FGAI4H-O-047

Berlin, 31 May – 2 June 2022

Source:	Institute for Molecular Medicine Finland – FIMM
Title:	Workshop: TG-POC & TG-Histo - When is AI good enough for implementation in diagnostics?
Purpose:	Discussion

Contact: Johan Lundin
Institute for Molecular Medicine Finland – FIMM
University of Helsinki, Finland
Karolinska Institutet, Sweden

E-mail: johan.lundin@helsinki.fi

Abstract: This PPT contains a presentation from the TG-POC & TG-Histo workshop on "Validation of annotations for AI models within the scope of point-of-care diagnostics (POC)"

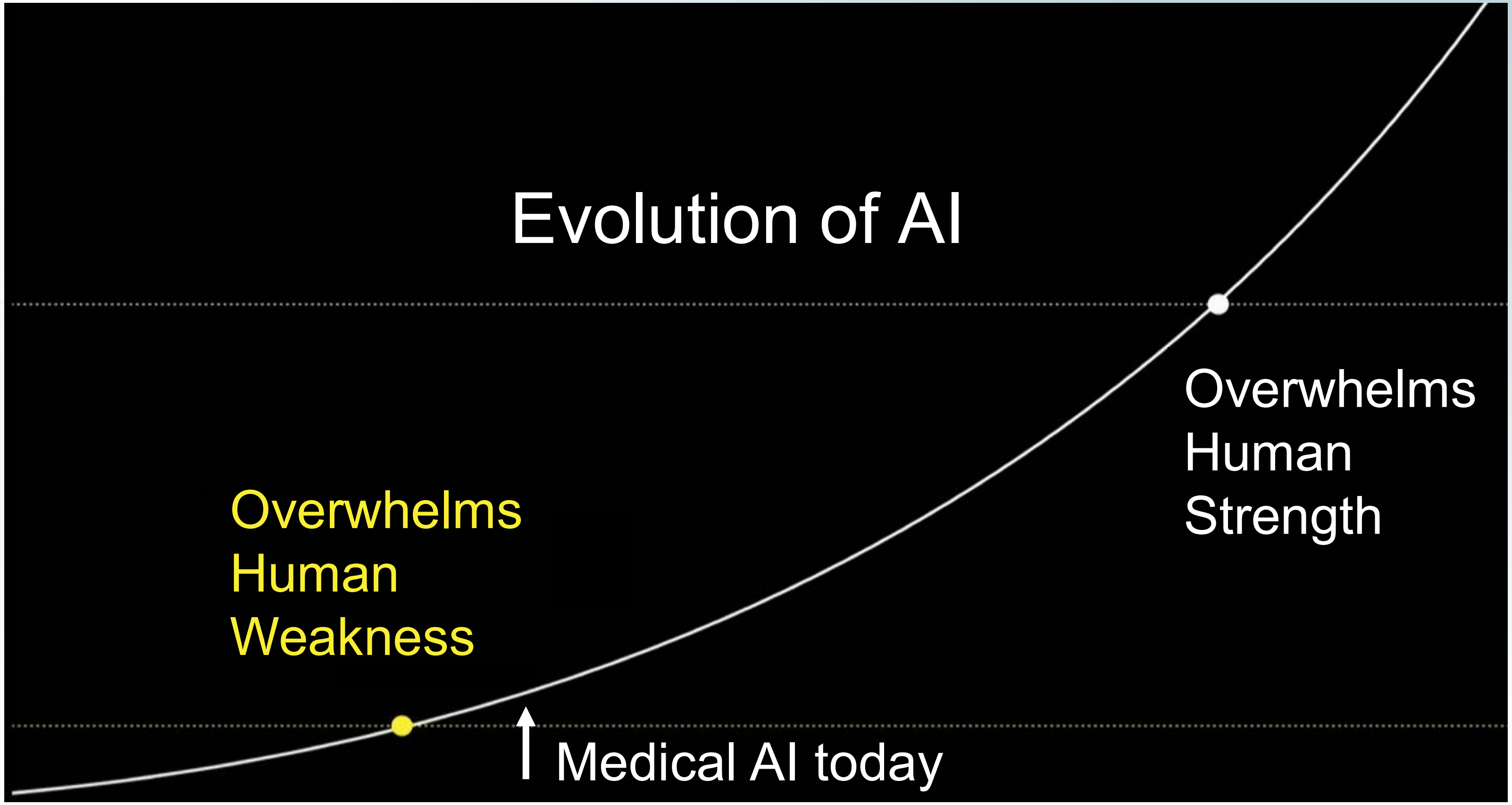
FIMM
Institute for Molecular Medicine Finland
Nordic EMBL Partnership for Molecular Medicine

When is AI good enough for implementation in diagnostics?

Johan Lundin, MD, PhD
Research Director
Institute for Molecular Medicine Finland – FIMM
University of Helsinki, Finland
and
Professor of Medical Technology
Karolinska Institutet, Sweden

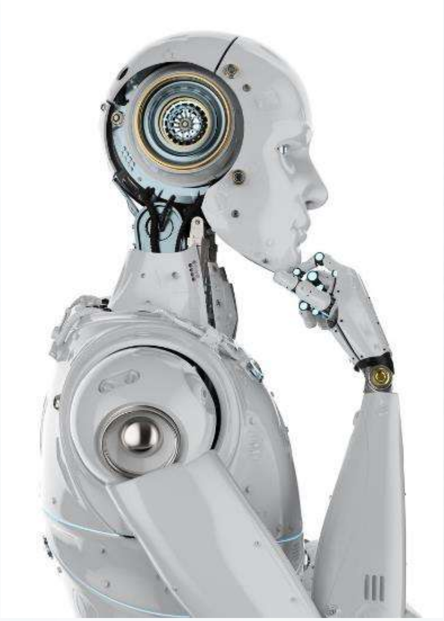
Disclaimer

- Founder, co-owner and board member of Aiforia Technologies



Courtesy of Tristan Harris, Center for Humane Technology

Human chess world champion learns from games played by AI



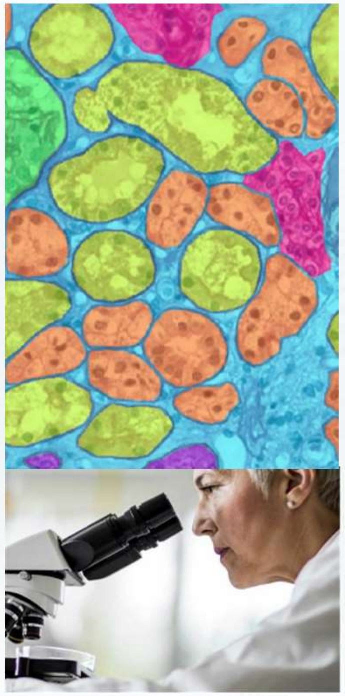
AlphaZero, self-taught AI-based world champion of chess*

