

Yandex

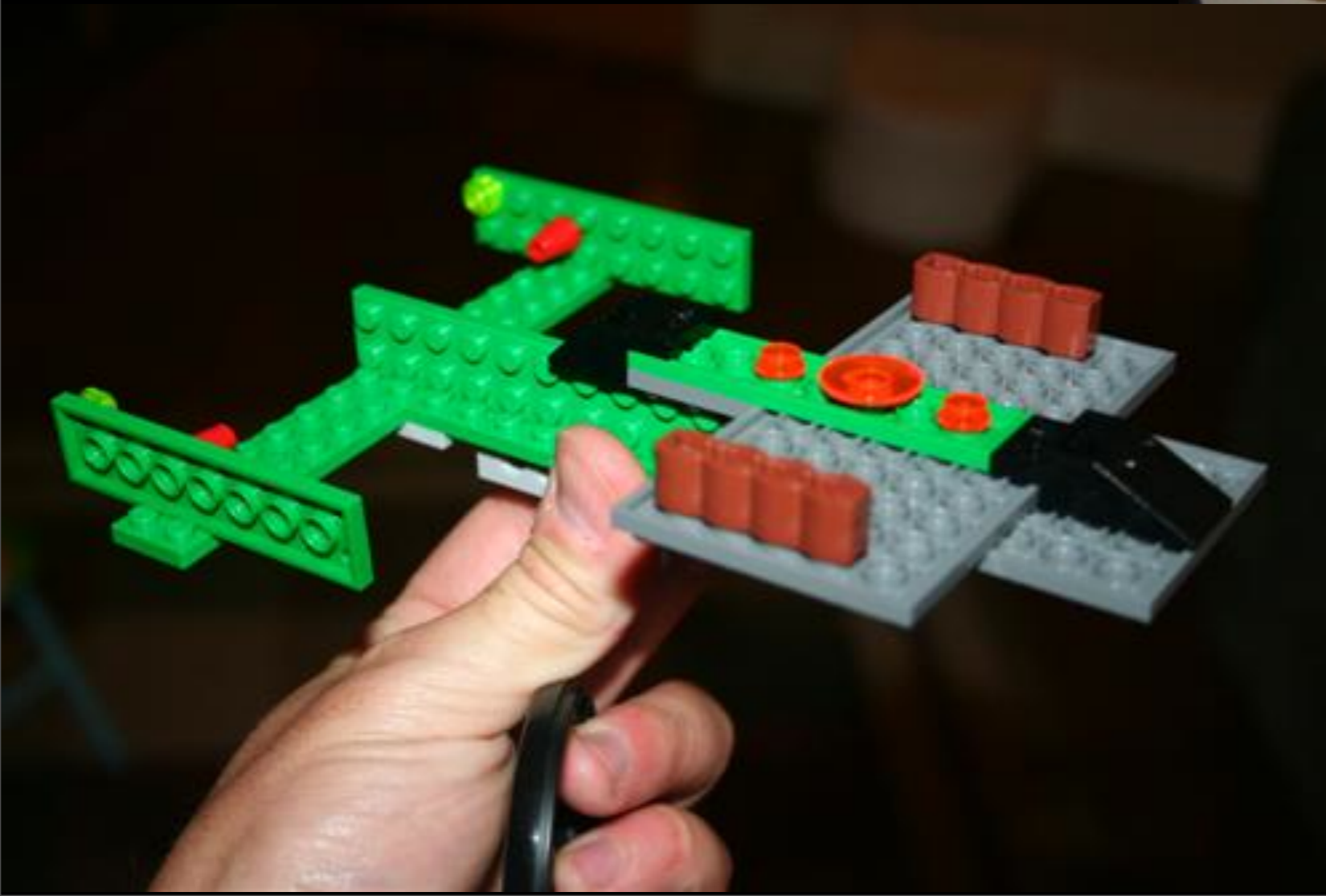


Yandex

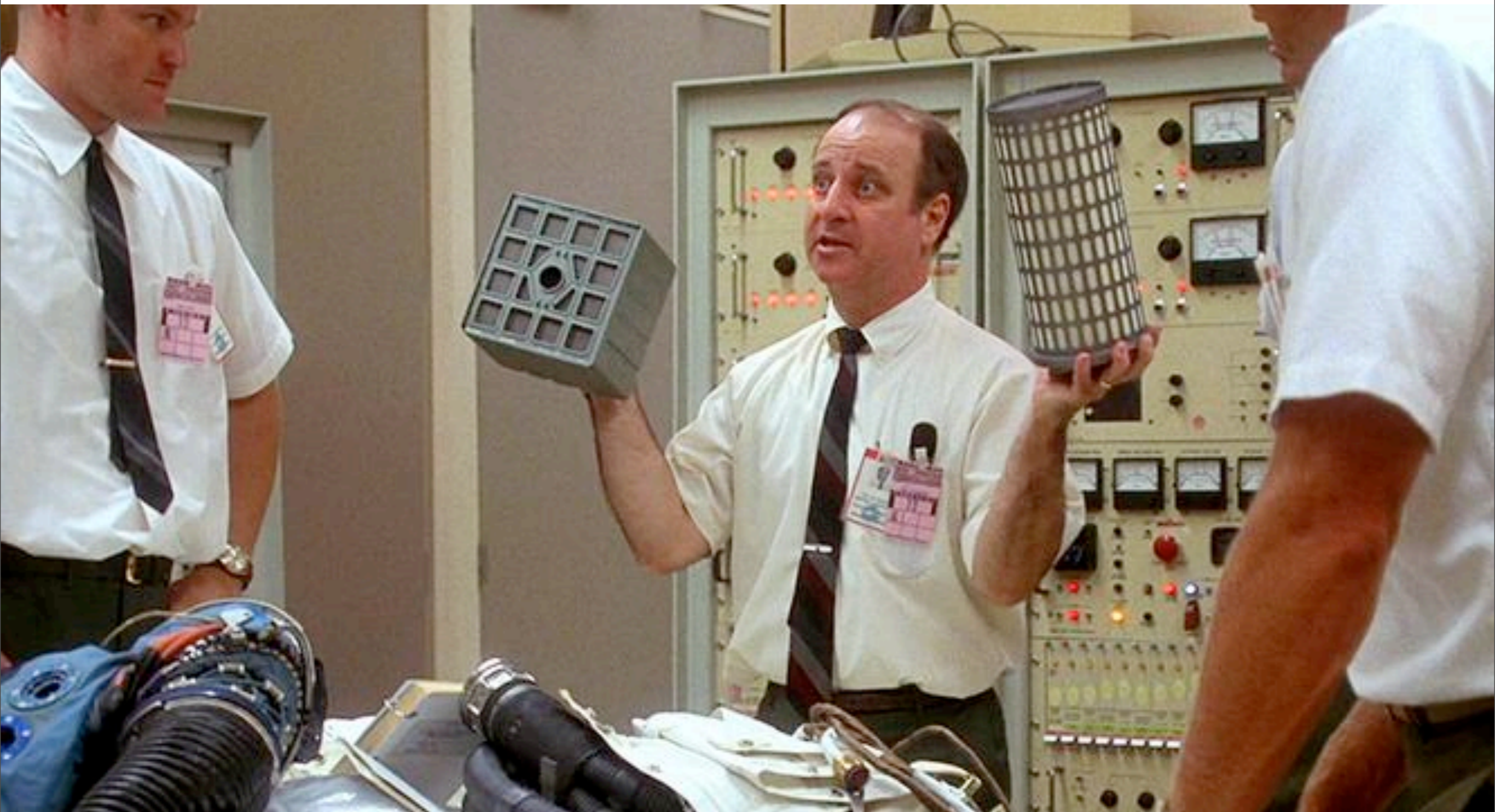
Towards reproducibility of research

Andrey Ustyuzhanin









Apollo 13 problem

Reuse, Replicability, Reproducibility

Replicability ←————→ Reproducibility

Reproduction of the original results using the same tools

by the original author on the same machine

by someone in the same lab/using a different machine

by someone in a different lab

Reproduction using different software, but with access to the original code

Completely independent reproduction based only on text description, without access to the original code

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”

