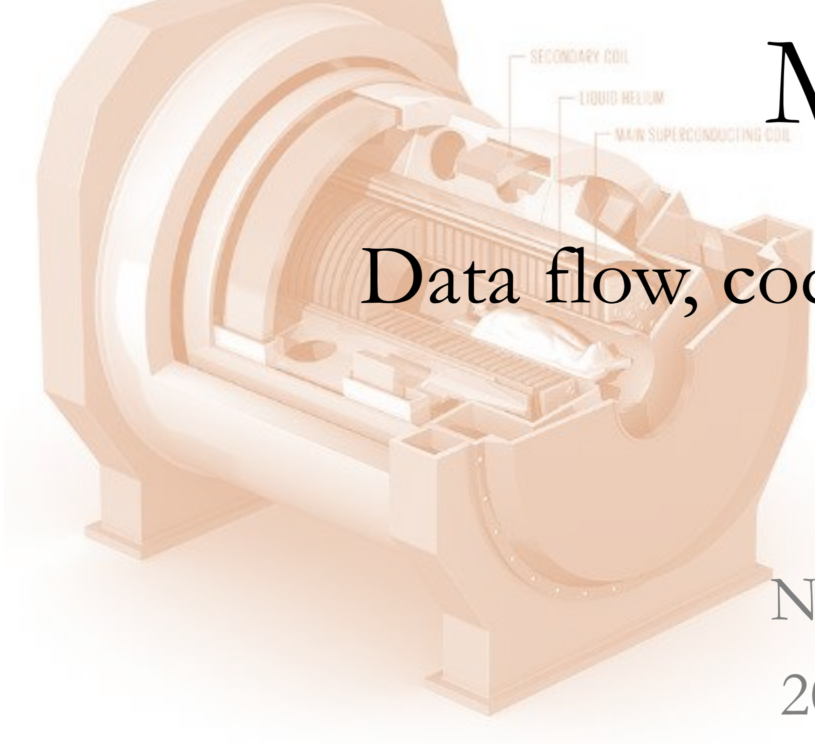
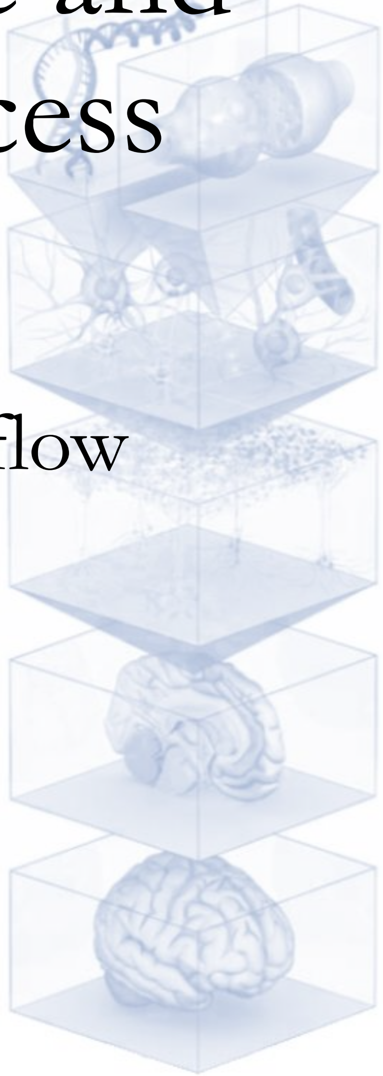
Computer Age Neuroscience and The Future of Science Process Management

Data flow, code flow, and collaborative flow





Daniel Labbé

Neuropsychology Retreat

2017-07-03 - 2017-07-05

daniel.labbe@rub.de

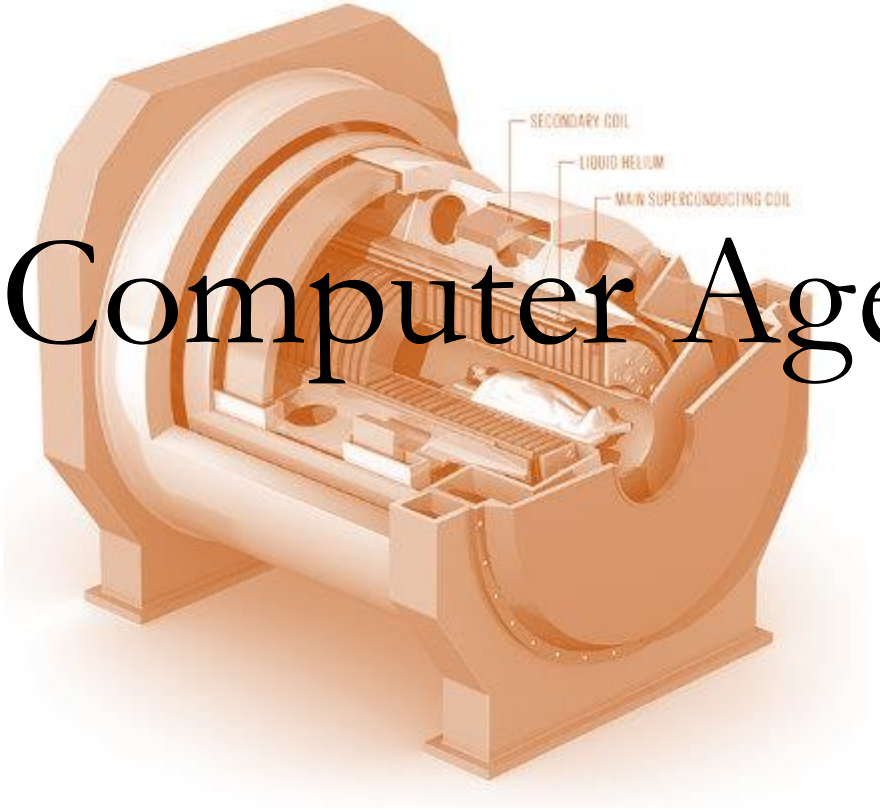
“Data sharing on a massive scale transformed the field of astronomy. Is neuroscience next?”

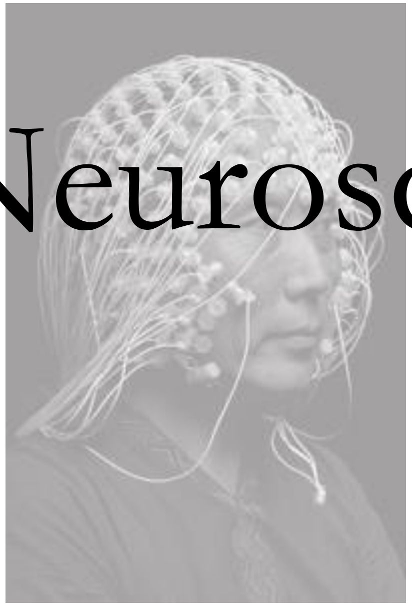
—Lindsay Borthwick, 2016, Neurodata Without borders

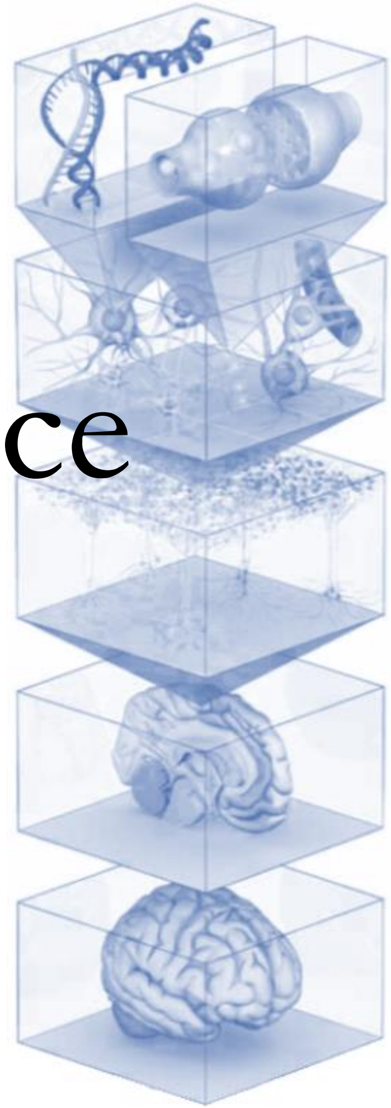
Outline

- Part I: Computer Age Neuroscience
- Part II: Science Process Management
- Discussion

Computer Age Neuroscience



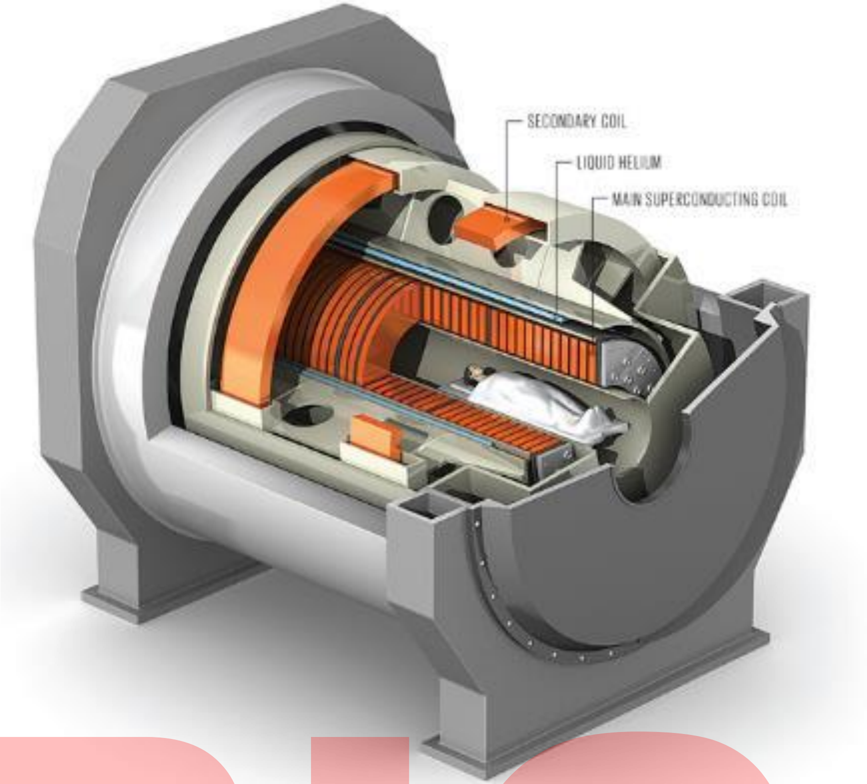




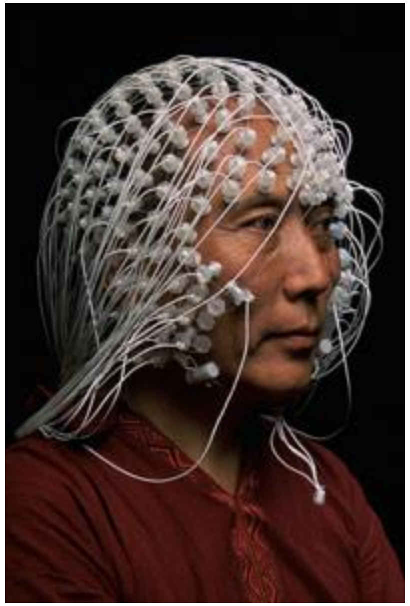
Computer Age Neuroscience

- How The End of Moore’s Law Prompts Parallel Computing
- The Inescapable Imminence of the Cloud
- Cloud Neuroscience
- The International Brain Station (TIBS)

