Be boundless

Knowledge and solutions
for a changing world

Advancing data-intensive
discovery in all fields

Reproducibility: failures & futures

David A. C. Beck

Chemical Engineering & eScience Institute

Reproducibility

- Can an experimental result be reproduced?
- Reproducibility comes in different flavors
  - Same data, same analyses (Reproducible)
  - Similar data, same analyses (Replicability)
  - Same data, similar analyses (Robustness)
  - Others?

  - Today I'll use **Reproducibility** to cover all of these

Reproducibility

- Can an experimental result be reproduced?
  - Medical science
    - Drug trial, Does a drug provide a benefit? Is it harmful?
    - Is there a genetic association with a cancer?
  - Economics
    - Is austerity the best way to get a national economy out of recession?
    - Is a 2 billion dollar industrial plant a financially sensible investment?

Reproducibility

- Can an experimental result be reproduced?
  - Social science
    - Does an in-person conversation change views on marriage equality?
  - Engineering
    - Does a waste water treatment strategy remove micro-pollutants down to a safe level?

Reproducibility

- Can an experimental result be reproduced?
  - The above examples all have data science components

Isn’t just academic science & engineering!

Reproducibility

- Can an experimental result be reproduced?
  - Marketing
    - Do loyalty programs alter buyer behavior?
    - Does removing fields from a registration form increase user completion?
    - Does a web page layout increase purchasing?
    - Sidebar:
      - To see some of how this works, check out this how to:
        - » https://webdesign.tutsplus.com/articles/split-testing-with-google-analytics-experiments--webdesign-7879
    - Other examples?

Epic fail Schadenfreude* parade

*a feeling of joy that comes from seeing or hearing about another person's troubles or failures. - Wikipedia

Epic fail

- In 2011, Bayer (pharmaceuticals) tried to replicate 67 important papers
  - Oncology
  - Women's health
  - Cardiovascular medicine

Only about 21% were reproducible

Begley, C. G.; Ellis, L. M. (2012). "Drug development: Raise standards for preclinical cancer research". *Nature* **483** (7391): 531–533.

Epic fail, part 2

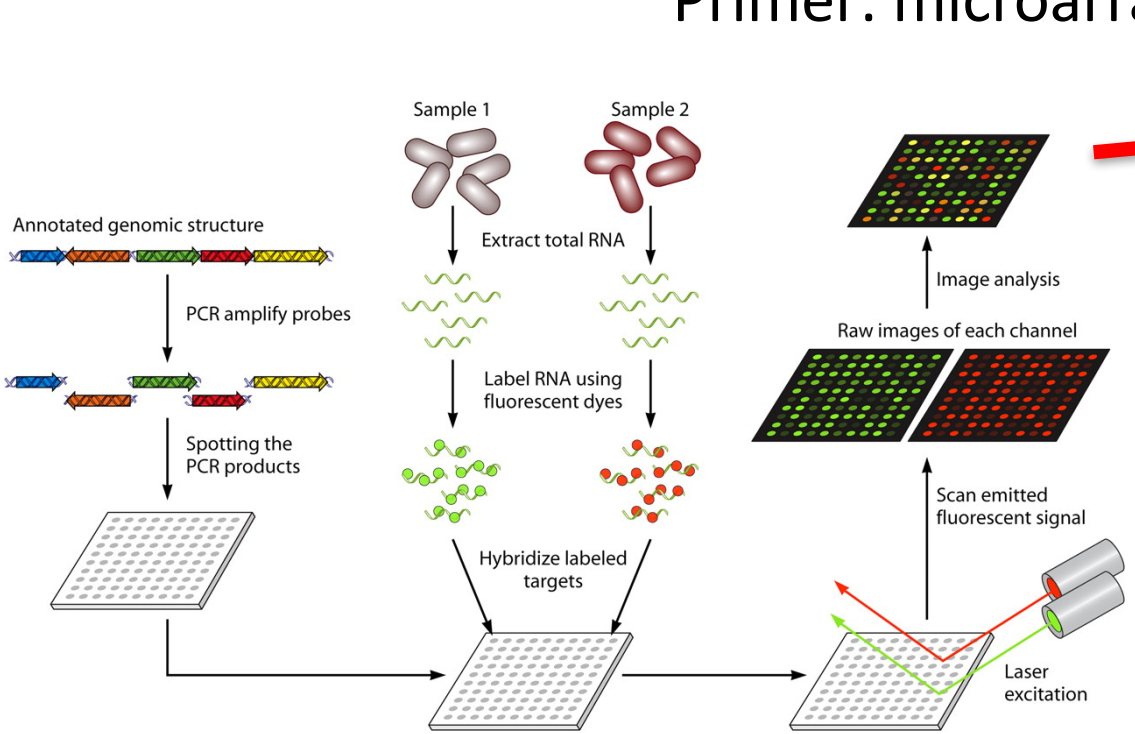
- In 2012, Amgen published a report in Nature
  - Examined 53 landmark studies in cancer

