Biopsia liquida e neoplasie ematologiche: disegnando il futuro
Sessione 6: Le tecnologie di ultima generazione

Le sfide dell'analisi informatica di CNAs, SNVs e single-cell data

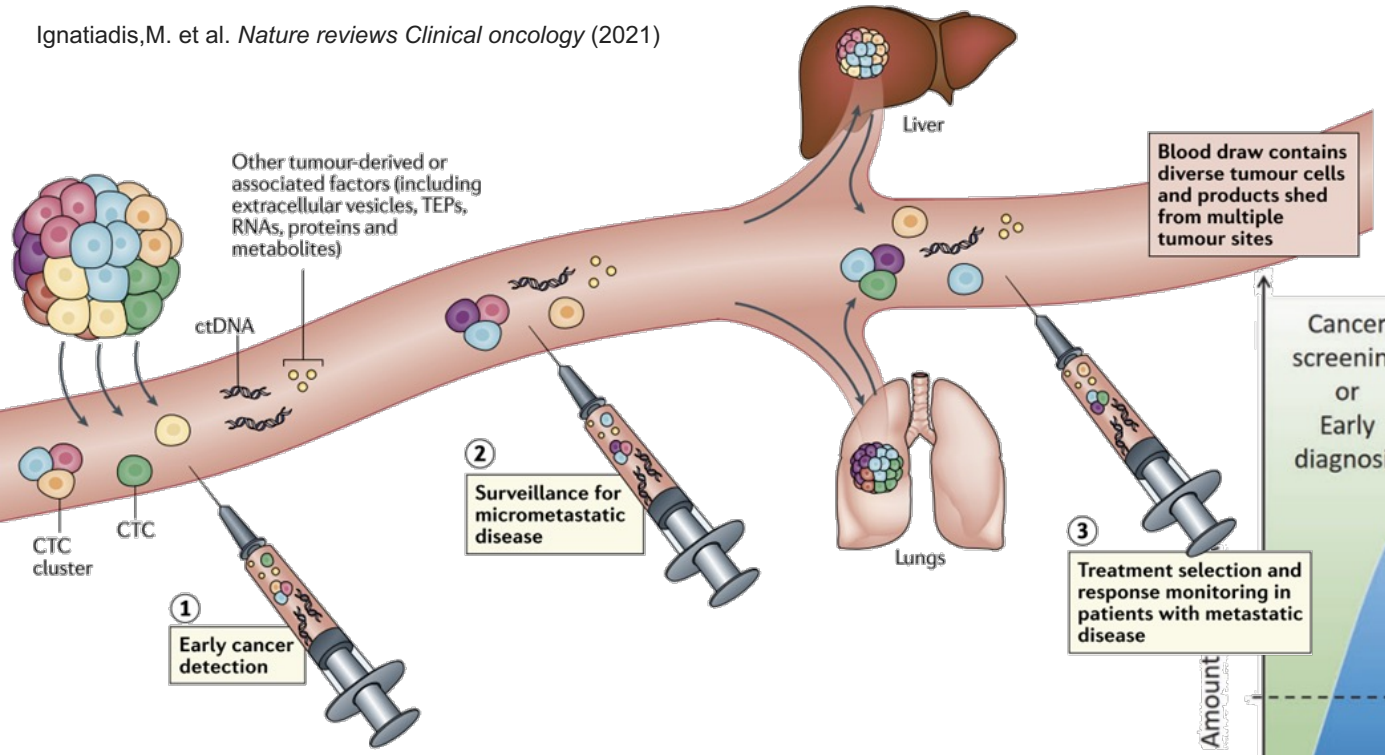
Andrea Poletti, PhD
IRCCS Azienda Ospedaliero Universitaria di Bologna S. Orsola
University of Bologna, DIMEC

Disclosures of Name Surname

Company name	Research support	Employee	Consultant	Stockholder	Speakers bureau	Advisory board	Other

Analysis of liquid-biopsy data: a multiple biomarkers dynamic scenario

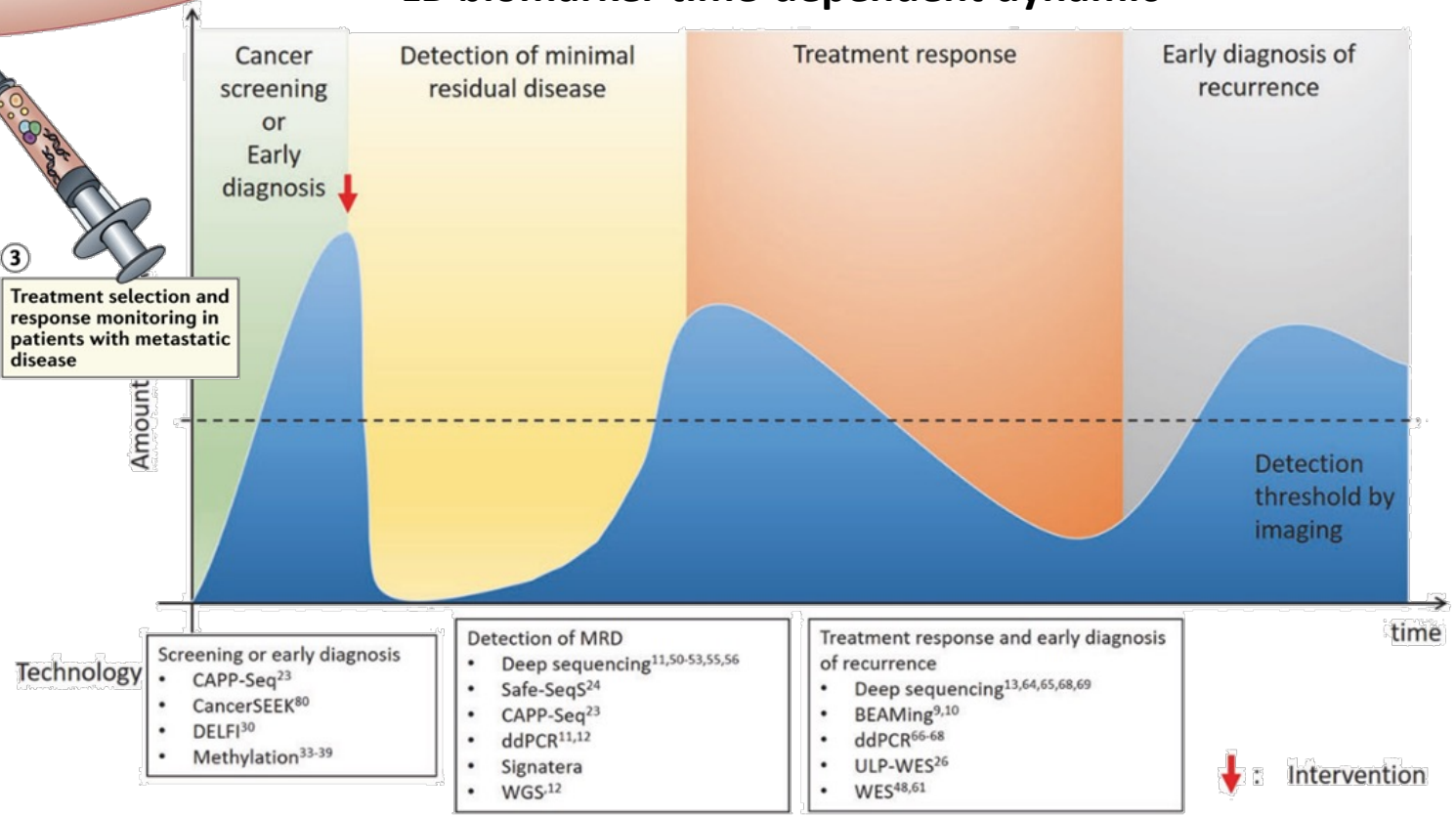
Ignatiadis,M. et al. *Nature reviews Clinical oncology* (2021)



