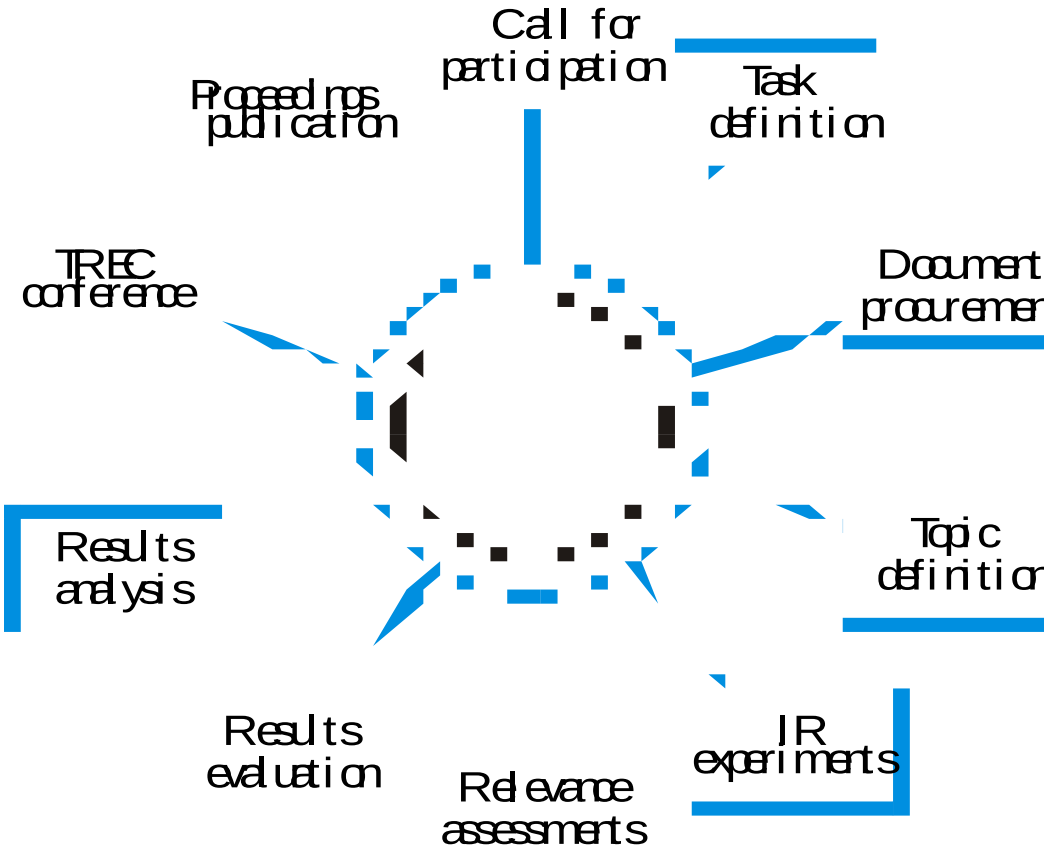
Scientific challenges: history, platforms, and perspectives



