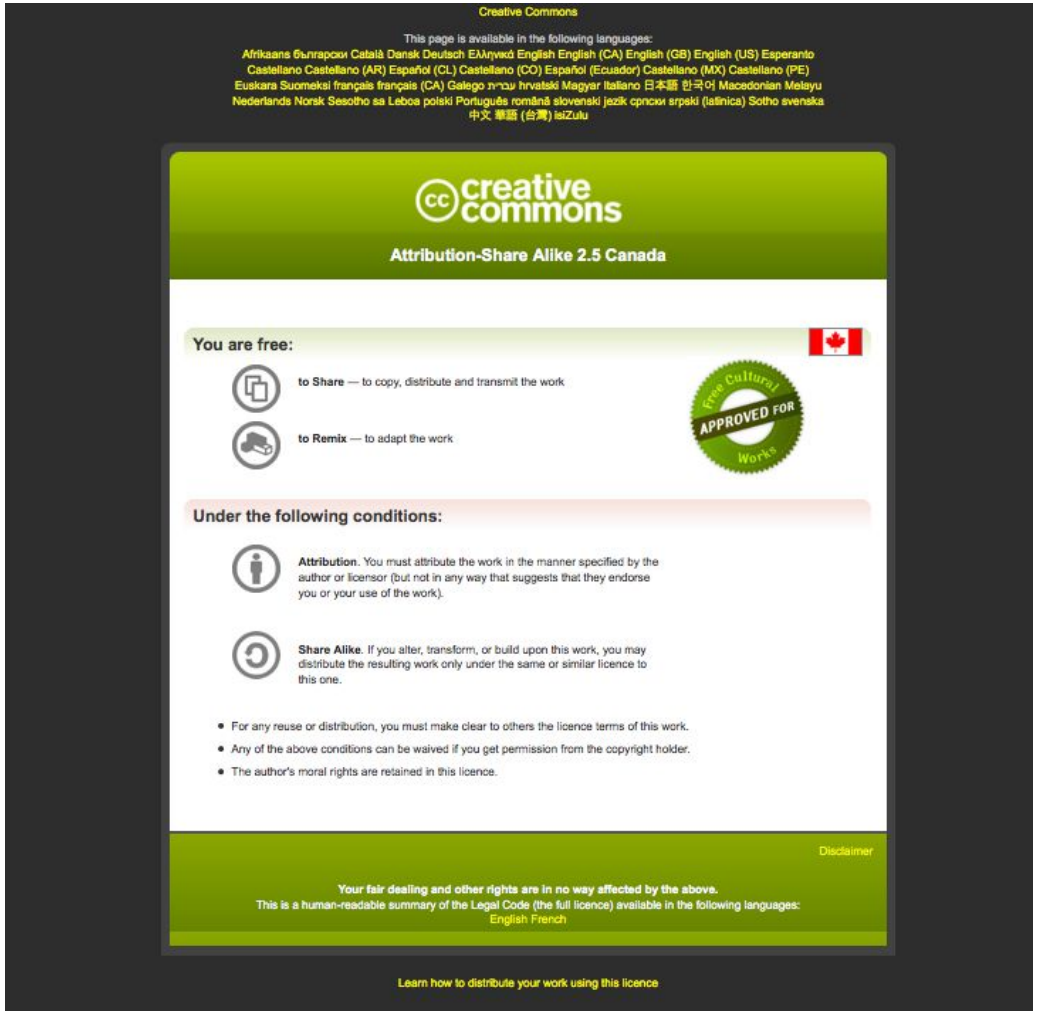
bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io


Creative Commons
This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu
creative commons
Attribution-Share Alike 2.5 Canada
You are free:
to Share — to copy, distribute and transmit the work
to Remix — to adapt the work
Free Cultural
APPROVED FOR
Works
Under the following conditions:
Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.
For any reuse or distribution, you must make clear to others the licence terms of this work.
Any of the above conditions can be waived if you get permission from the copyright holder.
The author's moral rights are retained in this licence.
Disclaimer
Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French
Learn how to distribute your work using this licence

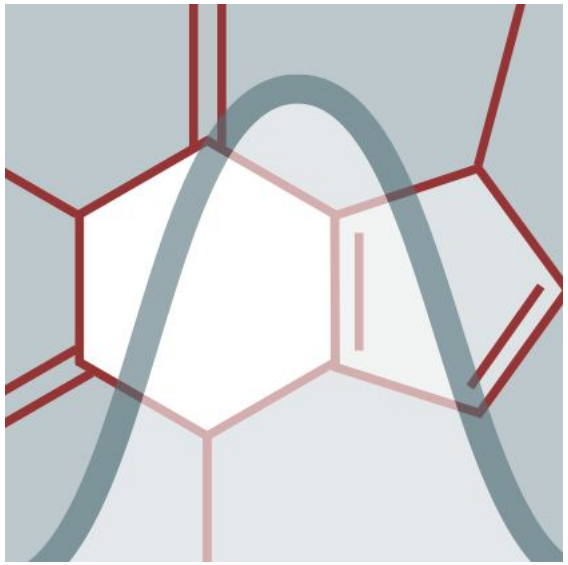
Backgrounder in omics data science

Jianguo (Jeff) Xia

Metabolomics Analysis

July 6-7, 2023

bioinformatics.ca



McGill

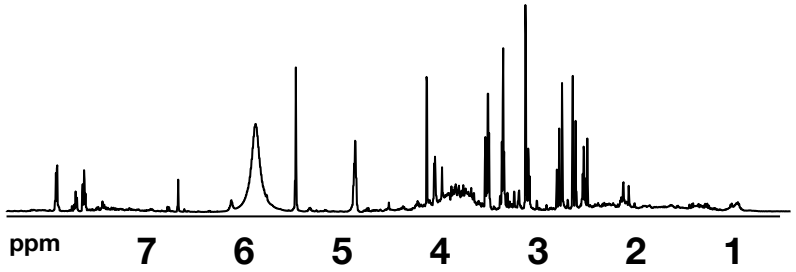
Schedule For July 7, 2023

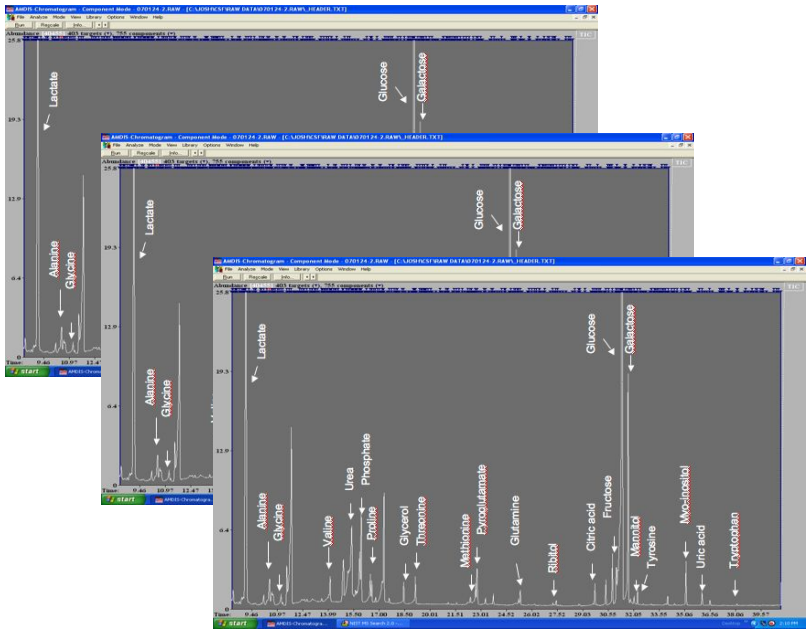
Time	Module
9:00 (MST)/11:00 (EST)	Module 5: Backgrounder in Omics Data Science (Jeff Xia)
10:30 (MST)/12:30 (EST)	Break/Lunch (45 min)
11:15 (MST)/13:15 (EST)	Module 6: Data Analytics for Untargeted Metabolomics (Jeff Xia)
12:15 (MST)/14:15 (EST)	Lunch/Break (45 min)
13:00 (MST)/15:00 (EST)	Module 7 (Lab): Metabolomics Data Analysis using MetaboAnalyst 5.0 (Jeff Xia)
15:00 (MST)/17:00 (EST)	Break (30 min)
15:30 (MST)/17:30 (EST)	Module 8: Integrating Metabolomics with other Omics (Jeff Xia)
17:00 (MST)/19:00 (EST)	Finish

Learning Objectives

1. Learn about basic concepts in ‘omics data analysis
2. Learn about p values calculation and interpretation
3. Learn about multivariate statistics (PCA and PLS-DA)
4. Learning about machine learning concepts (clustering, classification and performance evaluation)

Yesterday

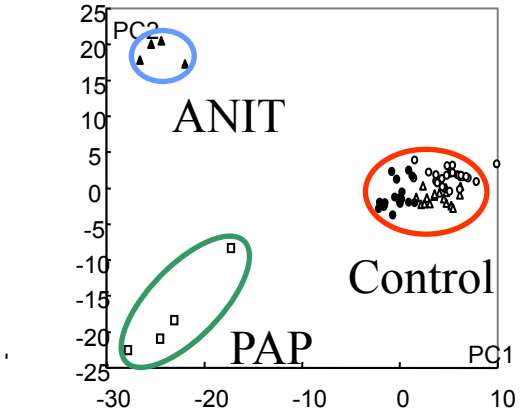


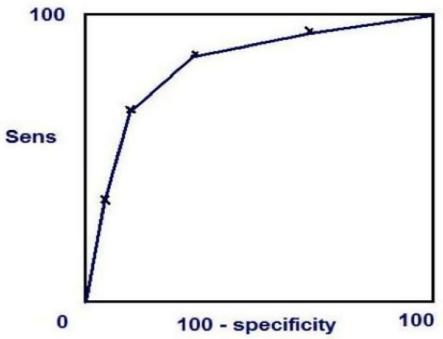


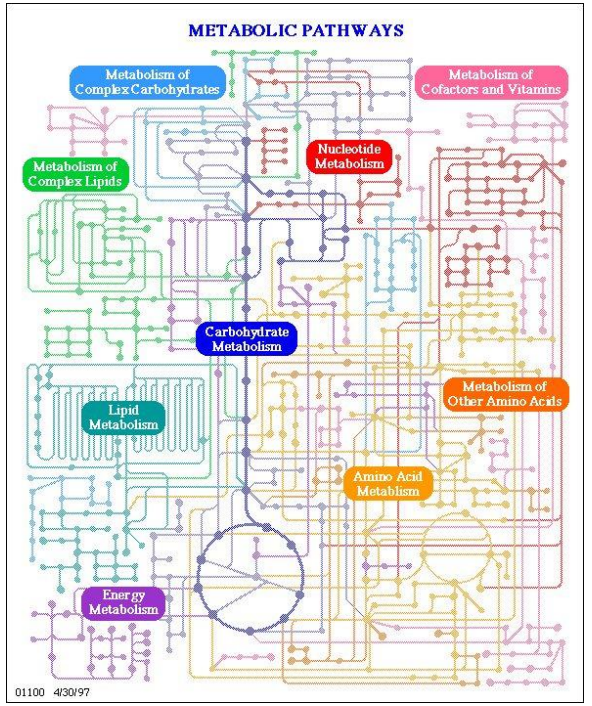
Compound	Retention Time (min)	Conc. in Urine (μM)	Compound	Retention Time (min)	Conc. in Urine (μM)
Dns-o-phospho -L-serine	0.92	<D.L.*	**Dns-Ile**	6.35	25
Dns-o-phospho -L-tyrosine	0.95	<D.L.	**Dns-3-aminosalicylic acid**	6.44	0.5
Dns-adnosine monophosphate	0.99	<D.L.	**Dns-pipecolic acid**	6.50	0.5
Dns-o-phosphoethanolamine	1.06	16	**Dns-Leu**	6.54	54
Dns-glucosamine	1.06	22	**Dns-cystathionine**	6.54	0.3
Dns-o-phospho -L-threonine	1.09	<D.L.	**Dns-Leu-Pro**	6.60	0.4
Dns-6-dimet hylamine purine	1.20	<D.L.	**Dns-5-hydroxylysine**	6.65	1.6
Dns-3-methyl -histidine	1.22	80	**Dns-Cystine**	6.73	160
Dns-taurine	1.25	834	**Dns-N-norleucine**	6.81	0.1
Dns-carnosine	1.34	28	Dns-5-hydroxydopamine	7.17	<D.L.
Dns-Arg	1.53	36	**Dns-dimethylamine**	7.33	293
Dns-Asn	1.55	133	**Dns-5-HIAA**	7.46	18
Dns-hypotaurine	1.58	10	**Dns-umbelliferone**	7.47	1.9
Dns-homocarnosine	1.61	3.9	Dns-2,3-diaminopropionic acid	7.63	<D.L.
Dns-guanidine	1.62	<D.L.	**Dns-L-ornithine**	7.70	15
Dns-Gln	1.72	633	**Dns-4-acetyamidophenol**	7.73	51
Dns-allantoin	1.83	3.8	**Dns-procaine**	7.73	8.9
Dns-L-citrulline	1.87	2.9	**Dns-homocystine**	7.76	3.3
Dns-1 (or 3 -)-methylhistamine	1.94	1.9	**Dns-acetaminophen**	7.97	82
Dns-adenosine	2.06	2.6	**Dns-Phe-Phe**	8.03	0.4
Dns-methylguanidine	2.20	<D.L.	**Dns-5-methyo xysalicylic acid**	8.04	2.1
Dns-Ser	2.24	511	**Dns-Lys**	8.16	184
Dns-aspartic acid amide	2.44	26	Dns-aniline	8.17	<D.L.
Dns-4-hydroxy -proline	2.56	2.3	**Dns-leu-Phe**	8.22	0.3
Dns-Glu	2.57	21	**Dns-His**	8.35	1550
Dns-Asp	2.60	90	Dns-4-thialysine	8.37	<D.L.
Dns-Thr	3.03	157	Dns-benzylamine	8.38	<D.L.
Dns-epinephrine	3.05	<D.L.	**Dns-1-ephedrine**	8.50	0.6
Dns-ethanolamine	3.11	471	**Dns-tryptamine**	8.63	0.4
Dns-aminoadipic acid	3.17	70	Dns-pyridoxamine	8.94	<D.L.
Dns-Gly	3.43	2510	Dns-2-methyl -benzylamine	9.24	<D.L.
Dns-Ala	3.88	593	**Dns-5-hydroxytrptophan**	9.25	0.12
Dns-aminolevulinic acid	3.97	30	**Dns-1,3-diaminopropane**	9.44	0.23
Dns-r-amino-butyric acid	3.98	4.6	**Dns-putrescine**	9.60	0.5
Dns-p-amino-hippuric acid	3.98	2.9	**Dns-1,2-diaminopropane**	9.66	0.1
Dns-5-hydroxymethyluricil	4.58	1.9	**Dns-tyrosinamide**	9.79	29
Dns-tryptophanamide	4.70	5.5	**Dns-dopamine**	10.08	140
Dns-isoguanine	4.75	<D.L.	**Dns-cadaverine**	10.08	0.08
Dns-5-aminopentanoic acid	4.79	1.6	**Dns-histamine**	10.19	0.4
Dns-sarcosine	4.81	7.2	**Dns-3-methoxy -tyramine**	10.19	9.2
Dns-3-amino-isobutyrate	4.81	85	**Dns-Tyr**	10.28	321
Dns-2-aminobutyric acid	4.91	17	Dns-cysteamine	10.44	<D.L.