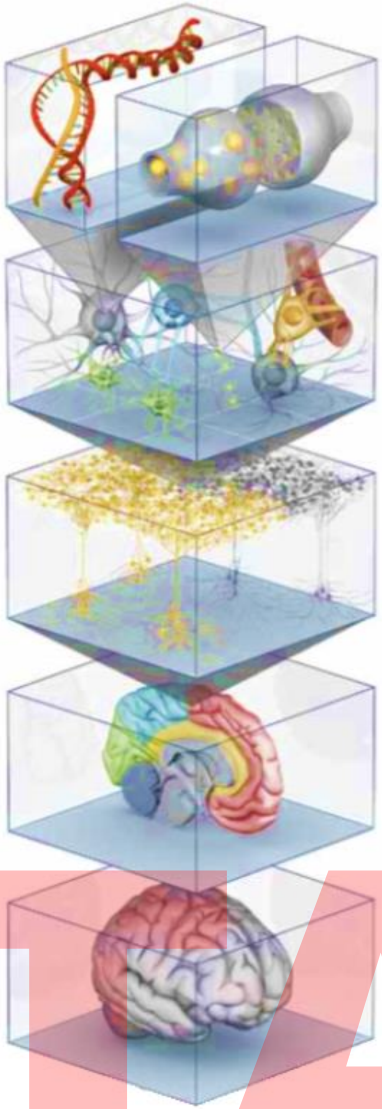
More Voxels, More Channels, and Multiple Levels of Simulation



11.7 Tesla



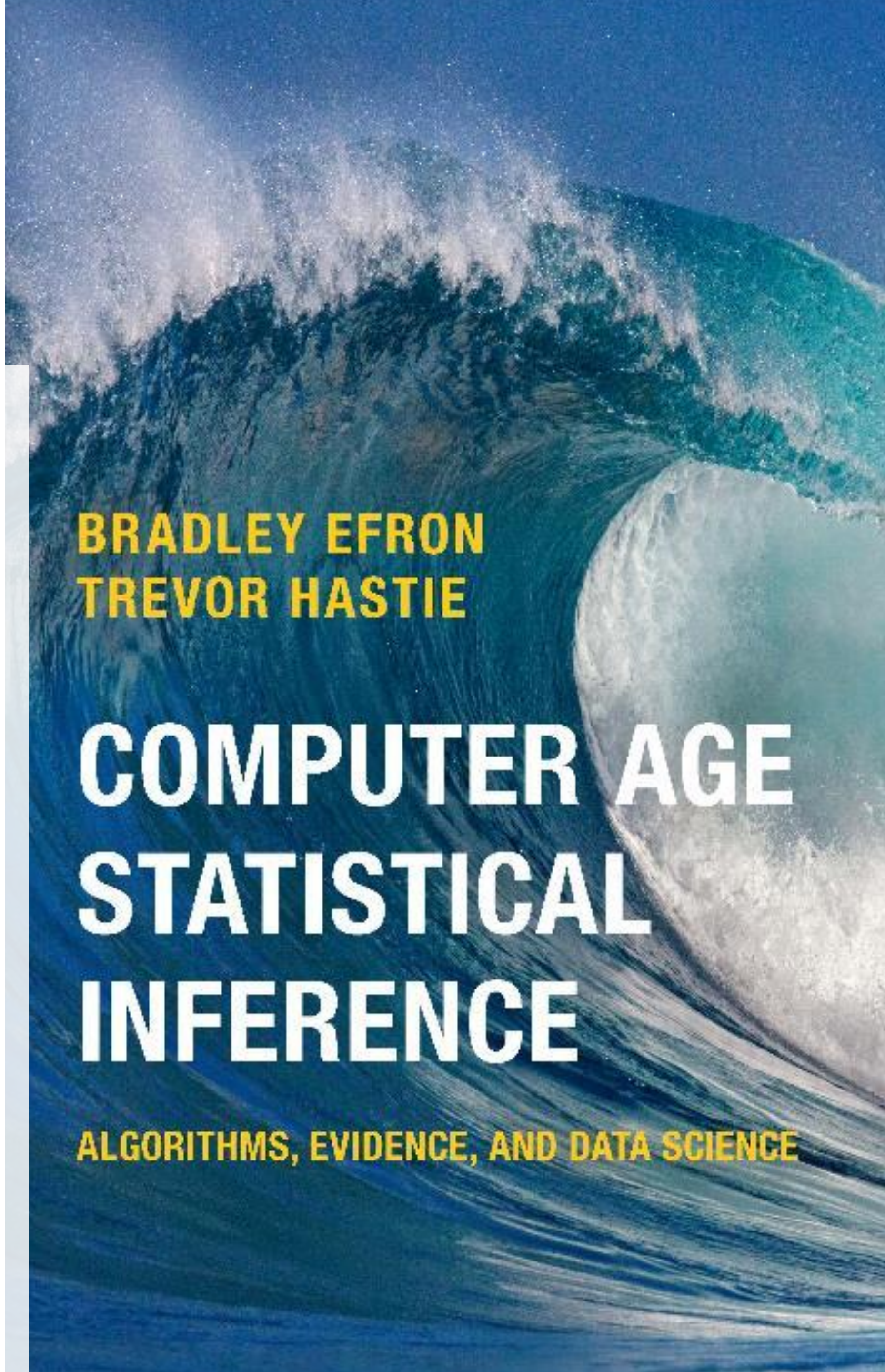
256 Electrodes



BIG DATA

Computer Age Neuroscience

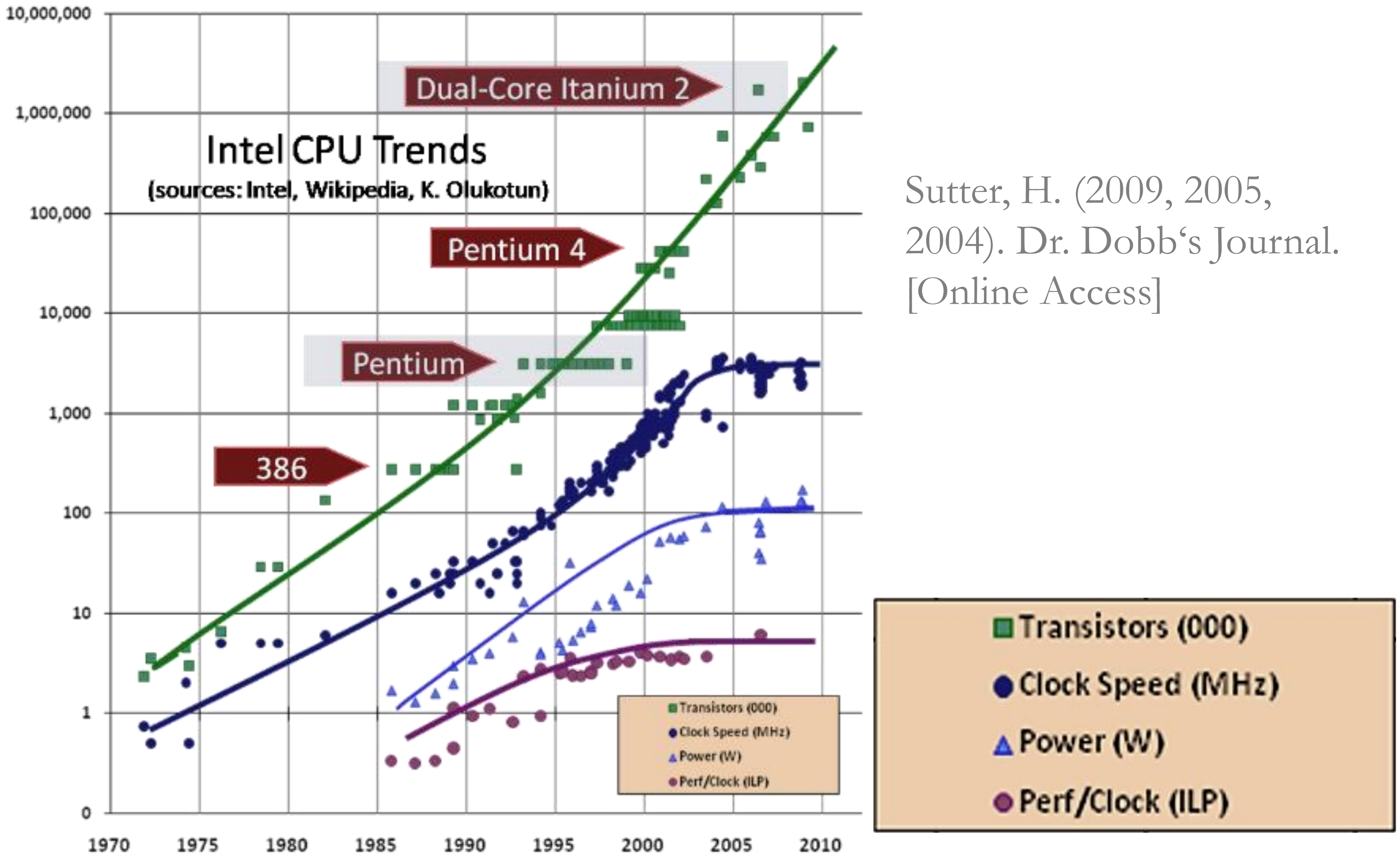
- ”...computer-based technology allows scientists to collect enormous data sets, orders of magnitude larger than those that classical statistical theory was designed to deal with; **huge data demands new methodology**, and the demand is being met by a burst of innovative computer-based statistical algorithms.” —Bradley Efron and Trevor Hastie, 2016, p.4 *Computer Age Statistical Inference.* NY: Cambridge University Press. (My emphasis.)
- ”**large-scale prediction algorithms**” such as ”neural nets, deep learning, boosting, random forests, and support-vector machines” **became prominent due to the commercial value of prediction** (Efron & Hastie, 2016, p.446, my emphasis)
- **Data science** became more about the **perfomance** of algorithms for big data (Efron & Hastie, 2016)



