

How do we use data science in EEB?

Dr. Tomomi Parins-Fukuchi

tomo.fukuchi@utoronto.ca

Can we better understand the potential for life on other planets?

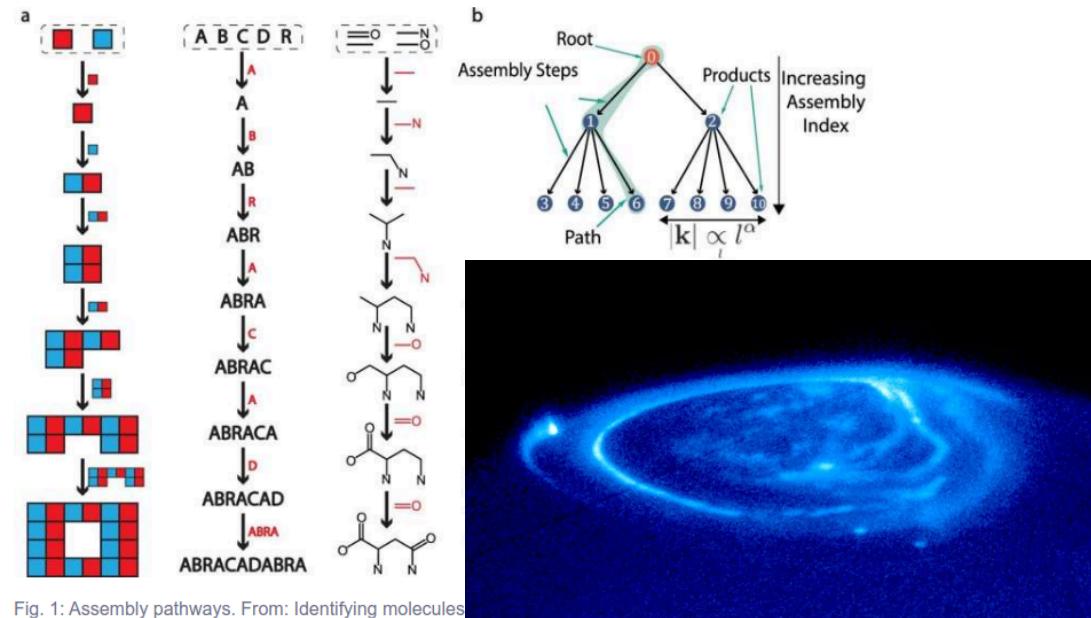
[Home](#) / [Astronomy & Space](#) / [Astrobiology](#)



J MAY 24, 2021

Complex molecules could hold the secret to identifying alien life

by University of Glasgow



"...the team used their method to assign MA numbers to a database containing about **2.5 million molecules**..."

Can we better understand emotional expressiveness in music?

Cornell University

We gratefully acknowledge support from the Simons Foundation and member institutions.

arXiv.org > physics > arXiv:2103.16737

Search... All fields Search
Help | Advanced Search

Physics > Popular Physics

[Submitted on 31 Mar 2021]

I Knew You Were Trouble: Emotional Trends in the Repertoire of Taylor Swift

Megan Mansfield, Darryl Seigman

As a modern musician and cultural icon, Taylor Swift has earned worldwide acclaim via pieces which predominantly draw upon the complex dynamics of personal and interpersonal experiences. Here we show, for the first time, how Swift's lyrical and melodic structure have evolved in their representation of emotion over the volume of the relevant discography, and that uniquely identifying a song that optimally describes a highly specific mood. To do this, we separate the criteria into the level of optimism (H) and the strength of commitment to a relationship (R). We find an overall trend toward positive emotions in stronger relationships, with the entire repertoire. We find an overall trend toward positive emotions in stronger relationships, with happiness (H) within individual albums over time. The mean relationship score (R) shows trends with blue eyes and/or bad reputations may lead to overall less positive emotions, while those with green eyes and/or good reputations may lead to overall more positive emotions. It is important to note that these trends are based on small sample sizes, and more data are necessary to validate them. For example, the song "I Knew You Were Trouble" is highly representative of the overall emotional range of Taylor Swift's music.

Comments: 11 pages, 8 figures. Submitted to Acta Prima Aprila. taylorswift code available at [this http URL](#)
Subjects: Popular Physics (physics.pop-ph); Earth and Planetary Astrophysics (astro-ph.EP)
Cite as: arXiv:2103.16737 [physics.pop-ph]
(or arXiv:2103.16737v1 [physics.pop-ph] for this version)

Submission history

From: Megan Mansfield [[view email](#)]

[v1] Wed, 31 Mar 2021 00:21:15 UTC (286 KB)

Download:

- [PDF](#)
- [Other formats](#)

 BY-NC-ND

Current browse context:
[physics.pop-ph](#)

< prev | next >

The Chicago Maroon

GREY CITY / July 24, 2021 / 2:27 p.m.

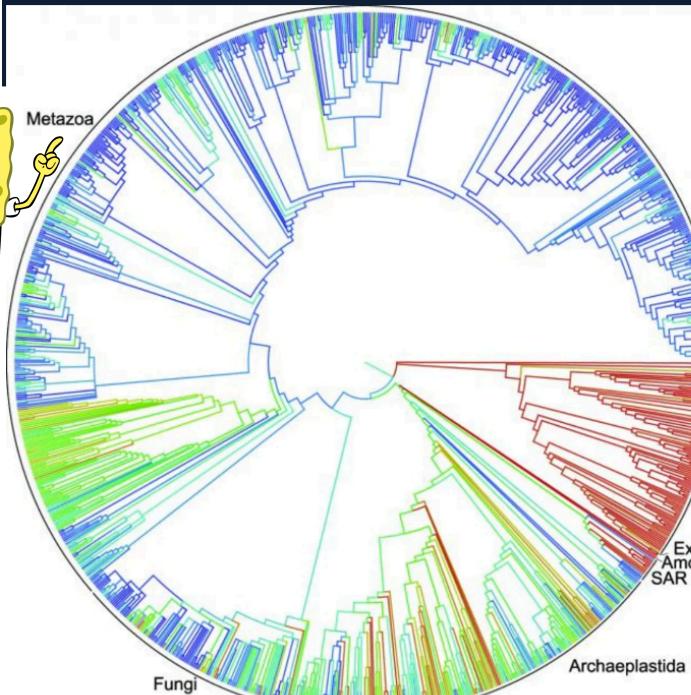
"Sad, Beautiful, Tragic": UChicago Researchers Analyze the Emotional Range of Taylor Swift's Music



Can we better understand the shared history across all life?

Online 'Open Tree of Life' Traces Origins of 2.3 Million Species

The combined efforts of thousands of scientists worldwide have produced the most complete yet "tree of life," available online for free.



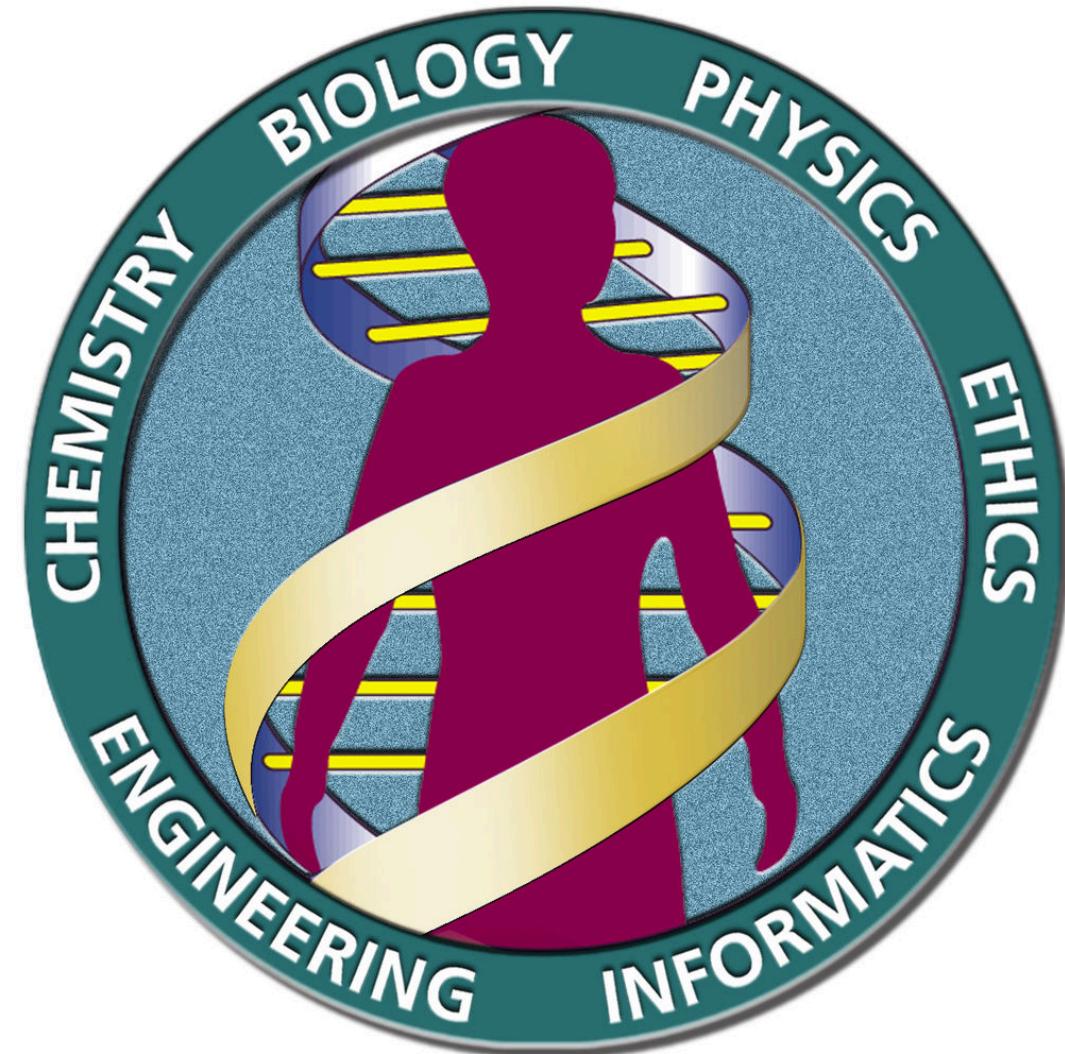
These are all **data science** questions

These are all **data science** questions

Data science has *revolutionized* biology

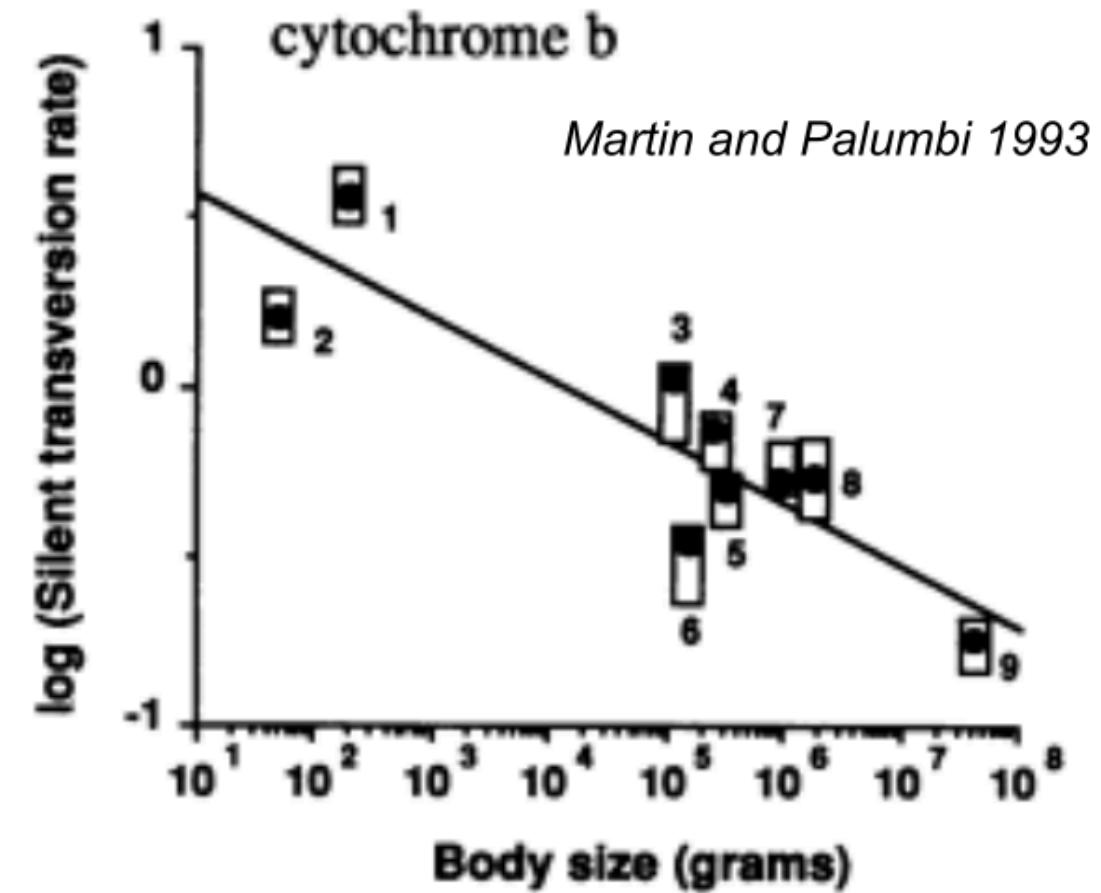
The Human Genome Project (1990-2003)

- Reshaped everything:
 - biology
 - medicine
 - psychology
 - computation
 - statistics



Scientific datasets have become massive

The field used to focus mostly on targeted questions using small datasets



'Genetics' -> 'Genomics'

- 'Genes' are comprised of nucleotides
- **Genome:** All nucleotides possessed by an individual



Pre-2003:

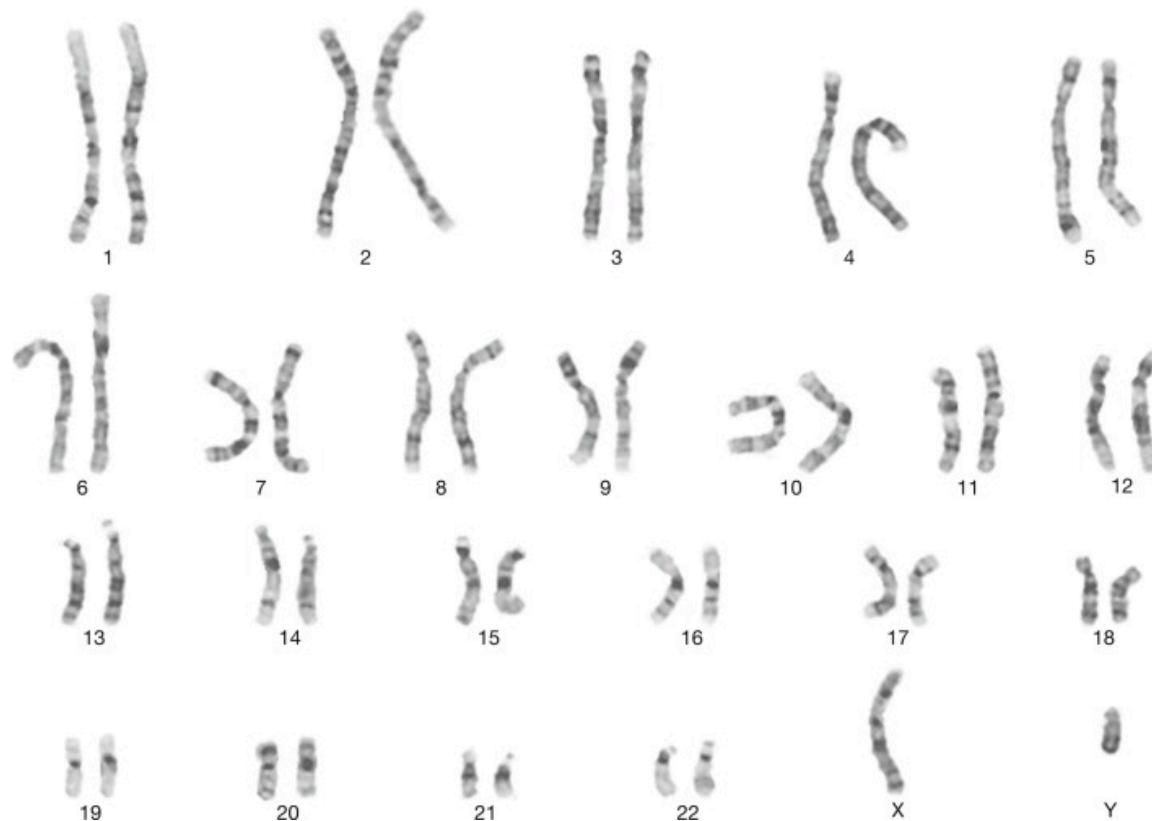
- Analyze ~3 thousand nucleotides at once



Imagine this as a single gene

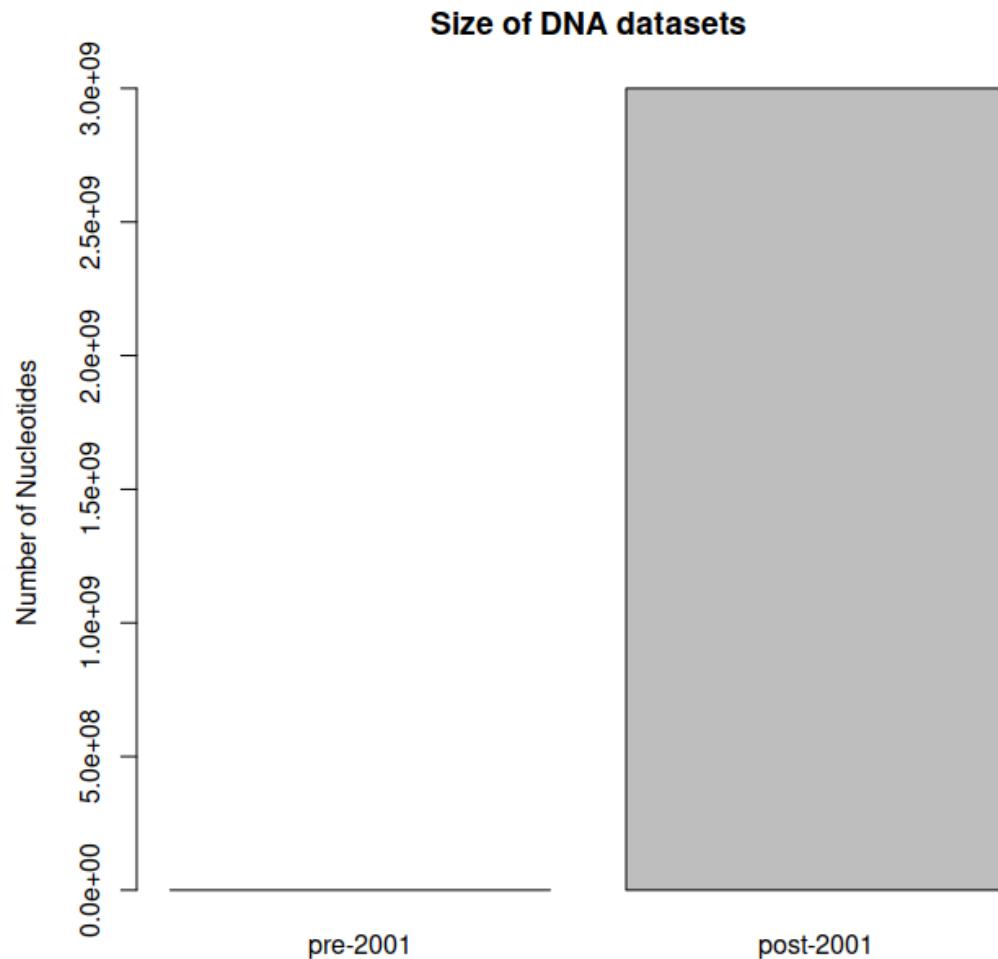
Post-2003:

- Analyze ~3 billion nucleotides at once



That is a **million times** bigger.

That is a million times bigger.



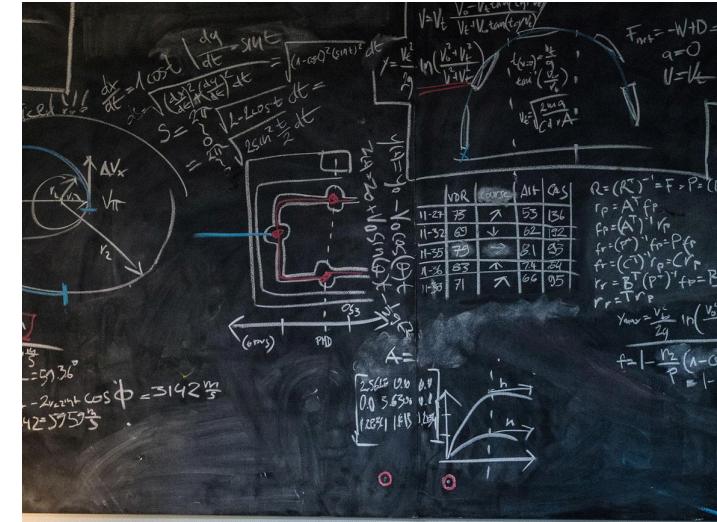
These discoveries depend on new approaches for dealing with data

These discoveries depend on new approaches for dealing with data

We often call this mix of new approaches "data science"

How do we practice data analysis?

- Develop new statistical approaches



Statistics/Math



Computation

How do we practice data analysis?

- Develop new statistical approaches
 - Write computer code to analyze data using stats



Statistics/Math



Computation

How do we practice data analysis?

- Develop new statistical approaches
- Write computer code to analyze data using stats
- Apply code to address scientific questions



Statistics/Math



Computation

Much of the science you encounter in your daily life is data science!

Mum's a Neanderthal, Dad's a Denisovan: First discovery of an ancient-human hybrid

Genetic analysis uncovers a direct descendant of two different groups of early humans.

[Matthew Warren](#)



Data science breakthroughs

We have learned a lot about human biology using data science



Denny inherited one set of chromosomes from her Neanderthal ancestors, depicted in this model. Credit: Christopher Rynn/University of Dundee

A female who died around 90,000 years ago was half Neanderthal and half Denisovan, according to genome analysis of a bone discovered in a Siberian cave. This is the first time scientists have identified an ancient individual whose parents belonged to distinct human groups. The findings were published on 22 August in *Nature*¹.

Data science breakthroughs

We have learned a lot about human biology using data science

Our Neanderthal genes linked to risk of depression and addiction



LIFE 11 February 2016

By Colin Barres



Having sex with Neanderthals meant some of us still carry their DNA, and with it, a higher risk of depression and nicotine addiction
Nikola Solic/Reuters

Dealing with data is hard

Neanderthal genes are *risk-inducing* for severe COVID?

Article | [Published: 30 September 2020](#)

The major genetic risk factor for severe COVID-19 is inherited from Neanderthals

[Hugo Zeberg](#)  & [Svante Pääbo](#) 

[Nature](#) 587, 610–612 (2020) | [Cite this article](#)

719k Accesses | 141 Citations | 4994 Altmetric | [Metrics](#)

Dealing with data is hard

Neanderthal genes are *protective* against severe COVID?

RESEARCH ARTICLE



A genomic region associated with protection against severe COVID-19 is inherited from Neandertals

Hugo Zeberg and Svante Pääbo

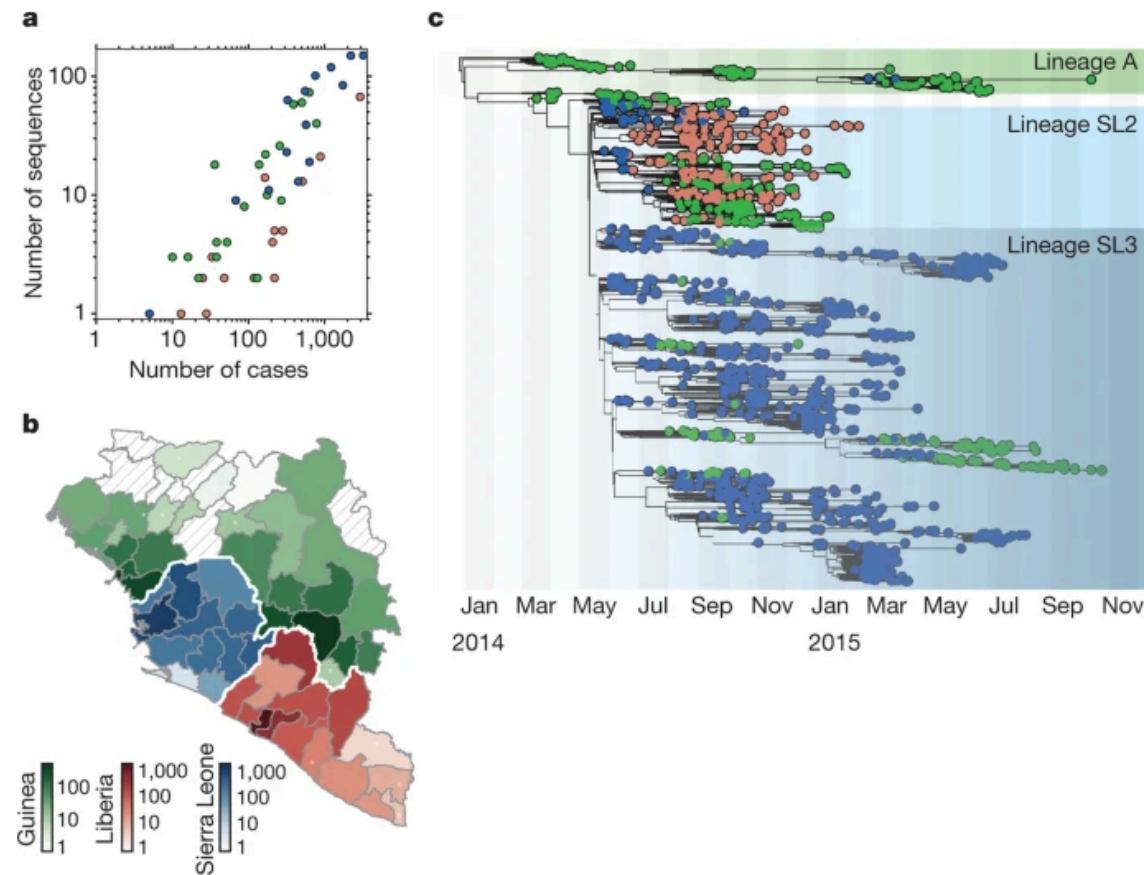
[+ See all authors and affiliations](#)

PNAS March 2, 2021 118 (9) e2026309118; <https://doi.org/10.1073/pnas.2026309118>

Contributed by Svante Pääbo, January 22, 2021 (sent for review December 21, 2020; reviewed by Tobias L. Lenz and Lluís Quintana-Murci)

Data science helps us understand human health

Figure 1: Evolution of EBOV during the 2013–2016 outbreak showing the extent and location of virus sampling.

