netherlands eScience center

Benchmarking Algorithms for Science

Adriënne Mendrik

Price, Keith. "Anything you can do, I can do better (No you can't)…." *Computer Vision, Graphics, and Image Processing* 36.2-3 (1986): 387-391.

What are the problems?

- How to evaluate the work of others if you don't have their programs?
- How do you use the algorithms without being able to get in touch with the original creator?
- What does it mean when a re-implementation does not work? Who failed, the algorithm or the implementation?
- How do you compare results? Mine works 80% of the time by some measure, yours 80% by another measure, and they seem to agree 40% of the time?
- How do you control parameter tuning of the algorithms?
- How do you present the central idea? Key ideas in papers are sometimes lost in details…

Price, Keith. "Anything you can do, I can do better (No you can't)…."
Computer Vision, Graphics, and Image Processing 36.2-3 (1986): 387-391.

- **How to evaluate the work of others if you don't have their programs?**
- How do you use the algorithms without being able to get in touch with the original creator?
- What does it mean when a re-implementation does not work? Who failed, the algorithm or the implementation?
- **How do you compare results? Mine works 80% of the time by some measure, yours 80% by another measure, and they seem to agree 40% of the time?**
- How do you control parameter tuning of the algorithms?
- How do you present the central idea? Key ideas in papers are sometimes lost in details…

Price, Keith. "Anything you can do, I can do better (No you can't)…."
Computer Vision, Graphics, and Image Processing 36.2-3 (1986): 387-391.

Benchmark =

- A standard or point of reference against which things may be compared.
- A test designed to evaluate or compare the performance of computer hardware or software.

https://www.lexico.com/en/definition/benchmark

“Voor organisaties binnen dezelfde branche met vergelijkbare activiteiten (bijvoorbeeld musea of toneelgezelschappen) is het interessant om de kwaliteit van deze activiteiten met elkaar te vergelijken door middel van een benchmark. Hierdoor kun je van elkaar leren: Wat lukt de één wel en de ander niet?“

https://www.claudiadegraauw.nl/een-benchmark-wat-het-en-waarom-het-zinvol/

Benchmark =

Open online benchmark that uses data, truth and metrics to evaluate the performance of automatic algorithms, submitted by participants, with respect to a research problem.

Challenge

Competition

Shared Task

The mission: to field a team of robots capable of winning against the human soccer World Cup champions by 2050.

Each year robots improve dramatically using the tricks of the best robots from the year before.

https://www.robocup.org

Challenges in Medical Image Analysis

https://grand-challenge.org/challenges/

Grand Challenge All Challenges Sign in / Register

All Challenges

Here is an overview of all challenges that have been organized within the area of medical image analysis that we are aware of. If you know any study that would fit in this overview, or want to advertise your challenge, please send an email to support@grand-challenge.org and we will add the challenge to the list on this page.

Active filters: 0

Filters

Host

Modality

Task type

Structure

Displaying 192 of 192

2020

EndoCV2020

Endoscopy computer vision challenge (EndoCV2020) introduces two core sub-themes in endoscopy: 1) artefact detection and segmentation (EAD2020) and 2) disease detection and segmentation (EDD2020).

Participants: 112
Workshop: April 3, 2020
Associated with: IEEE ISBI 2020
Hosted on: grand-challenge.org

HEROHE

Unlike previous challenges, this proposes to find an image analysis algorithm to identify HER2-positive from HER2-negative breast cancer specimens evaluating only the morphological features present on the HE slide, without the staining patterns of IHC.

Participants: 517
Workshop: May 13, 2020
Associated with: European Congress on Digital Pathology (ECDP 2020)
Hosted on: grand-challenge.org

2019

AASCE

Accurate Automated Spinal Curvature Estimation

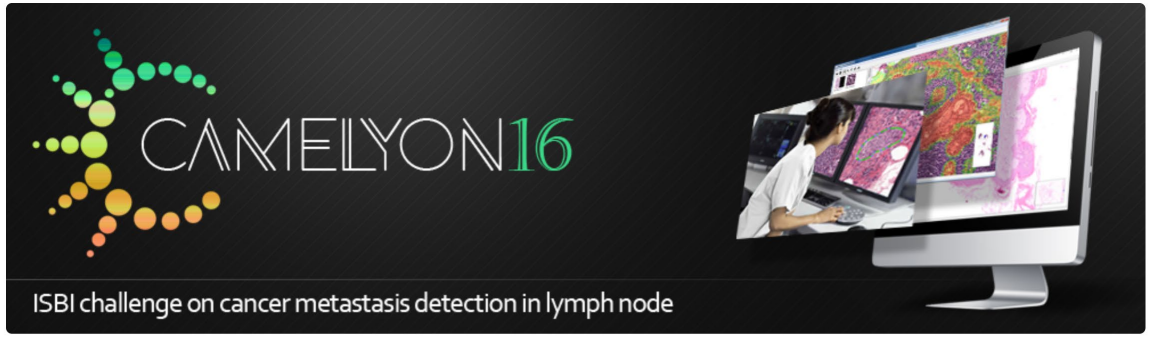
Lymphocyte Assessment Hackathon

Lymphocyte Assessment Hackathon in conjunction with the MICCAI COMPAY 2019 Workshop on Computational Pathology

ODIR-2019

北京大学国际眼底图像智能识别竞赛
Peking University International Competition on Ocular Disease Intelligent Recognition

Challenges on cancer metastasis detection in lymph node sections (Radboud UMC, Nijmegen)



Home | Background | Rules | Register | Data | Evaluation | Submit | Results | Organizers | Download | Forum | Program

The CAMELYON16 challenge has ended in November 2016

PLEASE CHECK OUT CAMELYON17:

https://camelyon17.grand-challenge.org

Overview

The goal of this challenge is to evaluate new and existing algorithms for automated detection of metastases in hematoxylin and eosin (H&E) stained whole-slide images of lymph node sections. This task has a high clinical relevance but requires large amounts of reading time from pathologists. Therefore, a successful solution would hold great promise to reduce the workload of the pathologists while at the same time reduce the subjectivity in diagnosis. This will be the first challenge using whole-slide images in histopathology. The challenge will run

Home | Background | Rules | Data | Evaluation | Forum | Organisers | Workshop | **Join** | Results

Overview

The CAMELYON17 challenge is still open for submissions!

Built on the success of its predecessor, CAMELYON17 is the second grand challenge in pathology organised by the Diagnostic Image Analysis Group (DIAG) and Department of Pathology of the Radboud University Medical Center (Radboudumc) in Nijmegen, The Netherlands.

The goal of this challenge is to evaluate new and existing algorithms for automated detection and classification of breast cancer metastases in whole-slide images of histological lymph node sections. This task has high clinical relevance and would normally require extensive microscopic assessment by pathologists. The presence of metastases in lymph nodes has therapeutic implications for breast cancer patients. Therefore, an automated solution would hold great promise to reduce the workload of pathologists while at the same time reduce the subjectivity in diagnosis.

Last year at ISBI, we organised the highly successful CAMELYON16 grand challenge, in which 32 submissions from as many as 23 research groups were received. This was the first

JAMA paper about the Camelyon Challenge

Original Investigation

December 12, 2017

Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer

Babak Ehteshami Bejnordi, MS[1]; Mitko Veta, PhD[2]; Paul Johannes van Diest, MD, PhD[3]; Bram van Ginneken, PhD[1]; Nico Karssemeijer, PhD[1]; Geert Litjens, PhD[4]; Jeroen A. W. M. van der Laak, PhD[4]; and the CAMELYON16 Consortium

Author Affiliations

JAMA. 2017;318(22):2199-2210. doi:10.1001/jama.2017.14585

Key Points

Question What is the discriminative accuracy of deep learning algorithms compared with the diagnoses of pathologists in detecting lymph node metastases in tissue sections of women with breast cancer?

Camelyon Challenge in the news

Voor het eerst is aangetoond dat een zelflerend computersysteem beter in staat is om uitgezaaide borstkanker te ontdekken dan een patholoog die onder normale tijdsdruk werkt. Dat blijkt uit een onderzoek van het Radboud UMC.

In de studie werden plakjes lymfeklier van patiënten ingevoerd in verschillende computersystemen om die te controleren op uitzaaiingen. Normaal gesproken onderzoekt een patholoog deze onder een microscoop.

GESCHREVEN DOOR

Machteld Veen

Redacteur Nieuwsuur

Van Ginneken, Bram, et al. "Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study." *Medical image analysis* 14.6 (2010): 707-722.

Results show a substantial performance difference between algorithms, and demonstrate that combining the output of algorithms leads to marked performance improvements.

