Data Sharing:
Why is it still so hard?

John Muschelli
Assistant Scientist
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Data Types

Raw

- GT3X
- CWA
- Bin files

Raw-ish

- HDF5
- CSV - gzipped vs. not

Some need applications/licenses - ideal to have open source solutions to open

Data Types

Interpolated/Filtered/Aggregated

- What epoch?
    - 1 second?
    - 5 second?
- Measure to use?
    - $r = \sqrt(x^2 + y^2 + z^2)$
    - Mean of x, y, z?
    - Variance?

Can't invert these - can't go back!

RAW

Interpolated/Filtered/Aggregated

- What epoch?
  - 1 second?
  - 5 second?
- Measure to use?
  - r = √(x^2 + y^2 + z^2)
  - Mean of x, y, z?
  - Variance?

Can't invert these - can't go back!

Journals: Guidelines

- Data Availability Upon Request

- Have it "somewhere"

OR

But where?

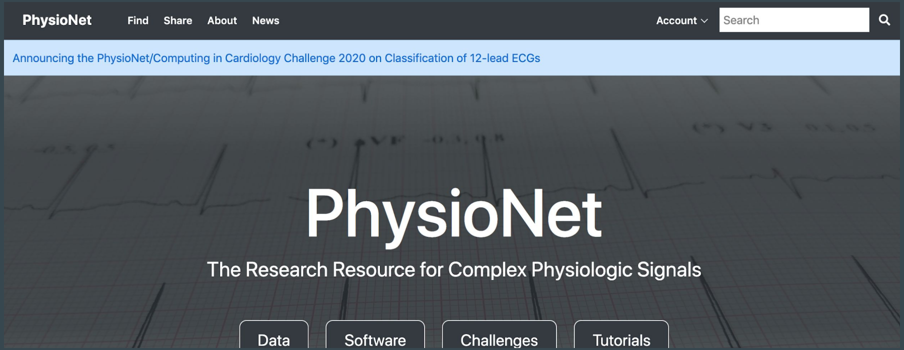
Example of biologically-related repos

- Physionet

General data

- Figshare
- Zenodo
- Synapse
- Datacite

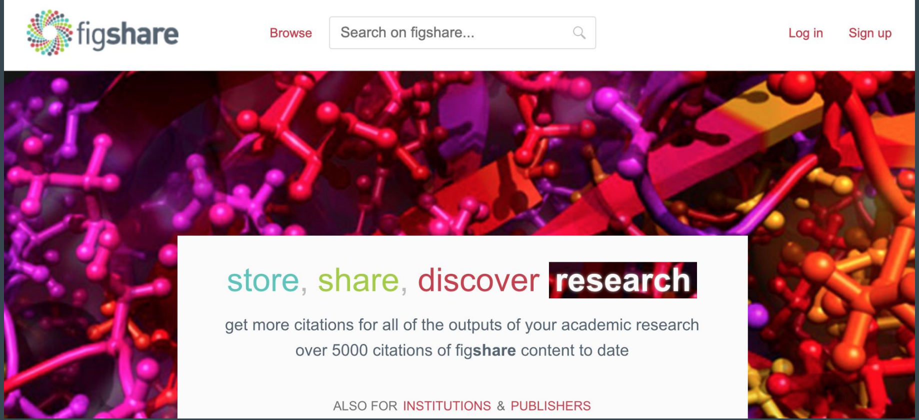
What are your other favorites?



Announcing the PhysioNet/Computing in Cardiology Challenge 2020 on Classification of 12-lead ECGs

PhysioNet

The Research Resource for Complex Physiologic Signals



store, share, discover **research**

get more citations for all of the outputs of your academic research
over 5000 citations of fig**share** content to date

ALSO FOR INSTITUTIONS & PUBLISHERS

Get a Digital Object Identifier (DOI)

Make sure it's

- FAIR - Findable, Accessible, Interoperable, and Reusable

But our commodity is paper/citations (it's true, so needs):

