Automated and Reproducible Data Analytics

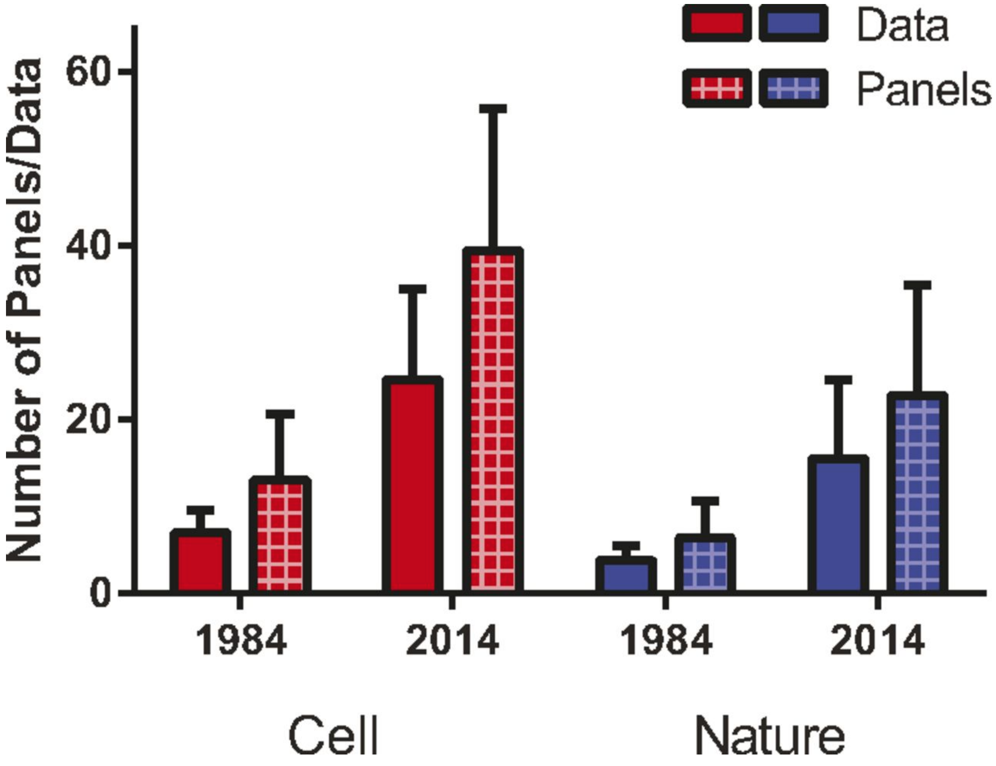
Tools and services for the life sciences and beyond.

