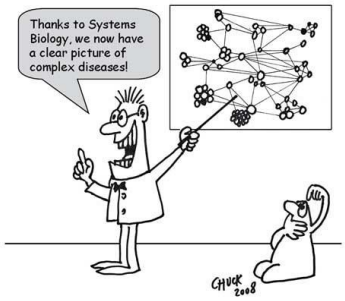
# SIMON says: Machine learning for everyone!




**Adriana Tomic**
Systems Immunology | Oxford Vaccine Group

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

*SIMON training course @Big Data Institute*
*Part I - May 5th, 2021*

1

## Training course - overview

***Part I – SIMON, pattern recognition and knowledge extraction platform (May 5th 2021)***

- Machine learning and AI – what is all the fuss about?
- What is SIMON?
- Case study – example 1 (dealing with missing values, overfitting, model performance)

***Theoretical part – 15min***

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

***Hands-on – 30min***

***Questions? 15min***

***Part II – Exploratory analysis (May 12th 2021)***

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

***Hands-on – 30min***

- Feature processing methods to avoid '*curse of dimensionality*'
- Case studies – example 2

***Theoretical part – 15min***

***Questions? 15min***

2

**Part I. SIMON, pattern recognition and knowledge extraction platform**

3

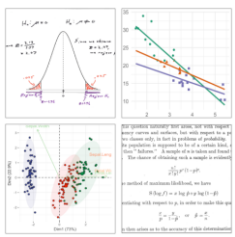
*Hypothesis*

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


*T cells*

*Data analysis*

*Comparison, statistics*

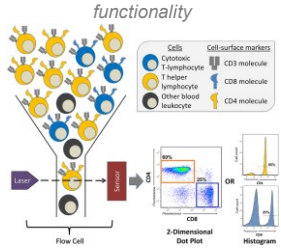


*https://statistics.rutgers.edu/*

Hypothesis-driven research

*Experiments*

*Assays to evaluate frequency, phenotype and functionality*

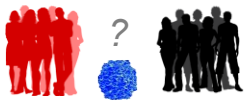



*Verschoor C et al, Front Immunol, 2015*

4

### *Hypothesis*

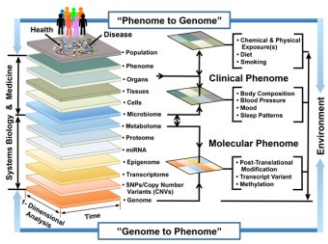
*Why is frequency of T cells increased among healthy vs infected person?*




*T cells*

### *Data analysis*

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

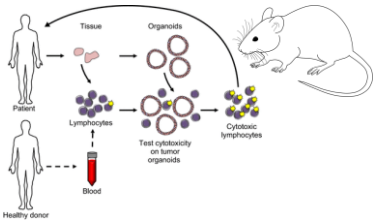



*https://medicine.umich.edu/dept/dcmb/research*

Data-driven research

### *Experiments*

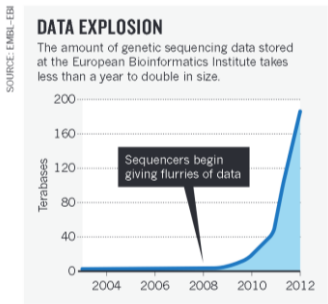
*Assays to confirm phenotype and reveal new mechanisms*



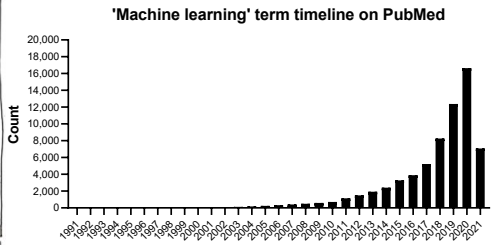

*Drost and Celevers, Development, 2017*

5

## Biology's Big Problem: From data to knowledge







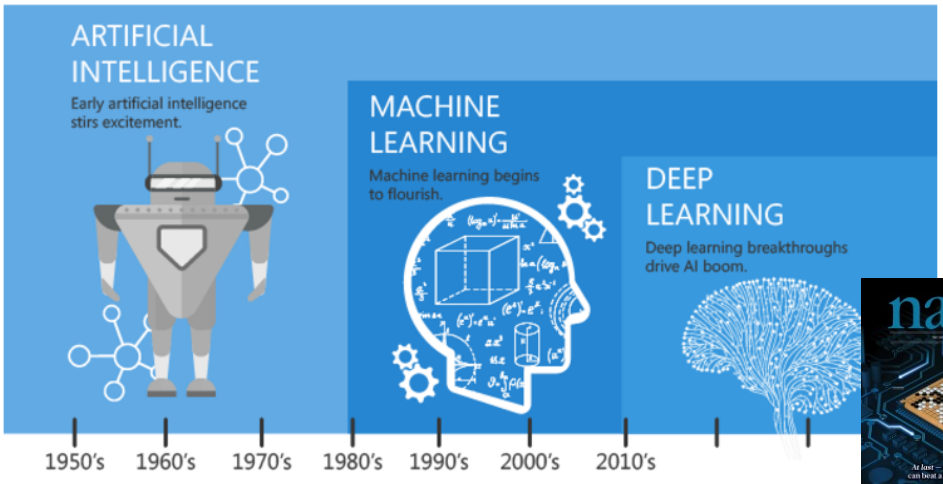

**273 petabytes of data**

(1 petabyte is 1000 terabytes or million gigabytes)

