-Japan Funds	21%	22%	27%	14%	-9%	21%	10%	-7%	3%	0.2%
Total Bond Funds	14%	4%	8%	-2%	-10%	24%	16%	4%	11%	1.5%
International Bond Funds	12%	12%	10%	-2%	-24%	25%	23%	3%	6%	1.1%
Corporate High Yield Bond Funds	NA	-18%	-2%	-4%	-5%	40%	15%	4%	18%	1.4%
US Bond Funds	NA	-17%	-9%	4%	-2%	23%	10%	6%	12%	2.2%
Western Europe Bond funds	NA	1%	58%	-8%	-46%	29%	-7%	-28%	2%	-3.4%
Germany Bond funds	NA	NA	NA	NA	NA	NA	29%	25%	-13%	-5.7%
Switzerland Bond funds	NA	NA	NA	NA	NA	NA	-65%	-19%	-2%	-2.0%
United Kingdom Bond funds	NA	22%	-17%	-141%	-26%	64%	8%	-3%	0%	-4.1%
Emerging Markets Debt Funds	12%	24%	18%	9%	-21%	19%	54%	7%	25%	2.4%
Asia ex-Japan Bond funds	NA	4%	3%	16%	-10%	2%	71%	25%	12%	2.2%
Emerging Europe Bond funds	NA	40%	-12%	-18%	-37%	-19%	-8%	-39%	-9%	0.1%
Lat-Am Bond funds	NA	-28%	-22%	-33%	-30%	19%	46%	38%	68%	2.8%
Money Market Funds	NA	NA	NA	NA	31%	-17%	-15%	-4%	-1%	-2.7%

Colour Legend:

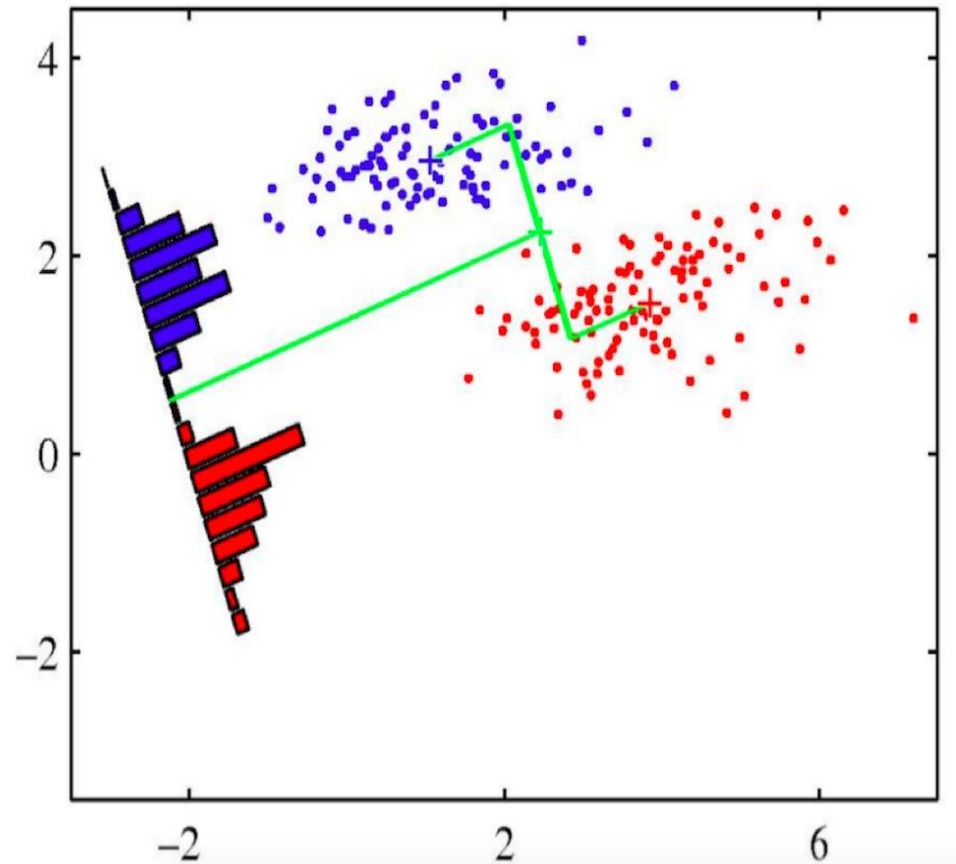
13%	8%	3%	-3%	-8%	-13%
>10%	5% to 10%	0% to 5%	-5% to 0%	-10% to -5%	<-10%

