Yandex SCHOOL OF DATA ANALYSIS

HIGHER SCHOOL OF ECONOMICS

Bridging the gap between scientific research and data-science

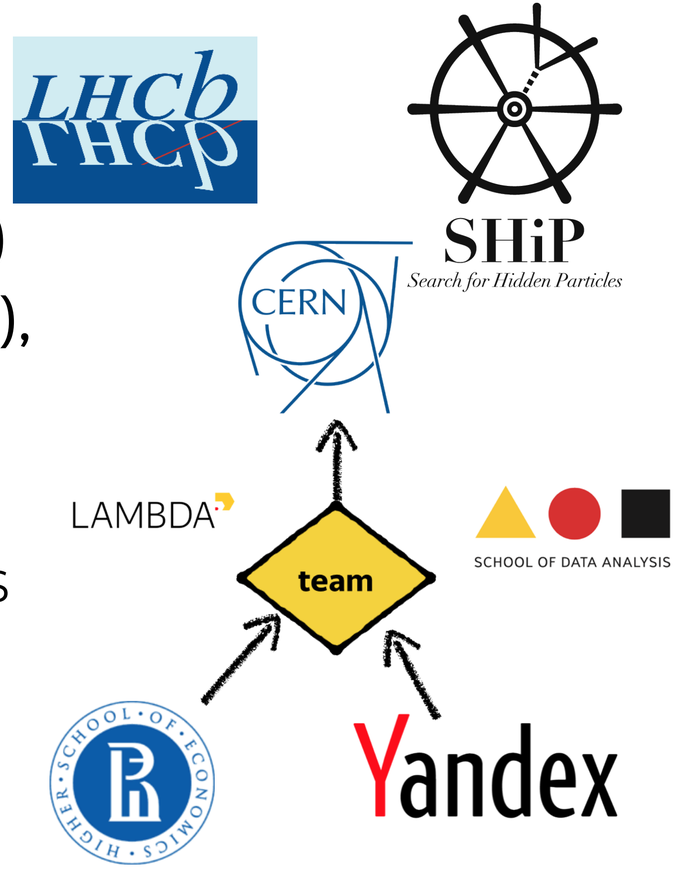
Andrey Ustyuzhanin

NRU HSE

YSDA

ICL

Quick self-intro



- Head of LHCb team in Yandex School of Data Analysis (YSDA)
- Head of Laboratory (link) at Higher School of Economics (HSE), YSDA (since 2007):
  - Joint master's degree in data science
  - Solving High Energy Physics problems with ML approaches
  - member of LHCb, SHiP, CRAYFIS
- HSE, Laboratory (since 2015):
  - focuses on applying ML to **natural science challenges**
  - HSE has joined LHCb this summer!
  - Collaborates with industry as well
- Education activities (MLHEP, ML at ICL,ClermonFerrand, LaSAL, Coursera)

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

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/2vzllWT

Divergent thinking

http://bit.ly/2vzllWT

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

DataScience competition: Netflix Prize

Netflix prize – prediction of DVD titles renting (1M USD)

- training data set of 100,480,507 ratings that 480,189 users gave to 17,770 movies
- Each training rating is a quadruplet of the form <user, movie, date of grade, grade>
- The user and movie fields are integer IDs, while grades are from 1 to 5 (integral) stars
- The qualifying data set contains over 2,817,131 triplets of the form <user, movie, date of grade>, with grades known only to the jury
- A participating team's algorithm must predict grades on the entire qualifying set, but they are only informed of the score for half of the data, the **quiz** set of 1,408,342 ratings. The other half is the **test** set of 1,408,789, used to find winners.
- Submitted predictions are scored against the true grades in terms of root mean squared error (RMSE), and the goal is to reduce this error