Lars Malmstroem

More data per paper

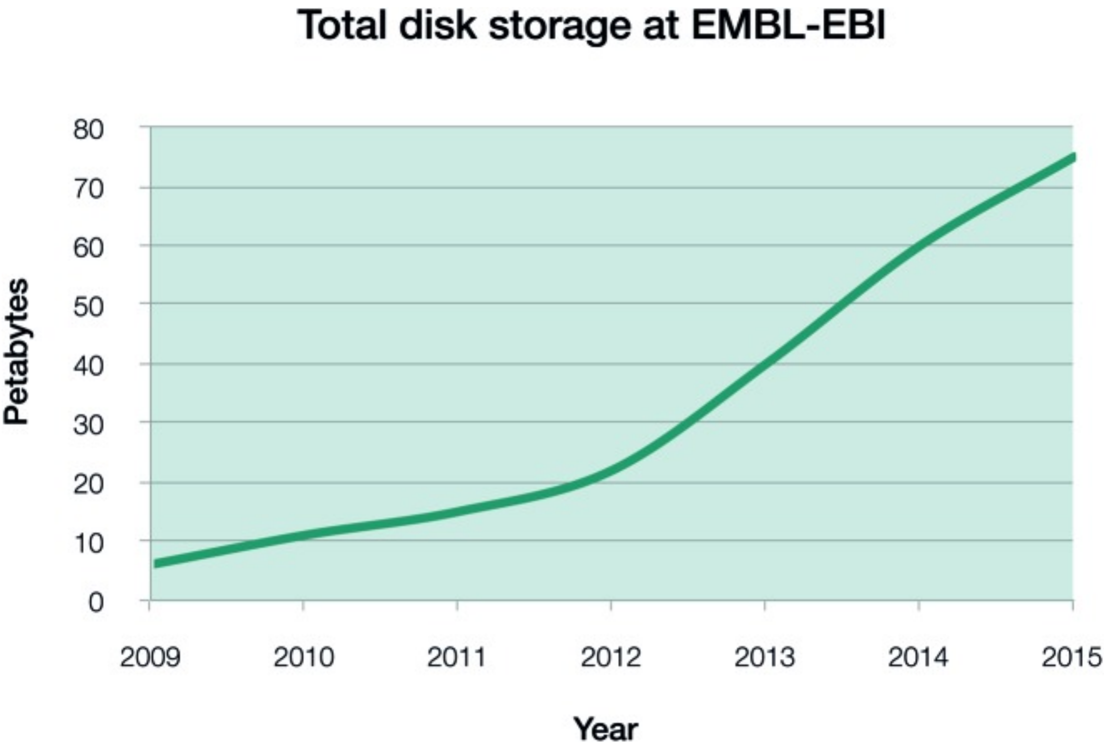
Estimated amount of data supporting each paper and the total number of figure panels in each paper

- Accelerating scientific publication in biology. Ronald D. Vale PNAS November 3, 2015. 112 (44) 13439-13446



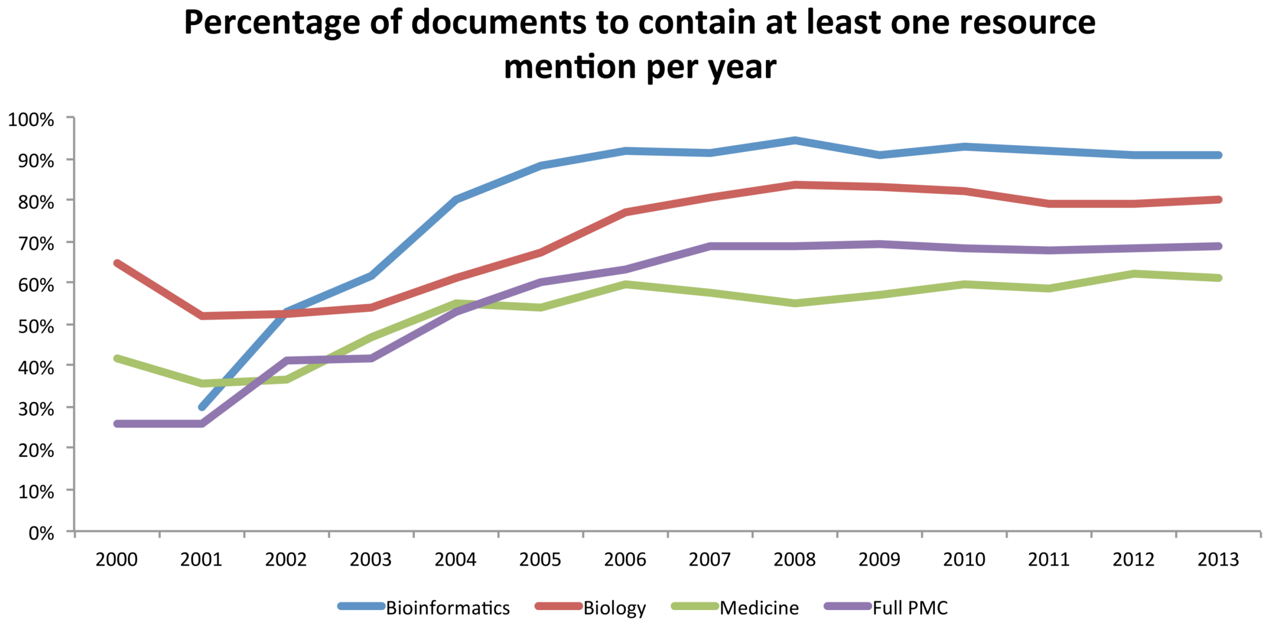
Many sciences, including biology, is changing and becoming more data intensive

