ARAMIS LAB
BRAIN DATA SCIENCE

UKRAINIAN CATHOLIC UNIVERSITY

FACULTY OF
APPLIED SCIENCES

23 – 25 July 2018

Lviv Data Science Summer School 2018

Machine Learning for Medical Applications:

Feature Extraction

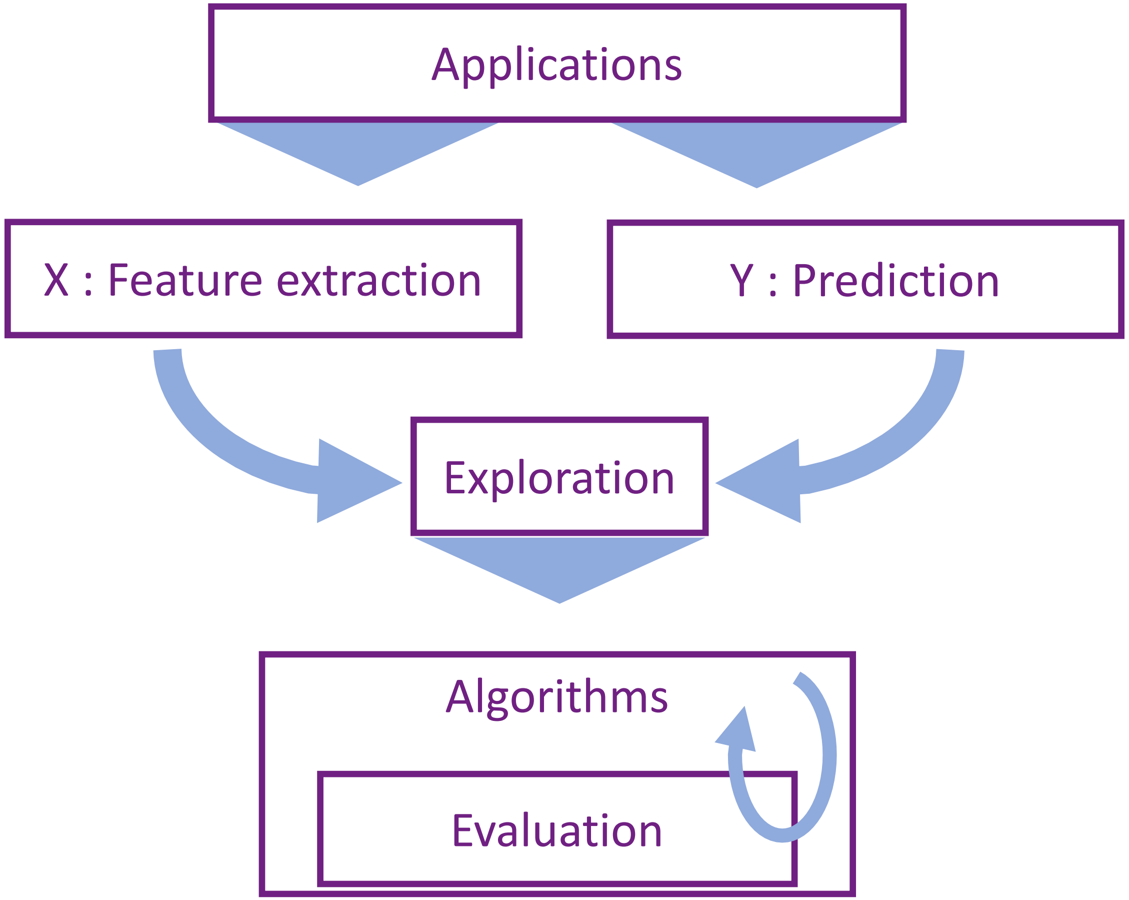
Igor Koval

PhD Student in Applied Mathematics

Brain and Spine Institute, Pitié Salpétrière Hospital, Paris, France
& Mathematical Laboratory of Ecole Polytechnique

