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

Know Your Data: Exploratory Data Analysis

Shraddha Pai

Analysis Using R

June 13, 2024

bioinformatics.ca

OICR
Ontario Institute for Cancer Research

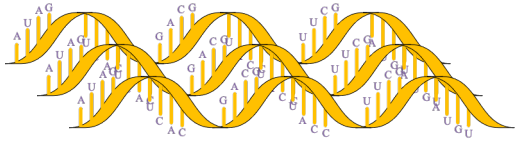
UNIVERSITY OF TORONTO

Learning Objectives

- By the end of this lecture, you will:
  - Know the basic terminology for terms in a statistical model
  - Know how to perform systematic exploratory data analysis to to identify sources of biological and technical variation
  - Appreciate the value of exploring missingness in your data
  - Have a high level understanding of clustering and be able to cluster your data

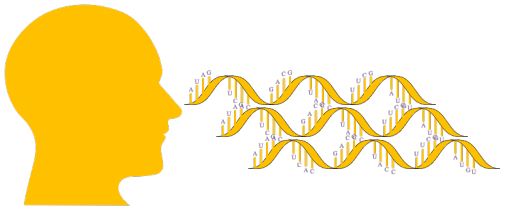
Sources of variation in a study

Goal: “Find transcriptomic biomarkers of disease”



Sources of variation in a study

Basic terminology in statistical model building



$$xpr = \beta_0 + \beta_1(disease) + \epsilon$$

Response variable

Dependent variable

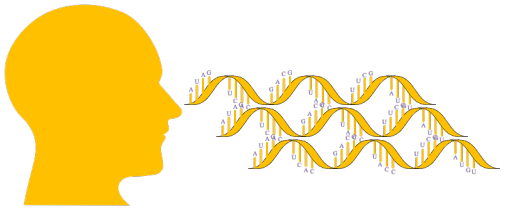
Explanatory variable

Independent variable

Residual; unmodelled variation

Sources of variation in a study

Basic terminology in statistical model building

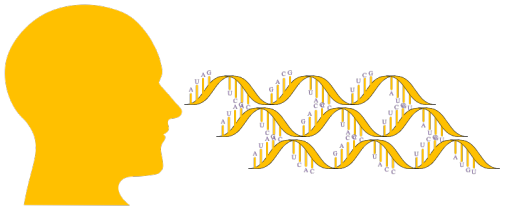


coefficients; weights

$$xpr = \beta_0 + \beta_1(disease) + \epsilon$$

Intercept

Goal: “Find transcriptomic biomarkers of disease”



$$xpr = \beta_0 + \beta_1(disease) + \epsilon$$

Response variable

Explanatory variable

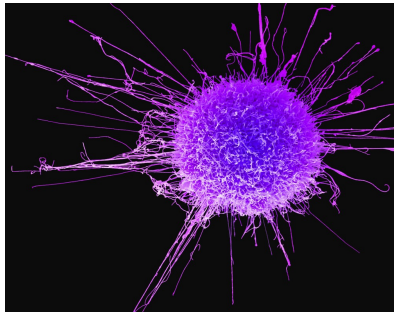
Residual; unmodelled variation

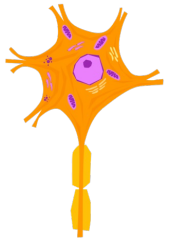
Dependent variable

Independent variable

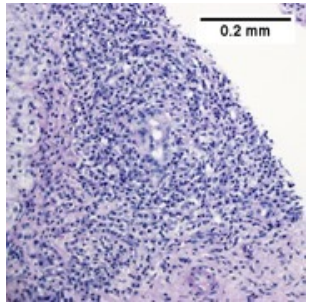
What could be affecting your outcome?

Biological sources of variation





Technical sources of variation



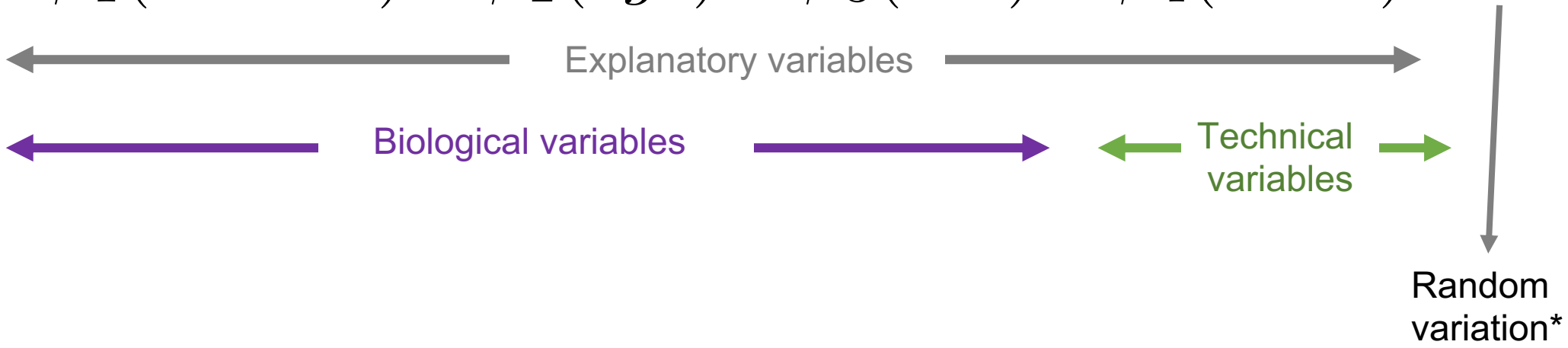
"Which of these is affecting my data?"

Visualize & quantify sources of variation:
* clustering
* dimensionality reduction
  (PCA, UMAP, tSNE)

$$xpr = \beta_0 + \beta_1(disease) + \beta_2(age) + \beta_3(sex) + \beta_4(batch) + \epsilon$$

Final model

$$xpr = \beta_0 + \beta_1(disease) + \beta_2(age) + \beta_3(sex) + \beta_4(batch) + \epsilon$$



- "Random": Values sampled from defined statistical distribution (e.g., Normal distribution, Binomial, Poisson etc.,)

Missingness

Missingness happens!

