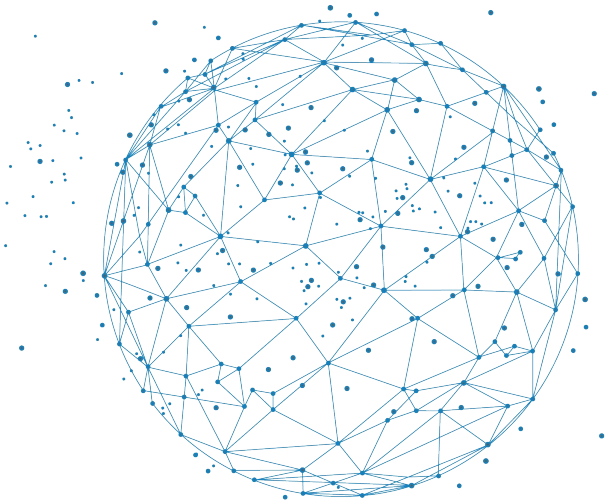
Second edition 2024 in Fréjus

ÉCOLE THÉMATIQUE ETBII BIOINFORMATIQUE INTÉGRATIVE IFB



Omics integration - General aspects

Jimmy Vandel (PLBS - Bililleu)
Arnaud Gloaguen (CNRGH-CEA)
Vincent Guillemot (Institut Pasteur)

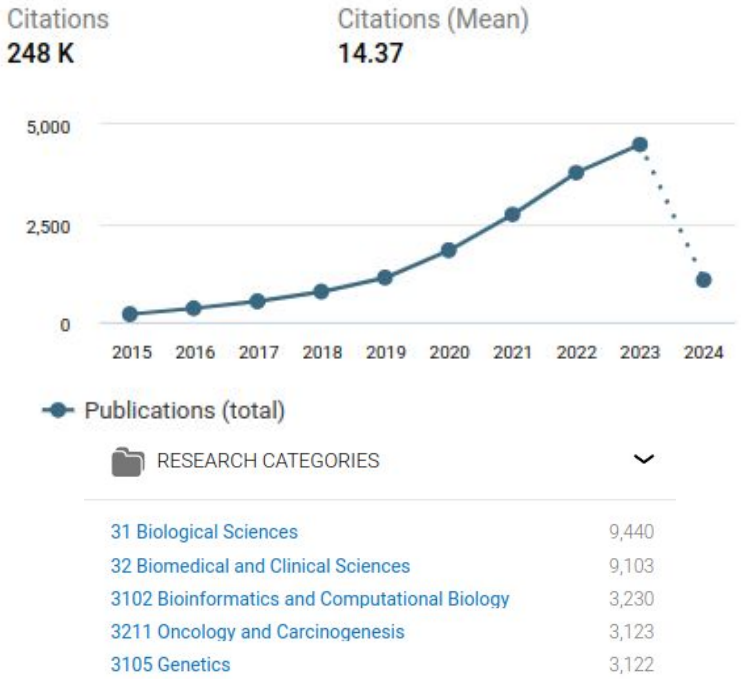
DOI final version

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

cnrs INRAE Inserm cea INVESTIR L'AVENIR

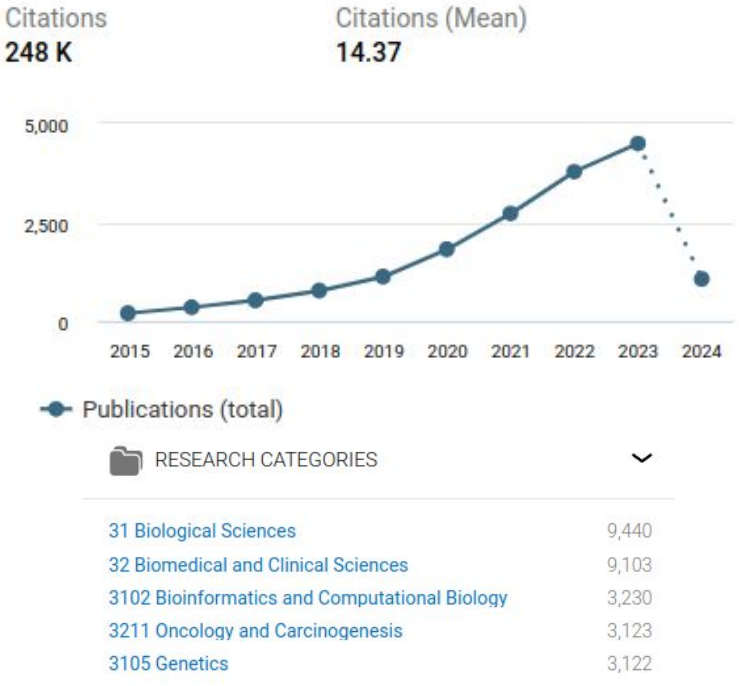
“Multi-omics” citations



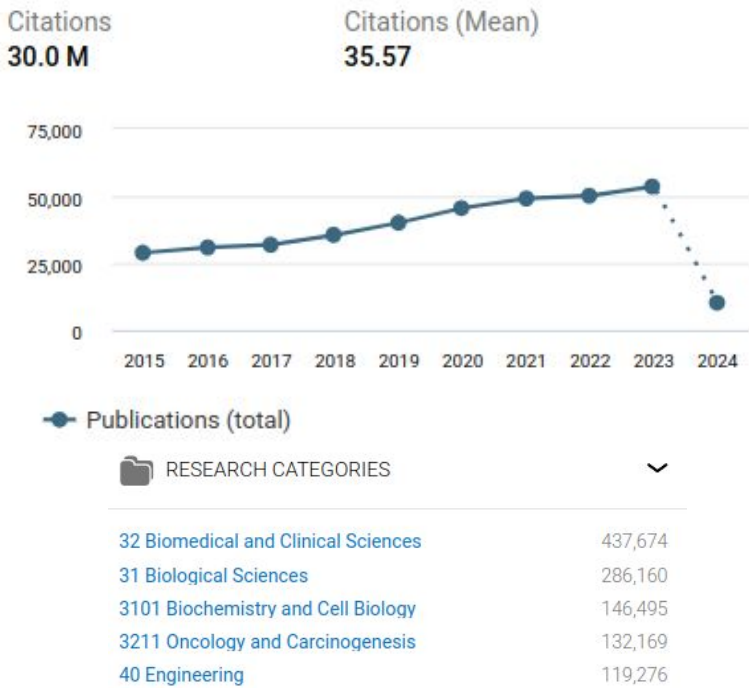
https://app.dimensions.ai/discover/publication *(15th Mar. 2024: 143,523,222 referenced publications)*

