

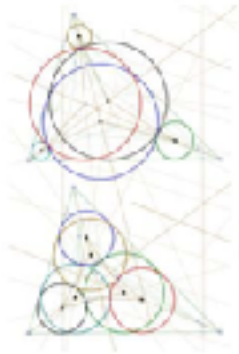


Open Science @ EPFL

Prof. Pierre Vandergheynst

**Data Management and Open Data Workshop
Lausanne, May 22th, 2017**

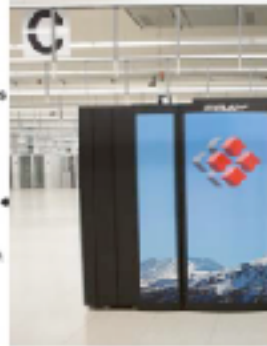
Deductive



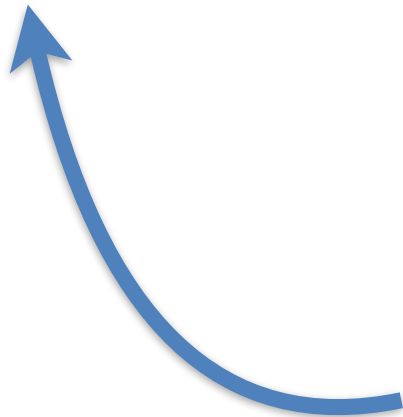
Empirical



Computational



Data driven (aka "Big Data")



Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Lipp propose how methods, publications and incentives must change if patients are to benefit.

47 (out of 52) foundational cancer studies are not reproducible

A Pragmatic Action Plan

Data sets

Electronic Labbooks

Code

Preprints

Education

Data Sets

Platform to archive / version data sets

Support Data Management Plans

Swiss Data Science Center: a pilot project



EMBEDDED R&D COLLABORATION

We engage in academic and industrial collaborations requiring large-scale distributed data processing (Big & Fast Data) and/or advanced analytics (machine learning & statistics) combined with an in-depth knowledge in select domains



DOMAIN-SPECIFIC INSIGHTS AS A SERVICE

We provide secure access to our cloud-hosted analytics platform - the Insights Factory, a highly scalable open software platform offering a one-stop-shop for hosting and exploring curated, calibrated and possibly anonymized data at scale, at-rest or in-motion



OPEN (DATA) SCIENCE

The Insights Factory offers user-friendly tooling and services to help with the adoption of Open Science, fostering research productivity and excellence

