

# Tuesday: Key concepts & alpha diversity

## Demo

Microbiome data exploration

## Lecture

Key concepts in microbiome data science

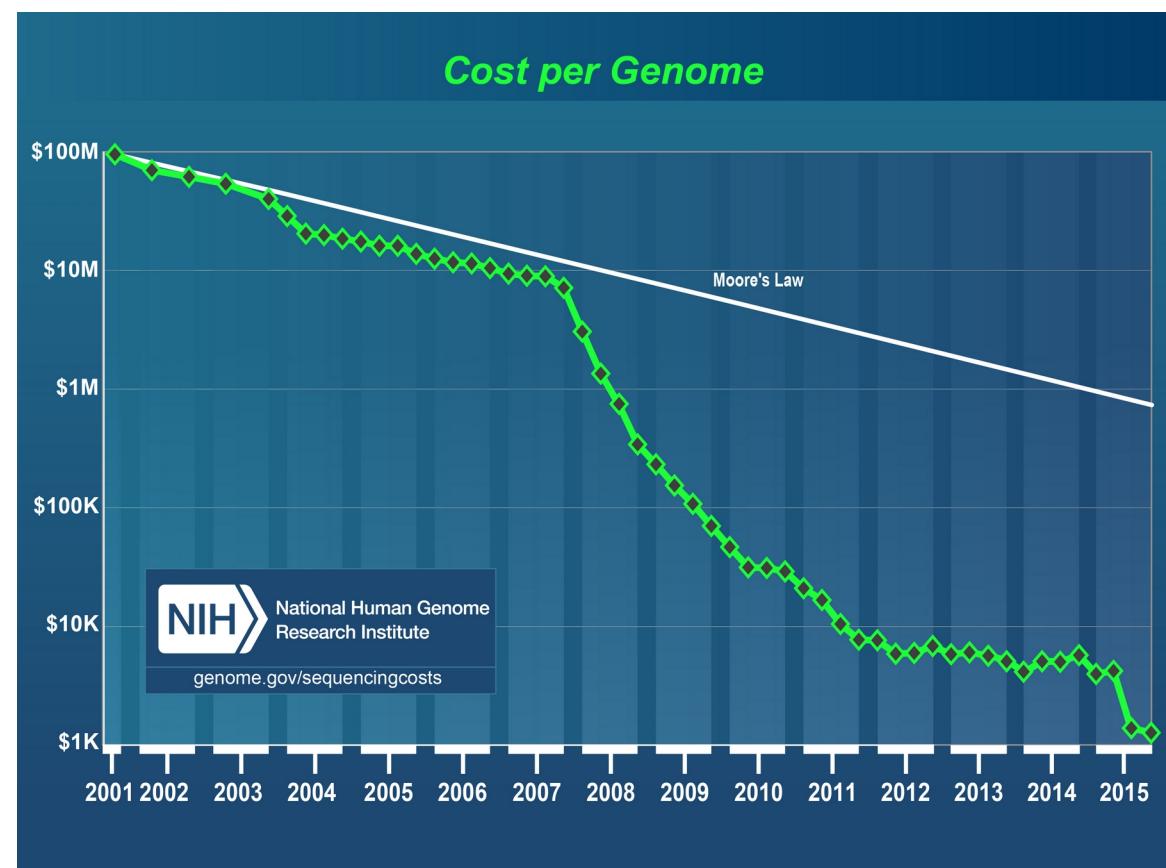
## Practical

Alpha diversity: estimation, analysis, and visualization



# Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions

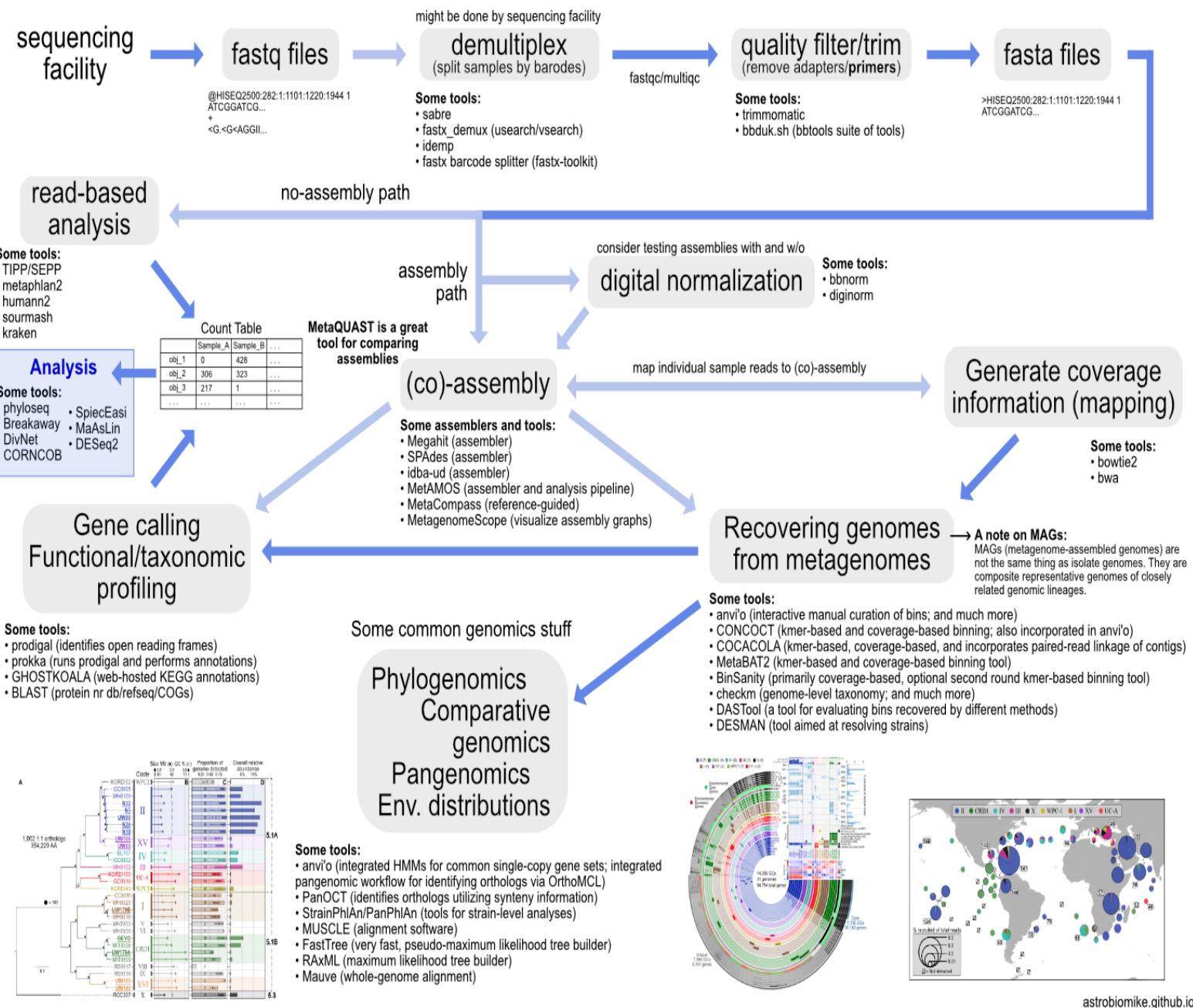
Isabel Moreno-Indias<sup>1,2\*</sup>, Leo Lahti<sup>3</sup>, Miroslava Nedyalkova<sup>4</sup>, Ilze Elbere<sup>5</sup>, Gennady



# Overview of generic\* metagenomics workflow

\*This is generic; specific workflows can vary on the order of steps here and how they are done.

