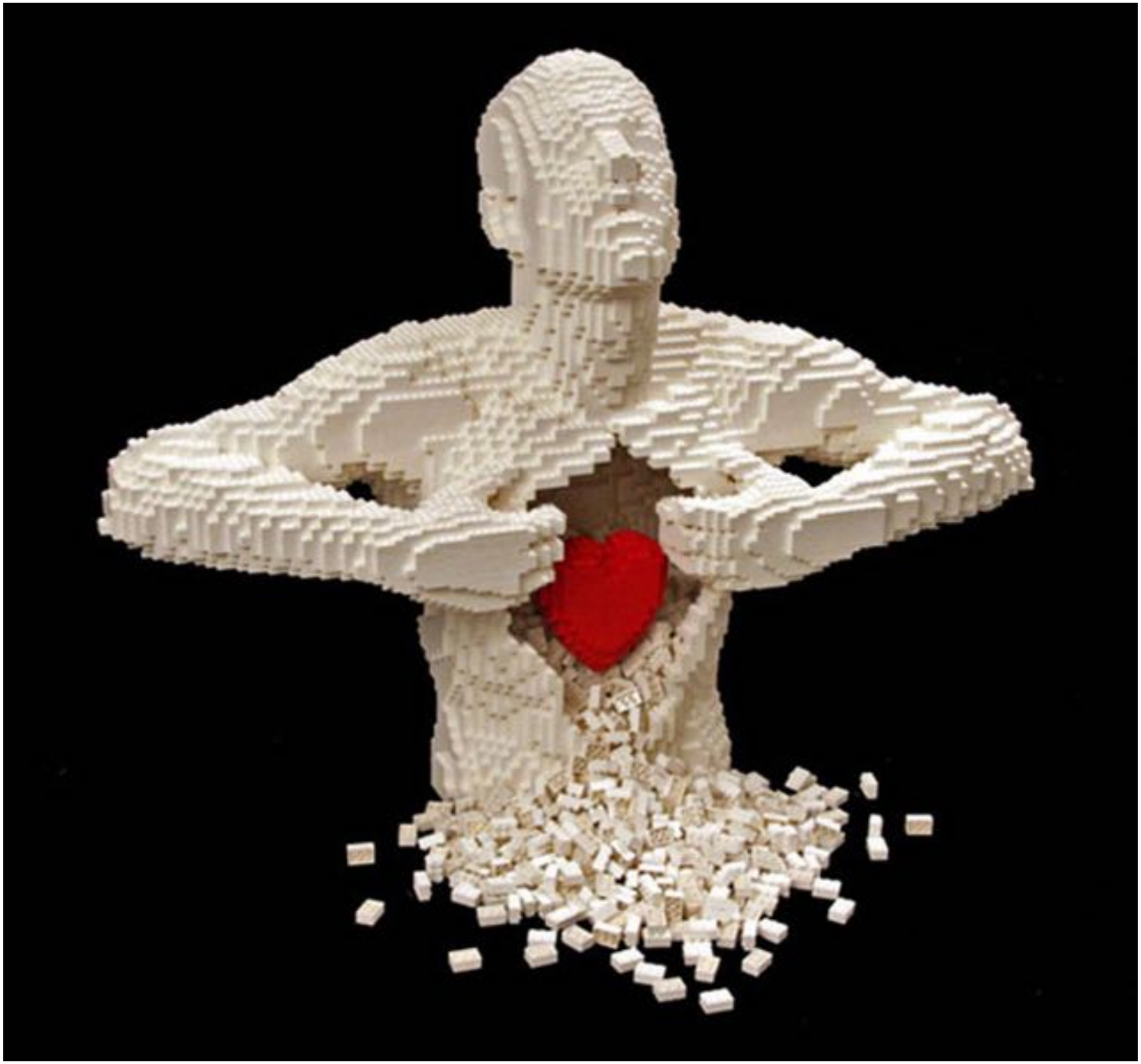
Introduction to tools for reproducible research

