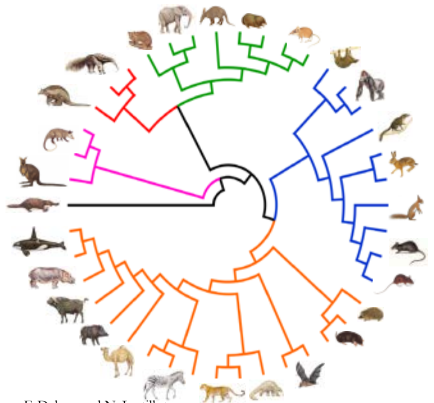
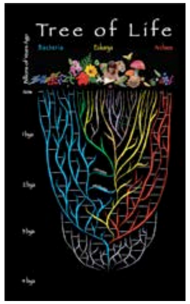
Darwin's regret: what maths tells us about the evolution of life



from F. Delsuc and N. Lartillot



Mike Steel

Biomathematics Research Centre

AWC public lecture

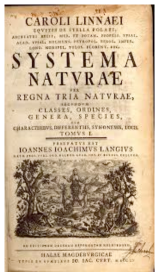
ALLAN WILSON CENTRE

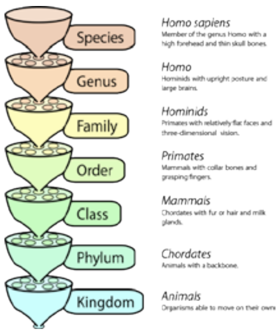
Talk outline

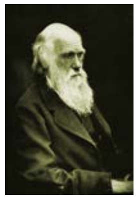
- Why trees?
- Why maths?
- What are some problems?
- Probability models

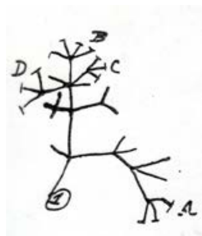
Pre-Darwin

Carl Linnaeus 1701-1778

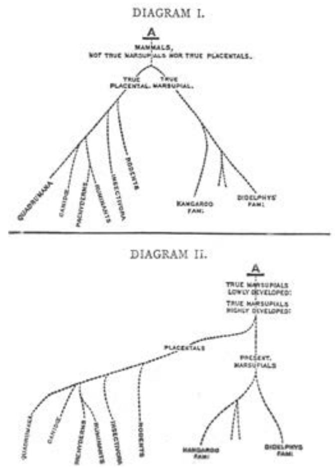




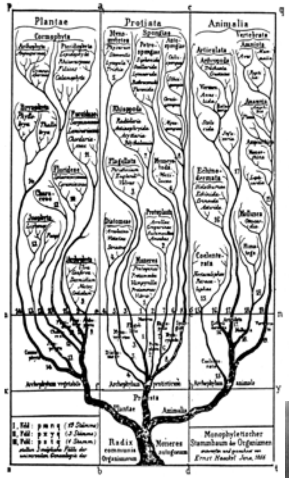


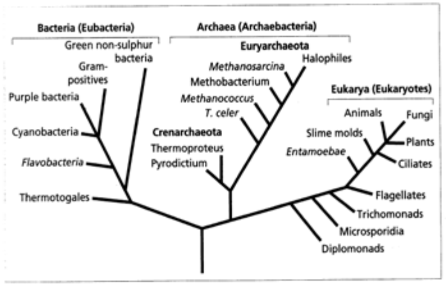


First Notebook on Transmutation of Species, 1837.



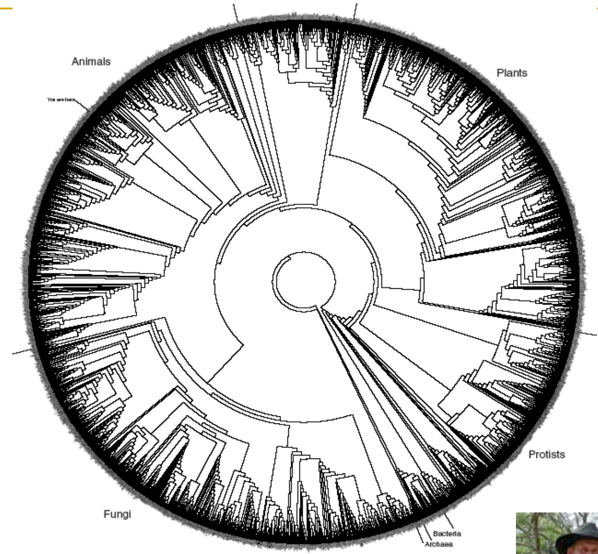
Letter from Darwin to Lyell, 1860.