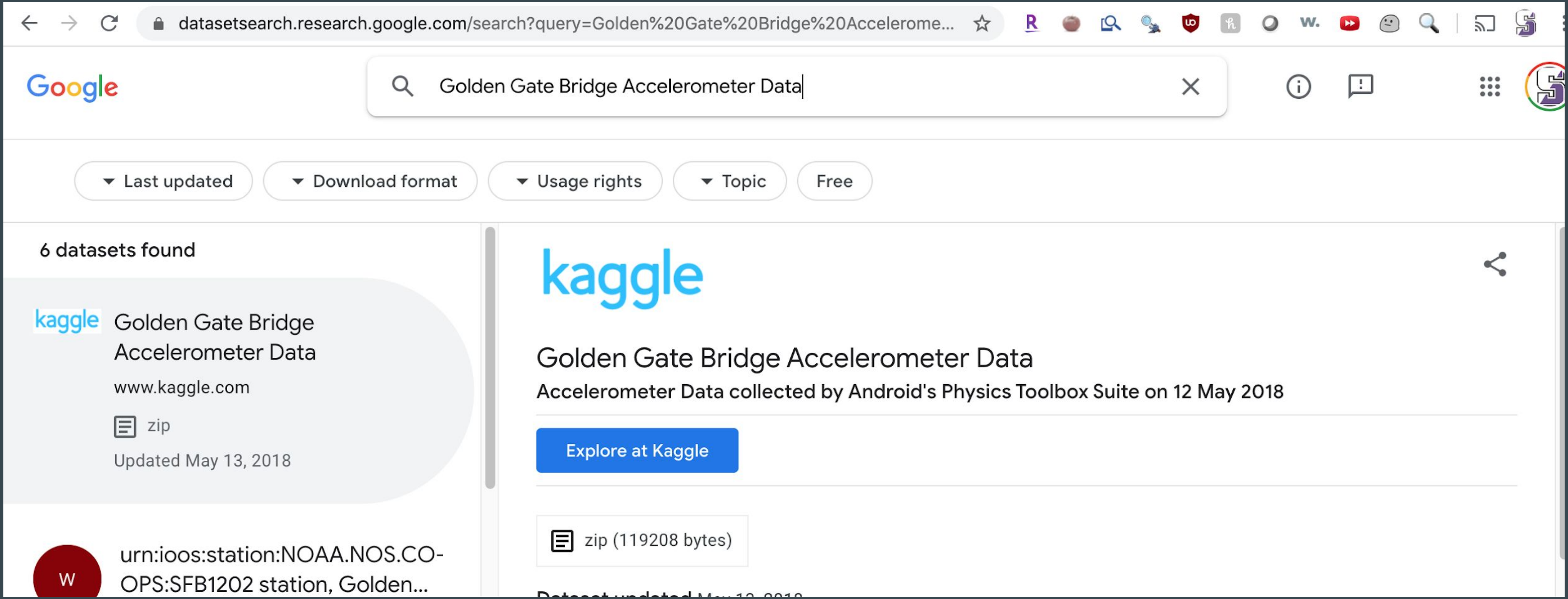
- Citable

Funding Agencies

- No central repository
- Other fields (e.g. genomics) have one that you ***need*** to put in
  - dbGaP - https://www.ncbi.nlm.nih.gov/gap/

One of the main issues starting one: funding

Searching can be hard: https://datasetsearch.research.google.com/



datasetsearch.research.google.com/search?query=Golden%20Gate%20Bridge%20Accelerome...

Golden Gate Bridge Accelerometer Data

6 datasets found

kaggle Golden Gate Bridge Accelerometer Data
www.kaggle.com
zip
Updated May 13, 2018

urn:ioos:station:NOAA.NOS.CO-OPS:SFB1202 station, Golden...

kaggle

Golden Gate Bridge Accelerometer Data

Accelerometer Data collected by Android's Physics Toolbox Suite on 12 May 2018

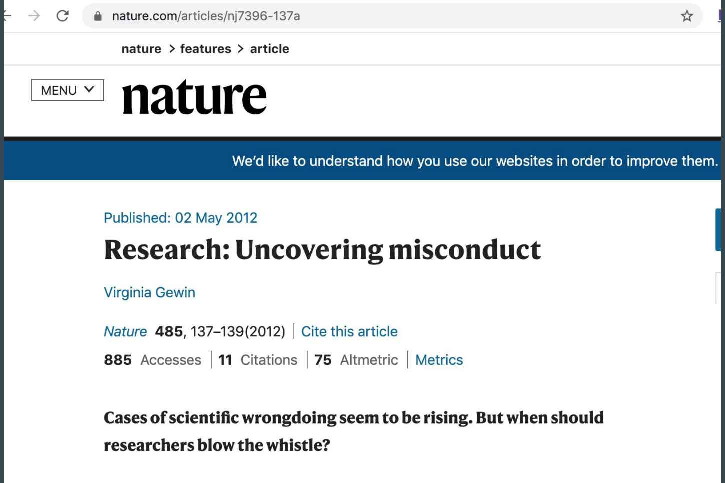
zip (119208 bytes)

Dataset updated May 13, 2018

Downsides: Why people don't share

Issue with non-replicability

- People see if your results replicate
- High-profile examples of errors
  - https://www.theverge.com/2013/4/17/4234136/excel-calculation-error-infamous-economic-study
  - https://www.nature.com/articles/nj7396-137a
- It's a Legitimate concern, but
  - Want to do good science or not be wrong?



