Leveraging user-interaction and auxiliary data to learn from small data.

Romain Mormont

Systems and modeling unit,
Department of EE & CS,
University of Liège, Belgium

6th October 2016

Context

- **Machine learning** (ML) has recently gained popularity through several successes (DeepMind, Google and Tesla self-driving cars, IBM Watson,...)
- Those were mostly possible thanks to:
  - **Plenty of data** (*big data*)
  - High computational power
- Those criterion are essentials to get the best performances out of common ML methods

Problem

- **In practice, those two criteria are not always met**.
- Especially, useful (i.e. labelled) data can be scarce. One can then talk about **small data** !

Small data: "the amount of data is not large enough with respect to the complexity of the task at hands"

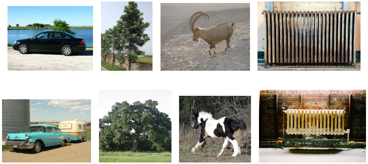
N.B.: in *small data* settings, the amount of data is not small in absolute terms

Small data: illustration

Big data
ImageNet[1]

Dataset:

- 14M labelled images
- Enough data to learn inter- and intra-class variabilities

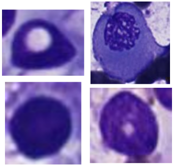


Typical images from ImageNet

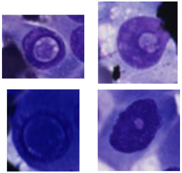
Small data
Thyroid nodule malignancy

Dataset:

- 6K labelled images
- Most important objects (i.e. malignant) are rare



(a) Healthy



(b) Malignant

[1] Deng & al. *ImageNet: A Large-Scale Hierarchical Image Database*. In CVPR09, 2009.

Objectifs

Explore and develop new methods adapted to *small data* problems.

Tracks and research questions:

1. *Interactice machine learnin (iML)* : how to **integrate human operators to the learning process** in order to improve its performances ?
2. *Transfer learning*: how to **leverage any available auxiliary data** for reaching the same goal ?

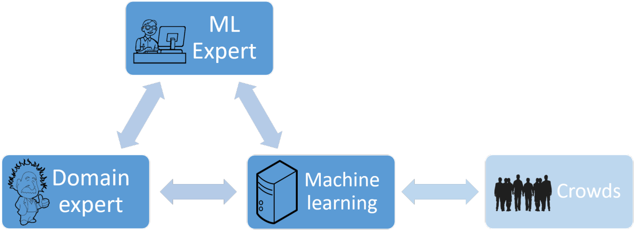
Methodology:

- Gather data
- Develop solution
- Test and improve it on benchmark problems
- Validate on real problems

Question 1: human in the loop ?

From ...

To ...



Question 1: human in the loop ?

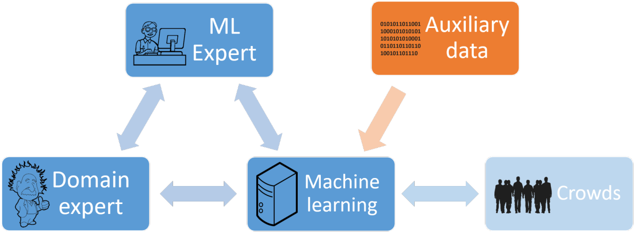
State of the art:

- Mostly **active learning** (i.e. smart example labelling)
- Few developments around richer forms of feedback

Methodological challenges:

- How to **minimize the number of interaction** ? (the method/system must *ask the right questions*)
- How to **minimize time between each interaction** ? (the method/system must *be fast and reactive*)
- How to **integrate the feedbacks** to the learning process ?

Question 2: transfer learning ?

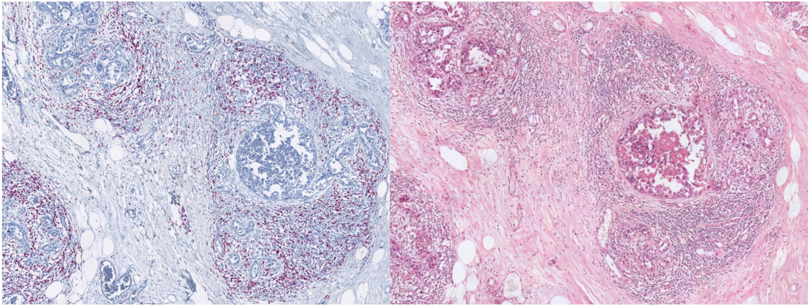


Methodological challenges:

- **What information should be transferred** to improve performances and avoid negative transfer?
- **How to integrate this information** into the methods ?

