Networked data-science for research, academic communities and beyond

Andrey Ustyuzhanin

NRU HSE

YSDA

ICL

Abridged history of Science

1000+ years - empirical (Aristotle, Democritus,)

100+ years – theoretical (Newton, Kepler,)

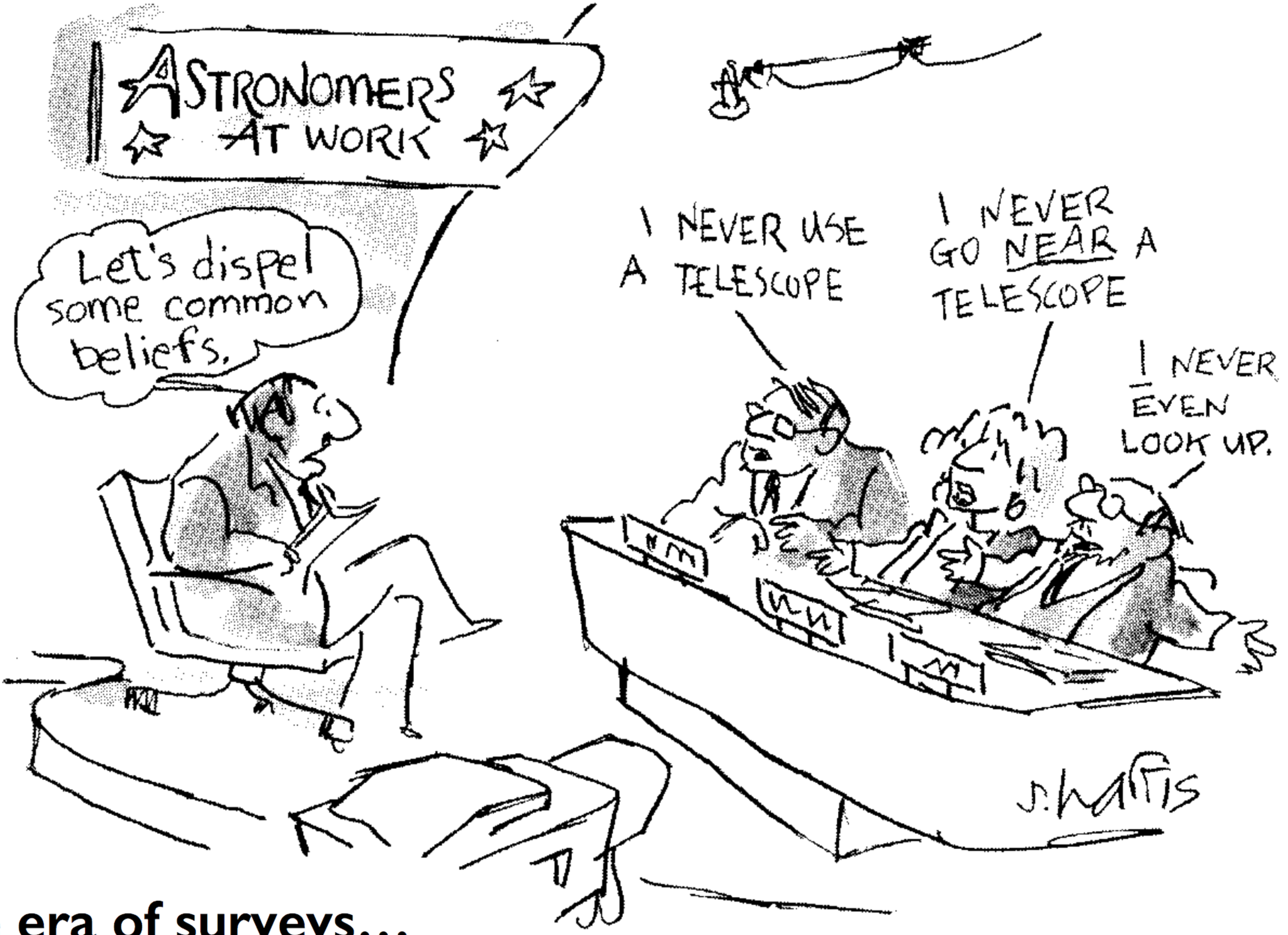
50+ years – computational (John von Neumann,)

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 algocian
SPACE_KNOW Captricity
netra deepomatic

CAPIO Expect Labs
Clover Mobvoi
Qurious.AI pop up archive

UPTAKE imubit Preferred Networks
thingworx KONUX Alluvium

Alation sapho Outlier
Digital Reasoning

Bottlenose MOTIVA
enigma CB INSIGHTS
Tracxn predata

DOMINO SPARKBEYOND rapidminer
kaggle DataRobot yhat AYASDI
dataiku seldon yseop bigml

MACHINE LEARNING

CognitiveScale GoogleML context relevant
Cycorp HyperScience nara logics minds.ai H2O.ai
SCALED INFERENCE sparkcognition loop GEOMETRIC INTELLIGENCE
deepsense.io reactive skymind bonsai

NATURAL LANGUAGE

agolo AYLIEN LEXALYTICS
Narrative Science loop ai labs 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
H2O DEEPLEARNING4J theano torch
DSSTNE Scikit-learn AzureML neon
MXNet DMTK Spark PaddlePaddle WEKA

HARDWARE

KNUPATH TENSTORRENT Cirrascale
NVIDIA intel nervana Movidius
tensilica GoogleTPU 10^{18} Labs Qualcomm

ENTERPRISE FUNCTIONS

CUSTOMER SUPPORT

DigitalGenius Kasisto
ELOQUENT Wise.io
ACTIONIQ zendesk
Preact CLARABRIDGE

SALES

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

MARKETING

MINTIGO Lattice RADIUS
Liftigniter [PERSADO]
brightfunnel retention science
COGNICOR AirPR mogul

SECURITY

CYLANCE DARKTRACE
ZIMPERIUM deepinstinct
Sentinel DEMISTO
graphistry drawbridge
SignalSense AppZen

RECRUITING

textio entelo
Wade & Wendy hiQ
unitive SpringRole
GIGSTER HireVue

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 ROBOTICS fetch robotics
KINDRED
HARVEST AUTOMATION rethink robotics

AGENTS

PERSONAL

amazon alexa
Cortana Allo
facebook M
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
TerrAvion AGRI-DATA
Descartes Labs udio abundant robotics

