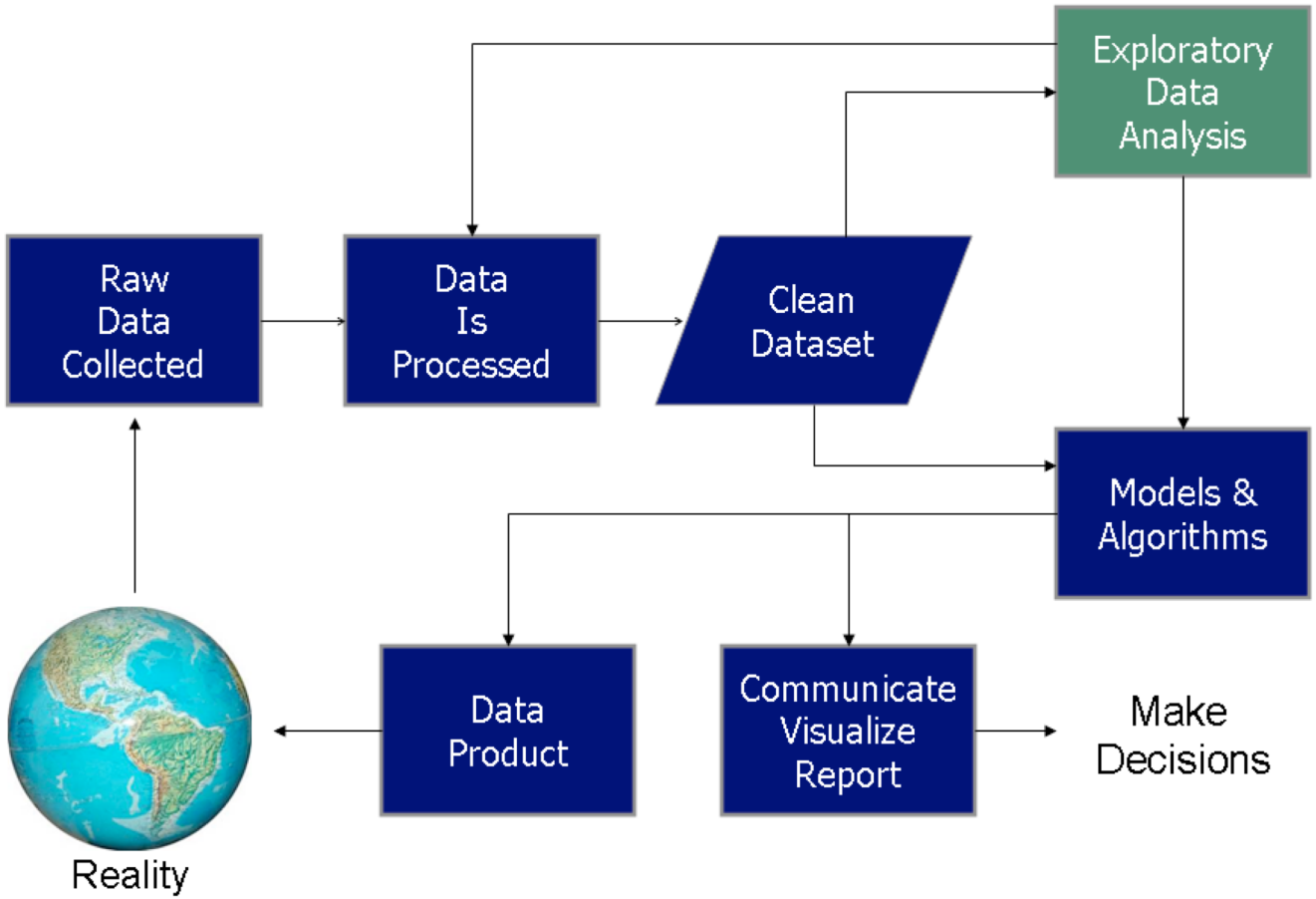
Exploratory Data Analysis

Hubert Rehrauer

Data Science Process



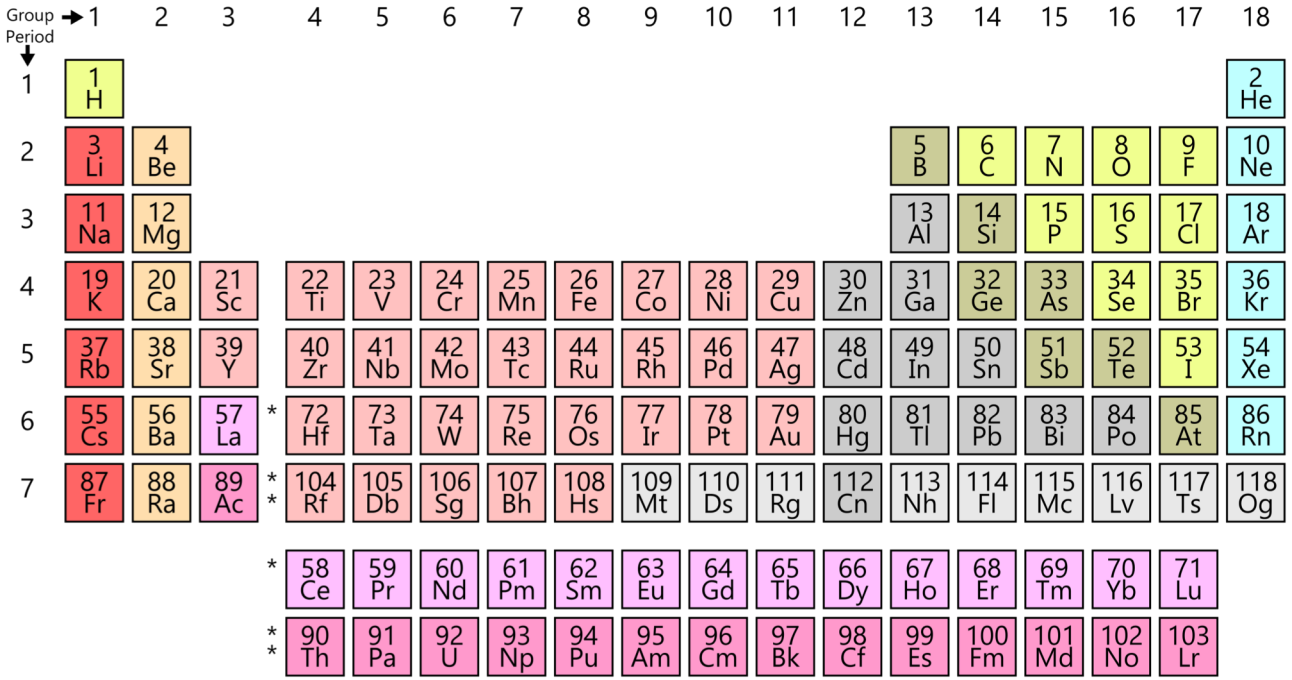
https://en.wikipedia.org/wiki/Exploratory_data_analysis

Dot map of cholera cases

John Snow, 1850, Source: https://devopedia.org/exploratory-data-analysis

Periodic Table

- made periodic structure apparent
- discovery of missing elements
- discovery of underlying principles



Goals and Methods of EDA

- Discover Patterns and Spot Anomalies
  - detection of trends but also mistakes
- Frame Hypothesis
  - assess direction and rough size of relationships between explanatory and outcome variables
- Check Assumptions
  - noise characteristics
  - preliminary selection of appropriate models

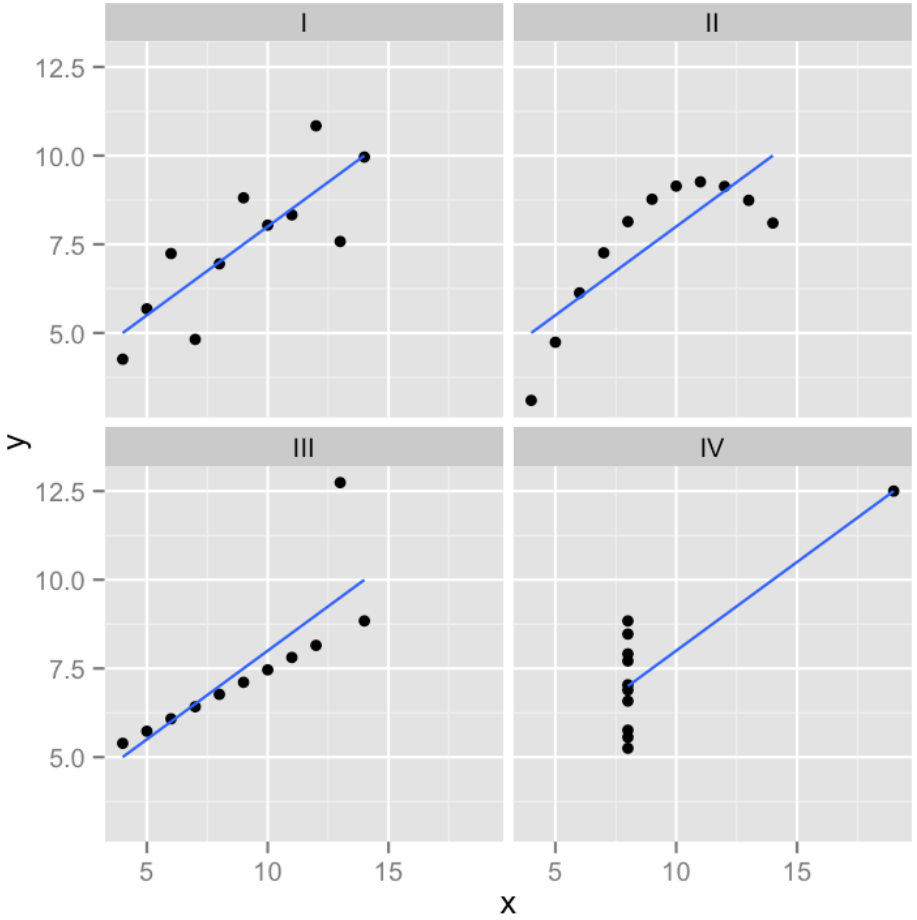
Methods:

