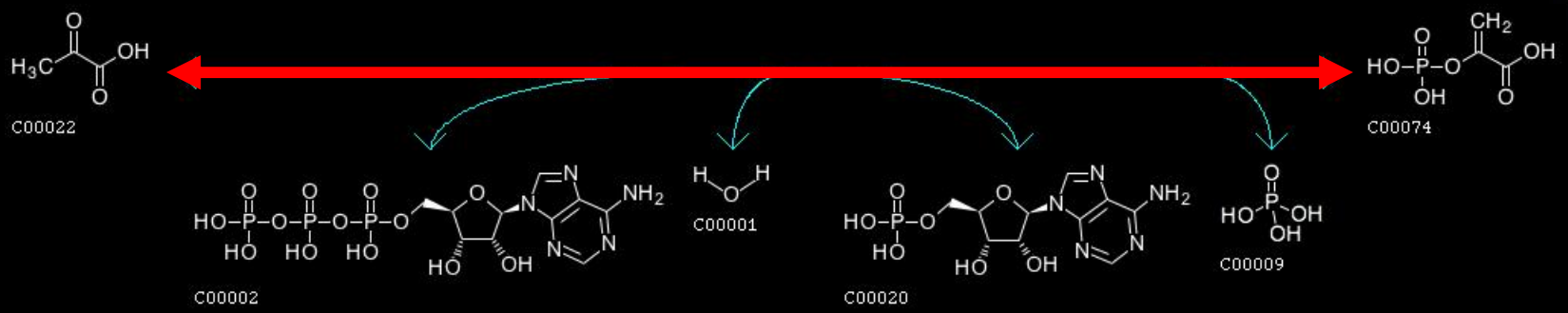
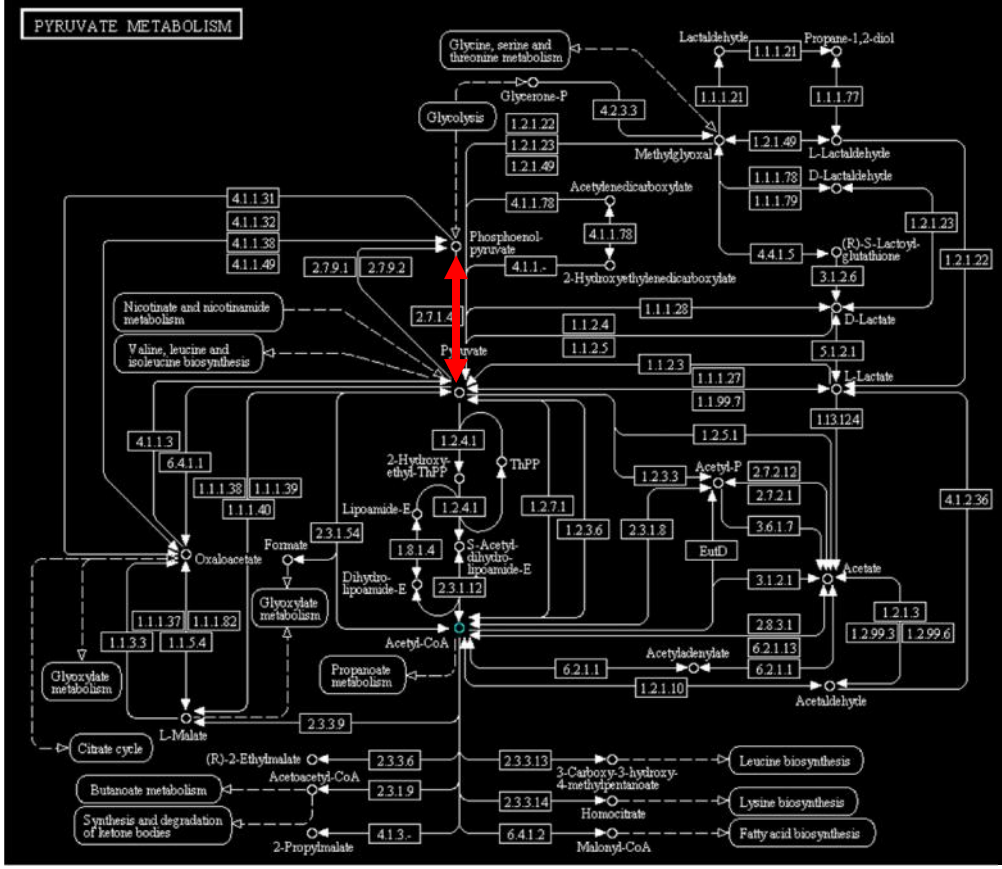
NIH: West Coast Metabolomics Center

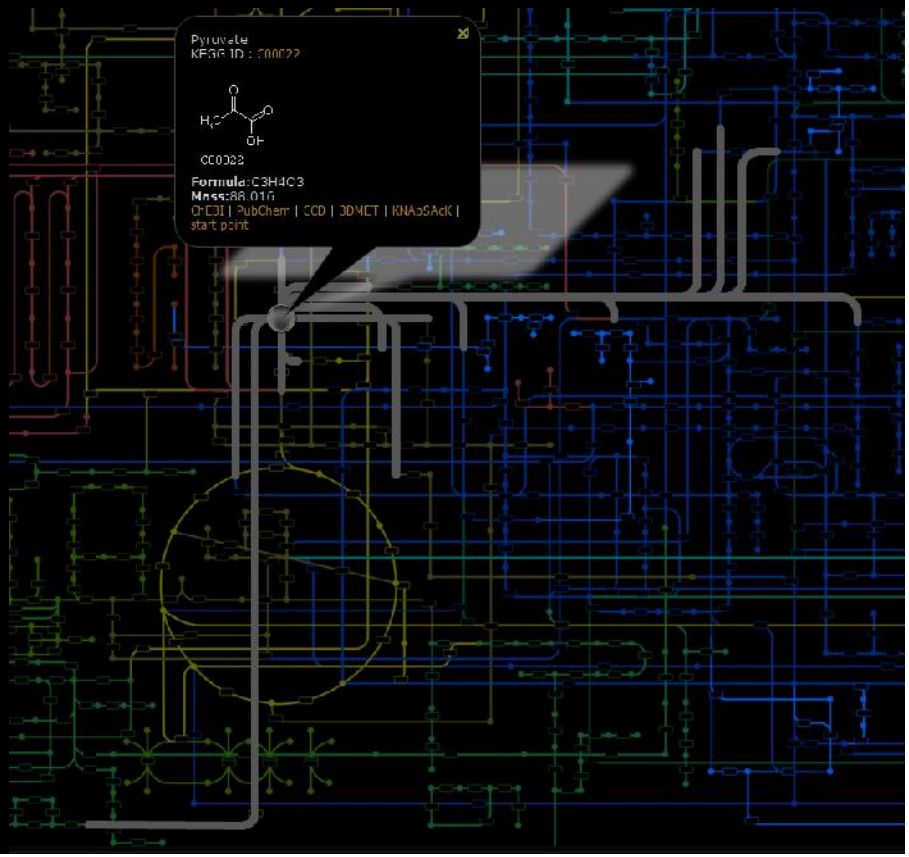
Strategies for Metabolomic Data Analysis

Dmitry Grapov, PhD

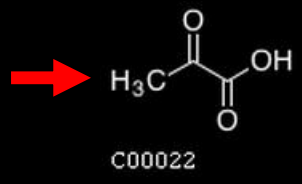
Goals?

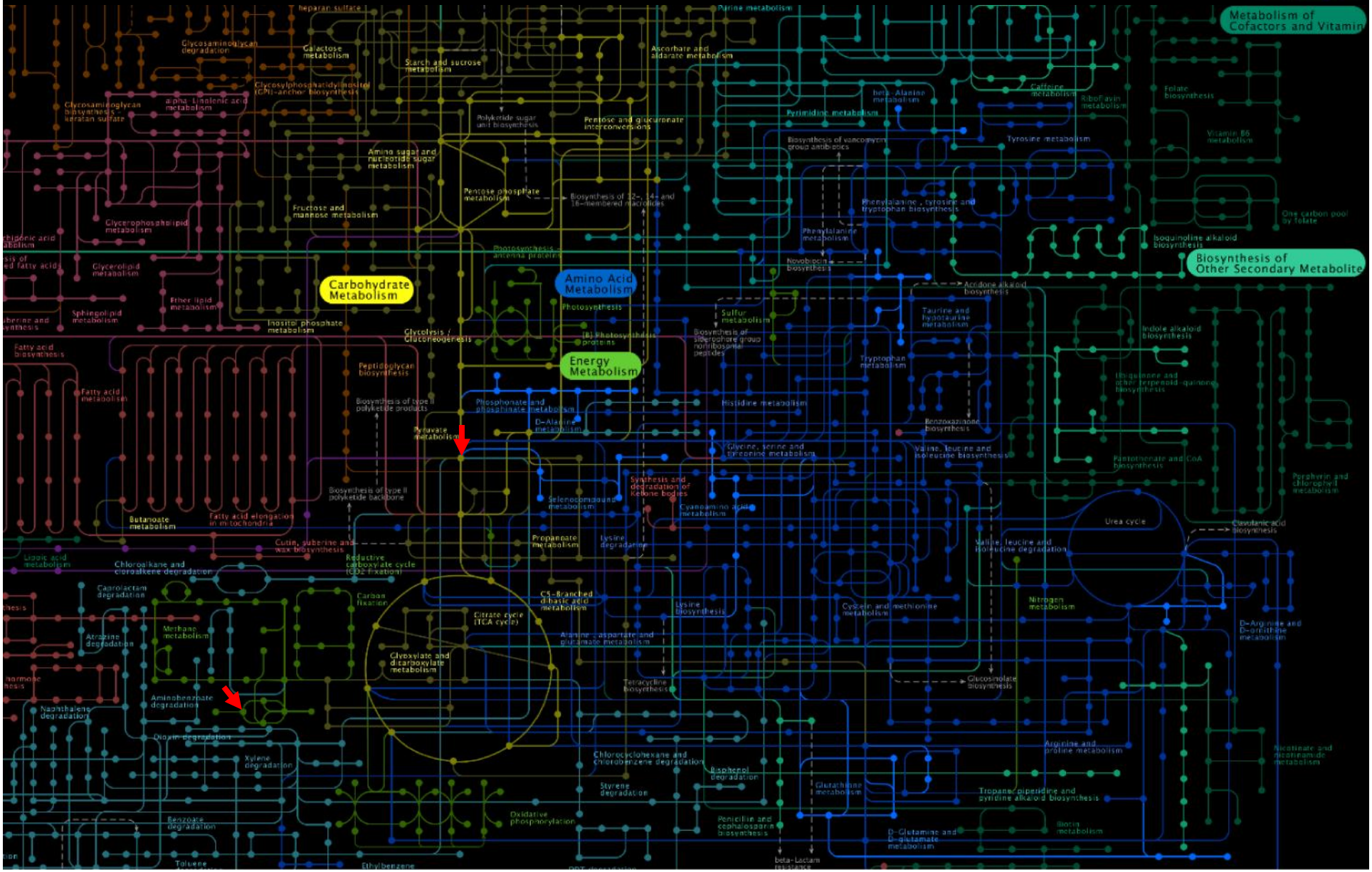

C00022
C00074
C00002
C00001
C00020
C00009


PYRUVATE METABOLISM
Glycine, serine and threonine metabolism
Lactaldehyde
Propane-1,2-diol
Glycerone-P
Glycolysis
Methylglyoxal
L-Lactaldehyde
D-Lactaldehyde
Acetylenedicarboxylate
Phosphoenol-pyruvate
2-Hydroxyethylenedicarboxylate
(R)-S-Lactoyl-glutathione
Nicotinate and nicotinamide metabolism
Valine, leucine and isoleucine biosynthesis
Pyruvate
D-Lactate
L-Lactate
2-Hydroxy-ethyl-ThPP
ThPP
Lipoamide-E
S-Acetyl-dihydro-lipoamide-E
Dihydro-lipoamide-E
Acetyl-P
EutD
Acetate
Oxaloacetate
Formate
Glyoxylate metabolism
Acetyl-CoA
Acetyladenylate
Propanoate metabolism
Acetaldehyde
L-Malate
Citrate cycle
(R)-2-Ethylmalate
Acetoacetyl-CoA
Butanoate metabolism
Synthesis and degradation of ketone bodies
2-Propylmalate
3-Carboxy-3-hydroxy-4-methylpentanoate
Homocitrate
Malonyl-CoA
Leucine biosynthesis
Lysine biosynthesis
Fatty acid biosynthesis


