A Robust Estimate of Performance of Reproducible Analytical Models for DREAM Challenges

Witold R. Rudnicki, Wojciech Lesiński, Radosław Piliszek, Institute of Informatics, University of Białystok
Julio Saez Rodriguez, Faculty of Medicine of RWTH-Aachen University,

DREAM

Dialogue for **R**everse **E**ngineering **A**ssessments and **M**ethods

DREAM

- Community effort that aims at answering important questions in biology and medicine
- Crowdsourcing research - open for anyone willing to participate
- Open science - challenge data* and algorithms developed by participants are open
- Objective evaluation of algorithms

* with limitations due to legal concerns

DREAM

Data preparation

Stating the questions

Technical setup, preparing the scoring system

Evaluation of results

Evaluation of results

Timeline of the DREAM Challenge

The open phase - competition

The community phase - collaboration

Writing paper

All data published

X months | 3-6 months | ~3 months | > 12 months

People involved

10

200

20

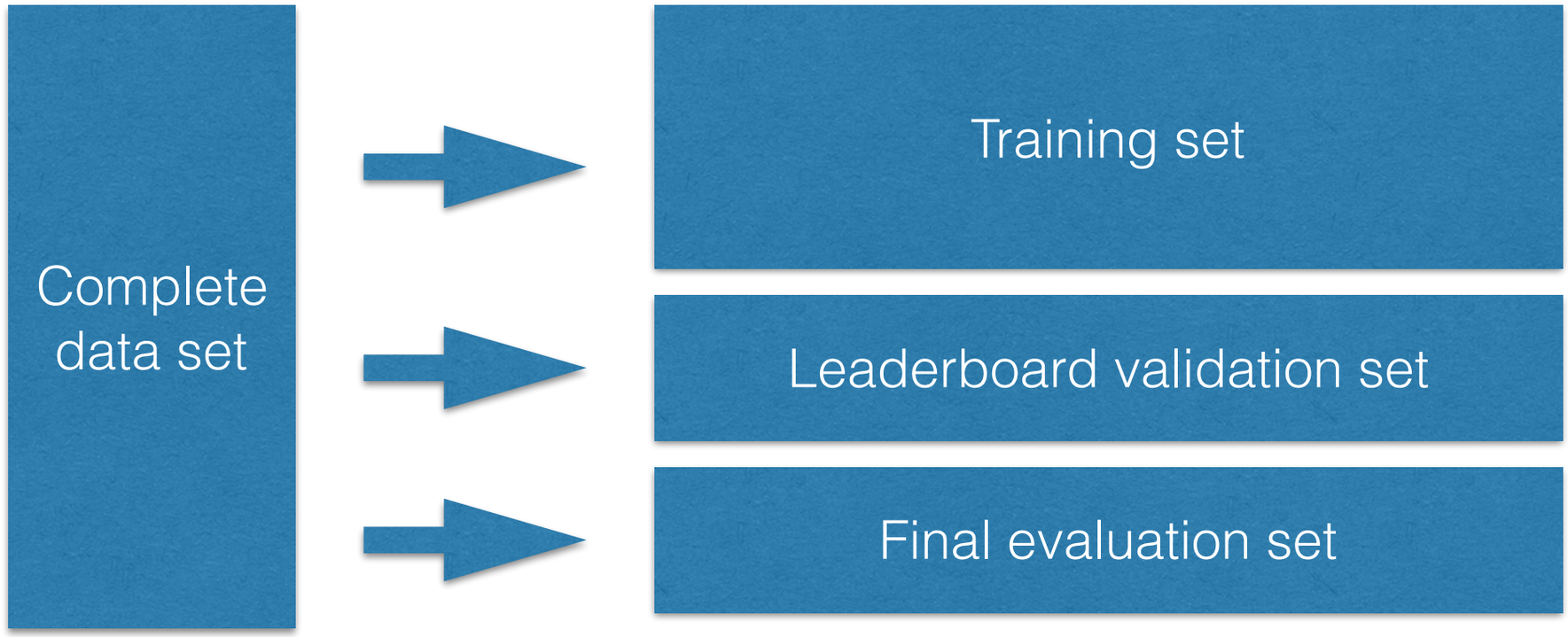
20

2

0

DREAM

Data split



DREAM

- Big success -
  - 100’s of groups participating
  - ~3 challenges a year
  - joint DREAM / RECOMB conferences a year
  - papers in a prestigeous journals:
    - Nature Methods, Nature Biotechnology, Nature Reviews Genetics, PLoS Computational Biology, Nature Genetics, Science Translational Medicine, Genome Research et.

