

TÉCNICO  
LISBOA



# Coopetition

Collaborative solutions for research problems

Giles Strong

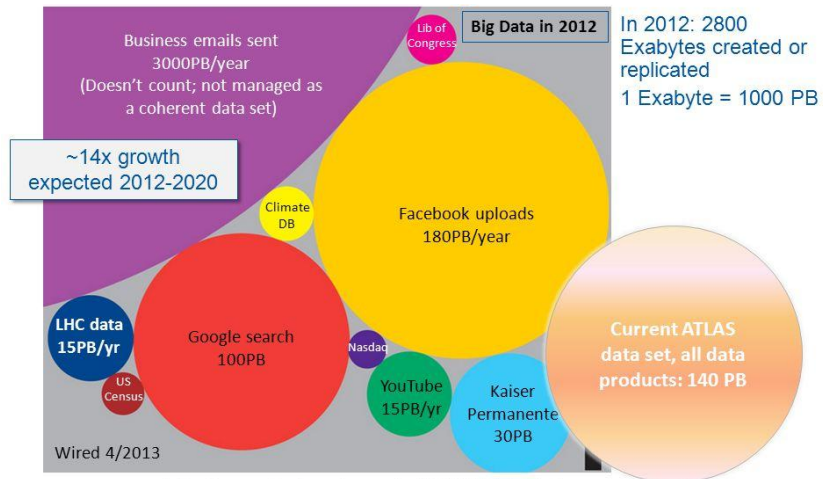
Yandex/HSE secondment summary - 29/04/19



# Background and motivation

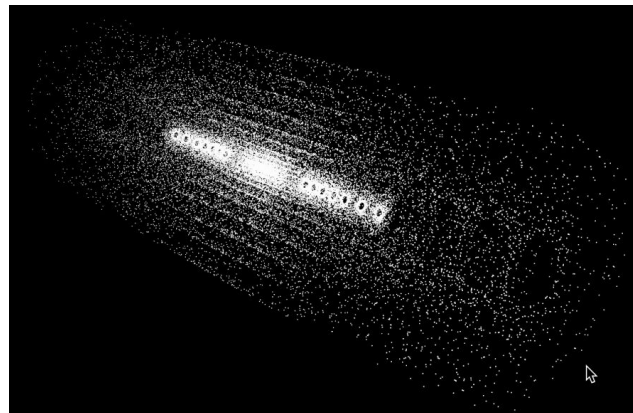
# Data growth

- Modern, state-of-the-art research facilities are increasingly producing vast amounts of data at unprecedented levels
  - E.g. (HL-)LHC, LIGO, European Solar Telescope, Cherenkov Telescope Array
- The data from these experiments are normally analysed using domain-knowledge-driven techniques
  - Good theoretical motivation
  - Easy to grasp by fellow domain experts



# Scaling problems

- The paradigm shift to 'Big Data' isn't always reflected in these domain-driven techniques:
  - The techniques struggle, or outright fail, to utilise the available data beyond the reduction of statistical uncertainties
  - Or the approaches cannot scale to be applied to the data beyond a certain size or rate
- E.g. HL-LHC track reconstruction:
  - Accurately find all 10k tracks from 100k hits
  - Do this 10-100  $\times 10^9$  times / year
  - Your computing budget won't be increased  
= Your current algorithm is 10 times too slow



# Knowledge problems

- In other cases, when domain experts do try to use extra-domain approaches, lack of experience and misunderstandings can hamper performance
- E.g. 2014 HiggsML challenge
  - Top-scoring physicist made good use of his domain knowledge
  - But was overfitting to public results (589 submissions!)
  - Was warned several times it was a bad idea, but he defended the approach
  - Once final scores revealed he dropped from 1st to 8th
  - = the data-scientists won










#	Team Name	Kernel	Team Members	Score	Entries	Last
1	Luboš Motl's team			3.85059	589	5y
2	Gábor Melis			3.85058	110	5y
3	Tim Salimans			3.84427	57	5y
4	nhlx5haze			3.80654	254	5y
5	Owen			3.80074	99	5y
6	Anton Bakhtin (Yandex)			3.79126	122	5y
7	Tatiana Likhomanenko, Alex R...			3.78725	61	5y
8	Stanislav Semenov & Co (HSE...			3.78533	68	5y