How The End of Moore's Law Prompts Parallel Computing

http://www.nature.com/news/the-chips-are-down-for-moore-s-law-1.19338

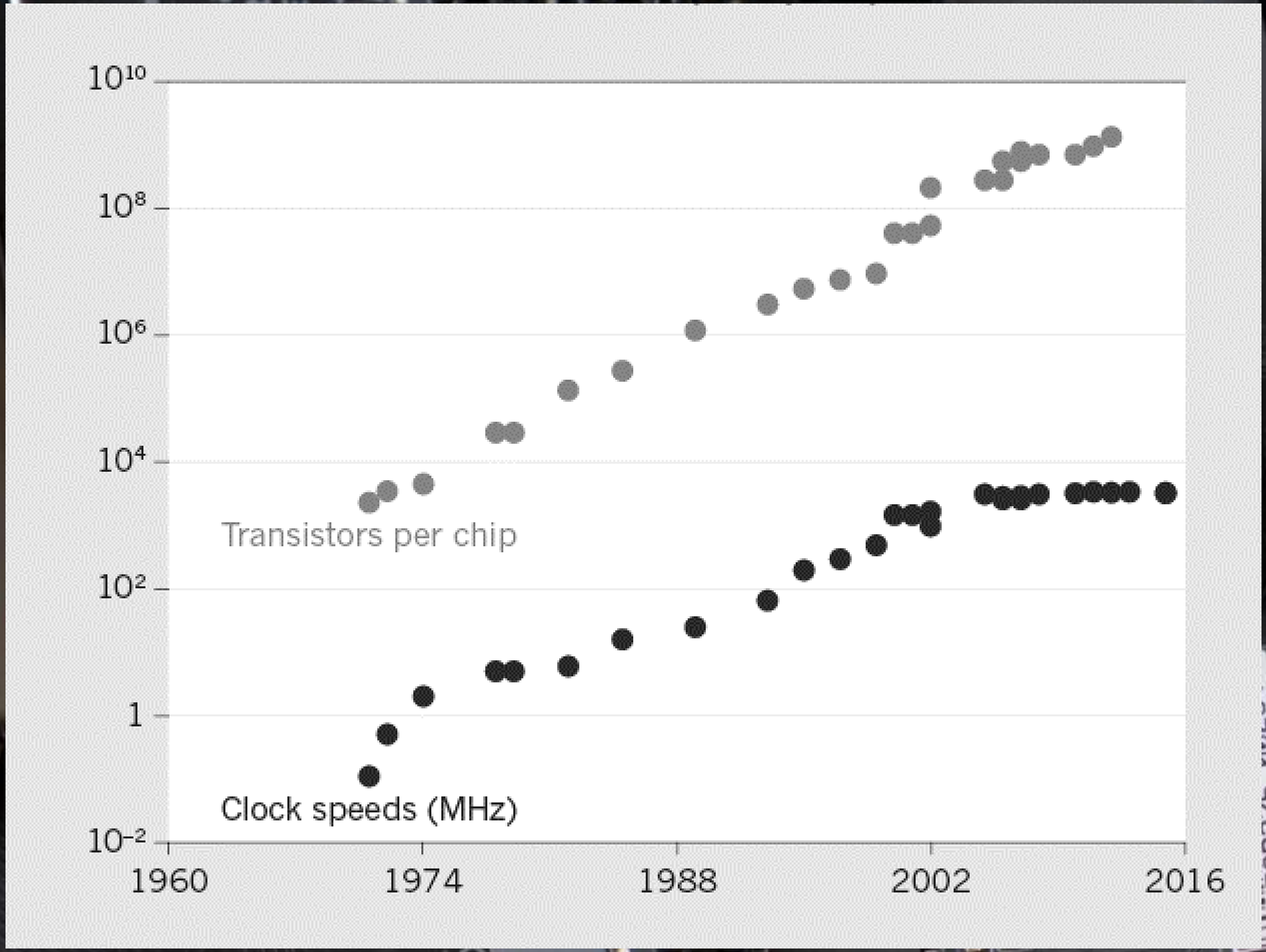
CPU Clock Speed Peaked in the Early 2000nds



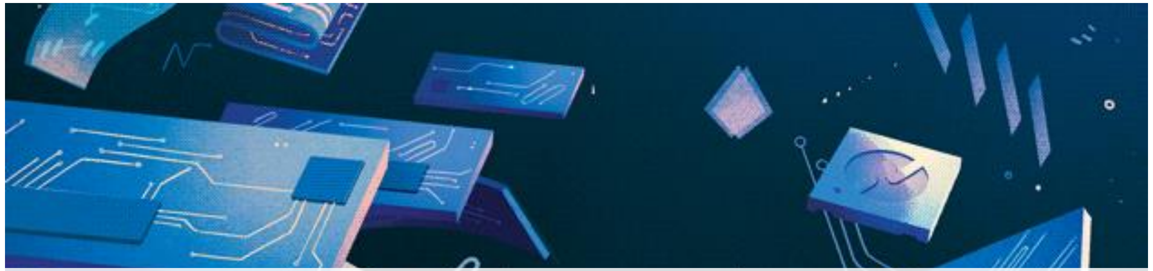
Sutter, H. (2009, 2005, 2004). Dr. Dobb‘s Journal. [Online Access]

Instruction-level parallelism (ILP)

Heat does not Dissipate Fast Enough in High Density Processing Arrays: Using multiple cores is a workaround



Waldrop, M. M. (2016).[News Feature] The chips are down for Moore's law, ***Nature***.



"Applications will increasingly need to be concurrent if they want to fully exploit continuing exponential CPU throughput gains"—Herb Sutter, 2009/2004

The Inescapable Imminence of the Cloud

"[for] an application that runs well on today's hardware and will just naturally run faster or better on tomorrow's hardware, you need to write an app with lots of juicy latent parallelism…

Costs

- Extra development effort
- Extra code complexity
- Extra testing effort

The Inescapable Imminence of the Cloud

The Good News

"(cont.) For many classes of applications the extra effort will be worthwhile, because concurrency will let them fully exploit the exponential gains in compute throughput that will continue to grow strong and fast long after Moore's Law has gone into its sunny retirement, as we continue to mine the cloud for the rest of our careers."—Sutter, H. (2012) (My emphasis.)

Cloud Neuroscience

"The kernel of the idea is based on a view of the scientific process as an 'upward spiral': a collective effort where each new experiment yields data, upon which analysis is performed, leading to new or refined models, which suggest novel experiments. The goal of The International Brain Station (TIBS) is to build technology that would democratize brain science, so that all brain scientists—professional and otherwise—can build on the shoulders of one another. To make that a reality, TIBS would enable "cloud neuroscience," meaning that the **data**, the **code** and the **analytic results** from neuroscience experiments all live in the cloud together, where it could be accessed by anyone."

—Joshua T. Vogelstein [Interview], *Neurodata Without borders*. My emphasis.
http://www.nwb.org/2016/11/11/unlocking-the-brain-with-open-data/

The International Brain Station (TIBS)

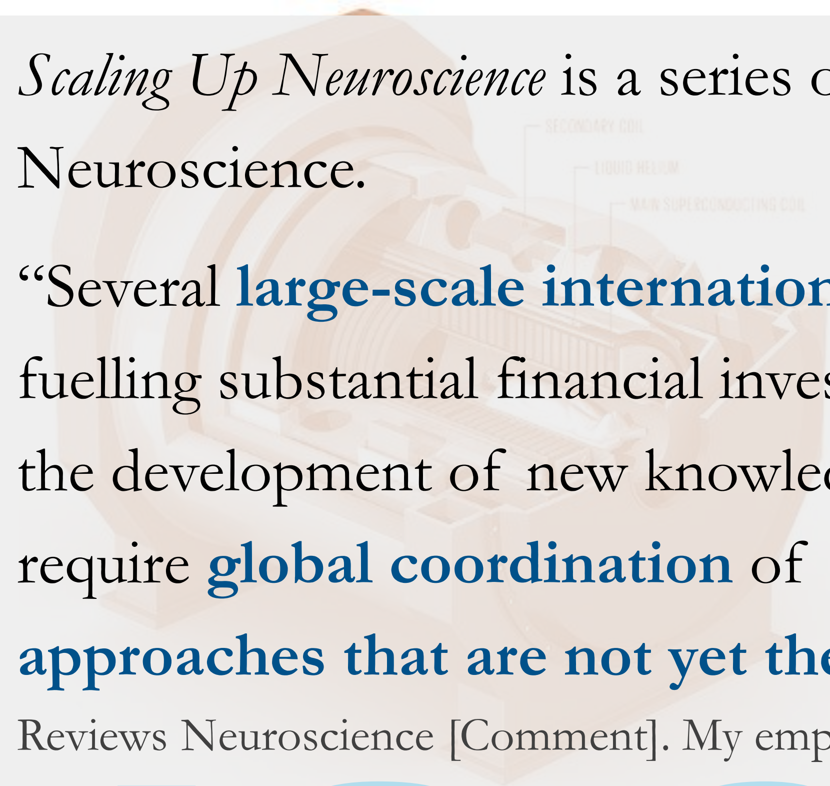
"[A] global neuroscience collaboration that would link [neuroscientists'] efforts and rival big science investments in astronomy and physics"—Underwood, E., Science, [In Depth][Summary] International Brain Projects Proposed. doi: 10.1126/science.352.6283.277

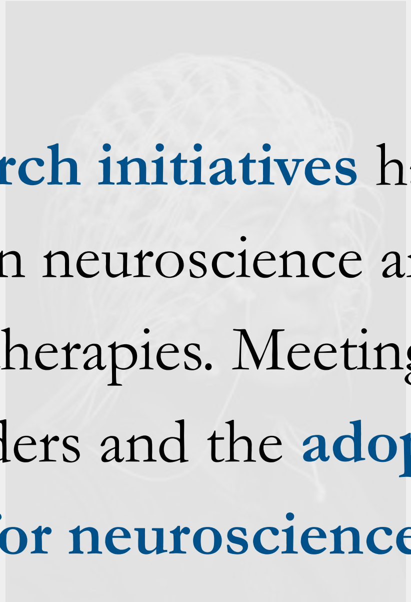
"This resource will realize a new era of brain sciences, one in which the bottlenecks to discovery transition away from data collection and processing to data enriching exploring, and modeling."—Global Brain Workshop 2016 Attendees, Grand Challenges for Global Brain Sciences. http://brainx.io/

Scaling Up Neuroscience

Scaling Up Neuroscience is a series of commentaries published in Nature Reviews Neuroscience.

"Several **large-scale international research initiatives** have recently been launched, fuelling substantial financial investments in neuroscience and raising expectations for the development of new knowledge and therapies. Meeting these expectations will require **global coordination** of stakeholders and the **adoption of team-based approaches that are not yet the norm for neuroscience.**"—Olds, J. L., [Abstract] Nature Reviews Neuroscience [Comment]. My emphasis.





BIG SCIENCE

Summary Part I: Computer Age Neuroscience

- Big Data emphasizes performance
- Vertical limit for processing units (CPUs, GPUs) reached
- Horizontal programming (parallelism, concurrency) required to future-proof software
- Cloud Neuroscience
  - Costs extra
  - Exponential gains
  - You cannot run from the cloud → Embrace the cloud
- Big Science requires a collaborative effort (team neuroscience)

How do we get there?

Science Process Management

Optimizing & Maintaining Data Flow, Code Flow, and Collaborative Flow

Three Challenges

- Data Flow: Parallelization & Scalability
- Code Flow: Sharing Code and Analysis in the Clouds
- Collaborative Workflow: Team Neuroscience

Part II: Science Process Management

- Definitions
- Software Examples
  - Pipelines—How to Achieve Data Flow Like a Molecular Geneticist
  - Version Control—Code Flow with Git
  - Lean Process Management—NASA Style Process Management using ZenHub: Integrating code flow with workflow
  - The Digital Notebook—Transparent Knowledge Transfer

Science Process Management

- **Scientific workflow**
  ”...a formal description of a process for accomplishing a scientific objective, usually expressed in terms of tasks and their dependencies” (Ludäscher et al. 2009, section 13.1)
- **Scientific workflow system** (Ludäscher et al., 2009)
  - Automatization
  - Documentation
  - Monitoring and recording (the process is requires regular self-regulation)
  - Optimization
  - Reusability
- Example: A paralell processing pipeline

The Human Process

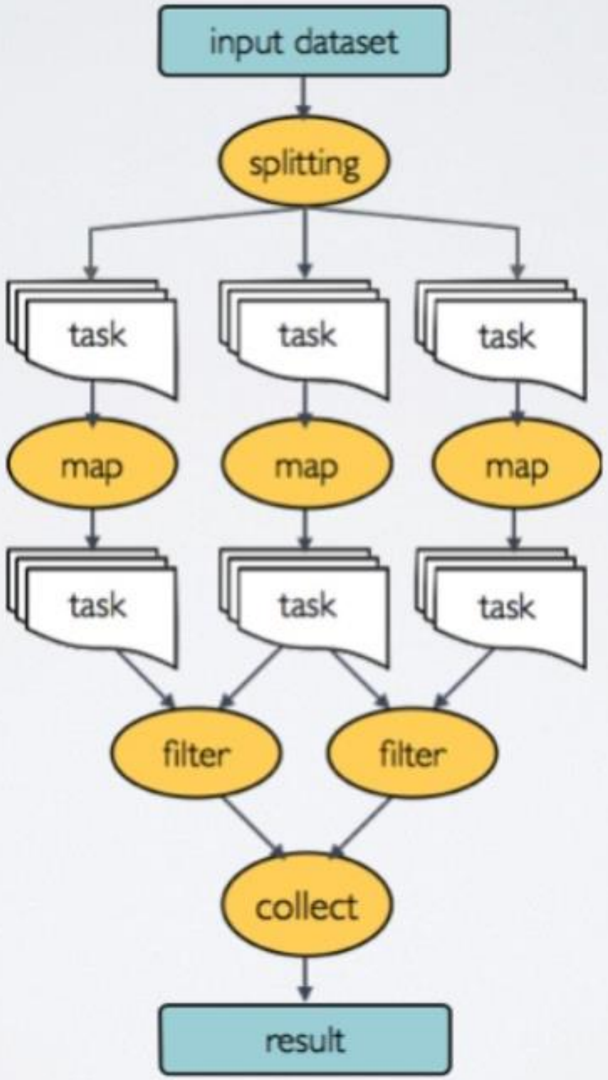
" 'human cycles' are a sometimes neglected resource, which can and should be optimized as well. For example, the use of data and workflow provenance information can be used for traditional purposes (such as optimizing system performance or improving fault-tolerance [citation]), but also to enhance the scientist's insights when trying to understand or debug scientific workflow results [citation]."