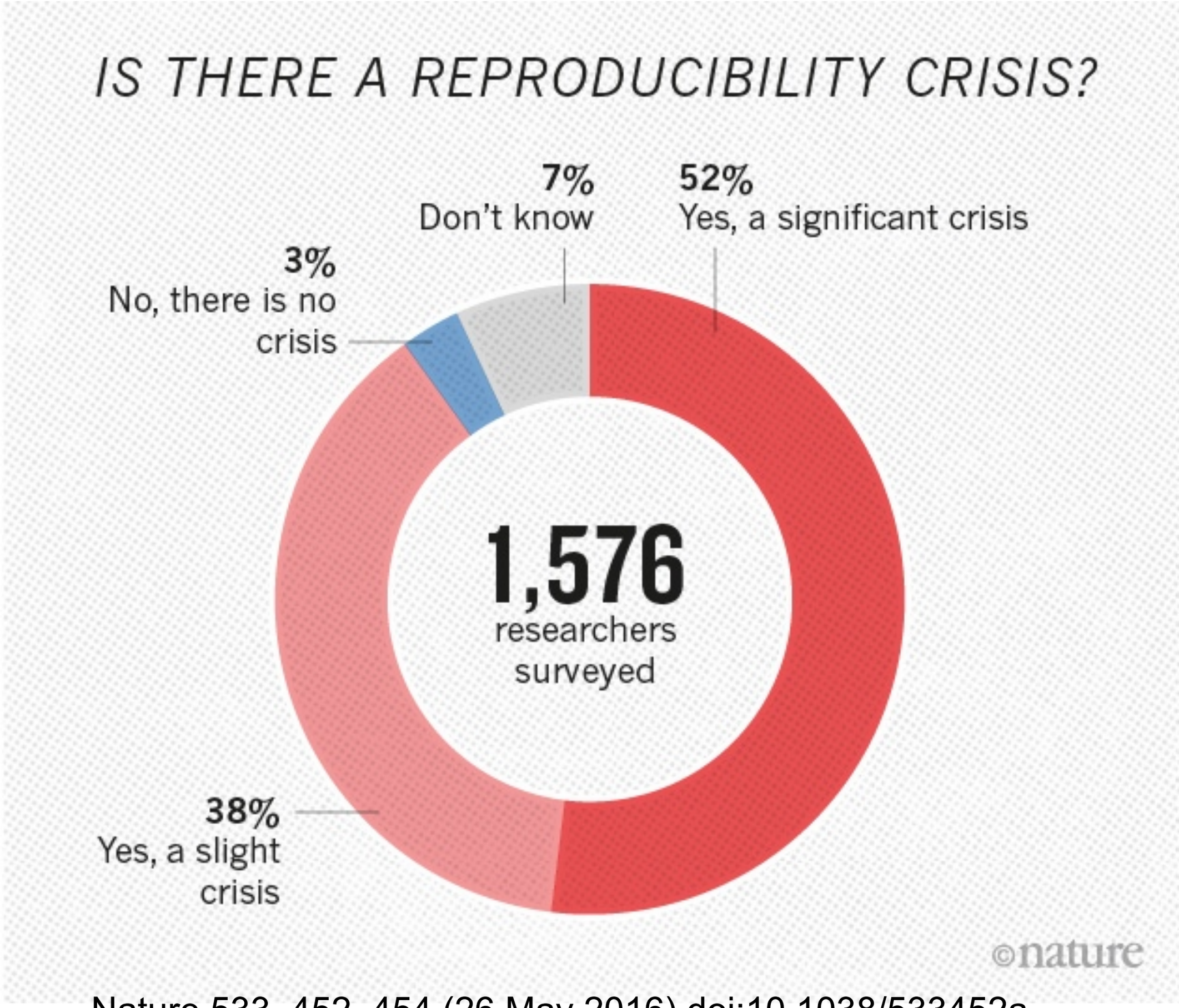
Tom Pollard and Felipe Torres

Laboratory for Computational Physiology,
Massachusetts Institute of Technology (MIT), USA

- Why care about reproducibility?
- Introduction to MIMIC-III, the critical care database
- Executable notebooks
- Code publishing platforms.
- Version control systems.

Why should we care about reproducibility?