Magnus Carlsen, the current human champion of chess

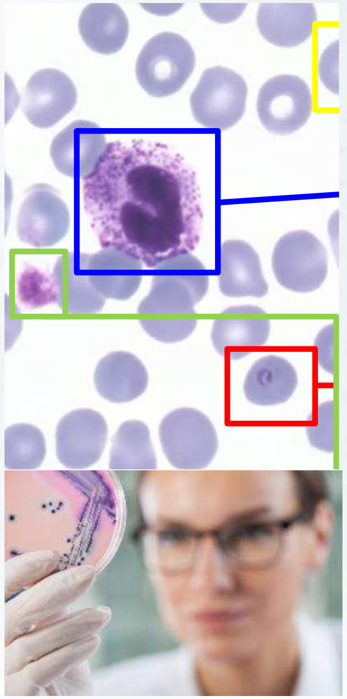
Medical experts are likely to benefit from AI in a similar way

*Silver, D. et al. Science. 362, 1140 (2018).

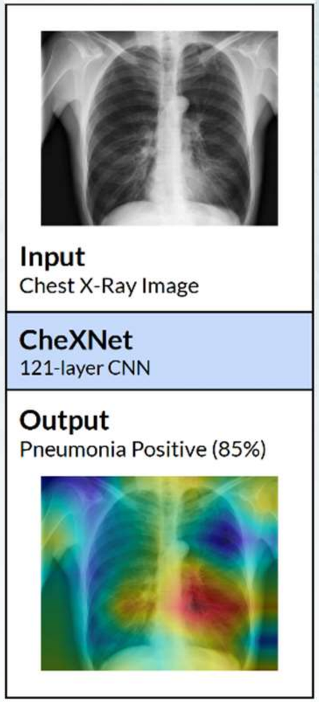
AI will impact all medical fields where an expert makes a visual interpretation



Pathology



Microbiology

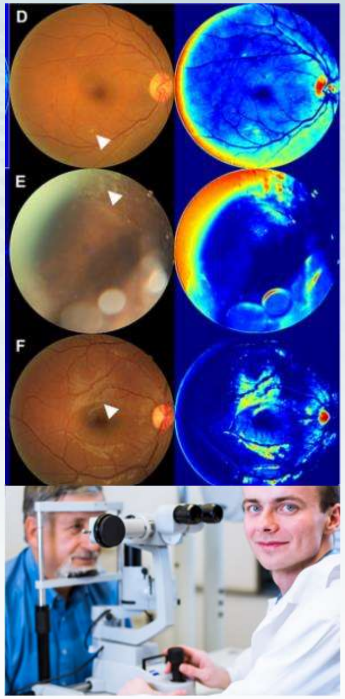


Input
Chest X-Ray Image

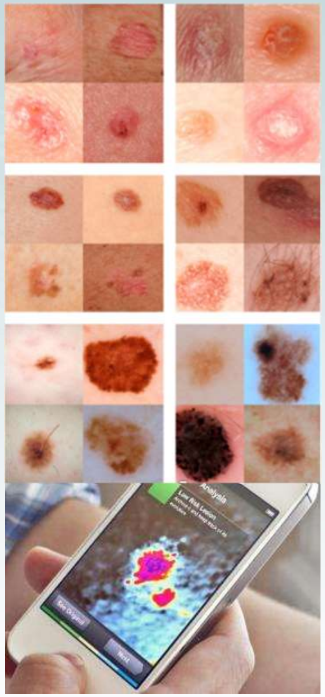
CheXNet
121-layer CNN

Output
Pneumonia Positive (85%)

Radiology



Ophthalmology



Dermatology

Crude estimate: minimum >7-8 billion visual diagnostic assessments globally per year

When is AI good enough for implementation in diagnostics?

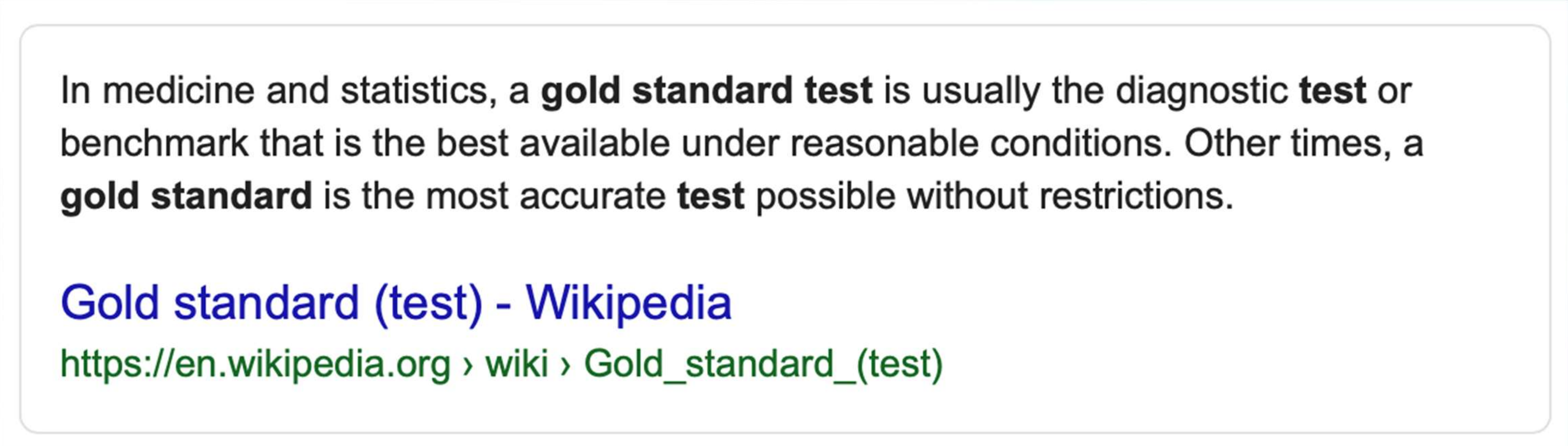
- When AI achieves the same accuracy or exceeds the current gold standard?
- When AI analyses an order of magnitude more samples than a human expert within a particular time period?
- When AI complements the human expert and finds a significant number of targets that otherwise would have been missed?
- When AI is the only alternative in a setting with shortage of experts?

Ground truth and gold standard tests

- The term ground truth refers to the underlying absolute state of information
- The gold standard strives to represent the ground truth as closely as possible.
- In machine learning and information retrieval, "ground truth" is the preferred term even when classifications may be imperfect
- The gold standard is assumed to be the ground truth

Text from Wikipedia

Definition of a gold standard test



In medicine and statistics, a **gold standard test** is usually the diagnostic **test** or benchmark that is the best available under reasonable conditions. Other times, a **gold standard** is the most accurate **test** possible without restrictions.