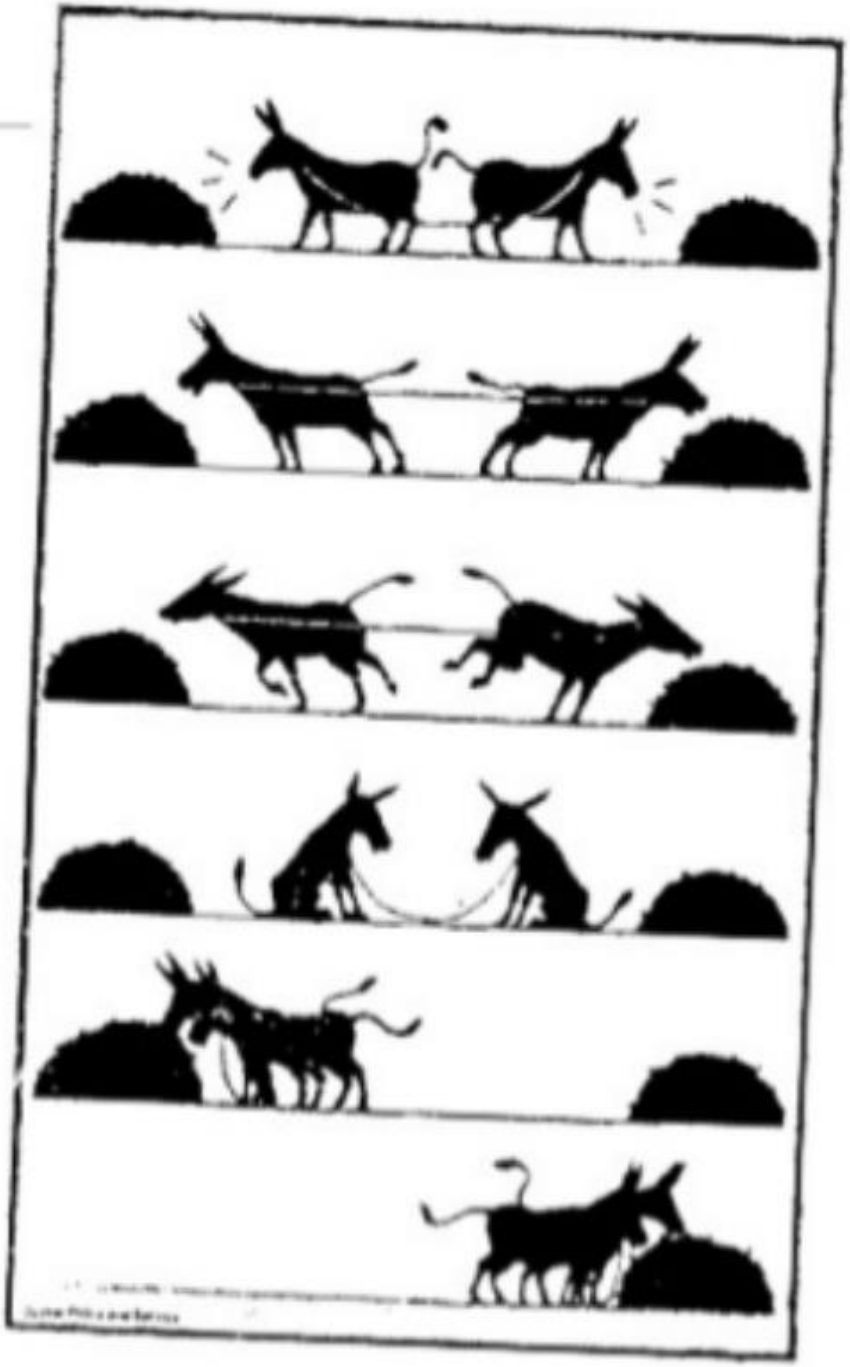
Henning Müller
HES-SO & UNIGE
9.5.2023

Henning Müller

- Medical informatics studies in Heidelberg, Germany (1992-1997)
  - Exchange with Daimler Benz research, USA
- PhD in image processing, image retrieval, Geneva, Switzerland (1998-2002)
  - Exchange with Monash University, Melbourne, AUS
- Professor in radiology and medical informatics at the University of Geneva (2014-)
- Professor in Computer Science at the HES-SO, Sierre, Switzerland (2007-)
- Visiting faculty at Martinos Center (2015-2016)
- Member of the Swiss National Research Council