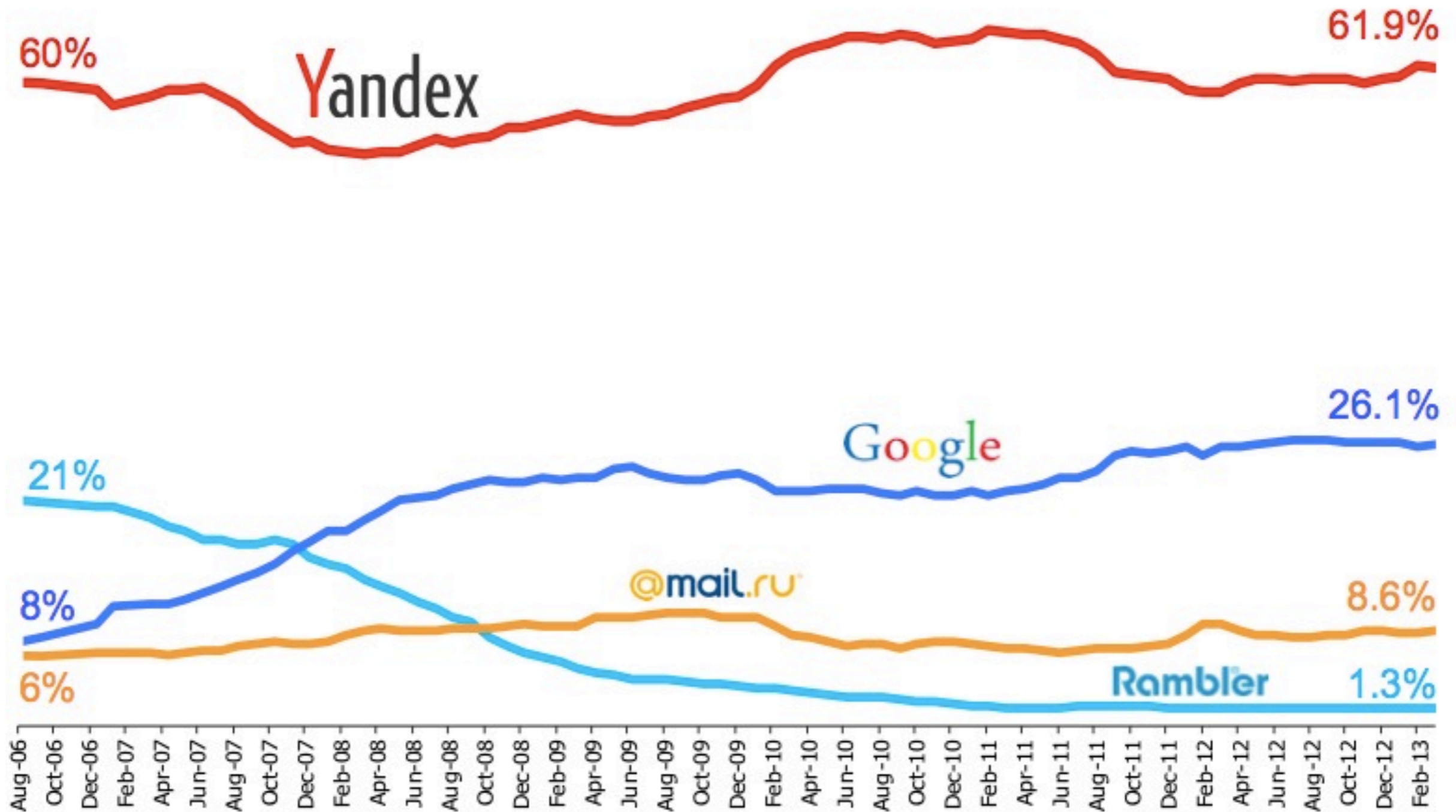
“Why did I do that?”