- Visualize raw data
- Visualize measures of central tendency
  - mean, median, mode
- Visualize measures of dispersion
  - range, standard deviation, …
- Visualize relationships

Anscombe's quartet

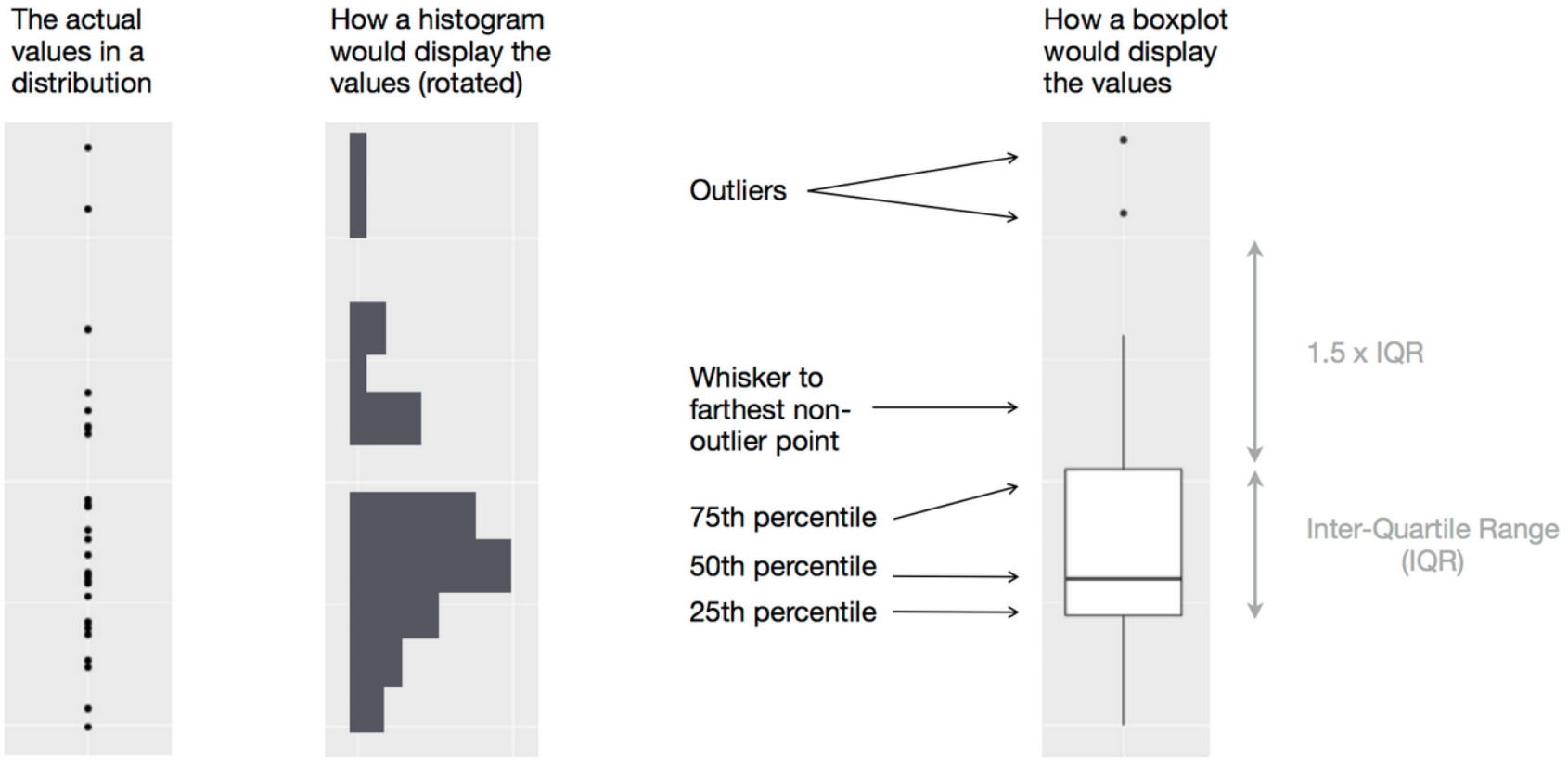
- All 4 datasets have nearly identical
  - mean
  - standard deviation
  - skewness & kurtosis

- Summarizing the data is not enough
- You have to visualize!