- Nucleic Acids Res. 2016 Jan 4; 44(Database issue): D20–D26. The European Bioinformatics Institute in 2016: Data growth and integration Charles E. Cook, Mary Todd Bergman,* Robert D. Finn, Guy Cochrane, Ewan Birney, and Rolf Apweiler



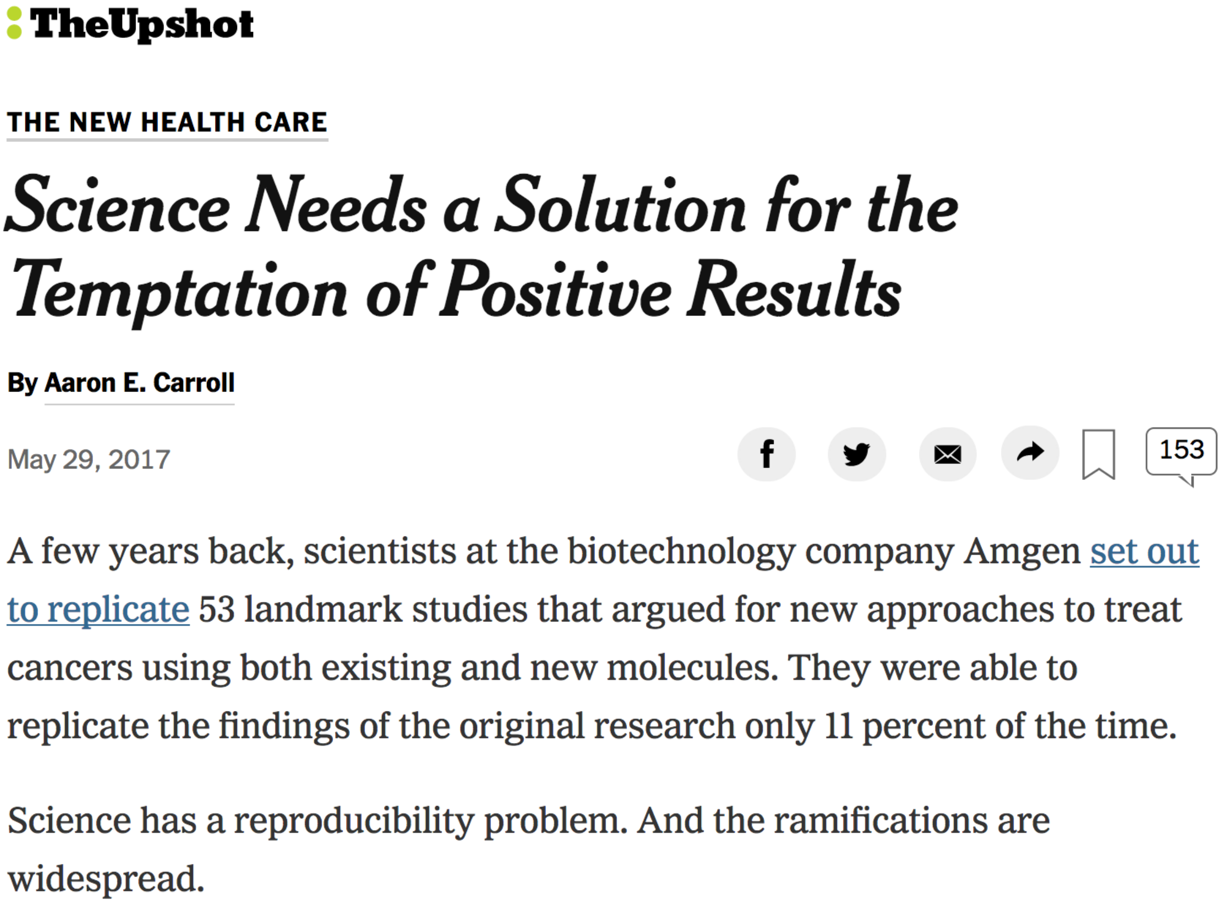
More articles are using bioinformatics or data resources