Pyruvate
KEGG ID : C00022
C00022
Formula:C3H4O3
Mass:88.016
start point

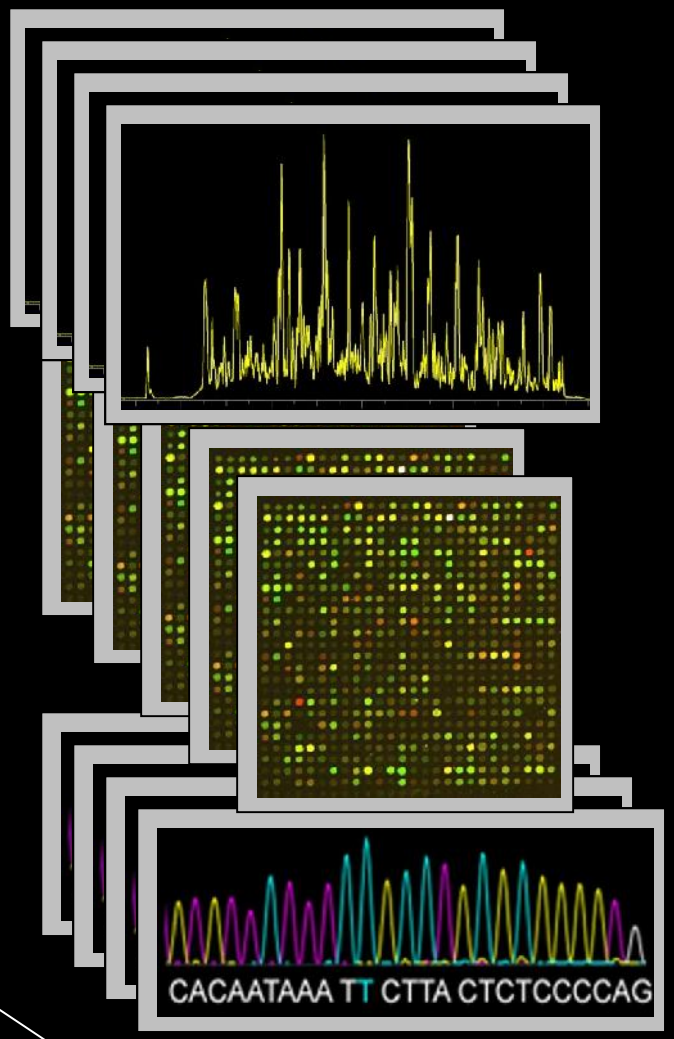
Metabolomics

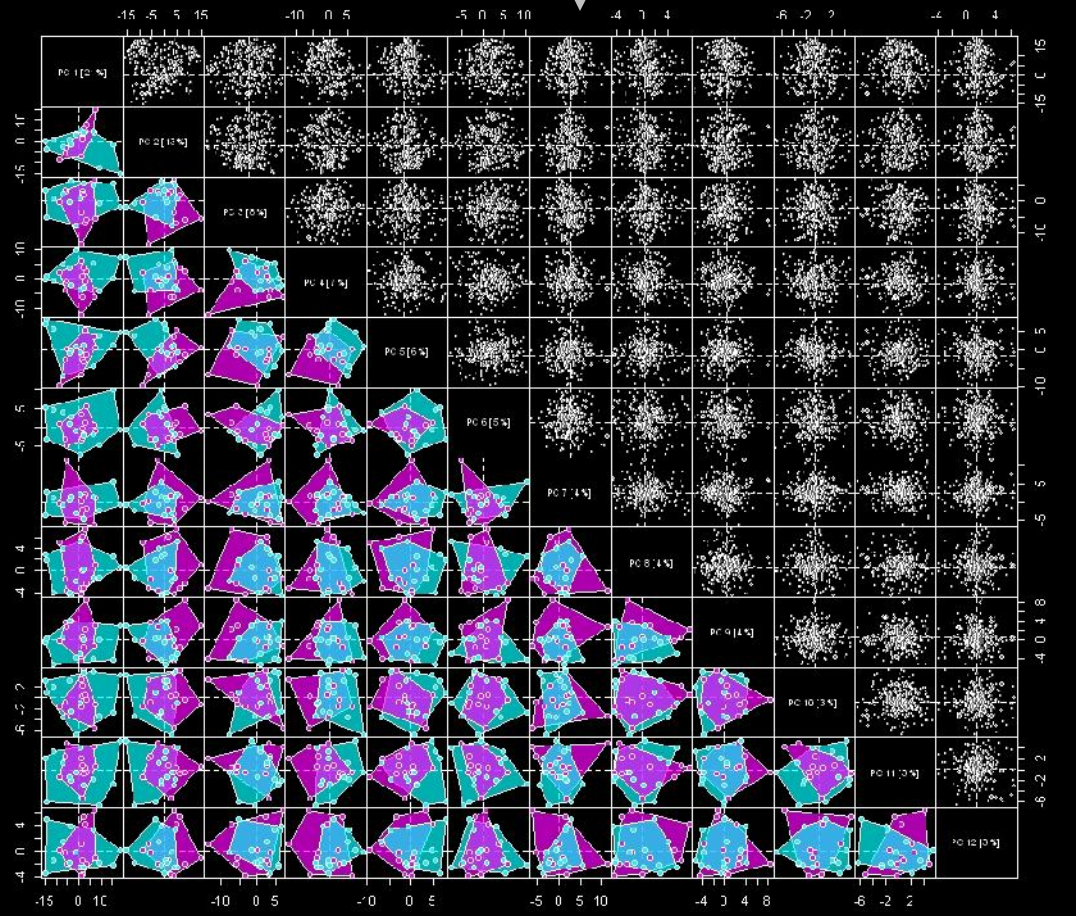




Analytical Dimensions

variables

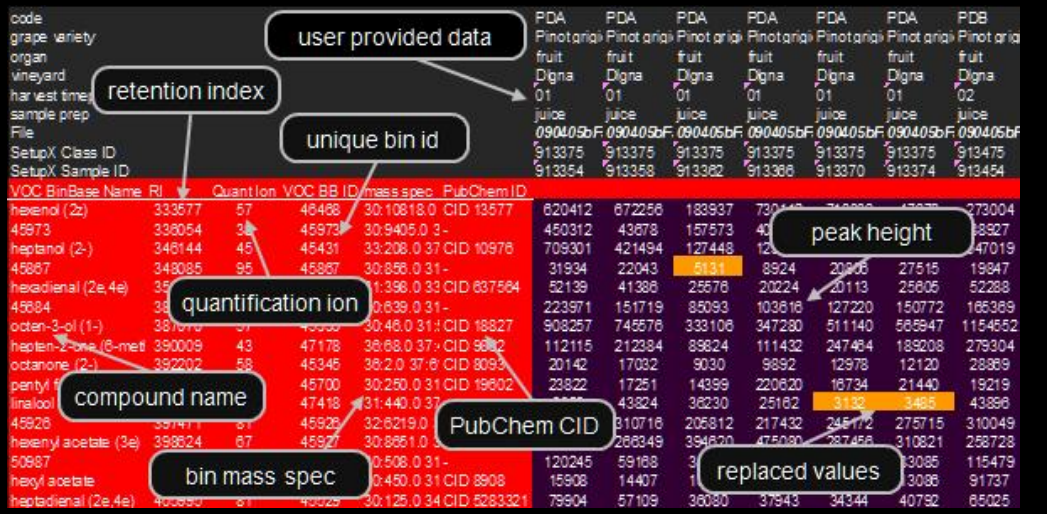




Samples

Metabolomics@ucdavis.edu Metabolomics.ucdavis.edu

Analyzing Metabolomic Data



- **Pre-analysis**
- **Data properties**
- **Statistical approaches**
- **Multivariate approaches**
- **Systems approaches**

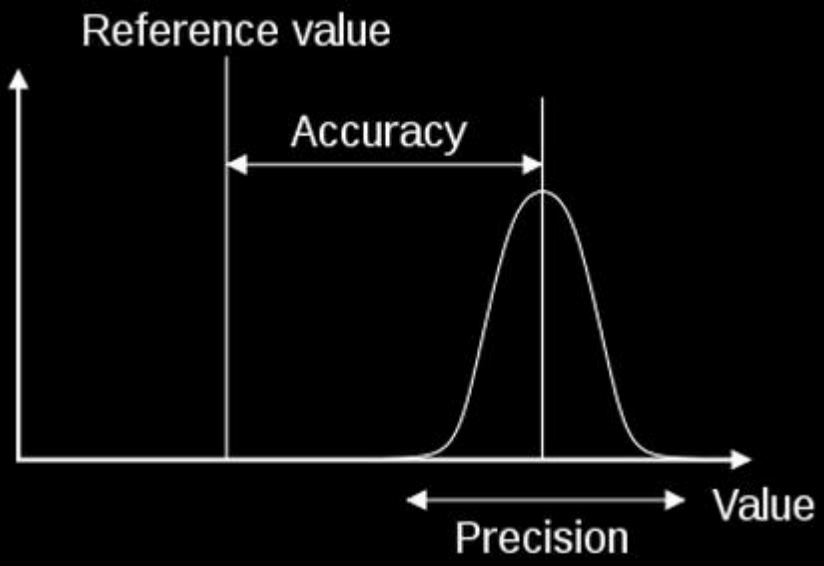
Pre-analysis

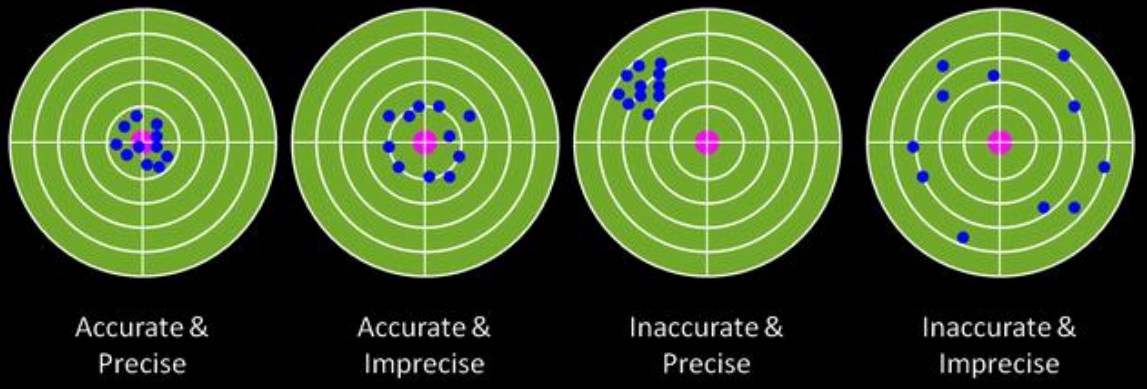
Data quality metrics

- **precision**
- **accuracy**

Remedies

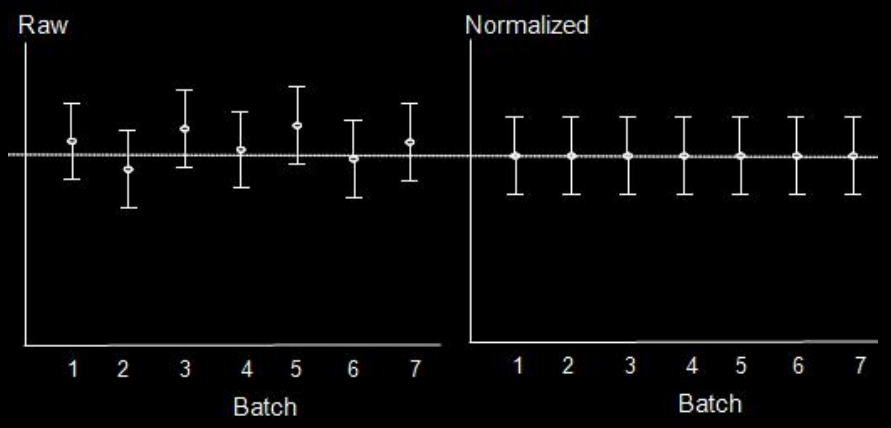
- **normalization**
- **outliers detection**
- **missing values imputation**

