



# From online education to gig science and augmented intelligence

Oct, 2019, AISIS

Andrey Ustyuzhanin

NRU HSE

YSDA

ICL

# Co-research process for applied data analysis challenges.

Failures, values, instrumentation and open calls.

Oct, 2019, AISIS

Andrey Ustyuzhanin

NRU HSE

YSDA

ICL

# Trends and Goals

- › Online teaching, blended learning
- › Interdisciplinarity (AI challenges for Particle Physics: [nature.com/articles/s41586-018-0361-2](https://www.nature.com/articles/s41586-018-0361-2))
- › AI hype attracts bright students from other science domains

Goal: How can we teach students practical data science  
AND advance domain science at the same time at no cost?

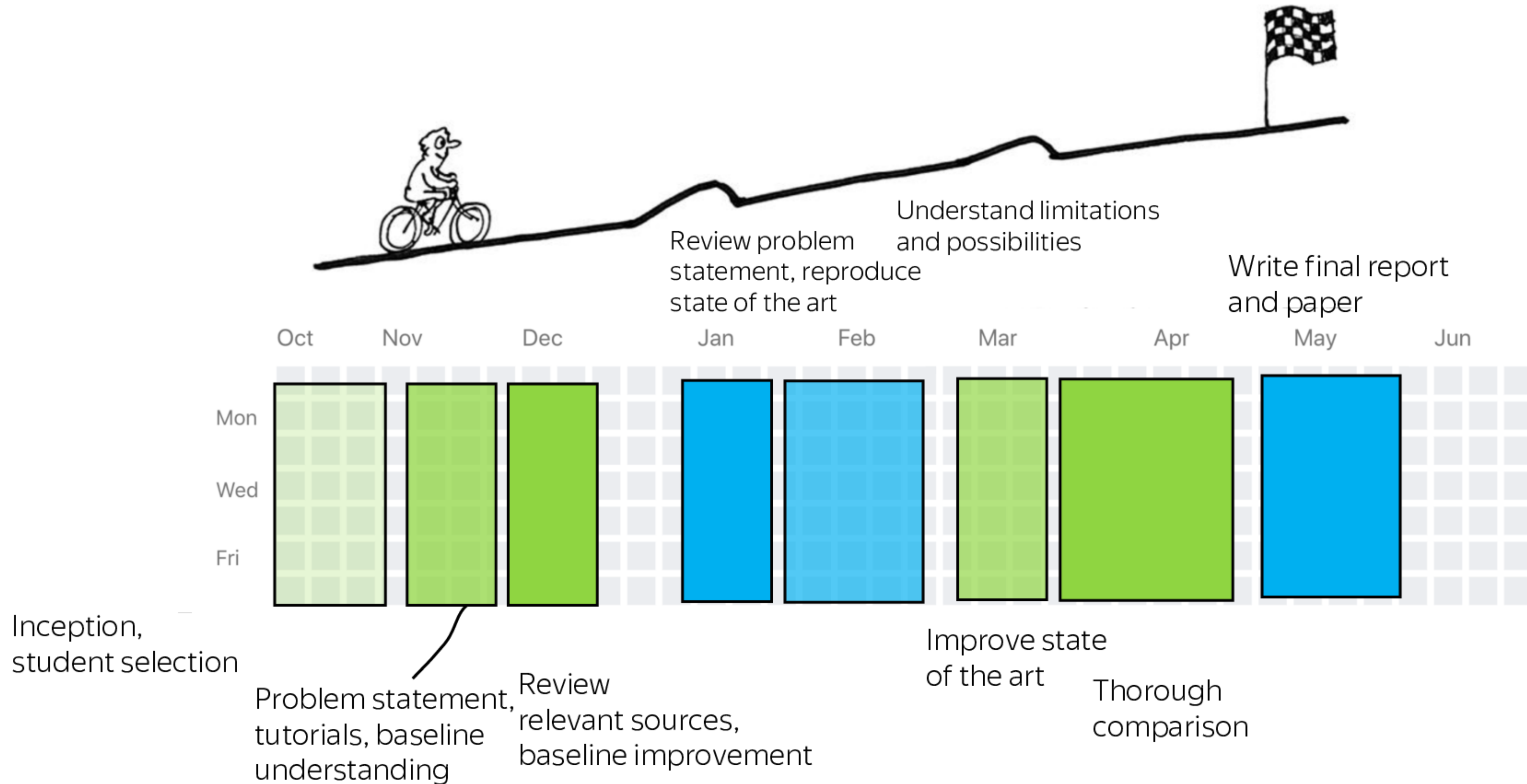
improve junior student involvement  
improve efficiency of research (predictability of results)  
reduce cost per student



# Lab Project Highlights 2018/19

1. OPERA shower reconstruction and generation (ACAT'19)
2. Generation of LHCb calorimetry images (ACAT'19)
3. FERMI telescope image denoising (JINST paper draft)
4. JUNO event reconstruction (JINST paper draft)
5. Transfer learning for fast Monte-Carlo event generation (ACAT'19)
6. Cityscape – autonomous vehicle camera images reconstruction

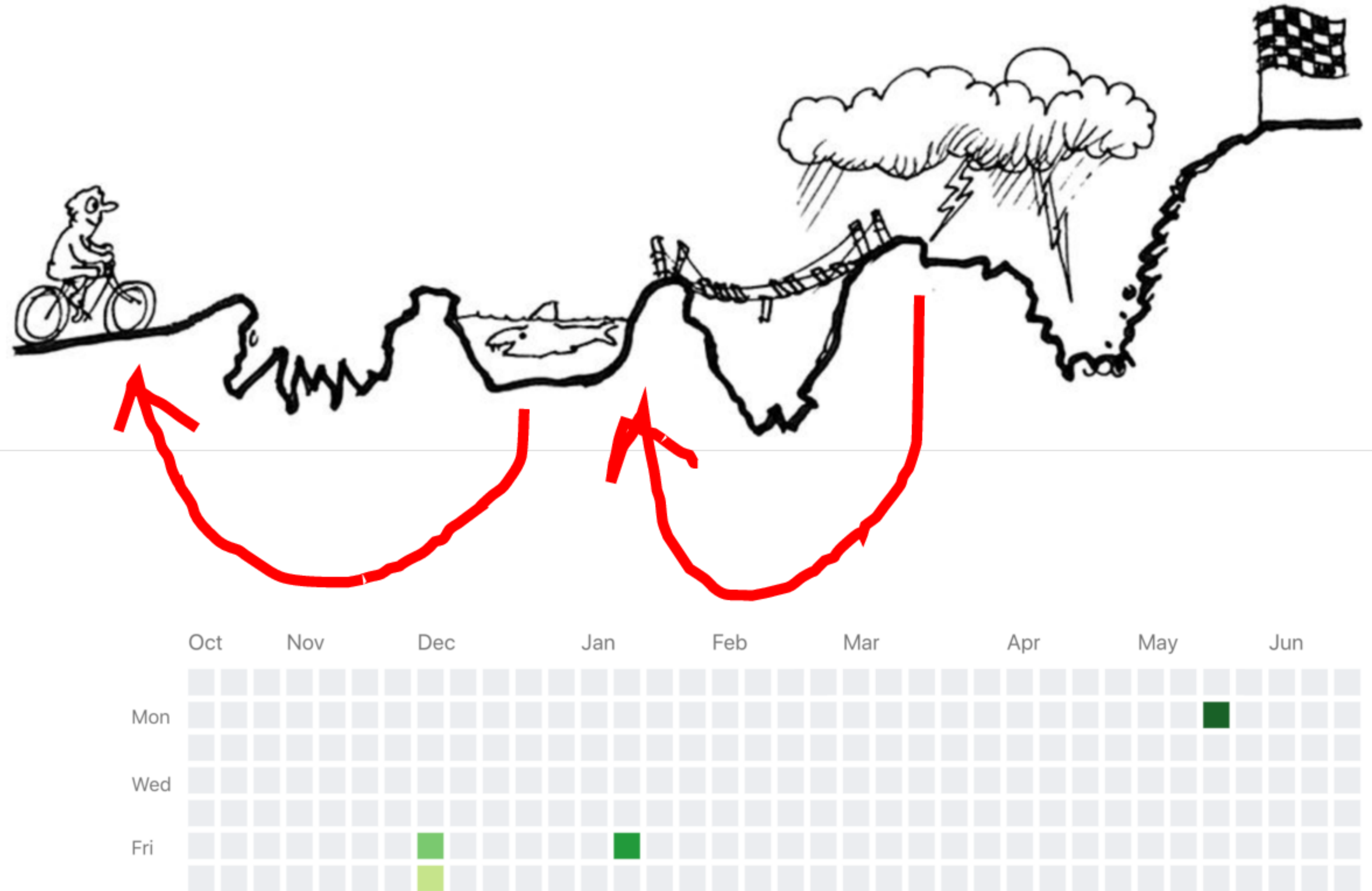
# Research timeline - ideal



# Research timeline - Real

## Main issues

- Unknown unknowns (data, design, others' work, computations)
- Underestimated risks
- Lack of proper communication
- Lack of supervisor time/attention
- Bad luck



# Can it be improved?

Traditional approach: 1-on-1 supervision

proven by ages, but  
doesn't scale well

Challenge-based approach

scales better, but  
a bit away from science  
difficult to explain, students consider it as homework they skip until deadline

# Machine Learning + Science examples

The image is a screenshot of the Kaggle website's competition page for the TrackML Particle Tracking Challenge. At the top, the Kaggle logo is on the left, followed by a search bar and navigation links for Competitions, Datasets, Kernels, Discussion, Learn, and a Sign In button. The main banner features a dark blue background with a complex network of yellow and red lines representing particle tracks. Text on the banner includes 'Featured Prediction Competition', 'TrackML Particle Tracking Challenge', 'High Energy Physics particle tracking in CERN detectors', '\$25,000 Prize Money', and 'CERN · 516 teams · a month to go (a month to go until merger deadline)'. Below the banner are navigation tabs: Overview (selected), Data, Kernels, Discussion, Leaderboard, and Rules. The 'Overview' section is expanded, showing a table of contents with links to Description, Evaluation, Prizes, About The Sponsors, and Timeline. The 'Description' text explains that scientists at CERN are colliding protons to recreate mini big bangs and observing these collisions with intricate silicon detectors. It notes that analyzing the enormous amounts of data produced is becoming an overwhelming challenge. To the right of the text is a 3D visualization of a particle detector, showing a series of stacked cylindrical layers with a green and orange color scheme.

kaggle Search kaggle Q Competitions Datasets Kernels Discussion Learn ... Sign In

