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

MAY 24, 2021

Complex molecules could hold the secret to identifying alien life

by University of Glasgow

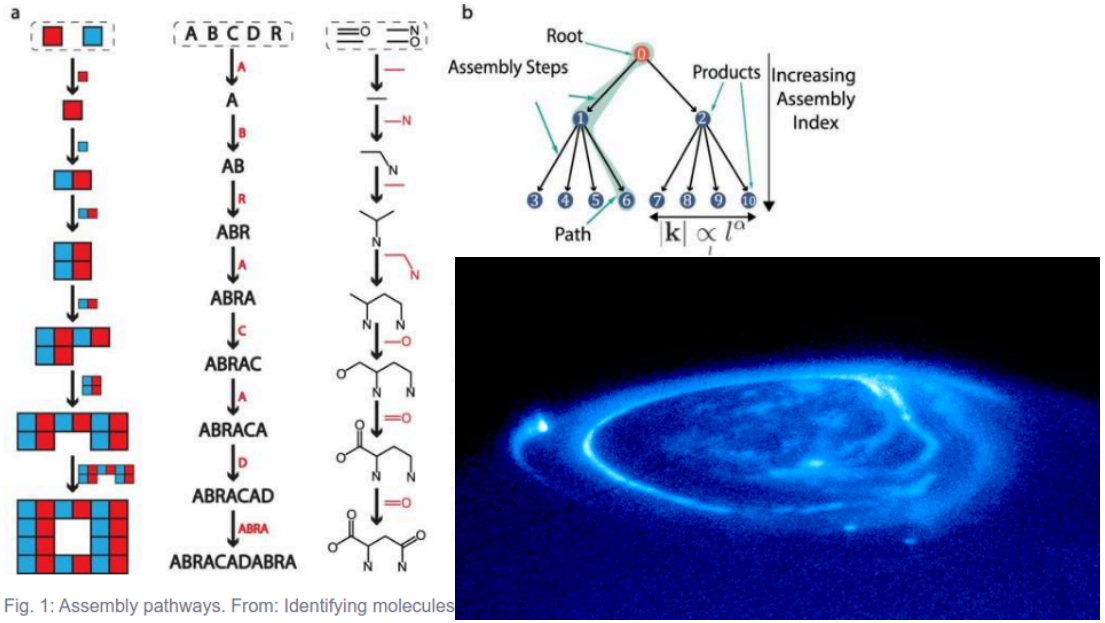


Fig. 1: Assembly pathways. From: Identifying molecules

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

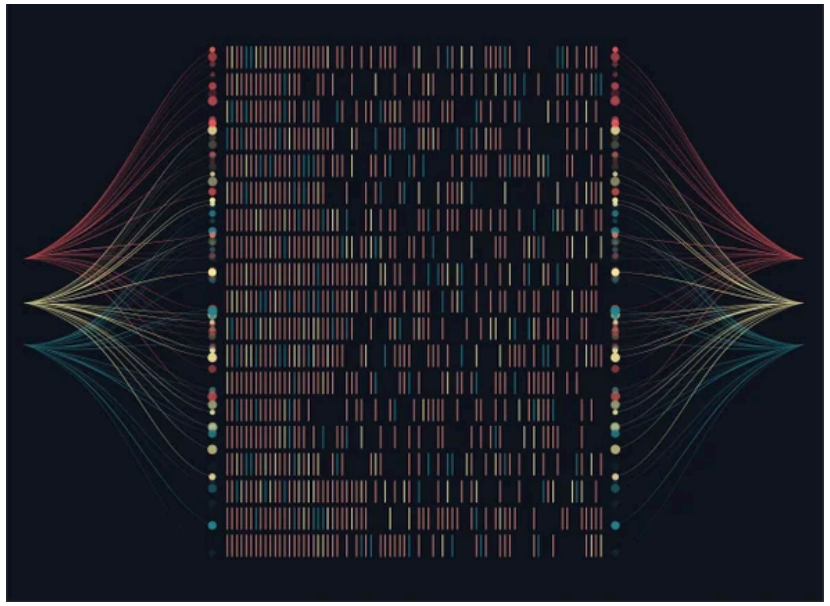
Can we better understand human neurological diversity?

NEUROLOGY | OPINION

How Big Data Are Unlocking the Mysteries of Autism

Better genetic insights can help support people across the spectrum

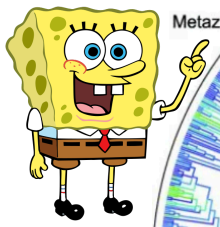
By Wendy Chung on April 30, 2021

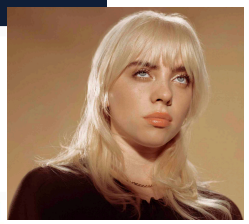


Can we better understand the shared history across all life?

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.





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 Seligman

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 emotio volume of the relevant discography, and that uniquely identifying a song that optimally describes a hy song, we separate the criteria into the level of optimism (H) and the strength of commitment to a rel 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 wh with blue eyes and/or bad reputations may lead to overall less positive emotions, while those with gr that these trends are based on small sample sizes, and more data are necessary to validate them. F specific mood.

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

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

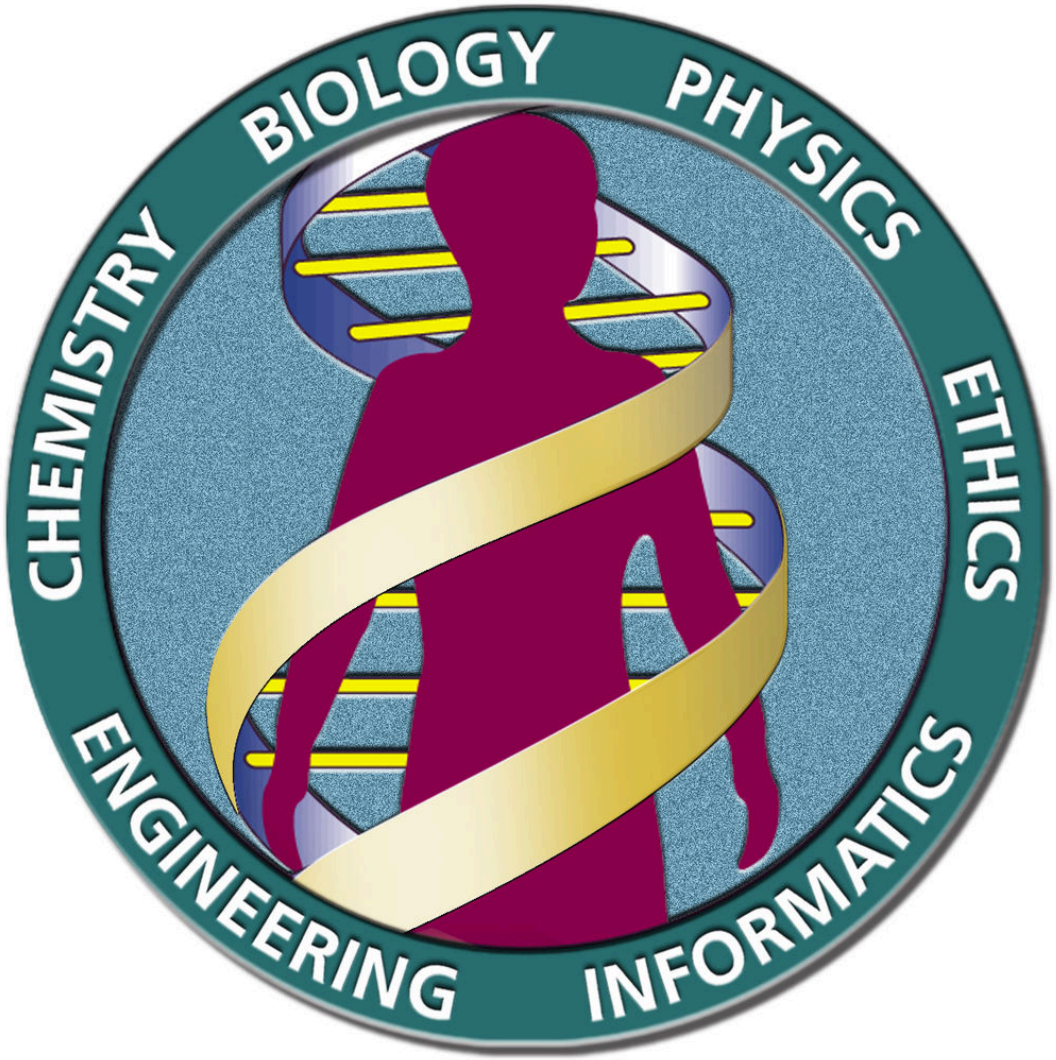
These are all **data science** questions

These are all **data science** questions

Data science has *revolutionized* biology

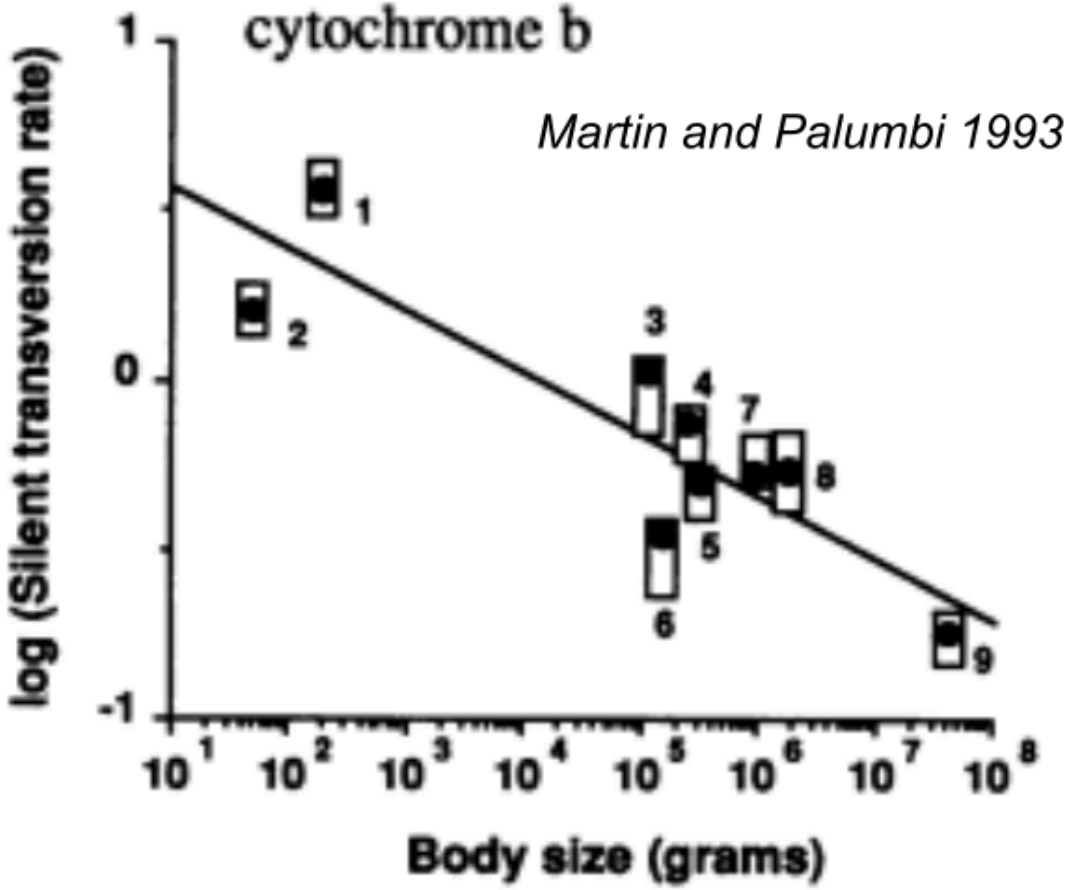
The Human Genome Project (1990-2003)

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



Scientific datasets have become massive

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

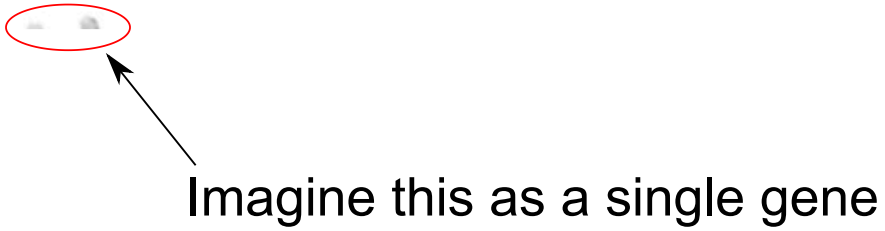


'Genetics' -> 'Genomics'

