Module ECUE
« Applied AI »

AI & Biomedical : Big data in bio imagery

Xavier Descombes
Morpheme team
INRIA/I3S/iBV

High Spatial Resolution Multiscale

- Microscopy images :
  - Spatial resolution in x/y : lower than 1μm
  - 2D or 3D datasets : up to several hundreds of slices

Example : mice brain image on light-sheet microscopy

X = 0.75μm, Y = 0.75μm, Z = 1.99μm

6000 x 6000 x 1000 voxels

Voxels coded on 16 bit

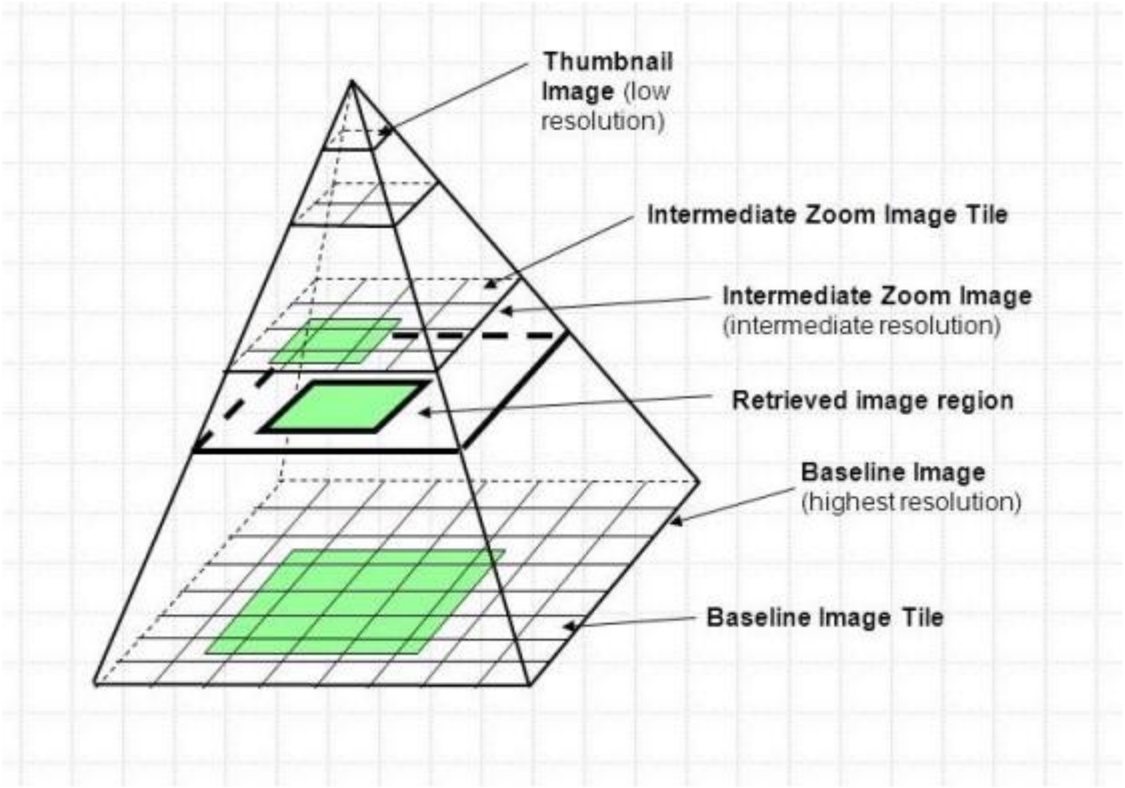
Study on 2 channels + time course

20-300 Gb

about 40 brains

Histopathology data

Pyramidal structure of data (Whole_Slide Imaging format)

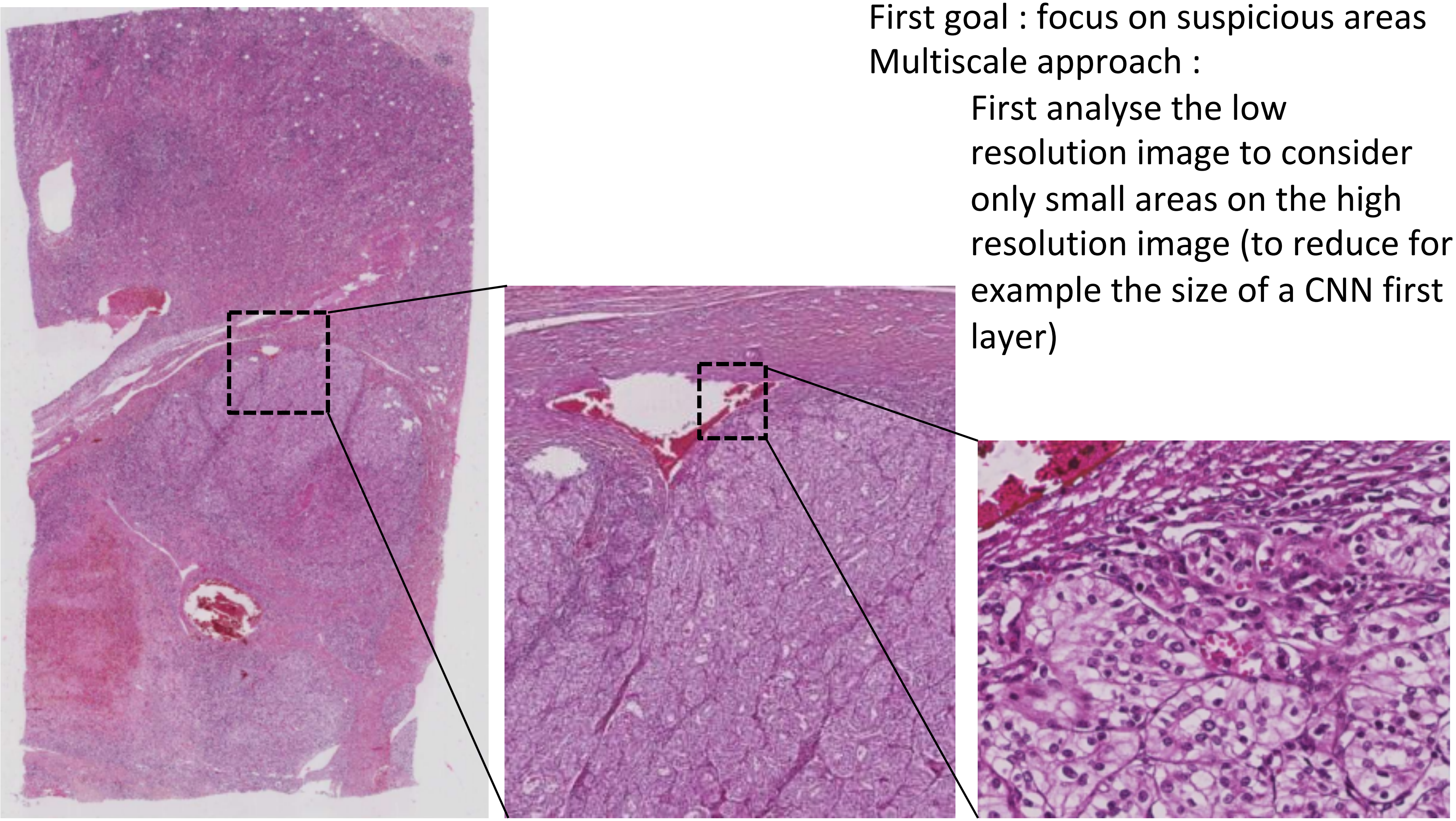

Thumbnail Image (low resolution)
Intermediate Zoom Image Tile
Intermediate Zoom Image (intermediate resolution)
Retrieved image region
Baseline Image (highest resolution)
Baseline Image Tile

Histopathological data

First goal : focus on suspicious areas

Multiscale approach :

First analyse the low resolution image to consider only small areas on the high resolution image (to reduce for example the size of a CNN first layer)



(822 x 1.365) *(13.152 x 21.840)* *(52.608 x 87.360)*

Histopathology data analysis

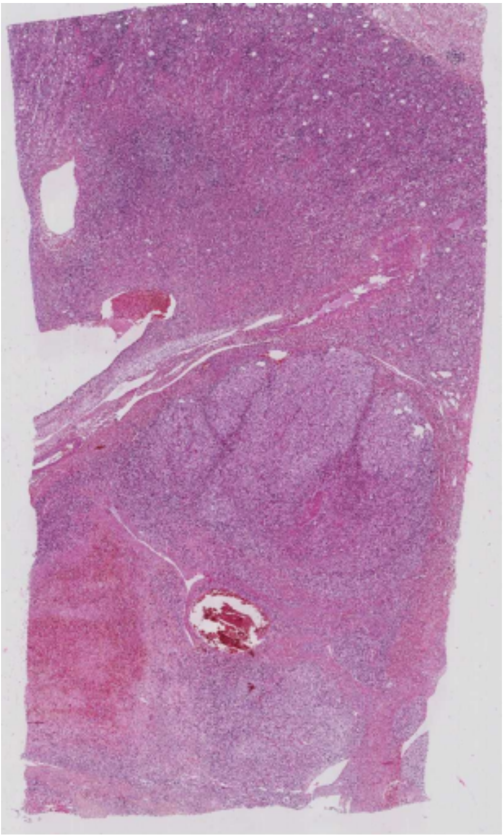
- Goal : classify and grade the cancer
- Focuss on ROI (tumor zones)
- Medical decision based on local patterns
- Histopathologist analysis :
  - Screen the low resolution image to detect ROIs
  - Zoom on these ROIs (and come back)