Featured Prediction Competition

**TrackML Particle Tracking Challenge**

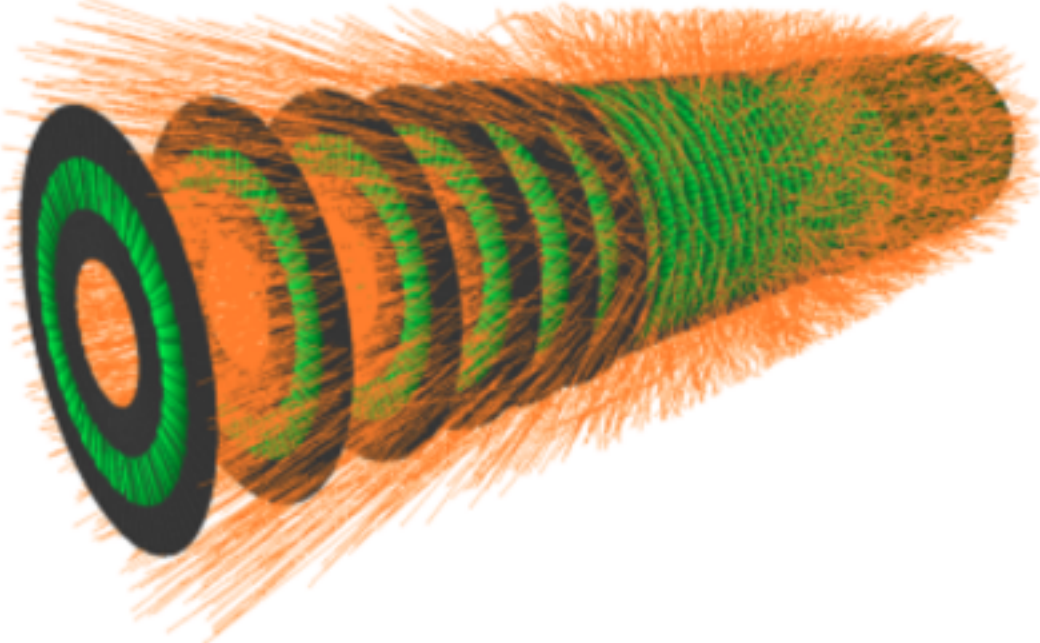
High Energy Physics particle tracking in CERN detectors

**\$25,000**  
Prize Money

CERN · 516 teams · a month to go (a month to go until merger deadline)

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#)

Overview

<b>Description</b>	To explore what our universe is made of, scientists at CERN are colliding protons, essentially recreating mini big bangs, and meticulously observing these collisions with intricate silicon detectors.	
<b>Evaluation</b>		
<b>Prizes</b>		
<b>About The Sponsors</b>		
<b>Timeline</b>	While orchestrating the collisions and observations is already a massive scientific accomplishment, analyzing the enormous amounts of data produced from the experiments is becoming an overwhelming challenge.	



# “Kaggle” style shortcomings

Domain specificity are difficult to check automatically:

- › Non-trivial metrics
- › Constraints

Participation Incentive

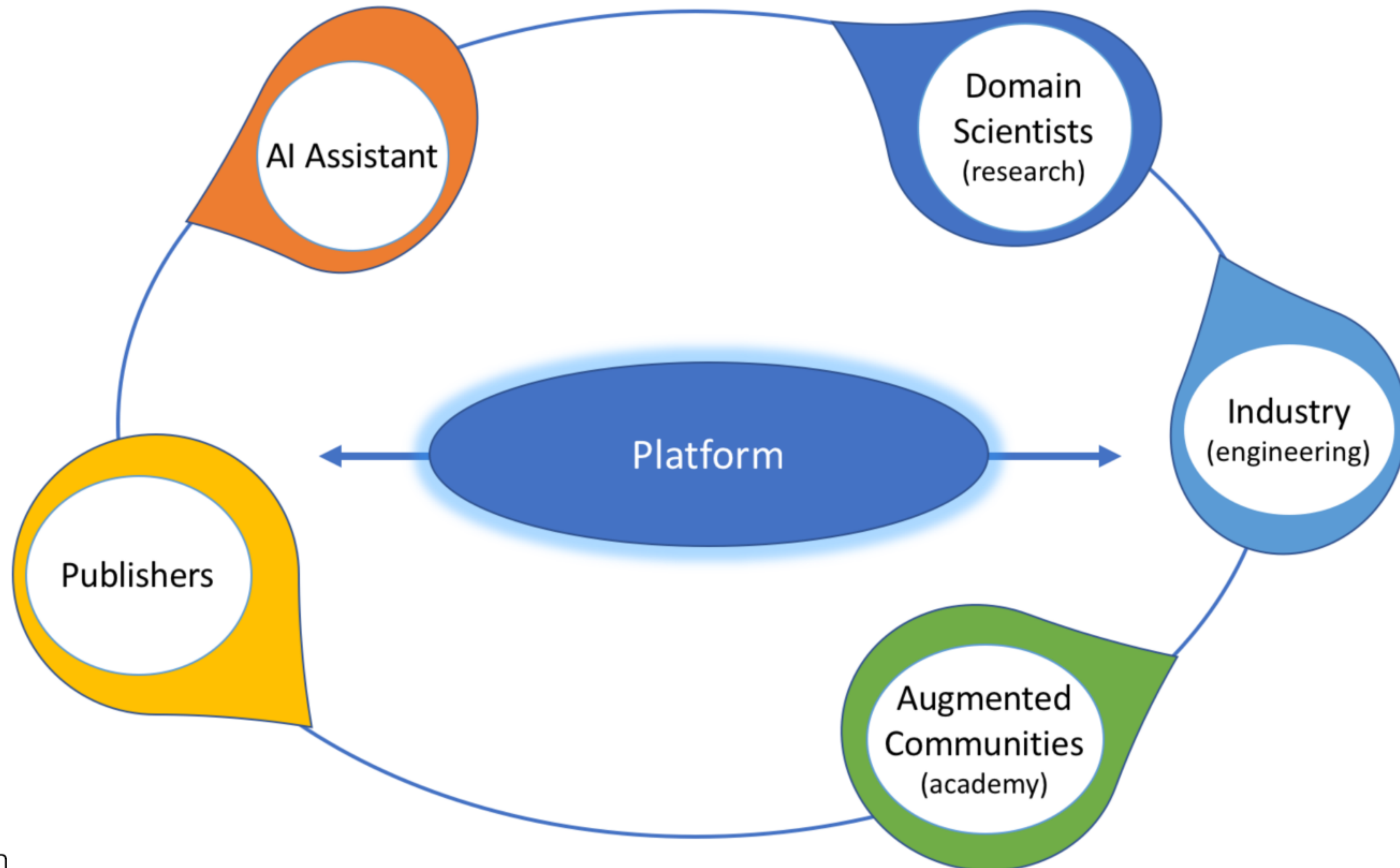
ML-practitioners sometimes consider it as real science

Beyond iid assumption (the test set differs from the training set)

Maybe we need to go deeper? Explain and support **values**, explain and run the **process**, have **infrastructure** in place to support it.



# Co-research holistic view



# Co-research values

Transdisciplinary approach	#transdiscip
Data quality	#data
Reproducible results	#reproducible
Scientific ethics	#ethics
Innovative efficient research	#innovation
Amazing results	#coolresearch
Social good	#socialgood
Strong AI-scientific community	#community
(International) collaboration	#teamwork
Science is not done until communicated	#PR

# Project big lifecycle (co-research)

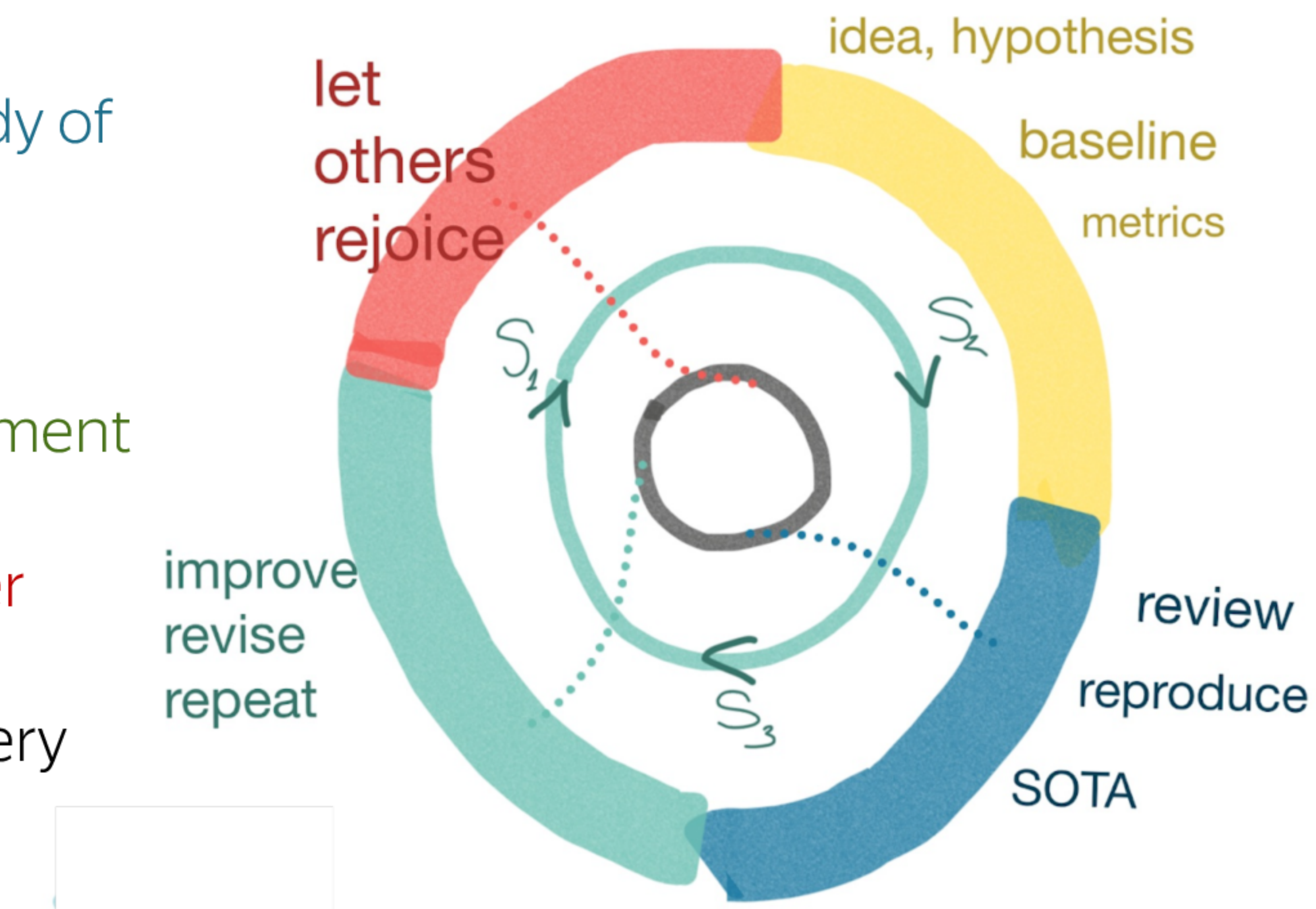
## Inception stage

Learning: doing papers review, study of «state-of-the-art» methods

individual solution: iterative search, model understanding and improvement

Writing final report, slides and paper

NB: iterations may be nested on every stage.