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

6

## Artificial intelligence (AI) to the rescue



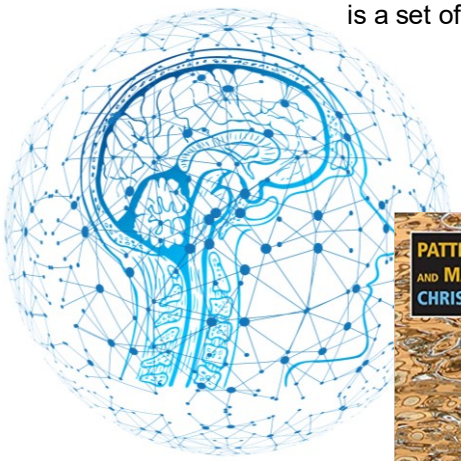


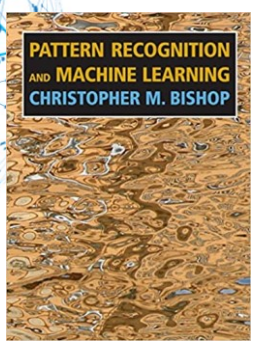
*Kapil Tandon. AI & Machine Learning: The evolution, differences and connections.*

7

**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*



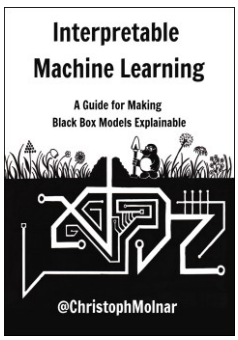



Christopher Bishop; Springer-Verlag New York: 2006

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

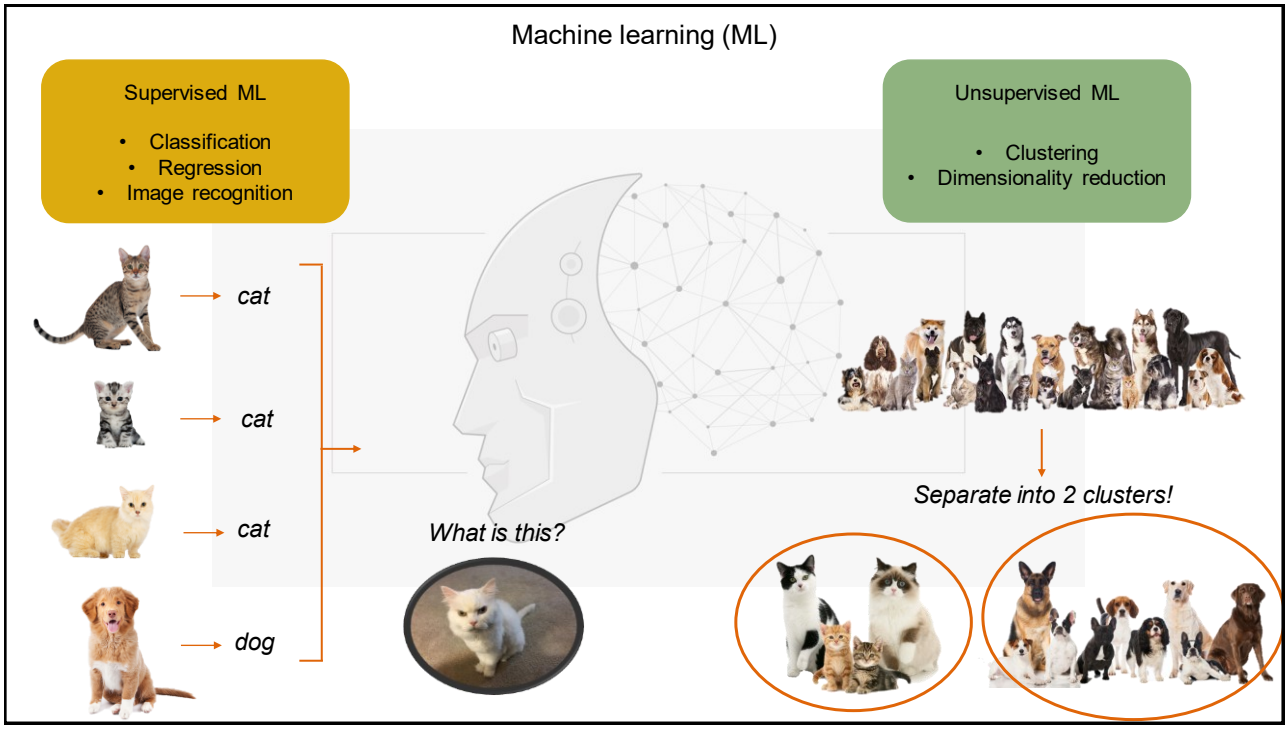

Max Kuhn and Kjell Johnson; Chapman & Hall/CRC Data Science Series: 2019

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

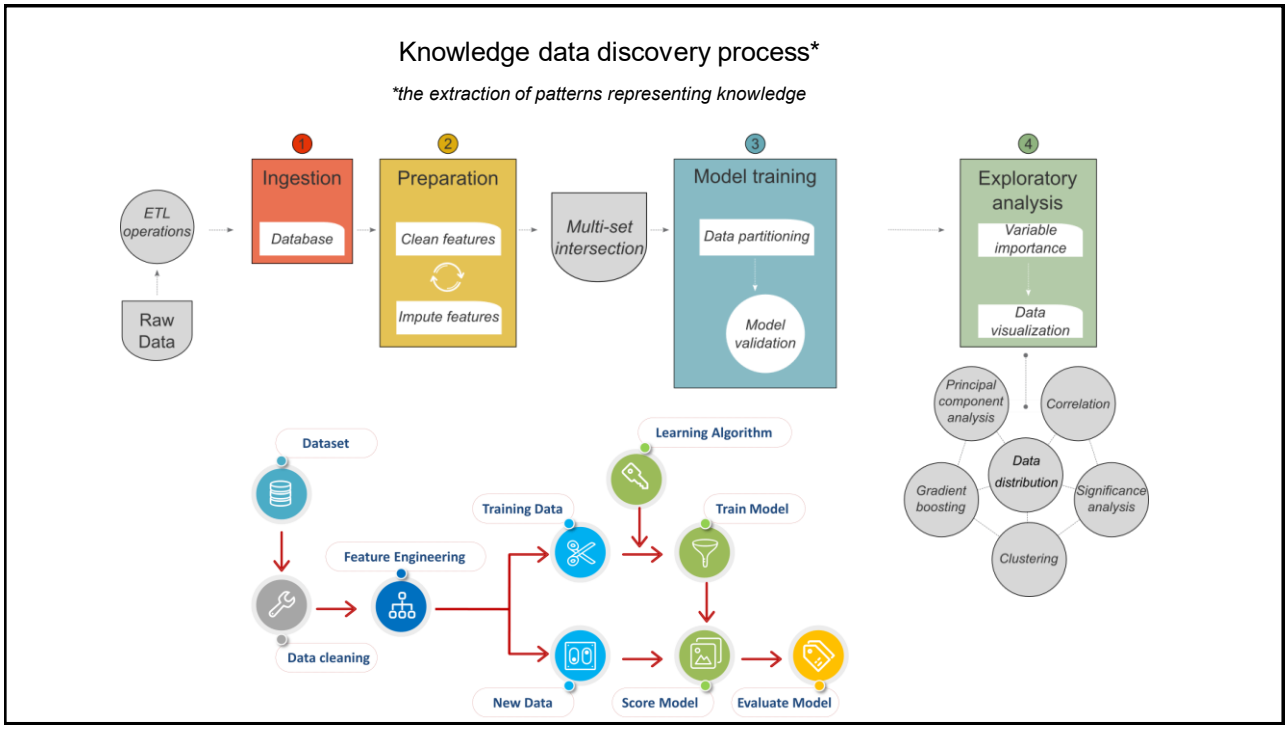



Christopher Molnar; 2021

8


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!

9


Knowledge data discovery process*
*the extraction of patterns representing knowledge
ETL operations
Raw Data
1
Ingestion
Database
2
Preparation
Clean features
Impute features
Multi-set intersection
3
Model training
Data partitioning
Model validation
4
Exploratory analysis
Variable importance
Data visualization
Principal component analysis
Correlation
Data distribution
Gradient boosting