Gold standard (test) - Wikipedia

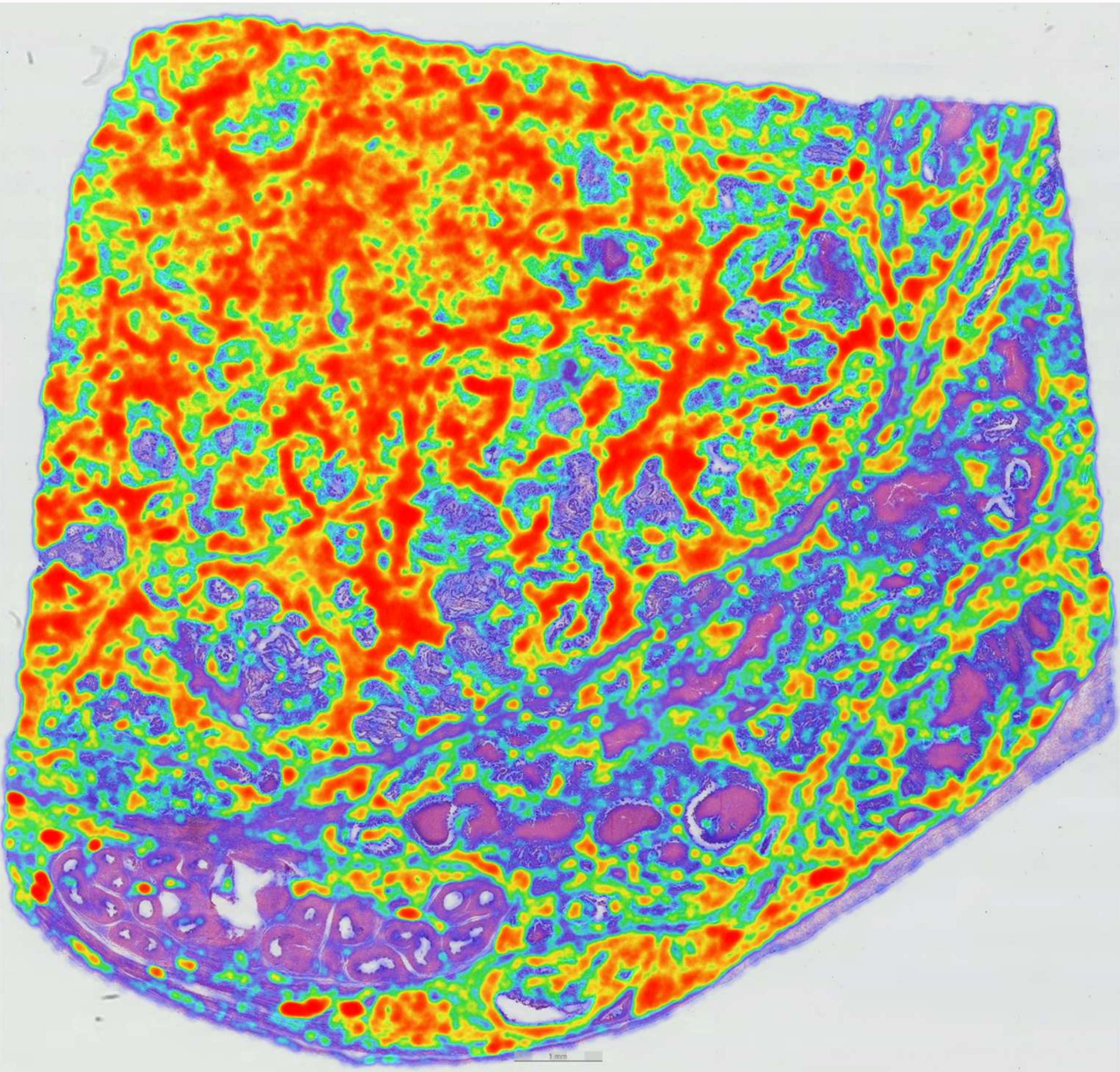
https://en.wikipedia.org › wiki › Gold_standard_(test)

Performance of a gold standard test

- A hypothetical ideal "gold standard" test has a sensitivity of 100% with respect to the presence of the disease and a specificity of 100%.
- In practice, there are sometimes no true gold standard tests.
- Currently, no gold standard tests exist for deep learning applied to pathology or microscopy
- According to the literature, AI-based algorithms typically reach a good to excellent diagnostic accuracy as compared to the ground truth, but the ground truth is rarely a real gold standard

Challenges in the development of gold standard for machine learning in pathology

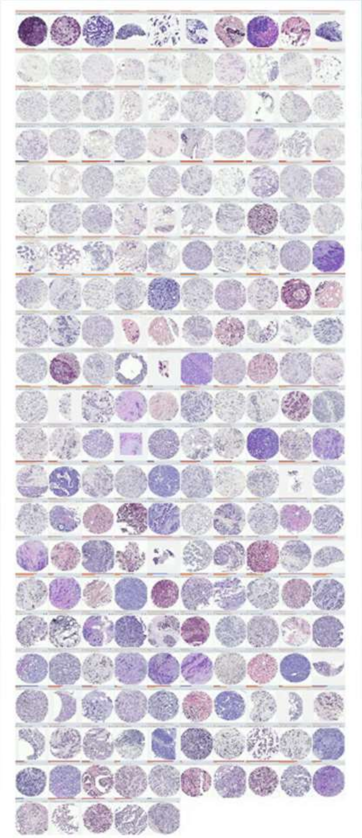
- Samples and data
- Annotations
- Algorithms
- Other challenges