- Clinical data may be incomplete (e.g., participant didn't answer questionnaire)
- Multi-'omic data, some participants missing an assay
- Some measures didn't pass QC

Solutions:

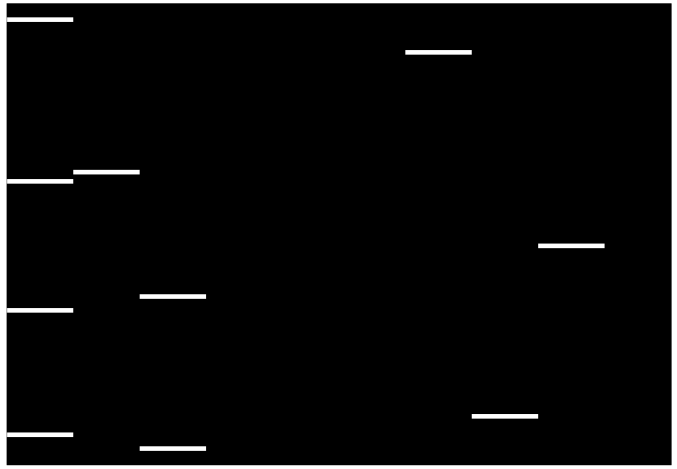
- Remove rows/cols with "excessive" missingness – use field convention where possible
- Use imputation to "guess" at missing values

Ensure your approach is defensible! Careful what you "guess".

What are trade-offs in each solution?

Checkered view of data table. White shows missing (NA).

Unstructured missingness



Structured missingness

Biased, structured missingness

Extreme situation: What if only one group is missing data, and we blindly impute?

Lesson: Where possible, look at your data.

Goals of exploratory data analysis are to:

1. Identify magnitude of KNOWN biological and technical variation

2. Identify sources of UNKNOWN variation

3. Detect OUTLIER samples

4. Characterize MISSINGNESS

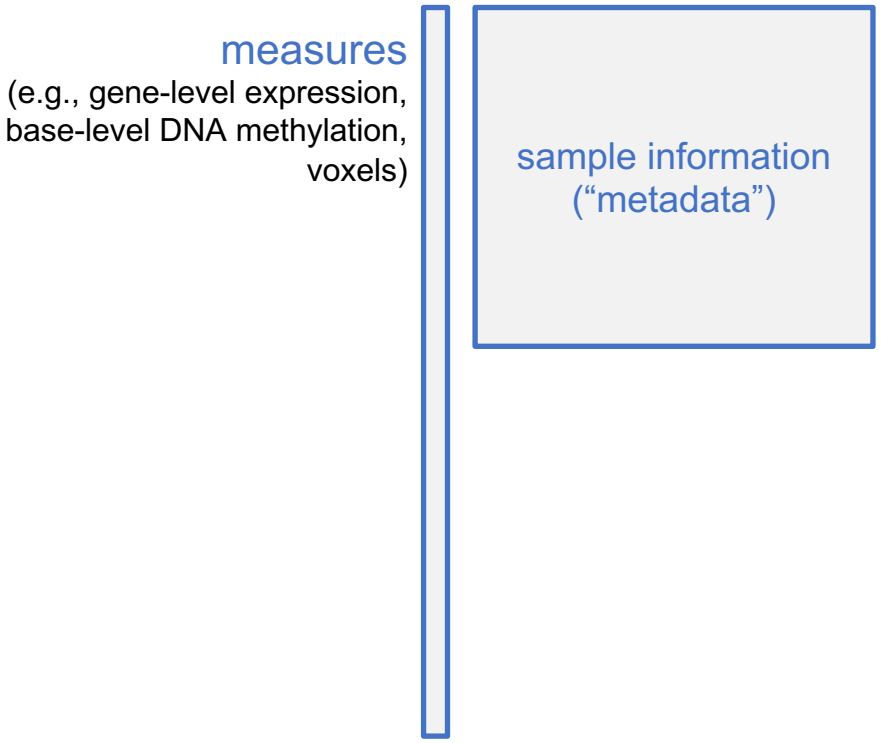
Goals of exploratory data analysis are to:

What you learn	How you learn it	What you do about it
1. Identify magnitude of KNOWN biological and technical variation	PCA*, clustering, prior knowledge	Add terms to model
2. Identify sources of UNKNOWN variation	PCA*, clustering, surrogate value analysis	Add terms to model
3. Detect OUTLIER samples	PCA*, clustering	 Exclude samples from analysis
4. Characterize MISSINGNESS	Plot missingness	• Exclude measures / samples • Define imputation strategy

** Or other dimensionality reduction technique*

Exploratory data analysis: Best practices checklist

Input: 2 tables: DATA and METADATA



Initial exploration

✅ ***How many*** *measures in metadata?*
✅ ***How many*** *measures in data?*
✅ *What are the* ***distribution*** *of the measures?*

✅ Is there **missingness**?

Decide solution

✅ Examine data **correlation** structure
* Does the data have natural groupings?

✅ How many major **sources of variation** are there?
✅ Do they map to biological or technical variables?

Expand model

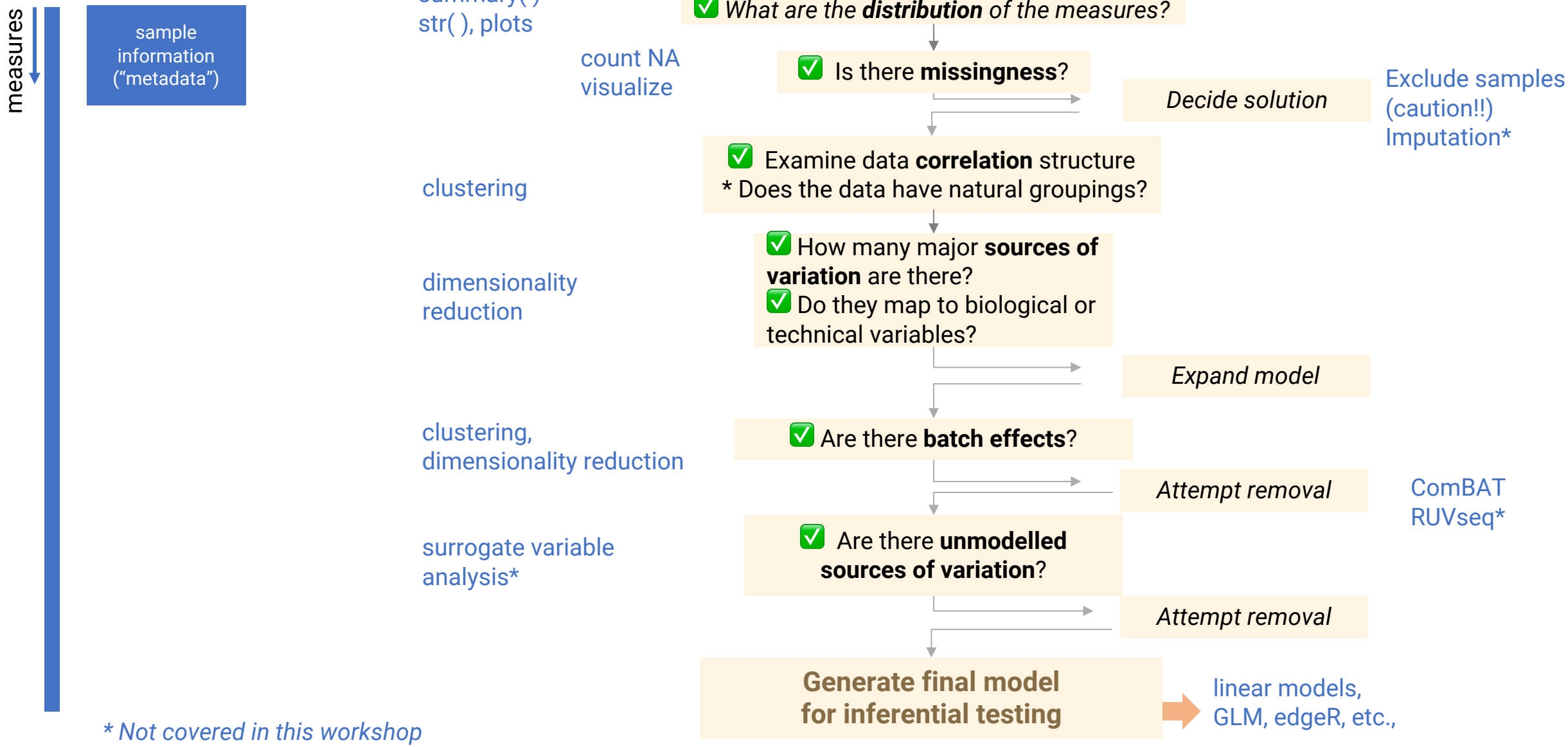
✅ Are there **batch effects**?

