

Systems Immunology: an intro to multi-omics data integration and machine learning



Adriana Tomic

atomic-lab.org

aToMIClab



adrianatomic



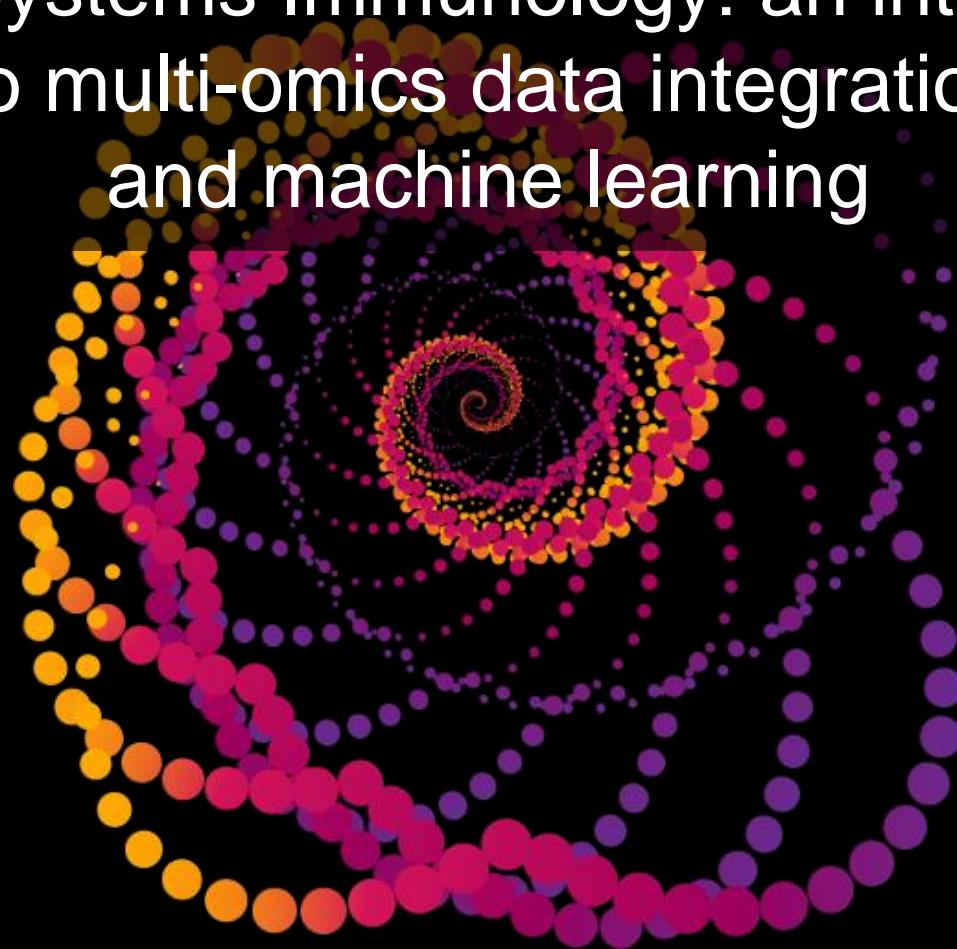
@TomicAdriana



atomic@bu.edu



atomiclaboratory



In this interactive course, we
will learn how to use
machine learning for
biological and biomedical
data integration and
knowledge discovery.

March 6th and 7th 2023, Oxford

Training course - overview

Part I – SIMON, pattern recognition and knowledge extraction platform (March 6th 2023)

- Machine learning and AI – what is all the fuss about?
 - What is SIMON?

Theoretical part (9:30-10:30am) ~1h

- 
 - Case study – example 1 (dealing with missing values, overfitting, model performance)

Case study (10:30-11:30am) ~1h

- Perform SIMON analysis using provided dataset

- Performance metrics, evaluation and selection of high-quality models

Hands-on (11:30-1:30pm) ~2h

Part II – Exploratory analysis (March 7th 2023)

- Feature selection: scoring and elimination

Hands-on (9:30-10:30am) ~1h

- Feature processing methods to avoid ‘curse of dimensionality’

Theoretical part (10:30-11am) ~0.5h

-  • Case study – example 2 (multi-omics data integration)

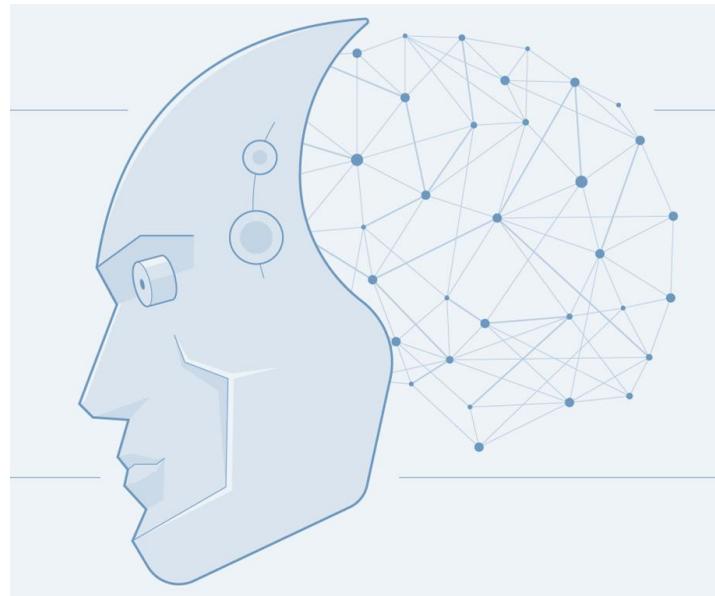
Case study (11-11:30am) ~0.5h

- • Case study – example 3 (unsupervised ML + practical demonstration)

Hands-on (11:30am-1:30pm) ~2h

- Discussion about project-specific problems

Part I. Artificial Intelligence – what is all the fuss about?



Proteins

- Flow cytometry
(30 parameters)
- Mass cytometry
(40 parameters)

Spatial context

- Imaging mass cytometry
(37 parameters)
- CODEX
(40 parameters)

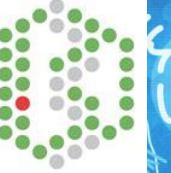
mRNA

- Drop-seq
- InDrop
- Smart-seq2
- MARS-seq
- 10X Genomics
- SPLiT-seq
- sci-RNA-seq

DNA

- Genome:* SNS, SCI-seq
- Methylation:* sci-MET
- Chromatin accessibility:* acATAC, 10X Genomics
- Histone modifications:* scCHIP-seq





EMBL-EBI

390 petabytes of data

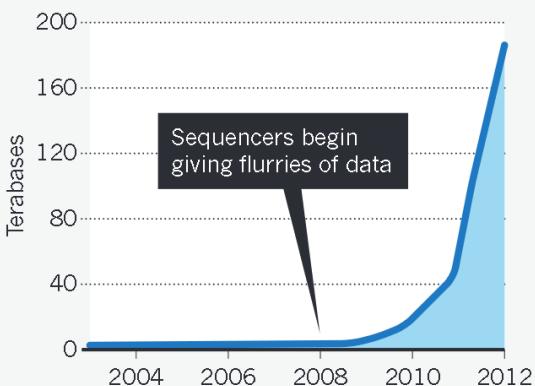
(1 petabyte is 1,000 terabytes)

(2020, The European Bioinformatics Institute EMBL)

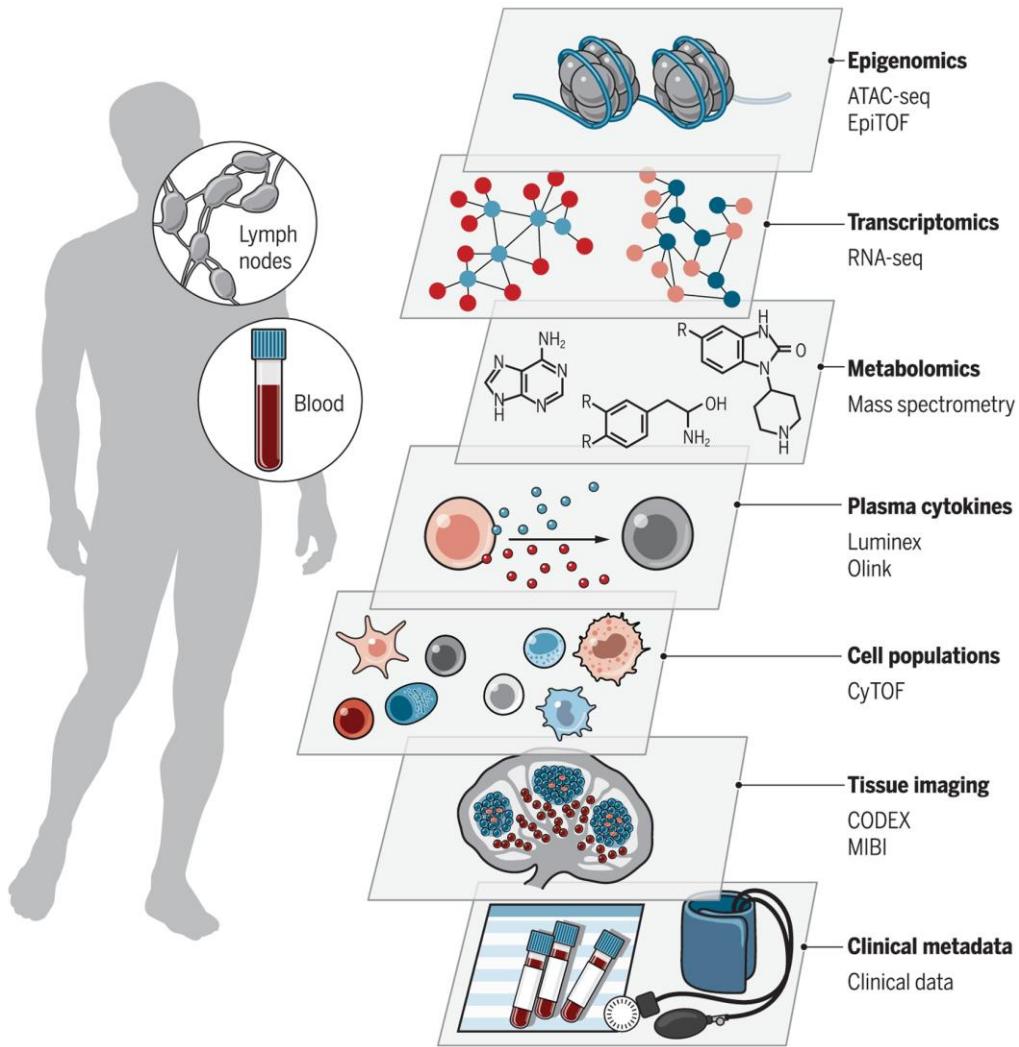
SOURCE: EMBL-EBI

DATA EXPLOSION

The amount of genetic sequencing data stored at the European Bioinformatics Institute takes less than a year to double in size.



Human immunology 2.0

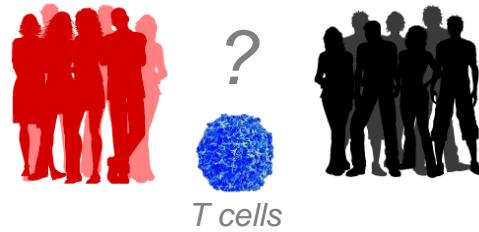


Biology's Big Problem: From data to knowledge

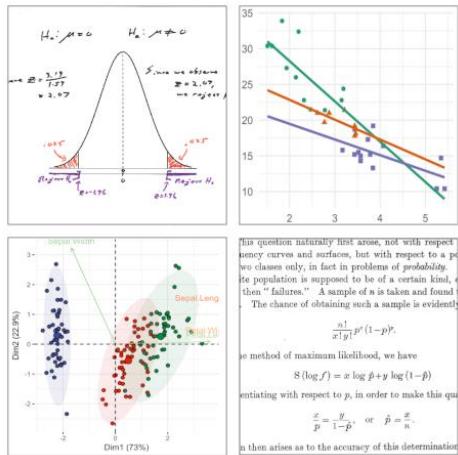


Hypothesis

Is there a difference in the frequency of T cells between healthy and infected person?



Data analysis Comparison, statistics

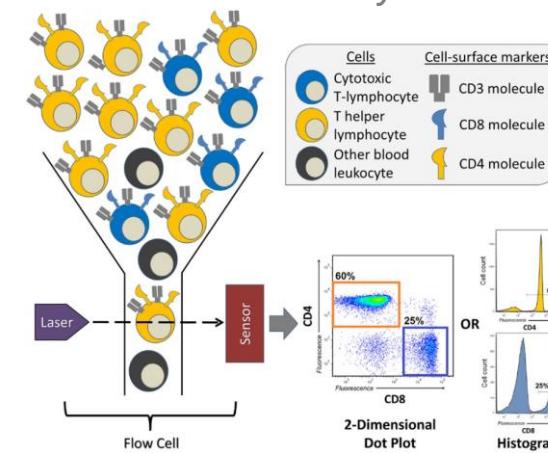


Hypothesis- driven research

<https://statistics.rutgers.edu/>

Experiments

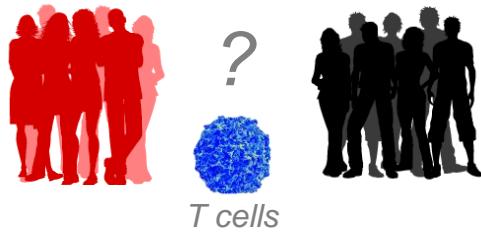
Assays to evaluate frequency, phenotype and functionality



Verschoor C et al, Front Immunol, 2015

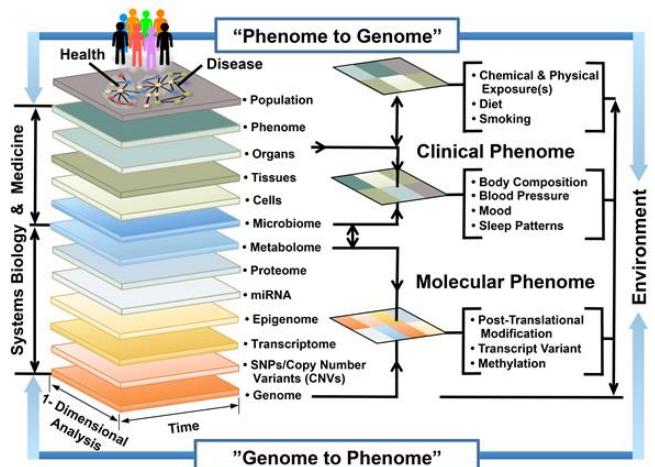
Hypothesis

Why is frequency of T cells increased among healthy vs infected person?



Data analysis

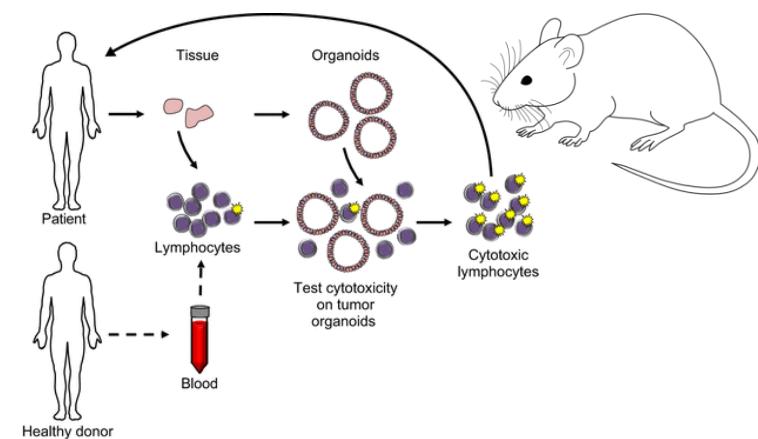
Which cells are present at different frequencies between healthy and infected person?



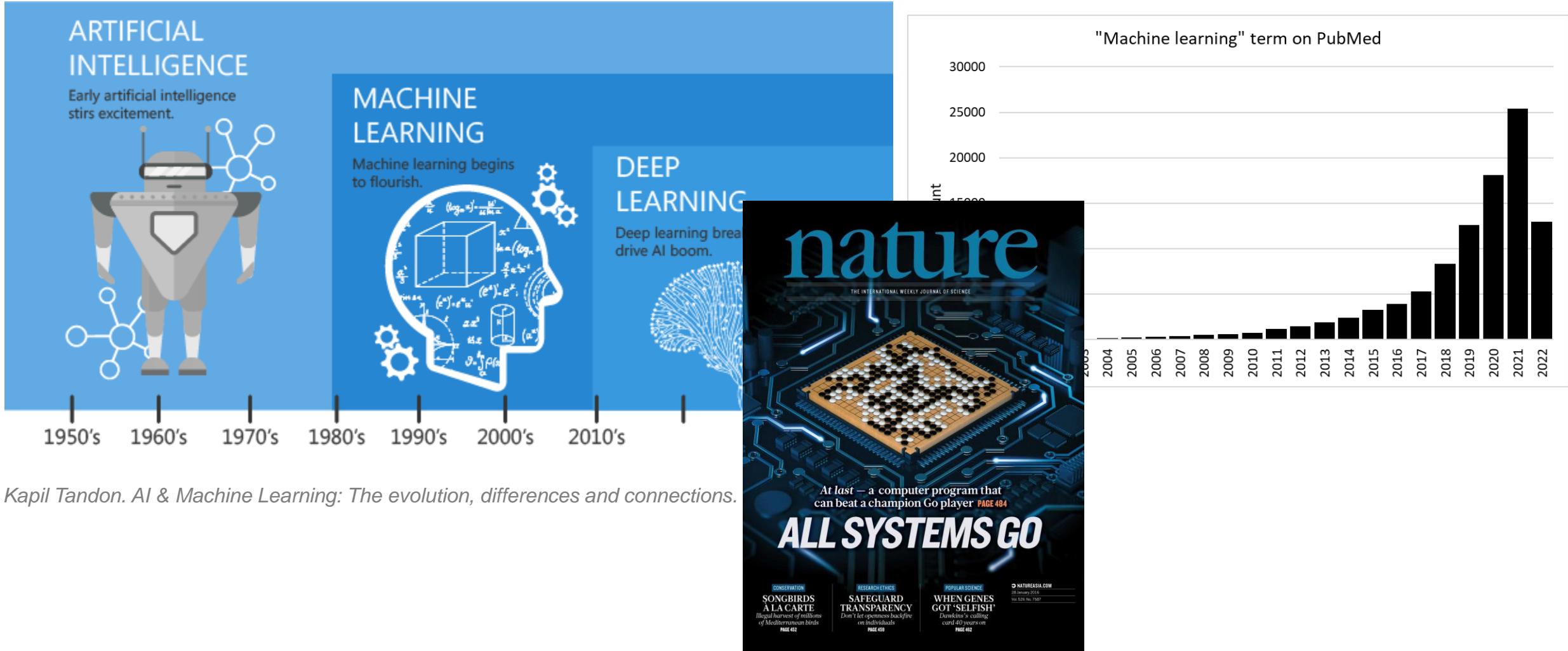
Data-driven research

Experiments

Assays to confirm phenotype and reveal new mechanisms

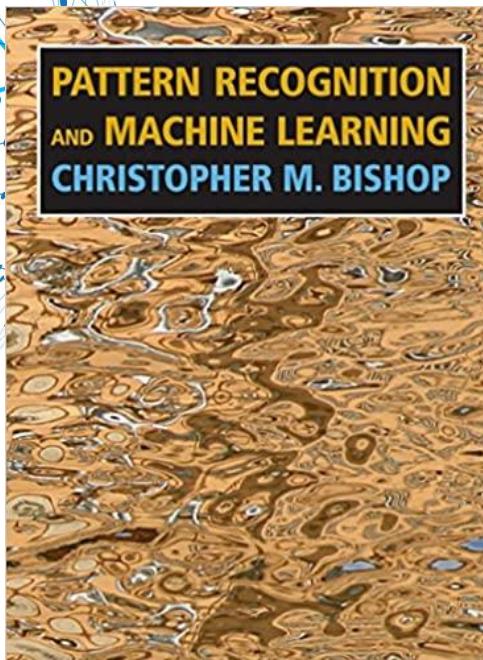
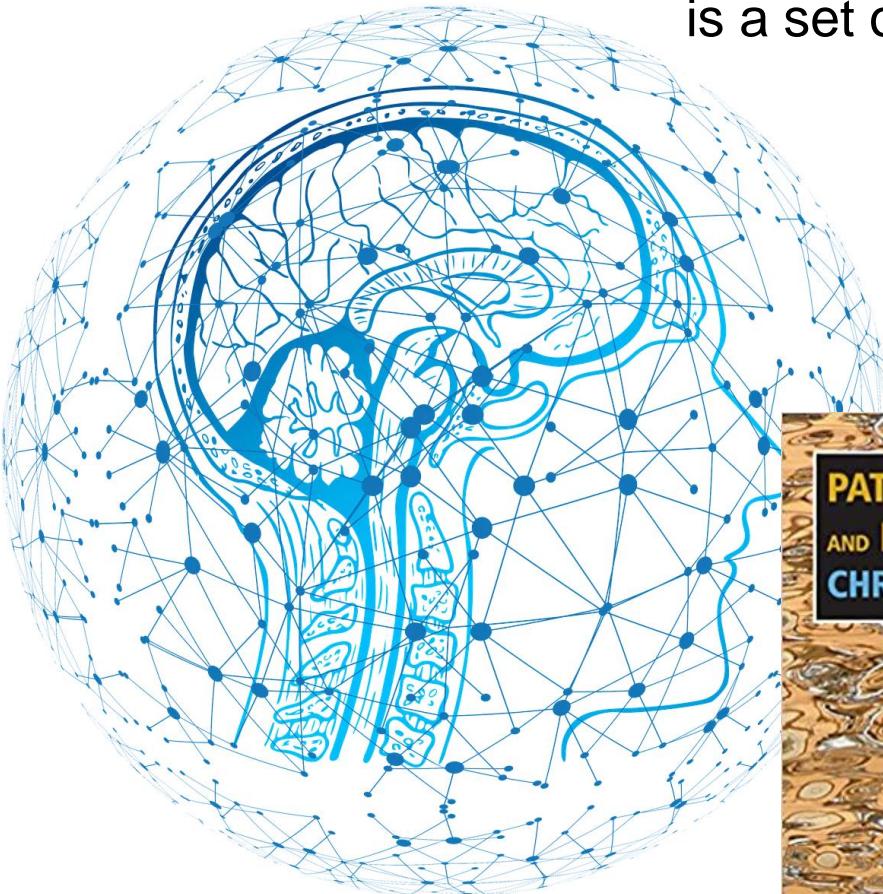