PROBLEM

- DREAM Challenges give opportunity to create gold standards for comparison of algorithms

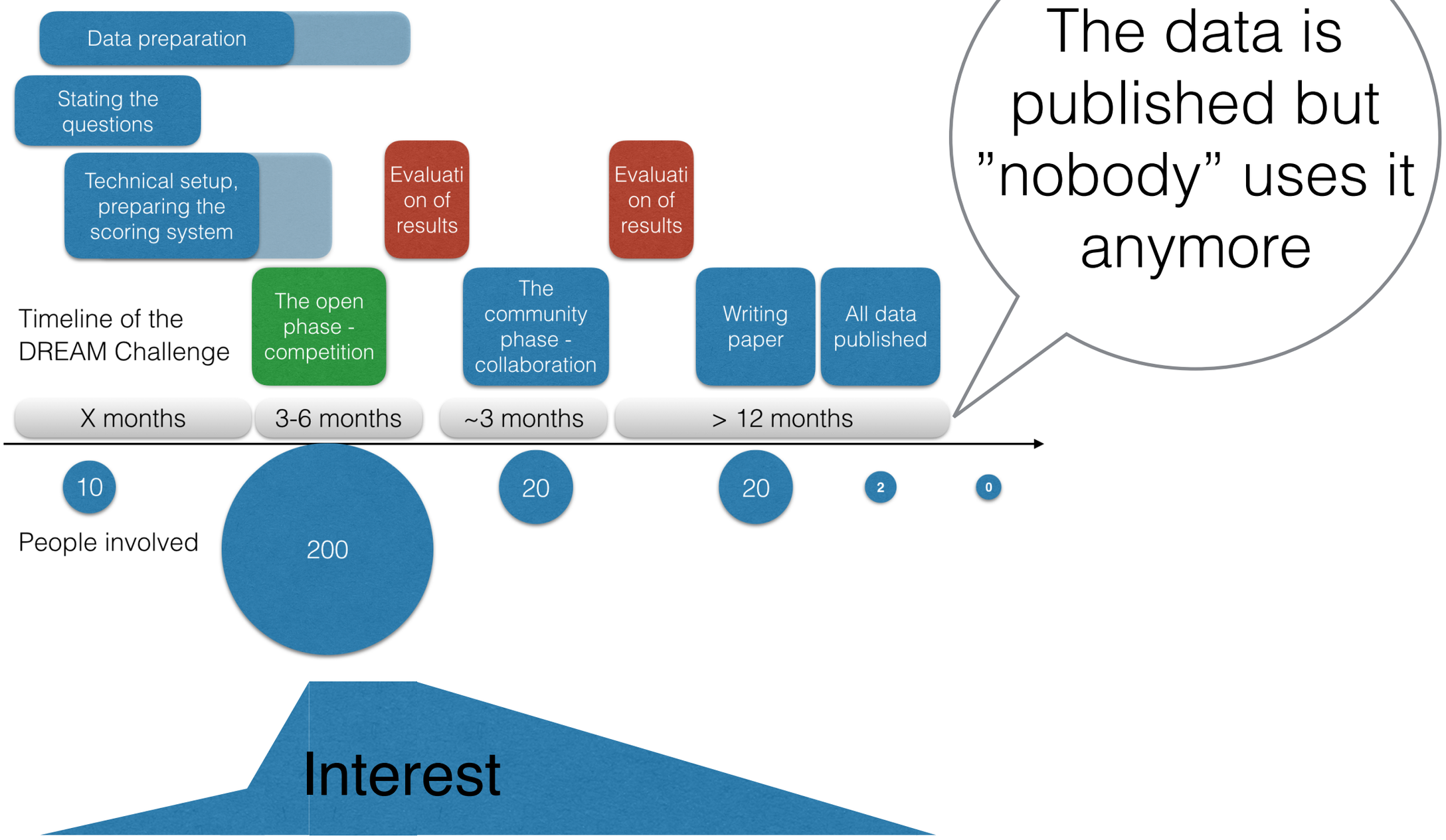
  however

- the opportunity is mostly untapped due to

- time gap between competitive phase and release of the data

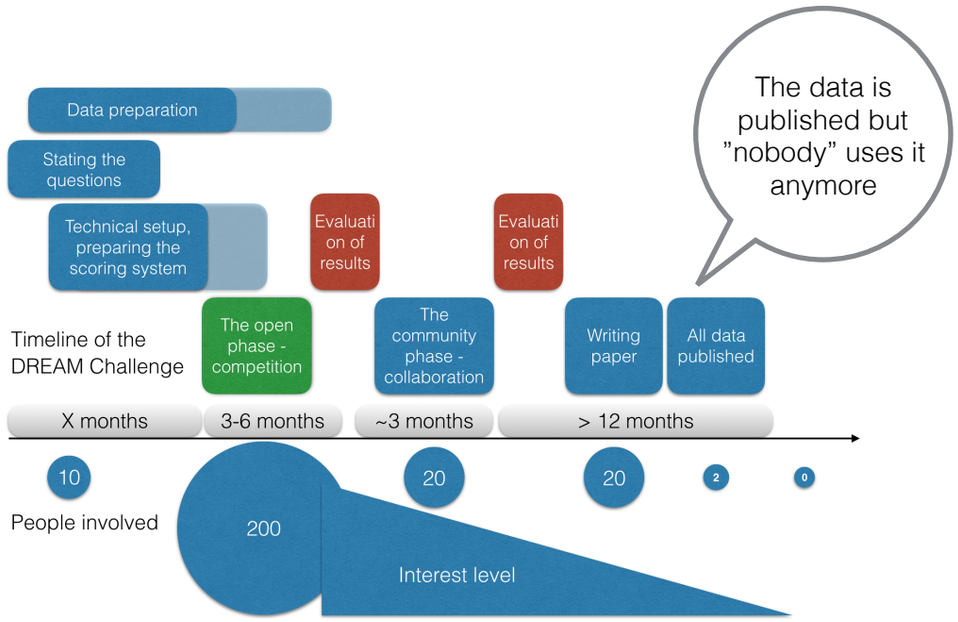
- release of the complete data makes the standard evaluation scheme irrelevant