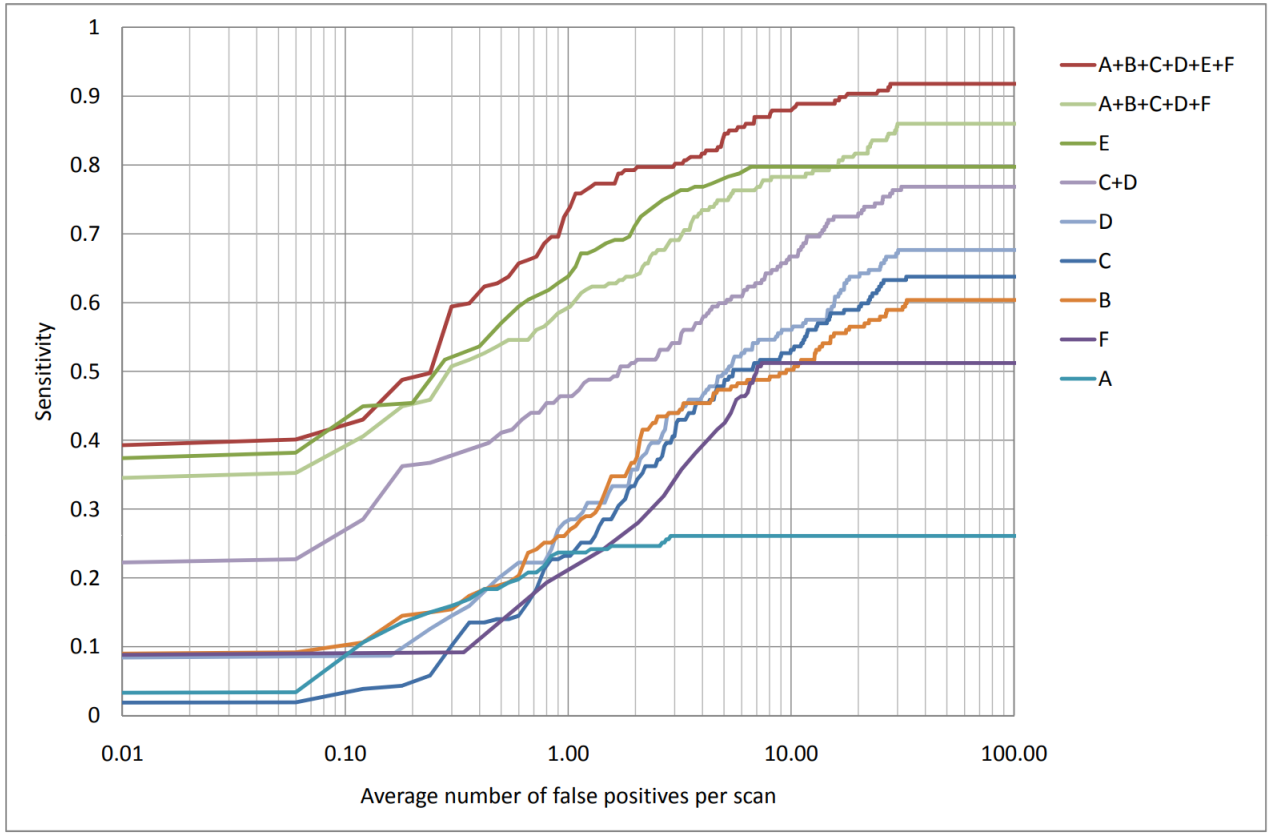
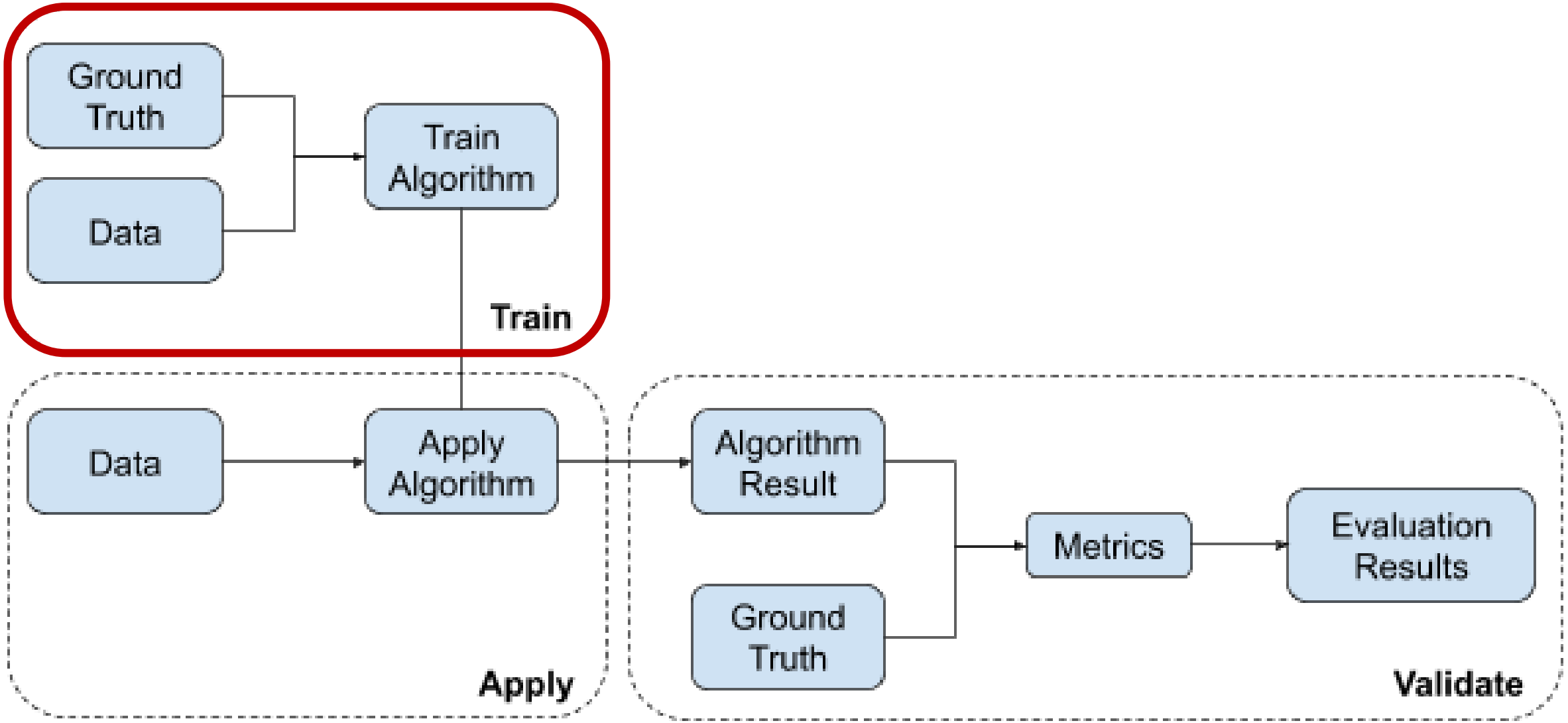
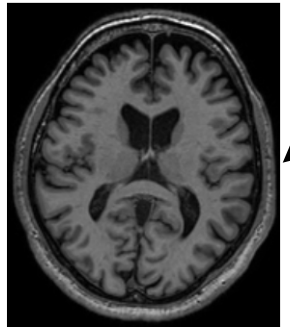


Figure 2: FROC curves of all six systems and three combinations. The horizontal axis is logarithmic and covers four orders of magnitude.

How does it work?