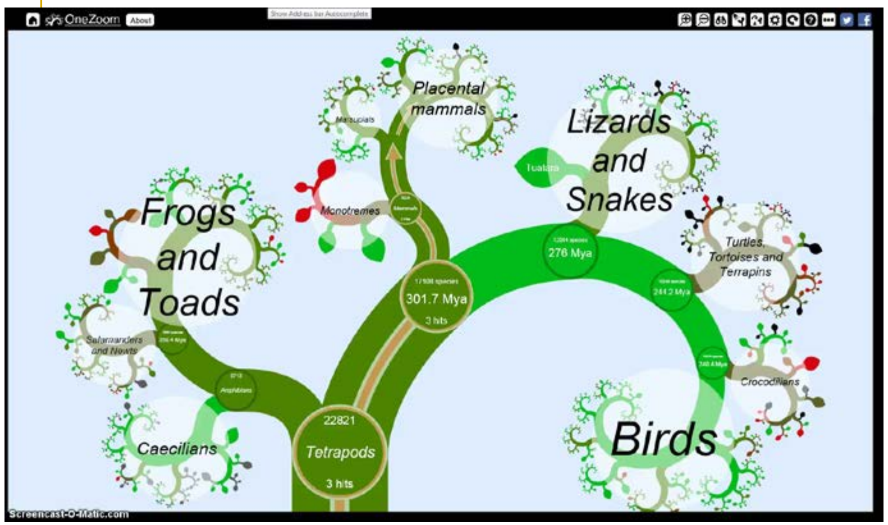


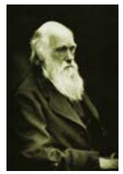
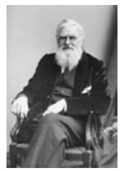
Ernst Haeckel (1866)

Olsen and Woese (1993)



David Hillis lab ~3000 species rRNA sequences



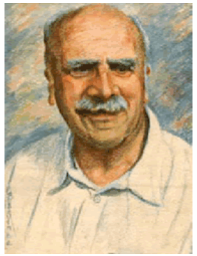
evolution

process **mechanism**

mechanism – *the theory*

Sir Ronald Aylmer Fisher FRS (1890 –1962)

Sewall Green Wright (1889 – 1988)



J.B.S ('Jack') Haldane [1892-1964]

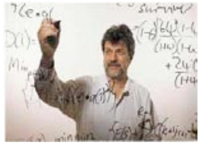
George Robert Price (1922 –1975)

William Donald "Bill" Hamilton FRS (1936 -2000)

John Maynard Smith FRS (1920 –2004)

Process: *phylogenetic trees*

- 1967 Walter Fitch and Emil Margoliash constructed phylogenetic trees from cytochrome c sequences from vertebrates that agreed well with the vertebrate fossil record.



David Sankoff (mathematician)

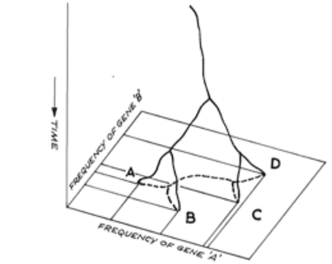
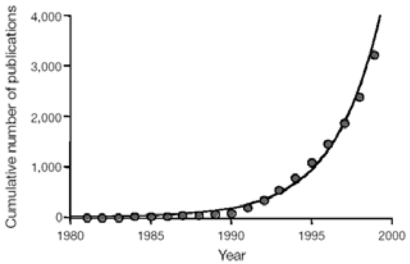
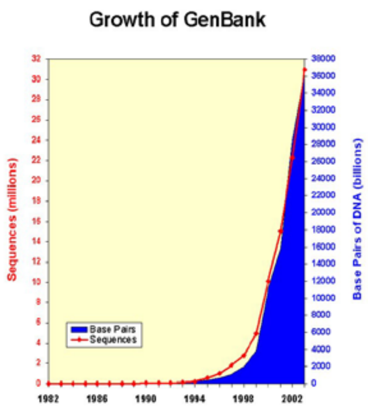


FIG. 1. An evolutionary tree and its projection onto the "now" plane.

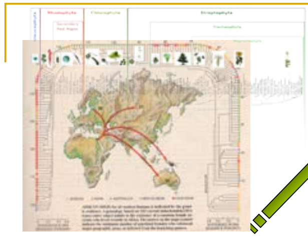
L.L. Cavalli-Sforza and Anthony Edwards (late 1960s)

