Inference, Models and Simulation for Complex Systems — CSCI 7000-001
Lectures 8 and 9 — 22 and 27 September 2011
Prof. Aaron Clauset

1 Random walk models of evolution

Evolution is the natural mechanism that produces biological diversity. The two basic components of evolution are (i) descent with variation, meaning that organisms produce offspring that are similar to but not exactly the same as themselves, and (ii) natural selection, meaning that on average organisms that are better adapted to their environment produce more offspring than those less well adapted.[1] Today, much of the discussion of "evolution" in the scientific literature focuses on understanding its molecular mechanisms and dynamics, that is, the information encoded in DNA, the way variations in that information arise and the impact they have on biological function, and the way molecular mechanics can shed light on natural selection, e.g., by looking at bits of DNA that appear in species separated by large amounts of evolutionary time.

Note that the first assumption of evolution (descent with variation) is very much like the notion of serial correlation that we encountered last week, in which $x_{t+1} = x_t + \lambda$. If we let x_t represent the particular "state" of an organism (e.g., its genome and various epigenetic factors) at time t, its descendent state x_{t+1} is simply its parent's state plus whatever random variation occurs during the reproduction process. From this perspective, a chain of ancestors and descendants—a lineage—can be viewed as a random walk within some state space.[2] If we omit selective effects, we call the evolutionary dynamics *neutral* because the variations we observe as some point in time are due largely to diffusion effects rather than selective effects.[3]

Macroevolution is a branch of evolution theory that focuses on the patterns of evolution at large spatial or temporal scales, for instance, evolution over millions of years (or more!) or evolution of entire species or clades.[4] There is no clean distinction between "microevolution" and "macroevolu-

[1] "Well adapted" is an ill-defined term and is often inferred by looking at what happened historically. The reason being that the "fitness" of one organism is not defined independently. That is, fitness is not an inherent characteristic of an organism, but rather implicitly defined by the internal structure and dynamics of the organism and its various interactions with its environment, which almost always includes other organisms.

[2] In genetics, the state space is the set of all DNA strings with n base pairs. The usual assumption is that all strings are equally likely and thus the random walk is unbiased within the n-dimensional hypercube. This is not always an accurate assumption.

[3] "Neutral" models of evolution are most commonly used as null models. Rejecting such a model via a standard hypothesis test is usually interpreted as evidence of selection. However, other effects, such as physical constraints on the space of possible states confound this conclusion. More recently, researchers have begun exploring neutral models of evolution at large scales and have found surprisingly good agreement with some empirical data.

[4] In evolutionary theory, the notion of a *phylogenetic tree* is fundamental. Mathematically, these are equivalent to the discrete object called a "tree". A "clade" is just a subtree. There's some tension between the classic taxonomic

tion," although a rough rule-of-thumb is that microevolution focuses on variation below the level of species, while macroevolution focuses on variation at or above the level of species. In microevolution, the "unit of selection" is an individual organism or genes within organisms. In macroevolution, the unit of selection is an entire species or clade.

Macroevolution is generally the domain of paleontologists (fossils!) and molecular biologists (DNA "archeology"). The types of questions these scientists investigate often focus on topics like body size or mass, metabolism and its variation across the tree of life, species diversity (defined broadly: number of species, morphological variation, taxonomy, genetics, behavior, etc.), complexity (for instance, the major transitions in evolution: prokaryote to eukaryote to multi-cellularity to sociality), carrying capacities, mass extinctions and recoveries, and adaptive radiations.

1.1 Geological time

Before moving on, it will be useful to have an intuitive sense of the raw timescales of macroevolution. For instance, Figure 10 shows the "geological clock" of the Earth as a radial diagram. The general timescale is enormous, with most of the 4 billion years being dominated by microbial evolution. Animals, in all their resplendent complexity and their charisma, are comparatively recent arrivals, appearing about 530 million years ago in the Cambrian explosion; most of evolutionary time is dominated by microbial species. Unfortunately, microbes (and many animals) don't leave nice fossils and we must use molecular techniques to dig around in DNA to make inferences about what microbes were like in the past.

2 Cladogenetic models of mammalian macroevolution

One place where random-walk models of macroevolution have been particularly successful is in modeling the morphological variation of animals, and in particular, the types of animals that leave good fossils.[5] The fossil data allows us to extract estimates of ancestral states and thereby construct time series that can be analyzed using random walk models. The simplest and most universal of all morphological variables is the animal's body size or mass, and this attribute is one of the easiest to recover or estimate empirically.[6] Moreover, we can measure the current distribution for living

structures (kingdoms, phyla, orders, families, etc.) and the notion of a phylogenetic tree or a "tree of life," but generally each clade in the tree corresponds to some taxonomic group. For instance, *Mammalia* is the subtree that contains all descendant individuals or species from the last common ancestor that possessed the defining characteristics of mammals (in this case, possession of hair, three middle ear bones, and mammary glands functional in mothers with young).

[5]These are mainly the vertebrates, whose calcified bones preserve well, and species with hard shells. Mammals are especially good because their teeth, which fossilize very well, reveal much about them. Soft-bodied species like sharks (cartilage doesn't fossilize well) and worms are much harder to work with.

[6]Ecologists, studying extant or living species, have shown that body size correlates with a large number of other species attributes, including lifespan, metabolic rate, range size, reproductive rate, etc., which means that you can

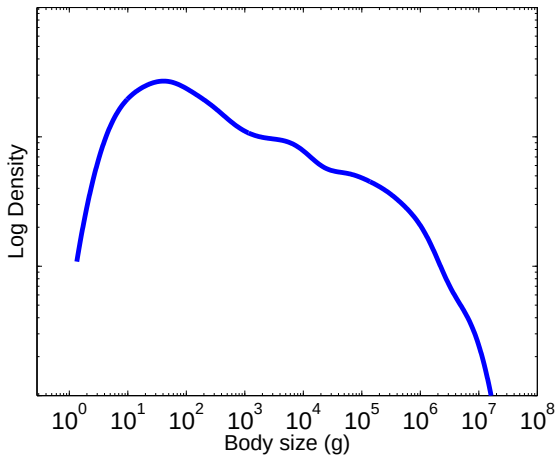

Figure 1: The distribution of 4002 living (extant) terrestrial mammal body sizes, smoothed using a Gaussian kernel. (Data from Felisa A. Smith et al.) Density smoothing is a nice non-parametric way to avoid choosing an arbitrary binning scheme in order to visualize the pdf of your empirical data; the cost of smoothing, however, is a loss of absolute scale for the densities (note the missing scale on the y-axis).

(extant) species, which is shown in Figure 1.