Example Brain Tissue Segmentation: Training



MRI Scan

Ground Truth

Data

Train Algorithm

Train

Neural Network

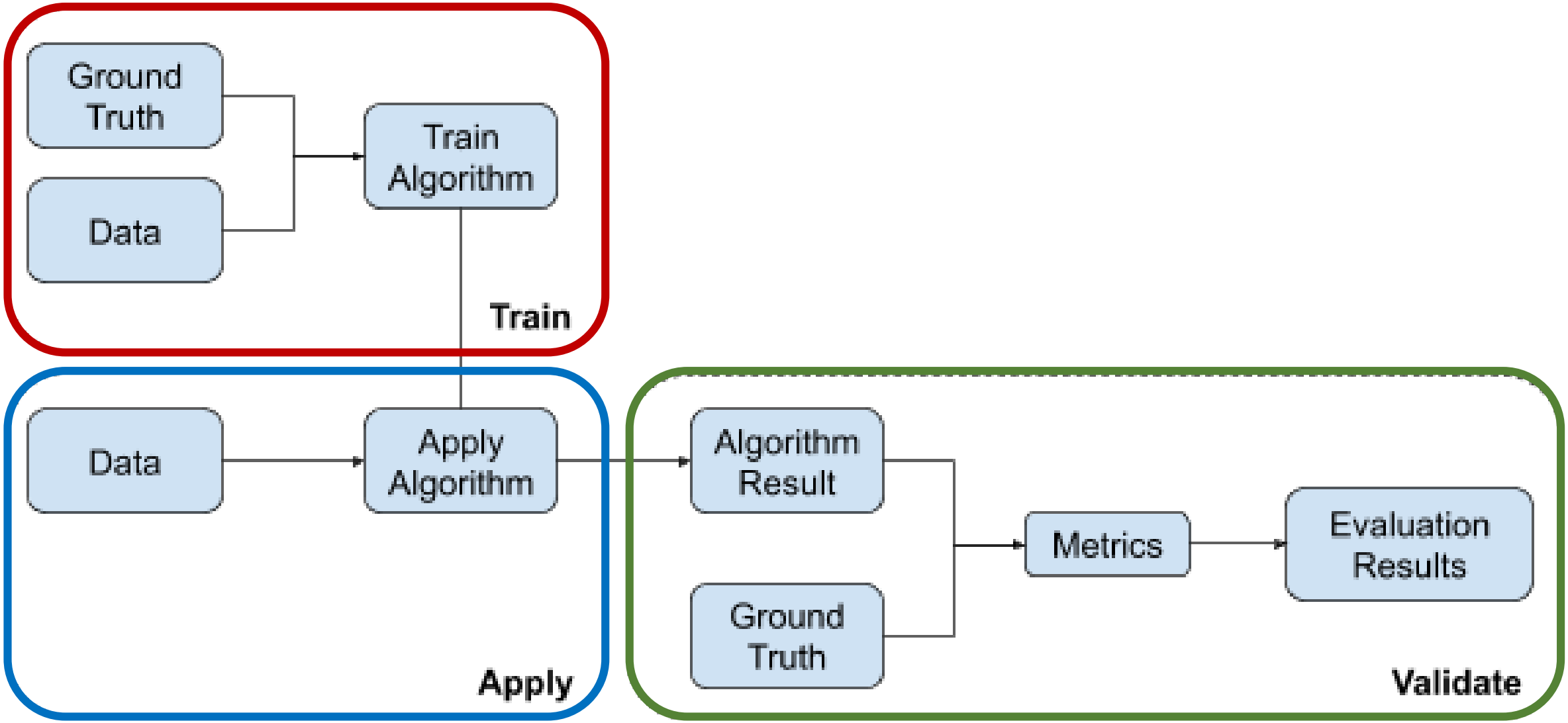
Atlas-based Method

Random Forest

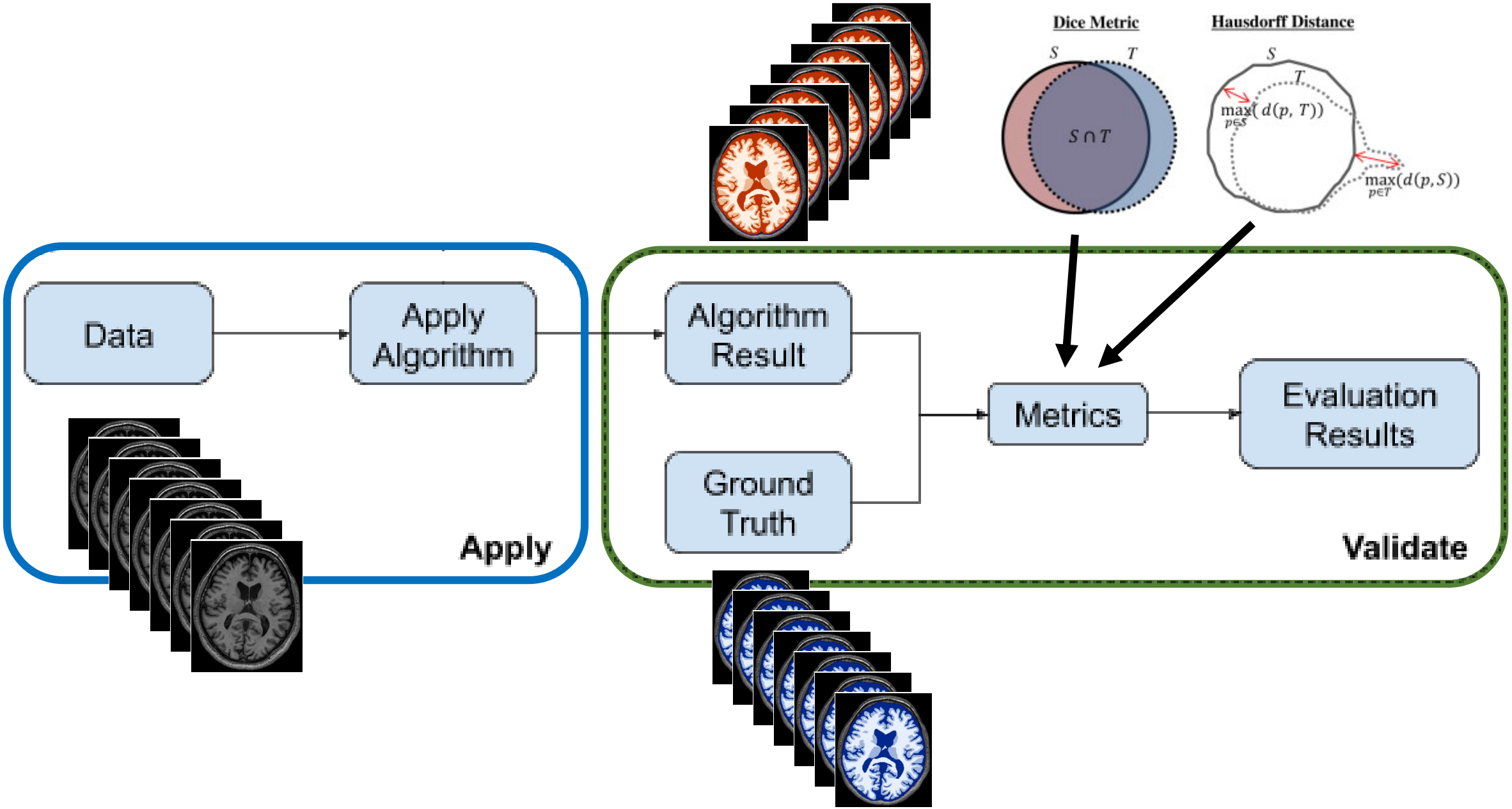
Support Vector Machine

…

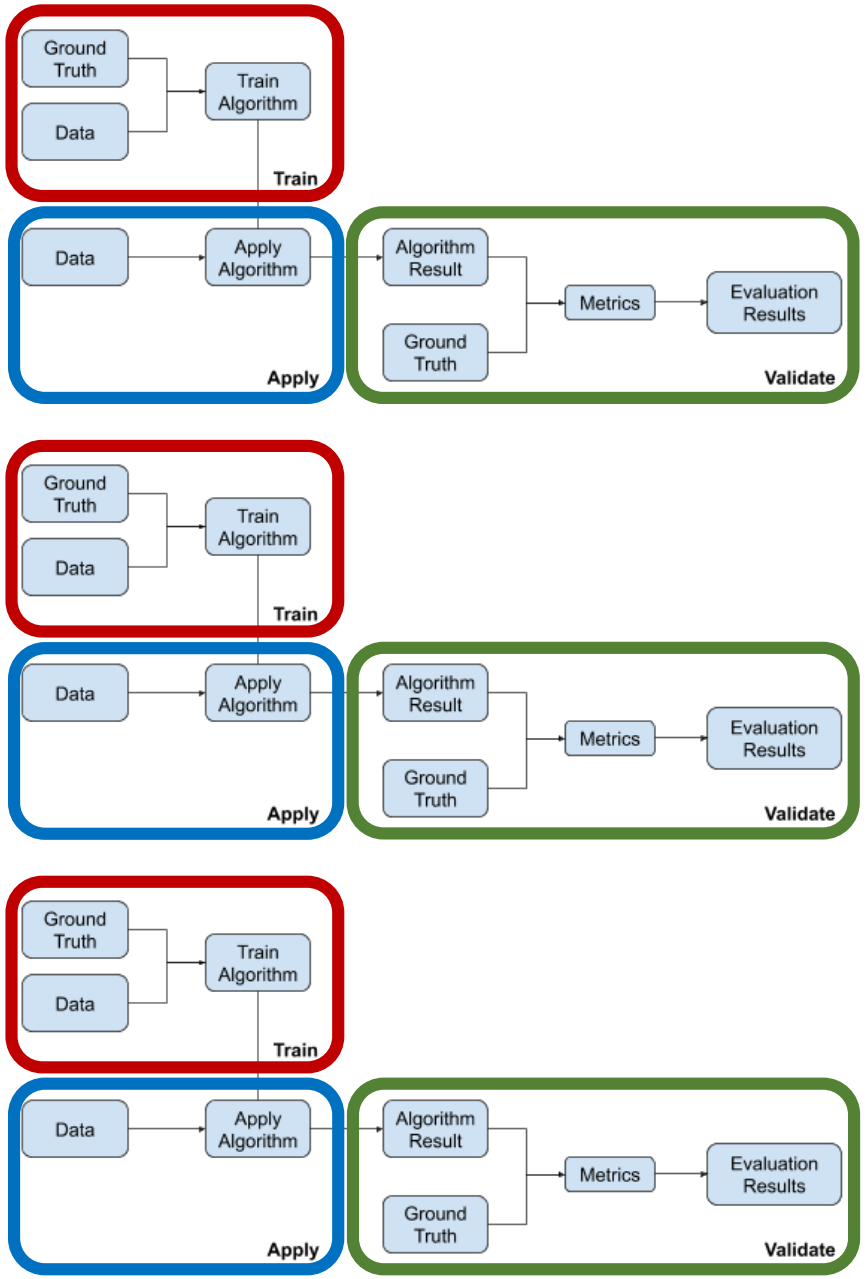
How does it work?

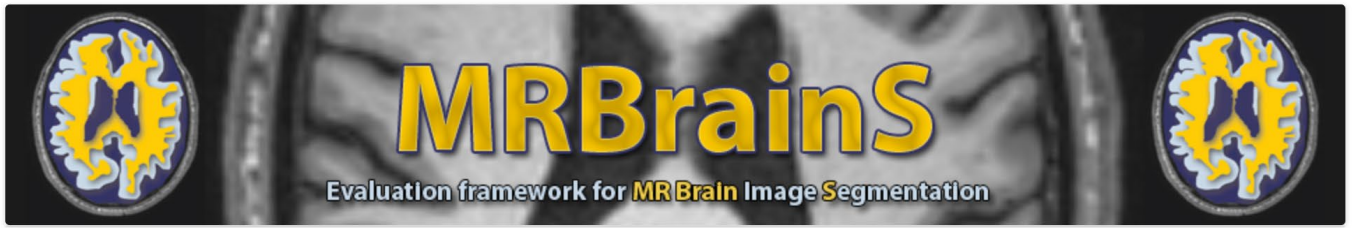


Example Brain Tissue Segmentation: Validation



Benchmarking Algorithm Performance





HOME THE CHALLENGE WORKSHOP DATA SUBMIT **RESULTS** ORGANIZERS — LOGIN REGISTER

Results

When teams submit their segmentation results, the evaluation results will be sent to the team contact person by e-mail and will be listed below.

Search:

RANK	TEAM NAME	SUBMISSION NAME	SUBMISSION DATE	SCORE	SEQUENCES USED	DURATION
1	XMU_SmartDSP2	3D Spatial Weighted U-Net for Multi-modality Brain MRI Segmentation	29-08-18	39	T1; T1_IR; FLAIR	10 min
2	Smartdsp3713	3D weighted U-shape fully convolutional network	26-07-19	53	T1; T1_IR; FLAIR	2 min
3	TailHot	Multi-modality aggregation network[3]	13-04-18	68	T1; T1_IR; FLAIR	13 sec
4	WTA2	3D Cascade convolutional architecture - Method 2[2]	23-05-18	84	T1; T1_IR; FLAIR	2 min
5	XMU SmartDSP	3D CNN with a Cross-modality Channel Attention Scheme[3]	17-08-18	90	T1; T1_IR; FLAIR	10 min
6	XLab	3D Fully CNN with Multi-Modality Feature Fusion[3]	10-08-18	101	T1; T1_IR; FLAIR	2 min

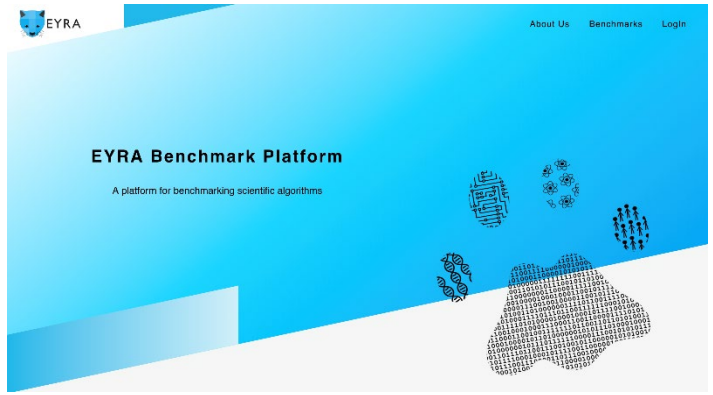
Enhance Your Research Alliance (EYRA) Benchmark Platform

Alliance between SURF and the Netherlands eScience Center (2018/2019)

