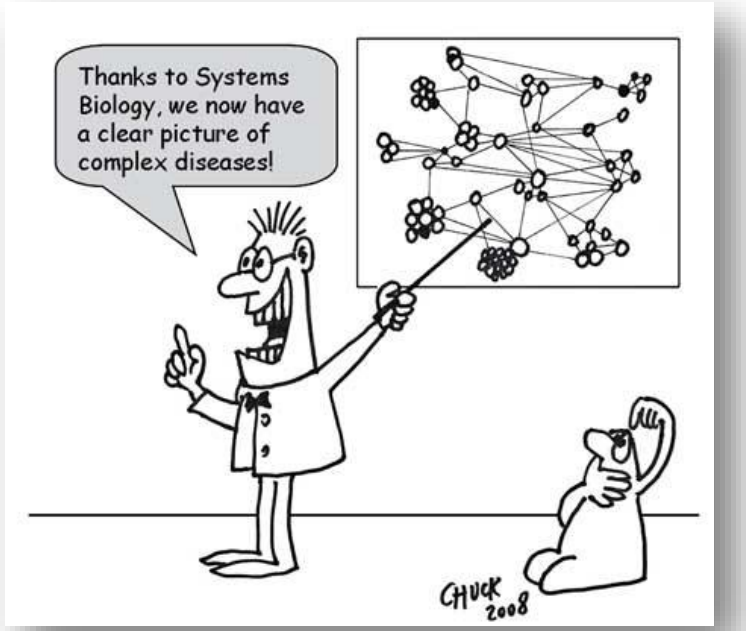
Systems immunology: an intro to multi-omics data integration and machine learning



Adriana Tomic
Systems Immunology | Oxford Vaccine Group

@TomicAdriana
atomiclaboratory
adriana.tomic@paediatrics.ox.ac.uk

SIMON training course @MSD

Part I – March 28th, 2022

Training course - overview

Part I – SIMON, pattern recognition and knowledge extraction platform (March 28th 2022)

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

Theoretical part (10-11am) ~1h

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

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

- SIMON installation
- Perform SIMON analysis using provided dataset
- Performance metrics, evaluation and selection of high-quality models

Hands-on (1-3pm) ~2h

Part II – Exploratory analysis (March 29th 2022)

- Feature selection: scoring and elimination
- Correlation and clustering analysis

Hands-on (9:30-11:30am) ~2h

- Feature processing methods to avoid '*curse of dimensionality*'

Theoretical part (1-1:30pm) ~0.5h

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

Case study (1:30-2pm) ~0.5h

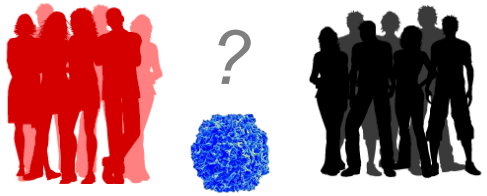
- Discussion about project-specific problems

Discussion with practical examples ~1h

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

Hypothesis

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

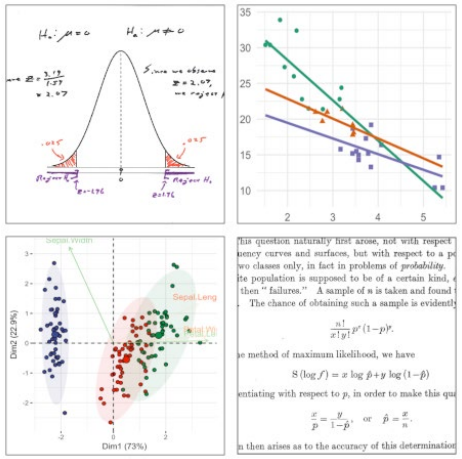


T cells

Hypothesis-driven research

Data analysis

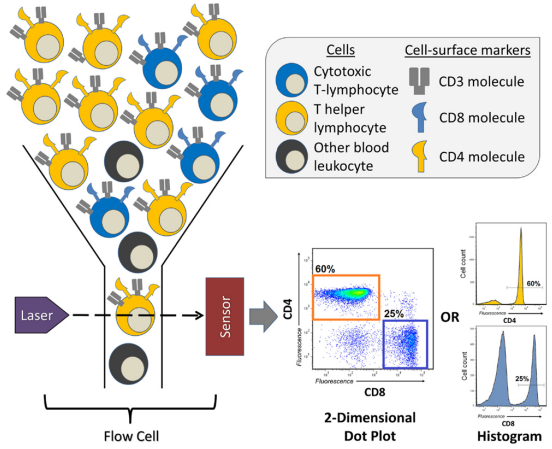
Comparison, statistics



https://statistics.rutgers.edu/

Experiments

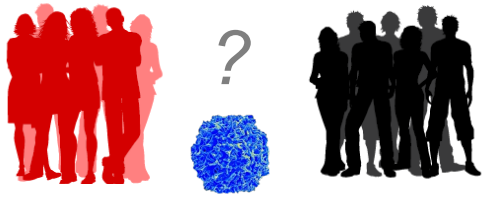
Assays to evaluate frequency, phenotype and functionality



Verschoor C et al, Front Immunol, 2015

Hypothesis

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



T cells

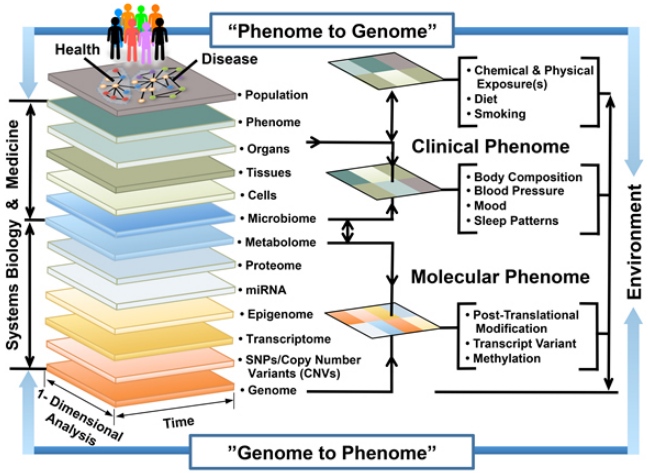
Data analysis

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

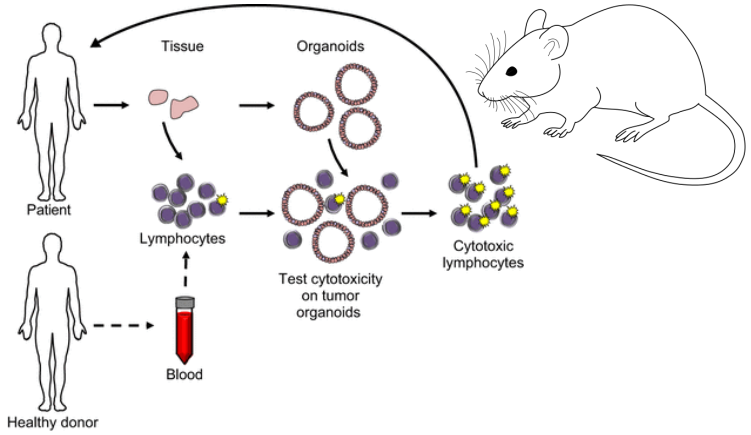
Data-driven research

Experiments

Assays to confirm phenotype and reveal new mechanisms

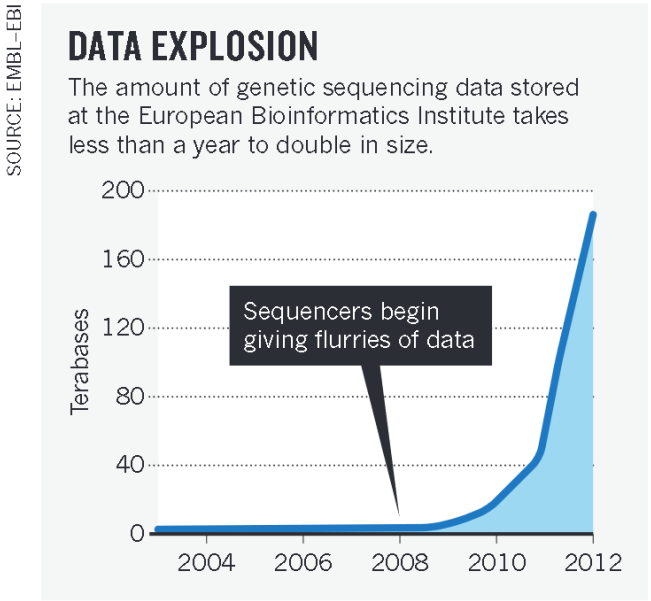


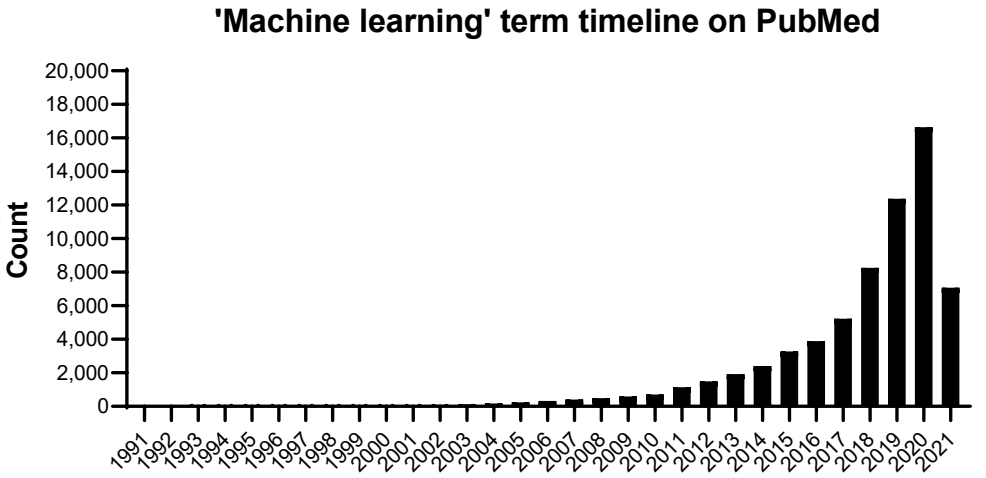
https://medicine.umich.edu/dept/dcmb/research



Drost and Celevers, Development, 2017

Biology's Big Problem: From data to knowledge



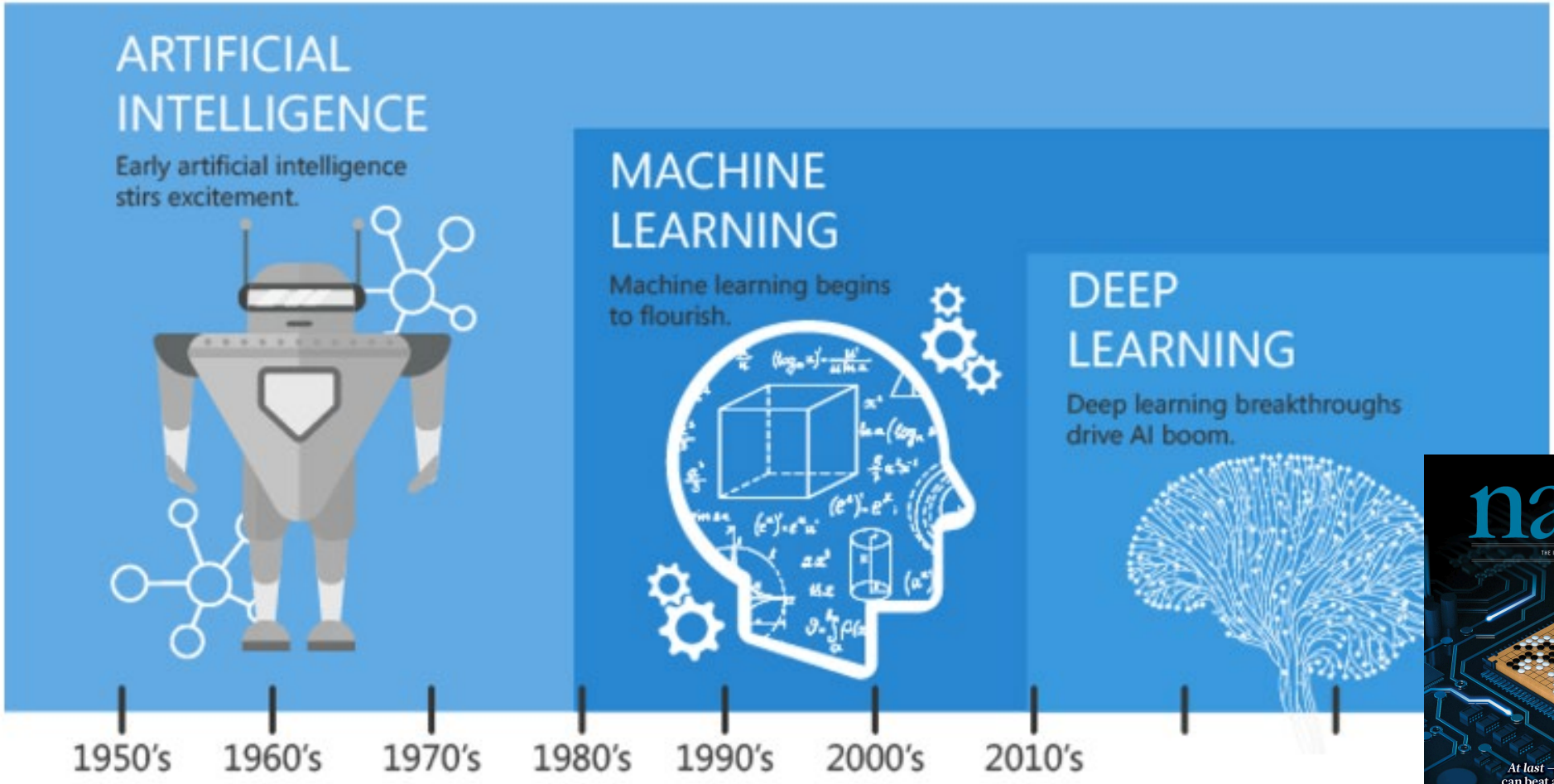


273 petabytes of data

(1 petabyte is 1000 terabytes or million gigabytes)

(2018, The European Bioinformatics Institute, EMBL in UK)

Artificial intelligence (AI) to the rescue

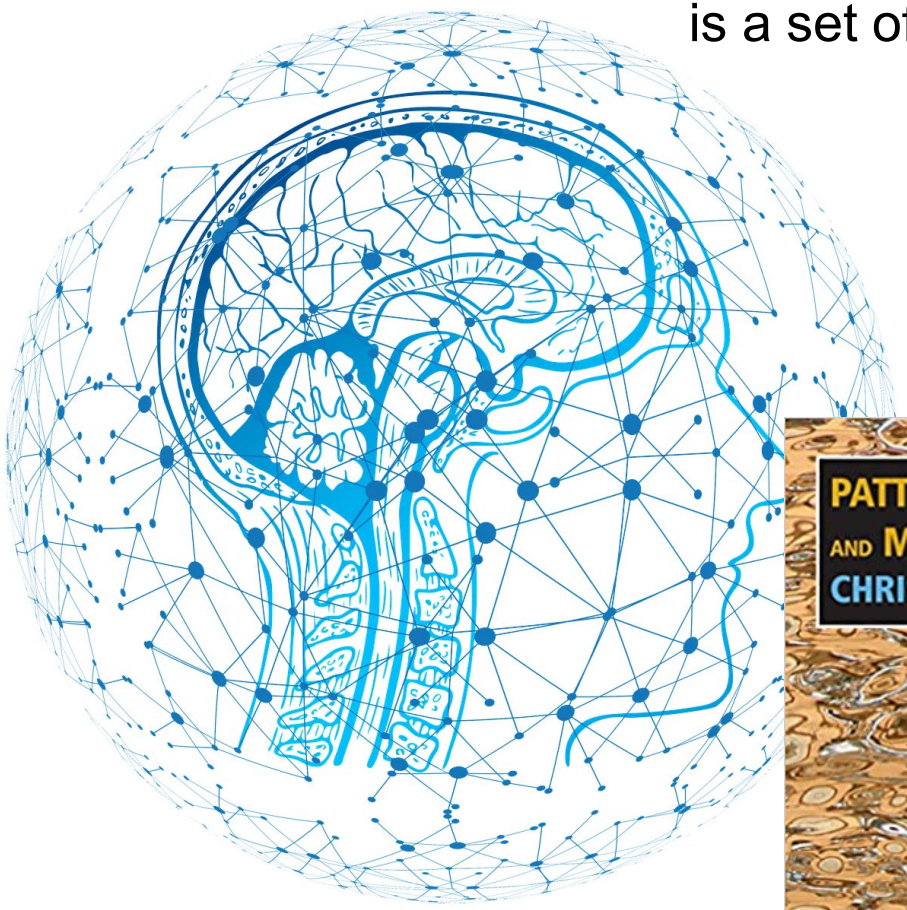


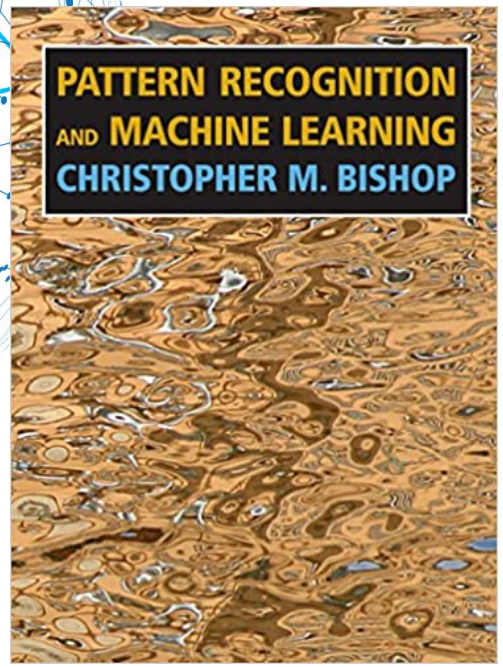
Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

Kapil Tandon. AI & Machine Learning: The evolution, differences and connections.

