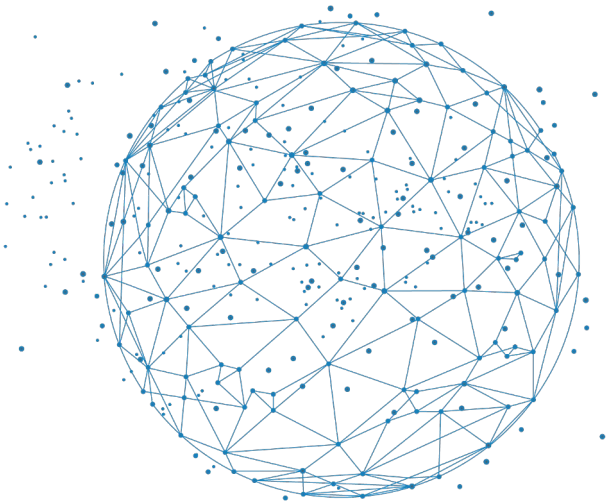
First edition 2023 in Fréjus

ÉCOLE THÉMATIQUE ETBII BIOINFORMATIQUE INTÉGRATIVE IFB



Omics integration - General aspects

Jimmy Vandel

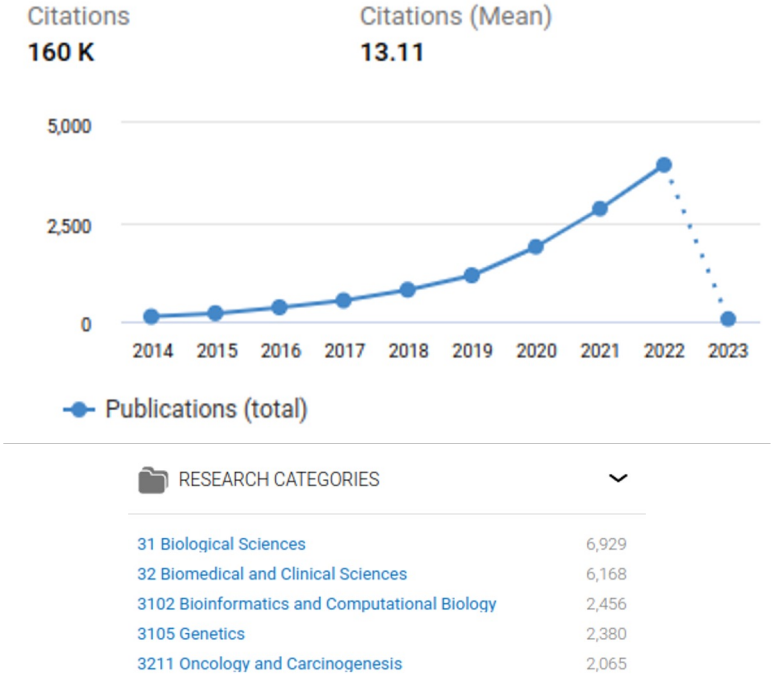
DOI version final

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

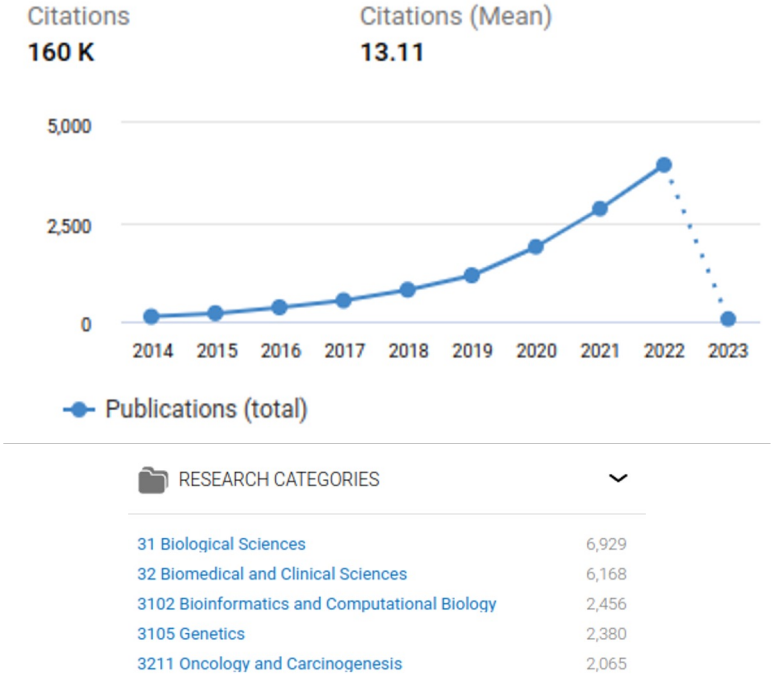
cnrs INRAE Inserm cea

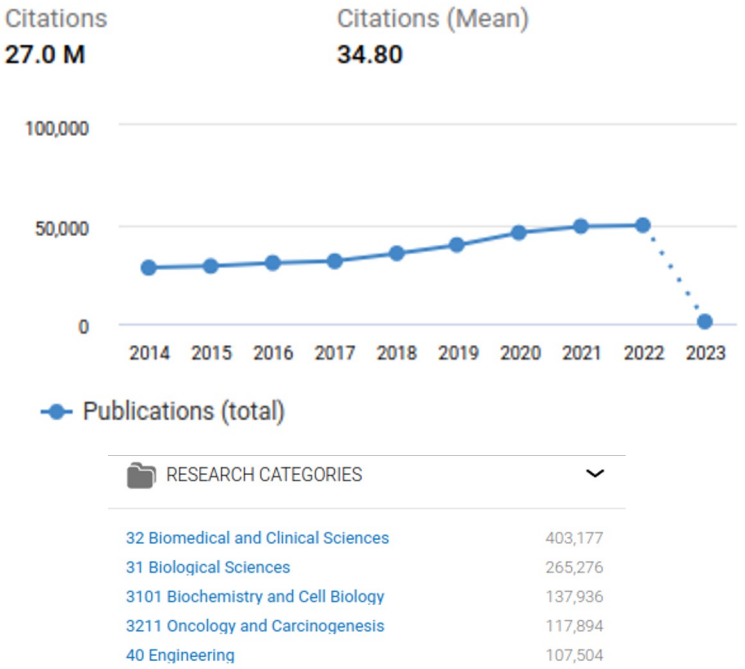
“Multi-omics” citations



https://app.dimensions.ai/discover/publication (8th Jan. 2023 : 132,863,611 referenced publications)

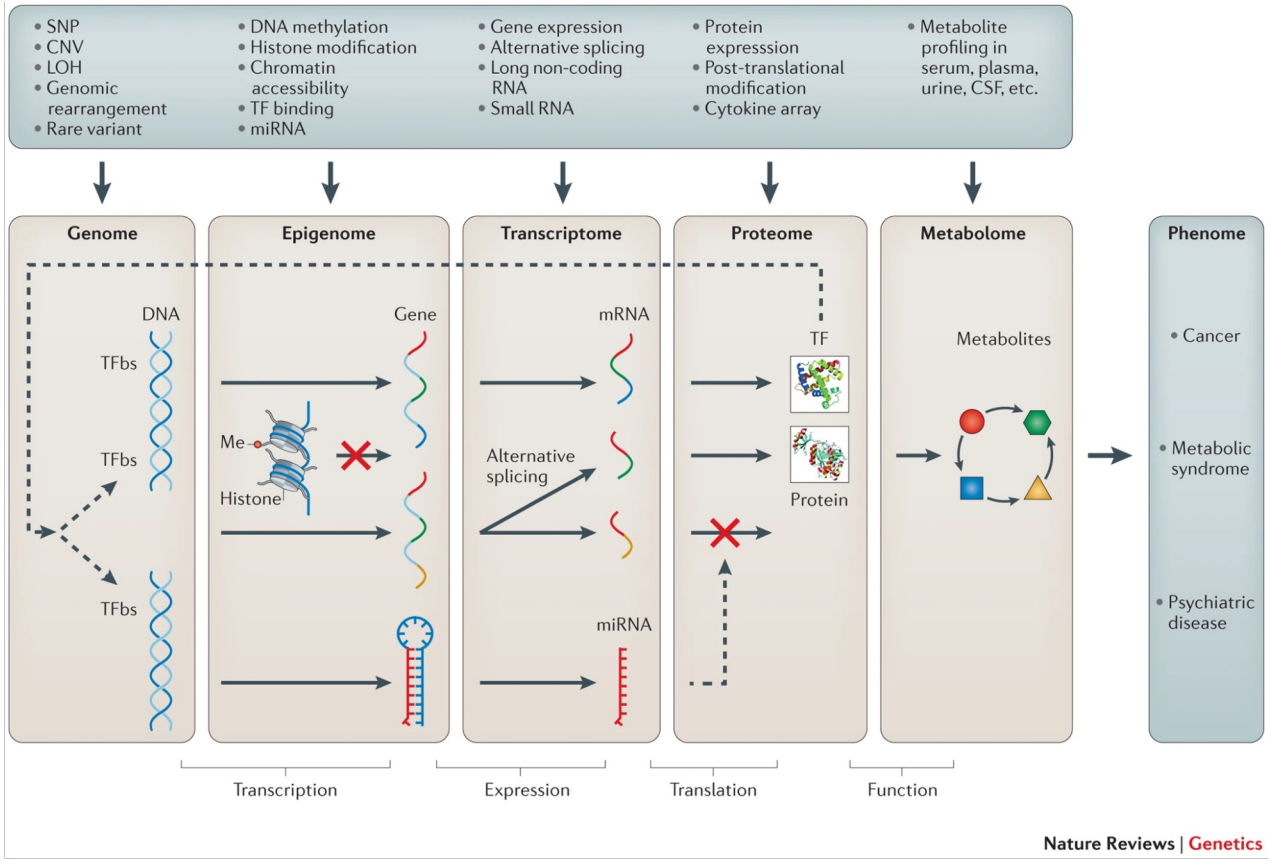
“Multi-omics” citations





https://app.dimensions.ai/discover/publication *(8th Jan. 2023 : 132,863,611 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 data ?

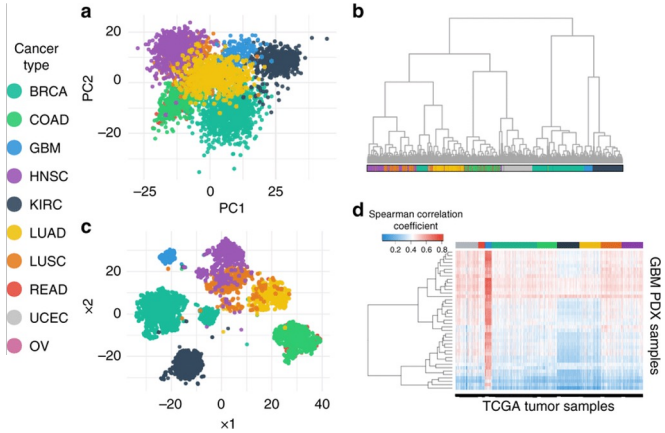
- clinical data
- imaging data (full data or extracted characteristics)
- new omics fields : fluxomics, ionomics, microbiomics, glycomics…

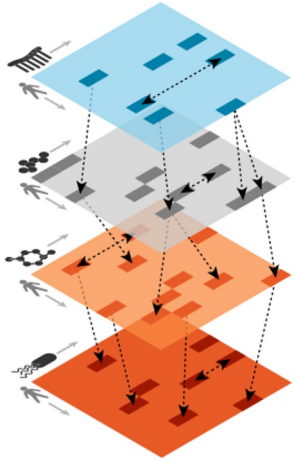
- biological knowledge : DNA/protein, protein/protein interactions

→ a priori in model definition/construcion

Integration: why?

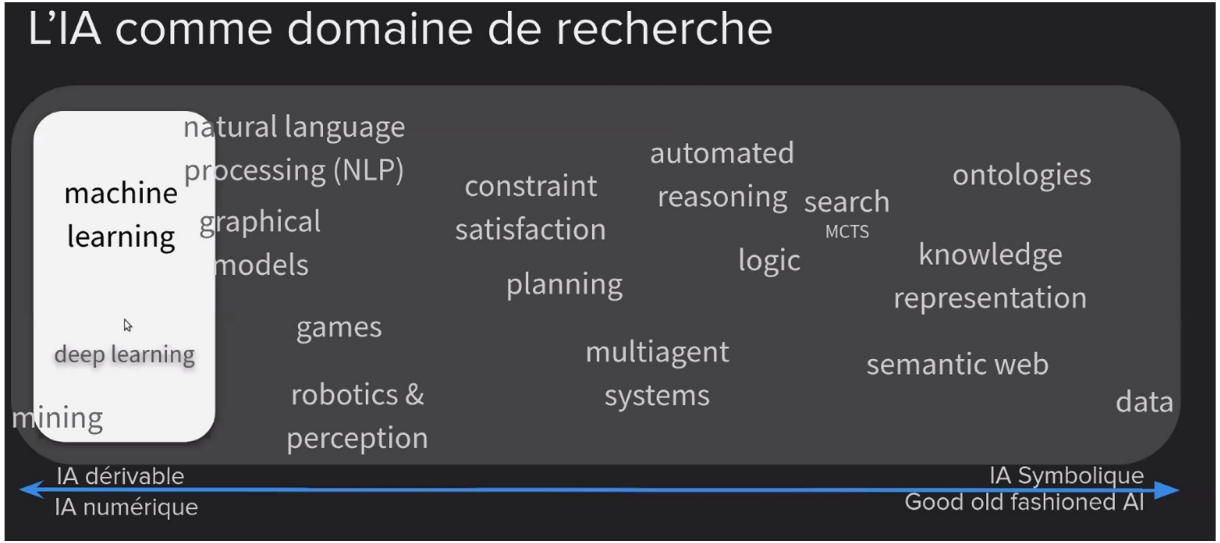
- Disease subtyping and classification
- Biomarkers prediction : diagnostic, disease drivers
- Deep insights into disease biology

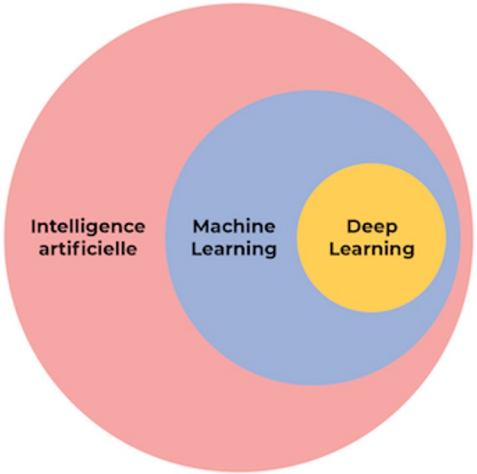




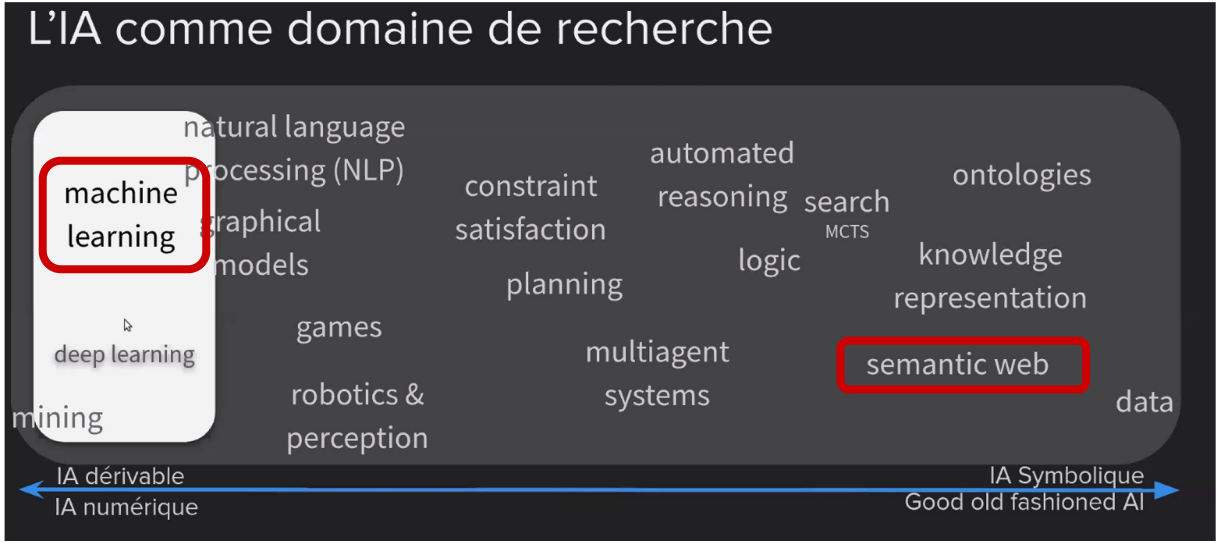
Vasileios et al (2018). Drug and disease signature integration identifies synergistic combinations in glioblastoma. Nature Communications. 9.

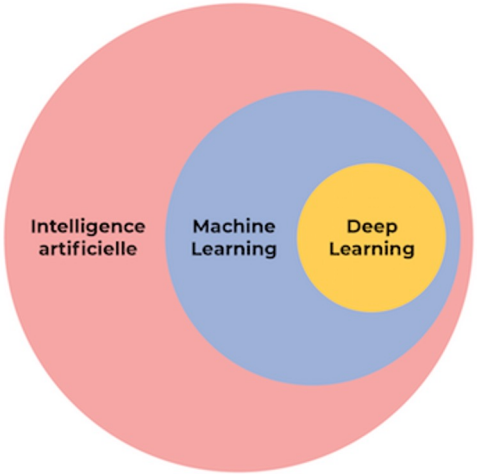
Artificial intelligence of course … and so ?



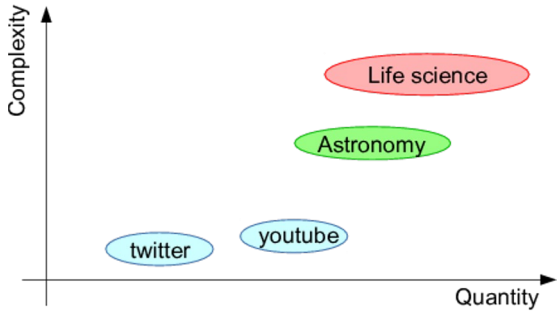


Artificial intelligence of course … and so ?





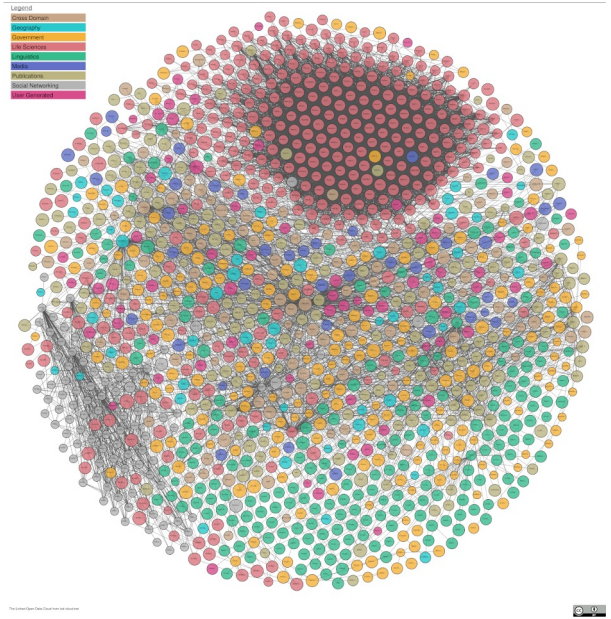
Integration: how? (hint: with the Semantic Web)



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]

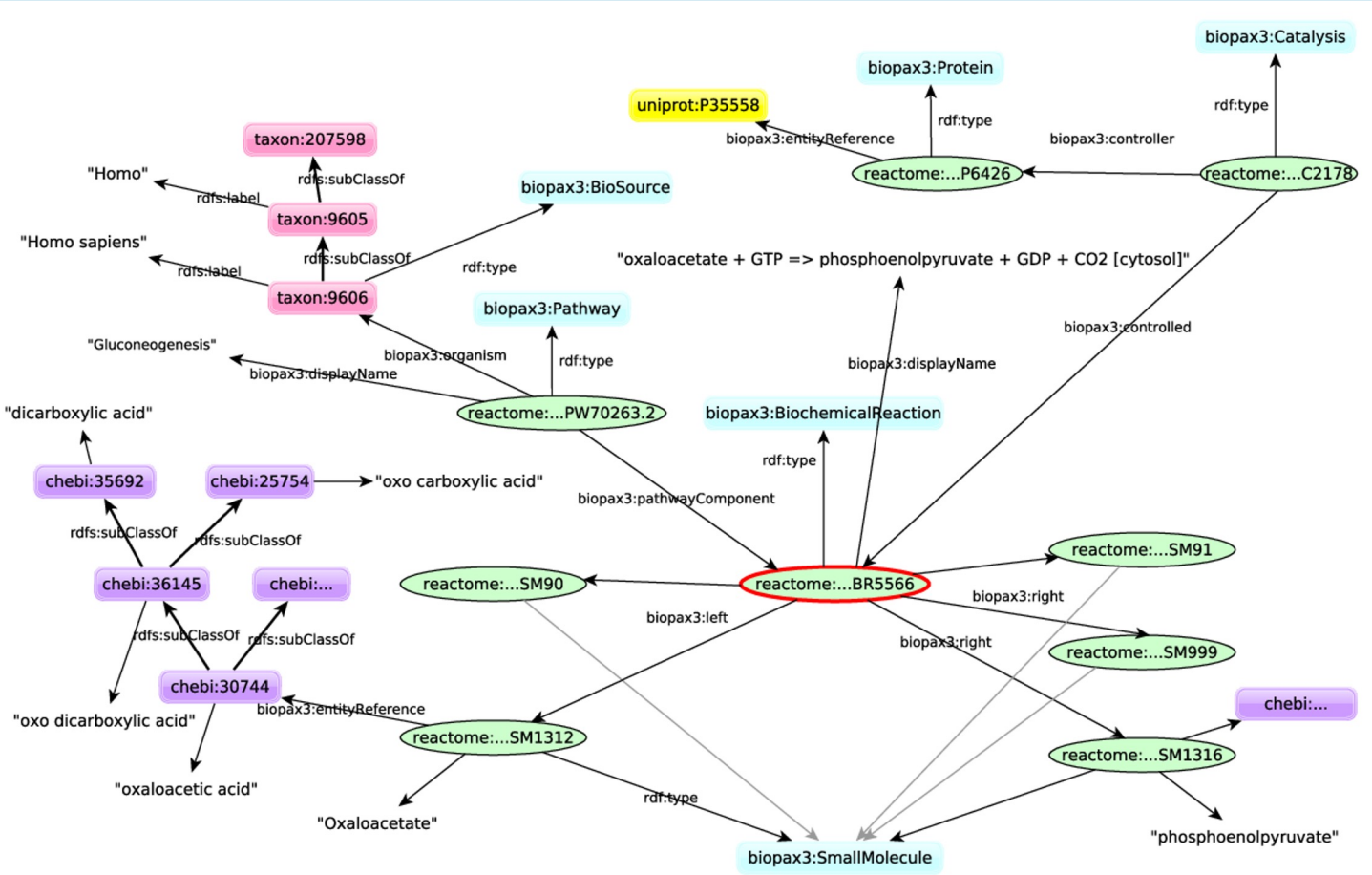
Affiliations + expand

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

Semantic Web = framework for:

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

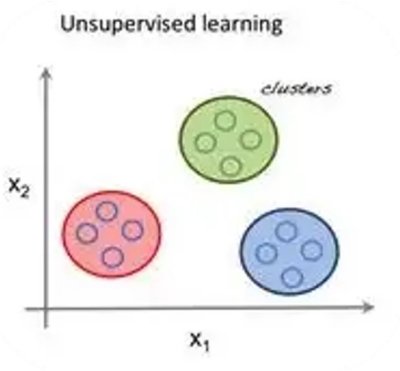
Integration: how? (hint: with the Semantic Web)



Unsupervised learning

find hidden patterns, analyze and organize unlabelled datasets.

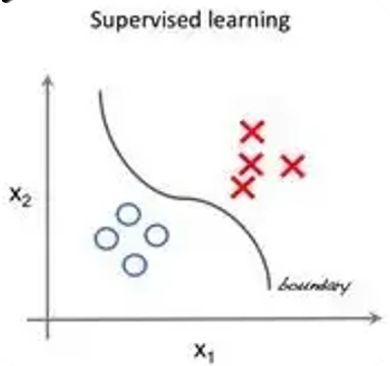
ex : clustering, dimension reduction, density estimation



Supervised learning

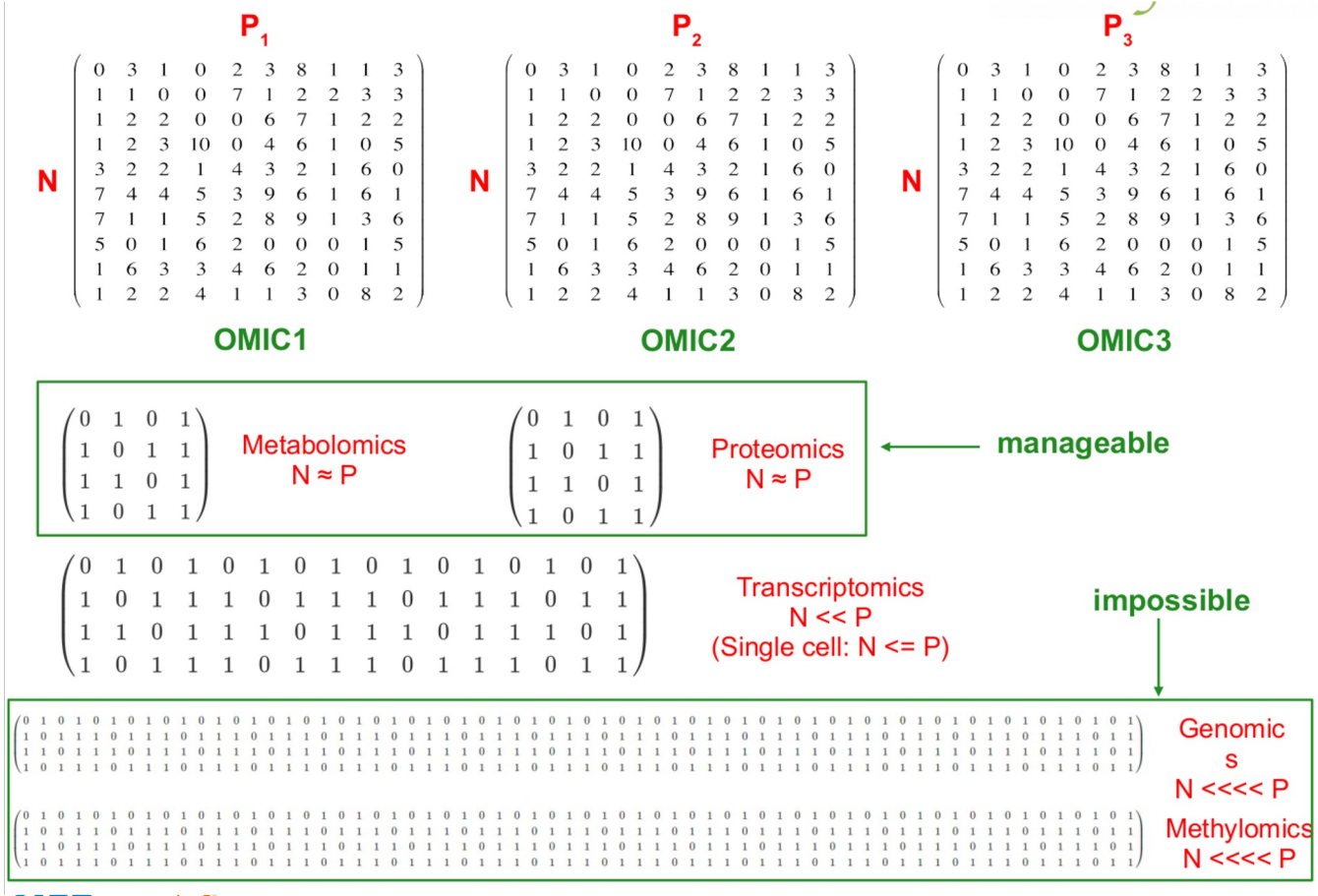
use labeled datasets and previous outputs to guess outcomes in advance (predictive model).

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



Semi-supervised

Curse of dimensionality



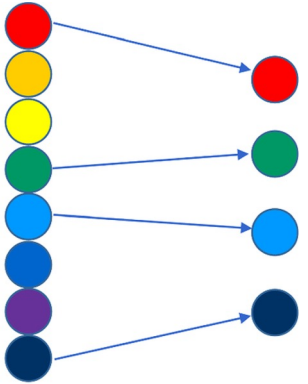
P : number of features (genes, proteins, genetic variants…)

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

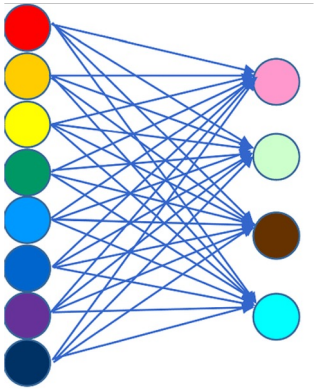


Feature extraction

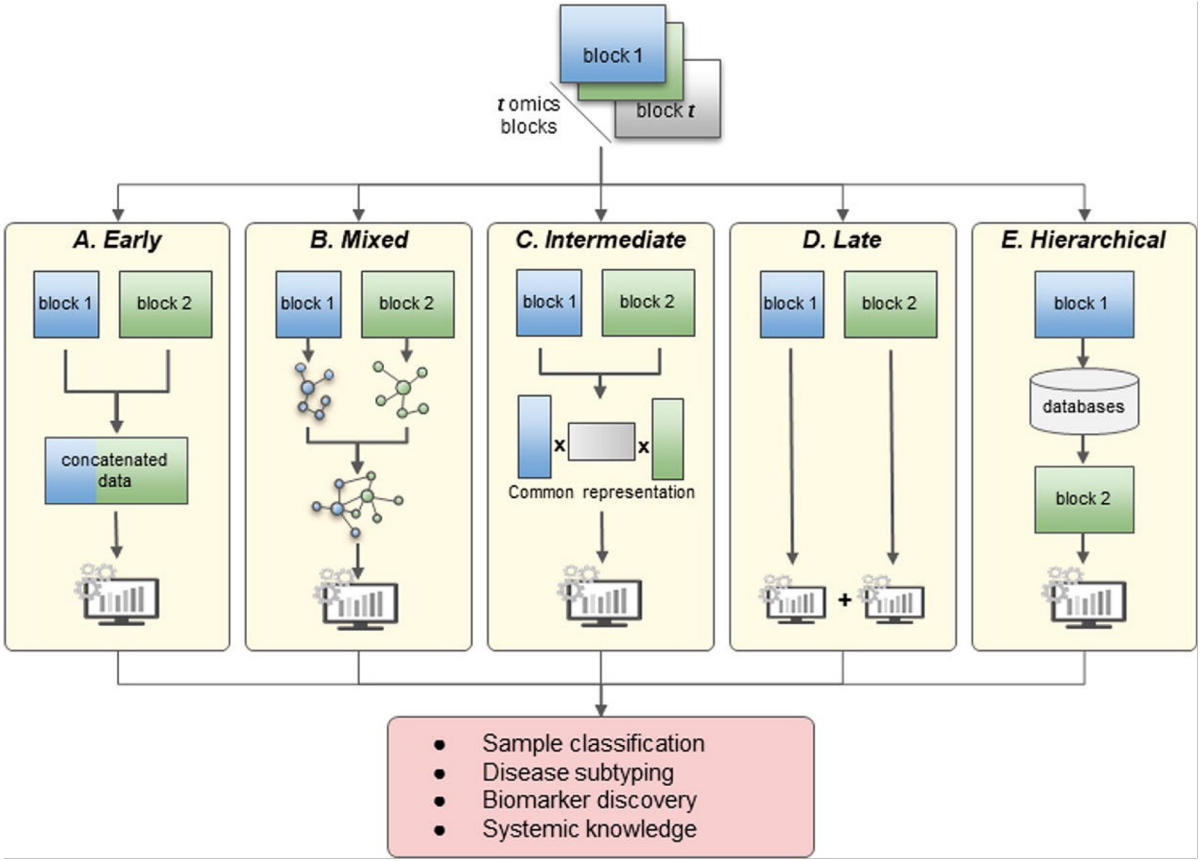
→ combine the input features into another set of variables in a linear or non-linear fashion

ex : PCA, PCoA, ICA…

+ regularization for sparse methods : sPCA, sNMF

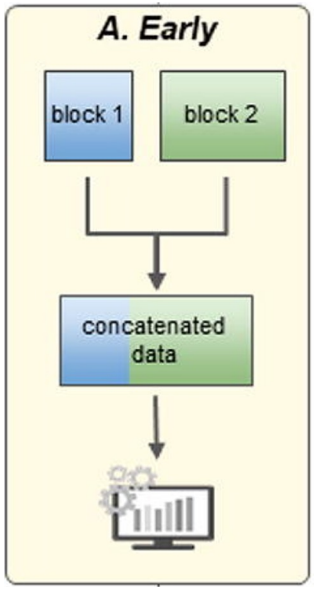


Integration strategies



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 :

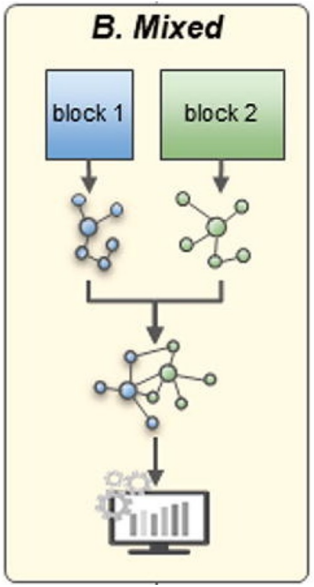
- conceptually simple
- easy implementation
- directly uncovers interactions between omics

Cons :

- technically complicated (noisy and high dimensional concatenated matrix)
- imbalanced omics datasets
- ignores the specific data distribution of each omics

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 :

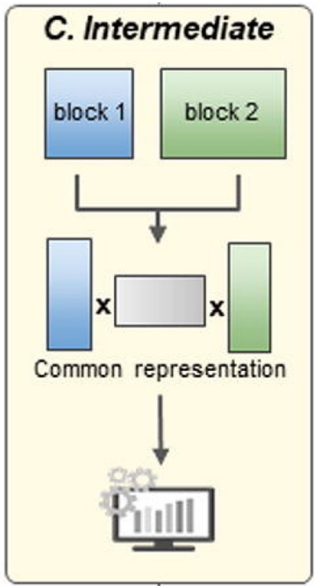
- new representation is less dimensional and less noisy
- less heterogeneity between omics
- classical approaches can be used on combined representation

Cons :

- choice of the transformation method is not trivial
- information loss during transformation

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 :

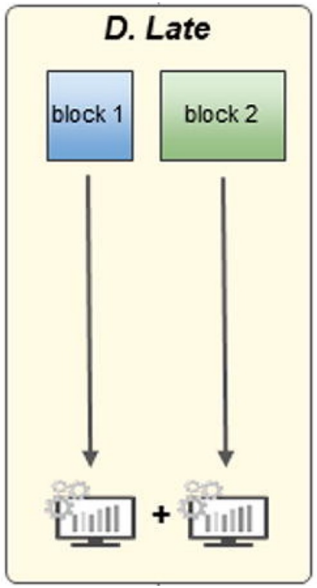
- reduce information loss
- discover the joint inter-omics structure
- highlight the complementary information in each omics

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 :

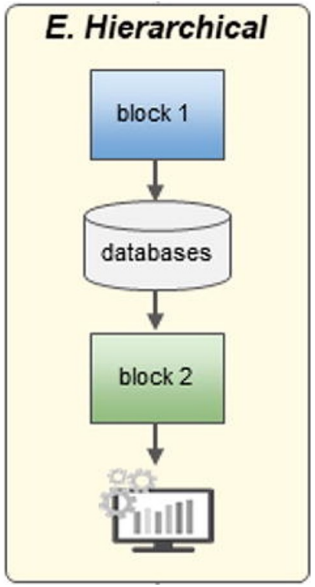
- avoid (numerous) challenges of direct omics integration
- use tools designed specifically for each omics
- classical approaches can be used to combine results

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 :

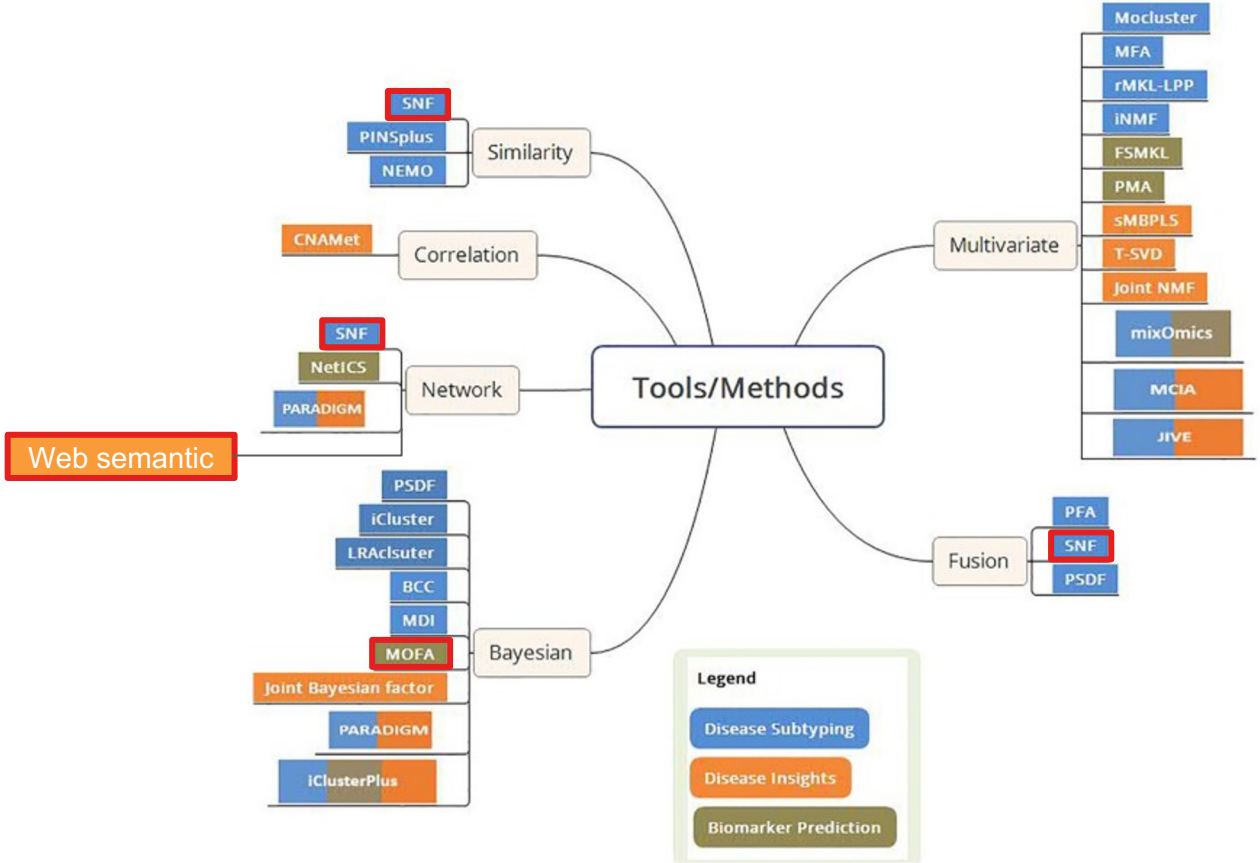
- reduced complexity (sequential integration)
- integrate external knowledge

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 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

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.)*

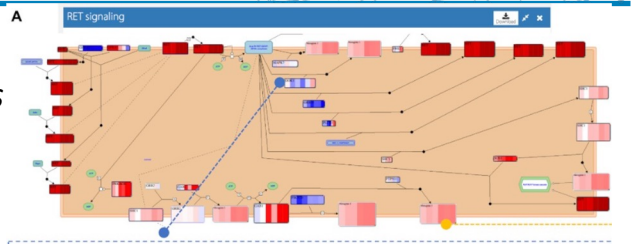
3Omics *(K. Tien-Chueh et al. 3Omics: A web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC systems biology. 7. 64, 2013)*

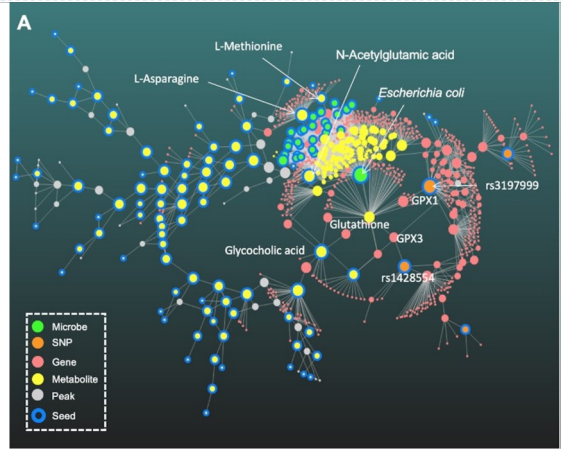
XCMSOnline *(EM. Forsberg et al. Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. Nat Protoc. 13(4):633-651, 2018)*

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.