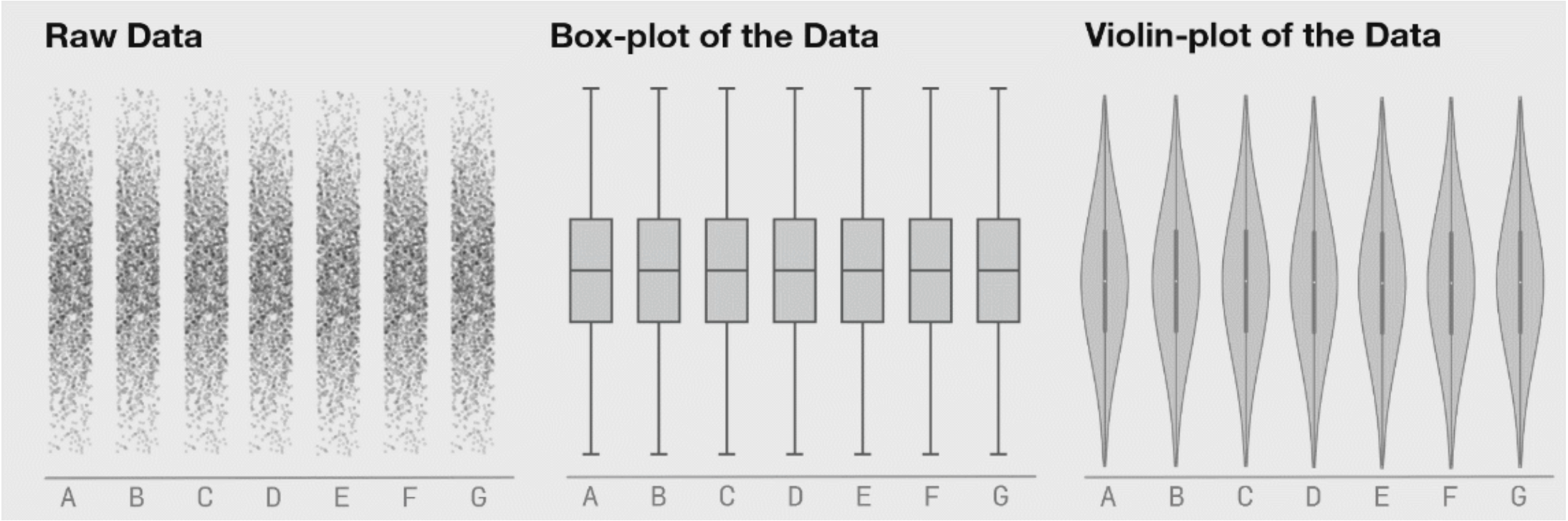


Turner 2016

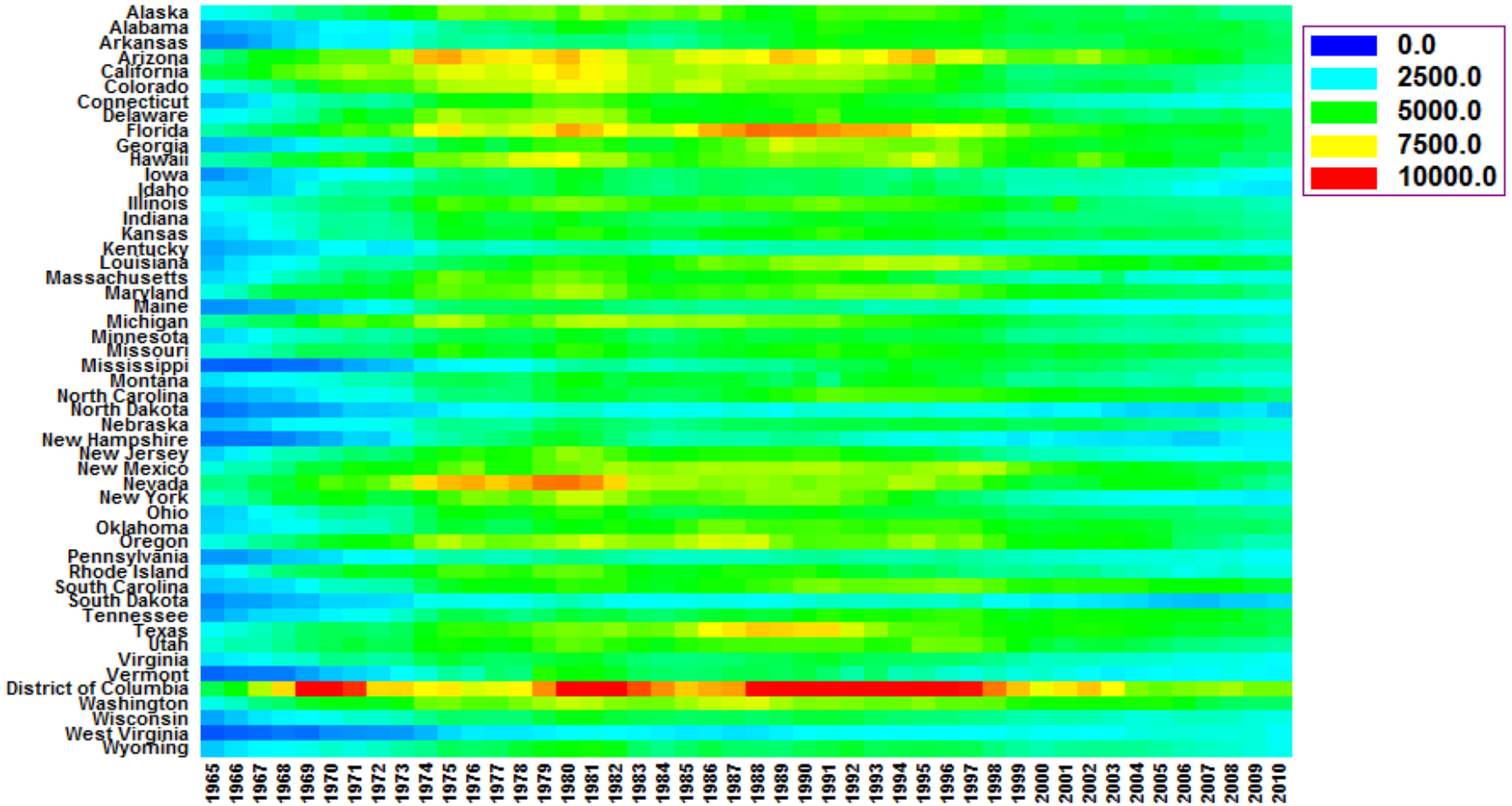
Example of misleading boxplot



https://r4ds.had.co.nz/exploratory-data-analysis.html


Raw Data
A B C D E F G
Box-plot of the Data
A B C D E F G
Violin-plot of the Data
A B C D E F G

Heat Map for Total Crime Rate



Data Generation Workflow

Experiment

⇓

Omics method

⇓

Raw “machine” data

⇓

Raw quantitative data →

- Normalization
- Transformation
- Visualization
- Quality Control
- Outlier Detection

→

- Quantitative sample differences

Quantitative Omics Data

- Gene expression data:
  - Quantification of relative mRNA abundance in cells/tissues
- Protein expression data
  - Quantification of relative protein abundance in cells/tissue
- Methylation status
- …

- Characteristics
  - obtained after analog signal transduction and amplification
  - not calibrated; no physical units
  - thousands or millions of measurements
  - measurements are at molecular scale

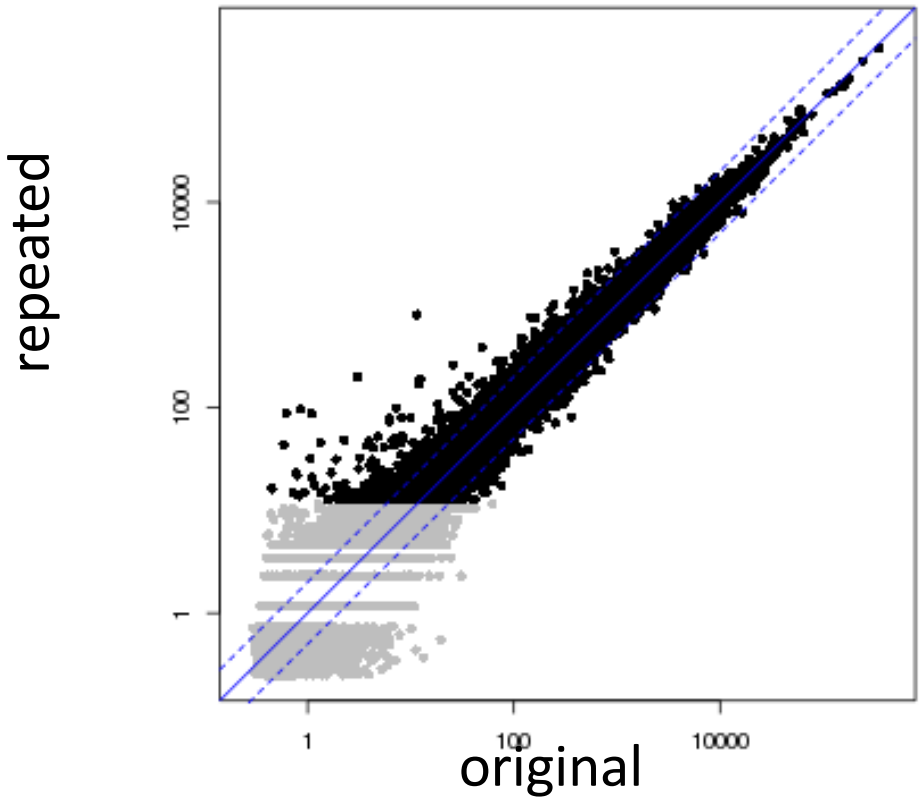
Technical characteristics

- Dynamic Range
  - ~10^4 to 10^6: largest value can be a million times higher than lowest values
  - non-linearity of measurement device

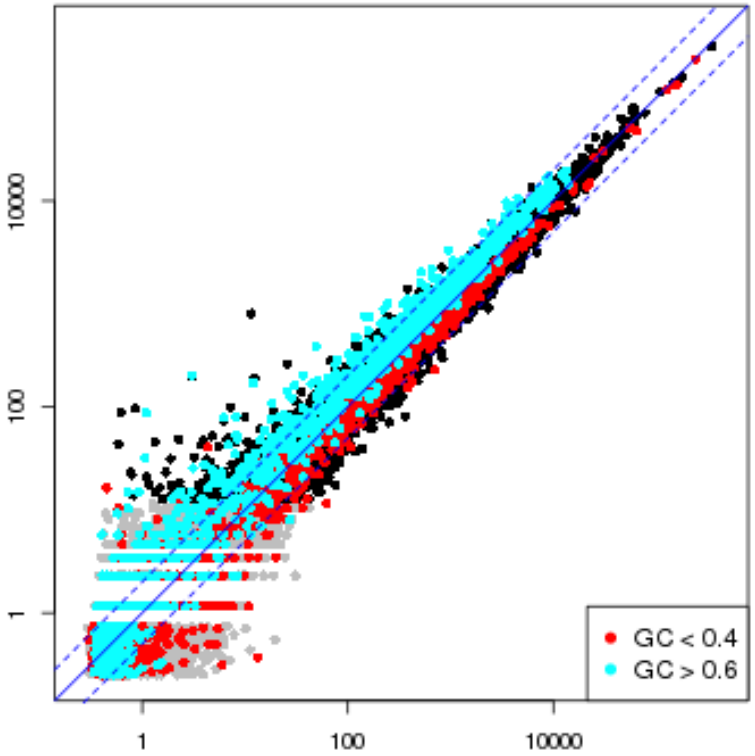
- Zero Measurements
  - additive background signal
  - zero value can have different explanations
    - technical failure to detect
    - value is truly zero
- Variability
  - Non-Gaussian noise

Example of Quantitative Data

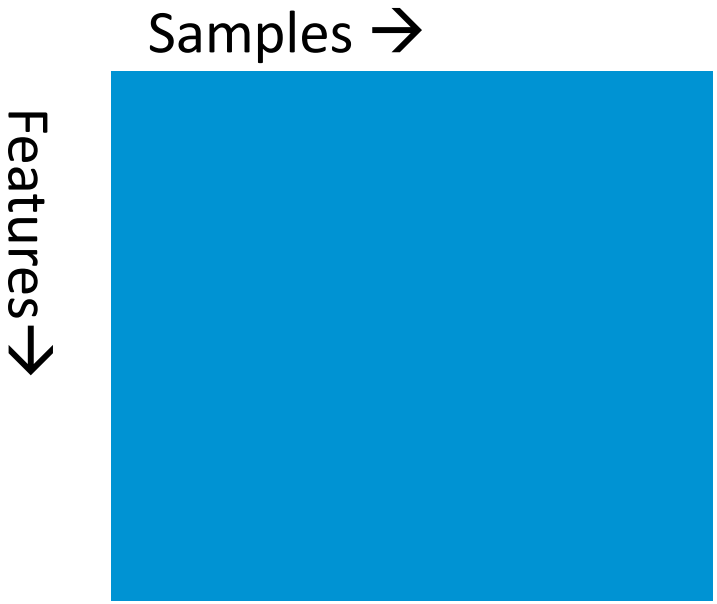
Comparison of a repeated experiment



Systematic effect of GC content on quantitative values



Representation as Data Matrix



Number of reads ≠ Expression level

	Sample 1	Sample 2	Sample 3
Gene A	5	3	8
Gene B	17	23	42
Gene C	10	13	27
Gene D	752	615	1203
Gene E	1507	1225	2455

- Gene D in the sample 3 has about twice as many reads aligned to it as in sample 2

Gene in Sample 2 Gene in Sample 3

1) The gene is two times more expressed in sample 3 than in sample 2

2) Difference in sequencing depth between samples – sequencing depth

3) Longer isoform was expressed in sample 3 – transcript length

Gene in Sample 2 Gene in Sample 3

Normalization & Transformation

- Scale samples so that all samples have the same total sum of expression values
- We will cover more sophisticated methods later

- Methods that assume that the majority of genes has no expression change are usually more accurate

Transformation

- Because of the large dynamic range of the expression values, the data should always be visualized at the log-scale

→ Log Transform with an additive constant

Data Exploration

- How similar are the expression profiles of the samples?
- Do the similarities match the experimental design and the anticipated effect sizes?
- Are samples within an experimental condition more similar than across conditions?
- Are there outliers?