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

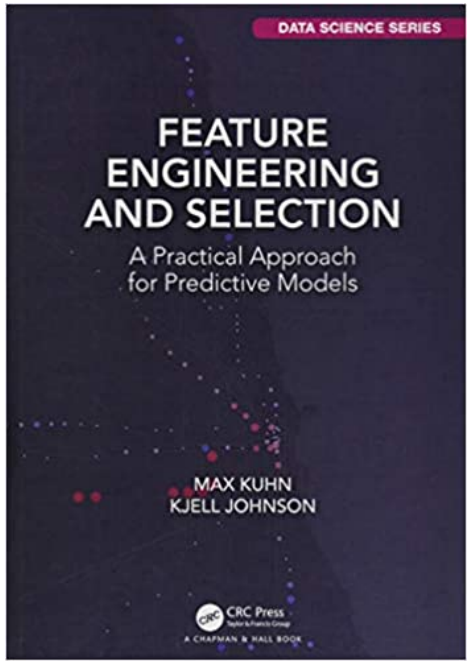
**even when the expert knowledge is incomplete*





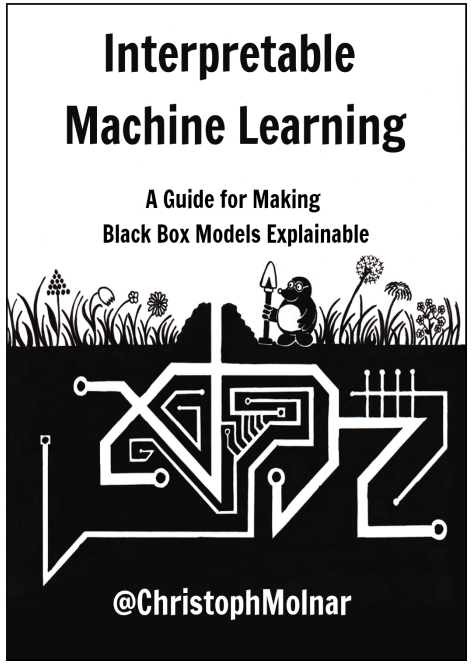
Christopher Bishop; Springer-Verlag New York: 2006

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



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

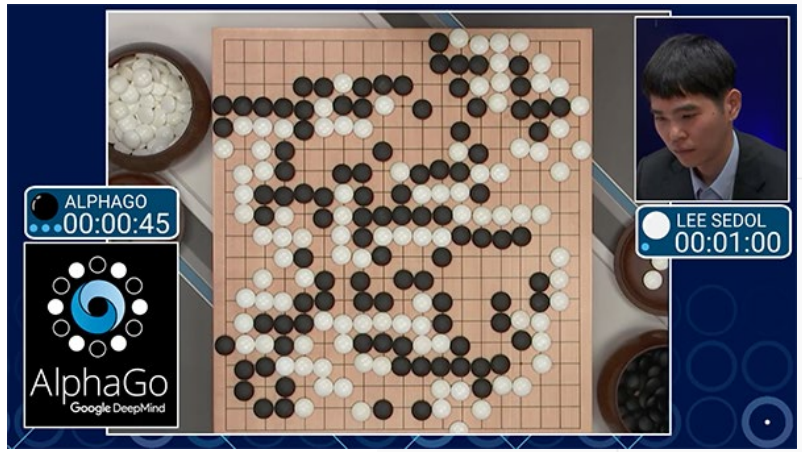
Free book online:
https://christophm.github.io/interpretable-ml-book/index.html



Christopher Molnar; 2021

We can teach computers to …

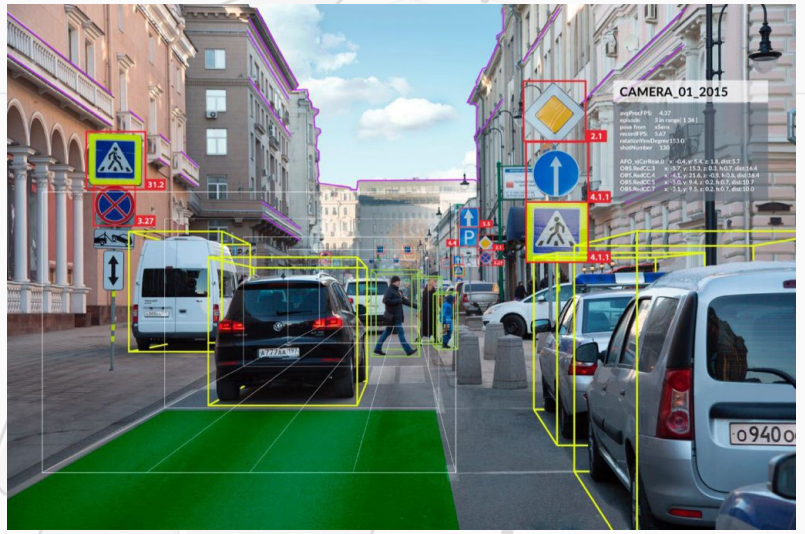
… play (and win) game of GO



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

Search The Guardian UK edition
For 200 years
News website of the year

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

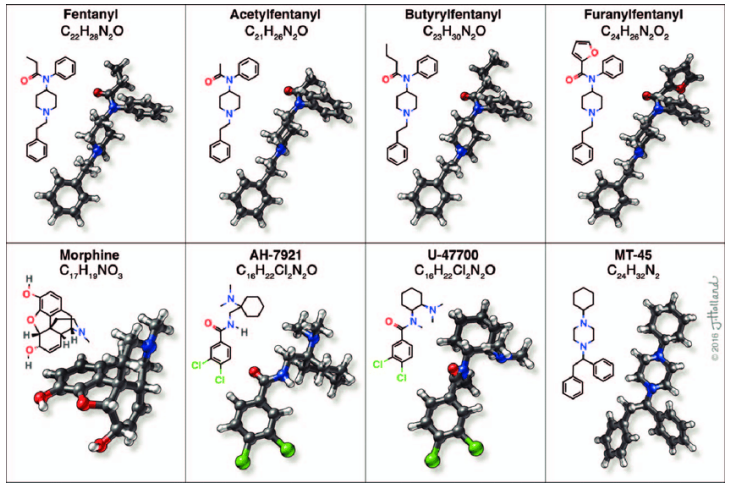
'We are not plotting to take over the human populace.' Photograph: Volker Schlichting/Getty Images/EyeEm

I 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)

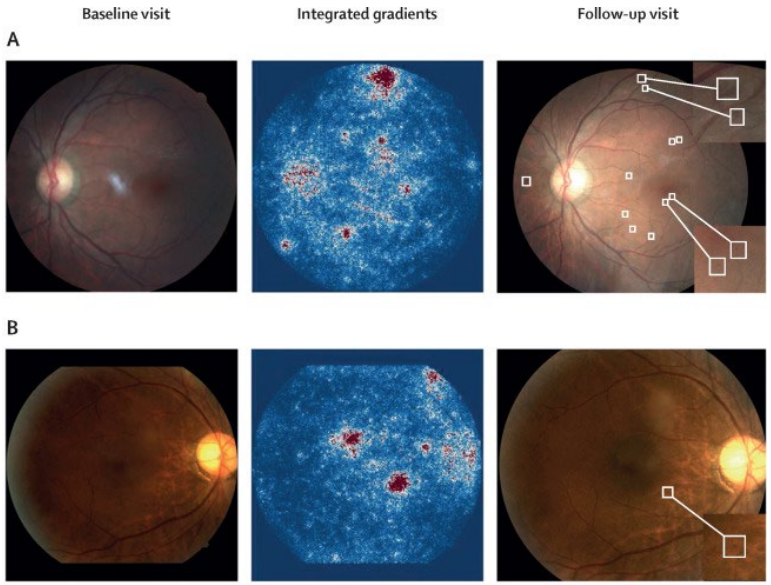
Machine learning outperforms human experts

Toxicity prediction from chemical structures



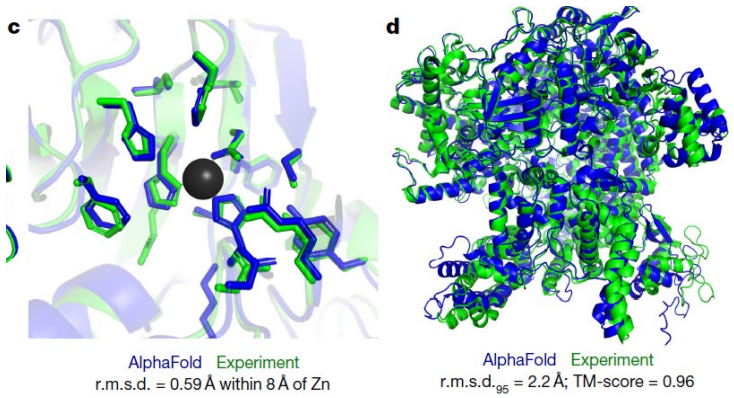
Eduati et al, Nat Biotech, 2015

Diabetic retinopathy prediction



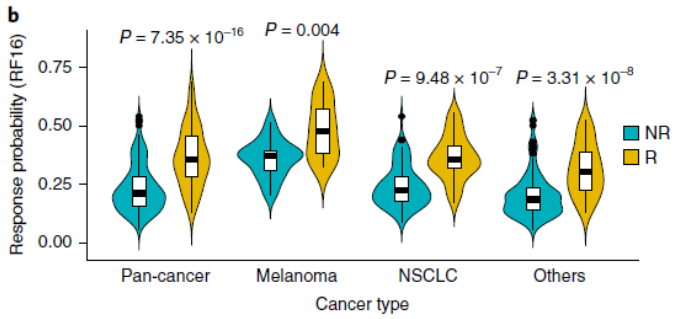
Bora et al, Lancet, 2021

Protein structure prediction (AlphaFold)



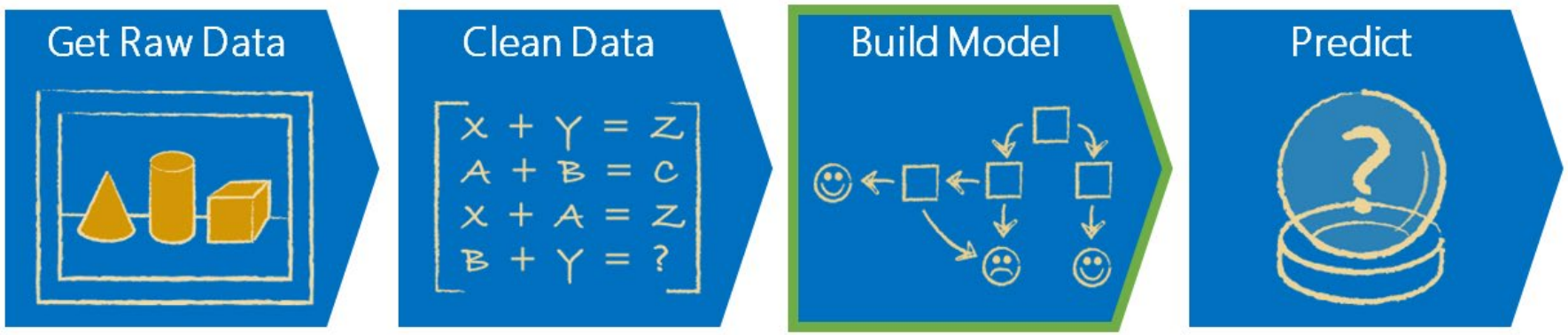
Jumper et al, Highly accurate protein structure prediction with AlphaFold, Nature 2021

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

Unsupervised ML

- Clustering
- Dimensionality reduction

cat

cat

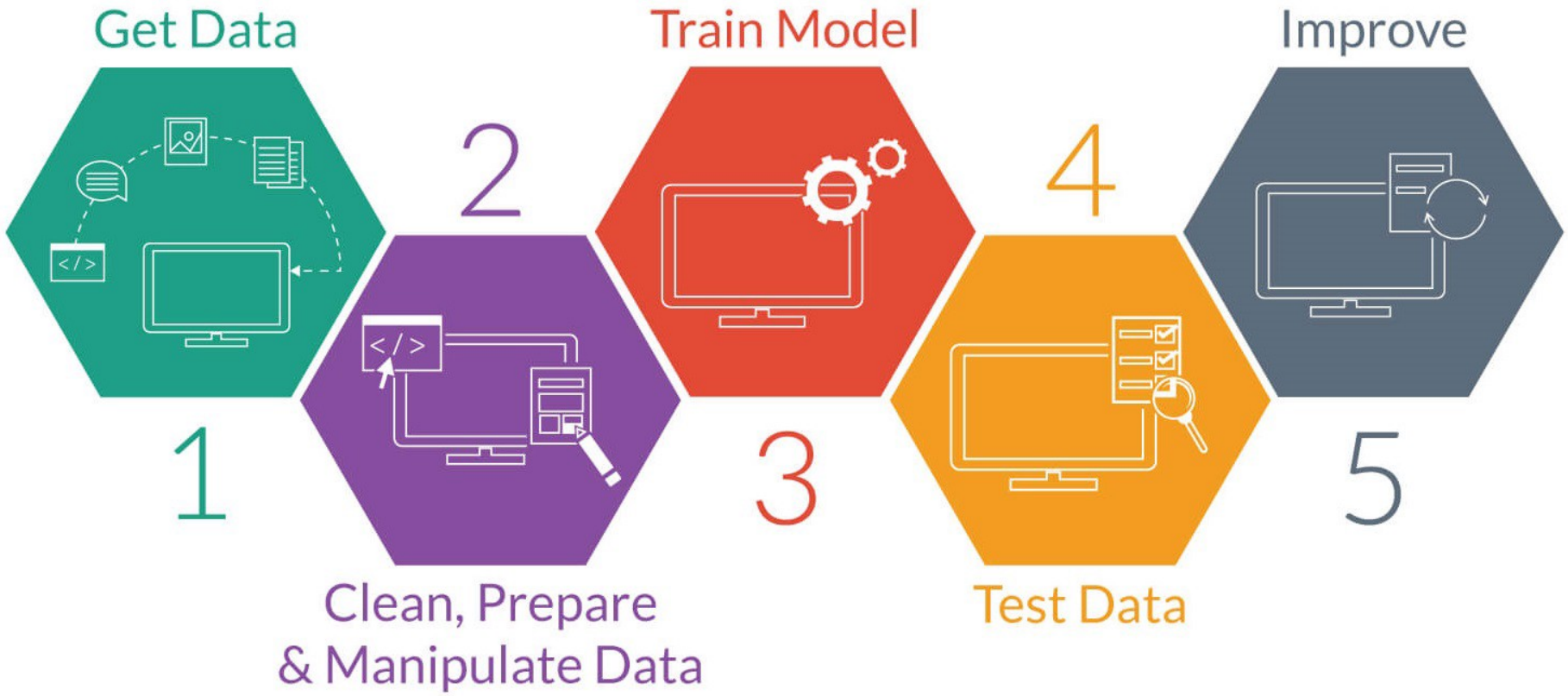
cat

dog

What is this?

Separate into 2 clusters!

Machine learning process

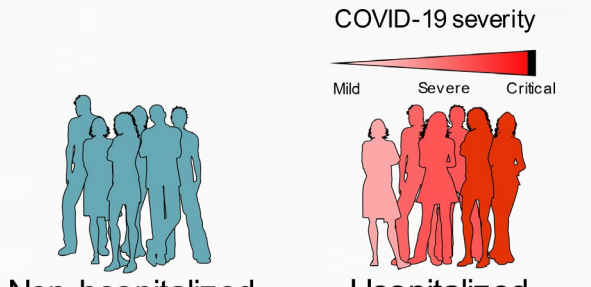


Formulate the research question!
Define the problem!

Research question dictates the data analysis

Outcome 1

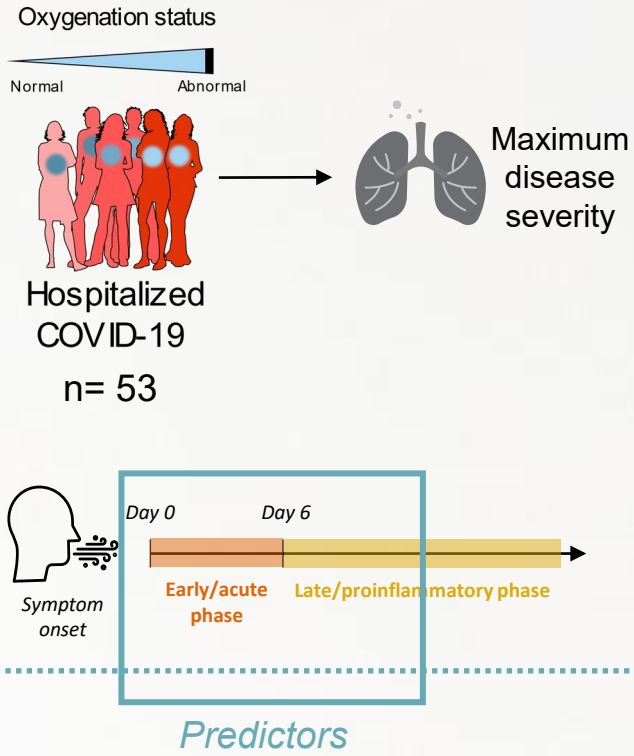
Difference between Sars-CoV-2-infected individuals?