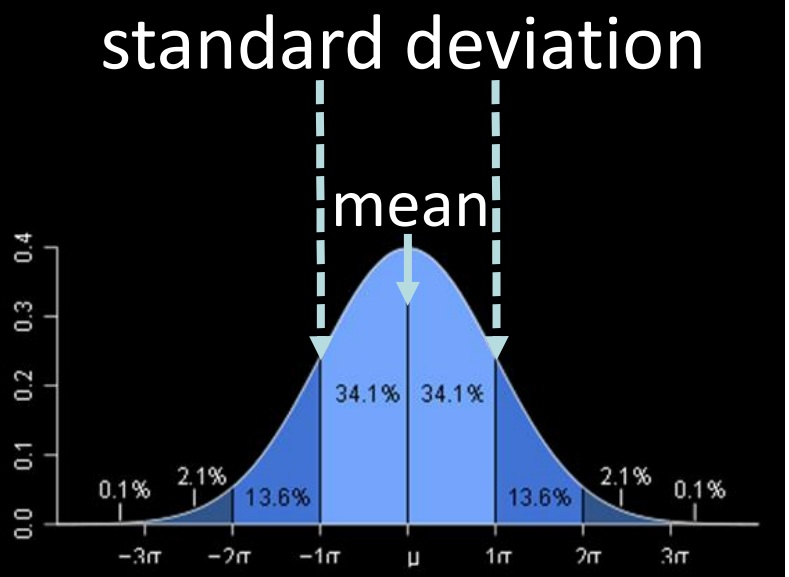




Normalization

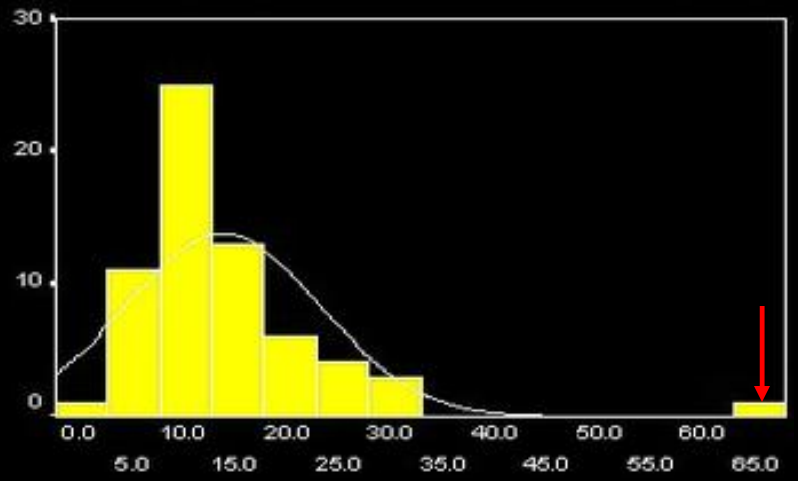
- sample-wise
  - sum, adjusted
- measurement-wise
  - transformation (normality)
  - encoding (trigonometric, etc.)



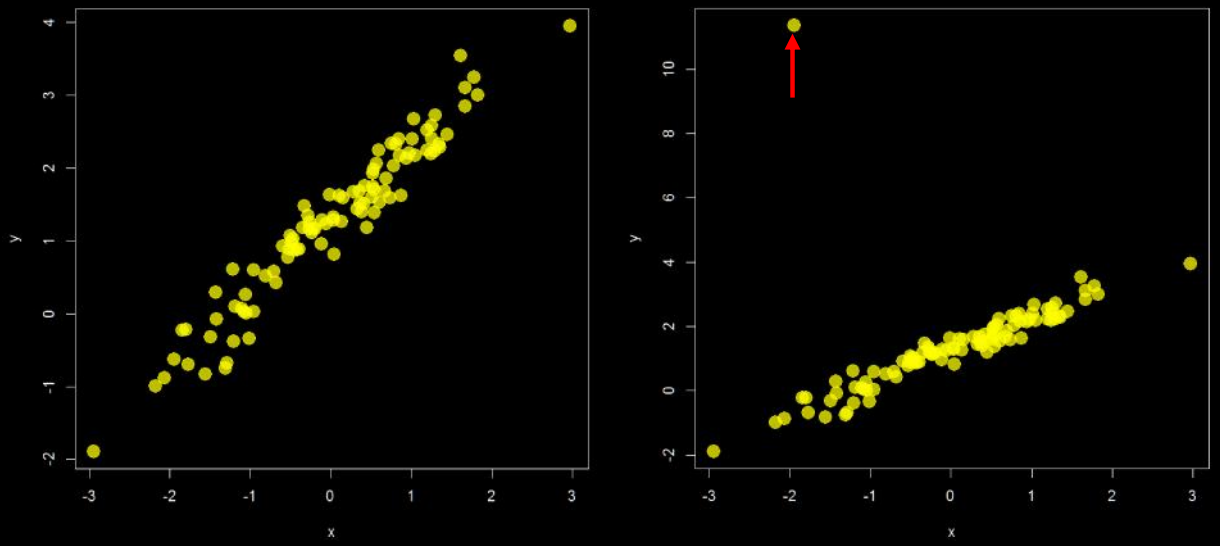


Outliers

- **single measurements (univariate)**

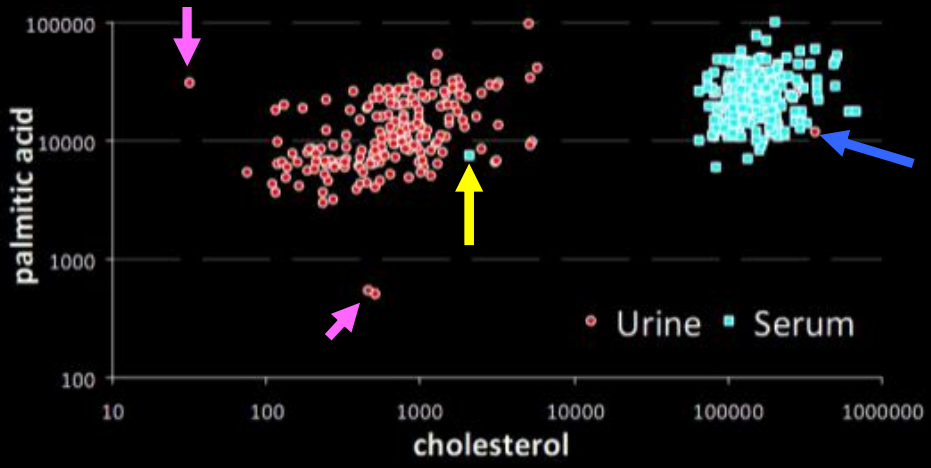


- **two compounds (bivariate)**



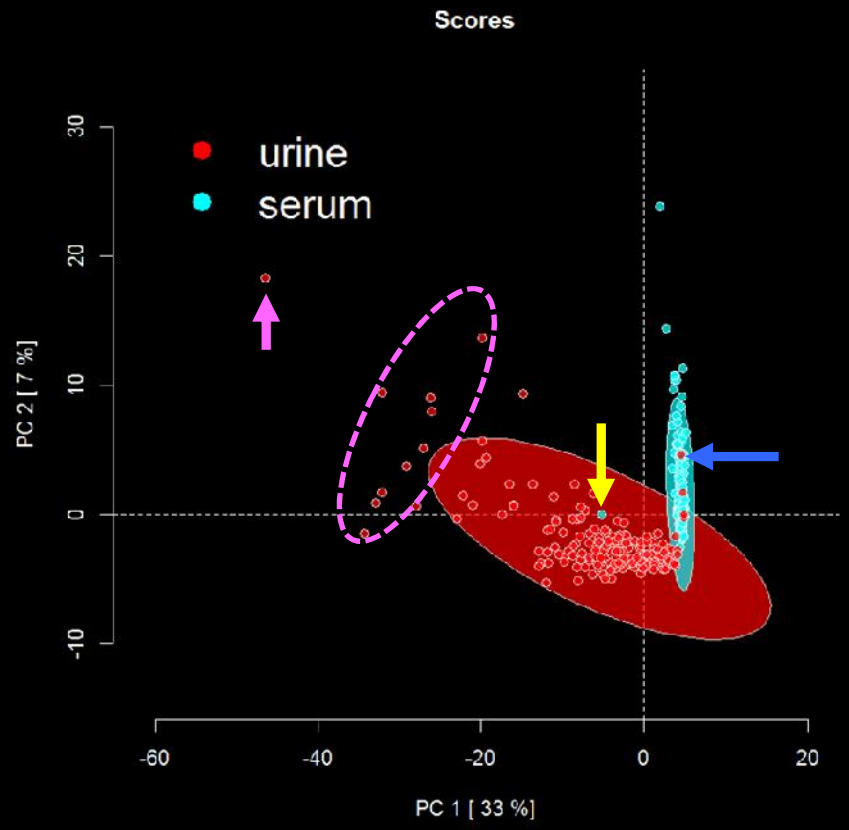
Outliers

univariate/bivariate



vs.

multivariate



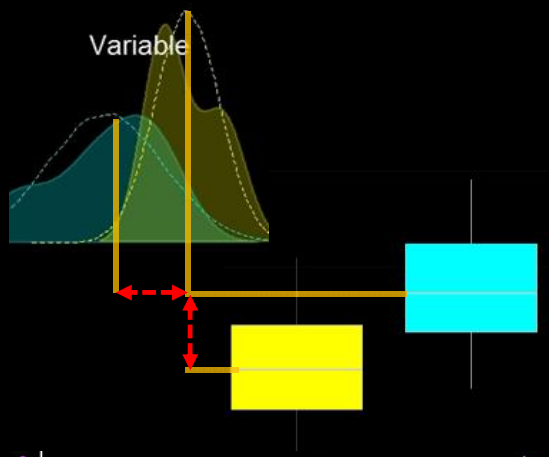
→ outliers?