  -> multiscale analysis

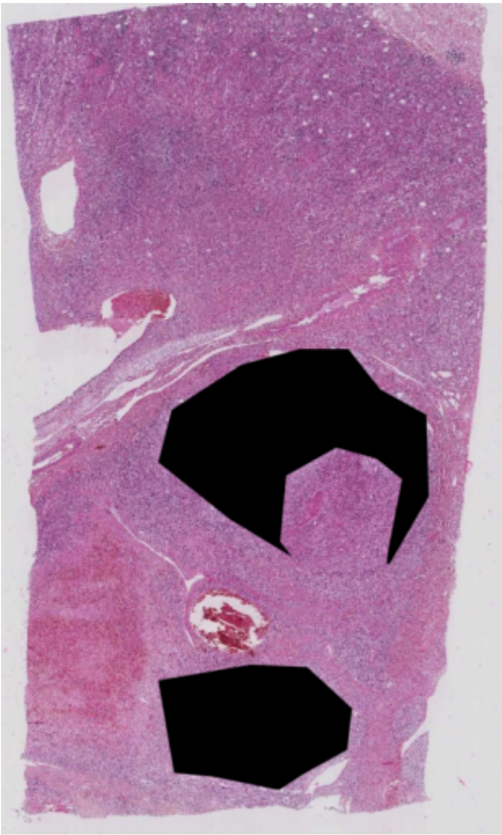
  -> qualitative analysis

A first machine learning task

- Classify ROI (tumor areas)



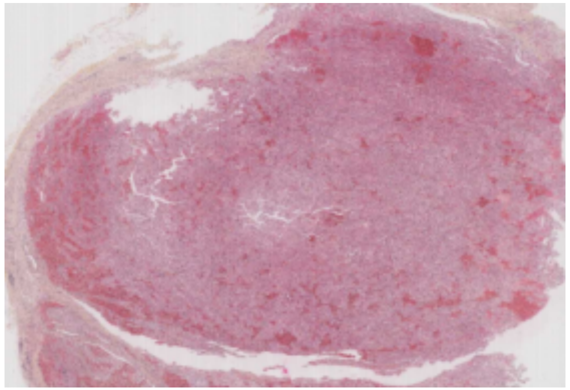
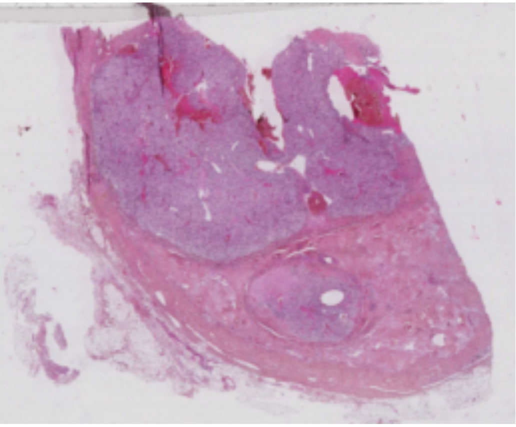
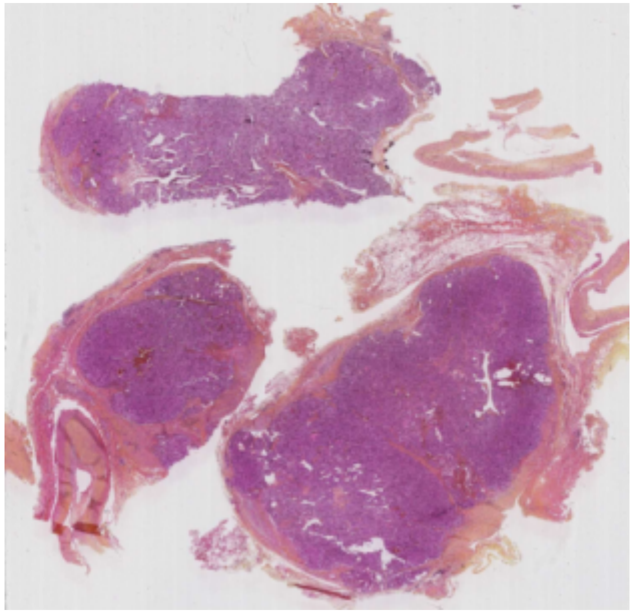
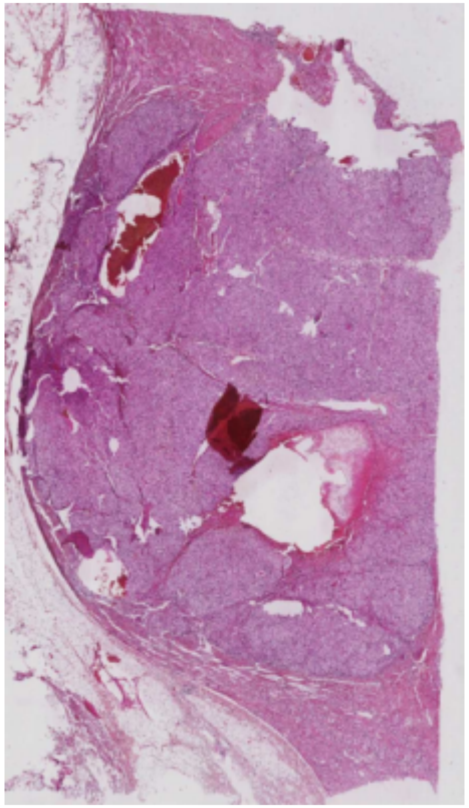
Initial Image



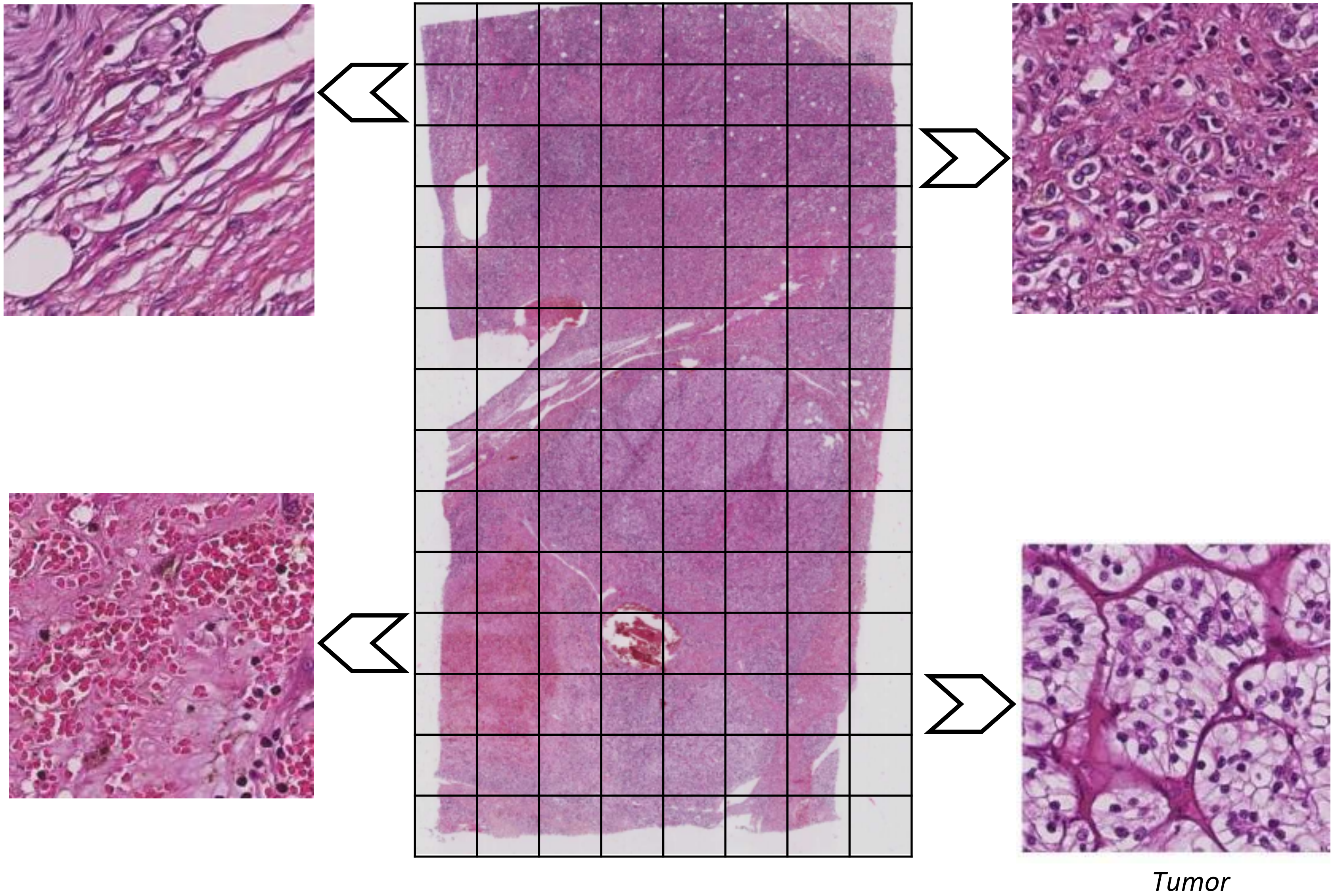
Tumor areas (ROI)

Challenge

- Variability between and within images
- Non informative areas (fat, blood…)
- Huge datasets (12 Go ≅ 100 000 pixels per axis)