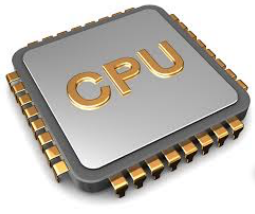
6 of 53 (11%) were reproducible

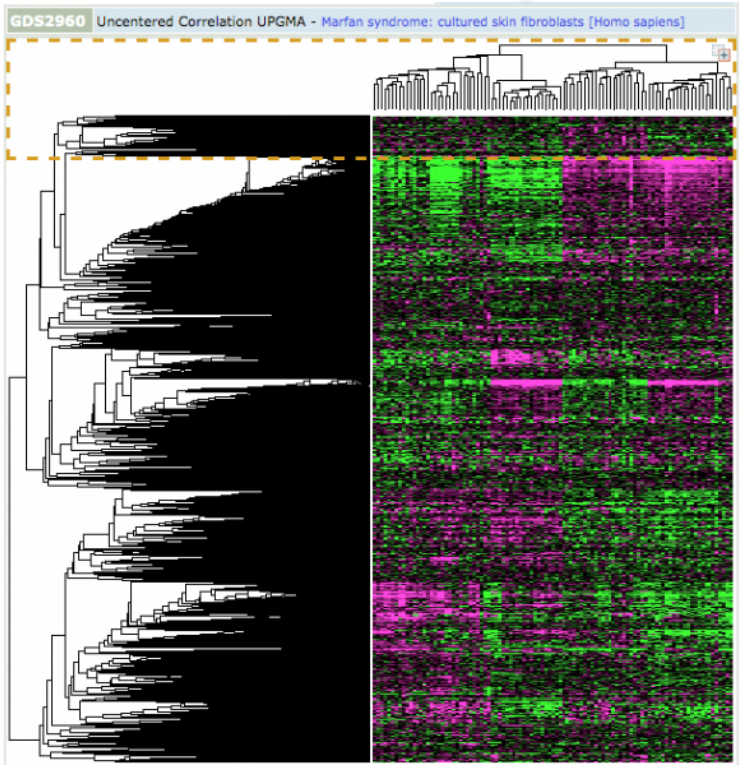
Begley, C. G.; Ellis, L. M. (2012). "Drug development: Raise standards for preclinical cancer research". *Nature* **483** (7391): 531–533.