Researchers

- Research Question
- Data
- Ground Truth
- Evaluation metrics

Benchmarks



Algorithms

Algorithm Developers

Adriënne Mendrik

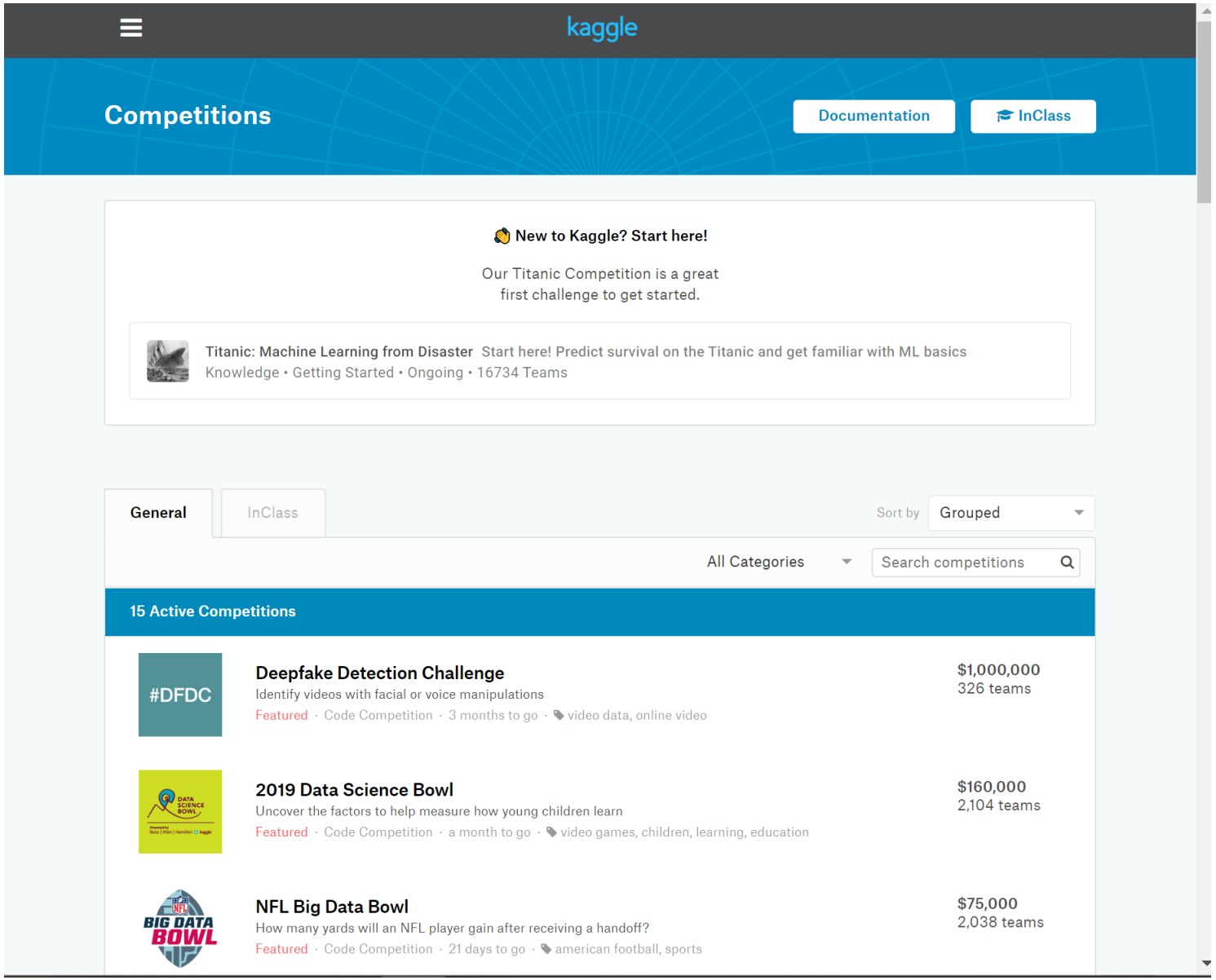
netherlands eScience center

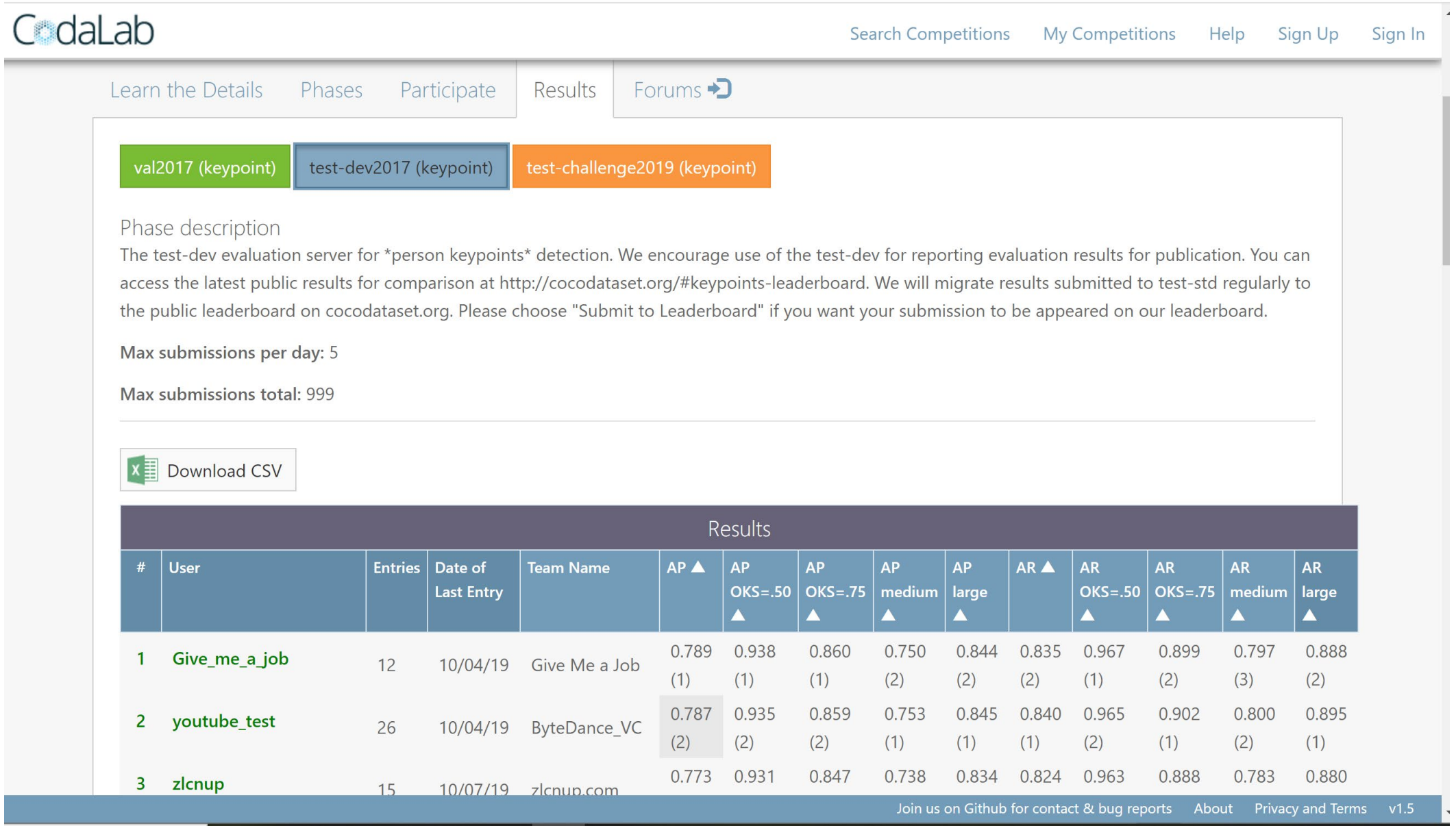
SURF

Mary Hester

Annette Langedijk

Kaggle


kaggle
Competitions
Documentation
InClass
New to Kaggle? Start here!
Our Titanic Competition is a great first challenge to get started.
Titanic: Machine Learning from Disaster Start here! Predict survival on the Titanic and get familiar with ML basics
Knowledge • Getting Started • Ongoing • 16734 Teams
General
InClass
Sort by
Grouped
All Categories
Search competitions
15 Active Competitions
#DFDC
Deepfake Detection Challenge
Identify videos with facial or voice manipulations
Featured · Code Competition · 3 months to go · video data, online video
$1,000,000
326 teams
2019 Data Science Bowl
Uncover the factors to help measure how young children learn
Featured · Code Competition · a month to go · video games, children, learning, education
$160,000
2,104 teams
NFL Big Data Bowl
How many yards will an NFL player gain after receiving a handoff?
Featured · Code Competition · 21 days to go · american football, sports
$75,000
2,038 teams



CodaLab

Search Competitions | My Competitions | Help | Sign Up | Sign In

Learn the Details | Phases | Participate | Results | Forums

val2017 (keypoint) | test-dev2017 (keypoint) | test-challenge2019 (keypoint)

Phase description

The test-dev evaluation server for *person keypoints* detection. We encourage use of the test-dev for reporting evaluation results for publication. You can access the latest public results for comparison at http://cocodataset.org/#keypoints-leaderboard. We will migrate results submitted to test-std regularly to the public leaderboard on cocodataset.org. Please choose "Submit to Leaderboard" if you want your submission to be appeared on our leaderboard.

Max submissions per day: 5

Max submissions total: 999

Download CSV

Results

#	User	Entries	Date of Last Entry	Team Name	AP ▲	AP OKS=.50 ▲	AP OKS=.75 ▲	AP medium ▲	AP large ▲	AR ▲	AR OKS=.50 ▲	AR OKS=.75 ▲	AR medium ▲	AR large ▲
1	Give_me_a_job	12	10/04/19	Give Me a Job	0.789 (1)	0.938 (1)	0.860 (1)	0.750 (2)	0.844 (2)	0.835 (2)	0.967 (1)	0.899 (2)	0.797 (3)	0.888 (2)
2	youtube_test	26	10/04/19	ByteDance_VC	0.787 (2)	0.935 (2)	0.859 (2)	0.753 (1)	0.845 (1)	0.840 (1)	0.965 (2)	0.902 (1)	0.800 (2)	0.895 (1)
3	zlcnup	15	10/07/19	zlcnup.com	0.773	0.931	0.847	0.738	0.834	0.824	0.963	0.888	0.783	0.880

Join us on Github for contact & bug reports | About | Privacy and Terms | v1.5

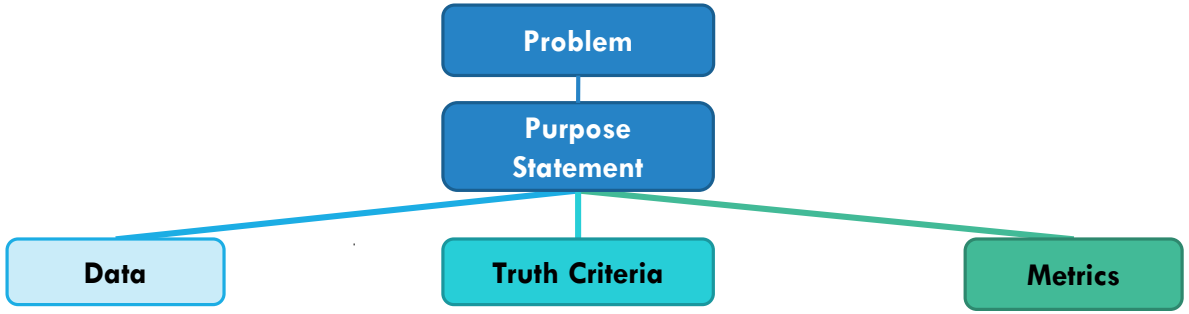
Benchmarking = Open Science

Can we learn from eachother?

Build on-top of each others knowledge?

“Voor organisaties binnen dezelfde branche met vergelijkbare activiteiten (bijvoorbeeld musea of toneelgezelschappen) is het interessant om de kwaliteit van deze activiteiten met elkaar te vergelijken door middel van een benchmark. Hierdoor kun je van elkaar leren: Wat lukt de één wel en de ander niet?“

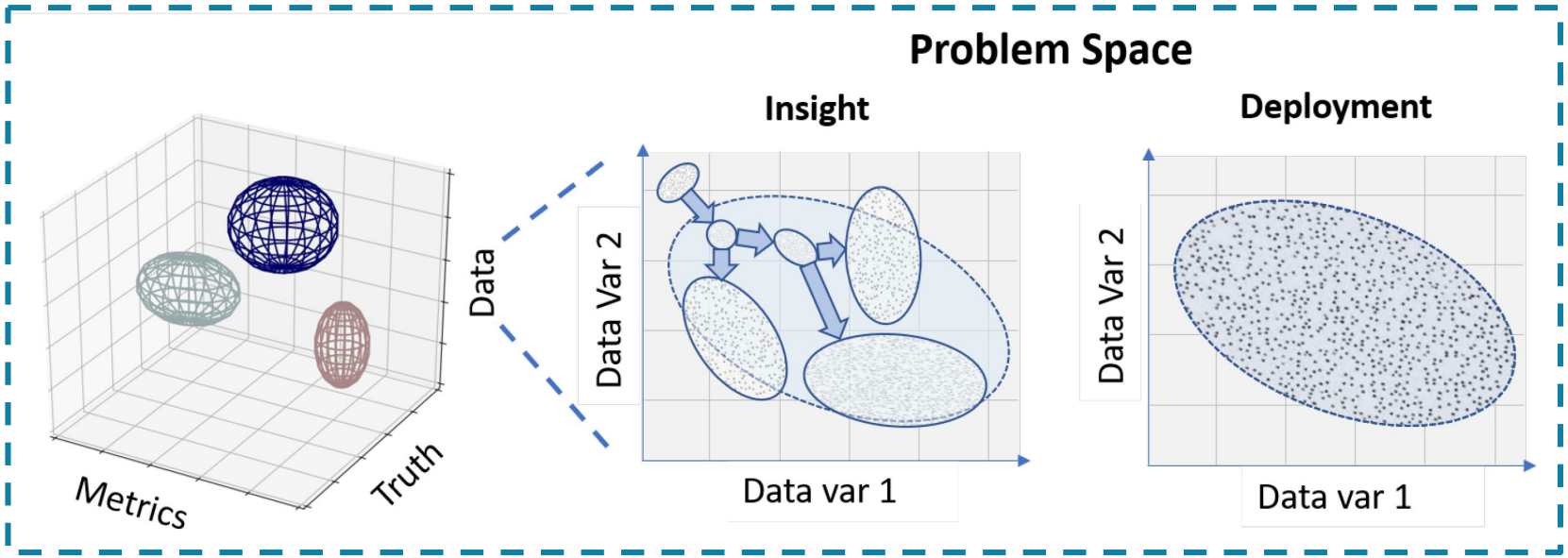
https://www.claudiadegraauw.nl/een-benchmark-wat-het-en-waarom-het-zinvol/



