RESEARCH DATA UNIT

Scientific AI
Hamprecht Lab, IWR,
Heidelberg University

SCIENTIFIC SOFTWARE CENTER

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

AI in research software: Best practices

Research Data Unit: Dr. Georg Schwesinger and Dr. Sebastian Zangerle
Scientific AI group: Peter Lippmann
Scientific Software Center: Dr. Inga Ulusoy

February 2025

1. Requirements of “ML-based science”

What this course is not

- An introduction to data science
- An introduction to machine learning
- A course about different ML algorithms
- A course about different ML training approaches and libraries
- …

What this course is

- A best practices guide to creating machine learning based research software (MLBRS)
- A recommendation on how to manage and prepare your data
- A recommendation on how to train your models
- An introduction to software engineering best practices for MLBRS
- A guideline on how to generate independently reproducible scientific results using data-based approaches
- A guideline on how to publish your data and your models

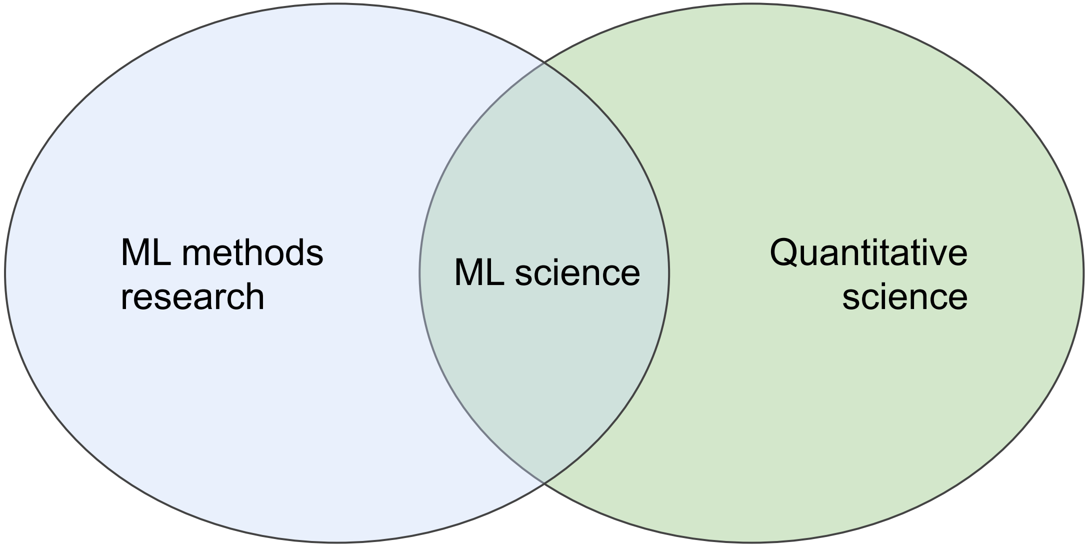
What is special about research software based on data? (“ML-based science”)

ML science

- Scientific research that uses machine learning models to extend scientific knowledge
- Answers a scientific question by using ML
- No restriction on algorithm, method, library, domain

Contrary to:

- ML methods research: Research on ML methods and algorithms with the goal to improve the field of ML



Sayash Kapoor *et al*, http://arxiv.org/abs/2308.07832

Research software

“... software that is developed and used in the context of research…”

Shifting requirements

A scientific question is answered using computation/simulation, but the way the problem is solved changes as part of the research process.

