ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI FISICA E ASTRONOMIA - DIFA

PhD Thesis discussion

Implementation and optimization of algorithms in Biomedical Big Data Analytics

Nico Curti

Supervisor:
Prof. Daniel Remondini

Co-Supervisor:
Prof. Gastone Castellani
Prof. Armando Bazzani

March 16, 2020

Purposes

Biomedical Big Data Analytics

Feature Selection
DNetPRO

- **Topic:** Gene Expression Analysis;
- **Characteristics:** Novel algorithm & optimization on distributed computing;
- **Applications:**
  - Synthetic dataset;
  - Benchmark TCGA datasets;
  - Cytokinoma;
  - Bovine Signature;
  - Pedestrian mobility in Venice;
- **URL:** https://github.com/Nico-Curti/DNetPRO;

Deep Learning
Byron

- **Topic:** Neural Network models;
- **Characteristics:** Code Optimization and multi-threading parallelization;
- **Applications:**
  - Super Resolution;
  - Object Detection;
  - Image Segmentation;
- **URL:**
  - https://github.com/Nico-Curti/NumPyNet;
  - https://github.com/Nico-Curti/Byron;

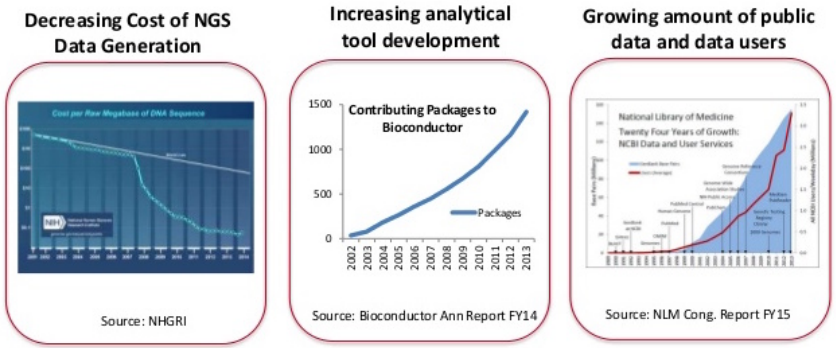
Big Data
CHIMeRA

- **Topic:**
  - Biomedical Big Data;
  - Database service;
  - Network analysis;
  - Natural Language Processing;
  - Web scraping;
  - Data harmonization;
- **Characteristics:** Merging of public available knowledge into network structure;
- **Applications:** Leukemia;
- **URL:** Not available yet;

The Big Data era

Biological Big Data

Considering the annual growth of data generation, the digital universe (i.e. data we generate annually) will reach 44 zettabytes by the year 2020, which is ten times the size of the digital universe in 2013. This trend in rising data volume is supported by:



Single Sequenced Human Genome

The size of a single sequenced human genome is approximately 200 Gb for less than US $ 1 000 (National Human Genome Research Institute, 2016).

Datasets available online

The amount of available data from large project such as 1 000 Genomes (http://www.1000genomes.org) will collectively approach the petabyte scale for the raw information alone.

A new research paradigm

Biological Big Data

The problems concerning the Big Data Analytics must be faced on two fronts:

Algorithm and Models:

- Data visualization
- Dimensionality reduction
- Parallel computing
- Optimization

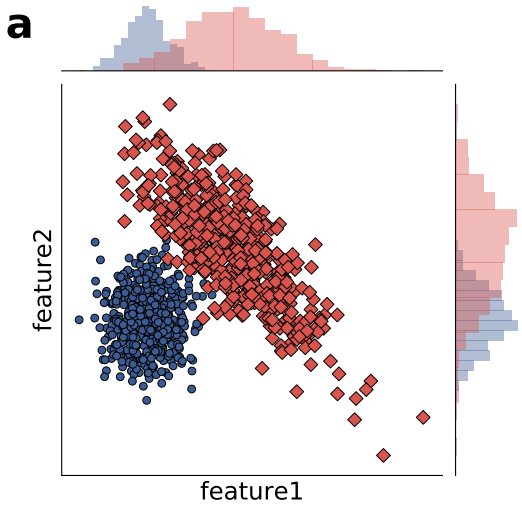
Hardware and Technology:

- HPC
- Cloud
- Storage
- Data exchange

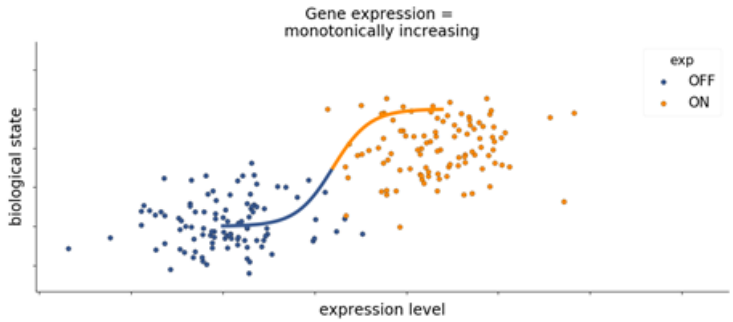
A **new paradigm** is emerging and it needs new hybrid sciences as Bioinformatics, Computational Sociology and new hybrid tools as Hybrid Cloud Architecture, Data Mining Algorithms.

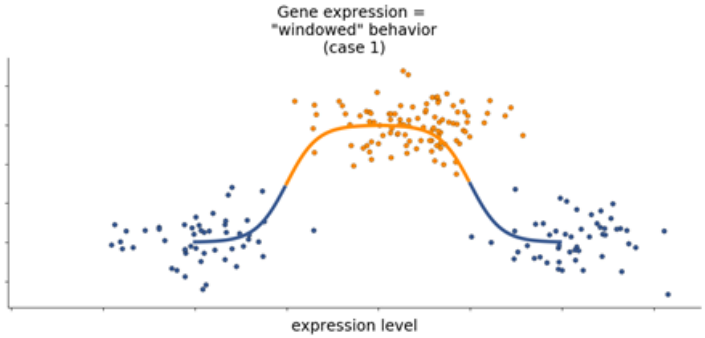
Biological features

Gene expression Big Data



- single parameter classification fails in predicting higher dimension classification performance;
- feature1 and feature2 classify badly in 1D, but have a very good performance in 2D;
- Activity of a biological feature as a function of its expression level: monotonically increasing; "windowed" behavior.





Classification and dimensionality reduction

DNetPRO algorithm

The potential clinical utility of genomic, transciptomic, epigenomic and proteomic data in aggregate remains largely unknown.
The goal of these type of analyses is to identify the **smallest number of probes** able to hold the most information about classes.

DNetPRO (Discriminant analysis with Network PROcessing) algorithm:

Computational Field:

- Feature selection
- Dimensionality reduction
- Classification

Biological Field:

- Evaluation of genes interaction
- Easy interpretation
- Extraction of Signature

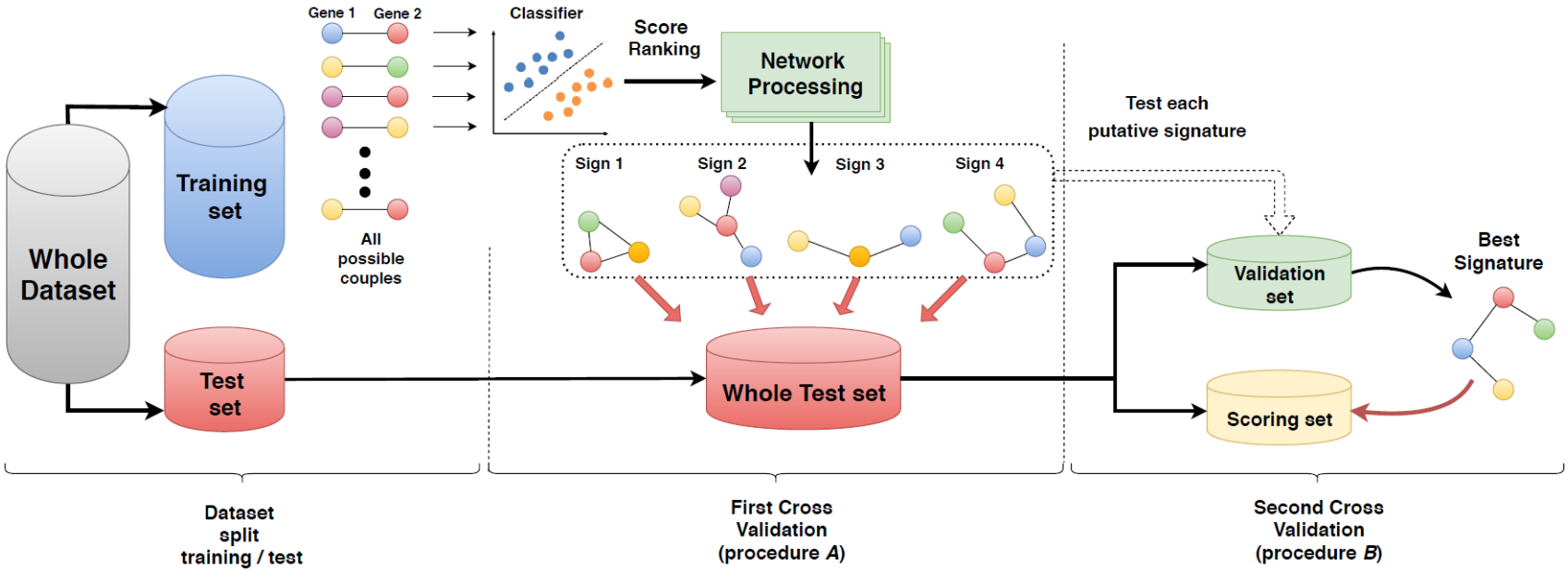
DNetPRO algorithm

Pipeline description

Data: Data matrix (N, S)
Result: List of putative signatures
Divide the data into training and test by an Hold-Out method;
for $couple \leftarrow (feature\_1, feature\_2) \in Couples$ **do**
    Leave-One-Out cross validation;
    Score estimation using a Classifier;
end
Sorting of the couples in ascending order according to their score;
Threshold over the couples score (Kbest couples);
for $component \in connected\_components$ **do**
    if *reduction* **then**
        Iteratively pendant node remotion;
    else
        continue;
    end
    Signature evaluation using a Classifier;
end

DNetPRO algorithm

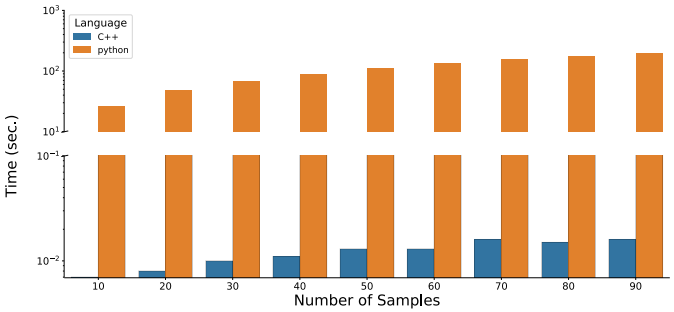
Pipeline description

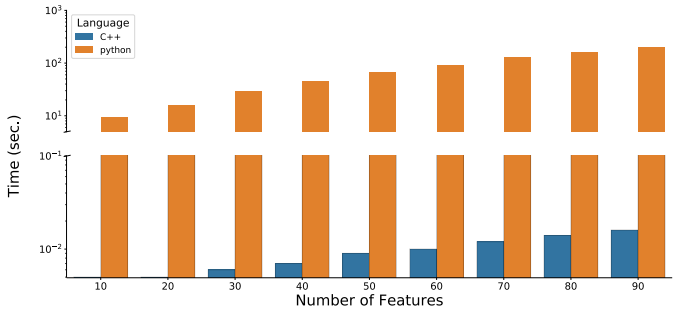


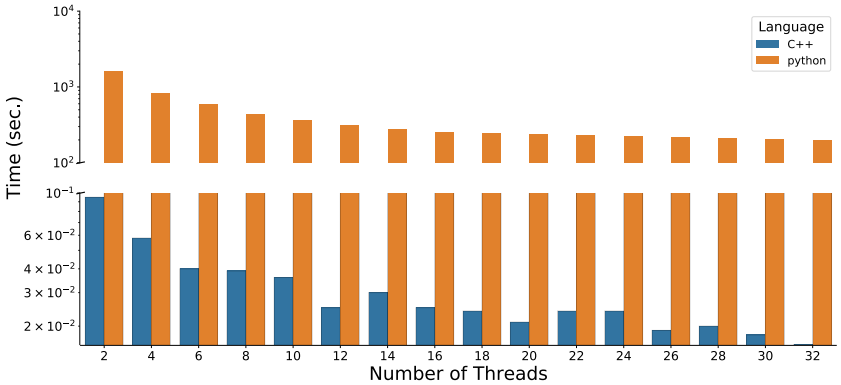
DNetPRO algorithm

Algorithm optimization

Porting on a parallel environment: C++ multi-threading implementation for network construction.
Distributed computation of each pipeline step (snakemake).



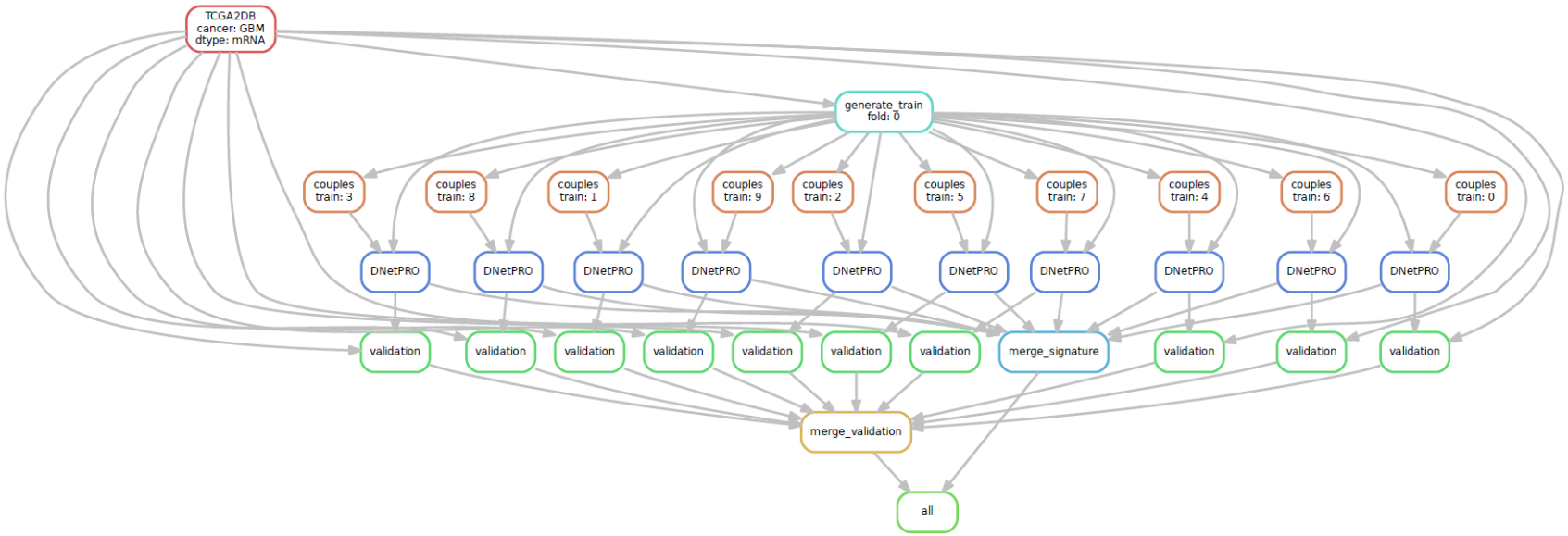




DNetPRO algorithm

Algorithm optimization

Porting on a parallel environment: C++ multi-threading implementation for network construction.
Distributed computation of each pipeline step (snakemake).

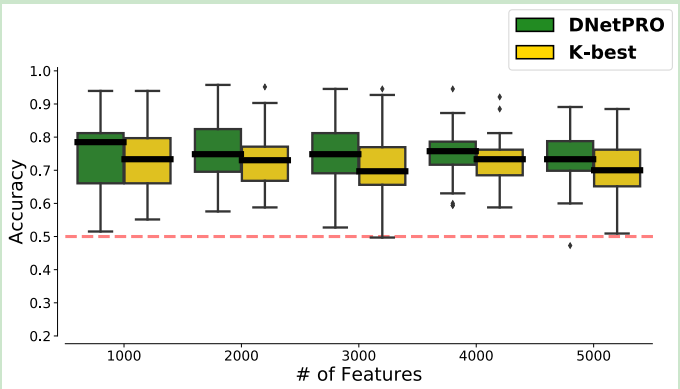


DNetPRO algorithm

Applications

Synthetic dataset

Study of the characteristics of the algorithm with different number of samples and features.



Cytokinome dataset

- Study on Alzheimer's disease;
- 26 cytokine expression levels;
- 289 subjects (189 female and 100 male);

Bovine dataset

- Study on Paratuberculosis disease;
- 15 mRNA samples of 3 bovine classes;
- 13529 features each;

Benchmark state-of-art

- Application on Synapse dataset;
- mRNA, miRNA and RPPA samples;
- 4 different tumors (KIRC, GBM, LUSC and OV);
- benchmark on state-of-art (Yuan et al., *Nature methods*, 2014);

Synapse dataset

DNetPRO benchmark

TCGA dataset - The Cancer Genome Atlas

Dichotomization: normal/tumor

Cancer	mRNA	miRNA	Protein	Number of samples
GBM	AgilentG4502A 17814	H-miRNA_8x15k 533	RPPA [a]	210
KIRC	HiseV2 20530	GA+Hiseq 1045	RPPA 166	243
OV	AgilentG4502A 17814	H-miRNA_8x15k 798	RPPA 165	379
LUSC	HiseqV2 20530	GA+Hiseq 1045	RPPA 174	121

4 cancer types:

- KIRC (*Kidney Renal clear cell Carcinoma*)
- GBM (*Glioblastoma Multiforme*)
- LUSC (*Lung Squamous Cell Carcinoma*)
- OV (*Ovarian serous cystadenocarcinoma*)

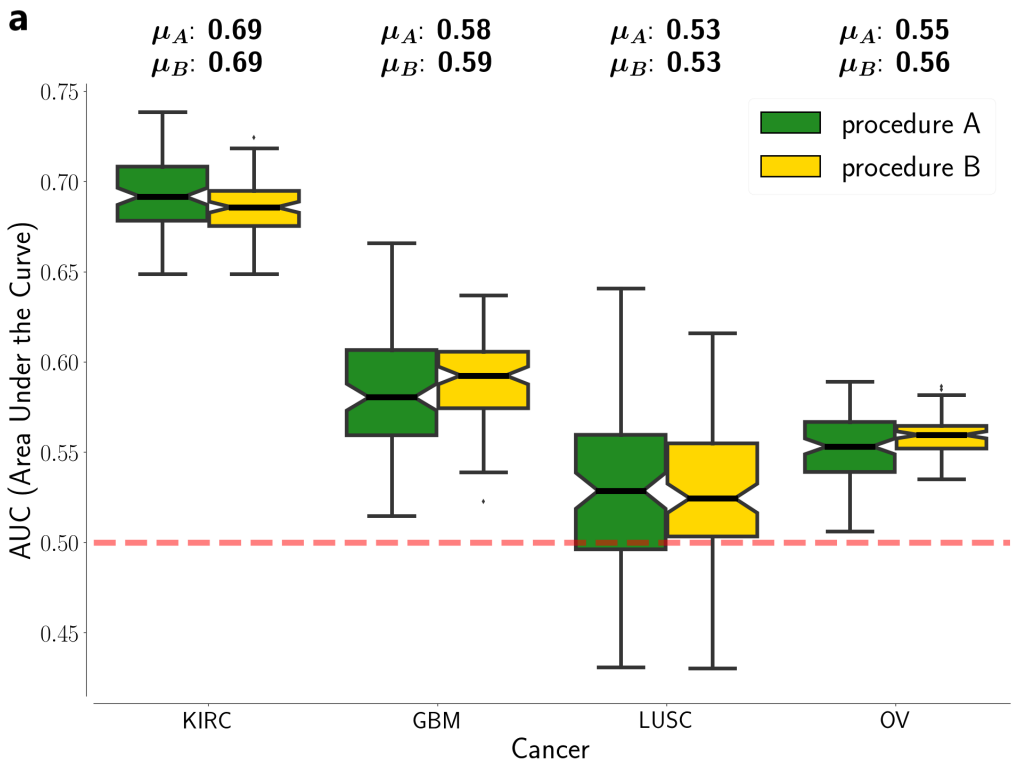
3 molecular data types:

- mRNA
- miRNA (*micro RNA*)
- RPPA (*Reverse Phase Protein Array*)

Synapse dataset

mRNA Results

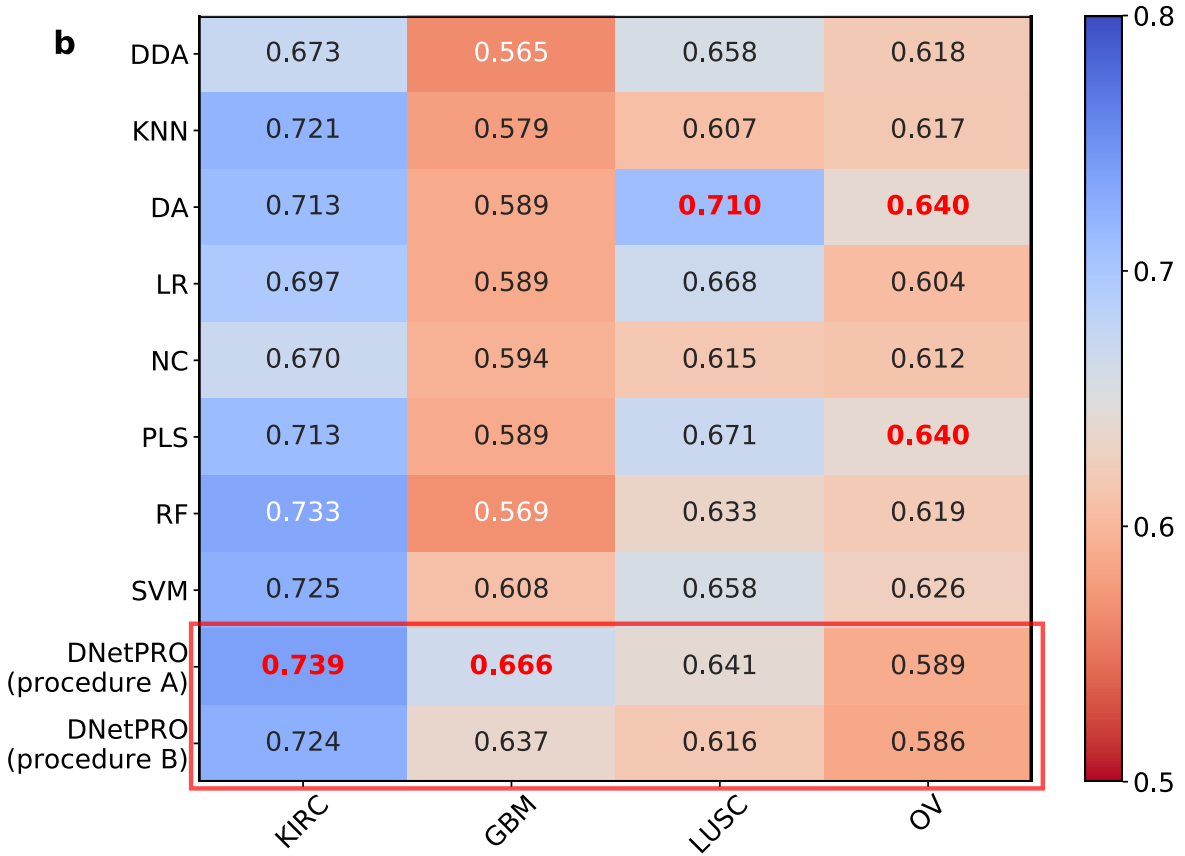
Benchmark with 8 classifiers with different features selection methods: **DDA** (diagonal discriminant analysis); **KNN** (K-nearest neighbor); **DA** (discriminant analysis); **LR** (logistic regression); **NC** (nearest centroid); **PLS** (partial least squares); **RF** (random forest); **SVM** (support vector machine).



