Multivariate analysis of 'omics data

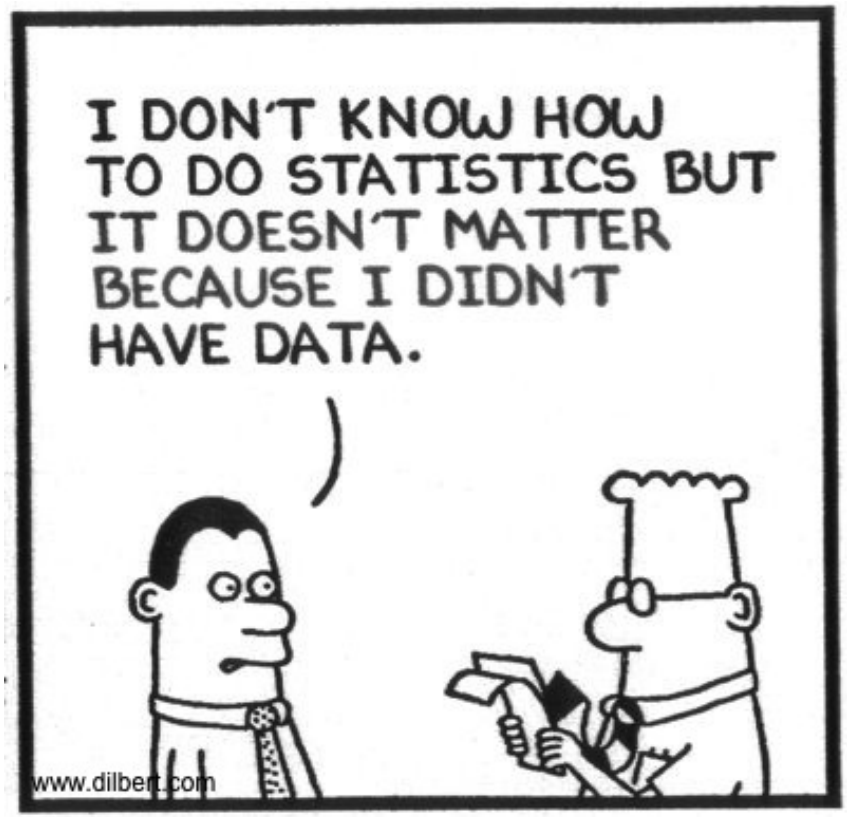
Background: Data Exploration with Principal Component Analysis

A/Prof. Kim-Anh Lê Cao

Melbourne Integrative Genomics
School of Mathematics & Statistics
University of Melbourne

kimanh.lecao@unimelb.edu.au

Learning objectives of this course



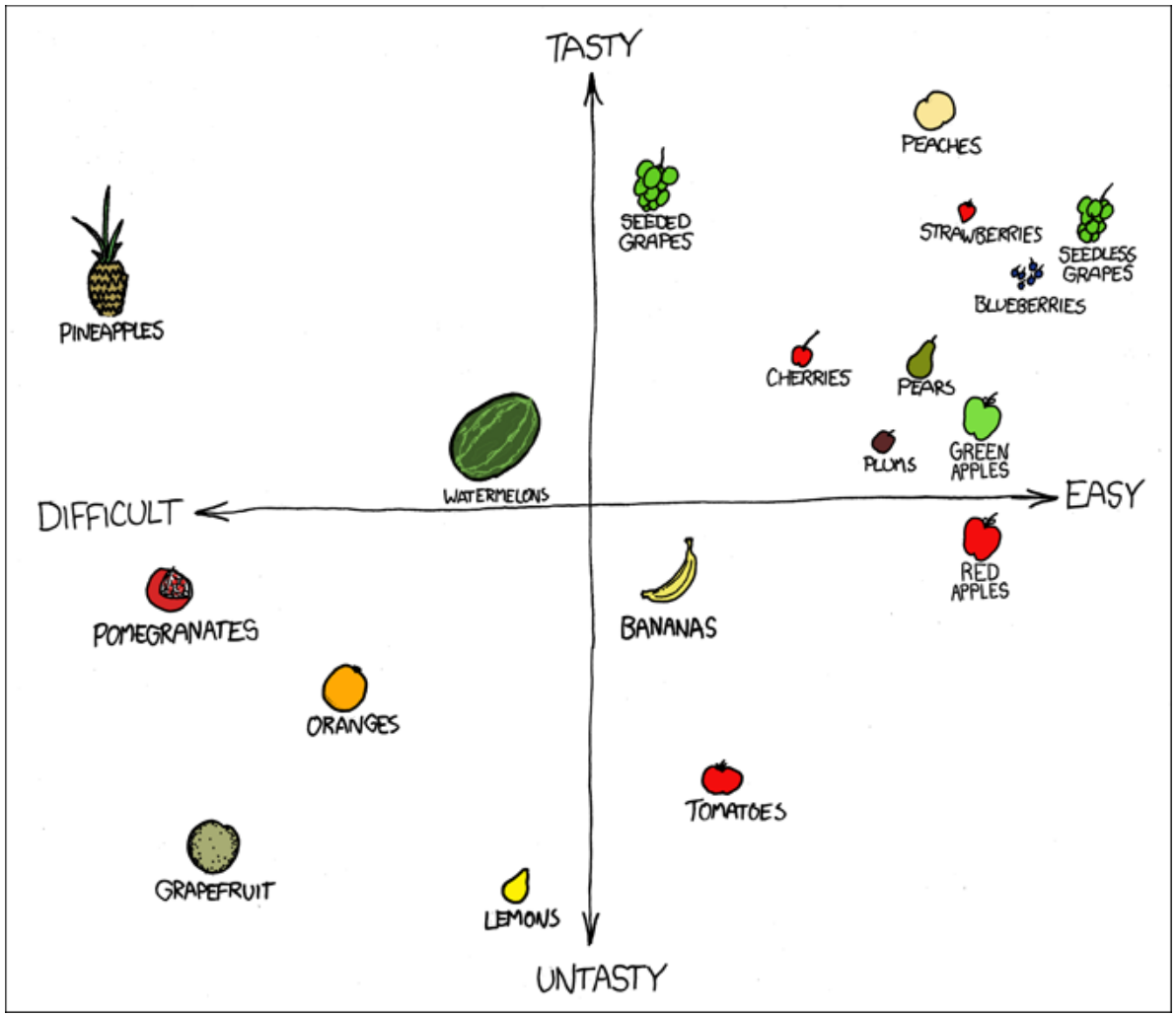
Learning objectives of this course

Theory

- Understand the main concepts of multivariate dimension reduction methods
- Choose the 'right' method for the 'right' biological question
- Be aware of the benefits and limitations of all methods presented
- Interpretation of the visualisation outputs

Practice

- Ability to use provided `R` code on own data
- Perform several types of multivariate analyses ranging from data exploration to biomarker selection using `mixOmics`
- Be critical of the results obtained

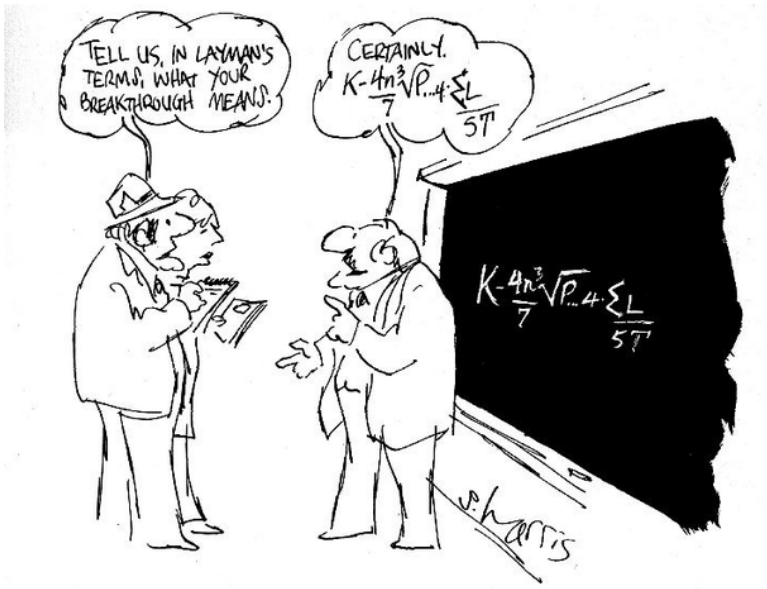


Component-based multivariate methods reduce data dimension whilst extracting the most 'relevant' information from the data.