—Bertram Ludäscher et al., 2009, section 13.7, Scientific Process Automation and Workflow Management. In: Scientific Data Management. Chapman and Hall/CRC. doi: 10.1201/9781420069815

Pipelines

Paralellism and Concurrency using NextFlow

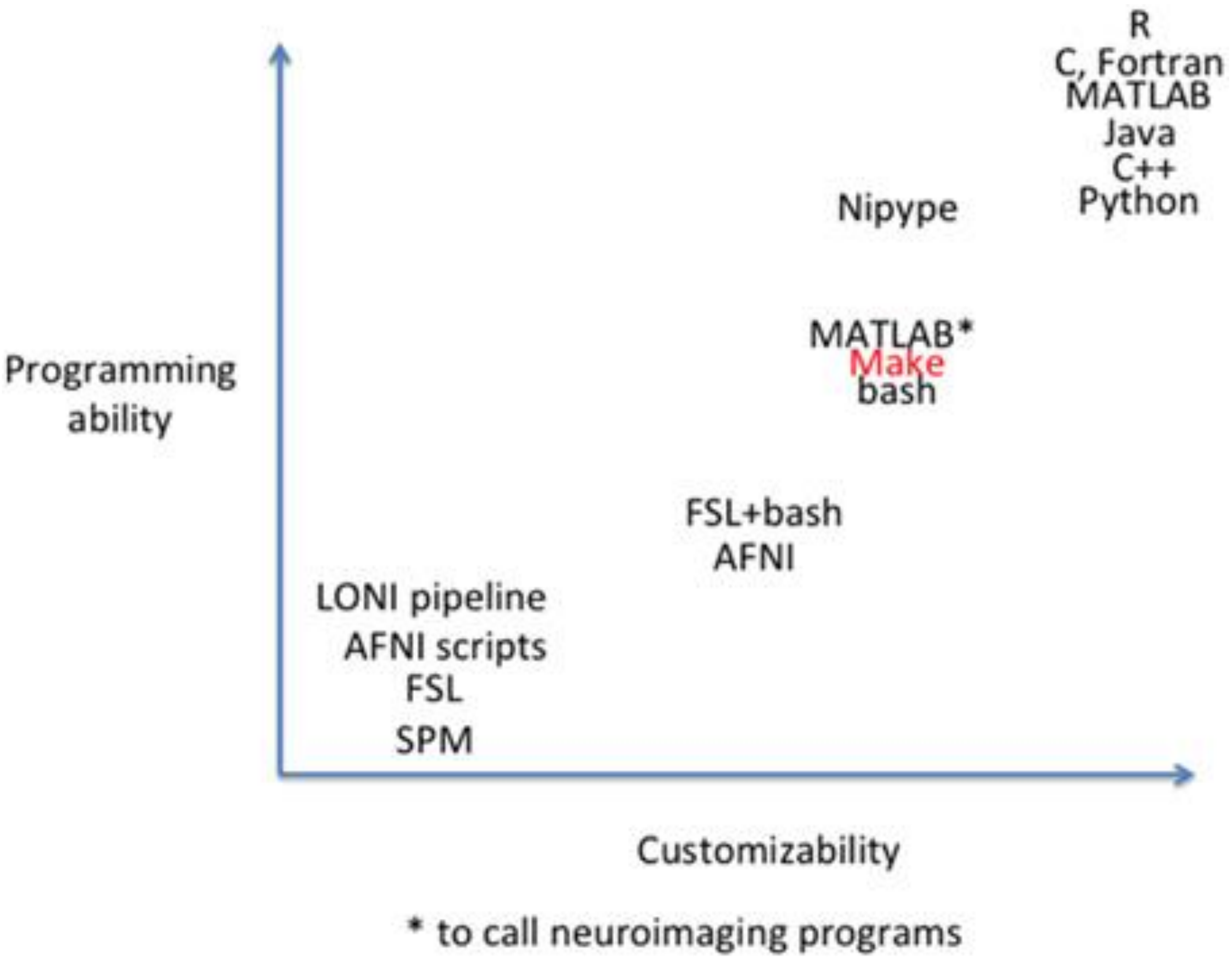
REACTIVE NETWORK



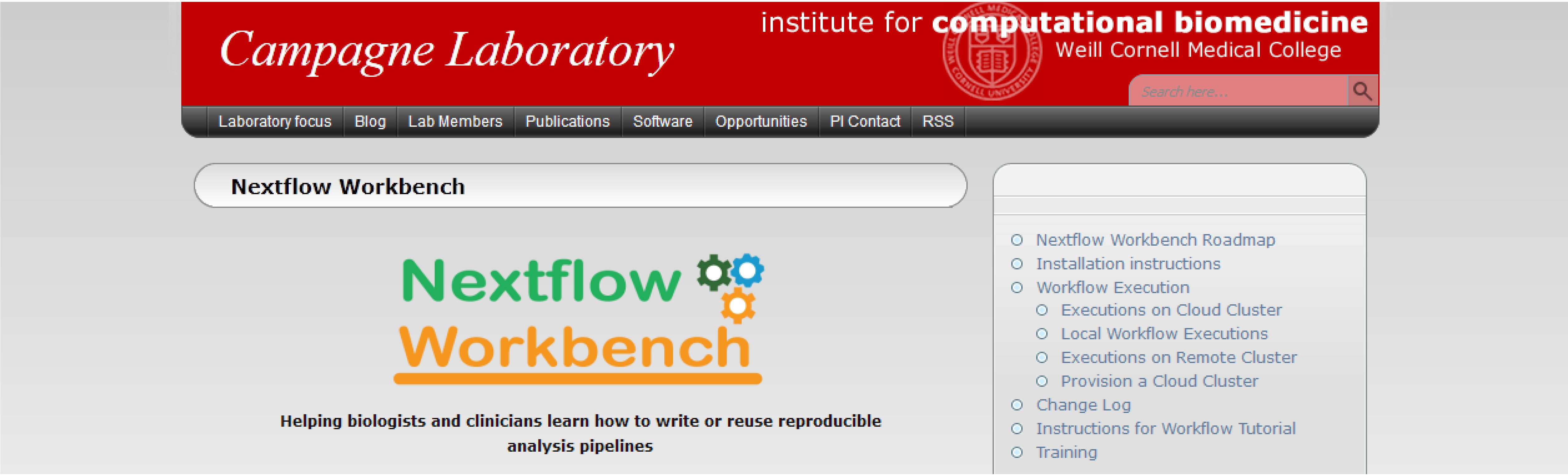
nextflow

Zero config	Polyglot	Concurrency	Scale easily

The next few slides are mainly from Paolo Di Tommaso, Center for Genomic Regulation, Barcelona
(Switzerland HPC Conference 2016.)



Askren et al. (2016). Using Make for Reproducible and Parallel Neuroimaging Workflow and Quality-Assurance. *Frontiers in Neuroinformatics.* https://doi.org/10.3389/fninf.2016.00002


Campagne Laboratory
institute for computational biomedicine
Weill Cornell Medical College
Search here...
Laboratory focus
Blog
Lab Members
Publications
Software
Opportunities
PI Contact
RSS
Nextflow Workbench
Nextflow
Workbench
Helping biologists and clinicians learn how to write or reuse reproducible analysis pipelines
Nextflow Workbench Roadmap
Installation instructions
Workflow Execution
Executions on Cloud Cluster
Local Workflow Executions
Executions on Remote Cluster
Provision a Cloud Cluster
Change Log
Instructions for Workflow Tutorial
Training

```
>MCHU - Calmodulin
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*
```

WHO IS USING NEXTFLOW?