Framework for Benchmark Design

Insight & Deployment Benchmarks

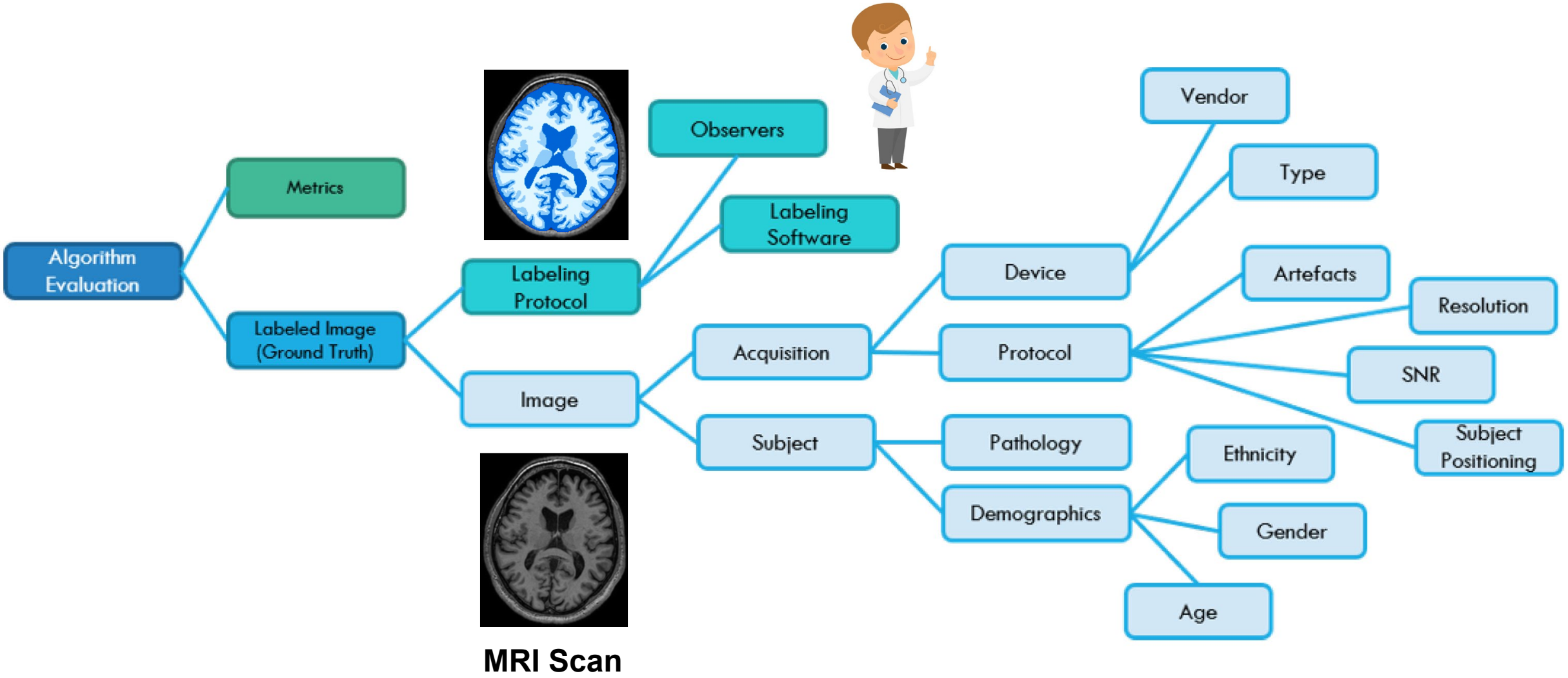
Experimental Research Study Design



A.M. Mendrik and S. R. Aylward.
"A Framework for Challenge Design: Insight and Deployment Challenges to Address Medical Image Analysis Problems."
arXiv preprint arXiv:1911.08531 (2019).

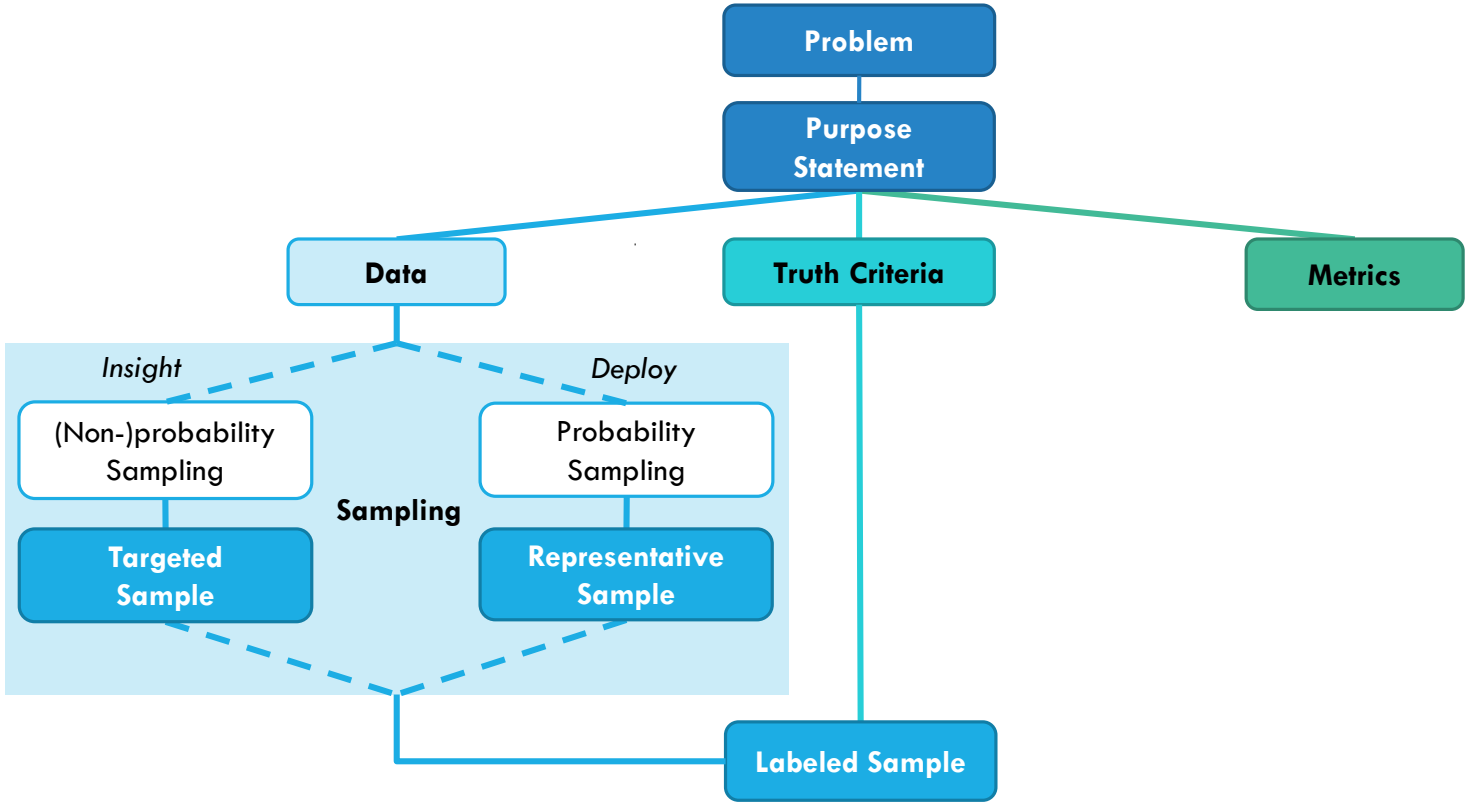
https://arxiv.org/pdf/1911.08531.pdf

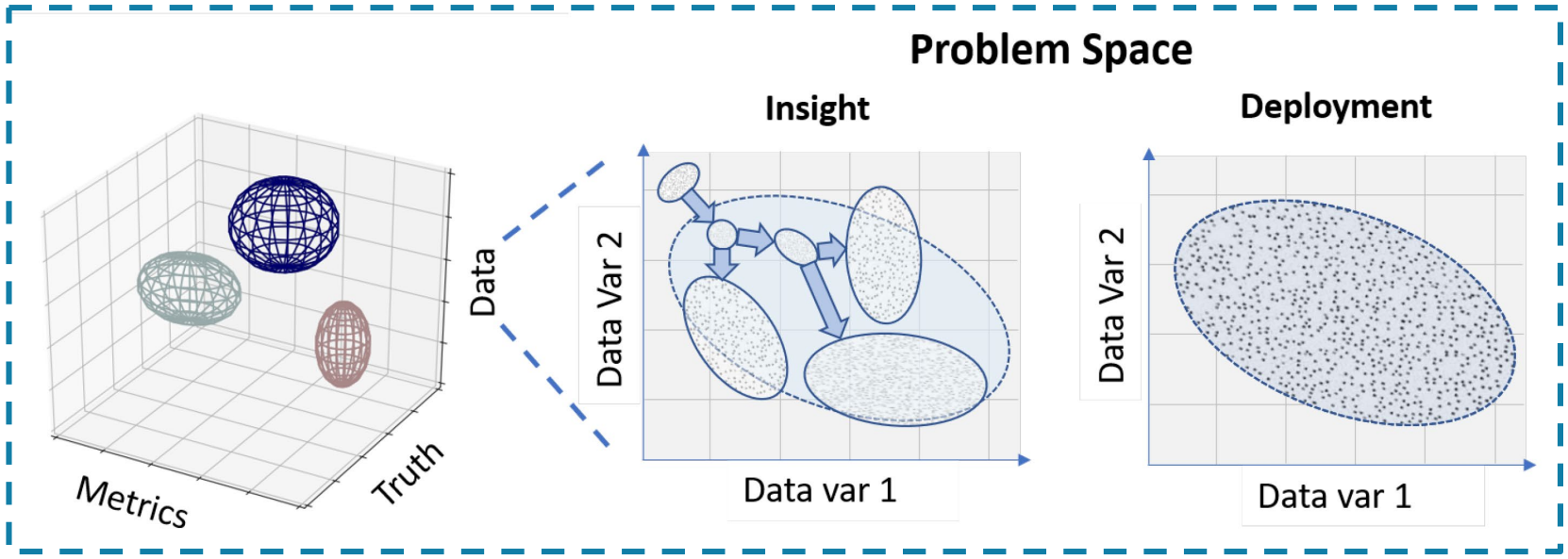
Statistical Population (Problem Space)