We'd like to understand how you use our websites in order to improve them.

Published: 02 May 2012

Research: Uncovering misconduct

Virginia Gewin

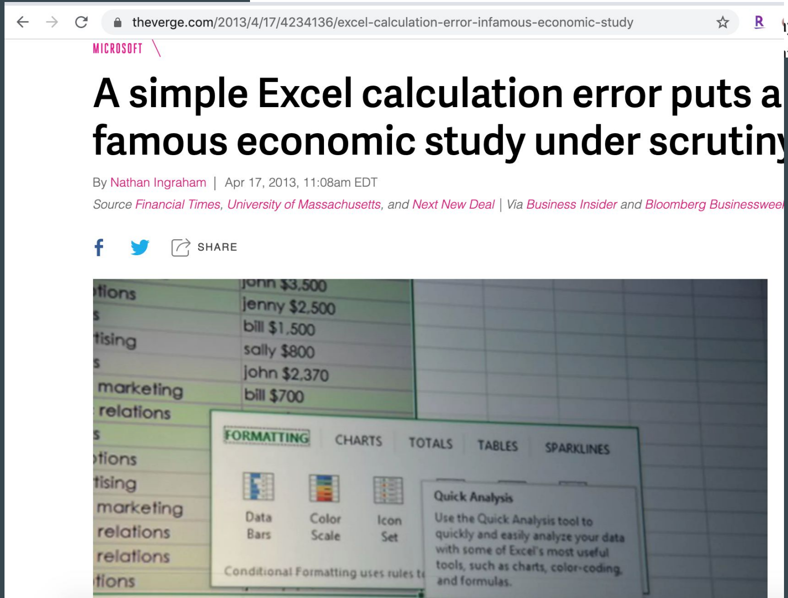
Nature **485**, 137–139(2012) | Cite this article

885 Accesses | **11** Citations | **75** Altmetric | Metrics

Cases of scientific wrongdoing seem to be rising. But when should researchers blow the whistle?

ly, were

ts by


theverge.com/2013/4/17/4234136/excel-calculation-error-infamous-economic-study
MICROSOFT
SHARE
john $3,500
jenny $2,500
bill $1,500
sally $800
john $2,370
bill $700
marketing
relations
FORMATTING
CHARTS
TOTALS
TABLES
SPARKLINES
Data Bars
Color Scale
Icon Set
Quick Analysis
Use the Quick Analysis tool to quickly and easily analyze your data with some of Excel's most useful tools, such as charts, color-coding, and formulas.
Conditional Formatting uses rules t

A simple Excel calculation error puts a famous economic study under scrutiny

By Nathan Ingraham | Apr 17, 2013, 11:08am EDT

Source Financial Times, University of Massachusetts, *and* Next New Deal | *Via* Business Insider *and* Bloomberg Businesswee