PROBLEM


Data preparation
Stating the questions
Technical setup, preparing the scoring system
Evaluation of results
Evaluation of results
Timeline of the DREAM Challenge
The open phase - competition
The community phase - collaboration
Writing paper
All data published
X months
3-6 months
~3 months
> 12 months
10
200
20
20
2
0
People involved
Interest
The data is published but ”nobody” uses it anymore

PROBLEM

- The gold standard data is there
- But
  - Most teams have lost interest and walked away
- No way to objectively asses performance of new algorithms against the challenge result

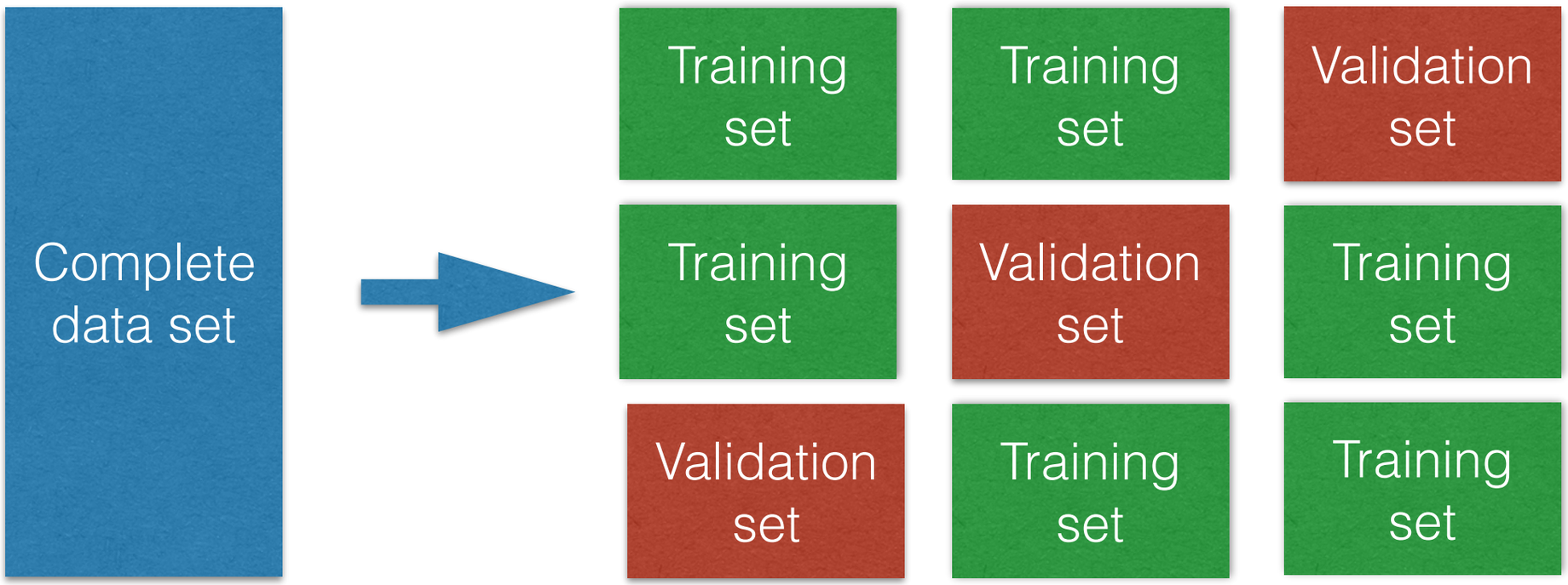


Solution

- Participants send their algorithms not responses
    - algorithms need to learn from random sample of the data and build predictive models
    - that are evaluated on the withheld sample
        - in the cross-validation loop
- Procedure is repeated numerous times