Source: EPFR, Deutsche Bank calculations

CERN School of Computing

Rotations - Fisher Discriminant

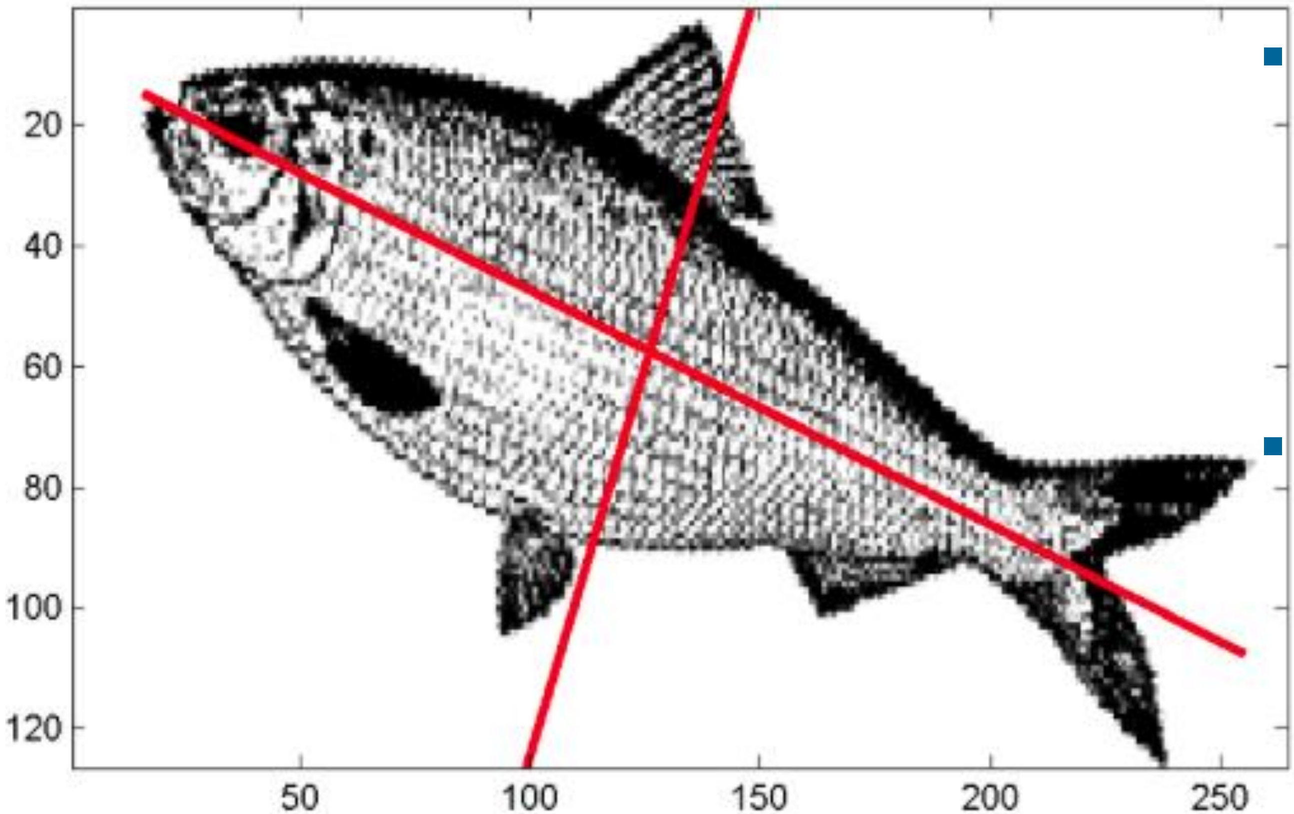
- **Rotating the axis of a 2d plot.**
- **Used to separate two distributions.**
- **For example signal and background.**
- **0 axis is defined as line best separating two distributions.**
- **This line doesn't have to be Straight…**
- **Other transformations?**