Aims:

1) Early Cancer detection
2) MRD detection
3) Treatment response monitoring
4) Early derection of disease relapse

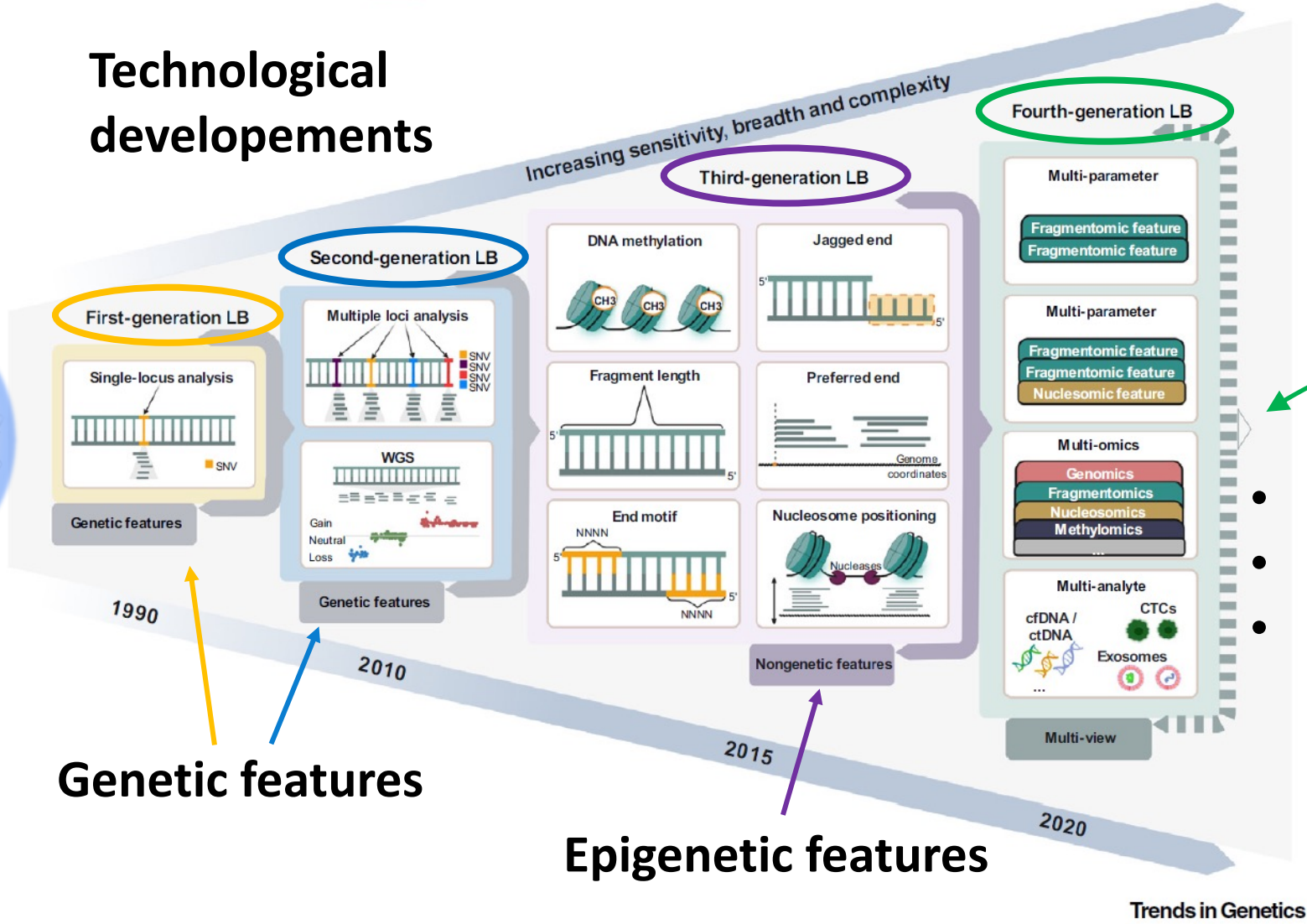
LB biomarker time-dependent dynamic



Watanabe, K., et al. *Journal of Human Genetics* (2021)

The technological (r)evolution of Liquid Biopsy analysis methods

Technological developements



Genetic features

Epigenetic features

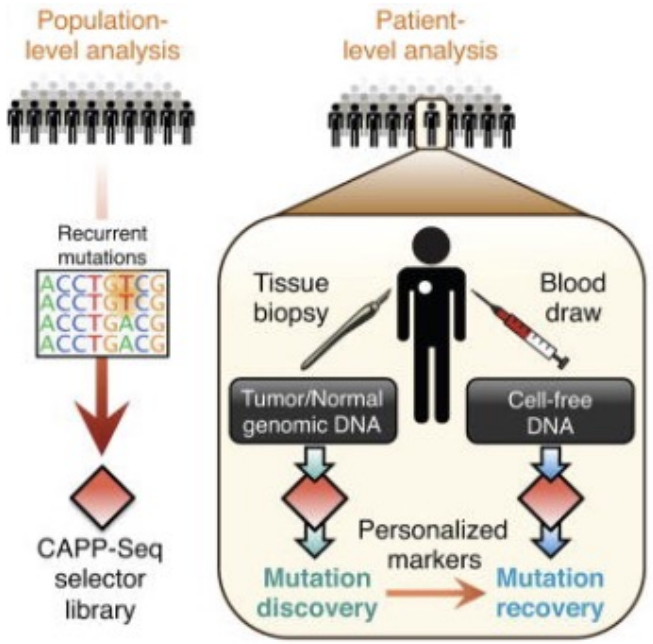
Multi-omics

- Multi-features
- Genetic + epigenetic
- Multi-analyte (e.g. cfDNA + CTCs…)

Moser, T., et al. *Trends in Genetics* (2023)

SECOND generation Liquid Biopsy methods

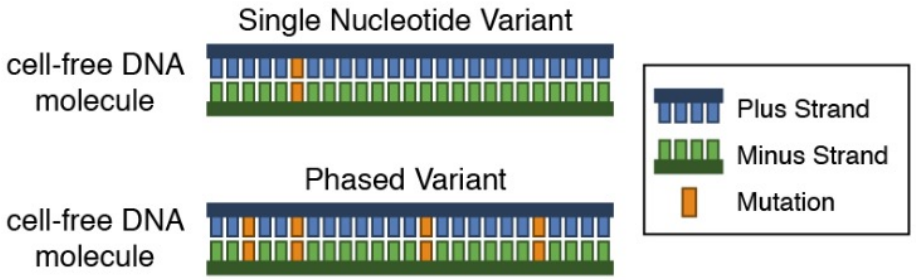
CAPP-Seq



Feature: genomic (mutations)
Strength: good LOD (10^-4)
Limit: requires prior information (population and patient)

Newman, A.M., et al. *Nature medicine* (2014)