- Duck G, Nenadic G, Filannino M, Brass A, Robertson DL, Stevens R (2016) A Survey of Bioinformatics Database and Software Usage through Mining the Literature. PLoS ONE 11(6): e0157989. https://doi.org/10.1371/journal.pone.0157989 (https://doi.org/10.1371/journal.pone.0157989)



We are unfortunately not handling this situation well...

- https://www.nytimes.com/2017/05/29/upshot/science-needs-a-solution-for-the-temptation-of-positive-results.html (https://www.nytimes.com/2017/05/29/upshot/science-needs-a-solution-for-the-temptation-of-positive-results.html)



TheUpshot

THE NEW HEALTH CARE

Science Needs a Solution for the Temptation of Positive Results

By Aaron E. Carroll

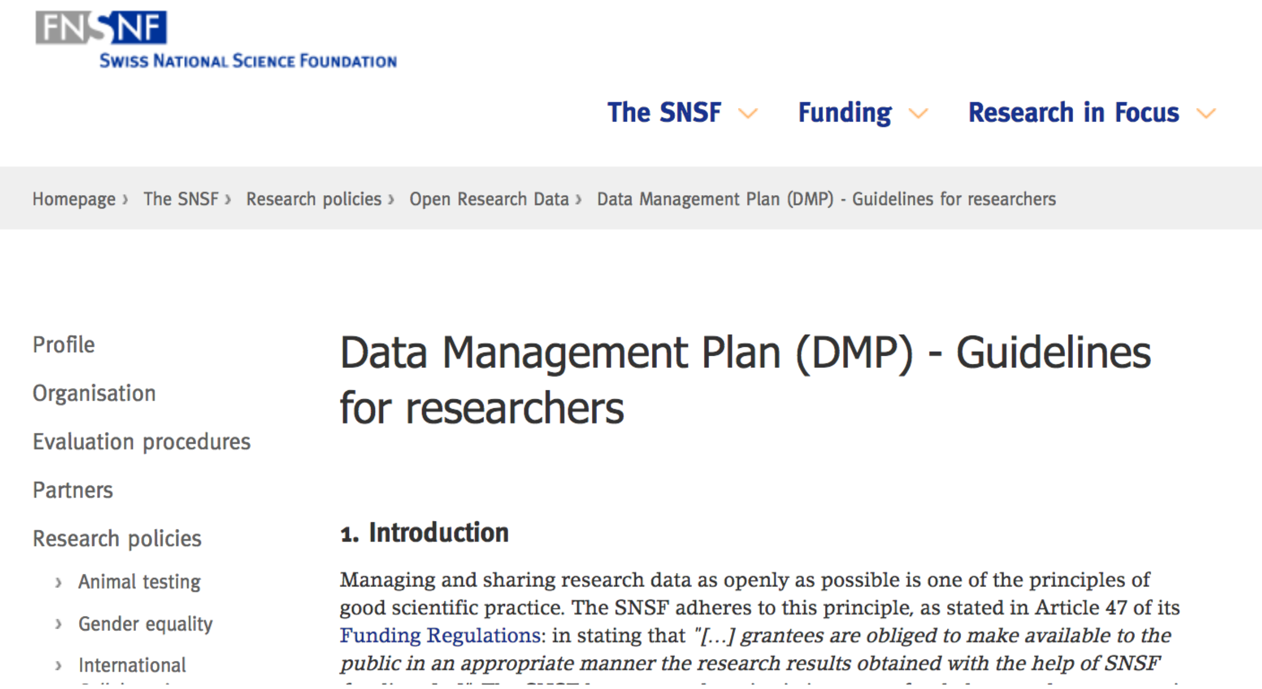
May 29, 2017

A few years back, scientists at the biotechnology company Amgen set out to replicate 53 landmark studies that argued for new approaches to treat cancers using both existing and new molecules. They were able to replicate the findings of the original research only 11 percent of the time.