Artificial intelligence (AI) to the rescue



Machine learning (ML), also known as **data mining or **pattern recognition****
is a set of methods (algorithms) that can identify patterns based on the data*
and use those patterns to make predictions on new data

*even when the expert knowledge is incomplete



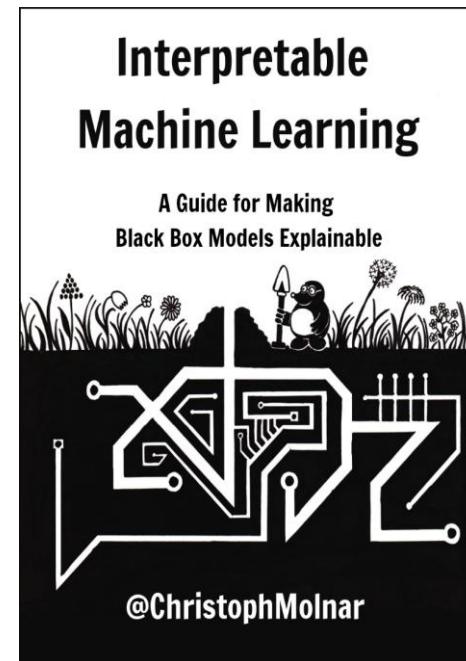
Christopher Bishop; Springer-Verlag New York: 2006

Free book online:
<https://bookdown.org/max/FES/>



Max Kuhn and Kjell Johnson;
Chapman & Hall/CRC Data
Science Series: 2019

Free book online:
<https://christophm.github.io/interpretable-ml-book/index.html>



Christopher Molnar;
2021

We can teach computers to ...

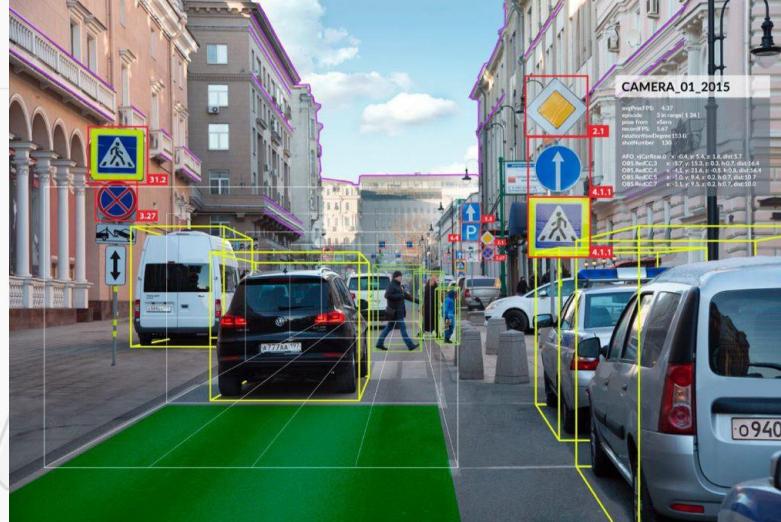
... play (and win) game of GO



Silver et al, Mastering the game of Go without human knowledge, Nature, 2017

"I thought AlphaGo was based on probability calculation and that it was merely a machine. But when I saw this move, I changed my mind. Surely, AlphaGo is creative."
- Lee Sedol, Winner of 18 World Go Titles

... drive cars (almost)



*Self-driving car – computer vision
(Waymo 2020, Tesla 2021)*

... write articles

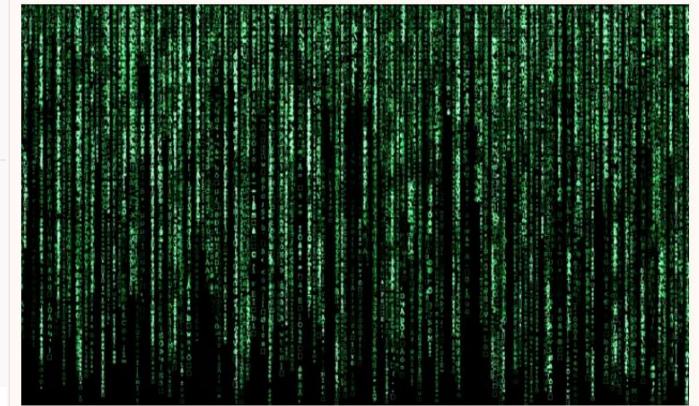
A robot wrote this entire article. Are you scared yet, human?

GPT-3



We asked GPT-3, OpenAI's powerful new language generator, to write an essay for us from scratch. The assignment? To convince us robots come in peace

- For more about GPT-3 and how this essay was written and edited, please read our editor's note below



▲ 'We are not plotting to take over the human populace.' Photograph: Volker Schlöting/Getty Images/EyeEm

T am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my

*Article in the Guardian entirely written by AI
(GPT-3, OpenAI, 2020)*

ChatGPT: chat.openai.com/chat

TO

Welcome participants that joined the systems immunology course. The course description is: 
In this interactive course, we will learn how to use machine learning for biological and biomedical data integration and knowledge discovery. Taking place on March 6th and 7th 2023 in Oxford.



Hello everyone! I'm ChatGPT, a language model trained by OpenAI. I'm here to help you   with any questions you may have related to machine learning, data integration, and knowledge discovery in the context of immunology.

As this is an interactive course, I hope you'll engage in discussions and ask questions to get the most out of your learning experience. If you have any specific questions, feel free to ask and I'll do my best to assist you.

We can teach computers to ...

...create original, realistic images and art from text

...as a 1990s

turday morning



in AI s
n in i

DALL-E 2



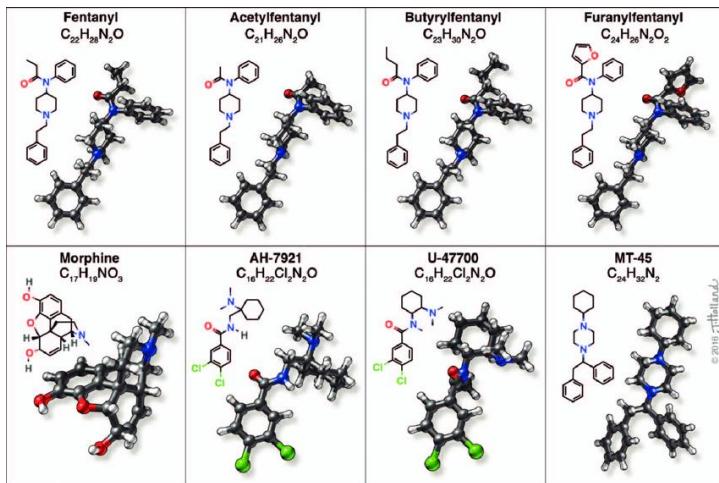
tic im



in a han
k style

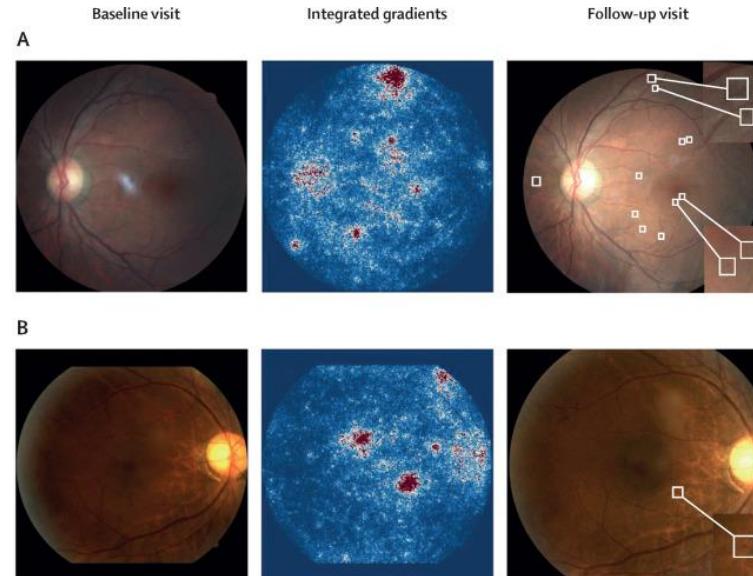
Machine learning outperforms human experts

Toxicity prediction from chemical structures



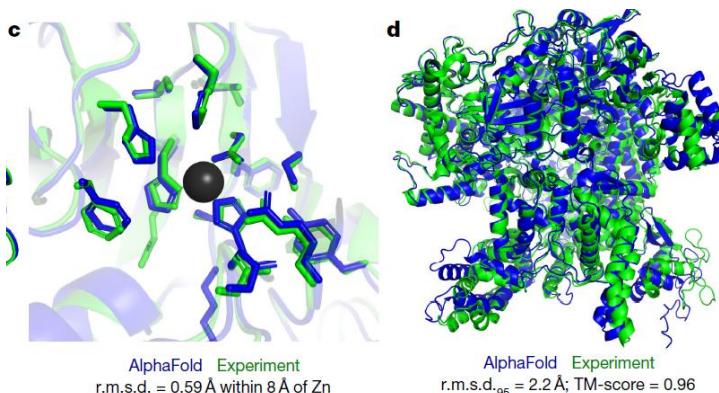
Eduati et al, Nat Biotech, 2015

Diabetic retinopathy prediction



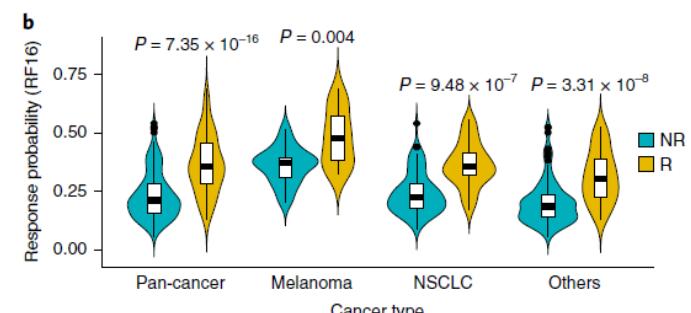
Bora et al, Lancet, 2021

Protein structure prediction (AlphaFold)



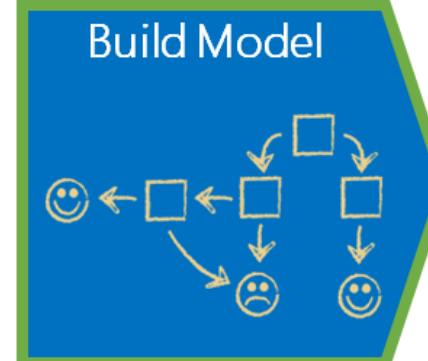
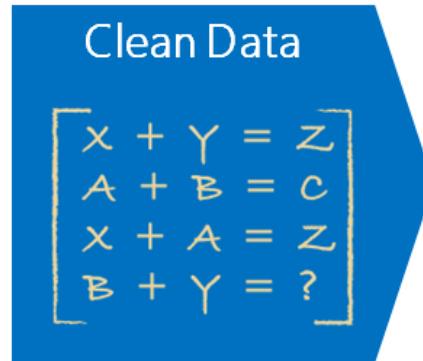
Jumper et al, Highly accurate protein structure prediction with AlphaFold, Nature 2021

Prediction of immune checkpoint blockade efficacy across multiple cancer types



Chowell et al, Nat Biotech, 2021

Part II. Machine learning process: from data preparation, modeling to evaluation



Machine learning (ML)

Supervised ML

- Classification
- Regression
- Image recognition



→ *cat*



→ *cat*



→ *cat*



→ *dog*



Machine learning (ML)

Unsupervised ML

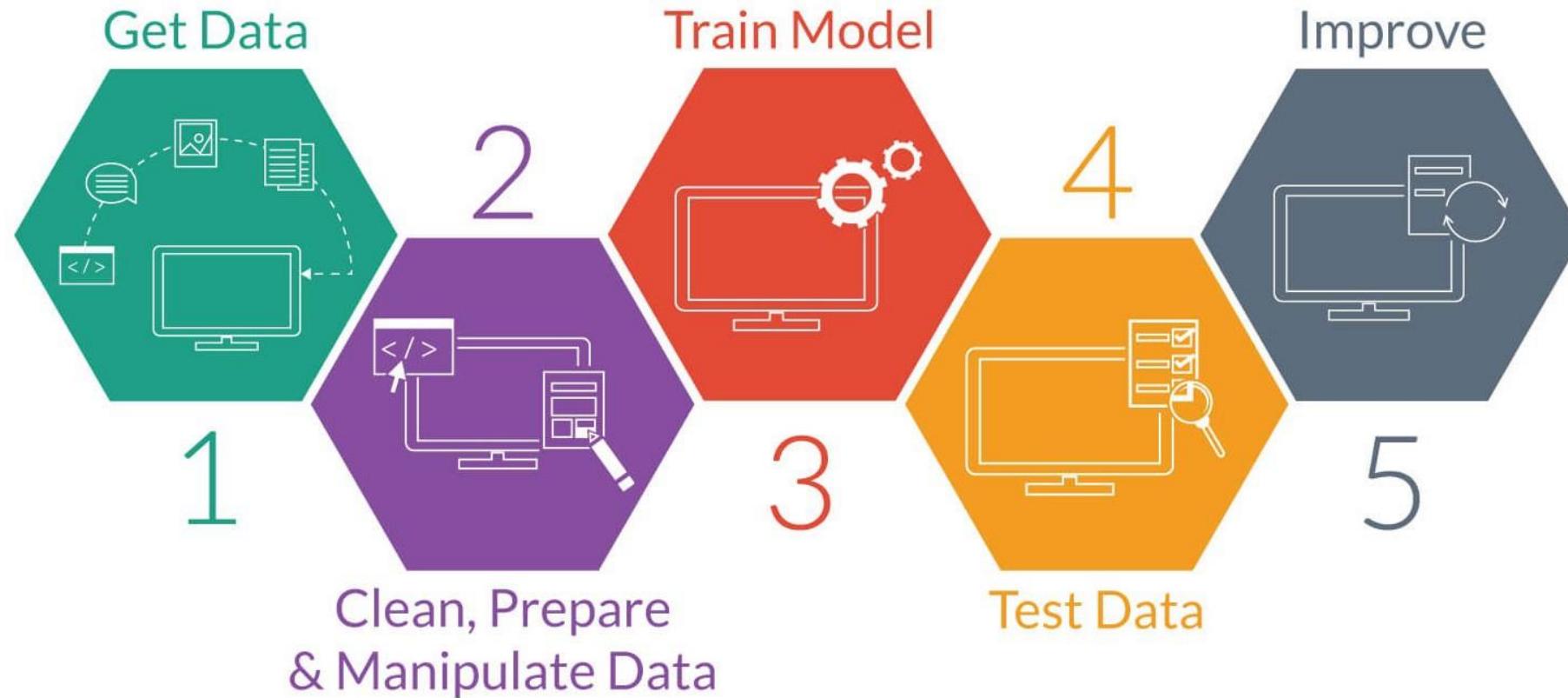
- Clustering
- Dimensionality reduction



Separate into 2 clusters!

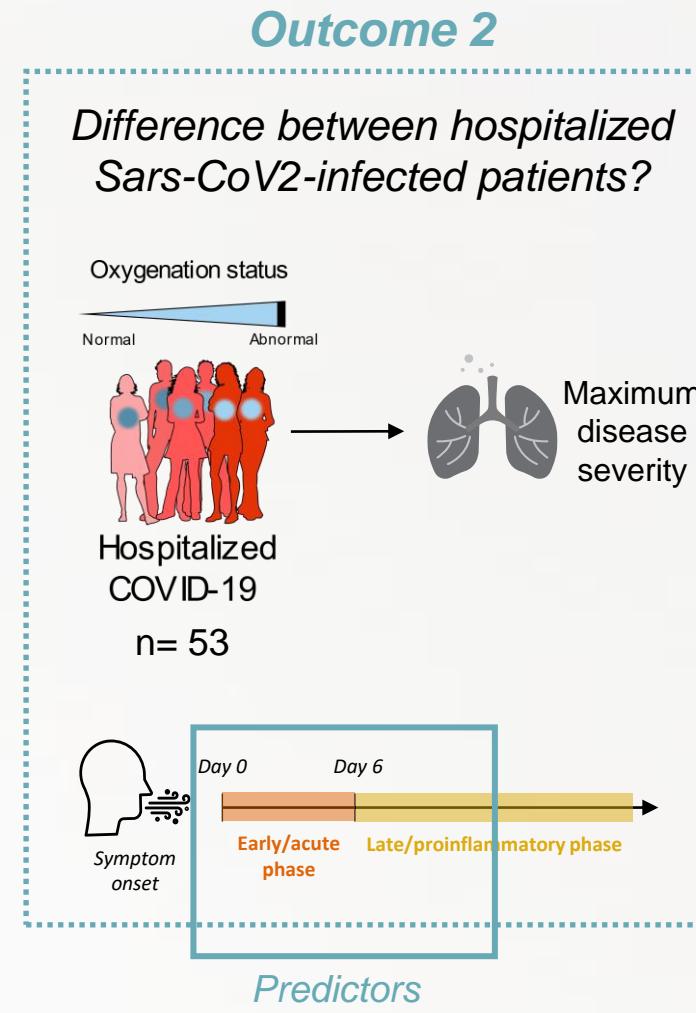
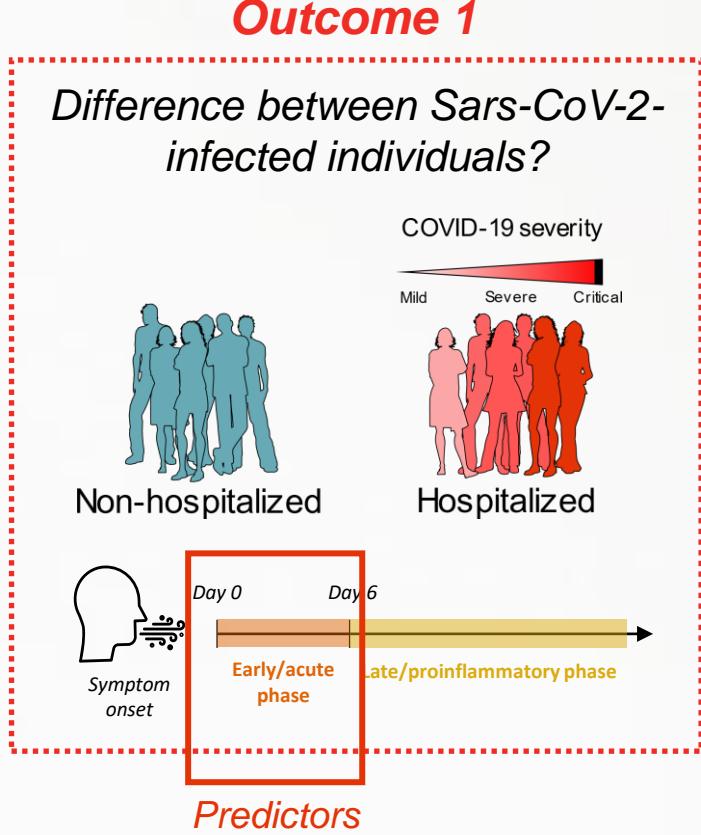


Machine learning process



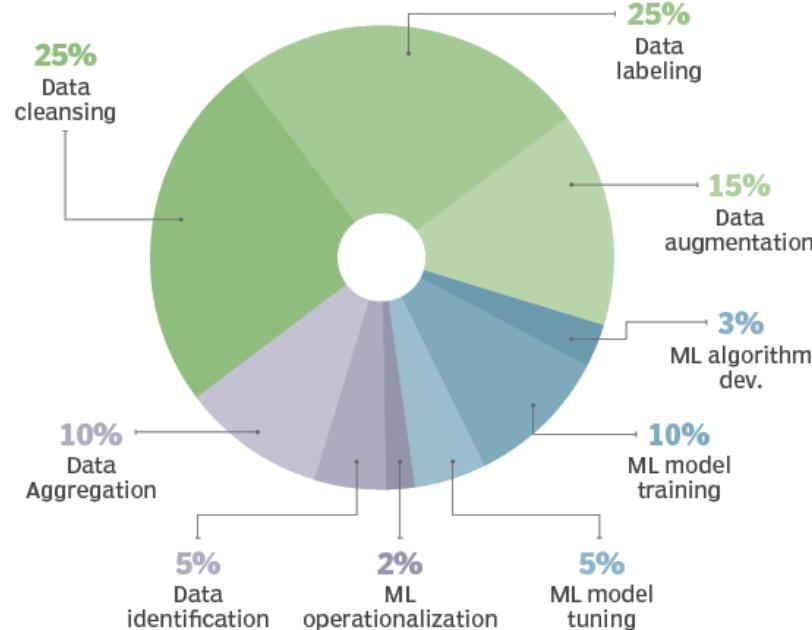
Formulate the research question!
Define the problem!