Synapse dataset

mRNA Results

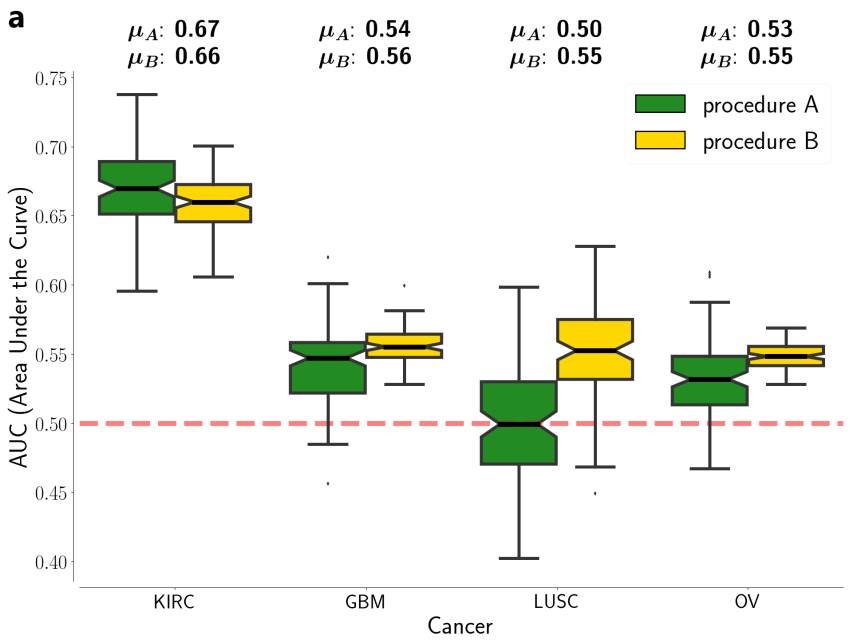
Benchmark with 8 classifiers with different features selection methods: **DDA** (diagonal discriminant analysis); **KNN** (K-nearest neighbor); **DA** (discriminant analysis); **LR** (logistic regression); **NC** (nearest centroid); **PLS** (partial least squares); **RF** (random forest); **SVM** (support vector machine).

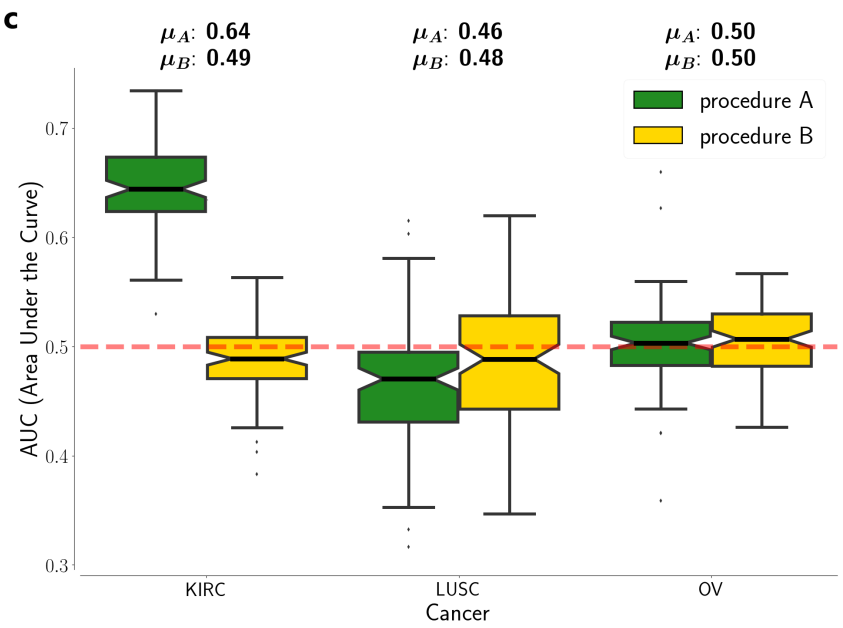


Synapse dataset

miRNA - RPPA Results

Benchmark with 8 classifiers with different features selection methods: **DDA** (diagonal discriminant analysis); **KNN** (K-nearest neighbor); **DA** (discriminant analysis); **LR** (logistic regression); **NC** (nearest centroid); **PLS** (partial least squares); **RF** (random forest); **SVM** (support vector machine).

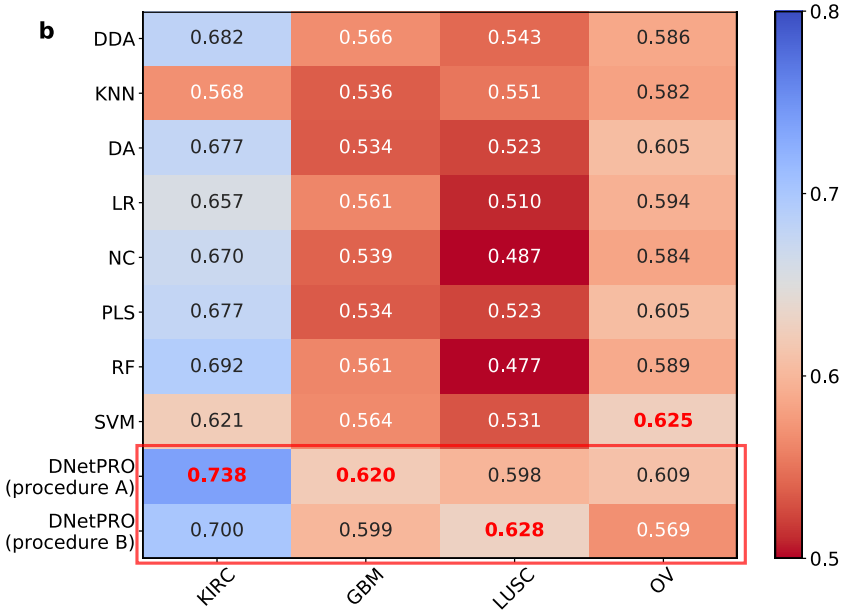


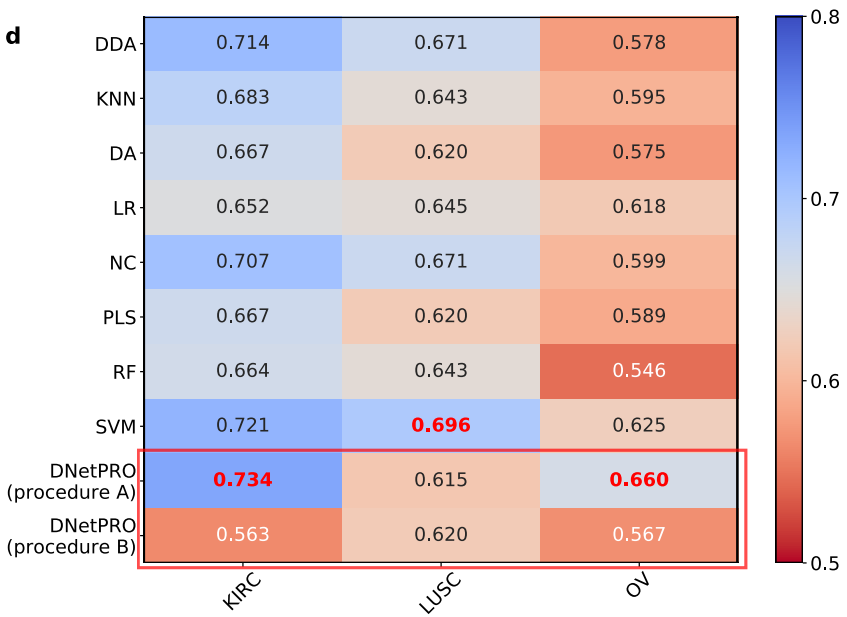


Synapse dataset

miRNA - RPPA Results

Benchmark with 8 classifiers with different features selection methods: **DDA** (diagonal discriminant analysis); **KNN** (K-nearest neighbor); **DA** (discriminant analysis); **LR** (logistic regression); **NC** (nearest centroid); **PLS** (partial least squares); **RF** (random forest); **SVM** (support vector machine).

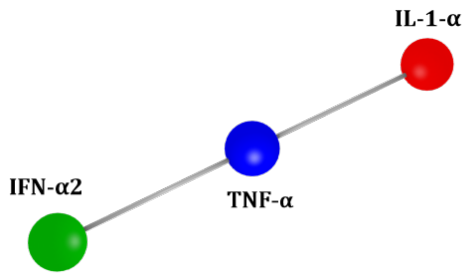




Cytokinome dataset

DNetPRO Applications

- Alzheimer's disease data.
- 26 number of cytokines.
- 3 classes: CTL (control), MCI (transient), AD (positive).
- 289 old-age subjects.



We extract the signature using the MCI class as AD-like samples.

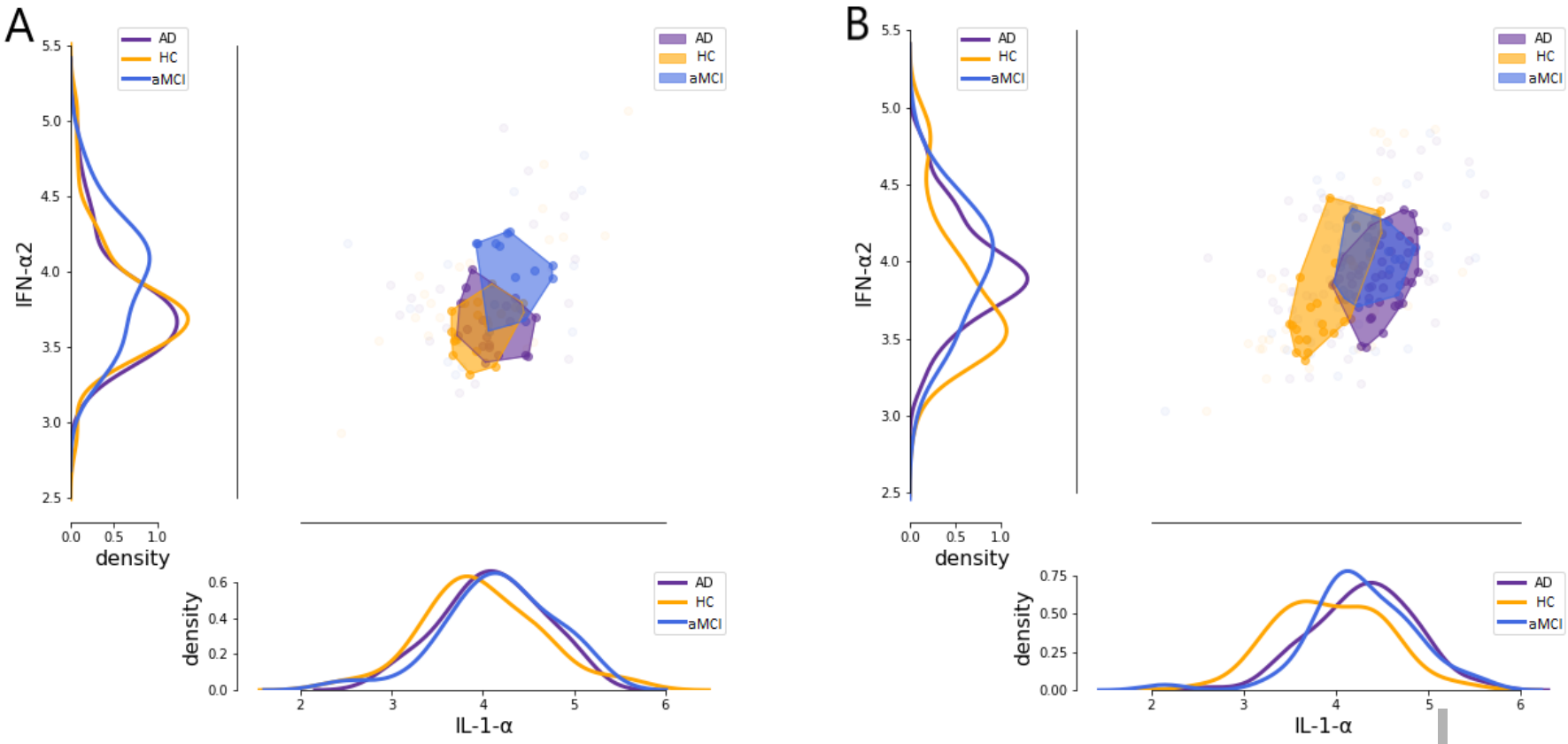
	Accuracy	Sensitivity	Specificity
	AD vs. CTL	AD	CTL
Men	16/25	8/12	8/13
	(64.00%)	(66.67%)	(61.54%)
Women	33/48	27/38	6/10
	(68.75%)	(71.05%)	(60.00%)
Total	47/72	36/54	11/18
	(65.24%)	(66.66%)	(61.11%)

Prediction	Sensitivity	Specificity
MCI as non-CTL	MCI	CTL
15/26	15/26	24/36
(57.69%)	(57.69%)	(66.67%)
41/47	41/47	23/51
(87.23%)	(87.23%)	(45.09%)
62/73	62/73	36/87
(84.93%)	(84.93%)	(41.38%)

Cytokinome dataset

DNetPRO Applications

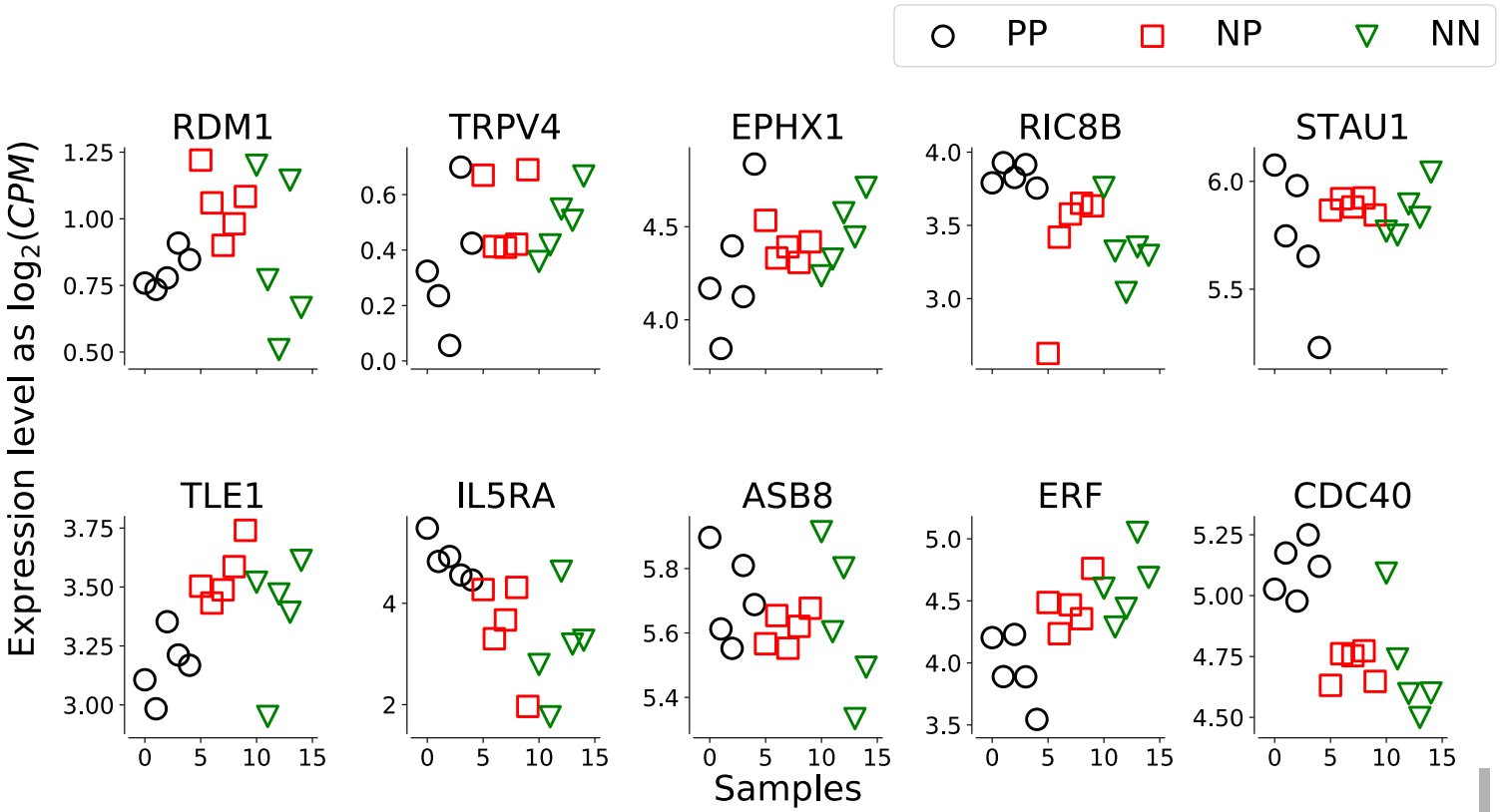
We extract the signature using the MCI class as AD-like samples.
Stratification according to the sex.
We observed a different behavior between **males** (A) and **females** (B) in CTL samples.



Bovine dataset

DNetPRO Applications

- Mycobatcterium avium subsp. paratuberculosis data.
- 15 036 number of genes.
- 3 classes: NN (negative), NP (exposed), PP (positive).
- 15 samples per probe.



Bovine dataset

DNetPRO Applications

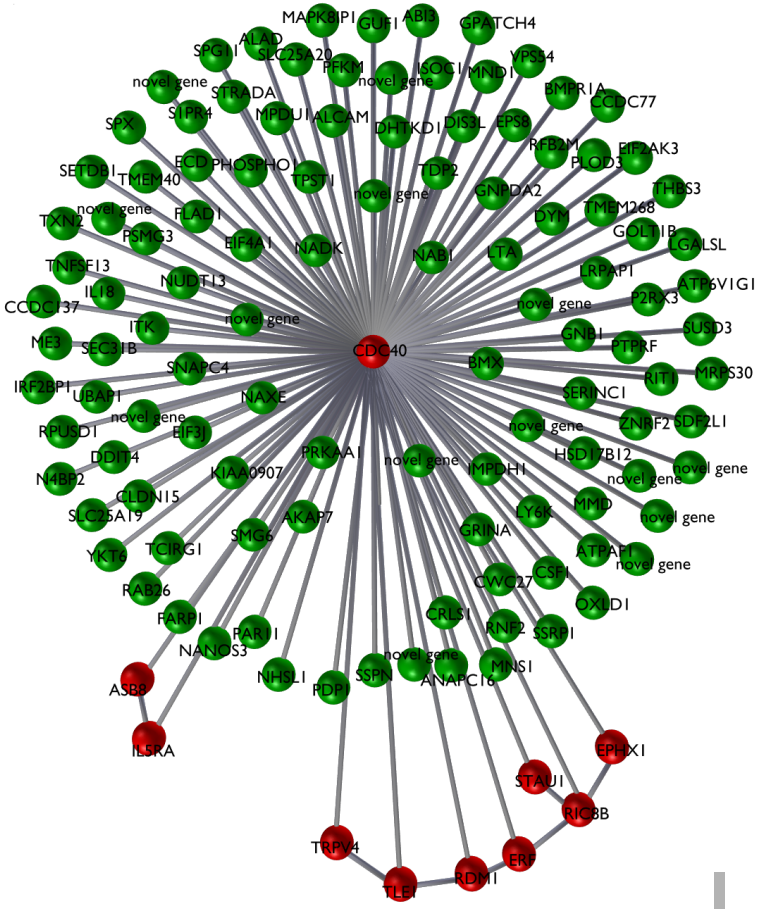
We extract the signature using the PP class as NP-like samples.

Results:

- Top performing signature (123 probes)
- Average performance 90%
- Matthews Correlation Coefficient 0.82

Pendant node remotion:

- Top performing signature (10 probes)
- Average performance 100%
- Matthews Correlation Coefficient 1.00

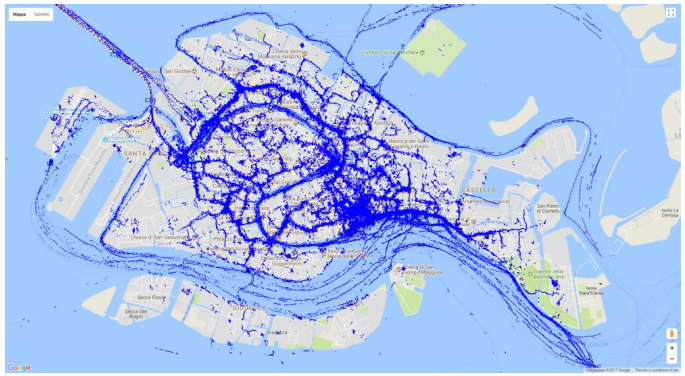


Social Application

Network pedestrian mobility

During major events the behavior of the crowd can lead to mobility and/or security issues.
ICT allows for the collection of a huge number of geolocalized timestamped data which can be analysed for:

- Crowd behavior detection and management
- Crowd mobility reconstruction and prediction



Data are provided within the Geosynthesis framework in collaboration with TIM Spa and Nokia.

