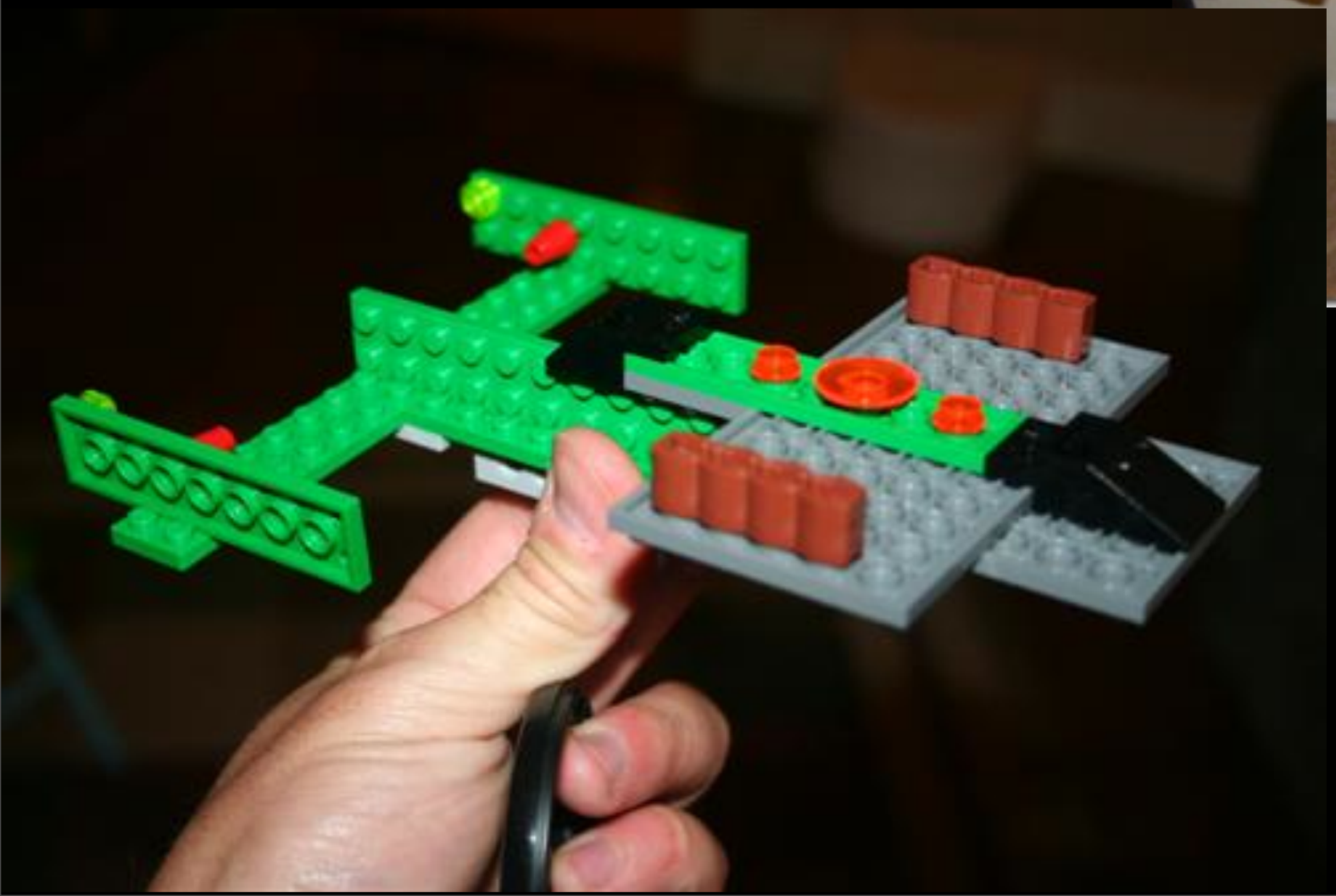
Yandex

Yandex

Towards reproducibility of research

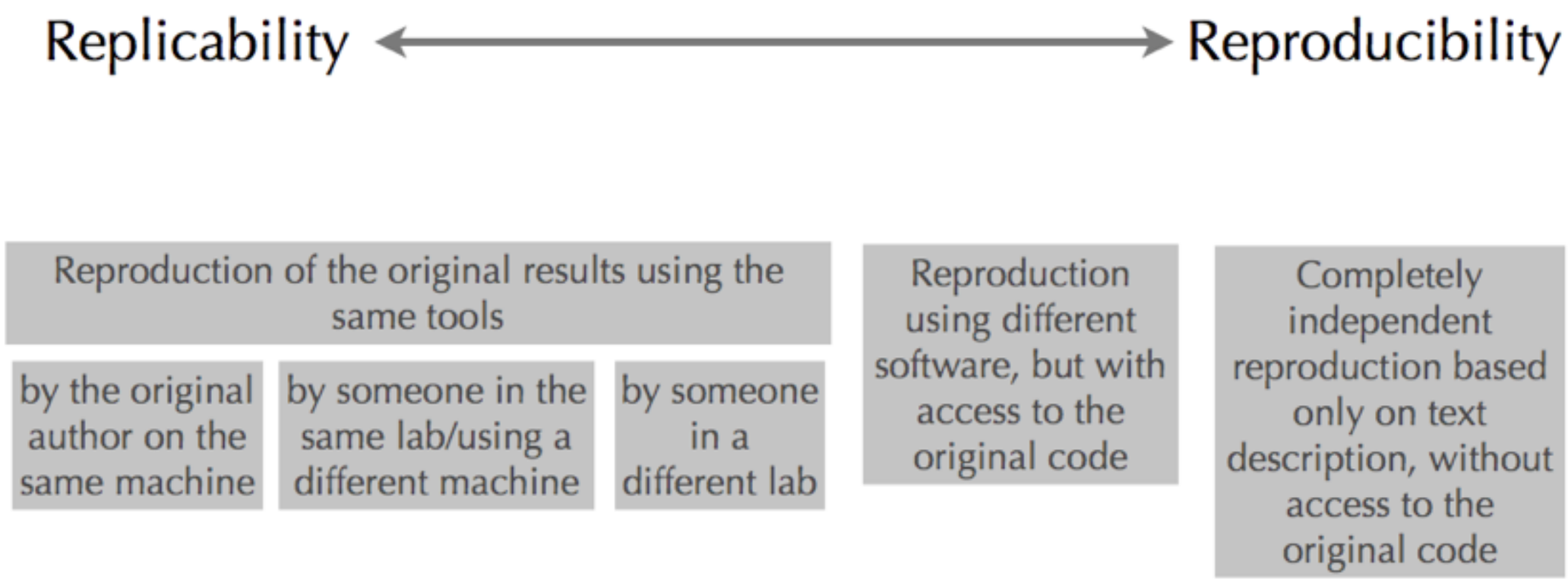
Andrey Ustyuzhanin




hov pod .com
hov pod
ACX

Apollo 13 problem

Reuse, Replicability, Reproducibility



Reproduce own research?