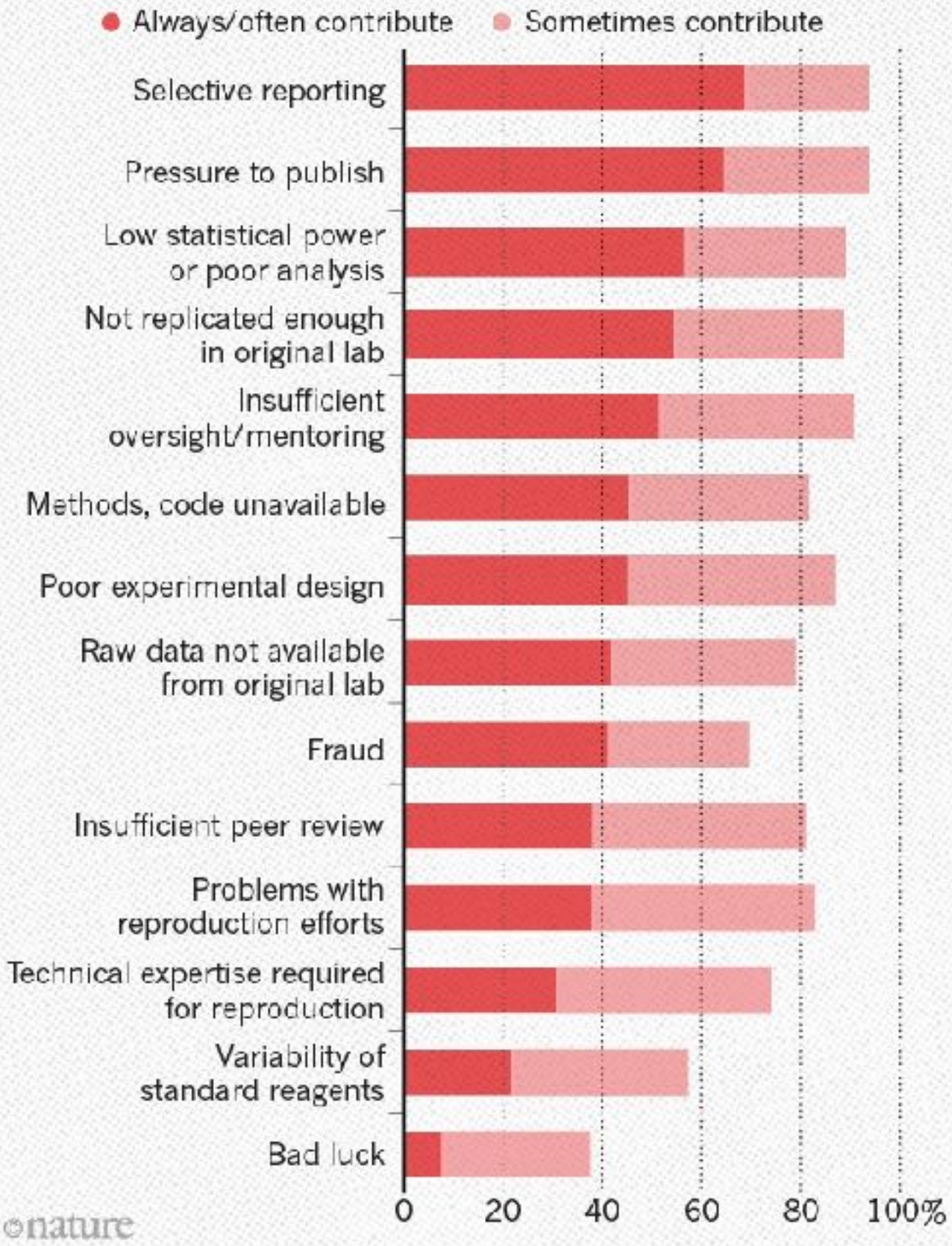




Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a

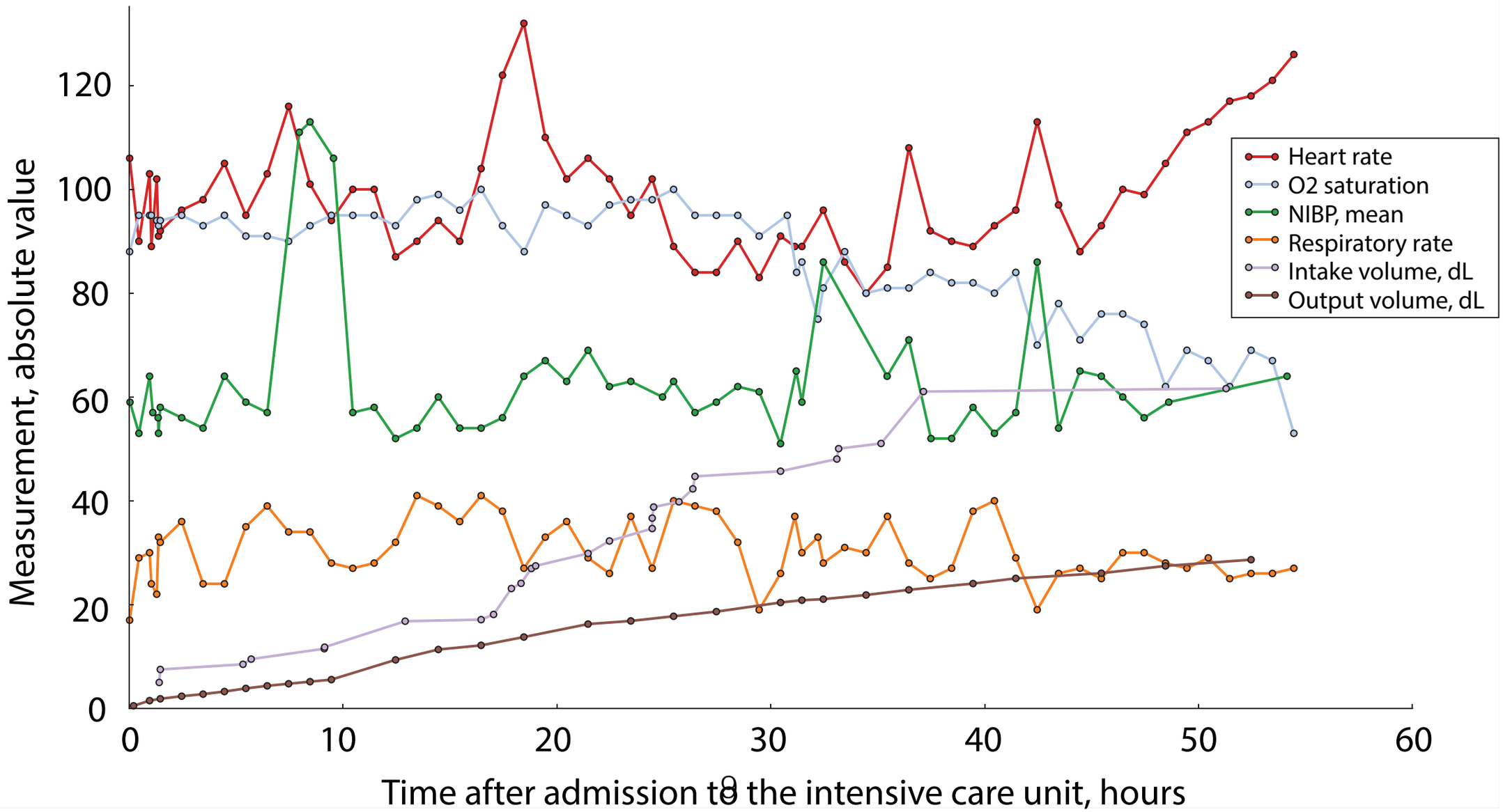
WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.



http://mimic.physionet.org

Code status	Full code					Comfort measures
GCS: Verbal GCS: Moto GCS: Eye	Oriented Obeys commands Spontaneously	Oriented Obeys commands Spontaneously	Oriented Obeys commands Spontaneously	Confused Obeys commands To speech	Confused Obeys commands To speech	Incomprehensible sounds Flex-withdraws None
Platelet, K/uL Creatinine, mg/dL White blood cell, K/uL Neutrophil, %	48 53 0.7 9.1 12.4 37	46 0.7 16.8			45 0.8 23.2	
Morphine Sulfate Vancomycin (1 dose) Piperacillin (1 dose) NaCl 0.9% Amiodarone Dextrose 5%	10.0mL/hour	10.0mL/hour 1mg/min 50mL/hour	0.5mg/min 25mL/hour	0.5mg/min 25mL/hour	10.0mL/hour	



Admission Date: [**2952-11-3**] Discharge Date: [**2952-11-9**]

Date of Birth: [**2887-7-23**] Sex: F

Service: MEDICINE

Allergies:
No Known Allergies / Adverse Drug Reactions

Attending:[**First Name3 (LF) 3925**]
Chief Complaint:
Sepsis, respiratory distress

Major Surgical or Invasive Procedure:
None

History of Present Illness:
F w/ h/o metastatic breast cancer to breast and lungs
currently receiving CMT, brought to the ED by
rehab for abnormal labs. She was found to be neutropenic, anemia
and thrombocytopenic. At the rehab, vitals were reportedly T
100.4, HR 107, BP 92/42. There is also a concern for possible...

Widely used internationally


THE LANCET
Respiratory Medicine

Articles

Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study

Dr Romain Pirracchio, MD, Maya L Petersen, MD, Marco Carone, PhD, Matthieu Resche Rigon, MD, Prof Sylvie Chevret, MD, Prof Mark J van der Laan, PhD