https://wiki2.org/en/Netflix_Prize

Netflix Prize timeline

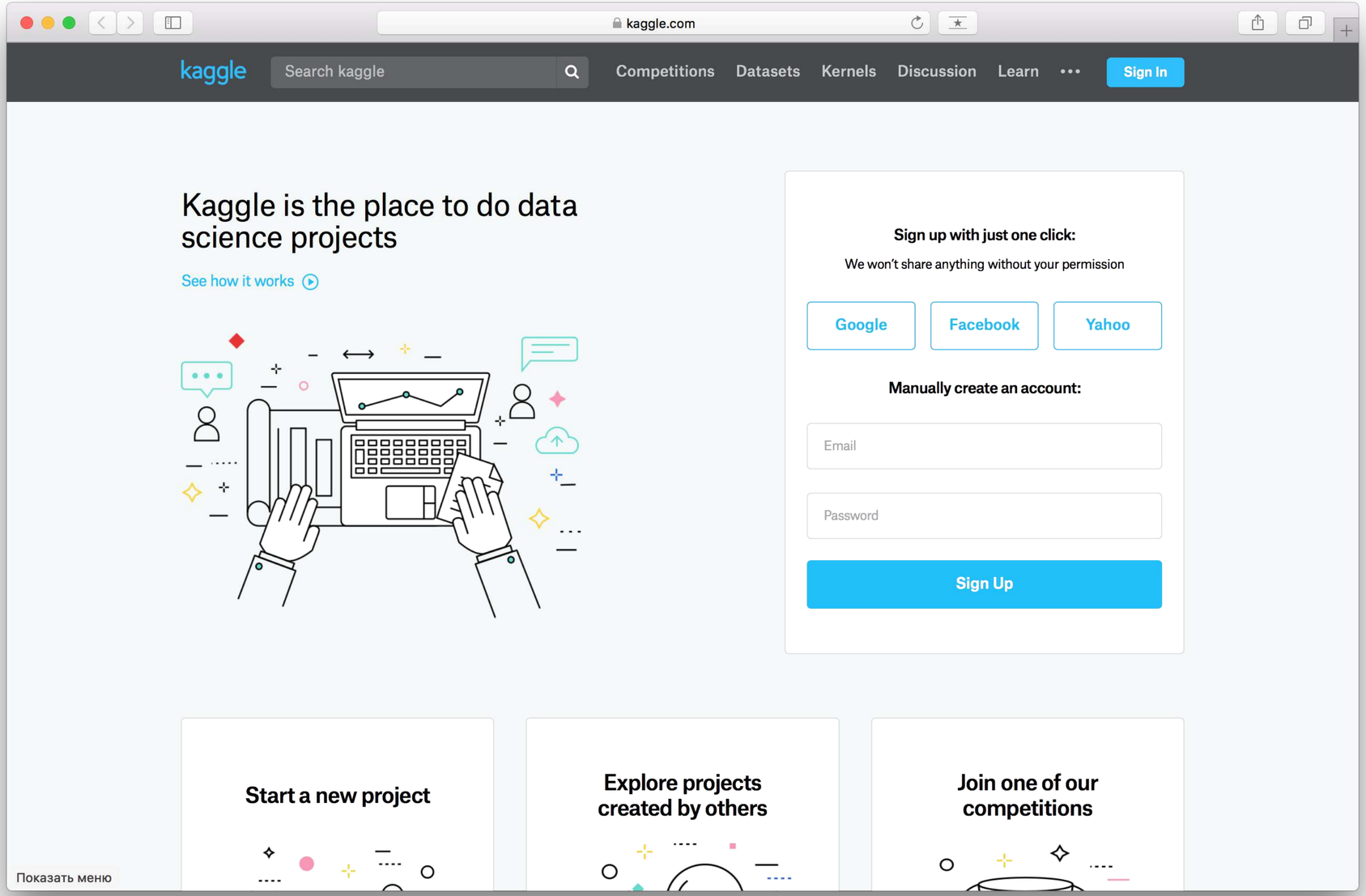
Netflix prize – prediction of DVD titles renting (1M USD for improving baseline by 10%)

- Baseline algorithm – Cinematch (linear model)
- 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 final test! The prize was awarded to BellKor's Pragmatic Chaos.

Second challenge was cancelled due to privacy concerns.

https://wiki2.org/en/Netflix_Prize

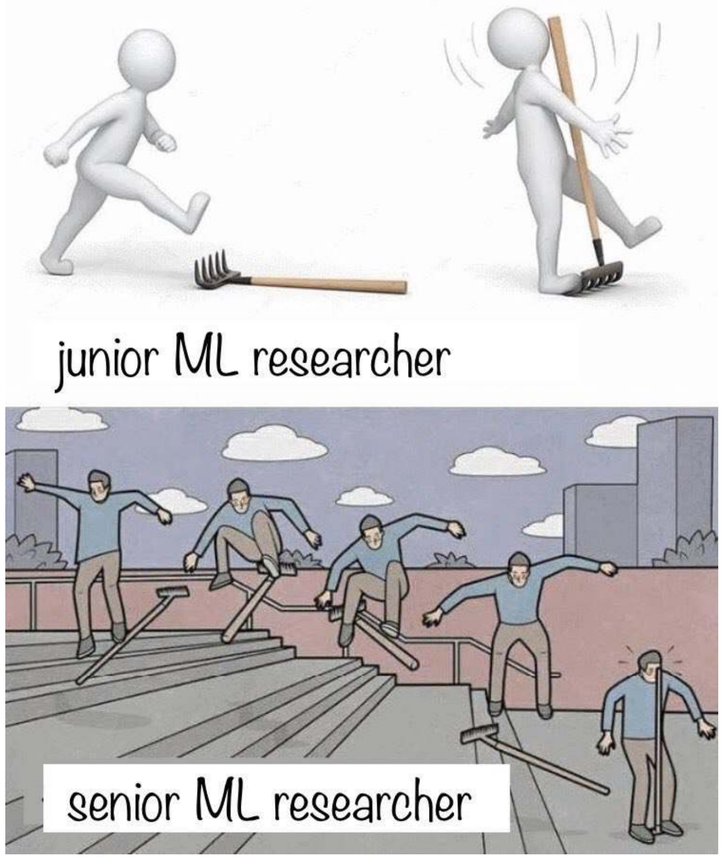


- $O(10^4)$ public datasets
- $O(10^3)$ competitions
- $O(10^4)$ users
- $O(10^8)$ submissions