Attempt removal

✅ Are there **unmodelled sources of variation**?

Attempt removal

Generate final model for inferential testing



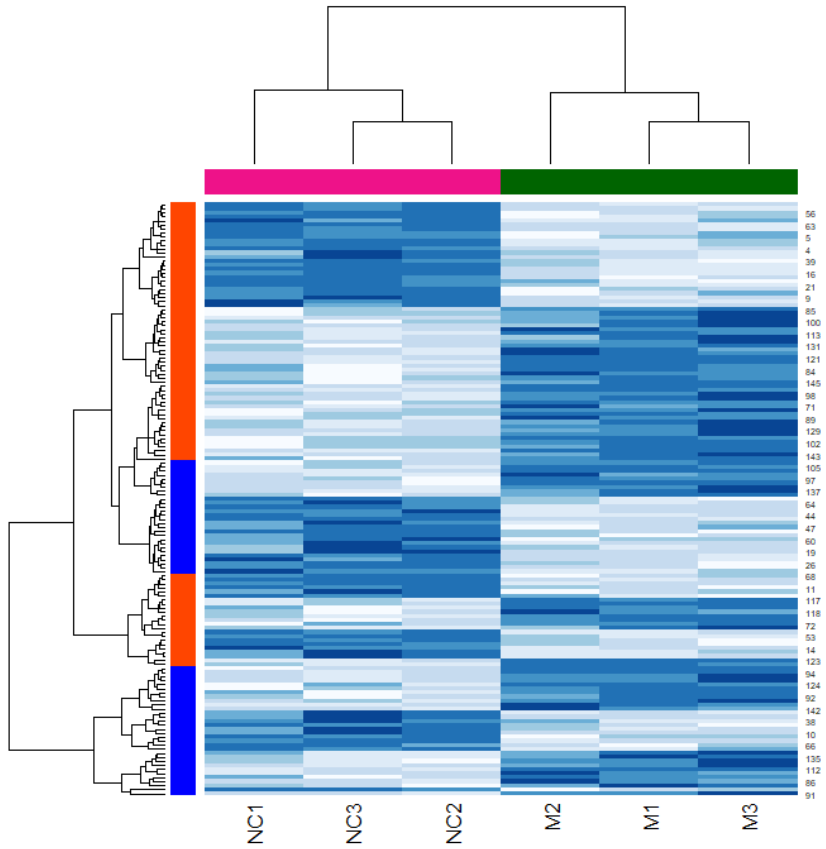
** Not covered in this workshop*

Clustering

High-level purpose: find groups in your data

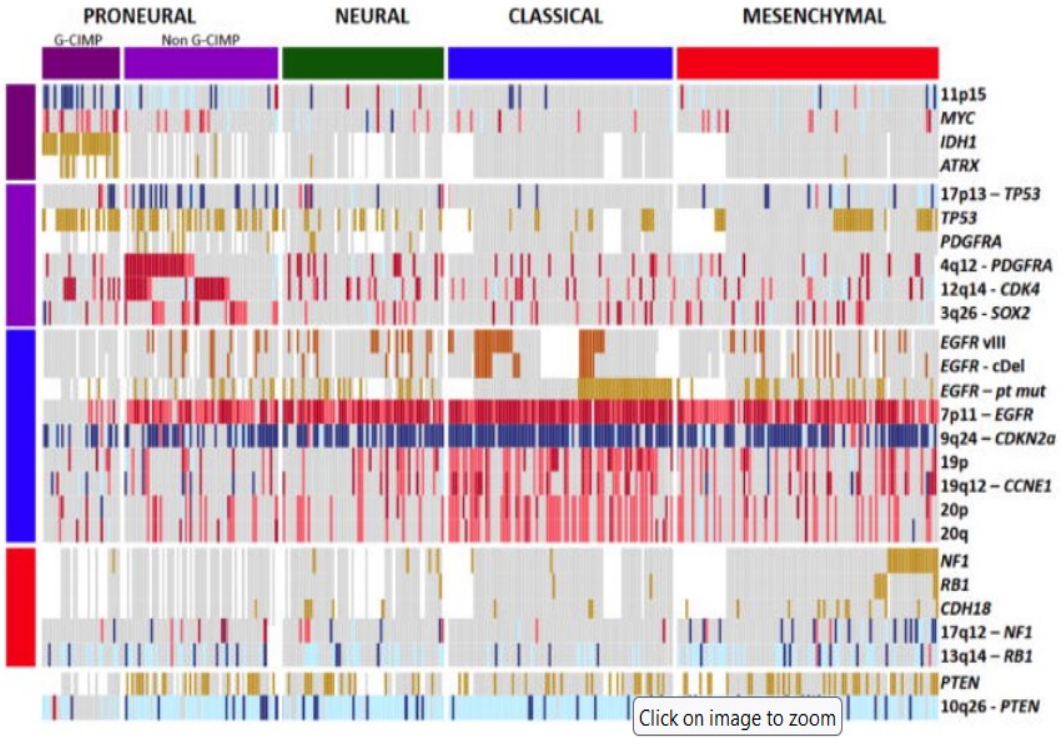
Your particular purpose (may be):

- Identify batches in your data
- Identify patient subtypes
- Identify groups of coexpressed genes

