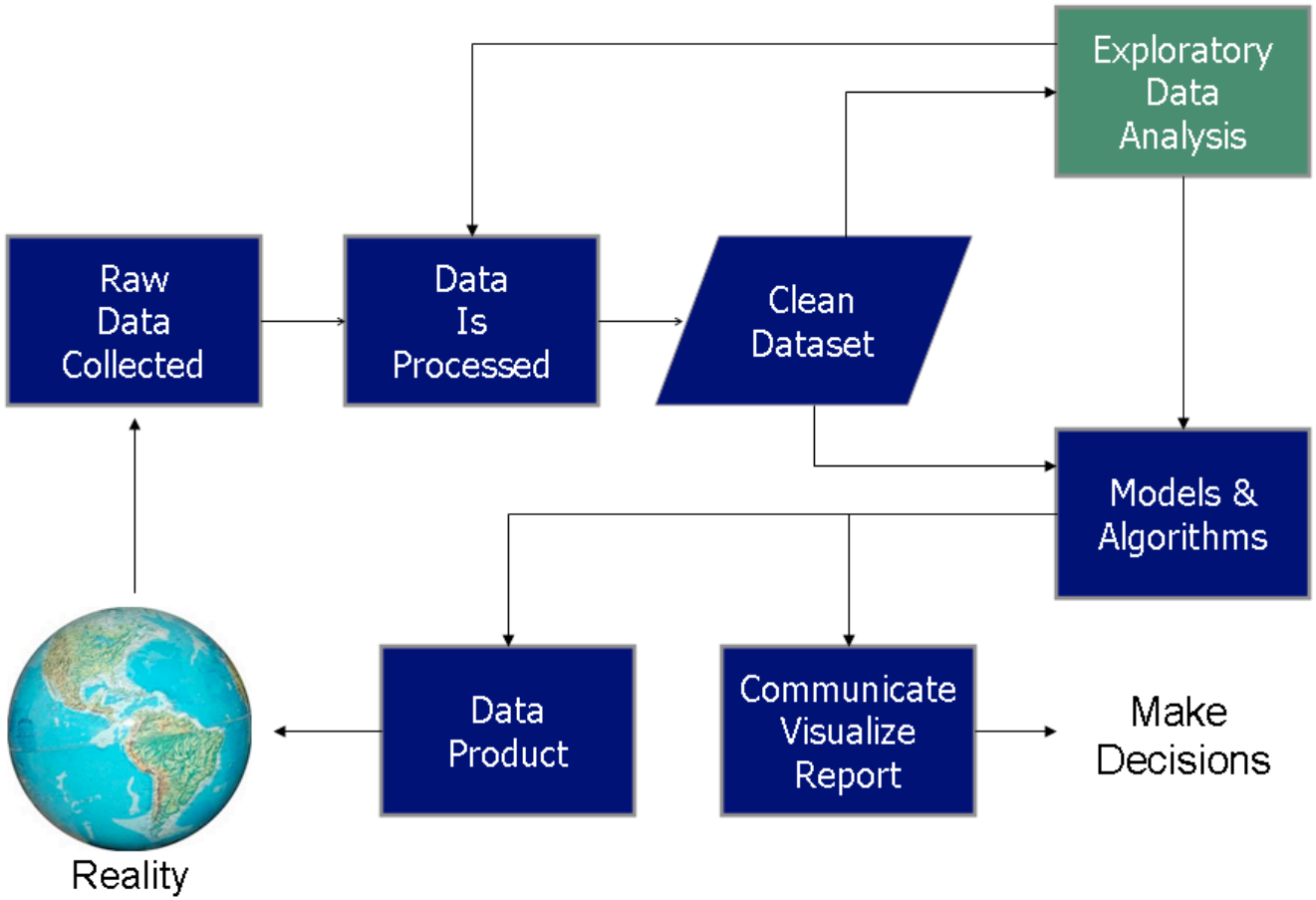
# Exploratory Data Analysis

Hubert Rehrauer

# Data Science Process



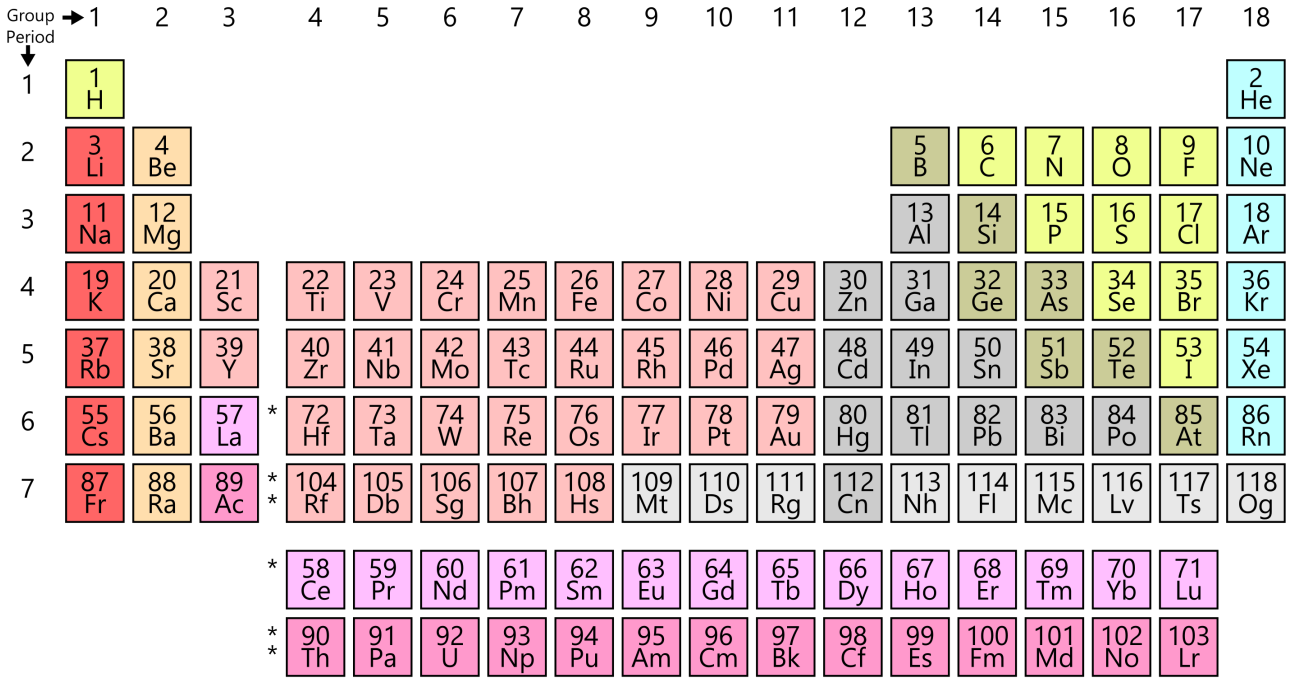

https://en.wikipedia.org/wiki/Exploratory_data_analysis

# Dot map of cholera cases

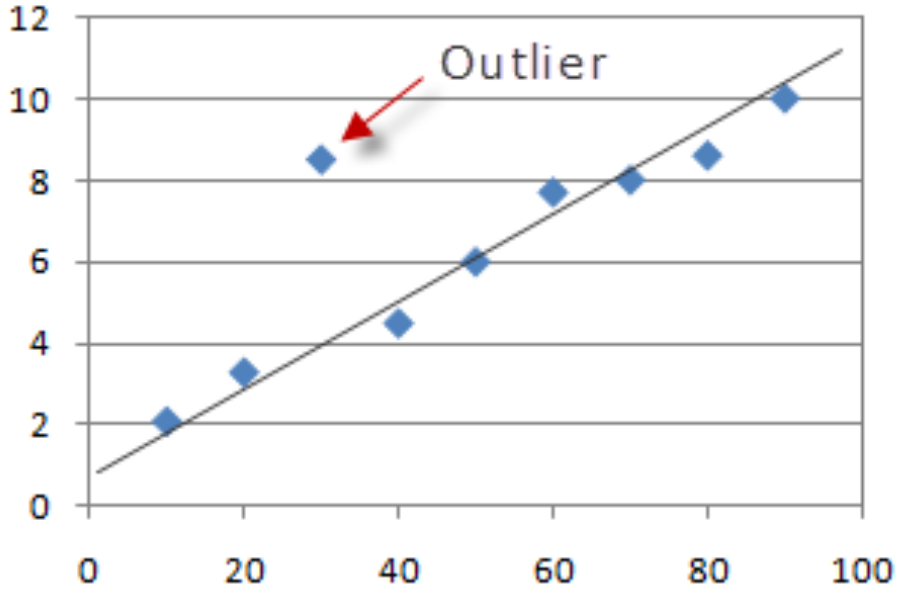

John Snow, 1850, Source: https://devopedia.org/exploratory-data-analysis

# Periodic Table

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



# Simple visualization to find outliers



# Goals and methods of EDA

**Goals:**

- 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

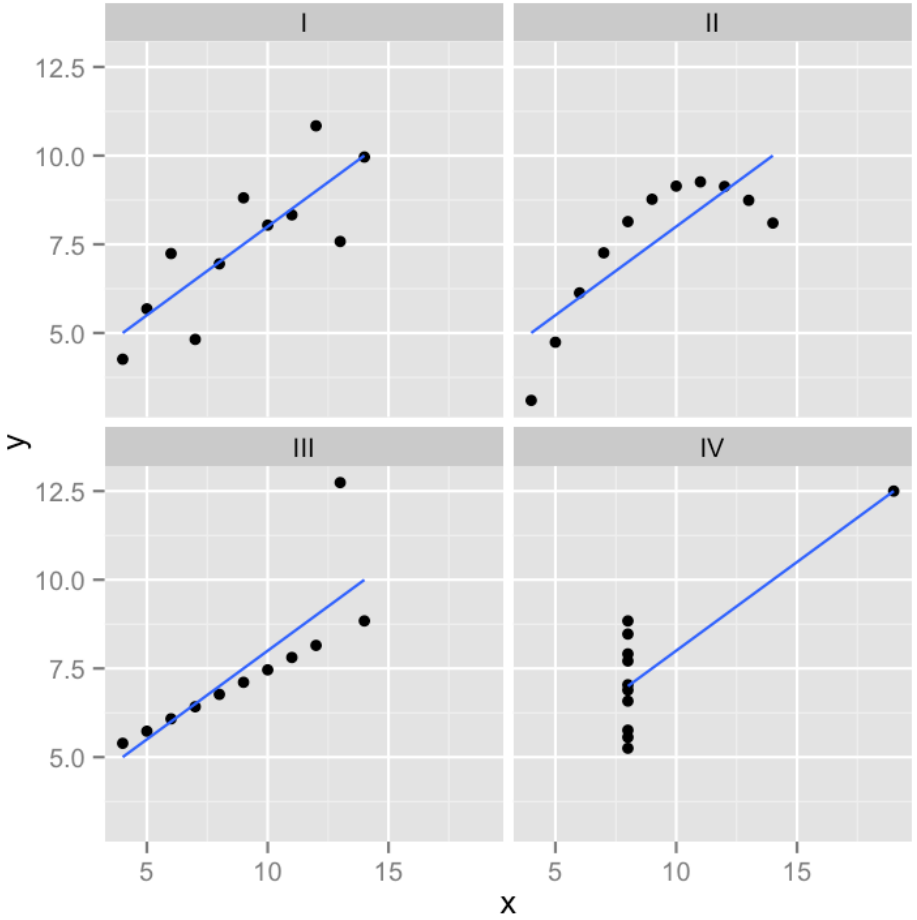
**Methods**:

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

# Anscombe's quartet

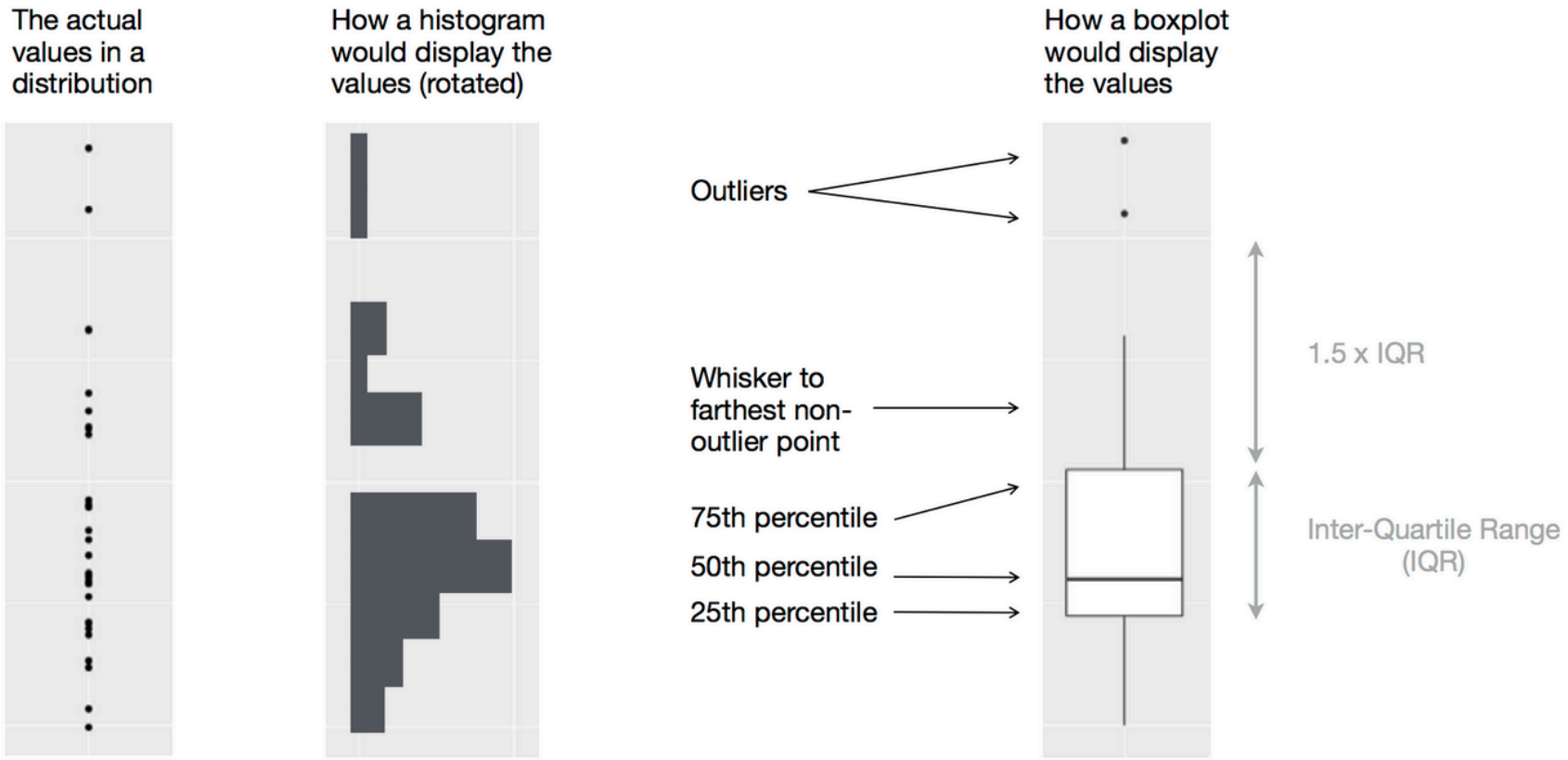
- All 4 datasets have nearly identical
  - mean
  - standard deviation
  - skewness & kurtosis