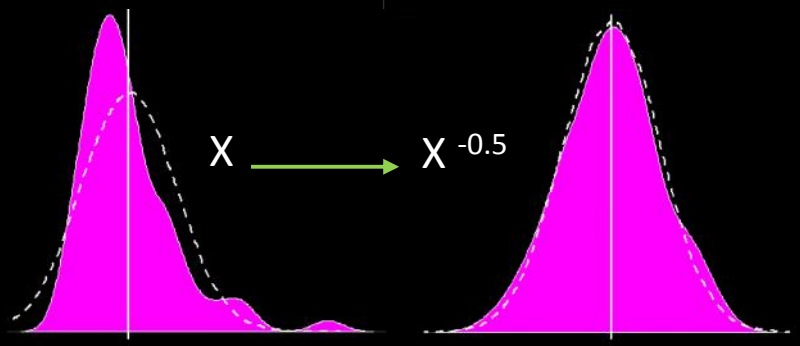
⇉ mixed up samples

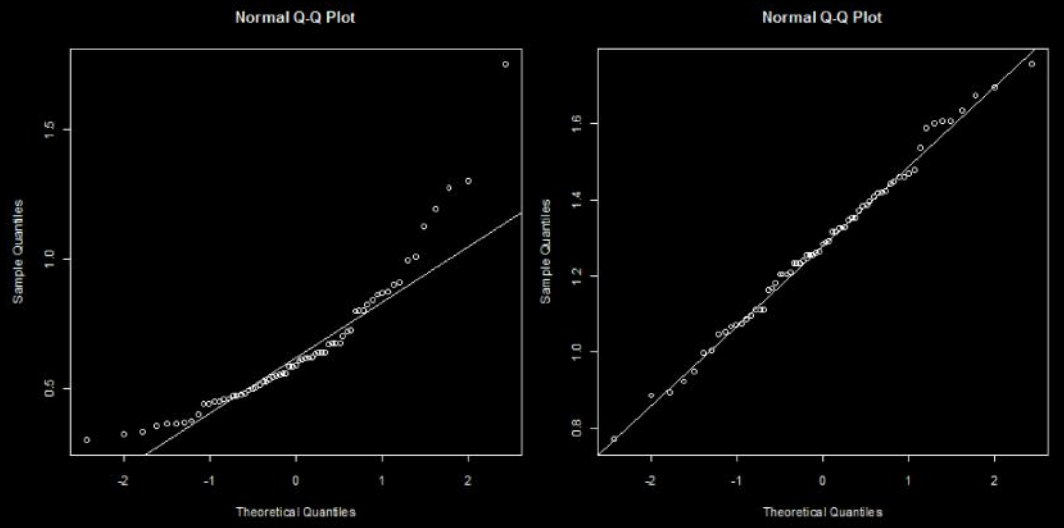
Transformation

- **logarithm (shifted)**
- **power (BOX-COX)**
- **inverse**

Quantile-quantile (Q-Q) plots are useful for visual overview of variable normality







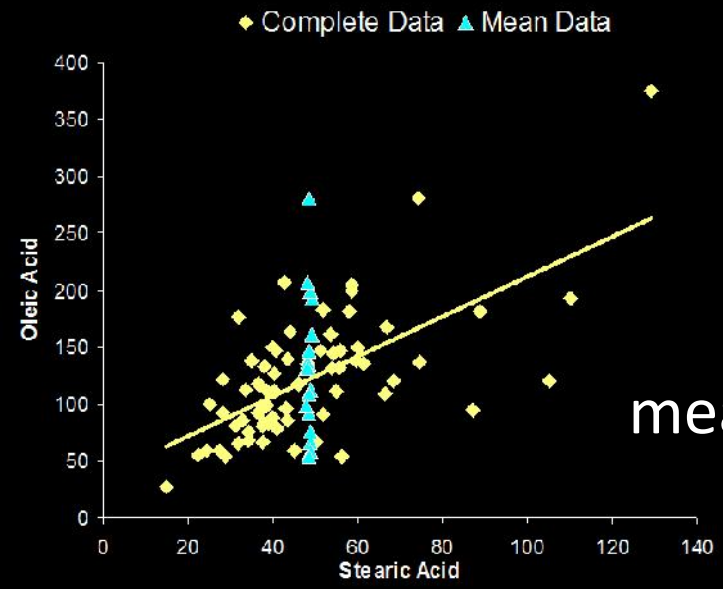
Missing Values Imputation

Why is it missing?

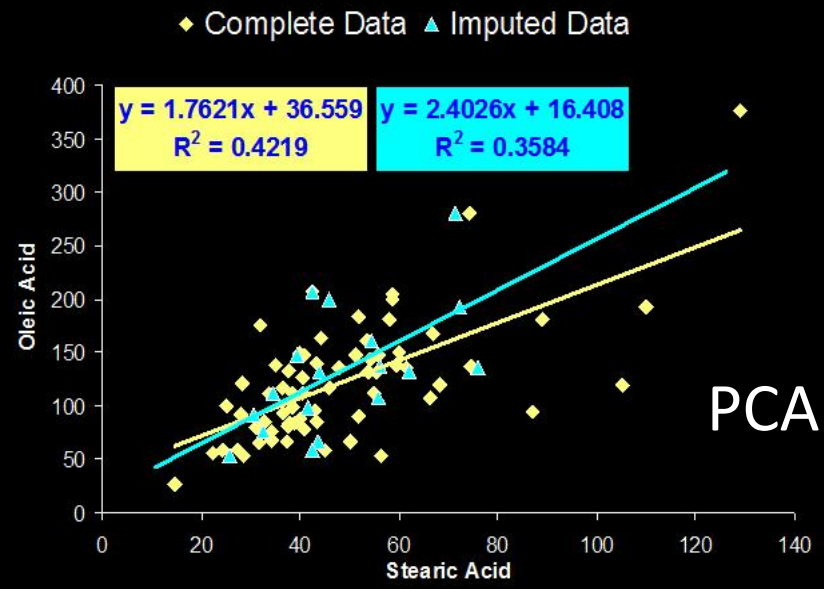
- random
- systematic
  - analytical
  - biological

Imputation methods

- single value (mean, min, etc.)
- multiple
- multivariate



mean



PCA

Goals for Data Analysis

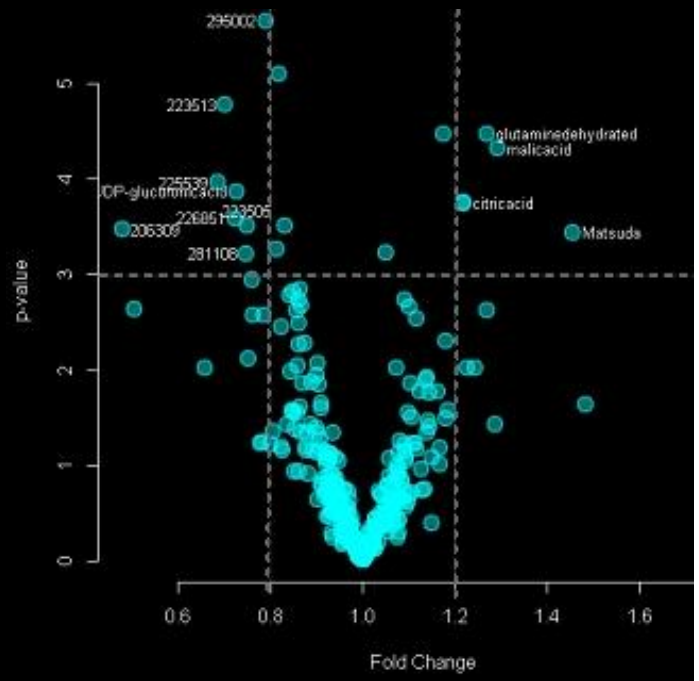
Exploration Classification Prediction

- Are there any trends in my data?
  - analytical sources
  - meta data/covariates

- Useful Methods
  - matrix decomposition (PCA, ICA, NMF)
  - cluster analysis

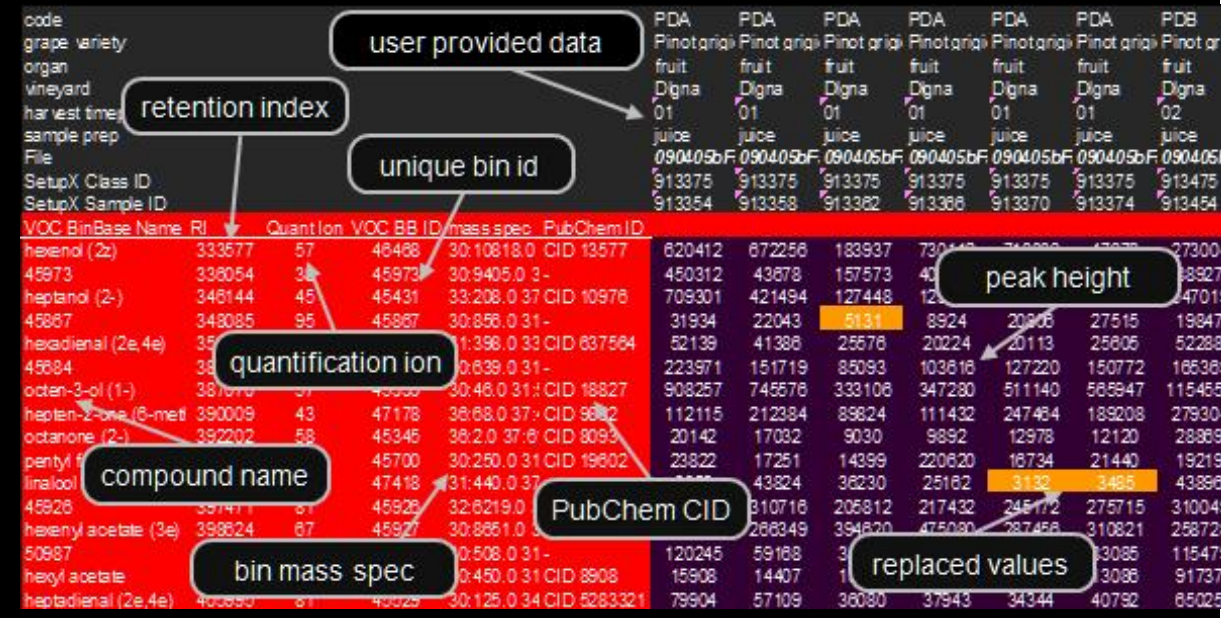
- Differences/similarities between groups?
  - discrimination, classification, significant changes
- Useful Methods
  - analysis of variance (ANOVA)
  - partial least squares discriminant analysis (PLS-DA)
  - Others: random forest, CART, SVM, ANN

- What is related or predictive of my variable(s) of interest?
  - regression
- Useful Methods
  - correlation