To construct a macroevolutionary random walk model, we first need to specify a few basic properties of the process. First, species body size tends to vary multiplicatively from ancestor to descendent—that is, proportionally, rather than additively—and thus we will use a multiplicative random walk model. Second, to mimic the way individual species can give rise to new species, we will allow individual walkers to generate new random walkers (descendants) whose initial state is equal to the ancestor's state at the point of reproduction. In the parlance of random walks, this is a *branching* process; in the parlance of macroevolution, we call it a "cladogenetic" random walk to represent the fact that new clades are being generated. Before describing more of the model, let's briefly examine branching random walks.

2.1 Branching random walks

Recall that the simplest kind of random walk has the form

$$x_{t+1} = x_t + \lambda \ , \tag{1}$$

learn a great deal about a species by simply estimating its body size.

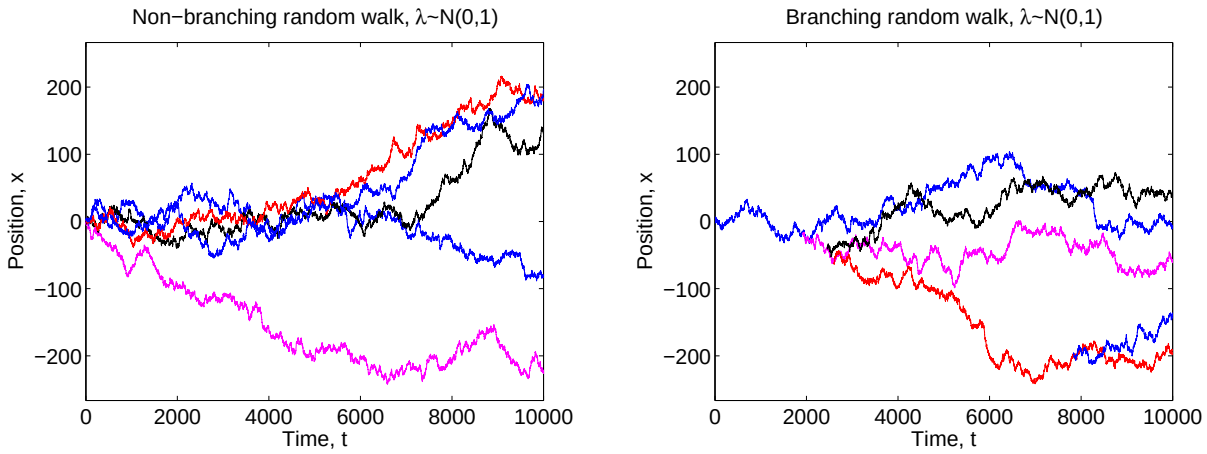

Figure 2: (a) Time series showing $T = 10^4$ steps of $N = 5$ random walkers with $\lambda \sim N(0, 1)$ and $x_0 = 0$. (b) The same, but where initially we have one random walker and at each time step with probability $p = N/T$ we allow one of the existing random walkers to branch.

where λ is a random variable drawn iid from a distribution $\Pr(\lambda)$ and where x_0 is the initial position of the walker. In the conventional version, this equation describes the trajectory of a single walker.

In a branching process, the number of walkers can increase over time; the dynamics are still the same, but the choice of x_0 for the new walker is made in such a way as to imitate a process in which one walker splits into two. Suppose that for some random walker A, a branch event happens at time t_{branch}, when A is located at $x^A_{t_{\text{branch}}}$, producing a new random walker B. From the perspective of B, its random walk has just begun and so we set its initial location to A's current location: $x^B_0 = x^A_{t_{\text{branch}}}$. If we draw λ values from $\Pr(\lambda)$ independently for A and B, their subsequent trajectories will evolve independently. In this case, their overall histories are correlated up to t_{branch} but completely independent thereafter.

Figure 2 shows the simulated trajectories of 5 independent random walkers (with $\lambda \sim N(0, 1)$ and $x_0 = 0$) and 5 branching random walkers (same parameters). Matlab code for this simple simulation is given at the end of these notes, in Section 5.1. (Should the population variance be the same, less or greater in the branching case versus the independent case? In what other ways is this system similar or dissimilar to a set of independent, non-branching random walkers?)

We will use a model very much like this one to represent body size variation on macroevolutionary timescales. Above, every species moves at every time step, regardless of whether it speciates (branches) or not. In evolutionary theory, this kind of dynamic, in which the "size" of a species

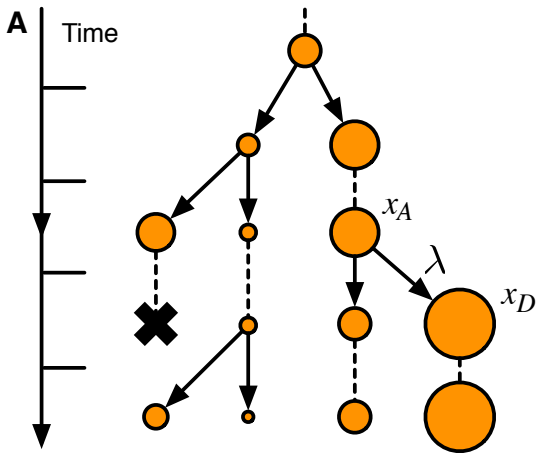

Figure 3: A schematic of a discrete-time, pure cladogenetic model of body size variation. At each time step, a single species is chosen uniformly at random and branched. Its descendants each take a step drawn from the diffusion kernel. Additionally, at each time step, each species becomes extinct (is removed from the population) with some probability (represented by the big "X").

changes between branch events, is called *anagensis*. If instead we only allow species to change their sizes at branch events, we have a pure *cladogenesis* model. The American paleontologist Stephen J. Gould called this kind of behavior "punctuated equilibrium," after the idea that our definition of a "species" is inherently based on the notion of stasis or a lack of change. The typical pattern observed in the fossil record is one of long periods of stasis that are "punctuated" by brief periods of rapid change, followed by more stasis. In fact, the differences between anagenesis and pure cladogenesis are partly an illusion obscured by the vast time and size scales being considered. A process that looks like pure cladogenesis on the multi-million-year time scale may look more like anagenesis on the thousand-year time scale, especially given the fact that the observed data are actually a sampling of the states rather than continuous monitoring.[7]

In our model of a branching random walk, we have only allowed the creation (birth) of new walkers. *Extinction* (death) is the process by which species disappear from the population. If we allow random walkers to disappear with some probability, we have what's called a "birth-death" process, which can exhibit an equilibrium or steady-state population level when the average birth rate equals

[7]Perhaps unsurprisingly, this confusion has led to great and largely useless debates in the macroevolutionary literature about whether macroevolution is best modeled as anagenesis plus cladogenesis or as cladogenesis alone. In fact, pure cladogenetic models are largely equivalent to anagenesis plus cladogenesis models as we can wrap up all the anagenetic variation that happens to a species between branch points by simply increasing the variance in the distribution of changes that occur at a speciation event.

the average death rate. Figure 3 shows an schematic of a pure cladogenetic diffusion process with extinction. This will be our basic model of macroevolution and species body size changes.