Table of Contents

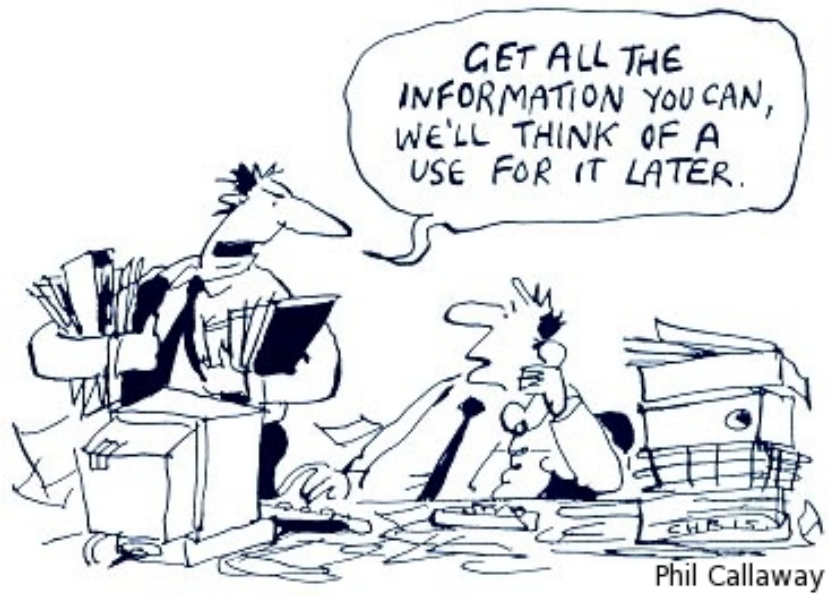
Context: when biology and statistics meet

"Data don't make any sense,
we will have to resort to statistics."



Context: when biology and statistics meet

A close interaction between statisticians, bioinformaticians and molecular biologists is essential to provide meaningful results



- Unlimited quantity of data from multiple and heterogeneous sources
- Computational issues to foresee
- Biological interpretation for validation
- Keep pace with new technologies

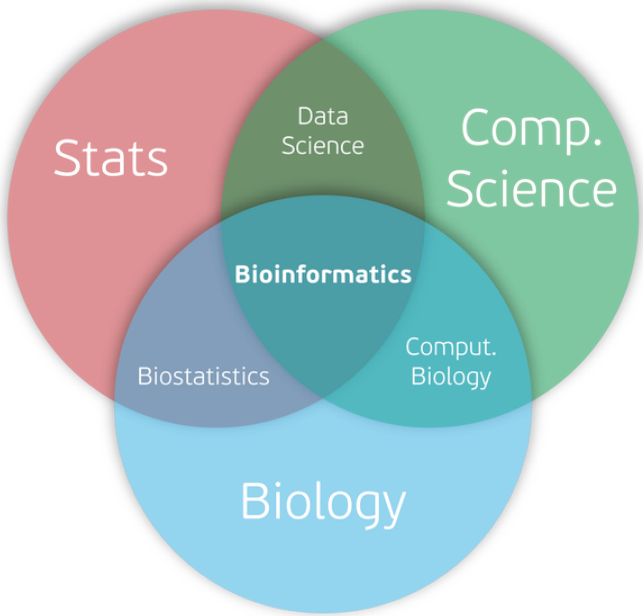
Bioinformatics and Statistics: same difference?

Bioinformatics ranges across:

- computer science applied to biological studies
- data management and databases design, dedicated hardware
- processing pipelines & data analysis

Statistics:

- (can be) part of bioinformatics
- inference based on population distribution
- methods development, data analysis



(Bio)informatics operates on data; Statistics infers from data

A holistic view of a biological system

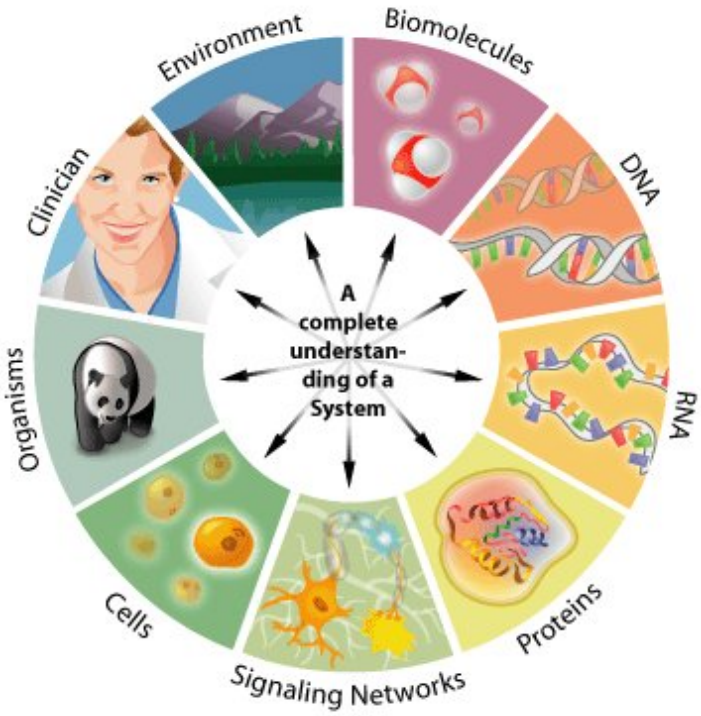
From reductionism ...

1 gene = 1 hypothesis = 1 statistical test

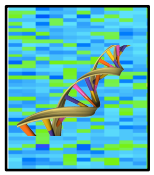
$\Downarrow$

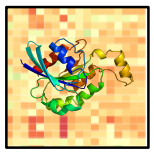
... to holism:

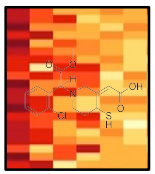
Thousands of molecules = **??**

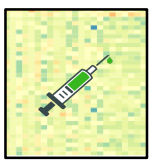


A holistic view of a biological system









Transcriptomics
$\sim 100,000$ transcripts

$\downarrow$

Proteomics
$\sim 1,000,000$ proteins

$\downarrow$

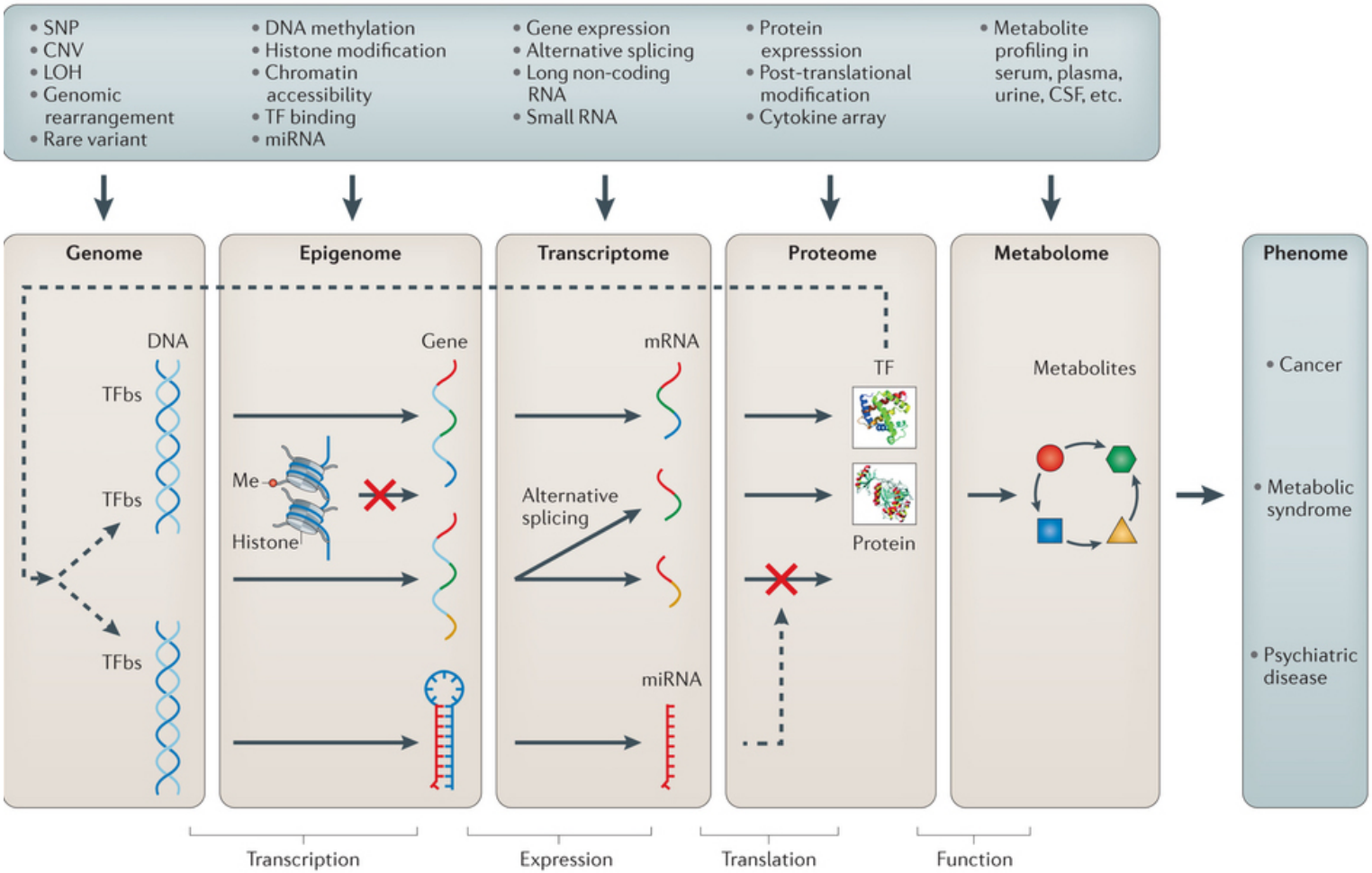
Metabolomics
$\sim 2,500$ compounds

Microbiome
$\sim 1,500$ trillion microbes

$\vdots$

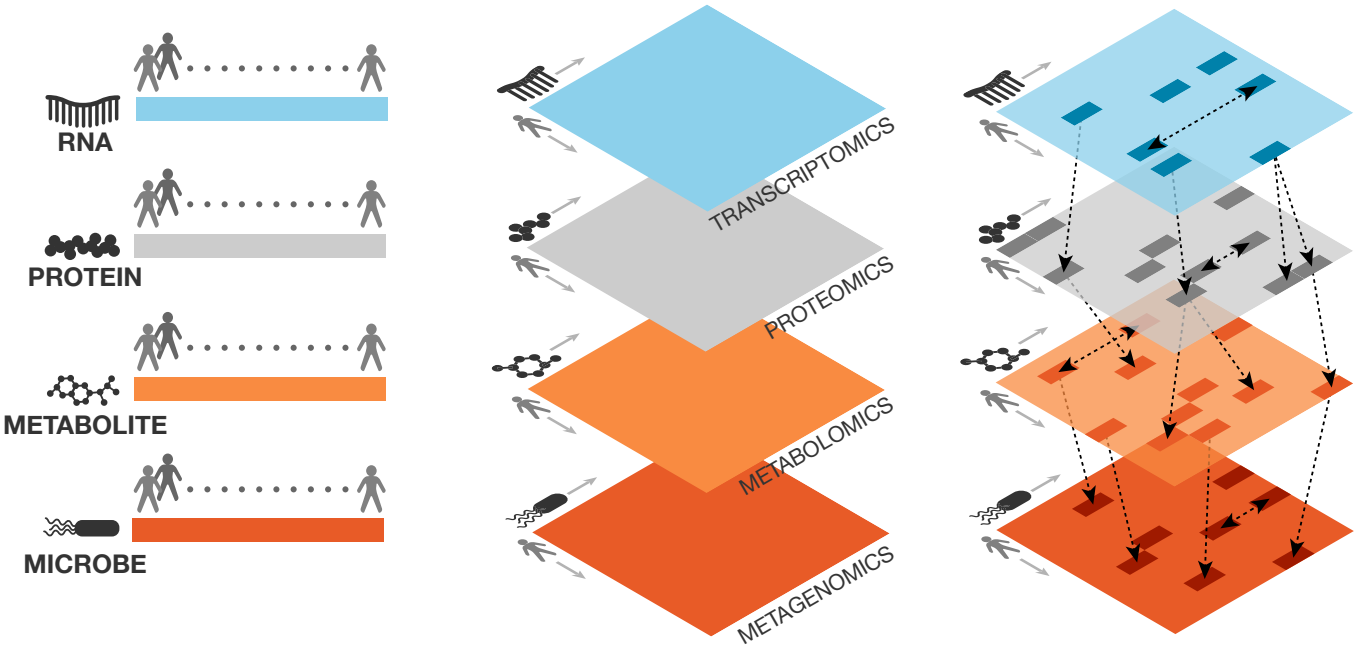
Clinical data
many

~~Biological dogma~~ with omics data: not that straighforward



Ritchie et al., 2015, *Nature Genetics* **16**

Why integrating 'omics data?



Conventional molecular biology
(Reductionist)

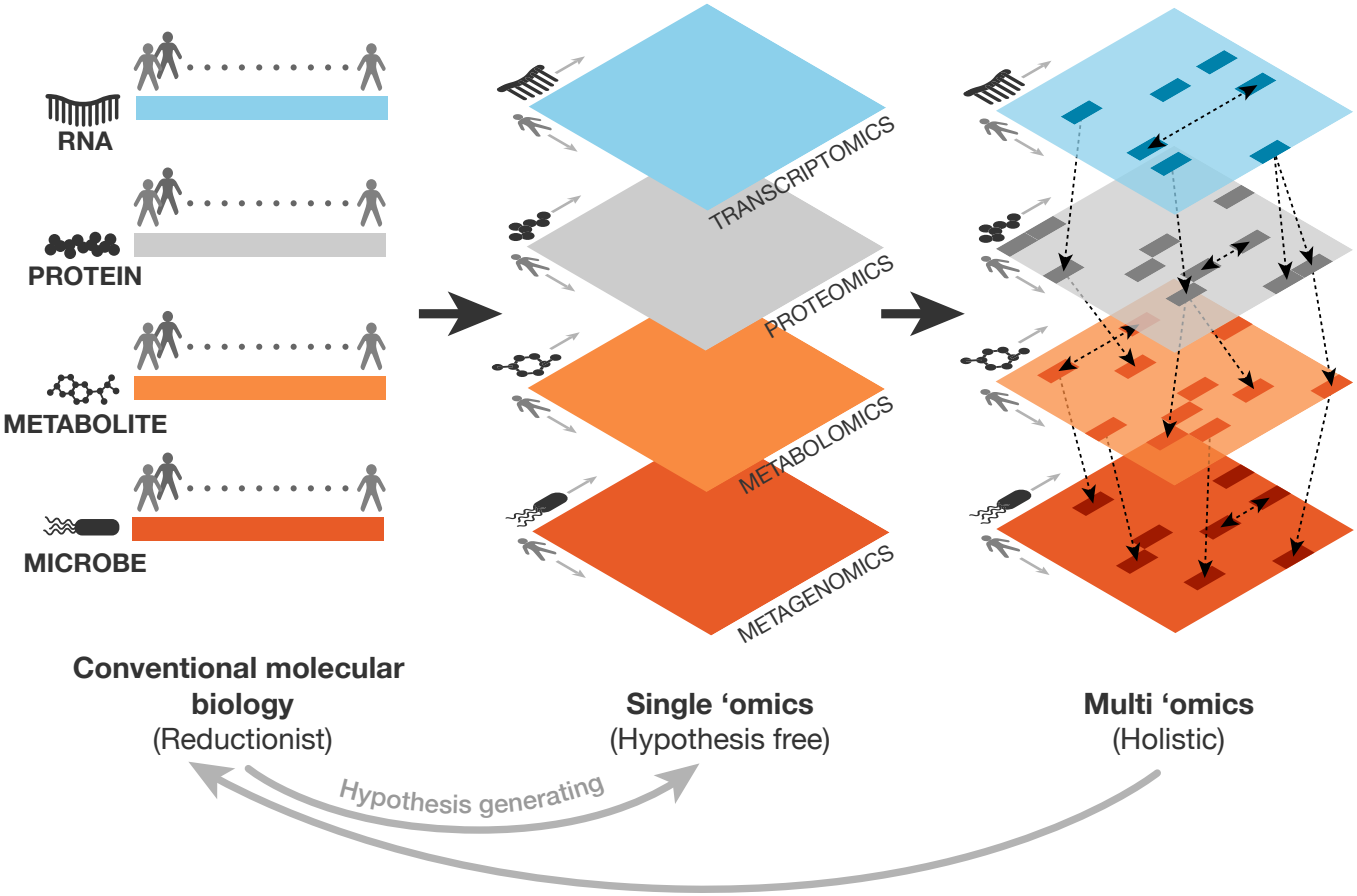
Single 'omics
(Hypothesis free)

Multi 'omics
(Holistic)

$\rightsquigarrow$ to better understand a biological system

(we won't cover data integration here, but the concepts we will cover constitute the basis for multivariate data integration)

Why integrating 'omics data?