Scientific environment

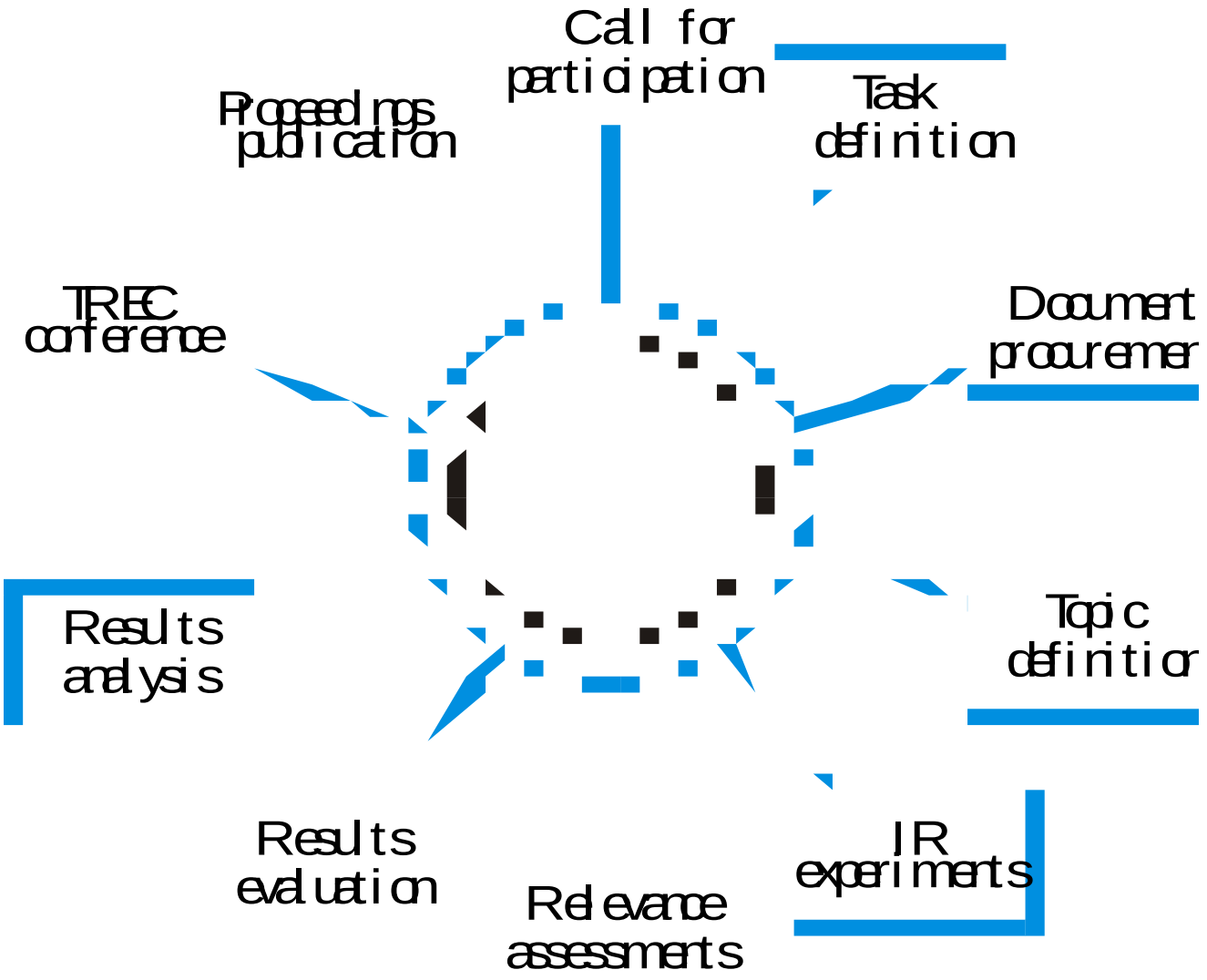
- Competition
- Coopetition
- Cooperation



History of scientific challenges

- Cranfield tests (1960-1967)
  - Cyril Cleverdon, found that automatic indexing is as good or better than manually attached keywords
- SMART retrieval system experiments (1971)
  - Gerald Salton, comparisons of several keyword weightings, such as tf/idf
- TREC – Text REtrieval Conference (1992)
  - Donna Harman, Ellen Vorhees, systematic comparison of information retrieval systems
    - Funded by NIST, …
    - Large-scale, video, multi-lingual, …
    - Yearly circle of events

