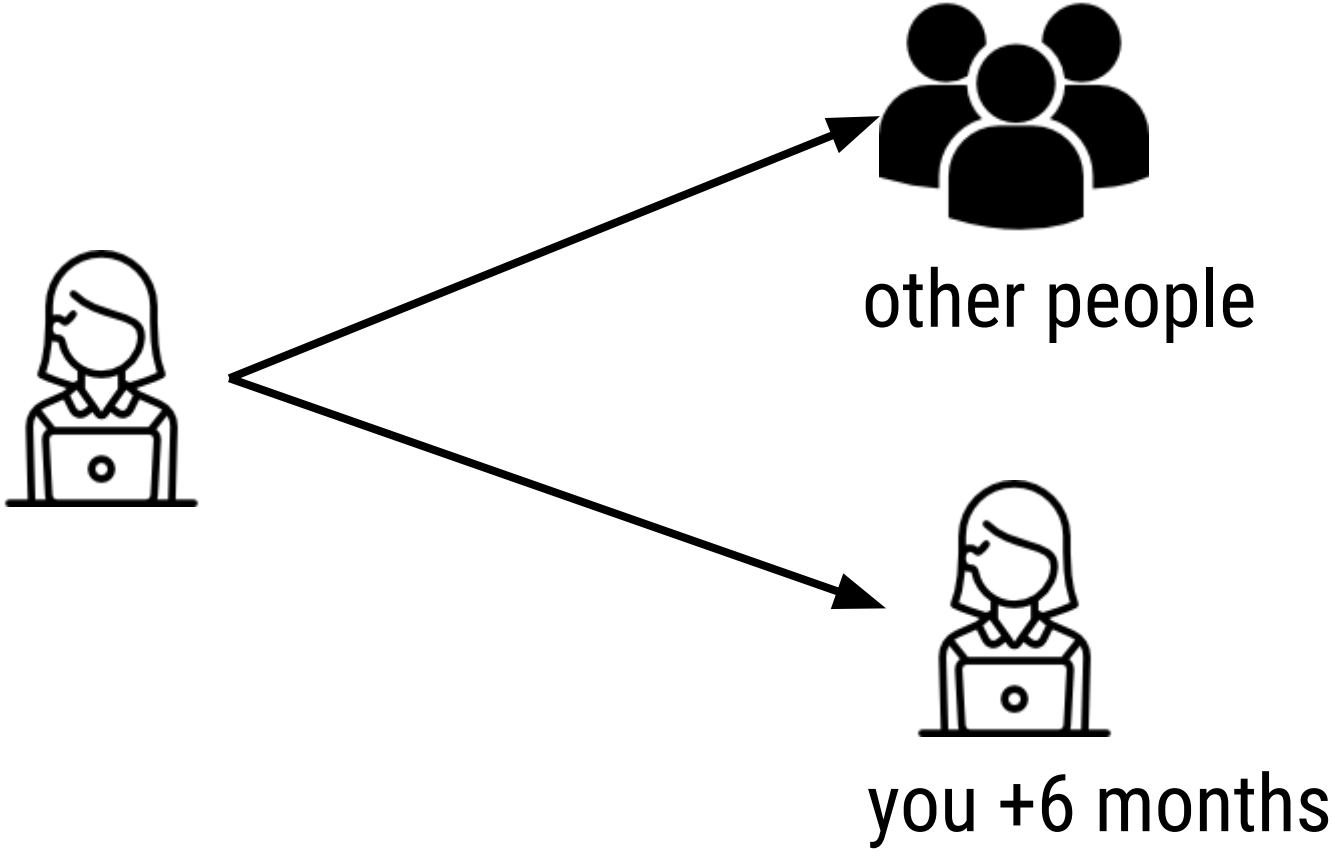
Organizing Projects

Organizing Data Science Projects

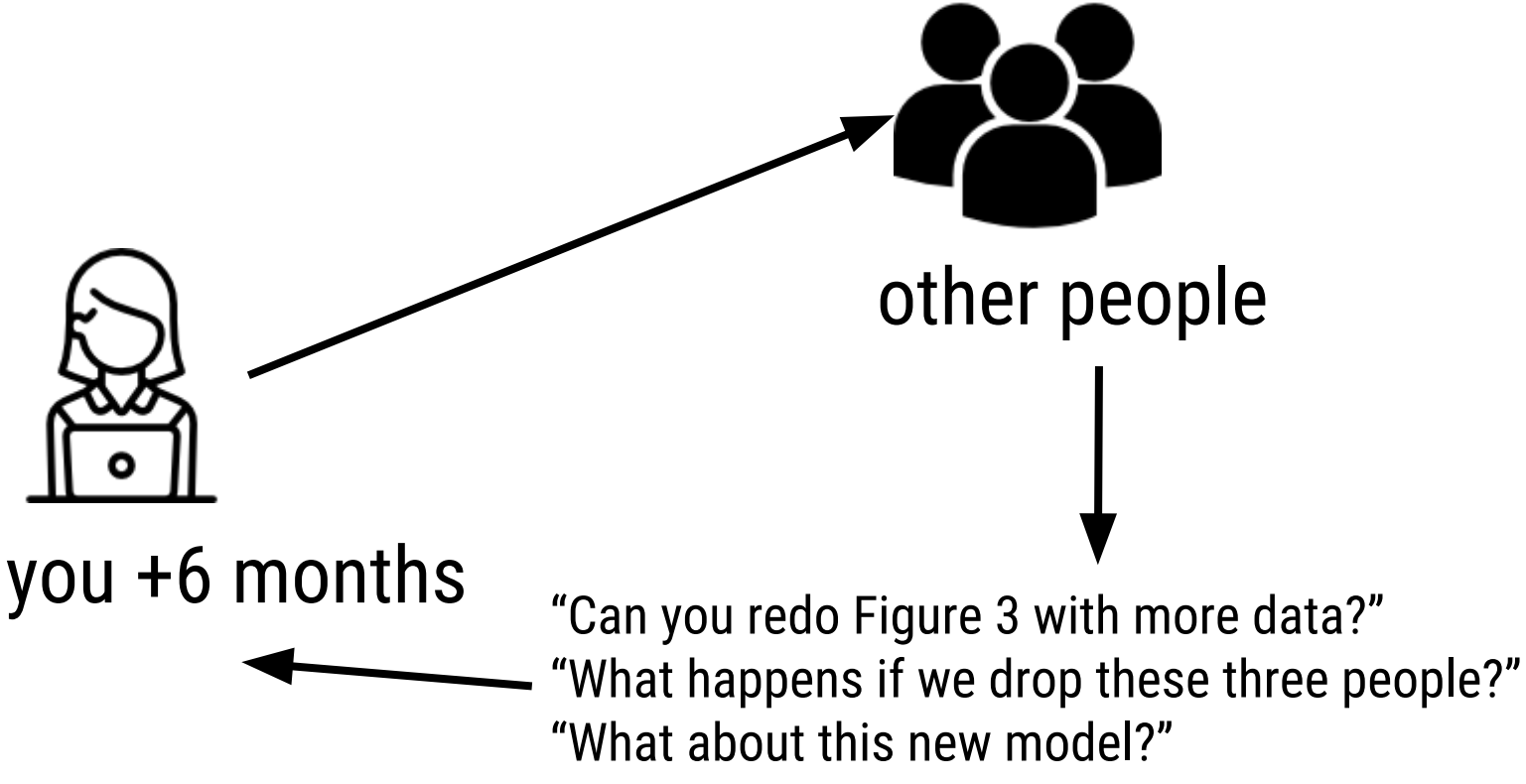
Data Science Files Should Be...

- Easy to find
- Easy to share
- Easy to understand
- Easy to update

Your audience



Why you are your own audience



Article

Genomic signatures to guide the use of chemotherapeutics

Anil Potti, Holly K Dressman, Andrea Bild, Richard F Riedel, Gina Chan, Robyn Sayer, Janiel Cragun, Hope Cottrill, Michael J Kelley, Rebecca Petersen, David Harpole, Jeffrey Marks, Andrew Berchuck, Geoffrey S Ginsburg, Phillip Febbo, Johnathan Lancaster & Joseph R Nevins ✉

Nature Medicine **12**, 1294–1300 (2006)
doi:10.1038/nm1491
Download Citation

Received: 11 July 2006
Accepted: 12 September 2006
Published: 22 October 2006

Corrigendum: 01 November 2007
Corrigendum: 01 August 2008
RETRACTED: 07 January 2011
Corrected online 27 October 2006
Corrected online 21 July 2008
Retracted online 07 January 2011

