

Towards gig science and augmented intelligence

July, 2019, DESY

Andrey Ustyuzhanin

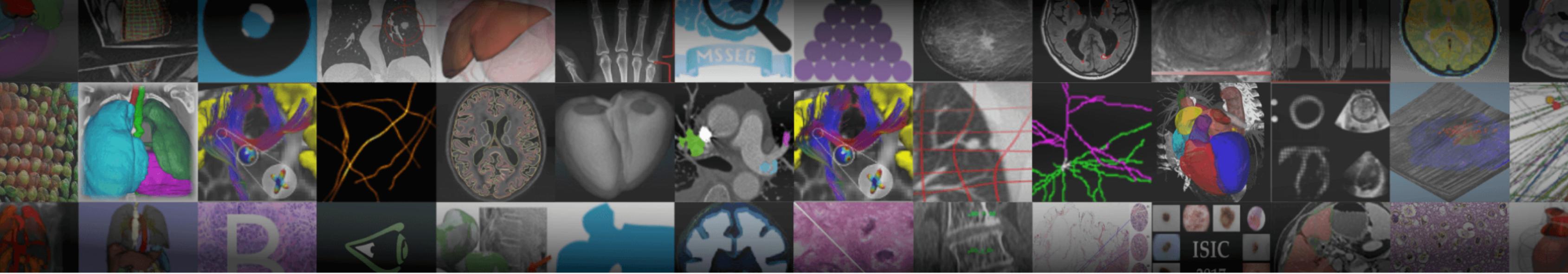
NRU HSE

YSDA

ICL

Machine Learning + Science examples

Grand-Challenges ALL CHALLENGES SIGN IN / RE



WHY CHALLENGES?
HOST A
CHALLENGE
CONTRIBUTORS

Grand Challenges in Biomedical Image Analysis

Every year, thousands of papers are published that describe new algorithms to be applied to medical and biomedical images, and various new products appear on the market based on such algorithms. But few papers and products provide a fair and direct comparison of the newly proposed solution with the state-of-the-art. We believe that such comparisons can help the research community and industry to develop better algorithms. We support the organization of these comparative studies and the dissemination of their results.

Organizing and participating in challenges is not the only way to facilitate better comparisons between new and existing solutions. If it were easy to publish and share your data, and the code you used to evaluate your algorithm's performance on that data, and possibly the algorithm itself, others could directly compare their approach to yours, using the same test data and the same evaluation metrics. With this site we provide tools to make it as easy as possible for you to publish your data and your evaluation for any paper you've written.

Machine Learning + Science examples

The image displays two side-by-side screenshots of scientific citizen science websites.

Foldit (Left): The homepage features a large green banner at the top with the Foldit logo and the text "Solve Puzzles for Science". Below the banner is a protein structure model. A callout box on the protein structure says: "Click to learn how you contribute to science by playing Foldit." The navigation menu includes links to PUZZLES, BLOG, CATEGORIES, FEEDBACK, GROUPS, FORUM, PLAYERS, WIKI, RECIPES, FAQ, CONTESTS, ABOUT, and CREDITS. The bottom left shows a "What's New" section with a link to "Paper Authorship: Calling all protein designers!" and a user login form.

Galaxy Zoo (Right): The homepage has a dark background with a large image of a spiral galaxy. The title "GALAXY ZOO.org" is prominently displayed. The navigation menu includes links to Home, The Science, How to Take Part, Galaxy Analysis, Forum, Press, Blog, FAQ, Links, and Contact Us. The main content area features a "Galaxy Analysis" section with text about the service and a large image of a spiral galaxy. To the right, there is a sidebar with a "Galaxy Ref:" code (587729387677679742) and instructions to choose a galaxy profile. Below this are four categories: "CLOCK", "ANTI", "SPRAL GALAXY", "ELLPTICAL GALAXY", "STAR DON'T KNOW", and "MERGERS".



Is there a space for gig-science?

Science as a free-lance:

- › Kolabtree.com
- › <http://labmate.us/>
- › nature.com/nature/journal/v550/n7676/full/nj7676-419a.html



Maybe [Kaggle.com](https://www.kaggle.com) could help?

