



Exploratory Data Analysis

Hubert Rehrauer

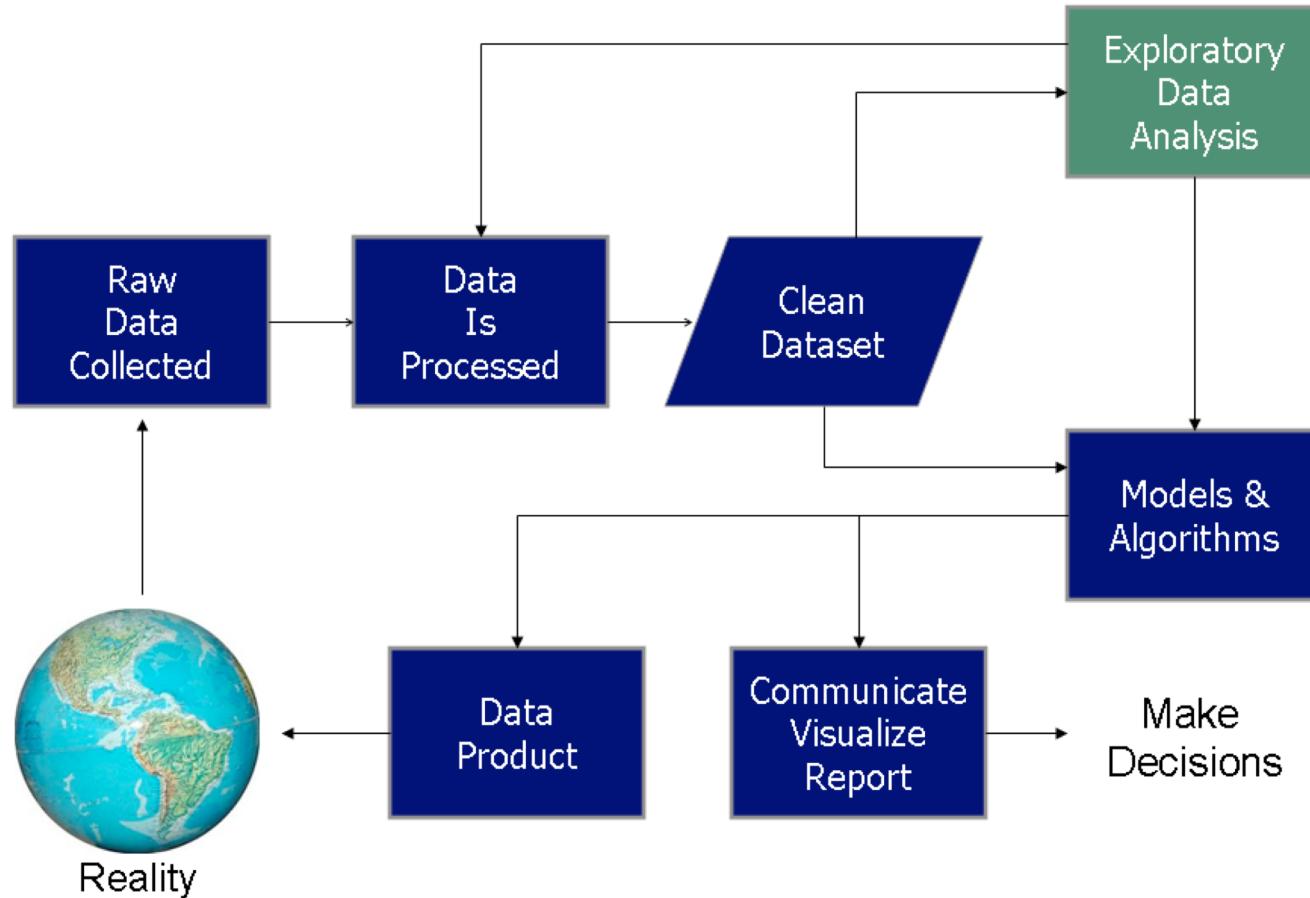


University of
Zurich^{UZH}

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Data Science Process



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



Periodic Table

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

Group ↓	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
1	1 H																2 He		
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne	
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar	
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr	
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe	
6	55 Cs	56 Ba	57 La	*	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	86 Rn	
7	87 Fr	88 Ra	89 Ac	*	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Nh	114 Fl	115 Mc	116 Lv	117 Ts	118 Og
	*	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu				
	*	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr				



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

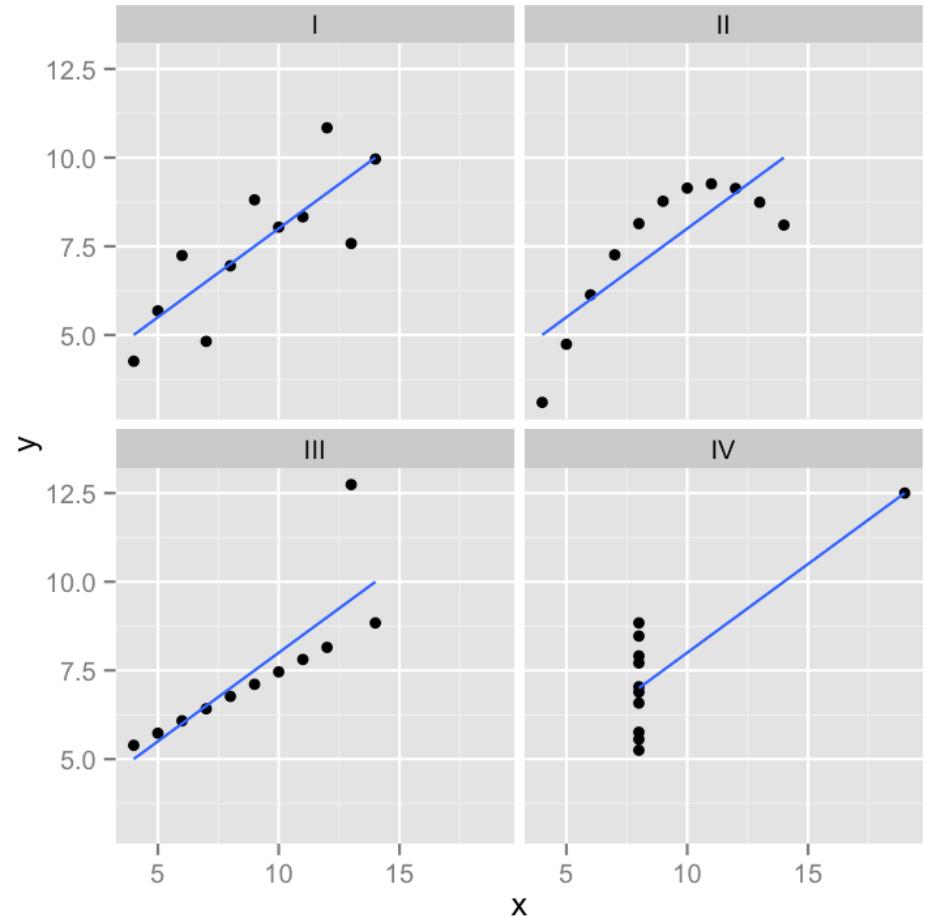
Methods:

- 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!





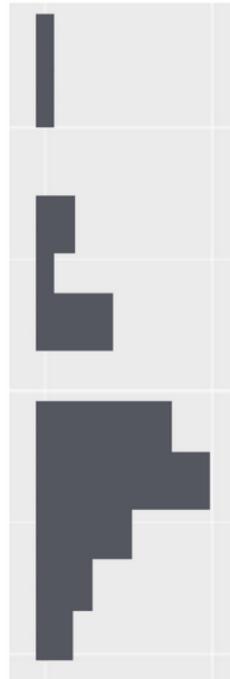
functional genomics center zurich
0
1
0
010 01
101 10
010 01
01 1
0 1
0 0

