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

28th October 2016

Context

- Supervised **machine learning** (ML) is about learning input-output models from data
- It has recently gained popularity through several successes with important media coverage
- Those were mostly possible thanks to:
  - **Plenty of data** (*big data*)
  - High computational power
- Those criteria are essential to get the best performance out of common ML methods

Problem

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

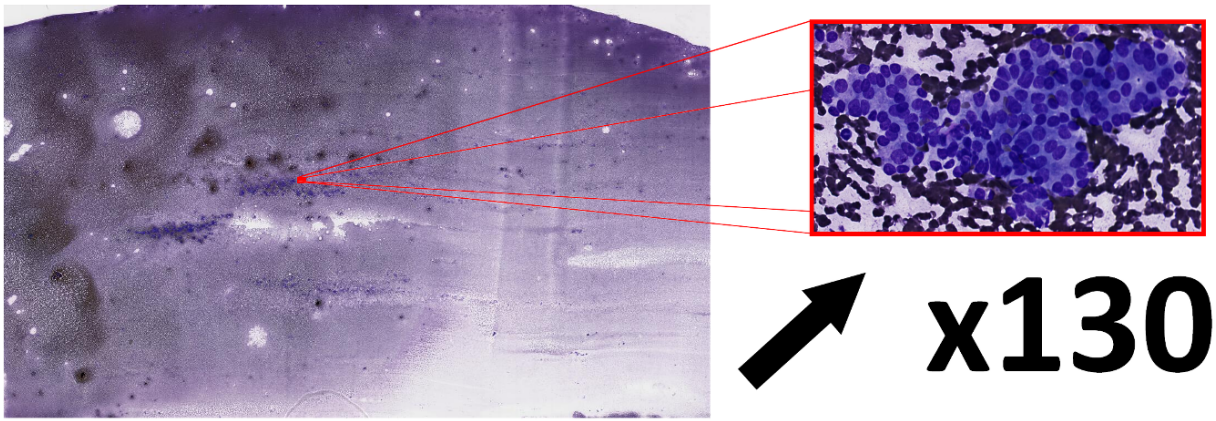
Small data: "the amount of data is not large enough with respect to the complexity of the task at hands"

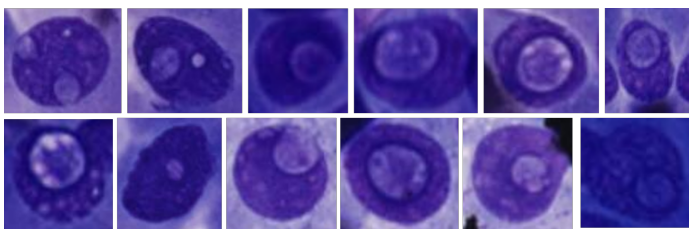
N.B.: in *small data* settings, the amount of data is not necessarily small in absolute terms

Case study

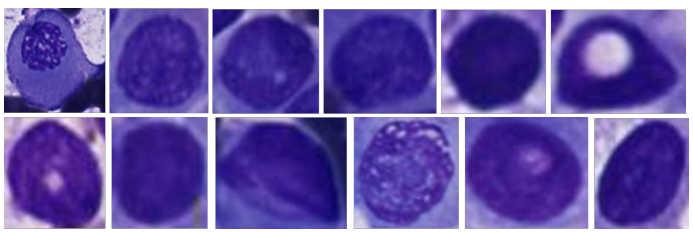
Rare object detection and categorization in high-resolution tissue images.