"I thought I used the same parameters but I'm getting different results"

"I can't remember which version of the code I used to generate figure 6"

"The new student wants to reuse that model I published three years ago but he can't reproduce the figures"

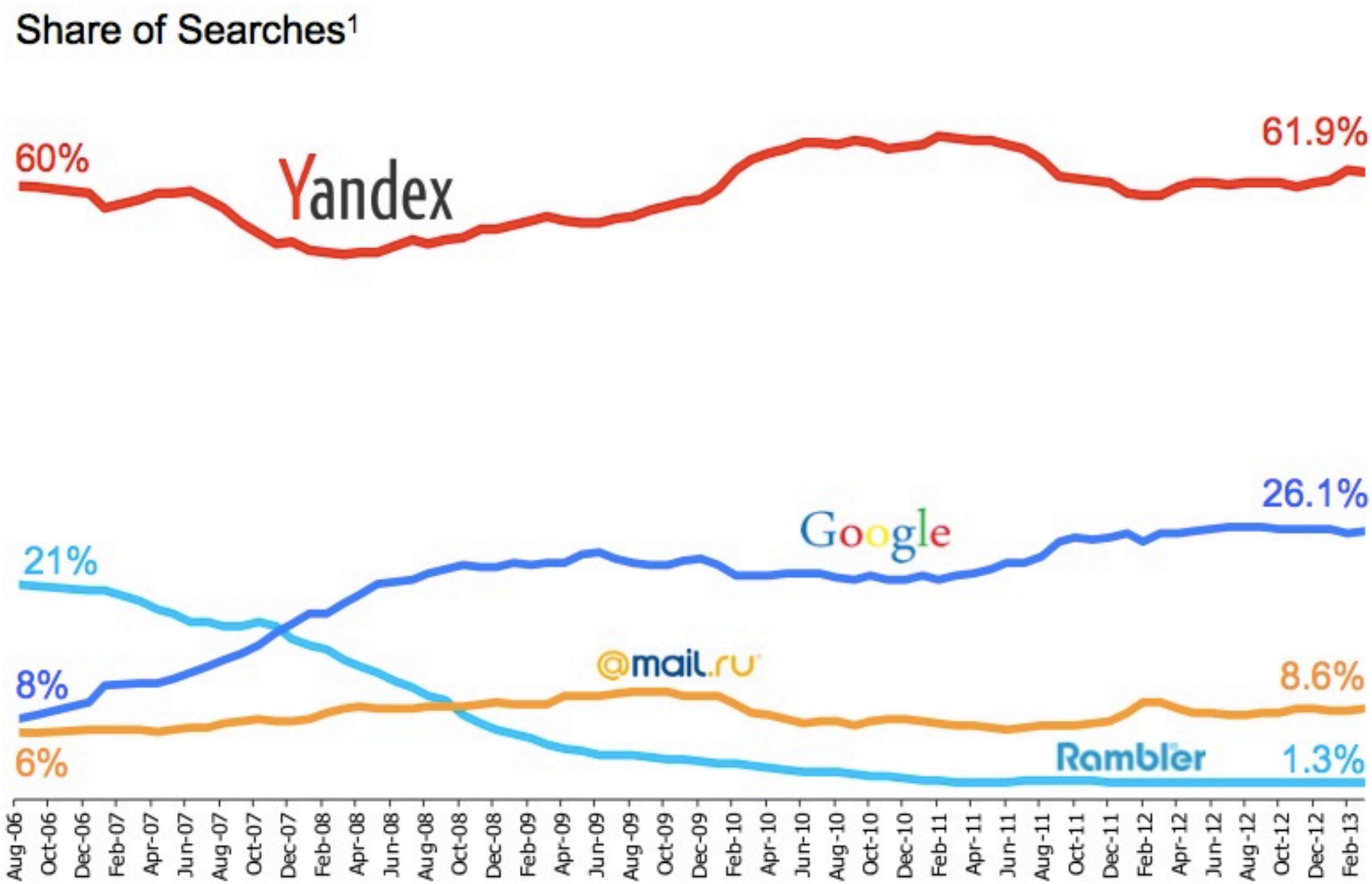
"It worked yesterday"

"Why did I do that?"