http://bit.ly/2JDMo8j

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



HEP Caveats

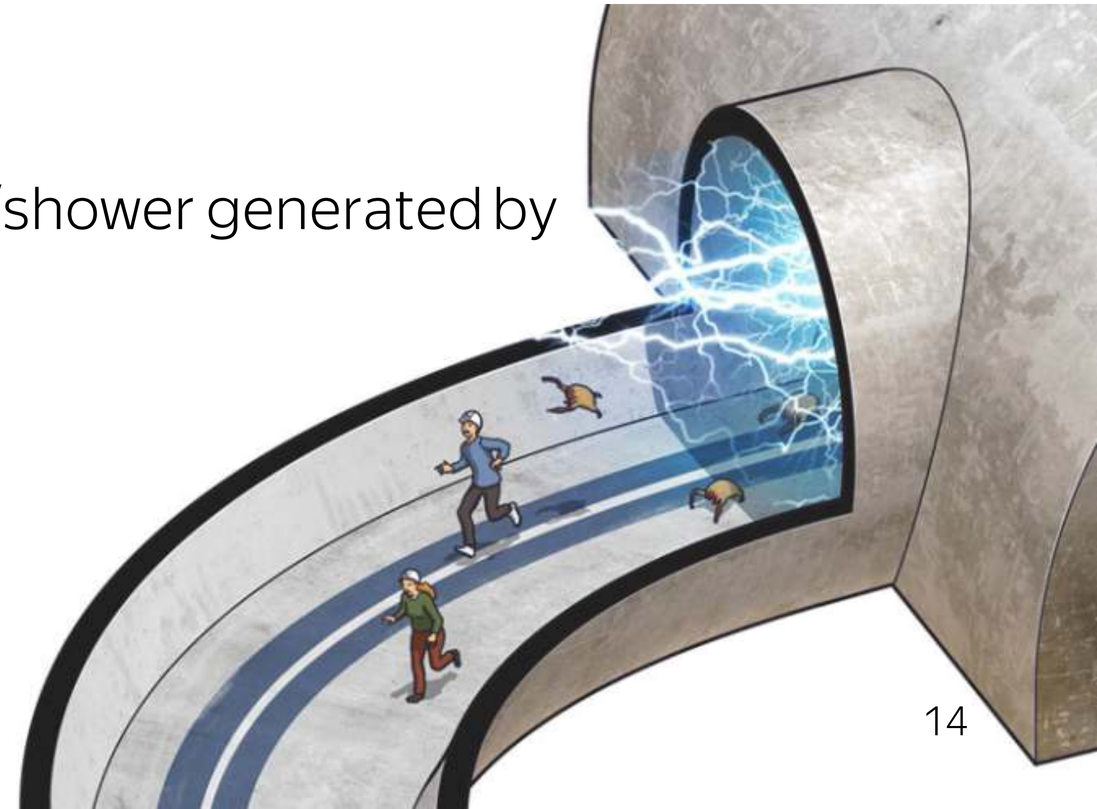
Domain-specific barriers

- 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
- Enormous data flows
- No obvious metrics for ’sanity’ checks (is a jet/shower generated by NN looks realistic enough?)

Reproducibility/traceability of results

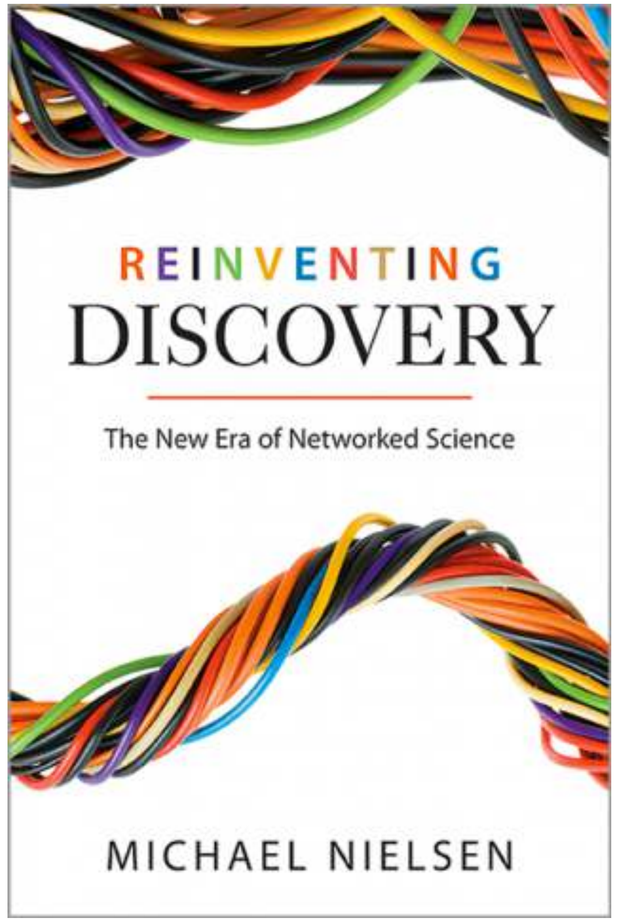
Cross-checks?

Motivation for DS people?



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



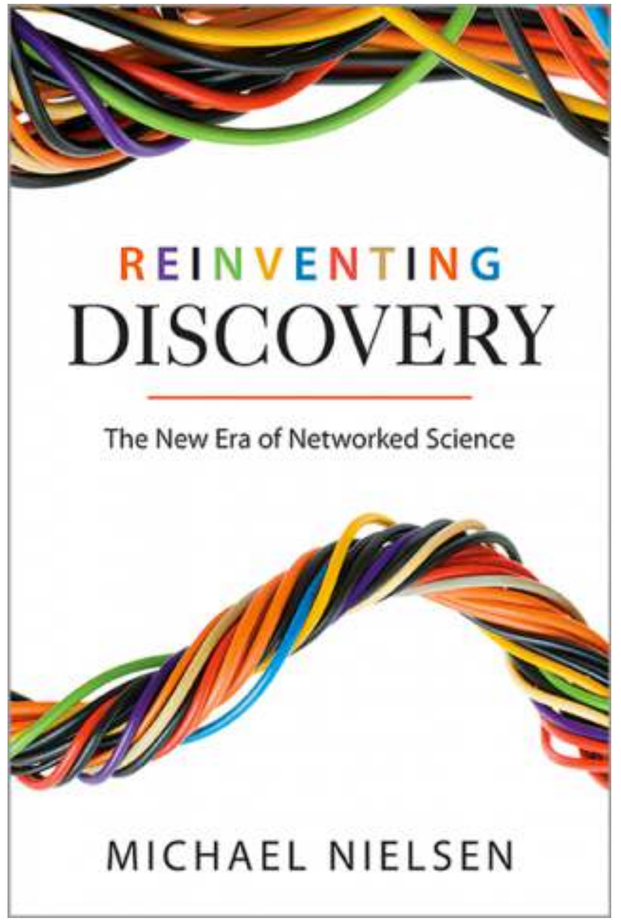
Michael Nielsen, Reinventing the Discovery, 2014

