# RANDOMNESS COMPLEXITY CORRELATIONS

Andrea Giansanti
Dipartimento di Fisica, Sapienza Università di Roma

DA_2022 L2  Rome march 7, 2022

DIPARTIMENTO DI FISICA

SAPIENZA
UNIVERSITÀ DI ROMA

# L2 OUTLINE

- Logistics (mailing list, e-learning platform, Bioarxiv, Pubmed)
- structure of a scientific paper, peer review, open access
- from Galilei's "simplicity" to Parisi's "complexity"
- evolutionary time, randomness
- complexity/simplicity
- the computational dimension (advent of computers. Moore's law)
- Computational medicine/computational phychiatry
- correlation is not causation
- Naomi Altman's set of papers: Points of Significance
- https://www.nature.com/collections/qghhqm/pointsofsignificance

see also a blog page on data visualization:
http://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html

- quantitative approach: Cell Biology by the numbers (Milo & Phillips)
- http://book.bionumbers.org/how-does-metabolic-rate-scale-with-size/

# DA_2022 TO DO

Enroll to the DA_2022 Course on the Sapienza Moodle Platform

https://elearning.uniroma1.it/course/view.php?id=14921#section-0

password: cwoese2

# Let us collect a set of original data

- We want to collect real data about the sample of the human population constituted by this class: the people in this course, then please let us discuss NOW how to build up a small database collecting 3 info per person:
- BODY MASS (kg, with one decimal)
- BODY HEIGHT (metres, with 2 decimals)
- SEX (M/F)

I propose to have two volunteers…

# Let us make our own dataset: Body Mass Index (BMI)

- The ‘average man’ of Adolphe Quetelet (1796-1874)
- BMI =[ Weight(Kg)/(height(m))$^2$]

# Another experiment to be done on the internet

- Collect the properties of several mineral waters available on the Italian Market…
- http://acqueminerali.it/

again: volunteers…

- The galilean paradigm: based on the “removal of the animal”
- What is physics: the study of material bodies, localized in space and time
- Reference frames + clocks (newtonian time, not percolating, it uniformly flows, always at the same rate
- Biology is based on the careful observation of single cases, then correlated in a qualitative way, based on senses, into classes (classification) (e.g. species)
- **The darwinian shift**: biological time vs physical time
- **The molecular revolution** (Watson & Crick) macroscopic genetic laws can be explained by looking at molecular materials and information

# EVOLUTION HAS TO WITH RANDOMNESS

A machine – now obsolete to generate random events (numbers)

- **The universal structure of a scientific paper:**

1. Introduction (what is the problem, where it comes from: what is the scientific question, reference to previous works)
2. Materials and methods: written in the style of a cooking recipe (with the aim of sharing a pleasure, willing to be reproduced)
3. Results ( Which facts we are proposing to the attention of the community, are they valid ? Data analysis play a major role)
4. Discussion (validity of the results, DA)

# Science dissemination process

Publish or perish
globalization of publishing enterprises (pros/cons)
peer review process
impact factors (reputation/cost)
open access

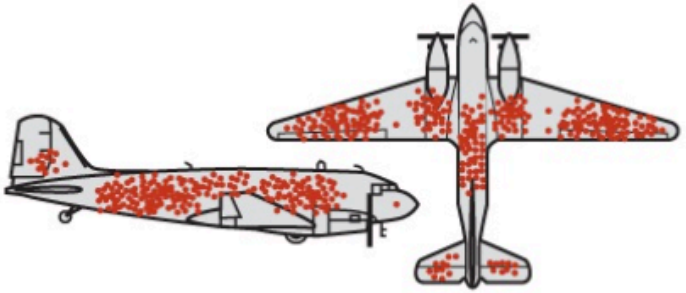
- preprint sharing (bioRxiv)https://www.biorxiv.org/ Yale Biology
- https://www.medrxiv.org/ Yale Medicine
- https://arxiv.org/ Cornell Physics

open access open review
https://f1000research.com/

PUBMED: https://pubmed.ncbi.nlm.nih.gov/

# Bias or large sampling error?

- During World War II, the British Royal Air Force estimated the density of bullet holes on different sections of planes returning to base from aerial sorties. Their goal was to use this information to determine which plane sections most needed additional protective shields. (It was not possible to reinforce the whole plane, because it would weigh too much.) They found that the density of holes was highest on the wings and lowest on the engines and near the cockpit, where the pilot sits (their initial conclusion, that therefore the wings should be reinforced, was later shown to be mistaken). What is the main problem with the sample: bias or large sampling error? What part of the plane should have been reinforced?