Reasons

- Complexity
- Entropy
- Human memory limitation
- No code access
- No library access
- No data access

Quick intro

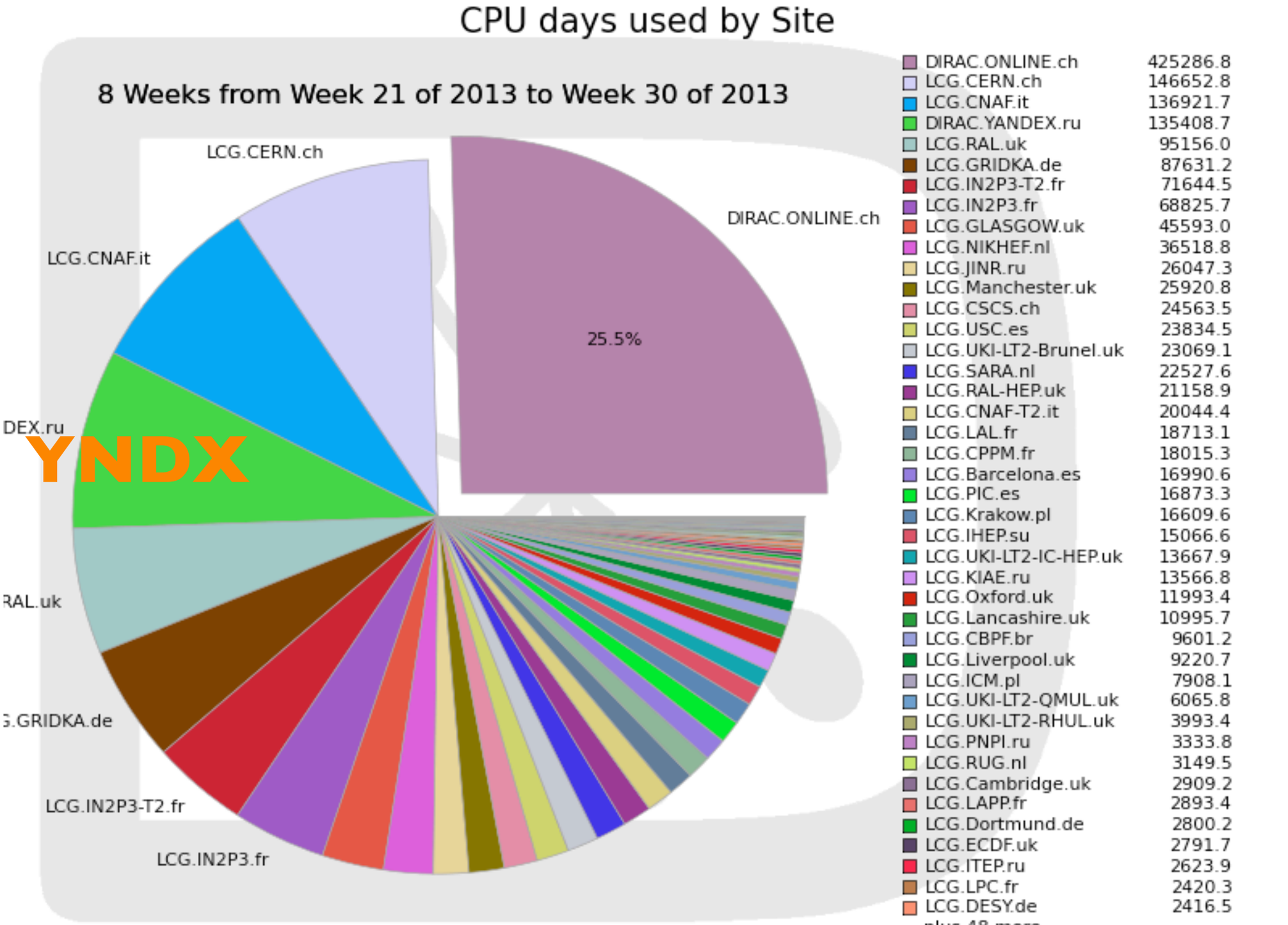


[1] Source: LiveInternet.ru, February 2013. Search traffic reflects Russian users to Russian websites and includes desktop and mobile

Areas of expertise

- Information retrieval
- Advertising
- Maps
- News aggregation
- Music online
- Video online
- Market online
- Web Browser
- Money online
- ...