Demand for a platform

Goal-oriented
Metric-based, flexibility to change the metric
Micro-contributions

- Track records
- Peer-reviews
- Micro-rewards

Communication (forum, wiki)
Open-sourced
Global-scale
“Mechanical Turk for science”



Michael Nielsen, Reinventing the Discovery, 2014

Research Collaboration Platform Candidates

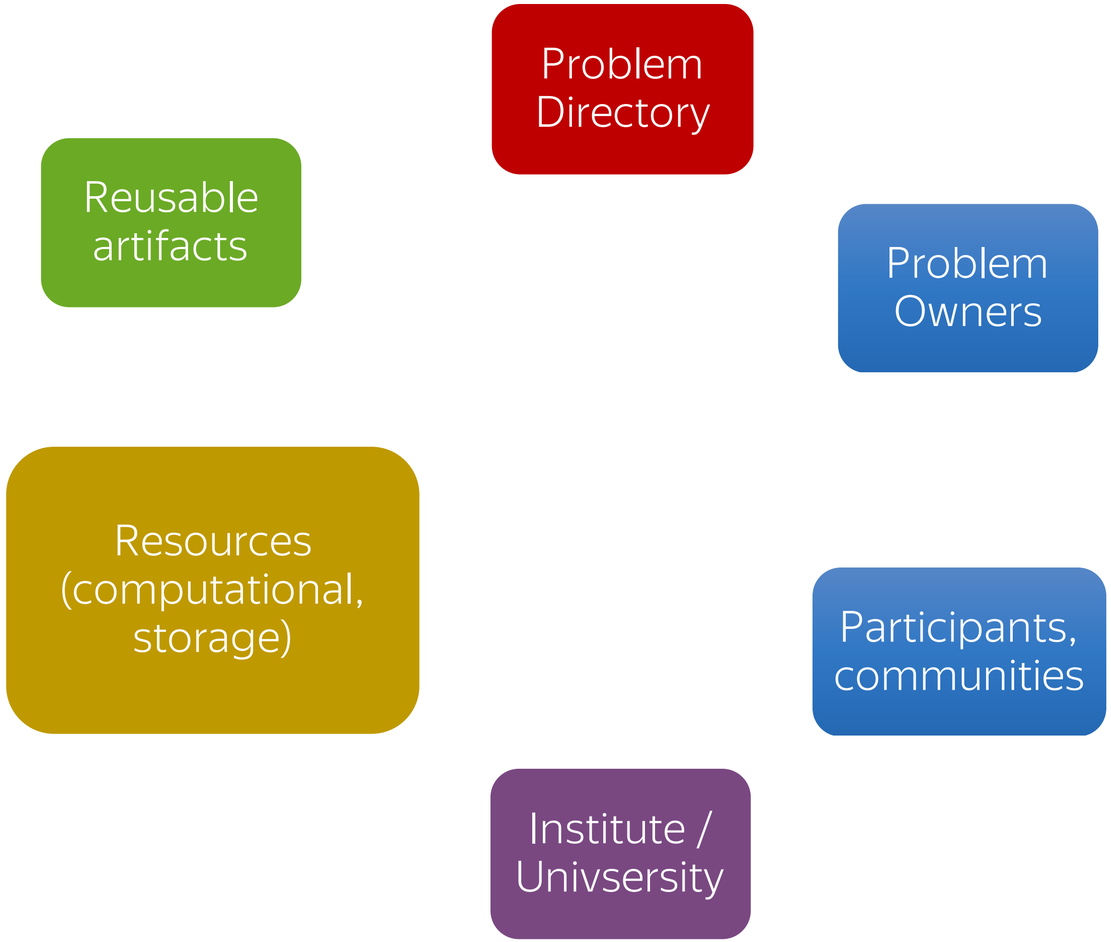
- **Github (belongs to Microsoft)**
  - No reward mechanism, too generic
- **Kaggle (belongs to Google)**
  - No micro-reward motivation, no reward for popular contribution, single metric from pre-defined list
- **CodaLab**
  - No micro-reward motivation, single metric, no means of publishing / reuse / peer review