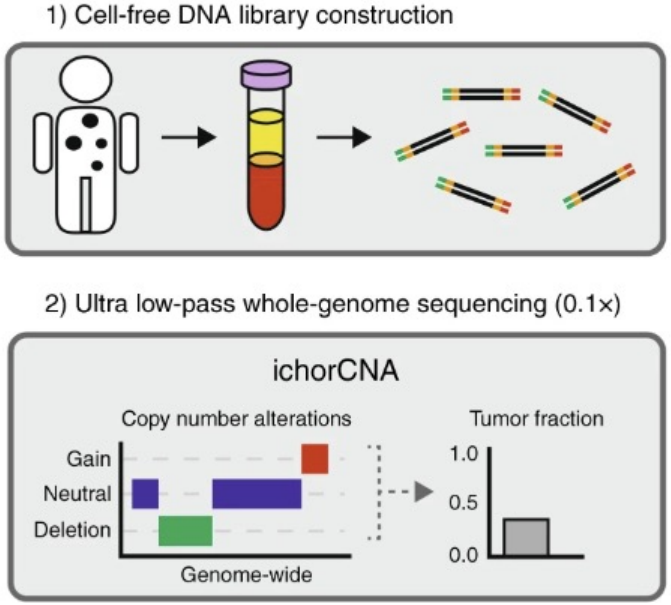
PhasED-Seq



Feature: genomic (mutations)
Strength: excellent LOD (10^-6)
Limit: requires very specific prior information (population and patient)

Kurtz, D. et al. *Nature biotechnology* 39.12 (2021)

IchorCNA



Feature: genomic (CNA)
Strength: cheap, ULP-WGS ≥0.1X
Limit: LOD ≥ 3%, probabilistic results

Adalsteinsson, A., et al. *Nature communications* (2017)

THIRD generation Liquid Biopsy methods

Features:
Epigenomic
(methylation, fragmentomic, etc.)

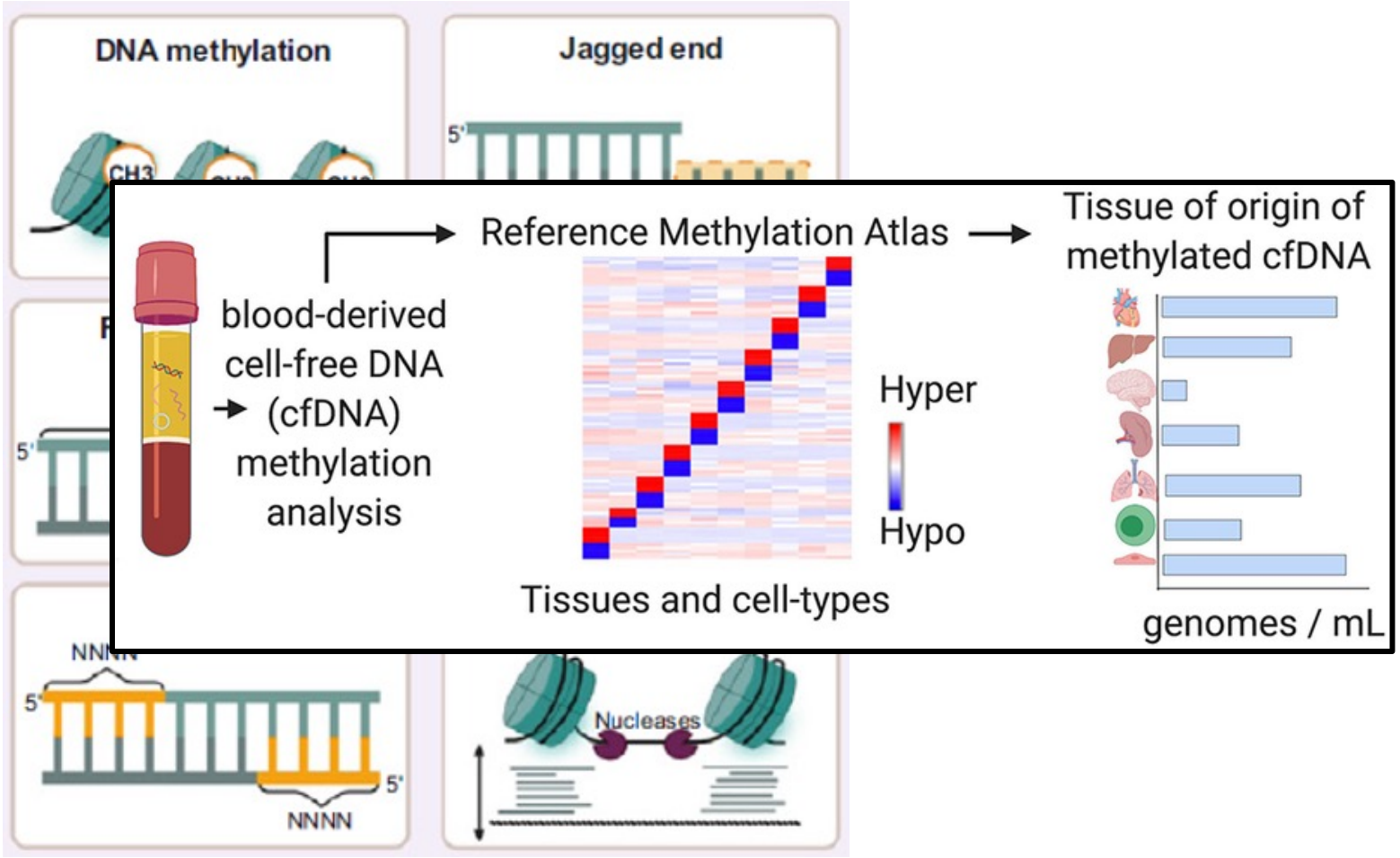
Strengths:

- **Deconvolution** of cfDNA data to obtain cell-of-origin (COO).
- **High-dimensional**, different features in a single assay.

Limits:

- **Standardization** challenges
- COO reference atlases are still evolving.
- **Biological confounders**: comorbidities, inflammation, lifestyle, etc… causing cfDNA release in blood.

Multiple epigenetic features of cfDNA



Deconvolution

Stanley, Kate E., et al. *Nature Communications* (2024)

FOURTH generation Liquid Biopsy methods

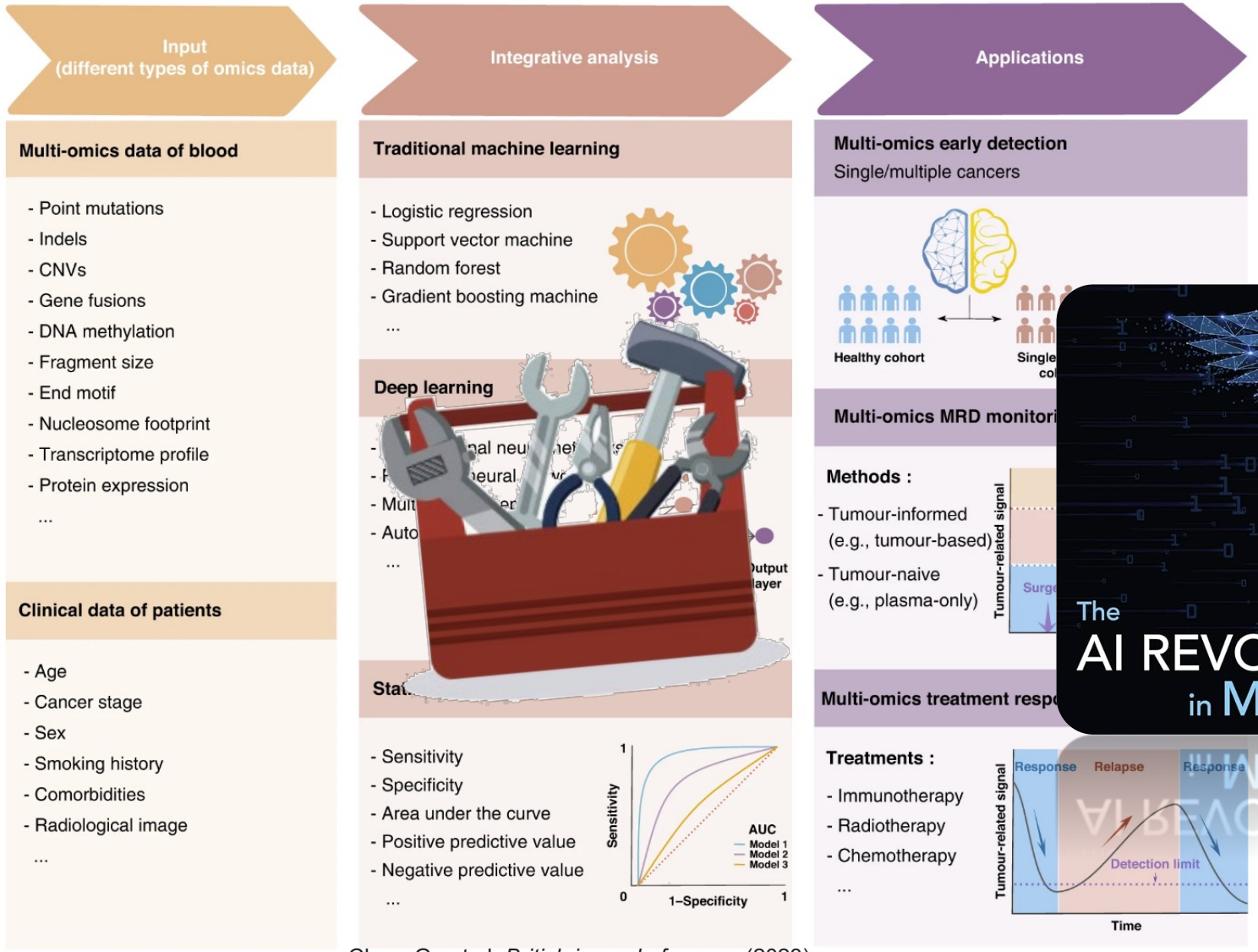
Features:
multi-omics
(genomics + epigenomics)