LHCb Grid Tier 2



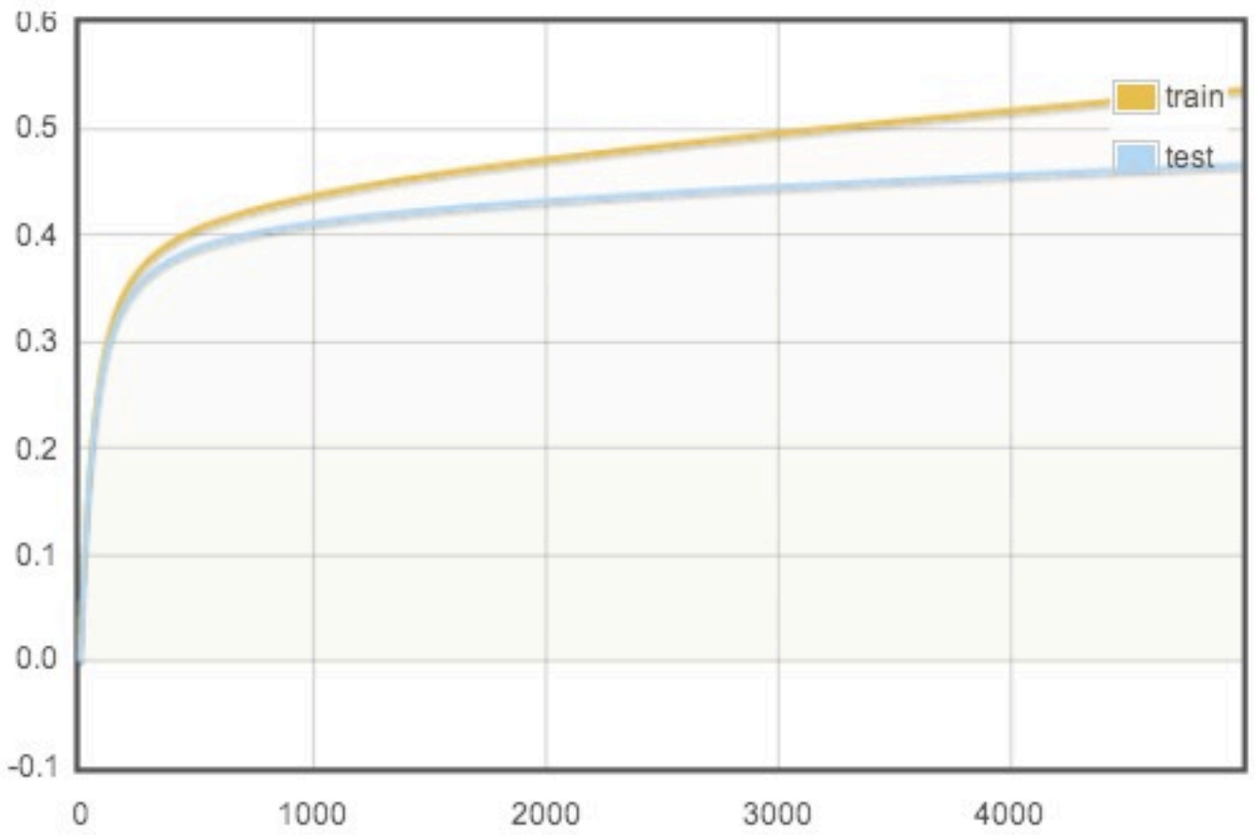
Event Filter

Home | Datasets | Pools | **Formulas** | Predictions | Tasks | About

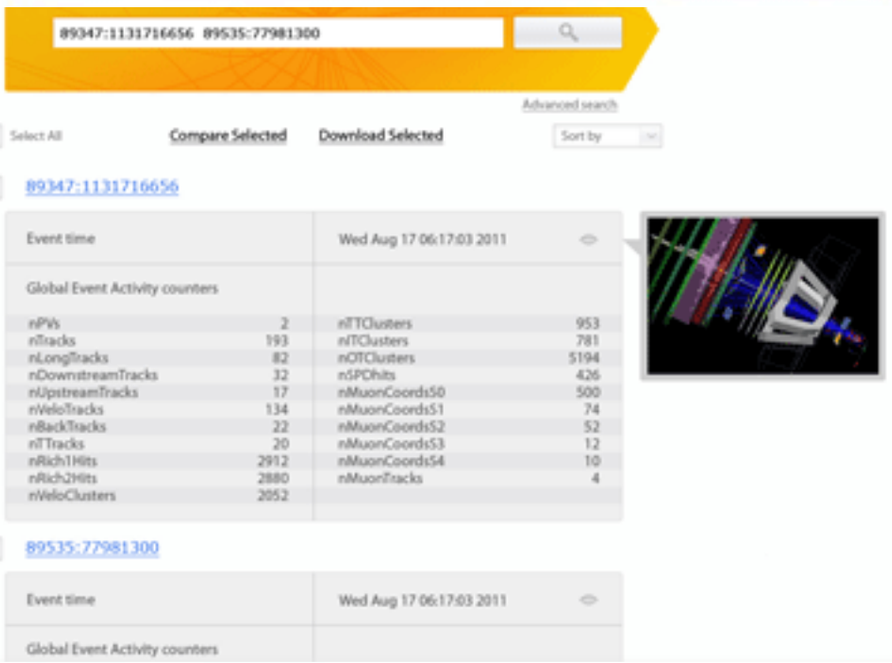
BDT9_Flavio1_i10000_w001_x64	Flavio1	B_s0_TAU_ps, mu_MINIPS, ...	336903	1	2013-04-19 10:12:49	-0.0543
D0mumu-test					013-05-06 14:45:23	1.4670
default					013-03-21 11:34:52	0.2430
form_1					013-02-23 18:00:32	-0.0007
marco_new_test					013-03-21 15:18:07	0.3142
mc_Bu2X-B2psiX_2					013-07-30 17:43:10	0.6070
mc_Bu2X-B2psiX_2a					013-07-30 17:53:19	0.4662
merger_3					013-08-01 09:37:05	0.0453
MN_13_Flavio1_i10000_w0					013-04-11 22:25:23	-0.3826

Formula «mc_Bu2X-B2psiX_2a»