Target Audience

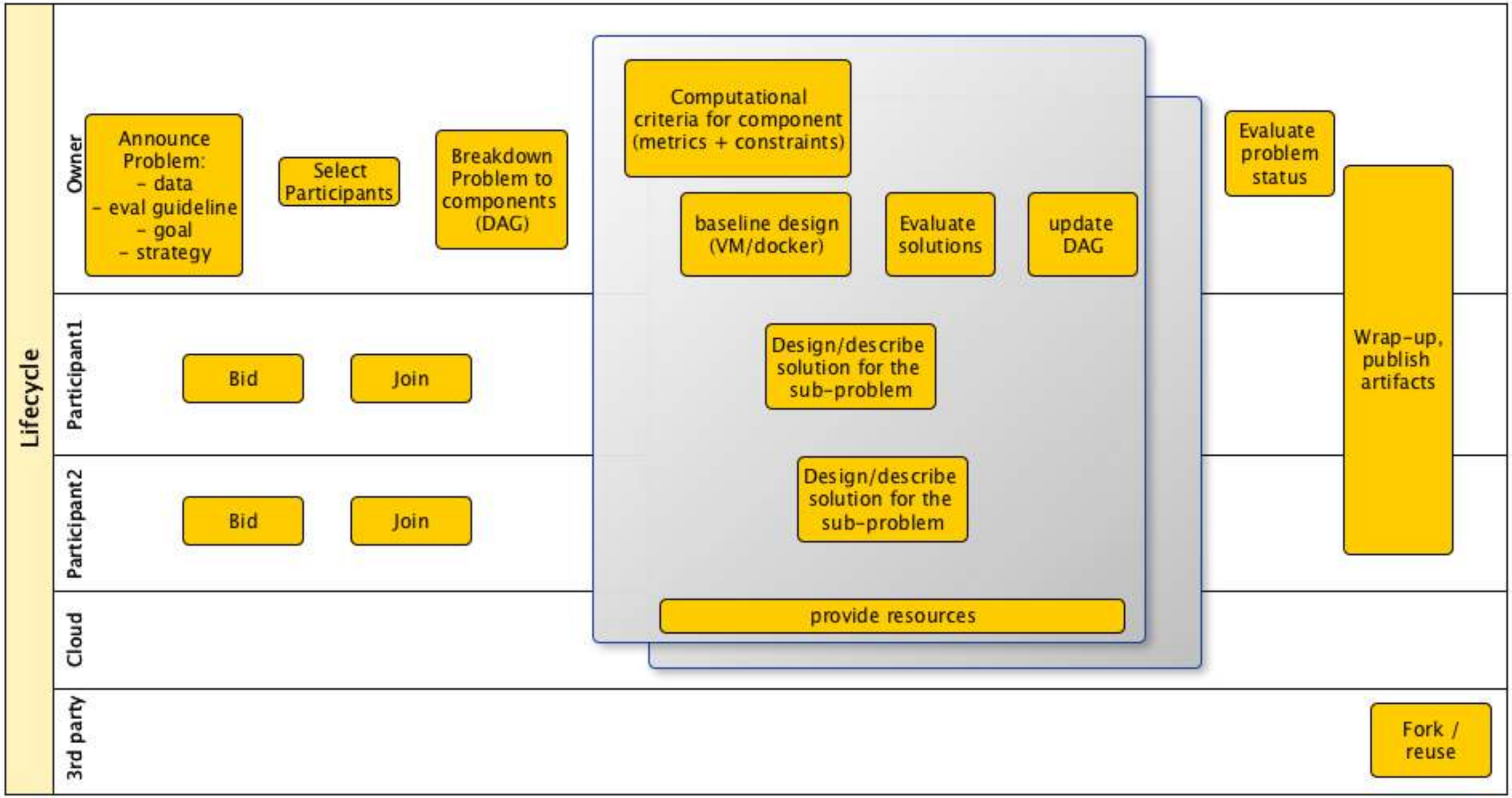
There are numerous people passing online ML courses, looking for decent problems to test their skills on

- Low-responsibility contribution
- Need for computational resources
- No time/resources for deep problem understanding
- Hungry for scoring records

High-level platform Components



Collaboration Lifecycle



What about diversity?

Blockchain - A Distributed Ledger Technology

A blockchain is a linked list where each node is connected to its predecessor by a cryptographic hash

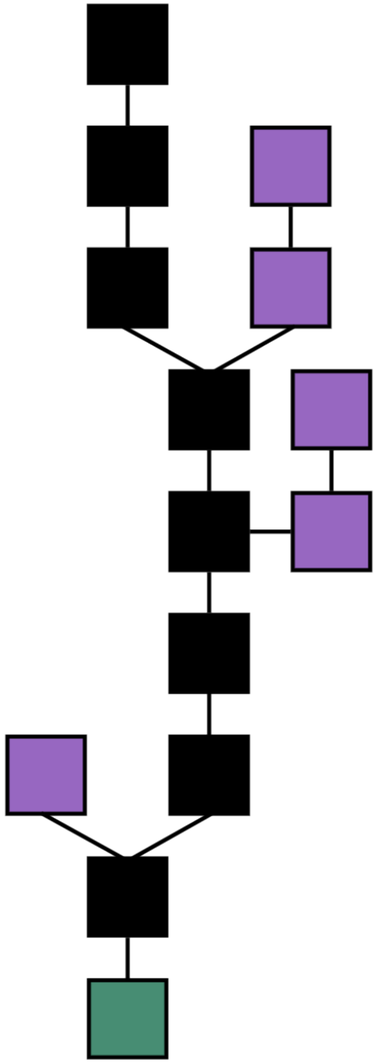
- All pointing back to the “genesis” block (right, in green) which may contain defining information about the rules for the blockchain protocol
- In this way a blockchain comprises a verifiable public ledger

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)