Example of misleading boxplot

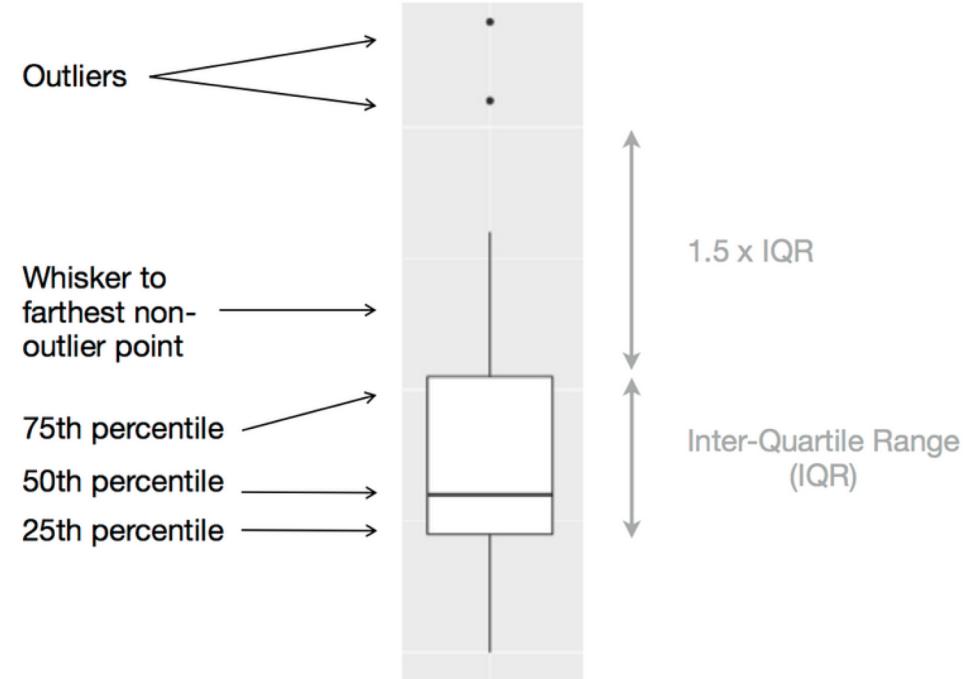
The actual
values in a
distribution



How a histogram
would display the
values (rotated)

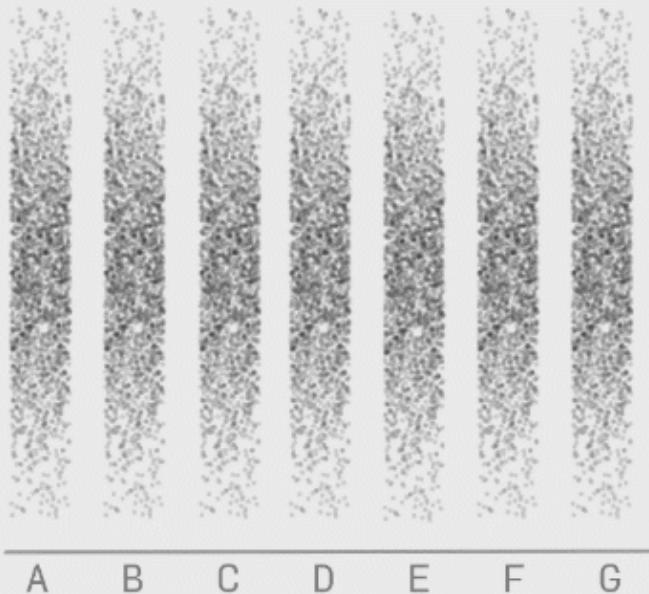


How a boxplot
would display
the values

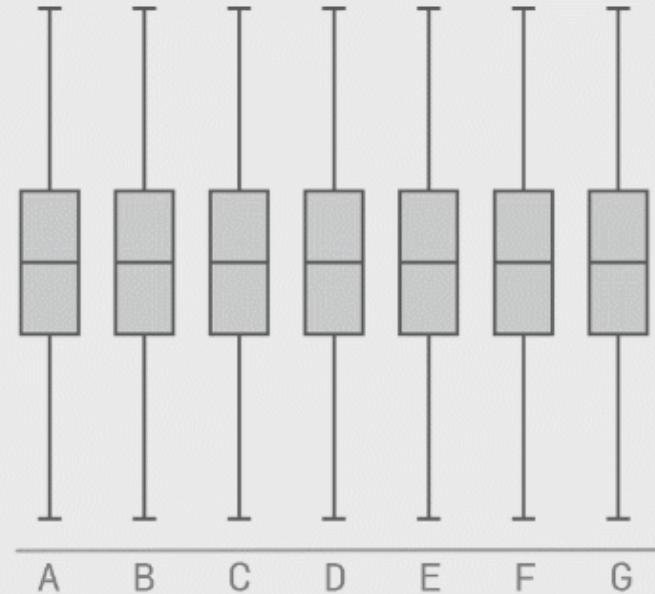




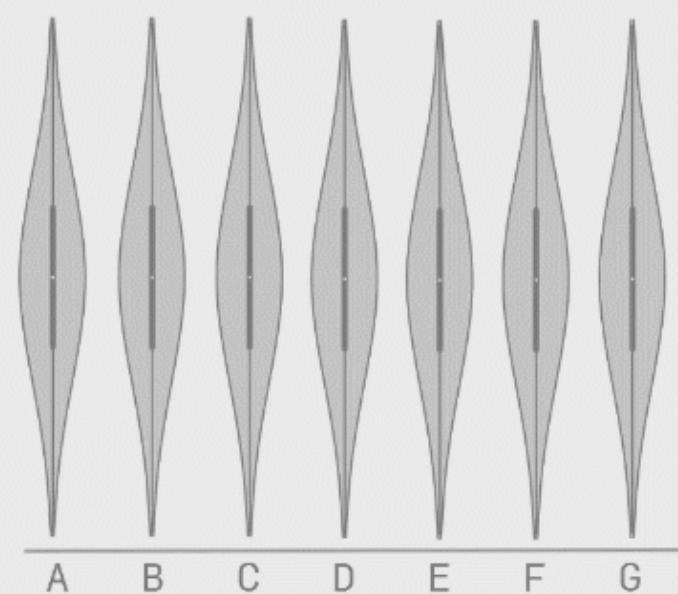
Raw Data



Box-plot of the Data

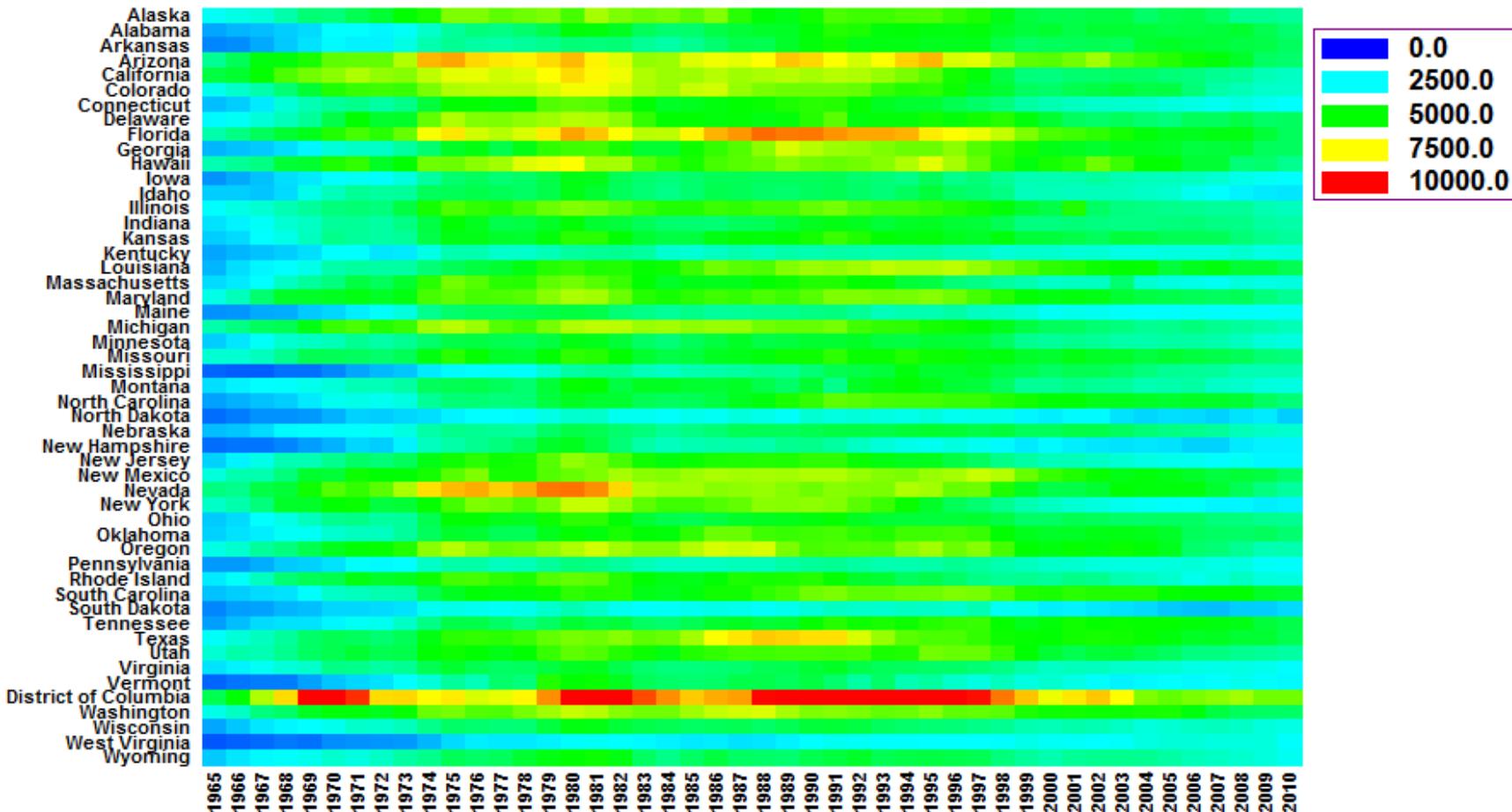


Violin-plot of the Data



⇒ Different distributions can lead to the same boxplot, but Violin plots can reflect different densities