NATIONAL MARROW DONOR PROGRAM®

Karolinska Institutet

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

International Agency Research on Cancer

World Health Organization

Weill Cornell Medical College

UT Southwestern Medical Center

UCSF

wellcome trust sanger institute

Universität Bielefeld

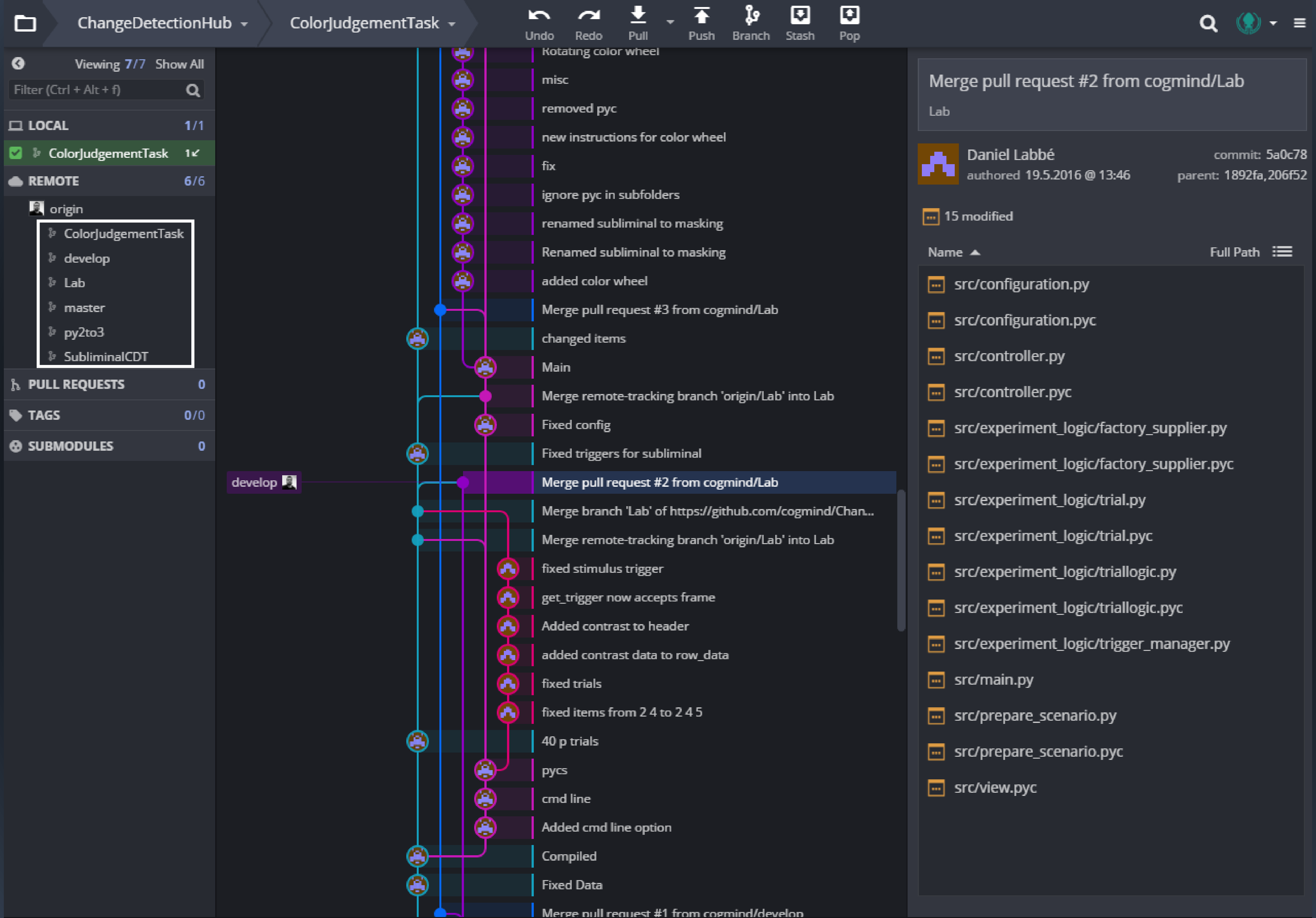
SciLifeLab

UiO : University of Oslo

CRG Centre for Genomic Regulation

Version Control

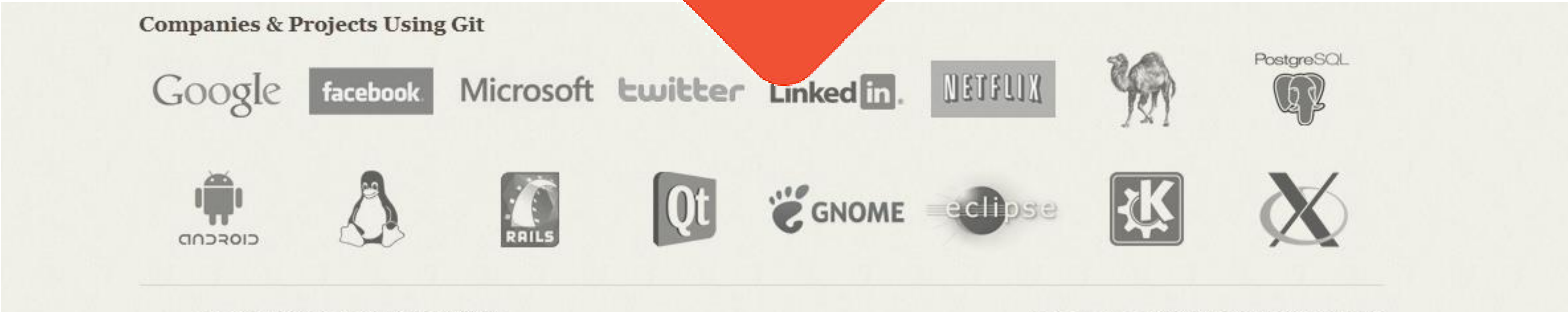
Using Git


ChangeDetectionHub
ColorJudgementTask
Undo
Redo
Pull
Push
Branch
Stash
Pop
Viewing 7/7 Show All
Filter (Ctrl + Alt + f)
LOCAL 1/1
ColorJudgementTask 1
REMOTE 6/6
origin
ColorJudgementTask
develop
Lab
master
py2to3
SubliminalCDT
PULL REQUESTS 0
TAGS 0/0
SUBMODULES 0
develop
Rotating color wheel
misc
removed pyc
new instructions for color wheel
fix
ignore pyc in subfolders
renamed subliminal to masking
Renamed subliminal to masking
added color wheel
Merge pull request #3 from cogmind/Lab
changed items
Main
Merge remote-tracking branch 'origin/Lab' into Lab
Fixed config
Fixed triggers for subliminal
Merge pull request #2 from cogmind/Lab
Merge branch 'Lab' of https://github.com/cogmind/Chan...
Merge remote-tracking branch 'origin/Lab' into Lab
fixed stimulus trigger
get_trigger now accepts frame
Added contrast to header
added contrast data to row_data
fixed trials
fixed items from 2 4 to 2 4 5
40 p trials
pycs
cmd line
Added cmd line option
Compiled
Fixed Data
Merge pull request #1 from cogmind/develop
Merge pull request #2 from cogmind/Lab
Lab
Daniel Labbé
authored 19.5.2016 @ 13:46
commit: 5a0c78
parent: 1892fa,206f52
15 modified
Name
Full Path
src/configuration.py
src/configuration.pyc
src/controller.py
src/controller.pyc
src/experiment_logic/factory_supplier.py
src/experiment_logic/factory_supplier.pyc
src/experiment_logic/trial.py
src/experiment_logic/trial.pyc
src/experiment_logic/triallogic.py
src/experiment_logic/triallogic.pyc
src/experiment_logic/trigger_manager.py
src/main.py
src/prepare_scenario.py
src/prepare_scenario.pyc
src/view.pyc

Version Control

Git

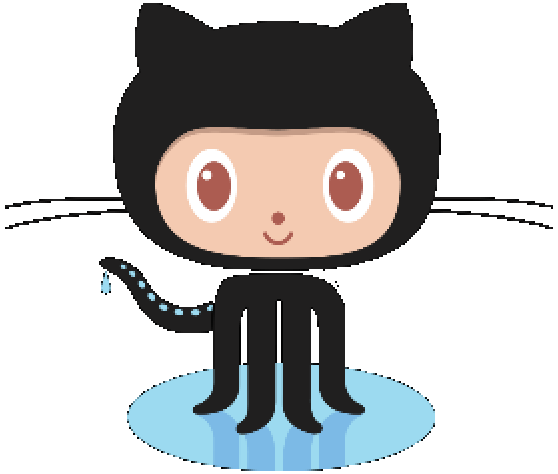
created by Linus Torvalds



Git

- Enables
  - Versioning
  - Branching
  - Distributed code repository
  - Secure connection to server
  - Intermediate documentation for changes (commit messages)
- Accessible through Git client

GitHub



- GitHub is an online code repository service that uses Git
- The biggest of several providers (other examples: GitLab, Bitbucket)
- Open source projects are free
- Private repositories for a monthly fee
- GitHub can also be connected to the Open Science Framework (OSF), an open data platform for publishing data, code, and for managing collaborations

Open Science Framework

A scholarly commons to connect the entire research cycle

Lean Process Management

Using ZenHub

ZenHub

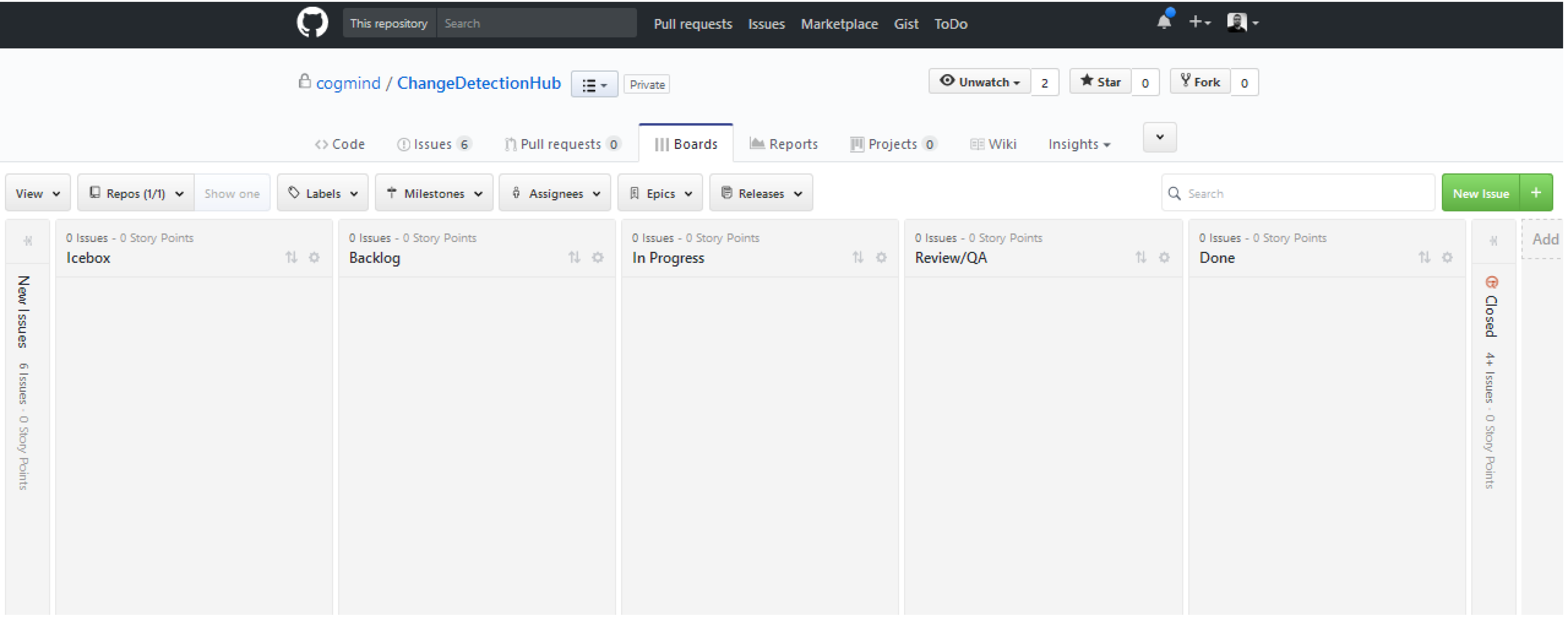
- A browser/API add-on for GitHub
- Suitable for science process management
- Scalable from project level to organization

