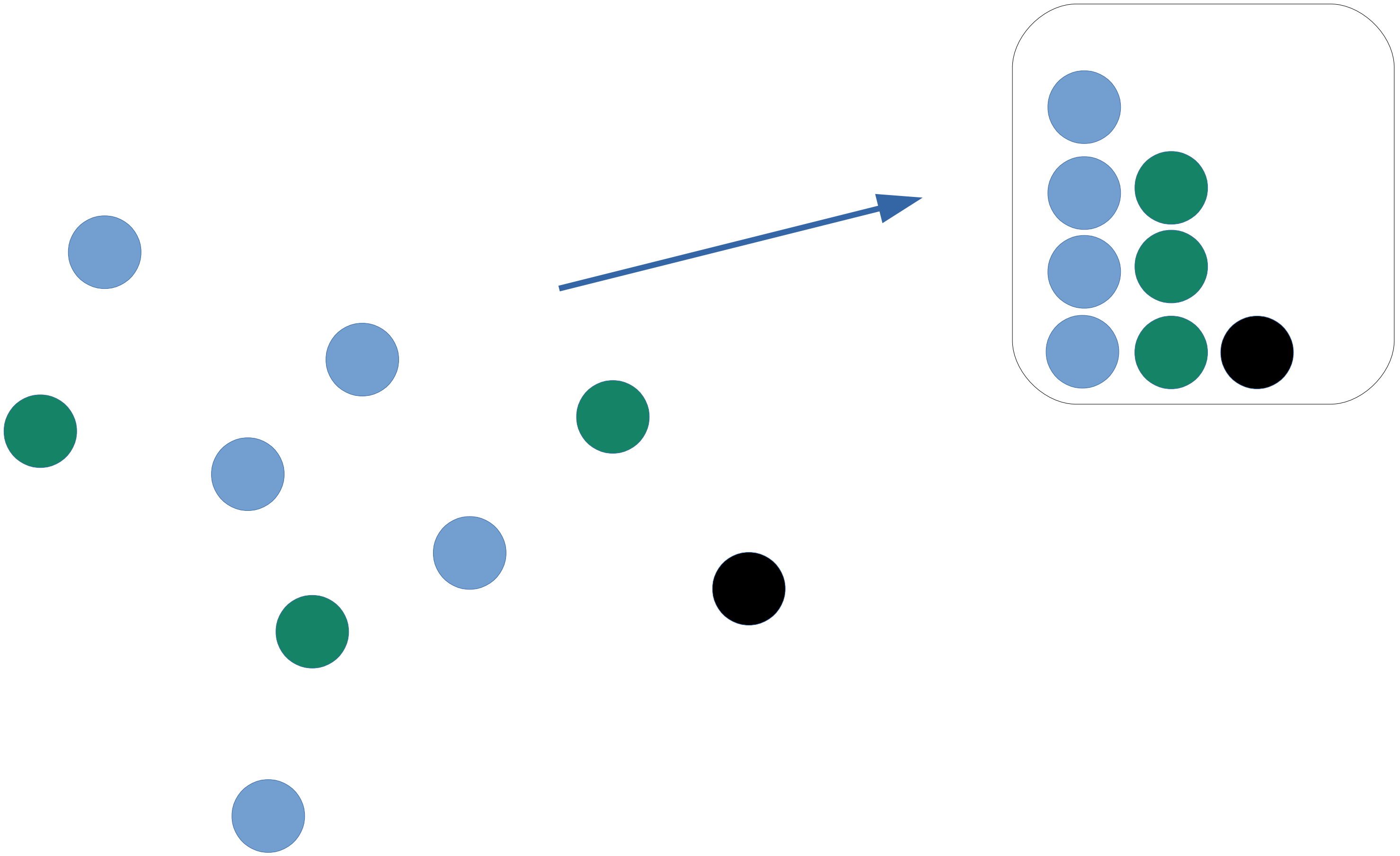
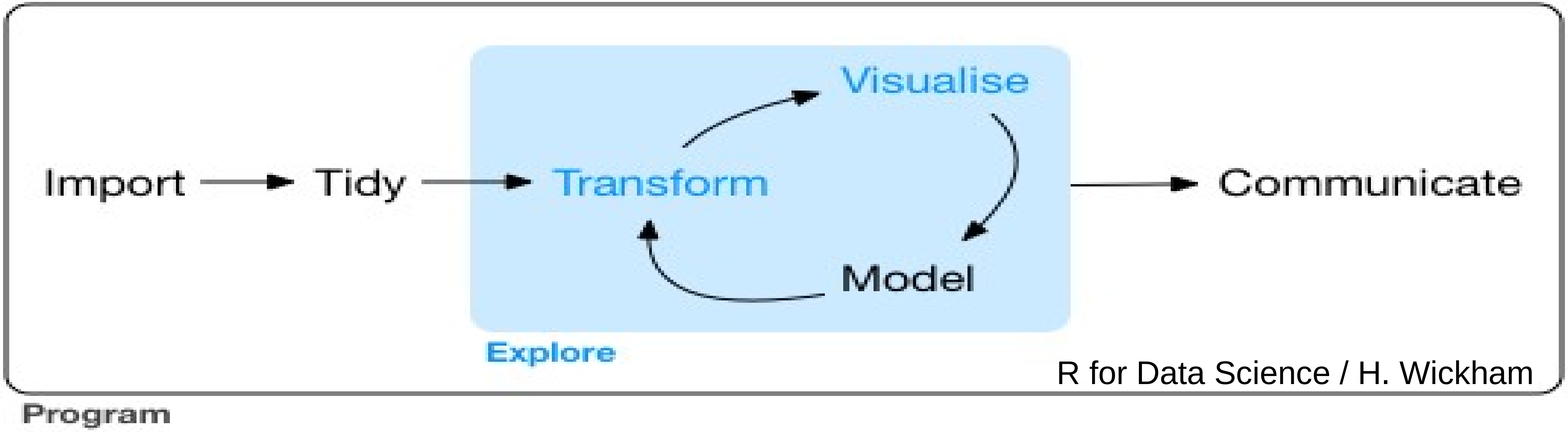


Data science workflow



R for Data Science / H. Wickham

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 → quality filter/trim → fasta files

fastq files

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

might be done by sequencing facility

demultiplex
(split samples by barodes)

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

fastqc/multiqc

quality filter/trim
(remove adapters/**primers**)

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

Count Table

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

Analysis

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

assembly path

consider testing assemblies with and w/o

digital normalization

Some tools:
- bbnorm
- diginorm

MetaQUAST is a great tool for comparing assemblies

(co)-assembly

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

map individual sample reads to (co)-assembly

Generate coverage information (mapping)

Some tools:
- bowtie2
- bwa

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)

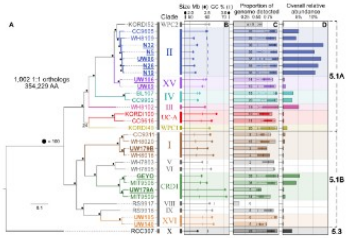
Gene calling
Functional/taxonomic profiling

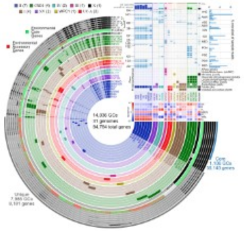
Some tools:
- prodigal (identifies open reading frames)
- prokka (runs prodigal and performs annotations)
- GHOSTKOALA (web-hosted KEGG annotations)
- BLAST (protein nr db/refseq/COGs)

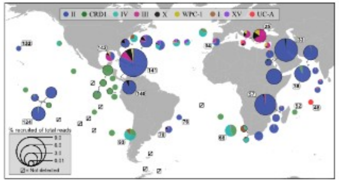
Some common genomics stuff

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

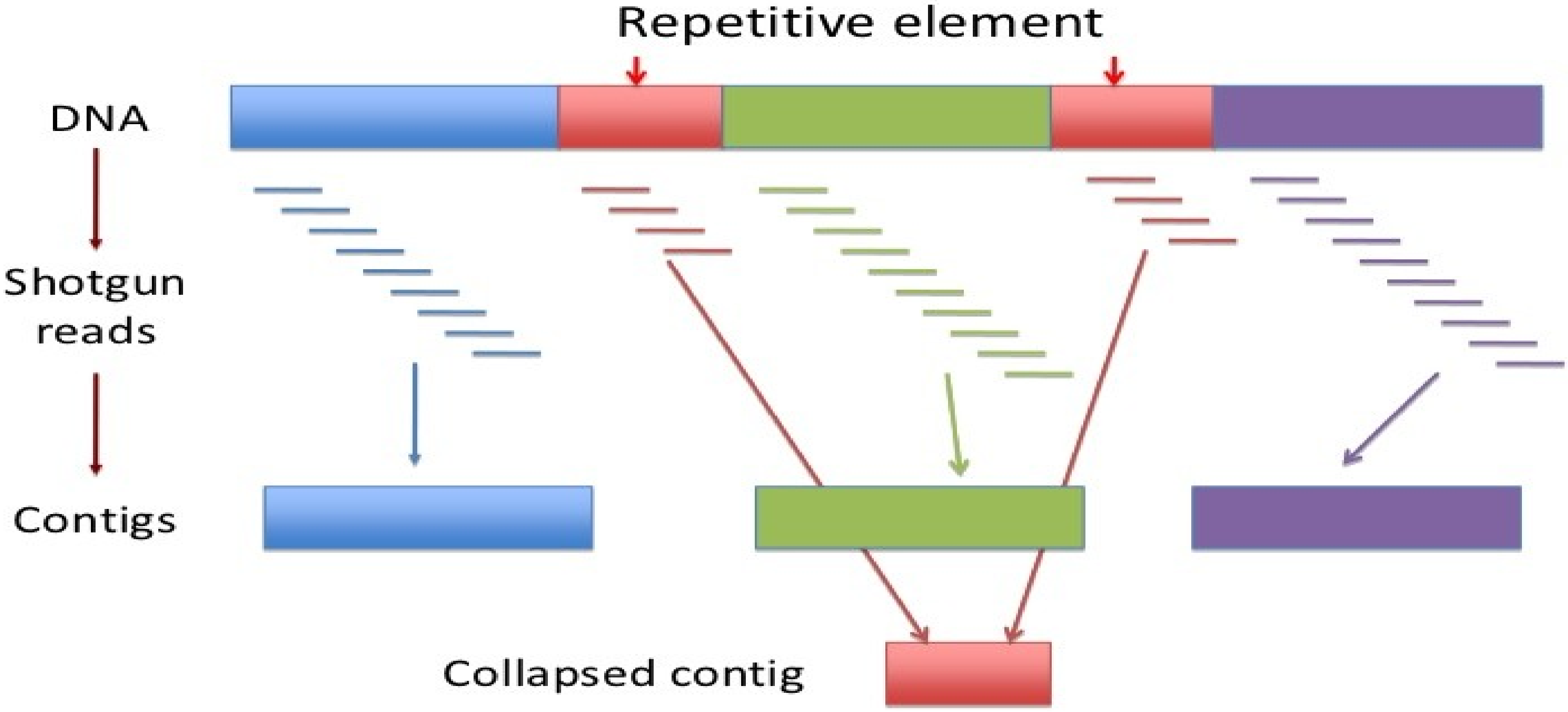
JOSE 10.21105/jose.00053

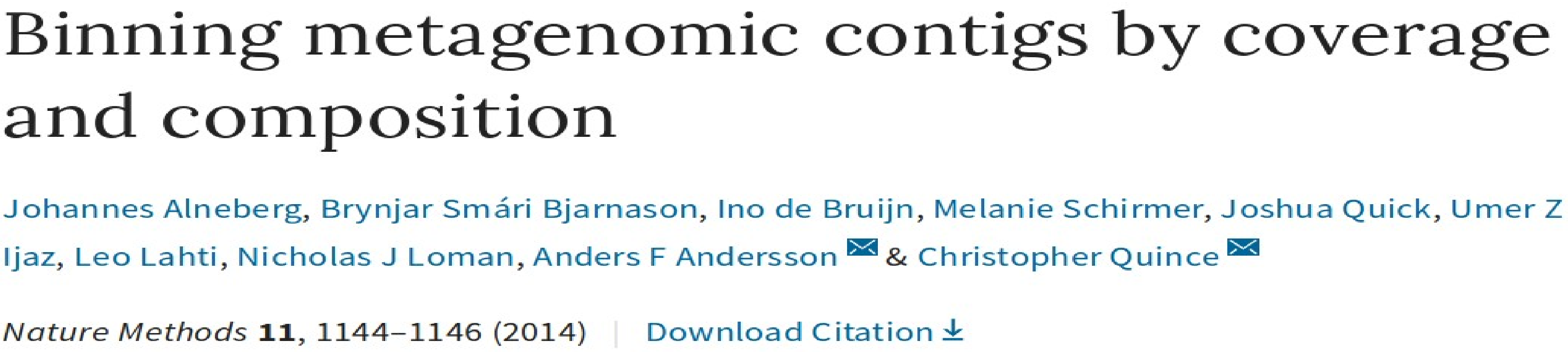
AstrobioMike

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

Application:
contig clustering for metagenome assembly





Binning metagenomic contigs by coverage and composition

Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson ✉ & Christopher Quince ✉

Nature Methods **11**, 1144–1146 (2014) | Download Citation

Cluster contigs by *co-occurrence* and *sequence composition*
in order to recover metagenomic species

CONCOCT

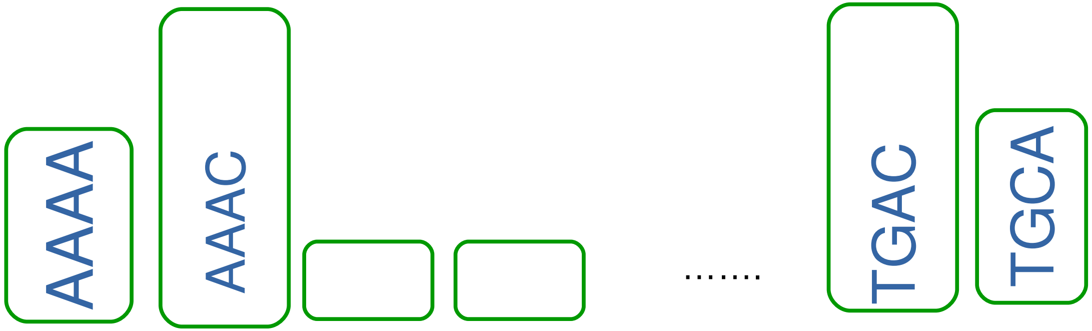
Contig composition (4-bp strings)

Contig: TCGAAATTACGGTCGATTTTAAACTCGGTCTGGA...

k-mer content: “TCGA”

k=4: 136 tetramers (excl. palindromic sequences)

Each contig has a profile over the 136 tetramers



CONCOCT

Cluster contigs by *co-occurrence* and *sequence composition*
in order to recover metagenomic species

Coverage
53 samples

Composition
136 tetramers

Coverage

Composition

High-dimensional feature space (53 + 136)!

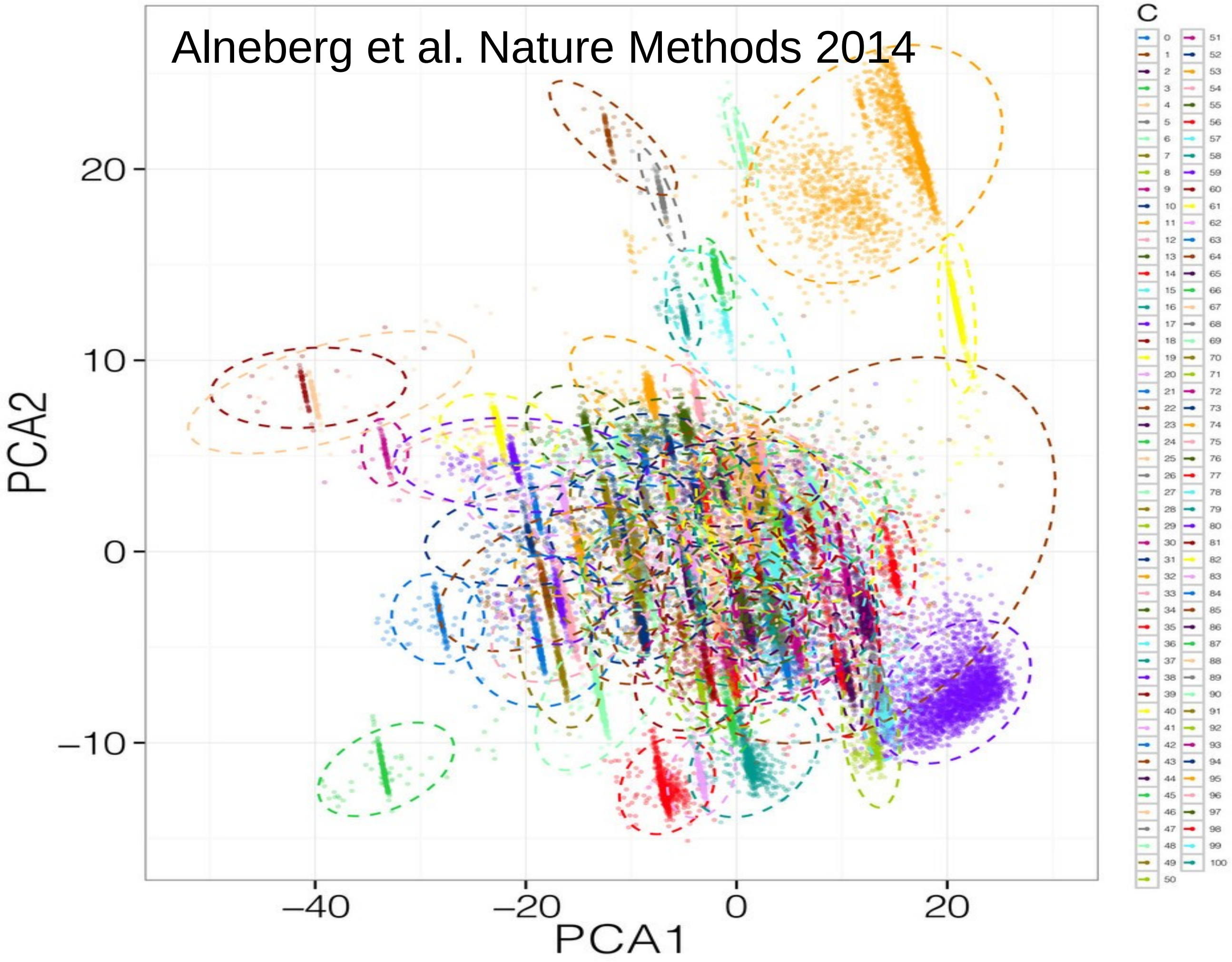
→ drop dimensionality with PCA (d~20; 90%)

→ multivariate gaussian clusters

→ infer the cluster number automatically

Clustering contigs by coverage and composition (CONCOCT)

Variational Dirichlet Process multivariate mixture models work very well in practical applications with high-dimensional data



Alneberg et al. Nature Methods 2014

Core microbiota variation (N = 5005)

Z-score across subjects: red – high abundance & blue – low abundance

Core microbiota shows remarḱable variation across population.