Reasons

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

Quick intro

Share of Searches¹



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

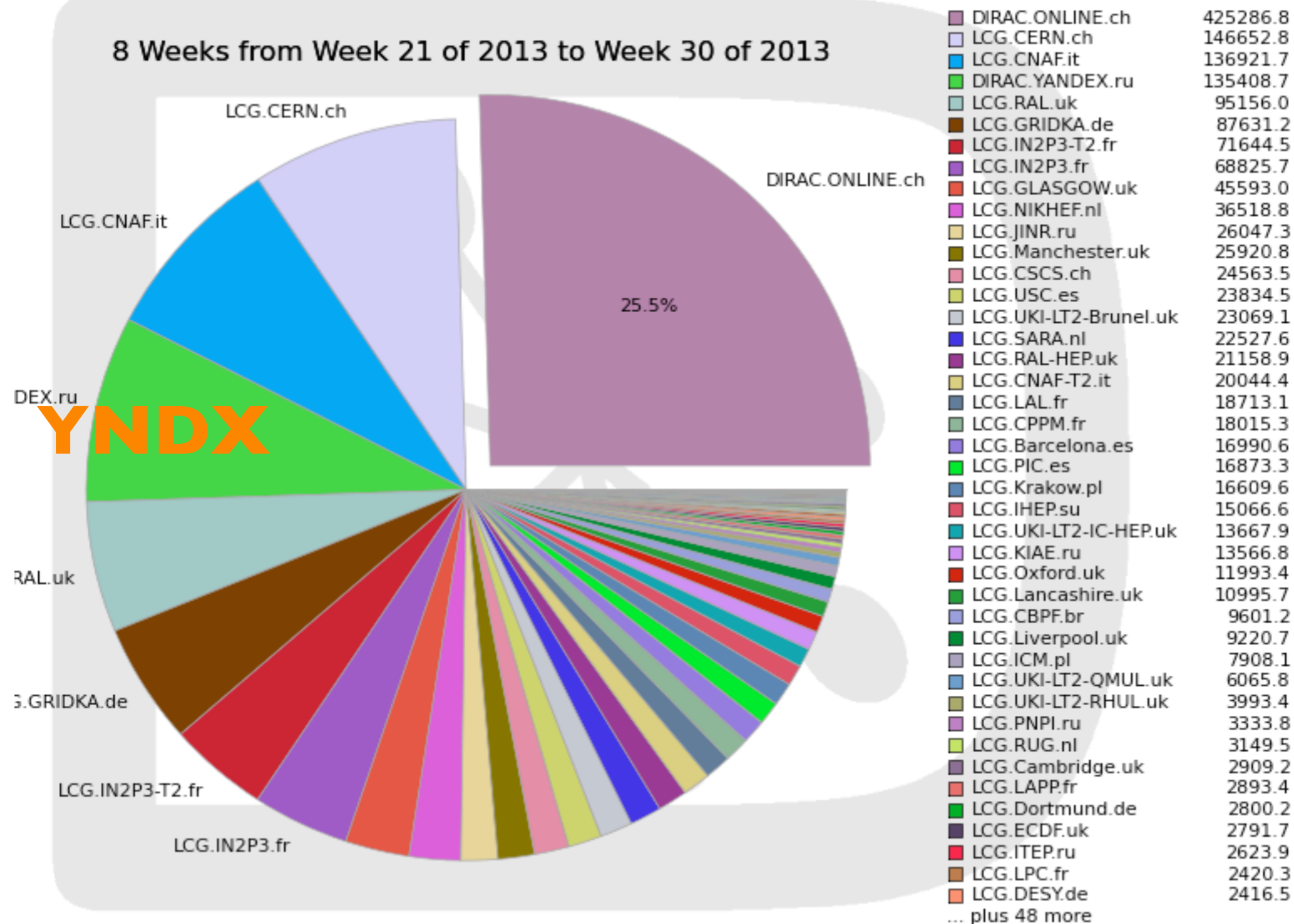
Areas of expertise

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

LHCb Grid Tier 2

CPU days used by Site

8 Weeks from Week 21 of 2013 to Week 30 of 2013



Generated on 2013-08-07 07:16:40 UTC

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					2013-05-06 14:45:23	1.4670		
default					2013-03-21 11:34:52	0.2430		
form_1					2013-02-23 18:00:32	-0.0007		
marco_new_test					2013-03-21 15:18:07	0.3142		
mc_Bu2X-B2psiX_2					2013-07-30 17:43:10	0.6070		
mc_Bu2X-B2psiX_2a					2013-07-30 17:53:19	0.4662		
merger_3					2013-08-01 09:37:05	0.0453		
MN_13_Flavio1_i10000_w01					2013-04-11 22:25:23	-0.382		

Formula «mc_Bu2X-B2psiX_2a»

Learning Factors Export

Time	Train	Test
0	0.0	0.0
1000	0.45	0.40
2000	0.48	0.43
3000	0.50	0.44
4000	0.52	0.45

LHCb Event Indexer



89347:1131716656 89535:77981300

Select All Compare Selected Download Selected Sort by

89347:1131716656

Event time	Wed Aug 17 06:17:03 2011		
Global Event Activity counters			
nPvs	2	nTClusters	953
nTracks	193	nOTClusters	781
nLongTracks	82	nOTClusters	5194
nDownstreamTracks	32	nSPChits	426
nUpstreamTracks	17	nMuonCoords50	500
nWeloTracks	134	nMuonCoords51	74
nBackTracks	22	nMuonCoords52	52
nITTracks	20	nMuonCoords53	12
nRich1Hits	2912	nMuonCoords54	10
nRich2Hits	2880	nMuonTracks	4
nWeloClusters	2052		

89535:77981300

Event time	Wed Aug 17 06:17:03 2011
Global Event Activity counters	

Comparison of events

All fields Unique fields

Run Number	98881	98881	98881
Event Number	32481496	32481496	32481496
Event time	Wed Aug 17 06:17:03 2011	Wed Aug 17 06:17:03 2011	Wed Aug 17 06:17:03 2011
Tags	('DODB', 'head-20110914') (LHCBCOND, 'head-20110914') (DQFLAGS, 'tt-20110126') (ONLINE, 'HEAD')	('DODB', 'head-20110914') (LHCBCOND, 'head-20110914') (DQFLAGS, 'tt-20110126') (ONLINE, 'HEAD')	('DODB', 'head-20110914') (LHCBCOND, 'head-20110914') (DQFLAGS, 'tt-20110126') (ONLINE, 'HEAD')
Application for Reconstruction	Brunel v41r1	Brunel v41r1	Brunel v41r1
Fired Stripping tags & number of candidates	('StrippingD2hhh_KKPLineDecision': 1L, 'StrippingDsForPromptCharmDecision': 1L, 'StrippingStreamCharmDecision': 0L)	('StrippingD2hhh_KKPLineDecision': 1L, 'StrippingDsForPromptCharmDecision': 1L, 'StrippingStreamCharmDecision': 0L)	('StrippingD2hhh_KKPLineDecision': 1L, 'StrippingDsForPromptCharmDecision': 1L, 'StrippingStreamCharmDecision': 0L)