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

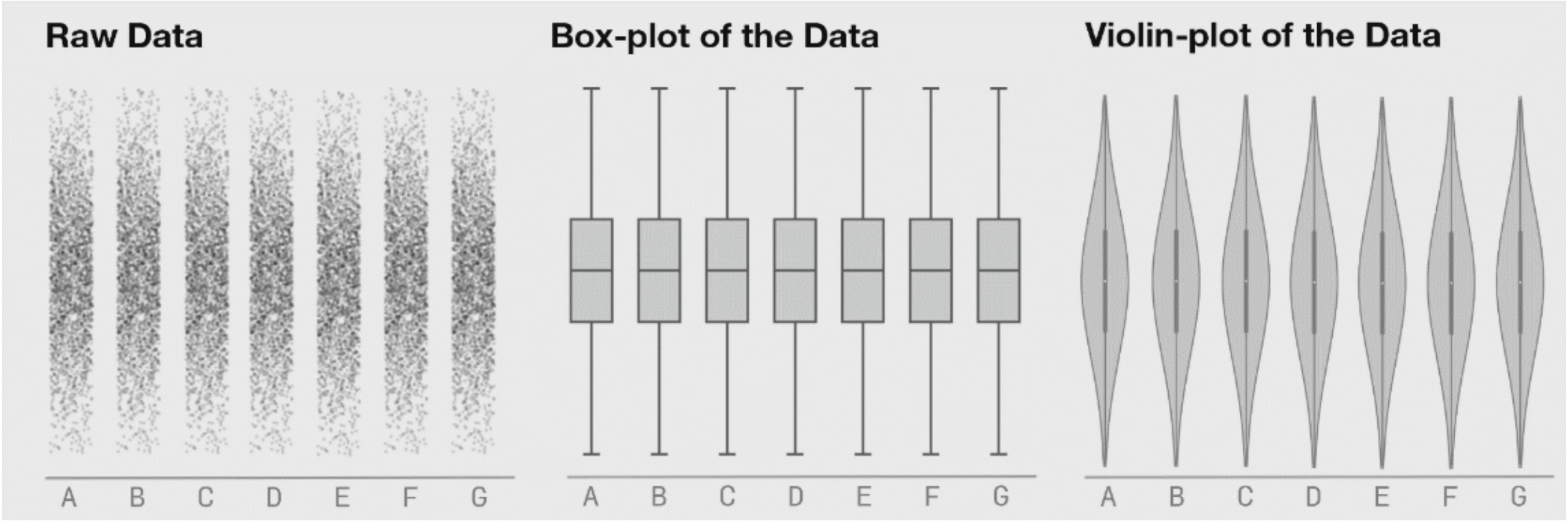



Turner 2016

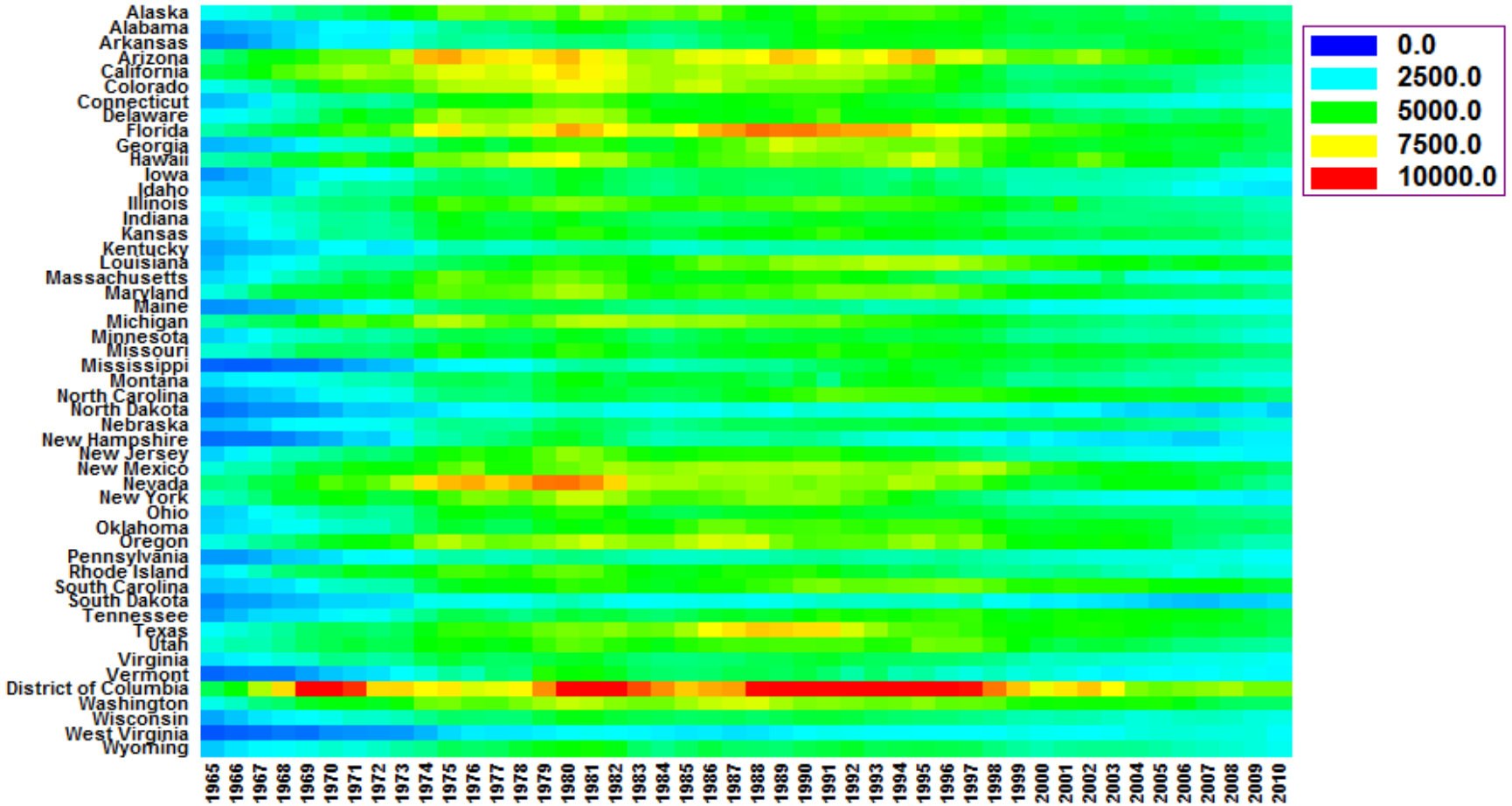
# Example of misleading boxplot



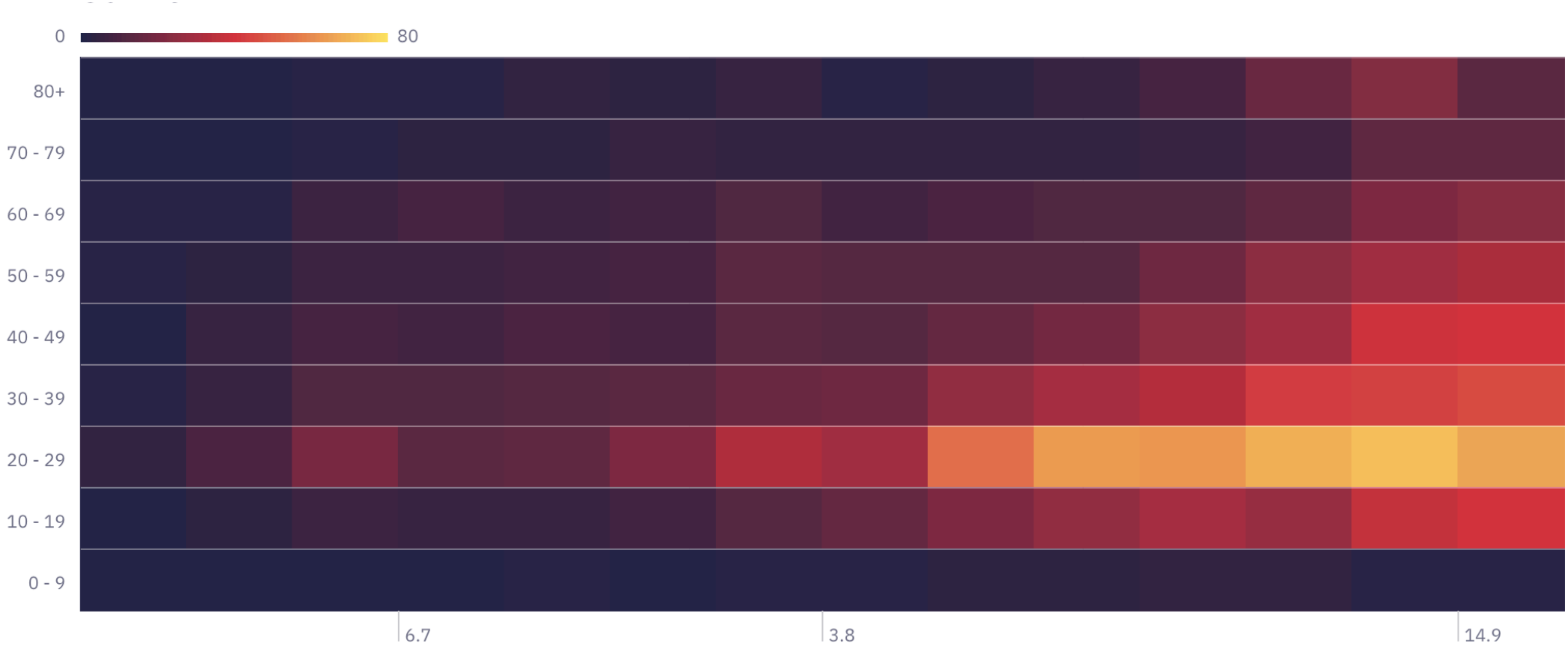

https://r4ds.had.co.nz/exploratory-data-analysis.html


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



# Age distribution of COVID cases over time

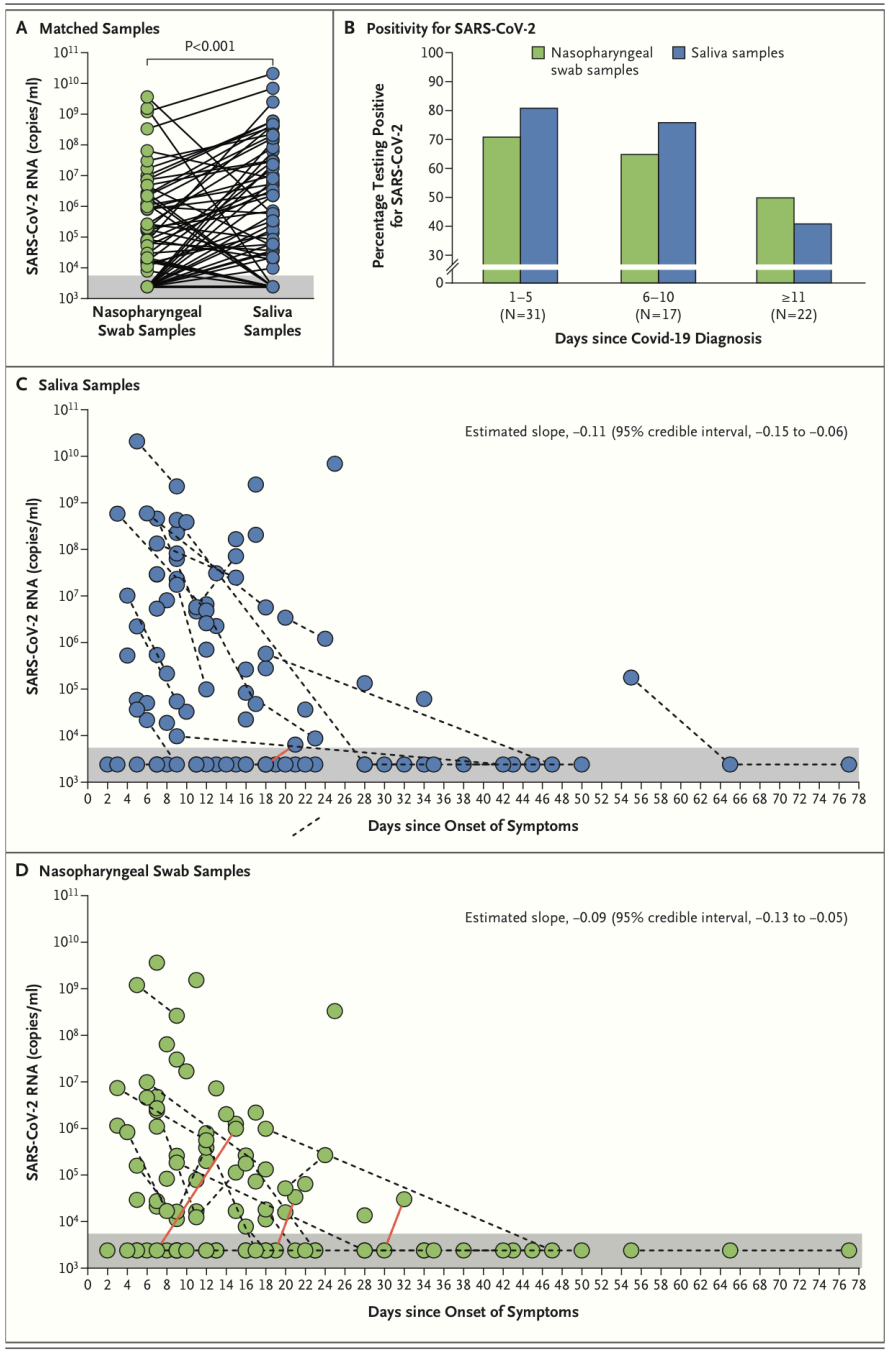


# Age distribution of COVID cases over time

# Covid: Nasal swabs vs saliva

- Can SARS-Cov2 be detected in Saliva?

- Both methods detect the virus
- Plot A triggers a data exploration question: Why are there so many subjects where the two methods disagree so strongly.