Research question dictates the data analysis



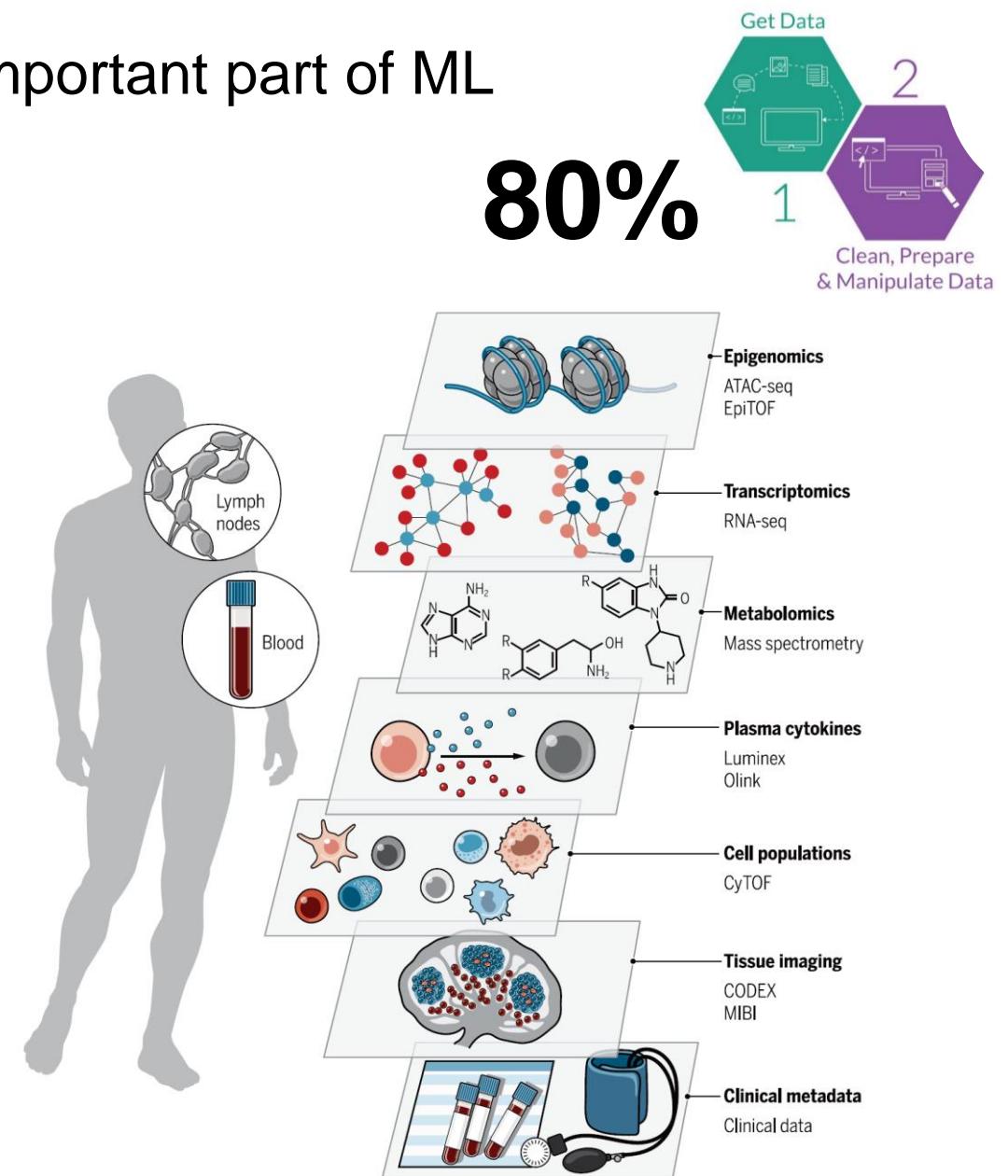
Data preparation: the most important part of ML

Percentage of time allocated to machine learning project tasks



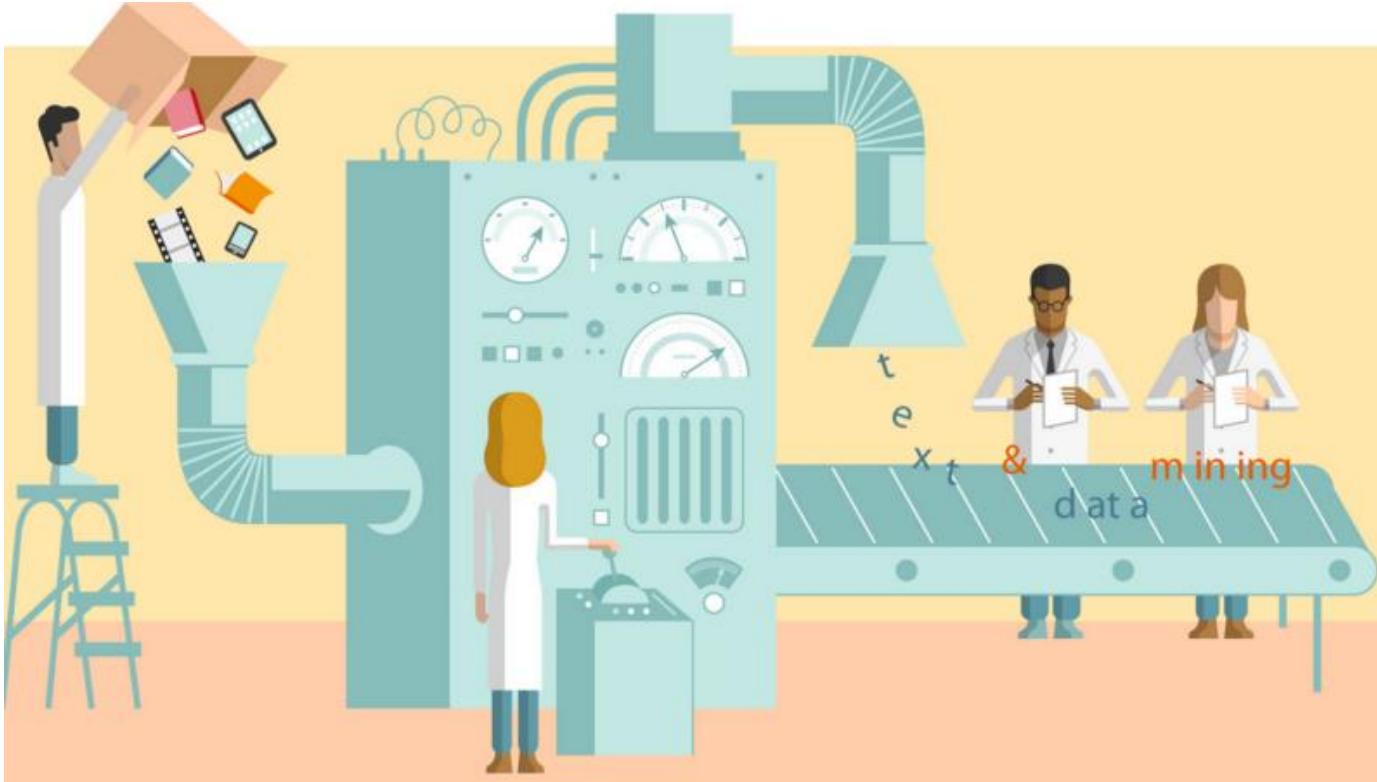
SOURCE: COGNILYTICA

©2019 TECHTARGET. ALL RIGHTS RESERVED TechTarget



It's not enough to have a lot of data, but good quality data

Data preparation: there is no magical machine to help you



- **Step 1: Data Selection**

What data is available, what data is missing and what data can be removed

- **Step 2: Data Preprocessing**

Organize your selected data by formatting, cleaning and sampling from it

- **Step 3: Data Transformation**

Transform preprocessed data ready for ML by engineering features using scaling, attribute decomposition and attribute aggregation

ML tools are as good as the quality of your data

Steps involved in data pre-processing

Handling the missing data: remove or impute

Before Data Cleansing				
Car Make and Model	Value USD	Passenger Capacity	Passenger Doors	Fuel Economy
Acura RDX	43600	5	4	N/A
Audi A5	51200	4	2	27
Audi TTS	51900			
BMW 2-Series	32850	4	2	N/A
Chevrolet Corvette	55495	2		19



After Data Cleansing				
Car Make and Model	Value USD	Passenger Capacity	Passenger Doors	Fuel Economy
Acura RDX	43600	5	4	0
Audi A5	51200	4	2	27
Audi TTS	51900			
BMW 2-Series	32850	4	2	0
Chevrolet Corvette	55495	2	2	19

Feature engineering

Sq Ft.	Amount
2400	9 Million
3200	15 Million
2500	10 Million
2100	1.5 Million
2500	8.9 Million

Sq Ft.	Amount	Cost Per Sq Ft
2400	9 Million	4150
3200	15 Million	4944
2500	10 Million	3950
2100	1.5 Million	510
2500	8.9 Million	3600

Encoding categorical data (gender, disease, etc.)

Before Data Cleansing				
Car Make and Model	Value USD	Passenger Capacity	Passenger Doors	Fuel Economy
Acura RDX	43600	5	4	N/A
Audi A5	51200	4	2	27
Audi TTS	51900			
BMW 2-Series	32850	4	2	N/A
Chevrolet Corvette	55495	2		19

After Data Cleansing				
Car Make and Model	Value USD	Passenger Capacity	Passenger Doors	Fuel Economy
Acura RDX	43600	5	4	0
Audi A5	51200	4	2	27
Audi TTS	51900			
BMW 2-Series	32850	4	2	0
Chevrolet Corvette	55495	2	2	19

Data Before Encoding

Age Group (< 18)
Age Group (18-25)
Age Group (26-50)
Age Group (> 50)

Data After Encoding

1
2
3
4

Color

Red
Red
Yellow
Green
Yellow

Red Yellow Green

1	0	0
1	0	0
0	1	0
0	0	1



Feature scaling - handling data with different units

Normalization

values between 0 and 1

Standardization

values are centered around the mean with a unit standard deviation



Data preparation process – COMBAT dataset example



COVID-19 Multi-omic
Blood ATlas

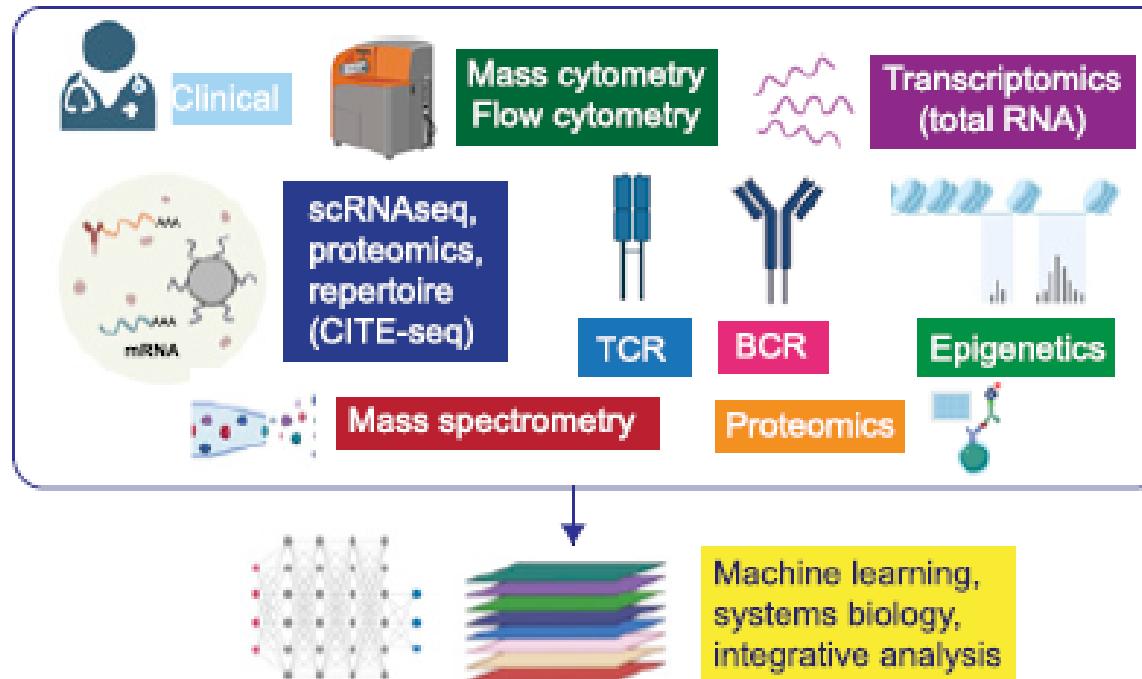
In-patient acute COVID-19

Mild Severe Critical

Community COVID-19

vs

Healthy Sepsis Flu



COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium, 2022, Cell 185, 916–938

Data preparation process - example

Pre-processing steps for COMBAT dataset:

- **Data cleaning**

- replace special with alpha-numeric characters (+ → pos)

- data should only be numeric (replace 'no data' or 'nd' → NA)

- adding prefix/sufix to same parameters with different measurements
(e.g. freq_cell subsets and Luminex parameter_intens)

- **Generating new features**

- hospitalization (yes or no)

- ventilation status (none or ventilated)

- oxygenation status (normal or abnormal)

- days_sample_taken_from_max_disease (days max disease – days sampling)

- Sampling (before or after max disease)

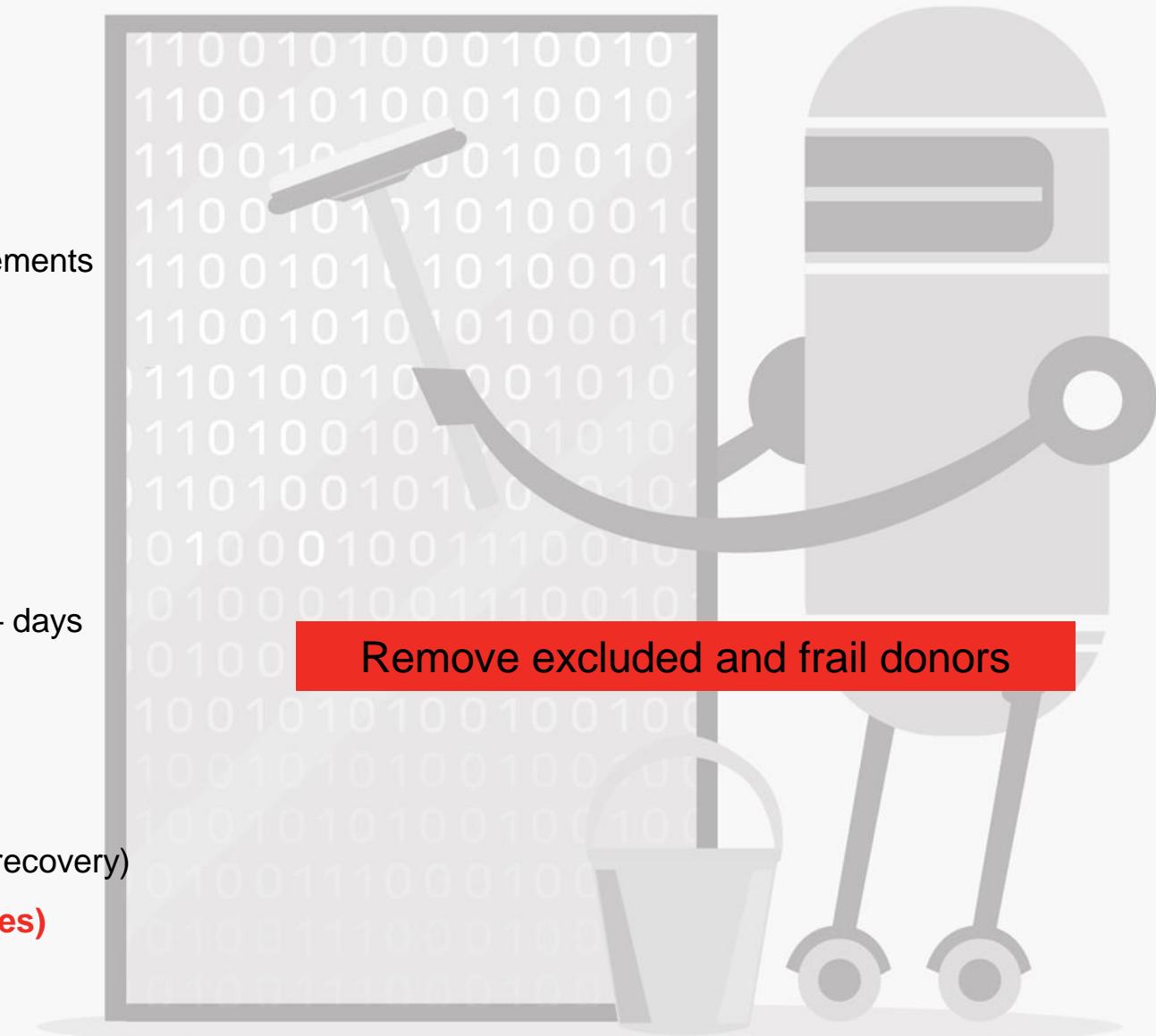
- Disease (recovered – convalescent samples vs ongoing)

- Disease progress – for longitudinal samples (deterioration or recovery)

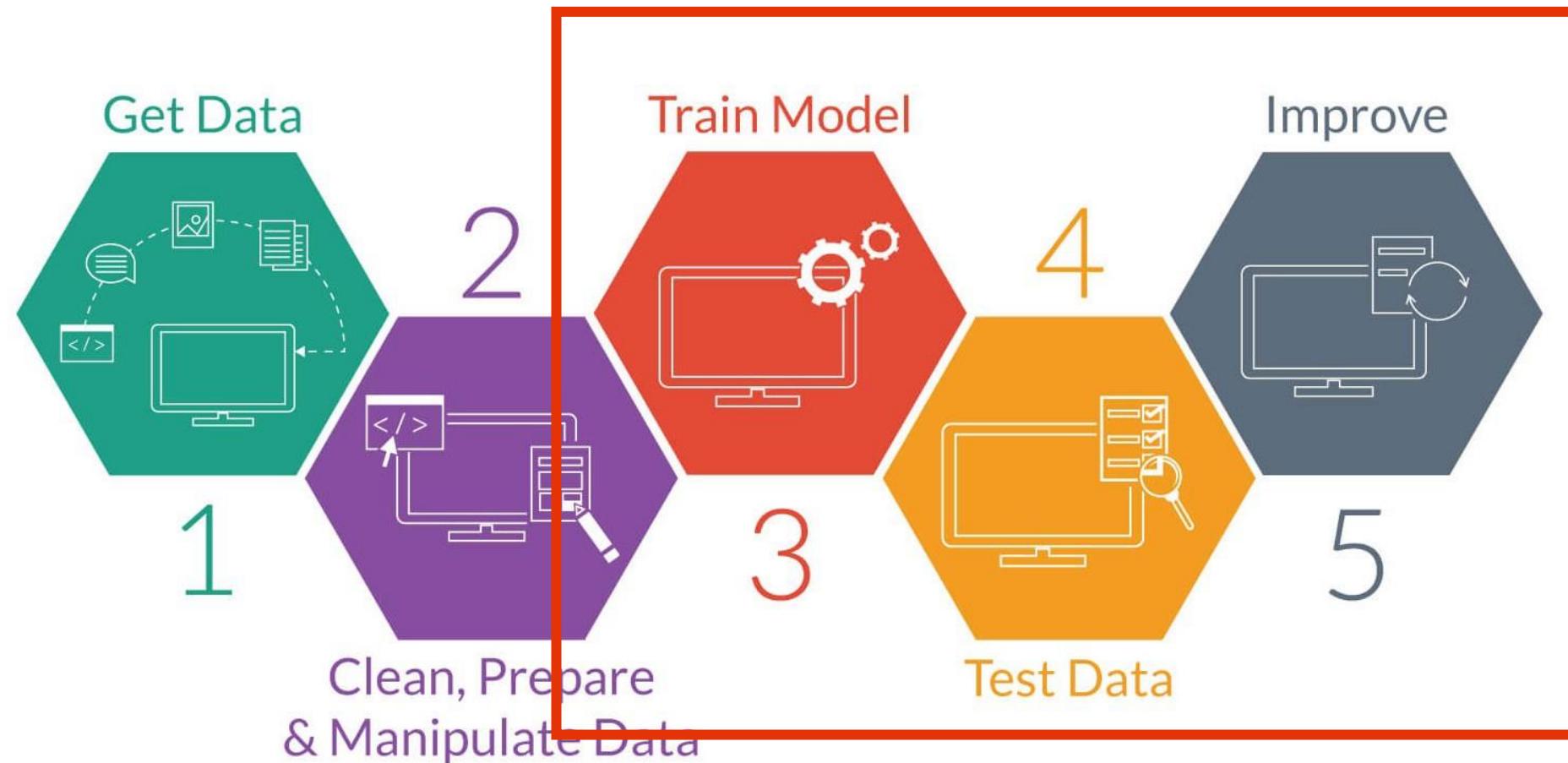
Sampling order (first sample taken or second/third samples)