Competition ends - private results released

#	Δpub	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲1	Gábor Melis			3.80581	110	5y
2	▲1	Tim Salimans			3.78912	57	5y
3	▲1	nhlx5haze			3.78682	254	5y
4	▲38	ChoKo Team			3.77526	216	5y
5	▲35	cheng chen			3.77383	21	5y
6	▲16	quantify			3.77086	8	5y
7	▲1	Stanislav Semenov & Co (HSE...			3.76211	68	5y
8	▼7	Luboš Motl's team			3.76050	589	5y

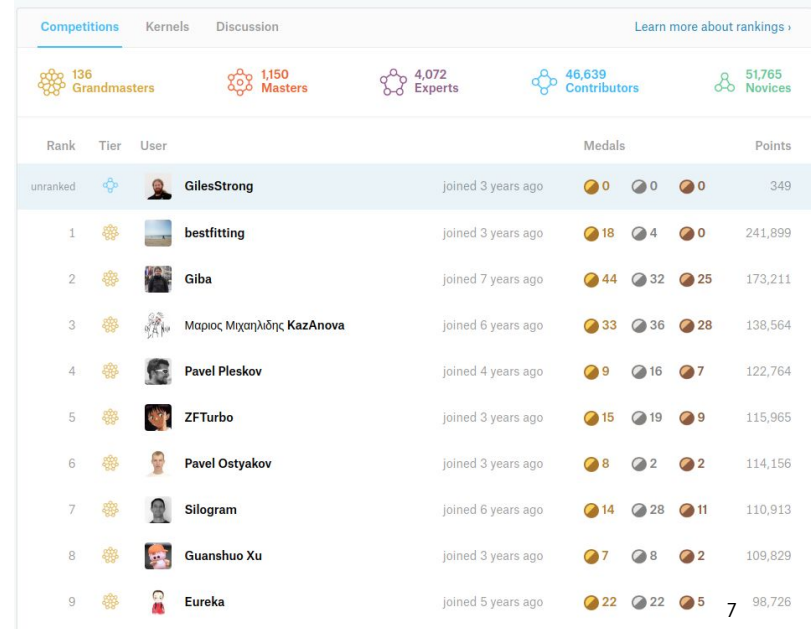
# Citizen science - research outsourcing

- This all points to a need for:
  - The sharing of knowledge expertise between domains
  - Access to innovative thinking from people with a wide range of backgrounds
- Luckily there are communities of highly motivated and intelligent people who are:
  - Eager to tackle new challenges
  - Hungry for unique datasets
  - Trying to improve their skills (and CV) through practice
  - Looking to do their part to help advance science

	<b>Two Sigma: Using News to Predict Stock Movements</b> Use news analytics to predict stock price performance <b>Featured</b> - Kernels Competition - 3 months to go - 📰 news agencies, time series, finance, money	\$100,000 2,927 teams
	<b>Jigsaw Unintended Bias in Toxicity Classification</b> Detect toxicity across a diverse range of conversations <b>Featured</b> - Kernels Competition - 3 months to go - 🗣️ biases, nlp, text data	\$65,000 188 teams
	<b>Santander Customer Transaction Prediction</b> Can you identify who will make a transaction? <b>Featured</b> - 9 days to go - 🏦 banking, tabular data, binary classification	\$65,000 8,410 teams
	<b>LANL Earthquake Prediction</b> Can you predict upcoming laboratory earthquakes? <b>Research</b> - 2 months to go - 🌍 earth sciences, physics, signal processing	\$50,000 2,204 teams
	<b>Gendered Pronoun Resolution</b> Pair pronouns to their correct entities <b>Research</b> - 21 days to go - 🗣️ nlp, text data	\$25,000 591 teams
	<b>PetFinder.my Adoption Prediction</b> How cute is that doggy in the shelter? <b>Featured</b> - Kernels Competition - 17 days to go - 🐾 image data, text data	\$25,000 2,010 teams
	<b>Google Cloud &amp; NCAA® ML Competition 2019-Women's</b> Apply Machine Learning to NCAA® March Madness® <b>Featured</b> - 8 days to go - 🏀 basketball, sports	\$25,000 502 teams
	<b>Google Cloud &amp; NCAA® ML Competition 2019-Men's</b> Apply Machine Learning to NCAA® March Madness® <b>Featured</b> - 7 days to go - 🏀 basketball, sports	\$25,000 868 teams
	<b>Data Science for Good: CareerVillage.org</b> Match career advice questions with professionals in the field	\$15,000 <sup>6</sup>

