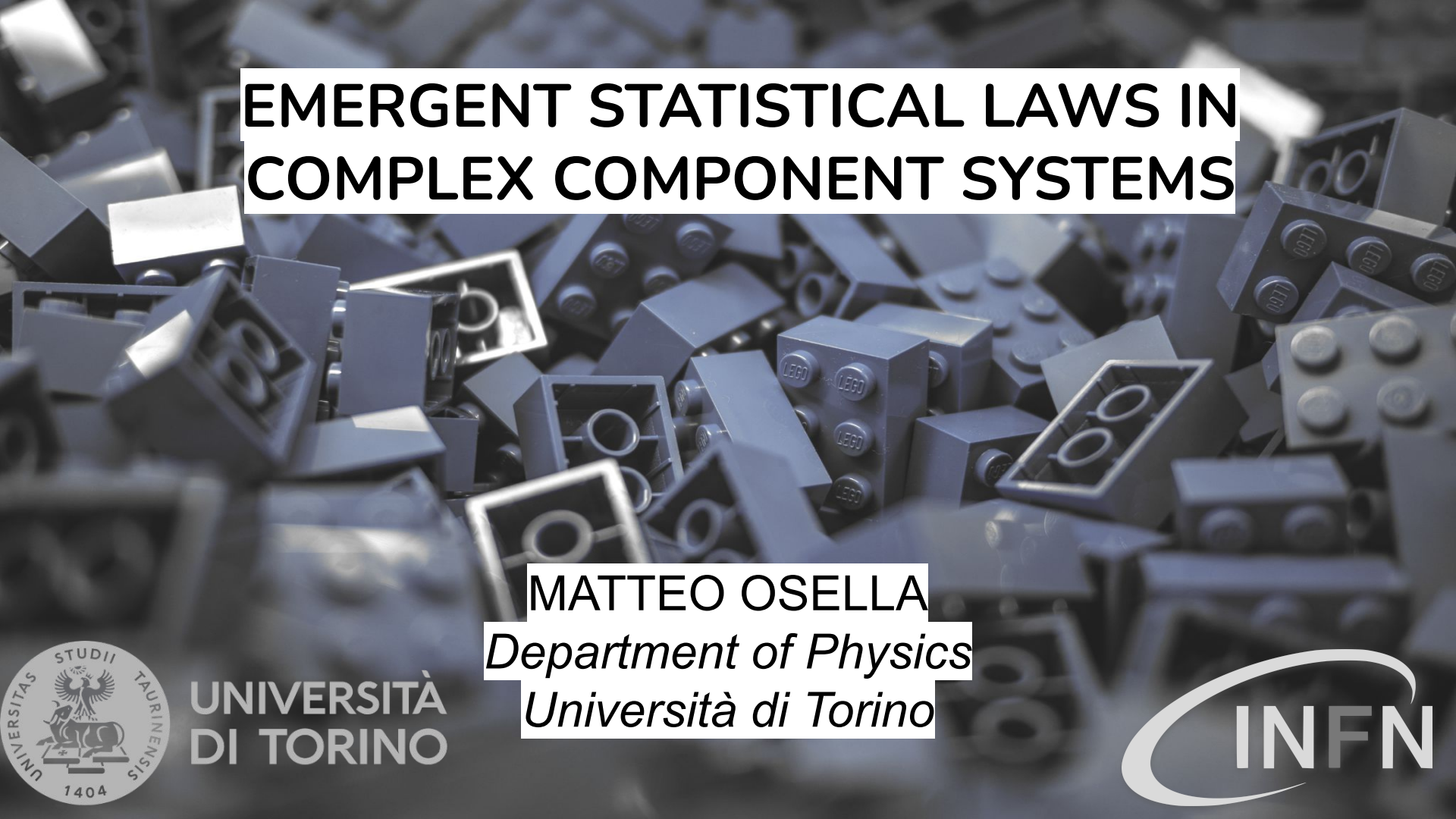
EMERGENT STATISTICAL LAWS IN COMPLEX COMPONENT SYSTEMS



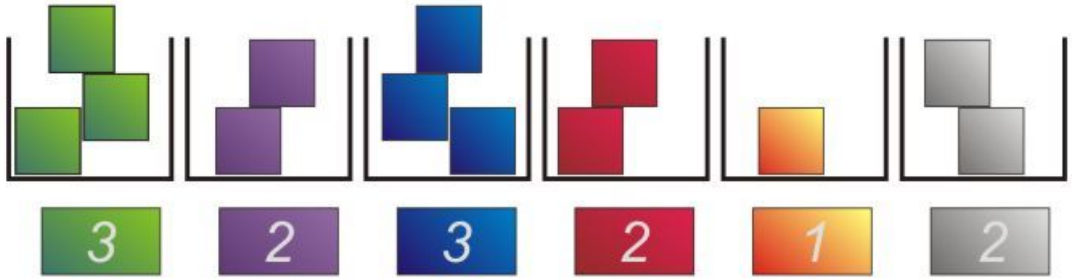
MATTEO OSELLA
Department of Physics
Università di Torino

UNIVERSITÀ DI TORINO

INFN

Andrea Mazzolini, Marco Cosentino Lagomarsino, Marco Gherardi, Jacopo Grilli, Michele Caselle, Filippo Valle, Silvia Lazzardi

Genome



Gene count

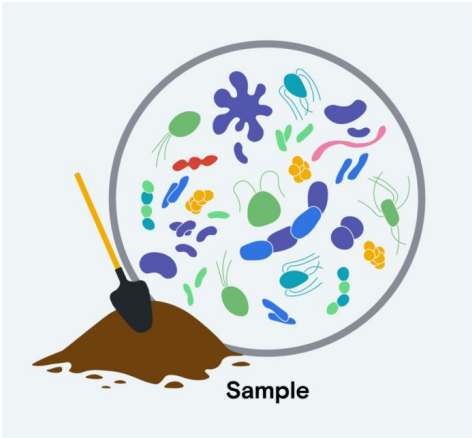
genomes

gene families

	G1	G2	G3	G4	G5
	2	1	3	1	1
	1	0	0	1	0
	0	3	0	3	0
	1	1	1	1	1
	0	0	0	0	1

TRANSCRIPTOMICS

METAGENOMICS



CELL / SAMPLE

GENE / SPECIES

	G1	G2	G3	G4	G5
	2	1	3	1	1
	1	0	0	1	0
	0	3	0	3	0
	1	1	1	1	1
	0	0	0	0	1

LINGUISTICS

“TOOLBOXES”