Predictors

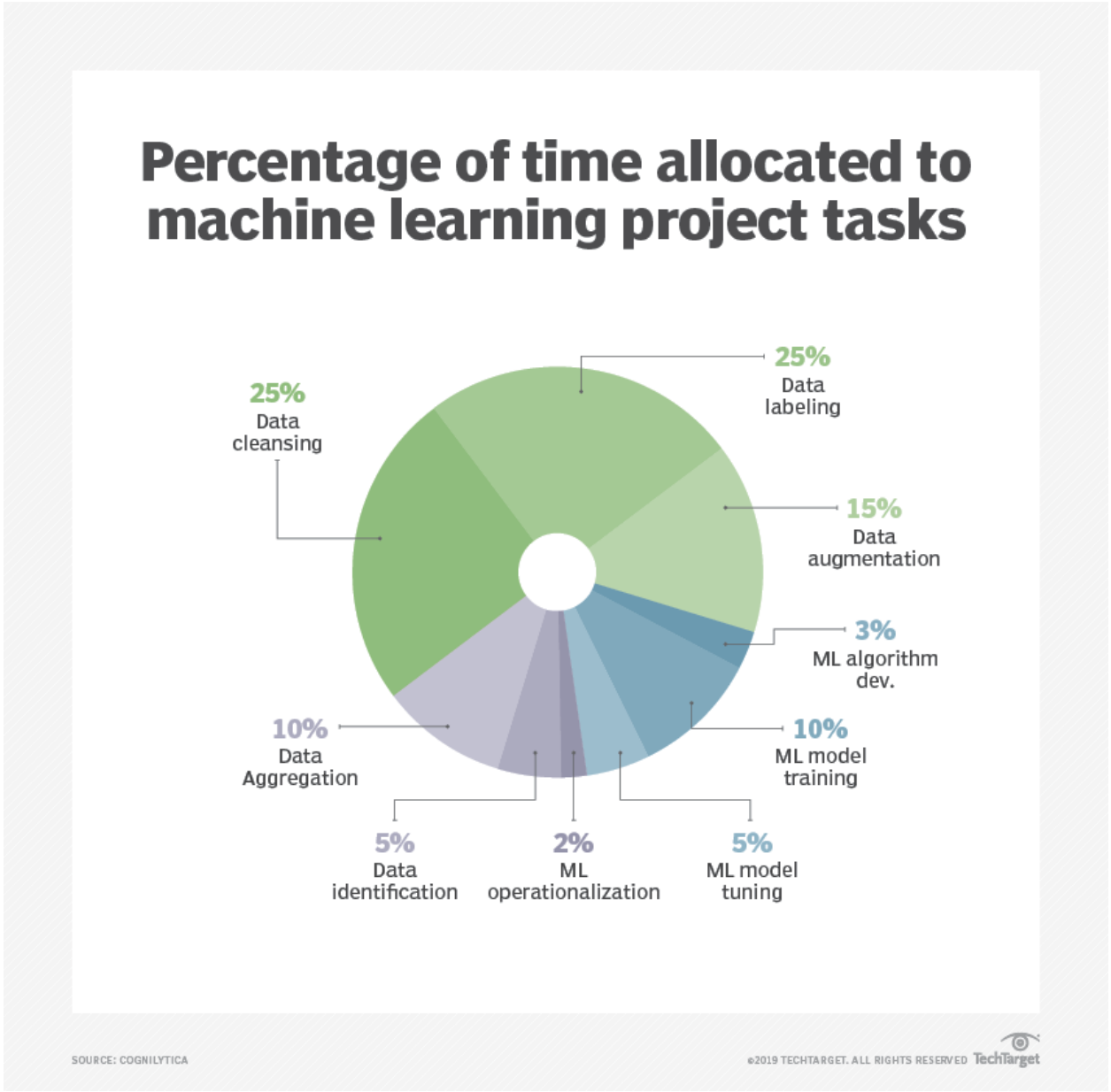
Outcome 2

Difference between hospitalized Sars-CoV2-infected patients?

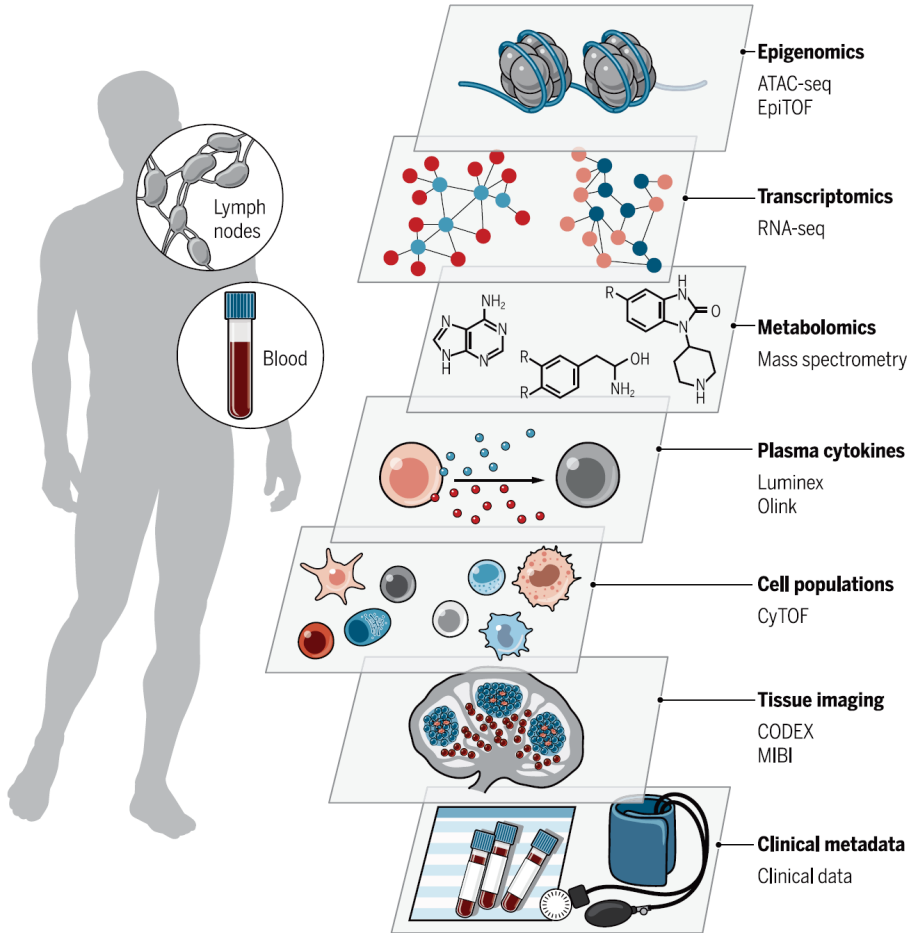


Predictors

Data preparation: the most important part of ML



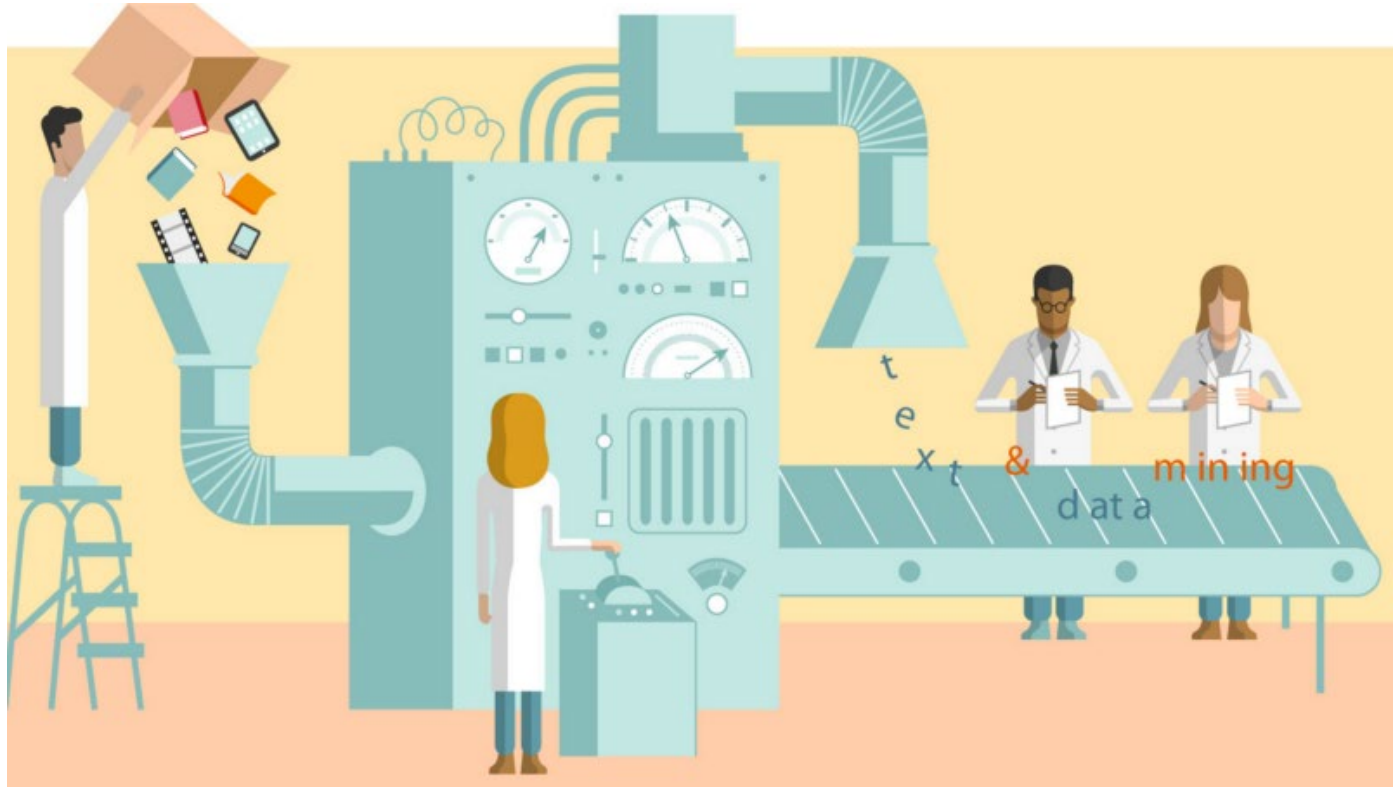
80%



Pulendran B and Davis M. Science, 2020

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

Encoding categorical data (gender, disease, etc.)

Data Before Encoding	Data After Encoding
Age Group (< 18)	1
Age Group (18-25)	2
Age Group (26-50)	3
Age Group (> 50)	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

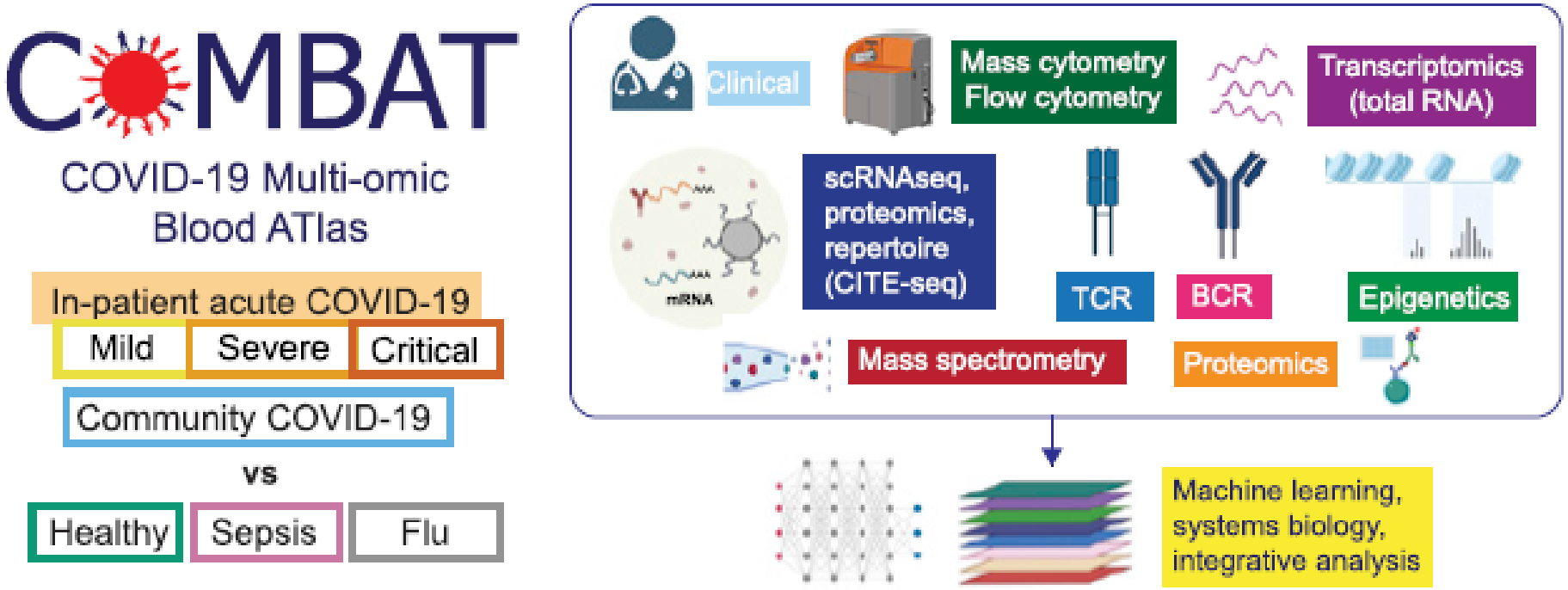
Feature engineering

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

https://medium.datadriveninvestor.com/data-preprocessing-3cd01eefd438

Data preparation process – COMBAT dataset example



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)

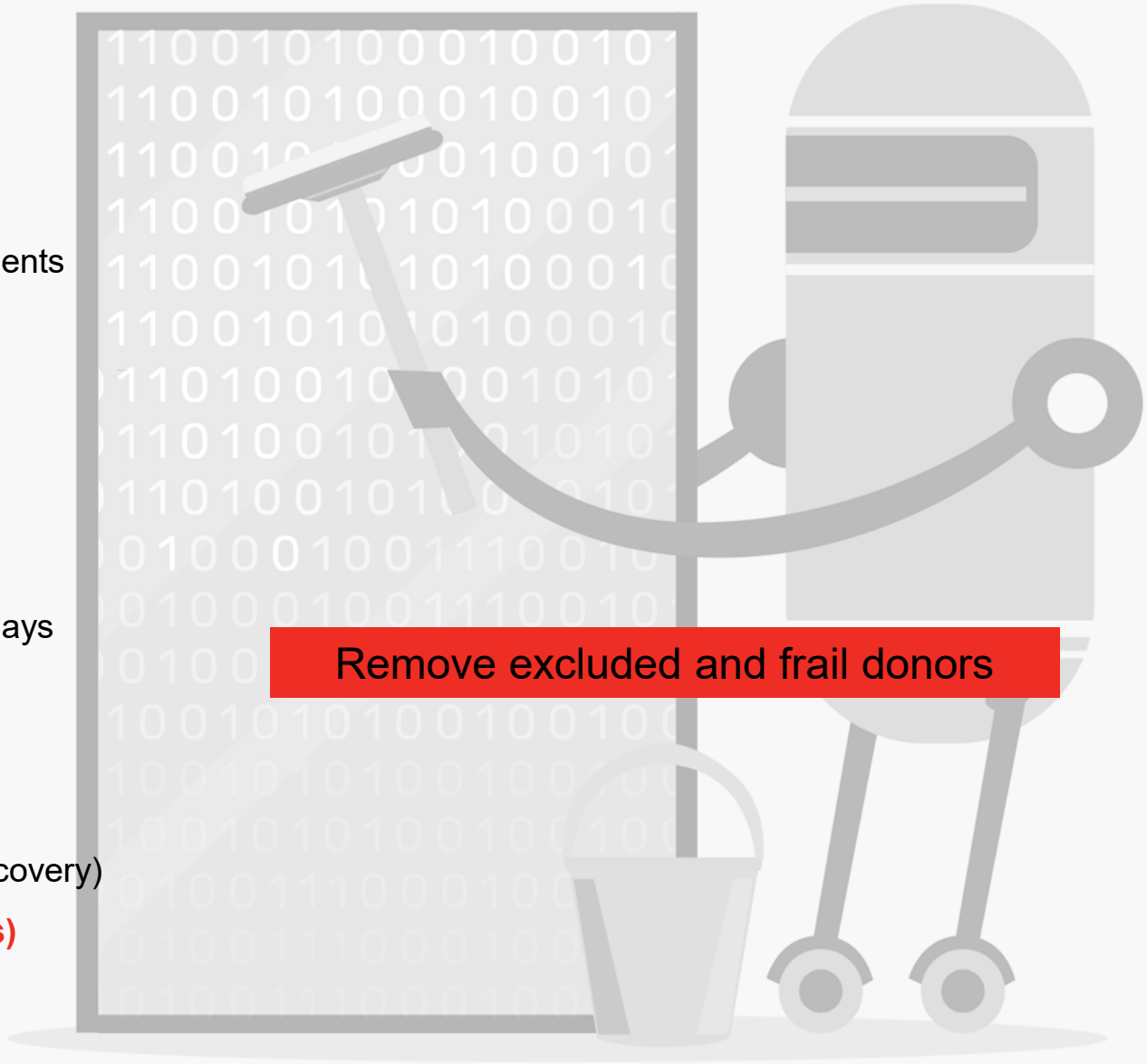
  Disease progress – for longitudinal samples (deterioration or recovery)