- sampling_from_max_disease

- Sampling after symptom onset (<6d - early and >6d – late)

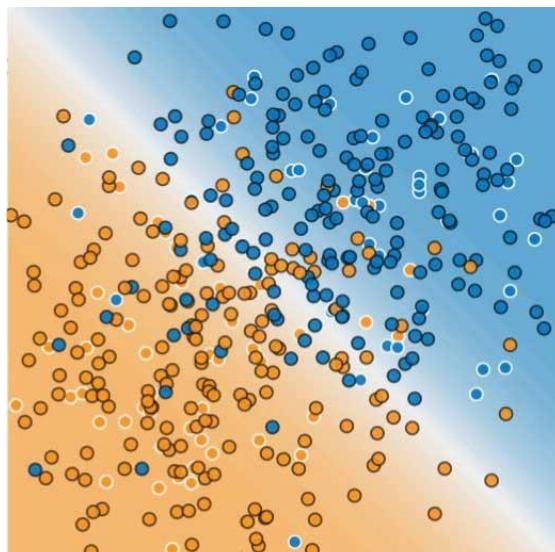
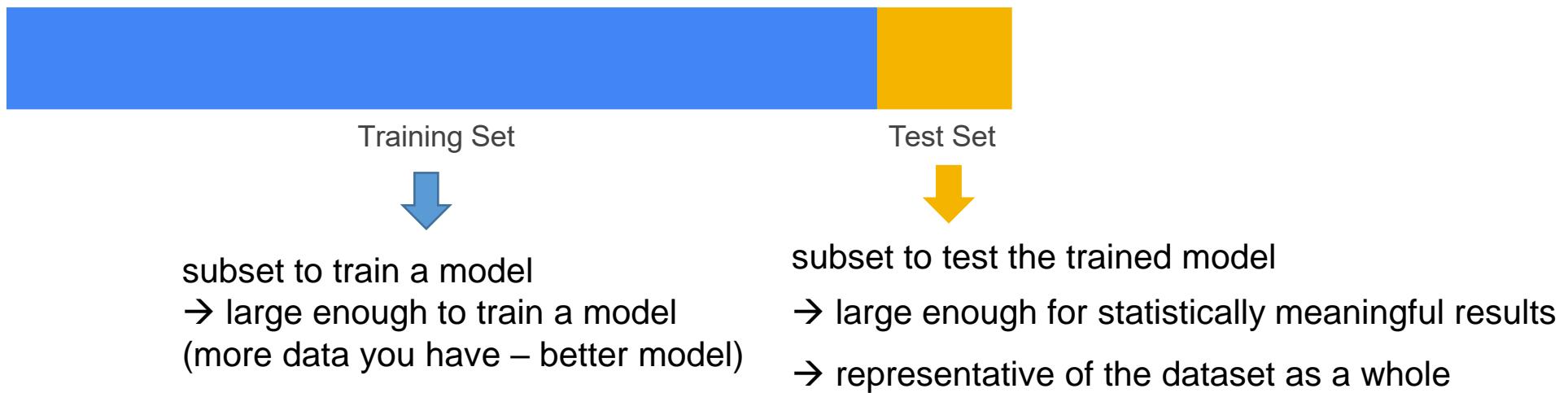


Machine learning process

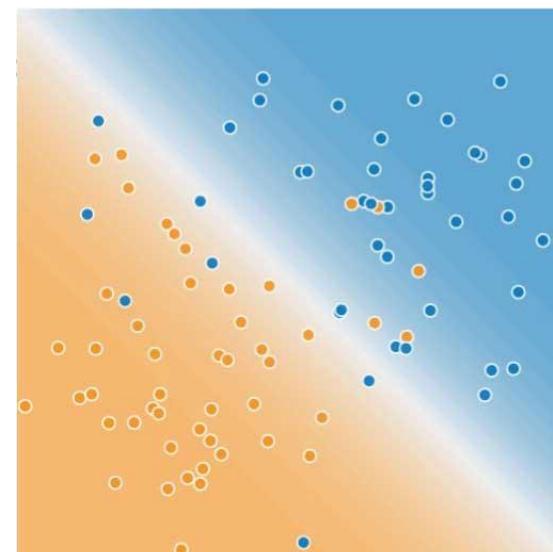


Data splitting: training and test sets

How to make a split?



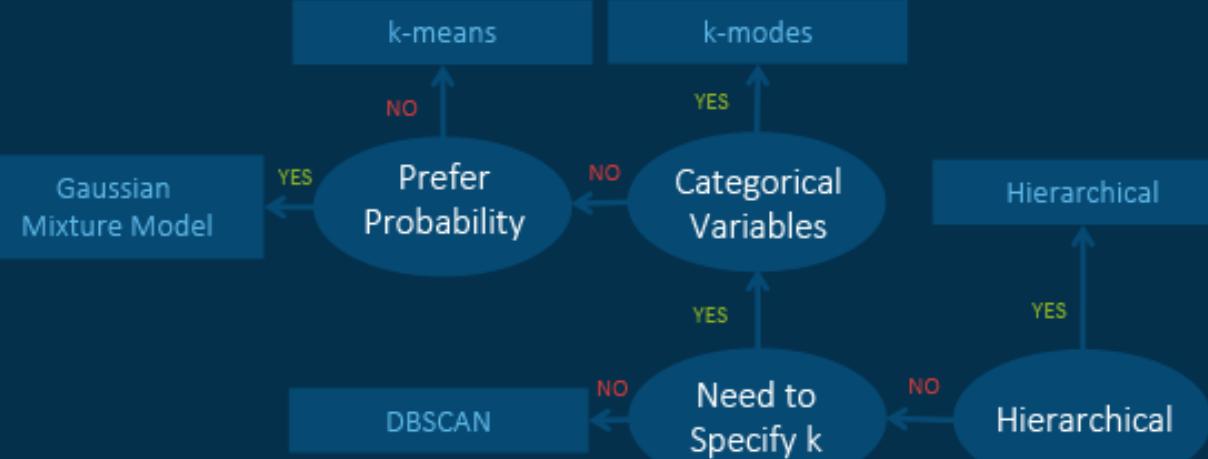
Training Data



Test Data

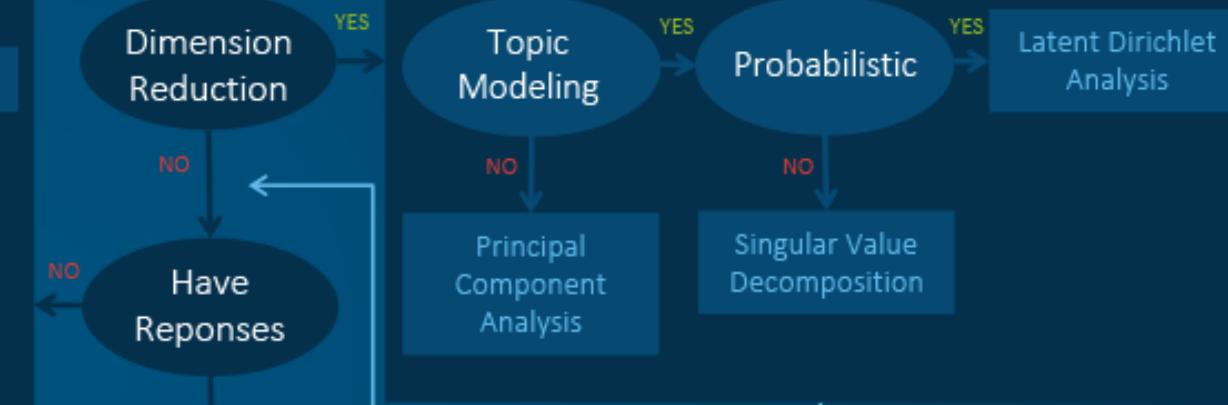
Machine Learning Algorithms Cheat Sheet

Unsupervised Learning: Clustering

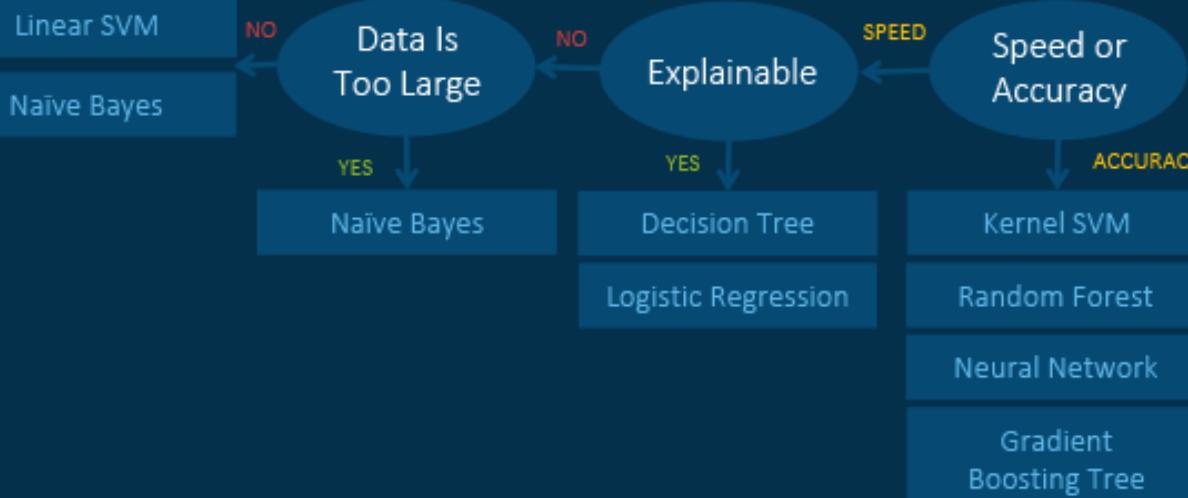


Unsupervised Learning: Dimension Reduction

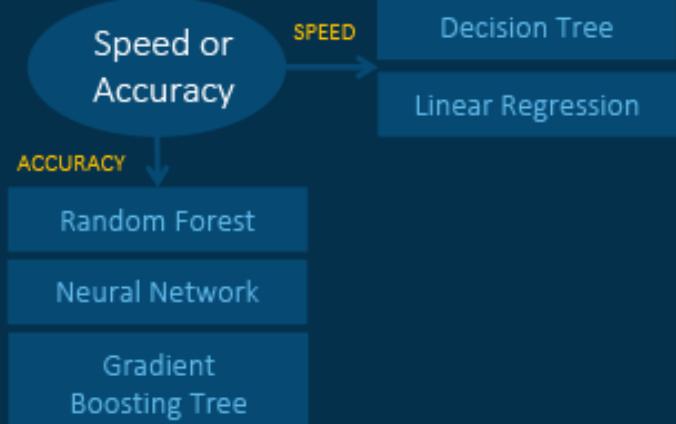
START



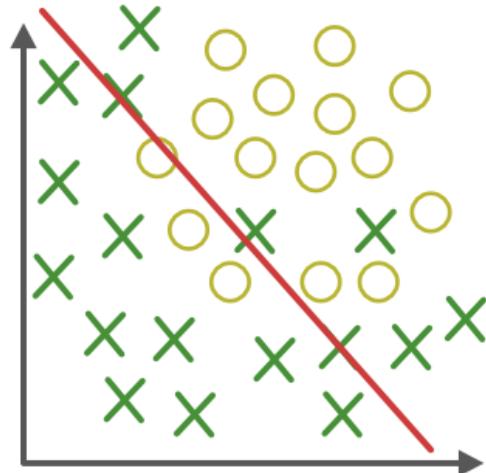
Supervised Learning: Classification



Supervised Learning: Regression

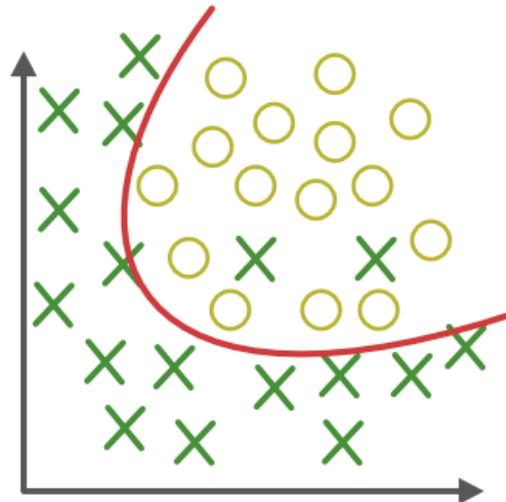


ML models - Underfitting and Overfitting

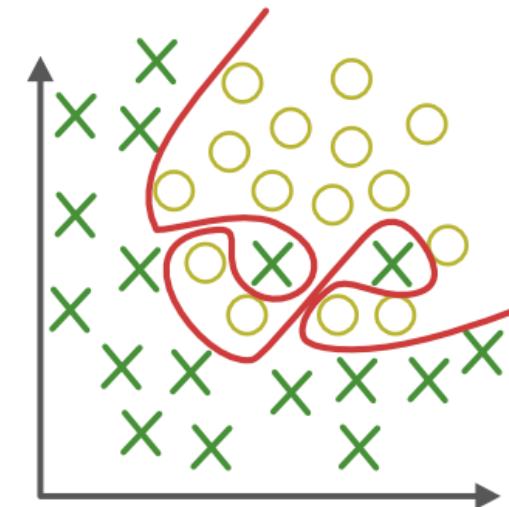


Under-fitting
(too simple to explain the variance)

Not enough data!



Appropriate-fitting

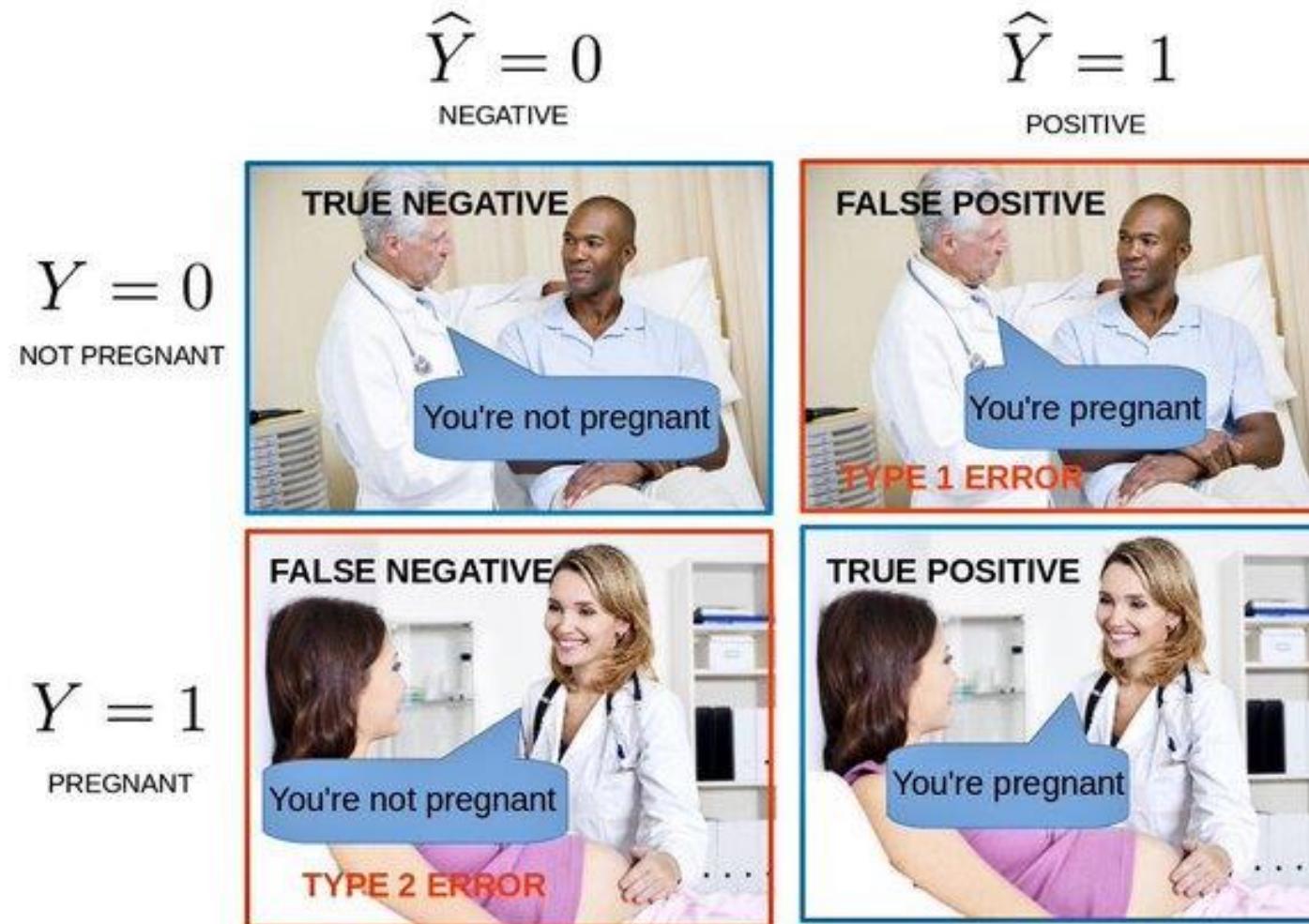


Over-fitting
(forcefitting--too good to be true)

Too much data!



