Big Data: metodi e applicazioni DALLA fisica PER la complessità

Nicola Amoroso

Prof. Modelling of Complex Systems
Curriculum: Theoretical Physics and Complex Systems

nicola.amoroso@uniba.it

March 15, 2022

Outline

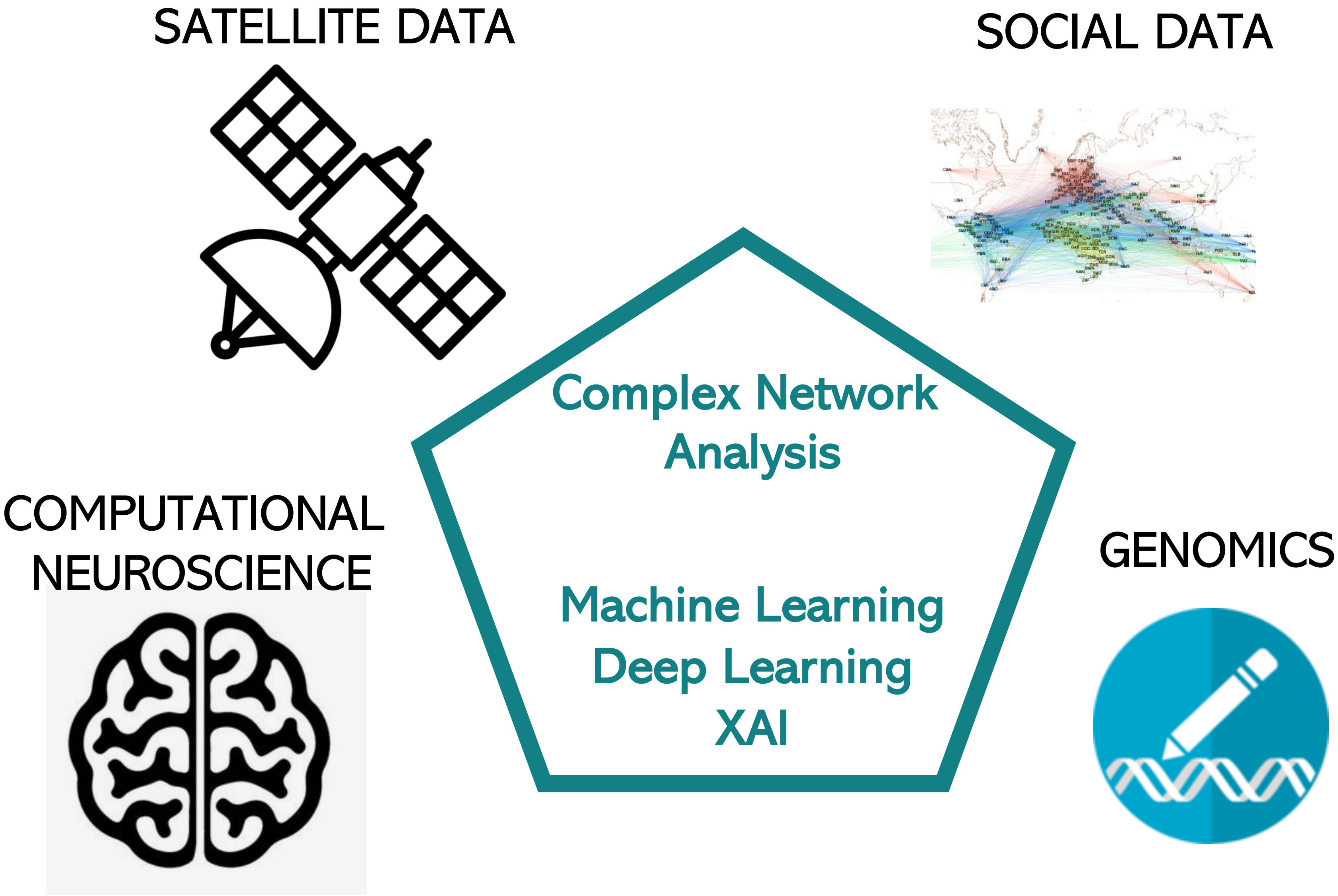
✓Big Data

✓Sistemi complessi

✓Soluzioni dalla Fisica

✓Prospettive

WHAT WE WORK ON



SATELLITE DATA

SOCIAL DATA

Complex Network Analysis

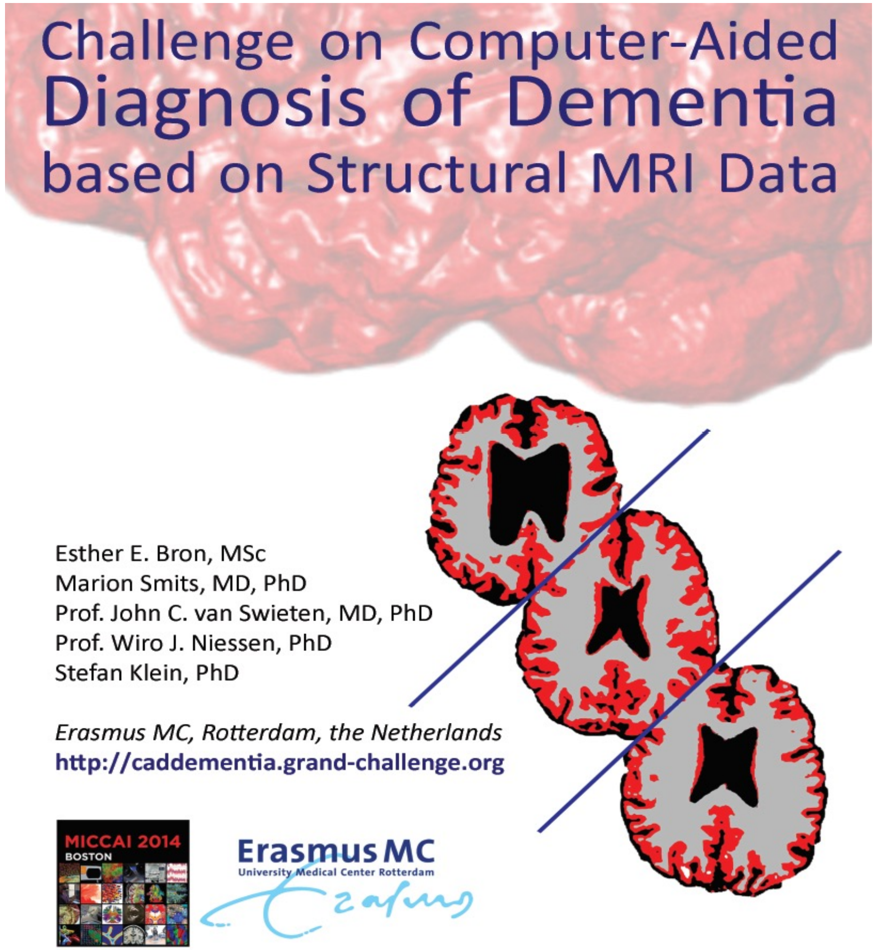
Machine Learning
Deep Learning
XAI

COMPUTATIONAL NEUROSCIENCE

GENOMICS

PARTICIPATION IN INTERNATIONAL CHALLENGES

1



Challenge on Computer-Aided Diagnosis of Dementia based on Structural MRI Data

Esther E. Bron, MSc
Marion Smits, MD, PhD
Prof. John C. van Swieten, MD, PhD
Prof. Wiro J. Niessen, PhD
Stefan Klein, PhD

Erasmus MC, Rotterdam, the Netherlands
http://caddementia.grand-challenge.org

MICCAI 2014
BOSTON
Erasmus MC
University Medical Center Rotterdam

2

MICCAI2016 Athens GREECE

19TH INTERNATIONAL CONFERENCE ON MEDICAL IMAGE COMPUTING & COMPUTER ASSISTED INTERVENTION

3

Alzheimer's Disease Big Data DREAM Challenge

SANOFI · BrightFocus Foundation · Ray and Dagmar Dolby Family Fund · frontiers in NEUROSCIENCE · nature neuroscience · The AddNeuroMed Study · EUROPEAN MEDICINES AGENCY · RUSH UNIVERSITY MEDICAL CENTER · Takeda · ADNI · Alzheimer's Research UK · ALZHEIMER'S PROJECT · Pfizer

INDUSTRIAL RESEARCH PROJECTS

GUNNEBO Industries

TEBAKA
TERRITORIAL BASIC KNOWLEDGE ACQUISITION

network.contacts
Passion to Care

FLAG

TOD'S

DTA
DISTRETTO
TECNOLOGICO
AEROSPAZIALE

SER&P
Software Engineering Research & Practices

btsconsulting

ENGINEERING

3rdPlace
user & customer AI

exprivia

LEONARDO

PON
RICERCA
E INNOVAZIONE
2014 - 2020

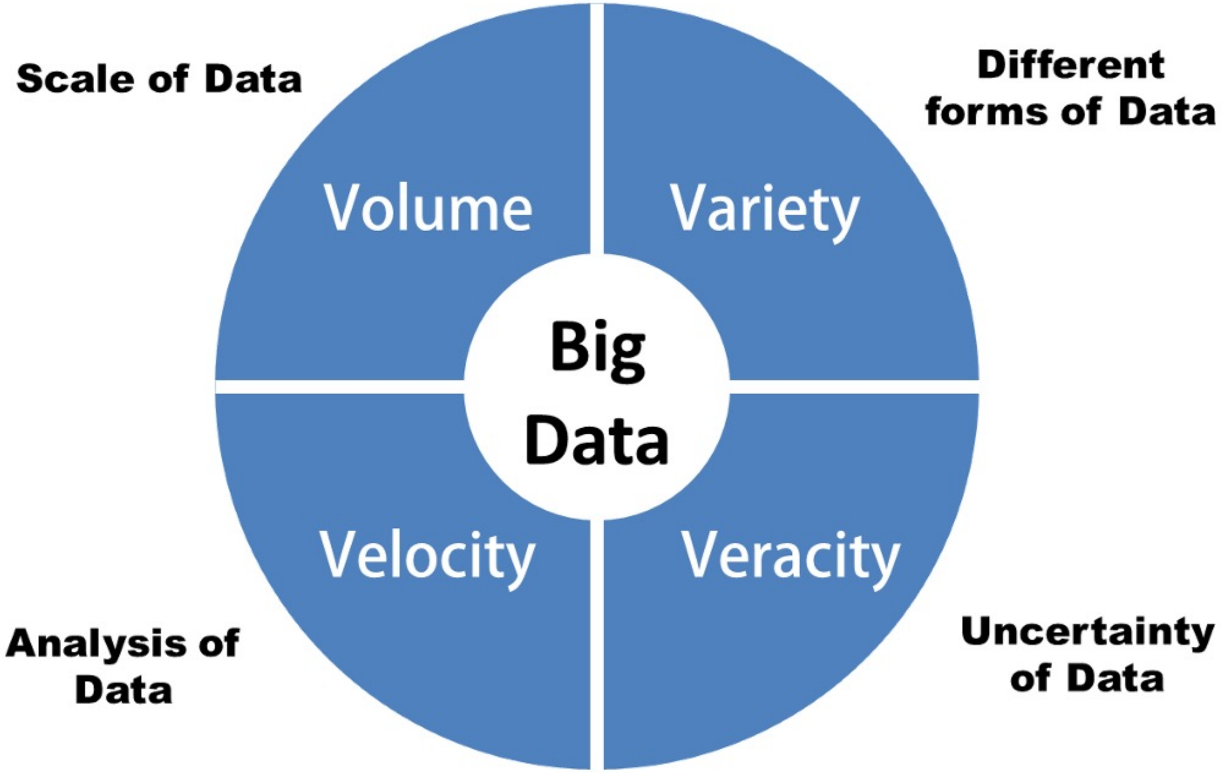
UNIONE EUROPEA
FondoEuropeodiSviluppoRegionale

Unione Europea
FONDI STRUTTURALI EUROPEI
pon 2014-2020
MIUR
PER LA SCUOLA - COMPETENZE E AMBIENTI PER L'APPRENDIMENTO (FSE-FESR)

MINISTERO
DELLA DIFESA

BIG DATA

Big data is high Volume, high Velocity, and/or high Variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. Additionally, a new "V" Veracity is added (IBM) to take into account data consistency.



The volume of data that enterprises acquire every day is increasing exponentially.

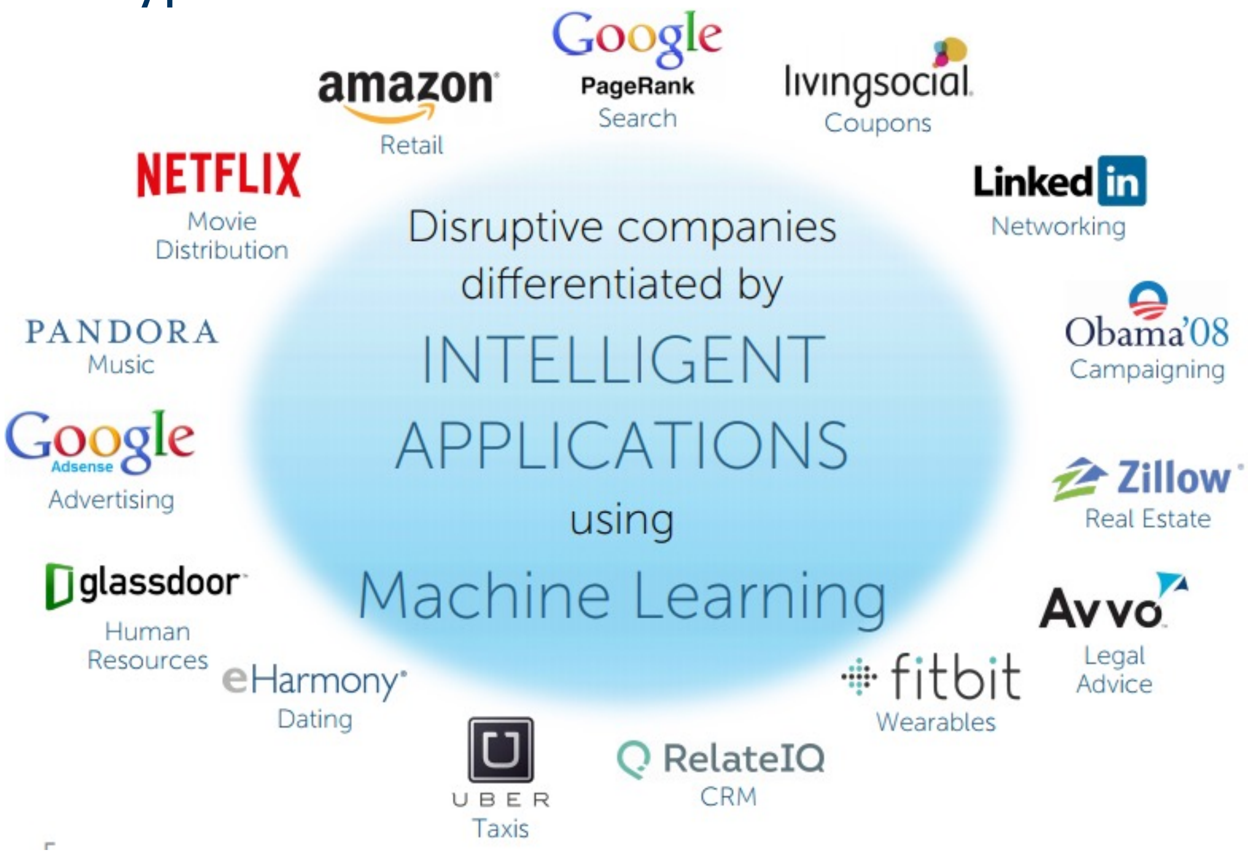
The challenge these organizations now face is what to do with all this data and how to get insights from it.

Thus, R comes into picture. R is a very amazing tool to run advanced statistical models on data.

BIG DATA

Interest in **Big Data** has exploded over the past decade. You see Big Data analytics in computer science programs, industry conferences, and the Wall Street Journal almost daily.

Fundamentally, these algorithms aim at extracting information from raw data and represent it in some type of model.



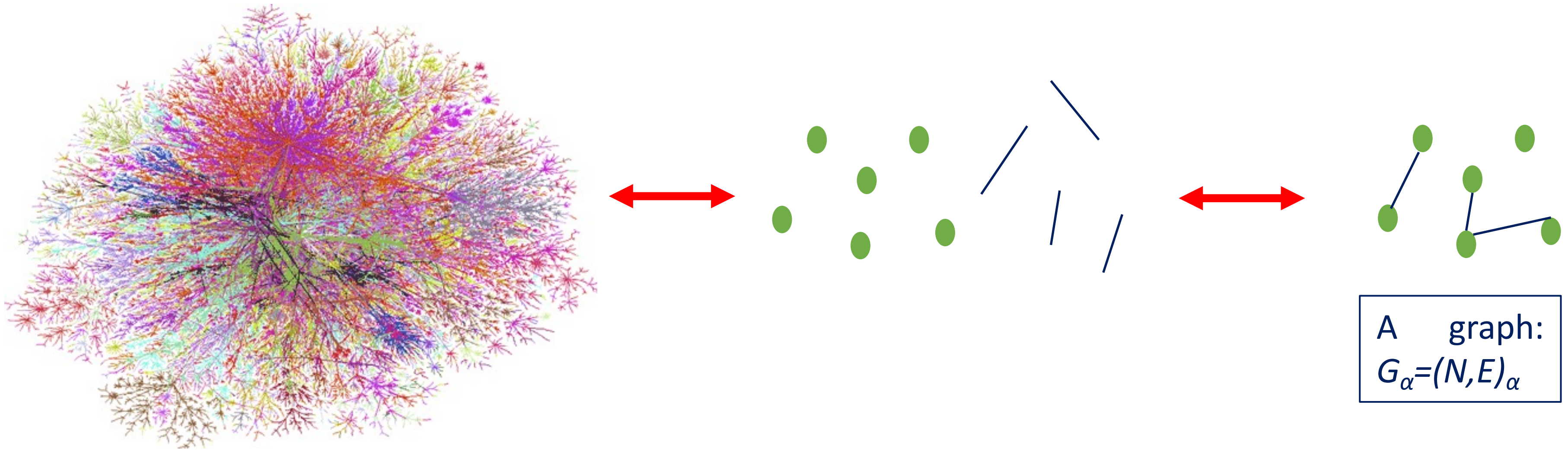
Top applications

1. Medical diagnosis
2. Finance and fraud detection
3. Retail
4. Travel
5. Social networks

A «Physical» choice: Complex Networks

- Complex networks are a direct application of the Theory of Graphs

- Complex Networks offer a set of quantitative measures to characterize a (Physical) system at both a global level and AT the level of its components (nodes and links)

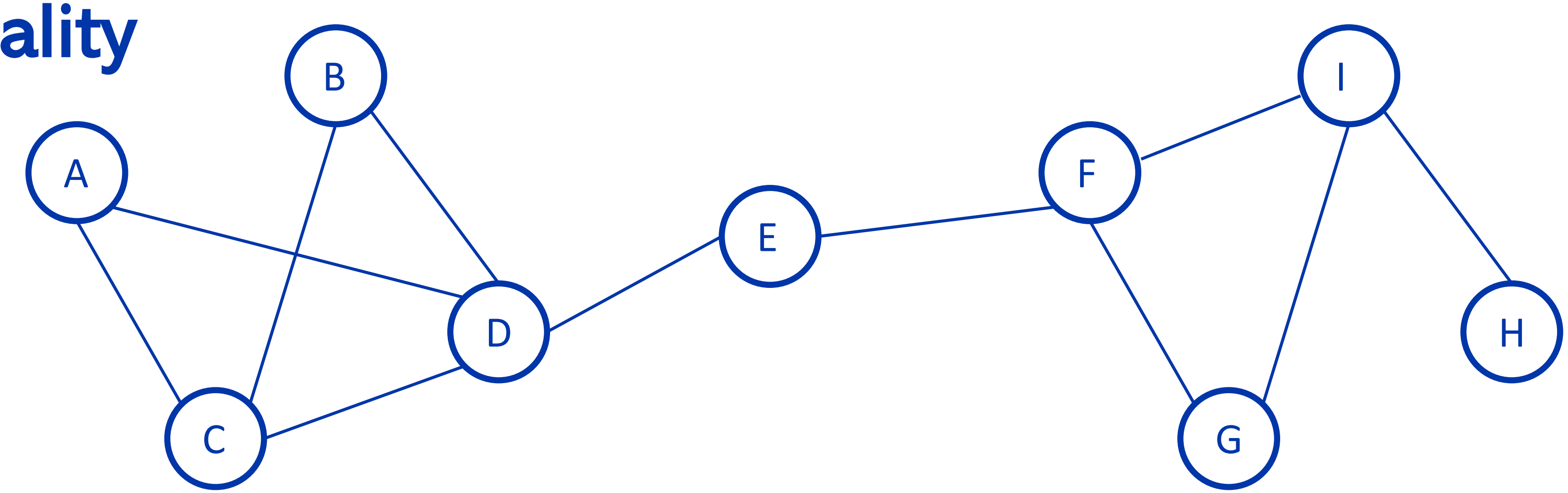


A graph: $G_{\alpha}=(N,E)_{\alpha}$

Statistical Mechanics makes the properties of a system explainable looking at it as a whole instead of looking at its parts (hint. Thermodynamics).