Downsides: Why people don't share

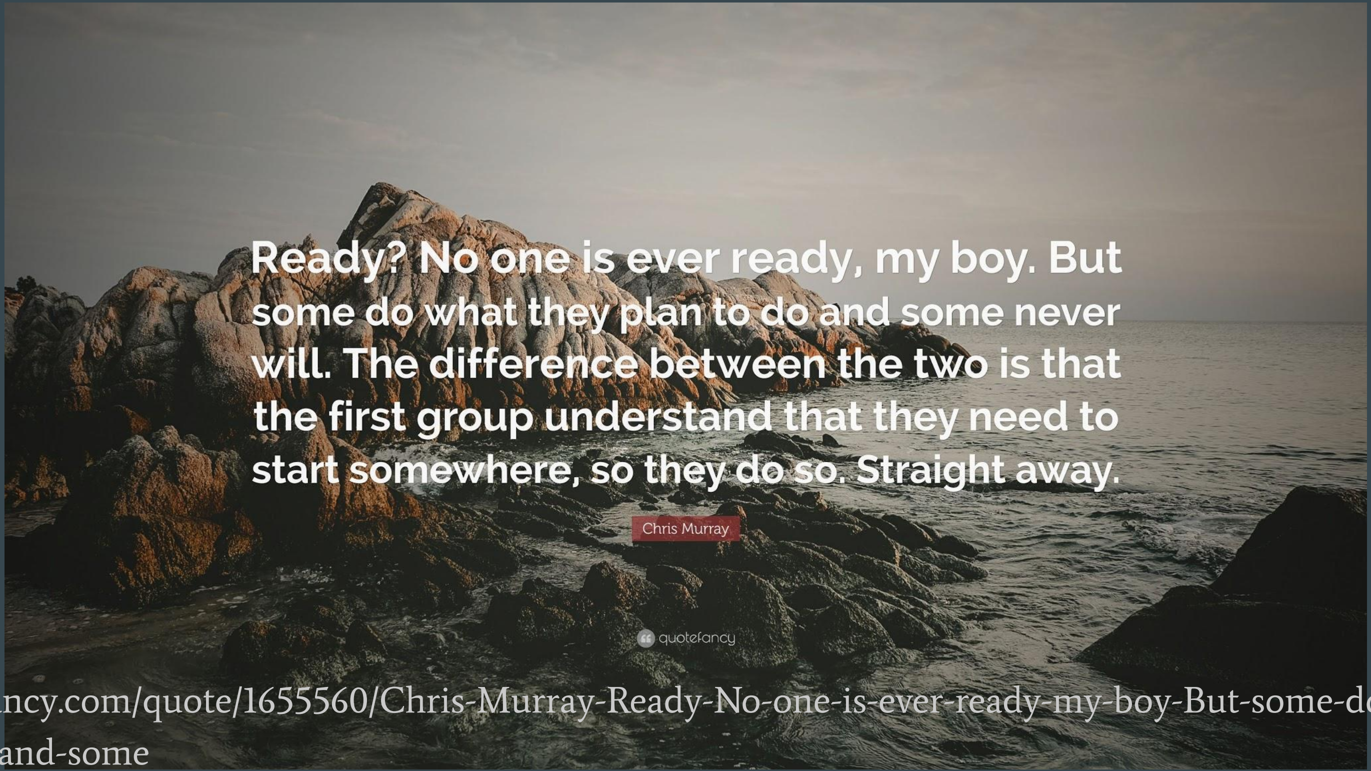
Many excuses

- **"I wanted to do that analysis", but many times →**
- **They'll scoop me**

PRIMARY DATA COLLECTORS SHOULD GET PRIORITY

- But the embargo time can't be ∞
- 5 years?
- What if publicly-funded?

Why people don't share: I'm not ready yet



Ready? No one is ever ready, my boy. But some do what they plan to do and some never will. The difference between the two is that the first group understand that they need to start somewhere, so they do so. Straight away.

Chris Murray

https://quotefancy.com/quote/1655560/Chris-Murray-Ready-No-one-is-ever-ready-my-boy-But-some-do-what-they-plan-to-do-and-some

Upside: Why should you still share

Replication

- Other researchers are helping identifying issues in your dataou with your data

More citations

- Hopefully good ones or people finding new things
- Consortia - UK Biobank

Pushing science

Upside: Why should you still share

Data Aggregation

- You have data X, I have data Y
- Can show better out-of-sample validity, or .... you can have MORE POWER:

Consent

Make sure your study documents have this in the language

- Talk to your local IRB
- Ask others
- Ask the repository
- Ask the funders

De-identification/Anonymization

LOADS of open questions - what is "identifiable"?

- HIPAA stuff obviously, but what about

Gait?

GPS?

Data also hides in things sometimes (e.g. header of gt3x)