Today

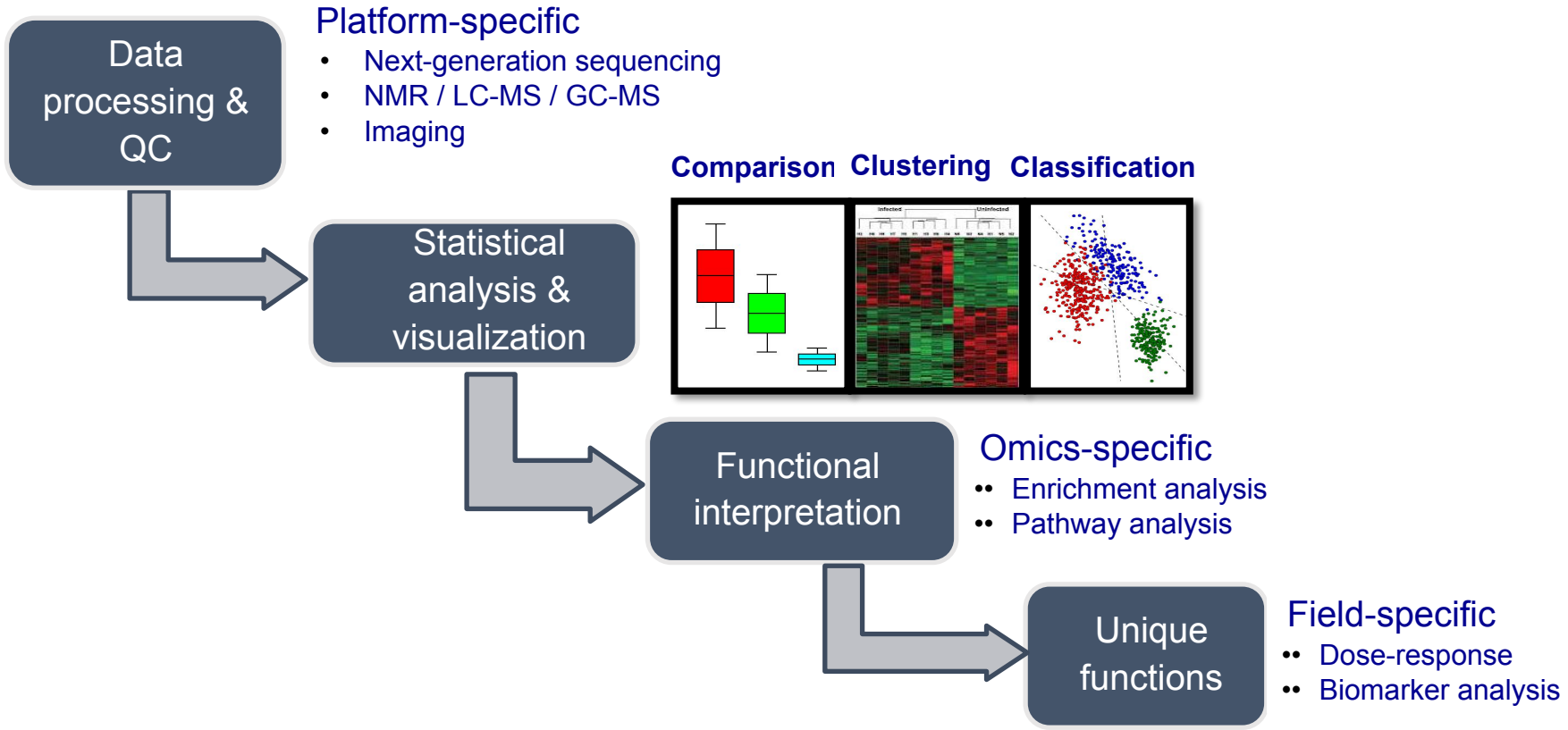
Compound	Retention Time (min)	Conc. in Urine (μM)	Compound	Retention Time (min)	Conc. in Urine (μM)
Dns-o-phospho -L-serine	0.92	<D.L.*	**Dns-Ile**	6.35	25
Dns-o-phospho -L-tyrosine	0.95	<D.L.	**Dns-3-aminosalicylic acid**	6.44	0.5
Dns-adnosine monophosphate	0.99	<D.L.	**Dns-pipecolic acid**	6.50	0.5
Dns-o-phosphoethanolamine	1.06	16	**Dns-Leu**	6.54	54
Dns-glucosamine	1.06	22	**Dns-cystathionine**	6.54	0.3
Dns-o-phospho -L-threonine	1.09	<D.L.	**Dns-Leu-Pro**	6.60	0.4
Dns-6-dimet hylamine purine	1.20	<D.L.	**Dns-5-hydroxylysine**	6.65	1.6
Dns-3-methyl -histidine	1.22	80	**Dns-Cystine**	6.73	160
Dns-taurine	1.25	834	**Dns-N-norleucine**	6.81	0.1
Dns-carnosine	1.34	28	Dns-5-hydroxydopamine	7.17	<D.L.
Dns-Arg	1.53	36	**Dns-dimethylamine**	7.33	293
Dns-Asn	1.55	133	**Dns-5-HIAA**	7.46	18
Dns-hypotaurine	1.58	10	**Dns-umbelliferone**	7.47	1.9
Dns-homocarnosine	1.61	3.9	Dns-2,3-diaminoproprionic acid	7.63	<D.L.
Dns-guanidine	1.62	<D.L.	**Dns-L-ornithine**	7.70	15
Dns-Gln	1.72	633	**Dns-4-acetyamidophenol**	7.73	51
Dns-allantoin	1.83	3.8	**Dns-procaine**	7.73	8.9
Dns-L-citrulline	1.87	2.9	**Dns-homocystine**	7.76	3.3
Dns-1 (or 3 -)-methylhistamine	1.94	1.9	**Dns-acetaminophen**	7.97	82
Dns-adenosine	2.06	2.6	**Dns-Phe-Phe**	8.03	0.4
Dns-methylguanidine	2.20	<D.L.	**Dns-5-methyo xysalicylic acid**	8.04	2.1
Dns-Ser	2.24	511	**Dns-Lys**	8.16	184
Dns-aspartic acid amide	2.44	26	Dns-aniline	8.17	<D.L.
Dns-4-hydroxy -proline	2.56	2.3	**Dns-leu-Phe**	8.22	0.3
Dns-Glu	2.57	21	**Dns-His**	8.35	1550
Dns-Asp	2.60	90	Dns-4-thialysine	8.37	<D.L.
Dns-Thr	3.03	157	Dns-benzylamine	8.38	<D.L.
Dns-epinephrine	3.05	<D.L.	**Dns-1-ephedrine**	8.50	0.6
Dns-ethanolamine	3.11	471	**Dns-tryptamine**	8.63	0.4
Dns-aminoadipic acid	3.17	70	Dns-pyridoxamine	8.94	<D.L.
Dns-Gly	3.43	2510	Dns-2-methyl -benzylamine	9.24	<D.L.
Dns-Ala	3.88	593	**Dns-5-hydroxytrptophan**	9.25	0.12
Dns-aminolevulinic acid	3.97	30	**Dns-1,3-diaminopropane**	9.44	0.23
Dns-r-amino-butyric acid	3.98	4.6	**Dns-putrescine**	9.60	0.5
Dns-p-amino-hippuric acid	3.98	2.9	**Dns-1,2-diaminopropane**	9.66	0.1
Dns-5-hydroxymethyluricil	4.58	1.9	**Dns-tyrosinamide**	9.79	29
Dns-tryptophanamide	4.70	5.5	**Dns-dopamine**	10.08	140
Dns-isoguanine	4.75	<D.L.	**Dns-cadaverine**	10.08	0.08
Dns-5-aminopentanoic acid	4.79	1.6	**Dns-histamine**	10.19	0.4
Dns-sarcosine	4.81	7.2	**Dns-3-methoxy -tyramine**	10.19	9.2
Dns-3-amino-isobutyrate	4.81	85	**Dns-Tyr**	10.28	321
Dns-2-aminobutyric acid	4.91	17	Dns-cysteamine	10.44	<D.L.



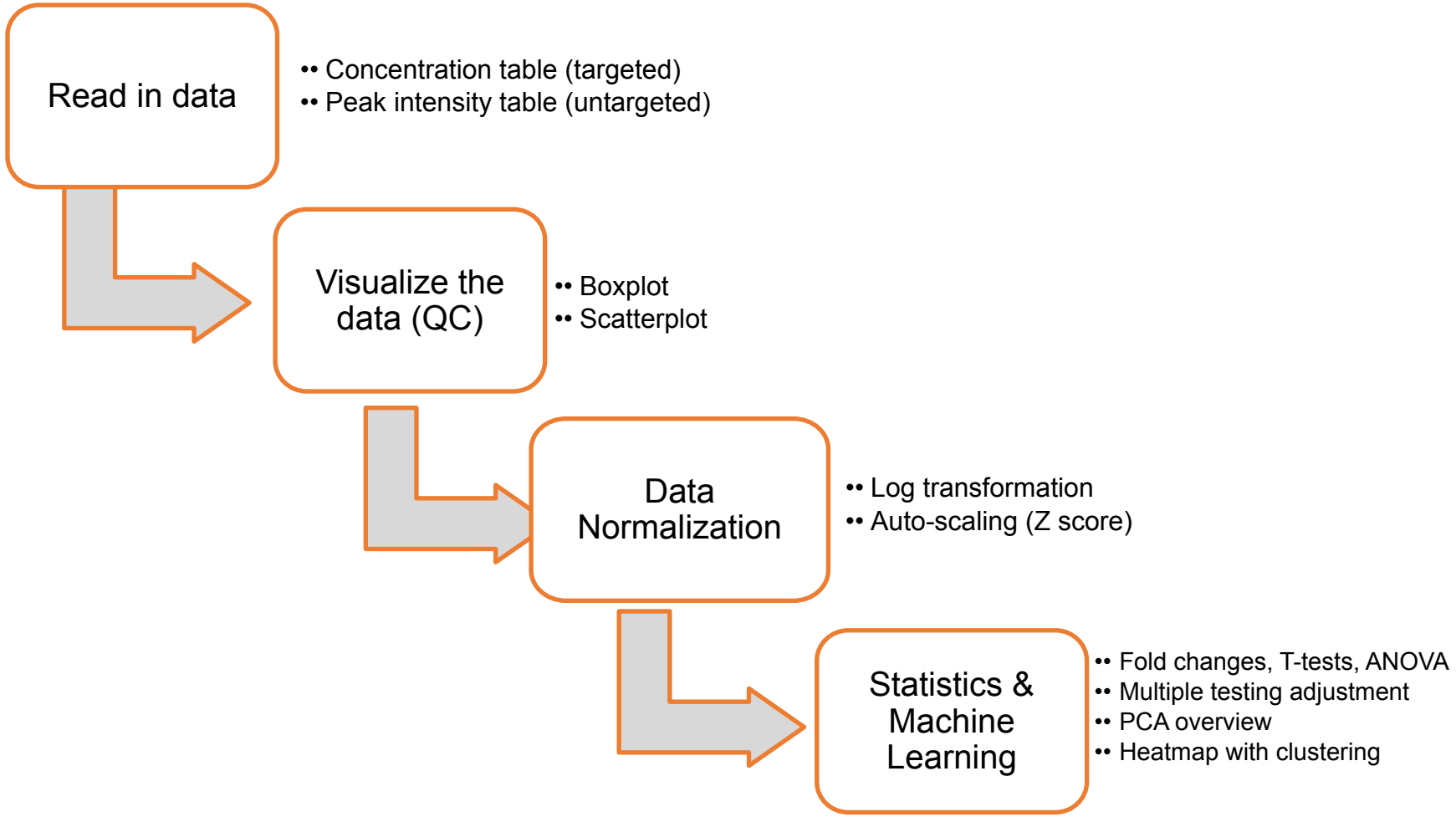




General Steps in Omics Data Analysis



General Steps in Statistical Analysis



Input & Output

Input: metabolomics data

- X: a table containing numerical values
  - Concentrations, peak intensities
- Y: meta-data: data about data
  - Class labels, experimental factors

Output: useful information

- Significant features
- Clustering patterns
- Rules (for prediction)
- Models ……

Preparing Data

X: quantitative data

- Continuous
  - Microarray intensities
  - Metabolite concentrations
- Discrete
  - Read counts (RNAseq)
- Need to be treated with different statistical models
  - Normal distribution
  - Poisson distribution

ENSMUSG00000025577	2	0	0
ENSMUSG00000055612	5	5	0
ENSMUSG00000022516	118	78	48
ENSMUSG00000022404	180	98	60
ENSMUSG00000022765	134	95	44
ENSMUSG00000022982	115	83	43
ENSMUSG00000050106	519	124	50
ENSMUSG00000056366	135	61	28
ENSMUSG00000020422	724	348	258
ENSMUSG00000051703	10	10	2
ENSMUSG00000025465	104	72	33
ENSMUSG00000040760	227	162	97
ENSMUSG00000029253	97	61	40
ENSMUSG00000030742	8	1	0
ENSMUSG00000055333	14	2	2
ENSMUSG00000047757	12	31	8
ENSMUSG00000078719	38	21	7
ENSMUSG00000079470	16	18	4
ENSMUSG00000020806	6294	3284	1988
ENSMUSG00000006378	33	72	34

Sample	Label	Acetate	Acetone	Alanine	Betaine	Carnitine	Choline
Patient1_d0	1	189.07	24.24	266.27	298.95	304.62	305.61
Patient2_d0	2	386.52	26.29	612.02	313.5	122.06	78.46
Patient3_d0	3	506.28	27.84	347.32	101.49	88.82	35.88
Patient4_d0	4	51	177.56	468.42	172.65	133.28	0
Patient5_d0	5	733.45	33.84	1080.82	239.52	89.21	187.98
Patient1_d3	-1	315.12	19.95	169.05	37.78	32.04	122.4
Patient2_d3	-2	325.39	13.29	295.53	85.1	89.36	44.93
Patient3_d3	-3	231.59	12.91	226.63	0	0	35.26
Patient4_d3	-4	285.55	0	217.43	0	77.79	0
Patient5_d3	-5	353.51	15.87	699.81	98.7	458.81	112.21

Y: metadata (data describing X)

- Binary data
  - 0/1, Y/N, Case/Control
- Nominal Data (> two groups)
  - Single = 1, Married = 2, Divorced = 3, Widowed = 4
  - Order is not important
- Ordinal data
  - Time series
  - Disease severities
  - Dose responses

```
Sample  Disease
CD-6KUCT.zip    CD
CD-9OS5Y.zip    CD
CD-9WOBP.zip    CD
CD-77FXR.zip    CD
HC-9SNJ4.zip    Control
HC-9X47O.zip    Control
HC-AMR37.zip    Control
HC-AUP8B.zip    Control
QC_PREFA02.zip  QC
QC_PREFB02.zip  QC
```

Sample	Diagnosis	Gender	Treatment	Age
S1	COVID	Male	non_Treated	62
S2	COVID	Male	non_Treated	44
S3	COVID	Male	Treated	54
S4	COVID	Male	non_Treated	62
S5	COVID	Male	Treated	82
S6	COVID	Male	Treated	65
S7	COVID	Female	Treated	49
S8	COVID	Female	Treated	42
S9	COVID	Female	Treated	56
S10	COVID	Female	Treated	56
S11	COVID	Female	Treated	69
S12	HC	Male	non_Treated	24
S13	HC	Female	non_Treated	38
S14	HC	Female	non_Treated	42
S15	HC	Female	non_Treated	40
S16	HC	Female	non_Treated	56
S17	HC	Male	non_Treated	57
S18	HC	Male	non_Treated	57
S19	HC	Male	non_Treated	60
S20	HC	Male	non_Treated	62
S21	HC	Male	non_Treated	55

Data Formatting for MetaboAnalyst

- Put together (X+Y). Requested by most modules

Sample	Label	Acetate	Acetone	Alanine	Betaine	Carnitine	Choline	Citrate	Creatine
Control_01	0	189.07	24.24	266.27	298.95	304.62	305.61	3969.16	366.7
Control_02	0	386.52	26.29	612.02	313.5	122.06	78.46	2075.6	435.99
Control_03	0	506.28	27.84	347.32	101.49	88.82	35.88	745.5	83.39
Control_04	0	51	177.56	468.42	172.65	133.28	0	10797.57	77.69
Control_05	0	733.4				21	187.98	1016.97	83.39
Disease_01	1	315.1				04	122.4	1486.6	152.5
Disease_02	1	325.3				36	44.93	3327.06	263.41
Disease_03	1	231.59	12.91	226.63	0	0	35.26	758.18	50.36
Disease_04	1	285.55	0	217.43	0	77.79	0	609.23	0
Disease_05	1	353.51	15.87	699.81	98.7	458.81	112.21	3415.49	229.79

Sample in rows

Sample	ko15	ko16	ko18	ko19	wt15	wt16	wt18	wt19
Label	KO	KO	KO	KO	WT	WT	WT	WT
200.1/2926	147887.53	451600.71	65290.38	56540.93	175177.08	82619.48	51951.61	69198.22
205/2791	1778569	1567038	1482796	1039130	1950287	1466781	1572679	1275313
206/2791	237993.6	269714	201393.4	150107.3	276541.8	222366.2	211717.7	186850.9
207.1/2719	380873	460629.7	351750.1	219288	417169.6	324892.5	277990.7	220972.4
219.1/2524	235544.92	173623.				5	72029.38	75096.99
231/2516	117649.77	48960.				4	96841.46	240261.21
233/3023	399145.3	35695'				1	334459.9	181901.3
234/3024	76880.87	99526.27	97493.76	53461.71	65215.64	55952.44	73781.01	45211.66
235.1/2695	171995.22	128945.16	155442.48	115286.25	199981.49	30028.6	156968.3	52596.48
236.1/2524	252282.04	206031.93	71763.79	73602.47	253791.07	187225.65	79389.63	90012.64

Sample in columns

- A separate metadata table (requested by **Statistical Analysis [metadata table]** module)

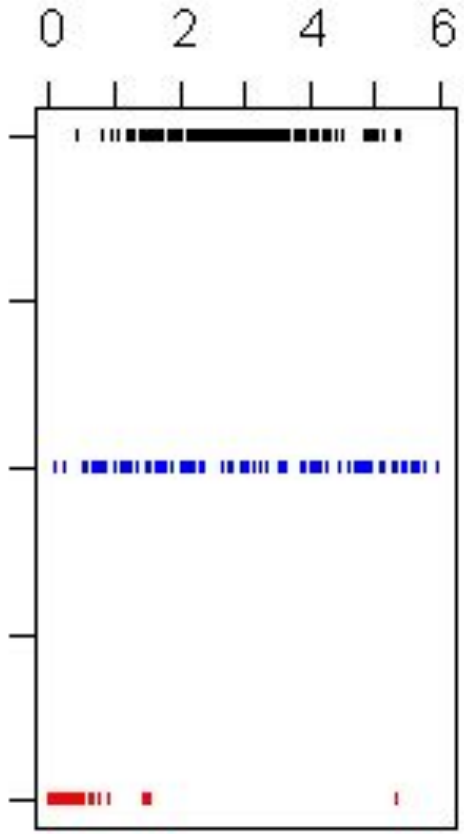
Sample	Diagnosis	Gender	Treatment	Age
S1	COVID	Male	non_Treated	62
S2	COVID	Male	non_Treated	44
S3	COVID	Male	Treated	54
S4	COVID	Male	non_Treated	62
S5	COVID	Male	Treated	82
S6	COVID	Male	Treated	65
S7	COVID	Female	Treated	49
S8	COVID	Female	Treated	42
S9	COVID	Female	Treated	56
S10	COVID	Female	Treated	56
S11	COVID	Female	Treated	69
S12	HC	Male	non_Treated	24
S13	HC	Female	non_Treated	38
S14	HC	Female	non_Treated	42
S15	HC	Female	non_Treated	40
S16	HC	Female	non_Treated	56
S17	HC	Male	non_Treated	57
S18	HC	Male	non_Treated	57
S19	HC	Male	non_Treated	60
S20	HC	Male	non_Treated	62
S21	HC	Male	non_Treated	55

Common Terms

- Feature / variable
  - Metabolites, peaks
- Dimension
  - The number of variables (metabolites, peaks)
- Univariate:
  - Measuring one variable per subject
- Multivariate
  - Measuring many variables per subject
  - Omics data are usually high-dimensional data

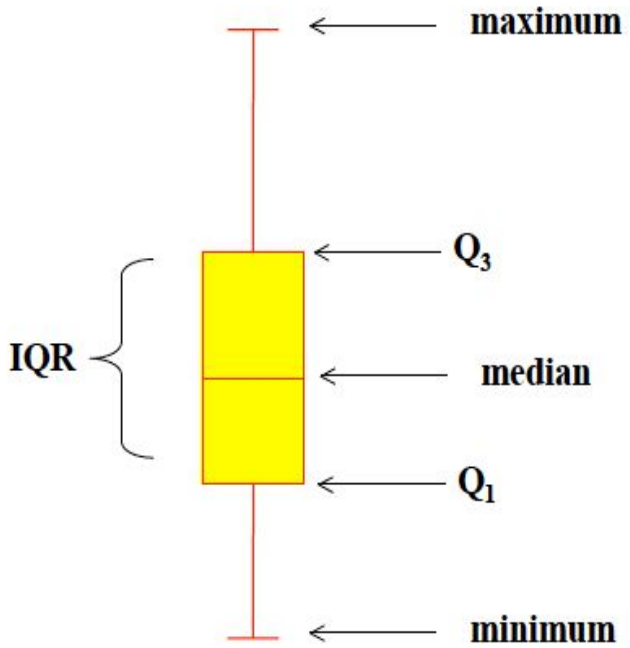
How Do We Describe Data?