A yearly circle



Visual retrieval evaluation

- Little systematic evaluation in first years of research (1990-2000)
  - Some papers on methodologies
  - Benchathlon to foster discussions
- Since then, evaluation has come a long way!!
- TRECVID, ImageCLEF, INEX MM, ImageEval, …
  - Improvement in performance can be shown
  - Techniques can be compared
  - Strong baselines, massive impact
- Methodologies and user models can be criticized
  - Not all research can be benchmarked
  - Innovation instead of pure performance

CLEF

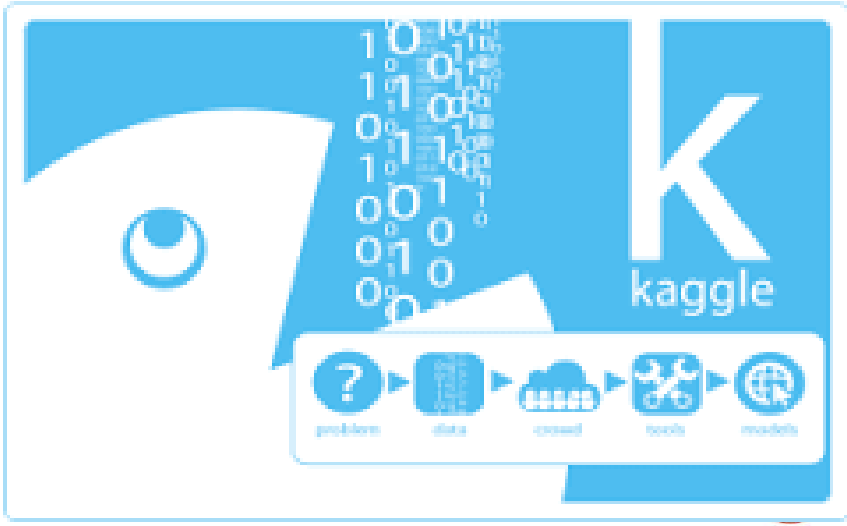
- Cross Language Evaluation Forum
  - Started as track in TREC (Text Retrieval Conference,1997)
- Independent workshop since 2000
- Multilingual information retrieval
  - Collections are multilingual
  - Queries are in a language different from the collection
- Good framework, registration, legal issues, proceedings in Springer LNCS, …

Early history of ImageCLEF

- 2003: first image retrieval task, 4 participants
- 2004: 17 participants for three tasks (~200 runs)
  - Medical task for visual image retrieval added
- 2005: 24 participants for fours tasks (~300 runs)
  - Two medical tasks
- 2006: 30 participants for four tasks (~300 runs)
  - LTU database of objects for object classification
- 2007: 35 participants (>1000 runs)
  - Hierarchical classification
- 2008: 45 participants submitted results (>2000 runs)
  - 63 registrations, wiki task

Changes for scientific challenges

- Proposed at most conferences and workshops in machine learning and medical imaging
  - RSNA is following on this as well, ICPR, MICCAI, …
- Same data, same evaluation methodology, workshop to discuss results among persons
  - Really make results comparable, discuss future ideas
- Commercial platforms
  - Kaggle, TopCoder
  - Codalab, CrowdAI,
  - …



Too much of a good thing?

- Many challenges now attract only few participants besides a few really big ones
  - A minimum is required for meaningful evaluation
- Each conference has its own challenges, plus benchmarks and professional platforms …
- Much effort is invested in challenge organization!
- Platforms such as crowdAI could make things easier, as they make benchmarks and results visible and allow post-conference participation
- Benchmarks in related domains should join forces, share data etc.