When working with your own data you should never follow any pipeline blindly. There can be critical differences based on your data.



Happy Belly Bioinformatics

JOSE 10.21105/jose.00053

AstroBioMike

Orcid: 0000-0001-7750-9145

Lee, (2019). Happy Belly Bioinformatics: an open-source resource dedicated to helping biologists utilize bioinformatics. Journal of Open Source Education, 4(41), 53, <https://doi.org/10.21105/jose.00053>

astrobioinfo.github.io

# open data science ecosystems

**mothur**

Download Wiki Forum Blog GitHub [facebook](#)

Welcome to the website for the mothur project, initiated by Dr. Patrick Schloss and his research group at the Department of Microbiology & Immunology at The University of Michigan. This project seeks to develop a single piece of open-source, expandable software to fill the bioinformatics needs of the microbial ecology community. mothur is a command-line version of mothur, which had accelerated versions of the popular DOTUR and SONS programs. mothur has gone on to become one of the most cited bioinformatics tool for analyzing 16S rRNA gene sequences. Step inside the wiki and user forum and learn how you can use mothur to process data generated by Sanger, Pacific, Ion, 454, and Illumina platforms. If you would like to contribute code to the project feel free to download the source code and make your own improvements. Alternatively, if you have an idea or a need, but lack the programming expertise, let us know through the forum and we'll add it to the queue of features we would like to add.

[Subscribe to the mothur mailing list](#)

Department of Microbiology & Immunology  
The University of Michigan Medical School  
The University of Michigan

This site is maintained by Pat Schloss  
© 2008-2019

QIIME 2™ is a next-generation microbiome bioinformatics platform that is extensible, free, open source, and community developed.

[Code of Conduct »](#) [Citing QIIME 2 »](#) [Learn more »](#)

Automatically track your analyses with decentralized data provenance — no more guesswork on what commands were run!

Interactively explore your data with beautiful visualizations that provide new perspectives.

Easily share results with your team, even those members without QIIME 2 installed.

Plugin-based system — your favorite microbiome methods all in one place.

[PeerJ >](#)

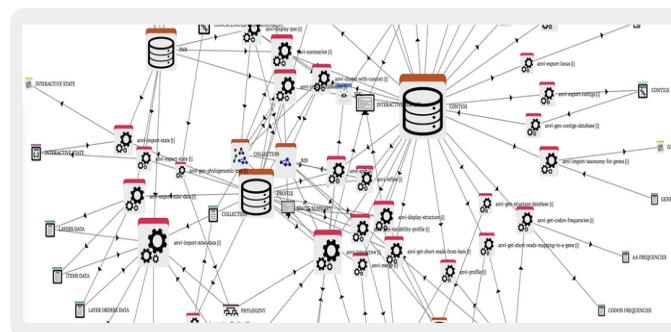
## Anvi'o: an advanced analysis and visualization platform for 'omics data

[Research article](#) Bioinformatics Biotechnology Computational Biology Genomics Microbiology

A. Murat Eren<sup>✉ 1,2</sup>, Özcan C. Esen<sup>1</sup>, Christopher Quince<sup>3</sup>, Joseph H. Vineis<sup>1</sup>, Hilary G. Morrison<sup>1</sup>, Mitchell L. Sogin<sup>1</sup>, Tom O. Delmont<sup>1</sup>