Significance analysis
Clustering
Dataset
Learning Algorithm
Training Data
Train Model
Feature Engineering
Data cleaning
New Data
Score Model
Evaluate Model

10

## 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/

11

## 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

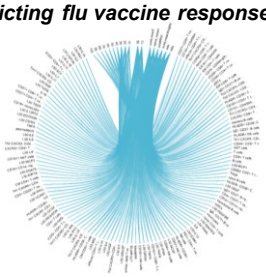
- Waikato Environment for Knowledge Analysis (WEKA) (https://www.cs.waikato.ac.nz/~ml/weka/),
- Orange (https://orange.biolab.si/)
- Konstanz Information Miner (KNIME) https://www.knime.com/)
- ELKI (https://elki-project.github.io/)

**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)

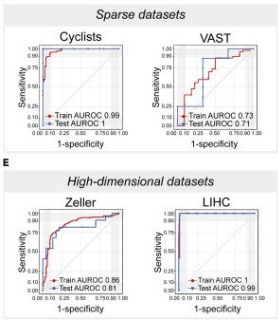
12

***Integrative analysis of different data types – predicting flu vaccine responses***

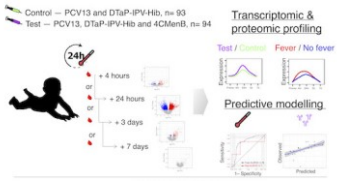


*Tomic et al, JI, 2019*

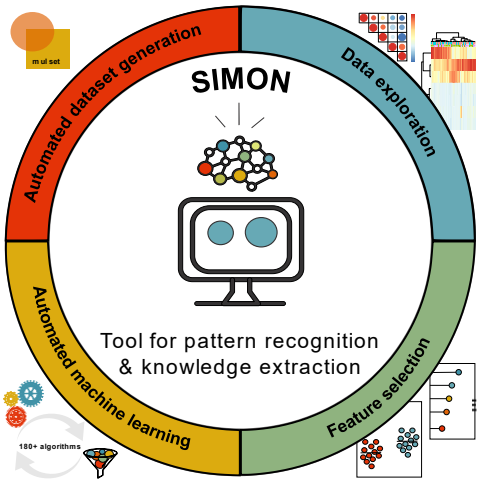
***Datasets with high sparsity or high-dimensionality (transcriptome, microbiome)***




***Transcriptome data***

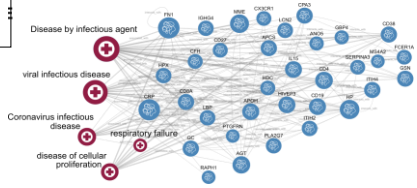



*O'Connor et al, Mol Syst Biol, 2020*




*Tomic et al, Patterns, 2021*

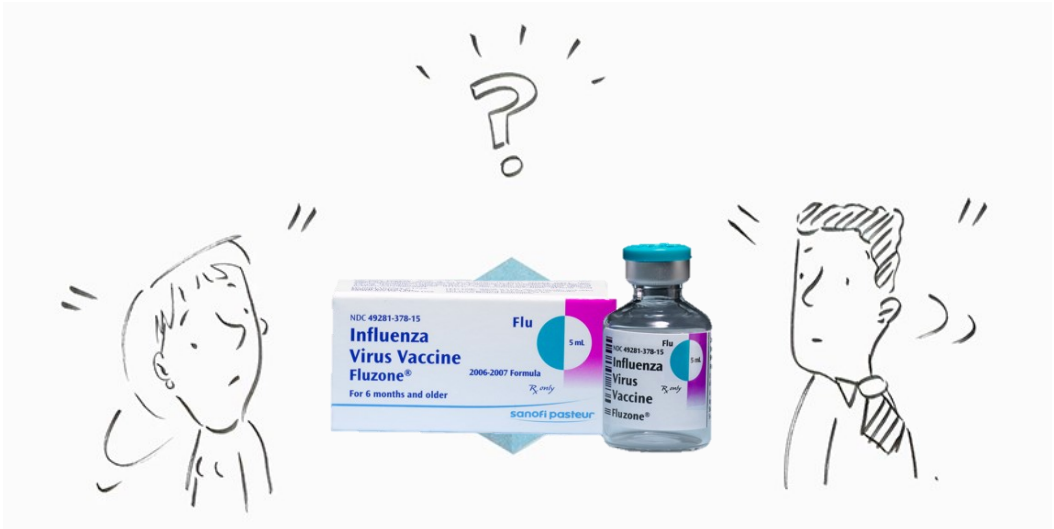
***Multi-omics integrative analysis – COVID-19 COMBAT project***