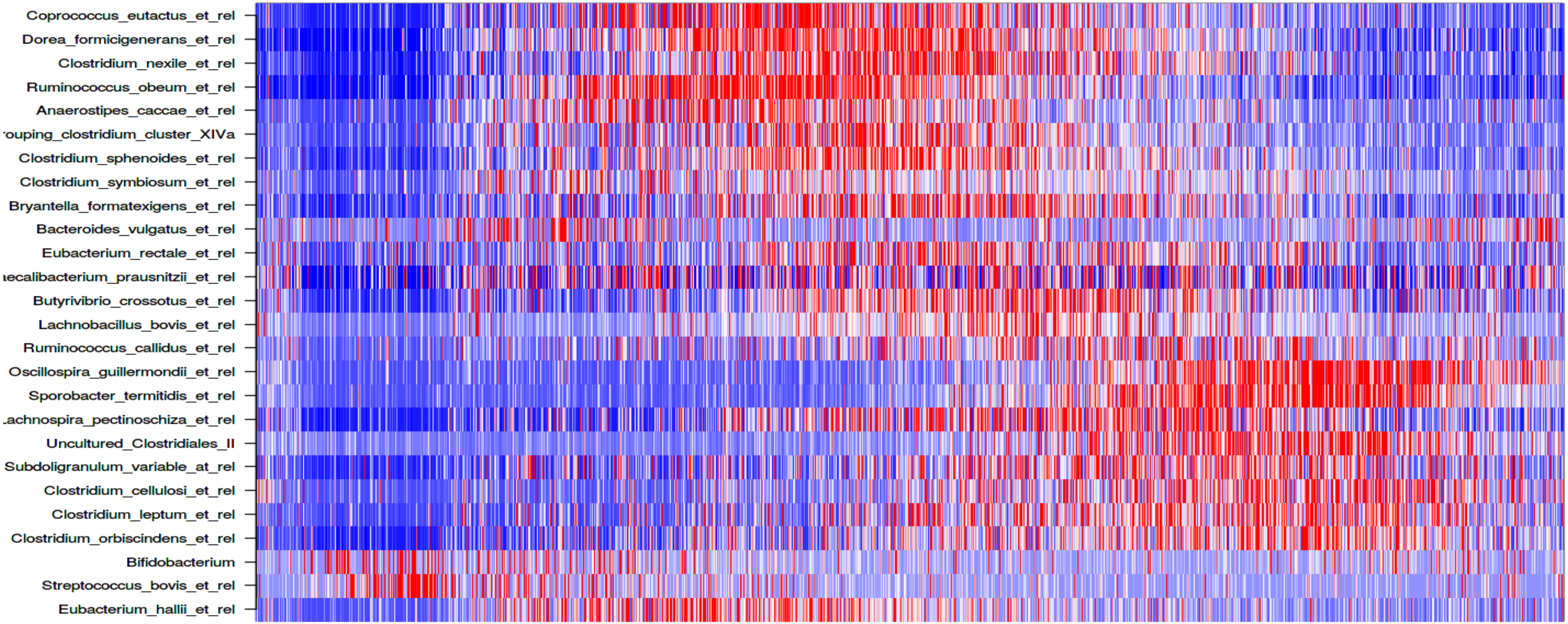
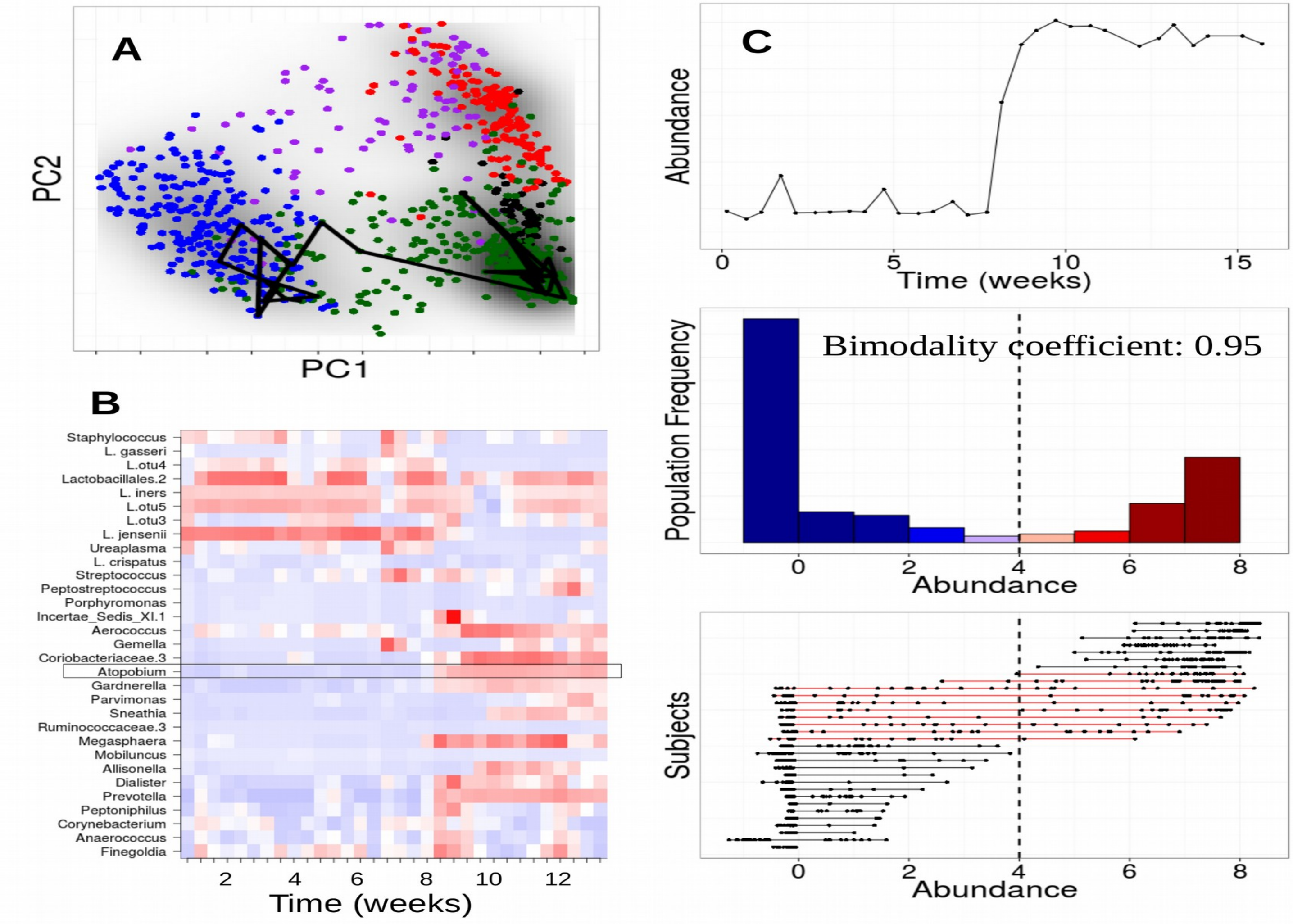
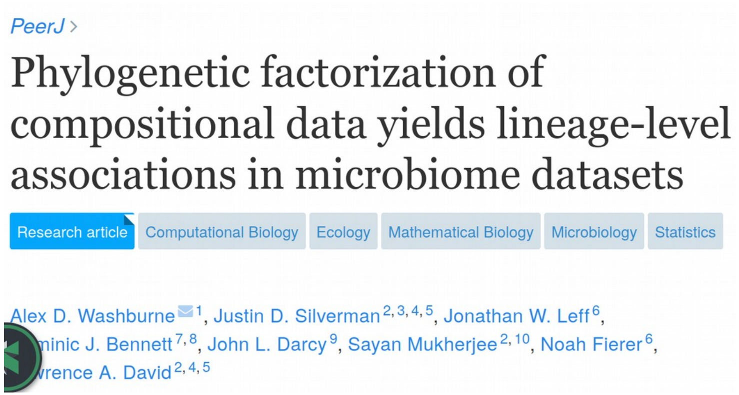


Figure: Faust et al. Curr. Op. Microbiol. 2015
Data: Gajer et al. 2012

Community states in vaginal microbiome of reproductive-age women



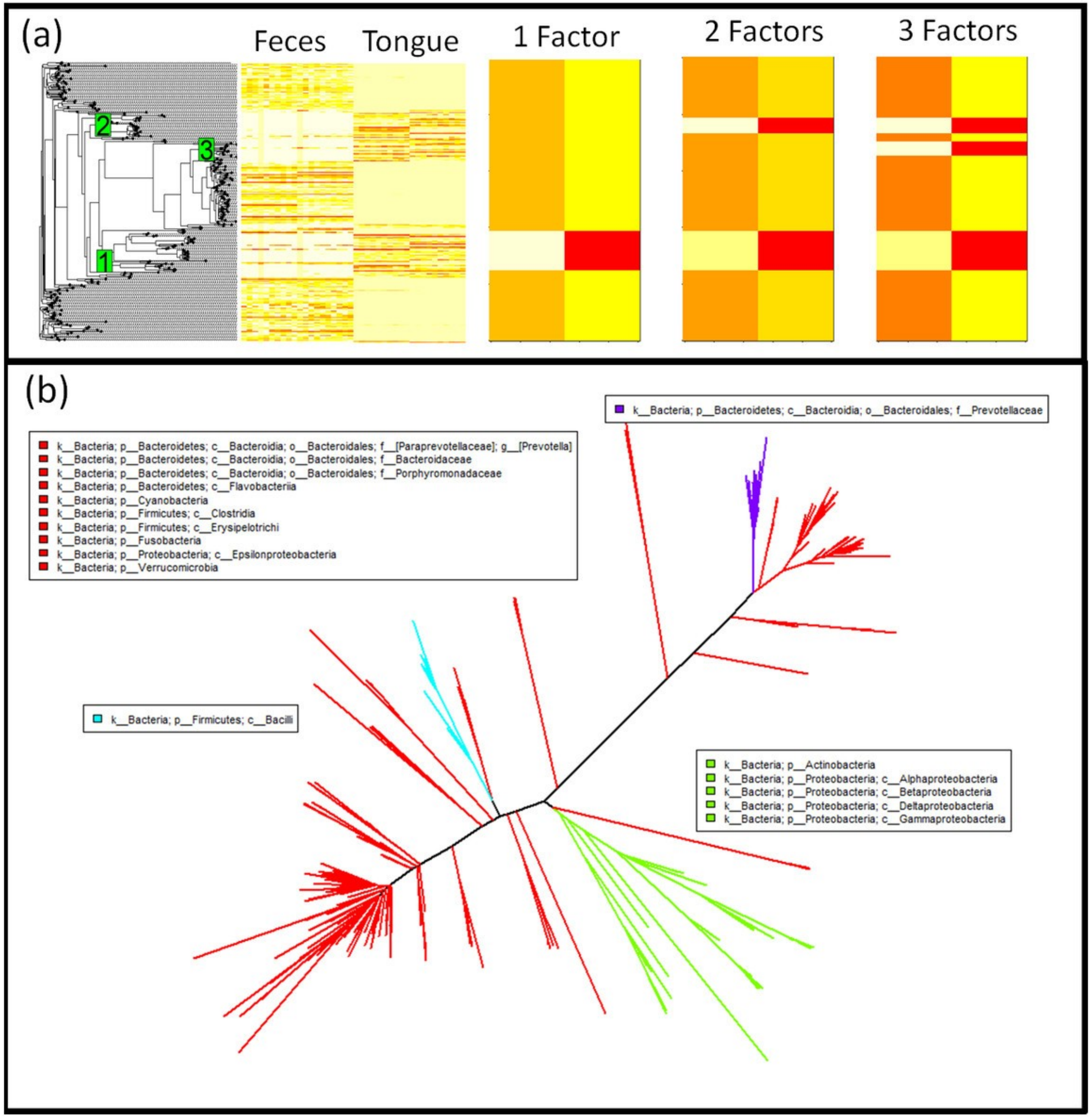


PeerJ >

Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets

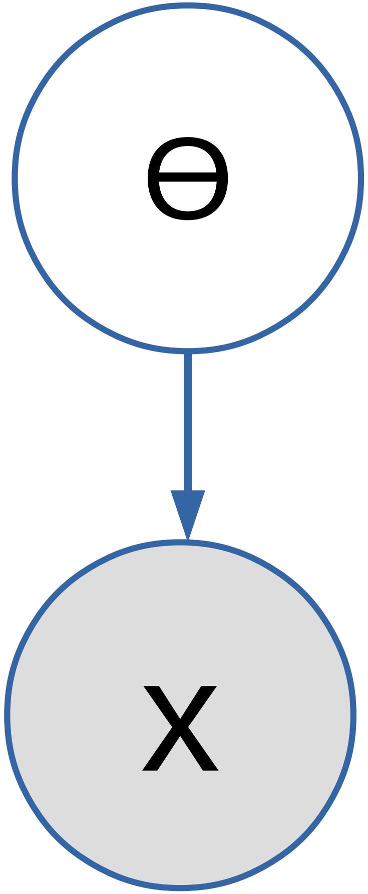
Research article | Computational Biology | Ecology | Mathematical Biology | Microbiology | Statistics

Alex D. Washburne[1], Justin D. Silverman[2,3,4,5], Jonathan W. Leff[6], minic J. Bennett[7,8], John L. Darcy[9], Sayan Mukherjee[2,10], Noah Fierer[6], wrence A. David[2,4,5]

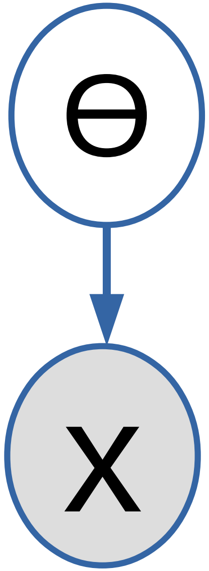


Model

Observations
(Data)



$$\overset{\text{Posterior}}{P(\boldsymbol{\Theta}|data)} \propto \overset{\text{Likelihood}}{P(data|\boldsymbol{\Theta})} \times \overset{\text{Prior}}{P(\boldsymbol{\Theta})}$$



$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Gaussian Likelihood

N(x|μ,σ)

Likelihood

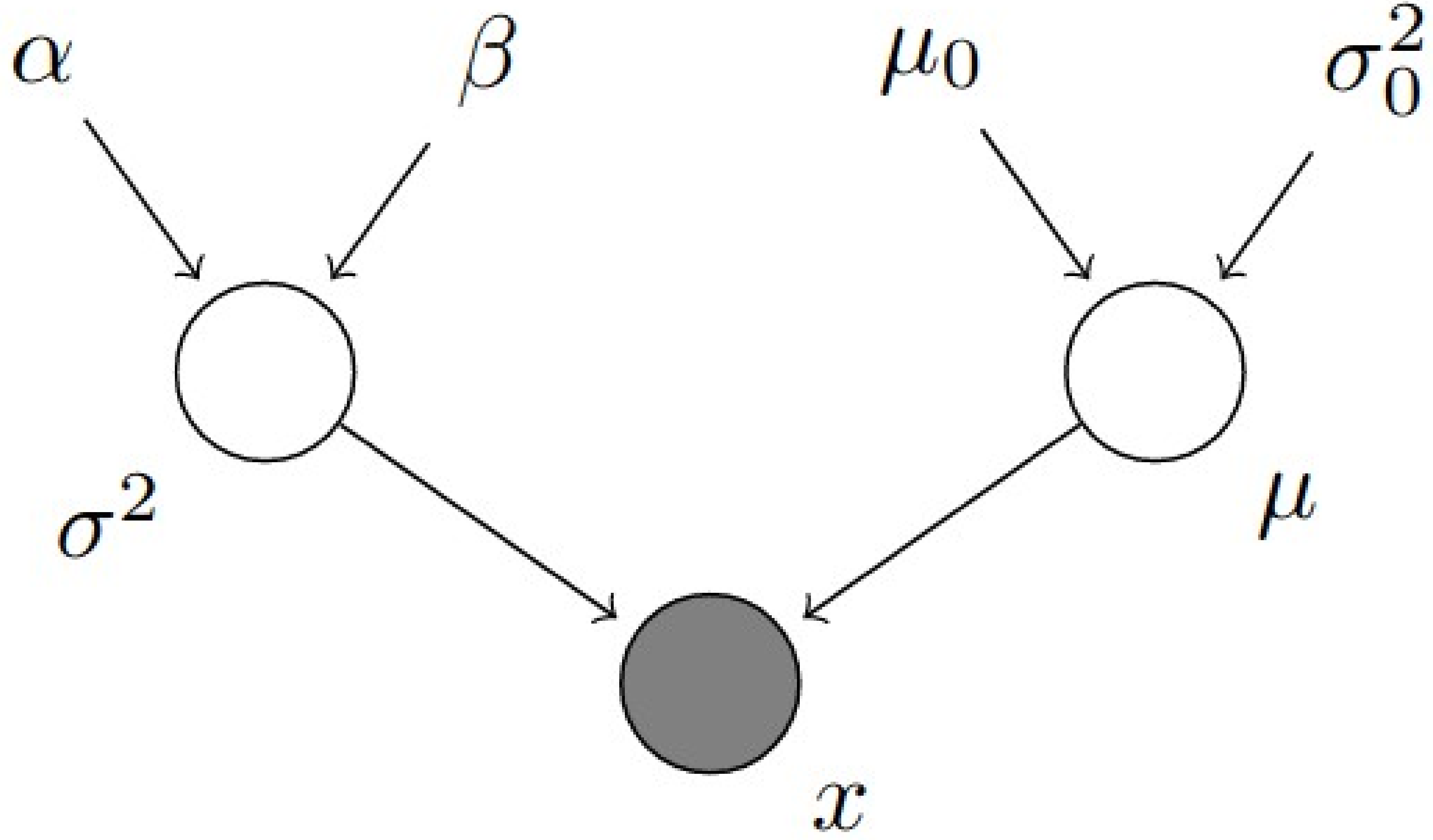
$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$N(x|\mu,\sigma)\ N(\mu|\mu_{0,}\ \sigma_0)$$

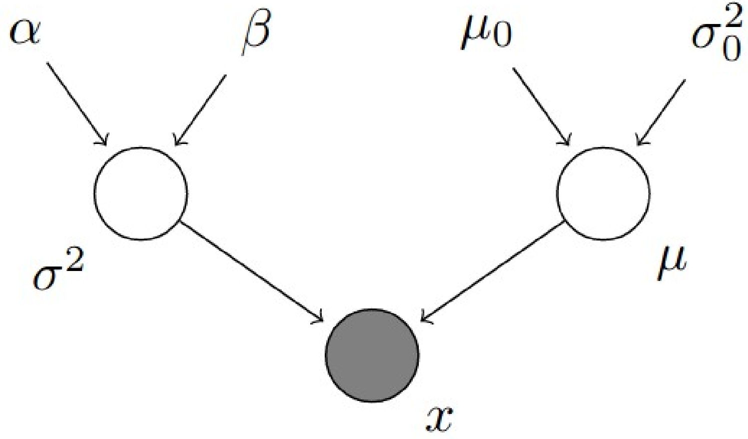
Likelihood Prior

$$N(\theta|x) \sim N(x|\mu,\sigma)\ N(\mu|\mu_{0,}\ \sigma_0)$$

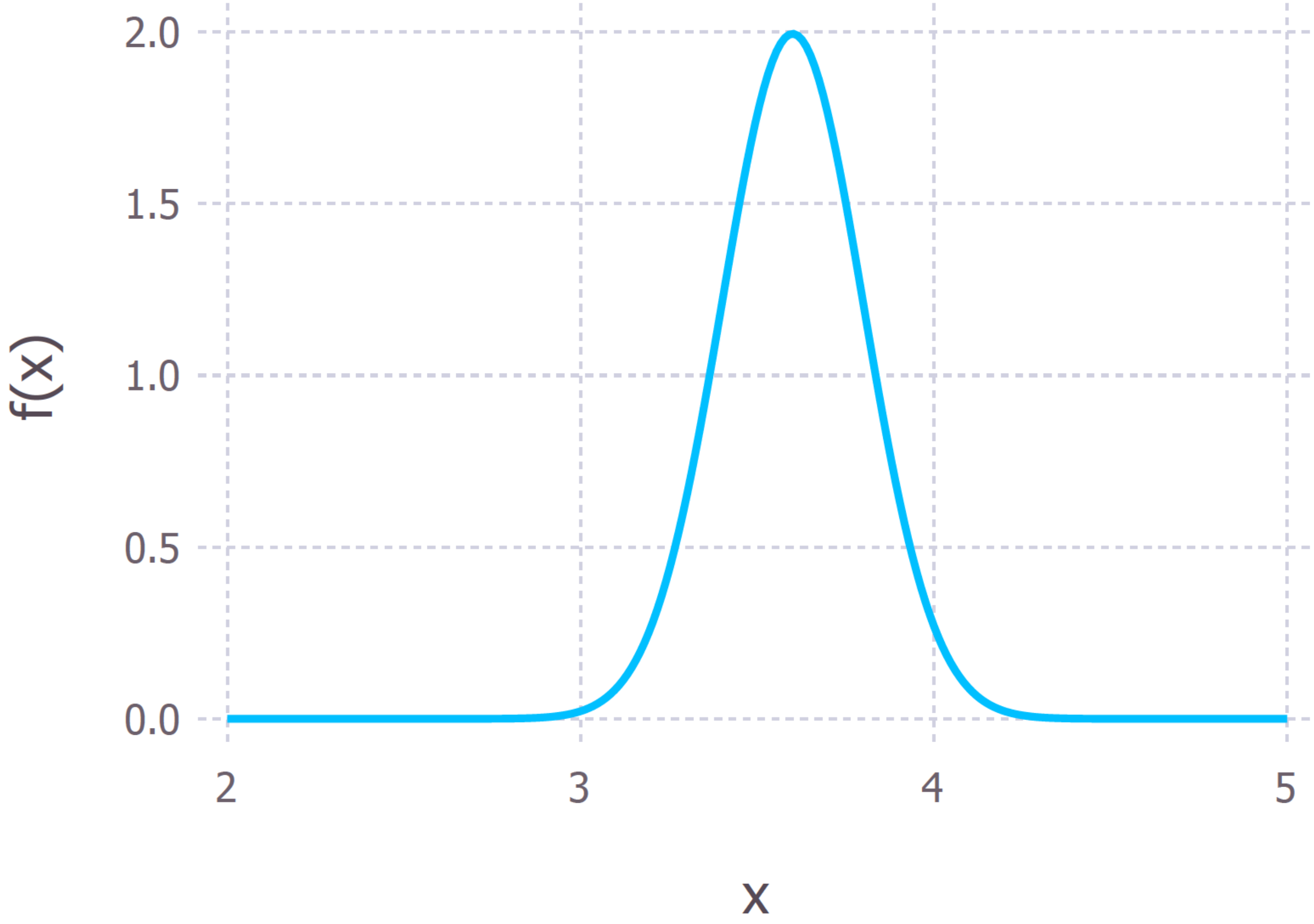
Posterior Likelihood Prior


α
β
μ_0
σ_0^2
σ^2
μ
x

$$
\begin{aligned}
x_i \mid \mu, \tau &\sim \mathcal{N}(\mu, \tau) \quad \text{i.i.d.} \\
\mu \mid \tau &\sim \mathcal{N}(\mu_0, n_0 \tau) \\
\tau &\sim \text{Ga}(\alpha, \beta)
\end{aligned}
$$