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

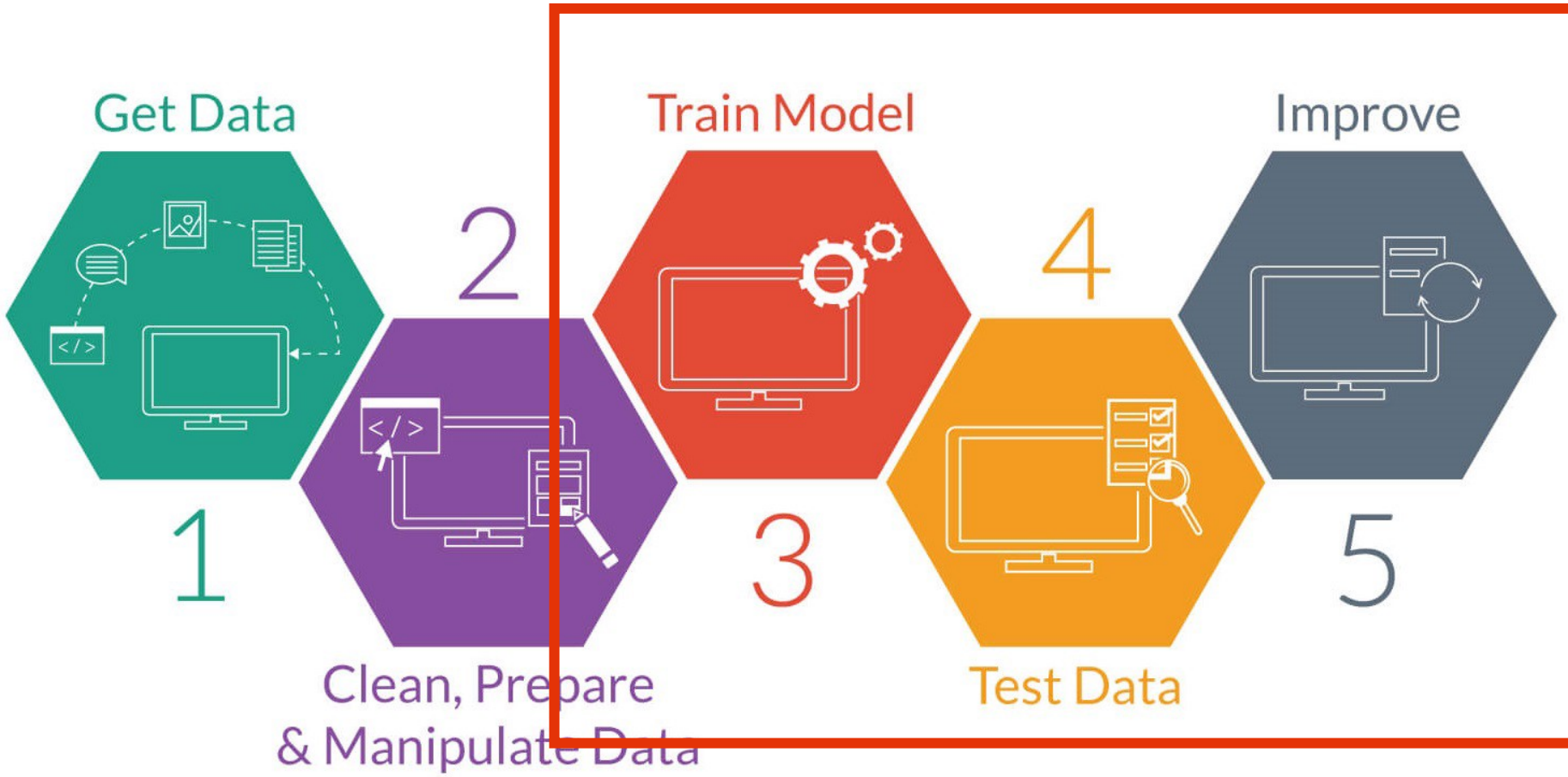
  sampling_from_max_disease

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

Written as factors (0 or 1)



Machine learning process

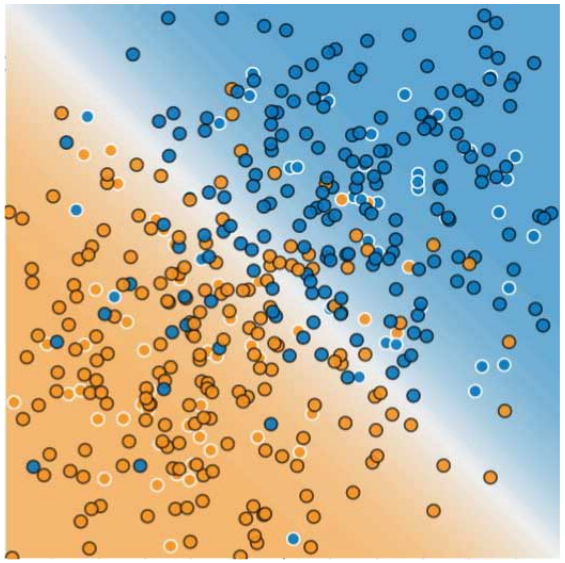


Data splitting: training and test sets

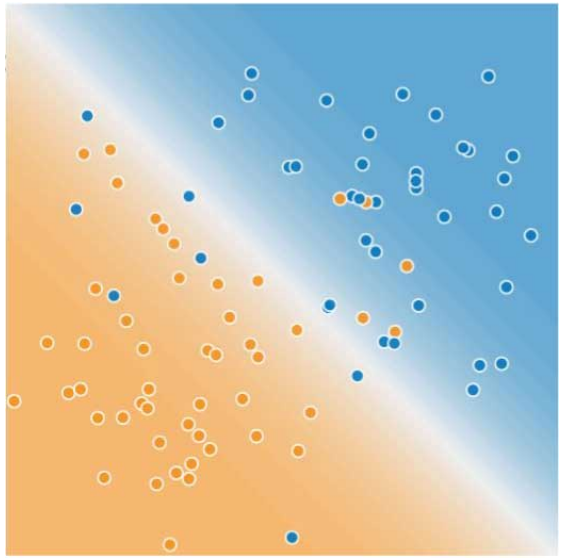
How to make a split?

subset to train a model
→ large enough to train a model
(more data you have – better model)

subset to test the trained model
→ large enough for statistically meaningful results
→ representative of the dataset as a whole

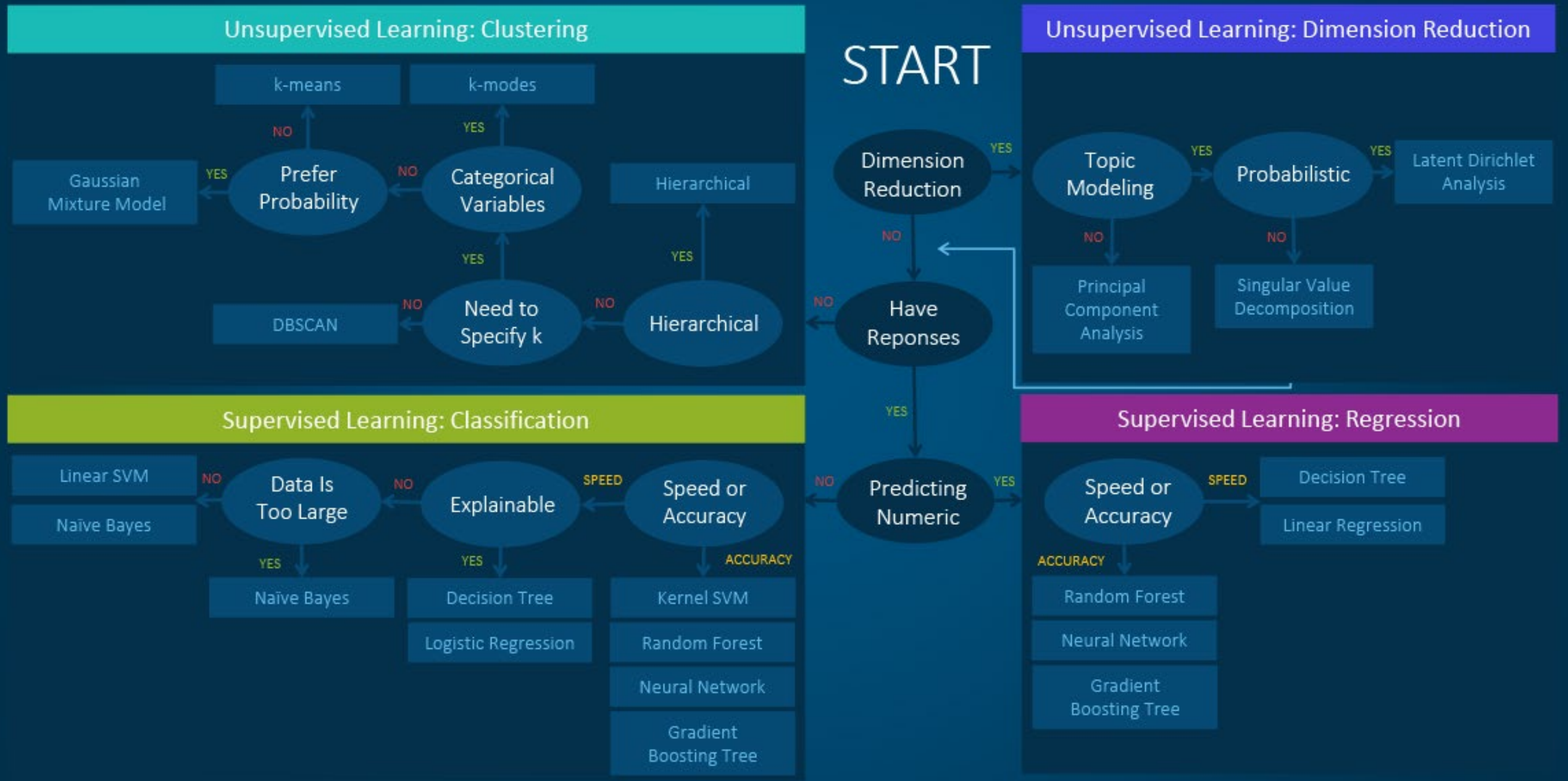


Training Data

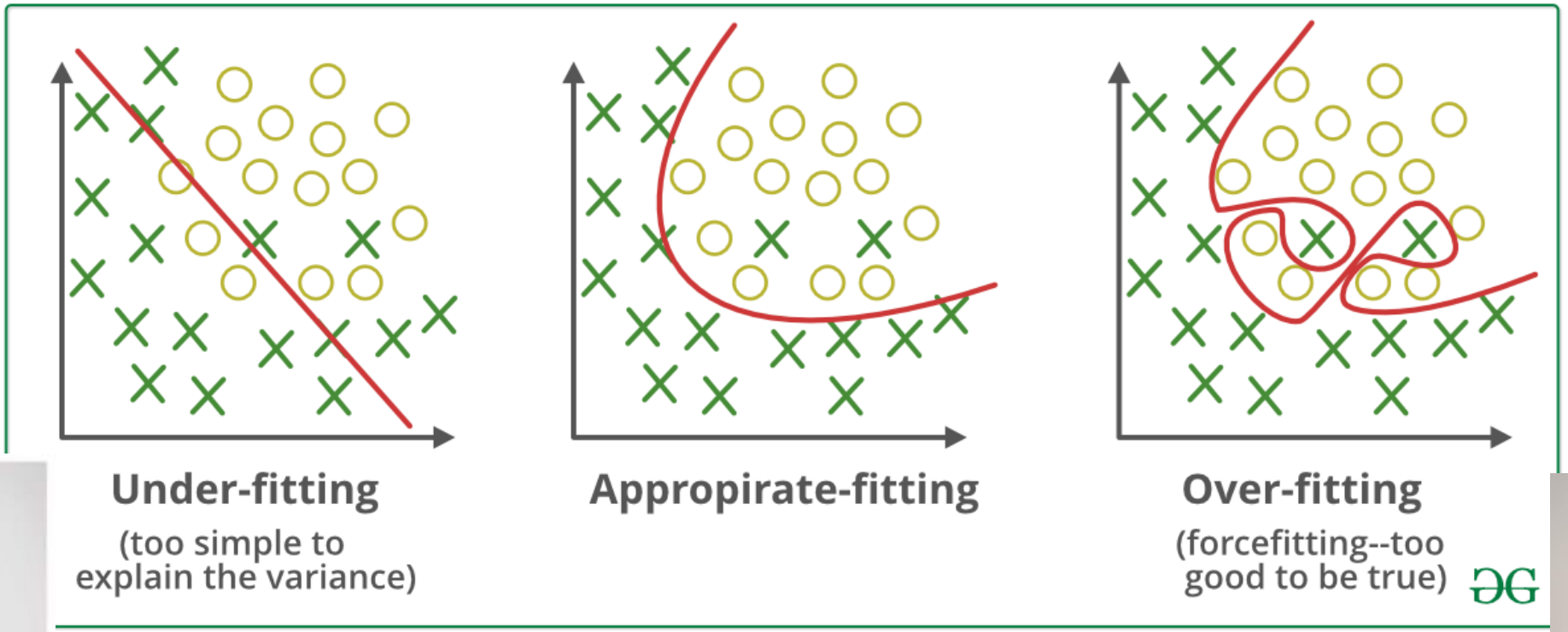


Test Data

Machine Learning Algorithms Cheat Sheet

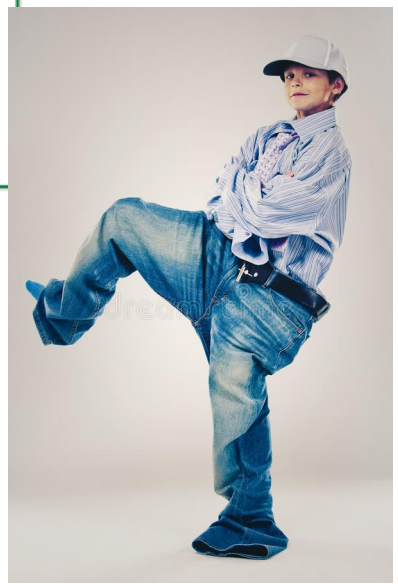


ML models - Underfitting and Overfitting

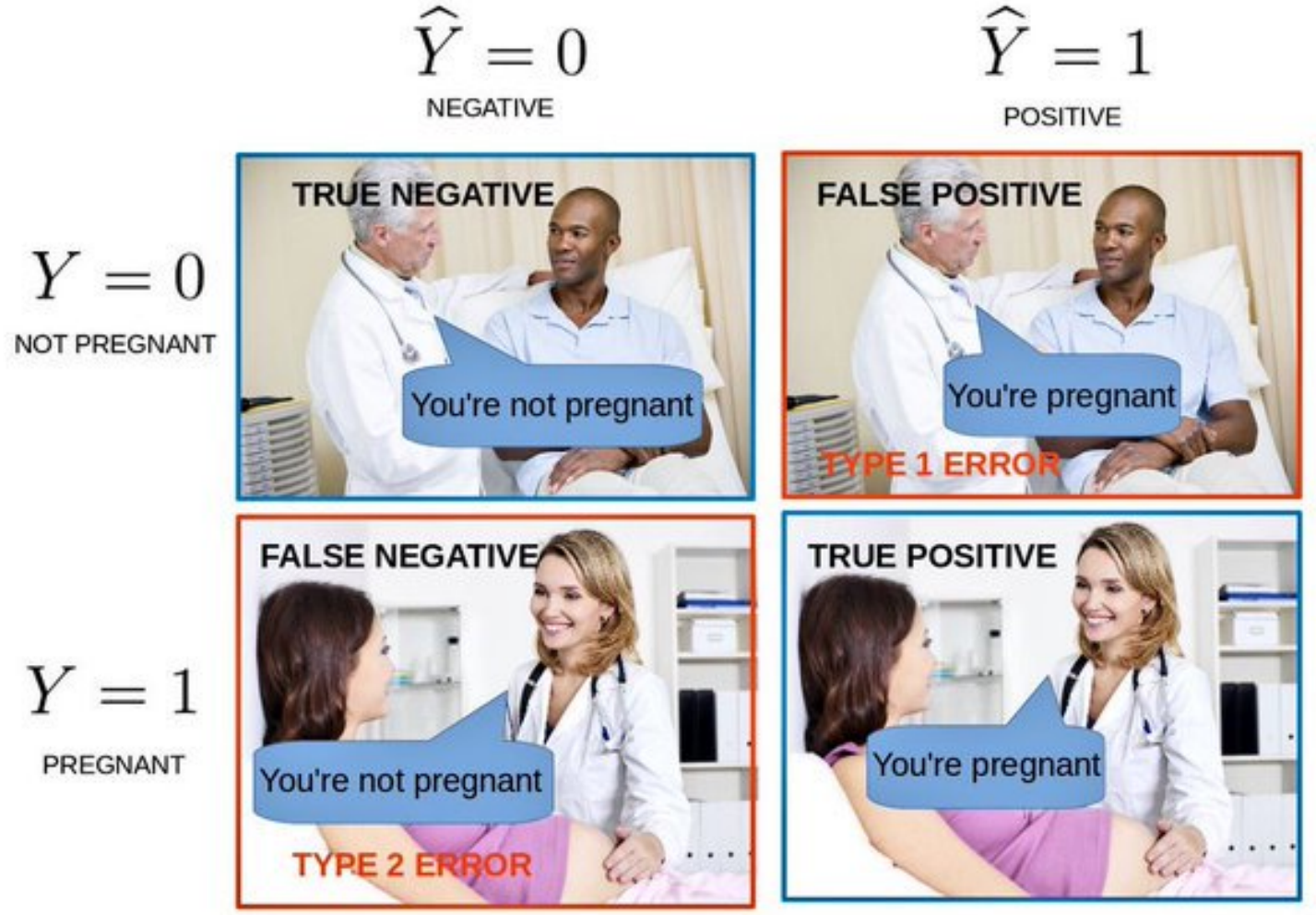


Not enough data!

Too much data!