Published Online: 23 November 2014

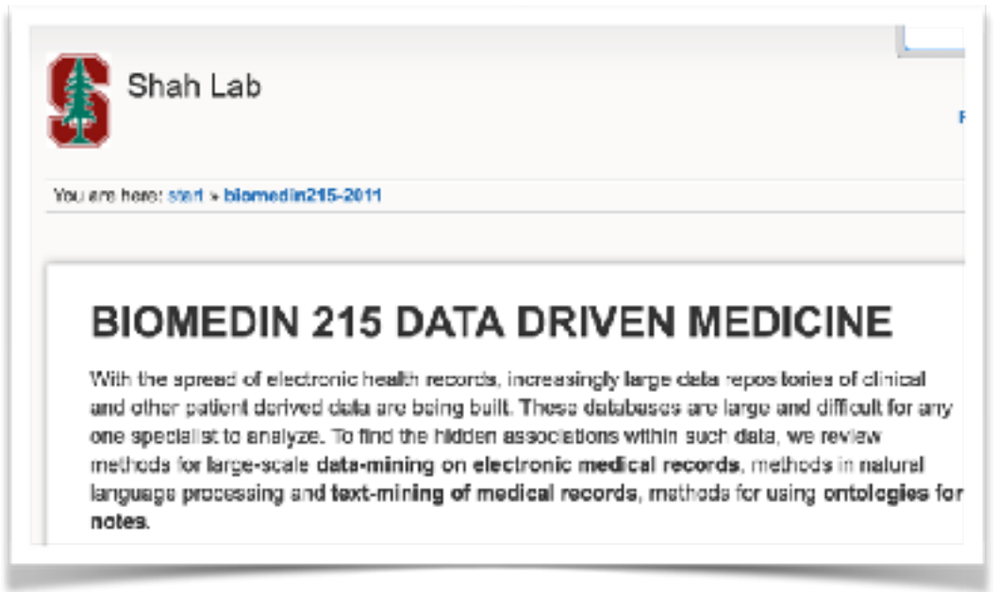
Altmetric 18

DOI: http://dx.doi.org/10.1016/S2213-2600(14)70239-5 | CrossMark

Article Info

This article can be found in the following collections: Respiratory medicine-other

Research


Shah Lab

You are here: start » biomedin215-2011

BIOMEDIN 215 DATA DRIVEN MEDICINE

With the spread of electronic health records, increasingly large data repositories of clinical and other patient derived data are being built. These databases are large and difficult for any one specialist to analyze. To find the hidden associations within such data, we review methods for large-scale **data-mining on electronic medical records**, methods in natural language processing and **text-mining of medical records**, methods for using **ontologies** for **notes**.

Education

SHARE

PERSPECTIVE | REPRODUCIBILITY

Science Translational Medicine

A "datathon" model to support cross-disciplinary collaboration

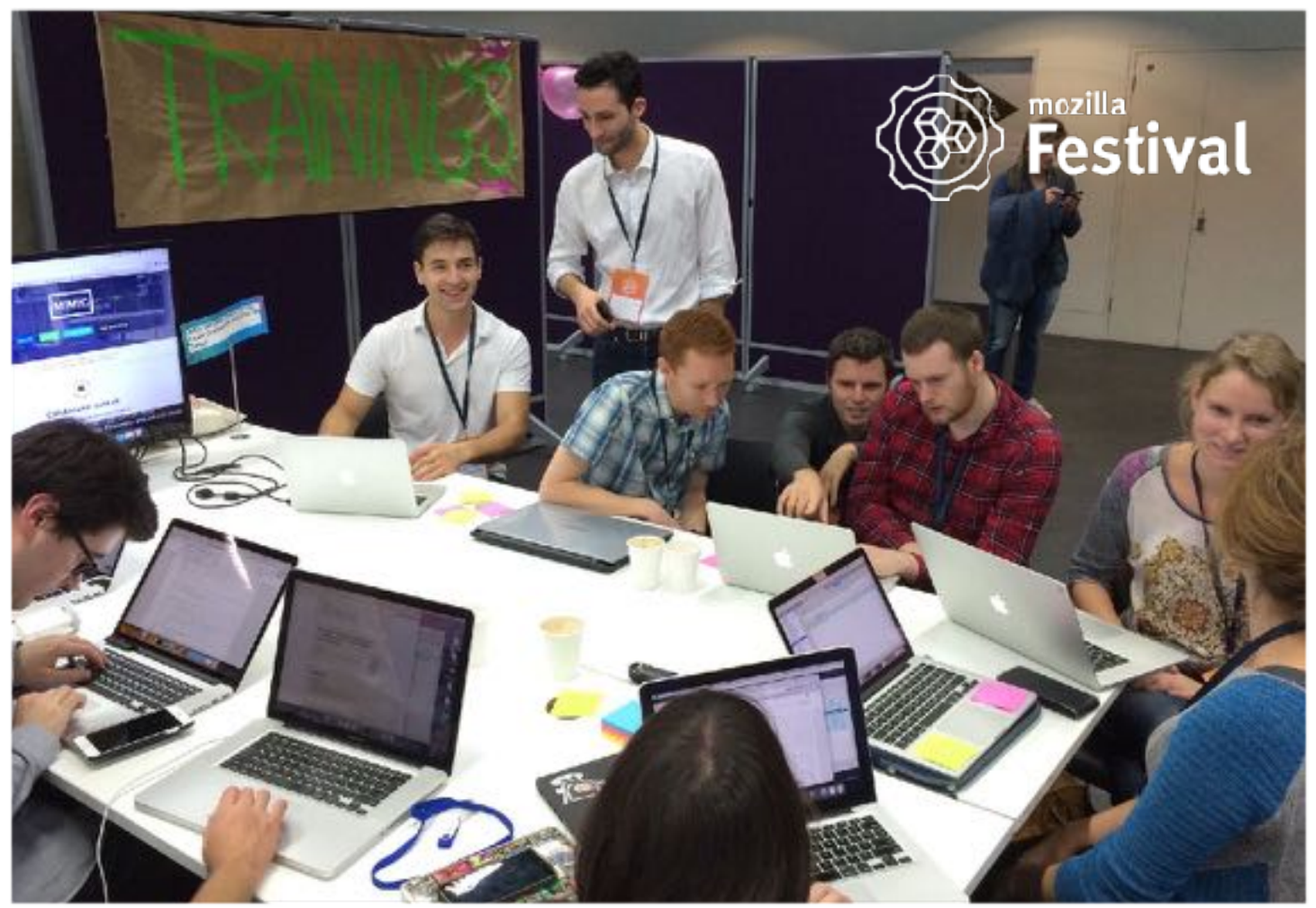
Jerôme Aboab[1,*], Leo Anthony Celi[1], Peter Charlton[1], Mengling Feng[1], Mohammad Ghassemi[1], Dominic C. Marshall[1,†], Louis Mayaud[1], Tristan Naumann[1], Ned McCague[1], Kenneth E. Paik[1], Tom J. Pollard[1], Matthieu Resche-Rigon[1], Justin D. Salciccioli[1] and David J. Stone[2,3]