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

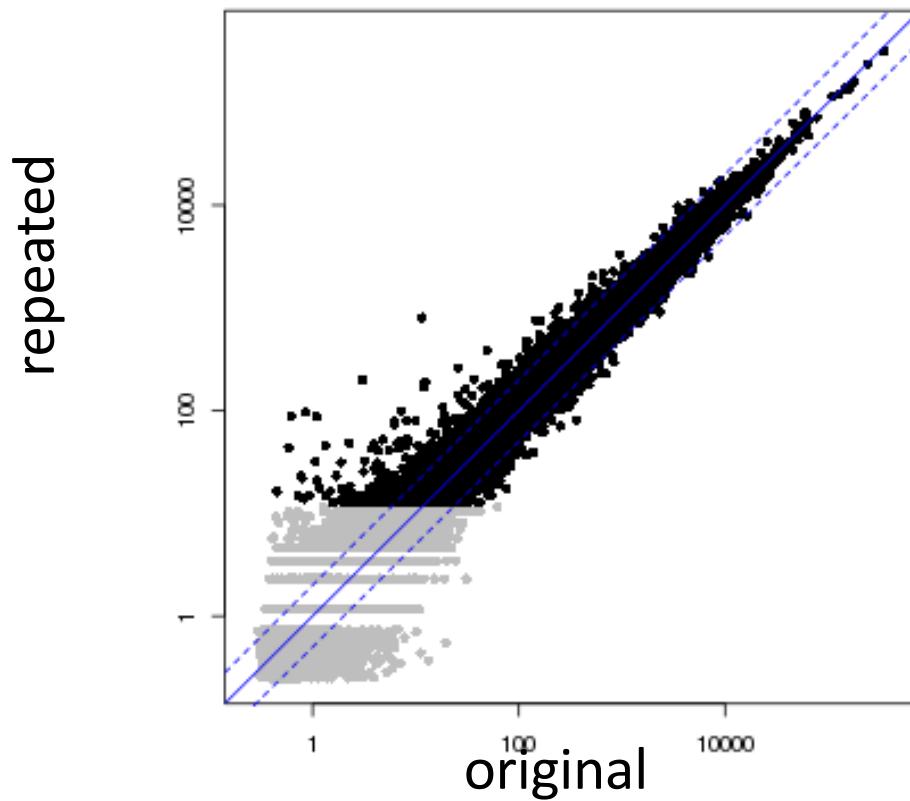
- Gene expression data:
 - Quantification of relative mRNA abundance in cells/tissues
 - Protein expression data
 - Quantification of relative protein abundance in cells/tissue
 - Methylation status
 - ...
 - Characteristics
 - obtained after analog signal transduction and amplification
 - not calibrated; no physical units
 - thousands or millions of measurements
 - measurements are at molecular scale

Technical characteristics

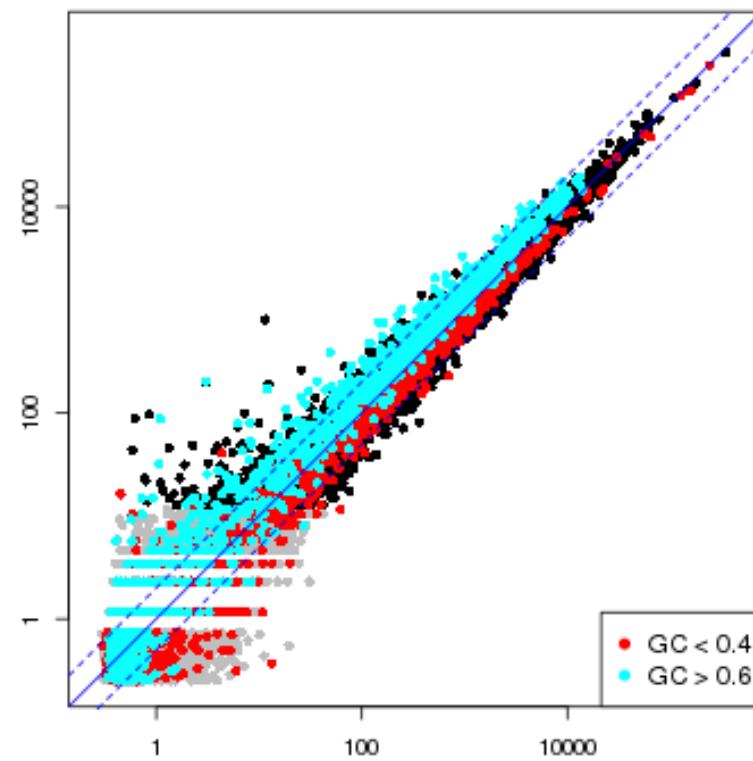
- Dynamic Range
 - $\sim 10^4$ to 10^6 : largest value can be a million times higher than lowest values
 - non-linearity of measurement device
 - Zero Measurements
 - additive background signal
 - zero value can have different explanations
 - technical failure to detect
 - value is truly zero
 - Variability
 - Non-Gaussian noise

Example of Quantitative Data

Comparison of a repeated experiment



Systematic effect of GC content on quantitative values



Representation as Data Matrix

Samples →

Features →



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



- The gene is two times more expressed in sample 3 than in sample 2
- Difference in sequencing depth between samples – sequencing depth
- Longer isoform was expressed in sample 3 – transcript length





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:
 - $1 - \text{correlation}(X, Y)$
 - 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



10

01

101



Correlation measure

Correlation of two profiles with p genes:

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

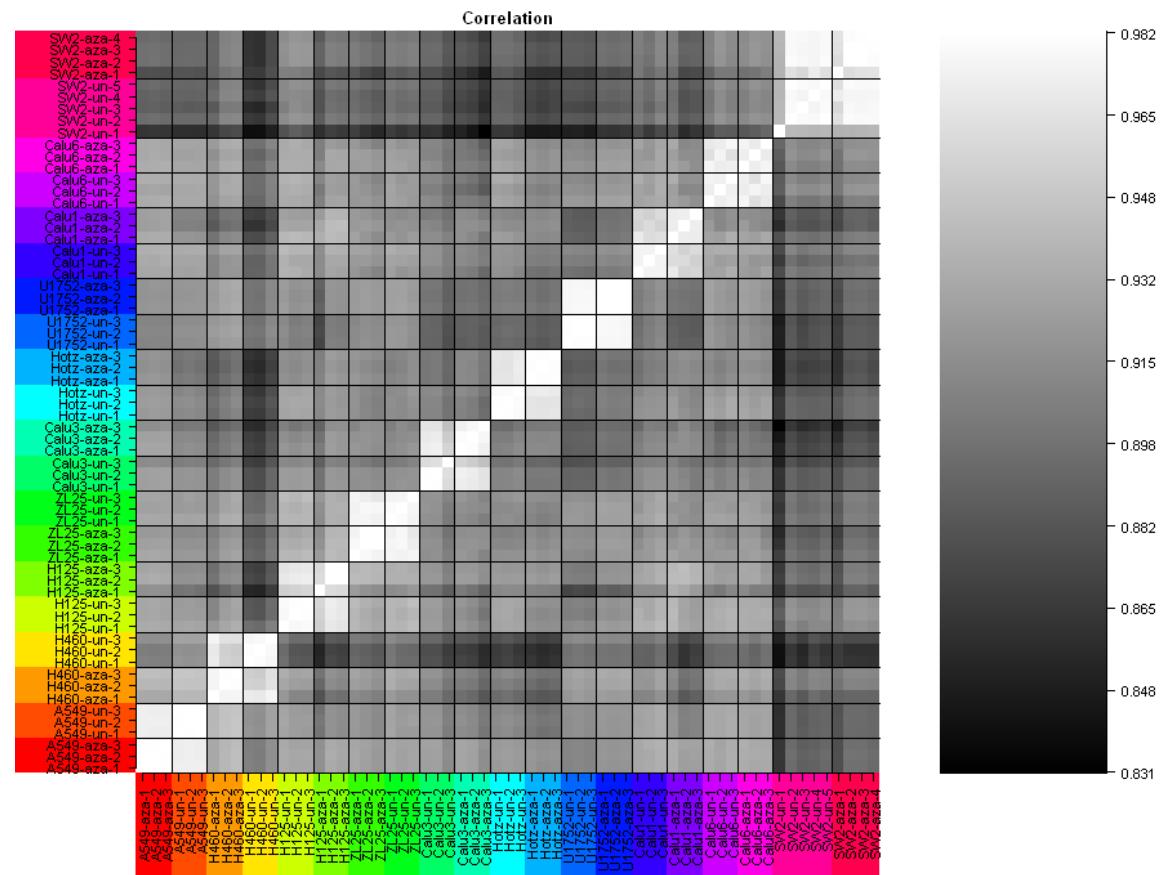
averages: $\bar{x} = \frac{1}{p} \sum_{i=1}^p x_i$ and $\bar{y} = \frac{1}{p} \sum_{i=1}^p y_i$.