SOFTWARE PACKAGES

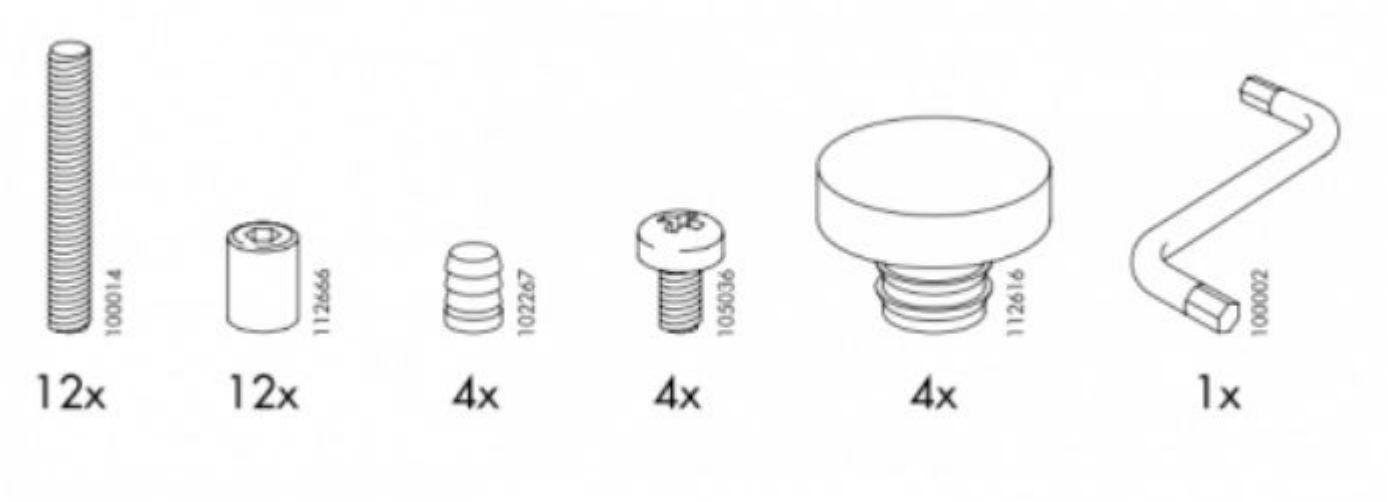
COMPONENT SYSTEMS

REALIZATIONS

COMPONENTS

	G1	G2	G3	G4	G5
	2	1	3	1	1
	1	0	0	1	0
	0	3	0	3	0
	1	1	1	1	1
	0	0	0	0	1

COMPONENT STATISTICS AND FUNCTION

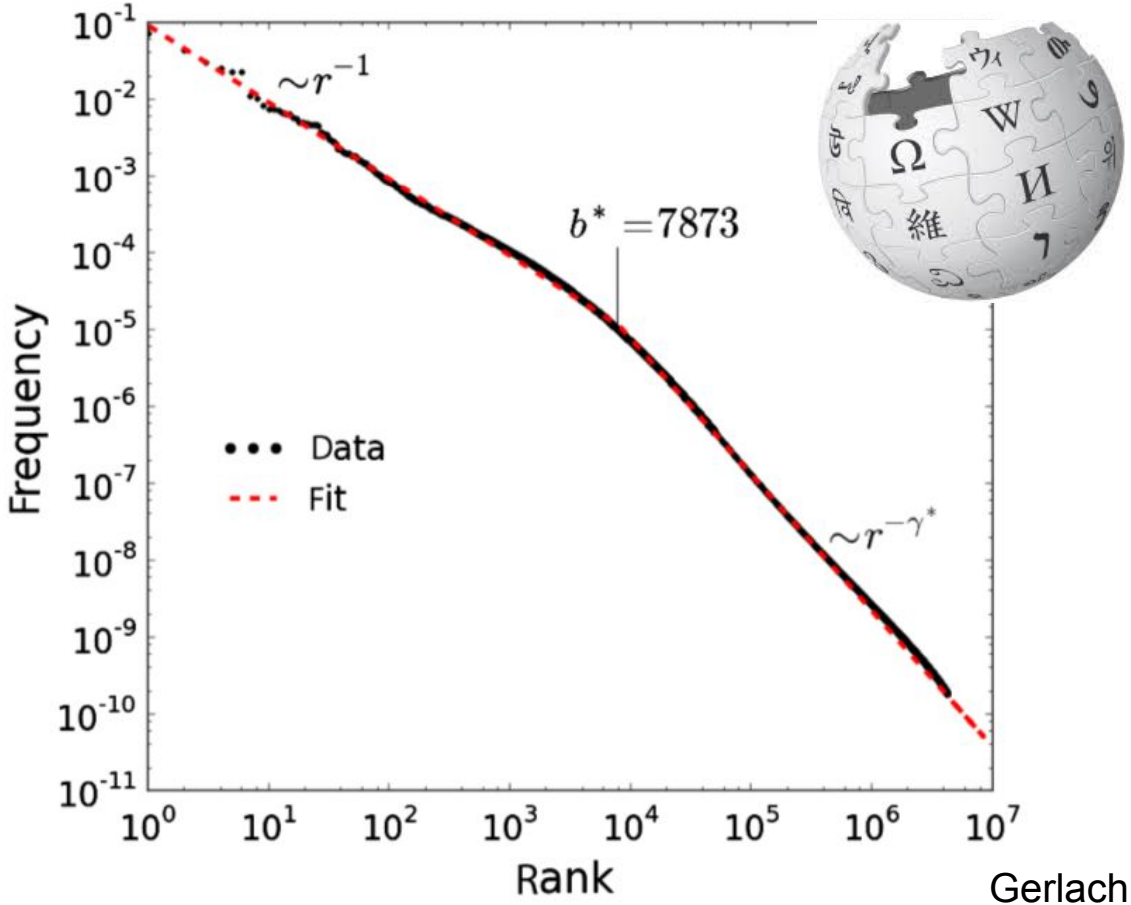


EMPIRICAL STATISTICAL PATTERNS

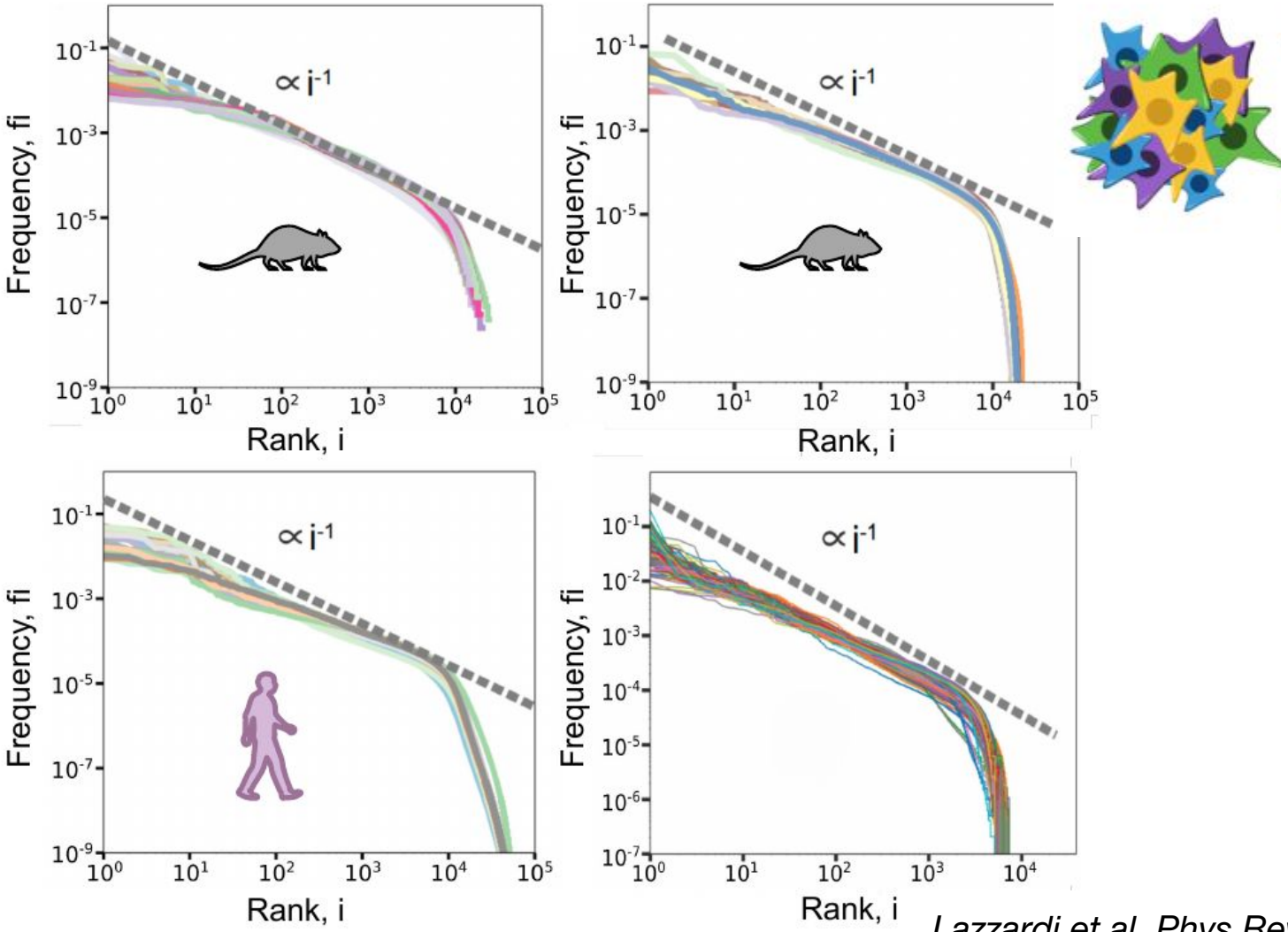
COMPONENT ABUNDANCES

	G1	G2	G3	G4	G5	
	2	1	3	1	1	5
	1	0	0	1	0	2
	0	3	0	3	0	2
	1	1	1	1	1	5