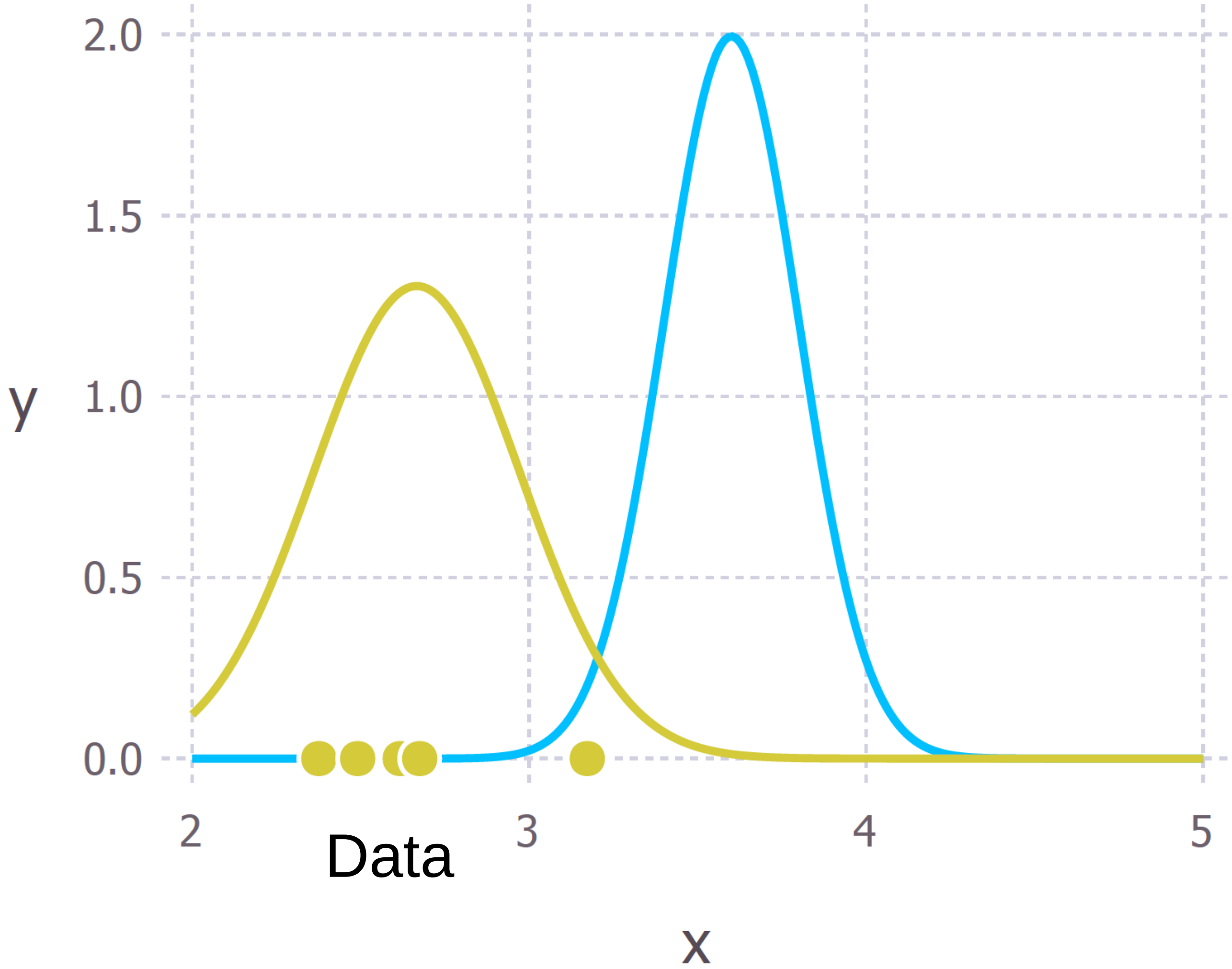


Prior

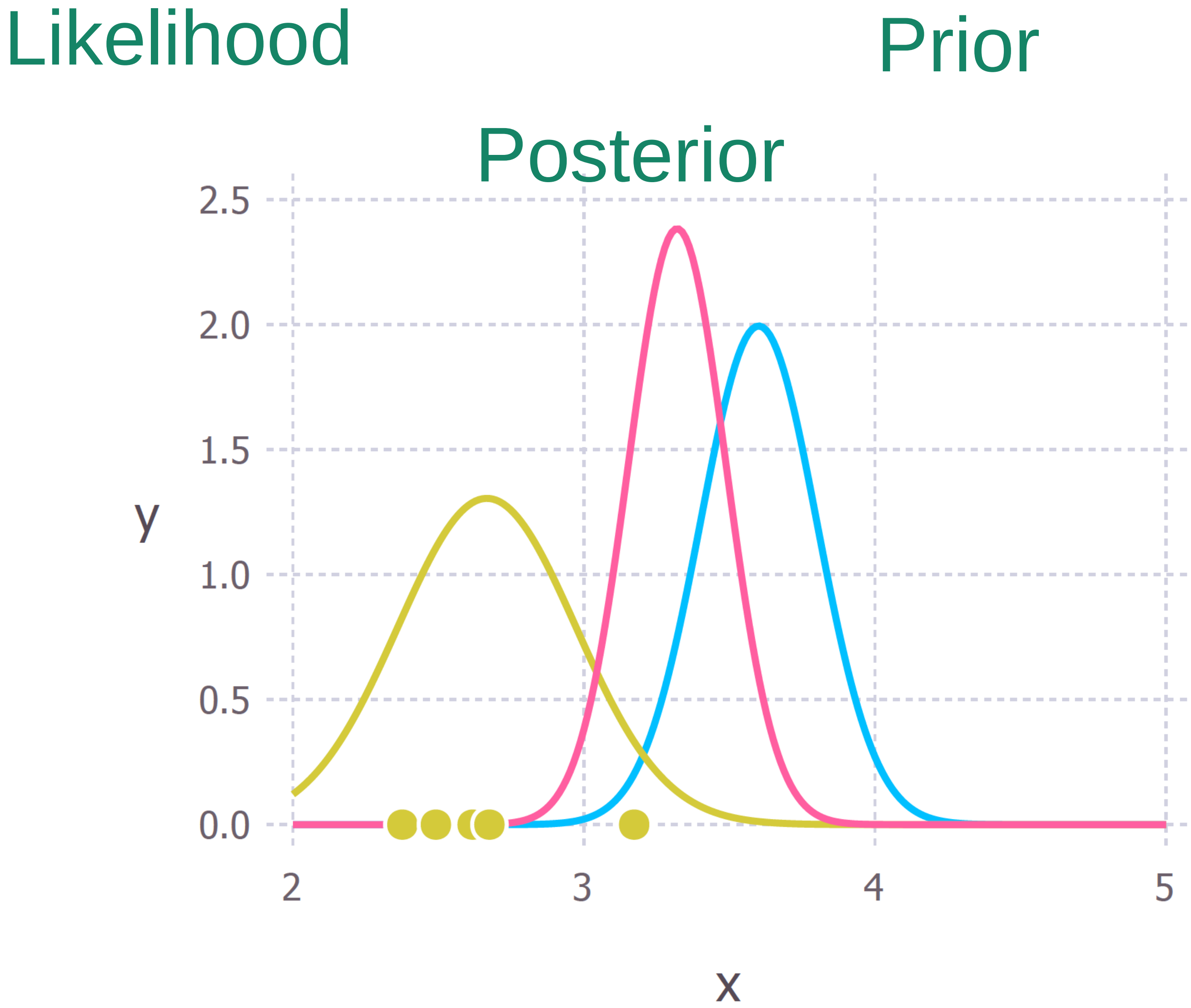


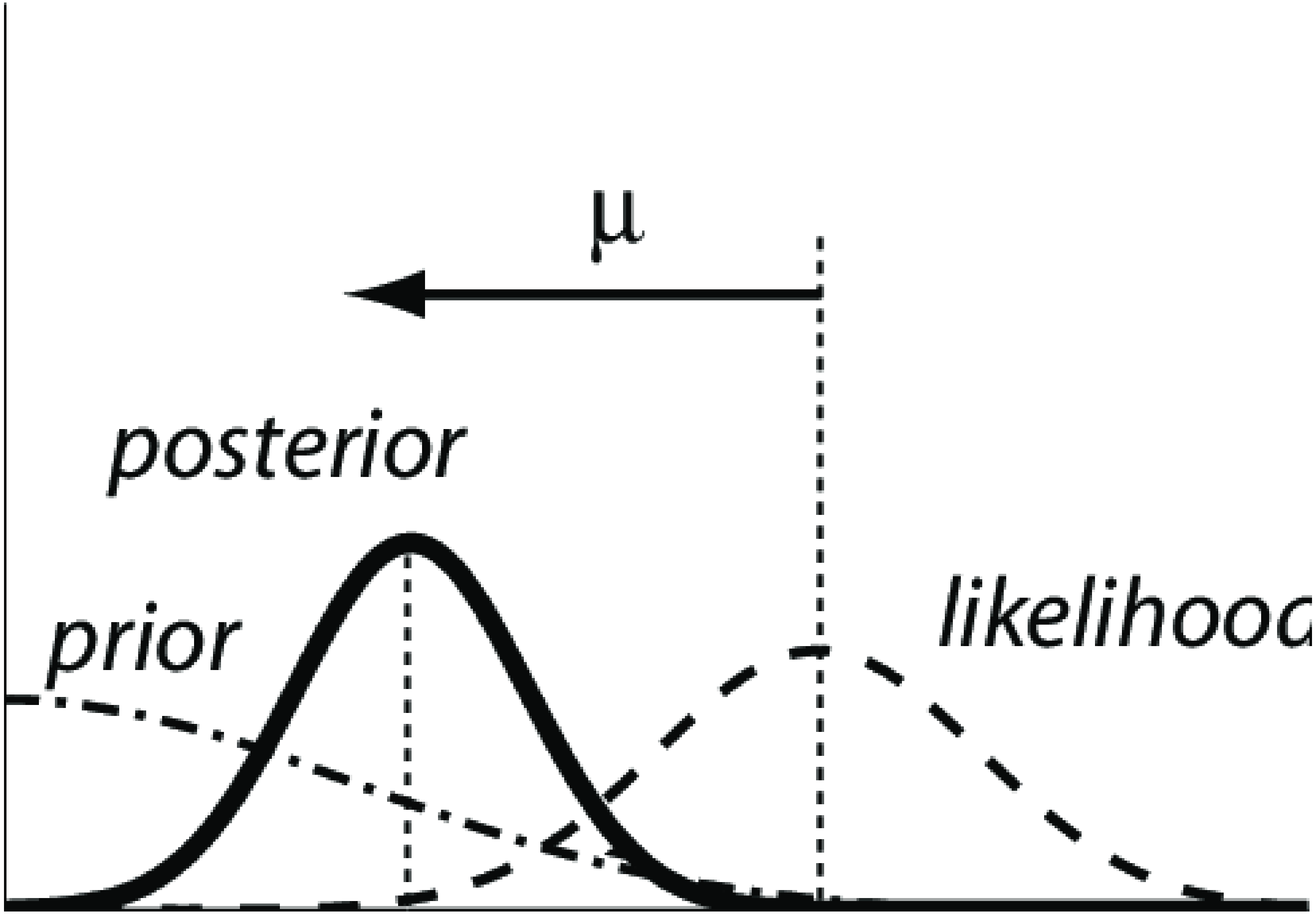
Likelihood

Prior



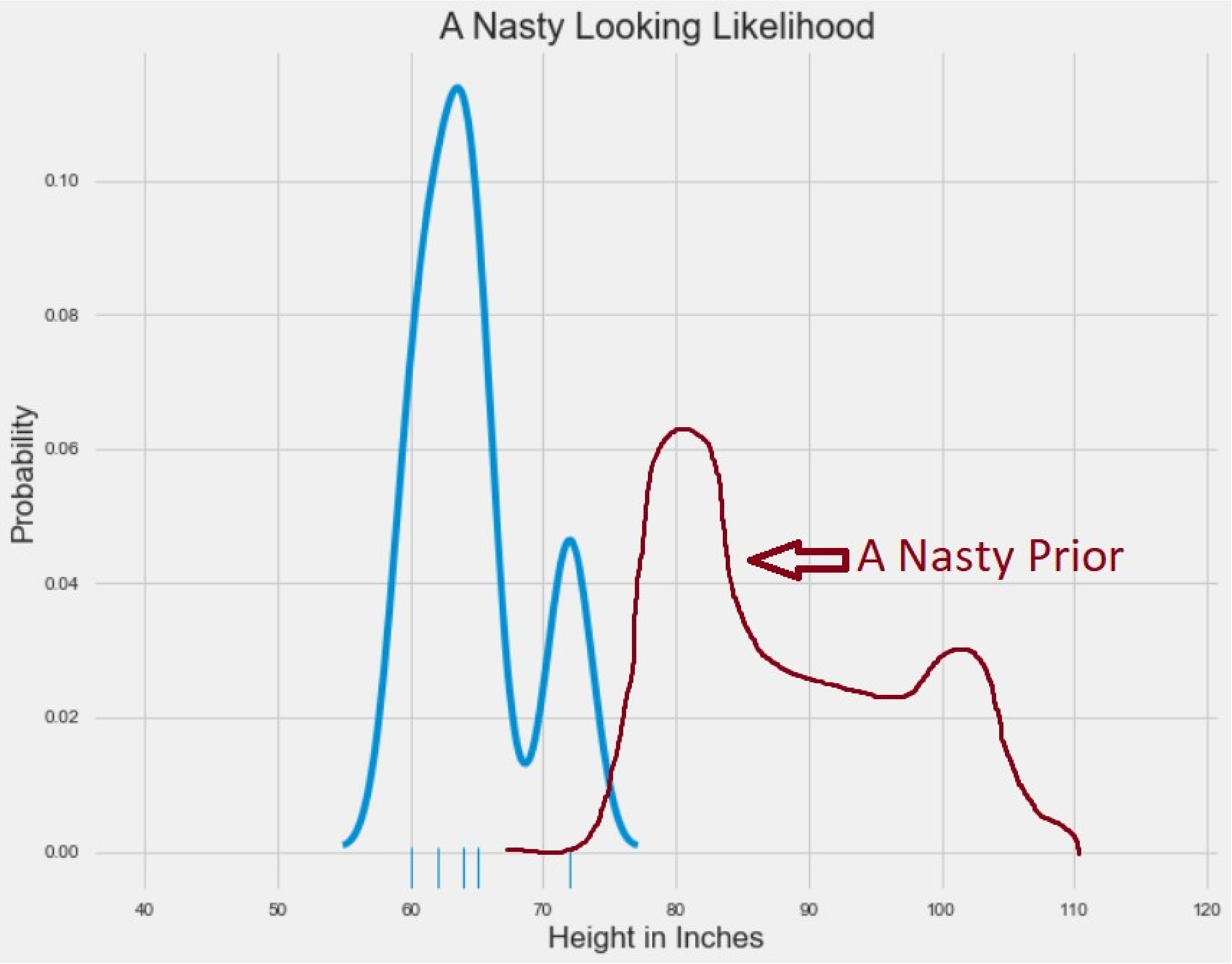
Data


Likelihood
Prior
Posterior
y
2.5
2.0
1.5
1.0
0.5
0.0
2
3
4
5
x

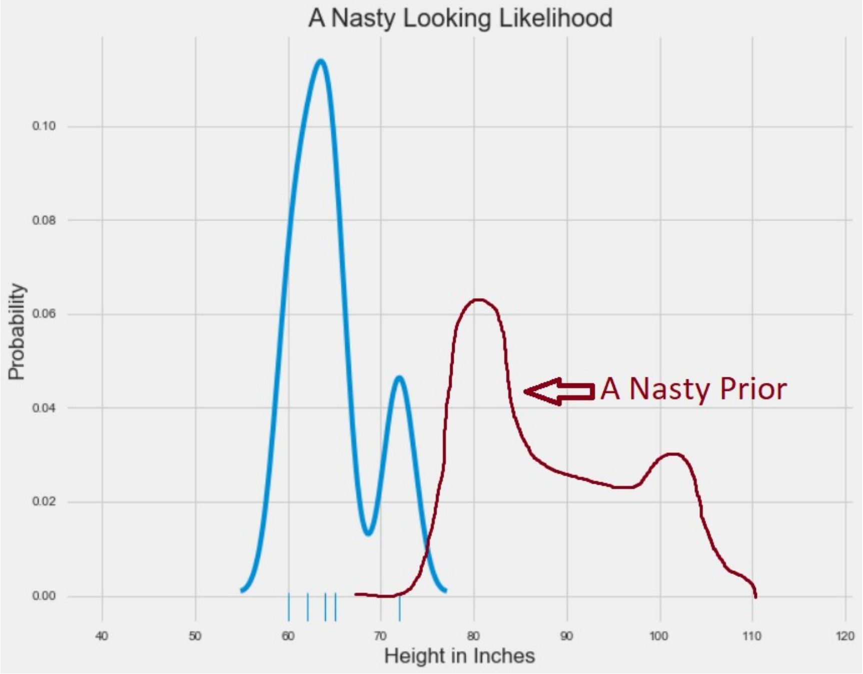

μ
posterior
prior
likelihooc

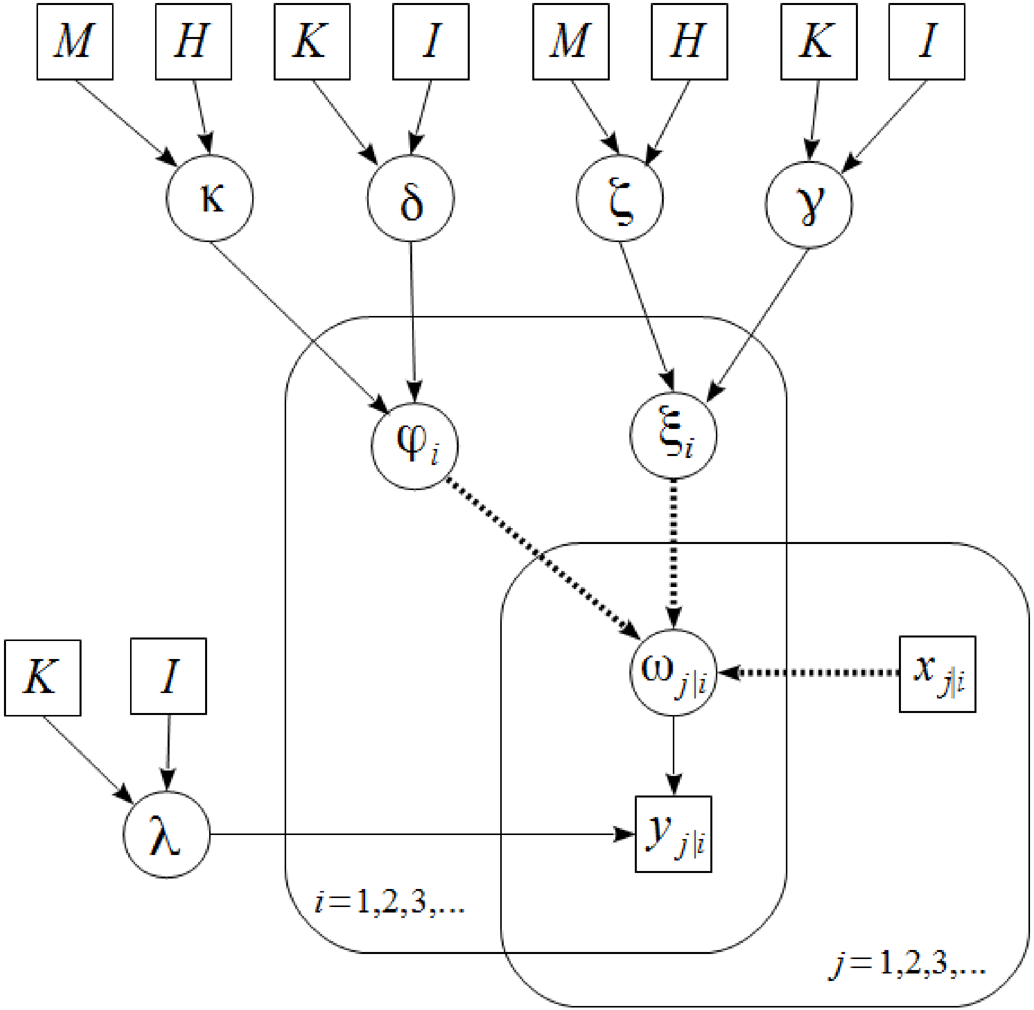
Posterior for Gaussian mean

$$\mu \mid x, \tau \sim \mathcal{N}\left(\frac{n\tau}{n\tau + n_0\tau}\bar{x} + \frac{n_0\tau}{n\tau + n_0\tau}\mu_0 \quad , \quad n\tau + n_0\tau \right)$$



https://towardsdatascience.com/a-zero-math-introduction-to-markov-chain-monte-carlo-methods-dcba889e0c50