2.2 Constraints on body-size diffusion

By balancing the speciation and extinction rates at the population level, we can induce an equilibrium in the number of species, such that it fluctuates around some particular value. But, these constraints say nothing about the positions of the walkers. In the previous example, the branching process is unbiased and unbounded, meaning that with non-zero probability a walker will escape to $+\infty$ or $-\infty$, which is not realistic behavior for biological evolution.[8] [9]

By introducing constraints on the diffusion process—that is, by restricting in some way the range of positions a walker can take—we can induce a stationary or equilibrium probability distribution over some region of the state space. If we choose the constraints properly, we can force the equilibrium distribution to have a particular shape. Choosing constraints in order to produce a specific target distribution is akin to fitting a model to the data and thus should be done carefully. A principled approach to designing the diffusion model would be to first consider reasonable biological or physical constraints on body-size variation and then test whether, when added to the diffusion model, they produce realistic results.

To be concrete, we'll investigate a constrained cladogenetic diffusion model (with extinction) for the way mammal species body size (mass) varies.[10] In principle, however, the choice of mammals is arbitrary and this model could be adapted or applied to any clade of species, e.g., fish, birds, lizards, dinosaurs, etc.[11]

The basic model has three main components, which we'll investigate in turn:

1. Diffusion of species body sizes, in which variation is governed by a simple multiplicative

[8] Recall from Lectures 6 and 7 the examples of a finite population whose variance grew monotonically with time.

[9] The dimensionality of the state space also plays an important role here. For $d > 2$, the probability that a walker will escape to infinity converges on a non-zero constant. The intuitive explanation for this weird behavior—that random walks in 3 dimensions tend to escape to infinity—is that as we increase the number of dimensions, a greater fraction of the random steps take us, on average, away from the point of origin. What's surprising here is that the dimensionality does not have to be large in order to produce this behavior. For biology, variation occurs along many many directly, and thus an unconstrained branching random walk will produce wildly divergent clusters of walkers over finite periods of time.

[10] The term "body size" is somewhat ambiguous, and is used to mean either the mass or the volume. Because the densities of flesh tend not to vary much, even across large evolutionary distances on the tree of life, this assumed equivalence is not terrible.

[11] There are some subtleties in choosing the clade, of course. The group of species should be *monophyletic*, meaning that they all stem from the same common ancestor species, and their environmental constraints should be fundamentally the same. For instance, in studying mammals, we will omit the fully aquatic species (members of *Cetacea*, like dolphins and whales) because their environmental constraints are fundamentally different from terrestrial mammals.

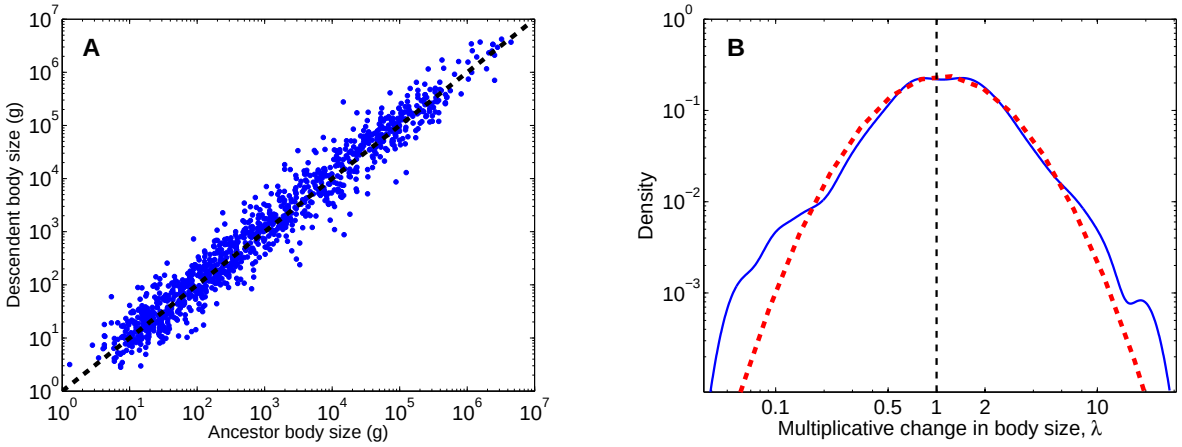

Figure 4: (a) A plot of estimated (x_t, x_{t-1}) pairs for North American mammal fossil species (data from John Alroy), showing a strong linear correlation, but with some scatter in both directions around the $y = x$ line. (b) The distribution of multiplicative residuals from (a), i.e., the distribution of ratios x_t/x_{t-1}, showing a nice wide diffusion kernel. The dashed line shows the maximum likelihood log-normal distribution.

random walk $x_t = \lambda\, x_{t-1}$ and where steps in the walk are taken only at branch points (as in punctuated equilibrium).

2. A fixed minimum size $x_{\min}$ allowed for a mammal species.

3. An extinction "gradient" in which the probability that a species becomes extinct increases with species size x.

2.2.1 Body size diffusion (or: fossils!)

If a random walk is a reasonable model for body-size variation, then fossils should show some evidence of a distribution of size changes $\Pr(\lambda)$, centered roughly at the "no change" value. It's possible (but mildly tedious) to extract estimates of species body sizes from the paleontological literature. Paleontologists are increasing interested in precisely these types of macroevolutionary questions, and have thus already started assembling databases of species-level information like their body size.

Figure 4 shows data from John Alroy's North American terrestrial mammal fossil database, which contains 1106 estimated changes in body size.[12] These changes are given in pairs (x_t, x_{t-1}), and

[12]Estimating these changes is itself a non-trivial inference problem. Alroy used a rough but non-phylogenetic ran-