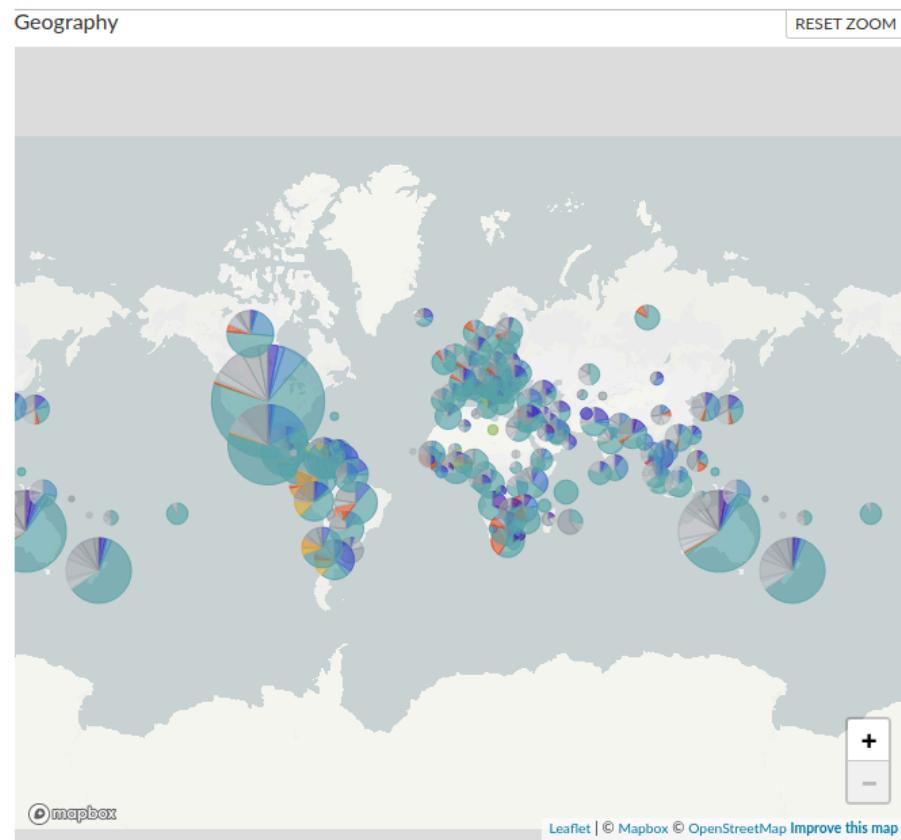
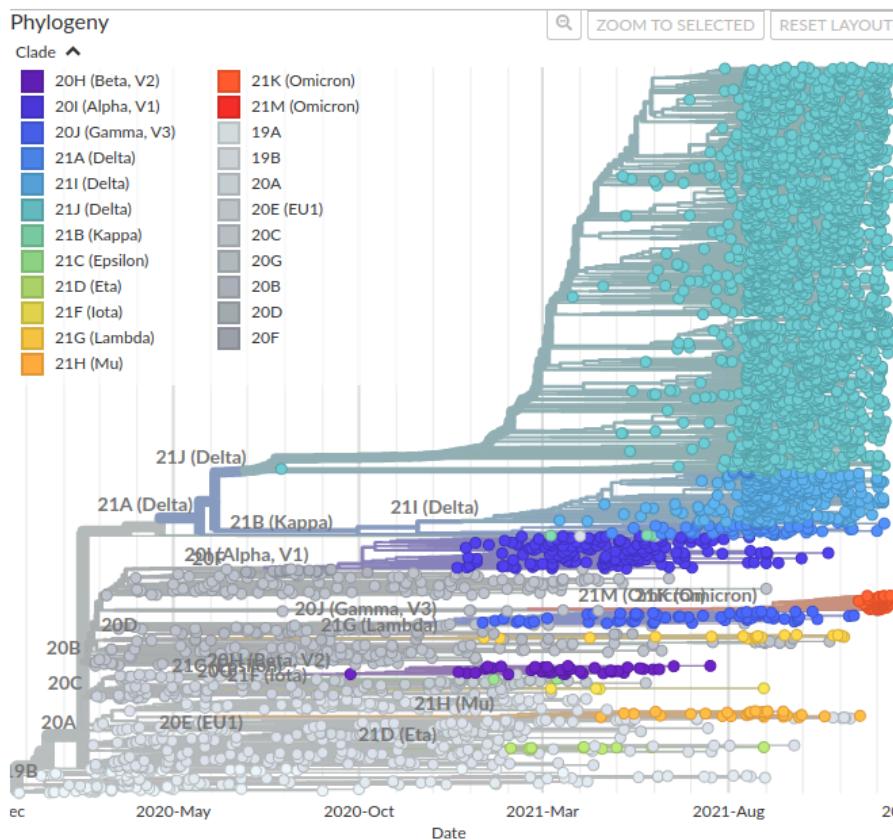


Data science helps us understand human health

Genomic epidemiology of novel coronavirus - Global subsampling

Built with [nextstrain/ncov](#). Maintained by the Nextstrain team. Enabled by data from [GISAID](#).

Showing 3589 of 3589 genomes sampled between Dec 2019 and Dec 2021.



Data science tools are also marketed to the public



ANCESTRY BREAKDOWN

Dig deeper into your ancestry.

It's the most complete genetic breakdown on the market, and the most comprehensive portrait of you yet.

- **Ancestry Composition**

Discover where in the world your DNA is from across 2000+ regions — in some cases, down to the county level.

- [Ancestry Detail Report](#)

[See all regions](#)



How does data science fit into the life sciences?

- The influx of new data has changed the landscape of research

How does data science fit into the life sciences?

- The influx of new data has changed the landscape of research
- Understanding these tools is *essential* to understanding modern science

Who am I?

- Evolutionary paleobiologist

Computational paleobiology

Paleobiologist

- a scientist who studies fossils to understand the biology of past organisms
- basically synonymous with paleontologist

Computational paleobiology

What do you think of when you imagine a paleontologist?

computational paleobiology

what do you think of when you imagine a paleontologist?

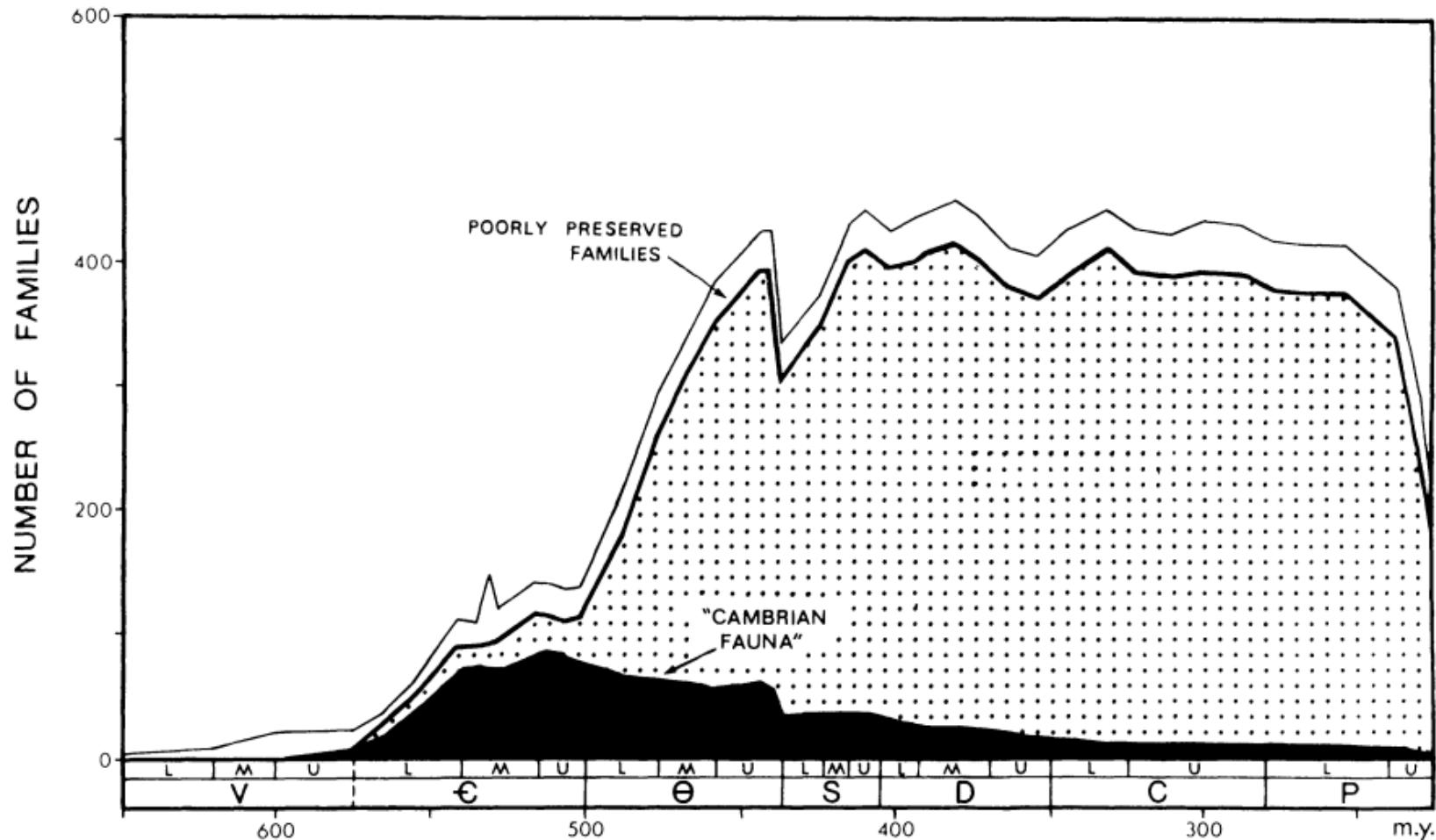


Computational paleobiology

- Paleontology has transformed into field where large datasets are used to ask general patterns about the history of life