EDUCATION

KNEWTON volley
gradescope
CTI coursera
UDACITY altschool

INVESTMENT

Bloomberg sentient
iSENTIUM KENSHO
alphasense Dataminr
CEREBELLUM CAPITAL Quandl

LEGAL

blue J BEAGLE
Everlaw RAVEL
seal ROSS
LEGAL ROBOT

LOGISTICS

NAUTO Acerta
PRETECKT clearmetal
Routific
MARBLE PITSTOP

INDUSTRIES CONT'D

MATERIALS

zymergen Citrine
Eigen Innovations

RETAIL FINANCE

TALA zest finance
Lenddo earnest

HEALTHCARE

PATIENT

PULSE CareSkore
ZEPHYR IBM Watson

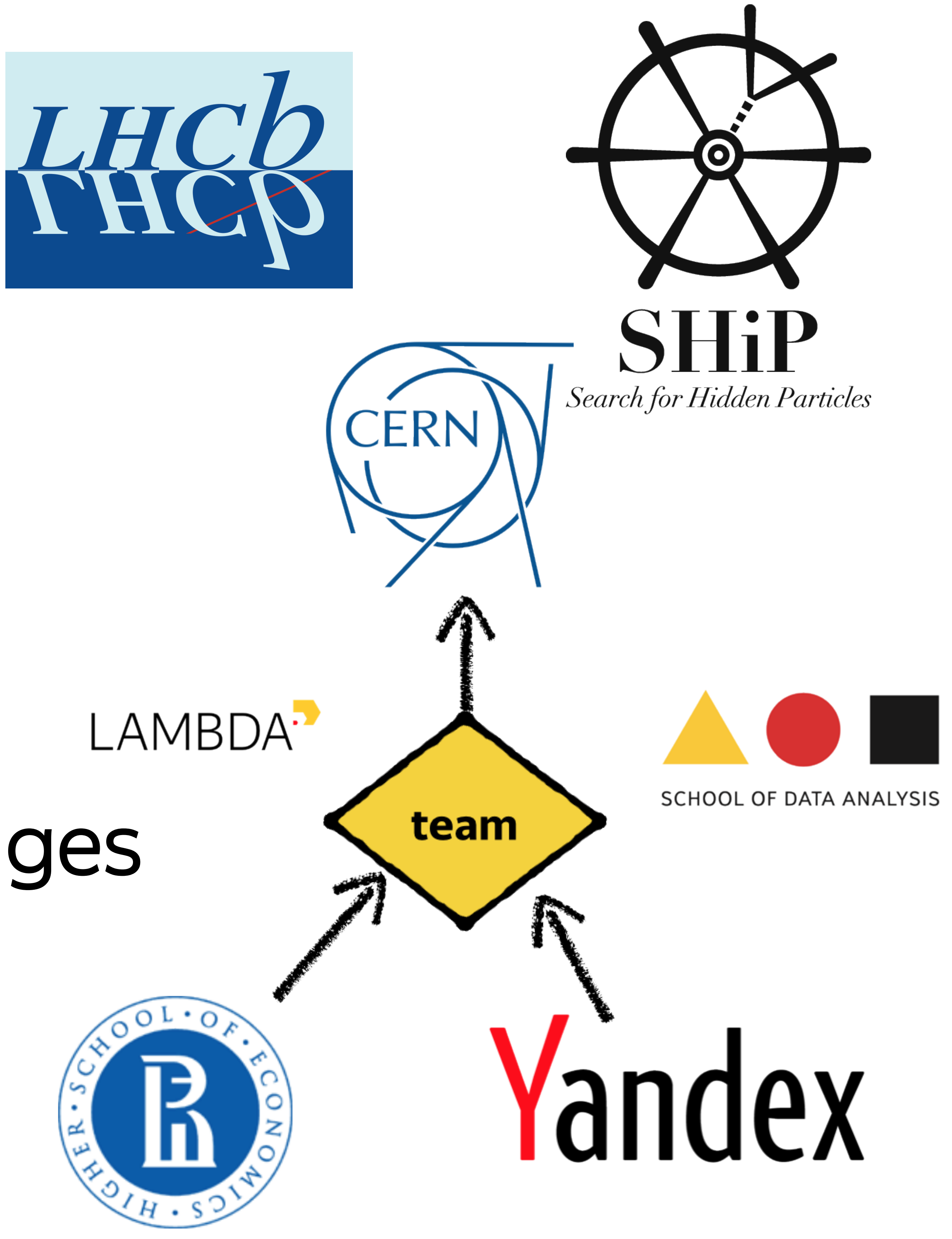
IMAGE

BUTTERFLY 3SCAN
ARTERYS enlitic

BIOLOGICAL

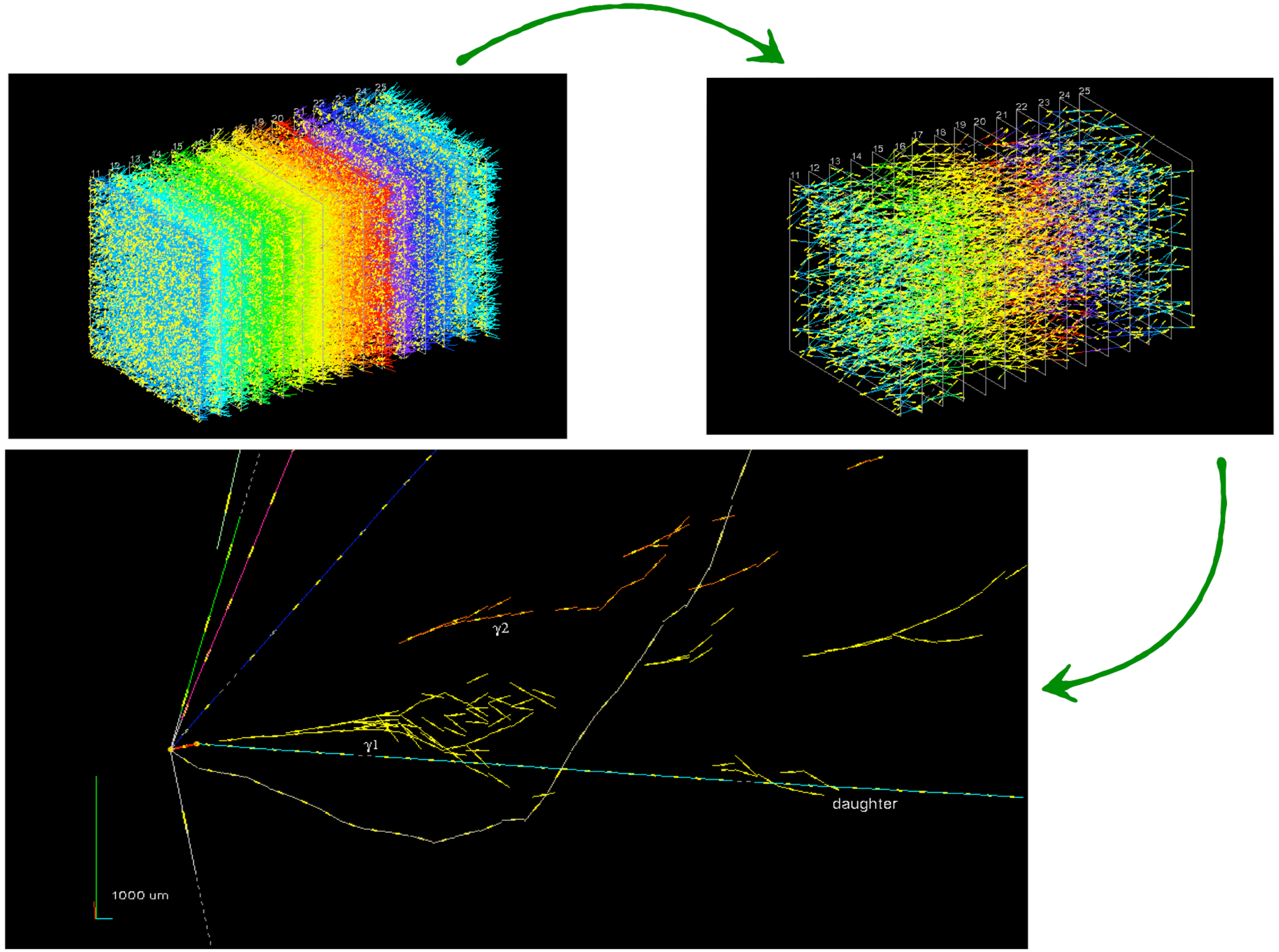
iCarbonX color GRAIL
deep genomics RECURSION

Quick self-intro



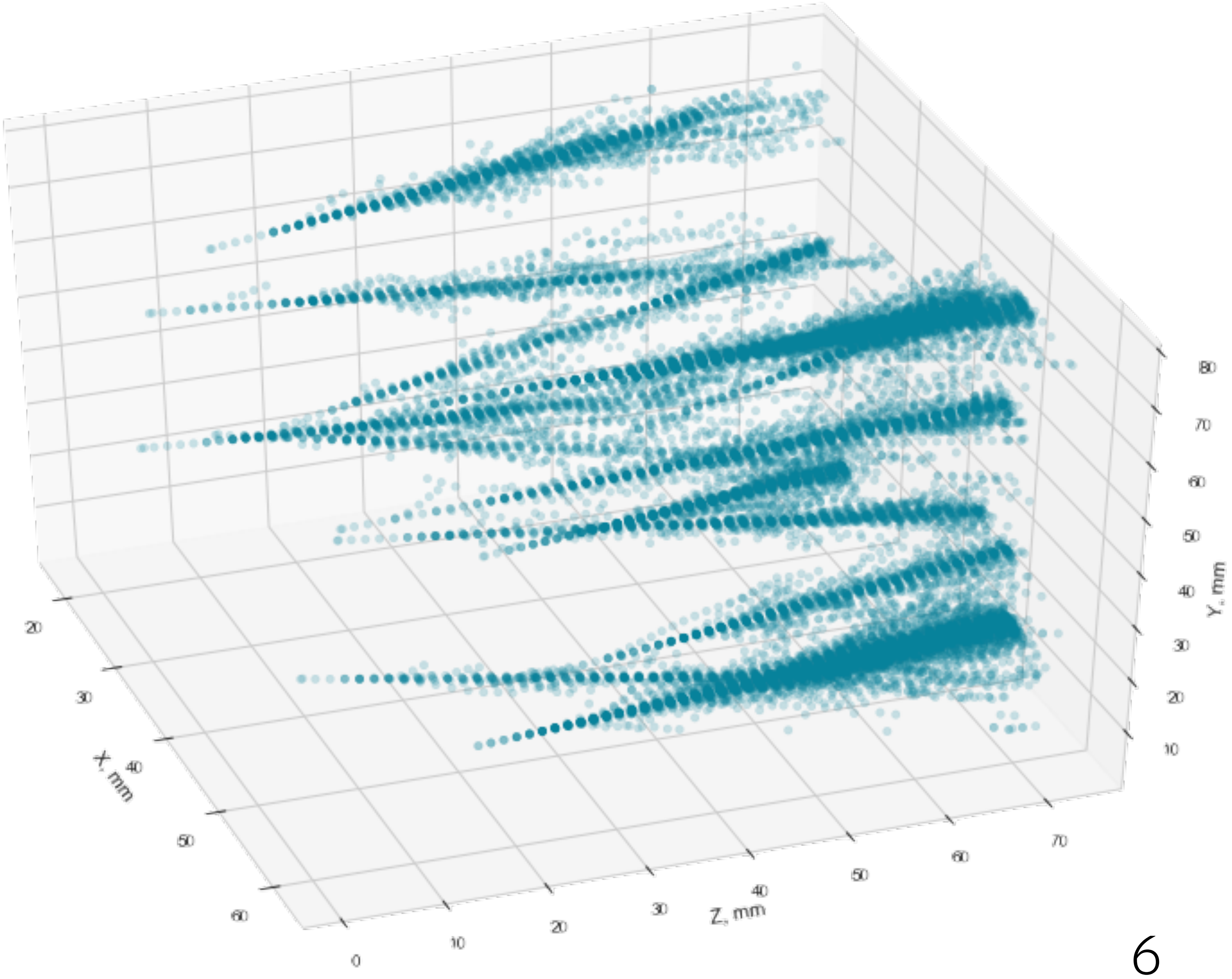
- Head of LHCb Yandex School of Data Analysis (YSDA) team
- Head of Laboratory (link) of methods for Big Data Analysis at Higher School of Economics (HSE),
  - Applications of Machine Learning to **natural science challenges**
  - HSE has joined LHCb this summer!
- Education activities (MLHEP, ML at ICL,ClermonFerrand, LaSAL, Coursera)
- One of organizers of Flavours of Physics Kaggle competitions (2015)
- One of organizers of TrackML challenge (2018)

Case: OPERA em-showers identification

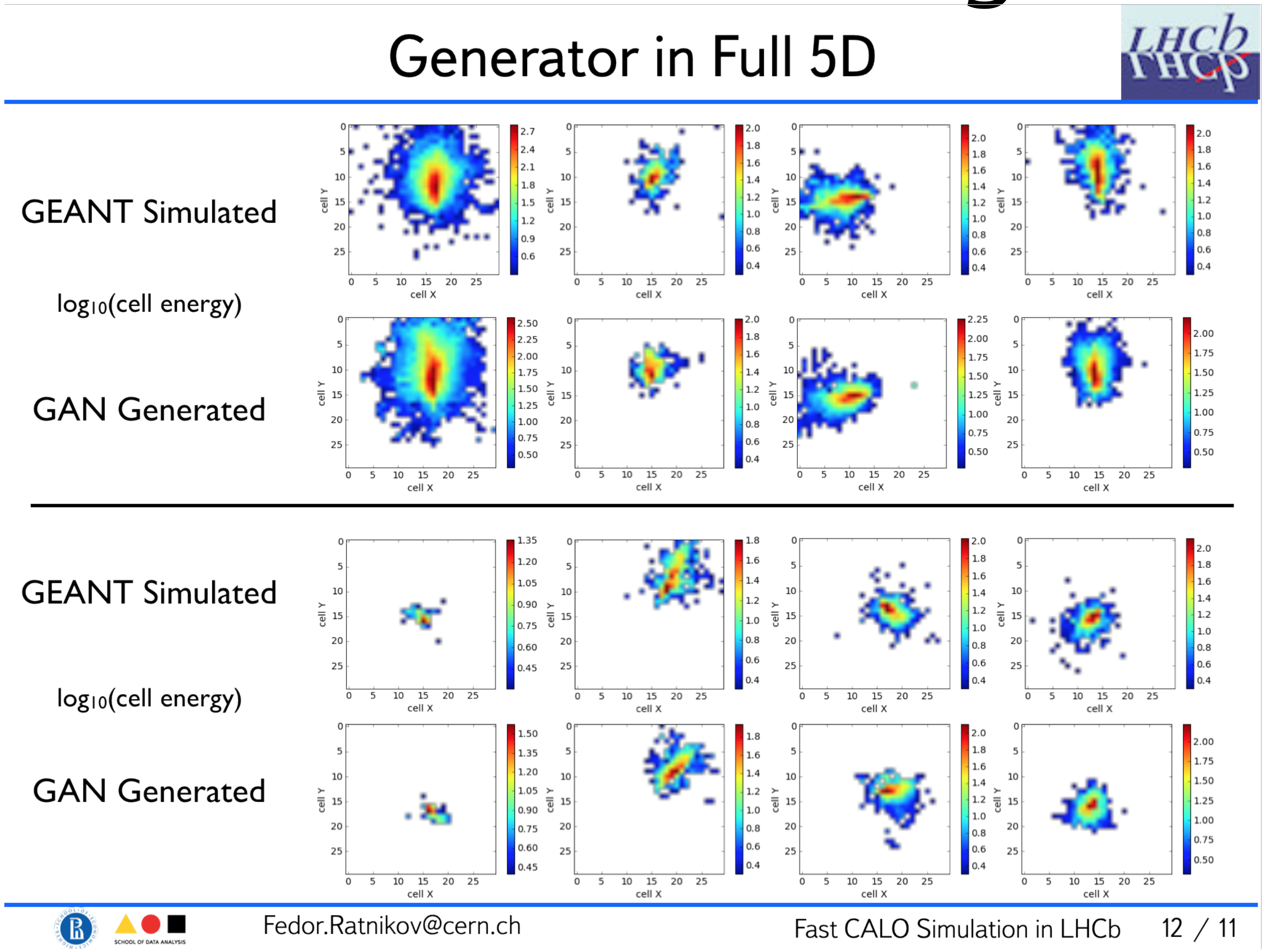


Metric: energy resolution, can be approximated by precision/recall

Difficulties: overlapping showers



Case: Machine Learning for Fast Simulation



Generator in Full 5D

GEANT Simulated

log10(cell energy)

GAN Generated

GEANT Simulated

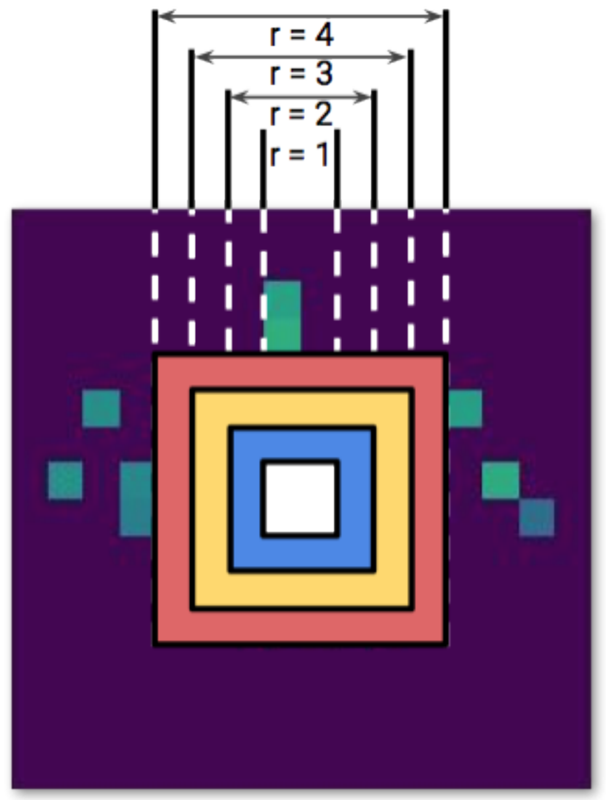
log10(cell energy)