Epic fail, part 3

Primer: microarrays





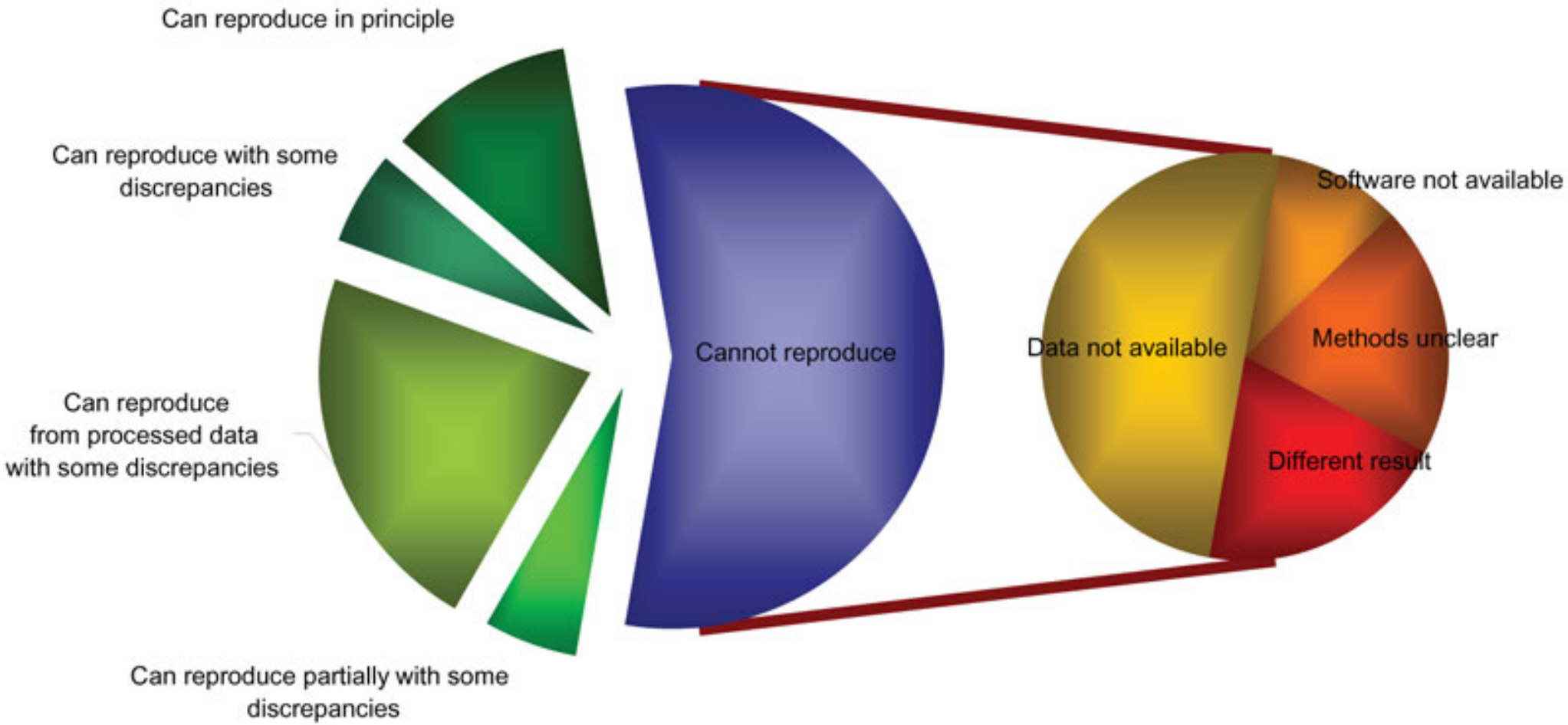


FIGHT CANCER

Miller, M. B. and Y. W. Tang (2009). "Basic concepts of microarrays and potential applications in clinical microbiology." Clin Microbiol Rev **22**(4): 611-633.

Epic fail, part 3

Attempt to reproduce 18 tables and figures papers published in Nature Genetics using microarrays



Ionnidis, P. et al. *Repeatability of published microarray gene expression analyses. Nat Gen , 41:2, Feb 2009*