Search is supported by



Research @ Yandex

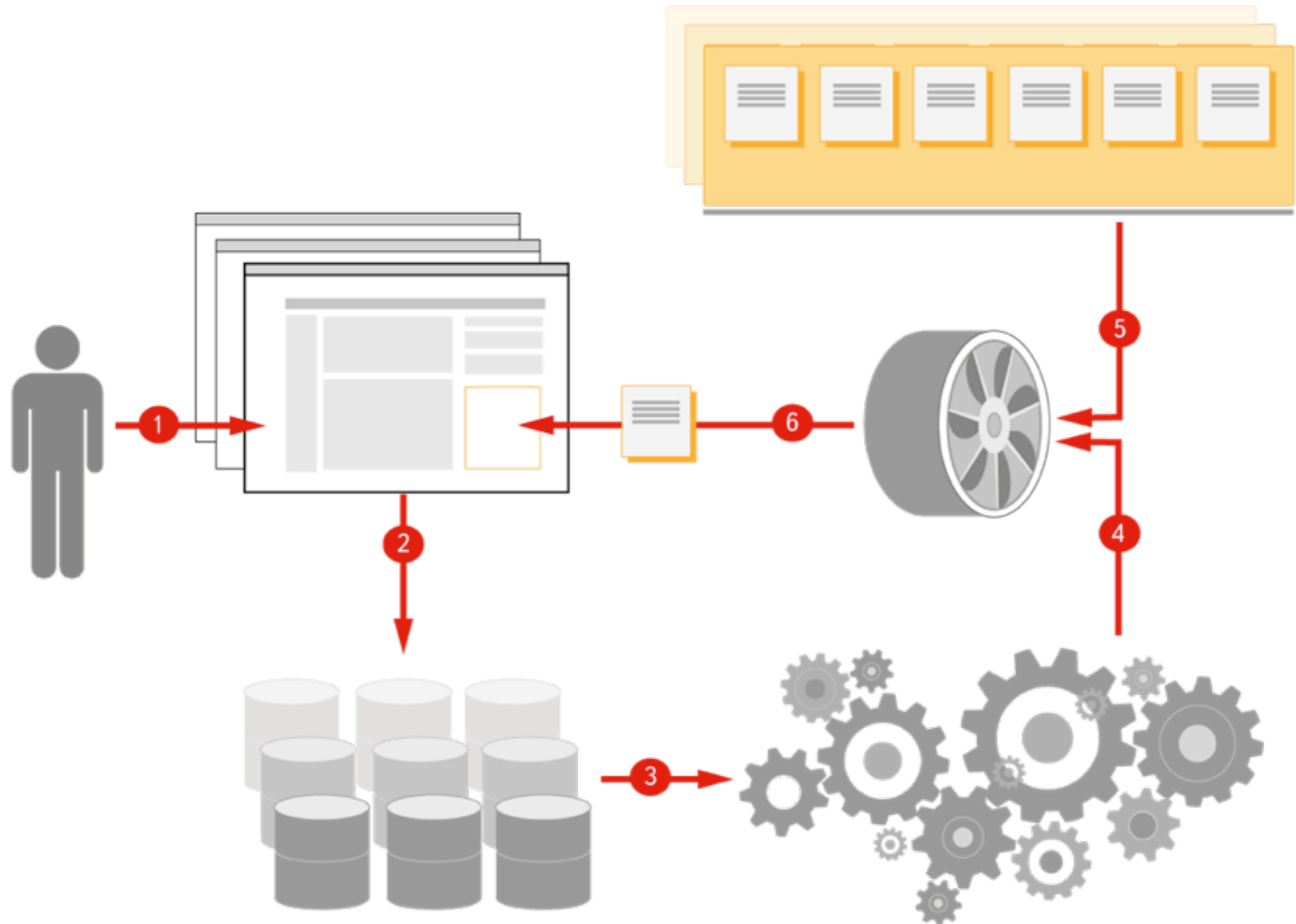
ACADEMIC RESEARCH

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 & Verification
- Continuous Integration
- Wheel reinventing

