Option Bio-Info/Bio-Stat

S. Granjeaud

Licence professionnelle Métiers du décisionnel et de la statistique

PARCOURS : INFORMATIQUE DÉCISIONNELLE, STATISTIQUES ET BIG DATA

LP MDS

CRCM Cancer Research Center of Marseille 50 years of research and innovation

Aix*Marseille université Socialement engagée

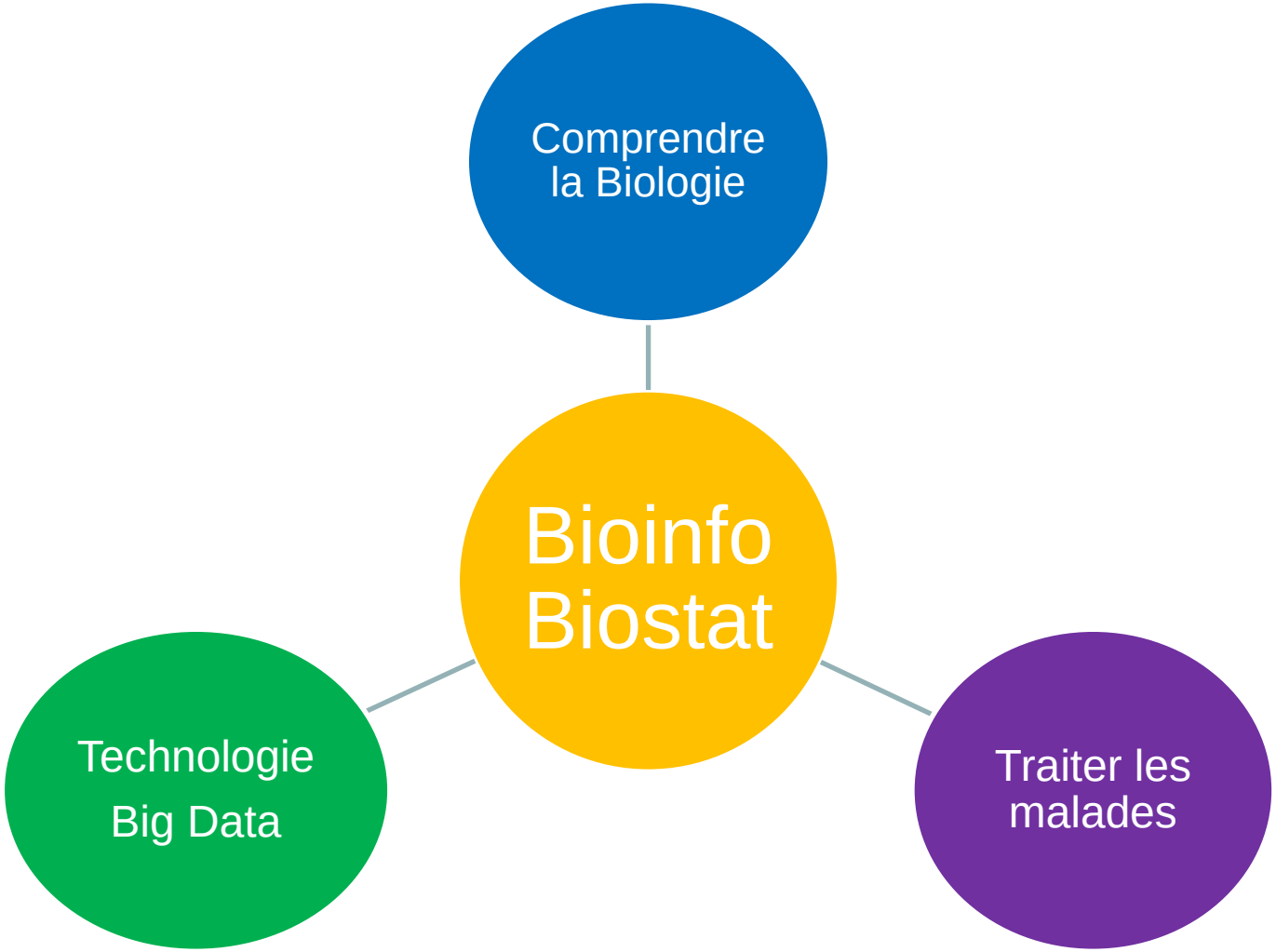
Tour de table

Qu’est-ce que vous avez fait ?

Qu’est-ce que vous voulez faire ?