Jet Propulsion Laboratory
California Institute of Technology

ZenHub

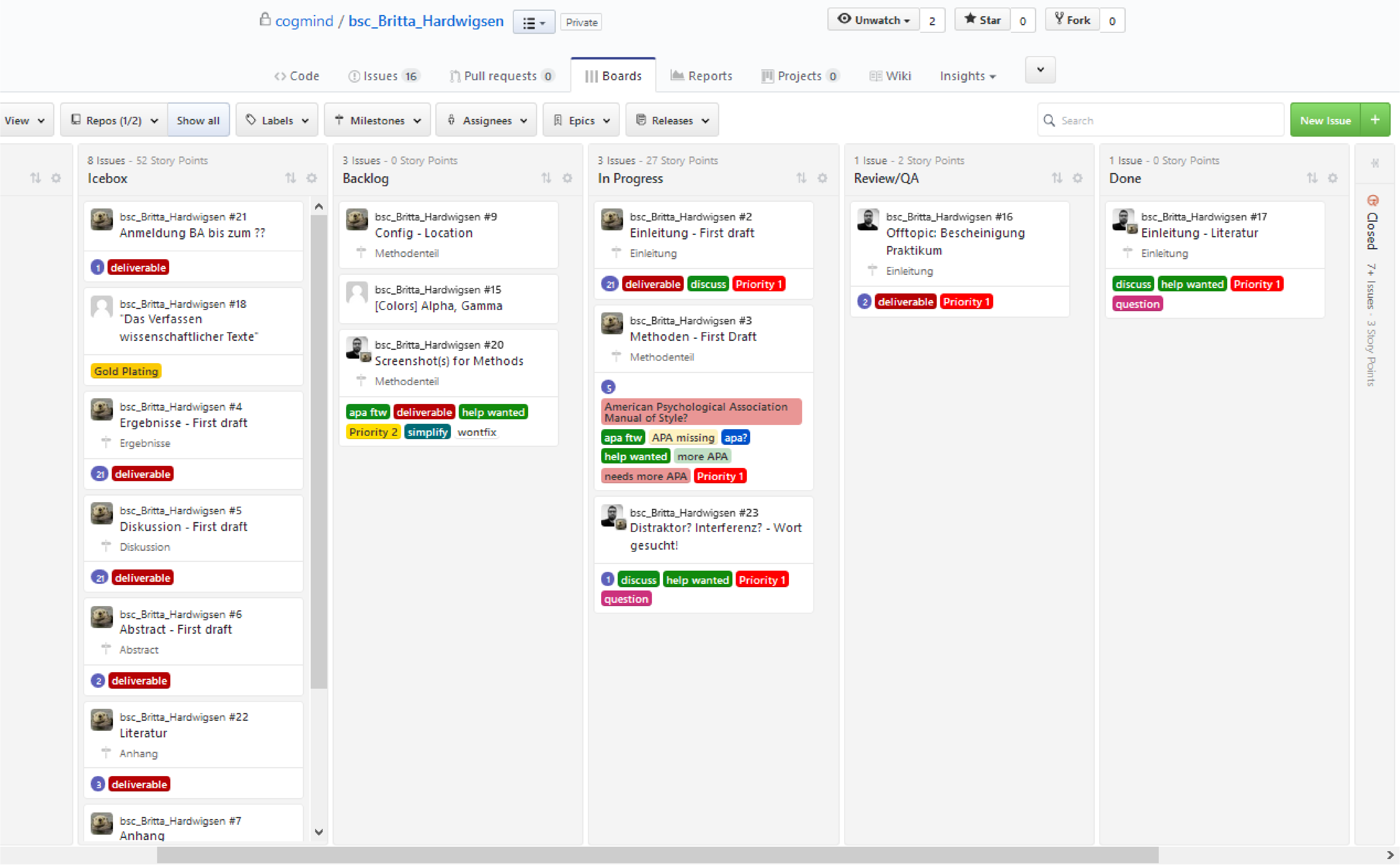
- Lean process management
  - Register tasks, assign people
  - Implements an advanced, and customizable, Kanban system
  - Set **milestones**, estimate relative effort/**story points**, and review progress in **burn-down charts**
- Shared planning
  - Collaborations
  - Individual projects (you can merge planning from two different repositories)
  - Team Neuroscience
- Horizontal knowledge transfer

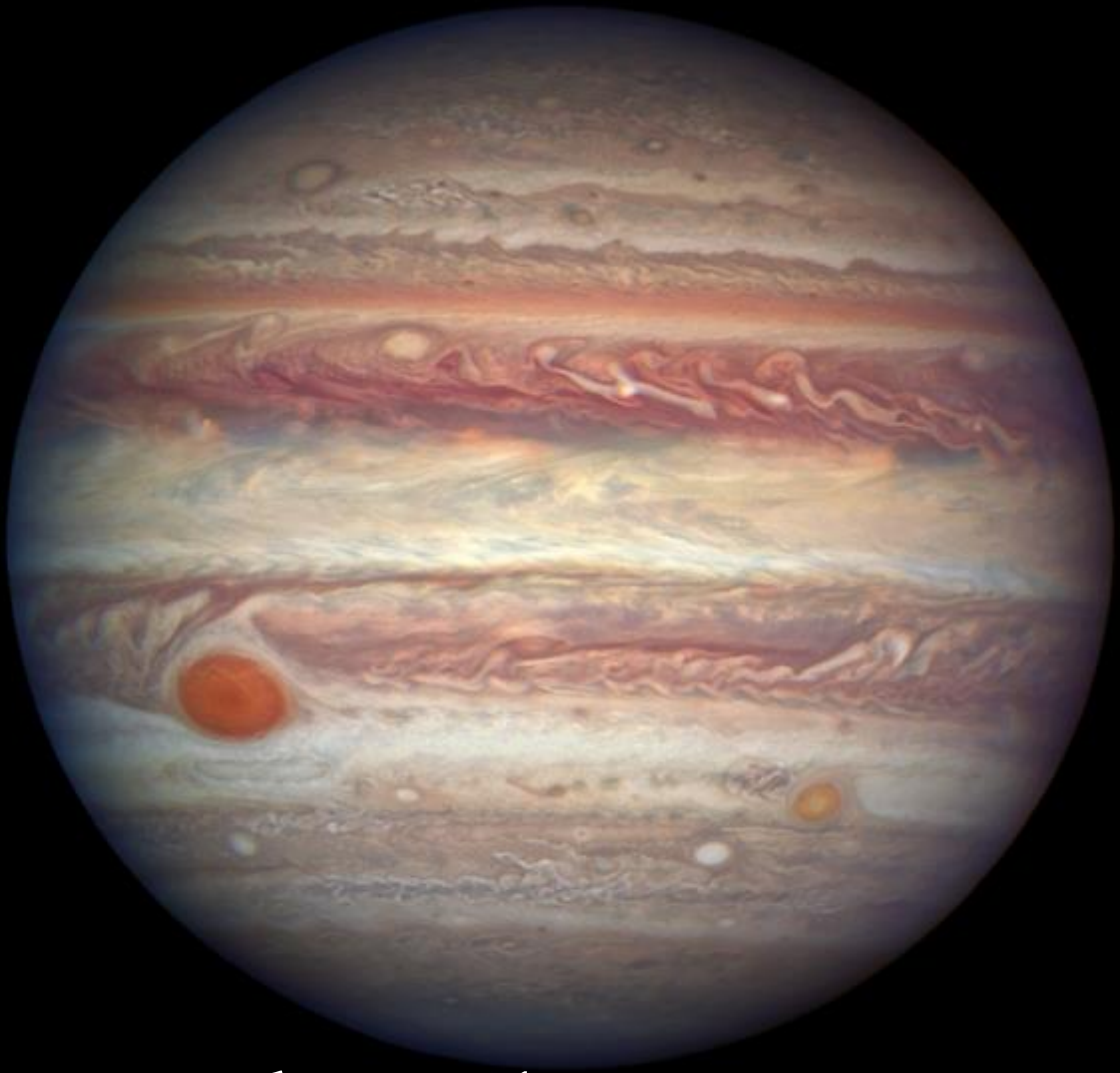
Kanban System



ZenHub Example

- Case Study: Supervision of a BSc student
- Two repositories
  - Writing up
  - Analysis


cogmind / bsc_Britta_Hardwigsen
Private
Unwatch 2
Star 0
Fork 0
Code
Issues 16
Pull requests 0
Boards
Reports
Projects 0
Wiki
Insights
View
Repos (1/2)
Show all
Labels
Milestones
Assignees
Epics
Releases
Search
New Issue
8 Issues - 52 Story Points
Icebox
bsc_Britta_Hardwigsen #21
Anmeldung BA bis zum ??
1 deliverable
bsc_Britta_Hardwigsen #18
"Das Verfassen wissenschaftlicher Texte"
Gold Plating
bsc_Britta_Hardwigsen #4
Ergebnisse - First draft
Ergebnisse
21 deliverable
bsc_Britta_Hardwigsen #5
Diskussion - First draft
Diskussion
21 deliverable
bsc_Britta_Hardwigsen #6
Abstract - First draft
Abstract
2 deliverable
bsc_Britta_Hardwigsen #22
Literatur
Anhang
3 deliverable
bsc_Britta_Hardwigsen #7
Anhang
3 Issues - 0 Story Points
Backlog
bsc_Britta_Hardwigsen #9
Config - Location
Methodenteil
bsc_Britta_Hardwigsen #15
[Colors] Alpha, Gamma
bsc_Britta_Hardwigsen #20
Screenshot(s) for Methods
Methodenteil
apa ftw
deliverable
help wanted
Priority 2
simplify
wontfix
3 Issues - 27 Story Points
In Progress
bsc_Britta_Hardwigsen #2
Einleitung - First draft
Einleitung
21 deliverable discuss Priority 1
bsc_Britta_Hardwigsen #3
Methoden - First Draft
Methodenteil
5
American Psychological Association Manual of Style?
apa ftw
APA missing
apa?
help wanted
more APA
needs more APA
Priority 1
bsc_Britta_Hardwigsen #23
Distraktor? Interferenz? - Wort gesucht!
1 discuss help wanted Priority 1
question
1 Issue - 2 Story Points
Review/QA
bsc_Britta_Hardwigsen #16
Offtopic: Bescheinigung Praktikum
Einleitung
2 deliverable Priority 1
1 Issue - 0 Story Points
Done
bsc_Britta_Hardwigsen #17
Einleitung - Literatur
Einleitung
discuss help wanted Priority 1
question
Closed
7+ Issues - 3 Story Points



Digital Notebooks

Enabling Knowledge Transfer

jupyter

Currently in use at

Google · Microsoft · IBM · Bloomberg · O'REILLY

CONTINUUM ANALYTICS · rackspace · SOUNDCLOUD · Quantopian · NetApp

software carpentry · hhmi janelia Research Campus · CODE NEURO · N-Site LLC · COCALC

BRYN MAWR COLLEGE · CAL POLY SAN LUIS OBISPO · Berkeley UNIVERSITY OF CALIFORNIA · The University Of Sheffield. · THE GEORGE WASHINGTON UNIVERSITY WASHINGTON, DC

CLEMSON UNIVERSITY · MICHIGAN STATE UNIVERSITY · Northwestern University · NYU · NASA