# The problem with Kaggle

- The largest of these communities is Kaggle
- Competitions provide:
  - 1000s of teams
  - Cash prizes
  - Familiar / recycled problems
- Large user base provides:
  - Good visibility at higher ranks
  - Degree of reliability achievement trust for headhunters / recruiters



The screenshot shows the Kaggle leaderboard for a competition. At the top, there are navigation tabs for 'Competitions', 'Kernels', and 'Discussion', along with a link to 'Learn more about rankings'. Below this, statistics for different user tiers are displayed: 136 Grandmasters, 1,150 Masters, 4,072 Experts, 46,639 Contributors, and 51,765 Novices. The main table lists the top 9 users, including their rank, tier, name, join date, and medal counts (Gold, Silver, Bronze), along with their total points.

Rank	Tier	User	Medals	Points	
unranked		GilesStrong	joined 3 years ago	0 Gold, 0 Silver, 0 Bronze	349
1	Grandmaster	bestfitting	joined 3 years ago	18 Gold, 4 Silver, 0 Bronze	241,899
2	Grandmaster	Giba	joined 7 years ago	44 Gold, 32 Silver, 25 Bronze	173,211
3	Grandmaster	Μαριος Μιχαηλιδης KazAnova	joined 6 years ago	33 Gold, 36 Silver, 28 Bronze	138,564
4	Grandmaster	Pavel Pleskov	joined 4 years ago	9 Gold, 16 Silver, 7 Bronze	122,764
5	Grandmaster	ZFTurbo	joined 3 years ago	15 Gold, 19 Silver, 9 Bronze	115,965
6	Grandmaster	Pavel Ostyakov	joined 3 years ago	8 Gold, 2 Silver, 2 Bronze	114,156
7	Grandmaster	Silogram	joined 6 years ago	14 Gold, 28 Silver, 11 Bronze	110,913
8	Grandmaster	Guanshuo Xu	joined 3 years ago	7 Gold, 8 Silver, 2 Bronze	109,829
9	Grandmaster	Eureka	joined 5 years ago	22 Gold, 22 Silver, 5 Bronze	98,726

# The problem with Kaggle

- But Kaggle is:
  - Fairly inflexible - most experiments can't be reduced to a single metric
    - E.g. Unable to rank by accuracy and time
  - Solutions not always appropriate / applicable to actual problem
    - E.g. 2nd place for HL-LHC tracking took one day per event
- More flexible platforms exist, but lack of user-base & renown deters participation
  - 1st phase of TrackML hosted on Kaggle = 653 teams
  - 2nd phase hosted on Codalab = only 26 teams!