*COMBAT project, submitted*

13

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




14

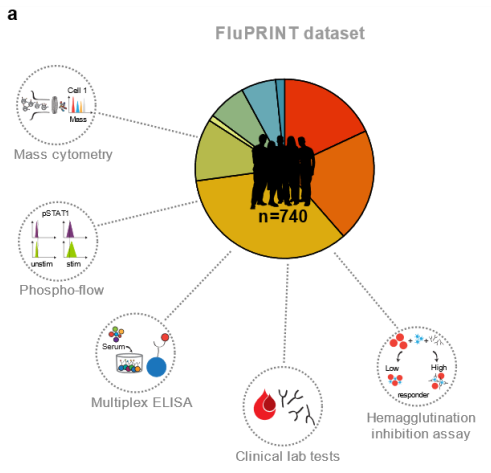
## Hidden treasures at Stanford

*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, …

15

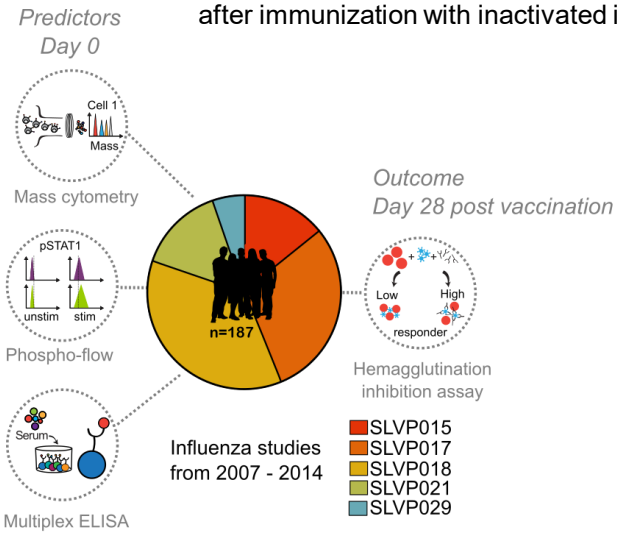
## FluPRINT database – 740 individuals




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

16

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

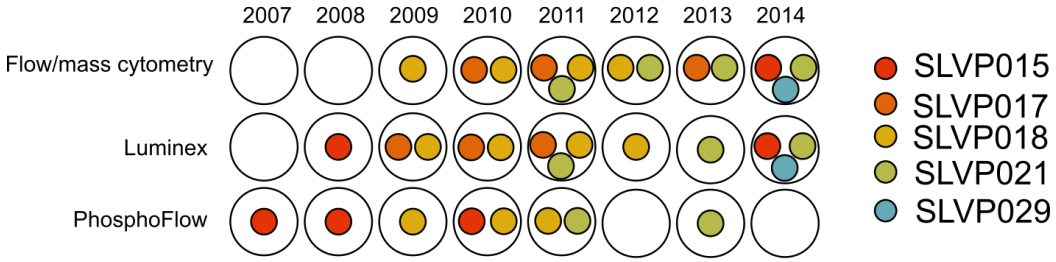



Stanford Human Immune Monitoring Center


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

17

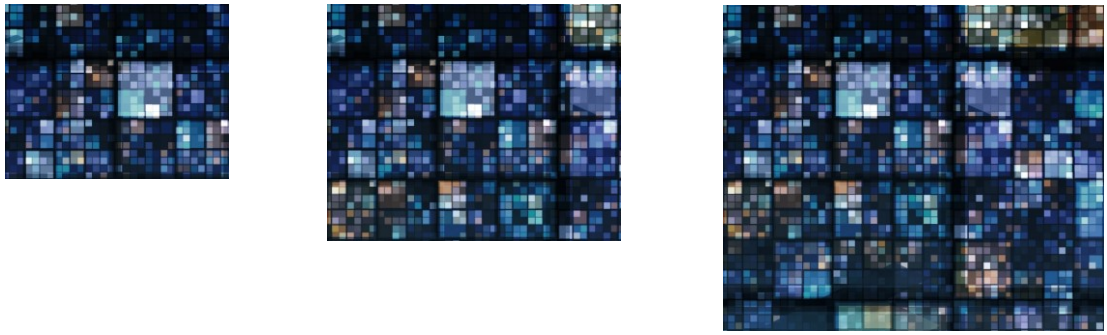
Dealing with missing values




18

The "BIG" problem: Highly percentage of missing data

*2007* *2010* *2014*



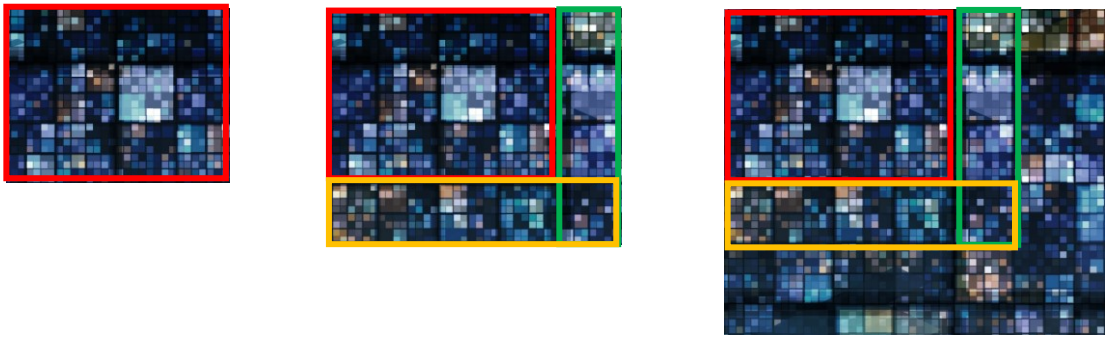
How to select optimal number of donors and optimal number of features?

