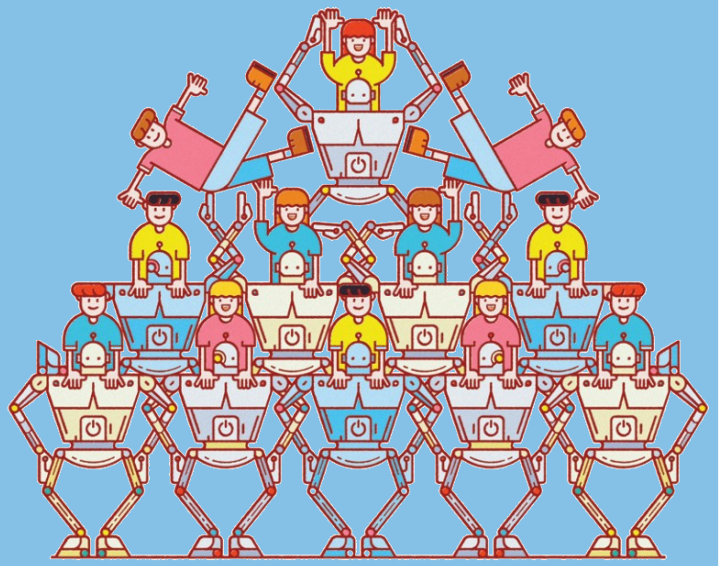
TRAIL
TRUSTED AI LABS
BY DIGITALWALLONIA / SPW-RECHERCHE

WP1: Human-AI interaction
Kickoff meeting



WP leader: Christine Decaestecker

Post-doc: Alberto Franzin

Laboratory of Image Synthesis and Analysis

LISA, ULB

.

Supported by Service public de Wallonie – Recherche under grant n°2010235 « ARIAC BY DIGITALWALLONIA4.AI »

Ocetic Your Connection to ICT Research · Cenaero · Agence du Numérique · sirris · Multitel INNOVATION CENTRE · digital wallonia 4.ai · Université de Namur · UMONS University of Mons · Liège université · ULB Université Libre de Bruxelles · UCLouvain · AI4Belgium · Wallonie recherche SPW

WP1 kickoff meeting: goals

Get to know each other and start building a network of the TRAIL researchers working on topics related to human-AI interaction

Foster collaborations among researchers, in particular PhD students, working on similar topics.

WP1 kickoff meeting: program

14h: Alberto Franzin: presentation of WP1

14h10: Adrien Foucart: DL for histopathology with multiple annotations

14h40: Contributed presentations (10m each): Elias Fernández, Antonio García-Díaz, Romain Mormont, Alberto Franzin

15h20: Question time for contributed presentations, with coffee

15h50: Contributed presentations (3m each + questions): Charlotte Nachtegael, Anaïs Halin, Natarajan Chidambaram, Laura Gálvez Jiménez, Alexandre Englebert

16h25: Christine Decaestecker: bias in AI

17h: Networking and informal discussions

Human-AI interaction: what, why, how

What: Establish and ease relations and interactions between AI and humans (experts or users)

Why: Improve AI performance and open up realistic perspectives on the application of AI in sensitive areas (medicine, justice, etc.)