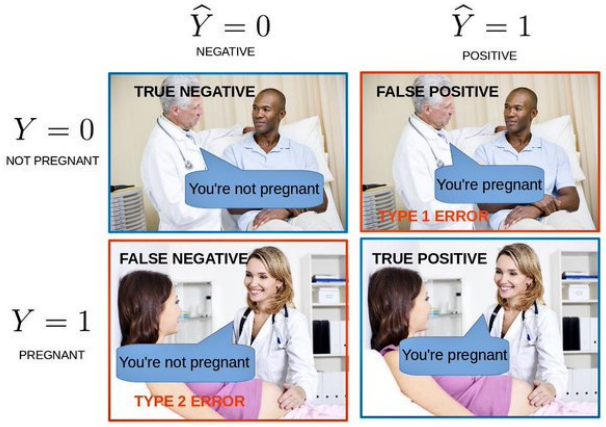
Evaluation of the machine learning algorithms performance – **confusion matrix**



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 + false positive)

FALSE DISCOVERY RATE

True negative/ (true + false negative)

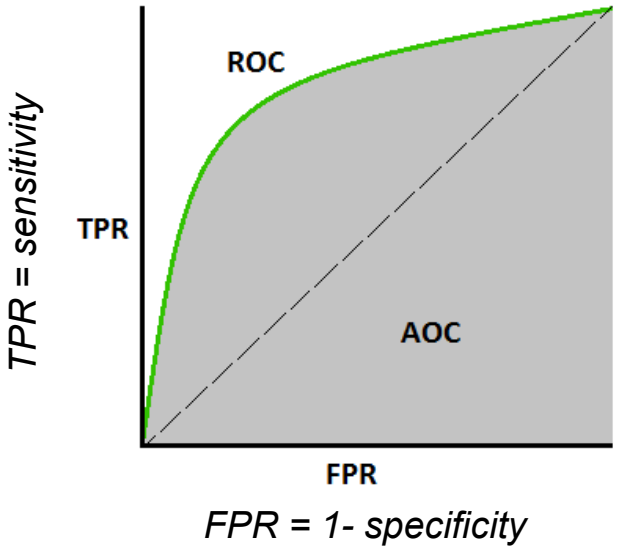
False positive/ Positive predictive value

ACCURACY= How often the classifier is correct? True positive + true negative/ sum of all

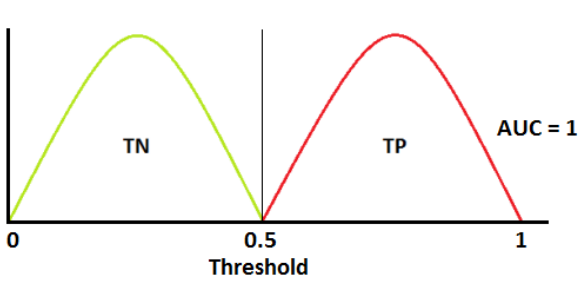
True positive/ (true positive + false negative)

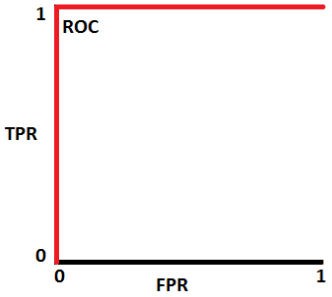
True negative/ (true negative + false positive)

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

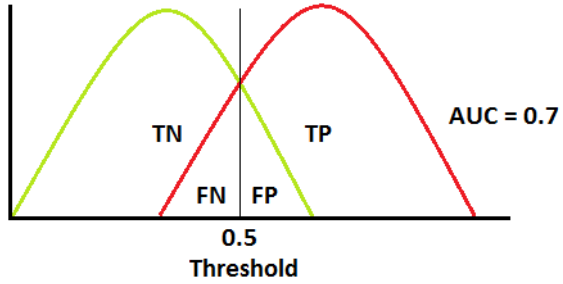


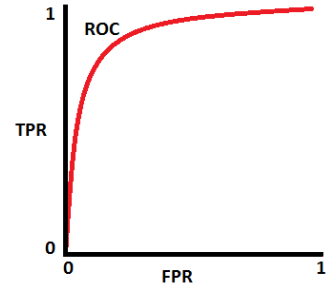
Perfect model: AUROC = 1



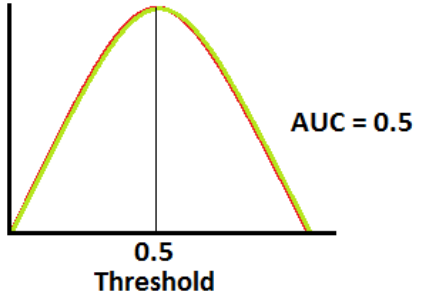


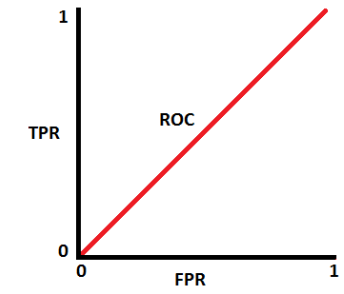
Good model: AUROC = 0.70





Random model: AUROC = 0.50



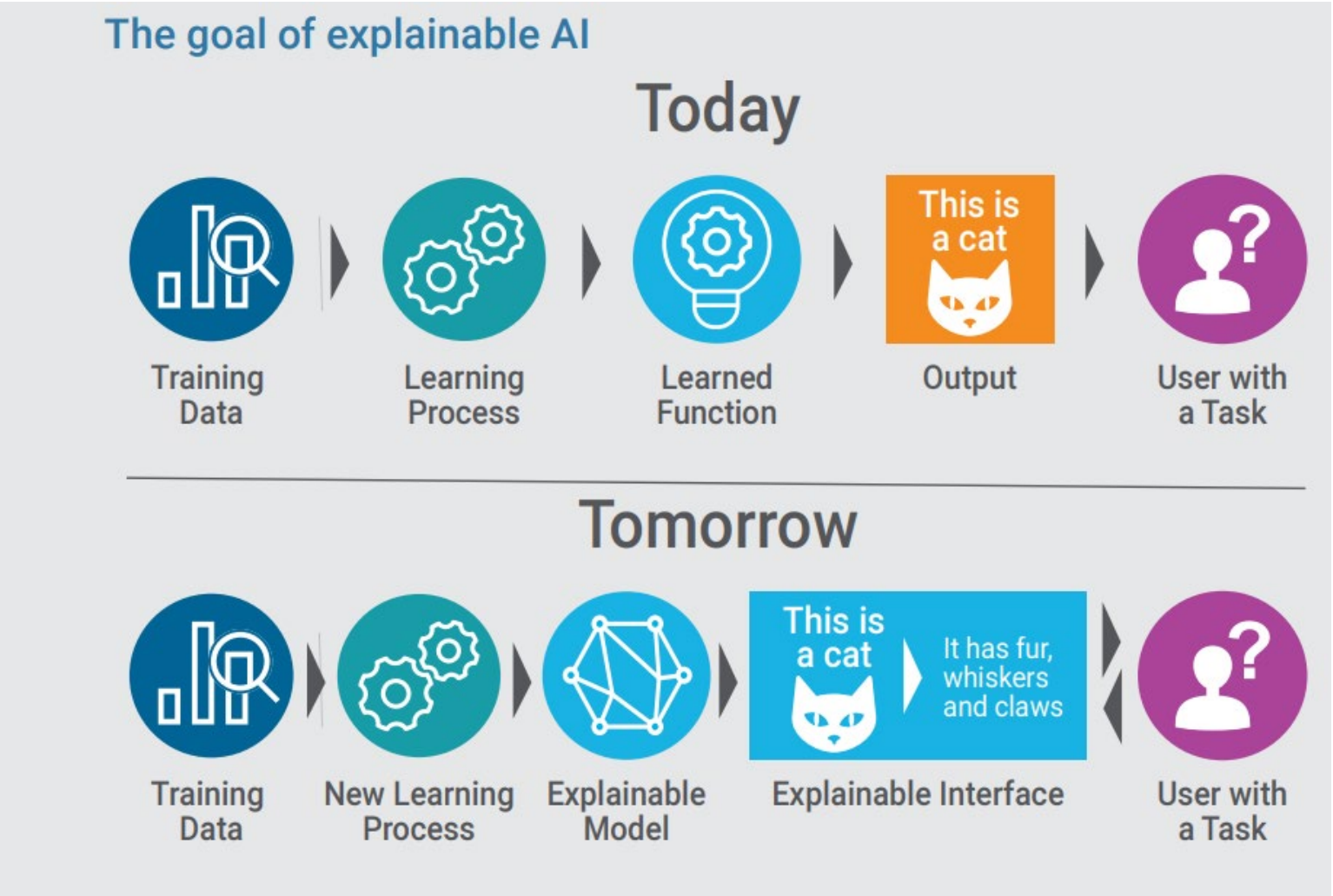


Note: Red distribution curve is of the positive class (patients with disease) and green distribution curve is of negative class(patients with no disease).

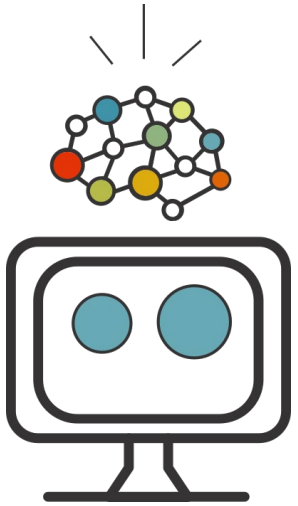
*AUROC (Area Under the Receiver Operating Characteristics)

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

Understanding the model

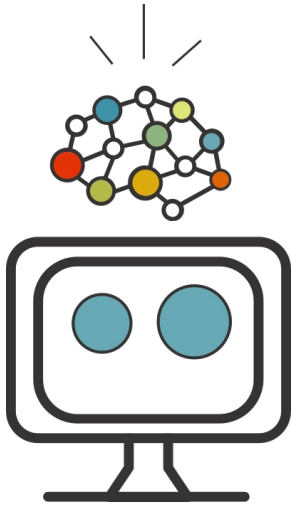


Part III. SIMON, pattern recognition and knowledge extraction software

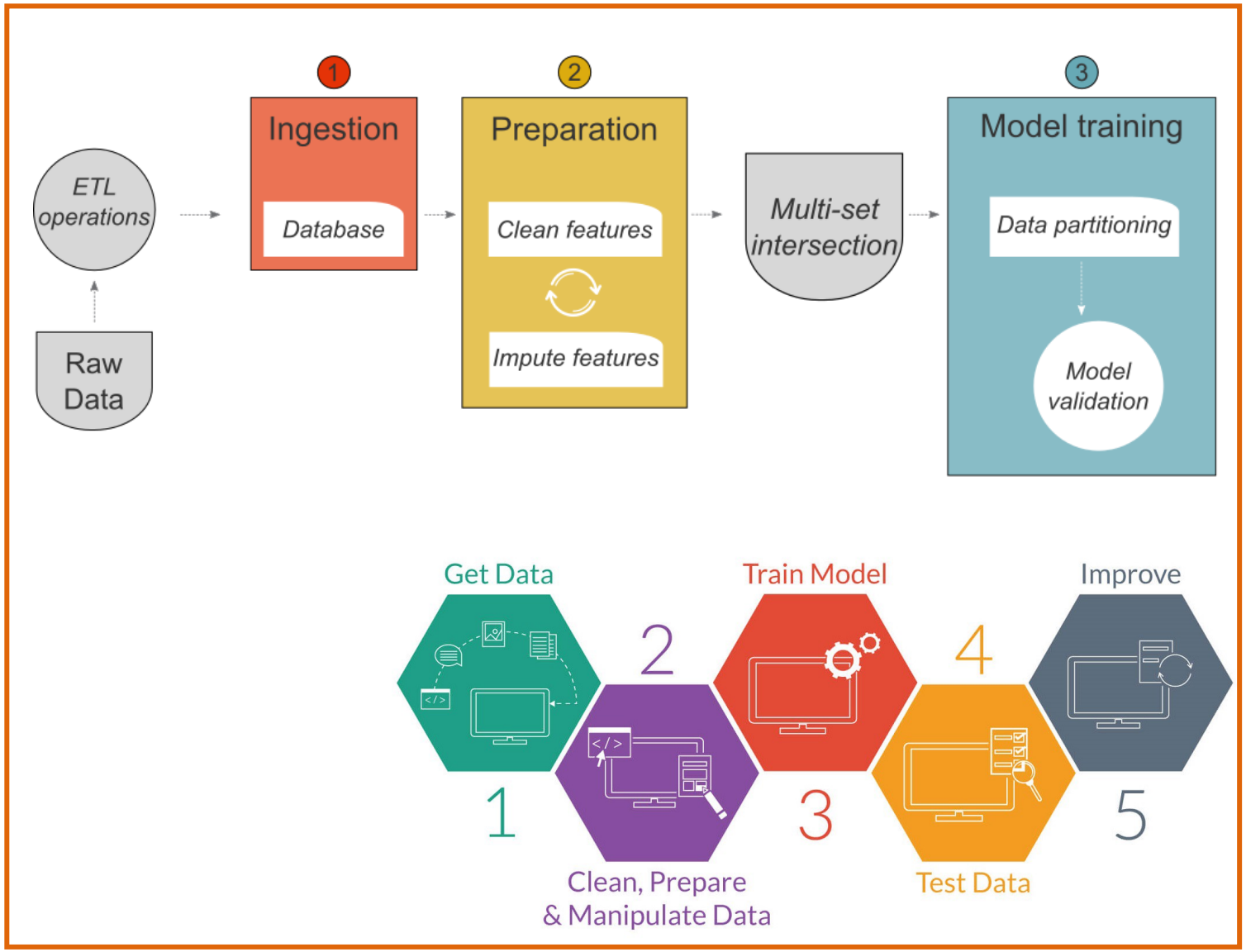



KUKA
KNEXT
KNEXT barista robot
Perfect cofee every time

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!)

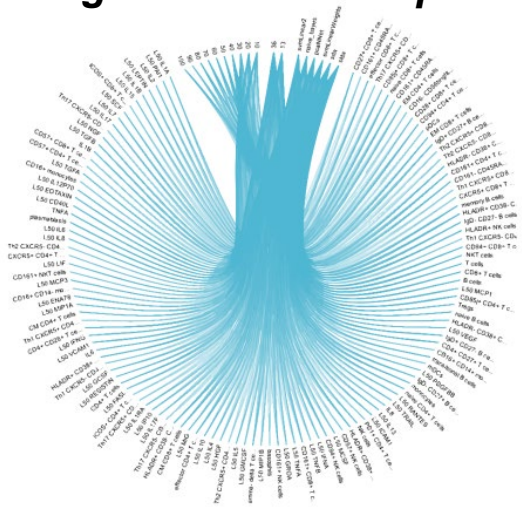
Academia-released software

- 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

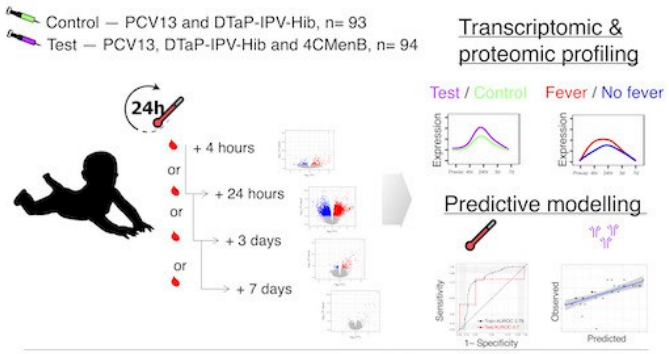
- Free and open source – explained/published ML methods and algorithms
- Requires knowledge of ML process
- Lack some of the advance features of commercial software (autoML)