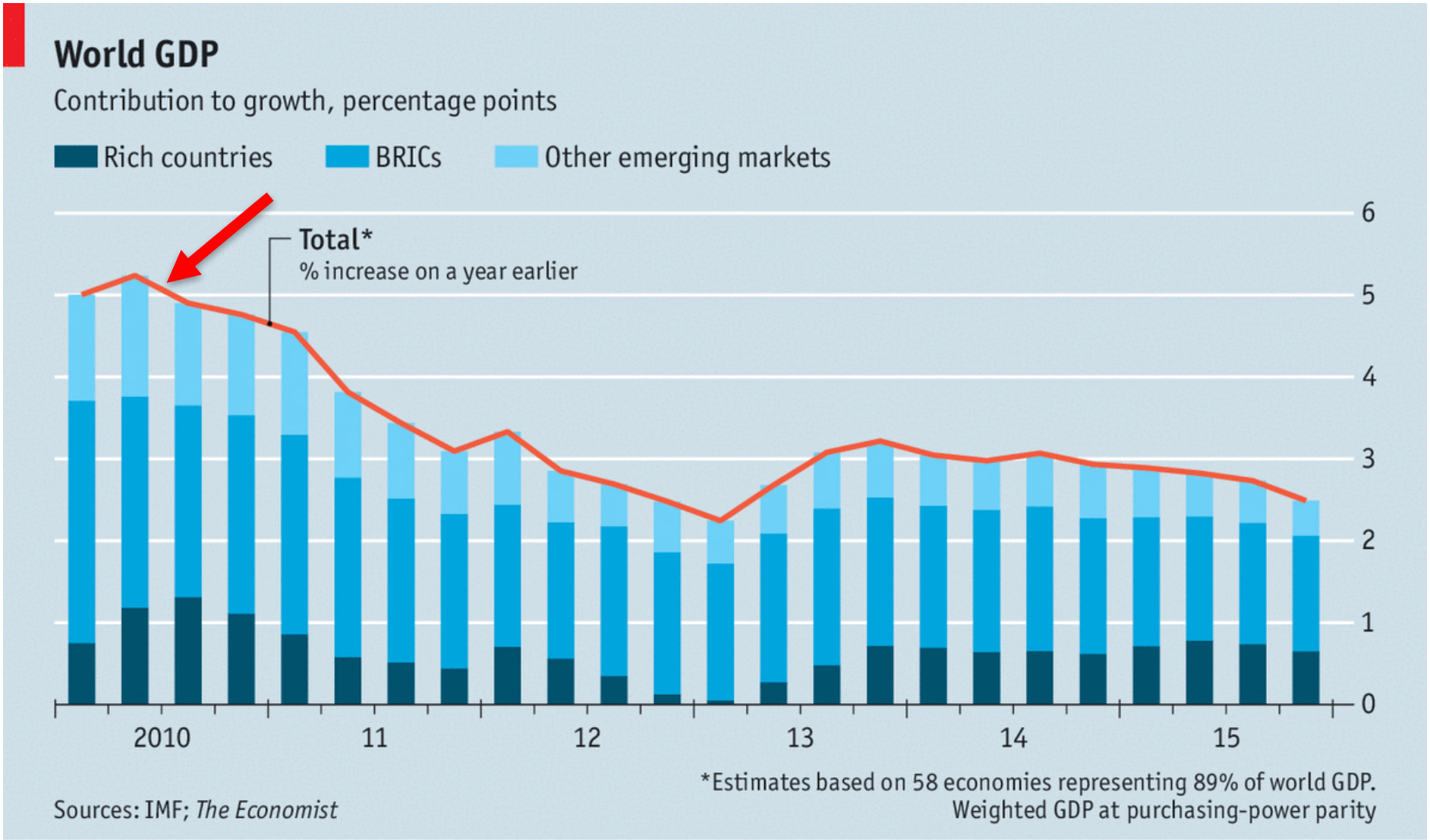
Epic fails in medicine

- What are the repercussions of irreproducible results in medicine?
  - Biotech companies
  - Government
  - People?

Epic fail, global impact

- Grab your way-back hat and put it on!

Epic fail, global impact



Epic fail, global impact

- 2010 paper by Reinhart & Rogoff “Growth in a Time of Debt”
  - **…high debt/GDP levels (90 percent and above) are associated with notably lower growth outcomes.**
  - **Debt to GDP ratios over 90% have read GDP growth of -0.1%**
  - **Seldom do countries “grow” their way out of debts.**

Reinhart, Carmen M., and Kenneth S. Rogoff. 2010. "Growth in a Time of Debt." *American Economic Review*, 100(2): 573-78.

Epic fail, global impact

- Paper was widely cited by
  - Political parties
  - Governments
  - International lending agencies
- To show that **austerity** was the solution to the global recession
- Even part of the 2012 US presidential election!

Reinhart, Carmen M., and Kenneth S. Rogoff. 2010. "Growth in a Time of Debt." *American Economic Review*, 100(2): 573-78.

Epic fail, global impact

- UMass Amherst Graduate student Thomas Herndon
  - Tried to reproduce the results of the paper for a class: **couldn't**
  - Requested the 'code' for the computations from R&R: got an Excel spreadsheet
  - Found multiple errors

Reinhart, Carmen M., and Kenneth S. Rogoff. 2010. "Growth in a Time of Debt." *American Economic Review*, 100(2): 573-78.
Thomas Herndon, Michael Ash & Robert Pollin, Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff

Epic fail, global impact

- UMass Amherst Graduate student Thomas Herndon
  - Found multiple errors

Coding errors, selective exclusion of available data, and unconventional weighting of summary statistics lead to serious errors that inaccurately represent the relationship between public debt and GDP growth.