$\rightsquigarrow$ Adopt a holistic view, rather than a traditional, reductionist, hypothesis-driven view

Research hypothesis

> Molecular entities act together to trigger cells' responses. We need to shift the 'one-gene hypothesis' paradigm to obtain deeper insight into biological systems.

$\rightsquigarrow$ shift the univariate statistics paradigm to obtain deeper insight into biological systems

Multivariate statistical methods to:

- Identify a combination of biomarkers rather than univariate biomarkers
- Reduce the dimension of the data for a better understanding of complex biological
- Integrate multiple sources of biological data

Disclaimer

"All models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind."

George E. P. Box, statistician
1919 - 2013

Table of Contents

Principal Component Analysis

PCA: the workhorse for linear multivariate statistical analysis is an (almost) compulsory first step in exploratory data analysis to:

- Understand the underlying data structure
- Identify bias, experimental errors, batch effects.

Original variables are replaced by artificial variables (principal components) which explain as much information as possible from the original data.

In PCA, the variance == information contained in the data
$\rightarrow$ data compression

A linear combinations of variables

	Height	Weight
1	174.0	65.6
2	175.3	71.8
3	193.5	80.7
4	186.5	72.6
5	187.2	78.8
6	181.5	74.8
7	184.0	86.4
8	184.5	78.4
9	175.0	62.0
10	184.0	81.6

Let's assign two coefficients $w_1 = 0.5$ and $w_2 = 2$ to the variables Height and Weight respectively.

Linear combination of the variables Height and Weight

	Height		Weight		Linear combination
	174.0		65.6		218.20
	175.3		71.8		231.25
	193.5		80.7		258.15
	186.5		72.6		238.45
$0.5 \times$	187.2	$+ \quad 2 \times$	78.8	$=$	251.20
	181.5		74.8		240.35
	184.0		86.4		264.80
	184.5		78.4		249.05
	175.0		62.0		211.50
	184.0		81.6		255.20

Two variables are summarised into a single variable - a linear combination, a.k.a component.

$\rightarrow$ efficient algorithms for big data

$\rightarrow$ challenge: identify the coefficients assigned to each variable.

Now a 'bigger' data set

data =

	V1	V2	V3	V4	V5
1	3.97	3.16	3.54	3.89	2.11
2	2.05	2.36	3.89	3.50	1.58
3	3.36	3.11	4.20	2.26	-0.09
4	4.15	4.79	5.43	3.30	2.00
5	2.91	3.05	3.87	4.99	1.69
6	3.44	3.96	3.54	3.89	2.11
7	3.65	2.22	3.89	3.50	1.58
8	4.40	3.60	4.20	2.26	-0.09
9	2.68	2.29	5.43	3.30	2.00
10	3.85	3.89	3.87	4.99	1.69

Components =

	C1	C2	C3	C4	C5
1	0.62	0.31	-0.45	0.71	-0.15
2	0.41	-1.53	0.05	-0.34	-0.47
3	-1.91	-0.53	-0.35	-0.39	-0.07
4	-0.33	1.72	1.31	-0.08	-0.11
5	1.36	-0.24	-0.38	-0.52	0.35
6	0.69	0.64	-0.29	0.02	-0.69
7	0.07	-0.71	-0.29	0.73	0.26
8	-2.16	0.46	-0.53	0.04	0.18
9	0.14	-1.10	1.45	0.13	0.27
10	1.11	0.97	-0.52	-0.31	0.42

e.g. PCA: original variables are replaced by artificial variables (components) explaining as much information as possible from the original data, with no overlap of information.

Not all components are needed to summarise the information.

PCA and the variance covariance matrix

data =

	V1	V2	V3	V4	V5
1	3.97	3.16	3.54	3.89	2.11
2	2.05	2.36	3.89	3.50	1.58
3	3.36	3.11	4.20	2.26	-0.09
4	4.15	4.79	5.43	3.30	2.00
5	2.91	3.05	3.87	4.99	1.69
6	3.44	3.96	3.54	3.89	2.11
7	3.65	2.22	3.89	3.50	1.58
8	4.40	3.60	4.20	2.26	-0.09
9	2.68	2.29	5.43	3.30	2.00
10	3.85	3.89	3.87	4.99	1.69

COV(data) =

	V1	V2	V3	V4	V5
V1	0.52	0.39	0.00	-0.12	-0.14
V2	0.39	0.69	0.08	0.06	0.03
V3	0.00	0.08	0.48	-0.22	0.03
V4	-0.12	0.06 -	0.22	0.87	0.54
V5	-0.14	0.03	0.03	0.54	0.71

Sum of variances on the diagonal:

$0.52+0.69+0.48+0.87+0.71 = 3.27$

Ref Appendix Covariance matrix

PCA and the variance covariance matrix

Original variables are replaced by a few principal components which explain as much information (variance) as possible from the original data. PCs are orthogonal to each other (covariance = 0).

PCs =

	PC1	PC2	PC3	PC4	PC5
1	0.62	0.31	-0.45	0.71	-0.15
2	0.41	-1.53	0.05	-0.34	-0.47
3	-1.91	-0.53	-0.35	-0.39	-0.07
4	-0.33	1.72	1.31	-0.08	-0.11
5	1.36	-0.24	-0.38	-0.52	0.35
6	0.69	0.64	-0.29	0.02	-0.69
7	0.07	-0.71	-0.29	0.73	0.26
8	-2.16	0.46	-0.53	0.04	0.18
9	0.14	-1.10	1.45	0.13	0.27
10	1.11	0.97	-0.52	-0.31	0.42

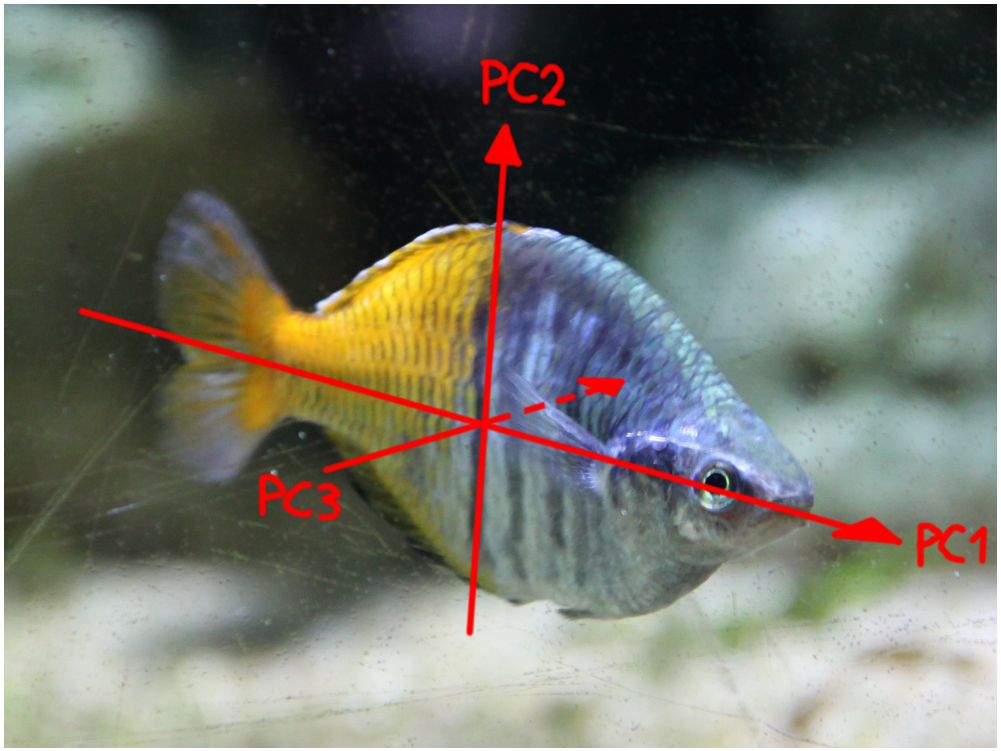
COV(PCs) =

	PC1	PC2	PC3	PC4	PC5
PC1	1.39	0.00	0.00	0.00	0.00
PC2	0.00	1.00	0.00	0.00	0.00
PC3	0.00	0.00	0.56	0.00	0.00
PC4	0.00	0.00	0.00	0.19	0.00
PC5	0.00	0.00	0.00	0.00	0.13