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

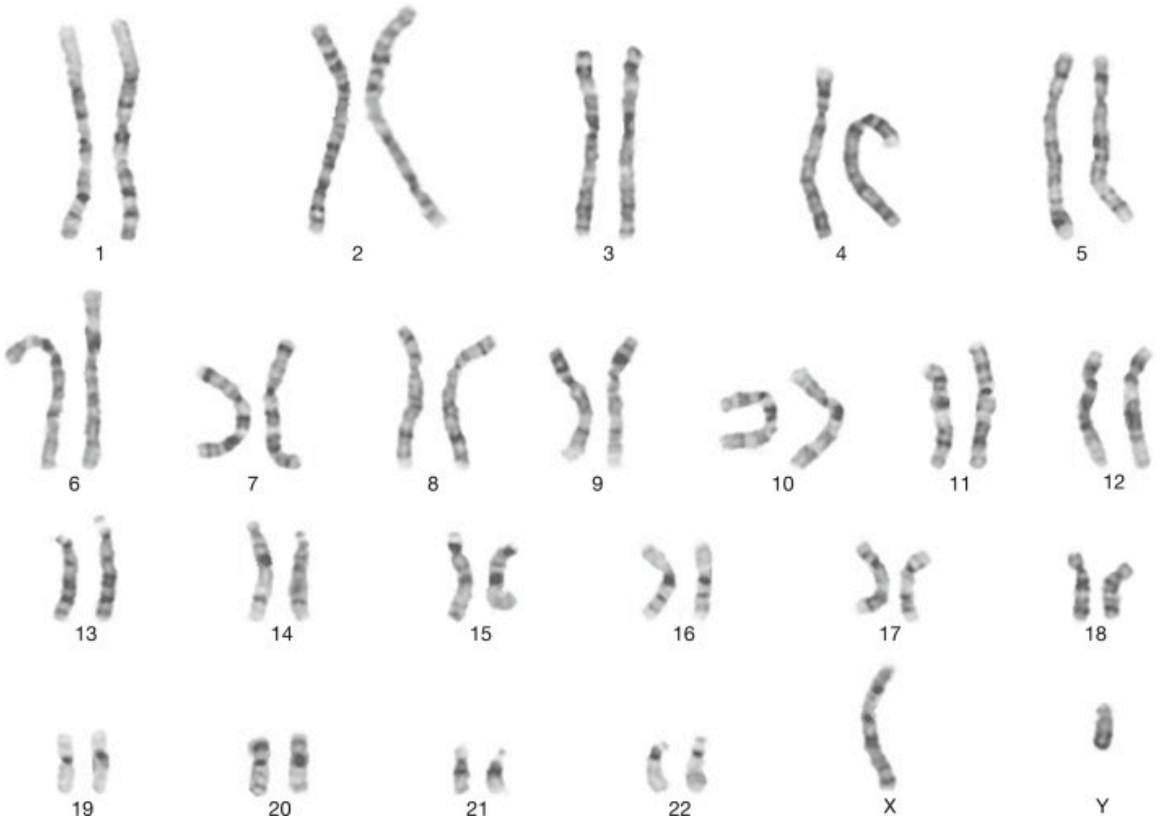
Pre-2003:

- Analyze ~3 **thousand** nucleotides at once



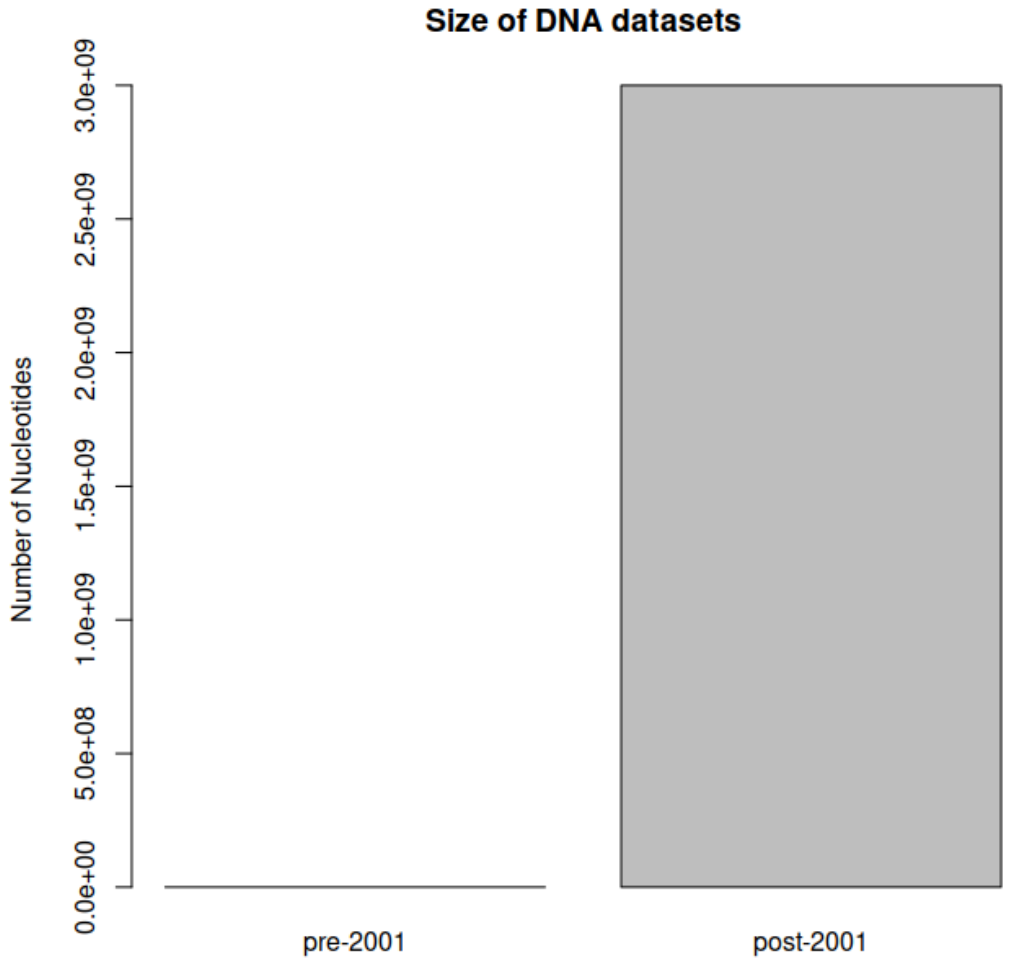
Post-2003:

- 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!

Data science breakthroughs

We have learned a lot about human biology using data science

NEWS | 22 August 2018 | Correction 23 August 2018

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

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 Barras

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

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

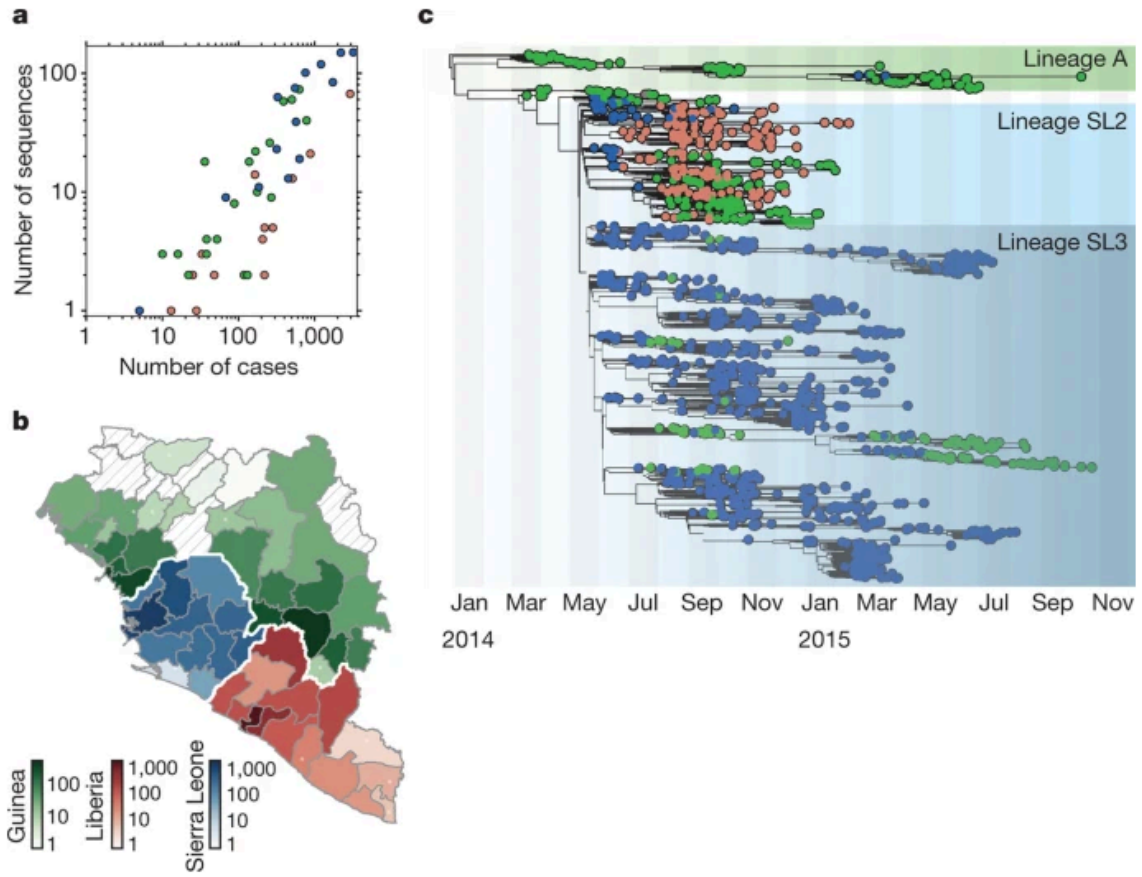
+ See all authors and affiliations

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

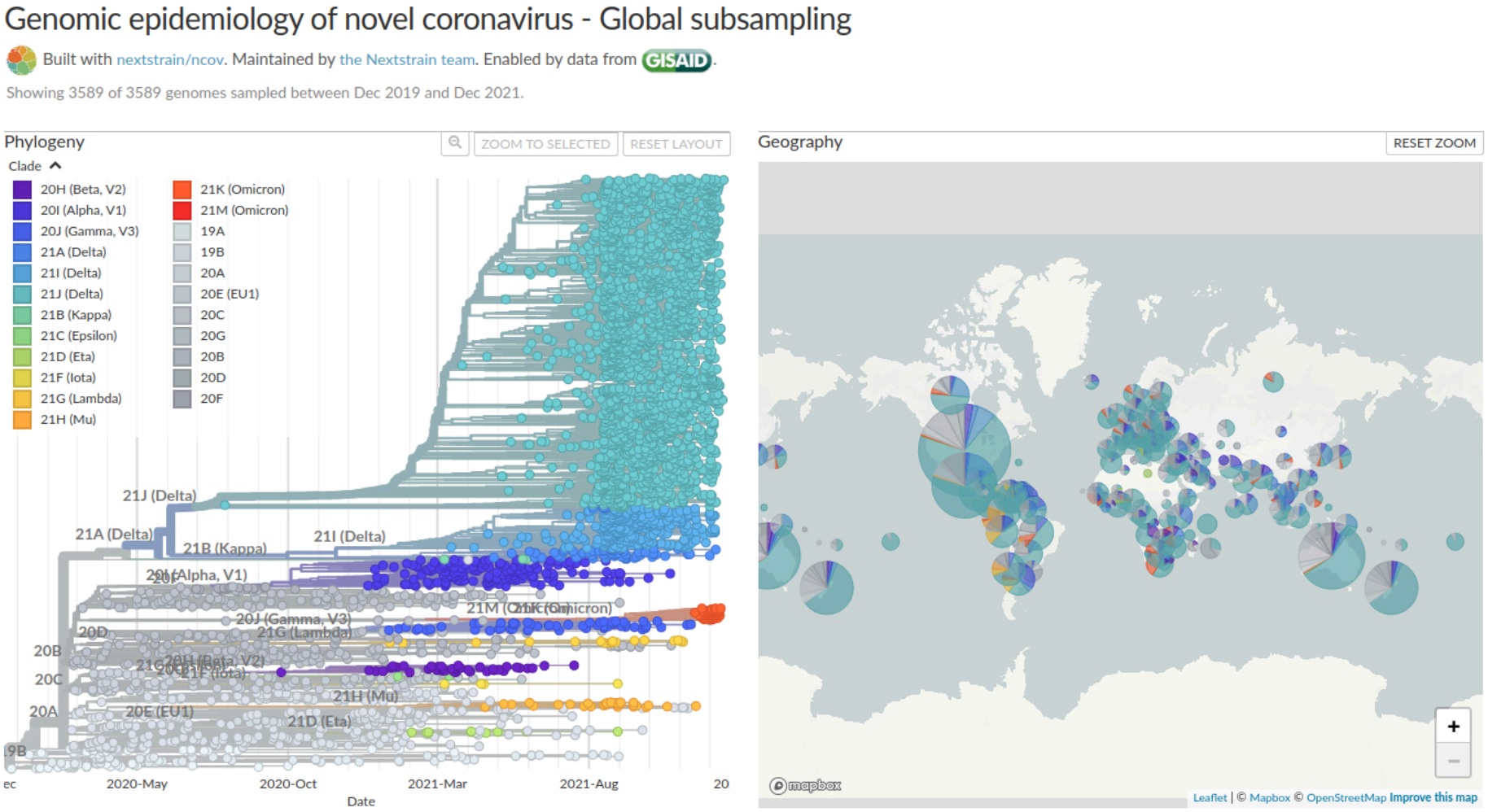
Contributed by Svante Pääbo, January 22, 2021 (sent for review December 21, 2020; reviewed by Tobias L. Lenz and Lluis Quintana-Murci)

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



Data science tools are also marketed to the public

23andMe®

ANCESTRY BREAKDOWN

Dig deeper into your ancestry.