Social Application

Network pedestrian mobility

Typical dataset size : 10^6 records/day
Typical users number : 5 000 users/day

Acquisition periods

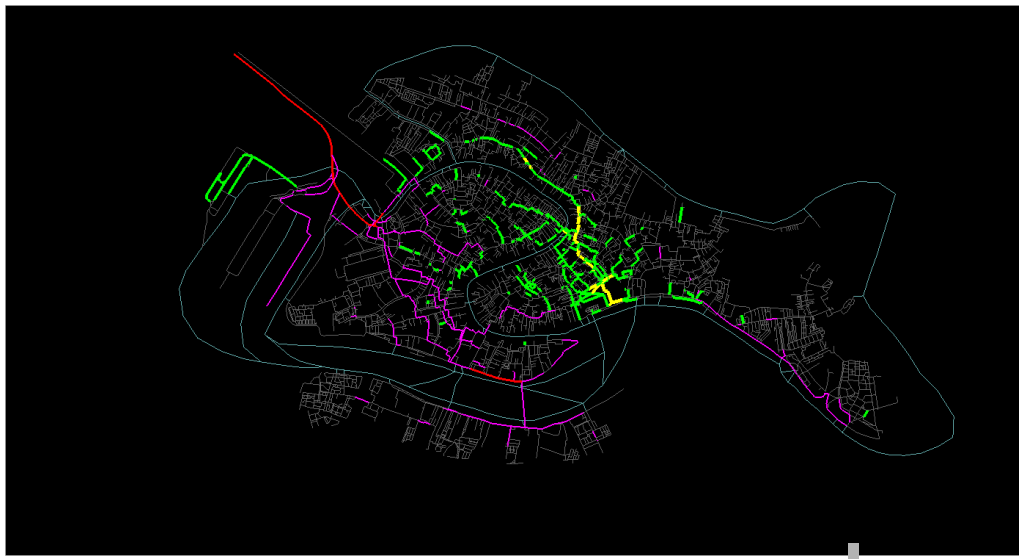
- Carnevale from 23/02/2017 to 02/03/2017
- Festa del Redentore 14,15,16/07/2017

Nationality disaggregation preferred paths:

- Italians (red, purple)
- Foreigners (yellow, green)

Statistics:

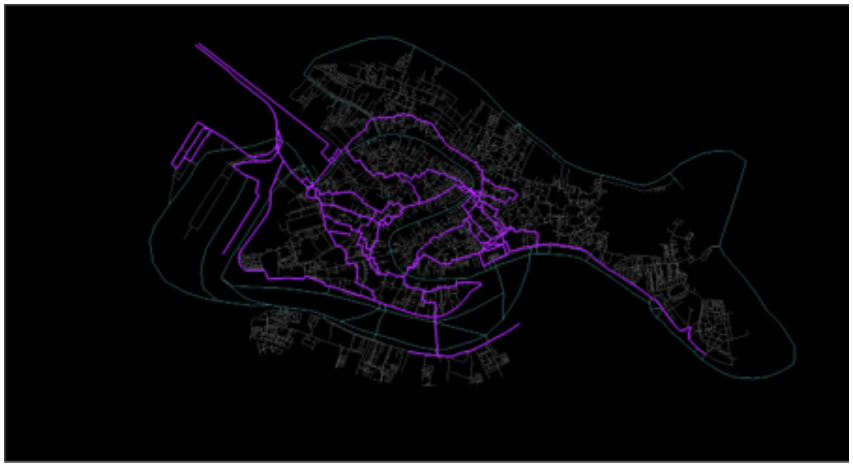
- Total lines : 11 596 039
- Errors : 92 713 (0.8%)
- Not georef : 8 715 928 (75.2%)
- Out of ROI : 1 773 013 (15.3%)
- Valid : 1 014 385 (8.7%)



Social Application

Network pedestrian mobility

Festa del Redentore 15/07/2017



Road network length fraction : 13% Mobility explained : 65%

Carnevale 26/02/2017

Road network length fraction : 15% Mobility explained : 64%

Conclusion

DNetPRO Algorithm Applications

- We proposed a novel technique of feature selection based on graph analysis;
- The method is optimized from a computational point of view and a it represents a valid alternative to standard methods;
- We proved the efficiency of DNetPRO algorithm on synthetic data proving its pros and cons;
- We highlighted its efficiency also in relation to state-of-art results on gene expression data;
- We applied the DNetPRO on several (real) dataset obtaining interesting results in several topics;
- We generalize the application of the DNetPRO algorithm also in non-biological data;

Deep Learning

Neural Network models

Deep Learning model is becoming equivalent to Neural Network architecture/model, i.e a more or less complex pipeline of functions which takes in input a sample and it applies a series of transformations to obtain the desired result.

Objectives

- Optimization and extension of state-of-art Neural Network libraries;
- Development of two novel libraries, for education and analytic purposes, respectively;
- Improve parallelization strategies to reach the best computational performances in cluster machine without GPUs supports;

All the algorithm and codes developed are **tested**, **documented** and **public available**!

Reference: https://nico-curti2.gitbook.io/phd-thesis (GitBook documentation).

What is a Deep Learning model?

Neural Network models

- Neural Networks are mathematical models **commonly used** in data analysis.
- From a theoretical point-of-view we can define a Neural Network as **a series of non-linear multi-parametric functions**.
- Neural Networks are considered as **Universal Function Approximators**.

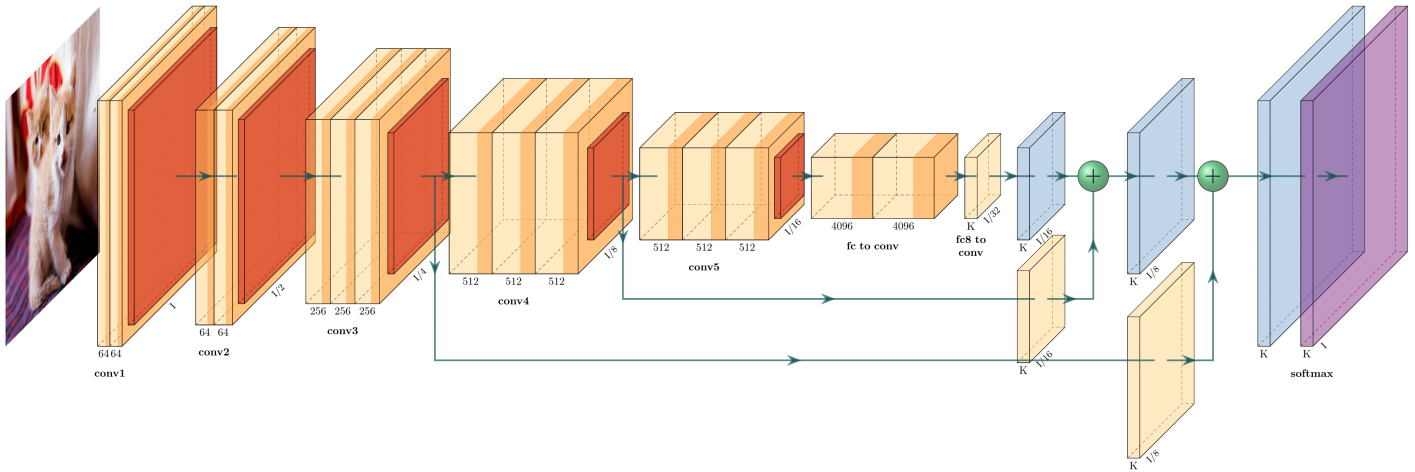


Figure: Neural Network model scheme. The model is composed by a series of layers made by interconnected computational units. The connections between layers can be serials or parallels and each layer models a pre-determined mathematical function

State-of-art libraries

Why other two libraries?

A wide range of documentations and implementations have been written on Deep Learning and it is more and more hard to move around the different sources. Leader on this topic are became the multiple open-source 'Python' libraries available on-line as Tensorflow, Pytorch and Caffe.

Pros

- Simplicity in writing complex models in a minimum number of code lines;
- Simplicity of the Python language;
- Extremely efficient in GPU environments;

Cons

- Steep learning curve for complex applications or algorithm modifications;
- Hard to manage as services;
- They are not designed for CPU performances;

Only a small part of the research community uses more deeper implementation in C++ or other low-level programming languages. About them it should be mentioned the darknet project of Redmon J. et al. which created a sort of standard in object detection applications using a pure Ansi-C library.

NumPyNet & Byron

Why other two libraries?

NumPyNet

- **LANGUAGE:** **Python**;
- **INSTALLATION:** **Setup** (**PyPI** upload wip);
- **OS:** Linux & MacOS & Windows;
- **DOCUMENTATION:** fully documented;
- **PURPOSE:** educational / study;
- **CI:** Travis & Appveyor;
- **USAGE:** model design / algorithm improvements;
- **MISCELLANEOUS:** all functionality are tested against **Keras** counterpart;
- **URL:** https://github.com/Nico-Curti/NumPyNet

Byron

- **LANGUAGE:** **C++** Standard 17;
- **INSTALLATION:** **CMake** & **Make** & **Docker**;
- **OS:** Linux & MacOS & Windows;
- **DOCUMENTATION:** wide set of usage examples;
- **PURPOSE:** multi-threading CPU optimization;
- **CI:** Travis & Appveyor & CircleCI;
- **USAGE:** Super Resolution & Object Detection & Image Segmentation;
- **MISCELLANEOUS:** extension and optimization of the **darknet** project;
- **URL:** https://github.com/Nico-Curti/Byron

From the union of this two project **Pyron** is born, i.e the completely wrap of the Byron library in Python using Cython.
* Both libraries are already used in different works of thesis and research projects!

NumPyNet

Neural Networks in Pure NumPy

NumPyNet

- The library is written in pure Python and the only **external library** used is Numpy (a based package for the scientific research).
- Despite all common libraries are correlated by a wide documentation is often difficult for novel users to move around the many hyper-links and papers cite in it. NumPyNet tries to overcome this problem with a **minimal mathematical documentation associated to each script** and a **wide range of comments inside the code**.
- Libraries like Tensorflow are efficient by a computational point-of-view and they have an extremely simple user interface. On the other hand the **deeper functionalities** of the code and the **implementation strategies** used **are unavoidably hidden behind tons of code lines**. In this way the user can performs complex computational tasks using the library as **black-box package**. NumPyNet wants to overcome this problem using **simple Python codes with extremely readability also for novel users** to better understand the symmetry between mathematical formulas and code.
- NumPyNet uses Python threads to easily manage independent tasks and optimize the computation.

* The library was developed in collaboration with Dott. Ceccarelli.

Byron

Build YouR Own Neural Network

Byron

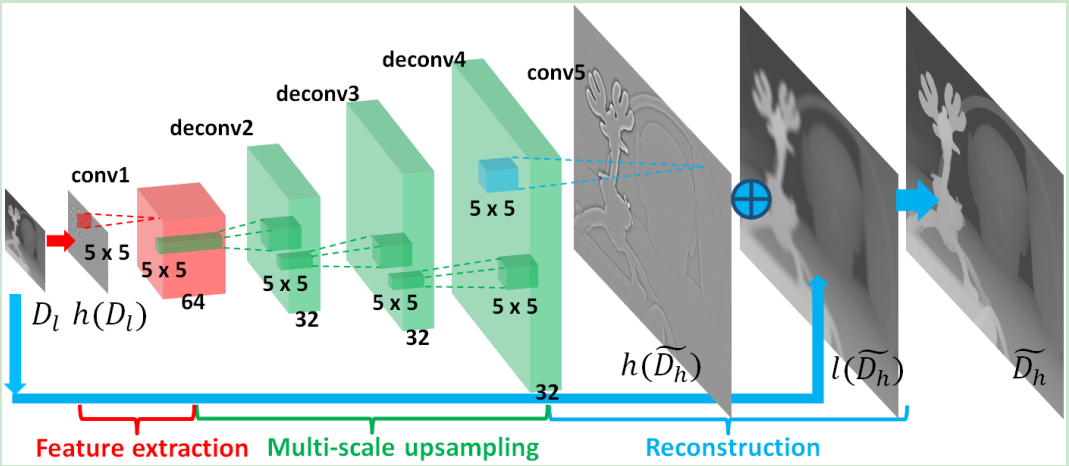
- The library is written in **pure C++** with the support of the modern standard 17.
- The library is **optimized for image processing** (probably the most common task in biomedical research).
- Starting from the darknet project backbone, Byron proposes **numerous improvements and fixes**.
- Byron works in a **fully parallel section** (OpenMP) in which each single computational function is performed using the full set of available cores. To further reduce the time of thread spawn and so optimize as much as possible the code performances, the library **works using a single parallel section** which is opened at the beginning of the computation and closed at the end.
- The library is also **completely wrapped** using Cython to enlarge the range of users also to the Python ones.

* The library was developed in collaboration with Dott. Baroncini.

Applications

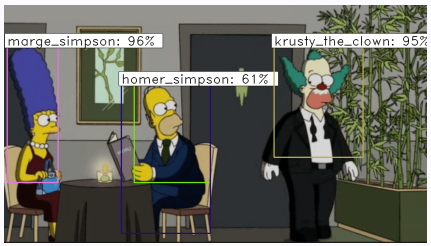
Deep Learning model tasks

Single Image Super Resolution



Improve spatial image resolution and reconstruction from low-resolution (LR) to high-resolution (HR).

Object Detection



Identify single or multiple objects into a picture or video stream.

Image Segmentation

Extract the exact pixels which belong to an object into a given image.

Single Image Super Resolution

What is Super Resolution?

We commonly think about Neural Network models as tools to **reduce** the problem dimensionality, i.e starting from high-dimensional data (e.g $w \times h \times c$ image) the model predicts a class (single number).
In the Super Resolution problem we have no classes but the desired output is a image.
Single Image Super Resolution tasks starts from a image aiming to reconstruct its HR counterpart.

Super resolution procedure

- We start from a prior-known HR image;
- We rescale this image obtaining its LR counterpart;
- We feed a super-resolution model with the LR image aiming to obtain in output its HR version;
- With the trained model we can re-apply the up-sampling procedure and further increase the HR quality*.

If this pipeline is performed feeding a Neural Network model using one-by-one a series of images, we talk about Single Image Super Resolution (SISR) algorithm.

* This is even more efficient if we think about a **zoom** procedure.

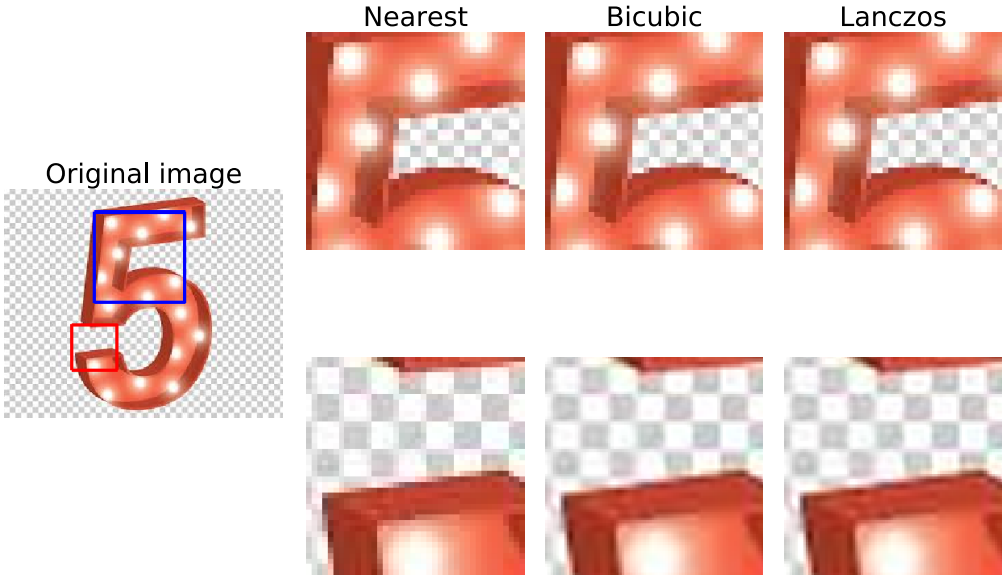
Standard up/down scaling algorithms

Single Image Super Resolution

Interpolation algorithms:

- **Nearest**;
- **Bicubic**;
- **Lanczos**;

The best compromise between computation cost and efficiency is given by the Bicubic interpolation algorithm. Given a pixel, the interpolation function evaluates the 4 pixels around it applying a filter given by the equation:



$$k(x) = \frac{1}{6}\begin{cases} (12-9B-6C)|x|^3 + (-18+12B+6C)|x|^2 + (6-2B) & \text{if} |x| < 1 \\ (-B-6C)|x|^3 + (6B+30C)|x|^2 + (-12B-48C)|x| + (8B+24C) & \text{if} 1 \leq |x| < 2 \\ 0 & \text{otherwise} \end{cases}$$

where x identifies each pixel below the filter. Commonly values used for the filter parameters are $B = 0$ and $C = 0.75$ (used by OpenCV library) or $B = 0$ and $C = 0.5$ used by Matlab.
NOTE: the nearest interpolation could be achieved also using *Pooling* algorithm.

Image Quality

Single Image Super Resolution

The most common image quality evaluator is given by our eyes.
Quantitative evaluations:

PSNR - Peak Signal to Noise Ratio

- It establishes the compression lossless of an image;
- Equation:

$$PSNR = 20 \cdot \log_{10}\left(\frac{\max(I)}{\sqrt{(MSE)}}\right)$$

where $\max(I)$ is the maximum value which can be taken by a pixel in the image and MSE is the Mean Square Error evaluated between the original and reconstructed images.

SSIM - Structural SIMilarity index

- It evaluates the structural similarity between two images taking into account also the visible improvement seen by human eyes.
- Equation:

$$SSIM(I,K) = \frac{1}{N}\sum_{i=1}^{N}\frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{({\mu_x}^2 + {\mu_y}^2 + c_1)({\sigma_x}^2 + {\sigma_y}^2 + c_2)}$$

where N is the number of arbitrary patches which divide the image and c_1 and c_2 parameters are fixed to avoid mathematical divergence.

	PSNR (dB)	SSIM ($\in [-1,1]$)
Nearest	25.118	0.847
Bicubic	27.254	0.894
Lanczos	26.566	0.871

EDSR - Enhanced Deep Super Resolution

Super Resolution Models

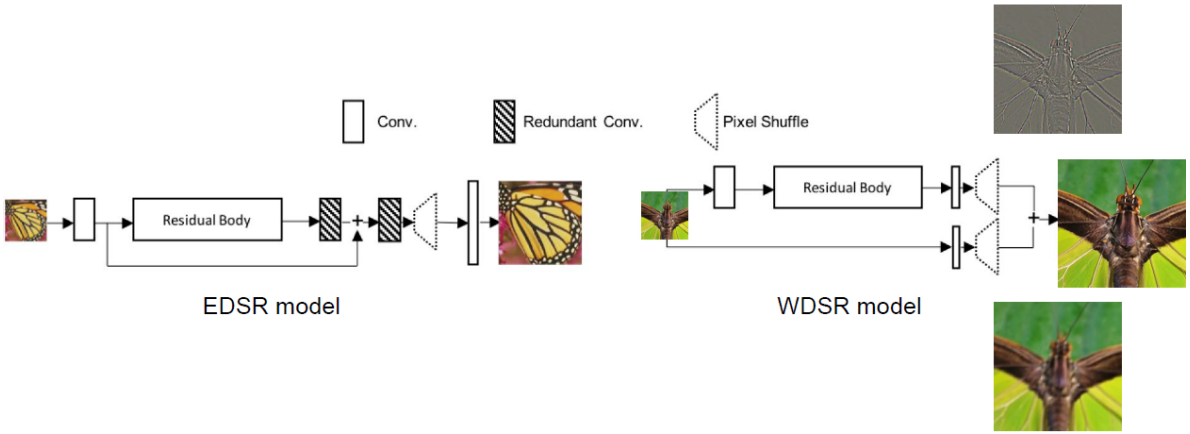
Architecture

Layer	Channels input/output	Filter dimensions	Number of Parameters
Conv. input	3/256	3×3	6912
Conv. (residual block)	256/256	3×3	589824
conv. (pre-shuffle)	256/256	3×3	589824
Conv. (upsample block)	256/1024	3×3	2359296
Conv. output	256/3	3×3	6912

Average Time on 510×339 image: 576.92 s.
Winner of the **NTIRE 2017**. More than **3 million** of parameters!

WDSR - Wide Deep Super Resolution

Super Resolution Models



Architecture

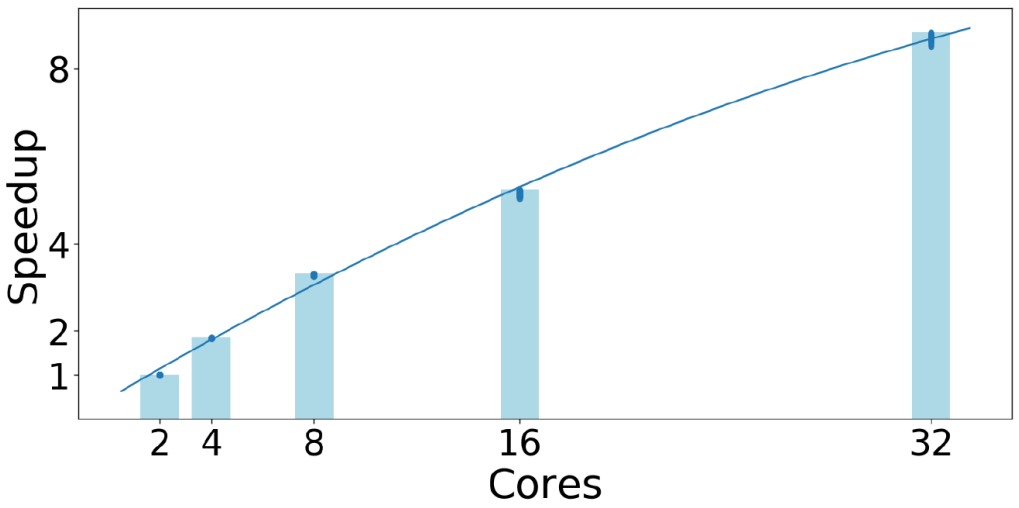
Layer	Channels input/output	Filter dimensions	Number of Parameters
Conv. input 1	3/32	3×3	864
Conv. 1 (residual block)	32/192	3×3	55296
conv. 2 (residual block)	192/32	3×3	55296
Conv. (pre-shuffle)	32/48	3×3	13824
Conv. input 2 (pre-shuffle)	3/48	5×5	3600

Average Time on 510×339 image: 46.35 s.
Winner of the **NTIRE 2018**. $\sim$ 100K parameters, **less than** 10% **of EDSR model**!

Computational Performances

Super Resolution Models

Computational performances of WDSR model on 510 × 339 images (100 runs). Environment: Bioinformatics server grade machine (128 GB RAM memory and 2 CPU E5-2620, with 8 cores each.



N° threads	Average Time (s)
2	403.478
4	219.268
8	122.986
16	78.571
32	46.348

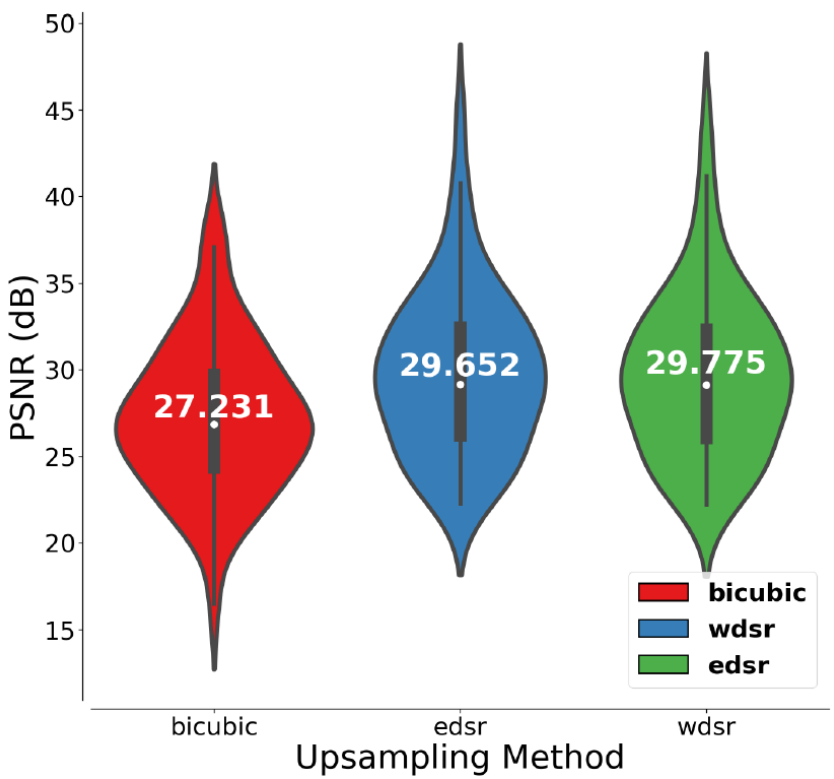
Sub-linear trend:

- Hyperthreading;
- NO affinity;

Quality Performances

Super Resolution Models

Evaluation of PSNR and SSIM on DIV2K validation set.