CERN School of Computing

Rotations - PCA

- **Principle Component Analysis**



- **Rotates axis to show maximum variance. This axis is referred to as the principle axis**
- **Other axis are defined in accordance**

Clustering

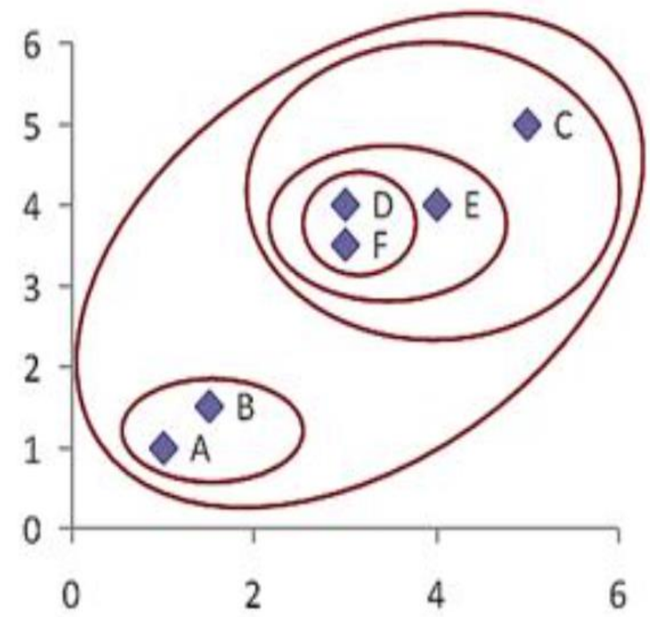
- **The notion of clusters is intuitive. A grouping of objects.**
- **Clusters can be formed from:**
  - Objects close together
  - Objects with similar properties
  - Objects that fit a particular distribution
- **Clustering can include all data points**
  - Automatically characterising groups of data.
  - Generalizes information for quicker processing
- **Clustering can highlight regions of interest**
  - Removing data that doesn't represent some underlying process.
  - Cleans data.

Defining Distance

- **Euclidean Distance (x,y)**
  - Simple. Intuitive. Easy to visualise
- **Density**
- **Correlations**
  - Shows similarity between variables
- **Mahalanobis distance (standardised statistical distance)**
  - Accounts for differences in scales between variables
  - Ignores effects from highly correlated variables
  - Ignores effects from variables with high variance
- **Many others.**
  - E.g. binary distance, like manhattan distance.