The growth of genomics/phylogenetics

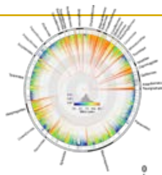




Publications with "molecular" and "phylogenetic" in abstract

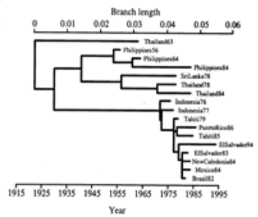


Systematic Biology



Epidemiology

+ medicine



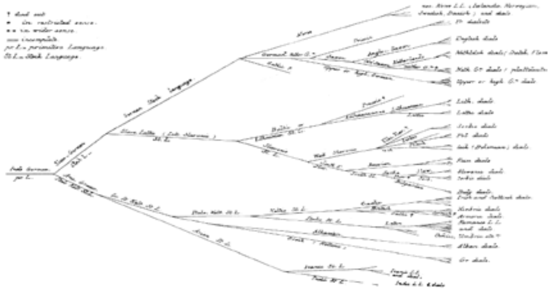
Phylogenetics

Evolutionary processes

Phylogeography, co-phylogeny, Biodiversity

Further applications of phylogenetics

linguistics, stematology,
psychology, whisky



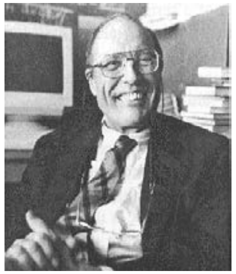
Stammbaum for Indo-Europena. From *Die Darwinshce Theorie und die Sprachwissenschaft* (Schleicher's 1863)

Figure 5: Cladogram for 46 manuscripts of the Wife of Bath's Prologue

2. Why maths?

"Unreasonable effectiveness of mathematics" in physics (1960).

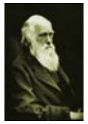
– Eugene Paul Wigner

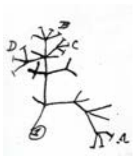


"The lack of real contact between mathematics and biology is either a tragedy, a scandal or a challenge, it is hard to decide which."

– Gian-Carlo Rota, (1986, in *Discrete thoughts*)

Why maths?

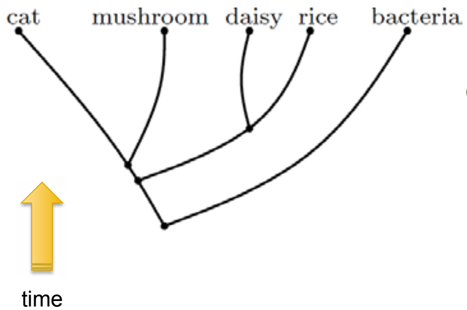




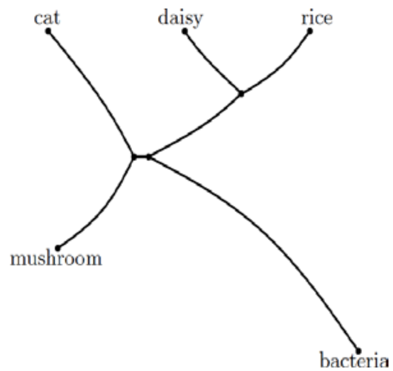
"I hope it arises from your being 10 fathoms deep in the **Mathematics**, & if you are God help you, for so am I, only with this difference: **I stick fast in the mud at the bottom and there I shall remain** "
– C. Darwin to Rev. William .D. Fox 29 July, 1828

- *Analysing existing methods*
- Developing better methods
- Help answer questions:
  - Why do some methods lead to different estimated trees?
  - How can we have confidence in a given tree? (or *any* tree?)
  - Is evolution really 'tree-like'?

First things about trees

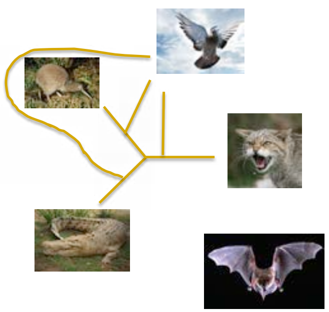


Rooted (binary) tree



Unrooted (binary) tree

Counting trees



Erwin Schroeder, 1870

Species
Kangaroo
Chimpanzee
Human
Gorilla
Hippopotamus
Whale
Lion
Tiger

trees = 1x 3 x 5

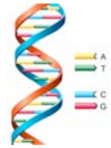
trees = 1x 3 x 5 x7 x 9 x11 = 10,395 trees

trees for 60 species is more than the number of atoms in the known universe

How do biologists build trees?

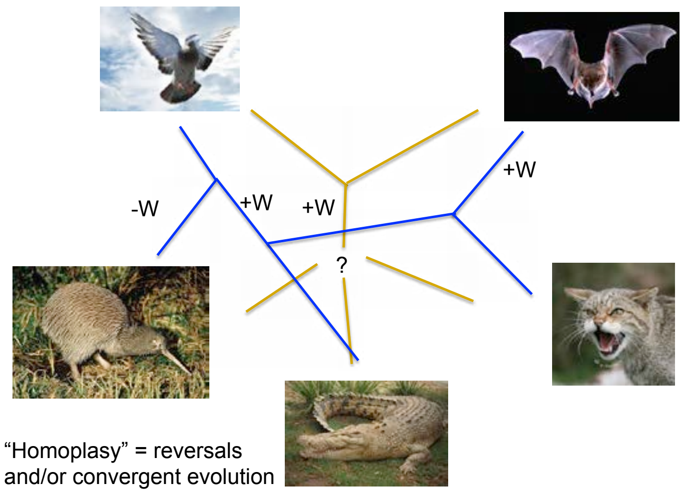
Old days (pre 1970s)
- Fossils
- Morphology, behavior, physiology

1970s onwards
- Amino acid/DNA sequences
- Gene order on chromosomes
- Presence/absence of genes/markers



Basic Idea: Compare species for difference in 'characteristics' they share.