INFORMATIQUE ET STATISTIQUES POUR LA BIOLOGIE ET LA MÉDECINE

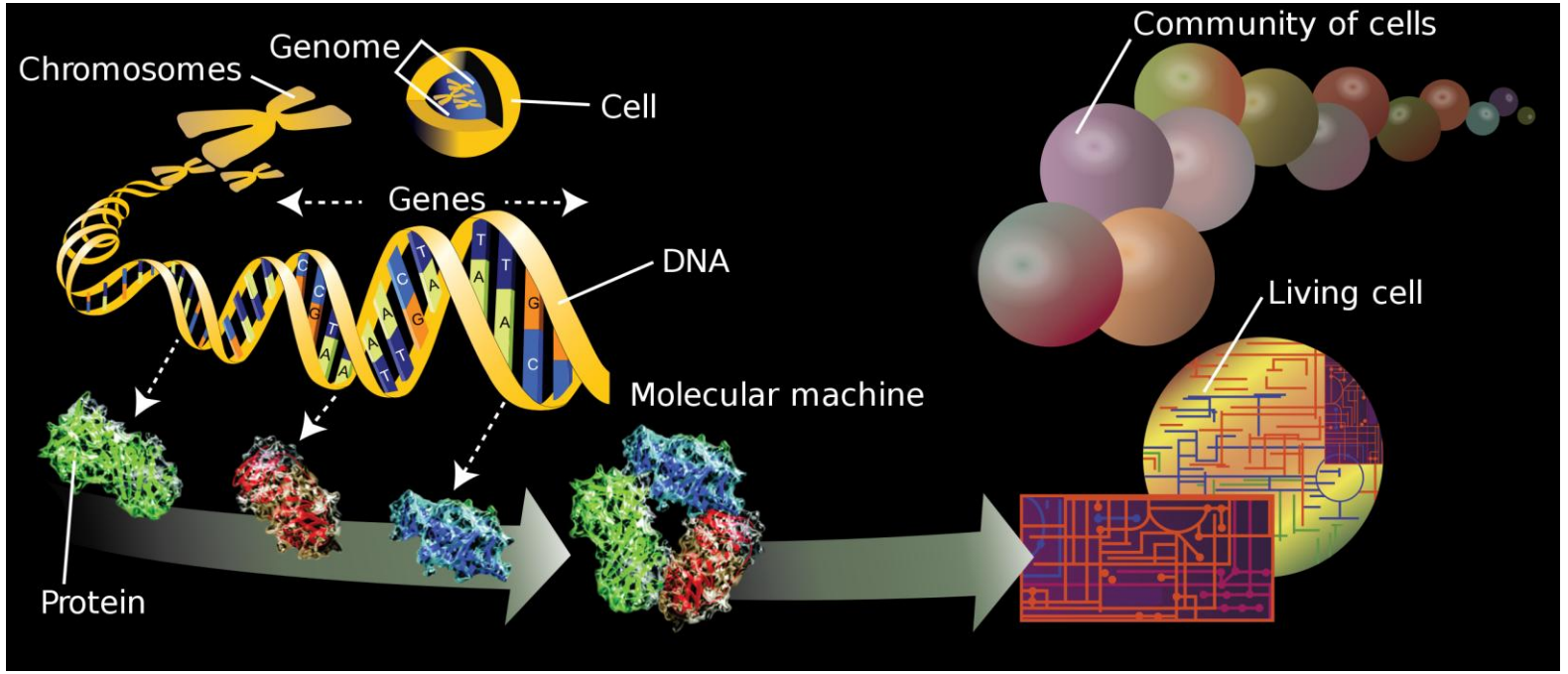
Contexte


Comprendre
la Biologie
Bioinfo
Biostat
Technologie
Big Data
Traiter les
malades

Bioinformatique

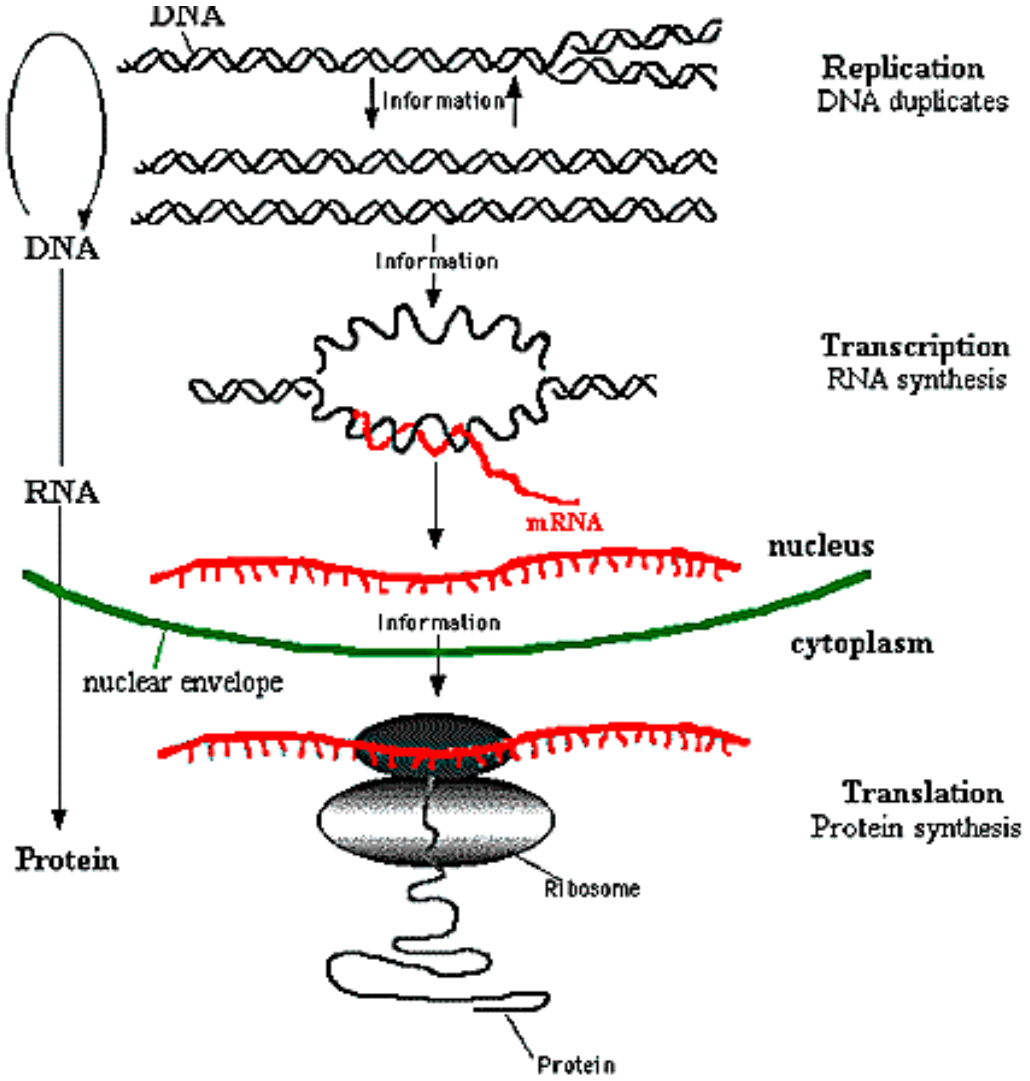
- BIO logie
- INFOR mation
- traitement auto MATIQUE

Biology



Wikipedia omics

Different Kinds of “Omes”

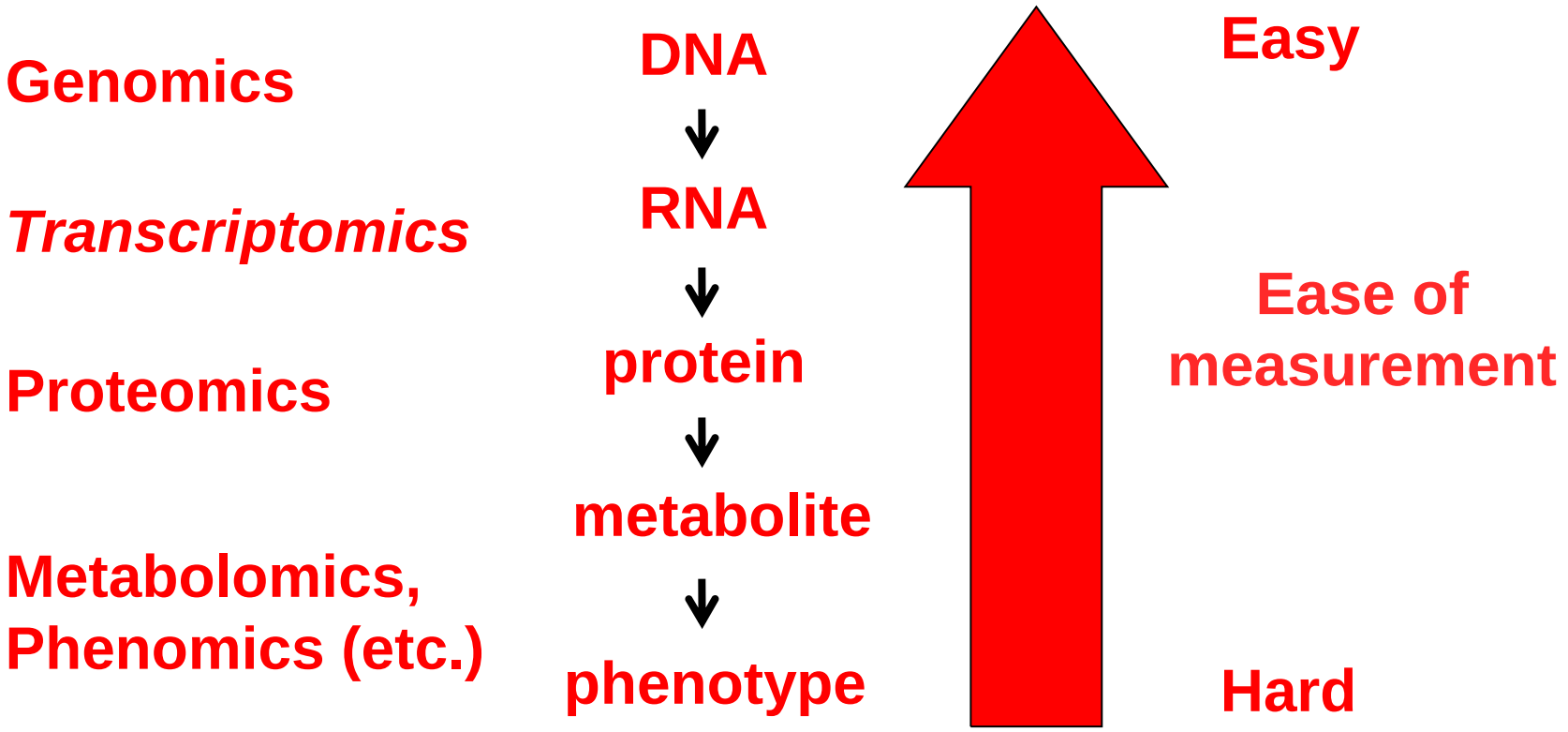


Genome

Transcriptome

Proteome

High Throughput Measurement



Toutes ces analyses sont complémentaires !