- Distance measures for two expression profiles X und Y:
  - 1 – correlation(X, Y)
  - Euclidian distance
  - …

- Most frequently the correlation is used

Euclidean Distance

Euclidean distance of two profiles $\mathbf{x}$ and $\mathbf{y}$ with p genes

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

Note: Expression values should be at log scale

Correlation measure

Correlation of two profiles with p genes:

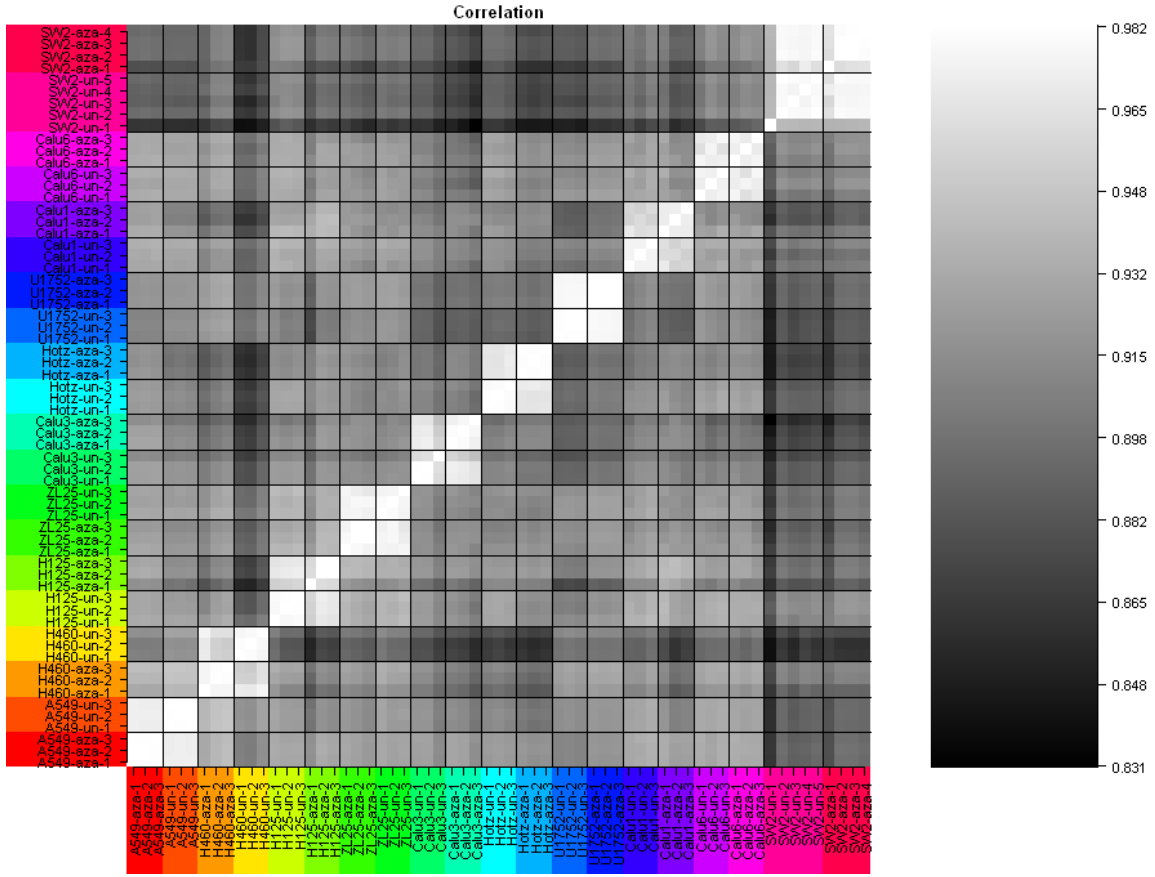
$$s(x,y) = \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i - \bar{x})^2 \times \sum_{i=1}^{p}(y_i - \bar{y})^2}}$$

averages: $\bar{x} = \frac{1}{p}\sum_{i=1}^{p} x_i$ and $\bar{y} = \frac{1}{p}\sum_{i=1}^{p} y_i$.

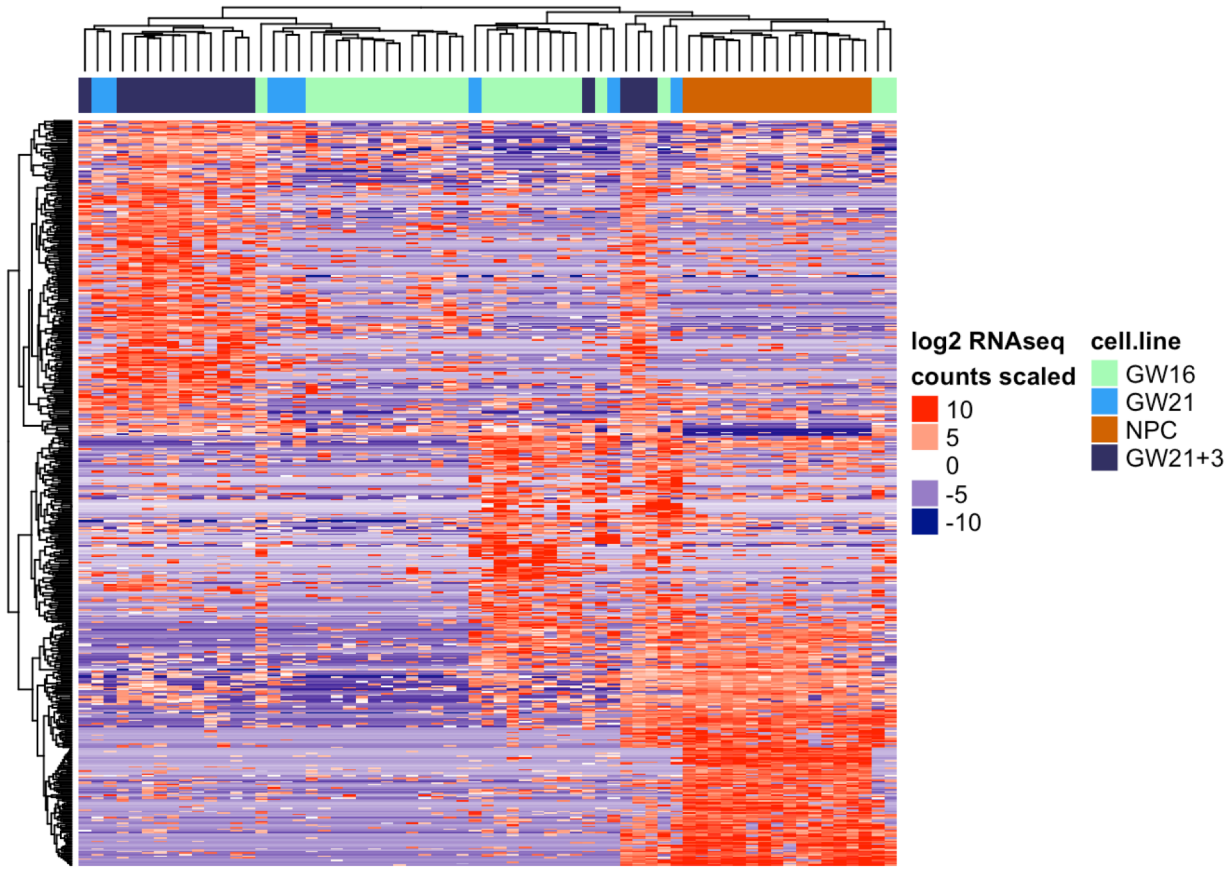
Correlation matrix for samples

The matrix shows the correlation for all sample pairs.

This gives a quick overview on the presence of outliers.

