# Why and how Zipfian power-law behaviour dominates (nearly) all discrete systems


Prof. Les Hatton
lesh@cantab.net



Prof. Greg Warr
Greg.Warr@cantab.net


10-Jan-2024

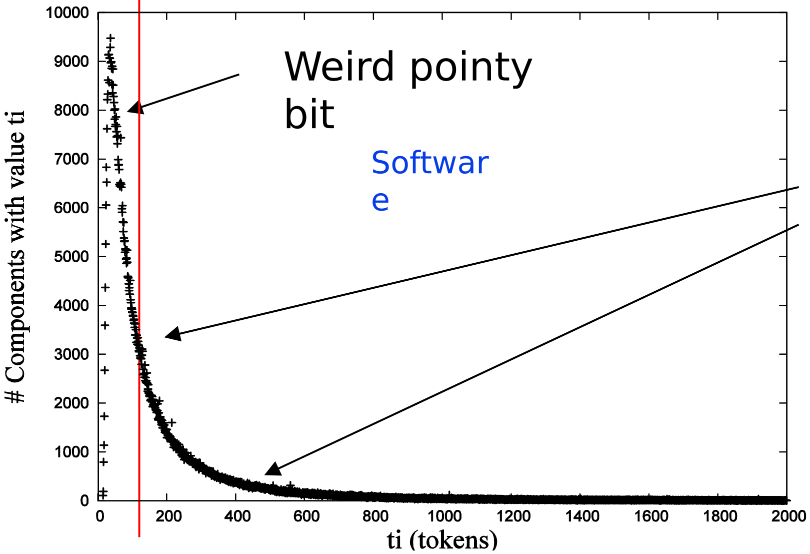
# Unexpected emergent properties in all discrete systems

**Emergent properties** are properties of the "system" as a whole.

A discrete system is one made up of distinguishable pieces, for example amino acids in proteins.  Meaning is irrelevant.

# Software: programming tokens 2009-2012

pdf – prob. dist. fn.



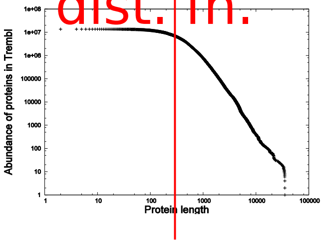

Weird pointy bit

Software

Almost perfect power-law

$a_i \sim t_i^{-1/\beta}$

ccdf – comp. cum. dist. fn.