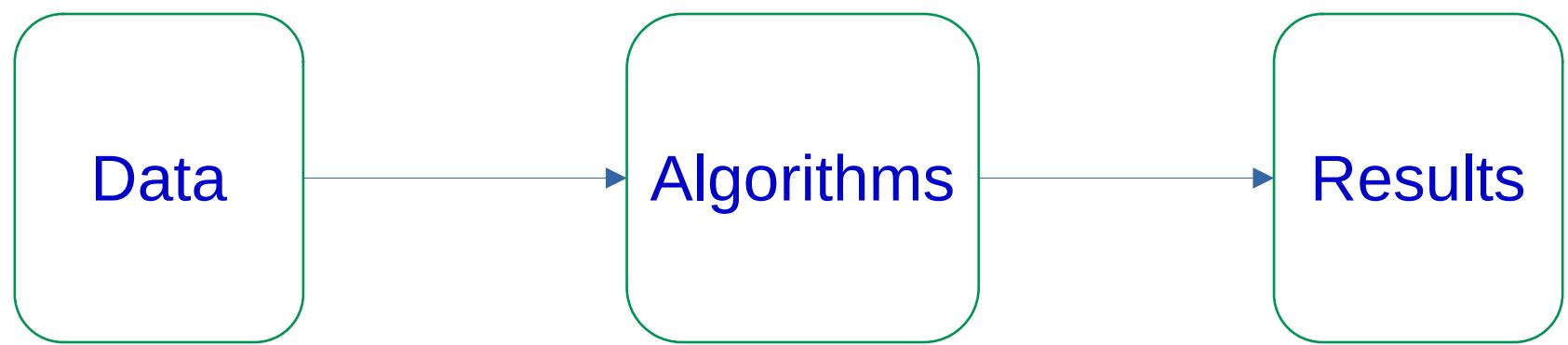
Published October 8, 2015

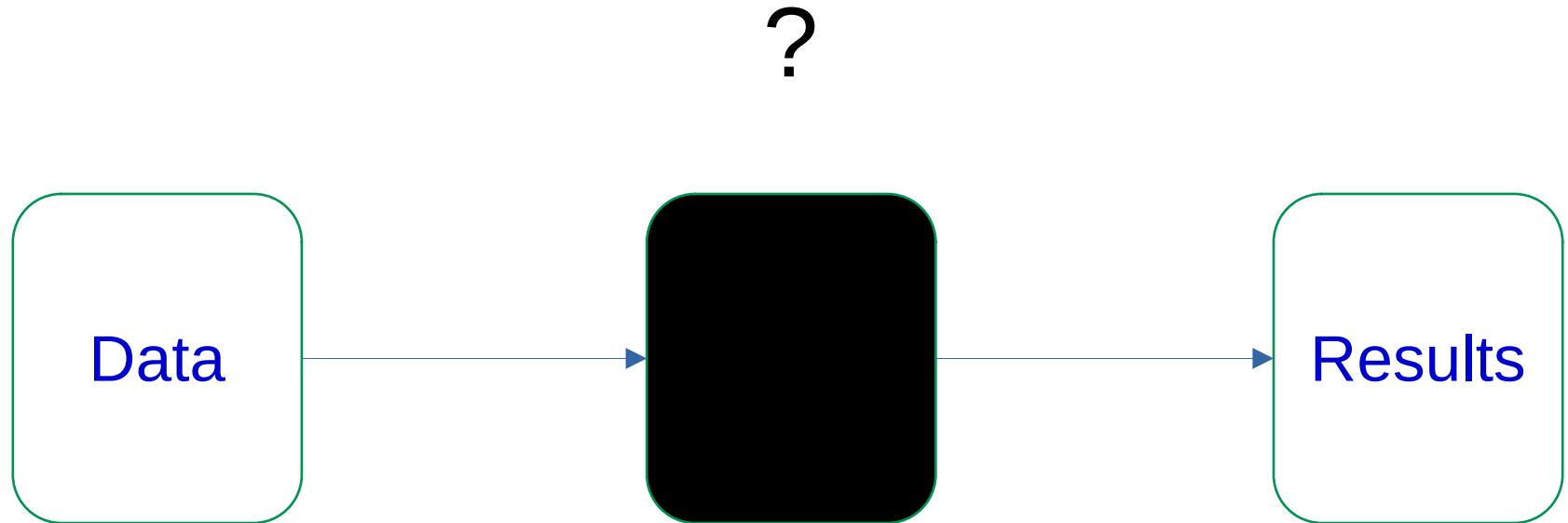
### Anvi'o in a nutshell



Anvi'o is an [open-source](#), community-driven analysis and visualization platform for 'omics data.



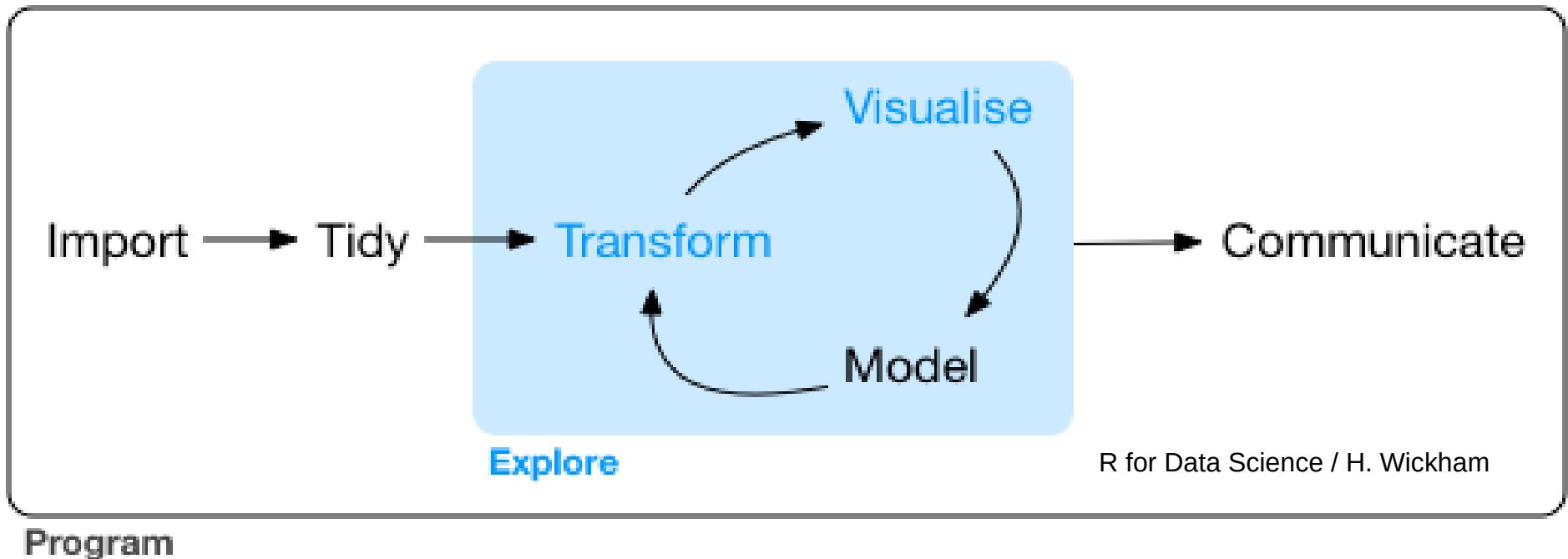




RESEARCH PRIORITIES  
**Shining Light into Black Boxes**

A. Morin<sup>1</sup>, J. Urban<sup>2</sup>, P. D. Adams<sup>3</sup>, I. Foster<sup>4</sup>, A. Sali<sup>5</sup>, D. Baker<sup>6</sup>, P. Sliz<sup>1,\*</sup>

# Data science workflow



**REVISED** Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses [version 2; peer review: 3 approved]

Ben J. Callahan<sup>1</sup>, Kris Sankaran<sup>1</sup>, Julia A. Fukuyama<sup>1</sup>, Paul J. McMurdie<sup>2</sup>, Susan P. Holmes



This article is included in the [Bioconductor](#) gateway.

# The influence of hidden researcher decisions in applied microeconomics

Nick Huntington-Klein , Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, Yaniv Stopnitzky

First published: 22 March 2021

<https://doi.org/10.1111/ecin.12992>

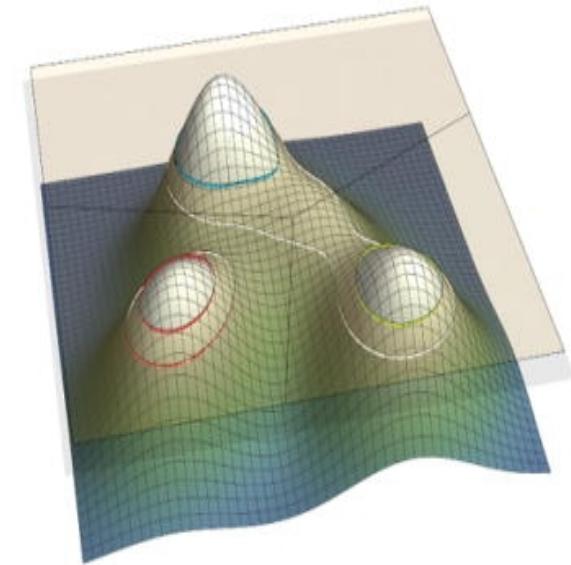
Researchers make hundreds of decisions about data collection, preparation, and analysis in their research. We use a many-analysts approach to measure the extent and impact of these decisions. Two published causal empirical results are replicated by seven replicators each. We find large differences in data preparation and analysis decisions, many of which would not likely be reported in a publication. No two replicators reported the same sample size. Statistical significance varied across replications, and for one of the studies the effect's sign varied as well. The standard deviation of estimates across replications was 3–4 times the mean reported standard error.