10
01
101

Correlation matrix for samples

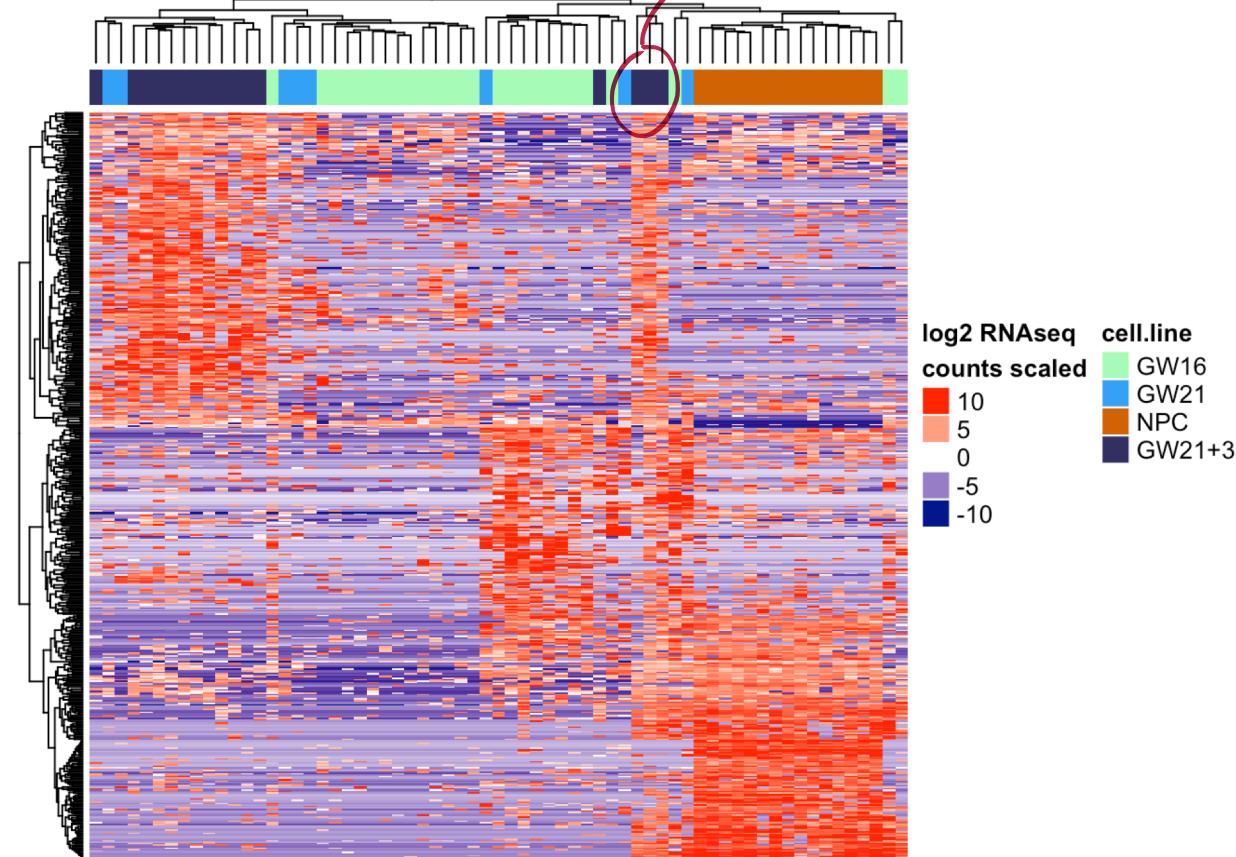
The matrix shows the correlation for all sample pairs.

This gives a quick overview on the presence of outliers.