600,000 functions from 80 million lines of C

# Power Laws $p_i \sim t_i^{-\beta}$
# Ubiquitous and Often Regarded with Scepticism

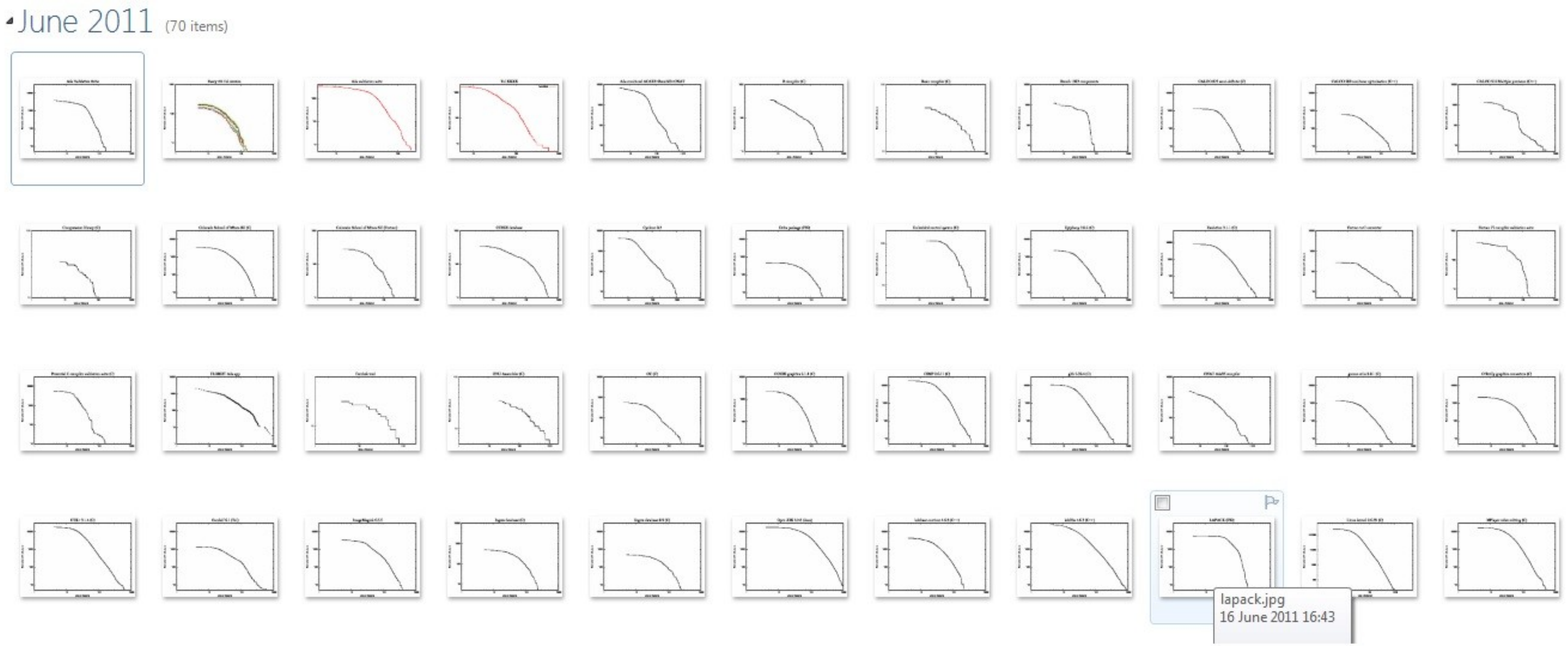
Necessary Condition

Linearity on log-log plot, adj $R^2 > 0.95$, $p < 10^{-12}$

Sufficiency Condition – Method of Clauset *et al*. 2009

$p > 0.1$, implying that we cannot reject the hypothesis of a power law

# A singular pattern …



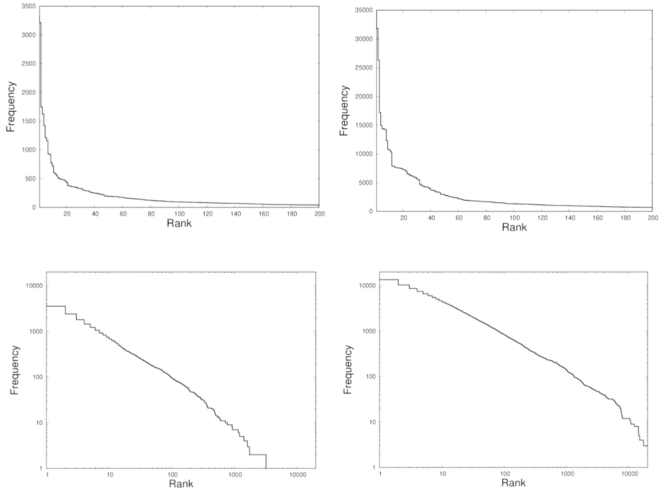

Some of the many software ccdfs I analysed in 2009-12

# Then along came Zipf …

Unordered systems such as word counts in books

Three Men in a Boat

Swedish Bible




Discovered empirically by George Zipf (1935)

# Pause for thought 2012-13

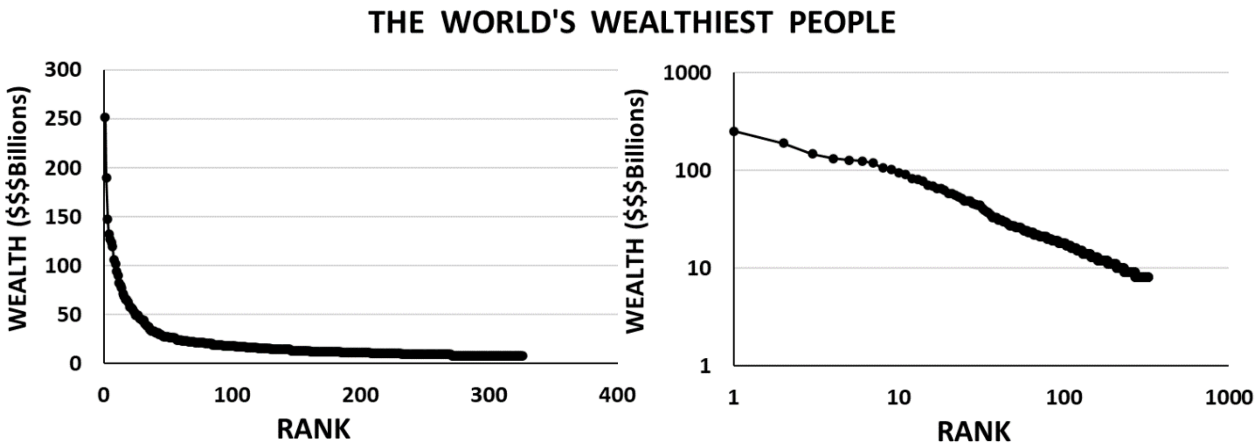
The software distributions are **independent** of language and application area

Zipf is **independent** of language and subject matter

Hmm.