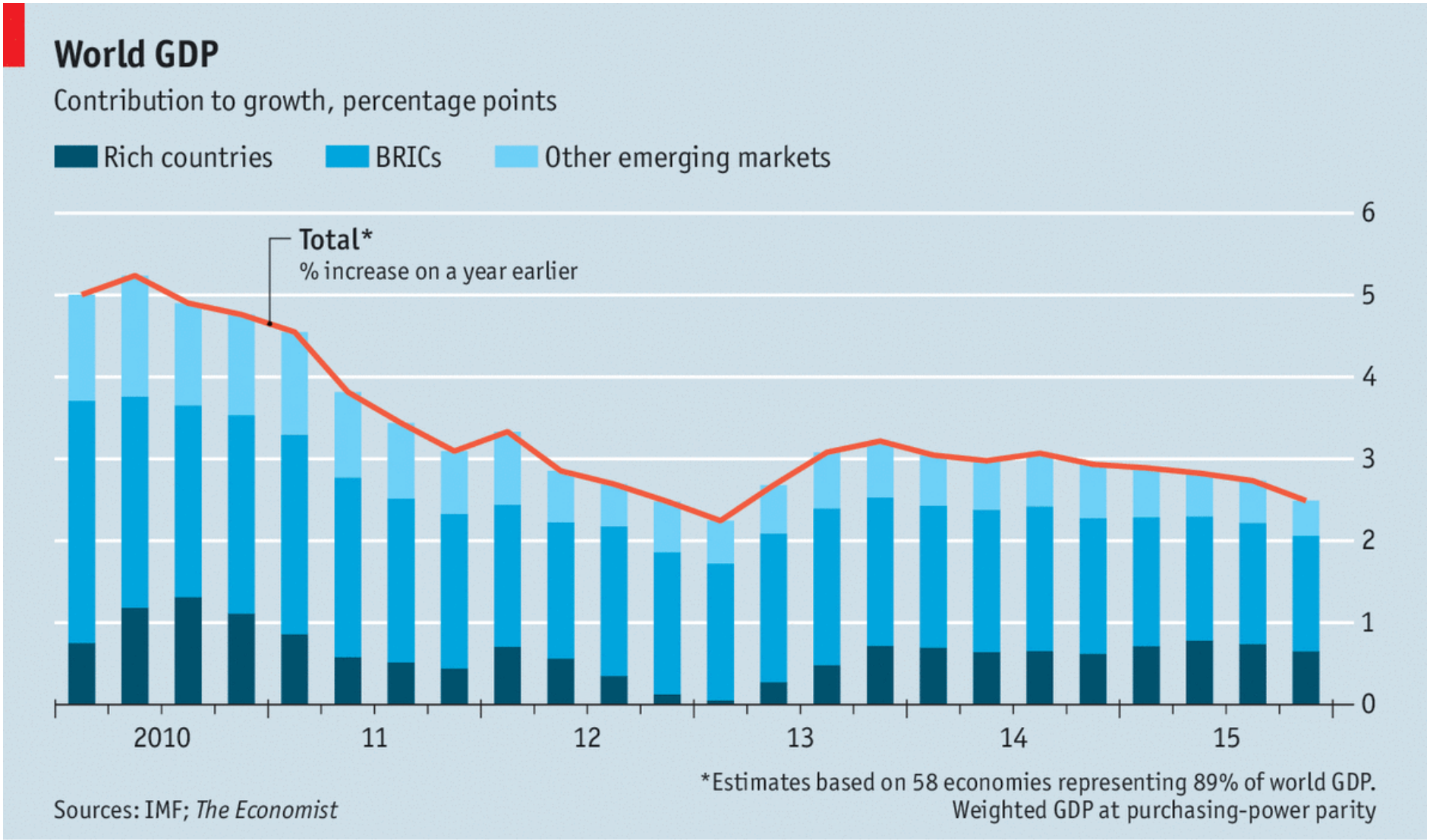
Reinhart, Carmen M., and Kenneth S. Rogoff. 2010. "Growth in a Time of Debt." *American Economic Review*, 100(2): 573-78.
Thomas Herndon, Michael Ash & Robert Pollin, Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff

Epic fail, global impact

- Herndon fixed the errors and reexamined claims
- Original claims
  - Debt to GDP ratios over 90% have real GDP growth of **-0.1%**
  - In a recession: Austerity good, spending bad
- Modified claims
  - Debt to GDP ratios over 90% have real GDP growth of **2.2%**
  - In a recession: Spending good

Reinhart, Carmen M., and Kenneth S. Rogoff. 2010. "Growth in a Time of Debt." *American Economic Review*, 100(2): 573-78.
Thomas Herndon, Michael Ash & Robert Pollin, Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff

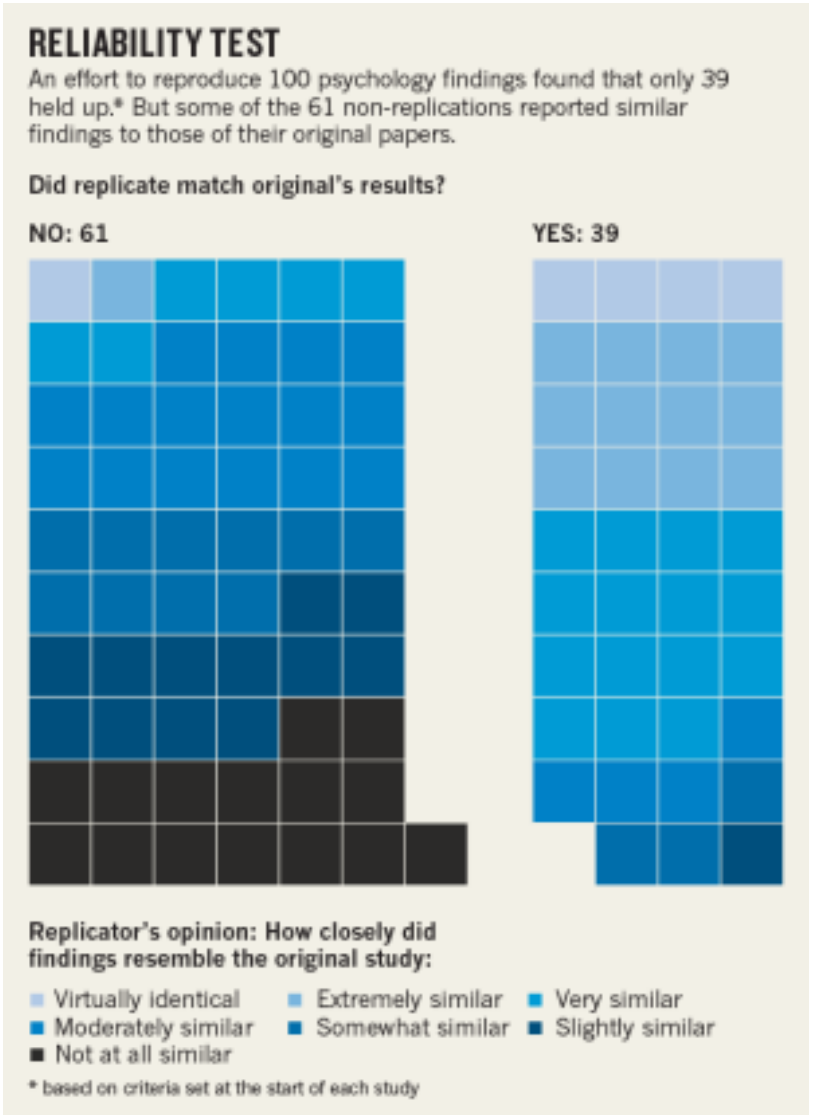
Epic fail, global impact



Epic fail, global impact

- What effect did the incorrect R&R paper have?