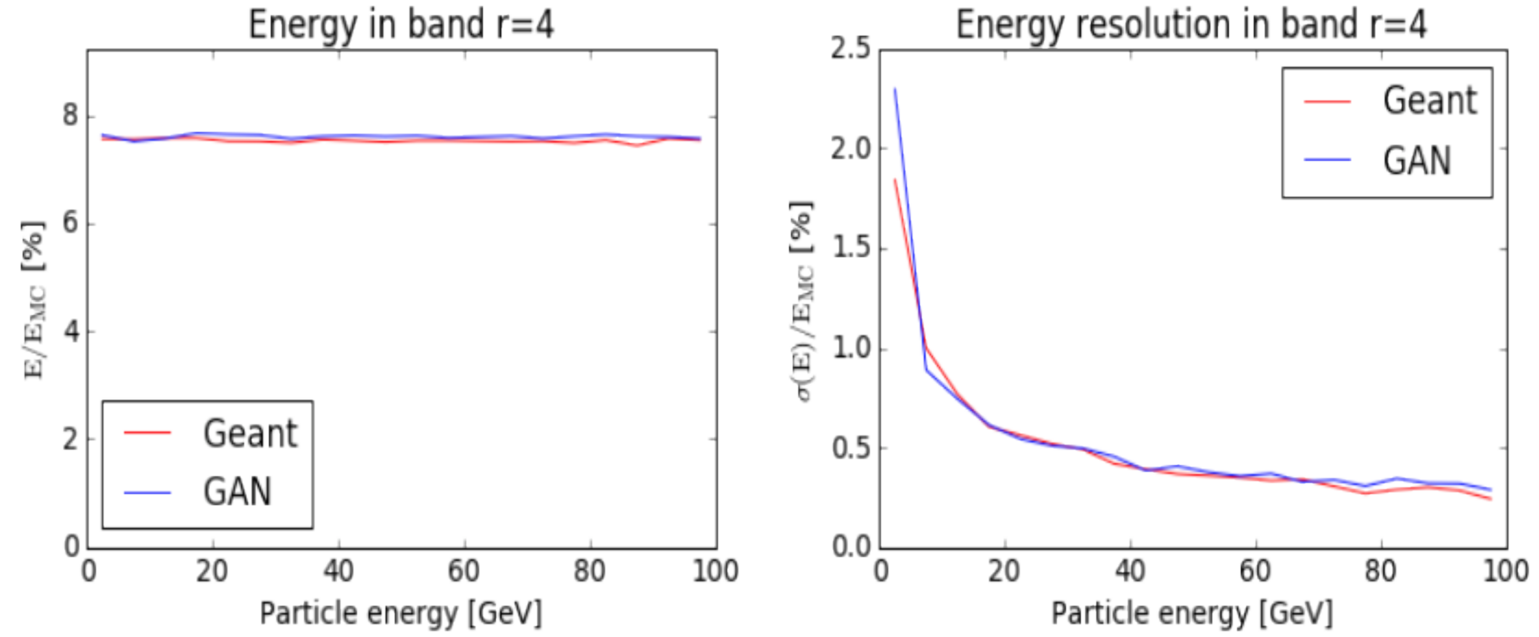
GAN Generated

Fedor.Ratnikov@cern.ch

Fast CALO Simulation in LHCb 12 / 11

Non-standard quality metric:





How to bridge the gap?

- Invite/hire person into the team
  - Grant? Motivation? Training?
- Collborate with external experienced team (like YSDA, HSE)
  - Motivation?
- Use crowd “wisdom”
  - Motivation? Transparency? Training? Communication?

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

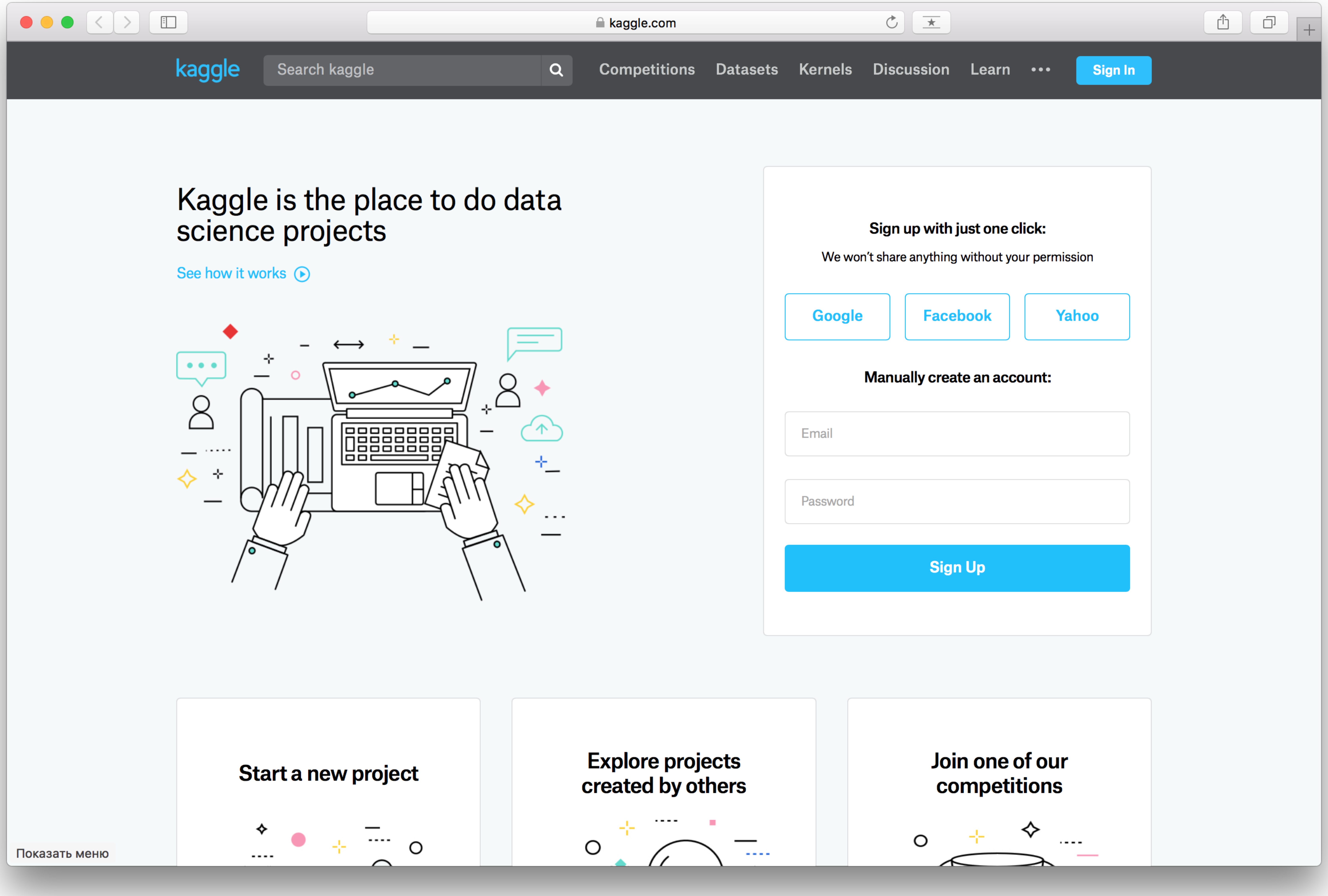
Netflix Prize timeline

- 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



Kaggle is the place to do data science projects

See how it works

Sign up with just one click:

We won't share anything without your permission

Google

Facebook

Yahoo

Manually create an account:

Email

Password

Sign Up

Start a new project

Explore projects created by others

Join one of our competitions

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

http://bit.ly/2JDMo8j

> Maybe we could harness
> a fraction of
> the crowd intelligence?