Electronic Labbooks

1665



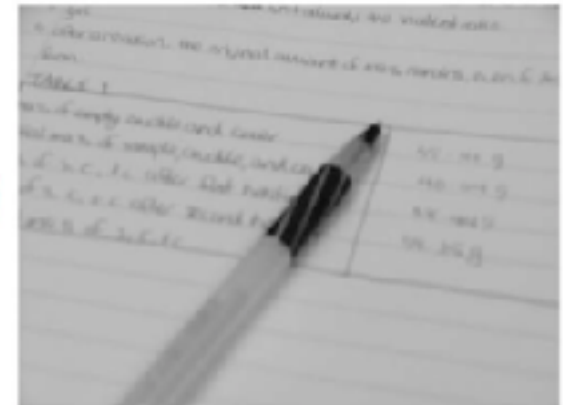
2017



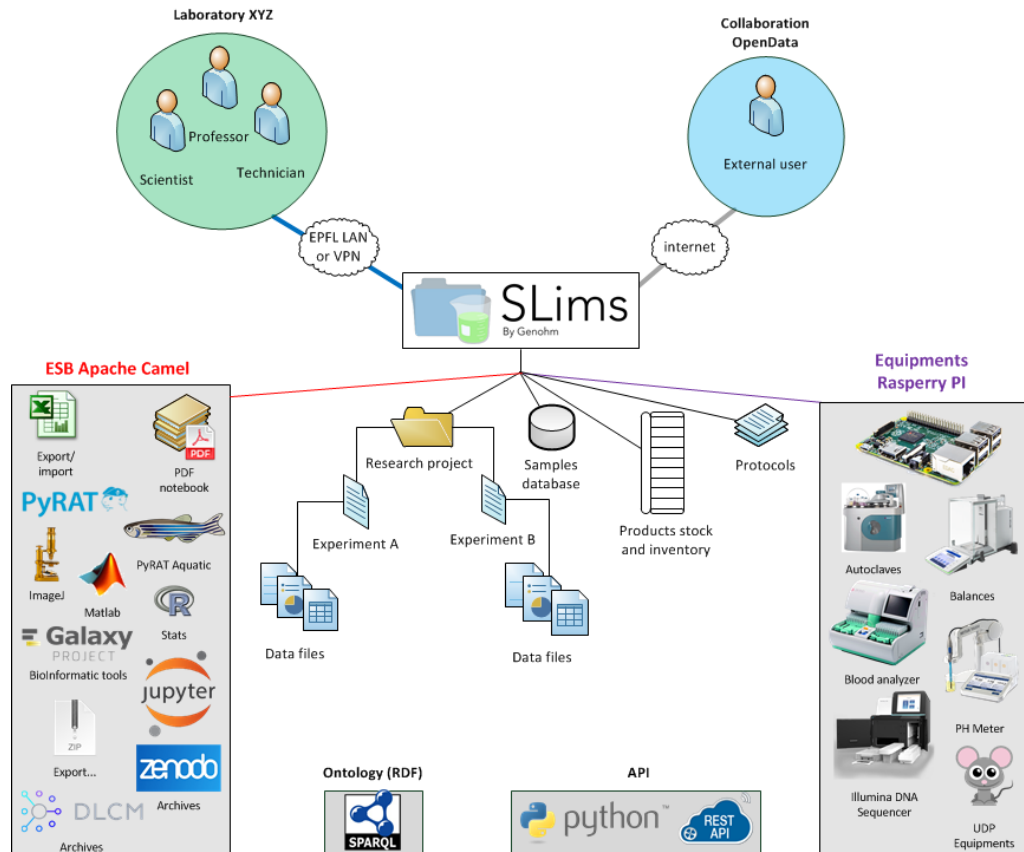
1876



2017

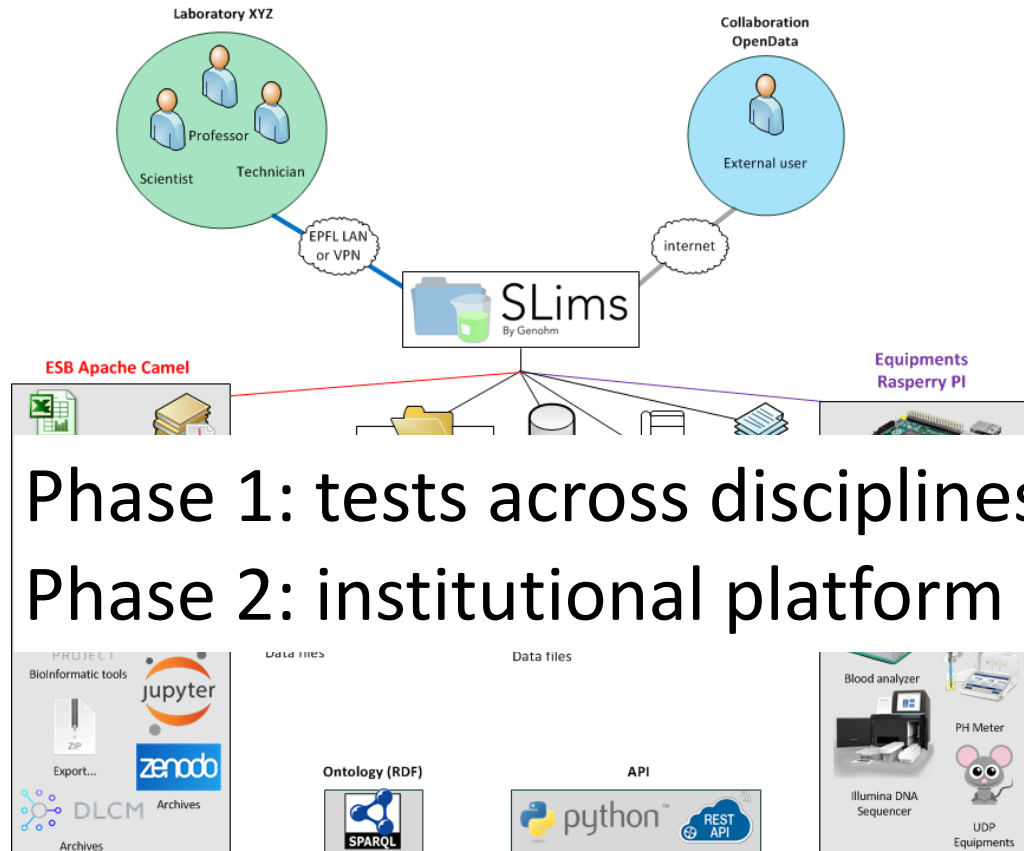


Pilot projects using SLIMS



in-house solutions also tested

Pilot projects using SLIMS



in-house solutions also tested