Whitlock et al., *The Analysis of Biological Data*, 2e, © 2015
W. H. Freeman and Company

# Phylogenetic trees: evolutionary vs newtonian time

(A)

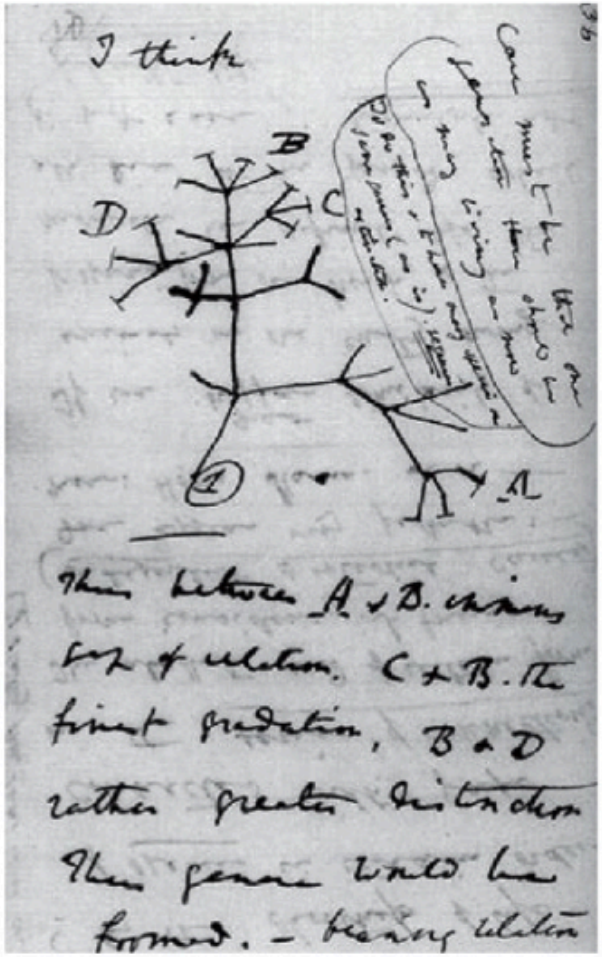



Figure 3.4 Physical Biology of the Cell (© Garland Science 2009)

(B)