Sum of variances on the diagonal:

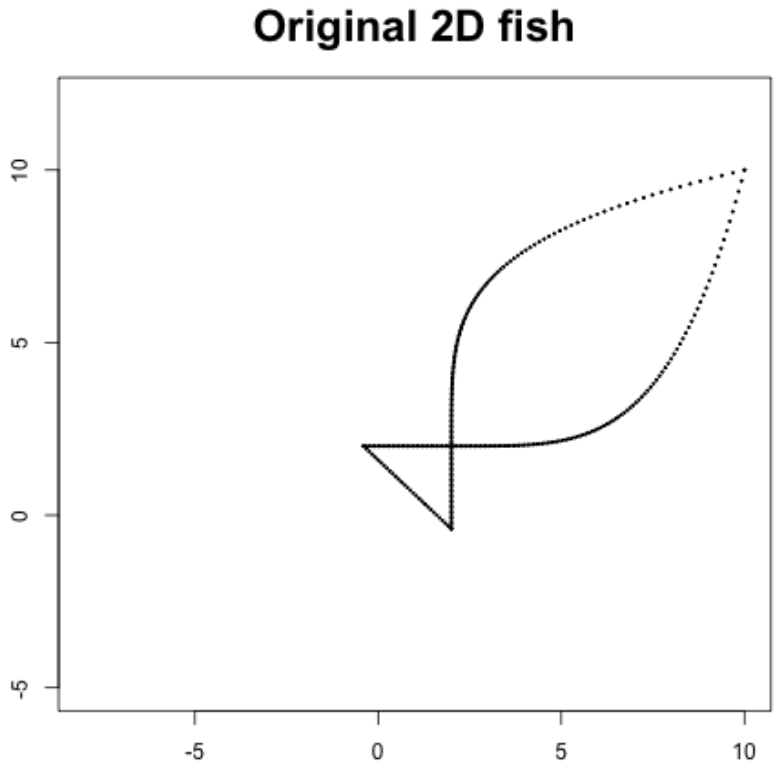
$1.39+1+0.56+0.19+0.13 = 3.27$

Data compression is used for visualisation



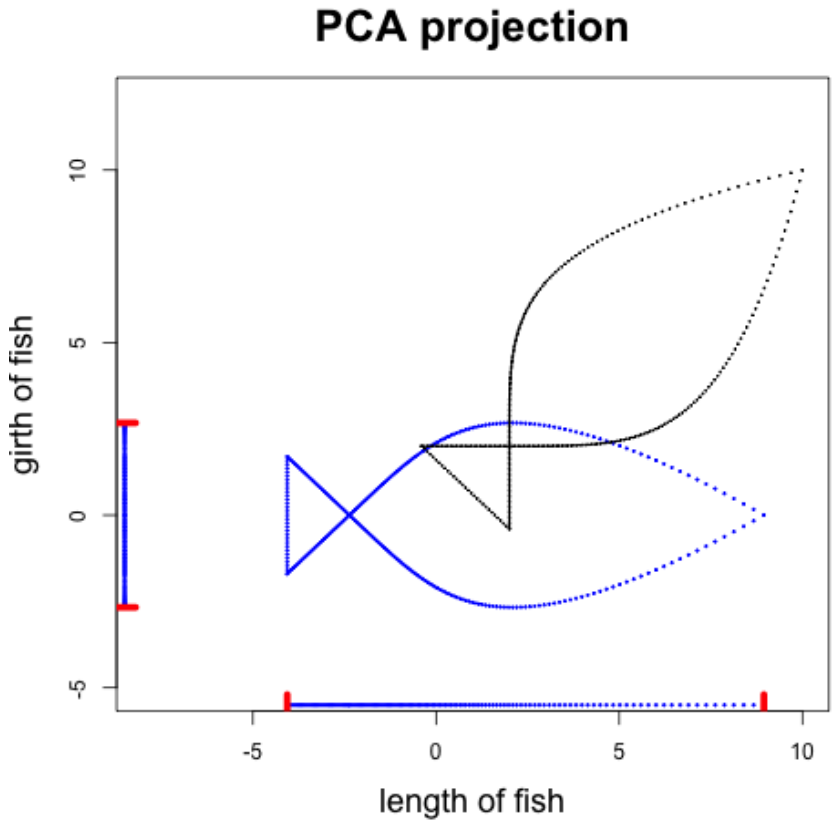
How to summarise a fish into a smaller dimension?

Data compression is used for visualisation



Each component (axis) will explain a certain information related to the fish, such as length, width (or something less explicit)

Data compression is used for visualisation



If we plot those components, we visualise the compressed summarised information in a projected space

PCA maximises the variance of the components

Seek for the best directions in the data that account for most of the variability. Objective function:

$$\max_{||\mathbf{a}||=1} var(X\mathbf{a})$$

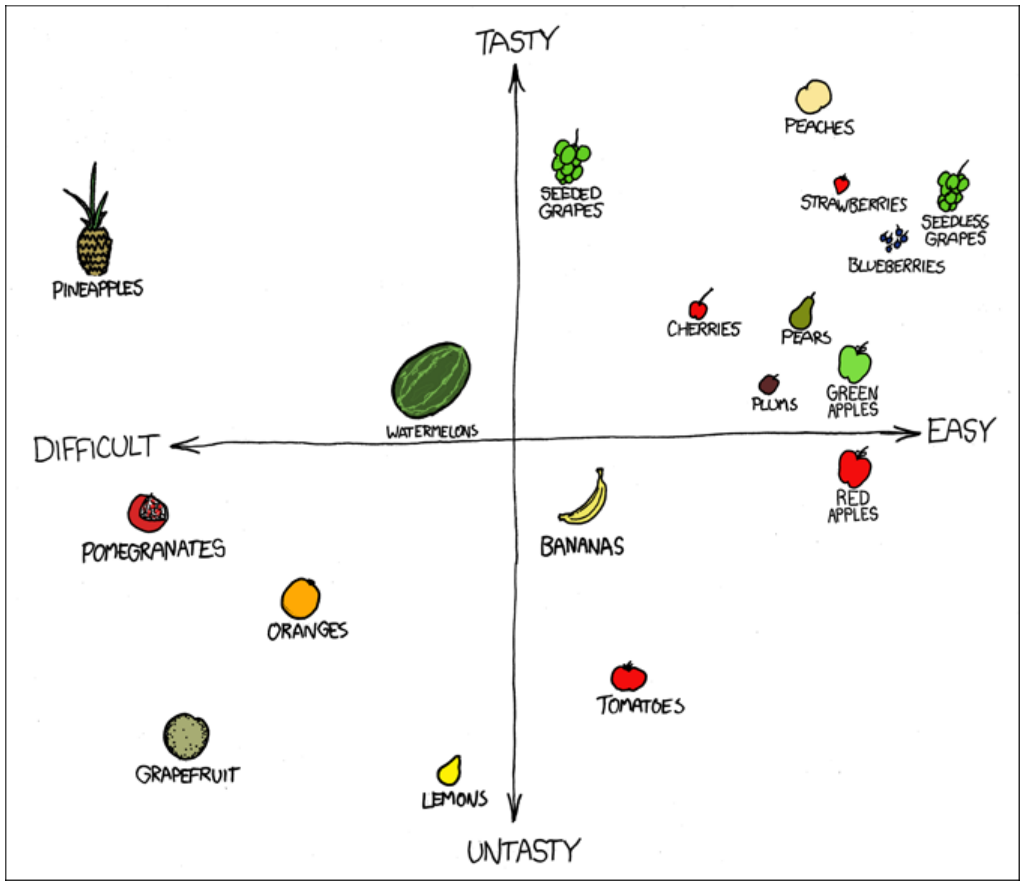
Each principal component $\mathbf{t}$ is a linear combinations of the original variables ($\mathbf{t} = X\mathbf{a}$):

