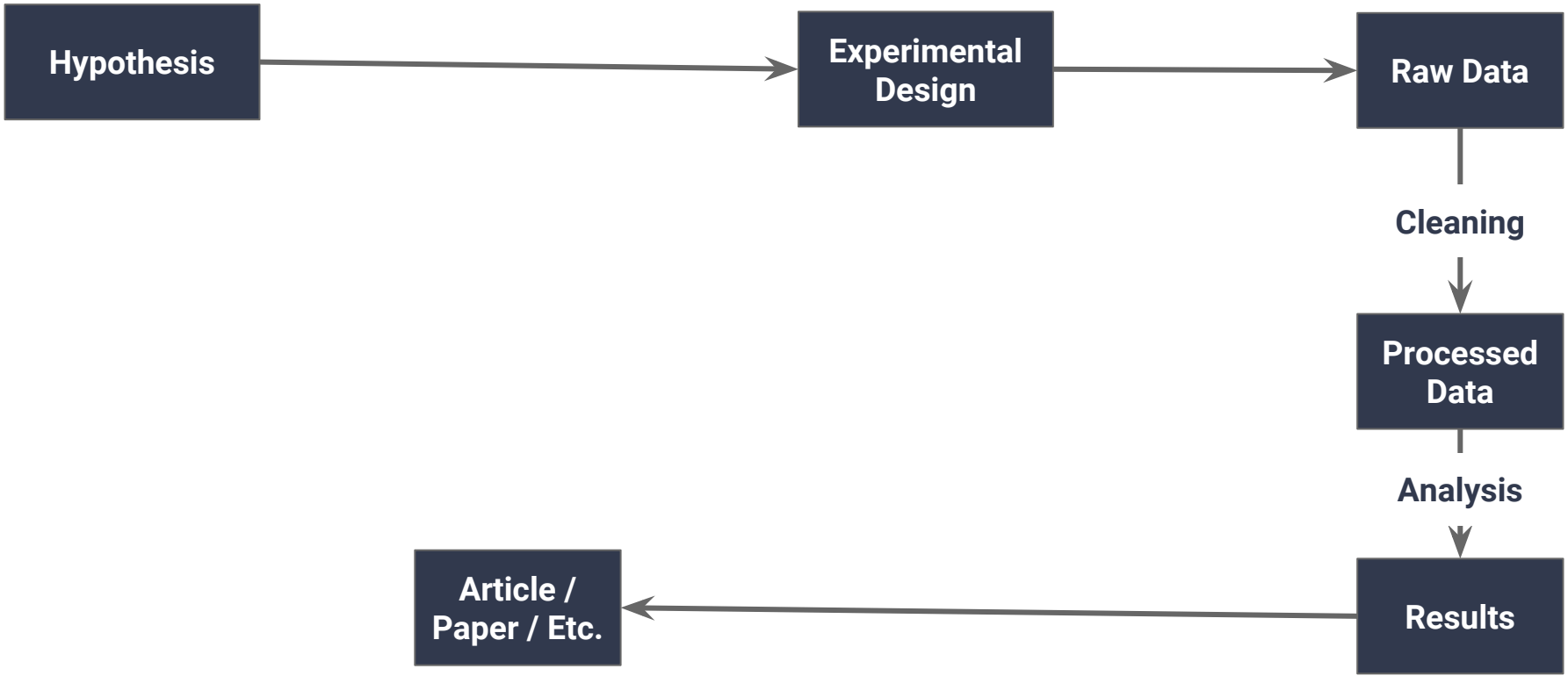
Reproducible Research:

What it is...
Why it's important...
How we can implement it

Adapted from:
https://www.slideshare.net/CTobinMagle/intro-to-reproducible-research and
https://www.slideshare.net/sahirbhatnagar/rrslides

Traditional Project Cycle



Reproducible Research:

Is the practice of distributing all data, source code, and tools required to reproduce the results discussed in a research publication.

Reproducible Research =

Data (with Metadata)