# Lifecycles

## Weekly lifecycle

- › Iterate through current tasks,
- › report results at weekly meeting / trello
- › Choose tasks for the next week
- › Artifacts: git commits, trello cards, models

## Bi-monthly lifecycle (stages)

- › Team report to wider audience (lab meeting)
- › Artifacts: slides, trained models, re-evaluated datasets and metric scripts

# Process Roles

## Domain experts

- › Problem statement
- › Dataset, explanation of dataset, initial metric, state of the art explanation

## ML experts

- › Supervise mentors, students
- › Refine metric in ML terms, methods landscape

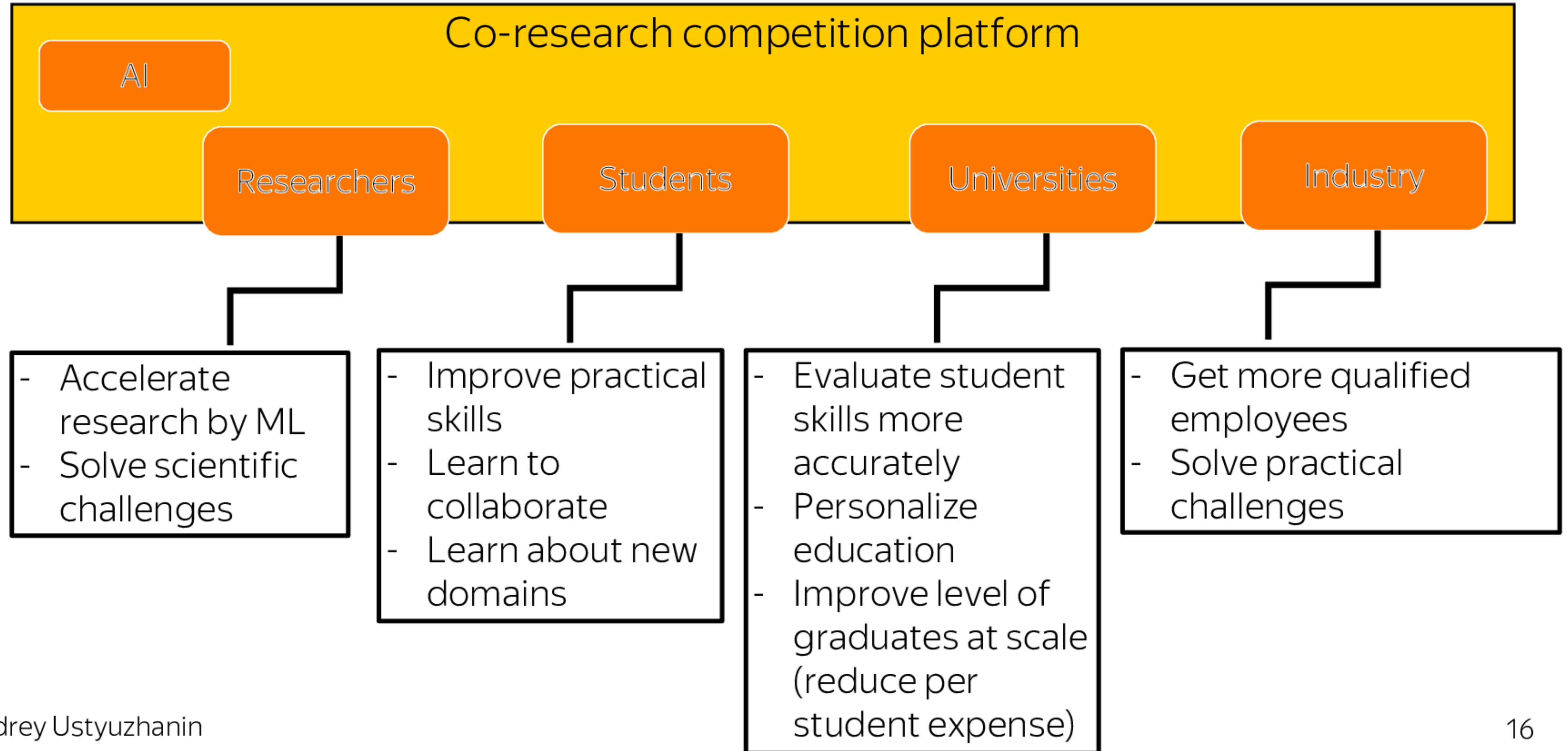
## Mentors

- › Implement baseline, setup infrastructure
- › Facilitate weekly/monthly lifecycles with students

## Students

- › Understand the problem
- › Understand and follow the process (implement, write, present)
- › Understand the final solution
- › Help and learn from others

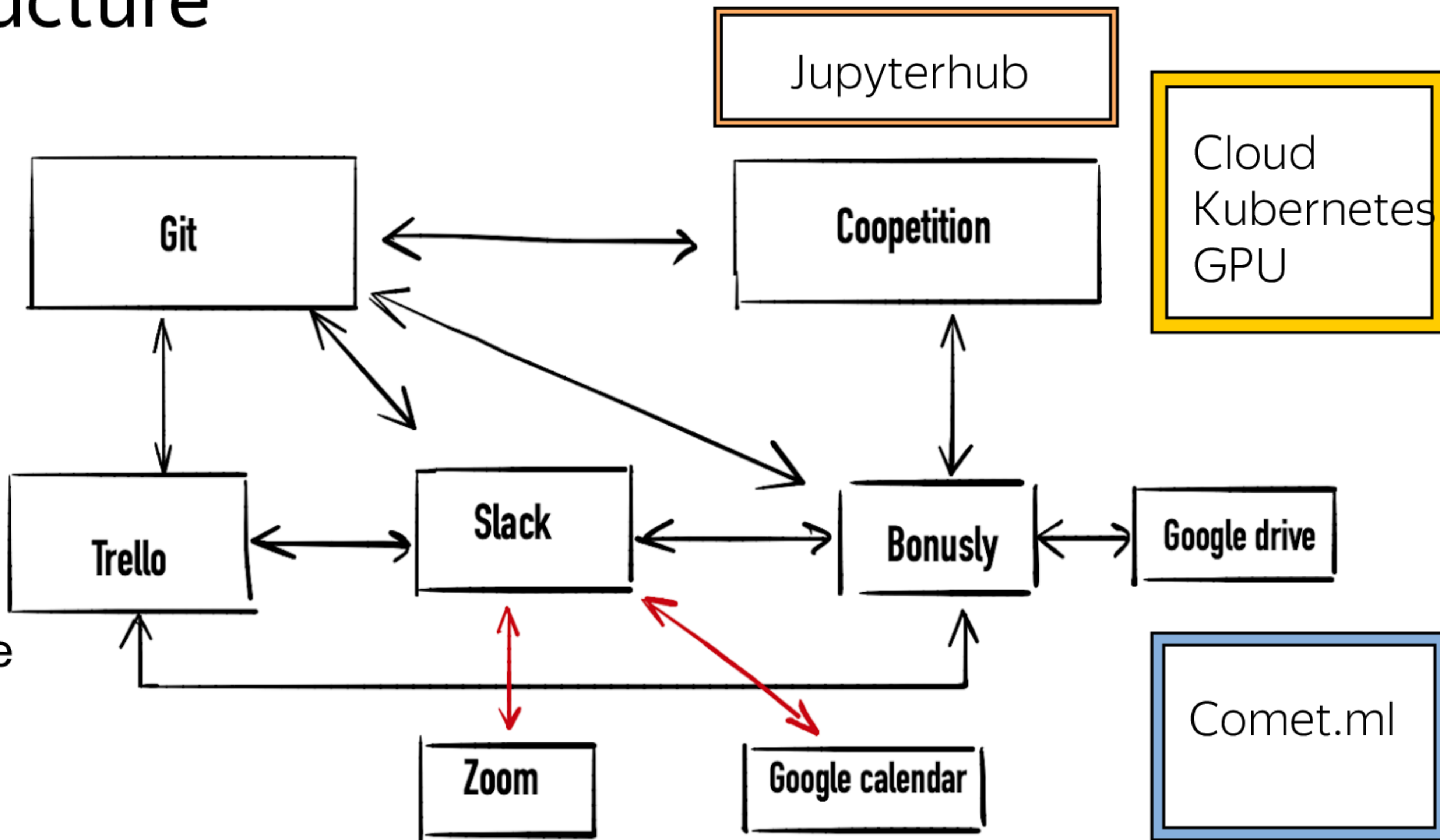
# Co-Research Platform - Coopetition





# Infrastructure

1. Github
2. Trello
3. Coopetition
4. Bonus.ly
5. Slack
6. JupyterHub
7. Cloud/Azure
8. Comet.ml
9. ....



# Trello – project roadmap

Project roadmap (links to info, who's doing what, what's going on now):

A board with the current tasks or «cards»

Card may contain checklists, comments, deadlines, etc.

Cards are grouped into columns

The screenshot shows a Trello board for a project named "lambda-summer-2019-galaxymass". The board is organized into four columns: "backlog", "planned", "doing", and "test".

- backlog:** Contains five cards with tasks in Russian, such as "выгрузить данные из оптического каталога по всем скоплениям галактик" and "предложить алгоритм объединения оптического каталога с рентгеновскими".
- planned:** Currently empty, with a "Добавить карточку" button.
- doing:** Contains one card with a scatter plot titled "g\_list\_s, iGrid = 149473". The plot shows "z\_group" on the y-axis (ranging from -0.015 to 0.015) and "distance from center in degrees" on the x-axis (ranging from 0.0 to 2.5). Below the plot is a detailed description of the plot and a list of URLs for data sources.
- test:** Currently empty, with a "Добавить карточку" button.

The interface includes a top navigation bar with "Доски", a search icon, the Trello logo, and user avatars. The board title "lambda-summer-2019-galaxymass" is prominently displayed at the top of the workspace.

# Coopetition – submission checker

- Tracks user submissions from github
- Verifies the submission code
- Give recommendations \*)

The screenshot shows the 'MLHEP2019 0 stage' competition page. At the top, there is a navigation bar with 'COOPETITION', 'COOPETITIONS', 'REWARDS', and 'FAQ' on the left, and 'MY COOPETITIONS' and a login icon on the right. Below the navigation bar is a breadcrumb trail: 'Home / Coopetitions / MLHEP2019 0 stage'. The main heading is 'MLHEP2019 0 stage', followed by the text 'You are registered in the competition'. Below this is a sub-navigation bar with 'OVERVIEW', 'LEADERBOARD', 'MY SUBMISSIONS', and 'STATISTICS'. The 'LEADERBOARD' tab is active. Underneath, there is a 'Development' filter button. The main content is a table with the following columns: RANK, NAME, DATE, PREDICTION SCORE, VIEW, RATING, TECHNOLOGIES, and a heart icon for likes. The table lists four entries:

RANK	NAME	DATE	PREDICTION SCORE	VIEW	RATING	TECHNOLOGIES	LIKES
1	fcostanza	2019-06-27T23:34:56.168Z	6184573.09	0.00	View	NUMPY	0
2	mbieker	2019-06-28T15:13:09.672Z	432971.74	0.00	View	NUMPY	0
3	SWuchterl	2019-06-29T15:41:11.007Z	4845.71	0.00	View	NUMPY	0
4	rishabh	2019-06-27T08:23:09.858Z	65.00	0.00	View	NUMPY	0

# Bonus.ly – employee engagement

The image displays the Bonus.ly interface, which is used for employee engagement and rewards. It is divided into several sections:

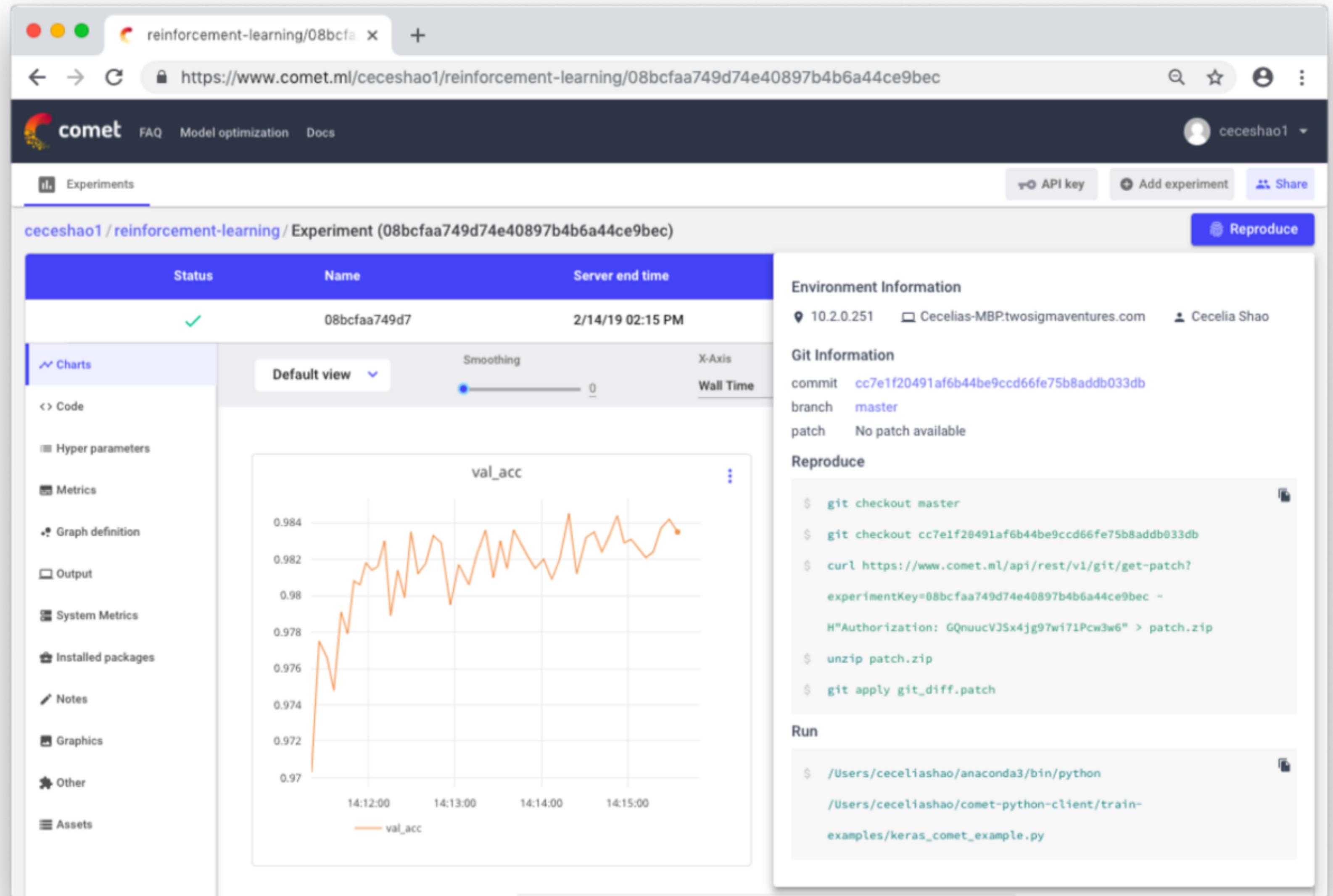
- Claim an Award:** A yellow header section with three award categories:
  - #Interscience:** холистический (трансдисциплинарный) подход к научным знаниям и методам работы с данными
  - Reproducible:** воспроизводимые результаты исследования
  - Best\_Score:** Best Codab score
- Point Redemption:** A yellow section with the text "You have 10 points to give away". It includes input fields for "+ Amount", "@ Recipient", and "# Hashtag". Below these is a text input field containing: "+5 @ktrofimova for helping me launch a marketing campaign so that we can generate new business #teamwork". To the right, a red-bordered box highlights the text "You have 40 points to redeem" and a "Pick a reward" button.
- Recommended for you:** A section showing a list of rewards with their point values:
  - pizza:** 100 points (Only 4 left)
  - Рюкзак:** 650 points
  - Чехол на ноутбук:** 550 points
  - Кружка:** 250 points
- Navigation and Admin:** A sidebar on the left contains navigation links (SEARCH, HOME, REWARDS, AWARDS, ANALYTICS, INTEGRATIONS) and an ADMIN section (COMPANY, USERS, BONUSES, REWARDS). A notification states "You have 22 days left in your trial" with an "Add payment" button.
- Header and Footer:** The top left shows the "LAMBDA" logo. The bottom right shows "Russian Federation" and "My Rewards".

# Comet.ml – model quality monitor

Tracks of model training progress

Keeps track of dependence between params, data and results

Presents results in nice plots



# Process - infrastructure matching

## Preparation

- › All students have access to small Jupyter server with data and limited GPU resources
- › Student forks baseline git repository

## Iteration

- › Student clones his/her repository and plays with it on Jupyter server or locally on data subsample until he/she wants to test it on big dataset
- › Student commits/pushes to git and it automatically runs training on Cloud
- › Comet.ml evaluates solution metrics while training (updates constantly)
- › Coopetition evaluates model automatically, updates leaderboard and bonuses
- › Coopetition suggests recommendations (materials, code of others)

# Open challenges towards gig science

- Problem selection adaptive to student level
- Low-code workflow language
- Addressing core ML (more abstract) problems
- Addressing out-of-iid settings (e.g. data has been generated under the same causal assumptions but with different initial settings)
- Student code analysis and improving recommendations
- AI Research assistant (meta-learning agent)

# Our Lab Project Highlights 2019/20

1. X-ray tomography
2. DUNE event reconstruction speed-up by deep neural architectures
3. X-ray Laser simulation and tuning
4. Estimation of population size by genomes analysis
5. Study of community polarization by search/recommendation output system analysis
6. Leadership style identification by social network profile
7. AI Researcher
8. Natural Language to Machine Learning corpus assembly
9. Improving Earth satellites position prediction by simulation tuning
10. Galaxy cluster mass estimation and registry update
11. Warm and cold dark matter hypothesis comparison

<http://bit.ly/2VNGoll>



# Conclusion

Huge demand for collaboration between ML & Domain scientists

› That also can help students to improve practical skills

Co-research process defines 2-level iterations to help participants to stay tuned:

› relies on explicit collaborative **values** and infrastructure

› implies **reproducibility** and **motivation** that leads to better results and PR

› more **scalable** than peer-to-peer supervision (less cost per student)

› **flexible** to adjust for your team needs

› way towards cross-domain multi-level AI community and gig science

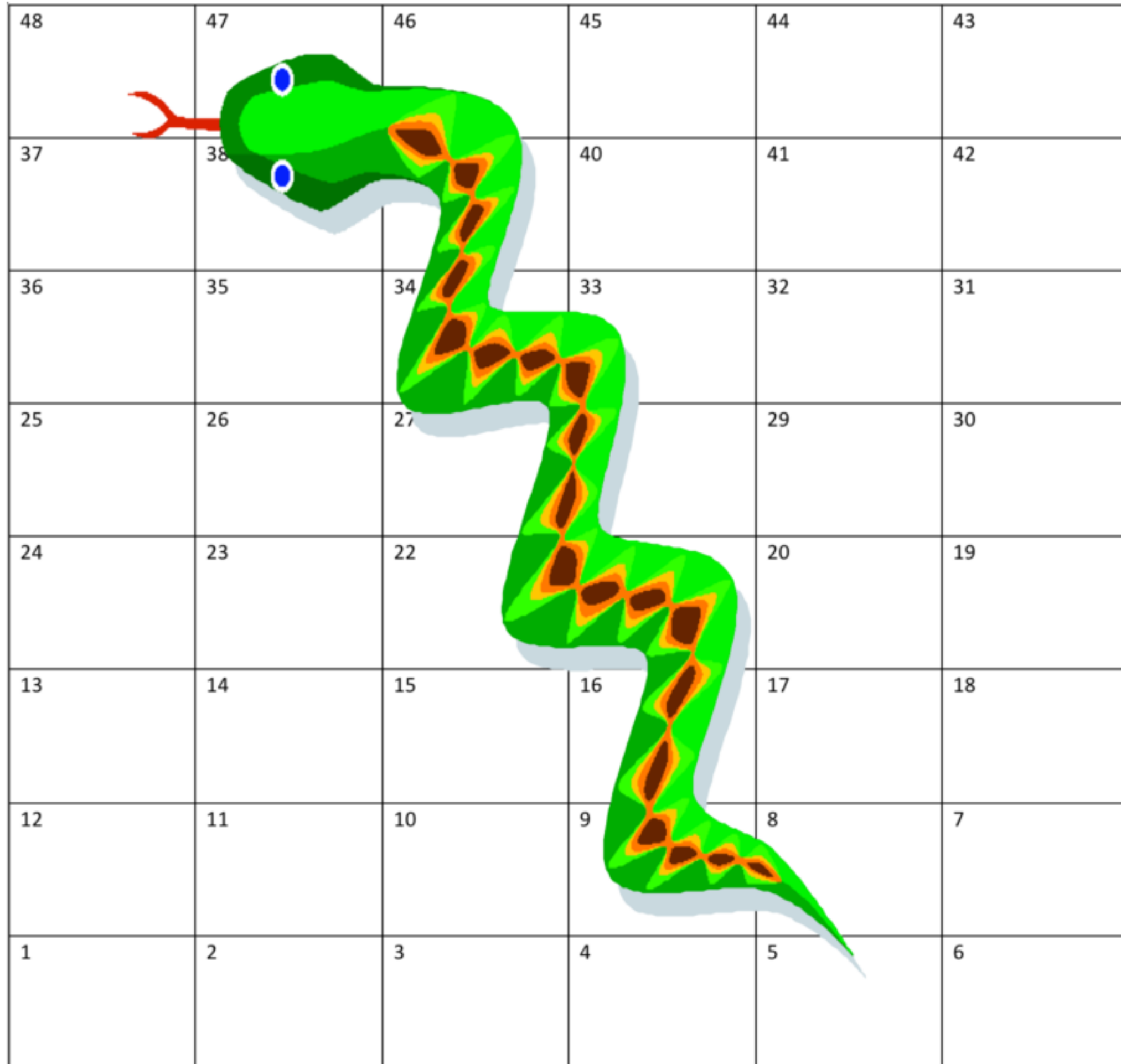
We are open for collaboration with your group and university 🤗