$$\mathbf{t} = a_1\mathbf{x}^1 + a_2\mathbf{x}^2 + \cdots + a_p\mathbf{x}^p$$

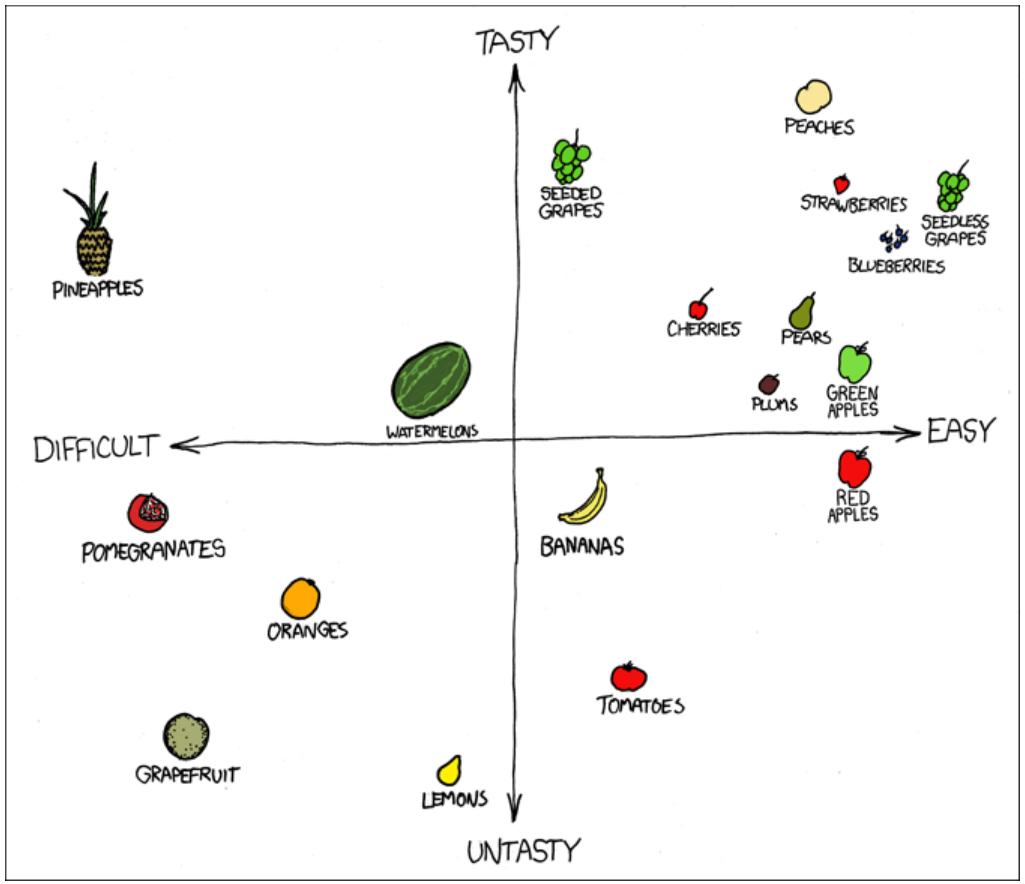
- $\mathbf{X}$ is a $n \times p$ data matrix with $\{\mathbf{x}^1, \ldots, \mathbf{x}^p\}$ the p variable profiles.
- $\mathbf{t}$ is the **first** principal component with max. variance
- $\{a_1, \ldots, a_p\}$ are the weights in the linear combination

The data are projected into a smaller subspace

- Each principal component is orthogonal to each other to ensure that no redundant information is extracted.
- The new PCs form a smaller subspace of dimension $<< p$.
- Each value in the principal component corresponds to a score for each sample
  $\rightarrow$ we project each sample into a new subspace spanned by the PCs
- Approximate representation of the data points in a low dimensional space
- Summarize the information related to the variance

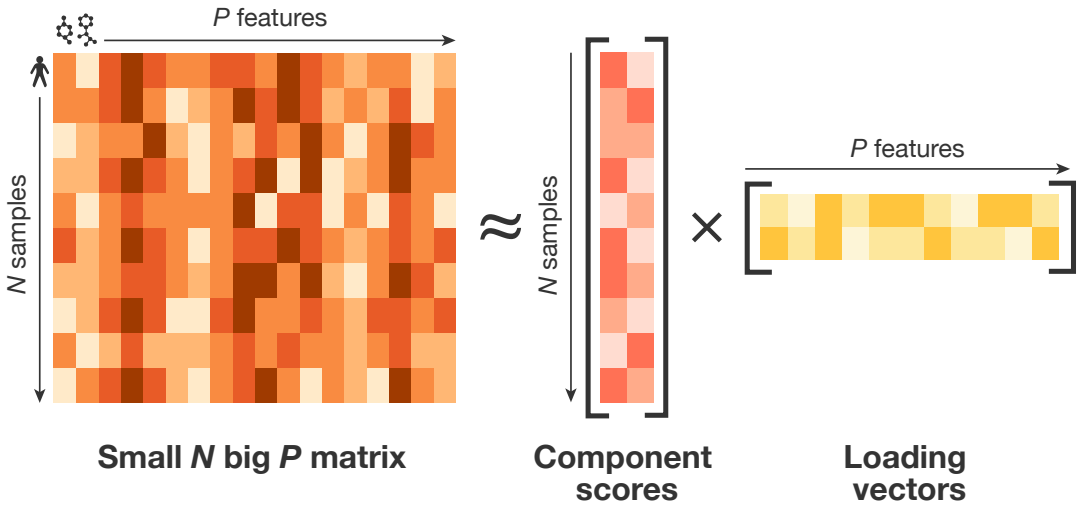


Summarise the data into 2 components (dimensions) then project the data in the space spanned by those 2 components.



Each component has a 'meaning' as it tends to spread the observations (here fruit) according to their measured characteristics.

PCA is a matrix decomposition



PCA solved using
Singular Value Decomposition:

$$\mathbf{X} = \mathbf{U}\mathbf{\Delta}\mathbf{A}^T$$

- $\mathbf{\Delta}$ diagonal matrix with δ_h
- $\mathbf{T} = \mathbf{U}\mathbf{\Delta}$, $\mathbf{T}$ contains the PCs $\mathbf{t}^h$
- $\mathbf{A}$ contains the loading vectors $\mathbf{a}^h$
- $h = 1..H$ is the number of PCs

The variance of the first principal component $\mathbf{t}^1$ is equal to its associated eigenvalue $\delta_1^2/(n-1)$, and so on for the other PCs.

Computing PCA

Several ways of solving PCA

- Eigenvalue decomposition: the old way, does not work well in high dimension $S\mathbf{a} = \lambda\mathbf{a}$; $\mathbf{t} = \tilde{X}\mathbf{a}$

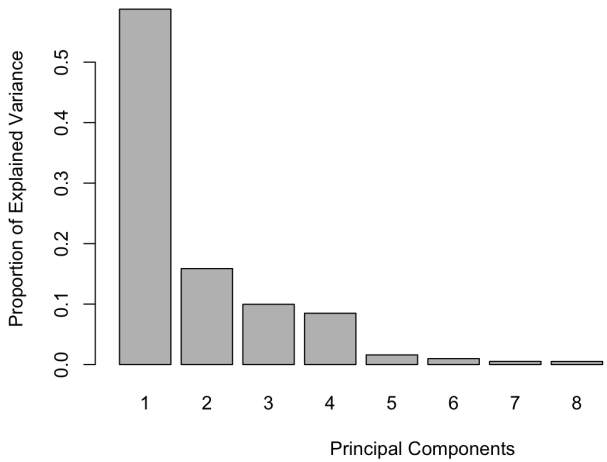
  S = variance covariance matrix or correlation matrix if $\tilde{X}$ is scaled
- NIPALS algorithm: long to compute but works for missing values and enables least squares regression framework (useful for variable selection)
- Singular Value Decomposition (SVD): the easiest and fastest, implemented in most software (`svd()`).

Choosing the parameters in PCA

How many principal components to choose to summarize most of the information?