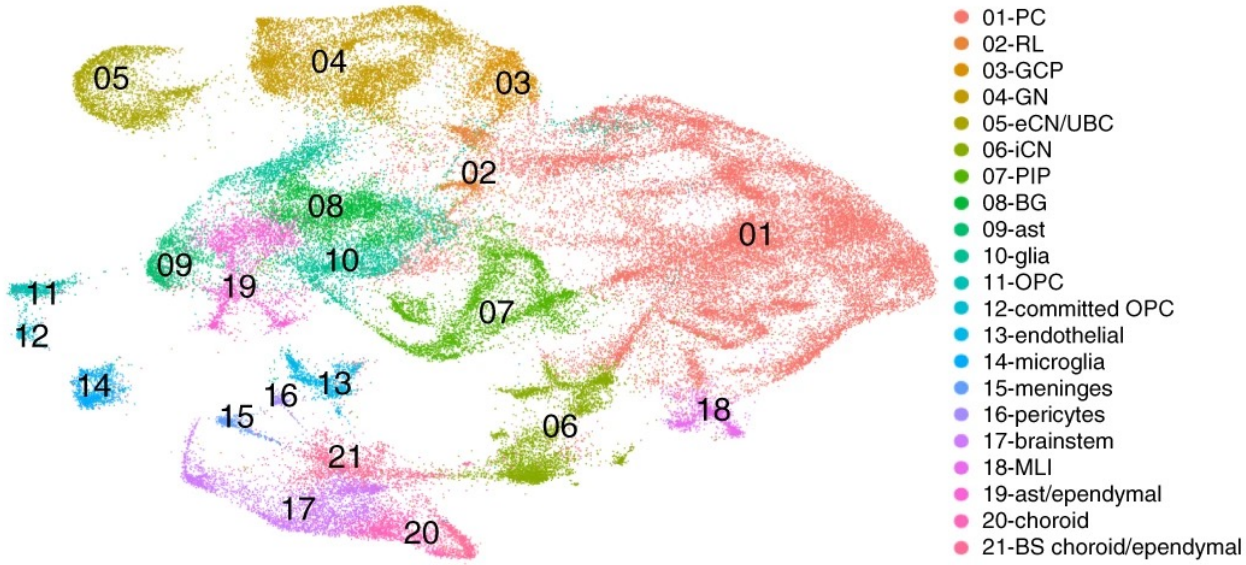


Applications of clustering

Molecular subgroups of glioblastoma



Cell clusters in single-cell genomics



Brennan et al (2013). *Cell*.
Aldinger et al. (2021). *Nat Neurosci*.

The heart of clustering: Defining distance between samples

When clustering, you need to find a way to quantify how similar/dissimilar observations are from one another.

This quantity is your **"distance metric"**

Different data types require different distance metrics

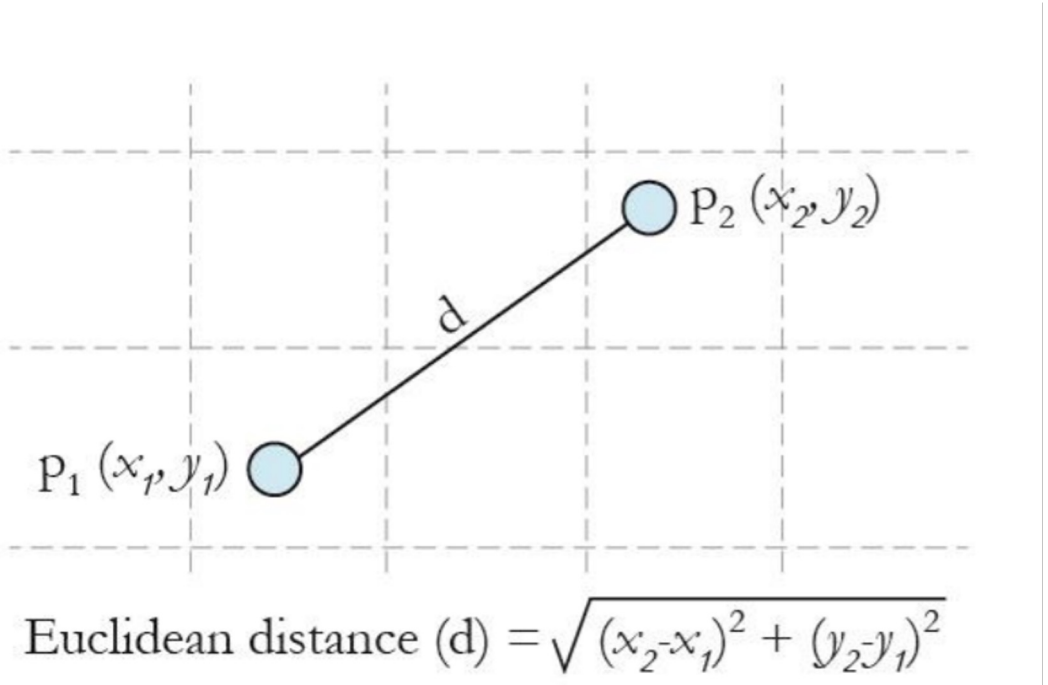
Some data types have many distance metrics, all which come with their own properties.

Distance metrics

Continuous variables:

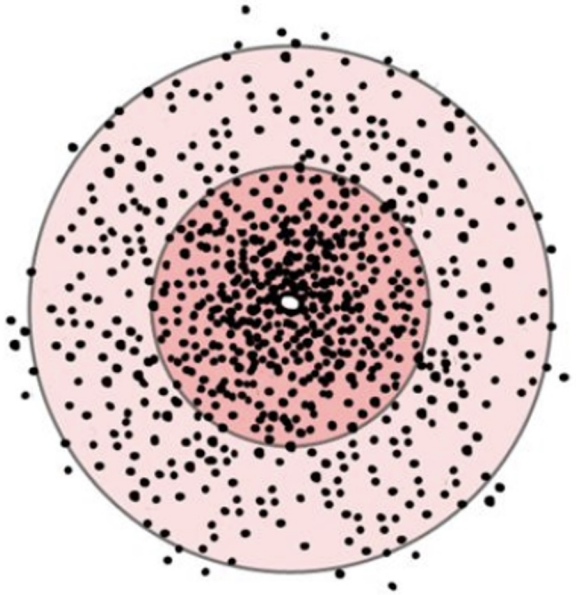
- Euclidean distance: Root squared error

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$



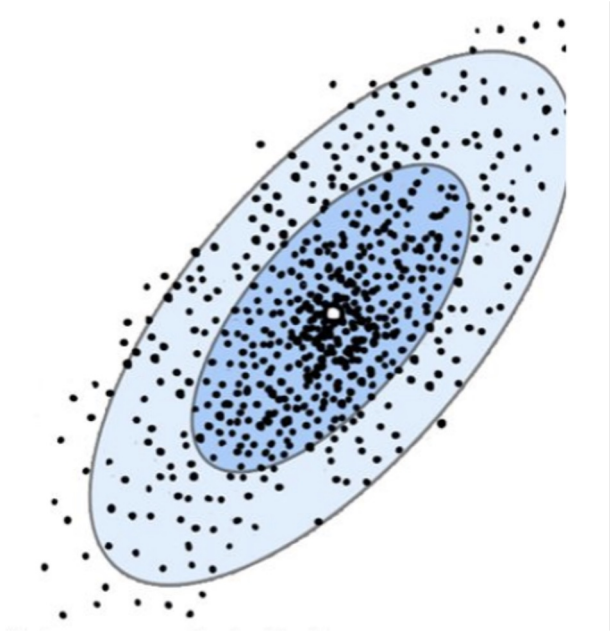
Distance metrics

Euclidean Distance
Ignores correlation between the variables; best if you know data are uncorrelated
(*i.e.,* knowing x does not tell us anything about y).