https://www.nejm.org/doi/full/10.1056/NEJMc2016359

# Mutations in Vitamin D Receptor and sex hormone levels in elderly people

| Attribute name | Description |
|---|---|
| AGE | age in years |
| HEIGHT | height given in [cm] |
| WEIGHT | weight in [kg] |
| WAISTLINE | waistline given in [cm] |
| HIP.GIRTH | hip girth given in [cm] |
| BMI | the body to mass index [kg/m$^2$] |
| FAT | Amount of body fat as percentage of body weight [%] |
| CHOL.HDL | Cholesterol serum level—High Density Lipoprotein [mg/dl] |
| CHOL.LDL | cholesterol serum level—Ligh Density Lipoprotein [mg/dl] |
| CHOL.TOTAL | total level of cholesterol [mg/dl] |
| TGC | serum level of triglycerides [mg/dl] |
| GLUCOSE | Serum Glucose level [mg/dl] |
| INS | serum level of insulin [μIU/ml] |
| TESTOSTERONE | serum level of testosterone [nmol/l] |
| ESTRADIOL | serum level of Estradiol [pmol/l] |
| DHEA.S | serum level of Dehydroepiandrosteron [ng/dl] |
| SHGB | serum level of sex hormone binding globulin [pmol/l] |
| FAI | Free Androgen Index defined as the ratio of total testosterone to SHBG × 100 [19] |
| FEI | Free Estradiol Index defined as the ratio of total estradiol to SHBG × 100 [19] |
| FSH | Serum Follicle-Stimulating Hormone level [IU/l] |
| ICTP | serum level of carboxy-terminal cross-linked telopeptide of type I collagen [mg/l] |
| OPG | serum level of osteoprotegerin [pmol/l] |
| VITAMIN.D | serum level of Vitamin D [ng/ml] |

https://doi.org/10.1371/journal.pone.0201950.t001

| Attribute name | Description |
|---|---|
| AGE.GROUP | age in discretized groups (5 year bins) |
| CG1.IDENTIFIED.DIABETES.YES | binary; 1 if observed |
| CITY.SIZE | city size bins: countryside, population < 20 thousand, 20–50 th |
| HYPERANDROGENISM.YES | binary; 1 if observed |
| HYPERTENSION.YES | binary; 1 if observed |
| INSOLATION.YES | binary; 1 if in summer and spring |
| MACROREGION | 6 binary attributes: 'north', 'east', 'south', 'central', 'north-west', |
| OBESTIY_PHENO_FLMHO | binary; obesity phenotype—metabolic healthy obesity [20] |
| OBESTIY_PHENO_FLMONW | binary; obesity phenotype—methabolic obesity normal weight [ |
| OBESTIY_PHENO_FLOMWD | binary; obesity phenotype—obesity methabolic weist disease [2 |
| OBSETITY_PHENOOBZM | binary; obesity phenotype–adjustment of FLMHO for Polish pc |
| OBSETITY_PHENOOZZM | binary; obesity phenotype–adjustment of FLMONW for Polish |
| YEAR_SEASON | 4 binary attributes: 'winter', 'spring', 'summer', 'autumn' |

https://doi.org/10.1371/journal.pone.0201950.t002

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0201950

# Mutations in Vitamin D Receptor and sex hormone levels in elderly people

- 515 participants ⟸ Subjects
- 44 attributes ⟸ Explanatory variables

- Which attributes have an impact on the hormone levels?
- Are there attributes that are correlate with the receptor mutations? If yes, these are confounding factors

| | AGE | SHGB | FSH | TESTOSTERONE | ESTRADIOL | DHEA.S | FAI | FEI | HEIGHT | WEIGHT | FAT | WAISTLINE | HIP.GIRTH | BMI | C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 92 | 117.8 | 53 | 0.12 | 4.4 | 300 | 0.353174 | 0.0137 | 151 | 41.1 | 25.7 | 94 | 90 | 18.025525 | |
| 3 | 58 | 67.2 | 76.5 | 0.15 | 3.7 | 656 | 0.773883 | 0.0202 | 170 | 69 | 34 | 92 | 102 | 23.875432 | |
| 5 | 67 | 53.7 | 94.9 | 0.2 | 4 | 109 | 1.291247 | 0.0273 | 143 | 60.1 | 33.9 | 106 | 90 | 29.39019 | |
| 6 | 75 | 51.6 | 9.6 | 3.31 | 10.2 | 609 | 22.23986 | 0.0725 | 174 | 74.1 | 29 | 100 | 104 | 24.474832 | |
| 8 | 59 | 34.5 | 46.8 | 0.32 | 3.6 | 565 | 3.215768 | 0.0383 | 164 | 70 | 38 | 89 | 103 | 26.026175 | |
| 9 | 83 | 55.5 | 56.3 | 0.17 | 3.9 | 750 | 1.061963 | 0.0257 | 152 | 61.2 | 31.3 | 87 | 99.5 | 26.48892 | |
| 10 | 77 | 73.8 | 21.5 | 7.56 | 19.7 | 1298 | 35.51560 | 0.0979 | 162 | 54.9 | 17.3 | 80 | 96 | 20.919067 | |
| 13 | 57 | 102.4 | 116.9 | 0.41 | 5.8 | 632 | 1.388154 | 0.0207 | 154 | 61.5 | 33.8 | 83 | 104 | 25.93186 | |

# 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
- …

- 10s – 100s subjects
- 1 – 10s attributes

Additional:

- ~25’000 genes as explanatory variables

| gene_name | Untreated_0( | Untreated_0( | Untreated_0( | TSA_001 | TSA_002 | TSA_003 | TFNA_001 | TFNA_002 | TFNA_003 | DMSO_001 | DMSO_002 | DMSO_003 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Actb | 893495 | 837711 | 832581 | 1472528 | 981134 | 1120678 | 991944 | 1101636 | 1229135 | 1139451 | 1051921 | 1207426 |
| Cd74 | 995494 | 787673 | 875016 | 949493 | 825139 | 818025 | 779110 | 901570 | 890373 | 875348 | 797970 | 746900 |
| Psap | 208993 | 174402 | 271321 | 327882 | 329209 | 287670 | 328943 | 345473 | 316682 | 443061 | 422382 | 375204 |
| H2-Ab1 | 426933 | 343484 | 417642 | 449921 | 392453 | 392767 | 393112 | 439128 | 425410 | 414268 | 385411 | 344702 |
| Wdfy4 | 179716 | 159346 | 203653 | 247430 | 204304 | 232487 | 207341 | 233608 | 226601 | 180468 | 192222 | 206202 |
| Eef1a1 | 199487 | 170600 | 184701 | 237333 | 203035 | 214763 | 187098 | 236834 | 225352 | 204775 | 182308 | 199842 |
| H2-Eb1 | 216846 | 199178 | 230637 | 265651 | 229147 | 227300 | 215171 | 255286 | 242801 | 238764 | 218298 | 192494 |
| Tmsb4x | 173784 | 146600 | 178973 | 254074 | 193715 | 218451 | 183551 | 212768 | 222196 | 184576 | 177274 | 183352 |

# Data Generation Workflow

Experiment

⇓

Omics method

⇓

Raw “machine” data

⇓

Raw quantitative data → 

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

→ 

- Quantitative sample differences

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

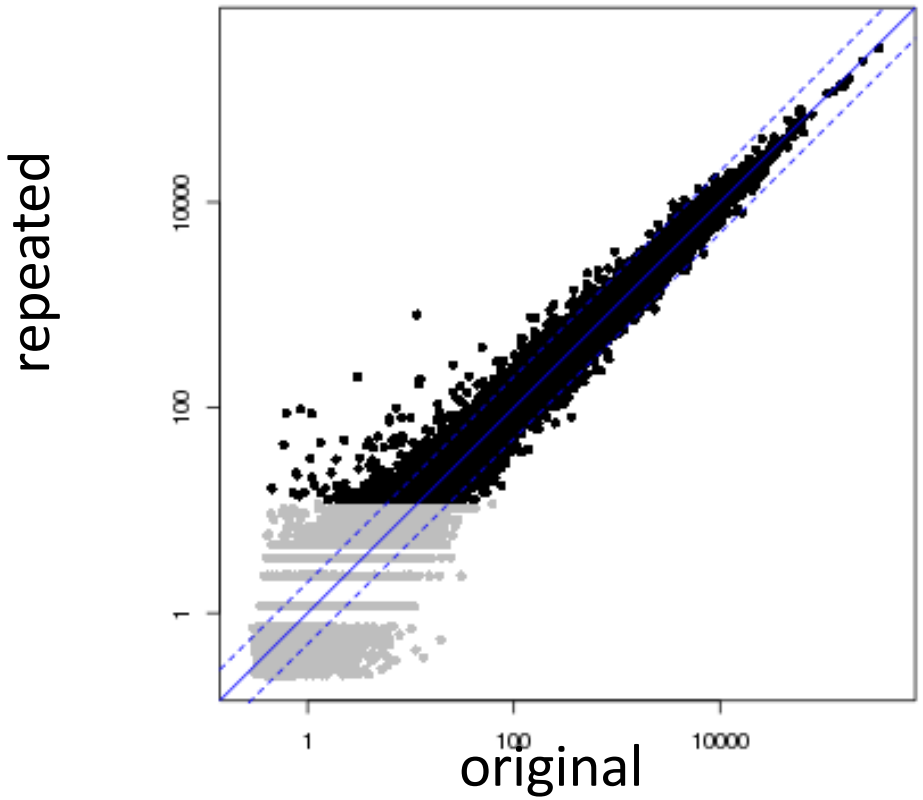
# Technical characteristics

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

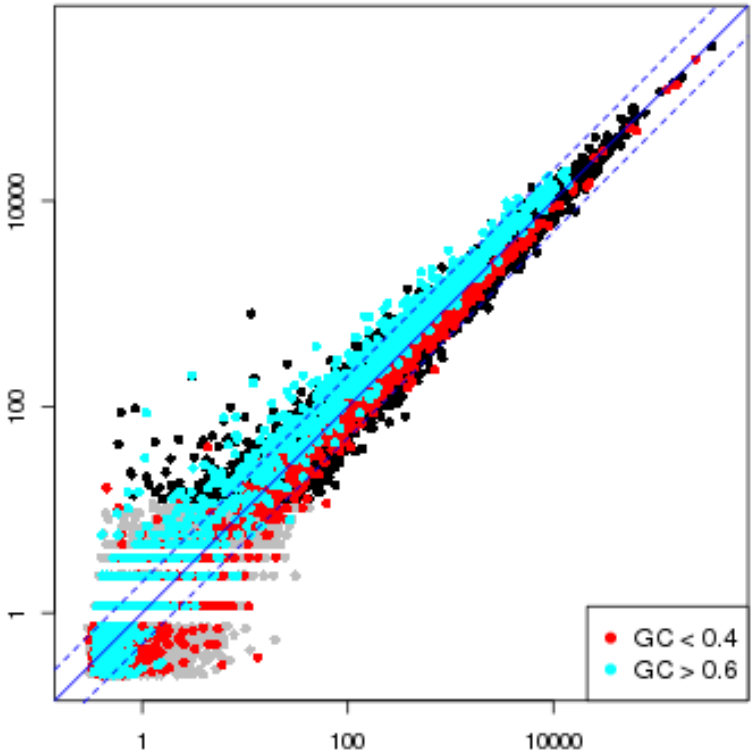
- Zero Measurements
  - additive background signal
  - zero value can have different explanations
    - technical failure to detect
    - value is truly zero
- Variability
  - Non-Gaussian noise

# Example of Quantitative Data

Comparison of a repeated experiment




Systematic effect of GC content on quantitative values



# 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 of RNA-seq data

- 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:

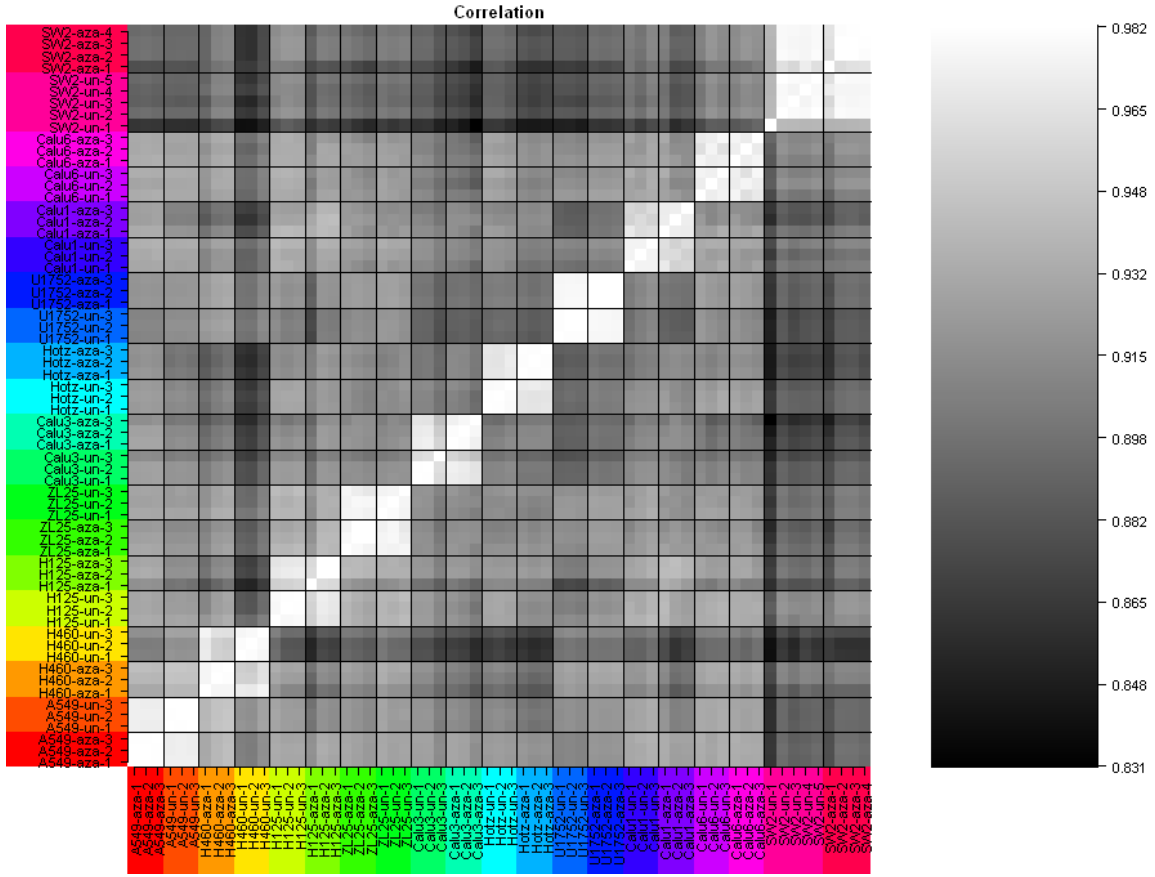
$$s(x,y) = \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i - \bar{x})^2 \times \sum_{i=1}^{p}(y_i - \bar{y})^2}}$$

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

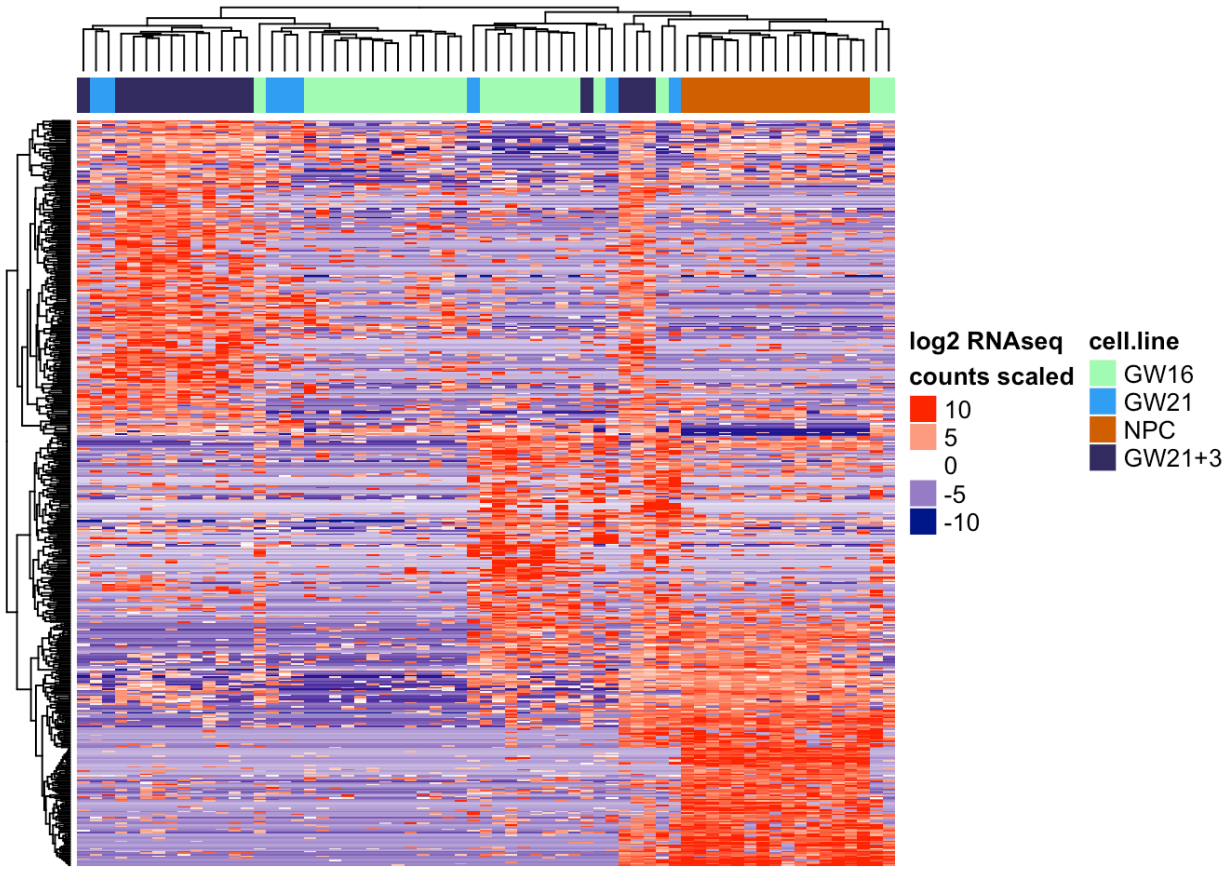
# Correlation matrix for samples

The matrix shows the correlation for all sample pairs.

This gives a quick overview on the presence of outliers.