RESULTS									
#	User	Entries	Date of Last Entry	score ▲	accuracy_mean ▲	accuracy_std ▲	computation time sec ▲	computation speed percent ▲	Duration ▲
1	sgorbuno	9	03/12/19	1.1727 (1)	0.944 (2)	0.00 (14)	28.06 (1)	0.36 (1)	64.00 (1)
2	fastrack	53	03/12/19	1.1145 (2)	0.944 (1)	0.00 (15)	35.51 (16)	1.11 (16)	91.00 (6)
3	cloudkitchen	73	03/12/19	0.9007 (3)	0.926 (3)	0.00 (13)	364.00 (10)	7.28 (10)	407.00 (8)
4	cubus	8	09/13/16	0.7719 (4)	0.895 (4)	0.01 (9)	675.35 (19)	13.51 (19)	724.00 (9)
5	Taka	11	01/13/19	0.5930 (5)	0.875 (5)	0.01 (12)	2068.50 (23)	53.37 (23)	2758.00 (13)
6	Vicennial	27	02/24/19	0.5634 (6)	0.815 (6)	0.01 (10)	1270.73 (20)	25.41 (20)	1339.00 (10)
7	Sharad	57	03/10/19	0.2916 (7)	0.674 (7)	0.02 (4)	1902.20 (22)	38.04 (22)	1966.00 (12)
8	WeizmannAI	5	03/12/19	0.0000 (8)	0.133 (11)	0.01 (11)	88.06 (17)	1.76 (17)	124.00 (7)
9	harshakoundinya	2	03/12/19	0.0000 (9)	0.005 (13)	0.01 (8)	49.22 (8)	0.96 (8)	86.00 (3)
10	iwit	6	03/10/19	0.0000 (9)	0.002 (15)	0.01 (8)	46.23 (3)	0.96 (3)	85.00 (2)
11	yangguo	1	04/01/19	0.0000 (8)	0.002 (15)	0.01 (8)	48.65 (4)	0.97 (4)	86.00 (3)
12	alexander_liao	13	02/14/19	0.0000 (9)	0.170 (10)	0.01 (5)	8271.47 (25)	165.43 (25)	9473.00 (15)
13	datomi	7	02/14/19	0.0000 (9)	0.005 (14)	0.01 (7)	50.17 (15)	1.00 (15)	87.00 (4)
14	hansl	1	01/21/19	0.0000 (8)	0.002 (15)	0.01 (8)	49.37 (6)	0.96 (6)	86.00 (3)
15	BimBomBom	3	12/20/18	0.0000 (9)	0.002 (15)	0.01 (8)	45.96 (5)	0.96 (5)	87.00 (4)
16	khavo	3	10/29/16	0.0000 (9)	0.304 (8)	0.03 (1)	10015.06 (26)	380.30 (26)	15419.00 (16)
17	traffic_congestion	2	10/21/18	0.0000 (8)	0.002 (15)	0.01 (8)	49.67 (4)	0.99 (14)	86.00 (3)
18	emb	3	10/20/18	0.0000 (8)	0.123 (12)	0.02 (3)	1004.97 (21)	37.30 (21)	1940.00 (11)
19	karadbara	1	10/17/18	0.0000 (8)	0.002 (15)	0.01 (8)	49.19 (7)	0.96 (7)	87.00 (4)
20	sanjaykr10	1	10/17/18	0.0000 (8)	0.002 (15)	0.01 (8)	49.35 (10)	0.99 (10)	86.00 (3)
21	EdmonWales	1	10/14/18	0.0000 (8)	0.002 (15)	0.01 (8)	49.23 (9)	0.96 (9)	86.00 (3)
22	dcoldeira	1	10/13/18	0.0000 (8)	0.002 (15)	0.01 (8)	49.66 (13)	0.99 (13)	86.00 (3)
23	brunoseznc	1	10/06/18	0.0000 (8)	0.002 (15)	0.01 (8)	49.35 (11)	0.99 (11)	87.00 (4)
24	mikhail94321	1	09/11/18	0.0000 (8)	0.173 (8)	0.02 (2)	4945.60 (24)	95.91 (24)	3060.00 (14)
25	droussea_naif	1	09/07/18	0.0000 (8)	0.002 (15)	0.01 (8)	48.21 (2)	0.96 (2)	85.00 (2)
26	Tester_91	1	09/07/18	0.0000 (8)	0.002 (15)	0.01 (8)	49.62 (12)	0.99 (12)	87.00 (4)





# The problem with Kaggle

- With such large prizes at stake, cooperation between teams (sharing of domain expertise) is disincentivised
- When knowledge sharing does occur, there's no official record of who did what

# Problem solutions

- *Fairly inflexible - most experiments can't be reduced to a single metric*
  - Use a range of smaller, specialised platforms with configurable (multi)metrics
- *Solutions not always appropriate / applicable to actual problem*
  - Provide guidelines / restrictions
  - Have a range of rewarded achievements
    - E.g. solution uses a single model to get above a given score, solutions runs on limited hardware
  - Run challenge in phases - allows organisers to fix exploits without restarting challenge

# Problem solutions

- *More flexible platforms exist, but lack of user-base deters participation*
  - Share user-base between smaller platforms via a single user-profile which shows full user-history
- *With such large prizes at stake, cooperation between teams (sharing of domain expertise) is disincentivised*
  - Move the challenge from *competitor vs. competitor* to *competitors vs. problem*
  - Competitors work together to solve the problem and are rewarded as a whole