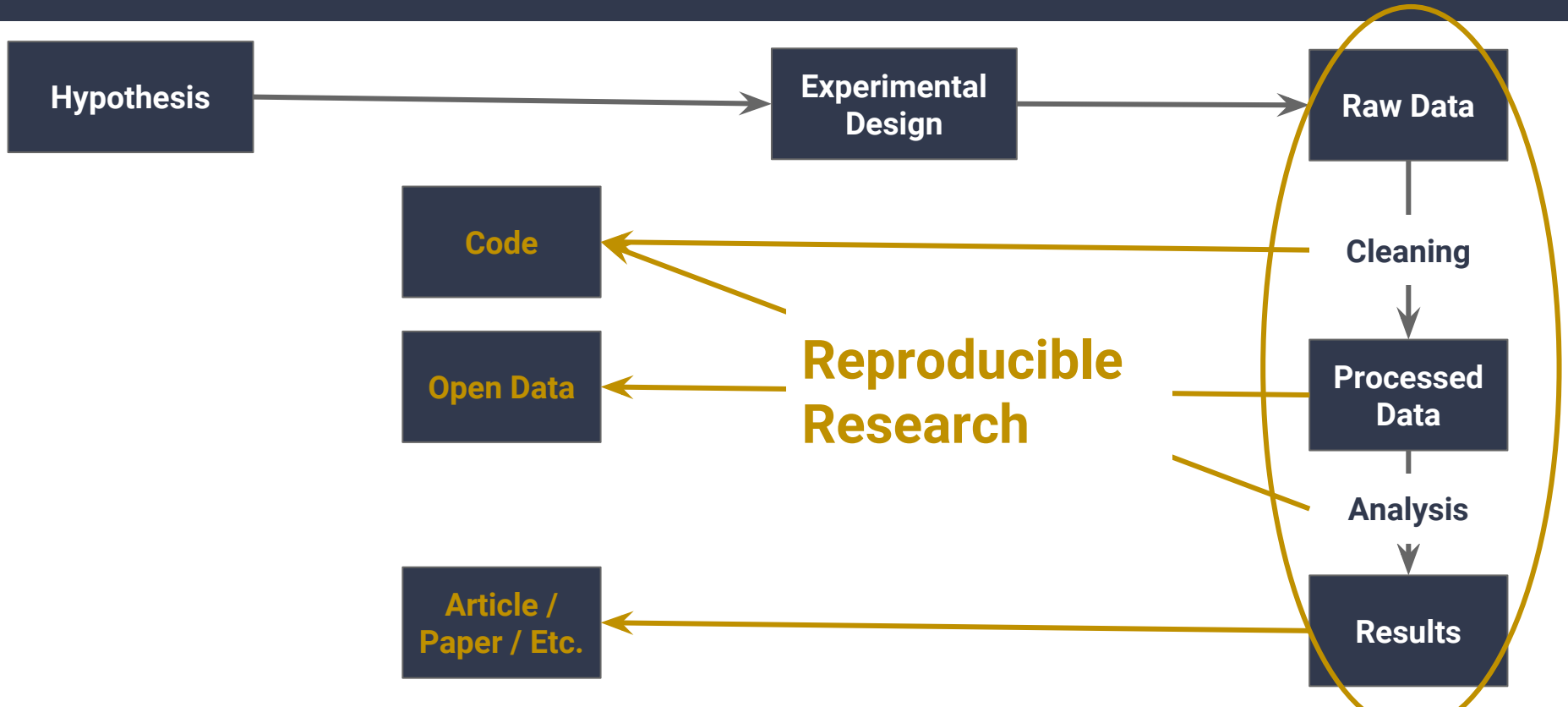
+

Code/Software

Reproducible Research =

Transparency

Reproducible Project Cycle



Why does it matter?

"Reproducibility Crisis" in Psychology

- Only 36% of replicated studies show statistically significant results.

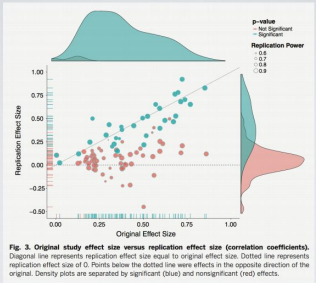


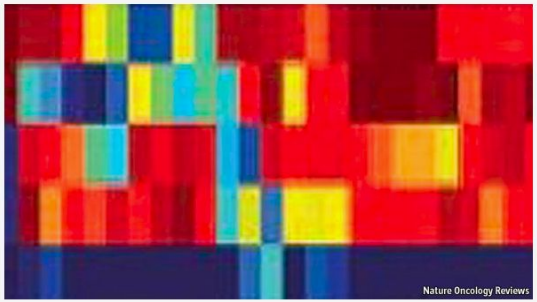
Fig. 3. Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

Nosek et al. for the Open Science Collaboration (2015, *Science*)

www.jihoonkim.org

DECEPTION AT DUKE: FRAUD IN CANCER CARE?

Were some cancer patients at Duke University given experimental treatments based on fabricated data? Scott Pelley reports.



ANIL POTTI, Joseph Nevins and their colleagues at Duke University in Durham, North Carolina, garnered widespread attention in 2006. They reported in the *New England Journal of Medicine* that they could predict the course of a patient's lung cancer using devices called expression arrays, which log the activity patterns of thousands of genes in a sample of tissue as a colourful picture (see above). A few months later, they wrote in *Nature Medicine* that they had developed a similar technique which used gene expression in laboratory cultures of cancer cells, known as cell lines, to predict which chemotherapy would be most effective for an individual patient suffering from lung, breast or ovarian cancer.