Thyroid nodule malignancy diagnosis





Malignant



Benign

Objectives

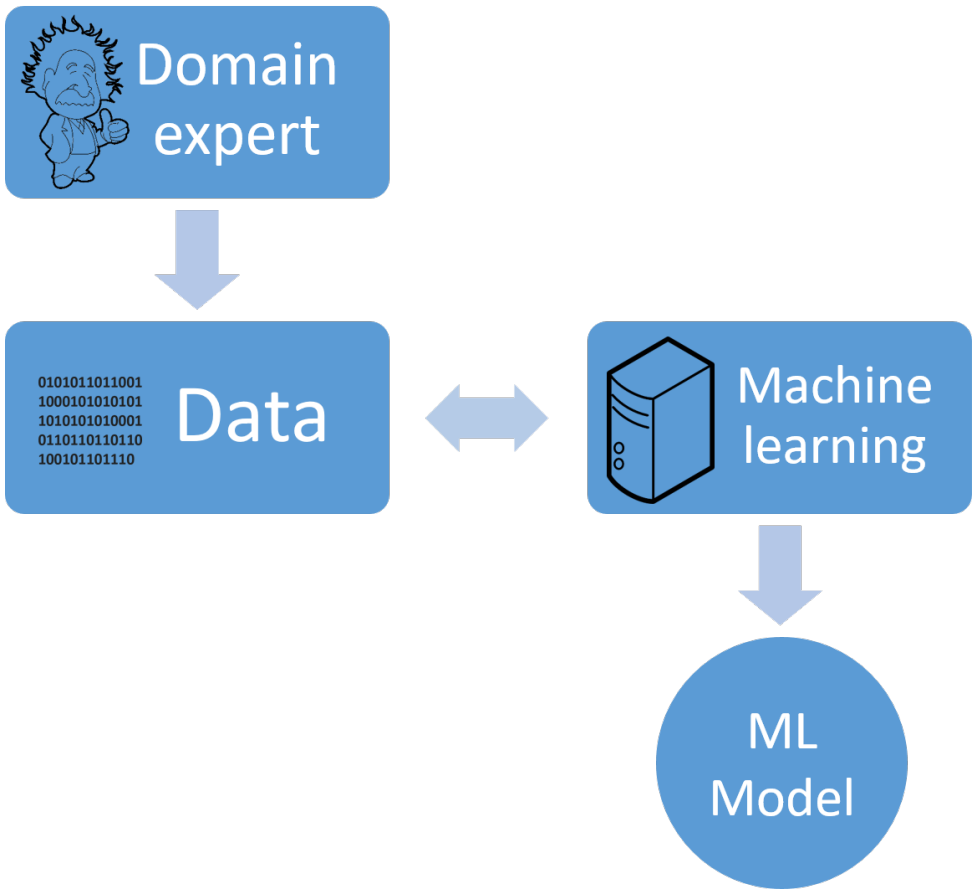
Explore and develop new methods adapted to *small data* problems.

Tracks and research questions:

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

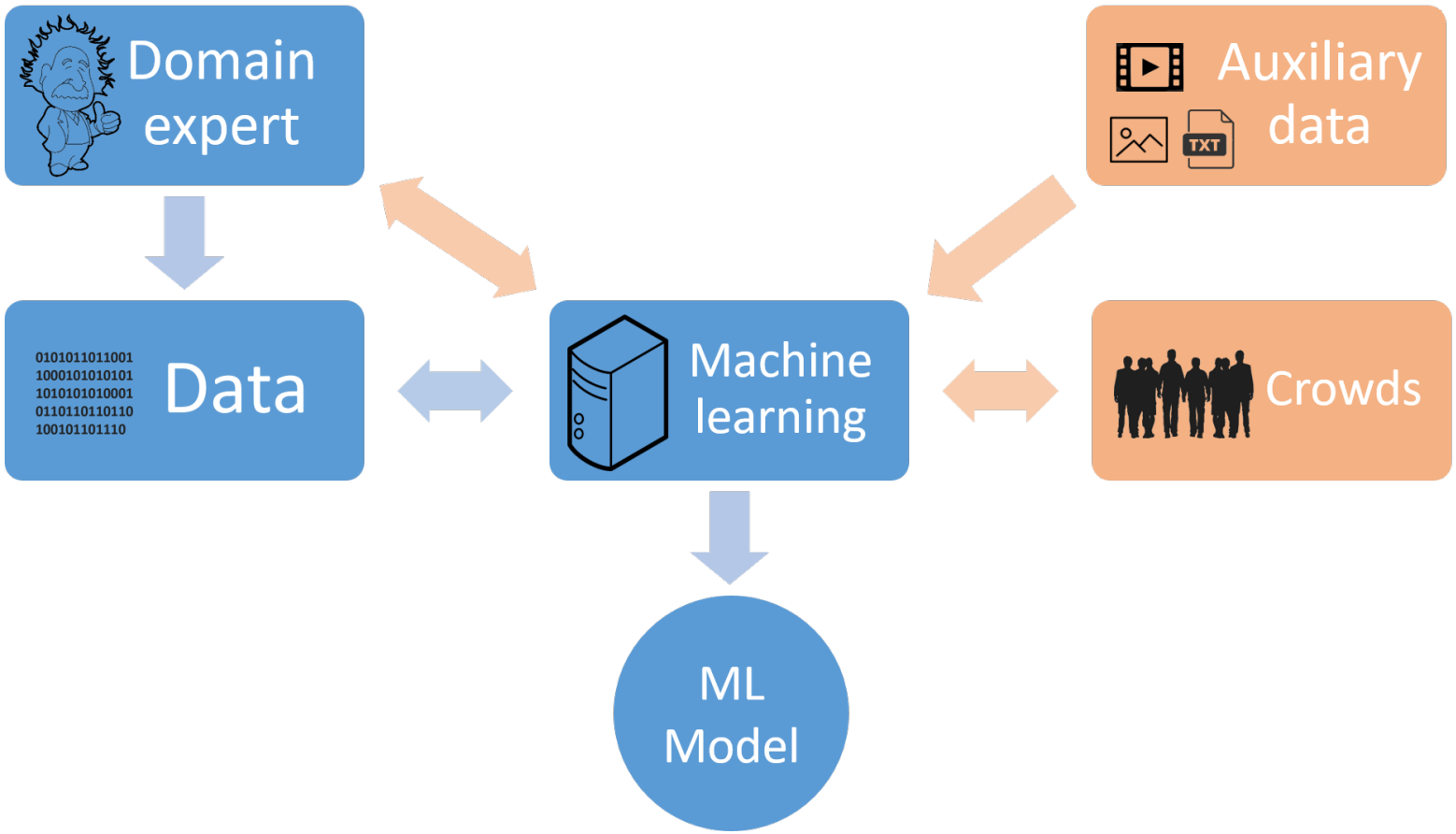
Typical machine learning

From ...



Integrating human(s) and auxiliary data

To ...



Interactive machine learning

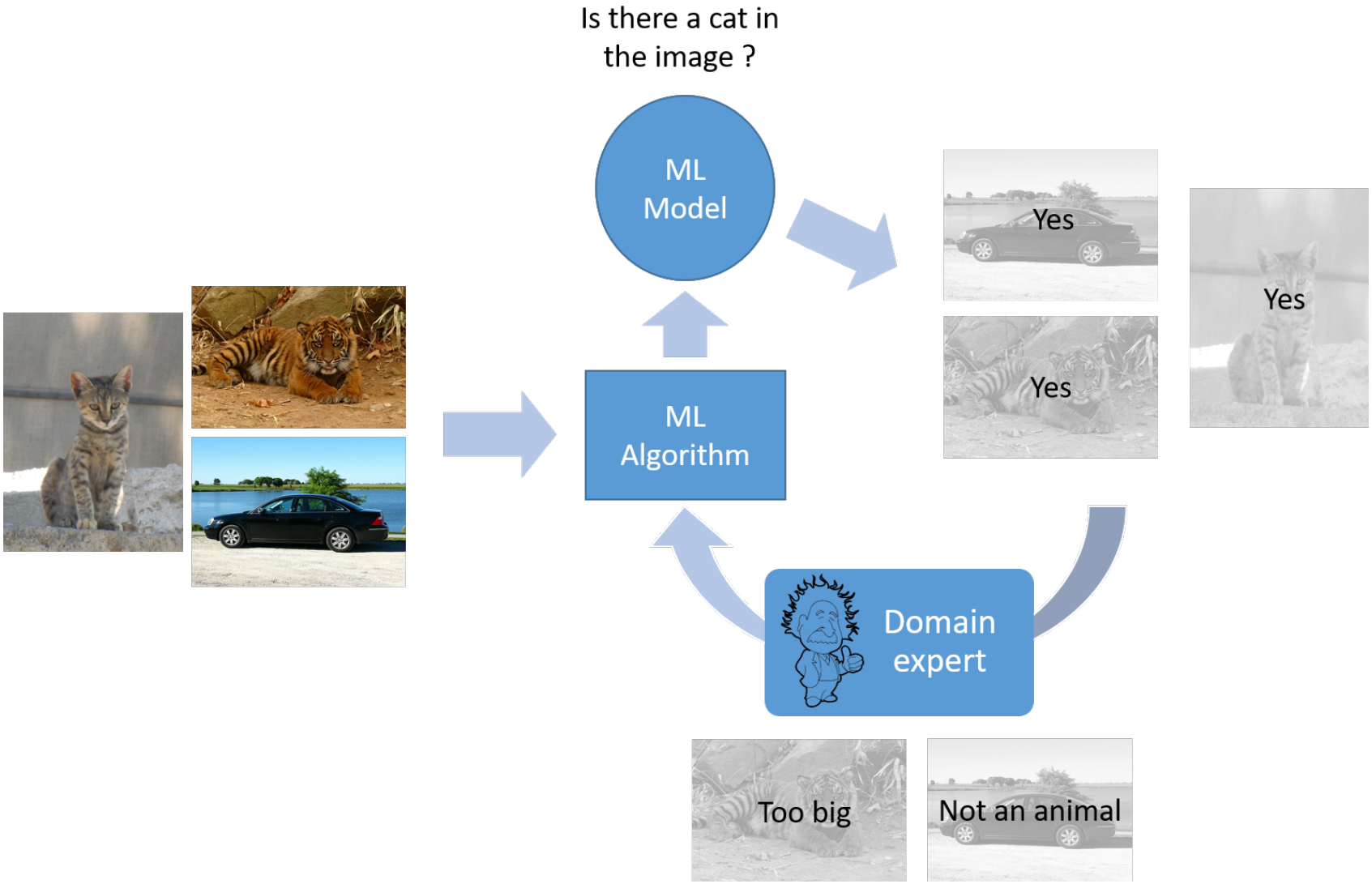
State of the art:

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

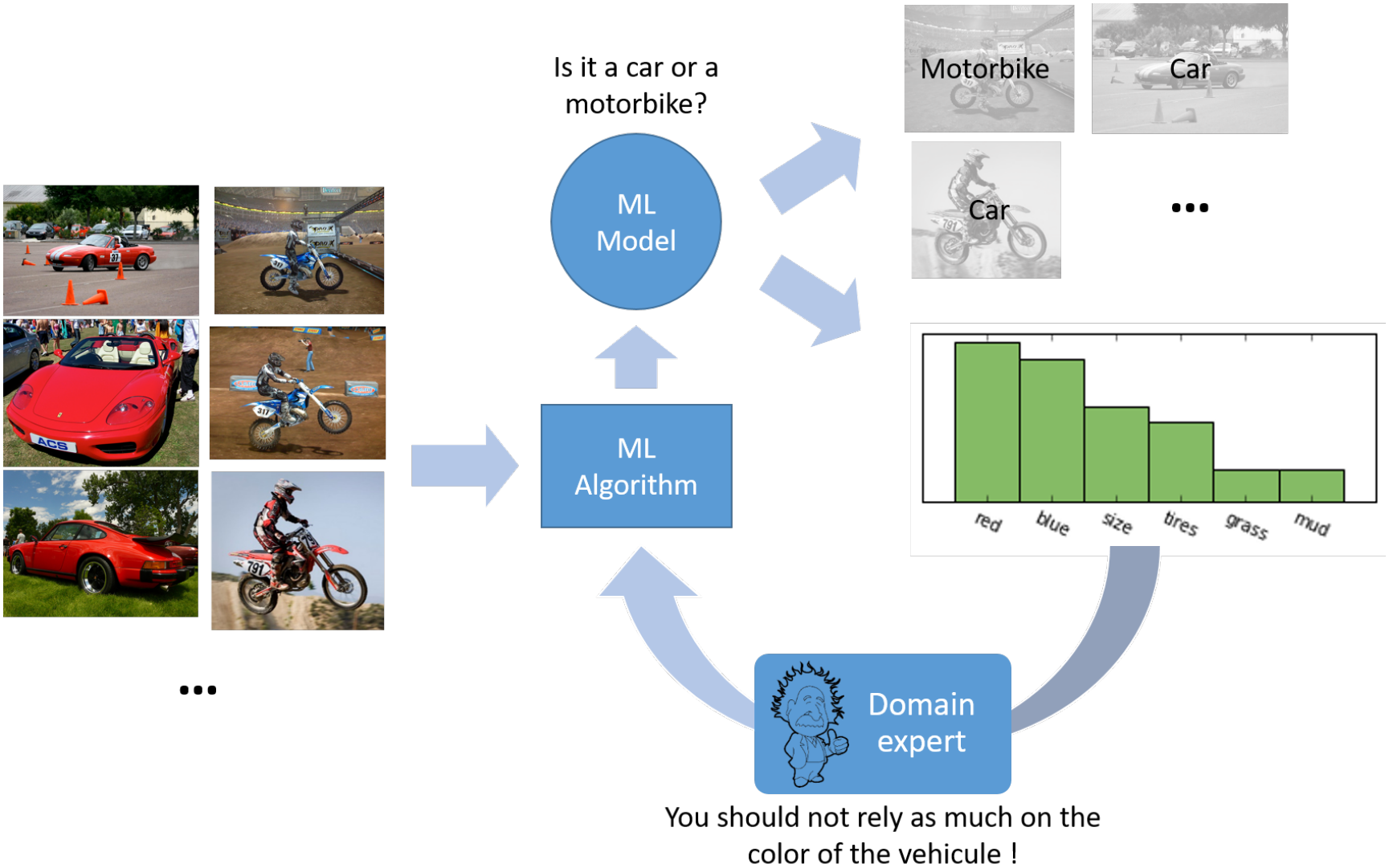
Methodological challenges:

- How to **minimize the number of interactions** ? (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 feedback** to the learning process ?

Error labelling

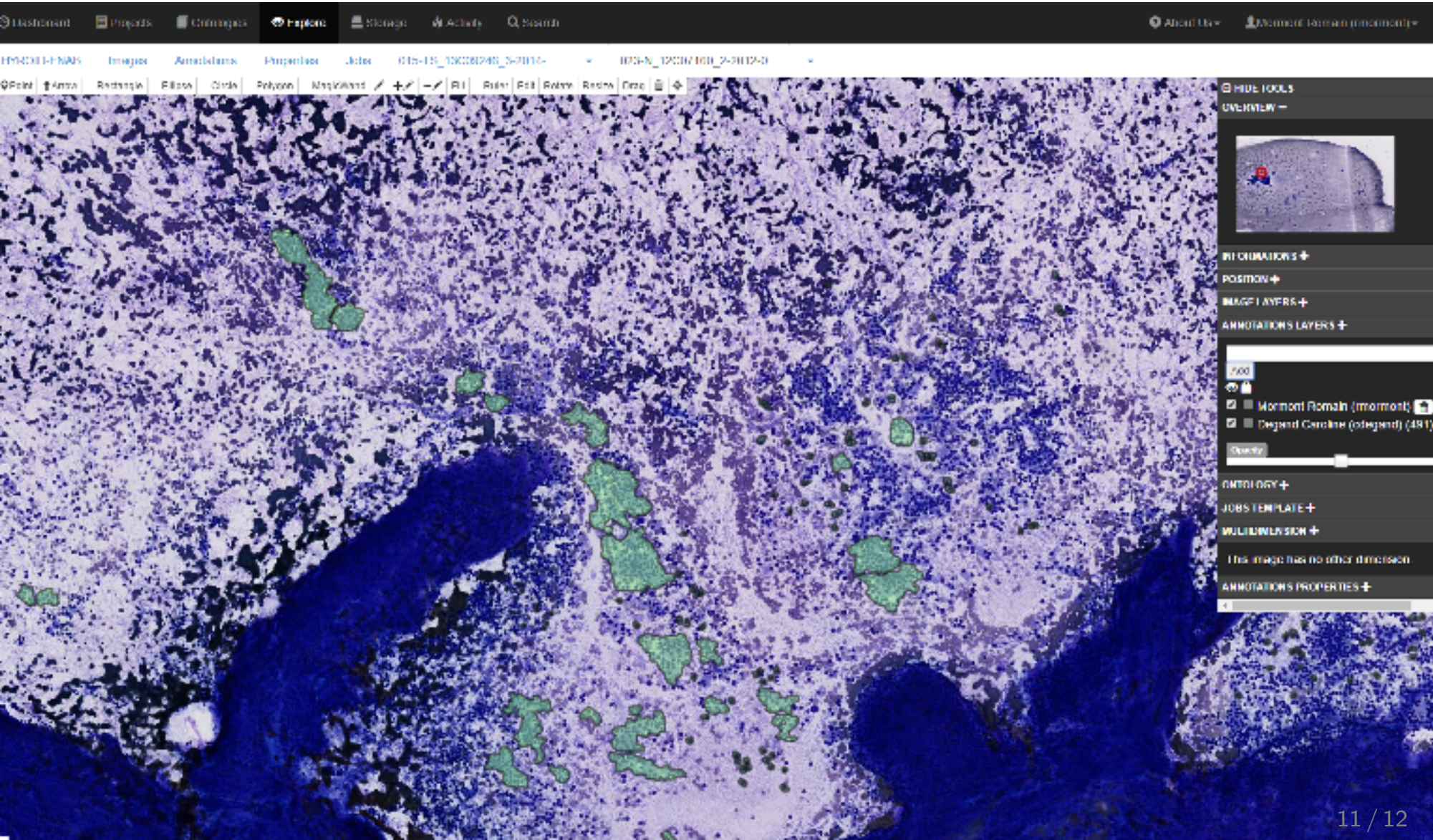