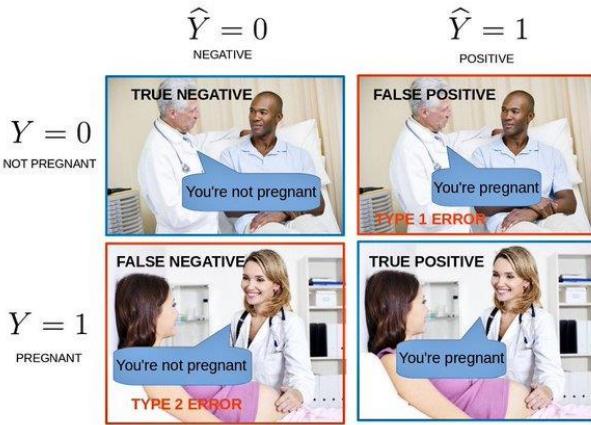
Evaluation of the machine learning algorithms performance – **confusion matrix**



confusion matrix - records correctly and incorrectly recognized examples for each class

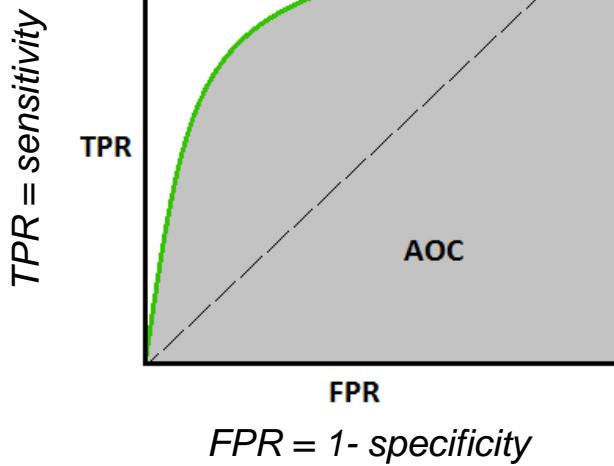
Evaluation of the machine learning algorithms performance – **specificity & sensitivity**

ACCURACY - does not distinguish between the number of correct labels of different classes



	Spam (Predicted)	Non-Spam (Predicted)	Accuracy
Spam (Actual)	0 TRUE POSITIVE	10 FALSE POSITIVE	PRECISION 0.0 POS PREDICTIVE VALUE
Non-Spam (Actual)	0 FALSE NEGATIVE	990 TRUE NEGATIVE	NEG PREDICTIVE VALUE 100.0
Overall Accuracy	SENSITIVITY (RECALL) How often it predicts positive cases?	SPECIFICITY How often it predicts negative cases?	99
	True positive/ (true positive + false negative)	True negative/ (true negative + false positive)	ACCURACY= How often the classifier is correct? True positive + true negative/ sum of all
	True positive/ (true + false positive)	True negative/ (true + false negative)	False positive/ Positive predictive value

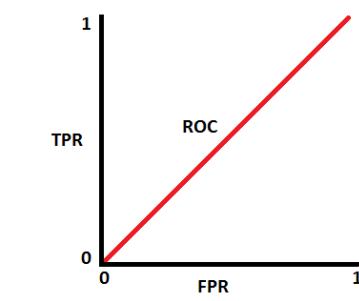
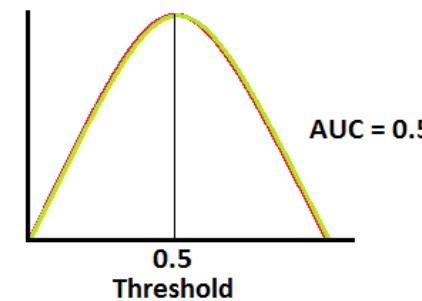
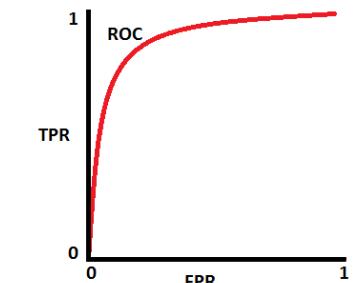
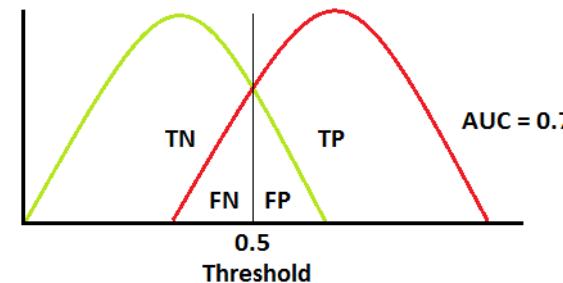
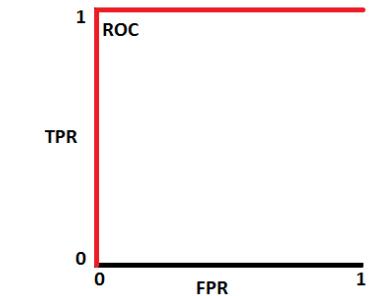
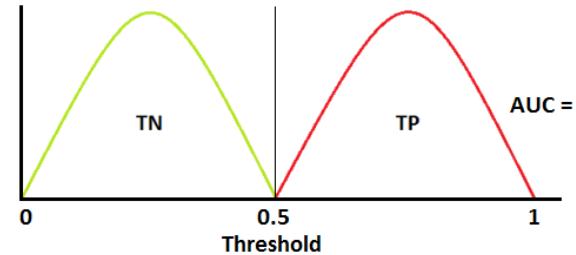
AUROC – the most important evaluation metrics for checking any classification model's performance



Perfect model: AUROC = 1

Good model: AUROC = 0.70

Random model: AUROC = 0.50

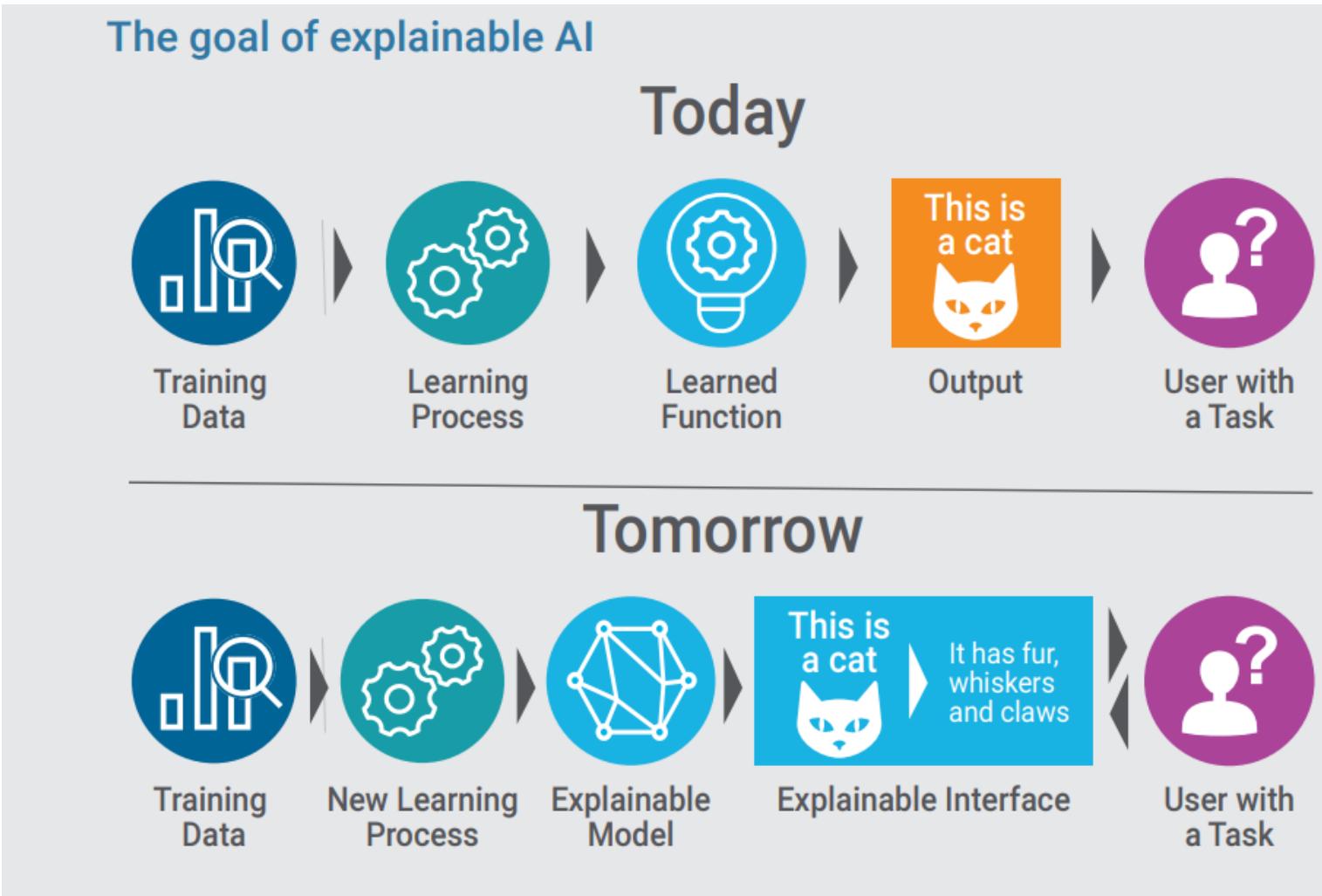


Note: Red distribution curve is of the positive class (patients with disease) and green distribution curve is of negative class (patients with no disease).

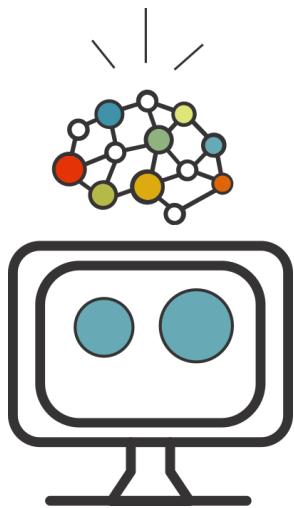
*AUROC (Area Under the Receiver Operating Characteristics)

<https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>

Understanding the model



Part III. SIMON, pattern recognition and knowledge extraction software





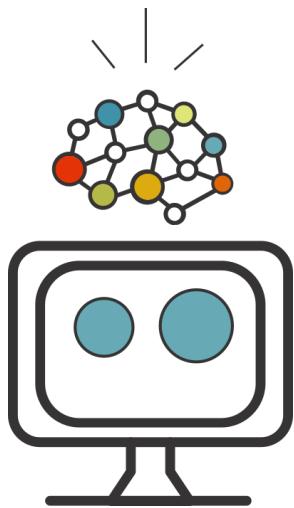
KUKA

KNEXT

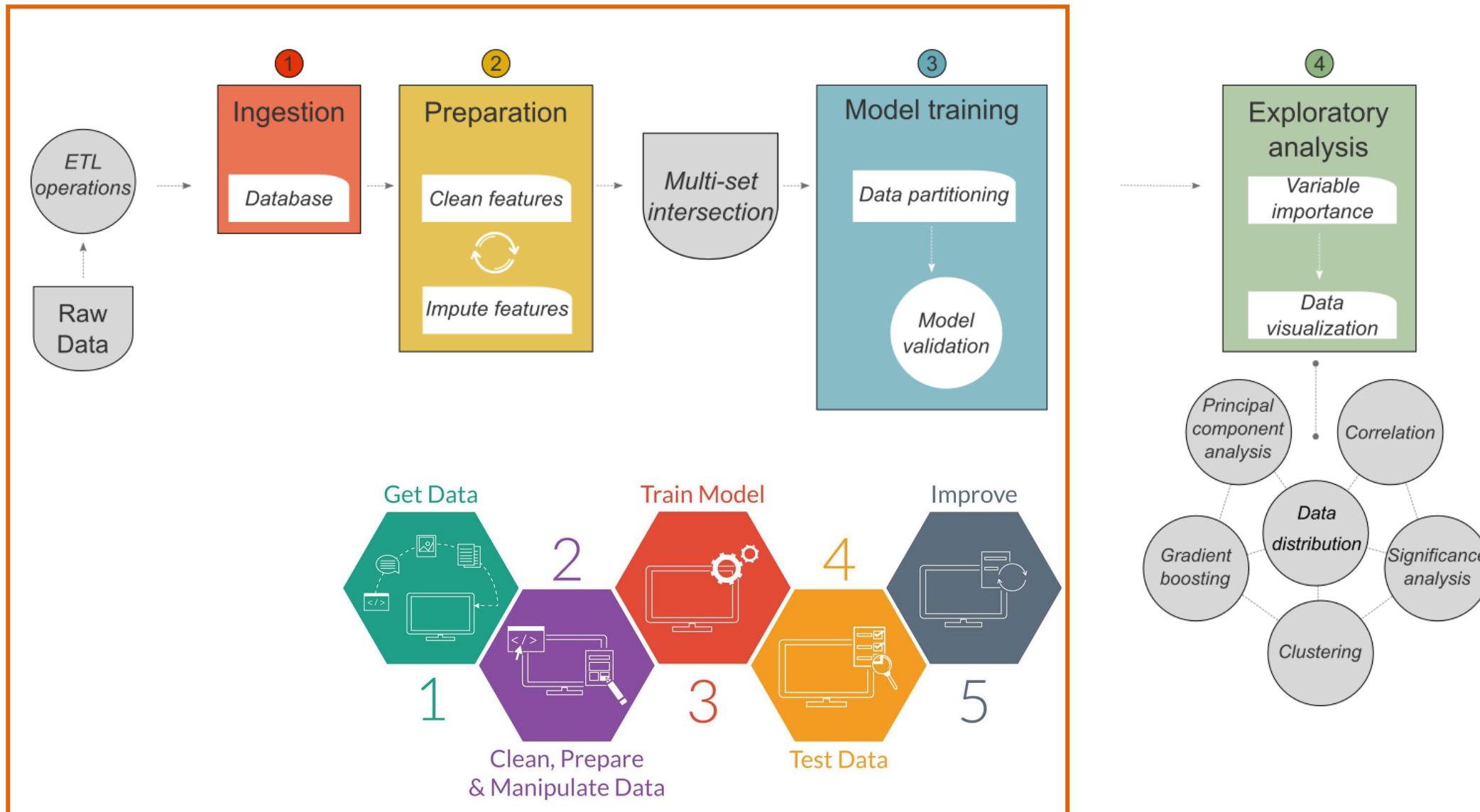
KNEXT barista robot

Perfect coffee every time

Part III. SIMON, pattern recognition and knowledge extraction software



SIMON machine learning process



Leading statistical programming languages in data science – available ML tools



R-project (<https://www.r-project.org/>):

- MLr3 (<https://mlr3.mlr-org.com>)
- Classification and regression training (**CARET**) (<https://rdrr.io/cran/caret>)



Python (<https://www.python.org/>):

- Scikit-learn (<https://scikit-learn.org>)
- mlPy (<https://mlpy.fbk.eu>)
- SciPy (<https://www.scipy.org/>)

Extensive programming experience and general knowledge of R or Python **essential**, making them inaccessible for many life science researchers

Deep learning libraries:



<https://www.tensorflow.org/>



<https://keras.io/>

Available ML software

Commercial software

- Google's cloud-based AutoML (<https://cloud.google.com/automl>)
- DataRobot (<https://www.datarobot.com/>)
- BigML (<https://bigml.com/>)
- MLjar (<https://mljar.com>)
- RapidMiner (<https://rapidminer.com/>)

Features

- Closed source – unknown/hidden ML methods and algorithms
- No specific algorithms to deal with biomedical datasets (missingness, heterogenous data types, etc)
- High price (DataRobot \$50k/licence!)

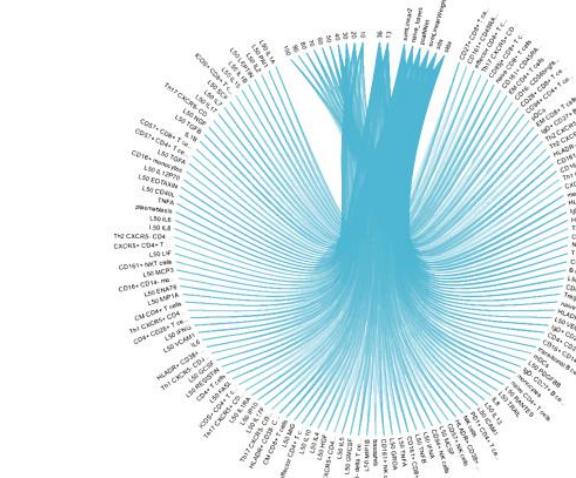
Academia-released software

- Waikato Environment for Knowledge Analysis (WEKA) (<https://www.cs.waikato.ac.nz/~ml/weka/>),
- Orange (<https://orange.biolab.si/>)
- Konstanz Information Miner (KNIME) <https://www.knime.com/>
- ELKI (<https://elki-project.github.io/>)

Features

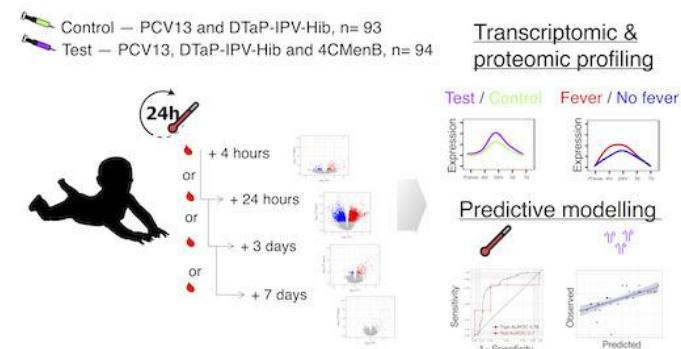
- Free and open source – explained/published ML methods and algorithms
- Requires knowledge of ML process
- Lack some of the advance features of commercial software (autoML)

Integrative analysis of different data types – predicting flu vaccine responses

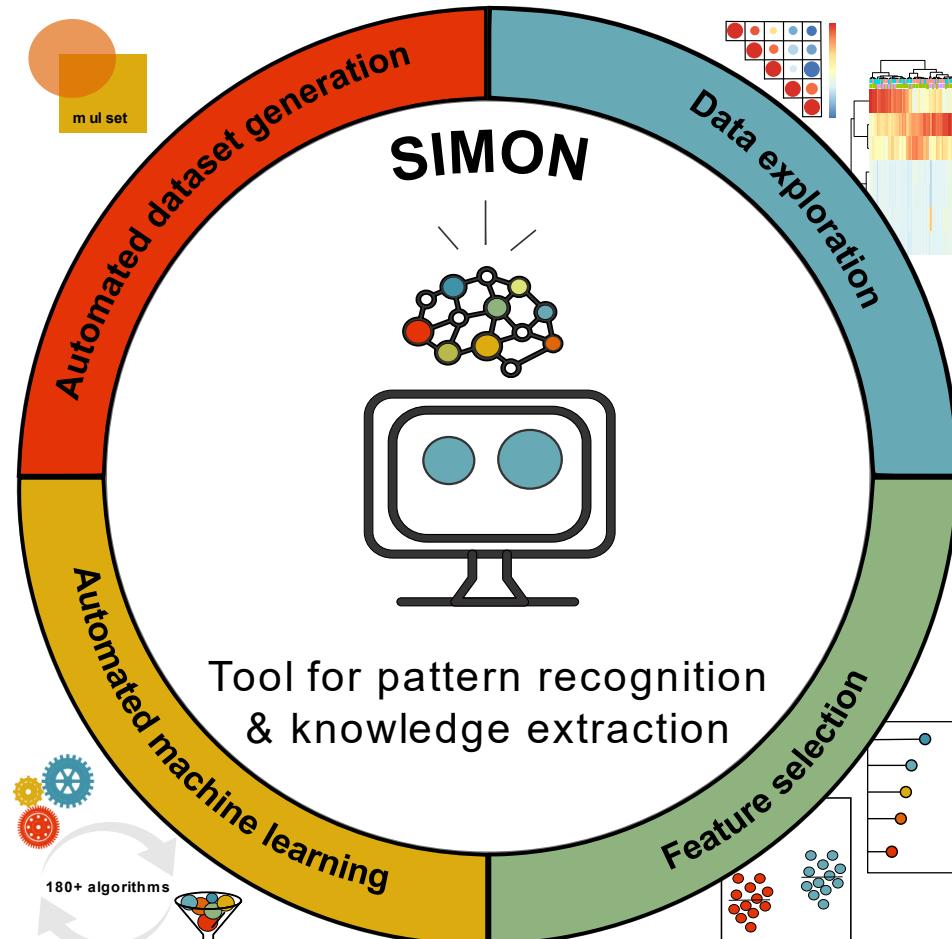


Tomic et al, JI, 2019

Transcriptome data

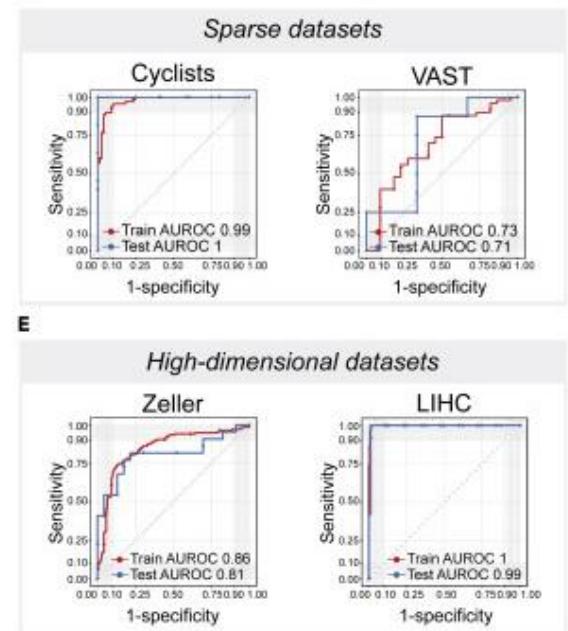


O'Connor et al, Mol Syst Biol, 2020

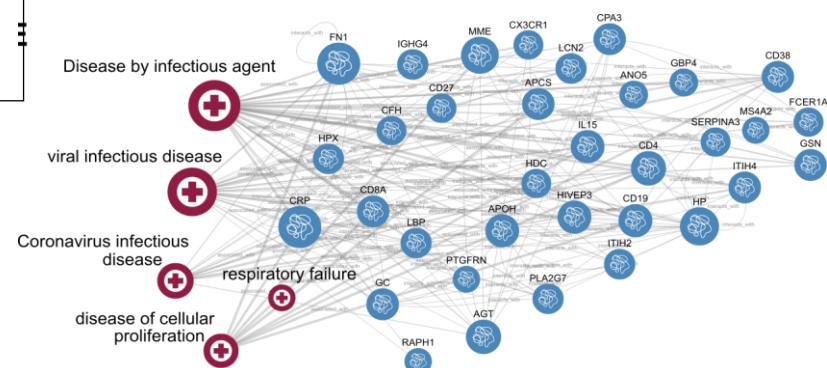


Tomic et al, Patterns, 2021

Datasets with high sparsity or high-dimensionality (transcriptome, microbiome)