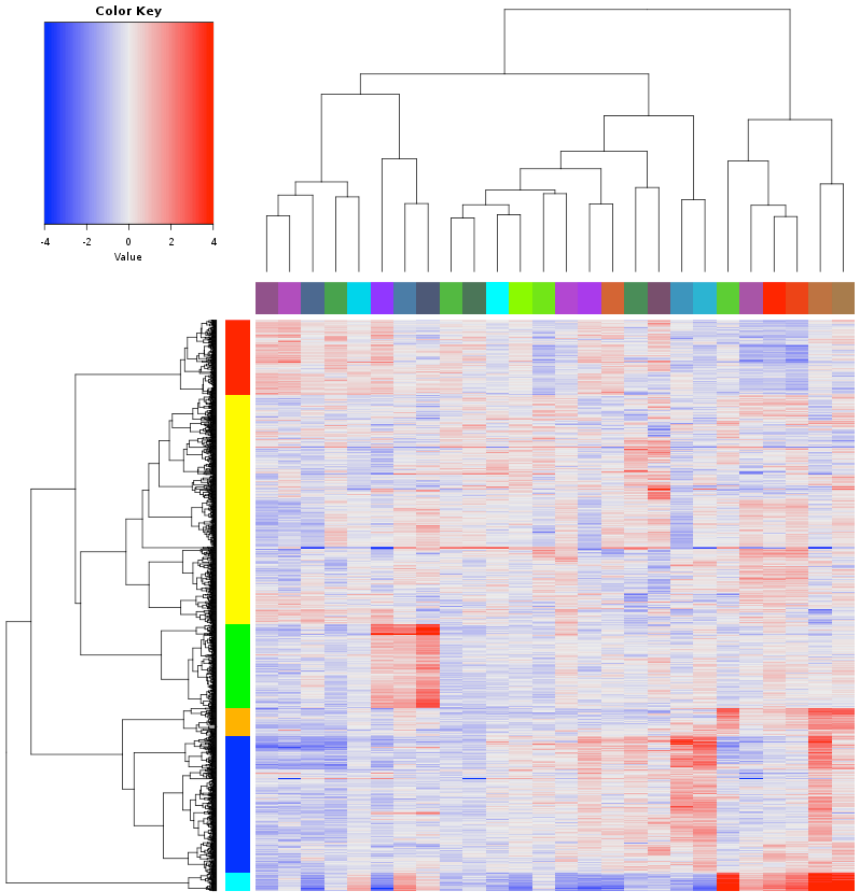


# Example: Identifying Outlier Samples and Expression Structure




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

# Hierarchical Clustering

## Goal

- Grouping of samples according to similarity

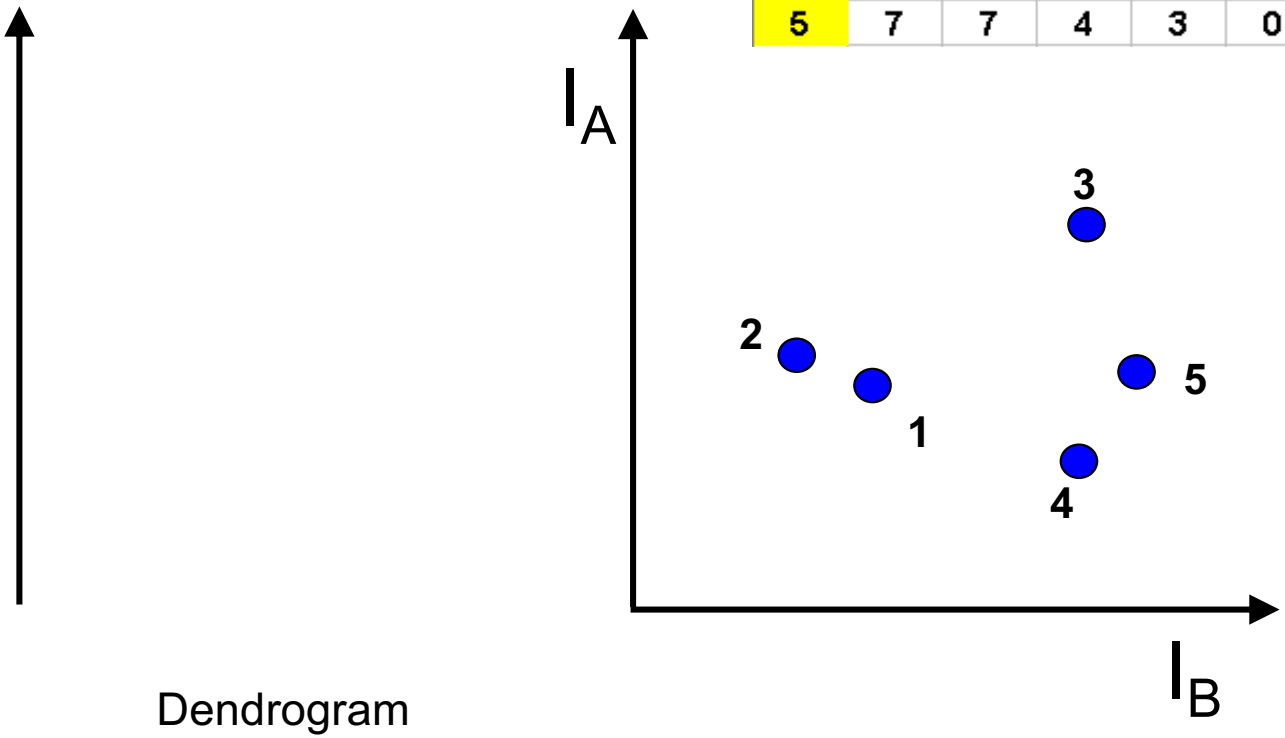
## Procedure

- 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 |

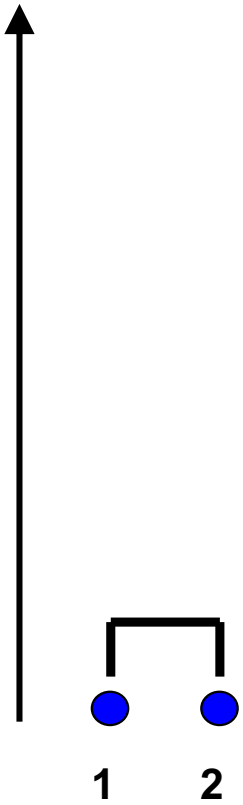



Dendrogram

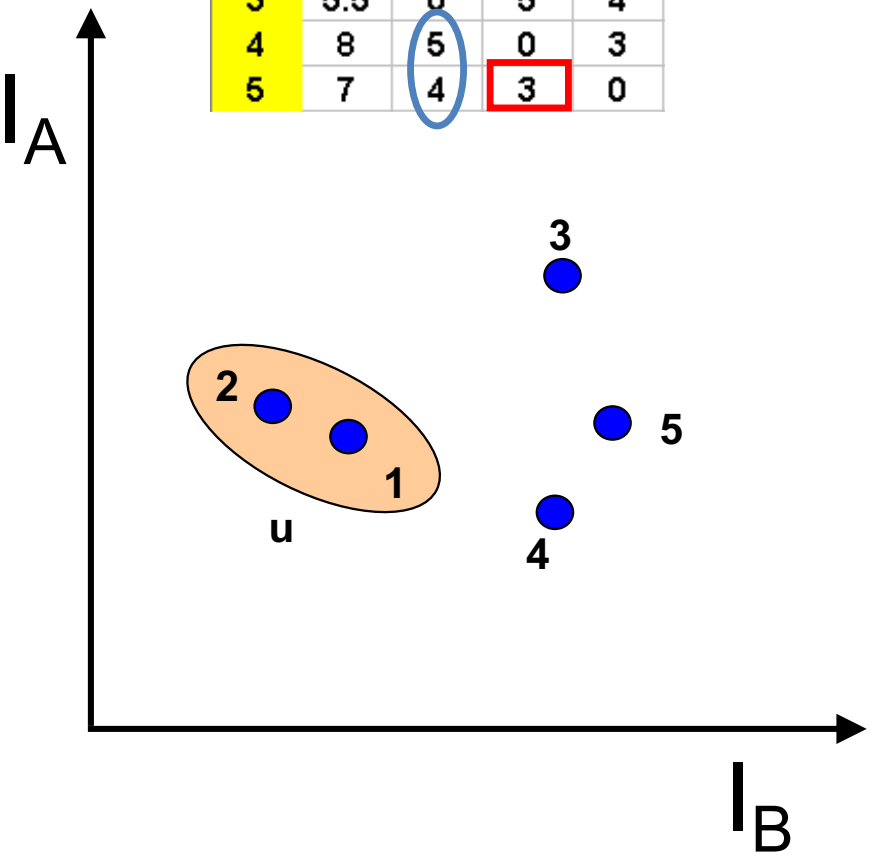
## Example

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 |



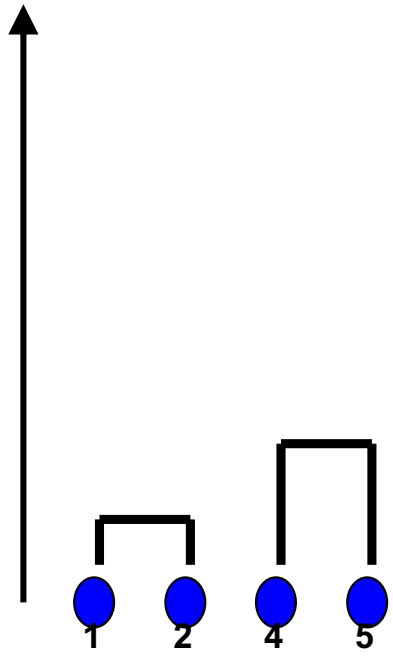

Dendrogram



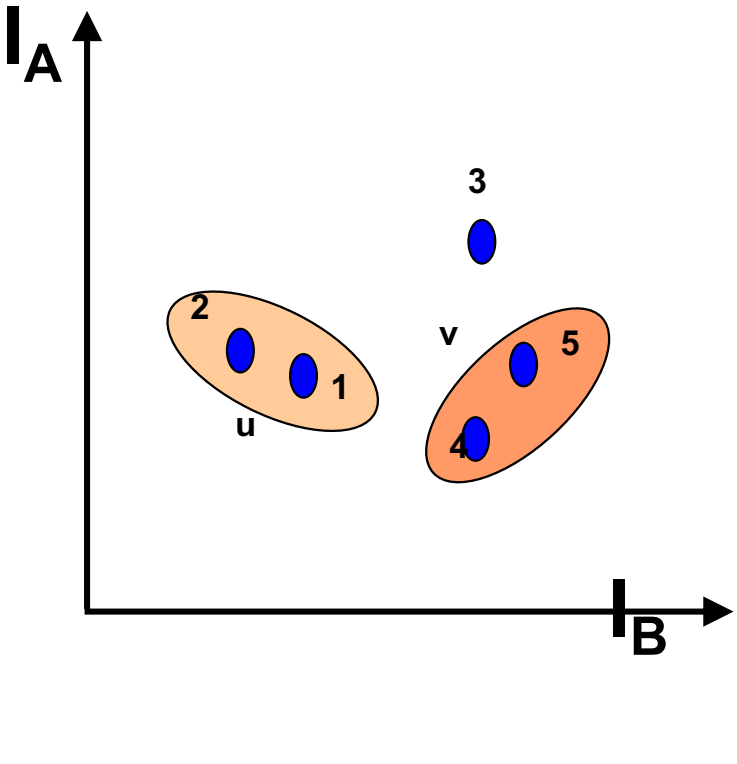
## Example

### 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 |



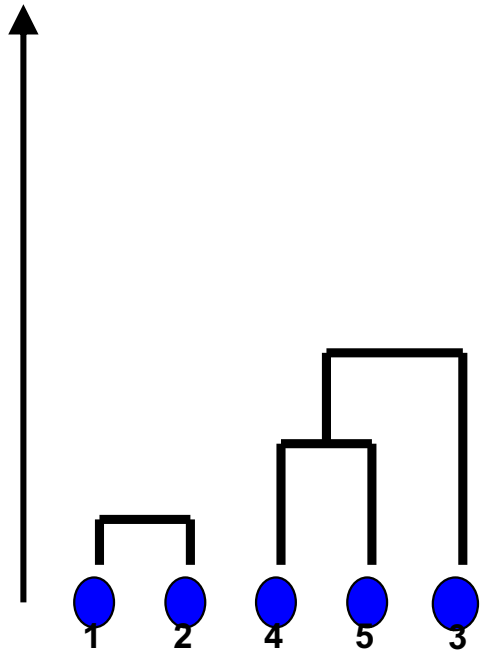

Dendrogram



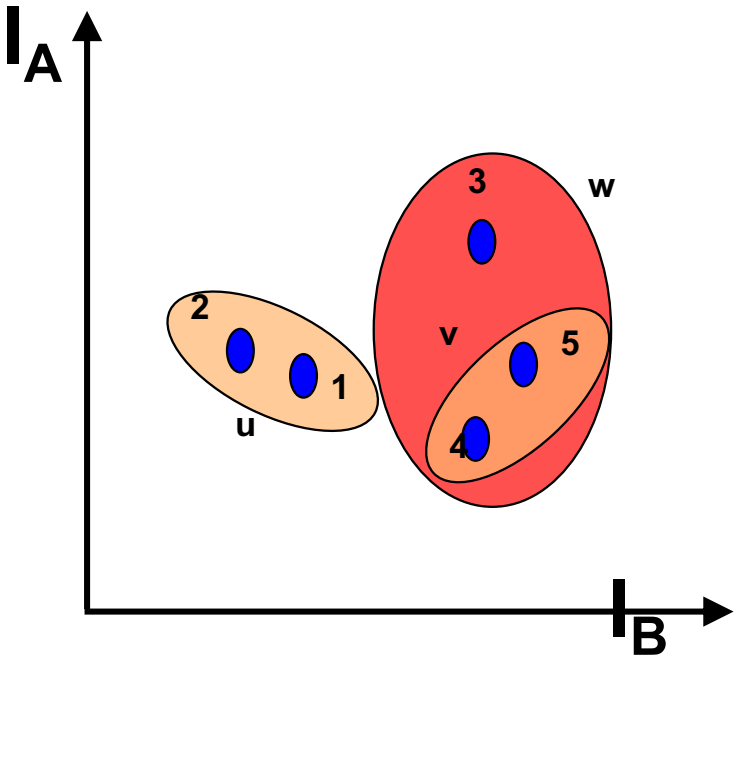