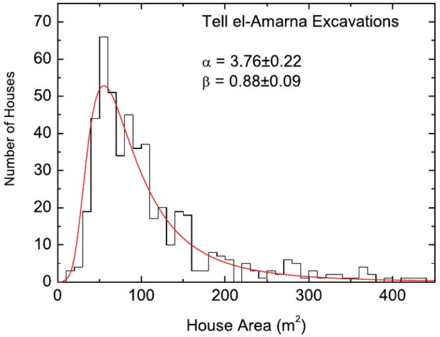
# Pareto and wealth in all its forms …

Wealth by rank

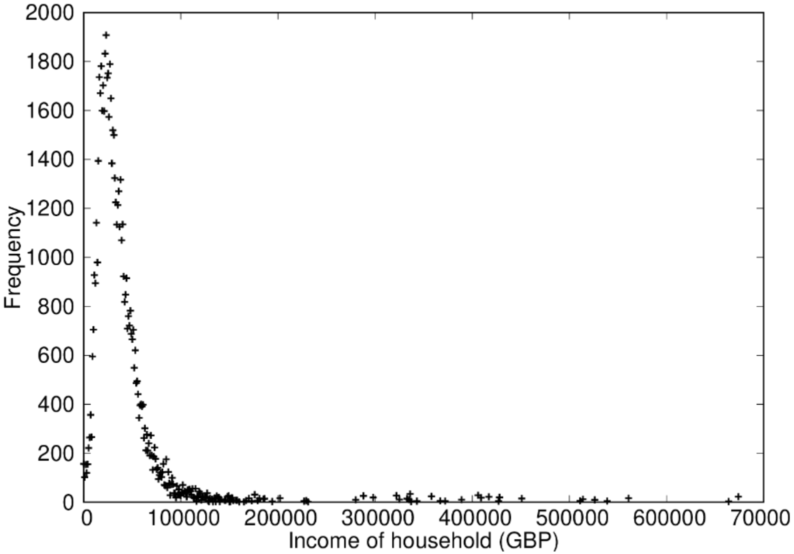


# Pareto and wealth in all its forms …

Wealth by amount – floor area Akhetaten (Abul-Magd)




Wealth by amount – UK household income)



# Pause for thought 2013-14

So wealth shows up as both a Zipfian rank order distribution

AND

A spiky software length distribution

Hmmm

# 2014-2016

- There appear to be two kinds of distribution, one the same as Zipf and a new one with a spike at the start.  Both are dominated by power-law behaviour.
- They are both independent of application area, provenance and symbol representation.  Proxies have the same property
- Evidence of scale-free behaviour (the systems have very different sizes).

*Anybody smell a conservation principle?*

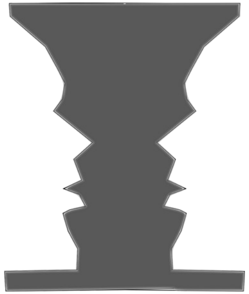
# Bring on the beads … 2017-2019

We can identify two kinds of discrete system depending on whether the **order** of pieces is important or not.

Word counts in a book ...

**Or ...**

Boxes of unordered coloured beads



Unfolded proteins ...

**Or ...**

Strings of ordered coloured beads

# Two Types of Discrete System

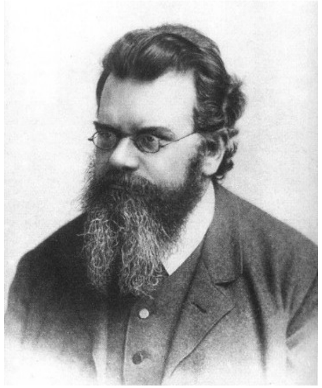
## Heterogeneous and Homogeneous

In **heterogeneous** systems the tokens (beads in this example) are ordered sequentially into strings (each string is a *component*).  Examples include proteins, genomes, software, texts & music

In **homogeneous** systems the tokens are grouped by type. The order or sequence of tokens is irrelevant (there is no implied order) and each bin (*component*) has tokens of a different property.