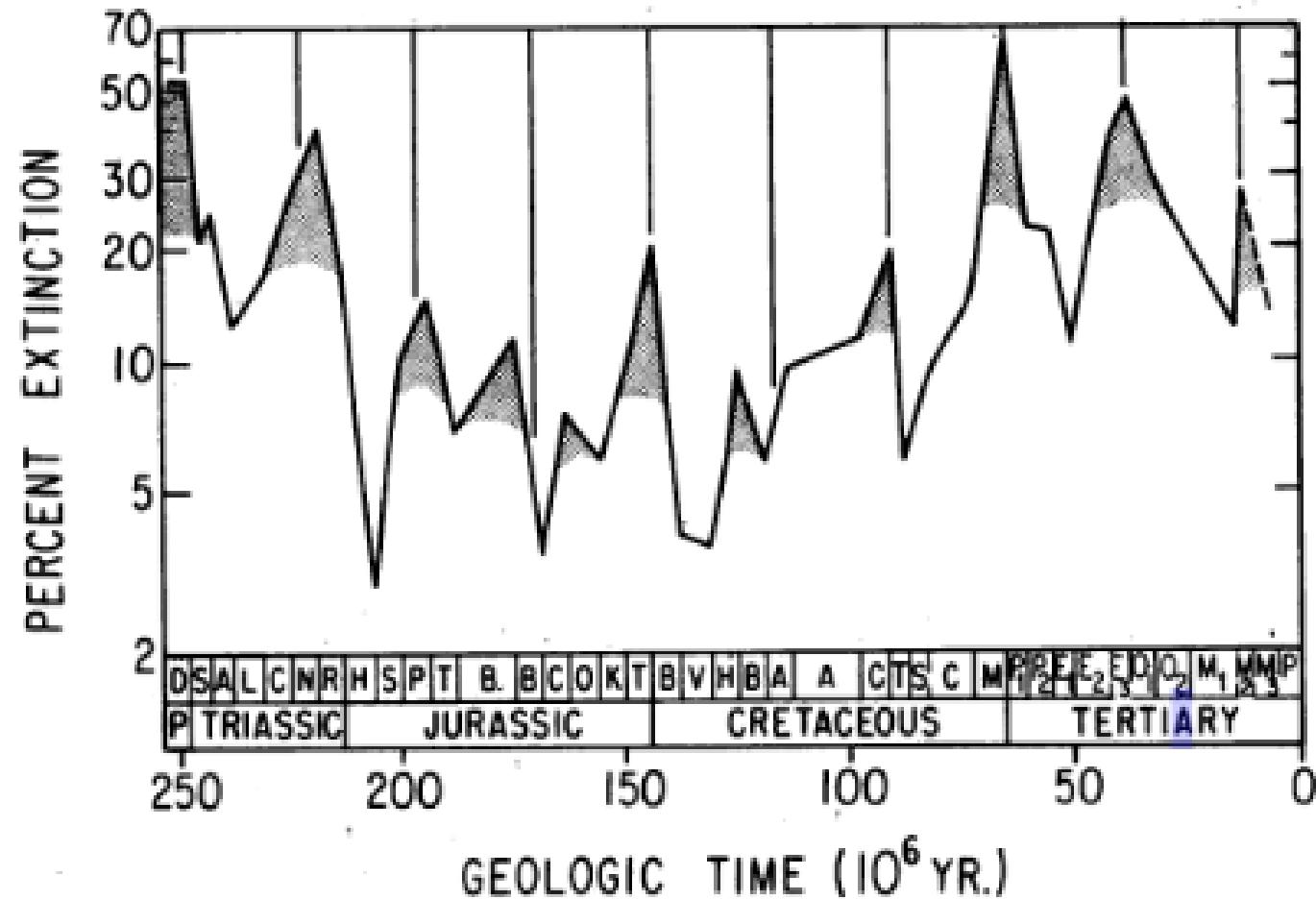
Computational paleobiology

- Large-scale analyses of thousands of fossils representing thousands of species



Computational paleobiology

- Large-scale analyses of thousands of fossils representing thousands of species



Computational paleobiology

The Paleobiology Database

Revealing the history of life

Learn ▾ User Guide Data ▾ Join & Support ▾

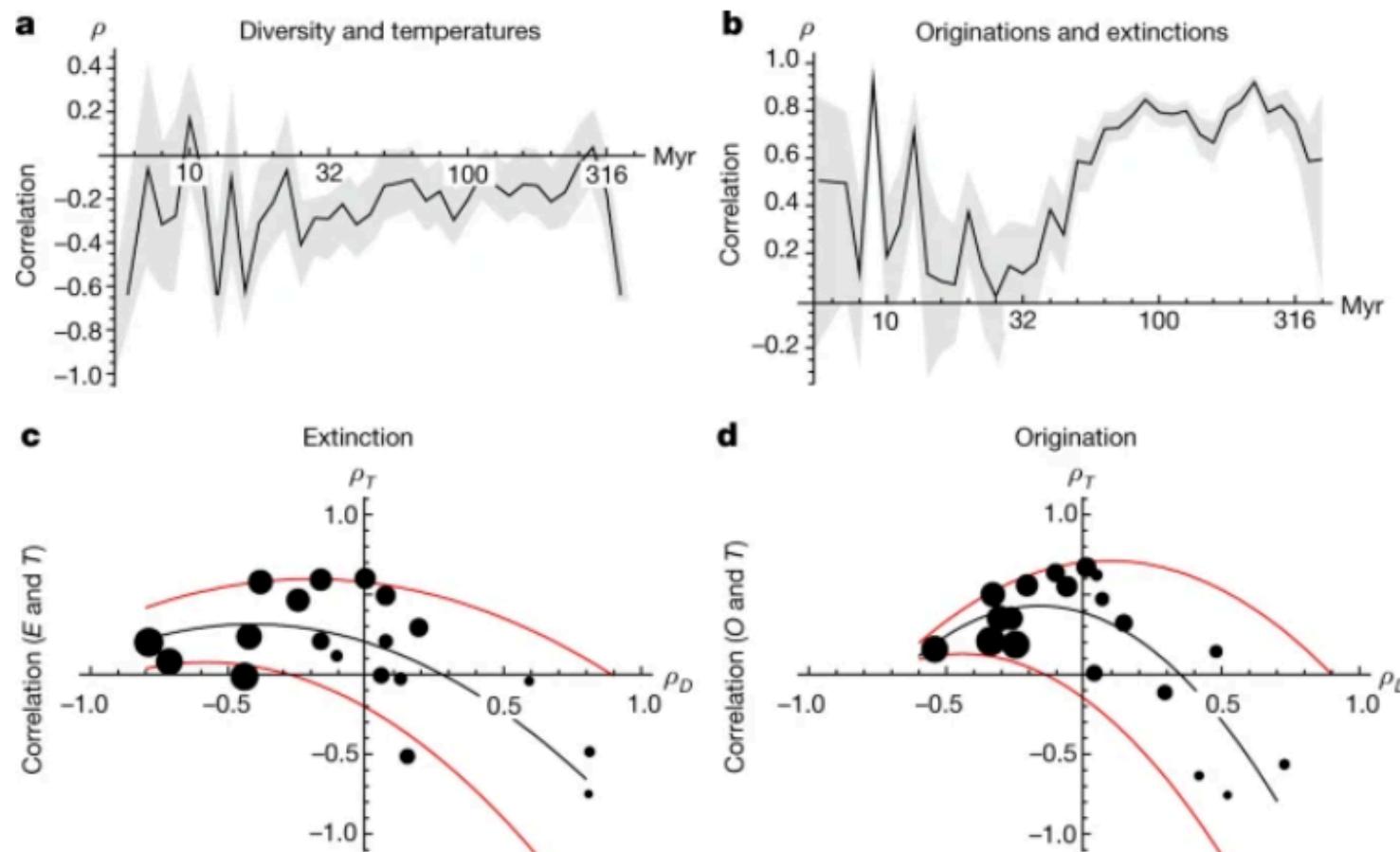
View recent changes

90,182 references	501,381 taxa	967,883 opinions	237,839 collections	1,636,514 occurrences	1,014 contributors
----------------------	-----------------	---------------------	------------------------	--------------------------	-----------------------

Computational paleobiology

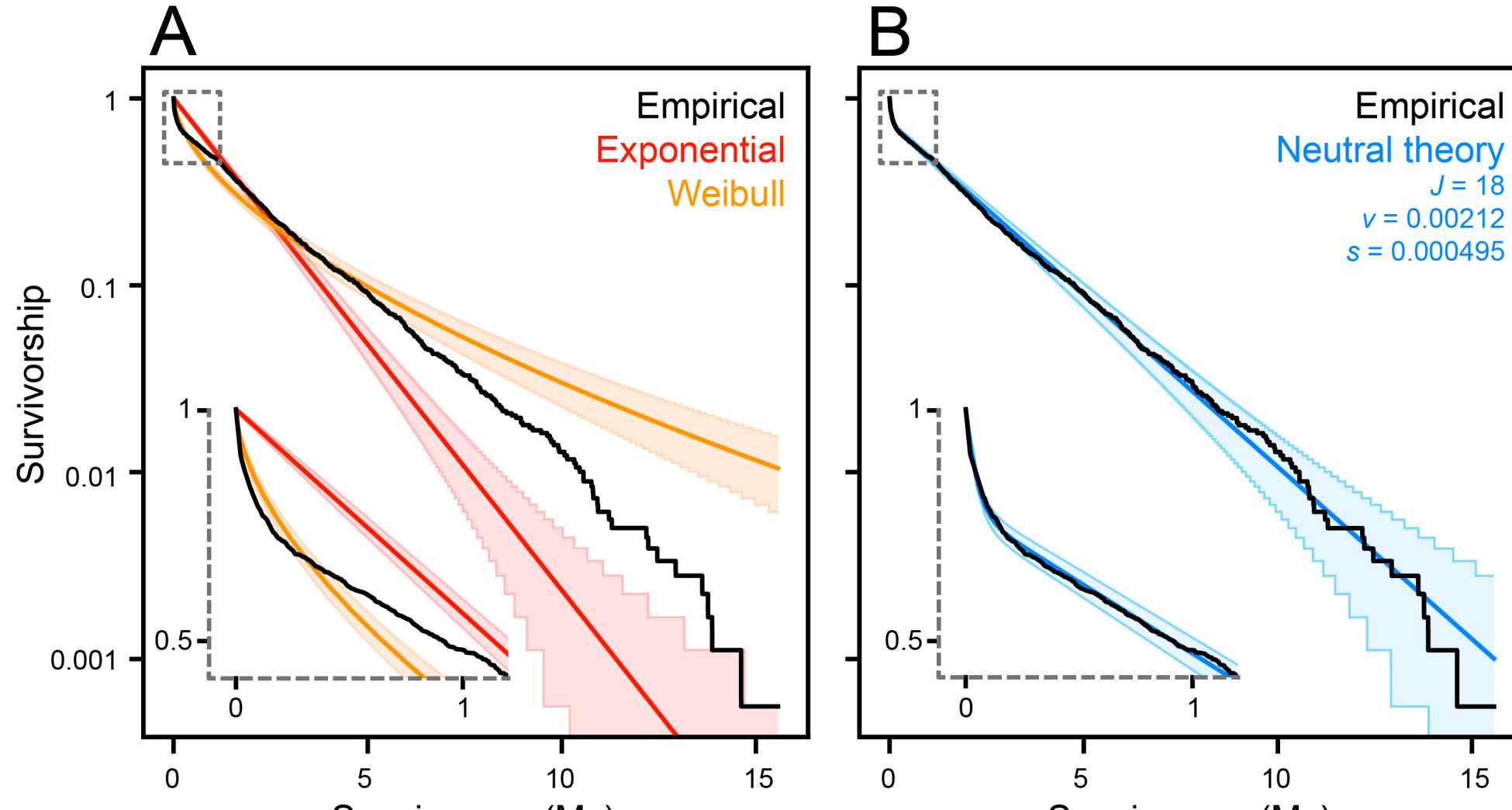
- Large-scale analyses of thousands of fossils representing thousands of species

Fig. 2: Scale-dependant correlations of macroevolutionary and palaeoclimate variables.



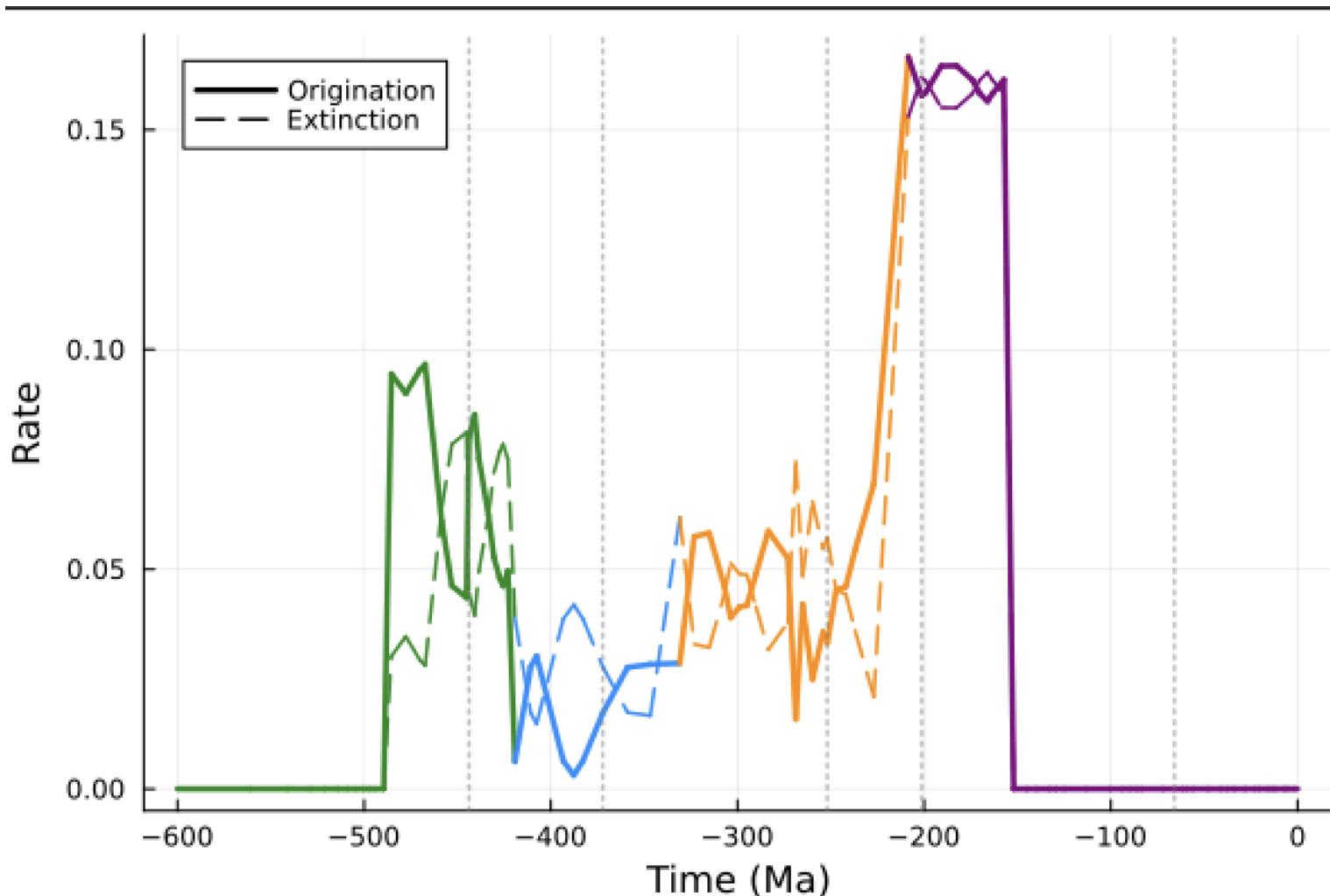
Computational paleobiology

- Large-scale analyses of thousands of fossils representing thousands of species



Computational paleobiology

- Large-scale analyses of thousands of fossils representing thousands of species



Who am I?

- Evolutionary paleobiologist

Who am I?

- Evolutionary paleobiologist
- I develop new computational approaches to ask 'big picture' questions about evolution

Who am I?

- Evolutionary paleobiologist
- I develop new computational approaches to ask 'big picture' questions about evolution
- For example...

A population of small, burrowing mammals gave rise to all of this ecological diversity

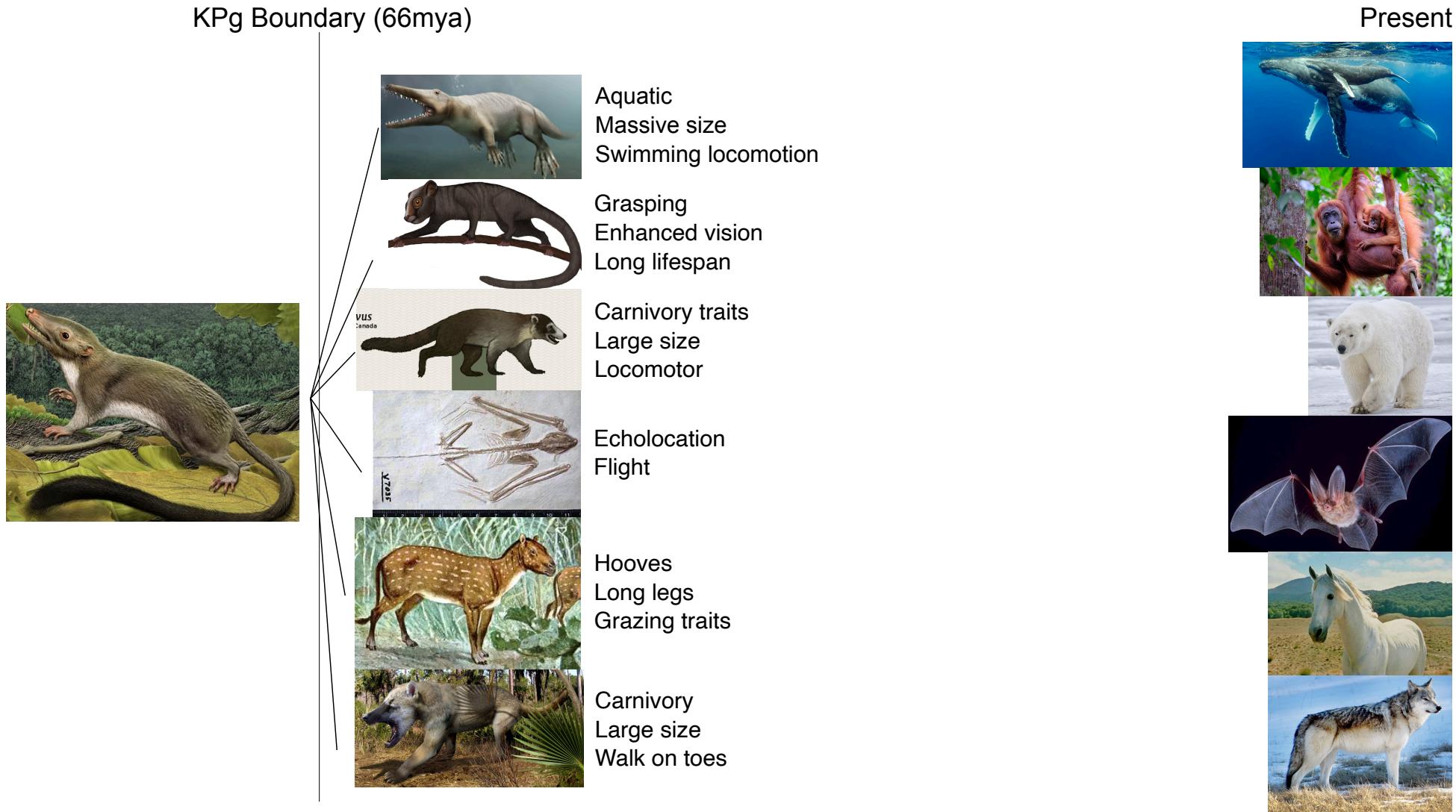
KPg Boundary (66mya)