*SUBSAMPLING*

19

The alternative solution to cope with high sparsity

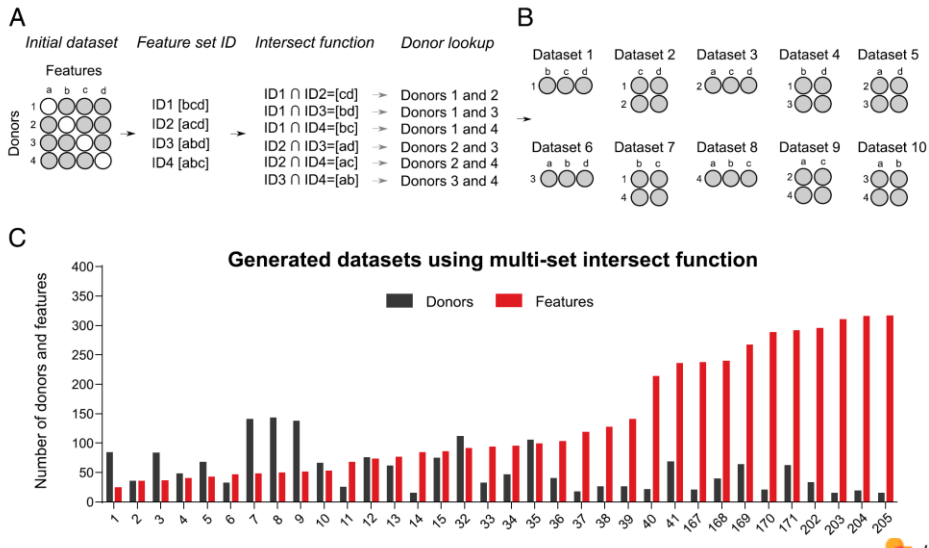
*2007* *2010* *2014*



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

20

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



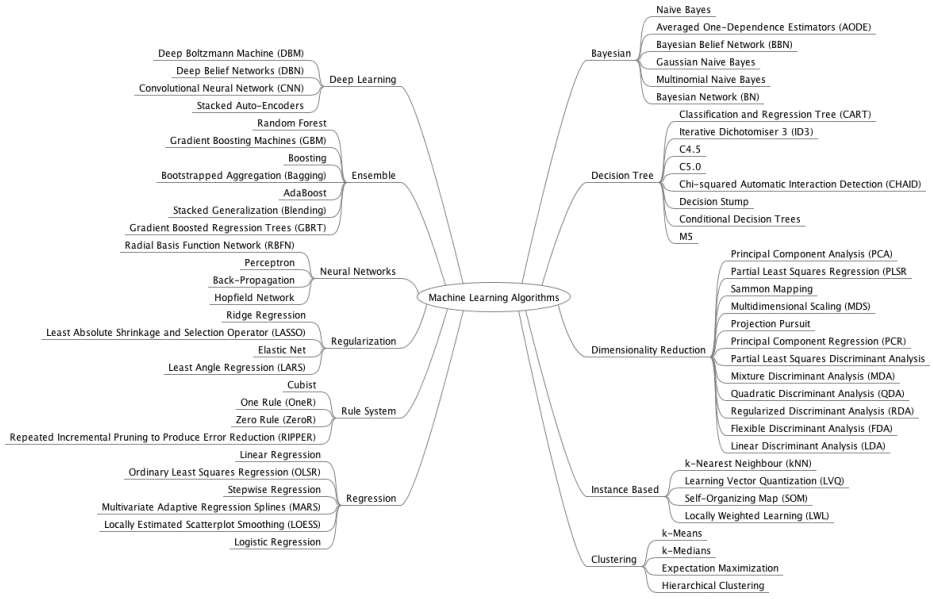

*Tomic et al. SIMON, an Automated Machine Learning System, Reveals Immune Signatures of Influenza Vaccine Responses, JI, 2019*

# mulset

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

21

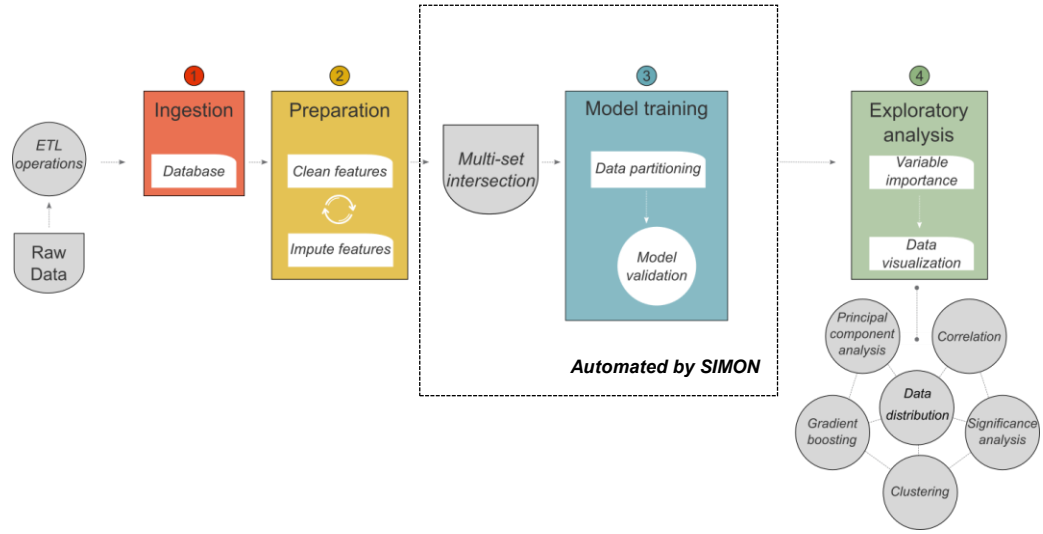
## *Machine learning models: Which one to use? Use all of them!*