- › Strong community
- › Numerous challenges

Machine Learning + Science examples

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

Featured Prediction Competition

TrackML Particle Tracking Challenge

High Energy Physics particle tracking in CERN detectors

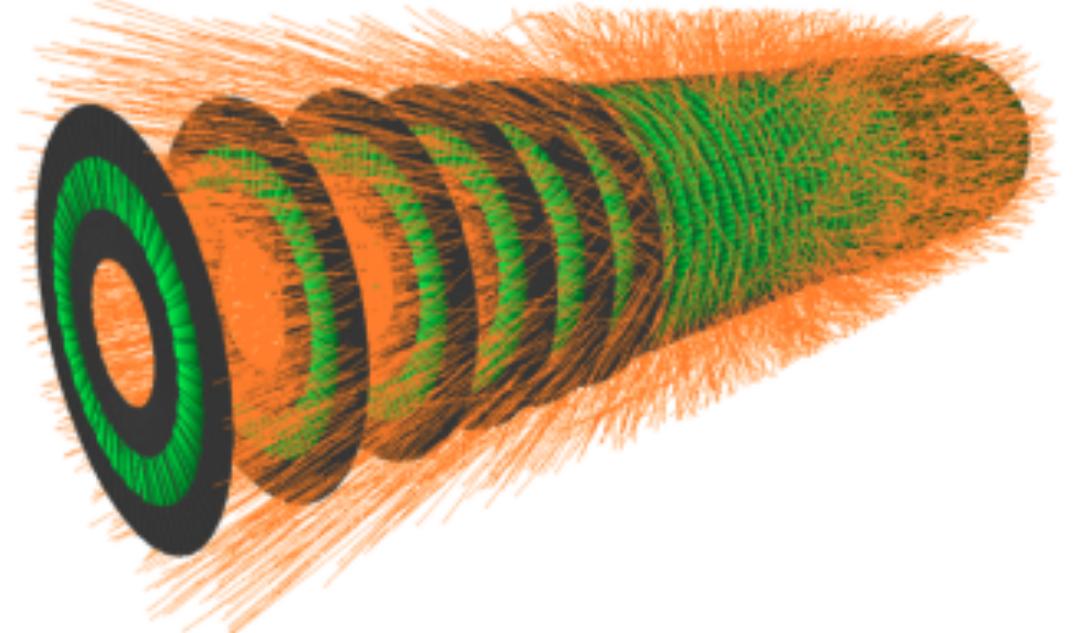
CERN · 516 teams · a month to go (a month to go until merger deadline)

\$25,000 Prize Money

Overview Data Kernels Discussion Leaderboard Rules

Overview

Description	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.
Evaluation	
Prizes	
About The Sponsors	
Timeline	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.