Signal in data (and why it can be misleading…)

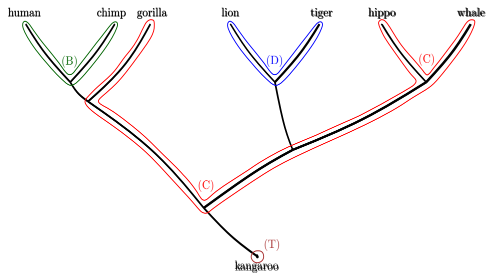


"Homoplasy" = reversals and/or convergent evolution

Life without homoplasy

10,395 trees

Species	Attribute 1	2	3	4
Kangaroo	T	R	U	E
Chimpanzee	B	R	E	T
Human	B	R	O	E
Gorilla	C	O	E	E
Hippopotamus	C	A	P	O
Whale	C	A	U	P
Lion	D	R	A	O
Tiger	D	R	U	G



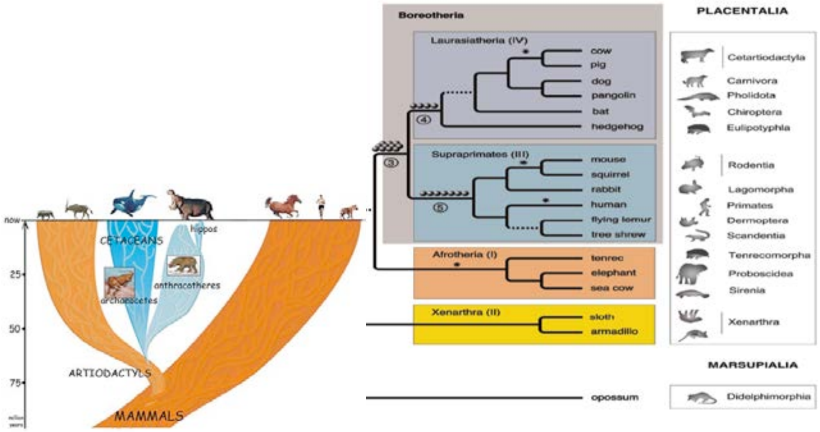
A 'perfect phylogeny'

… and the only one

The 'four is enough' theorem

Every tree – on any number of species – is the *only* perfect phylogeny for some choice of *four* characters

Example of low-homoplasy data (SINEs)



[Kreigs *et al.* PLoS biology, 2006. Tree of placental mammals]

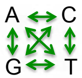
Part 3: What are some problems?

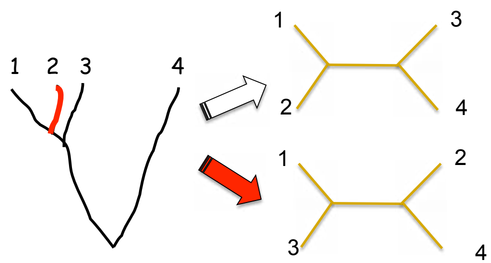
- The 'Felsenstein Zone'
- Lineage sorting
- Deep divergences – site saturation

The 'Felsenstein Zone'

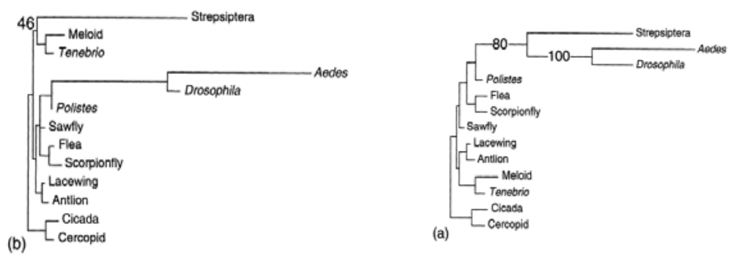
Joseph Felsenstein

Species 1:….ACAACGT….
Species 2…. ACGACTC….