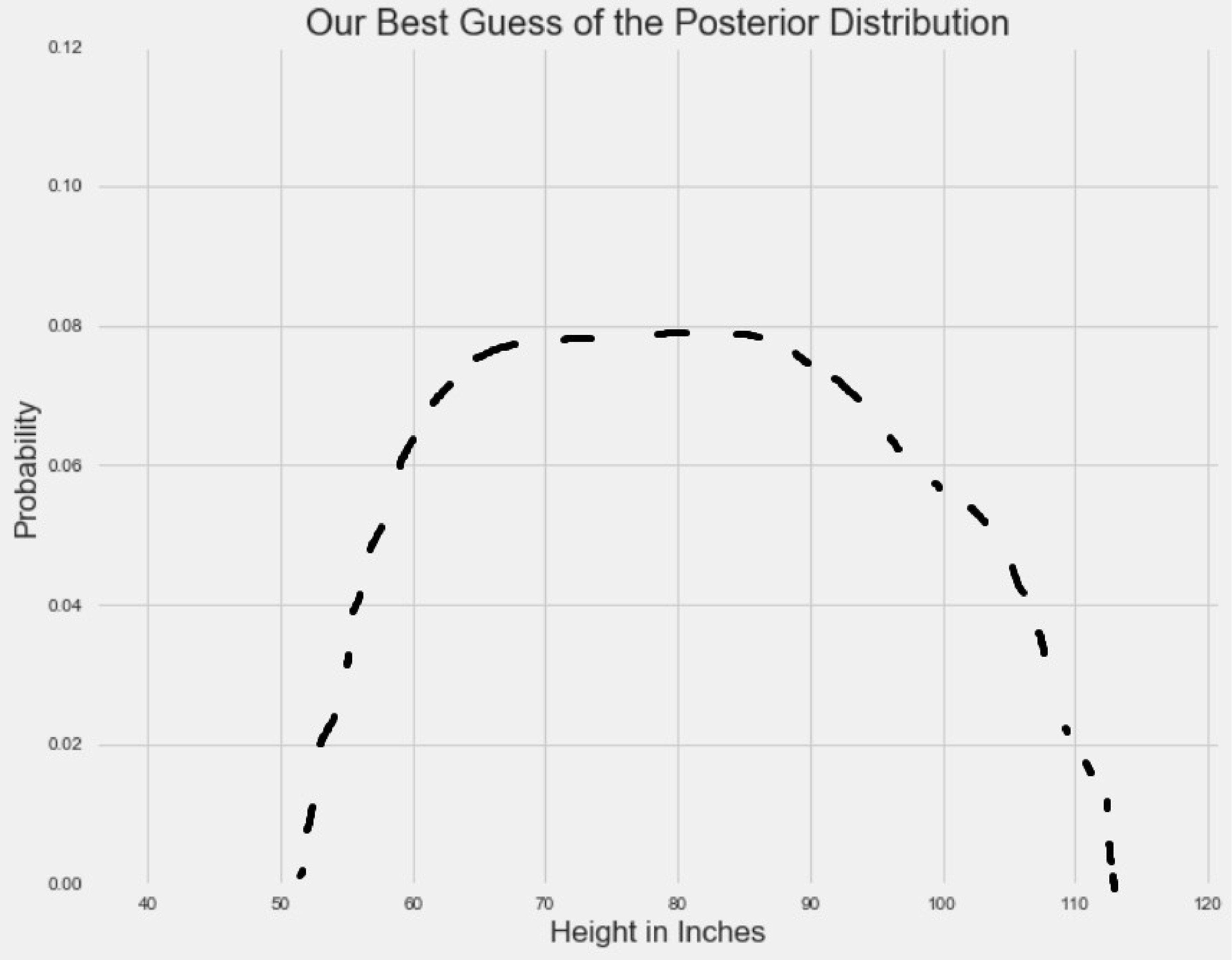




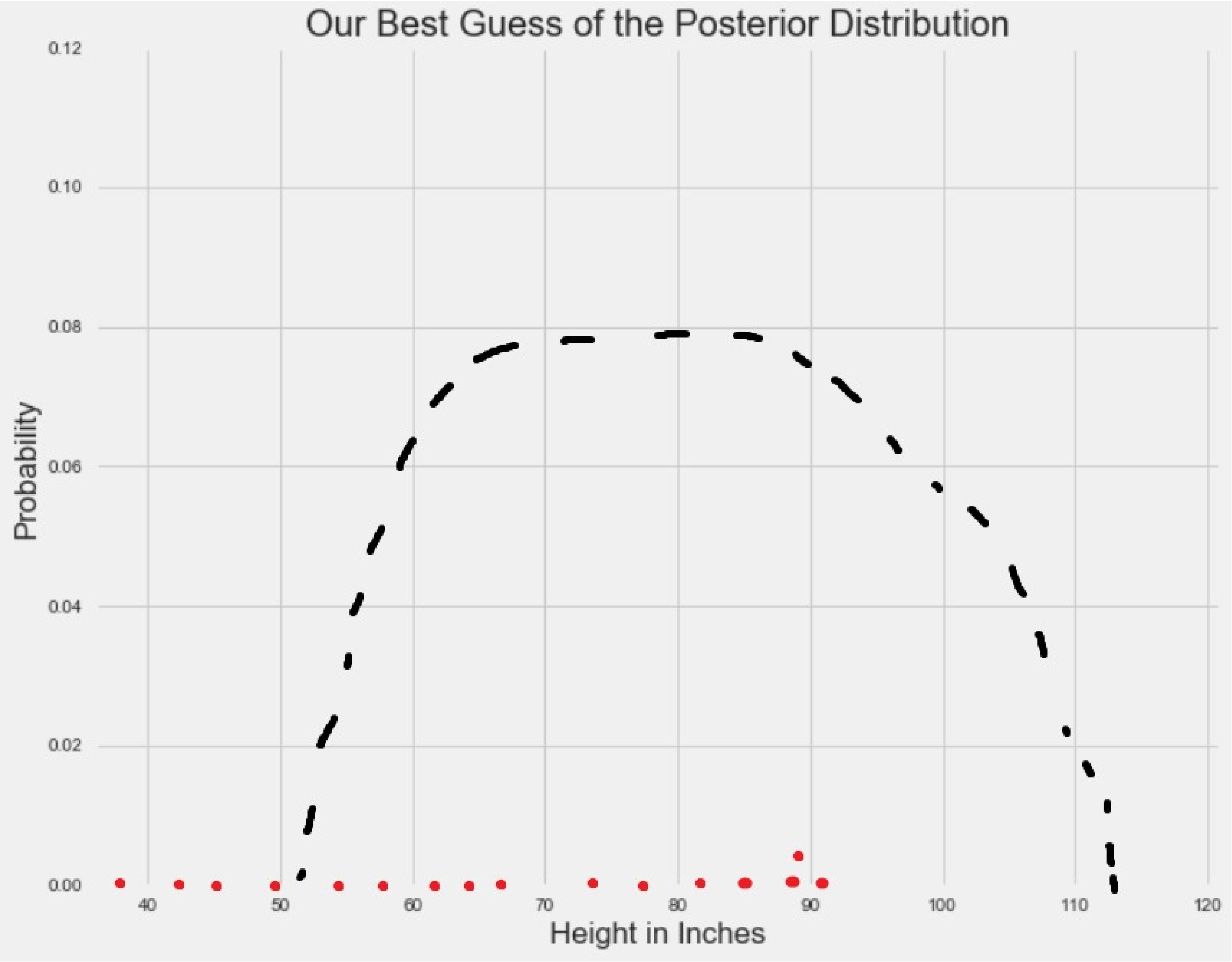
http://doingbayesiandataanalysis.blogspot.com/2013/10/diagrams-for-hierarchical-models-we.html

Estimating the posterior?

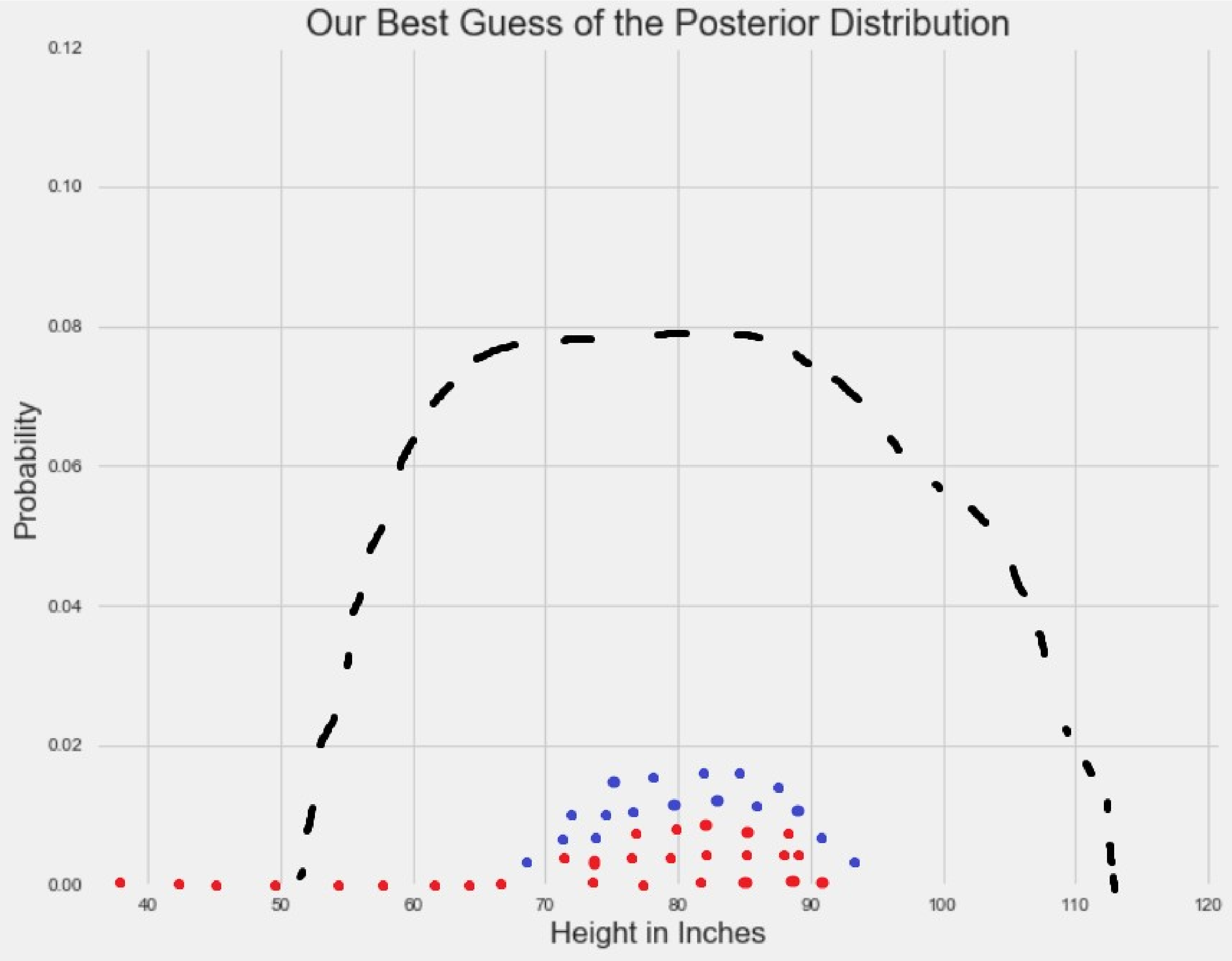
- Analytical solution..
- Gibbs sampling, Markov Chain Monte Carlo (MCMC)
- Variational Inference (VI)
- Hamiltonian Monte Carlo (HMC)
- Approximate Bayesian Computation (ABC)
- ...



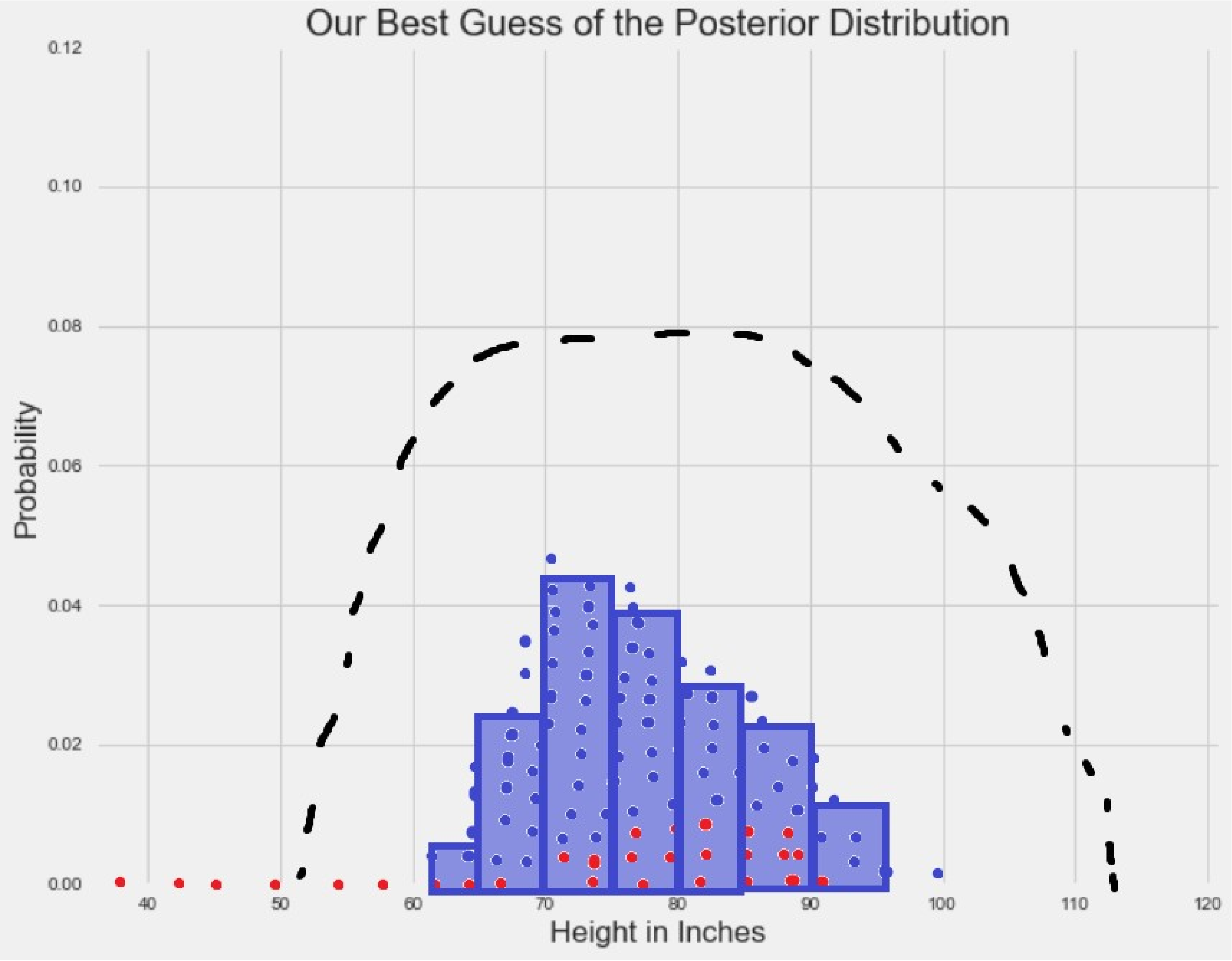
https://towardsdatascience.com/a-zero-math-introduction-to-markov-chain-monte-carlo-methods-dcba889e0c50



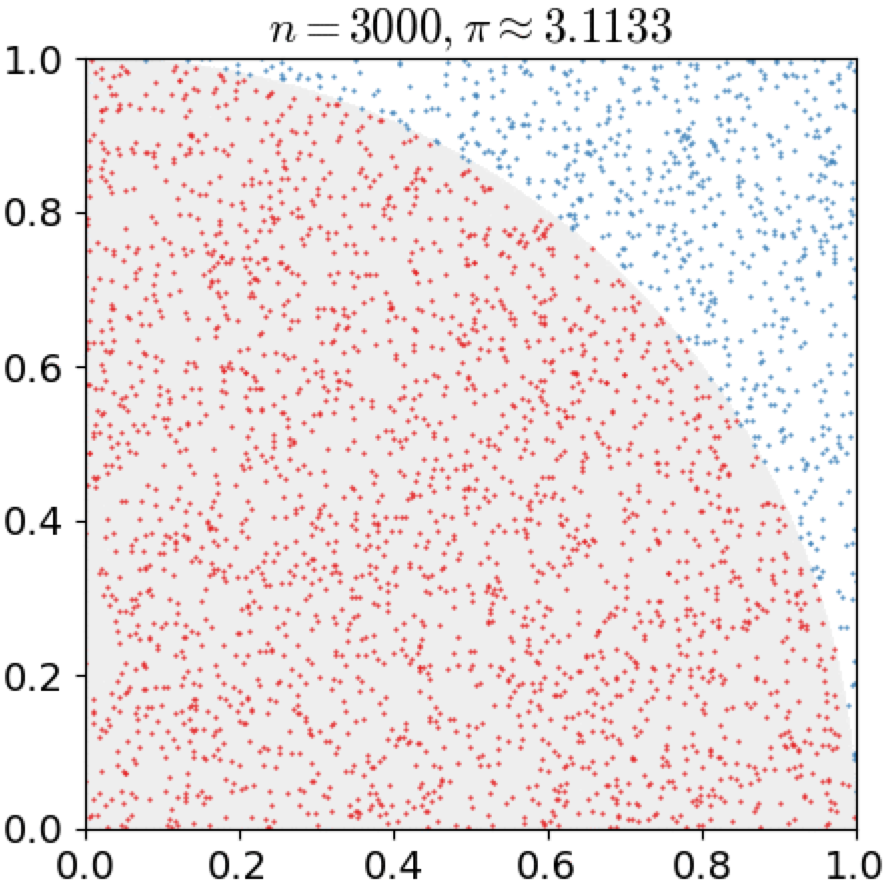
https://towardsdatascience.com/a-zero-math-introduction-to-markov-chain-monte-carlo-methods-dcba889e0c50



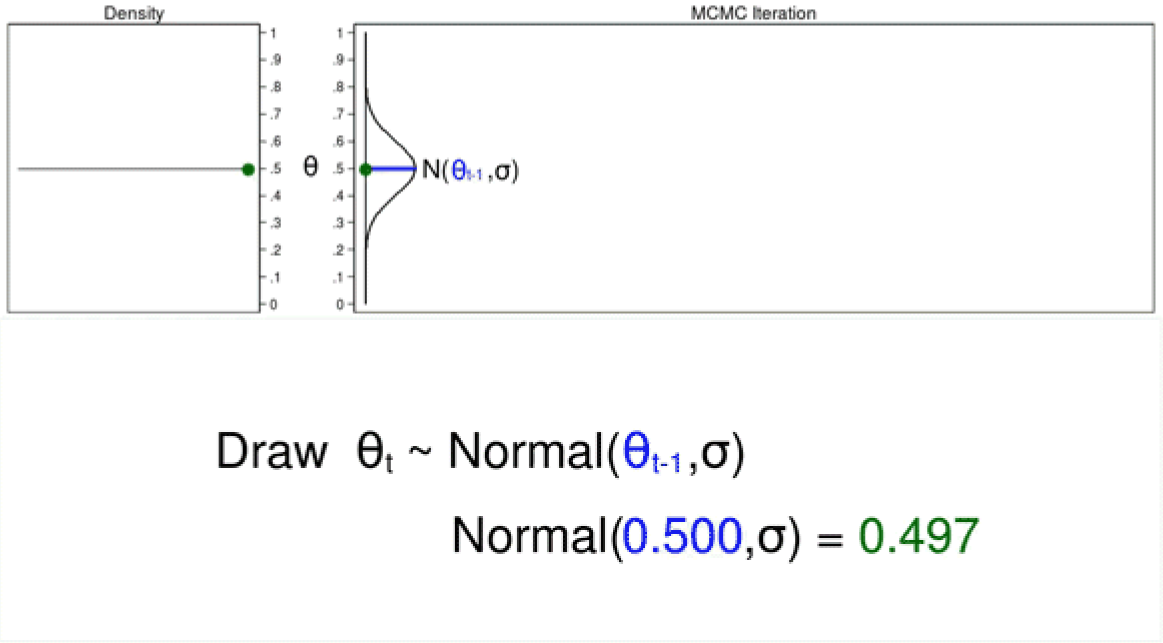
https://towardsdatascience.com/a-zero-math-introduction-to-markov-chain-monte-carlo-methods-dcba889e0c50



https://towardsdatascience.com/a-zero-math-introduction-to-markov-chain-monte-carlo-methods-dcba889e0c50



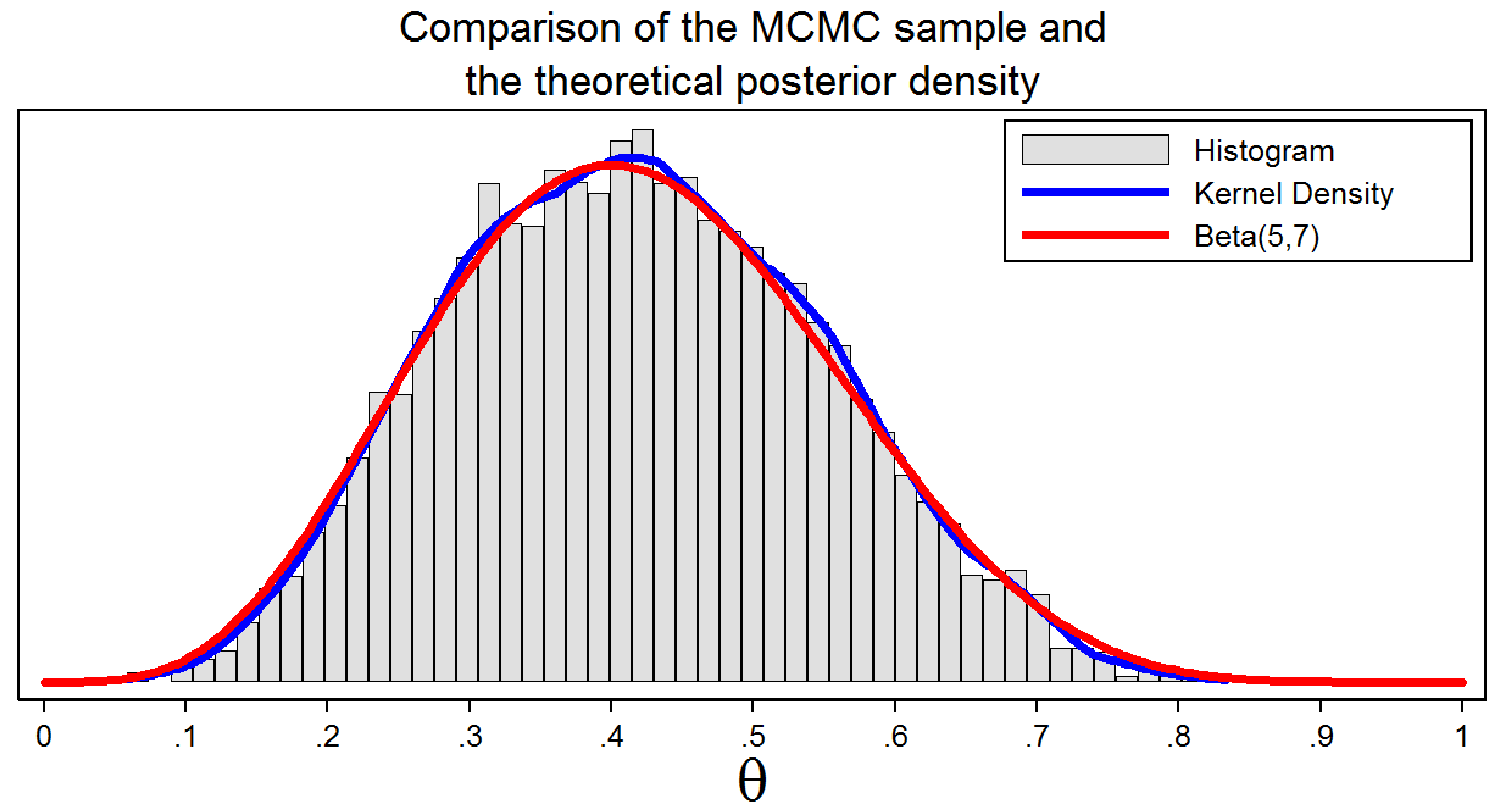
https://towardsdatascience.com/a-zero-math-introduction-to-markov-chain-monte-carlo-methods-dcba889e0c50



Draw $\theta_t \sim \text{Normal}(\theta_{t-1},\sigma)$

$\text{Normal}(0.500,\sigma) = 0.497$

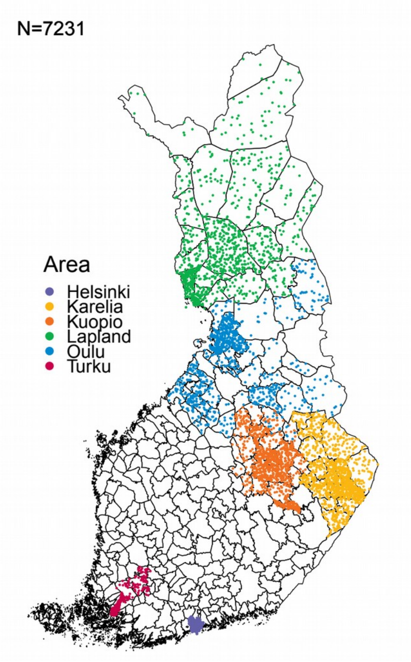
https://blog.stata.com/2016/11/15/introduction-to-bayesian-statistics-part-2-mcmc-and-the-metropolis-hastings-algorithm/


Comparison of the MCMC sample and
the theoretical posterior density
Histogram
Kernel Density
Beta(5,7)
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
θ

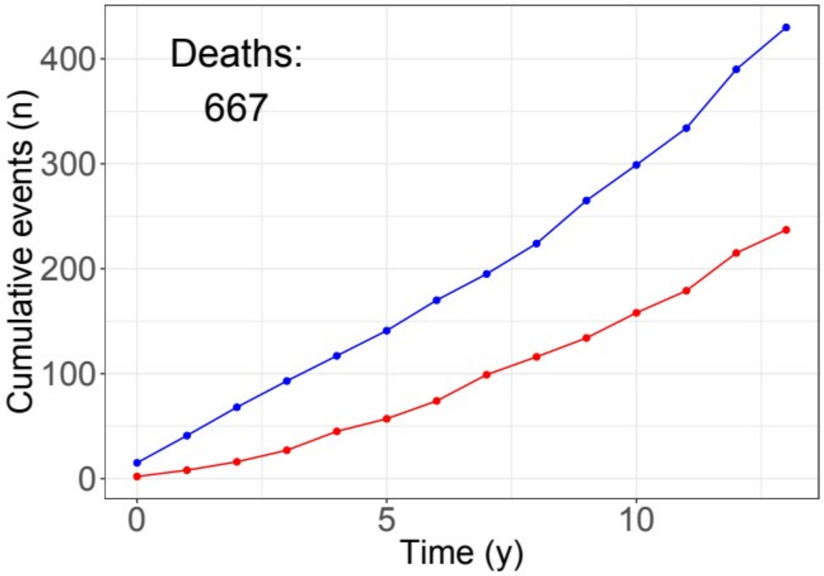
15 year **prospective** view (Finland / FINRISK2002)

2002 ~7000+ stool samples: omics and clinical measurements.