Andrey Ustyuzhanin 6

Kaggle shortcomings

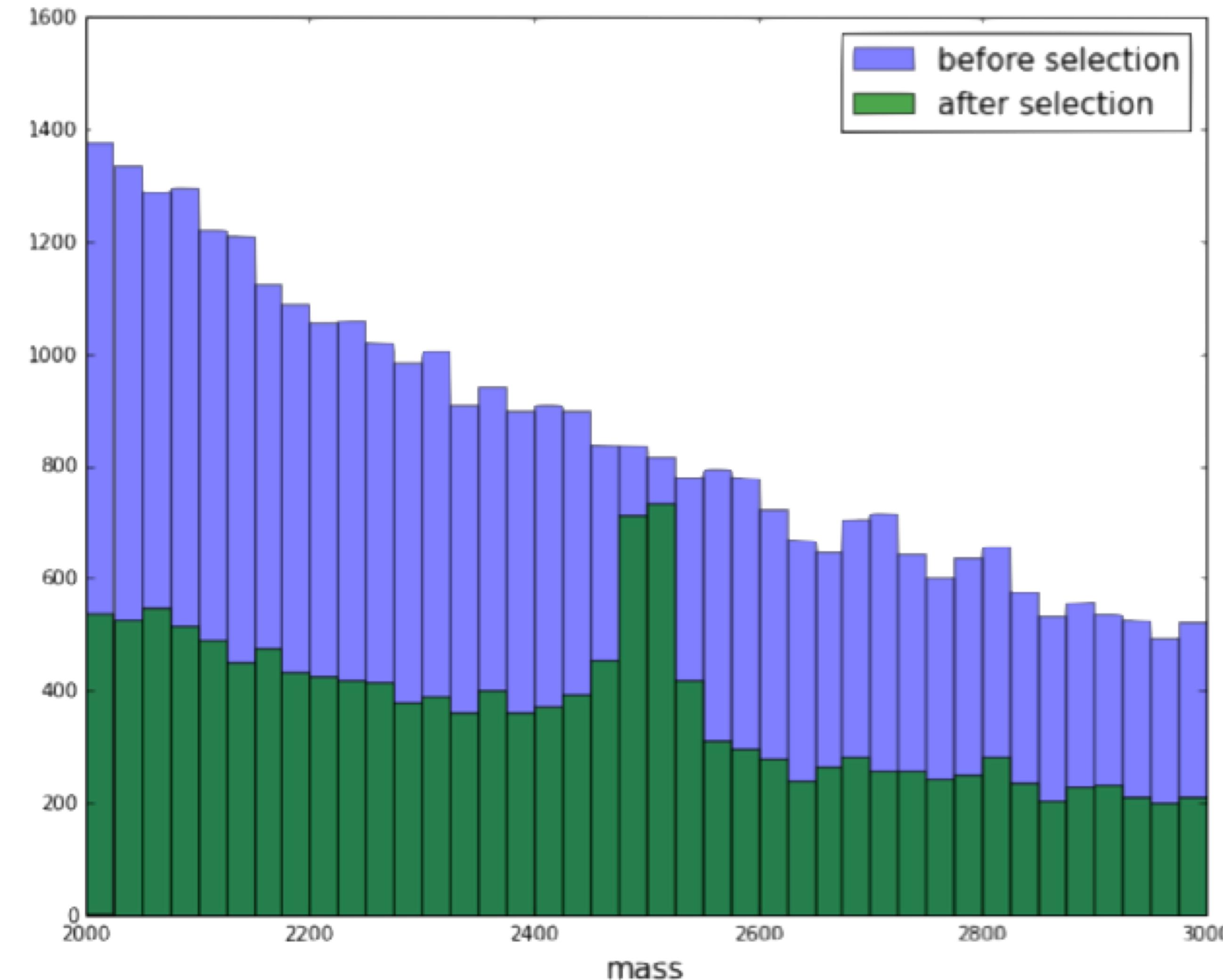
Domain specifics:

- › Non-trivial metrics
- › Constraints

AI beyond iid assumption (test set is not the same as training set)