Framework for Benchmark Design

Insight & Deployment Benchmarks

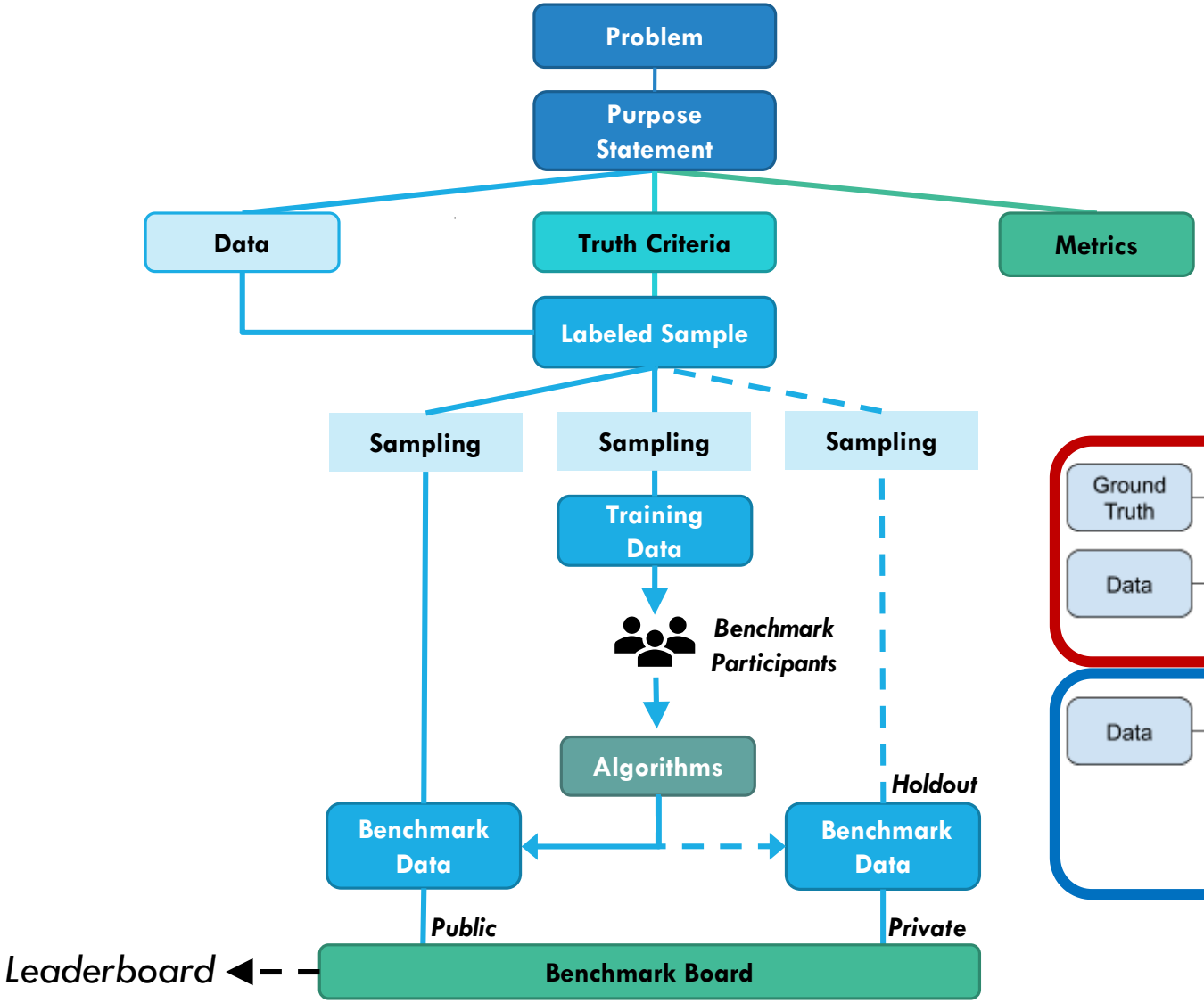


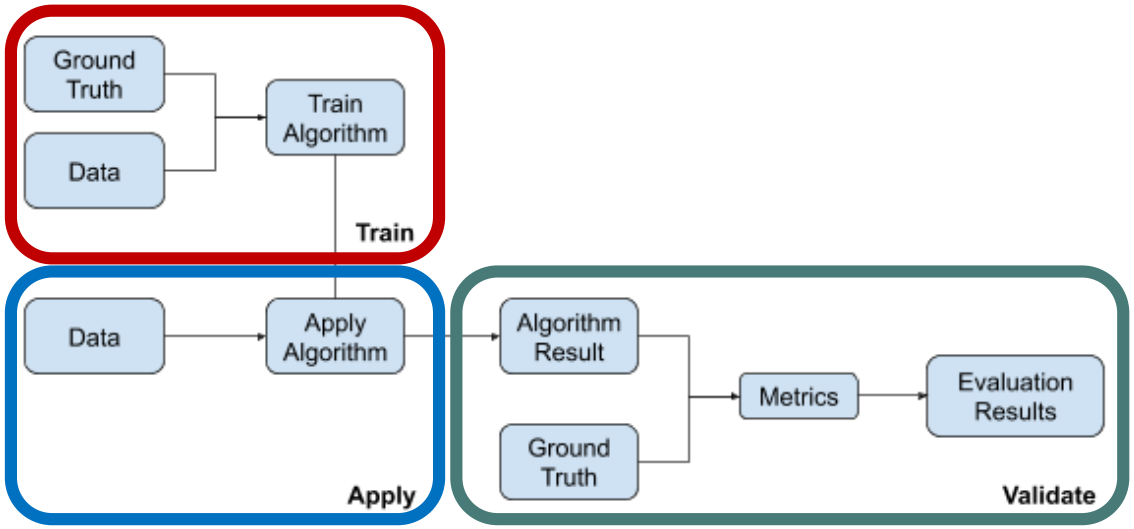


A.M. Mendrik and S. R. Aylward.
"A Framework for Challenge Design: Insight and Deployment Challenges to Address Medical Image Analysis Problems."
arXiv preprint arXiv:1911.08531 (2019).

https://arxiv.org/pdf/1911.08531.pdf

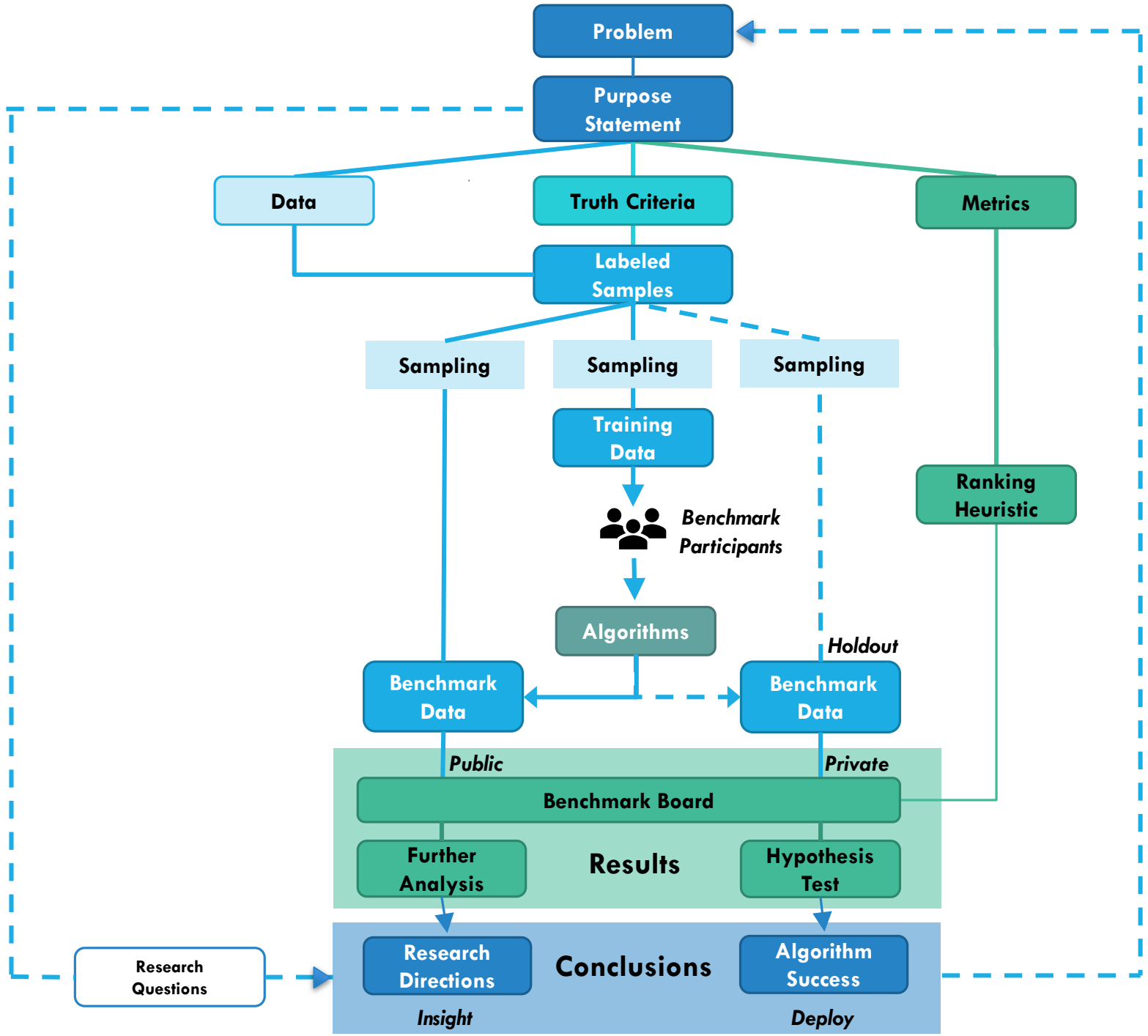
Framework for Benchmark Design





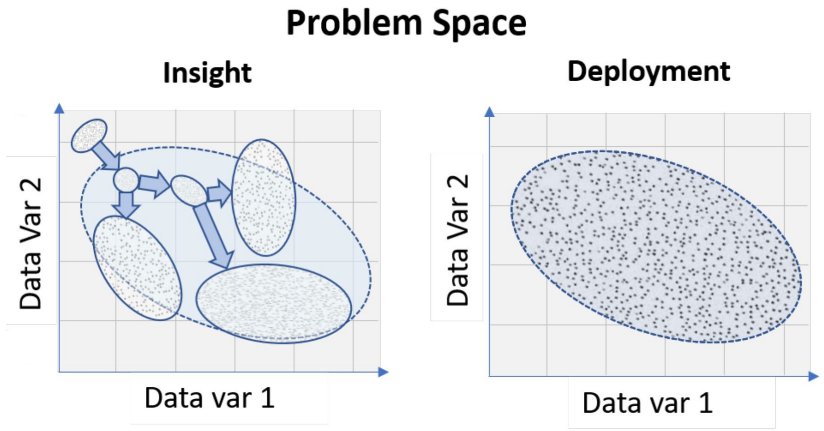
A.M. Mendrik and S. R. Aylward.
"A Framework for Challenge Design: Insight and Deployment Challenges to Address Medical Image Analysis Problems."
arXiv preprint arXiv:1911.08531 (2019).

https://arxiv.org/pdf/1911.08531.pdf



Framework for Benchmark Design

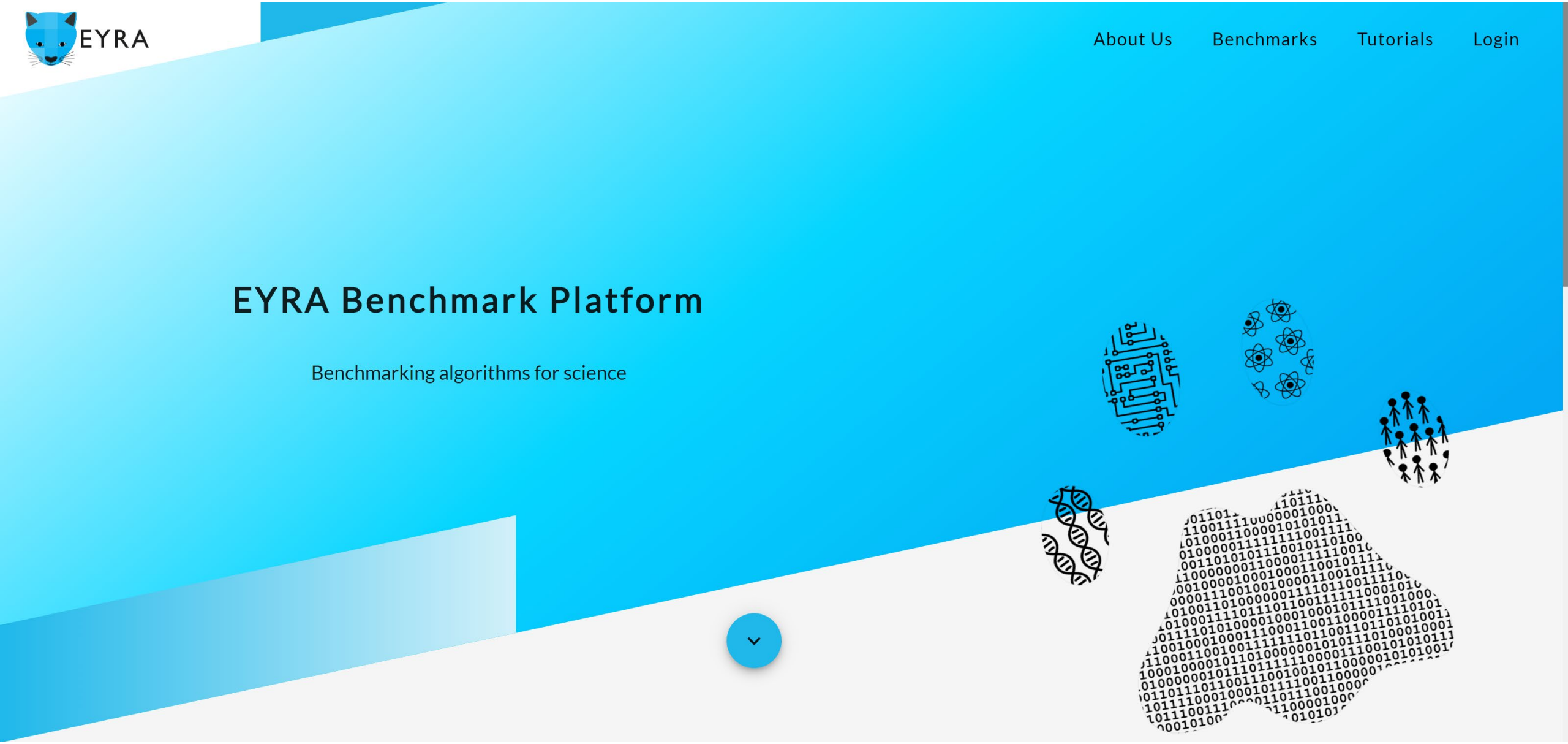
Insight & Deployment Benchmarks



A.M. Mendrik and S. R. Aylward.
"A Framework for Challenge Design: Insight and Deployment Challenges to Address Medical Image Analysis Problems."
arXiv preprint arXiv:1911.08531 (2019).