Complex Network Analysis

Node centrality



Node	Degree
A	1
B	2
C	3
D	**4**
E	2
F	3
G	3
H	2
I	3

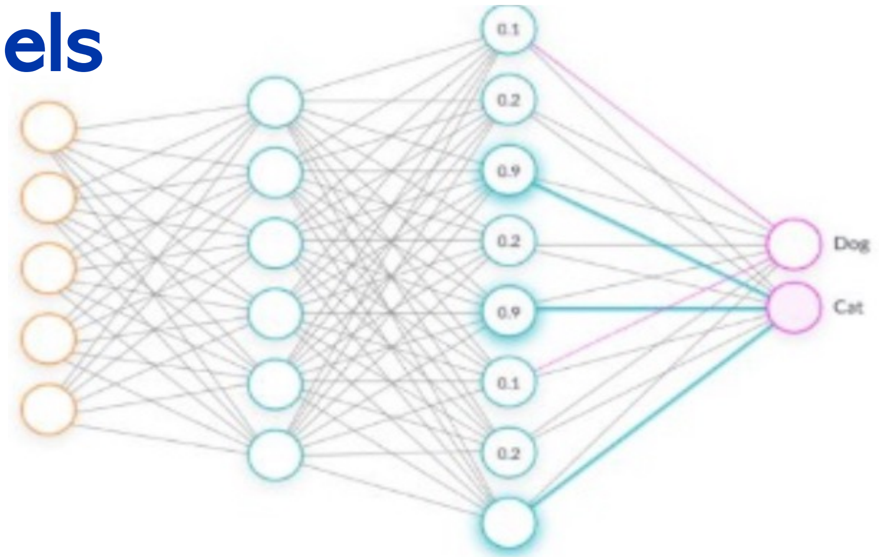
Node D has the most links, however node E, with less links, connects the two sides of the networks. This kind of centrality is measured by an indicator called **betweenness (centrality)**.

Betweenness measures the number of "shortest paths" (between any couple of nodes in the graphs) that passes through the target node.

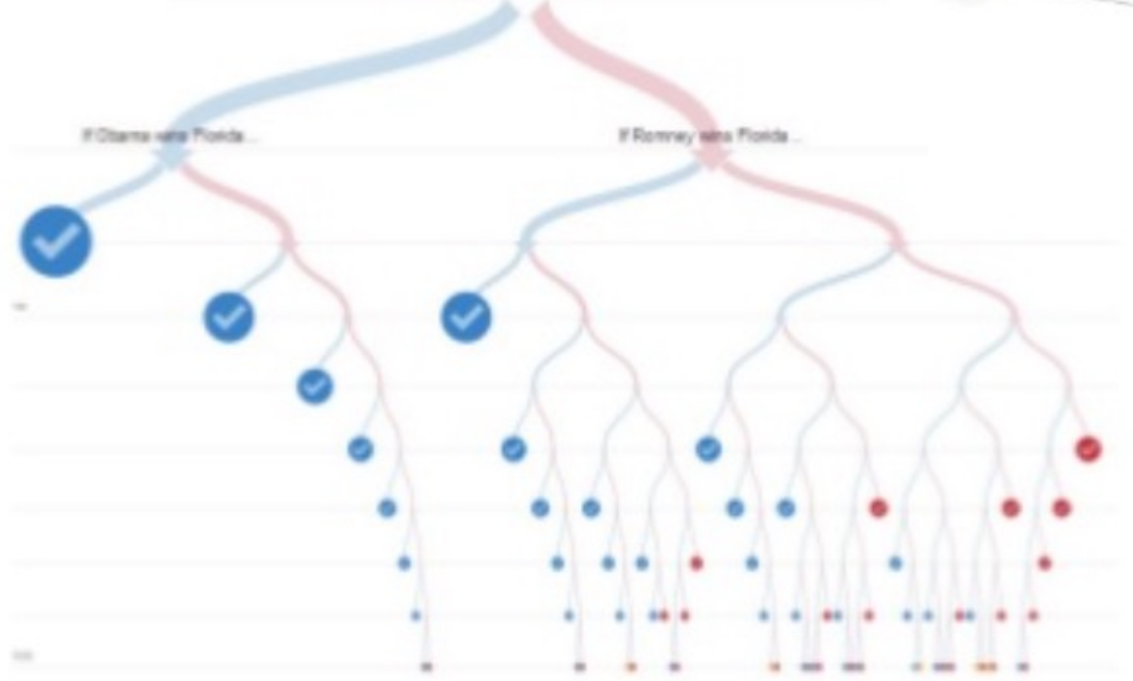
Node	Betweenness
A	0
B	0
C	0.5
D	15.5
E	**16**
F	15.5
G	0
H	0
I	7

From models to reality: ML, DL, XAI

Accuracy VS Interpretability of ML models

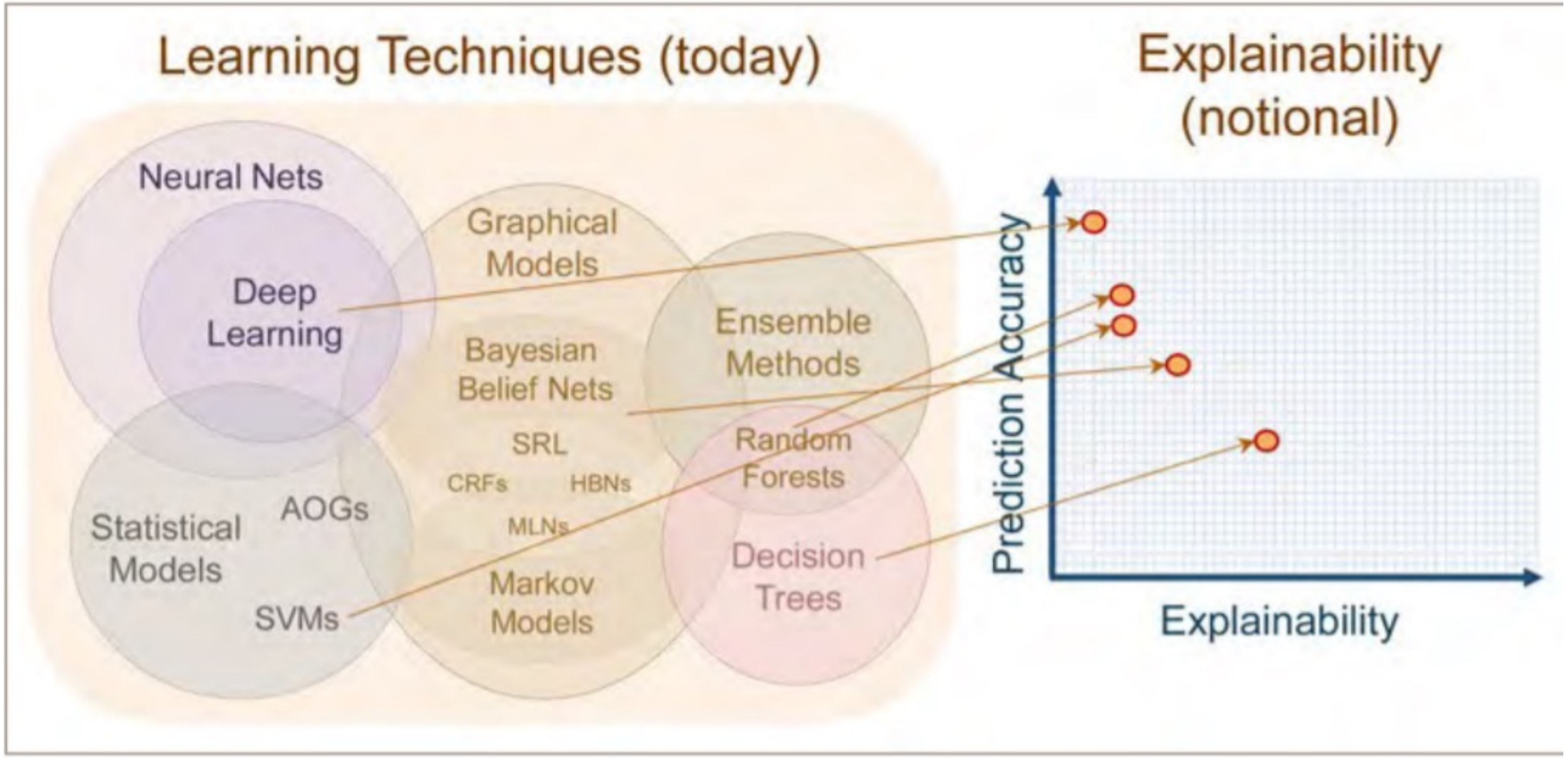


ML creates functions that combine features in sophisticated ways.



It is difficult to disaggregate the final predictions to single feature contributions and untangle interactions among features.

Accuracy VS Interpretability of ML models



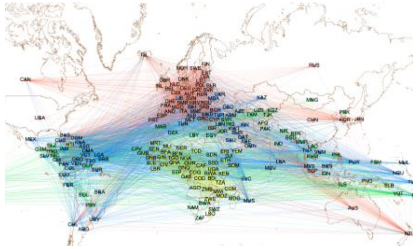
Source: DARPA

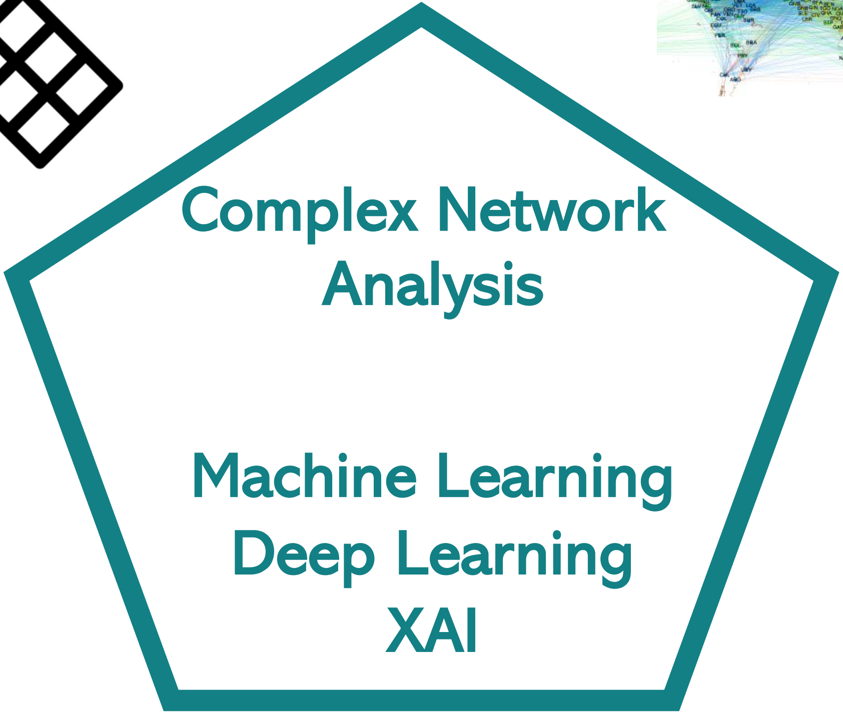
In Physics we are interested in understanding how systems work.
In this context we are interested in understanding how **each predictor** is **contributing to the model** and how the different predictors **interact.**

WHAT WE WORK ON

SATELLITE DATA

SOCIAL DATA





Complex Network Analysis

Machine Learning
Deep Learning
XAI

COMPUTATIONAL NEUROSCIENCE

GENOMICS

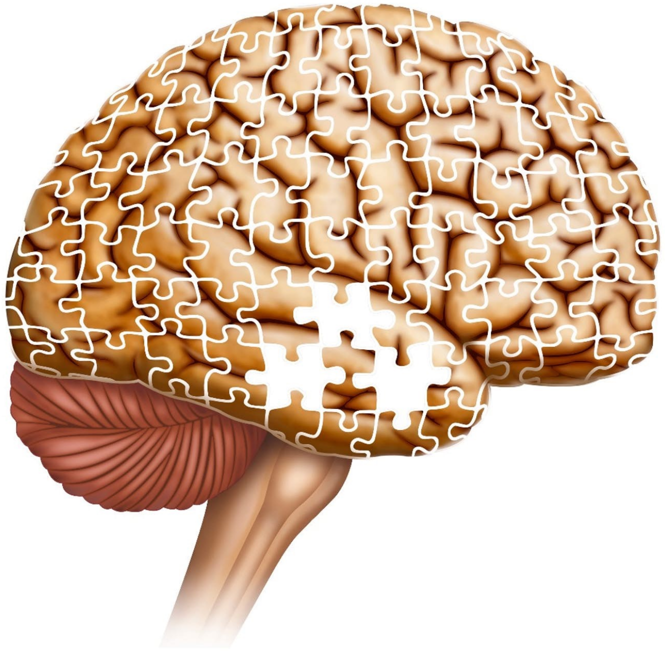
MORE …

COMPUTATIONAL NEUROSCIENCE

Alzheimer's disease

Alzheimer's disease is a degenerative brain disease and the most common cause of dementia.

It is characterized by a decline in memory, language, problem-solving and other cognitive skills that affects a person's ability to perform everyday activities.



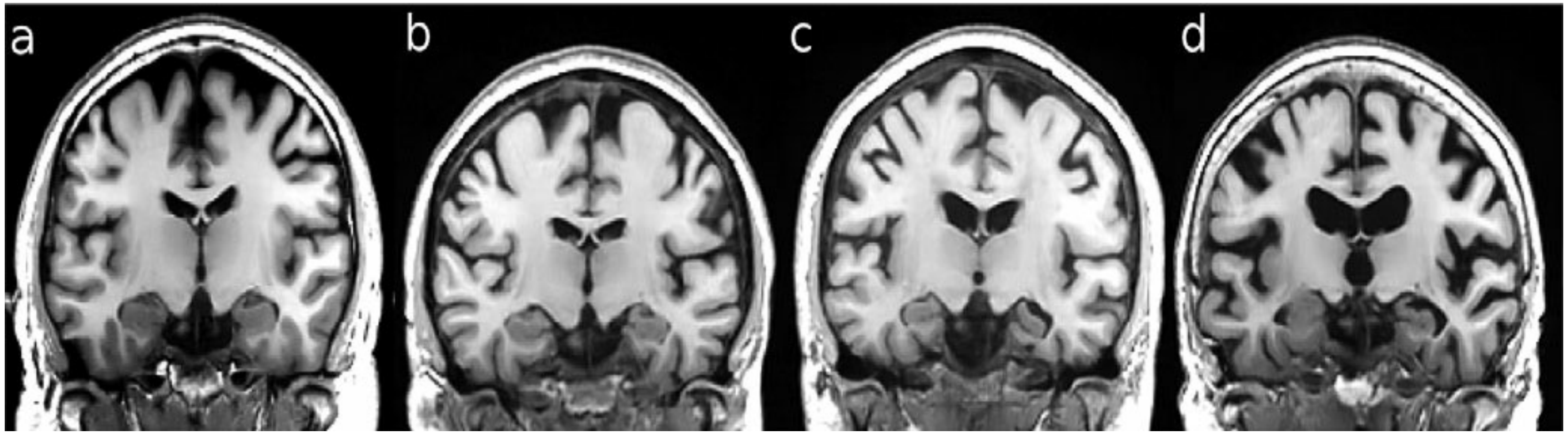
This decline occurs because neurons in parts of the brain involved in cognitive function have been damaged and no longer function normally.

People in the final stages of the disease are bed-bound and require around-the-clock care.

Clinical practice

Structural imaging based on magnetic resonance is an integral part of the clinical assessment of patients with suspected Alzheimer dementia.

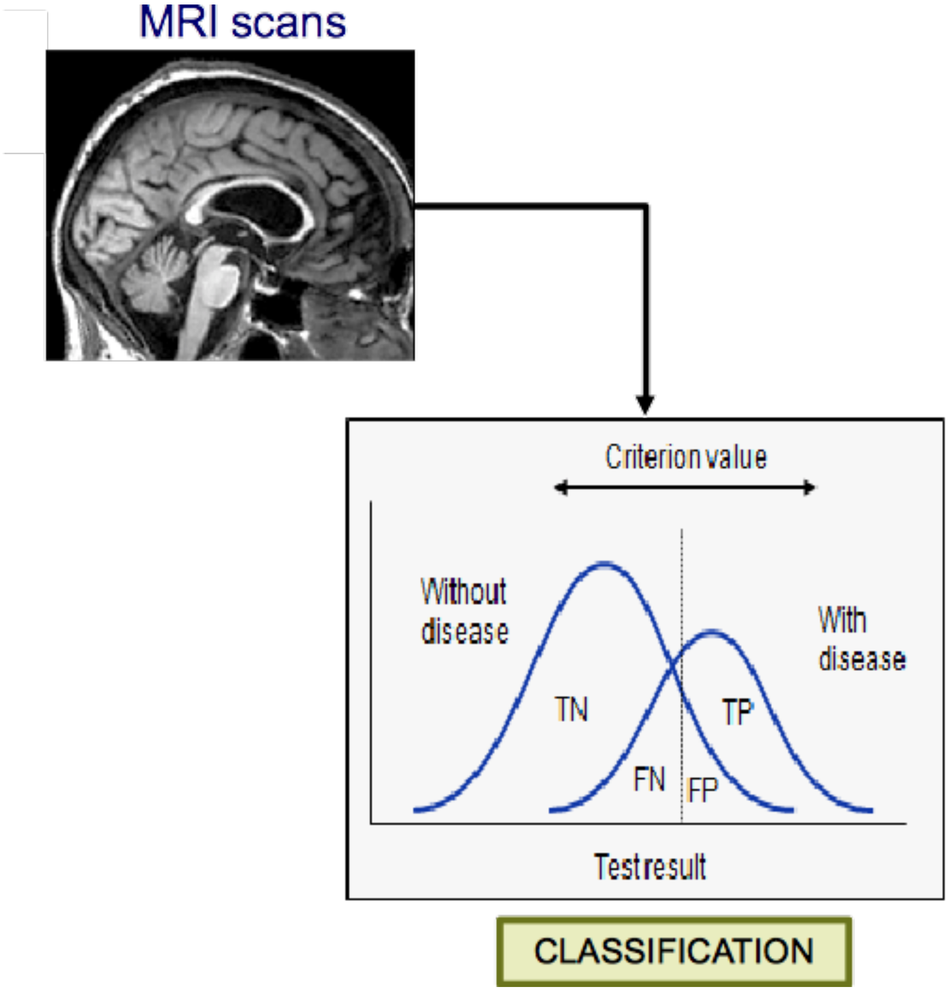
Rates of whole-brain and hippocampal atrophy are sensitive markers of neurodegeneration, and are increasingly used as **outcome measures** in trials of potentially disease-modifying therapies.



Coronal view of a sample progression of pathological brain atrophy in intensity and spatially normalized MRI scans. Subjective visual rating of the medial temporal lobe atrophy can be assessed on these MR films: (a) absent, (b) minimal, (c) moderate, and (d) severe.