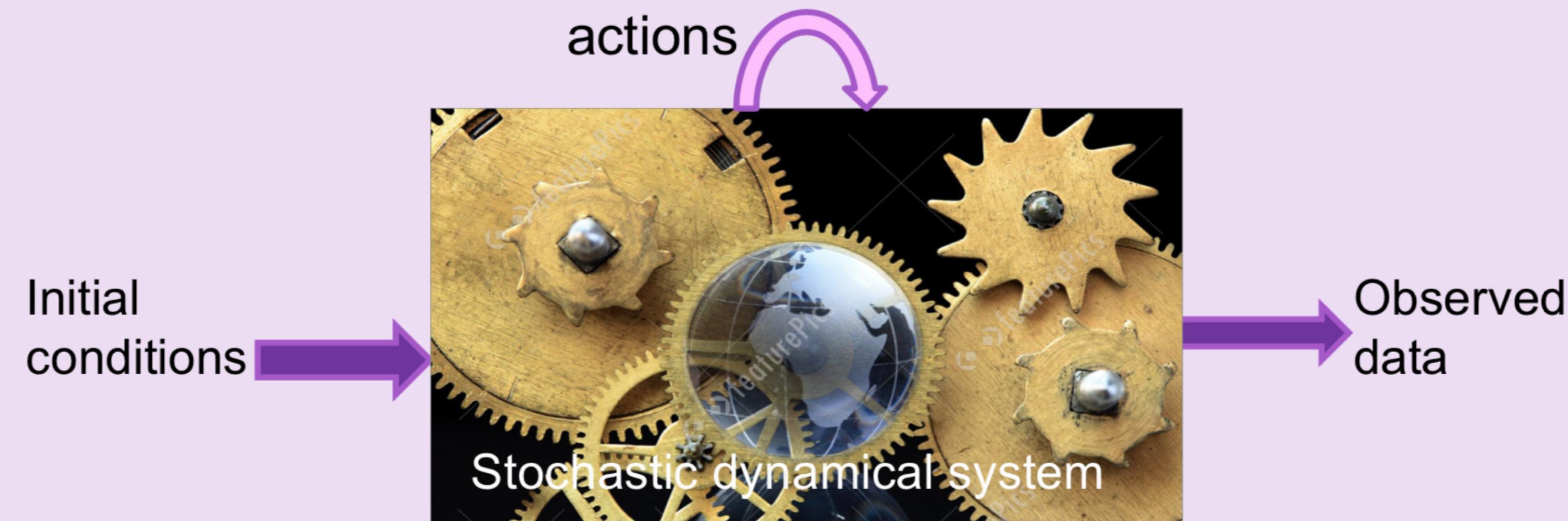
Participation Incentive

2. Domain specific example



Beyond iid assumption: causal mechanisms

- The assumption that the test data is from the same distribution as the training data is too strong, and it is often violated in practice, leading to poor out-of-distribution generalization.
- Consider relaxed assumptions: the test data was generated under the same *causal dynamics*, but from different initial conditions (which may be unlikely under the training distribution) and agents' actions.



Beyond iid assumption

RoboCup

OpenAI gym

NeurIPS CML workshop highlights

- › AI Driving Olympics: Duckie Town
- › Adversarial Attacks
- › Chat bots
- › TrackML

BabyAI

3. Kaggle incentive: winner takes it all



Can it be more fair?



Micro-rewards

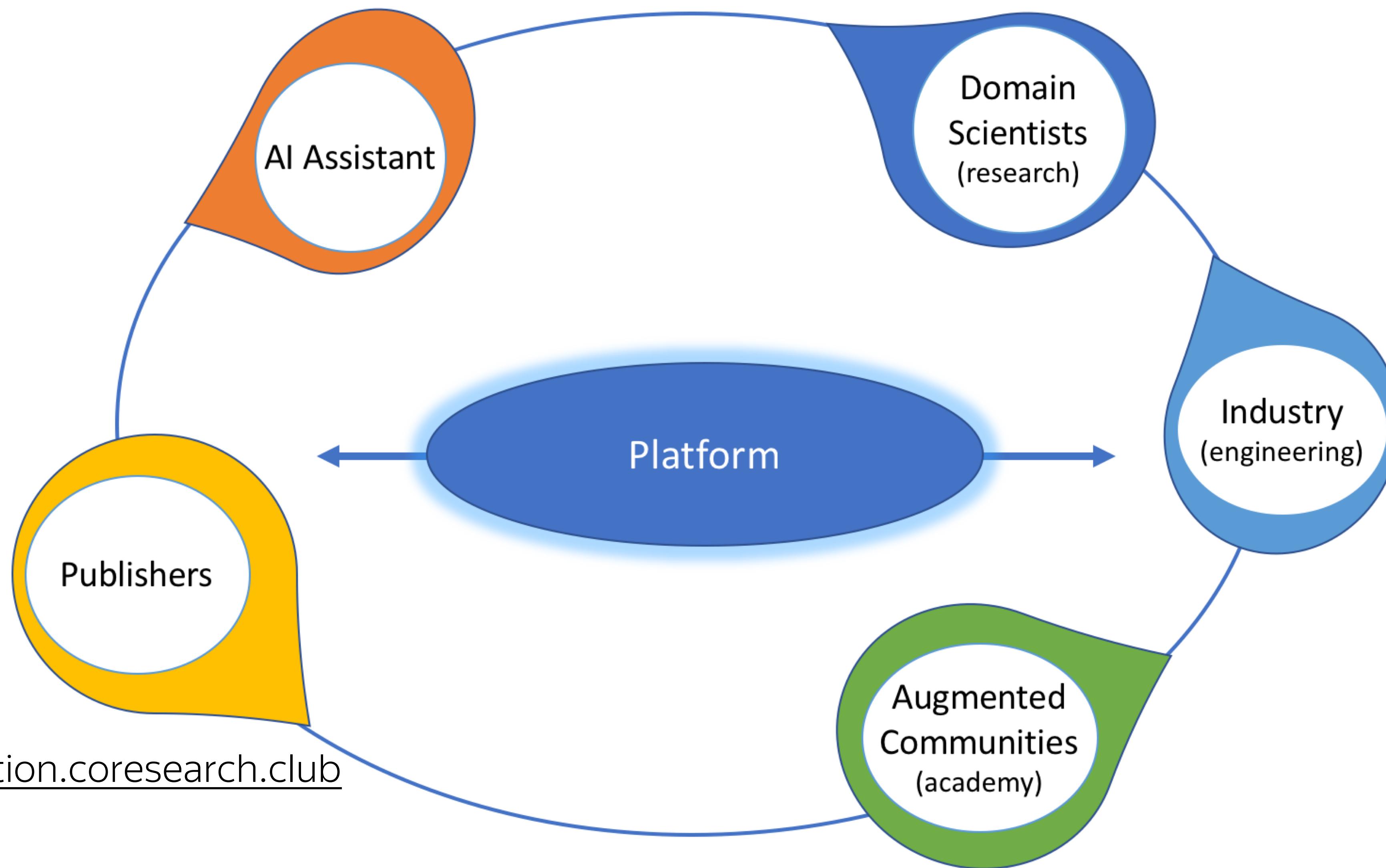
Metric-based:

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

Source code-based:

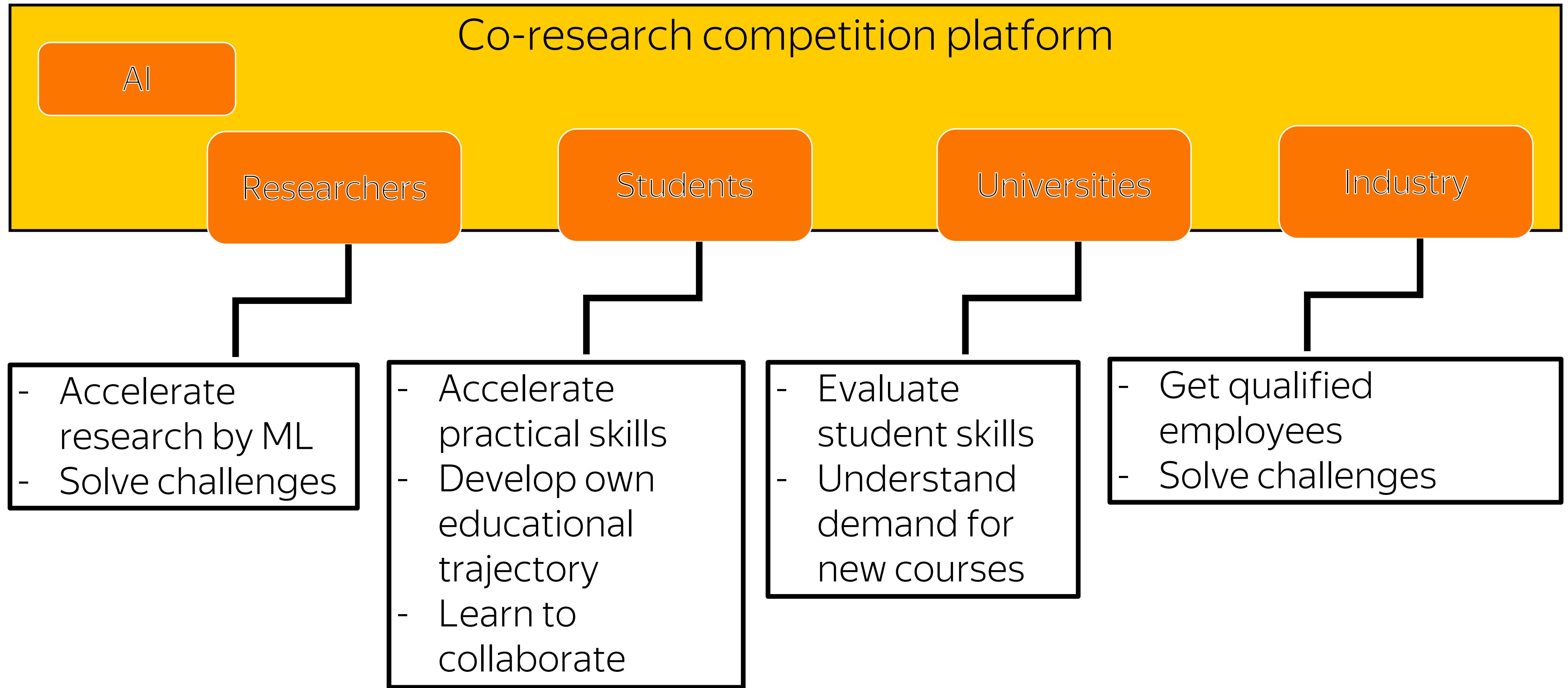
- › Distinctive solution
- › Source of other's inspirations
- › Your code is re-used by others