## Example

### Cluster distances

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

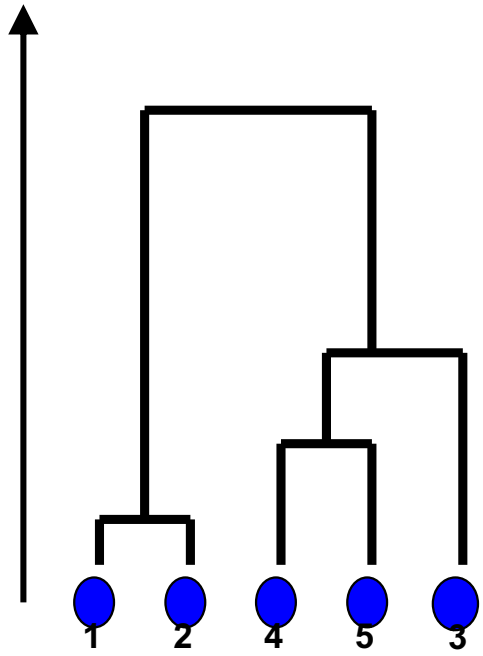



Dendrogram

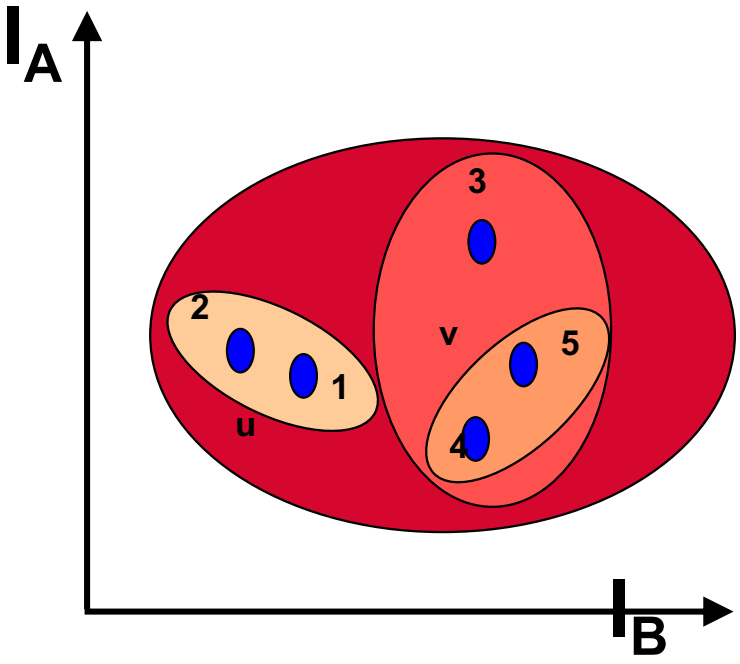


## Example

### Cluster distances




Dendrogram



# 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:

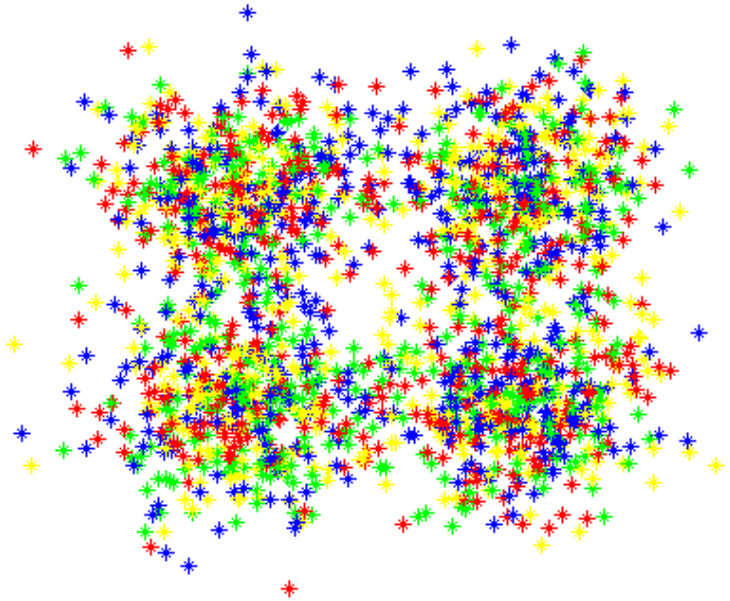
- Finds clusters that minimize intra-cluster variance

Parameters:

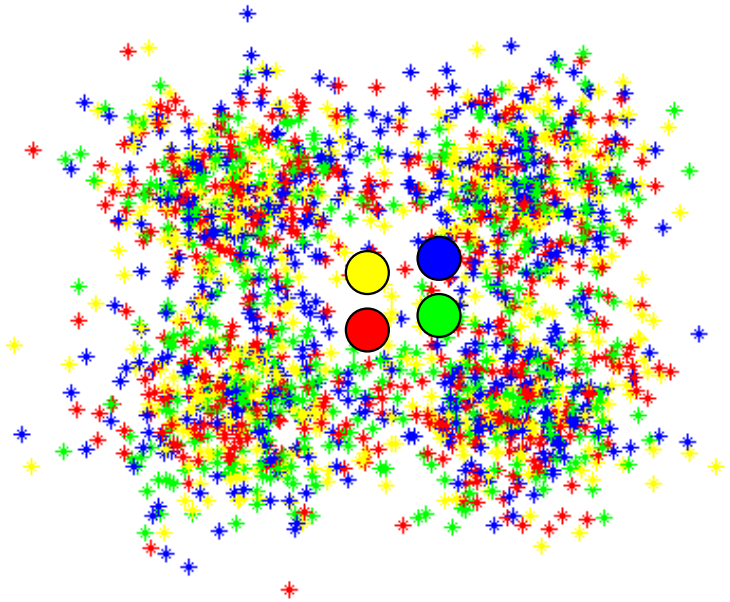
- number of clusters
- distance measure

# K-means Example

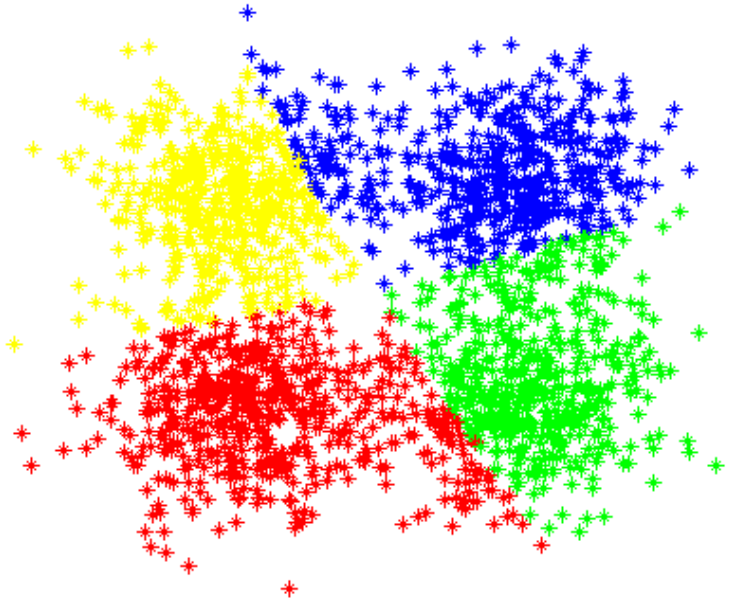
# K-means: Initialization



# 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 )

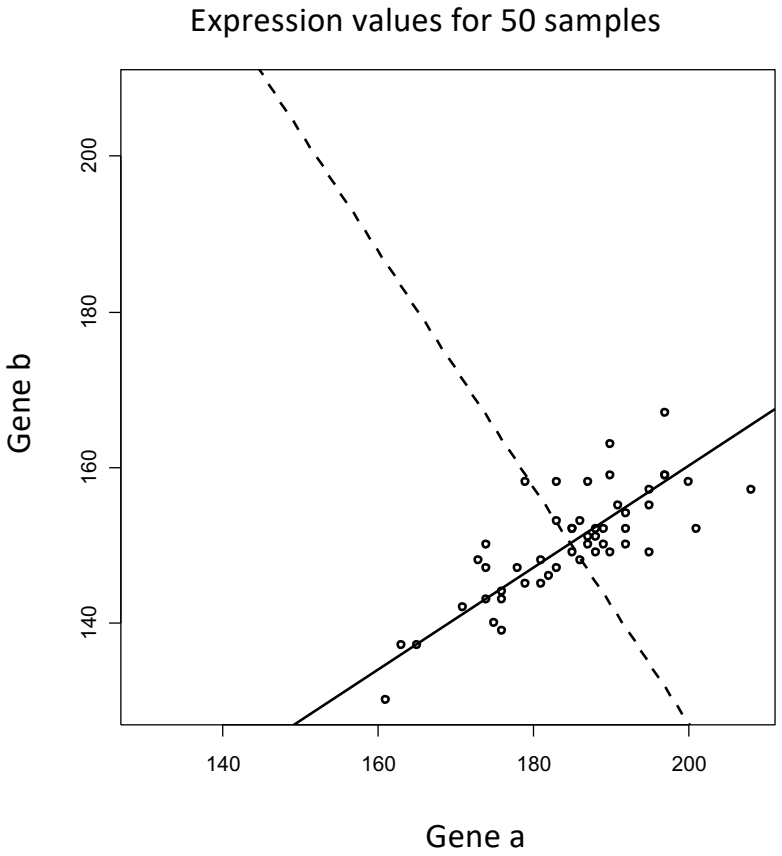
- Memory requirements:
- Hierarchical clustering
  - O( n2 )
- K-means clustering
  - O( kn )