Patches classification



Pre-Processing : color deconvolution

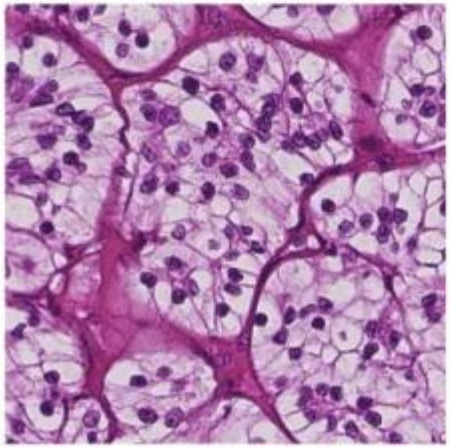
$$\forall c \in \mathbb{N}^2, H(c) = \frac{Rouge(c)}{C3(c)},$$

$$\forall c \in \mathbb{N}^2, E(c) = \frac{Vert(c)}{Rouge(c)}.$$

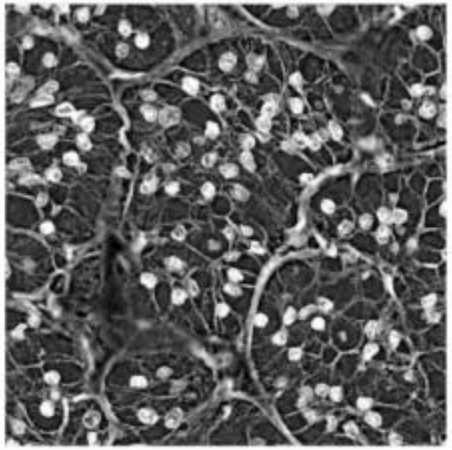
$$\forall c \in \mathbb{N}^2, C1(c) \quad = arctan\left[\frac{Rouge(c)}{max\Big(Vert(c), Bleu(c)\Big)}\right]$$

$$\forall c \in \mathbb{N}^2, C2(c) \quad = arctan\left[\frac{Vert(c)}{max\Big(Rouge(c), Bleu(c)\Big)}\right]$$

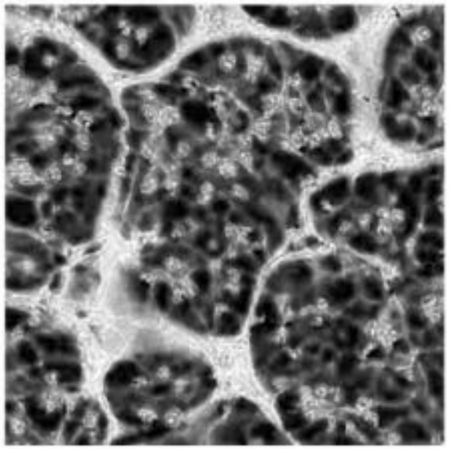
$$\forall c \in \mathbb{N}^2, C3(c) \quad = arctan\left[\frac{Bleu(c)}{max\Big(Rouge(c), Vert(c)\Big)}\right]$$