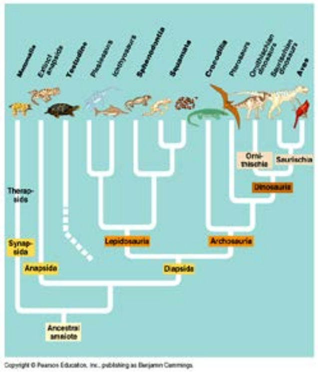


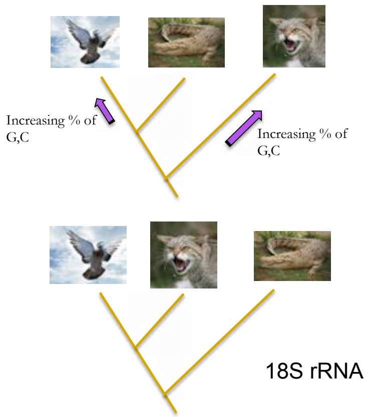
Does it happen?



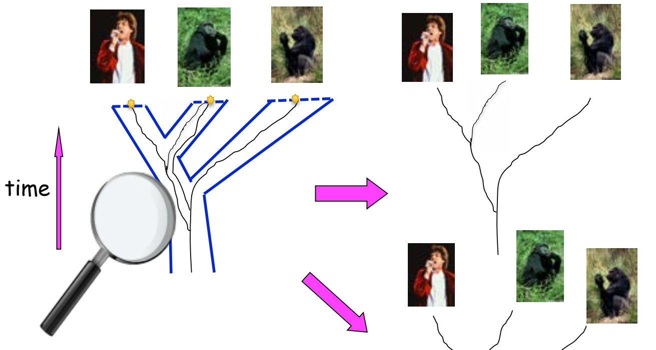
J. Huelsenbeck 1998: Is the Felsenstein Zone a fly trap?

A different type of 'zone'



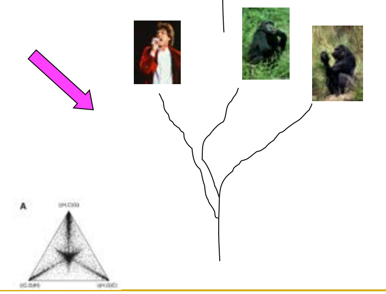


Problem 2: lineage sorting: 'gene trees vs. species trees'



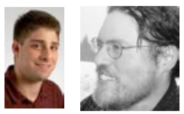
The good (math) news….

(HC)G ~78%
(GC)H ~11% (HG)C ~11%



(Ebersberger et al. MBE 2007)

The plot thickens…



Theorem 2006 Noah Rosenberg and James Degnan

Whenever you have **five or more** species, the most likely gene tree can be different from the species tree

Why bad?

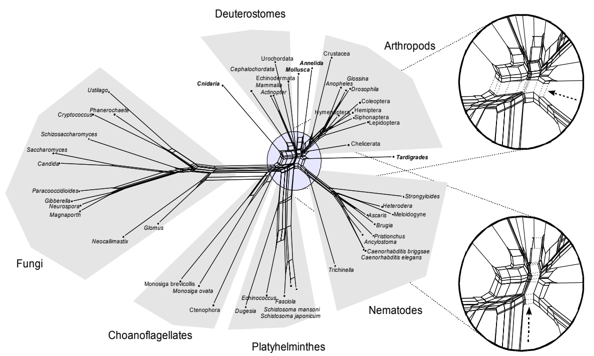
But…

"Theorem" ~2008

There are simple ways that will find the species tree from gene trees that don't get fooled!

Problem 3: Deep divergences

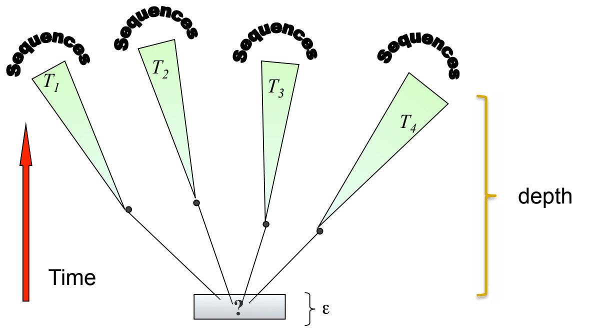
(the Metazoan phylogeny)



From Huson and Bryant, Applications of phylogenetic networks in evolutionary studies, MBE. 2006

Coelomata vs Ecdysozoa

Deep divergences



"Theorem":
The amount of data needed to resolve ? is at least $\frac{e^{depth}}{\varepsilon^2}$

Does it matter?

Part 4: Probability models

- The 'shape' of trees
- Predicting the loss of 'evolutionary heritage'
- Where is the root of a tree?