Bases of retrieval benchmarks

- Tasks/topics (with motivation and user model)
  - Experts for the relevance judgements
- Data set (large, diverse, realistic)
- Participants with their techniques
- Ground truth and a gold standard
- Performance measures
  - Valid, so measure precisely what is supposed to be measured

Difficulties of benchmarks

- Lack of funding (as this is not research in itself)
- Access to data
- **Motivate participants** & create a community
- Partners from industry for relevance
- Realistic tasks and user models
- Ground truth and annotations
- Professional organization
- Prove advances and benefits
  - Impact analysis exists: TREC, TRECVid, CLEF, ImageCLEF

The VISCERAL project

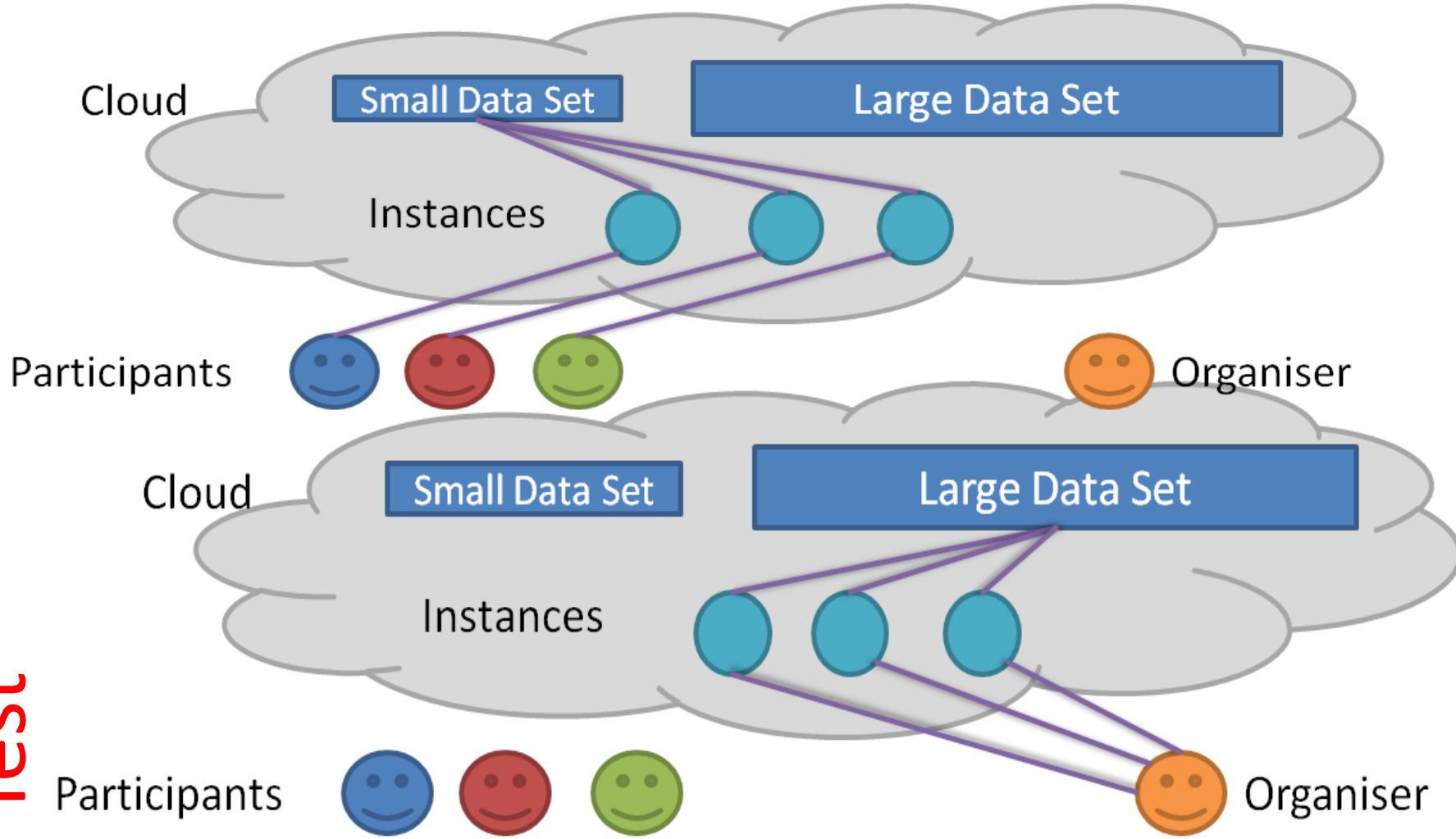
- Visual Concept Extraction challenge in Radiology
- Partners:
  - Technical University of Vienna, Aust
  - Medical University of Vienna, Austria
  - HES-SO, Sierre, Switzerland
  - ETHZ, Zürich, Switzerland
  - University of Heidelberg, Germany
  - Catalonia Health Authority, Barcelona,
  - 1.11.2012-30.4.2015 (30 months)
- Run challenges on medical organ segmentation, similar case retrieval and lesion detection