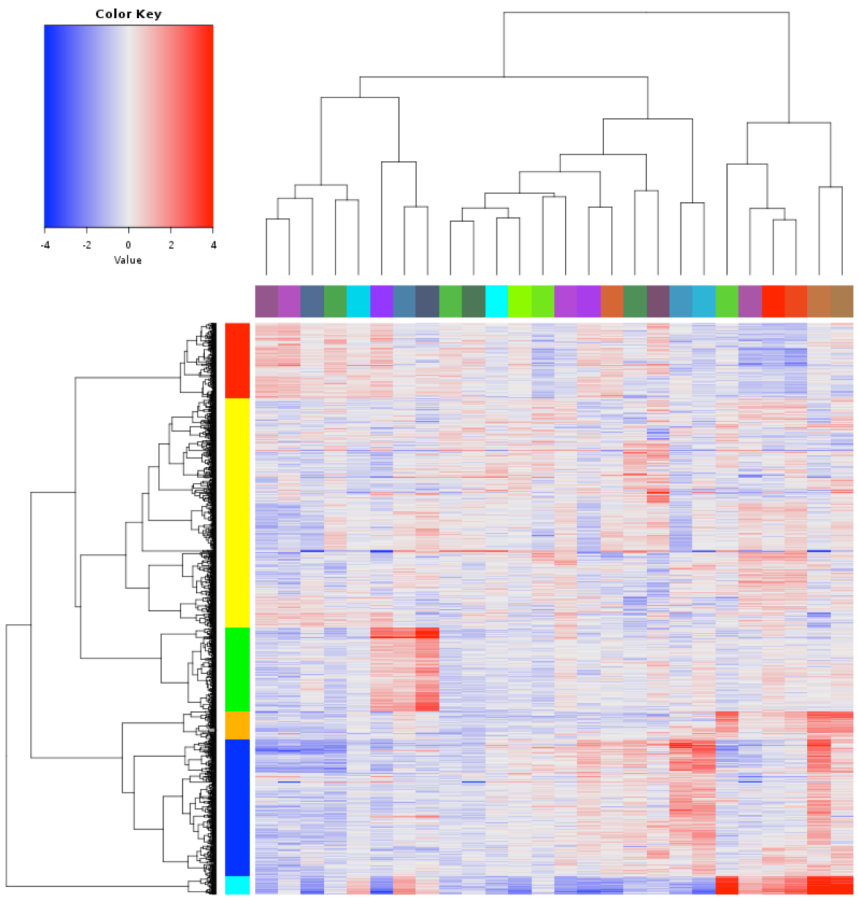


Example: Identifying Outlier Samples and Expression Structure



https://rpubs.com/crazyhottommy/heatmap_demystified

Simultaneous clustering of samples and genes



- Columns are samples
- Samples have different
  - genotype
  - gender
  - body mass index

- Green cluster: Neutrophil degranulation, monocytes, macrophages
- Blue cluster: Ppar signaling; lipid particle; adipocytes; adipose signaling;
- red cluster: oxidative stress??
- lightblue+orange: liver

23

Hierarchical Clustering

Goal

- Grouping of samples according to similarity

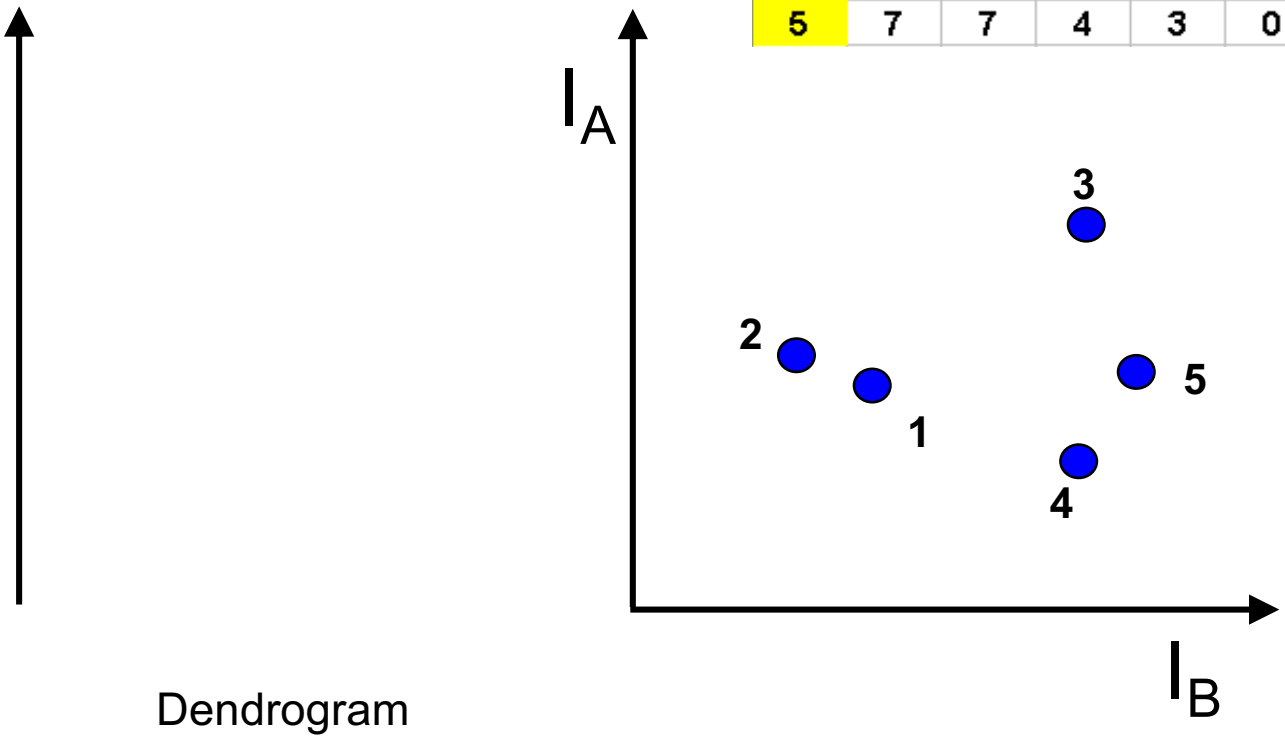
Procedure

- Initialization
  - Every sample is a cluster
- Iteration:
  - Recursive joining of the most similar clusters