Learning | Factors | Export



LHCb Event Indexer



Search is supported by Yandex

Research @ Yandex

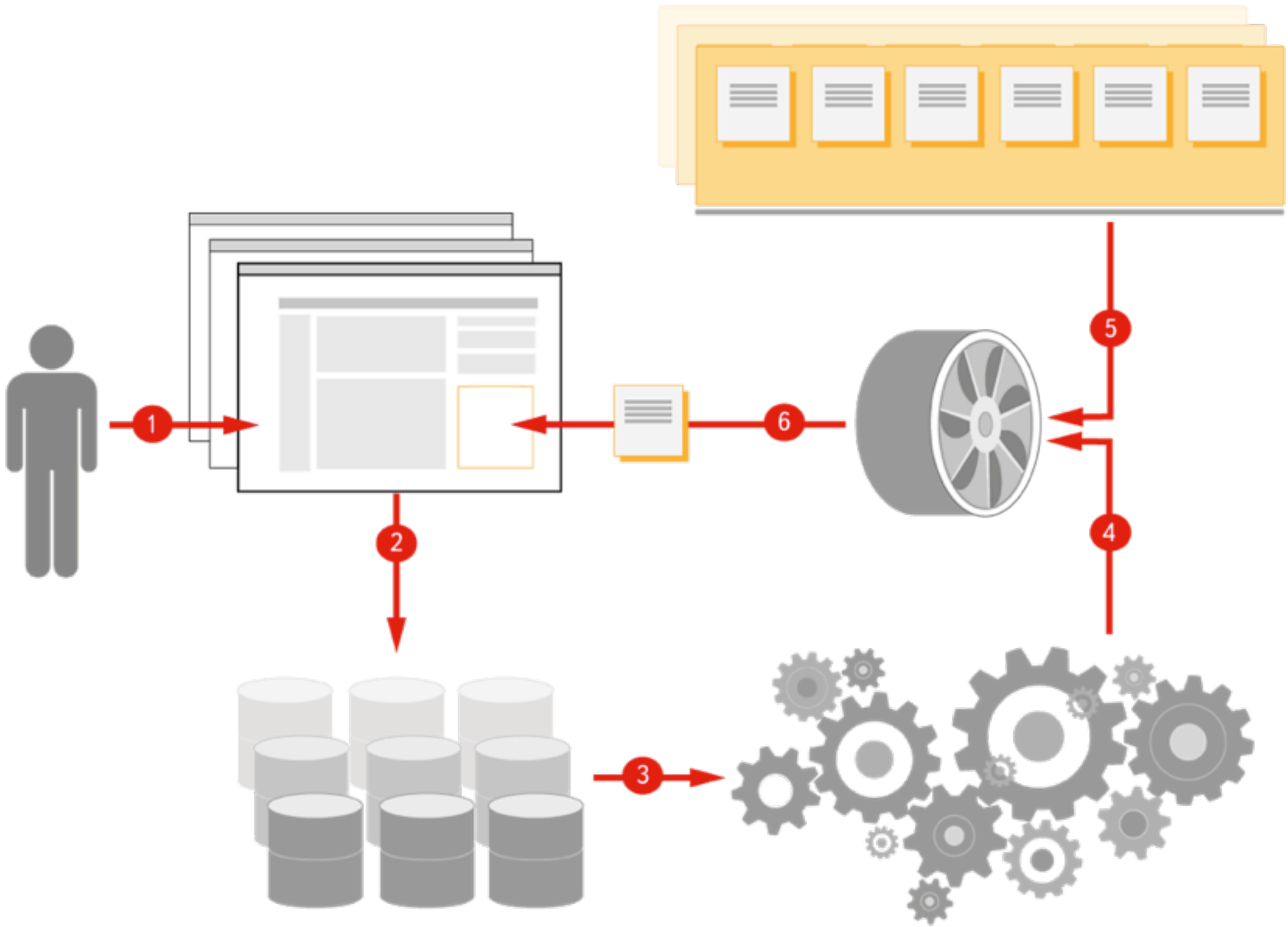
ACADEMIC RESEARCH

Publications

2013

- "Click Model-Based Information Retrieval Metrics". Aleksandr Chuklin, Pavel Serdyukov, Maarten de Rijke. 36th Annual ACM SIGIR Conference, SIGIR 2013.
- "User model-based metrics for offline query suggestion evaluation". Eugene Kharitonov, Craig Macdonald, Pavel Serdyukov, Iadh Ounis. 36th Annual ACM SIGIR Conference, SIGIR 2013.
- "Fresh BrowseRank". Maksim Zhukovskii, Andrey Khropov, Gleb Gusev, Pavel Serdyukov. 36th Annual ACM SIGIR Conference, SIGIR 2013.
- "Studying Page Life Patterns in Dynamical Web". Alexey Tikhonov, Gleb Gusev, Ivan Bogatyy, Pavel Burangulov, Liudmila Ostroumova, Vitaliy Koshelev. 36th Annual ACM SIGIR Conference, SIGIR 2013.
- "Predicting the Audience Size of a Tweet". Andrey Kupavskii, Alexey Umnov, Gleb Gusev and Pavel Serdyukov. The 7th International AAAI Conference on Weblogs and Social Media (ICWSM 2013). [pdf]
- "Introducing search behavior into browsing based models of page's importance". Maksim Zhukovskiy, Andrey Khropov, Gleb Gusev and Pavel Serdyukov. The 22nd International World Wide Web Conference (WWW 2013). [pdf]

Applied research



Principles

- Modularity
- Computational Measurability
- Automation
- Transparency & Sharing

Processes

- Agile
  - Scrum
  - Kanban
- Testing & Verificaiton
- Continuous Integration
- Wheel reinventing

Practices

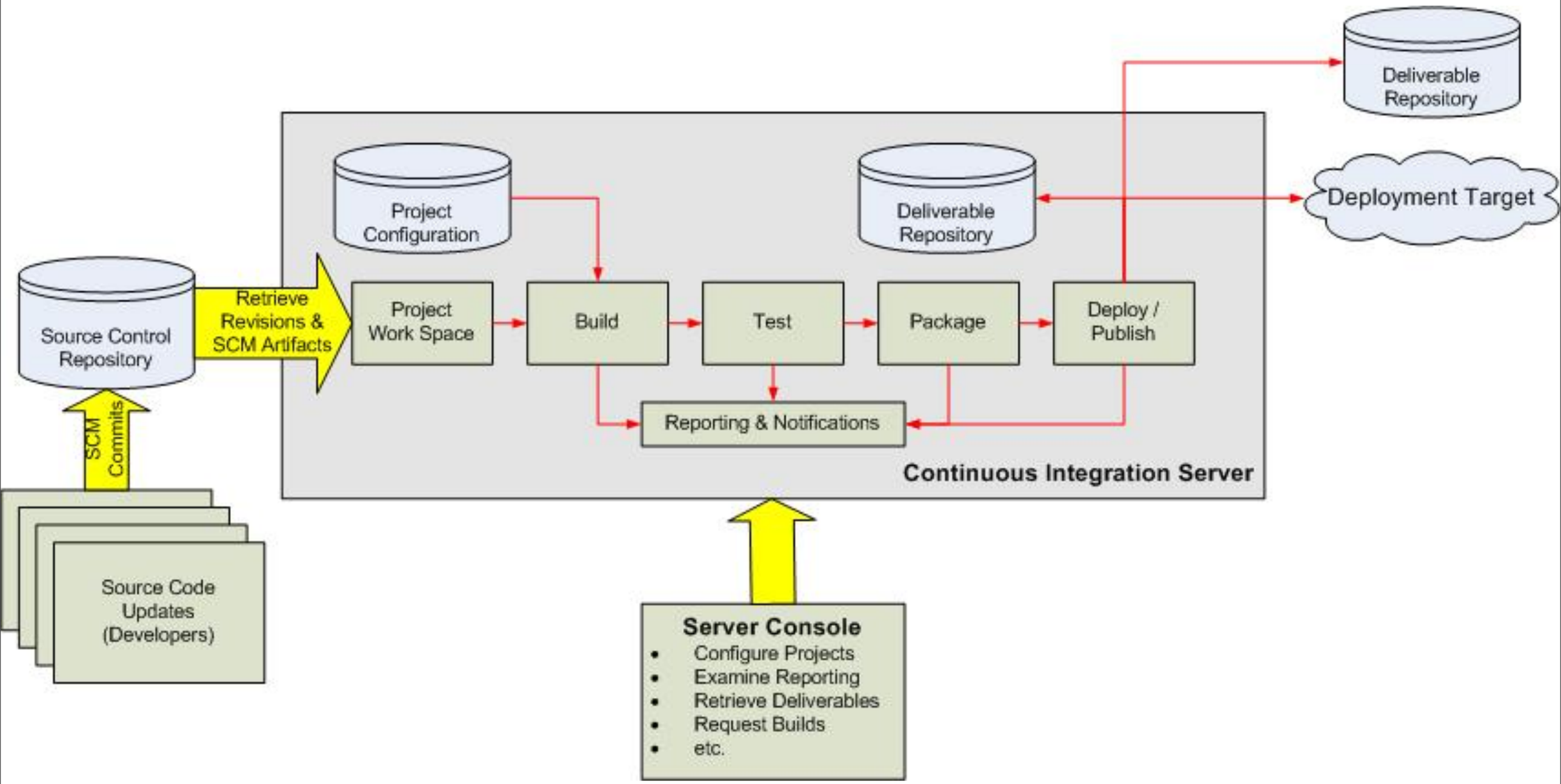
- Use version control
- Write tests & use it
- Prioritize code robustness
- Maintain consistent, repeatable computer environment
- Separate code from configuration
- Separate model definition from experiment
- Share your code