Mahalanobis Distance
Metric takes into account correlation between variables. Best used if you suspect variables are correlated.

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$$



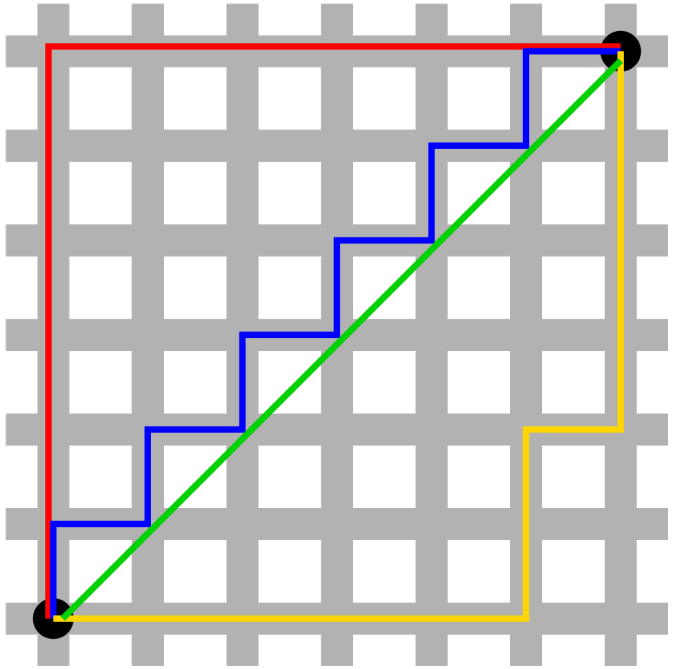
Related term: “similarity”

Distance metrics

Continuous variables:

- Euclidean distance
- Mahalanobis distance
- Manhattan or “block” distance:

$$d_M(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$



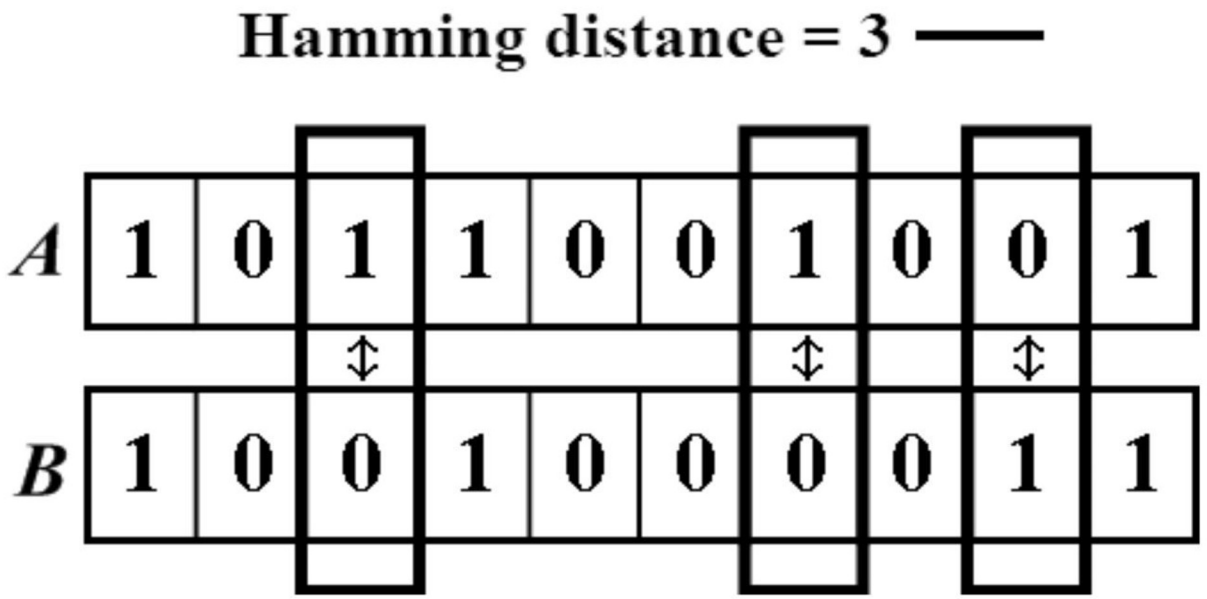
Distance metrics

Continuous variables:

- Euclidean distance: Root squared error
- Mahalanobis distance (Normalized Euclidean Distance)
- Manhattan distance

Categorical variable:

- Hamming distance (number of mismatches)