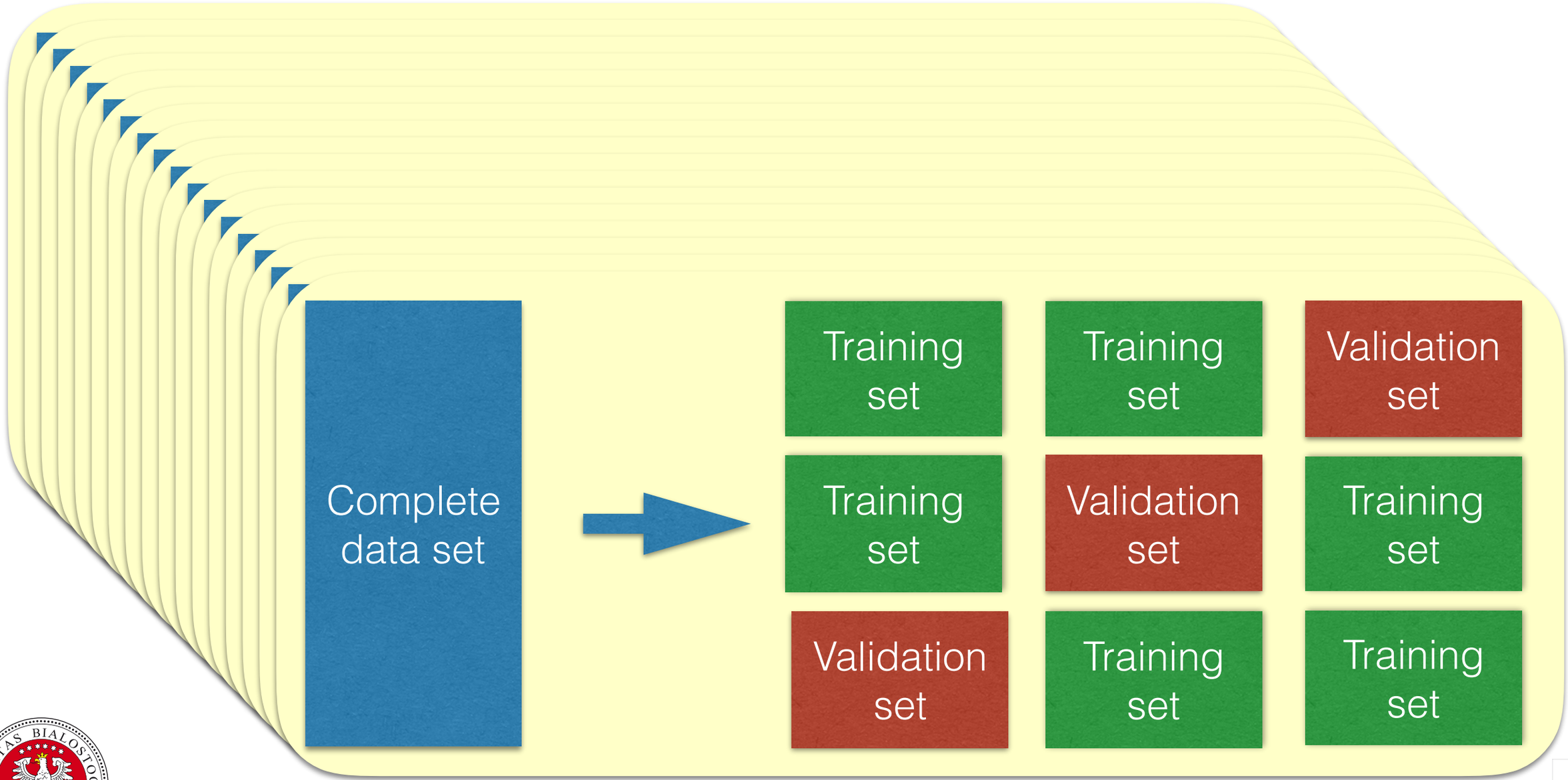
Solution

- 3 -fold cross validation scheme



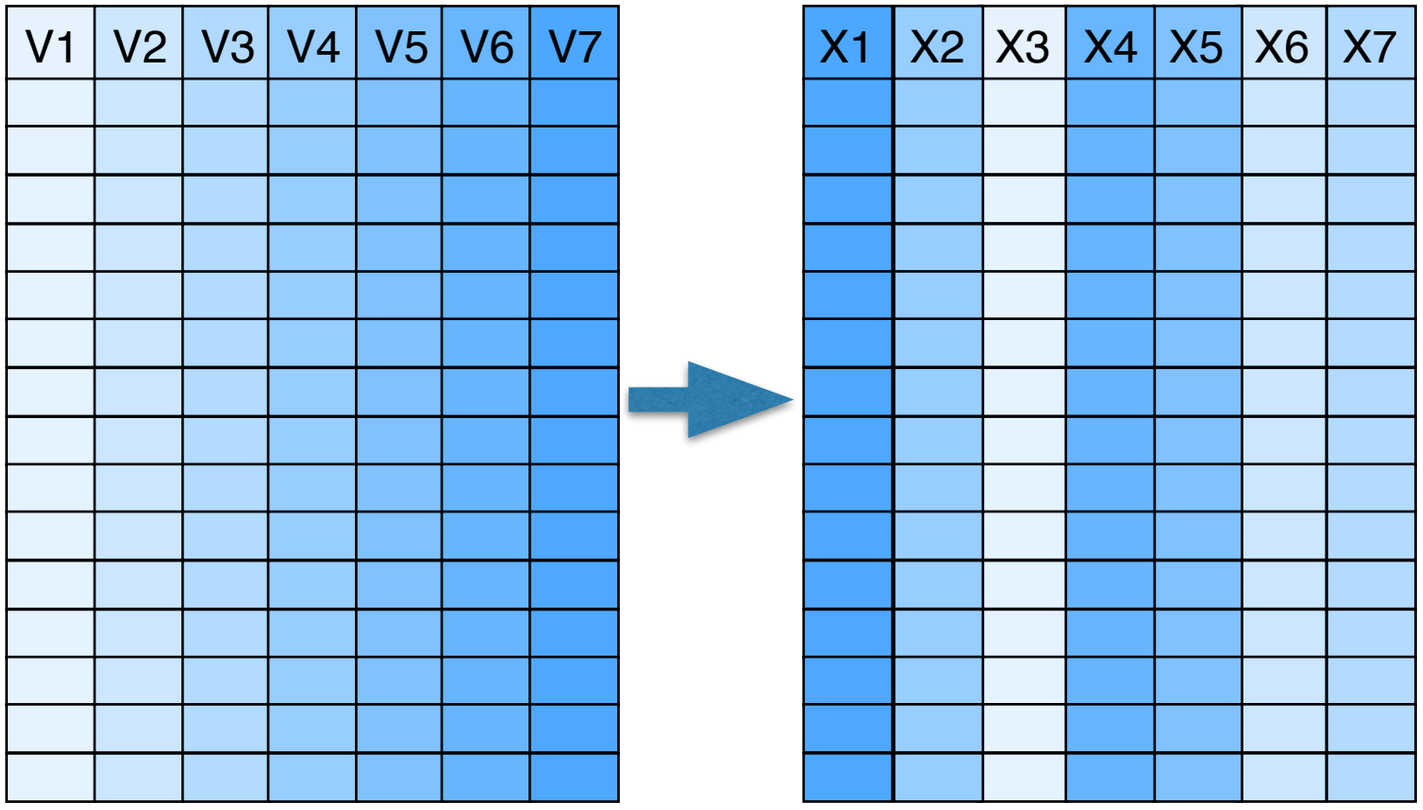
Solution

- 3 -fold cross validation scheme
- repeated 30 times and averaged