# Enter Boltzmann and Hartley

- Conservation principles means Ludwig Boltzmann's extraordinary 19th century invention of statistical mechanics
- We need something logarithmically additive *which is independent of any meaning in the tokens* – Ralph Hartley's 1928 *information content.*

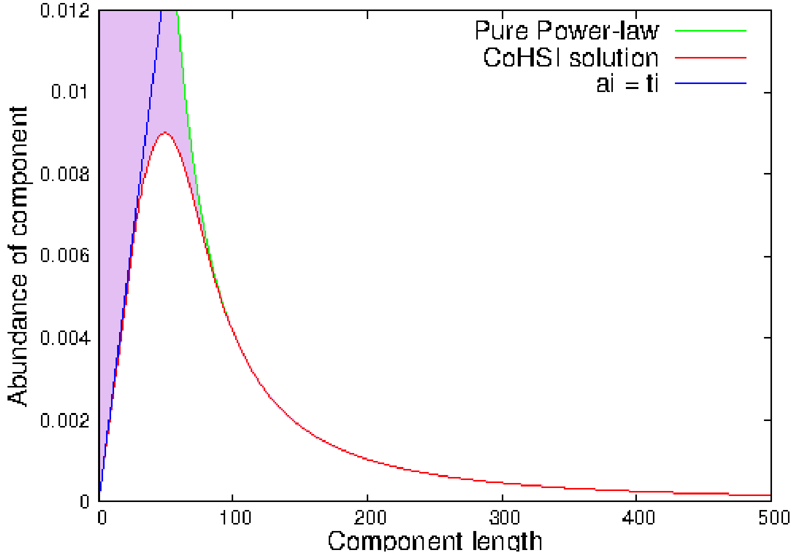


# Boltzmann + Hartley = CoHSI

$$\log t_i = -\alpha - \beta \left( \frac{d}{dt_i} \log N(t_i, a_i; a_i) \right)$$

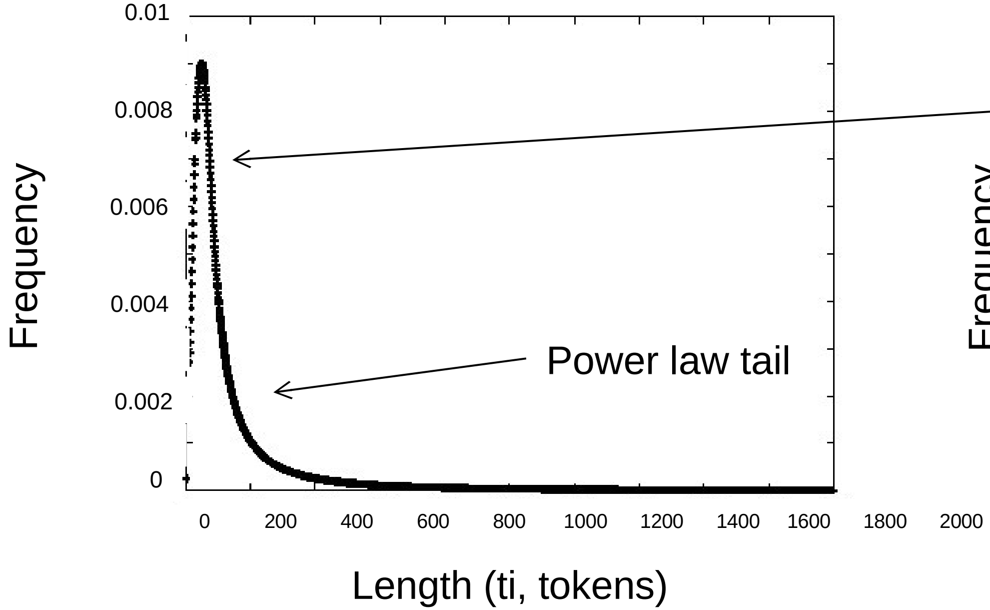
*α* and *β* are the undefined Lagrangian parameters
is the number of tokens in a component (= 9 for ***professor***)
of a component (= 6 for ***professor***)
*N* = number of ways of arranging tokens from
a unique alphabet so that there is at least one
of each and their order is distinguishable.
https://royalsocietypublishing.org/doi/full/10.1098/rsos.191101