An early example…

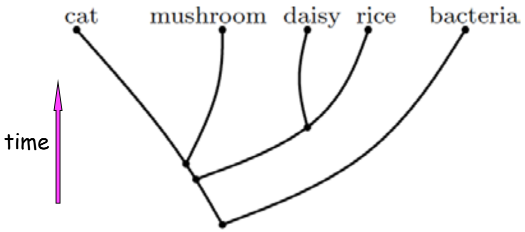
Number of species in genus	Number of genera: Observed	Calculated
1	131	130.9
2	35	47.2
3	28	25.2
4	17	16.0
5	16	11.2
6	9	8.3
7	8	6.5
8	8	5.2
9 to 11	13	11.1
12 to 14	3	7.2
15 to 20	7	8.8
21 to 34	14	9.2
35 upwards	4	6.2
Total	293	293.0

$$Pr(N = n) \approx \frac{1}{n^{1+r}}$$

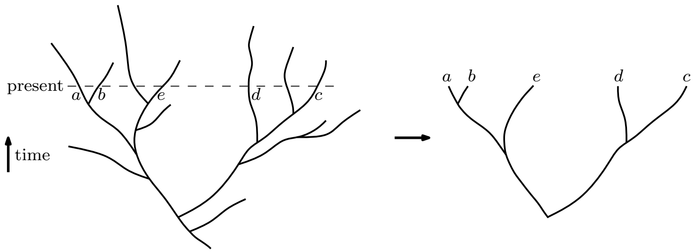
[1]G. U. Yule, A mathematical theory of evolution. Based on the Conclusions of Dr. J.C. Willis, F.R.S. Phil. Trans. Roy. Soc. 213 (1925), 21-87.

Where do evolutionary trees comes from?


Forestry Unit men tree-felling in Southern Italy



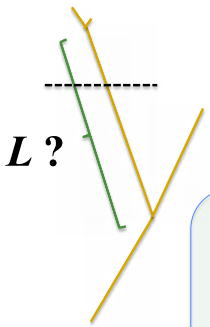
Another viewpoint



Evolutionary tree

Reconstructed tree

Question 1: How long are the branches in a tree?



Speciation rate = 1/million years

so the expected value of L equals 1 million years

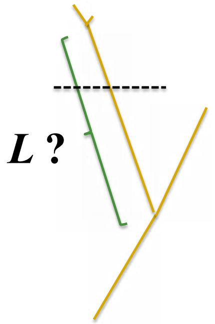
The bus 'paradox'

"It wouldn't hurt to wait around for a little while."

You turn up at a bus stop, with no idea when the next bus will arrive.

- If buses arrive regularly every 20 mins what is your expected waiting time?
- If buses arrive randomly every 20 mins what is your expected waiting time?

Length of a new branch

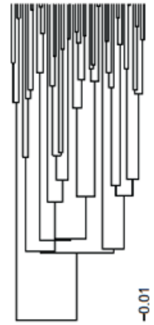


"It wouldn't hurt to wait around for a little while."

Expected value of *L* is 1 million years

Tree puzzle (I):

A tree evolves with each lineage randomly generating a new lineage on average once every **1 million years** (no extinction).



Look at the tree when it has 100 tips

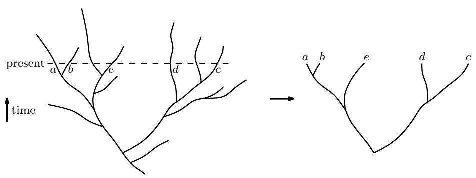
What is the expected length of a randomly selected *extant* branch?

Answer 1: 1 million years?

Answer 2: 500,000 years?

What about the *interior* lineages?

Tree puzzle (II):

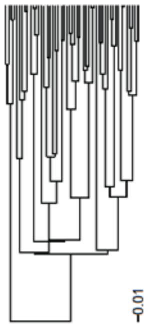


Now suppose extinction occurs at the same rate as speciation.

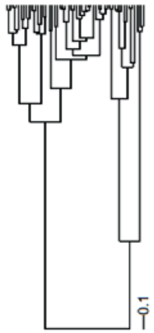
Suppose the 'reconstructed tree' we see today has 100 tips.

Does this tree look just like a tree grown with a slower speciation rate and zero extinction?

No



(A) Extinction=0



(B) Extinction = speciation

Real reconstructed trees generally look much more Model A than B trees

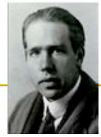
(Morlon et al. 2010; McPeek 2008)

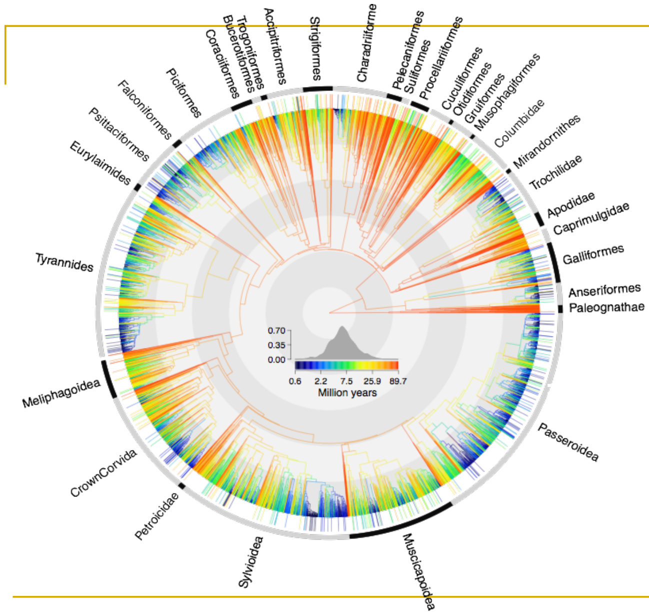
Why does this matter?