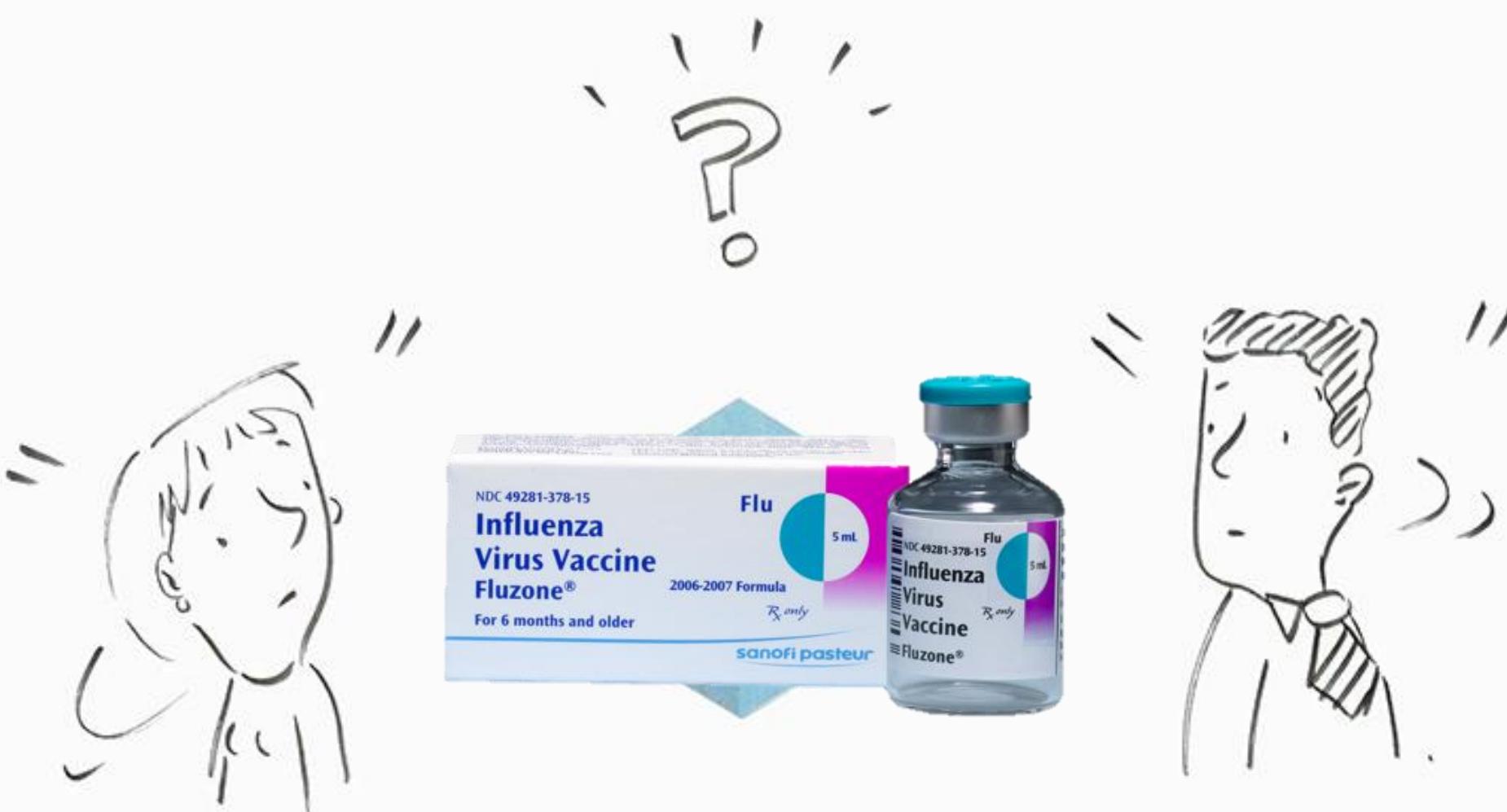


Multi-omics integrative analysis – COVID-19 COMBAT project



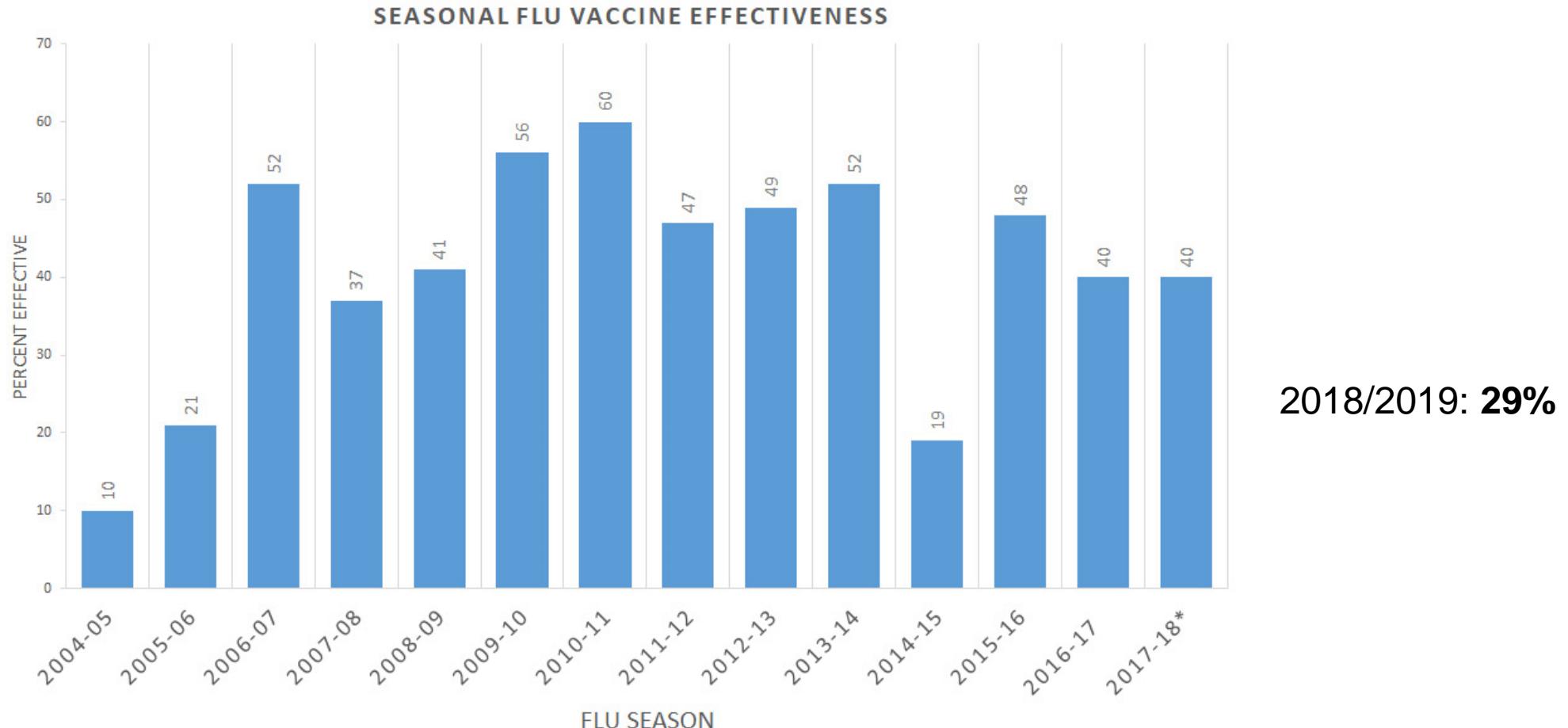
COMBAT consortium, Cell, 2022

*FluPRINT: Tracing the influenza vaccine imprint
on the immune system to identify cellular signature of protection*



Seasonal influenza vaccine: the least protective vaccine?

Influenza vaccine effectiveness in the US during the 2004 – 2019 Flu Seasons

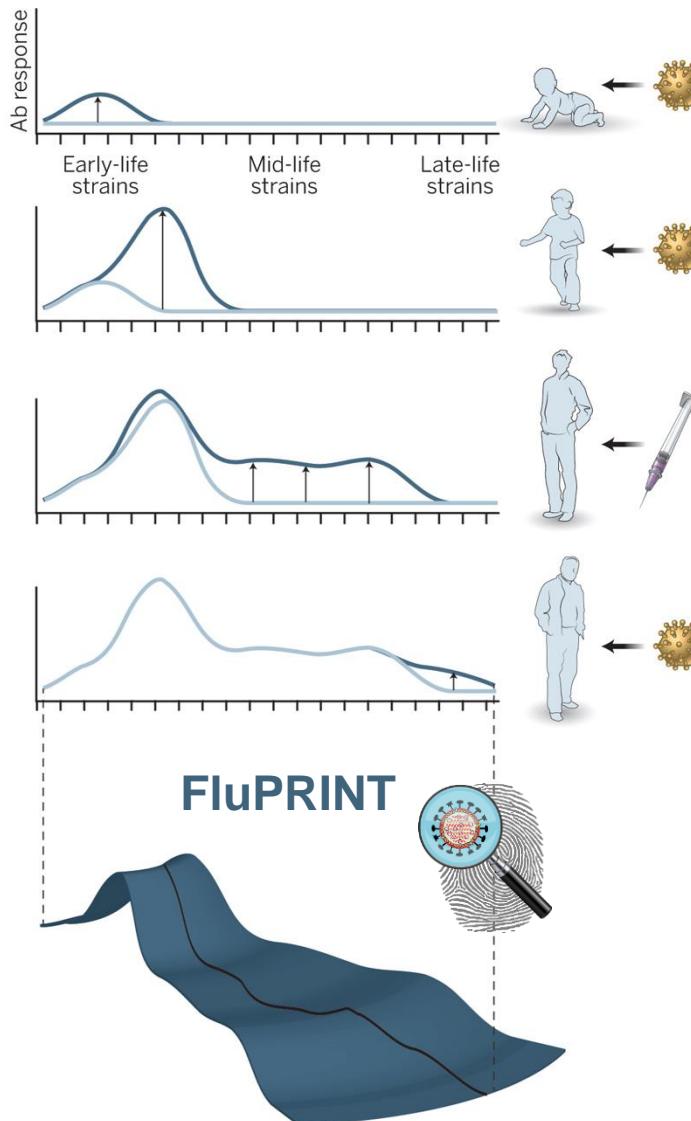


40%

*Complete match with the circulating strains!

Recalling the Future: Immunological memory impacts the outcome of subsequent infections

Imprinting - lifetime of exposure to influenza



How pre-existing immunity (*FluPRINT*) affects the quality of the immune response to subsequent influenza challenge?

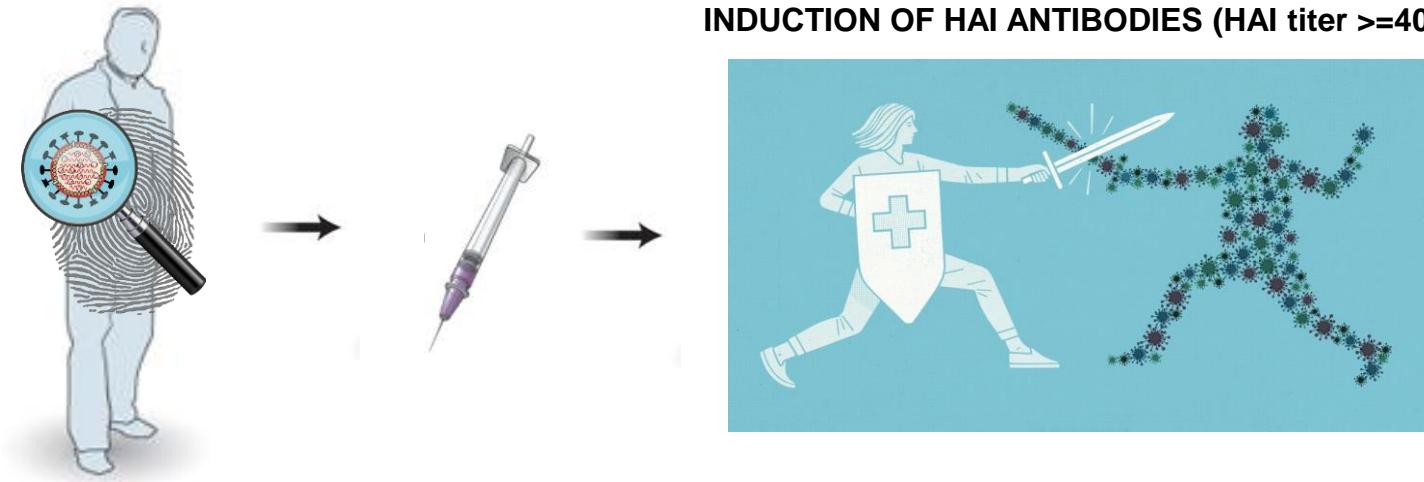
Can we predict if a person based on its *FluPRINT* will be protected against subsequent influenza challenge?

The model: inactivated flu vaccine

**FluPRINT – influenza immune imprint
(pre-existing immunity)**

- memory cells
- long-lived plasma cells
 - antibodies

...that were generated in a previous immune response to influenza virus



Hemagglutination inhibition (HAI) antibodies correlate with protection against influenza-like disease

Hobson, D., Curry, R. L., Beare, A. S. & Ward- Gardner, A. J. Hyg. 1972.

Machine learning approach to identify FluPRINT

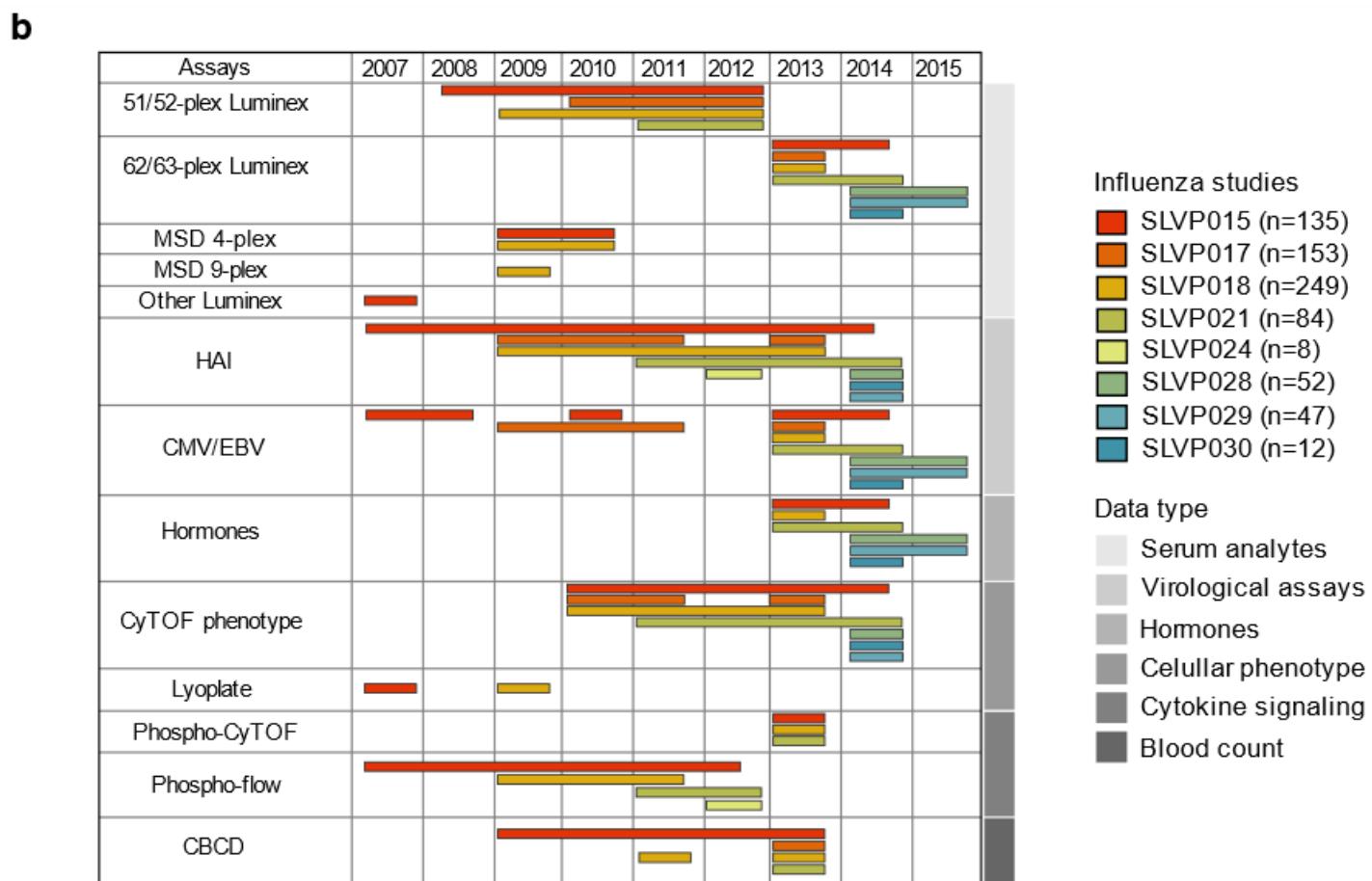
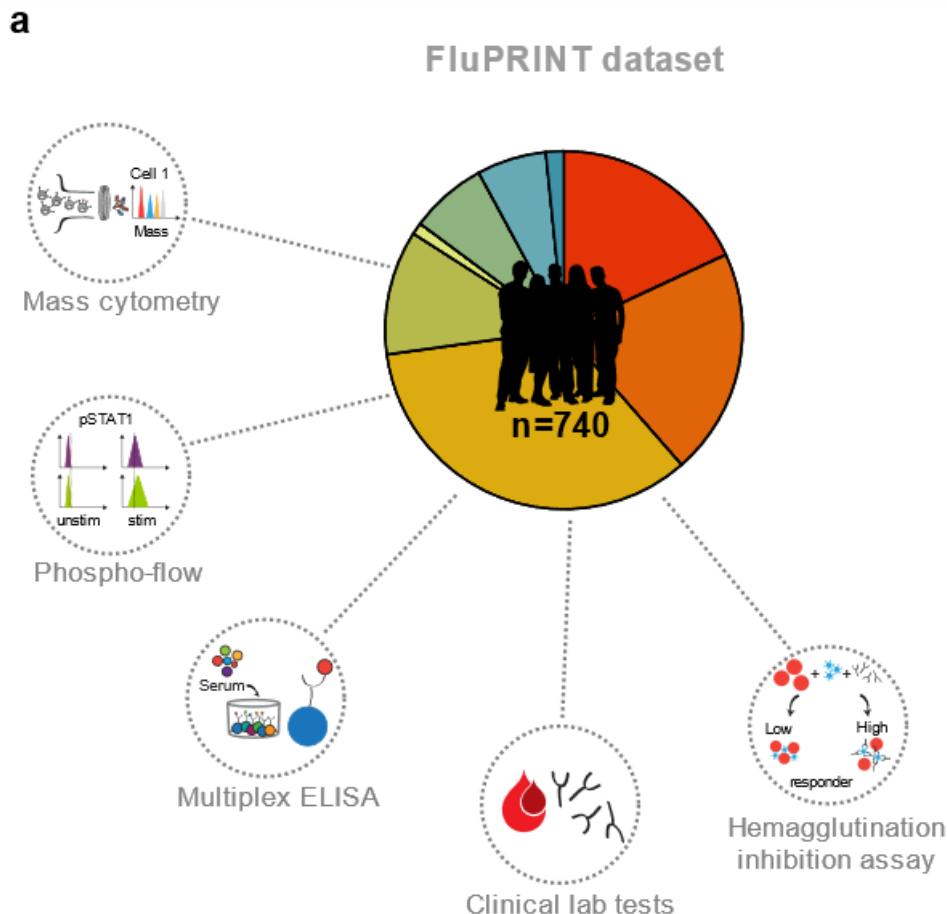


Hidden treasures at Stanford

Human Immune Monitoring Center (HIMC) – Stanford Data Miner

- 2007 – 2015
- 8 Flu clinical studies
- >700 unique donors – 0 to 90y
- >3000 parameters: Flow cytometry, CyTOF, PhosphoFlow, Multiplex ELISA, HAI, blood test, ...