# CoHSI for heterogeneous systems

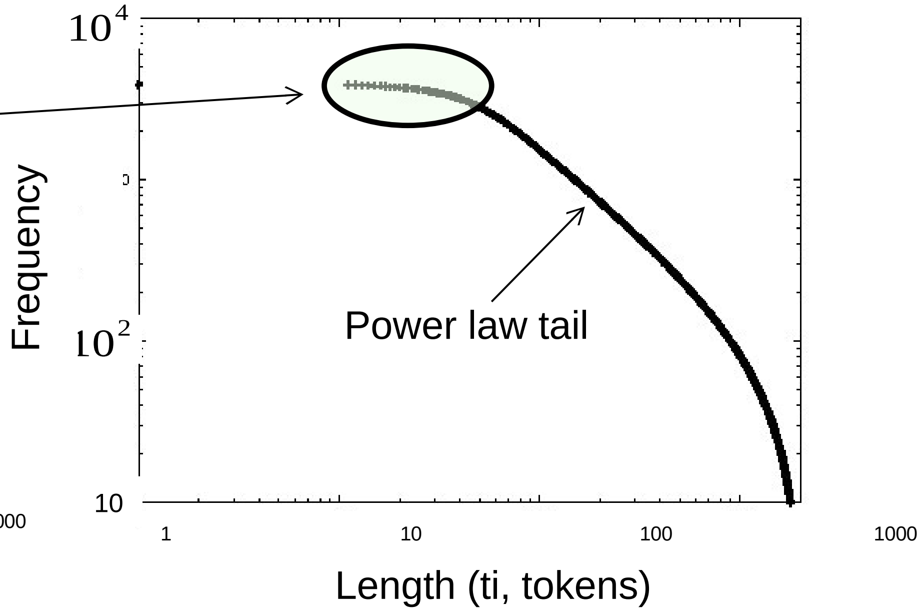


# CoHSI for heterogeneous systems

PDF




CCDF

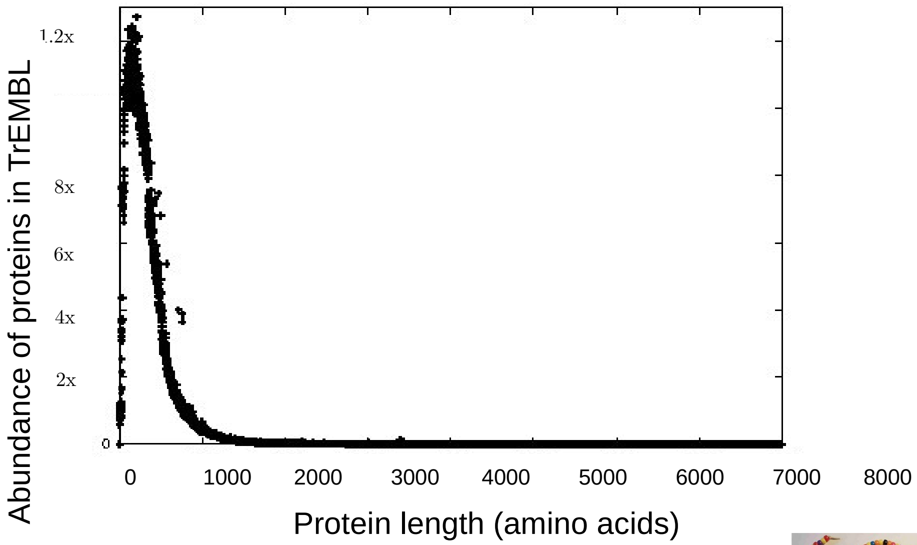


# CoHSI for heterogeneous systems: some predictions

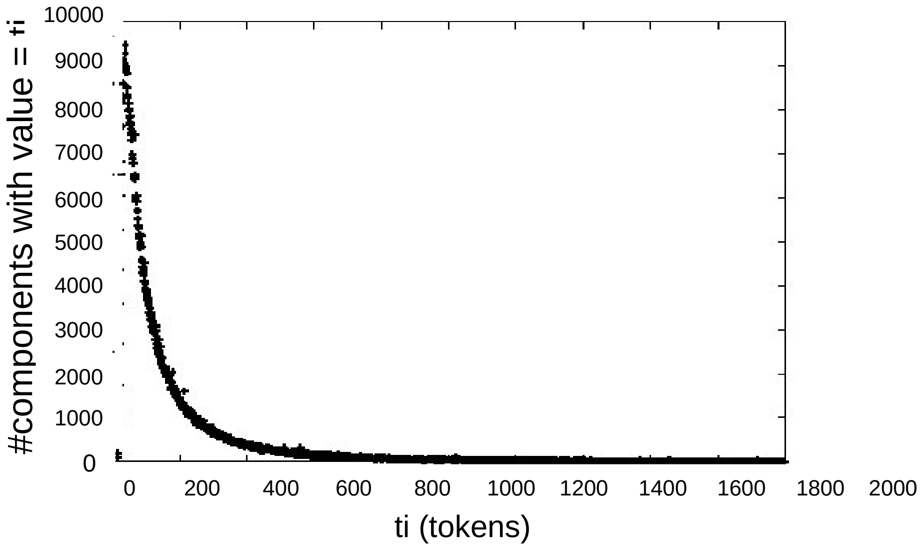
- The length distributions of proteins should be identical to software component distributions.
- Very long proteins are inevitable and its nothing to do with Natural Selection.