JPMorgan Discloses $2 Billion in Trading Losses

By JESSICA SILVER-GREENBERG and PETER EAVIS

Mario Tama/Getty Images

Jamie Dimon, the chief executive of JPMorgan Chase.

Figure 3: The hedging strategy operated through a series of Excel spreadsheets, which had to be completed manually, by a process of copying and pasting data from one spreadsheet to another

Why does it matter?

For Science:

- Assists with replication (when possible)
  - Findings cannot be considered genuine contributions independently verified
- Enables the cumulative growth of future scientific knowledge

For You:

- Better work habits
- Better teamwork
  - Bring current and future collaborators up to speed with ease
- Changes are easier
  - No research process is linear
- Higher research impact
  - Others more willing to read, learn, build and cite

Replication vs. Reproducibility

- **Replication**: Same conclusion new study (gold standard)

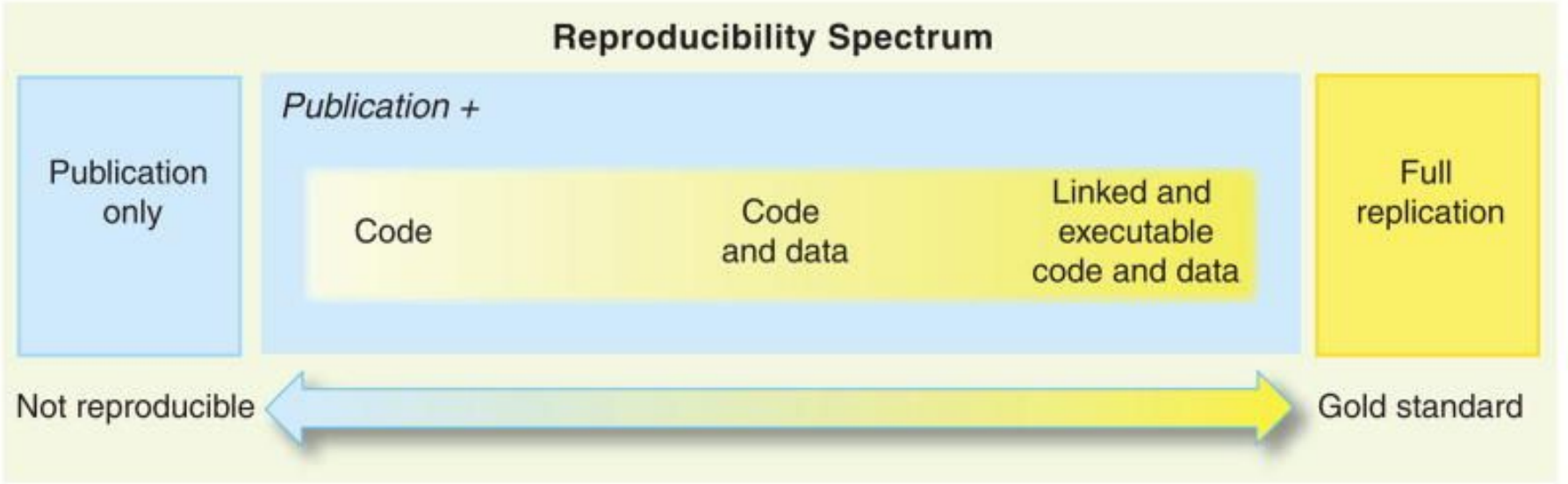
  “Again, and Again, and Again ...” BR Jasny et. al. Science, 2011. 334(6060) pp. 1225 DOI: 10.1126/science.334.6060.1225

- **Replication isn’t always feasible**: Too big, too costly, too time consuming, one time event, rare samples

- **Reproducibility**: Same results from same data and code (minimum standard for validity)

  “Reproducible Research in Computational Science”. RD Peng Science, 2011. 334 (6060) pp. 1226-1227 DOI: 10.1126/science.1213847

Reproducibility Spectrum



"Reproducible Research in Computational Science". RD Peng Science, 2011. 334 (6060) pp. 1226-1227 DOI: 10.1126/science.1213847

Documenting cleaning & analysis process

- **Optimal**: The instructions should be an automated script (i.e., “code”)
- **Minimum**: Written instructions that allow for the complete reproduction of your analysis

Additional details to document

- What software was used? (i.e., Jupyter Notebook)
- What version # and settings were used? (i.e., Python 3.6)
- What else does the software need to run?
  - Computer Architecture
  - Operating System
  - External Databases

Automate as much as possible

Doing things by hands is...

- Slow
- Difficult to document
- Hard to repeat

Automation is...

- Fast
- Simple to document
- Easy to repeat

How do we actually do it?