RGB image



H channel



E channel

Reduce dataset

- Remove patches with few cells

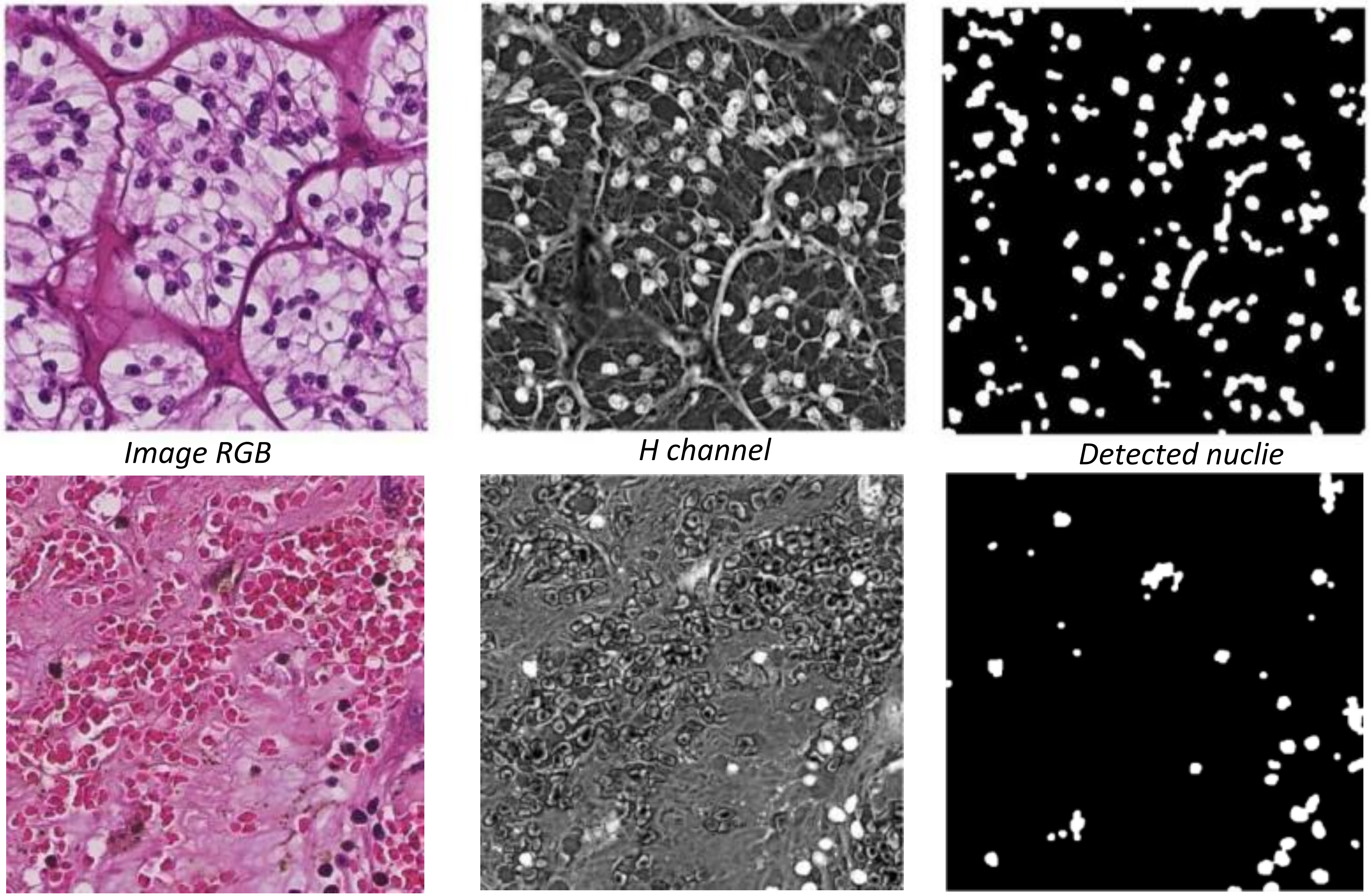


Image RGB

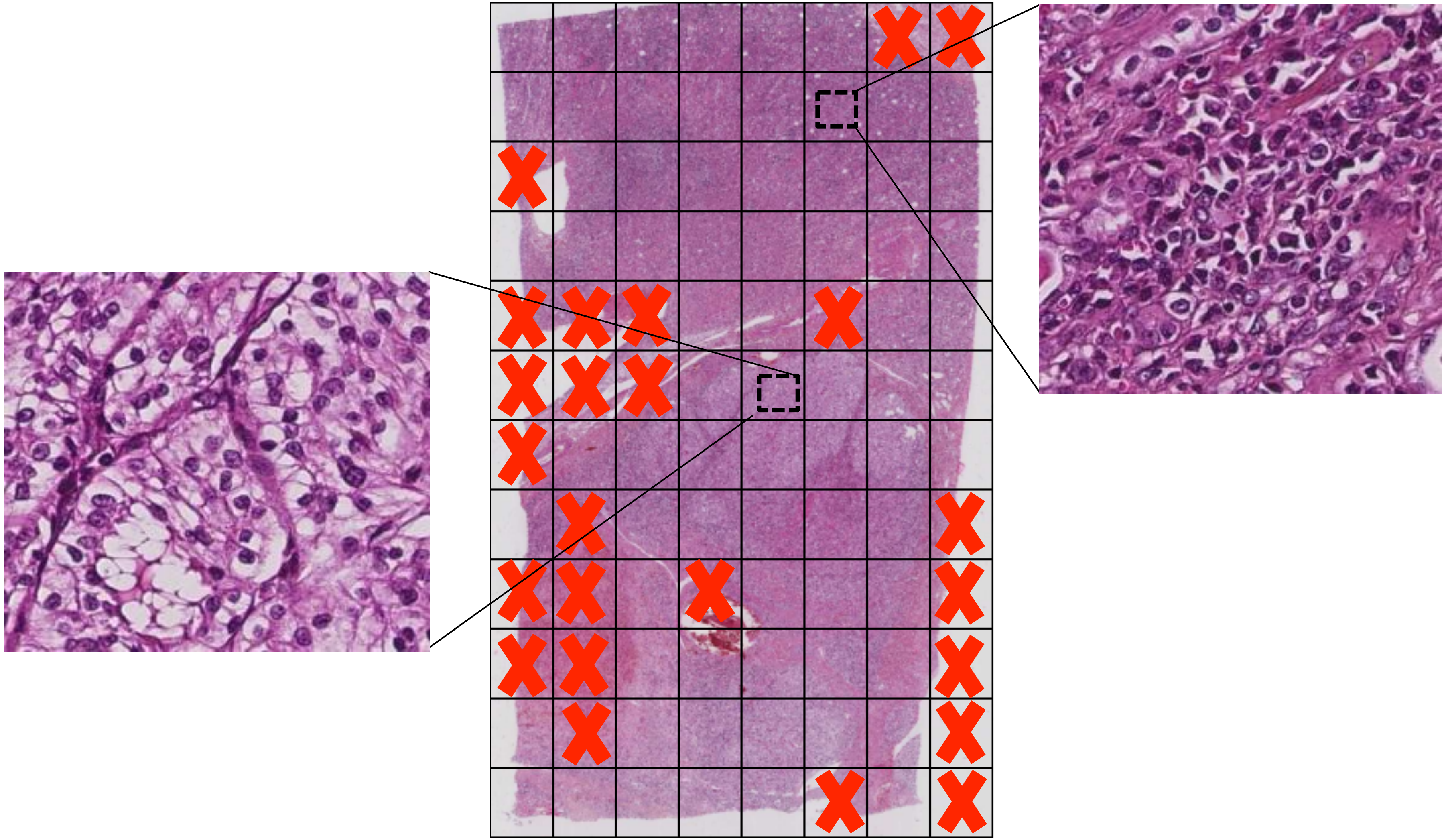
H channel

Detected nuclie

120 nuclei

32 nuclei

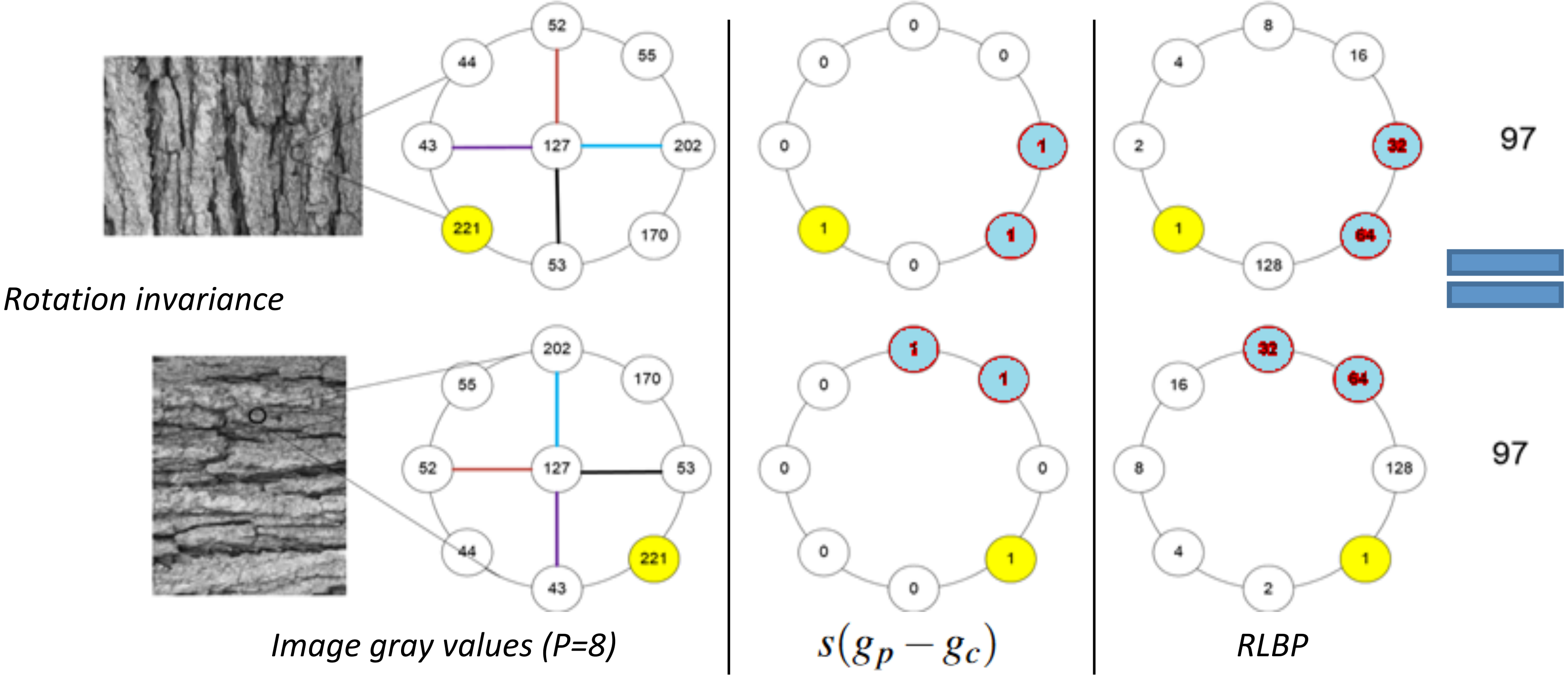
Reduce dataset



Features extraction : local binary patterns

$$D = \underset{p \in (0,1 \ldots P-1)}{\arg\max} |g_p - g_c|$$

$$RLBP_{R,P} = \sum_{p=0}^{P-1} s(g_p - g_c) \cdot 2^{mod(p-D,P)}$$



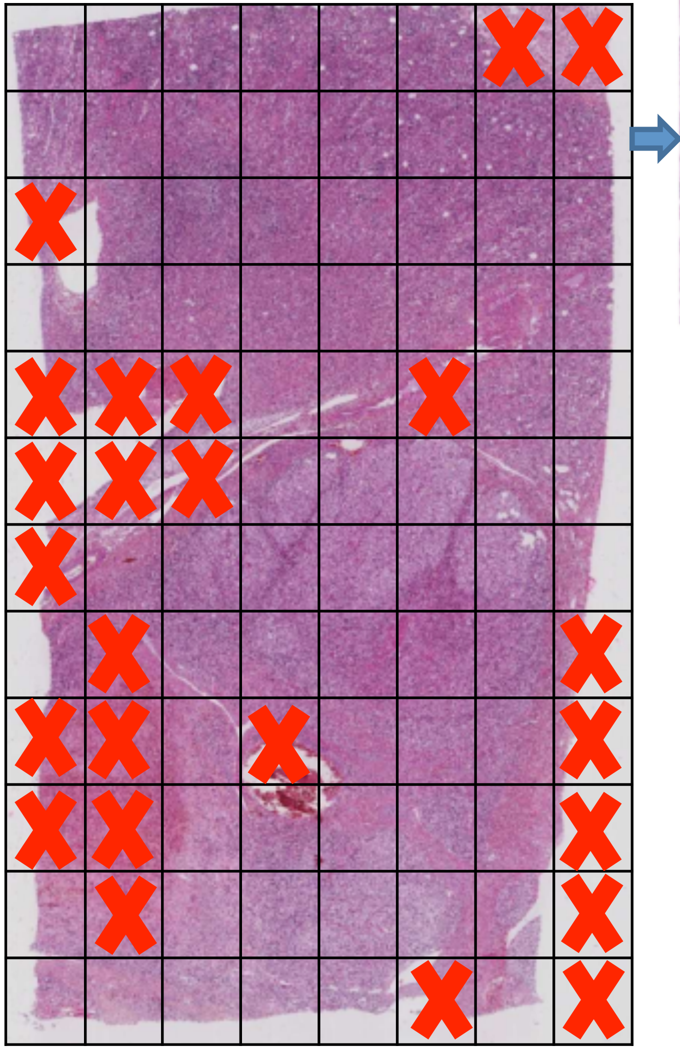
Rotation invariance

Image gray values (P=8)

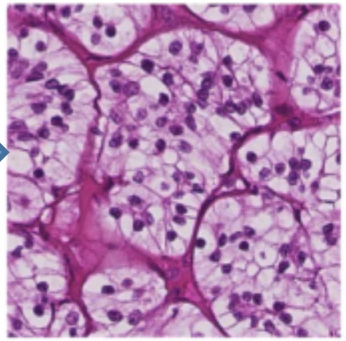
$s(g_p - g_c)$

RLBP

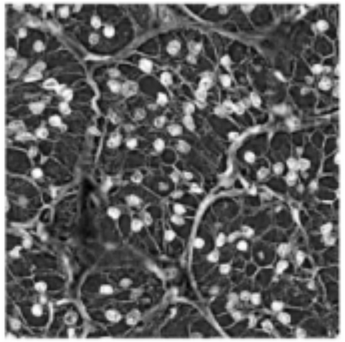
Classification : k-means



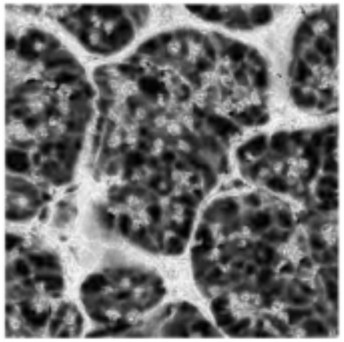
Result after nuclei detection



RGB image



H channel



E channel

RLBP- H	14	440	8700	...	...	...	...	...	...	745
RLBP- E	945	560	163	...	...	...	...	...	...	12

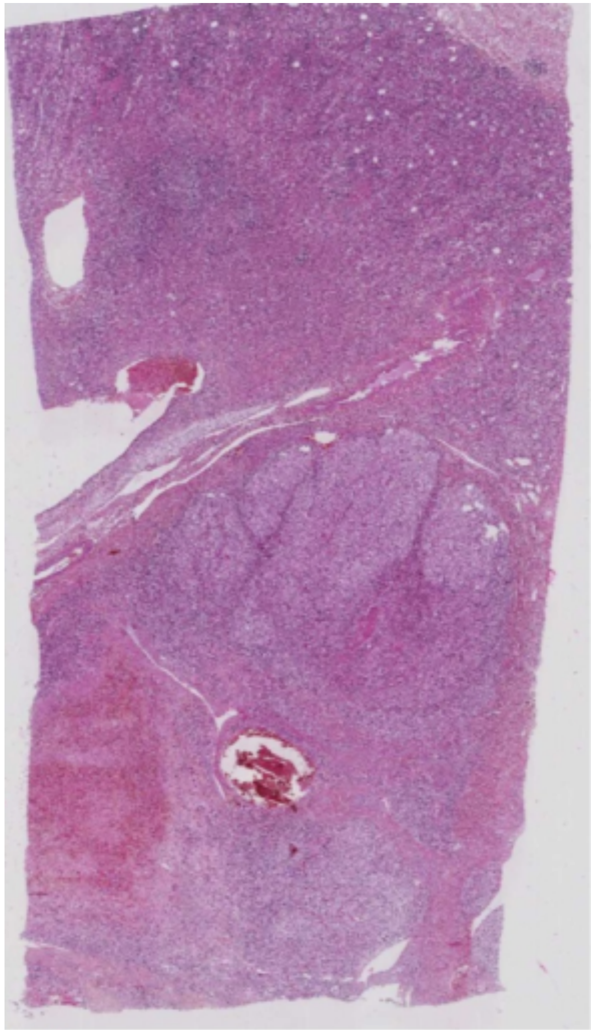
P=16, R=3 (65.536 x 2 patterns)