How: 5 tasks

Human-AI interaction: an example

- Supervised learning makes use of labeled data

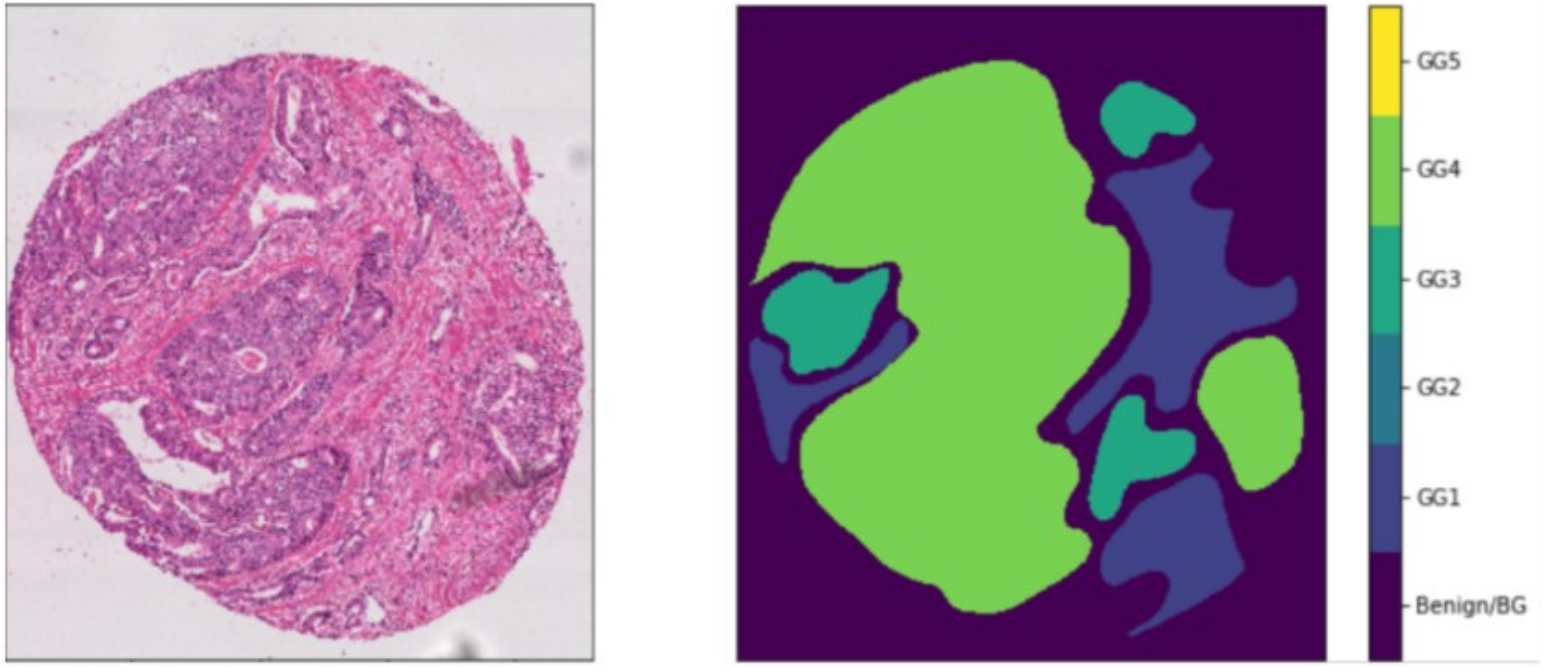


Figure 2.3: Semantic segmentation of a PCa sample, core image (left) and region segmentation and classification (right)

Human-AI interaction: an example

- But how can we be sure the labels are correct?