Associated Content

Nature Medicine | Correspondence
Microarrays: retracing steps
Kevin R Coombes, Jing Wang & Keith A Baggerly

Sections Figures References

Abstract
Change history
Accessions
References
Acknowledgements
Author information
Supplementary information

DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

By Keith A. Baggerly* and Kevin R. Coombes†

U.T. M.D. Anderson Cancer Center

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report, we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

From the article:
Cancer trial errors revealed

2006 Anil Potti, a cancer geneticist at Duke University in Durham, North Carolina, and others file patent applications on the idea of using gene-expression data to predict sensitivity to cancer drugs. Potti is first author on a paper in *Nature Medicine*[1].

2007 Potti is last author on a paper in the *Journal of Clinical Oncology* (*JCO*)[2]. Duke begins three clinical trials to test Potti's predictors in patients with breast or lung cancer.

SEPTEMBER 2009 Keith Baggerly and Kevin Coombes, statisticians at the University of Texas M. D. Anderson Cancer Centre in Houston, publish a paper in *Annals of Applied Statistics*[3] stating that they could not replicate Potti's claims. Duke suspends the trials and asks a review panel to investigate.

NOVEMBER 2009 Potti places data underlying the *JCO* paper online. Baggerly writes to Sally Kornbluth, Duke vice-dean for research, and Michael Cuffe, Duke vice-president for medical affairs, to point out differences from raw data.

DECEMBER 2009 An unredacted copy of the report by Duke's review panel, later obtained by *Nature*, shows that the panel replicated Potti's claims using his data, but were unaware that those data contained discrepancies.

JANUARY 2010 Duke restarts clinical trials.

JULY 2010 *The Cancer Letter* reveals that Potti made false claims about his CV. Trials are suspended and an investigation begins. Harold Varmus, director of the National Cancer Institute in Bethesda, Maryland, asks the Institute of Medicine to review Duke's trials.

NOVEMBER 2010 *JCO* paper is retracted. Duke closes the trials permanently. Potti resigns.

DECEMBER 2010 Institute of Medicine study begins, but will now focus more generally on criteria for genomics predictor.

JANUARY 2011 *Nature Medicine* paper is retracted.

American Economic Review: Papers & Proceedings 100 (May 2010): 573–578
http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.2.573

Growth in a Time of Debt

By CARMEN M. REINHART AND KENNETH S. ROGOFF*

In this paper, we exploit a new multi-country historical dataset on public (government) debt to search for a systemic relationship between high public debt levels, growth and inflation.[1] Our main result is that whereas the link between growth and debt seems relatively weak at "normal" debt levels, median growth rates for countries with public debt over roughly 90 percent of GDP are about one percent lower than otherwise; average (mean) growth rates are several percent lower. Surprisingly, the relationship between public debt and growth is remarkably similar across emerging markets and advanced economies. This is not the case for inflation. We find no systematic relationship between high debt levels and inflation for advanced economies as a group (albeit with individual country exceptions including the United States). By contrast, in emerging market countries, high public debt levels coincide with higher inflation.

Our topic would seem to be a timely one, especially against the backdrop of graying populations and rising social insurance costs? Are sharply elevated public debts ultimately a manageable policy challenge?

Our approach here is decidedly empirical, taking advantage of a broad new historical dataset on public debt (in particular, central government debt) first presented in Carmen M. Reinhart and Kenneth S. Rogoff (2008, 2009b). Prior to this dataset, it was exceedingly difficult to get more than two or three decades of public debt data even for many rich countries, and virtually impossible for most emerging markets. Our results incorporate data on 44 countries spanning about 200 years. Taken together, the data incorporate over 3,700 annual observations covering a wide range of political systems, institutions, exchange rate and monetary arrangements, and historic circumstances.

We also employ more recent data on external debt, including debt owed both by governments

http://www.cc.com/video-clips/dcyvro/the-colbert-report-austerity-s-spreadsheet-error

"File organization and naming are powerful weapons against chaos."

- Jenny Bryan

How you should approach organization

- **Step 1** slow down & make lots of notes for yourself.
- **Step 2** have sympathy for your future self.
- **Step 3** have a standard system that you understand