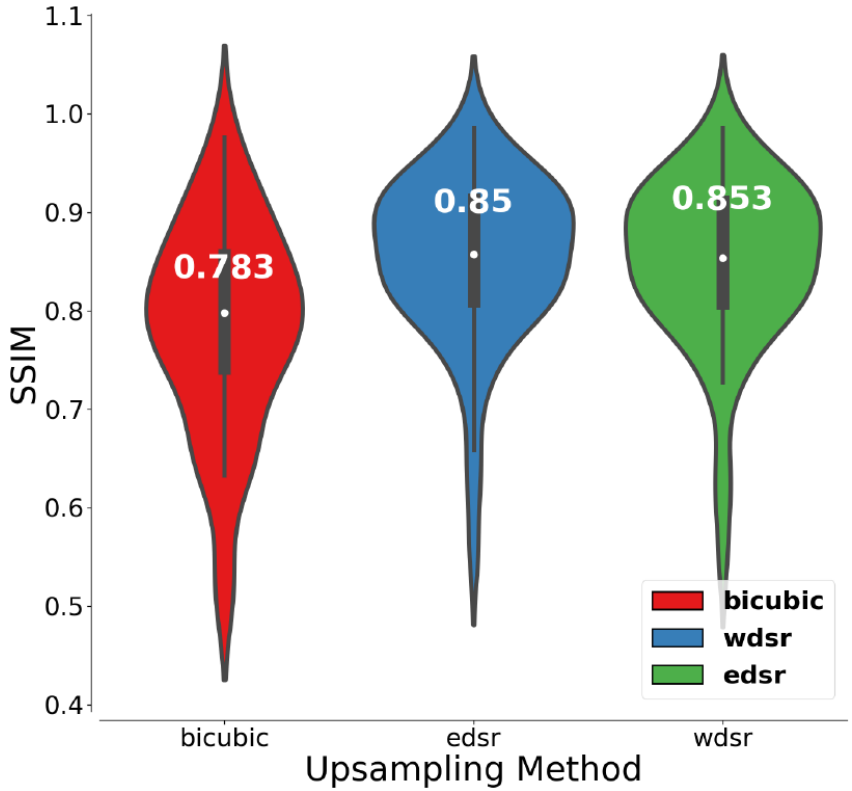
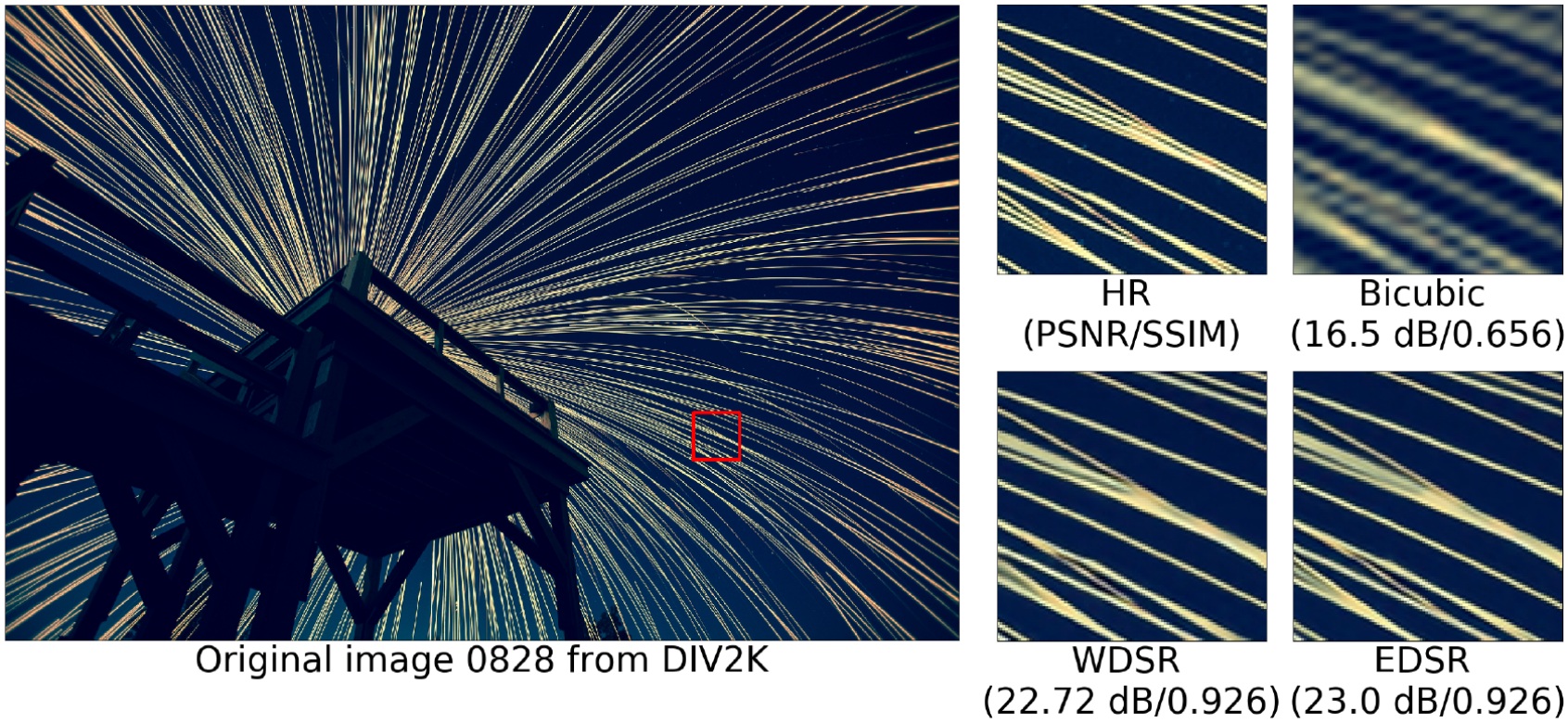


Image Results

Single Image Super Resolution



The model is trained on the **DIV2K**(*DIVerse 2K resolution high quality images*) dataset.
The dataset contains 800 high-resolution images as training set and their corresponding low-resolution ones, obtained by different down-sampling methods and different scale factors (2, 3, and 4).

Image Results

Single Image Super Resolution

Original image 0861 from DIV2K

HR
(PSNR/SSIM)

Bicubic
(21.23 dB/0.648)

WDSR
(23.17 dB/0.776)

EDSR
(23.36 dB/0.783)

The model is trained on the **DIV2K**(*DIVerse 2K resolution high quality images*) dataset.
The dataset contains 800 high-resolution images as training set and their corresponding low-resolution ones, obtained by different down-sampling methods and different scale factors (2, 3, and 4).

Image Results

Single Image Super Resolution

Original image 0845 from DIV2K

HR
(PSNR/SSIM)

Bicubic
(21.5 dB/0.669)

WDSR
(24.58 dB/0.801)

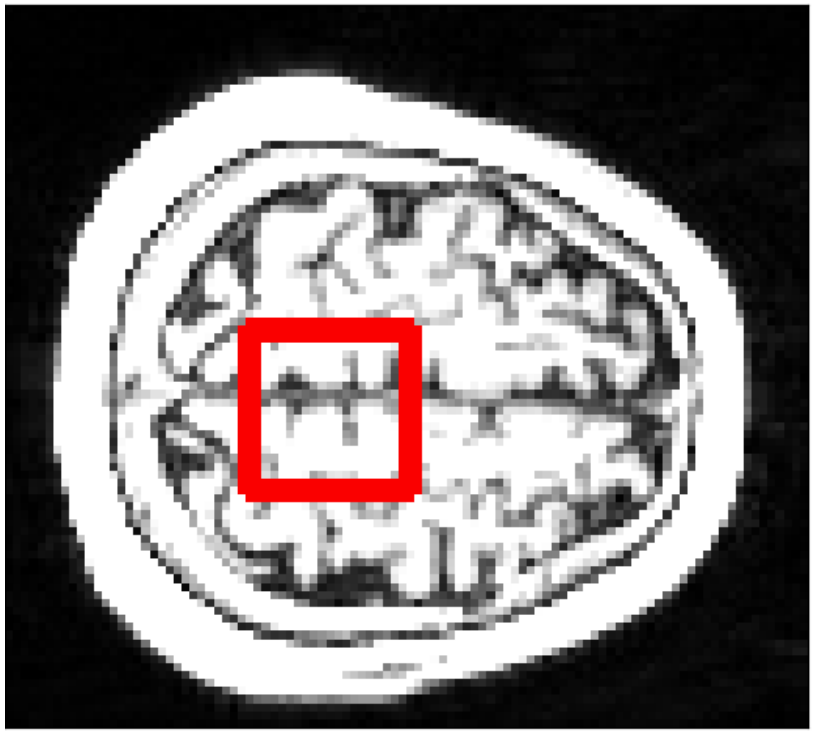
EDSR
(25.04 dB/0.811)

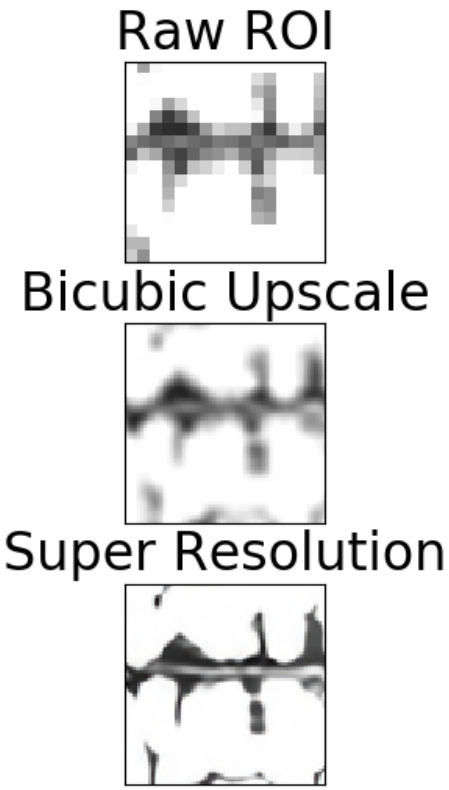
The model is trained on the **DIV2K**(*DIVerse 2K resolution high quality images*) dataset.
The dataset contains 800 high-resolution images as training set and their corresponding low-resolution ones, obtained by different down-sampling methods and different scale factors (2, 3, and 4).

Model Extrapolation

NMR Application

Test on T1 weighted NMR images ($256 \times 256 \rightarrow$ **REAL TIME**).
The images were down-sampled ($128 \times 128 \rightarrow$ 2x and $64 \times 64 \rightarrow$ 4x) and then re-up-sampled.

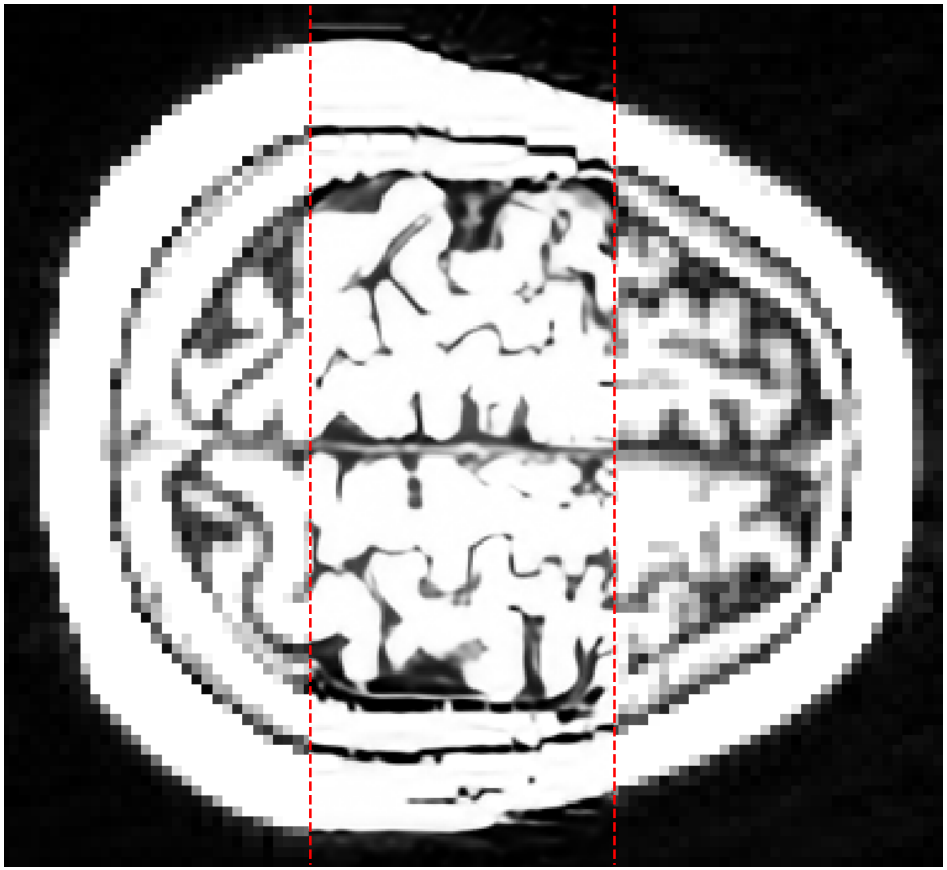




Model Extrapolation

NMR Application

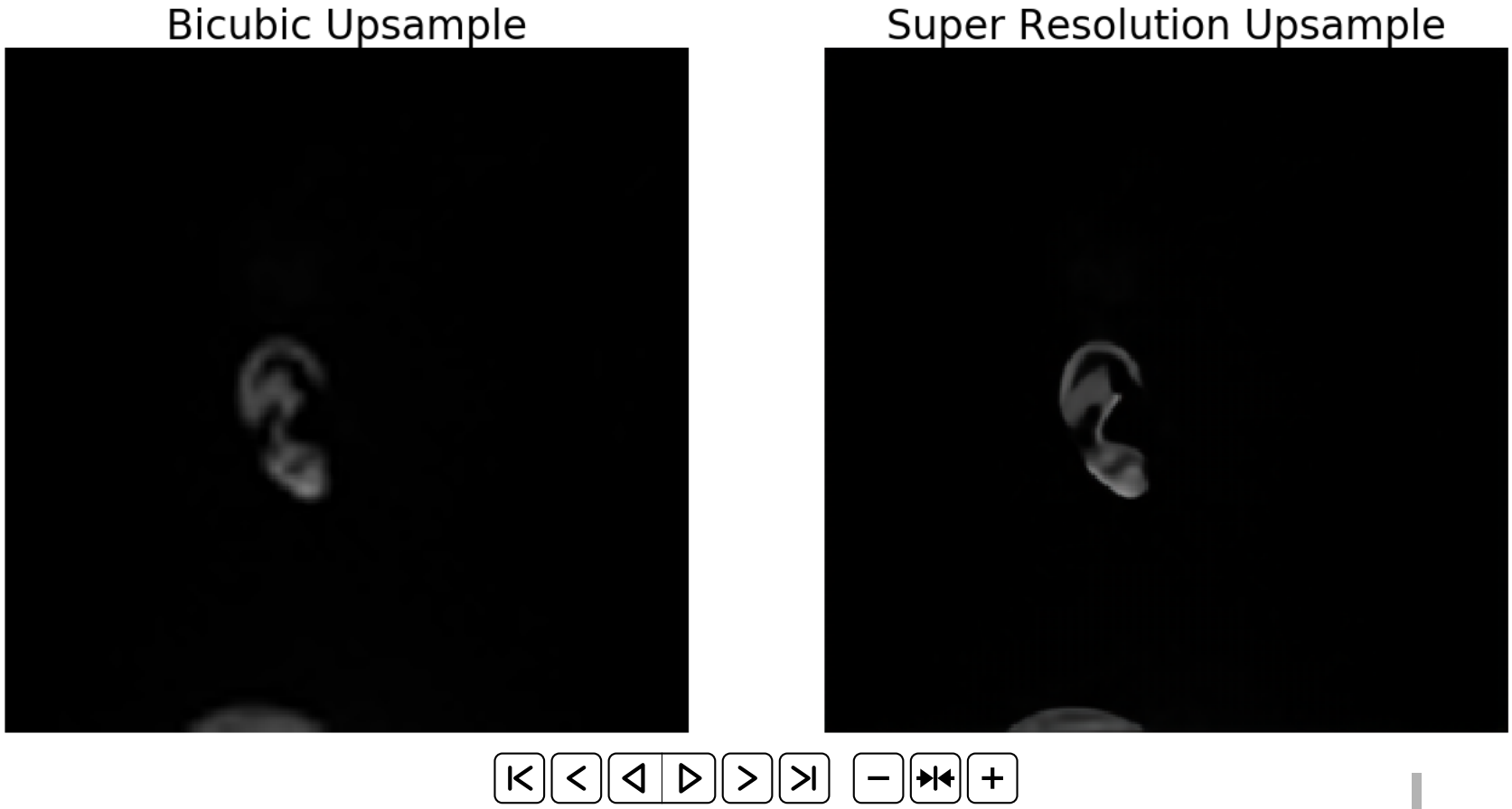
Test on T1 weighted NMR images ($256 \times 256 \rightarrow$ **REAL TIME**).
The images were down-sampled ($128 \times 128 \rightarrow$ 2x and $64 \times 64 \rightarrow$ 4x) and then re-up-sampled.



Visual score

NMR Application

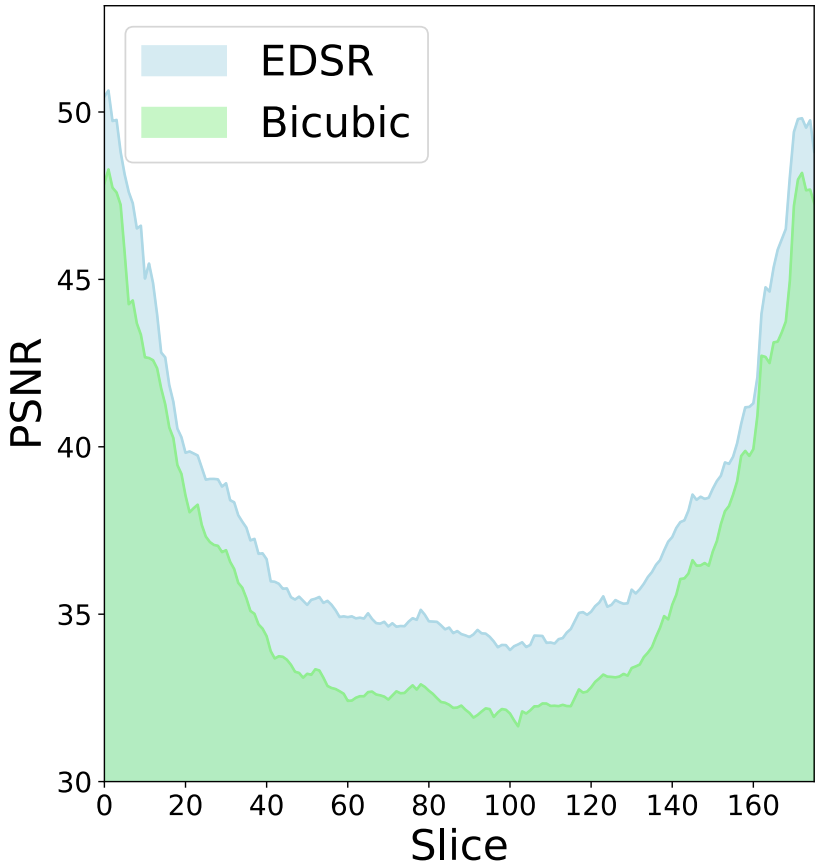
Test on T1 weighted NMR images ($256 \times 256 \rightarrow$ **REAL TIME**).
The images were down-sampled ($128 \times 128 \rightarrow$ 2x and $64 \times 64 \rightarrow$ 4x) and then re-up-sampled.

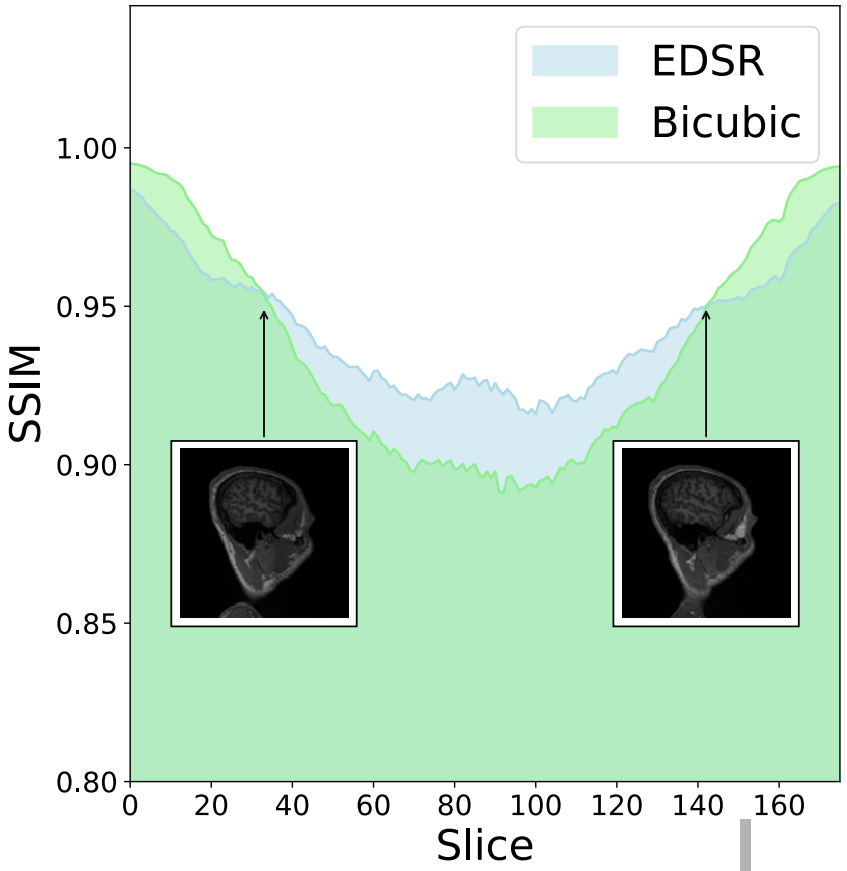


Quality score

NMR Application

Test on T1 weighted NMR images ($256 \times 256 \rightarrow$ **REAL TIME**).
The images were down-sampled ($128 \times 128 \rightarrow$ 2x and $64 \times 64 \rightarrow$ 4x) and then re-up-sampled with SR models.