Integrative analysis of different data types – predicting flu vaccine responses

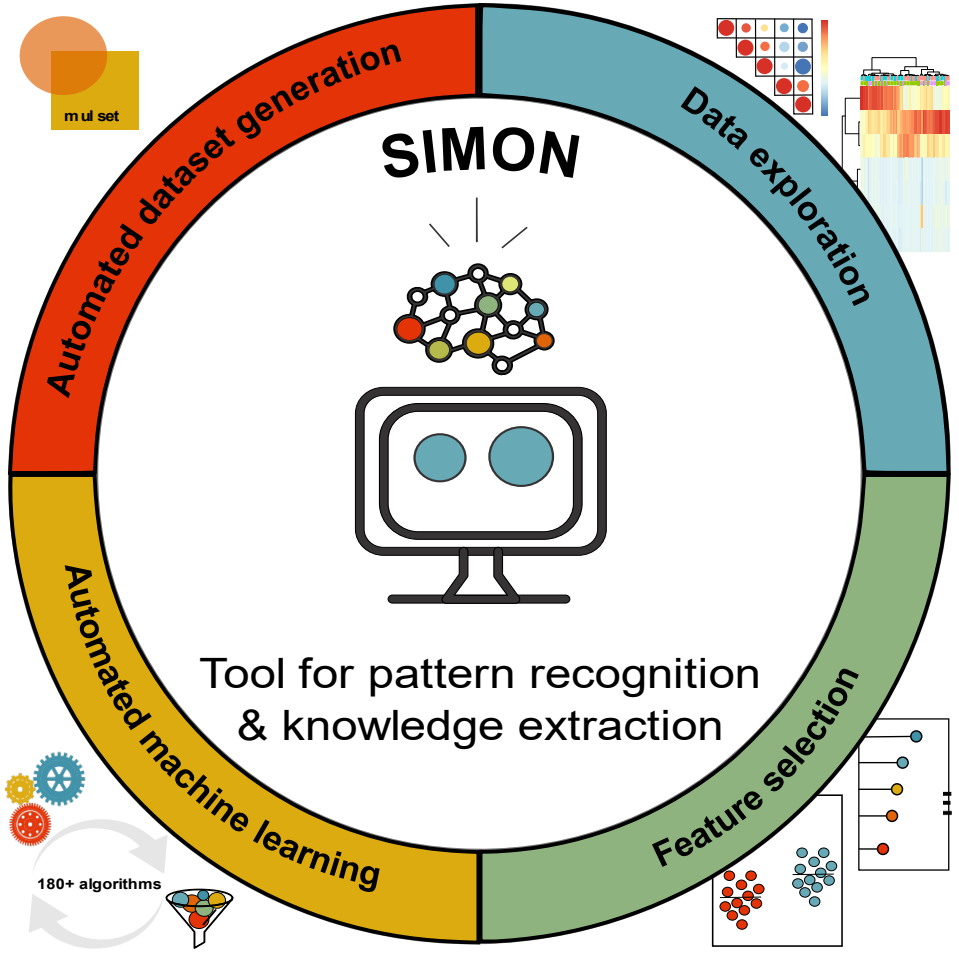


Tomic et al, JI, 2019

Transcriptome data

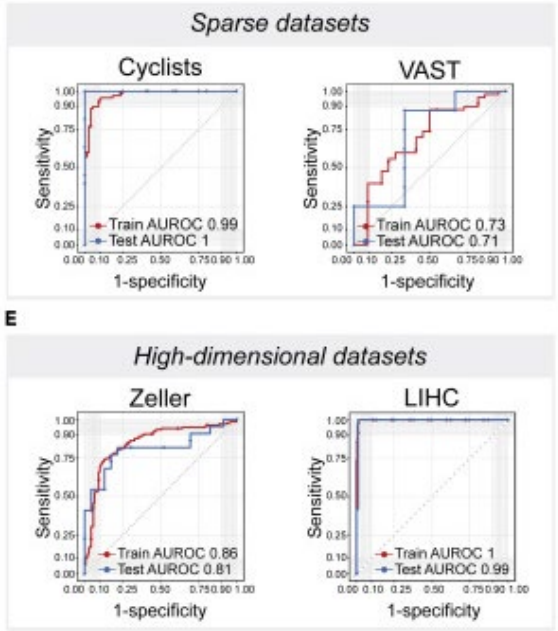


O'Connor et al, Mol Syst Biol, 2020

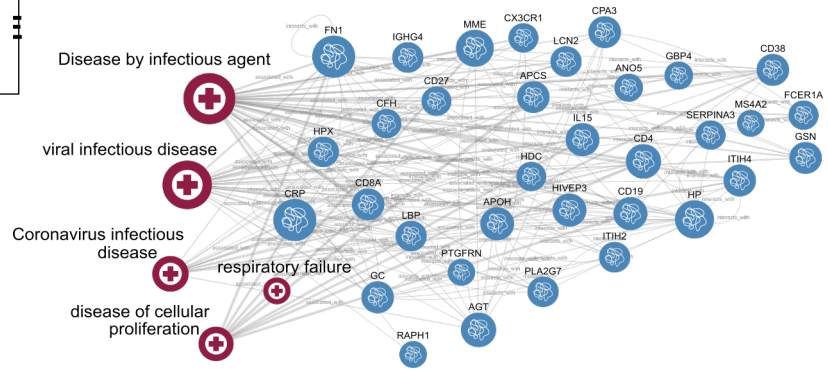


Tomic et al, Patterns, 2021

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

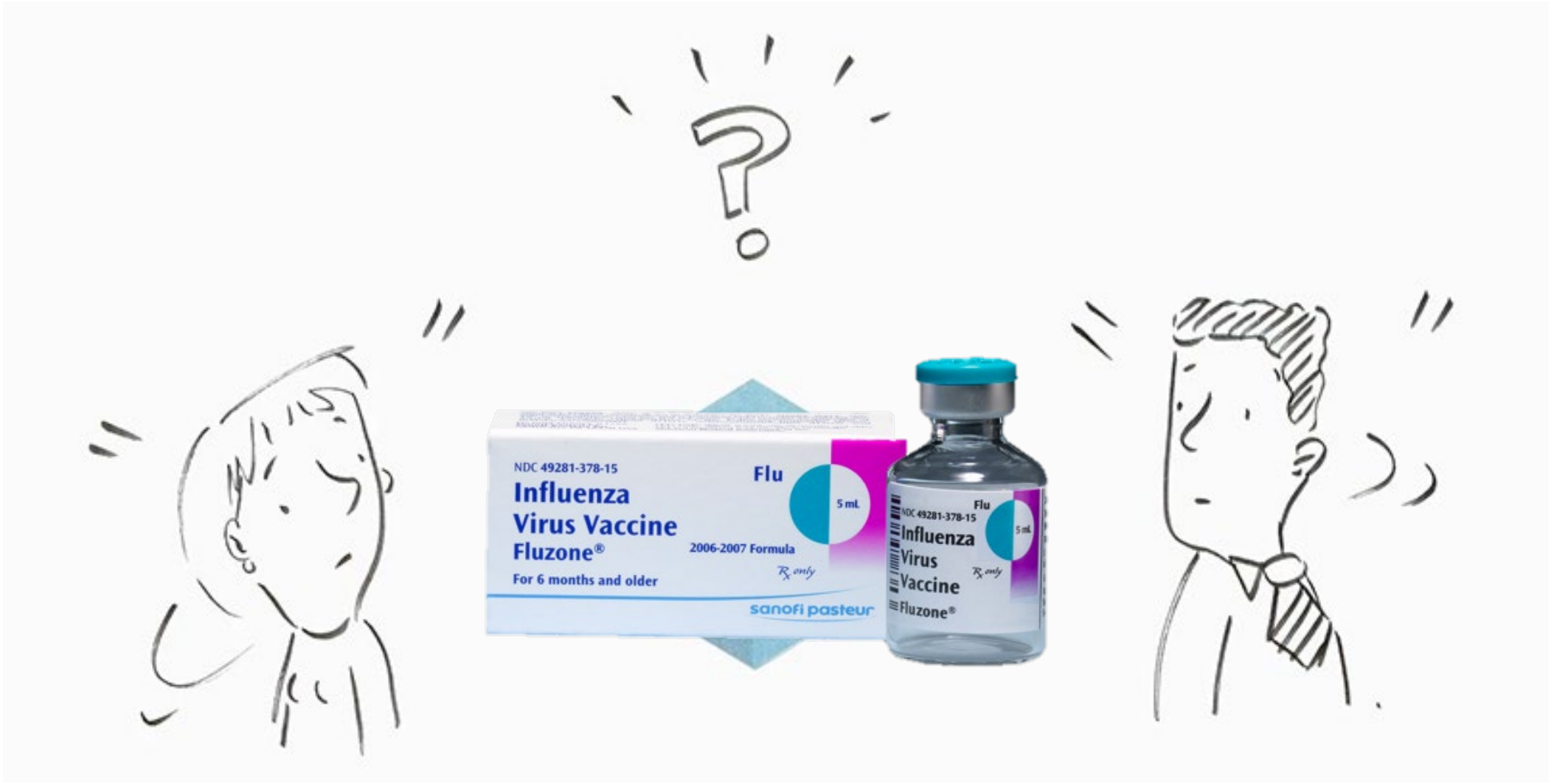


Multi-omics integrative analysis – COVID-19 COMBAT project



COMBAT consortium, Cell, 2022

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

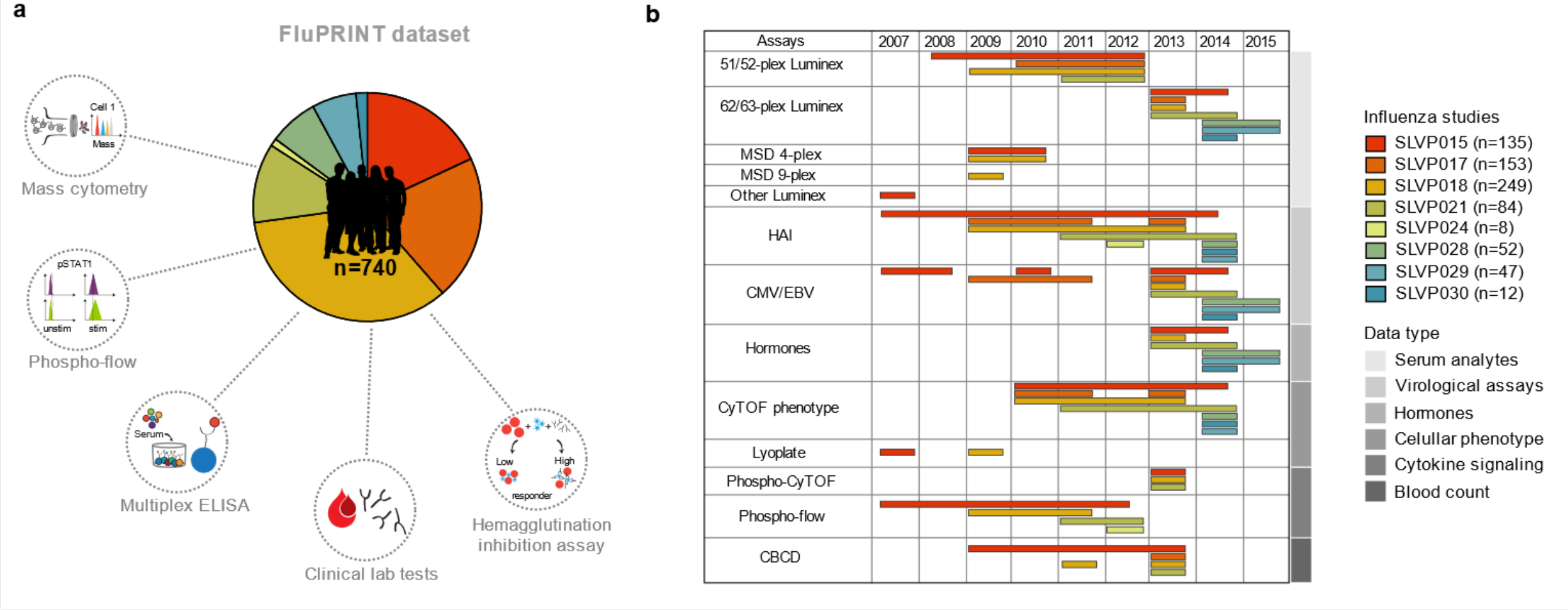


Hidden treasures at Stanford

Human Immune Monitoring Center (HIMC) – Stanford Data Miner

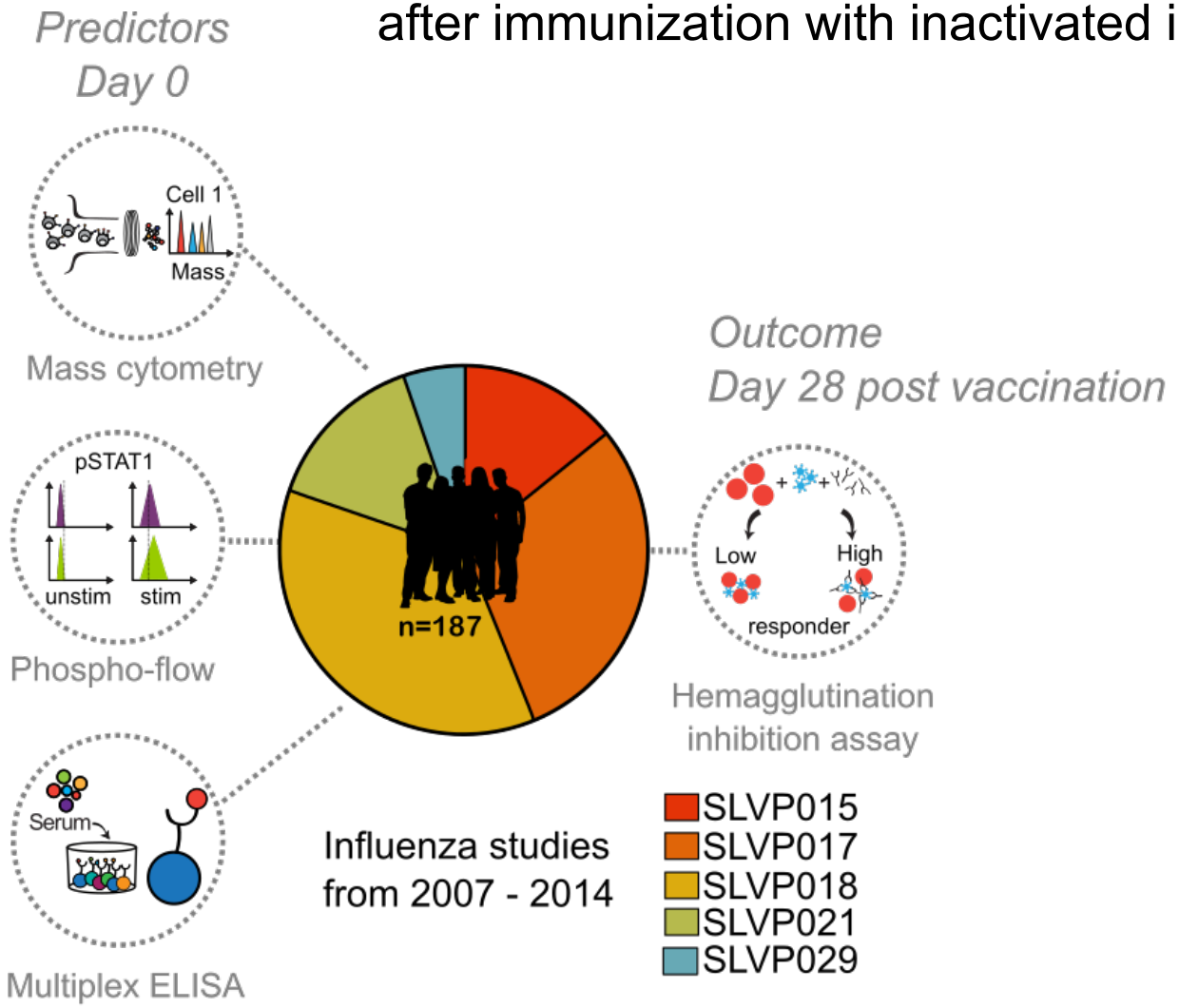
- 2007 – 2015
- 8 Flu clinical studies
- >700 unique donors – 0 to 90y
- >3000 parameters: Flow cytometry, CyTOF, PhosphoFlow, Multiplex ELISA, HAI, blood test, …

FluPRINT database – 740 individuals

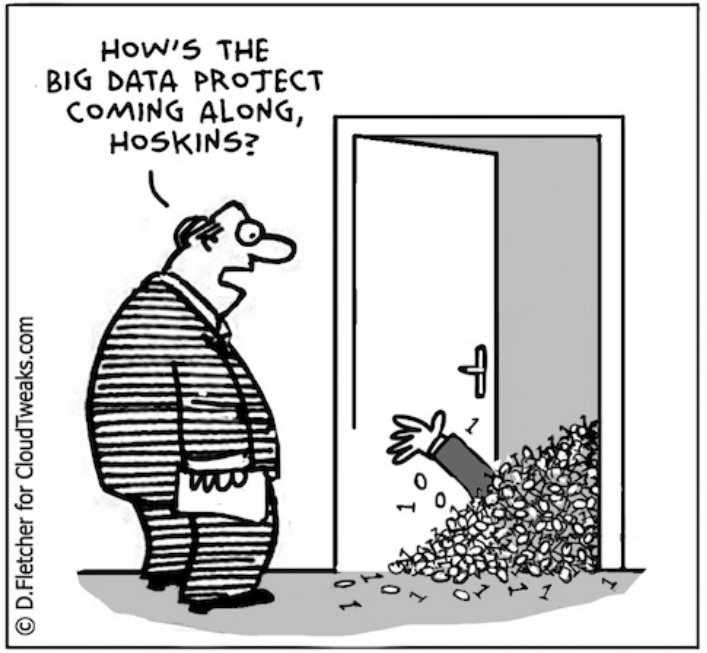


Tomic et al. The FluPRINT dataset, a multidimensional analysis of the influenza vaccine imprint on the immune system. Sci Data, 2019

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

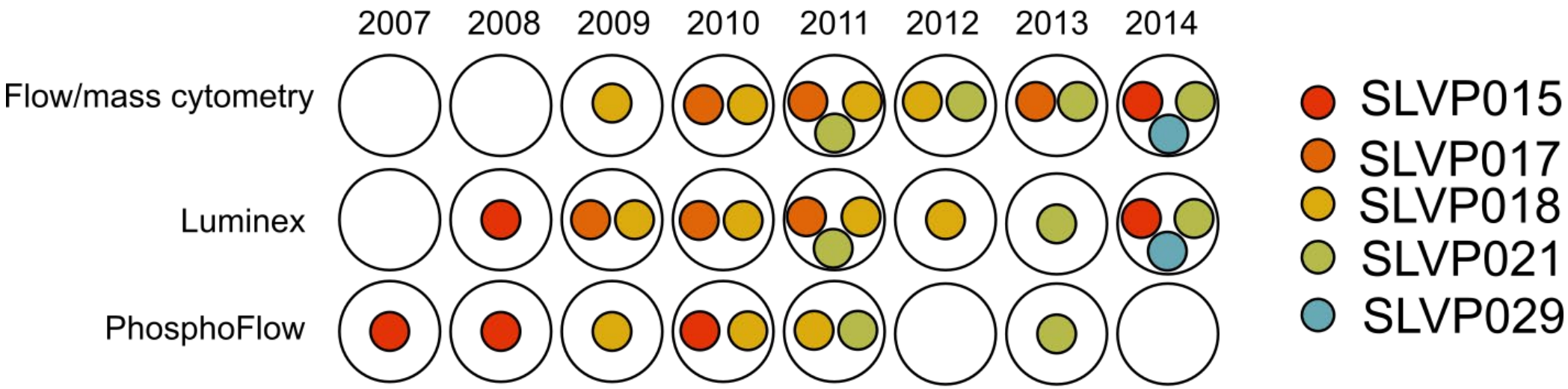


Stanford Human Immune Monitoring Center



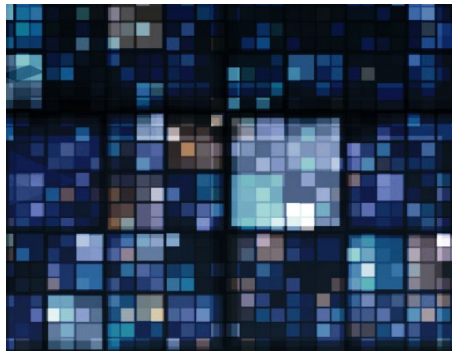
Tomic et al. SIMON, an Automated Machine Learning System, Reveals Immune Signatures of Influenza Vaccine Responses, JI, 2019

Dealing with missing values

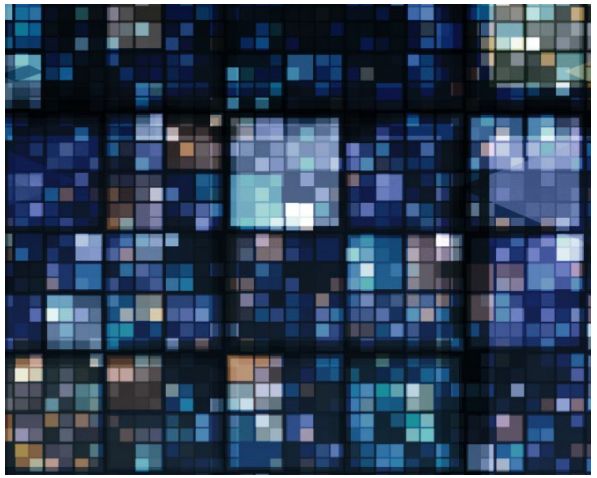


The “BIG” problem: Highly percentage of missing data

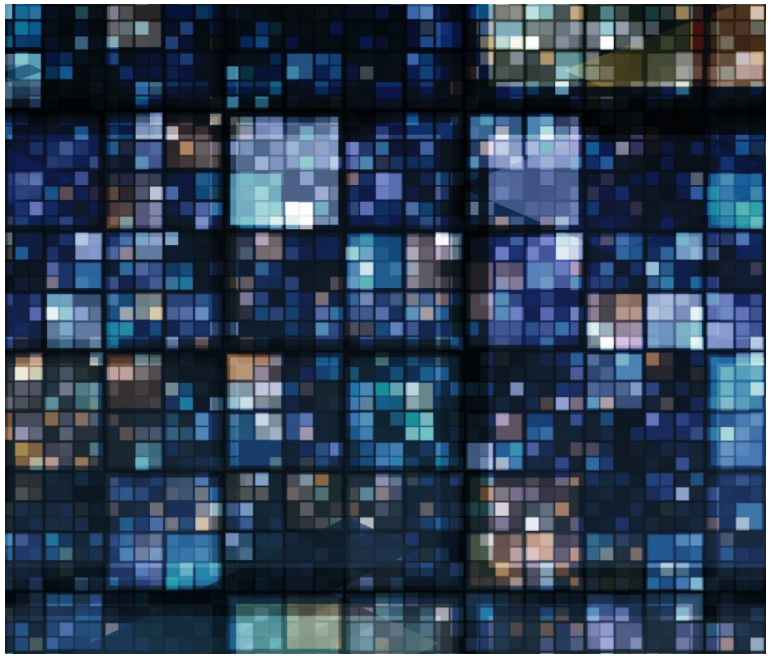
2007



2010



2014

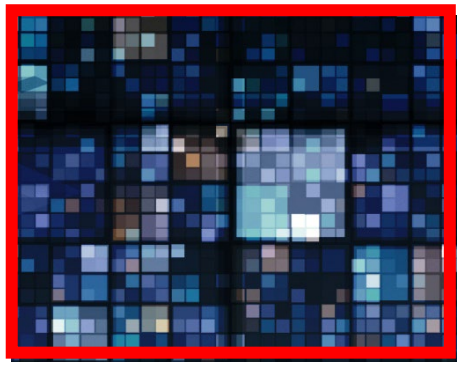


How to select optimal number of donors and optimal number of features?