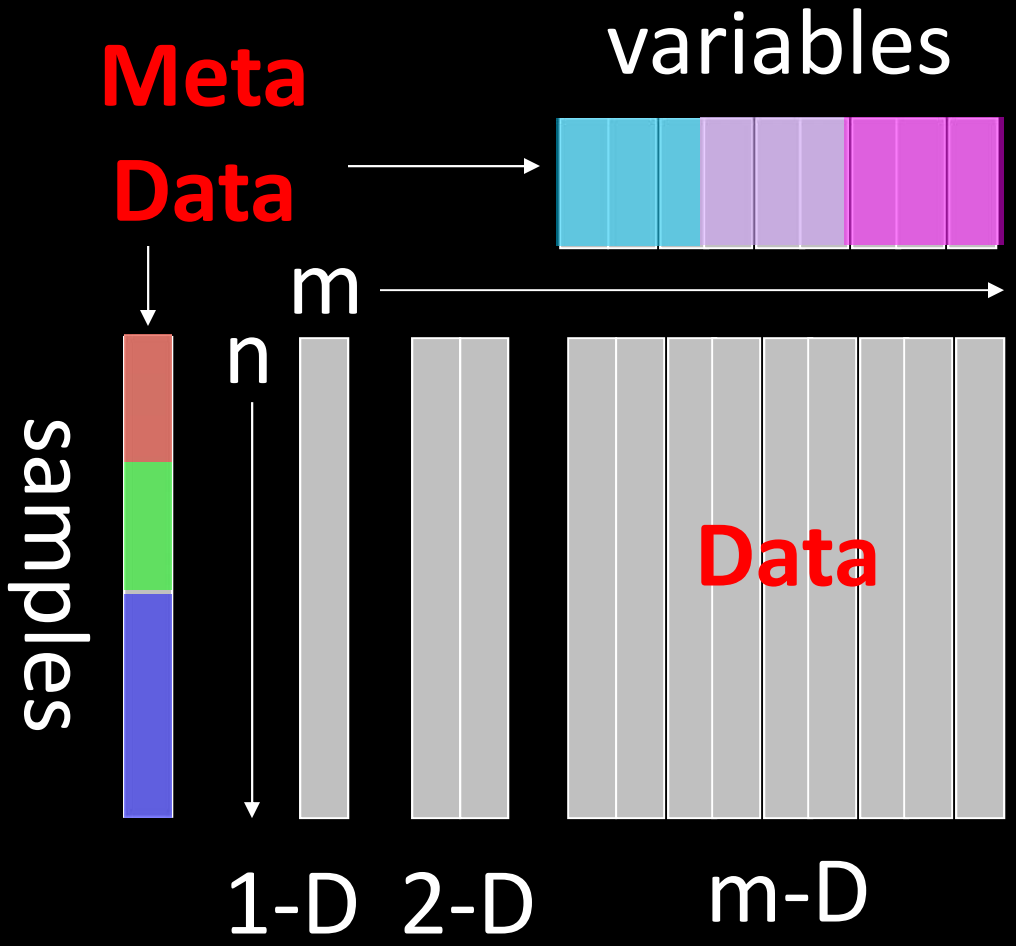


Data Structure

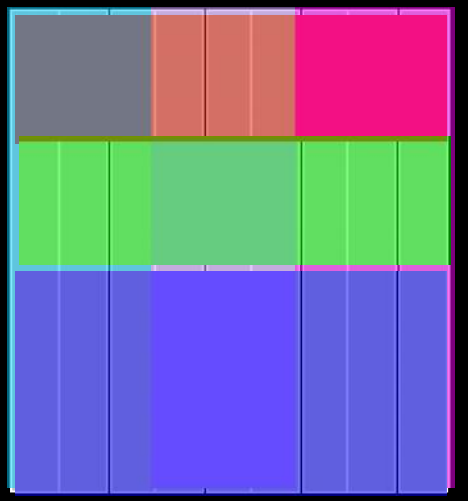
- **univariate: a single variable (1-D)**
- **bivariate: two variables (2-D)**
- **multivariate: 2 > variables (m-D)**
  - **Data Types**
    - continuous
    - discreet
    - binary



Data Complexity



Experimental Design = complexity

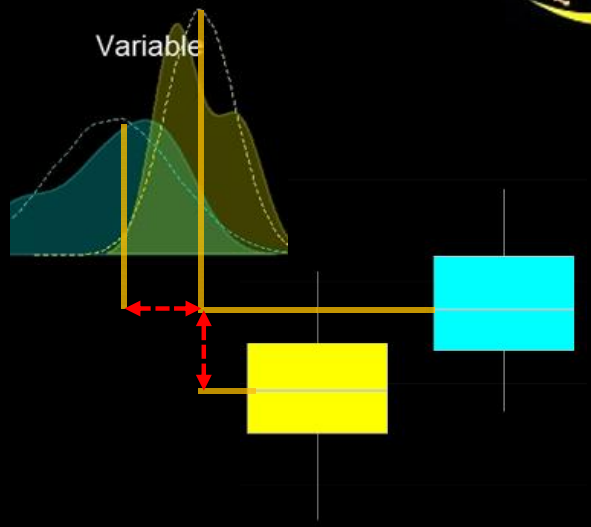


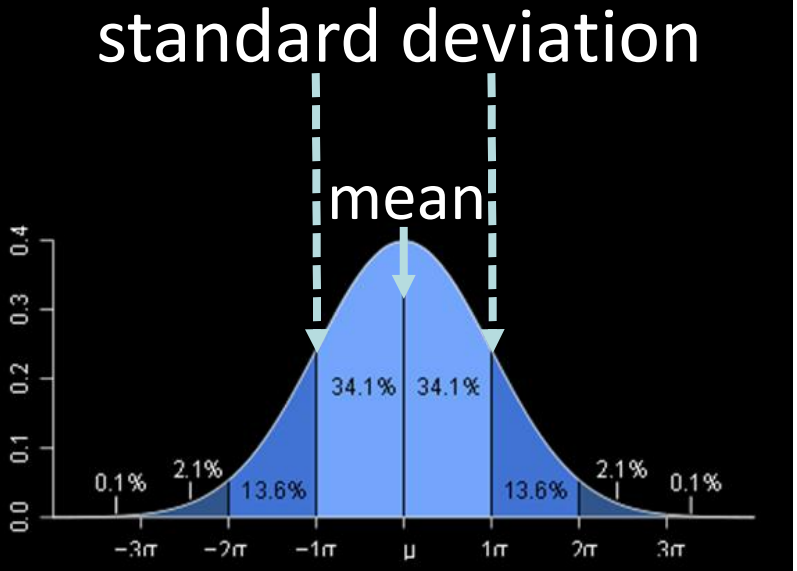
Variable # = dimensionality

Univariate Analyses

univariate properties

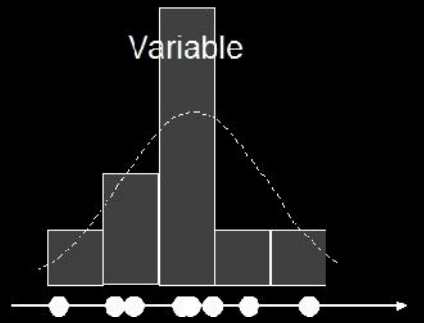
- length
- center (mean, median, geometric mean)
- dispersion (variance, standard deviation)
- Range (min / max)

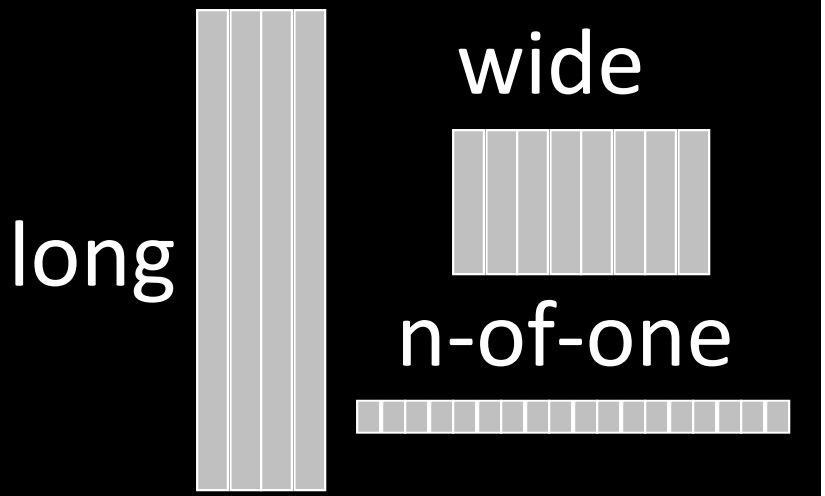




Univariate Analyses

- sensitive to distribution shape
  - parametric = assumes normality
- error in Y, not in X (Y = mX + error)
- optimal for long data
- assumed independence
- false discovery rate



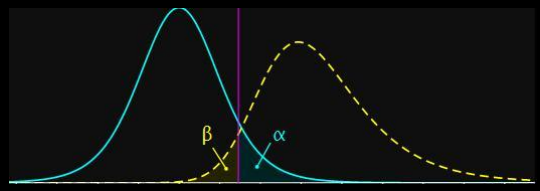


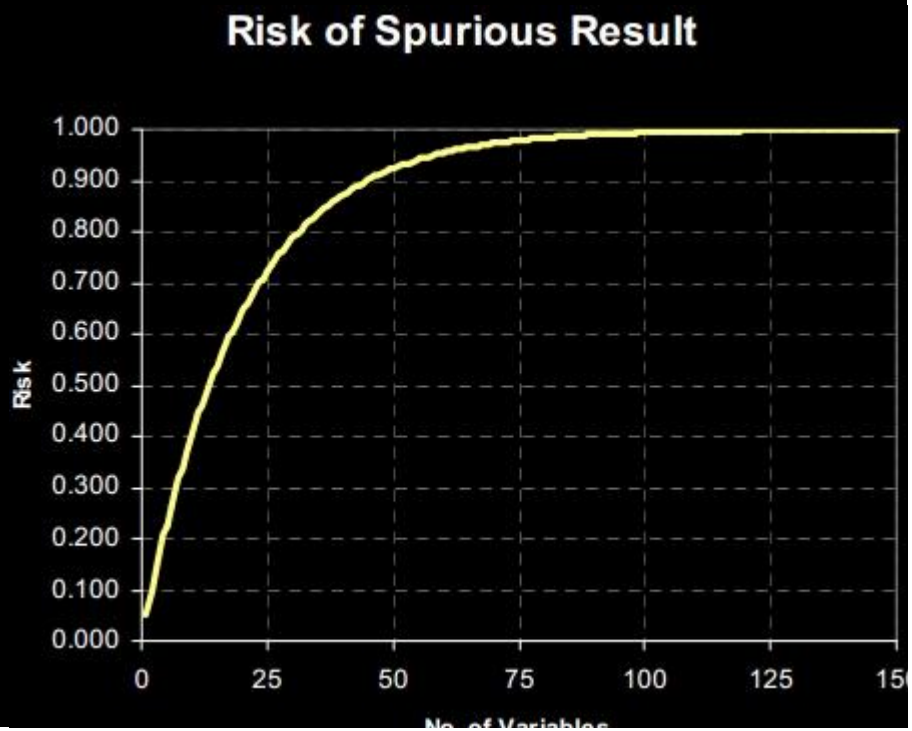
False Discovery Rate (FDR)

univariate approaches do not scale well

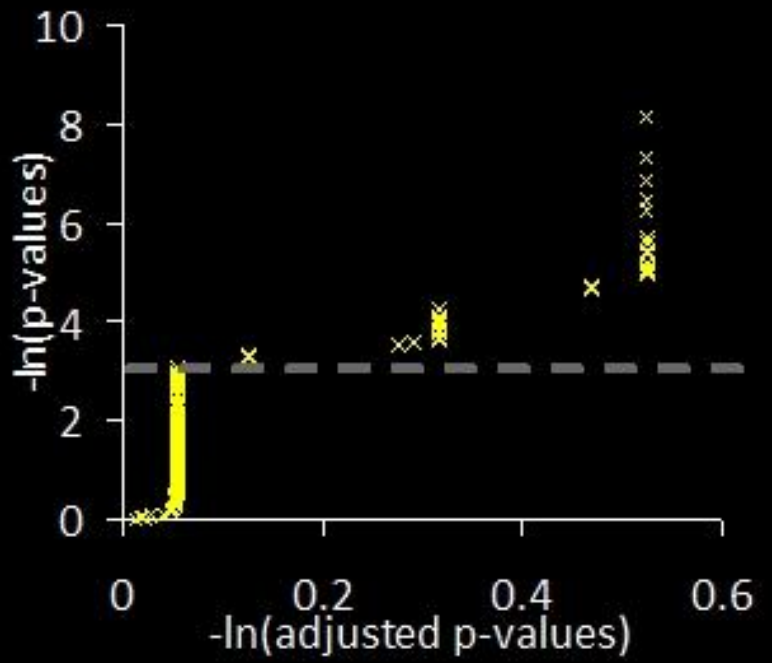
- **Type I Error: False Positives**
- **Type II Error: False Negatives**
- **Type I risk =**
  - **$1-(1-p.value)^m$**

m = number of variables tested





FDR correction



Example:

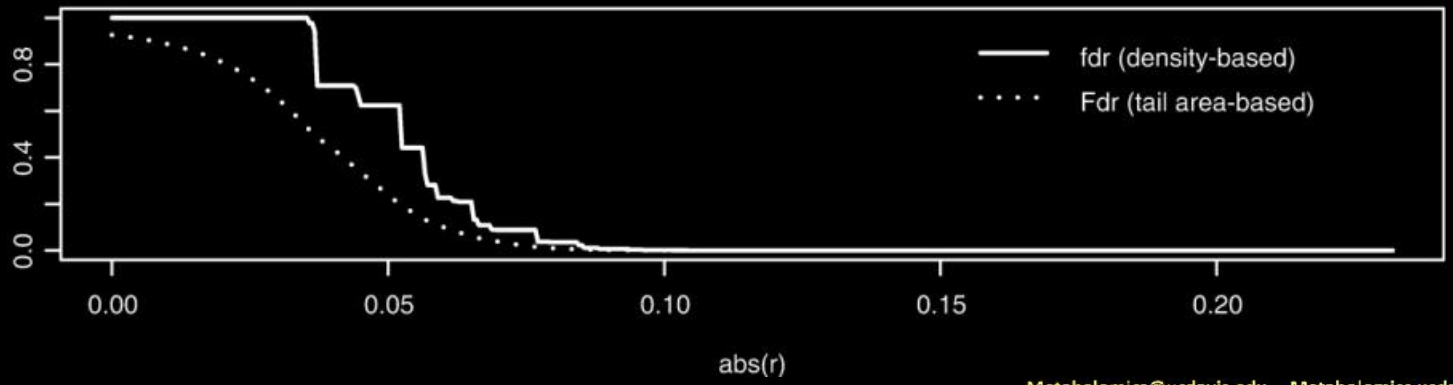
Design: 30 sample, 300 variables

Test: t-test

FDR method: Benjamini and Hochberg (fdr) correction at q=0.05

Results

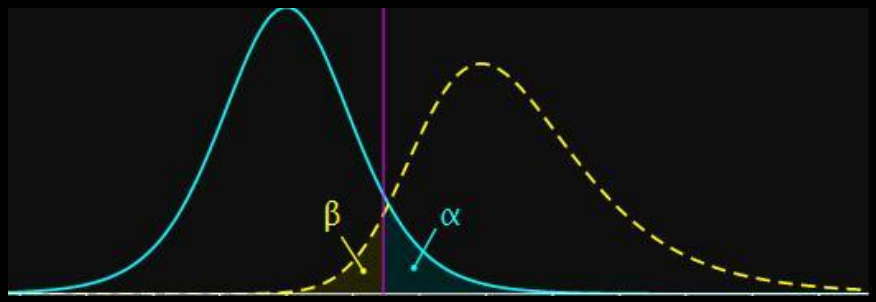
FDR adjusted p-values (fdr) or estimate of FDR (Fdr, q-value)



Bioinformatics (2008) 24 (12):1461-1462

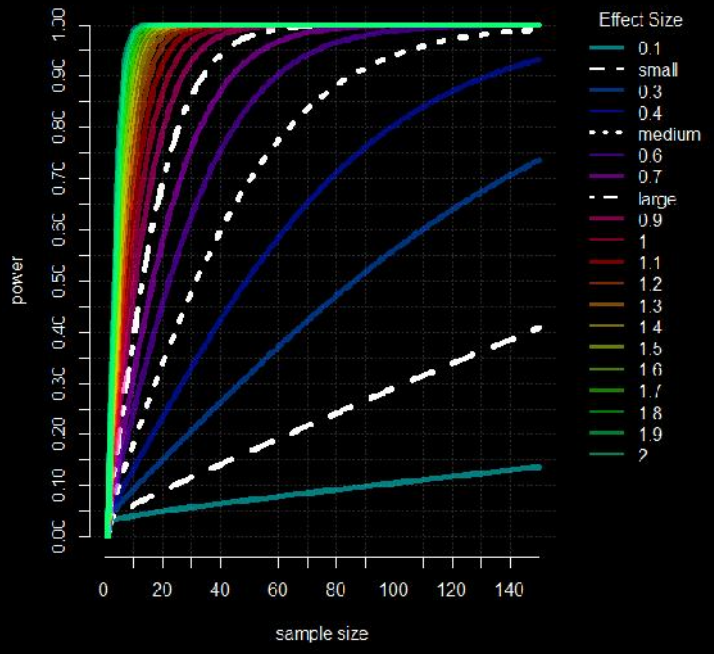
Achieving “significance” is a function of:

significance level (α) and power (1-β)



effect size (standardized difference in means)

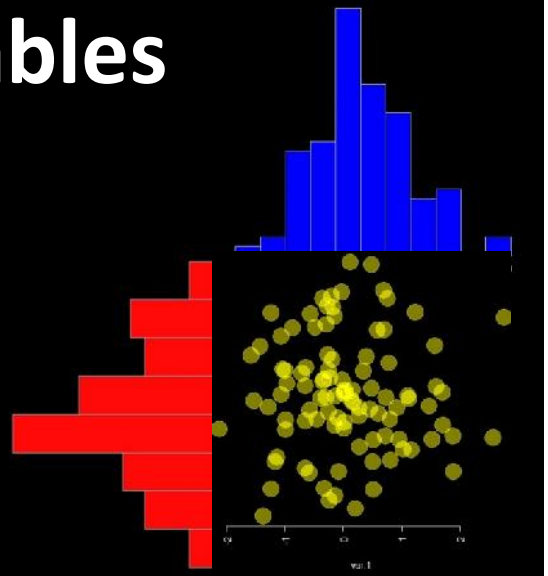
sample size (n)