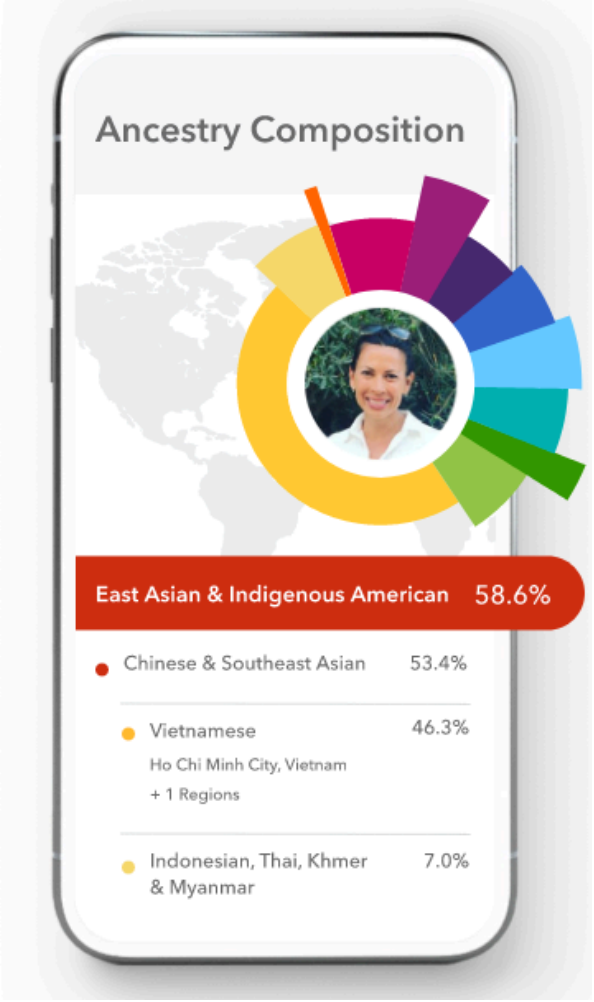
It's the most complete genetic breakdown on the market, and the most comprehensive portrait of you yet.

- **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

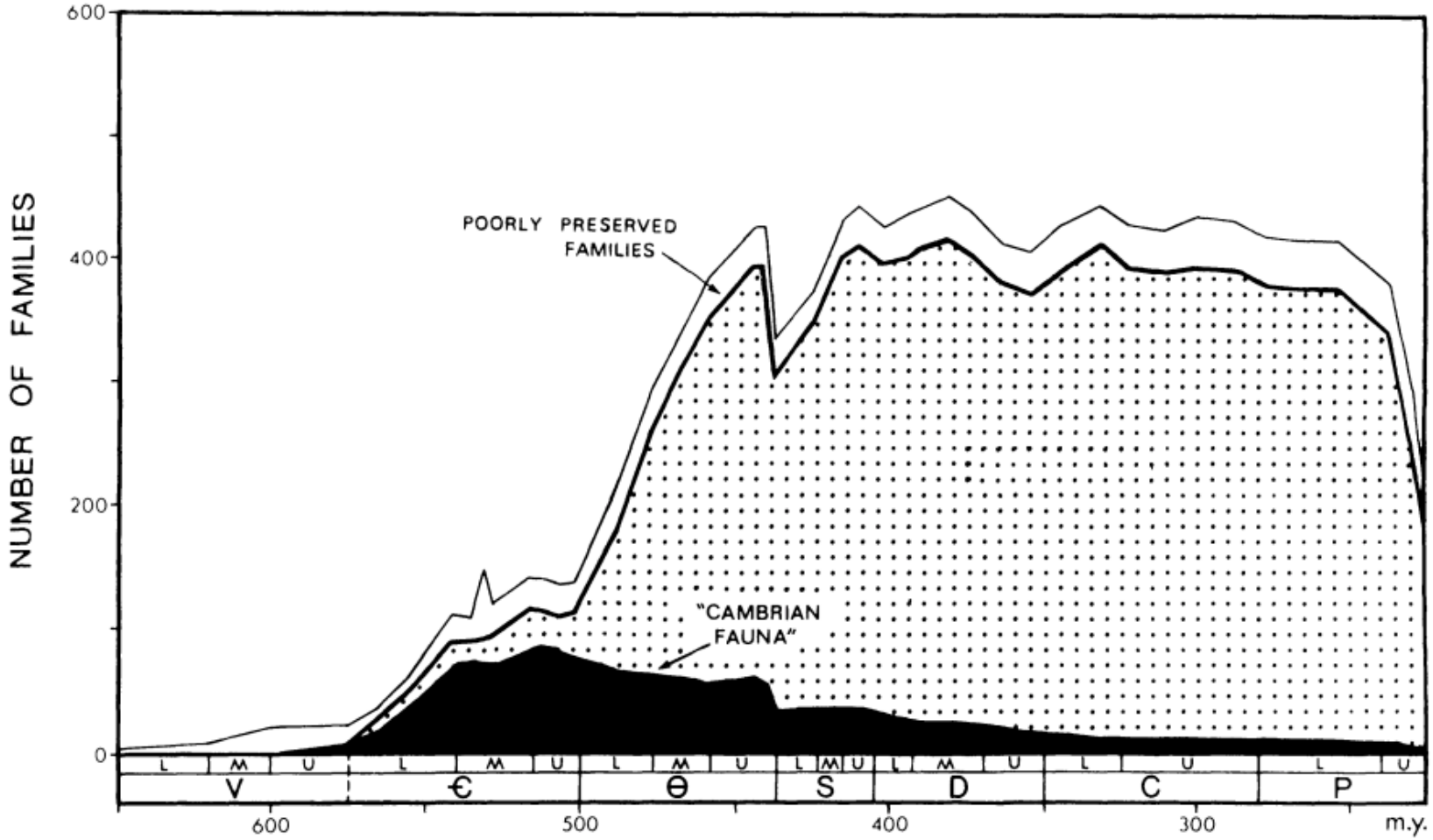
what do you think of when you imagine a paleontologist?