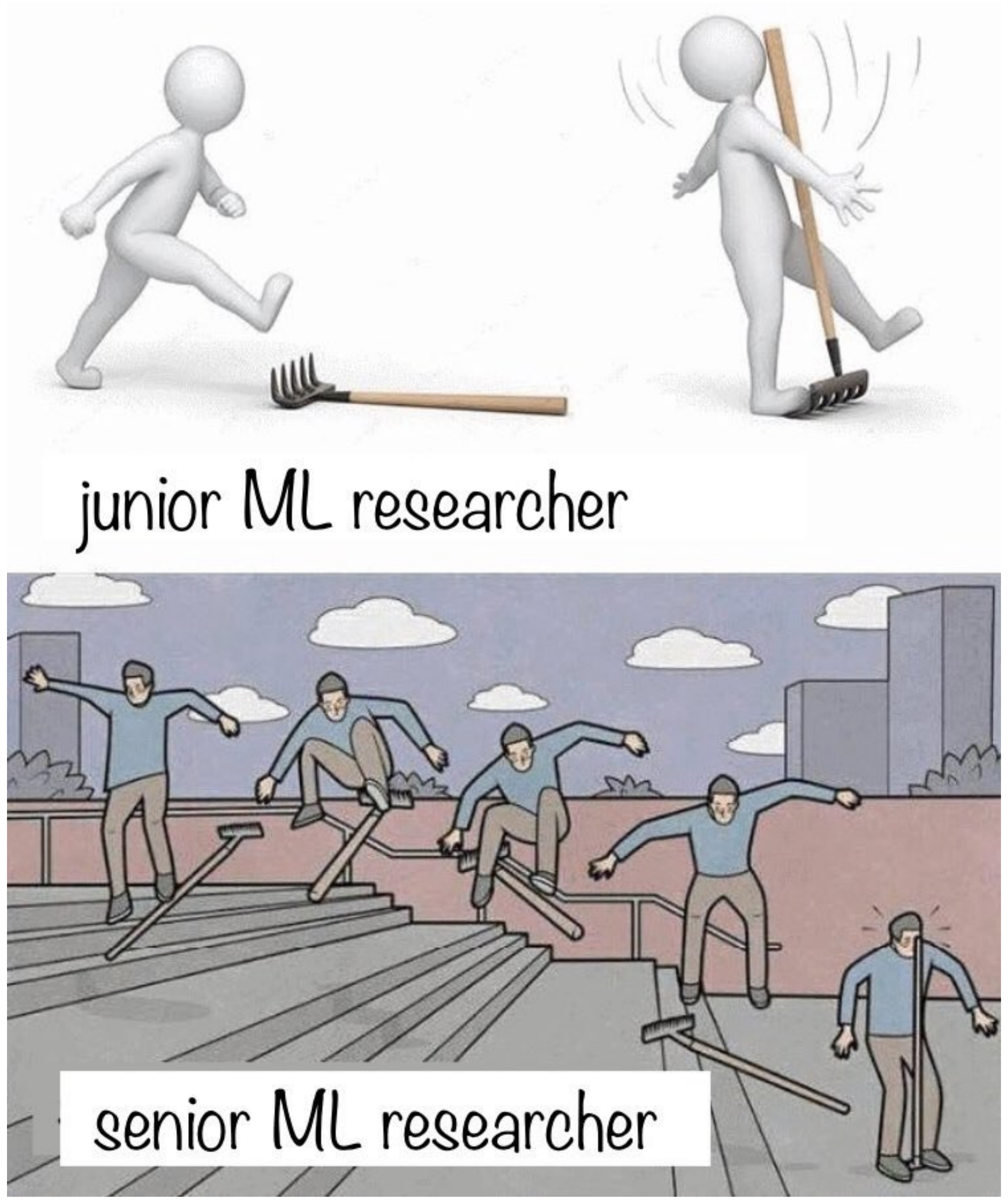
Wishful Thinker

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

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



Particle Physics Caveats

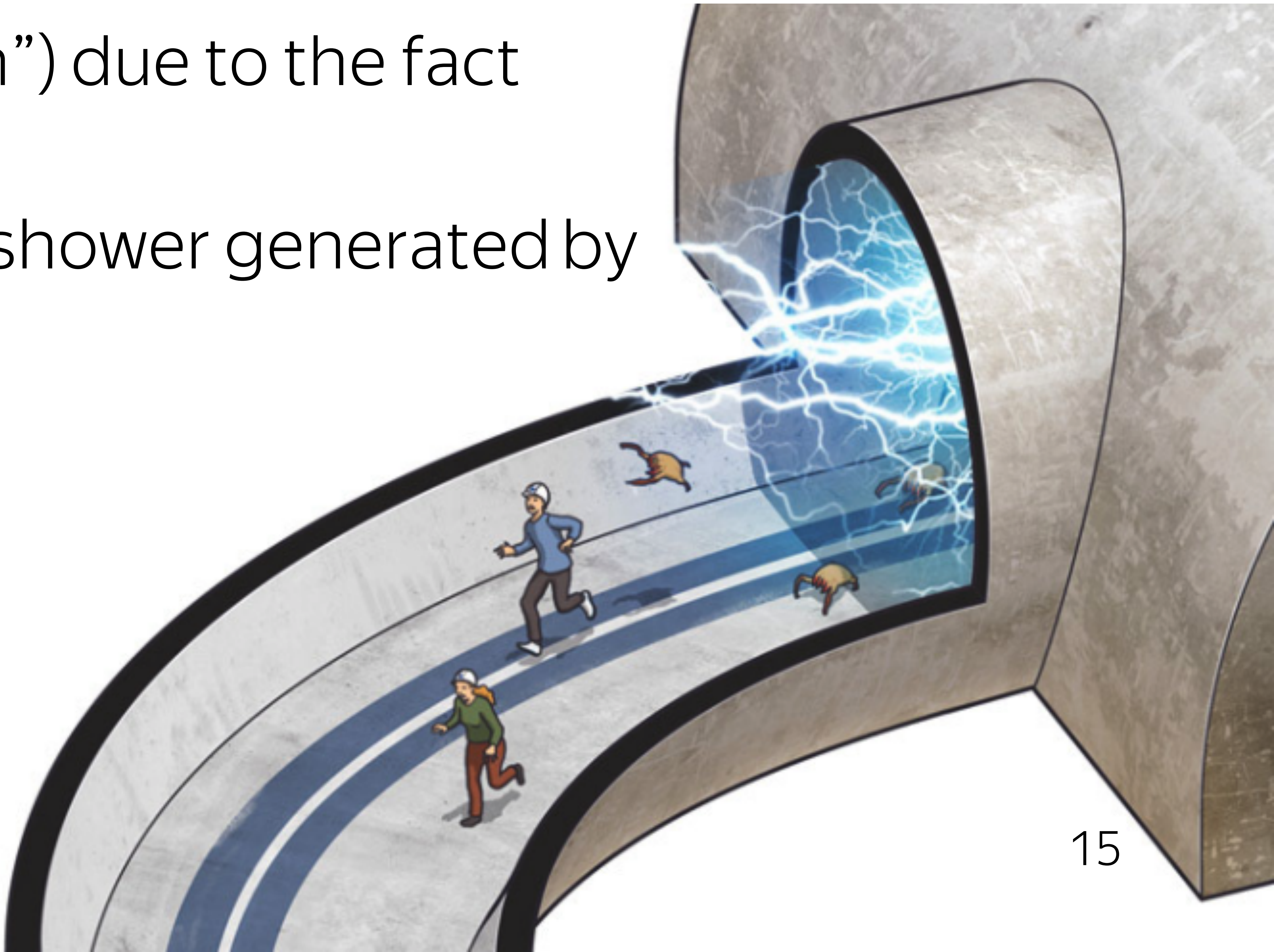
Domain-specific barriers

- 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?

Motivation for DS people?

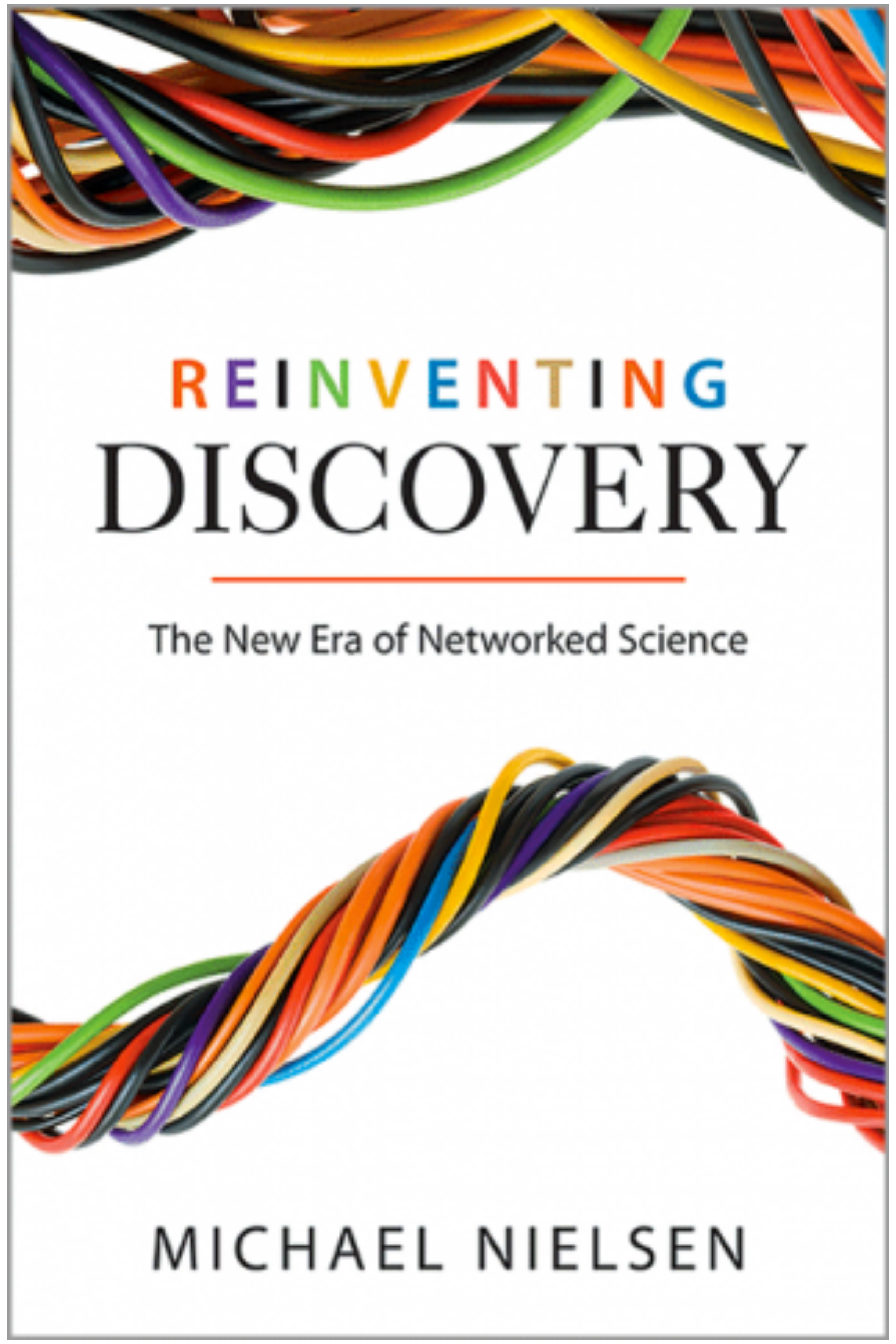


Research Collaboration Platform Candidates

- 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
  - No micro-reward motivation, no contribution-tracking, no means for publishing / reuse / peer review

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



REINVENTING

DISCOVERY

The New Era of Networked Science

MICHAEL NIELSEN

Michael Nielsen, Reinventing the Discovery, 2014

Demand for a platform

“Mechanical Turk for science”

Flexibility to change the metric, even during research

Micro-contributions

- Track all the records
- Peer-reviews
- Profile building

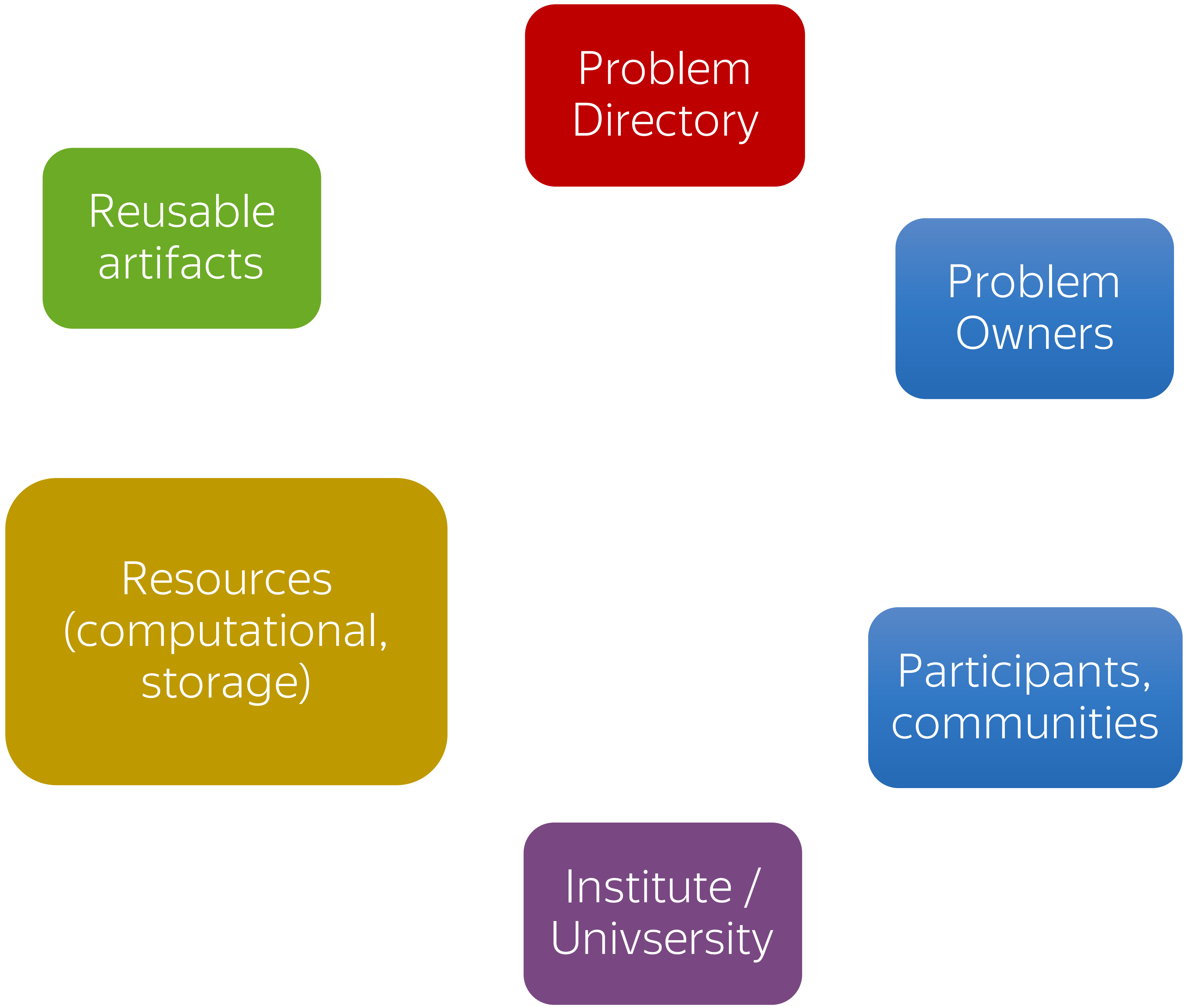
Reusable (publicable) results

Communication (goal, manifest, fast bootstrap)

Global-scale, transparency

Motivation for micro-contributions

High-level platform Components



Problem Directory

Reusable artifacts

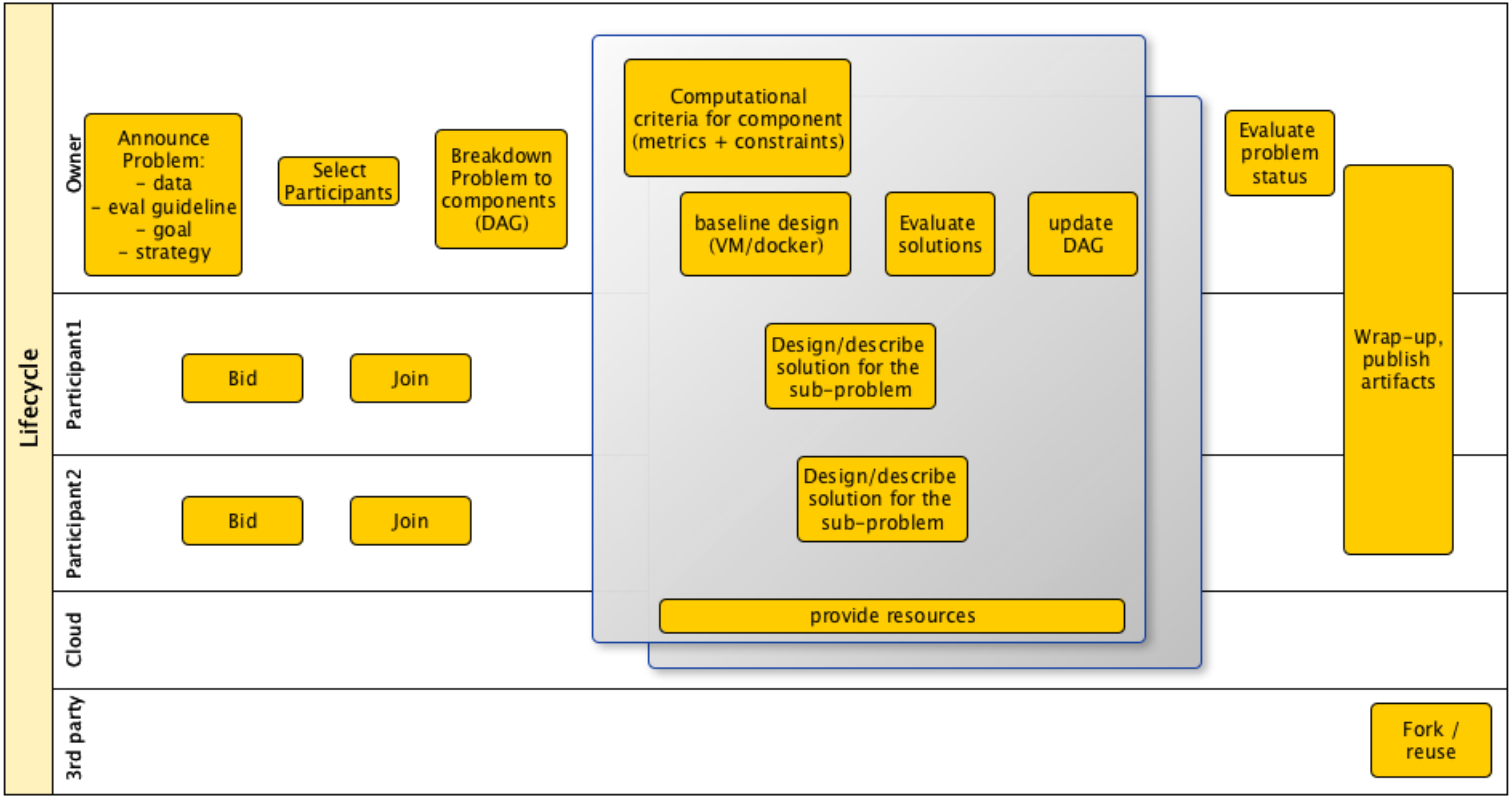
Problem Owners

Resources (computational, storage)