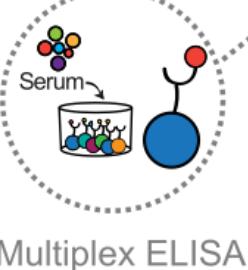
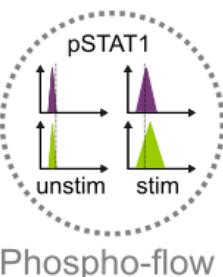
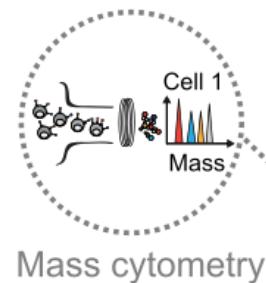
FluPRINT database – 740 individuals



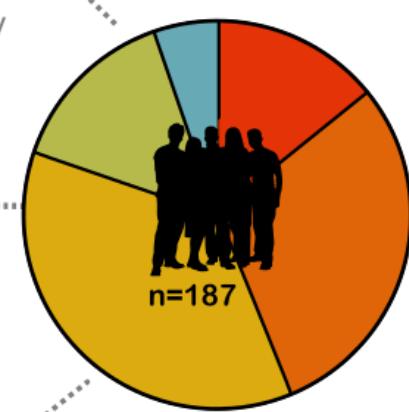
Which parameters correlate with increased antibody responses after immunization with inactivated influenza vaccine?

Predictors

Day 0

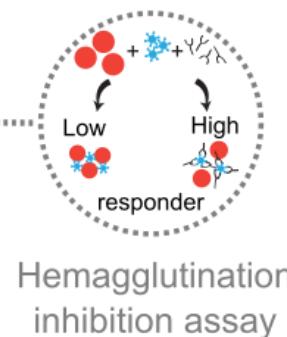


Influenza studies
from 2007 - 2014



- SLVP015
- SLVP017
- SLVP018
- SLVP021
- SLVP029

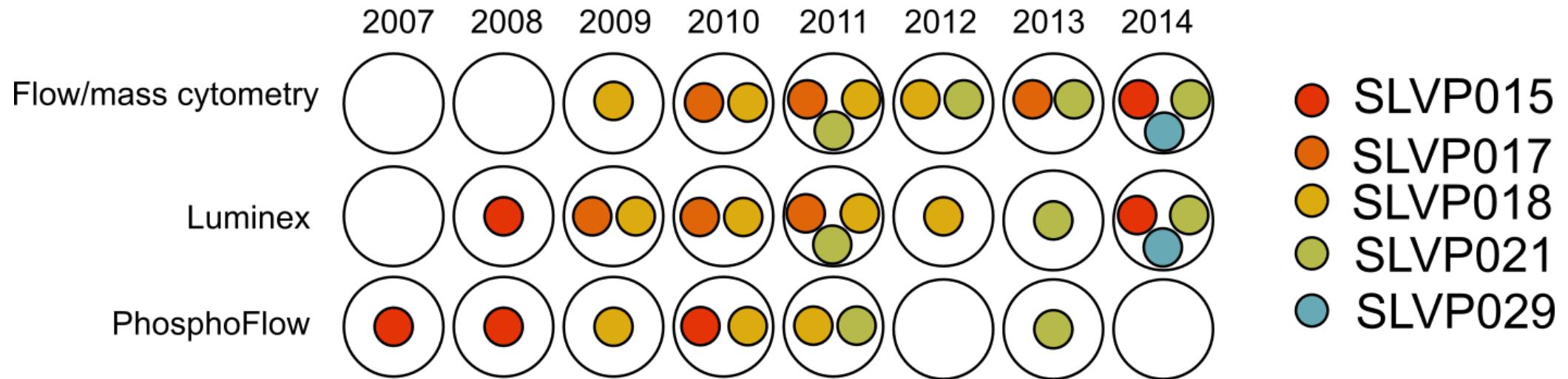
Outcome
Day 28 post vaccination



Stanford Human Immune Monitoring Center

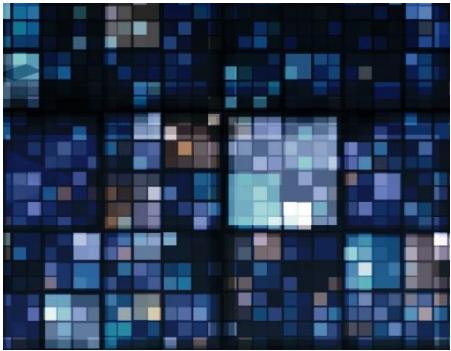


Dealing with missing values

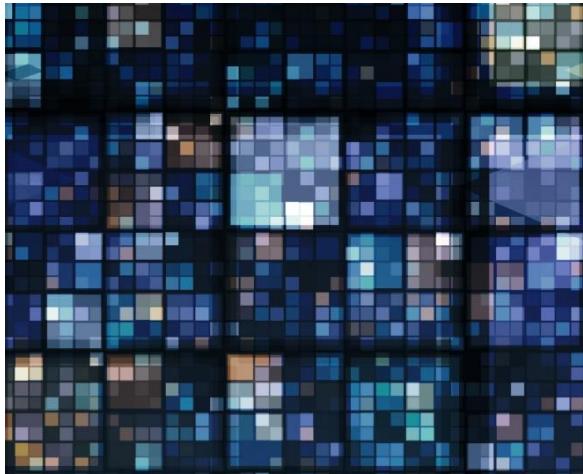


The “BIG” problem: Highly percentage of missing data

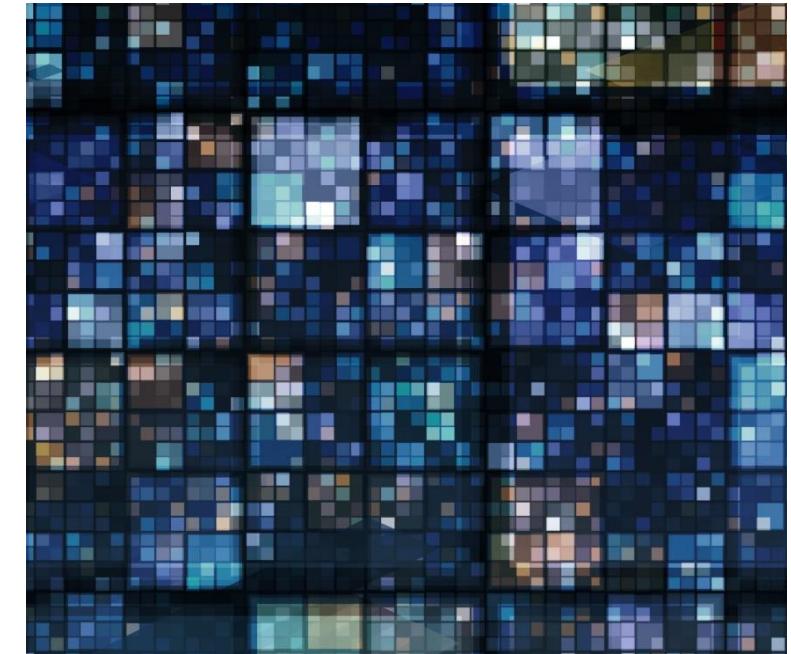
2007



2010



2014



How to select optimal number of donors and optimal number of features?

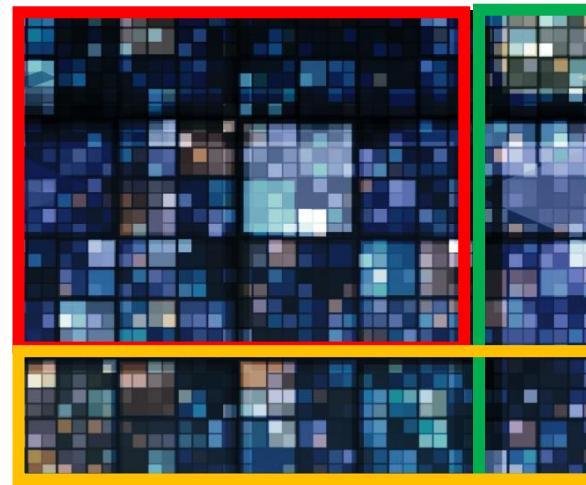
SUBSAMPLING

The alternative solution to cope with high sparsity

2007



2010



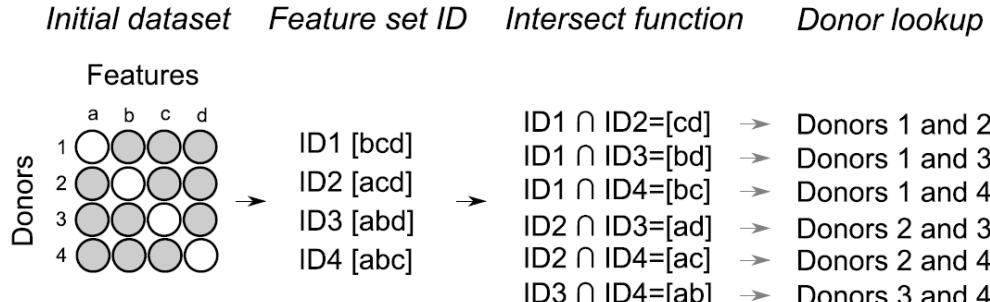
2014



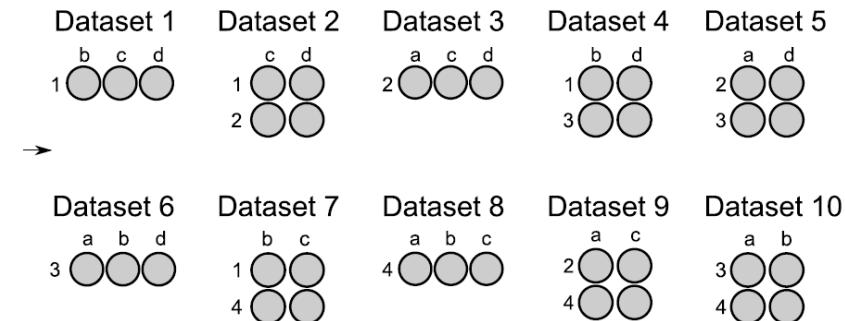
A fully automated script for feature subset selection, dimensionality reduction and data sampling

An R package *mulset*: A multi-set intersection function

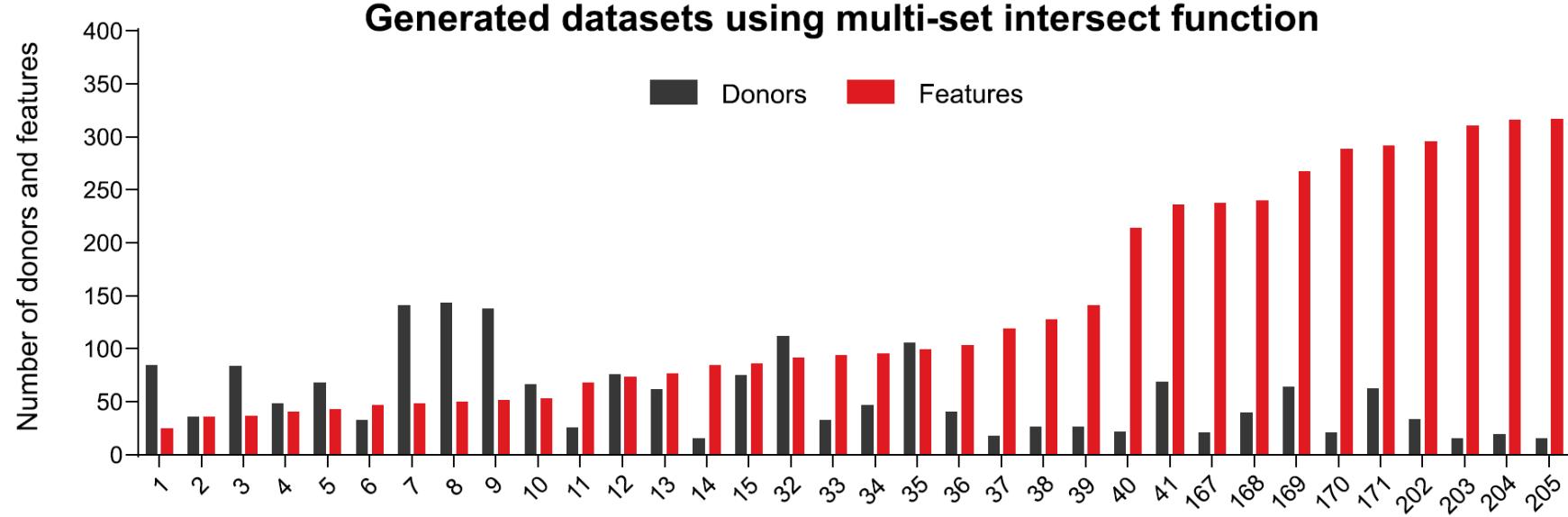
A



B



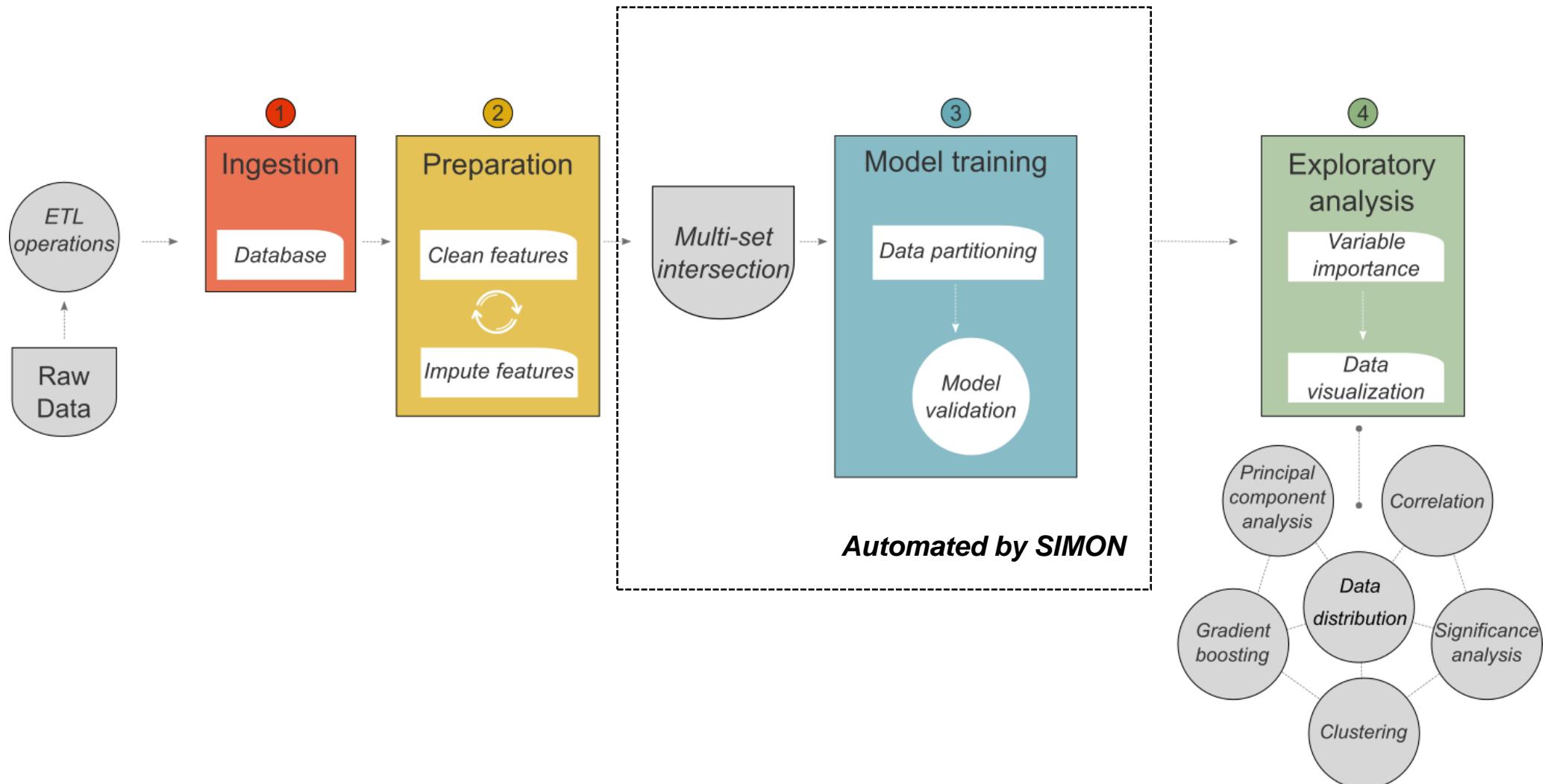
C



Machine learning models: Which one to use? Use all of them!

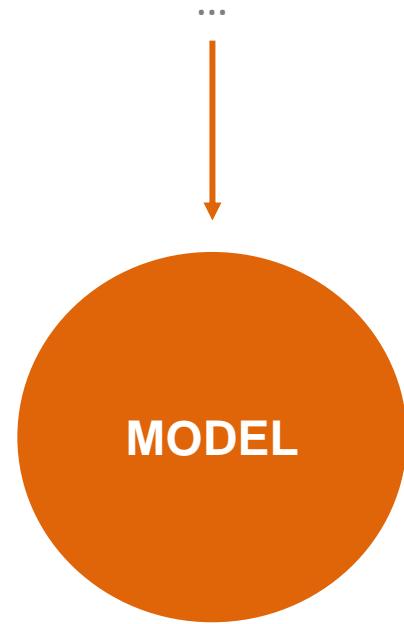
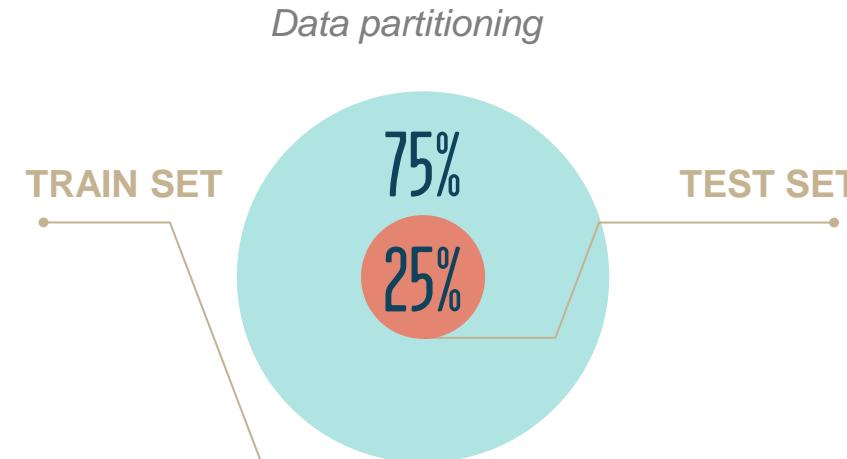
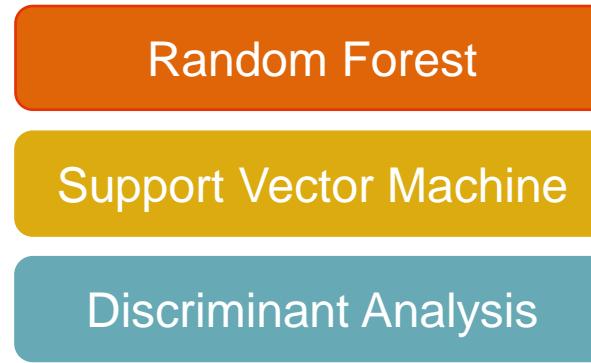


SIMON: Sequential Iterative Modelling OverNight

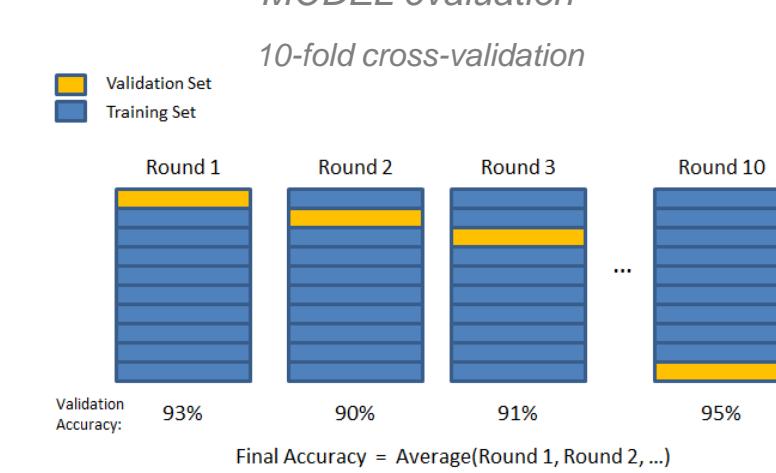


SIMON: Sequential Iterative Modelling OverNight

180+ machine learning algorithms



Validation Set
Training Set



Data partitioning

75%

25%

TEST SET

MODEL evaluation

10-fold cross-validation

Validation Set

Training Set

Round 1

Round 2

Round 3

Round 10

Validation
Accuracy:

93%

90%

91%

95%

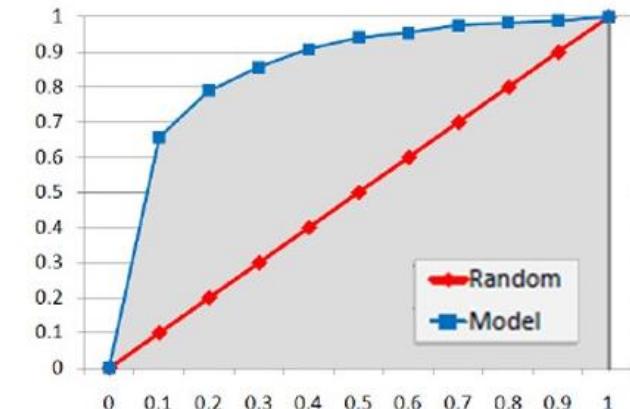
Final Accuracy = Average(Round 1, Round 2, ...)

