FGAI4H-O-050

Berlin, 31 May – 2 June 2022

Source:	Institute for Molecular Medicine Finland – FIMM
Title:	Workshop: TG-POC & TG-Histo - Improving the quality of annotations in digitized whole slide images
Purpose:	Discussion

Contact:	Antti Suutala Institute for Molecular Medicine Finland (FIMM) University of Helsinki, Finland	E-mail: antti.suutala@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)"

FG-AI4H Workshop
May 30th 2022

Improving the quality of annotations in digitized whole slide images

Antti Suutala (M.Sc.)
Digital Diagnostics Group
Institute for Molecular Medicine Finland (FIMM)
University of Helsinki, Finland

Institute for Molecular Medicine Finland
Nordic EMBL Partnership for Molecular Medicine
HiLIFE UNIT

UNIVERSITY OF HELSINKI

Annotations in supervised learning

- Data annotation is an essential part of supervised learning in artificial intelligence

- Requires reliable visual interpretation of the digitized microscope images

- Quality and quantity of annotated data is reflected in the performance of the trained AI models (bad quality data – poor AI model performance)

- Cost of annotation: a lot of effort is needed to achieve required amount of reliable data and to maintain the quality of the data

The nature of whole slide datasets

- Very large areas to be examined
  - Digitized whole slide (50 mm x 25 mm) @ 40X Mag. => 20 gigapixels (200k x 100k)
- Very large number of findings for experts to be manually annotated in reasonable time
  - Inadequate annotations
- Individual cases maybe too difficult for an expert to make correct decisions
  - False labels
- Juggling with several different types of decisions at the same time is mentally tasking
  - Source of human errors

Challenges and limitations

- Experts are prone to make simple errors (fatigue, carelessness, subjectivity)
- Limitation of personal competence causing biased decisions
- Overconfidence / Underconfidence
- The requirement of correctly labelled data for training is too vast
- Evaluating the quality of the annotations objectively can be very difficult

Objectives to overcome the challenges

- To minimize cognitive bias and subjectivity
- To achieve balanced training datasets
- To produce more trustworthy data in less time
- To reduce the gap between the requirements that practitioners often have in mind when they build an AI model, and the requirements that are actually enforced by the AI pipeline

How to reach the objectives

- Utilize AI assisted micro tasking
- Exploit AI driven selection of training data
- Using multiple annotators with diverse skillsets
- Use of evaluation tools for quality assurance and quality management to improve annotations as well as to analyse and enhance annotators' performance

Micro tasking

- Dissecting annotation processes into easily manageable small tasks
  - data discovery, object tracing/alignment, image quality assessment, object labelling, etc.
- Microtasks turns unstructured data efficiently into structured data
- Serializing the tasks by the task type to avoid multi decision situations

Structuring micro tasking

- Utilizing different levels of expertise and diverse skillsets for different tasks for more accurate and reliable outcome

- Visualizing the digitized images in the best way for human perception, the turnaround time can be significantly improved

- Experts' valuable time should be focused on making considered decisions rather than learning and using complex tools that are inefficient

Multiple annotators

- Wisdom of the crowd*

- The collective opinion of a diverse independent group of individuals is likely to make certain type of decisions and predictions better than that of a single expert