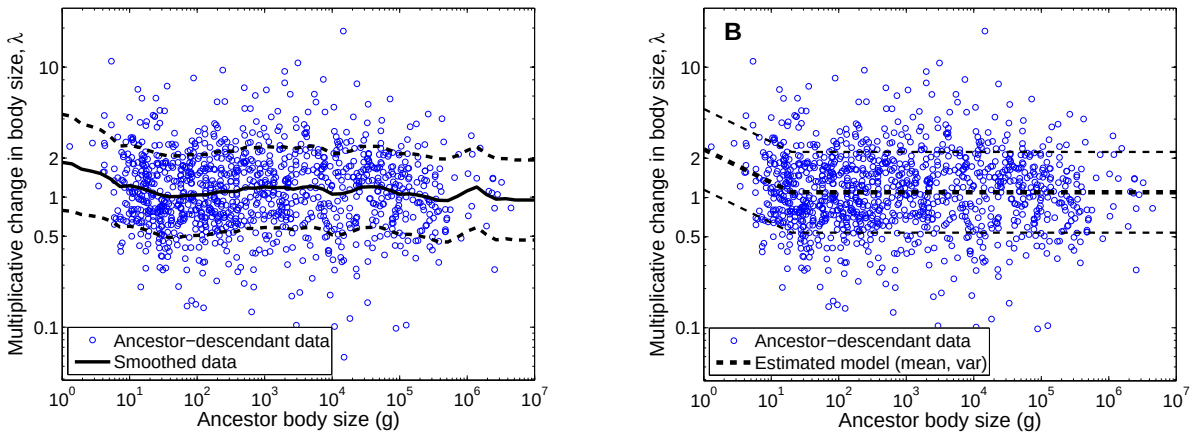

Figure 5: (a) Changes in size as a function of ancestor size, along with the mean (solid line) and variance (dashed) of the scatter, as a function of size (via an non-parametric smoother). (b) The same scatter, but now the model is a piece-wise linear parametric model of changes on a log-scale. The point here is to show that the non-parametric pattern is very similar to the parametric model we fit to the data.

because we assume a multiplicative random walk, the changes are simply the ratios x_t/x_{t-1}. Fig. 4a shows the raw scatter of the pairs and 4b shows the distribution of changes. These data thus allow us to directly estimate the structure of the diffusion "kernel" $\Pr(\lambda)$ from fossil data, e.g., modeled by a log-normal distribution with parameters μ and σ or as a log-normal "double Pareto" distribution (a log-normal distribution with power-law tails) with parameters μ, σ and α.[13]

We can go slightly further, however, and consider the fluctuations as a function of size. Figure 5 shows the scatter data, i.e., the points $(x_D/x_A, x_A)$, along with two types of analysis. On the left, we compute the mean and variance of the scatter as a function of ancestor size x_A using an exponential smoother.[14] On the right, we show the results of fitting via maximum likelihood a log-

domization technique for choosing plausible ancestor-descendent pairs, given only taxonomic and temporal ordering information. A more powerful technique would use inference on both the taxonomic structure *and* the size changes; that is, large size-changes should be less likely than small size changes, and thus when given a choice between two inferred phylogenetic trees, we should favor the one with smaller changes to size. This would be a good class project.

[13]Because we have no good reason to prefer or believe one of these distributional models over the other, we can simply try both and check if they behave similarly within the simulation. In fact, they mostly do, so the overall results do not depend sensitively on the choice of tail behavior.

[14]That is, the mean at a particular value x_A is the weighted average where the weight assigned to a point (λ_i, x_i) is an exponential function in $|x_i - x_A|$. This is a non-parametric way to extract the general variation in the mean fluctuation size.

normal distribution of the form $\mathrm{LN}(\mu(x_A), \sigma)$, i.e., the mean of the log-normal varies as a function of location x_A while the variance is constant. The structure of $\mu(x_A)$ is a piece-wise linear function on the log-scale and has the form:

$$\mu(x_A) = \begin{cases} (c_1/c_2)\log x_A + c_1 + \delta & \text{if } \log x_A < c_2, \\ \delta & \text{otherwise.} \end{cases} \tag{2}$$

That is, the function has two linear pieces, one which is flat above $\log x_A = c_2$, while below that point, the function can be sloped. In this case, we require that the two pieces meet at c_2, i.e., that the value returned by the upper condition equals the value returned by the lower condition at $x_A = c_2$.

Notably, the structure of this piecewise linear model is visually very similar to the results of the non-parametric smoother, which is reassuring. Furthermore, note that in both cases we see a slight up-tick in the mean change-in-size as $x_A \to x_{\min}$, which is exactly what we would expect for a hard lower limit on the size mammals can achieve.

2.2.2 A minimize size for mammals

Evidence for a minimum body size for mammals comes mainly from metabolic experiments and theory. The best evidence comes from a simple paper in *Science* in 1948, which showed that as the body mass of mammal species approaches 2 grams, the metabolic rate skyrockets. Figure 6 reprints the main finding of this paper.

We can interpret this empirical finding as support for the existence of a hard minimum size $x_{\min}$ in our diffusion model. Theoretically, this minimum is caused by the physical effect of heat loss and the need for mammals to maintain a stable internal body temperature. That is, as mass decreases, so do both surface area and volume, but the volume, which varies roughly like r^3 where r is half a body length, decreases faster than surface area, which varies roughly like r^2. Thus, if heat loss to the environment is roughly proportional to surface area, smaller mammals will lose a greater fraction of their energy to the environment due to radiative effects. For mammals, this is problematic because they are *endothermic*, i.e., their cellular machinery requires a constant internal temperature. From the modeling perspective, this implies that the smaller a mammal, the more oxygen and food it must consume in order to stay warm. Upper limits on consumption thus impose a lower limit on size.

2.2.3 Extinction gradients

For a cladogenetic process, the relevant parameter is not the absolute extinction rate but rather the ratio of the speciation (birth) rate to the extinction (death) rate as a function of body size. In the absence of the diffusion mechanism, when the speciation rate equals the extinction rate, then

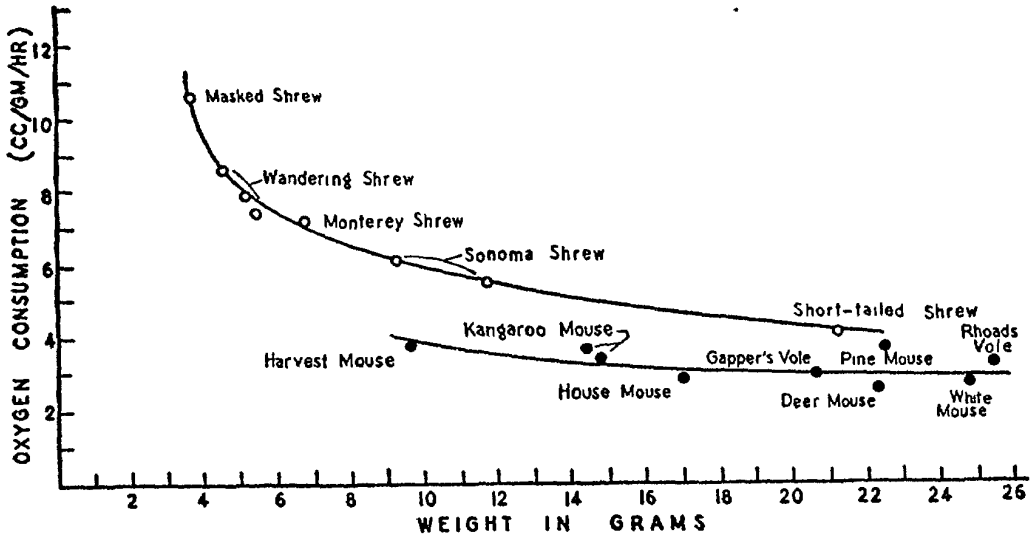

Figure 6: From Pearson (1948) in *Science*, the metabolic rate (measured as oxygen consumption) as a function of body size, for various small mammal species.

at equilibrium, the number of species at that size will remain roughly stable. (How does diffusion change this picture?) Thus, without loss of generality, we can fix the speciation rate for all species (note: we have already done this by assuming that at each time step there is exactly one speciation event) and only consider how the effective extinction rate might vary.