# Problem solutions

- *When knowledge sharing does occur, there's is no official record of who did what*
  - Competitors submit via commits to open-source solutions
  - Full history of user contributions visible via Git and impact matched to changes in score metrics
- *Good visibility at higher ranks & degree of reliability achievement trust for headhunters / recruiters*
  - Provide customisable searching and ranking of users via their single user profile
  - Allows recruiters to better see the specific skills of users, rather than a single aggregated score
  - Doesn't penalise users for focussing on specific types of problems

# Problem solutions

- *Familiar / recycled problems* → cutting edge, domain-specific tasks
  - The use of challenge phases can be used to:
    - Step users through unfamiliar and difficult problems
    - Gradually increase the challenge difficulty from simplified to real-world
  - Rewarded achievements can:
    - Encourage users to explore the full solution space
    - Provide achievable goals to prevent users from becoming discouraged and leaving



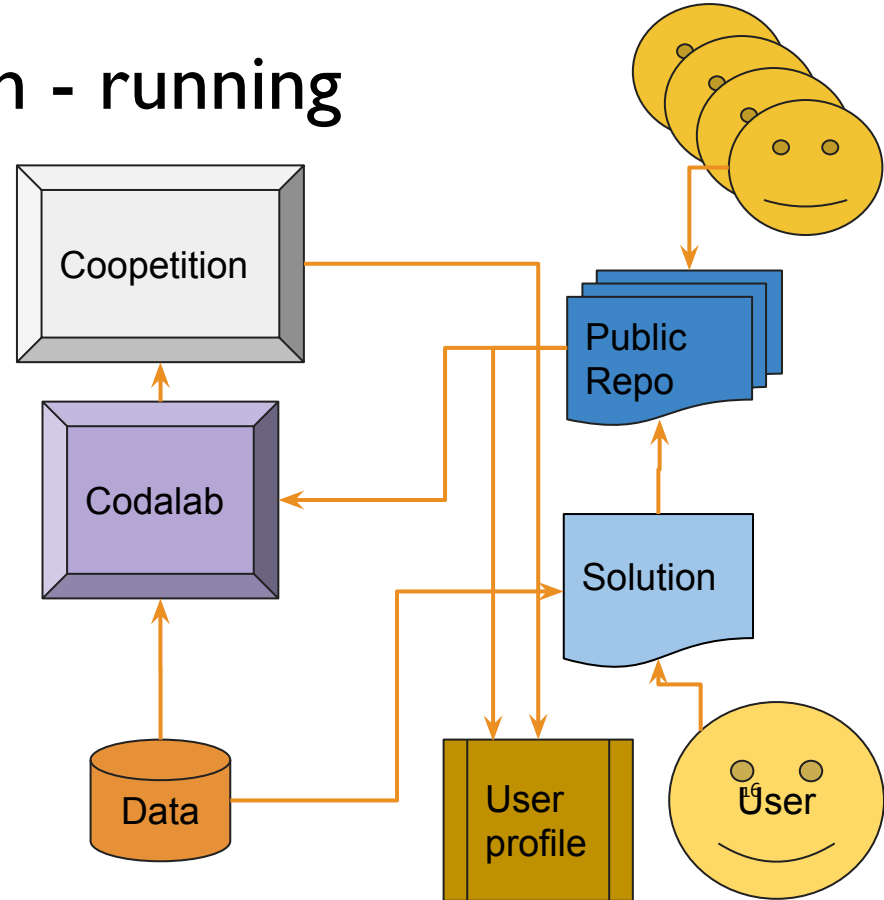
## Secondment activities

# Coopetition

- My secondment involved working on the development of ‘Coopetition’
- A portmanteau of *cooperation* and *competition*
- Coopetition is a challenge platform which sits on top of a fork of Codalab
- Users cooperate and work together to solve problems
- In the near future it might be one of a range of platforms which together form a federated group sharing a single user-profile