Rise in popularity

“Multi-omics” citations

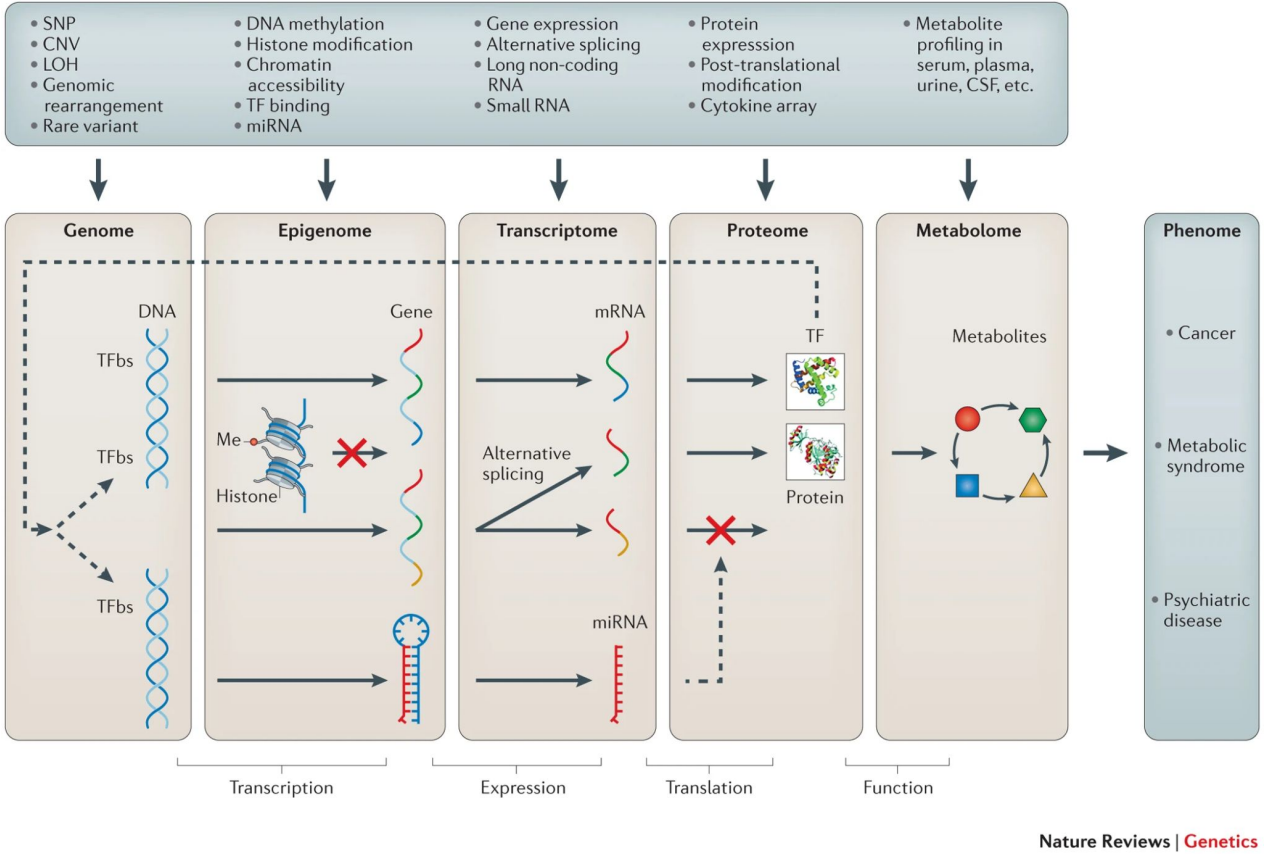


“Single-cell” citations



https://app.dimensions.ai/discover/publication *(15th Mar. 2024: 143,523,222 referenced publications)*

Omics… which ones ?



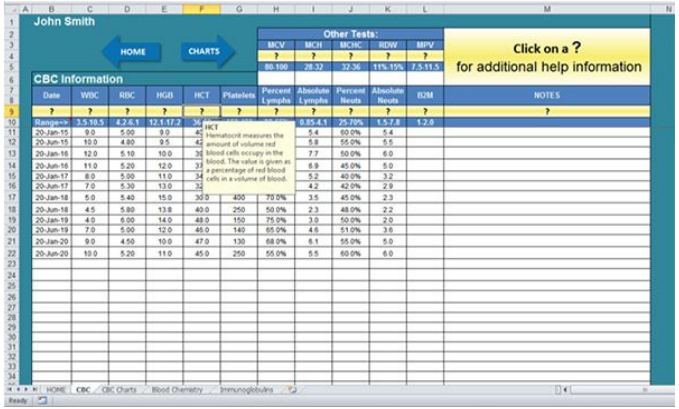
Ritchie, M., Holzinger, E., Li, R. et al. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet 16, 85–97 (2015).

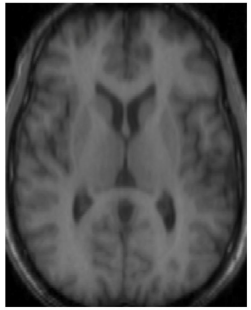
But also ?

Other related data ?

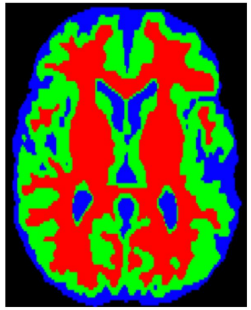
- clinical data
- imaging data (full data or extracted characteristics)
- new omics fields : fluxomics, ionomics, microbiomics, glycomics…
- biological knowledge : DNA/protein, protein/protein interactions, DNA recombination

→ a priori in model definition/construction





(a) Axial slice

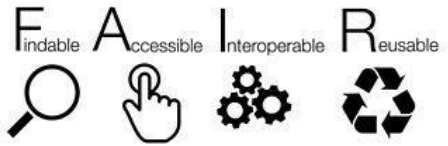


(b) Tissue segmentation

Integration: motivation and challenges

- Take advantage of the vast amount of available data
  - Data access (local/national regulation, infrastructures...)
  - Data representation (structuration, ontologies...)

  -> Need of common representation framework

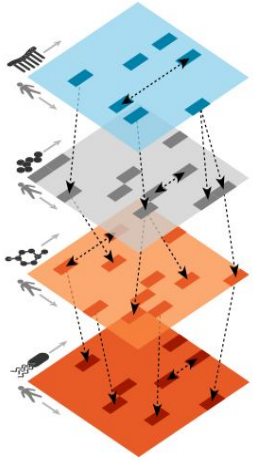


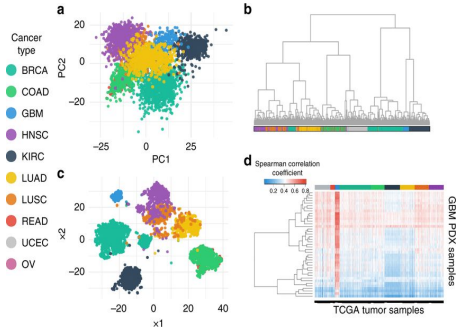
- Improve our understanding of biological phenomena
  - Data heterogeneity (technology, format, biological meaning, stat. distribution...)
  - Data complexity (dependances/independances, ad-hoc assumptions...)
  - Amount a data (time/memory consuming)

  -> Need of new analysis methods/algorithms

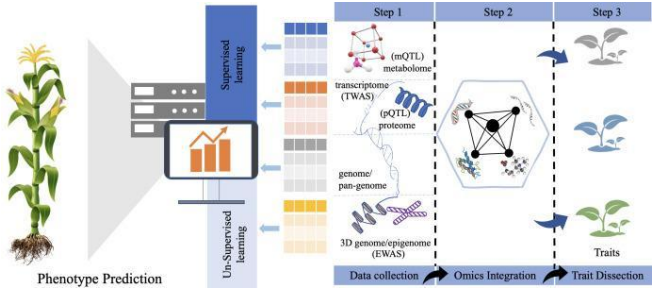
Integration: biological motivation

- Deep insights into biology phenomenon
- Subtyping and classification (disease, species, varieties)
- Biomarkers prediction : diagnostic, disease drivers, plant/animal selection...