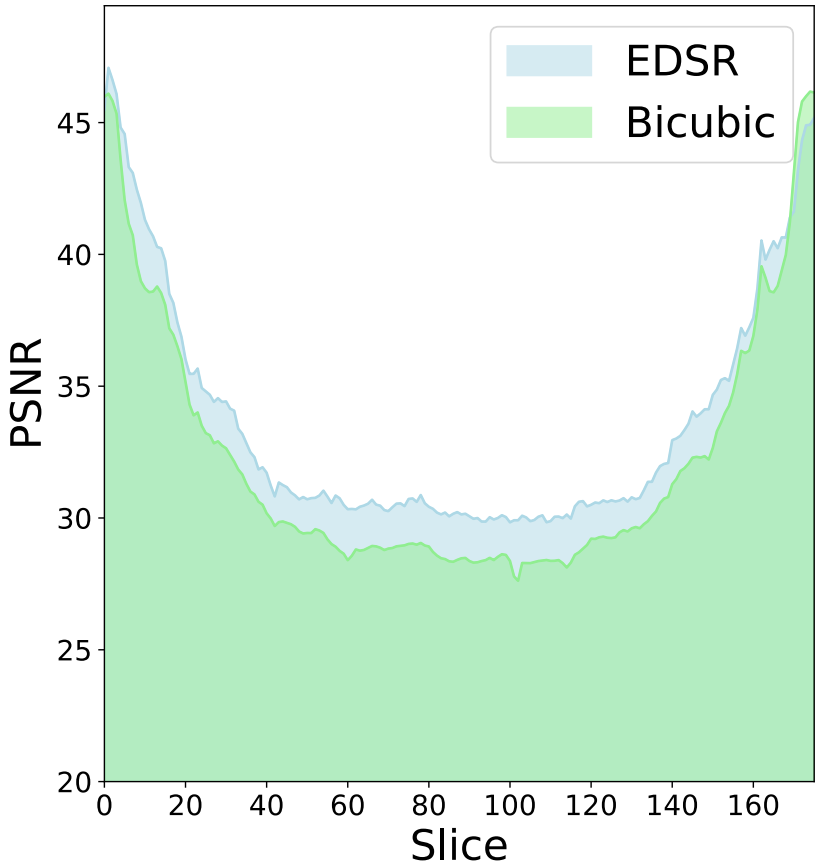


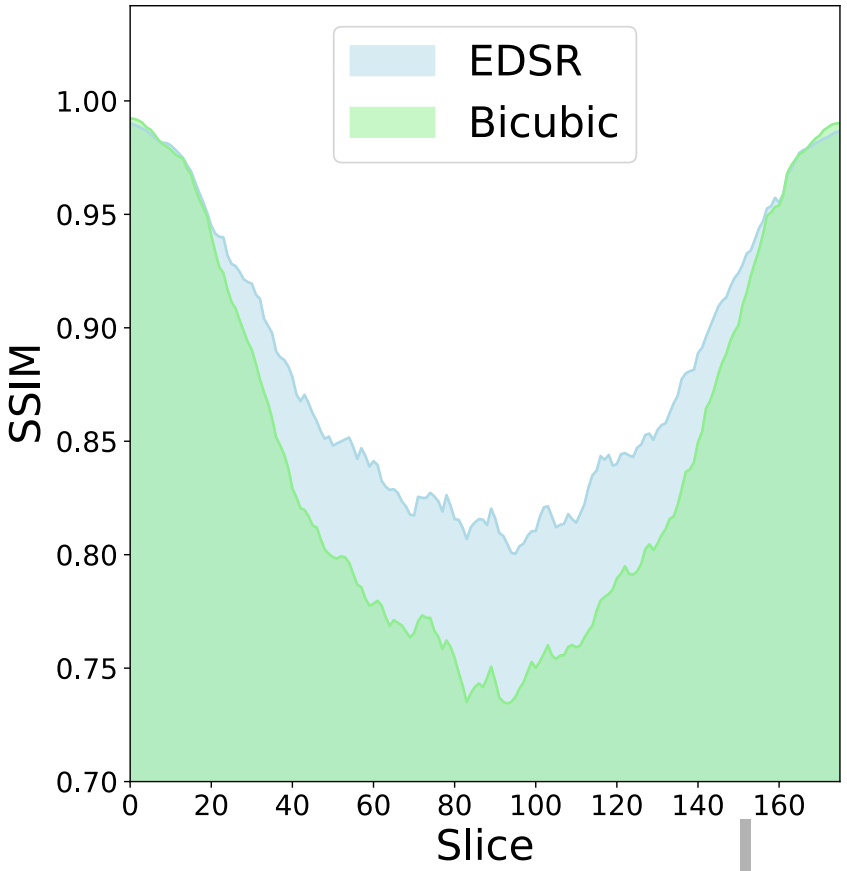


Quality score

NMR Application

Test on T1 weighted NMR images ($256 \times 256 \rightarrow$ **REAL TIME**).
The images were down-sampled ($128 \times 128 \rightarrow$ 2x and $64 \times 64 \rightarrow$ 4x) and then re-up-sampled with SR models.

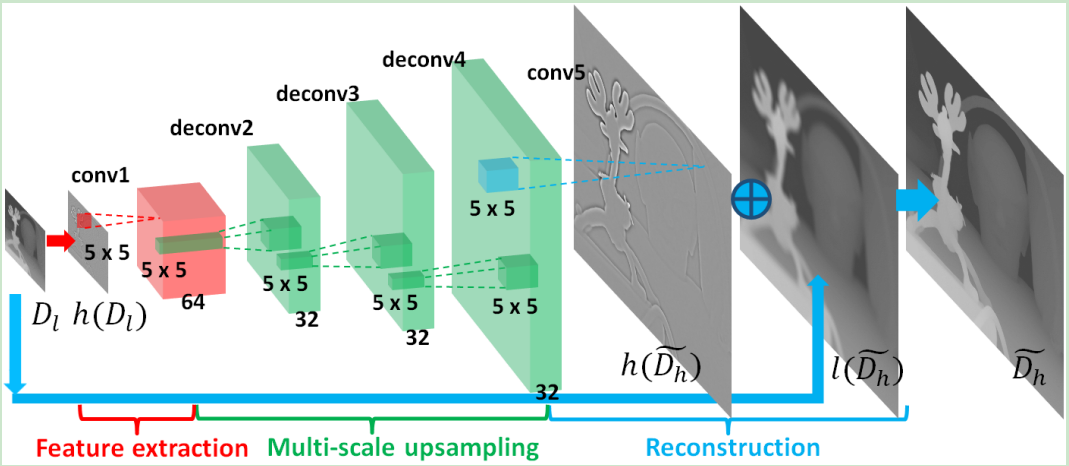




Applications

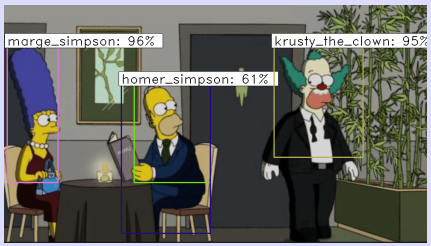
Deep Learning model tasks

Single Image Super Resolution



Improve spatial image resolution and reconstruction from low-resolution (LR) to high-resolution (HR).

Object Detection



Identify single or multiple objects into a picture or video stream.

Image Segmentation

Extract the exact pixels which belong to an object into a given image.

Object Detection

What is Object Detection?

Object detection is one of the larger deep learning sub-discipline, especially when we talk about Neural Network models.
This kind of problems aim to identify single or multiple objects into a picture or video stream.

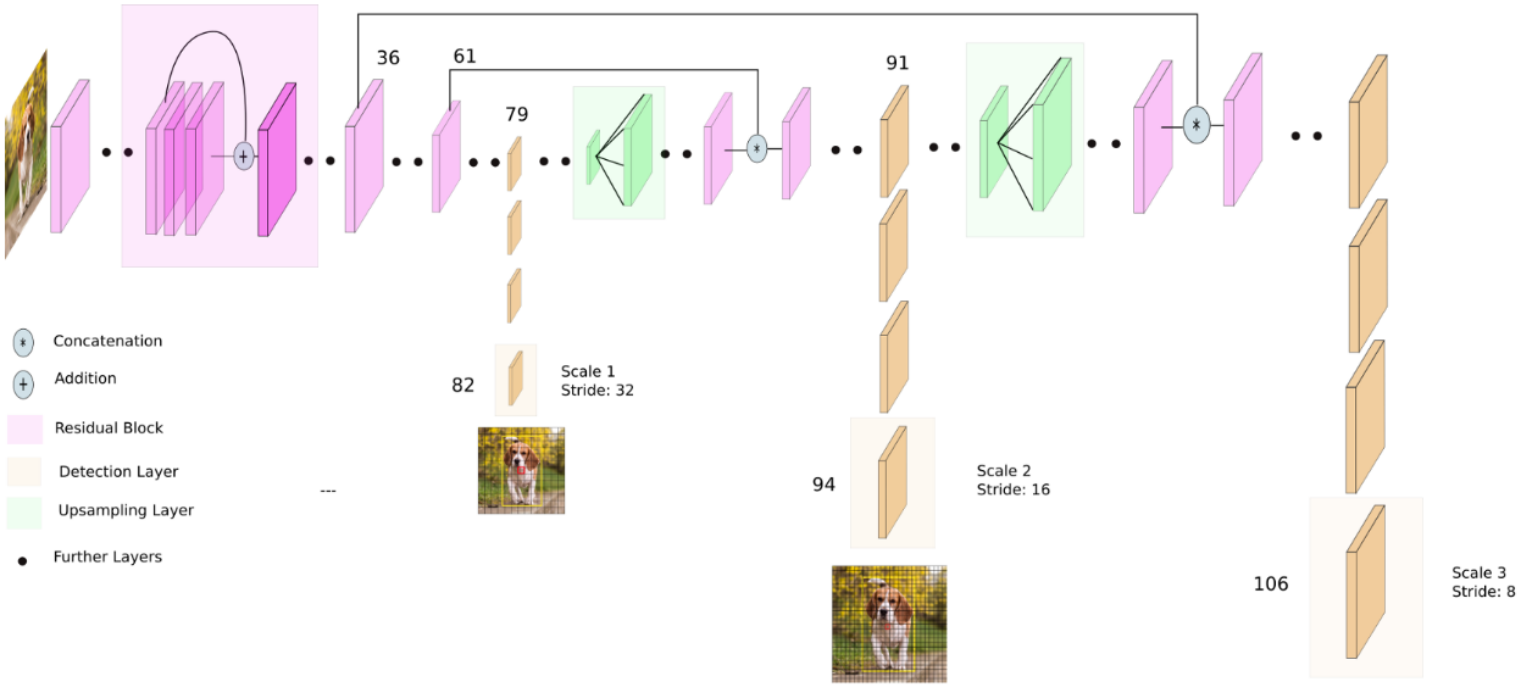
Possible applications:

- **Object tracking**;
- Video surveillance;
- **Pedestrian detection**;
- Anomaly detection;
- **People counting**;
- Self-driving cars;
- Face detection;
- . . . ;

YOLO - You Only Look Once

Object Detection Model

YOLO is a deep Neural Network model with more than 100 layers and more than 62 million of parameters.
The original implementation of the model is provided by Joseph Redmon (https://github.com/pjreddie/darknet) into the darknet project (ANSI C).

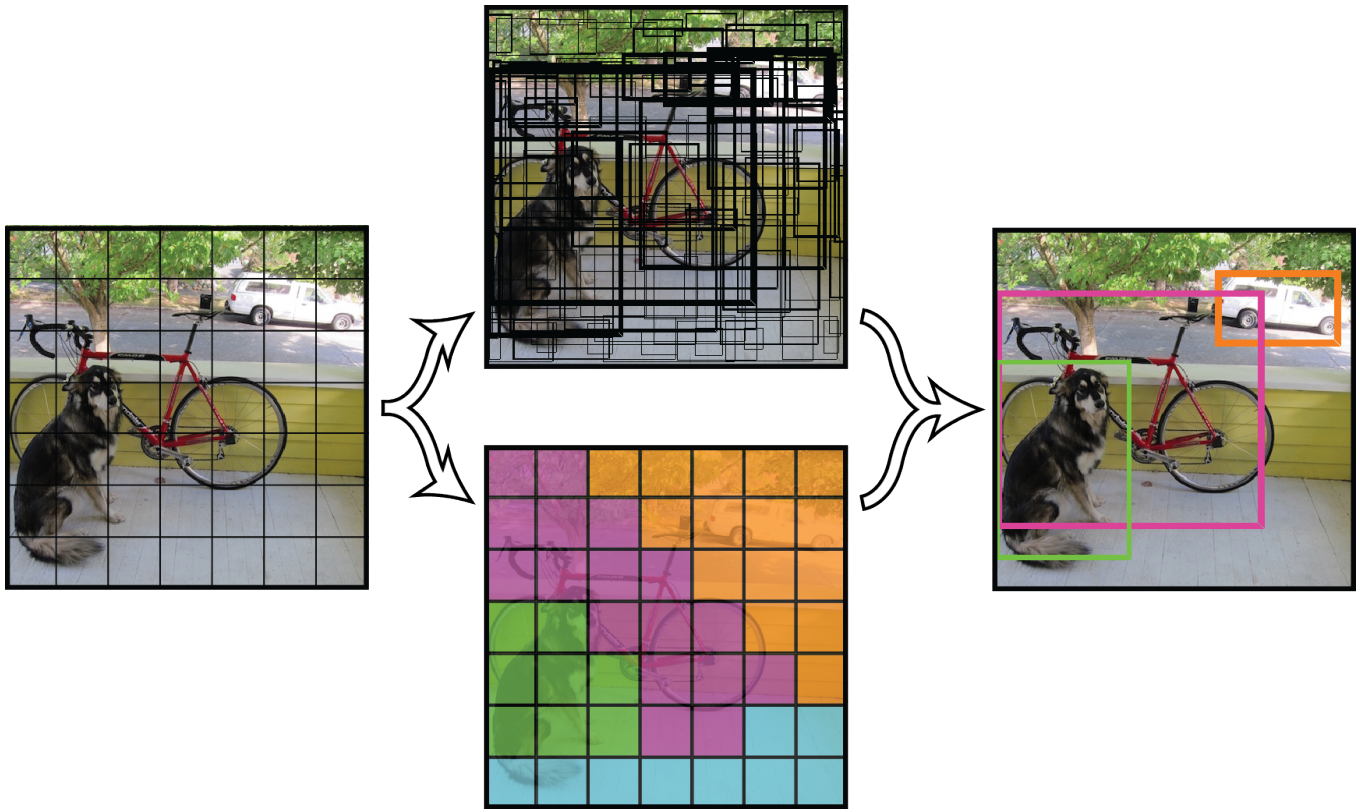


YOLO v3 network Architecture

YOLO - You Only Look Once

Object Detection Model

YOLO is a deep Neural Network model with more than 100 layers and more than 62 million of parameters.
The original implementation of the model is provided by Joseph Redmon (https://github.com/pjreddie/darknet) into the darknet project (ANSI C).

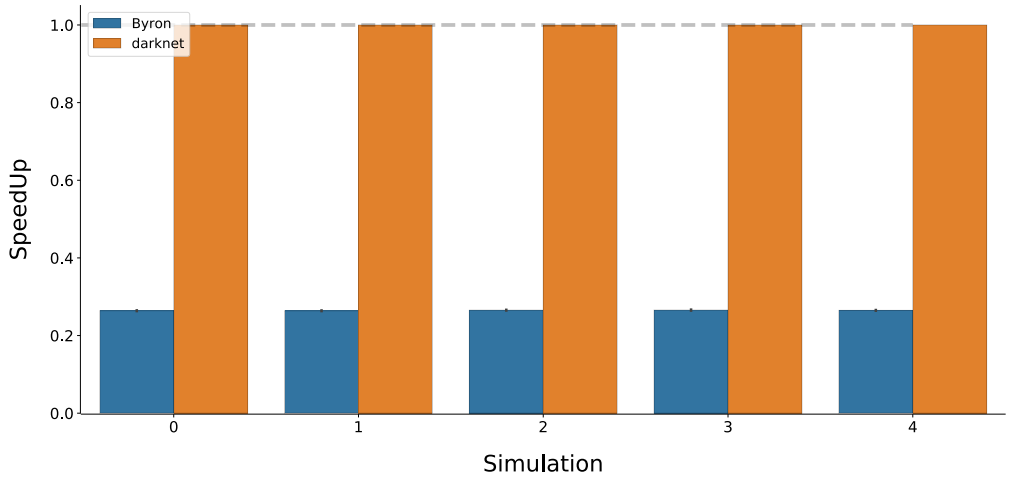


Computational Performances

Object Detection Models

Byron is inspired to darknet and its first implementation was about object detection task.
darknet → benchmark → but its efficiency is based on GPU support!

Evaluation performed on 100 runs without multi-threading!



Framework	Average Time (s)
Byron	5.18
darknet	19.58

Time evaluated on 416×416 images.
Environment: Bioinformatics server grade machine (128 GB RAM memory and 2 CPU E5-2620, with 8 cores each.

Model Extrapolation

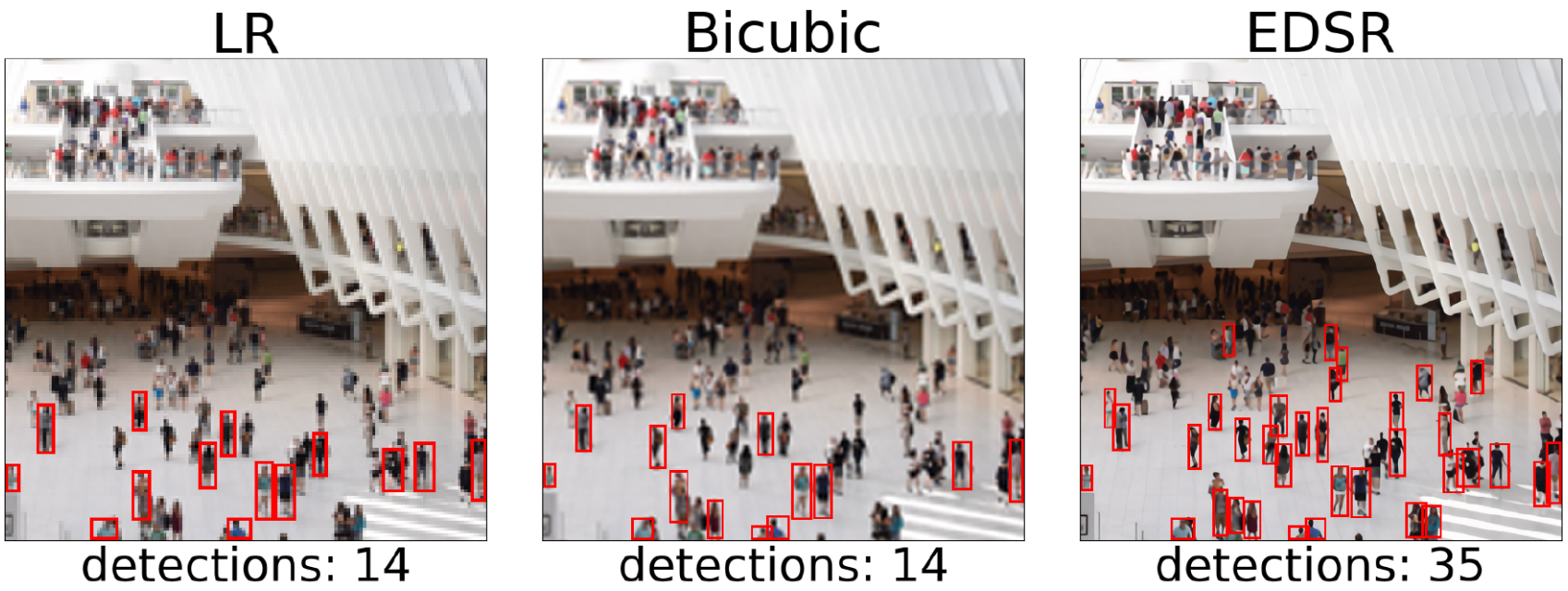
People counting Application

Despite the model is incredibly efficient in object detection also with low quality images, there is a sort of **limit** in the number of pixels needed for object identification. We had the opportunity to empirically verify its limit working on a people tracking project for real time applications.
The project was developed in collaboration with the Complex Systems (**PhySyCom**) group.

Model Extrapolation

People counting Application

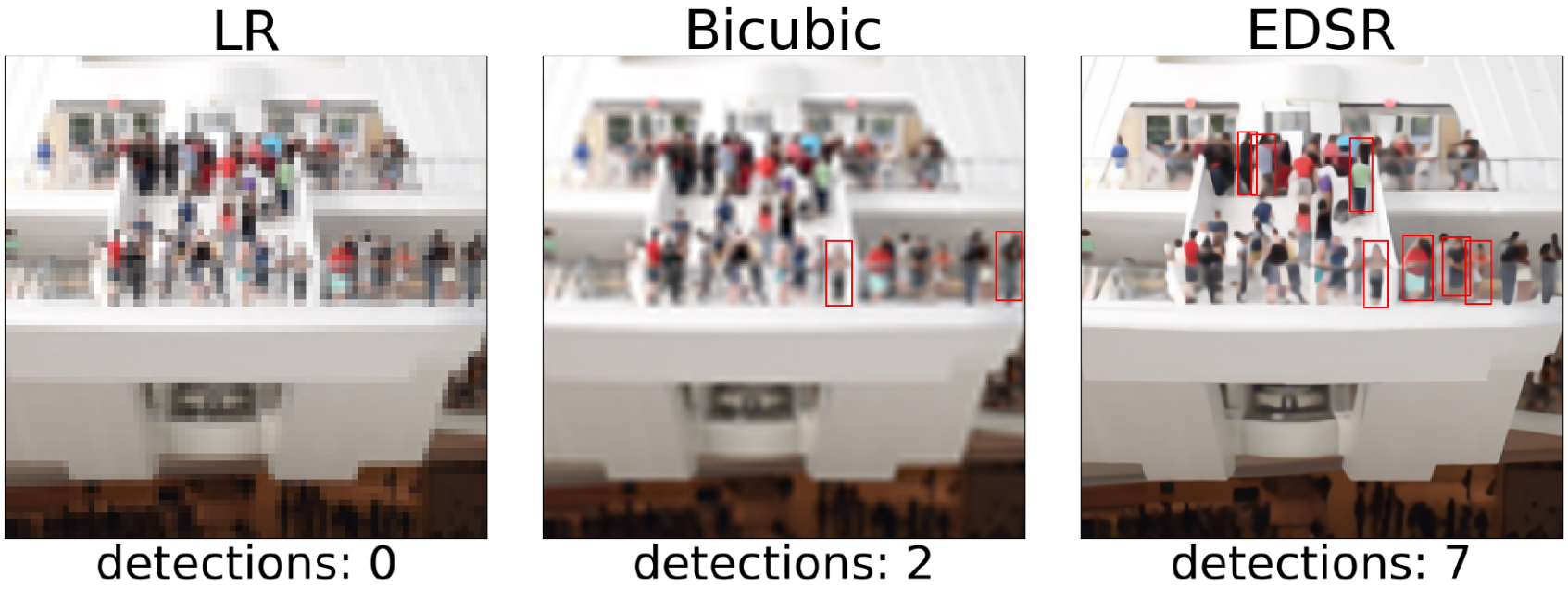
Despite the model is incredibly efficient in object detection also with low quality images, there is a sort of **limit** in the number of pixels needed for object identification. We had the opportunity to empirically verify its limit working on a people tracking project for real time applications.
The project was developed in collaboration with the Complex Systems (**PhySyCom**) group.