Passed along researchers

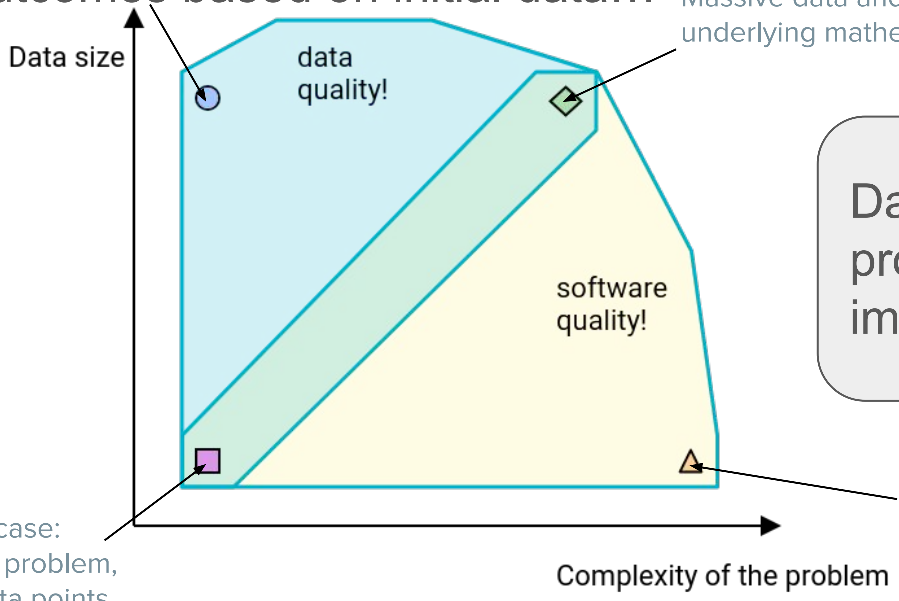
Initially developed for one purpose but then often organically extended depending on the researcher’s needs.

Development Practices

Often created by researchers that have no fundamental training in software engineering and inherit practices from those around them.

ML-based research software

“... software that is developed and used in the context of research and predicts outcomes based on initial data…”



Mathematically simpler problem, but large amount of data

Massive data and complex underlying mathematical structure

Trivial case: simple problem, few data points

Mathematically complex problem, but simple data

Data ***and*** software quality provide limits to quality and impact of research!

MLBRS: Data

Data is foundation for..

…model training, decision making and/or predictions.

Different kinds of data

For example, numerical data, textual data, images, audio, video.

Metadata

What is relevant metadata and should be included on the data card?

Availability and licensing

Will the data be publicly available to the community? What license does/will the dataset have?

Legal considerations

Where does the data come from? Is it licensed? Is it public or private data? In what form is the data stored and processed?

Ethical considerations

Does the data exploit work of others? Does it break some sort of confidentiality? Will it impact in a possible harmful way or can it be misconstrued to do harm?

Bias

Is there an inherent bias in the data itself, due to the data collection approach, or other reasons?

MLBRS: Software

Purpose

Will the software be more widely used, be an in-house code, or one-person software?

Software engineering best practices

Does the software follow software engineering best practices (version control, testing, documentation, …)?

Usability and reproducibility

Does the software include documentation on how models can be trained, and keeps track of training parameters? Does the software help to generate model cards and provide models in transferable format?

Accuracy and reliability

Does the software create robust and consistent results, even though it is based on a non-deterministic process?

Legal considerations

Does the software incorporate third-party models and/or code?

Legal considerations

What license is the software published under?

What license are models published under?

Security

Is the software secure against data injection?

Reproducibility

- Provide data to enable others to reproduce findings
- Provide code to enable others to reproduce findings

➔ ***Computational reproducibility (i)***

- Make sure your findings are true findings, and do not arise from problems with your data/code

➔ ***Independent reproducibility (ii)***

Research software engineering generally targets (i), but with MLBRS we target (ii)

Why should you care?

Your research integrity, scientific best conduct (malpractice), can have long-lasting detrimental effect on science (impact on others and the field), affects society!

Key aspects

- Legal aspects
- Documentation on model training, hyperparameter tuning, model testing
- Model bias
- Software quality
- Reproducibility of the model’s predictions
- Robustness of the model(s)
- Software security
- Reproducibility of the model training
- Ethical aspects
- Data bias
- Documentation on data collection, data cleaning, feature selection
- Data leakage