Evidence for extinction risk being positively correlated with species body size is largely indirect because extinction removes from the fossil record the species we're most interested in. Recent studies of the impact of human activities on extant mammal species makes it clear that larger body size is the most important predictor for whether a mammal species is threatened with extinction.[15] Even some studies of the fossil record suggest that the durations of fossil species is longer for smaller-bodied species than for larger ones.

Finally, there is a powerful theoretical argument for why we should expect larger species to become extinct more quickly than smaller species. Recall that question 3 on Problem Set 2 focuses on the "drunkard's random walk" in which a random walker is walking near a cliff. If the walker crosses the cliff, the walker "dies." Let the location of the walker be the number of individuals in a species

[15]The other factor is living close to human settlements, although with global climate change that factor is becoming less relevant.

and let changes to that number be year-to-year fluctuations in within-population birth and death rates. Clearly, the smaller the population, the more likely a series of "bad" reproductive years is to yield a population too small to reproduce. If larger species have smaller populations—and they do, just think about elephants versus rats—then within-species population dynamics provide a simple theoretical explanation for why extinction rates should increase with body size.

To capture this effect, at each time step, we let each species become extinct with a small probability that depends on both a "background" extinction rate and a size-dependent extinction rate:

$$p_{\text{extinct}}(x) = \rho \log x + \beta \ , \tag{3}$$

where ρ parameterizes the strength of the extinction gradient and β is the background extinction rate (which would hold even if there were no gradient).[16] We use the $\log x$ in order to make the gradient very gentle. In principle, ρ and β could both be estimated from fossil data on the duration of species, but instead we'll simply estimate ρ by fitting the model to the empirical data on extant species.

2.3 Putting the pieces together

Now that we've specified the three basic components of the model, we can combine them in a simulation to investigate the type of distributions of species body sizes they produce. Matlab code is given at the end of these notes, in section 5.2.

Before we can do this, however, we need to choose parameter values for the various processes. Table 1 gives the 12 parameters, the values chosen and the way they were chosen. It's important to note that a simulation with 12 parameters is fairly complicated. Fortunately, we estimate all but one of the parameters directly from fossil or extant data. The one parameter we were unable to estimate from data was the slope of the extinction risk curve ρ in Eq. (3).

Instead, we can choose ρ so as to make the model produce results that are similar to the empirical data. Were this a standard probabilistic model, we would use maximum likelihood techniques to choose its value. However, the model is complicated and the probability of observing a particular species body size x_i is not an iid random variable. In some cases, such correlated structures can be handled mathematically, but this is not one of them. An alternative approach to estimating ρ is to choose the value that minimizes the distance between the equilibrium distribution (at the

[16] Equation (3) is effectively an exponential increase function for the extinction risk curve. If we exchange the $p_{\text{extinct}}(x)$ on the left-hand side for $\log p_{\text{extinct}}(x)$, we instead have a power-law function. Because it's not clear which of these, or some other model, is most natural to use, we can use both and compare the results. As it turns out, when we choose ρ to make the overall model's behavior match the extant size distribution, the value chosen under the exponential model makes a functional shape almost exactly the same as the one chosen for the power-law model. In both cases, the ρ parameter is small but positive, indicating only a gradual increase to the extinction rates.

parameter		value	source
lower bound	$x_{\min}$	1.8g	fossil and extant data
baseline extinction rate	β	$1/n$	derived from other parameter
rate of extinction increase	ρ	0.025	fitted to extant data
$\log \lambda$-intercept	c_1	0.33	fossil data
$\log x$-intercept	c_2	1.30	fossil data
systematic bias	δ	0.04	fossil data
variance	σ	0.63	fossil data
power-law tail	α	3.3(1)	fossil data
founder body size	x_0	40g	modal size of extant data
species at equilibrium	n	5000	extant data
mean species lifetime	ν	1.60(1) My	fossil data
years in equilibrium	τ	60 My	fossil data

Table 1: Simulation parameters, their estimated values and the sources of those values. The parameters can be grouped according to mechanism: the physiological lower limit of the terrestrial mammalian body size ($x_{\min}$); the extinction risk function (β and ρ); the distribution $\Pr(\lambda)$ of within-lineage changes to body size (c_1, c_2, δ, σ and α), where δ denotes the systematic bias away from smaller body sizes (Cope's rule) and c_1 and c_2 denote the additional bias for small-bodied species; the initial conditions and duration of the simulation (x_0, τ, ν and n).

end of the simulation) and the empirical data. In this case, we use a variation on the classic Kolmogorov-Smirnov statistic that we used in Lecture 3. The tail-weighted Kolmogorov-Smirnov statistic reweights the classic measure to give more weight to deviations that occur in the left- or right-tails of the distribution, and is defined like this:

$$\text{wKS} = \max_x \frac{|S(x) - P(x)|}{\sqrt{P(x)[1 - P(x)]}} \ , \tag{4}$$

where $S(x)$ is the cdf of the simulated data and $P(x)$ is the cdf of the empirical data. We then simply choose $\hat{\rho}$ as the value that minimizes the wKS statistic over all possible values of ρ.

With the parameters chosen, we can now simulate the entire process. Figure 7 shows four snapshots from the simulation. We initialize the simulation with a single founder species with mass $x_0 = 40$g and then let the dynamics proceed as described above. Because an equilibrium exists under this model both for the number of species and their distribution of sizes, the particular choice of initial mass does not matter, so long as we run the simulation to convergence on the equilibrium state. As the simulation progresses, it first fills out the left-side of the distribution, where species range from 2g up to about 100g. This is a transient effect cause by seeding the simulation with a founder size close to the left boundary. The agreement between the simulated distribution and the empirical