Strengths:

- **Multi-analyte & multi-modal**
- **Detection of weak/hidden signals**
- Analysis of complementary data
- Complexity reduction (automatic feature learning)

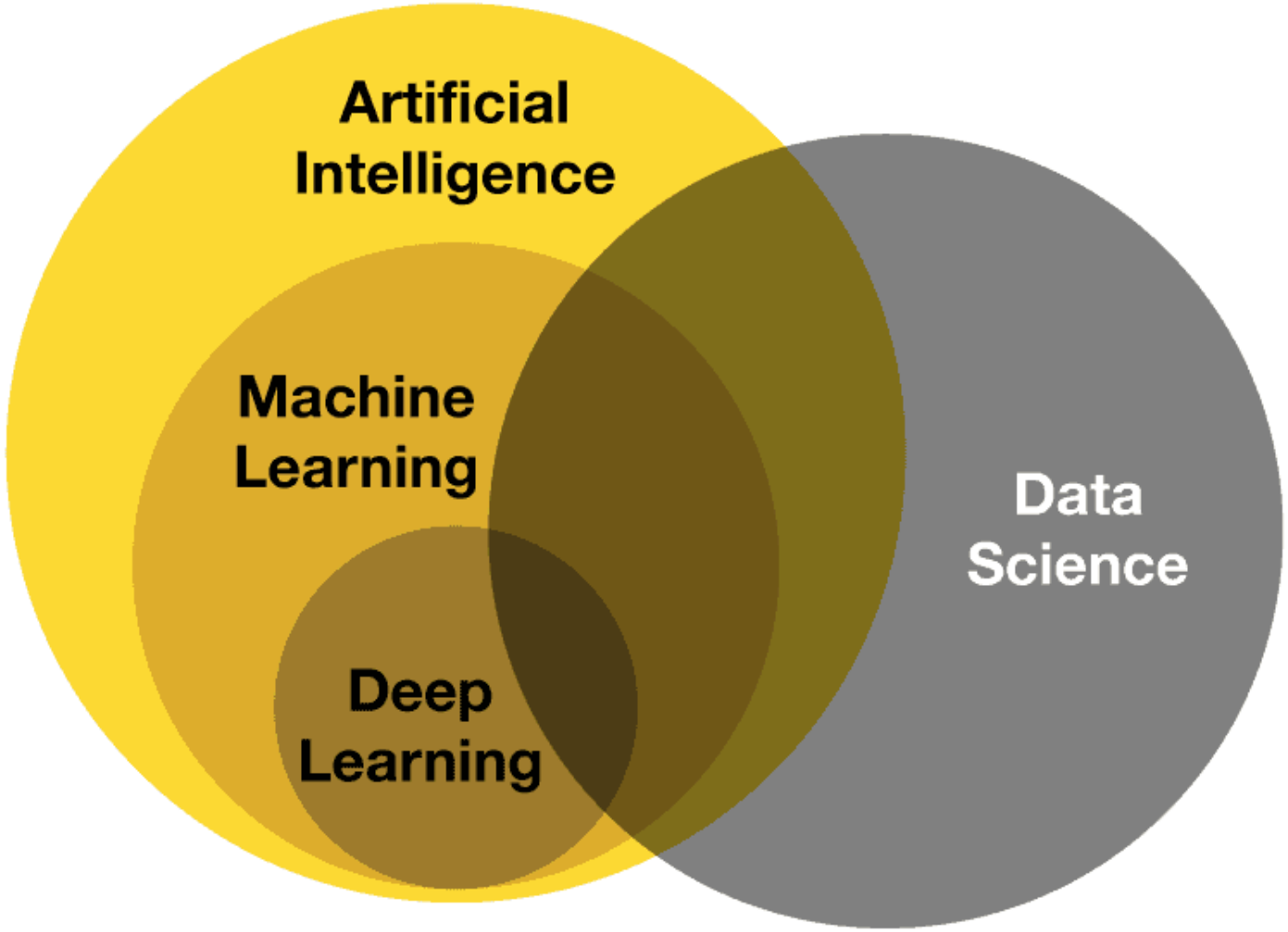
Limits:

- **Data quantity**
- **AI explainability**
- Standardization Issues
- Computational requirements



Chen, G., et al. *British journal of cancer* (2023)

DATA SCIENCE : AI, Machine Learning, Deep-learning



AI definition:

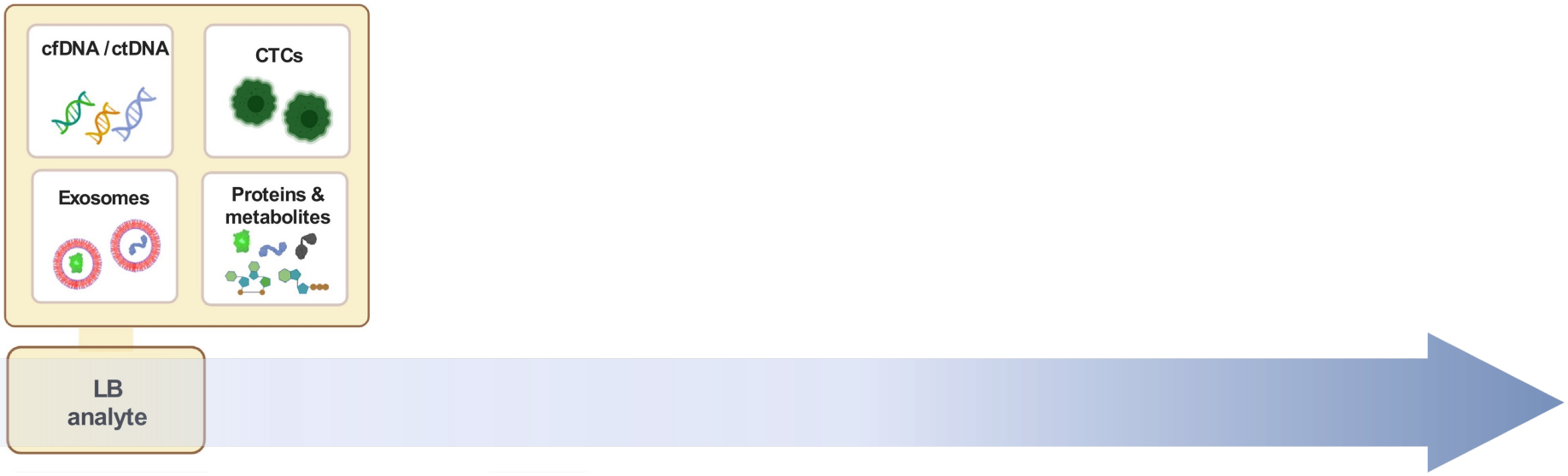
"Computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages."