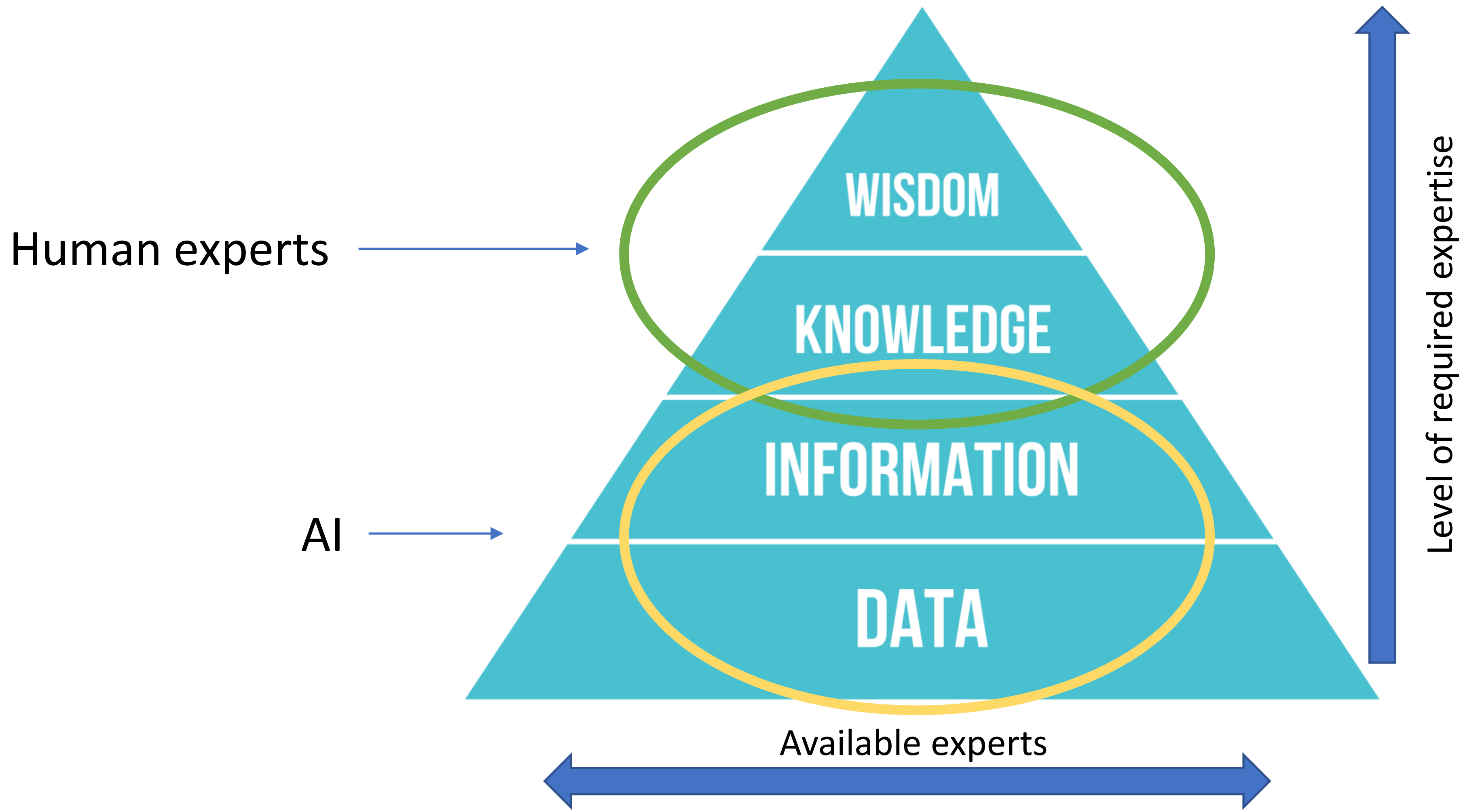
- Averaging results removes the noise associated with each individual judgment

*Surowiecki, James. *The Wisdom of Crowds*. Anchor Books, 2004.

Multiple annotators

- Highlights outliers and discrepancies
- Measure experts' decisions among the peer group
- Failure in cases when the majority is wrong