2017 comprehensive health information from Finnish registers



15+ year follow up

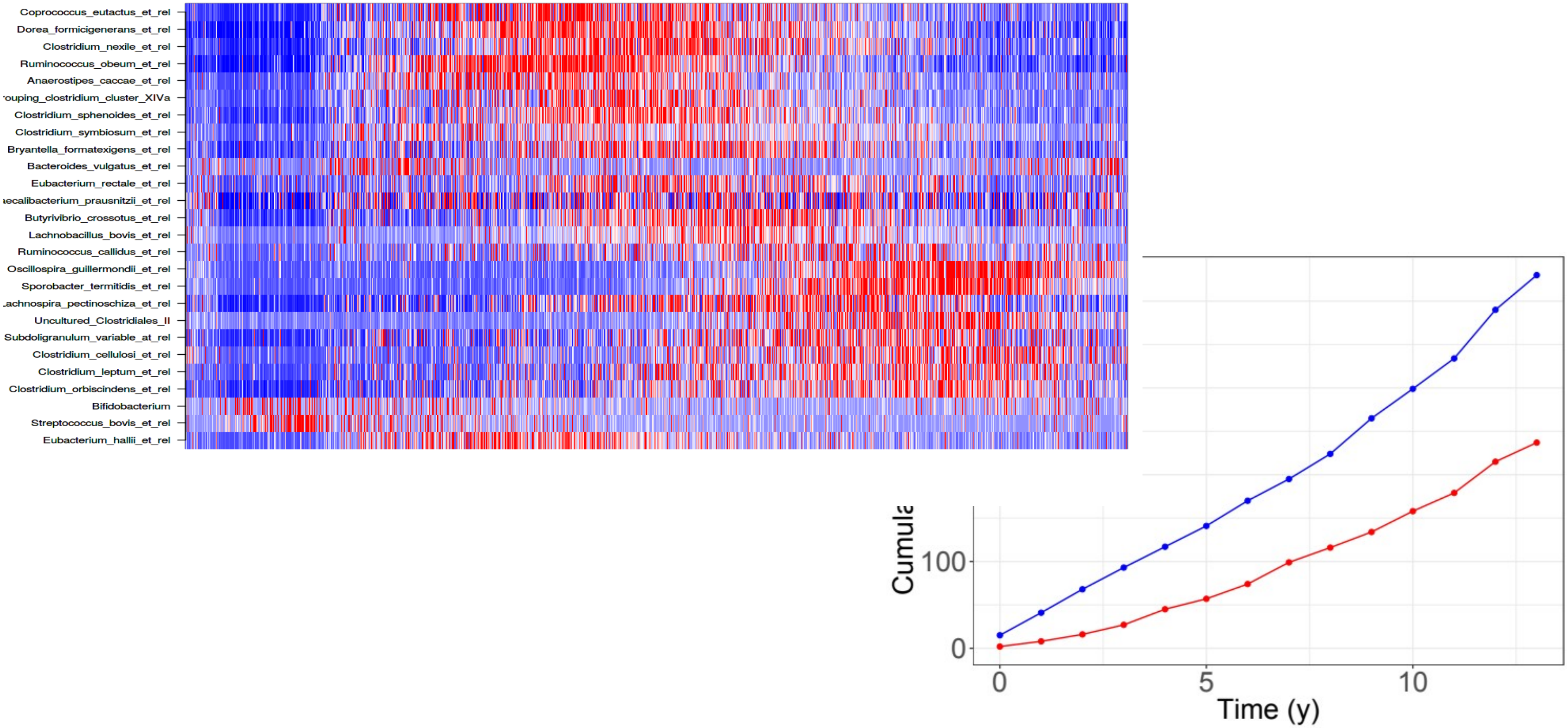


NATIONAL INSTITUTE FOR HEALTH AND WELFARE

With: Teemu Niiranen, Veikko Salomaa, Rob Knight, Mohit Jain, Mike Inouye, Aaro Salosensaari, Ville Laitinen, and others

Taxonomic profiles (N = 5005)

Z-score across subjects: red – high abundance & blue – low abundance - HITChip Atlas



Cox proportional hazards

Hazard function

Baseline hazard

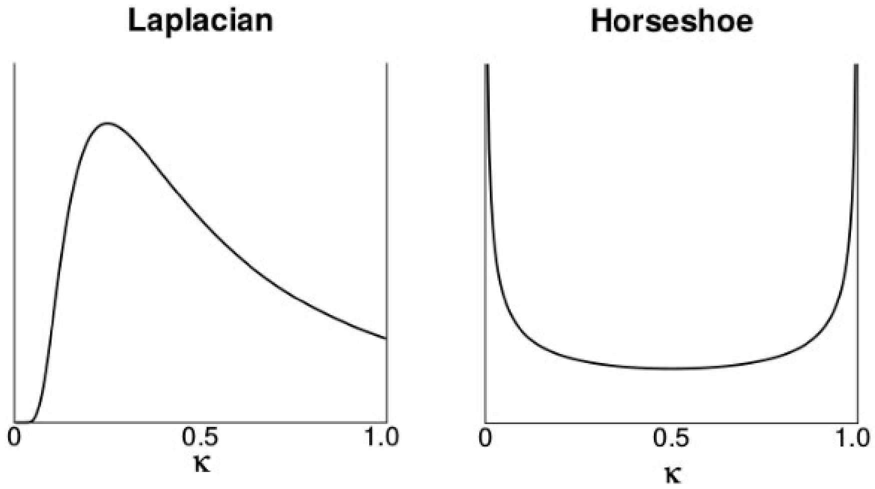
Data

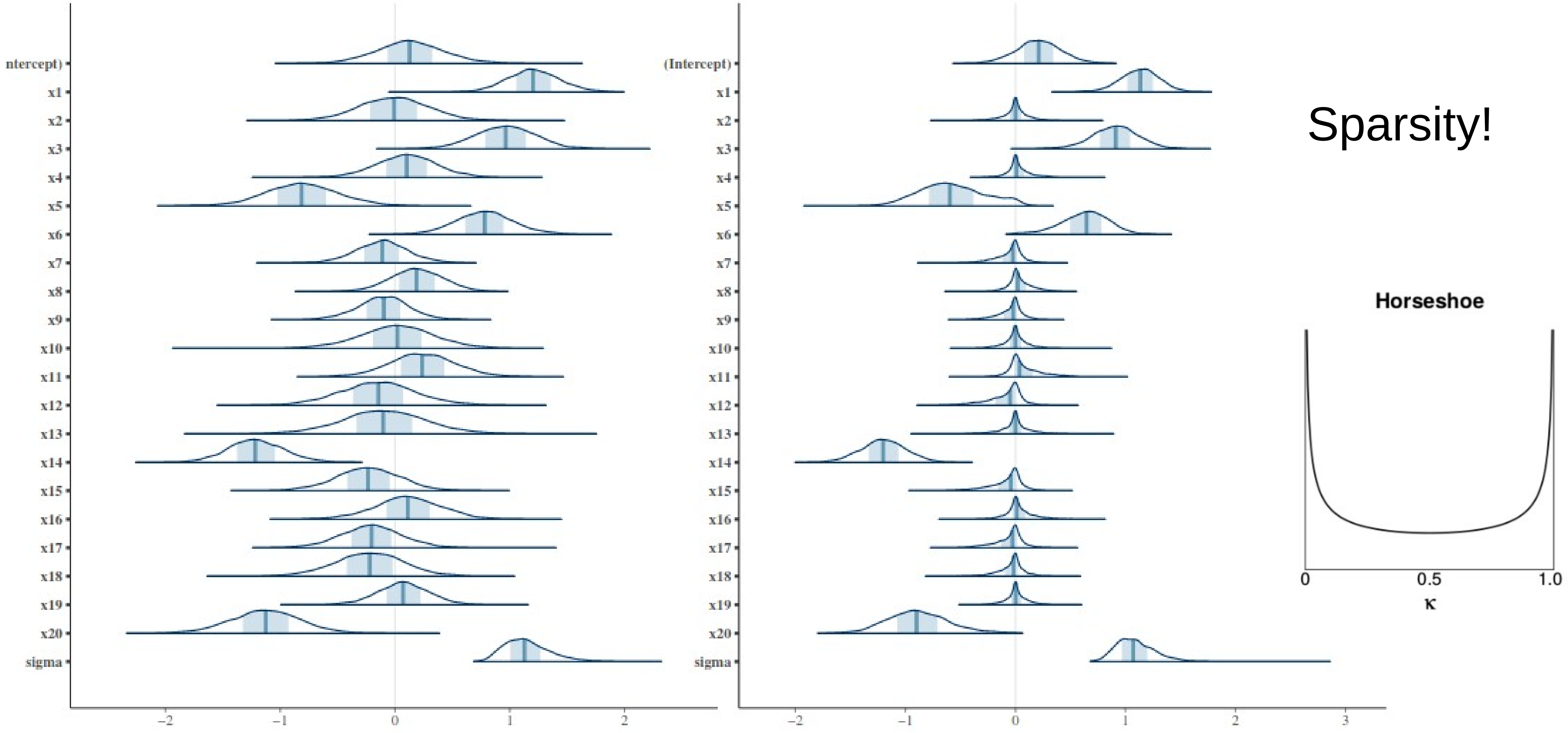
Coefficients

$$\lambda(t) = \lambda_0(t)\exp(\mathbf{x}\beta).$$

$$\lambda(t) = \lambda_0(t)\exp(\mathbf{x}\beta).$$

Alternative priors: P(β)





Gaussian prior

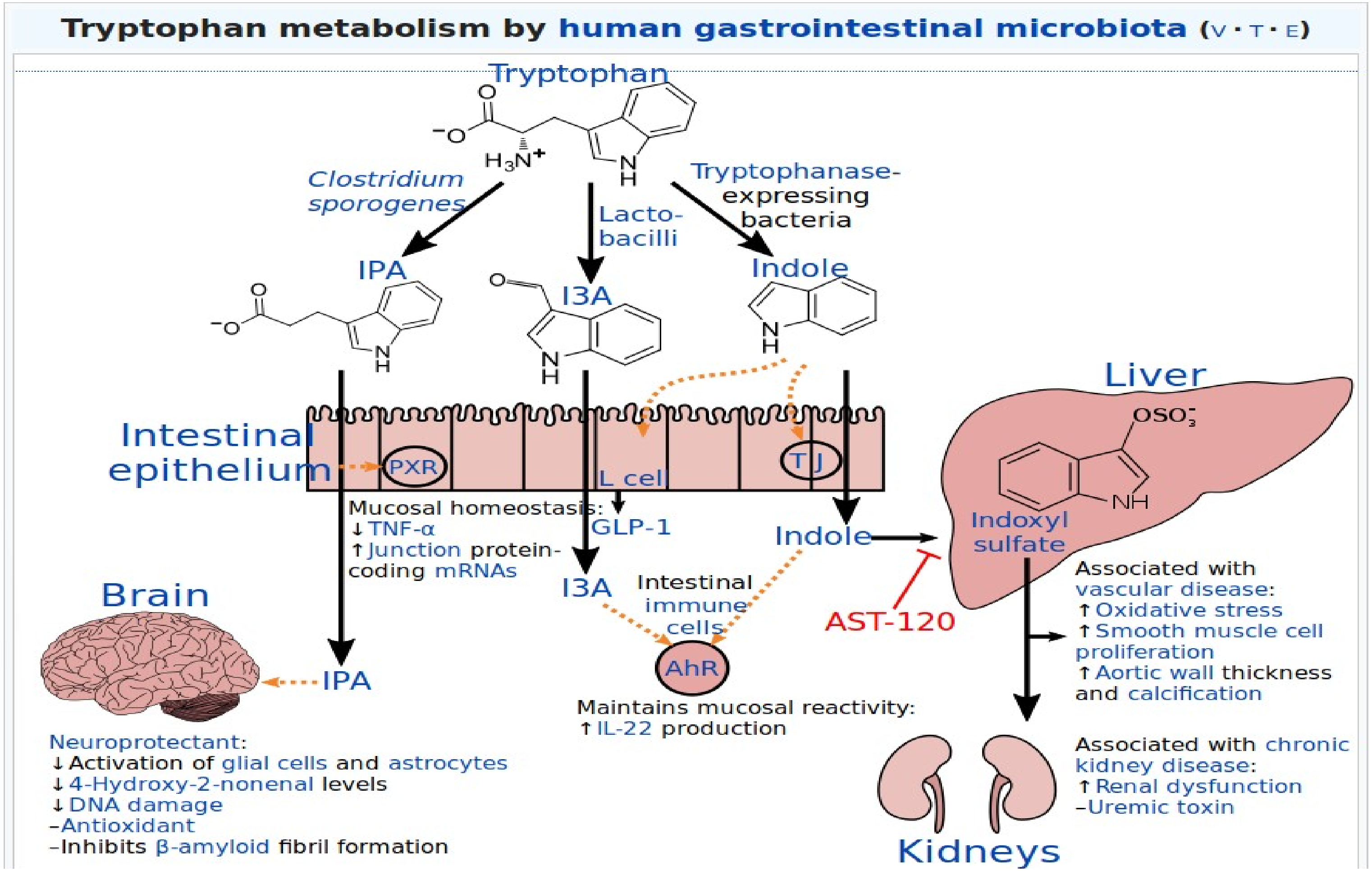
Horseshoe prior

Sparsity!

Aki Vehtari

Function, multi-omics, causality, mechanisms

Tryptophan metabolism by human gastrointestinal microbiota (V · T · E)



Mucosal homeostasis:
↓TNF-α
↑Junction protein-coding mRNAs

Maintains mucosal reactivity:
↑IL-22 production

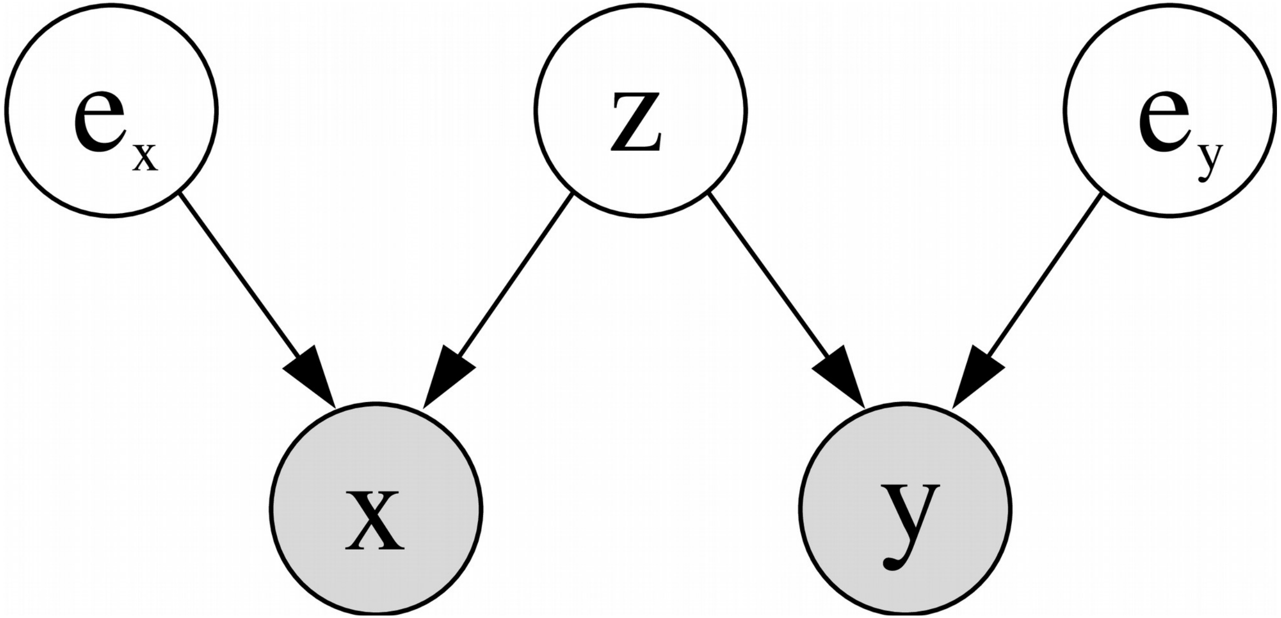
Neuroprotectant:
↓Activation of glial cells and astrocytes
↓4-Hydroxy-2-nonenal levels
↓DNA damage
–Antioxidant
–Inhibits β-amyloid fibril formation

Associated with vascular disease:
↑Oxidative stress
↑Smooth muscle cell proliferation
↑Aortic wall thickness and calcification

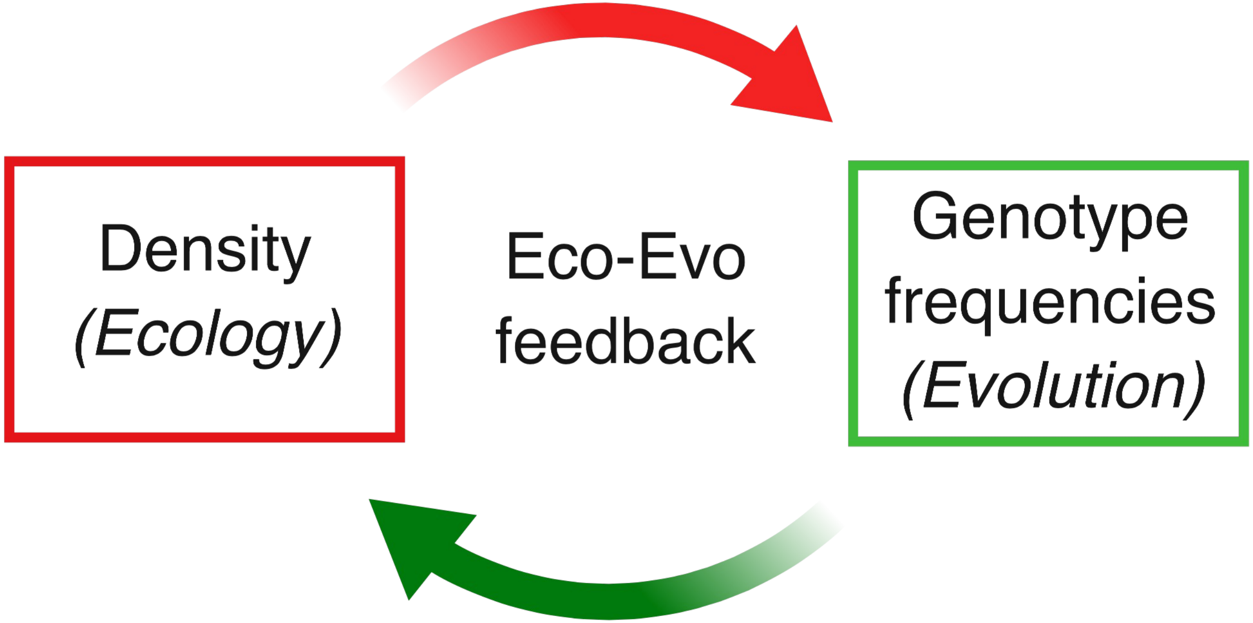
Associated with chronic kidney disease:
↑Renal dysfunction
–Uremic toxin