Example

Cluster distances

d(ij)	1	2	3	4	5
1	0	2	6	9	7
2	2	0	5	7	7
3	6	5	0	5	4
4	9	7	5	0	3
5	7	7	4	3	0

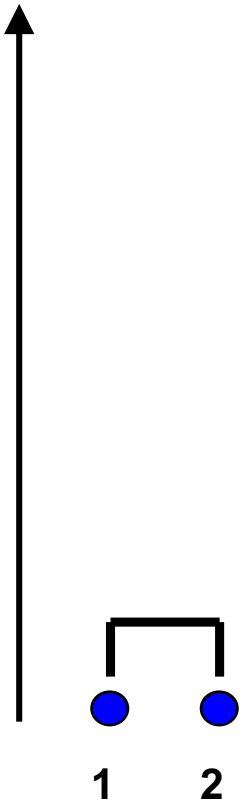


Dendrogram

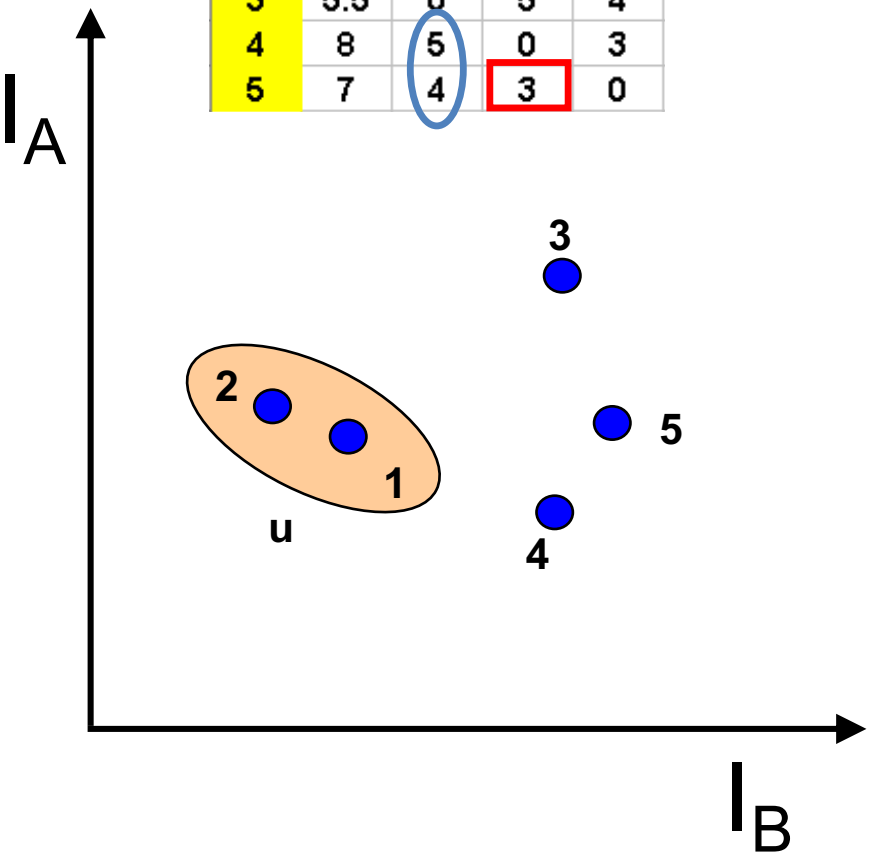
Example

Cluster distances

d(ij)	u	3	4	5
u		5.5	8	7
3	5.5	0	5	4
4	8	5	0	3
5	7	4	3	0



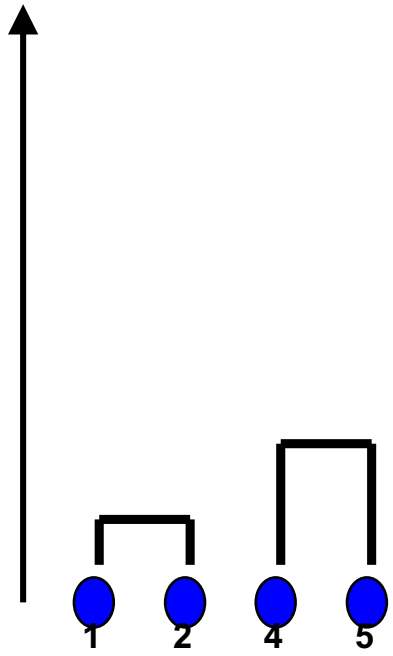
Dendrogram



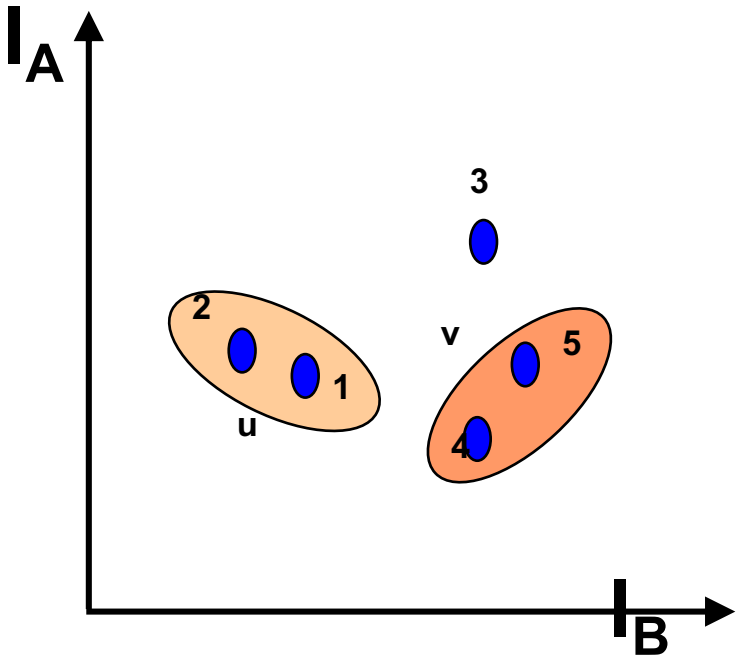
Example

Cluster distances

d(ij)	u	3	v
u	0	5.5	7.6
3	5.5	0	4.5
v	7.5	4,5	0



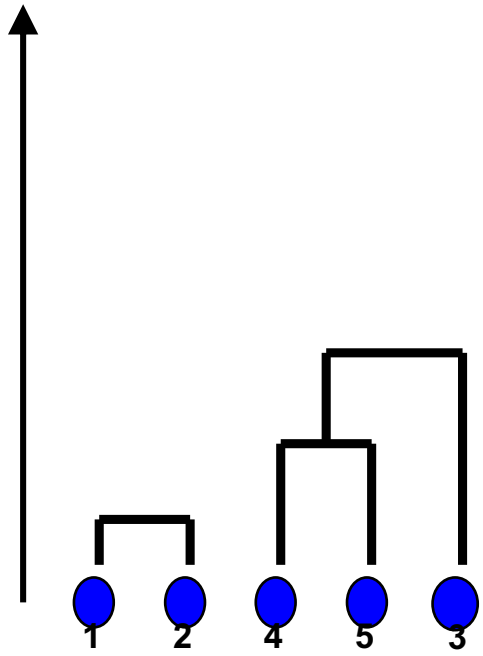
Dendrogram



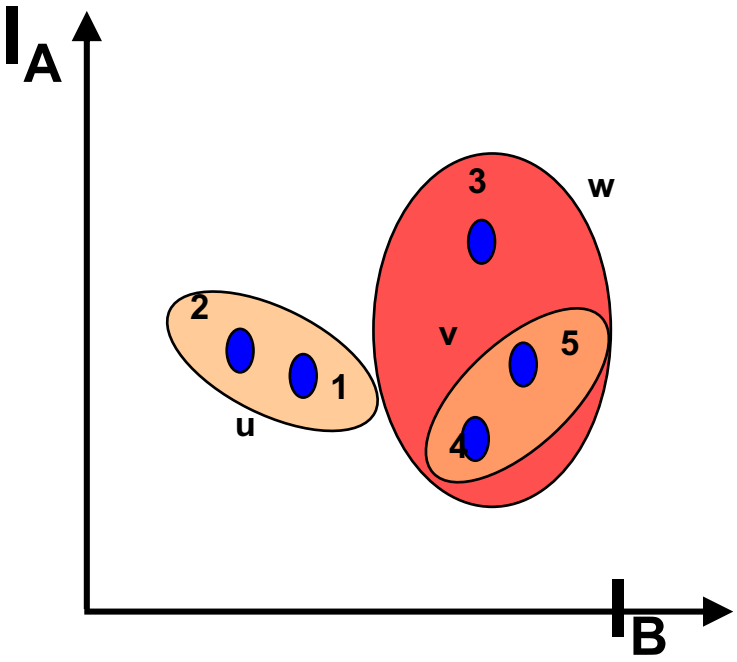
Example

Cluster distances

d(ij)	u	w
u	0	6.5
w	6.5	0

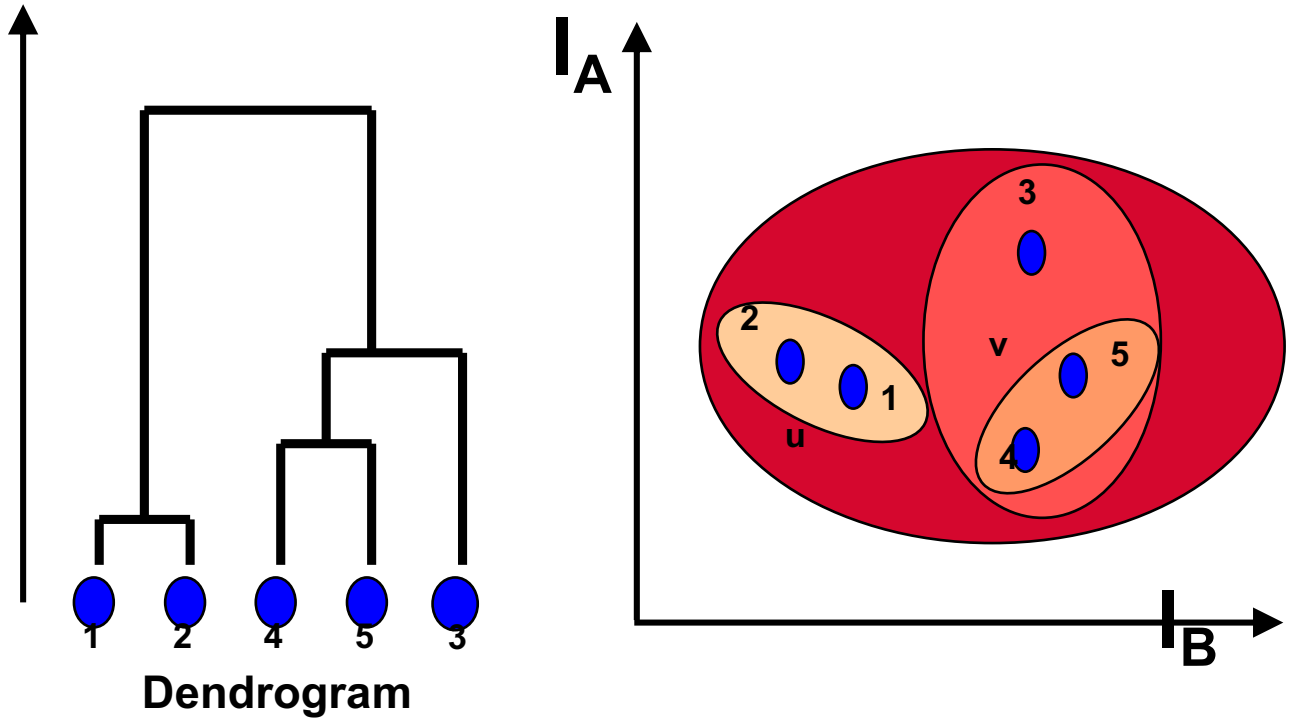


Dendrogram



Example

Cluster distances



Hierarchical Clustering: Algorithm

Algorithm

1. Compute matrix of pair-wise distances
2. Find pair with minimal distance and merge them
3. Update distance matrix
4. Continue with step 3 until a single cluster is left

Parameters:

- Distance measure for individual samples
  - For gene expression this is typically the correlation
- Distance measure for clusters of samples (linkage rule)

Hierarchical Clustering

Linkage Rules:
How to compute the distance of two clusters (groups of samples)

- Single linkage: minimal distance of two members
- Complete linkage: maximal distance of two members
- Average linkage: average distance of all members
- Ward‘s linkage: minimal increase in intra-cluster variance
- …

Hint:
- The above cluster distances can be derived directly from the distance matrix of the samples
- Cluster algorithm only needs the distance matrix as input not the measurements of the individual samples

K-means Clustering

1. Initialization: Randomly assign each sample to a cluster
2. Compute the cluster centers as the average of the assigned samples
3. Assign each sample to the closest cluster center
4. Repeat steps 2 and 3 until convergence or maximum number of steps is reached