Computational paleobiology

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

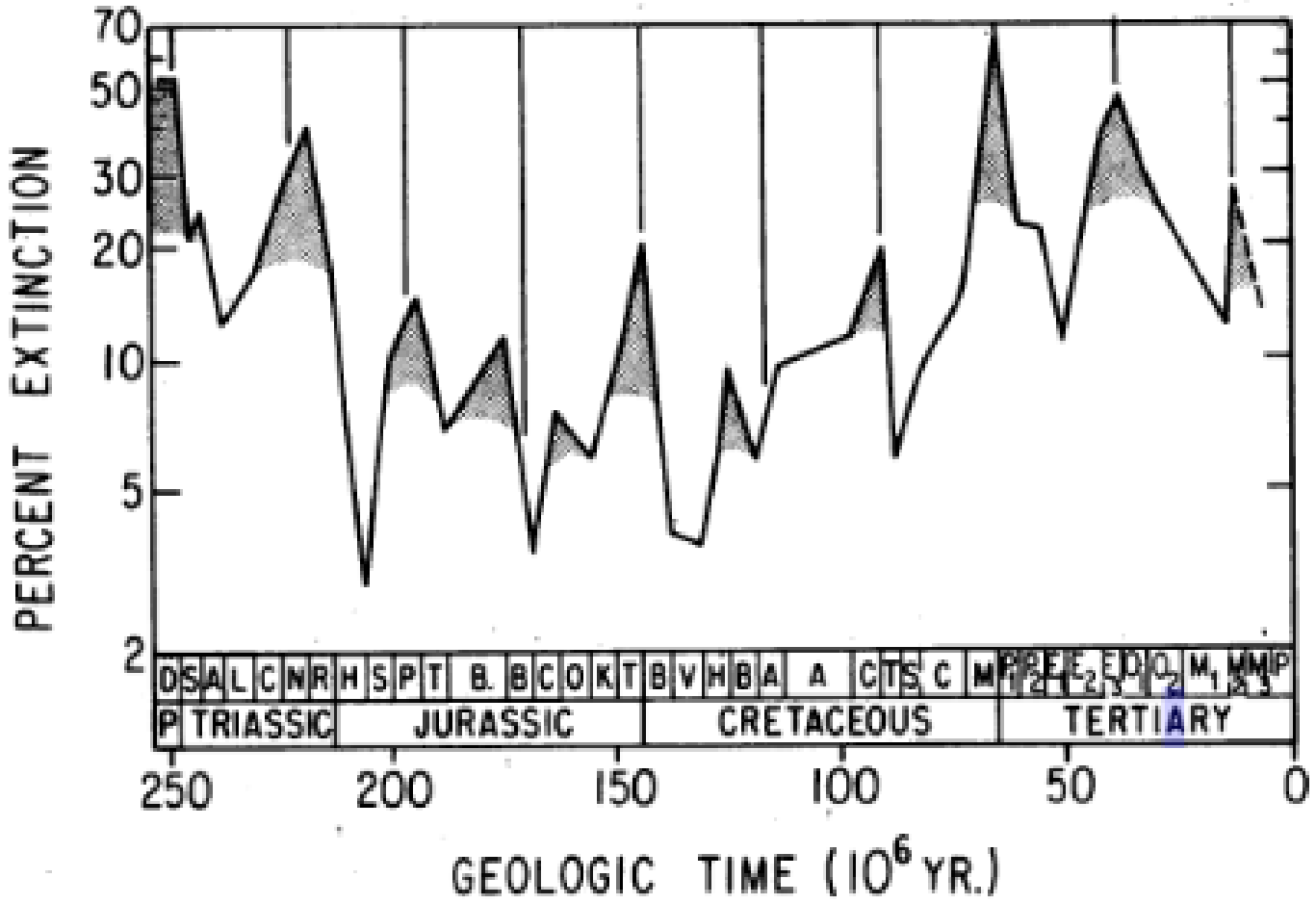
Computational paleobiology

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

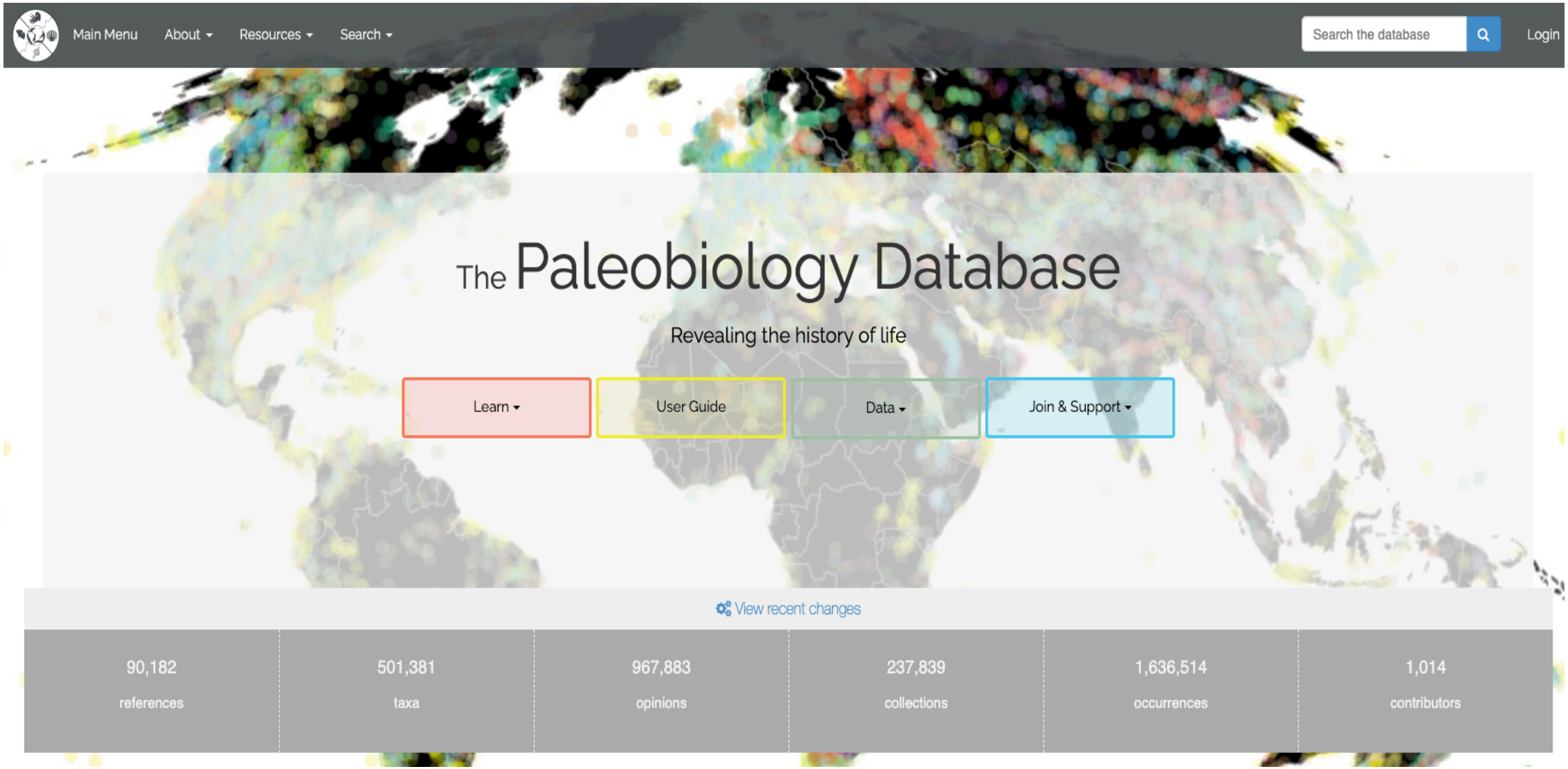


Computational paleobiology

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



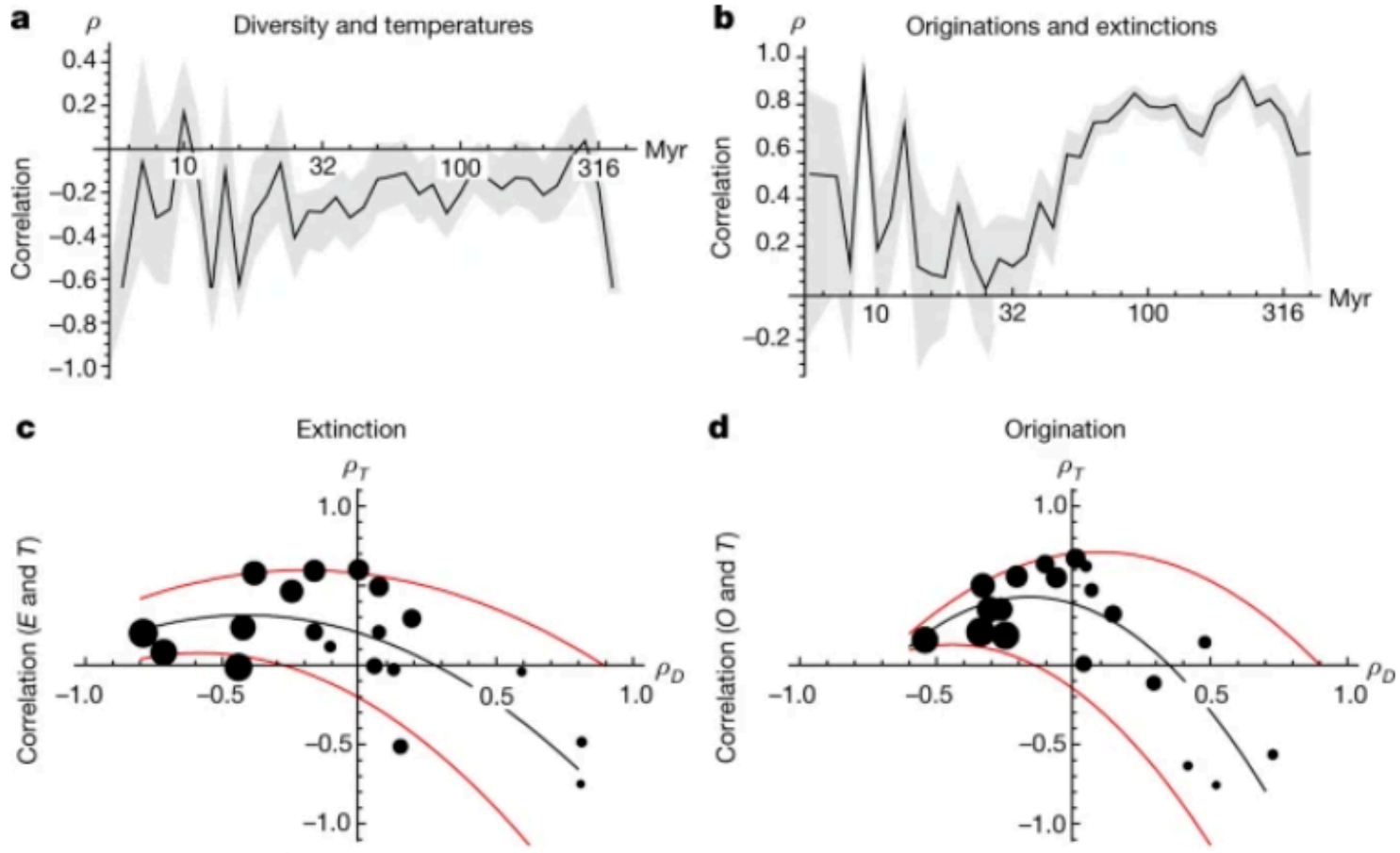
Computational paleobiology



Computational paleobiology

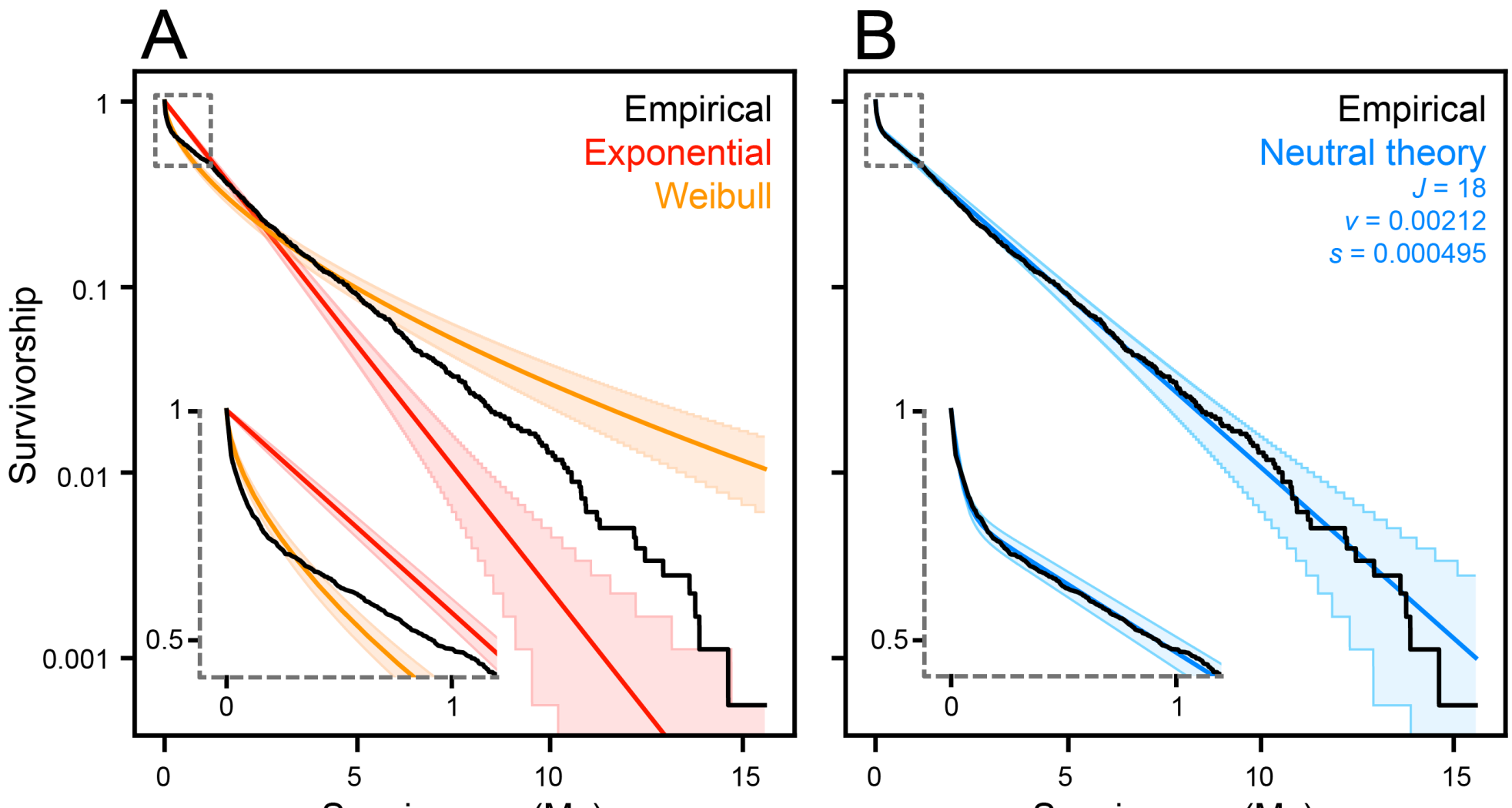
- Large-scale analyses of thousands of fossils representing thousands of species

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



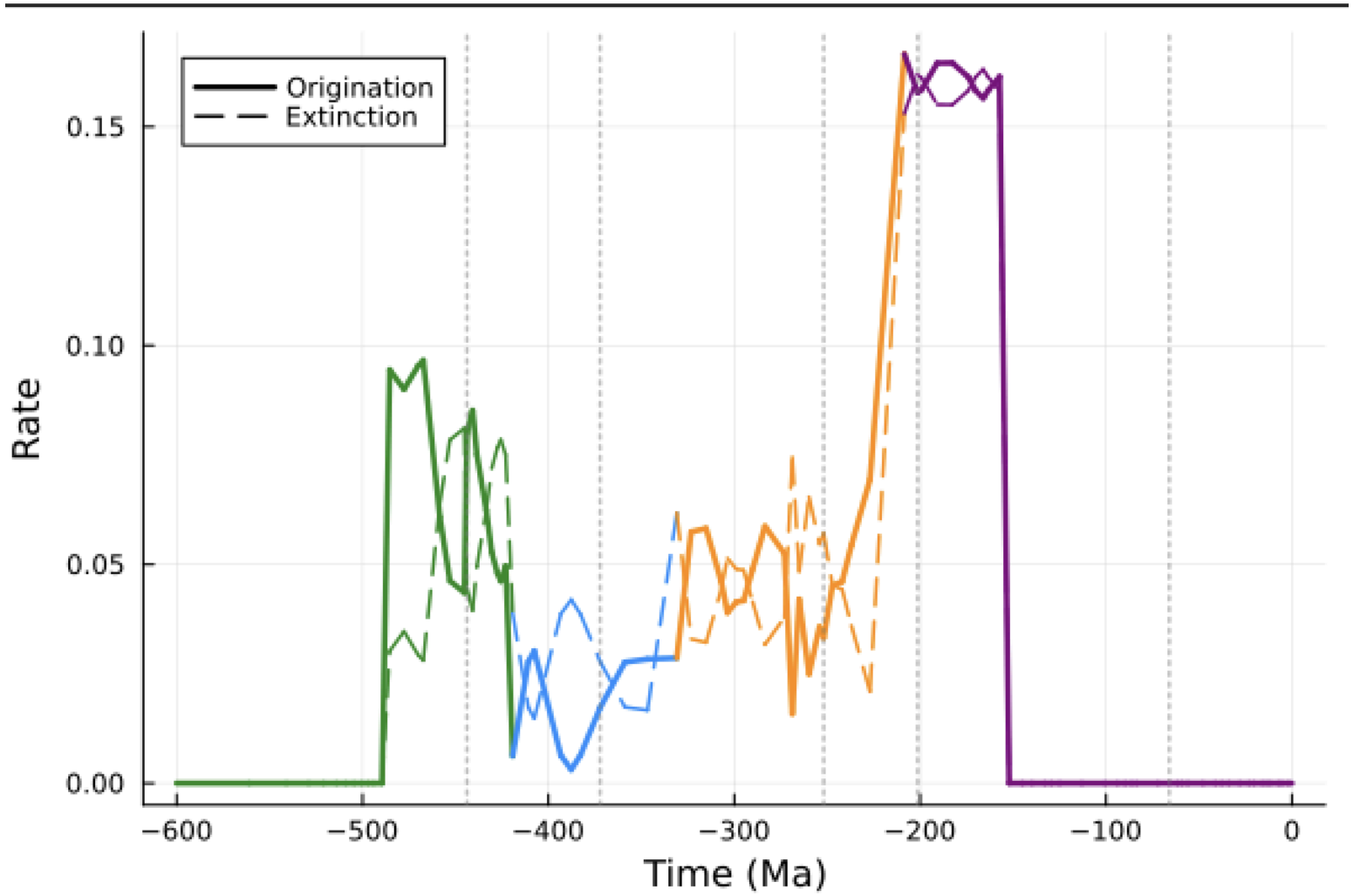
Computational paleobiology

- 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

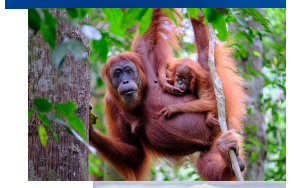
Who am I?