Challenges with challenges

- Difficult to distribute very big datasets
  - Sending around hard disks? risky, expensive
- Sharing confidential data
  - Big data is impossible to anonymize automatically
- Quickly changing data sets
  - Outdated when a test collection is being created
- Optimizations on the test data are possible
  - Manual adaptations, etc.
  - Often hard to fully reproduce results
- Groups without large computing infrastructures are disadvantaged

Visceral

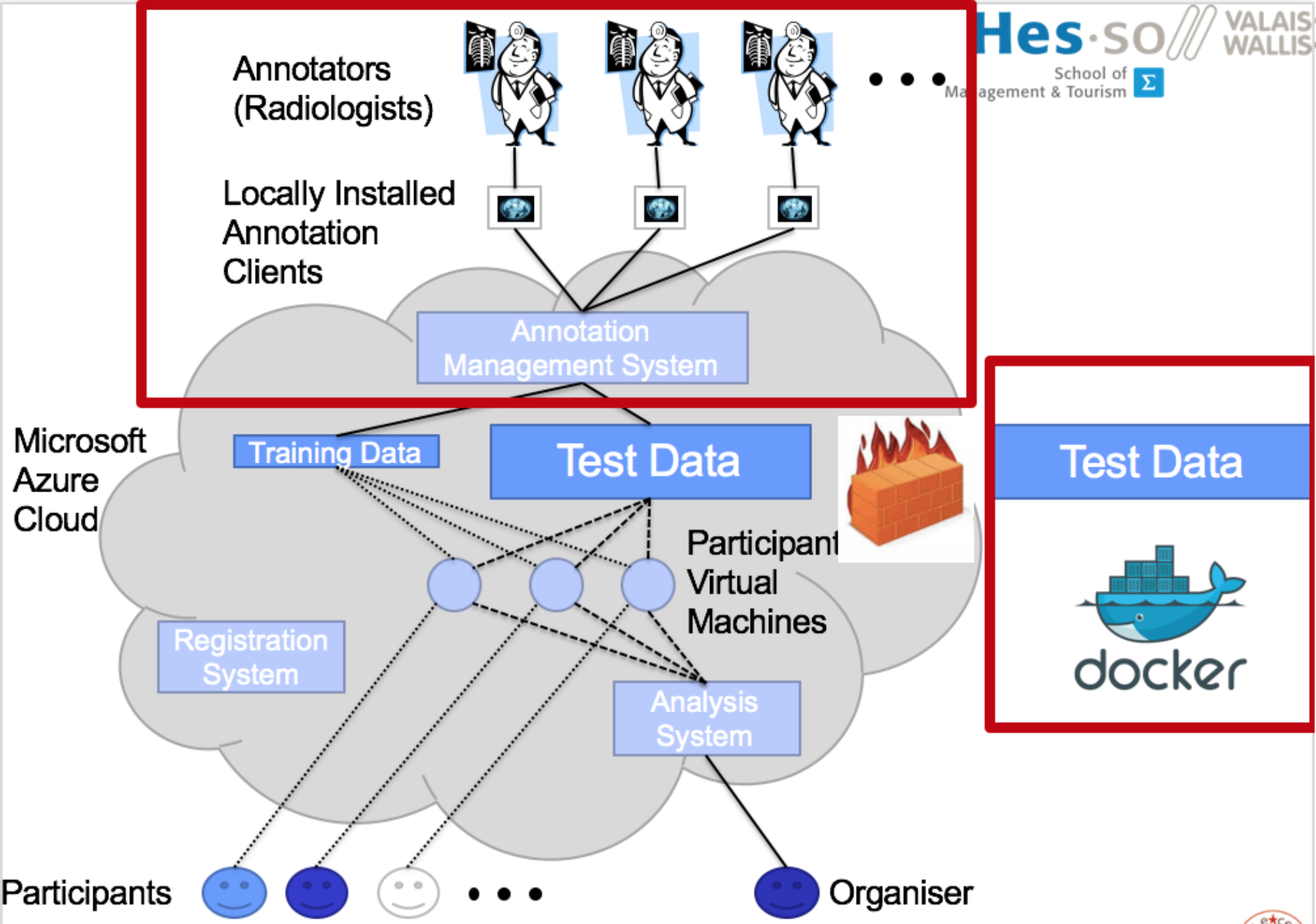
Hes·so // VALAIS WALLIS
School of Management & Tourism