+ Author Affiliations

↵†Corresponding author. E-mail: domini

↵* All authors contributed equally to thi

Science Translational Medicine 06 Apr 2
Vol. 8, Issue 333, pp. 333ps8
DOI: 10.1126/scitranslmed.aad9072

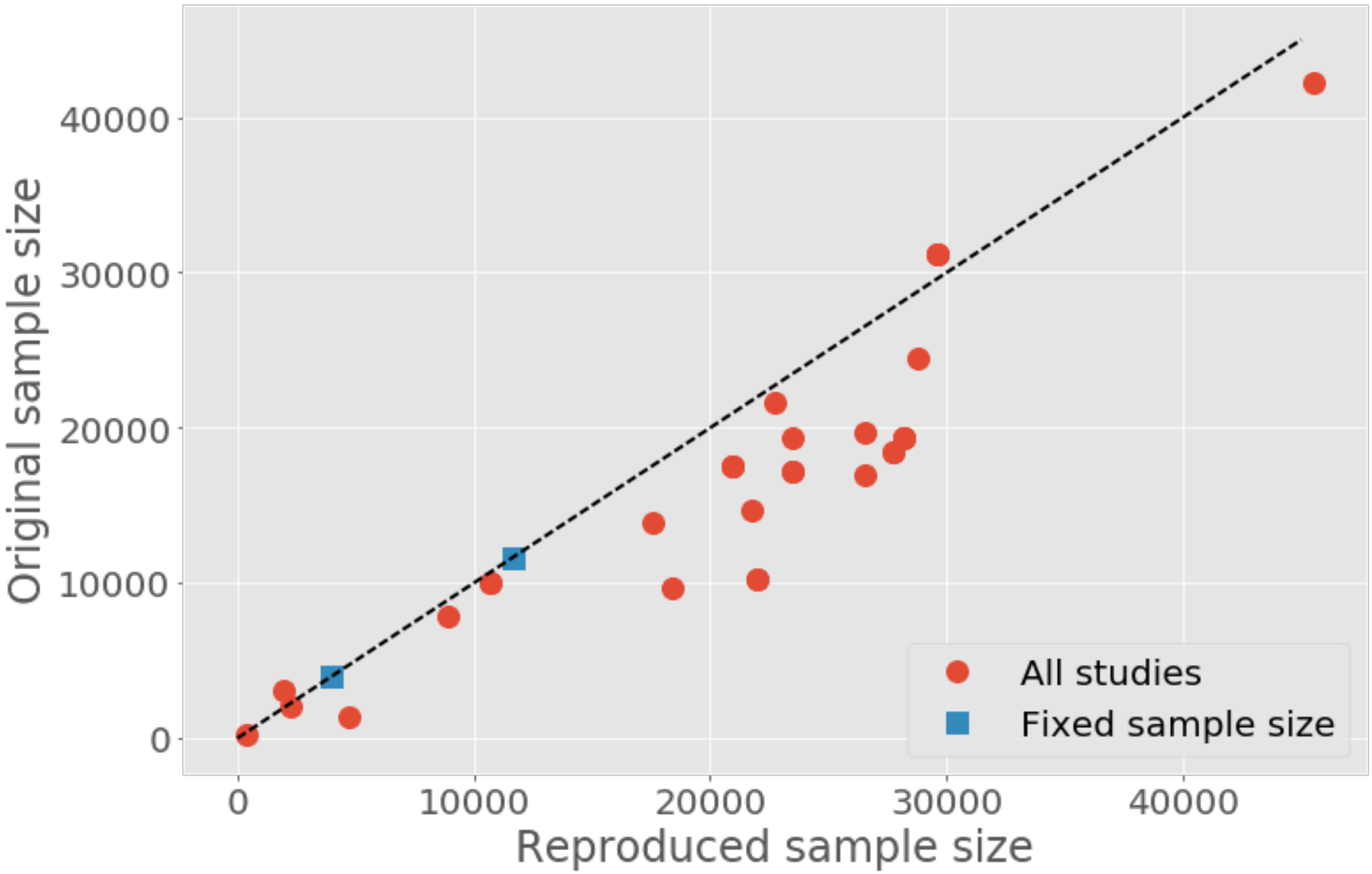


MIMIC is freely available critical care database

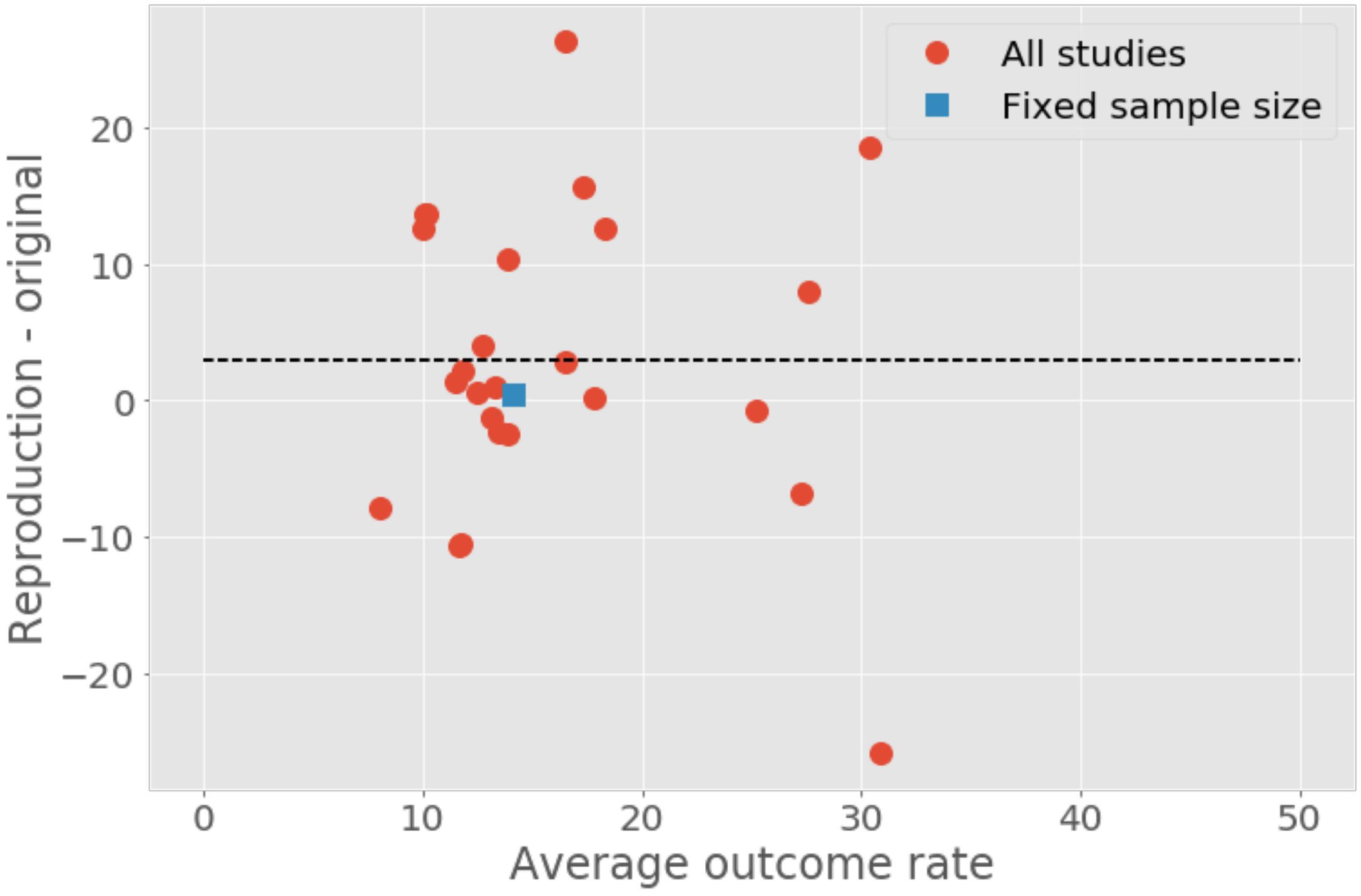