	0	0	0	0	1	1

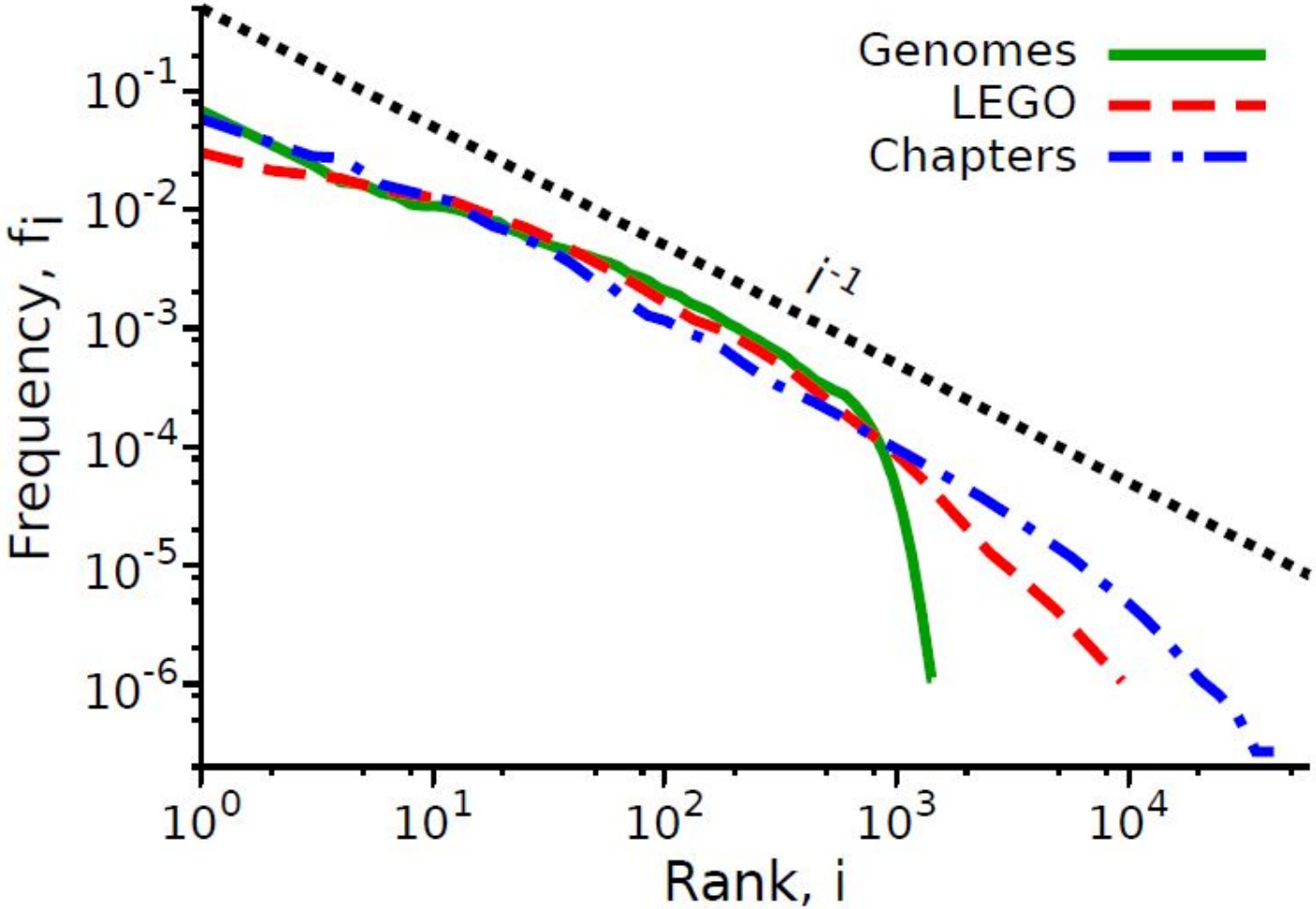
ZIPF'S LAW (1930s)



Gerlach, Altmann PRX 2013



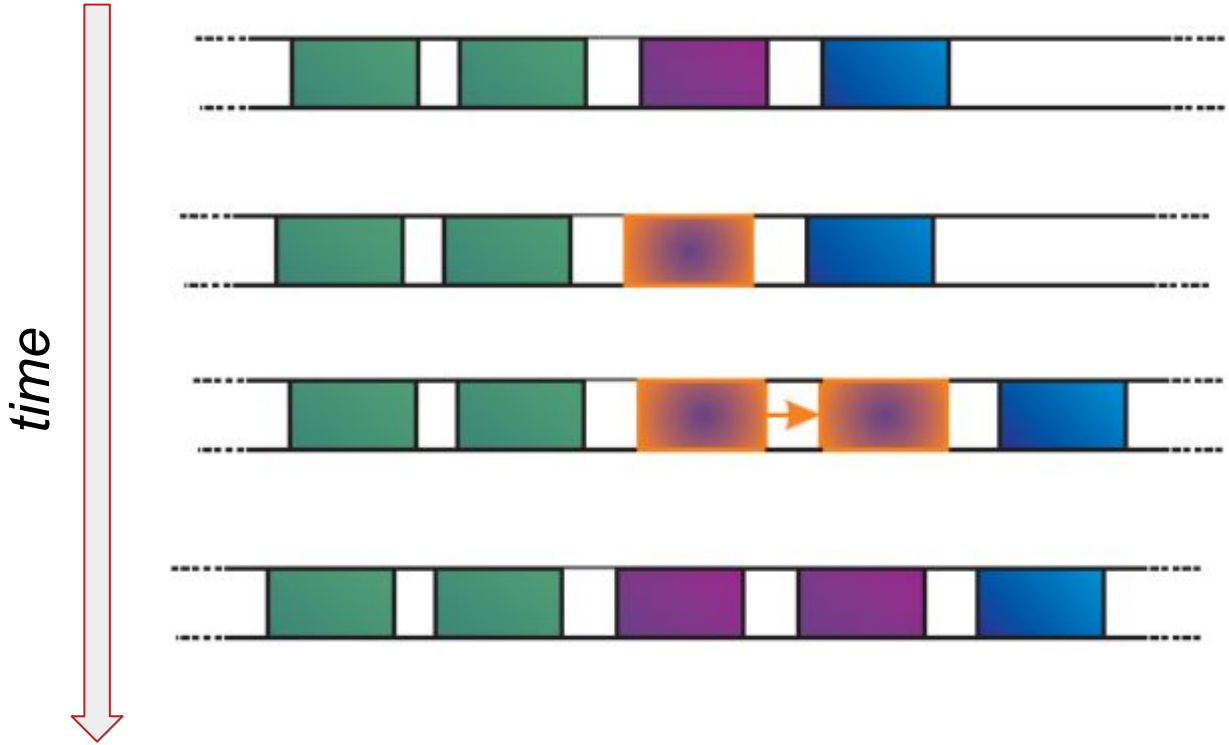
Lazzardi et al. Phys Rev E 2023



Mazzolini et al. Phys Rev X 2018

MODELS FOR THE EMERGENCE OF ZIPF'S LAW

Evolution by gene duplication



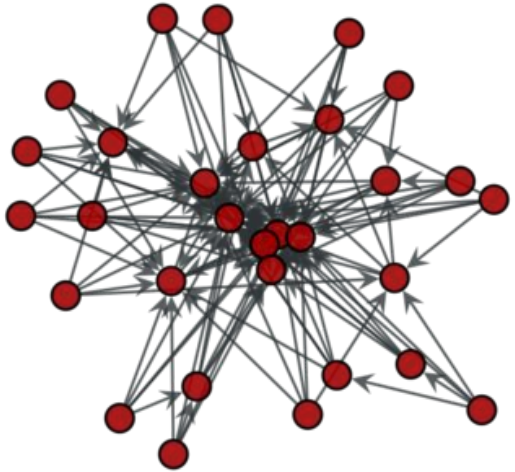
$$p^i_{old} \sim \frac{n^i}{m}$$

Random Duplications

“The rich get richer”

