Networked data-science for research, academic communities and beyond

2018 Oct 24, DESY

Andrey Ustyuzhanin

NRU HSE

YSDA

ICL

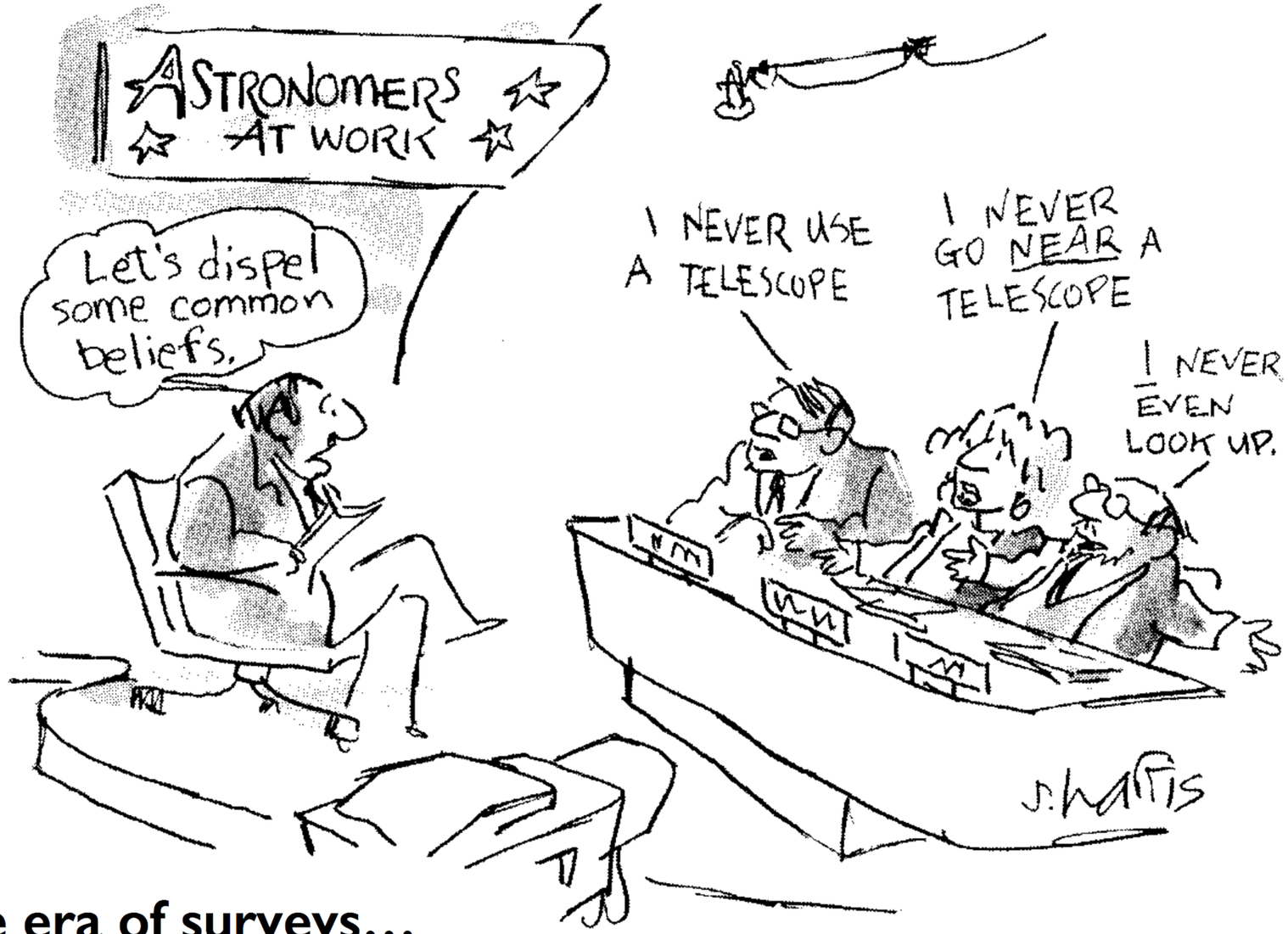
Abridged history of Science

1000+ years - empirical (Aristotle, Democritus,)
100+ years – theoretical (Newton, Kepler,)
50+ years – computational (John von Neumann,)
10+ years – data driven (the “Fourth paraditm”, Jim Gray,)

- Unify theory, experiment and simulation
- Data is captured or simulated
- Processed by software
- Information/knowledge is stored in computer
- Scientists analyzes database/files using data management and statistics

The Fourth Paradigm

From Zeljko Ivezic



The era of surveys…

"Ask Not What Data You Need To Do Your Science, Ask What Science You Can Do With Your Data."

cortica, Igocion, SPACE_KNOW, Optricity, netra, deepomatic
CAPIO, Expect Labs, Clover, Mobvoi, Qurious.AI, popUP archive
UPTAKE, INSIGHT, Preferred Networks, thingworx, KONUX, Alluvium
Alation, sapho, Outlier, Digital Reasoning
Bottlenose, MOTIVA, enigma, CB INSIGHTS, Tracxn, predata

ENTERPRISE FUNCTIONS

CUSTOMER SUPPORT
DigitalGenius, Kasisto, ELOQUENT, Wisedo, ACTIONIQ, zendesk, Preact, CLARABRIDGE

SALES
collective[i], 6sense, fuse|machines, AVISO, salesforce, INSIDE SALES .COM, clari, Zensight

MARKETING
MENTIGO, Lattice, RADIUS, Liftigniter, [PERSADO], brightfunnel, retention science, COGNICOR, AIRPR, mag-al

SECURITY
CYLANCE, DARKTRACE, ZIMPERIUM, darktrace, Sentinel, DEMISTO, graphistry, drawbridge, SignalSense, AppZen

RECRUITING
textio, entelo, Wade & Wendy, hiQ, unitive, SpringRole, GIGSTER, Hire Vue

AUTONOMOUS SYSTEMS

GROUND NAVIGATION
drive.ai, AdasWorks, ZOOX, Mobileye, UBER, Google, TESLA, nuTonomy, Auro Robotics

AERIAL
SKYDIO, SHIELD AI, Airware, dji, LILY, DroneDeploy, pilot.ai, SKYCATCH

INDUSTRIAL
JAYBRIDGE, OSARO, CLEARPATH, fetch, KINDRED, HARVEST, rethink robotics

AGENTS

PERSONAL
amazon alexa, Cortana, Allo, facebook, Siri, Replika

PROFESSIONAL
butter.ai, pogo, SKIPFLAG, clara, x.ai, slack, talla, Zoom.ai, sudo

INDUSTRIES

AGRICULTURE
BLUE RIVER, mavrx, tule, TRACE GENOMICS, Pivot Bio, Terrafusion, AGRI-DATA, Descartes Labs, udio, abundant

EDUCATION
KNEWTON, volley, gradescope, CTI, coursera, UDACITY, altschool

INVESTMENT
Bloomberg, sentient, SENTIUM, KENSHO, alphasense, Dataminr, CEREBELLUM CAPITAL, Quandl

LEGAL
blue J, BEAGLE, Everlaw, RAVEL, seal, ROSS, LEGAL ROBOT

LOGISTICS
NAUTO, Acerta, PRETECKT, Routific, clearmetal, MARBLE, PITSTOP

INDUSTRIES CONT'D

MATERIALS
zymergen, Citrine, Eigen Innovations

RETAIL FINANCE
TALA, zest finance, Lendo

HEALTHCARE

PATIENT
PULSE, CareSkore, ZEPHYR, IBM Watson

IMAGE
BUTTERFLY, 3SCAN, ARTERYS, enlitic

BIOLOGICAL
iCarbonX, color, GRAIL, deep genomics, RECURSION

DOMINO, SPARKBEYOND, rapidminer, kaggle, DataRobot, yhat, AYASDI, data iku, seldon, yseop, bigml

MACHINE LEARNING
CognitiveScale, GoogleML, context relevant, Cycorp, HyperScience, naralogics, minds.ai, H2O.ai, SCALED INFERENCE, sparkcognition, Loop, GEOMETRIC INTELLIGENCE, deepsense.io, reactive, skymind, bonsai

NATURAL LANGUAGE
agolo, RAYLIEN, LEXALYTICS, Narrative Science, spaCy, LUMINOSO, cortical.io, MonkeyLearn

DEVELOPMENT
SIGOPT, HyperOpt, fuzzy.io, kite, rainforest, lobe, Anodot, Signifai, LAYER 6, bonsai