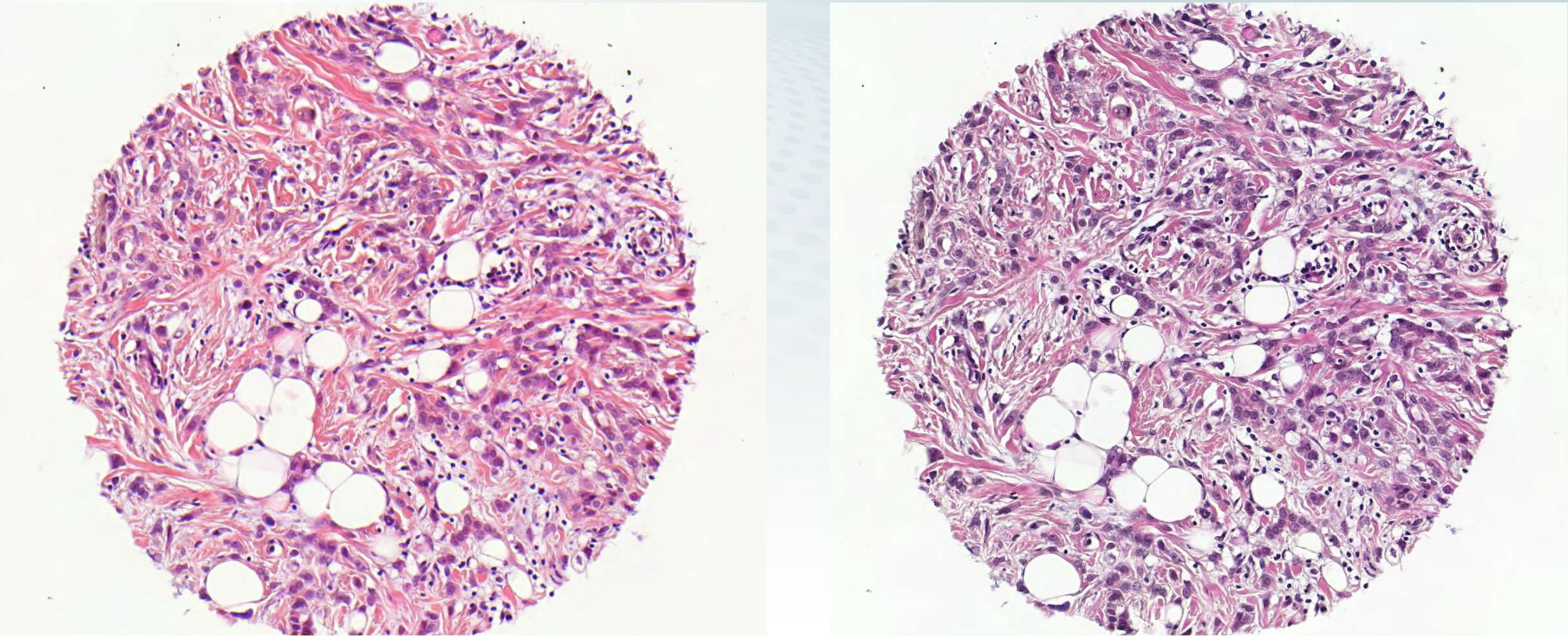
Challenges related to samples, data and annotations

- Limited access to sample images with associated clinical data
  - Share data, form joint projects, use federated or swarm learning
- Lack of annotated images
  - Create public libraries and common repositories of annotated images
- Biased data due to incompleteness or lack of diversity
  - Strive for completeness of data, collect from many centers
- Variable quality, artifacts and heterogeneity of samples
  - Perform quality control (with AI?), re-cut, re-stain, re-scan, color calibrate
  - ..or include all types of artifacts and variabilities in the training set

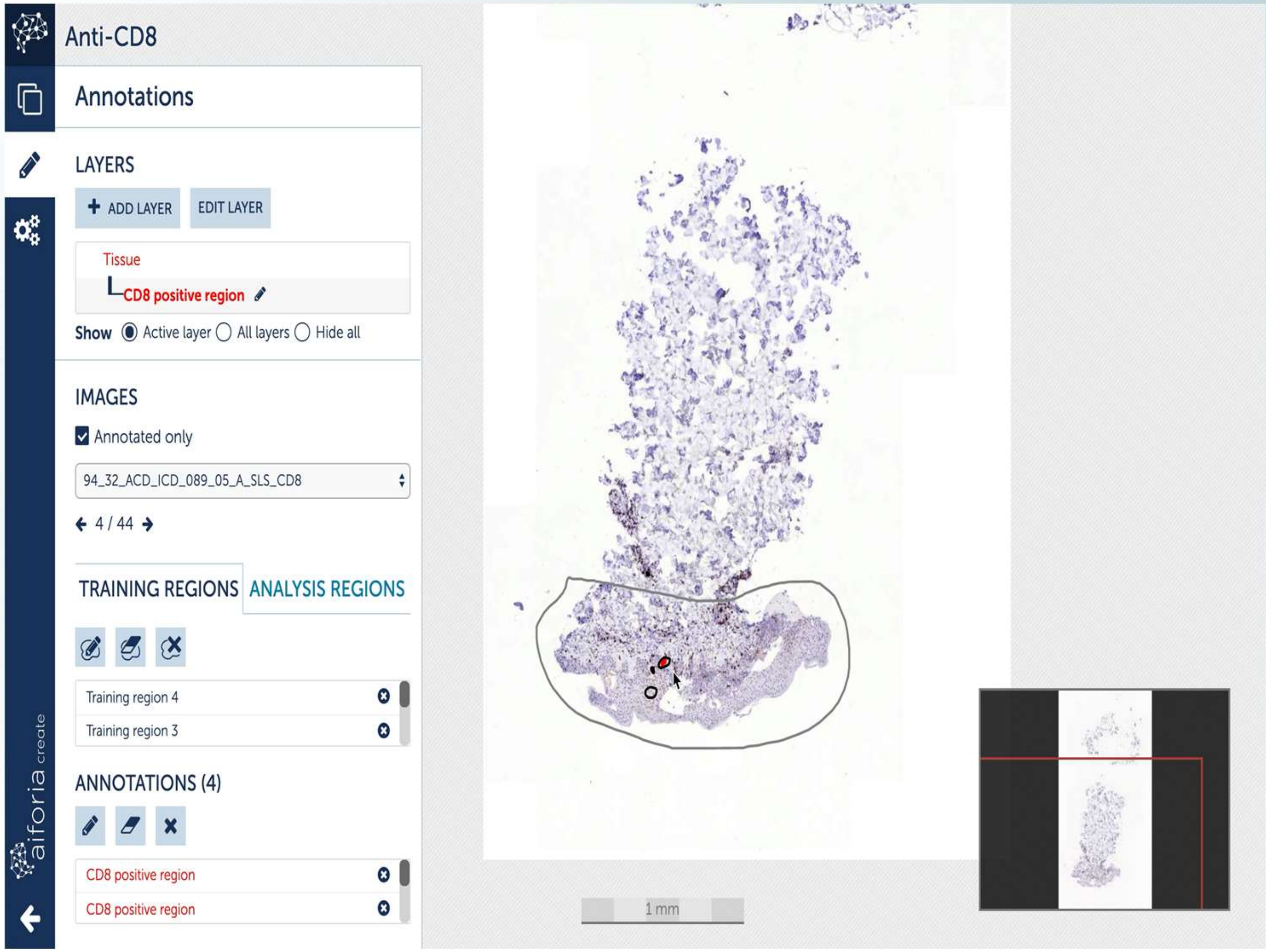
Sample variability in a breast cancer tissue microarray