Find the Most Frequent patterns from the training images

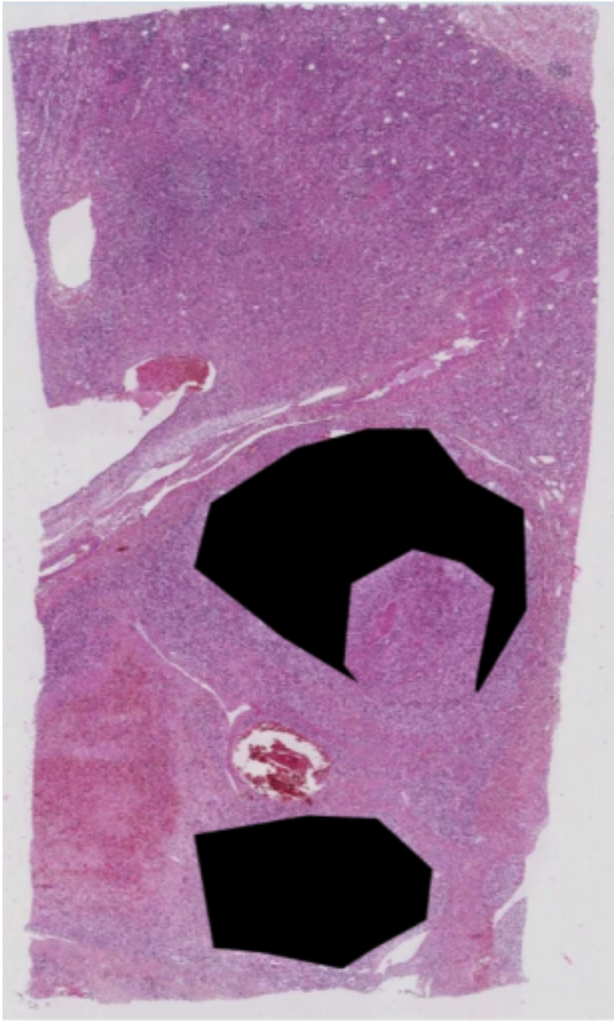
Classification K-NN

- *Classes number: 02 (Tumor / Not-tumor)*
- *Learning : 10 slides ➔ 850 patches*
- *Test : 05 slides*