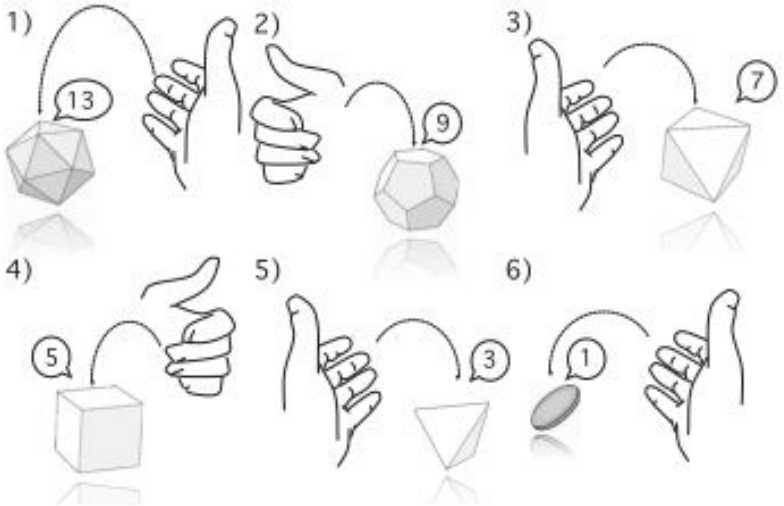
Zipf’s law

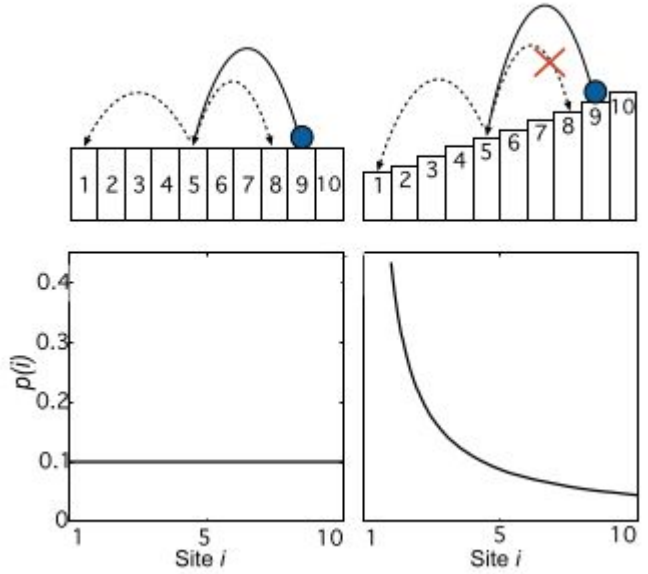
Polya Urn Model



Preferential Attachment

Sample-space-reducing process

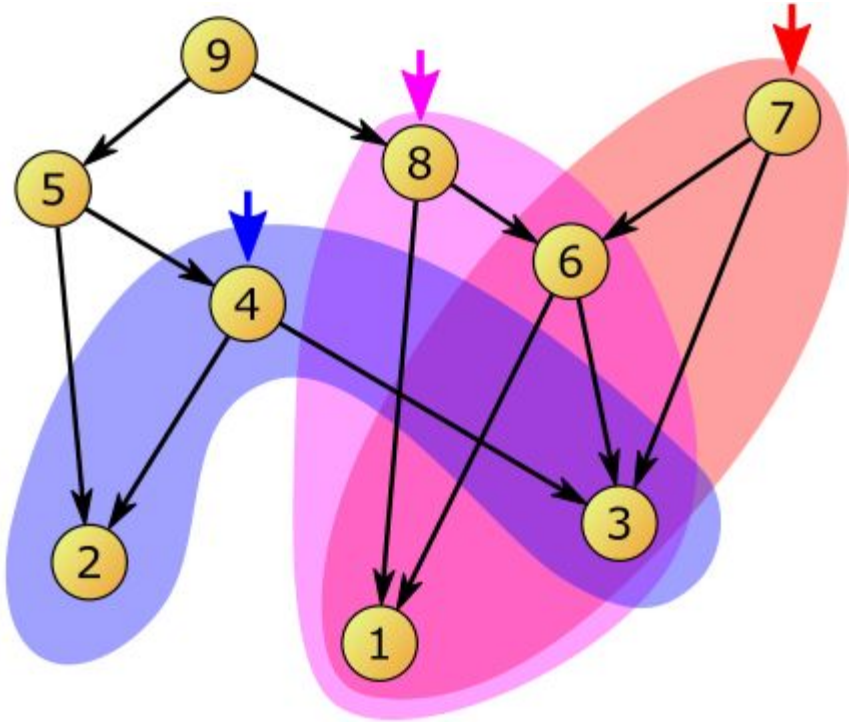


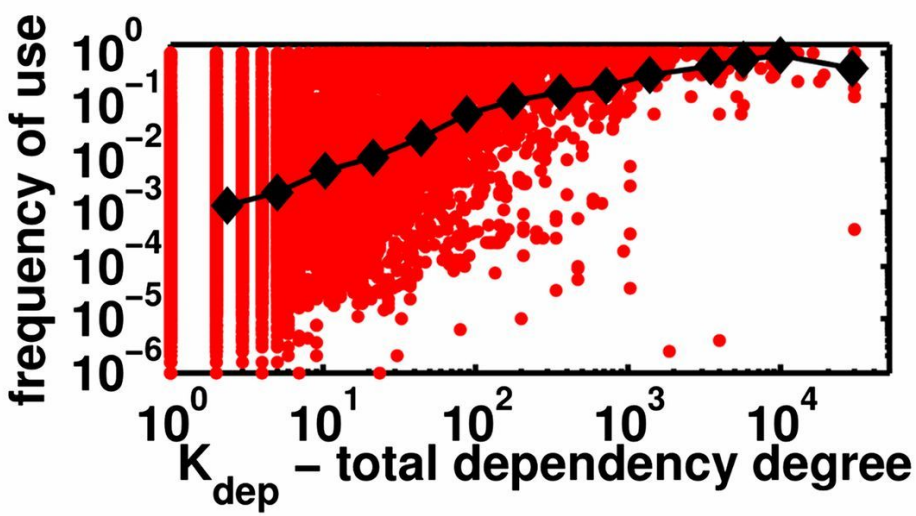


$$f_i = \frac{1}{\alpha} i^{-\gamma}$$

Corominas-Murtra et al PNAS 2015
Mazzolini et al Phys Rev E 2018

Models with dependency structures



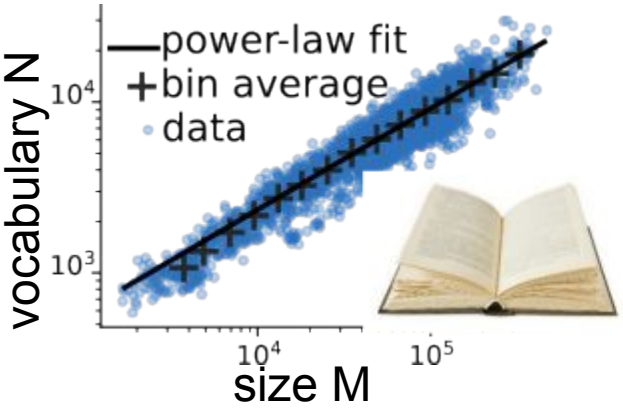


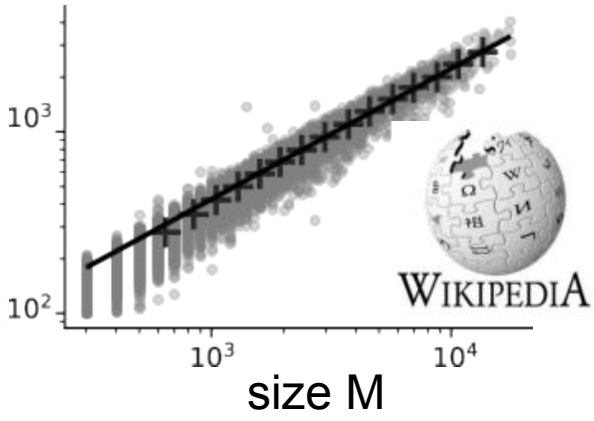
Pang, Maslov PNAS 2013
Mazzolini et al. Phys Rev E 2018

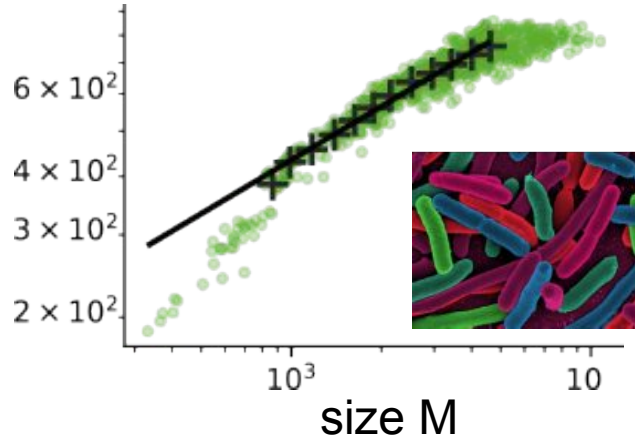
and alternative models based on latent variables