Present



...in just a few million years



Most biologists are interested in the origin of cool features

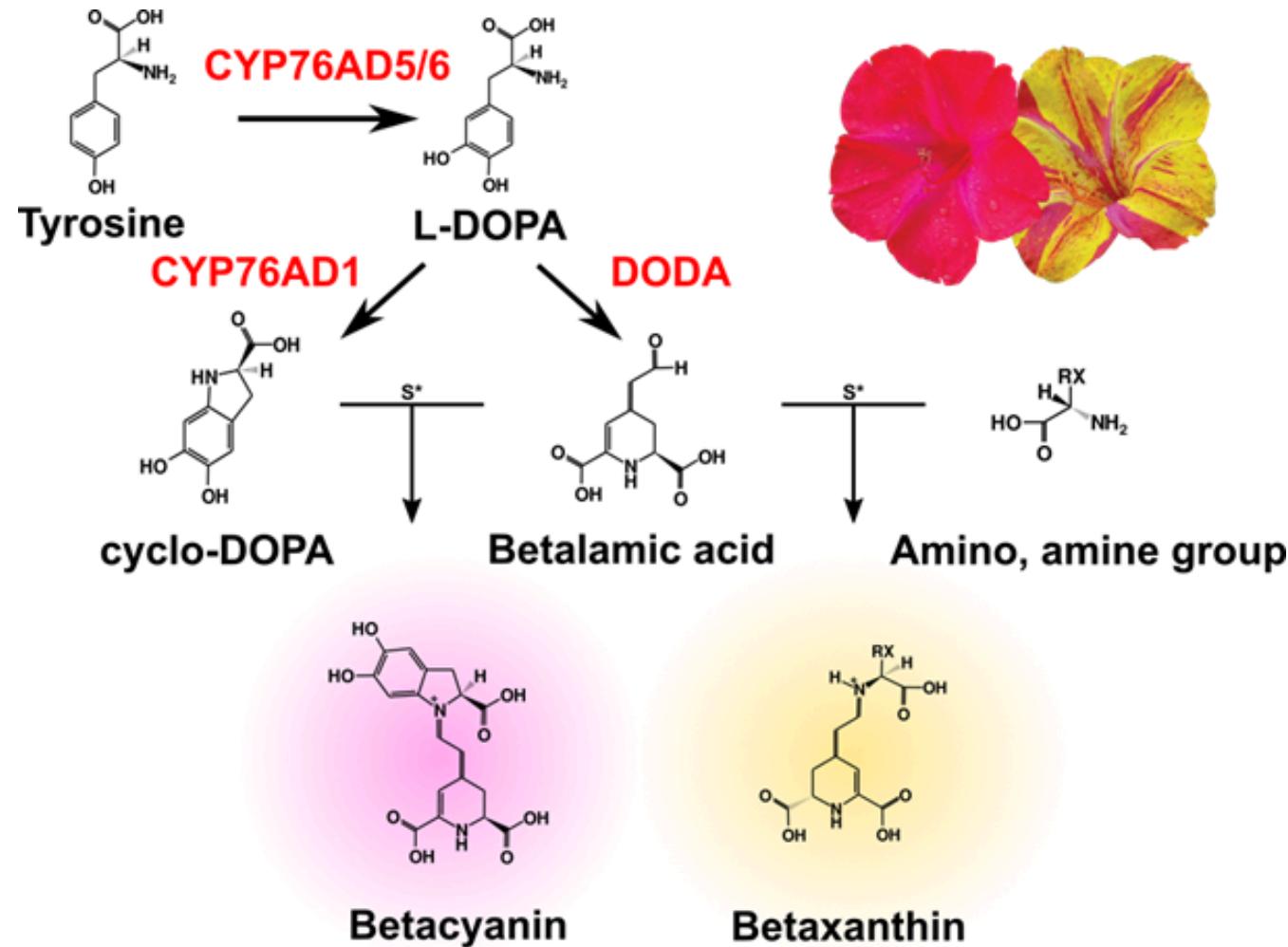
Most biologists are interested in the origin of cool features

- flowers



Most biologists are interested in the origin of cool features

- novel biochemical pathways (C4, betalain, etc)



Most biologists are interested in the origin of cool features

- skeletal features associated with novel locomotor modes



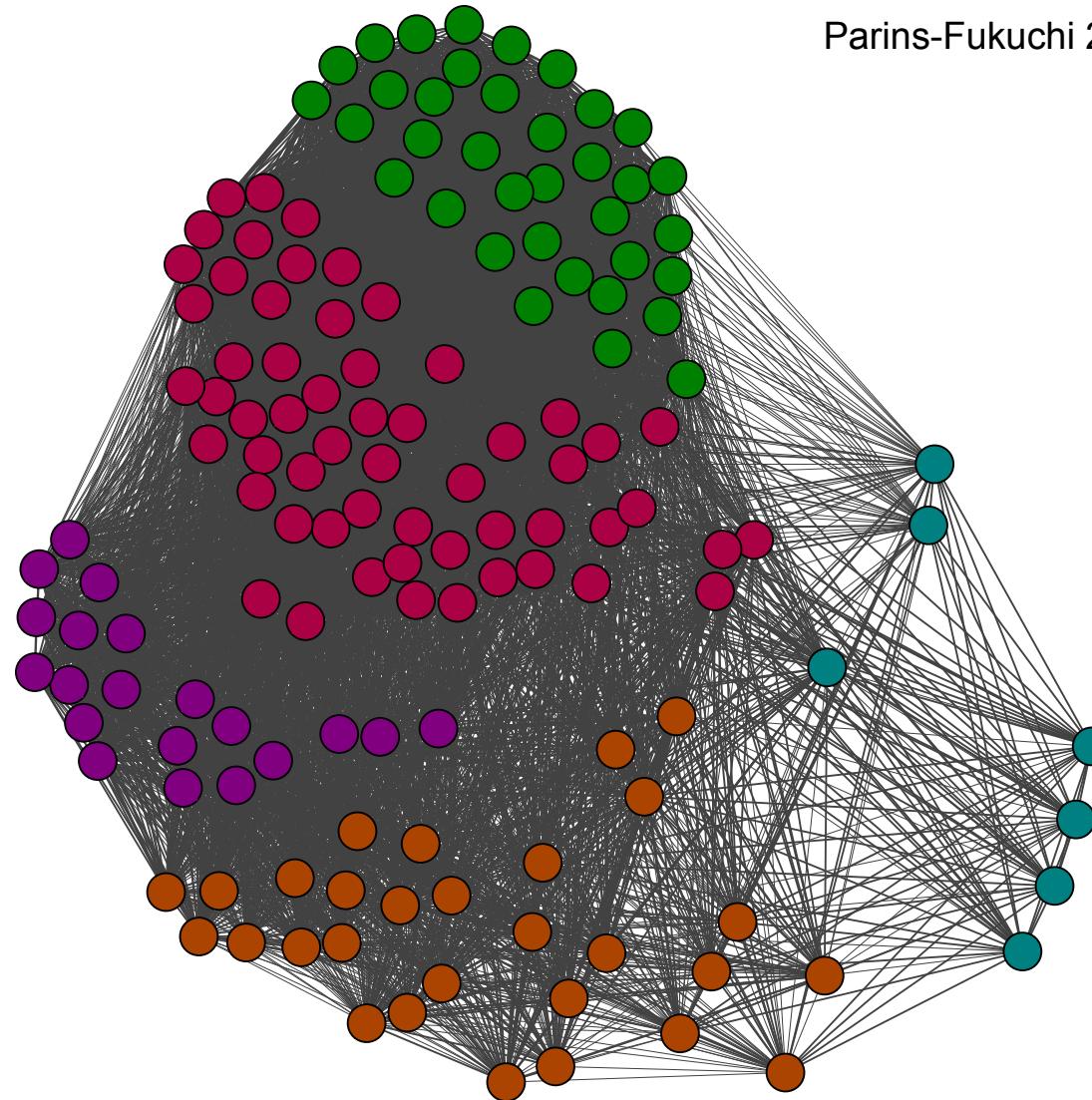
Apes

Ape skeletons display many locomotor innovations



Skeletal features display complex functional and genetic relationships

Parins-Fukuchi 2020, Evolution



fossil crinoids

- how does variation in groups of sea creatures affect their evolution?



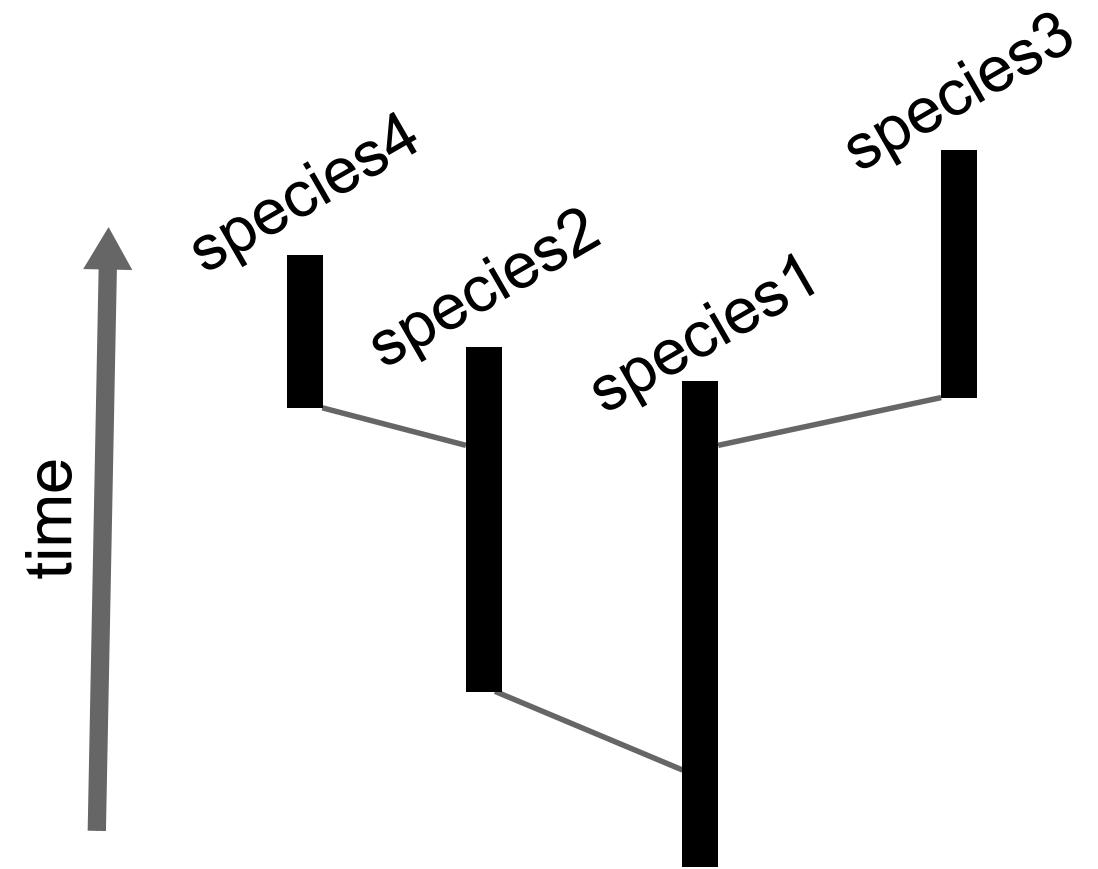
fossil sea urchins

- how does variation in groups of sea creatures affect their evolution?