*Jason Brownlee, machinelearningmastery.com*

22

## *SIMON*: Sequential Iterative Modelling OverNight




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

23

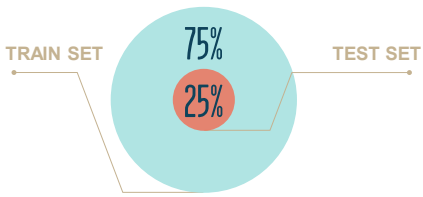
## *SIMON*: Sequential Iterative Modelling OverNight

*180+ machine learning algorithms*

Random Forest
Support Vector Machine
Discriminant Analysis
...

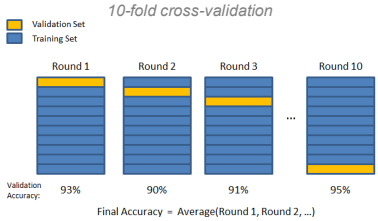
MODEL

*Data partitioning*




*MODEL evaluation*

*10-fold cross-validation*

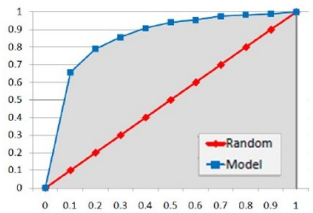



*Model selection*

*Confusion matrix*

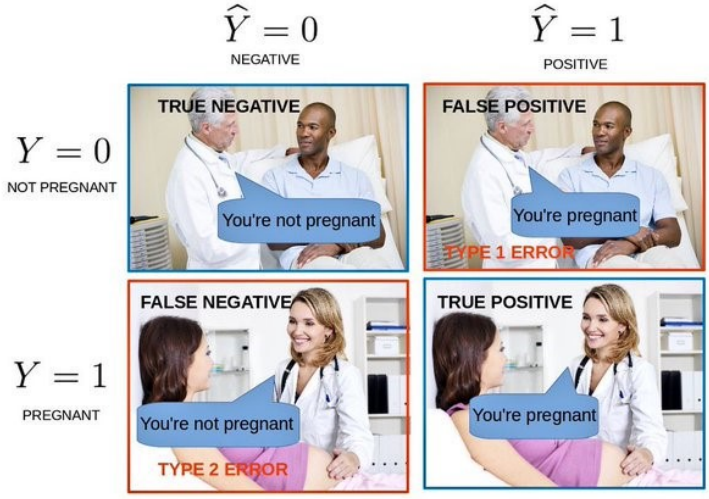
| n=165 | Predicted: High responders | Predicted: Low responders |
|---|---|---|
| Actual: High responders | 50 | 10 |
| Actual: Low responders | 5 | 100 |

*AUROC*




24

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

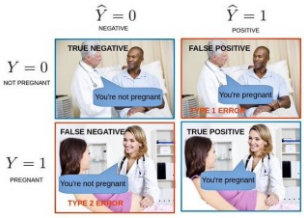



**confusion matrix -** records correctly and incorrectly recognized examples for each class

25

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

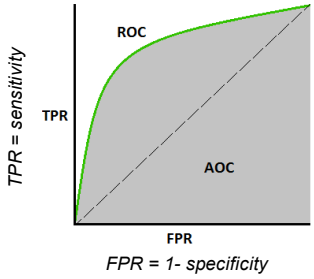
ACCURACY=How often the classifier is correct? True positive + true negative/ sum of all

True positive/ (true positive + false negative)

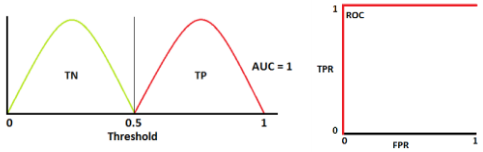
True negative/ (true negative + false positive)

26

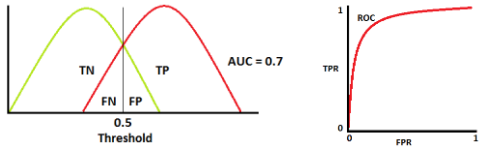
**AUROC** – the most important evaluation metrics for checking any classification model's performance



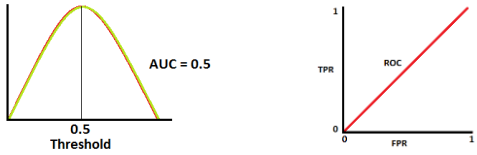

**Perfect model: AUROC = 1**




**Good model: AUROC = 0.70**




**Random model: AUROC = 0.50**



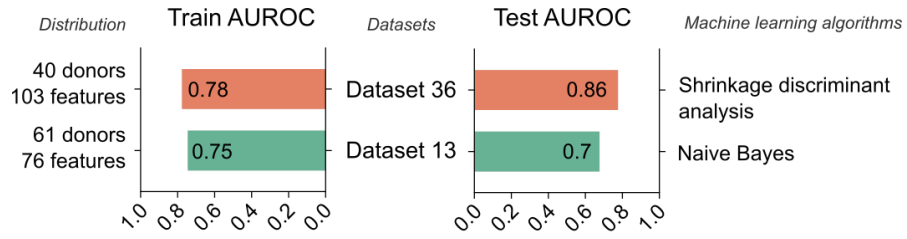

*Note: Red distribution curve is of the positive class (patients with disease) and green distribution curve is of negative class(patients with no disease).*