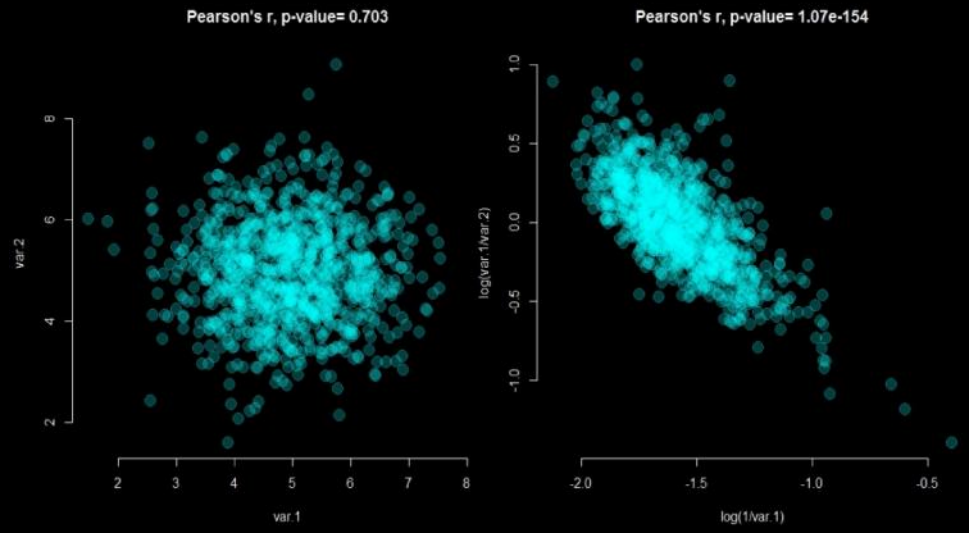


Bivariate Data

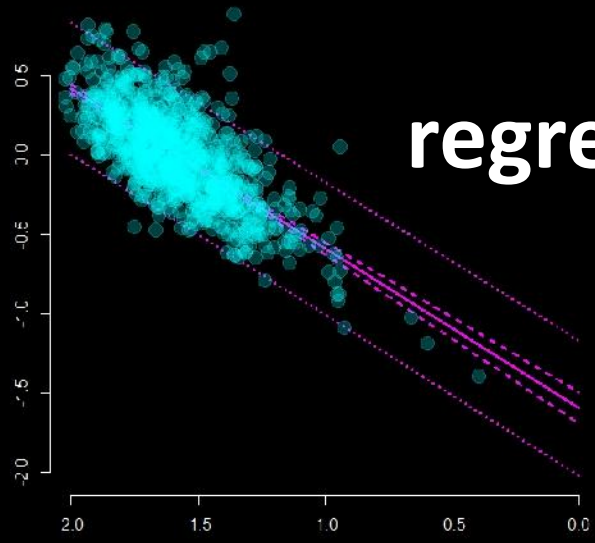
relationship between two variables

- **correlation (strength)**
- **regression (predictive)**





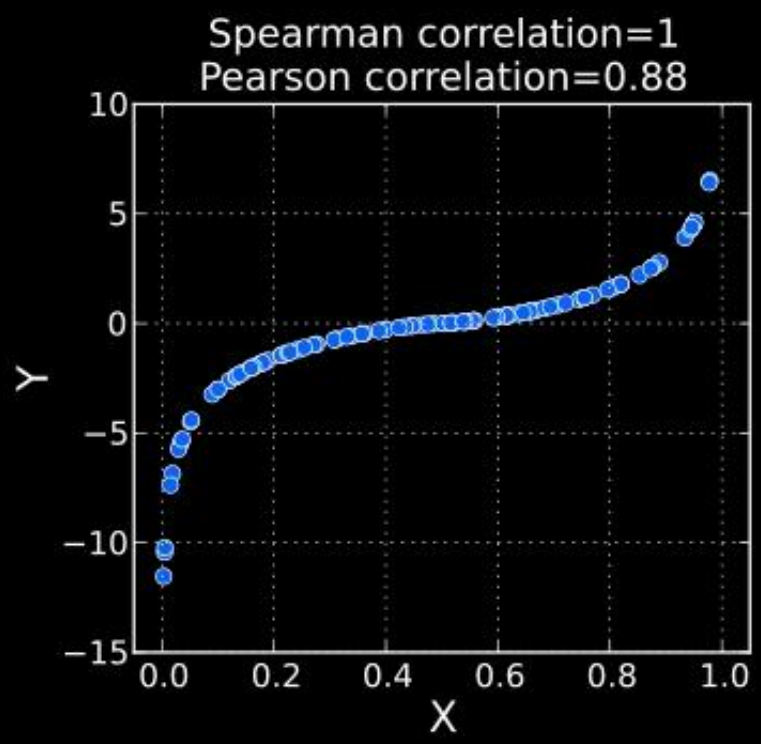
correlation

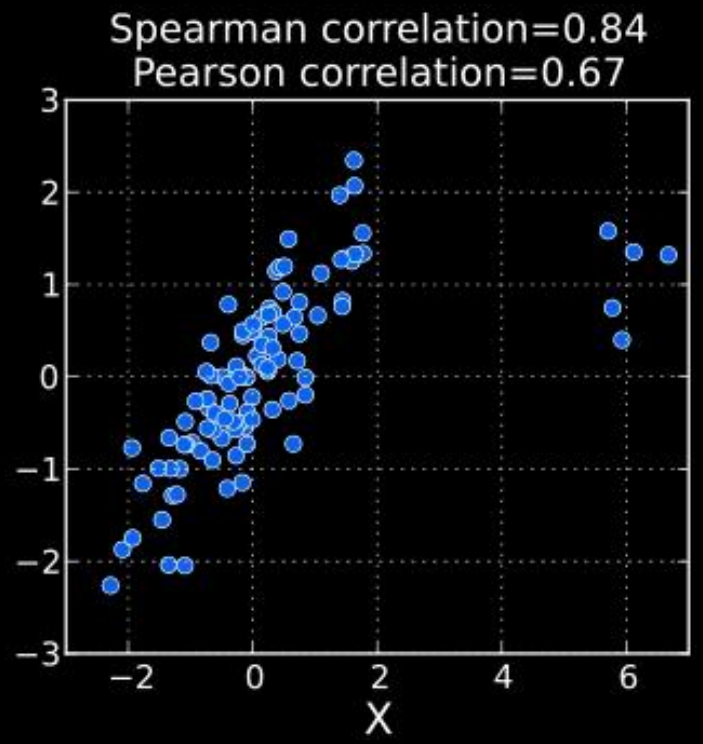


regression

Correlation

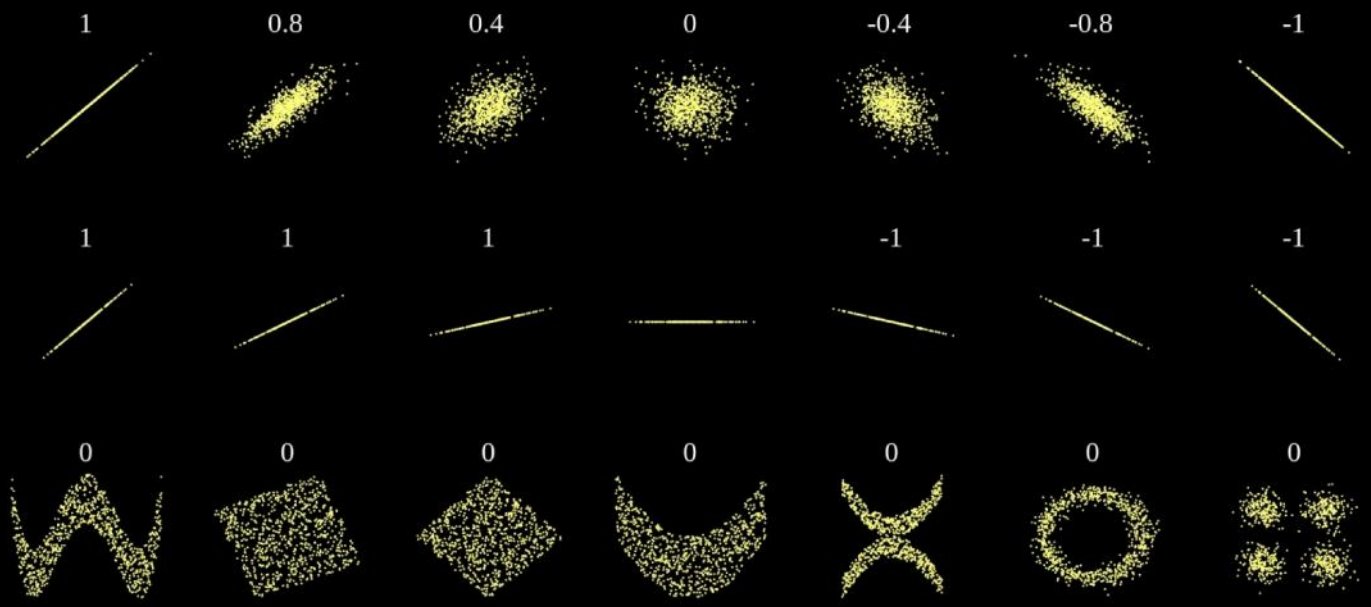
- Parametric (Pearson) or rank-order (Spearman, Kendall)

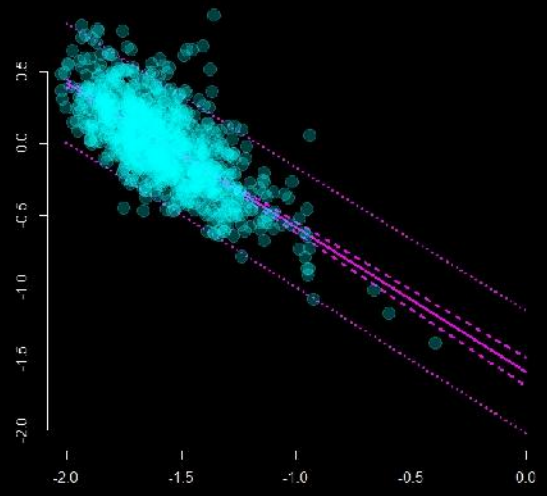




- correlation is covariance scaled between -1 and 1

Correlation vs. Regression





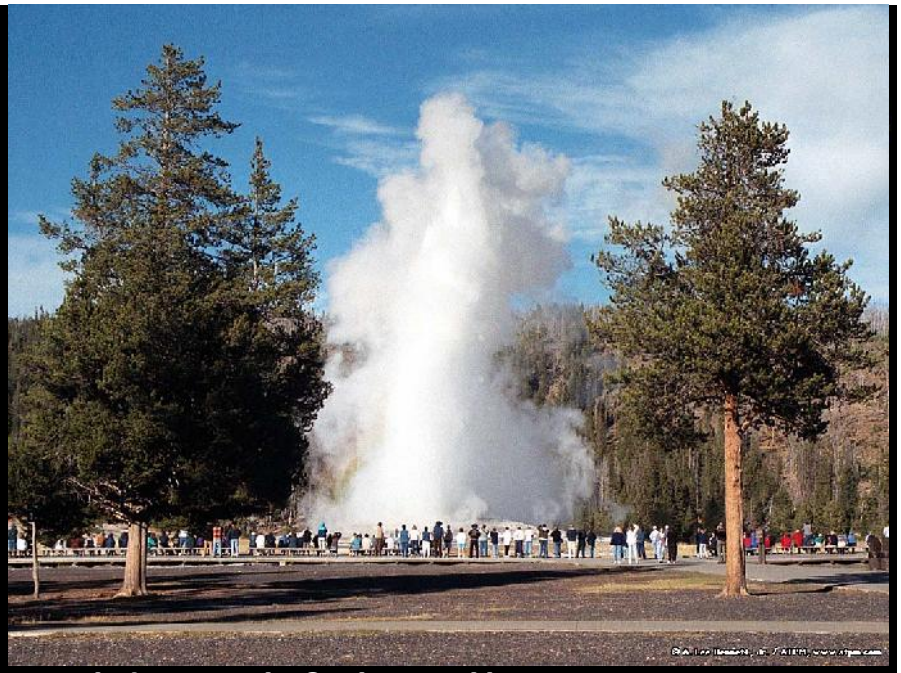
Regression describes the least squares or best-fit-line for the relationship (Y = m*X + b)

Bivariate Example

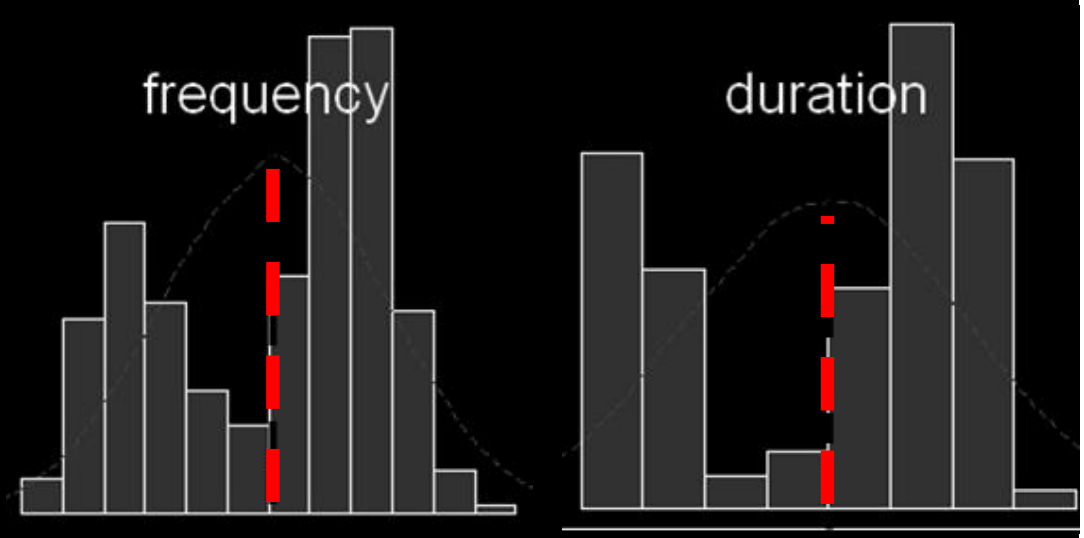
Goal: Don't miss eruption!