DATA CAPTURE
CrowdFlower, diffbot, CrowdAI, import.io, Paxata, DATASIFT, amazon mechanical turk, enigma, WorkFusion, DATALOGUE, TRIFACTA, parsehub

OPEN SOURCE LIBRARIES
Keras, Chainer, CNTK, TensorFlow, Caffe, H20, DEEPLEARNING4J, theano, torch, DSSTNE, Scikit-learn, AzureML, neon, MXNet, DMTK, Spark, PaddlePaddle, WEKA

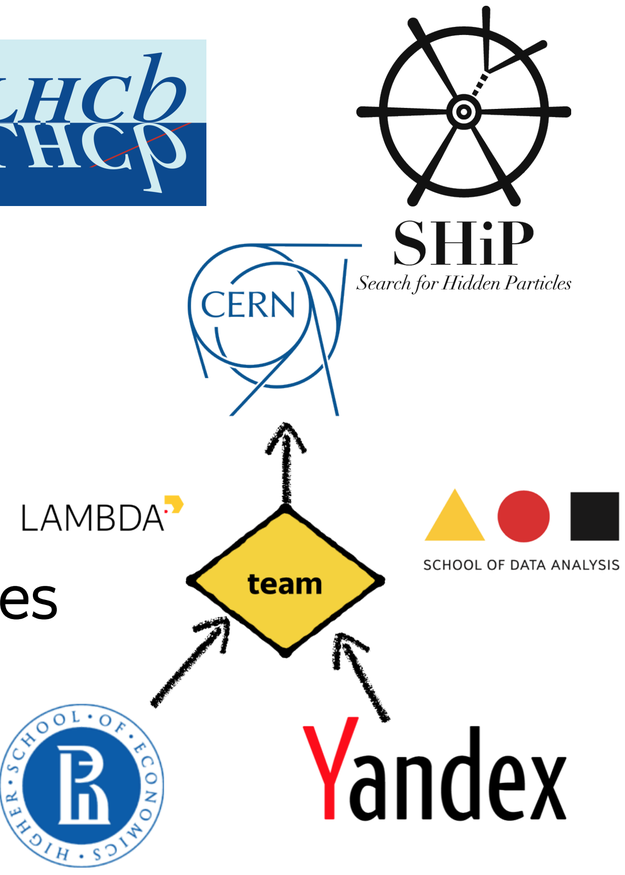
HARDWARE
KNUPATH, TENSTORRENT, Cirrascale, NVIDIA, nervana, Movidius, tensilica, GoogleTPU, 10^28 Labs, Qualcomm

The Gap

ML expertise is needed to deal with data, but
Core of ML expertise is outside of specific science domain

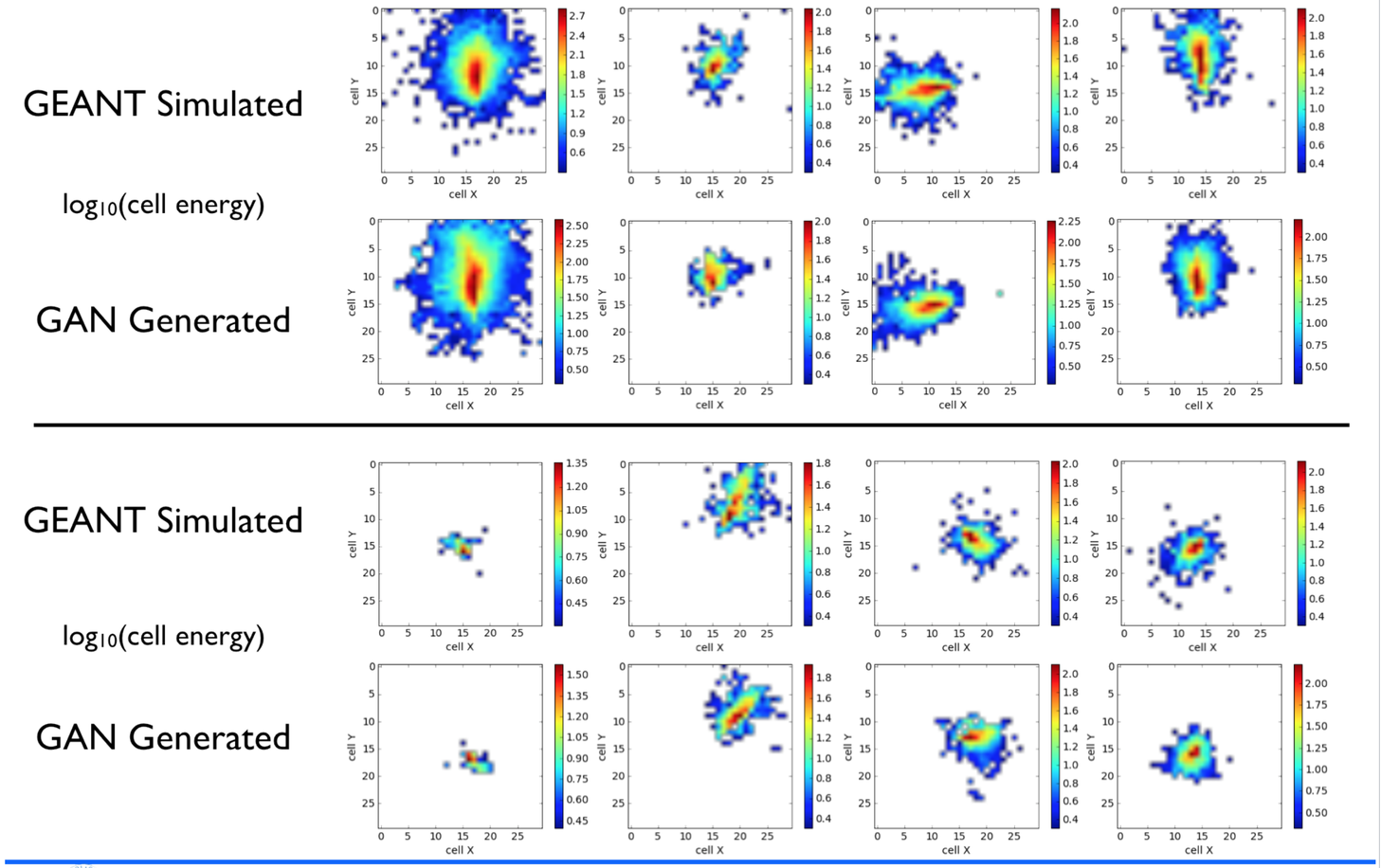
Quick self-intro

- 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 this summer!
- Education activities (MLHEP, ML at ICL,ClermonFerrand, LaSAL, Coursera)
- One of organizers of Flavours of Physics Kaggle competitions (2015)
- One of organizers of TrackML challenge (2018)



Example: Machine Learning for Fast Simulation

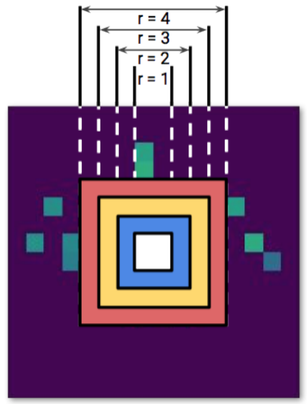
Generator in Full 5D

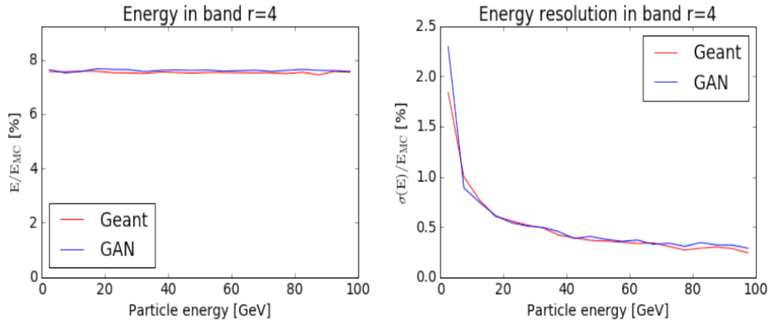


Fedor.Ratnikov@cern.ch

Fast CALO Simulation in LHCb 12 / 11

Non-standard quality metric:





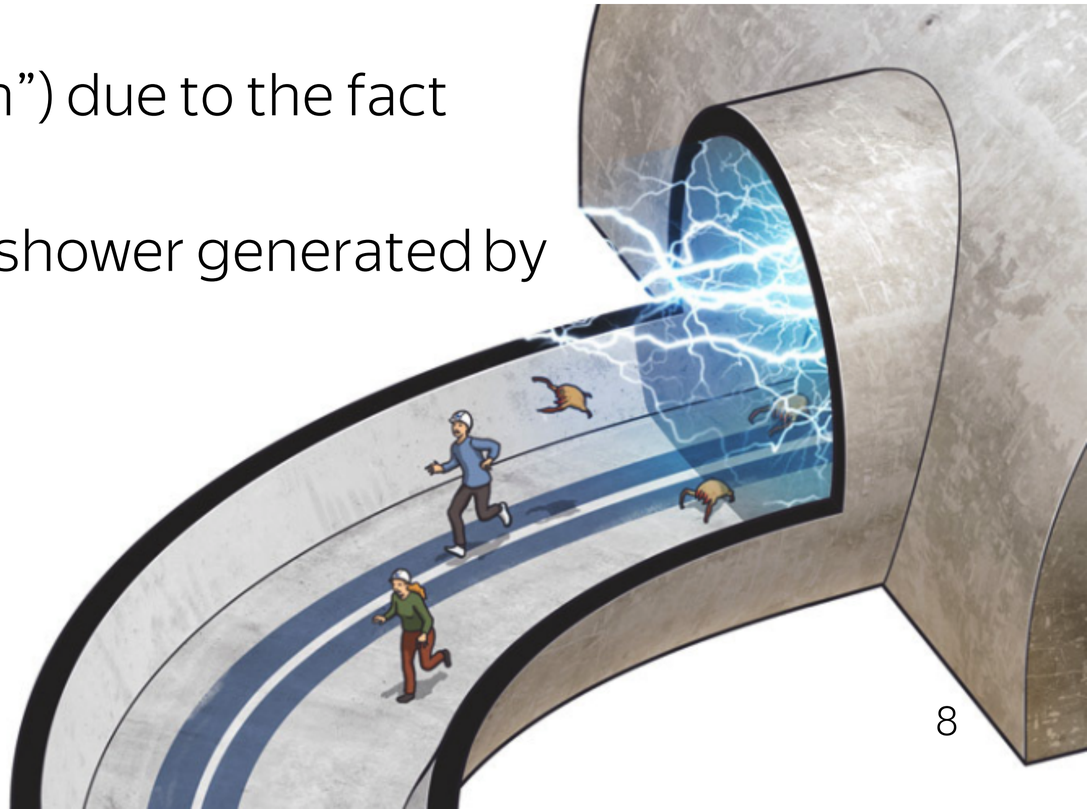
Cross-domain caveats (Particle Physics)

Domain-specific barriers

- Developed terminology and mindset
- Structured and semantically-rich data
- “Weird” constraints (“systematics”, “calibration”) due to the fact that ML part is just a step of a bigger picture
- No obvious metrics for ’sanity’ checks (is a jet/shower generated by NN looks realistic enough?)

Reproducibility/traceability of results

Cross-checks?



How to bridge the gap?

- **Train own expertise**
  - Mission impossible
- **Invite/hire person into the team**
  - Motivation? Domain-specific training? Communication?
- **Collborate with external experienced team (like YSDA, HSE, etc.)**
  - Motivation??? Domain-specific training? Communication?
- **Use crowd “wisdom”**
  - Motivation? Training? Communication? Reproducibility?

DataScience competition: Netflix Prize

Netflix prize – for improving baseline accuracy by 10%, 1M USD

- Training data set of 100,480,507 ratings that 480,189 users gave to 17,770 movies: `<userID, movieID, date, grade>`, where `grades` are from 1 to 5 “stars”
- The qualifying data set: 2,817,131 triplets of the form `<userID, movieID, date>`
- Goal: accurately predict grades on the entire qualifying set:

1. Accuracy for the **quiz** set of (50% of the whole set) is publicly available
2. The other half is the **test** set to identify the winners.

- Quality metric: root MSE between predicted and actual grades
- Baseline: Cinematch (linear model)

https://wiki2.org/en/Netflix_Prize

Netflix Prize timeline

- Aug 2007 – international conference, announcement
- Oct 2007 – BellKor FTW – 8.43% improvement! (among 20k teams)
- Oct 2008 – Big Chaos took lead
- Late Oct 2008 – BellKor + Big Chaos – 9.43% impovement
- June 2009 – BellKor's Pragmatic Chaos – 10.05%
- 26 July 2009 18:18:28 – BellKor's Pragmatic Chaos – 10.09%
- 26 July 2009 18:38:22 – Ensemble – 10.10%

Got same result on the **test**! The prize was awarded to BellKor's Pragmatic Chaos.

Second challenge was cancelled due to privacy concerns.

https://wiki2.org/en/Netflix_Prize

Sources of crowd intelligence

- Participants of Machine Learning (ML) courses, looking for decent problems to test their skills on
  - Low-responsibility contribution
  - Need for computational resources
  - No time/resources for deep problem understanding
  - Hungry for scoring records
- Teams like YSDA, HSE that are interested in extending ML for domain sciences

Demand for a collaborative platform

"Mechanical Turk for (open) science"

Flexibility to define and update challenge (metric, dataset)
Micro-contributions (commits) for

- Tracking meaningful contributions
- Peer-reviews
- Researcher profile

Motivation for micro-contributions (micro-rewards)
Reusable (reproducible) results
Communication (goal, manifest, fast bootstrap)
Global-scale, transparency
Unified hardware access

Research Collaboration Platforms Candidates

- Github (belongs to Microsoft)
  - No reward mechanism, too generic
- Kaggle (belongs to Google)
  - No micro-reward motivation, no contribution-tracking, single metric from pre-defined list, limited flexibility
- CodaLab (https://codalab.org)
  - Allows staged competitions, custom metrics. No micro-reward motivation, no contribution-tracking, no reuse / peer review
- RAMP (https://ramp.studio)
  - Builds ensembles. No easy way to start own compeition. No micro-reward motivation, no user profile.