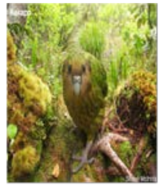
Question:

If a random 10% of species from some clade were to disappear in the next 100 years due to current high rates of extinction, how much 'evolutionary diversity' would be lost?

Prediction is very difficult, especially about the future. Niels Bohr, Danish physicist (1885-1962)

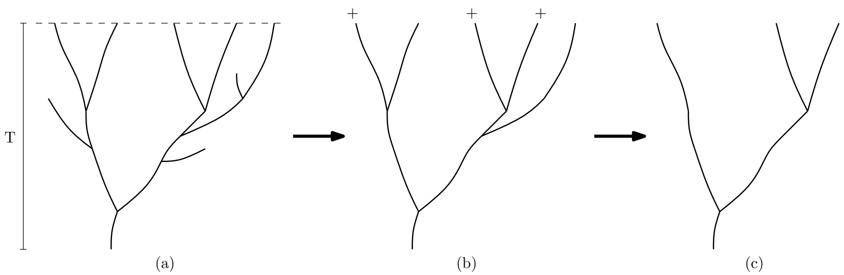






2014 'Tree of birds' ~(10,000 species) with ages colour coded (and with distribution of 575 impelled species at tips (rep. 2.7 billions years of evolution) [Jetz. et al. 2014]

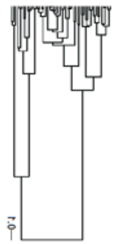
Diversity

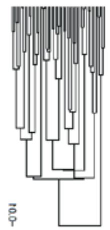


Predict the proportion of diversity that remains if each species (leaf) survives with probability p (independently of others).

A much cited paper

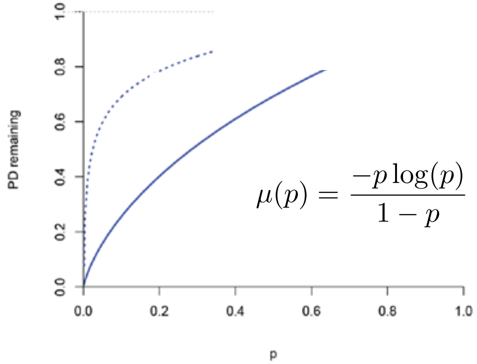
"...80 percent of the underlying tree can survive even when approximately 95 percent of species are lost."
Nee and May, ***Science***, 1997





$$\mu(p) = \frac{\text{Expected future diversity}}{\text{Expected present diversity}}$$

$$\mu(p) = \frac{-p\log(p)}{1-p}$$

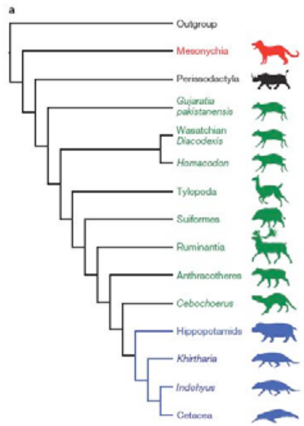


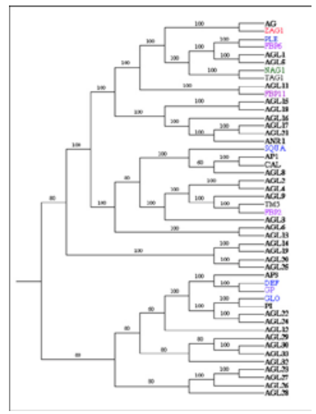
$$\mu(p) = \frac{-p\log(p)}{1-p}$$

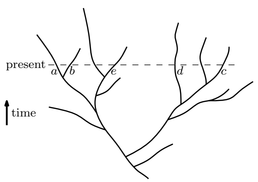
"…80 percent of the underlying tree can survive even when approximately 95 percent of species are lost."

"…84 percent of the underlying tree **is lost** when approximately 95 percent of species are lost."

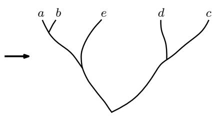
Question 2: Can we explain tree balance?