We can have as many components as the rank of the matrix X

- Proportion of explained variance / cumulative prop.
- Screeplot of eigenvalues. Any elbow?
- Look at sample plot. Makes sense?
- Some stat tests exist to estimate the 'intrinsic' dimension, but limitations



Cumulative proportion of explained variance for the first 8 principal components:

PC1	PC1 to 2	PC1 to 3	PC1 to 4	PC1 to 5	PC1 to 6	PC1 to 7	PC1 to 8
0.59	0.75	0.84	0.93	0.946	0.956	0.961	0.966

PCA is a visualisation tool

The PCA sample plot is the most well known PCA output.

Other important graphical outputs:

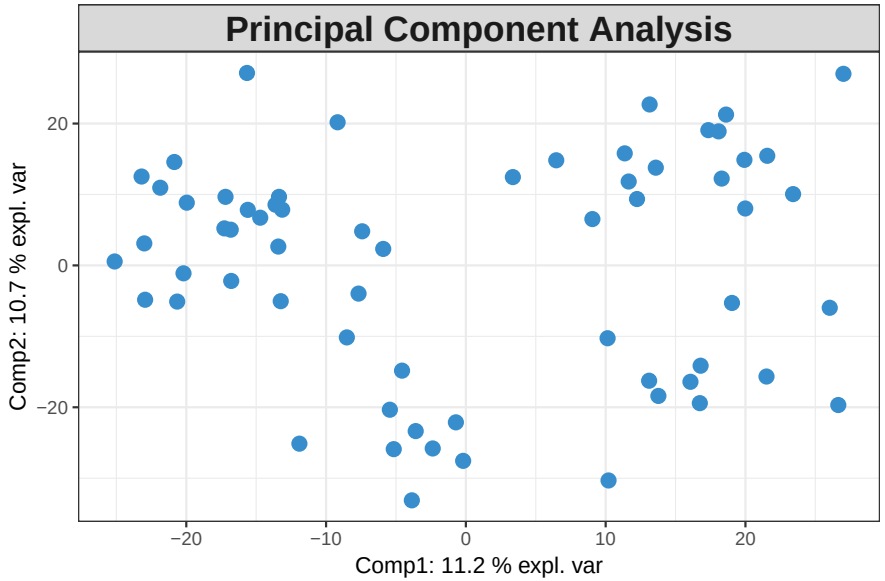
- Correlation circle plots to represent variables and their correlations*.
- Biplot to represent sample and variables in the same plot.

* González, I., Lê Cao, K.-A., Davis, M. J., Déjean, S. (2012). Visualising associations between paired 'omics data sets. BioData mining, 5(1), 19.

Sample representation: `plotIndiv` function

A principal component is a vector of length n representing the projection of each sample onto the space spanned by that PC.

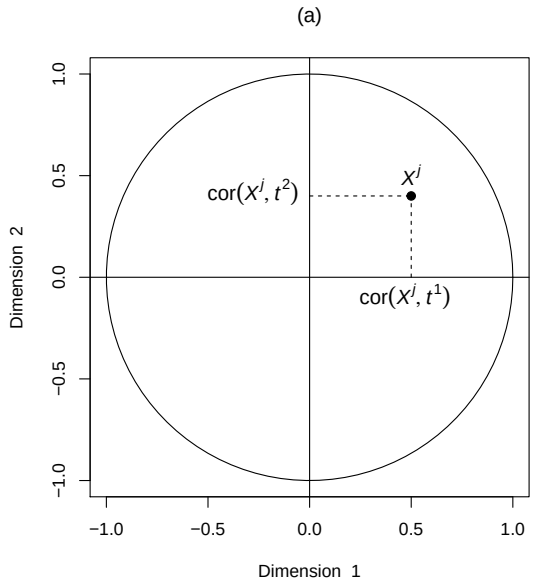
- `plotIndiv` is the scatterplot of the PCs and represents the samples into a smaller dimensional space



Variable representation: `plotVar` function

A correlation circle plot represents the correlation between variables and each variable's contribution to each component.

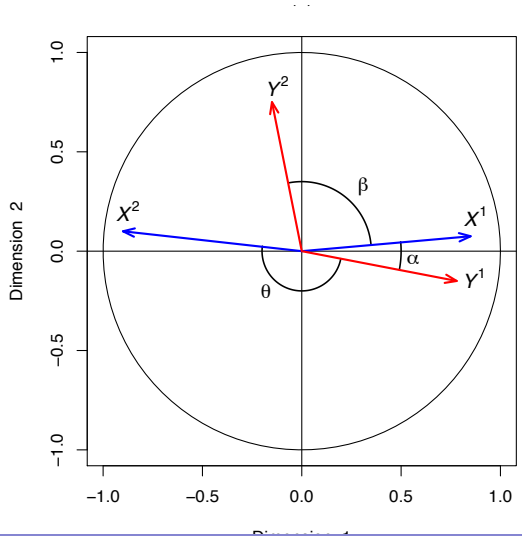
To obtain the coordinate of each variable: calculate the correlation between each variable and each PC: $\text{cor}(\mathbf{X}^j, \mathbf{t}^1), \text{cor}(\mathbf{X}^j, \mathbf{t}^2)$



Variable representation: `plotVar` function

A correlation circle plot represents the correlation between variables and each variable's contribution to each component.

- correlation between the variable and the PC = cos(angle) btw the variable vector and the PC
- correlation between two variables = cos(angle) between 2 vectors

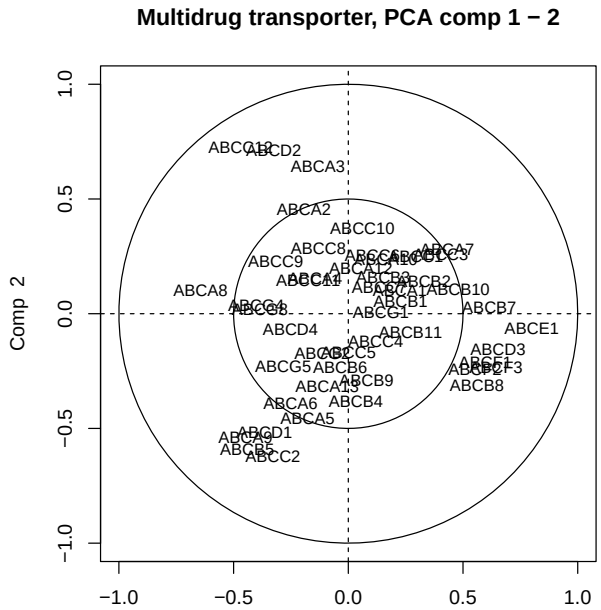


- data centered and scaled in PCA
- $cos(\alpha)$ close to 1 $\rightarrow > 0$ corr
- $cos(\beta)$ close to 0 $\rightarrow$ no corr
- $cos(\beta)$ close to -1 $\rightarrow < 0$ corr

Variable representation: `plotVar` function

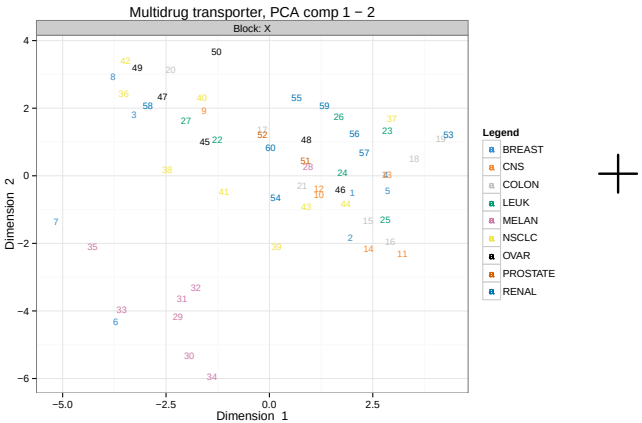
A correlation circle plot represents the correlation between variables and each variable's contribution to each component.

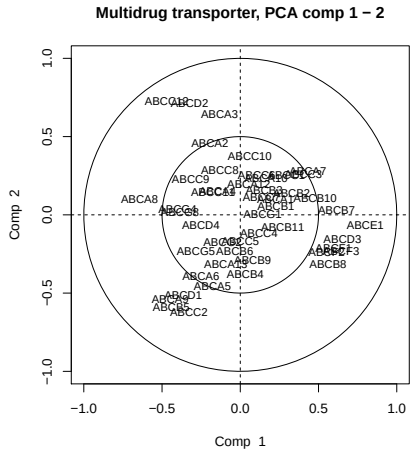
In some software the variables are represented as a dot or a name (no arrow)



The biplot

The biplot overlays both sample and variable plots to understand how they relate.