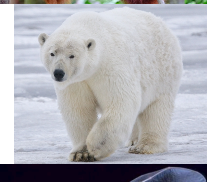
- Evolutionary paleobiologist
- I develop new computational approaches to ask 'big picture' questions about evolution
- For example...

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

KPg Boundary (66mya)

Present





...in just a few million years

KPg Boundary (66mya)

Aquatic
Massive size
Swimming locomotion

Grasping
Enhanced vision
Long lifespan

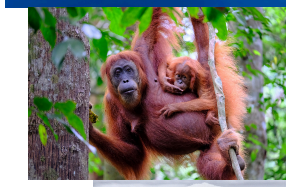
Carnivory traits
Large size
Locomotor

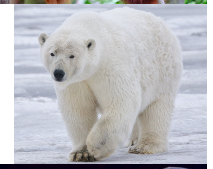
Echolocation
Flight

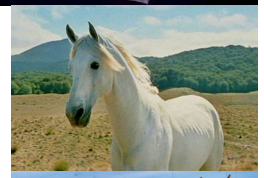
Hooves
Long legs
Grazing traits

Carnivory
Large size
Walk on toes

Present







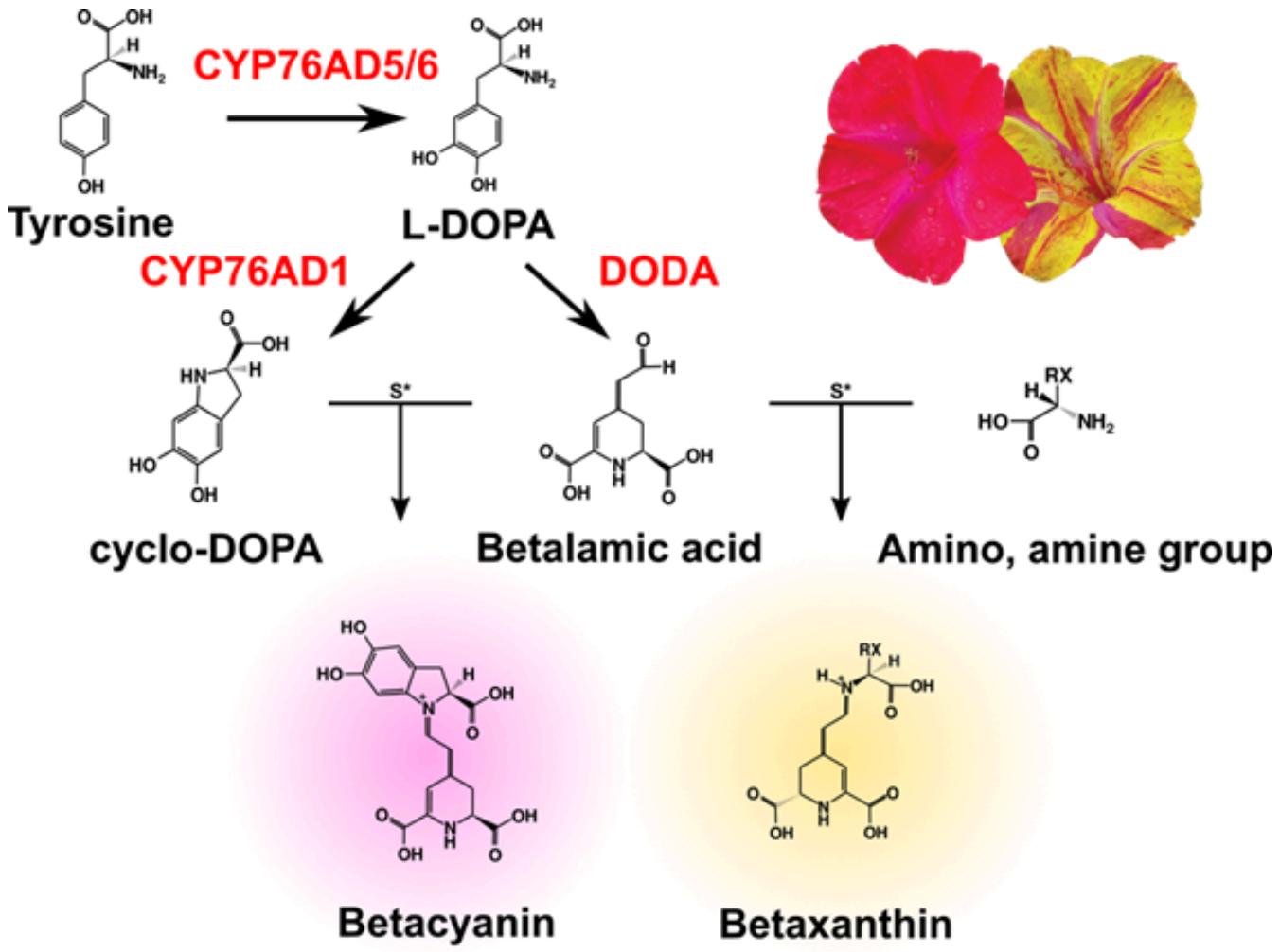
Most biologists are interested in the origin of cool features

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

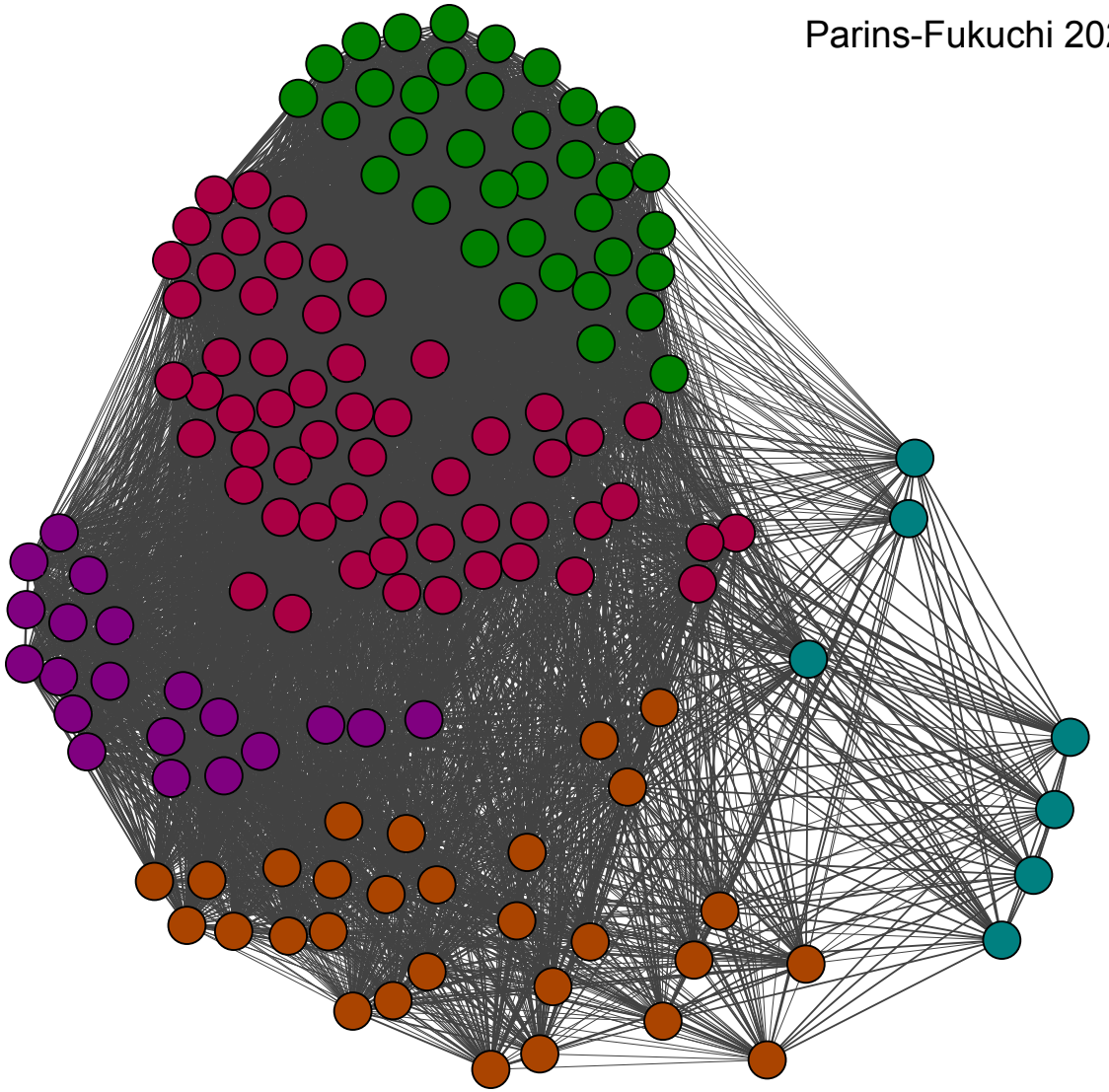
- skeletal features associated with novel locomotor modes

Apes

Ape skeletons display many locomotor innovations

Skeletal features display complex functional and genetic relationships

Parins-Fukuchi 2020, Evolution



fossil crinoids

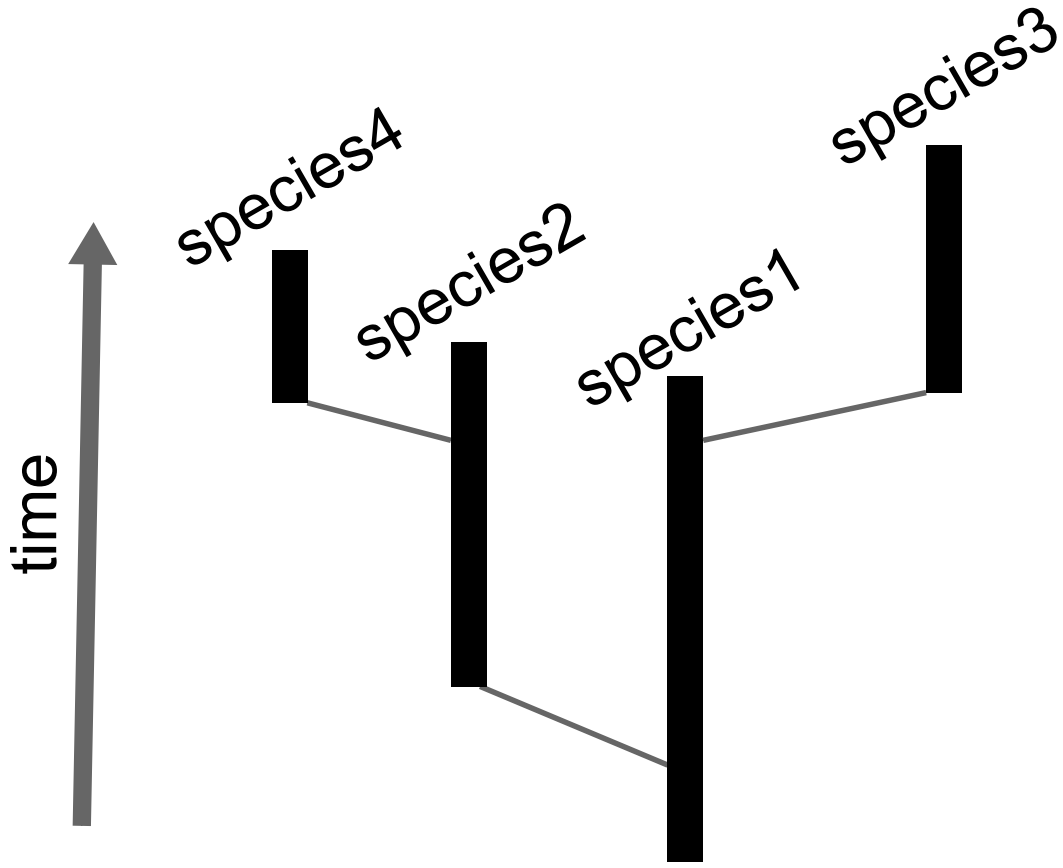
- how does variation in groups of sea creatures affect their evolution?