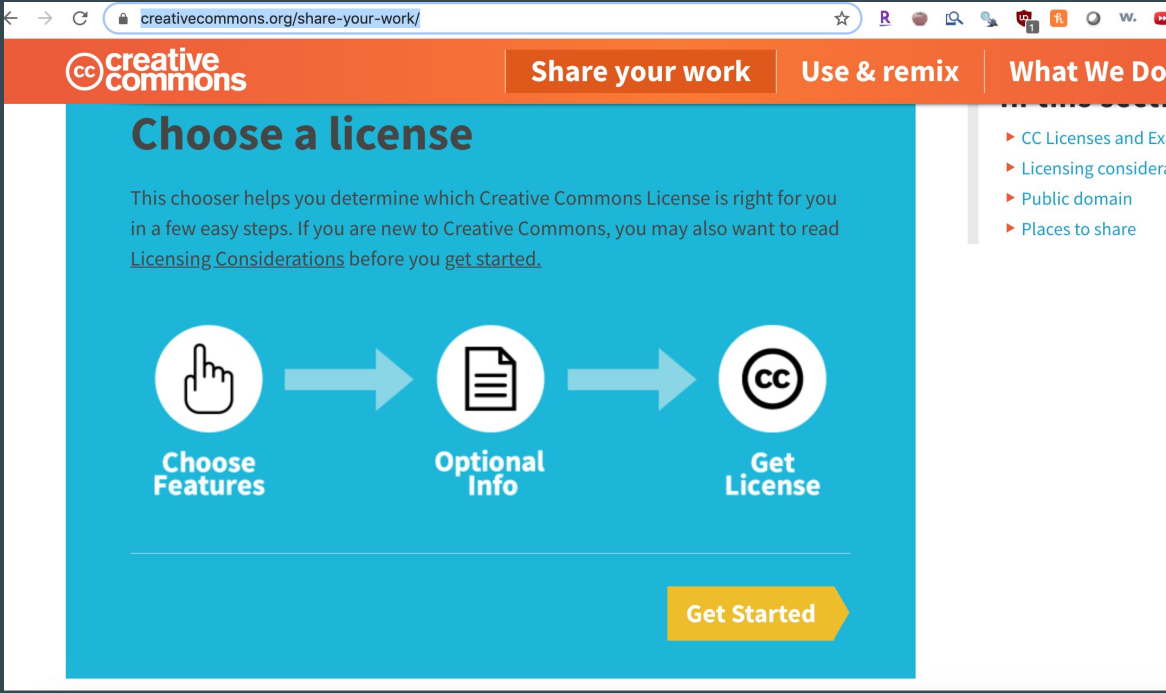
Licenses:

See https://help.data.world/hc/en-us/articles/115006114287-Common-license-types-for-datasets for help

- Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

- MIT
- Apache
- BSD
- GPL

https://creativecommons.org/share-your-work/



Choose a license

This chooser helps you determine which Creative Commons License is right for you in a few easy steps. If you are new to Creative Commons, you may also want to read Licensing Considerations before you get started.

Open Data

Upper limb activity of twenty myoelectric prosthesis users and twenty healthy anatomically intact adults.

- https://springernature.figshare.com/collections/Upper_limb_activity_of_twenty_myoelectric_prosthesis_users_and_twenty_healthy_anatomically_intact_adults_/4457855

Newcastle polysomnography and accelerometer data

- https://zenodo.org/record/1160410

Open Data: Why I care

Having open data allows us to make tutorials/workflows for:

1. Our Students
2. Students in our courses
3. Software packages
4. Collaborators
5. The rest of science

(But maybe gives away "secret sauce")

If you're releasing data, see if

- Can you read it open source
  - gt3x files (R: AGread, read.gt3x, Python: ActiGraph-ActiWave-Analysis)
  - Most other file types: GGIR (e.g. Axtivity CWA)
  - Everything can read CSV
  - Mostly everything can read HDF5 - can do ALL of these
- Data Structure/naming convention: BIDS
  - https://www.openmhealth.org/schemas/omh_acceleration/

Gifs/Images

- https://imgur.com/gallery/Cr2L8s1
- https://painepublishing.com/measurementadvisor/wp-content/uploads/sites/4/2018/03/too-much-data.jpg
- https://media1.giphy.com/media/G5X63GrrLjjVK/giphy.gif?cid=ecf05e4780a131d84268014d0b5604c95823e8fc843a62e2&rid=giphy.gif
- https://media.giphy.com/media/9LZpYKd2CmV6ICxs17/giphy.gif
- https://media.giphy.com/media/3oz8xRD1irfRHMnVD2/giphy.gif
- https://media.giphy.com/media/3orifdO6eKr9YBdOBq/giphy.gif
- https://media.giphy.com/media/l41Yr7JNU9U7upjNu/giphy.gif
- https://media.giphy.com/media/FqAwoNjVneJxK/giphy.gif
- https://media.giphy.com/media/3oKIP5F6l0hfcdOn6M/giphy.gif
- https://media.giphy.com/media/26AHPxxnSw1L9T1rW/giphy.gif
- https://media.giphy.com/media/WUCeN3kP3X3JZu52vg/giphy.gif
-
-

Thanks!