COMPUTATIONAL NEUROSCIENCE

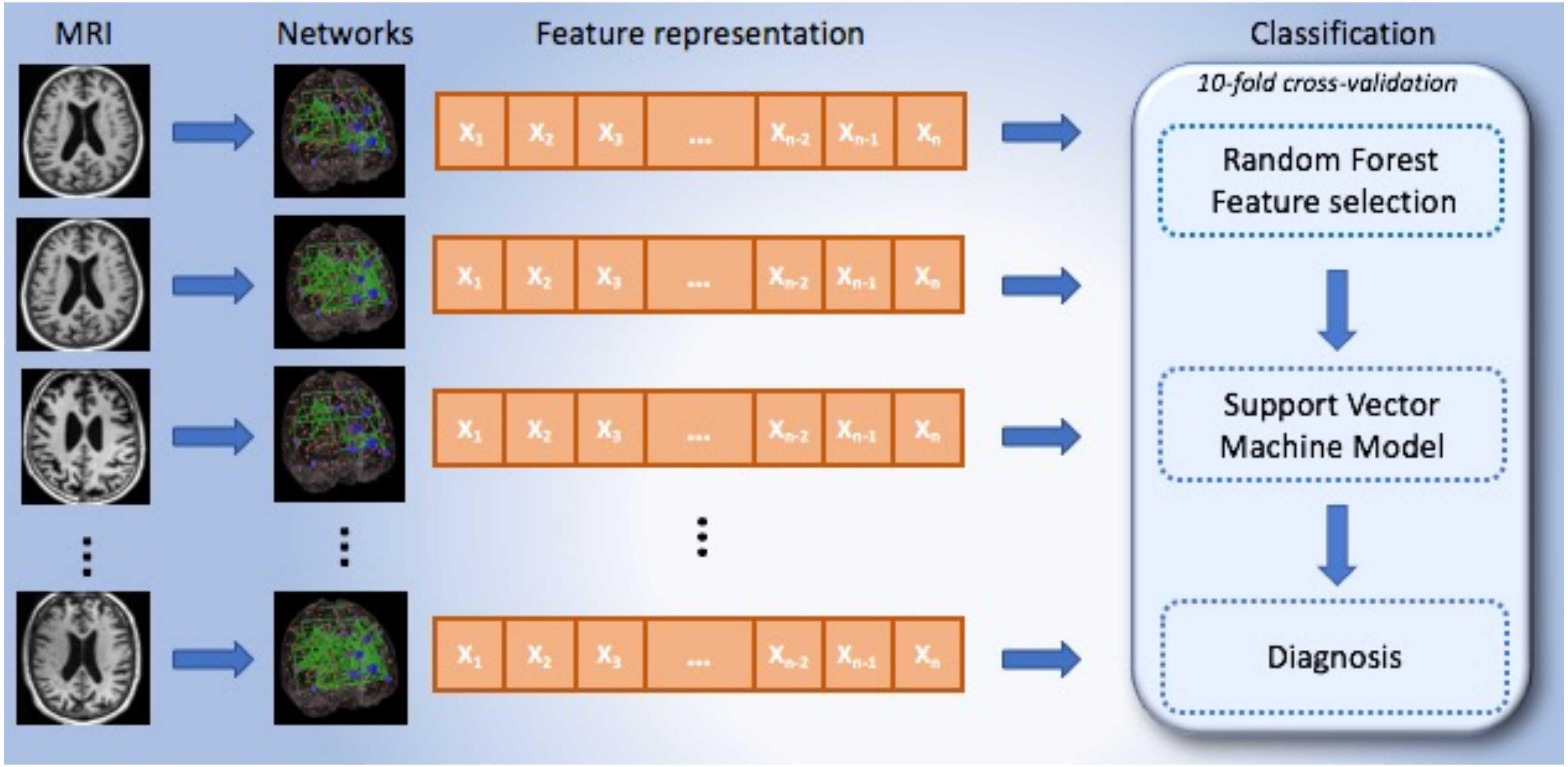
What are we aiming at?



We aim at a **quantitative measure** of a significant biomarker/criterion and its **error**.

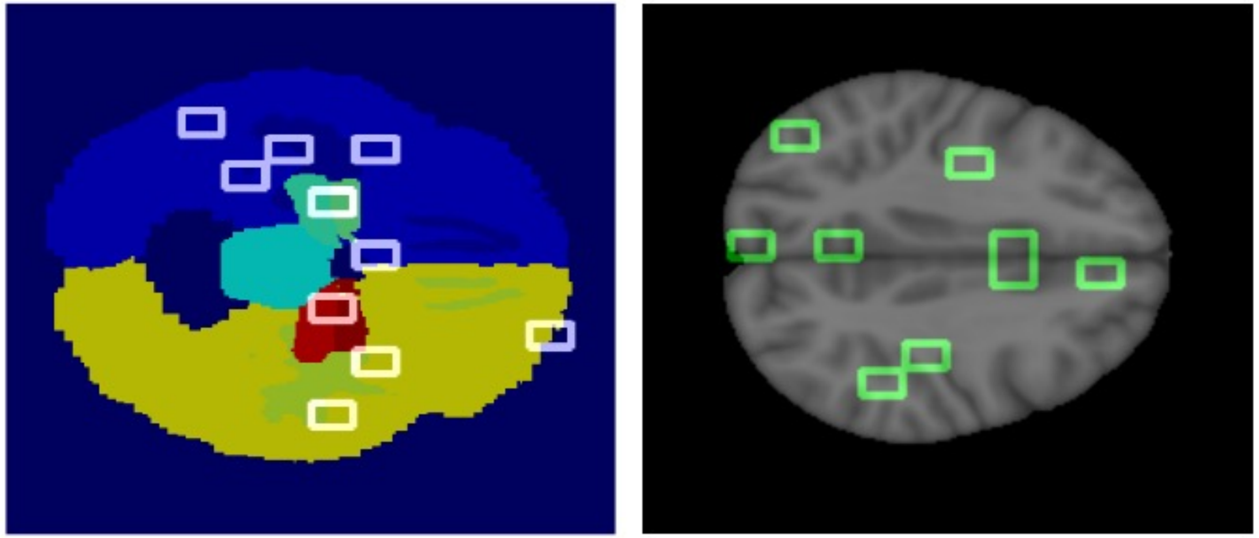
- Classification
  - Normalcy vs. *probable* Pathology
  - Pathology “X” vs. pathology “Y”
- “Continuous” index
  - Proportional to pathology degree
  - Usable for follow-ups, decline rate, subject ranking, drug trials, …

Alzheimer


MRI
Networks
Feature representation
Classification
x_1 x_2 x_3 ... x_{n-2} x_{n-1} x_n
10-fold cross-validation
Random Forest Feature selection
Support Vector Machine Model
Diagnosis

Explainining and Predicting

COMPUTATIONAL NEUROSCIENCE



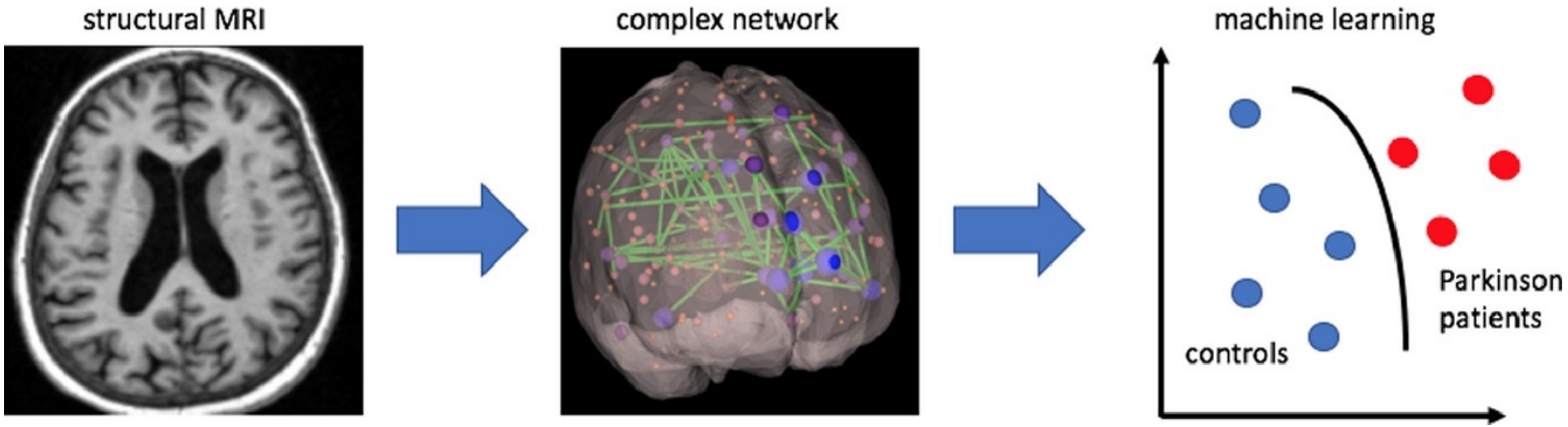
Starting from MRI, we can outline **important regions** for Alzheimer's diagnosis.

We highlighted these important areas and confirmed that the selected regions have previously been reported to be connected to Alzheimer.

COMPUTATIONAL NEUROSCIENCE

Parkinson’s disease

Parkinson’s disease is associated with mild cognitive impairment and dementia, so that it is possible to hypothesize that its diagnosis can be related to brain structure as Alzheimer’s disease.



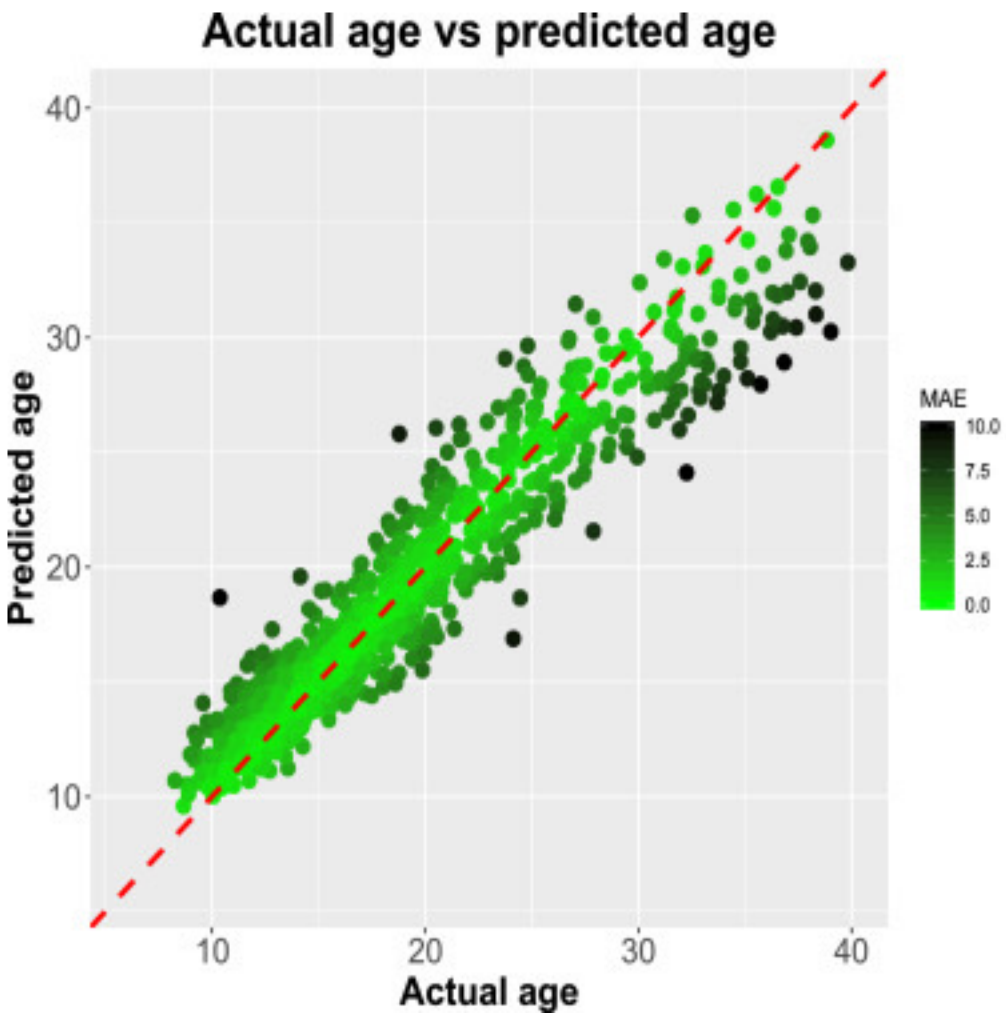
Our model compares favorably with existing state-of-the-art MRI approaches (AUC = 0.97 ± 0.02).

COMPUTATIONAL NEUROSCIENCE

The Brain Age Gap as a biomarker

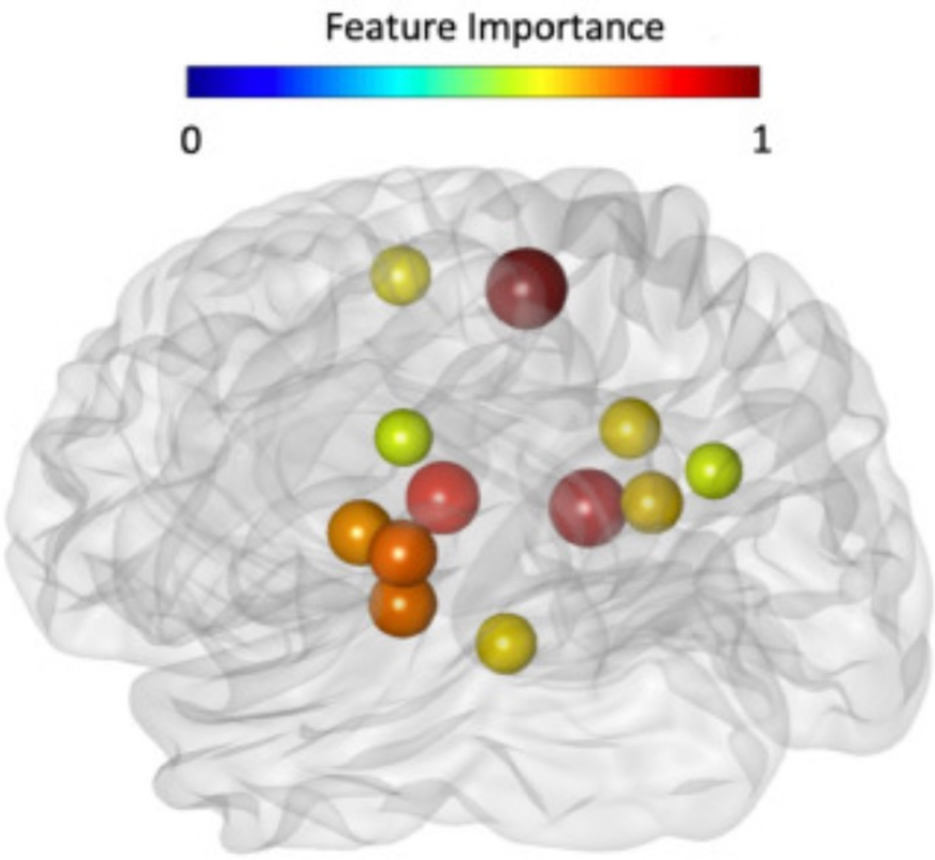
The "brain age gap" estimation method (*BrainAGE*) uses MRI images to quantify accelerations or decelerations of individual brain aging. The rationale:

- Establish **reference aging curves** for *healthy brain* maturation during childhood into young adulthood and for *healthy brain* aging during adulthood into senescence;
- **Quantify the deviation** of a new (test) subject from the reference curve, using the ***BrainAGE gap*** index*;*
- Capture multidimensional maturation/aging **patterns of this index** as potential biomarkers of **neurodevelopmental** disorders (e.g. autism) or **neurodegenerative** diseases (e.g. Alzheimer's).



The Brain Age Gap as a biomarker

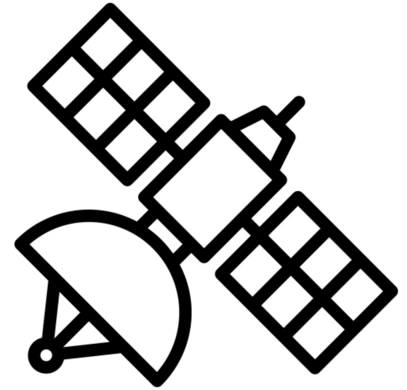
COMPUTATIONAL NEUROSCIENCE



Even in this case, it is important to establish which are the brain areas most involved in the process to «explain» how aging works.

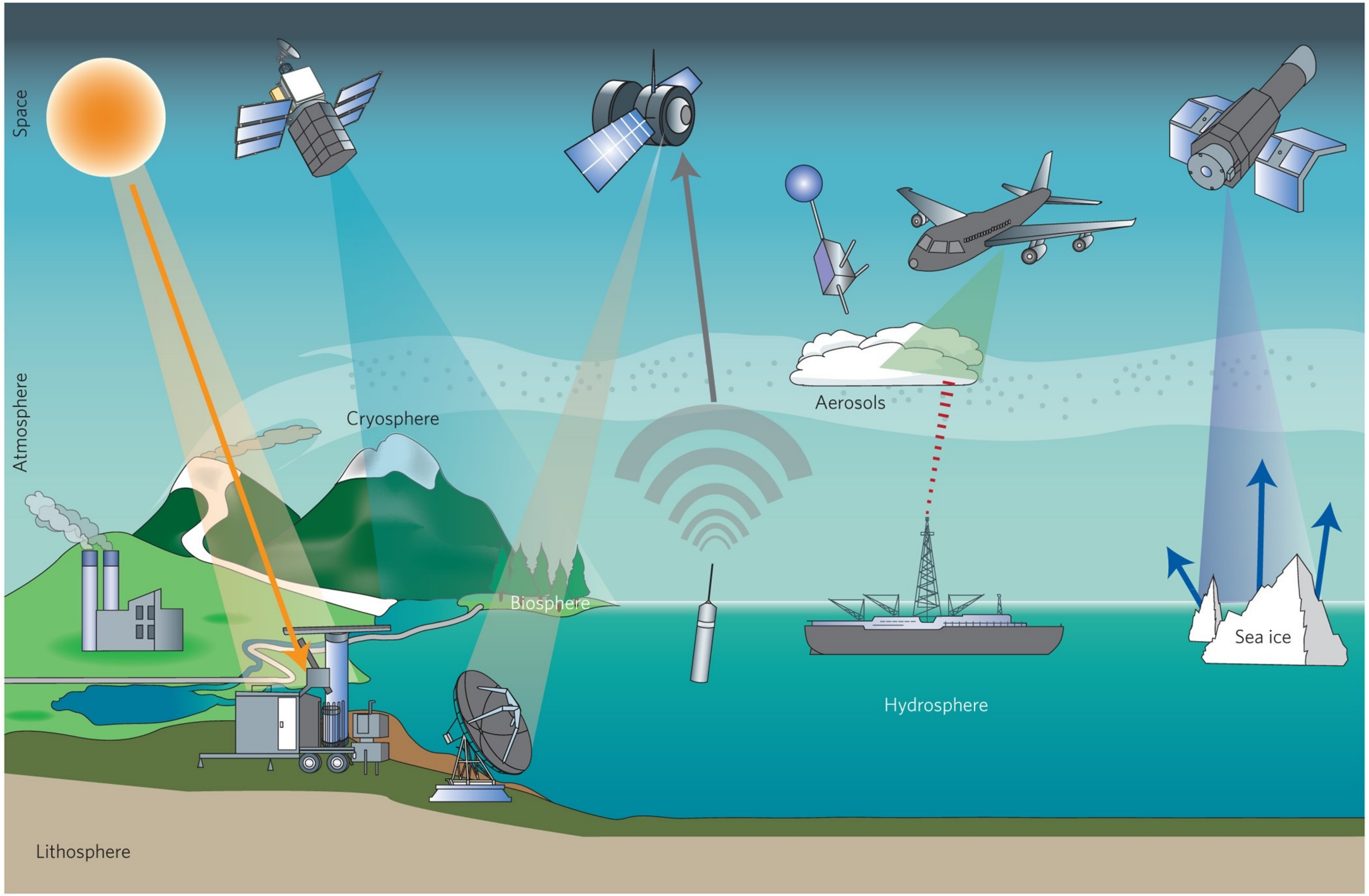
Among them… try to guess!.

