# Data integration in cancer research

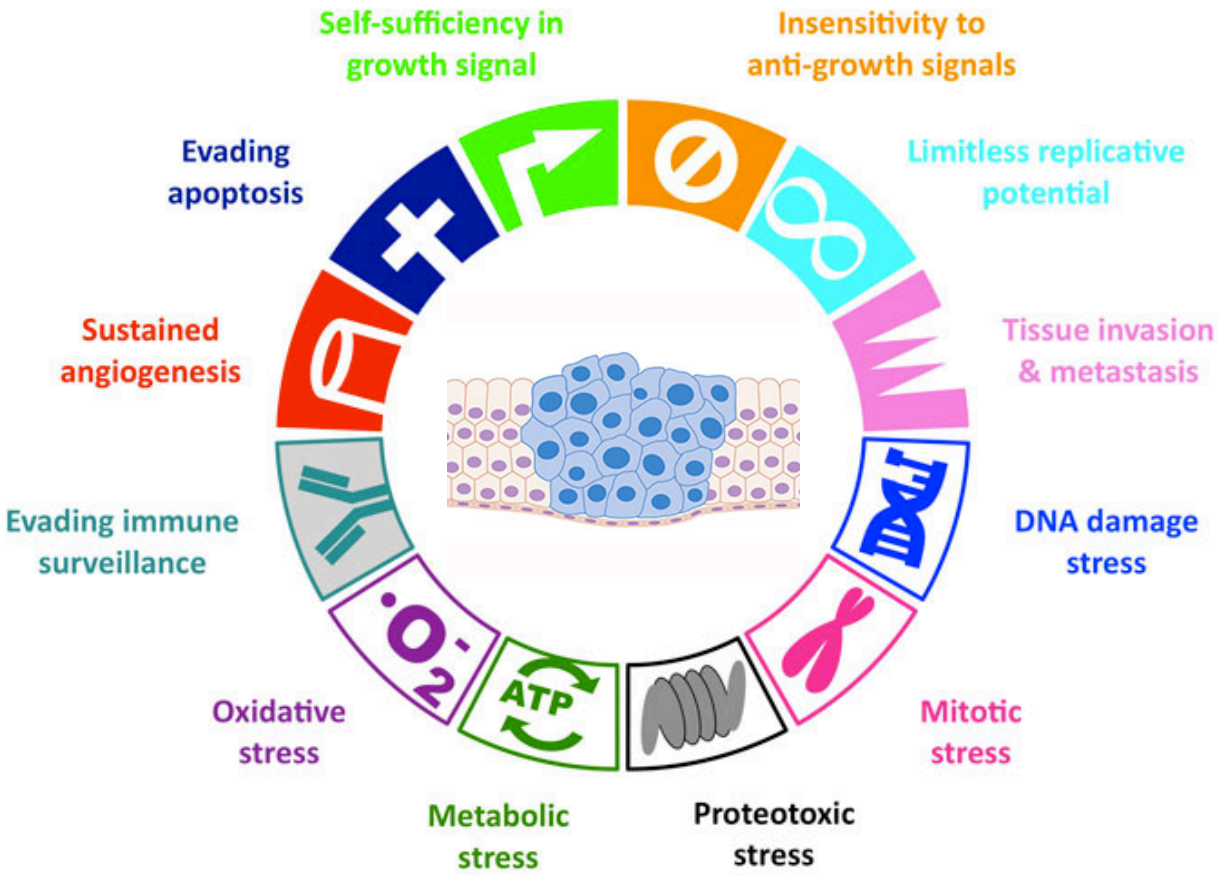
An overview of the existing approaches

**Laura Cantini**
Computational Systems Biology Team
IBENS, Paris

26-02-2019

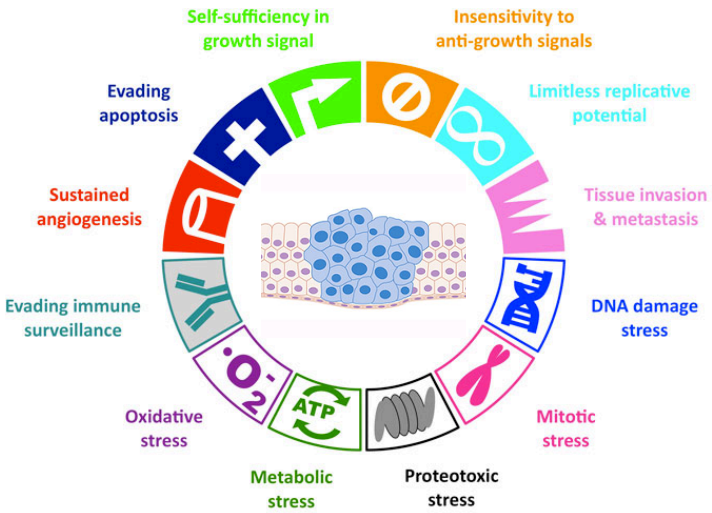
DU-Bii Integrative bioinformatics

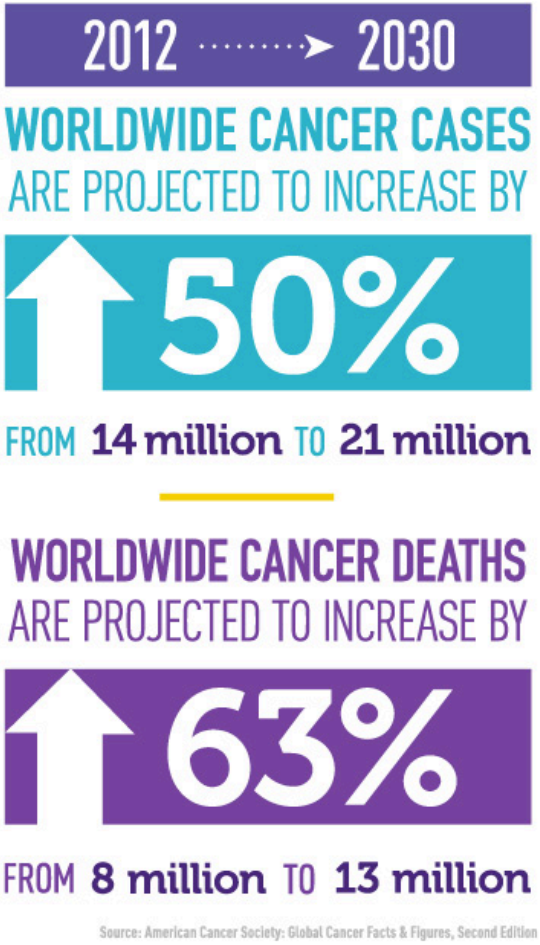
# Cancer Hallmarks




Hanahan, Douglas, and Robert A. Weinberg. "Hallmarks of cancer: the next generation." cell 144.5 (2011): 646-674.

# Cancer Hallmarks



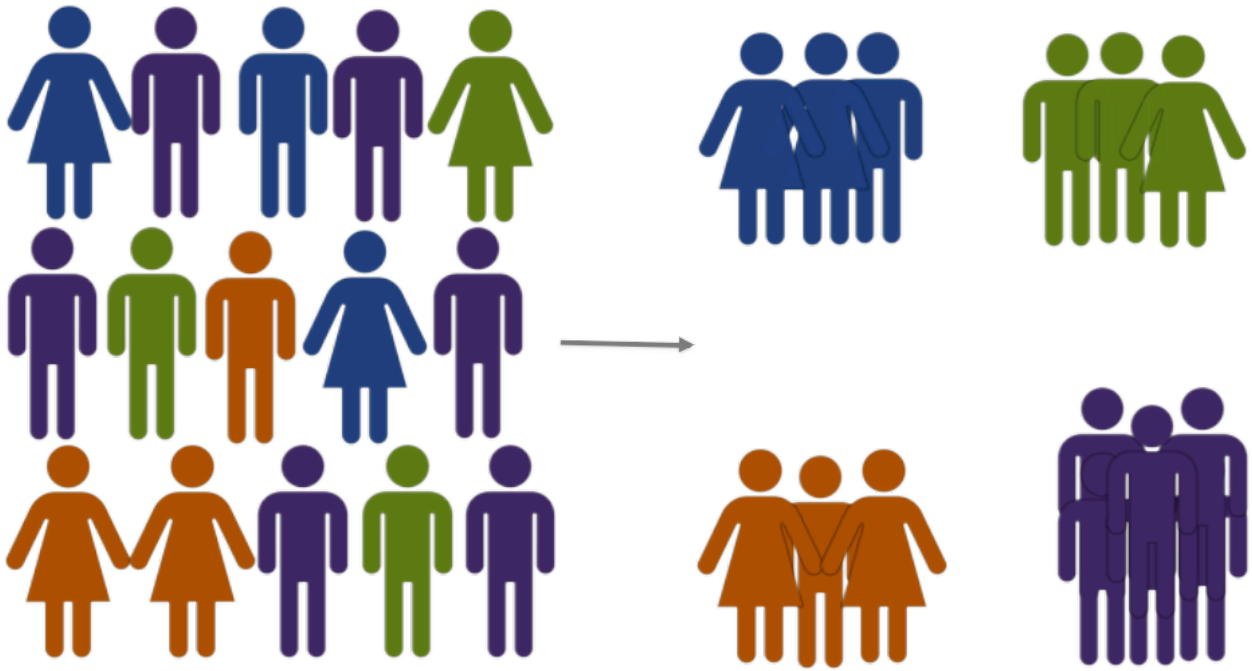

Hanahan, Douglas, and Robert A. Weinberg. "Hallmarks of cancer: the next generation." cell 144.5 (2011): 646-674.

NIH
https://www.cancer.gov/

# Cancer precision medicine

Patients suffering from the same cancer can present:

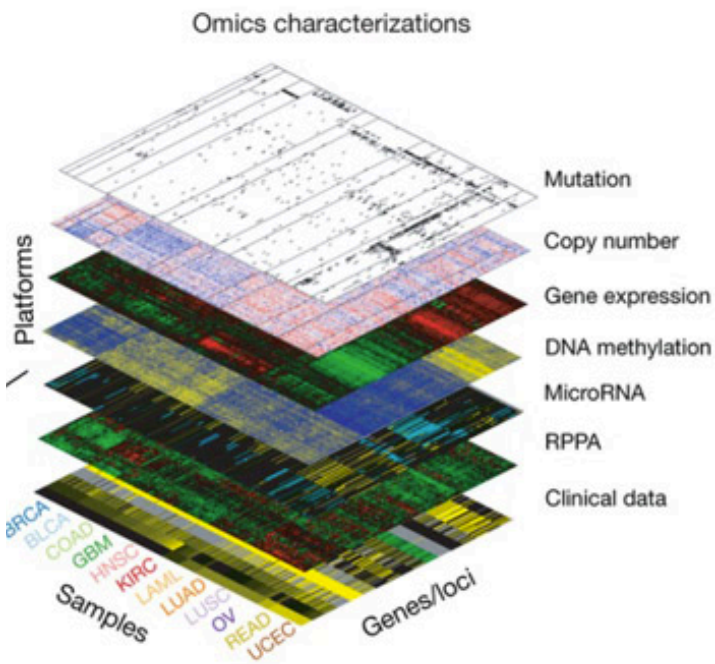
- Different prognosis
- Different response to the same treatment



→ Precision medicine is needed

# Cancer « omics » data

## The Cancer Genome Atlas (TCGA)



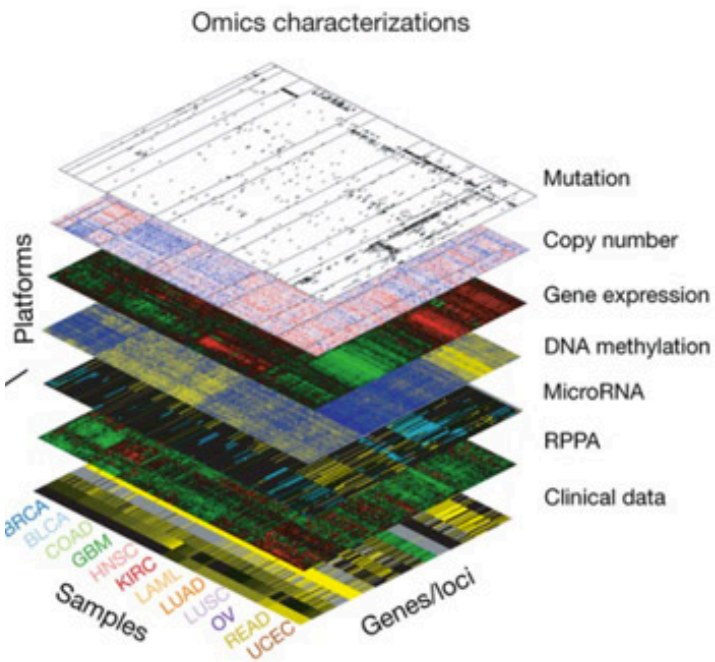

From single omics to **Multi-omics**

TCGA is the largest collection of multi-omics data:

- 10.000 cancer patients
- 33 cancer types
- 6 omics, plus clinical data

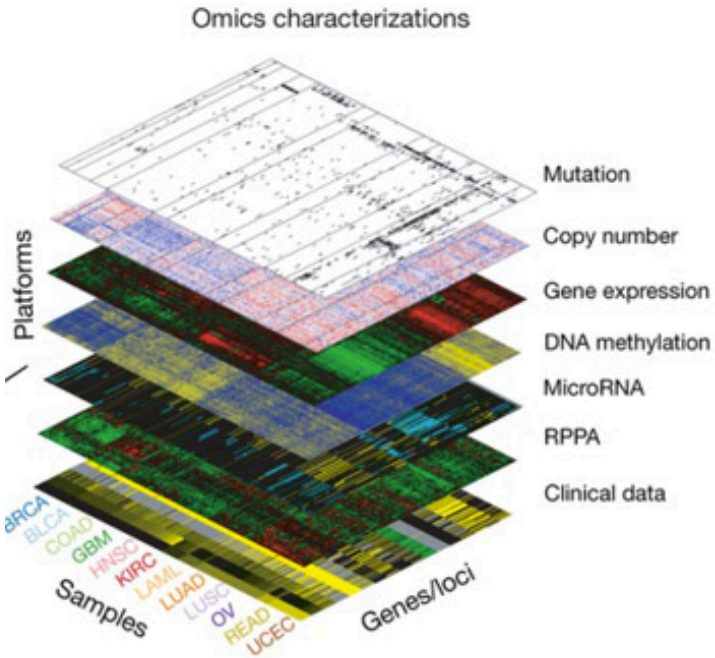
The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi: 10.1038/ng.2764

# Multi-omics integration in cancer



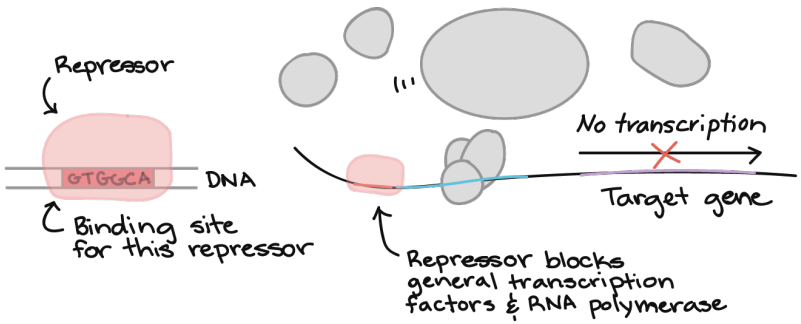

The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi: 10.1038/ng.2764

# Multi-omics integration in cancer



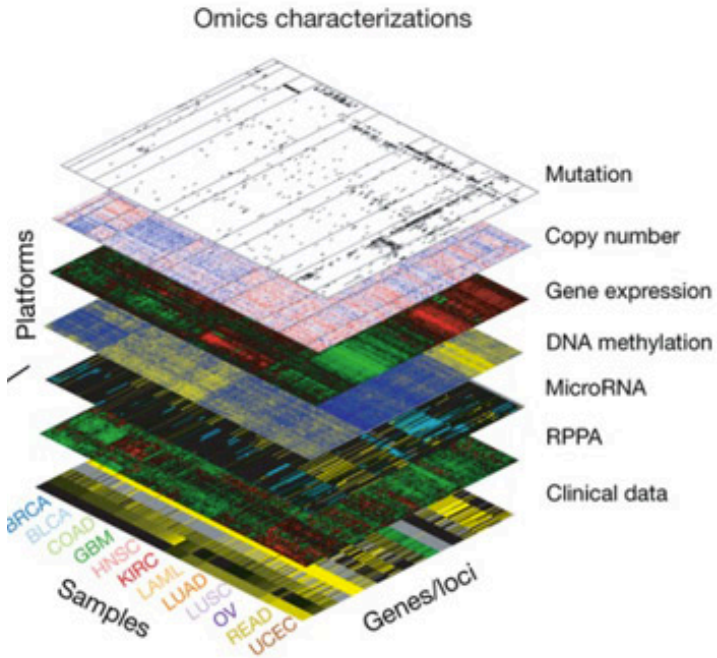

RPPA → Gene expression

(Transcription Factors)



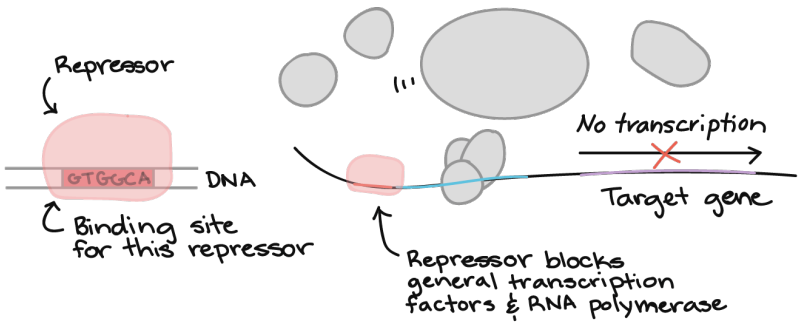

The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi: 10.1038/ng.2764

# Multi-omics integration in cancer



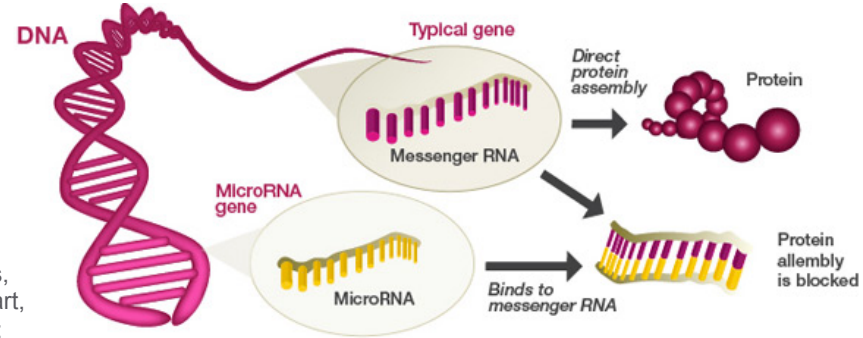

RPPA → Gene expression

(Transcription Factors)



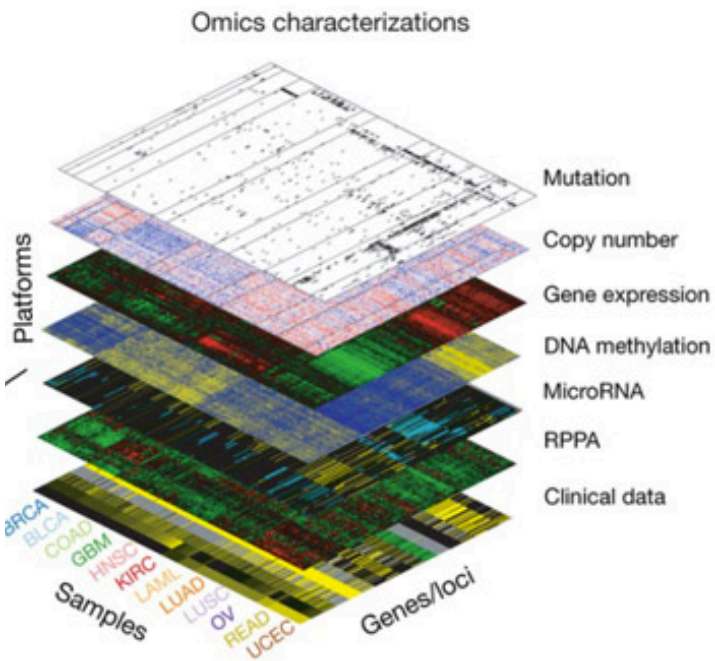

miRNA → RPPA

(microRNAs)



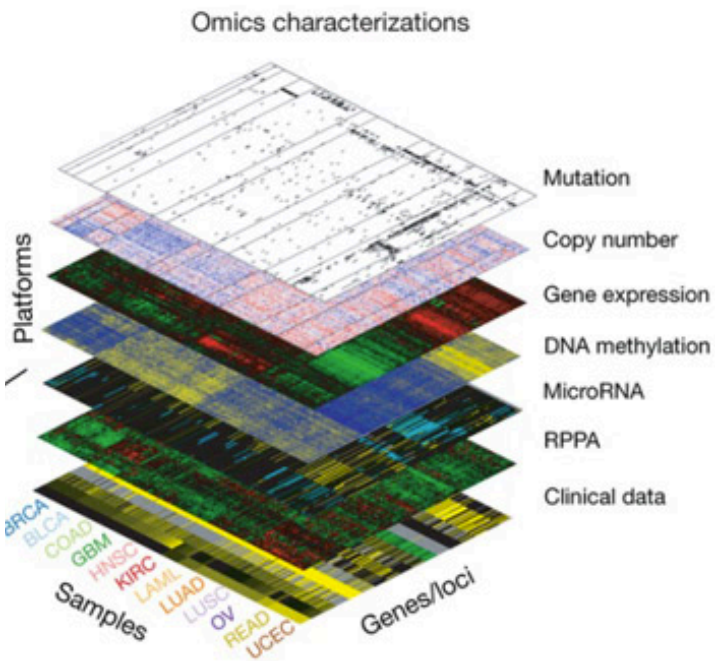

The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi: 10.1038/ng.2764

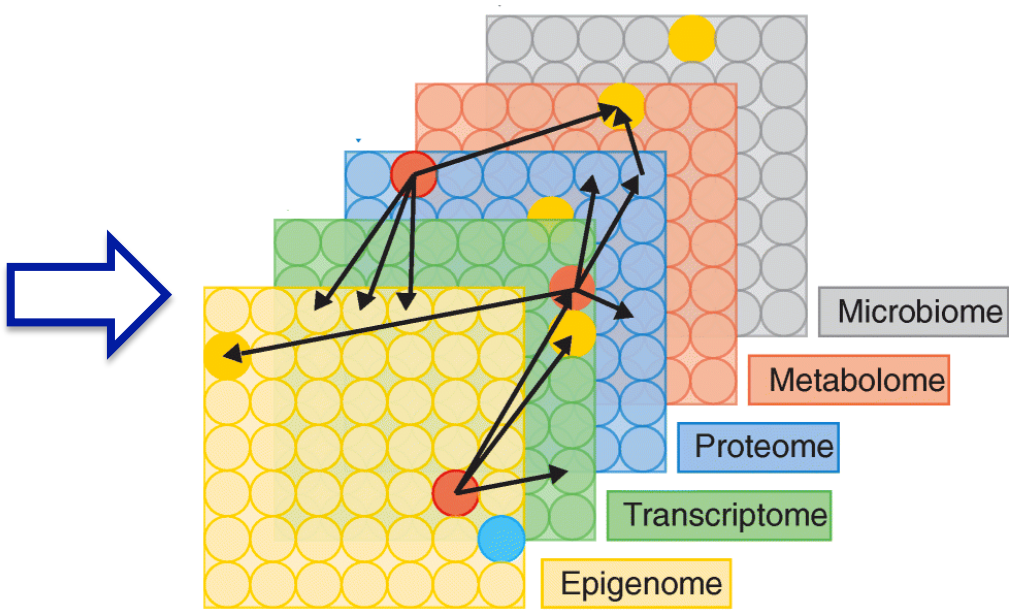
# Multi-omics integration in cancer




The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi: 10.1038/ng.2764

# Multi-omics integration in cancer




The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi: 10.1038/ng.2764

Hasin, Yehudit, Marcus Seldin, and Aldons Lusis. "Multi-omics approaches to disease." Genome biology 18.1 (2017): 83.

# More omics is better, but how many more?

# Is it always good to consider ALL the available omics?

# Choosing which omics to integrate

**Aim:** predicting drug response

**Available input data:**

- Mutations
- Copy Number Alterations (CNA)
- Methylation
- Gene expression
- Proteomics
- Cancer types
- Drug response

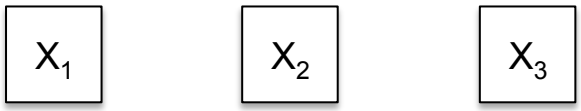
ABEN, Nanne, et al. iTOP: inferring the topology of omics data. Bioinformatics, 2018, 34.17: i988-i996.

# Choosing which omics to integrate

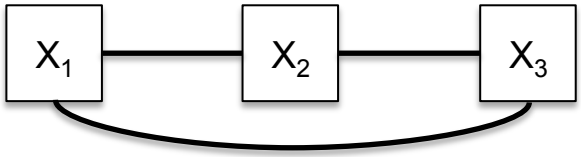
**Aim:** predicting drug response

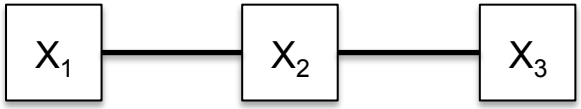
**Available input data:**

- Mutations
- Copy Number Alterations (CNA)
- Methylation
- Gene expression
- Proteomics
- Cancer types
- Drug response




**Using correlation:**



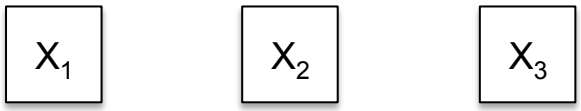

ABEN, Nanne, et al. iTOP: inferring the topology of omics data. Bioinformatics, 2018, 34.17: i988-i996.

# Choosing which omics to integrate

**Aim:** predicting drug response

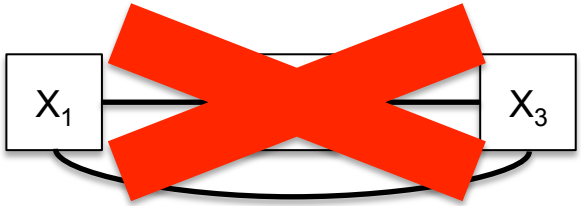
**Available input data:**

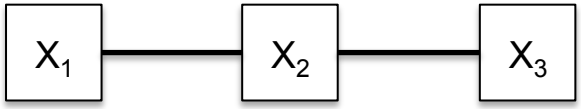
- Mutations
- Copy Number Alterations (CNA)
- Methylation
- Gene expression
- Proteomics
- Cancer types
- Drug response




**Using partial correlation (iTOP):**

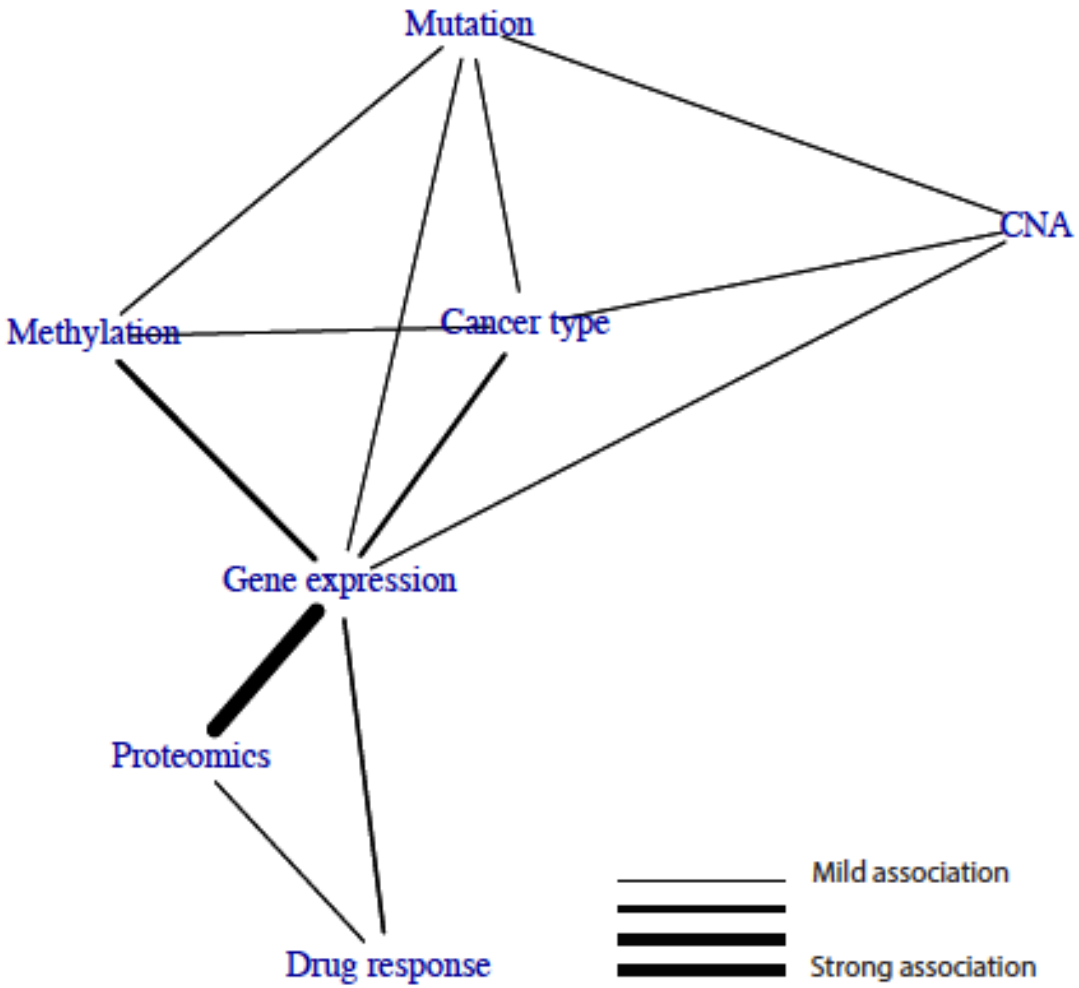
e.g. $cor(X_1, X_3 \mid X_2) \approx 0$



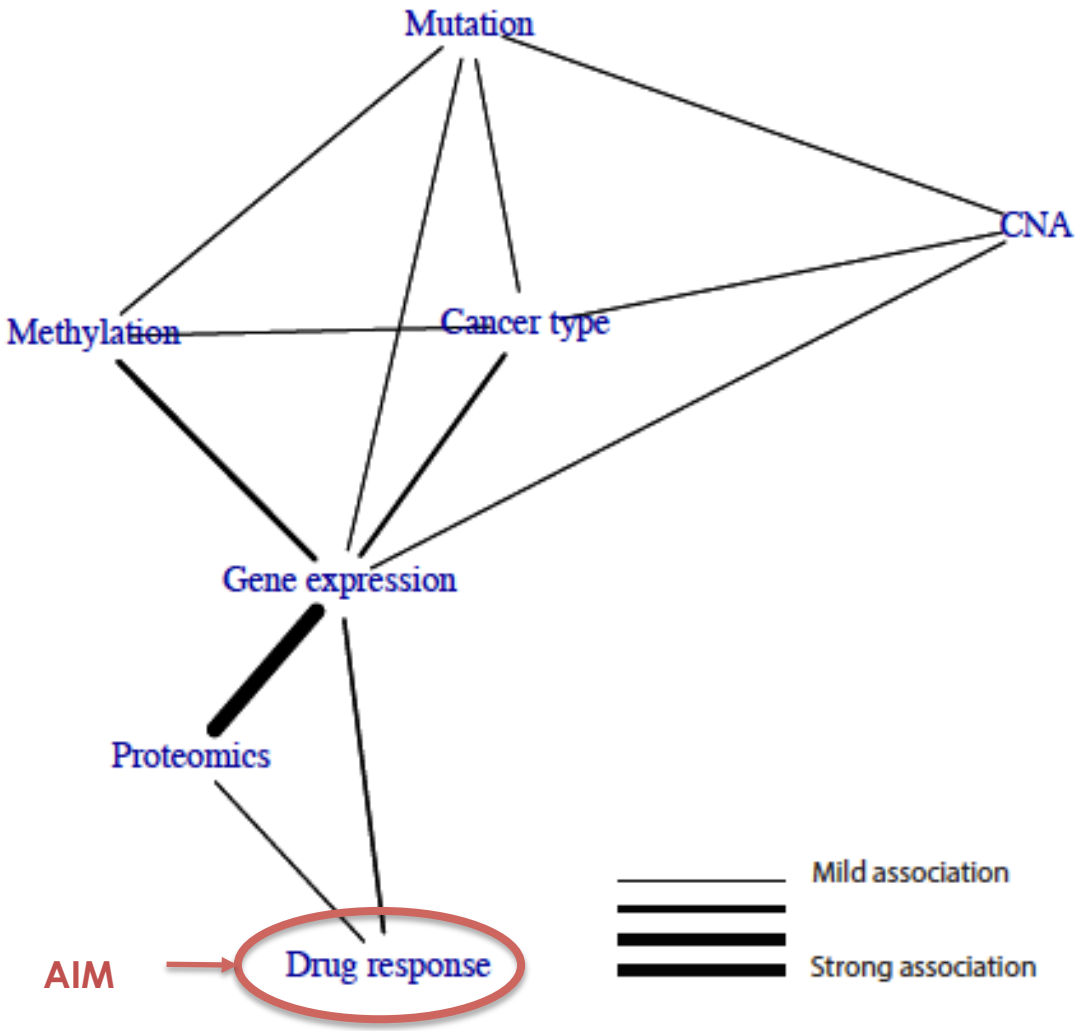

ABEN, Nanne, et al. iTOP: inferring the topology of omics data. Bioinformatics, 2018, 34.17: i988-i996.

# Choosing which omics to integrate



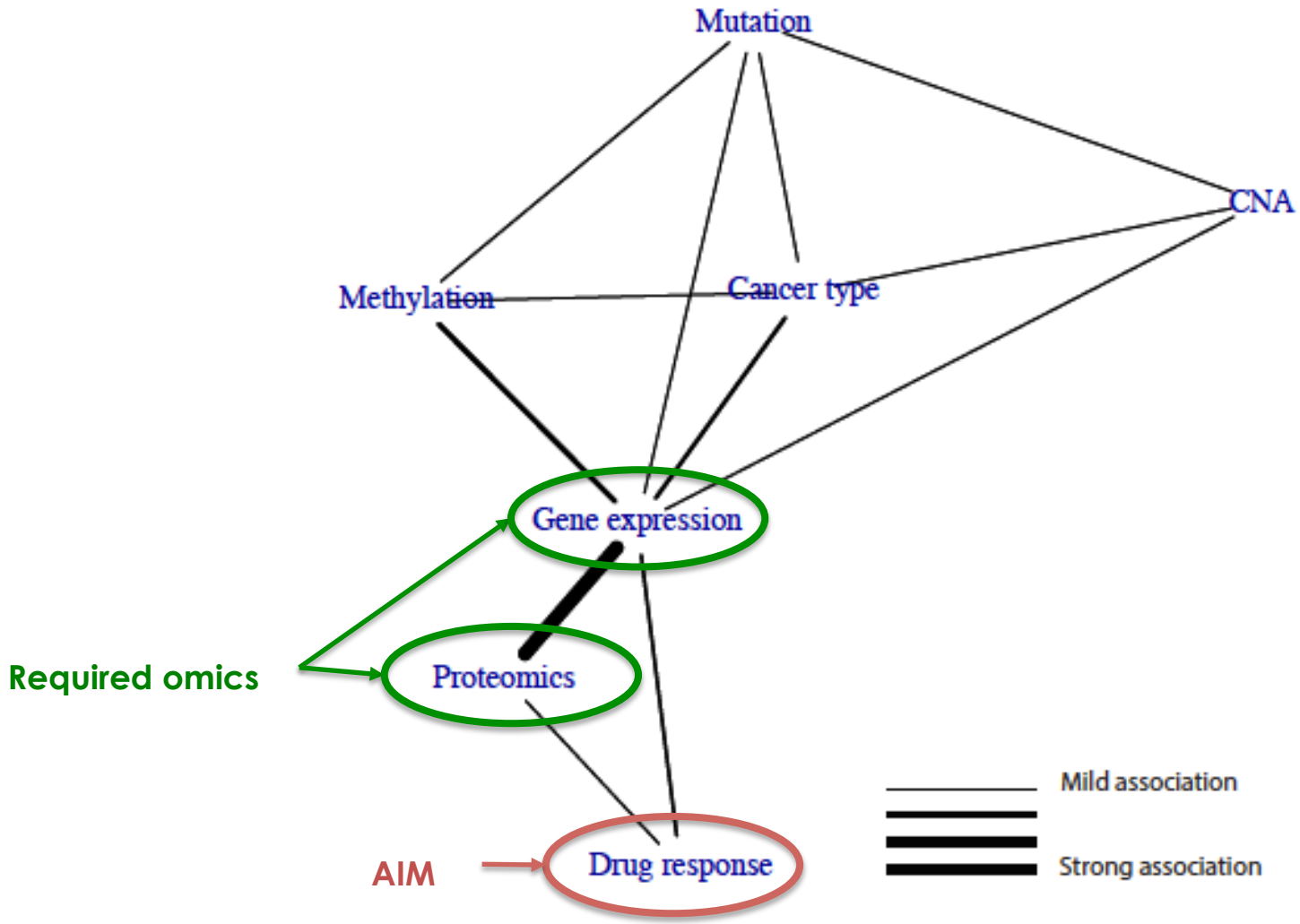

ABEN, Nanne, et al. iTOP: inferring the topology of omics data. Bioinformatics, 2018, 34.17: i988-i996.

# Choosing which omics to integrate




ABEN, Nanne, et al. iTOP: inferring the topology of omics data. Bioinformatics, 2018, 34.17: i988-i996.

# Choosing which omics to integrate

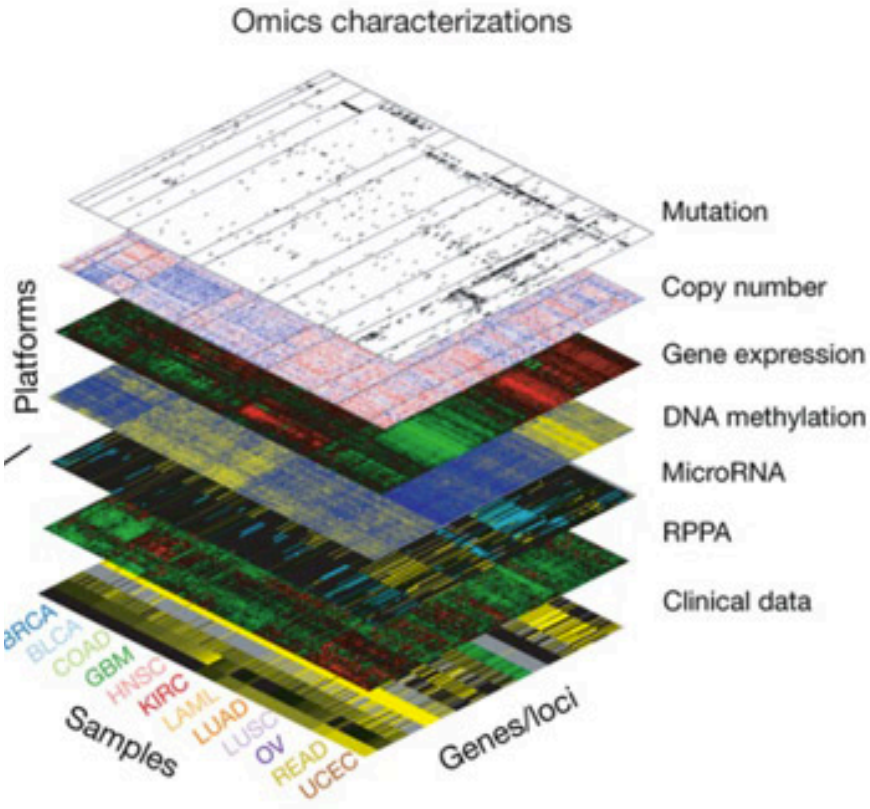



ABEN, Nanne, et al. iTOP: inferring the topology of omics data. Bioinformatics, 2018, 34.17: i988-i996.

# How omics should be combined?

# Challenge: Multi-omics data are heterogeneous

- continuous (e.g. gene expression) vs. discrete (e.g. CNV).
- They have different dimension
- They have different ranges of variability
- Involving heterogeneous entities (e.g. genes, proteins, CpGs). -> matching entities drastically reduce the information

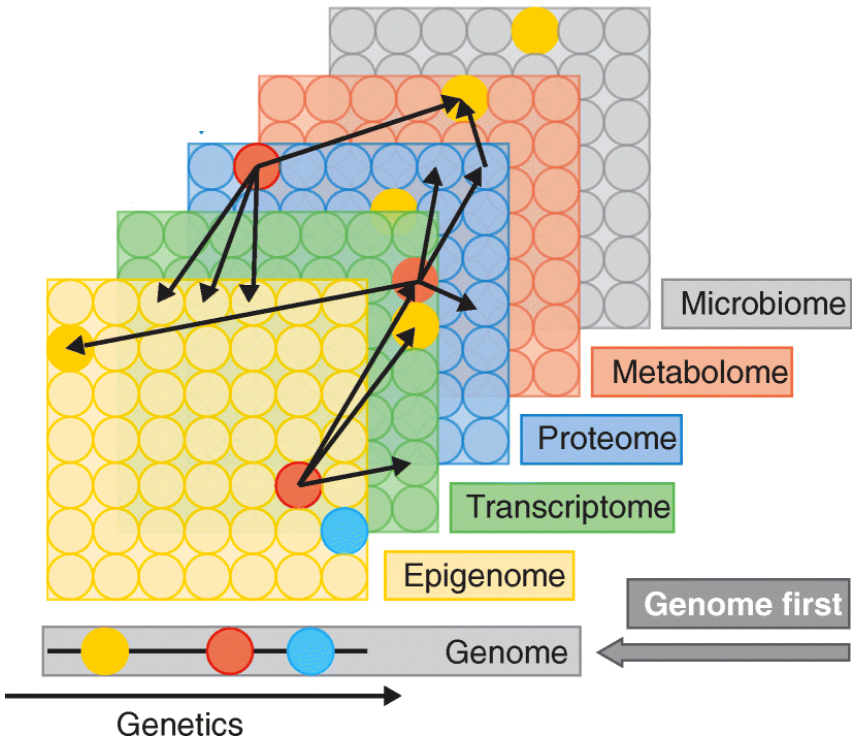


# Integrating multi-omics data

**Approach "Genome First"**

Priority given to genome

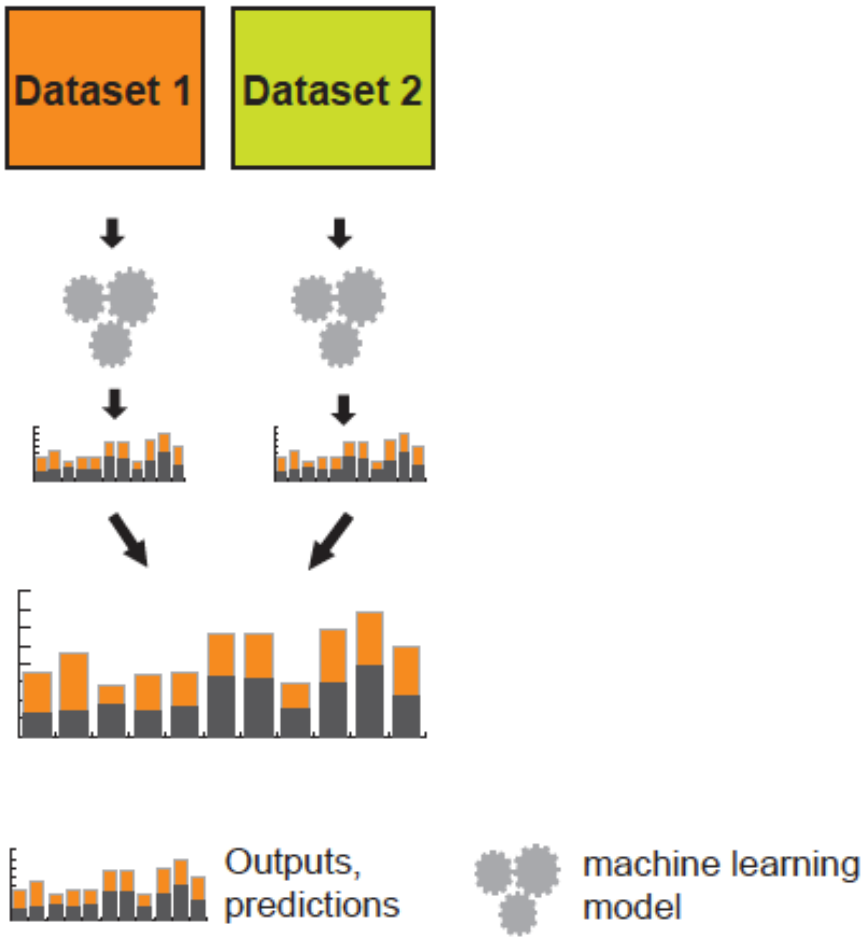
Other omics are only used for interpretation




Hasin, Yehudit, Marcus Seldin, and Aldons Lusis. "Multi-omics approaches to disease." Genome biology 18.1 (2017): 83.

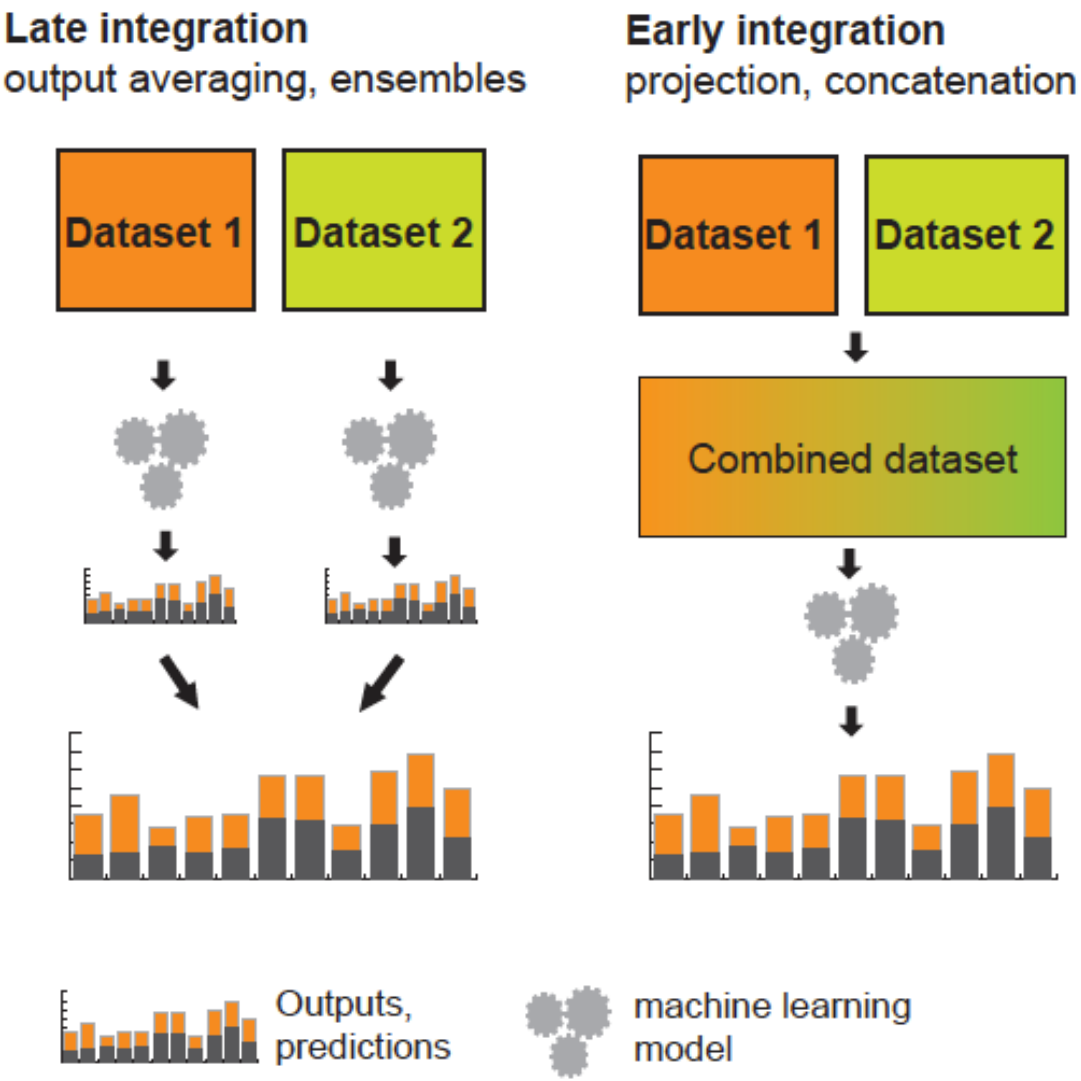
# Integrating multi-omics data

**Late integration**
output averaging, ensembles



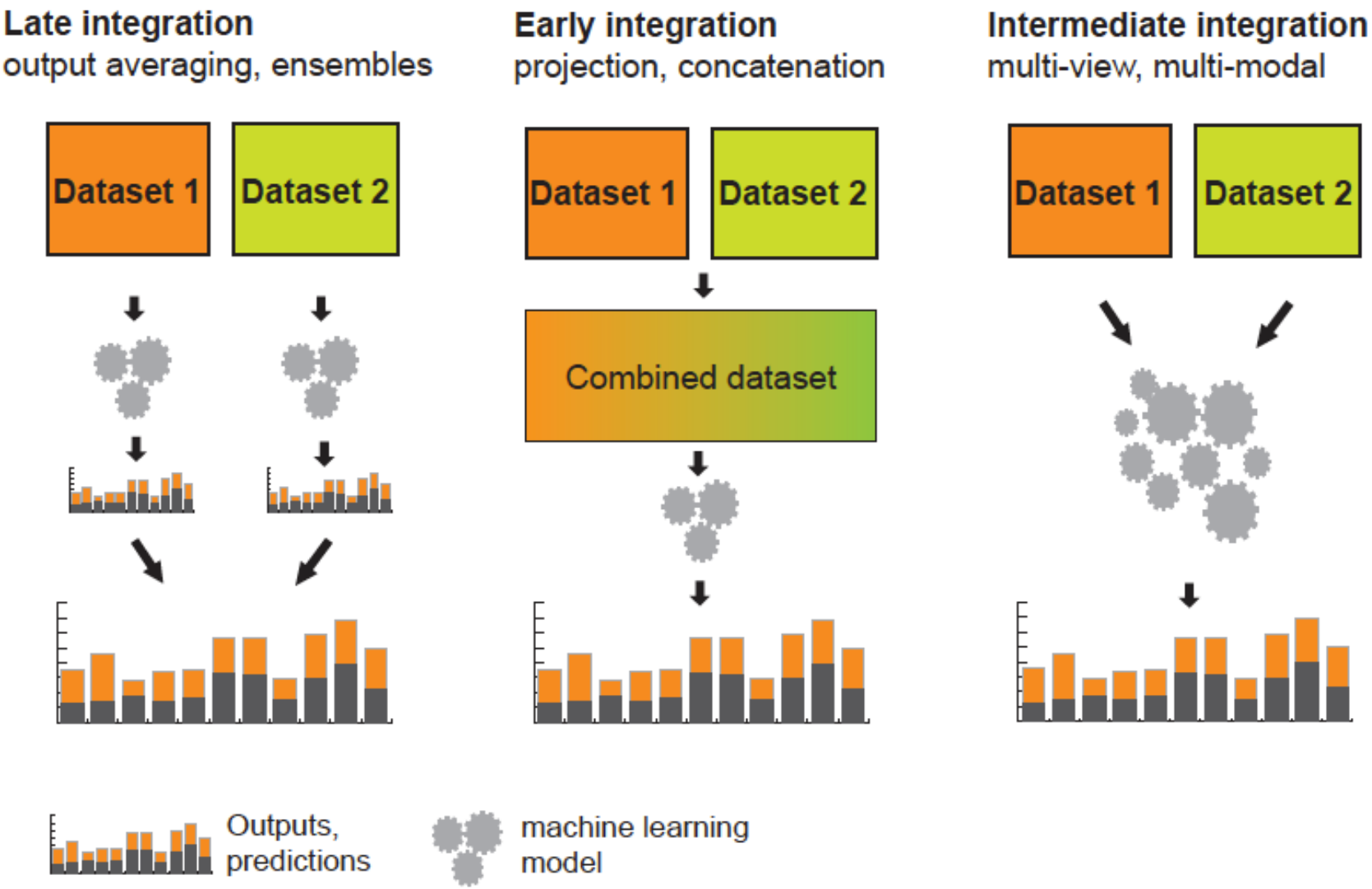

Zitnik, Marinka, et al. "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities." Information Fusion 50 (2019): 71-91.

# Integrating multi-omics data




Zitnik, Marinka, et al. "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities." Information Fusion 50 (2019): 71-91.

# Integrating multi-omics data

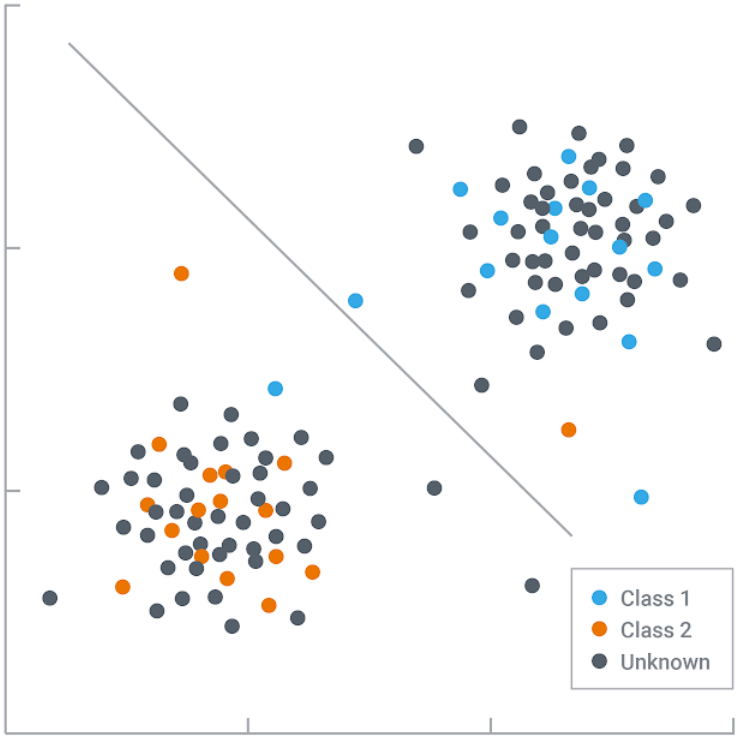



Zitnik, Marinka, et al. "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities." Information Fusion 50 (2019): 71-91.

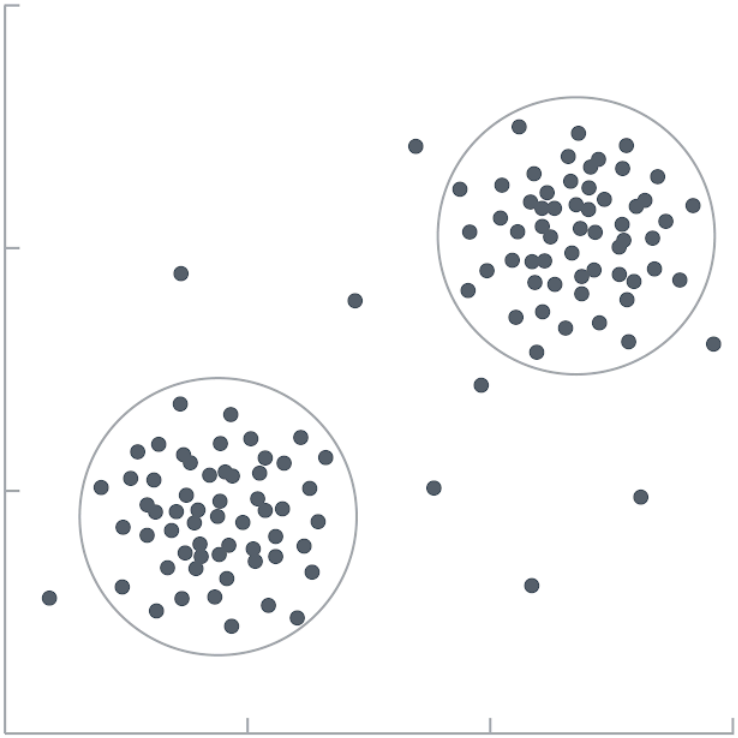
# Main categories of existing multi-omics integrative approaches

# Main categories of integrative approaches

## Supervised methods

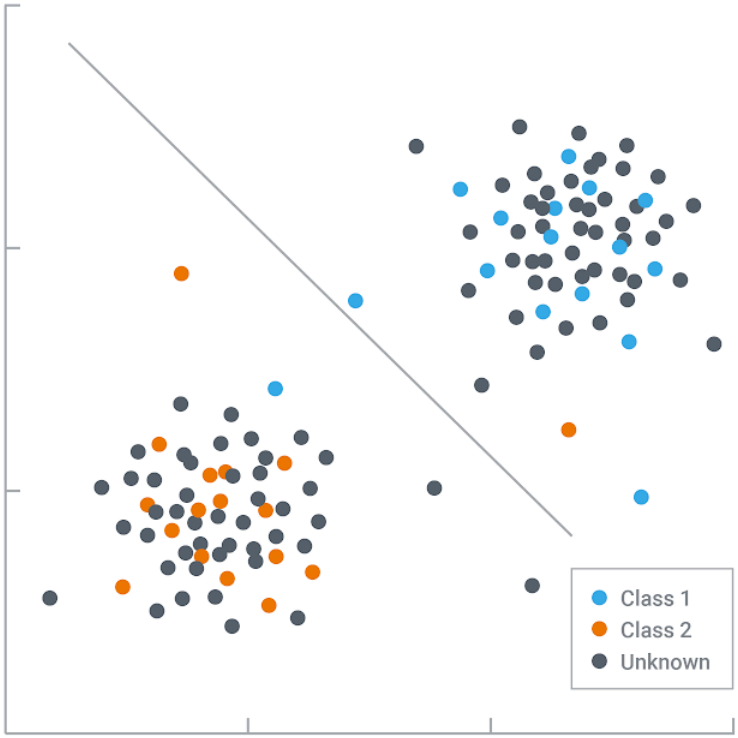



## Unsupervised methods



# Main categories of integrative approaches

## Supervised methods

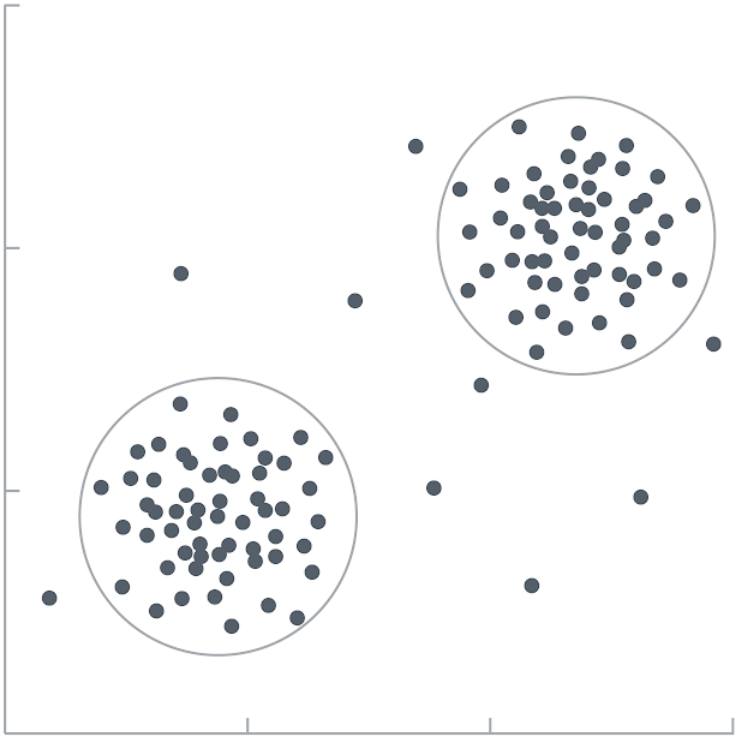



- They require 2 datasets in input: training and test datasets
- Labels must be avilable for the training dataset
- This information is used to infer labels on the test dataset

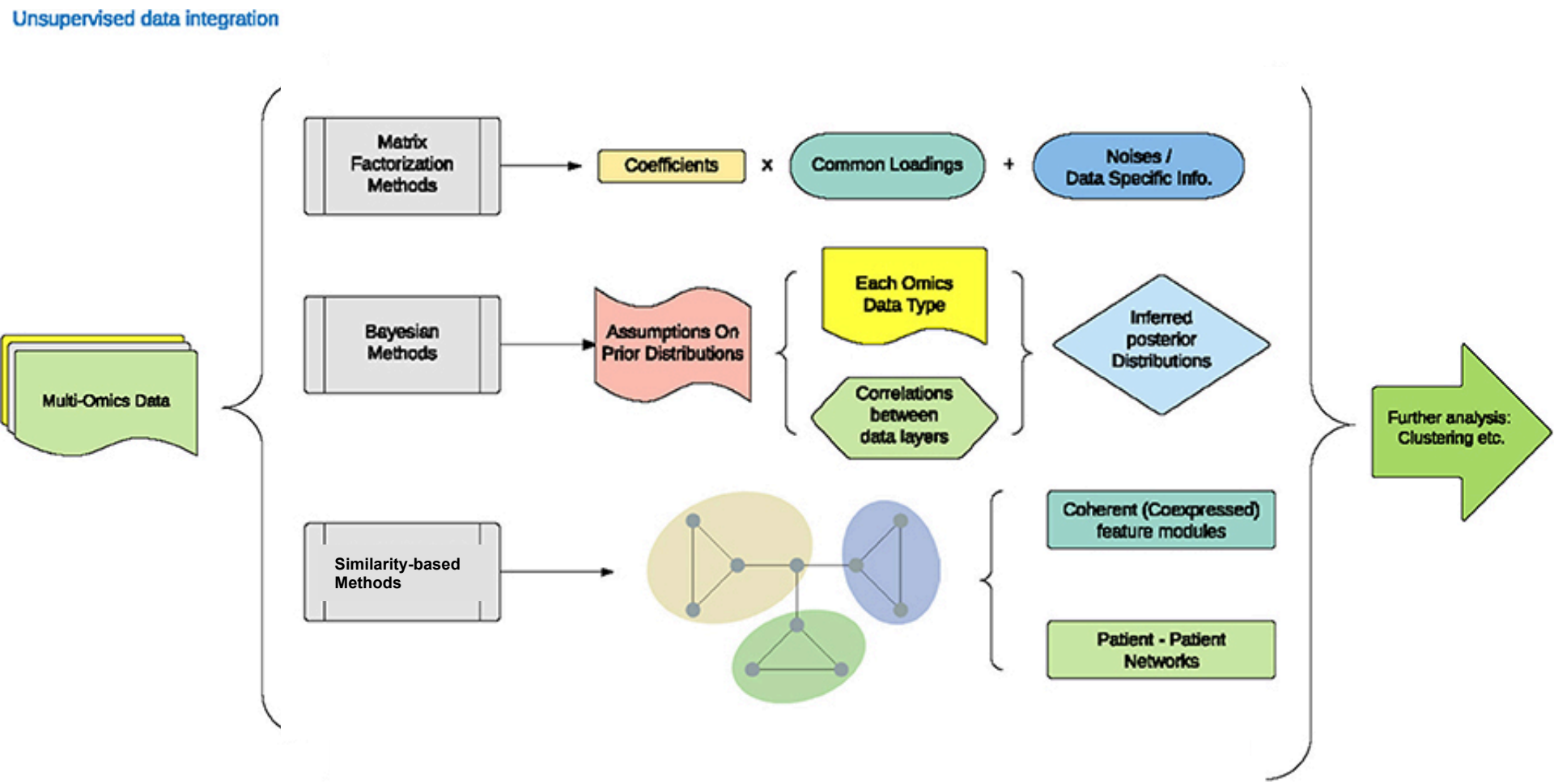
# Main categories of integrative approaches

## Unsupervised methods

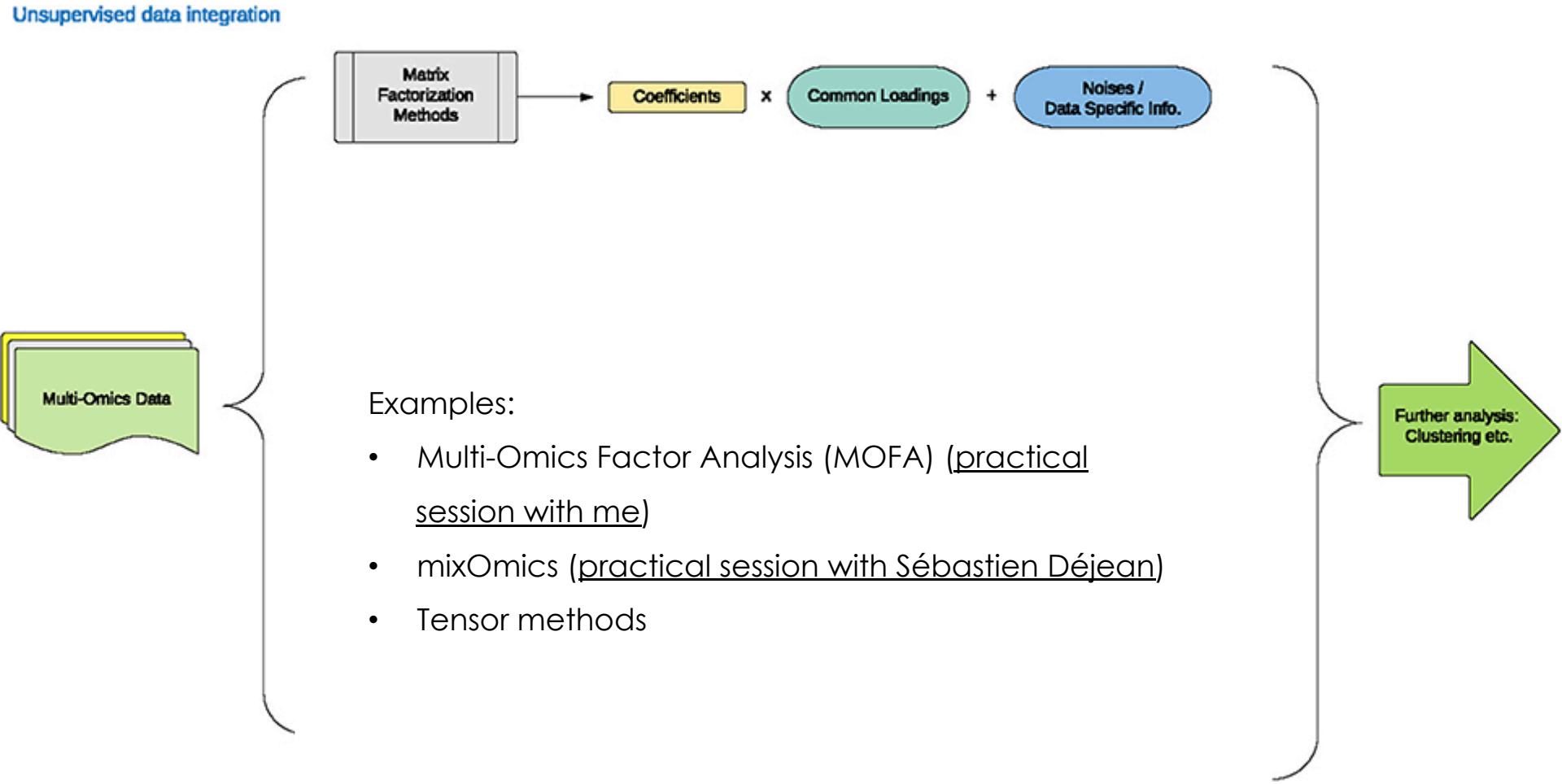
- The methodology is directly applied to one dataset
- They infer information from the structure of the data without any label information



# Unsupervised integrative approaches



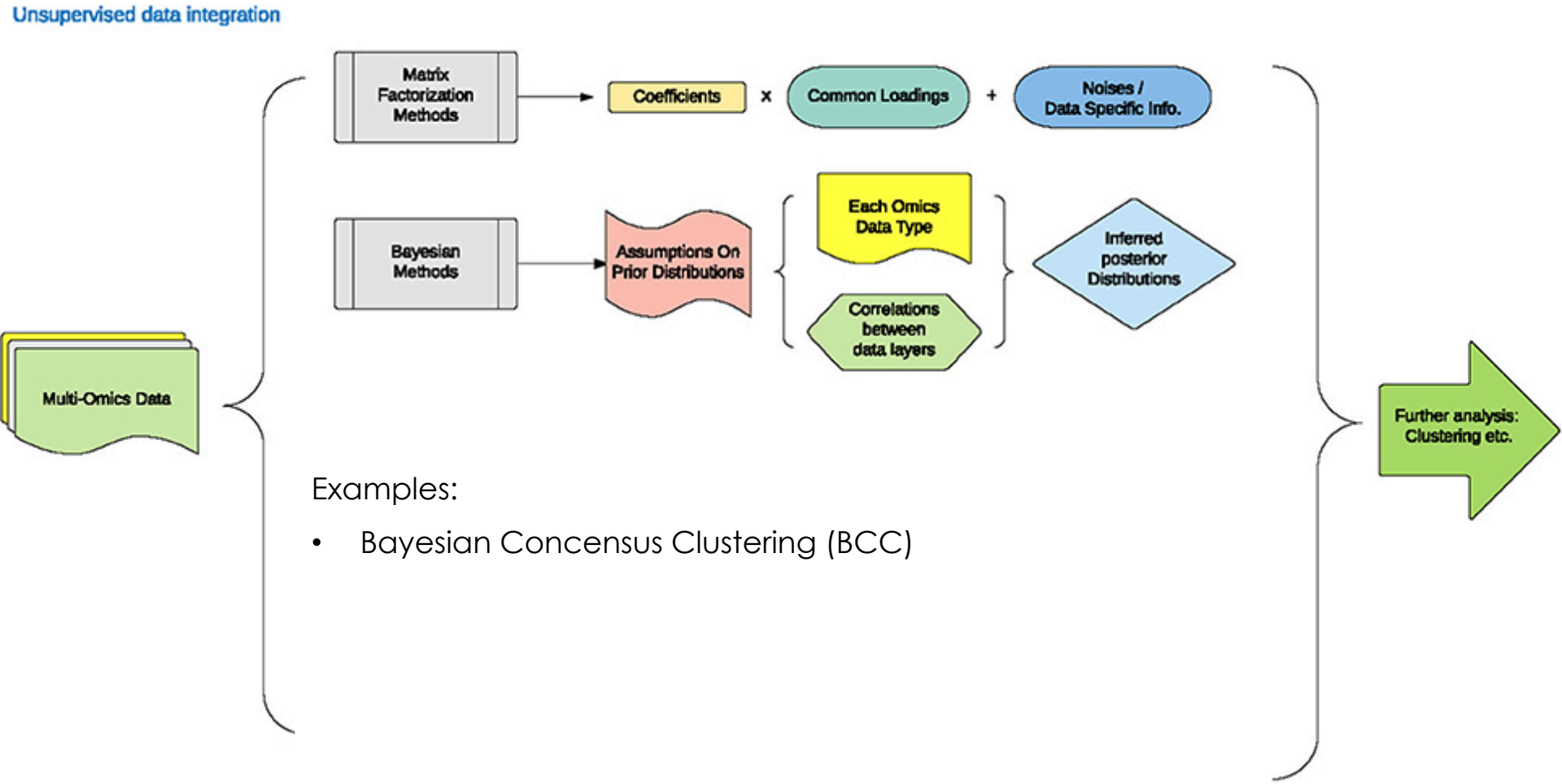
# Unsupervised integrative approaches




Examples:

- Multi-Omics Factor Analysis (MOFA) (practical session with me)
- mixOmics (practical session with Sébastien Déjean)
- Tensor methods

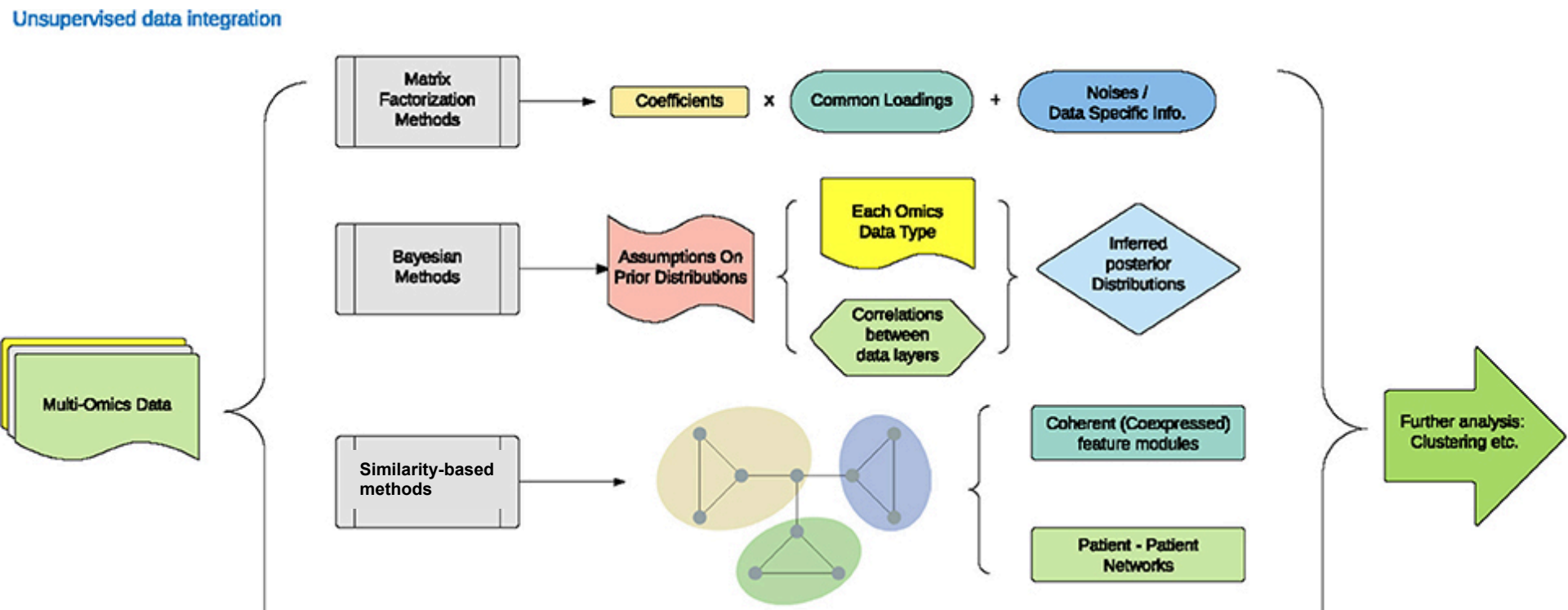
# Unsupervised integrative approaches




Examples:

- Bayesian Concensus Clustering (BCC)

# Unsupervised integrative approaches




Examples:

Network-based methods:

- Similarity between macromolecules: Anais Baudot and Costas Bouyioukos tomorrow
- Similarity between samples: Similarity Network Fusion SNF

Kernel methods: practical session Jérôme Mariette

# Unsupervised integrative approaches

**Of Note:**

This last category of methods give the possibility to integrate heterogeneous sources of information, such as: data matrices, phylogenetic trees and networks

| | Sample_a | Sample_b | Sample_c |
|---|---|---|---|
| OTU_1 | 3 | 3 | 3 |
| OTU_2 | 5 | 2 | 3 |
| OTU_3 | 2 | 0 | 2 |
| OTU_4 | 4 | 3 | 0 |
| OTU_5 | 0 | 2 | 2 |
| OTU_6 | 0 | 2 | 2 |


Examples:

Network-based methods:

- Similarity between macromolecules: Anais Baudot tomorrow
- Similarity between samples: Similarity Network Fusion SNF

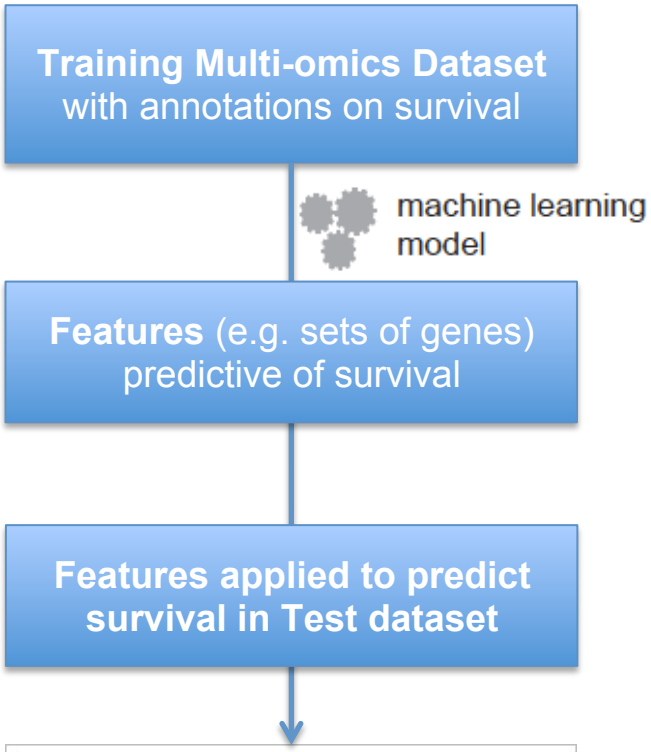
Kernel methods: practical session Jérôme Mariette

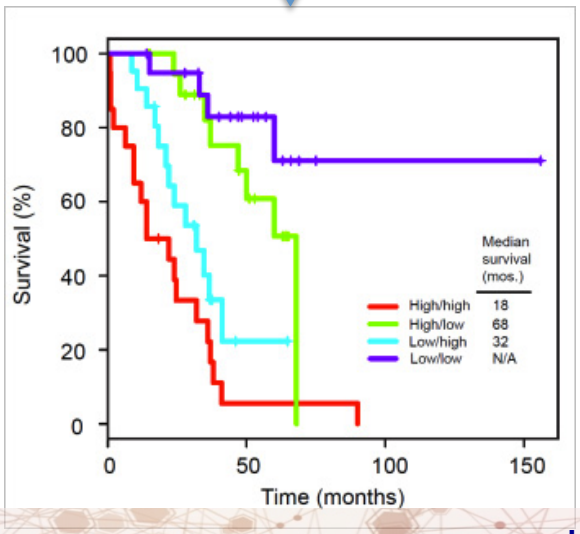
# Cancer insights from data integration methods

# Patients survival prediction

Given a set of cancer we want to patients predict their survival.

This problem is generally approached with supervised approaches.

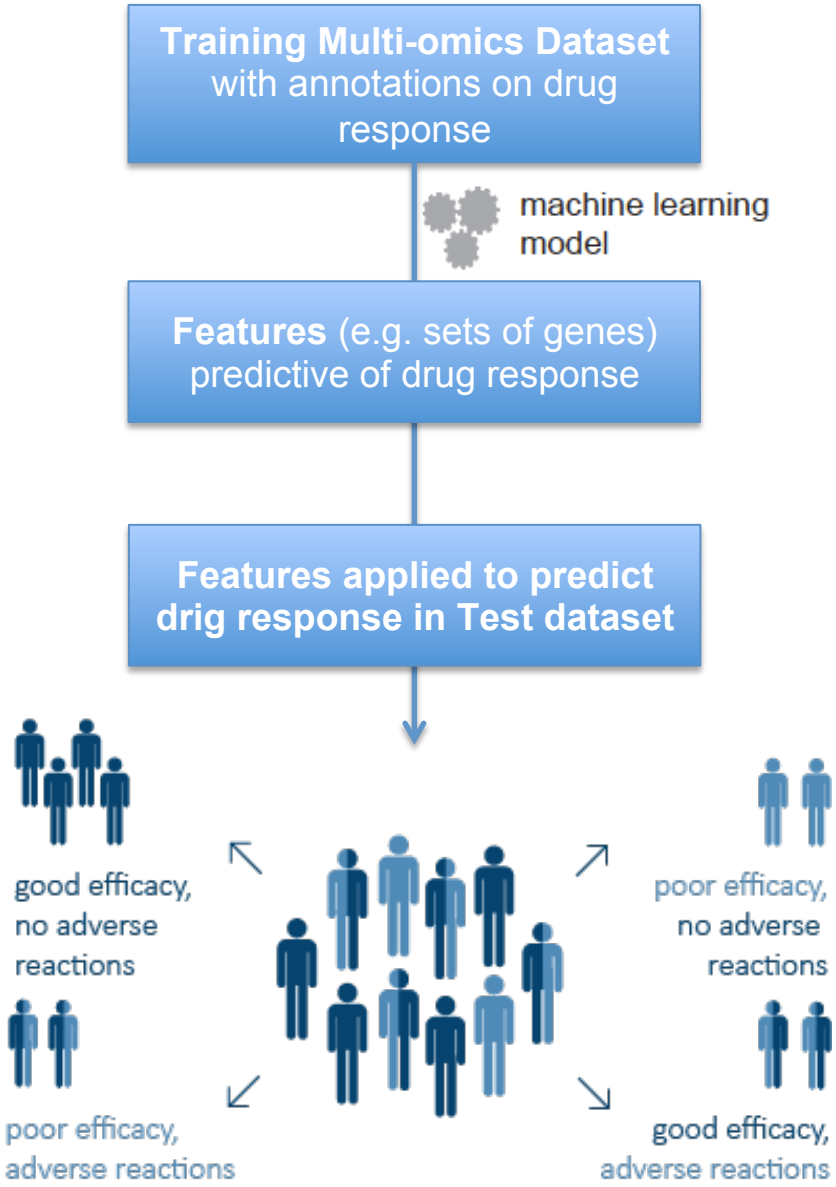


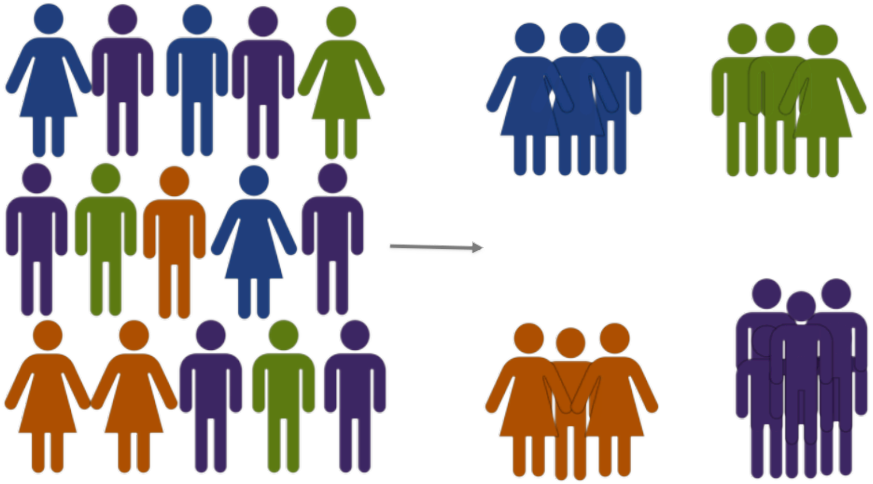
# Drug responce prediction

Given a set of cancer we want to patients predict which will respond to a given therapy.

This problem is generally approached with supervised approaches.



# Cancer subtyping

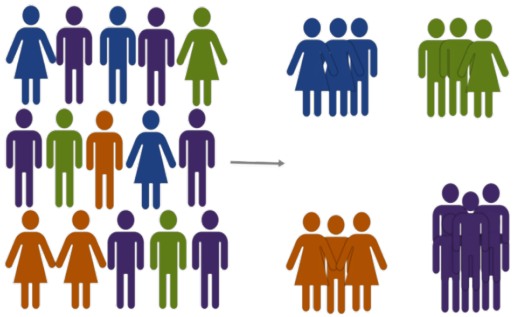


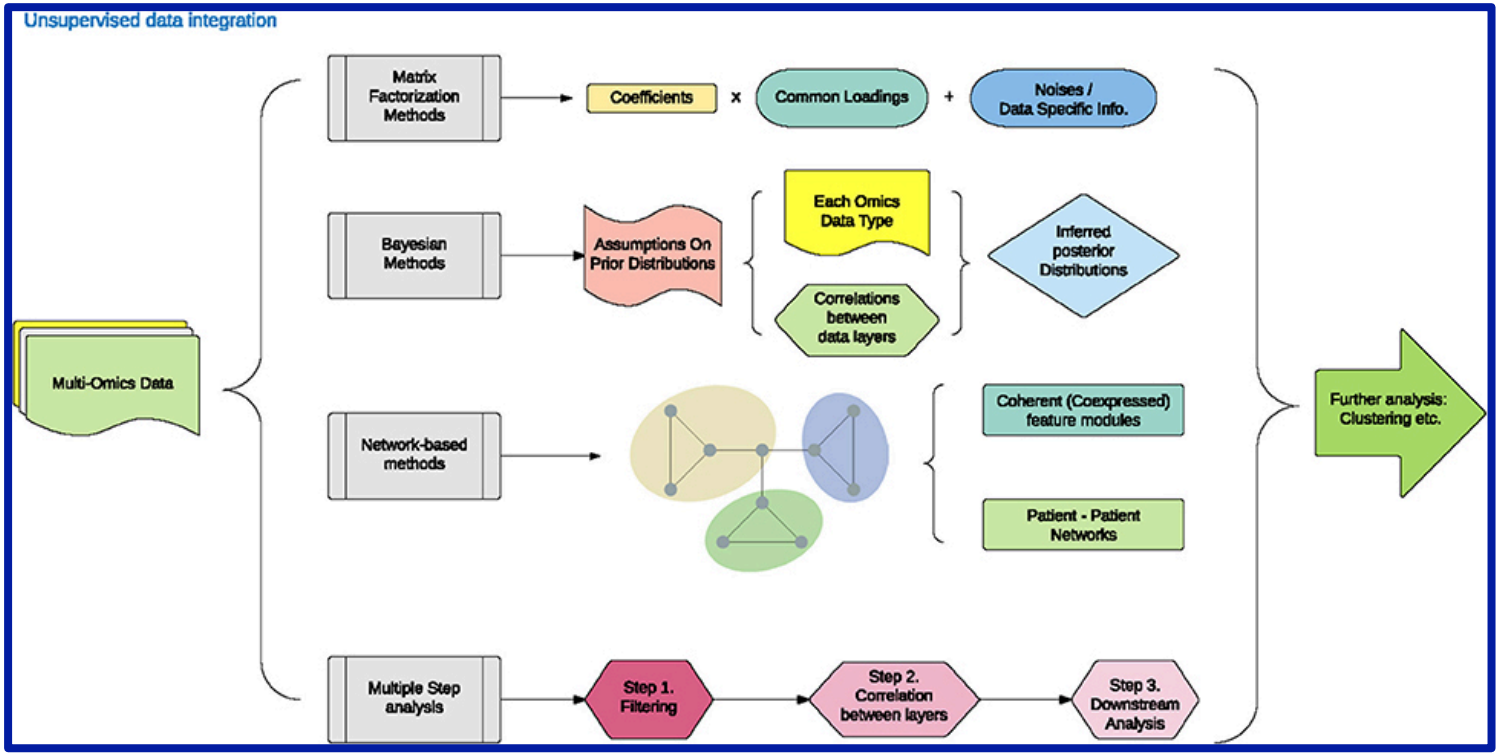
| CMS1 (13%) | CMS2 (35%) | CMS3 (11%) | CMS4 (20%) | Unclassified (21%) |
|---|---|---|---|---|
| • Right colon, female<br>• MSI, *BRAF* mut, hypermutated<br>• Immune activation<br>• Worse survival after relapse | • Left colon<br>• MSS, CIN, *BRAF* wt, *TP53* mut<br>• Epithelial, WNT/Myc pathway activation<br>• Better survival after relapse | • *KRAS* mut<br>• Epithelial, IGFBP2 overexpression | • Mesenchymal, TGFβ pathway activation, NOTCH3 overexpression<br>• Worse relapse free survival and overall survival | • Immune and stroma infiltration<br>• Variable epithelial - mesenchymal activation |
| C2, Subtype 1.2, A-type, CCS2, C, Inflammatory | C1-C5-C6, B-type, Subtype 2.2, B, CCS1, Enterocyte-TA | C3, Subtype 2.1, Globet-like, A | C4, C-type, Subtype 1.1-1.3, CCS3, D-E, Stem-like | |

Santos, Cristina, et al. "Intrinsic cancer subtypes-next steps into personalized medicine." Cellular oncology 38.1 (2015): 3-16.

# Cancer subtyping

This problem is generally approached with unsupervised approaches.

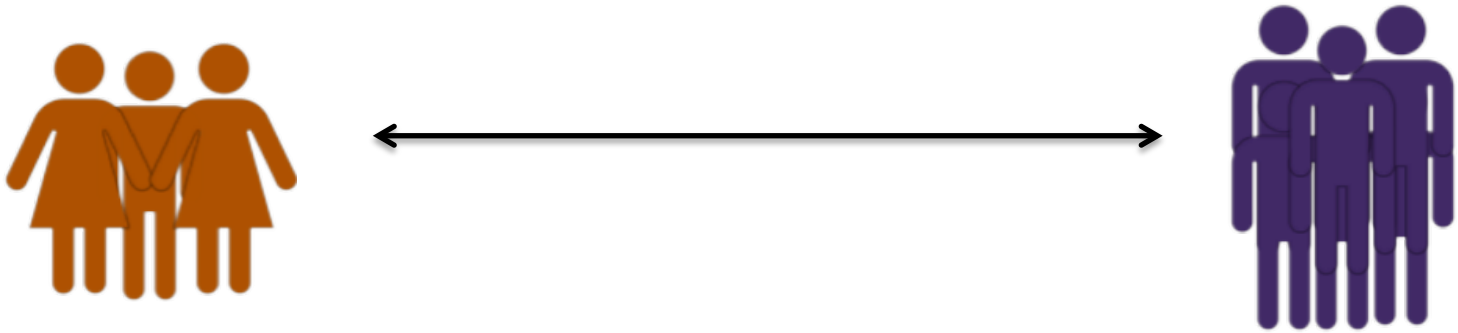




# Gene modules identification

Drug rensponding

Drug resistent

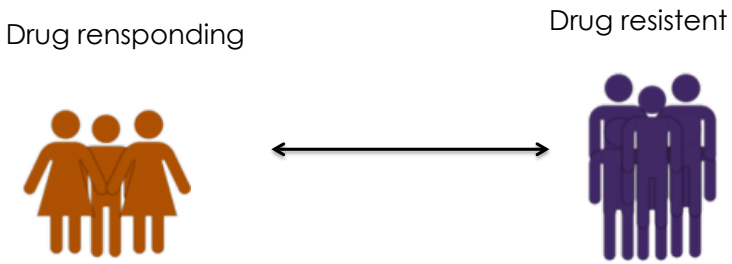


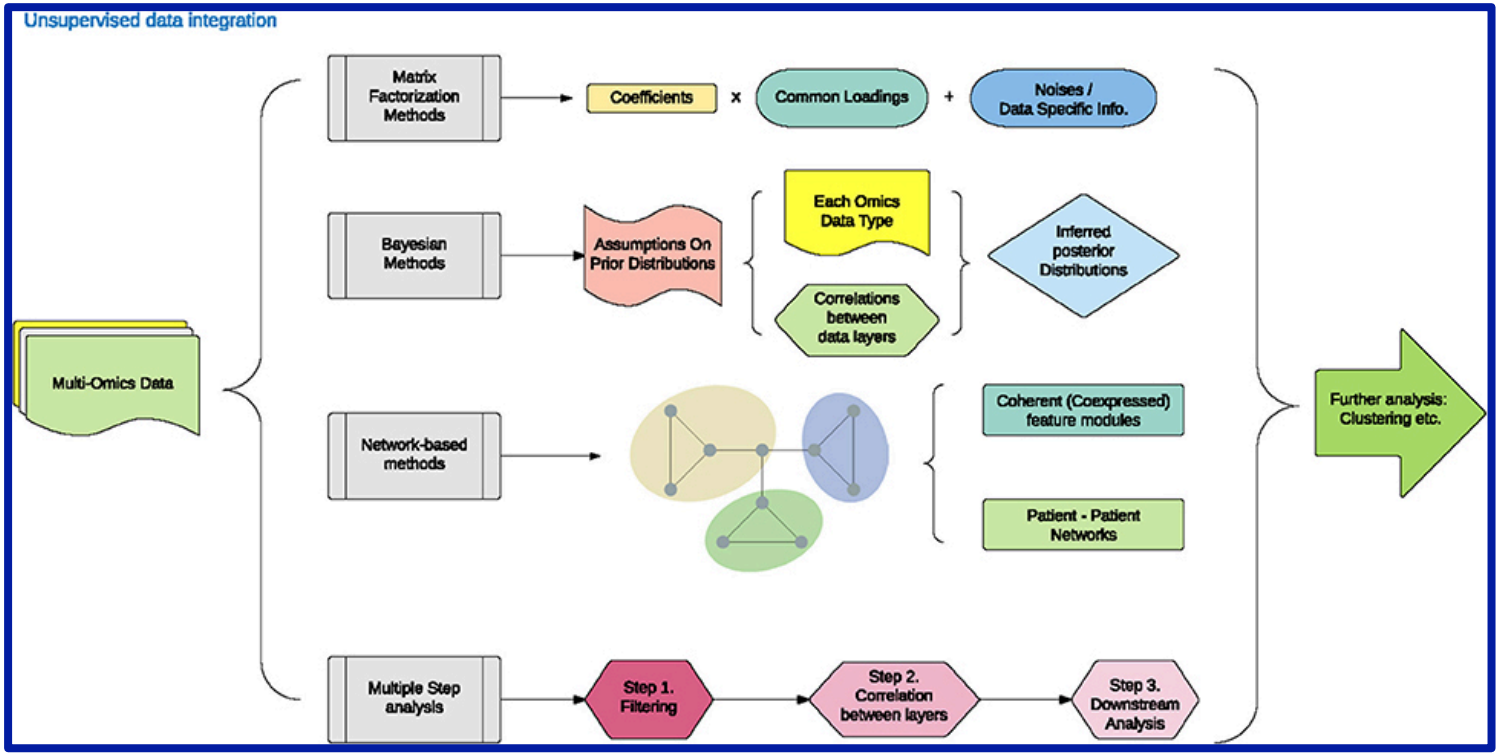
Which are the molecular mechanisms that make these two groups of patients having a different behaviour?

Can we identify a driver that can alter the behaviour of a set of patients?

# Gene modules identification

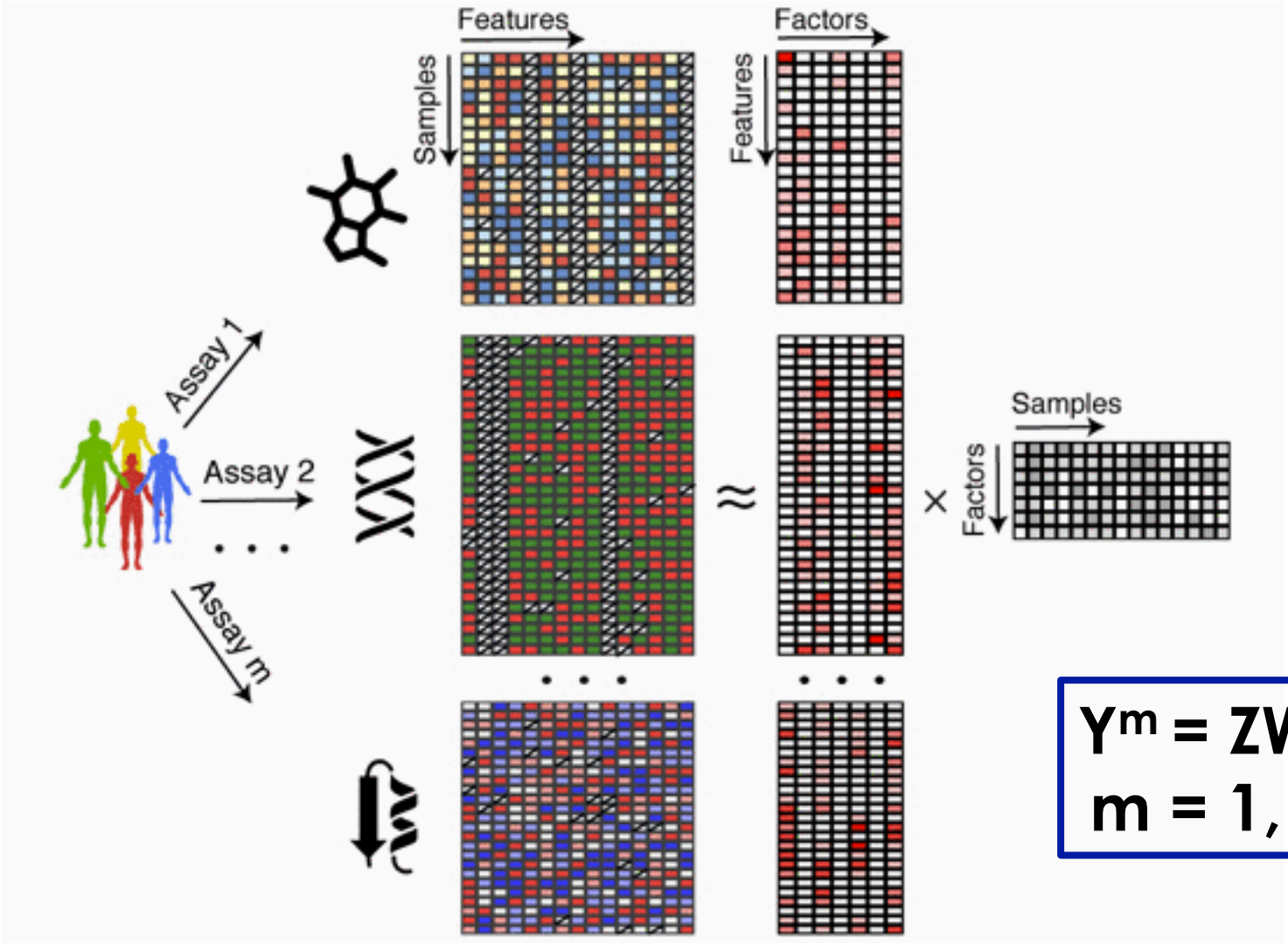
This problem is generally approached with unsupervised approaches.



# Multi-Omics Factor Analysis (MOFA)

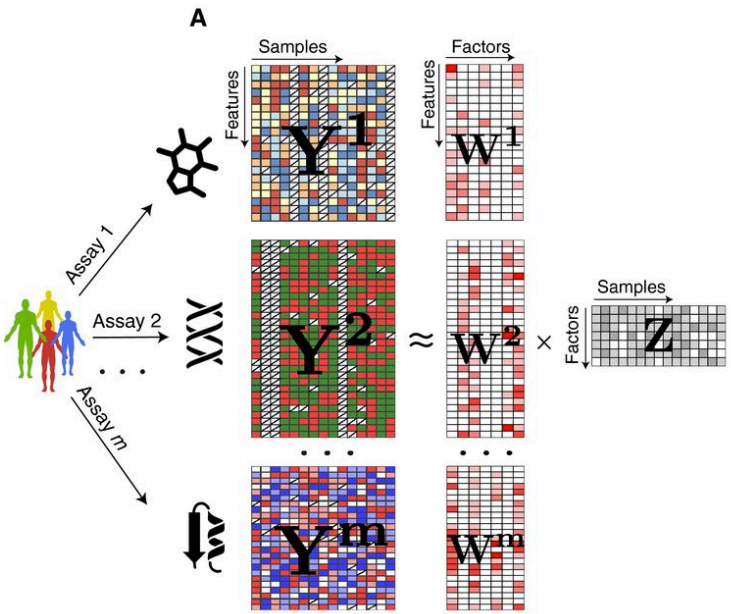
# MOFA model

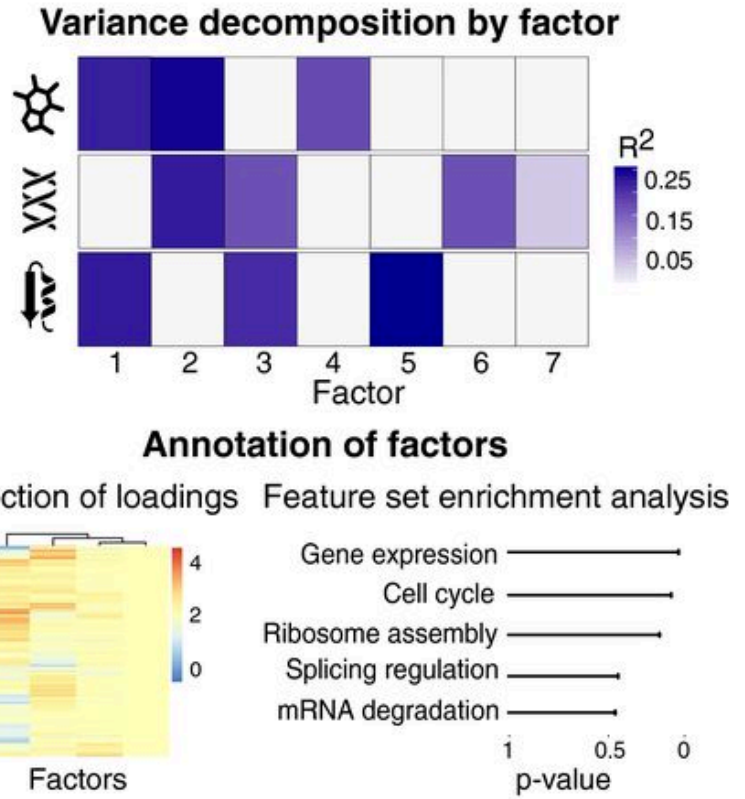


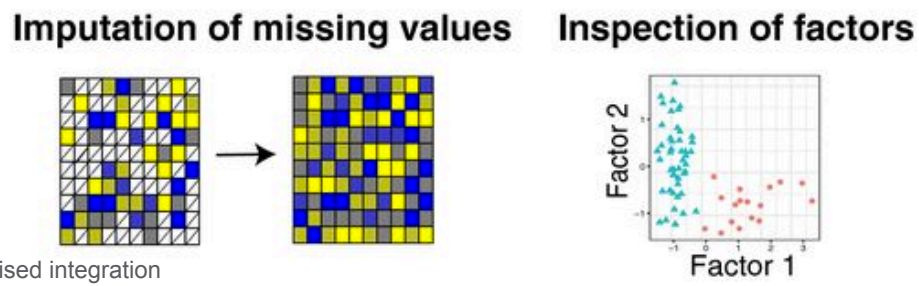

$$Y^m = ZW^m + e^m$$
$$m = 1, \ldots, M$$

Argelaguet, Ricard, et al. "Multi−Omics Factor Analysis—a framework for unsupervised integration of multi−omics data sets." Molecular systems biology 14.6 (2018): e8124.

# MOFA advantage: interpretability of factors




Argelaguet, Ricard, et al. "Multi−Omics Factor Analysis—a framework for unsupervised integration of multi−omics data sets." Molecular systems biology 14.6 (2018): e8124.