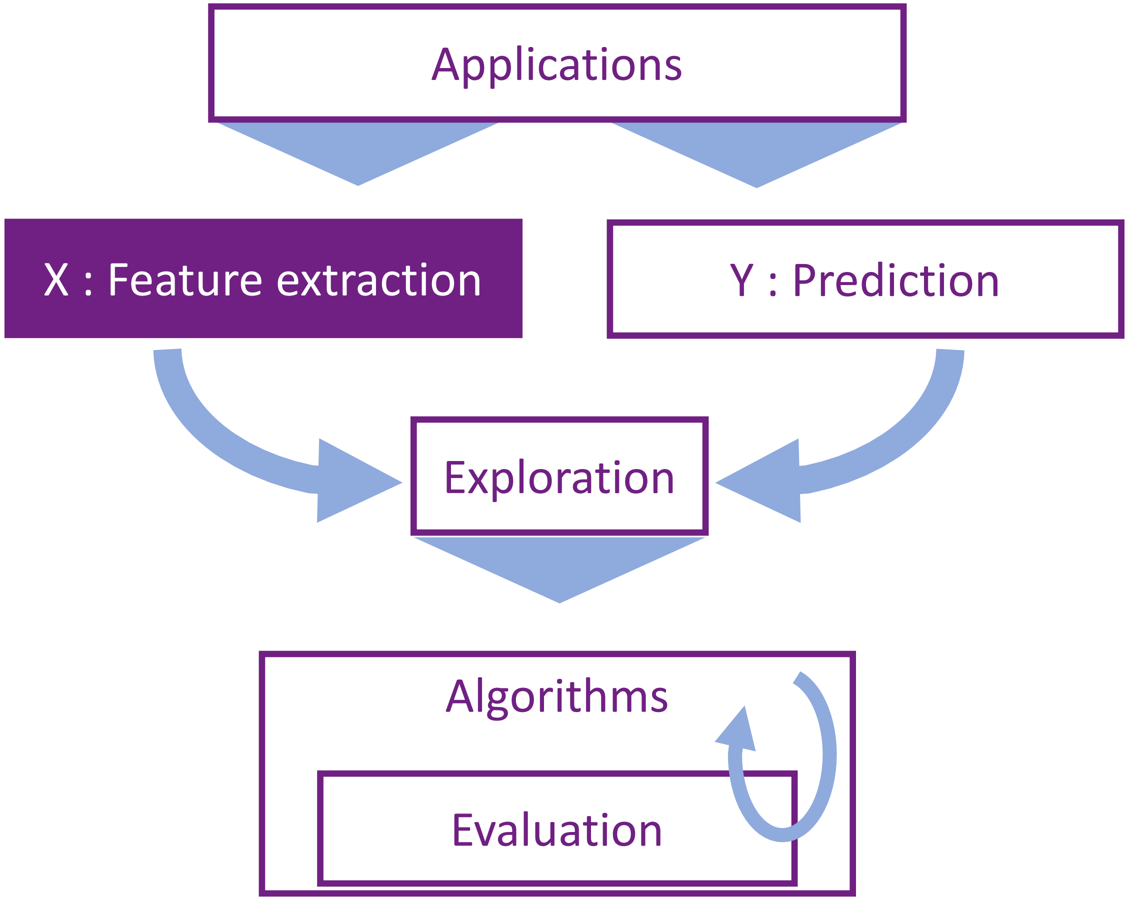
igor.koval@icm-institute.org

Plan



Applications
X : Feature extraction
Y : Prediction
Exploration
Algorithms
Evaluation

Plan


Applications

X : Feature extraction

Y : Prediction

Exploration

Algorithms

Evaluation

Medical data

Volume

PET scan or MRI :
- Millions of voxels

Humain brain :
- 33/86/100 billion neurons

Humain genom:
- 22,000 genes
- 3 billion base pairs

EEG :
- millisecond measurements
- sampling rates between 250 and 2000 Hz

Variety

Unstructured
- Text

Structured
- Time-series
- Medical imaging
- Biomarkers

Complexity

Multiple sources & protocols

Acquisition Noise & Outliers

Missing values

Multimodal data :
- Protocols include multiple data

No ground truth

What type of data ?

Biomarkers

Graph / Mesh / Network

Genomics

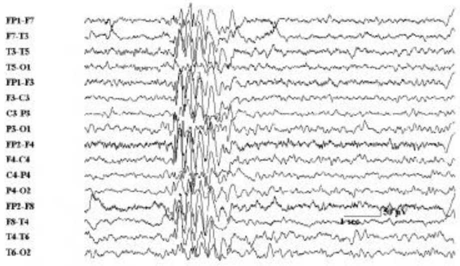
Medical check-up

Brain connectivity of surface

DNA

Doctor presciption

Röntgen first medical X-ray (of his wife)



EEG registration

Text

Medical Imaging

Time-series

Biomarkers

Examples ▶ Feature extraction ▶ Comments

Personal information

Medical check-ups

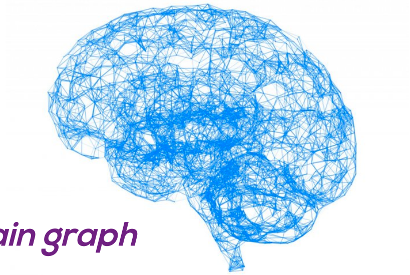
Smartphone data :
Quantified-self &
Personalized medecine