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

*AUROC (Area Under the Receiver Operating Characteristics)

27

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*

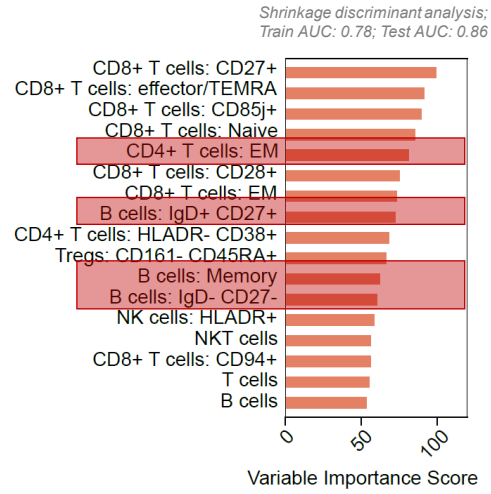
Naïve Bayes
Very simple but powerful. Known to outperform highly sophisticated classification methods

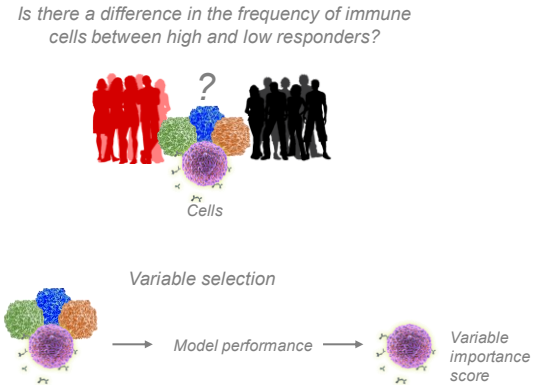
*1960s*

*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*

28

## Pattern recognition in influenza vaccine study using SIMON



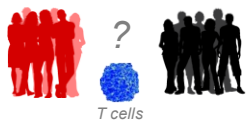




Tregs: CD25hi, CD127-

29

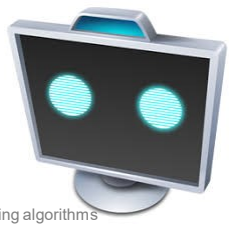
## Hypothesis

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




### Data analysis

*SIMON**

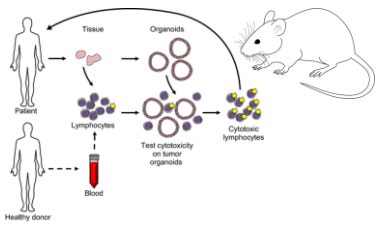


- Feed dataset
- 180+ machine learning algorithms
- Build 1000s of models in one click
- Explore top models
- Identify top important variables

Data-driven research

### Experiments

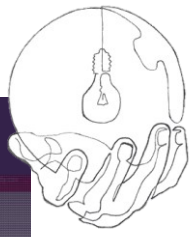
*Assays to confirm phenotype and reveal new mechanisms of T cells In both groups*

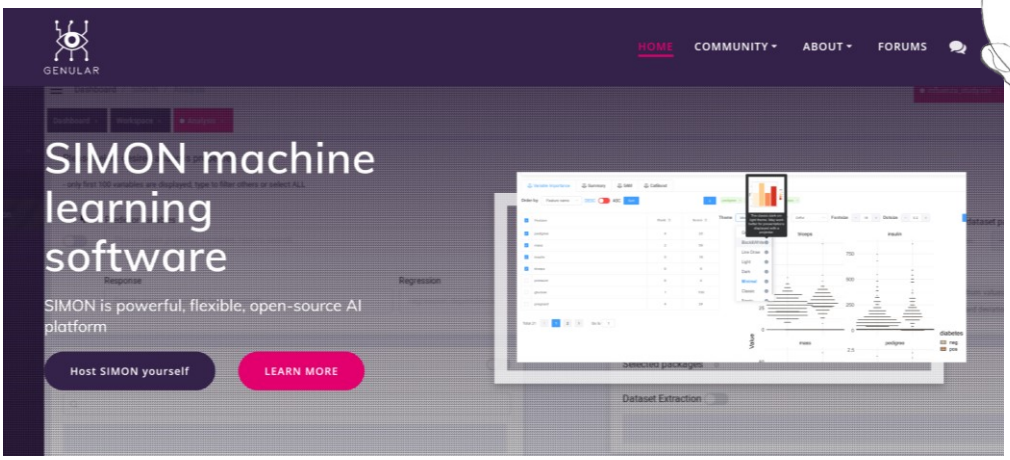



Drost and Celevers, Development, 2017

30

Join open-source community supporting SIMON!