# Testing Predictions – Proteins and Software lengths

200 million Proteins

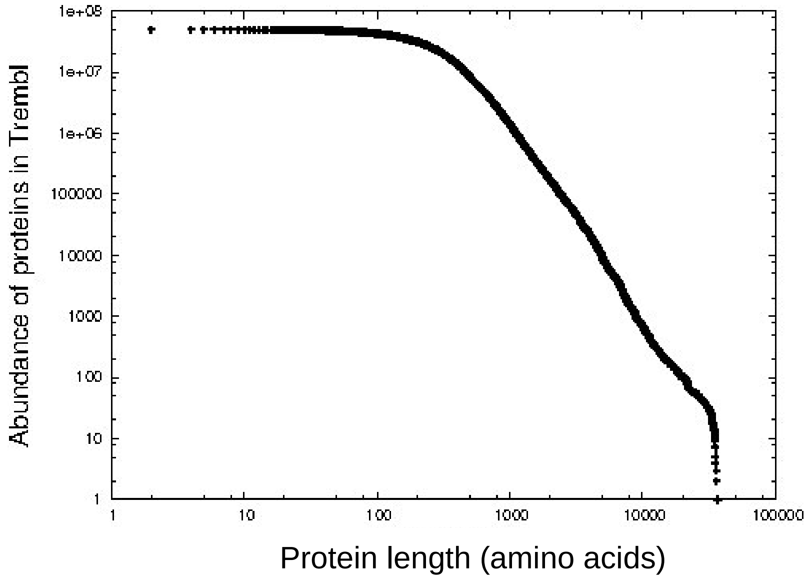



100 Million Lines of Code in 7 Programming Languages

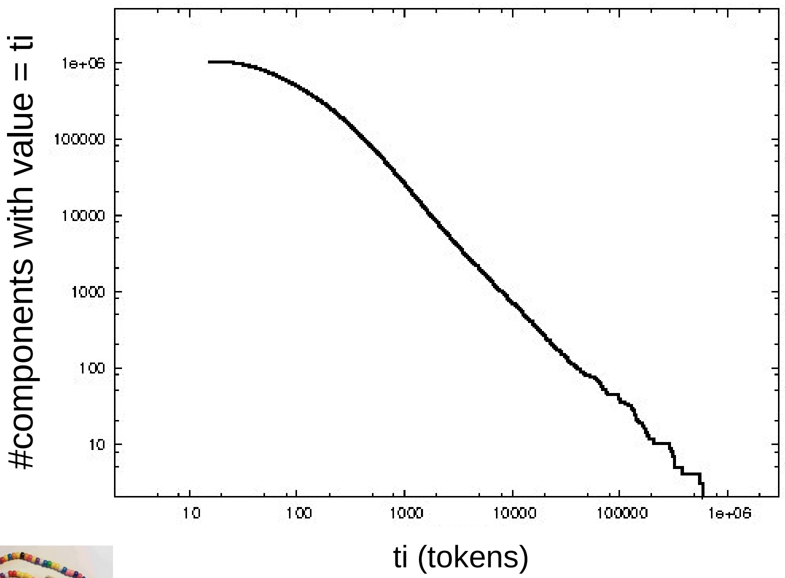



# Testing Predictions – Proteins and Software lengths in log-log

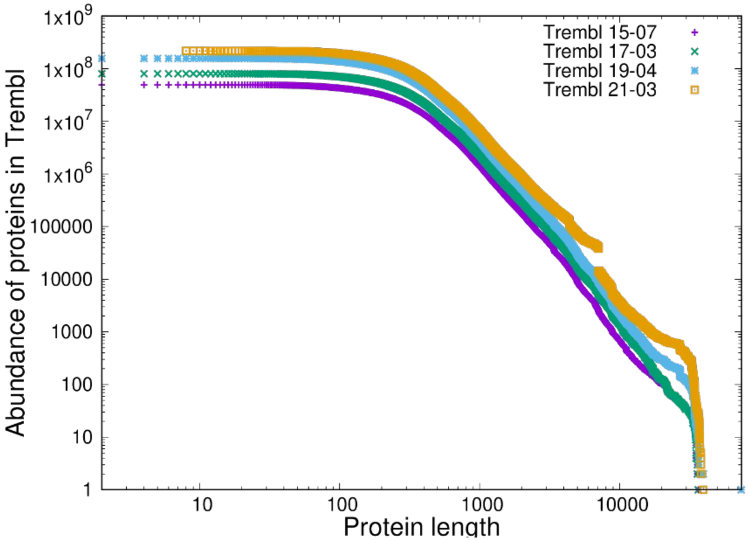
All Known Proteins




100 Million Lines of Code in 7 Programming Languages

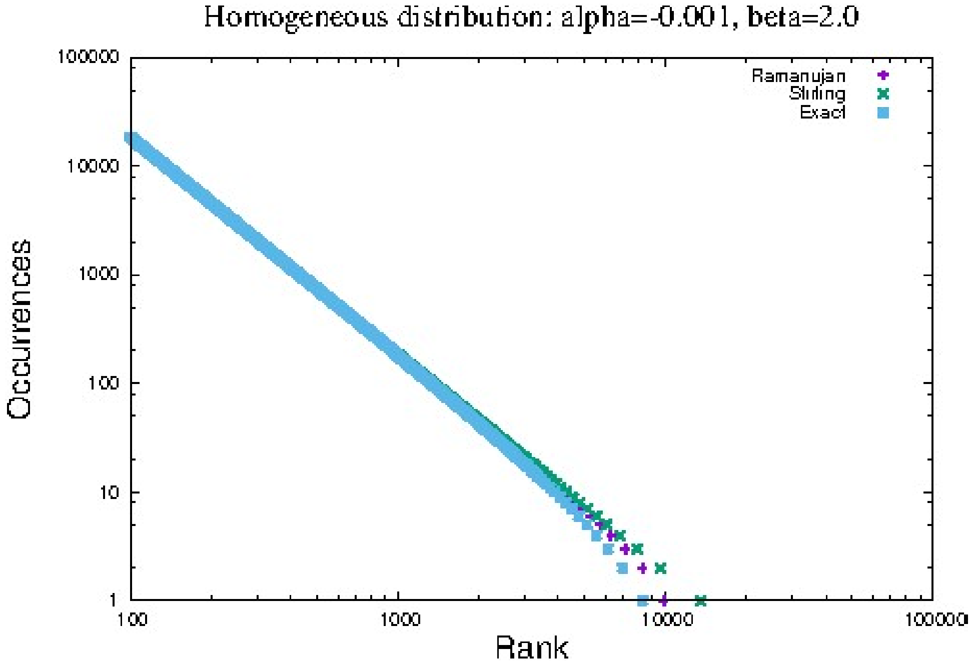



# Testing Predictions – Scale-independence and maximum length




Growth in known set of proteins
2015-2021

# CoHSI for homogeneous systems

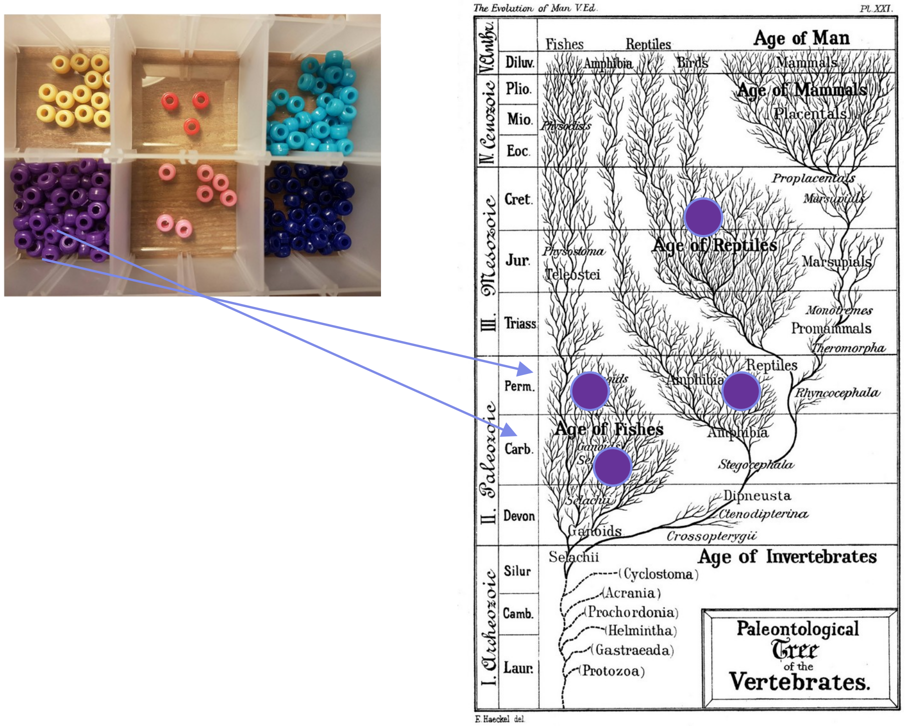



Some choice examples follow …

# HGT: Protein copying round the tree of life and viruses

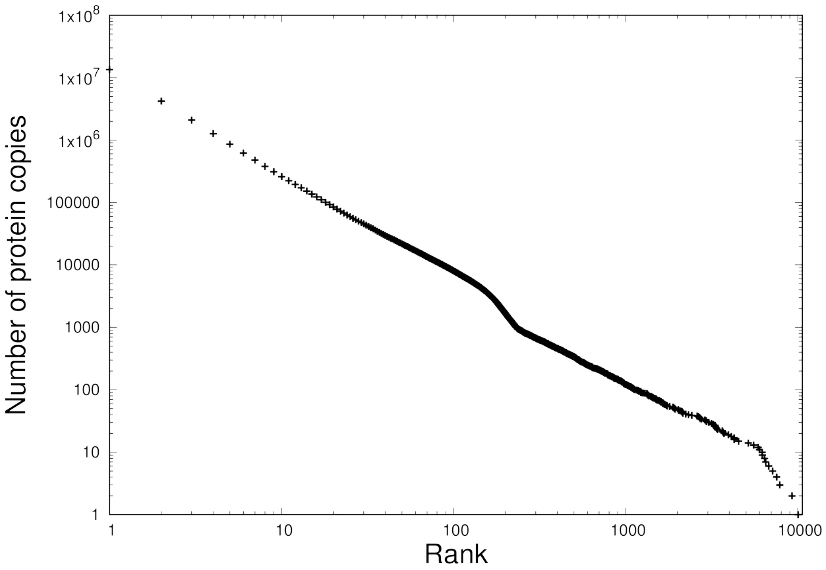
The only requirements for the homogeneous model is that bins do not overlap (each bead is distinguished only by its colour) and that the whole system is represented (1 bead for every discrete piece).