Common clustering approaches

- Hierarchical
- K-means
- Many more…
  - e.g., Spectral clustering for networks

Hierarchical Clustering

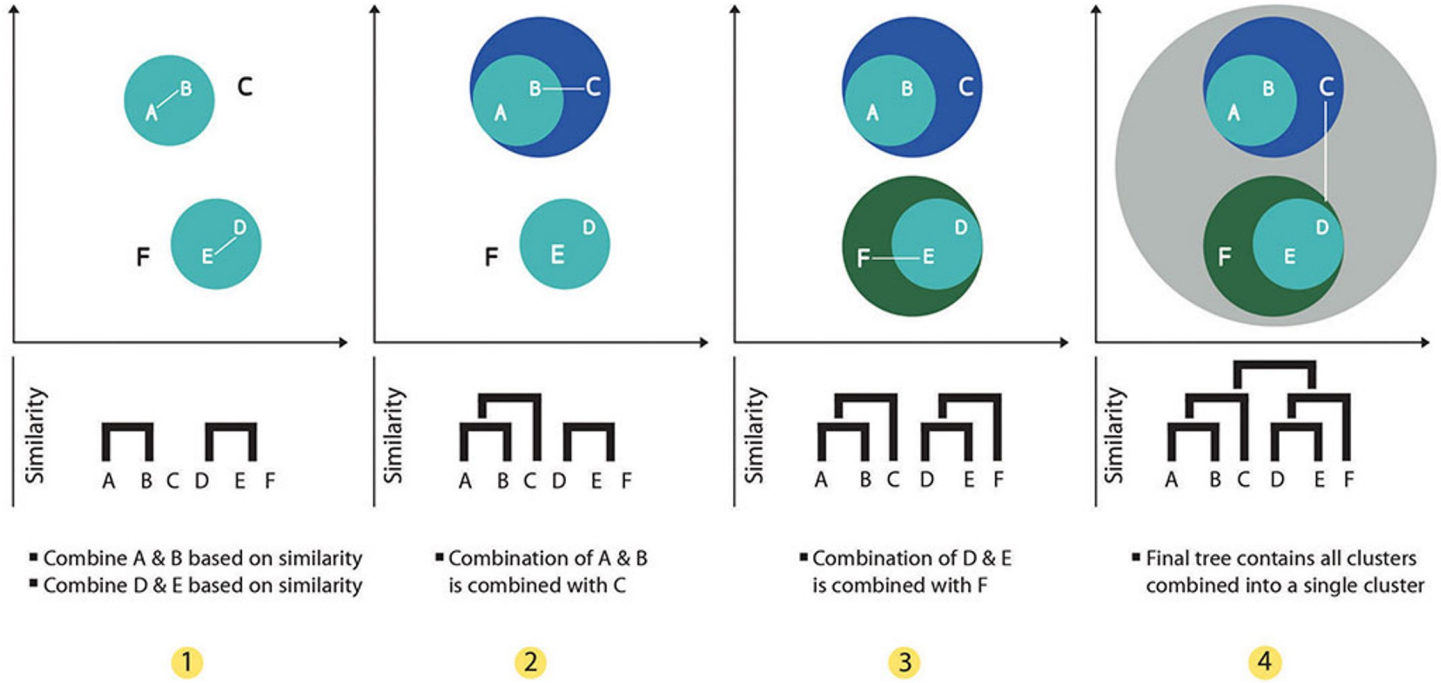
Steps:

- Build dendrogram
- Choose cut point (based on dendrogram or K)

Hierarchical Clustering

Steps:

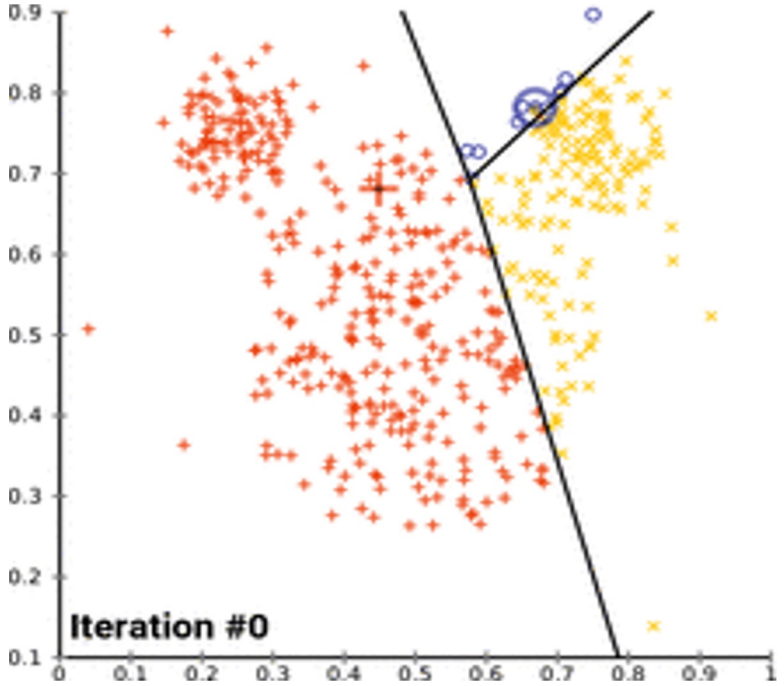
- Build dendrogram
- Choose cut point (based on dendrogram or K)



K-Means Clustering

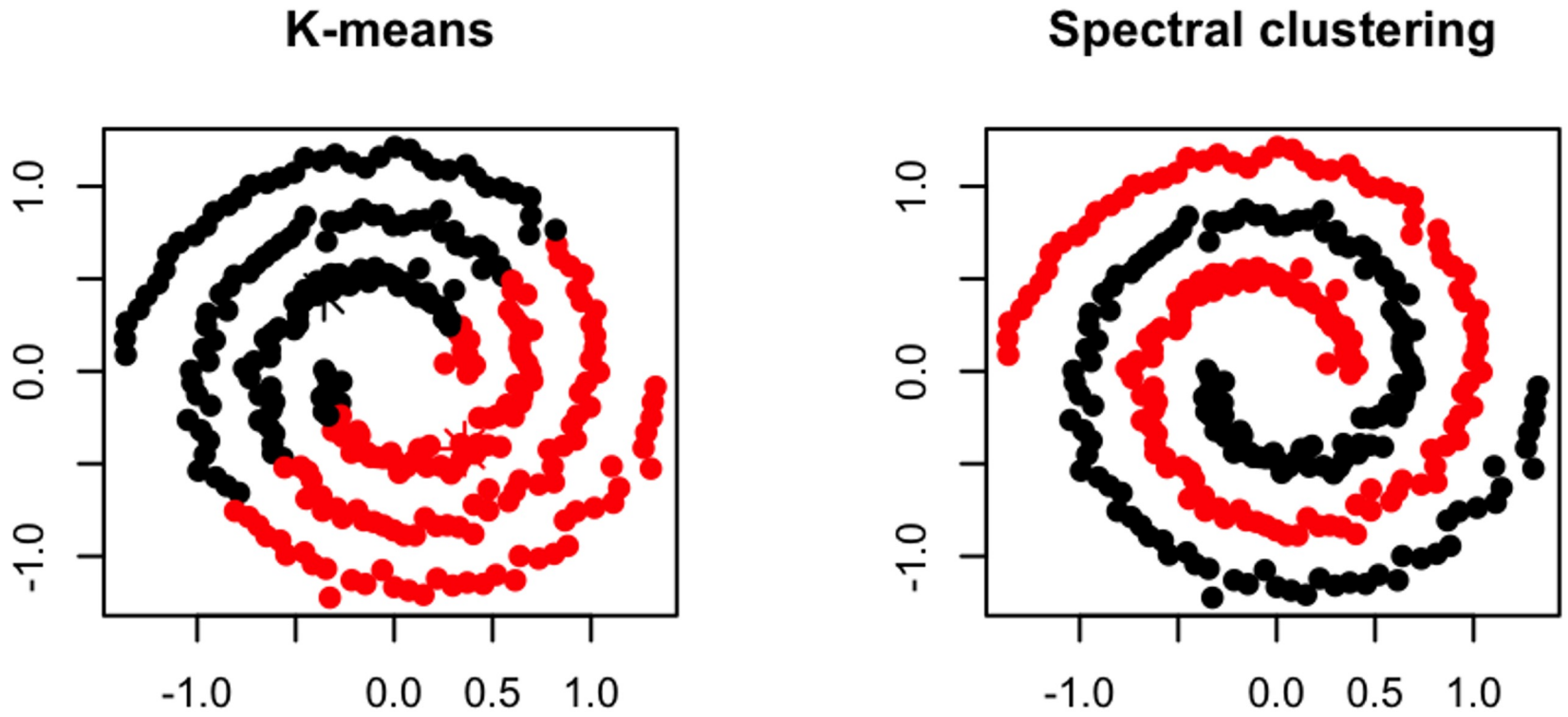
How it works:

- Choose number of clusters: k
- Set random cluster centers (“centroid”)
- For each point:
  - find closest centroid
  - assign it to that cluster
- Recompute the new centroid for each cluster
- Repeat until centroids stop moving (convergence)



Spectral Clustering

- Commonly used for networks/graphs
- Operates on pairwise sample similarity ("adjacency")



Deciding on the number of clusters

- Arbitrarily cutting the dendrogram (by eye)
- Silhouette statistic
- Dunn Index
- Connectivity

Measured in the ***clValid*** package

We will do this with ***hclust( )*** in the lab

And others…

Silhouette width

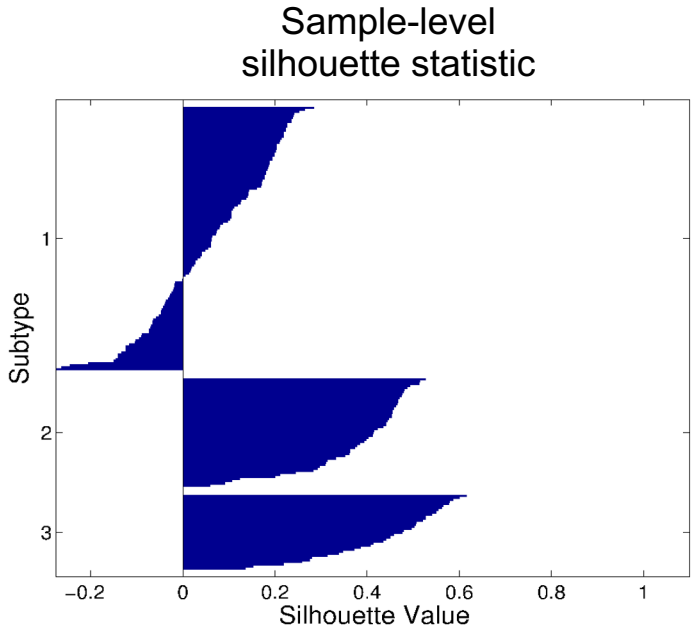
On average, how similar is a sample to its assigned cluster, compared to other clusters.

Requires identified clusters.

average distance to samples in nearest neighbouring cluster

average distance to within-cluster samples

$$S(i) = \frac{b_i - a_i}{max(b_i, a_i)}$$



Worst cluster separation

Best cluster separation

-1 0 +1

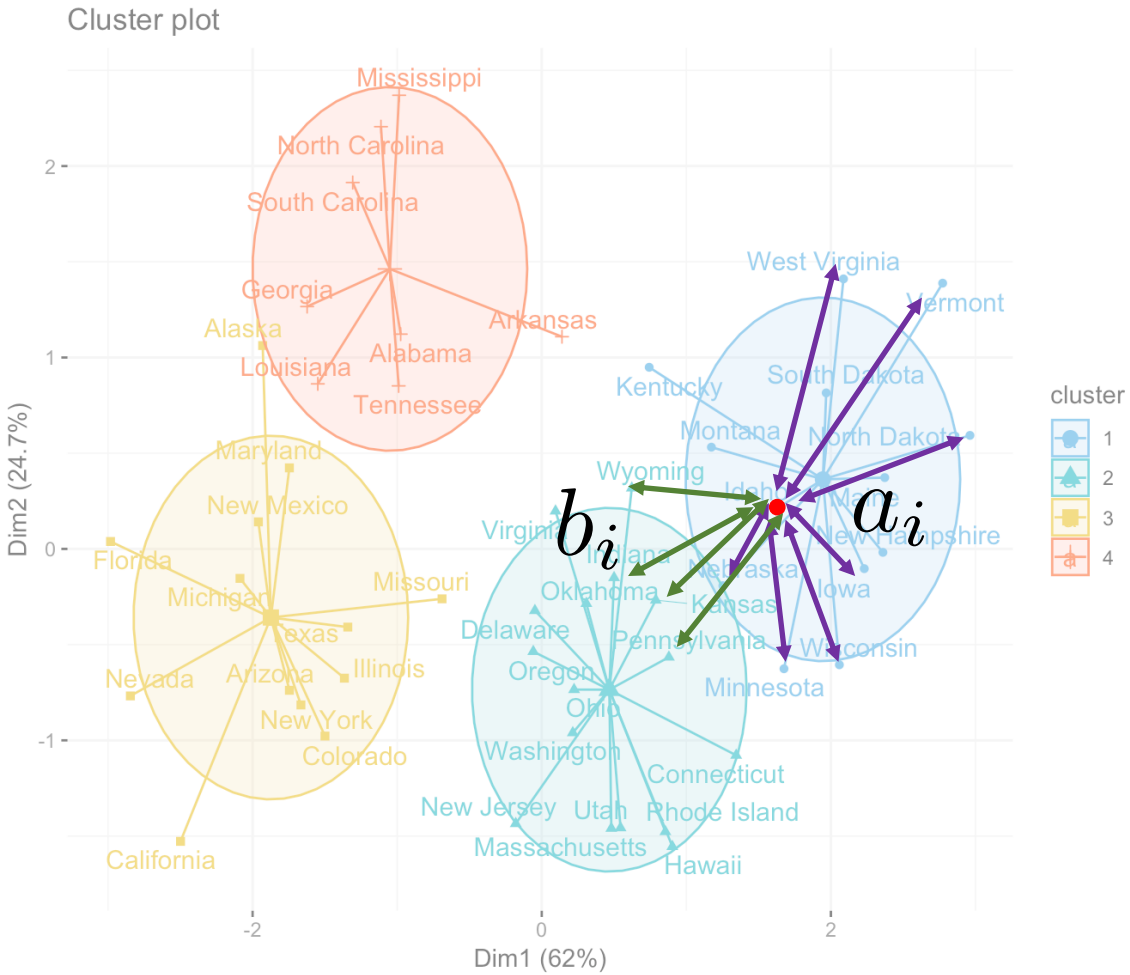


Image source: https://www.datanovia.com/en/lessons/k-means-clustering-in-r-algorith-and-practical-examples/

Dunn Index

Dunn Index

The Dunn Index is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance. It is computed as

$$D(\mathcal{C}) = \frac{\min\limits_{C_k, C_l \in \mathcal{C}, C_k \neq C_l} \left(\min\limits_{i \in C_k, j \in C_l} dist(i,j) \right)}{\max\limits_{C_m \in \mathcal{C}} diam(C_m)},$$

where $diam(C_m)$ is the maximum distance between observations in cluster C_m. The Dunn Index has a value between zero and ∞, and should be maximized.