- Central Tendency – center of the data location
  - Mean, Median, Mode
- Variability – the spread of the data
  - Variance
  - Standard deviation
- Relative standing – distribution of data within the spread
  - Quantiles
  - Range
  - IQR (inter-quantile range)



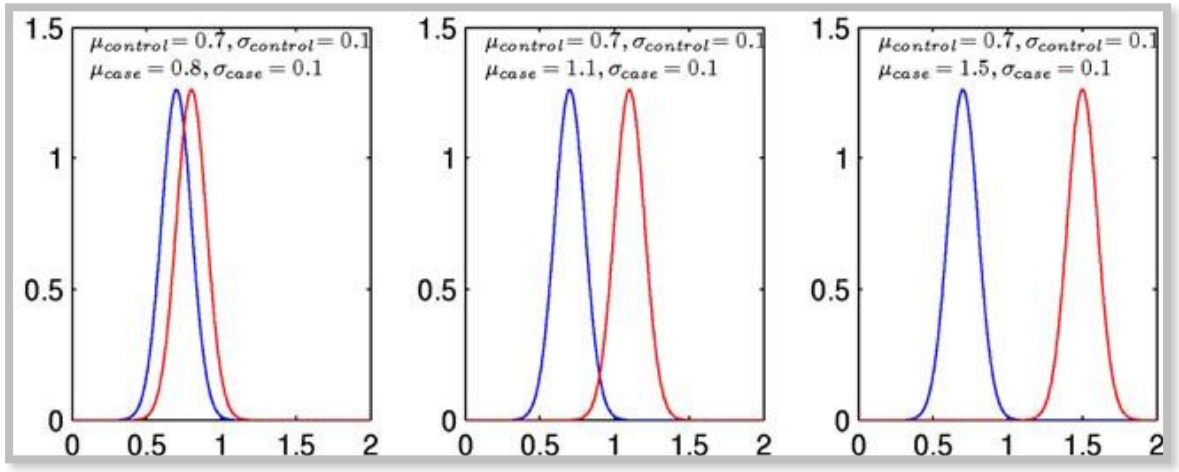
Box-and-whisker plot

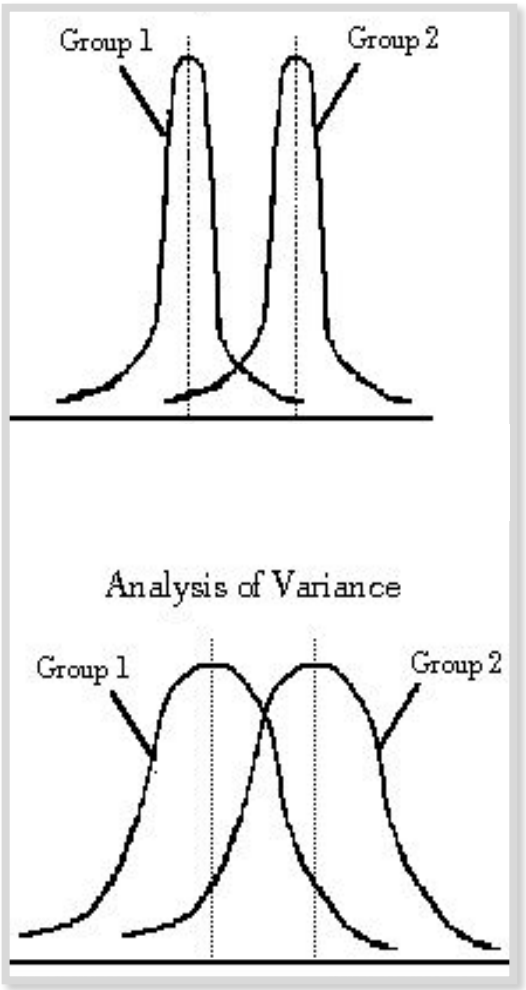
- The 1st quantile Q_1 is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as median (50% are smaller and 50% larger)
- Q_3 is the value that only 25% of the observations are larger
- Range is minimum to maximum



Mean vs. Variance

Most univariate tests compare the difference in the means, assuming equal variance

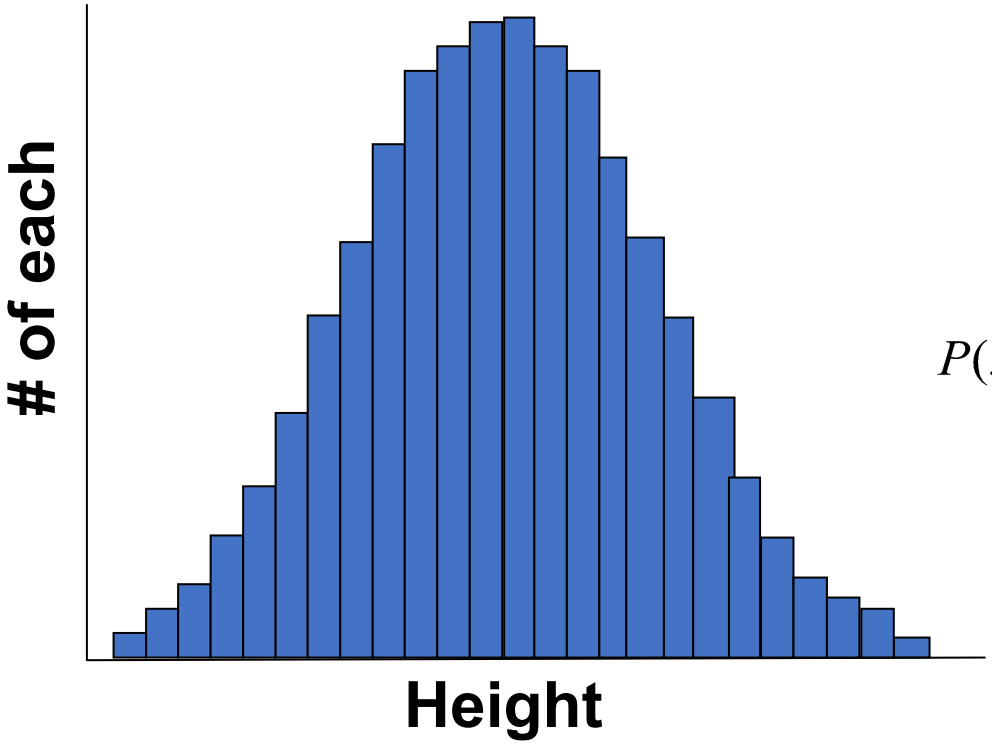




Normal or “Gaussian” Distribution

- Almost any set of biological or physical measurements will display some some variation and these will almost always follow a Normal distribution
- The larger the set of measurements, the more “normal” the curve
- Minimum set of measurements to get a normal distribution is 30-40

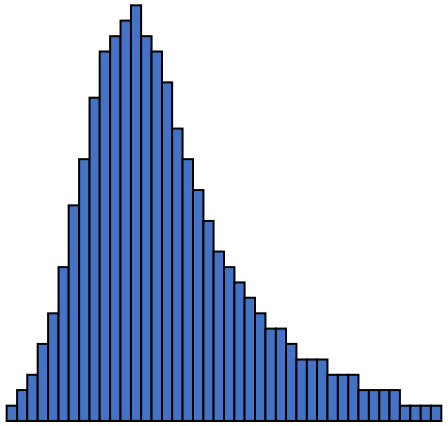
Normal Distribution -- a Bell Curve



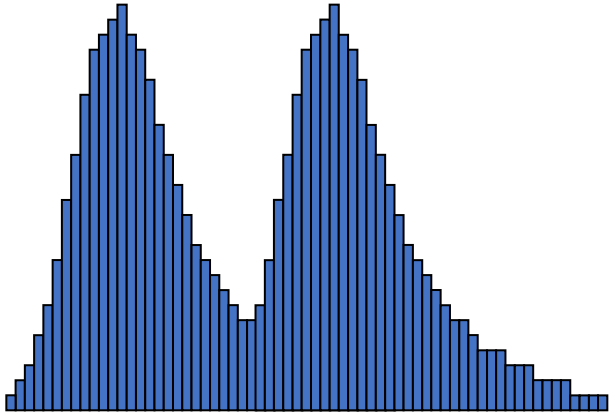
$$P(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Other distributions are common

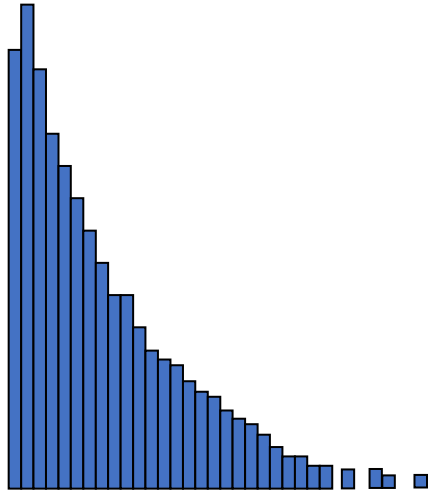
Unimodal



Bimodal



Skewed



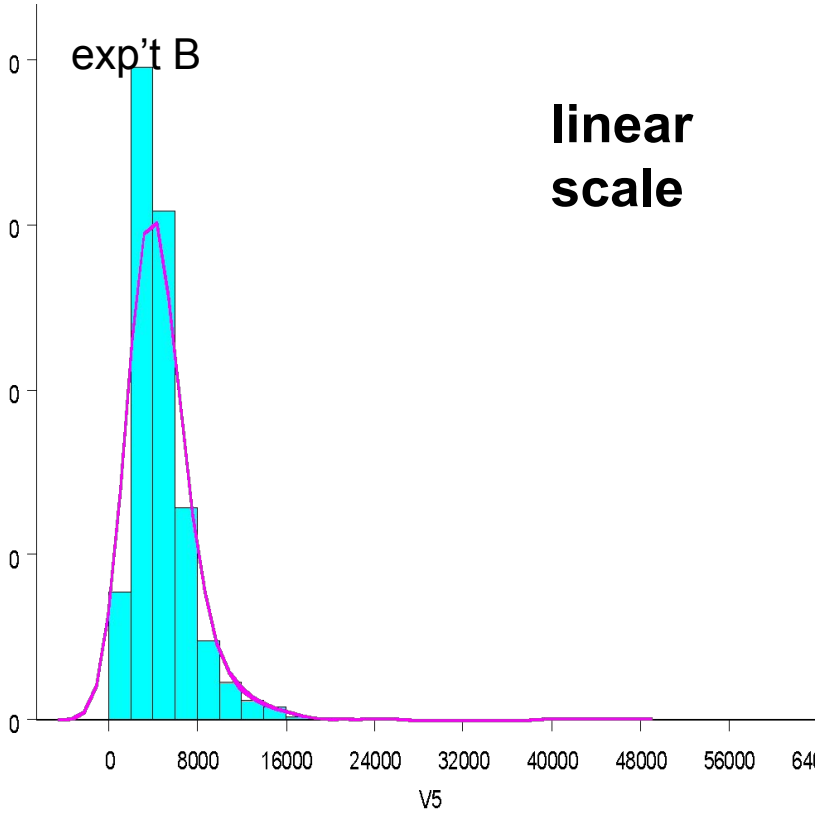
Data normalization

- Most statistics models have been developed based on the assumption that the underlying data are ”normally” distributed
  - They work suboptimal when this assumption is not met (i.e. high false positives)
- Solution - to transform data to close to normal distr.