So papers using the MIMIC are reproducible. Right?

- Collected all studies which attempted to predict mortality using MIMIC
- Attempted to regenerate the patient cohort using the published methodology
- Compared our cohort with the published cohort

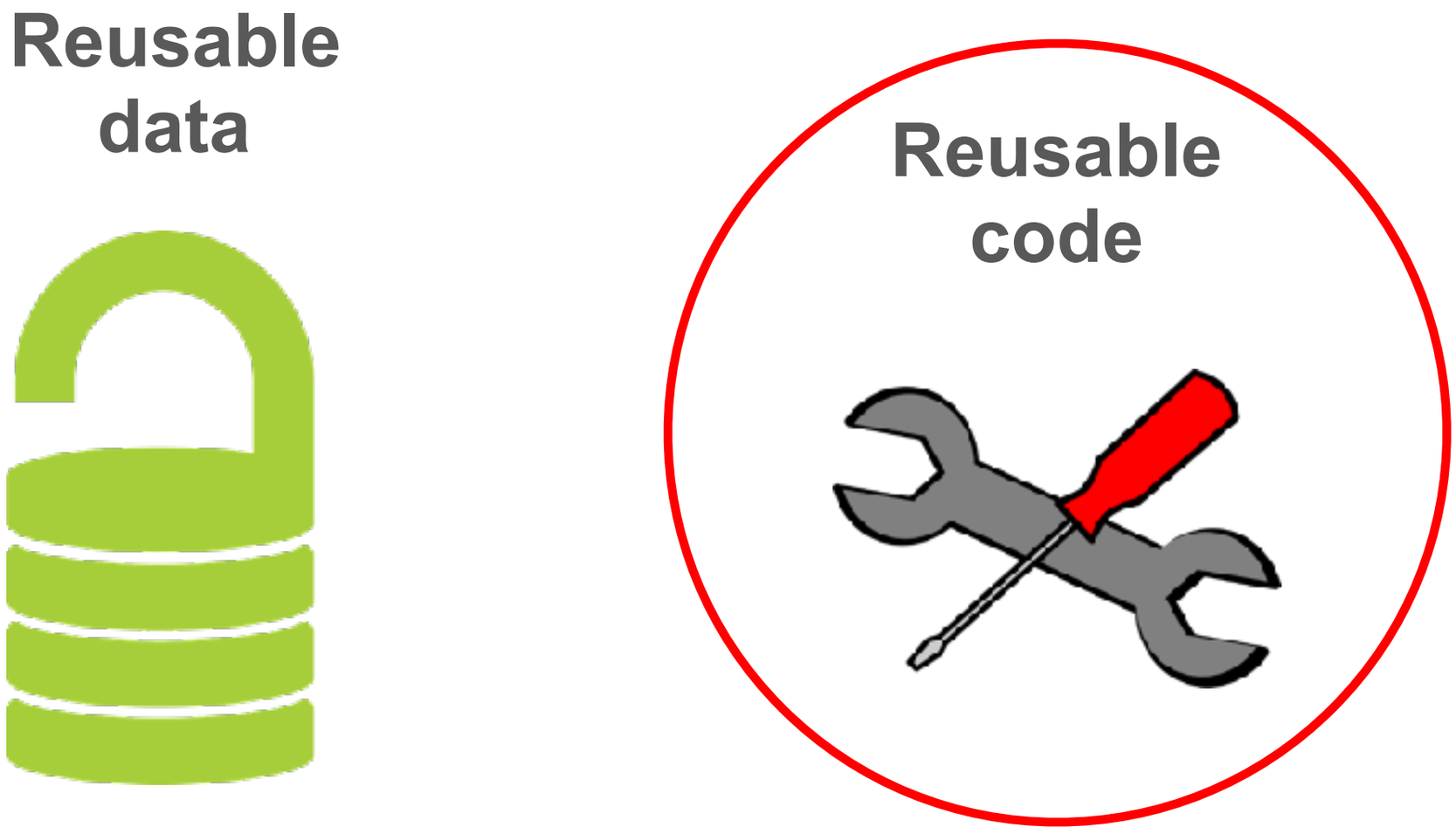
Results - sample size



Results - % mortality



Openly available data isn’t enough!