(Source: Oxford Languages)

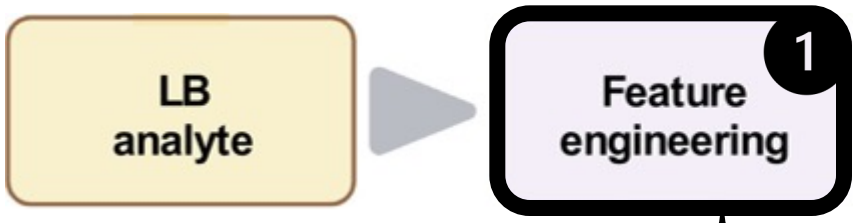
Data science:

the study of data to extract meaningful insights.

The CHALLENGES of the fourth-gen LB bioinformatic analyses



Moser, T., et al. *Trends in Genetics* (2023)



“Transforming Raw Signals into Meaningful Clues”

- In liquid biopsy, we collect many features (mutations, CNA, methylation, fragment lengths, etc.) from many analytes (like CTCs, exosomes, metabolites).
- Each signal is a **small piece of the overall puzzle**.
- **Challenge 1**: turn scattered data points (“raw materials”) into useful information (“meaningful clues”). We do this by extracting, combining, and encoding features that can capture underlying biology (e.g., nucleosome positioning or methylation blocks).

LB analyte

Feature engineering 1

Example

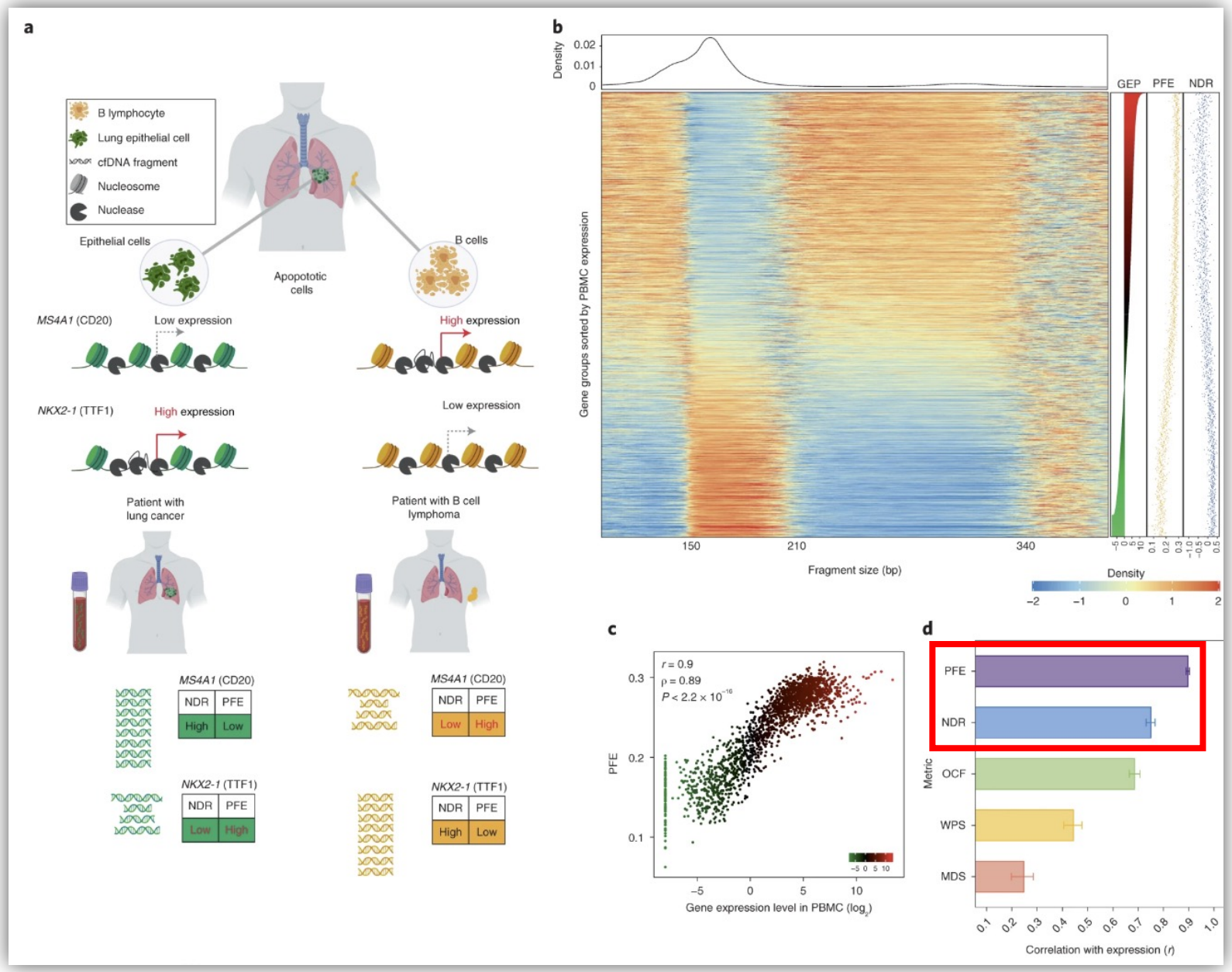
Esfahani, M., et al. Nature biotechnology (2022)

EPIC-seq method

Introduced new fragmentomic features!

- promoter fragmentation entropy (PFE)
- nucleosome-depleted regions (NDR)

Strongly correlated to gene-expression and clinical outcome



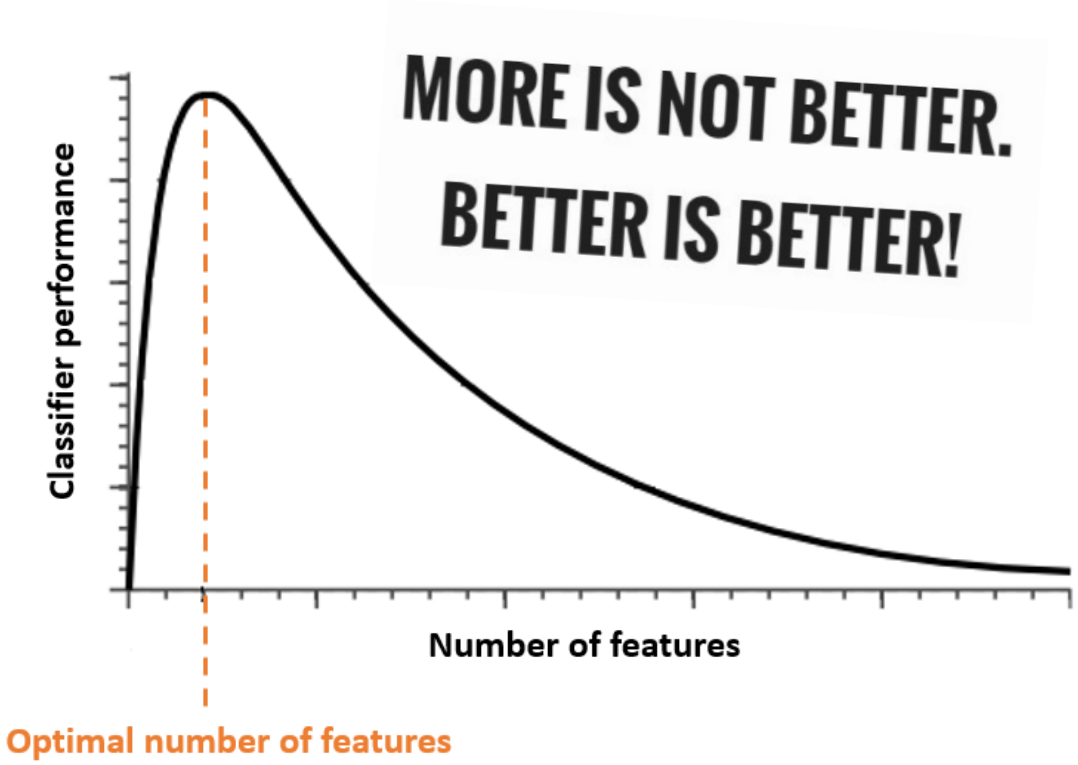
LB analyte → Feature engineering → Feature selection 2