Practices

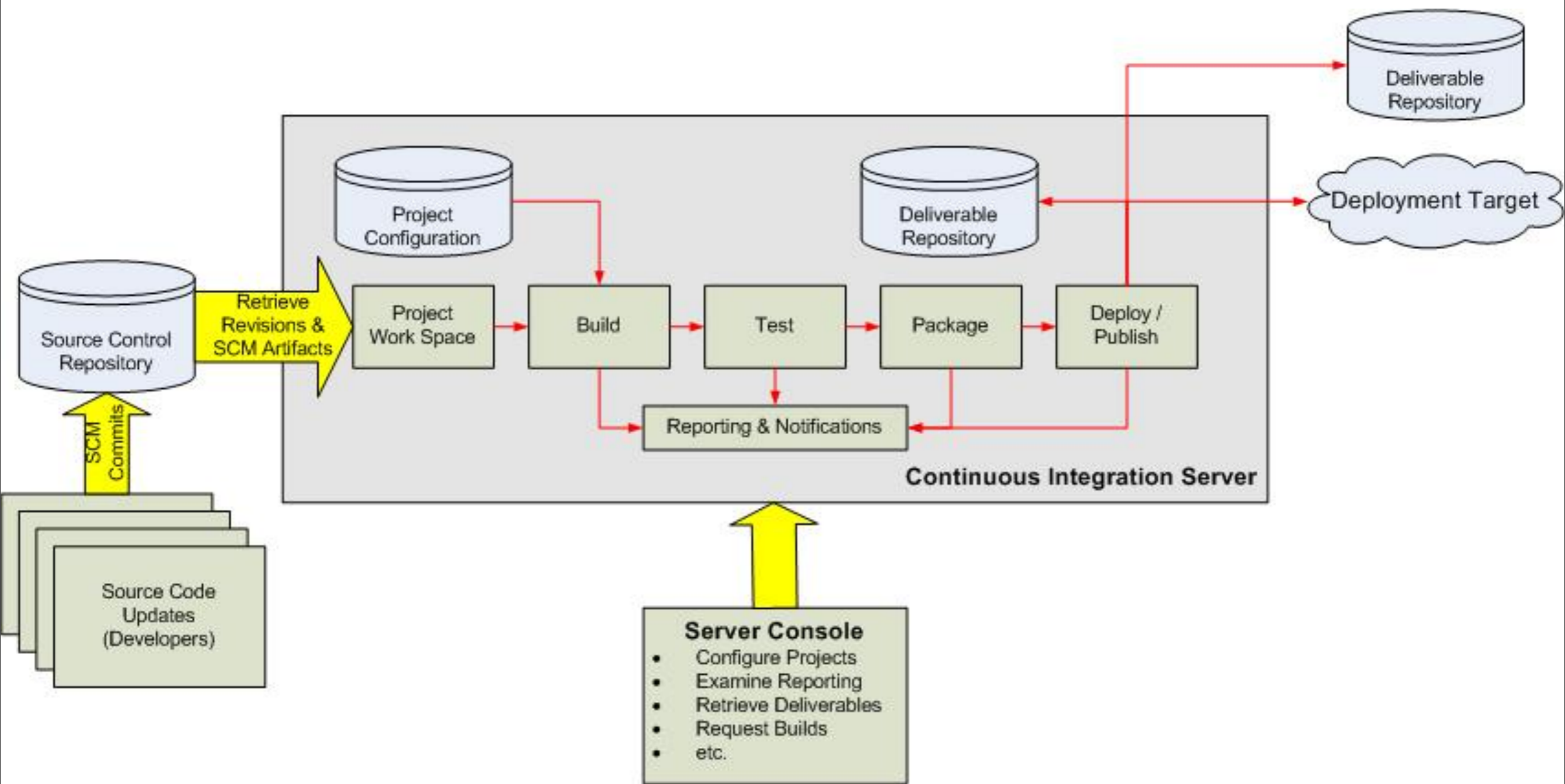
- Use version control
- Write tests & use it
- Prioritize code robustness
- Maintain consistent, repeatable computer environment
- Separate code from configuration
- Separate model definition from experiment
- Share your code

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/I30e2X3>

iPython

- Iteration programming
- Literate computation

Python Notebook Spectrogram Save Idle

New Open
Download ipynb
Print
Delete

Format Code Markdown
Output Toggle ClearAll
Insert Above Below
Move Up Down
Run Selected All
Autoindent:

Interrupt Restart
Kill kernel upon exit:

Python IPython
NumPy SciPy
MPL SymPy

run selected cell
run in terminal mode
show keyboard shortcuts

Simple spectral analysis

An illustration of the [Discrete Fourier Transform](#)

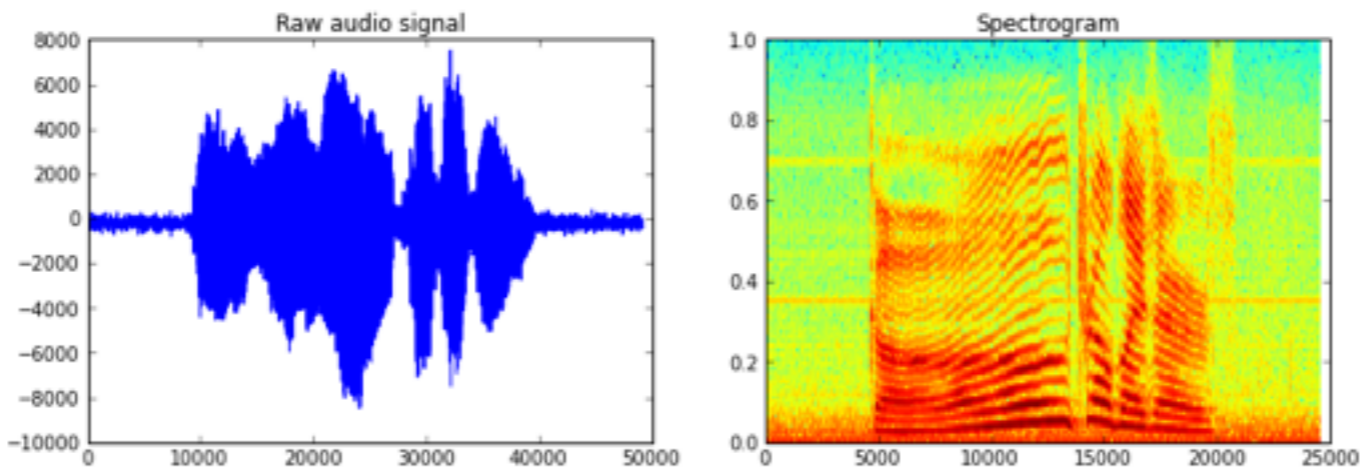
$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \quad k = 0, \dots, N-1$$

using windowing, to reveal the frequency content of a sound signal.
We begin by loading a datafile using SciPy's audio file support:

```
In [1]: from scipy.io import wavfile  
rate, x = wavfile.read('/home/fperez/teach/py4science/book/examples/test_mono.wav')
```