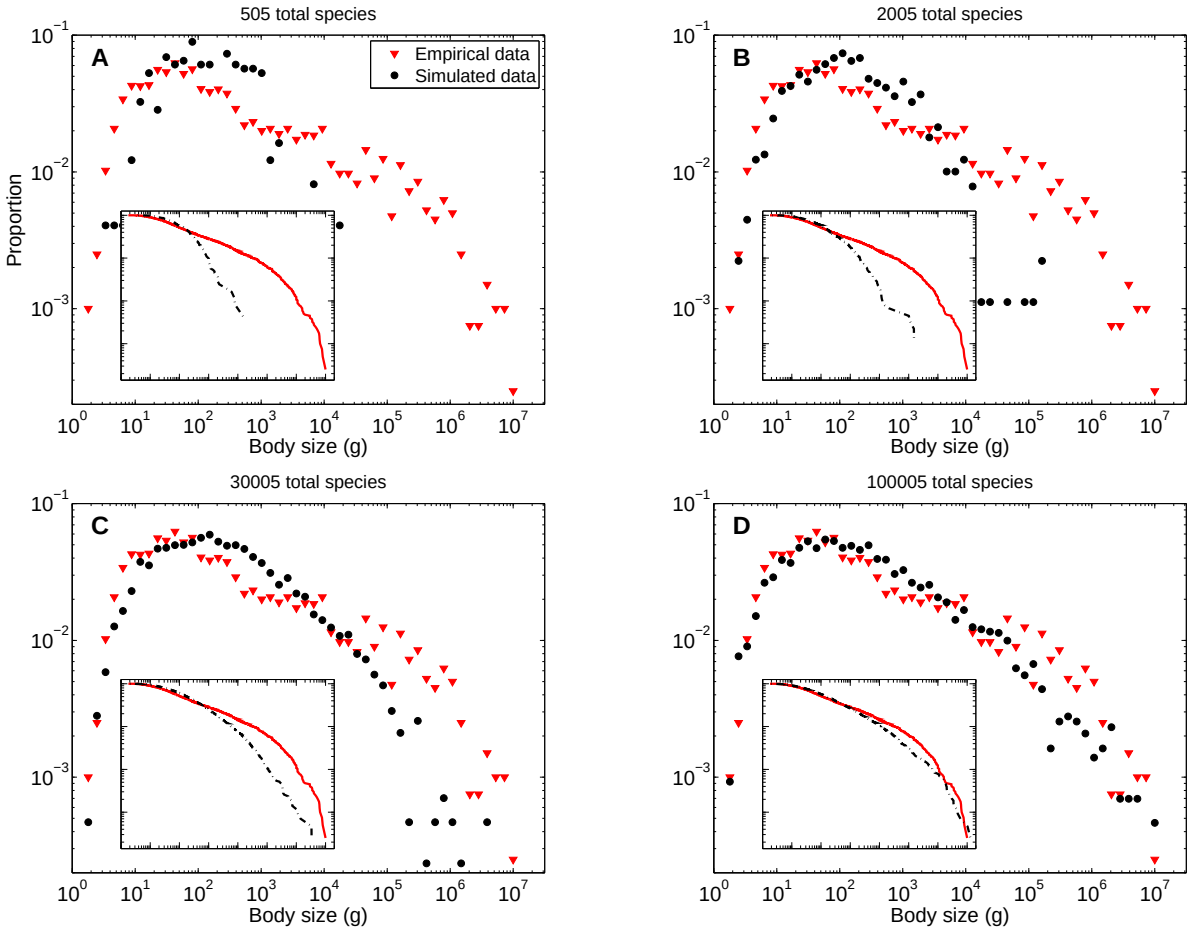

Figure 7: Snapshots from the simulation, showing the simulated histogram (black dots) of species body sizes, along with the empirical distribution (red triangles), for four values of the total number of species generated. In this case, the steady-state number of species is 4000, however, because of extinction, the simulation by the last snapshot has produced over 100,000 species. The insets show the ccdf on log-log axes.

data in the left tail (from 40g down to 2g) is excellent, which supports the interpretation that a hard lower limit under macroevolutionary diffusion is sufficient to explain why the modal body size (40g) is larger than the small body size (2g). As the simulation continues, the right-tail expands such that we see begin to see several very large-body species (100kg) fairly quickly, but megafuana (1000kg or more) remain relatively rare even after tens of thousands of species have become extinct. After roughly 100,000 species, the simulated distribution agrees fairly closely with the empirical

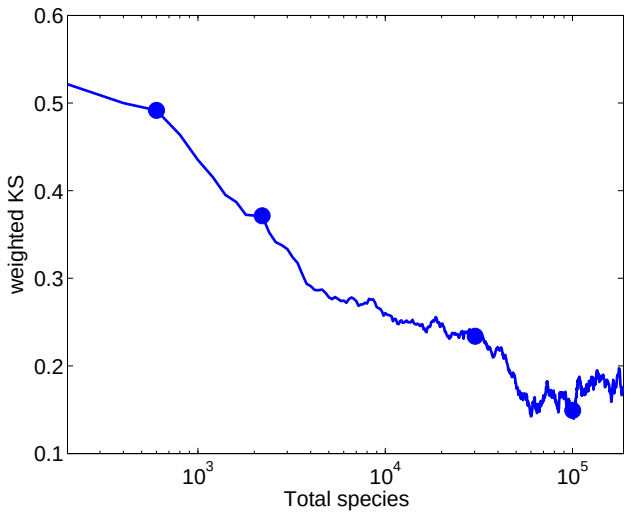

Figure 8: The distance (measured by the tail-weighted KS statistic) between the simulated distribution and the empirical one, as a function of the number of steps along in the simulation.

data, except for a few places you can spot with your eye, e.g., around 10^3 and between 10^5 and 10^6.

Figure 8 shows the goodness-of-fit statistic over the course of the simulation, showing that the simulated distribution converges on the empirical distribution, i.e., the wKS value decreases with time. The marked points on the figure correspond to the snapshots in Figure 7.

2.4 Variations on the model

Although the agreement of the model with the data is fairly good, an important question remains as to whether all of the mechanisms we assumed are necessary. The goodness of the fit shows that they are collectively sufficient, but could we get away with a simpler version of this model? That is, could we remove or simplify one of the three main mechanisms and get just as good a match to the empirical data?