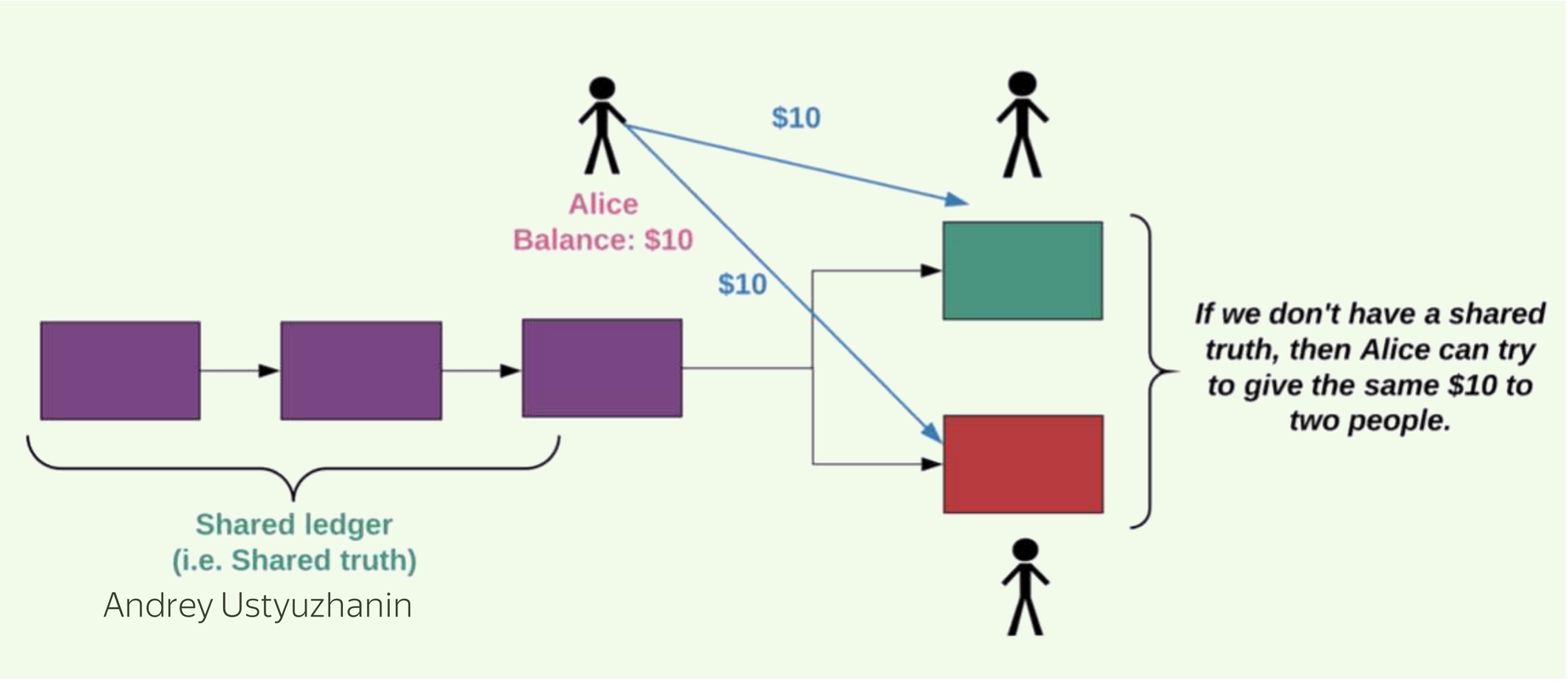
- However it’s up to the developers who define the protocol to determine the rules for consensus and evolution of the chain

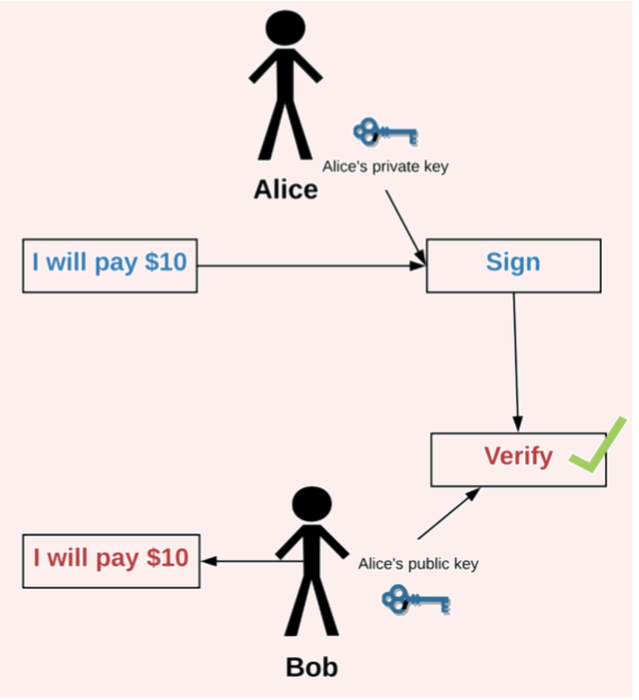
A variety of blockchains exist today, some exploring alternative architectures to test multiple aspects of scalability



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

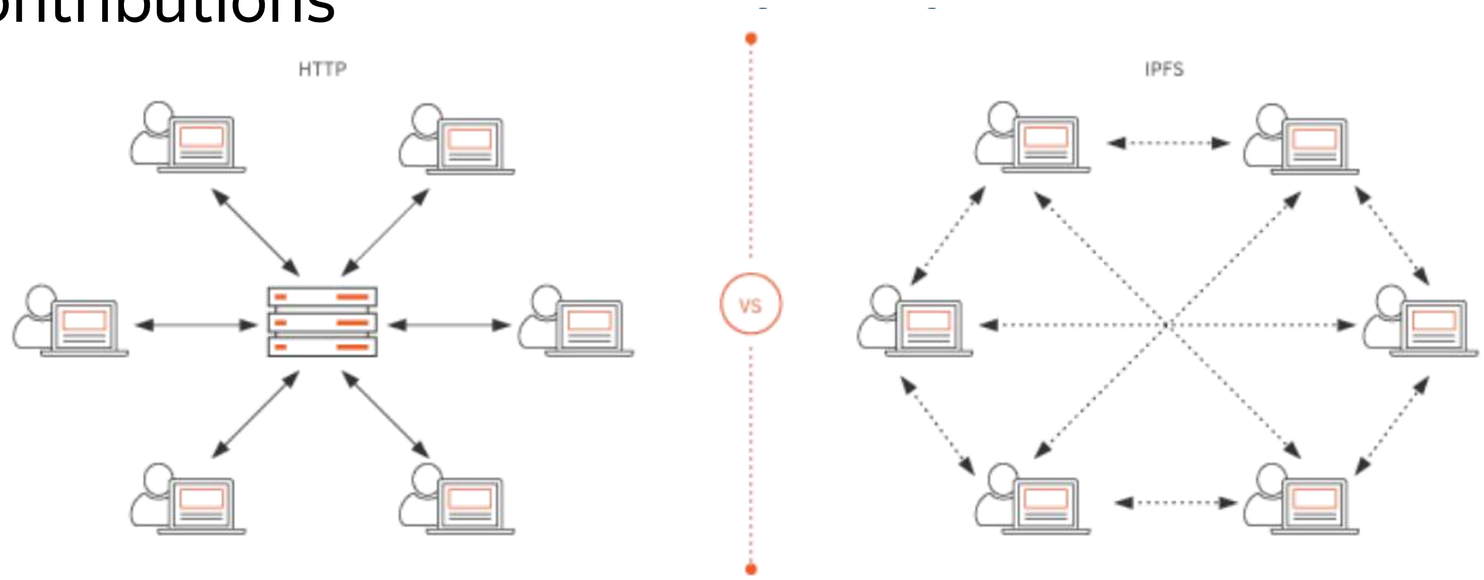
- Newer blockchains, Ethereum for instance, implement virtual machines that can execute byte code
- Smart contracts, implemented in this code allow binding between blockchain addresses and actions that are taken by the code
  - Typically the same code gets executed by all nodes in the network (extension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