Proceedings of Machine Learning for Healthcare 2017 JMLR W&C Track Volume 68

Reproducibility in critical care: a mortality prediction case study

Alistair E. W. Johnson AEWJ@MIT.EDU
Institute of Medical Engineering & Science
Massachussetts Institute of Technology
Cambridge, MA, USA

Tom J. Pollard TPOLLARD@MIT.EDU
Institute of Medical Engineering & Science
Massachussetts Institute of Technology
Cambridge, MA, USA

Roger G. Mark RGMARK@MIT.EDU
Institute of Medical Engineering & Science
Massachussetts Institute of Technology
Cambridge, MA, USA

Abstract

Mortality prediction of intensive care unit (ICU) patients facilitates hospital benchmarking and has the opportunity to provide caregivers with useful summaries of patient health at the bedside. The development of novel models for mortality prediction is a popular task in machine learning, with researchers typically seeking to maximize measures such as the area under the receiver operator characteristic curve (AUROC). The number of 'researcher degrees of freedom' that contribute to the performance of a model, however, presents a challenge when seeking to compare reported performance of such models.

In this study, we review publications that have reported performance of mortality prediction models based on the Medical Information Mart for Intensive Care (MIMIC) database and attempt to reproduce the cohorts used in their studies. We then compare

Some tools for reproducible research…

1. Executable notebooks
(“literate computing”)

Jupyter Notebook

jupyter mimic-lcs Last Checkpoint: 2 hours ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 2

Markdown

Length of stay in the ICU

This document shows how Jupyter Notebooks can be used to create a reproducible analysis using MIMIC-III. Let's calculate the median length of stay in the ICU and then include this value in our document.

Connect to the demo database

The demo database is a subset of the MIMIC-III database containing ~100 patients.

In [5]:
```
# Import libraries
import sqlite3
import pandas as pd
%matplotlib inline
```

In [6]:
```
# Connect to MIMIC demo
con = sqlite3.connect('../data/mimiciii_demo_100.sqlite')
```

Query the database and assign the results to a variable

Tables are described in the "Tables in eICU" section of the website. For example, the icustays table is described at:
https://mimic.physionet.org/mimictables/icustays/

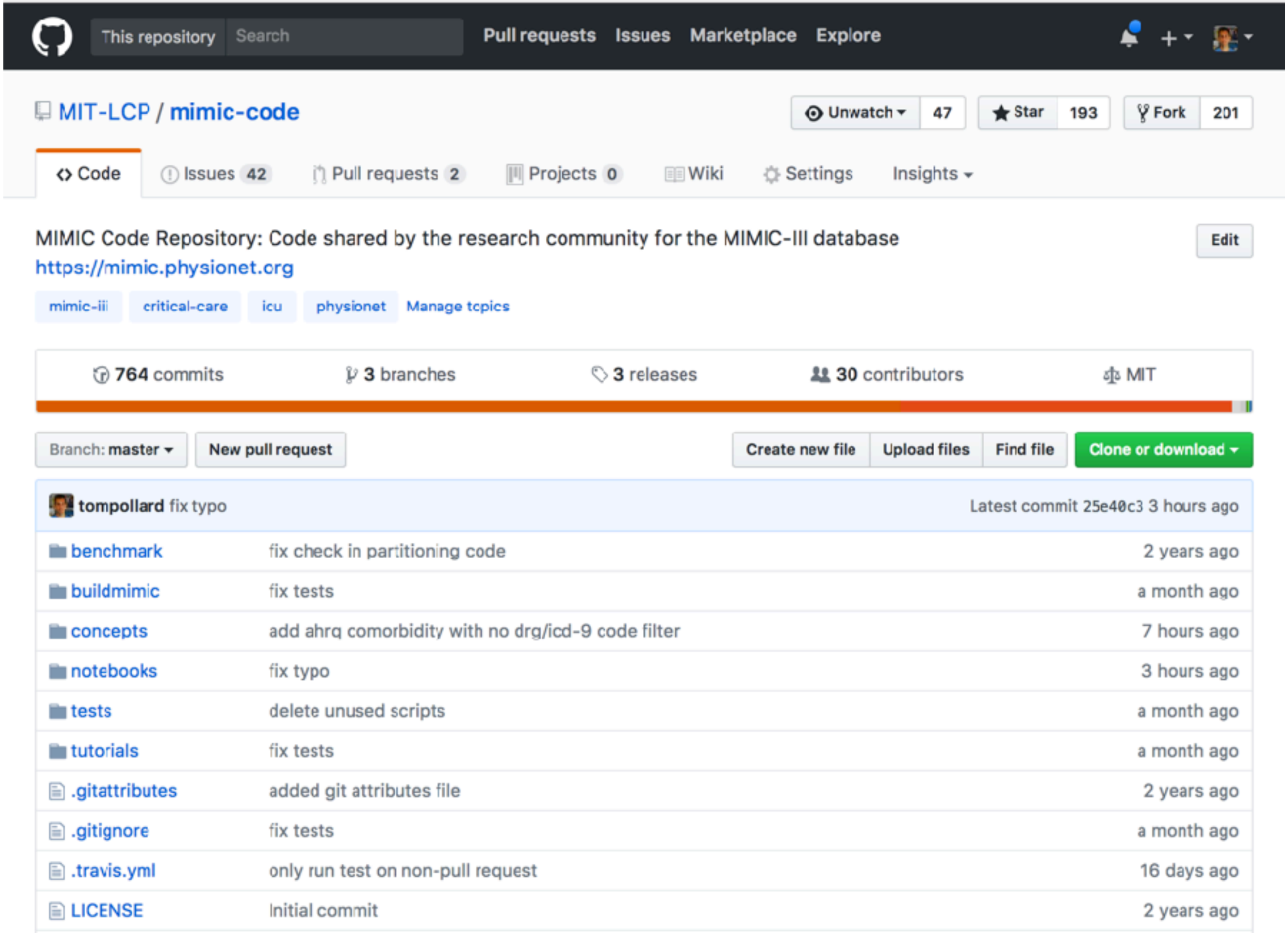
In [14]:
```
# Query the MIMIC demo dataset
query = \
"""
SELECT i.subject_id, i.hadm_id, i.los
FROM icustays i;
"""

data = pd.read_sql_query(query,con)
```