*“I have begun to think that no one ought to publish biometric results, without lodging a well arranged and well bound manuscript copy of all his data, in some place where it should be accessible, under reasonable restrictions, to those who desire to verify his work.”*

Francis Galton (1901), *Biometrika* 1:1, pp. 7-10.

# How to choose a correct model?

→ a community typing example



$$2 \times 6^6 = 93312$$

## Taxonomic level

- Phylum
- Family
- Order
- Genus
- Species
- Strain...

## Filtering

- None
- Prevalent
- Core
- Excl. outliers
- High variance
- Custom

## Normalization

- None
- TSS
- CSS
- ILR/ALR/CLR
- phILR
- Hellinger

## (Dis)similarity

- Eulidean
- Aitchison
- Bray-Curtis
- Jaccard
- weighted Unifrac
- unweighted Unifrac

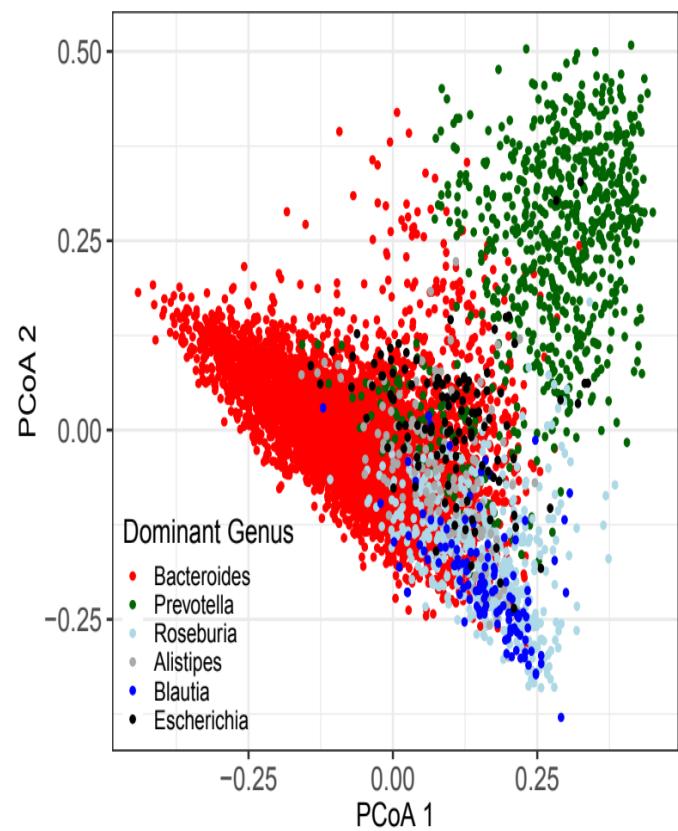
## Clustering method

- Hierarchical / Ward
- Hierarchical / Complete
- Gaussian mixture
- DMM
- PAMR
- K-means