# Backup

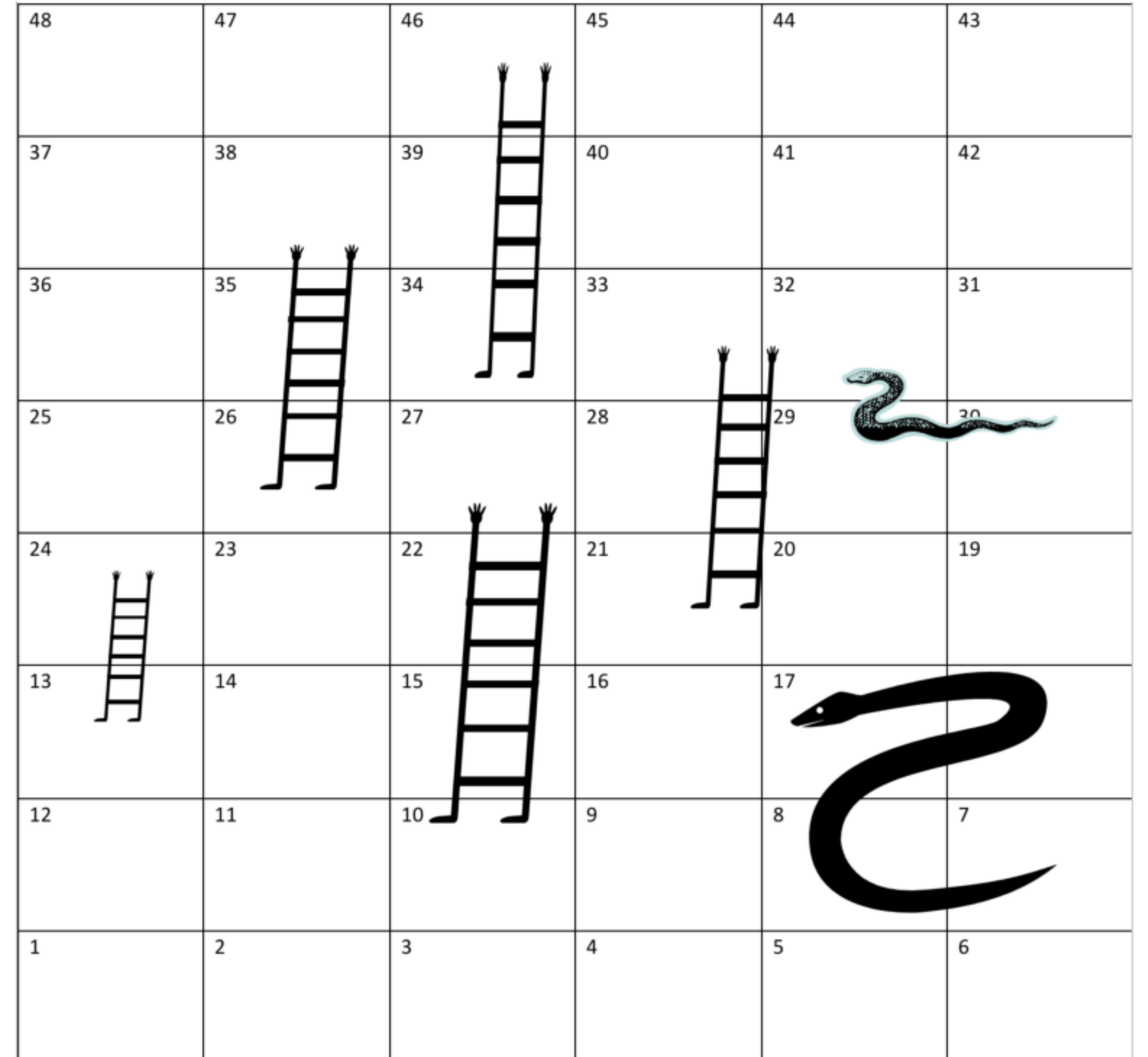


# Research process metaphor

Optimistic view:



In reality:

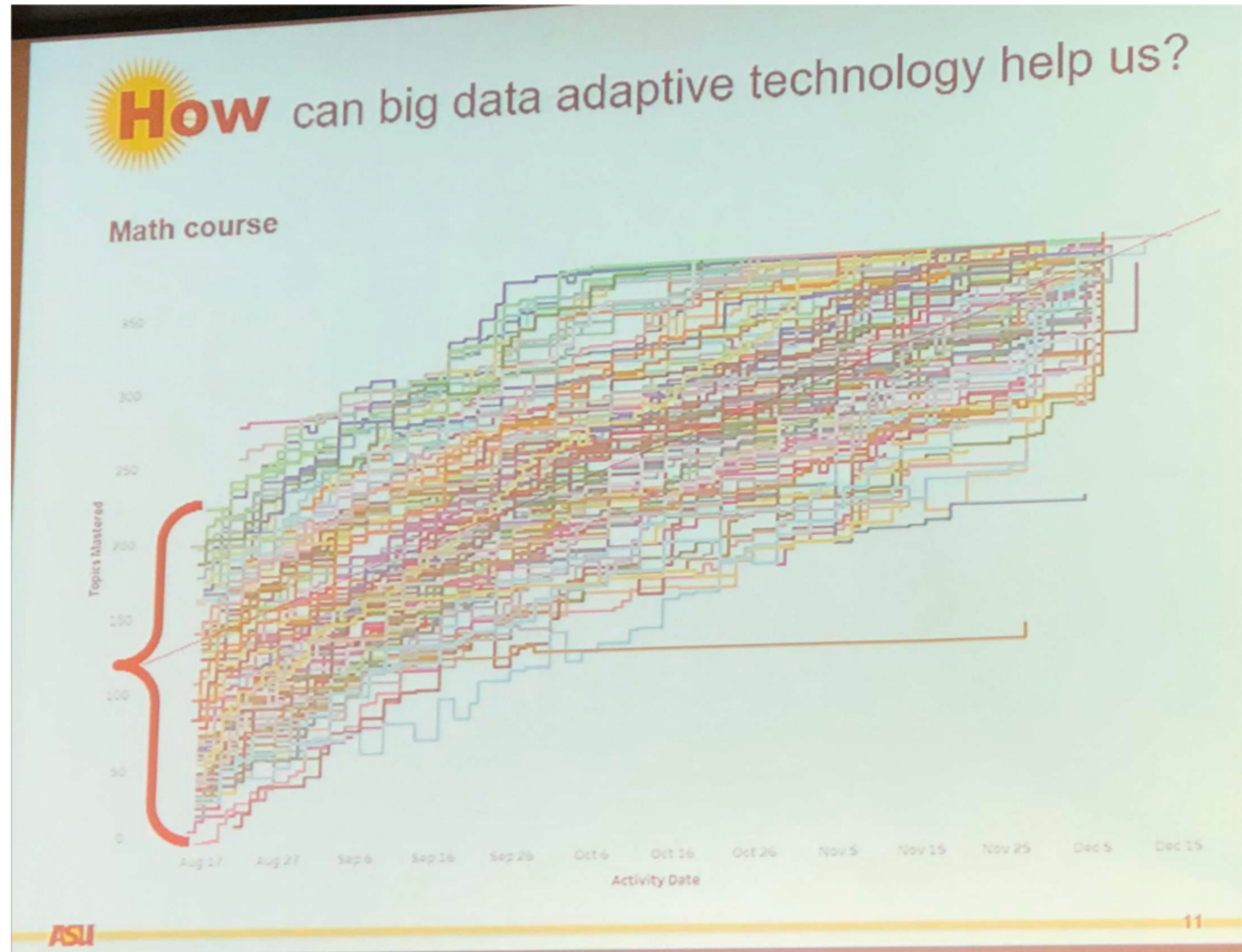


# Machine Learning + Science examples

The image shows two overlapping web pages. The top-left page is the Foldit website, featuring a green header with the 'foldit BETA' logo and the tagline 'Solve Puzzles for Science'. It includes a navigation menu with links for PUZZLES, BLOG, CATEGORIES, FEEDBACK, GROUPS, FORUM, PLAYERS, WIKI, RECIPES, FAQ, CONTESTS, ABOUT, and CREDITS. The main content area displays a 3D protein structure puzzle with a text box that says 'Click to learn how you contribute to science by playing Foldit.' Below this is a 'What's New' section with the headline 'Paper Authorship: Calling all protein designers!' and a short paragraph about a research paper on protein design.

The bottom-right page is the Galaxy Zoo website, which has a dark, space-themed background. The header features the 'GALAXY ZOO.org' logo and a navigation menu with links for Home, The Science, How to Take Part, Galaxy Analysis, Forum, Press, Blog, FAQ, Links, and Contact Us. A user profile for 'Hi starstryder' is visible. The main content area is titled 'Galaxy Analysis' and includes a 'Welcome to Galaxy Zoo's view of the Universe' message. To the right of the main image is a 'Galaxy Ref: 587729387677679742' and a prompt to 'Choose the Galaxy Profile by clicking the buttons below'. There are five buttons: 'CLOCK', 'ANTI', 'EDGE ON / UNCLEAR', 'ELLIPTICAL GALAXY', 'STAR / DON'T KNOW', and 'MERGERS'. A checkbox at the bottom of the main image area is labeled 'Show Grid Overlay on the next Image'.

# Arizona State University teaching math



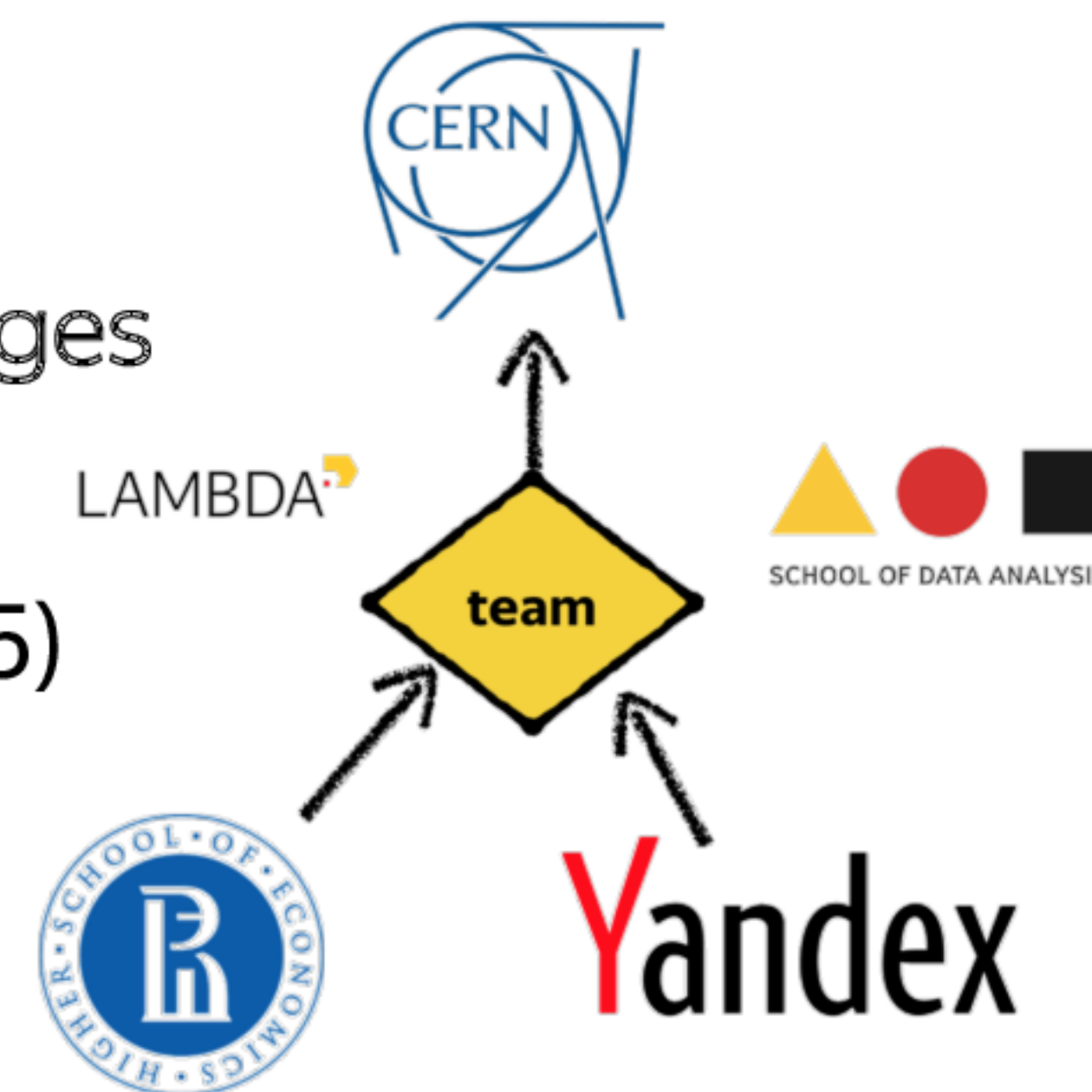
# Quick self-intro



Head of LHCb Yandex School of Data Analysis (YSDA) team  
Head of Laboratory [\(link\)](#) of methods for Big Data Analysis at  
Higher School of Economics (HSE),

- › Applications of Machine Learning to natural science challenges
- › HSE has joined LHCb in 2018!