SATELLITE DATA



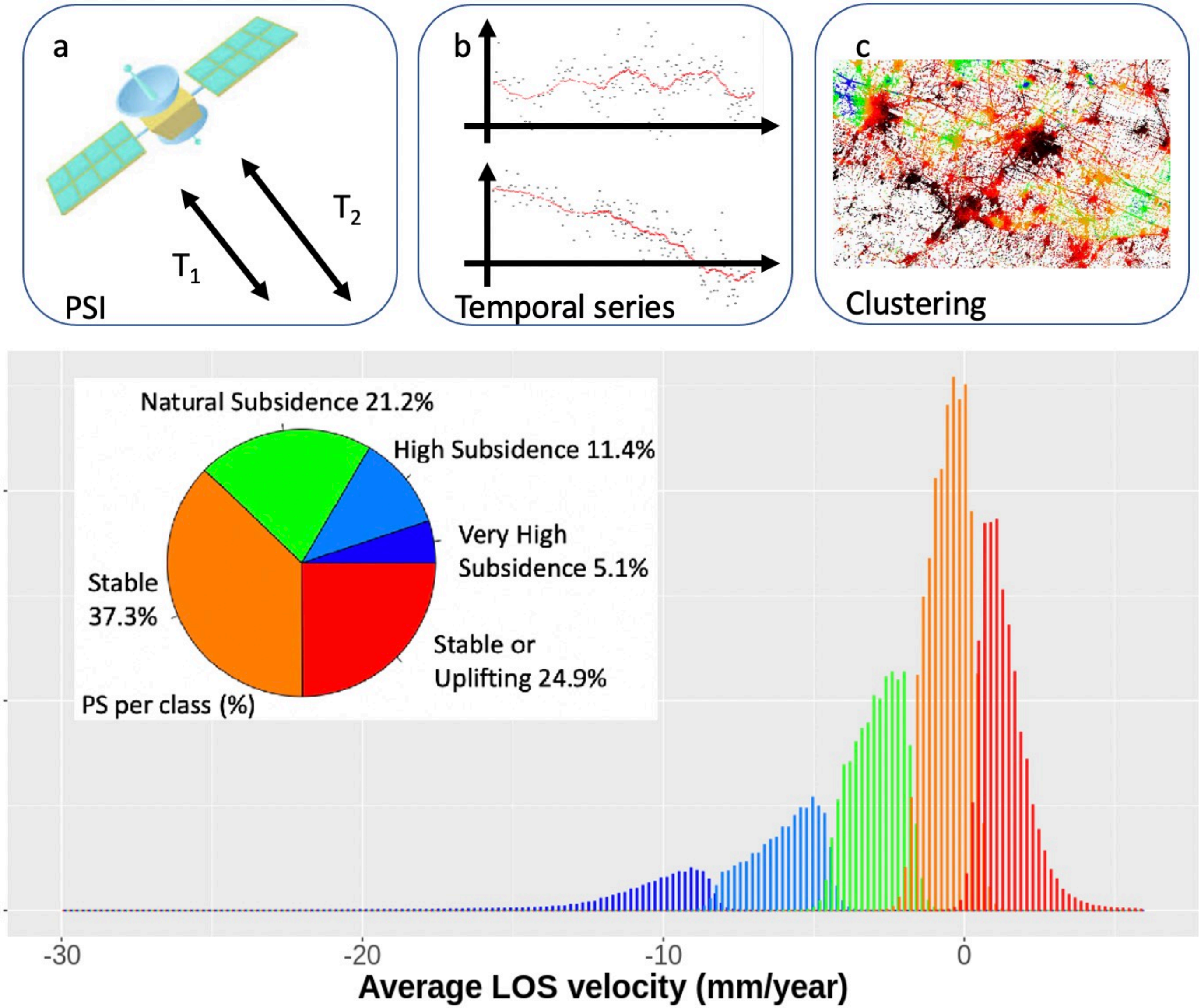
SATELLITE DATA

Earth Observation and Climate change



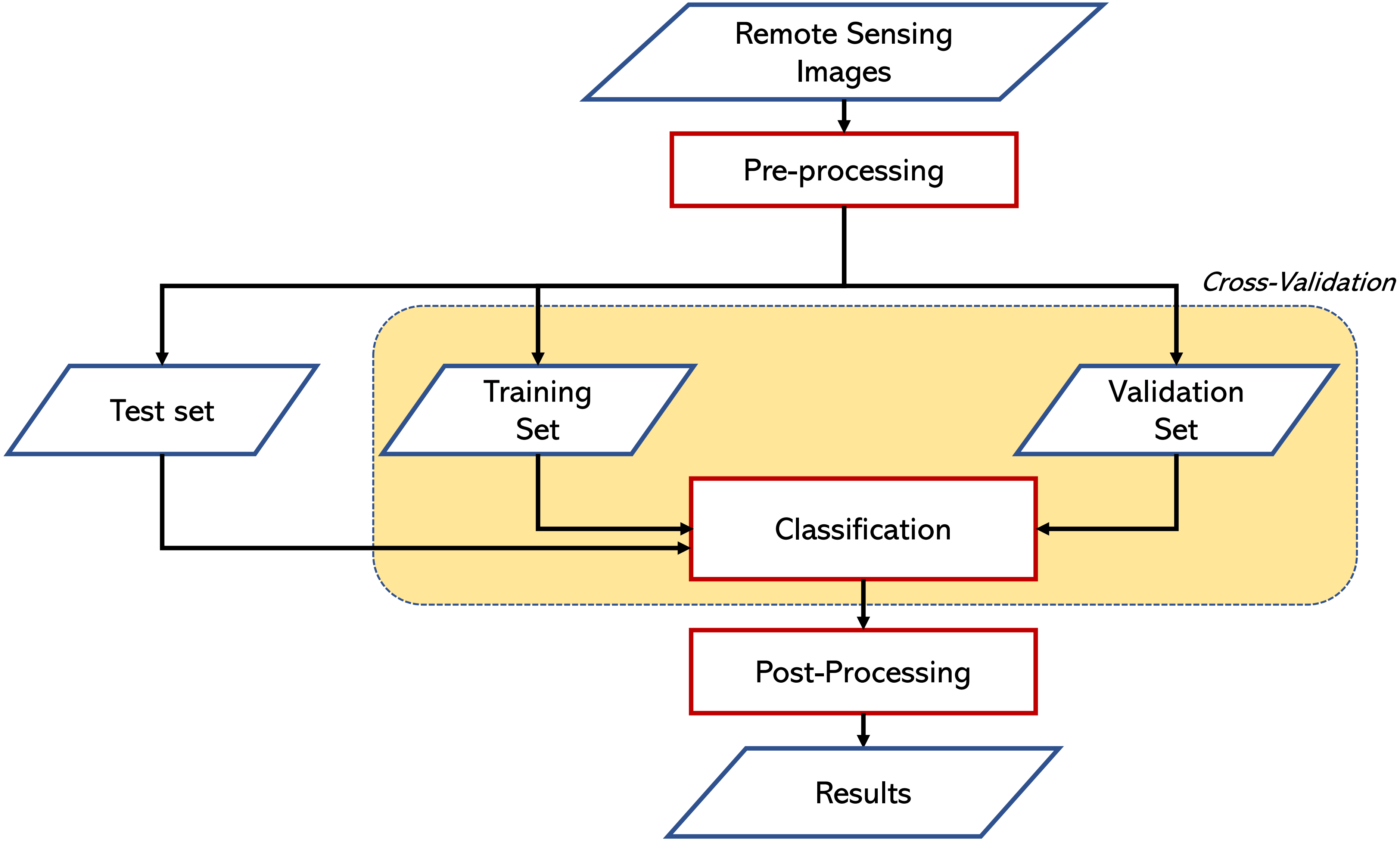
Structural Health Monitoring

SATELLITE DATA



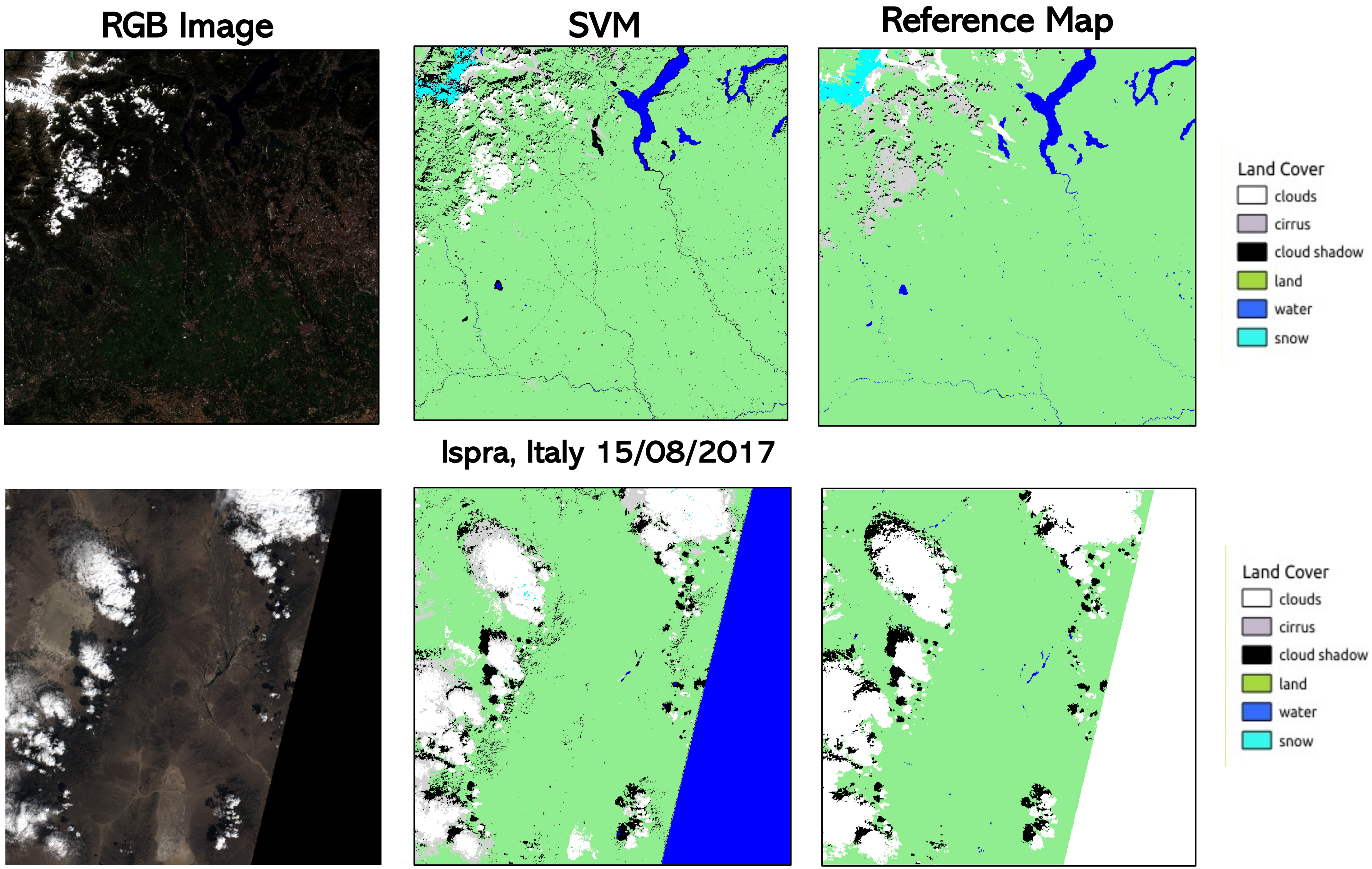
Workflow

SATELLITE DATA



SATELLITE DATA

Some examples: Cloud detection



RGB Image

SVM

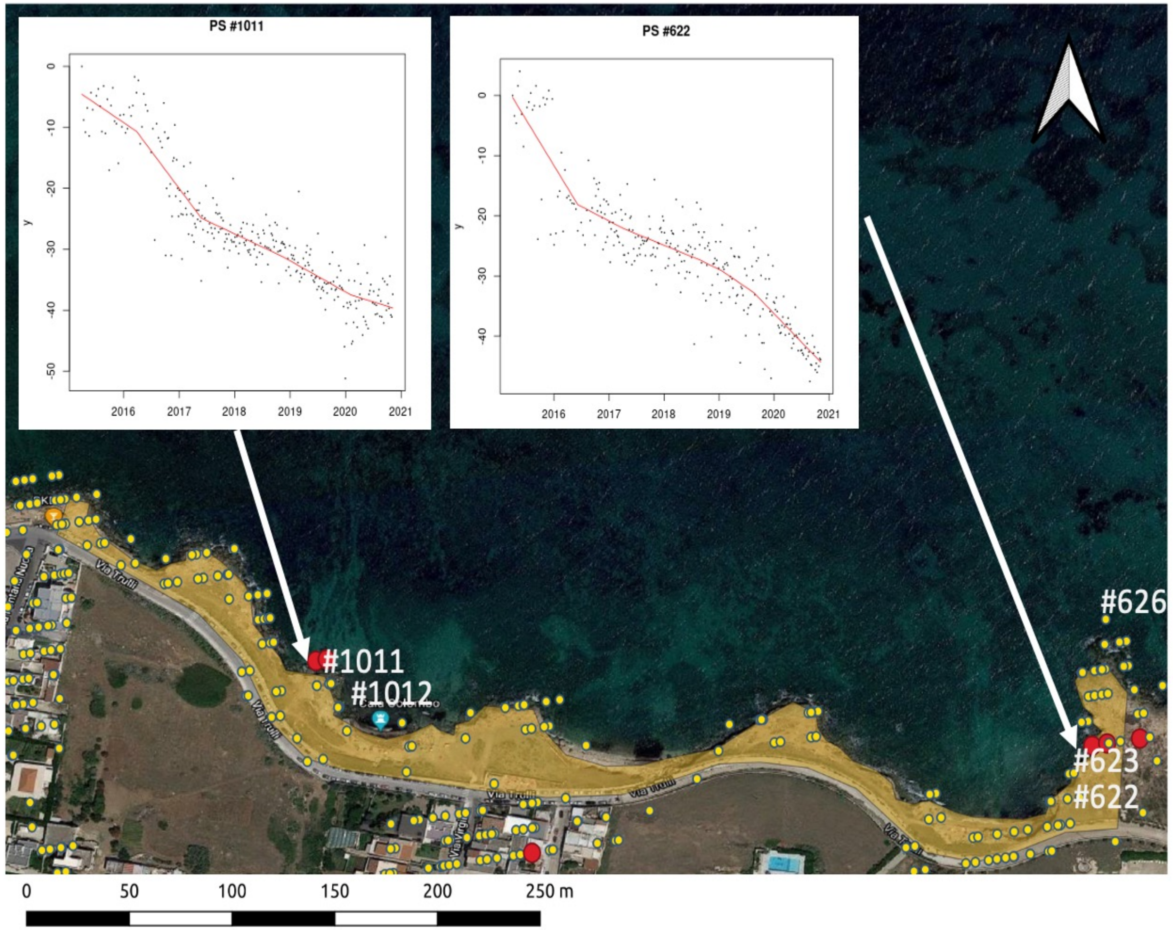
Reference Map

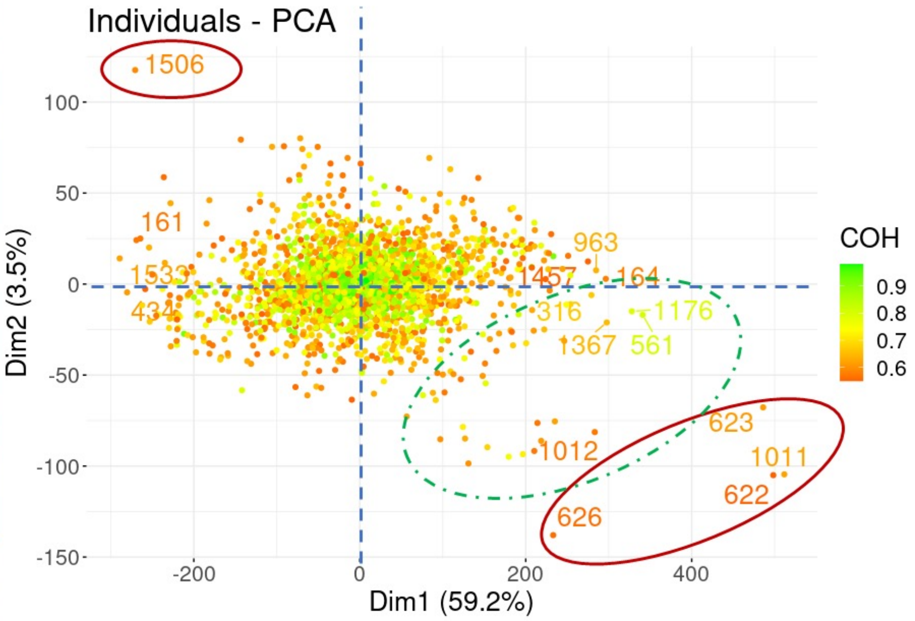
Ispra, Italy 15/08/2017

Railroad Valley, Nevada, 27/08/2017

SATELLITE DATA

Some examples: Monitoring of Coastal Cliffs

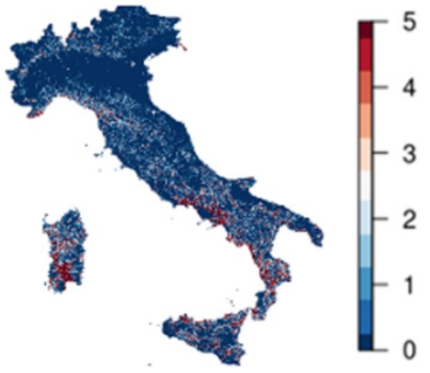


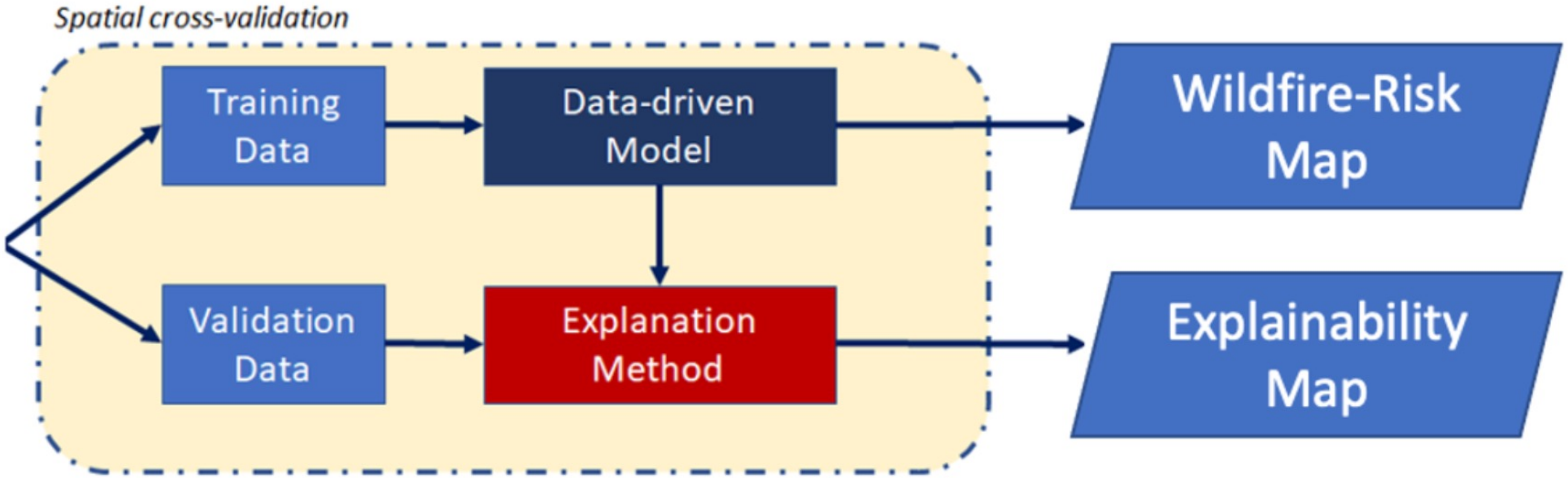


Our analysis found only one meaningful cluster of deformation behavior and some **anomalous** points related to cliff instabilities