Participants, communities

Institute / Univsersity

Collaboration Lifecycle


Lifecycle
Owner
Participant1
Participant2
Cloud
3rd party
Announce Problem:
- data
- eval guideline
- goal
- strategy
Select Participants
Breakdown Problem to components (DAG)
Computational criteria for component (metrics + constraints)
baseline design (VM/docker)
Evaluate solutions
update DAG
Evaluate problem status
Bid
Join
Design/describe solution for the sub-problem
Bid
Join
Design/describe solution for the sub-problem
provide resources
Wrap-up, publish artifacts
Fork / reuse

Collaboration artifacts

User profile:

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

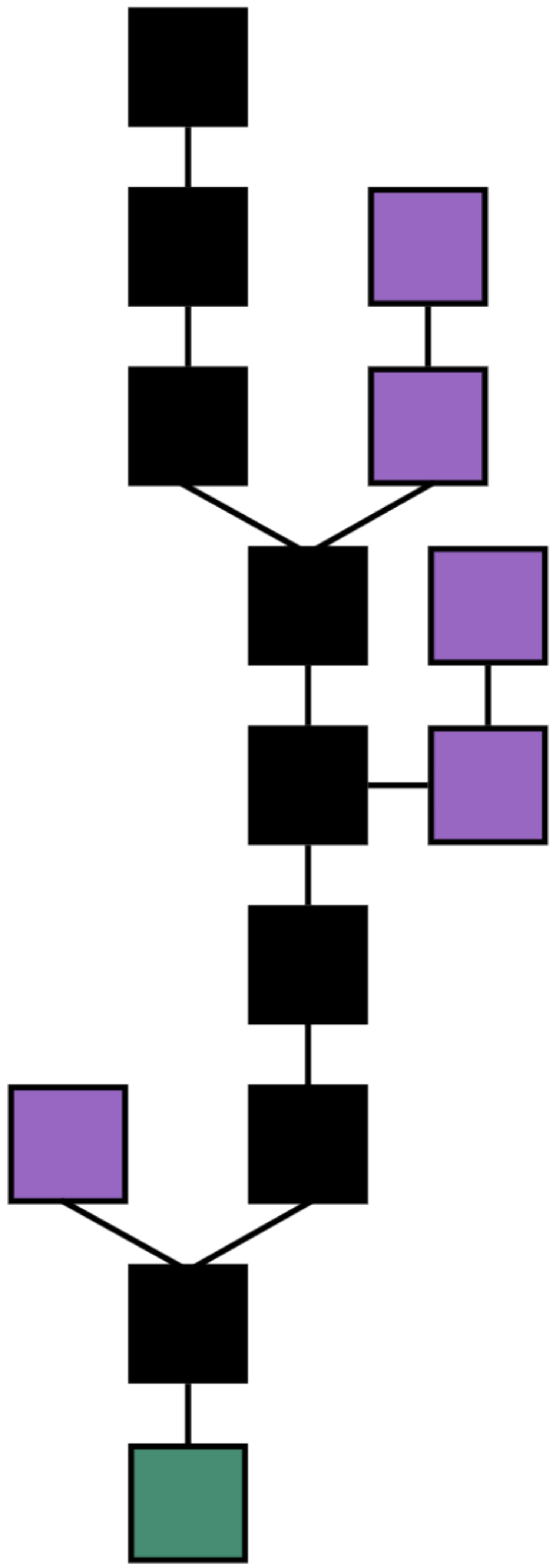
Competition profile:

- Baselie
- Metrics, leaderboard
- Re-usable models

What about trust and motivation?

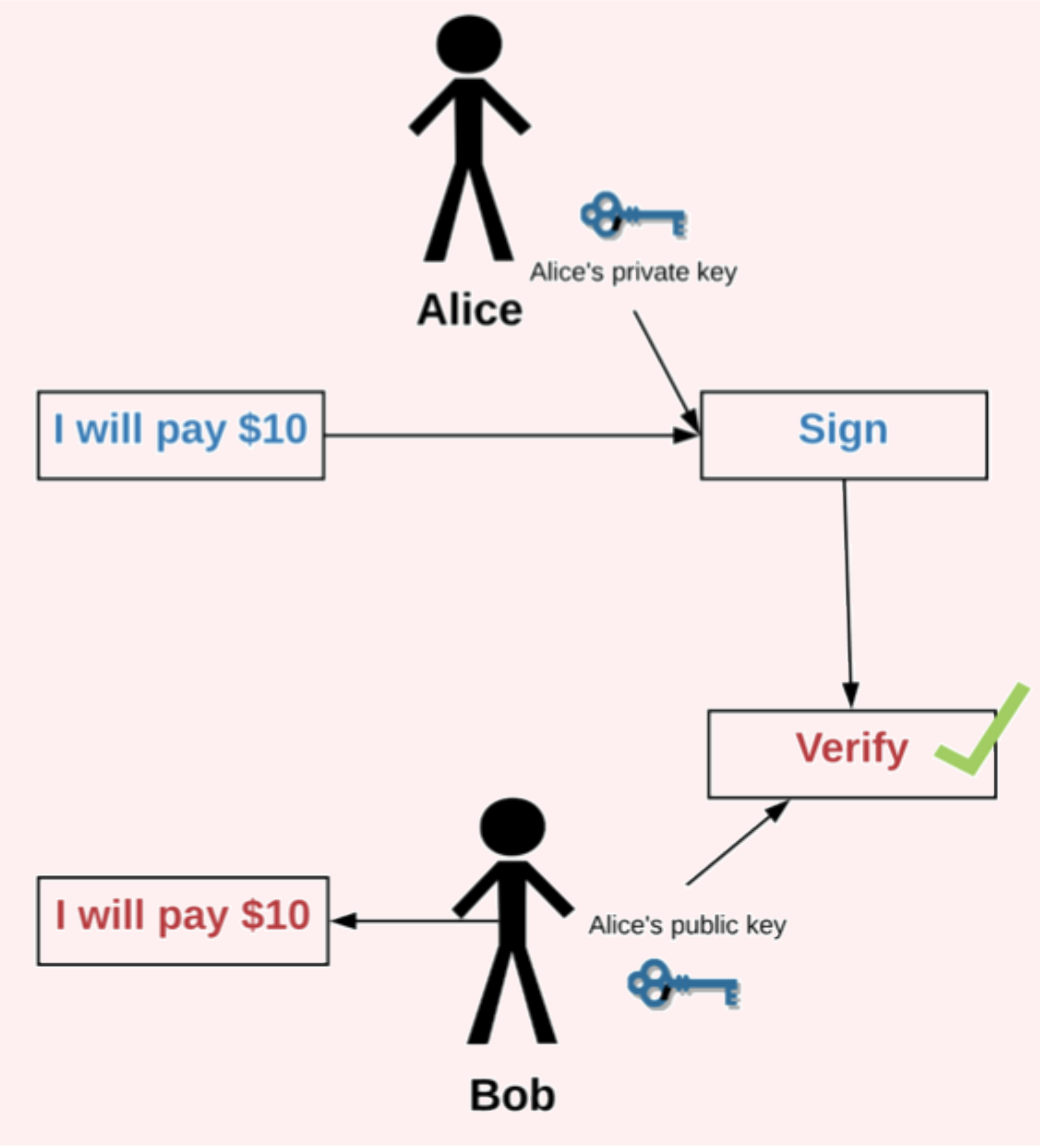
Blockchain - A Distributed Ledger Technology

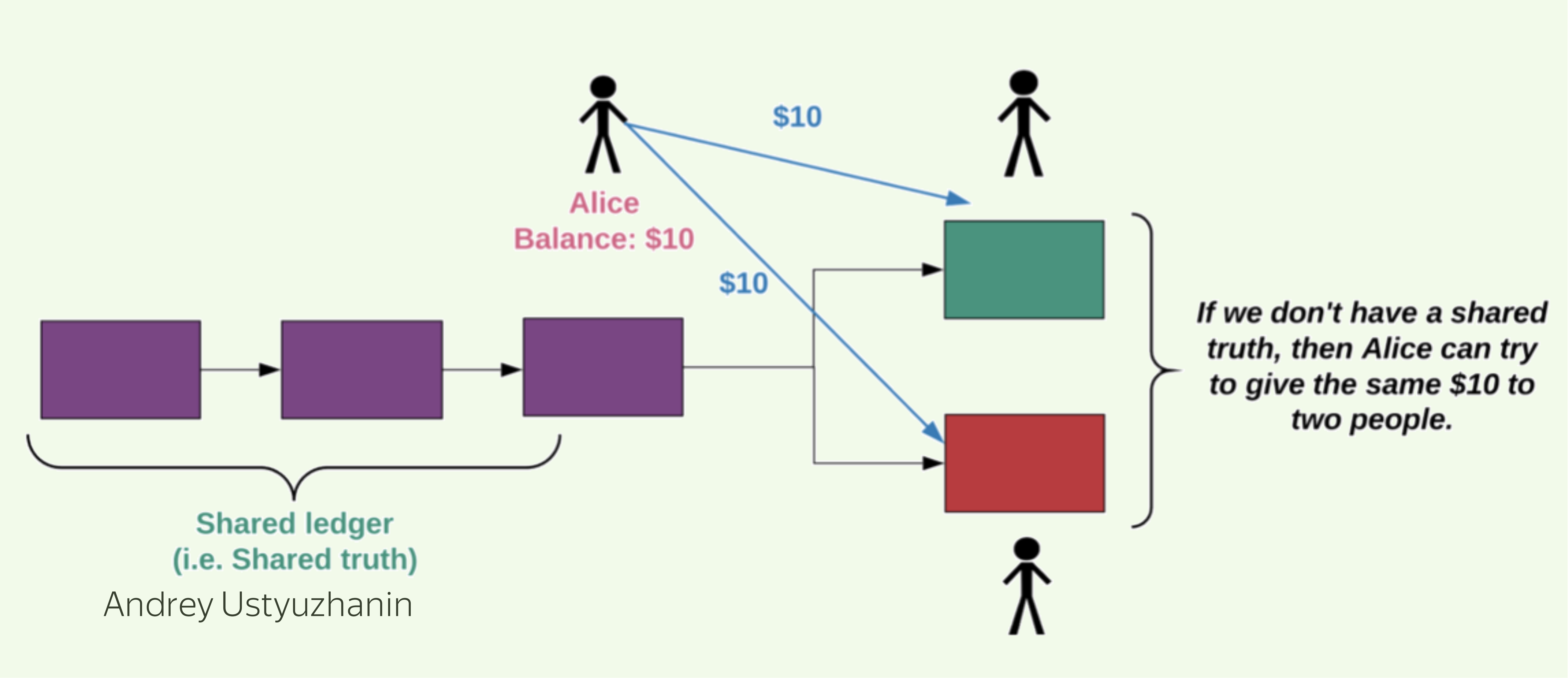
- A blockchain is a linked list where each node is connected to its predecessor by a cryptographic hash
  - All pointing back to the “genesis” block (right, in green) which may contain defining information about the rules for the blockchain protocol
  - In this way a blockchain comprises a verifiable 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