Co-organizer of Flavours of Physics Kaggle competitions (2015)  
Co-organizer of TrackML challenge (2018)  
Education activities (MLHEP, ML at ICL, Clermont Ferrand,  
URL Barcelona, Coursera)



# Outline

Applied ML research trends

Research process

Competition + Cooperation

- › Strengths
- › Issues: reproducibility, metrics, beyond iid, motivation

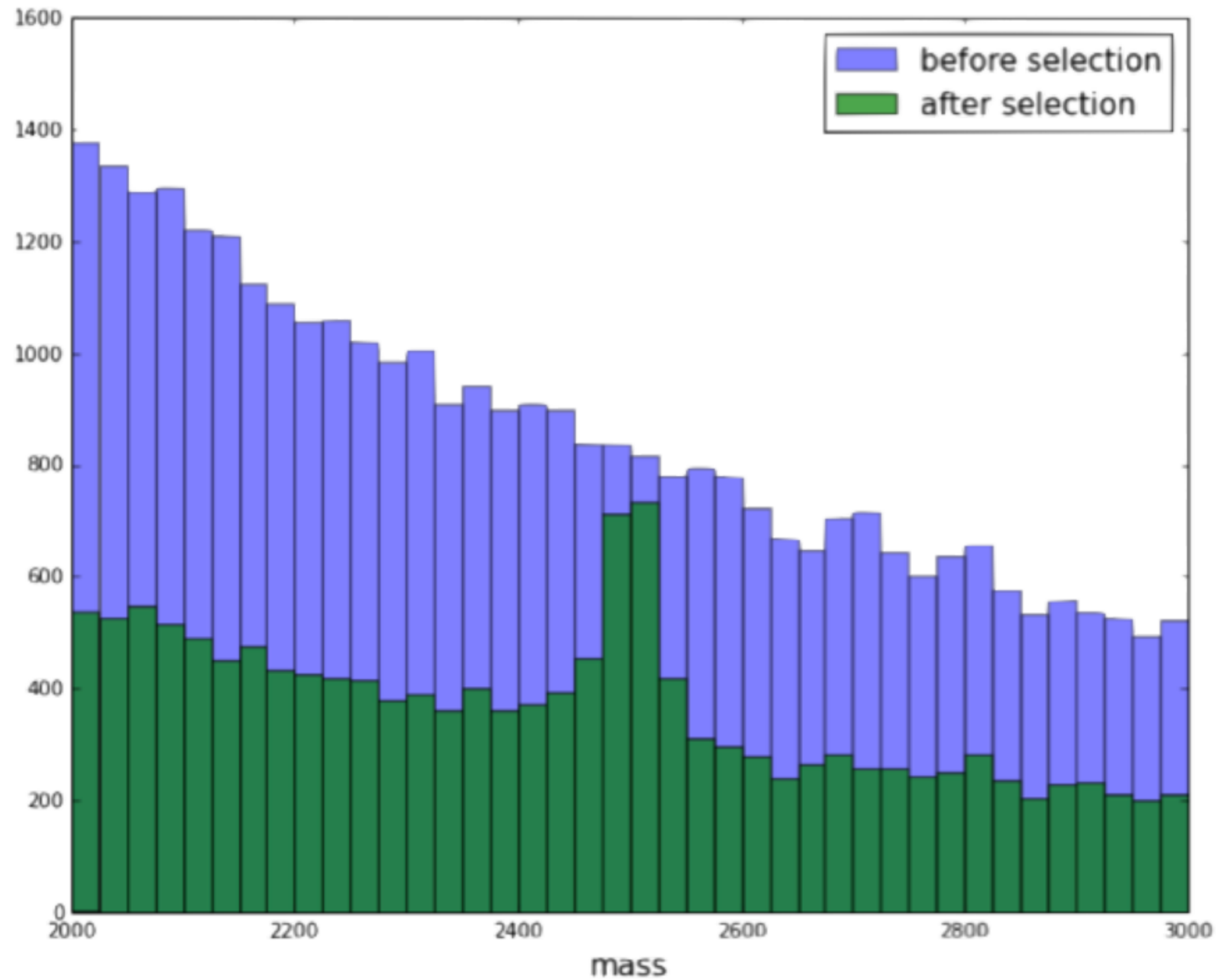
And beyond

- › Values, Lifecycle, Process
- › Roles, Infrastructure

Open issues and call for cooperation

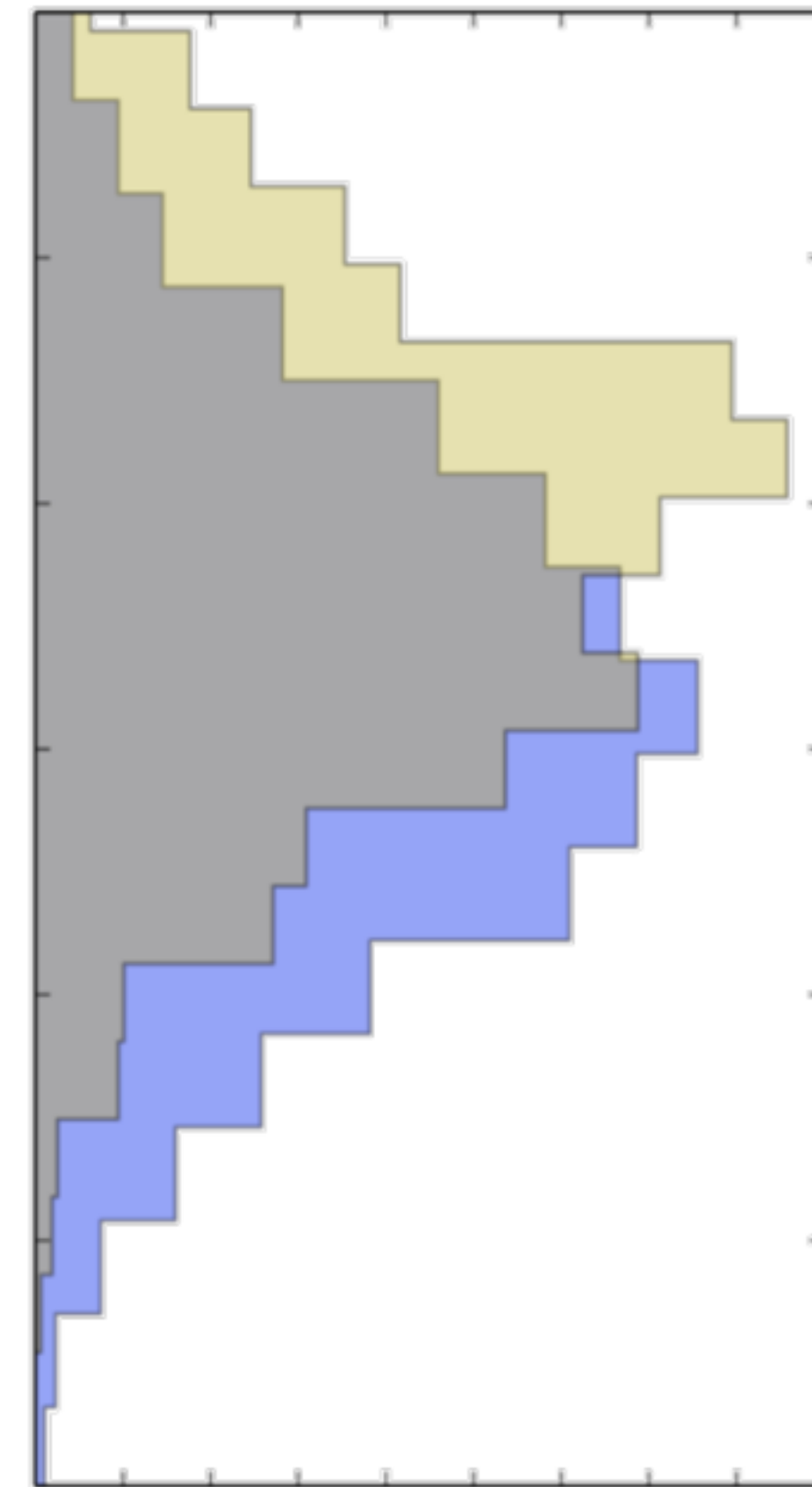
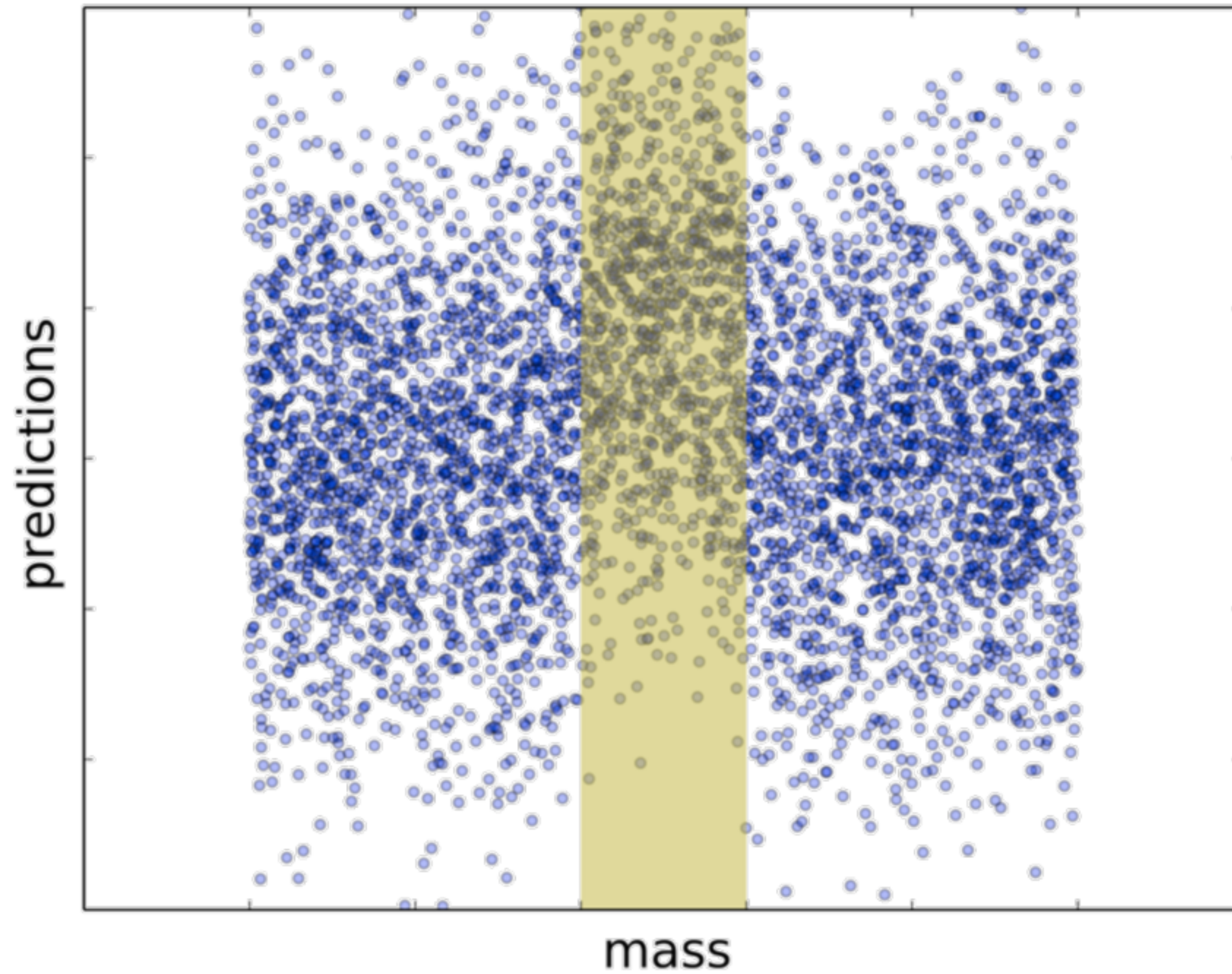
- › Project highlights 2019/20