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



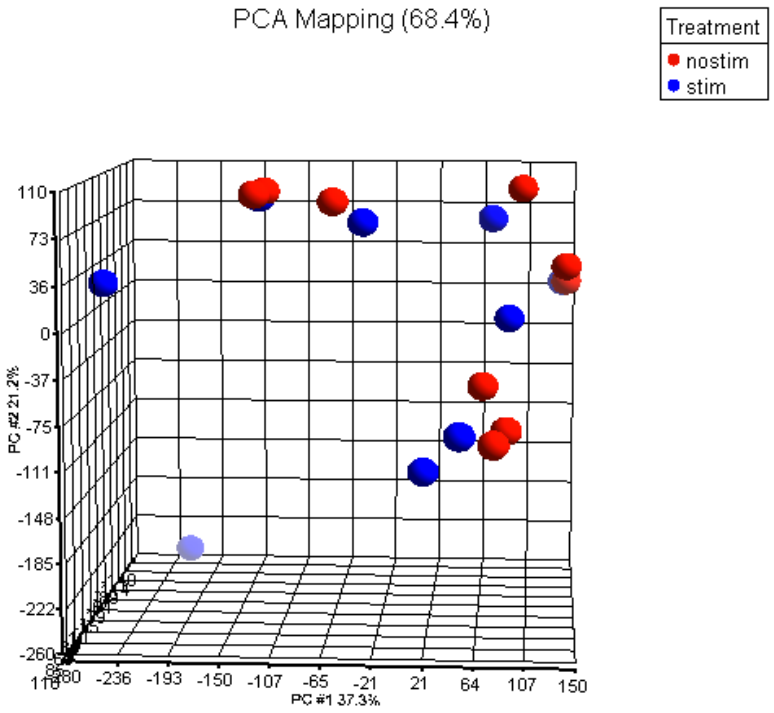
## Principal Component Analysis

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

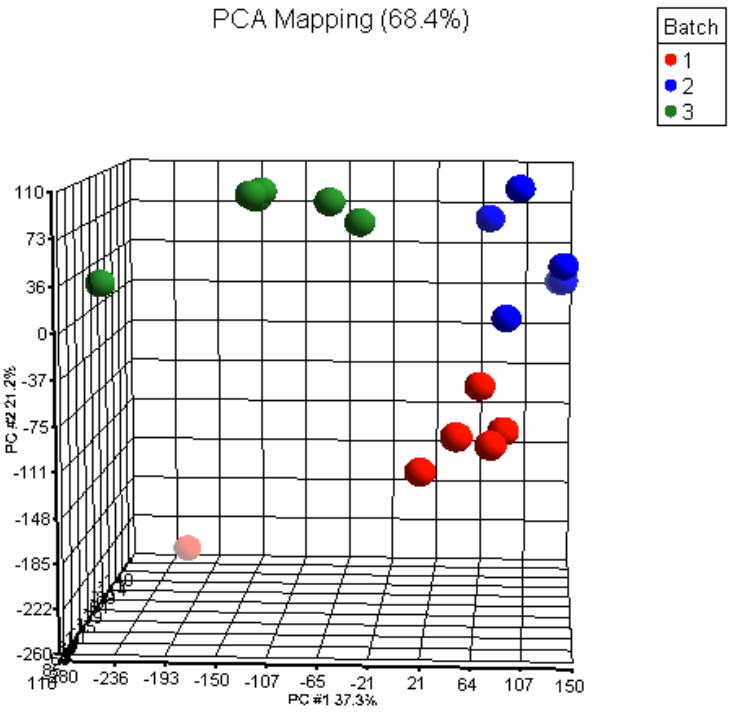
# Example: Stimulation experiment

- 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



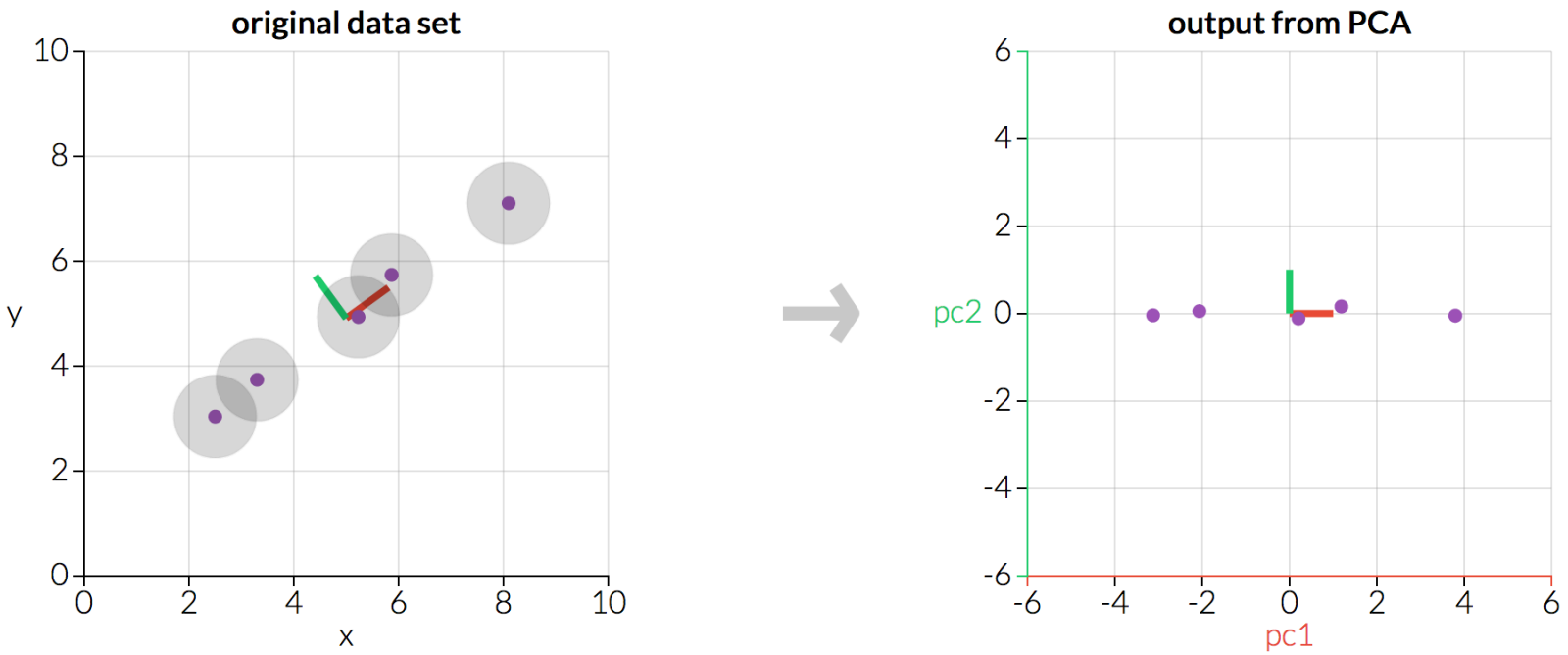
# Example: Stimulation experiment

- 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

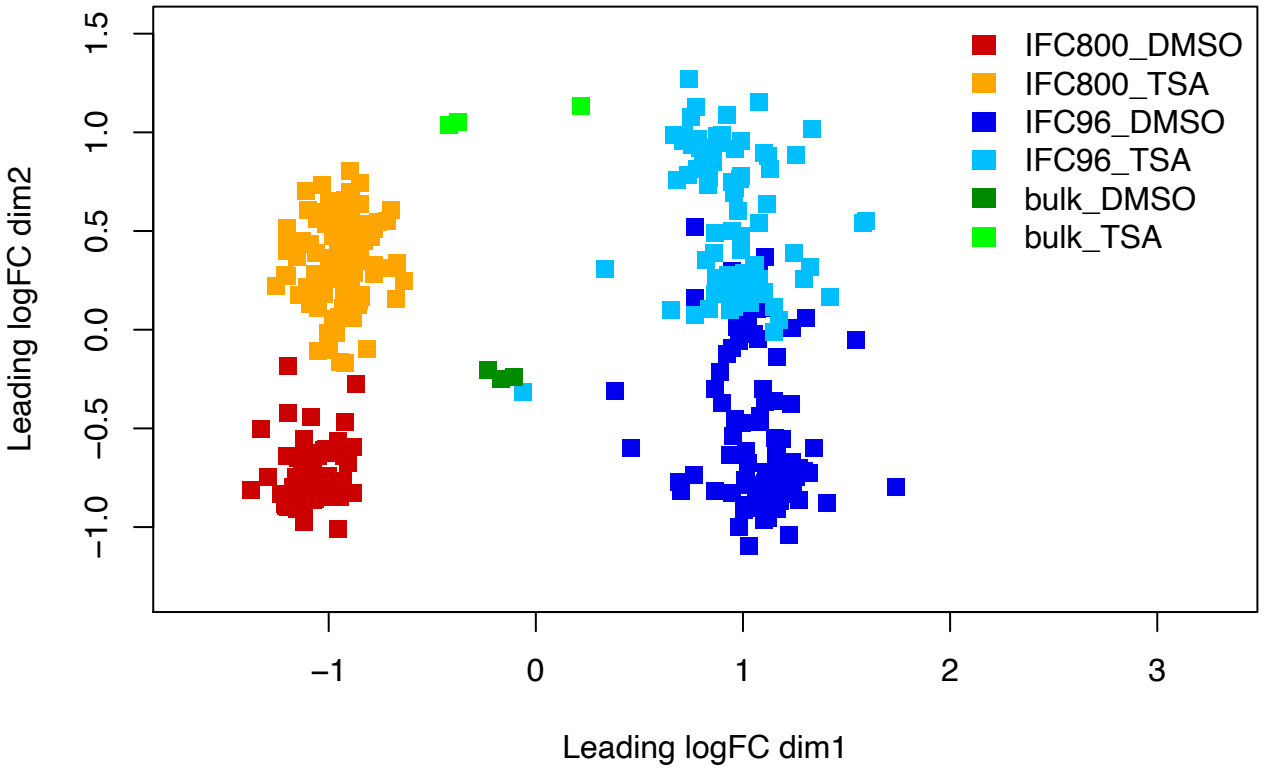
- http://setosa.io/ev/principal-component-analysis/



## Example of multi-dimensional scaling

- Samples treated with two compounds: TSA and DMSO
- Samples measured with three approaches: IFC800, IFC96, and bulk

- What do we learn about effect sizes?



# Example of exploratory analysis

- Fraction of low-quality reads increased in some sequencing results
- What is the driving factor?
  - Sequencer
  - Sample preparation protocol
  - Readlength
  - Input amount
  - ….
- Factors are correlated:
  - Key driving source: presence of long fragments