Characteristics:

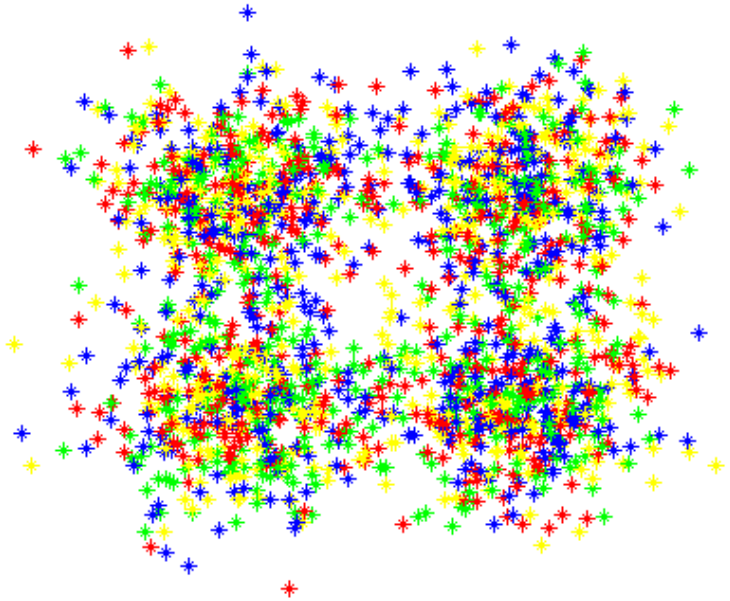
- Finds clusters that minimize intra-cluster variance

Parameters:

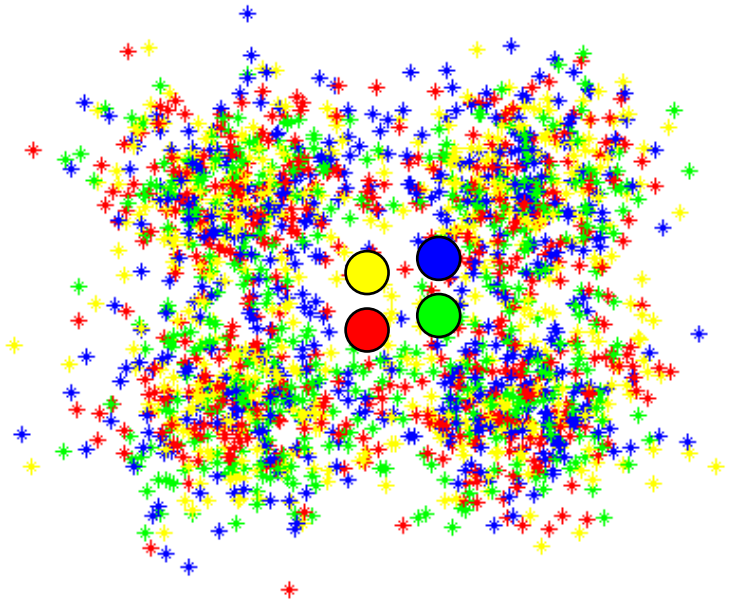
- number of clusters
- distance measure

K-means Example

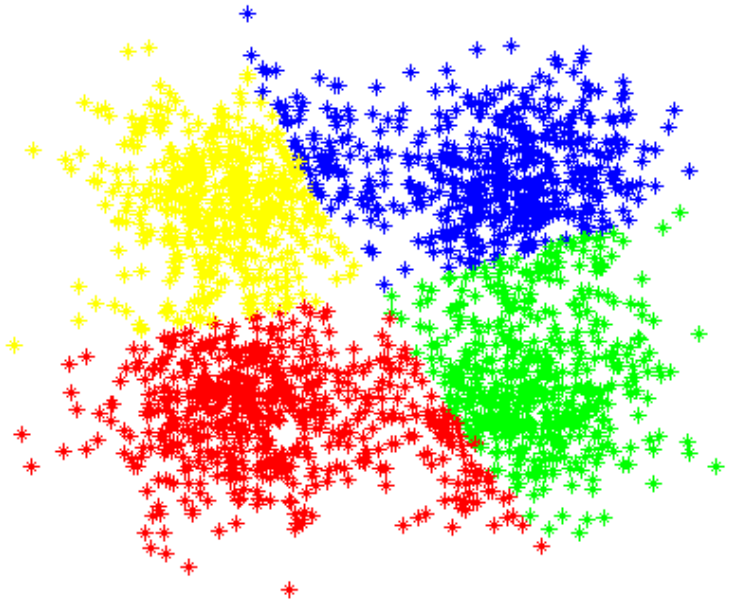
K-means: Initialization



K-means: Computing cluster centers



K-means: Reassignment of points



Comparison of Clustering Methods

- Computing time:
- Hierarchical clustering
  - O(n2 log(n))
- K-means clustering
  - t: number of iterations
  - k: number of clusters
  - O(k t n)

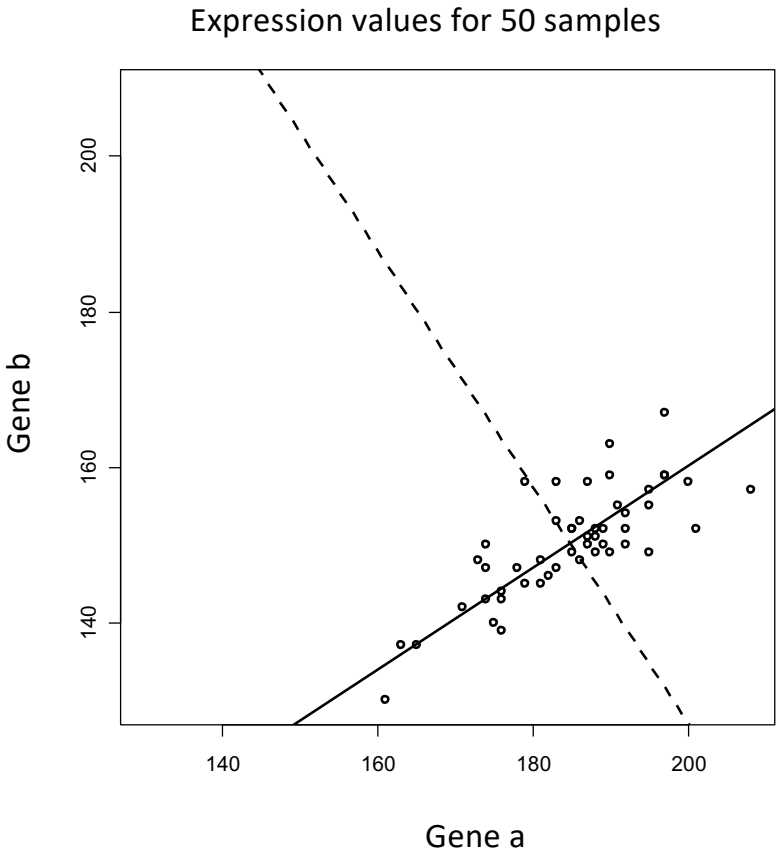
- Memory requirements:
- Hierarchical clustering
  - O(n2)
- K-means clustering
  - O(kn)

Note:

When clustering large numbers of genes (>1e4, hierarchical clustering becomes resource intensive)

Principal Component Analysis

- An expression profile characterizes the state of a sample with ~25 000 genes (variables)
- Can we get a representation that uses less variables? Reduction of dimensionality?
- Yes, genes that are highly correlated can be summarized without major loss of information content
- Goal is to represent the samples in a low-dimensional space where the distance relationships of the samples are similar to the relationships in the full space



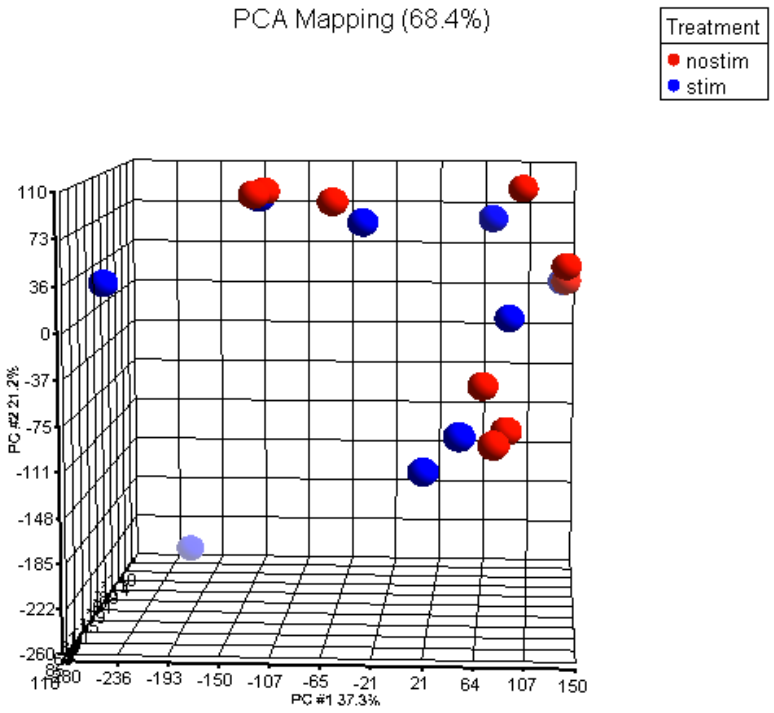
Principal Component Analysis

Procedure:

- Center the matrix: Subtract average
- Compute covariance matrix of the centered matrix (gives an n x n Matrix)
- Compute Eigenvalues and Eigenvectors of the covariance matrix
- Sort Eigenvectors according to the magnitude of the associated Eigenvalues
- Transform to the Eigenspace
- Only show the first k variables in the Eigenspace

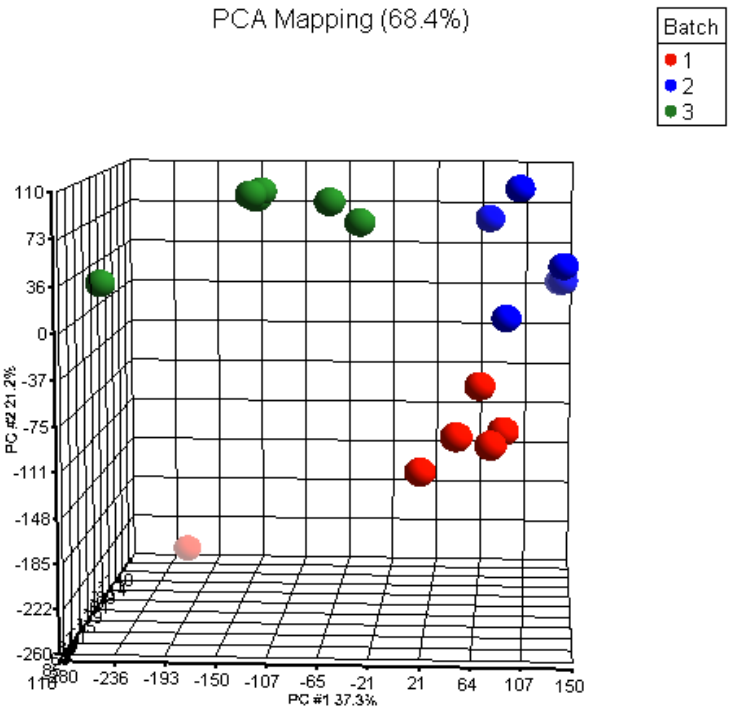
Example: Stimulation experiment

- Plot showing the 18 samples in the PCA coordinates
- Coloring by treatment of the samples
  - stimulated
  - not stimulated
- The samples do not separate
- There are two outliers on the left



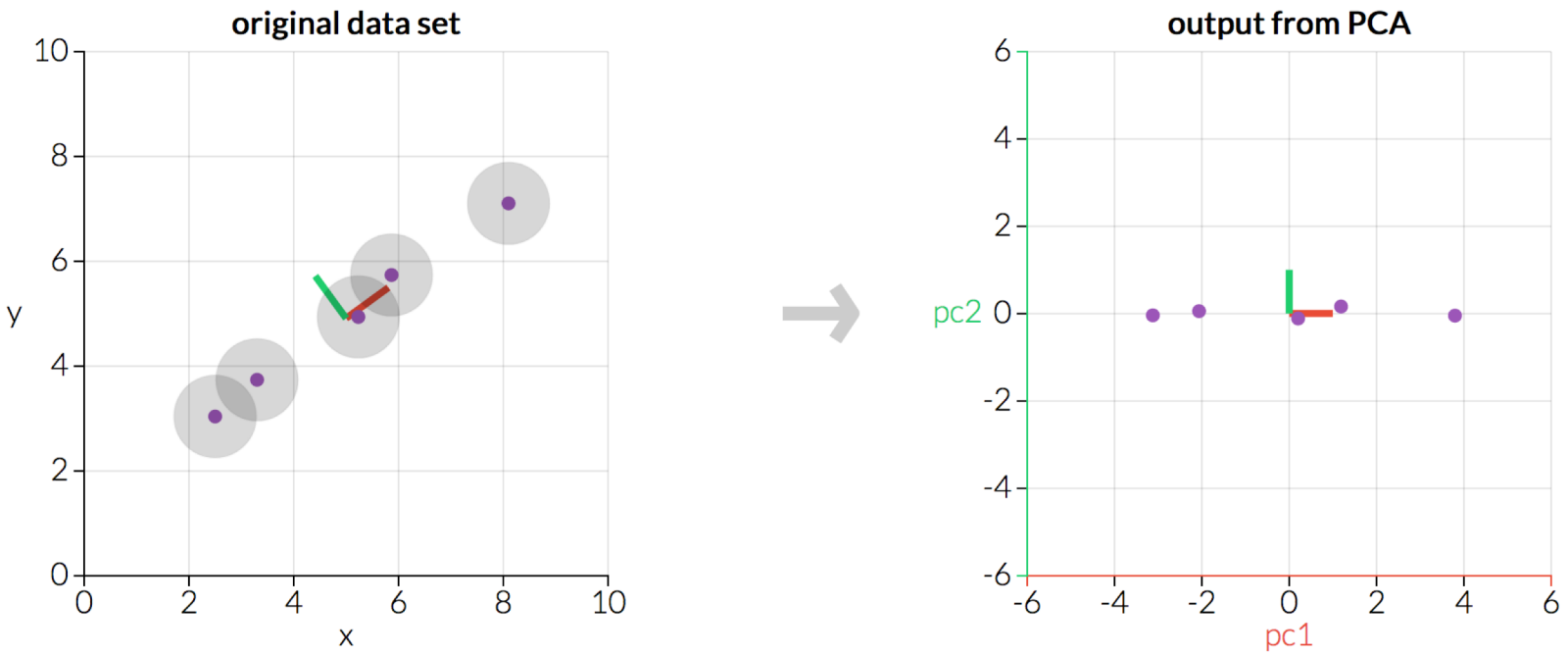
Example: Stimulation experiment

- Coloring by batch shows that the major effect is the batch effect
- Global expression profile is majorily determined by the batch
- Stimulation leads only to a minor modulation of the expression profile



PCA Explained Visually

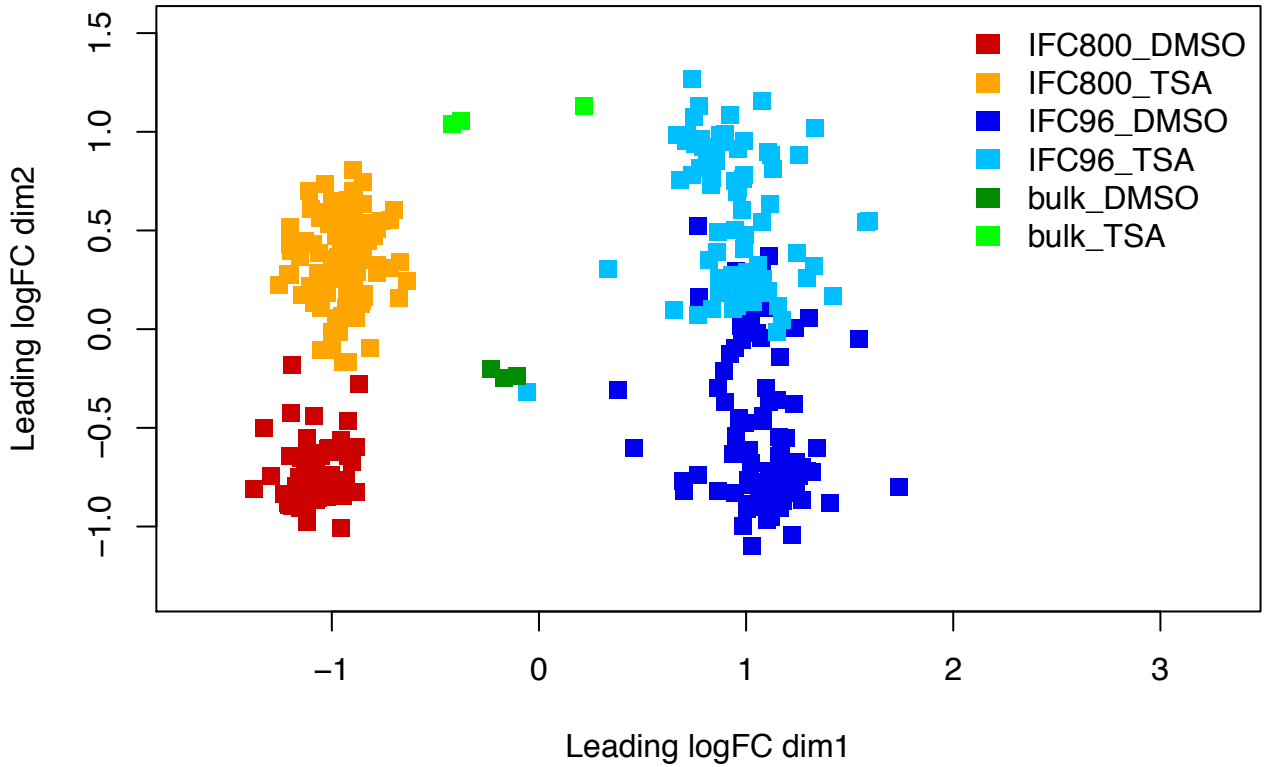
- http://setosa.io/ev/principal-component-analysis/



Example of multi-dimensional scaling

Identify consistency

Assess effect sizes



Example of exploratory analysis