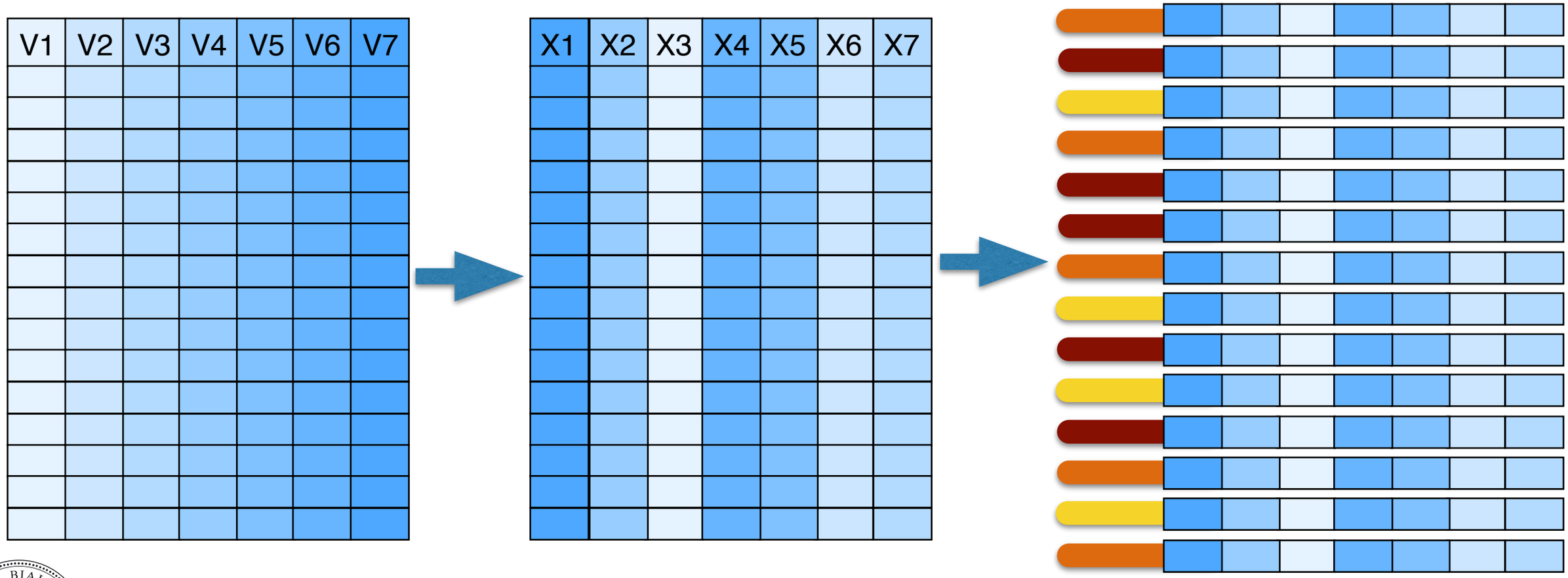
Solution

- Overfitting is still possible via feature selection performed on the entire data set
- We may force feature selection by scrambling feature names and order