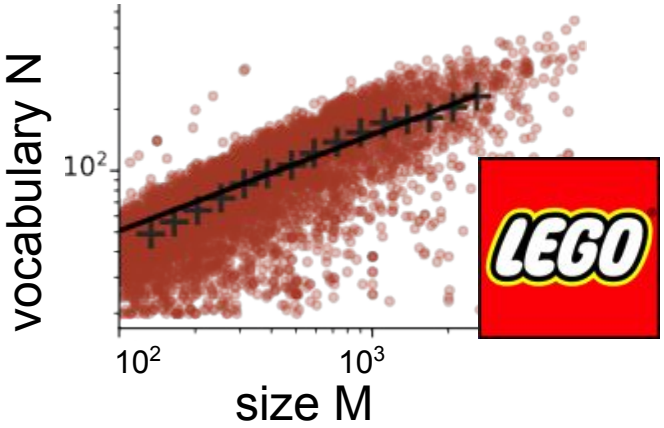
Schwab, Nemenman, Metha PRL 2014
Aitchison, Corradi, Latham PLOS CB 2014

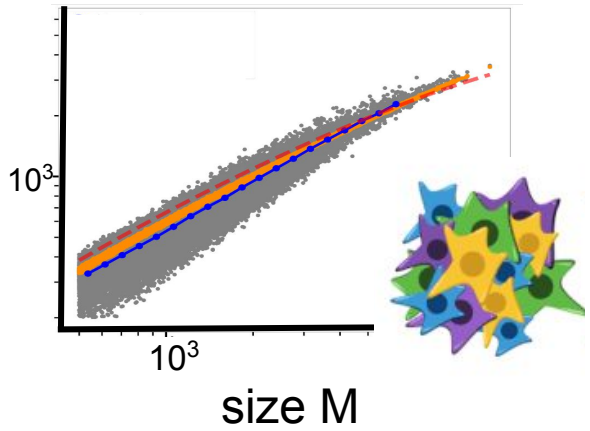
What about “innovation”?