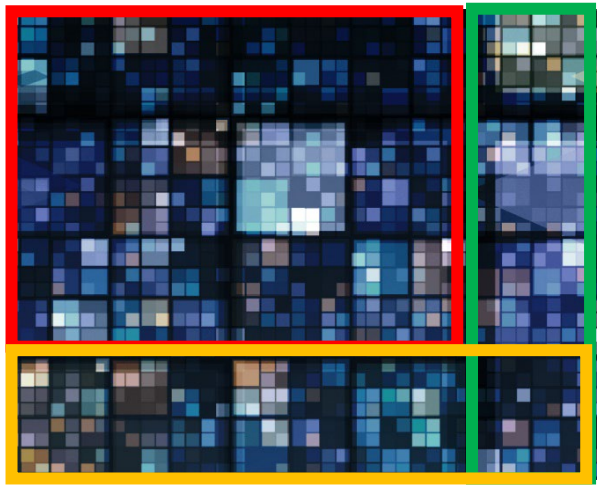
SUBSAMPLING

The alternative solution to cope with high sparsity

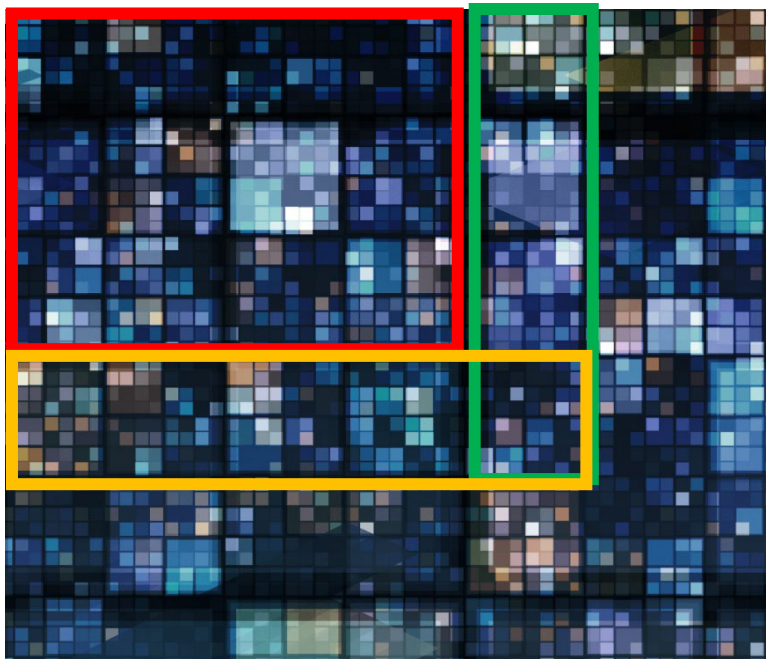
2007



2010

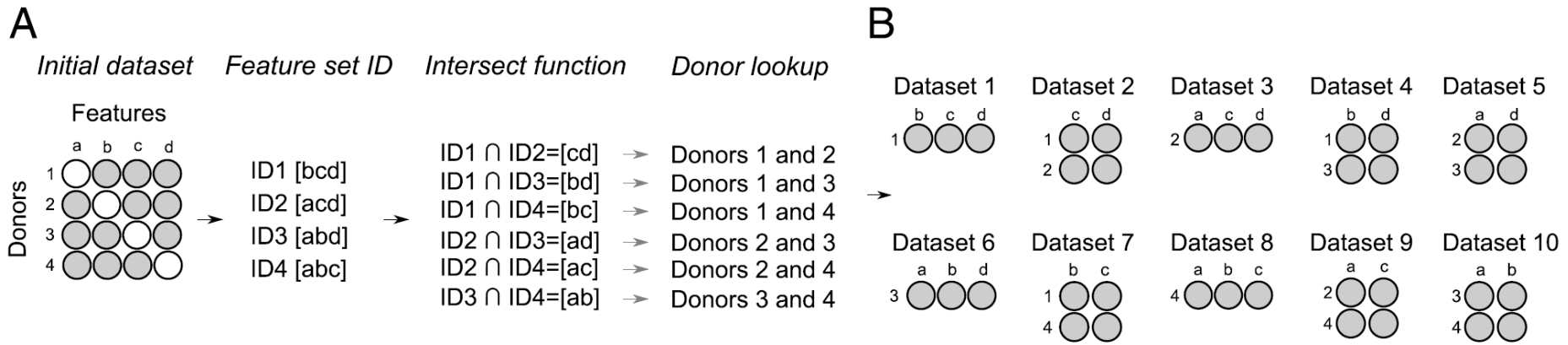


2014

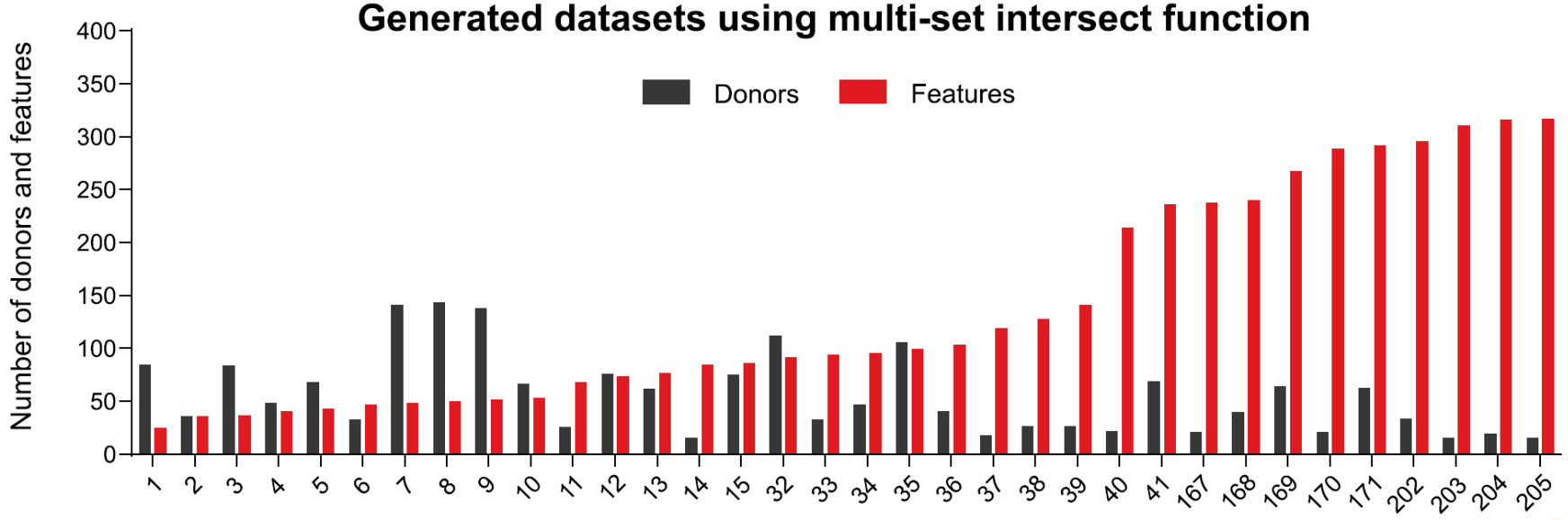


A fully automated script for feature subset selection, dimensionality reduction and data sampling

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



C

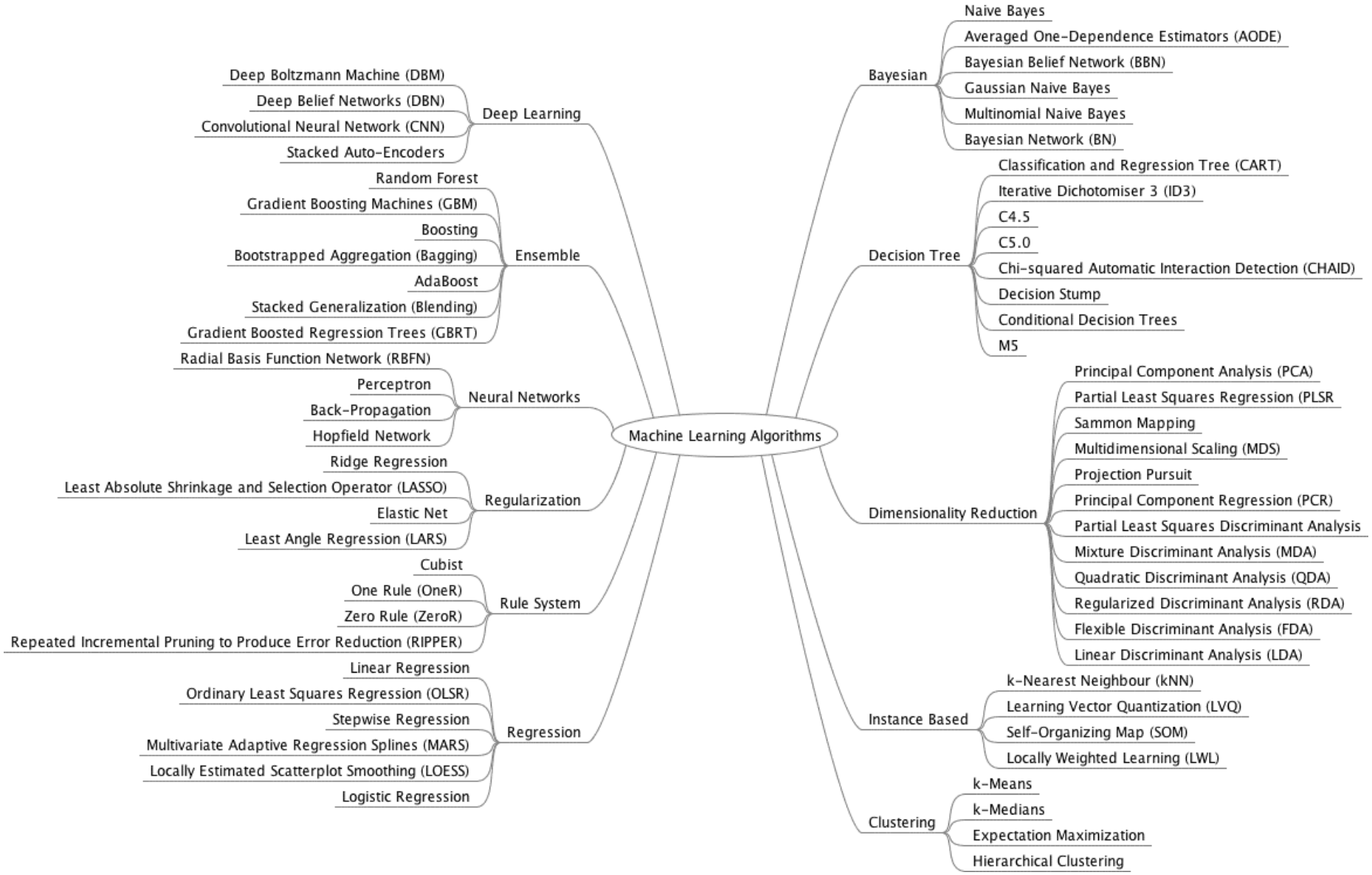


Tomic et al. SIMON, an Automated Machine Learning System, Reveals Immune Signatures of Influenza Vaccine Responses, JI, 2019

mulset

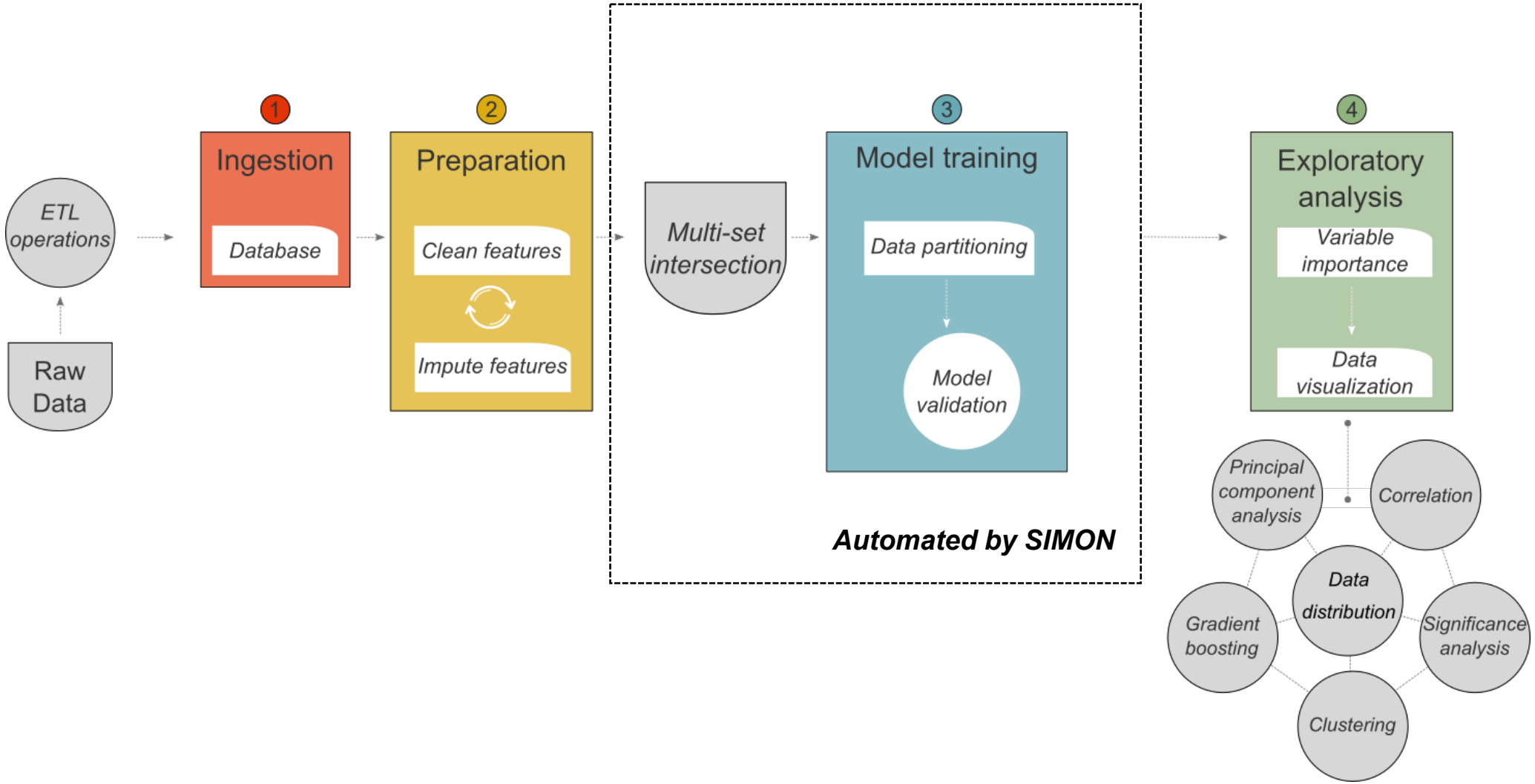
https://github.com/LogIN-/mulset

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



Jason Brownlee, machinelearningmastery.com

SIMON: **S**equential **I**terative **M**odelling **O**ver**N**ight



Tomic et al. SIMON, an Automated Machine Learning System, Reveals Immune Signatures of Influenza Vaccine Responses, JI, 2019

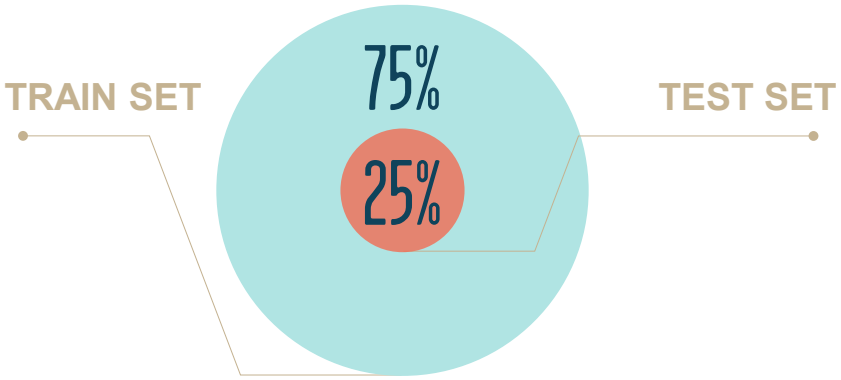
SIMON: **S**equential **I**terative **M**odelling **O**ver**N**ight

180+ machine learning algorithms

- Random Forest
- Support Vector Machine
- Discriminant Analysis
- ...