Solution

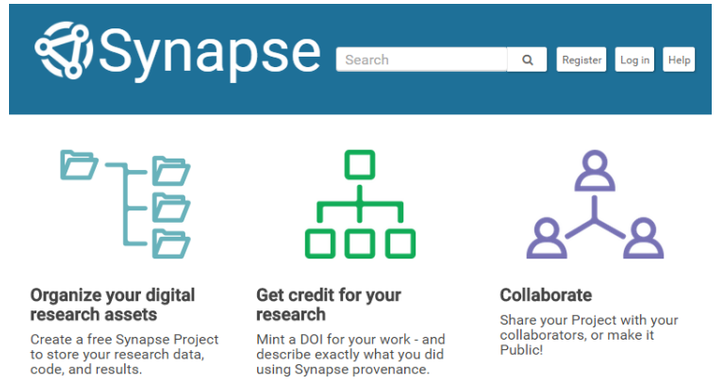
1. Anonimize variables
2. Randomly split data into three subsets
3. Build three cross-validated models
4. Repeat 30 times



Implementation

- Synapse interface

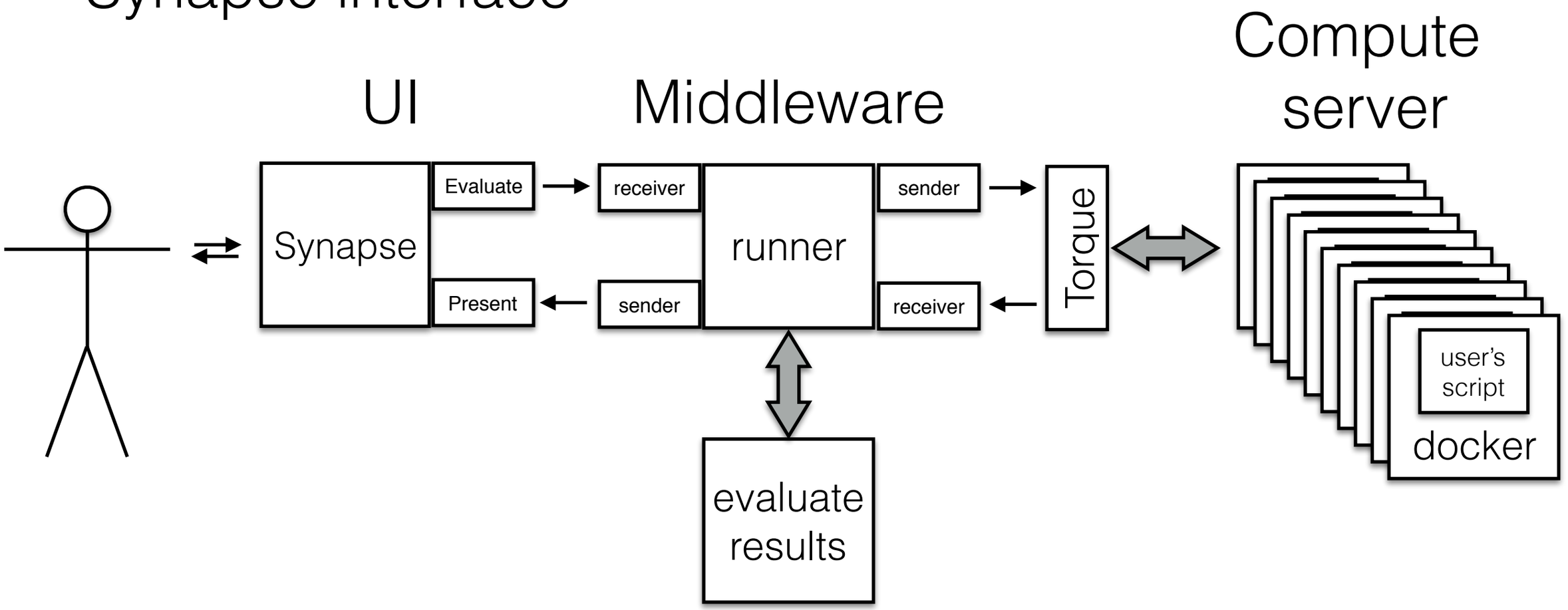
- Middleware based on Redis database



- Computational backend

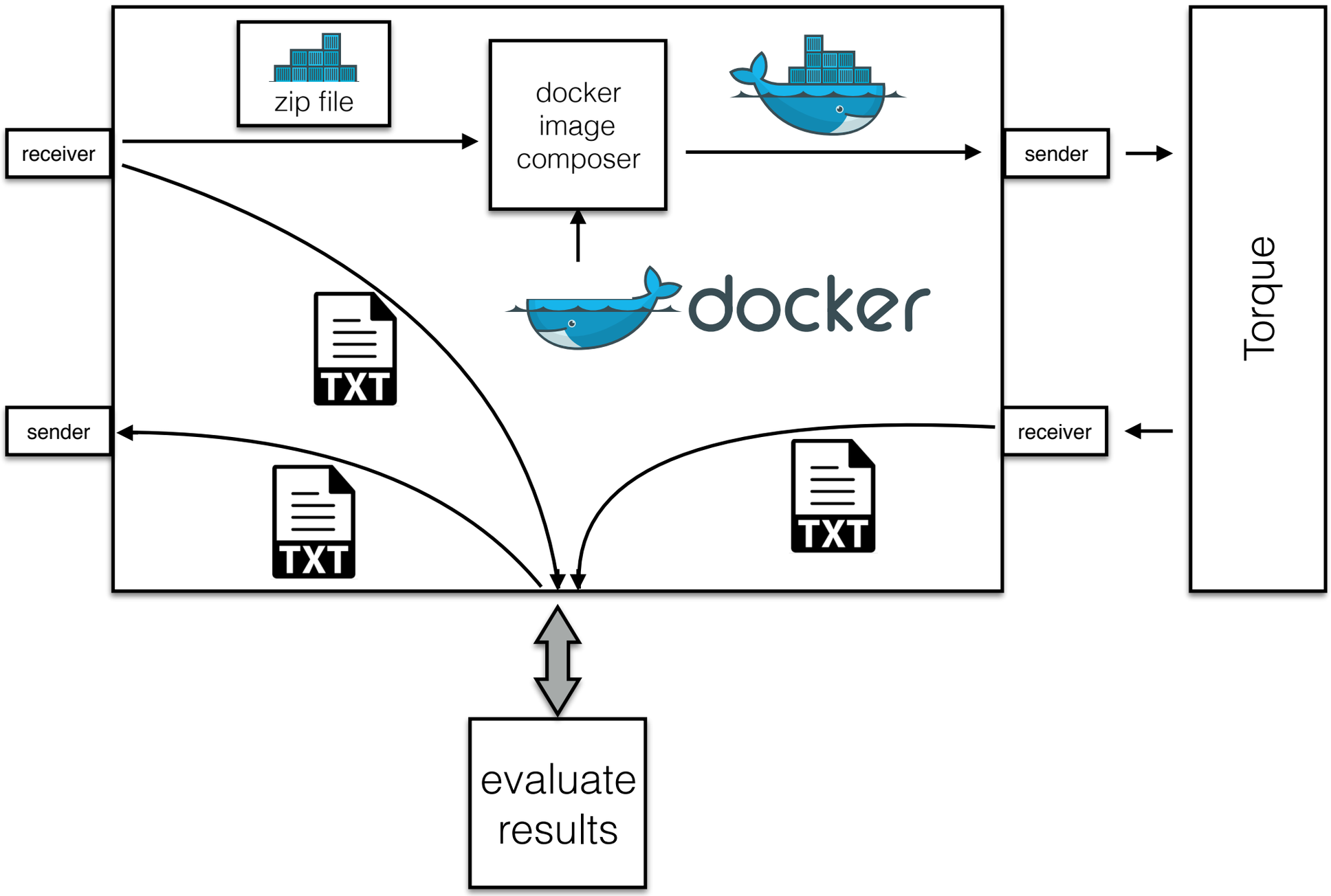
Implementation

- Synapse interface



Implementation

runner internals



Implementation

Transcription Factor Binding Post-challenge Evaluations

Search | Witold Rudnicki (Witold.Rudnicki) | Help

Synapse ID: syn6829411 **Storage Location:** Synapse Storage

Share | Annotations | Tools

Wiki | Files | Tables | Challenge | Discussion | Docker beta

This project is a continuation of DREAM5 Transcription-Factor, DNA Motif Challenge. The full description of that finished challenge is here.

Introduction

We provide two separate evaluations (queues):

- text results evaluation (like in the original challenge) where we expect a plain text file with one number per line which stands for the predicted response
- program results evaluation where we score results of submitted programs which we run on our machines in a Docker image

The submission button is at the bottom of the page. You have to register and join the team first to be able to use it. In case of problems, please see the FAQ on the left.