- **Qualitative**
  - Gender
  - Sex
  - Socioprofessional category
  - Environmental factors
  - Genetic mutation
  - Specific treatment
- **Quantitative**
  - Protein concentration
  - Blood pressure
  - Heart rate

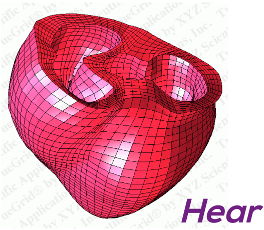
- **First data used, extensively in linear regressions**
- **They can be :**
  - Continuous
  - Discrete ordered
  - Discrete unordered
- **p-value (and statistical tests) regularization**

Graph / Mesh / Network

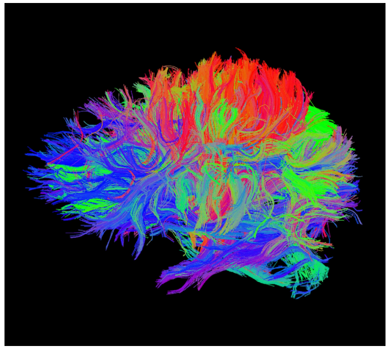
Examples



Brain graph



Heart mesh



Brain Network

Feature extraction

- **Graph topology**
  - Number of nodes
  - Connectivity distribution
  - Shortest path
  - Nearest neighbours & clicks
  - Directed graph features
  - Weighted graph features
- **Mesh spatial embedding**
  - Physical and geodesic distances
  - Spatial structure
- **Network**
  - Embeding time-series : (Anti)correlation matrix
  - Network specific features when the node is an object (an article, an individual, …)

Comments

- **Networks and meshes are particular graphs (same features)**
- **How to compare two graphs, or say that they belong to the same family?**
- **Trade-off between the resolution and the computational time**

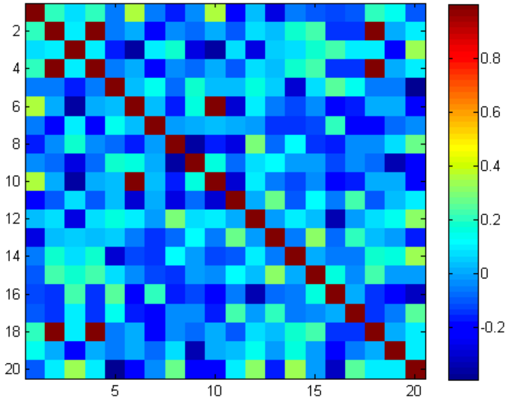
Genomics

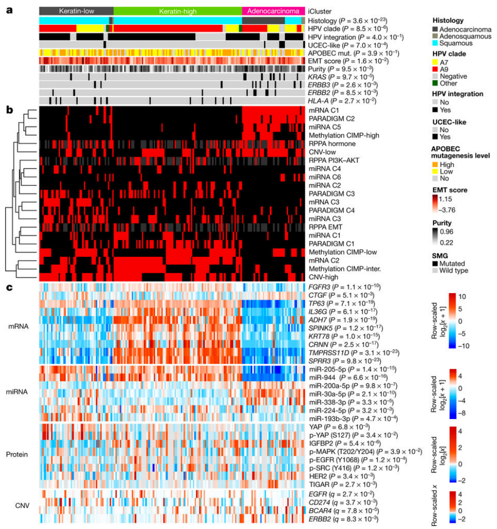
Examples ▶ **Feature extraction** ▶ **Comments**

DNA

- Allele number
- Gene number
- Mutation repetition
- Coexpression between genes

- Often in relation with other observations : causal relation
- Very high dimensionality
- Computational complexity
- Stability of the algorithms





Text

Examples

Prescription

Consultation & hospitalization reports

Ambulance reports

Feature extraction

- **Values**
  - Numbers : biomarkers, indicators, …
  - Dates
  - Evolution in time : different prescriptions, hospitalization, reports …
- **Words**
  - Occurences
  - Tf-Idf
- **Themes**
  - Latent Dirichlet Allocation (and other Probabilistic Graphical Models)
  - Word2Vec

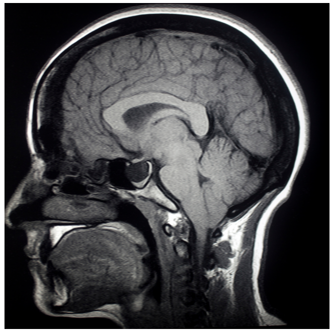
Comments

- **Are the texts transcripted in a numerical format? (Doctors handwriting recognition is not done yet …)**
- **Different models than the classical ML tasks**