Some examples: Wildfire detection

SATELLITE DATA





Spatial cross-validation

Training Data

Data-driven Model

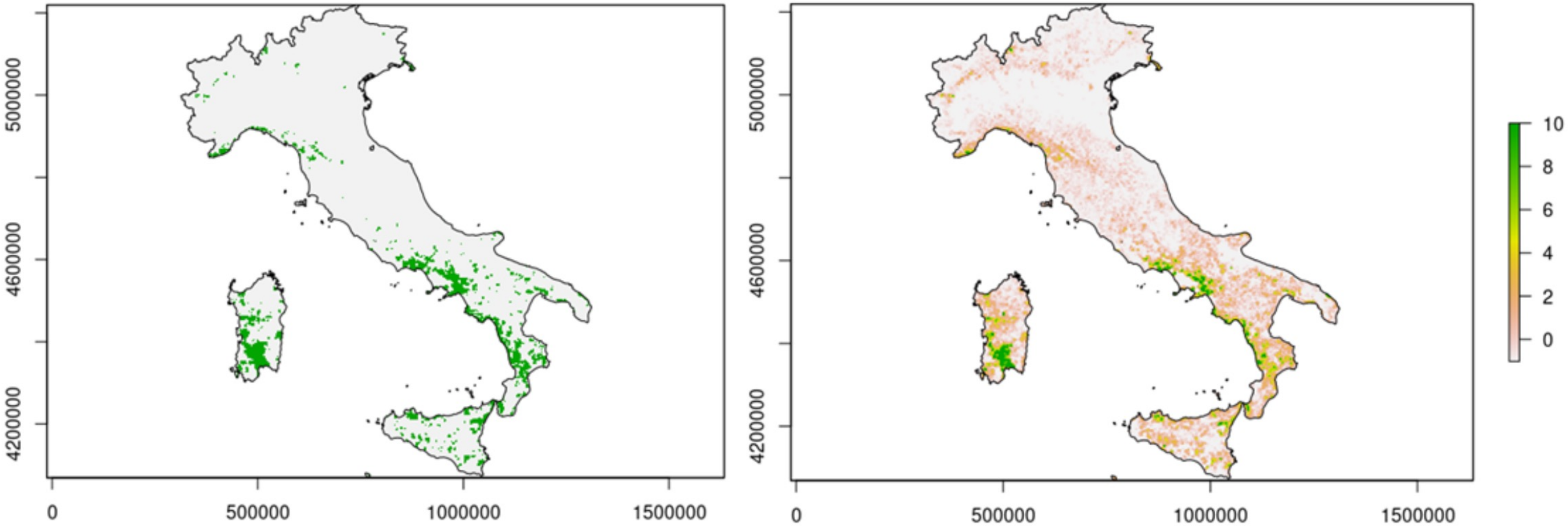
Validation Data

Explanation Method

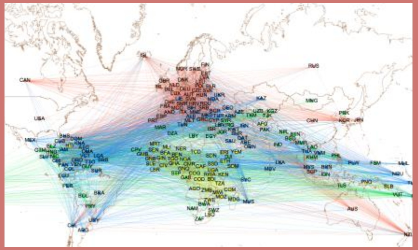
Wildfire-Risk Map

Explainability Map

Specific locations show **anomalous** patterns suggesting malicious or negligent causes.

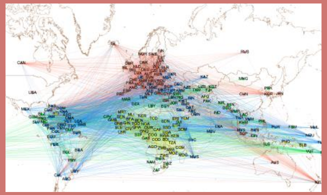


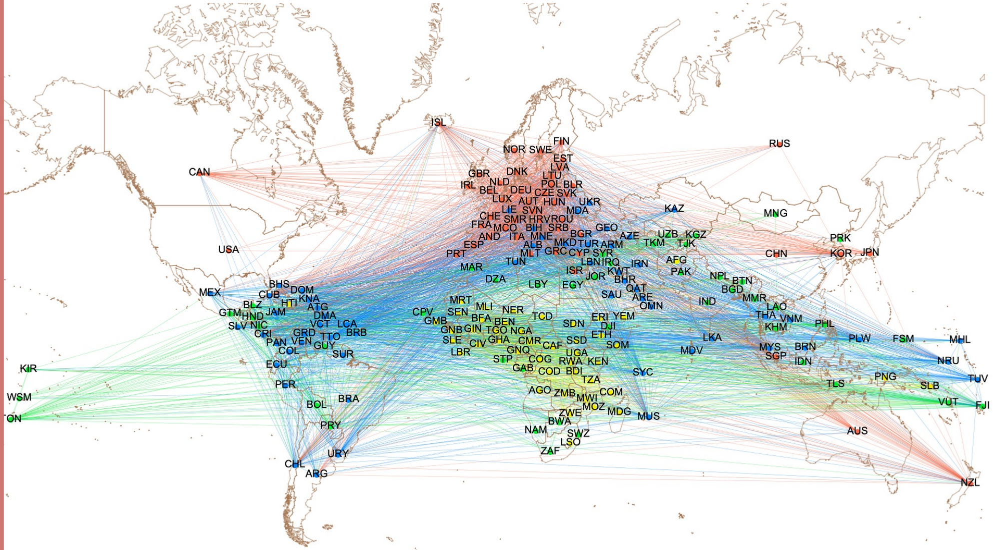
SOCIAL DATA



Complex Networks and World Rankings

SOCIAL DATA

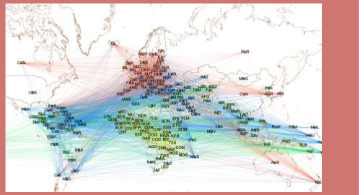




- **International Rankings** provide a concise evaluation of performances in a specific domain and guidelines for decision makers;

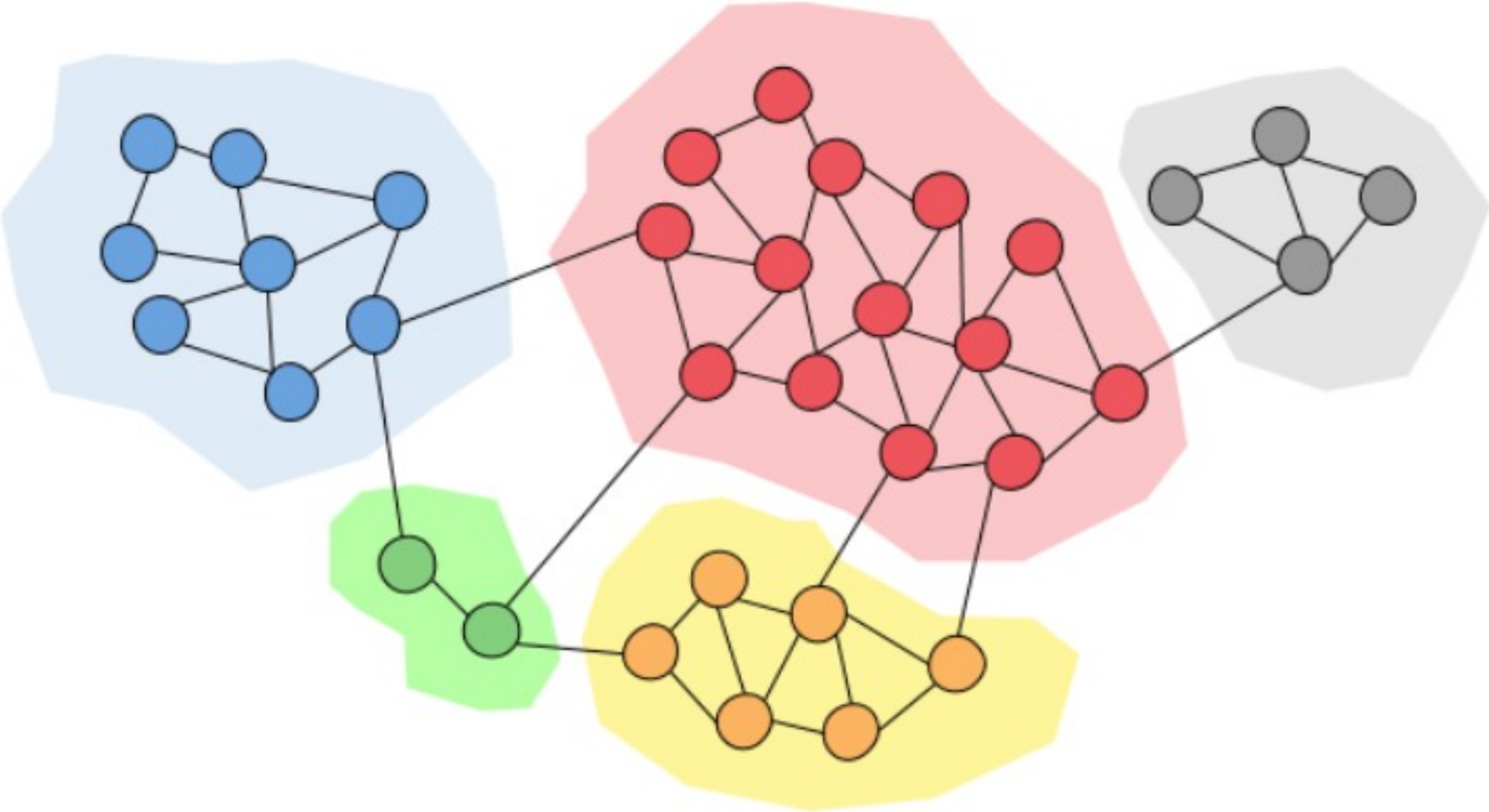
- However, they do not consider the differences between participants. Are they **fair**?

SOCIAL DATA

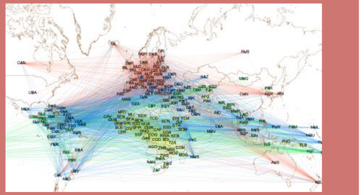


Community detection

Networks provide a suitable tool to evaluate similarities between the elements of a system. That is **community detection**.

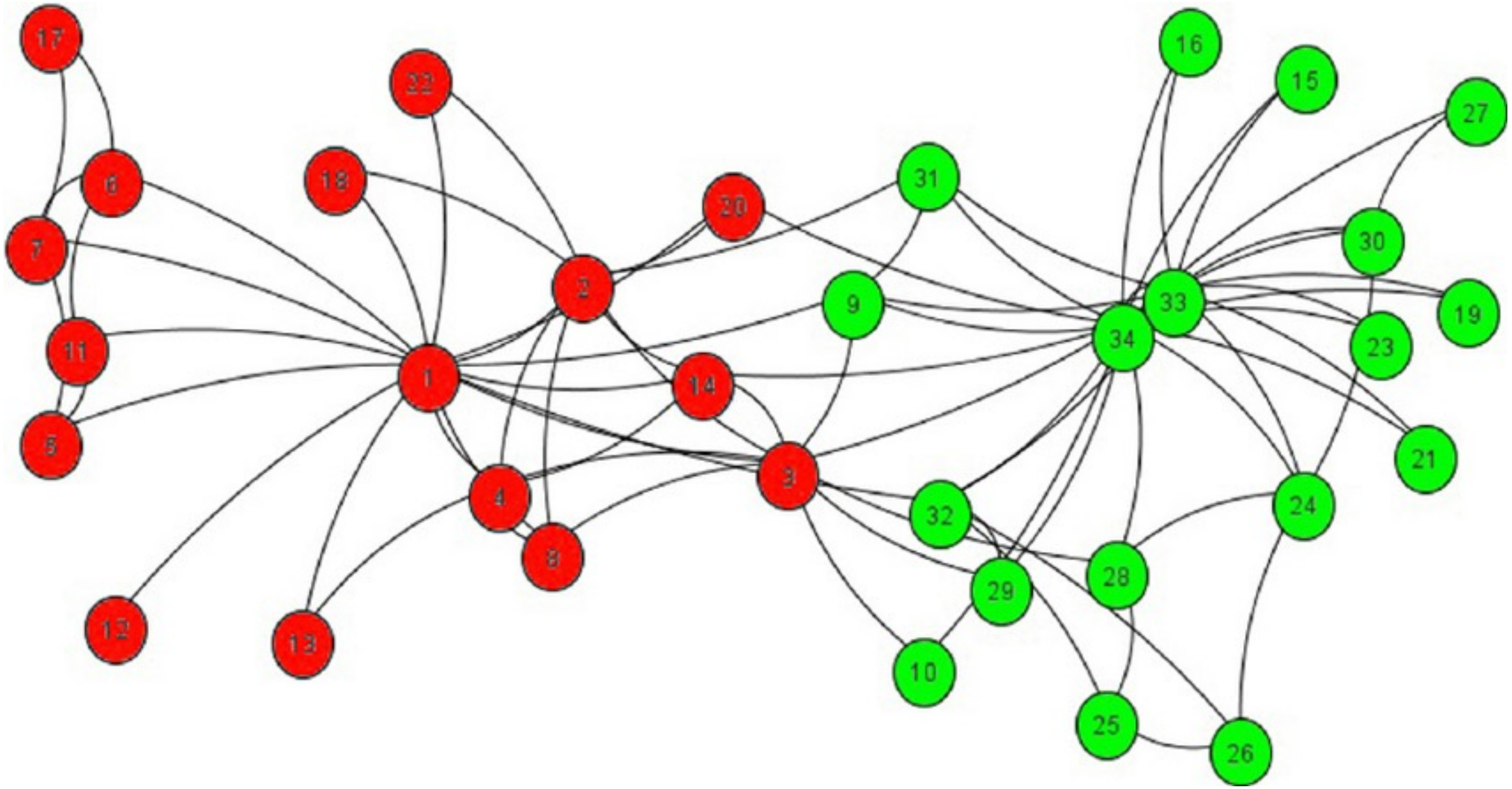


SOCIAL DATA



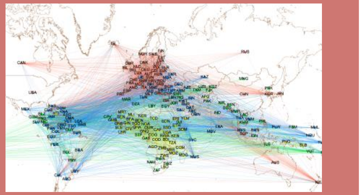
Zachary’s Karate club

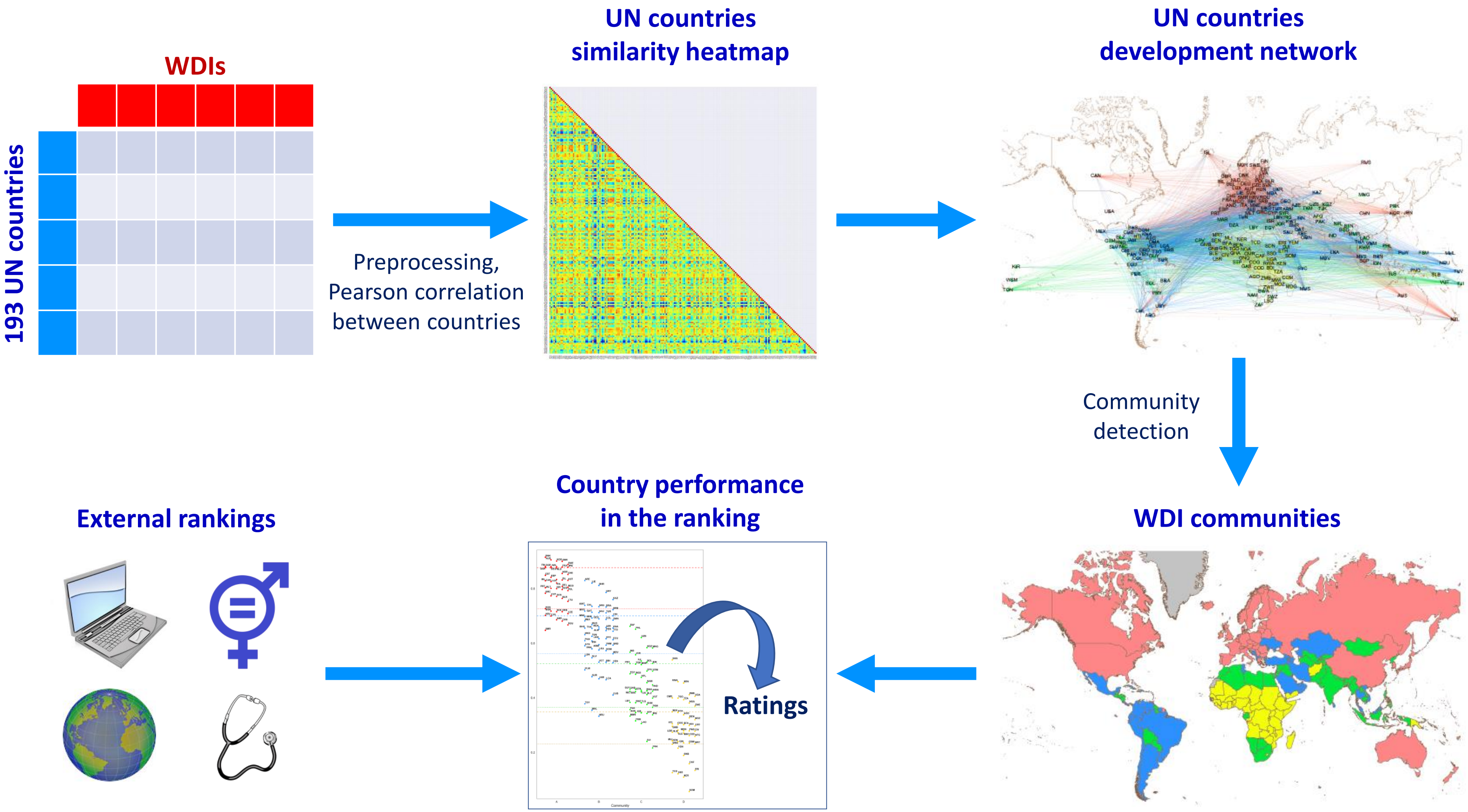
Zachary's karate club is a social network of a university karate club with 34 members. After a dispute the club split into two.