High-level platform Components

Challenge Directory

Reusable artifacts

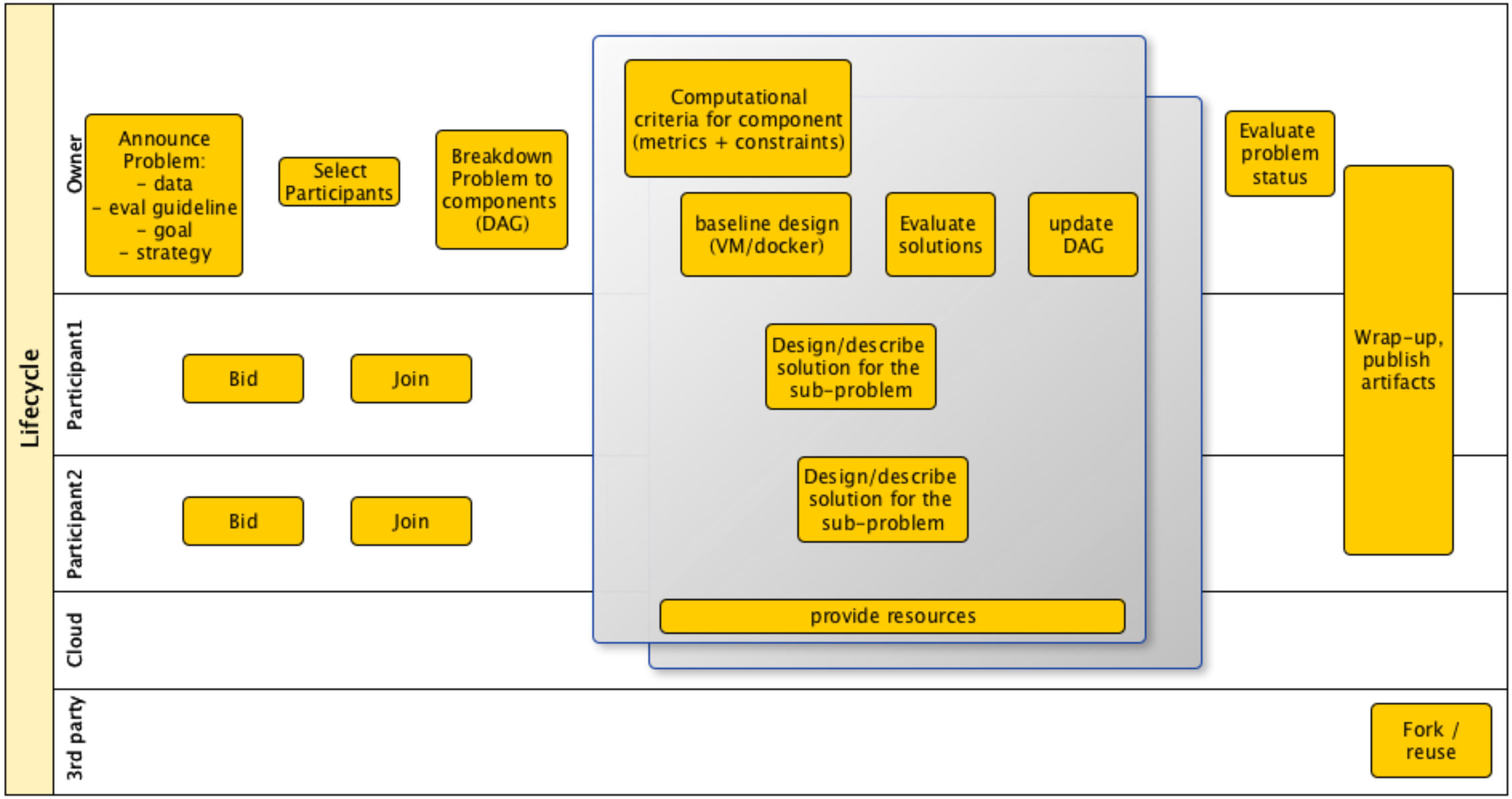
Challenge Owners

Resources (computational, storage)

Participants, communities

Institute / Univsersity

Collaboration Lifecycle



Collaboration artifacts

User profile:

- Track of user commits, linked to metrics improvements
- Track of source-code

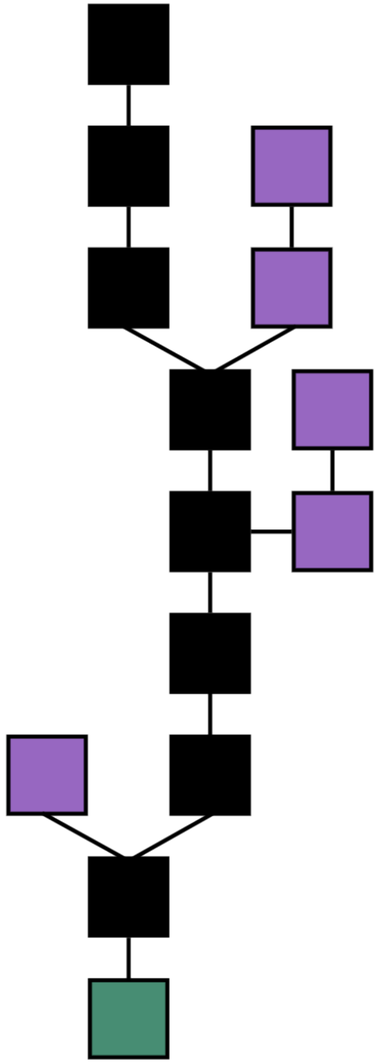
Competition profile:

- Baselie
- Metrics, leaderboard
- Re-usable, reviewable models

What about trust and motivation?

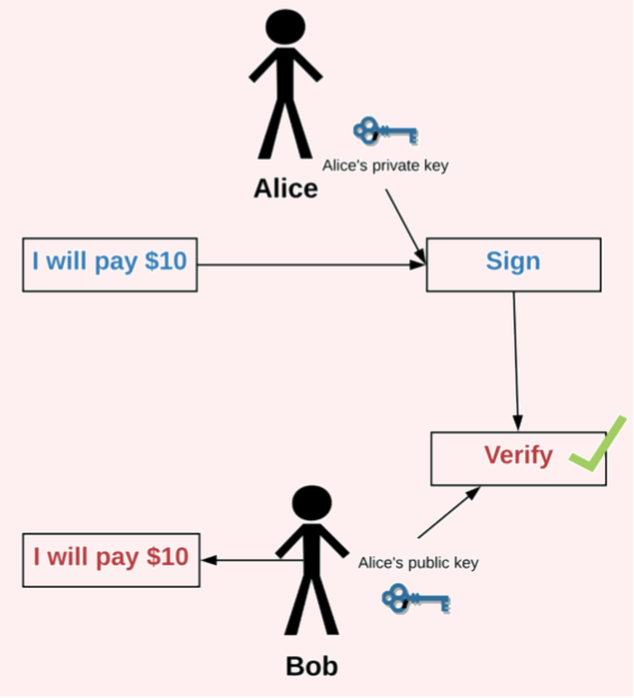
Blockchain - A Distributed Ledger Technology

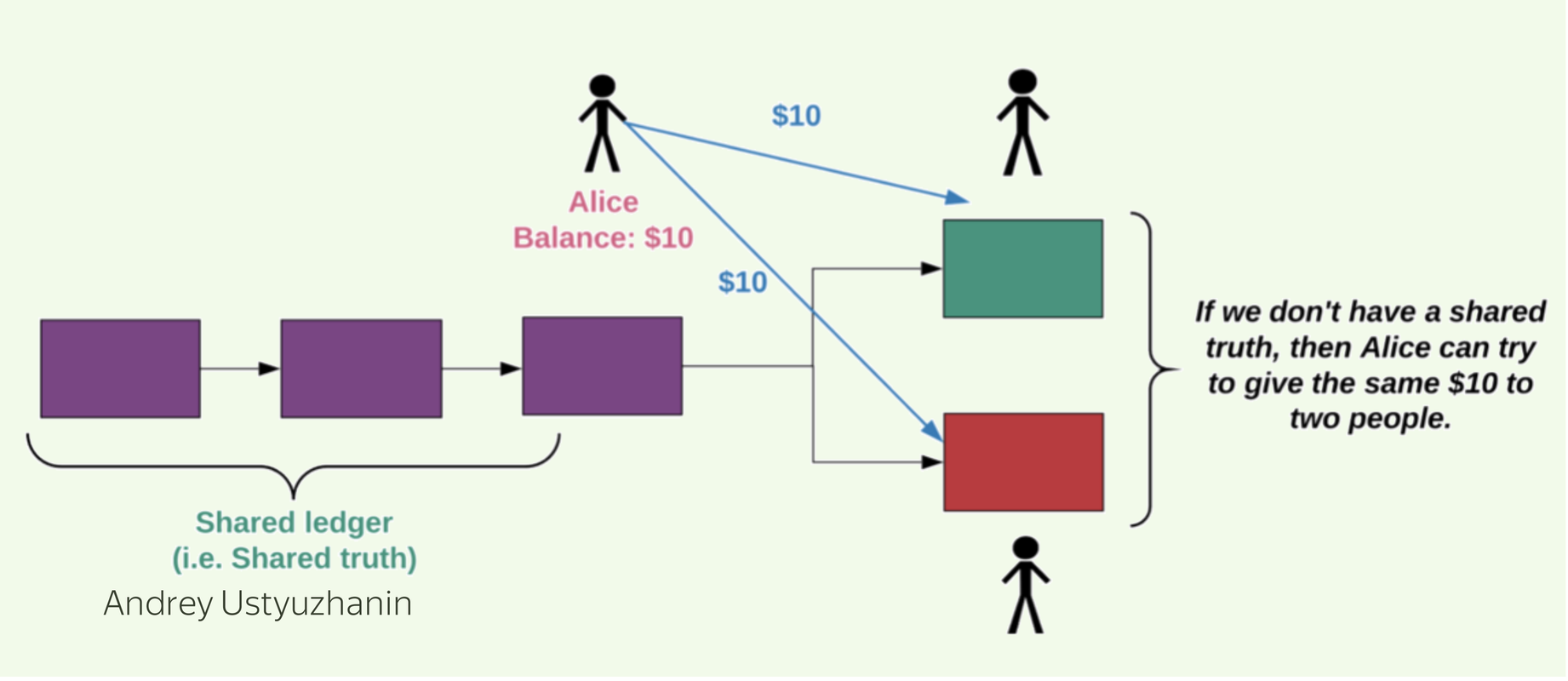
- A blockchain is a linked list where each node is connected to its predecessor by a cryptographic hash
  - All pointing back to the “genesis” block (right, in green) which may contain defining information about the rules for the blockchain protocol
  - In this way a blockchain comprises a verifiable public ledger
- Each node of the linked may contain additional transaction data (verifiable)
- Typically it’s the longest contiguous chain (right, in black) which is considered valid (purple are orphaned blocks)
  - However it’s up to the developers who define the protocol to determine the rules for consensus and evolution of the chain
- A variety of blockchains exist today, some exploring alternative architectures to test multiple aspects of scalability