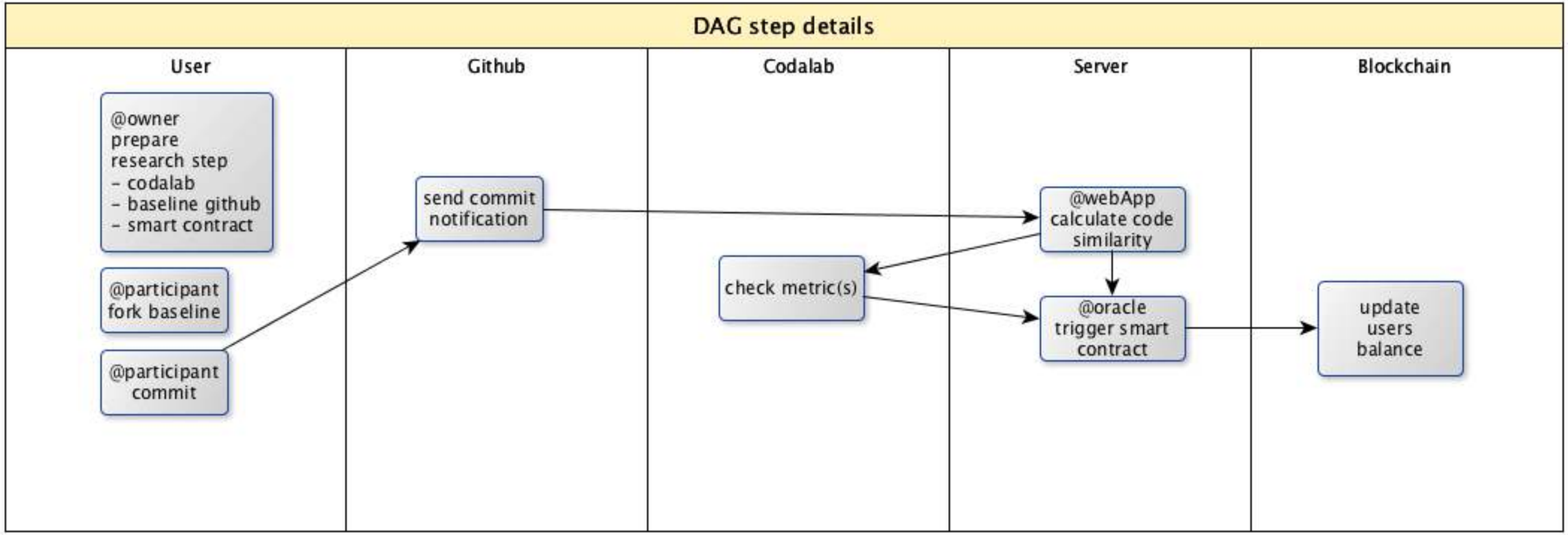
A simple example of a derived currency

Blockchain application

- Based on existing crypto-token
- Stores artifacts
- Manages computational resources allocation
- Records and rewards micro-contributions
  - Commit
  - Forks
- Removes bottle-neck and single vendor lock



Possible integration scenario for DAG step



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

Blockchain application. Challenges

Bootstrapping

- Organizational (institutes / online education systems)
- Marketing

Should there be feedback loop from solution running in production?

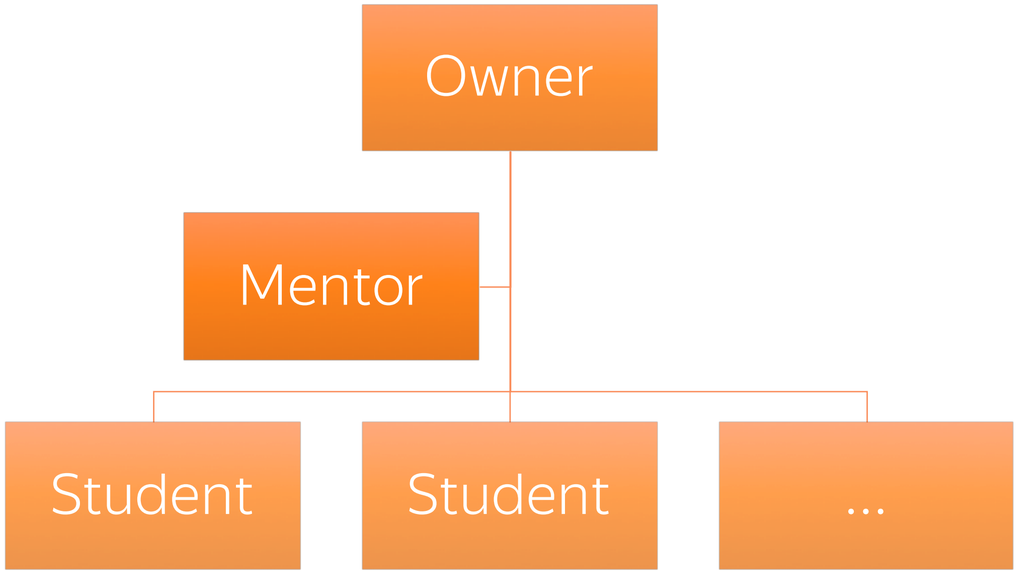
Translation of metrics into fair smart-contracts?