Workflow

SOCIAL DATA





WDIs

193 UN countries

Preprocessing, Pearson correlation between countries

UN countries similarity heatmap

UN countries development network

Community detection

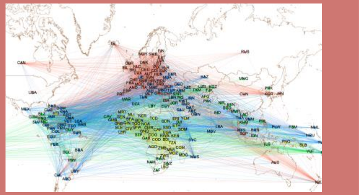
WDI communities

Country performance in the ranking

Ratings

External rankings

SOCIAL DATA



Rankings

E-Government Development Index (EGDI) 2020

Credits: United Nations Department of Economic and Social Affairs (UN-DESA)

Global Gender Gap Index (GGGI) 2020

Credits: World Economic Forum (WEF)

SDG Global Index Score 2019

SUSTAINABLE DEVELOPMENT GOALS

Credits: Bertelsmann Stiftung and Sustainable Development Solutions Network (SDSN)

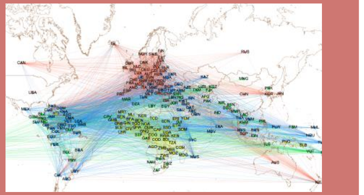
Environmental Performance Index (EPI) 2018

Credits: Yale Center for Environmental Law and Policy (YCELP), Yale Data-Driven Environmental Solutions Group, Center for International Earth Science Information Network (CIESIN), World Economic Forum (WEF)

Healthcare Access and Quality Index (HAQI) 2016

Credits: Global Burden of Disease (GBD) 2016 Healthcare Access and Quality collaboration

SOCIAL DATA



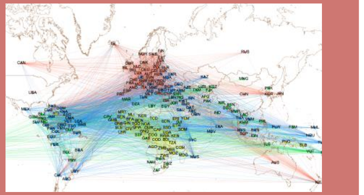
”New” Rankings

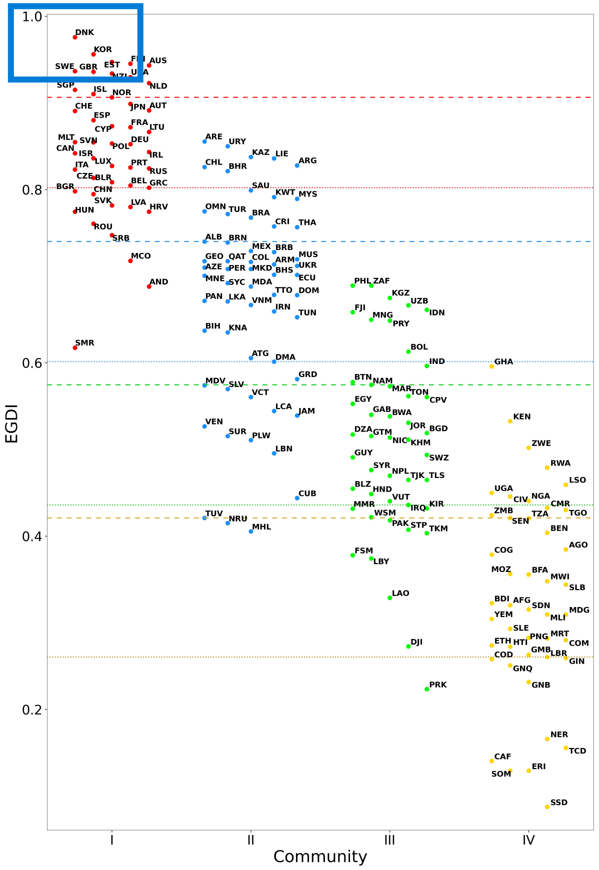
New rating system: quantify the discrepancy btw a country’s ranking and its expected ranking based on the belonging community

BENCHMARK COUNTRIES	*TOP-OF-THE-CLASS* COUNTRIES	*ROOM-FOR-IMPROVEMENT* COUNTRIES	*TRAILING* COUNTRIES
• Belong to community I • Rank in the 75th percentile of their community	• Rank in the 75th percentile of their community • Rank in the 25th percentile of at least one less developed community	• Rank in the 25th percentile of their community • Rank in the 75th percentile of at least one less developed community	• Belong to community IV • Rank in the 25th percentile of their community

E-Government Development Index 2020

SOCIAL DATA





BENCHMARK countries

Denmark

Rep. Korea

Estonia

Finland

Australia

Sweden

United Kingdom

New Zealand

United States

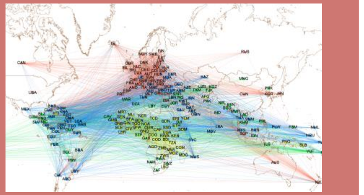
Netherlands

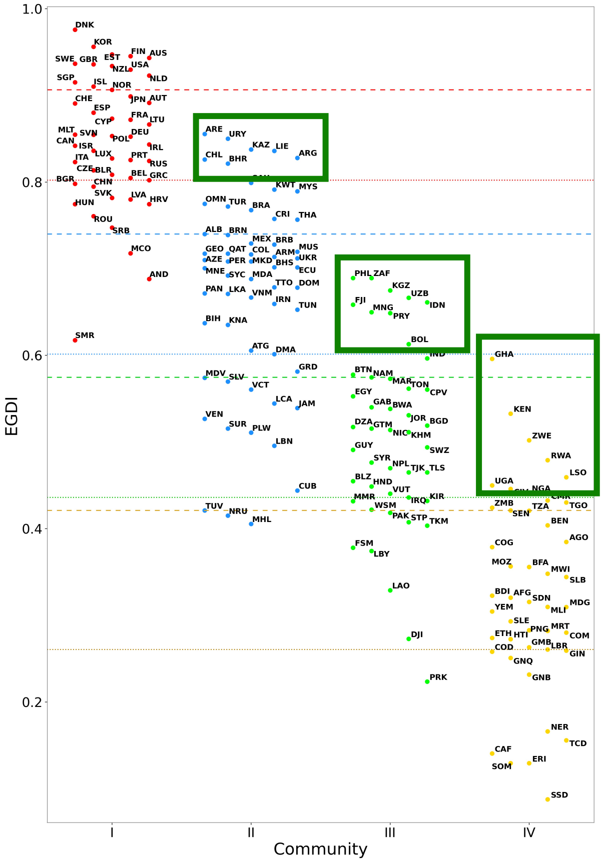
Singapore

Iceland

E-Government Development Index 2020

SOCIAL DATA





TOP OF THE CLASS COUNTRIES

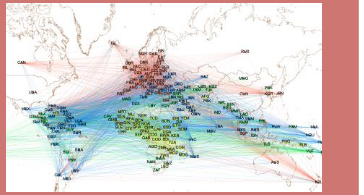
Community II: United Arab Emirates, Uruguay, Kazakhstan, Liechtenstein, Argentina, Chile, Bahrain

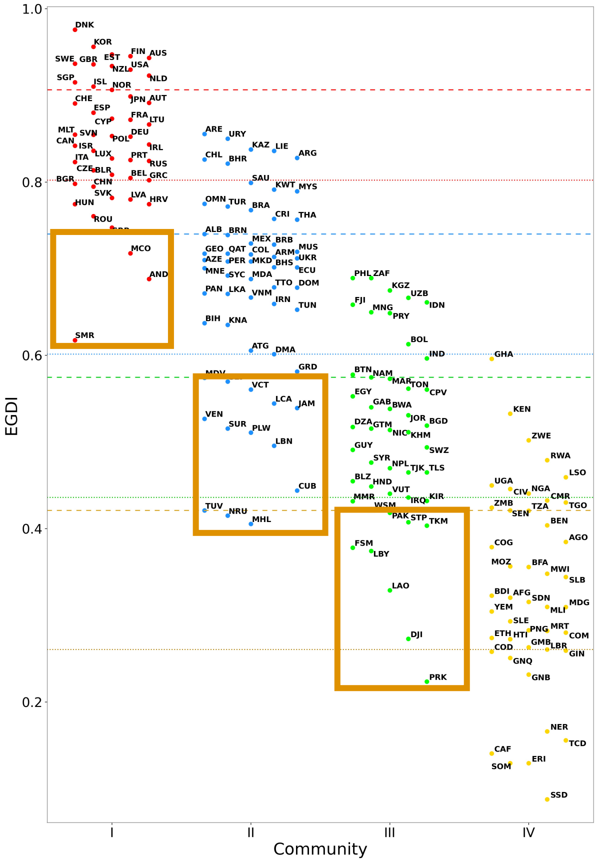
Community III: Philippines, South Africa, Kyrgyz Republic, Uzbekistan, Indonesia, Fiji, Mongolia, Paraguay, Bolivia

Community IV: Ghana, Kenya, Zimbabwe, Ruanda, Lesotho, Uganda, Cote d’Ivoire, Nigeria.

E-Government Development Index 2020

SOCIAL DATA





ROOM FOR IMPROVEMENT COUNTRIES

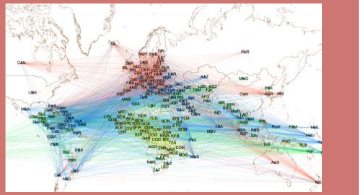
Community I: Monaco, Andorra, San Marino

Community II: Maldives, El Salvador, Saint Vincent and the Grenadines, St. Lucia, Jamaica, Venezuela, Suriname, Palau, Libano, Cuba, Tuvalu (**), Nauru (**), Marshall Islands (**)

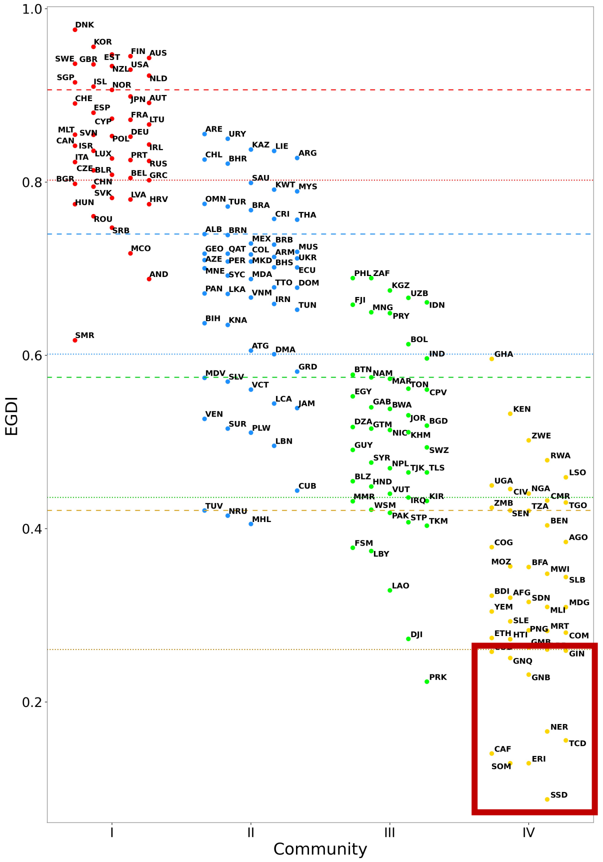
Community III: Pakistan, São Tomé and Príncipe, Turkmenistan, Fed. Sts. Micronesia, Libya, Lao PDR, Djibouti, Dem. People's Rep. Korea

(**) ranking in the 75th percentile of two less developed communities

SOCIAL DATA



E-Government Development Index 2020



TRAILING COUNTRIES

Guinea

Dem. Rep. Congo

Equatorial Guinea

Guinea-Bissau

Niger

Chad

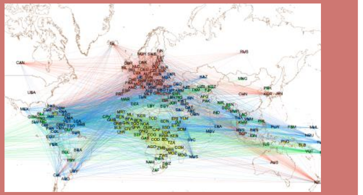
Central African Republic

Somalia

Eritrea

South Sudan

SOCIAL DATA



Take home messages

We developed a rigorous, transparent and reproducible pipeline that provides:

- the **unsupervised** identification of a **robust community structure** in the WDI complex network, that interpolates between the established UN and World Bank groupings;
- a targeted, **fair** and meaningful criterion to detect **country similarities**;
- a straightforward and validated method to **reinterpret rankings**, that evaluates country performances based on their **development** level;
- identification of both **leading countries**, that reach higher positions than expected from their general development levels and **trailing countries** that have worse-than-expected performances.

SOCIAL DATA

Complex Networks and Startup success factors

- **Business success:** Collected funds are an immediate measure of economic power, but do not represent the whole picture.
- There are only a **few quantitative analyses of the startup ecosystem**
- **Network analysis** can give quantitative insights into complex economic systems, can measure the importance of individual elements.
- **Network metrics** can predict the amount of funds collected by a startup??

Amoroso, N., Bellantuono, L., Monaco, A., De Nicolò, F., Somma E., Bellotti R. Economic Interplay Forecasting Business Success. Complexity 2021

Startup and funding data

SOCIAL DATA

CrunchBase

- One of the **most complete** and widely known datasets about startups and their funding data and history.
- Includes data from 550.000 firms in 160 countries (up to 2017).

SOCIAL DATA

Modeling funding interactions

- **Follow the money**: Each element is a node. Two nodes are linked if they have a funding relation??. Quando un nodo finanzia l’altro

- Three economically-interpretable network metrics: **indegree, outdegree, betweenness**

Indegree

Investor attractiveness

Outdegree

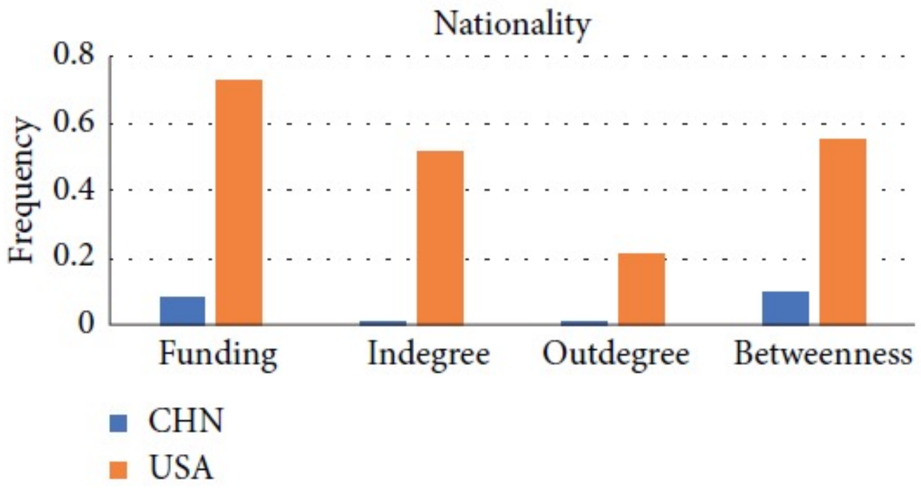
Financing Power

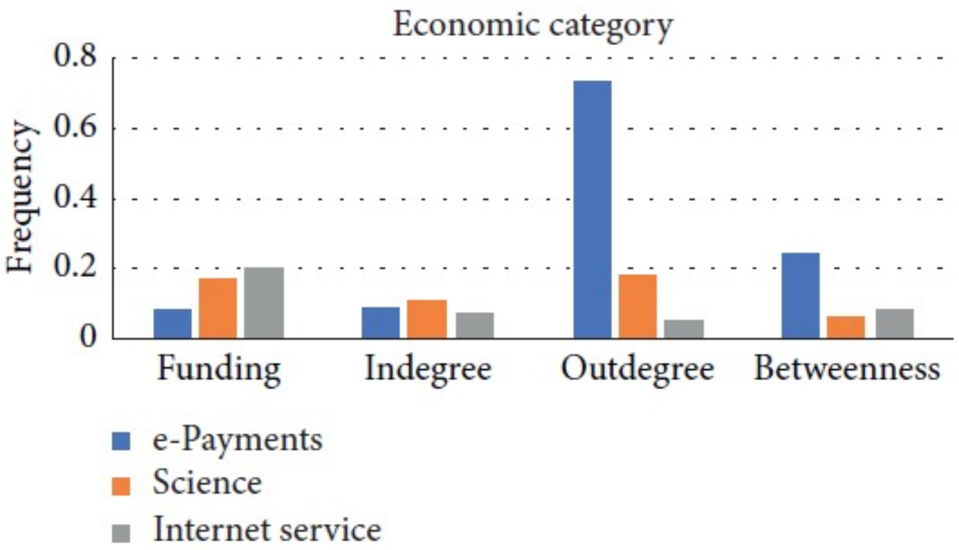
Betweenness

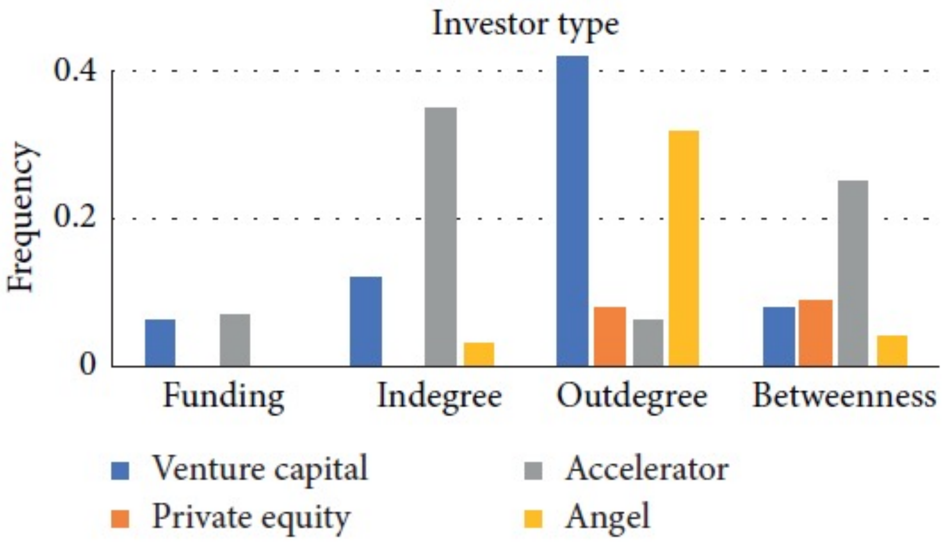
Capital Conveyance

Identifying strategic elements

SOCIAL DATA





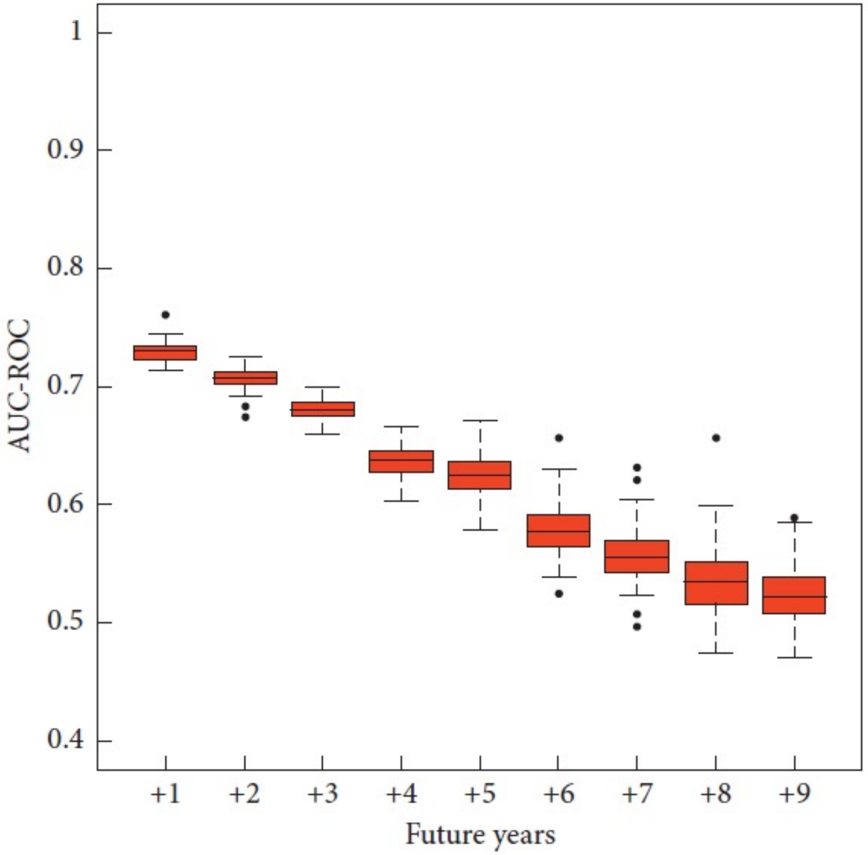


Network metrics give insights into how strategic individual elements are.

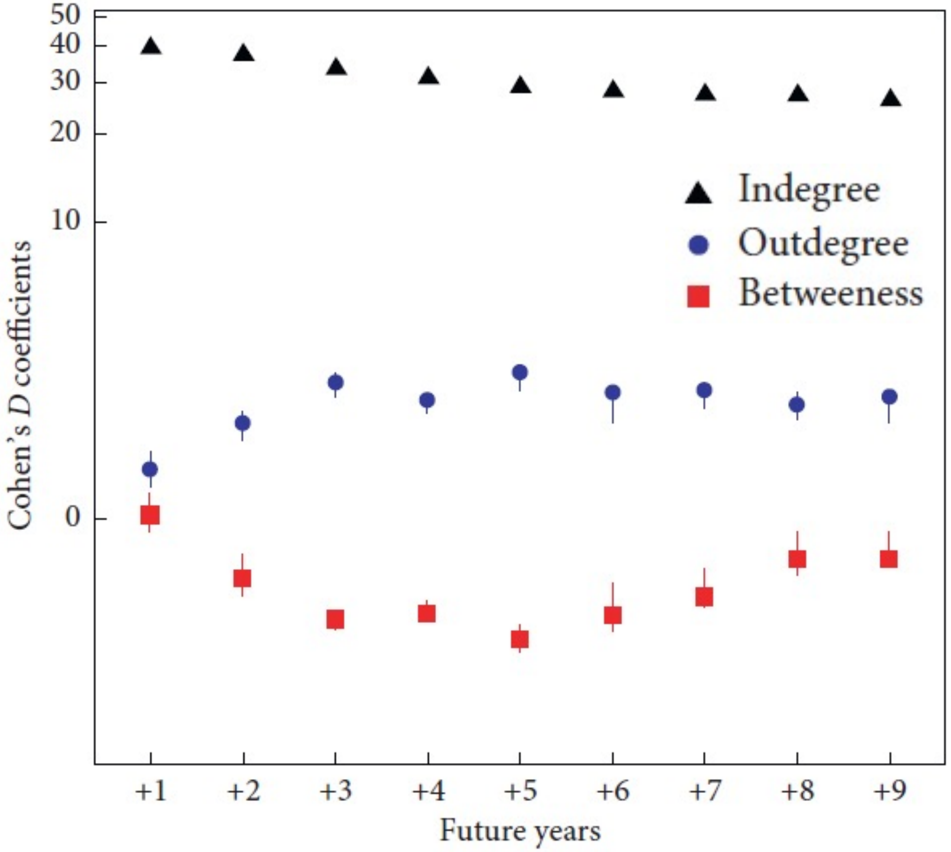
These findings could not have been retrieved if funds only had been considered.