Science has a reproducibility problem. And the ramifications are widespread.

Not explicity linked, but might be two sides of the same coin

- http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/data-management-plan-dmp-guidelines-for-researchers.aspx (http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/data-management-plan-dmp-guidelines-for-researchers.aspx)



FNSNF
Swiss National Science Foundation

The SNSF Funding Research in Focus

Homepage › The SNSF › Research policies › Open Research Data › Data Management Plan (DMP) - Guidelines for researchers

Profile
Organisation
Evaluation procedures
Partners
Research policies
› Animal testing
› Gender equality
› International

Data Management Plan (DMP) - Guidelines for researchers

1. Introduction

Managing and sharing research data as openly as possible is one of the principles of good scientific practice. The SNSF adheres to this principle, as stated in Article 47 of its Funding Regulations: in stating that *"[...] grantees are obliged to make available to the public in an appropriate manner the research results obtained with the help of SNSF*

Is reproducible computing / data analytics next?

- Reproducibility: A tragedy of errors. Allison DB, Brown AW, George BJ, Kaiser KA. **Nature**. 2016 Feb 4;530(7588):27-9. doi: 10.1038/530027a.
- Reproducibility of computational workflows is automated using continuous analysis. Beaulieu-Jones BK, Greene CS. **Nat Biotechnol** 2017 04; 35(4):342-346 PMID: 28288103 DOI: 10.1038/nbt.3780
- Nextflow enables reproducible computational workflows Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo & Cedric Notredame **Nature Biotechnology** volume 35, pages 316–319 (2017)

As a regular author and reviewer for Nature BioTech / Methods / Communcation, I am quite surprised that Nature Biotech published the continuous analysis and the Nextflow papers. I can only explain them going this high by there being some underlying agenda or a prediction that this topic is of future importance and therefore might be highly cites.

Reproducible computing is hard

- Enhancing reproducibility for computational methods. Stodden V, McNutt M, Bailey DH, Deelman E ... Hanson B, Heroux MA, Ioannidis JP, Taufer M. Science 2016 12 09; 354(6317):1240-1241

"Although some of these actions may be aspirational, we believe it is important to recognize and move toward ameliorating irreproducibility in computational research".

- **Share data, software, workflows, and details of the computational environment** that generate published findings in open trusted repositories.
- **Persistent links** should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.
- To enable credit for shared digital scholarly objects, citation should be standard practice.
- To facilitate reuse, **adequately document digital scholarly artifacts**.
- Use Open Licensing when publishing digital scholarly objects.

S3ITData: an integrated data and workflow manager

- web-based, **easy to use**
- retains data **long-term** (immutable - cannot change)
- workflows (i.e. digital lab protocols) are run in a **reproducible and sharable** way
- reports are **explicitly linked** to the input data
- reports are generated in a clear and **reproducible way**
- **reports can be extended**, stored and linked to the original report.
- Developed with input from several research groups
- Based on several establish open-source software packages such as openBIS, GC3PIE, Singularity Containers and Jupyter

These reports are Jupyter Notebooks

- Data visualization tools drive interactivity and reproducibility in online publishing. Perkel JM. Nature. 2018 Feb 1;554(7690):133-134. doi: 10.1038/d41586-018-01322-9.
- Interactive notebooks: Sharing the code. Shen H. Nature. 2014 Nov 6;515(7525):151-2. doi: 10.1038/515151a.
- Notebooks on GitHub (http://nbviewer.jupyter.org/github/parente/nbestimate/blob/master/estimate.ipynb)