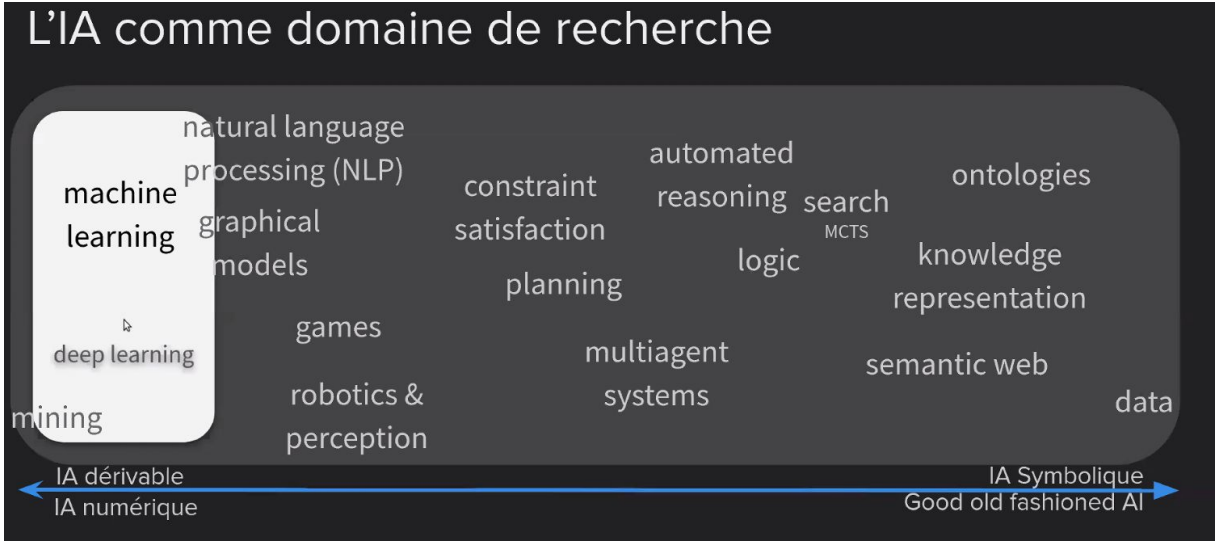


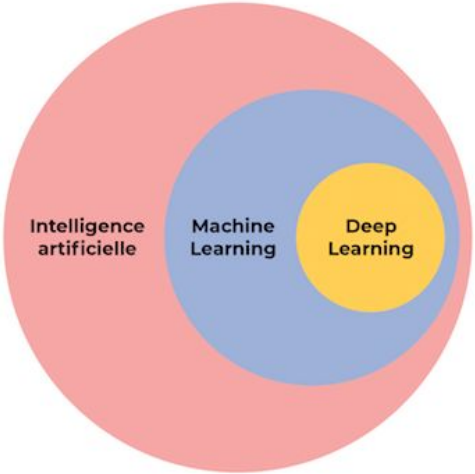
Vasileios et al (2018). Drug and disease signature integration identifies synergistic combinations in glioblastoma. Nature Communications. 9.



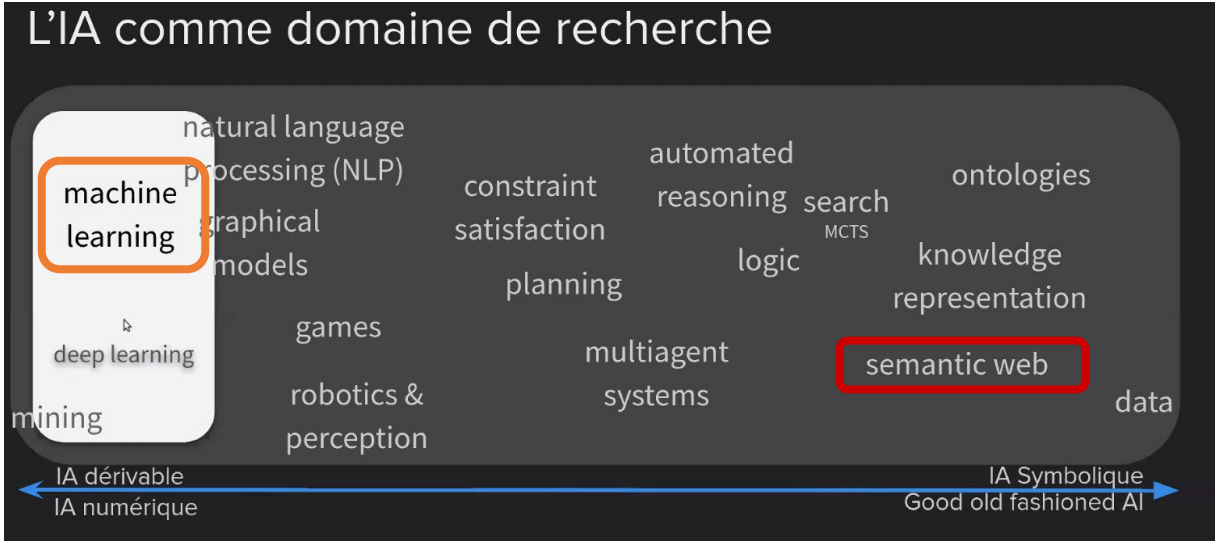
Mahmood et al (2022). Multi-omics revolution to promote plant breeding efficiency. Front Plant Sci.

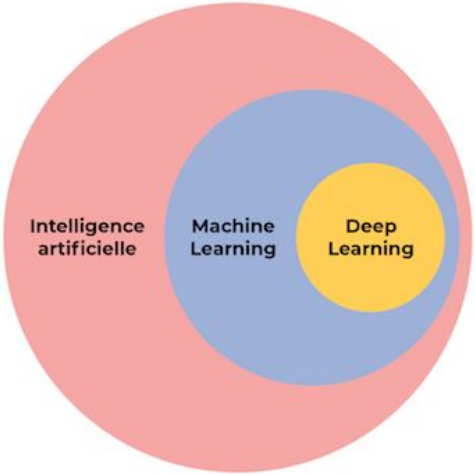
Artificial intelligence of course … and so ?





Artificial intelligence of course … and so ?

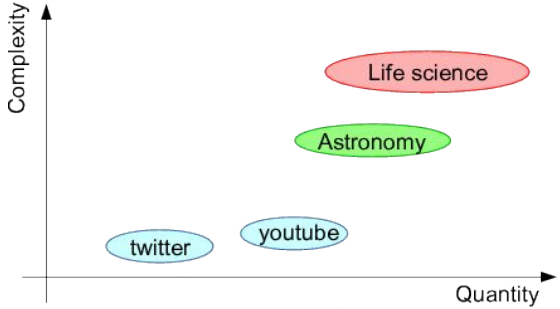




Improve our understanding of biological phenomena

Take advantage of the vast amount of available data

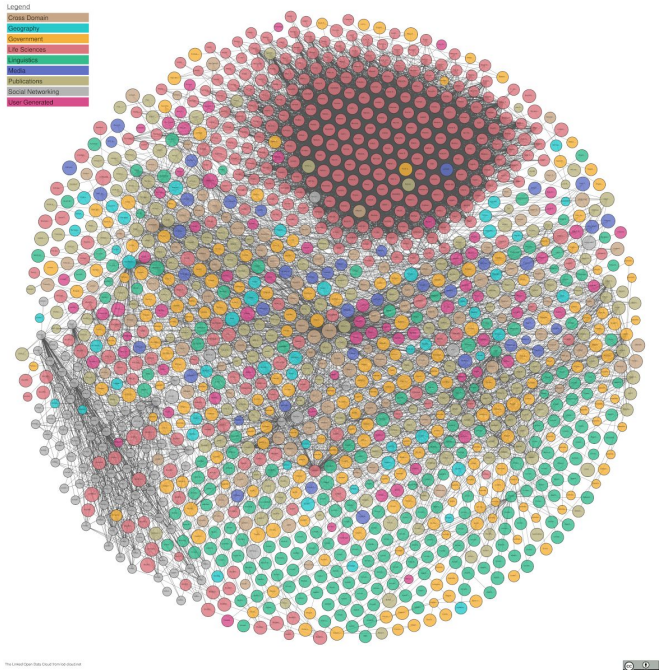
Take advantage of the vast amount of available data



Big Data: Astronomical or Genomical?

Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]*

PLOS Biology | DOI:10.1371/journal.pbio.1002195 July 7, 2015



Life science: 1600+ reference databases

→ integrating heterogeneous data and knowledge is (badly) needed!

Editorial > Nucleic Acids Res. 2022 Jan 7;50(D1):D1-D10. doi: 10.1093/nar/gkab1195.

The 2022 Nucleic Acids Research database issue and the online molecular biology database collection

Daniel J Rigden [1], Xosé M Fernández [2]

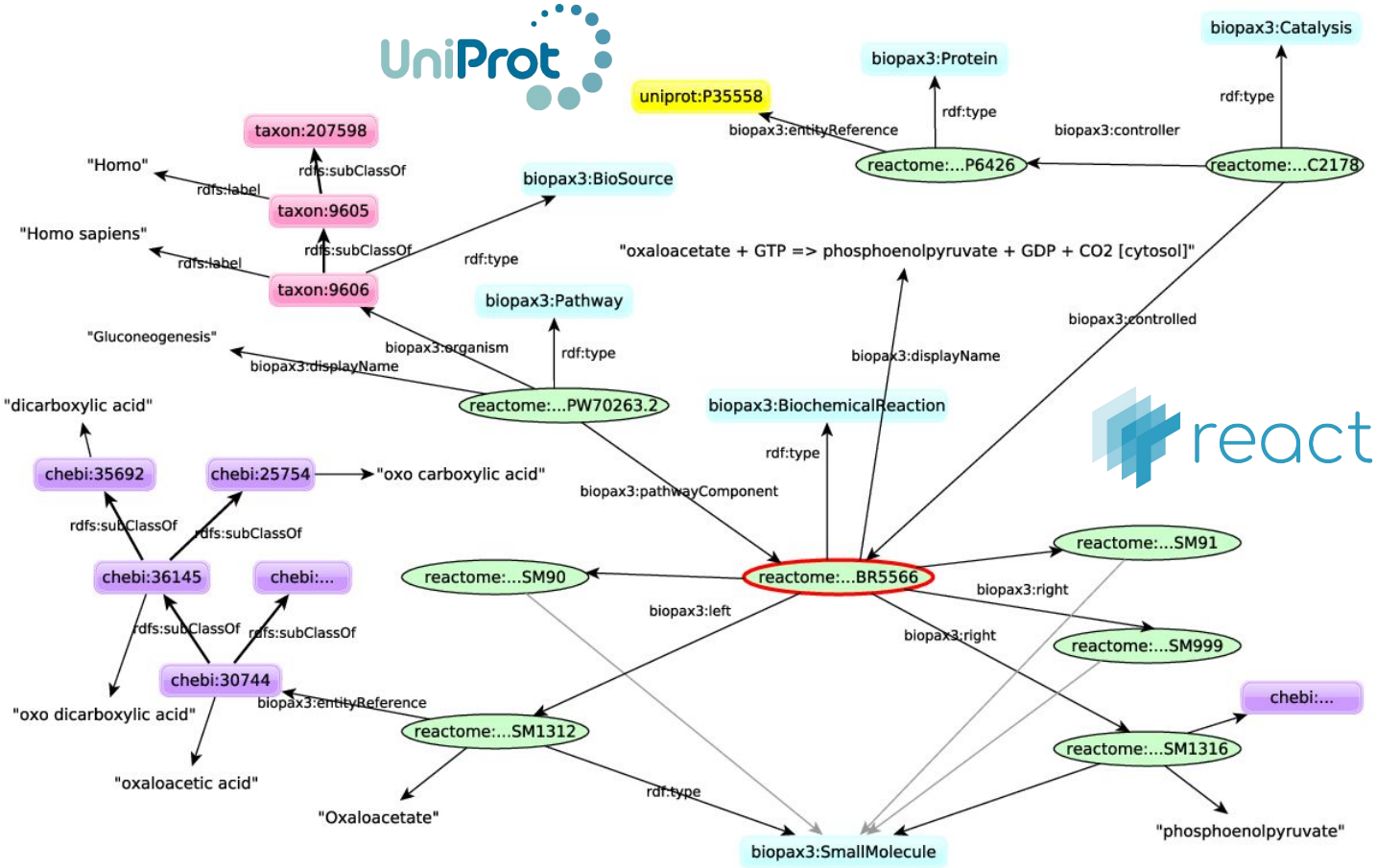
Affiliations + expand

PMID: 34986604 PMCID: PMC8728296 DOI: 10.1093/nar/gkab1195

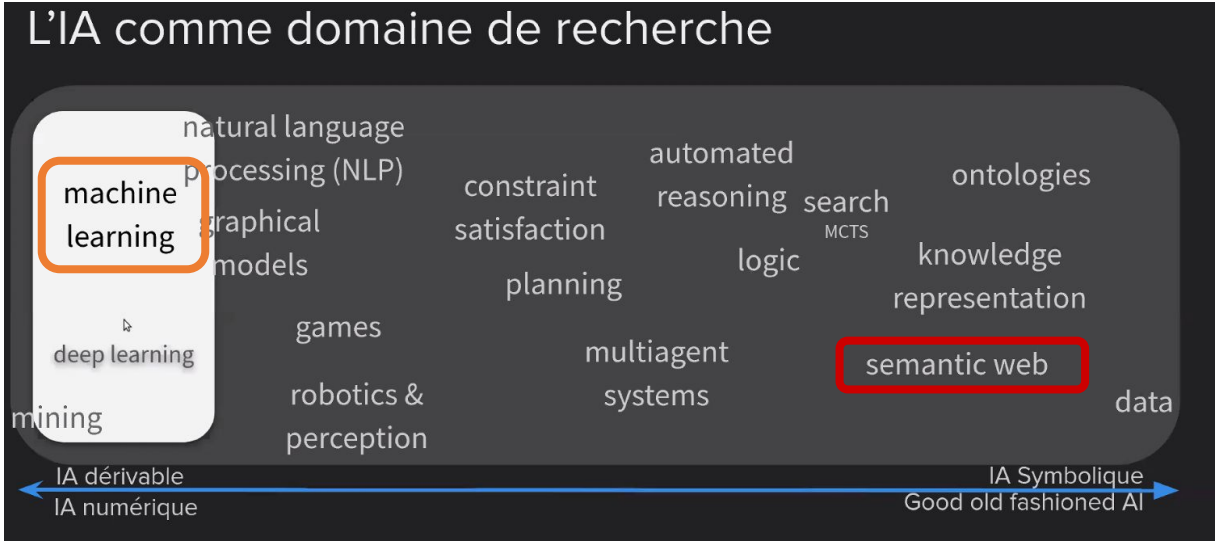
Semantic Web = framework for:

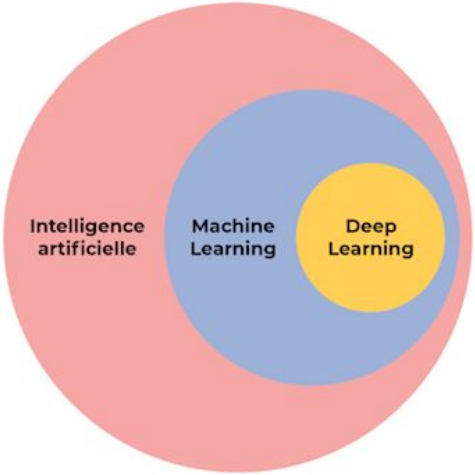
- **integrating** data and knowledge
- **querying**
- **reasoning**

Take advantage of the vast amount of available data



Artificial intelligence of course … and so ?