# Coopetition - running

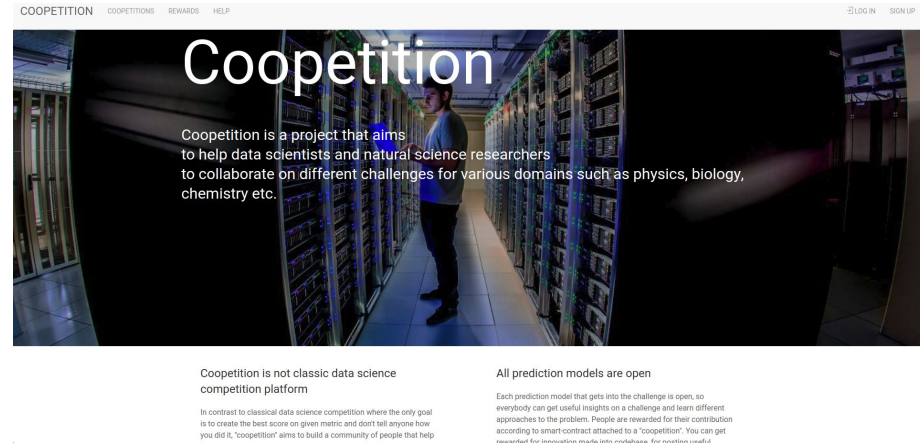
- Challenges are hosted on Codalab but submissions are made via Git commits to public repositories
- Webhooks allow solutions to be pulled and evaluated
- Configurable rules allow atomic rewards to be provided whilst the challenge is running





# Coopetition status

- Coopetition is currently live:  
<https://coopetition.coresearch.club/>
- And connected to Codalab:  
<https://codalab.coresearch.club/>



COOPETITION COOPETITIONS REWARDS HELP LOG IN SIGN UP

## Coopetition

Coopetition is a project that aims to help data scientists and natural science researchers to collaborate on different challenges for various domains such as physics, biology, chemistry etc.

Coopetition is not classic data science competition platform

In contrast to classical data science competition where the only goal is to create the best score on given metric and don't tell anyone how you did it, "coopetition" aims to build a community of people that help

All prediction models are open

Each prediction model that gets into the challenge is open, so everybody can get useful insights on a challenge and learn different approaches to the problem. People are rewarded for their contribution according to smart contract attached to a "coopetition". You can get rewarded for innovation made into marketplace. For innovation useful

# Coopetition status

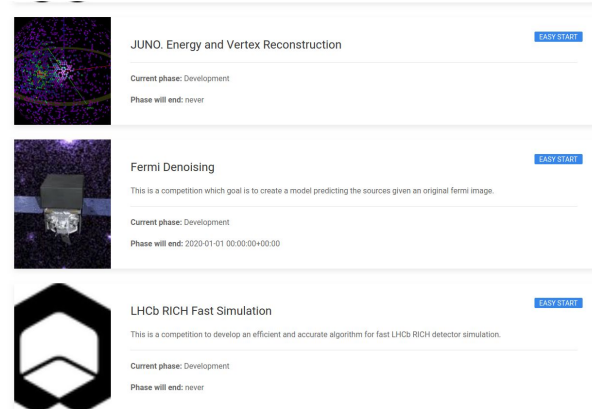
- Coopetitions are currently running within the Higher School of Economics
- Provides easy testing and feedback since the Yandex Lambda Group is situated within the HSE computer science faculty

## Cityscapes. Urban scenes segmentation

[Register on codalab](#)

Leaderboard

RANK	NAME	DATE	AP	MDU	VIEW	RATING
1	buntar29	2019-2-13	0.0	0.3688		♡0
2	mitrosh11	2019-2-17	0.0	0.6362	<a href="#">View</a>	♡1
3	alucard1177	2019-2-17	0.0	0.6362		♡0
4	human97	2019-2-17	0.0	0.6362		♡0
5	farran	2019-2-21	0.0	0.6362	<a href="#">View</a>	♡1



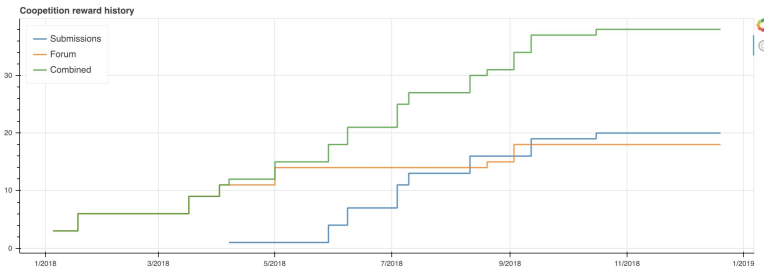
The screenshot displays a competition dashboard with three entries. Each entry includes a thumbnail image, the competition title, a 'EASY START' button, the current phase, and the phase end date.