Code

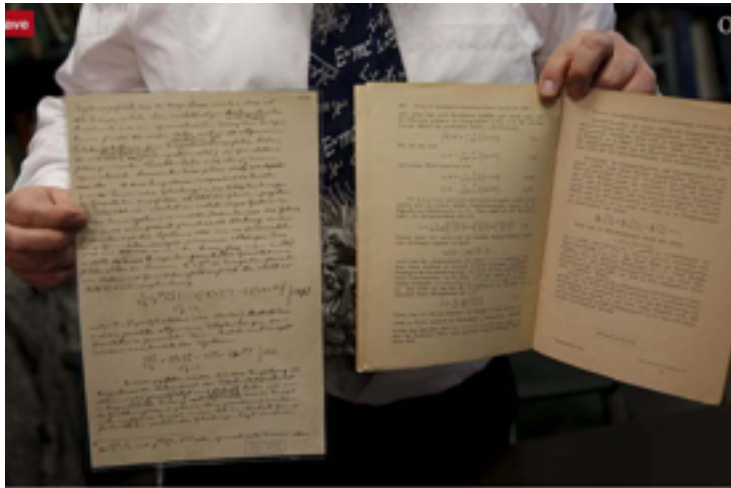
Code curation is well-known and tools exist
(GIT, ...)

But code alone has limited value: code+data,
parameters, dependencies, ...

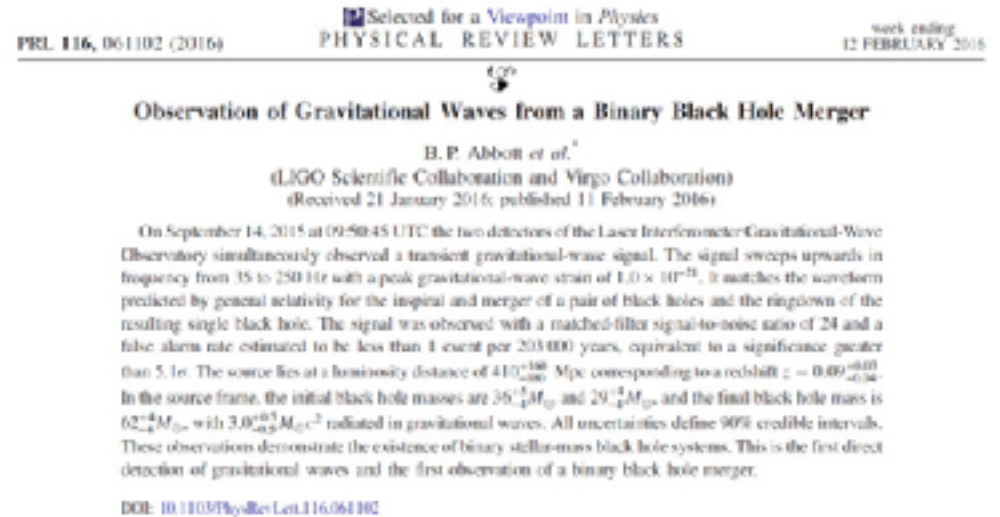
Follow “rich code” initiatives: C4Science, Beat-eu