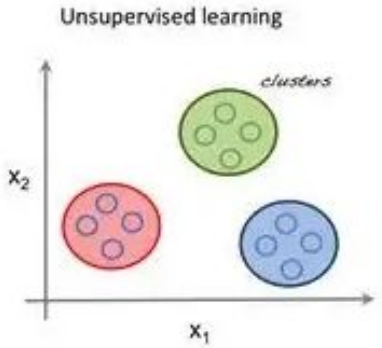
Improve our understanding of biological phenomena

Take advantage of the vast amount of available data

- **Unsupervised learning**

find hidden patterns, analyze and organize unlabelled datasets.

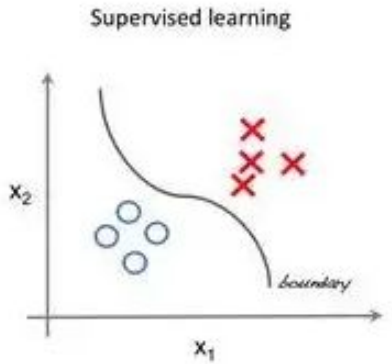
ex : clustering, dimension reduction, density estimation



- **Supervised learning**

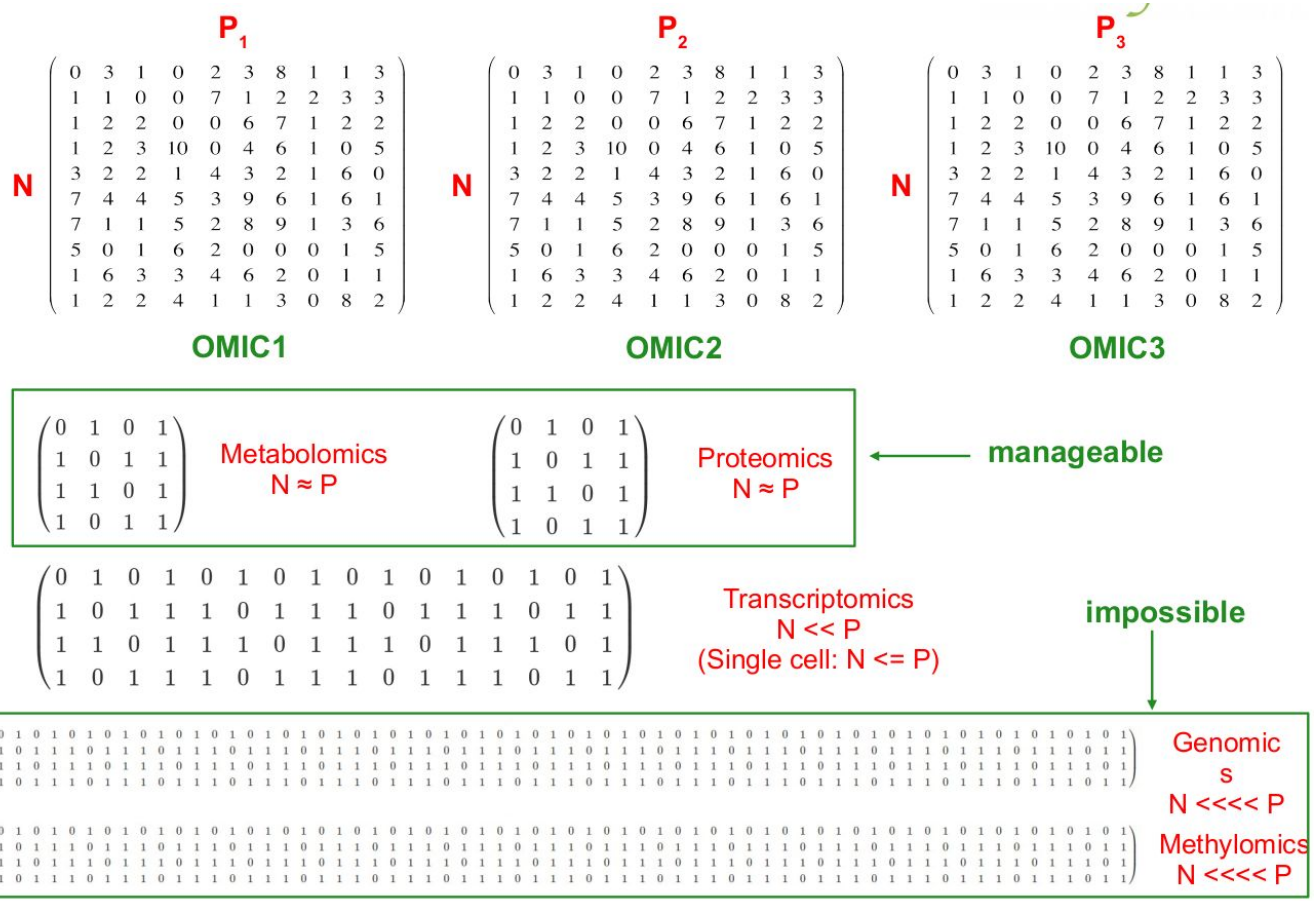
use labeled datasets and previous outputs to guess outcomes in advance (predictive model).

ex : classification task (categorical/numerical), regression (numerical)



- **Semi-supervised**

Curse of dimensionality



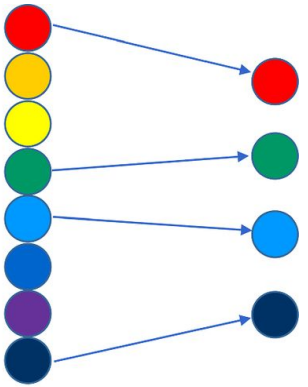
P : number of features (genes, proteins, genetic variants…)

N : number of observations (samples, cells, nucleotides…)

Feature selection

→ determine a smaller set of features minimizing (relevant) information loss

ex : filtering methods (correlation), recursive elimination, regularization

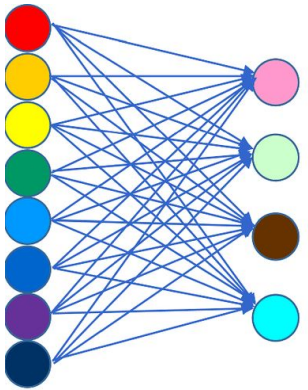


Feature extraction

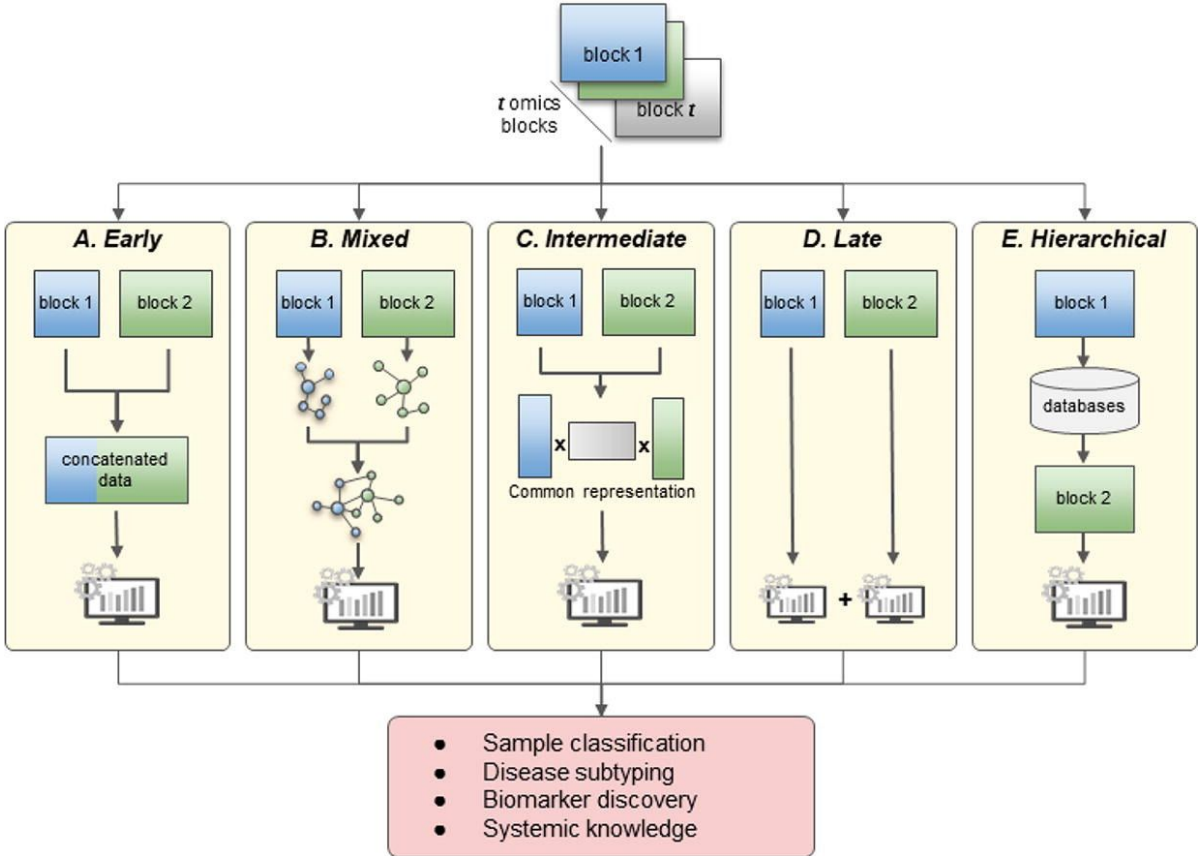
→ combine the input features into another set of variables in a linear or non-linear fashion

ex : PCA, PCoA, ICA…

+ regularization for sparse methods : sPCA, sNMF

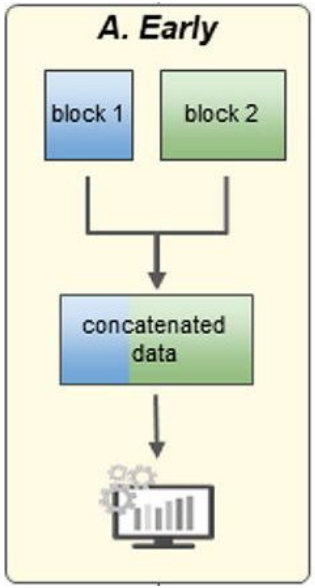


Integration strategies



Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies



Concatenate every omics datasets into a single large matrix.

Pros :

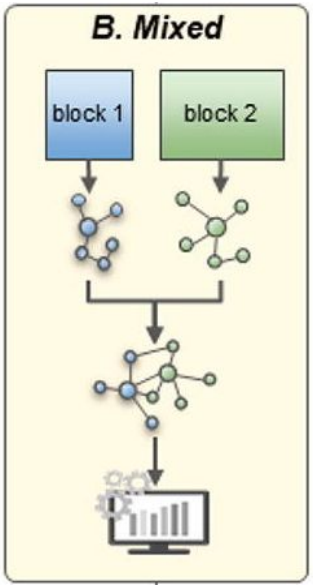
- conceptually simple
- easy implementation
- directly uncovers interactions between omics

Cons :

- technically complicated (noisy and high dimensional concatenated matrix)
- imbalanced omics datasets
- ignores the specific data distribution of each omics
- common definition space (rows or columns → samples or features)

Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies



Transform independently each omics dataset into a simpler representation before integration.

Pros :

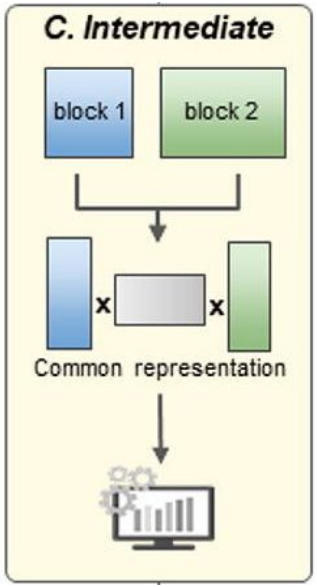
- new representation is less dimensional and less noisy
- less heterogeneity between omics
- classical approaches can be used on combined representation

Cons :

- choice of the transformation method is not trivial
- information loss during transformation
- correspondence between omics in the new representation space

Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies



Jointly integrate the multi-omics datasets without prior transformation.

Pros :

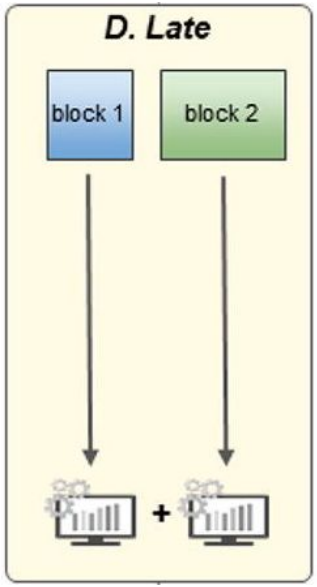
- reduce information loss
- discover the joint inter-omics structure
- highlight the complementary information in each omics

Cons :

- could require robust pre-processing step to reduce heterogeneity
- common latent space assumption

Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies



Apply machine learning models separately on each omics dataset and then combine results.

Pros :

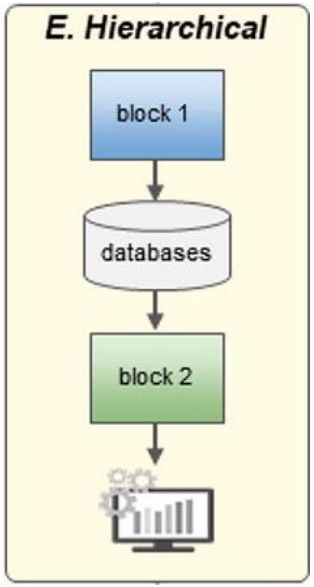
- avoid (numerous) challenges of direct omics integration
- use tools designed specifically for each omics
- classical approaches can be used to combine results

Cons :

- cannot capture inter-omics interactions
- complementarity information between omics is not exploited

Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration strategies



Include prior knowledge of omics relationships.

Pros :

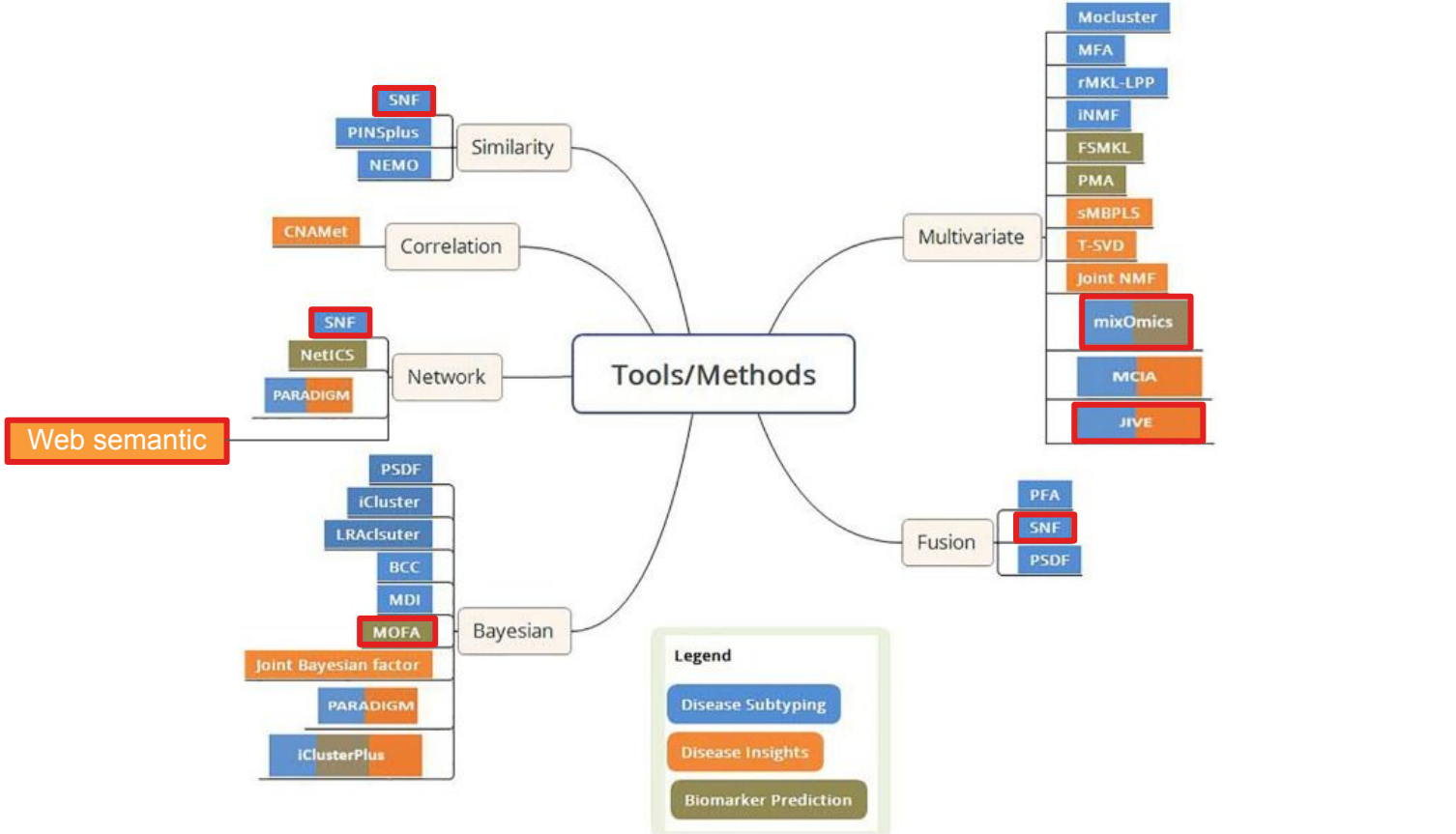
- reduced complexity (sequential integration)
- integrate external knowledge

Cons :

- less generic than previous strategies

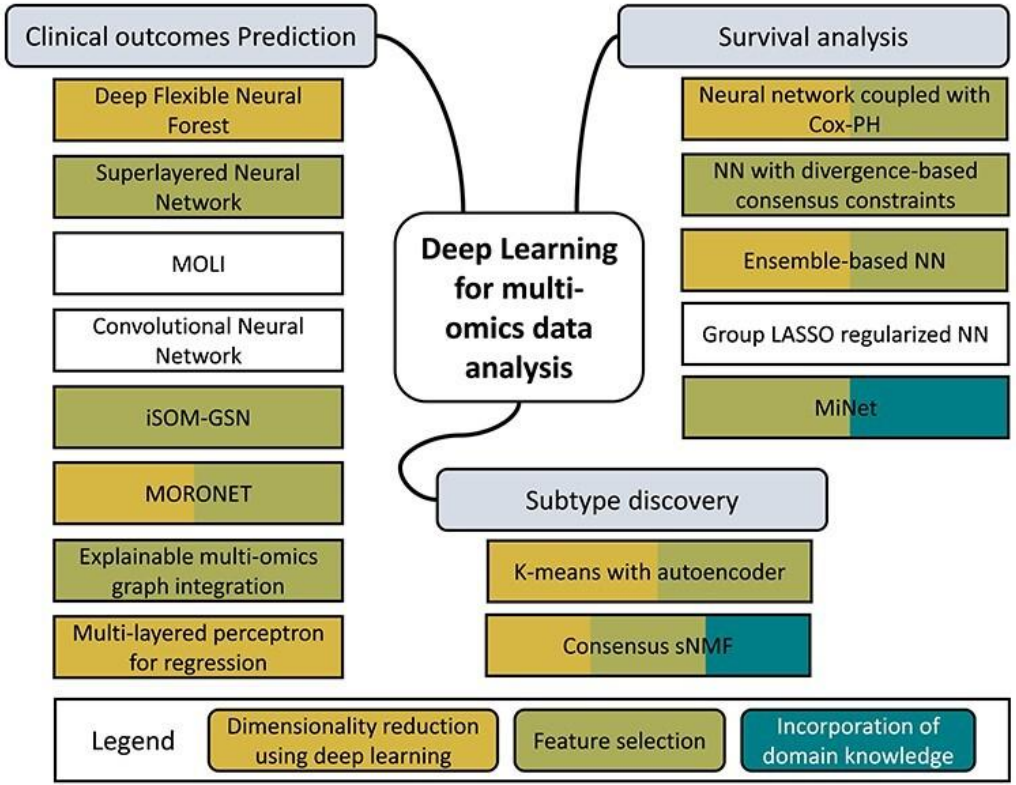
Picard M. et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021.

Integration approaches



Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights. 2020

Integration approaches : neural networks



Kang, M., Ko, E., & Mersha, T. B. (2022). A roadmap for multi-omics data integration using deep learning. Briefings in bioinformatics, 23(1).

Integration approaches : the good one ?

Integration methods are not unique

- comparisons exist... for a given application
- parametrization need expertise
- make your own comparisons/expertise
- keep an eye open

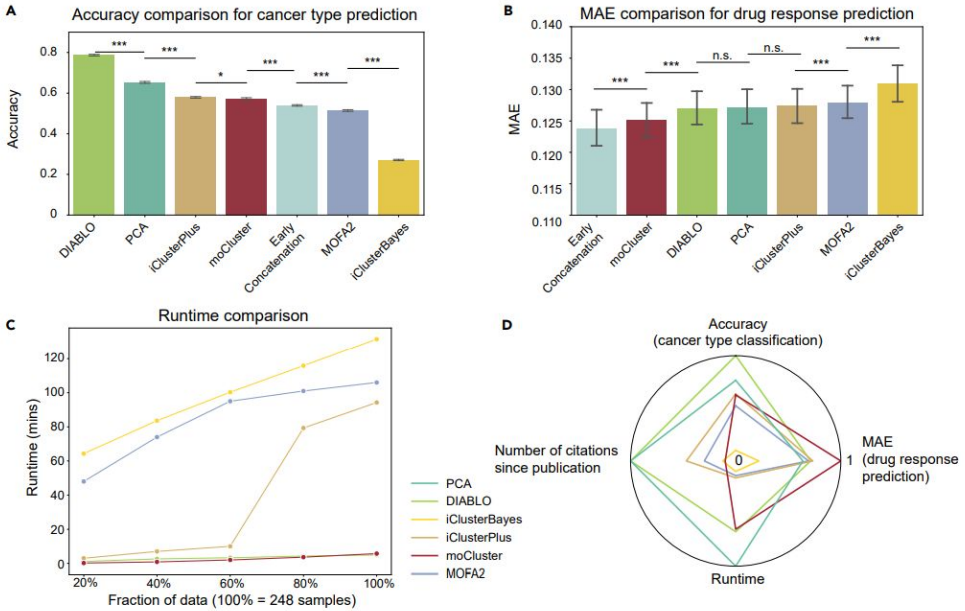


Figure 5. Benchmarking of machine learning-based integration tools using the CCLE multi-omics data

Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience. 2022

Integration methods are not magic!

You will still need to:

- carefully check design and confounding factors
- perform specific data pre-processing for each omic
- impute missing values* (different meaning → different strategy)
- choose your integration strategy based on your objective and your data (ex. matching between omics) → still no standard pipelines
- some omics bring more noise than answers

Multi-omics data portal

Table 1. Key portals for accessing publicly available multi-omics datasets

Name	URL	Omic and other data types	Notes
TCGA (Campbell et al., 2020)	https://portal.gdc.cancer.gov/	• Genomics • Epigenomics • Transcriptomics	• Tumor data • Large coverage of tumors
ICGC (Campbell et al., 2020)	https://dcc.icgc.org/	• Genomics • Transcriptomics	• Tumor data • Powerful online analytics tools
CPTAC	https://cptac-data-portal.georgetown.edu/cptacPublic/	• Proteomics	• Tumor data • The largest proteomic data portal
COSMIC Cell Lines (Iorio et al., 2016)	https://cancer.sanger.ac.uk/cell_lines	• Genomics • Epigenomics • Transcriptomics • Drug response • CRISPR-Cas9 screen	• Cancer cell line data • Manually curated • Large coverage of cell lines
DepMap (Broad, 2020)	https://depmap.org/portal/	• Genomics • Epigenomics • Transcriptomics • Proteomics • Drug response • CRISPR-Cas9 screen	• Cancer cell line data • Large coverage of omic types • Powerful online tools
COSMIC (Tate et al., 2019)	https://cancer.sanger.ac.uk/cosmic	• Genomics • Epigenomics • Transcriptomics	• Tumor data • Manually curated • Focus on genomics • Overlap with other portals

Cai Z, Poulos RC, Liu J, Zhong Q. Machine learning for multi-omics data integration in cancer. iScience. 2022

Web-applications

PaintOmics *(T. Liu et al. PaintOmics 4: new tools for the integrative analysis of multi-omics datasets supported by multiple pathway databases, Nucleic Acids Research, Volume 50, Issue W1, 2022.)*

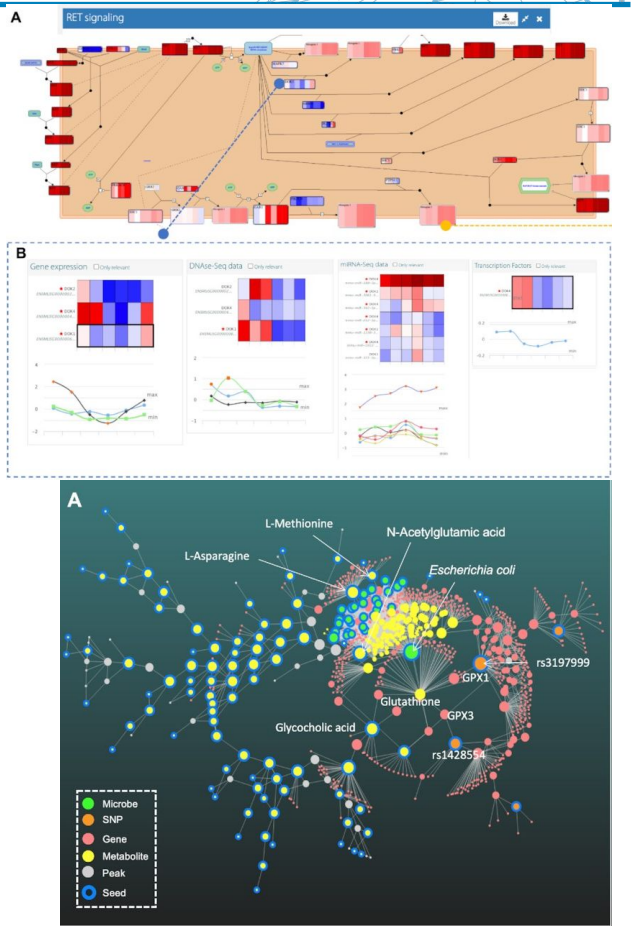
3Omics *(K. Tien-Chueh et al. 3Omics: A web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC systems biology. 7. 64, 2013)*

XCMSOnline *(EM. Forsberg et al. Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. Nat Protoc. 13(4):633-651, 2018)*

Galaxy-P project *(Galaxy-P Project. galaxyp.org.)*

OmicsNet *(G. Zhou et al., OmicsNet 2.0: a web-based platform for multi-omics integration and network visual analytics, Nucleic Acids Research, Volume 50, Issue W1, 5 , 2022.)*

...



References

Misra BB, Langefeld CD, Olivier M, Cox LA. Integrated Omics: Tools, Advances, and Future Approaches. J Mol Endocrinol, 2018.

Subramanian I, Verma S, Kumar S, Jere A, Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights, 2020.

Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J., 2021.

Benfeitas R, Viklund J, Ash706, Robinson J, Manoharan L, Fasterius E, Oskolkov N, Francis R, Anton M. (2020). NBISweden/workshop_omics_integration: Lund, 2020/10/05 (Version course2010). Zenodo. https://doi.org/10.5281/zenodo.4084627

Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, Milanesi L. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics, 17 Suppl 2(Suppl 2):15, 2016.

Ritchie, M., Holzinger, E., Li, R. et al. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet 16, 85–97, 2015.