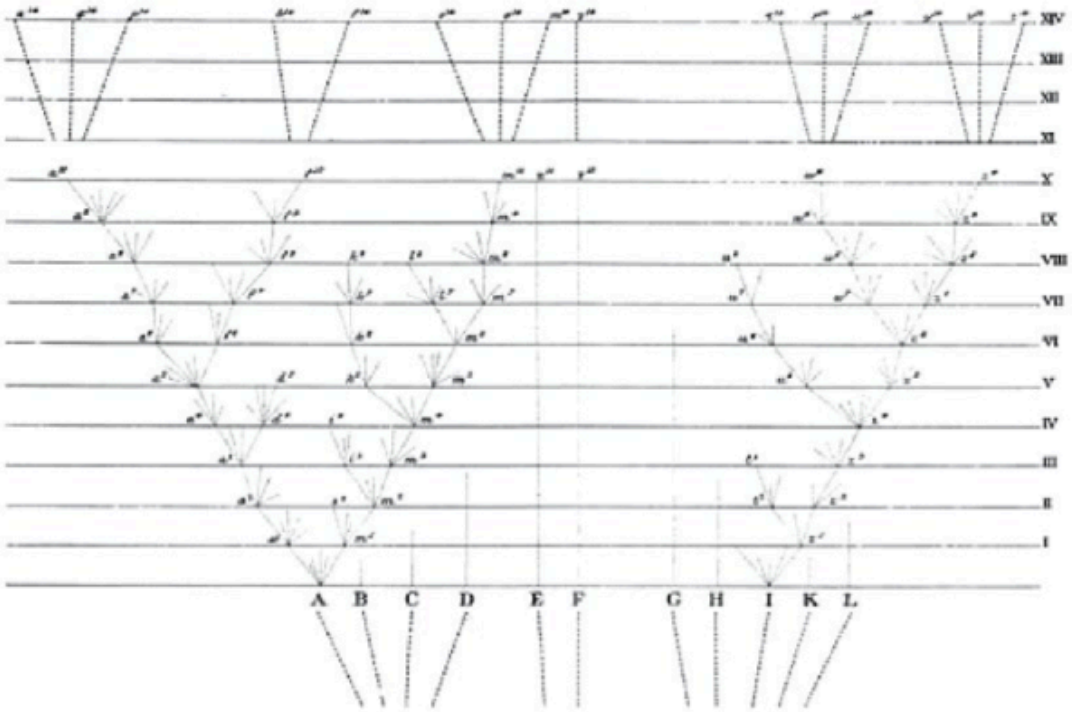


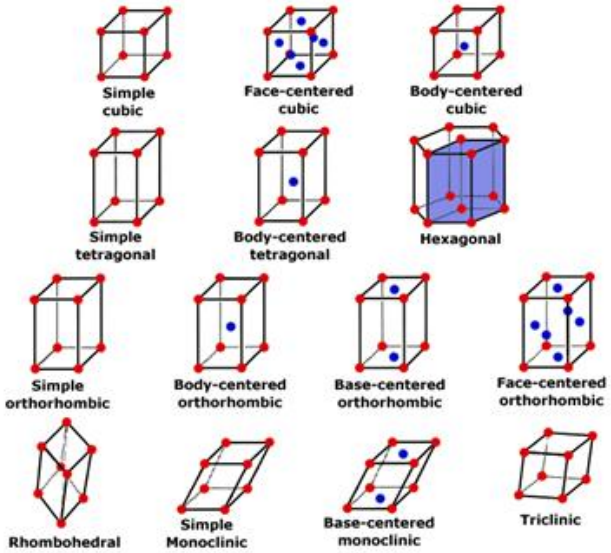

**Figure 3.4** Two versions of Darwin's phylogenetic tree. (A) In his notebooks, Darwin drew the first version of what we now recognize as a common schematic demonstrating the relatedness of organisms. He introduced this speculative sketch with the words "I think" as his theory was beginning to take form. (B) In the final published version of *On the Origin of Species*, the tree had assumed more detail showing the passage of time and explicitly indicating that most species have gone extinct. (Adapted from C. Darwin, *On the Origin of Species*, London, John Murray, 1859. Courtesy of The American Museum of Natural History.)

Evolution is not a deterministic process
but a random one

# Complexity

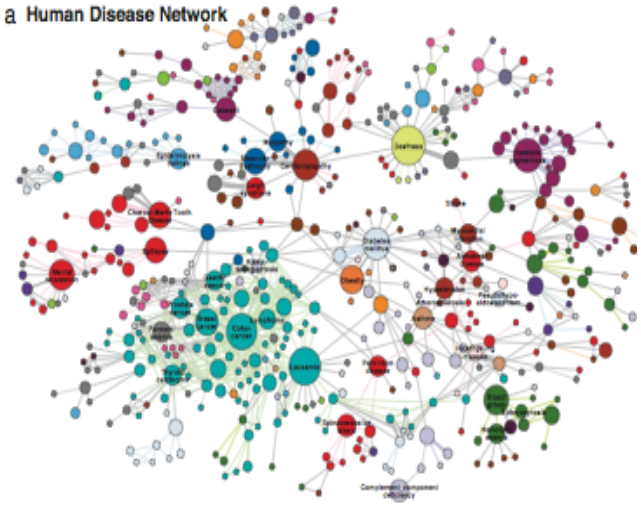
*Between randomness and order*

**LATTICES**



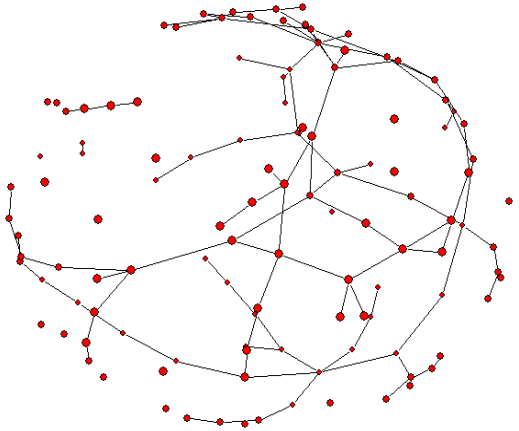

Regular networks
Symmetric

**COMPLEX NETWORKS**




Scale free networks
Small world
With communities
ENCODING INFORMATION
IN THEIR STRUCTURE

**RANDOM GRAPHS**



Totally random
Binomial degree
distribution

# Scale-free networks

**Technological networks**
Internet
World-Wide Web

**Biological networks**
Metabolic networks,
Protein-interaction networks,
Transcription networks

**Transportation networks**
Airport networks

**Social networks**
Collaboration networks
Citation networks
Facebook

**Economical networks**
Networks of shareholders
The World Trade Web

# IL CIELO SOPRA ROMA E NUOVI PARADIGMI DELLA FISICA, come dire, IL RITORNO DELL' ANIMALE
# (Animal REDUX the physics of Swarming, Flocking; Schooling)

CoBBS – Collective Behaviour in Biological Systems

**Diretto da Andrea Cavagna e Irene Giardina**
**(ispirazione di Giorgio Parisi, v. il suo libro UN VOLO DI STORNI)**
https://www.isc.cnr.it/groups/cobbs/

# Computational Medicine: a new paradigm?

**CM is an emerging discipline devoted to the development of quantitative approaches for understanding the mechanisms, diagnosis and treatment of human disease through applications of mathematics, engineering and computational science. The core approach of CM is to develop computational models of the molecular biology, physiology, and anatomy of disease, and apply these models to improve patient care. CM approaches can provide insight into and across many areas of biology, including genetics, genomics, molecular networks, cellular and tissue physiology, organ systems, and whole body pharmacology. At the Institute for Computational Medicine (ICM), the "birthplace" of CM, research focuses on four key areas: Computational Molecular Medicine, Computational Physiological Medicine, Computational Anatomy, and Computational Healthcare.**