Technologies to support

- Jira (T, Me)
- Jenkins & build automation (Me, A, Mo)
- Git/github (T, A)
- Wiki (T)
- iPython (Me, A, Mo, T)
- Monitoring (Me, Mo)
- FML (T, A, Me, Mo)

Jenkins

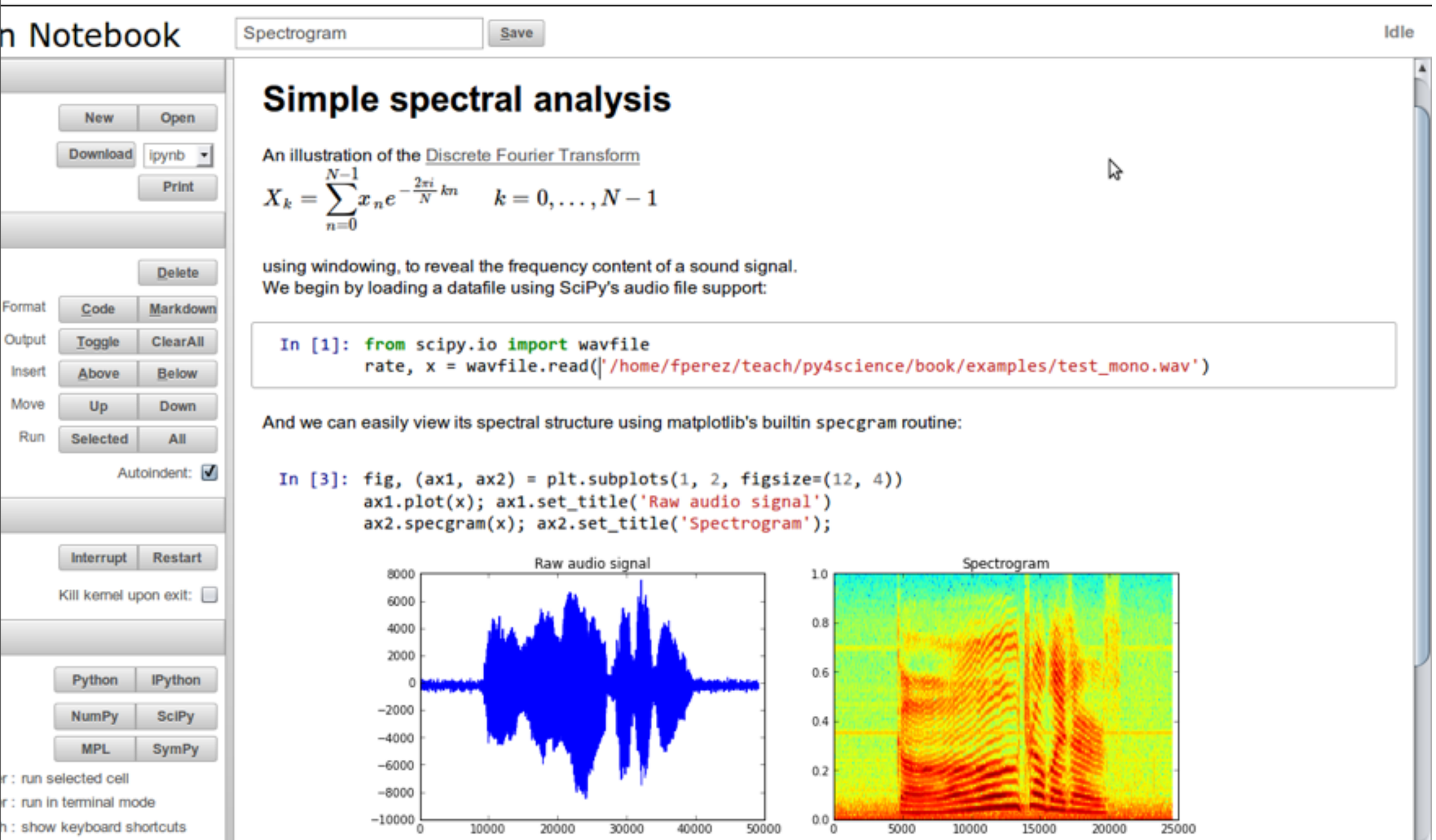
- Single source repository
- Commit often
- Make your Build self-testing
- Automate the Build
- Build fast