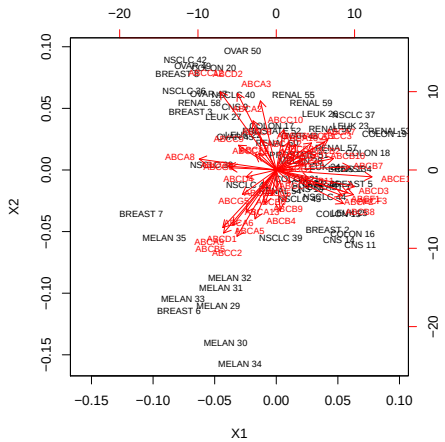


Table of Contents

Yeast metabolome

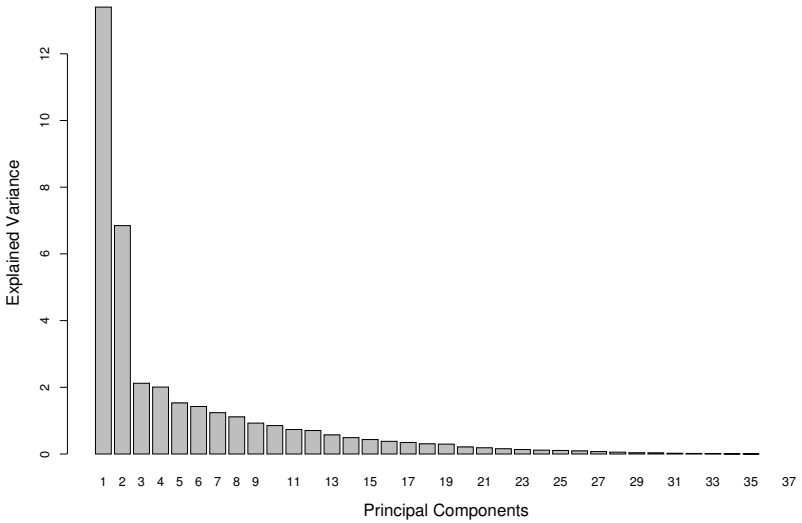
Yeast study from Villas-Boâs et al, 2005:

- Two Saccharomyces cerevisiae strains: WT and MT
- Two environmental conditions: aerobic (AER) and anaerobic (ANA)
- 37 metabolites and 55 samples
  (13 MT-AER, 14 MT-ANA, 15 WT-AER and 13 WT-ANA)

Questions:

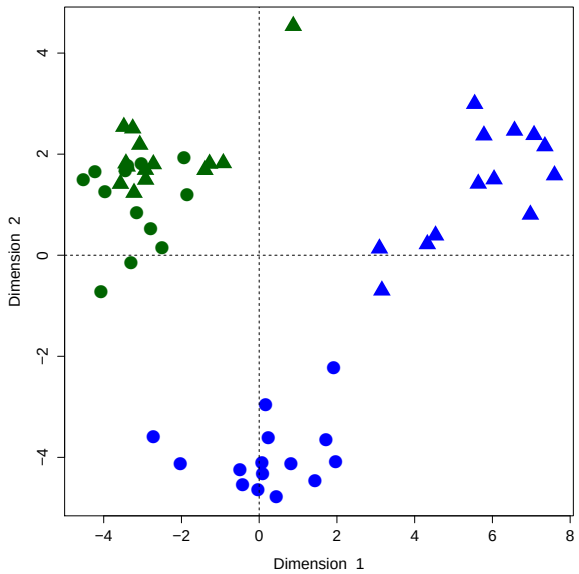
- Does the information provided by the metabolites spectra relate to the experimental conditions, or to some interfering signals?
- What is the strongest source of variation in the data: strain or environment?

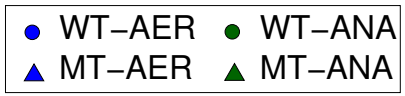
Yeast metabolome



Cumulative % of explained variance:

- Two PCs: 54.72% of total var
- Three PCs: 60.45% of total var





Yeast metabolome

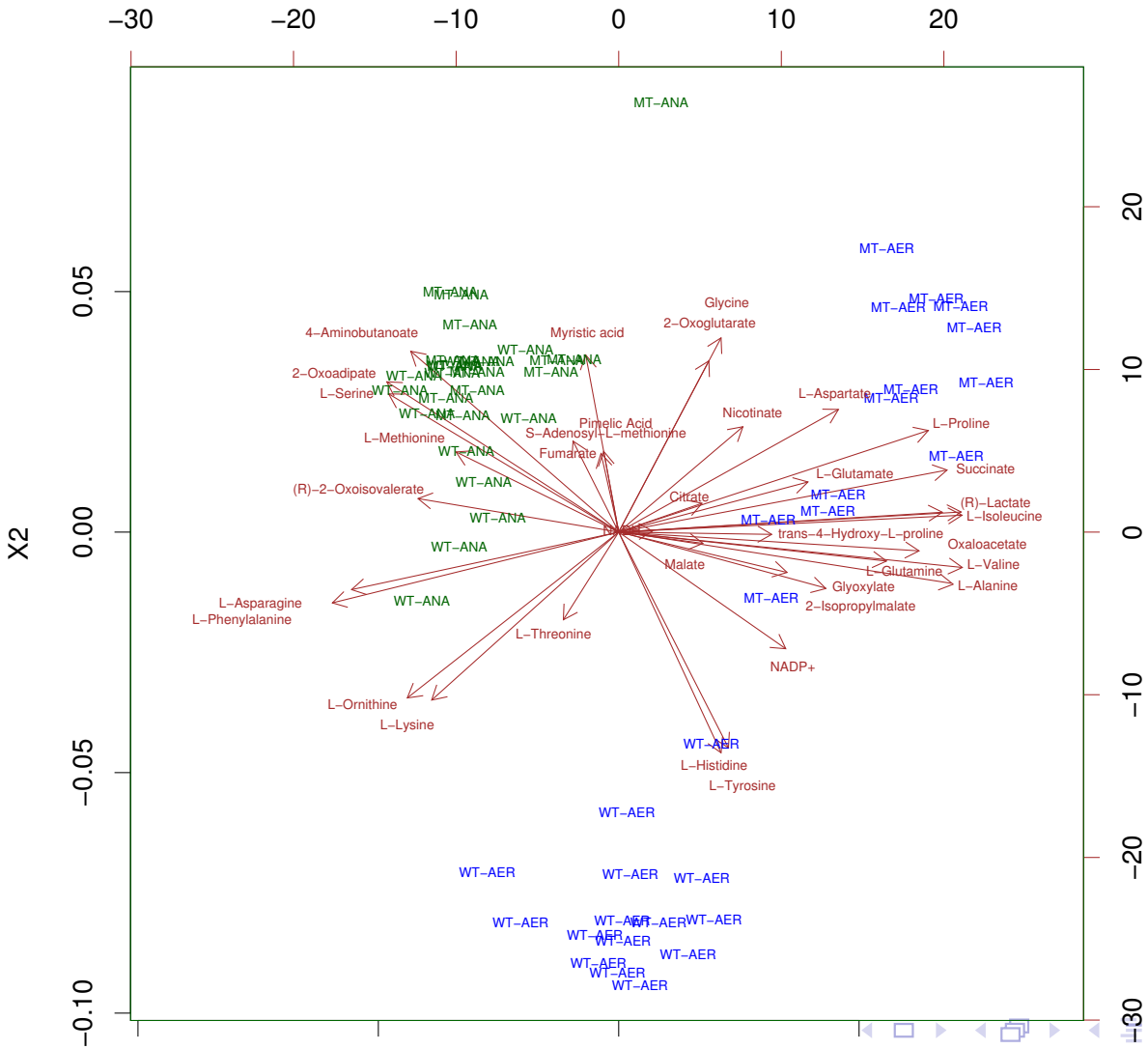


Table of Contents

PCA summary

- PCA is a matrix decomposition technique that allows for dimension reduction.
- Perform a PCA first to understand the sources of variation in your data.
- Always report the % explained variance per component.
- PCA can highlight 'batch effect' or unexpected sources of variation in the data.