BIG DATA PARADIGM

http://www.bdc4cm.org/index.php?title=Main_Page

https://www.bme.jhu.edu/graduate/mse/degree-requirements/computational-medicine/

*Computational Psychiatry* publishes original research articles and reviews that involve the application, analysis, or invention of theoretical, computational and statistical approaches to mental function and dysfunction. Topics include brain modeling over multiple scales and levels of analysis, and the use of these models to understand psychiatric dysfunction, its remediation, and the sustenance of healthy cognition through the lifespan. The journal also has a special interest in computational issues pertaining to related areas such as law and education.

https://www.technologyreview.com/s/608322/the-emerging-science-of-computational-psychiatry/

A YOUNG EMERGING COMMUNITY:
LONDON COMPUTATIONAL PSYCHIATRY

2 DAYS COURSE (2014-)

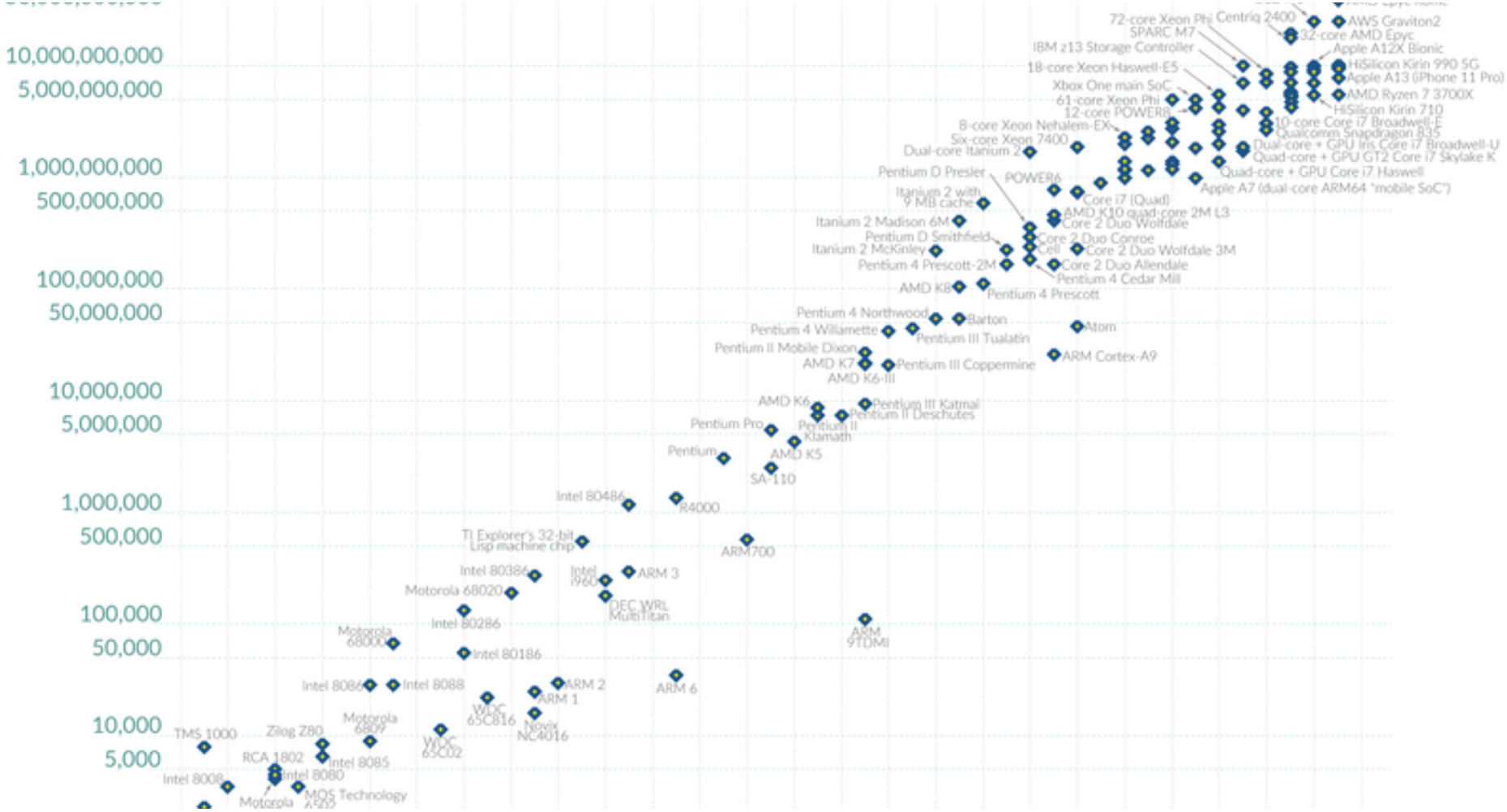
https://www.cpcourse.org

THE BAYESIAN BRAIN!