Co-research platform

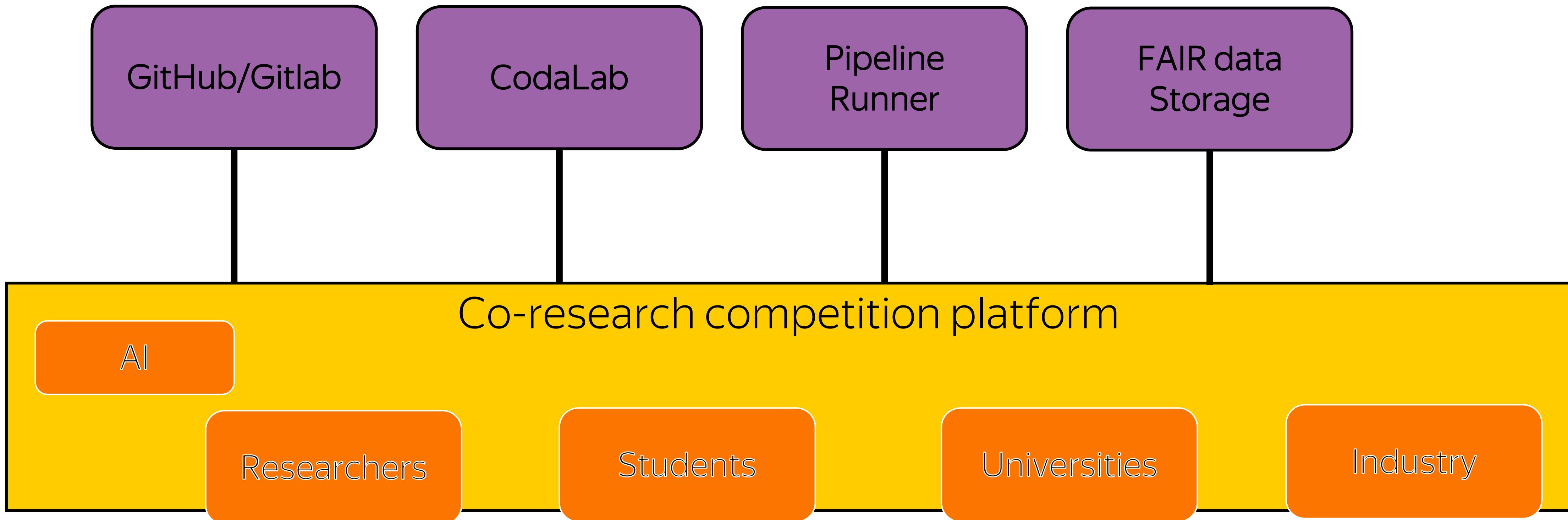


<https://coopetition.coresearch.club>

Co-Research Platform overview



Co-Research Platform overview



Regular Participant's Workflow

Register at the platform, chose a co-research project

- › Read materials, research map, get initial advise to try

Fork baseline or other's participant github repository

Iterate

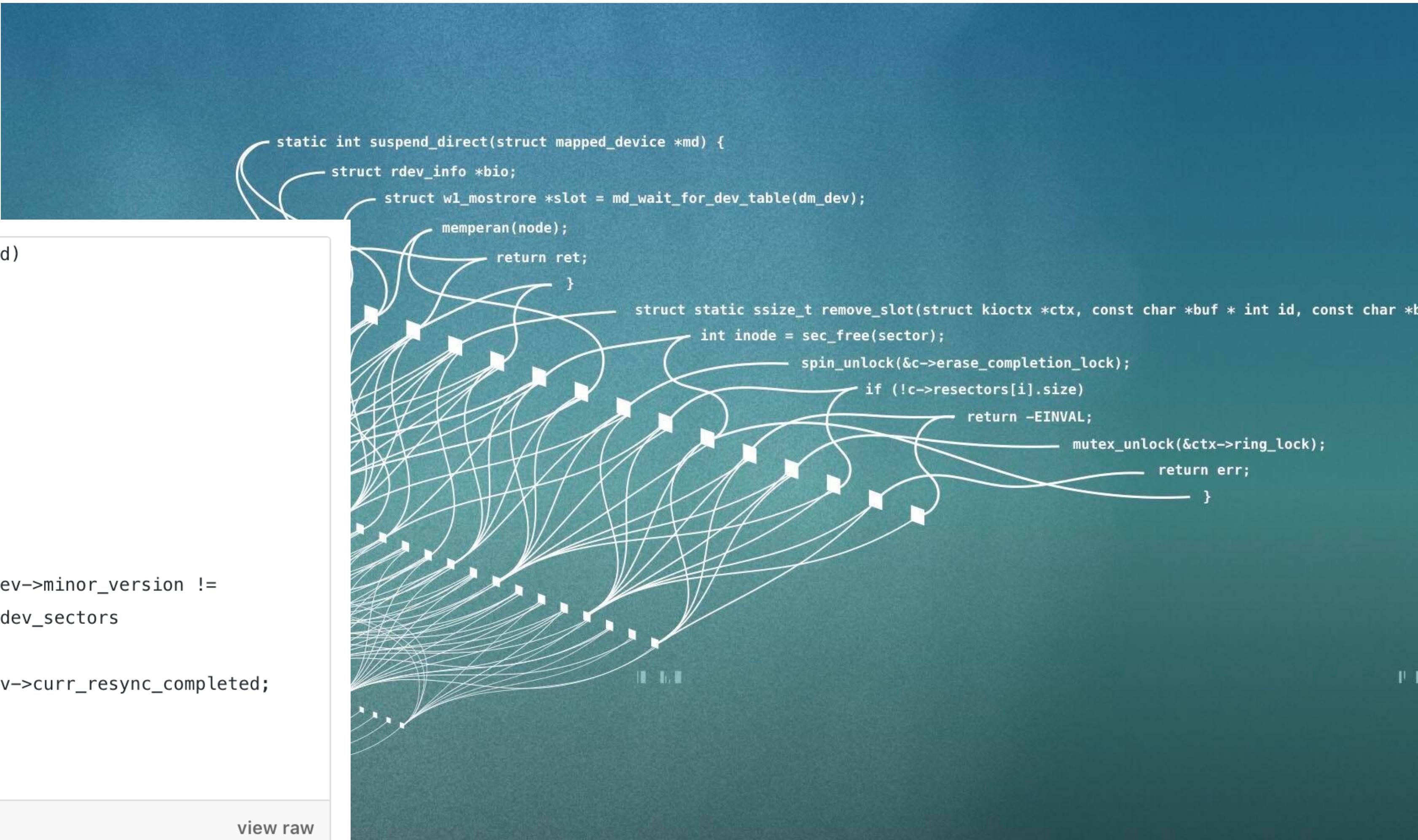
- › Read other's submissions / explore / communicate
- › Realize an idea to try
- › Write the code, commit to own repository
- › Get evaluation, recommendations

AutoML opportunity

```
1 static int super_fold(struct mddev *mddev, void __user **rd)
2 {
3     struct md_rdev *rdev;
4
5     if (!tryet & gcov_ntreef(*stint)) {
6         if (gc_th->max_sectors)
7             if (task)
8                 goto next_start;
9         if (!list_empty(&mddev->disks) {
10            if (mddev->dev_sectors == 0 ||
11                mddev->chunk_sectors == 0 && mddev->minor_version !=
12                    mddev->max_disks && mddev->dev_sectors
13                    rdev2->rescan_recovnr != 0)
14                rdev->recovery_offset = mddev->curr_resync_completed;
15        }
16    }
17 }
```

deep_generation_c_code_8.c hosted with ❤ by GitHub

[view raw](#)



Conclusion

Coopetition is a platform for multidisciplinary data-driven research

Motivation for at personal and scientific growth

Interaction between domain-experts and machine-learning experts

Source code-based submissions

AutoML for AI research assistant

<https://coopetition.coresearch.club>
anaderiRu@twitter
austyuzhanin@hse.ru

Backup