Model Extrapolation

People counting Application

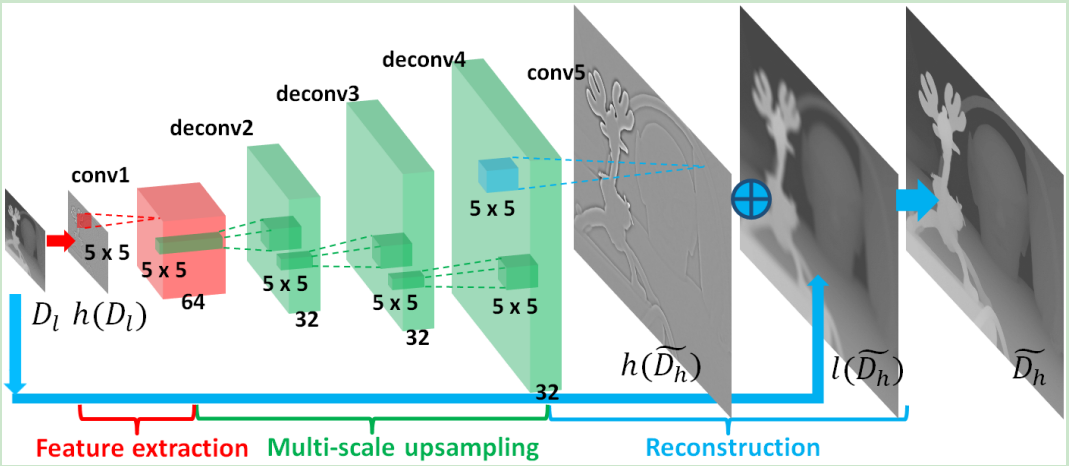
Despite the model is incredibly efficient in object detection also with low quality images, there is a sort of **limit** in the number of pixels needed for object identification. We had the opportunity to empirically verify its limit working on a people tracking project for real time applications.
The project was developed in collaboration with the Complex Systems (**PhySyCom**) group.



Applications

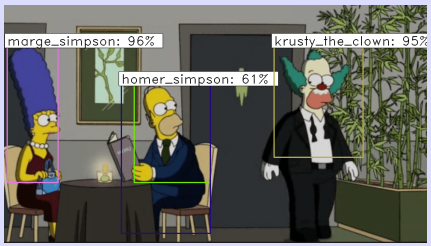
Deep Learning model tasks

Single Image Super Resolution



Improve spatial image resolution and reconstruction from low-resolution (LR) to high-resolution (HR).

Object Detection



Identify single or multiple objects into a picture or video stream.

Image Segmentation

Extract the exact pixels which belong to an object into a given image.

Image Segmentation

What is Image Segmentation?

Object Detection: give a label to **ROIs** of the input image.
Image Segmentation: give a label to **each pixel** of the input image.

3D Femur Structure

Rizzoli Hospital project

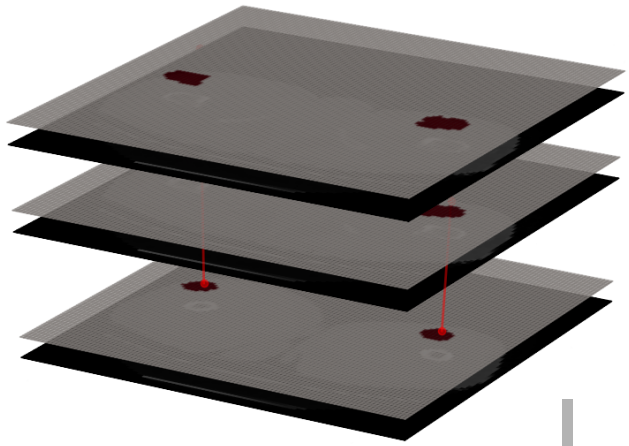
Improve the identification and segmentation of the femoral head, trying to discriminate this part of the bone from the articular cartilage and moreover from the acetabular fossa.

Data type: CT (*Computer Tomography*) images.

Dataset: 4 (not-annotated) patients.

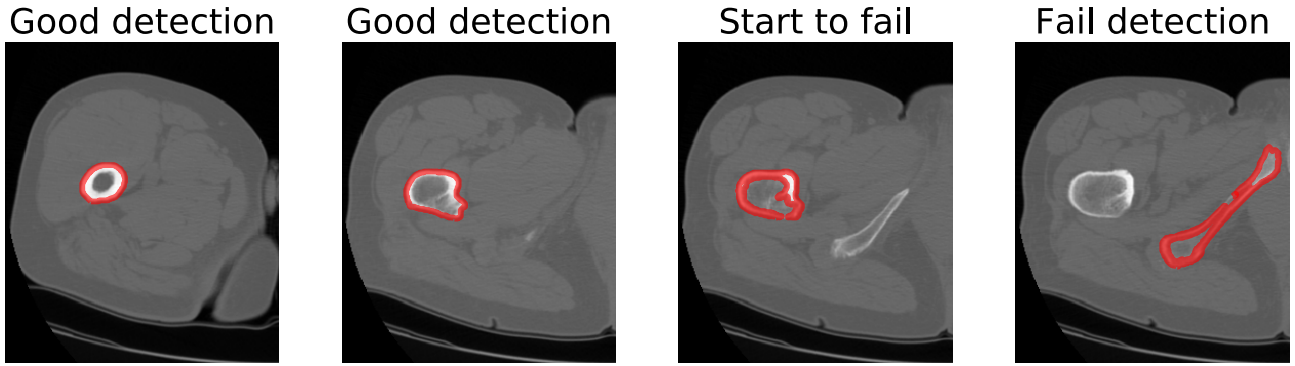
Semi-automatic pipeline:

- Edge enhancement;
- Thresholding;
- Morphological operations;
- Connected Components;
- Interpolation between consecutive frames;

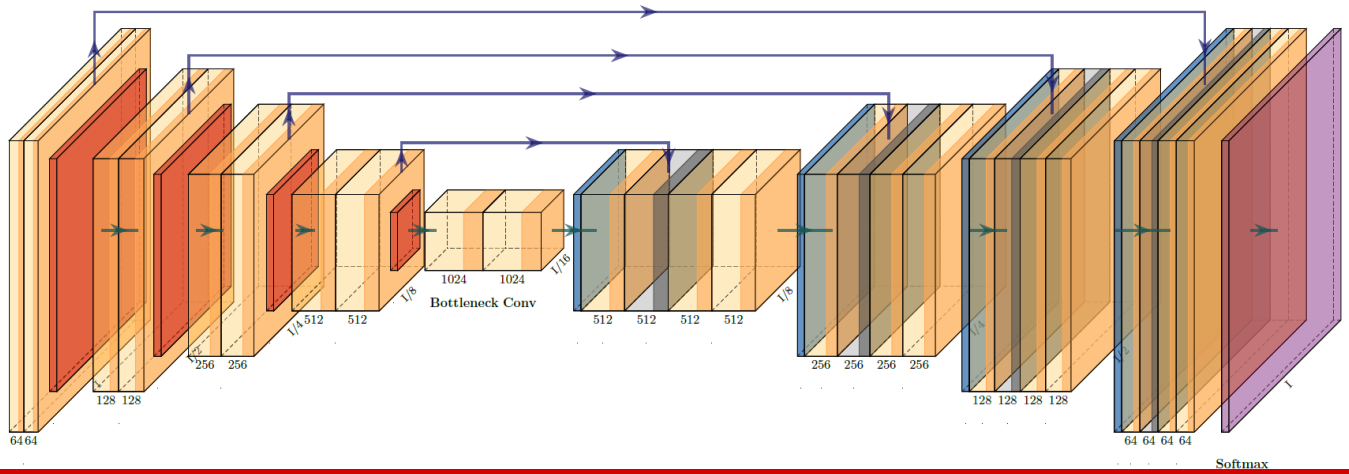


Segmentation Models

The method is too naive to perform a good segmentation on the full set of slices. However, it can be useful to **reduce** the quantity of slices to annotate manually ($\sim$ 400 slices per patient)!



Thus, we moved to a **Neural Network model**

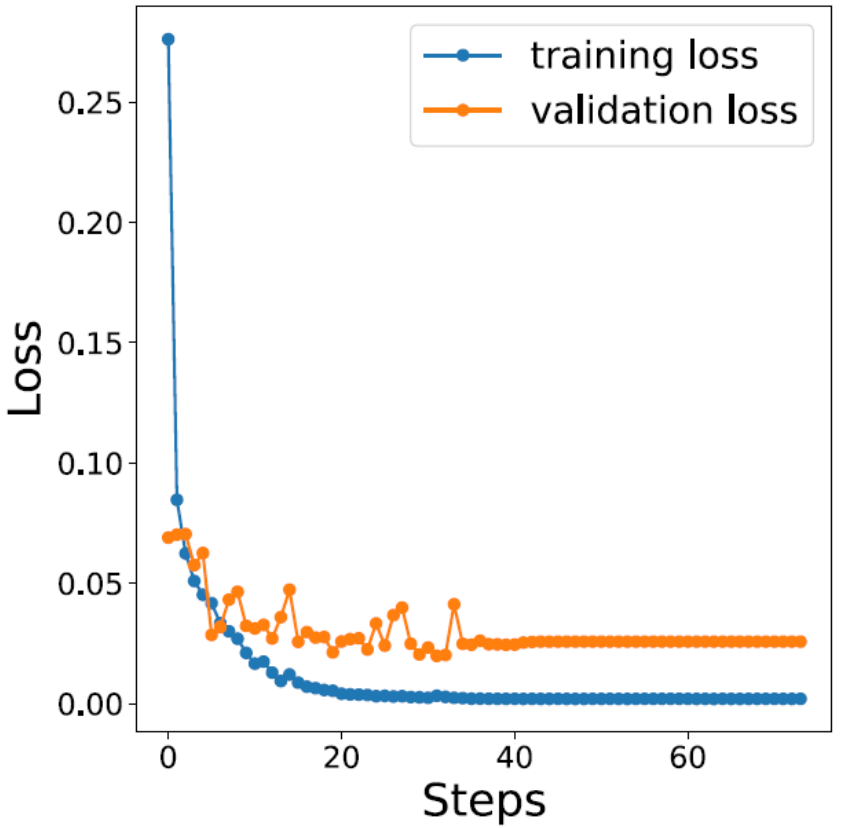


U-Net Model

Segmentation Models

Training parameters:

- Training Set: 96 images;
- Test Set: 40 images;
- Data augmentation: Rotated, shifted, mirrored.
- 40 epochs;
- loss: binary cross-entropy
- metrics: IoU (*Intersection over Union*)

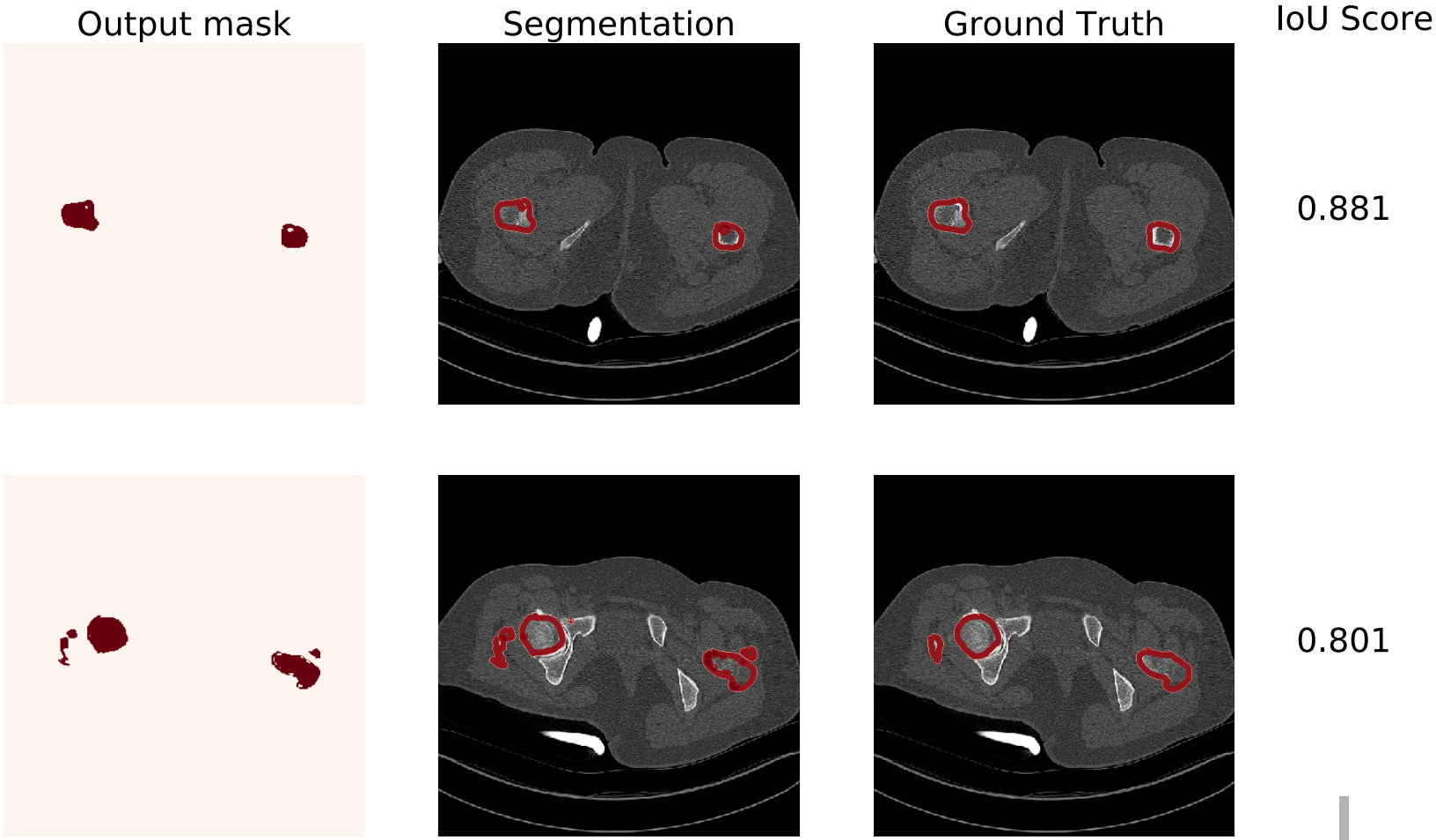


IoU Validation Set:
The 80% of the test set has obtained a IoU score greater than 0.8 and thus a good correspondence between our results and the ground truth.

IoU score

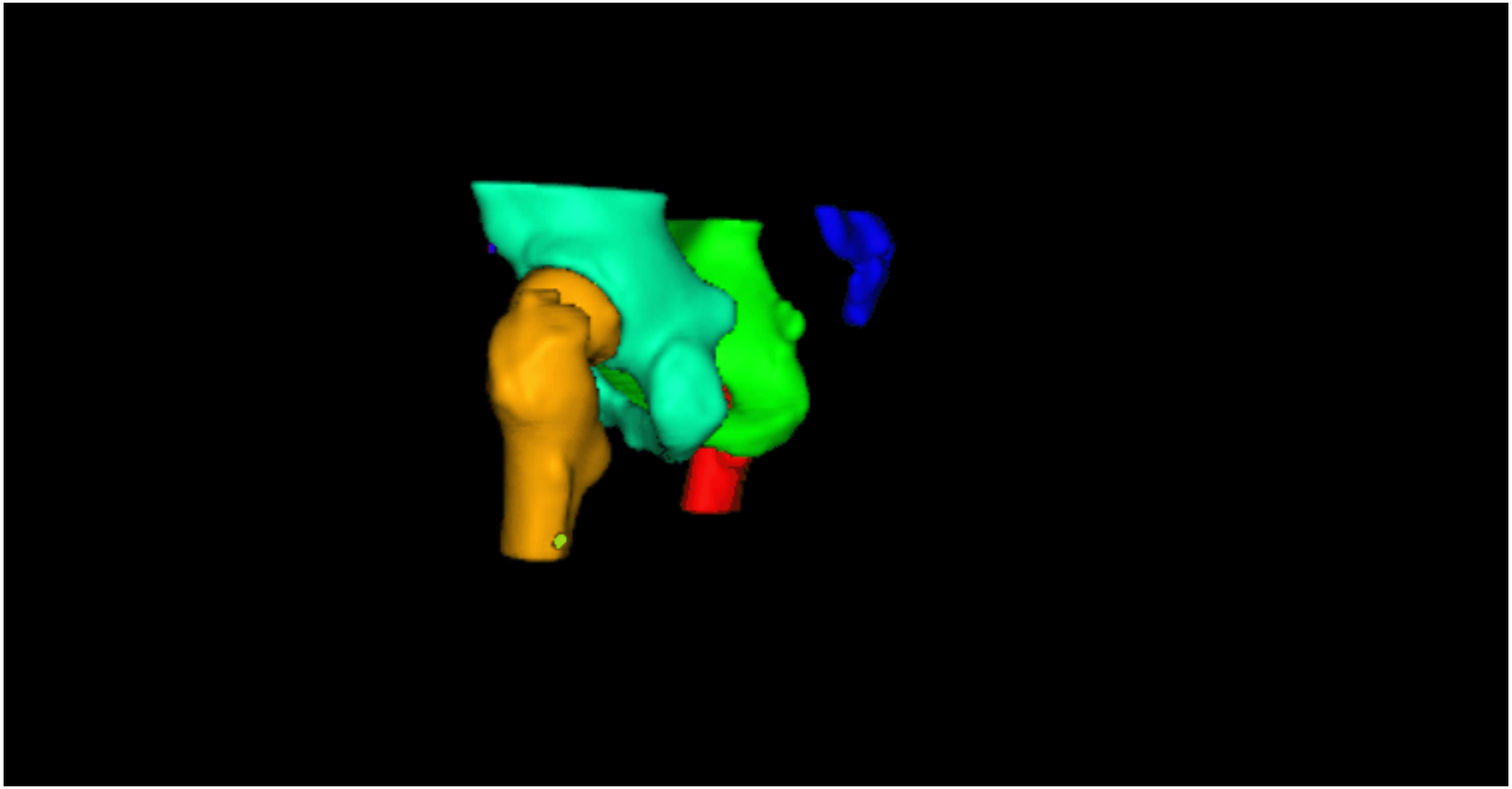
Segmentation Models

The model output were filtered using a thresholding of 10^{-2}, i.e values $\leq$ were turned off.



3D Reconstruction

Segmentation Models



Conclusion

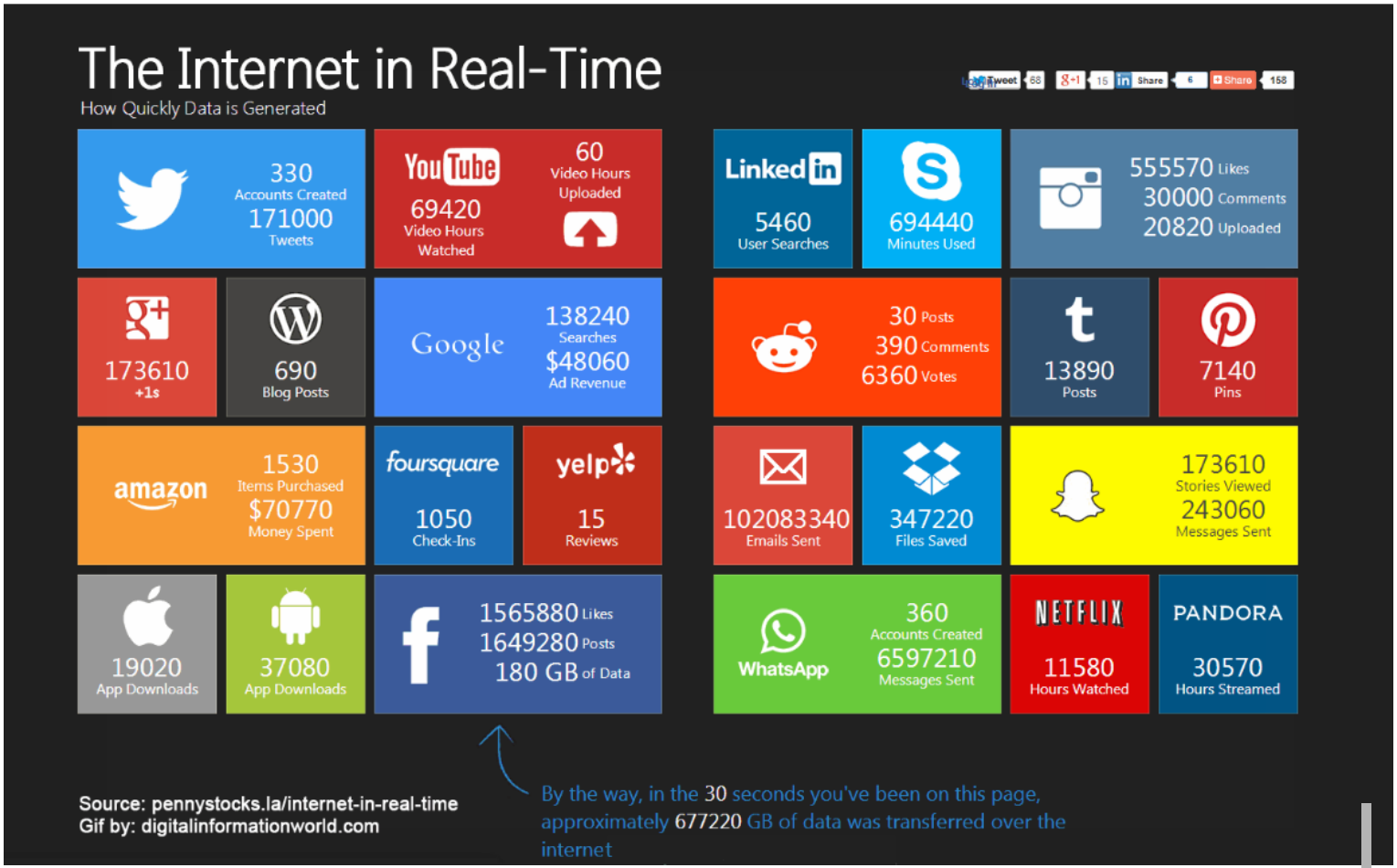
Deep Learning applications

- Two new libraries for Deep Learning applications were proposed:
  - NumPyNet: a tool for educational practice and modeling;
  - Byron: a tool for an efficient implementation of these models;
- Both libraries implement cutting edge numerical techniques to maximize computation efficiency;
- Byron optimize and extend the darknet project with several fixes and computational improvements.
- Byron is able to join in a single framework either super resolution and object detection tasks.
- The results obtained using Super Resolution and Segmentation models are very promising for the analysis of biomedical images.
- We have shown that deep learning models are capable of a very efficient generalization due to their vast amount of parameters and a well-programmed training section.
- The Byron YOLO version is approximately 3.8x faster than darknet in all the simulations;

The Big Data era

Data generation in Real Time

Big Data are the new microscopy to "measure" the society.