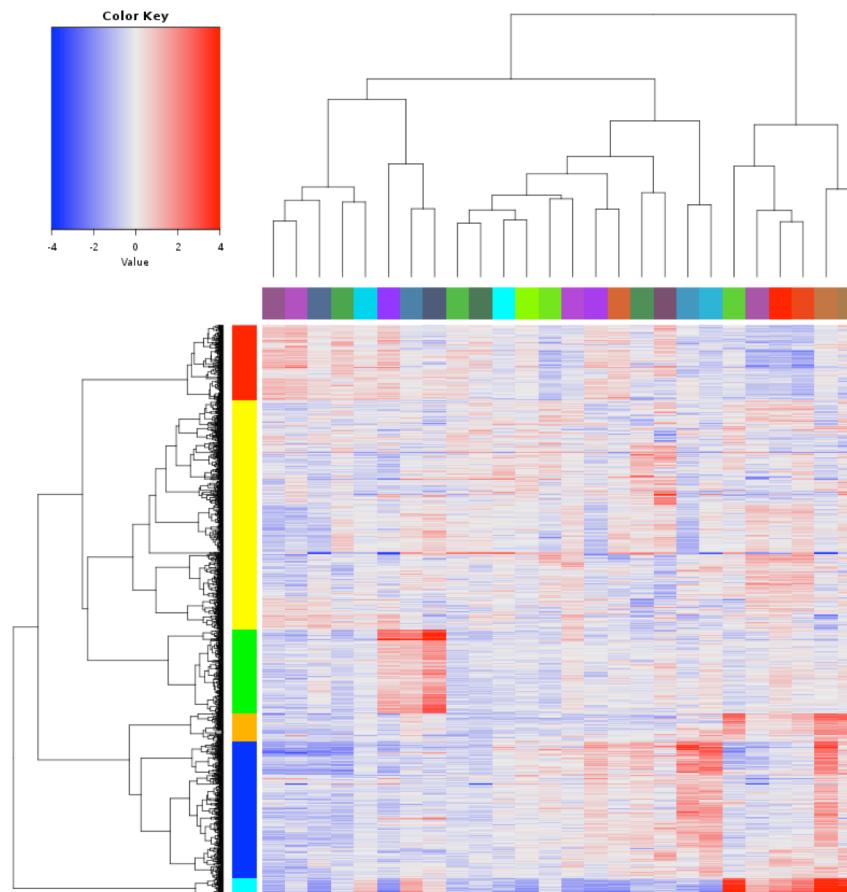


Example: Identifying Outlier Samples and Expression Structure

- Cluster red with the
rest of SW21^{ES} like



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

10
01
101

functional genomics center zurich

010 01
101 10
010 01
01 1f g c z
10 0
01 1

Hierarchical Clustering

Goal

- Grouping of samples according to similarity

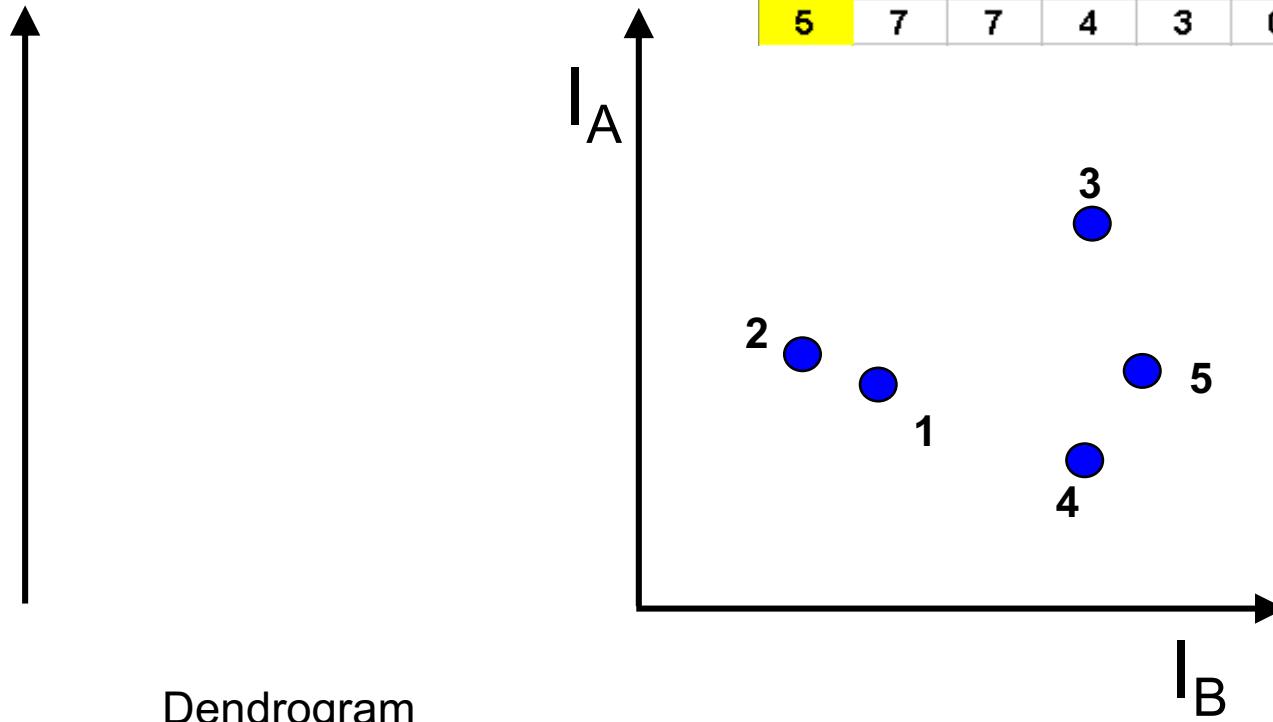
Procedure

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



Example

Cluster distances

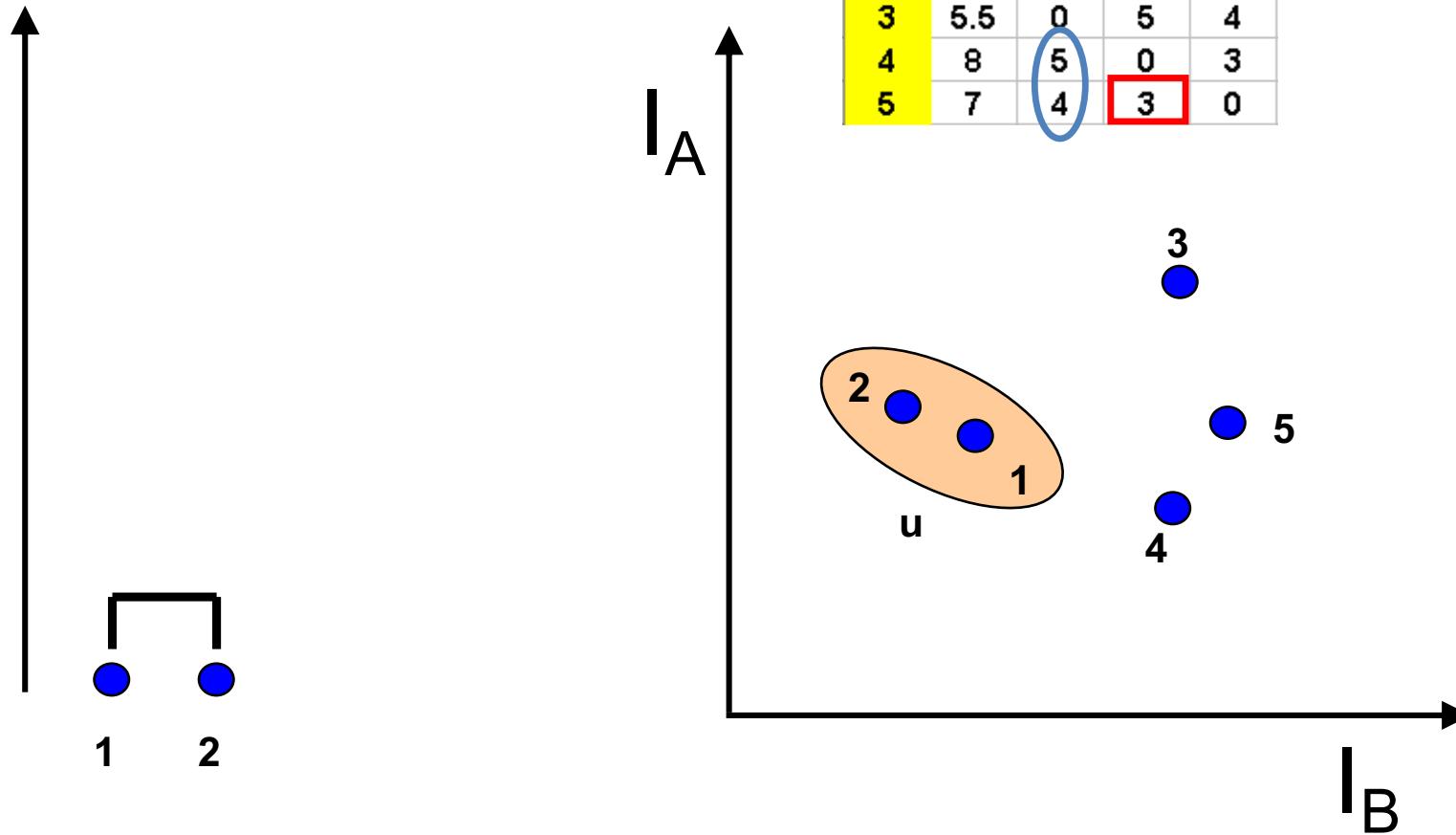


10
01
101

functional genomics center zurich
0
1
0
010 01
101 10
010 01
01 1
01 1

Example

Cluster distances

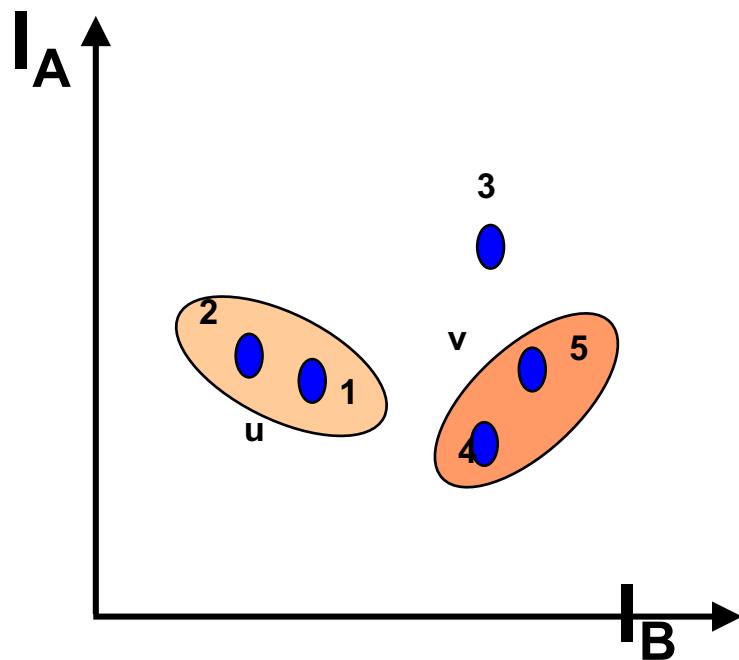
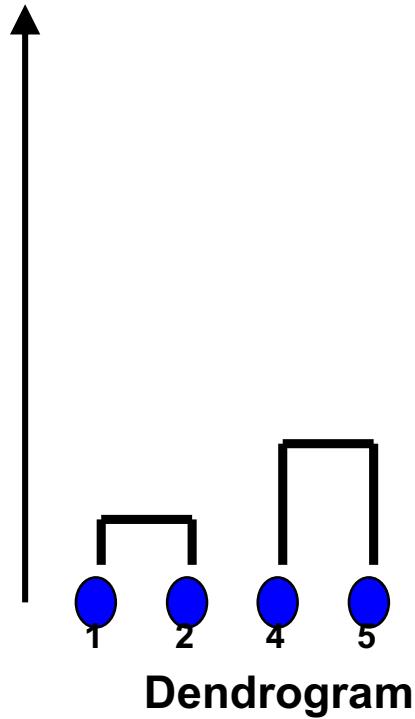


