Benchmarking in the Social Sciences

Paulina Pankowska, Adriënne Mendrik,
Daniel Oberski, and Javier Garcia-Bernardo

Funded by
ODISSEI

Universiteit Utrecht
Eyra

What is benchmarking?

- Creates a framework, which allows to compare different models, methods, approaches
- Enables to analyse the strengths and weaknesses of approaches – which method works better in which context?
- Commonly used in IT and data science/machine learning
- Usually focuses on predictions – how well can we predict Y given the X's?

Traditional statistics versus Benchmarking approach

Traditional research workflow

Data → Fit a model
Evaluate the performance (effect sizes / R^2)
Test significance

Benchmarking workflow

Participant data → Fit a model

Holdout data → Evaluate predictions

Typical problems:

p-hacking

Theory after data

Replicability crisis

Testing predictions:

Robust way of assessing the quality of models

Facilitates comparison of competing theories and models

Avoids overfitting

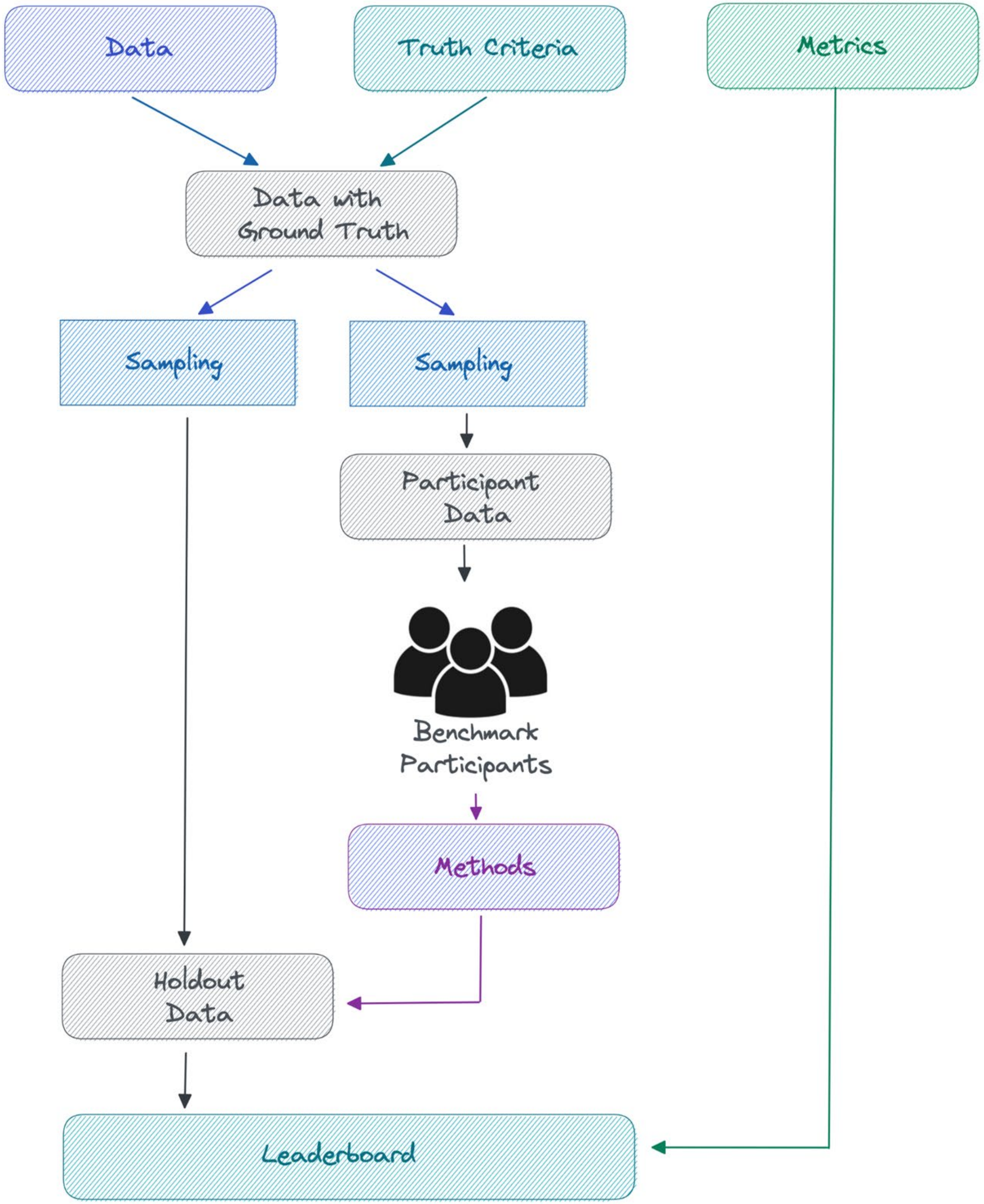
Benchmark challenge setup

Physical challenge (shorter period of time)

Ongoing online challenge (longer period of time)

e.g., https://www.kaggle.com/competitions

Benchmark challenge setup



An example of a well-known benchmark:

The Netflix Prize (2006 to 2009)

Open competition to search for the best method to predict what kind of movies a user would like, based on user ratings

Goal: Make the company's recommendation engine 10% more accurate

Participant data: over 100 million ratings of 17,770 movies from 480,189 customers.