IP[y]: IPython

Interactive Computing

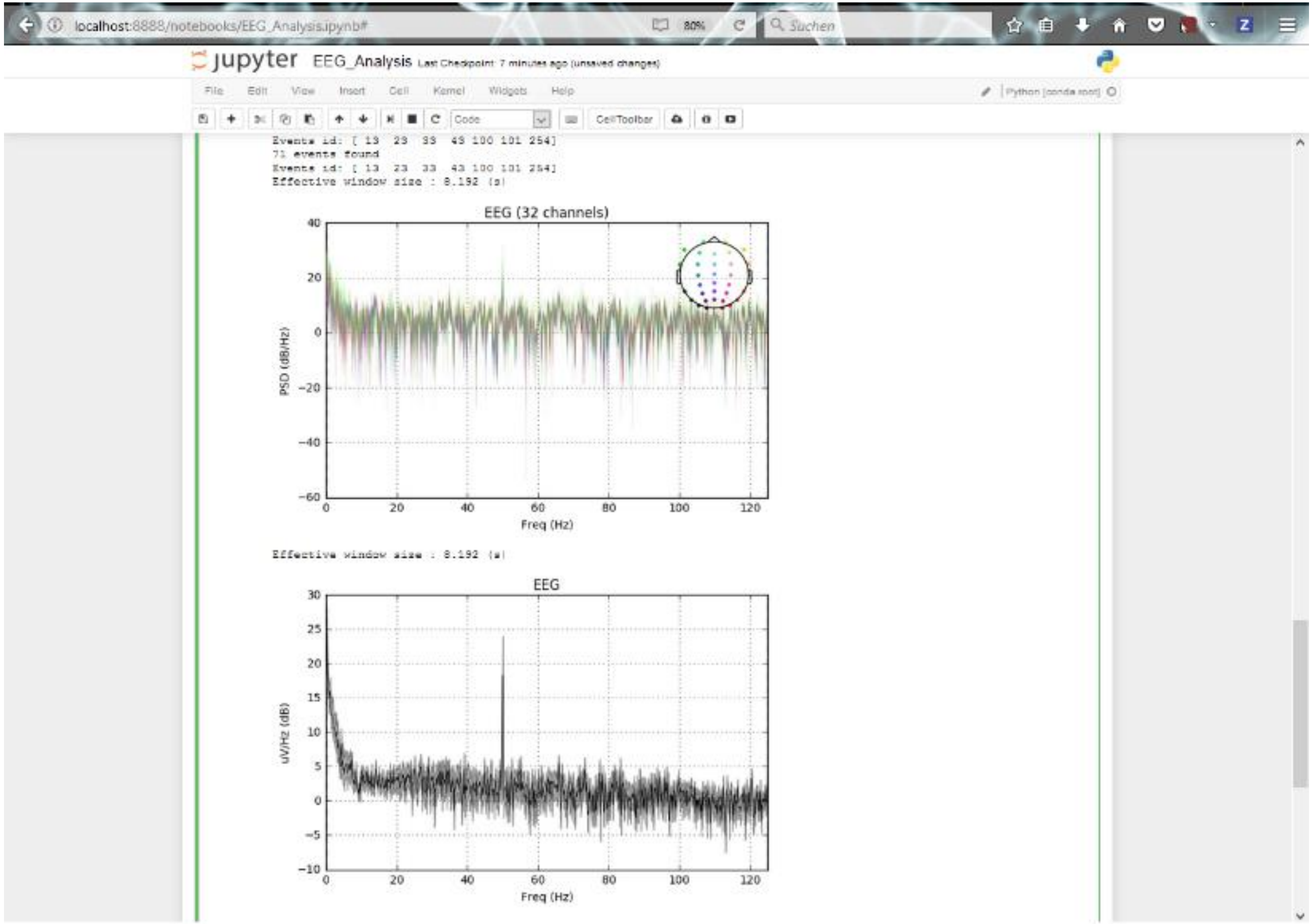
```
D:\analysis\alpha\eeg>cd mne

D:\analysis\alpha\eeg\mne>dir
 Volume in Laufwerk D: hat keine Bezeichnung.
 Volumeseriennummer: 860C-4E82

 Verzeichnis von D:\analysis\alpha\eeg\mne

2017-06-26  16:27    <DIR>          .
2017-06-26  16:27    <DIR>          ..
2017-06-26  16:59               922 mne_start.py
               1 Datei(en),            922 Bytes
               2 Verzeichnis(se), 471,515,373,568 Bytes frei

D:\analysis\alpha\eeg\mne>jupyter notebook
[I 14:17:10.352 NotebookApp] [nb_conda_kernels] enabled, 7 kernels found
[W 14:17:14.900 NotebookApp] Error loading server extension nbpresent
    Traceback (most recent call last):
      File "c:\Anaconda2\lib\site-packages\notebook\notebookapp.py", line 1046, in init_server_extensions
        mod = importlib.import_module(modulename)
      File "c:\Anaconda2\lib\importlib\__init__.py", line 37, in import_module
        __import__(name)
      File "c:\Anaconda2\lib\site-packages\nbpresent\__init__.py", line 5, in <module>
        from nbconvert.exporters.export import exporter_map
      File "c:\Anaconda2\lib\site-packages\nbconvert\__init__.py", line 4, in <module>
        from .exporters import *
      File "c:\Anaconda2\lib\site-packages\nbconvert\exporters\__init__.py", line 1, in <module>
        from .export import *
      File "c:\Anaconda2\lib\site-packages\nbconvert\exporters\export.py", line 9, in <module>
        import entrypoints
      File "c:\Anaconda2\lib\site-packages\entrypoints.py", line 16, in <module>
        from backports import configparser
    ImportError: cannot import name configparser
[I 14:17:14.961 NotebookApp] [nb_conda] enabled
[I 14:17:16.605 NotebookApp] [nb_anacondacloud] enabled
[I 14:17:17.344 NotebookApp] Serving notebooks from local directory: D:\analysis\alpha\eeg\mne
[I 14:17:17.345 NotebookApp] 0 active kernels
[I 14:17:17.345 NotebookApp] The Jupyter Notebook is running at: http://localhost:8888/
```


localhost:8888/notebooks/EEG_Analysis.ipynb#
Suchen
jupyter EEG_Analysis Last Checkpoint: 7 minutes ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Python [conda root]
Code
CellToolbar
Events id: [13 23 33 43 100 101 254]
71 events found
Events id: [13 23 33 43 100 101 254]
Effective window size : 8.192 (s)
EEG (32 channels)
PSD (dB/Hz)
Freq (Hz)
Effective window size : 8.192 (s)
EEG
uV/Hz (dB)
Freq (Hz)

Branch: master | vector / eeg_analysis.ipynb | Find file | Copy path

cogmind Add files via upload | 1e7aead 21 days ago

1 contributor

389 lines (388 sloc) | 11 KB | Raw | Blame | History

In [1]:
```
# Binomatic
#
# Description= Generates ERPLAB binlister files
#
# Author= Daniel Labbé
# Date= 2017-06-09

# Based on https://github.com/lucklab/erplab/wiki/Assigning-Events-to-Bins-with-BINLISTER
#
```

In [2]:
```
from itertools import count, product
```

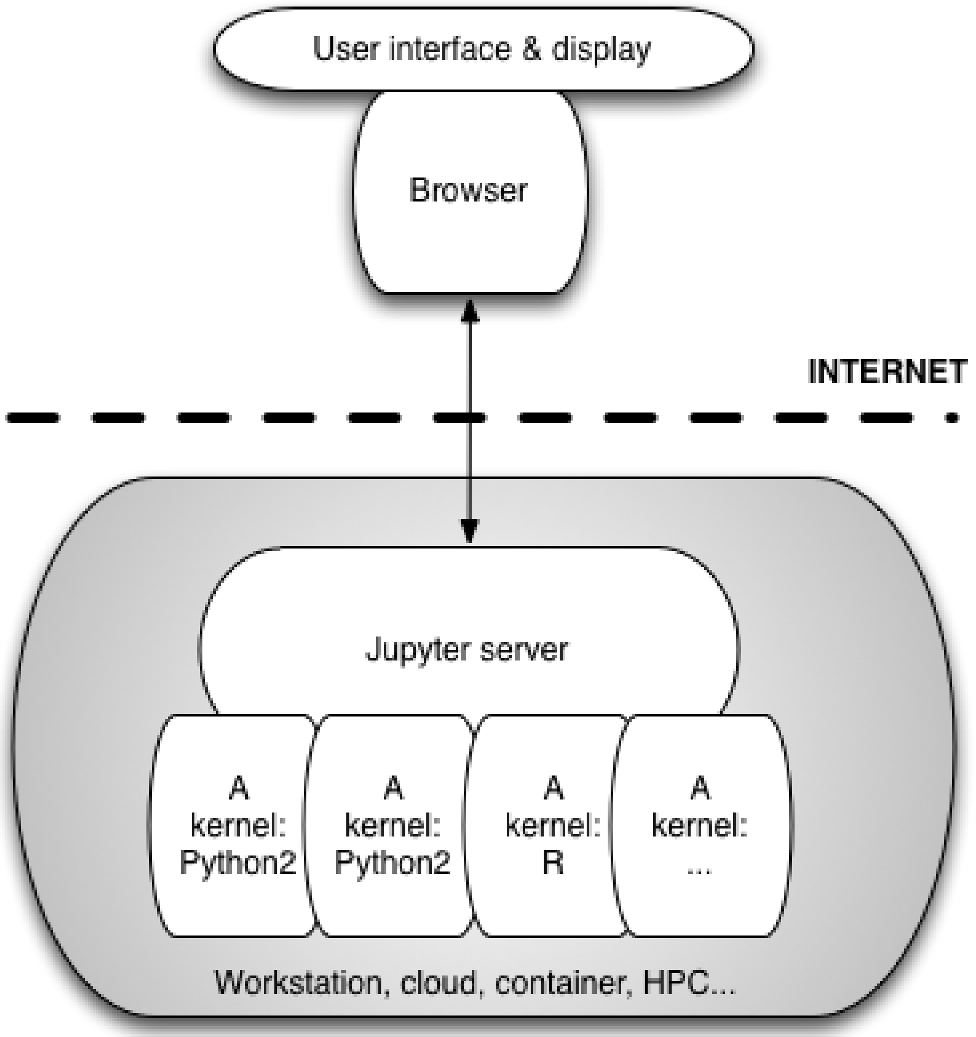
In [22]:
```
# GRAMMAR
_TAGS = '<' + '{}' + '>'

# Please note: "101;102" should be entered as a string and not a tuple

g = dict(
    EVENT = '{{{0}}}', # {{ acts as escape for {
    PRE = '.' + '{}',
    POST = '{}' +  '.',
    NOT = '~' + '{}',
    OR = '{}' + ';' + '{}',
    FROM_TO = '{}' + ':' + '{}',
    RANGE = '{}-{}',
    TAGS = _TAGS,
    TIME = 't' + _TAGS + '{}', # 't<' + '{}' + '>' + '{}'
    RT = '{}' + ':rt' + _TAGS,
    WA = '{}' + ':wa' + _TAGS,
    FA = '{}' + ':fa' + _TAGS,
    WB = '{}' + ':wb' + _TAGS,
    FB = '{}' + ':fb' + _TAGS
)
```

In [28]:
```
def parse(stimulus, syntax):
    #print(type(stimulus))
    assert isinstance(stimulus, str) or isinstance(stimulus, int) or isinstance(stimulus, tupl
e)
    assert isinstance(syntax, str)

    if isinstance(stimulus, str):
        p = g[syntax].format(stimulus)
    elif isinstance(stimulus, tuple):
```


User interface & display
Browser
INTERNET
Jupyter server
A kernel: Python2
A kernel: Python2
A kernel: R
A kernel: ...
Workstation, cloud, container, HPC...

Discussion

Technical Debt

DEFINITION

The extra work required to bring a rushed piece of code up to standard with current practises or your own understanding of the code.