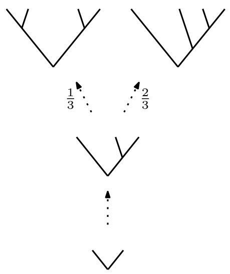




Evolutionary tree

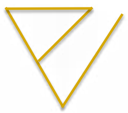


Reconstructed tree

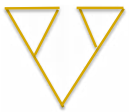


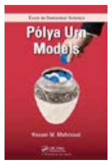
Yule model

Grow a Yule tree till it has 101 leaves.
Which is more likely?



OR

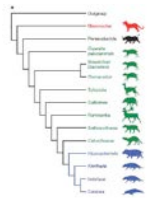




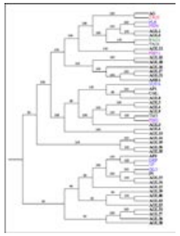
$$\frac{2}{3}\times\frac{3}{4}\times\frac{4}{5}\times\ldots\times\frac{99}{100}=\frac{2}{100}$$

Balance:

Yule model	1/3	2/3
All trees equal model	3/15 = 1/5	12/15=4/5



Real trees



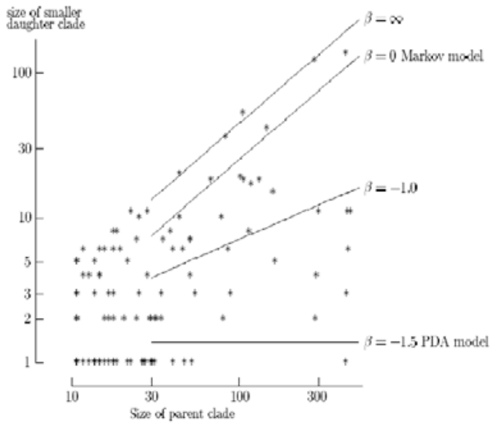
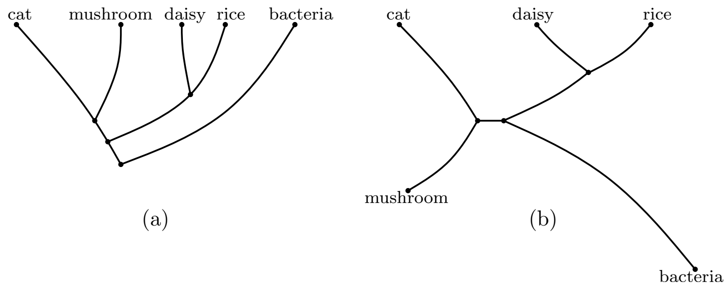


FIG. 3. *Splits in the tree of Chase et al (1993), and approximate median lines for the beta-splitting model. Note the log-log scale.*

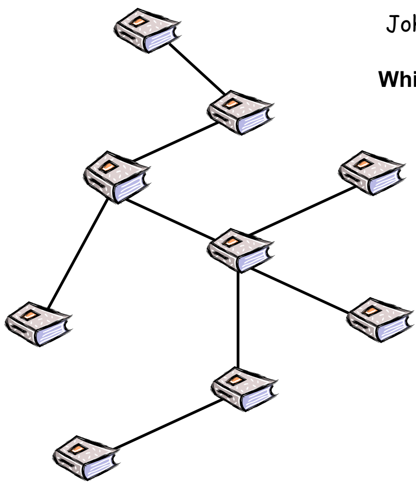
From: Aldous, D. (2001). Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today. *Statistical Science* 16: 23-34

Question 4: Where's the tree root?



Does the Yule model provide any clue?

A different (but related) question



John Haigh 1970

Which was the original version?

Select the 'most likely' one. What's the probability it is the original?

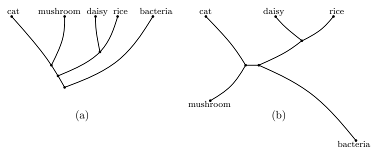
$$1 - \log(2) \sim .307$$

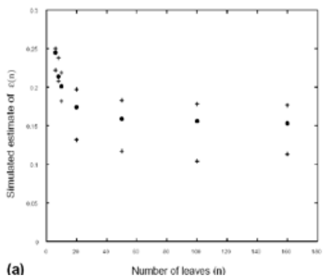
Nine nodes: probability is >99.6%

~Independent of *n*

Back to our trees

$$1 - \log(2) \sim .307$$





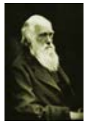
Theorem (McKenzie+S)

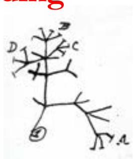
$$\Pr(e_{ML} = e_0) = 4\log(^4/_3) - 1$$
$$\sim 0.15$$

P(longer of longer < shorter)

Why maths? (again)…

I attempted mathematics, and even went during the summer of 1828 with a private tutor (a very dull man) to Barmouth, but I got on very slowly…. **This impatience was very foolish, and in later years I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics.**





Some questions for the future

- How did life start?
- How did photosynthesis evolve?
- How much were genes transfered in early life?