Bioinformatique & Analyse de données

- Emergence des études à l'échelle du génome complet (haut flux).
- Médecine personnalisée :
  - Altérations génomiques
  - Modulation du niveau transcriptomique
  - Dysfonctionnement protéique

- Besoin d'outils informatiques et de méthodes statistiques d'analyse et de décision
- Identifier les acteurs d'une perturbation et de ses conséquences
- Déterminer l'efficacité et les effets d'un traitement

Objectifs

- Diagnostic, détection précoce
- Pronostic de la réponse à un traitement, de survie
- Déterminer la thérapie optimale

- Mutations dans le génome (insertions, deletions)

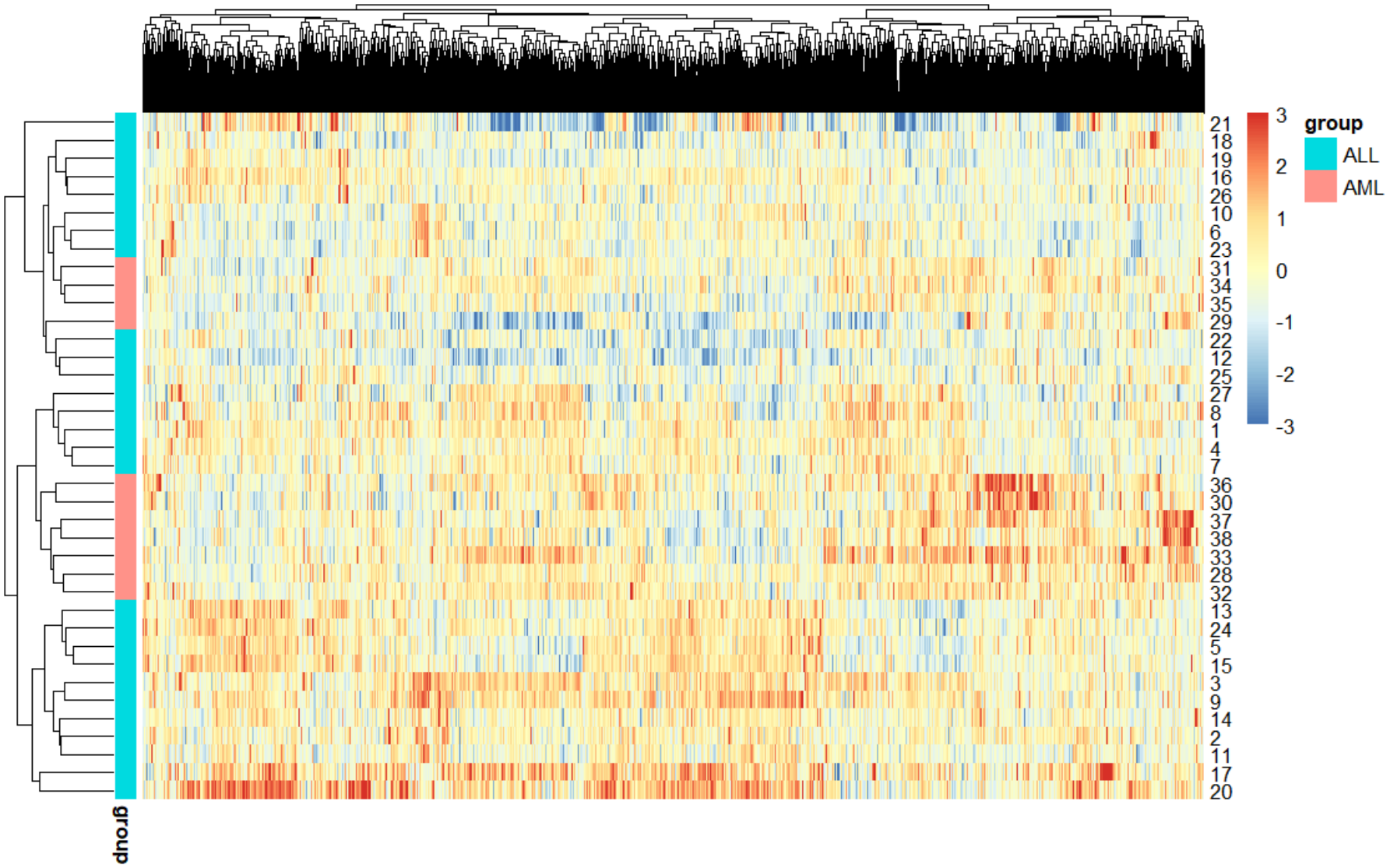
- Profil génomique, transcriptomique, protéique… des tumeurs

- Identifier les acteurs et comprendre les mécanismes

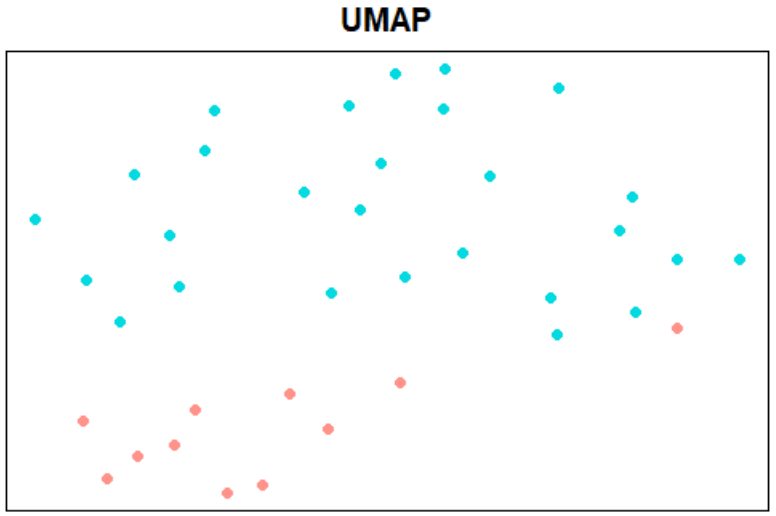
AML vs ALL, Golub et al., 1999

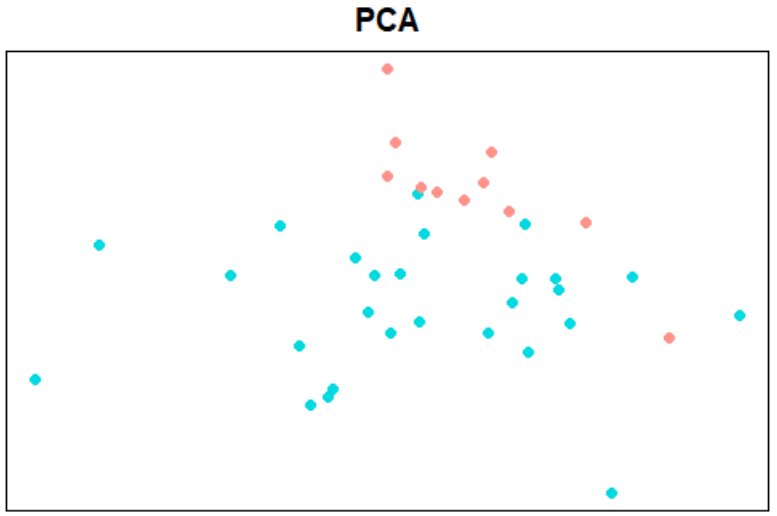
L’analyse de Golub a marqué l’avènement de la génomique clinique. En utilisant des puces à ADN pour mesurer l'expression de milliers de gènes, l'étude a démontré qu'il est possible de **classer des cancers (AML vs ALL) sans diagnostic humain préalable**. Grâce au clustering non supervisé (découverte de classes) et à un algorithme de vote pondéré (prédiction de classes), les chercheurs ont identifié une signature moléculaire de 50 gènes discriminants. Cette approche a prouvé que les pathologies peuvent être définies par leur profil transcriptomique, posant ainsi les bases de la **médecine personnalisée** et du **diagnostic moléculaire** moderne.

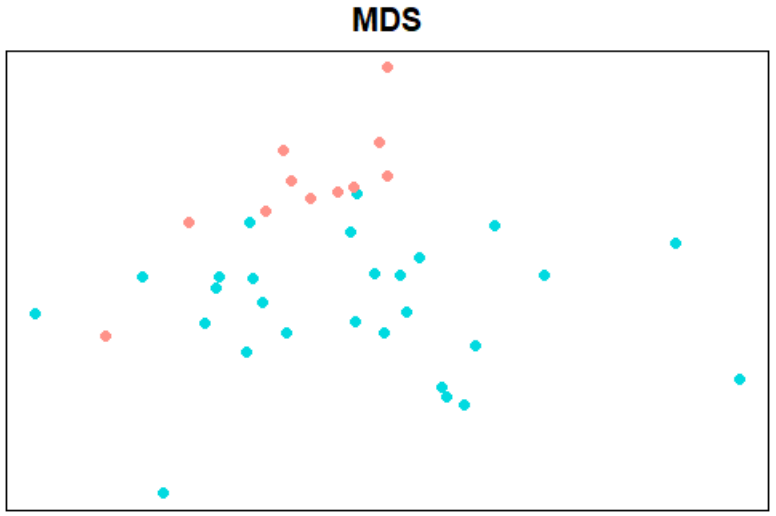
Heatmap & Clustering



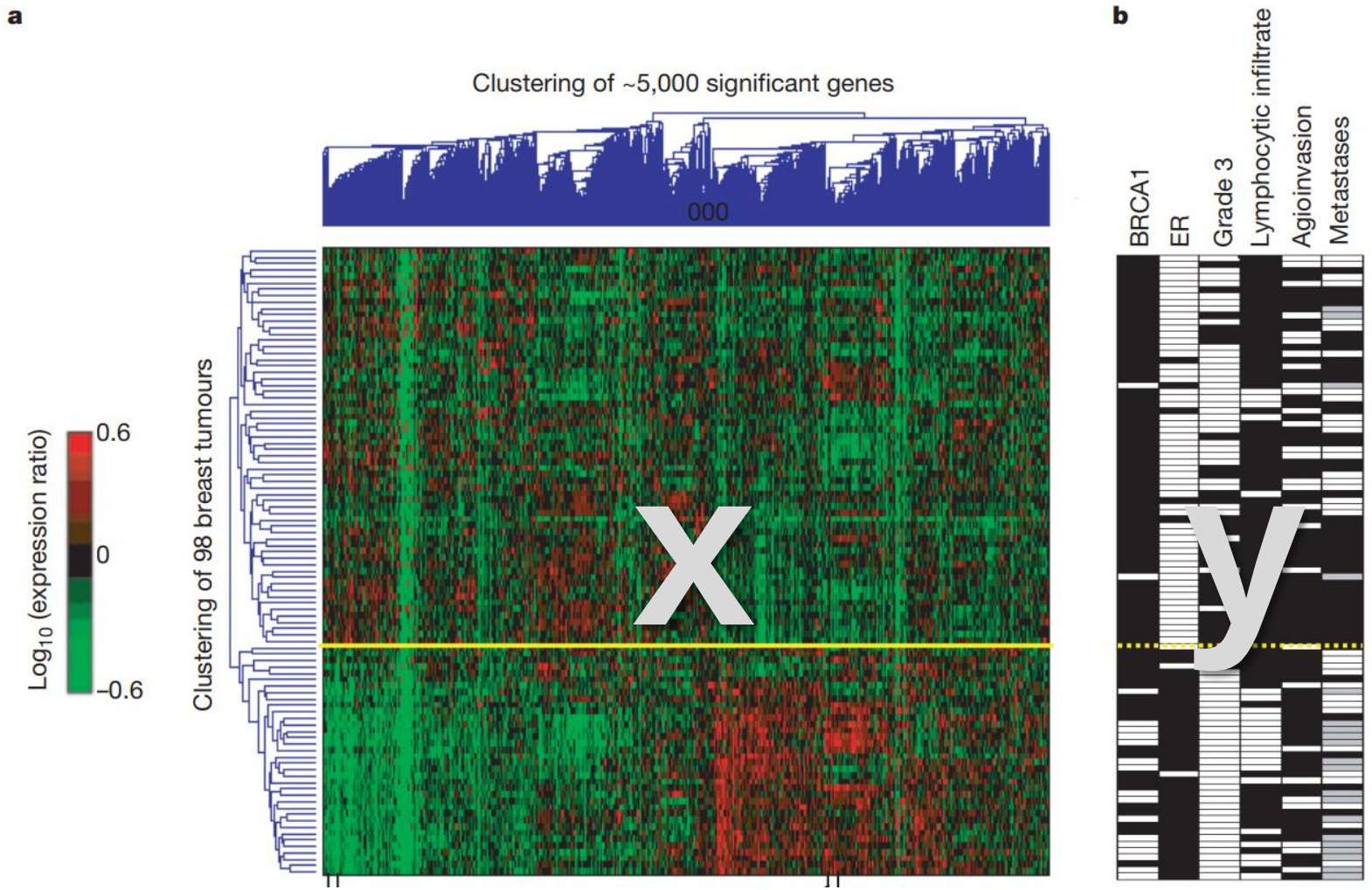
Réduction de dimensions







From matrix to image



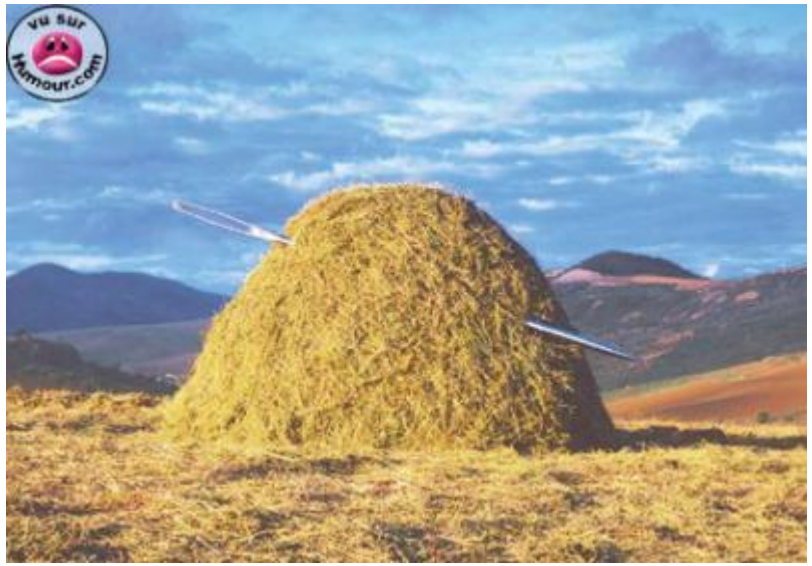
- supervised = explain y using x
  $y = f(x) \Leftrightarrow y \sim x$
- unsupervised = show patterns in x and overlay y
- statistics
  - rows = individuals
  - columns = variables

van't Veer et al., Nature, 2002

Objectifs

- Objectif : identifier les variables (gènes…) ou une combinaison liées à une caractéristique des individus ou un dysfonctionnement

- Grande échelle
  = variables innombrables
- Biologie/Médecine
  = peu d'échantillons
  car précieux, rares,
  longs à préparer
- Diagnostic/Pronostic
  = besoin de statistiques

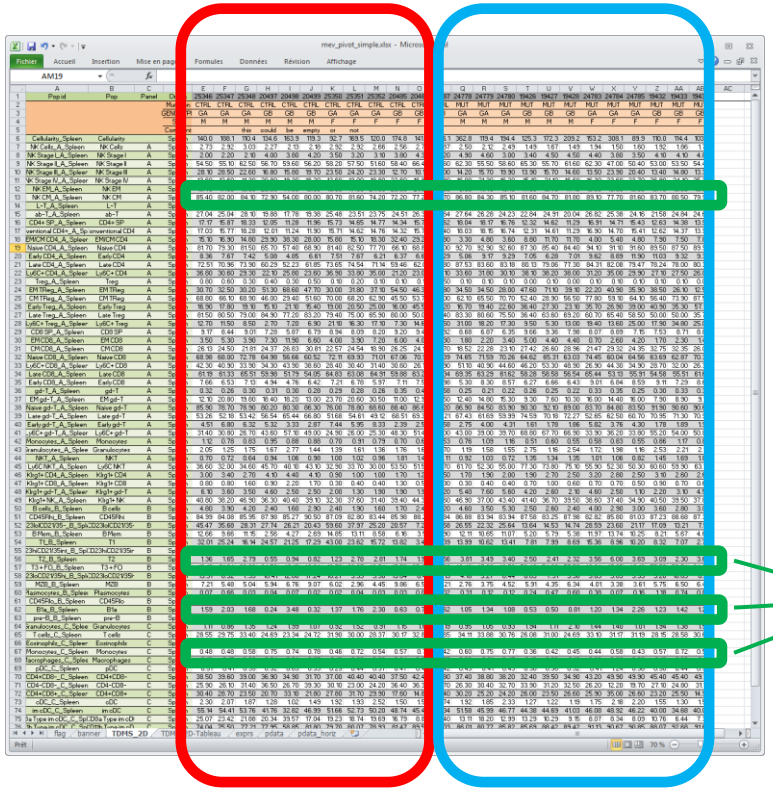


Grandes dimensions

Biologie/Médecine
= peu d'échantillons
car précieux, rares,
longs à préparer

colonne = individu = échantillon
ligne = variable = gène, protéine…

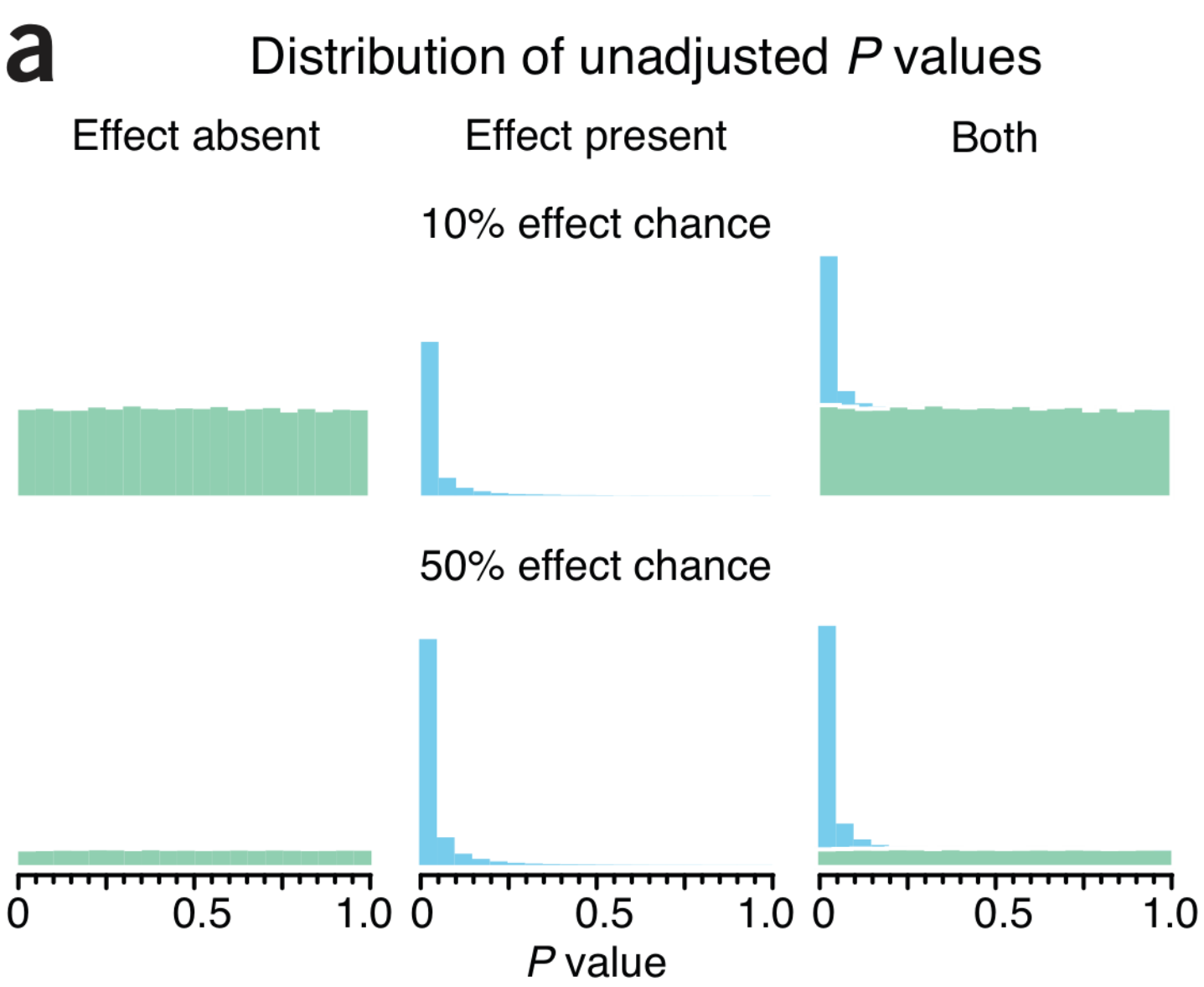
Omics = Grande échelle
= variables innombrables

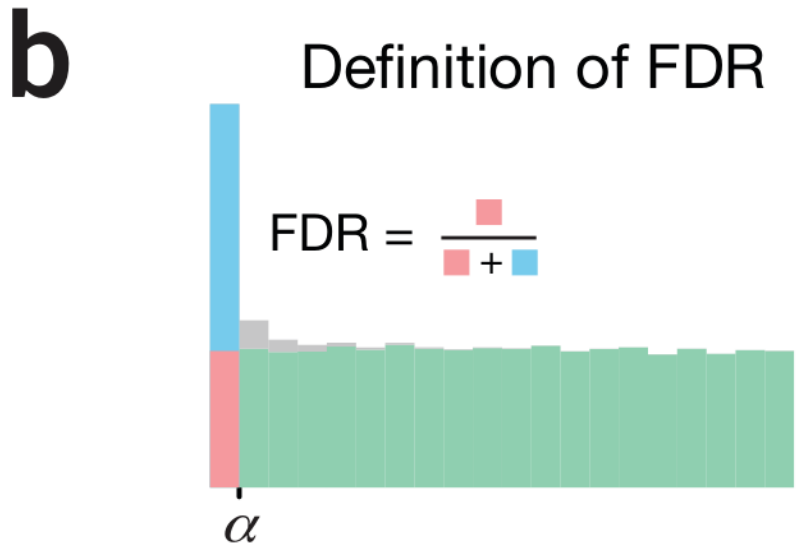


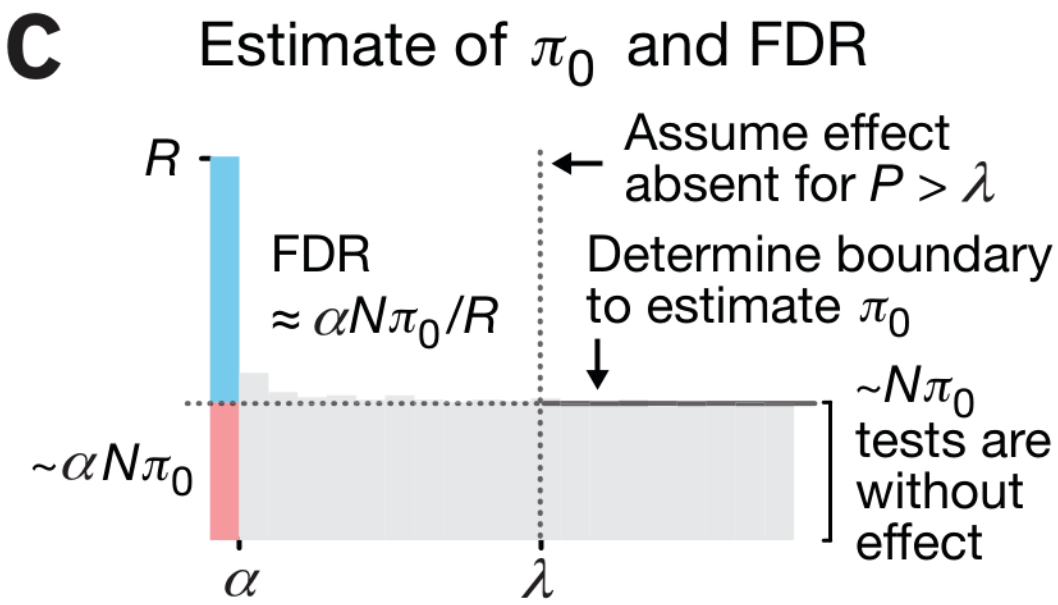
? Uni-variée

? Multi-variée

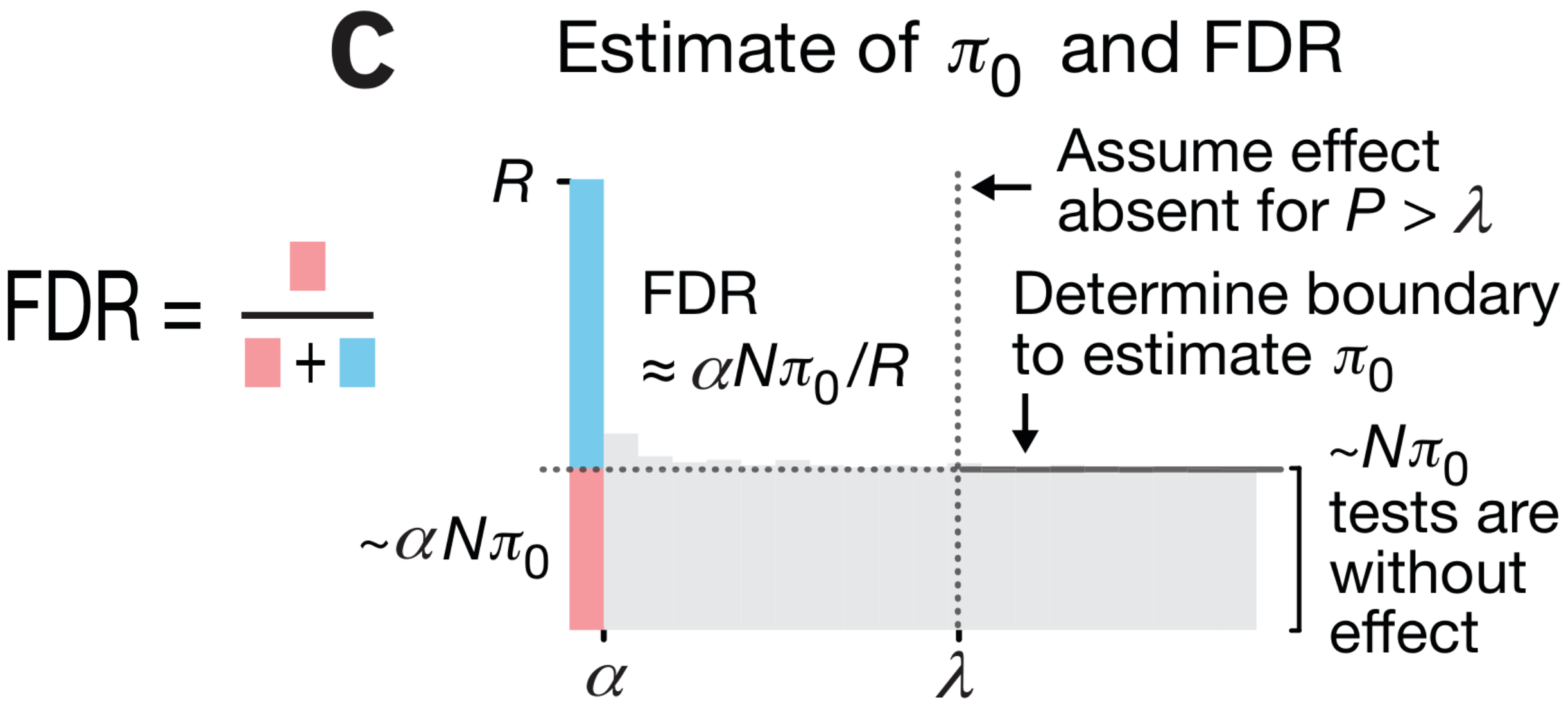
FDR : Ratio de Fausses Découvertes







FDR : Ratio de Fausses Découvertes



Bioinfo/Biostat : où ?

- Extraction de l'information des instruments
  Prétraitement de la mesure
  - Spécifique à l'instrument, à sa technologie

- Analyse numérique pour la sélection et la décision

- Aide à l'interprétation
  - Apport de la connaissance acquise

Analysis process

This is what you learn in school & textbooks

```
# (ideal) data analysis process
raw_data = GET(data)
proc_data = PROCESS(raw_data)
SUMMARY(proc_data)
PLOT(proc_data)
model = FIT_MODEL(proc_data)
prediction = PREDICT(model)
PRINT(prediction)
> "Woo-hoo! validated model =)"
```

This is what you learn in the real world

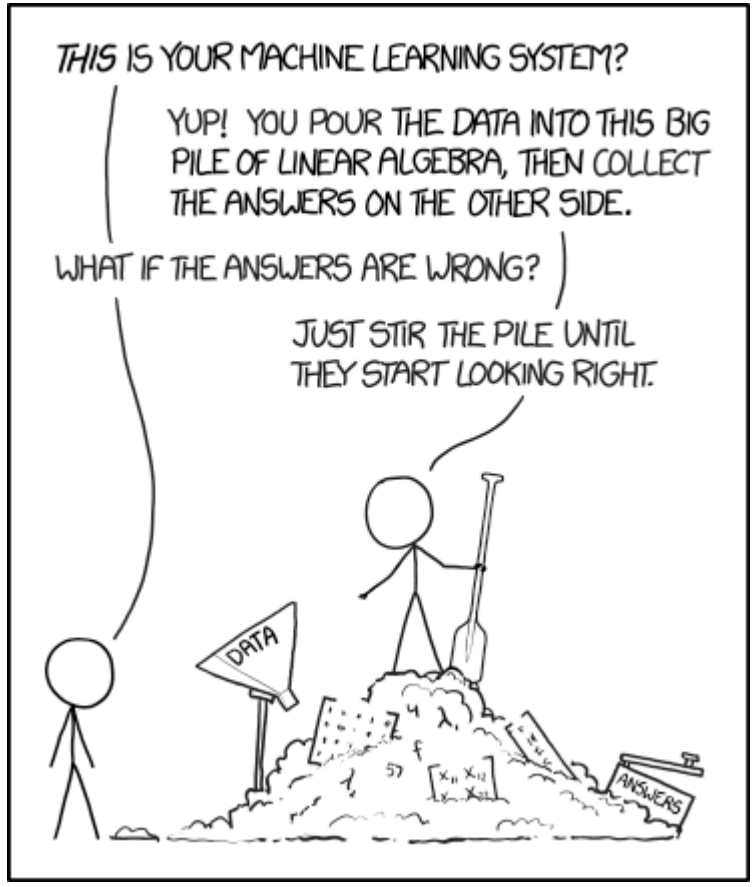
```
# (real) data analysis process
raw_data = GET(data)
clean_data = CLEAN(data)
proc_data = PROCESS(clean_data)
while (QUALITY(proc_data) != "good") {
   clean_data = CLEAN(proc_data)
   proc_data = PROCESS(clean_data)
   # while loop may run indefinitely
}
SUMMARY(proc_data)
PLOT(proc_data)
model = FIT_MODEL(proc_data)
prediction = PREDICT(model)
PRINT(prediction)
> "Ooops! model sucks =("
```

https://www.gastonsanchez.com/Data-analysis-process/

Garbage In = Garbage Out

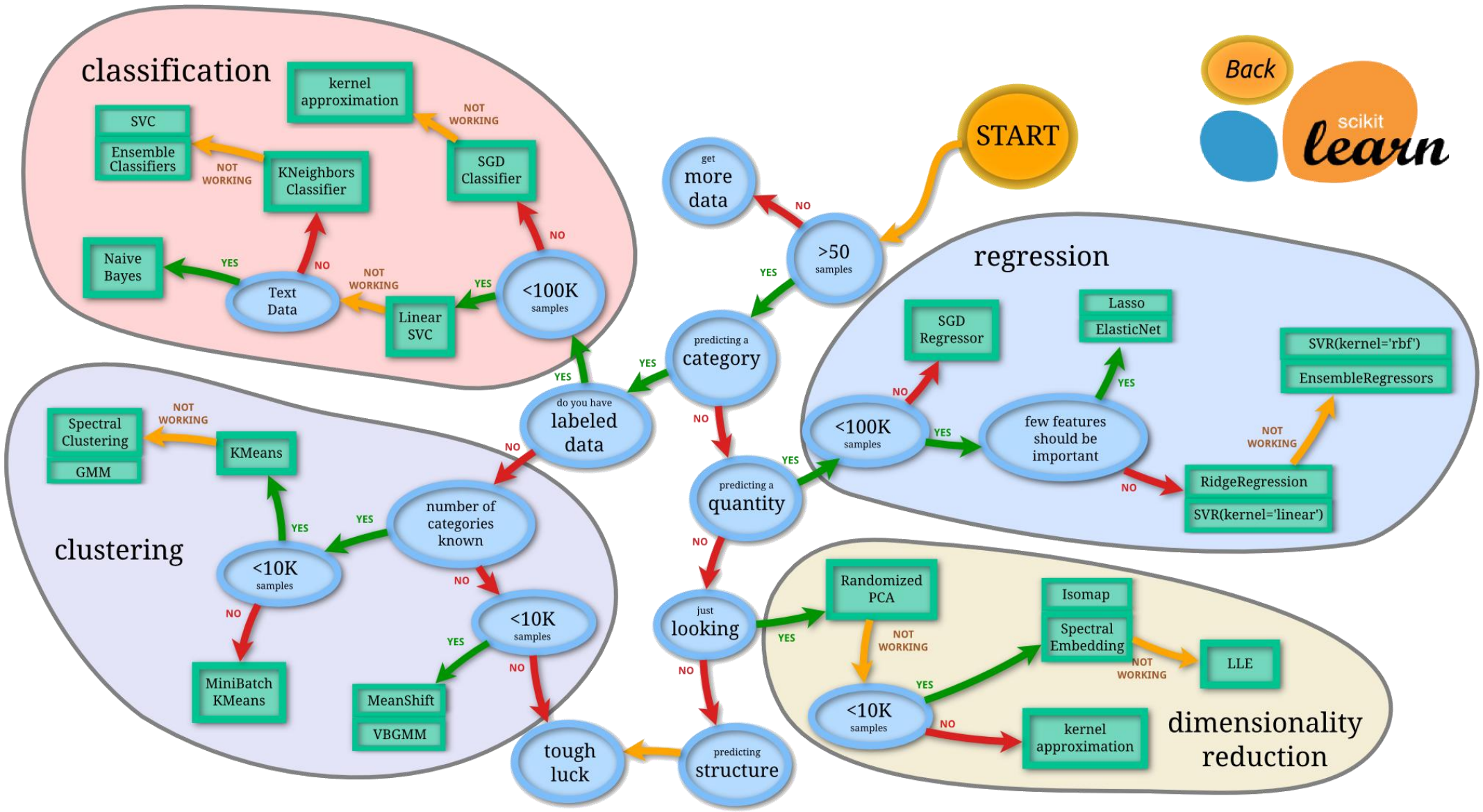
- Quality control is crucial!
- Important for manual analysis, but even more for automated analysis

"No Data Analysis Technique Can Make Good Data out of Bad Data" Howard Shapiro.



https://xkcd.com/1838/

Méthodes d'analyse



Analyse de données

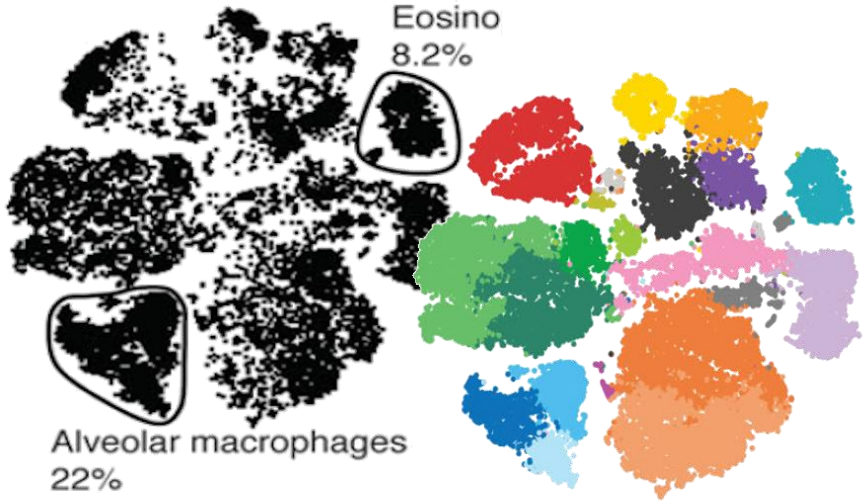
- ***Méthodes Supervisées***
  - Découverte de classes
  - Tests statistiques "classiques"
  - Utilisation des expressions corrélées
  - Score discriminant
  - Réseaux de neurones

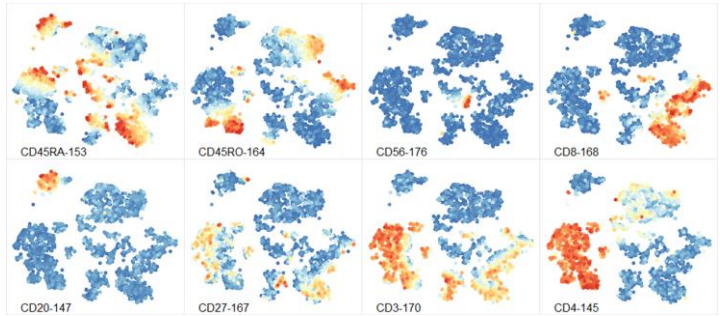
  - Valider les modèles…

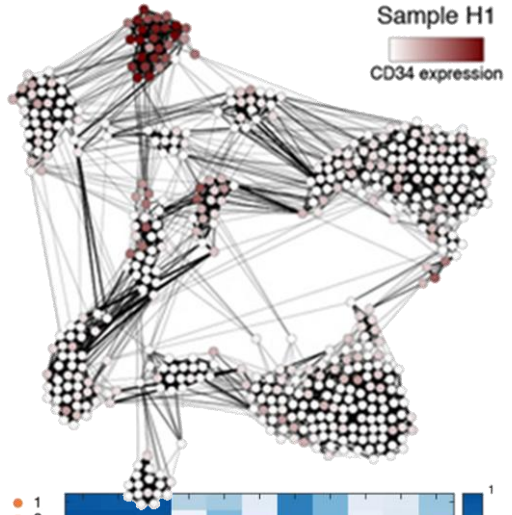
- ***Méthodes Non Supervisées***
  - Exploration des données
  - Grouper
    - Classifications hiérarchiques
    - Nuées dynamiques (K-means)
  - Réduire les dimensions
    - Analyse en composantes principales
    - tSNE, UMAP
  - Grouper et Réduire
    - Cartes de Kohonen (SOM)

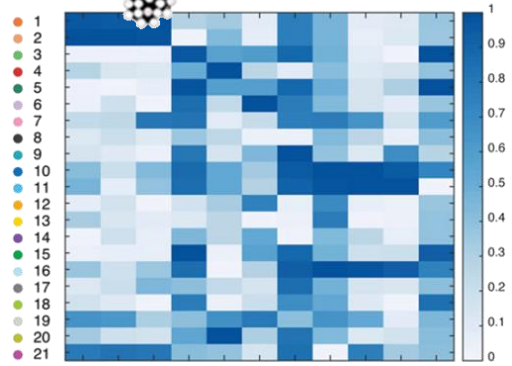
VISUALISATION DE DONNÉES

1 Picture vs 1000 Words



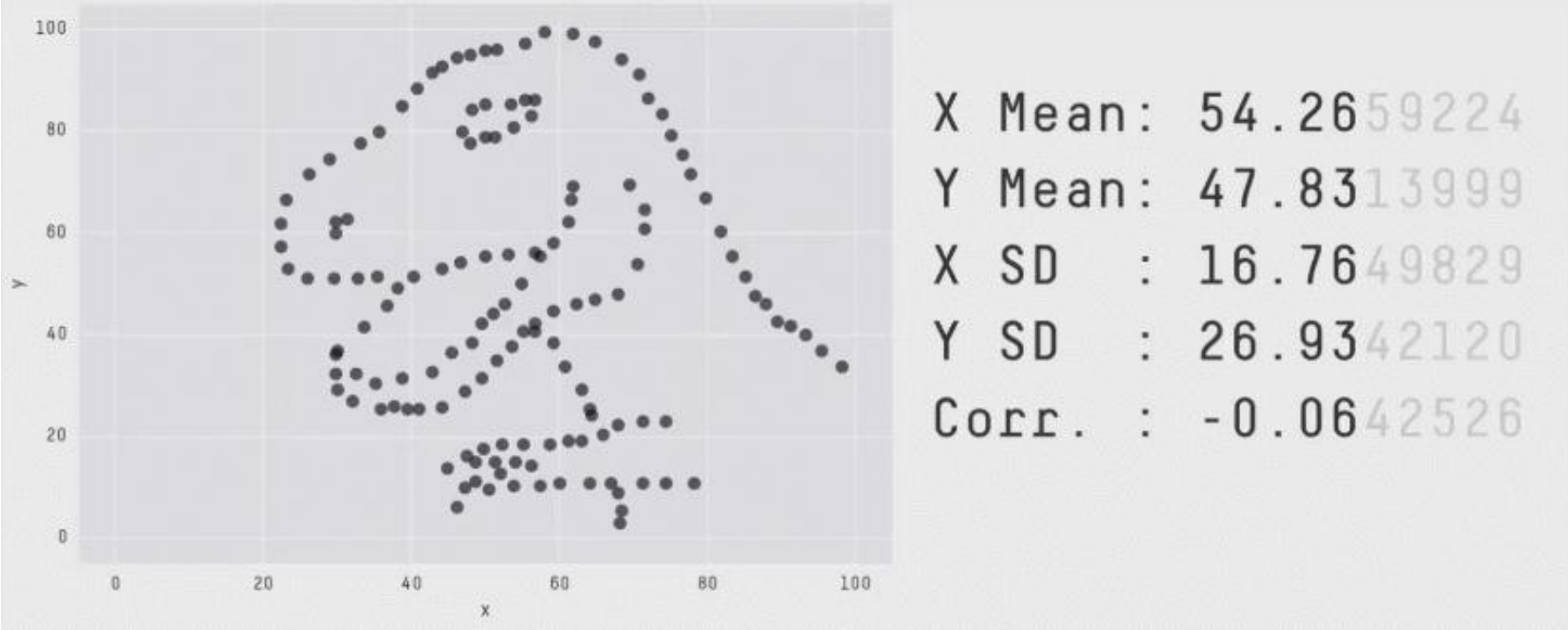






Most illustrations from Mair et al 2016 and Kimball et al 2017

Datasaurus



"... make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding."
F.J. Anscombe, 1973

https://www.research.autodesk.com/publications/same-stats-different-graphs/

Visualisation

- Always make intelligible figures
- Put axes, label the axes, set units
- Set a legend of symbols, colors, line types…
- Define a title

Visualisation, the basics

- Screen
- X, Y coord.
- Color
- Shape
- Size

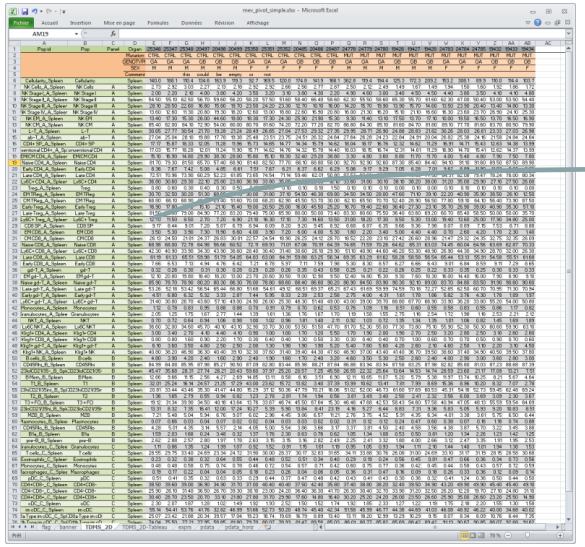
- Link the points => graph, tree

What is a dot plot?

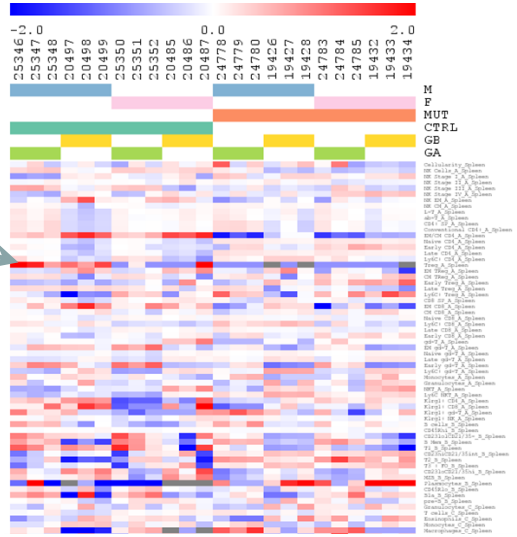
- What are those points on the scatter plot, graph?
  - Single cells or centers of cell group
- What is the meaning of the color
  - Group membership or expression level of a marker
  - Raw or transformed expression level
- Is the distance on the screen meaningful?
- How to evaluate the distance between dots?

From matrix to image

Heatmap



Scale
Convert



What is on a heatmap?

- What are those rectangles?
  - Columns = ? Rows = ?
  - Cell groups x markers or Cell groups x patients
- What is the meaning of the color?
  - Expression level of a marker or percentages
  - Raw or transformed value
  - What is the scaling?
- Criterion for arranging the rows/columns?
  - User defined or hierarchical clustering
  - How to evaluate the distance?