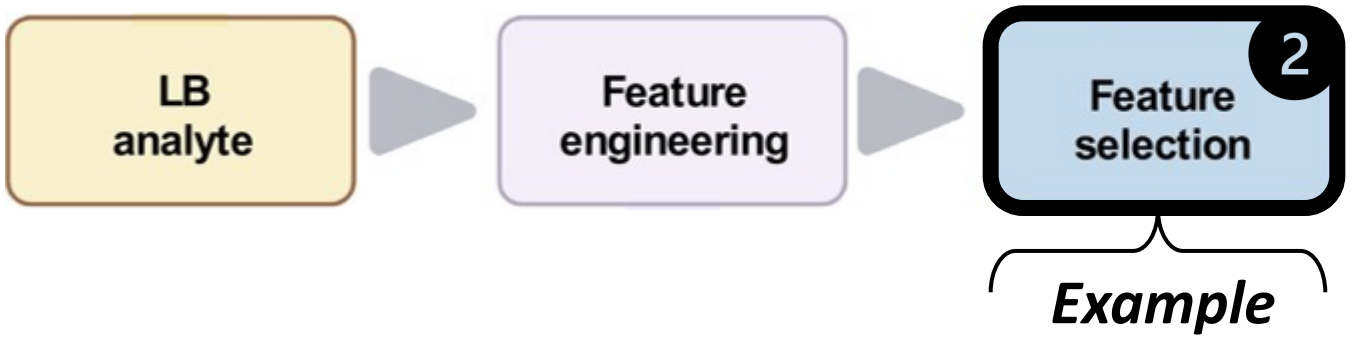
"Distilling the Essentials from Thousands of Candidates"

- After generating features, we face a problem called the "***curse of dimensionality***" : thousands of potential markers, but relatively smaller patient cohorts.

- **Challenge 2**: Filtering out uninformative or redundant features and keeping only those with real predictive power.

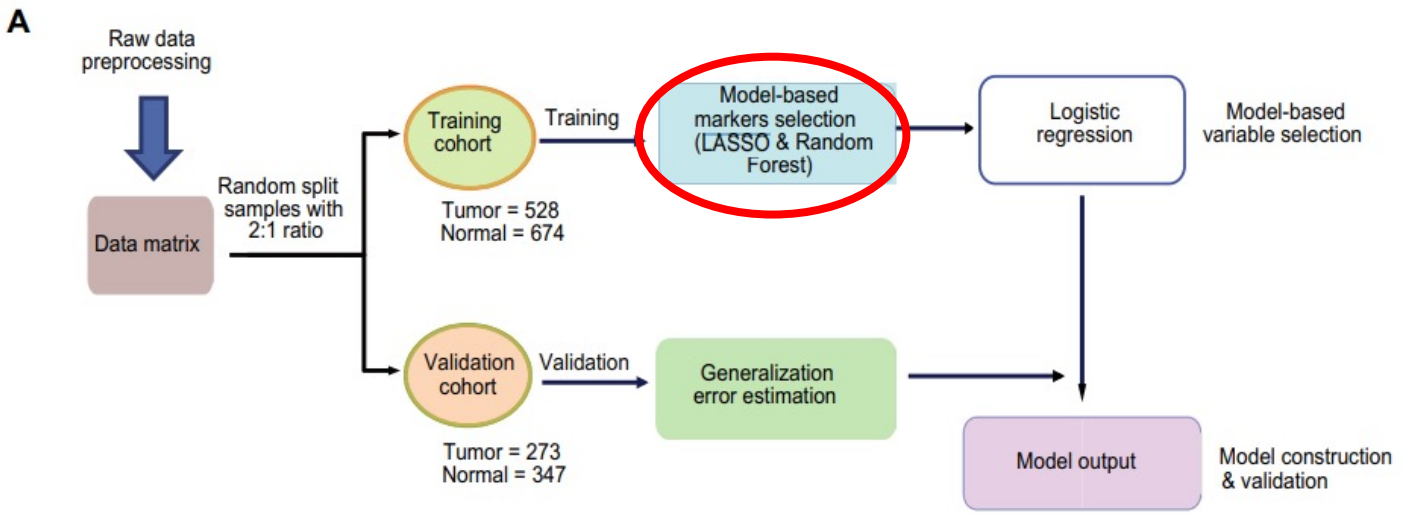
- **Techniques**: Common methods include statistical tests (correlation, linear regression) with regularization approaches (LASSO, ridge).
  These help selecting the most discriminative signals.

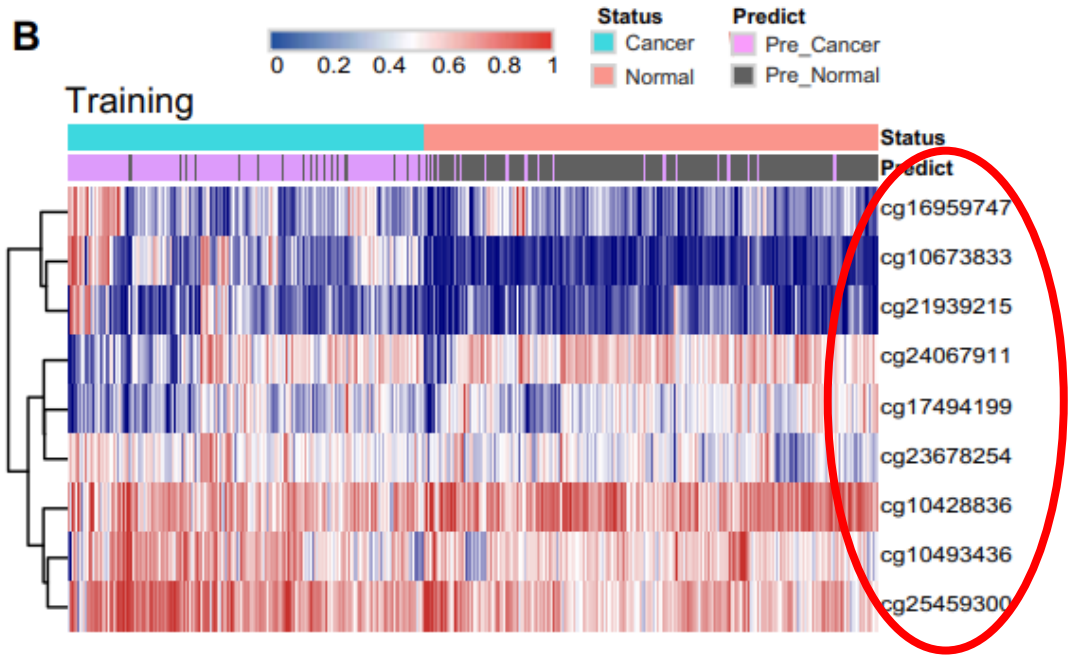




Luo, H., et al. «ctDNA methylation enable early diagnosis, prognosis prediction, and screening for colorectal cancer» Science transl. Med. (2020)

