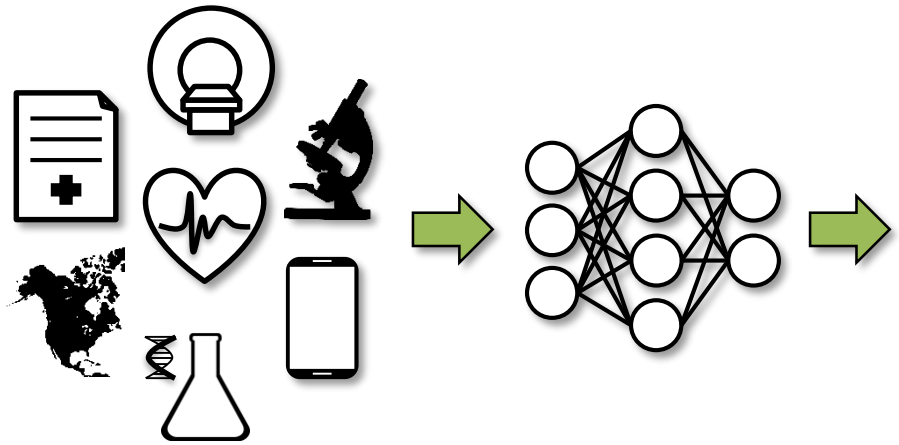
ITU KALEIDOSCOPE
ATLANTA 2019

Towards International Standards for the Evaluation of AI for Health

Markus Wenzel
Fraunhofer Heinrich Hertz Institute, Berlin, Germany
markus.wenzel@hhi.fraunhofer.de

4-6 December
Atlanta, Georgia, USA

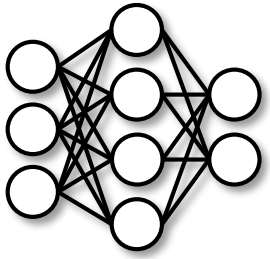
Artificial Intelligence/Machine Learning *for Health*



- Diagnostics, forecasting, triage, early detection, segmentation
- Classification: 0/1, ICD11
- Regression: 0.7
- ’19: Nature· FDA· NHS· NeurIPS

Safe, Effective? Accurate, Robust, Trustworthy?

- Clinical validation
  - Randomized controlled trials
  - Equator Network
    - TRIPOD-ML
    - CONSORT-AI
    - SPIRIT-AI



- Machine learning
  - Out-of-sample tests
  - Cross-validation
  - Competitions
  - **Benchmarking**
  - **AI standards**

AI Standards

- NIST: technical standards for AI · reliable, robust, trustworthy
- CTA: AI in healthcare
- CESI AI standardization
- CEN/CENELEC Focus Group AI
- DIN AI Roadmap
- ITU/WHO Focus Group AI for Health
- IEEE P2801/2
- ISO/IEC JTC1 SC42 AI
- Partnership on AI

ITU/WHO - Objectives

AI for Health
An ITU Focus Group
In collaboration with WHO

- Platform to facilitate global dialogue
- Standardized AI benchmarking
- Transparent AI evaluation
- Criteria, procedures, metrics, test data