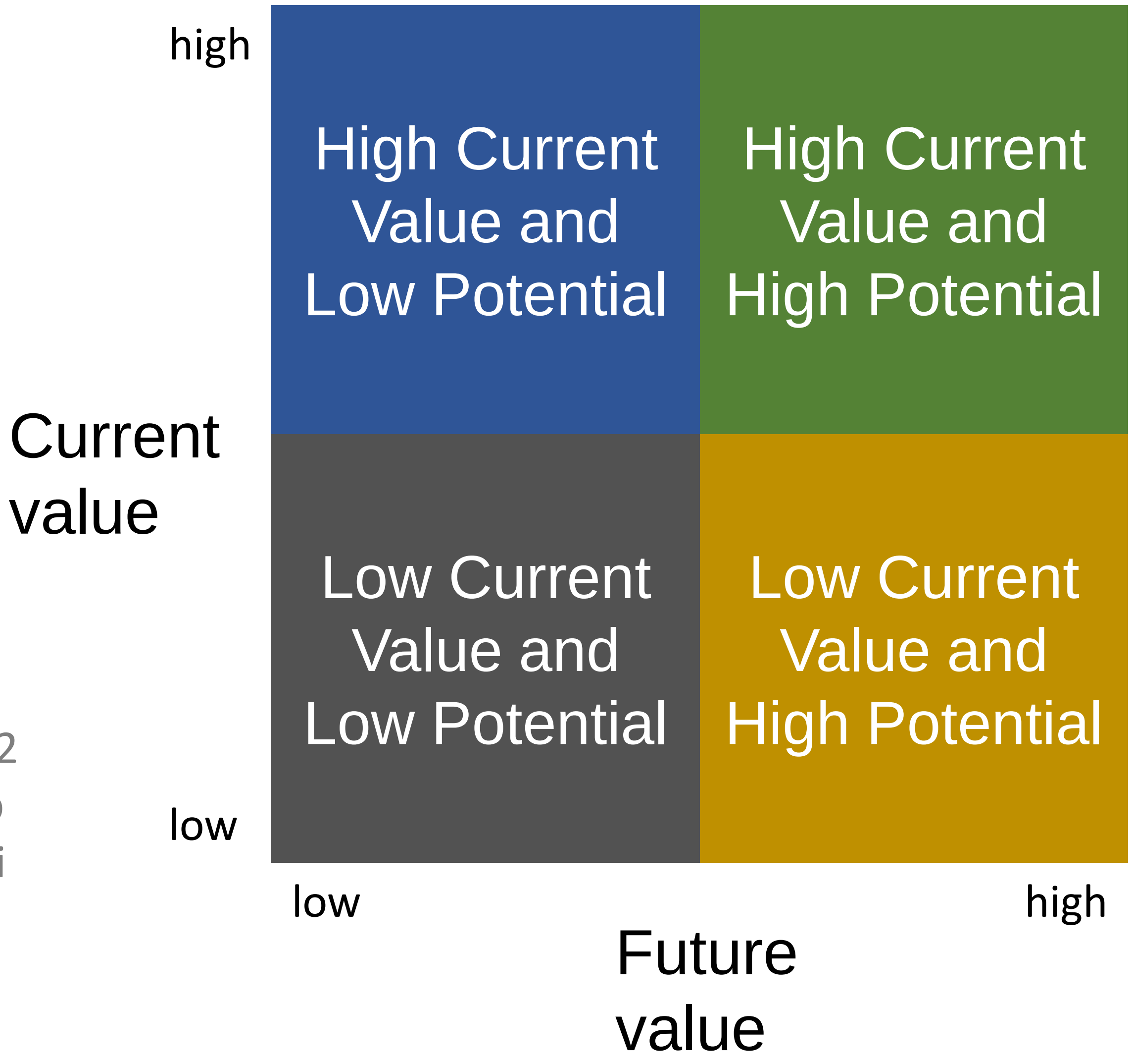
First coined by Ward Cunningham, inventor of the wiki.

vs.

YAGNI

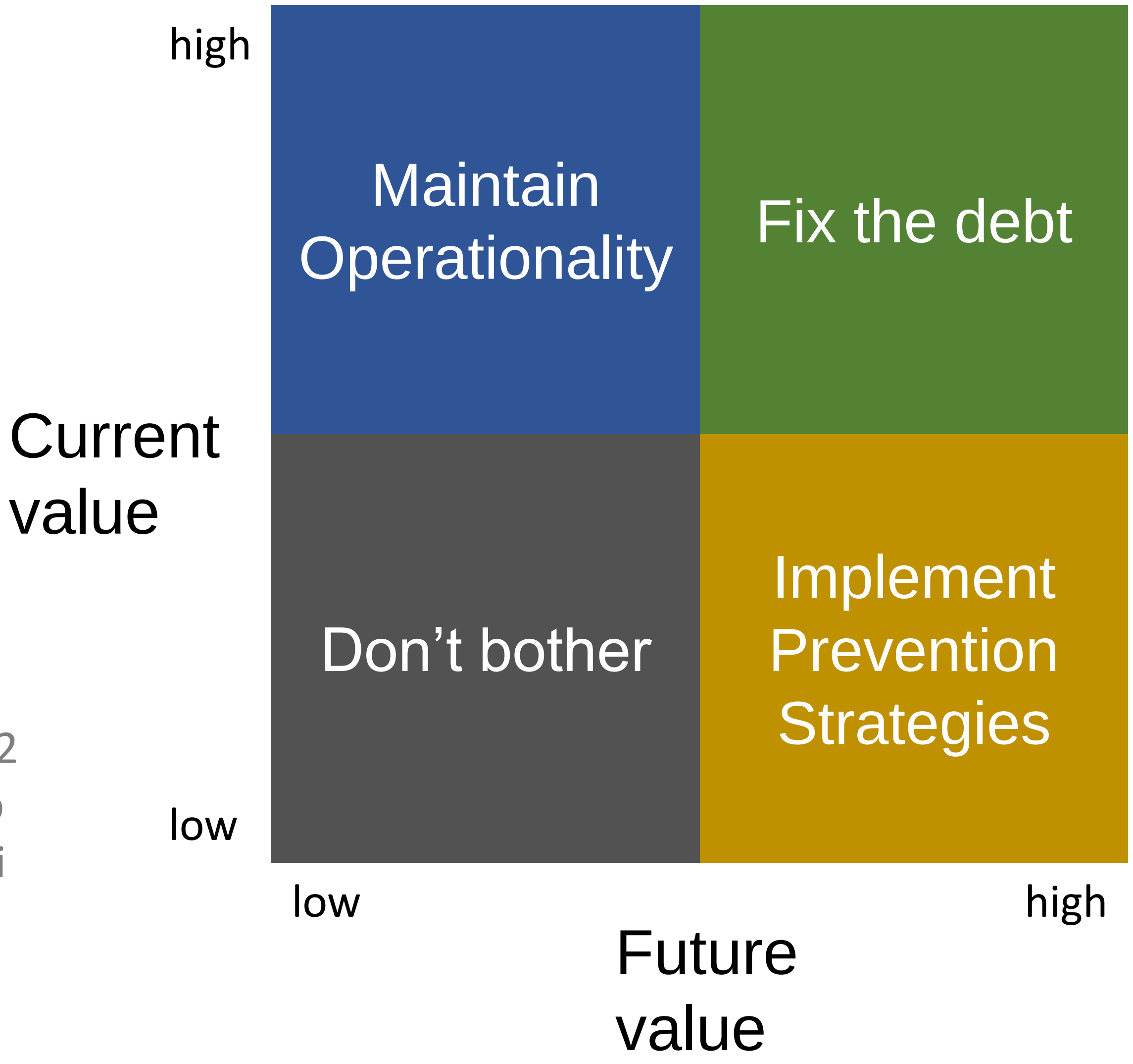
You ain't gonna need it…

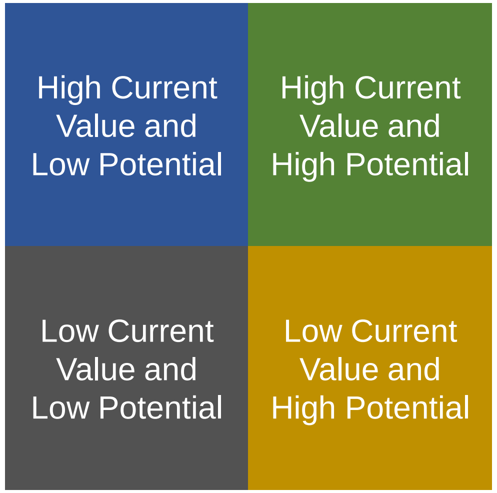
A Method to Determine when to Adapt Technology (Refactoring) by **Value**



http://blogs.ripple-rock.com/SteveGarnett/2013/03/05/TechnicalDebtStrategiesTacticsForAvoidingRemovingIt.aspx

A Method to Determine when to Adapt Technology (Refactoring) by **Value**





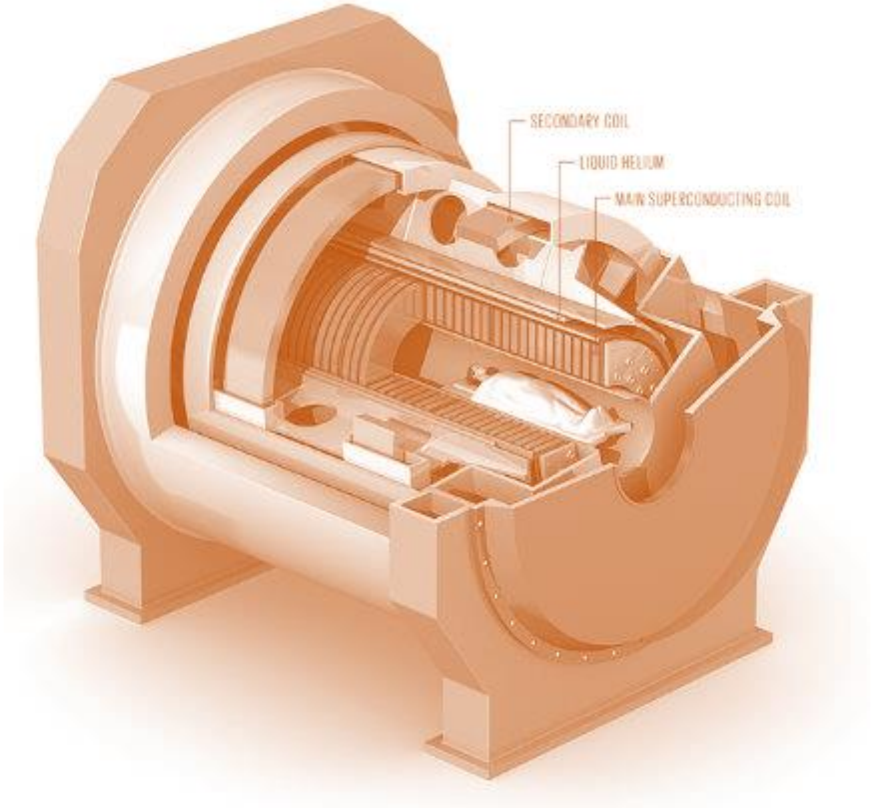
http://blogs.ripple-rock.com/SteveGarnett/2013/03/05/TechnicalDebtStrategiesTacticsForAvoidingRemovingIt.aspx

Six Open Questions

- Technical debt. Should we adapt?
- Are our efforts future proof?
- What are the specific challenges for cognitive neuroscience when it comes to sharing data and open data?
- How can cognitive neuroscientists improve collaborations with computational and basic/systems neuroscientists?
- What are the consequences for neuroscience education?
- What is our Mars expedition?

Main Literature

- Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, evidence, and data science.* NY: Cambridge University Press.
- Ludäscher, B et al. (2009). Scientific Process Automation and Workflow Management. In: Eds. Shoshani, A., & Rotem, D. (2009). *Scientific Data Management: Challenges, technology, and deployment.* Chapman and Hall/CRC. doi: 10.1201/9781420069815
- Sutter, H. The Free Lunch is Over: A fundamental turn toward concurrency in software. Accessed 2017-05-29. *http://www.gotw.ca/publications/concurrency-ddj.htm*
- Sutter, H. (2012, 2011) Welcome to the Jungle. Accessed 2017-05-29. *https://herbsutter.com/welcome-to-the-jungle/*


SECONDARY COIL
LIQUID HELIUM
MAIN SUPERCONDUCTING COIL