A NEW APPROACH TO MENTAL ILLNESS BASED ON MODERN PROBABILITY CALCULUS

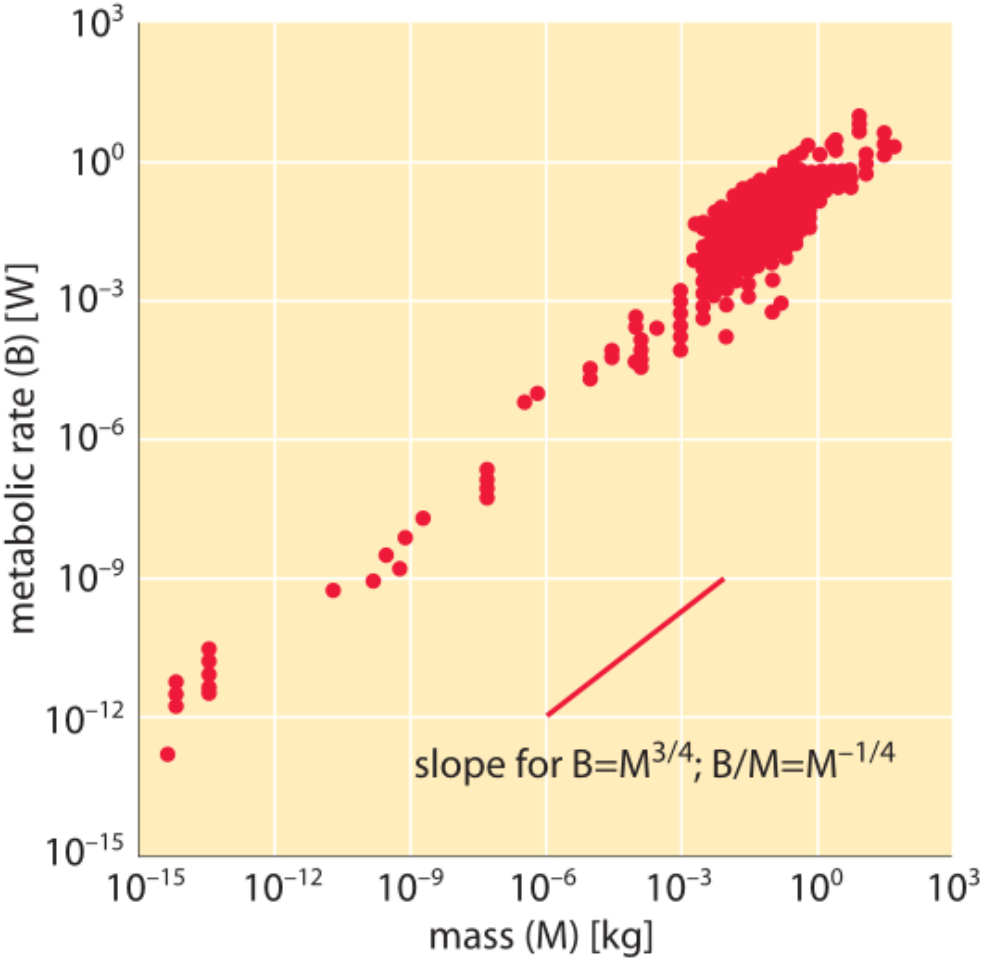
DO YOU KNOW WHAT IS BAYES THEOREM?

Moore's law (1965)

…the number of transistors in a dense integrated circuit (IC) doubles about every two years



Scaling arguments: regressions, histograms, models




http://book.bionumbers.org/how-does-metabolic-rate-scale-with-size/ /

# SCALING LAWS

example 1.2 METABOLISM / BODY MASS
(see: Cell BIOLOGY BY THE NUMBERS p. 205)

$$M \cong B^{\alpha}$$

This is a 'power law' dependence. If M is measured in Watts (Joule·sec$^{-1}$) and B in Kg, throug a constant K with the right units:

$$M = K B^{\alpha}$$ $\alpha$, exponent

To linearize this 'law' take Logs of both sides:

$$\text{Log } M = \alpha \text{ Log } B + \text{Log } K$$

$\alpha$ is then the 'slope' of the straight line we get in double-logarithmic graphs.

( EXPONENTIAL LAWS

$$Y = Y_0 e^{+KX}$$

$$\text{Log } Y = \text{Log } Y_0 + K X$$

Then we get a straight line in semilog plot
(Log-Lin plot) )