Epic failure, part 4



http://www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248

Reproducibility

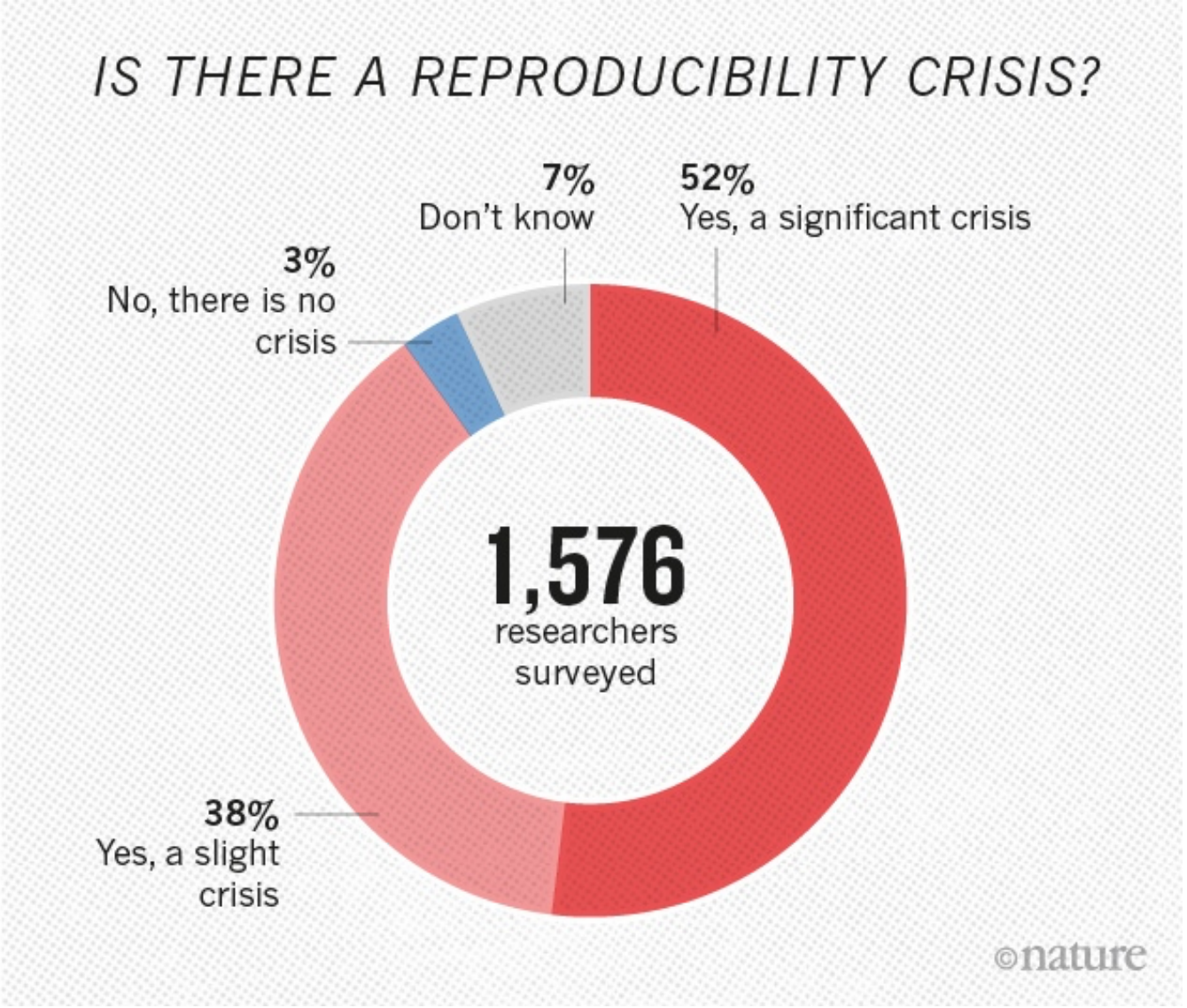
- Why do we care?

"Non-reproducible single occurrences are of no significance to science."

– Karl Popper

Popper, K. R. 1959. The logic of scientific discovery. Hutchinson, London, United Kingdom.

Science in crisis?



Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452-454 (2016).

Reproducibility: Things are bad

Why is this happening?

- Social factors, e.g.
  - Fraud, misconduct
  - Pressure to publish
- *p*-hacking
- Poor experimental design
  - Small effect size
  - Small sample size
- Data not disclosed
- Methods not disclosed or properly described
  - Software not available

Important but not Data Science related.
WE ARE WORKING ON THESE!

p-hacking

- Do a study to test some hypothesis
  - E.g. an apple a day keeps the Dr. away
- Use a *p*-value of 0.05
  - i.e. 5% chance of seeing a difference at least as big as we have, by chance alone
- Perform 1000s of statistical tests
- What happens?

~50 significant results by chance alone

1. Simmons, J.P., N.D. Nelson, and U. Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22(11):1359-1366.

p-hacking

- Test very large number of hypothesis on a data set searching for any statistically significant effect
- Goes by many names in different disciplines
  - Multiple comparisons (1950s, most statisticians),
  - File drawer problem (Rosenthal, 1979),
  - Significance questing (Rothman and Boice, 1979),
  - Data mining, dredging, torturing (Mills, 1993),
  - Data snooping (White, 2000),
  - Selective outcome reporting (Chan et al., 2004),
  - Bias (Ioannidis, 2005),
  - Hidden multiplicity (Berry, 2007),
  - Specification searching (Leamer, 1978), and
  - p-hacking (Simmons et al., 2011).

https://www.nap.edu/read/21915/chapter/4#43

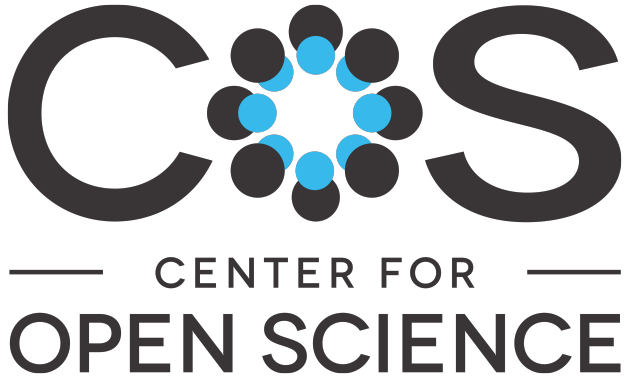
p-hacking

- Is this intentionally evil?
- Why isn't it misconduct?