Examples are provided on the MIMIC Code Repository:

https://mimic.physionet.org/

2. Code publishing platforms (GitHub/Bitbucket/ GitLab/…)

https://github.com/MIT-LCP/mimic-code



3. Version control systems that integrate with Github etc.

"FINAL".doc


FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012


```
$ git init
Initialized empty Git repository in /tmp/tmp.IMBYSY7R8Y/.git/
$ cat > README << 'EOF'
> Git is a distributed revision control system.
> EOF
$ git add README
$ git commit
[master (root-commit) e4dcc69] You can edit locally, and push
to any remote.
 1 file changed, 1 insertion(+)
 create mode 100644 README
$ git remote add origin git@github.com:cdown/thats.git
$ git push -u origin master
```

A command-line session showing repository creation, addition of a file, and remote synchronization

Original author(s)	Linus Torvalds[1]
Developer(s)	Junio Hamano and others[2]

MIT-LCP / mimic-code

Unwatch 47 | Star 193 | Fork 201

Code | Issues 42 | Pull requests 2 | Projects 0 | Wiki | Settings | Insights

Branch: master

Commits on Sep 21, 2017

fix typo
tompollard committed 4 hours ago ✓ — 25e40c3

add rmd example
tompollard committed 4 hours ago ✓ — cf4af9a

rename consort diagram folder
tompollard committed 4 hours ago ✗ — e9c297c

add ahrq comorbidity with no drg/icd-9 code filter
alistairewj committed 8 hours ago ✓ — faa276a

add ahrq comorbid with no drg filter
alistairewj committed 8 hours ago — 079ee75

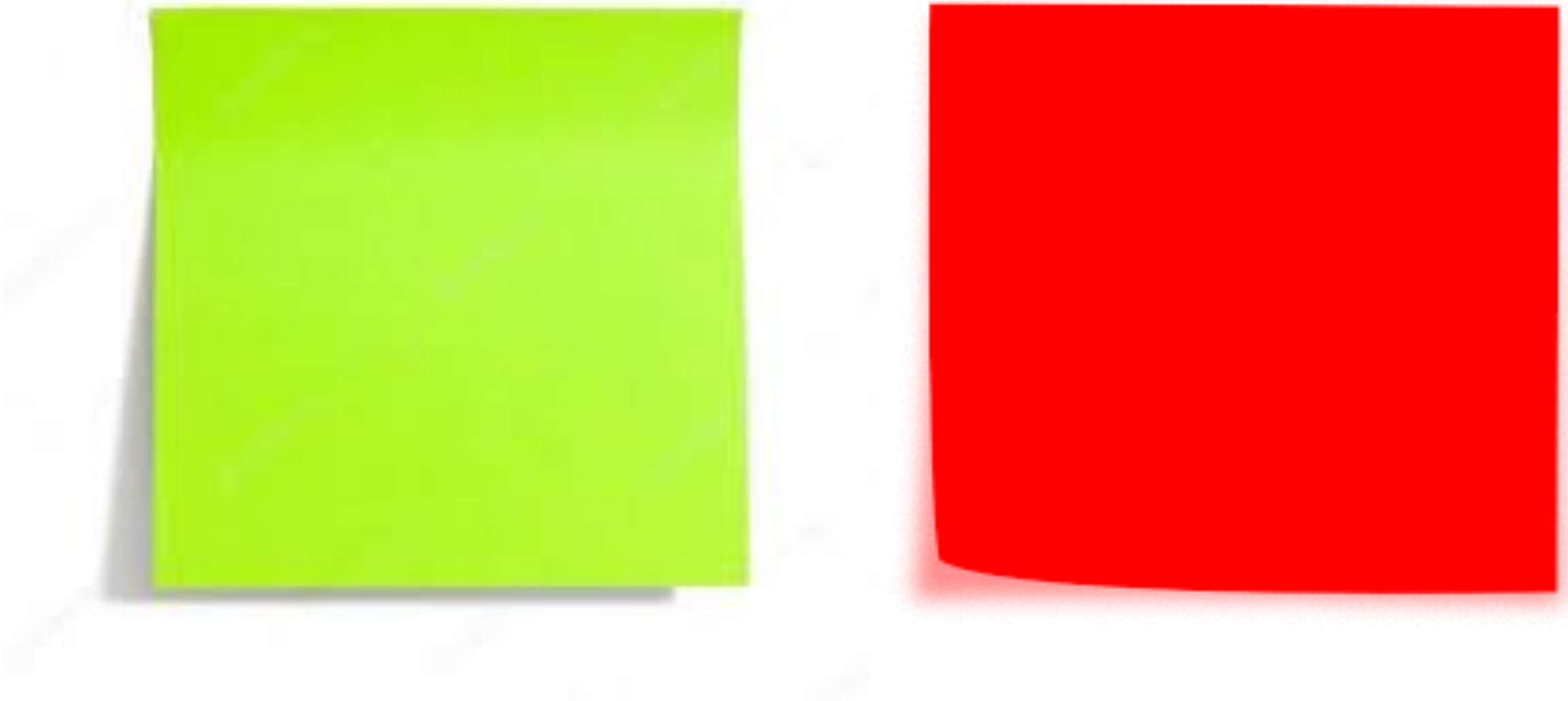
Commits on Sep 13, 2017

format syntax
tompollard committed 8 days ago ✓ — 57c3537

Commits on Sep 7, 2017

Merge pull request #307 from gitter-badger/gitter-badge-2 ...
alistairewj committed on GitHub 14 days ago ✓ — cca2066

Add Gitter badge
gitter-badger committed 14 days ago — bc914b7



https://github.com/tompollard/buenosaires2018