Variability due to the scanner and camera



Variable quality of annotations



Anti-CD8

Annotations

LAYERS

Tissue
└CD8 positive region

Show ● Active layer ○ All layers ○ Hide all

IMAGES

☑ Annotated only

94_32_ACD_ICD_089_05_A_SLS_CD8

4 / 44

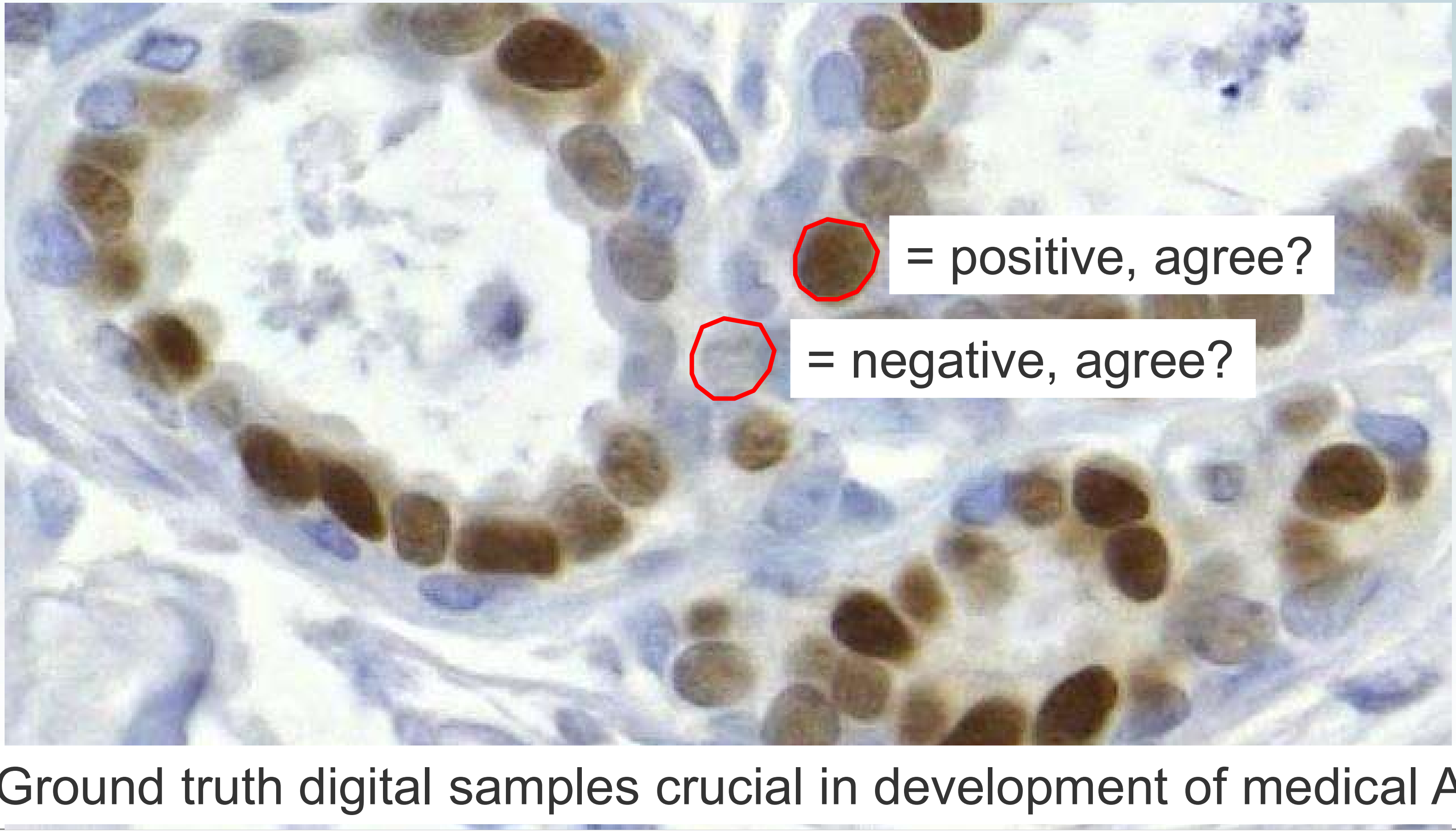
TRAINING REGIONS ANALYSIS REGIONS

- Training region 4
- Training region 3

ANNOTATIONS (4)

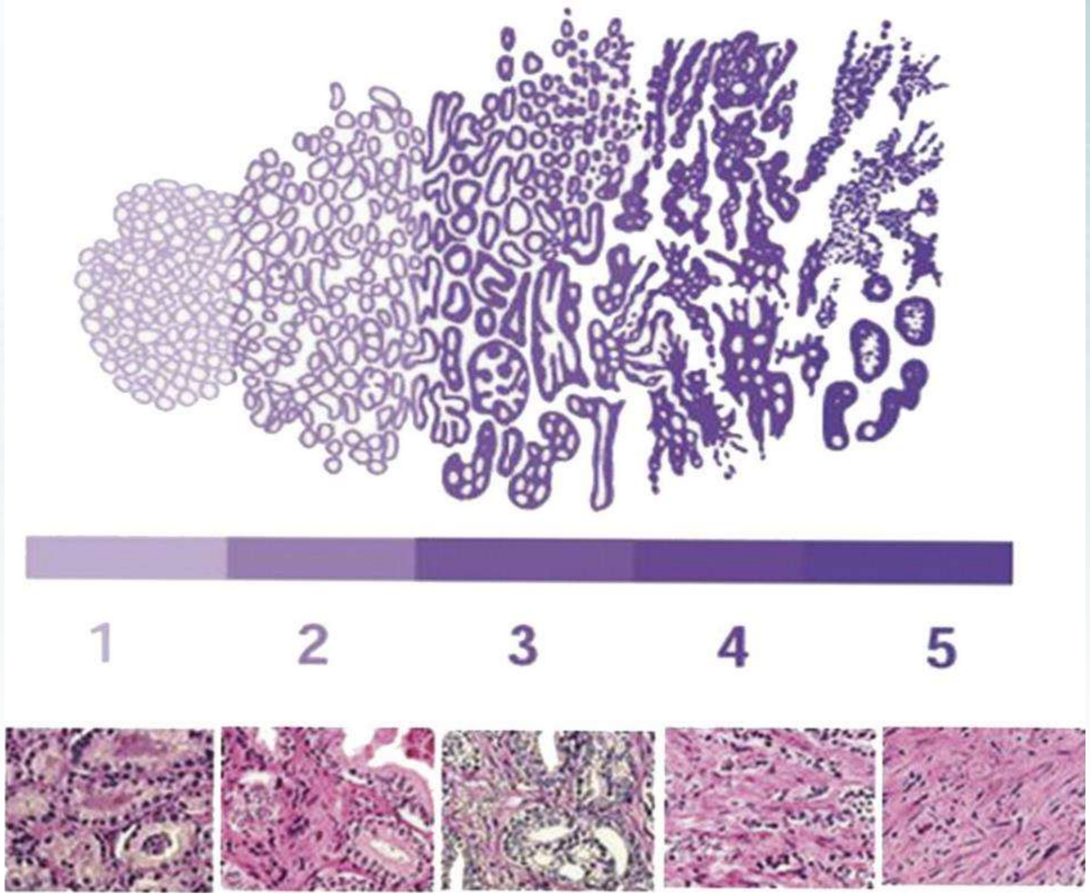
- CD8 positive region
- CD8 positive region

Consistency and representativeness of ground truth -deep learning algorithms just as good as their teacher?