Figure 9 shows the results of this kind of model-simplification exercise. The first panel shows the results for the full model. In the subsequent panels, we simplify the model by removing something from the full model: first we remove the increased bias in the diffusion kernel at very small body sizes (recall Figure 5), then we additionally remove the size-dependent extinction risk. These changes are made not by altering the simulation code, but by zeroing-out the appropriate parameters. In each case, we combine the results of 1000 repetitions of the simulation to derive a central

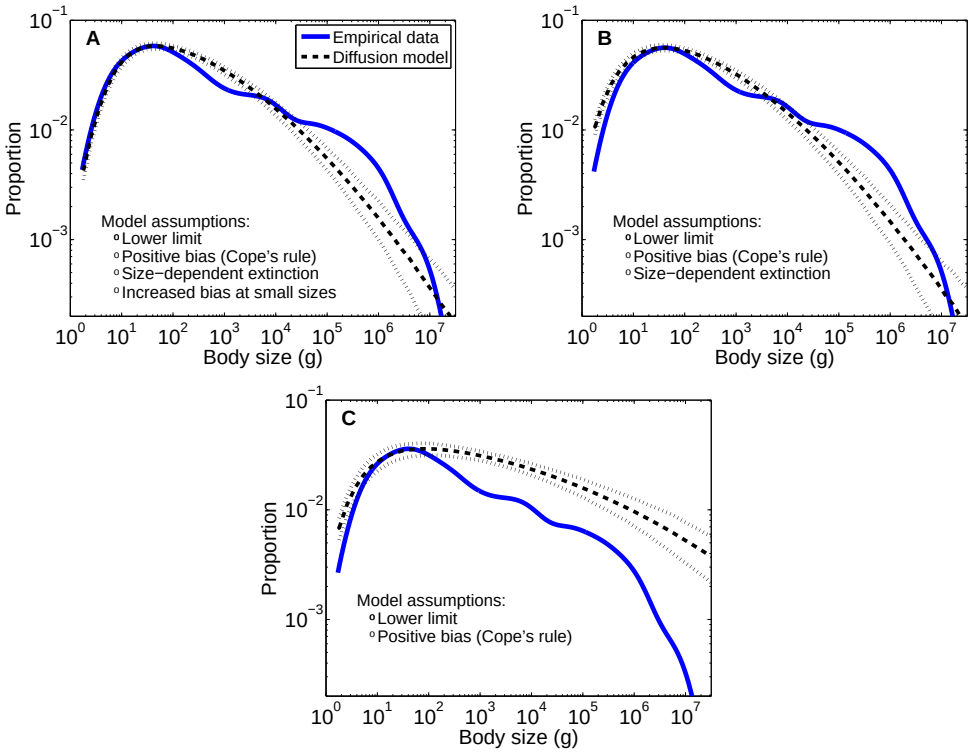

Figure 9: Simulated distributions of species body size (central tendency ± 95% confidence intervals from 1000 repetitions; all model parameters held fixed, but with ρ reëstimated to provide the best fit) and the empirical distribution of Recent terrestrial mammals. (a) The model described in the text. (b) The same model as before but with a bias $\langle \log \lambda \rangle$ that is independent of size. (c) The same model as before but with an extinction risk that is independent of size.

tendency of the predicted distribution of body sizes (which smooths out the fluctuations from any particular run of the model). We can also consider more dramatic changes, like removing the lower bound $x_{\min}$ (i.e., setting $\log x_{\min} \to -\infty$).

Given the lectures on random walks, it is perhaps unsurprising that removing the size-dependent extinction bias yields poor agreement with the empirical data — it produces far too many very large-bodied species because there is nothing to prevent lucky species from racing off to $+\infty$. Removing the increased positive bias at small sizes has only a marginal impact on the agreement with the empirical data — the form used in the full model makes the model agree perfectly with the data, while we get a slight over-estimate of the number of species if we remove this part of the

		lower boundary	Cope's rule	small-body bias		extinction dependence
Model description	⟨wKS⟩	$x_{\min}$	δ	c_1	c_2	ρ
Full model	0.181(1)	1.8	0.04	0.33	1.30	0.025
Full model with no small-size bias	0.244(1)	1.8	0.04	0	0.25	0.023
Unbiased diffusion with lower bound	2.97(3)	1.8	0	0	0.25	0
Cope's rule with size-dependent extinction	10.60(7)	10^{-8}	0.04	0	−8	−0.002
Cope's rule alone	11.72(9)	10^{-8}	0.04	0	−8	0
Size-dependent extinction alone	10.37(6)	10^{-8}	0	0	−8	−0.005

Table 2: A comparison of the full model with five simpler models. Each of these alternatives are special cases of the full model. Each was run 1000 times, from which we computed the central tendency of the simulated distribution and the average distance ⟨wKS⟩ from the empirical distribution. Results are for $\Pr(\lambda)$ with power-law tails and the power-law model of extinction risk, but similar results for log-normal tails or logarithmic extinction risk; the standard error in the last digit is quoted parenthetically. For models with $\rho \neq 0$, ρ was reëstimated by minimizing ⟨wKS⟩.

model. As a result, one could argue that the evidence supporting this piece of the full model is not as strong as the evidence supporting the size-dependent extinction curve.

Table 2 shows the results of a more systematic investigation of these choices. In each case, in order to be fair to the alternative models, we reëstimate the value of ρ in order to make the alternative model fit as closely as possible to the empirical data, just as we did for the full model. Among these alternative models, we can see clearly from the $\langle wKS \rangle$ values that the models without a lower boundary are simply terrible — like the models with no extinction curve, there is nothing to prevent the distribution from moving to very small masses, yielding a poor agreement with the empirical data.

3 Matlab code

3.1 Independent and branching random walkers

This code produces Figure 1.

```
% simulation parameters: number of walkers, number of steps
[N T] = deal(5,10^4);

% -- simulate N independent random walkers
x = zeros(N,T);
for i=2:T
    x(:,i) = x(:,i-1)+randn(N,1);        % everyone takes a step
end;

% simulate N branching random walkers
y = -Inf.*ones(N,T);
n = 1;  % 1 random walker, initially
y(n,1) = 0;
for i=2:T
    y(1:n,i) = y(1:n,i-1)+randn(n,1);    % everyone takes a step
    if rand(1)<(N/T) && n<N
        n = n+1;                         % branch
        parent   = ceil((n-1).*rand(1)); % choose random parent
        y(n,i-1) = y(parent,i-1);        % copy parent's location
        y(n,i)   = y(n,i-1)+randn(1);    % take a random step
    end;
end;

% make the pretty figures
cstr = {'r','b','m','k'};
xmax = 1.1*max(max(abs(x)));                    % choose a common y-scale for the walkers
ymax = max([xmax 1.1*max(max(abs(y(y>-Inf))))]); %

% plot trajectories for independent walkers
figure(1); clf;
for i=1:N
    plot((1:T),x(i,:),strcat(cstr{1+mod(i,length(cstr))},'-')); hold on;
end;
hold off;
set(gca,'YLim',[-ymax ymax],'FontSize',16);
xlabel('Time, t','FontSize',16);
ylabel('Position, x','FontSize',16);
title('Non-branching random walk, \lambda~N(0,1)','FontSize',16);

% plot trajectories for branching random walkers
```

```
figure(2); clf;
for i=1:N
    plot((1:T),y(i,:),strcat(cstr{1+mod(i,length(cstr))},'-')); hold on;
end;
hold off;
set(gca,'YLim',[-ymax ymax],'FontSize',16);
xlabel('Time, t','FontSize',16);
ylabel('Position, x','FontSize',16);
title('Branching random walk, \lambda~N(0,1)','FontSize',16);
```