Preprints

The format of “The Paper” has not changed much in 100 years



Gravitational waves, take 1



Gravitational waves, take 2

Only the medium has changed

Academic publishing scores VERY LOW on innovation

The Paper 2.0

Where is the code ? Where is the data ?

```
In [8]: # We start by suppressing the high frequencies with some tapering:
bb, aa = butter(4, [20.*2./fs, 330.*2./fs], btype='band')
strain_H1_whitebp = filtfilt(bb, aa, strain_H1_white)
strain_H1_whitebp = filtfilt(bb, aa, strain_H1_white)
RR_H1_whitebp = filtfilt(bb, aa, RR_H1_white)

# give the data a few samples
# first, shift H1 by 7 ms, and inverse. See the GW150914 detection paper for why!
strain_H1_shift = -np.roll(strain_H1_whitebp, int(0.007*fs))

plt.figure()
plt.plot(1000*times, strain_H1_whitebp, 'r', label='H1 strain')
plt.plot(1000*times+times*strain_H1_shift, 'g', label='H1 strain')
plt.plot(1000*times+times*strain_H1_shift, 'k', label='matched RR waveform')
plt.xlim([-0.1, 0.05])
plt.ylim([-5, 5])
plt.xlabel('time (s) since "a1" (msec)')
plt.ylabel('amplitude (m)')
plt.legend(loc='lower left')
plt.title('Advanced LIGO WHITENED strain data near GW150914')
plt.savefig('gw150914_strain_whitened.png')
```

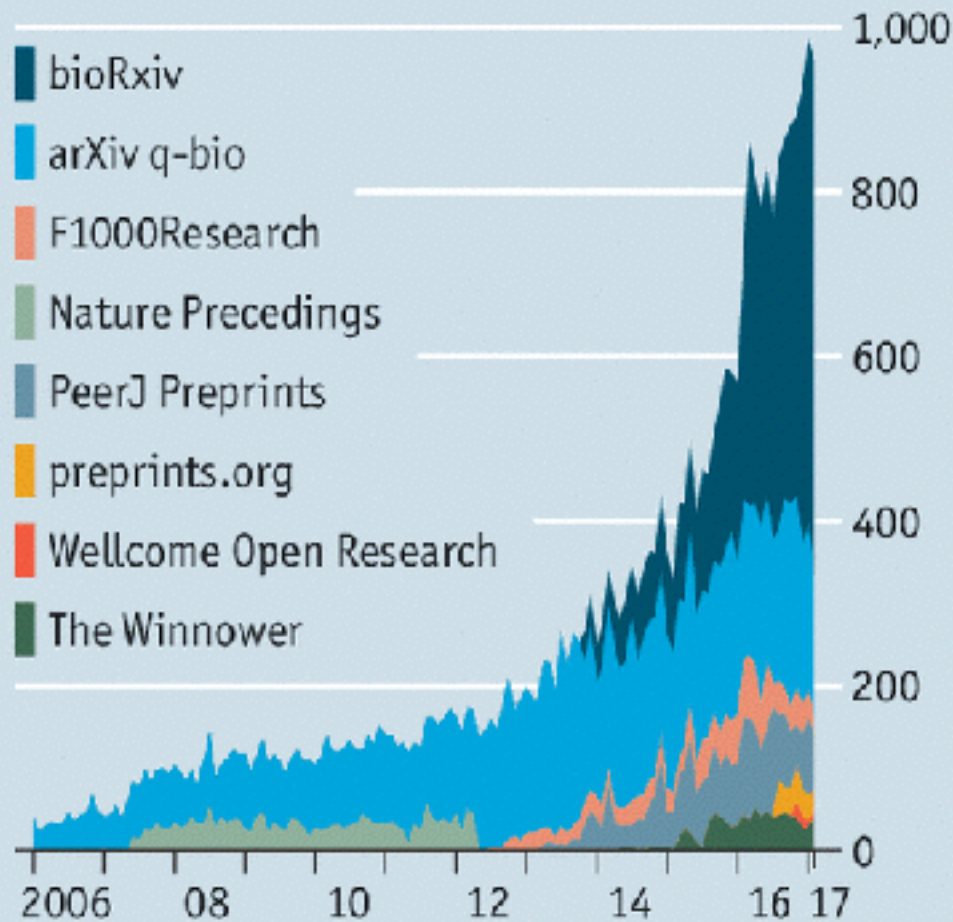


Gravitational waves, take 3

Fundamental issues in reviewing & reproducibility

Print first, ask questions later

Biomedical preprint submissions, by month, to



Source: Jordan Anaya

Data

Preprint

Code



Invenio Digital Library Framework

Build your own fully customised digital library, institutional repository, multimedia archive, or research data repository on the web.

Education

Awareness

Develop good practices

Integrate into your research workflow

An introduction to Open Science for all PhD students at EPFL