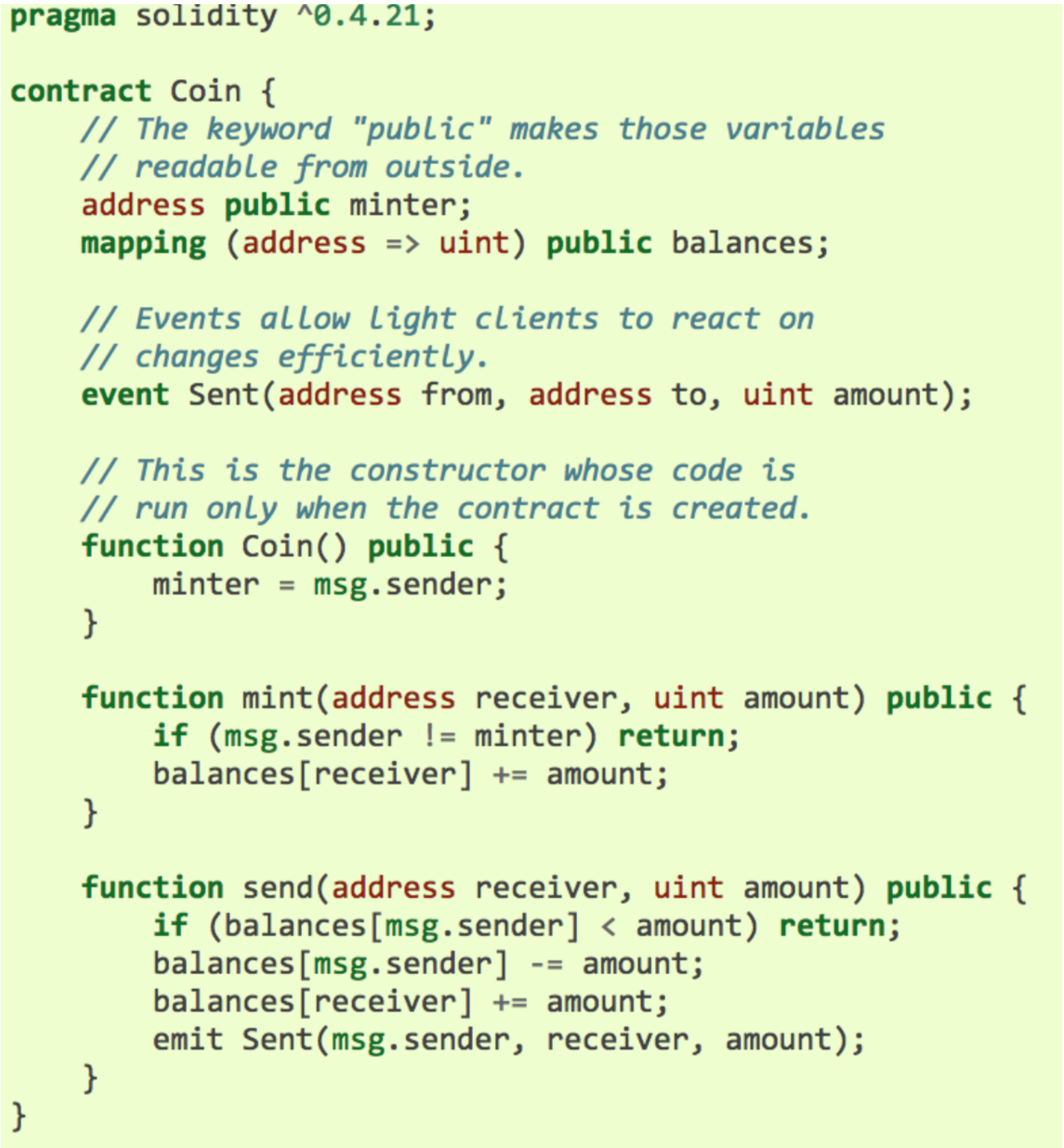


If we don't have a shared truth, then Alice can try to give the same $10 to two people.