Forecasting future fundings using network metrics

SOCIAL DATA



Network metrics are able to reliably forecast future fundings up to 5 years in the future



Indegree (investor attractiveness) is the most important feature to forecast future fundings

AI tools for tourism intelligence: the C-BAS Project

SOCIAL DATA

- Tourists influence each other through the reviews of their experiences

- **The C-BAS Project focuses on the Apulian tourist offer**

- Apulian tourism has witnessed an impressive growth in the last decade. What are the main drivers of this phenomenon?

- **Natural Language Processing (NLP) and AI models** are fundamental for the analysis of Apulian tourist offer through tourists' reviews.

network.contacts Passion to Care

3rdPlace user & customer AI

btsconsulting

SER&P Software Engineering Research & Practices

FLAG

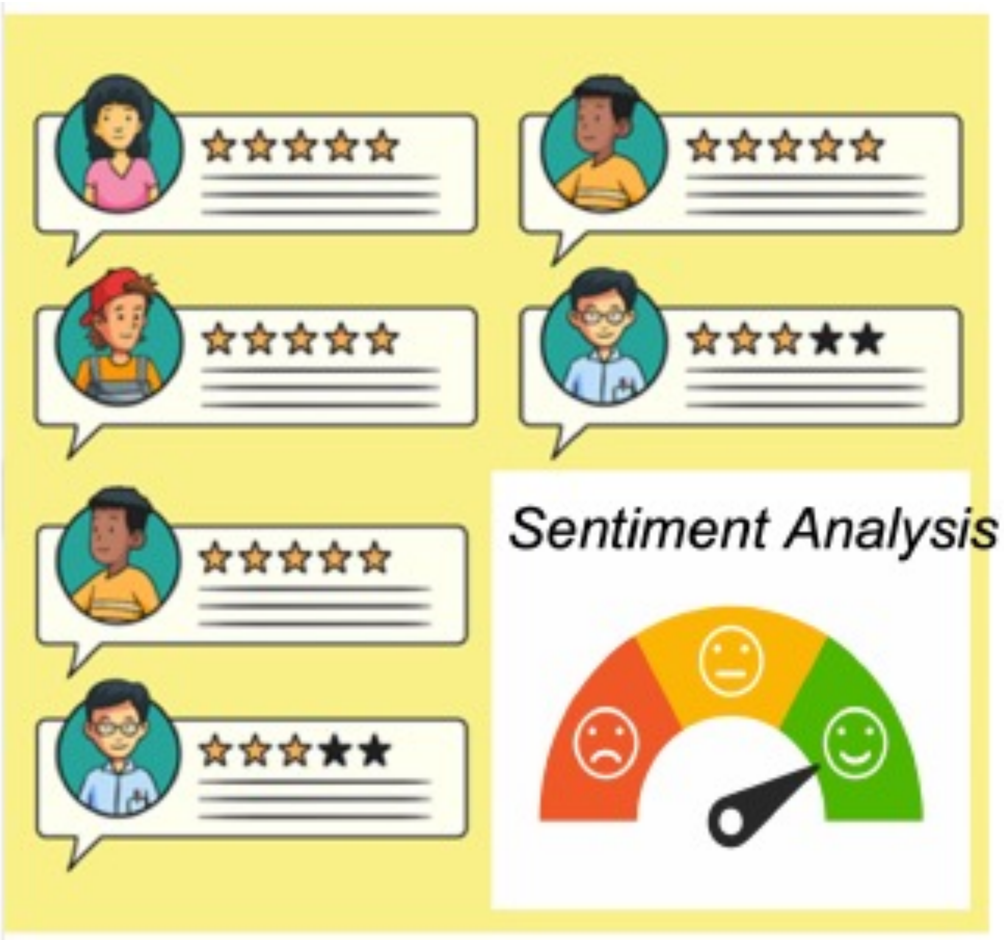
SOCIAL DATA

Tripadvisor reviews for Apulian Tourism

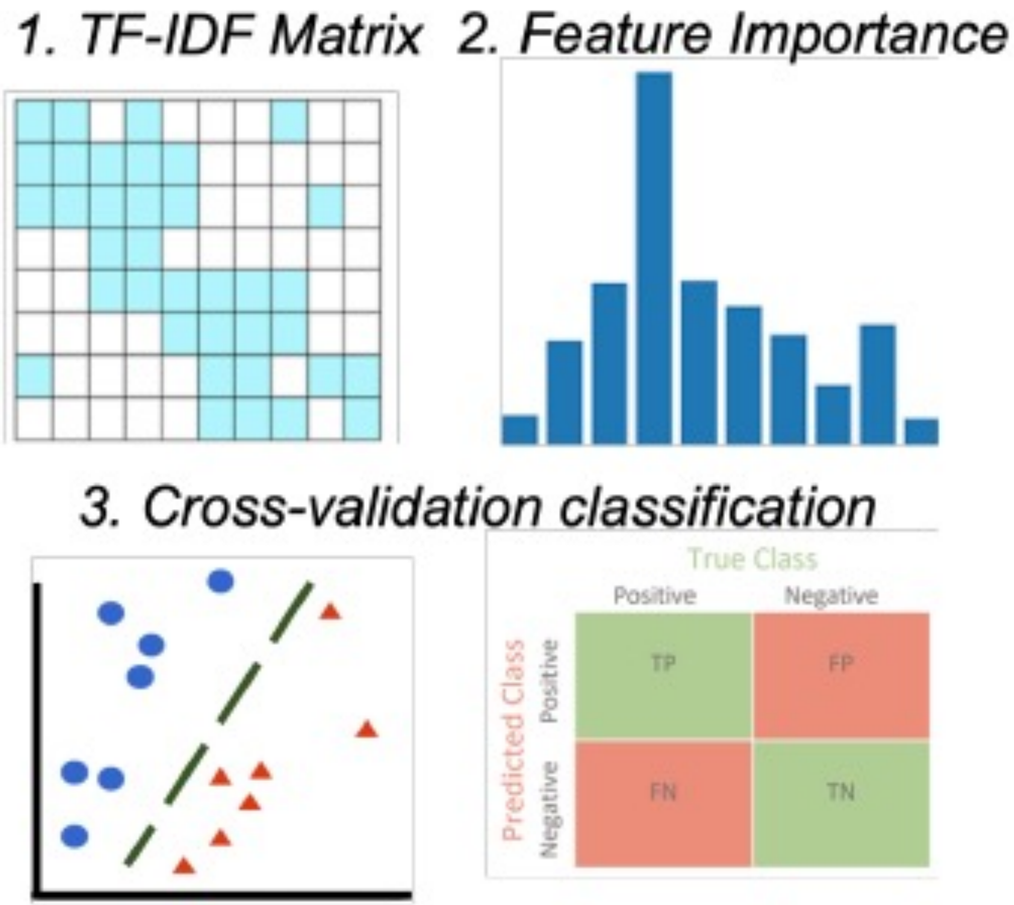
- 13.400 reviews; 974 Apulian tourist facilities; from May 2004 to June 2020.
- Data: Review text and rating (from 1 to 5)
- Positive reviews: rating ≥ 3; Negative reviews: rating < 3

Identifying the most important words through AI

SOCIAL DATA



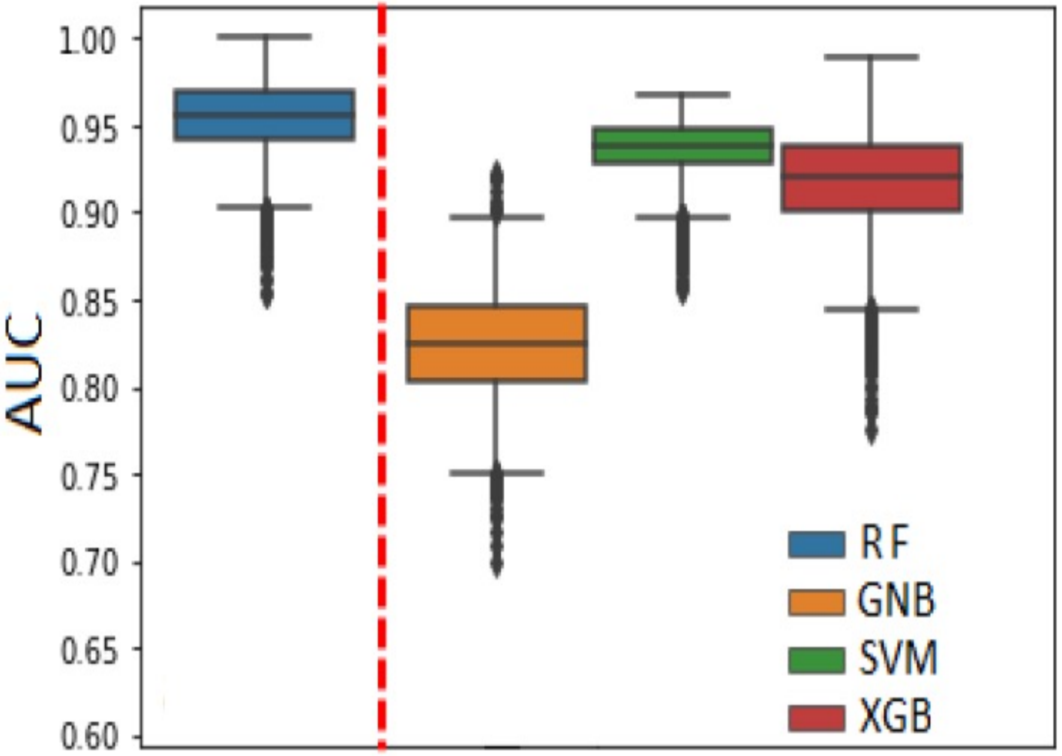
Detecting mis-matches between text content and rating



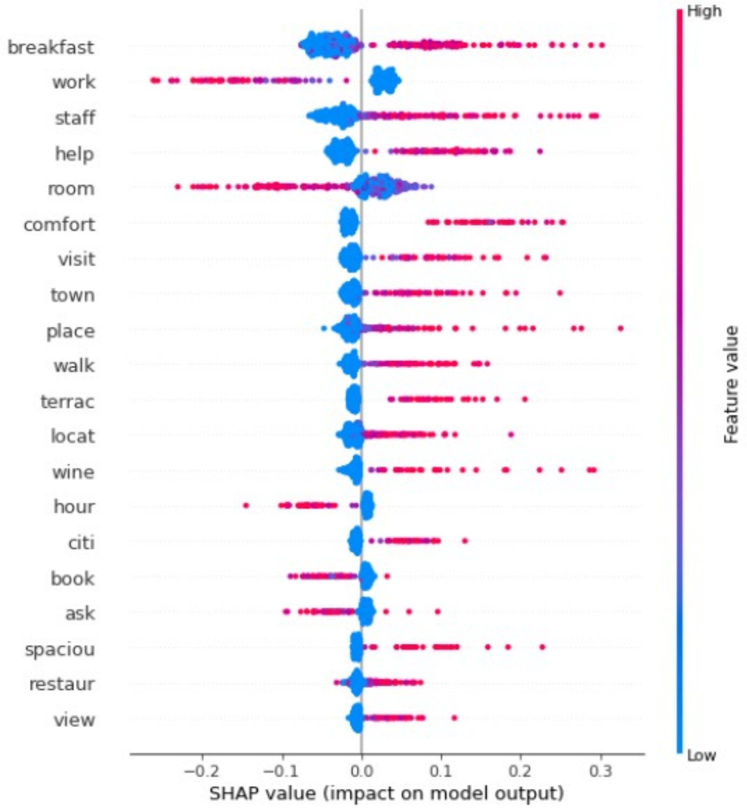
Classification of reviews rating using word frequencies

Explaining how the AI model works: Shapley values

SOCIAL DATA



Choosing the best classification model



Determining the most important words for classification using Shapley values

SOCIAL DATA

Future Perspectives

- Using n-grams for reviews’ classification.
- Considering reviewers’ attributes (e.g.: nationality, trip-type).
- Adding facilities’ attributes and observing differences in classification and word-importance.
- Developing a network where nodes are reviewers and links are determined for example by the semantic similarity of the review to identify communities of reviewers and eventually to model interactions among them.

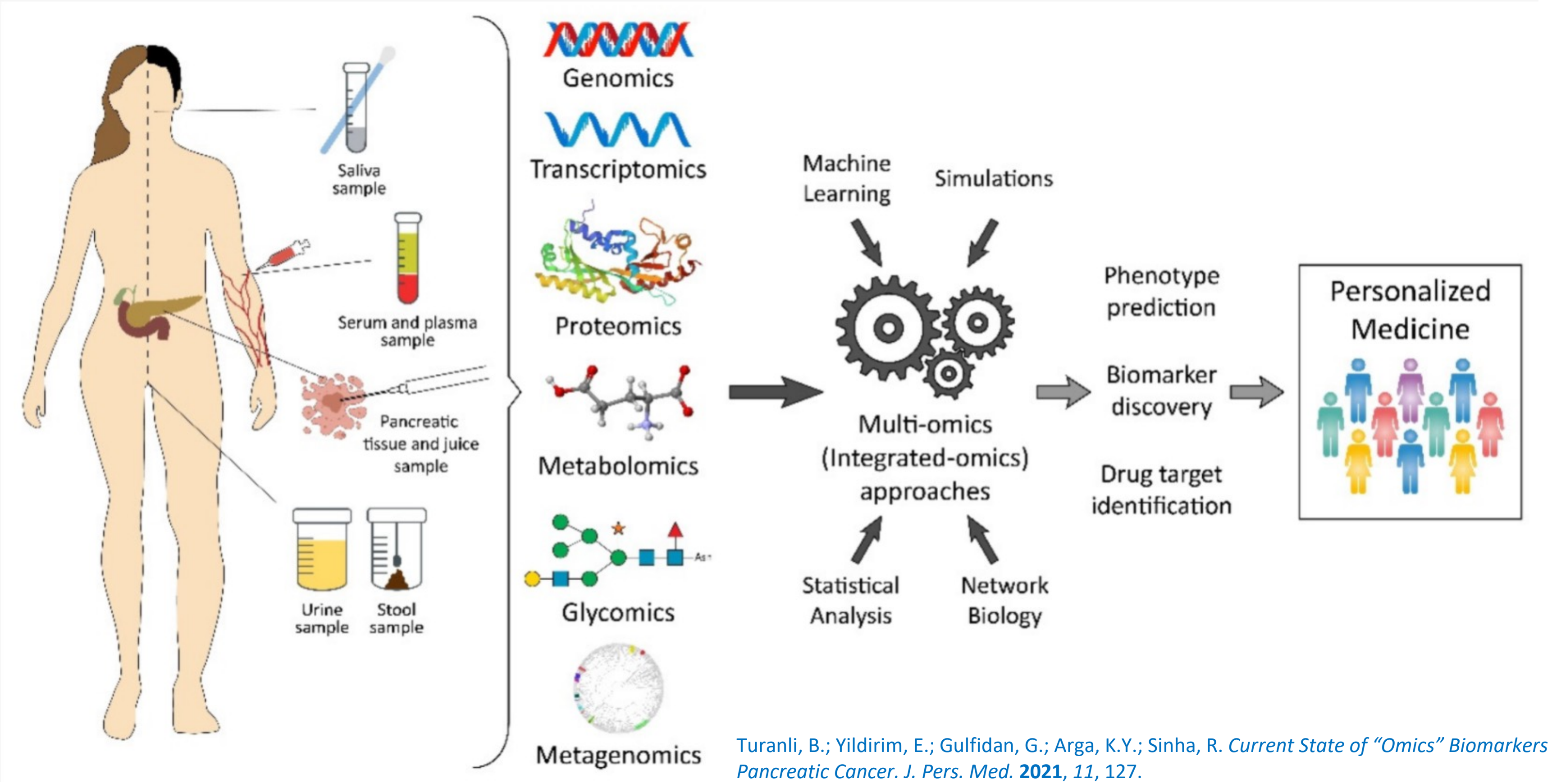
GENOMICS

GENOMICS

Managing, Analysing, and Integrating Big Data in Medical Bioinformatics

New high throughput technologies in Biology → **Big Data** → the new **OMIC SCIENCES** → **Personalized Medicine**

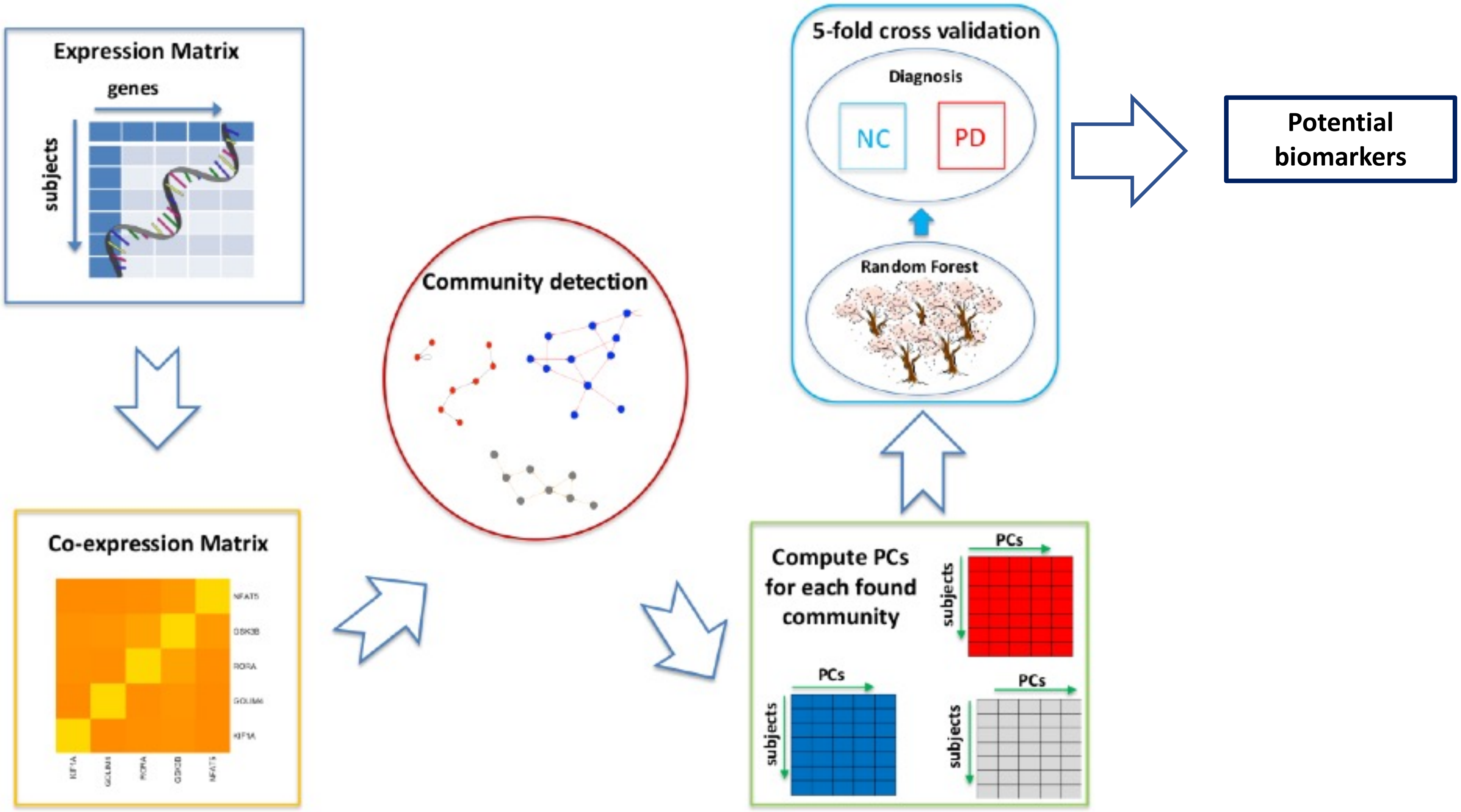
1 individual: 6 Gbases in the genome,
10^{13} cells
10^{14} cells in the microbiome



Turanli, B.; Yildirim, E.; Gulfidan, G.; Arga, K.Y.; Sinha, R. *Current State of "Omics" Biomarkers in Pancreatic Cancer. J. Pers. Med.* **2021**, *11*, 127.