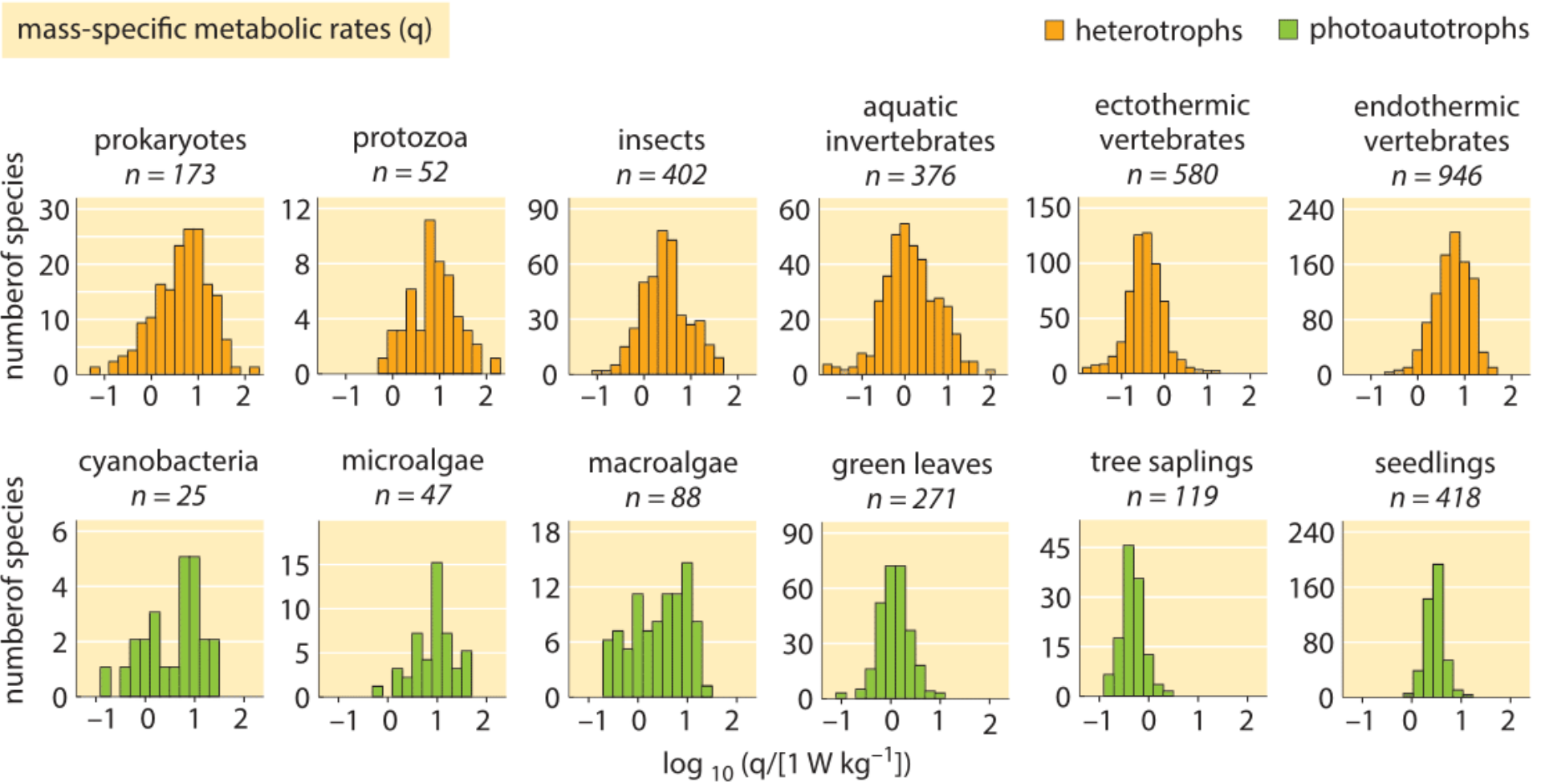
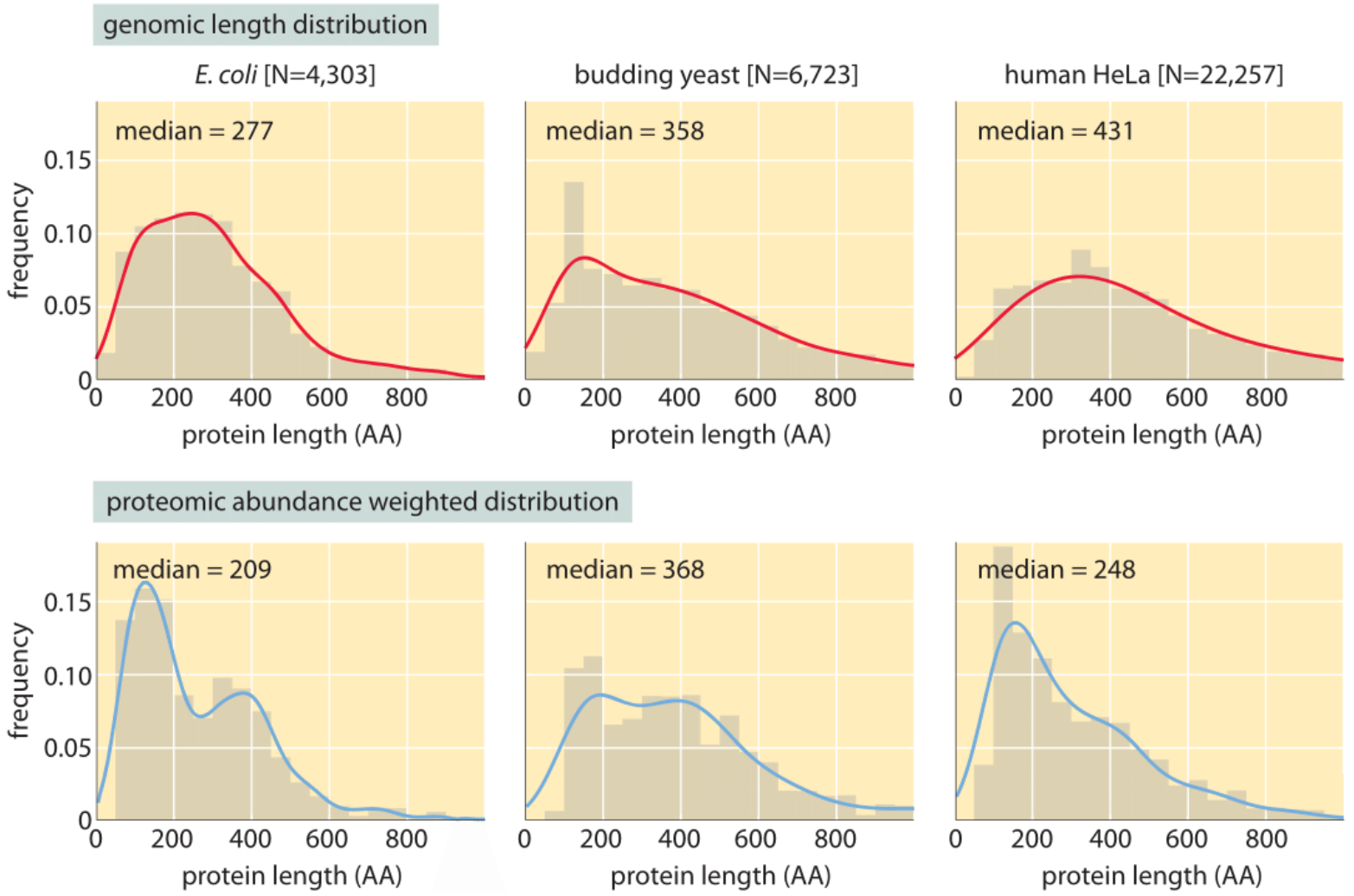