fossil sea urchins

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

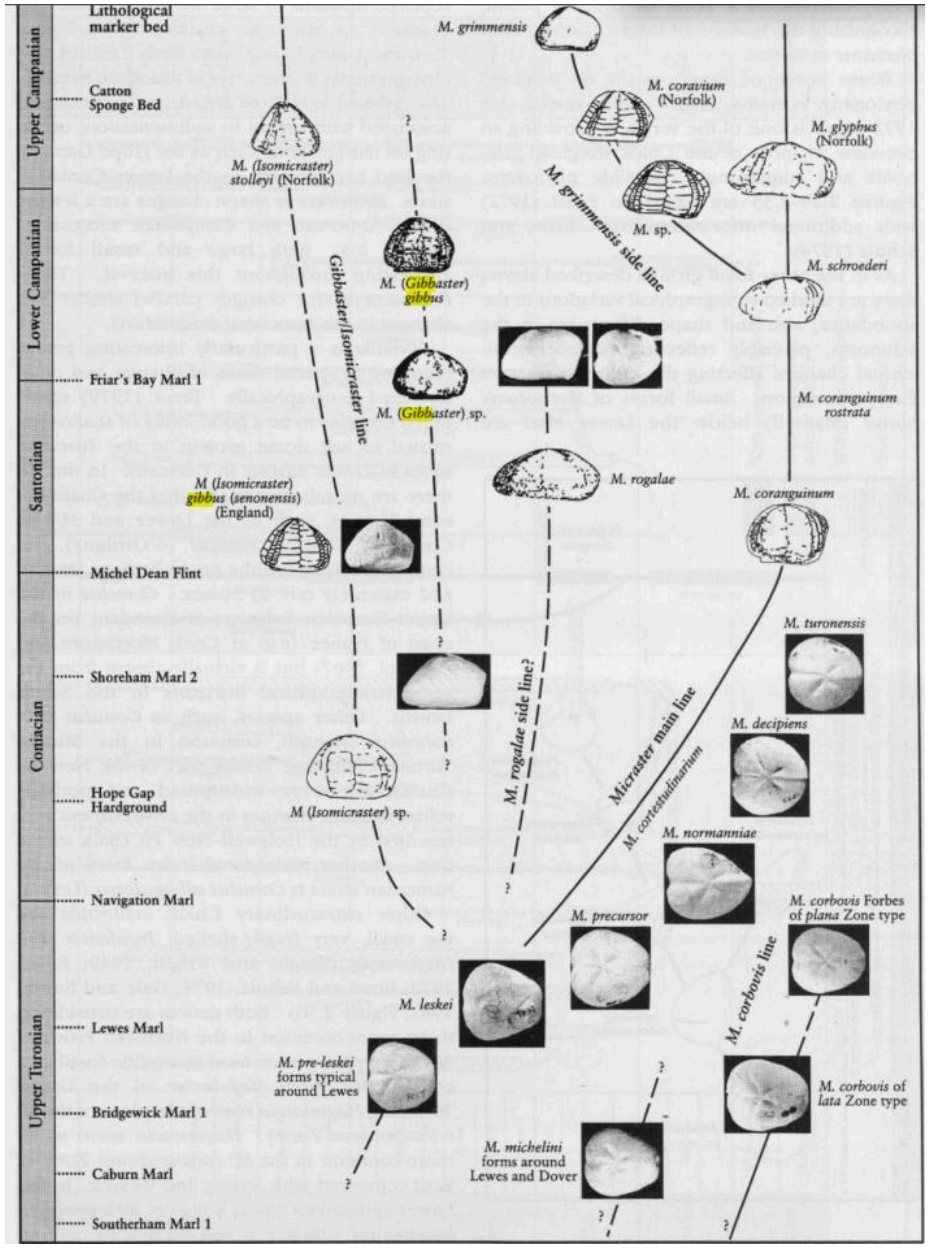
phylogenetic trees

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



Micraster

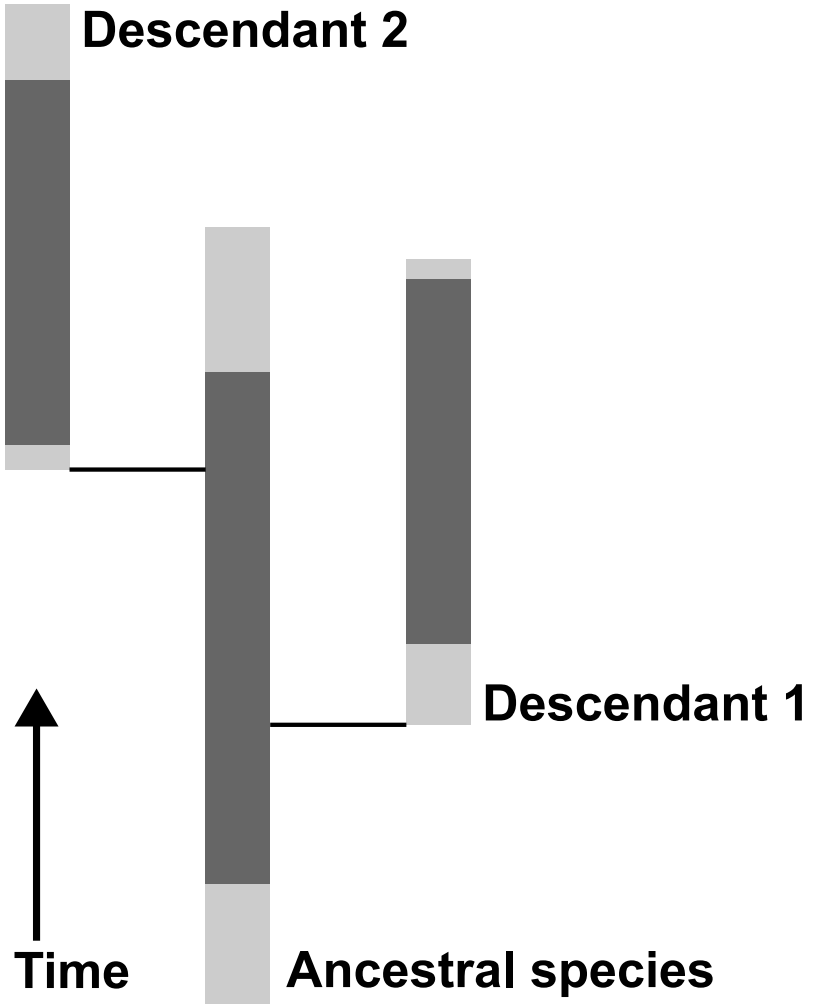
- evolutionary relationships used to be mainly hand-drawn