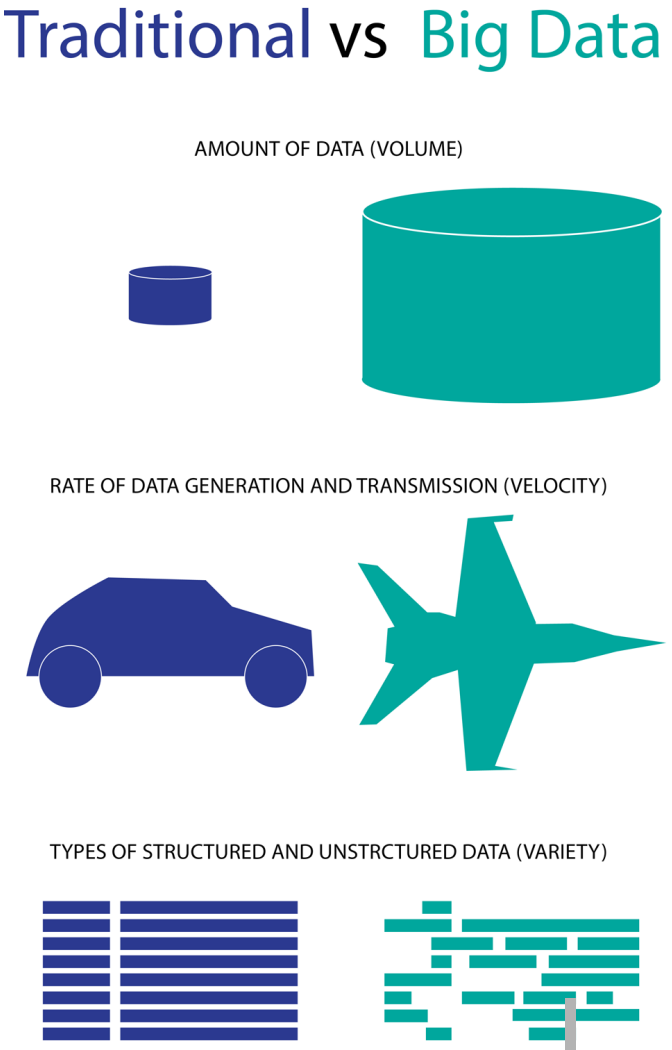
The 5 V's of Big Data

Big Data characteristics

- **Volume**: Terabytes to exabytes of existing data to process.
- **Velocity**: Streaming data, requiring msec to respond.
- **Variety**: different types of data we can now use

$$\rightarrow \begin{cases} Structured\ data \\ Semi-Structured\ data \\ Unstructured\ data \end{cases}$$

- **Veracity**: Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception.
- **Value**: It is all well and good having access to big data but unless we can turn it into value it is useless.



Structured data Vs Semi-Structured Data

Data formats

Structured Data


```
{
"HMDB0000001332-80-9": {
  "Location": [
    "Blood",
    "Cerebrospinal Fluid (CSF)",
    "Feces",
    "Saliva",
    "Urine"
  ],
  "name": "1-Methylhistidine",
  "Synonyms": [
    "(2S)-2-amino-3-(1-Methyl-1H-imidazol-4-yl)propanoic acid",
    "1-Methylhistidine",
    "Pi-methylhistidine",
    "(2S)-2-amino-3-(1-Methyl-1H-imidazol-4-yl)propanoate",
    "1 Methylhistidine",
    "1-Methyl histidine",
    "1-Methyl-histidine",
    "1-Methyl-L-histidine",
    "1-MHis",
    "1-N-Methyl-L-histidine",
    "L-1-Methylhistidine",
    "N1-Methyl-L-histidine",
    "1-Methylhistidine dihydrochloride"
  ],
  "SuperClass": "Organic acids and derivatives",
  "Class": "Carboxylic acids and derivatives",
  "Pathway": [
    "Histidine Metabolism"
  ],
```

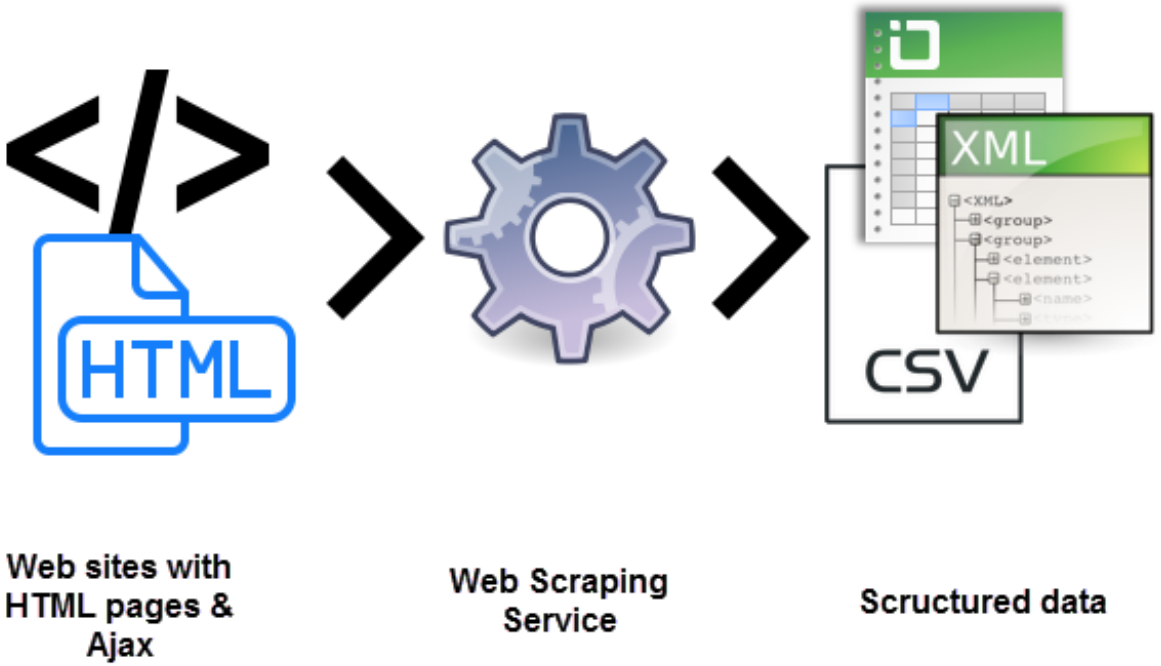

UnStructured Data


```
'<script>(window.NREUM||(NREUM={}
  )).loader_config={xpid:"VQUGVVJVGwQEUFRWBQ=="};window.NREUM||(
  ),__nr_require=function(t,n,e)'
'{function r(e){if(!n[e]){var o=n[e]={exports:{}};t[e][0].call(o
  o=t[e][1][n];return r(o||n)},o,o.exports)}'
'return n[e].exports}if("function"==typeof __nr_require)return _
  <e.length;o++)r(e[o]);return r}'
'({1:[function(t,n,e){function r(t){try{s.console&&console.log(t
  o,i=t("ee"),a=t(15),s={};try{o=localStorage.getItem("__nr_flag
'.split(","),console&&"function"==typeof console.log&&(s.console
&&(s.dev=!0),o.indexOf("nr_dev")!==-1&&(s.nrDev=!0))}'
'catch(c){}s.nrDev&&i.on("internal-error",function(t){r(
  t.stack)}),s.dev&&i.on("fn-err",function(t,n,e){r(e.stack)}),s
'DEVELOPMENT MODE"),r("flags: "+a(s,function(t,n){return t}).joi
  ],2:[function(t,n,e){function r(t,n,e,r,s){try{p?p-=1:o(s||new
'UncaughtException(t,n,e),!0)}catch(u){try{i("ierr",[u,c.now(),!
  }return"function"==typeof f&&f.apply(this,a(arguments))}functi
'UncaughtException(t,n,e){this.message=t||"Uncaught error with n
  information",this.sourceURL=n,this.line=e}function o(t,n){var
'e=n?null:c.now();i("err",[t,e])}var i=t("handle"),a=t(16),s=t("
os"),f=window.onerror,d=!1,l="nr@seenError",p=0;'
'c.features.err=!0,t(1),window.onerror=r;try{throw new Error}cat
  h&&(t(5),t(4),"addEventListener"in window&&t(3),c.xhrWrappable
'&&t(6),d=!0)}s.on("fn-start",function(t,n,e){d&&(p+=1)}),s.on("
  &&!e[1]&&(u(e,l,function(){return!0}),this.thrown=!0,'
'o(e))}),s.on("fn-end",function(){d&&!this.thrown&&p>
  0&&(p-=1)}),s.on("internal-error",function(t){i("ierr",[t,c.no
  ],3:[function'
```

Web Scraping Algorithm

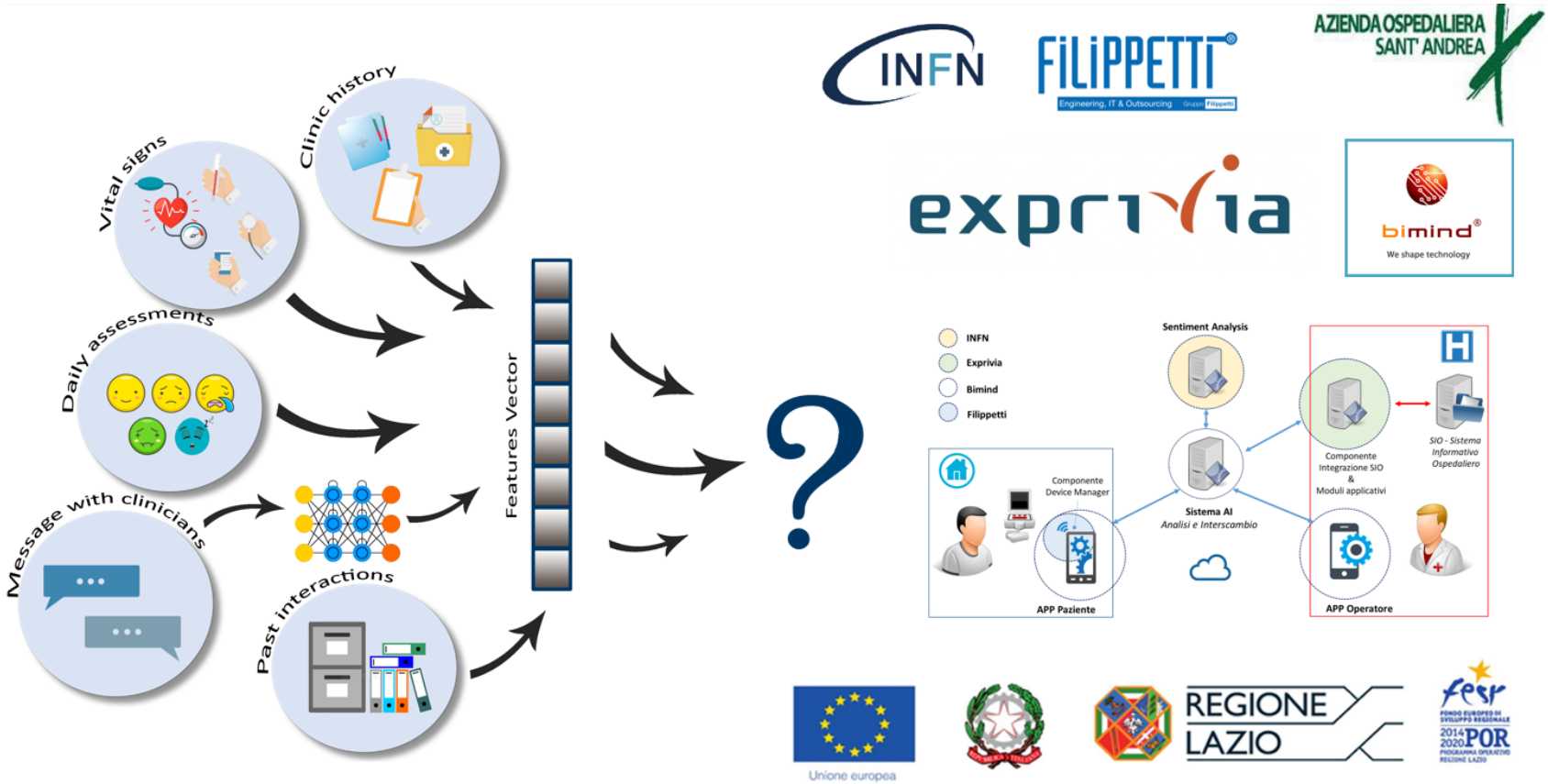
Mining the Web

Web scraping, **web harvesting**, or web data extraction is data scraping used for extracting data from websites.



FiloBLU Project

A INFN project



Disease Ontology

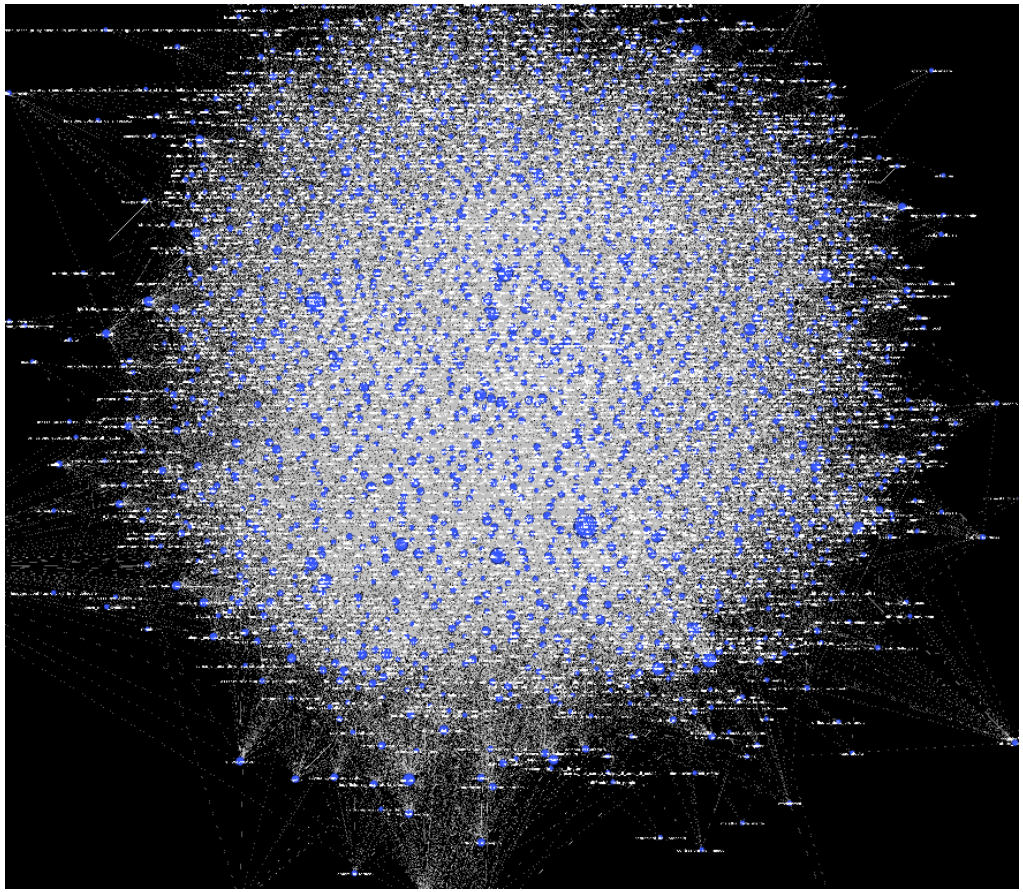
Sympotms Network

Mypersonaltrainer

- 1381 nodes;
- 14002 links;

SAPERE.IT

- 1200 nodes;
- 16035 links;



Symptoms Network:

- **nodes:** 2285;
- **links:** 29557;

Why CHIMeRA?

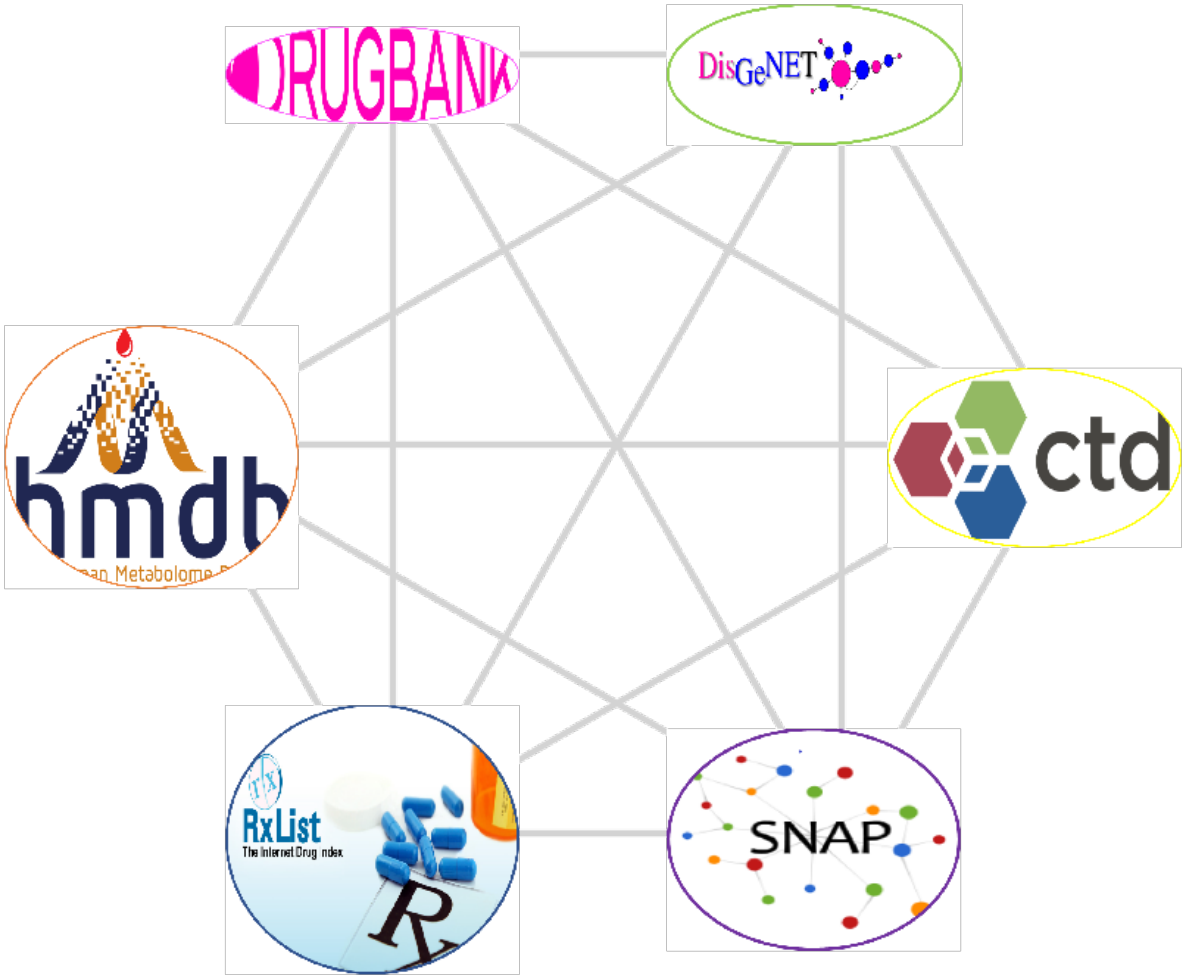
Issues

- Abysmal overlap between existing databases using medical ontology.
- Difficulties in propagating existing information between bipartite networks containing diseases.
- To find a way to let different type of information "talk" to each other through diseases.

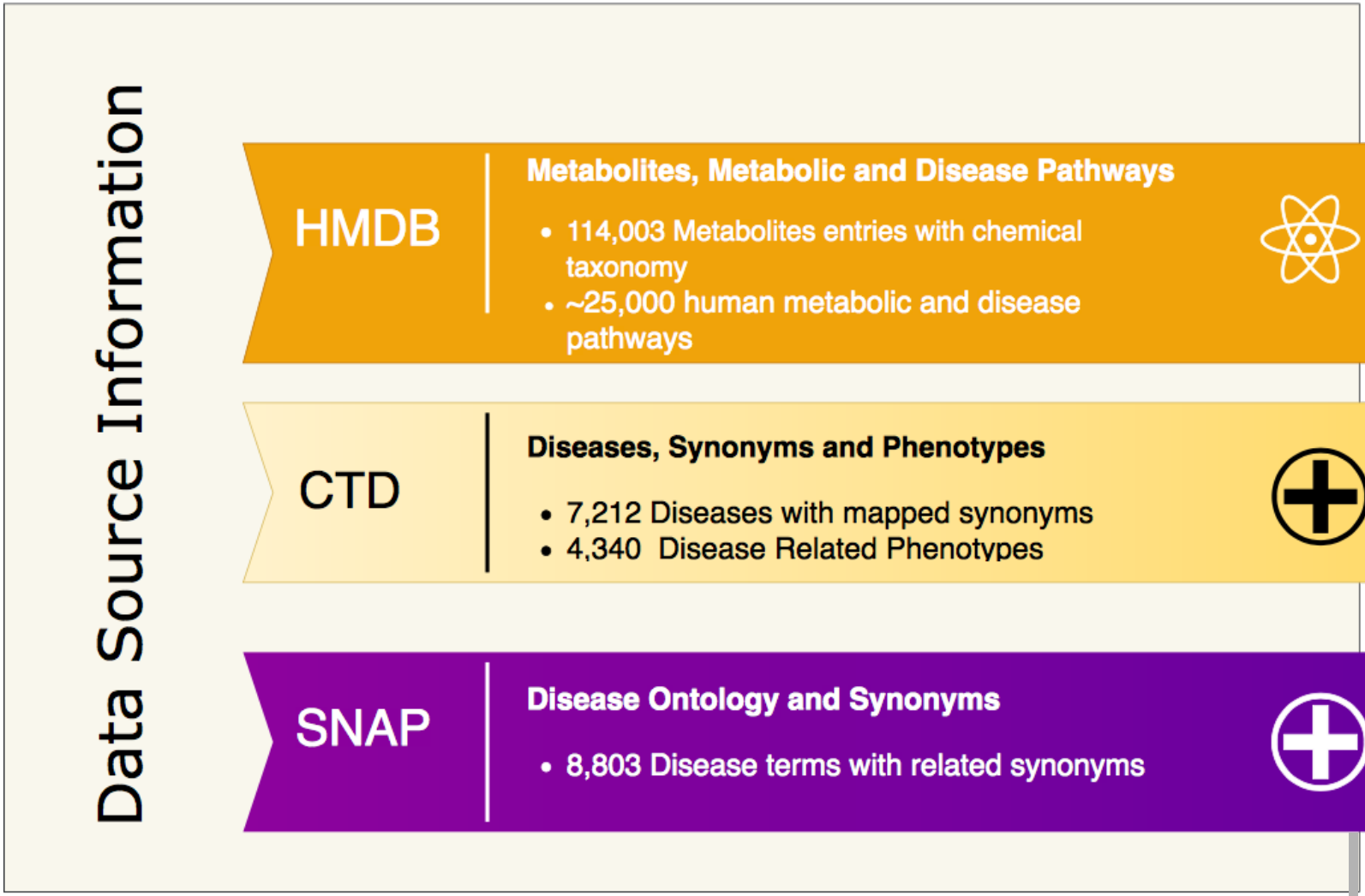
CHIMeRA Structure

Data sources interaction

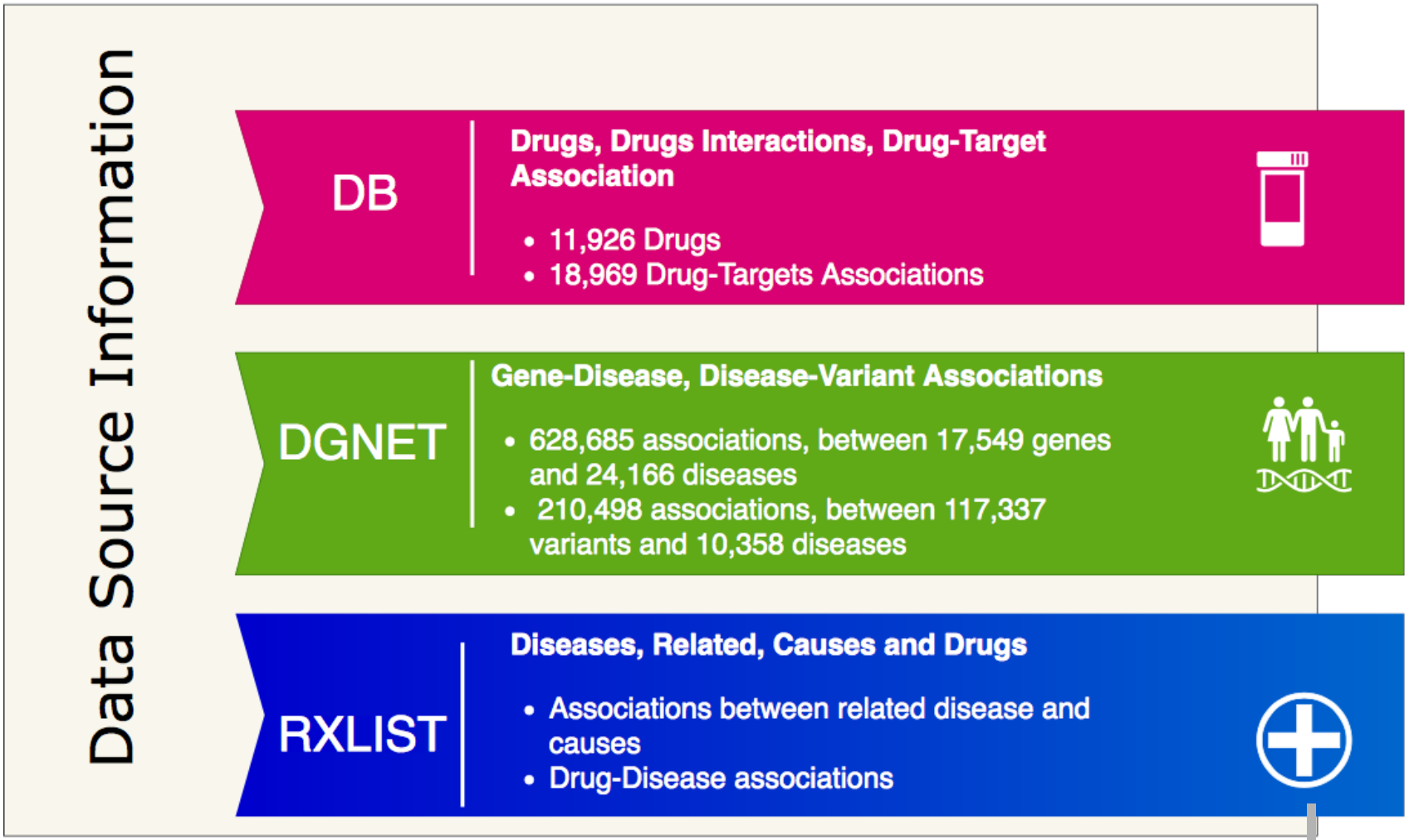
The 6 most popular bio-medical databases:



Database contents

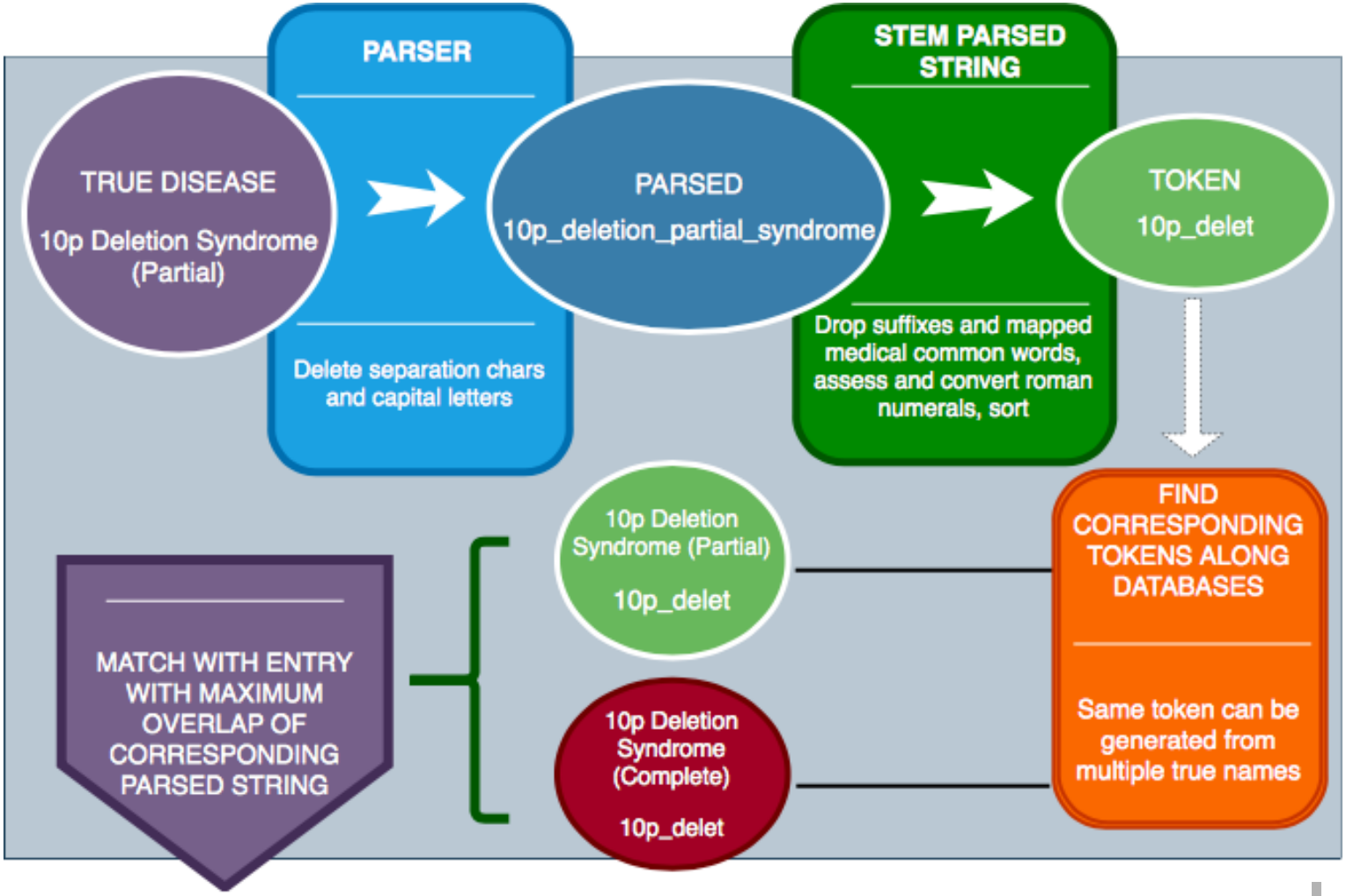


Database contents



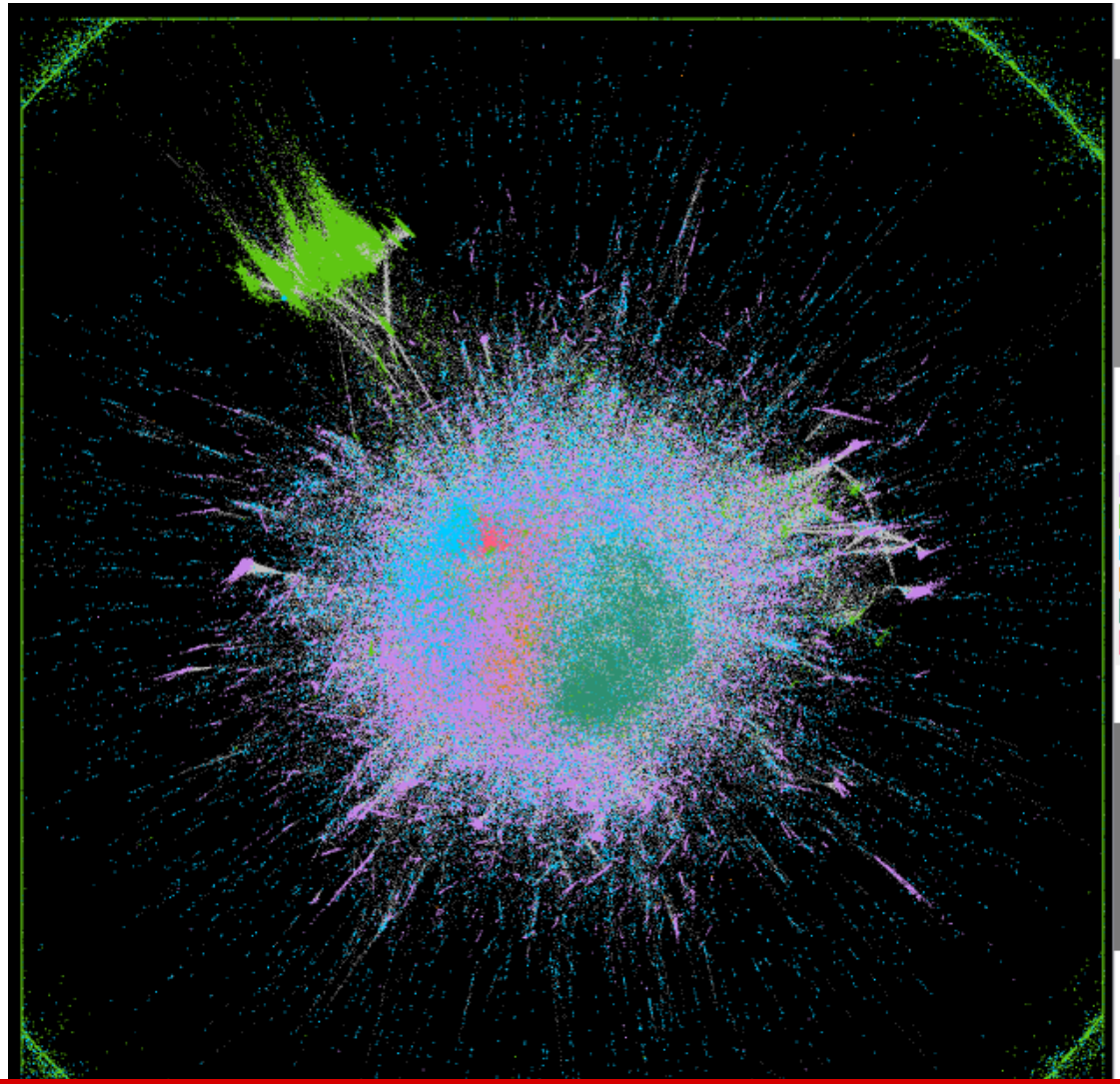
String Processing

Pre-processing Pipeline



CHIMeRA

Network visualization



Node Overlap

- CTD-SNAP = 24.17%
- RXLIST-HMDB = 8.03%
- SNAP-HMDB = 0.39%
- DGNET-RXLIST = 19.78%

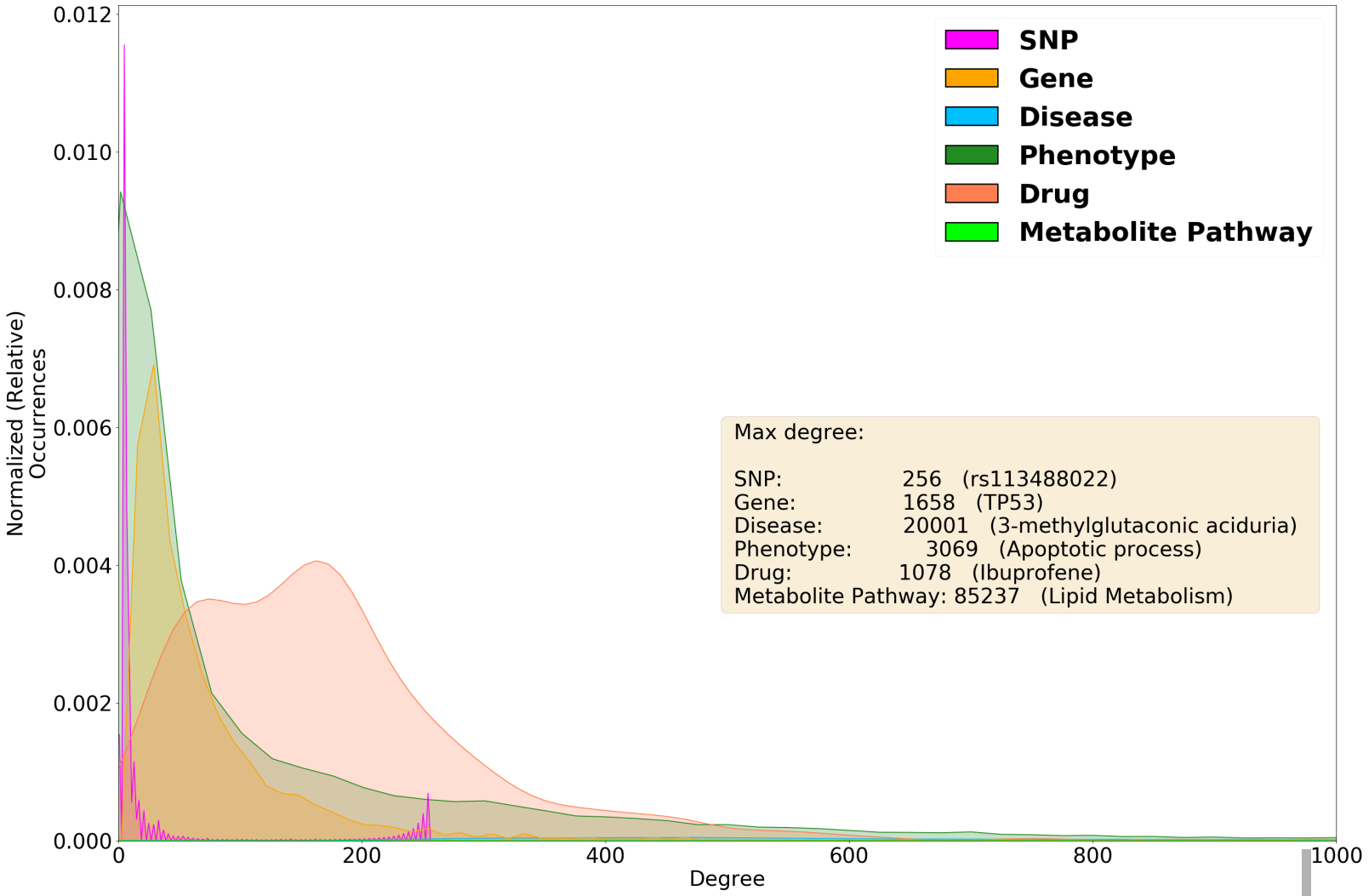
variant	(36,01%)
Metabolite_Pathway	(35,21%)
disease	(19,1%)
gene	(5,38%)
phenotype	(4,05%)
drug	(0,25%)

 # Nodes ~ 3x10^5

 # Edges ~ 2x10^6

Preliminary Analysis

Degree distributions



CHIMeRA as Service

CHIMeRA database: $\sim 3.6 \times 10^5$ nodes, $\sim 3.8 \times 10^7$ links

- Convert Network structure to quearable DataBase (**ArangoDB**);
- Search node according to its type (disease, gene, snp, ...):
  - Star graph: equivalent to single database search;
  - Breath first search: percolation from a given root node;
  - Multiple roots search: features intersection between nodes **(work in progress)**;
  - Shortest paths: minimum number of connections between two entries **(work in progress)**.
- Graphical visualization of resulting networks: visual interaction between entries;

Example 1

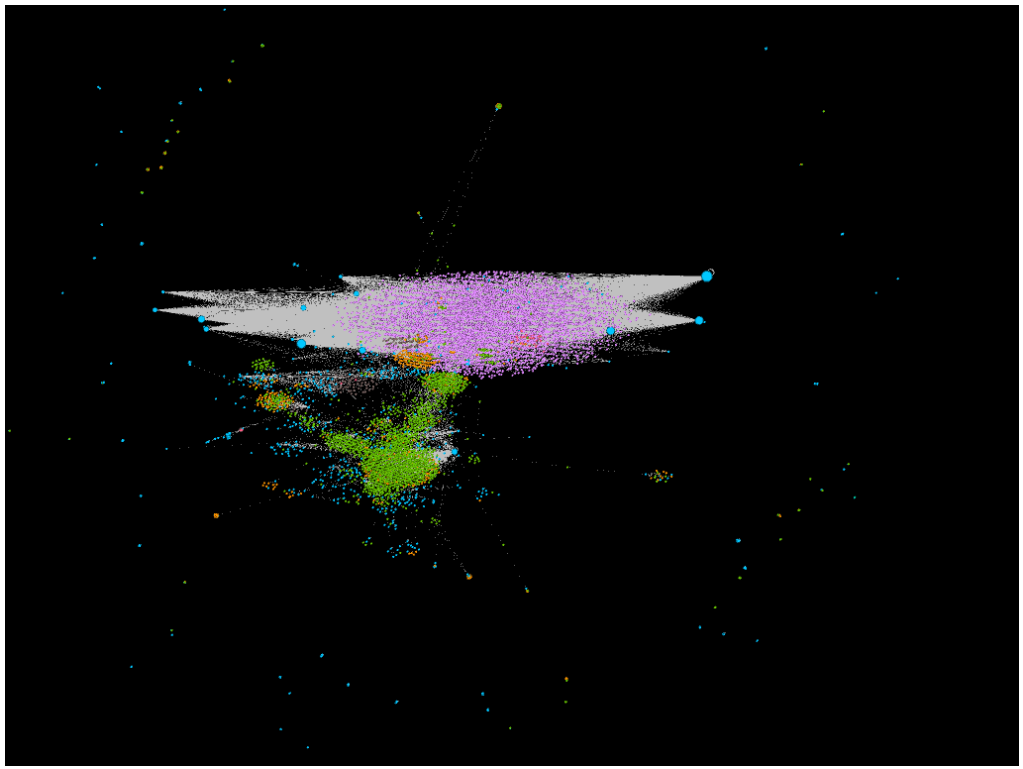
Leukemia disease

Looking for "Leukemia" disease into CHIMeRA db:

- 291 types of Leukemia;
- 82 connected components!
- 5270/9460 pendant nodes;

Fraction of types:

- 838 diseases;
- 2463 genes;
- 5195 phenotypes;
- 765 SNPs;
- 154 metabolite pathways;
- 40 metabolites;
- 5 drugs;



Example 2

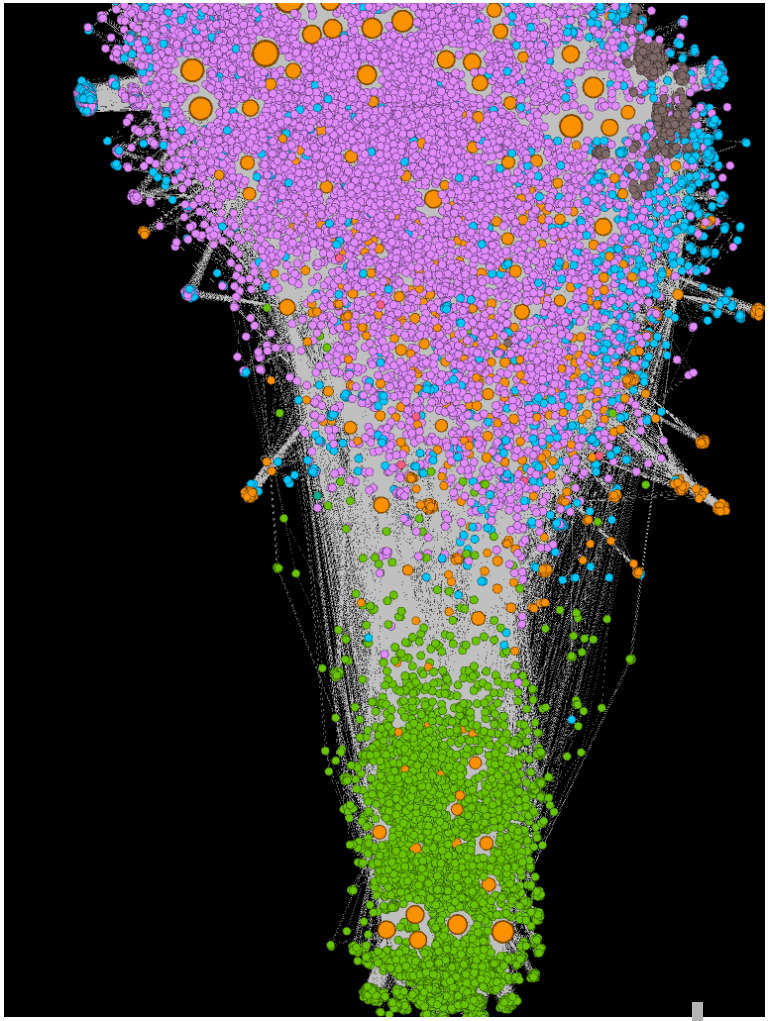
PRNP gene

Looking for "PRNP" gene into CHIMeRA db:

- 1 types of PRNP;
- 1 connected component;
- 19103 nodes;
- 139496 links;

Fraction of types:

- 777 diseases;
- 9 drugs;
- 10452 genes;
- 1 metabolites;
- 576 pathways;
- 3775 phenotypes;
- 3513 SNPs;



Work in Progress

Next steps

Future aims

- New build based on enhanced processing to increase information overlap.
- New features and information embedding from DrugBank and other data sources.
- Network of Networks analysis.
- Improve query efficiency and user interface.
- Release of the first public version as web service.

Conclusions

CHIMeRA project

- Interesting information seems to emerge from CHIMeRA, even at its alpha stage.
- Observational studies can be supported and observational biases mitigated by multi-level information patterns in the network.
- The goal is to propagate information using diseases as a “bridge”.

Project References

https://github.com/Nico-Curti

Byron: Build YouR Own Neural Network, http://github.com/Nico-Curti/Byron

NumPyNet: Neural Network library in Pure Numpy, http://github.com/Nico-Curti/NumPyNet

DNetPRO: Discriminant Analysis with Network PROcessing, http://github.com/Nico-Curti/DNetPRO

rFBP: Replicated Focusing Belief Propagation algorithm, http://github.com/Nico-Curti/rFBP

Scorer: Machine Learning Scorer Library with parallel DAG support, http://github.com/Nico-Curti/scorer

FiloBluService: FiloBlu project service manager for text message processing, http://github.com/Nico-Curti/FiloBluService

ShUt: Shell Utilities and Installers for no root users, http://github.com/Nico-Curti/shut

genetic: Examples about genetic algorithms for parallel and distributed computing, http://github.com/Nico-Curti/genetic

BlendNet: Network viewer with Blender support, http://github.com/Nico-Curti/BlendNet

CryptoSocket: TCP/IP Client Server with RSA cryptography, http://github.com/Nico-Curti/CryptoSocket

easyDAG: Simple template DAG scheduler in C++, http://github.com/Nico-Curti/easyDAG

Cardio: Pulse oximetry data processing and classification, http://github.com/Nico-Curti/cardio

SysDyn: System Dynamics script utilities, http://github.com/Nico-Curti/SysDyn

rSGD: Replicated Stochastic Gradient Descent algorithm, http://github.com/Nico-Curti/rSGD

Walkers: Random Walk and Optimizer Simulator, http://github.com/Nico-Curti/Walkers

Data-Analysis: Data Analysis Utilities (from Statistics to Machine Learning), http://github.com/Nico-Curti/Data-ANalysis

Thanks to the research group

Biophys and Physycom groups