Dendrogram

Example

Cluster distances

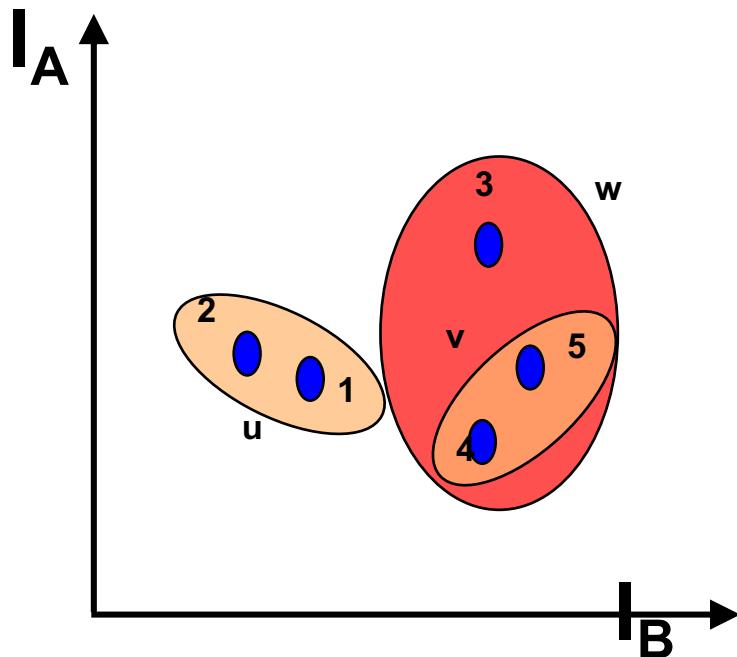
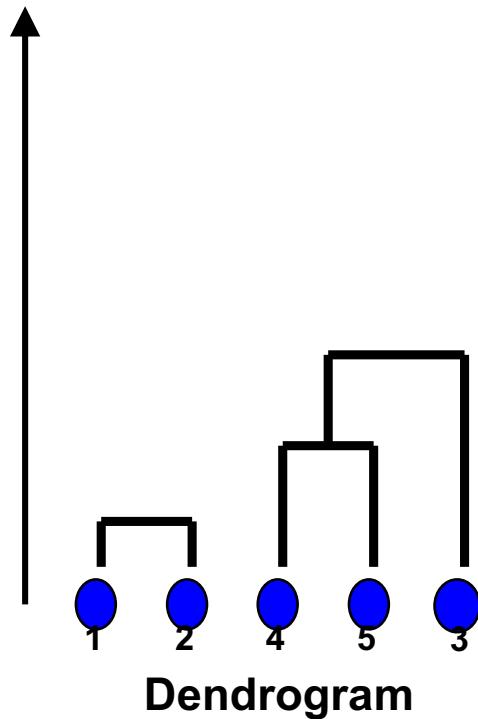
d(ij)	u	3	v
u	0	5.5	7.6
3	5.5	0	4.5
v	7.5	4,5	0



Example

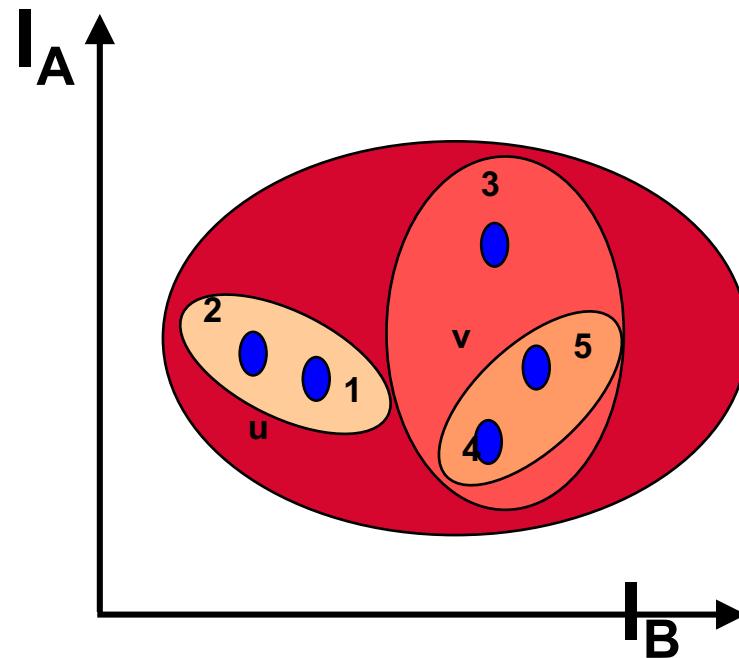
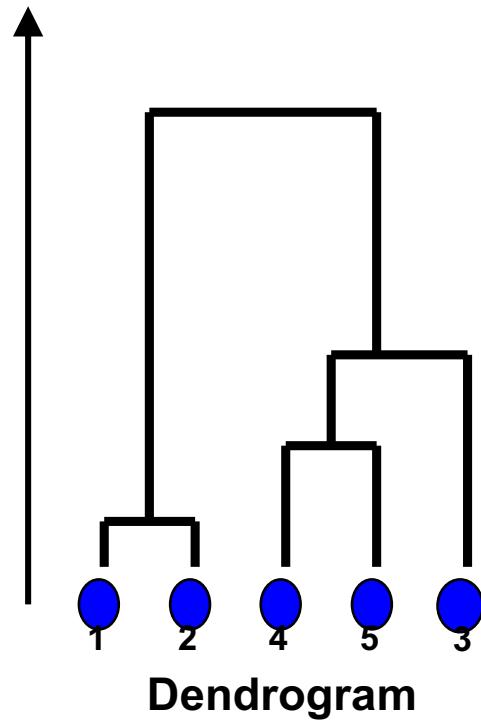
Cluster distances

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



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)



10
01
101



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:

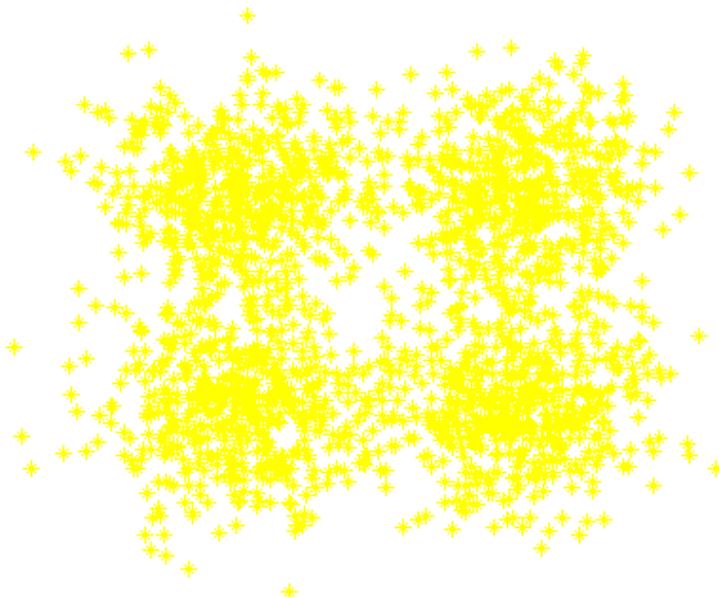
- number of clusters
 - distance measure



10
01
01
101

functional genomics center zurich
010 01 01
101 10 10
010 01 01
01 10 01
01 10 01
01 10 01