Figure 2: Histograms of resting metabolic rates normalized to wet weight. Across many orders of magnitudes of body size and widely differing phylogenetic groups the rates are very similar at about 0.3-9 W/kg wet weight. (Adapted from A. M. Makarieva, Proc. Nat. Acad. Sci., 105:16994, 2008.)

Let us look at distributions: remember the distiction between: Sample quantities and population quantities. What is the median?



Previous slide:
**Distribution of protein lengths in E. coli, budding yeast and human HeLa cells**.

(A) Protein length is calculated in amino acids (AA), based on the coding sequences in the genome.
(B) Distributions are drawn after weighting each gene with the protein copy number inferred from mass spectrometry proteomic studies (M. Heinemann in press, M9+glucose; LMF de Godoy et al. Nature 455:1251, 2008, defined media; T. Geiger et al., Mol. Cell Proteomics 11:M111.014050, 2012). Continuous lines are Gaussian kernel-density estimates for the distributions serving as a guide to the eye.

http://book.bionumbers.org/how-big-is-the-average-protein/

# Statistics for biologists:

Nature Methods Points of Significance series (Naomi Altman)

https://www.nature.com/collections/qghhqm/pointsofsignificance

POINTS OF SIGNIFICANCE

## Association, correlation and causation

Correlation implies association, but not causation. Conversely, causation implies association, but not correlation.

Most studies include multiple response variables, and the dependencies among them are often of great interest. For example, we may wish to know whether the levels of mRNA and the matching protein vary together in a tissue, or whether increasing levels of one metabolite are associated with changed levels of another. This month we begin a series of columns about relationships between variables (or features of a system), beginning with how pairwise dependencies can be characterized using correlation.

- Two variables are independent when the value of one gives no information about the value of the other.

- For variables X and Y, we can express independence by saying that the chance of measuring any one of the possible values of X is unaffected by the value of Y, and vice versa, or by using conditional probability, $P(X|Y) = P(X)$.

- For example, successive tosses of a coin are independent—for a fair coin, $P(H) = 0.5$ regardless of the outcome of the previous toss, because a toss does not alter the properties of the coin.