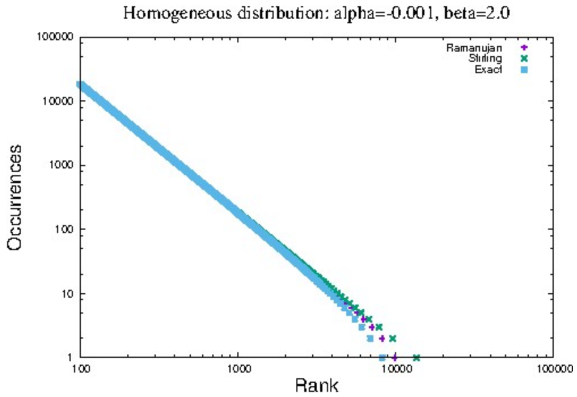
**Each bead is a protein and a unique colour is used if a protein appears N times in the trees of life and viruses**, N=1,..,10000 (or so in TrEMBL 21-03).



# HGT: Protein copying round the tree of life and viruses

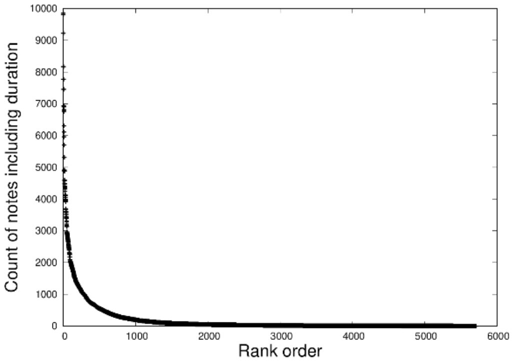
Version 21-03 of full protein database



# Music and homogeneous behaviour

Bach chorales + assorted other pieces





# Conclusions

- All discrete systems share the same asymptotic distributions **when tokens have no intrinsic meaning.**
- There are two obvious information models (bins and strings).
  - **Bins** lead to a *droopy Zipf* proving Zipfian behaviour is a general property.  This is found in distribution of moon craters, earthquake sizes, textual analysis of books, music – the list is endless.
  - **Strings** lead to a *spiky distribution following by an extremely precise power-law*. This is found in the distribution of wealth, extensively in the known set of genomes and proteins and in large distributions of software and many other systems.
- Both distributions are scaleable, can be used to make predictions and both feature gross inequality; **the Fundamental Law of Inequality**.

# References

Introduction ...


- Les Hatton and Gregory Warr (2019), "Strong evidence of an information-theoretical conservation principle linking all discrete systems", Royal Society Open Science, https://doi.org/10.1098/rsos.191101


General Background ...


- Les Hatton and Gregory Warr (2022), "Exposing Nature's Bias: The Hidden Clockwork behind Society, Life and the Universe", Bluespear Publishing, ISBN 978-1908422040


Images

- With grateful thanks to Wikipedia, reproduced either as public domain or Creative Commons.