- **Code Control**
  - Git
  - Subversion
- **Literate Programming**
  - Jupyter Notebooks
  - Pweave: Scientific Reports using Python (http://mpastell.com/pweave/)
  - Stitch: A knitr- RMarkdown- like library, in Python (https://github.com/pystitch/stitch)
- **Data Control**
  - Quilt: Python Data Registry (https://github.com/quiltdata/quilt; https://quiltdata.com/)

Questions???

Additional Resources

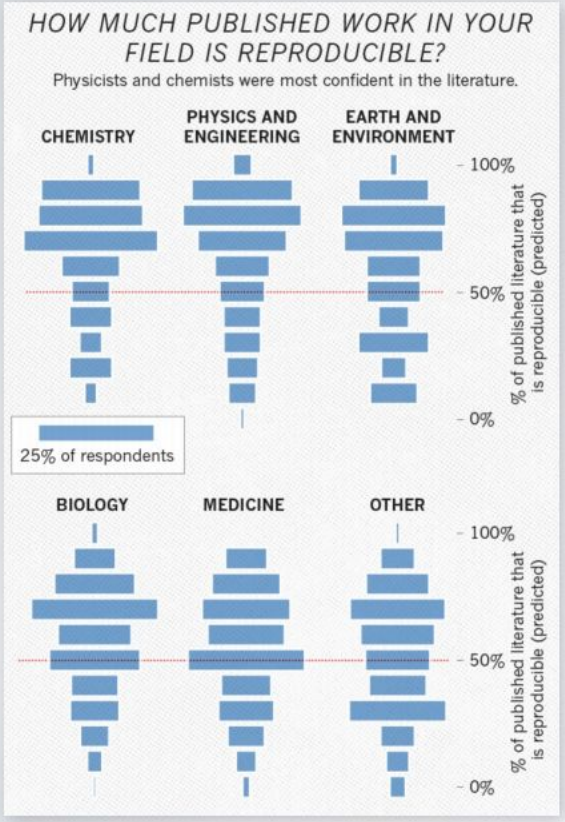
1. The Practice of Reproducible Research
   https://www.practicereproducibleresearch.org/
2. Reproducible Research in Computational Science
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3383002/
3. 1,500 scientists lift the lid on reproducibility
   https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970
4. Deception at Duke: Fraud in cancer care?
   https://www.cbsnews.com/news/deception-at-duke-fraud-in-cancer-care/2/
   https://www.economist.com/node/21528593
5. Reproducibility in Science
   http://ropensci.github.io/reproducibility-guide/

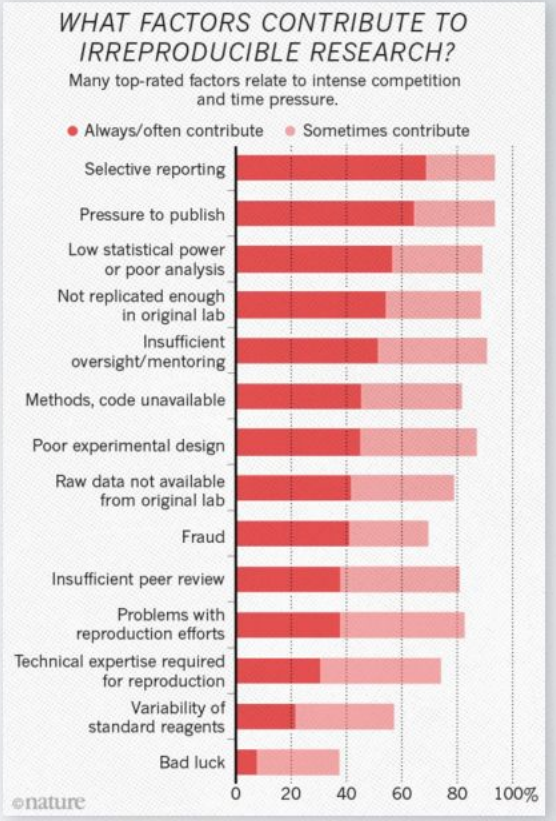
Reproducible Research Checklist

- **Think about the entire pipeline**: are all the pieces reproducible?
- **Is your cleaning/analysis process automated?**: guarantees reproducibility
  - Are you doing things “by hand”?: editing tables/figures; splitting/reformatting data
  - Does your software support log files or scripts?
  - If no, do you have a detailed description of your process?
- **Are you using version control?**
- **Are you keeping track of your software?**
  - Computer architecture;
  - OS/Software/tool/add ons (libraries/packages)/external databases
  - version numbers for everything (when available)
- **Are you saving the right files?**: if it’s not reproducible, it’s not worth saving
  - Save the data and the code
  - Data + Code = Output
- **Are your reports human and machine readable?**

Adapted from:
https://github.com/DataScienceSpecialization/courses/blob/master/05_ReproducibleResearch/Checklist/Reproducible%20Research%20Checklist.pdf

"Reproducibility Crisis" - *Nature* Magazine





Nature Survey of 1576 researchers (2016)