Blockchain - A Distributed Ledger Technology

- Original purpose of the blockchain:
  - Keep shared (consensus) state of the “truth”
  - For example balance on each participant’s account





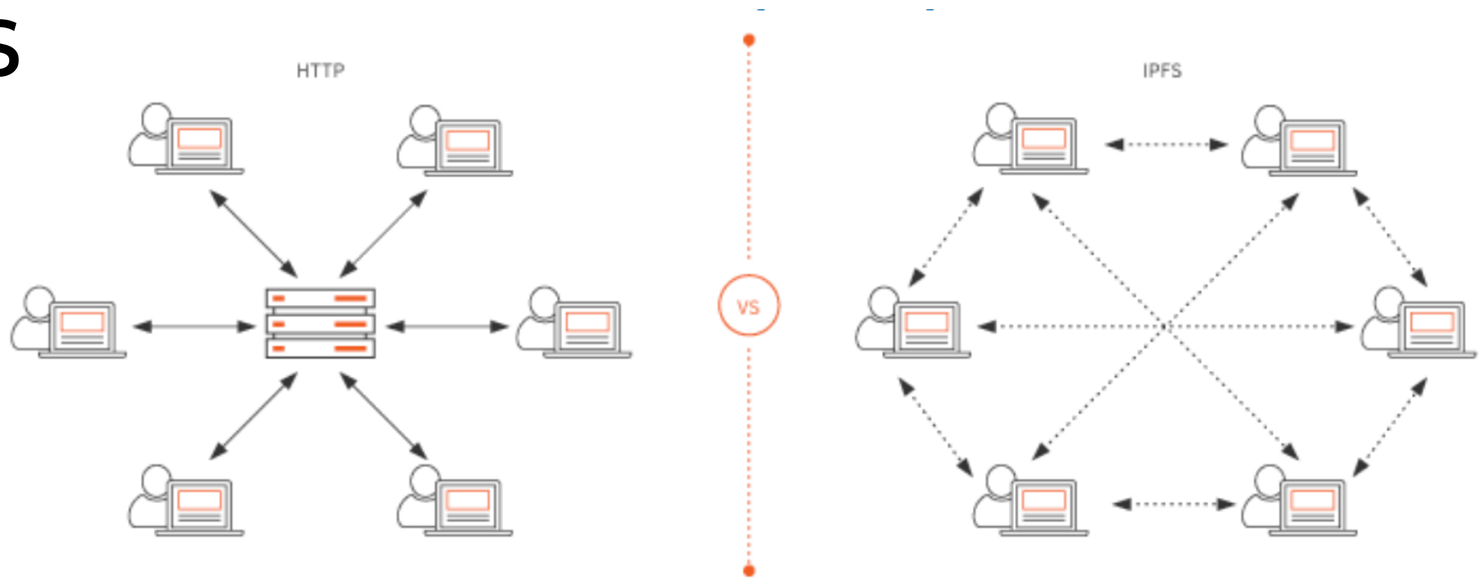
Blockchain – Smart Contract

- Newer blockchains, Ethereum for instance, implement virtual machines that can execute byte code
- Smart contracts, implemented in this code allow binding between blockchain addresses and actions that are taken by the code
  - Typically the same code gets executed by all nodes in the network (extension of Nakamoto consensus)
- This can be used to implement a huge range of tasks
  - sub-currencies
  - timed payments
  - running of mathematical proofs
- Limited by blockchain transaction speed

```
pragma solidity ^0.4.21;

contract Coin {
    // The keyword "public" makes those variables
    // readable from outside.
    address public minter;
    mapping (address => uint) public balances;

    // Events allow light clients to react on
    // changes efficiently.
    event Sent(address from, address to, uint amount);

    // This is the constructor whose code is
    // run only when the contract is created.
    function Coin() public {
        minter = msg.sender;
    }

    function mint(address receiver, uint amount) public {
        if (msg.sender != minter) return;
        balances[receiver] += amount;
    }

    function send(address receiver, uint amount) public {
        if (balances[msg.sender] < amount) return;
        balances[msg.sender] -= amount;
        balances[receiver] += amount;
        emit Sent(msg.sender, receiver, amount);
    }
}
```

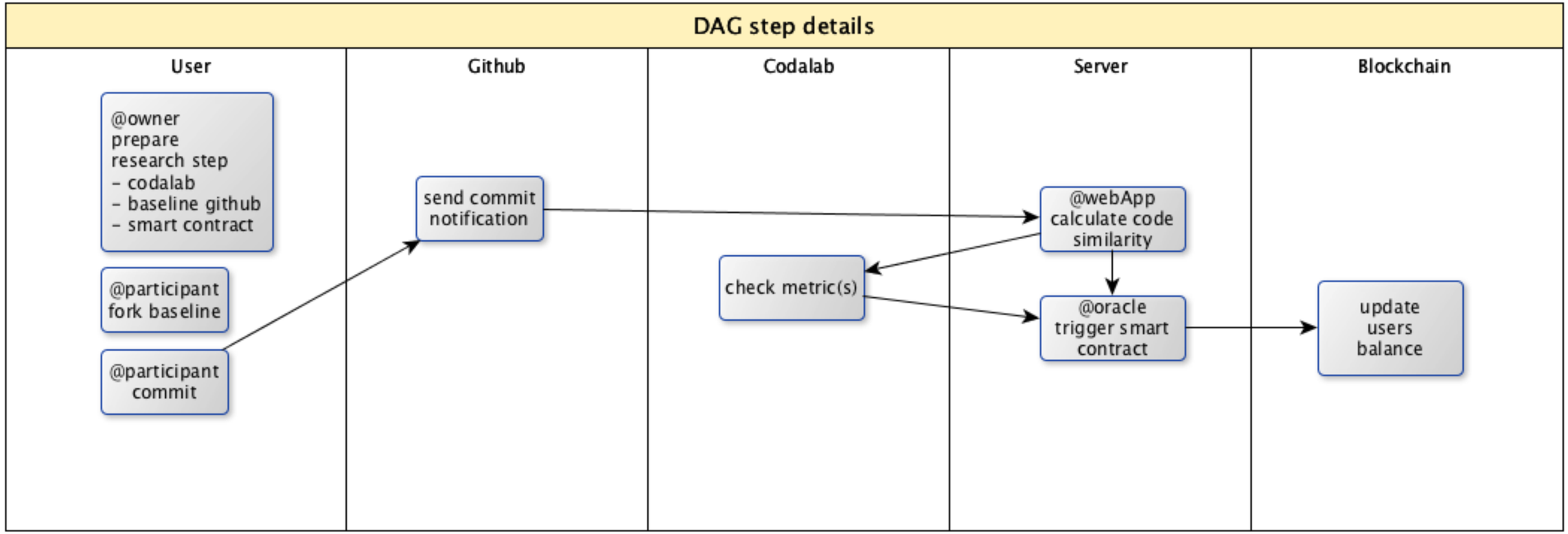
A simple example of a derived currency

Blockchain provides



- Shared state (knowledge)
- Time stamps for commits
- References to artifacts
- Personal portfolio
- Transparent rules from commits to rewards
  - Commit
  - Forks
- Removes bottle-neck and single vendor lock

Possible integration scenario for DAG step

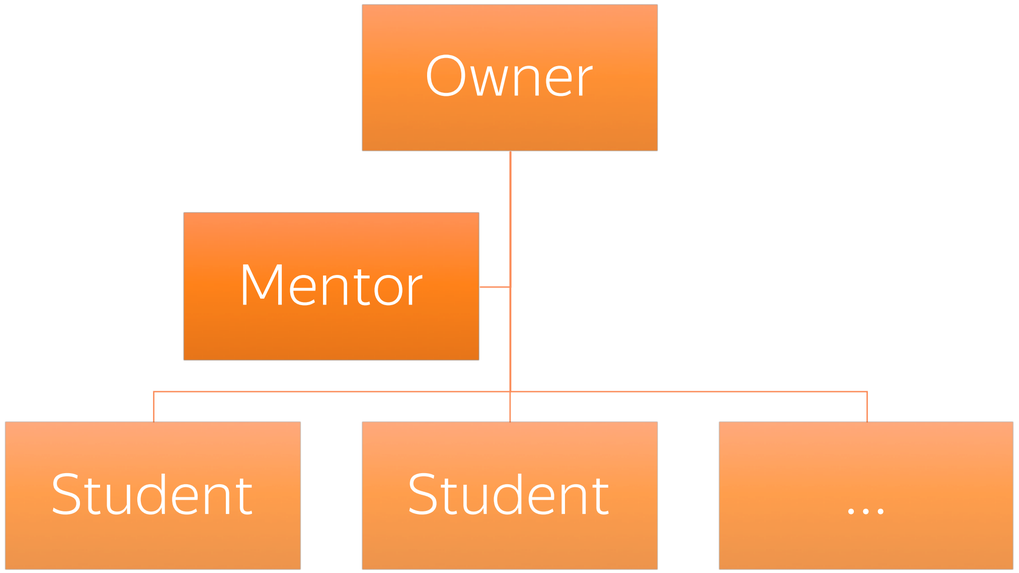


Coopetition Platform for Applied Data Science

- Target audience
  - DS-intensive courses / universities
  - Strudents/practitioners
  - Domain scientists
- Built on top of existing services
  - GitHub, CodaLab, Jupyter, etc
- Motivation for universities
  - Keep student's contribution, more adequate grading
- Motivation for students
  - Mini-grants to participants for computing access
  - Motivation through social dynamics of published code (likes/claps/forks)
  - Mini-grants for participants meeting evaluation criteria
- Motivation for problem owners
  - Many students may eventually improve well-formulated problems