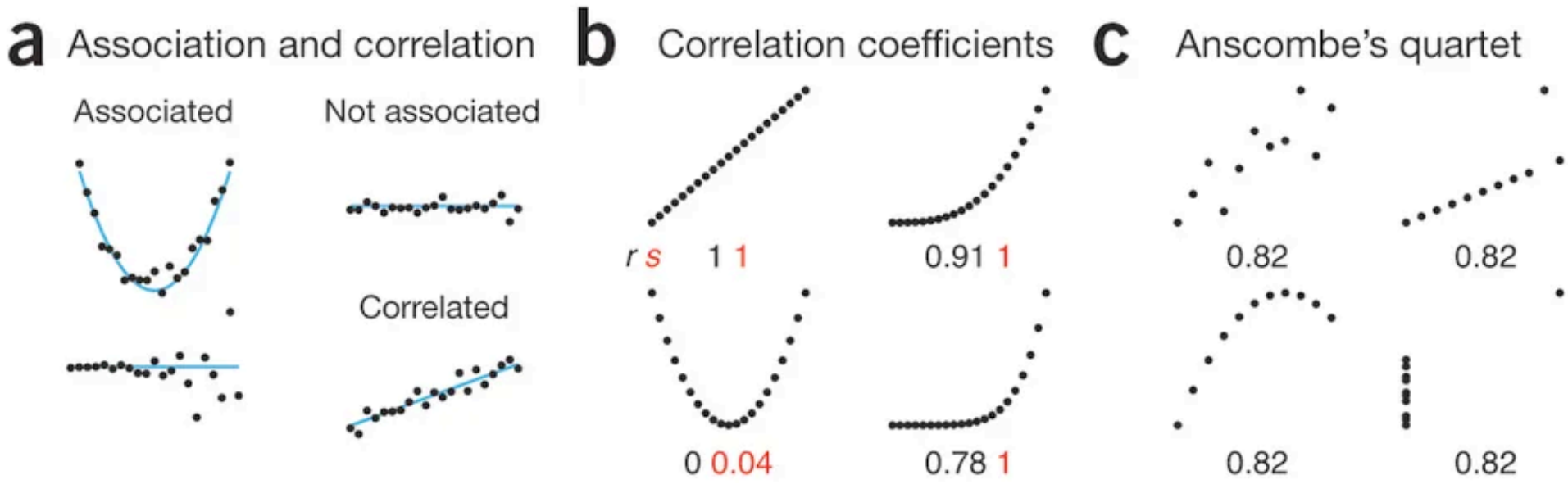
- In contrast, if a system is changed by observation, measurements may become associated or, equivalently, dependent. Cards drawn without replacement are not independent; when a red card is drawn, the probability of drawing a black card increases, because now there are fewer red cards.

- Association should not be confused with causality;
  if X causes Y, then the two are associated (dependent).

- However, associations can arise between variables in the presence (i.e., X causes Y) and absence (i.e., they have a common cause) of a causal relationship

As an example, suppose we observe that people who daily drink more than 4 cups of coffee have a decreased chance of developing skin cancer. This does not necessarily mean that coffee confers resistance to cancer; one alternative explanation would be that people who drink a lot of coffee work indoors for long hours and thus have little exposure to the sun, a known risk.

If this is the case, then the number of hours spent outdoors is a confounding variable—a cause common to both observations. In such a situation, a direct causal link cannot be inferred; the association merely suggests a hypothesis, such as a common cause, but does not offer proof. In addition, when many variables in complex systems are studied, spurious associations can arise. Thus, association does not imply causation.




(a) Scatter plots of associated (but not correlated), non-associated and correlated variables. In the lower association example, variance in y is increasing with x.

(b) The Pearson correlation coefficient (r, black) measures linear trends, and the Spearman correlation coefficient (s, red) measures increasing or decreasing trends. (c) Very different data sets may have similar r values. Descriptors such as curvature or the presence of outliers can be more specific.

- In everyday language, dependence, association and correlation are used interchangeably. Technically, however, association is synony-mous with dependence and is different from correlation (Fig. ıa). Association is a very general relationship: one variable provides information about another. Correlation is more specific: two vari-ables are correlated when they display an increasing or decreasing trend. For example, in an increasing trend, observing that X > μXimplies that it is more likely that Y > μY. Because not all associations are correlations, and because causality, as discussed above, can be connected only to association, we cannot equate correlation with causality in either direction.

- When "correlated" is used unmodified, it generally refers to Pearson's correlation, given by $\rho$ (X, Y) = cov(X, Y)/ $\sigma$ X $\sigma$ Y, where cov(X, Y) = E((X – μX)(Y – μY)). The correlation computed from the sample is denoted by r. Both variables must be on an interval or ratio scale; r cannot be interpreted if either variable is ordinal. For a linear trend, |r| = ı in the absence of noise and decreases with noise, but it is also possible that |r| < ı for perfectly associated nonlinear trends (Fig. ıb). In addition, data sets with very different associa-tions may have the same correlation (Fig. ıc). Thus, a scatter plot should be used to interpret r. If either variable is shifted or scaled, rdoes not change and r(X, Y) = r(aX + b, Y). However, r is sensitive to nonlinear monotone (increasing or decreasing) transformation.