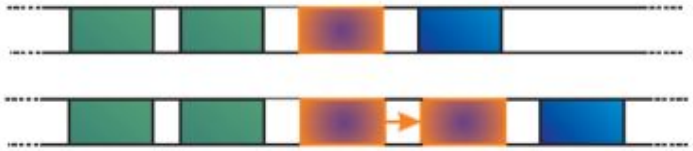


Heaps' law $N \propto M^{\alpha}$

Heaps’ law and innovation dynamics in growth models

duplication

innovation

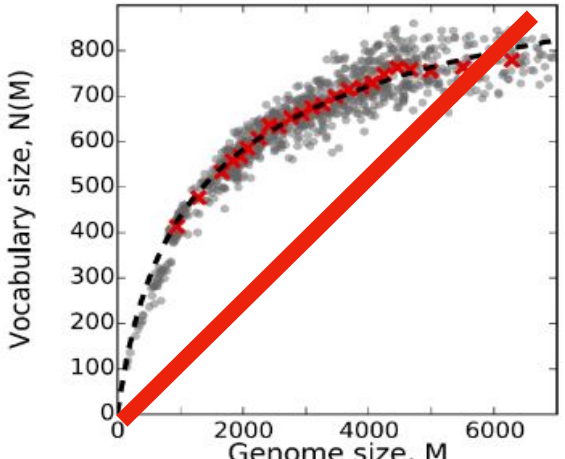


$$p_{new} \sim \text{CONSTANT}$$

$$p_{old}^{i} \sim \frac{n^i}{m}$$

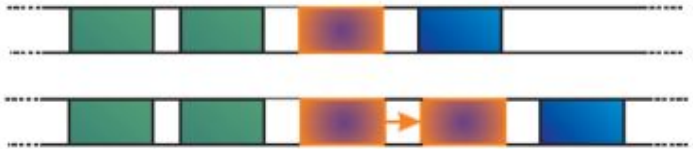
$$p_{old} = \sum_i p_{old}^{i}$$

$$p_{old} + p_{new} = 1$$



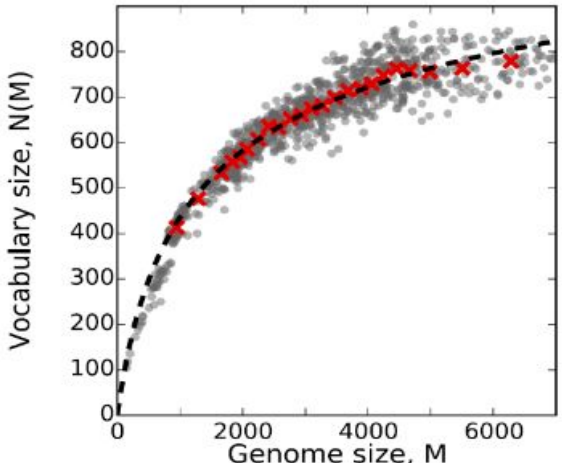
duplication

innovation

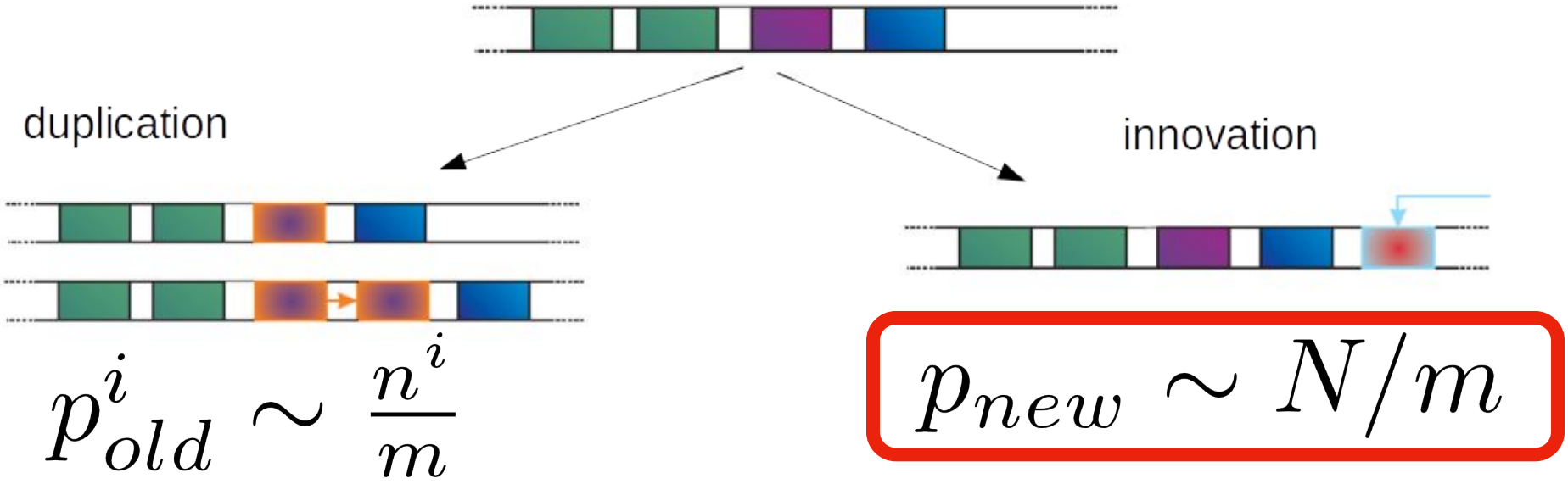


$$p^i_{old} \sim \frac{n^i}{m}$$

$$p_{new} \sim m^{-\mu}$$



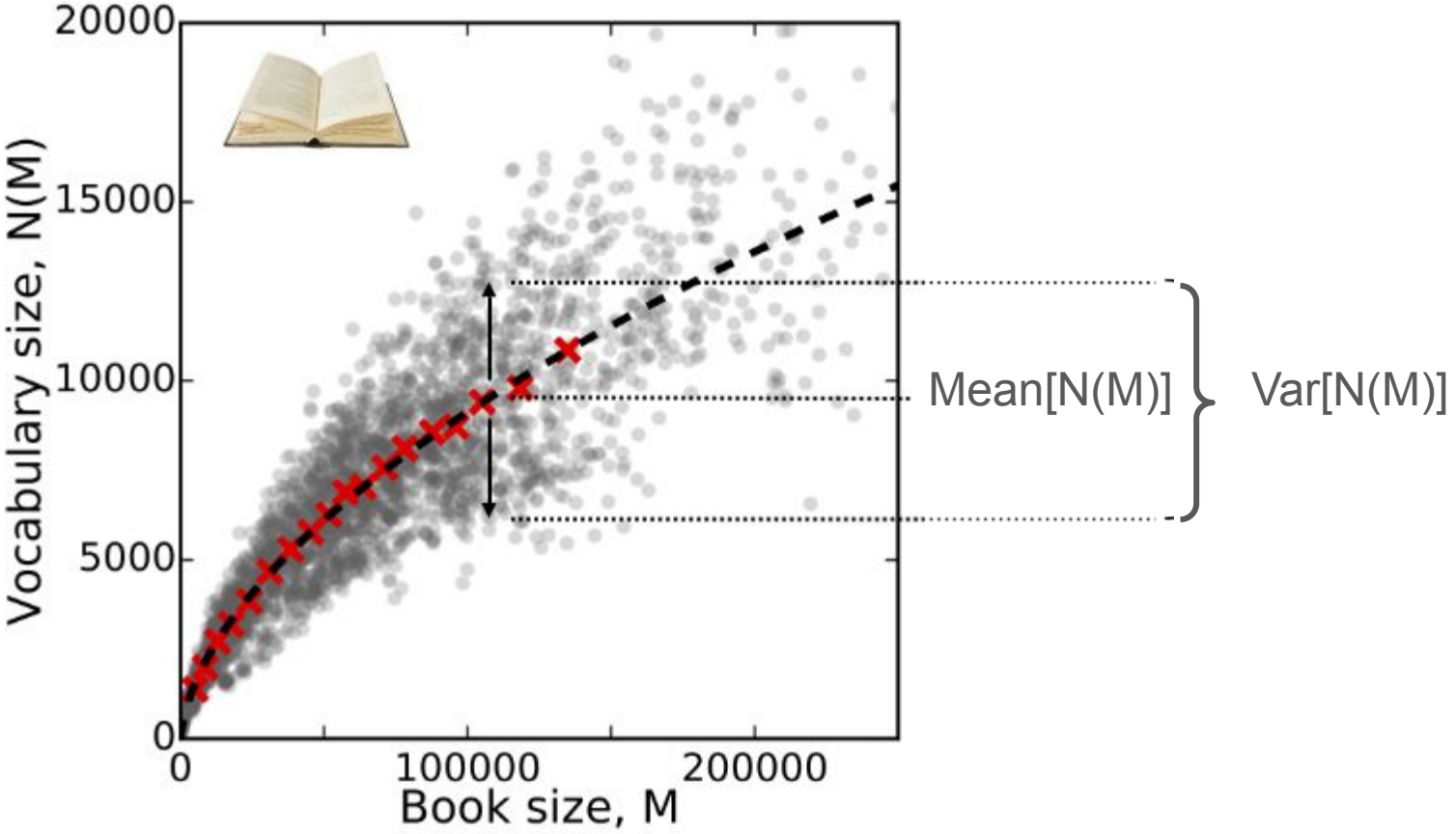
Generalized Simon's Models



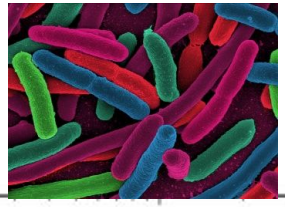
Chinese Restaurant Process;
Urn model with semantic triggering

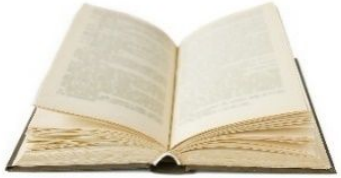
Cosentino Lagomarsino et al Genome Biol 2009
Tria et al. Sci Rep 2014

FLUCTUATION SCALING
(AKA “DIVERSITY”)

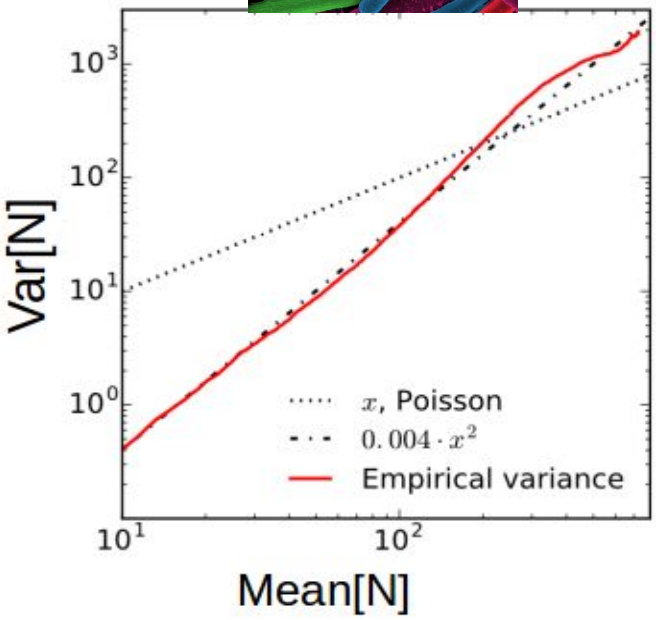

20000
15000
10000
5000
0
Vocabulary size, N(M)
0
100000
200000
Book size, M
Mean[N(M)]
Var[N(M)]