Data

- time between eruptions
  - 70 ± 14 min
- duration of eruption
  - 3.5 ± 1 min



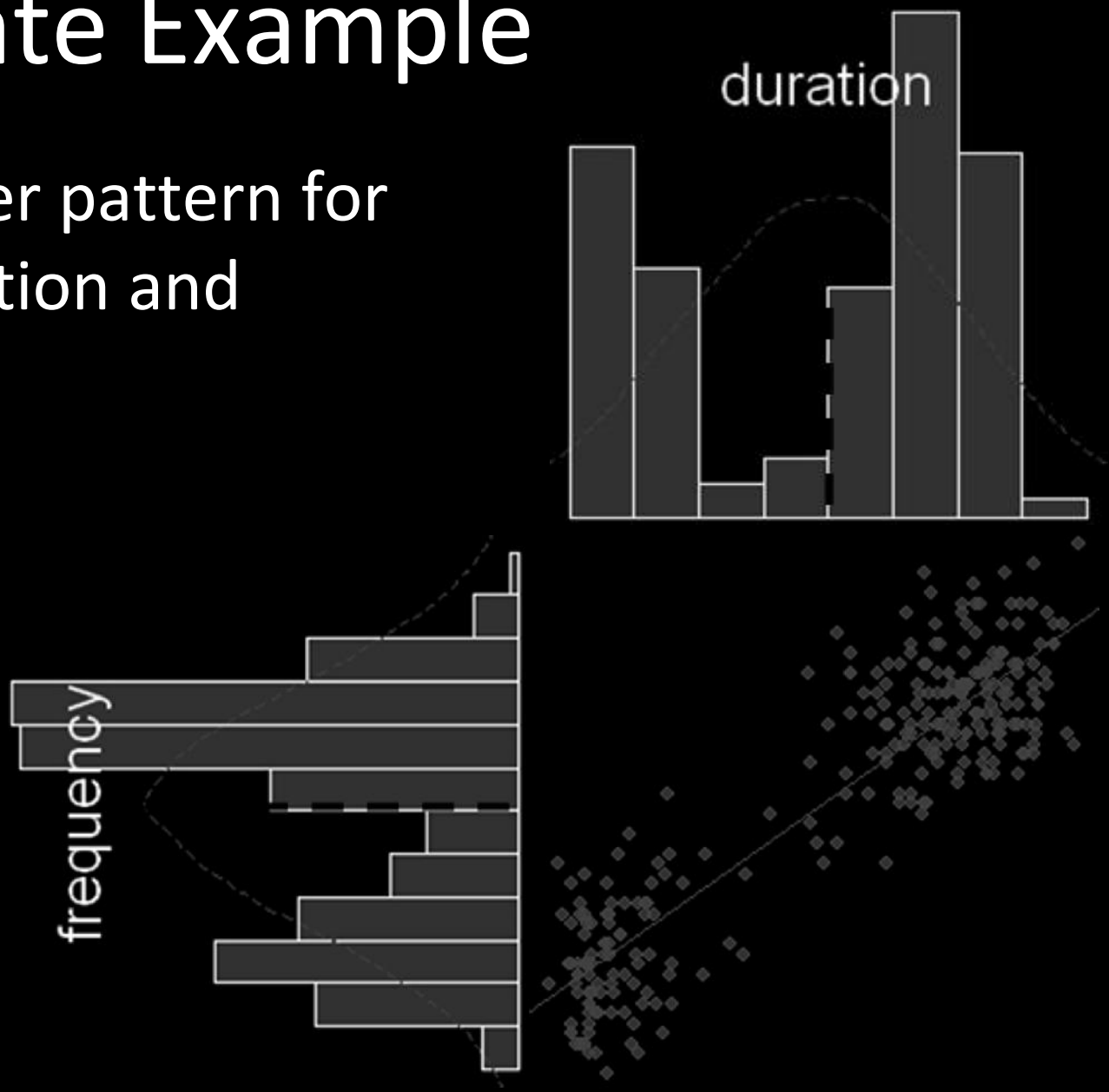
Old Faithful, Yellowstone, WY



Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics* **39**, 357–365

Bivariate Example

Two cluster pattern for both duration and frequency

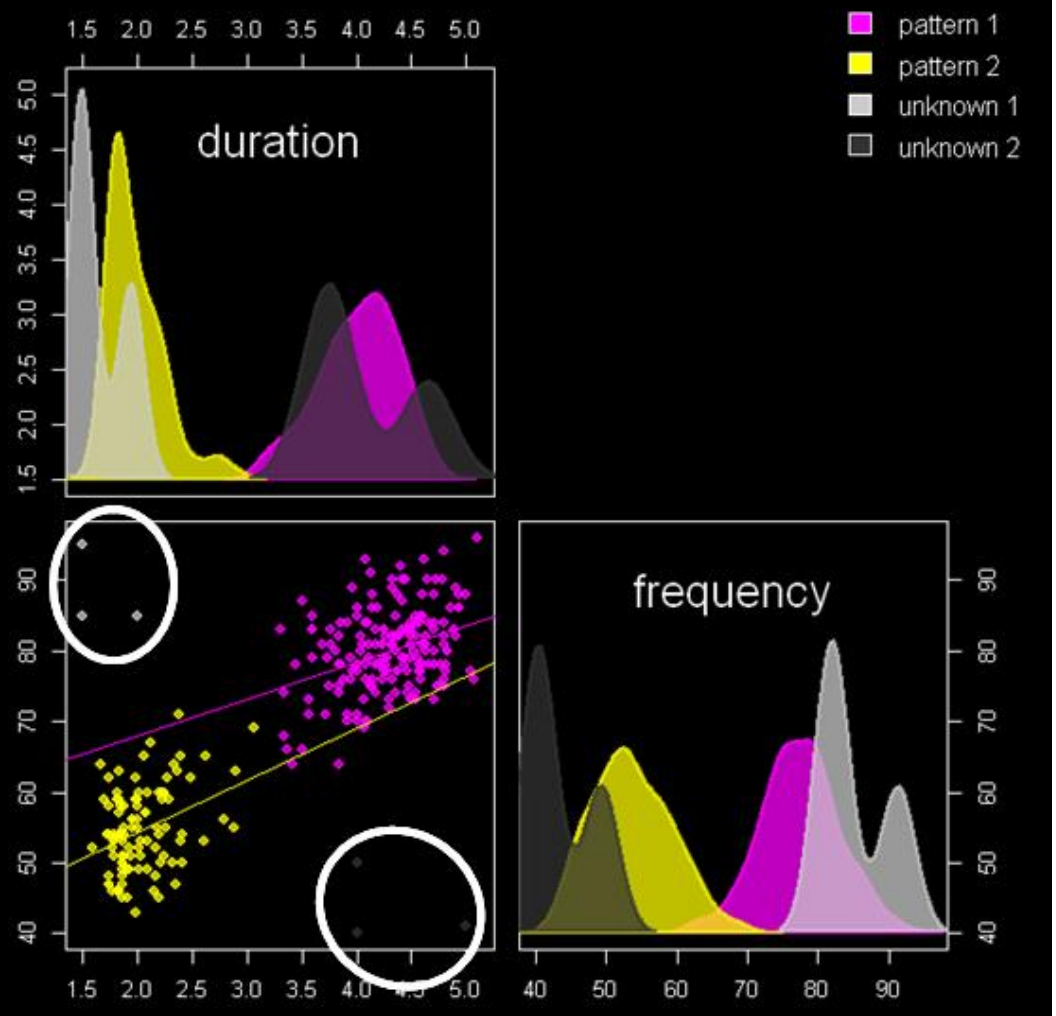


Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics* **39**, 357–365

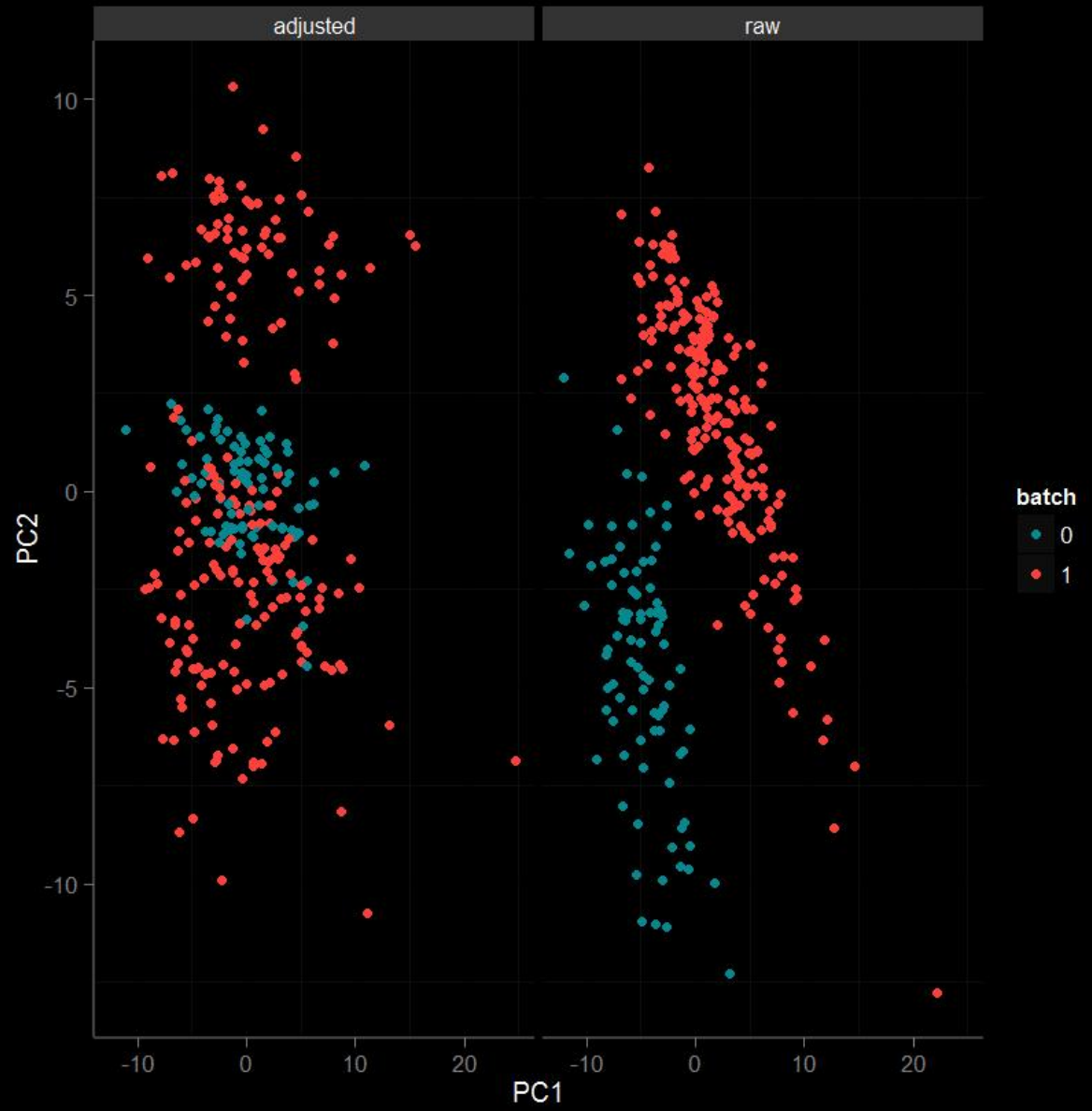
Bivariate Example

Noted deviations from two cluster pattern

- –Outliers?
- –Covariates?



Covariates



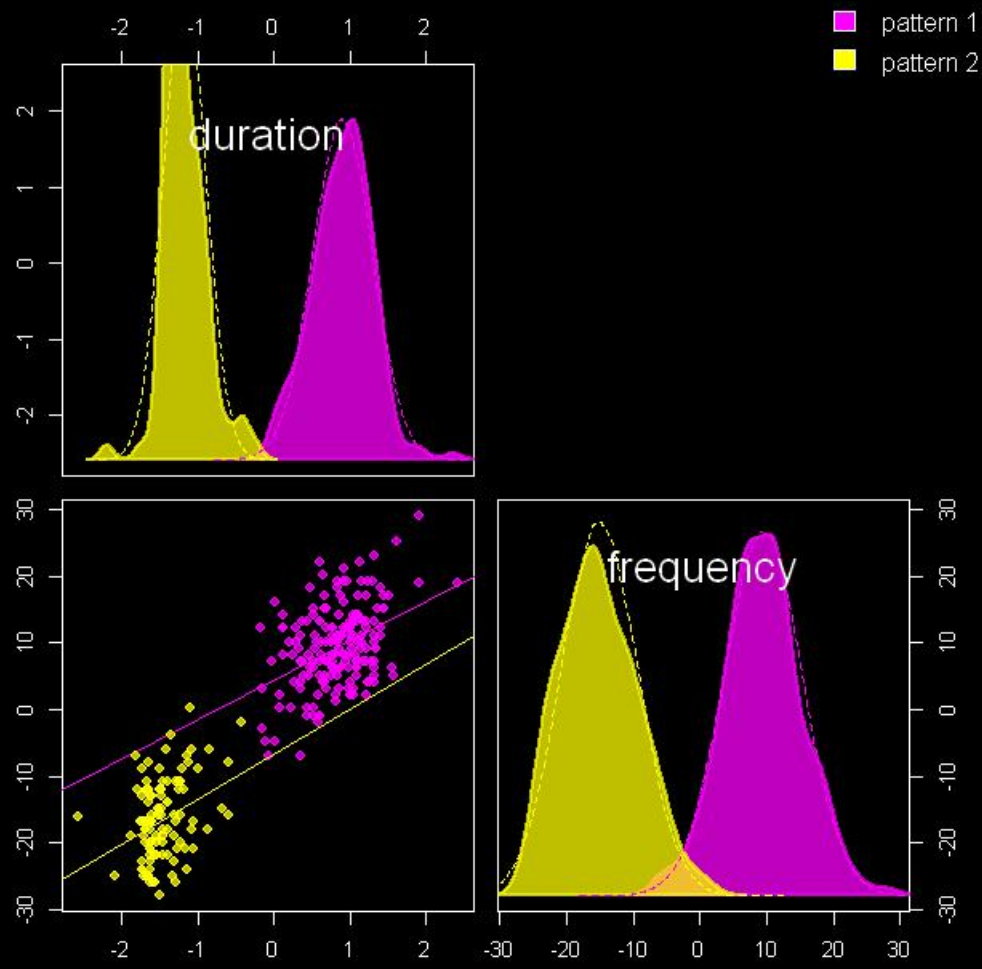
Trends in data which mask primary goals can be accounted for using covariate adjustment and appropriate modeling strategies

Bivariate Example

Noted deviations from two cluster pattern can be explained by covariate:

Hydrofraking ☹

Covariate adjustment is an integral aspect of statistical analyses (e.g. ANCOVA)



Summary

Data exploration and pre-analysis:

- increase robustness of results
- guards against spurious findings
- Can greatly improve primary analyses

Univariate Statistics:

- are useful for identification of statically significant changes or relationships
- sub-optimal for wide data
- best when combined with advanced multivariate techniques

Resources

Web-based data analysis platforms

- MetaboAnalyst(http://www.metaboanalyst.ca/MetaboAnalyst/faces/Home.jsp)
- MeltDB(https://meltdb.cebitec.uni-bielefeld.de/cgi-bin/login.cgi)

Programming tools

- The **R** Project for Statistical Computing(http://www.r-project.org/)
- Bioconductor(http://www.bioconductor.org/)

GUI tools

- imDEV(http://sourceforge.net/projects/imdev/?source=directory)