https://arxiv.org/pdf/1911.08531.pdf

https://www.eyrabenchmark.net/



EYRA Benchmark Platform Team

Tom Klaver, MSc
eScience Research Engineer

Maarten van Meersbergen, MSc
eScience Research Engineer

Dr. Adriënne Mendrik
eScience Coordinator

Pushpanjali Pawar, MSc
eScience Research Engineer

Evelien Schat, MSc
eScience Research Engineer

Dr. Roel Zinkstok
eScience Research Engineer

Dr. Janneke van der Zwaan
eScience Research Engineer

Haukur Pall Jonsson

Maurice Bouwhuis

Annette Langedijk

Giuseppe Gianquitto

Martin Brandt

Mary Hester

Ymke vd Berg

Based on the COMIC platform

https://github.com/comic/grand-challenge.org

Why GitHub? Enterprise Explore Marketplace Pricing Search / Sign in Sign up

comic / grand-challenge.org

Watch 10 Star 53 Fork 23

Code Issues 46 Pull requests 2 Security Insights

A framework to objectively evaluate the performance of machine learning algorithms in biomedical imaging https://grand-challenge.org

python django django-rest-framework docker machine-learning medical-imaging reproducible-research computer-vision ai challenges

4,383 commits 6 branches 4 releases 19 contributors Apache-2.0

Branch: master New pull request Find file Clone or download

jmsmkn Add GitHub actions for static code inspections (#977) Latest commit c275f2c 3 days ago

.github	Add GitHub actions for static code inspections (#977)	3 days ago
app	Remove images from reader study admin (#976)	3 days ago
dockerfiles	Adds flake8 checks (#959)	11 days ago
docs	Ignore flake8 error	3 days ago
.codeclimate.yml	Update .codeclimate.yml	7 months ago
.coveragerc	Update for coveralls (#370)	2 years ago
.csslintrc	Adds code climate	3 years ago

Dr. James Meakin
Radboud University Medical Center (Nijmegen)

Prof. dr. Bram van Ginneken
Radboud University Medical Center (Nijmegen)

Platform Requirements

Tutorial on Benchmarking Algorithm Performance

October 29th at the 14th IEEE International Conference on eScience 2018, Amsterdam, the Netherlands

View the Project on GitHub NLeSC/IEEE-eScience-Tutorial-Designing-Benchmarks

14th eScience IEEE INTERNATIONAL CONFERENCE

Benchmarking Algorithm Performance for Research

Kasper Marstal
Erasmus MC

Melvin Wevers
Digital Humanities Lab
KNAW Humanities Cluster

Liam Connor
UvA

Maria Eskevich
CLARIN ERIC

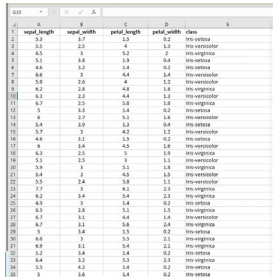
Marius Staring
LUMC

Mike Lees
UvA

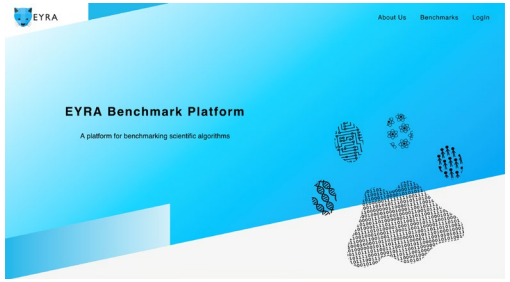
Example Workflow for a Benchmark on the EYRA Platform

Algorithm Developers

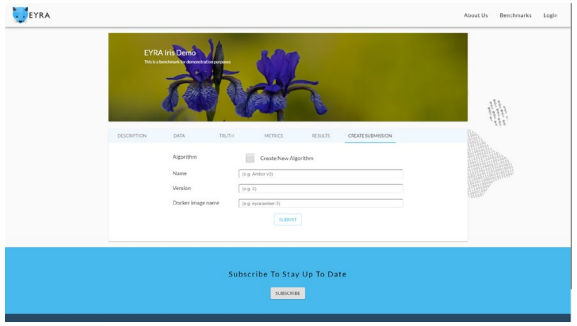
Download Training Data



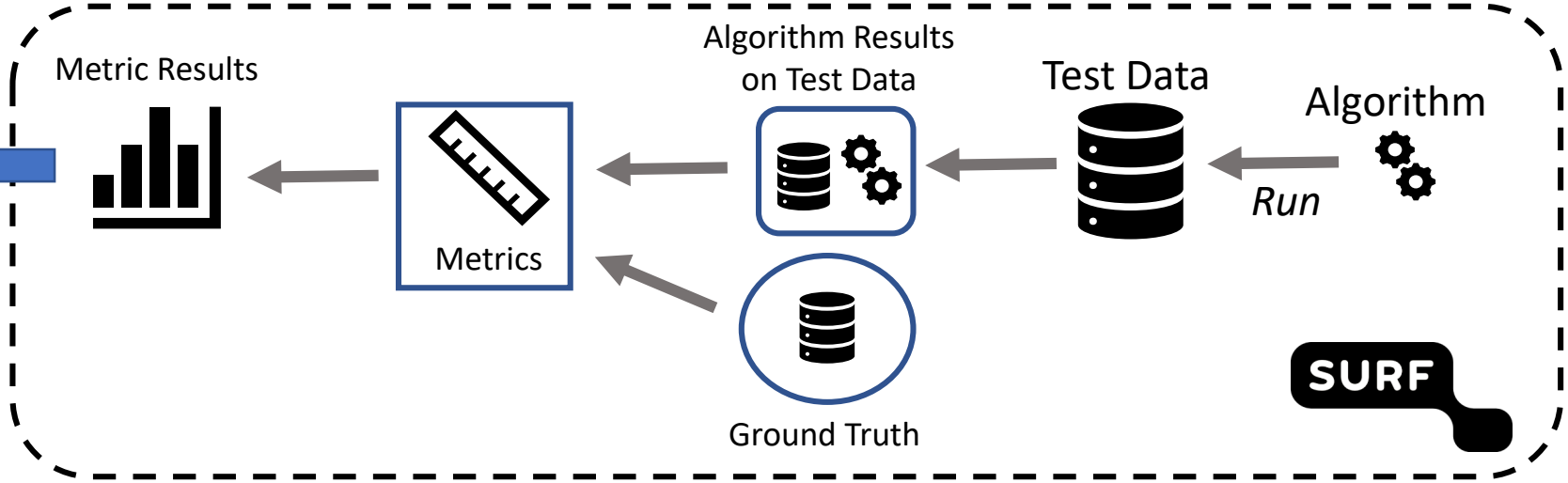
Login on EYRA Platform



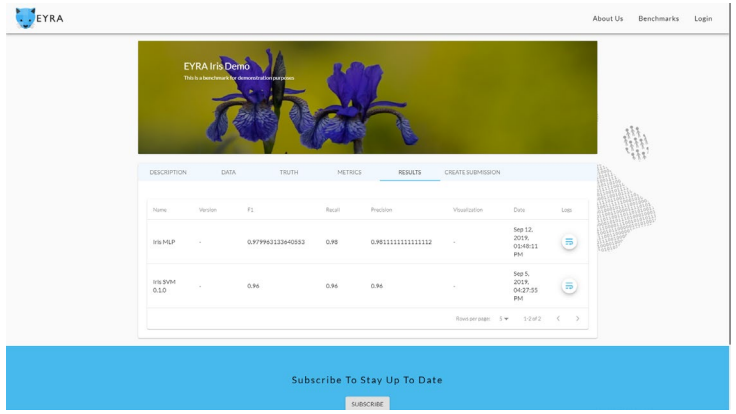
Submit Algorithm



Run Algorithm at SURF in the Cloud and Evaluate Results

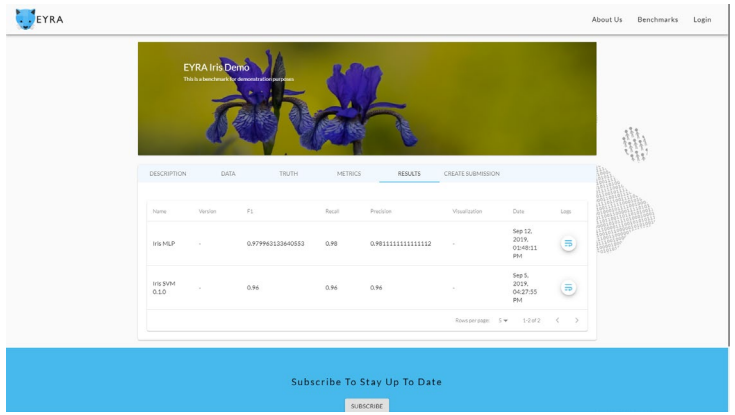


Show Algorithm performance on EYRA Platform



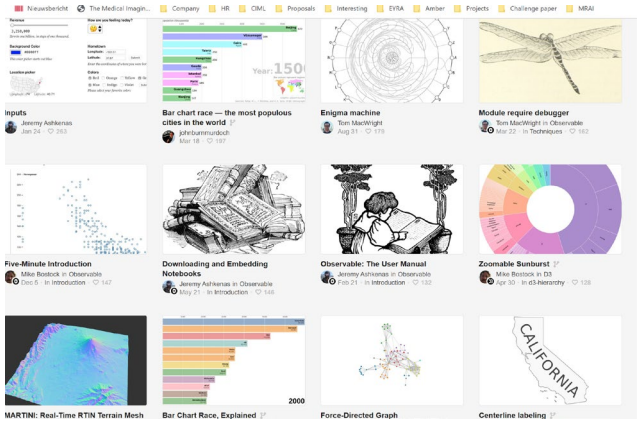
Example Workflow for a Benchmark on the EYRA Platform