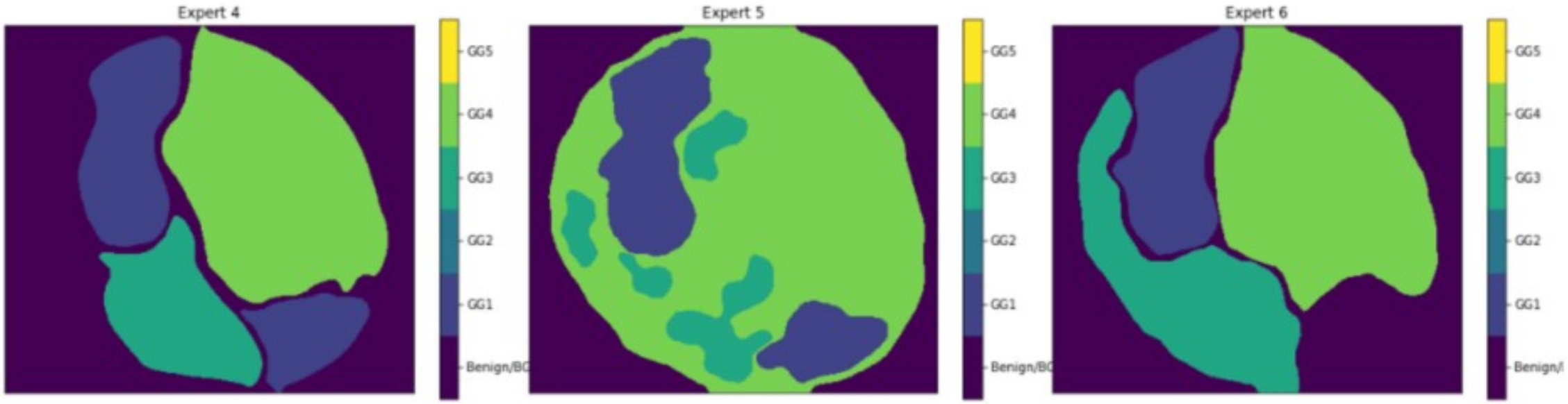


Figure 2.4: Illustration of the high-inter rater variability in PCa grading.

Human-AI interaction: an example

- The same image is labeled differently when different **experts** look at it
- More generally, real data, and particularly labels used as supervision in machine learning algorithms, can contain errors
- What kind of results can we expect from an algorithm fed with (possibly) contradictory or even wrong data?

Human-AI interaction: an example

- Possible solutions:
  - Collect correct dataset → infeasible or impossible?
  - Evaluate and correct a posteriori → infeasible or impossible?
  - **Involve human operators in the process, to control each step, to provide bidirectional feedback and explanations**
  - **Reduce the need of a single and very precise supervision and increase robustness to errors**

Human-AI interaction

- A necessary paradigm to ensure:
  - Correctness of results
  - Control of side effects
  - Adherence to goals
  - The possibility of taking action

- Hence, human-AI interaction is a key part of Trustworthy AI systems in real world applications (see also the other WPs)

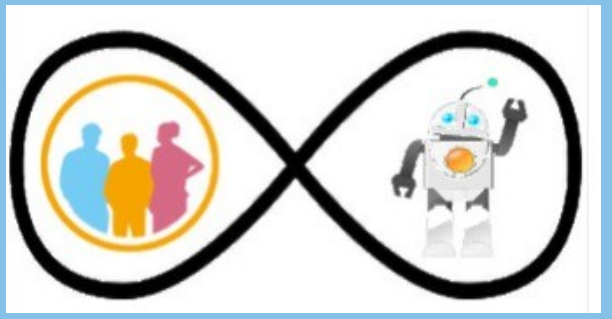
Human-AI interaction

High-risk AI systems will be subject to **strict obligations** before they can be put on the market:

- **Adequate risk assessment and mitigation systems;**
- **High quality of the datasets** feeding the system to minimise risks and discriminatory outcomes;
- **Logging of activity to ensure traceability of results**;
- **Detailed documentation** providing all information necessary on the system and its purpose for authorities to assess its compliance;
- **Clear and adequate information** to the user;
- **Appropriate human oversight** measures to minimise risk;
- High level of **robustness**, **security** and **accuracy**.

https://ec.europa.eu/commission/presscorner/detail/en/IP_21_1682

WP1: 5 tasks



1.1 Human-in-the-loop: human helps AI

- Provide continuous bidirectional feedback
- Progressively improve the datasets (labels, features, examples), their (pre)processing, and finally the AI performance

1.2 AI-in-the-loop: AI helps human

- to optimize complex systems by solving and organizing tasks in a complex workflow
- in the maintenance and adaptation of such systems in an evolving context

WP1: 5 tasks

1.3 Consensus mechanisms

- Managing conflicting opinions of human experts in an unbiased manner

1.4 Data from multiple sources and of varying quality

- Make the best use of such data, reduce the need of «perfect» annotations, increase robustness to errors

1.5 Explainable AI decision and knowledge extraction

- Increase trust in systems and their results (see also WP2)

WP1 contact

- alberto.franzin@ulb.be
- christine.decaestecker@ulb.be
- LISA: 02 650 2291
- ariacwp1@googlegroups.com
- Suggestions?