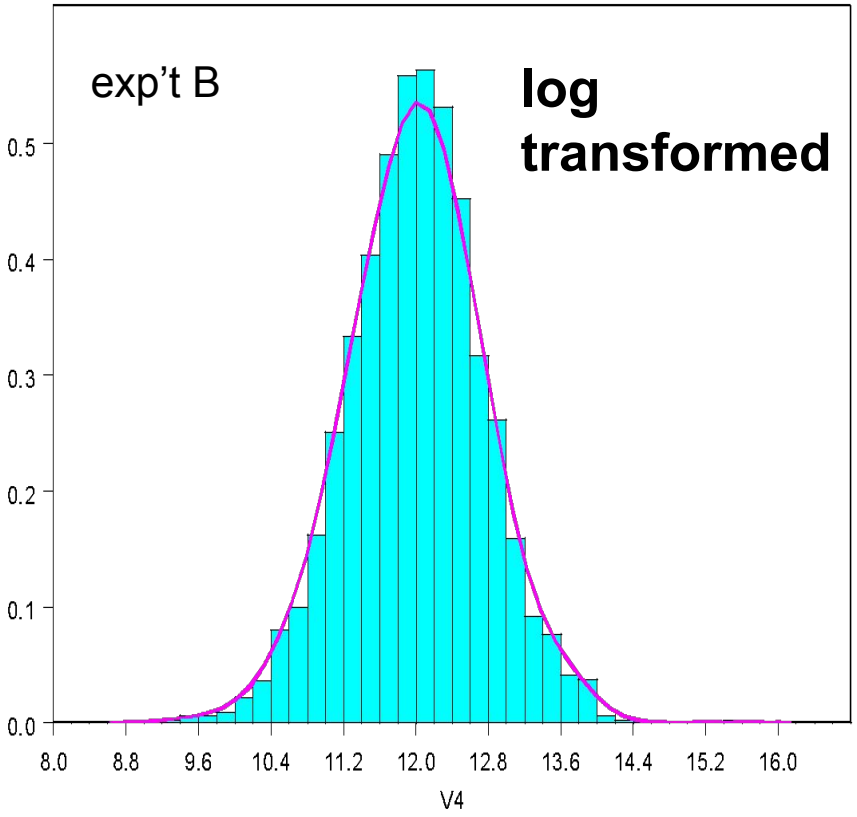
  normalization

Log Transformation

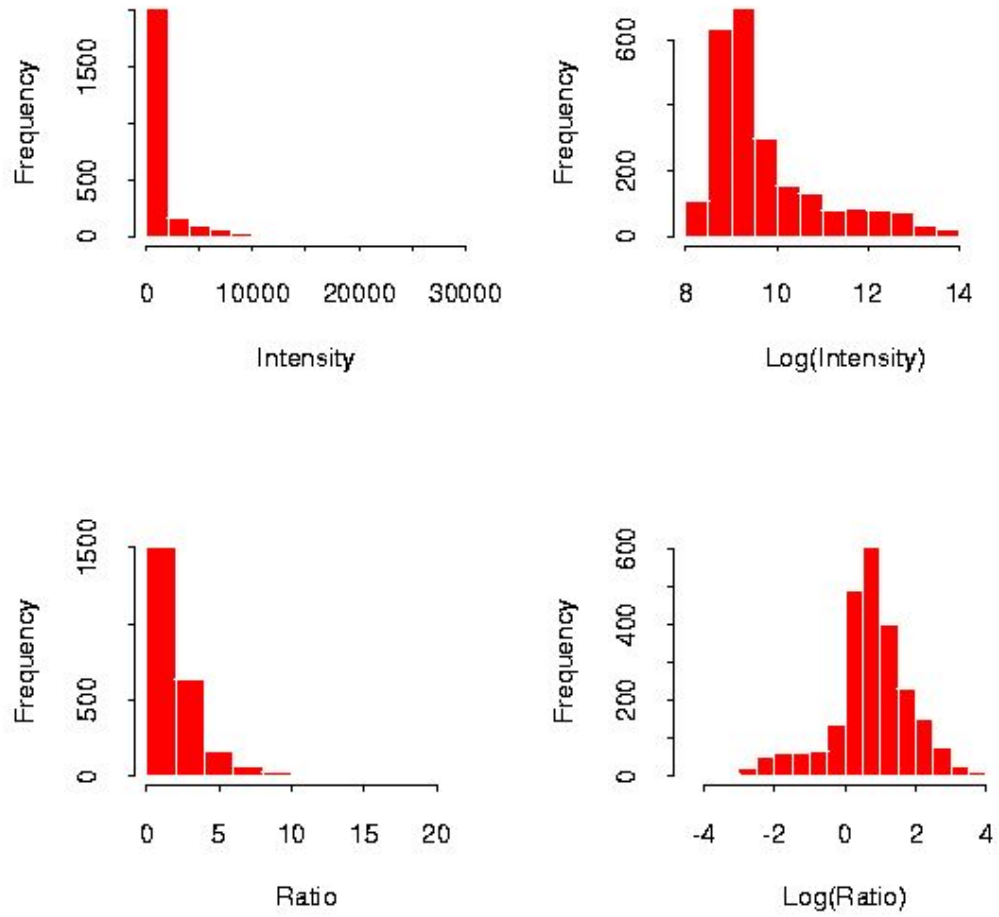
Skewed distribution



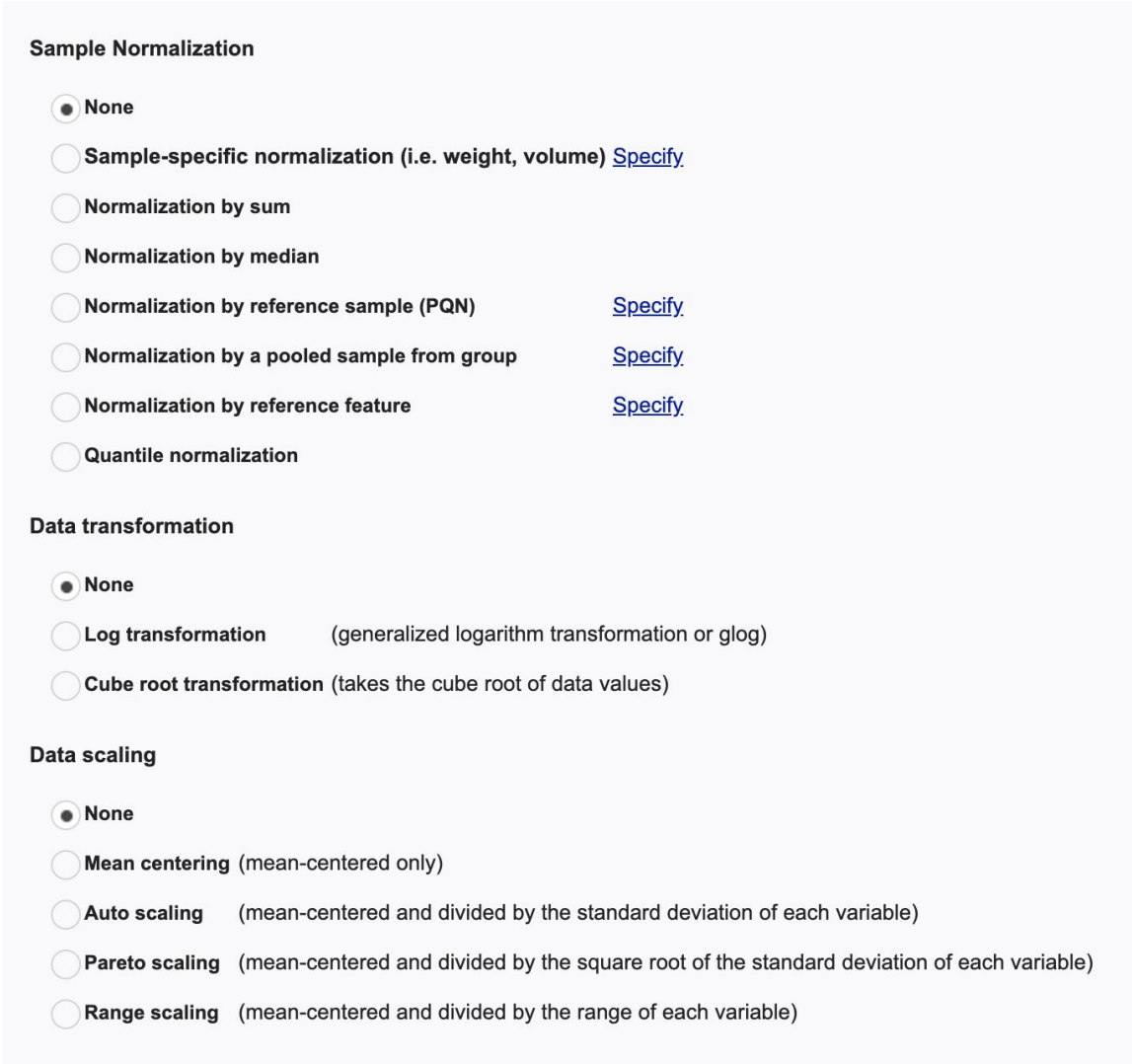
Normal distribution



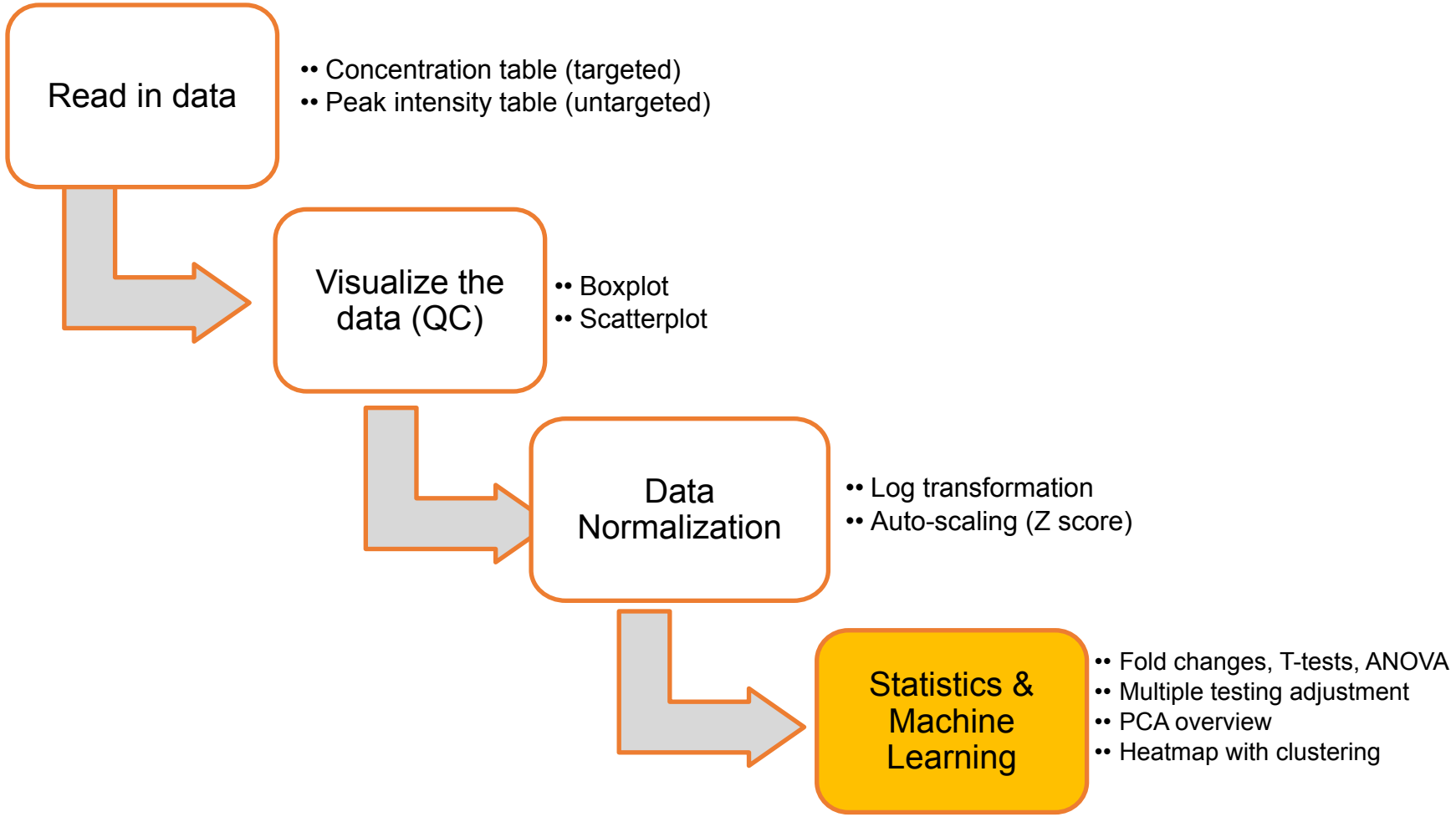
Log Transformation (real data)



Data normalization options in MetaboAnalyst

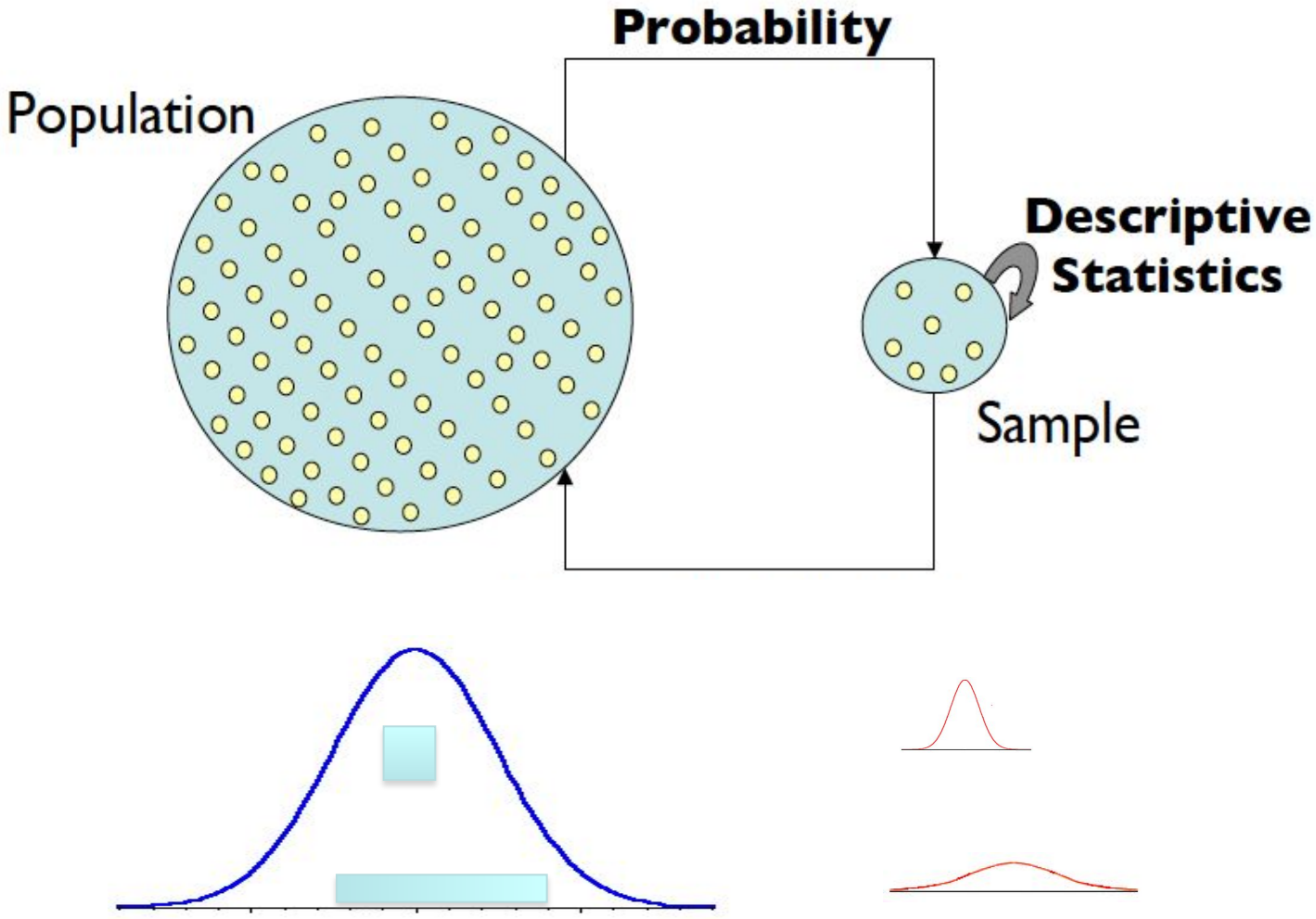


General Steps in Statistical Analysis



Understanding P values

Uncertainties in Parameter Estimation



From Samples to Population

- So how do we know whether the effect observed in our sample was genuine?
  - We don't

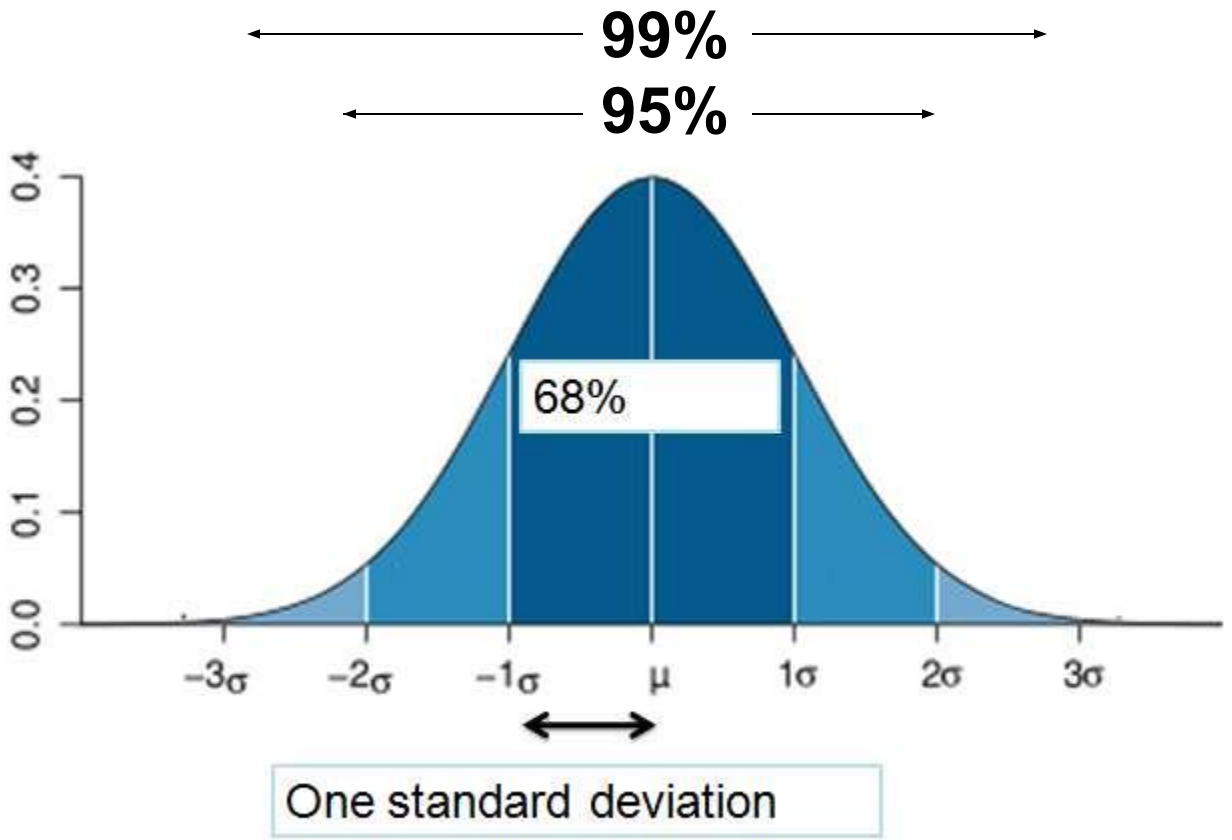
- Instead, we use *p values* to indicate our level of uncertainty that our results represent a genuine effect present in the whole population

P values

- P values = the probability that the observed result was obtained by chance
  - i.e. when the null hypothesis (H_0: *i.e.* no effect) is true
- If that probability (p-value) is small, it suggests the observed result cannot be easily explained by chance (i.e. random effects)

How do we calculate a p value?