- JUNO Energy and Vertex Reconstruction**: Current phase: Development, Phase will end: never.
- Fermi Denoising**: Current phase: Development, Phase will end: 2020-01-01 00:00:00+00:00.
- LHCb RICH Fast Simulation**: Current phase: Development, Phase will end: never.

# My contributions

- I was mainly involved in building the reward system:
  - a. Admins can set rewarders per competition phase
  - b. Solution submissions trigger the system to pull the result and check it against the configured rewarders
  - c. Any rewards the user received are then added to their profile and various logs are updated

## Reward History



COMPETITION COOPERATIONS REWARDS HELP MY COOPERATIONS LOG OUT

## COOPERATION ADMIN

SETTINGS REWARD HISTORY

### 1 Development

#### General Parameters

Baseline URL:

#### Rewards rules

Rewarder type:   Reward amount:  Threshold for reward:

SAVE

19

### 2 Final

#### General Parameters

Baseline URL:

# My contributions

- I was also involved in:
  - General testing
  - Code review
  - Bug fixes
  - Code deployment

- Overall, this required:

- Python
  - Docker
  - HTML
  - SQL
  - Postgres
  - Django
  - Nginx
  - Rabbit/Celery/Flower
  - Knowledge of databases, security, web design, code review, issue tracking, unit tests
- New

# Overall impressions

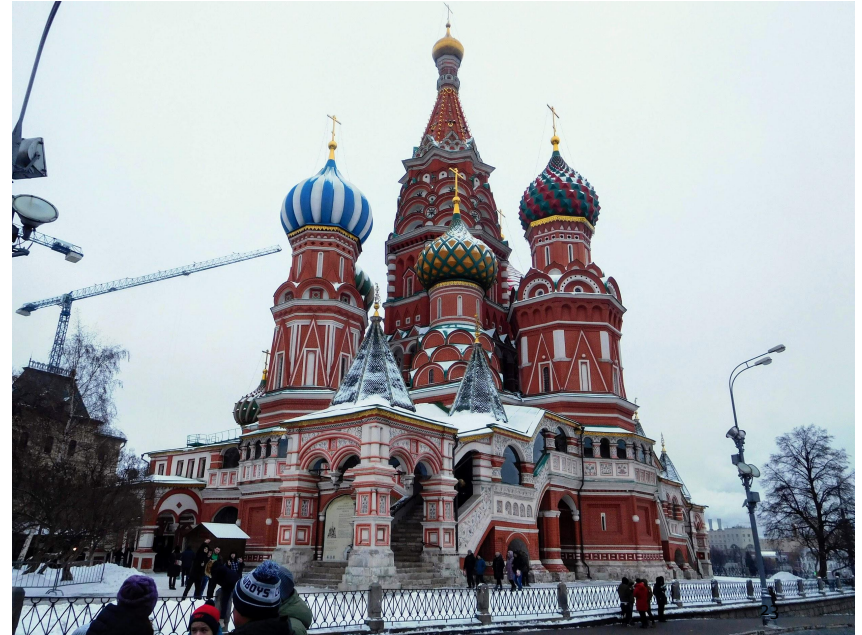
- I was lacking in many of the technical skills required, so got thrown in the deep-end a bit
  - Still, the group I was working with were able to bring me up to speed, and explain things which I didn't know
  - Certainly I learnt a lot
- Another new aspect was the Scrum Agile Development process:
  - Setting clear, short-term goals
  - Working on them during a *sprint* - with daily *standups* to report progress
  - Presenting the completed tasks
  - Retrospective review of the sprint

# Overall impressions

- It was also interesting to see the difference between how physicists and software developers approach problems, e.g. to build a complex system
  - Both would split the system into smaller components
  - My approach seemed to be to make the components complex and join them together with simple connections
  - Whereas the software developers would make the components very simple and then knot them together with complex connections
- The latter sounds like it should be more difficult to understand how the system works, as there are more interactions to keep track of
- But they seemed to have a predefined way of tying the components together, which meant they could quickly grasp what someone else's code was doing.

# Overall impressions

- Overall, it was a useful experience to see software development in a professional environment
- The skills I've picked up are already being put to good use in [my own code](#)





This Report is part of a project that has received funding from the **European Union's Horizon 2020 research and innovation programme** under grant agreement N°675440