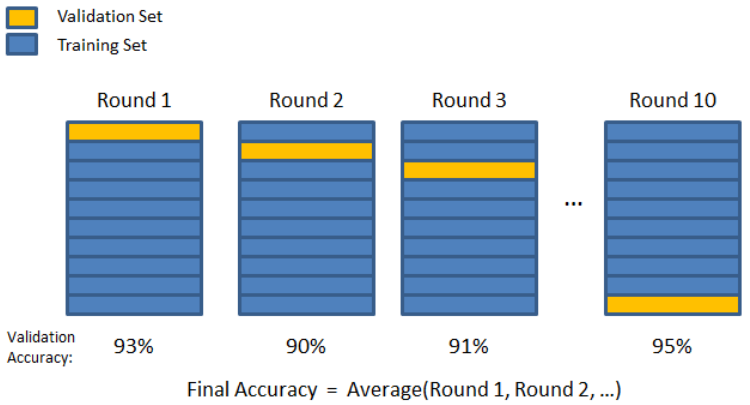
MODEL

Data partitioning



MODEL evaluation

10-fold cross-validation



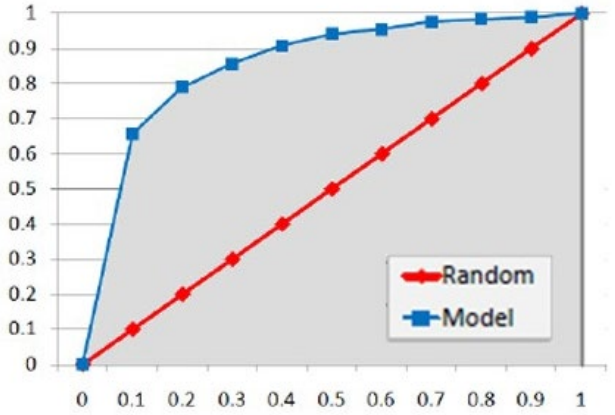
Final Accuracy = Average(Round 1, Round 2, ...)

Model selection

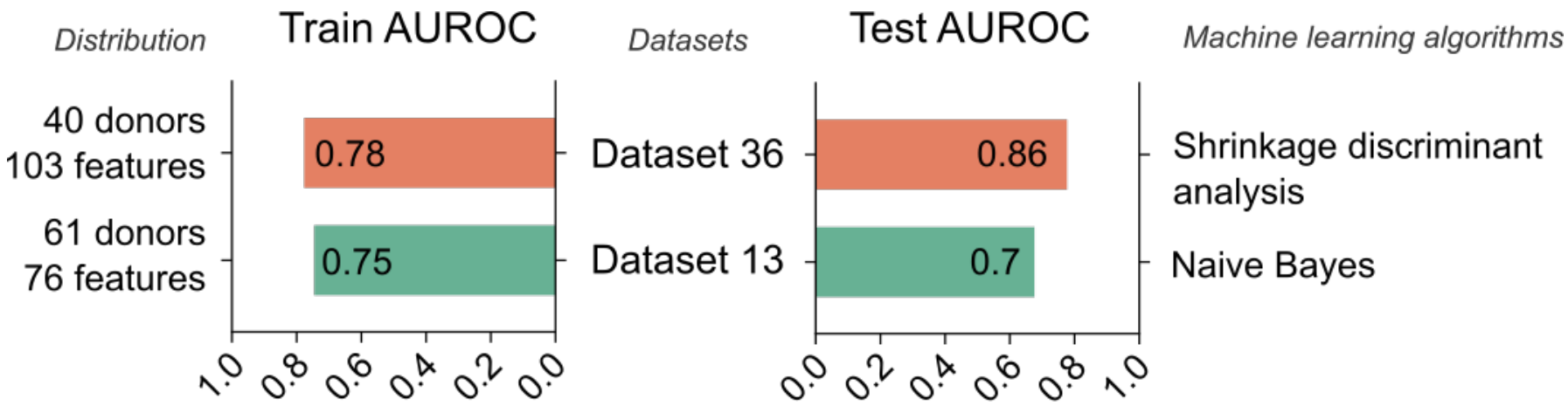
Confusion matrix

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

AUROC



SIMON results: 2 datasets with the highest accuracy



SDA
Works the best with the small sample size, but high-dimensional setting

Bankruptcy prediction

Image recognition

Marketing

Mkhadri A, Pattern Recognition Letter 1995

Text categorization

Medical diagnosis

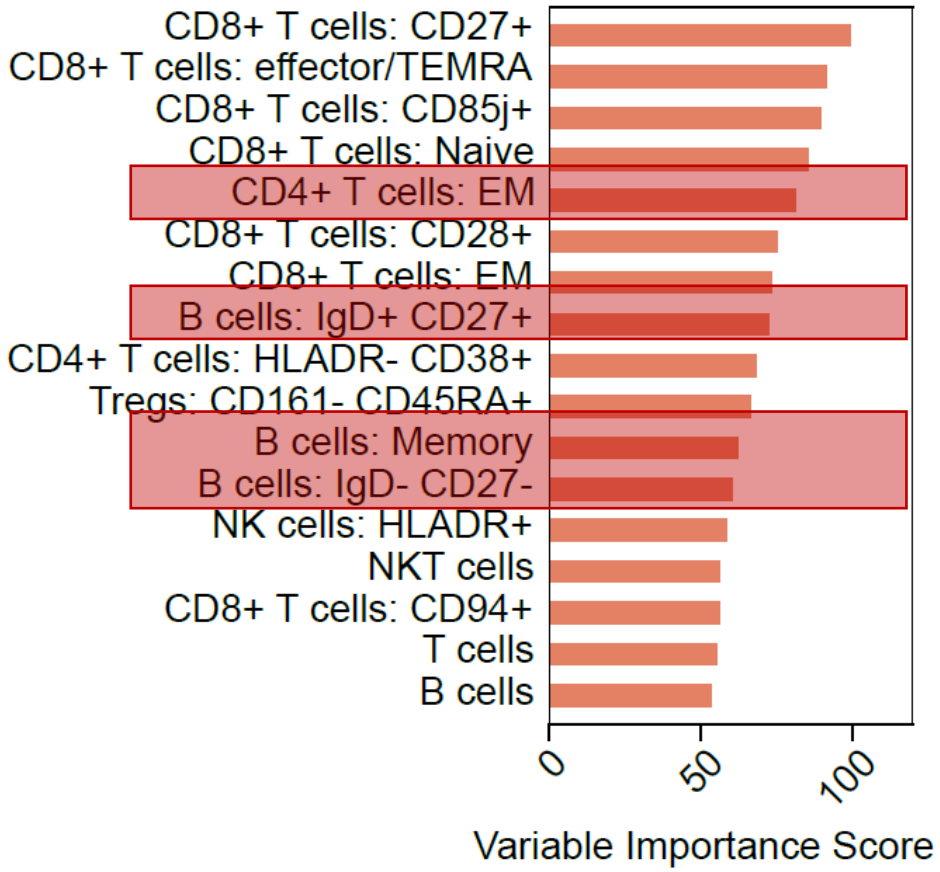
Naïve Bayes
Very simple but powerful. Known to outperform highly sophisticated classification methods

1960s

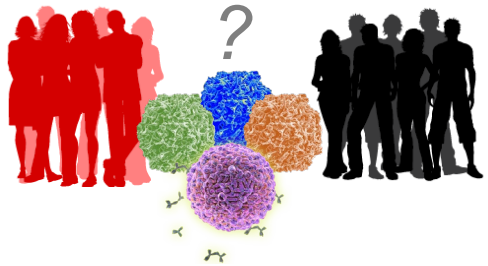
Train AUROC – mean value determined from confusion matrix after 10-fold cross-validation (repeated 3 times)
Test AUROC – evaluated from confusion matrix on independent test set

Pattern recognition in influenza vaccine study using SIMON

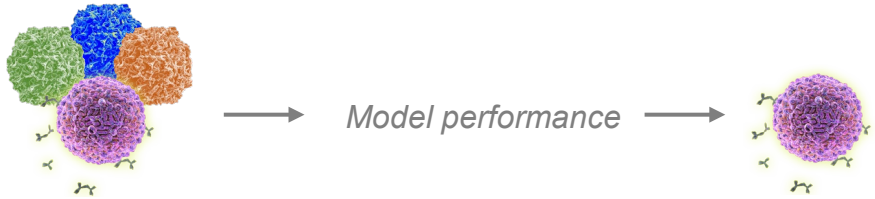
Shrinkage discriminant analysis;
Train AUC: 0.78; Test AUC: 0.86



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



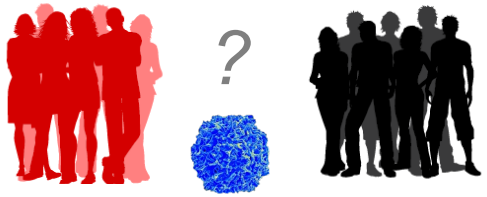
Variable selection



Tregs: CD25hi, CD127-

Hypothesis

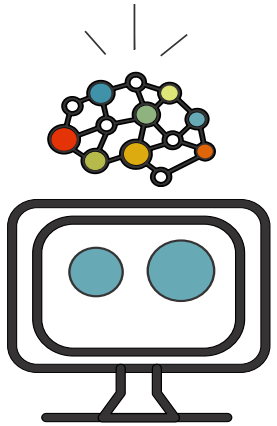
Why is frequency of T cells increased among healthy vs infected person?



T cells

Data-driven research

Data analysis

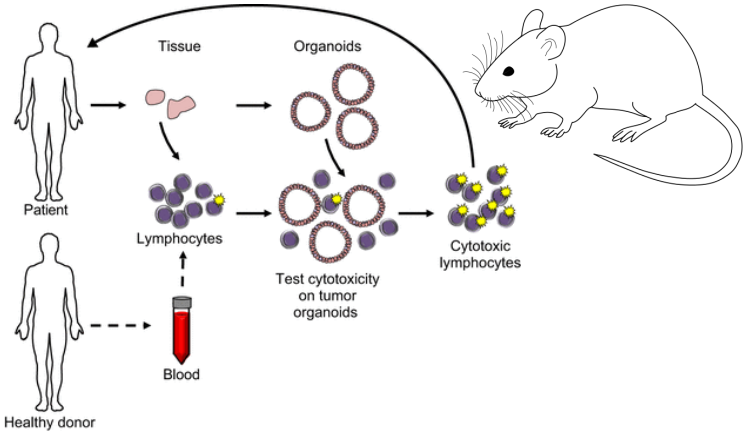


*SIMON**

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

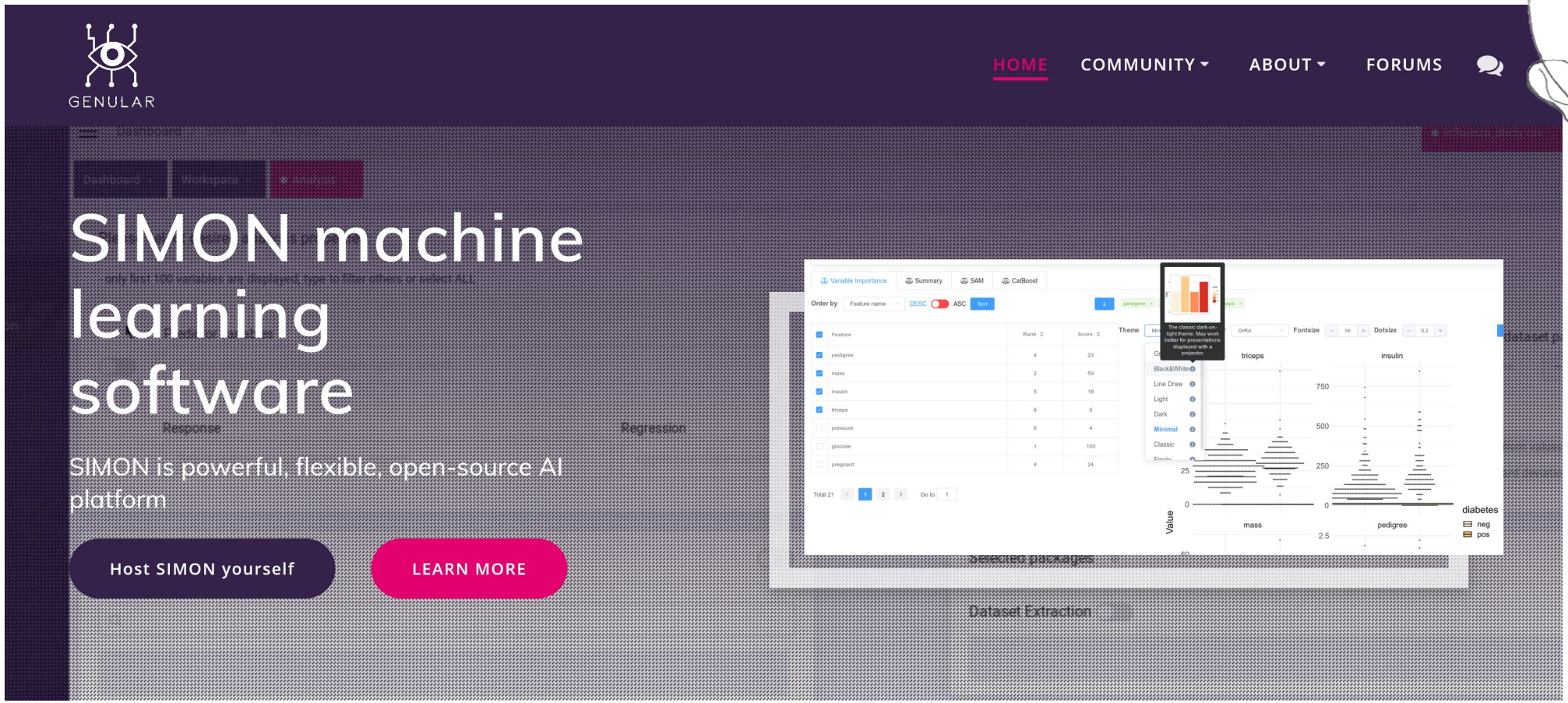
Experiments

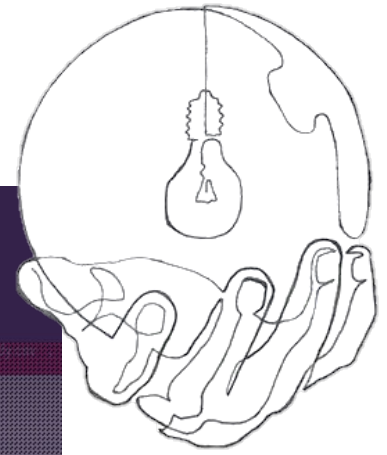
Assays to confirm phenotype and reveal new mechanisms of T cells In both groups



Drost and Celevers, Development, 2017

Join open-source community supporting SIMON!





AGPL v3 Free Software
Free as in Freedom

Check out SIMON at **genular.org**

Have a Question?

Search the documentation...

Search

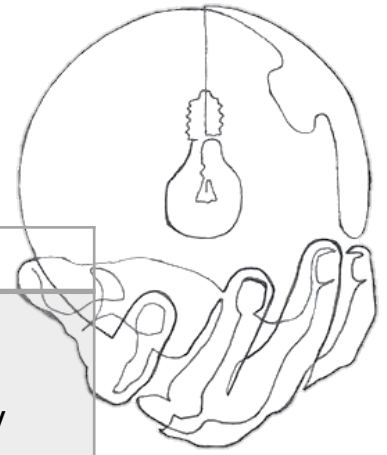
Installation

- Installation Quickstart

Machine Learning

- How to perform SIMON analysis?
- Instruction videos

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**

Thank you

Stanford MEDICINE | Mark M. Davis Lab
Department of Microbiology and Immunology

Elsa Sola Verges
Allison Nau
Lisa Wagar

Stanford MEDICINE | Stanford-LPCH Vaccine Program

Cornelia L. Dekker

Stanford MEDICINE | Institute for Immunity, Transplantation and Infection

The Human Immune Monitoring Center

Mike Leipold
Yael Rosenberg-Hasson
Janine Bodea Sung
Holden Maecker

Stanford MEDICINE | Institute for Immunity, Transplantation and Infection

Biomedical Data Science

Purvesh Khatri

Medawar pathogen research

Paul Klenerman
Susanna Dunachie
Ellie Barnes
Donal Skelly
Ane Ogbe

COMBAT
Oxford COVID-19 Multi-omic Blood ATlas

wellcome centre human genetics

Alex Mentzer
Julian Knight

Ivan Tomic

RStudio

Max Kuhn

UNIVERSITY OF OXFORD

oxford vaccine group

Daniel O'Connor
Chris Dold
Andrew Pollard
Elizabeth Clutterbuck

Duke UNIVERSITY

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

KING'S College LONDON

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

CUNY THE CITY UNIVERSITY OF NEW YORK

Levi Waldron
Ludwig Geistlinger

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Florent Ailloud

UKRI Medical Research Council

NIH National Institutes of Health
Turning Discovery Into Health

Marie Sklodowska-Curie grant (FluPRINT, Project No 796636)

UKCIC
UK Coronavirus Immunology Consortium

PITCH
PROTECTIVE IMMUNITY from T CELLS in HEALTHCARE WORKERS

HIC-Vac