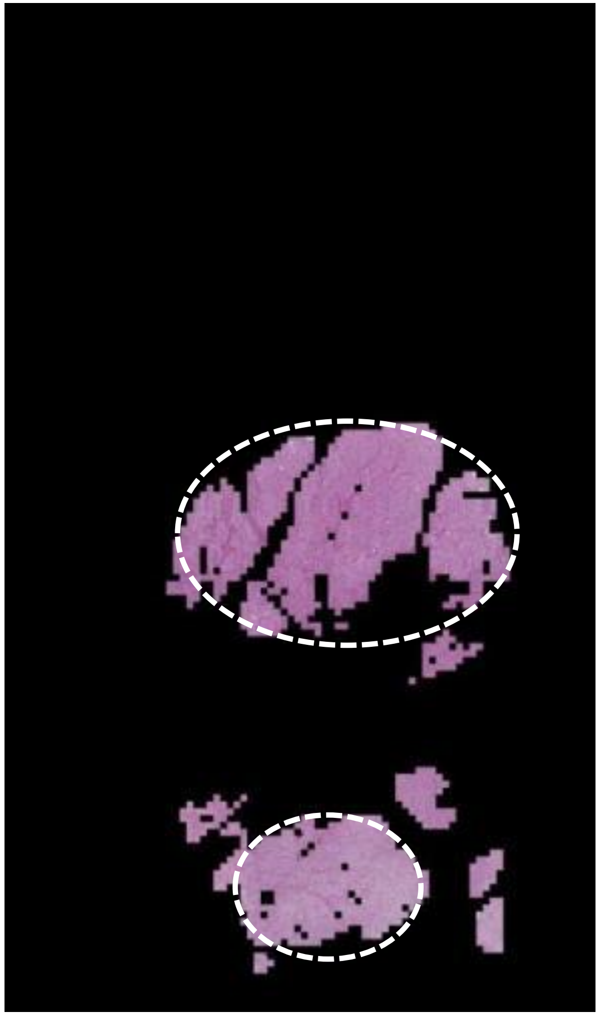
Result on learning set



Input image

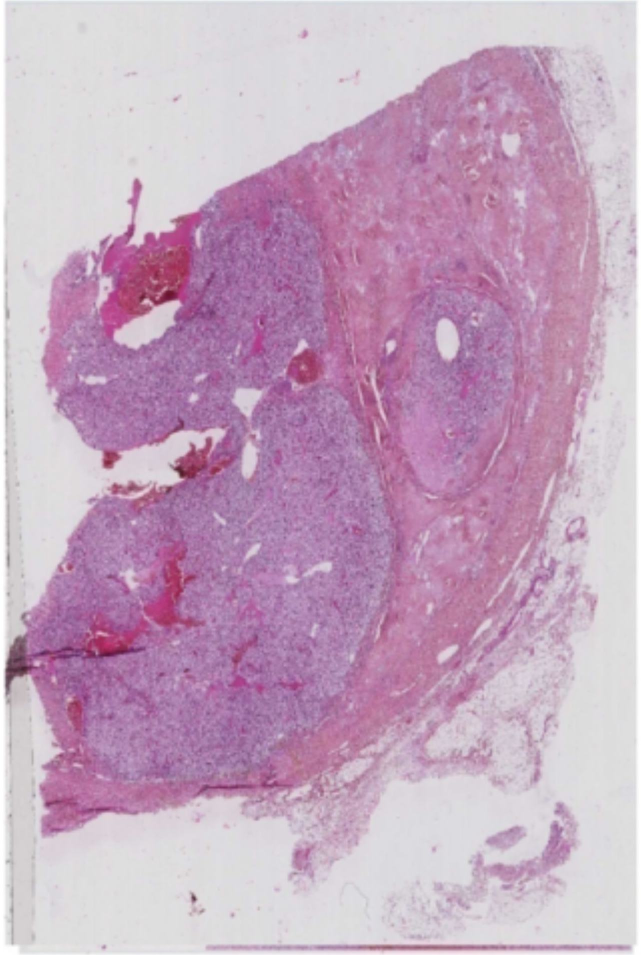


Ground Truth

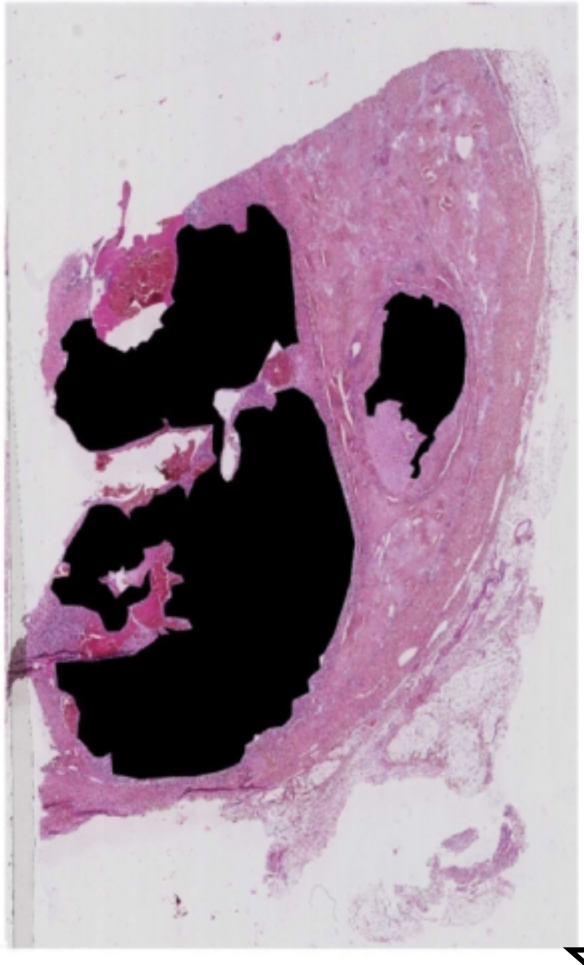


Result

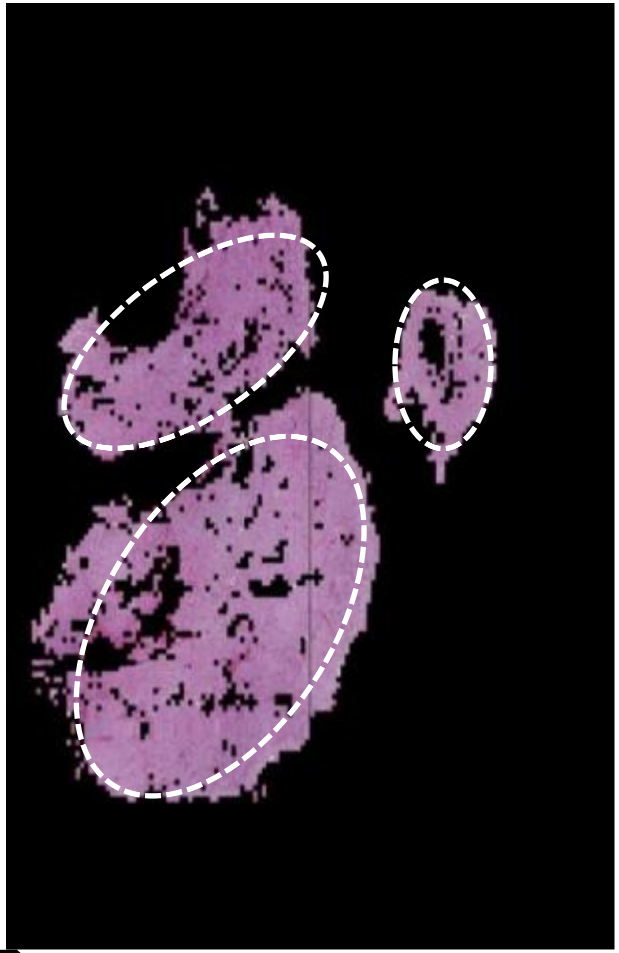
Result on test set



Input image



Ground Truth

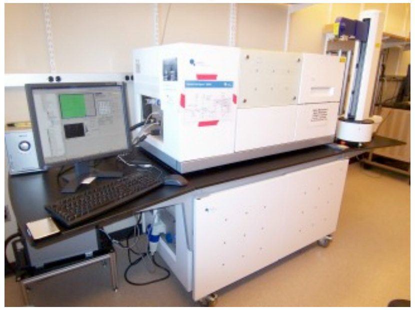


Result

High Throughput data

GFP-Imp + cells

RNAi



high
throughput
confocal
microscope

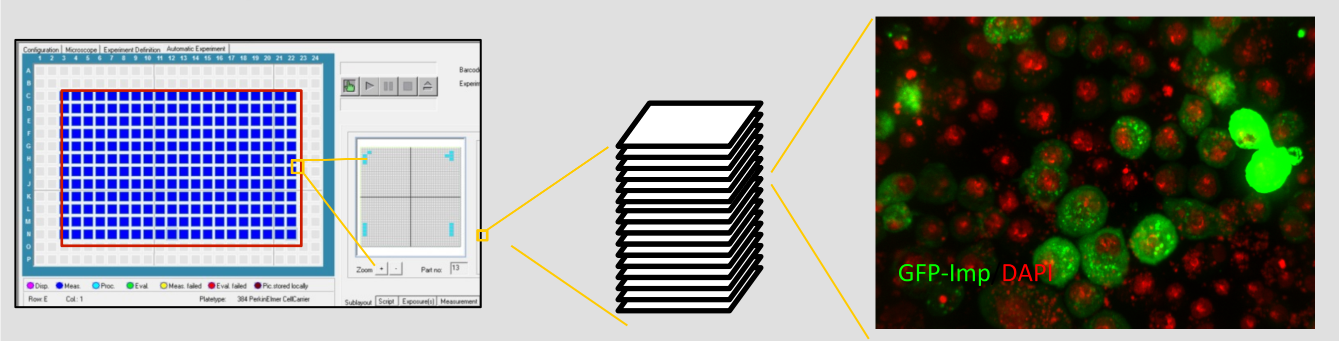
PILOT SCREEN

Kinases and phosphatases (563)

RNA binding proteins (406)

~1000 out of 13000 genes

DRSC & TRiP
Drosophila RNAi Screening Center at Harvard Medical School



3 millions of images !!!

High Throughput data

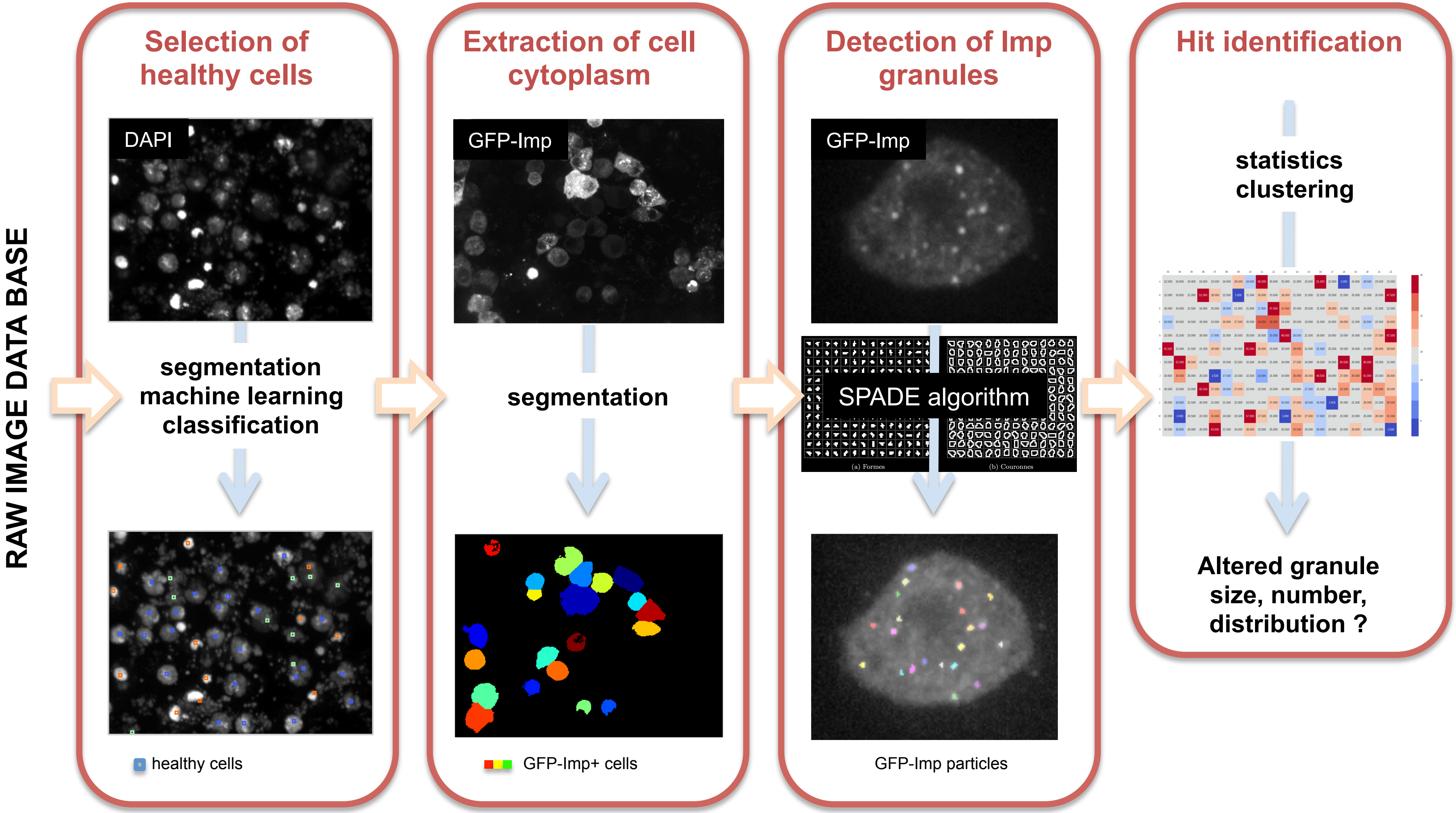
First consider a PILOT screen (subsample of well chosen genes) :

Kinases and phosphatases (563)

RNA binding proteins (406)

~1000 out of 13000 genes

A pipeline for analysing the screen



RAW IMAGE DATA BASE

Selection of healthy cells

segmentation
machine learning
classification

🟦 healthy cells

Extraction of cell cytoplasm

segmentation

🟥🟨🟩 GFP-Imp+ cells

Detection of Imp granules

SPADE algorithm

GFP-Imp particles

Hit identification

statistics
clustering

Altered granule
size, number,
distribution ?