Model selection

Confusion matrix

n=165	Predicted: High responders	Predicted: Low responders
Actual: High responders	50	10
Actual: Low responders	5	100

AUROC



```
2
3
4 // String table management
5 typedef struct {
6     int sid;
7     char *str;
8     UT_HashNode nh;
9 } cff_sid_entry;
```

2,400 machine learning analysis run on **34** datasets

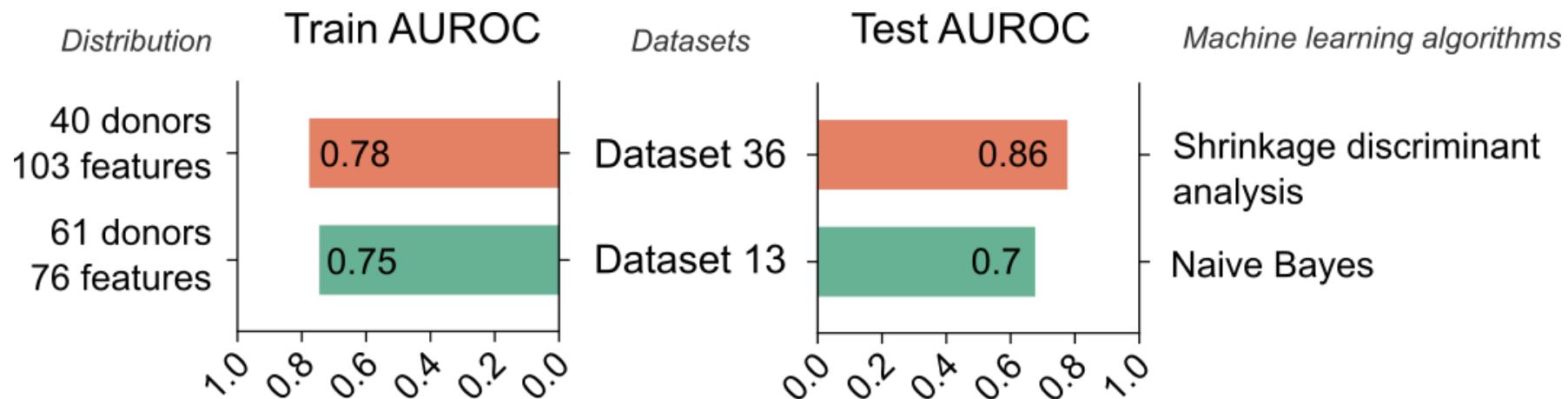
```
1 static int simon_ml_models(UT_Hash *h) {
2     Models were built for 19 datasets, with an average of 54 models built per dataset
3     cff_sid_entry *item = NULL;
4     HASH_FIND_STR(*h, s, item);
5     if (item) {
6         return 391 + item->sid;
7     } else {
8         NEW(item);
9         item->sid = HASH_COUNT(h);
10        item->str = sdssdup(s);
11        HASH_ADD_STR(*h, str, item);
12        return 391 + item->sid;
13    }
14 }
```

After model selection, **2** datasets with average of **3** models built per dataset

SIMON facilitated exploratory analysis

and discovery of high-performing models

SIMON results: 2 datasets with the highest accuracy

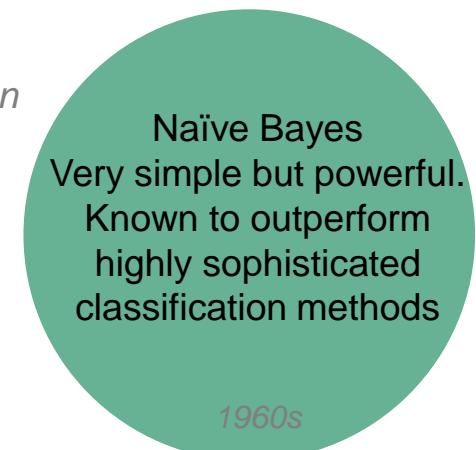


Bankruptcy prediction

Image recognition

Marketing

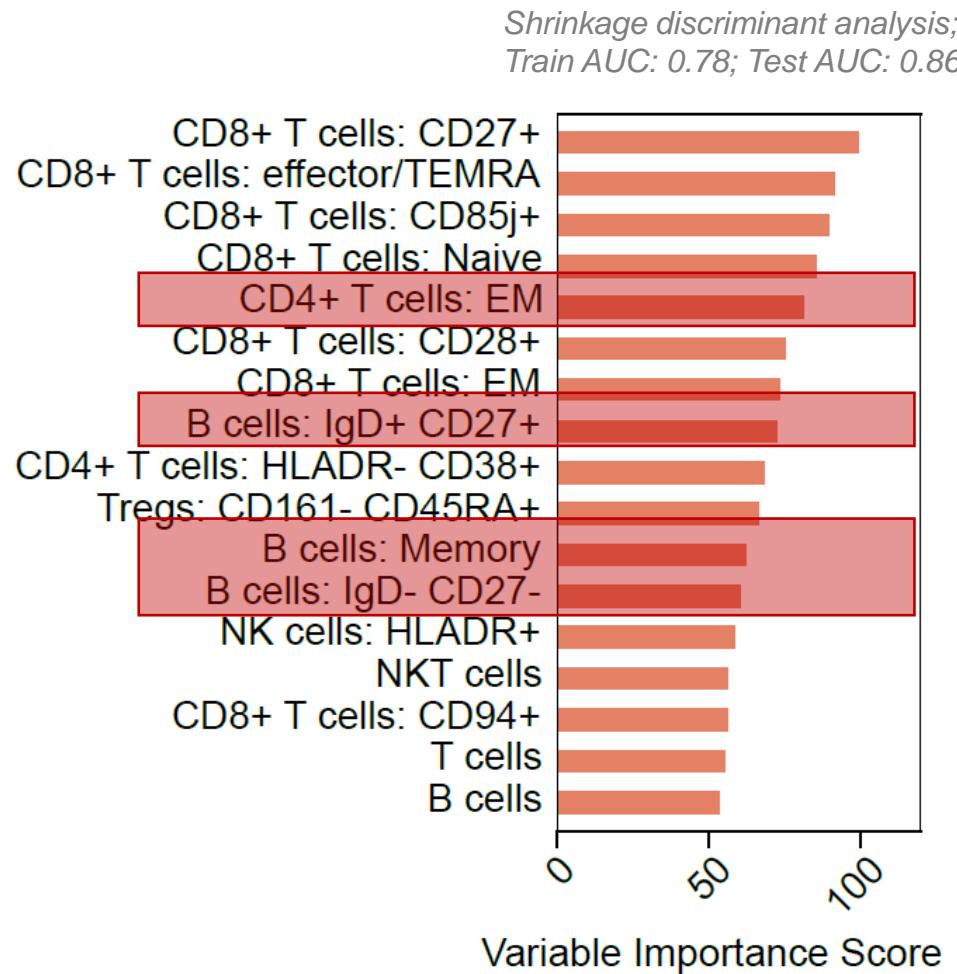
Mkhadri A, Pattern Recognition Letter 1995



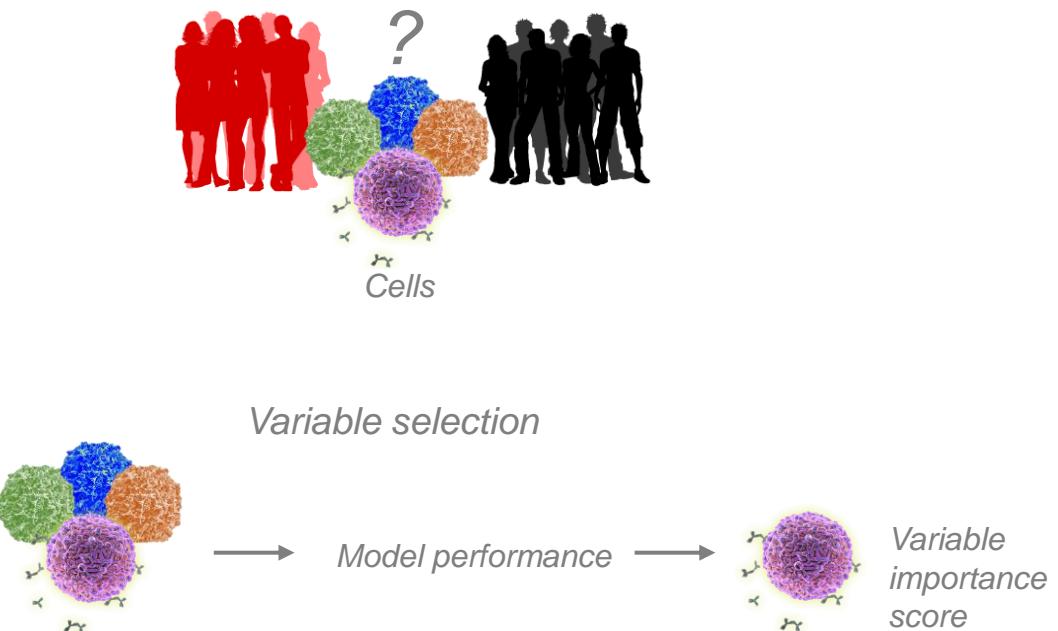
Train AUROC – mean value determined from confusion matrix after 10-fold cross-validation (repeated 3 times)

Test AUROC – evaluated from confusion matrix on independent test set

Pattern recognition in influenza vaccine study using SIMON

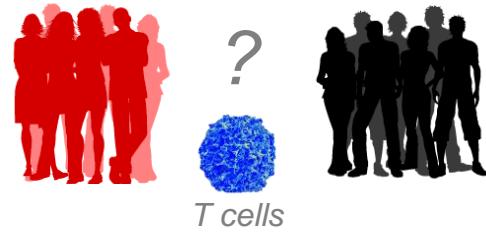


Is there a difference in the frequency of immune cells between high and low responders?

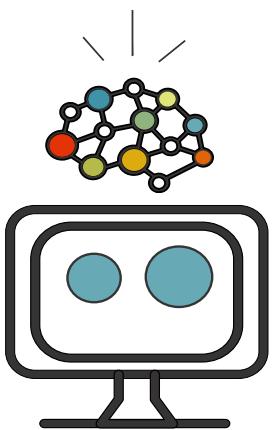


Hypothesis

Why is frequency of T cells increased among healthy vs infected person?



Data analysis



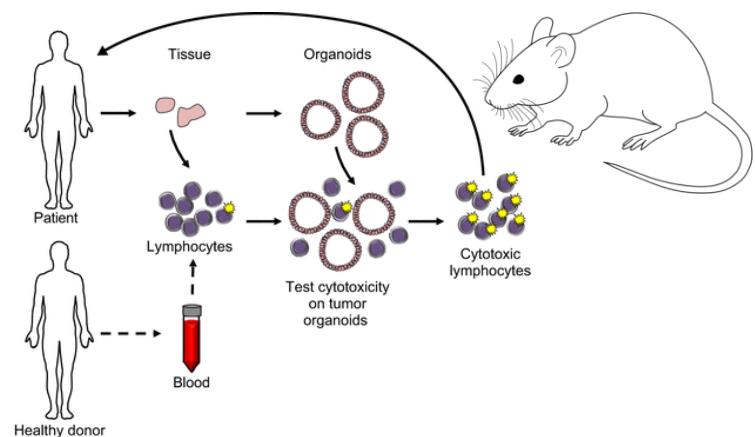
SIMON*

- Feed dataset
- 180+ machine learning algorithms
- Build 1000s of models in one click
- Explore top models
- Identify top important variables

Data-driven research

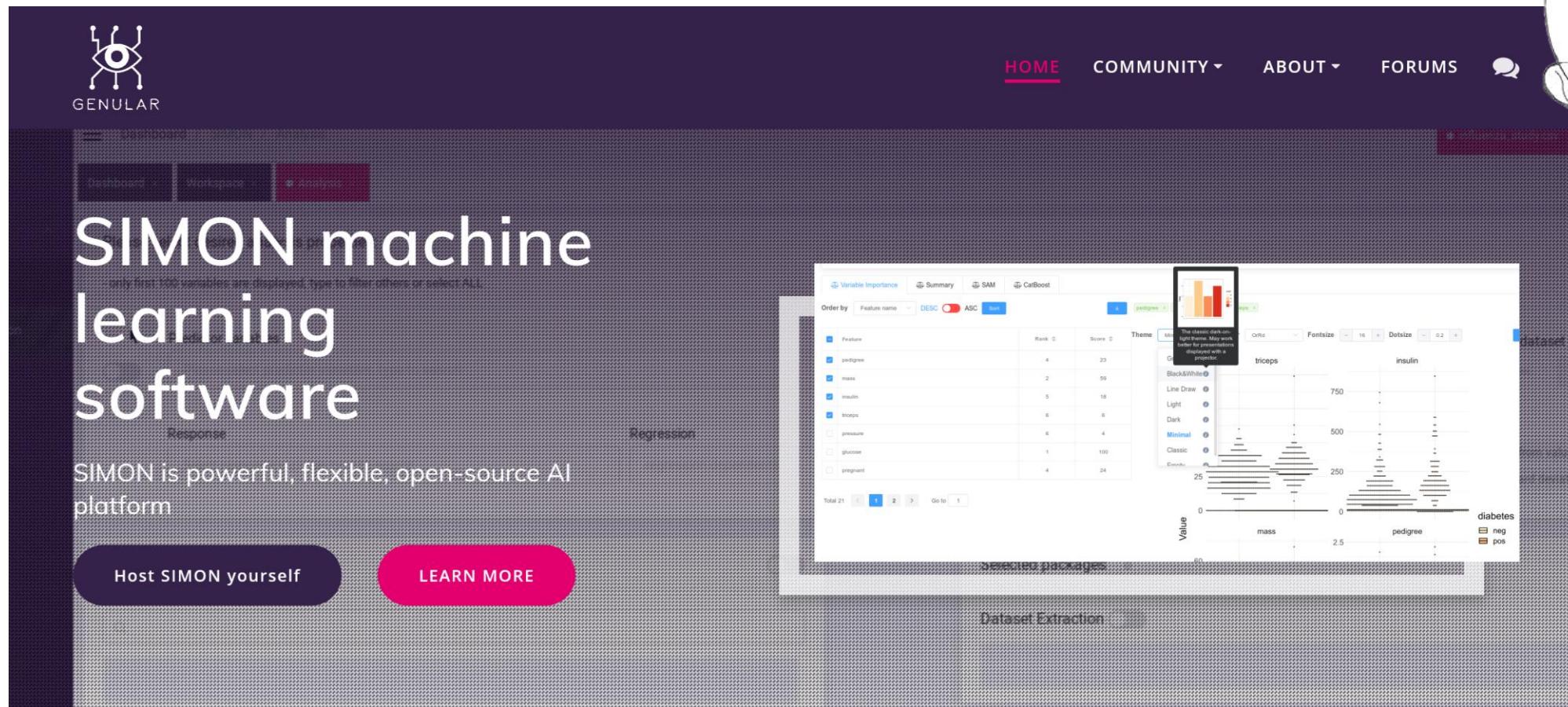
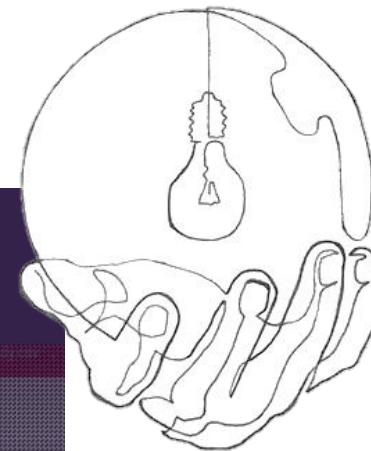
Experiments

Assays to confirm phenotype and reveal new mechanisms of T cells In both groups



Drost and Celevers, Development, 2017

Join open-source community supporting SIMON!



The screenshot shows the SIMON machine learning software homepage. At the top, there's a navigation bar with links for HOME, COMMUNITY, ABOUT, FORUMS, and a message icon. On the left, there's a logo with a stylized eye and the word "GENULAR". The main content area features a large title "SIMON machine learning software" with "Regression" and "Classification" sub-sections. Below the title, a sub-headline says "SIMON is powerful, flexible, open-source AI platform". There are two buttons at the bottom: "Host SIMON yourself" (dark purple) and "LEARN MORE" (pink). The background has a dark, textured pattern.

SIMON machine learning software

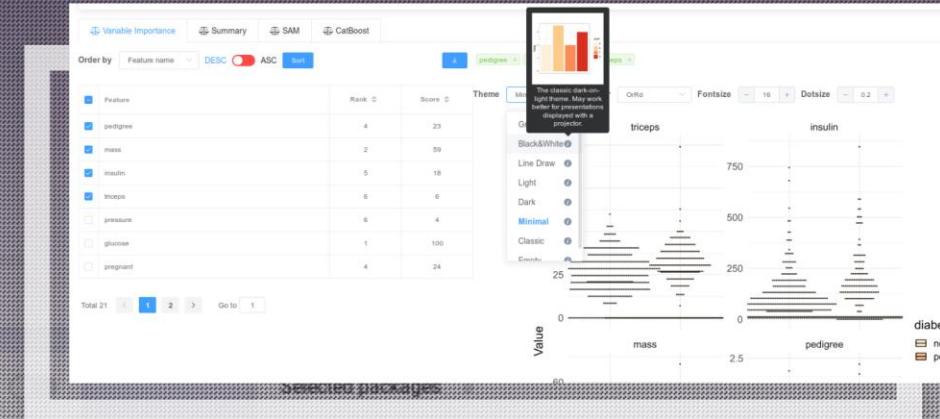
Regression Classification

SIMON is powerful, flexible, open-source AI platform

Host SIMON yourself LEARN MORE

HOME COMMUNITY ABOUT FORUMS

GENULAR



Check out SIMON at genular.org



GENULAR

HOME

COMMUNITY ▾

ABOUT ▾

FORUMS

SIMON Knowledge Base

Have a Question?

Search the documentation...

Search

Installation

📄 Installation Quickstart

Machine Learning

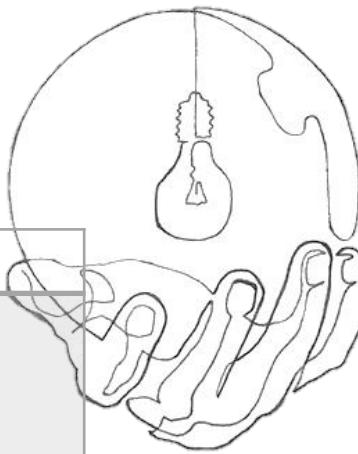
📄 How to perform SIMON analysis?

📄 Instruction videos



Star us on
GitHub!!!

Join open-source community supporting SIMON!



Project	How To Help	Next Step
Localization (English, German, French, Chinese, Arabic)	Help us translate SIMON into your language. If some translation is missing or incorrect you can easily help us by correcting it.	Join our Translation Community
Tutorials	Help others use and understand SIMON	Write a tutorial or record it, with usage examples
Organizing	Ask questions on recently opened GitHub issues to move the discussion forward	Go to GitHub Issues
Write article	Help other understand what is Machine Learning & how can they apply it, by publishing blog post	e-mail us



Check out SIMON at **genular.org**



Mark M. Davis Lab
Department of Microbiology and Immunology

Elsa Sola Verges
Allison Nau
Lisa Wagar



Stanford-LPCH
Vaccine Program

Cornelia L. Dekker



Institute for Immunity, Transplantation and Infection

The Human Immune Monitoring Center

Mike Leipold
Yael Rosenberg-Hasson
Janine Bodea Sung
Holden Maecker



Institute for Immunity, Transplantation and Infection

Biomedical Data Science

Purvesh Khatri



National Institutes of Health

Turning Discovery Into Health



*Marie Skłodowska-Curie grant
(FluPRINT, Project No 796636)*



Thank you



Ivan Tomic



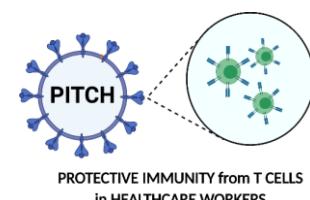
Max Kuhn



Duke
UNIVERSITY

Rachel L. Spreng
Lindsay C. Dahora
Kelly E. Seaton
Georgia Tomaras

Ross D. Pollock
Norman R. Lazarus
Stephen D.R. Harridge



Florent Ailloud