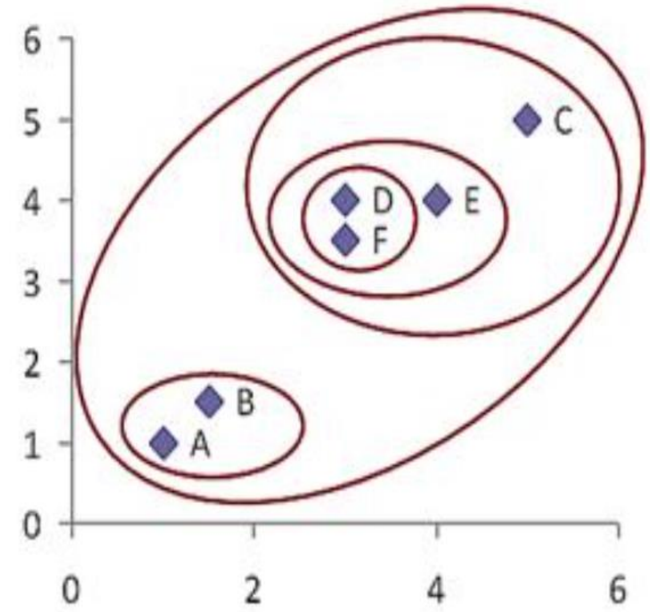
Hierarchical Clustering

- **Deterministic**
  - Results are always the same
- **Shows scale**
  - All points are clustered eventually
  - Needs stopping condition
- **Uses various distance metrics**
  - The closest two points are always the closest two, the two highest correlations are the two highest correlations



Hierarchical Clustering

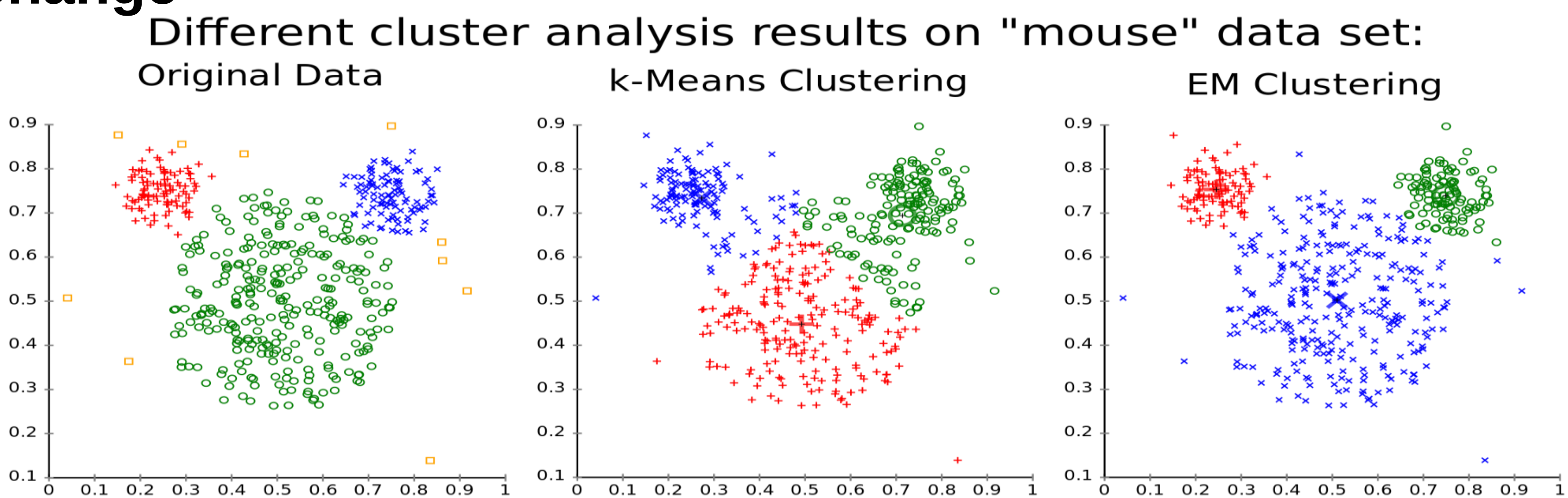
- **First find two closest points**
- **Merge into single cluster**
- **Find next two closest points**
- **Merge**
- **Continue until stop or all points are clustered**
- **Stopping conditions include:**
  - Number of clusters
  - Max distance
  - Fit to distribution