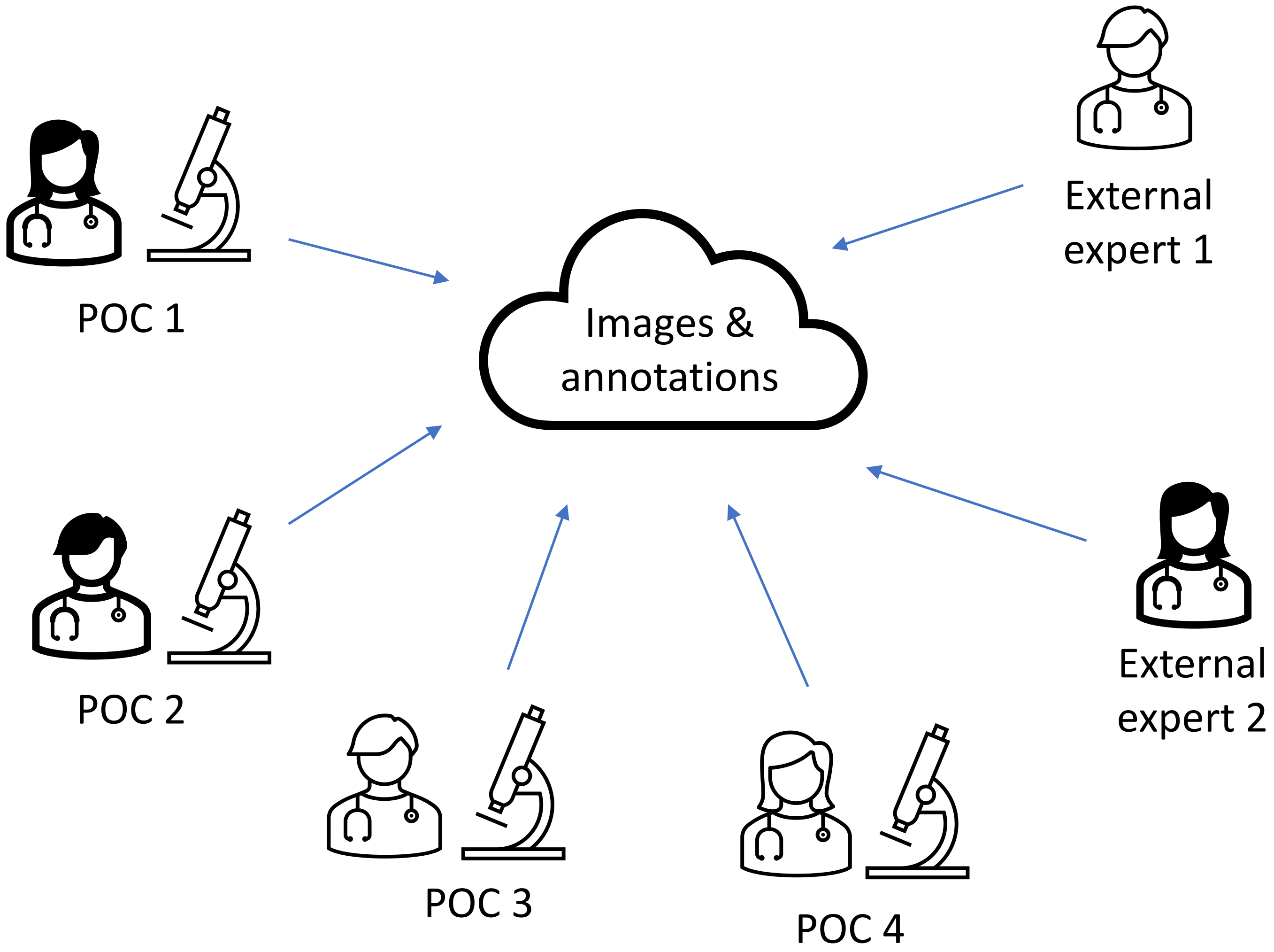
Effective targeting of expertise



Underlying figure source:
Wikipedia (DIKW pyramid)

Collaborative annotation process

- Ownership is shared over the whole dataset instead of local responsibility in each POC
- Utilizing external experts' participation with diverse skillsets (lack of certain type of skills on remote sites)
- Use of annotated digitized samples as an educational material to further increase the workforce competence



Collaborative annotation process

- Whole slide image data turned into AI assisted micro tasks across the target application domain

- Collectively but independently built ground truth

- Constant process of adjustment and readjustment for training more accurate AI models over time
  - ➢Intervene, review, correct, and verify the mislabelled data

Annotation platforms

- Several commercial annotation services and platforms available
  - V7, Labelbox, Scale AI, SuperAnnotate, Playment, Supervise.ly, Hive Data, DataLoop, etc

- Use cases: agriculture, autonomous driving, robotics, aerial imagery, NLP, healthcare, sports, financial services, insurance, security, life sciences, etc.

Key features

- AI assisted annotation and automation

- Ability to improve training data by labelling the same assets by different people independently

- Ability to manually review annotations side-by-side with the help of other experts

Key features

- Review annotation performance automatically using statistics, analytics and score metrics

- The collaboration and management of multiple distributed labelling workforces

- Multi-level quality management and quality assurance

Case Study: Autonomous Vehicle Startup

Use case: Produce training data for object perception models.