A. Hanbury, H. Müller, G. Langs, M. A. Weber, B. H. Menze, T. Salas Fernandez, Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis, CLEF conference, Springer Lecture Notes in Computer Science, 2012.

Workflow

- Ethics approval, obtaining, anonymizing data
- Annotations by health professionals in Debrecen, Hungary
  - Semi-automatic tools (Geos) for larger structures, 3Dslicer for small structures
- Quality control (personal profiles for annotators) and adaptations of annotation guidelines
  - Radiology partner checked all initial annotations
- Extremely detailed annotation guidelines
  - Positive and negative examples
  - Detailed descriptions limit variability
- Double annotations of the same structures to measure subjectivity of the segmentation task


Annotators
(Radiologists)
Locally Installed
Annotation
Clients
Annotation
Management System
Microsoft
Azure
Cloud
Training Data
Test Data
Participant
Virtual
Machines
Registration
System
Analysis
System
Participants
Organiser
Test Data
docker

Evaluation-as-a-Service (EaaS)

- Moving the algorithms to the data, not vice versa
  - Required when data are: very large, changing quickly, confidential (medical, commercial, …)
- Different approaches
  - Source code submission, APIs, VMs local or in the cloud, Docker containers, specific frameworks
- Allows for continuous evaluation, component-based evaluation, total reproducibility, updates, …
  - Workshop March 2015 in Sierre on EaaS
  - Workshop November 2015 in Boston on cloud-based evaluation (http://www.martinos.org/cloudWorkshop/)

Allan Hanbury, Henning Müller, Krisztian Balog, Torben Brodt, Gordon V. Cormack, Ivan Eggel, Tim Gollub, Frank Hopfgartner, Jayashree Kalpathy-Cramer, Noriko Kando, Anastasia Krithara, Jimmy Lin, Simon Mercer, Martin Potthast, Evaluation-as-a-Service: Overview and Outlook, ArXiv, 2015.