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

Multi-view learning



Ecology and evolution are connected processes



Density

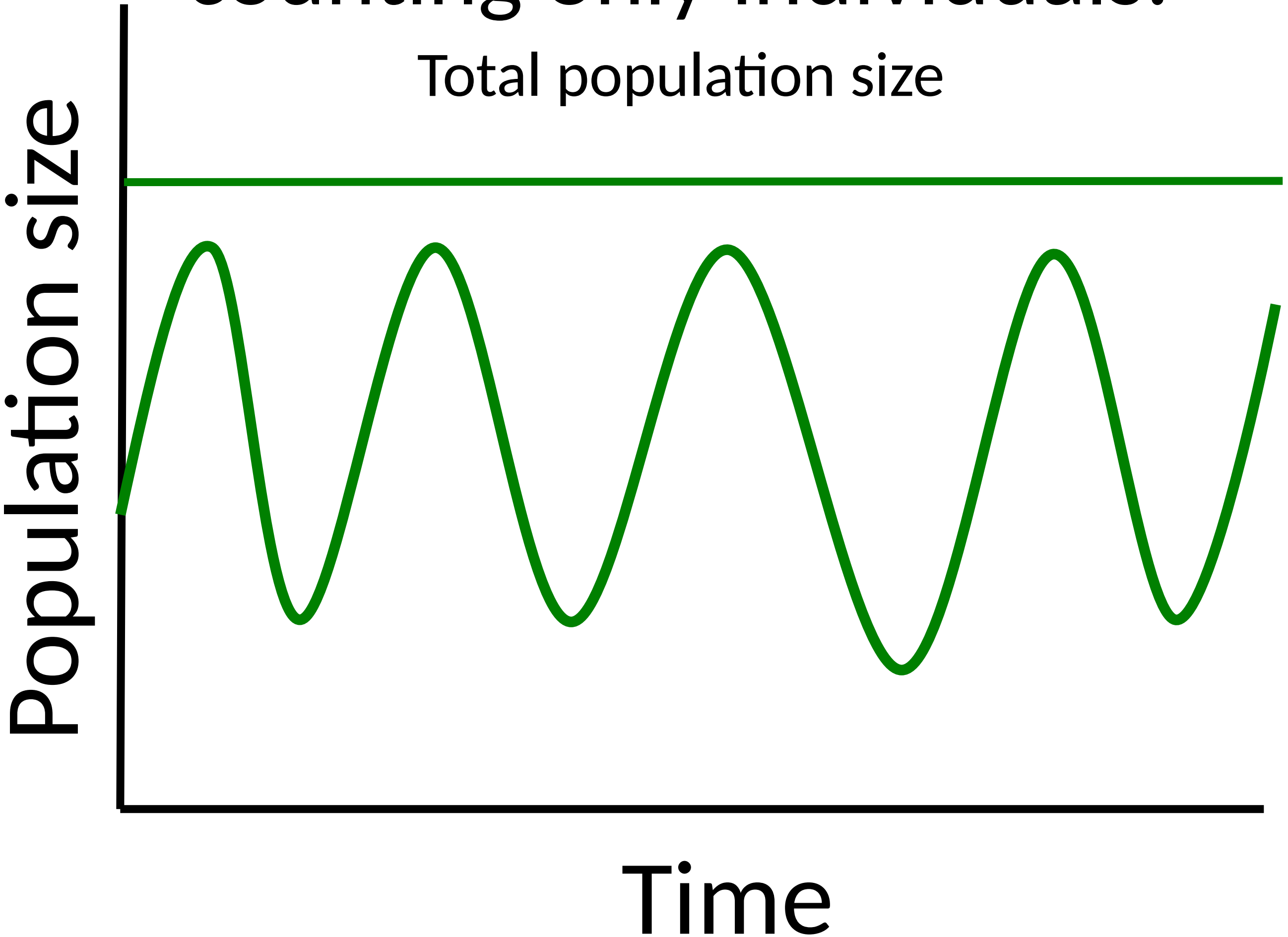
(Ecology)

Eco-Evo feedback

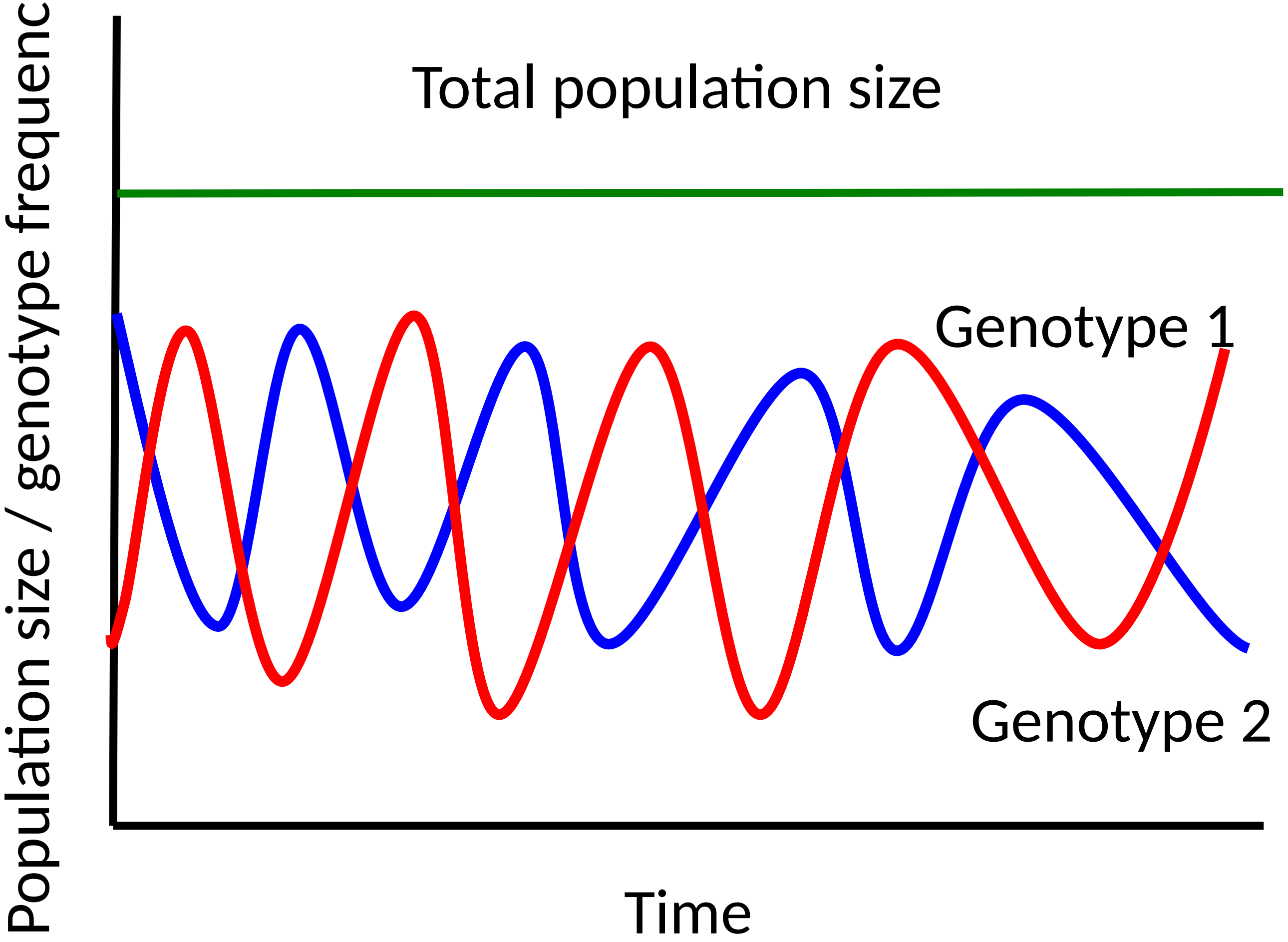
Genotype frequencies

(Evolution)

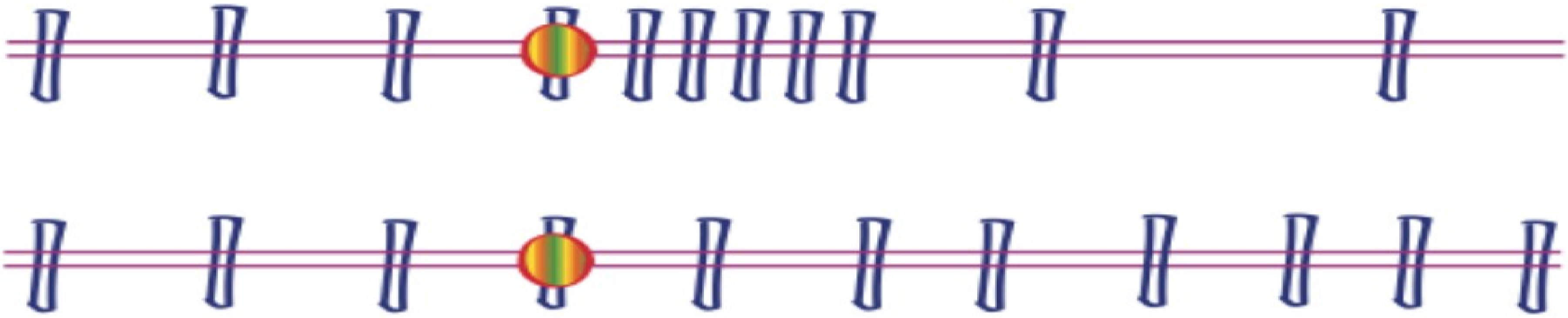
Traditional population level view when counting only individuals:



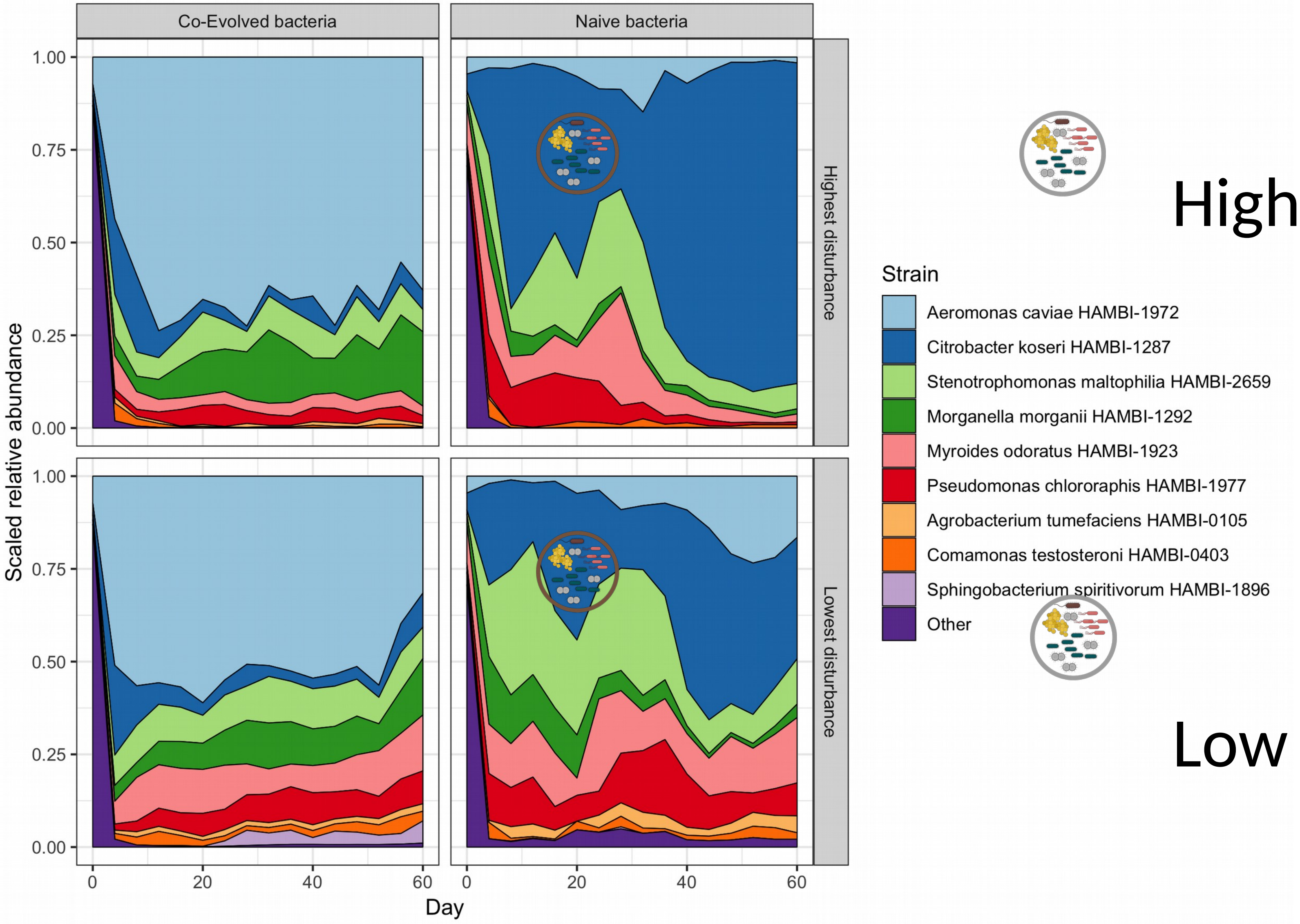
We need information about the underlying genotype dynamics as well



Longitudinal Data Analyses



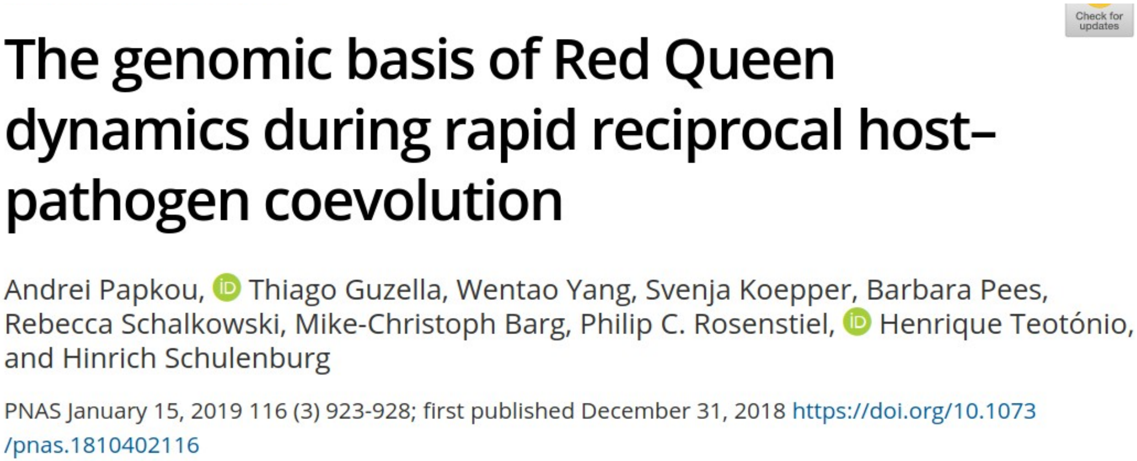
Source: Susan Holmes | http://web.stanford.edu/class/bios221/Short-Phyloseq-Resources.html



High disturbance

Low disturbance

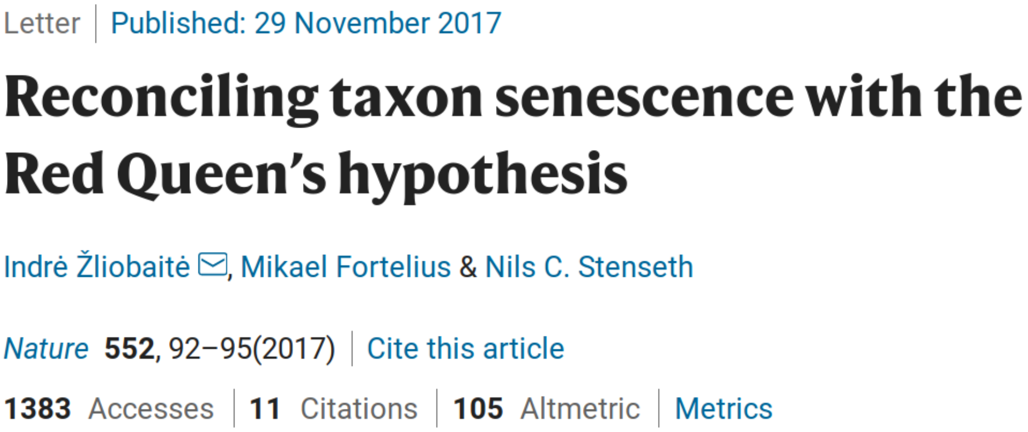
Eco-evo



The genomic basis of Red Queen dynamics during rapid reciprocal host–pathogen coevolution

Andrei Papkou, Thiago Guzella, Wentao Yang, Svenja Koepper, Barbara Pees, Rebecca Schalkowski, Mike-Christoph Barg, Philip C. Rosenstiel, Henrique Teotónio, and Hinrich Schulenburg

PNAS January 15, 2019 116 (3) 923-928; first published December 31, 2018 https://doi.org/10.1073/pnas.1810402116



Letter | Published: 29 November 2017

Reconciling taxon senescence with the Red Queen's hypothesis

Indrė Žliobaitė ✉, Mikael Fortelius & Nils C. Stenseth

Nature **552**, 92–95(2017) | Cite this article

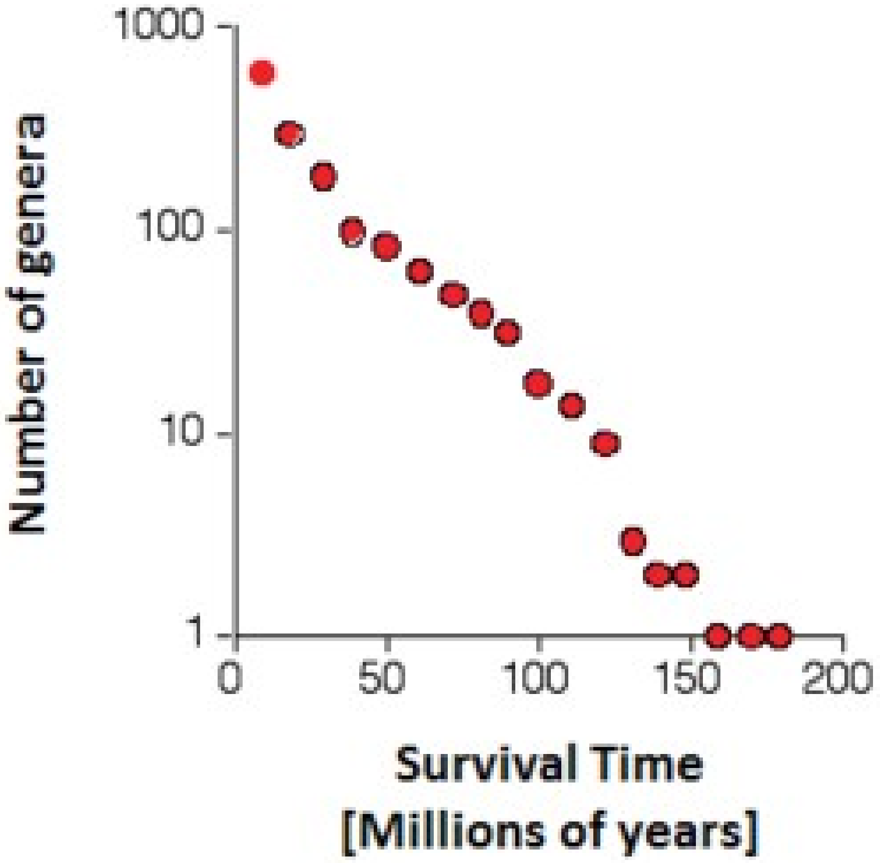
1383 Accesses | **11** Citations | **105** Altmetric | Metrics

Do species age?

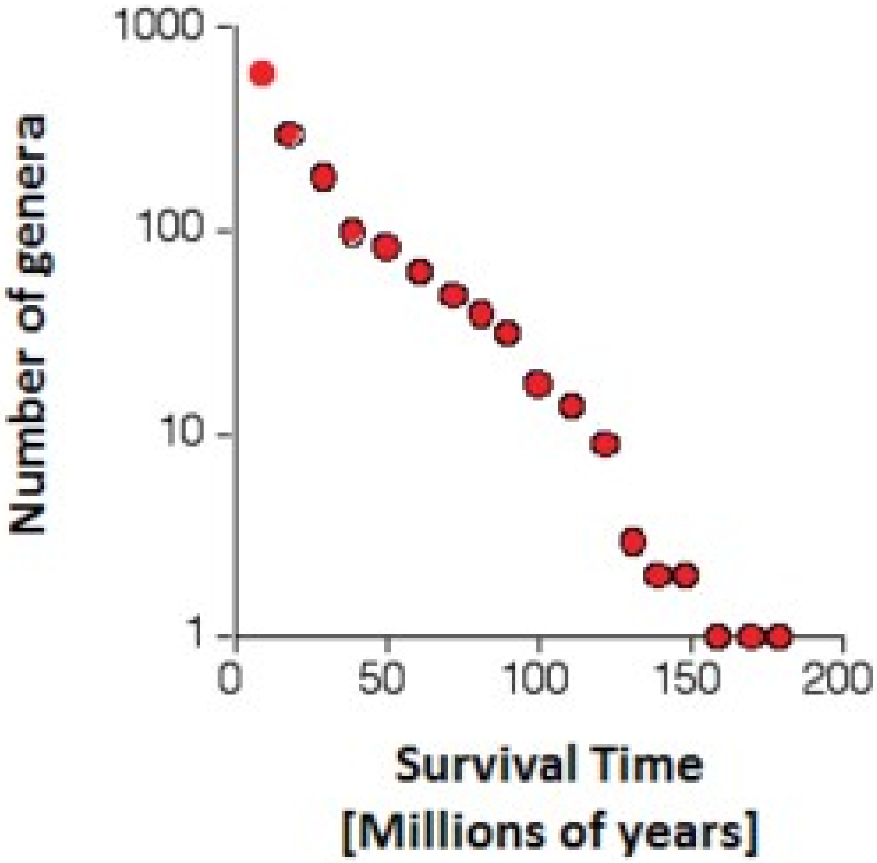
Hypothesis: the longer a species has been around, the more likely it will go extinct.