Show Algorithm performance on EYRA Platform



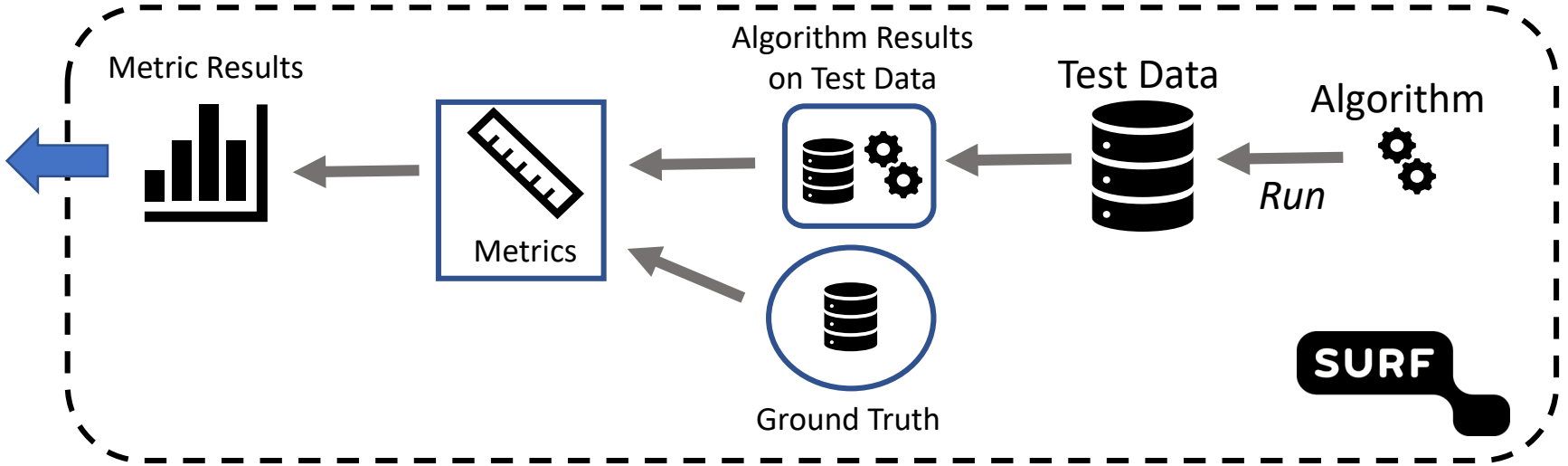
Further Analysis of Results

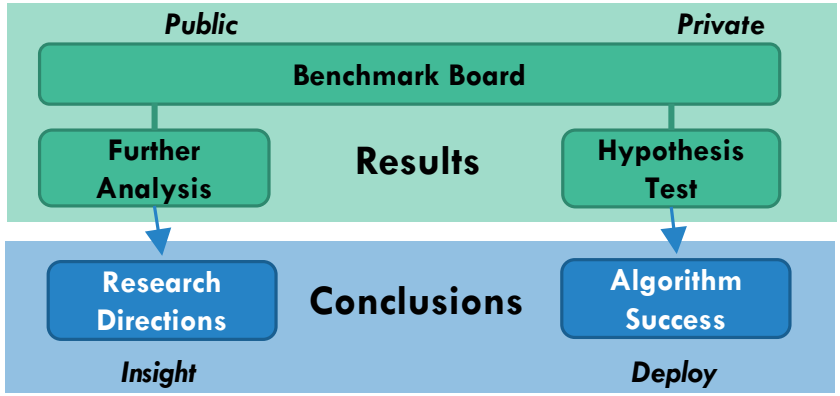
Observable



observablehq.com

Run Algorithm at SURF in the Cloud and Evaluate Results





Tom Klaver
Netherlands eScience Center

OVERVIEW
BENCHMARKS
SUBMISSIONS
ALGORITHMS

My Benchmarks

Fast Radioburst Detection

EYRA Iris Demo

My Submissions

Benchmark ↑	Started	Name	Status(impl.)	Run time(impl.)	Status(eval.)	Run time(e
EYRA Iris Demo	02-08-2019 15:39		⧖	aN:aN:aN	✗	aN:aN
Fast Radioburst Detection	11-06-2019 16:37	Amber v1 on FRB 1.5GB	✓	00:02:17	✓	00:0(
Fast Radioburst Detection	11-06-2019 16:32	Heimdall v1 on FRB 1.5GB	✓	00:02:22	✓	00:01

More Submissions >

My Algorithms

User Manual

If you want to participate in a benchmark on the EYRA Benchmark Platform, you need to create a model based on participant data, that can predict outcomes (e.g., class labels or numeric values) given test data. What you submit to the benchmark platform is a Docker container that does the predictions, given your model or algorithm. Benchmark organisers need to provide a Docker container for evaluating the results produced by the models/algorithms of participants. The EYRA tools can be used to generate boilerplate containers that set up as much as possible, so you can focus on implementing the prediction algorithm or evaluation metrics.

To be able to use the Docker container as an submission or evaluation, you need to publish it on Docker Hub. If you (or your organization) do(es) not yet have a Docker Hub account, you need to sign up for one.

Quickstart

1. Generate a boilerplate container by running: `eyra-generate [submission|evaluation] <name>`. A directory called `<name>` is created.
2. Put the input data in the `<name>/data/input/` directory.
   - Benchmark participants should download the public test data and put it in this directory. The file should be called `data/input/test_data`.
   - Benchmark organisers should create an example output file in the same format as the ground truth and put them in the `data/input/` directory. The files should be called

This is a benchmark for demonstration purposes.

This benchmark was created to help potential benchmark organizers and participants understand how the EYRA Benchmark Platform works. This benchmark uses the Iris dataset. And the task is to predict Iris species based on four measurements: sepal length, sepal width, petal length, and petal width.

Have a look at the EYRA Iris Demo benchmark tutorial. The data and implementations of the evaluation algorithm and an example submission can be found on github.

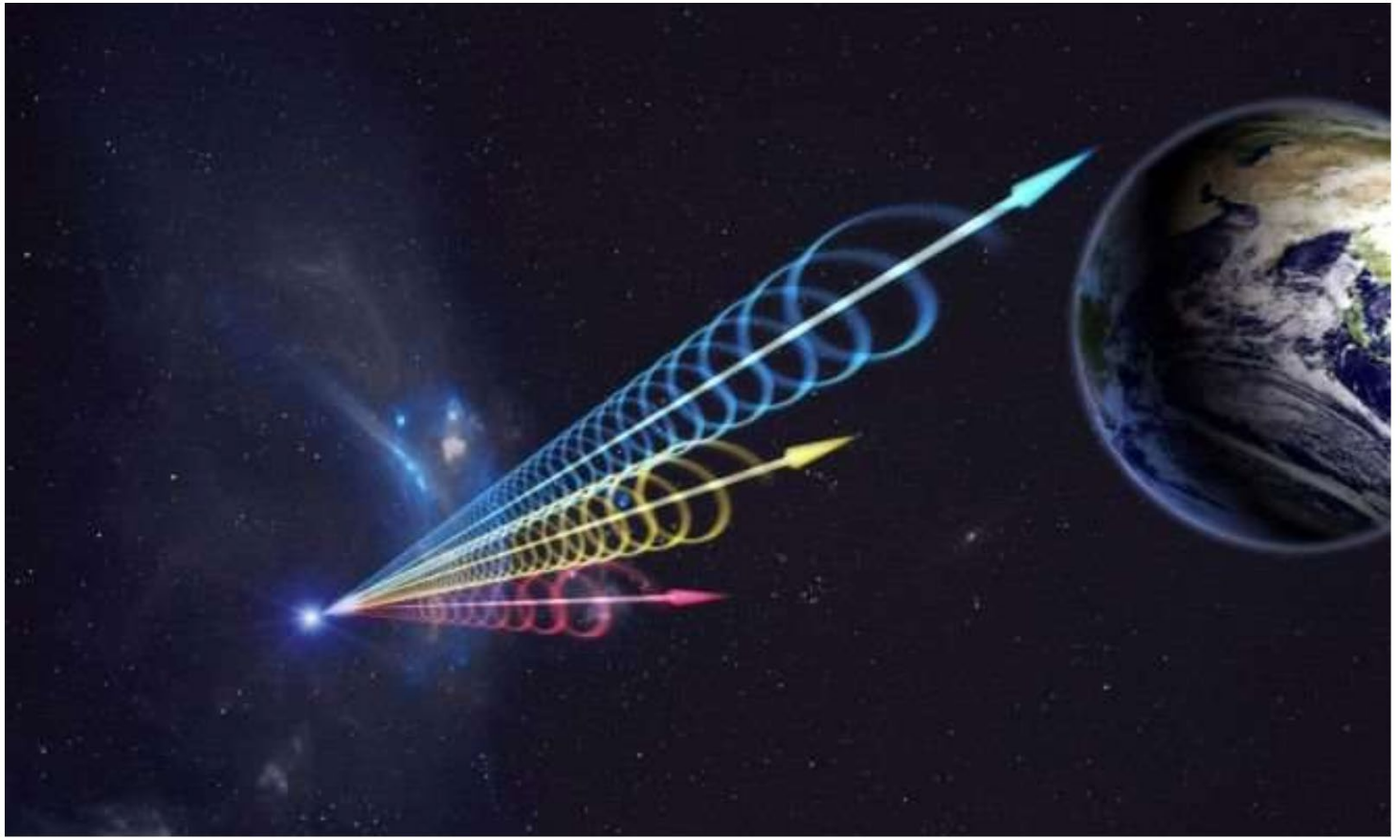
Submissions

Submission containers should output their data to `/data/output`, as a CSV file. The first row should be the column name (`class`), next should be a single predicted class per line for each row of test data.

example /data/output:

```
class
Iris-setosa
Iris-setosa
Iris-versicolor
Iris-versicolor
Iris-setosa
Iris-versicolor
```

- dr. Liam Connor (UvA), dr. Joeri v. Leeuwen (ASTRON), dr. Alessio Sclocco (eScience center)
- Fast Radio Burst Detection on Radio Telescope Data
- Several software packages world wide, but never compared on the same data
- Three benchmarks planned



Fast Radio Bursts are brief but powerful and as-yet-unexplained bursts of radio energy originating from deep space. Image via Jingchuan Yu/Beijing Planetarium, NRAO.

Future benchmark

University of Salford MANCHESTER

PRImA Research Lab

Email | Password | Log in

Not registered? - Request an account here

Home | Expertise | Research | People | Tools | Datasets | In the press | Contact

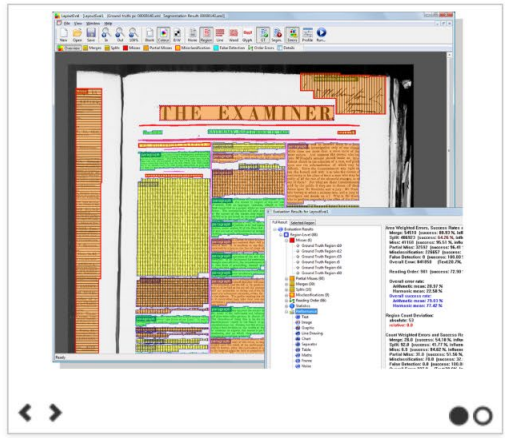
PRImA

Pattern Recognition & Image Analysis Research Lab

PRImA / Tools / Performance Evaluation /

- Tools Home
- Aletheia
- WebAletheia
- **Performance Evaluation**
- PAGE Libraries
- PAGE Converter and Validator
- PAGE Metadata Scanner
- Tesseract OCR to PAGE
- PAGE Viewer
- Extractor / Exporter

Performance Evaluation A framework for Performance Analysis of OCR methods



Layout Evaluation

Overview

This tool is part of a framework for evaluating the performance of layout analysis methods. It combines efficiency and accuracy by using a special interval based geometric representation of regions. A wide range of sophisticated evaluation measures provides the means for a deep insight into the analysed systems, which goes far beyond simple benchmarking. The support of user-defined profiles allows the tuning for practically any kind of evaluation scenario related to real world applications.

The framework has been successfully delivered as part of a major EU-funded project (IMPACT) to evaluate large-scale digitisation projects and has been validated applied within the past three ICDAR Page Segmentation Competitions.

Features

- Evaluation Profile editor with Novel Metrics
  - Fully customisable metrics for different Region Types (Text, Image, Table, ...)
  - Fully customisable metrics for different Error Types (Merge, Split, Miss, Partial Miss, False Detection and Missclassification)

KB LAB

Lotte Wilms

KB Koninklijke Bibliotheek Nationale bibliotheek van Nederland

https://www.eyrabenchmark.net/

EYRA

About Us Benchmarks Tutorials Login

EYRA Benchmark Platform

Benchmarking algorithms for science