* Similar to silhouette width. Want to maximize

clValid, an R package for cluster validation

Connectivity

Connectivity

Let N denote the total number of observations (rows) in a dataset and M denote the total number of columns, which are assumed to be numeric (e.g., a collection of samples, time points, etc.). Define $nn_{i(j)}$ as the jth nearest neighbor of observation i, and let $x_{i,nn_{i(j)}}$ be zero if i and j are in the same cluster and $1/j$ otherwise. Then, for a particular clustering partition $\mathcal{C} = \{C_1, \ldots, C_K\}$ of the N observations into K disjoint clusters, the connectivity is defined as

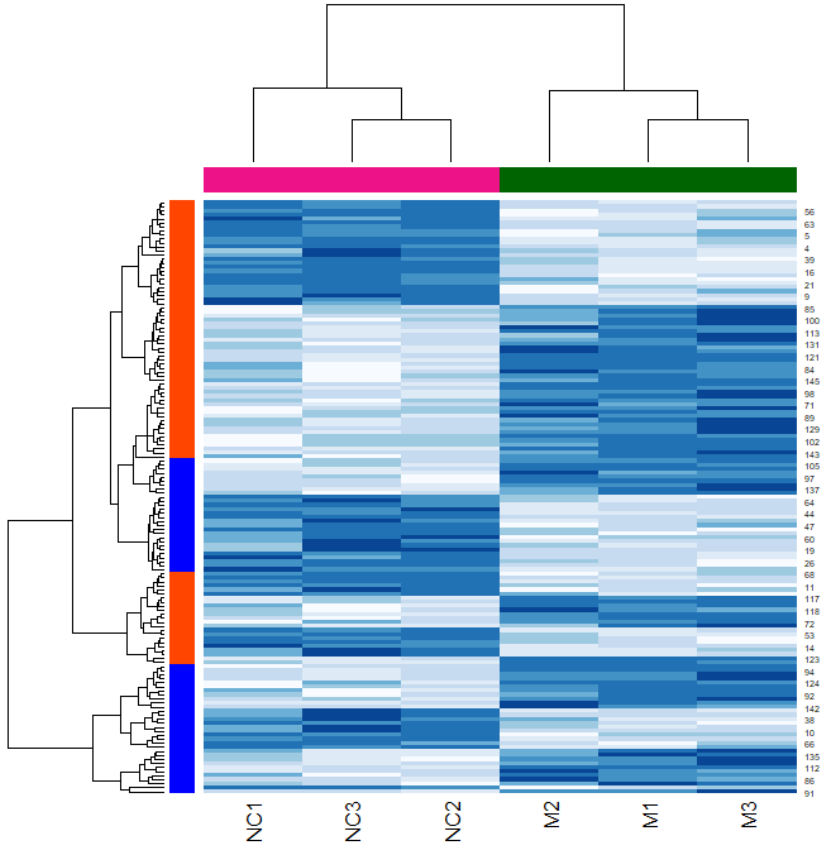
$$Conn(\mathcal{C}) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_{i(j)}} \ ,$$

where L is a parameter giving the number of nearest neighbors to use. The connectivity has a value between zero and ∞ and should be minimized.

* Counts what fraction of nearest neighbours are not in the same cluster. Note: Should be minimized.

clValid, an R package for cluster validation

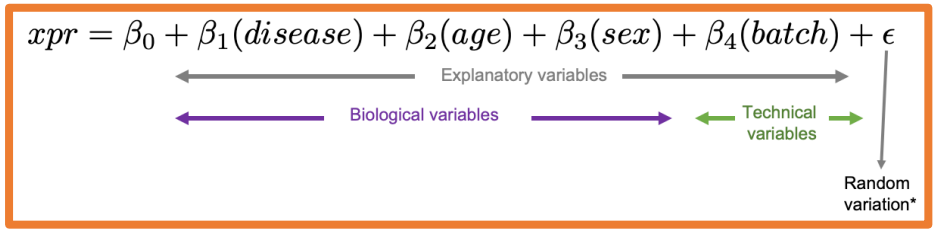
Clustering



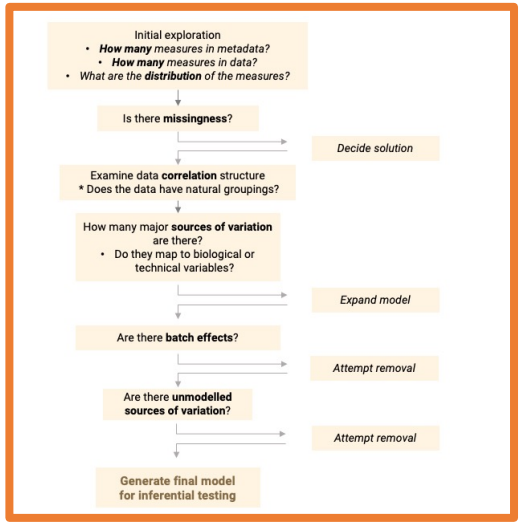
- Find groups in your data
- Requires a distance
- Multiple clustering methods exist, pick one appropriate to your application.
- Measure goodness of clustering with silhouette, Dunn score or connectivity

Let’s recap!

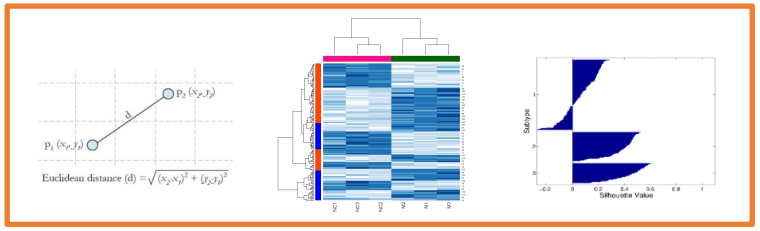
Main points:



The goals of exploratory data analysis are to identify major sources of variation and co-variation, and identify outliers / missing data



Perform exploratory data analysis in a systematic way using an approach like the one on the left.



Clustering can be used to find natural groupings in data. It requires a distance metric. Clustering can be validated with metrics.

Let’s look at how to achieve EDA and clustering using R.

We are on a Coffee Break & Networking Session

Workshop Sponsors: