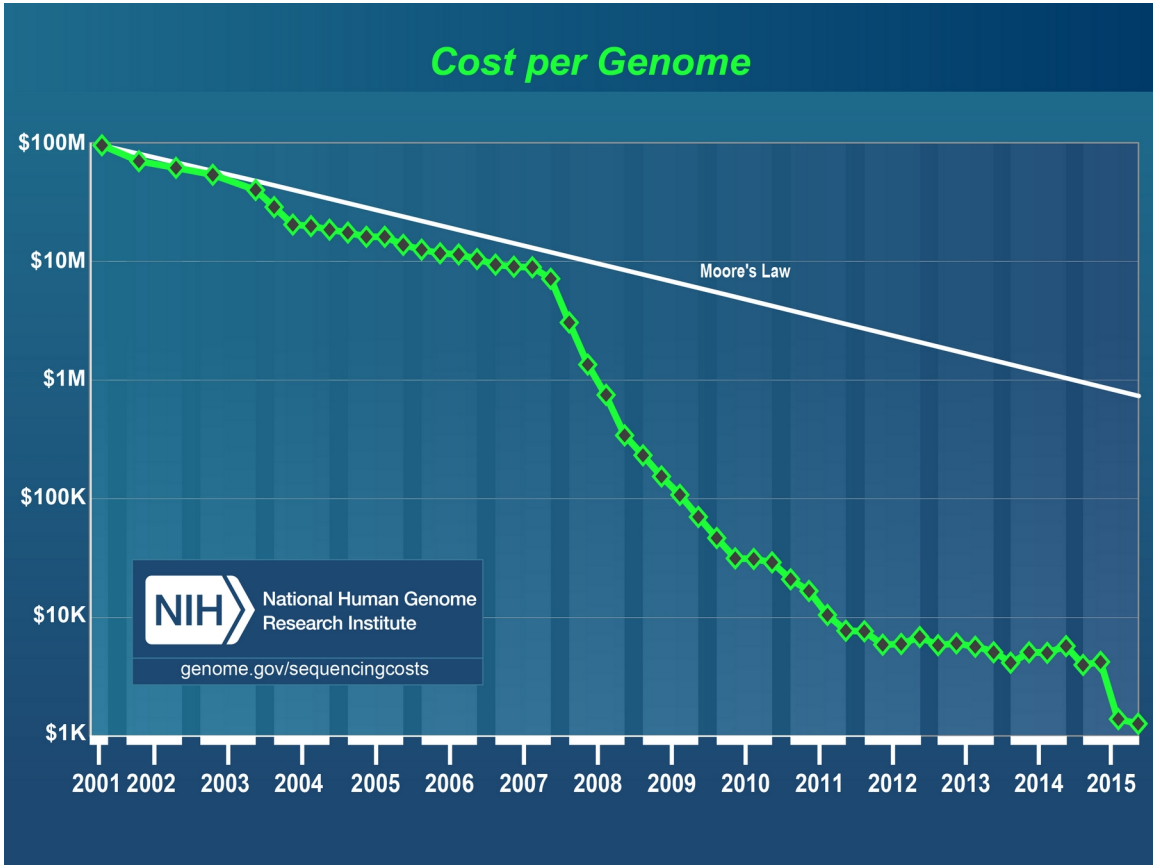
PERSPECTIVE ARTICLE

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

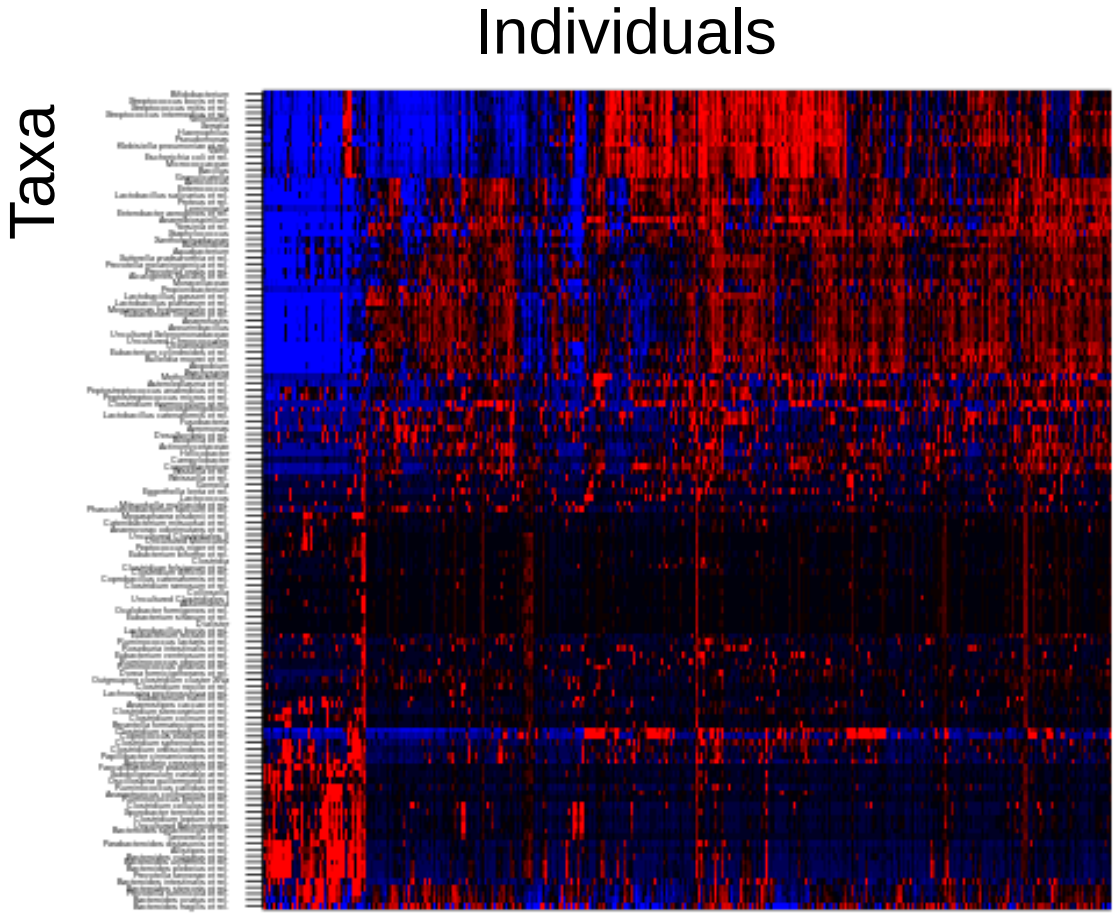
Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions

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



Human Intestinal Tract (HIT)Chip Atlas: 100+ genera ~ 10,000+ samples

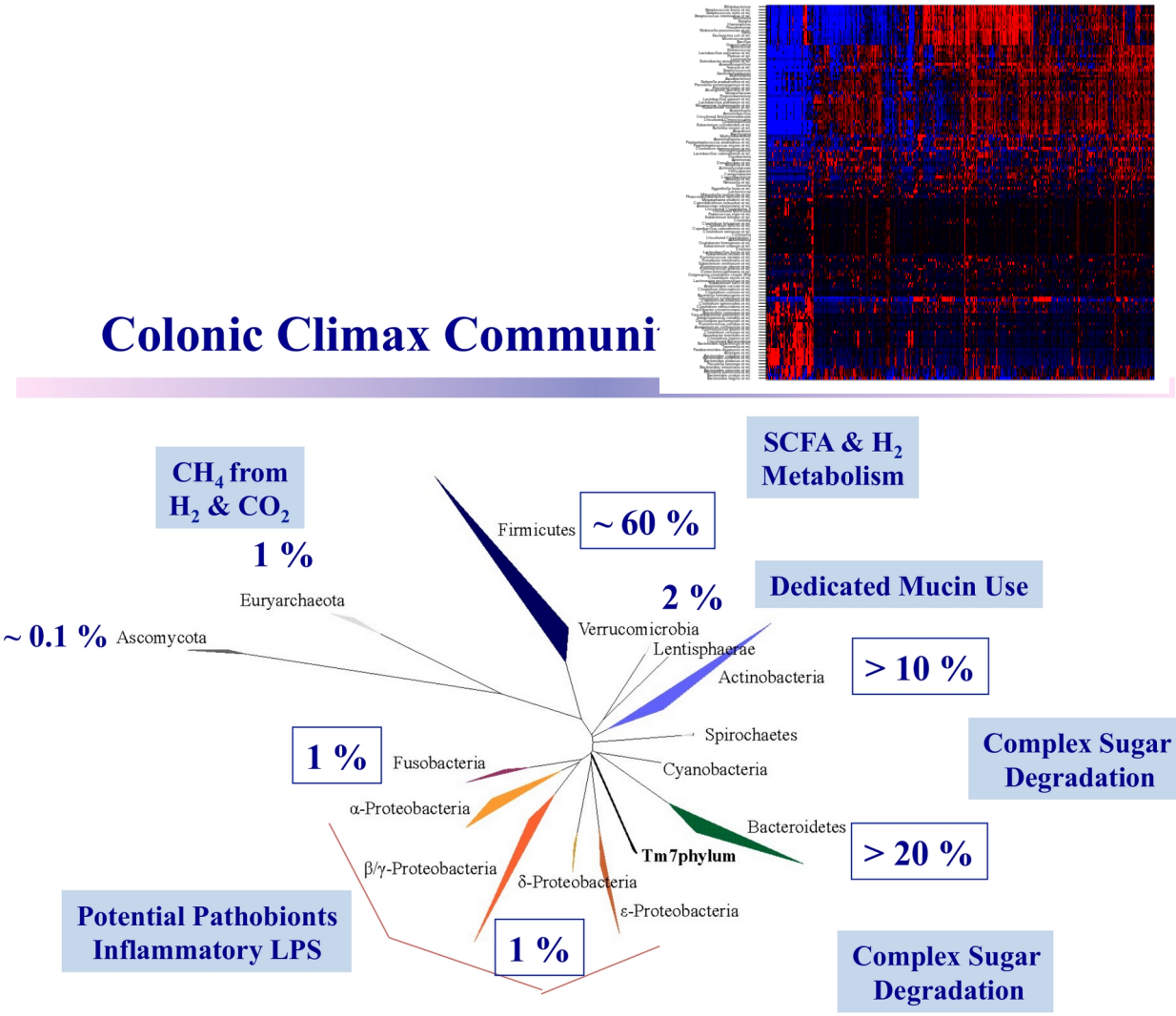
Gut microbiota: 1000 western adults (Lahti *et al.* Nature Comm. 2014)



Standardized – cost efficient – accurate at 0.1% relative abundance
Rajilic-Stojanovic et al. Env. Microbiol. 2009

Special properties of microbiome data

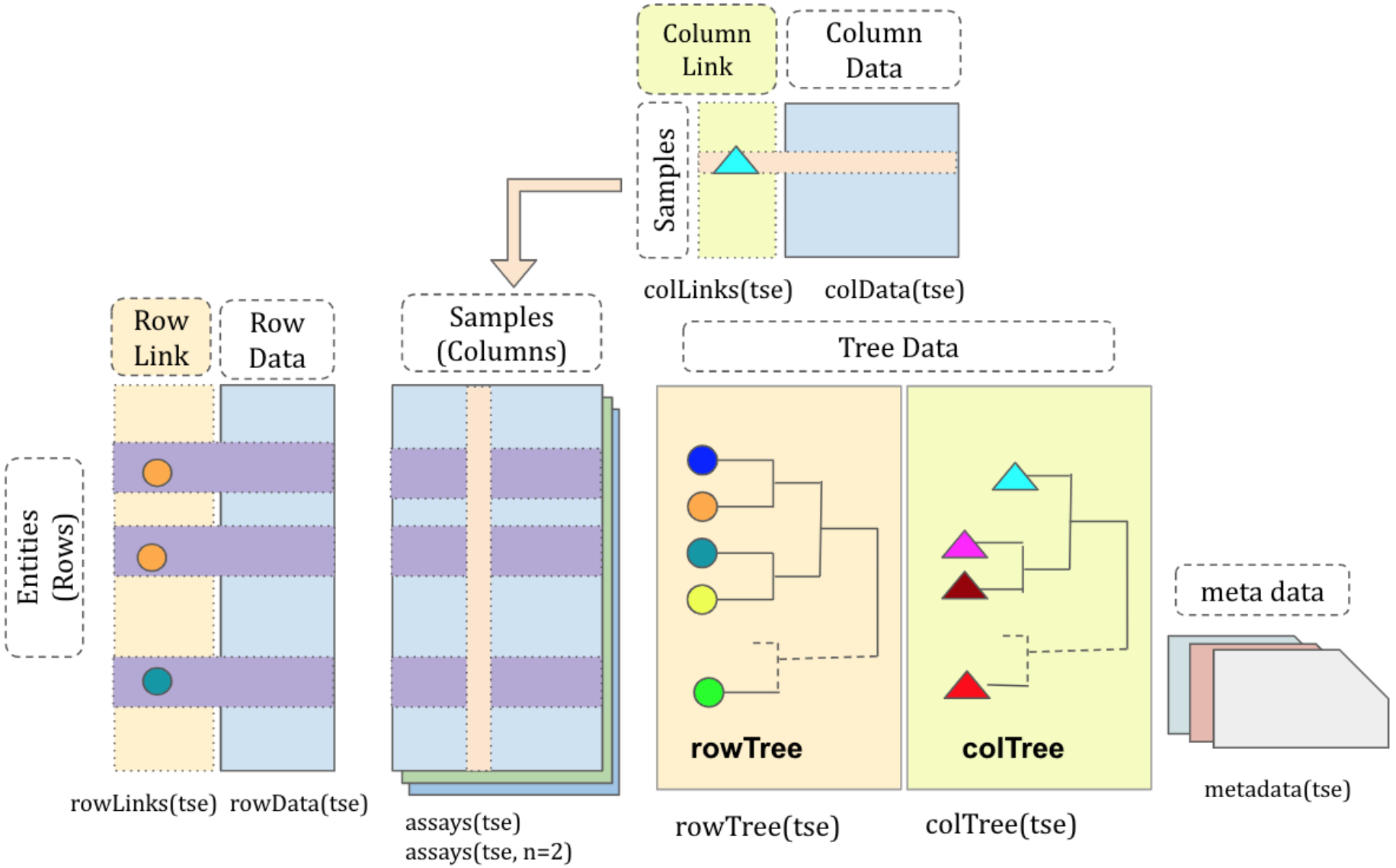
- Sparse
- Compositional
- Non-Gaussian
- Overdispersed
- Discrete
- Complex
- Stochastic
- Multi-level



Zoetendal EG, EE Vaughan & WM de Vos (2006) Mol Microbiol 59: 1639

Lay C, L Rigottier-Gois, K Holmstrom, M Rajilic, EE Vaughan, WM de Vos, MD Collins, R Their, P Namsolleck, M Blaut & J Dore (2005) AEM 71: 4153

Anatomy of TreeSummarizedExperiment

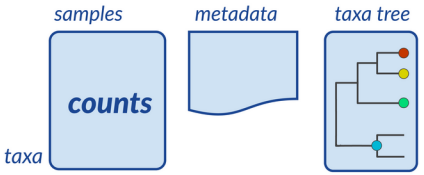


Example workflow

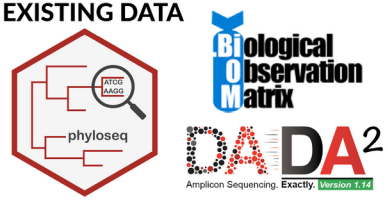
Import Data

This workflow starts with either raw data directly from relative abundance estimation or taxonomic classification OR pre-existing data objects from widely used software.

RAW DATA

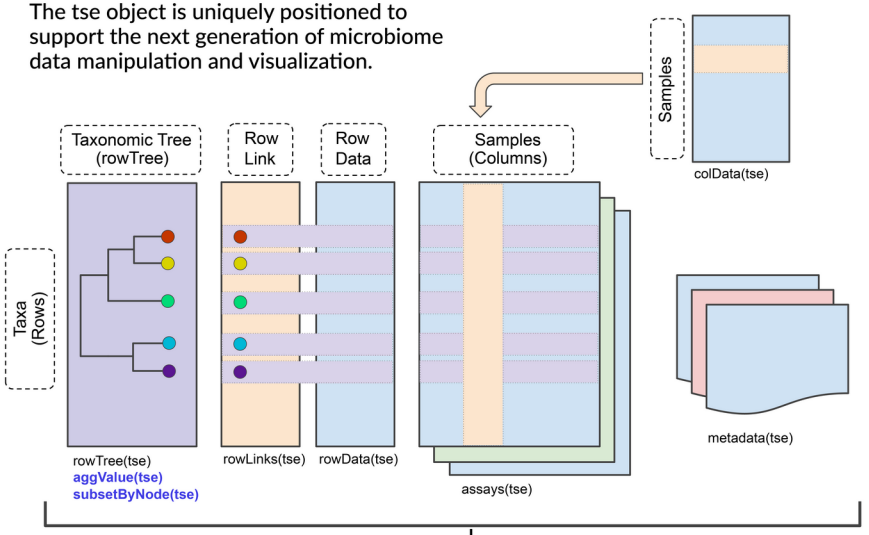


EXISTING DATA



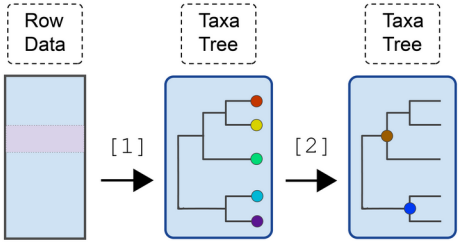
The `TreeSE` object

The tse object is uniquely positioned to support the next generation of microbiome data manipulation and visualization.

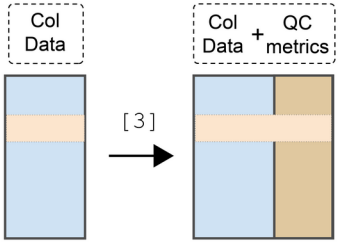


The `mia` Pipeline

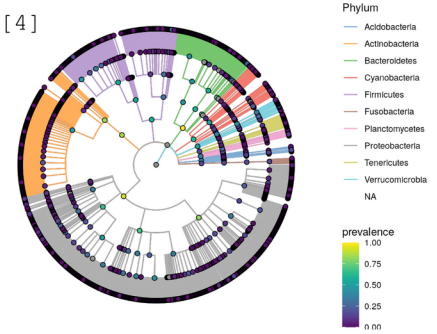
Accessing Taxonomic Info.



Quality Control



Visualizing with `miaViz`




```
[1] mia::addTaxonomyTree(tse)
[2] TreeSE::aggValue(tse)
[3] scatter::addPerCellQC(tse)
[4] miaViz::plotRowTree(tse)
```

Check the poster
F1000 / EuroBioC!

Typical study designs

Case-control studies

Interventions

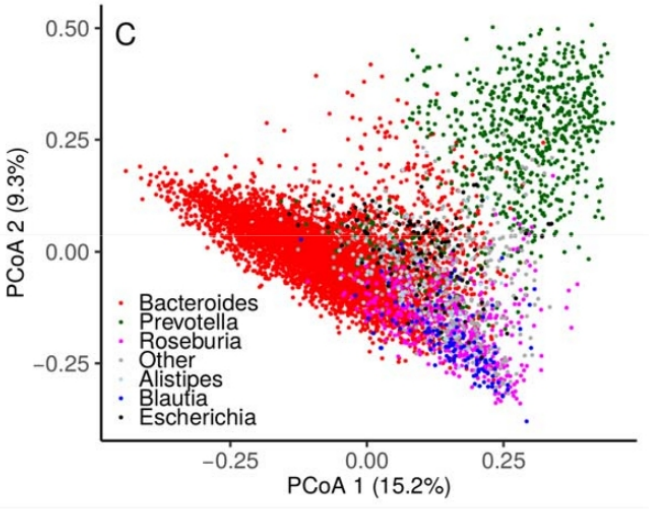
Cross-sectional population cohorts

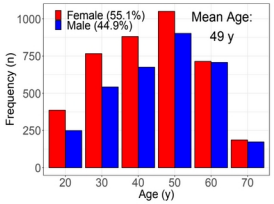
Prospective follow-ups

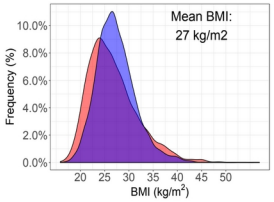
Longitudinal time series

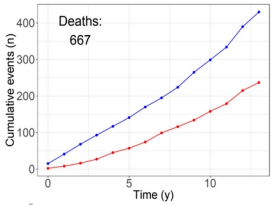
Multi-omics

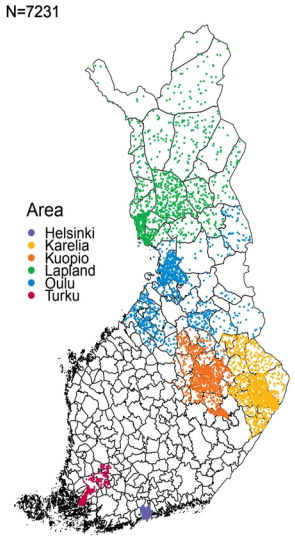
Common study types











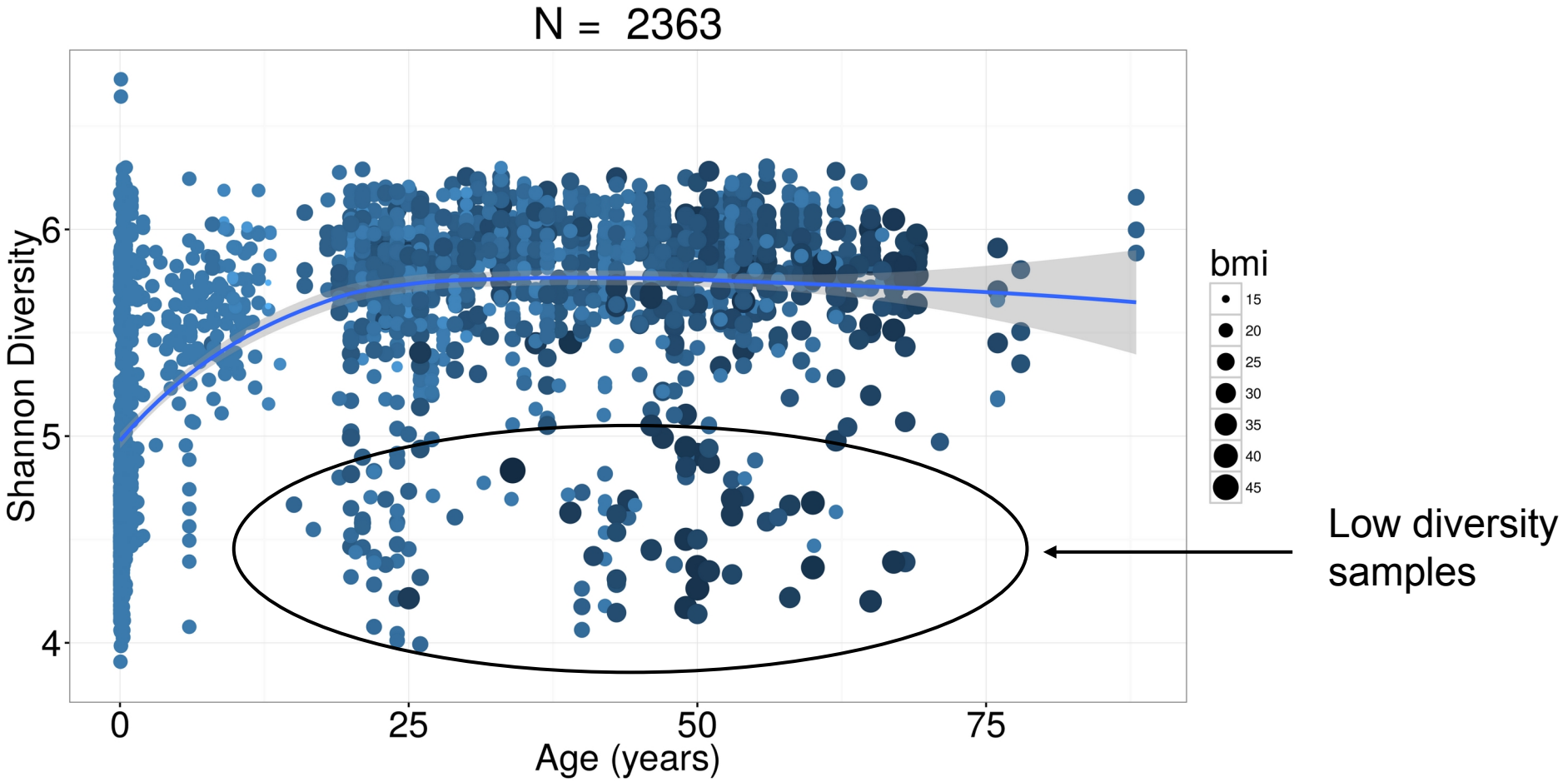
- Data preprocessing
- Case-control studies
- Interventions
- Cross-sectional analysis
- Prospective analysis
- Longitudinal dynamics
- Multi-omics

Lecture: Key concepts

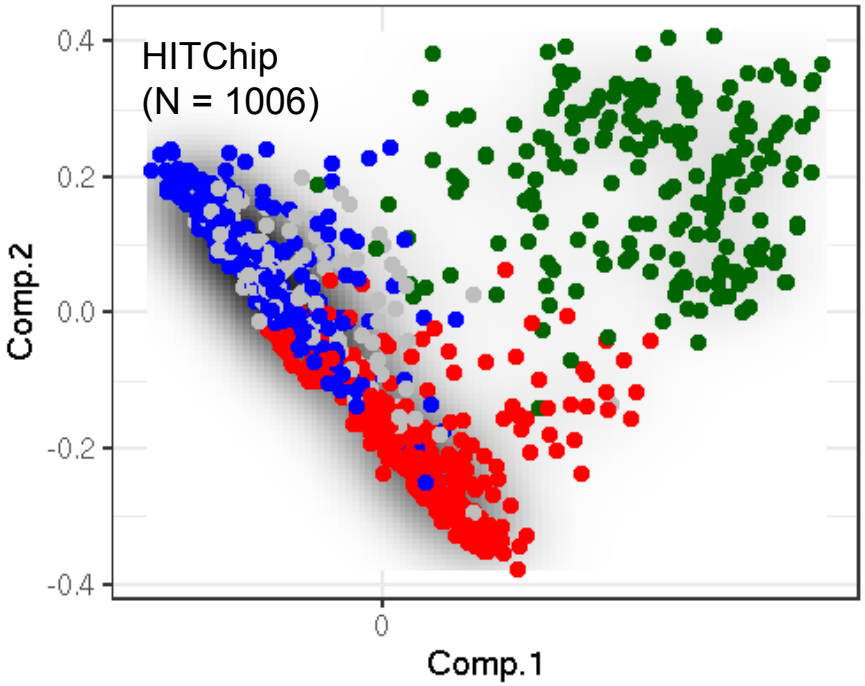
- special properties of microbiome data
- data science workflows

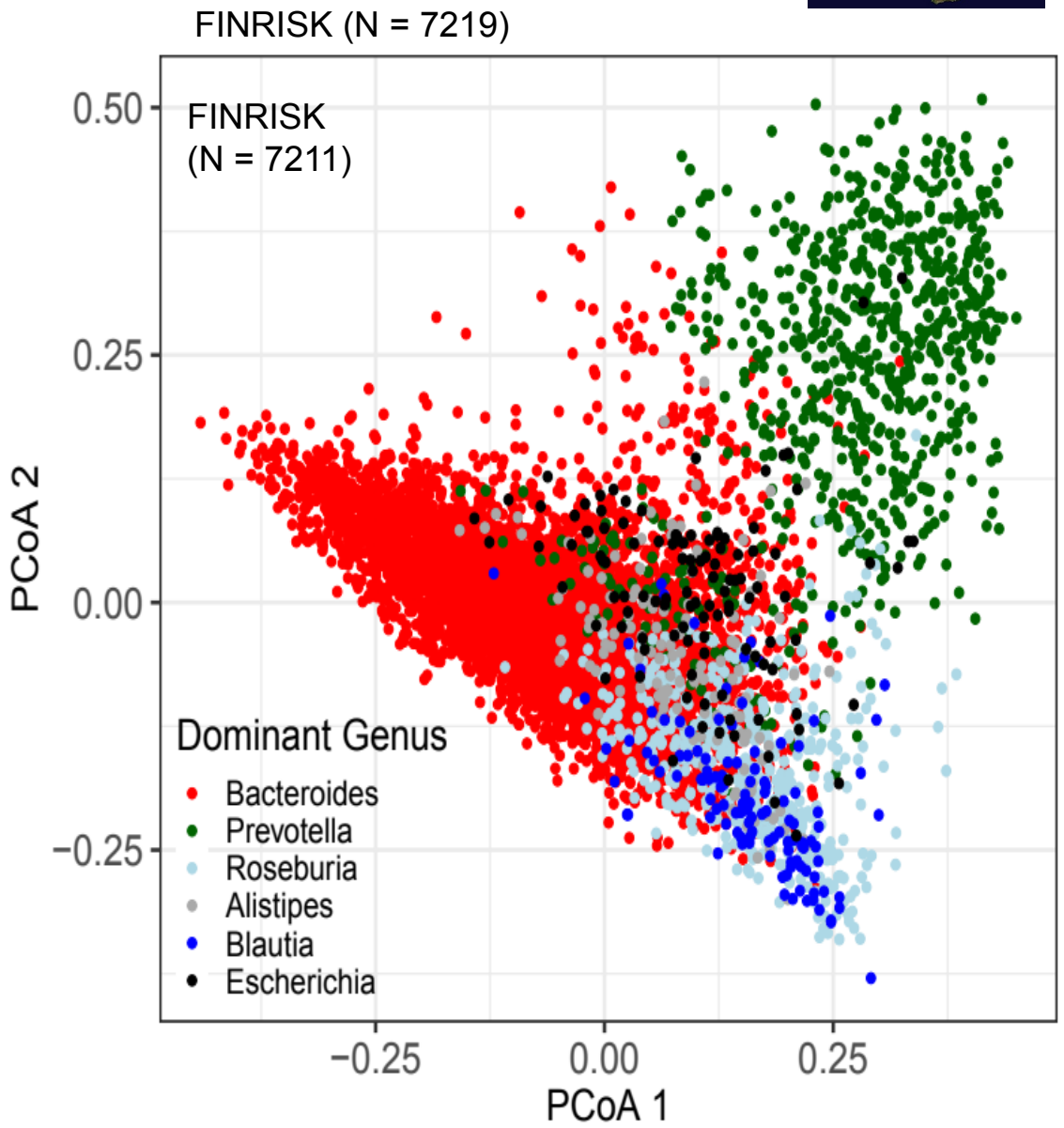
- alpha diversity
- beta diversity
- differential abundance

Alpha diversity & aging
healthy & normal obese subjects



Beta diversity & population landscape



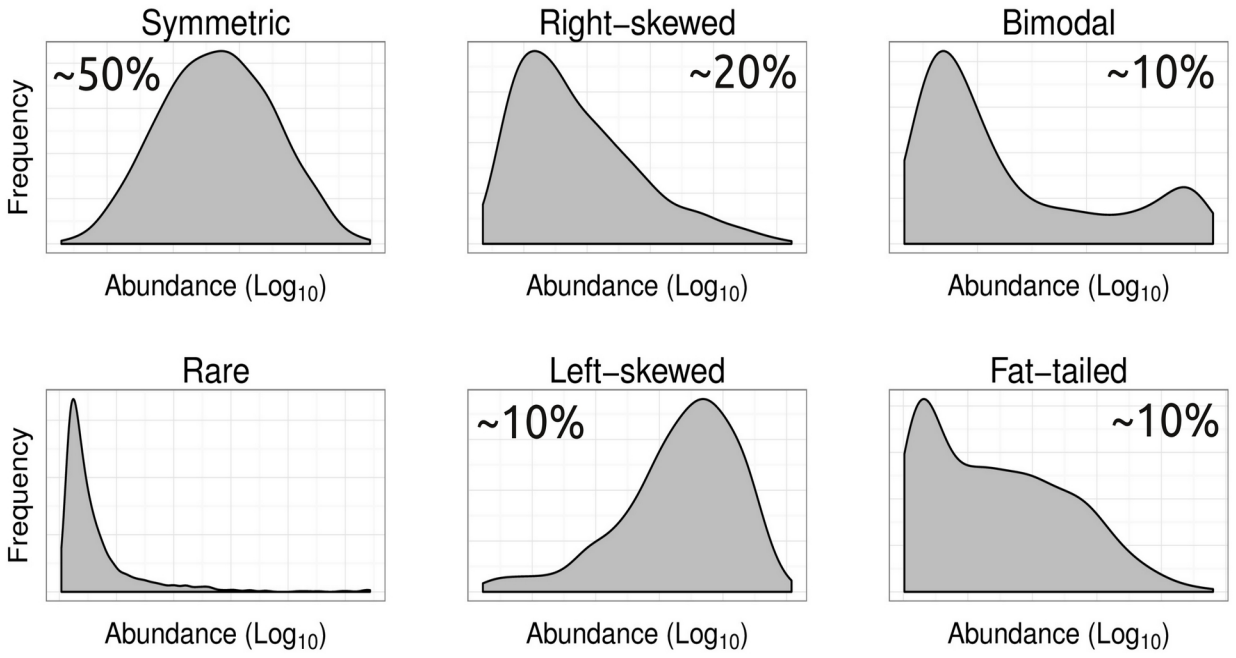


Differential abundance

Standard t-test for two-group comparison?

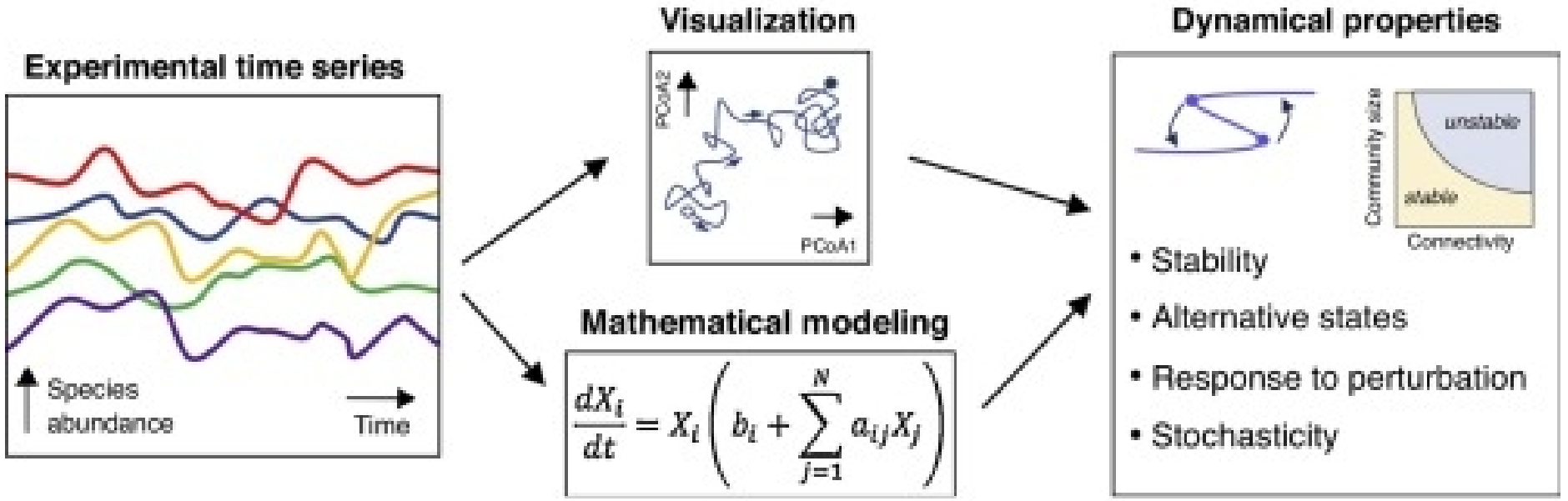
Problems:

- Few replicates
- Non-gaussian, discrete, positive, skewed..
- Multiple testing



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

fastqc/multiqc

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

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

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

>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

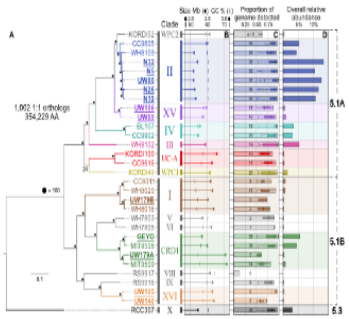
A note on MAGs:
MAGs (metagenome-assembled genomes) are not the same thing as isolate genomes. They are composite representative genomes of closely related genomic lineages.

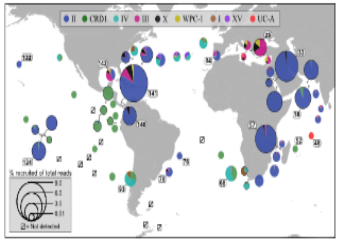
Some tools:
- anvi'o (interactive manual curation of bins; and much more)
- CONCOCT (kmer-based and coverage-based binning; also incorporated in anvi'o)
- COCACOLA (kmer-based, coverage-based, and incorporates paired-read linkage of contigs)
- MetaBAT2 (kmer-based and coverage-based binning tool)
- BinSanity (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)



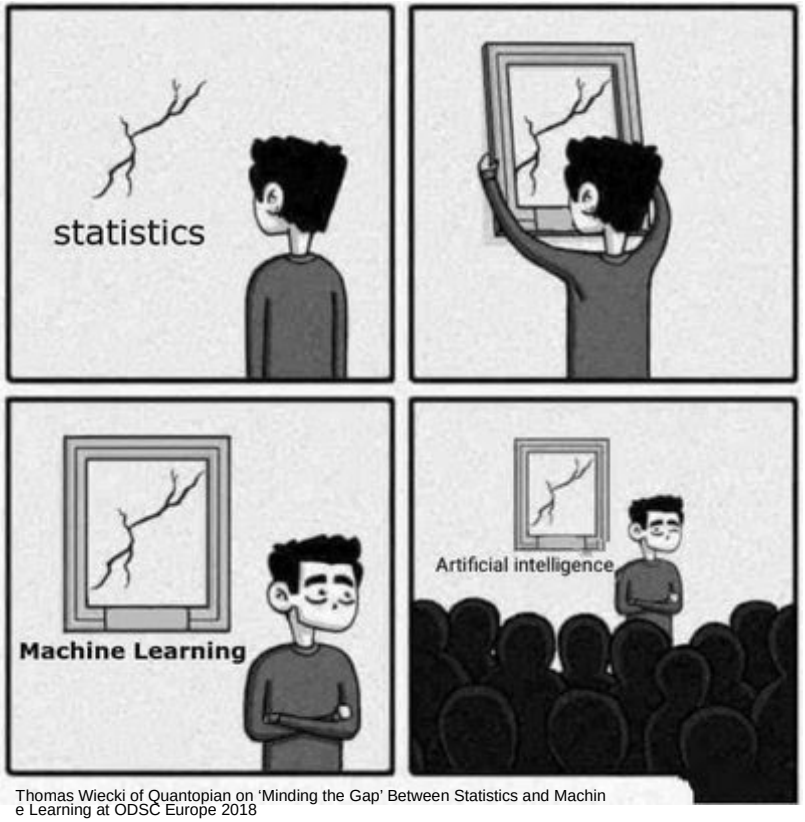


astrobiomike.github.io

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



Thomas Wiecki of Quantopian on 'Minding the Gap' Between Statistics and Machin
e Learning at ODSC Europe 2018

PERSPECTIVE ARTICLE

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

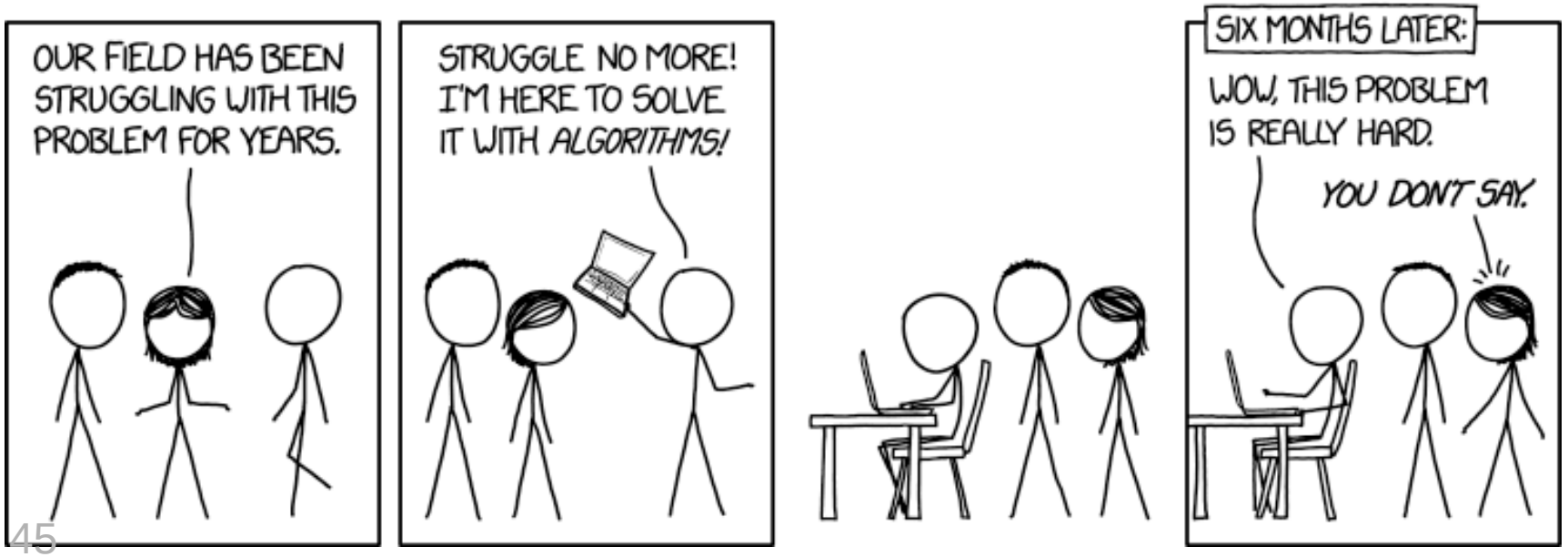
Statistical and Machine Learning Techniques in Human Microbiome Studies: Contemporary Challenges and Solutions

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

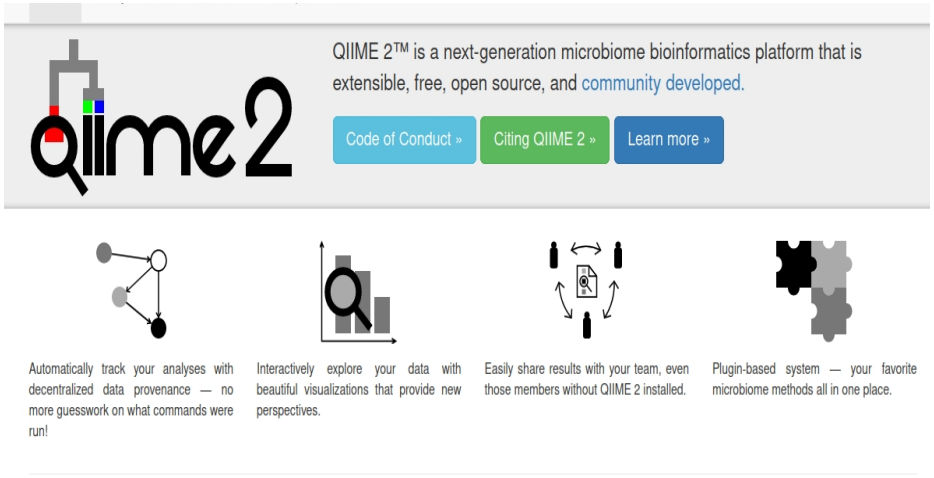
REVIEW ARTICLE

Front. Microbiol., 19 February 2021 | https://doi.org/10.3389/fmicb.2021.634511

Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment



open data science ecosystems



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

PeerJ ›

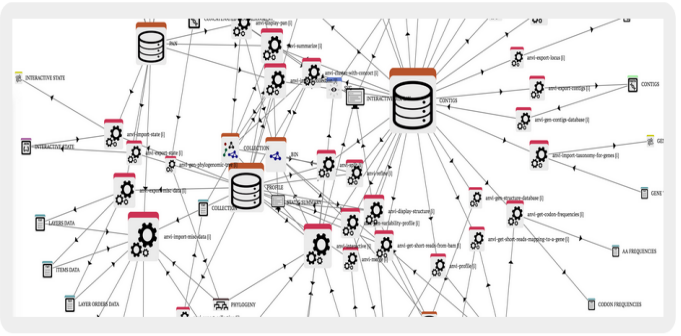
Anvi'o: an advanced analysis and visualization platform for 'omics data

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]

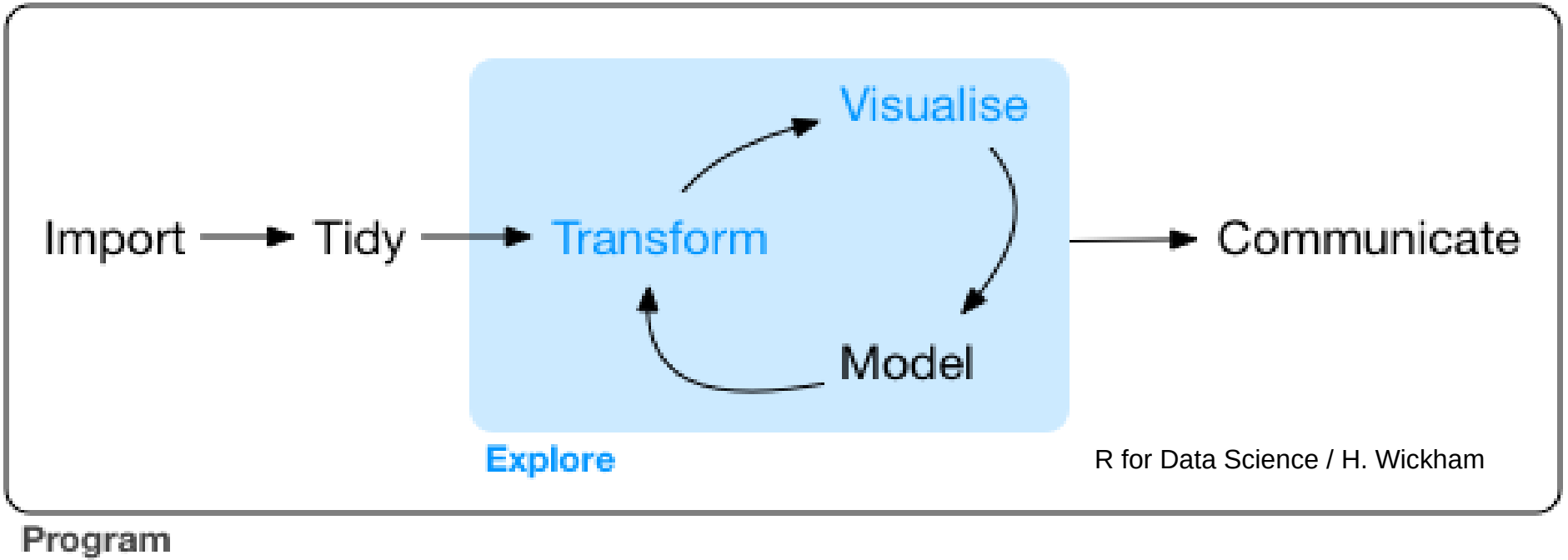
blished October 8, 2015

Anvi'o in a nutshell



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

Data science workflow

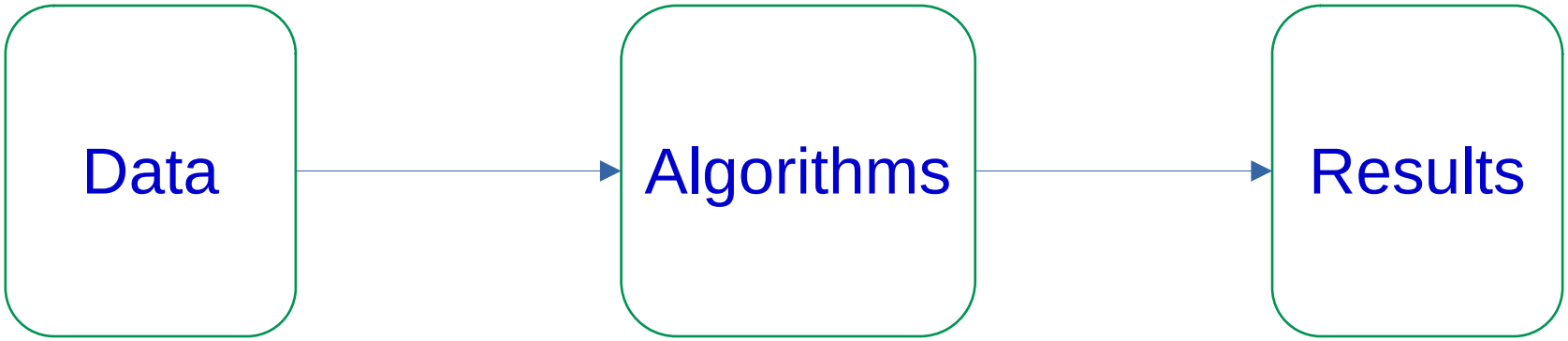


R for Data Science / H. Wickham

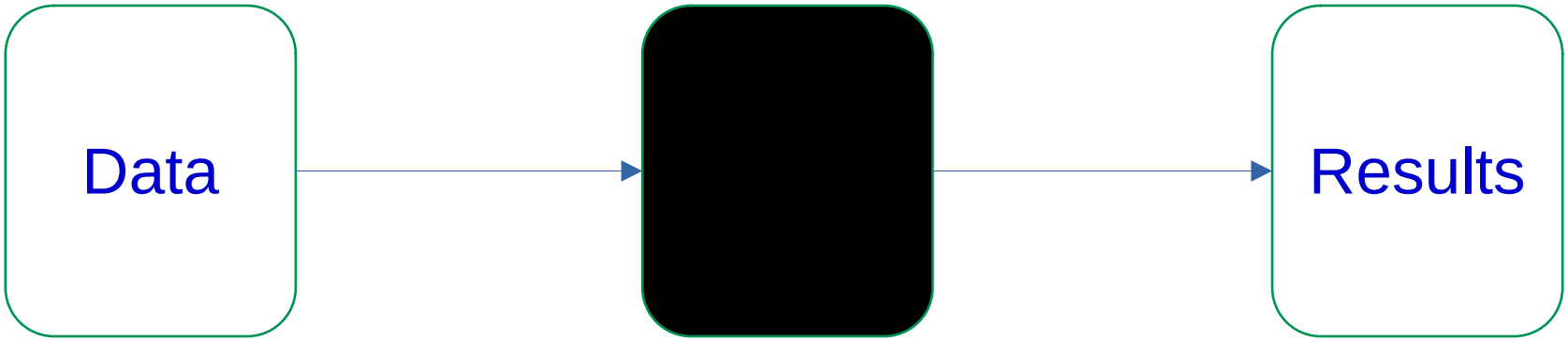
REVISED **Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses [version 2; peer review: 3 approved]**

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.


Data
Algorithms
Results

?



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,*]

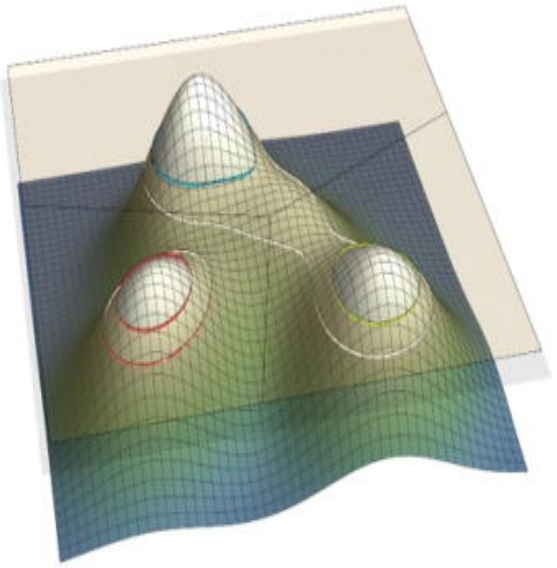
The influence of hidden researcher decisions in applied microeconomics

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.

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

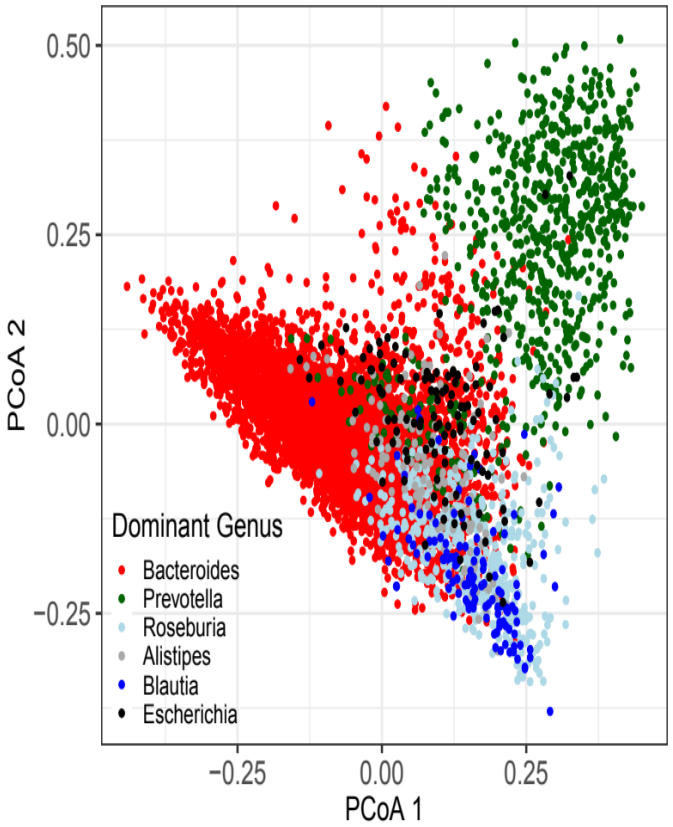
(Dis)similarity
- Eulidean
- Aitchison
- Bray-Curtis
- Jaccard
- weighted Unifrac
- unweighted Unifrac

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

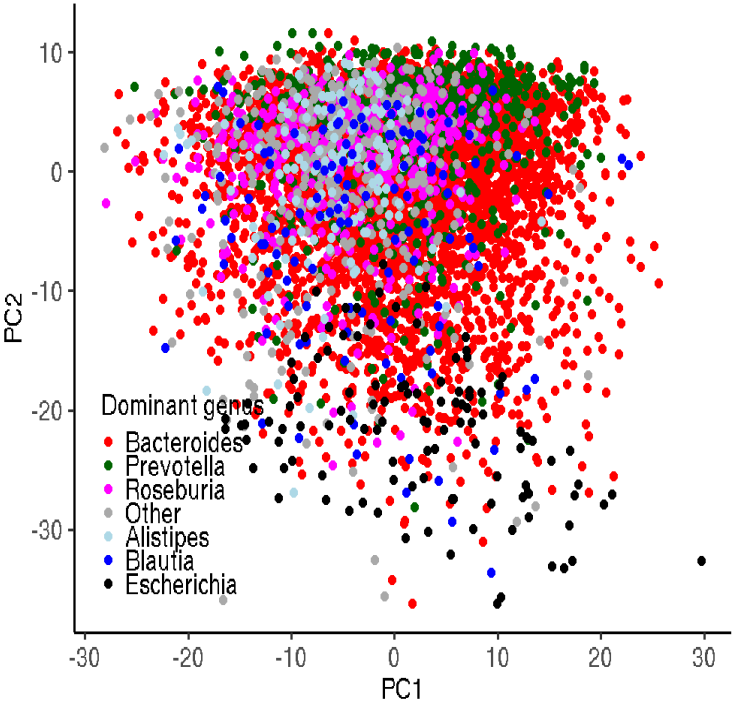
Regulation
- Calinski-Harabasz
- Dirichlet Process
- Silhouette Index
- AIC
- BIC
- DIC

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

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

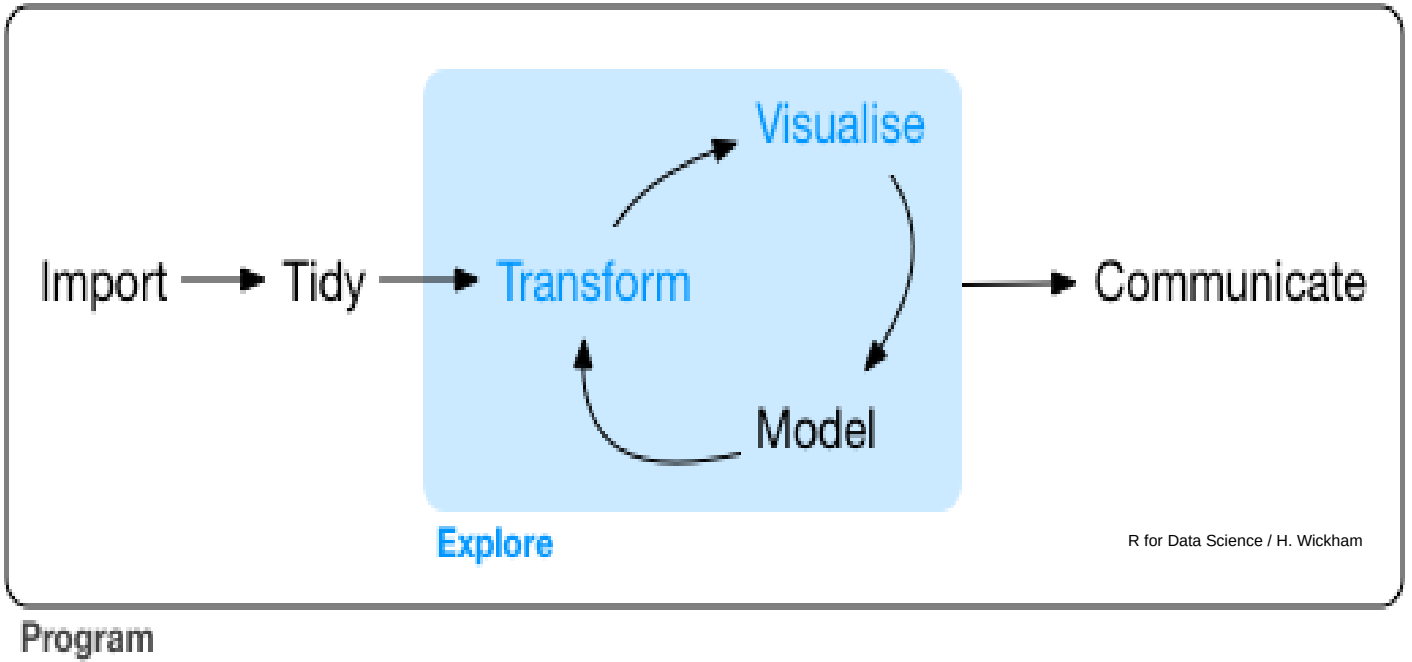
Aaro Salosensaari, Ville Laitinen, Aki Havulinna, Guillaume Meric, Susan Cheng, Markus Perola, Liisa Valsta, Georg Alfthan, Michael Inouye, Jeramie D. Watrous, Tao Long, Rodolfo Salido, Karenina Sanders, Caitriona Brennan, Gregory C. Humphrey, Jon G. Sanders, Mohit Jain, Pekka Jousilahti, Veikko Salomaa, Rob Knight, Leo Lahti, Teemu Niiranen

doi: https://doi.org/10.1101/2019.12.30.19015842

"*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.

Reproducible workflows improve transparency and robustness



Taxonomic level?
- Phylum
- Family
- Order
- Genus
- Species
- Strain..

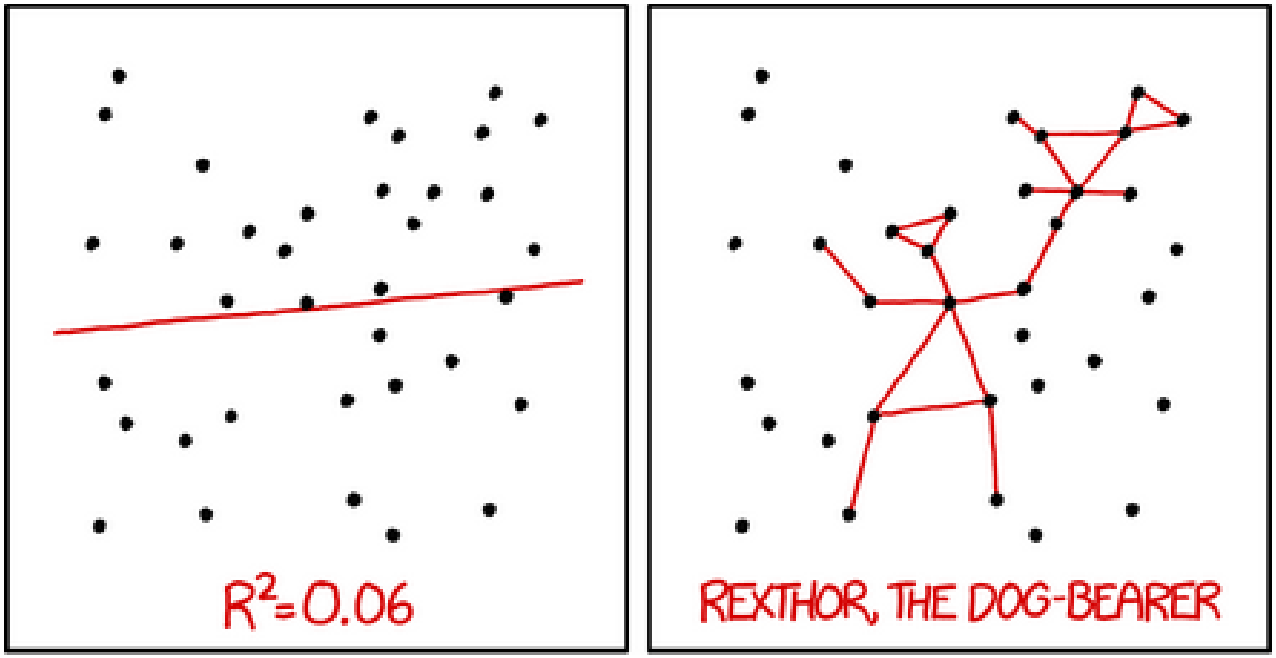
Normalization
- None
- TSS
- CSS
- ILR/ALR/CLR
- phILR
- Hellinger

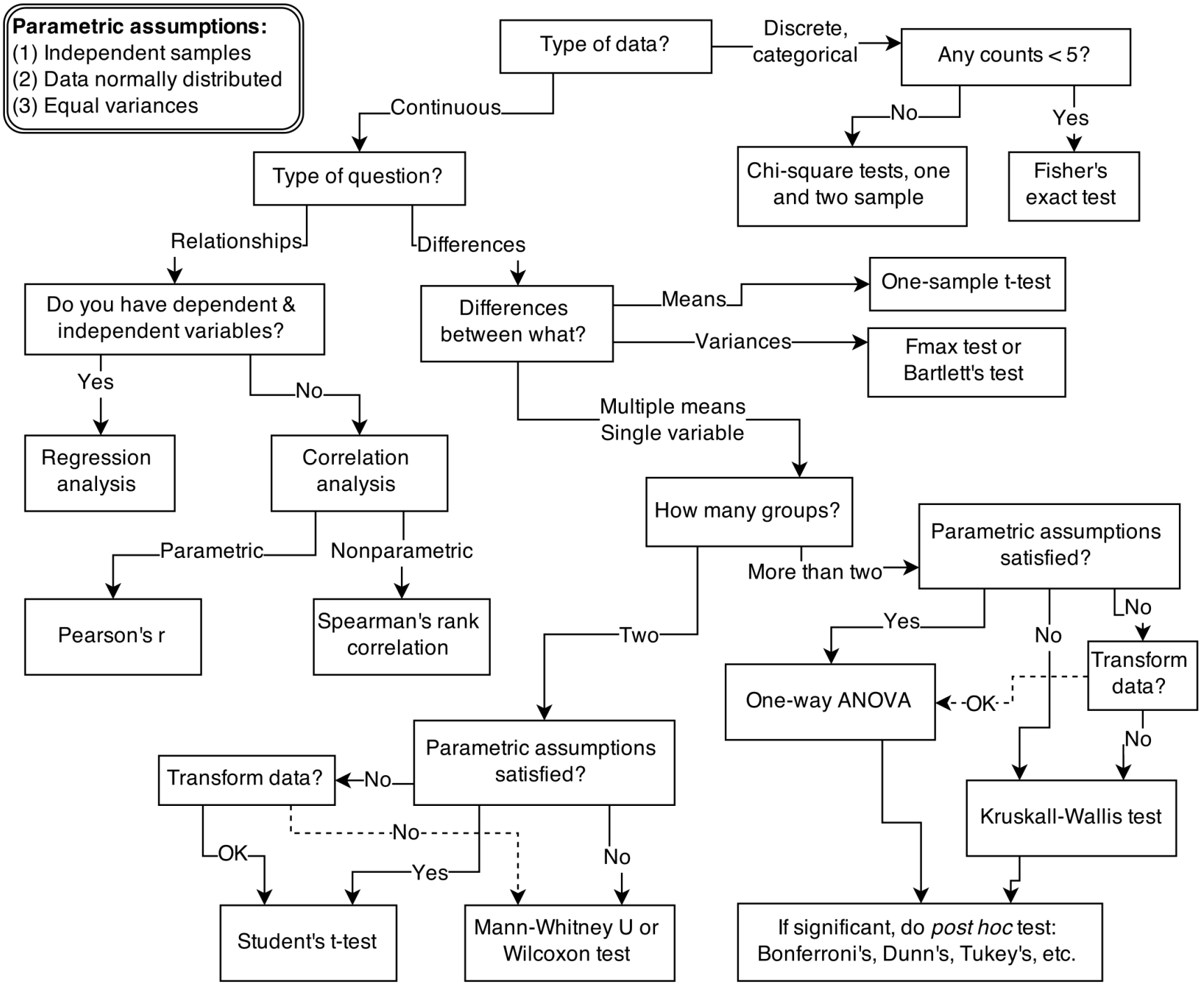
(Dis)similarity?
- Eulidean
- Aitchison
- Bray-Curtis
- Jaccard
- weighted Unifrac
- unweighted Unifrac

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

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

How we choose which model to apply?





Richard McElreath HBES_plenary_2015_page3.png

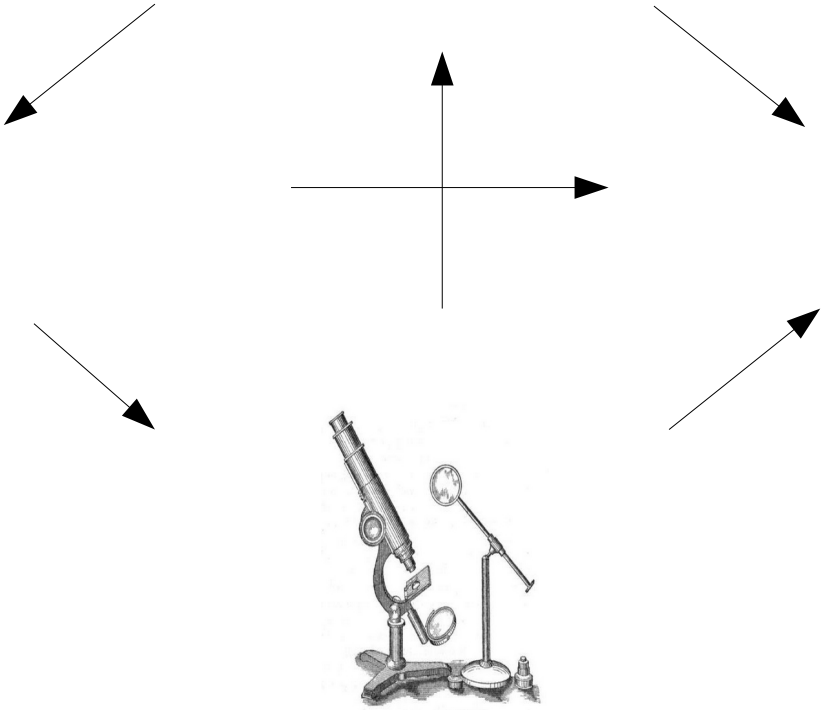
Hypothesis testing vs. hypothesis discovery?

?

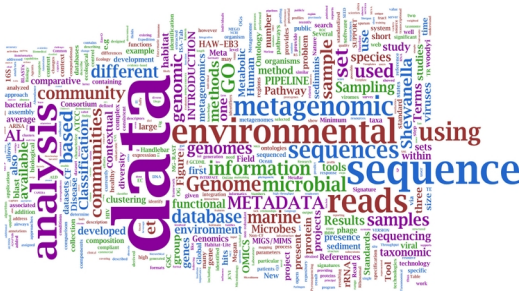
Hypothesis

Method

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



Tools



Data

(barely) not statistically significant (p=0.052)
a barely detectable statistically significant difference (p=0.073)
a borderline significant trend (p=0.09)
a certain trend toward significance (p=0.08)
a clear tendency to significance (p=0.052)
a clear trend (p<0.09)
a clear, strong trend (p=0.09)
a considerable trend toward significance (p=0.069)
a decreasing trend (p=0.09)
a definite trend (p=0.08)
a distinct trend toward significance (p=0.07)
a favorable trend (p=0.09)
a favourable statistical trend (p=0.09)
a little significant (p<0.1)
a margin at the edge of significance (p=0.0608)
a marginal trend (p=0.09)
a marginal trend toward significance (p=0.052)
a marked trend (p=0.07)
a mild trend (p<0.09)
a moderate trend toward significance (p=0.068)
a near-significant trend (p=0.07)
a negative trend (p=0.09)
a nonsignificant trend (p<0.1)
a nonsignificant trend toward significance (p=0.1)
a notable trend (p<0.1)
a numerical increasing trend (p=0.09)
a numerical trend (p=0.09)
a positive trend (p=0.09)
a possible trend (p=0.09)
a possible trend toward significance (p=0.052)
a pronounced trend (p=0.09)
a reliable trend (p=0.058)
a robust trend toward significance (p=0.0503)
a significant trend (p=0.09)
a slight slide towards significance (p<0.20)
a slight tendency toward significance(p<0.08)
a slight trend (p<0.09)
a slight trend toward significance (p=0.098)
a slightly increasing trend (p=0.09)
a small trend (p=0.09)
a statistical trend (p=0.09)
a statistical trend toward significance (p=0.09)
a strong tendency towards statistical significance (p=0.051)
a strong trend (p=0.077)
a strong trend toward significance (p=0.08)
a substantial trend toward significance (p=0.068)
a suggestive trend (p=0.06)
a trend close to significance (p=0.08)
a trend significance level (p=0.08)
a trend that approached significance (p<0.06)
a very slight trend toward significance (p=0.20)
a weak trend (p=0.09)
a weak trend toward significance (p=0.12)
a worrying trend (p=0.07)
all but significant (p=0.055)
almost achieved significance (p=0-065)
almost approached significance (p=0.065)
almost attained significance (p<0.06)
almost became significant (p=0.06)
almost but not quite significant (p=0.06)
almost clinically significant (p<0.10)
almost insignificant (p>0.065)
almost marginally significant (p>0.05)
almost non-significant (p=0.083)
almost reached statistical significance (p=0.06)
almost significant (p=0.06)
almost significant tendency (p=0.06)
almost statistically significant (p=0.06)
an adverse trend (p=0.10)
an apparent trend (p=0.286)
an associative trend (p=0.09)
an elevated trend (p<0.05)
an encouraging trend (p<0.1)
an established trend (p<0.10)
an evident trend (p=0.13)
an expected trend (p=0.08)
an important trend (p=0.066)
an increasing trend (p<0.09)
an interesting trend (p=0.1)
an inverse trend toward significance (p=0.06)
an observed trend (p=0.06)
an obvious trend (p=0.06)
an overall trend (p=0.2)
an unexpected trend (p=0.09)
an unexplained trend (p=0.09)
an unfavorable trend (p<0.10)
appeared to be marginally significant (p<0.10)
approached acceptable levels of statistical significance (p=0.054)
approached but did not quite achieve significance (p>0.05)
approached but fell short of significance (p=0.07)
approached conventional levels of significance (p<0.10)
approached near significance (p=0.06)
approached our criterion of significance (p>0.08)
approached significant (p=0.11)
approached the borderline of significance (p=0.07)
approached the level of significance (p=0.09)
approached trend levels of significance (p0.05)
approached, but did reach, significance (p=0.065)
approaches but fails to achieve a customary level of statistical significance (p=0.154)
approaches statistical significance (p>0.06)
approaching a level of significance (p=0.089)
approaching an acceptable significance level (p=0.056)
approaching borderline significance (p=0.08)
approaching borderline statistical significance (p=0.07)
approaching but not reaching significance (p=0.53)
approaching clinical significance (p=0.07)
approaching close to significance (p<0.1)
approaching conventional significance levels (p=0.06)
approaching conventional statistical significance (p=0.06)
approaching formal significance (p=0.1052)
approaching independent prognostic significance (p=0.08)
approaching marginal levels of significance p<0.107)
approaching marginal significance (p=0.064)
approaching more closely significance (p=0.06)
approaching our preset significance level (p=0.076)
approaching prognostic significance (p=0.052)
approaching significance (p=0.09)
approaching the traditional significance level (p=0.06)
approaching to statistical significance (p=0.075)
approaching, although not reaching, significance (p=0.08)
approaching, but not reaching, significance (p<0.09)
approximately significant (p=0.053)
approximating significance (p=0.09)
arguably significant (p=0.07)
as good as significant (p=0.0502)
at the brink of significance (p=0.06)
at the cusp of significance (p=0.06)
at the edge of significance (p=0.055)
at the limit of significance (p=0.054)
at the limits of significance (p=0.053)
at the margin of significance (p=0.056)
at the margin of statistical significance (p<0.07)
at the verge of significance (p=0.058)
at the very edge of significance (p=0.053)
barely below the level of significance (p=0.06)
barely escaped statistical significance (p=0.07)
barely escapes being statistically significant at the 5% risk level (0.1>p>0.05)
barely failed to attain statistical significance (p=0.067)
barely fails to attain statistical significance at conventional levels (p<0.10
barely insignificant (p=0.075)
barely missed statistical significance (p=0.051)
barely missed the commonly acceptable significance level (p<0.053)
barely outside the range of significance (p=0.06)
barely significant (p=0.07)
below (but verging on) the statistical significant level (p>0.05)
better trends of improvement (p=0.056)
bordered on a statistically significant value (p=0.06)
bordered on being significant (p>0.07)
bordered on being statistically significant (p=0.0502)
bordered on but was not less than the accepted level of significance (p>0.05)
bordered on significant (p=0.09)
borderline conventional significance (p=0.051)
borderline level of statistical significance (p=0.053)
borderline significant (p=0.09)
borderline significant trends (p=0.099)
close to a marginally significant level (p=0.06)
close to being significant (p=0.06)
close to being statistically significant (p=0.055)
close to borderline significance (p=0.072)
close to the boundary of significance (p=0.06)
close to the level of significance (p=0.07)
close to the limit of significance (p=0.17)
close to the margin of significance (p=0.055)
close to the margin of statistical significance (p=0.075)
closely approaches the brink of significance (p=0.07)
closely approaches the statistical significance (p=0.0669)
closely approximating significance (p>0.05)
closely not significant (p=0.06)
closely significant (p=0.058)
close-to-significant (p=0.09)
did not achieve conventional threshold levels of statistical significance (p=0.08)
did not exceed the conventional level of statistical significance (p<0.08)
did not quite achieve acceptable levels of statistical significance (p=0.054)
did not quite achieve significance (p=0.076)
did not quite achieve the conventional levels of significance (p=0.052)
did not quite achieve the threshold for statistical significance (p=0.08)
did not quite attain conventional levels of significance (p=0.07)
did not quite reach a statistically significant level (p=0.108)
did not quite reach conventional levels of statistical significance (p=0.079)
did not quite reach statistical significance (p=0.063)
did not reach the traditional level of significance (p=0.10)
did not reach the usually accepted level of clinical significance (p=0.07)
difference was apparent (p=0.07)
direction heading towards significance (p=0.10)
does not appear to be sufficiently significant (p>0.05)
does not narrowly reach statistical significance (p=0.06)
does not reach the conventional significance level (p=0.098)
effectively significant (p=0.051)
equivocal significance (p=0.06)
essentially significant (p=0.10)
extremely close to significance (p=0.07)
failed to reach significance on this occasion (p=0.09)
failed to reach statistical significance (p=0.06)
fairly close to significance (p=0.065)
fairly significant (p=0.09)
falls just short of standard levels of statistical significance (p=0.06)
fell (just) short of significance (p=0.08)
fell barely short of significance (p=0.08)
fell just short of significance (p=0.07)
fell just short of statistical significance (p=0.12)
fell just short of the traditional definition of statistical significance (p=0.051)
fell marginally short of significance (p=0.07)
fell narrowly short of significance (p=0.0623)
fell only marginally short of significance (p=0.0879)
fell only short of significance (p=0.06)
fell short of significance (p=0.07)
fell slightly short of significance (p>0.0167)
fell somewhat short of significance (p=0.138)
felt short of significance (p=0.07)
flirting with conventional levels of significance (p>0.1)
heading towards significance (p=0.086)
highly significant (p=0.09)
hint of significance (p>0.05)
hovered around significance (p = 0.061)
hovered at nearly a significant level (p=0.058)
hovering closer to statistical significance (p=0.076)
hovers on the brink of significance (p=0.055)
in the edge of significance (p=0.059)
in the verge of significance (p=0.06)
inconclusively significant (p=0.070)
indeterminate significance (p=0.08)
indicative significance (p=0.08)
is just outside the conventional levels of significance
just about significant (p=0.051)
just above the arbitrary level of significance (p=0.07)
just above the margin of significance (p=0.053)
just at the conventional level of significance (p=0.05001)
just barely below the level of significance (p=0.06)
just barely failed to reach significance (p<0.06)
just barely insignificant (p=0.11)
just barely statistically significant (p=0.054)
just beyond significance (p=0.06)
just borderline significant (p=0.058)
just escaped significance (p=0.07)
just failed significance (p=0.057)
just failed to be significant (p=0.072)
just failed to reach statistical significance (p=0.06)
just failing to reach statistical significance (p=0.06)
just fails to reach conventional levels of statistical significance (p=0.07)
just lacked significance (p=0.053)
just marginally significant (p=0.0562)
just missed being statistically significant (p=0.06)
just missing significance (p=0.07)
just on the verge of significance (p=0.06)
just outside accepted levels of significance (p=0.06)
just outside levels of significance (p<0.08)
just outside the bounds of significance (p=0.06)
just outside the conventional levels of significance (p=0.1076)
just outside the level of significance (p=0.0683)
just outside the limits of significance (p=0.06)
just outside the traditional bounds of significance (p=0.06)
just over the limits of statistical significance (p=0.06)
just short of significance (p=0.07)
just shy of significance (p=0.053)
just skirting the boundary of significance (p=0.052)
just tendentially significant (p=0.056)
just tottering on the brink of significance at the 0.05 level
just very slightly missed the significance level (p=0.086)
leaning towards significance (p=0.15)
leaning towards statistical significance (p=0.06)
likely to be significant (p=0.054)
loosely significant (p=0.10)
marginal significance (p=0.07)
marginally and negatively significant (p=0.08)
marginally insignificant (p=0.08)
marginally nonsignificant (p=0.096)
marginally outside the level of significance
marginally significant (p>=0.1)
marginally significant tendency (p=0.08)
marginally statistically significant (p=0.08)
may not be significant (p=0.06)
medium level of significance (p=0.051)
mildly significant (p=0.07)
missed narrowly statistical significance (p=0.054)
moderately significant (p>0.11)
modestly significant (p=0.09)
narrowly avoided significance (p=0.052)
narrowly eluded statistical significance (p=0.0789)
narrowly escaped significance (p=0.08)
narrowly evaded statistical significance (p>0.05)
narrowly failed significance (p=0.054)
narrowly missed achieving significance (p=0.055)
narrowly missed overall significance (p=0.06)
narrowly missed significance (p=0.051)
narrowly missed standard significance levels (p<0.07)
narrowly missed the significance level (p=0.07)
narrowly missing conventional significance (p=0.054)
near limit significance (p=0.073)
near miss of statistical significance (p>0.1)
near nominal significance (p=0.064)
near significance (p=0.07)
near to statistical significance (p=0.056)
near/possible significance(p=0.0661)
near-borderline significance (p=0.10)
near-certain significance (p=0.07)
nearing significance (p<0.051)
nearly acceptable level of significance (p=0.06)
nearly approaches statistical significance (p=0.079)
nearly borderline significance (p=0.052)
nearly negatively significant (p<0.1)
nearly positively significant (p=0.063)
nearly reached a significant level (p=0.07)
nearly reaching the level of significance (p<0.06)
nearly significant (p=0.06)
nearly significant tendency (p=0.06)
nearly, but not quite significant (p>0.06)
near-marginal significance (p=0.18)
near-significant (p=0.09)
near-to-significance (p=0.093)
near-trend significance (p=0.11)
nominally significant (p=0.08)
non-insignificant result (p=0.500)
non-significant in the statistical sense (p>0.05
not absolutely significant but very probably so (p>0.05)
not as significant (p=0.06)
not clearly significant (p=0.08)
not completely significant (p=0.07)
not completely statistically significant (p=0.0811)
not conventionally significant (p=0.089), but..
not currently significant (p=0.06)
not decisively significant (p=0.106)
not entirely significant (p=0.10)
not especially significant (p>0.05)
not exactly significant (p=0.052)
not extremely significant (p<0.06)
not formally significant (p=0.06)
not fully significant (p=0.085)
not globally significant (p=0.11)
not highly significant (p=0.089)
not insignificant (p=0.056)
not markedly significant (p=0.06)
not moderately significant (P>0.20)
not non-significant (p>0.1)
not numerically significant (p>0.05)
not obviously significant (p>0.3)
not overly significant (p>0.08)
not quite borderline significance (p>=0.089)
not quite reach the level of significance (p=0.07)
not quite significant (p=0.118)

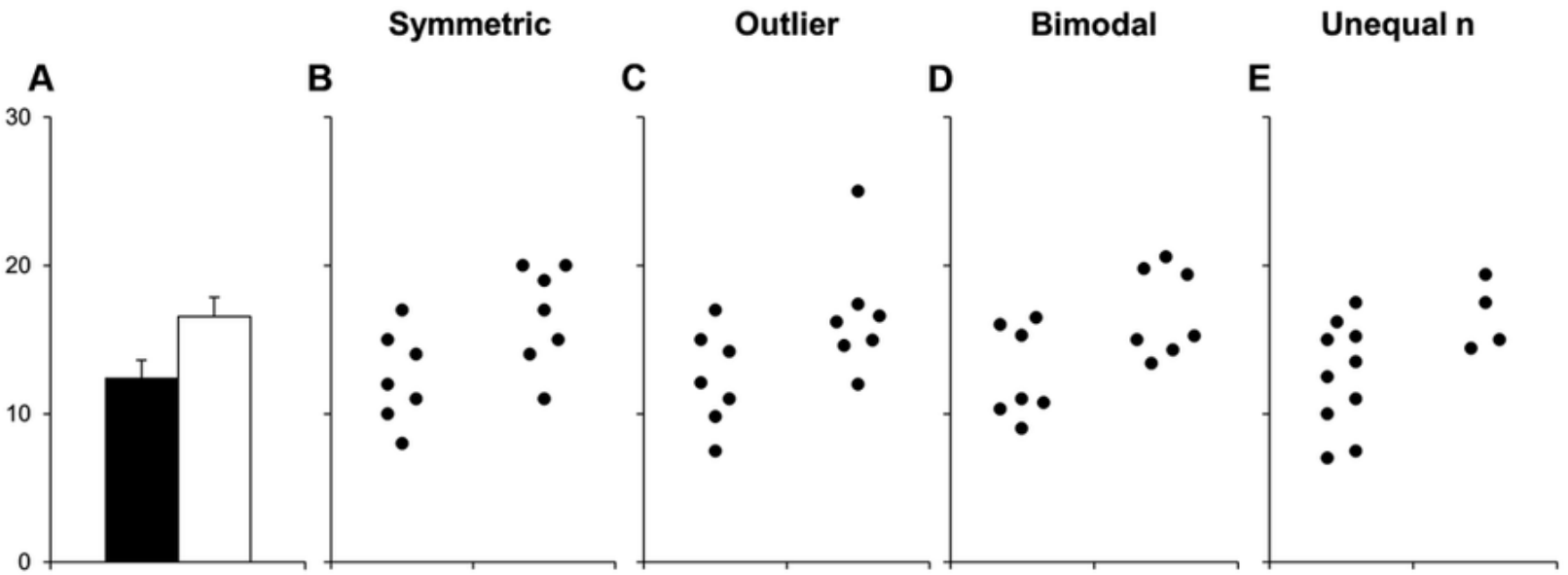
P < 0.04

Effect?

# P < 0.05

P < 0.06

No effect?



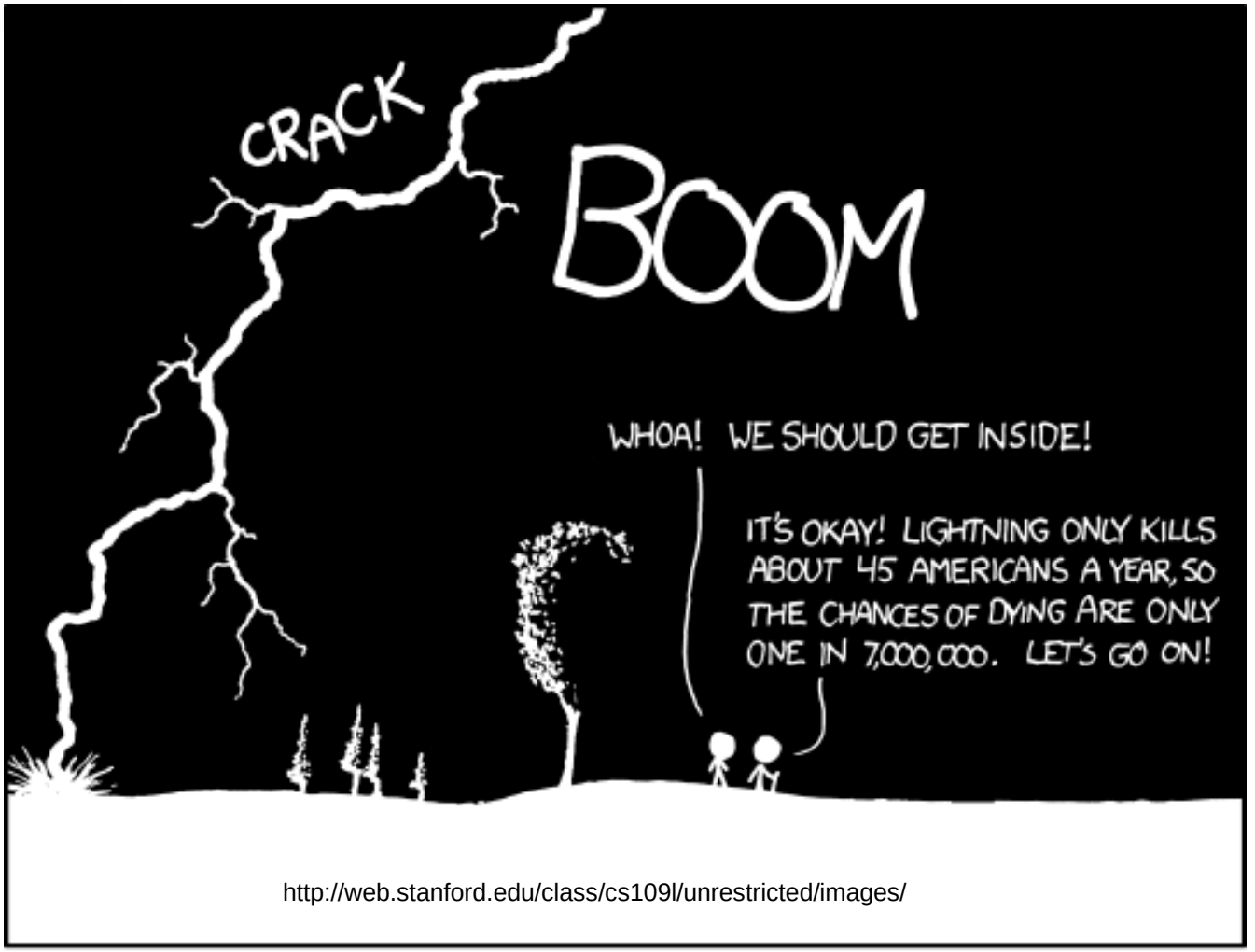
Test				
	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

absence of evidence
is not
evidence of absence

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

$$\underbrace{p(\theta|D,M)}_{\text{Posterior}} = \frac{\overbrace{p(D|\theta,M)}^{\text{Likelihood}} \times \overbrace{p(\theta)}^{\text{Prior}}}{\underbrace{p(D|M)}_{\text{Marginal Likelihood}}}$$

Bayesian analysis: role of prior knowledge?



Editorial

The ASA Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

Pages 129-133 | Accepted author version posted online: 07 Mar 2016, Published online:09 Jun 2016

Download citation https://doi.org/10.1080/00031305.2016.1154108 Check for updates

ASA advises researchers to avoid drawing scientific conclusions or making policy decisions based on P values alone. Researchers should describe not only the data analyses that produced statistically significant results, the society says, but all statistical tests and choices made in calculations. Otherwise, results may seem falsely robust. "the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation"

1) P-values can indicate how incompatible the data are with a specified statistical model.
2) P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3) Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
4) Proper inference requires full reporting and transparency.
5) A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6) By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

"statisticians often supplement or even replace p-values with other approaches. These include methods "that emphasize estimation over testing such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence such as likelihood ratios or Bayes factors; and other approaches such as decision-theoretic modeling and false discovery rates."

"The p-value was never intended to be a substitute for scientific reasoning," said Ron Wasserstein, the ASA's executive director. "Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post $p<0.05$ era.'"

PLOS | MEDICINE

Browse | For Authors | Abou

OPEN ACCESS

ESSAY

898,944
VIEWS

Why Most Published Research Findings Are False

John P. A. Ioannidis

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

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

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.

Data silo


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

What is Bioconductor ?

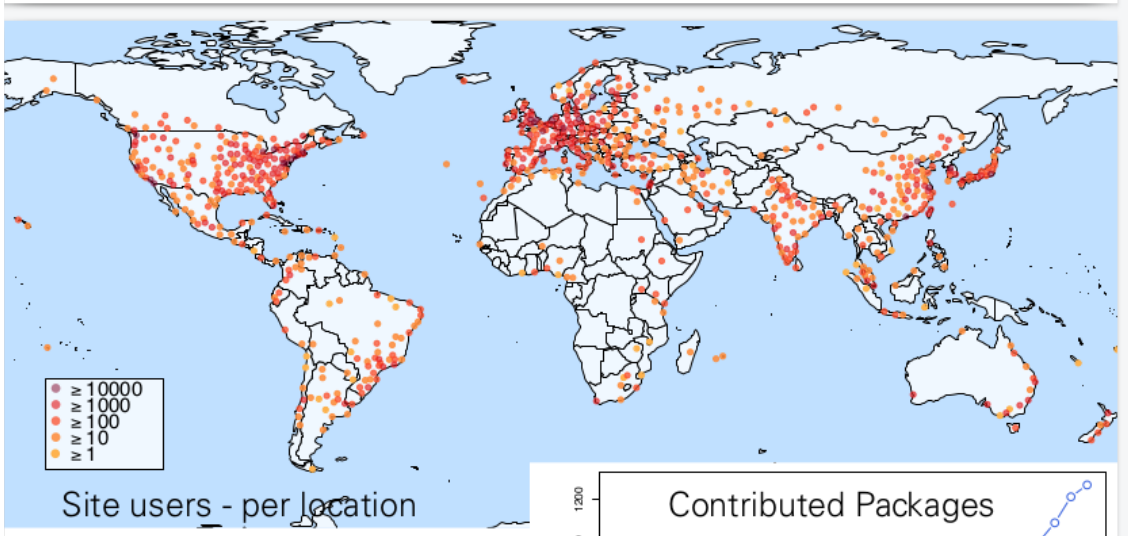
Started 2001 as a platform for analysis & understanding of microarray data

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 Bioconductor ?

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

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

PLOS COMPUTATIONAL BIOLOGY

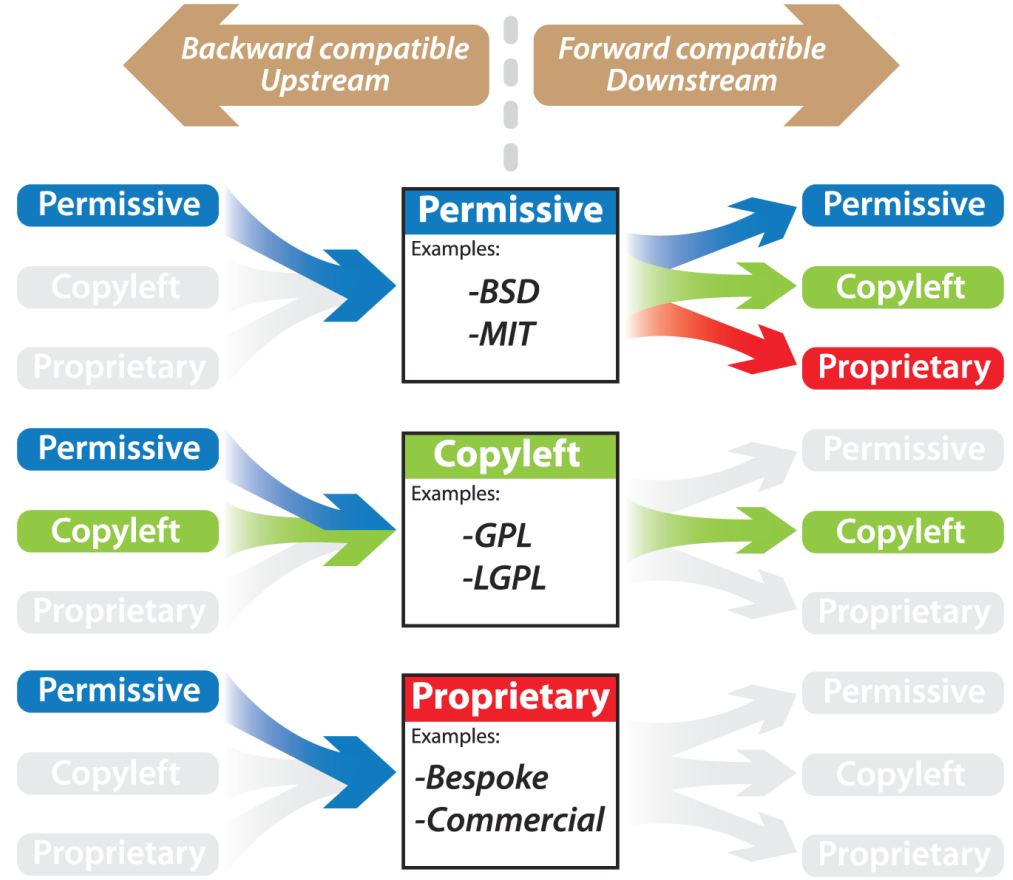
OPEN ACCESS

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

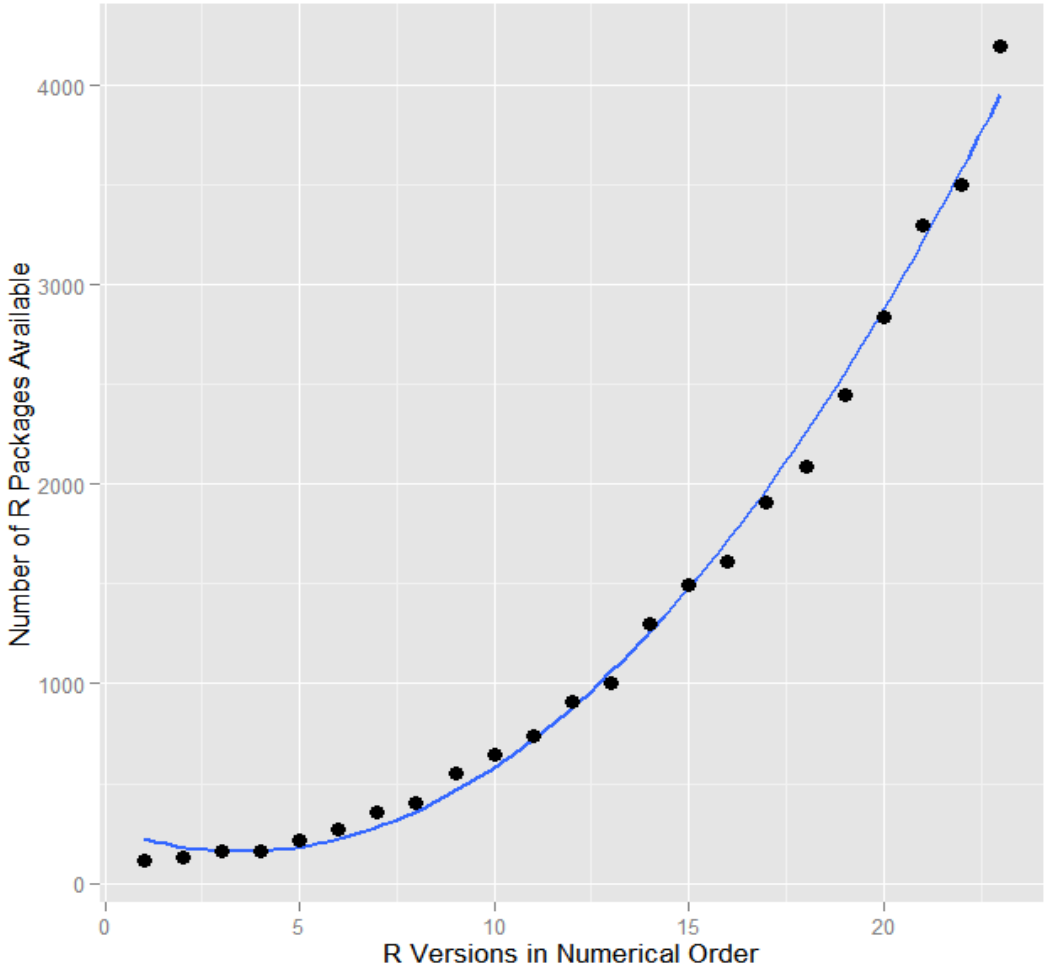
```
Copyright (c) <year> <copyright holders>

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.
```

Number of open analysis tools has grown exponentially



github SOCIAL CODING

R

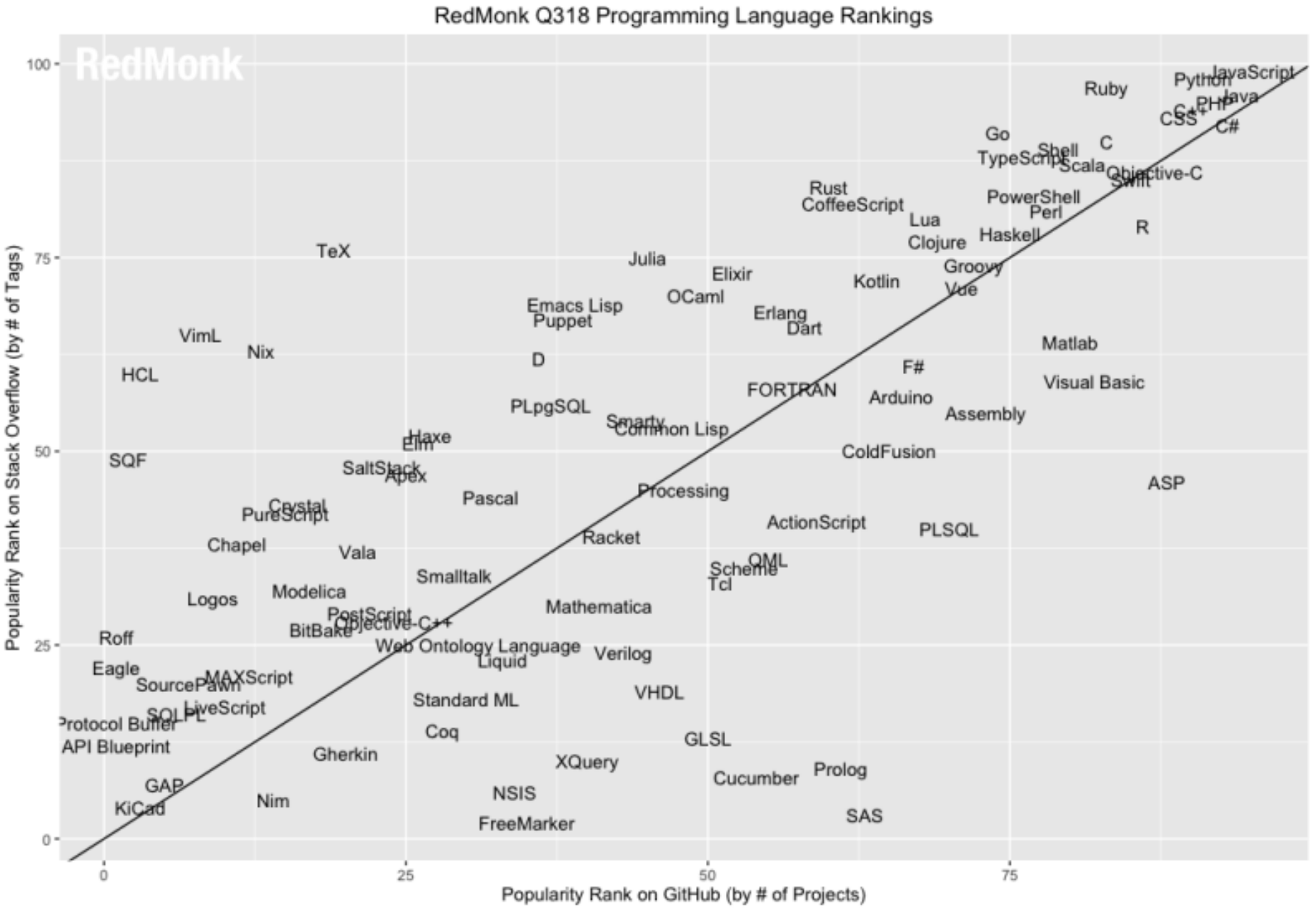
rOpenGov

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

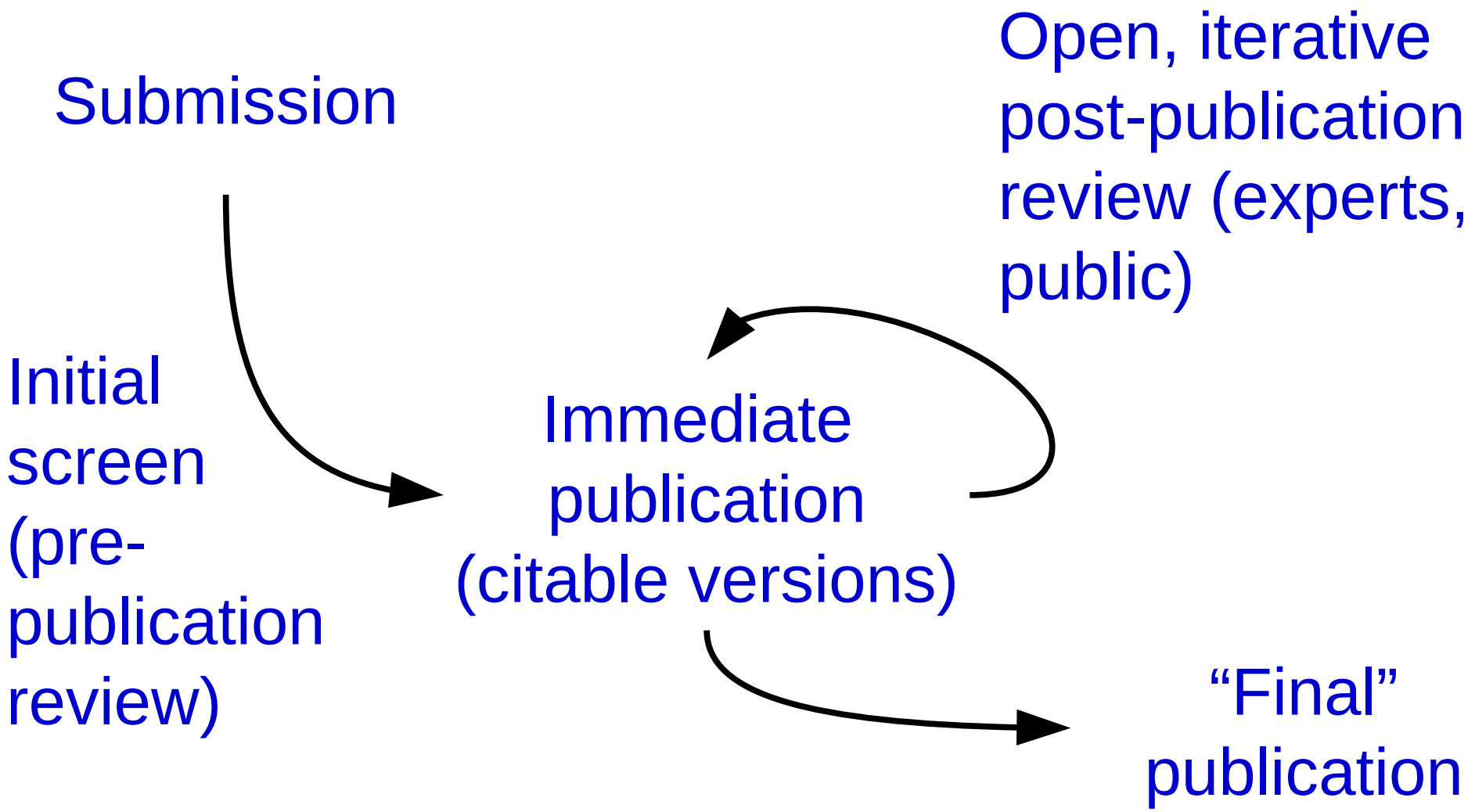
ROpenSci

Value of data can increase through sharing & use

Varying cultures of open collaboration



Pre- vs. post-publication review



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

Source: Wikimedia Commons / Public domain

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*

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

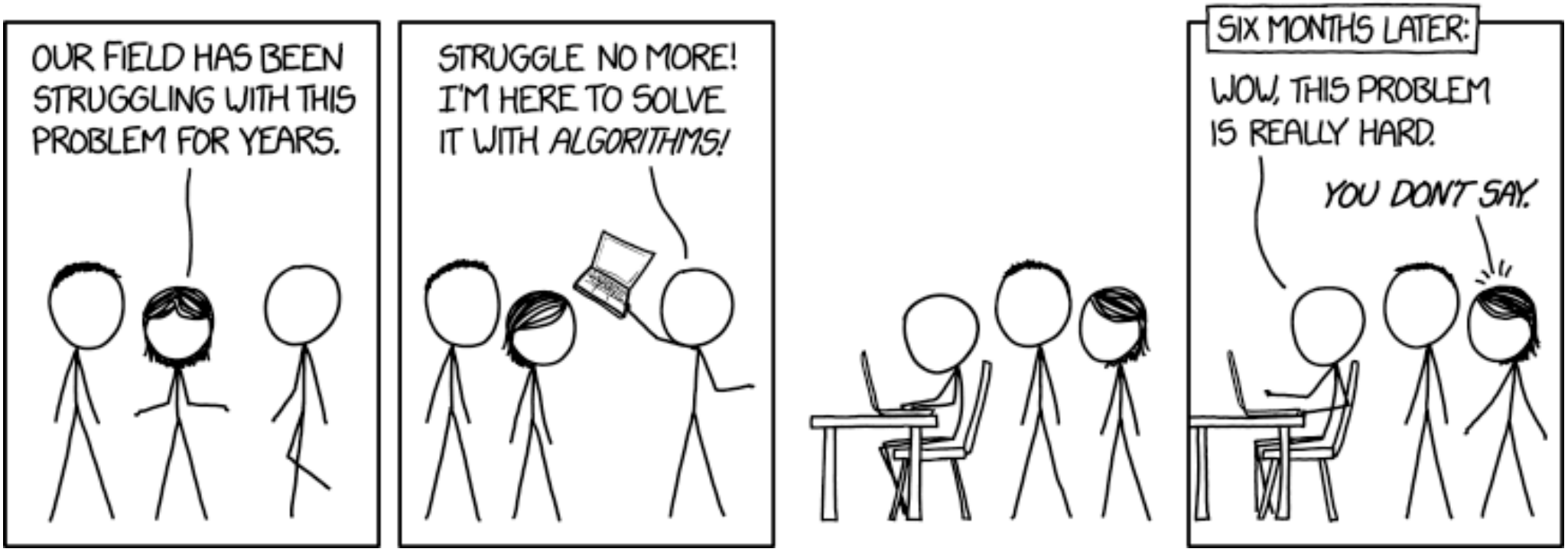
Research Ideas and Outcomes 3:e13593, 2017

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


OUR FIELD HAS BEEN STRUGGLING WITH THIS PROBLEM FOR YEARS.
STRUGGLE NO MORE! I'M HERE TO SOLVE IT WITH ALGORITHMS!
SIX MONTHS LATER:
WOW, THIS PROBLEM IS REALLY HARD.
YOU DON'T SAY.

Task

Create a clear reproducible report of the example data, including the following aspects:

- Data import
- Data exploration & summaries
- Alpha diversity
- Beta diversity
- Differential abundance analysis
- **Conclusions**