*"We analyzed the **methylation dataset of 544 markers** using LASSO and random forest algorithms to **reduce the number of markers**. […] **We obtained nine overlapping markers** from the two algorithms and constructed a diagnostic score"*

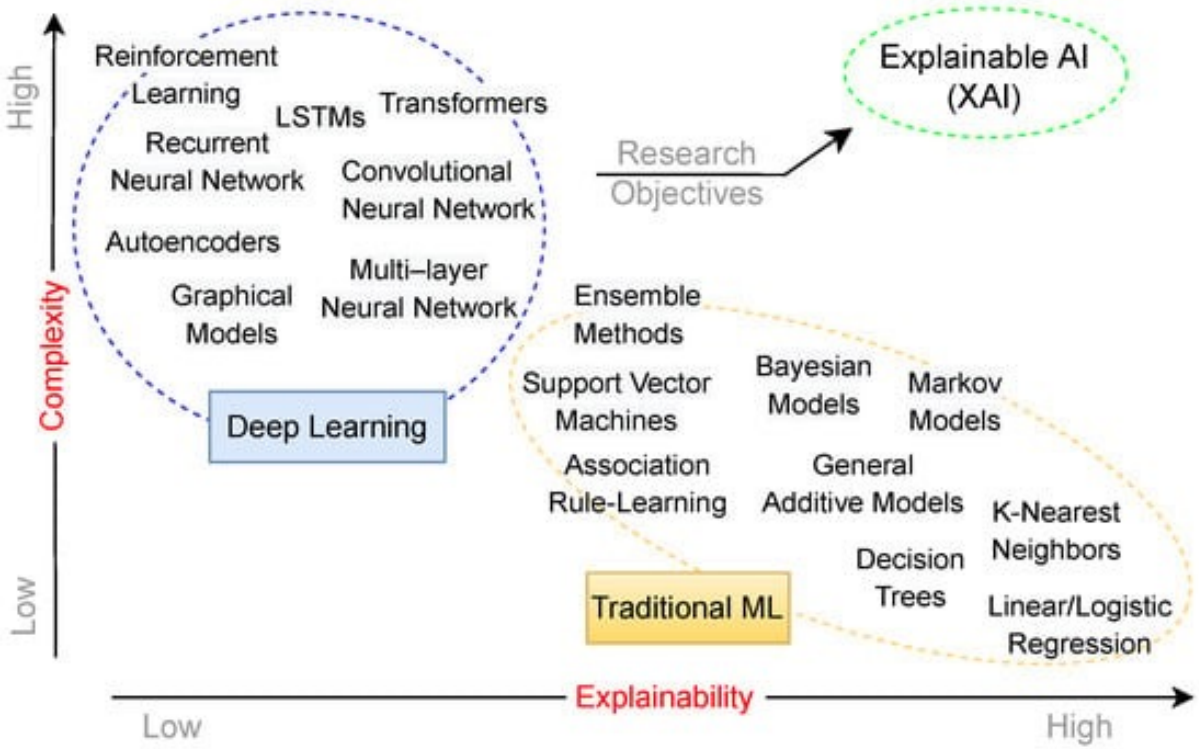




LB analyte → Feature engineering → Feature selection → Model selection 3

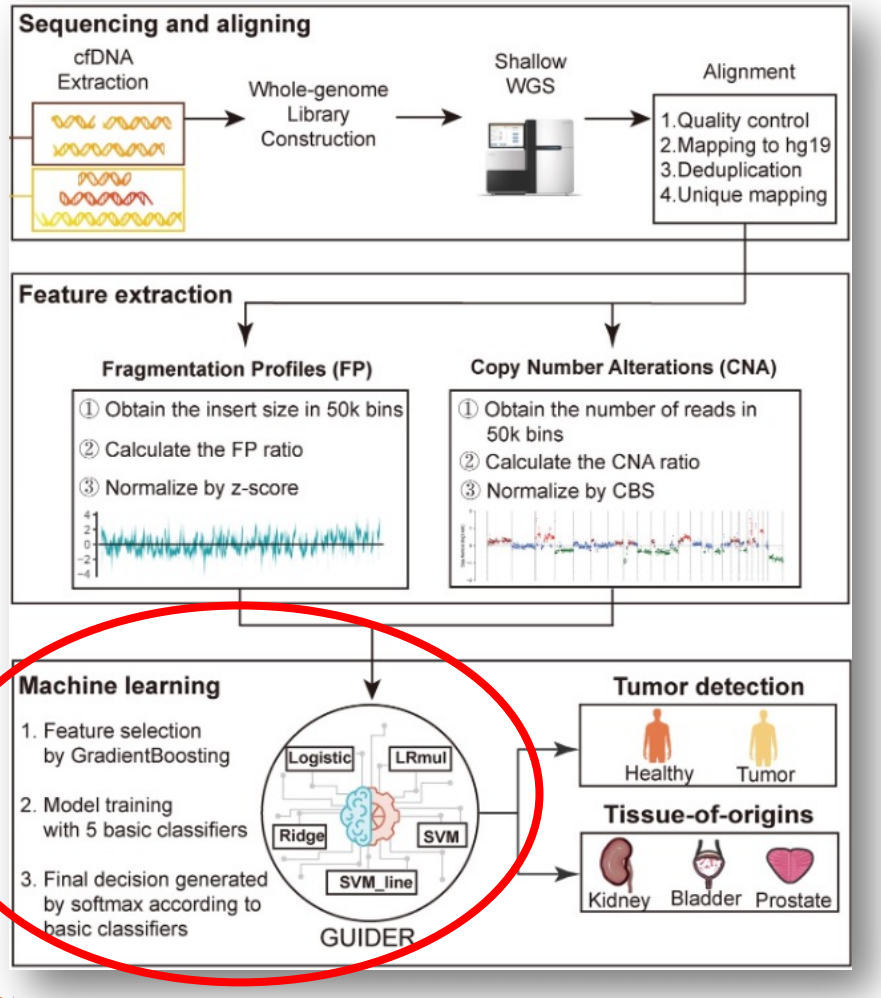
Choosing the Right Engine to Drive Accurate Predictions

- Once we've defined our feature set, we pick a model: simple and explainable methods (e.g., logistic regression) or complex "black-box" methods (e.g. neural networks).
- **Challenge 3**: Balancing explainability and complexity. Complex AI and *deep-learning* models can uncover intricate patterns but are much harder to explain.
- Clinicians often prefer transparent models if accuracy is not severely compromised.
- **No Universal Rule**: There's no single model optimal for all problems -> test various models, experimentation is key to find the one that performs best!



LB analyte → Feature engineering → Feature selection → Model selection 3

Example



Han, Yang, et al. "Enhanced detection of genitourinary cancers using fragmentation and copy number profiles obtained from urinary cfDNA" Clinical Chemistry (2021).