Feature evaluation

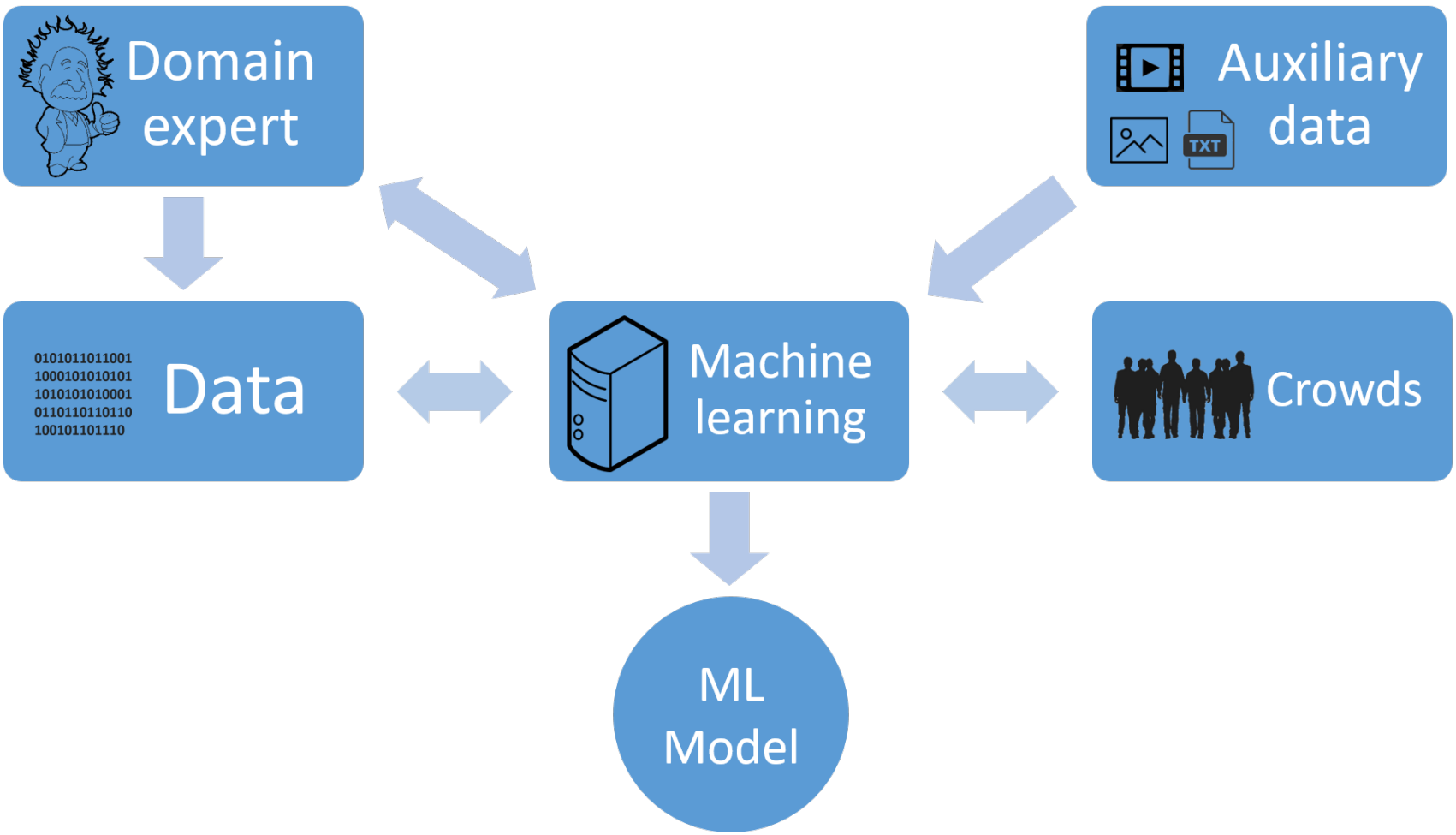


A platform ready for user-interaction

cytomine, an open-source web platform will be used for collecting **feedback** from experts and from the crowds.