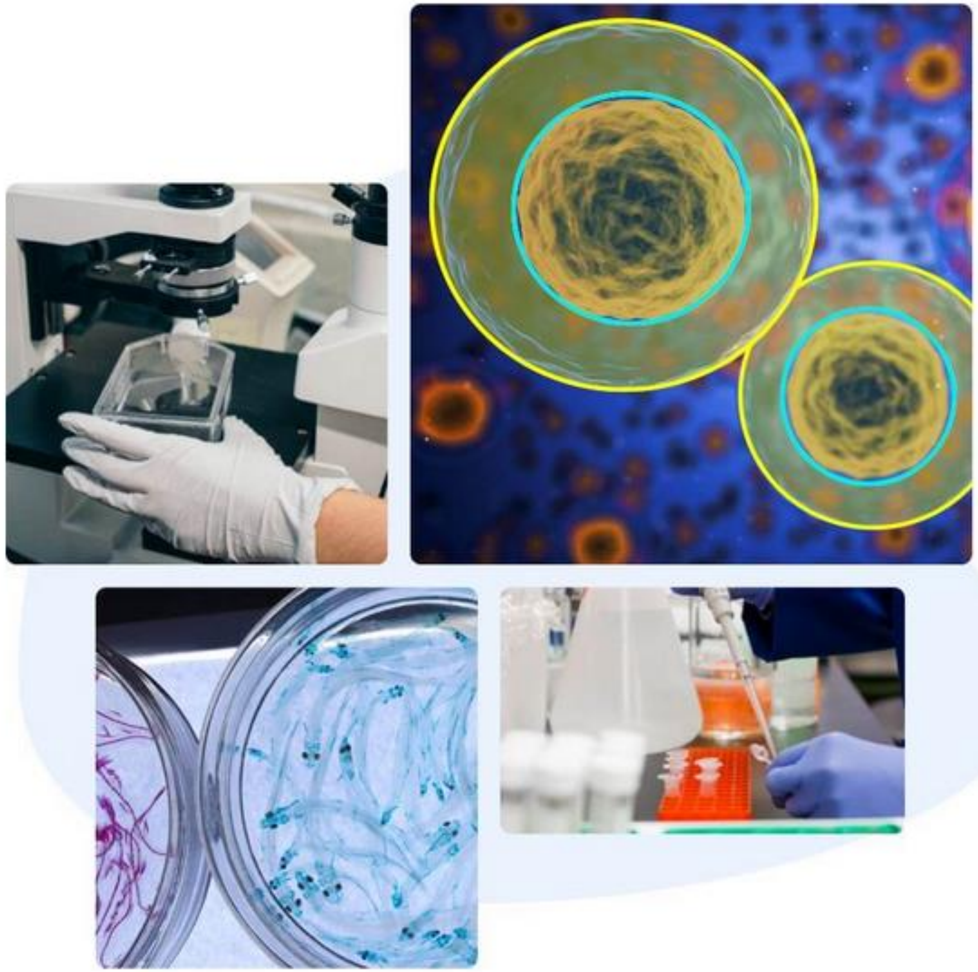
Source data: Car footage.

Data labelling type: Bounding boxes, cuboids, categorizations, semantic segmentations, polygons.

Volume: 1M bounding boxes, 600K cuboids, 1.6M+ categorizations.

Outcome: The company was able to optimize its autonomous vehicles' processing accuracy by significant margins. We surpassed expectations and the client is now working exclusively with us.

Source: Hive Data (thehiveai.com)



Case Study: Medical Research Center

Use case: Improve client AI model's ability to detect cells and organelles, reduce manual labor, and help come to research findings more efficiently.

Source data: Images of cells and organelles.

Data labelling type: Semantic segmentations.

Volume: 75K+ masks.

Outcome: The research center was able to use our data to test and improve its AI model, which enabled the organization to prioritize its research efforts.

Source: Hive Data (thehiveai.com)

Case Study: Energy Optimization Company

Use case: Profile residential home footprints to inform clients where heat is dispersed on their property and how they could optimize energy consumption. Classify home features including solar panels, pools, and trampolines.

Source data: Plane-sampled aerial imagery.

Data labelling type: Polygons.

Volume: 10K+ images, 400K+ geolocated polygons.

Outcome: We helped the client profile home energy use and guide targeted outreach to its residential customers.

Source: Hive Data (thehiveai.com)

Summary

- AI assisted microtasks can efficiently and reliably turn large unstructured data into structured data with minimal human labour

- Collectively, systematically, and consistently annotated datasets leads to more robust and objectively evaluated ground truth over time

- Collaborative annotation process reduces biases and allows decisions of individual experts with different skillsets to be compared, monitored, and evaluated

Thank you for your attention!

Digital Diagnostics group

Institute for Molecular Medicine Finland
Nordic EMBL Partnership for Molecular Medicine

Johan Lundin
Nina Linder
Dmitrii Bychkov
Hakan Kücükel
Mikael Lundin
Sebastian Stenman
Antti Suutala

Collaborators in POC diagnostics

Karolinska Institutet

KINONDO KWETU HEALTH SERVICES

UPPSALA UNIVERSITET

MUHAS