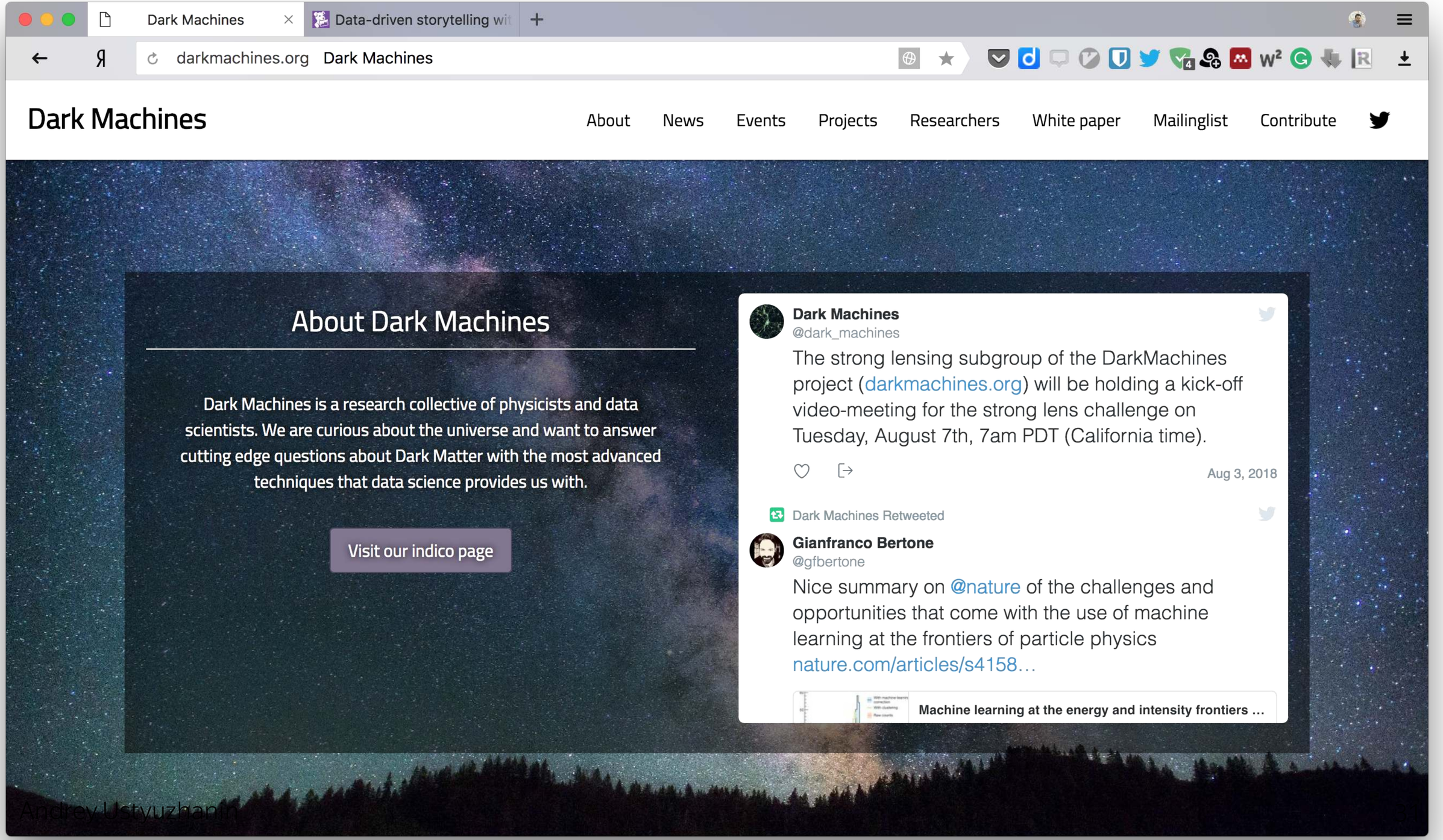
Social Uncertainties

- Would you outsource a challenge to such a platform?
  - Dataset?
  - Metric?
  - Mentors?
- Would you like to collaborate with unknown researchers on it?
- Would you accept decay of interest among participants?

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 state of the art**
- **More Challenges to solve:**
  - LHCb data compression
  - LArTPC 3D tracks identification
  - Quantum computer control





Andrey Ustyuzhanin

Questionnaire if you have a challenge to share

https://goo.gl/forms/PmYJBwyA3RVsPSHC2

Conclusion & Focus points

- **Plenty of cool stuff is driven by data**
  - in fundamental science and applied science
  - where Machine Learning can help
- **Machine Intelligence field is growing exponentially**
  - New algorithms and methods
  - Infrastructure
  - Driven by industry
- **Open-science, open-innovation. Demand for platform!**
  - Should be built on existing well-adopted services (i.e. github)
  - Should be flexible to support variety of processes used in scientific domains
  - Challenges: technological, sociological (communications), psychological
  - Distributed, diverse (blockchain)

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

Problem Structure