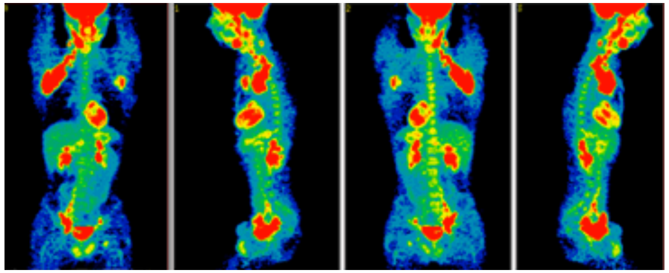
Medical imaging

Examples

X-ray



MRI : Magnetic Resonnance Imaging



PET : Positon Emission Tomography

Feature extraction

- **Structural**
  - Volume
  - Thickness
  - Ratio
  - Deformation
  - Distance
  - …
- **Functional**
  - Intensity
  - Distribution
  - Ratio
  - …
- **Deep Learning features**
  - Edges
  - Regions of interest
  - …

Comments

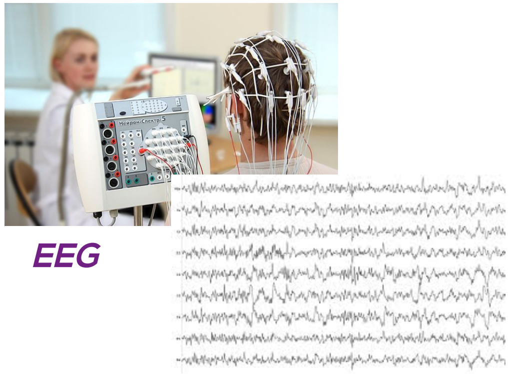
- Need an important preprocessing (PET cortex, Thickness value, …)
- Very high dimensionality
- Is it a structural of functional deformation?
- Realigned the patients
- Normalized the data to compare patients
- Same protocol & (hyper)parameters?
- Finite resolution, noise during the acquisition and noise during the extraction
- Cause of the inter-individual variability ?
- No operation (+, -, /, *) between images : not a Euclidean manifold

Time-series

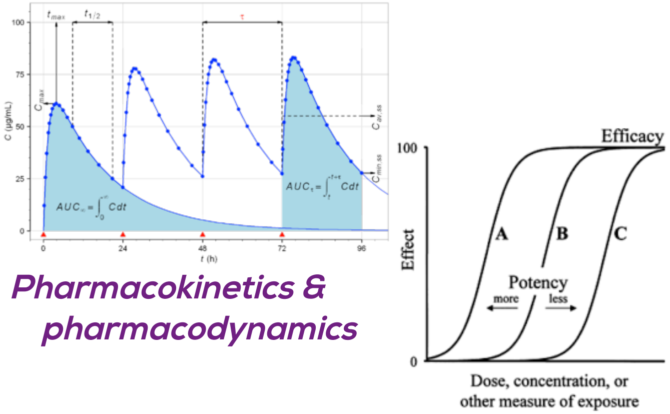
Examples

Feature extraction

Comments



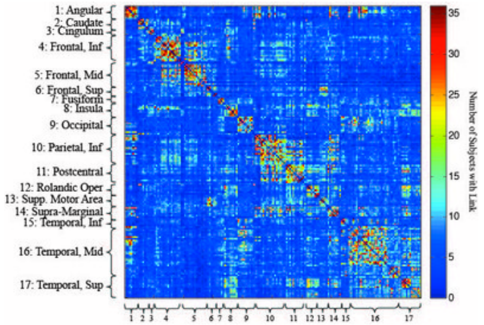
EEG



Pharmacokinetics & pharmacodynamics

Basically any previous feature

- **Time-series features**
  - Mean & standard deviation
  - Max / Min / Difference
  - Correlation / autocorrelation
  - Offset
  - Frequency domain features
  - …



Correlation matrix

- **All the previous problems**
- **Often needs preprocessing (time warping, normalization, noise removing by smoothing)**
- **Potential high dimensionality**
- **Different scales : milliseconds or years**

Practical session

Database :

MRI image of the brain at time t1
(256x256 pixels)

Follow-up of the brain at time t2
(256x256 pixels)

Prediction :

Brain tumor size
Evolution of the size

Objective :

1. Detection of a brain tumor within imaging data (no learning algorithms)
2. Estimation of the treatment effect based on the tumor evolution

- Part 1 : Thresholding Binarization
- Part 2 : Fuzzy C-means clustering

For both:

1. Estimation of the tumor size
2. Impact of the treatment
3. Effect of the hyperparameters