And we can easily view its spectral structure using matplotlib's builtin specgram routine:

```
In [3]: fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 4))  
ax1.plot(x); ax1.set_title('Raw audio signal')  
ax2.specgram(x); ax2.set_title('Spectrogram');
```



Wakari.io

Shutdown Add Compute Nodes Share

Terminals Plots webpl

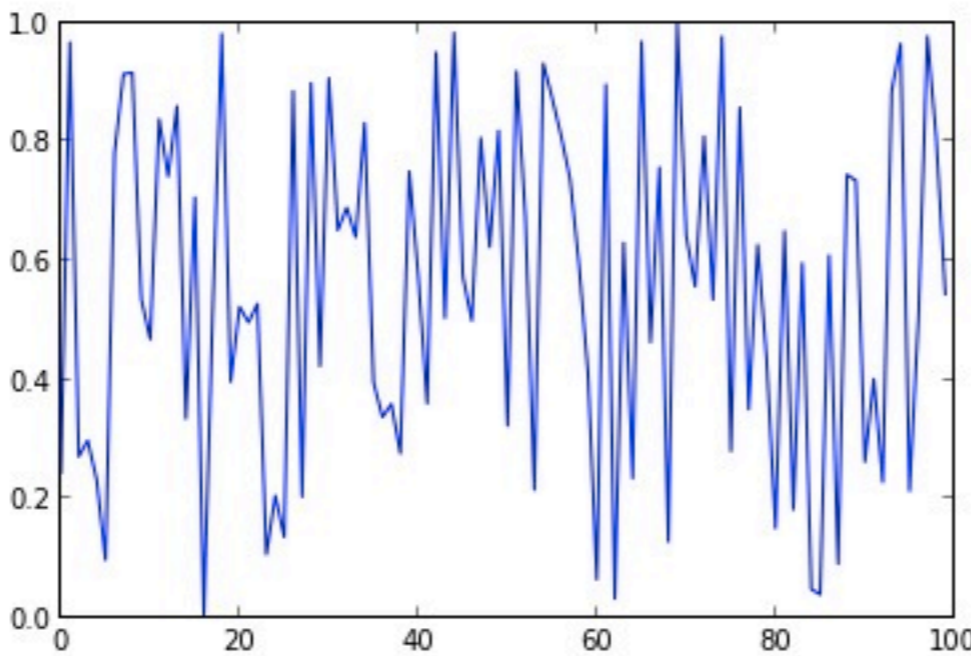
IP[y] 01_notebook_introduction Last saved: Aug 07 09:55

File Edit View Insert Cell Kernel Help

Markdown Cell Toolbar: None Enviro

```
In [6]: %pylab inline
        plot(rand(100))
```

Welcome to pylab, a matplotlib-based Python environment [backend: modu
For more information, type 'help(pylab)'.
Out[6]: [<matplotlib.lines.Line2D at 0x3a019d0>]