Disciplines involved : biology, machine learning, image processing, data base

Machine learning task

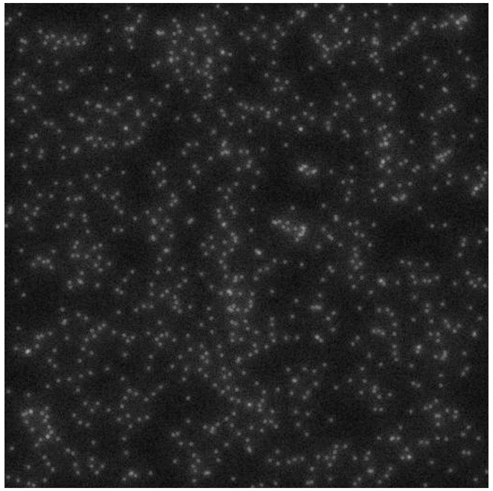
- Classify the cells w.r.t. the granule population
- Features : number, size, spatial repartition
- Challenge : unknown number of classes
- Rejection class
- Unbalanced classes
- Huge number of samples : checking difficult

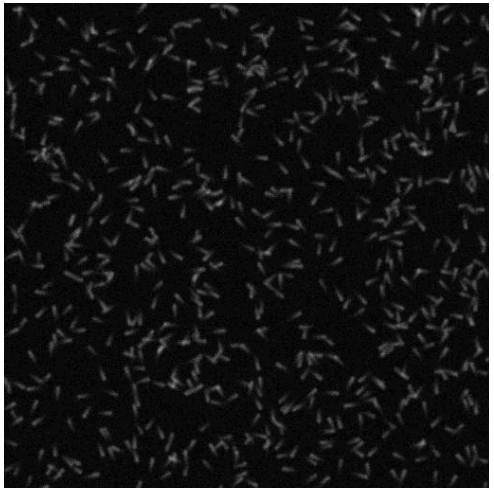
Time sequences

Time sequences

Multiple particles tracking

- Main approaches in two steps :
  1) Objects (particles) detection
  2) Objects (particles) linking
- Particular case : (for low speed)
  - » Trajectories detection in (Space + Time) domain

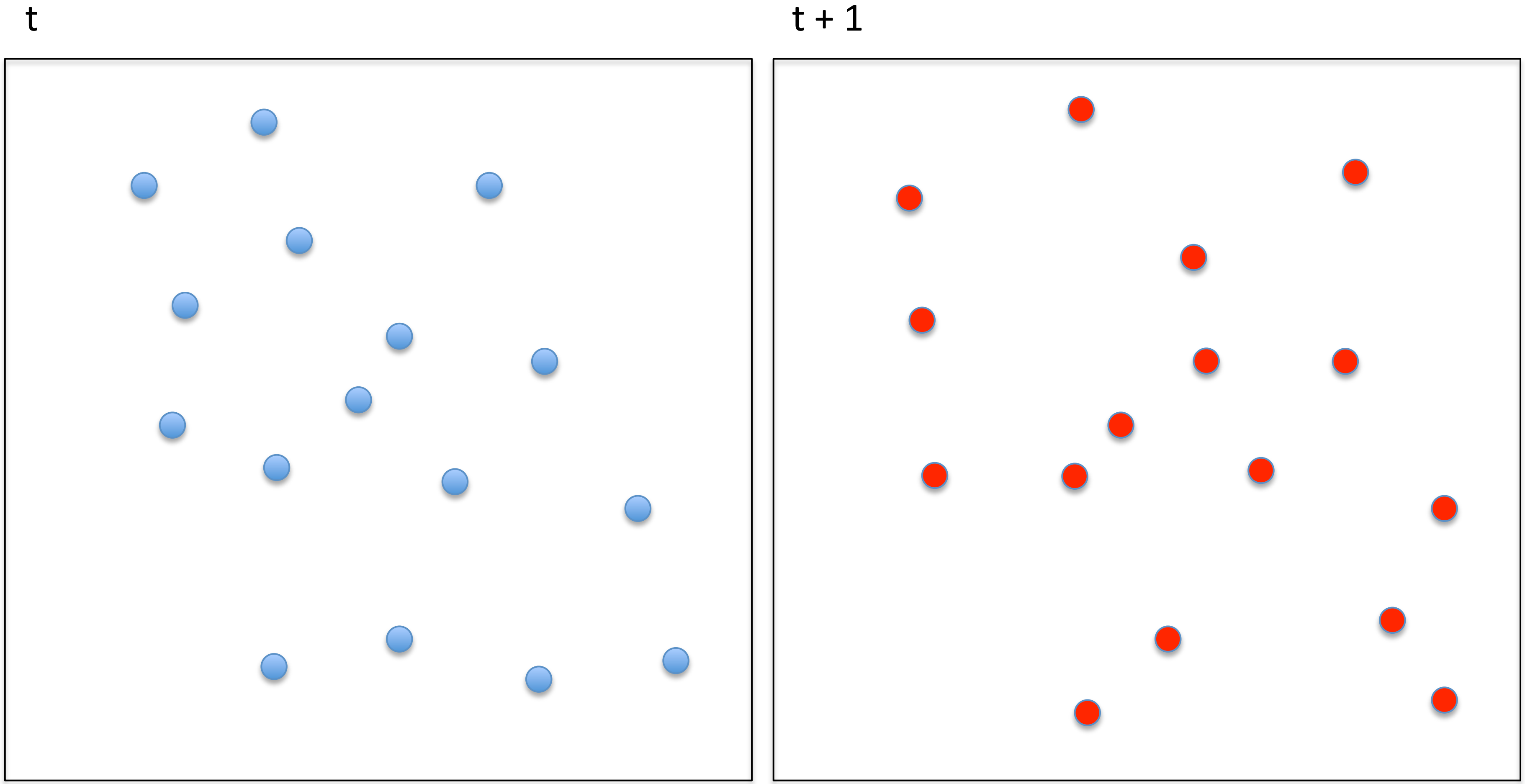




The challenges

- Detection (see corresponding course)
- Appearance / Disappearance of particles
- Crossing
- Occlusion
- Noise
- Location VS shape descriptor

Detect and Match

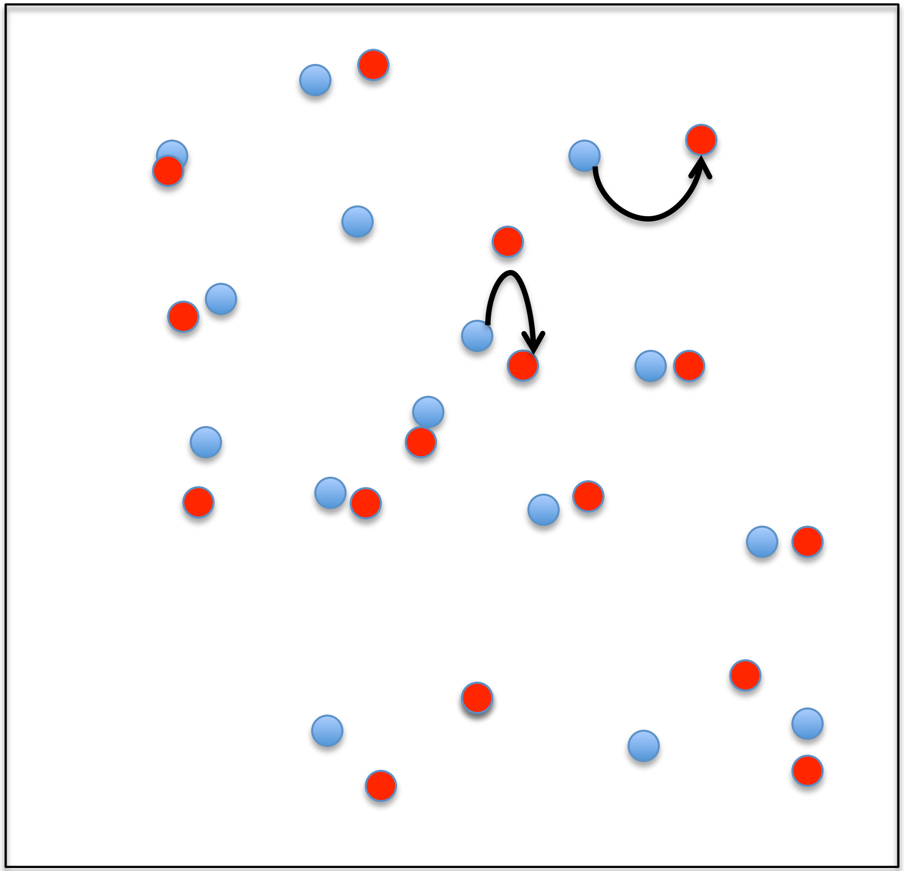


Match : nearest neighbor

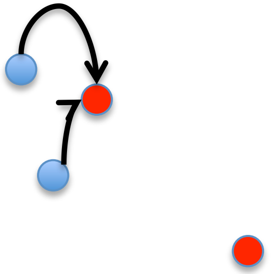
x_i

y_j

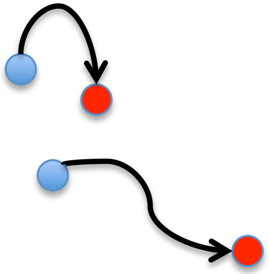
$$M(x_i) = \arg\min_{y_j} d(x_i,y_j)$$



Unicity constraint

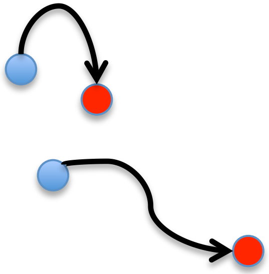


versus

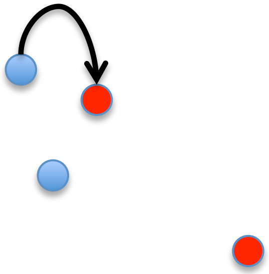


Matching matrix : $M(i,j) = 0 \text{ or } 1$ $\quad \sum_i M(i,j) = 0 \text{ or } 1$