# Domain specific example





# Non-uniformity check

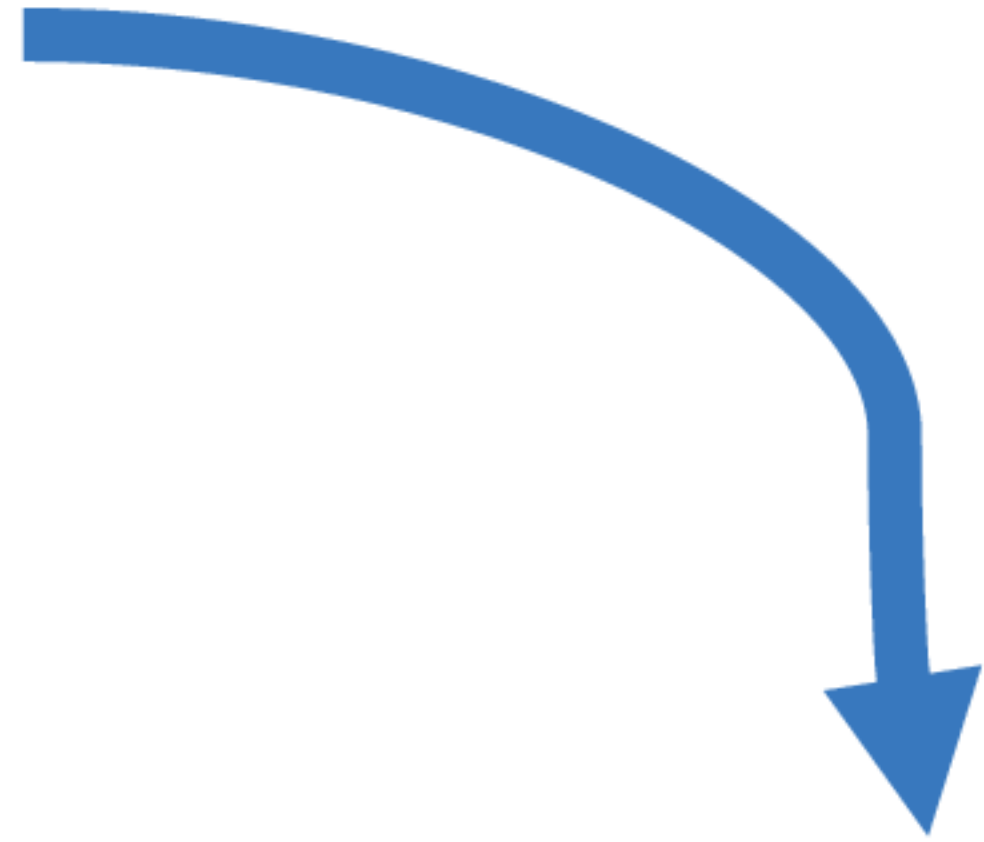


Non-uniform predictions  
(peak in highlighted  
region)

# Kaggle incentive: winner takes it all



# Can it be more fair?



# Micro-rewards examples

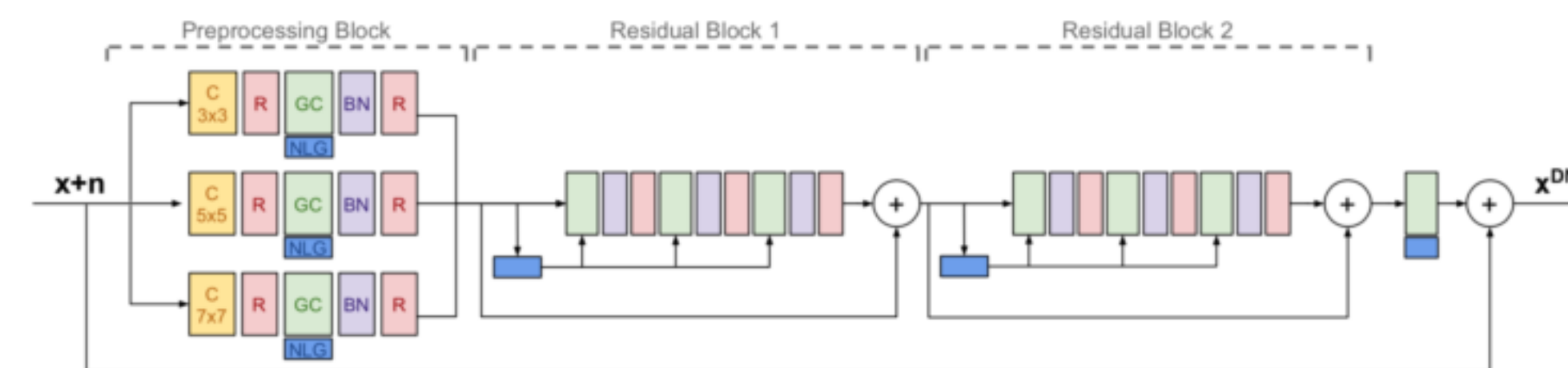
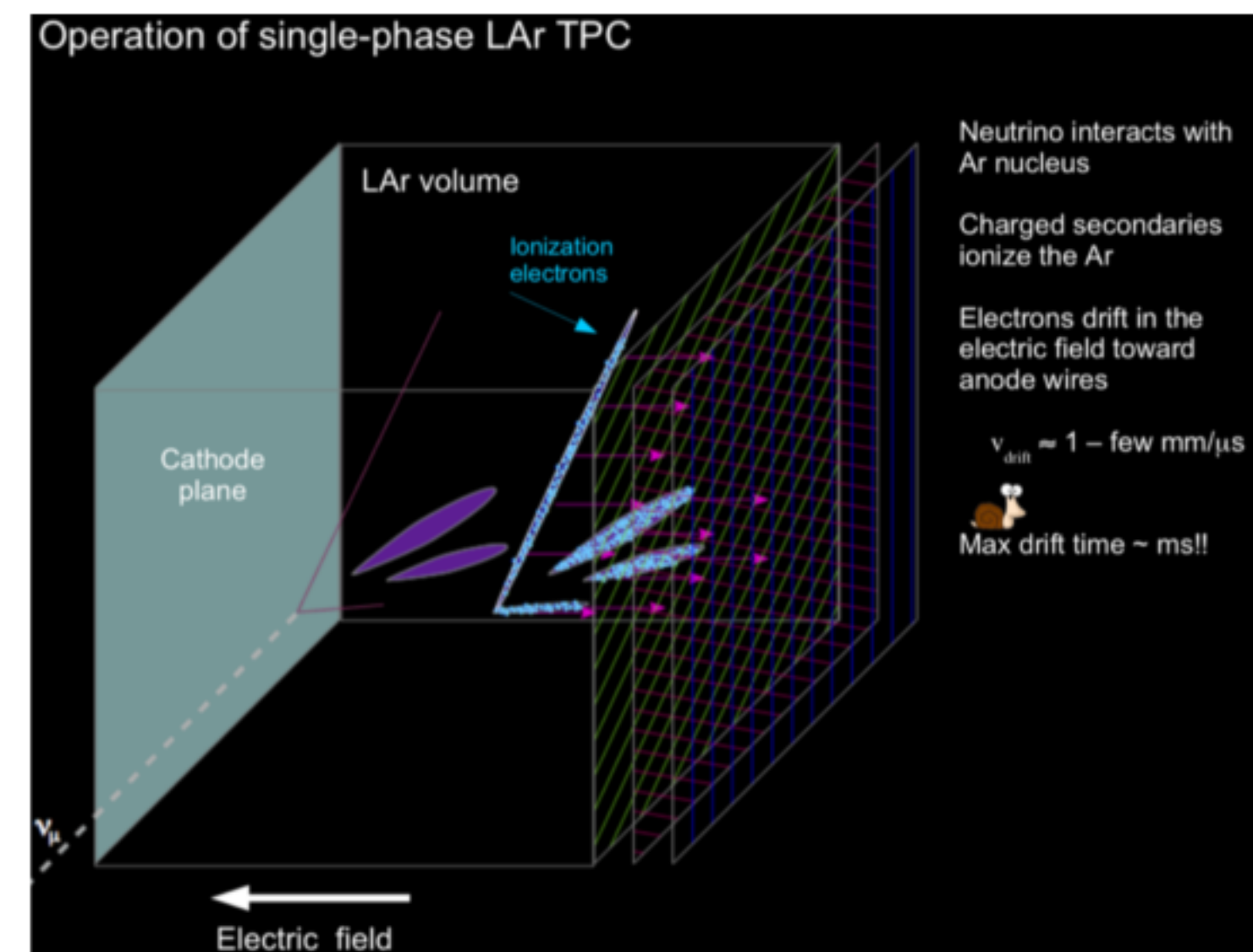
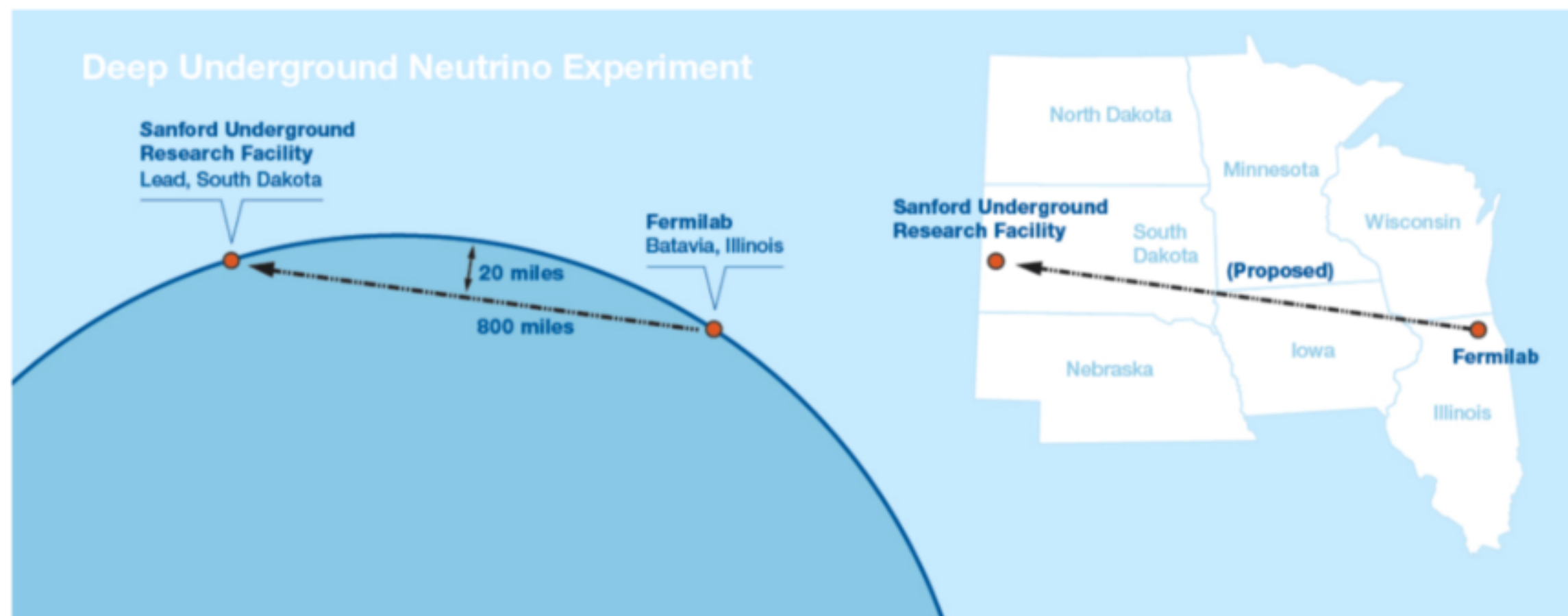
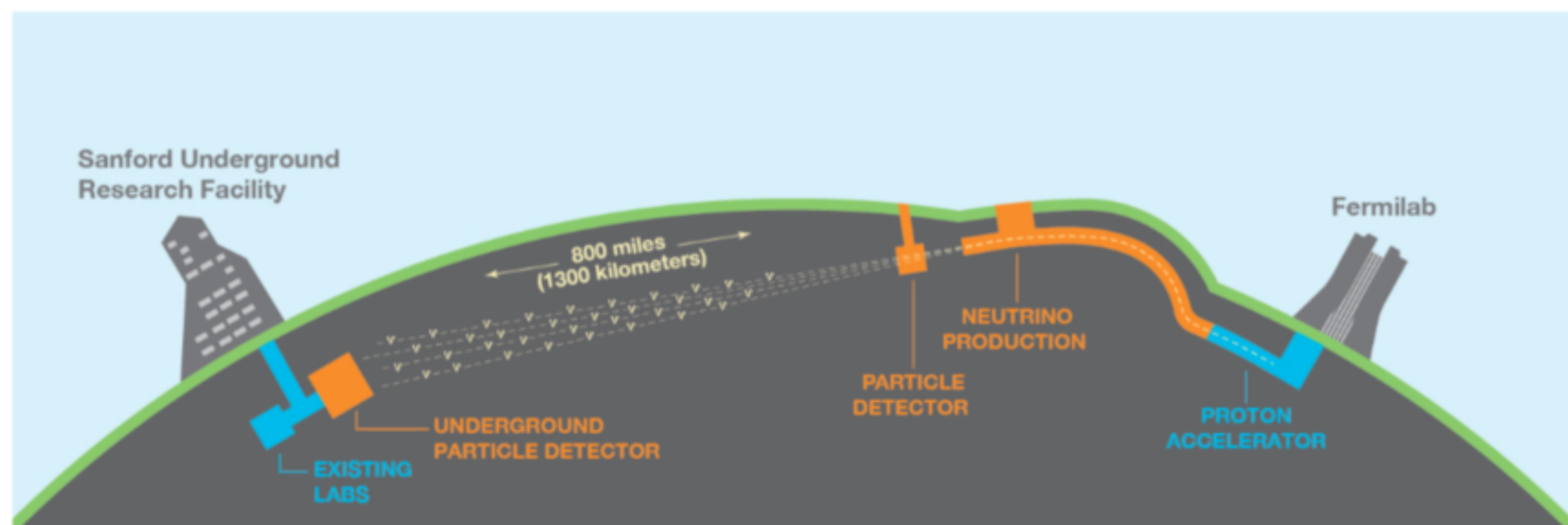
## Metric-based:

- › top X of the leaderboard,
- › time spent at the top,
- › persistence

## Source code-based:

- › Distinctive solution
- › Source of other's inspirations
- › Your code is re-used by others

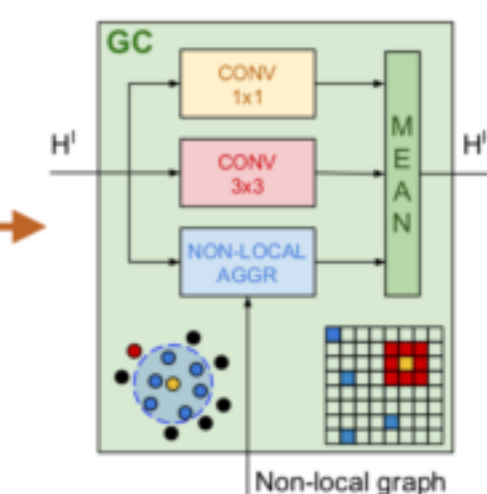
# Dune event reconstruction



C 3x3  $\rightarrow$  Convolutional Layer 3x3

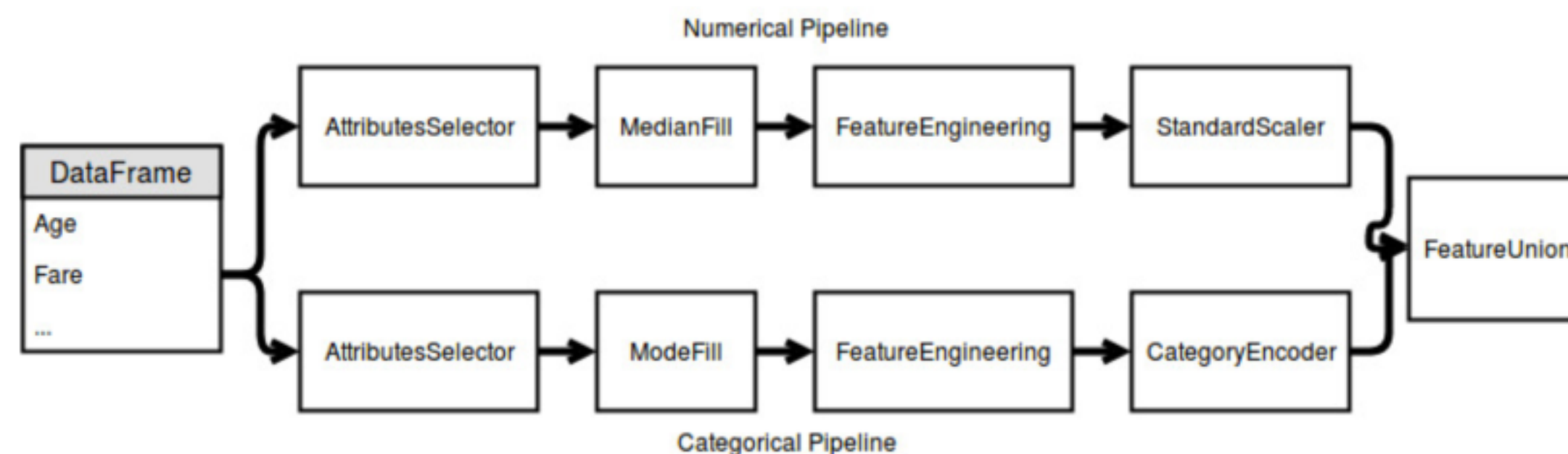
GC  $\rightarrow$  Graph-Convolutional Layer

$$H_i = \sigma \left( \underbrace{\sum_{j \in \mathcal{N}_i} \frac{F_m(H_j - H_i)}{|N_j|}}_{\text{neighborhood}} + \underbrace{W_i H_i + b}_{\text{node bias}} \right)$$



# Natural language to machine learning corpus

1. Collect code snippets of ML examples (e.g. spacy.io, uber.github.io/ludwig/, scikit-learn.org, stackexchange.com)
2. Annotate
3. Build knowledge graph of possible cases and map pairs into it
4. Build proof-of-concept classifier for certain kinds of cases



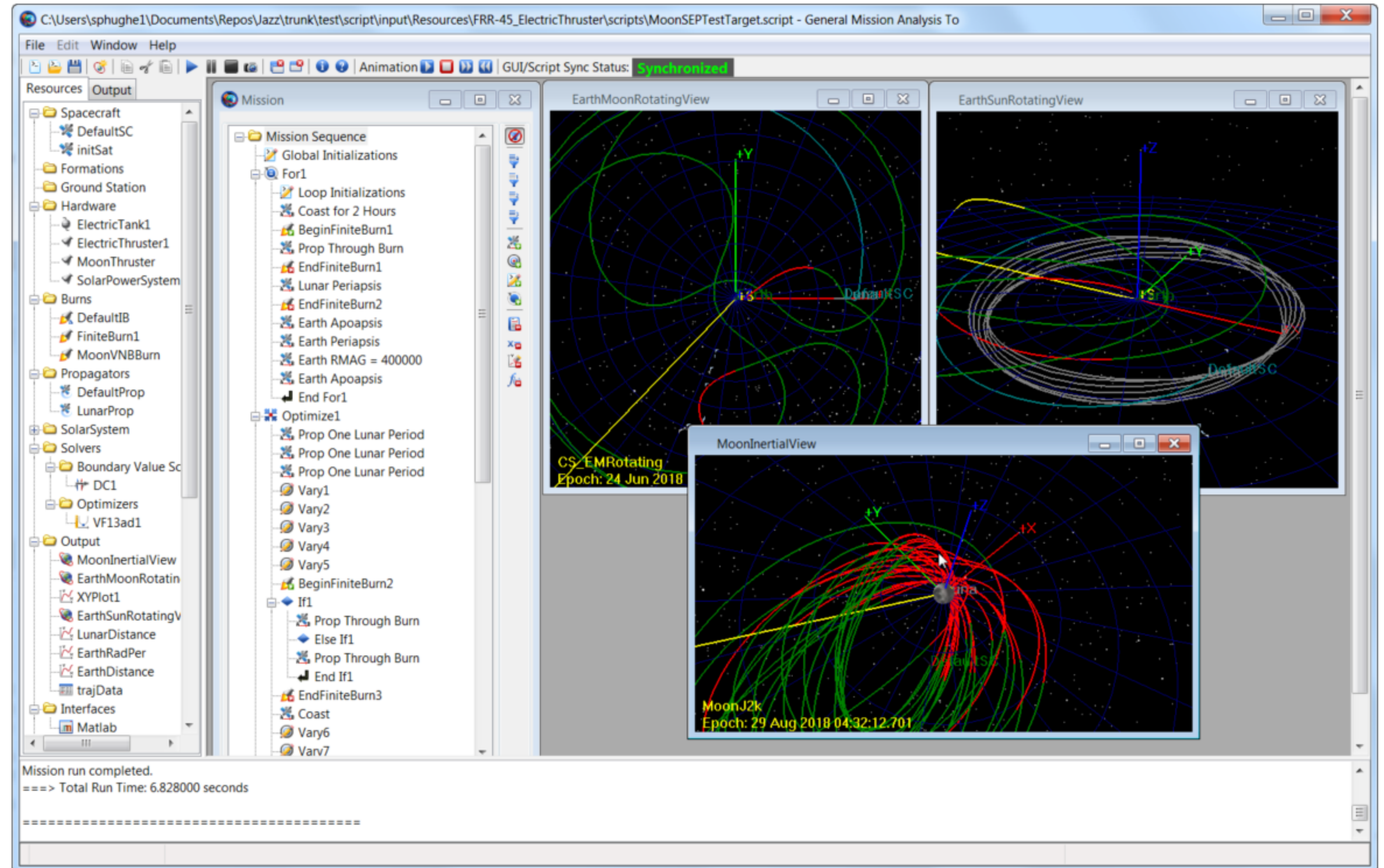
```
{
  "question_id": 36875258,
  "intent": "copying one file's contents to another in python",
  "rewritten_intent": "copy the content of file 'file.txt' to file 'file2.txt'",
  "snippet": "shutil.copy('file.txt', 'file2.txt')",
}

{
  "intent": "How do I check if all elements in a list are the same?",
  "rewritten_intent": "check if all elements in list `mylist` are the same",
  "snippet": "len(set(mylist)) == 1",
  "question_id": 22240602
}

{
  "intent": "Iterate through words of a file in Python",
  "rewritten_intent": "get a list of words `words` of a file 'myfile'",
  "snippet": "words = open('myfile').read().split()",
  "question_id": 7745260
}
```

# Satellite position prediction

1. Explore observation uncertainty propagation in time
2. Find satellite feature representation
3. Tune simulation model parameters to improve precision of satellite position predictions



# AI Physicist++

Unsupervised explorer algorithm