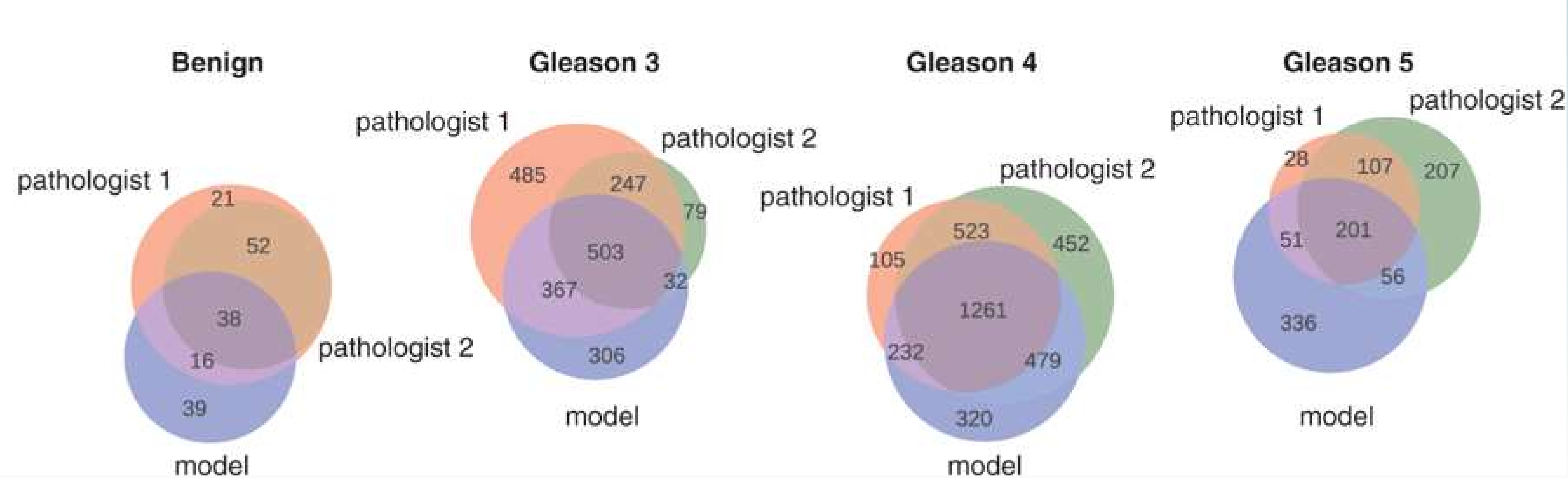


Ground truth digital samples crucial in development of medical AI

More complicated patterns – e.g. Gleason grade in prostate cancer



There is a risk that AI will be just another subjective ”expert” - example of automated grading of Gleason