EaaS aspects

Sharing images, research data

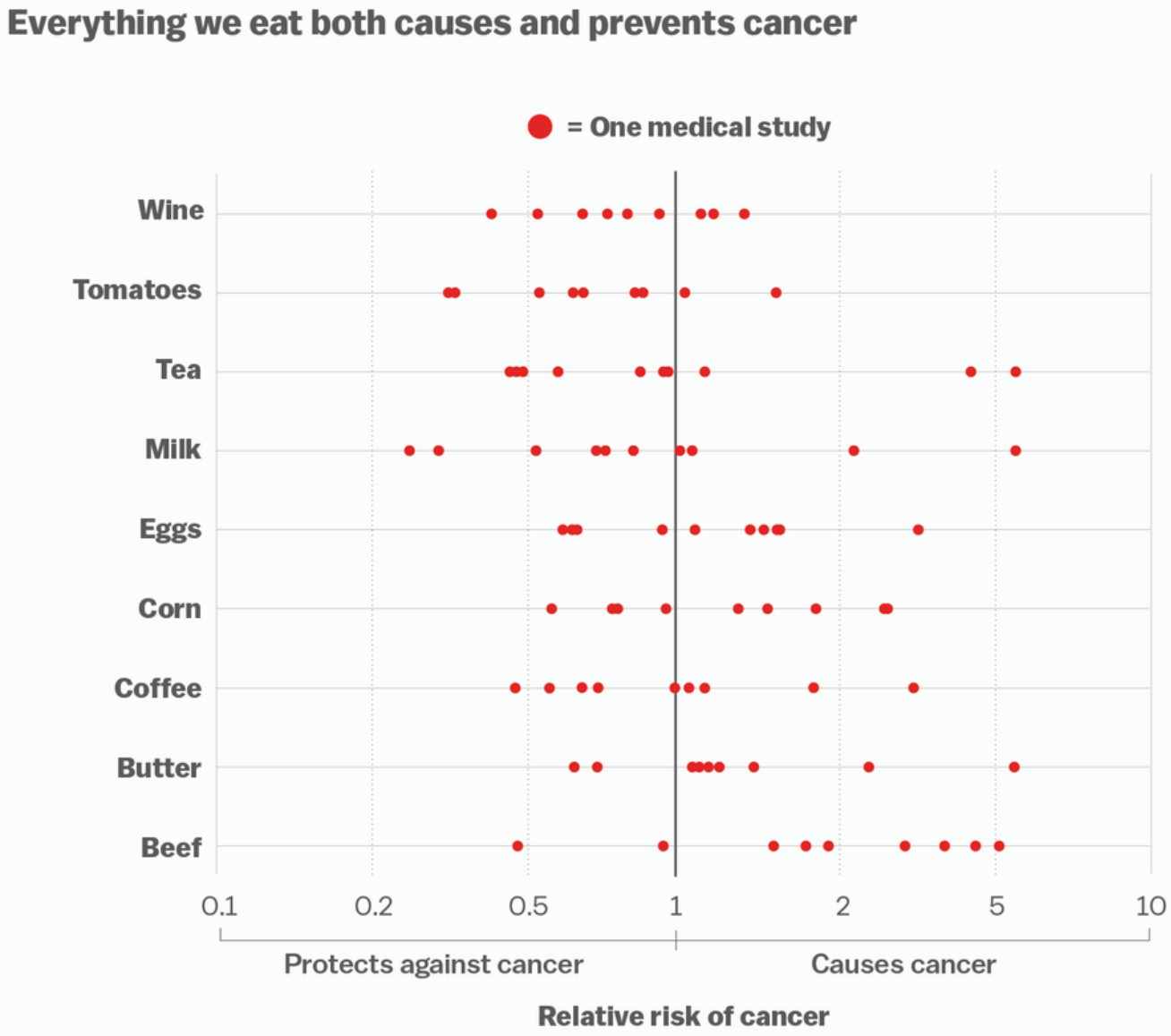
- Very important aspect of research is to have solid methods, data, large if possible
  - If data not available, results can not be reproduced
  - If data are small, results may be meaningless
- Many multi-center projects spend most money on data acquisition, often delayed no time for analysis
- Research is international!
- NIH is great to push data availability
  - But data can be made available in an unusable way


Why Most Published Research Findings Are False

John P. A. Ioannidis

factors that influence this problem and is characteristic of the field and can

Contradictory science



SOURCE: Schoenfeld and Ioannidis, *American Journal of Clinical Nutrition*

Perspectives and conclusions

- Scientific challenges can be an important part of the scientific work
  - Get strong baselines, improve reproducibility
  - Impact when organizing this
- Attention needs to be paid to get a critical mass and motivate participants
  - Have unique data sets and scientifically challenging tasks
  - Update the tasks and data regularly
- Platforms to automate the process are important
  - Particularly for Docker-based submissions

Contact

- More information can be found at
  - http://medgift.hevs.ch/
  - http://publications.hevs.ch/
- Contact:
  - Henning.mueller@hevs.ch