Andrey Ustyuzhanin

Blockchain – Smart Contract

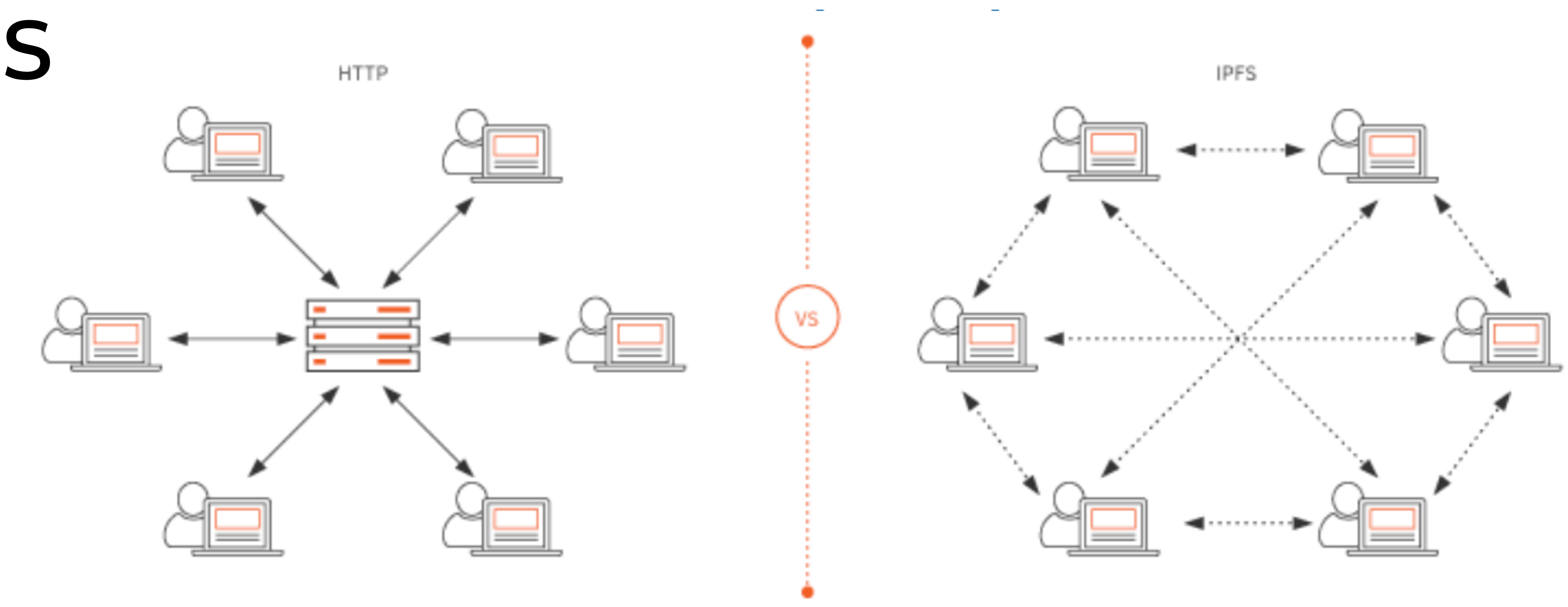
- Newer blockchains, Ethereum for instance, implement virtual machines that can execute byte code
- Smart contracts, implemented in this code allow binding between blockchain addresses and actions that are taken by the code
  - Typically the same code gets executed by all nodes in the network (extension 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);
    }
}
```

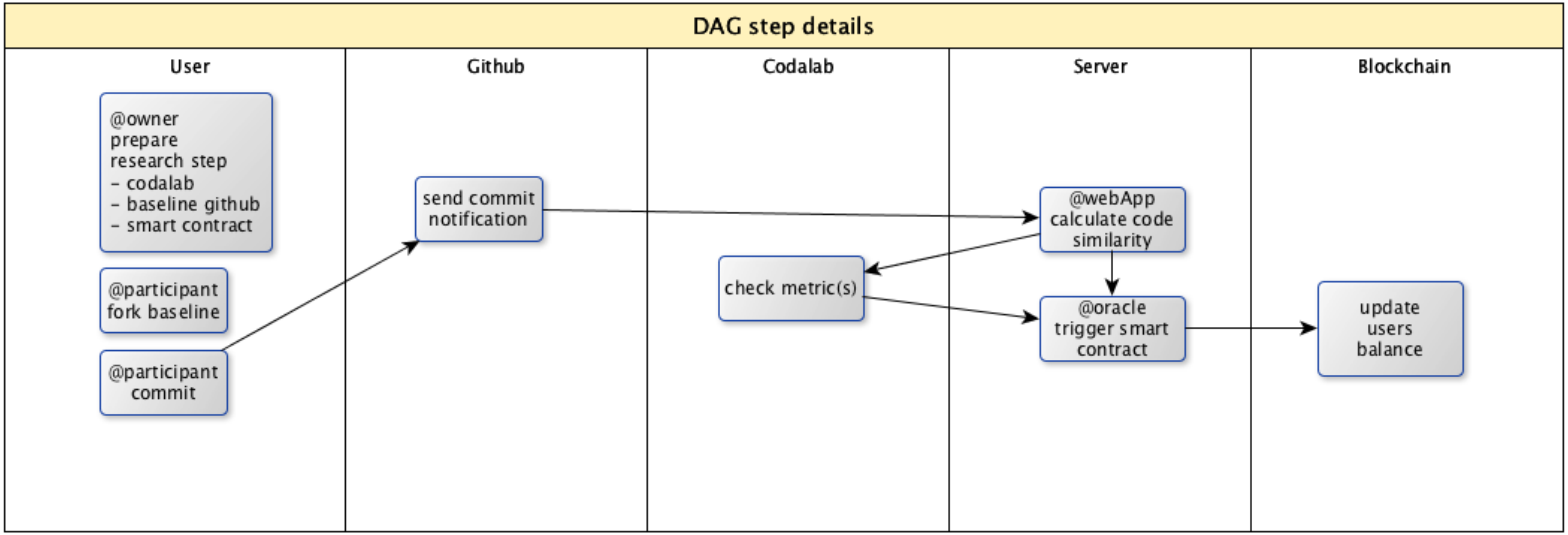
A simple example of a derived currency

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

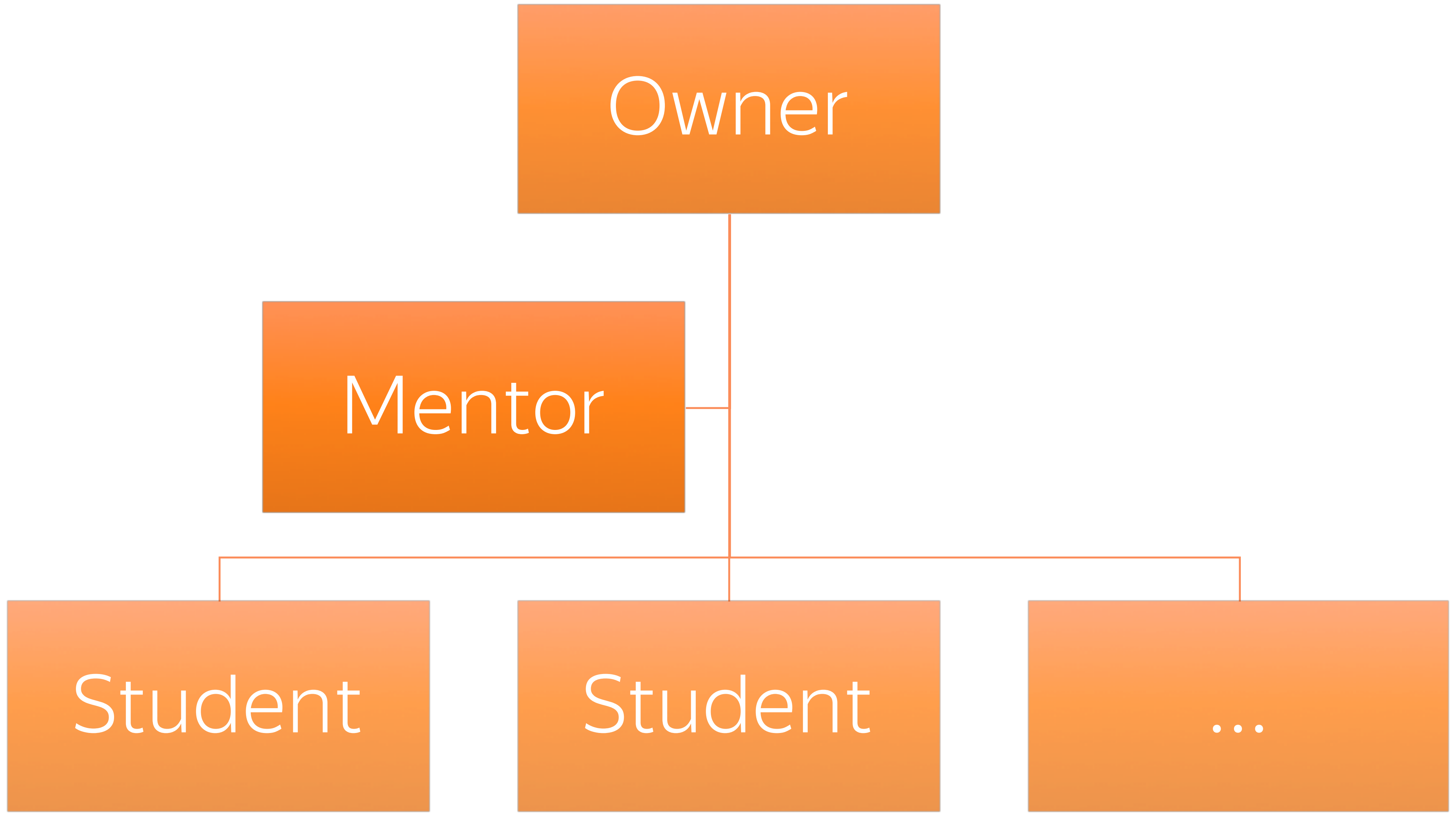
Technologies used:

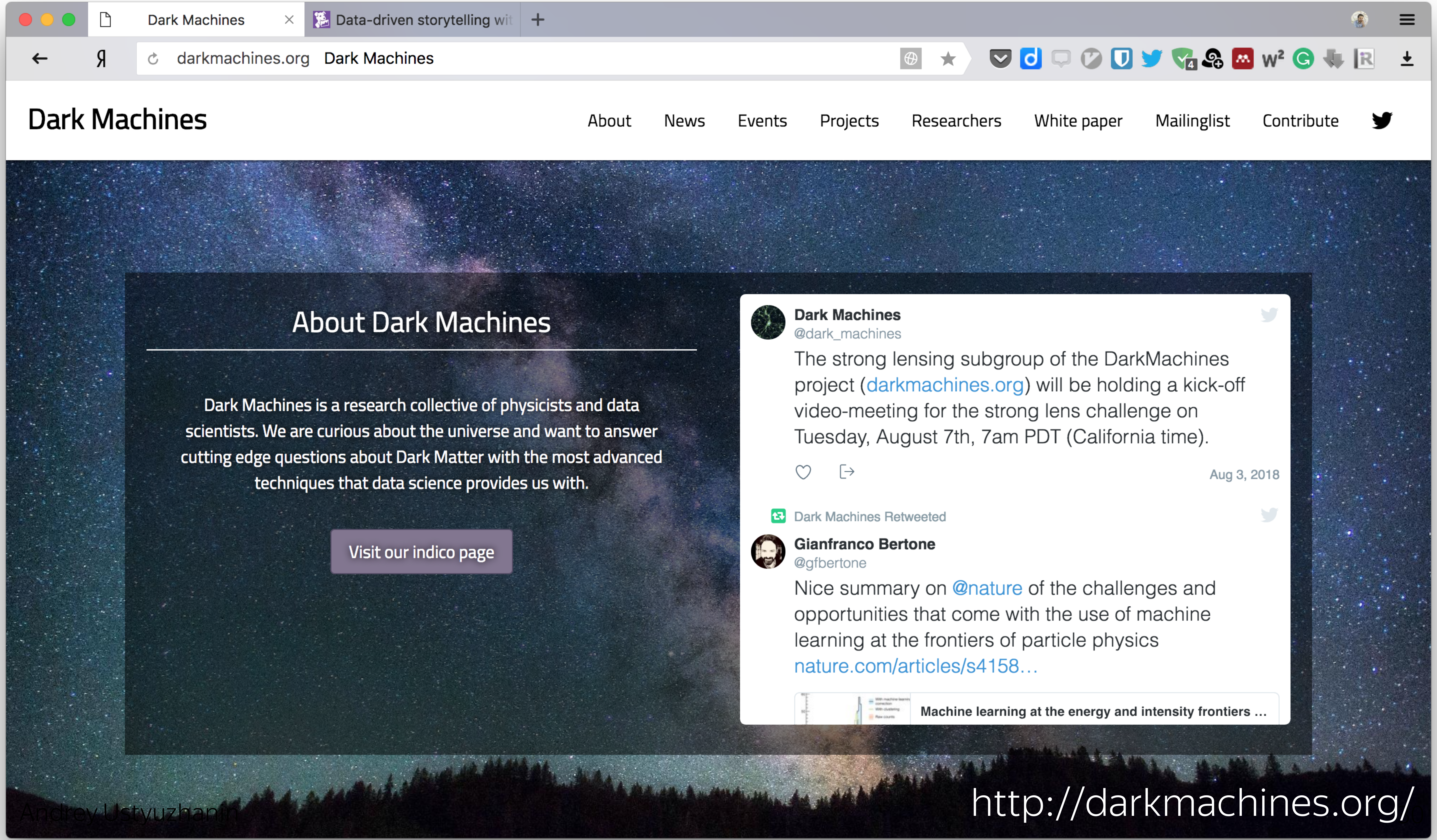
- Github, Kaggle

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





Dark Machines

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.

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 …

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

More ideas for collaboration

Triggers

Tracking

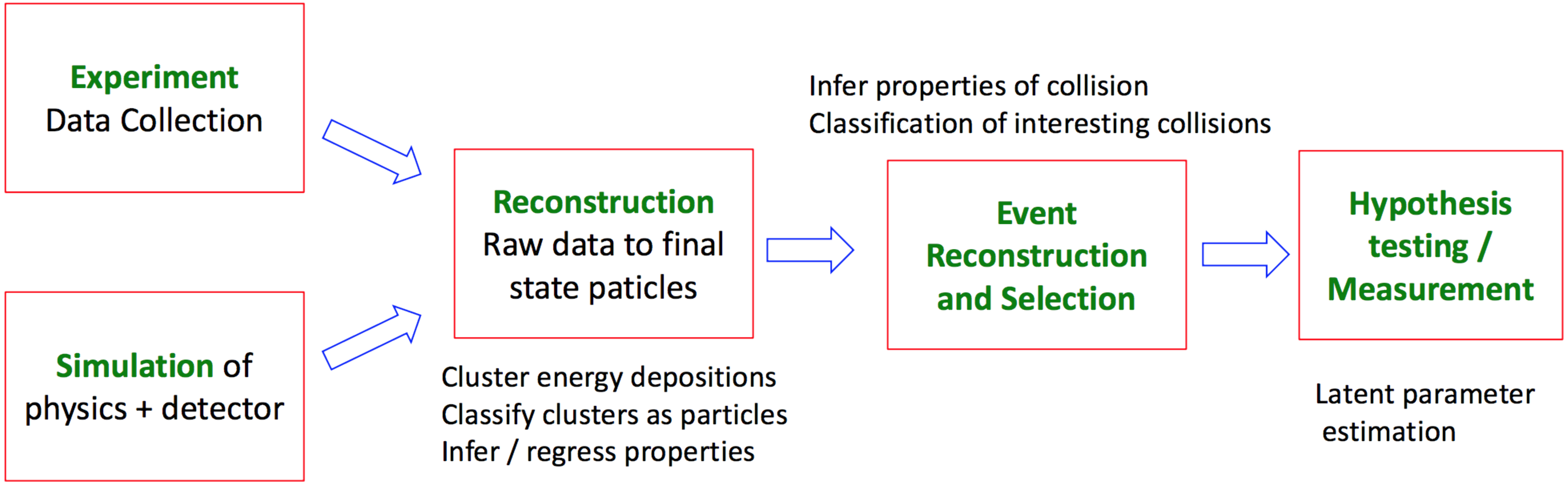
Object identification

- Particle, showers, jets

Fast MC generation

Model checking

Detector design optimization



Experiment Data Collection

Simulation of physics + detector

Reconstruction Raw data to final state paticles

Event Reconstruction and Selection

Hypothesis testing / Measurement

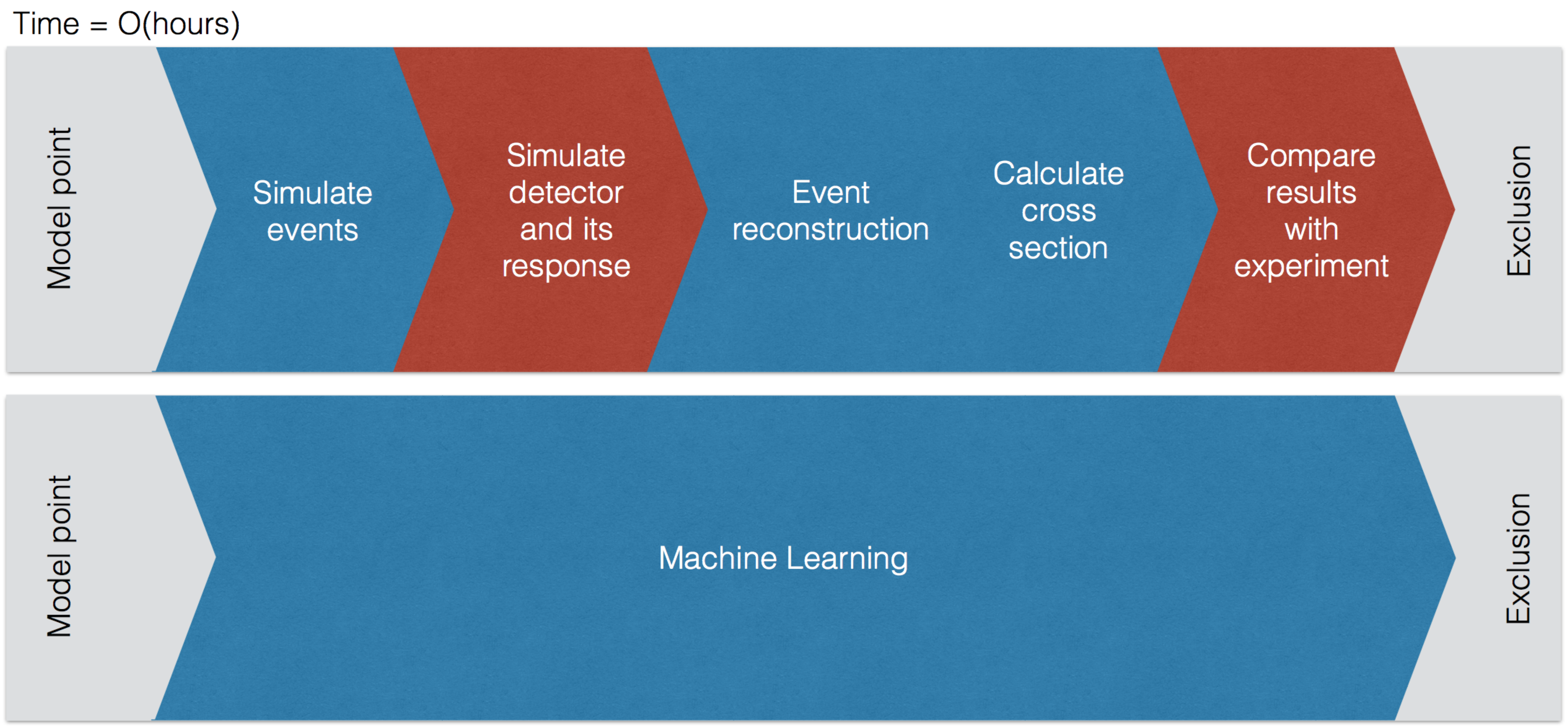
Infer properties of collision
Classification of interesting collisions

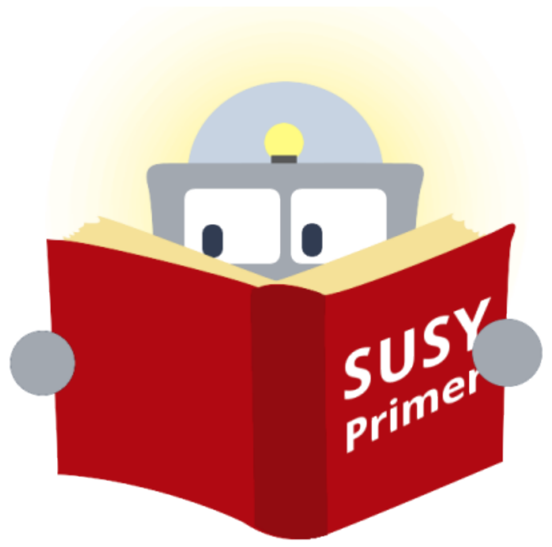
Cluster energy depositions
Classify clusters as particles
Infer / regress properties

Latent parameter estimation

Generative Model

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?
  - Dataset? (simulated, or generator itself)
  - 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?

Further ideas