When distribution is known



Empirical P-values

- Previously mentioned p-values are based on well defined models
  - *Gaussian* distributions, *Poisson* distribution
- What if we don't know the distribution?
  - The only thing we know is that the data does not follow a normal distribution
  - Poor performance using a model based on normal distribution
- We can find out the null distribution from the data itself, then calculate the p-value
  - Also known as empirical p-values

Basic steps

1. Under the null hypothesis (H_0), all data comes from the same distribution
2. We calculate our statistic, such as the mean difference, from the **original** data
3. We then shuffle the data with respect to group labels and recalculate the statistic (mean difference)
   - group labels do not matter under H_0
4. Repeat step 3 multiple times
5. Find out where our statistic lies in comparison to the null distribution

A Simple Example

To find out whether there is a mean difference between case vs. control

Mean difference .541

	case		control
1	-0.49274	10	1.471227
2	-0.30228	11	0.612679
3	0.093007	12	-0.47886
4	0.715722	13	0.746045
5	1.272872	14	0.871994
6	-1.37599	15	0.985237
7	-0.14798	16	-0.44421
8	-1.22195	17	0.246393
9	1.2812	18	0.68246
Mean	-0.01979		0.52144

Permutation #1

Note how the different labels have been swapped for the permutation

Mean difference = .329

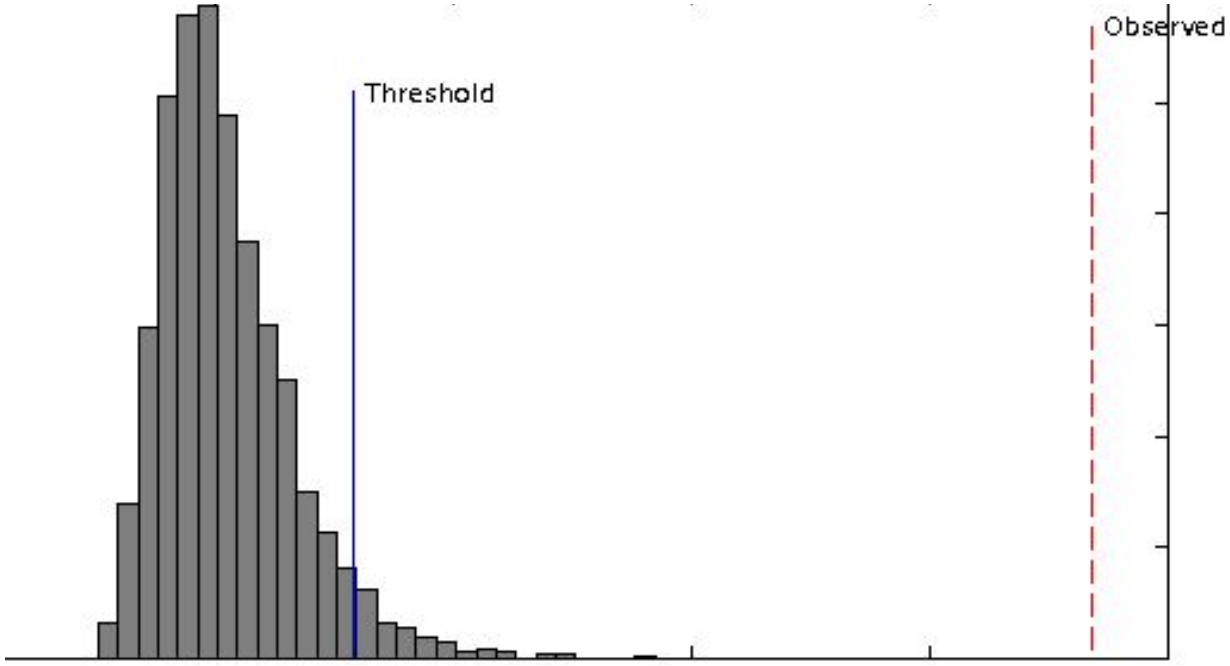
	case		control
9	1.2812	11	0.612679
3	0.093007	18	0.68246
17	0.246393	14	0.871994
15	0.985237	4	0.715722
16	-0.44421	6	-1.37599
1	-0.49274	2	-0.30228
7	-0.14798	5	1.272872
10	1.471227	12	-0.47886
13	0.746045	8	-1.22195
Mean	0.415354		0.086295

Permutations

Repeat many many times
(i.e. 1000 times)

	case		control
9	1.2812	11	0.612679
3	0.093007	18	0.68246
17	0.246393	14	0.871994
15	0.985237	4	0.715722
16	-0.44421	6	-1.37599
1	-0.49274	2	-0.30228
7	-0.14798	5	1.272872
10	1.471227	12	-0.47886
13	0.746045	8	-1.22195
Mean	0.415354		0.086295

Compute Empirical P-value



In 1000 times permutations

- There are three times the permuted data given large difference
  □ p = 0.003 (this is the chance under null hypothesis)
- If none of the permuted mean difference is bigger than the original one □ p < 0.001 or (1/1001) # prevent p-value equal to zero

General advantages

- Does not rely on distributional assumptions
- Corrects for hidden correlation
- Disadvantage?
  - Need relatively large number of samples
  - Computationally intensive
  - Not as powerful as when using the right model (if known)

Multiple Testing Issues

Omics yields high-dimensional data

- 100s to 10,000s of variables

Lots of hypothesis tests, with each one we accept a small chance

- Performing T-tests on typical metabolomic data might result in performing 10000 separate hypothesis tests. If we use a standard p value cut-off of 0.05, we would see 500 (10000*0.05) genes to be deemed “significant” **by chance alone!**

Multiple Testing Correction (I)

- Family Wise Error Rate (FWER) - e.g. *Bonferroni corrections*
  - Corrected P-value= p-value * n (number of genes in test) <0.05. If testing 1000 genes at a time, the corrected p-value is 0.00005 (0.05/1000).
- False Discovery Rate (FDR) - e.g. Benjamini-Hochberg
  - A FDR of 0.05 means that 5% among the significant metabolites are expected to be false positives

High-dimensional data

- So far, our analyses are dealing with a single variable (i.e. univariate analysis)
  - T-tests: one variable, two groups
  - ANOVA: one variable, > 2 groups
- First analyze a single variable, and then apply the procedure to all variables, finally do multiple test adjustment
- Visualization are limited to three dimensions
- How can we analyze & visualize these high-dimensional data **holistically (considering all data together)**?

Machine Learning & Multivariate Statistics

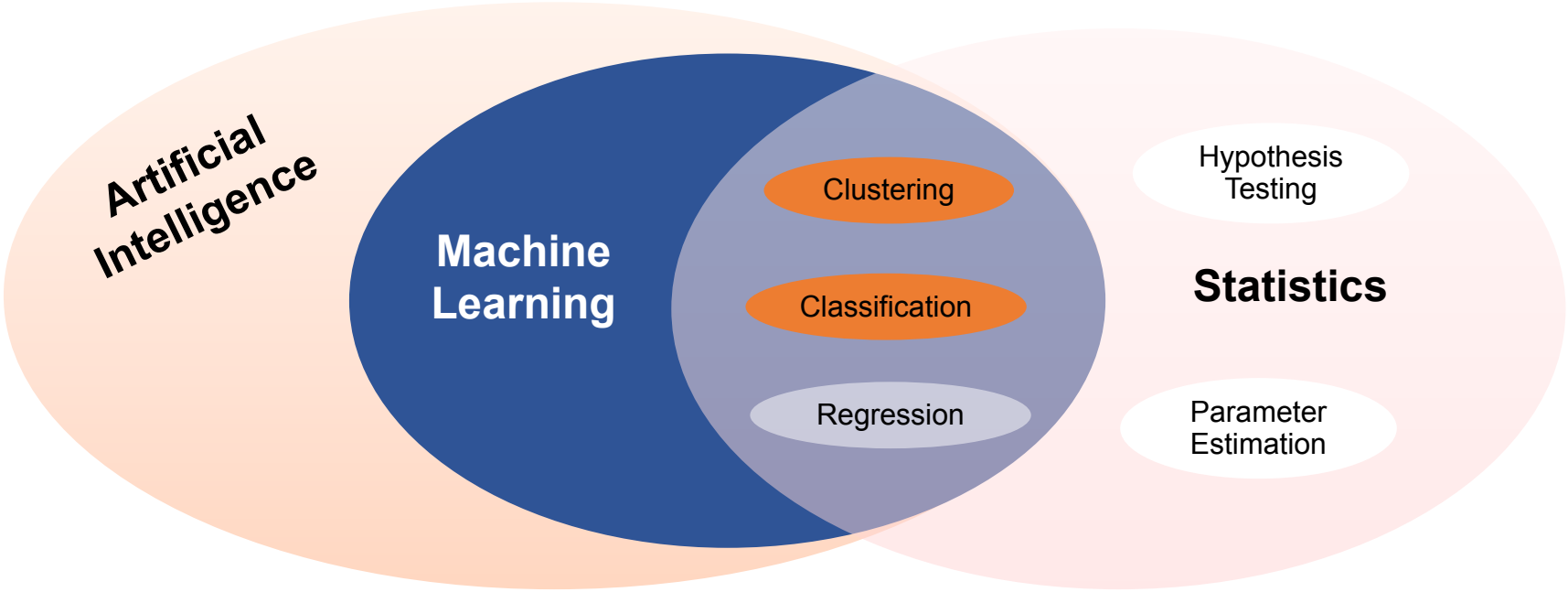
The Challenges

- Most statistical methods have been developed before the coming of the omics era
- Most statistical methods are designed for single or very few variables
  - T-tests, ANOVA
  - Linear/logistic regression
- These methods assume there are more samples (n) than the number of variables (p) (i.e. n > p) for parameter estimation
  - In omics data, n << p

Current Practices

- Classical multivariate statistics requires more complex, multidimensional analyses or dimensional reduction methods
  - Hard to use, hard to understand
- Omics data analyses borrow a lot from several other fields
  - Pattern recognition / machine learning
  - Dimensionality reduction

Key Areas in Data Science



Machine Learning

- **Unsupervised learning:** explore the data to find some intrinsic structures in them Disregard whether they are related to the class labels or not)
  - These patterns can be used to understand the key information within the data itself
- **Supervised learning:** discover features / patterns in the data that relate data attributes with related to a target (class) attribute.
  - These patterns can be utilized to predict the values of the target attribute in future data instances
- **Reinforcement learning:** task-oriented to maximize reward; used in game theory, auto-drive

Unsupervised Learning Methods for high-dimensional Data

- Clustering
  - Organize the 1000s of variables into blocks
  - Variables in each block are more homogenous, and treat these blocks as a unit
- Dimension reduction
  - Reduce the high-dimensional data into low-dimension i.e. from 1000s of variables to 2 ~ 3 variables
    - Principal component analysis (PCA)

Clustering

Definition - a process by which objects that are logically similar in characteristics are grouped together

- A method to measure similarity (a similarity matrix) or dissimilarity (a dissimilarity coefficient) between objects
- A threshold value with which to decide whether an object belongs with a cluster
- A way of measuring the “distance” between two clusters

Two Common Clustering Algorithms

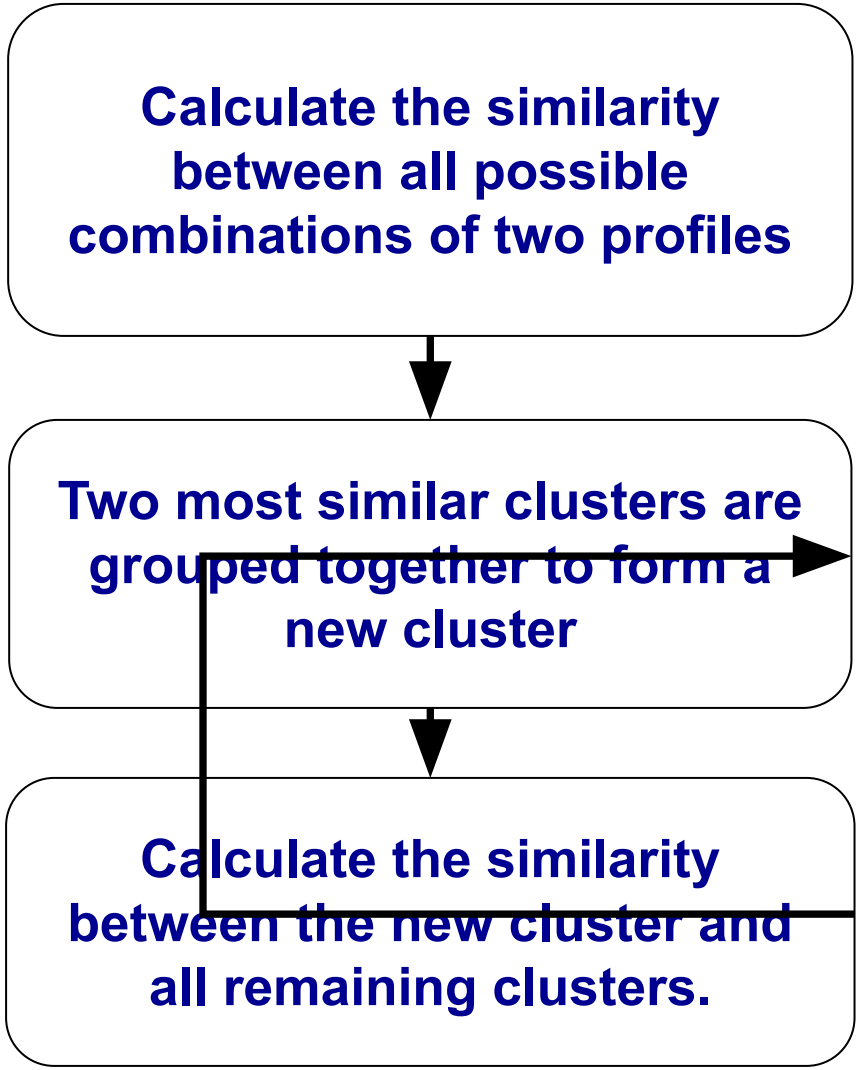
- Hierarchical Methods - produces a set of nested clusters in which each pair of objects is progressively nested into a larger cluster until only one cluster remains
- K-means or Partitioning Methods - divides a set of N objects into M clusters -- with or without overlap

Hierarchical Clustering

- Find the two closest objects and merge them into a cluster
- Find and merge the next two closest objects (or an object and a cluster, or two clusters) using some similarity measure and a predefined threshold
- If more than one cluster remains return to step 2 until finished

Key parameters

- Similarity between samples
- Similarity between clusters



Similarity Measurements

- Euclidean Distance: Absolute difference

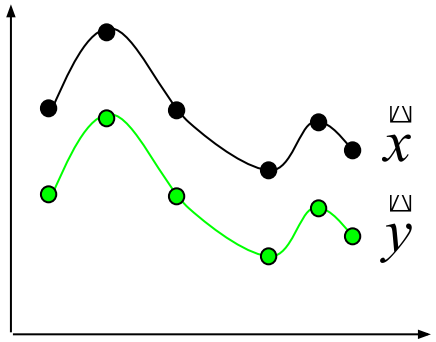
$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{n=1}^{N} (x_n - y_n)^2}$$

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

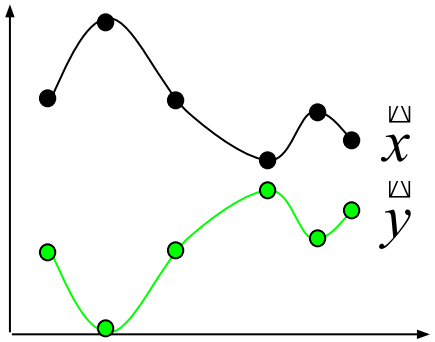
Similarity Measurements

Pearson Correlation: Trend similarity

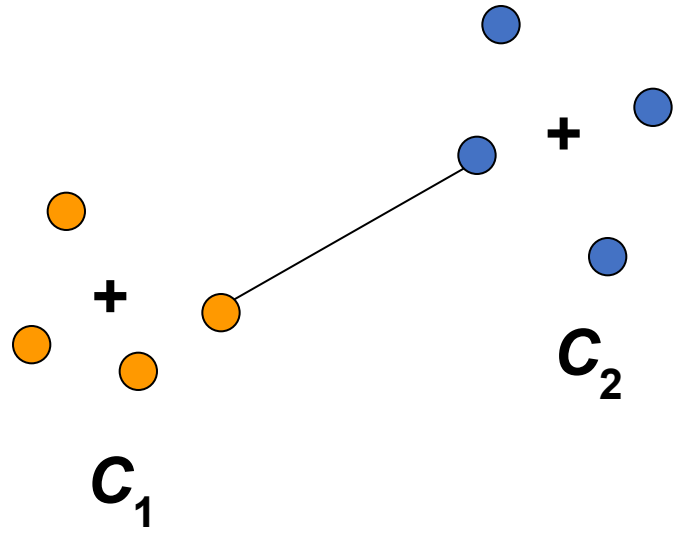
$$C_{pearson}(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{N}(x_i - m_x)(y_i - m_y)}{\sqrt{[\sum_{i=1}^{N}(x_i - m_x)^2][\sum_{i=1}^{N}(y_i - m_y)^2]}}$$



+1 ≥ Pearson Correlation ≥ – 1



Clustering



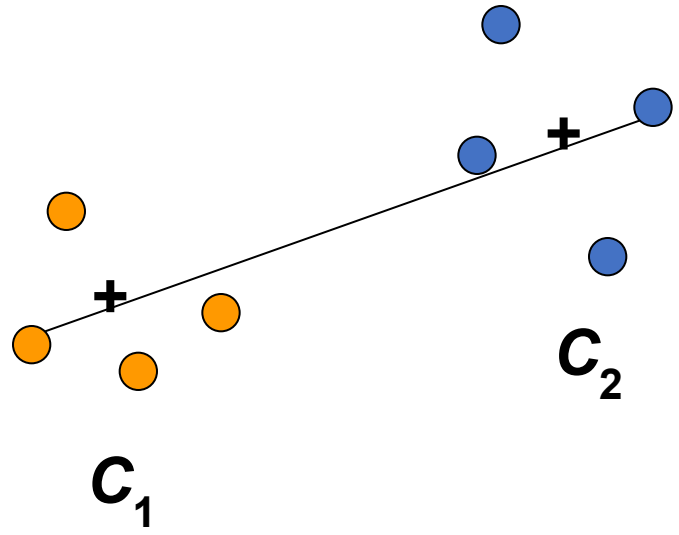
Single Linkage

Dissimilarity between two clusters = Minimum dissimilarity between the members of two clusters

Tend to generate “long chains”

Clustering

Complete Linkage

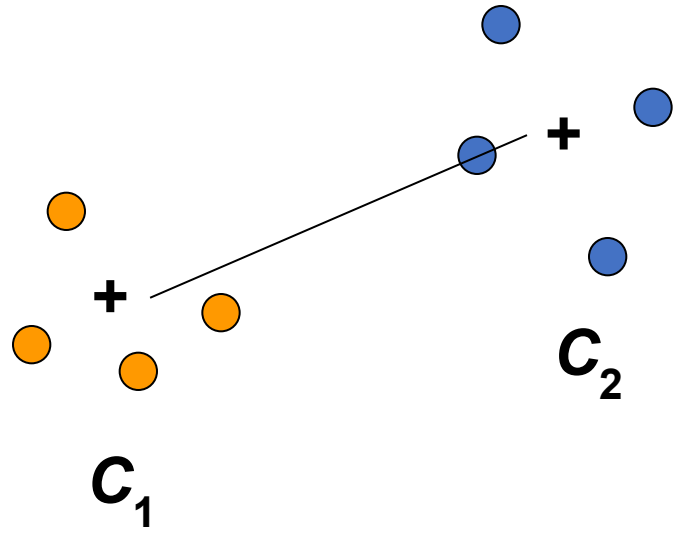


Dissimilarity between two clusters = Maximum dissimilarity between the members of two clusters

Tend to generate “clumps”

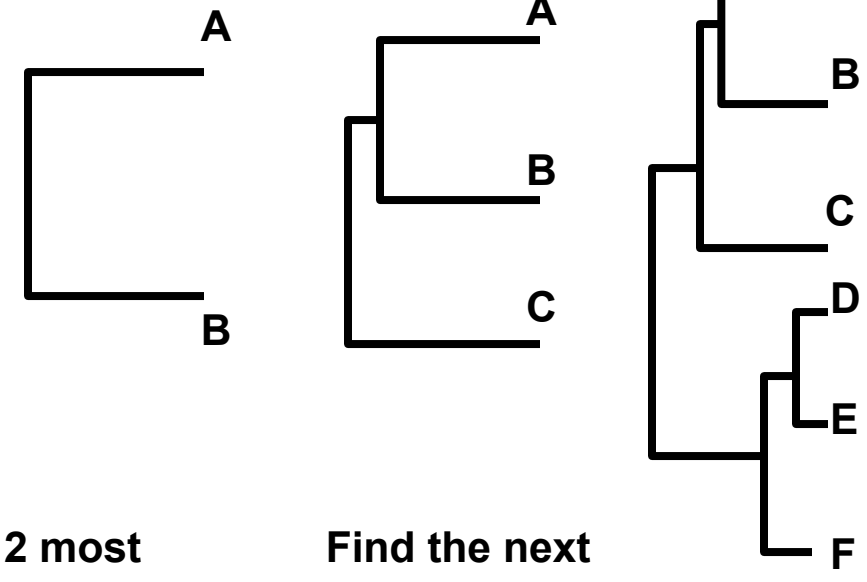
Clustering

Average Group Linkage



Dissimilarity between two clusters = Distance between two cluster means.