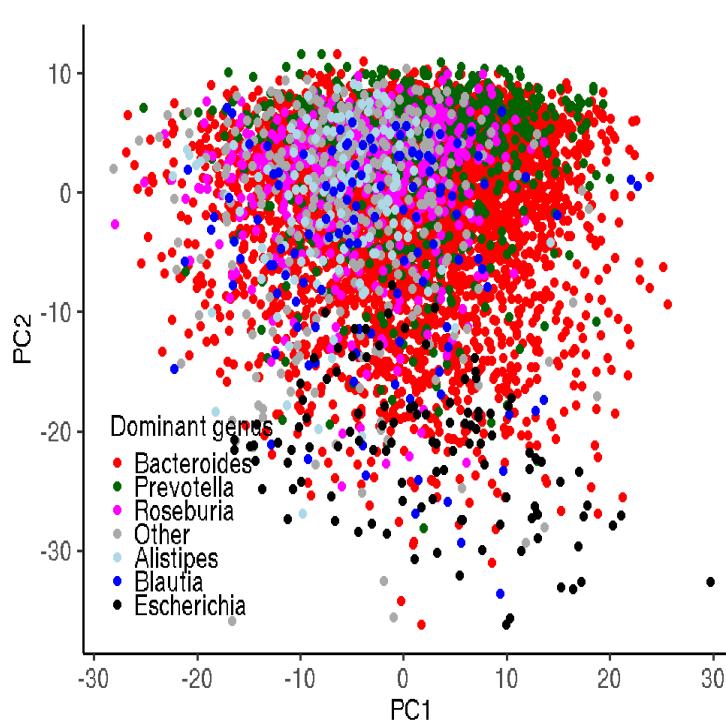
## Regulation

- Calinski-Harabasz
- Dirichlet Process
- Silhouette Index
- AIC
- BIC
- DIC

## PCoA + Bray-Curtis



## PCA + Aitchison



## Reproducible Research: Enterotype Example

Susan Holmes and Joey McMurdie

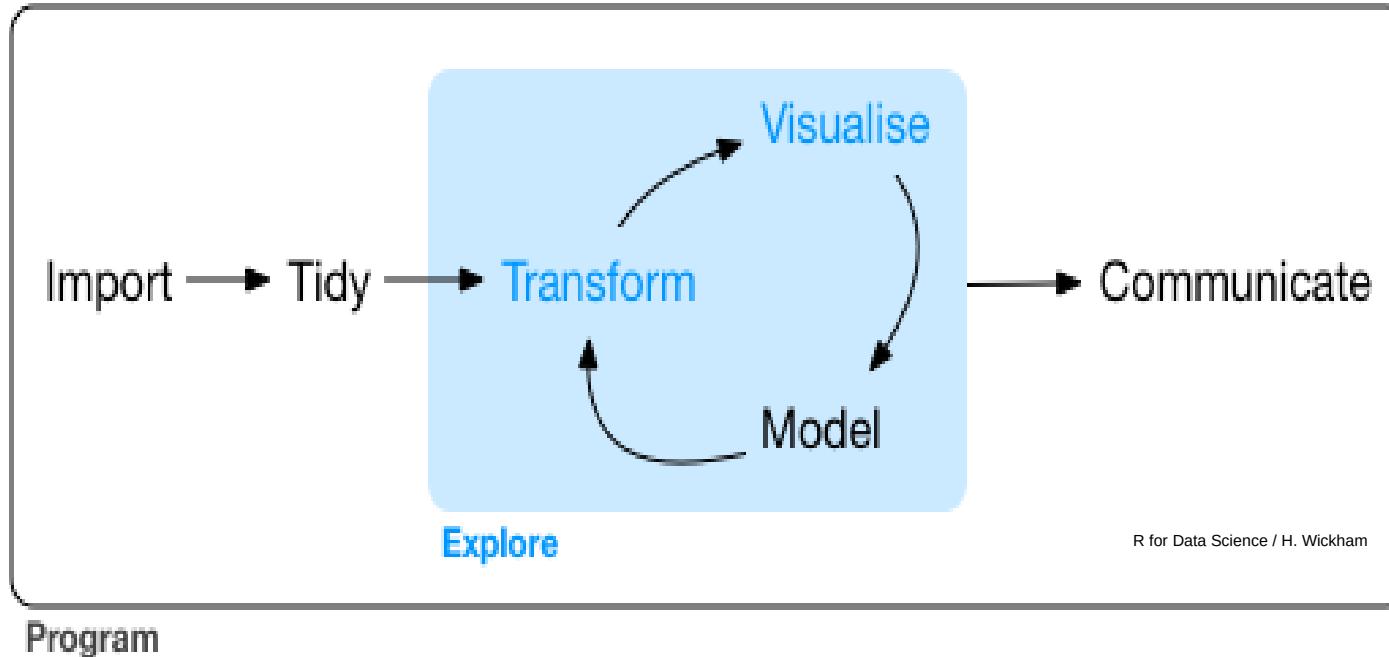
<http://statweb.stanford.edu/~susan/papers/EnterotypeRR.html>

[Comment on this paper](#)

Taxonomic Signatures of Long-Term Mortality Risk in Human Gut Microbiota

Aaro Saloensaa, Ville Laitinen, Aki Havulinna, Guillaume Meric, Susan Cheng, Markus Perola, Liisa Valsta, Georg Alfrhan, Michael Inouye, Jeremie D. Watrous, Tao Long, Rodolfo Salido, Karenina Sanders, Caitriona Brennan, Gregory C. Humphrey, Jon G. Sanders, Mohit Jain, Pekka Jousilahti, Veikko Salomaa, Rob Knight, Leo Lahti, Teemu Niiranen  
doi: <https://doi.org/10.1101/2019.12.30.19015842>

# Reproducible workflows improve transparency and robustness



## Taxonomic level?

- Phylum
- Family
- Order
- Genus
- Species
- Strain...

## Normalization

- None
- TSS
- CSS
- ILR/ALR/CLR
- phILR
- Hellinger

## (Dis)similarity?

- Euclidean
- Aitchison
- Bray-Curtis
- Jaccard
- weighted Unifrac
- unweighted Unifrac

## Regulation

- Calinski-Harabasz
- Dirichlet Process
- Silhouette Index
- AIC
- BIC
- DIC

## Clustering

- Hierarchical / Ward
- Hierarchical / Complete
- Gaussian mixture
- DMM
- PAMR
- K-means

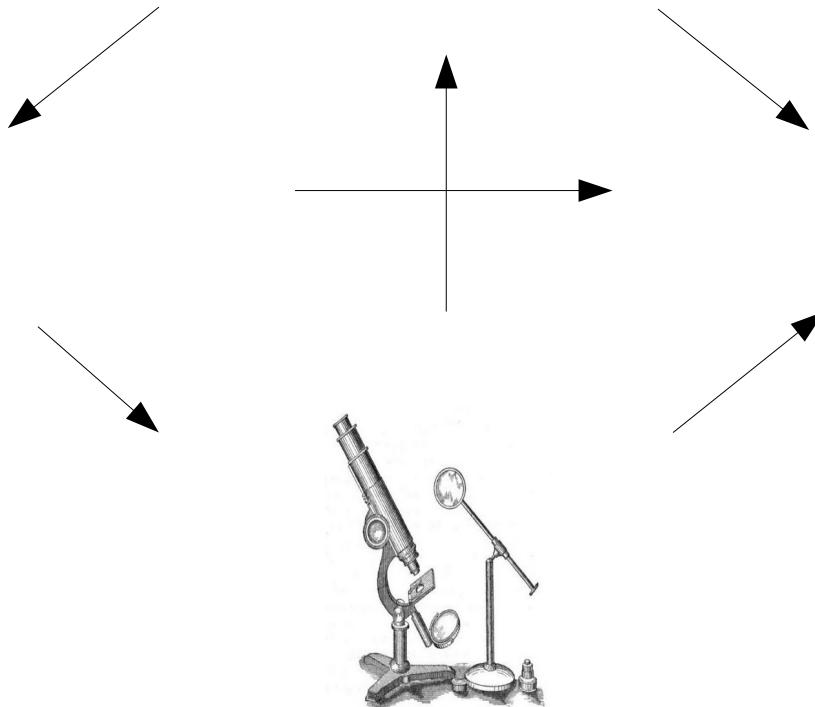
# Hypothesis testing vs. hypothesis discovery?

?

# Hypothesis

# Method

$$\begin{cases} X = W_x \mathbf{z} + \varepsilon_x \\ Y = W_y \mathbf{z} + \varepsilon_y \end{cases}$$



# Tools



# Data

OPEN ACCESS

ESSAY

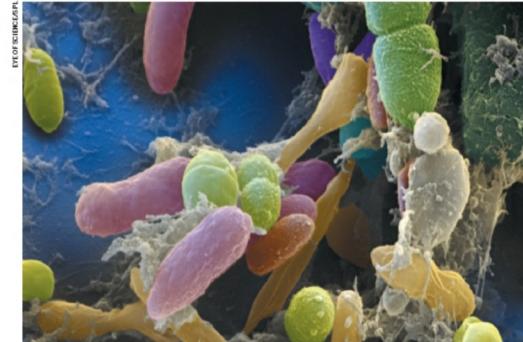
## Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: Aug 30, 2005 • DOI: 10.1371/journal.pmed.0020124

898,944

VIEWS



A scanning electron micrograph of bacteria in human faeces, in which 50% of species originate from the gut.

## Microbiome science needs a healthy dose of scepticism

To guard against hype, those interpreting research on the body's microscopic communities should ask five questions, says William P. Hanage.