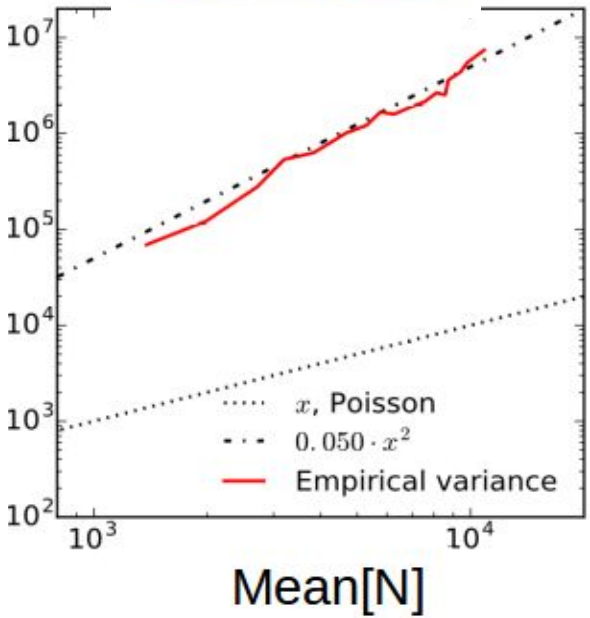
Taylor's Law

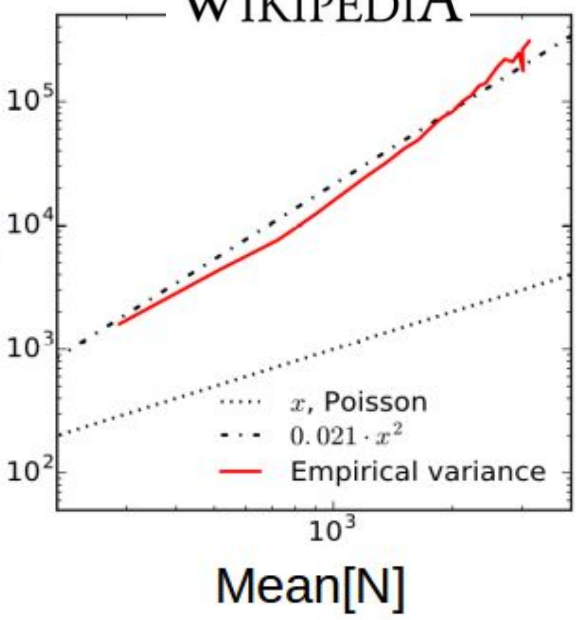


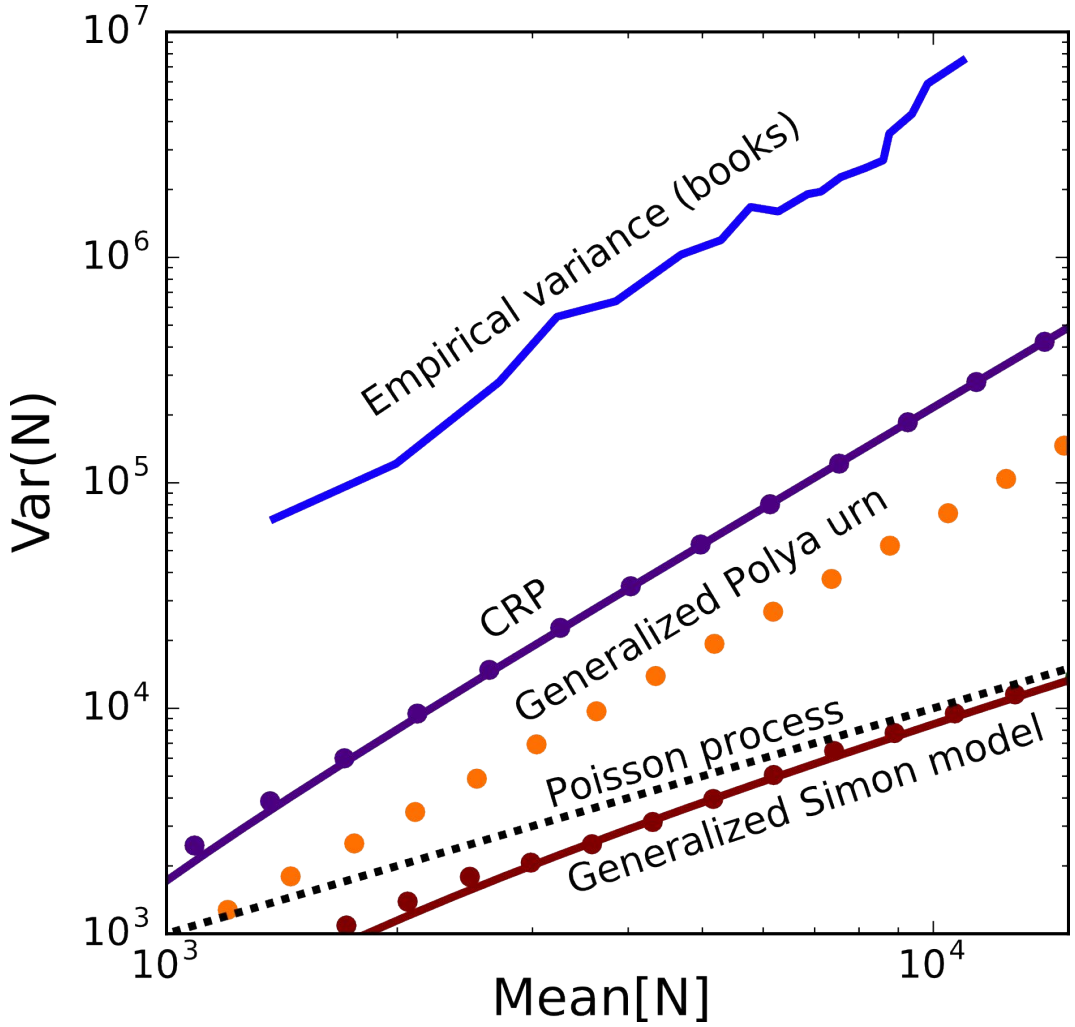



WIKIPEDIA


Var[N]
Mean[N]
x, Poisson
$0.004 \cdot x^2$
Empirical variance


Mean[N]
x, Poisson
$0.050 \cdot x^2$
Empirical variance


Mean[N]
x, Poisson
$0.021 \cdot x^2$
Empirical variance

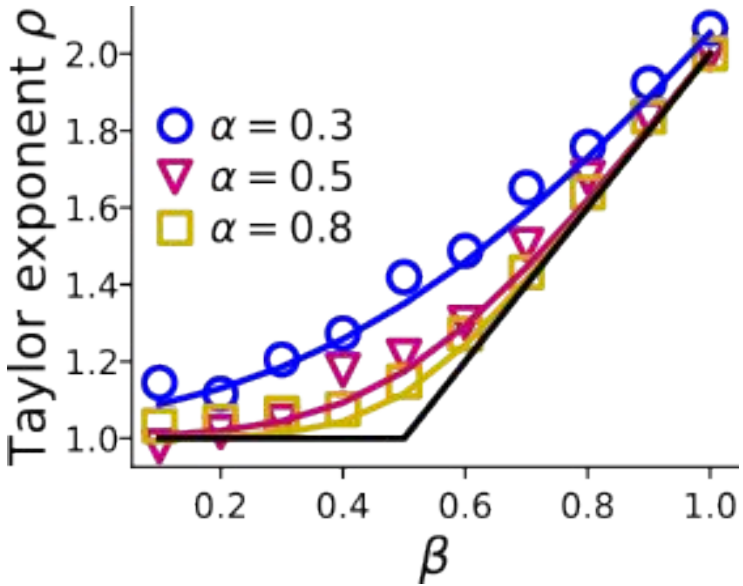

10^7
10^6
10^5
10^4
10^3
10^3
10^4
Var(N)
Mean[N]
Empirical variance (books)
CRP
Generalized Polya urn
Poisson process
Generalized Simon model

$$p_{new} \sim N/m$$

$$\frac{dp_{new}}{dN} = 0 \Rightarrow Var[N] \leqslant Mean[N]$$

A linear dependence is required for Taylor's law

$$P_{new} = a \frac{N^{\beta}}{m^{\gamma}}$$



Diversity generates diversity

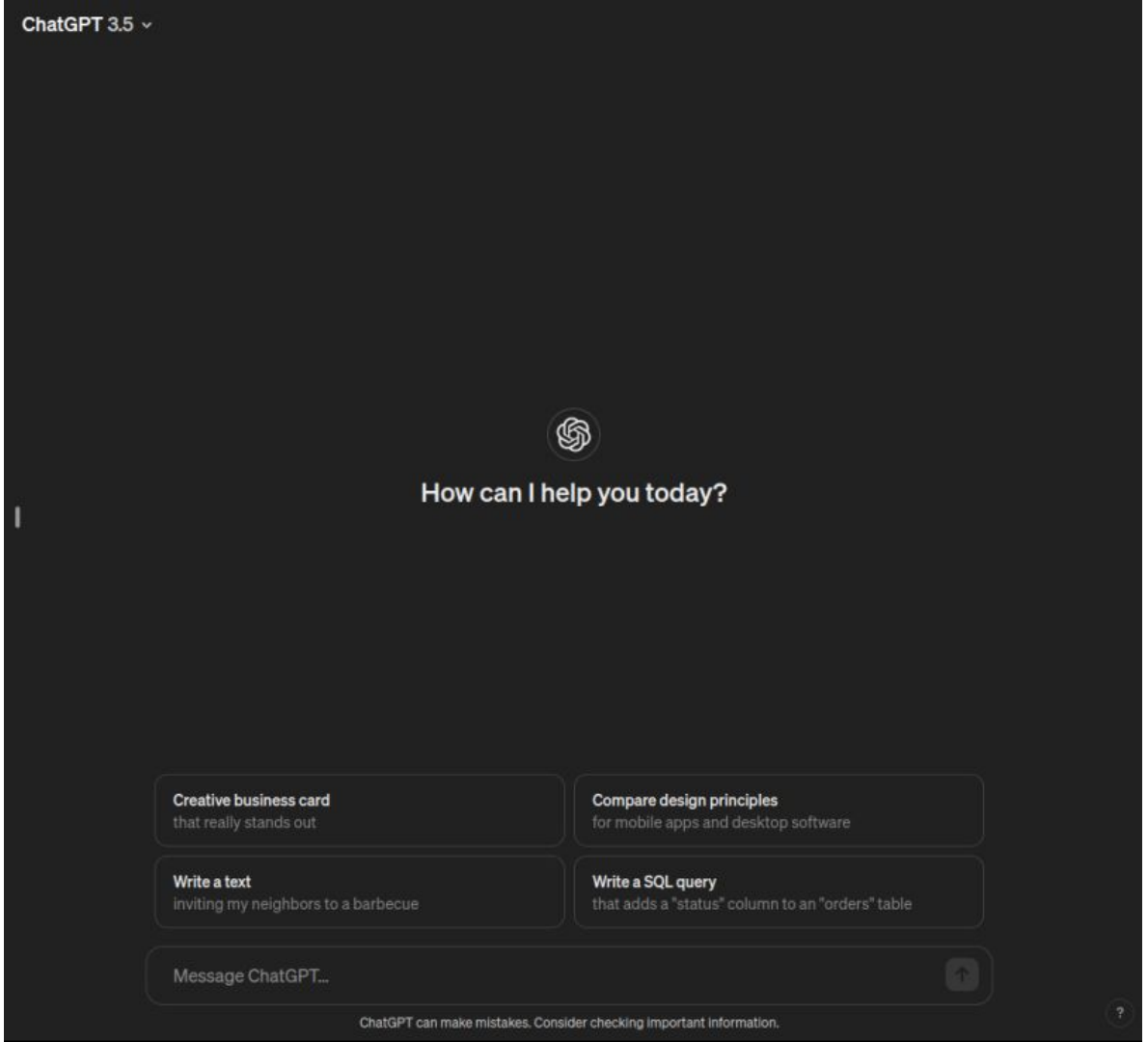
$$p_{new} \sim N$$

A general condition on stochastic processes to reproduce Taylors' law

(Connection with sampling models based on latent variables!)