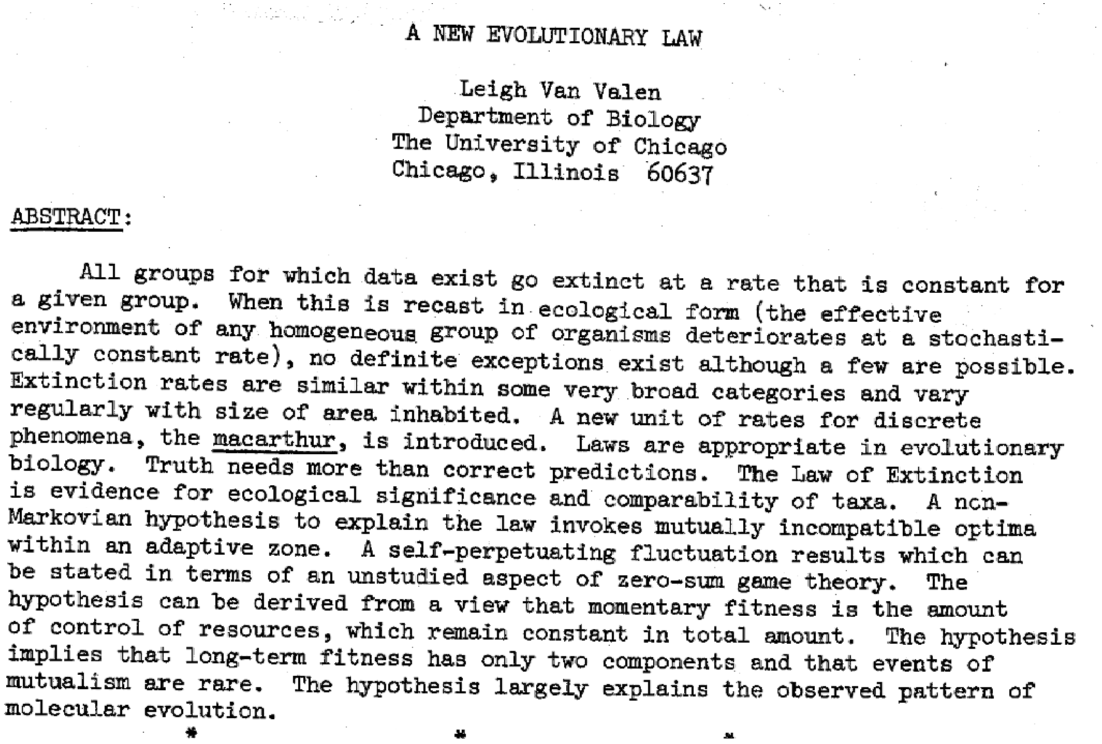
Rejection: Fossil data did not support this.

Old and young species have a similar probability to go extinct at any point.



(Leigh) Van Valen's "Law of Extinction": Probability of extinction does not depend on the lifetime of the population / "Half-life of species"





A NEW EVOLUTIONARY LAW

Leigh Van Valen
Department of Biology
The University of Chicago
Chicago, Illinois 60637

ABSTRACT:

All groups for which data exist go extinct at a rate that is constant for a given group. When this is recast in ecological form (the effective environment of any homogeneous group of organisms deteriorates at a stochastically constant rate), no definite exceptions exist although a few are possible. Extinction rates are similar within some very broad categories and vary regularly with size of area inhabited. A new unit of rates for discrete phenomena, the macarthur, is introduced. Laws are appropriate in evolutionary biology. Truth needs more than correct predictions. The Law of Extinction is evidence for ecological significance and comparability of taxa. A ncn-Markovian hypothesis to explain the law invokes mutually incompatible optima within an adaptive zone. A self-perpetuating fluctuation results which can be stated in terms of an unstudied aspect of zero-sum game theory. The hypothesis can be derived from a view that momentary fitness is the amount of control of resources, which remain constant in total amount. The hypothesis implies that long-term fitness has only two components and that events of mutualism are rare. The hypothesis largely explains the observed pattern of molecular evolution.

* * *

Van Valen, Leigh (1973). "A new evolutionary law". Evolutionary Theory. 1: 1–30.

Explanation?

Red Queen Hypothesis: organisms must constantly adapt, evolve, and proliferate in order to survive. A classic evolutionary theory (Van Valen, 1973).

Red Queen explains how her country differs from Alice’s:

“Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!"

A search for ‘**Red Queen**’ on Google Scholar gives over a million hits, reflecting the enormous influence this idea has had and continues to have in a wide sector of science.

Image credit: Ika Österblad

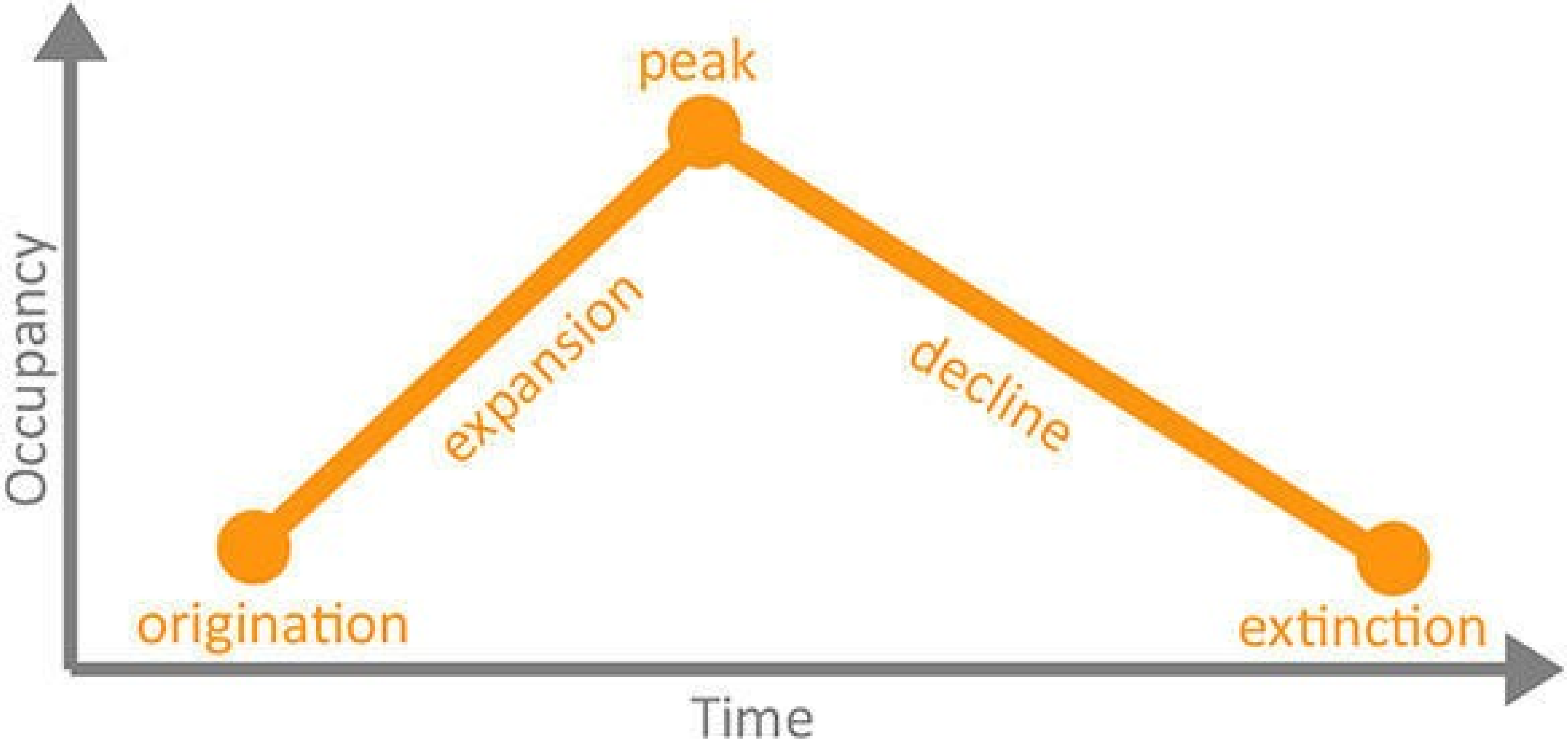
https://www.helsinki.fi/en/news/science-news/new-interpretation-of-the-red-queens-hypothesis-its-about-expansion

> Stenseth: “The Red Queen's Hypothesis has fascinated me from the very beginning since it, as an evolutionary hypothesis, explicitly brings in ecological interactions to explain large scale evolutionary patterns, such as rate of extinctions."

Image credit: Ika Österblad

Hat pattern

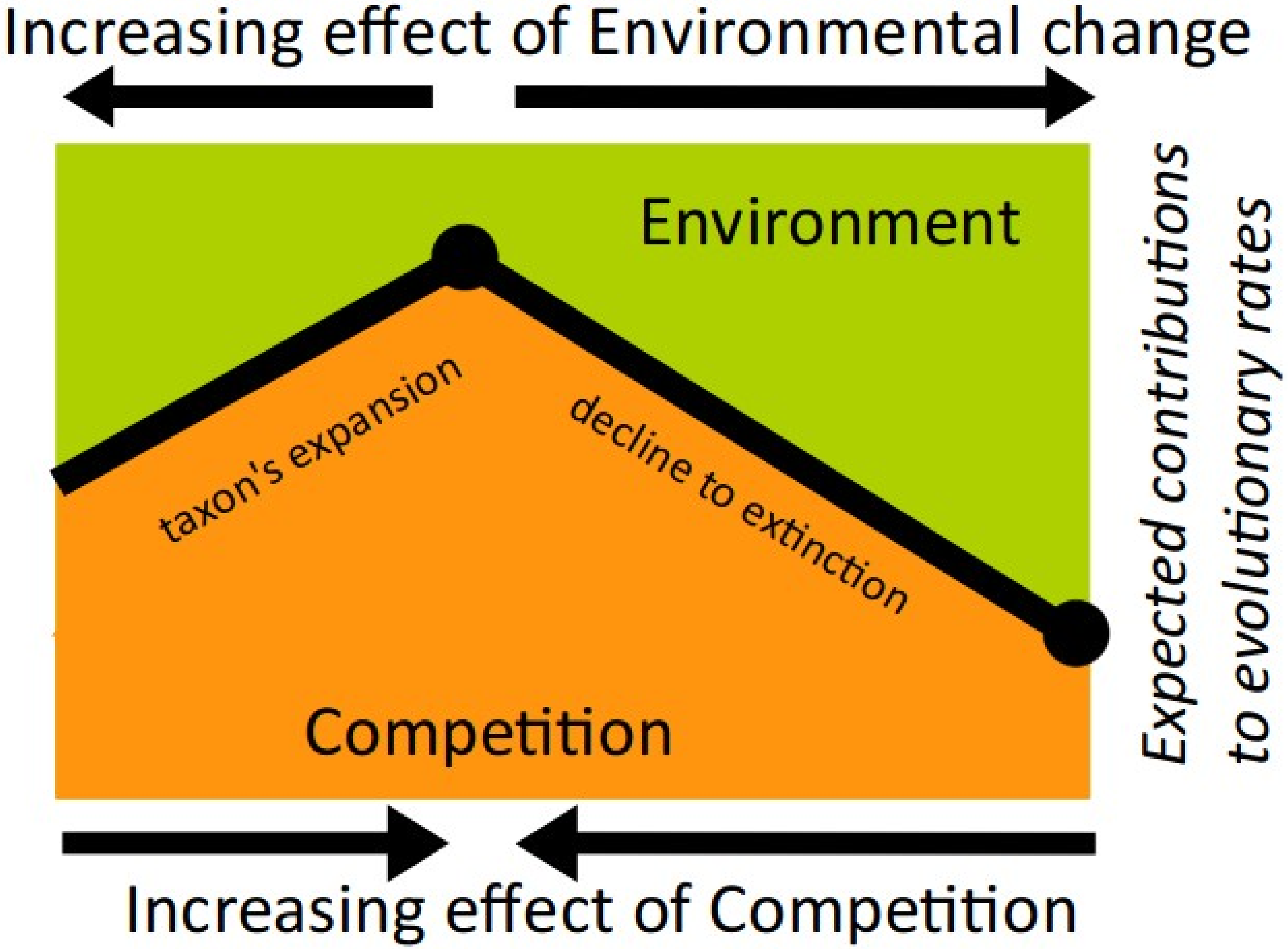
Contradiction! Jernvall & Fortelius (2004). Do species age or not?



Originally reported in: Jernvall, J.&Fortelius, M.Maintenance of trophic structure in fossil mammal communities: site occupancy and taxon resilience. The American Naturalist164, 614-624 (2004).

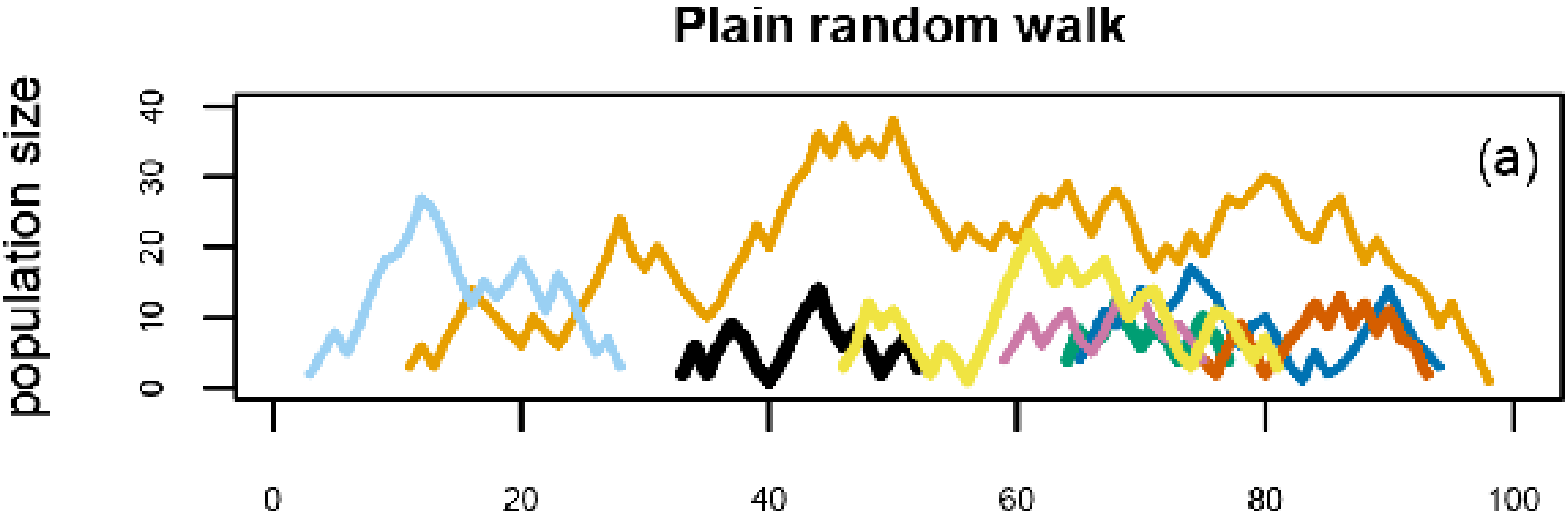
Expansion, stabilization, decline

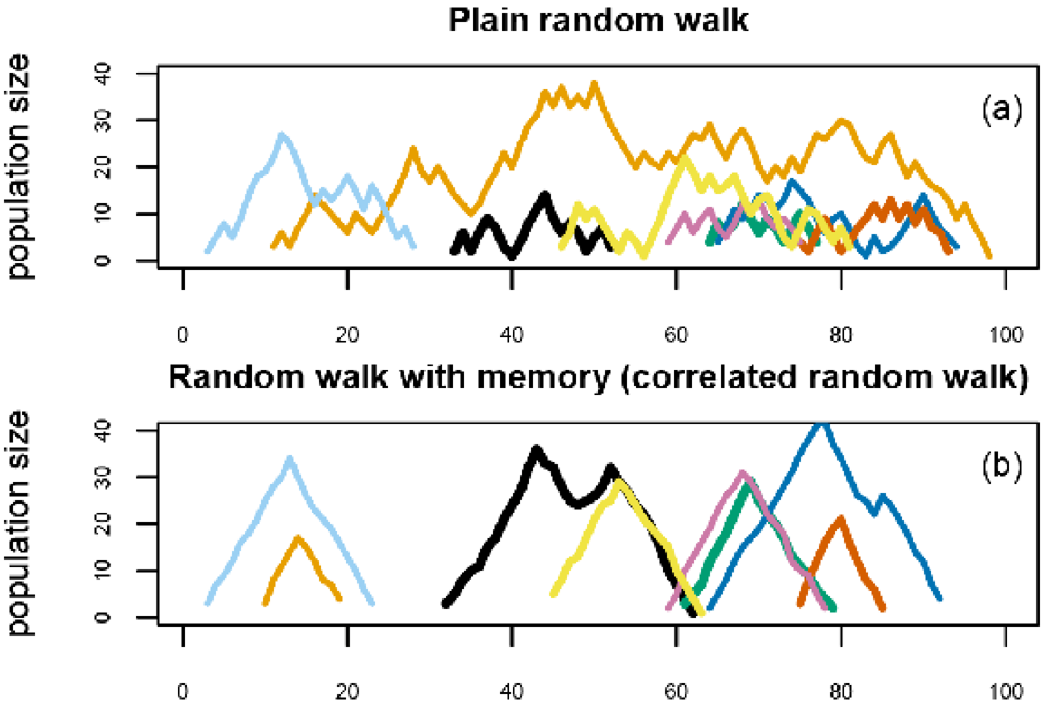
*"It takes all the running you can do, to keep in the same place**. If you want to get somewhere else, you must run at least twice as fast as that!"***


Increasing effect of Environmental change
Environment
taxon's expansion
decline to extinction
Competition
Expected contributions to evolutionary rates
Increasing effect of Competition

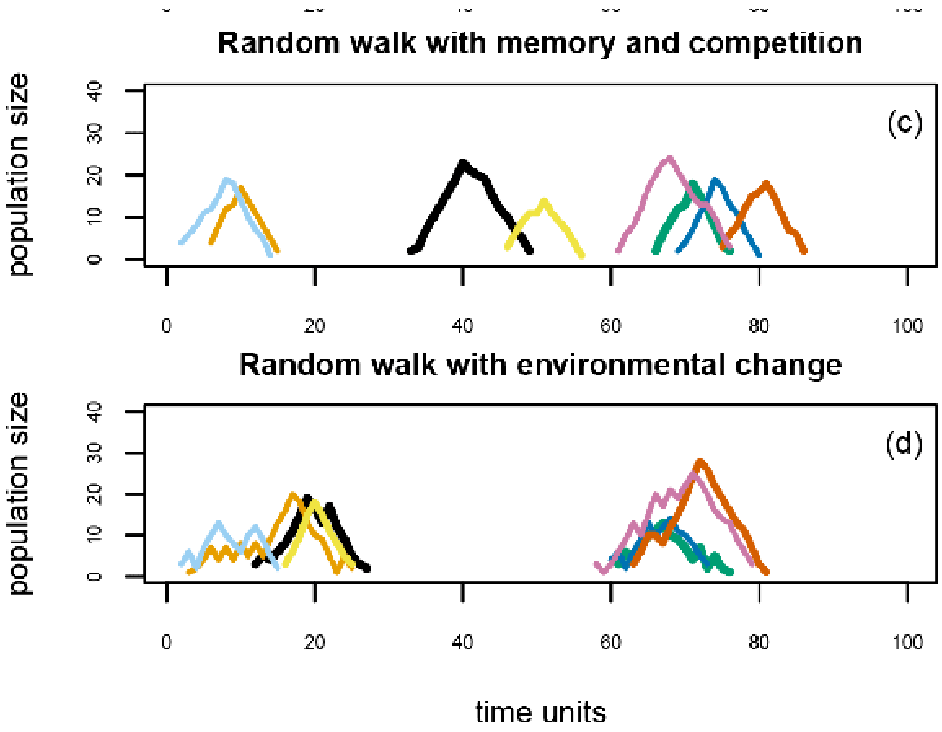
It is likely that a species will get extinct by random chance in the long run.

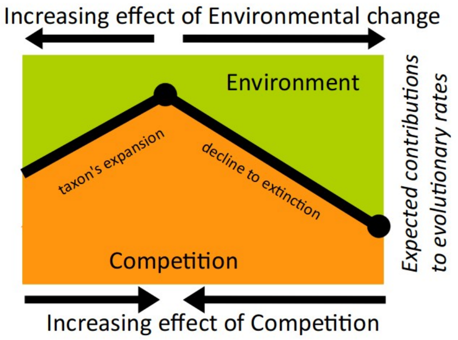
More than three peaks are very seldom observed in real data.





Memory effects emphasize the Red Queen effect.