Text results evaluation

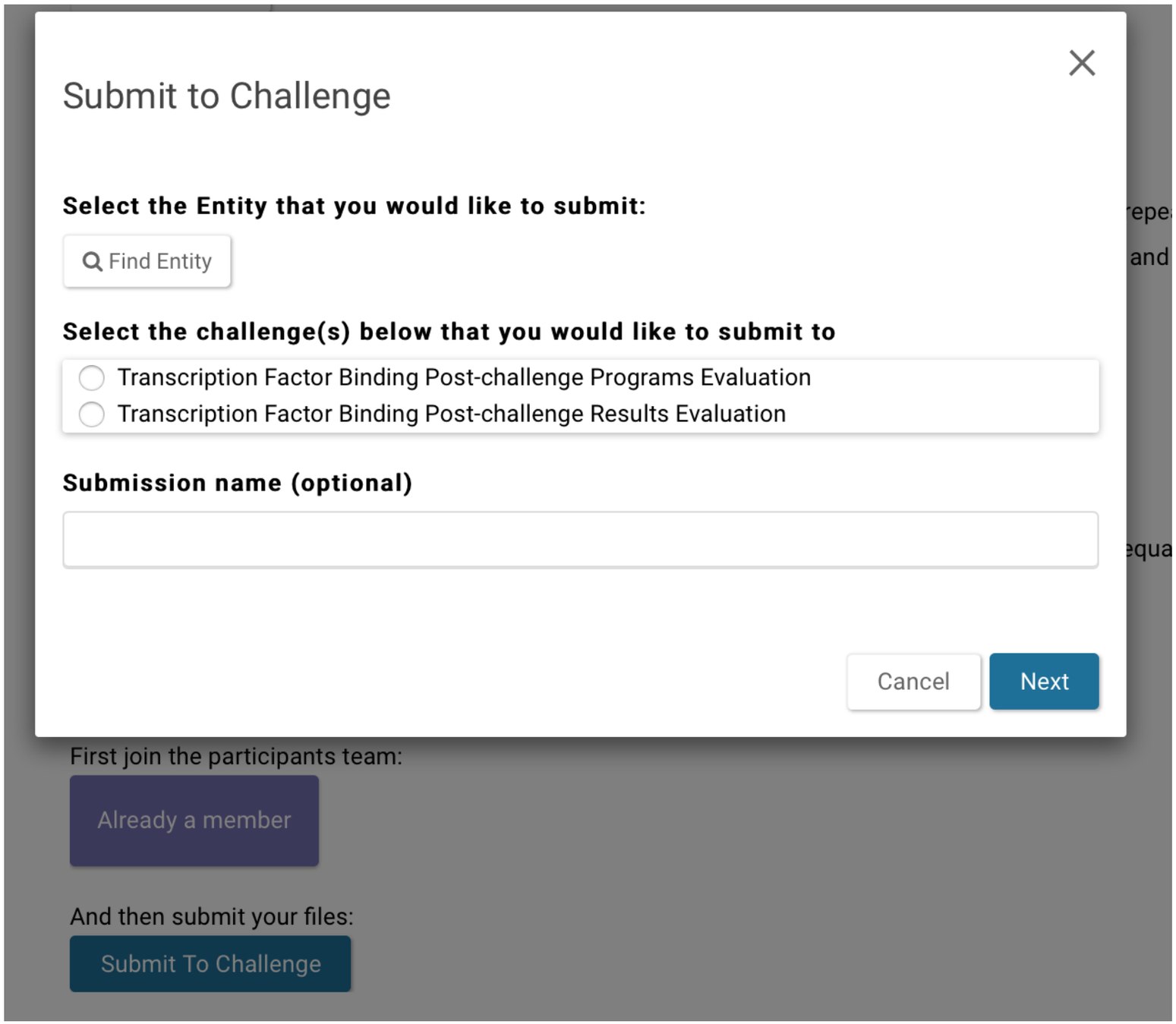
Please see the **DREAM5_Answers.txt** file for the expected format. It includes solely the responses (2 668 248 floating point numbers) in the order given by the **DREAM5_GoldStandard_probes.txt** file.
The predictor should be trained on the **DREAM5_PBM_Data_Needed_For_Predictions.txt** file but it can be tweaked on the **DREAM5_PBM_Data_TrainingSet.zip** file.

Go to files

Program results evaluation

Implementation

- Synapse interface



Implementation

Transcription Factor Binding Post-challenge Evaluations

Search | Witold Rudnicki (Witold.Rudnicki) | Help

Synapse ID: syn6829411 **Storage Location:** Synapse Storage

Share | Annotations | Tools

Wiki | Files | Tables | Challenge | Discussion | Docker beta

Edit Order <<

Transcription Factor Bin... » Scores of text results

Scores of text results

Submission ID	Entity ID	Pearson correlation	Spearman correlation	MSE	RMSE	AUC
7252743	syn7252736	1.0	1.0	0.0	0.0	1.0
7349516	syn7349511	1.0	1.0	0.0	0.0	1.0
7349517	syn7349510	0.999999999999157	0.999999999971669	0.250167222249427	0.500167194295495	1.0
7349518	syn7349509	0.999999999999156	0.999999999971657	100.000115145154	10.000005757256	1.0
7349519	syn7349508	0.999999999999165	0.999999999971787	10000.0007858864	100.000003929432	1.0

Created by Radosław Piliszek (Radoslaw.Piliszek) on Monday, October 3, 2016 8:20 AM

Modified by Radosław Piliszek (Radoslaw.Piliszek) on Monday, October 3, 2016 8:22 AM

Wiki History

Opportunities

1. Computing power is cheap, data transfer is and storage is expensive

2. Computer power is ubiquitous, sensitive data needs protection

- New paradigm for running challenges possible - no need to transfer large and/or sensitive datasets.

- Run models where the data is, without the need to see actual data itself.

Opportunities

- Current approach
  - Roughly half of the data is used for building the models
  - Results of the evaluation depend on the particular split of the data.

- New approach:
  - build cross-validated robust models (ensemble of repeats)

People

Radek Piliszek
University of Białystok

Wojtek Lesiński
University of Białystok

and

Julio Saez Rodriguez,
Technische Hochschule
Aachen