Benchmarking for the Social Sciences – Opportunities

How can benchmarking advance social sciences?

Understand which method works best for which research problems: regression, ABM, network analysis

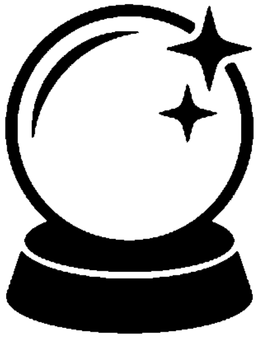
By making benchmark data accessible, more researchers can contribute to these research problems

Understand the limits of predictions for certain research problems

Benchmarking for the Social Sciences - Challenges

Social scientists are interested in causal mechanisms

However:



Causal models make predictions

We already use predictions to understand our models (e.g. R^2)

We are also increasingly interested in predicting outcomes

The Fragile Families Challenge

'A mass collaboration that combines predictive modeling, causal inference, and in-depth interviews to yield insights that can improve the lives of disadvantaged children in the United States'

Uses data from the *Fragile Families and Child Wellbeing Study* (FFCWS)

N ~ 5,000

T = 15 years (children followed from birth through age 15)

7 survey waves

The Fragile Families Challenge

Consists of two steps:

1. Participants built predictive models of six life outcomes (e.g., GPA) and the predictive performance was evaluated with holdout data

2. Using the individual models and the community model to conduct further substantive and methodological research

The Fragile Families Challenge

Diverse group of participants (i.e., social & data scientists)

Combined social science and data science approaches

Survey data with rather small N ⟶ less suitable for ML approaches

Based on (largely) publicly available data

SICSS-ODISSEI Summer School Benchmark

Goal: Predict career outcomes, namely **contract type** and **income level** prediction, with a focus on the prediction of temporary contracts & low income (*precarious employment*)

Data: CBS administrative data used to predict contract type and income level in 2020 based on 2010/11 data

Focus: Benchmark of statistical methods and predictors

SICSS-ODISSEI Summer School Benchmark – setup

Approx. 20 participants divided into 6 teams (3-4 participants per team)

Mainly social scientists & PhD candidates

5 days to prepare data and run analysis

Participants were provided with baseline dataset & could request additional data

The submissions incl. predictions, code and narrative

SICSS-ODISSEI Summer School Benchmark – evaluation criteria

Predictive accuracy, overall & of main categories of interest)

Innovativeness and embeddedness in theories and existing research

SICSS-ODISSEI Summer School Benchmark – results

The teams used the following methods

Team	Algorithm
Hamster	Gradient boosting
Team Blind and Deaf	Gradient boosting
The Black Box	Random forest
Team Trying	Random forest
srgd	Random forest
Run Forest Run	(Extreme) gradient boosting

Models included a wide range of features:
educational data, socio-economic status, parental background, and migration background

Model selection based on cross-validation

SICSS-ODISSEI Summer School Benchmark – results

Quantitative evaluation

Team	F1 Score precarious employment	Global accuracy	1-RMSE/4 of income	F1 of contract type
Hamster	**0.252**	0.379	0.693	0.454
Blind and Deaf	0.002	0.139	0.640	0.186
Black box	0.227	0.345	0.686	0.437
Trying	0.092	0.079	0.581	0.189
SRGD	0.000	0.000	0.000	0.000
Run Forest Run	**0.250**	0.369	0.683	0.438

SICSS-ODISSEI Summer School Benchmark – results

Qualitative evaluation

Team	Expert 1			Expert 2			Both experts		
	Innovative	Embedded	Combined score	Innovative	Embedded	Combined score	Reversed score	Combined score	Ranking
Hamster	5	5	5	5	2	3.5	3.5	4.25	1
Team Blind and Deaf	3	4	3.5	5	5	5	2	2.75	5
The Black Box	3	4	3.5	5	1	3	4	3.75	3
Team Trying	6	4	5	5	3	4	3	4	4
srgd	2	2	2	6	6	6	1	1.5	6
Run Forest Run	4	6	5	1	5	3	4	4.5	2

SICSS-ODISSEI Summer School Benchmark – conclusions

Main constraints

Computational performance of the secure access environment

- This constrained pre-processing and model choices significantly.

Identifying appropriate linkage variables

- Participants were unable to include all features they wanted.

Time

- Large amount of time was spent on processing and linkage.
- Participants would benefit from a wide range of pre- processing options and mappings.

SICSS-ODISSEI Summer School Benchmark – lessons learned

One week was insufficient for the teams to fully address the challenge.

Given the constraints, we are still short of understanding the full potential of benchmarking in the social sciences.

Nevertheless, feedback from participants was highly positive.

Thank you!

p.k.pankowska@uu.nl

a.m.mendrik@eyra.co

d.l.oberski@uu.nl

j.garciabernardo@uu.nl

Funded by

ODISSEI

Eyra

Universiteit Utrecht