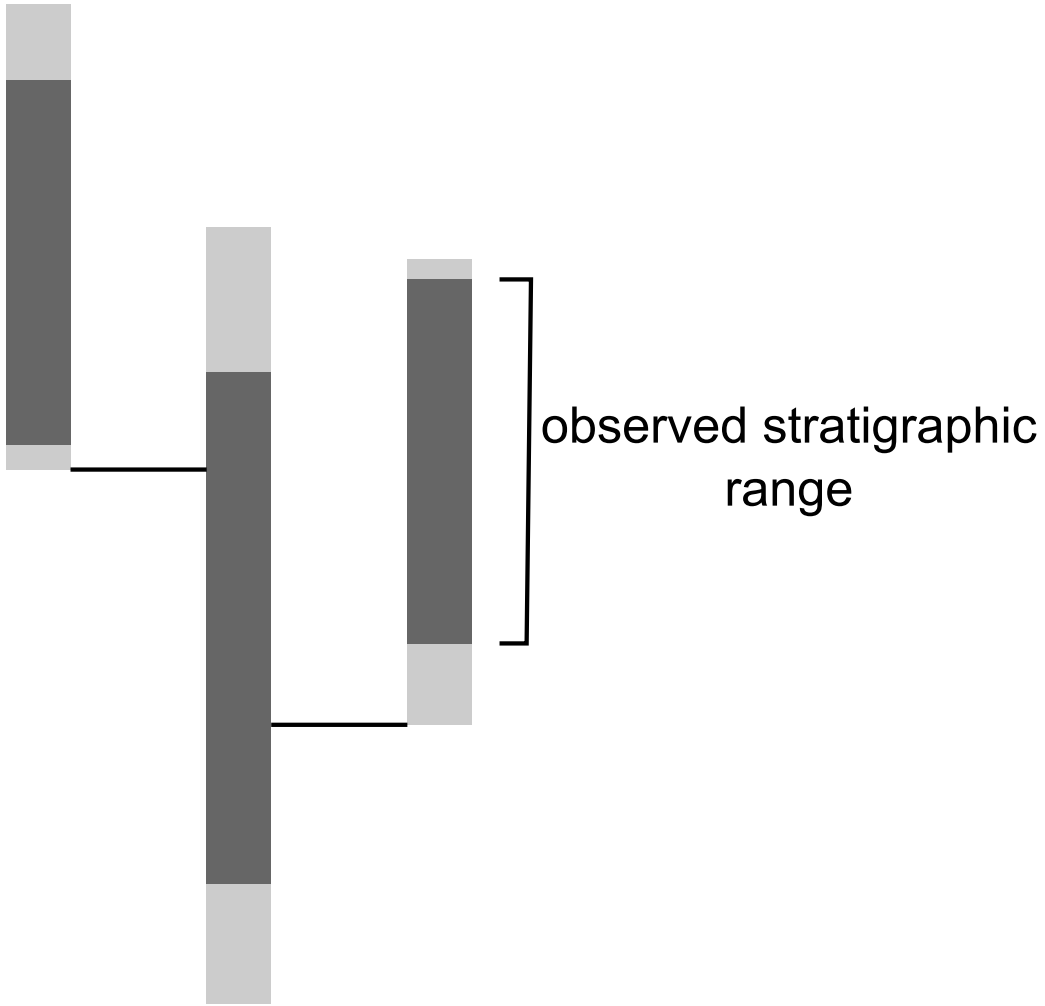
phylogenetic trees

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

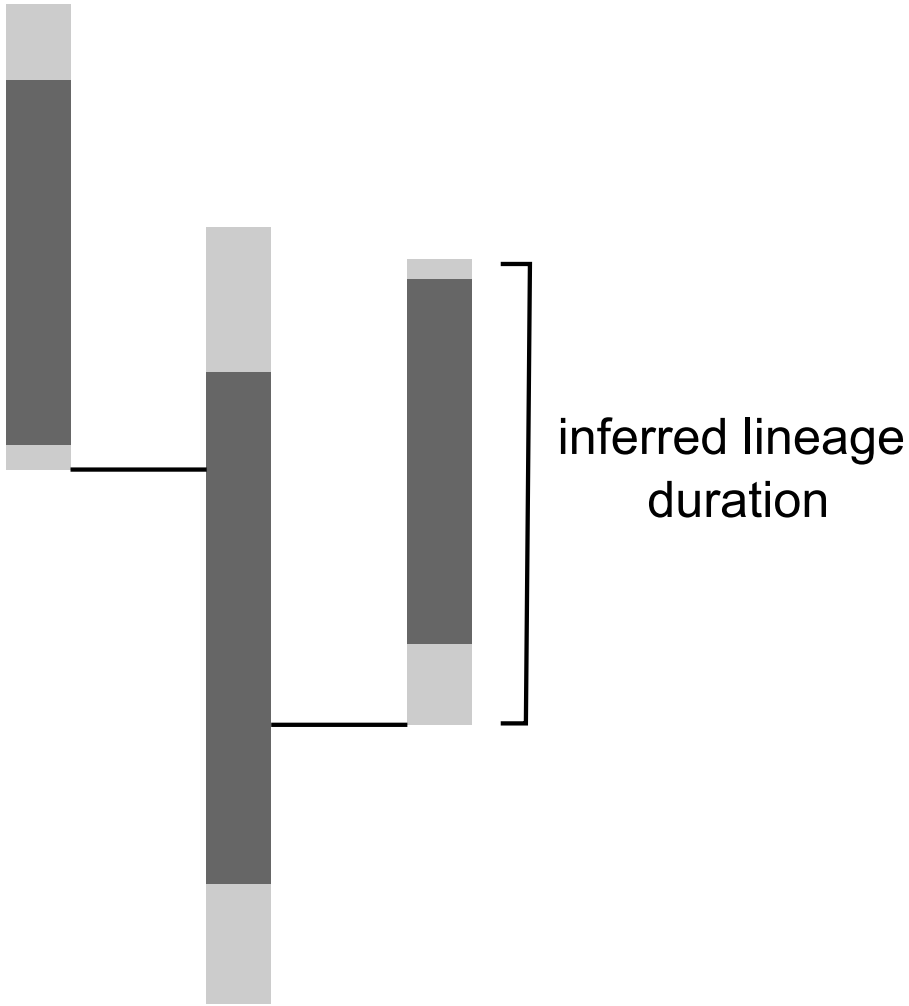
phylogenetic trees



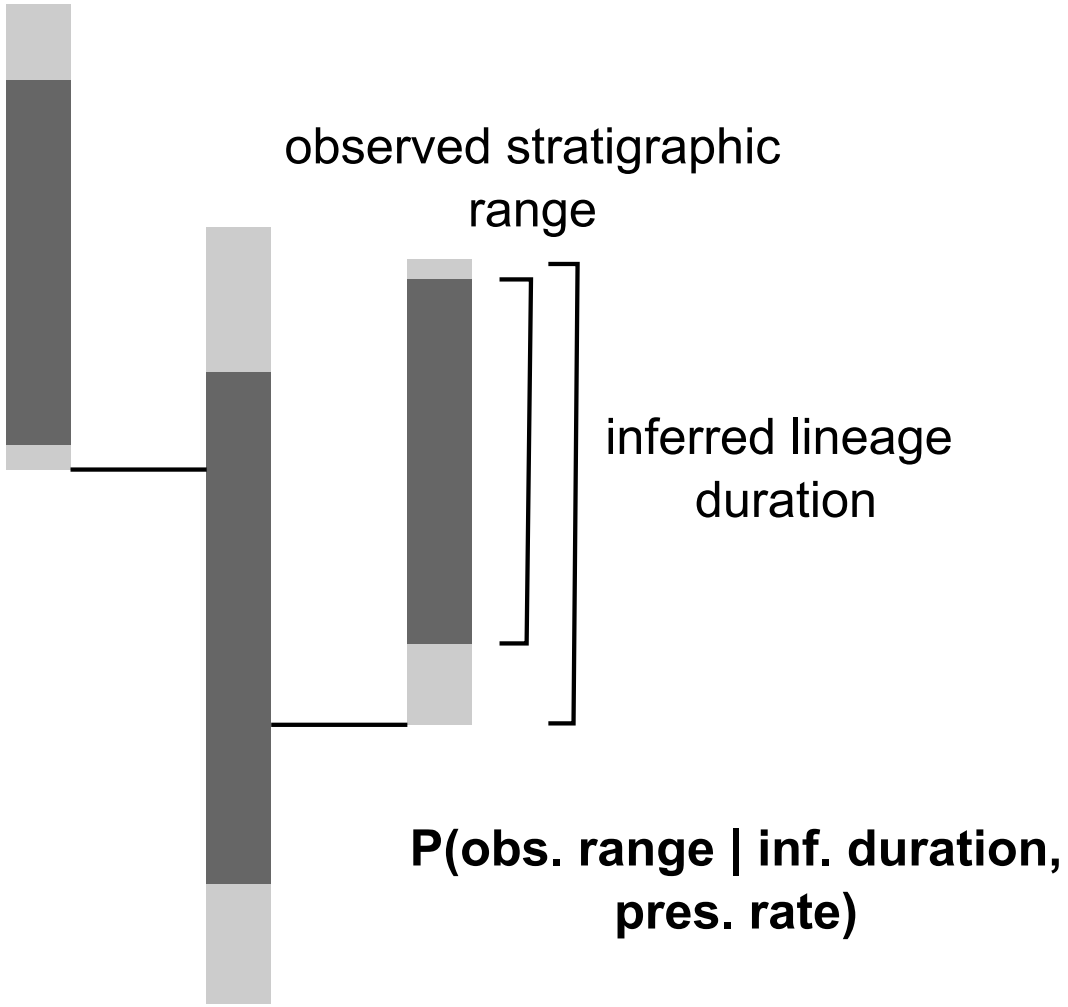
phylogenetic trees



phylogenetic trees

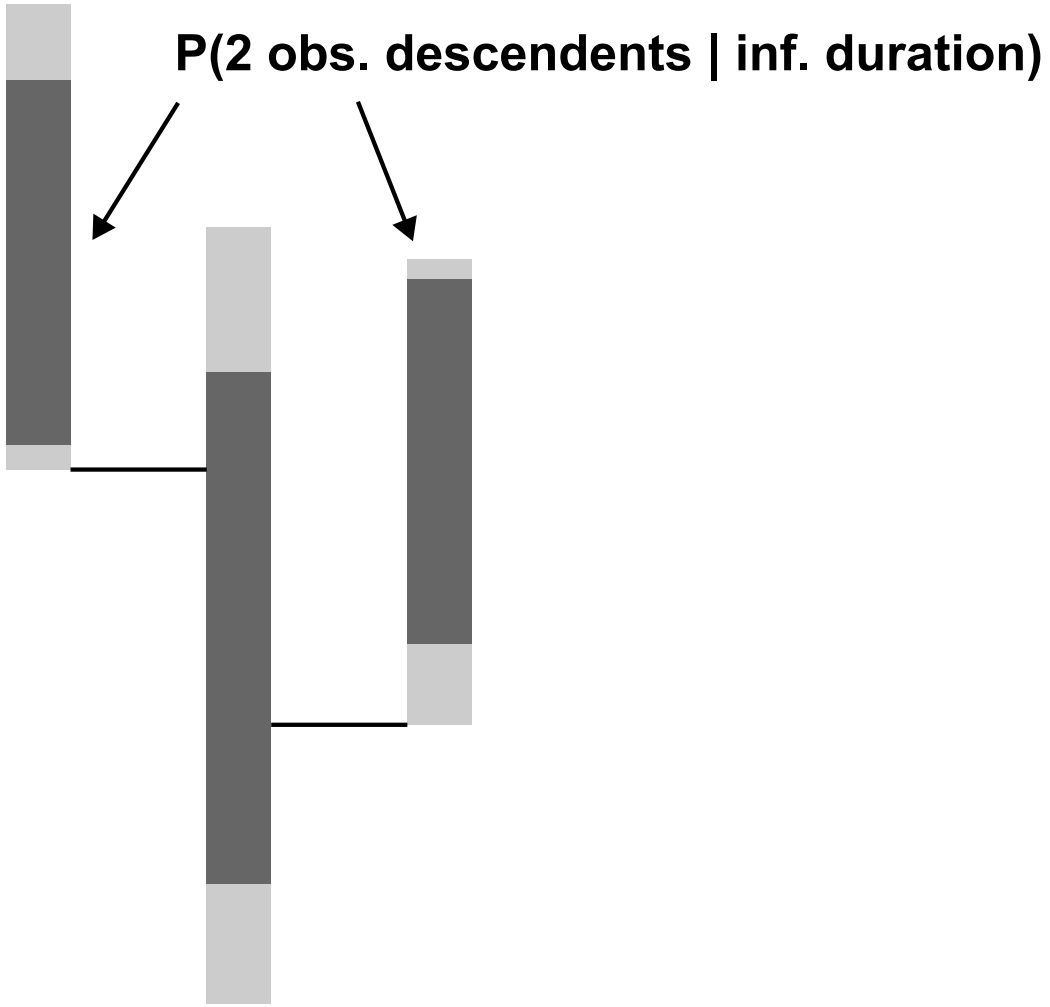


phylogenetic trees

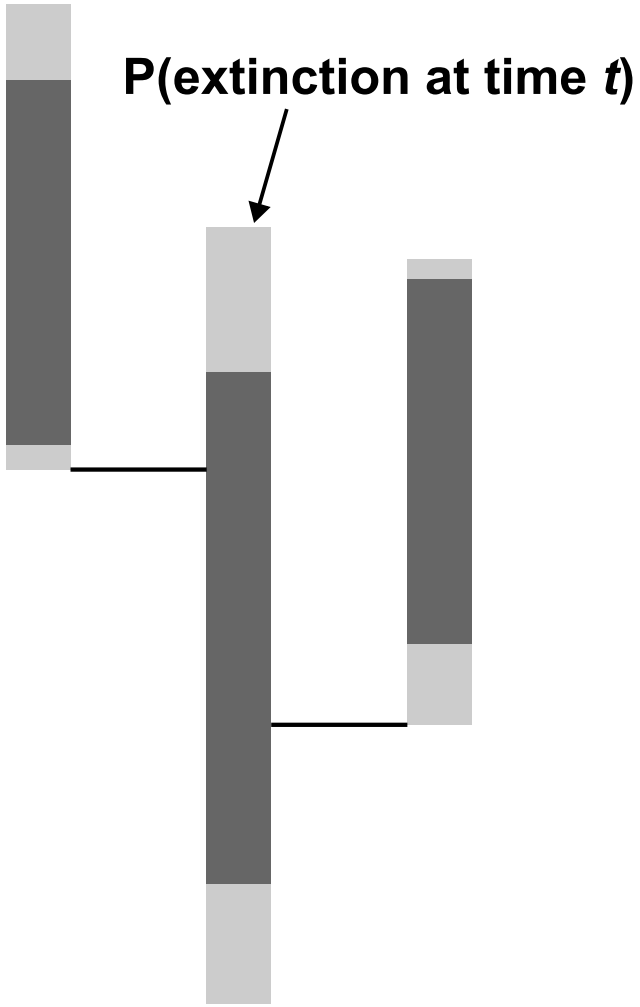


P(obs. range | inf. duration, pres. rate)

phylogenetic trees



phylogenetic trees

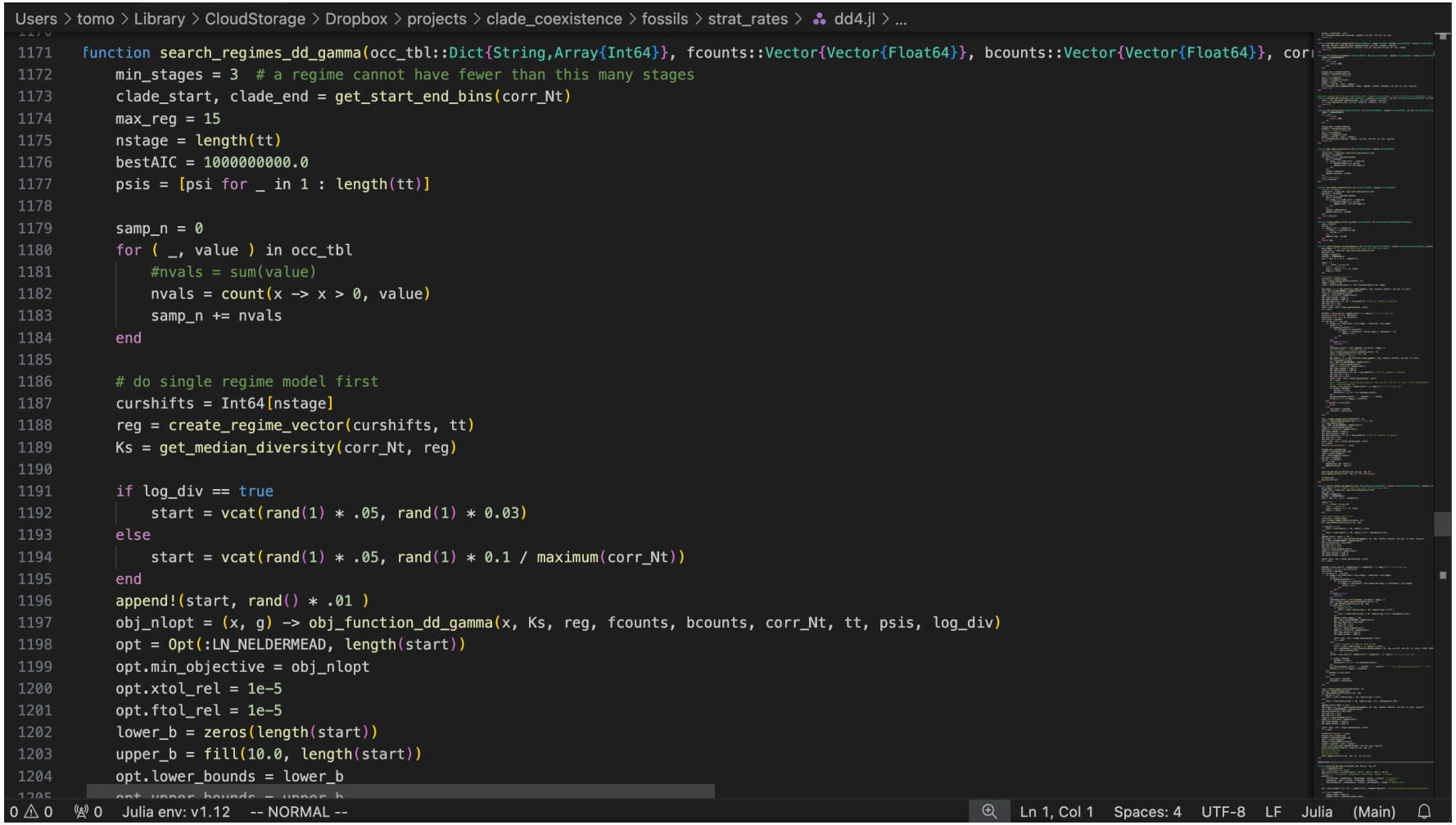


computational paleobiology

Instead of looking like this:

computational paleobiology

Things tend to look much more like:



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

With the power of AI, we now have the ability to

ACCELERATE
the discovery of novel targets

PREDICT
the effectiveness of treatments

IDENTIFY
potentially life-saving clin trials

DIAGNOSE
multiple diseases earlier

Commentaries

Data science in modern evidence-based medicine

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

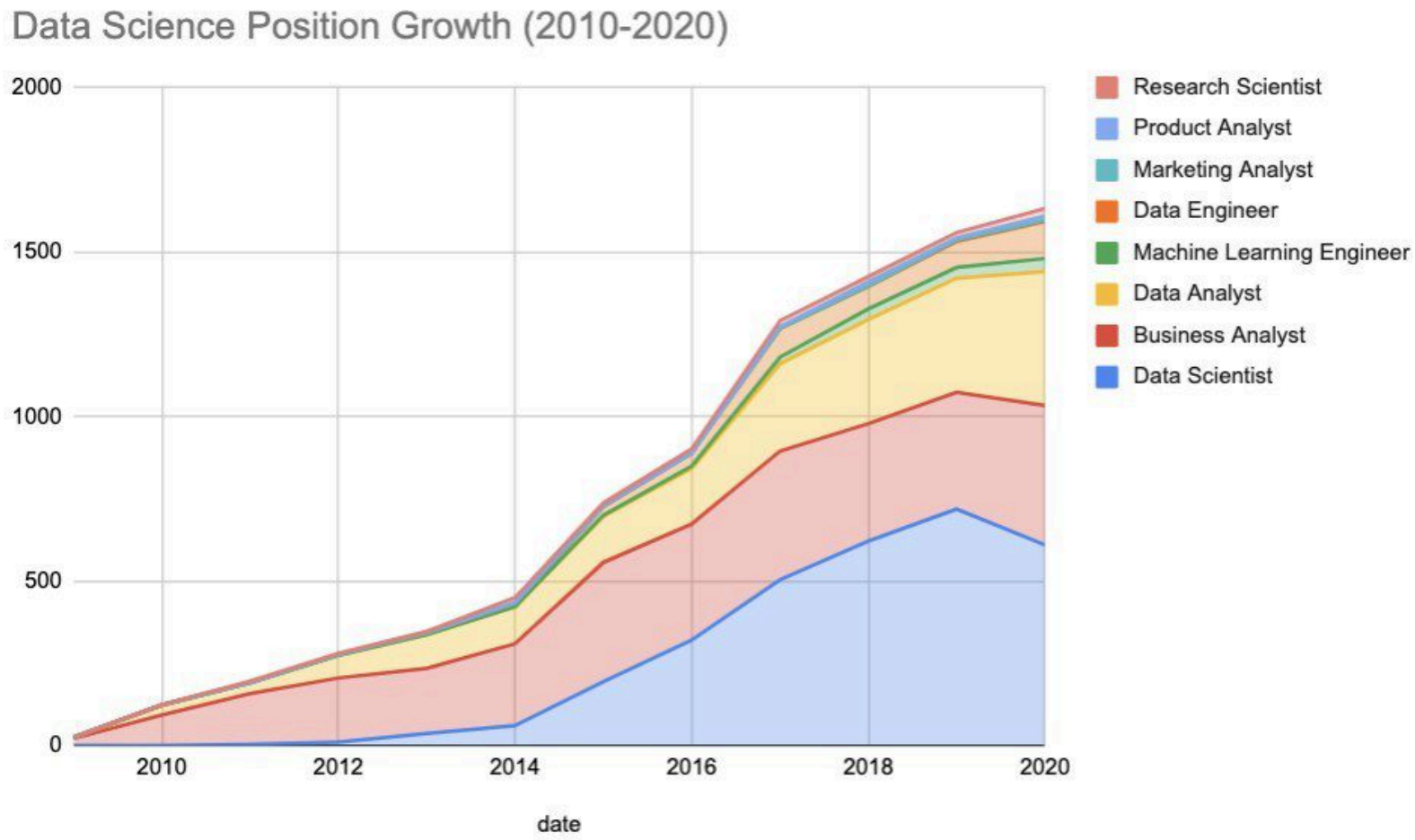
Predicting the Future — Big Data, Machine Learning, and Clinical Medicine

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

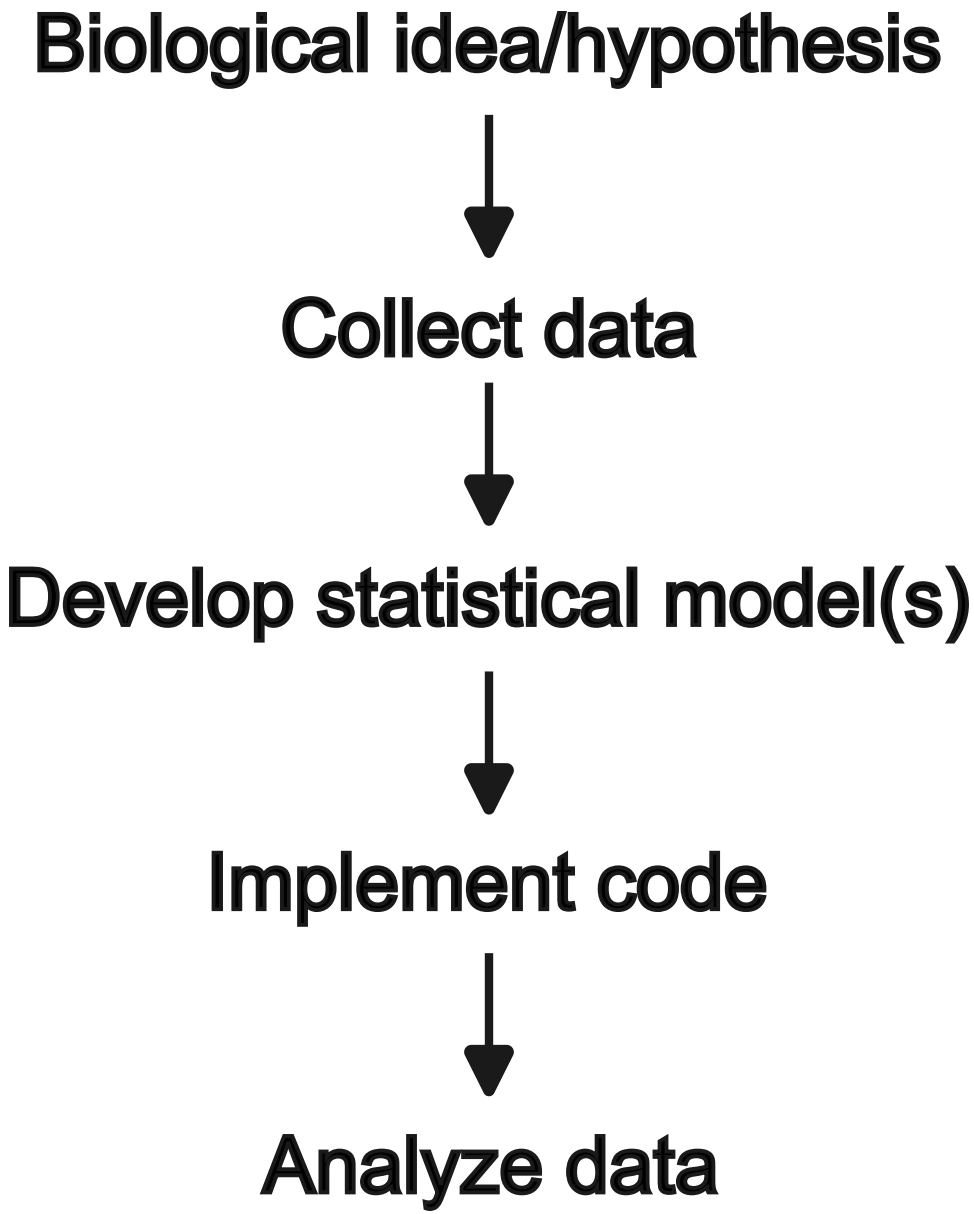
- This semester, we will focus on data science applications to problems and data in ecology and evolutionary biology
- However, the **fundamentals** that we learn are equally applicable to any other field

Computational Evo Bio

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

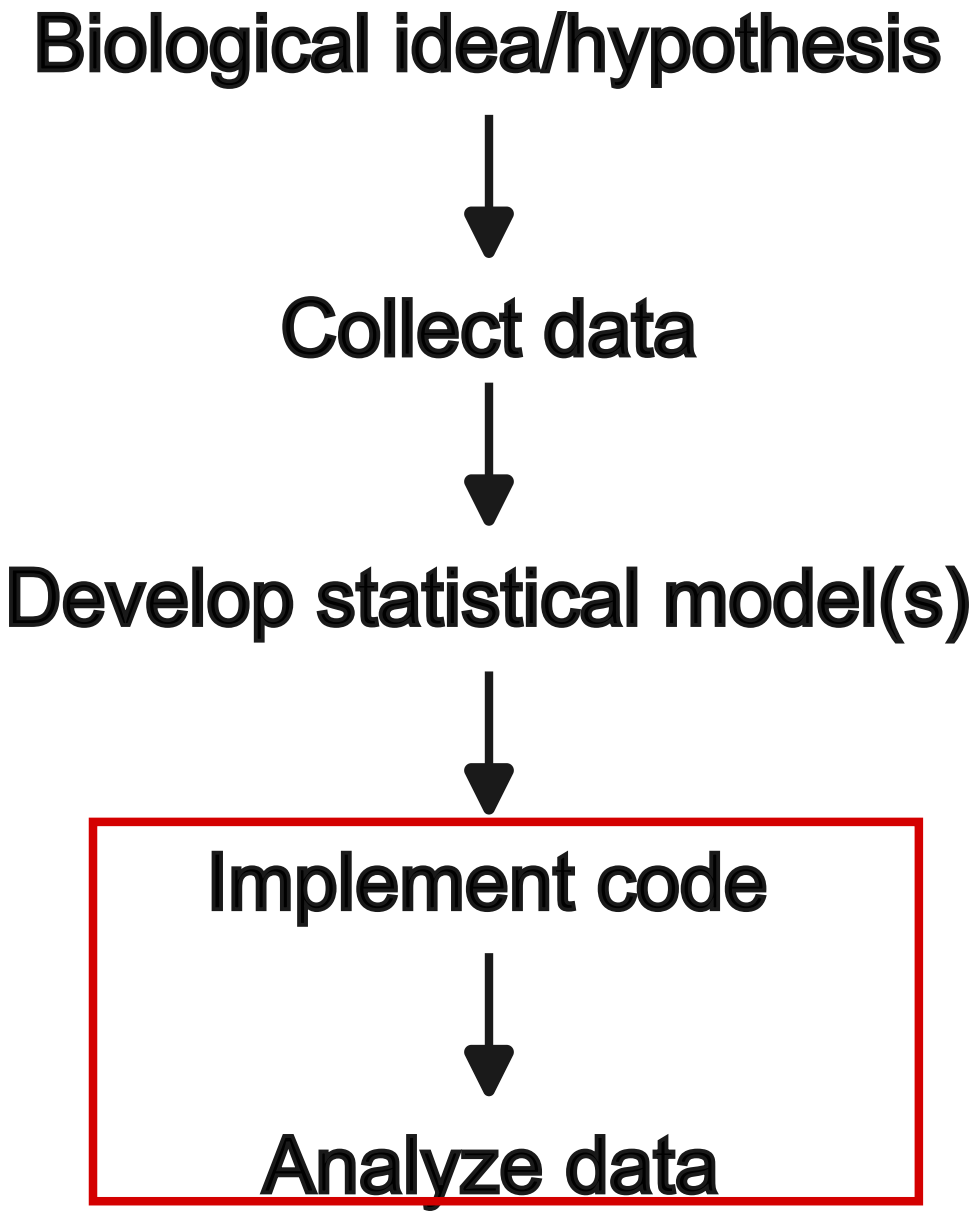
Computational Evo Bio

- Many projects have similar flow



Computational Evo Bio

- This course will focus on the last two steps



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 = []
                                                    9,0-1         Top
```

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