Should there be feedback loop from solution running in production?
What is the best way from metric to fair smart-contracts?

- Increase of metrics?
- Metric hacking?

Collect statistics of humans dealing with problems for training ML algorithm for automated improvements

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
  - Challenges: sociological (communications), psychological

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/

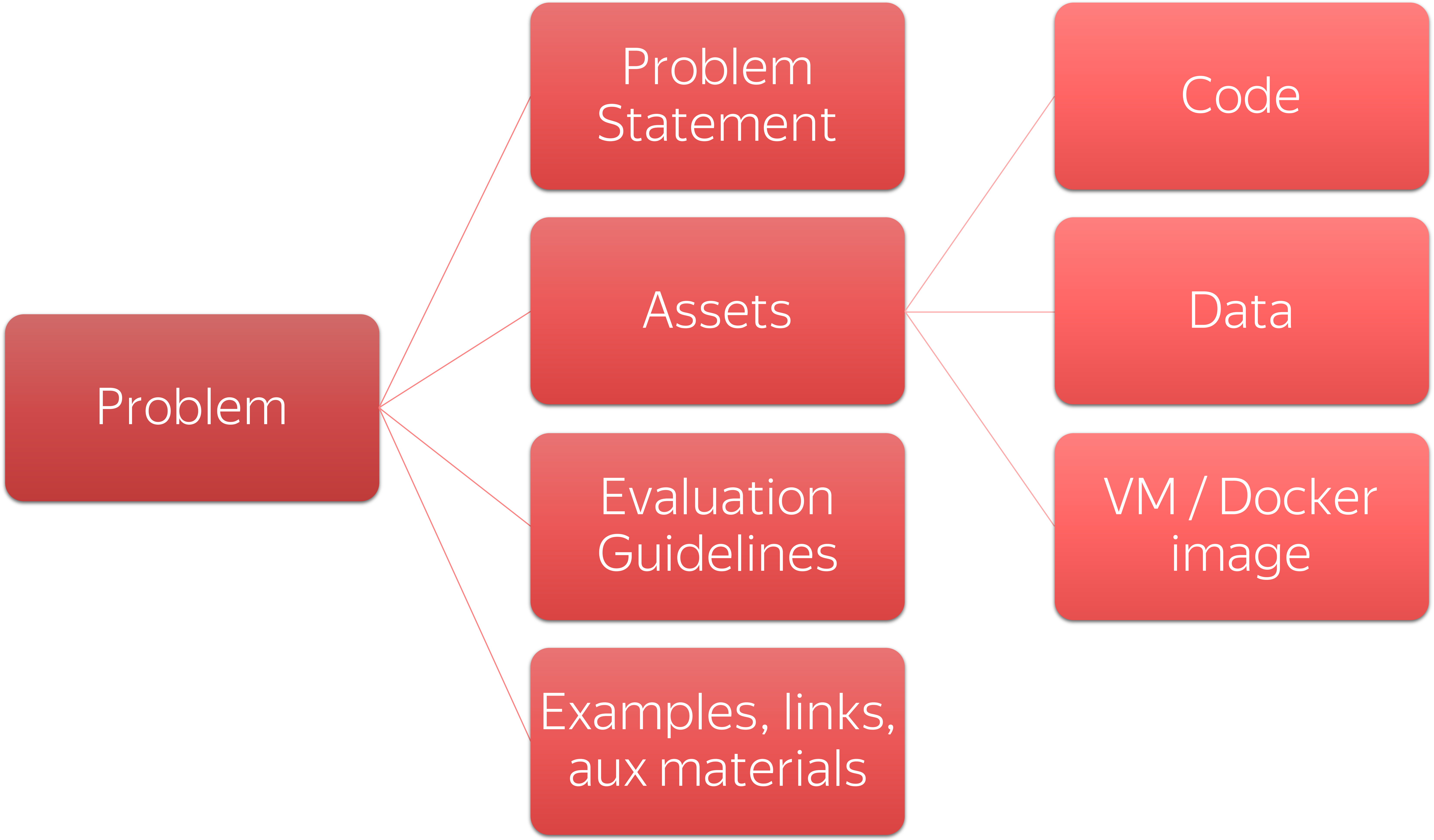
Questionnaire if you have a challenge to share

- https://goo.gl/forms/PmYJBwyA3RVsPSHC2

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

Problem Structure



Problem

Problem Statement

Assets

Evaluation Guidelines

Examples, links, aux materials

Code

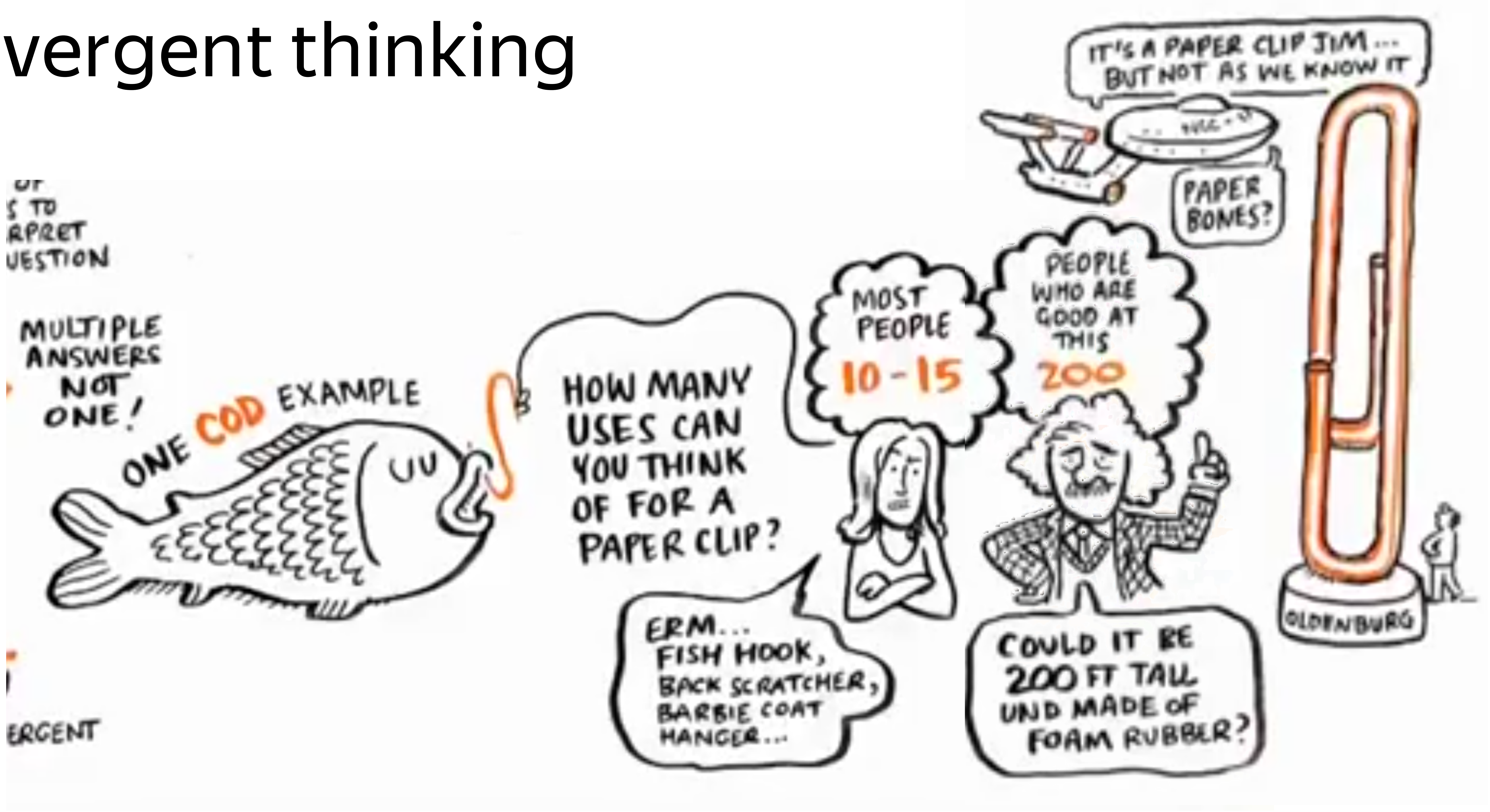
Data

VM / Docker image

Abridged history of Eductaion system

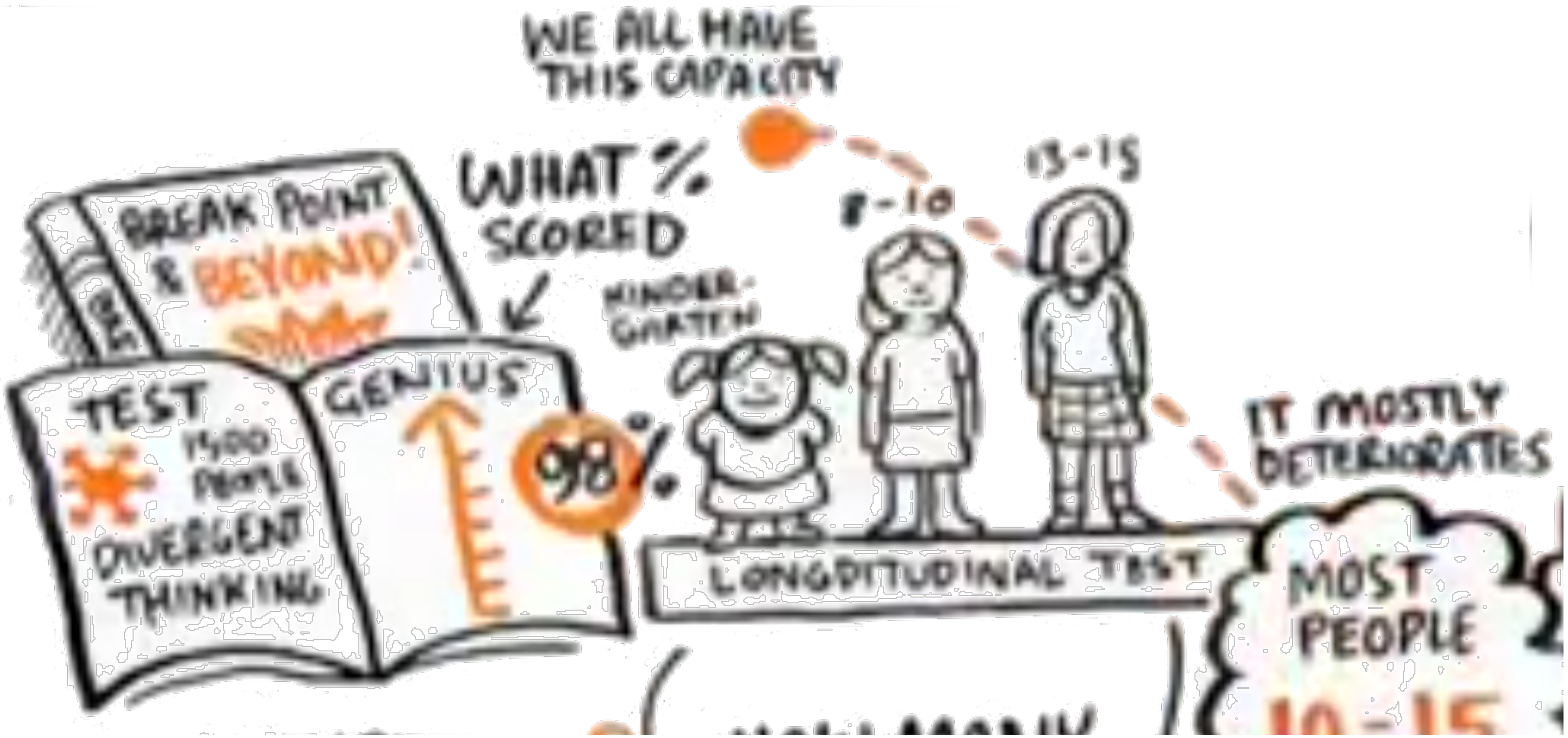
- **1000+ years – elite**
  - hollistic
- **200+ years – public**
  - Funded by state (from taxes)
  - Industry-oriented
  - There are life-long paths to take
- **10+ years – online**
  - Individual (no batches)
  - Limited practice
  - Limited credibility

Divergent thinking


UF
S TO
RPRET
UESTION
MULTIPLE ANSWERS NOT ONE!
ONE COD EXAMPLE
HOW MANY USES CAN YOU THINK OF FOR A PAPER CLIP?
MOST PEOPLE 10 - 15
PEOPLE WHO ARE GOOD AT THIS 200
IT'S A PAPER CLIP JIM... BUT NOT AS WE KNOW IT
PAPER BONES?
ERM... FISH HOOK, BACK SCRATCHER, BARBIE COAT HANGER...
COULD IT BE 200 FT TALL UND MADE OF FOAM RUBBER?
OLDENBURG
ERGENT

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/

GREAT LEARNING HAPPENS IN GROUPS
COLLABORATION IS THE STUFF OF GROWTH

One more trend in Science

Factors

- Reduced research funding
- Higher enternace barriers
- Higher interest in research for amateurs

Demand:

- Communication media for collaboraiton