CERN School of Computing

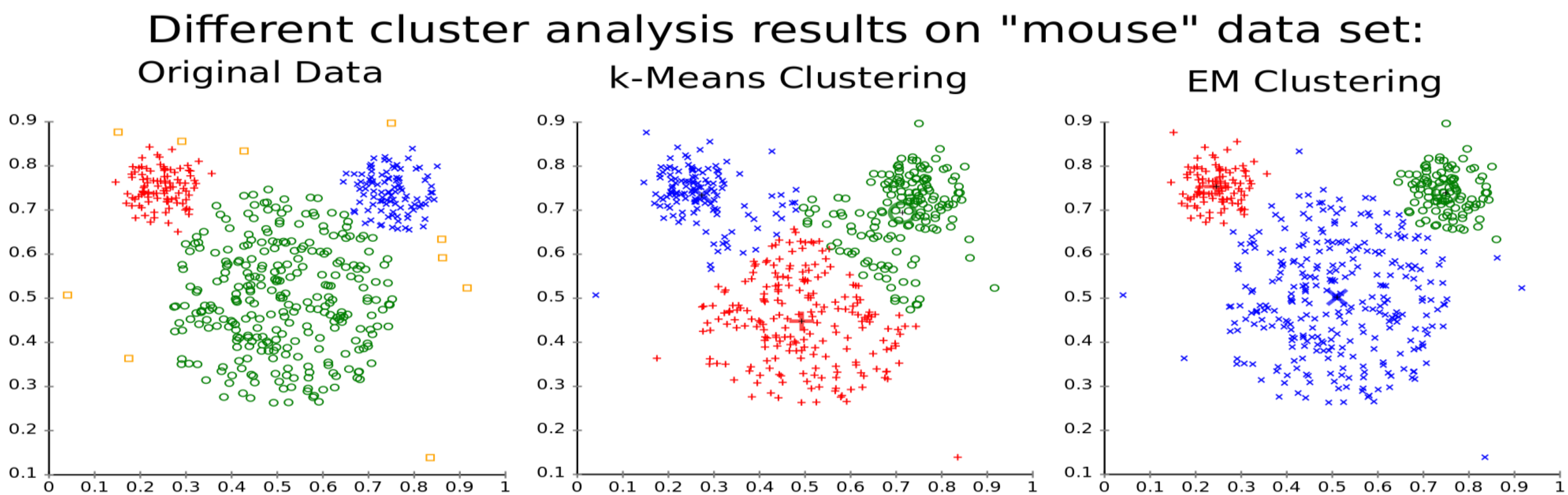
K-Means Clustering

- **K is the number of clusters**
  - This must be specified.
- **The initial properties of each centroid must be provided**
  - Often this must be guessed
- **Iterates over data until the position of the centroid doesn't change**



K-Means Clustering

- **Pick number of clusters**
- **Guess/assign centroids**
- **Assign points to the closest centroid**
- **Recalculate centroids**



Dimensional Reduction

- **Often we don't need all the information about a topic to characterise the underlying process.**
- **We can transform the data to summarise the data**
  - E.g SVD or PCA
- **We can cluster the data**
  - E.g. Hierarchical or k-means clustering
- **This can give us statistical information.**
- **This can also be used for data compression.**
  - (less variables=less data but with the same information)

Summary

- Data can show us lots of information.
- Information can be obtained from the inter-variable relationships. E.g. (PCA)
- Information can be obtained from the summaries of multivariate distributions.
- In Multivariate analysis adding variables and adding more data sometimes hides information rather than adds to it.
- By exploring the correlations, ranks and distributions of our data we can optimise the information contained for analysis.

Map Reduce

- **In MVA each additional variable reduces the density of information and increases processing time exponentially.**

- **MapReduce is a scalable programming model designed for processing very large data sets in a parallel distributed environment**

- **Two steps. (possibly iterated)**
  - Map Data
    - Filters and sorting
    - e.g. making clusters for each event
  - Reduce Data
    - Makes summary of data
    - E.g. combines clusters into histograms

      Use these to redefine clusters

Hadoop

- **Platform for distributed computing and parallelized computation whilst being scalable to meet exponential increases in data and cheap to implement.**
- **Inspired by Google research and Google File System**
- **Key implementation in analysis for Facebook, Yahoo, american express and many more.**