Increasing effect of Environmental change

Environment

taxon's expansion

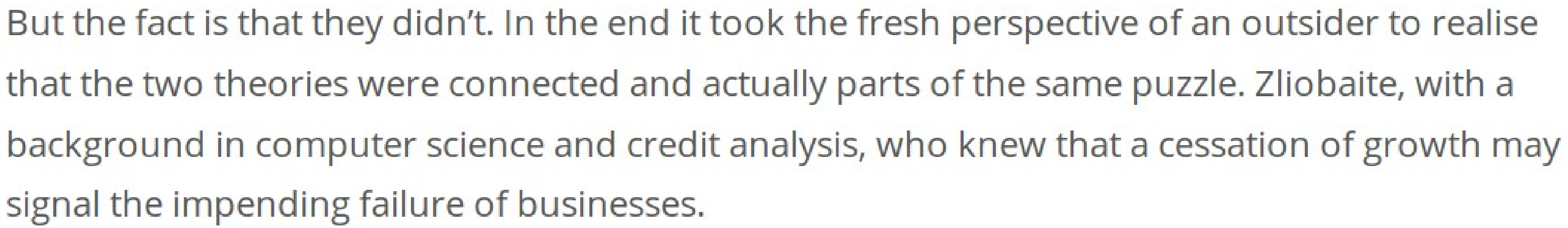
decline to extinction

Competition

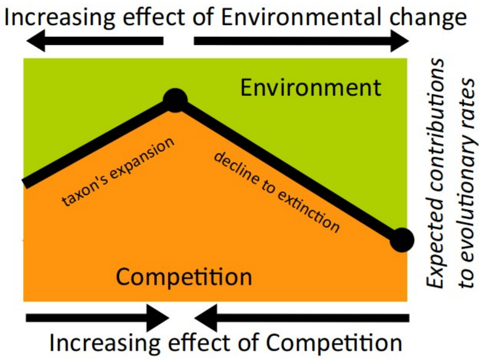
Expected contributions to evolutionary rates

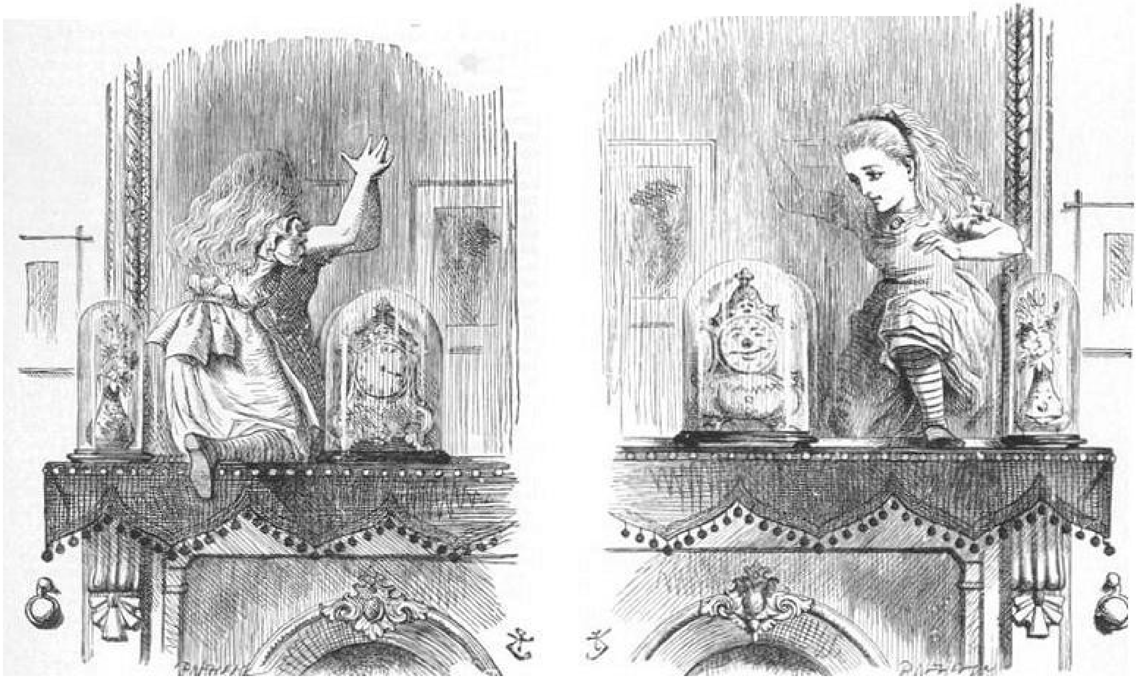
Increasing effect of Competition

Resolution: Hat Pattern *is* compatible with the Law of Constant Extinction, and predicted by the Red Queen’s Hypothesis.



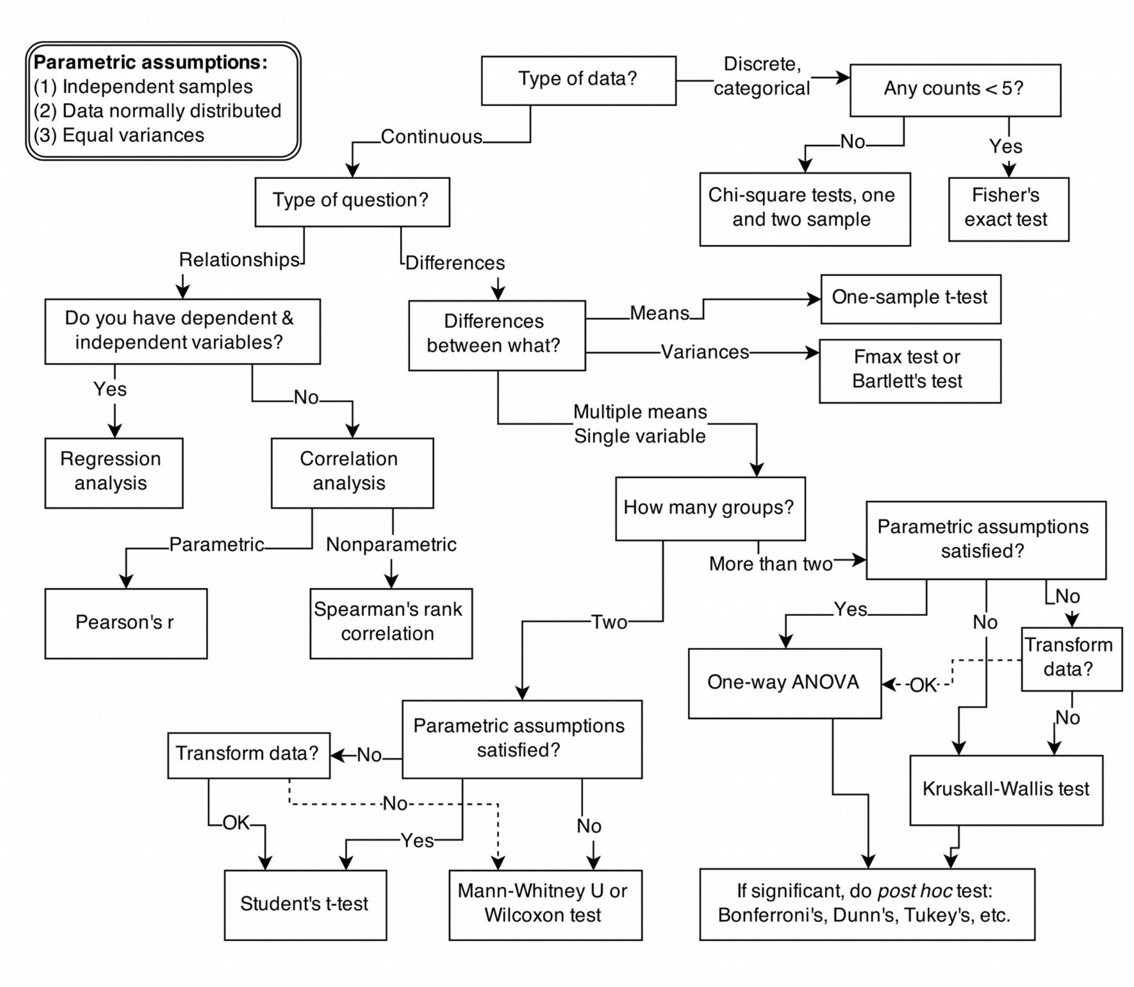
But the fact is that they didn’t. In the end it took the fresh perspective of an outsider to realise that the two theories were connected and actually parts of the same puzzle. Zliobaite, with a background in computer science and credit analysis, who knew that a cessation of growth may signal the impending failure of businesses.


Increasing effect of Environmental change
Environment
taxon's expansion
decline to extinction
Competition
Increasing effect of Competition
Expected contributions to evolutionary rates



Through the Looking-Glass, and What Alice Found There (Lewis Carroll, 1871)

How to choose a correct model?



Parametric assumptions:
(1) Independent samples
(2) Data normally distributed
(3) Equal variances

Type of data?

Discrete, categorical

Any counts < 5?

No

Yes

Chi-square tests, one and two sample

Fisher's exact test

Continuous

Type of question?

Relationships

Differences

Do you have dependent & independent variables?

Yes

No

Regression analysis

Correlation analysis

Parametric

Nonparametric

Pearson's r

Spearman's rank correlation

Differences between what?

Means

Variances

One-sample t-test

Fmax test or Bartlett's test

Multiple means Single variable

How many groups?

More than two

Two

Parametric assumptions satisfied?

Yes

No

No

Transform data?

OK

No

One-way ANOVA

Kruskall-Wallis test

If significant, do *post hoc* test: Bonferroni's, Dunn's, Tukey's, etc.

Parametric assumptions satisfied?

No

Transform data?

OK

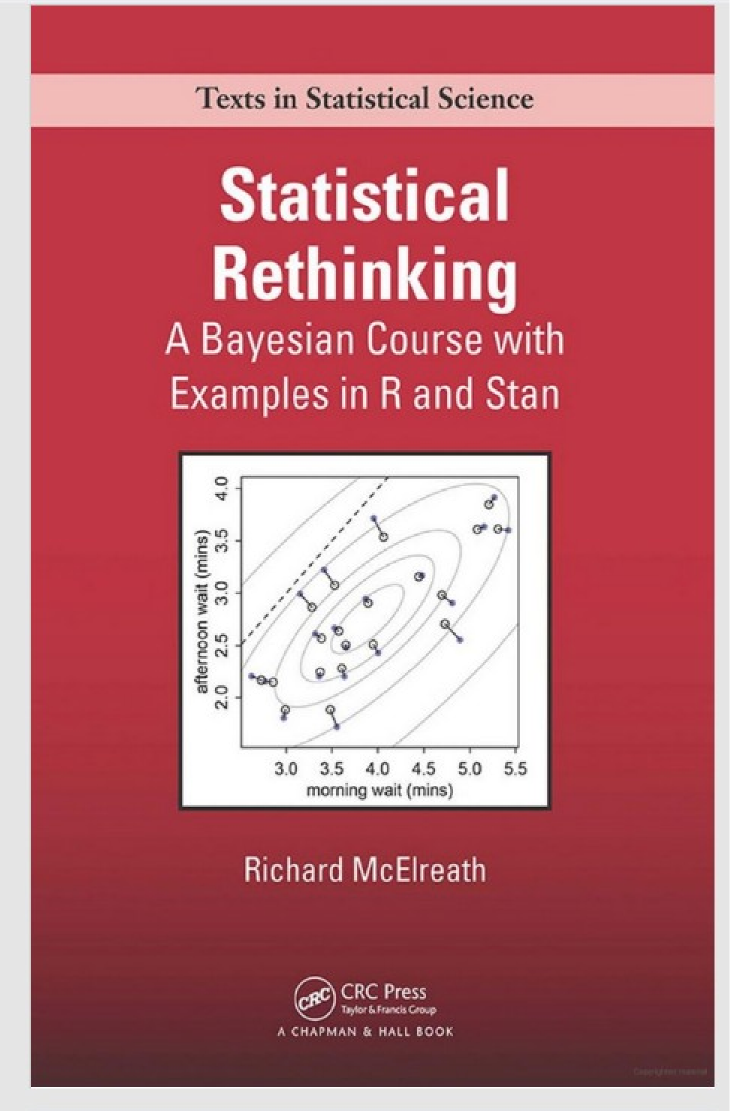
No

Yes

No

Student's t-test

Mann-Whitney U or Wilcoxon test



Texts in Statistical Science

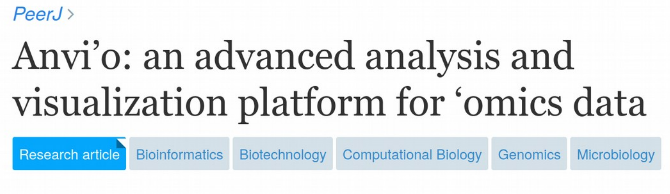
Statistical Rethinking

A Bayesian Course with Examples in R and Stan

Richard McElreath

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Anvi'o in a nutshell



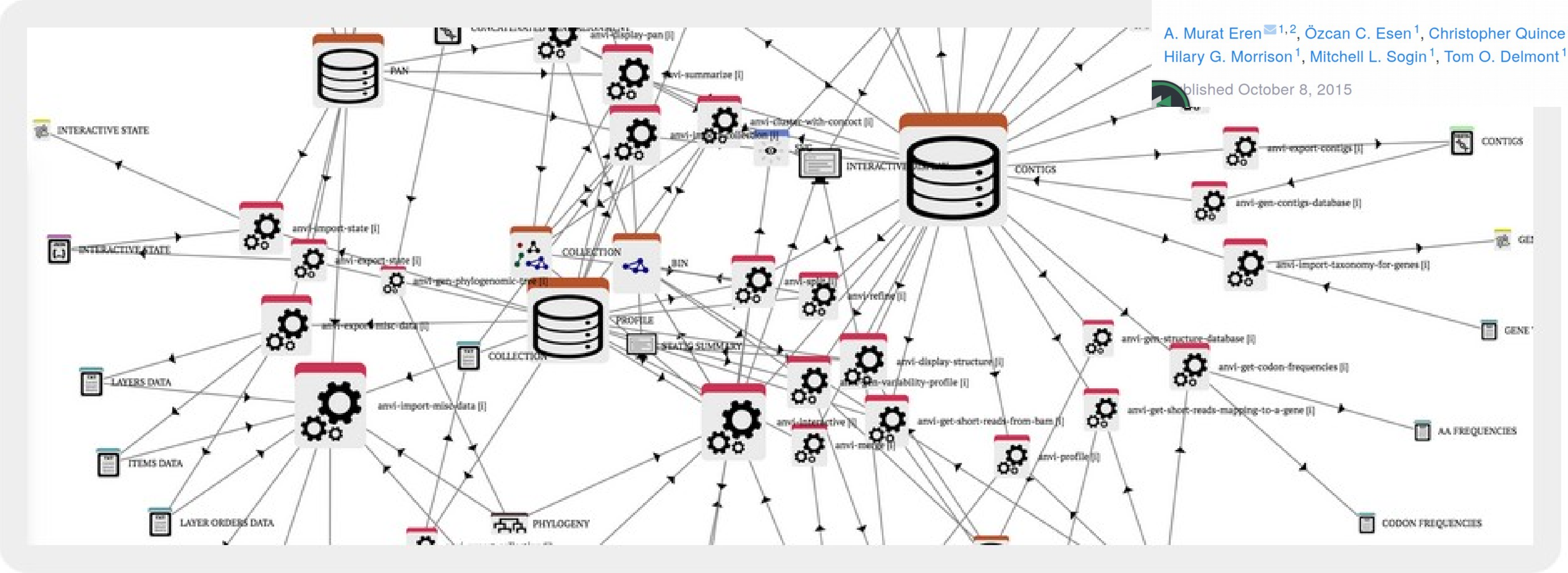
PeerJ ›

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-display-pan [I]
PAN
anvi-summarize [I]
INTERACTIVE STATE
anvi-cluster-with-contrast [I]
anvi-import-collection [I]
INTERACTIVE
CONTIGS
anvi-export-contigs [I]
CONTIGS
anvi-gen-contigs-database [I]
anvi-import-state [I]
INTERACTIVE STATE
anvi-export-state [I]
anvi-gen-phylogenomic-tree [I]
COLLECTION
BIN
anvi-split [I]
anvi-refine [I]
anvi-import-taxonomy-for-genes [I]
GENE
anvi-export-misc-data [I]
PROFILE
STATIC SUMMARY
COLLECTION
anvi-gen-structure-database [I]
anvi-display-structure [I]
anvi-gen-variability-profile [I]
anvi-get-codon-frequencies [I]
LAYERS DATA
anvi-import-misc-data [I]
anvi-interactive [I]
anvi-get-short-reads-from-bam [I]
anvi-get-short-reads-mapping-to-a-gene [I]
AA FREQUENCIES
ITEMS DATA
anvi-merge [I]
anvi-profile [I]
LAYER ORDERS DATA
PHYLOGENY
CODON FREQUENCIES

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

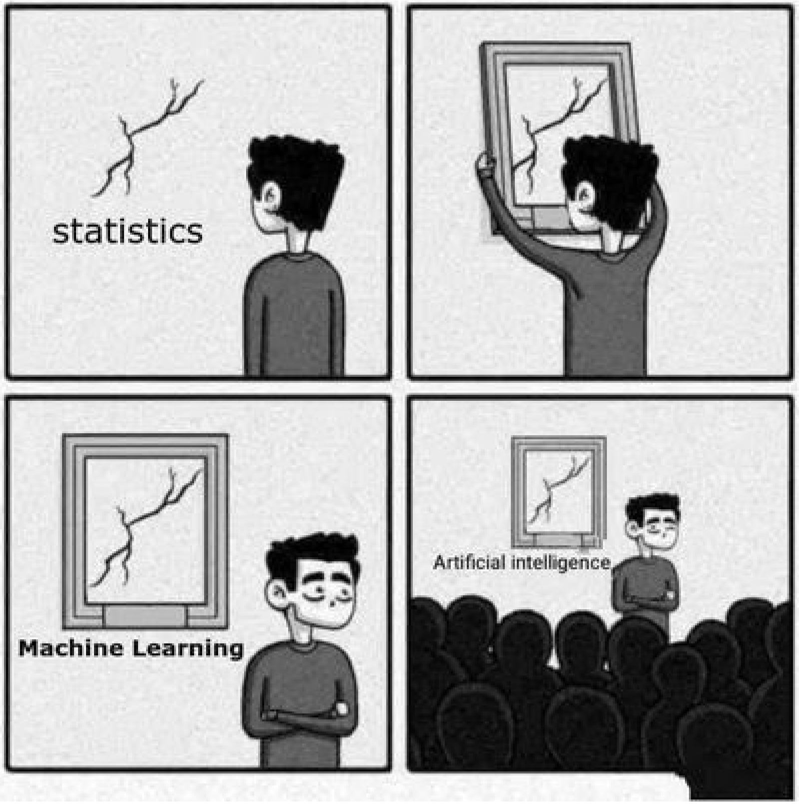
Summary & Conclusions

Statistical aspects

- supervised / unsupervised
- exploration / modeling
- mechanistic / non-mechanistic
- parametric / non-parametric
- deterministic / stochastic

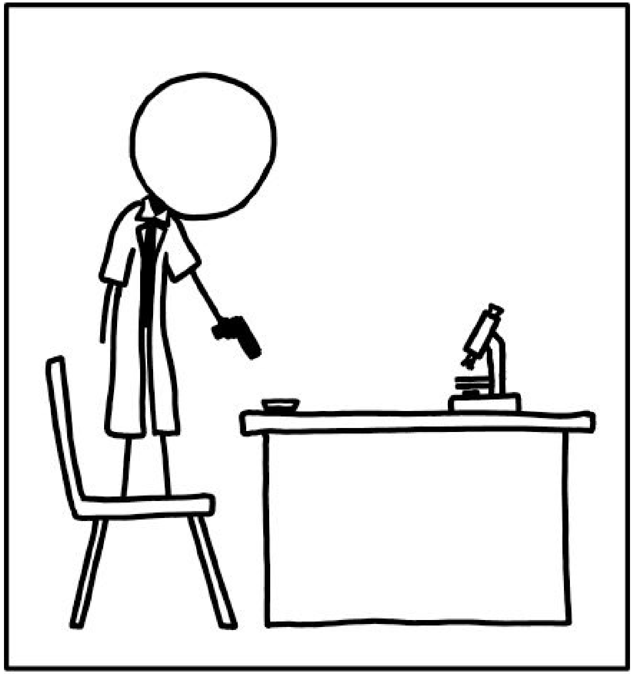
Learning goals

- Statistical thinking in microbiome studies
- The concept of open and reproducible research
- Familiarity with standard tools in amplicon profiling
- Looking at your own research problems in new ways
- Networking & collaboration!

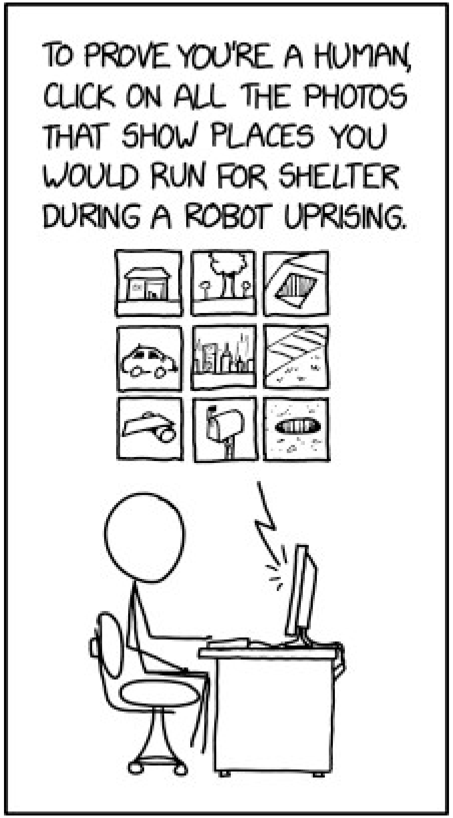

statistics
Machine Learning
Artificial intelligence

WHEN YOU SEE A CLAIM THAT A COMMON DRUG OR VITAMIN "KILLS CANCER CELLS IN A PETRI DISH,"

KEEP IN MIND:



SO DOES A HANDGUN.


TO PROVE YOU'RE A HUMAN, CLICK ON ALL THE PHOTOS THAT SHOW PLACES YOU WOULD RUN FOR SHELTER DURING A ROBOT UPRISING.