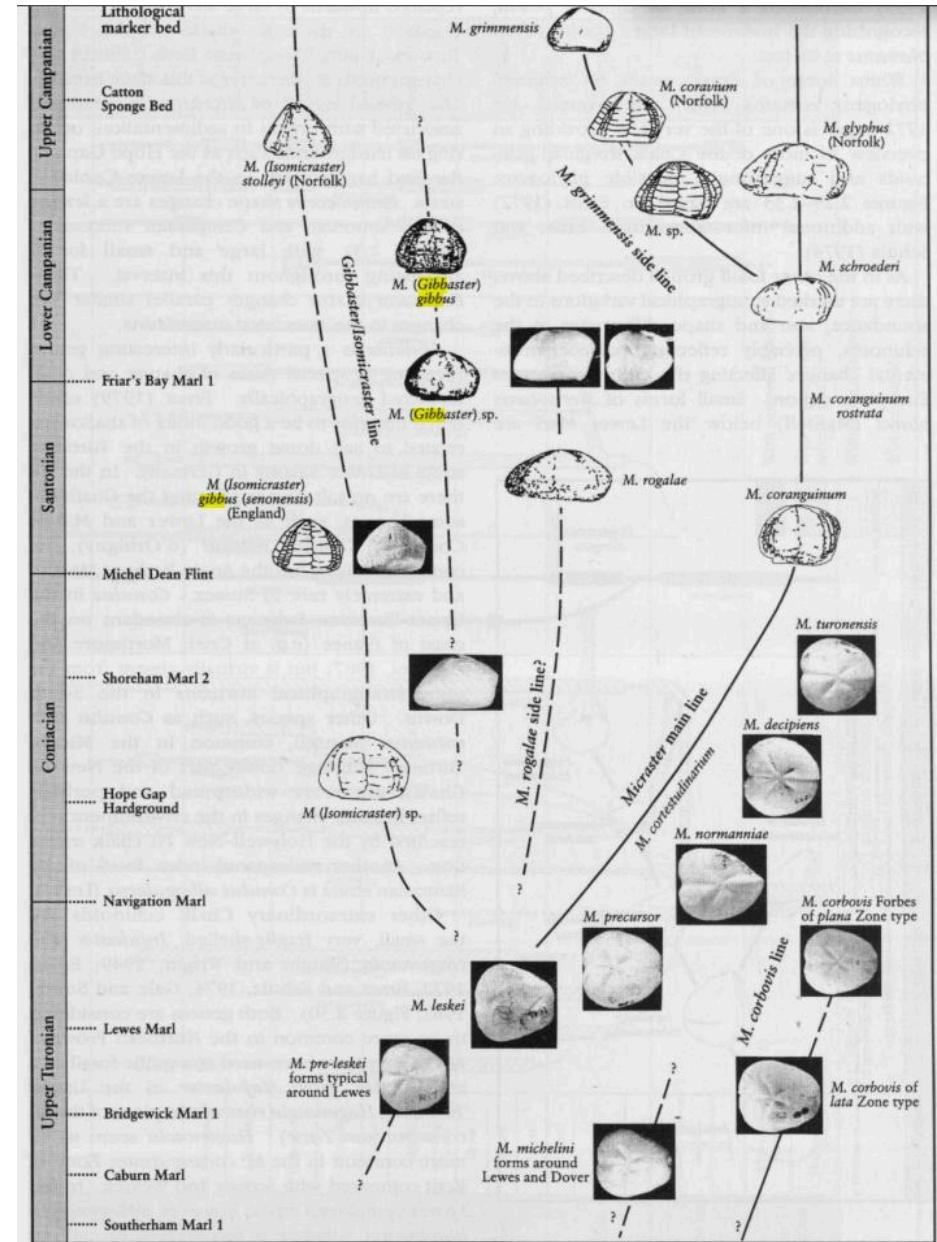
phylogenetic trees

- sequence of ancestor-descendent relationships
- when/how did species diverge from one another?



Micraster

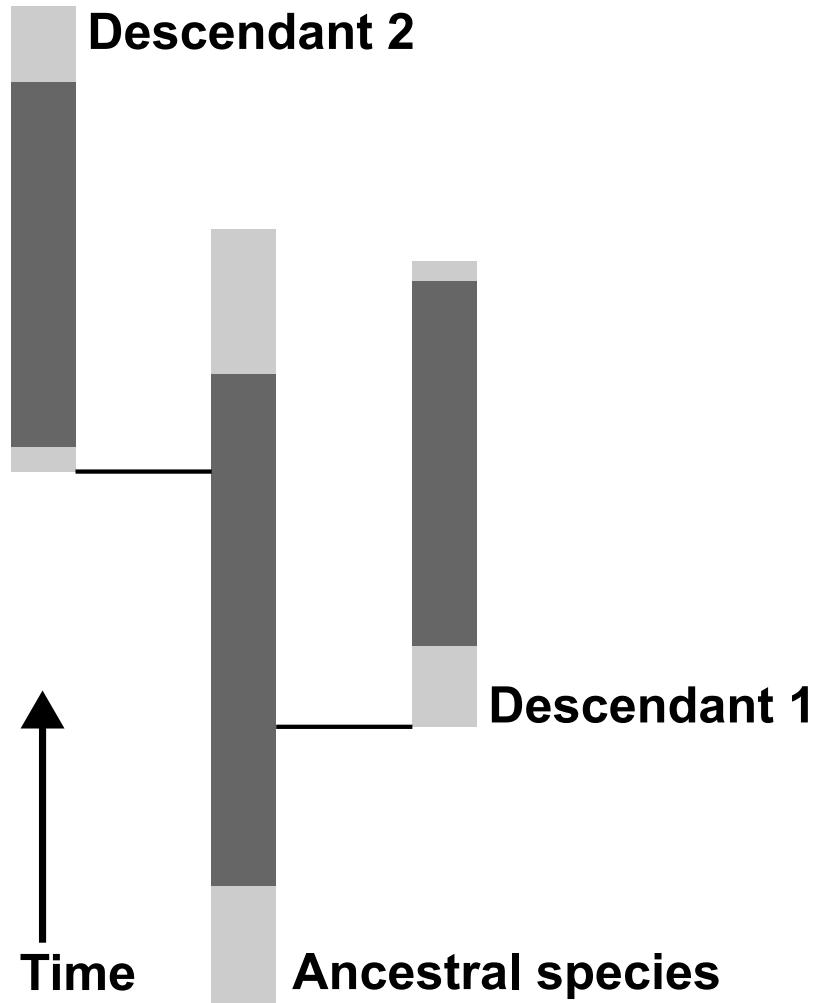
- evolutionary relationships used to be mainly hand-drawn



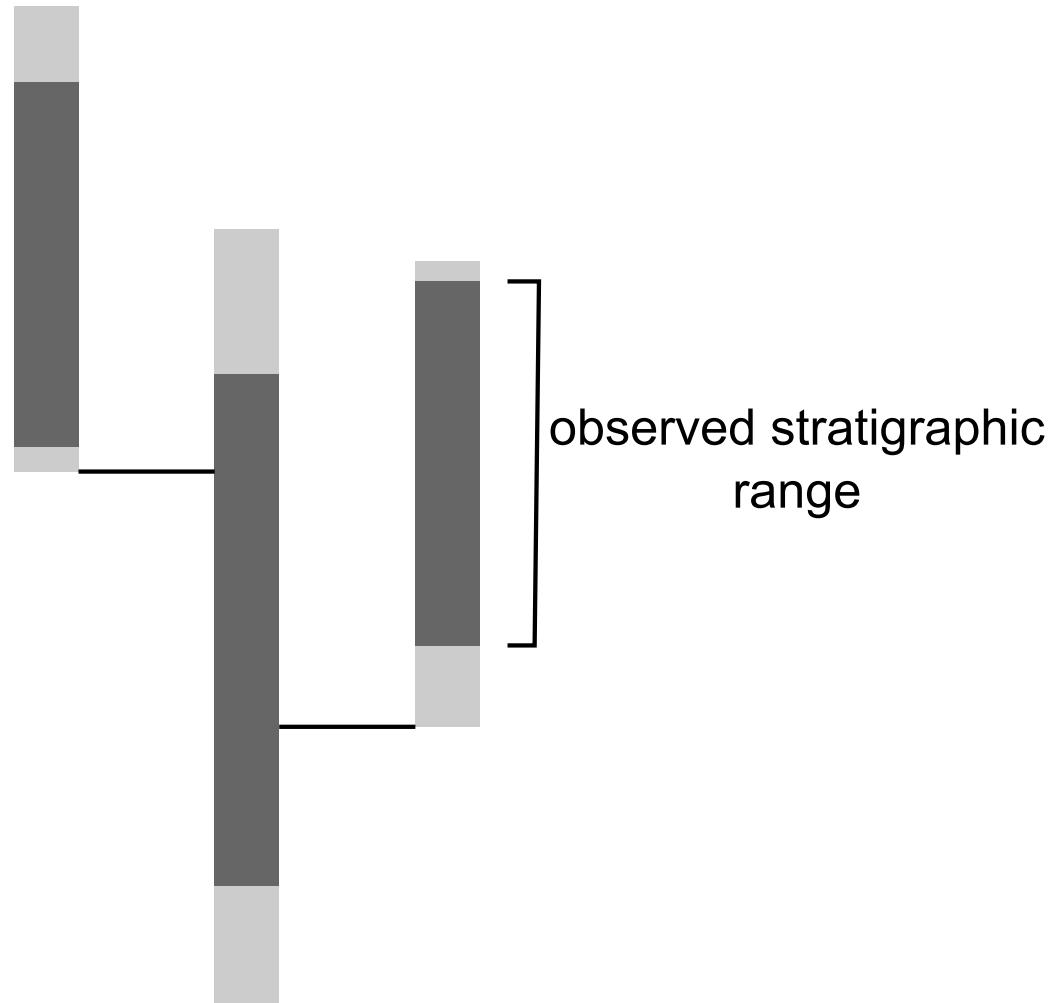
phylogenetic trees

- now we can write statistical programs to reconstruct relationships using statistical models based on probability