Personal experience in 2017/2018

- Challenges:
  - OPERA e-m shower identification
  - EEG signal compression
  - Calorimeter fast simulation
- Platforms used:
  - Github, Kaggle
- Practices
  - Tickets, time-boxed stages, variable metrics
  - Mediation
- Result: one of the projects has beaten the state of the art
- More Challenges to solve:
  - LHCb data compression
  - LArTPC 3D tracks identification
  - Quantum computer control



Example Challenges for 2018/2019

Our group pipeline

Topics

- LHCb data compression
- LArTPC 3D tracks segmentation
- Quantum computer control
- High-level fast MonteCarlo generation by Neural Networks
- Cherenkov Telescope Array particle identification

Students & Mentors

- 12 – MIPT
- 20 – HSE
- 15 – YSDA

Dark Machines | Data-driven storytelling wit

darkmachines.org Dark Machines

Dark Machines

About News Events Projects Researchers White paper Mailinglist Contribute

About Dark Machines

Dark Machines is a research collective of physicists and data scientists. We are curious about the universe and want to answer cutting edge questions about Dark Matter with the most advanced techniques that data science provides us with.

Visit our indico page

Dark Machines
@dark_machines

The strong lensing subgroup of the DarkMachines project (darkmachines.org) will be holding a kick-off video-meeting for the strong lens challenge on Tuesday, August 7th, 7am PDT (California time).

Aug 3, 2018

Dark Machines Retweeted

Gianfranco Bertone
@gfbertone

Nice summary on @nature of the challenges and opportunities that come with the use of machine learning at the frontiers of particle physics nature.com/articles/s4158…

Machine learning at the energy and intensity frontiers …

Andrey Ustyuzhanin

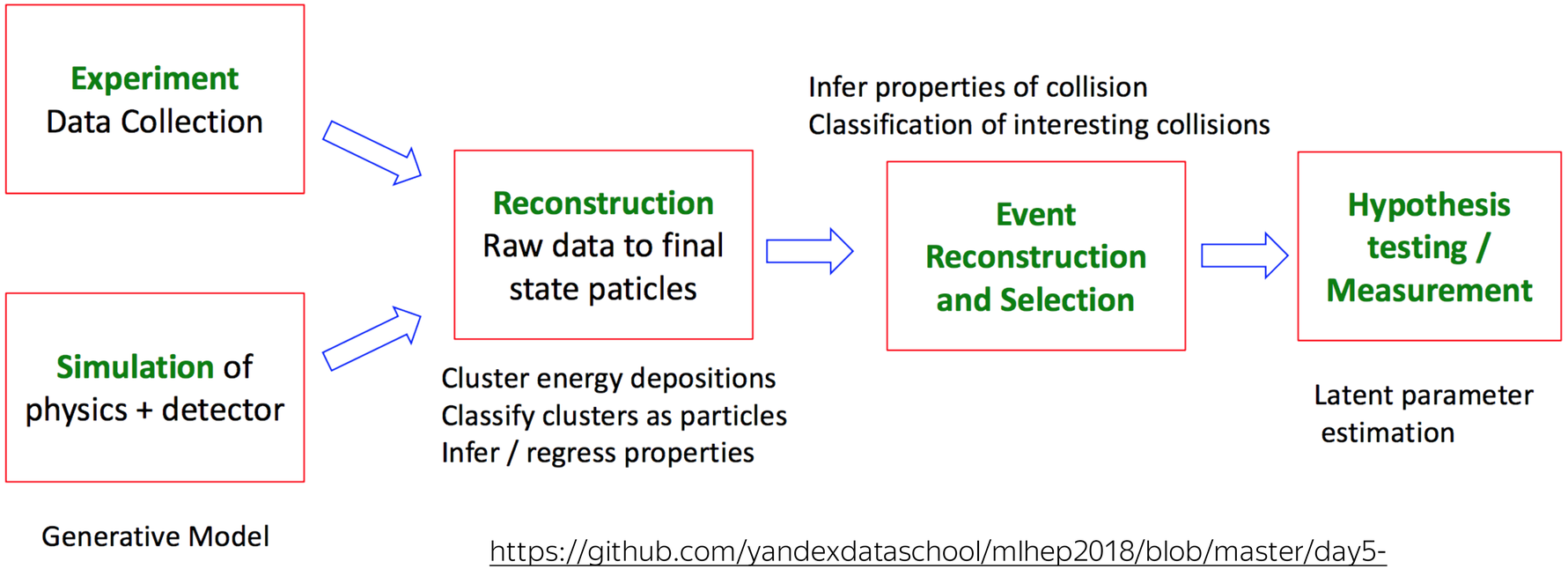
http://darkmachines.org/

DarkMachines projects

- Particle track reconstruction with ML
- Inclusive analysis of Fermi-LAT point sources
- Exploiting the full information on DM signals contained in multi-wavelength and multi-messenger observations
- Indirect detection & unsupervised learning
- Strong lensing & unsupervised learning
- Collider searches & unsupervised: or supervised or not-yet-thought-off learning
- Learning dark matter distributions in galaxies

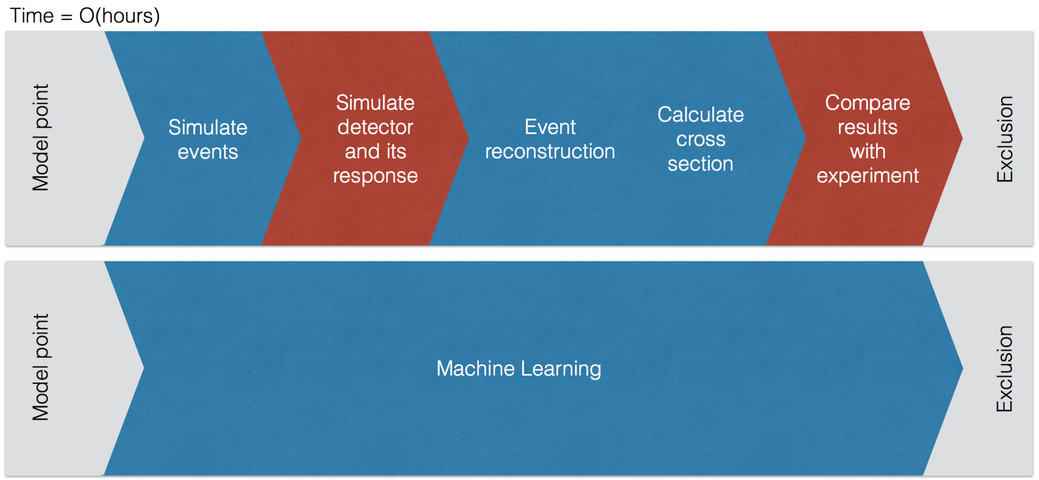
http://darkmachines.org/

More ideas for collaboration



- Triggers
- Tracking
- Object identification
  - Particle, showers, jets
- Fast MC generation
- Model checking
- Detector design optimization

https://github.com/yandexdataschool/mlhep2018/blob/master/day5-Sat/bartunov_few_shot_learning_ebook.pdf



Q&A for Domain Research

- Would you outsource a challenge to such a platform?
  - Does research goal look big/ambitious?
  - Do you have enough resources to solve it yourself?
  - Do you have a dataset? (simulated would work, or generator itself)
  - Can you express the challenge quality as a computable metric?
- Would you like to collaborate with unknown researchers on it? And even publish a joint paper with them?
- Are there people in your team willing to guide/communicate newcomers?