K-means Example



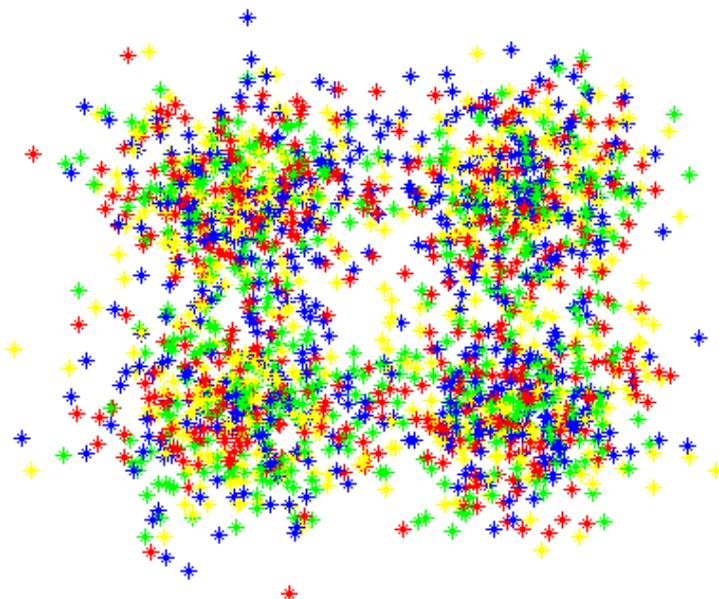


University of
Zurich UZH

10
01
101

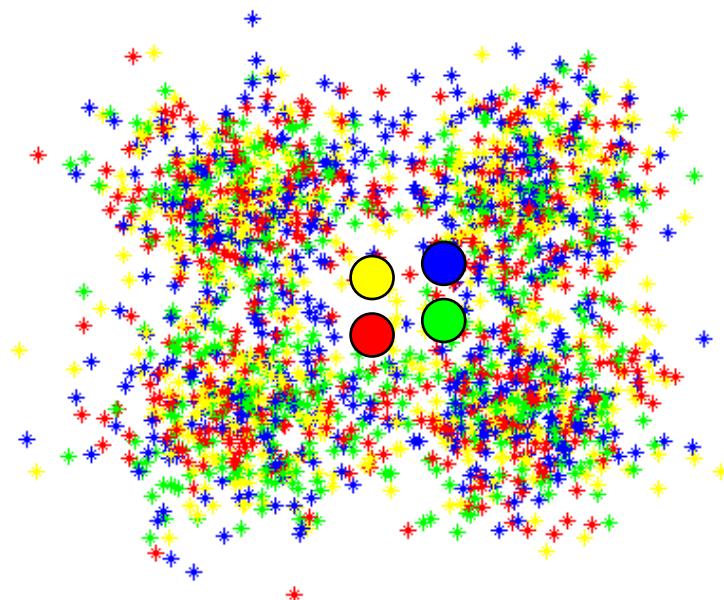
functional genomics center zurich
010 01
101 10
010 01
010 01
01 1
0 1
1 0
0 1 1

K-means: Initialization





K-means: Computing cluster centers





University of
Zurich UZH

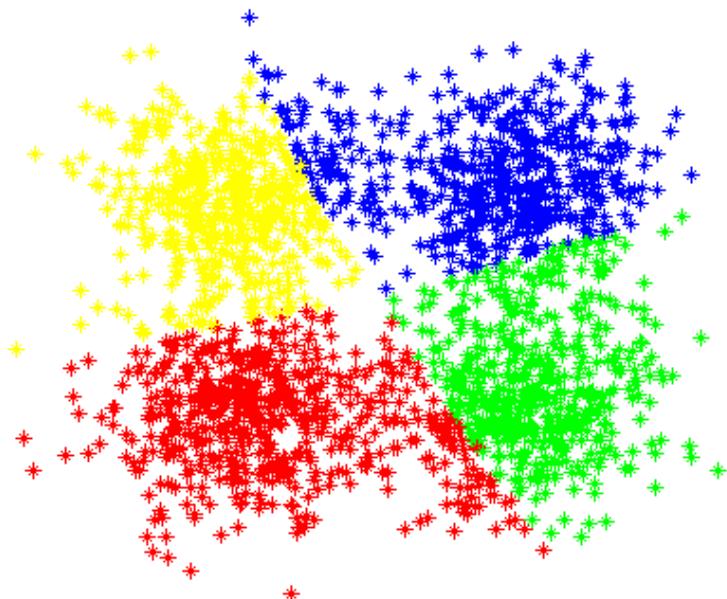
10
01
101

functional genomics center zurich

010 01
101 10
010 01
010 01

0
1
0
1
0
1
1

K-means: Reassignment of points





Comparison of Clustering Methods

- Computing time:
- Hierarchical clustering
 - $O(n^2 \log(n))$
- K-means clustering
 - t: number of iterations
 - k: number of clusters
 - $O(k t n)$
- Memory requirements:
- Hierarchical clustering
 - $O(n^2)$
- K-means clustering
 - $O(kn)$

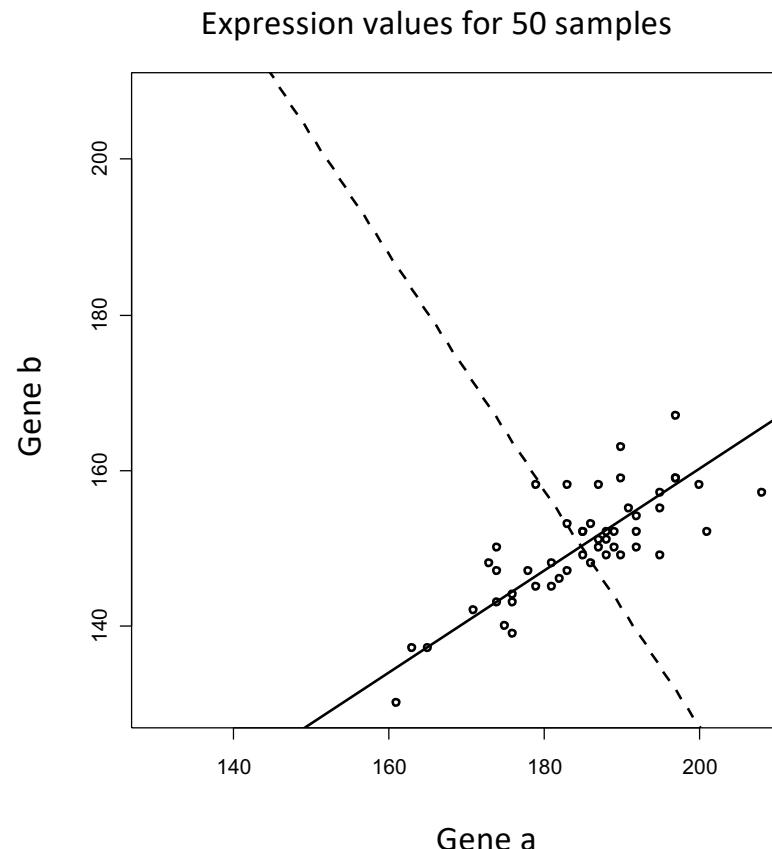
Note:

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



10
01
101

Principal Component Analysis

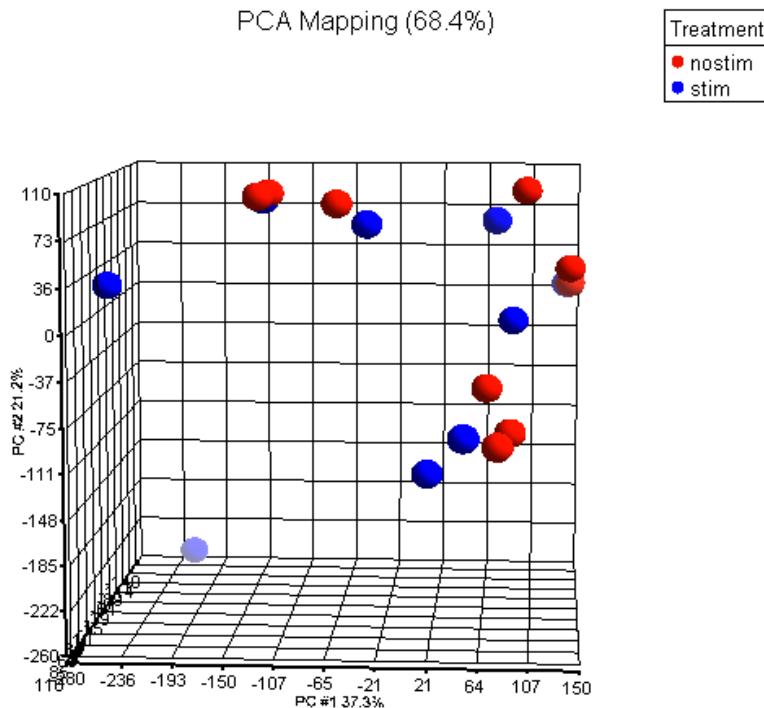
Procedure:

- Center the matrix: Subtract average
- Compute covariance matrix of the centered matrix (gives an $n \times n$ 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