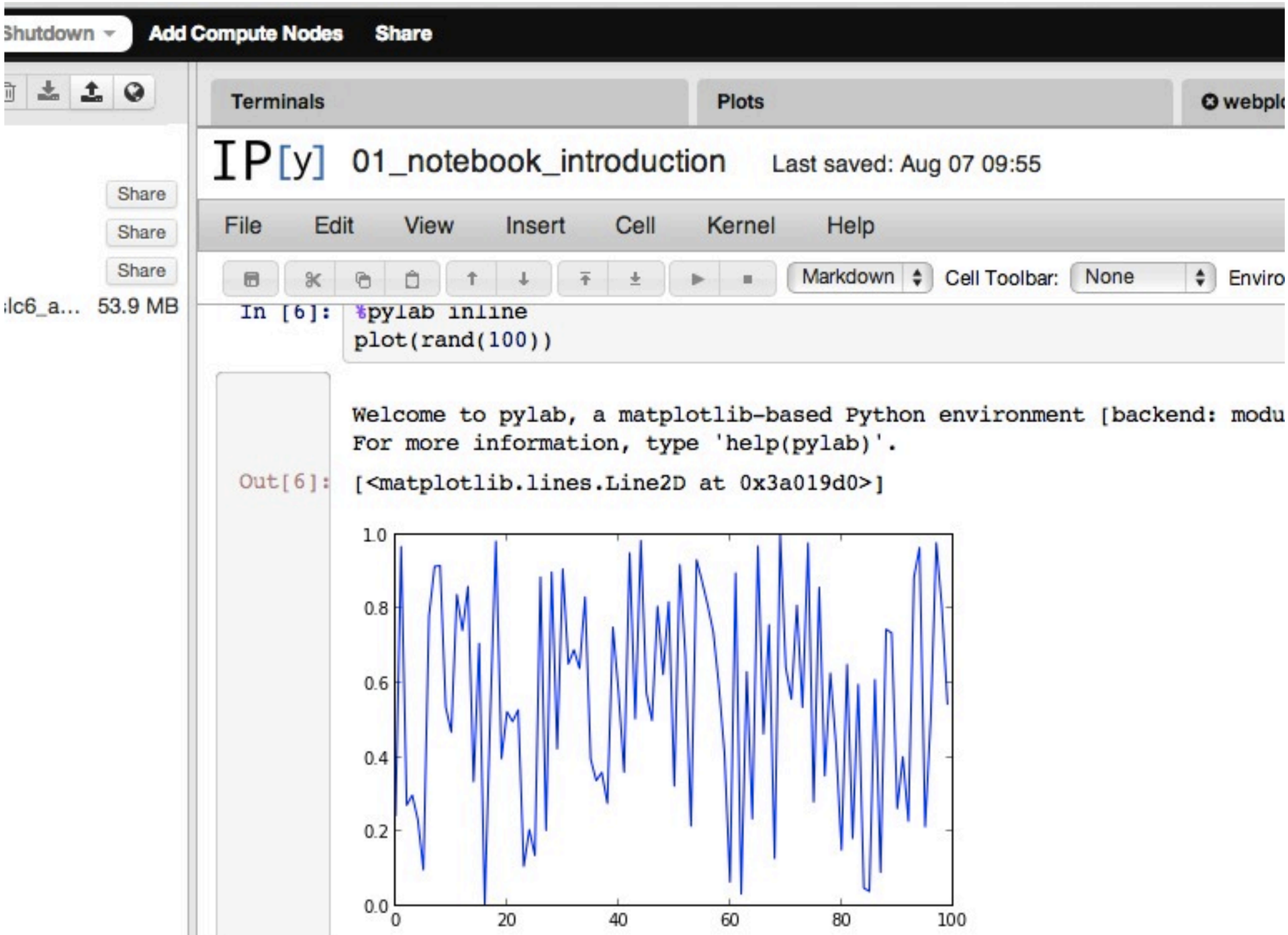
http://bit.ly/130e2X3

iPython

- Literation programming
- Literate computation



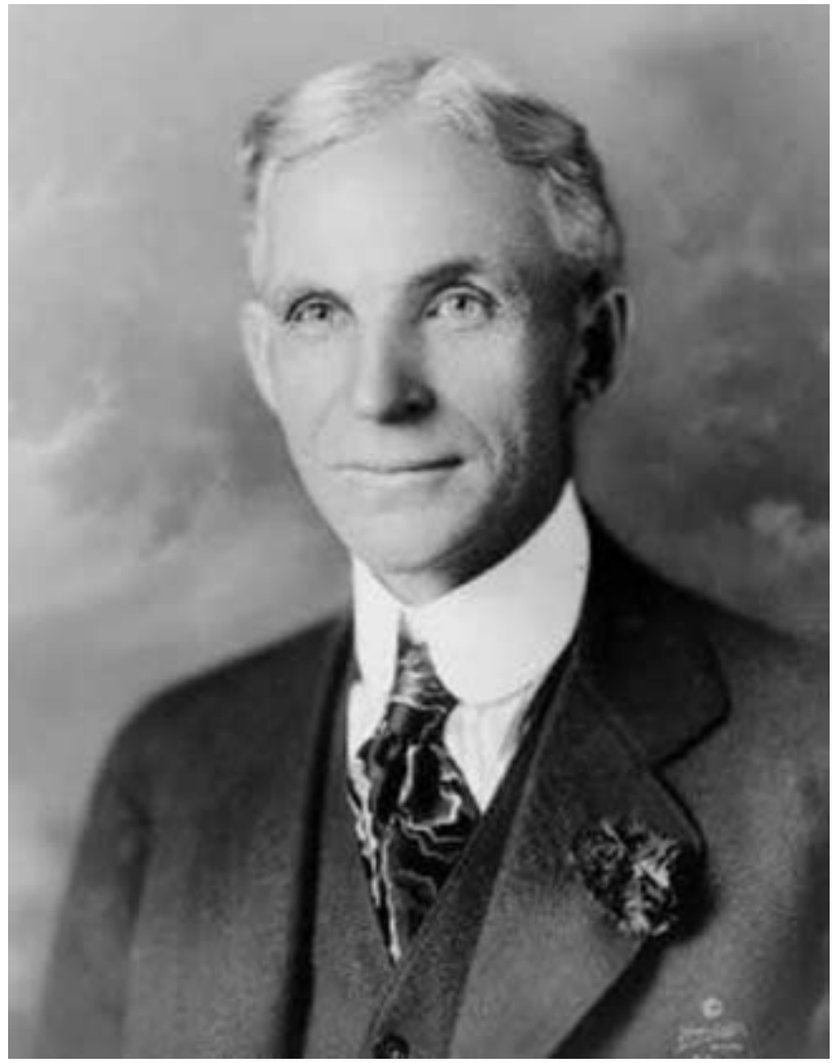
Wakari.io



Security

Another meta transition awaiting...

«How research & learning can be automated?»



fml ▾ R1 Index

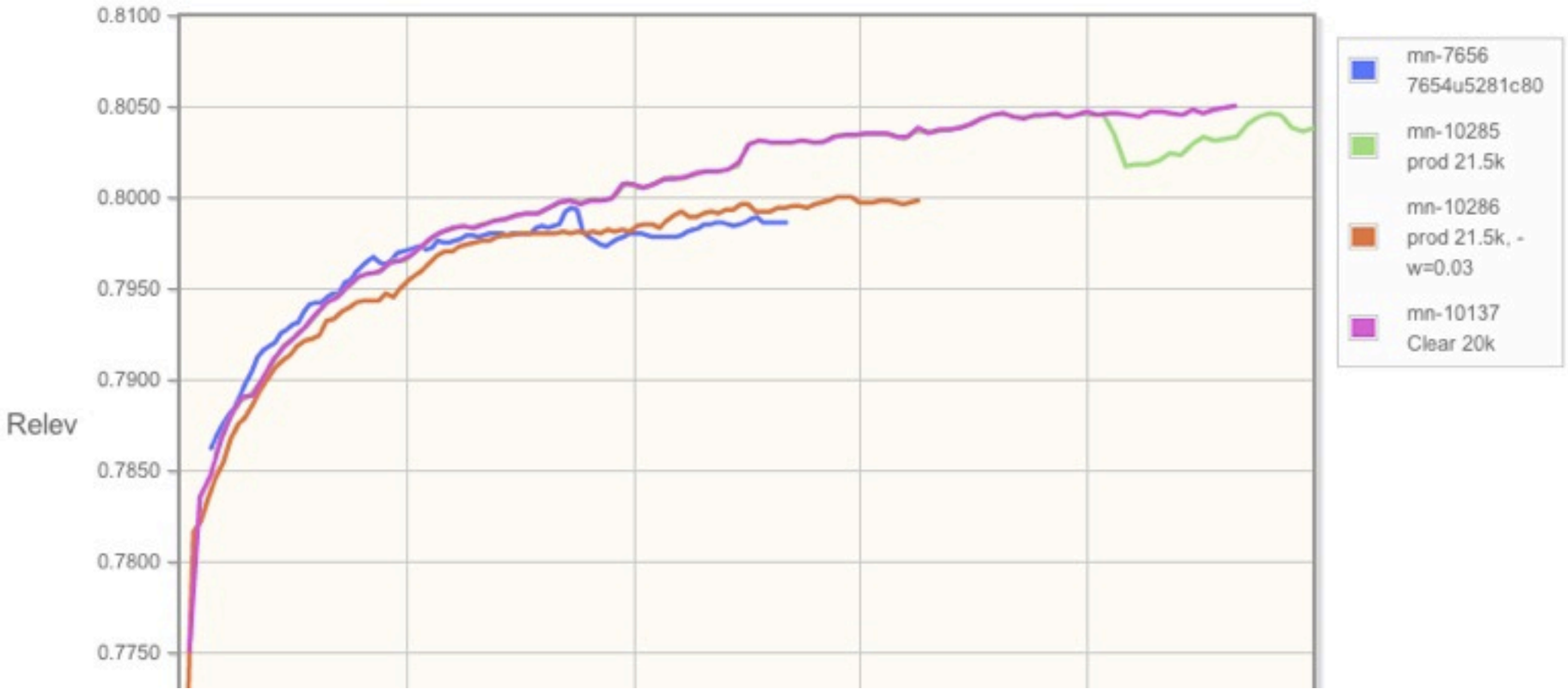
- Home
  - Download pool
  - Upload pool
  - Eval Feature
  - Make Pool
  - Join Features
  - Train Formula
  - Import Formula
- Overview
  - Eval Queue
  - Pools
  - Formulas
  - Comparisons
  - Wizard
- Experiments
  - Factor table
- Formulator
  - Kosher Uploads
  - Boosts
- Monitoring
  - Factors
  - Eval Feature Tests
- Tools

anaderi

Model comparison

Pool:	**Russia Common** / 2013-08-05-00-53 #121956
Test:	test.tsv.gz
Factors:	
Owner:	mklimushkin
Created:	2013-08-06-14-02
Code for wiki:	show

Formula Plot



Consequences

- easy code reading & understanding
- code reuse
- replicability
- reproducibility
- meta-transition
- smoother transition of knowledge & experience
- easier to crawl up on the shoulder of giants

Other points of interest

- Sumatra (http://neuralensemble.org/sumatra/)
- Workflows (e.g. VisTrails)
- Git-annex (http://git-annex.branchable.com/)
- PyCon 2013, Reproducible science track (http://pyvideo.org/search/?models=videos.video&q=reproducible)
- ttyrec (http://0xcc.net/ttyrec/)

Yandex

Andrey Ustyuzhanin

anaderi@yandex-team.ru

Thank you