3.2 Cladogenetic diffusion code

This code runs the basic cladogenetic diffusion simulation. It does not include visualization procedures, but simply plot the histogram of the masses stored in the vector `x` to do this. (Or rather, the non `-Inf` masses in that vector, since `-Inf` is the value used to denote an unused element of the array.)

```
% simulation parameters
xmin  = 1.8;   % lower bound
x0    = 40;    % founder body size
n     = 5000;  % num. species at equilbrium
beta  = 1/n;   % baseline extinction rate
rho   = 0.025; % rate of extinction increase
nu    = 1.6;   % mean species lifetime (My)
tau   = 60;    % total simulation time (My)
c(1)  = 0.33;  % log-lambda intercept
c(2)  = 1.30;  % log-size intercept
delta = 0.04;  % systematic bias (Cope's rule)
sigma = 0.63;  % variance
alpha = 0.30;  % power-law tail

% data structure set up
tmax = round((tau/nu)*n);
xmax = 10^15;
x    = -Inf*ones(ceil(1.5*n),1);
x(1) = x0;
[kdt ns nk kd] = deal(5000,1,0,1);
f_stop = 0;

% begin main loop
while ~f_stop

     % begin cladogenesis step
     pair = [ceil(ns*rand(1)) ns+1];
```

```
     mass = x(pair(1),1);
     L1 = mass/xmin;  % lower bound
     L2 = xmax/mass;  % upper bound

     % model of Cope's rule
     if log10(mass)<c(2)
          % increased bias for small sizes
          mu = (-c(1)/c(2))*log10(mass)+c(1)+delta;
     else
          % uniform bias for large sizes
          mu = delta;
     end;

     % Monte Carlo draw of growth factors
     tt = [0 0];
     while any(tt<1/L1 | tt>L2)
          % F(lambda) with power-law tails
          tt = exp(randn(2,1)*sigma+mu).* ((rand(2,1).* ...
               (1-1./L1)+1./L1).^alpha)./ ((rand(2,1).*(1-1./L2)+1./L2).^alpha);
     end;
     x(pair) = mass.*tt;
     kd = kd+2;
     ns = ns+1;
     % end cladogenesis step

     % begin extinction step
     % power-law model of extinction risk
     kl = rand(ns,1) < 10.^(rho*log10(x(1:ns))+log10(beta));
     if sum(kl)>0
          x(1:sum(~kl)) = x(~kl);
          x(sum(~kl)+1:ns) = repmat([-Inf],sum(kl),1);
          ns = sum(~kl);
          nk = nk+sum(kl);
     end;
     % end extinction step

     % begin check stop-criteria
     if kd>=tmax, f_stop = 1; end;
     % end check stop-criteria

end;
% end main loop
```

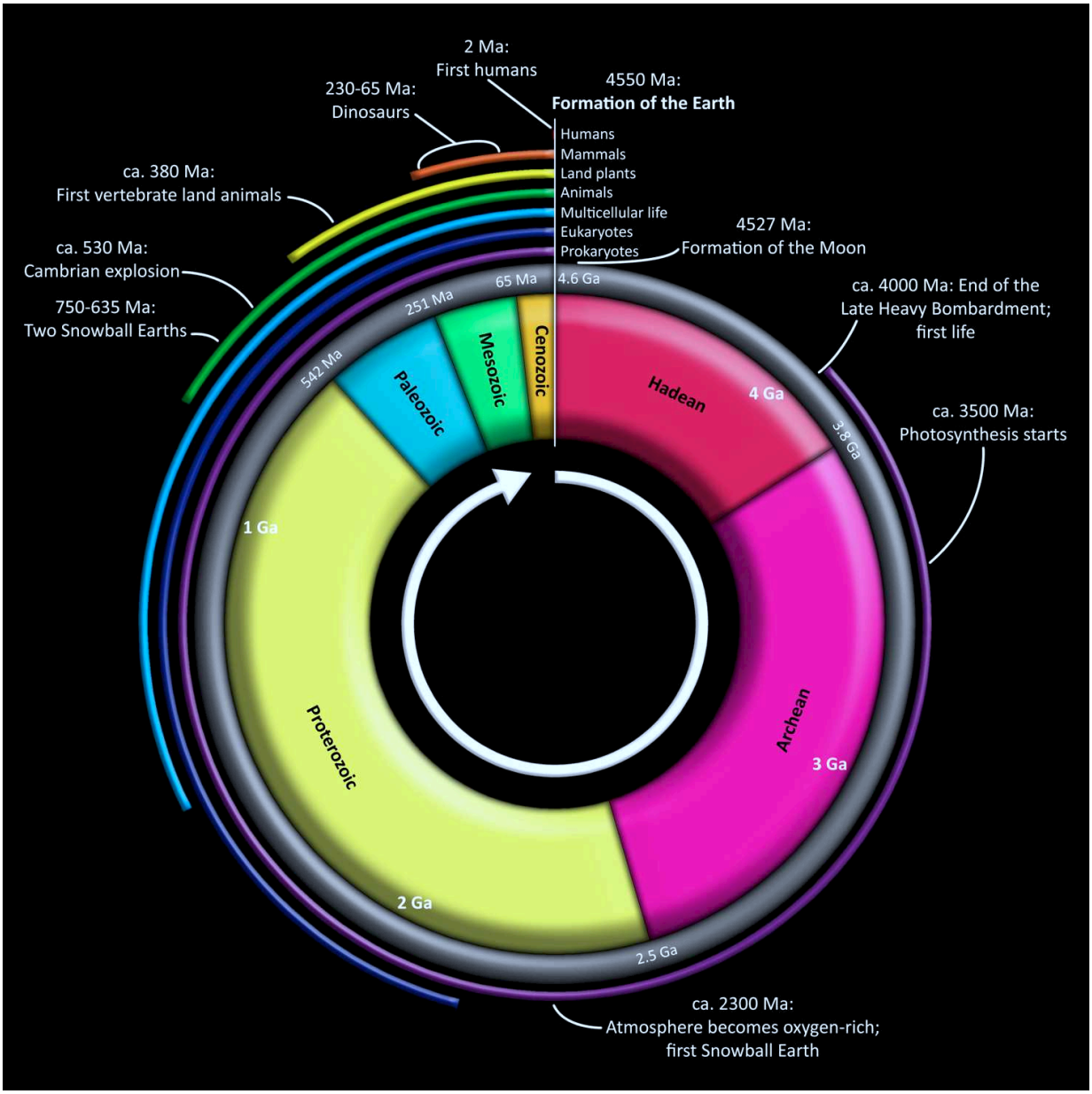

Figure 10: Geological (or, macroevolutionary) time for the Earth, annotated with important biological and other events. Models of macroevolution consider patterns in biological variation and change at roughly these timescales (that is, on the order of millions of years).