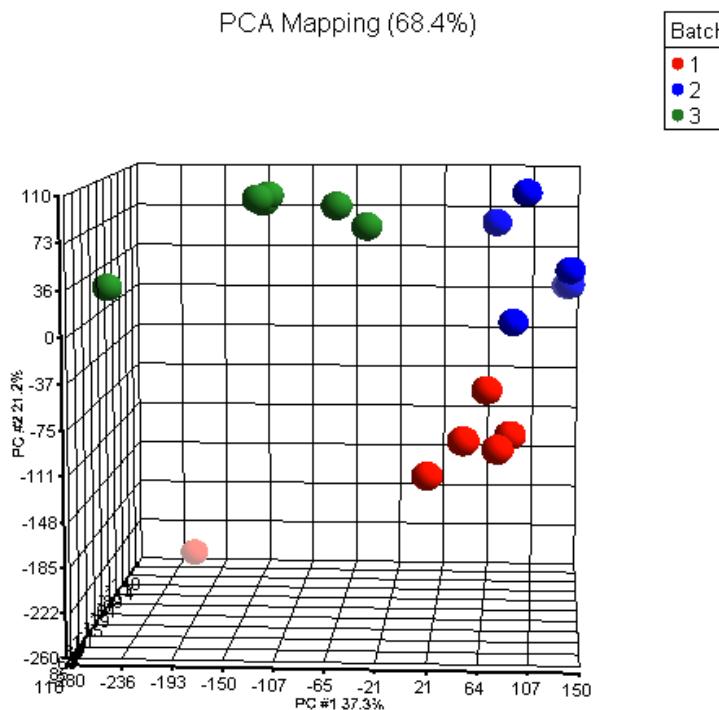
- Plot showing the 18 samples in the PCA coordinates
- Coloring by treatment of the samples
 - stimulated
 - not stimulated
- The samples do not separate
- There are two outliers on the left



10
01
101functional genomics center zurich
01 1
010 01
101 10
010 01
01 1
01 0
10 0

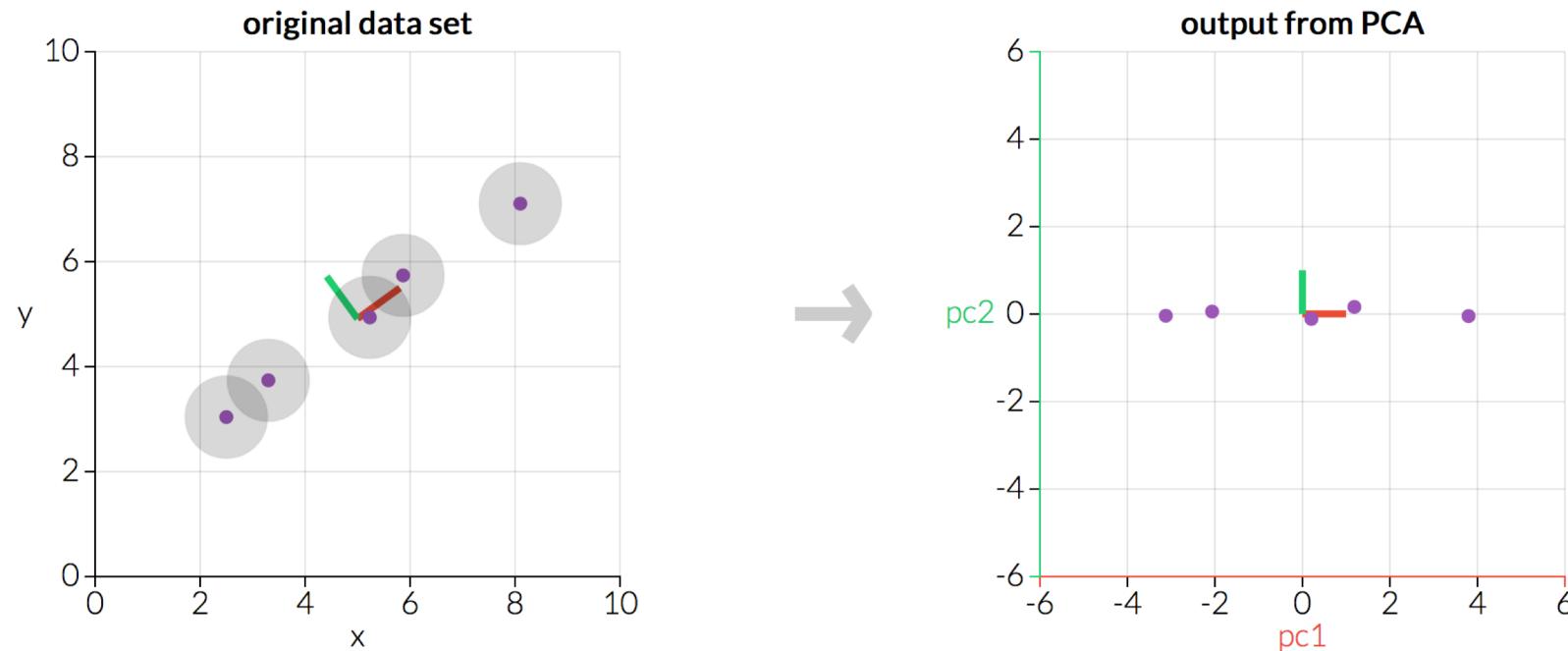
Example: Stimulation experiment

- 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

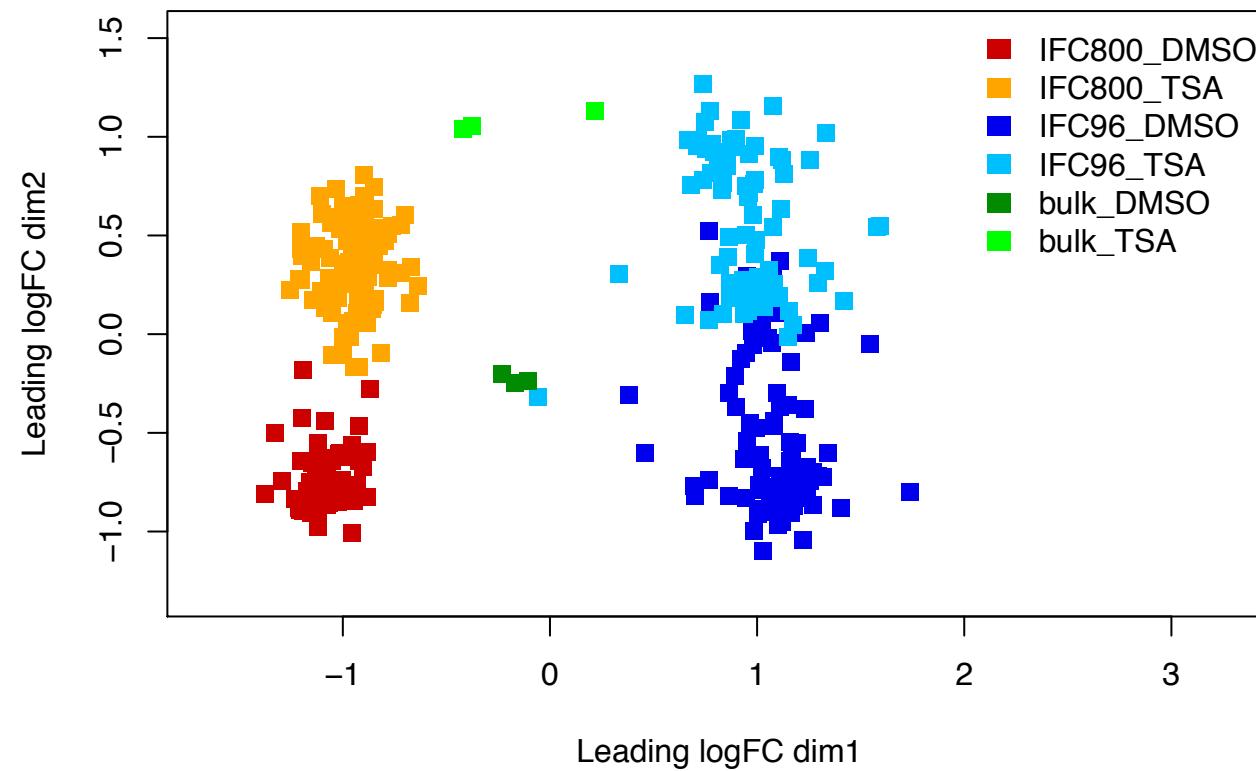
- <http://setosa.io/ev/principal-component-analysis/>



Example of multi-dimensional scaling

Identify consistency

Assess effect sizes



Example of exploratory analysis