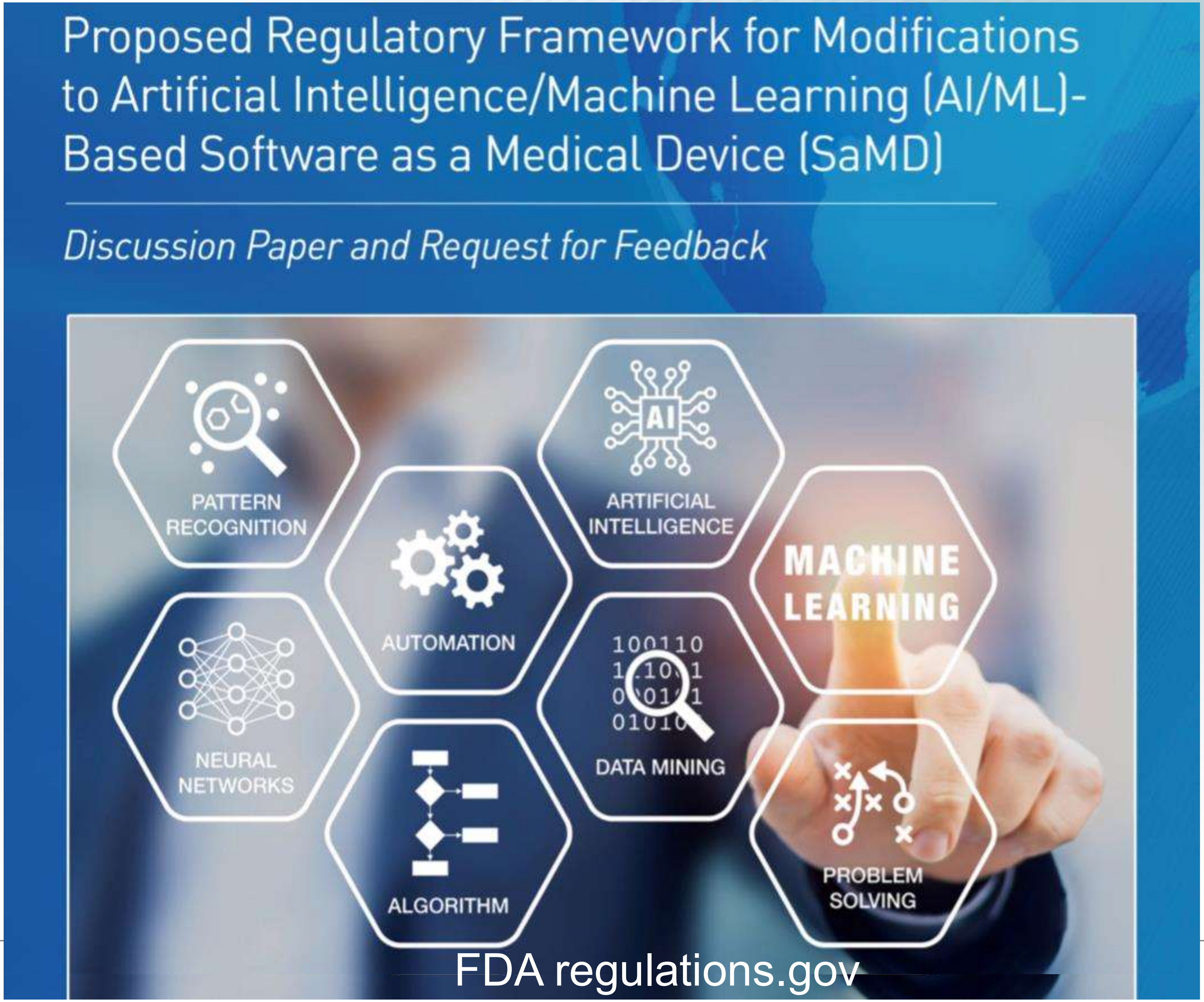


🟠 Pathologist 1 🟢 Pathologist 2 🔵 Model = AI

Arvaniti: Scientific Reports 2018; 8: 12054

Other challenges related to AI for diagnostics

- Ownership and access to the images
  - Create public libraries of annotated images for developers
- What to do with rapidly improving and updated algorithms?
  - Allow algorithm performance to be a moving target
  - FDA white paper
- How to handle tens or hundreds of AI:s for the same purpose?
  - Create consensus algorithms? Use swarm learning to adjust parameters?



Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)

Discussion Paper and Request for Feedback

FDA regulations.gov

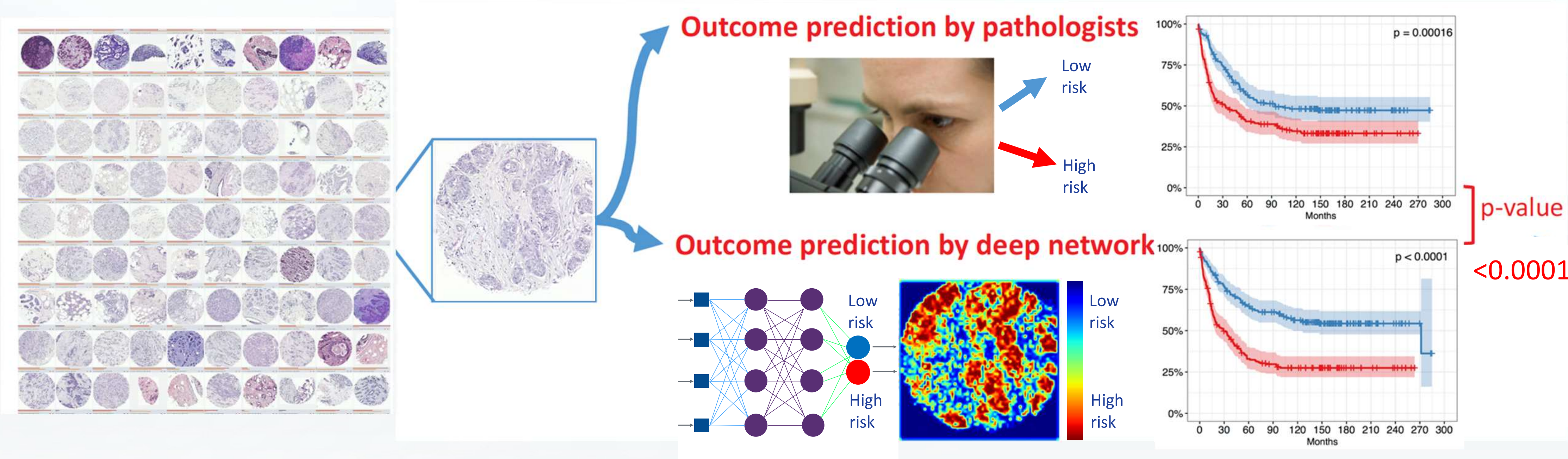
Human vs machine

	Human	Machine
› Understand context	1	2
› Reproduce assessment	2	1
› Handle outliers	1	2
› Find rare events	2	1
› Generalize	1	2
› Achieve high throughput	2	1

When is an AI algorithm good enough?

- 🤔?
- Needs to outperform or supplement human experts in at least one of the following: Sensitivity, speed, reproducibility
- Generalizability needs to be established
  - robustness to artifacts, outliers and local variations in protocols
- Superiority shown in prediction of clinical endpoints rather the replication of annotations?
  - Outcome and biomarker supervised learning

Outcome supervised learning in colorectal and breast cancer



Outcome prediction by pathologists

Low risk

High risk

Outcome prediction by deep network

Low risk

High risk

Low risk

High risk

p-value

<0.0001

Digitized tissue samples from cancer patients with known outcome of cancer i.e. survivor or non-survivor

Comparison of human expert-based and AI-based outcome prediction

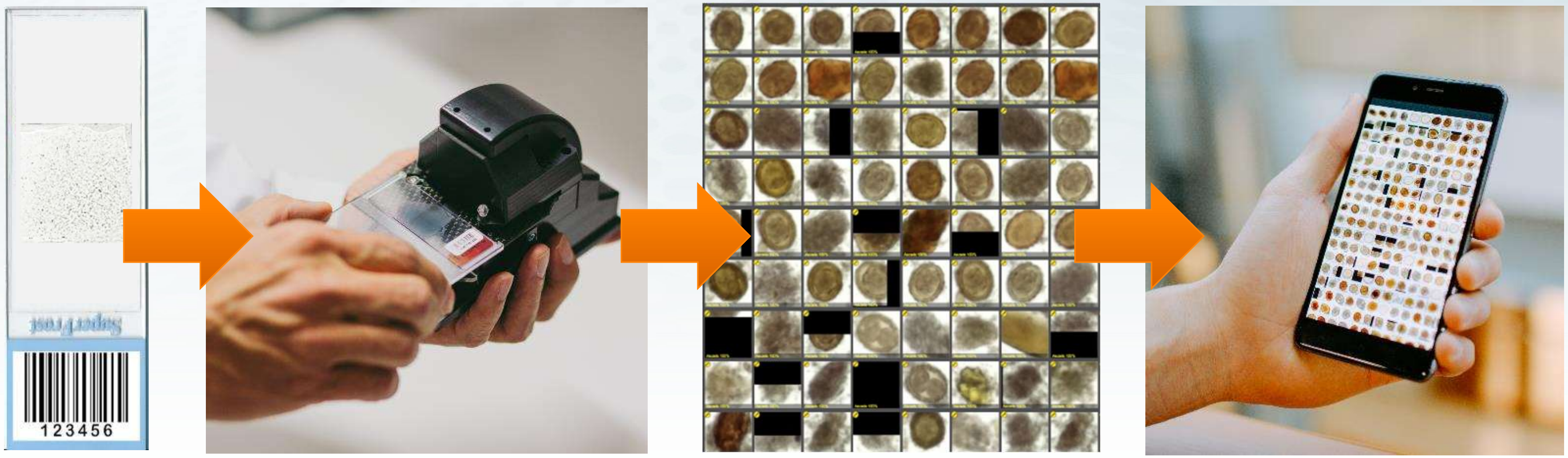
Deep learning outperformed experienced pathologists in outcome prediction [1-3]