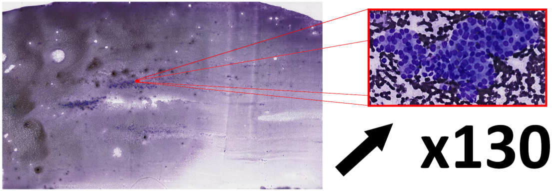
Question 2: transfer learning ?

Apply transfer learning on **multi-modal images**.



Case study

Rare object detection and categorization in high-resolution tissue images.



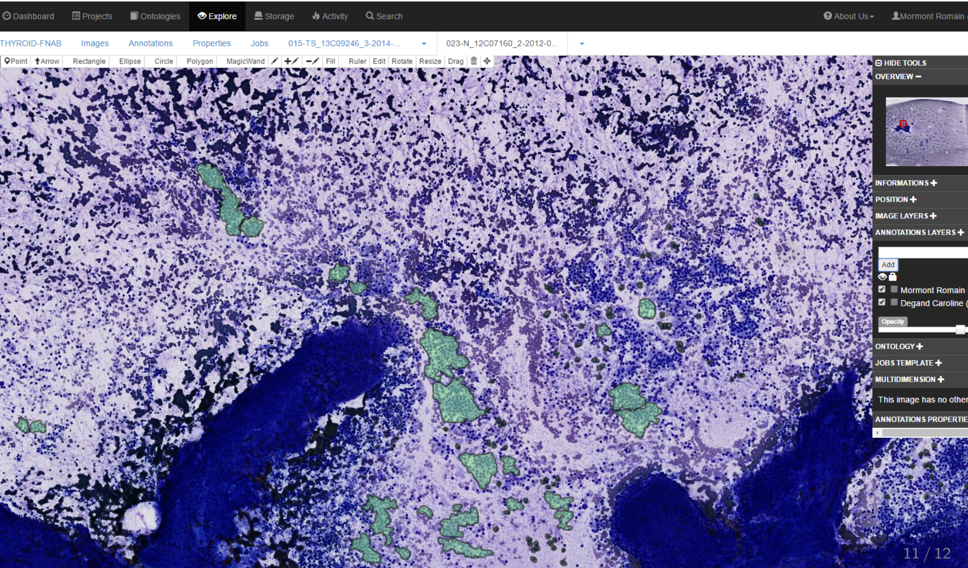
Thyroid nodule malignancy diagnosis

Related applications:

- Defect detection in images of manufactured components
- Large celestial body detection in astronomy

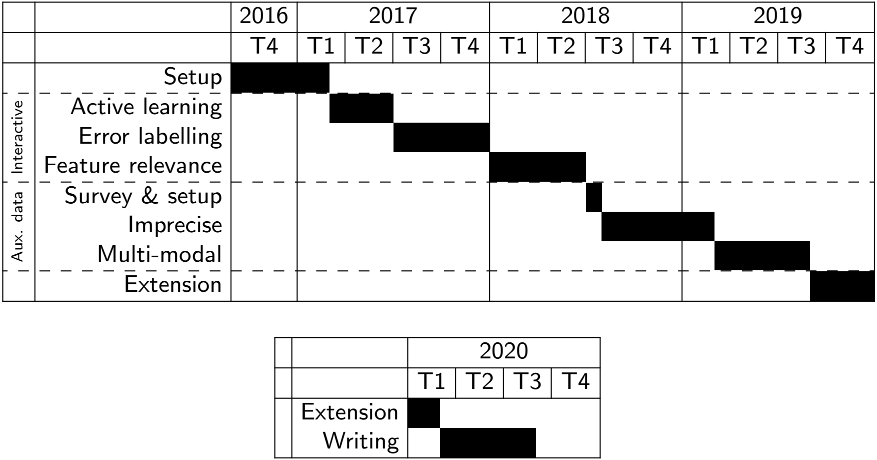
Case study: Cytomine

cytomine web platform will be used for collecting **feedback** from experts and from the crowds.



Thank you for your attention!
Any question ?

Work calendar



Backup slides

TODO:

- résultats du TFE,
- applications à moyen ou long terme,
- publications récentes. . .
- théorie (ML,...)
- thyroid whole slide image for example illustration