Maximum velocity



versus



$$d(x_i, y_j) > V_{max} \implies M(i,j) = 0$$

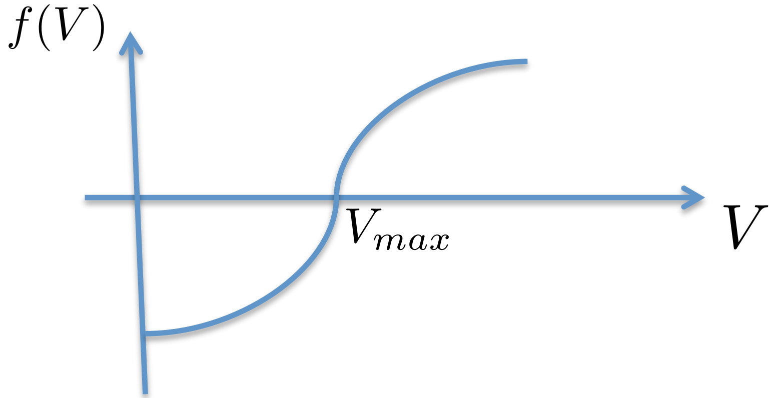
Global optimization

$$argmin_M \sum_{i \in I, j \in J} d(x_i, y_j) M(i,j)$$

$$\boxed{M(i,j) = 0 \text{ or } 1}$$

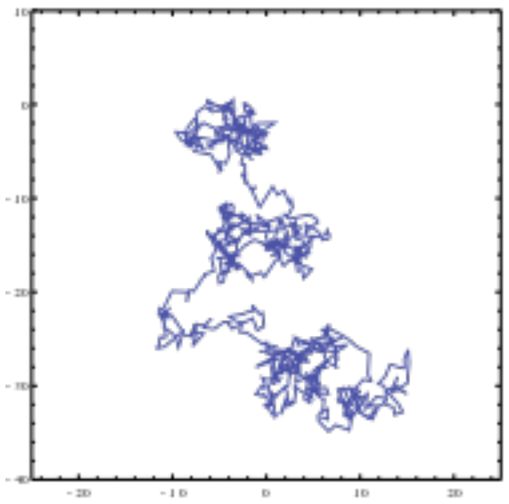
$$\forall i \sum_j M(i,j) = 0 \text{ or } 1$$

$$d(x_i, y_j) = f(||y_j - x_i||)$$



Movement modeling

- Brownian motion : random movement (big particle in a fluid)



$$P(x_{t+1}|x_t) = \frac{1}{2\pi\sqrt{\sigma^2}} \exp\left[-\frac{(x_{t+1}-x_t)^2}{2\sigma^2}\right]$$

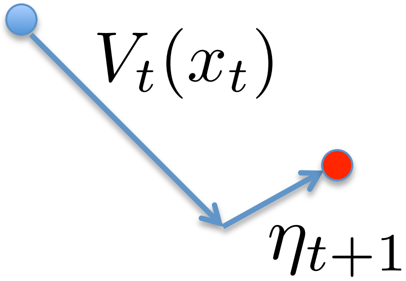
https://fr.wikipedia.org/wiki/Mouvement_brownien#Processus_d%E2%80%99Ornstein-Uhlenbeck

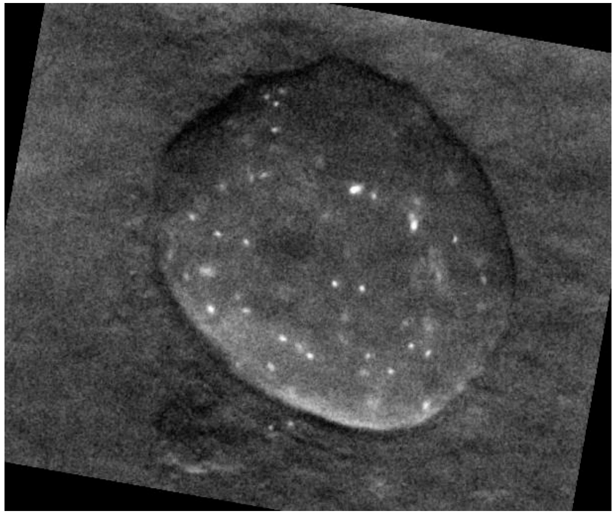
Deterministic speed model

$$x_{t+1} = x_t + V_t(x_t)dt + d\eta_{t+1}$$

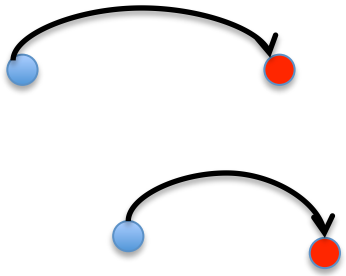
Deterministic model

Fluctuation

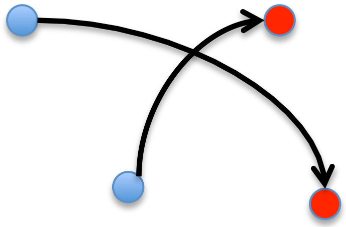


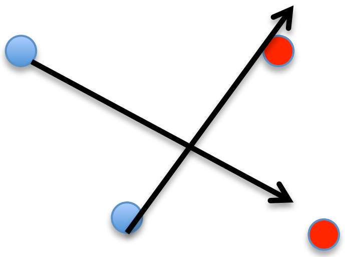


Advantage of a model



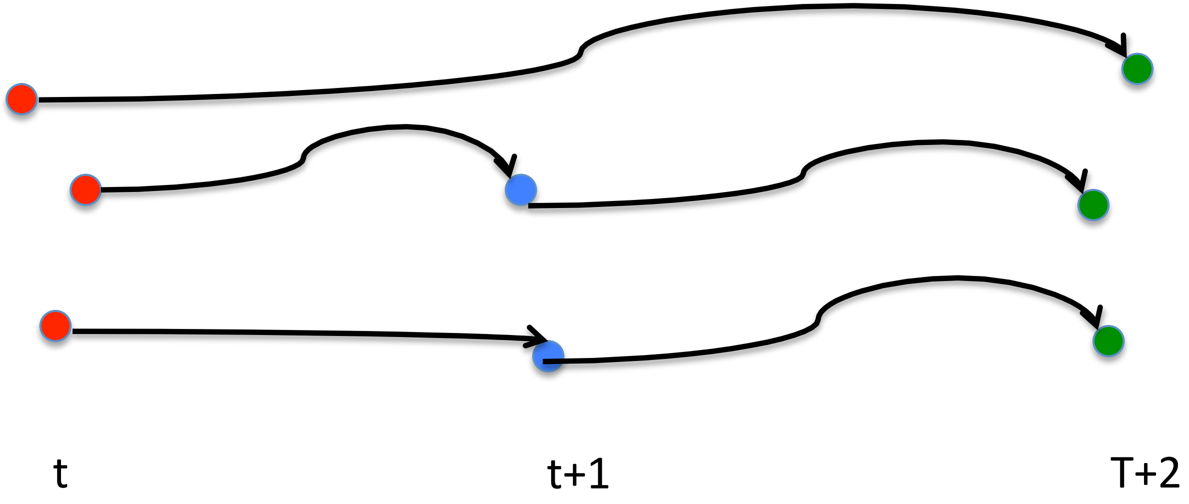
versus



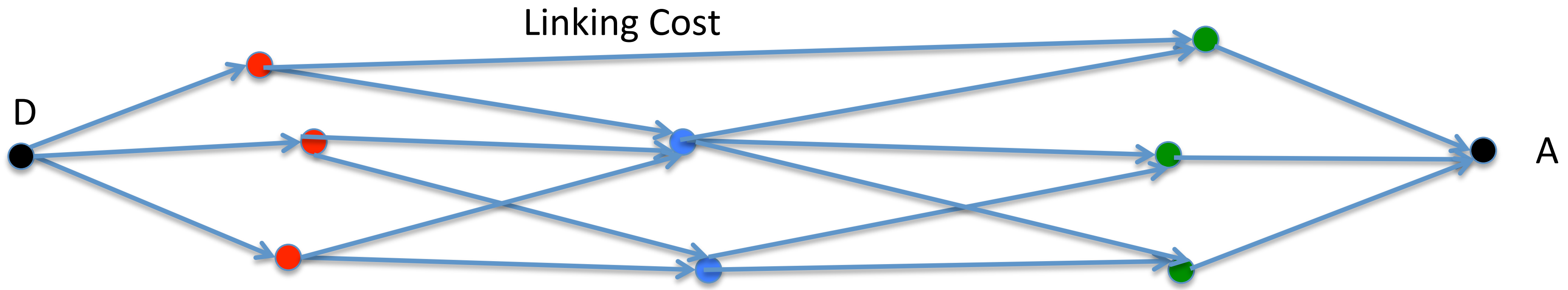


Speed : learnt from a model or estimated from past steps

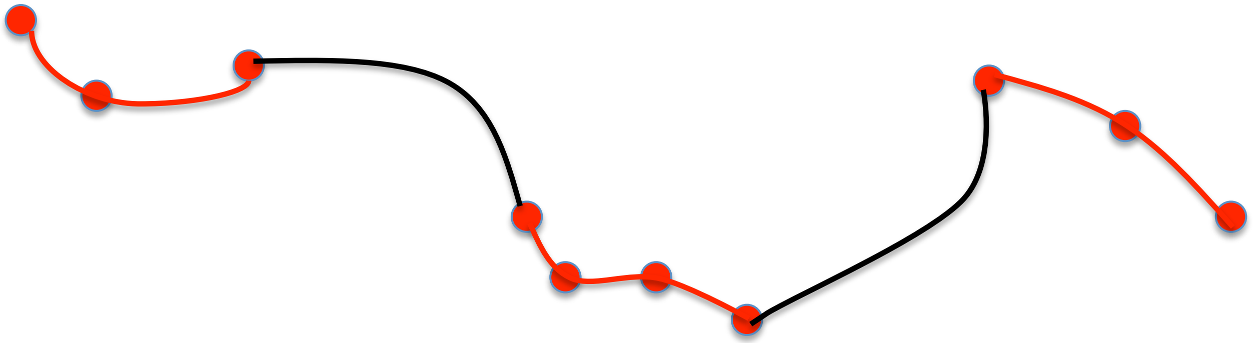
Gap filling



Graph model : Minimal Path



Tracklets



Two steps : Local (tracklets detection), tracklets merging

Pros : consider trajectory and/or speed models

Take home message

- Do not consider higher resolution than needed (for space, time and intensity)
- Consider a multiscale approach
- Adapt the processing to the size of data (compromise between accuracy and computaiton time)
- Consider parrallel programing