https://goo.gl/forms/P9OIjOfW1FcRbRIJ3

Wait, there is more

- Testbench for solutions for common domain problems
  - Jet identification,
  - B-tagging,
  - Tracking
- Evolution of metric/motivation design: the best way writing smart-contracts?
  - Increase of metrics?
  - Metric hacking?
  - Popular commit?
- Collect statistics of humans dealing with problems for training ML algorithm for automated improvements
- Call for proposal: INFRAEOSC-02-2019, “Prototyping new innovative services”

Conclusion & Focus points

- **Plenty of cool stuff is driven by data in Science**
  - in fundamental and applied sciences
  - ...where Machine Intelligence can help
- **Machine Intelligence field is growing exponentially**
  - New algorithms and methods, infrastructure
  - Driven by industry
- **To bridge the gap: demand for platform!**
  - Can be built on existing well-adopted services (i.e. github, codalab)
  - Should be flexible to support variety of processes used in scientific domains
  - Well-aligned with Open Science values
- **You are welcome to join and try!**

http://cs.hse.ru/lambda/en
anaderiRu@twitter
austyuzhanin@hse.ru

Backup

References

James Surowiecki, The Wisdom of Crowds, 2004
https://www.scienceroot.com/#science
https://indico.cern.ch/event/700917/
https://osf.io/
https://www.topcoder.com/
https://www.nature.com/articles/d41586-017-08589-4
https://www.nature.com/articles/s41586-018-0361-2
https://www.blockchainforscience.com/
https://www.theatlantic.com/science/archive/2018/04/the-scientific-paper-is-obsolete/556676/
https://distill.pub/
https://blog.acolyer.org/2018/03/30/the-surprising-creativity-of-digital-evolution/

Collaboration Highights

- Preparation-stage
  - Define the case goal(s), make it as independednt as possible
  - Specify reasoning model, make it as clear as possible
  - Produce dataset(s), describe the structure
  - Produce evaluation baseline
- Research-iterations
  - Describe Figures of Merit (FOM) and constraints clearly
  - Be comfortable with FOM evolution, repeat in cycles (sprints)
  - Cycles are time-boxed
  - For solution preparation and evaluation external resources are needed
- Wrap-up stage
  - Publish reusable artifacts + result communication
  - Generate track record for *each participant*, estimate impact of each contribution

Abridged history of Eductaion system

- **1000+ years – elite**
  - hollistic
- **200+ years – public**
  - Funded by state (from taxes)
  - Industry-oriented
  - There are life-long paths to take
- **10+ years – online**
  - Individual (no batches)
  - Limited practice
  - Limited credibility

Divergent thinking

http://bit.ly/2vzlIWT

Divergent thinking

http://bit.ly/2vzlIWT

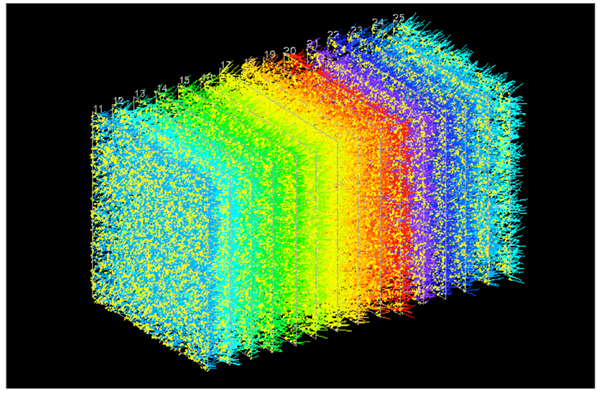
Examples of citizen-science collaborations

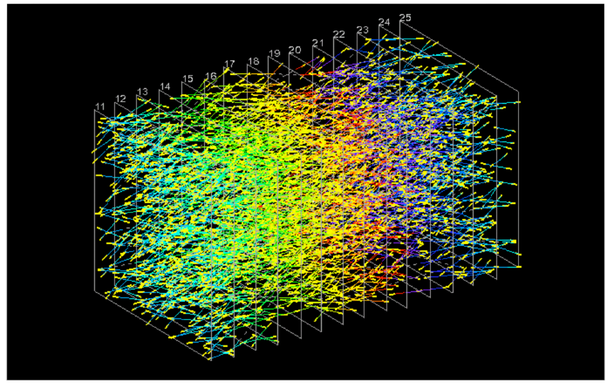
- Linux Kernel
- Galaxy Zoo – finding galaxy rotation pattern
- FoldIt – finding protein shape as a game
- Tim Gower's Polymath
- InnoCentive - https://www.innocentive.com/resources-overview/whitepapers/

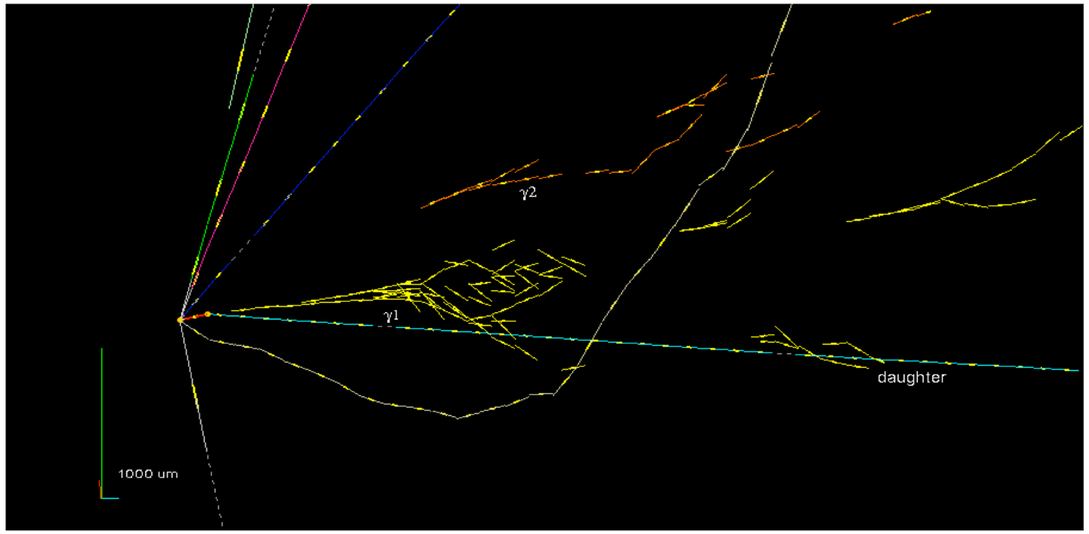
One more trend in Science

- Factors
  - Reduced research funding
  - Higher enternace barriers
  - Higher interest in research for amateurs
- Demand:
  - Communication media for collaboraiton

