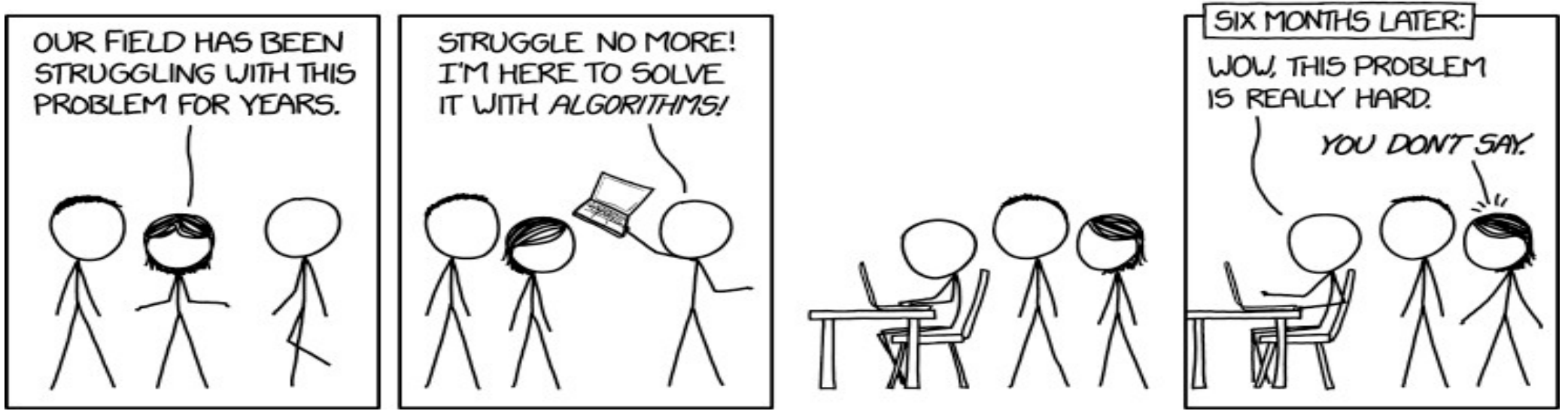
Open data science



Initial sequencing and analysis of the human genome

~2001

International Human Genome Sequencing Consortium*

** A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. **Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome**. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

PERSPECTIVE ARTICLE

Front. Microbiol., 22 February 2021 | https://doi.org/10.3389/fmicb.2021.635781

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

Isabel Moreno-Indias[1,2*], Leo Lahti[3], Miroslava Nedyalkova[4], Ilze Elbere[5], Gennady

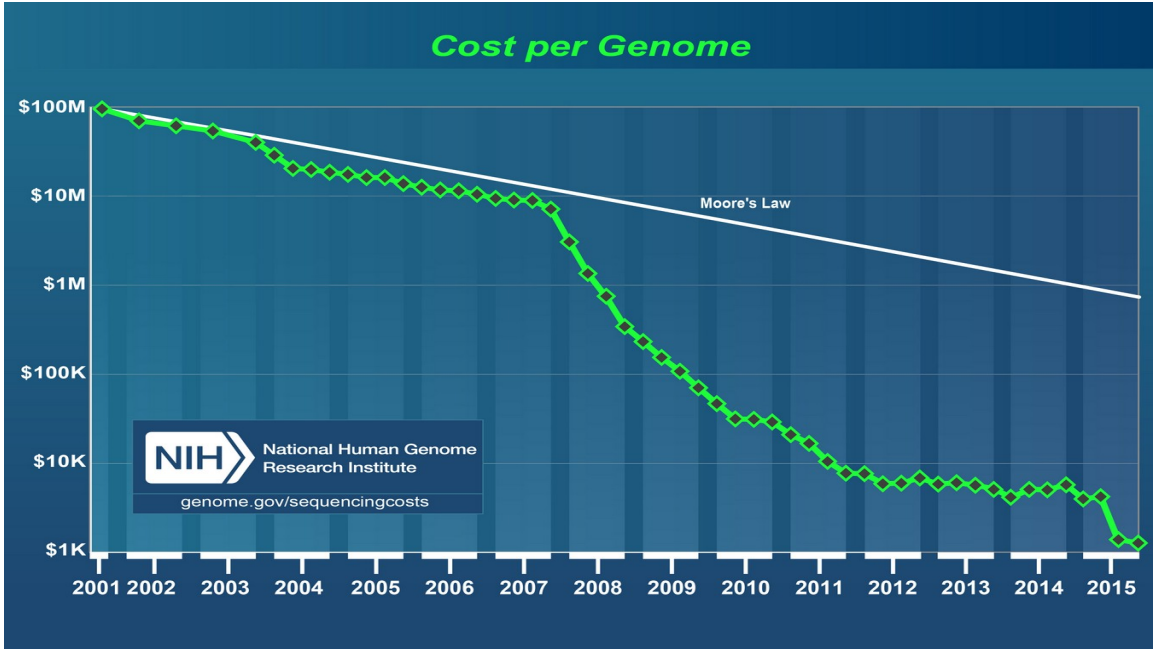
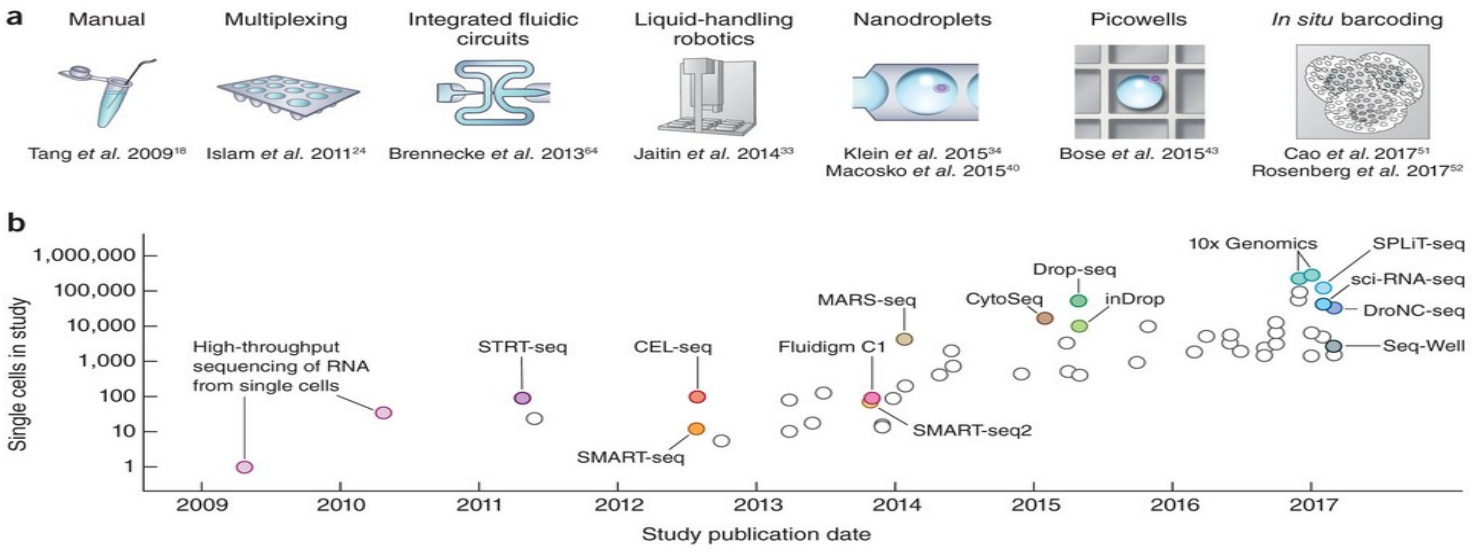


Figure 1: Scaling of scRNA-seq experiments.

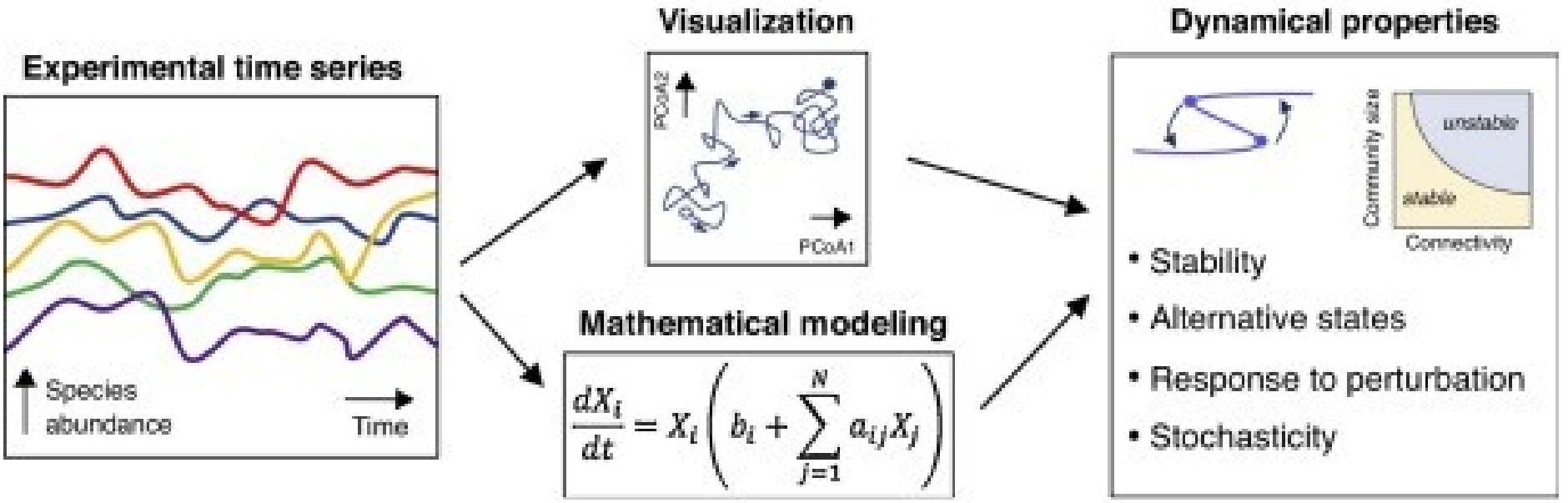
From: Exponential scaling of single-cell RNA-seq in the past decade



a) Key technologies that have allowed jumps in experimental scale. A jump to ~100 cells was enabled by sample multiplexing, and then a jump to ~1,000 cells was achieved by large-scale studies using integrated fluidic circuits, followed by a jump to several thousands of cells with liquid-handling robotics. Further orders-of-magnitude increases bringing the number of cells assayed into the tens of thousands were enabled by random capture technologies using nanodroplets and picowell technologies. Recent studies have used in situ barcoding to inexpensively reach the next order of magnitude of hundreds of thousands of cells. (b) Cell numbers reported in representative publications by publication date. Key technologies are indicated.

Microbial communities as dynamical systems

Didier Gonze [1, 2] ✉, Katharine Z Coyte [3, 4], Leo Lahti [5, 6, 7], Karoline Faust [5] ✉



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.

sequencing facility → fastq files → demultiplex (split samples by barodes) → quality filter/trim (remove adapters/**primers**) → fasta files

might be done by sequencing facility

fastq files:
@HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G.<G<AGGII...

fastqc/multiqc

Some tools:
- sabre
- fastx_demux (usearch/vsearch)
- idemp
- fastx barcode splitter (fastx-toolkit)

Some tools:
- trimmomatic
- bbduk.sh (bbtools suite of tools)

fasta files:
>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...

no-assembly path

read-based analysis

Some tools:
- TIPP/SEPP
- metaphlan2
- humann2
- sourmash
- kraken

assembly path

consider testing assemblies with and w/o

digital normalization

Some tools:
- bbnorm
- diginorm

Count Table

	Sample_A	Sample_B	. . .
obj_1	0	428	. . .
obj_2	306	323	. . .
obj_3	217	1	. . .
. . .	. . .	. . .	. . .

MetaQUAST is a great tool for comparing assemblies

Analysis

Some tools:
- phyloseq
- Breakaway
- DivNet
- CORNCOB
- SpiecEasi
- MaAsLin
- DESeq2

(co)-assembly

map individual sample reads to (co)-assembly

Generate coverage information (mapping)

Some tools:
- bowtie2
- bwa

Some assemblers and tools:
- Megahit (assembler)
- SPAdes (assembler)
- idba-ud (assembler)
- MetAMOS (assembler and analysis pipeline)
- MetaCompass (reference-guided)
- MetagenomeScope (visualize assembly graphs)

Gene calling Functional/taxonomic profiling

Some tools:
- prodigal (identifies open reading frames)
- prokka (runs prodigal and performs annotations)
- GHOSTKOALA (web-hosted KEGG annotations)
- BLAST (protein nr db/refseq/COGs)

Recovering genomes from metagenomes

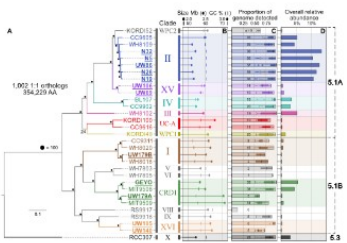
A note on MAGs:
MAGs (metagenome-assembled genomes) are not the same thing as isolate genomes. They are composite representative genomes of closely related genomic lineages.

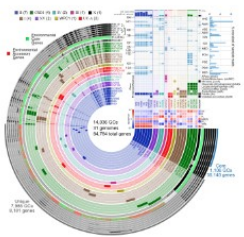
Some tools:
- anvi'o (interactive manual curation of bins; and much more)
- CONCOCT (kmer-based and coverage-based binning; also incorporated in anvi'o)
- COCACOLA (kmer-based, coverage-based, and incorporates paired-read linkage of contigs)
- MetaBAT2 (kmer-based and coverage-based binning tool)
- BinSanity (primarily coverage-based, optional second round kmer-based binning tool)
- checkm (genome-level taxonomy; and much more)
- DASTool (a tool for evaluating bins recovered by different methods)
- DESMAN (tool aimed at resolving strains)

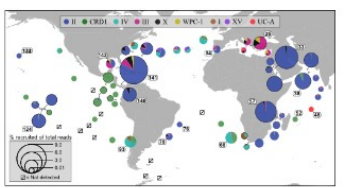
Some common genomics stuff

Phylogenomics
Comparative genomics
Pangenomics
Env. distributions

Some tools:
- anvi'o (integrated HMMs for common single-copy gene sets; integrated pangenomic workflow for identifying orthologs via OrthoMCL)
- PanOCT (identifies orthologs utilizing synteny information)
- StrainPhlAn/PanPhlAn (tools for strain-level analyses)
- MUSCLE (alignment software)
- FastTree (very fast, pseudo-maximum likelihood tree builder)
- RAxML (maximum likelihood tree builder)
- Mauve (whole-genome alignment)







astrobiomike.github.io

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

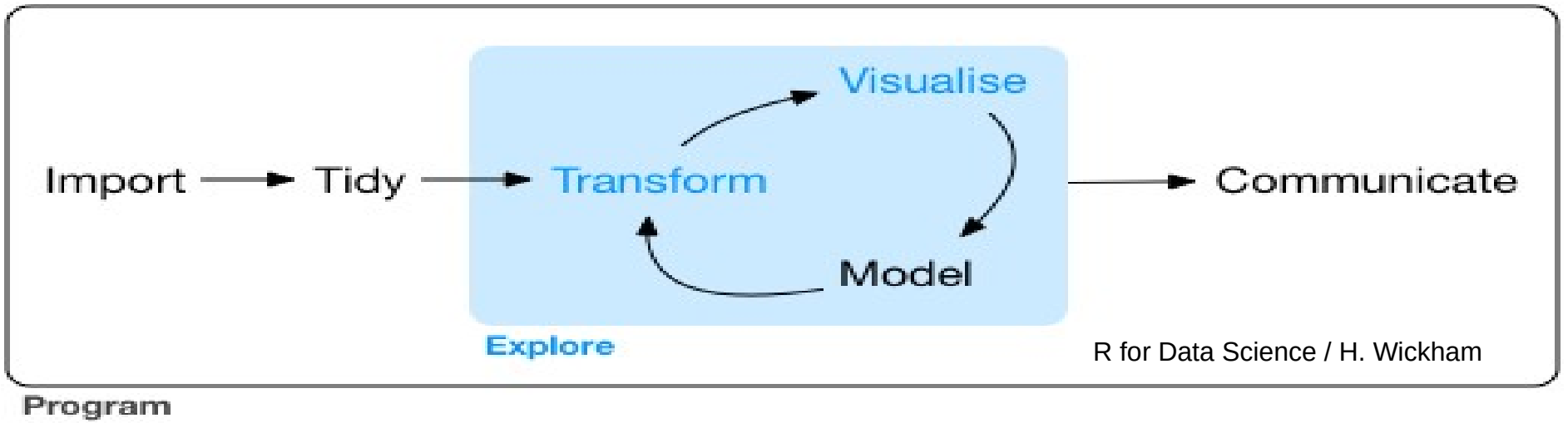
The influence of hidden researcher decisions in applied microeconomics

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.

Data science workflow



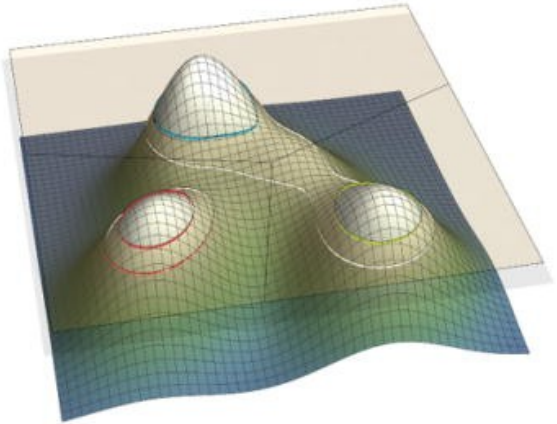
R for Data Science / H. Wickham

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

Ben J. Callahan[1], Kris Sankaran[1], Julia A. Fukuyama[1], Paul J. McMurdie[2], Susan P. Holmes

This article is included in the Bioconductor gateway.

How to choose a correct model?
→ a community typing example



Enterotypes in the landscape of gut microbial community composition. Costea *et al*. Nature 2018.

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

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

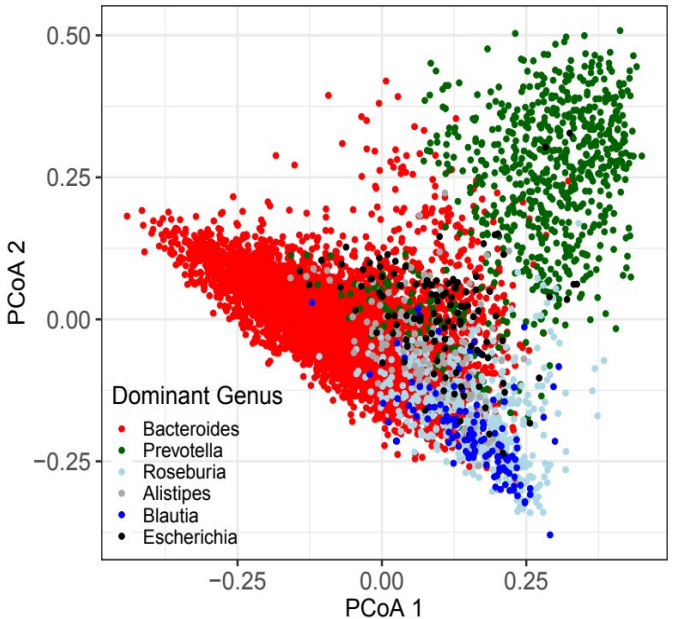
(Dis)similarity
- Eulidean
- Aitchison
- Bray-Curtis
- Jaccard
- weighted Unifrac
- unweighted Unifrac

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

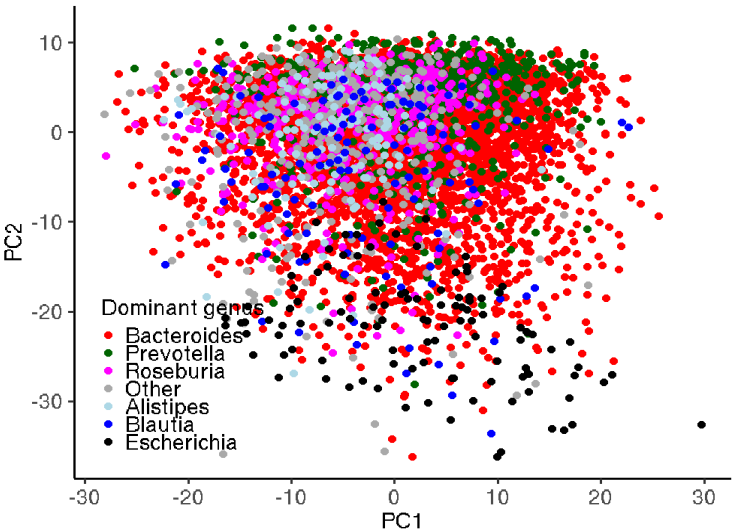
Regulation
- Calinski-Harabasz
- Dirichlet Process
- Silhouette Index
- AIC
- BIC
- DIC

Walk-through example in R/Bioc by Holmes & McMurdie
http://statweb.stanford.edu/~susan/papers/EnterotypeRR.html

PCoA + Bray-Curtis



PCA + Aitchison



Reproducible Research: Enterotype Example

Susan Holmes and Joey McMurdie

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

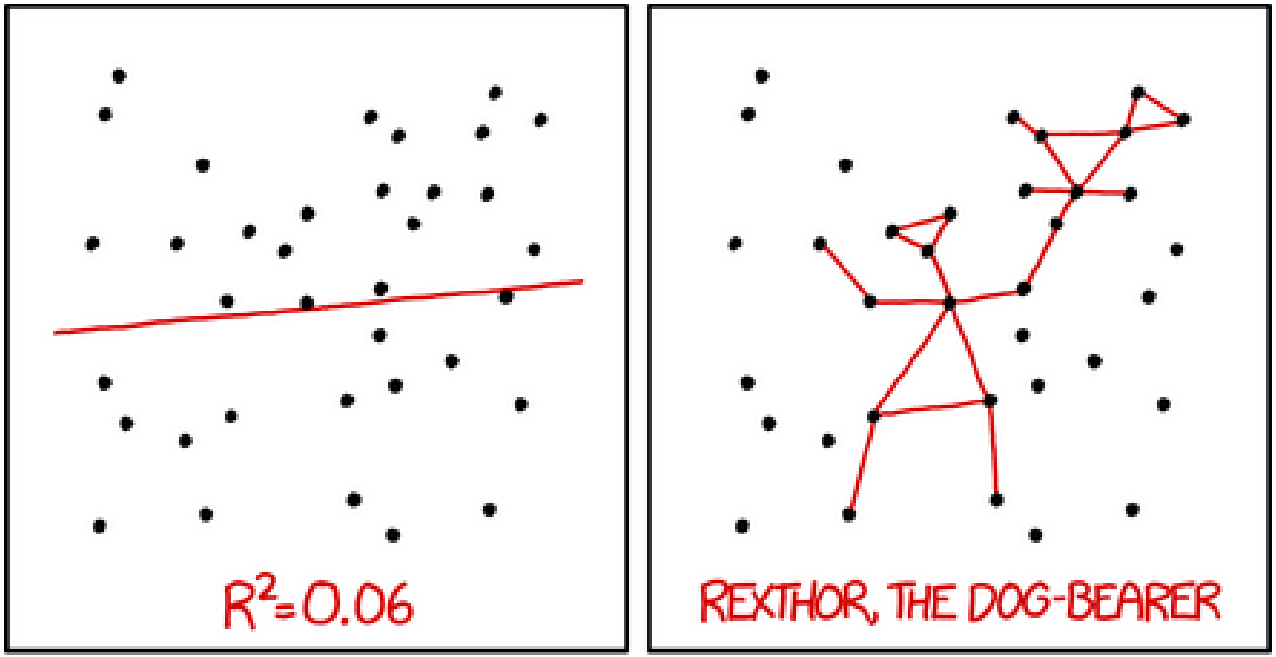
Comment on this paper

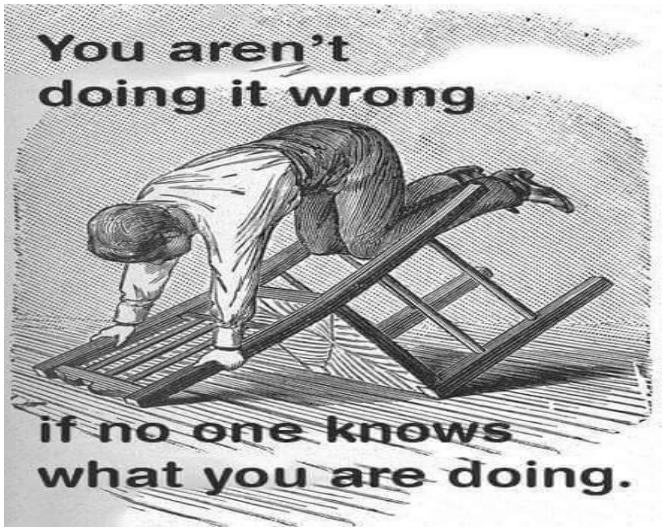
Taxonomic Signatures of Long-Term Mortality Risk in Human Gut Microbiota

Aaro Salosensaari, Ville Laitinen, Aki Havulinna, Guillaume Meric, Susan Cheng, Markus Perola, Liisa Valsta, Georg Alfthan, Michael Inouye, Jeramie 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

How we choose which model to apply?





"*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 whereit should be accessible, under reasonable restrictions, to those who desire to verify his work.*"

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

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

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

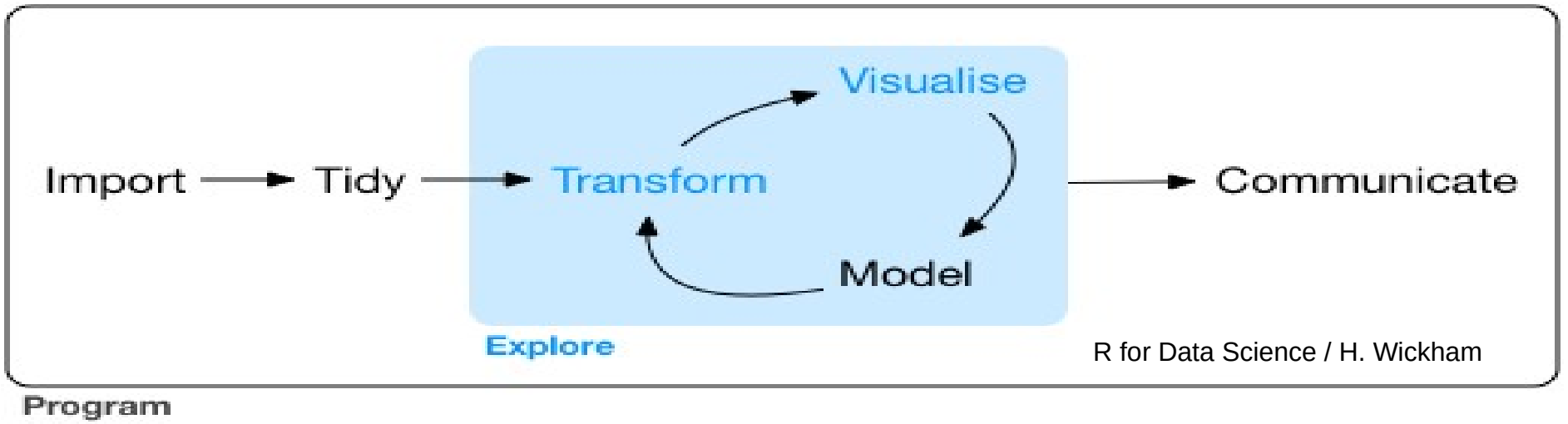
RESEARCH PRIORITIES

Shining Light into Black Boxes

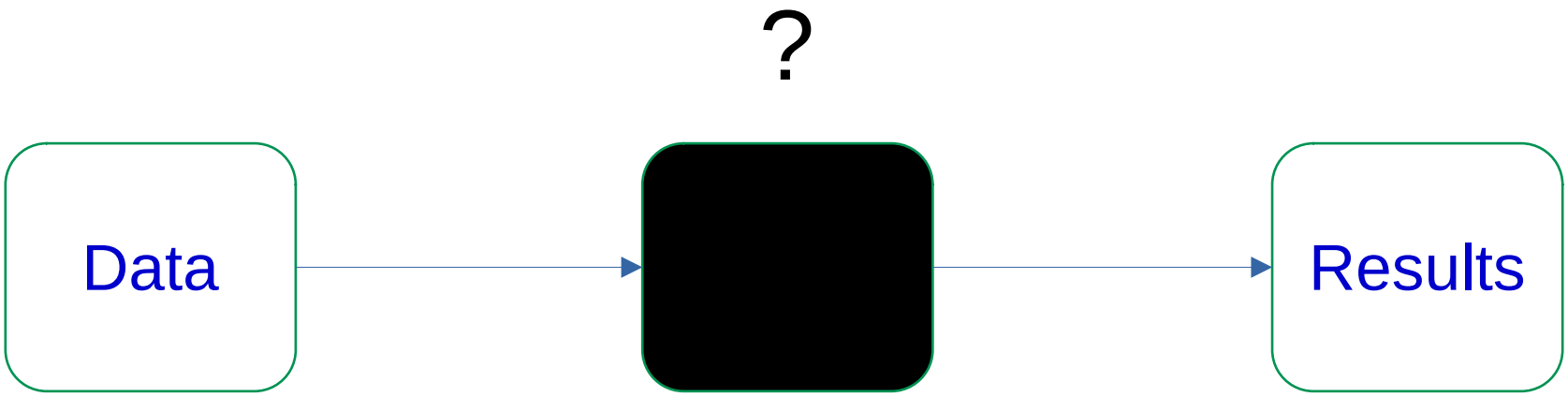
A. Morin[1], J. Urban[2], P. D. Adams[3], I. Foster[4], A. Sali[5], D. Baker[6], P. Sliz[1,*]

Data silo

Data science workflow



R for Data Science / H. Wickham



RESEARCH PRIORITIES

Shining Light into Black Boxes

A. Morin[1], J. Urban[2], P. D. Adams[3], I. Foster[4], A. Sali[5], D. Baker[6], P. Sliz[1,*]

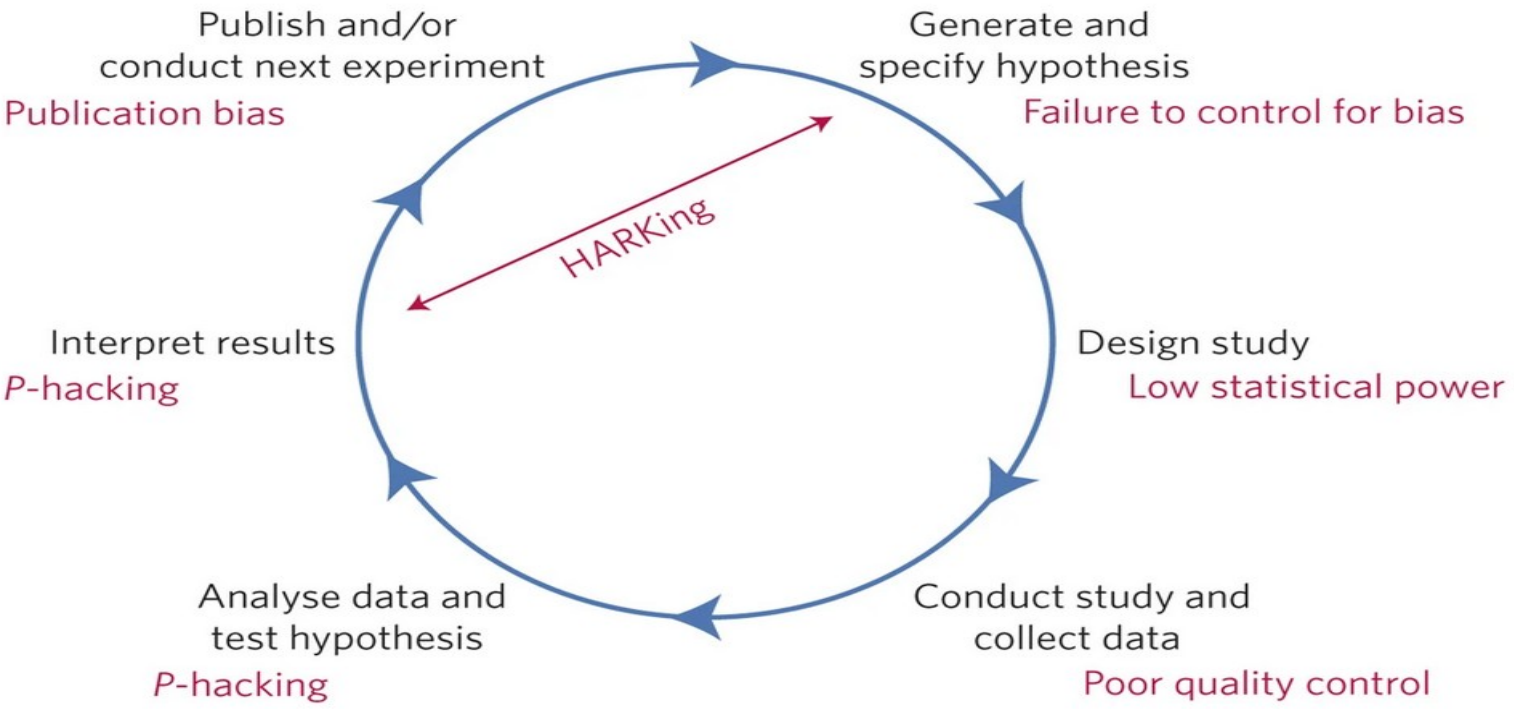
A manifesto for reproducible science

Marcus R. Munafò ✉, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis

Nature Human Behaviour **1**, Article number: 0021 (2017) | Cite this article

Figure 1: Threats to reproducible science.

From: A manifesto for reproducible science



An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

PLOS | MEDICINE

Browse | For Authors | About Us | S

OPEN ACCESS

ESSAY

898,944 VIEWS | 1,119 CITATIONS

Why Most Published Research Findings Are False

John P. A. Ioannidis

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

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

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

How to Make More Published Research True

John P. A. Ioannidis

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

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 Lähteenoja, Mikko Tolonen

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

Heidi Laine Leo Lahti Anne Lehto

Proceedings of the 19th International Academic Mindtrek Conference 2015. ACM New York, NY, USA. ISBN: 978-1-4503-3948-3

Open Science

UNESCO Recommendation on Open Science

The UNESCO Recommendation on Open Science was adopted by the General Conference of UNESCO at its 41st session, in November 2021.

- UNESCO Recommendation on Open Science
  English | Français

Home

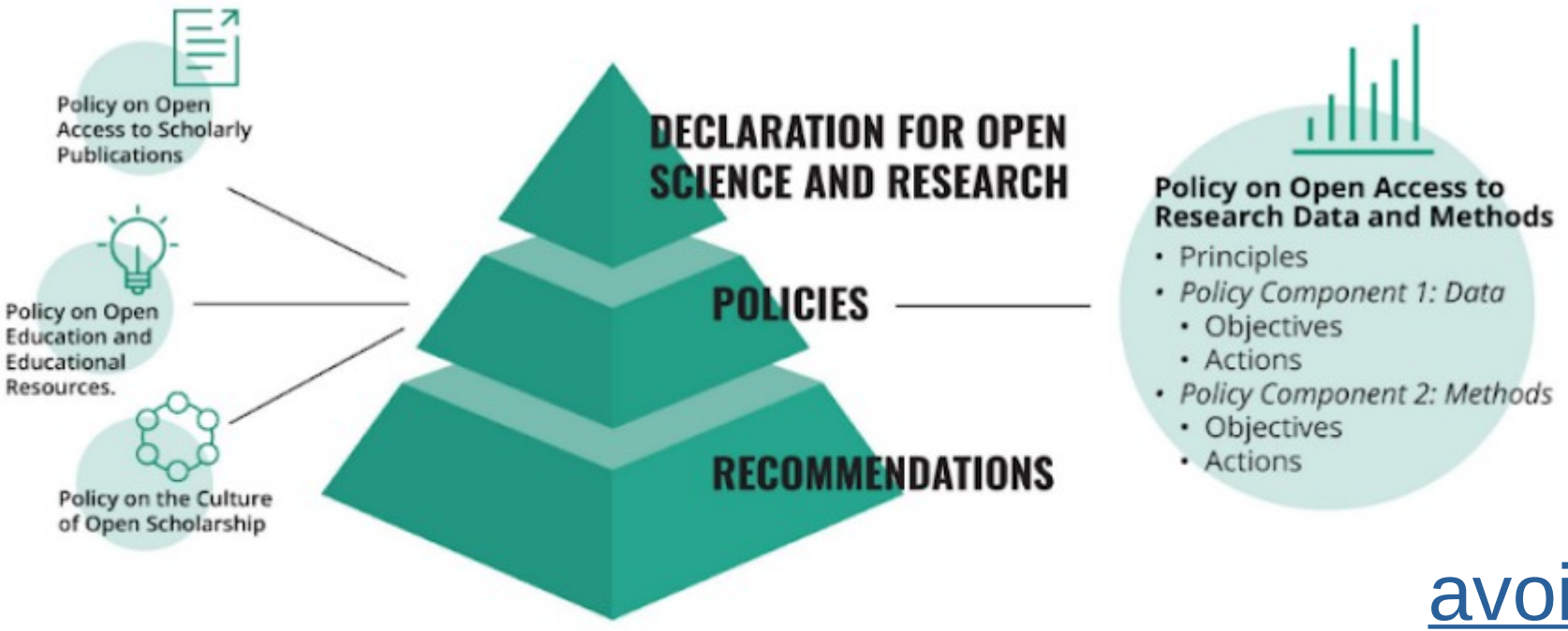
UNESCO Recommendation on Open Science

Multistakeholder Consultations on Open Science

Avoin tiede

National Policy on Open Science
National policy and executive plan by the higher education and research community for 2021-2025

Picture 1. The policy in relation to other national open science documents.



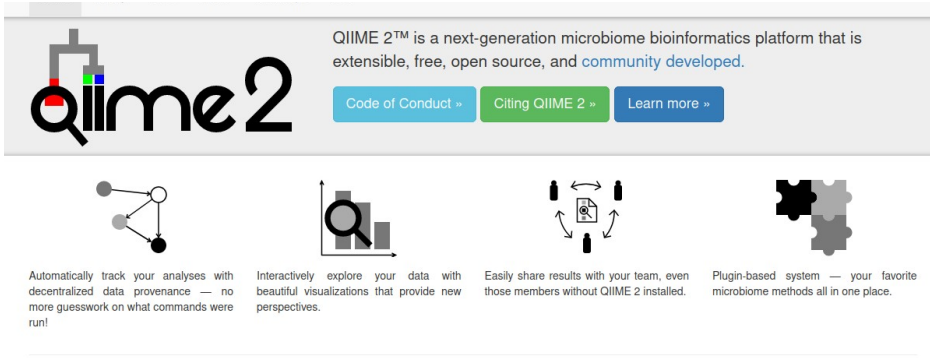
avointiede.fi

PRINCIPLE 1: RESEARCH DATA AND METHODS SHALL BE MANAGED, OPENED AND USED RESPONSIBLY AND APPROPRIATELY.

PRINCIPLE 2: RESEARCHERS HAVE ACCESS TO INFRASTRUCTURES AND SERVICES THAT ENABLE RESPONSIBLE DATA MANAGEMENT, AND THESE ARE DEVELOPED FURTHER IN AN ECONOMICALLY SUSTAINABLE WAY, TAKING INTO ACCOUNT THE RESEARCHERS' NEEDS.

PRINCIPLE 3: THE RESEARCHER'S MERITS IN THE PROMOTION OF GOOD DATA MANAGEMENT, WORK RELATED TO RESEARCH DATA AND METHODS, AND THE APPROPRIATE OPENING OF RESEARCH DATA AND METHODS ARE VALUED AND CAN SUPPORT THE RESEARCHER'S CAREER.

open data science ecosystems



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

PeerJ >

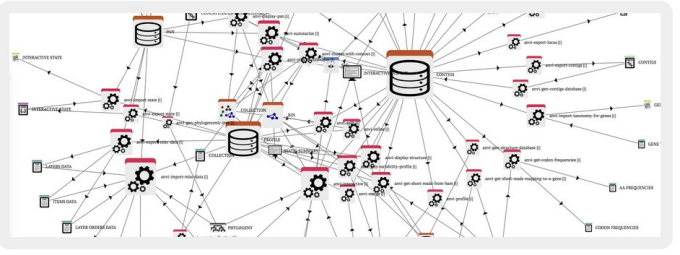
Anvi'o: an advanced analysis and visualization platform for 'omics data

Research article | Bioinformatics | Biotechnology | Computational Biology | Genomics | Microbiology

A. Murat Eren [1,2], Özcan C. Esen [1], Christopher Quince [3], Joseph H. Vineis [1], Hilary G. Morrison [1], Mitchell L. Sogin [1], Tom O. Delmont [1]

Published October 8, 2015

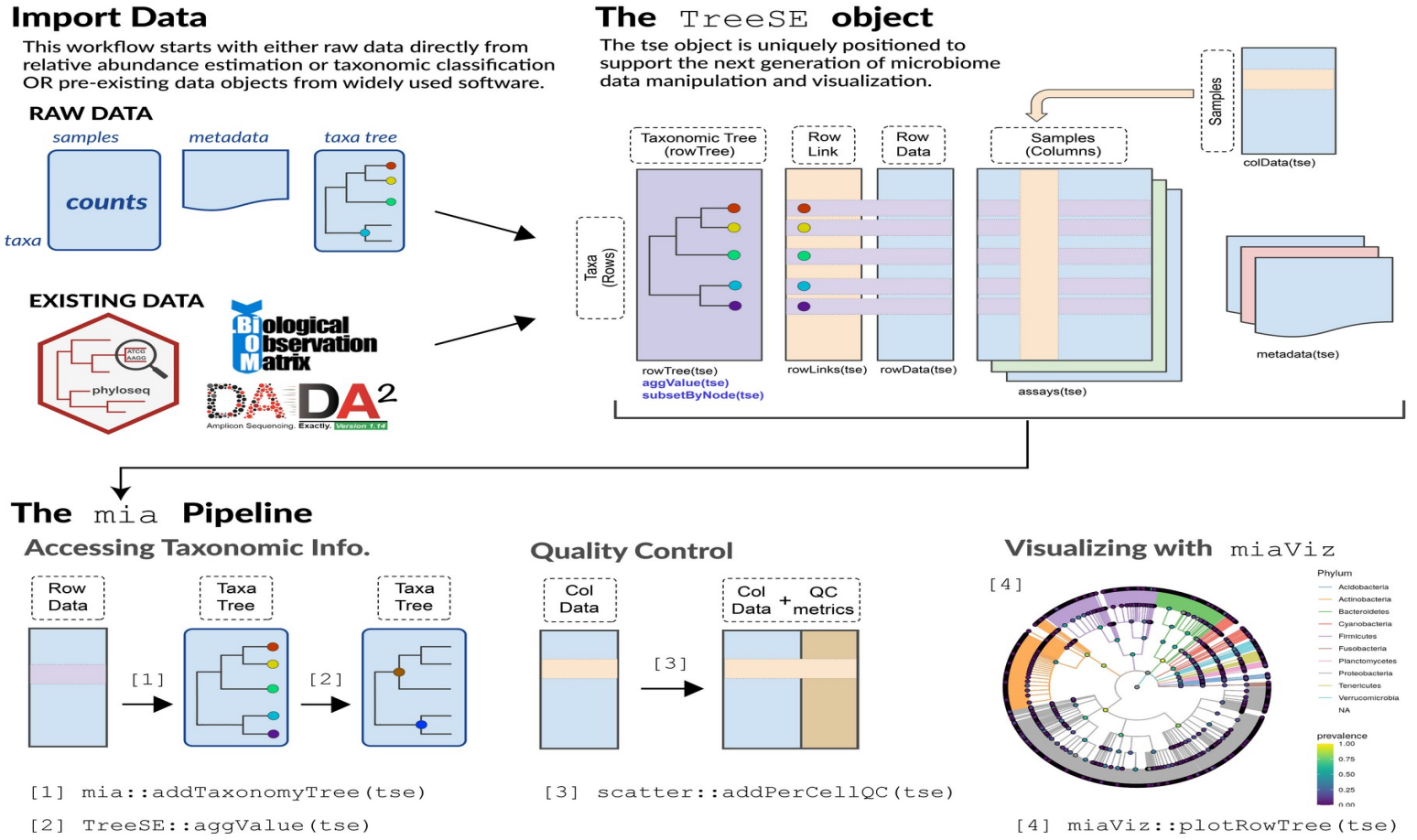
Anvi'o in a nutshell



Anvi'o is an open-source, community-driven analysis and visualization platform for 'omics data.

Example workflow – microbiome.github.io

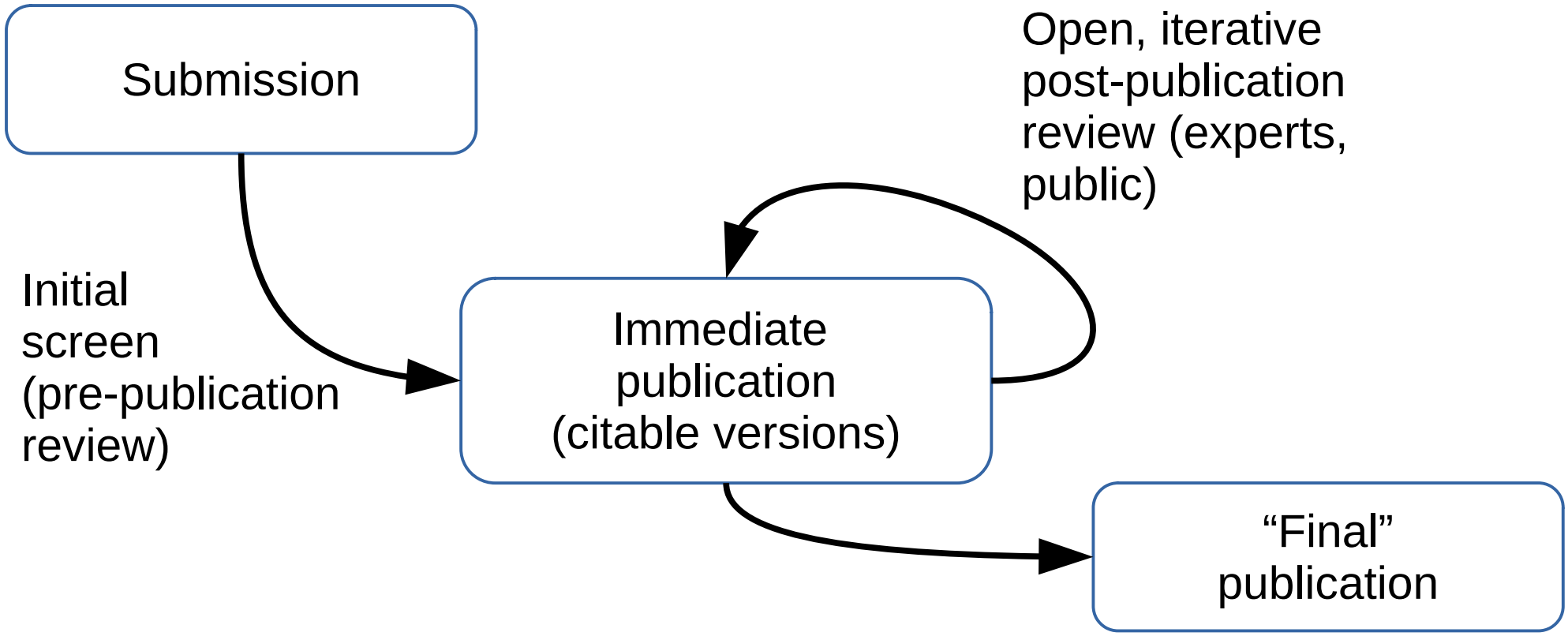
Figure by Domenick Braccia (EuroBioC 2020)

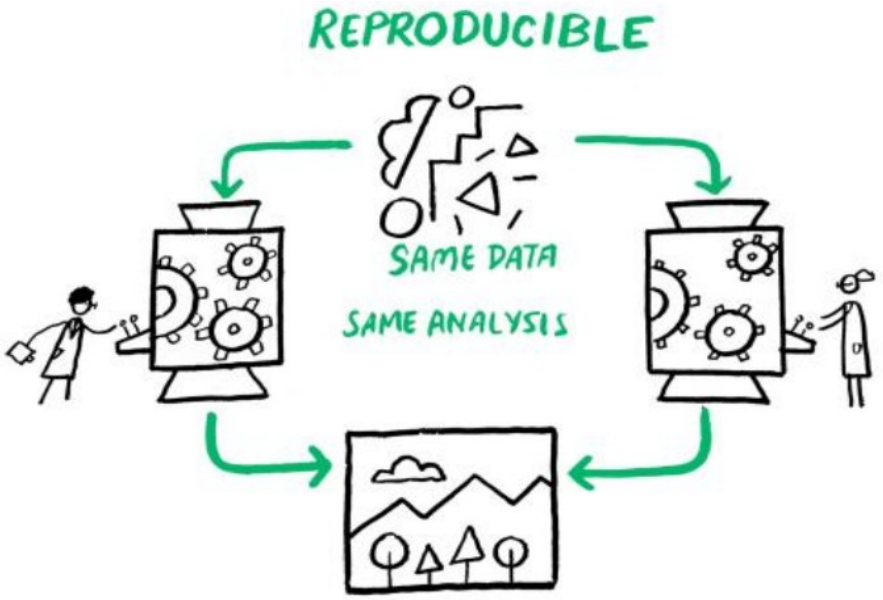


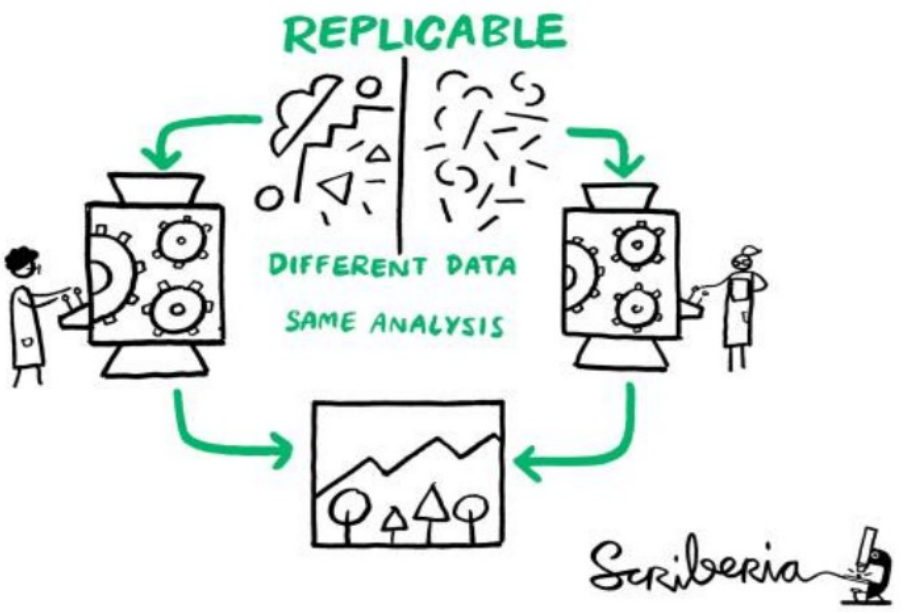
European *Bioconductor* Meeting 2020

- Where: Virtual Conference
- When: 14-18 December 2020
- On twitter: #EuroBioc2020

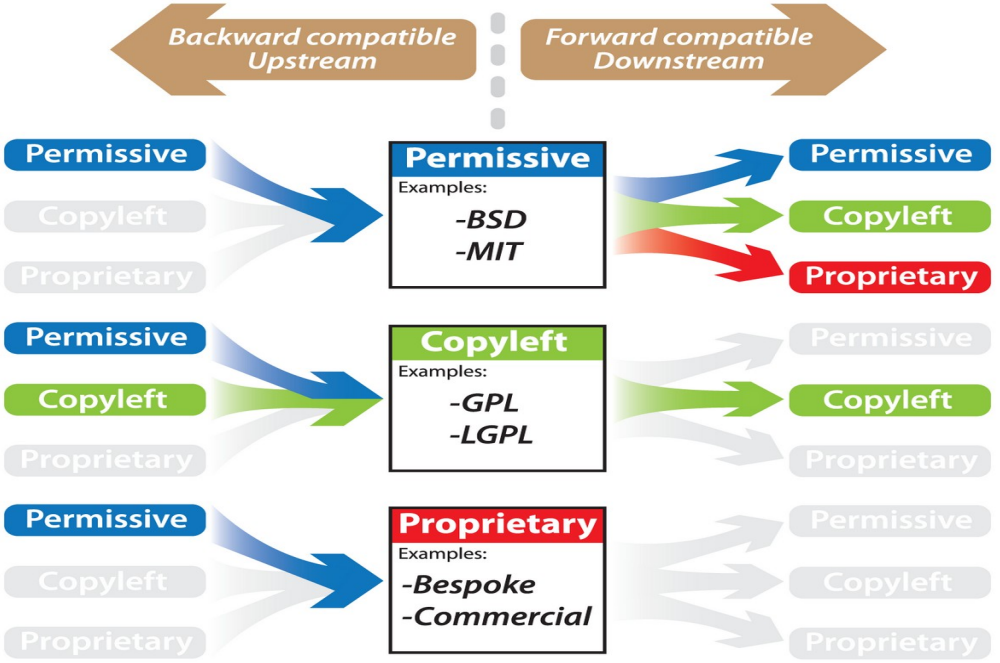
Transparency: pre- vs. post-publication review







Reusability



PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS

EDUCATION

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

PeerJ Computer Science

Software citation principles

Arfon M. Smith[1,*], Daniel S. Katz[2,*], Kyle E. Niemeyer[3,*]
FORCE11 Software Citation Working Group

[1] GitHub, Inc., San Francisco, California, United States
[2] National Center for Supercomputing Applications & Electrical and Computer Department & School of Information Sciences, University of Illinois at Urbana Urbana, Illinois, United States
[3] School of Mechanical, Industrial, and Manufacturing Engineering, Oregon Sta Corvallis, Oregon, United States
* These authors contributed equally to this work.

MIT License

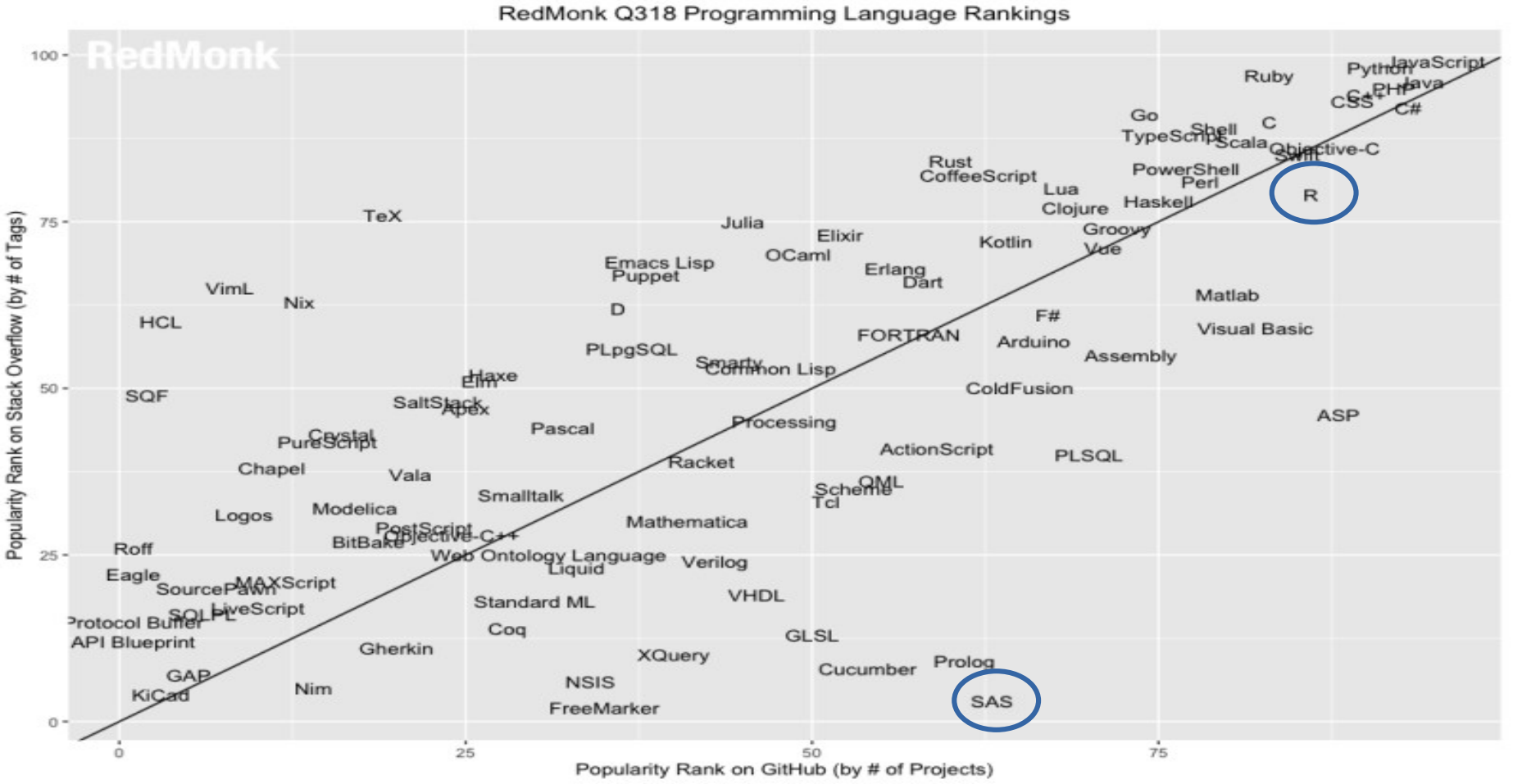
```
Copyright (c) <year> <copyright holders>

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.
```

Varying cultures of open collaboration



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.

transparency

replicability & reproducibility

reusability

collaborative research