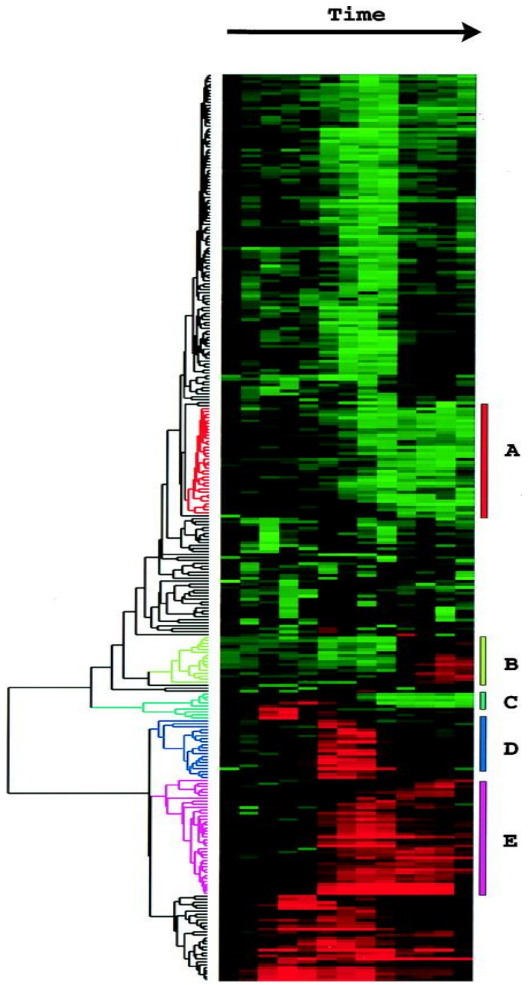
Hierarchical Clustering & Heatmaps



Find 2 most similar metabolite expression levels or curves

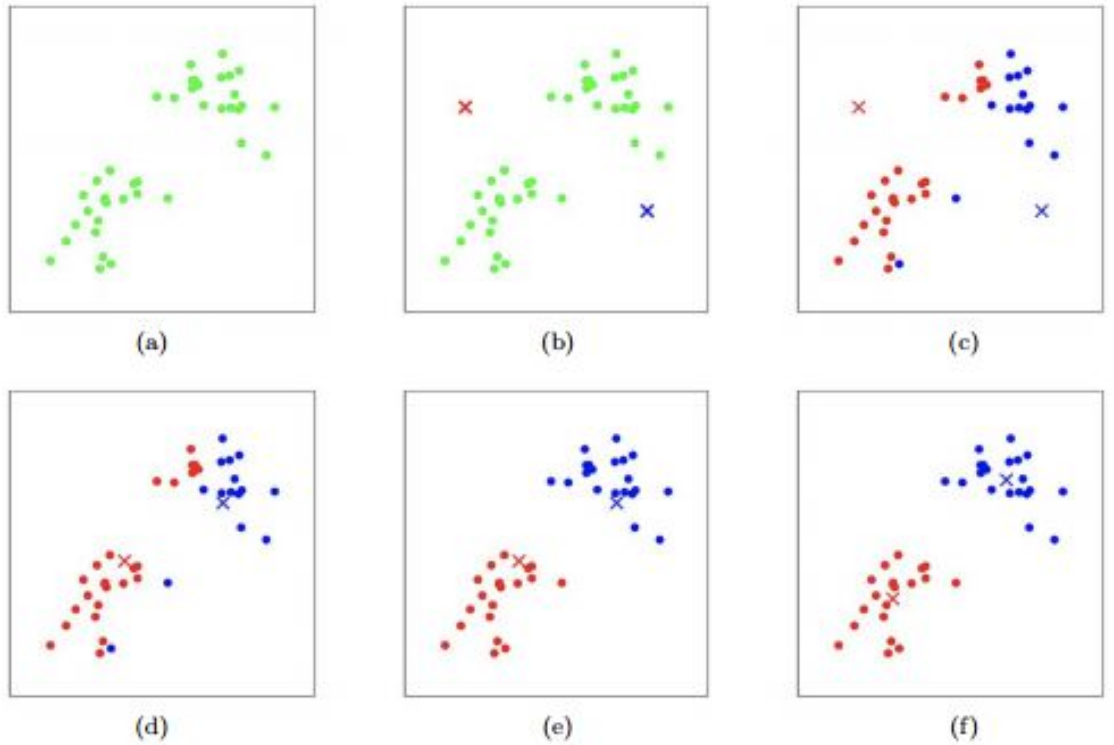
Find the next closest pair of levels or curves

Iterate



K-means clustering

1. Randomly chooses *k* observations from the dataset and uses these as the initial means
2. For the next object calculate the similarity to each existing centroid
3. If the similarity is greater than a threshold add the object to the existing cluster and re-compute the centroid, else use the object to start new cluster
4. Return to step 2 and repeat until done



K = 2

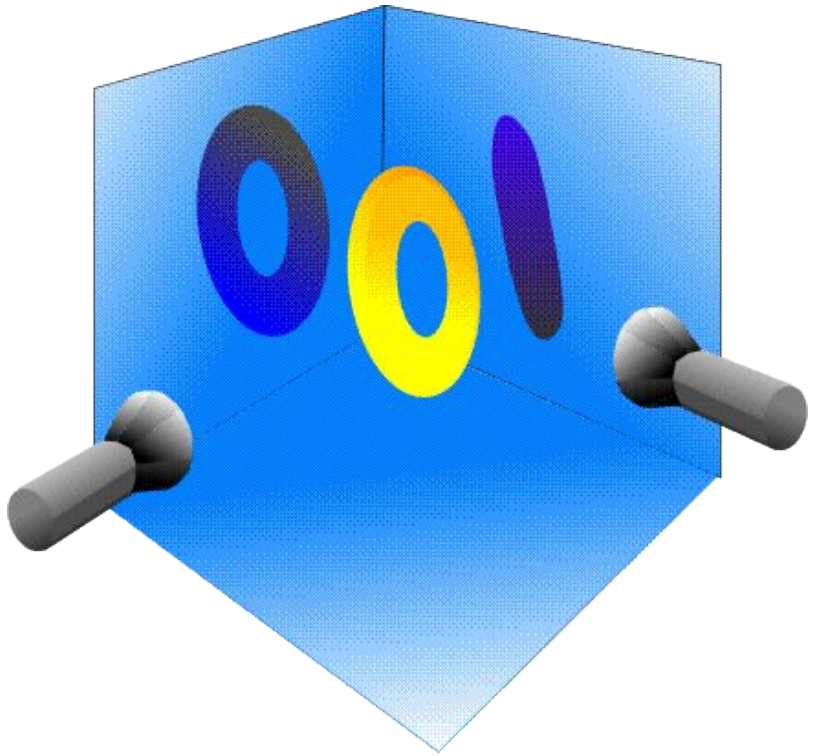
Principal Component Analysis (PCA)

- Project high-dimensional data into lower dimensions that capture the most variance of the data
- Assumption:

  Main directions of variance

  ≈ major data characteristics

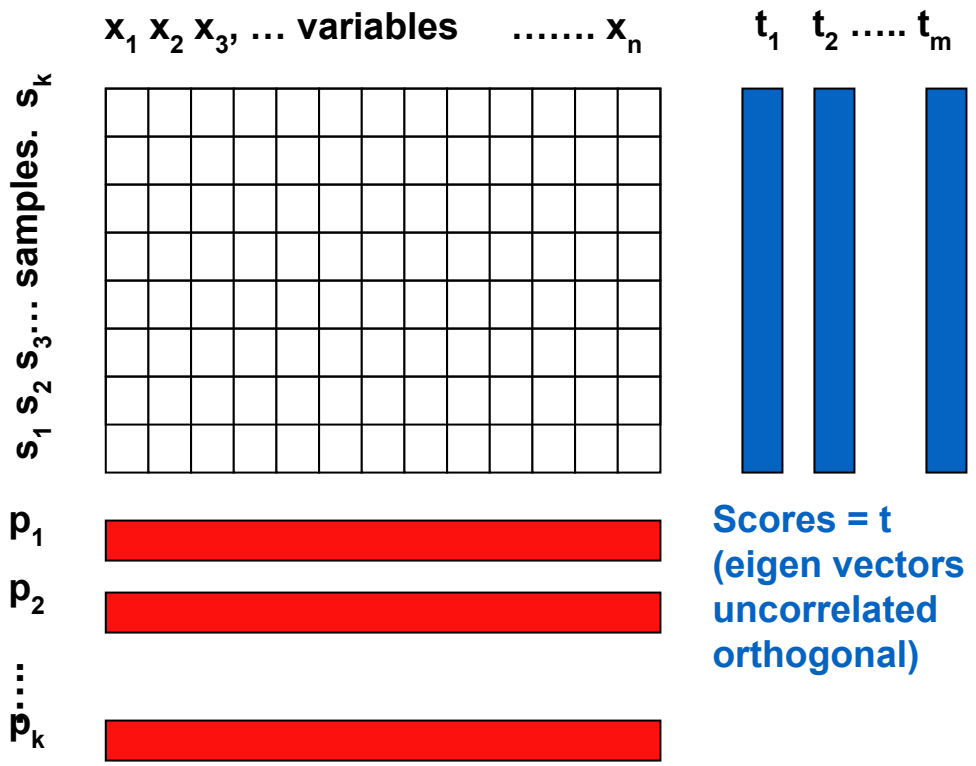
Visualizing PCA



- PCA of a “bagel”
- One projection produces a weiner (hotdog)
- Another projection produces an “O”
- The “O” projection captures most of the variation and has the largest eigenvector (PC1)
- The weiner projection is PC2 and gives depth info

PCA - The Details

- PCA involves the calculation of the eigenvalue (singular value) decomposition of a data covariance matrix
- PCA is an orthogonal linear transformation
- PCA transforms data to a new coordinate system so that the greatest variance of the data comes to lie on the first coordinate (1st PC), the second greatest variance on the 2nd PC etc.



Scores = Loadings x data

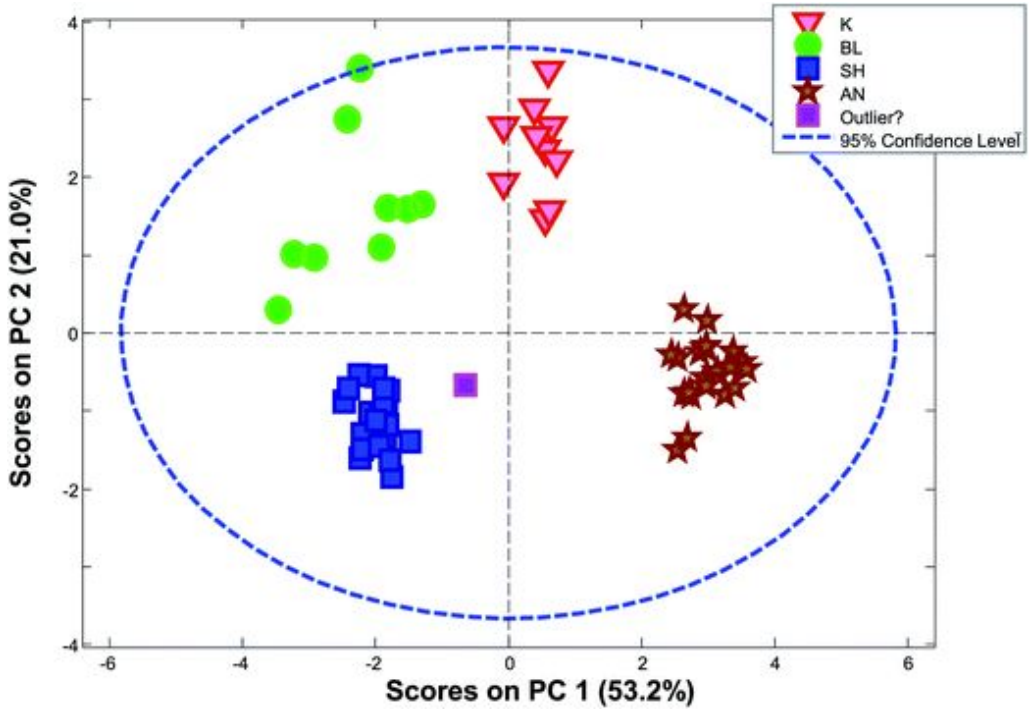
$$t_1 = p_1x_1 + p_2x_2 + p_3x_3 + \ldots + p_nx_n$$

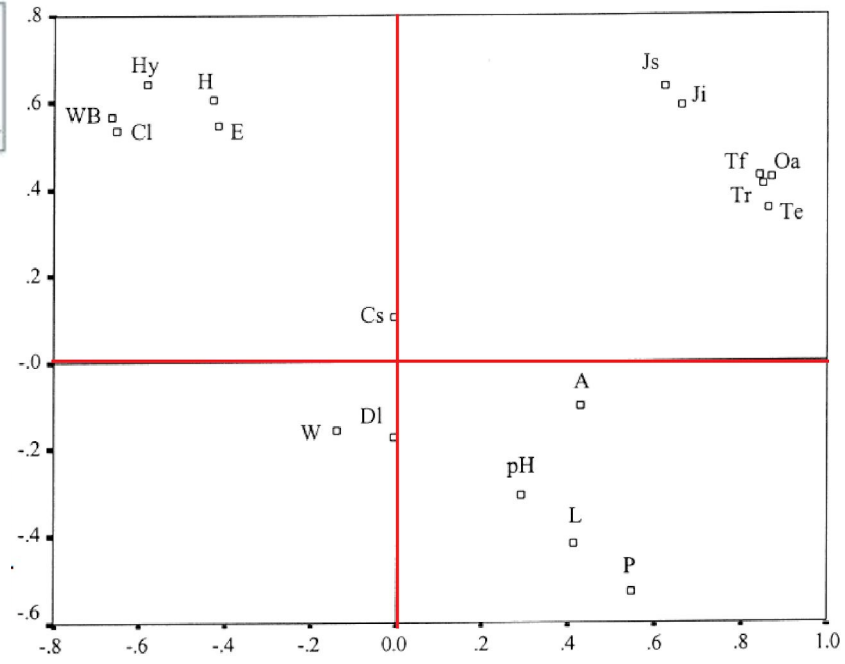
Principal Components Analysis On:

- Covariance matrix:
  - Variables must be in same units
  - Emphasizes variables with most variance

- Correlation matrix:
  - Variables are standardized (mean 0.0, SD 1.0)
  - Variables can be in different units
  - All variables have same impact on analysis

Scores & Loadings





Scores plot shows the overall patterns of similarities among samples. Samples close to each other are more similar

Loadings plot shows how much each of the variables contributed to the different principal components

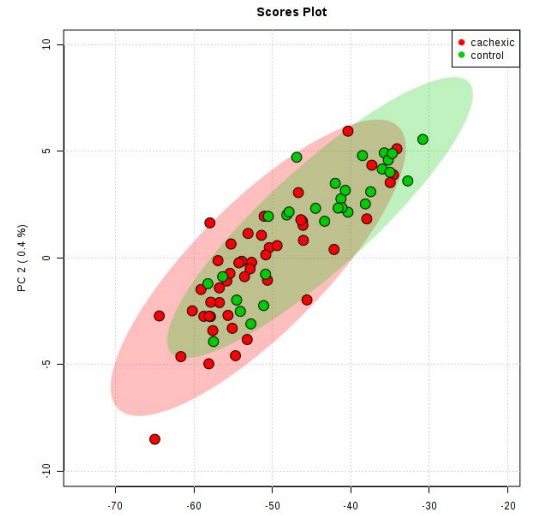
PCA Summary

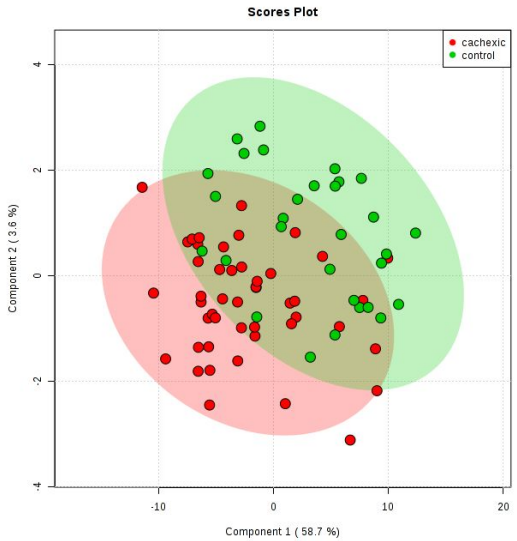
- Rotates multivariate dataset into a new configuration which is easier to interpret
- Purposes
  - Data overview
  - Outlier detection
  - Look at relationships between samples or variables