Comment August 2014 Nature

## How to Make More Published Research True

John P. A. Ioannidis 

Published: October 21, 2014 • DOI: 10.1371/journal.pmed.1001747

```
int getRandomNumber()
{
    return 4; // chosen by fair dice roll.
              // guaranteed to be random.
}
```

<http://web.stanford.edu/class/cs109l/unrestricted/images/>

#### RESEARCH PRIORITIES

#### Shining Light into Black Boxes

A. Morin<sup>1</sup>, J. Urban<sup>2</sup>, P. D. Adams<sup>3</sup>, I. Foster<sup>4</sup>, A. Sali<sup>5</sup>, D. Baker<sup>6</sup>, P. Sliz<sup>1,\*</sup>

You aren't  
doing it wrong



if no one knows  
what you are doing.



# Data silo



The demise of alchemy provides further evidence, if further evidence were needed, that what marks out modern science is not the conduct of experiments (alchemists conducted plenty of experiments), but the formation of a *critical community capable of assessing discoveries and replicating results*. Alchemy, as a clandestine enterprise, could never develop a community of the right sort. Popper was right to think that science can flourish only in an open society.

## The Invention of Science: A New History of the Scientific Revolution, by David Wootton



A family of alchemists at work, an engraving by Philip Galle, after a painting by Pieter Bruegel the Elder, published by Hieronymus Cock, c.1558.



# Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

# What is Bioconductor ?

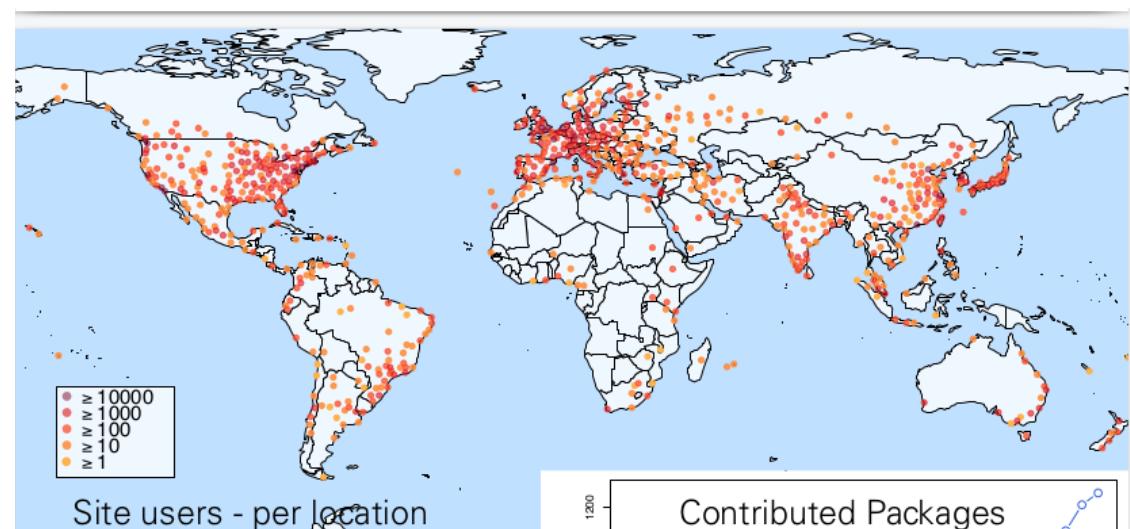
Started 2001 as a platform for analysis & understanding of microarray data

More than 1,600 packages. Domains of expertise:

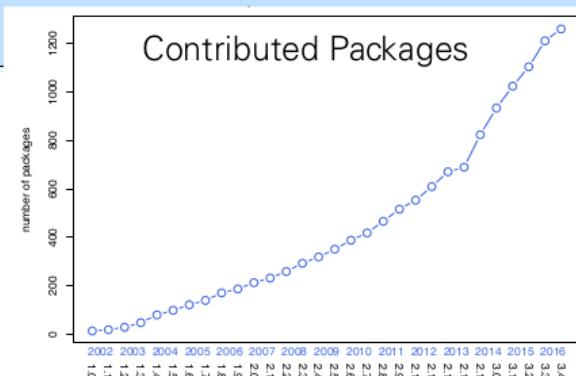
- Sequencing (RNASeq, ChIPSeq, single-cell, called variants, ...)
- Microarrays (methylation, expression, copy number, ...)
- Flow cytometry
- Proteomics
- Multi-Omics data integration

## Important themes

- Reproducible research
- Interoperability between packages & workflows  
... even from different authors
- Usability



World largest bioinformatics project  
10,000s users  
>18,000 papers in PubmedCentral



What is



?

Principally a collaborative software development project

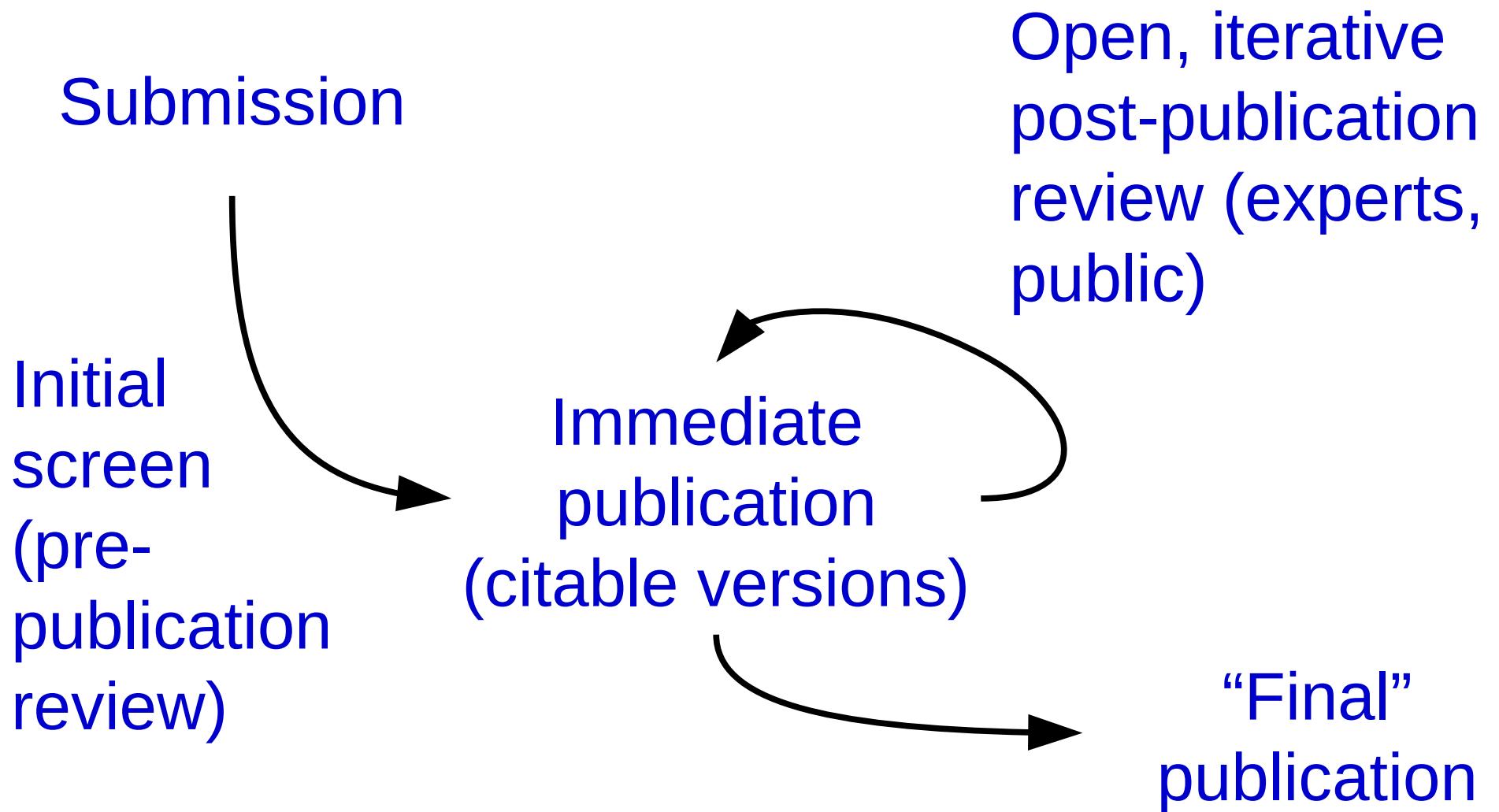
But it is also:

- a software repository
- a bioinformatics support site
- data repository
- publisher for supplementary materials
- source for tutorials and instructional documentation

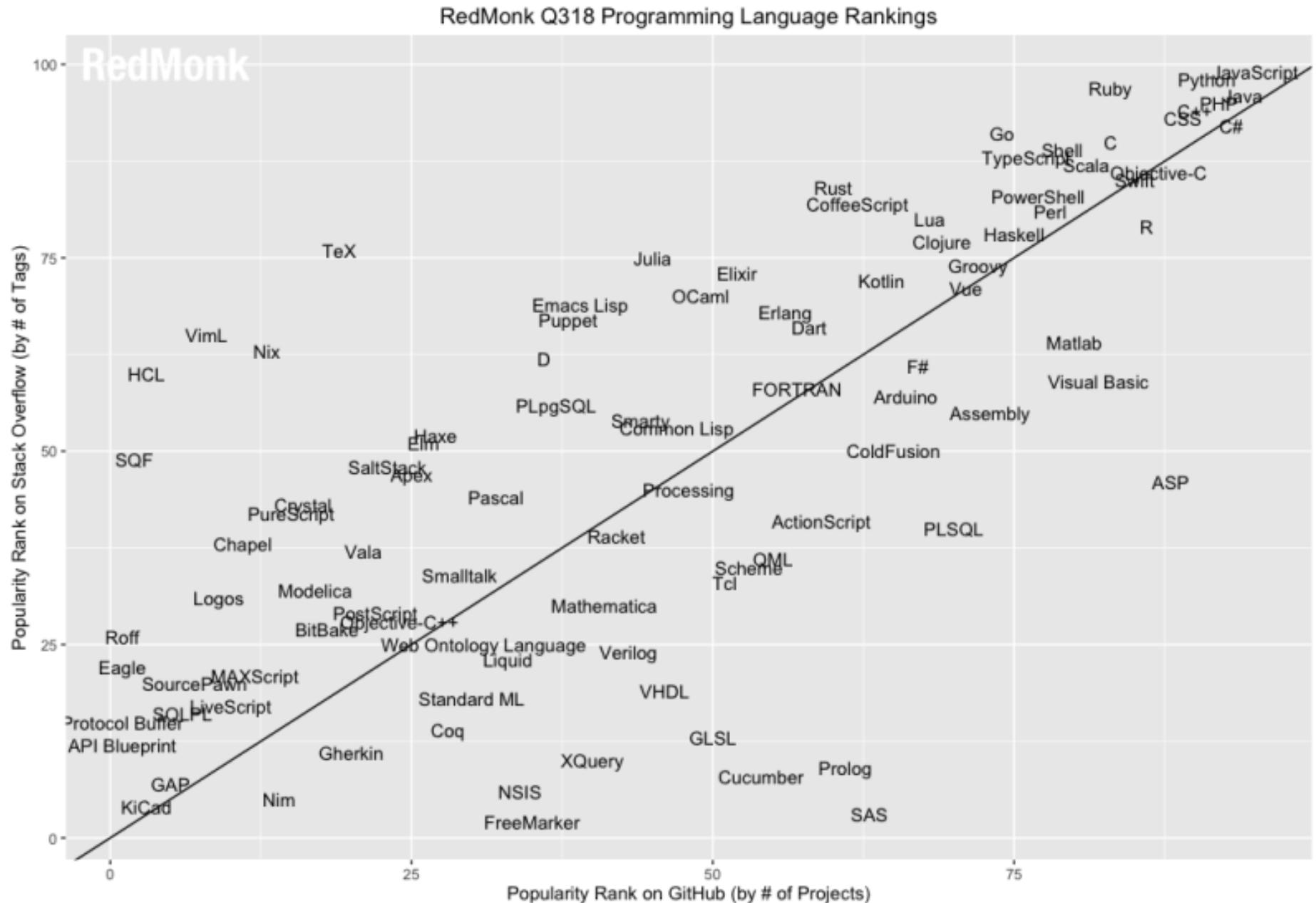
Managed and maintained  
by a core team of ~6  
people, with contributions  
coming from all over the  
world