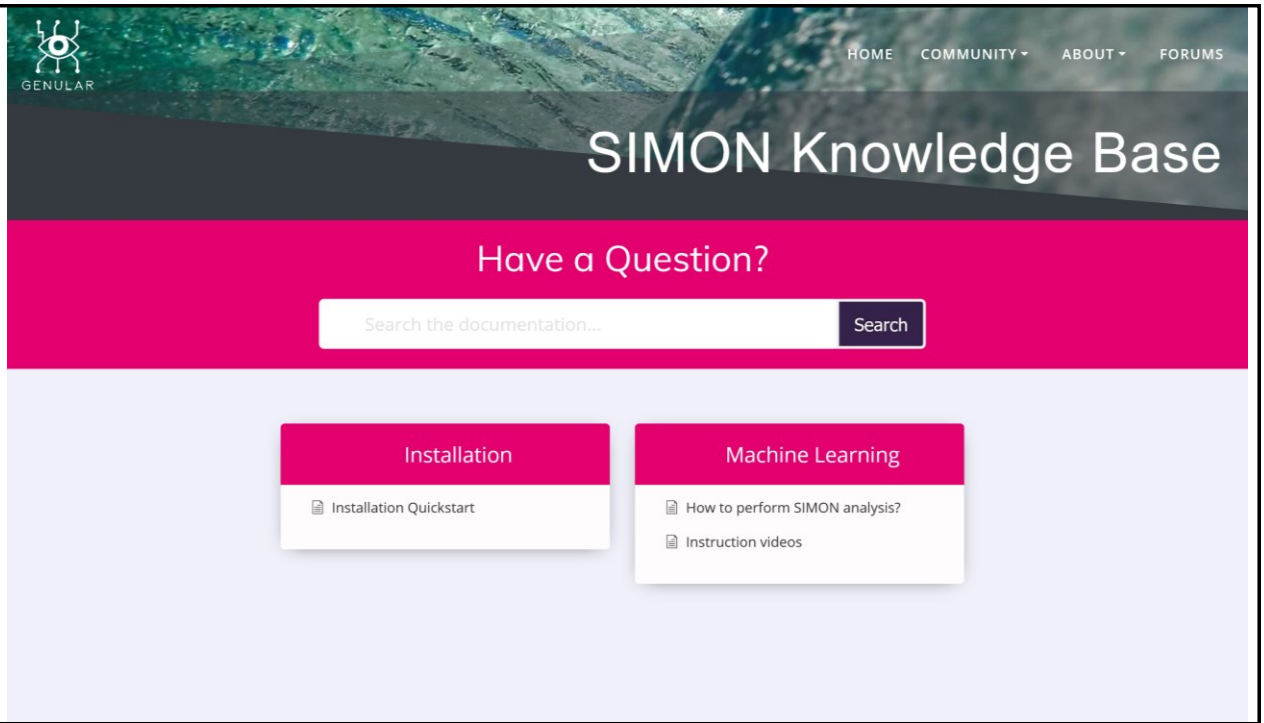





AGPL v3 Free Software
Free as in Freedom

*Check out SIMON at* **genular.org**

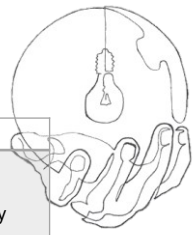
31




32

Star us on GitHub!!!

## Join open-source community supporting SIMON!



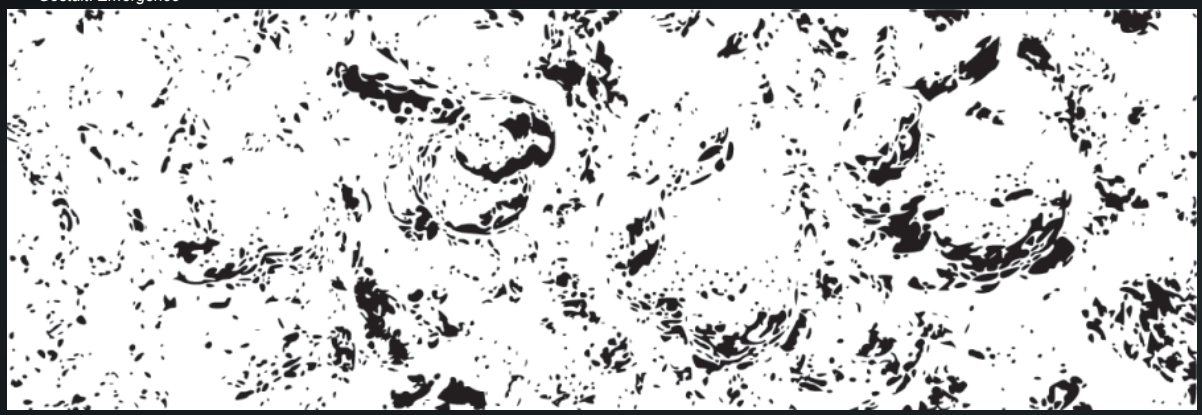
| Project | How To Help | Next Step |
|---|---|---|
| Localization (English, German, French, Chinese, Arabic) | Help us translate SIMON into your language. If some translation is missing or incorrect you can easily help us by correcting it. | Join our Translation Community |
| Tutorials | Help others 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**

33

*SIMON says: Look at the rabbits!*

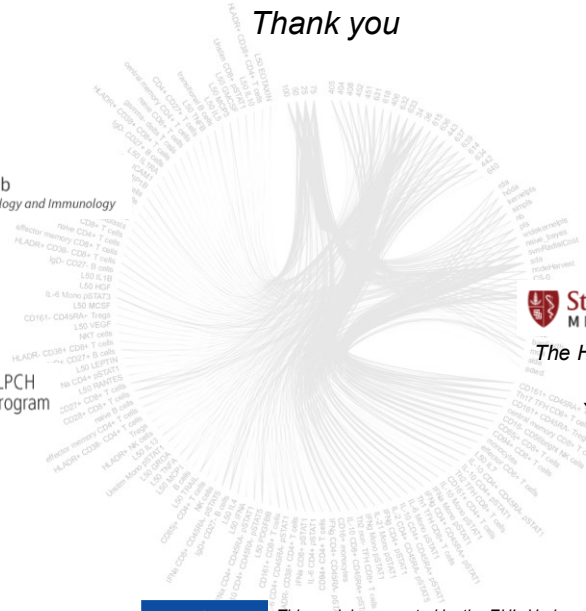
*Gestalt: Emergence*



*Niloy J. Mitra, Hung-Kuo Chu, Tong-Yee Lee, Lior Wolf, Hezy Yeshurun, Daniel Cohen-Or, Emerging images.*

34

# *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

Ivan Tomic

Stanford MEDICINE | Institute for Immunity, Transplantation and Infection

*The Human Immune Monitoring Center*

Mike Leipold
Yael Rosenberg-Hasson
Janine Bodea Sung
Holden Maecker

*This work is supported by the EU's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant (FluPRINT, Project No 796636)*

35