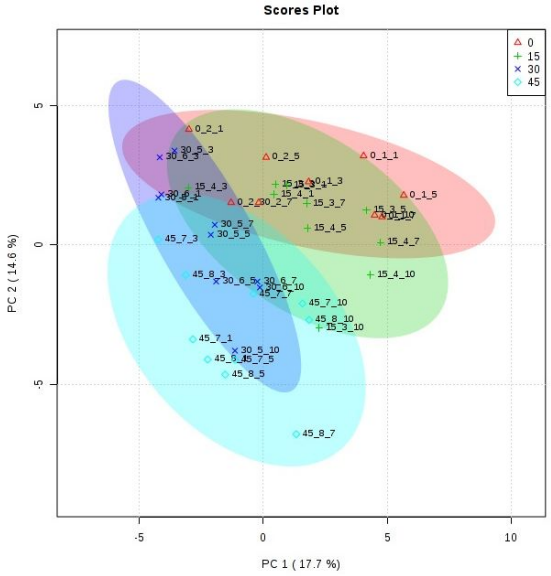
PLS-DA

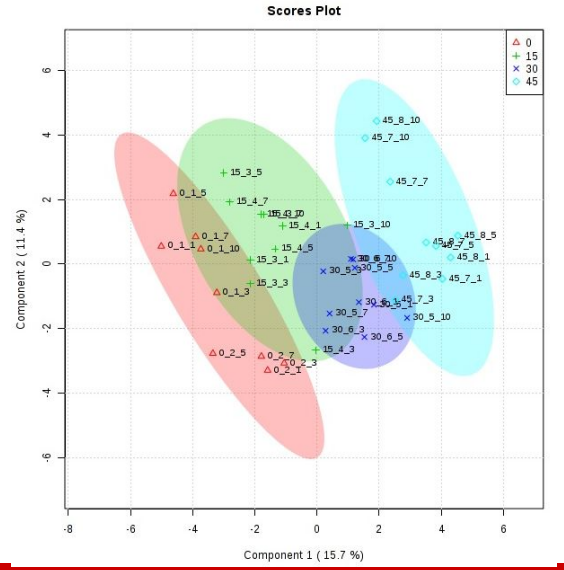
- When the experimental effects are subtle or moderate, PCA will not show good separation patterns
- PLS-DA is a supervised method, it is calculated by maximizing the co-variance between the data matrix (X) and the class labels (Y)
- PLS-DA always produces certain separation patterns with regard the conditions

PCA vs. PLS-DA

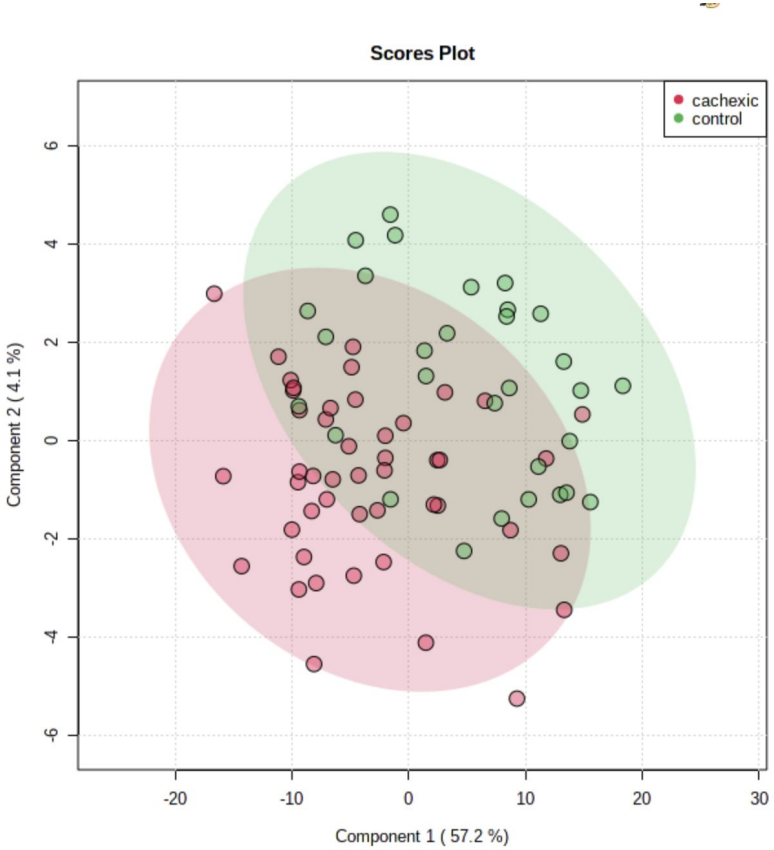

Scores Plot
cachexic
control
PC 2 (0.4 %)
-70
-60
-50
-40
-30
-20
10
5
0
-5
-10

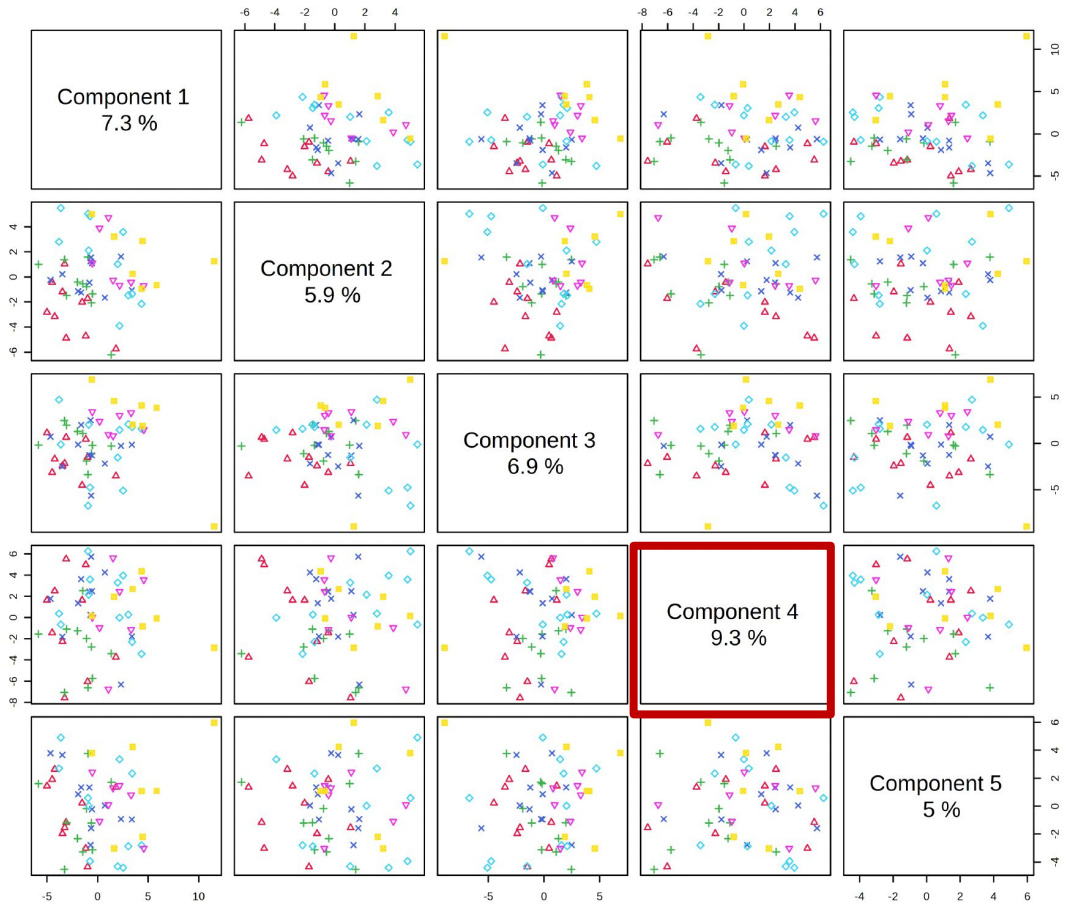

Scores Plot
cachexic
control
Component 2 (3.6 %)
Component 1 (58.7 %)
-10
0
10
20
4
2
0
-2
-4


Scores Plot
0
15
30
45
PC 2 (14.6 %)
PC 1 (17.7 %)
-5
0
5
10


Scores Plot
0
15
30
45
Component 2 (11.4 %)
Component 1 (15.7 %)
-8
-6
-4
-2
0
2
4
6

PLS-DA maximize covariance (not variance in X)





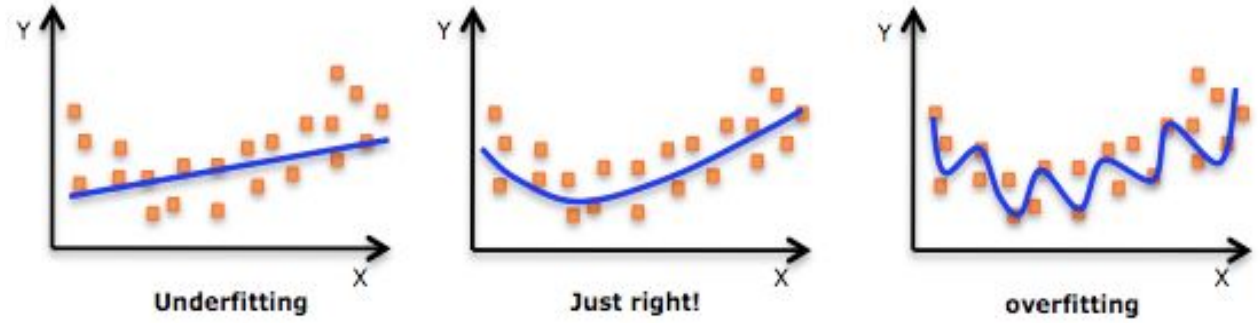
Note, PLS-DA maximizes the **covariance** between X (data) and Y (group). The variance displayed in the plot above is the **explained variance for X**. Covariance and x-varian may not agree with each other in some cases. For instance, the 1st component may not explain more X-variance than the 2nd component.

Use PLS-DA with Caution

- PLS-DA is susceptible to over-fitting by producing patterns of separation even for data randomly drawn from the same population
  - Need cross validation
  - Need permutation tests

Overfitting

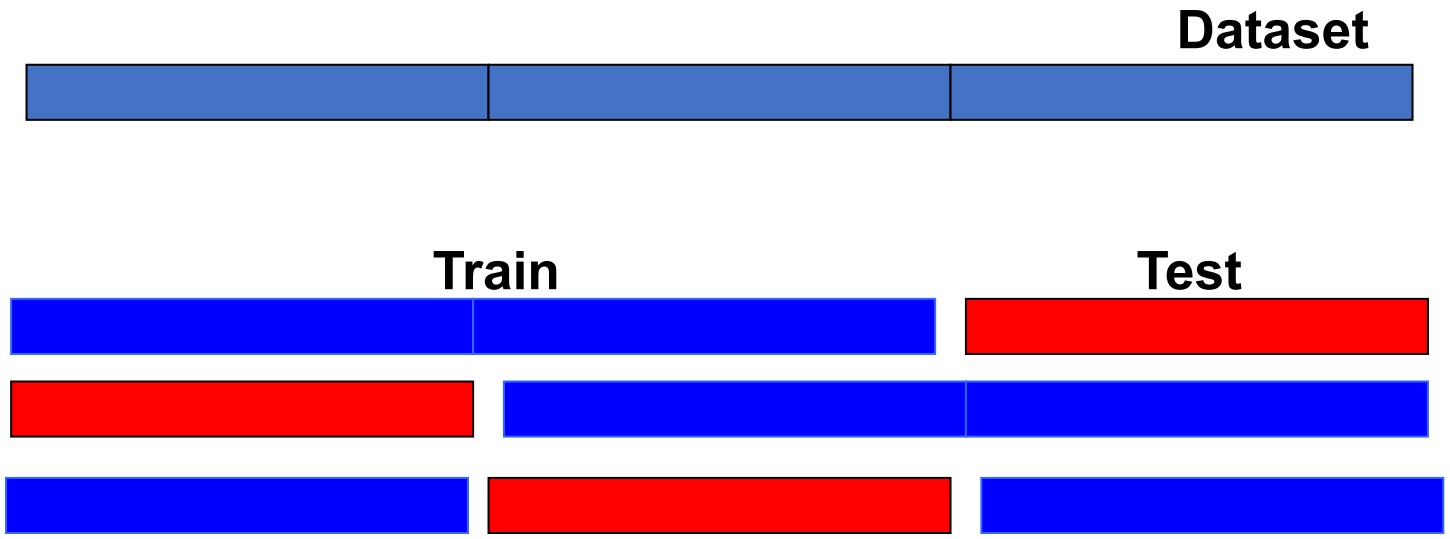
- If we put too many variables in the model, including some unrelated to the response, we are *overfitting*.
- Consequences are:
  - Fitted model is not good for prediction of new data – prediction error is underestimated
  - Model is too elaborate, models “noise” that will not be the same for new data



Cross Validations

Goal: test whether your model can predict class labels for new samples

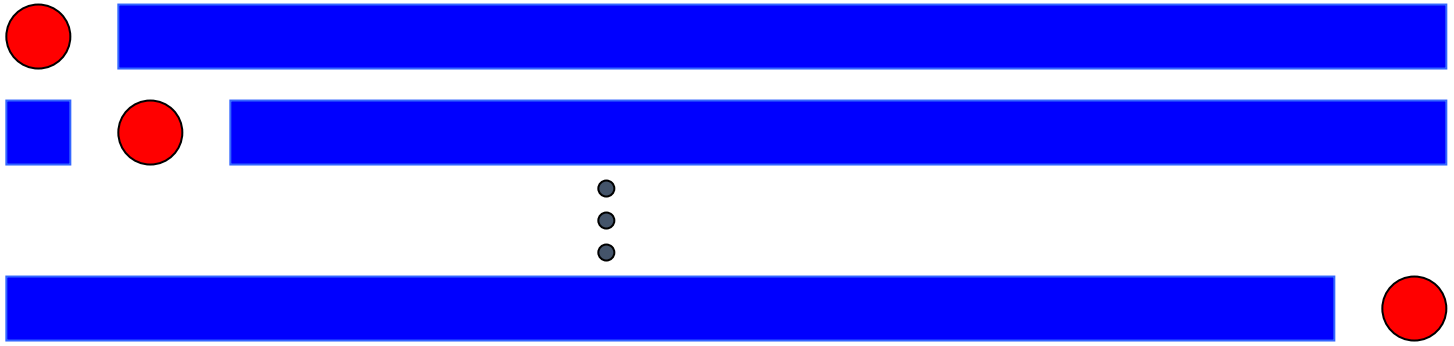
-- results may vary between splitting



Common Splitting Strategies

Leave-one-out (n-fold cross validation)

--- more stable



Components and Features

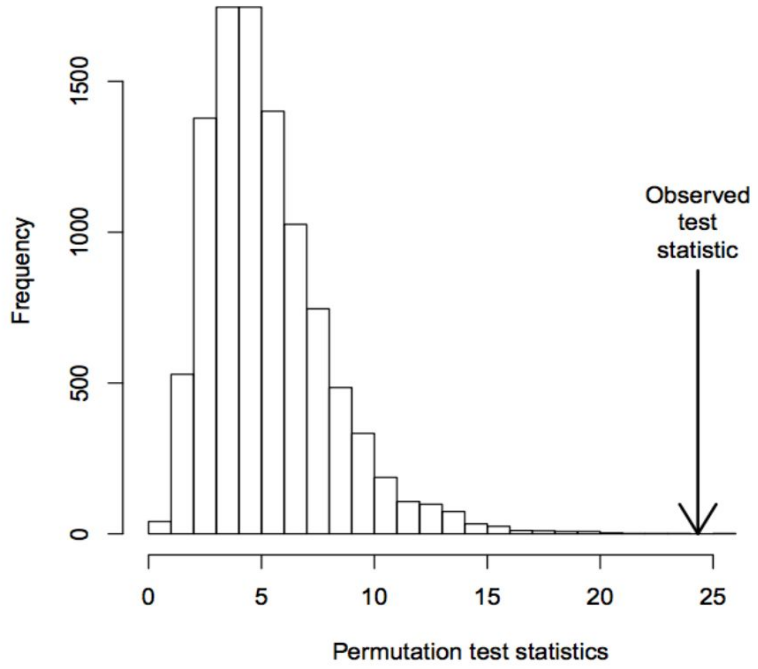
Cross validation is used to determine the optimal number of components needed to build the PLS-DA model. Three common performance measures:

- Sum of squares captured by the model (R^2)
- Cross-validated R^2 (also known as Q^2)
- Prediction accuracy

Permutation Tests

- Goal: to test whether your model is significantly different from the null models
  1. Randomly shuffle the class labels (y) and build the (null) model between new y and x;
  2. Test whether there is still the similar distances of separation;
  3. We can compute empirical p values
     - If the result is similar to the permuted results (i.e. null model), then we can NOT say y and x are significantly correlated

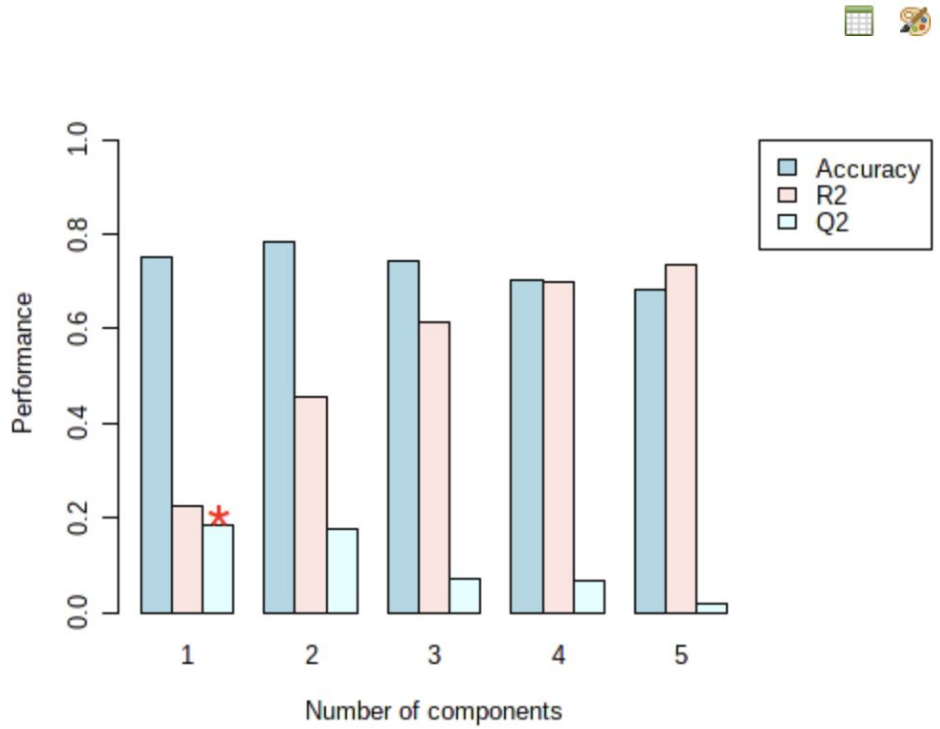
Compute Empirical P-values



In 1000 permutations, if none of the permuted models has a mean difference is bigger than the original one then:

- p < 0.001 or (1/1001) # prevent p value equal to zero

PLS-DA (R2 & Q2)

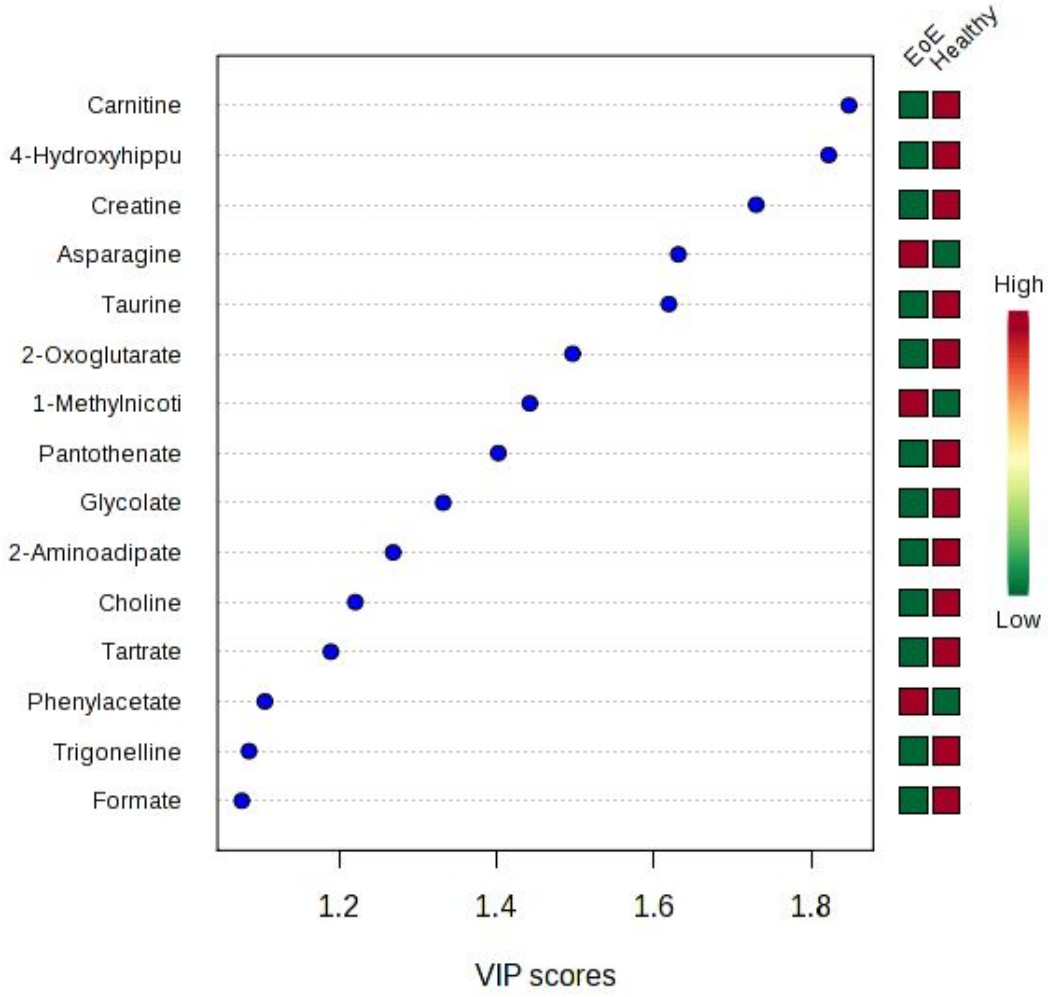


Q2 is an estimate of the predictive ability of the model, and is calculated via cross-validation (CV). In each CV, the predicted data are compared with the original data, and the sum of squared errors is calculated. The prediction error is then summed over all samples (Predicted Residual Sum of Squares or PRESS). For convenience, the PRESS is divided by the initial sum of squares and subtracted from 1 to resemble the scale of the R2. Good predictions will have low PRESS or high Q2. It is possible to have **negative Q2**, which means that your model is not at all predictive or is overfitted. For more details, refer to an excellent paper by (Szymańska, et al).

PLS-DA VIP Score

Variable Importance in Projection (VIP) scores estimate the importance of each variable in a PLS-DA model

- ✔ Weighted sum of the squared correlations between the PLS-DA components and the original variable
- ✔ Weights correspond to the percentage variation explained by the PLS-DA component in the model
- ✔ A common threshold: VIP > 1



Performance Measure & ROC curve analysis

Assessing Classification Model Performance

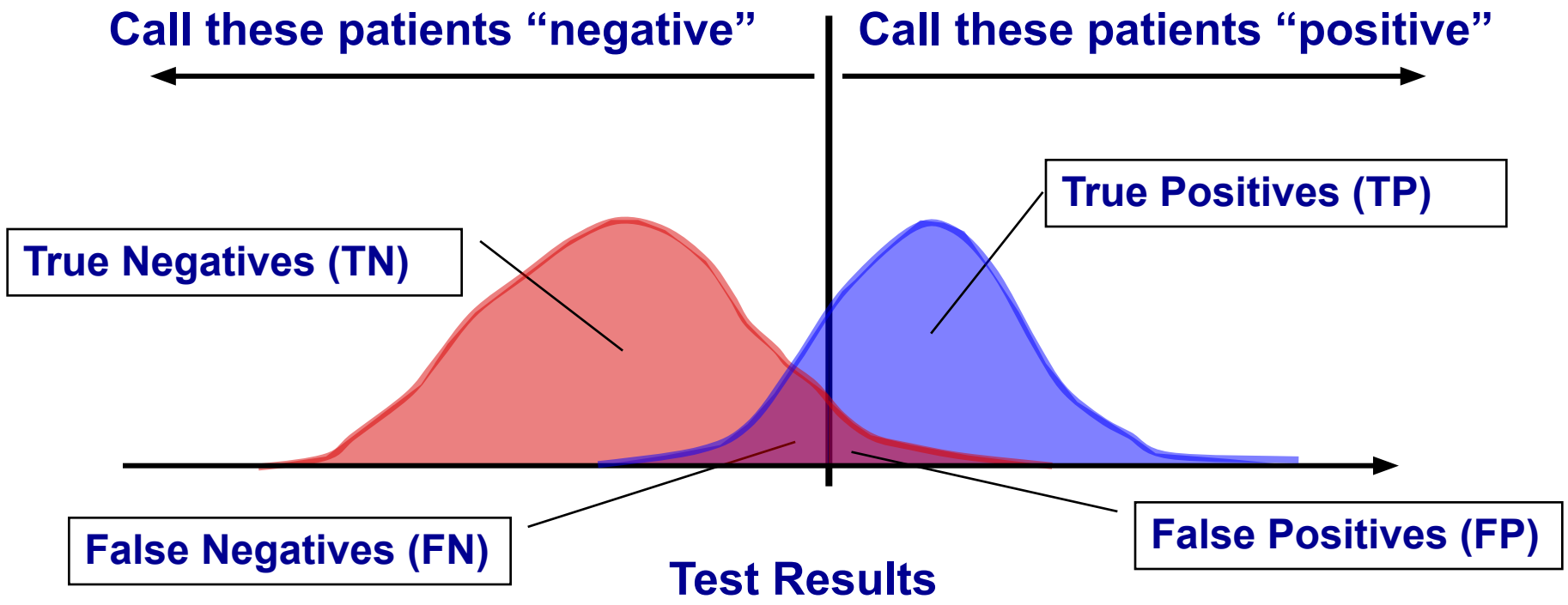
- For balanced data
  - Accuracy: 9/13 correct => 69% accuracy
  - Error rate: 1 – accuracy => 31%
- Not suitable for imbalanced data
  - In a population, cancer incidence is low: ~5 cases in 1000 people. If a classifier predict all people to be healthy, then it is 99.5% accurate (majority vote)

Evaluating Performance

- Basic concepts
  - True positives (TP)
  - True negatives (TN)
  - False positives (FP)
  - False negatives (FN).
  - Sensitivity (Sn)
  - Specificity (Sp)
- Sn (sensitivity) = True positive rate
- Sp (specificity) = True negative rate

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

An Example

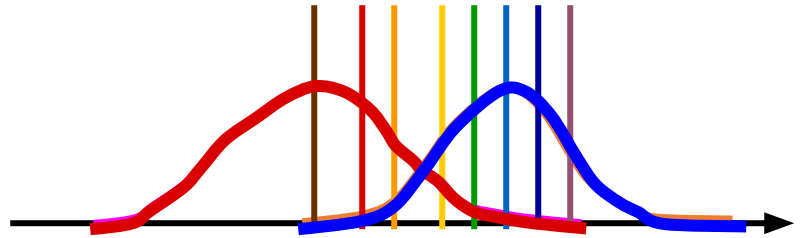


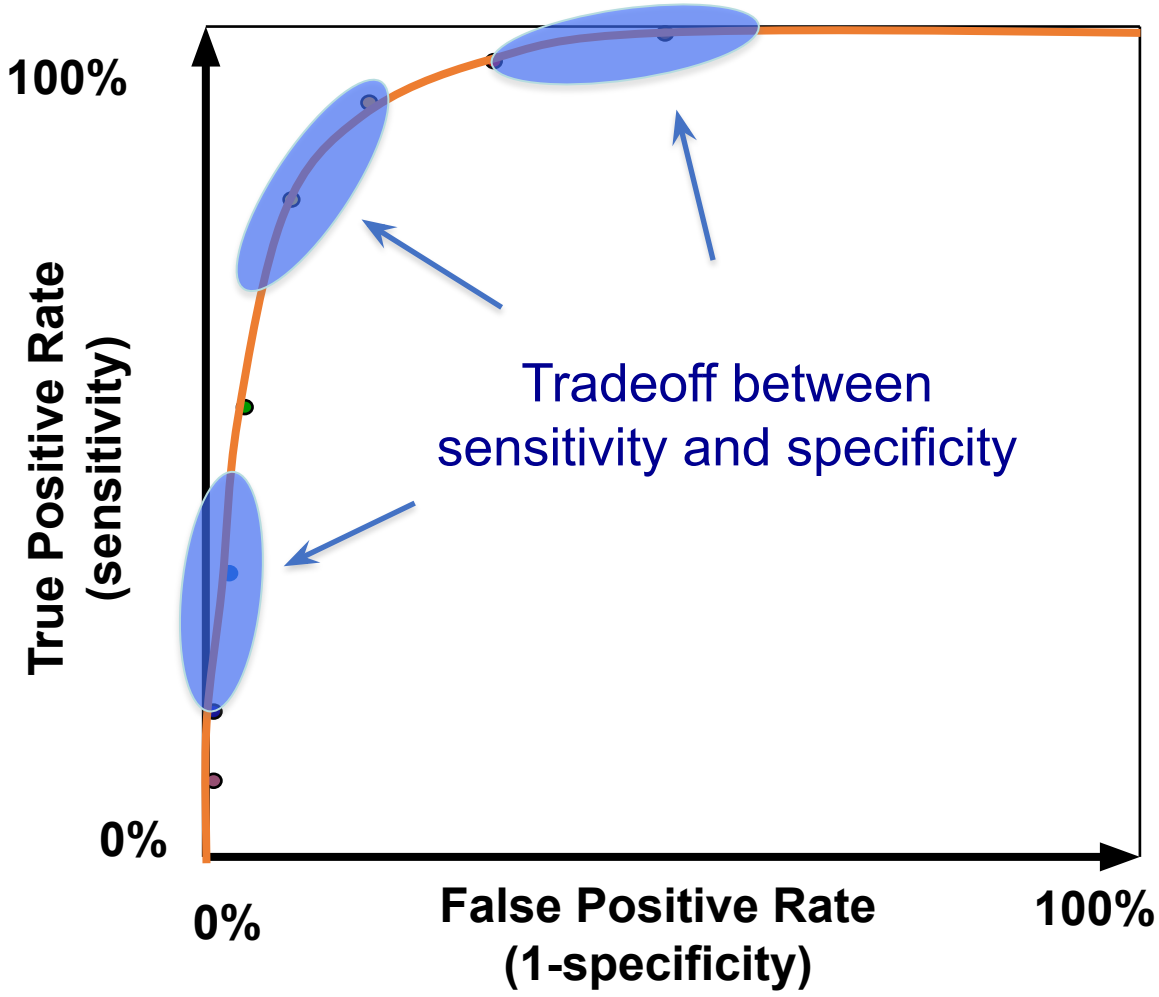
Sn = TP/(TP + FN)

Sp = TN/(TN + FP)

ROC Curves

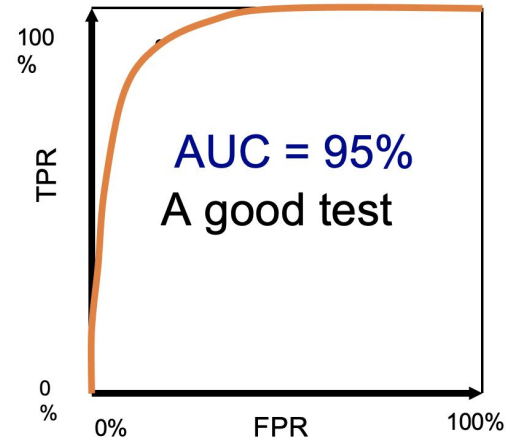
- ROC = Receiver Operating Characteristic
  - A historic name from radar studies
  - Very popular in biomedical applications
    - To assess performance of classifiers.
    - To compare different biomarker models
- A graphical plot of the true positive rate (TPR) vs. false positive rate (FPR), for a binary classifier (i.e. positive/negative) as its cutoff point is varied

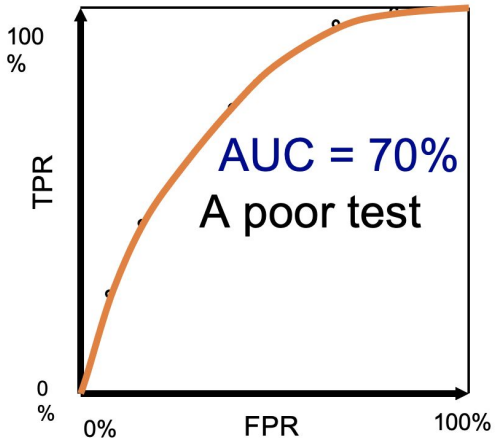


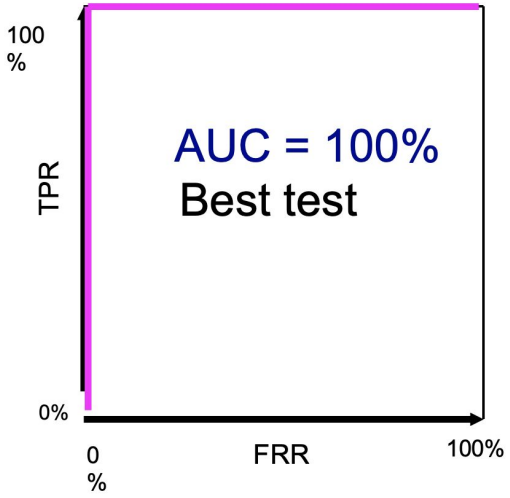


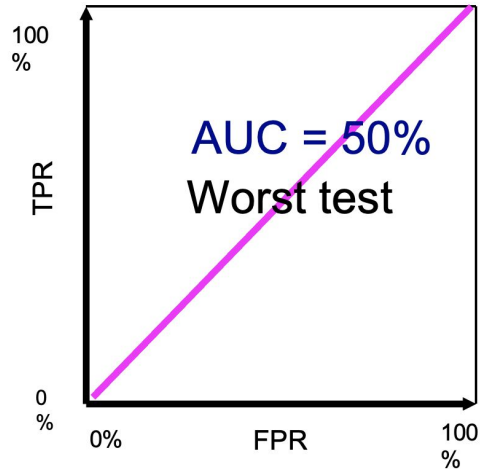
Area Under ROC Curve (AUC)

- Overall measure of test performance
- Comparisons between two tests based on differences between (estimated) AUC









Other Supervised Classification Methods

- SIMCA – Soft Independent Modeling of Class Analogy
- *OPLS – Orthogonal Projection of Latent Structures
- *Support Vector Machines
- *Random Forest
- *Logistic regression (Biomarker Analysis module)

* Implemented in MetaboAnalyst

Data Analysis Progression

- Unsupervised Methods
  - PCA or cluster to see if natural clusters form or if data separates well
  - Data is “unlabeled” (no prior knowledge)
- Supervised Methods
  - Data is labeled (prior knowledge)
  - Used to see if data can be classified
  - Helps separate less obvious clusters or features
  - Supervised methods always generate clusters/patterns
    - This can be very misleading
    - Cross validation and permutation are important
  - Need a large number of samples
    - At least 20~30 replicates for meaningful permutations or classifications
    - Small sample size
      - Univariate (t-tests/ ANOVA), PCA

We are on a Coffee Break & Networking Session

Workshop Sponsors:

Canadian Centre for Computational Genomics