Thank you for your attention!

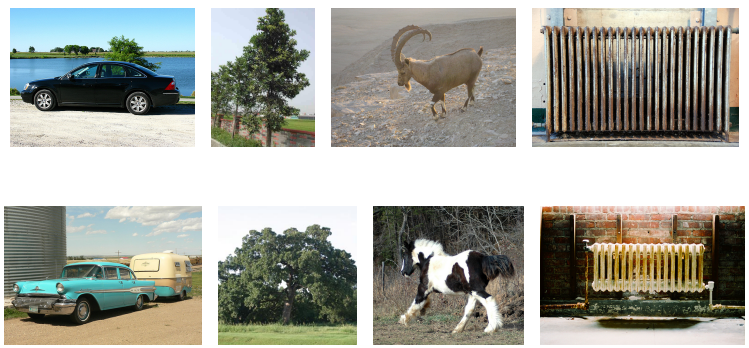


Small data: illustration

Big data
ImageNet[1]

Dataset:

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

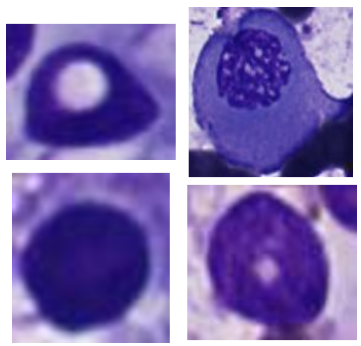


Typical images from ImageNet

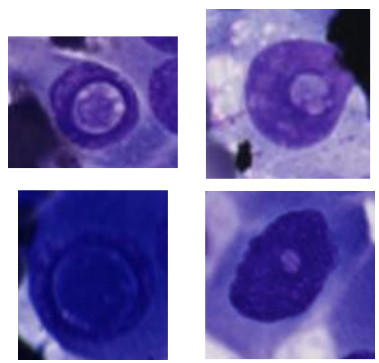
Small data
Thyroid nodule malignancy

Dataset:

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



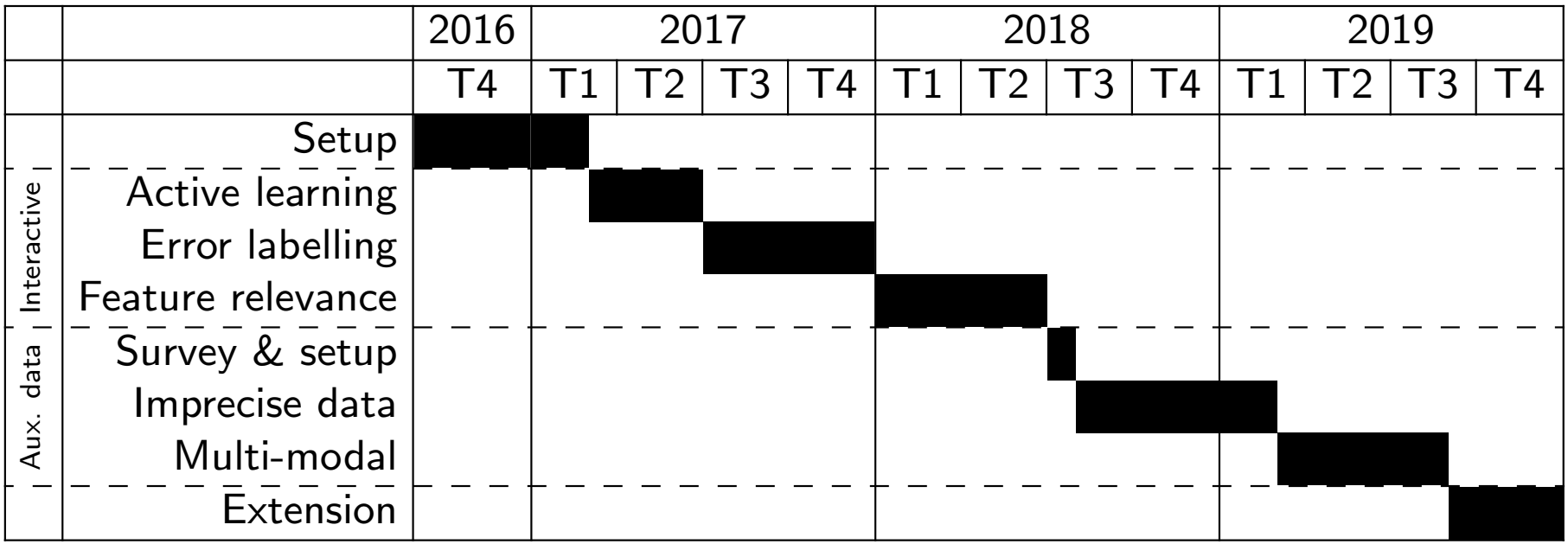
(a) Healthy

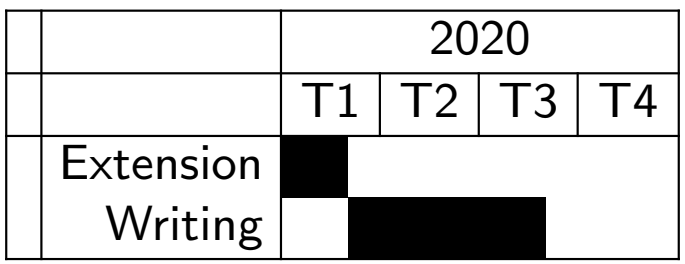


(b) Malignant

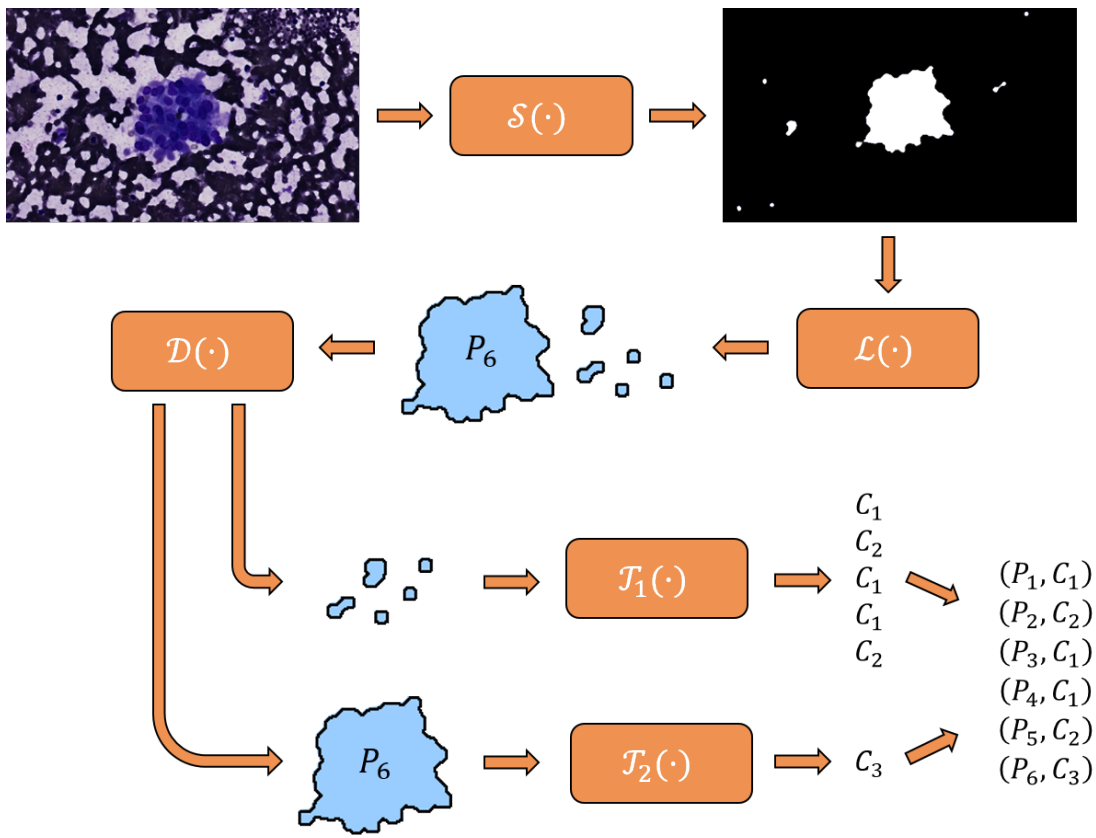
[1] Deng & al. *ImageNet: A Large-Scale Hierarchical Image Database*. In CVPR09, 2009.

Work calendar





A workflow for large-scale computer-aided cytology and its applications.



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 ?

Focus on two settings:

- Imprecise measurements
- Multi-modal images