GENOMICS

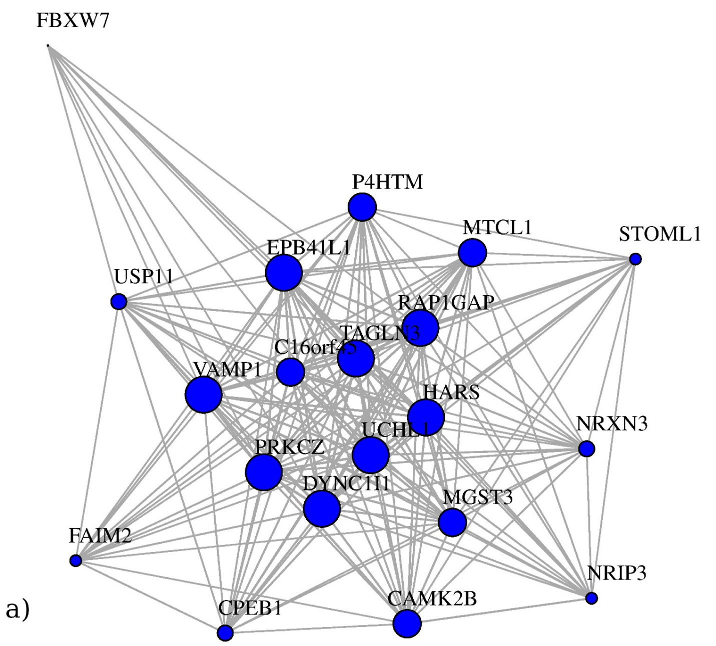
An information entropy approach to identify potential gene biomarkers for Parkinson’s Disease

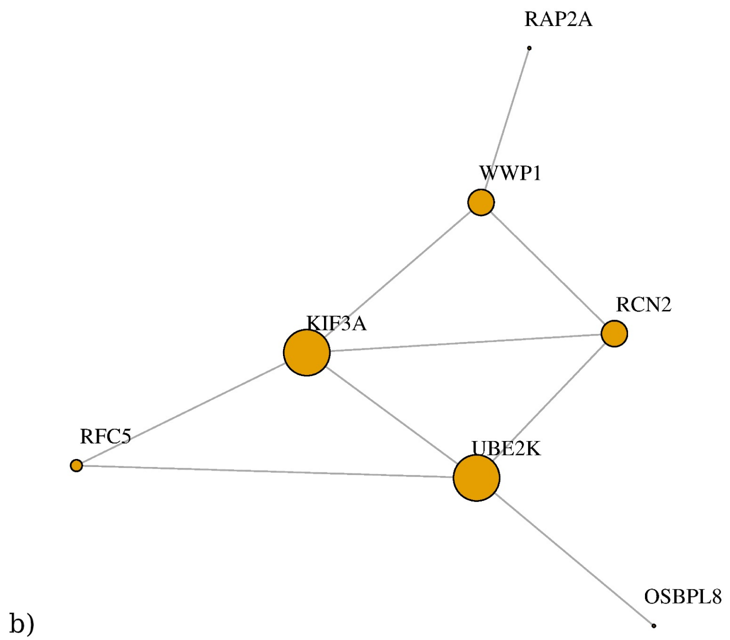


Monaco A, Pantaleo E, Amoroso N, Bellantuono L, Lombardi A, Tateo A, Tangaro S, Bellotti R. Identifying potential gene biomarkers for Parkinson's disease through an information entropy based approach. Phys Biol. 2020 Dec 1;18(1):016003

Two gene communities discriminate PD vs Healthy Control

GENOMICS





Monaco A, Pantaleo E, Amoroso N, Bellantuono L, Lombardi A, Tateo A, Tangaro S, Bellotti R. Identifying potential gene biomarkers for Parkinson's disease through an information entropy based approach. Phys Biol. 2020 Dec 1;18(1):016003

Applications: Further research focusing on the restricted number of genes belonging to the selected communities may reveal essential mechanisms responsible for PD at a network level and could contribute to the discovery of new biomarkers for PD → Personalized Medicine

MORE

Project and goals

ECHO SYSTEM

- Build a Decision Support platform to:

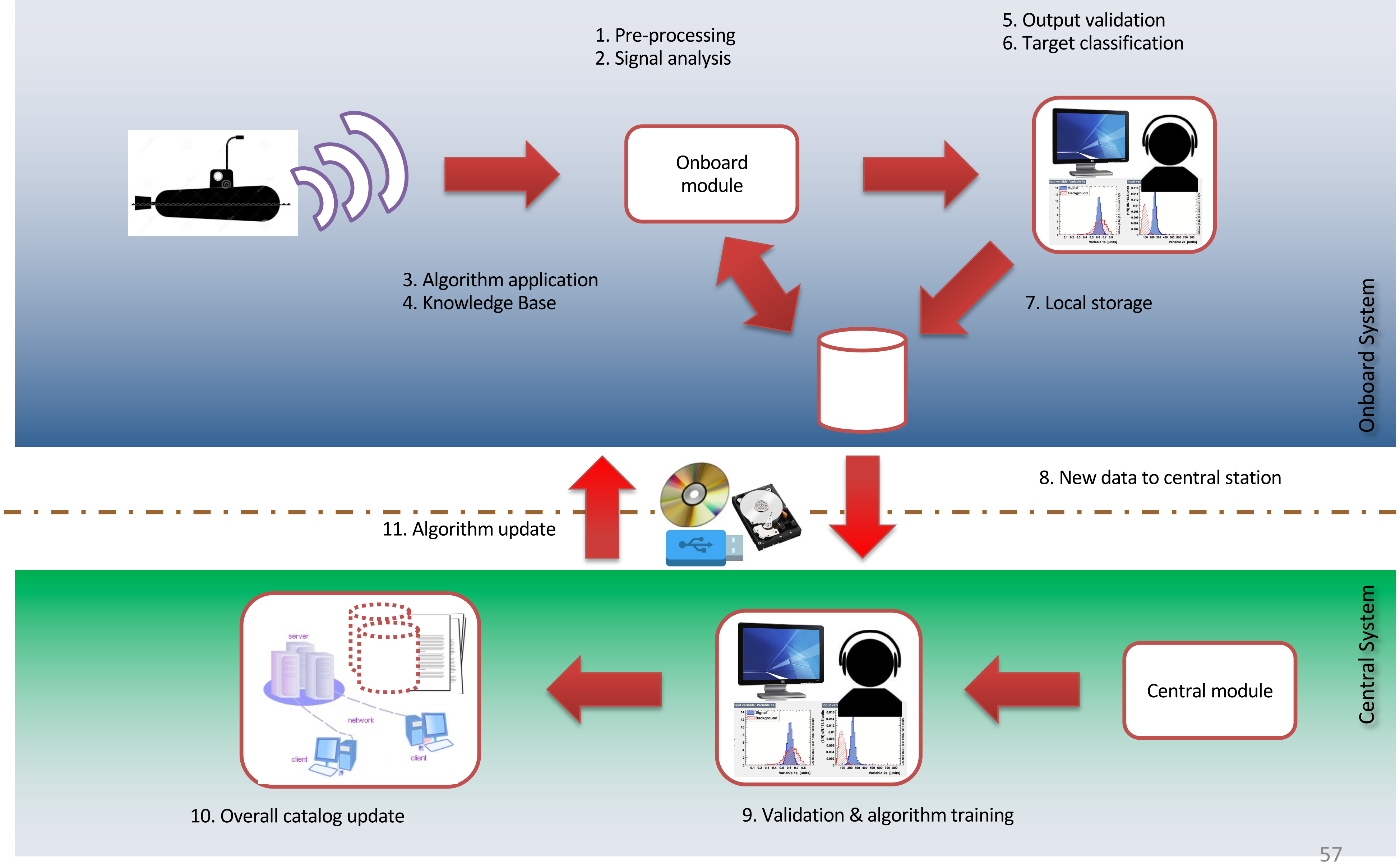
✓ process audio tracks acquired in an underwater environment;

✓ classify the detected target;

✓ provide the operator with the results of the processing for verification and validation.

Project Outline

ECHO SYSTEM


1. Pre-processing
2. Signal analysis

Onboard module

5. Output validation
6. Target classification

3. Algorithm application
4. Knowledge Base

7. Local storage

Onboard System

8. New data to central station

11. Algorithm update

Central module

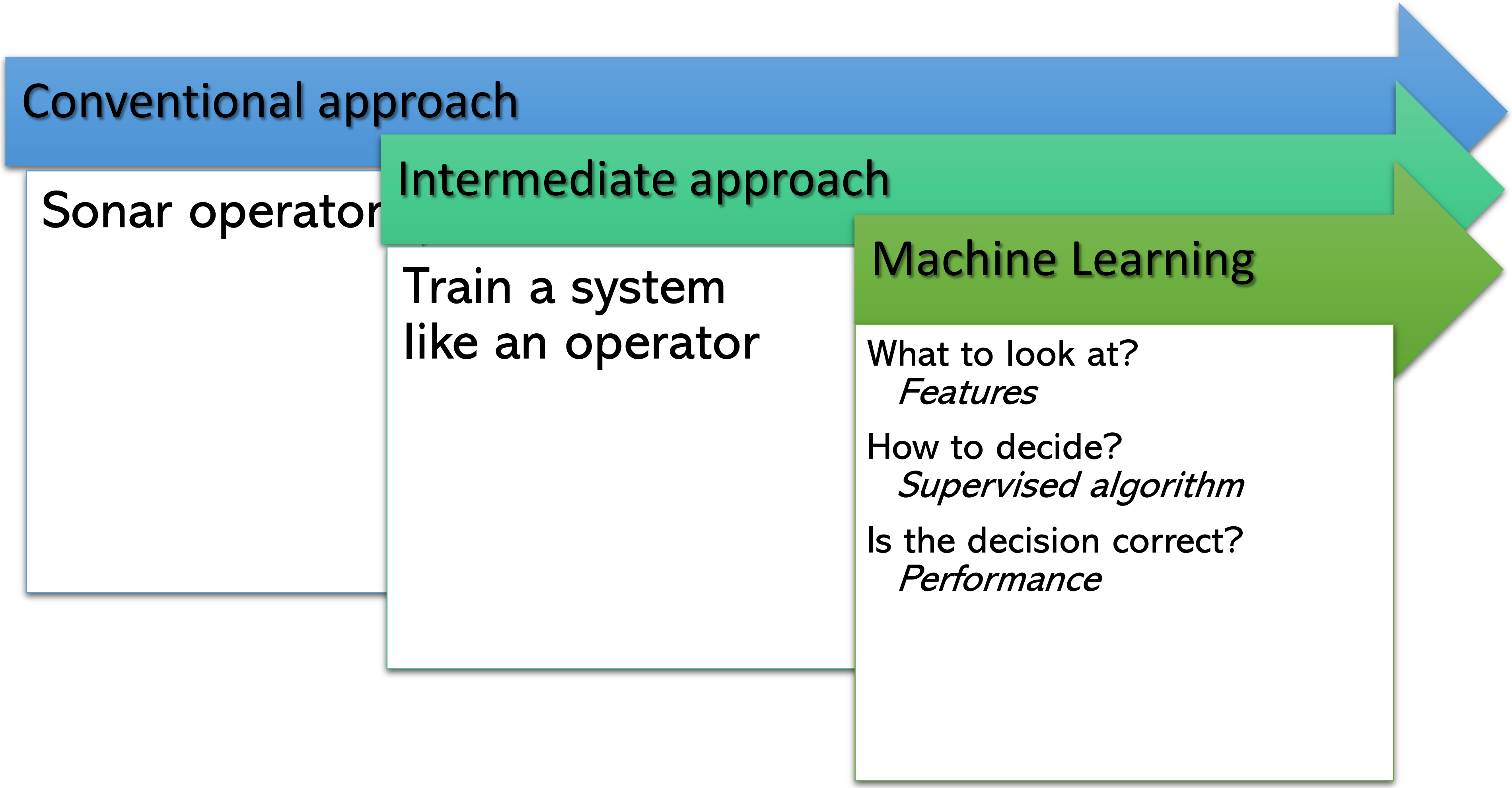
Central System

10. Overall catalog update

9. Validation & algorithm training

ECHO SYSTEM

Machine Learning approach



Conventional approach

Sonar operator

Intermediate approach

Train a system like an operator

Machine Learning

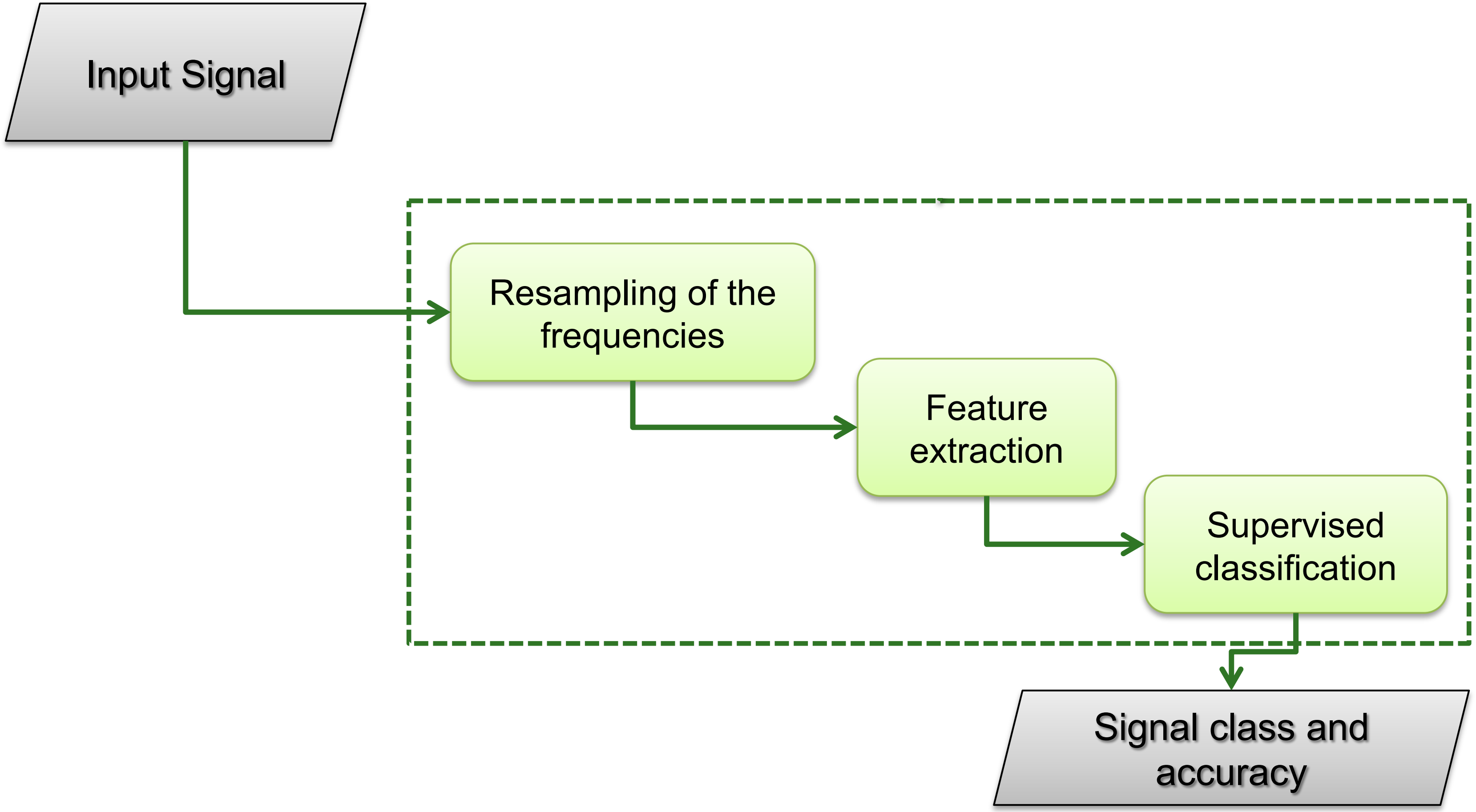
What to look at?
Features

How to decide?
Supervised algorithm

Is the decision correct?
Performance

ECHO SYSTEM

Flow chart

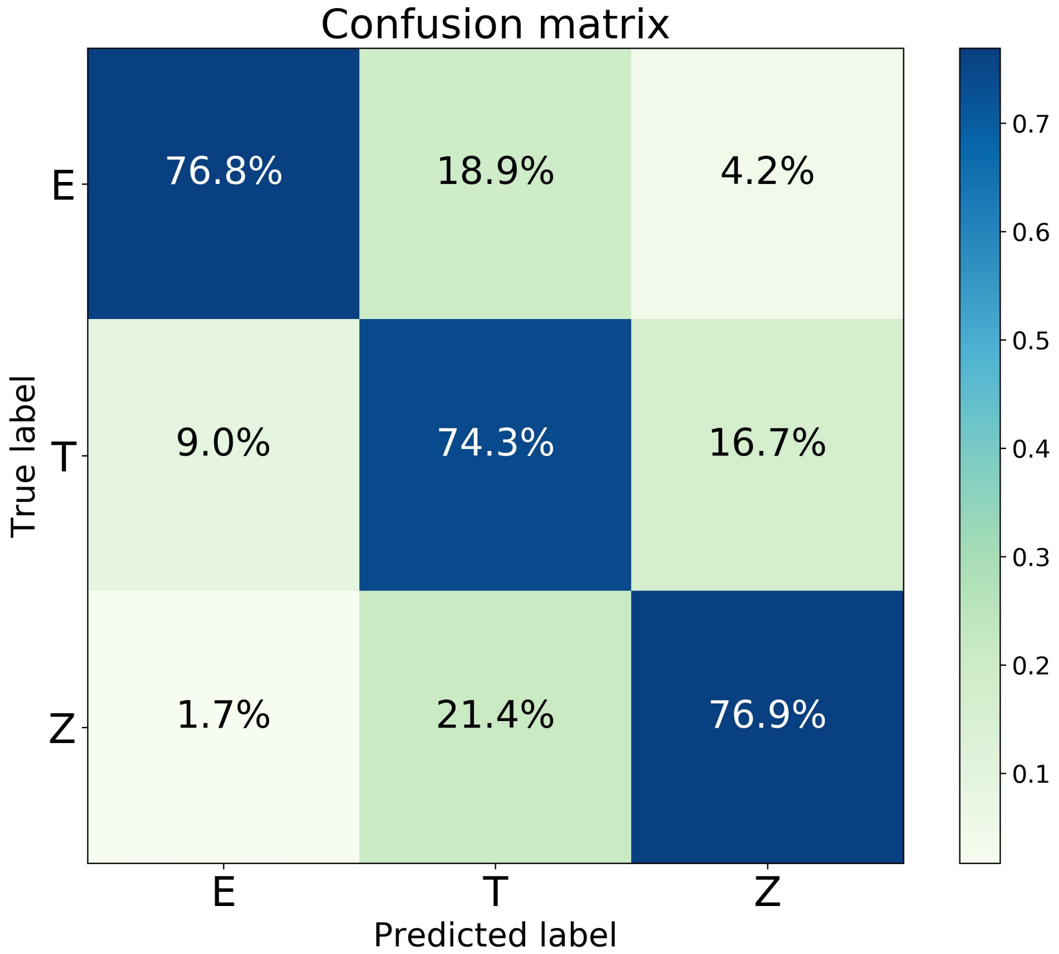

Input Signal
Resampling of the frequencies
Feature extraction
Supervised classification
Signal class and accuracy

Classification performances

ECHO SYSTEM

Output discarded if $\max(p_E, p_T, p_Z) < 0.4$

500 5-fold CV runs



Confusion matrix

True label \ Predicted label	E	T	Z
E	76.8%	18.9%	4.2%
T	9.0%	74.3%	16.7%
Z	1.7%	21.4%	76.9%

THANKS!