Global Dialogue

AI for Health
An ITU Focus Group
In collaboration with WHO





7/18
Formal
Creation

11/2018
Columbia University
NYC

4/2019
World Expo
Shanghai

9/2019
UCSAF
Zanzibar

1/2020
PAHO/WHO
Brasilia

5/2018
AI for Good
Geneva

10/2018
WHO HQ
Geneva

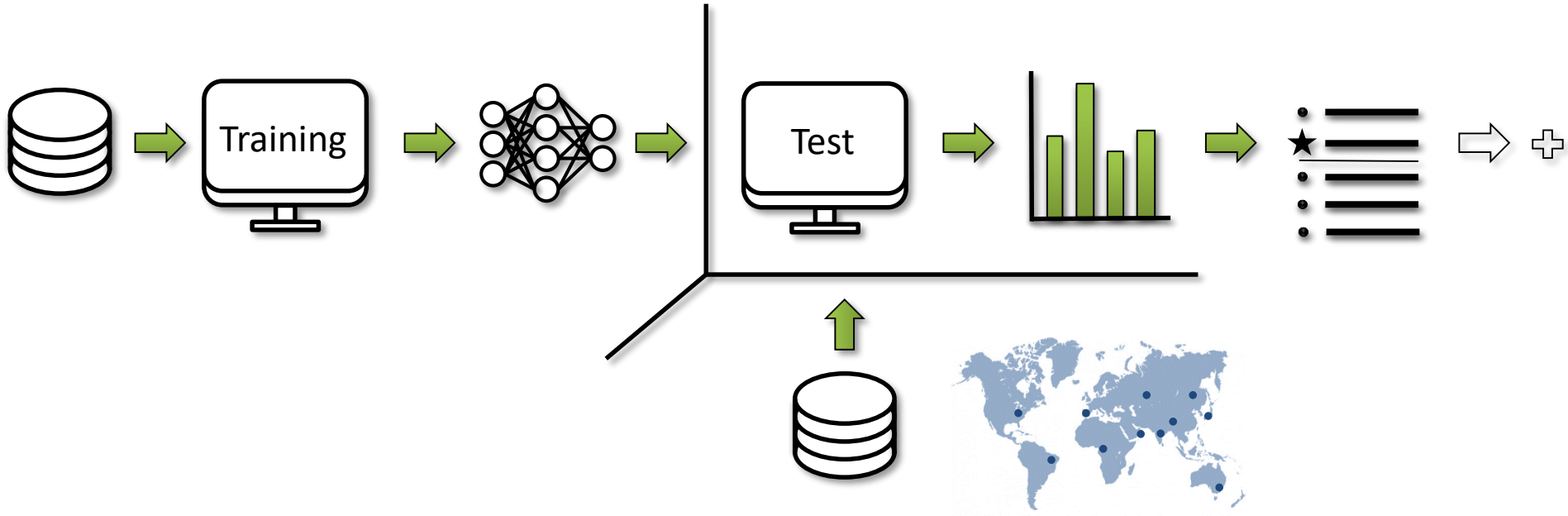
1/2019
EPFL
Lausanne

5/2019
AI for Good
Geneva

11/2019
ICMR & NICF
New Delhi

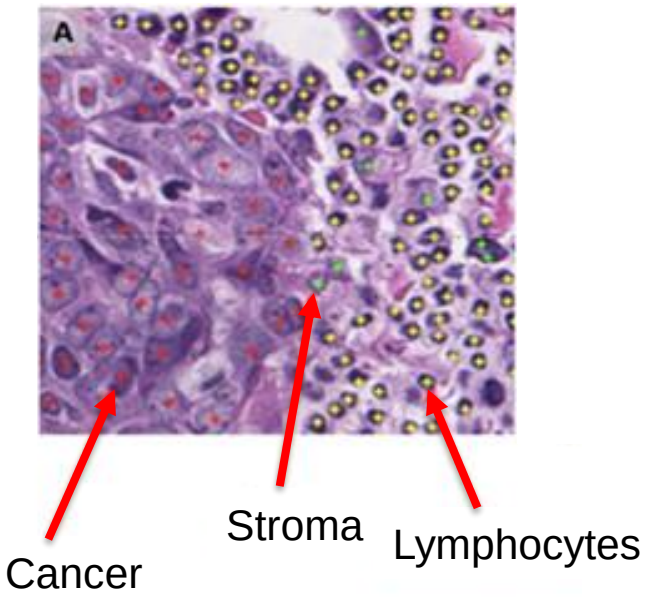
3/2020
AI in
Singapore

Benchmarking of AI Models



Proof-of-Concept Benchmark

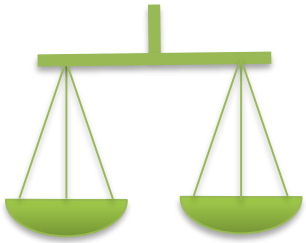
- Histopathology microscope images
- Detect lymphocytes
- AI model (Singapore Univ. Technology)
- Test data (Charité, Berlin): Unpublished 80 annotated images→ 232,000 patches
- Test (ITU/WHO FG AI4H)
- AI output vs. pathologists annotations
- Sensitivity: 0.91; Specificity: 0.88



Source: Prof. F. Klauschen (Charité)

Evaluation Criteria & Metrics

- **Performance:** sensitivity, specificity, accuracy, F1, AUC
- **Robustness:** interference, noise, manipulation
- **Explainability:** XAI, plausibility, Clever Hans cases
- **Generalizability:** other equipment, hospitals, regions
- **Confidence:** uncertainty quantification
- **Fairness:** FAT ML
- …



→ Agree on criteria, procedures, metrics, test data!

AI for Health
An ITU Focus Group
In collaboration with WHO

Approach I

1. CfP: AI for health topics
2. Community creation
3. Standard AI evaluation
   a) Define benchmarking tasks, criteria, metrics, test data
   b) Implement and run tests
   c) Report results
4. Next steps

Approach II

AI for Health
An ITU Focus Group
In collaboration with WHO

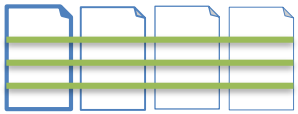
Topic Groups

- Ophthalmology
- Symptom assessment
- Histopathology
- Malaria detection
- Outbreak detection
- Dermatology
- Dental diagnostics
- Snakebite and snake identification
- Falsified medicine
- Neurological disorders
- Tuberculosis
- Volumetric chest CT
- Cardiovascular disease risk prediction
- Diagnosis of bact. infection & anti-microbial resist.
- Falls among the elderly
- Psychiatry
- Maternal and child health
- Radiotherapy
- …

Processes & Standards

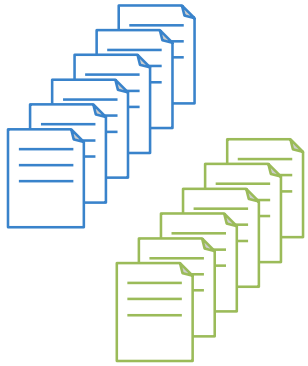
Working Groups

- Ethics
- Regulatory considerations
- Data & AI assessment
- Data & AI handling
- …



Work in progress: Deliverables – AI for Health

- Topic descriptions 18+
- Ethics considerations
- Regulatory considerations
- Software life cycle specification
- Training best practices specification
- Requirements specification
- Evaluation specification: process, AI test+metrics, clinical validation
- Data specification: requirements, acquisition, annotation, training/test, handling, sharing



Conclusion & Outlook

- Standards for reliable, robust, trustworthy AI
- Health AI benchmarking
- ITU/WHO
- Wenzel, Wiegand (2019). Towards int. standards for the evaluation of AI for health

- Anticipate cooperation - standard setting organizations
- Shared problems and data modalities *(robust image classification)*
- NIST · CTA · Partnership on AI
- Questions - Feedback?

ITU KALEIDOSCOPE ATLANTA 2019

Thank you

markus.wenzel@hhi.fraunhofer.de

itu.int/go/fgai4h