Case: OPERA em-showers identification

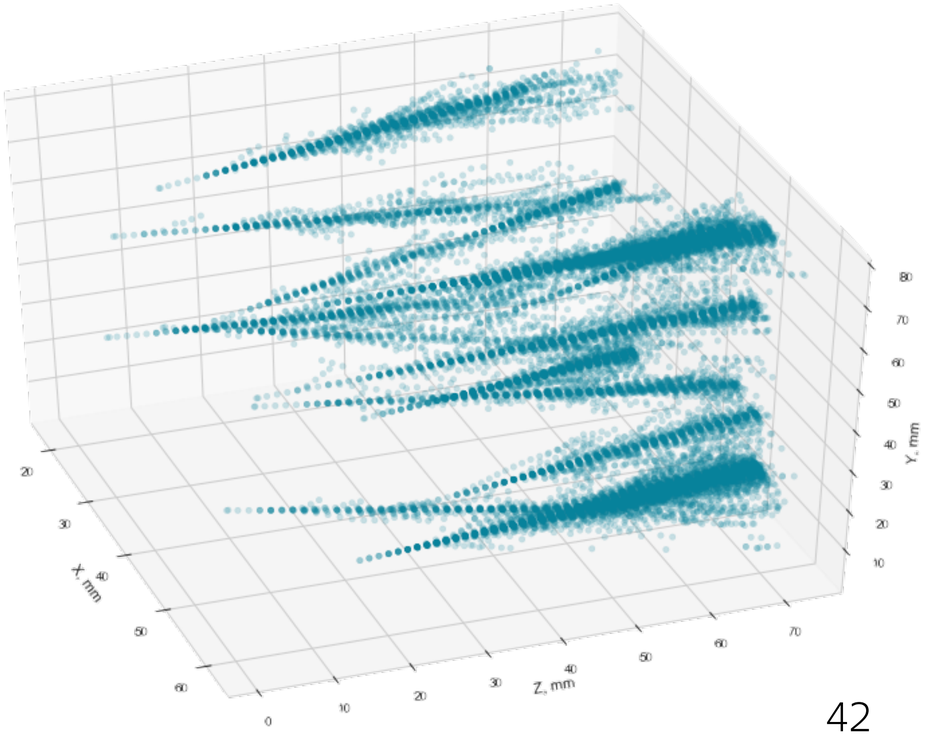






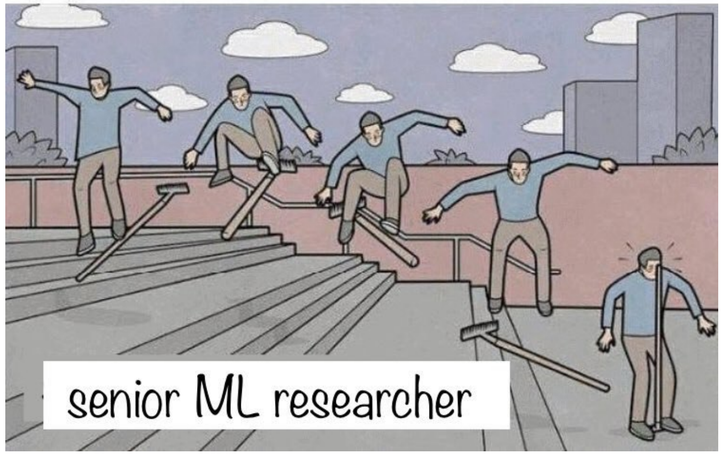
Metric: energy resolution, can be approximated by precision/recall

Difficulties: overlapping showers



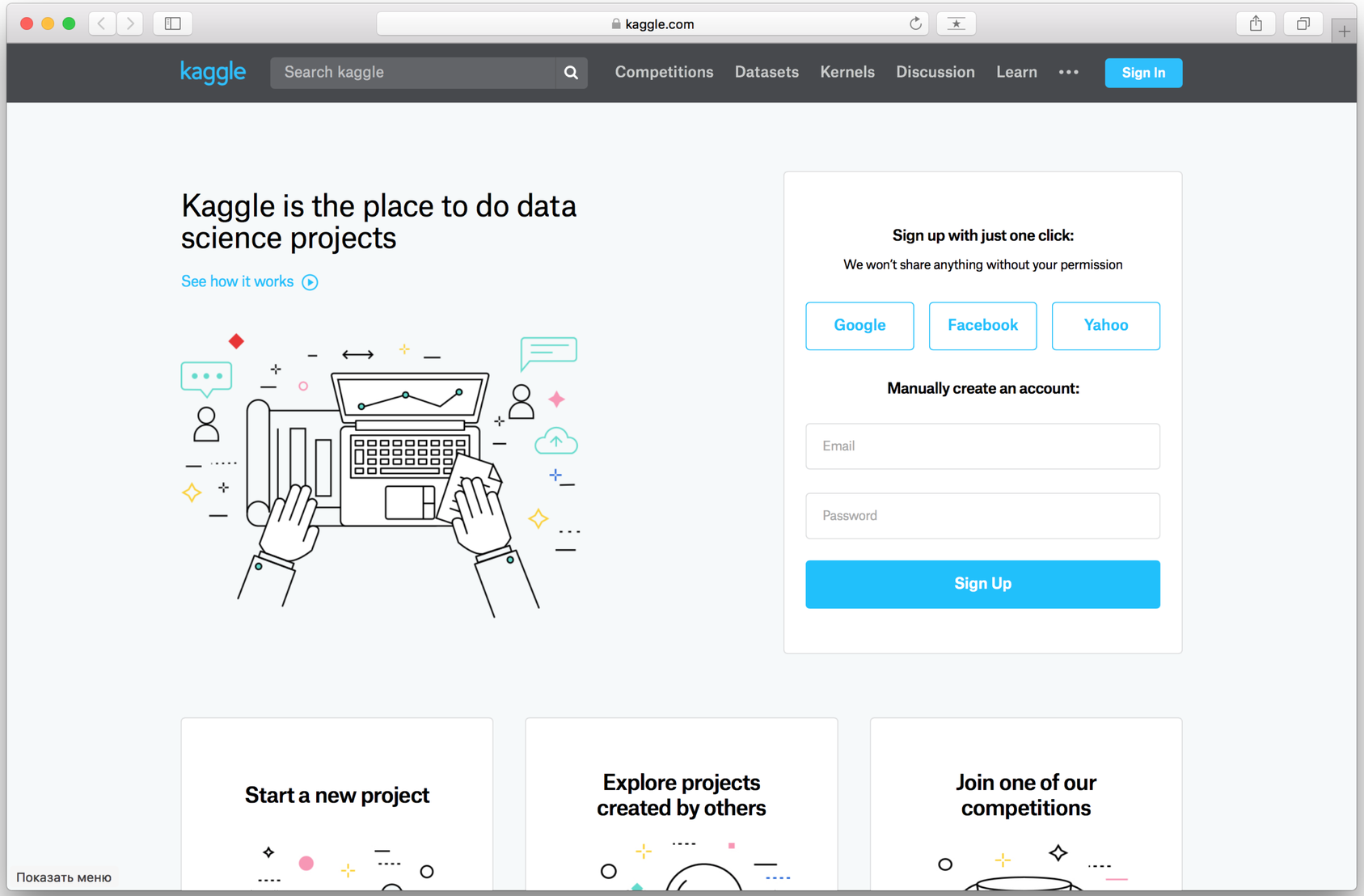
Collaboration with Data Science (DS)

- There is a plenitude of methods that has been developed in ‘data science’ and ‘deep learning’ fields during last 5-7 years
- Those are mainly developed by industry (Google, Apple, Facebook, Amazon, …)
- Domain science researches do not necessarily have required skills and background to properly adapt those methods (High Energy Physics, Astro Physics, Neuroscience, etc)
- Industry or Academic data scientists are eager to help, but sometimes it is difficult to cope with domain specificity



Maybe we could harness a fraction of the crowd intelligence?

Wishful Thinker

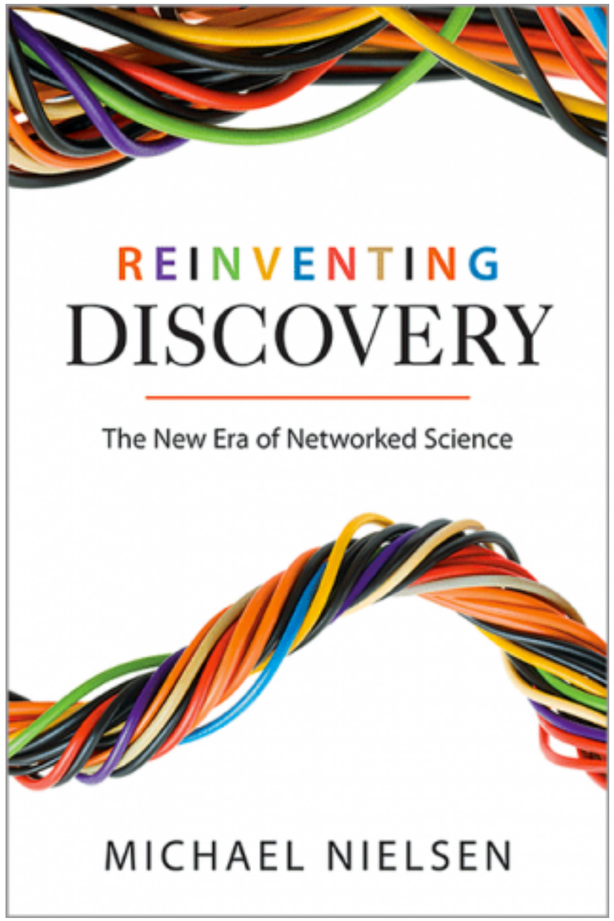


$O(10^4)$ public datasets
$O(10^3)$ competitions
$O(10^6)$ users
$O(10^9)$ submissions

http://bit.ly/2JDMo8j

Successful Citizen-Science project check list

- Clear goals, context and ambitions
  - marketing
- Explanatory materials, methodological manifest, research protocol/conventions
- If you want to eat an elephant do it one bite a time
  - Split big goal in feasible steps
- Participant's motivation even for weakly involved ones
- Specialist attention focus at percise moments
  - Progress announcemnts
  - Short contribution check cycle
- Check or reuse artifacts created by other participants



Michael Nielsen, Reinventing the Discovery, 2014