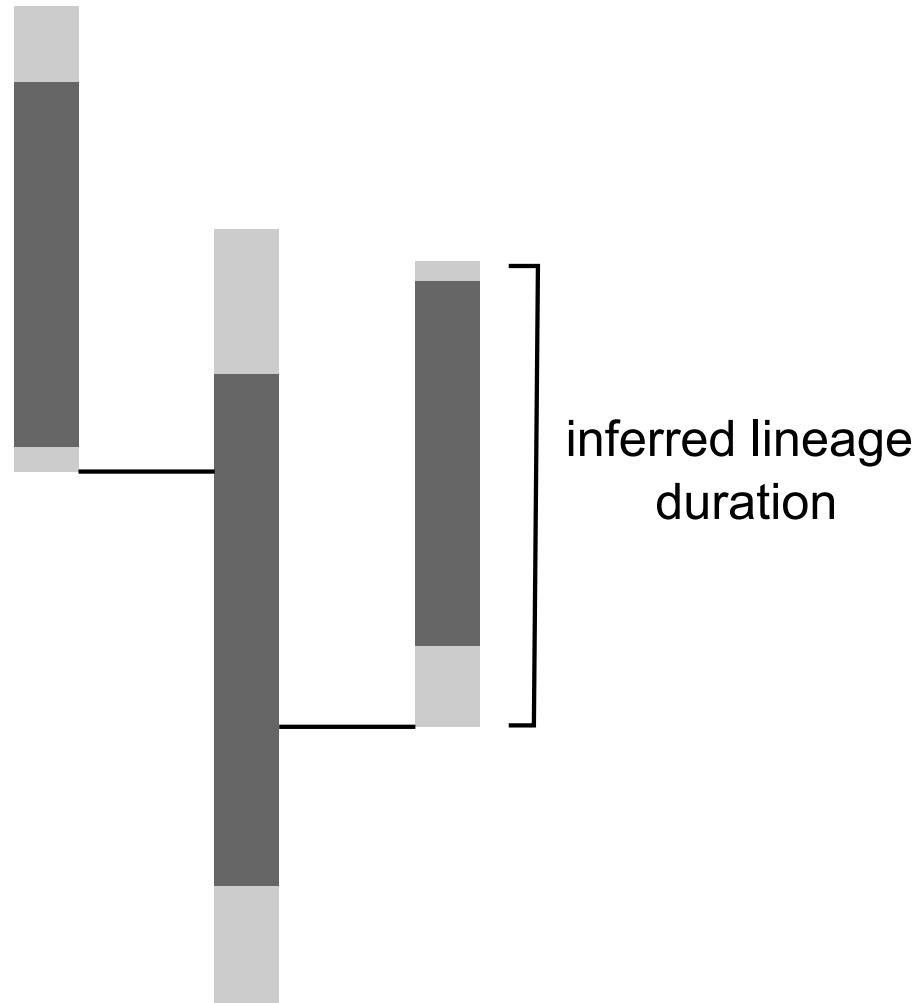
phylogenetic trees



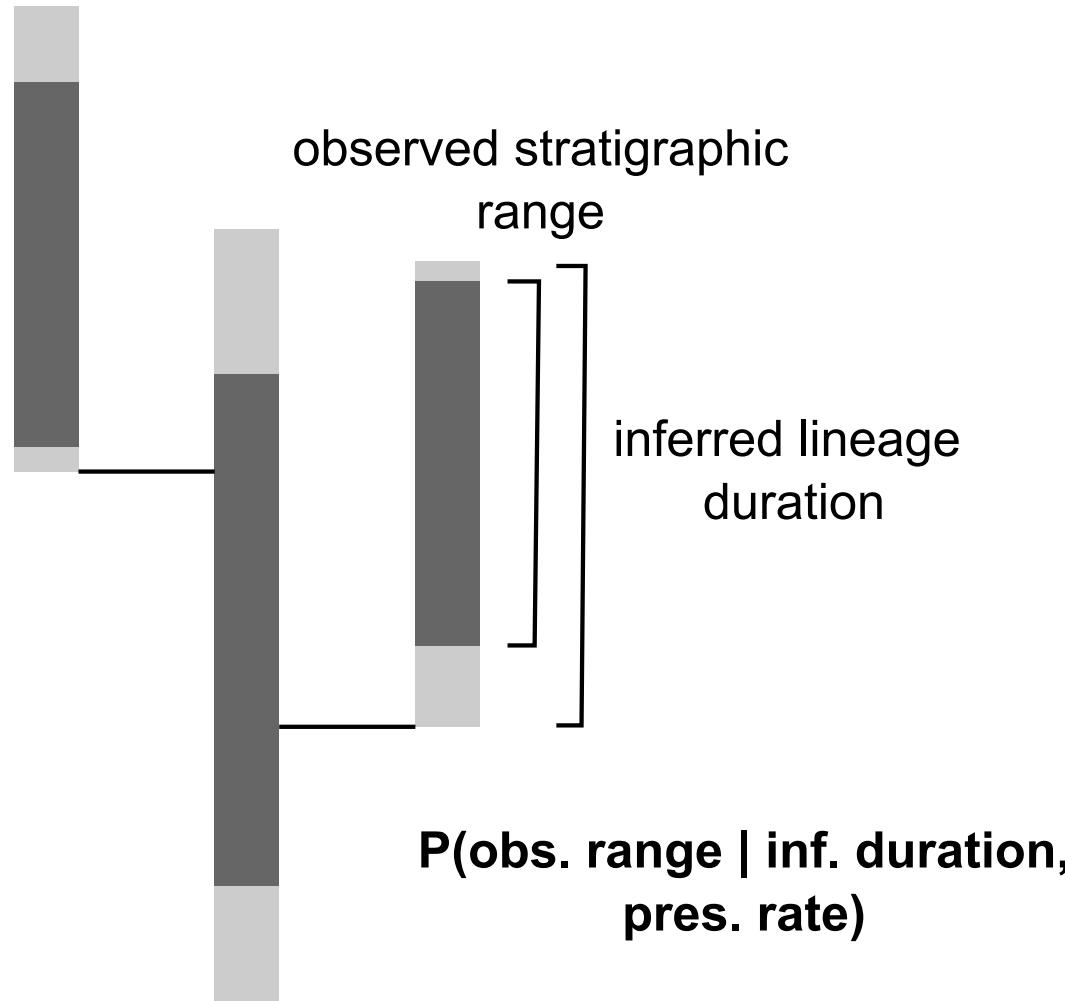
phylogenetic trees



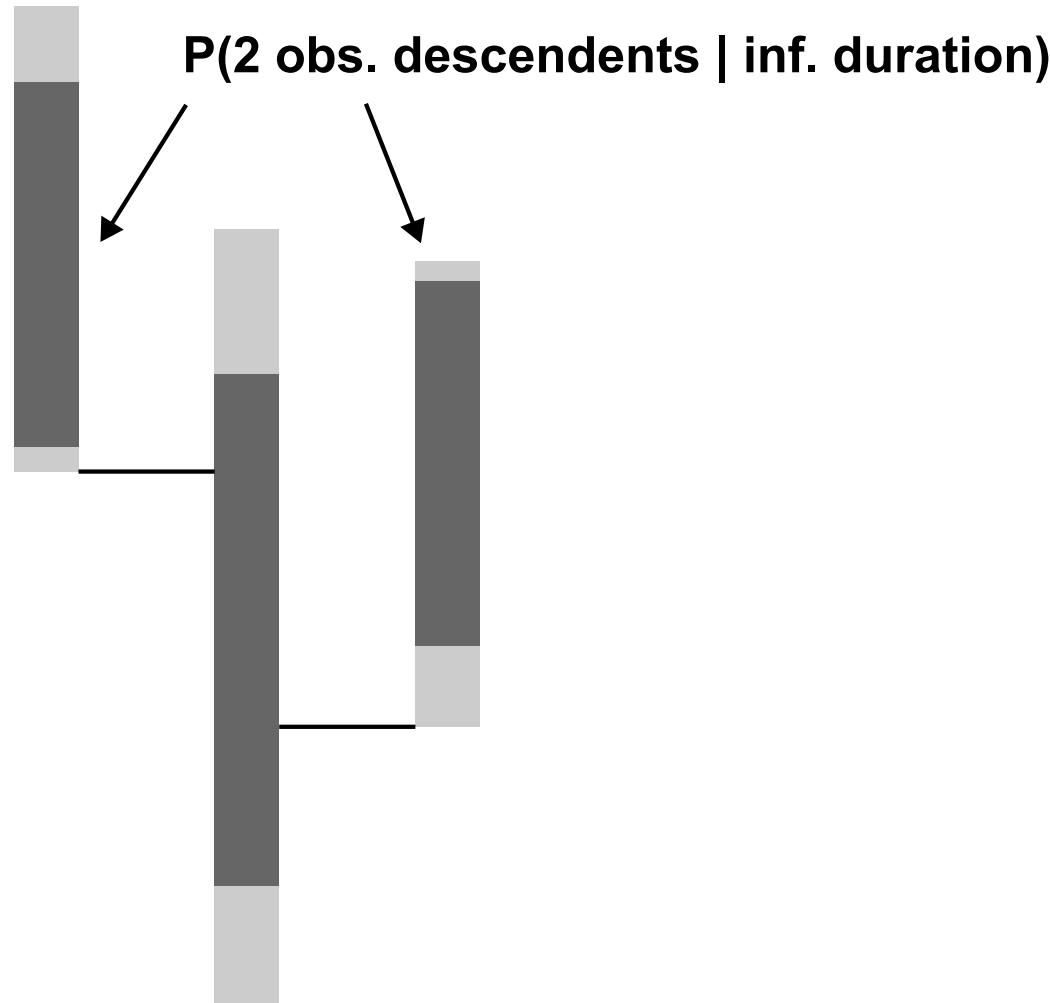
phylogenetic trees



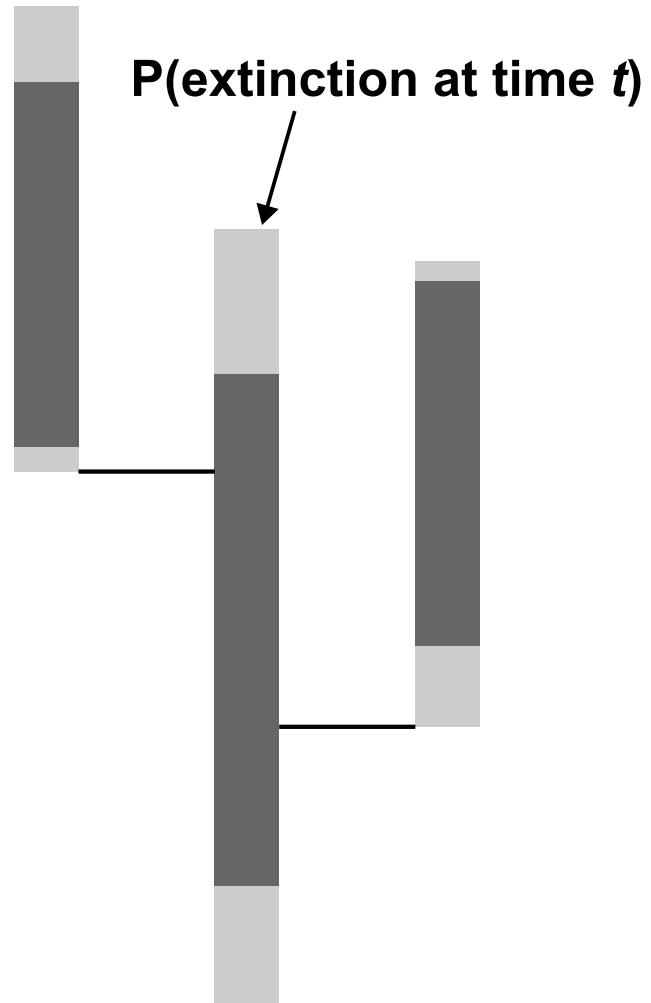
phylogenetic trees



phylogenetic trees



phylogenetic trees



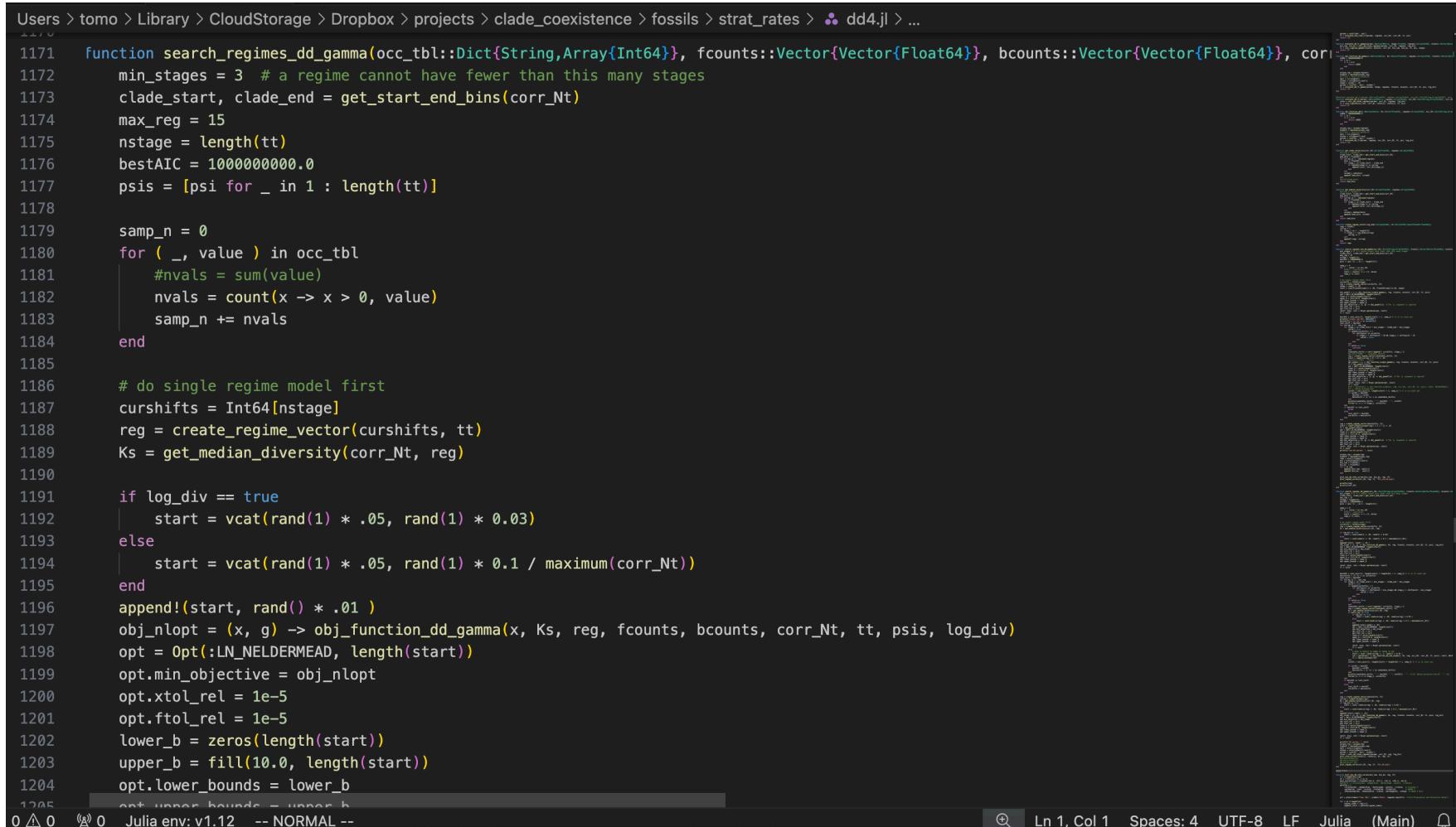
computational paleobiology

Instead of looking like this:



computational paleobiology

Things tend to look much more like:



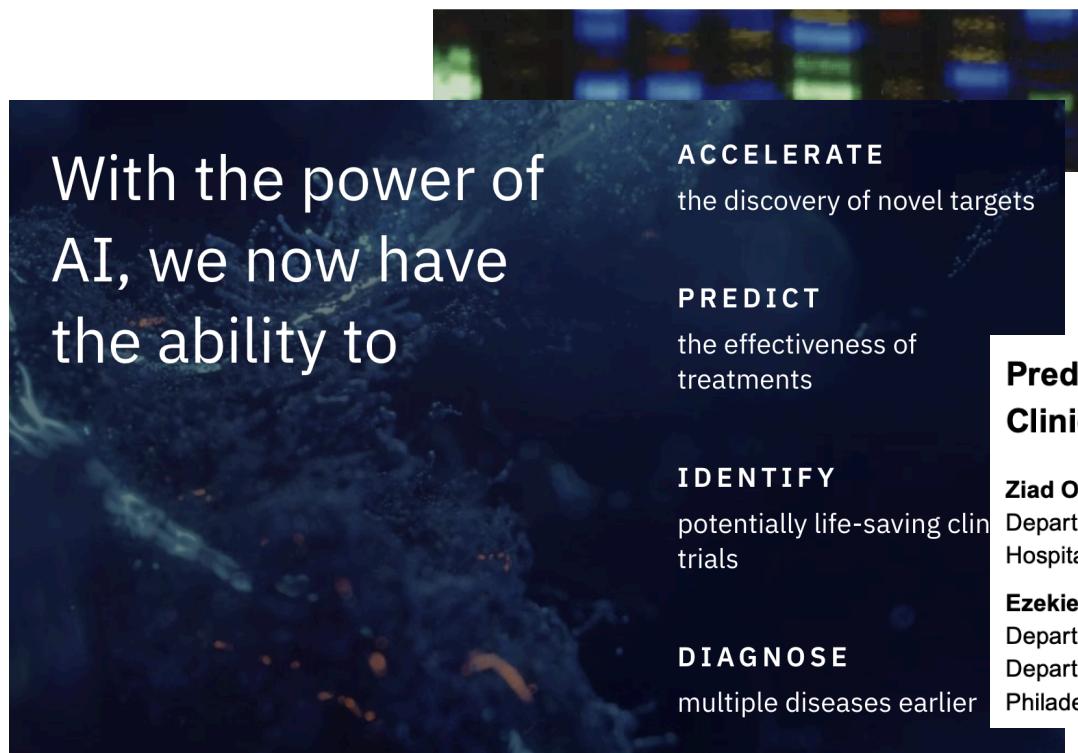
```
Users > tomo > Library > CloudStorage > Dropbox > projects > clade_coexistence > fossils > strat_rates > dd4.jl > ...  
1171 function search_regimes_dd_gamma(occ_tbl::Dict{String,Array{Int64}}, fcounts::Vector{Vector{Float64}}, bcounts::Vector{Vector{Float64}}, corr_Nt::Vector{Vector{Float64}}, log_div::Bool)  
1172     min_stages = 3 # a regime cannot have fewer than this many stages  
1173     clade_start, clade_end = get_start_end_bins(corr_Nt)  
1174     max_reg = 15  
1175     nstage = length(tt)  
1176     bestAIC = 1000000000.0  
1177     psis = [psi for _ in 1 : length(tt)]  
1178  
1179     samp_n = 0  
1180     for (_, value) in occ_tbl  
1181         #nvals = sum(value)  
1182         nvals = count(x -> x > 0, value)  
1183         samp_n += nvals  
1184     end  
1185  
1186     # do single regime model first  
1187     curshifts = Int64[nstage]  
1188     reg = create_regime_vector(curshifts, tt)  
1189     Ks = get_median_diversity(corr_Nt, reg)  
1190  
1191     if log_div == true  
1192         start = vcat(rand(1) * .05, rand(1) * 0.03)  
1193     else  
1194         start = vcat(rand(1) * .05, rand(1) * 0.1 / maximum(corr_Nt))  
1195     end  
1196     append!(start, rand() * .01)  
1197     obj_nlopt = (x, g) -> obj_function_dd_gamma(x, Ks, reg, fcounts, bcounts, corr_Nt, tt, psis, log_div)  
1198     opt = Opt(:LN_NELDERMEAD, length(start))  
1199     opt.min_objective = obj_nlopt  
1200     opt.xtol_rel = 1e-5  
1201     opt.ftol_rel = 1e-5  
1202     lower_b = zeros(length(start))  
1203     upper_b = fill(10.0, length(start))  
1204     opt.lower_bounds = lower_b  
1205     opt.upper_bounds = upper_b  
0 △ 0 ⌂ 0 Julia env: v1.12 -- NORMAL -- ⊕ Ln 1, Col 1 Spaces: 4 UTF-8 LF Julia (Main) ⊞
```