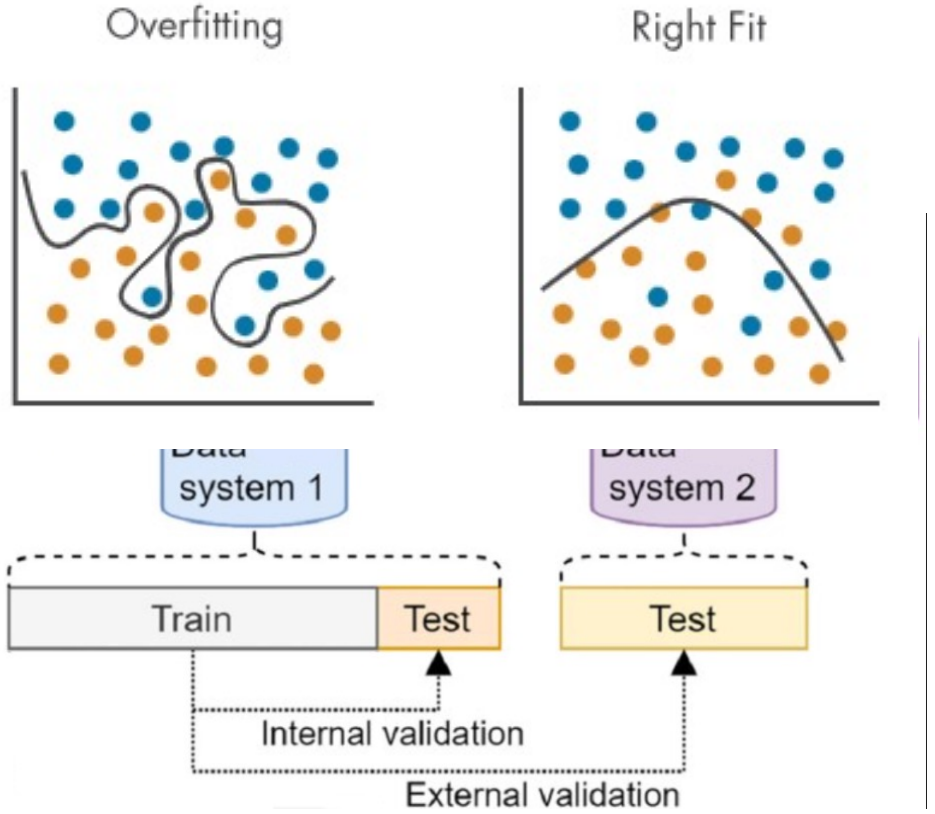
Multiple models involved

1. *Logistic Regression* (Logistic)
2. *Logistic Regression multinomial* (LRmul)
3. *Support Vector Machine* (SVM)
4. *Support Vector Machine con kernel lineare* (SVM_line)
5. *Ridge Regression* (Ridge)

Ensemble Approach: Each algorithm independently analyzes data and produces predictions. Then the outputs of these five models are consolidated using a probability-averaging strategy (softmax) which picks the prediction with the highest probability.

"From Proof-of-Concept to Clinical applications"

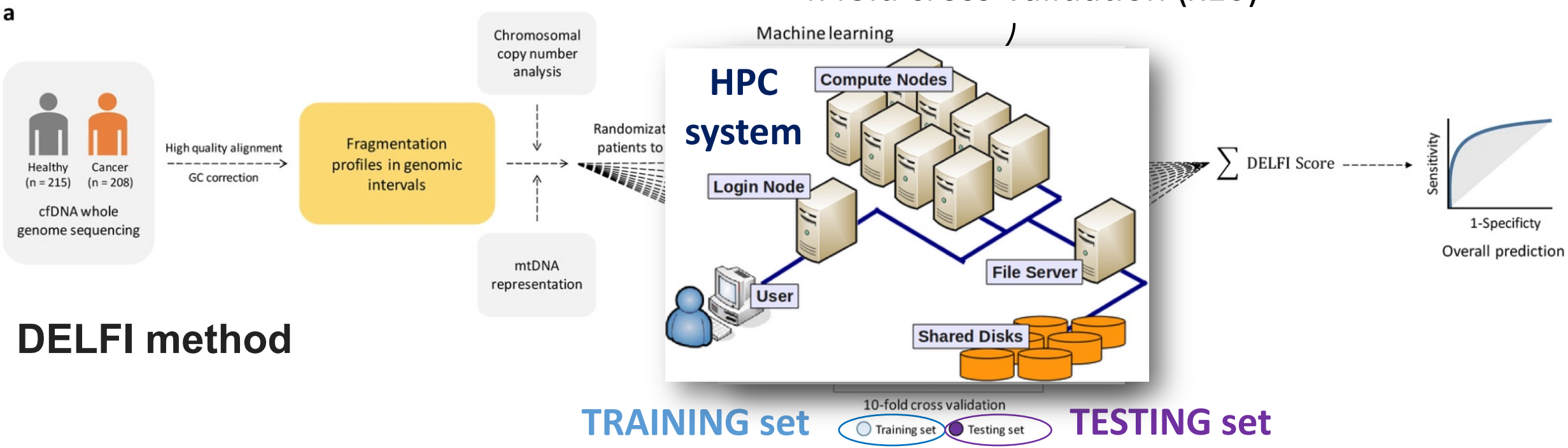
- Ultimately, the goal is to detect disease early, classify tumor, and predict relapse or response. This requires rigorous validation at every step.
- **Challenge**: Avoid **OVERFITTING**. Or more simply: ensuring that the model generalizes well beyond the initial *training cohort.*
- A machine-learning approach must undergo strict validation of performances (sensitivity, specificity, ROC) on a *test cohort* before clinical deployment.
- Validation can be **external** (best) or **internal** (more feasible)



Example

Cristiano, Stephen, et al.
"Genome-wide cell-free DNA fragmentation in patients with cancer."
Nature (2019)

Internal validation:
k-fold cross-validation (x10)



DELFI method

TRAINING set **TESTING set**

Conclusions

A **combination** of
- *Genomics*
- *Epigenomics*
- ***Multi-analyte** features*
- ***Other features…***
- ***Artificial Intelligence approaches***

will drive the **next generation of Liquid Biopsy biomarkers** for early cancer detection, cancer monitoring, therapy response assessment, etc..

Evolution of Liquid Biopsy

- **Third generation LB**: methilation, fragmentomic, nucleosome positioning, deconvolution in COO, etc…
- **Fourth generarion LB**: AI and Machine-learning to integrate all genomic and epigenomic features.

Complex Data-Analysis Challenges

- **Feature engineering**
- **Dimensionality reduction**
- **Feature selection**
- **Model selection**
- **Internal / external validation configurations**

Balancing Accuracy and Explainabilty

- **Advanced models** (DL, neural networks) **greatly enhance performances** for tumor detection.
- They are **intrinsically hard to explain**, efforts transparency remains essential for clinical adoption.

From Research to Clinical Application

- **Real-world impact**: rigorous **validation** with independent testing and external datasets is fundamental to achieve clinical-grade applications.

Future is in Data Integration

Thank you for the attention!

SERVIZIO SANITARIO REGIONALE
EMILIA-ROMAGNA
Azienda Ospedaliero - Universitaria di Bologna
IRCCS Istituto di Ricovero e Cura a Carattere Scientifico

POLICLINICO DI
SANT'ORSOLA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

AIRC
ASSOCIAZIONE ITALIANA
PER LA RICERCA SUL CANCRO

Multiple Myeloma Research Unit

Prof. Elena Zamagni

Bioinformatic Team

Andrea Poletti
Vincenza Solli
Gaia Mazzocchetti
Viola Meixian Vuong

Molecular Biology Lab

Carolina Terragna
Marina Martello
Enrica Borsi
Silvia Armuzzi
Ilaria Vigliotta
Barbara Taurisano
Ignazia Pistis
Alessia Croce

Clinical Research Unit

Paola Tacchetti
Lucia Pantani
Katia Mancuso
Serena Rocchi
Ilaria Rizzello

AIL Bologna