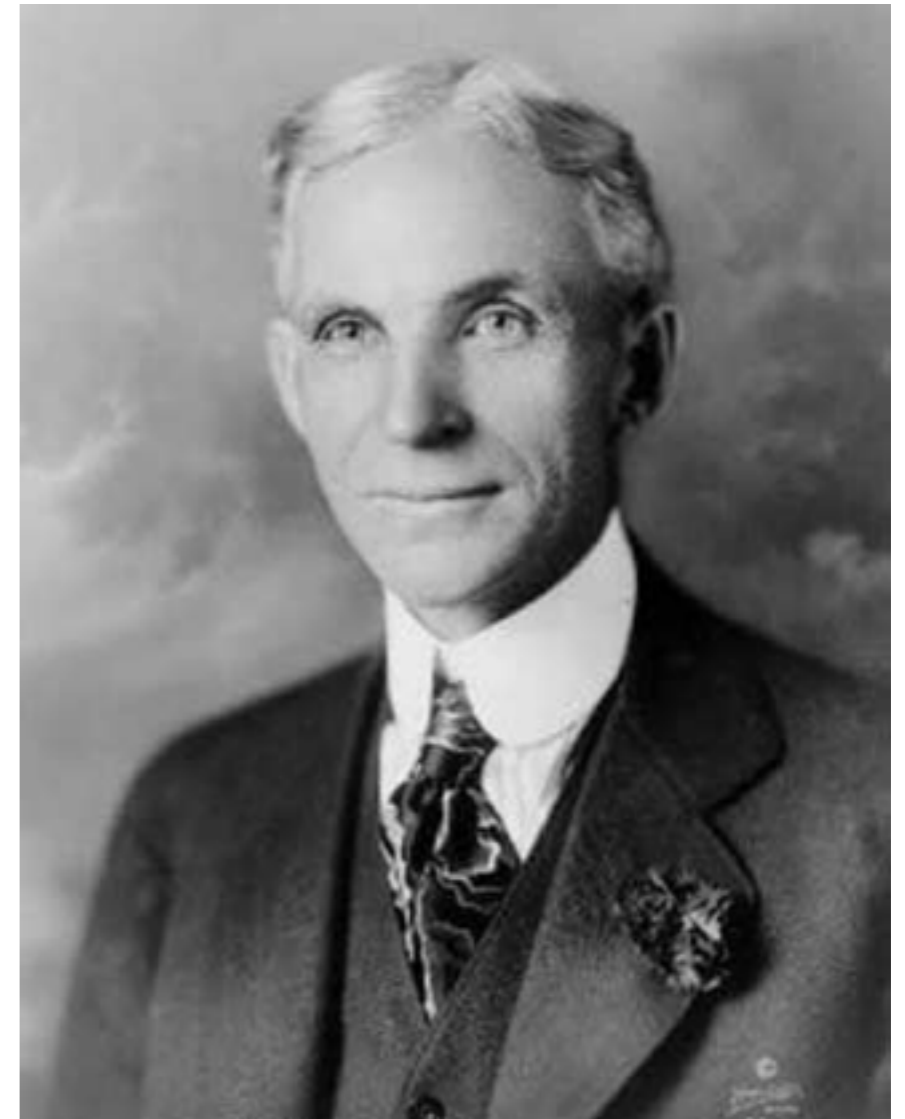


The plot displays a single blue line representing a random signal. The x-axis is labeled from 0 to 100 in increments of 20. The y-axis is labeled from 0.0 to 1.0 in increments of 0.2. The line fluctuates rapidly across the entire range of the x-axis, with values mostly between 0.2 and 1.0, and a few dips near 0.0.

Security

Another meta transition awaiting...

«How research & learning
can be automated?»



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](#)

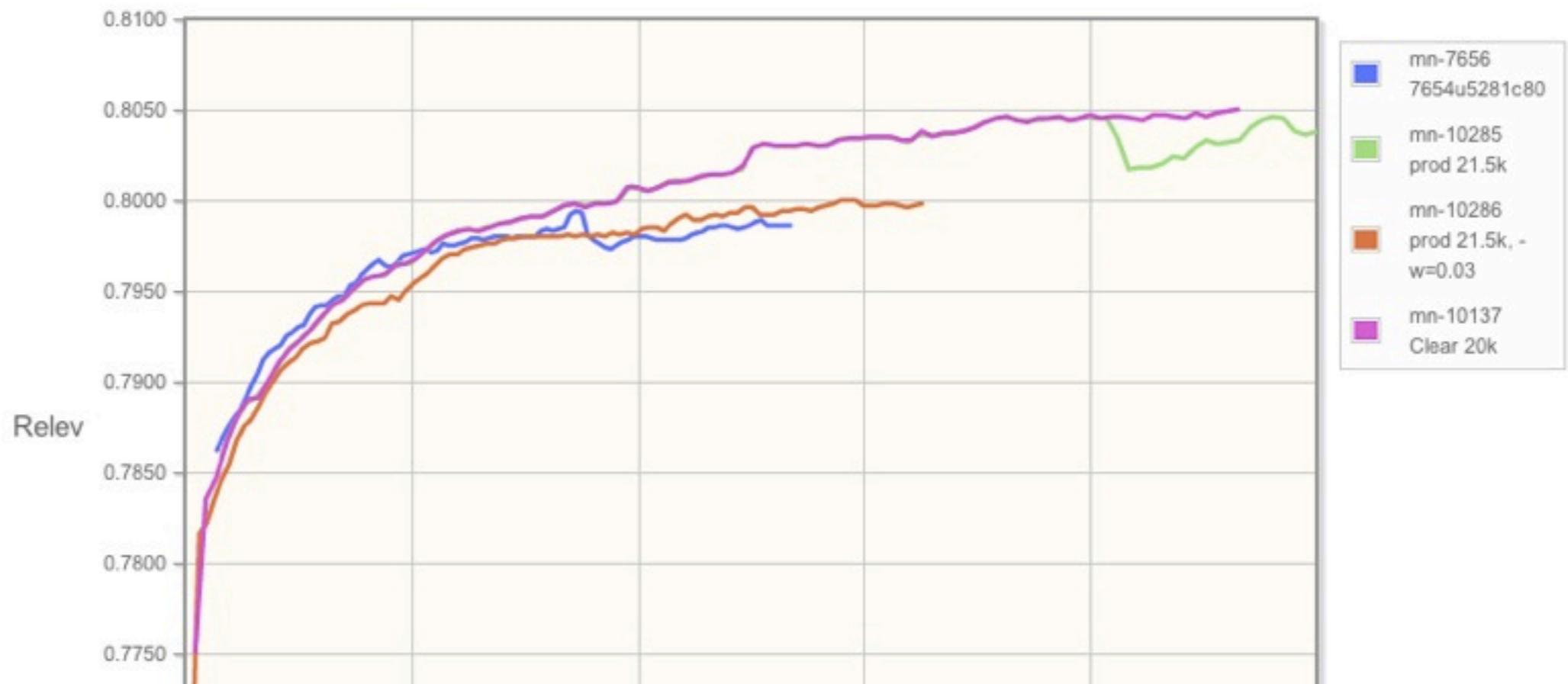
Tools

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