Computation and the Life Sciences

This trend has replicated across virtually all biological subfields

Biology and data science are on a collision course. Here's what you need to know

Mar 23, 2023



Commentaries

Data science in modern evidence-based medicine

Dina Radenkovic ¹, Sir Bruce Keogh ², and Mahiben Maruthappu ³

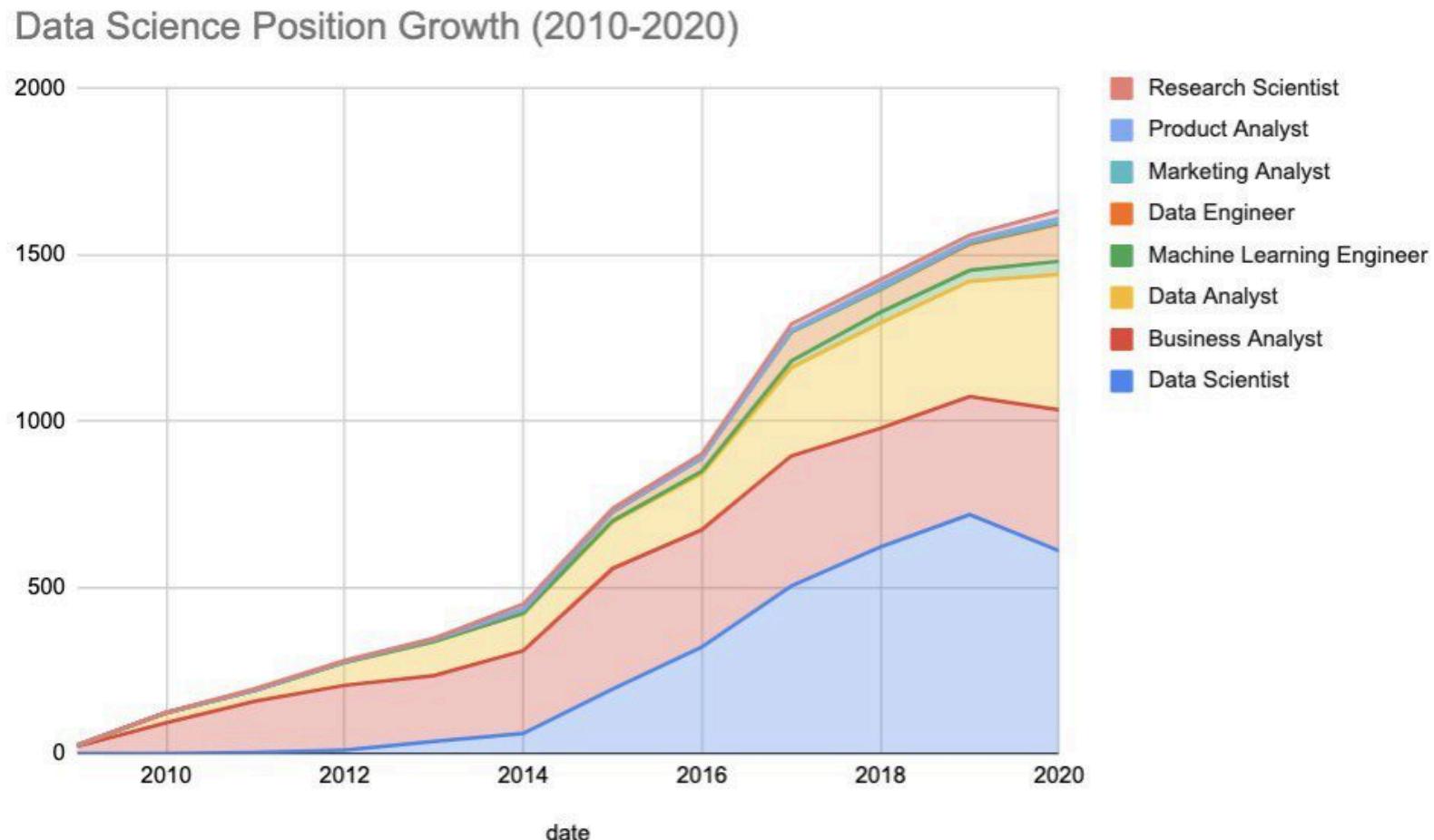
Predicting the Future — Big Data, Machine Learning, and Clinical Medicine

Ziad Obermeyer, M.D., M.Phil. and
Department of Emergency Medicine, Harvard Medical School and Brigham and Women's Hospital, and the Department of Health Care Policy, Harvard Medical School, Boston

Ezekiel J. Emanuel, M.D., Ph.D.
Department of Medical Ethics and Health Policy, Perelman School of Medicine, and the Department of Health Care Management, the Wharton School, University of Pennsylvania, Philadelphia

Data Science in the World

More broadly, this is the case across virtually all fields and industries



What about ChatGPT!?!

- GenAI could reshape what skills are most valued
- This could reduce demand for data science as a profession
- But fundamental skills in data analysis and computation will always be useful



Computation and the Life Sciences

- This semester, we will focus on data science applications to problems and data in ecology and evolutionary biology
- However, the **fundamentals** that we learn are equally applicable to any other field

Computational Evo Bio

- What does a biological data scientist's work actually look like?

Computational Evo Bio

- Many projects have similar flow

Biological idea/hypothesis



Collect data



Develop statistical model(s)



Implement code



Analyze data

Computational Evo Bio

- This course will focus on the last two steps

Biological idea/hypothesis

Collect data

Develop statistical model(s)

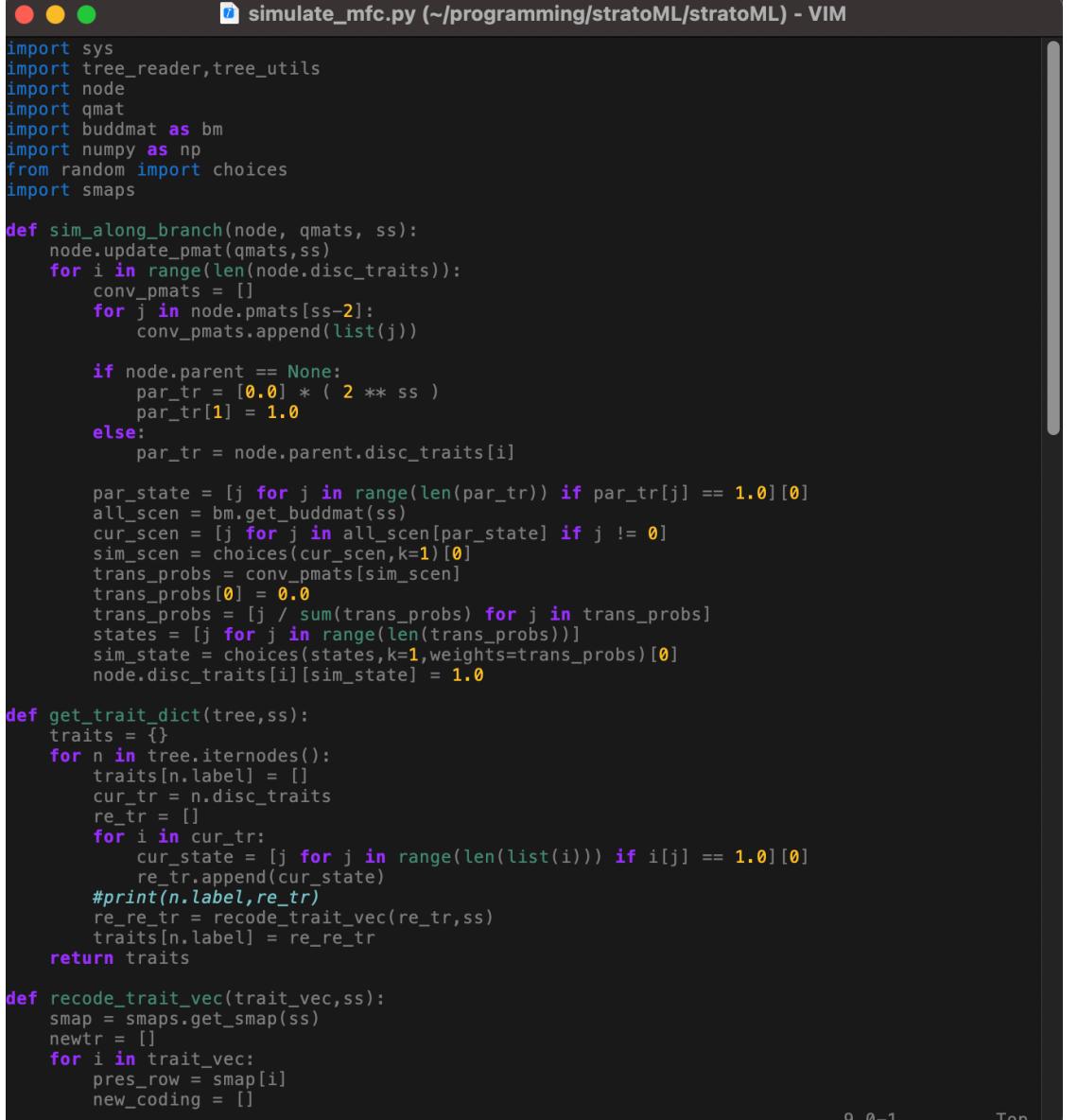
Implement code

Analyze data



Computational Evo Bio

- Data science code is used to manipulate data and test hypotheses



```
● ● ● simulate_mfc.py (~/programming/stratoML/stratoML) - VIM
import sys
import tree_reader,tree_utils
import node
import qmat
import buddmat as bm
import numpy as np
from random import choices
import smaps

def sim_along_branch(node, qmats, ss):
    node.update_pmat(qmats,ss)
    for i in range(len(node.disc_traits)):
        conv_pmats = []
        for j in node.pmats[ss-2]:
            conv_pmats.append(list(j))

        if node.parent == None:
            par_tr = [0.0] * ( 2 ** ss )
            par_tr[1] = 1.0
        else:
            par_tr = node.parent.disc_traits[i]

        par_state = [j for j in range(len(par_tr)) if par_tr[j] == 1.0][0]
        all_scen = bm.get_buddmat(ss)
        cur_scen = [j for j in all_scen[par_state] if j != 0]
        sim_scen = choices(cur_scen,k=1)[0]
        trans_probs = conv_pmats[sim_scen]
        trans_probs[0] = 0.0
        trans_probs = [j / sum(trans_probs) for j in trans_probs]
        states = [j for j in range(len(trans_probs))]
        sim_state = choices(states,k=1,weights=trans_probs)[0]
        node.disc_traits[i][sim_state] = 1.0

def get_trait_dict(tree,ss):
    traits = {}
    for n in tree.iternodes():
        traits[n.label] = []
        cur_tr = n.disc_traits
        re_tr = []
        for i in cur_tr:
            cur_state = [j for j in range(len(list(i))) if i[j] == 1.0][0]
            re_tr.append(cur_state)
        #print(n.label,re_tr)
        re_re_tr = recode_trait_vec(re_tr,ss)
        traits[n.label] = re_re_tr
    return traits

def recode_trait_vec(trait_vec,ss):
    smap = smaps.get_smap(ss)
    newtr = []
    for i in trait_vec:
        pres_row = smap[i]
        new_coding = []
        for j in range(len(pres_row)):
            if pres_row[j] == 1.0:
                new_coding.append(1)
            else:
                new_coding.append(0)
        newtr.append(new_coding)
```

Computational Evo Bio

- This course is really about learning to *think like a scientist*
 - Python is the *language* we use to engage with this
 - Statistical models and approaches are our *tools*
- You will learn these tools to see how scientists formulate and test ideas
- These skills are irreplaceable and useful no matter what you do

EEB 125 learning goals

By the end of this semester, you will be able to:

- Write general Python code that can be used in many settings
- Manipulate and clean real datasets using Python
- Create visualizations to summarize patterns in real datasets
- Perform statistical analyses to test scientific hypotheses