[1] Bychkov et al, Scientific Reports 2018;8:3395 [2] Turkki et al, Breast Can Res Tr 2019;177:41-52 [3] Bychkov et al, J Pathol Informatics 2022;13:9

ONCOSYS Research Program
Faculty of Medicine, UH

Human and machine combination:
Sensitive AI algorithm – specific human observer

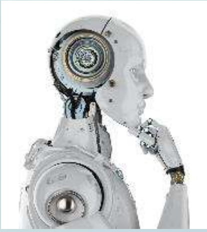
- Example of application to neglected tropical diseases for better access to diagnostics
- Assisted detection of helminth eggs in stool samples and verification by human expert



Stool sample — Scan with mobile microscope — Parasite egg candidates — Human expert reviews results

Global Health Action 2017; 10:1337325.

Human **AND** machine?



	Human	Machine	Human&Machine
› Understand context	1	2	1
› Reproduce assessment	2	1	1
› Handle outliers	1	2	1
› Find rare events	2	1	1
› Generalize	1	2	1
› Achieve high throughput	2	1	1

Some arguments and thoughts as a conclusion

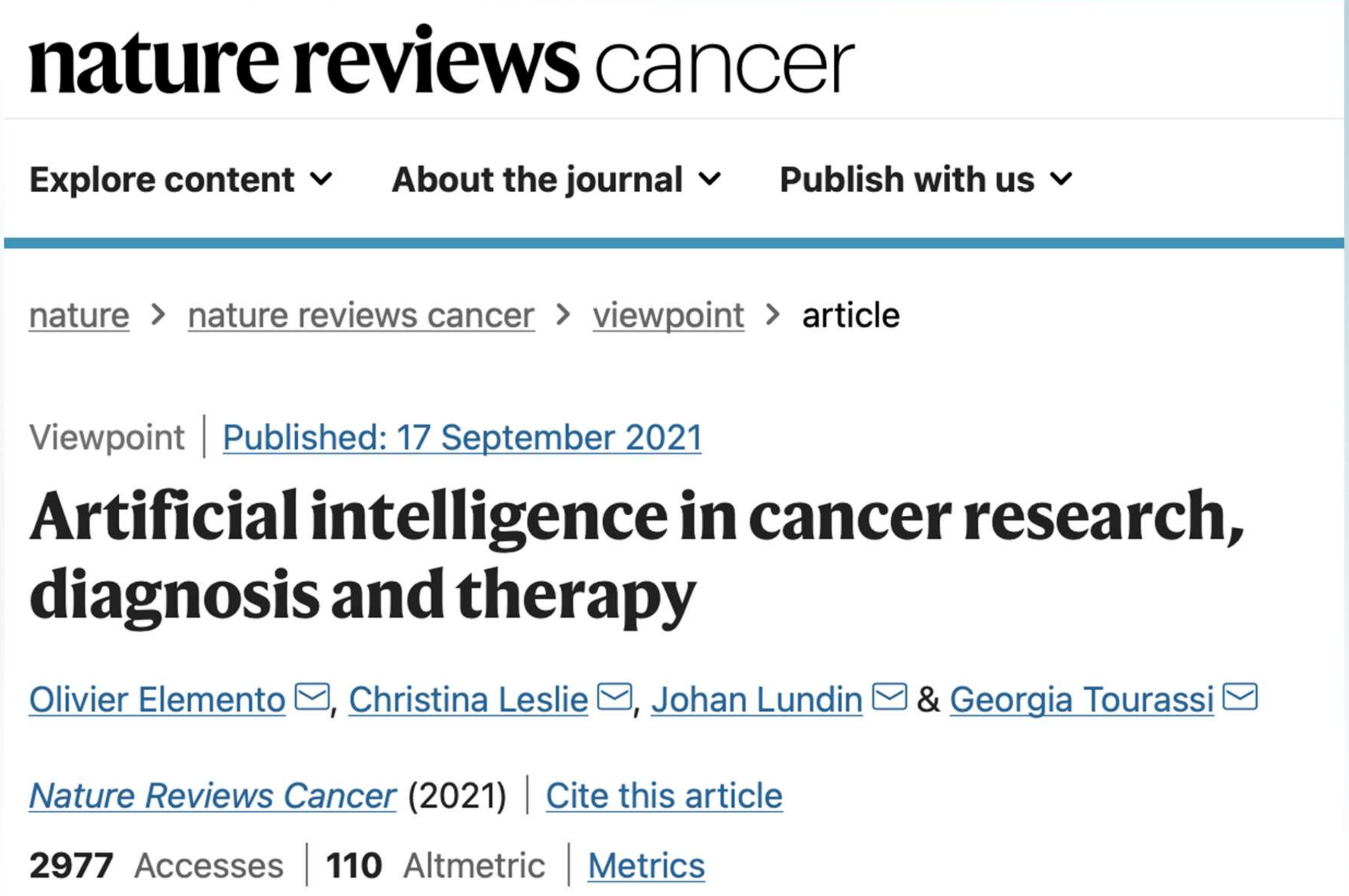
› The quality, selection and annotation of training data for AI is crucial

› How could we get more reliable ground truth?

› There is a risk that we end up with hundreds of AIs for the same purpose but with unclear accuarcy

› Ground truth based on human observation and annotation will always be subjective

› Alternative ground truths, such as sample level diagnosis, patient outcome and response to treatment should be explored



nature reviews cancer

Viewpoint | Published: 17 September 2021

Artificial intelligence in cancer research, diagnosis and therapy

Olivier Elemento, Christina Leslie, Johan Lundin & Georgia Tourassi

Nature Reviews Cancer (2021) | Cite this article

2977 Accesses | 110 Altmetric | Metrics

FIMM
Institute for Molecular Medicine Finland
Nordic EMBL Partnership for Molecular Medicine

Lundin Group, FIMM

Vetenskapsrådet

SIGRID JUSELIUS STIFTELSE

FAMILJEN ERLING-PERSSONS STIFTELSE

BF Biocenter Finland

HiLIFE HELSINKI INSTITUTE OF LIFE SCIENCE

Karolinska Institutet

iCAN

ONCOSYS Research Program
Faculty of Medicine, UH

FIMM

Institute for Molecular Medicine Finland
Nordic EMBL Partnership for Molecular Medicine

Building a bridge from discovery to medicine

www.fimm.fi