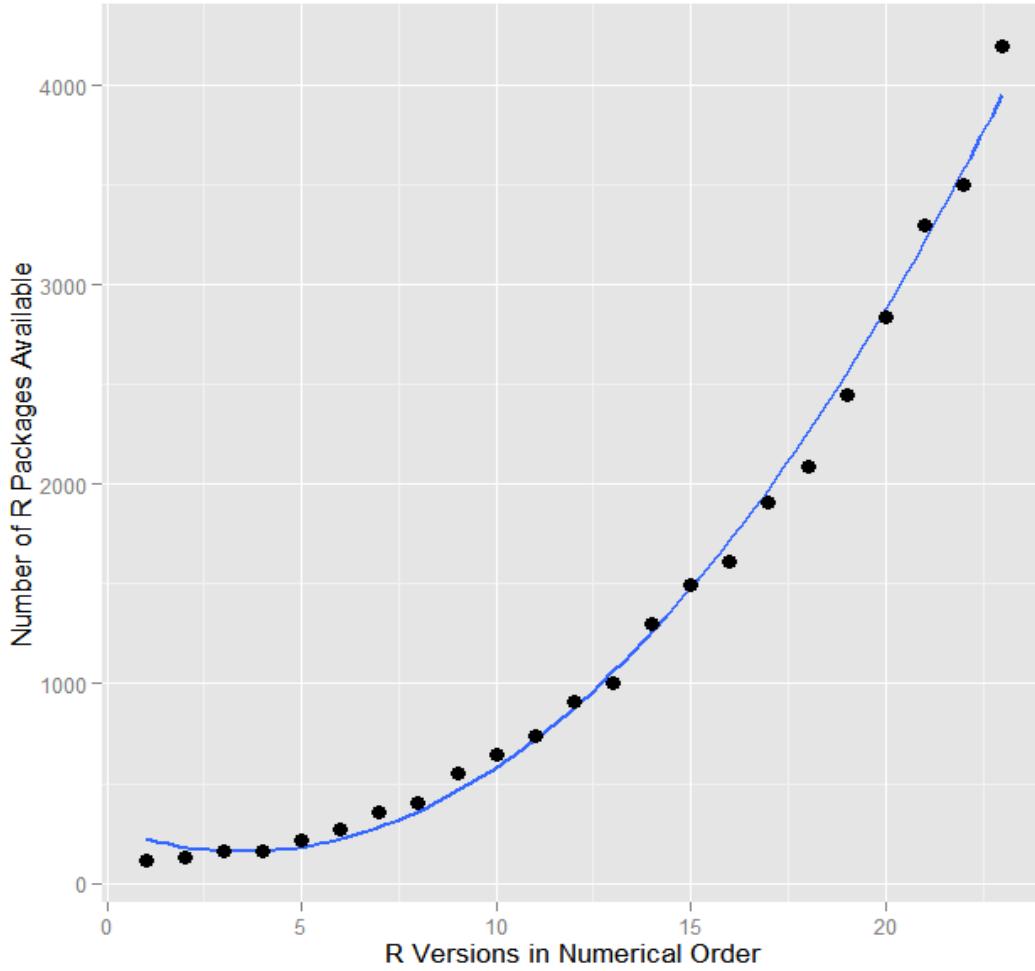
# Pre- vs. post-publication review



# Varying cultures of open collaboration



# Number of open analysis tools has grown exponentially



Value of data can increase through sharing & use

## A Quick Guide to Software Licensing for the Scientist-Programmer

Andrew Morin, Jennifer Urban, Piotr Sliz 

Published: July 26, 2012 • <https://doi.org/10.1371/journal.pcbi.1002598>



### Software citation principles

Arfon M. Smith<sup>1,\*</sup>, Daniel S. Katz<sup>2,\*</sup>, Kyle E. Niemeyer<sup>3,\*</sup>  
FORCE11 Software Citation Working Group

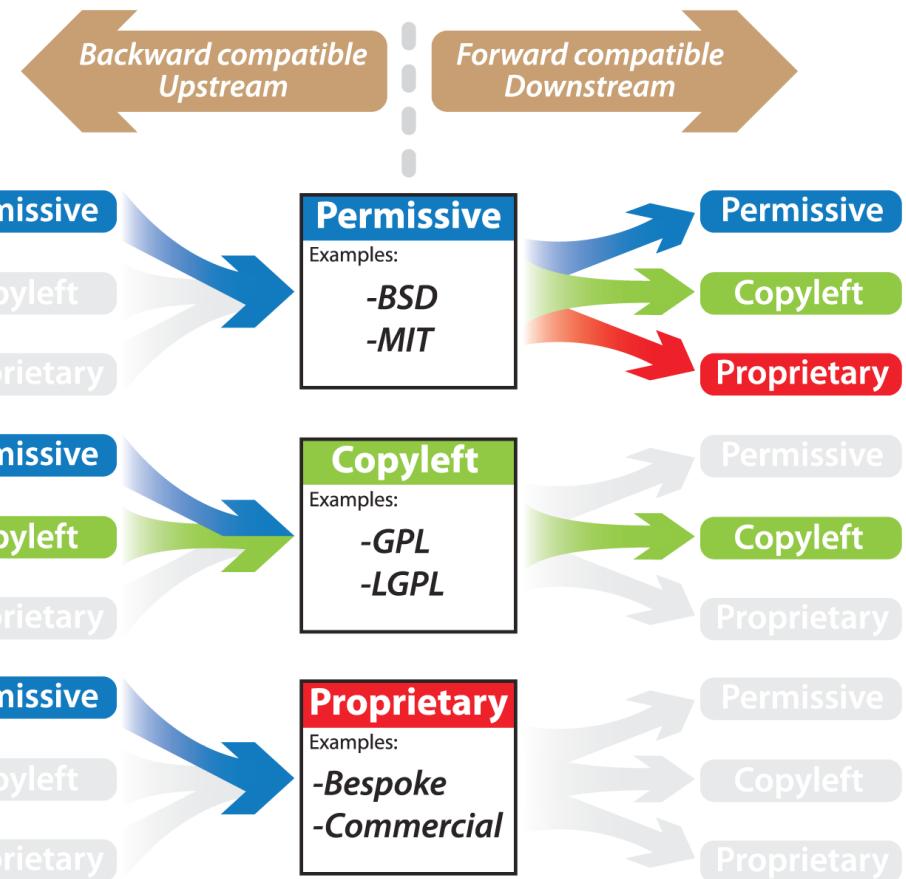
<sup>1</sup> GitHub, Inc., San Francisco, California, United States

<sup>2</sup> National Center for Supercomputing Applications & Electrical and Computer Department & School of Information Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States

<sup>3</sup> School of Mechanical, Industrial, and Manufacturing Engineering, Oregon State University, Corvallis, Oregon, United States

\* These authors contributed equally to this work.

## MIT License



Copyright (c) <year> <copyright holders>

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

# Open reporting and communication were part of academic culture since the early days

BJHS 45(2): 165–188, June 2012. © British Society for the History of Science 2012  
doi:10.1017/S0007087412000064 First published online 20 March 2012

Openness versus secrecy? Historical and historiographical remarks

KOEN VERMEIR\*



Source: Wikimedia Commons / Public domain

## Alchemy & algorithms: perspectives on the philosophy and history of open science

Research Ideas and Outcomes 3:e13593, 2017

▼ Leo Lahti, Filipe da Silva, Markus Petteri Laine, Viivi Lähteenaja, Mikko Tolonen

## Beyond Open Access - The Changing Culture of Producing and Disseminating Scientific Knowledge

Heidi Laine

Leo Lahti

Anne Lehto