- My opinion:
  - Most times, probably not
  - Reflects lack of understanding about hypothesis testing

p-hacking

- What is being done about it?
  - Register the study beforehand “Preregistration”
  - Let everyone know what the precise hypothesis being tested before data are collected



  - Get free from the tyranny of the *p*-value
  - Better statistics education

Poor experimental design

- Want to test toxicity of my new fluorescent brown dye

Poor experimental design

- Want to test toxicity of my new fluorescent brown dye
  - Feed some to 10 people
  - Watch how long they live

10 subjects, day 0

Poor experimental design

- What are some problems with this experimental design?
  - Control group?

10 subjects, no dye

Similar demographics

WHAT DO YOU MEAN YOU FORGOT THE CONTROL?

Poor experimental design

- Is it toxic?

*Average lifespan in us is 78 years
with a standard deviation of 15 years

10 subjects, day 0

10 subjects, day 1

Poor experimental design

- Is it toxic?

*Average lifespan in us is 78 years
with a standard deviation of 15 years

10 subjects, day 0

10 subjects, 50 years

Poor experimental design

- Is it toxic?

*Average lifespan in us is 78 years
with a standard deviation of 15 years

10 subjects, day 0

10 subjects, 50 years

Poor experimental design

- What are some problems with this experimental design?
  - What is the effect size you want to be able to measure? E.g. how many years difference?
  - What is the sample size required to see that effect?

- Small sample can see an effect due to chance
  - Won't be reproducible!

Poor experimental design

- What is being done about it?
  - Better statistics education
  - Replicate significant results with small effect size with way more samples

Data disclosure

- Data unavailable
  - Lost or destroyed
  - Streaming data too big to store
- Raw data not kept, only processed
- Data intentionally not shared
  - By law (FERPA, HIPPA)
  - Corporate data (e.g. twitter, JSTOR)
  - Some jerk just won't share

Data disclosure

- Data unavailable
  - Lost or destroyed
  - Streaming data too big to store
- Raw data not kept, only processed
- Data intentionally not shared
  - By law (FERPA, HIPPA)
  - Corporate data (e.g. twitter, JSTOR)
  - Some jerk just won’t share

Data disclosure

- What is being done about it?
  - Federal funding agencies now require data sharing
  - Science journals require open data
  - Deposit raw data as soon as collected
    - Similar to preregistration
  - Open data badges for researchers
  - Data sharing repositories
    - National Center for Biotechnology Information
    - Dryad (20GB limit, $100/10GB beyond)

Methods

- Poorly written methods
  - Steps missing
- Intentional methods omissions
  - To protect a monopoly on an experimental procedure
- The fix:
  - Better peer review in science
  - Better communication skills education in business

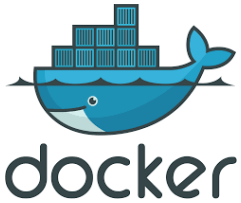
Software

- **Software unavailable**
  - Why?

- What are some other other software issues?
  - Un-runnable, i.e. broken
  - Not documented
  - Dependencies not known or given
  - Hardware constraints

Software

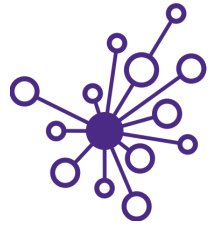
- What is being done about it?
  - Use open source software
  - Virtual environments
    - Use something that can **FREEZE** the state of the software and hardware
    - Docker images
    - Amazon Machine Images (AMI)
    - Virtual machines generally
  - Educating scientists in software engineering
    - Version control, documentation, testing, …



Resources

- eScience Institute Reproducbility Group
  - http://uwescience.github.io/reproducible/
- Berkeley Institute for Data Science Repro Stuff
  - https://bids.berkeley.edu/working-groups/reproducibility-and-open-science
- Center for Open Science
  - https://cos.io
- Coursera from JHU
  - https://www.coursera.org/learn/reproducible-research
- Other links in this presentation

Thank you!



- See you next week for last seminar!

- CSE 491 folks:
  - Don’t forget to take the quiz!
  - Don’t forget to take the quiz!
  - Don’t forget to take the quiz!
  - Don’t forget to take the quiz!
  - Don’t forget to take the quiz!
  - Don’t forget to take the quiz!