Diversity generates diversity

Can we test it with a “true” generative process?


ChatGPT 3.5
How can I help you today?
Creative business card
that really stands out
Compare design principles
for mobile apps and desktop software
Write a text
inviting my neighbors to a barbecue
Write a SQL query
that adds a "status" column to an "orders" table
Message ChatGPT...
ChatGPT can make mistakes. Consider checking important information.

Large Language Models (LLMs) trained on next word prediction

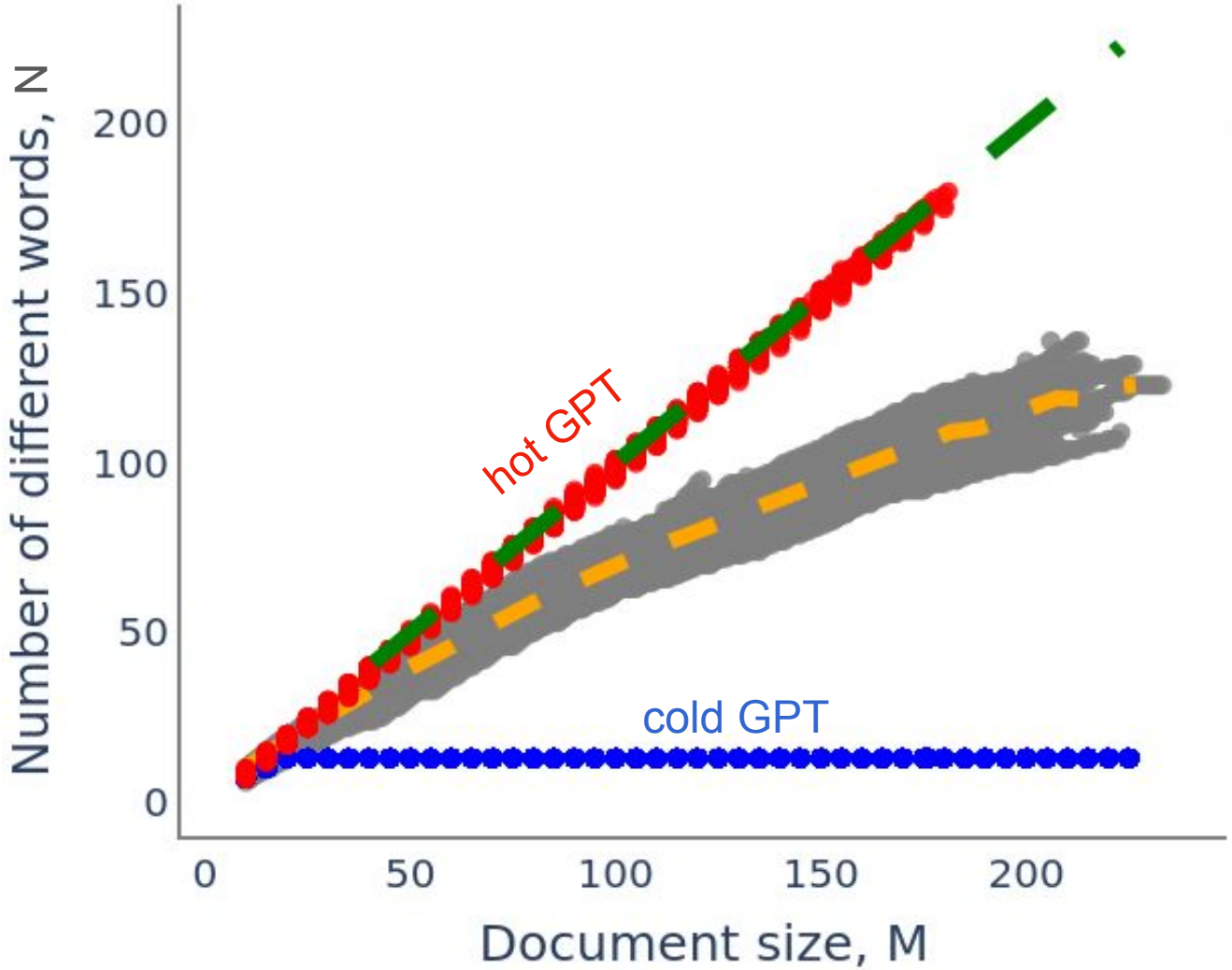
Generative Models

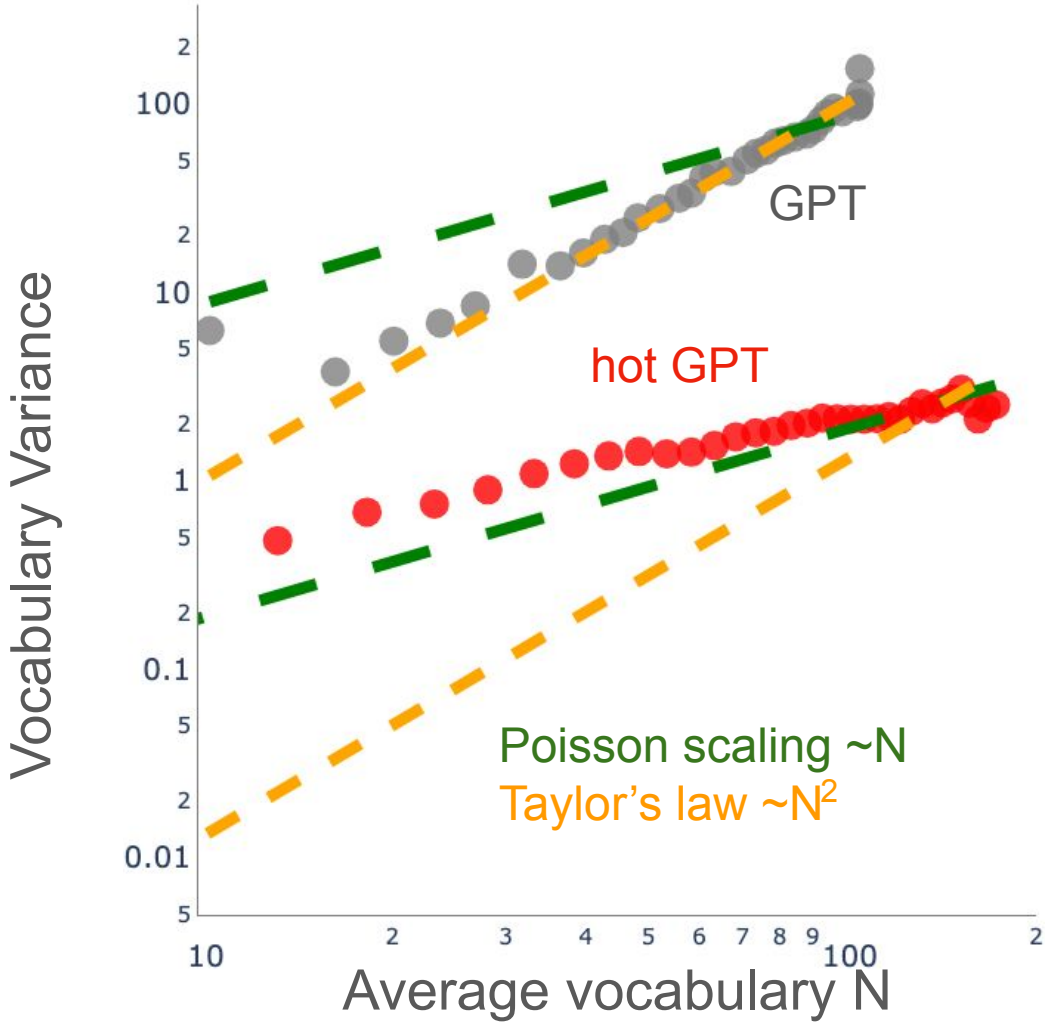
Next word probability distribution as a function of a window of previous tokens

$$\text{softmax}(z)_i = \frac{e^{\frac{z_i}{\tau}}}{\sum_{j=1}^{n} e^{\frac{z_j}{\tau}}}$$

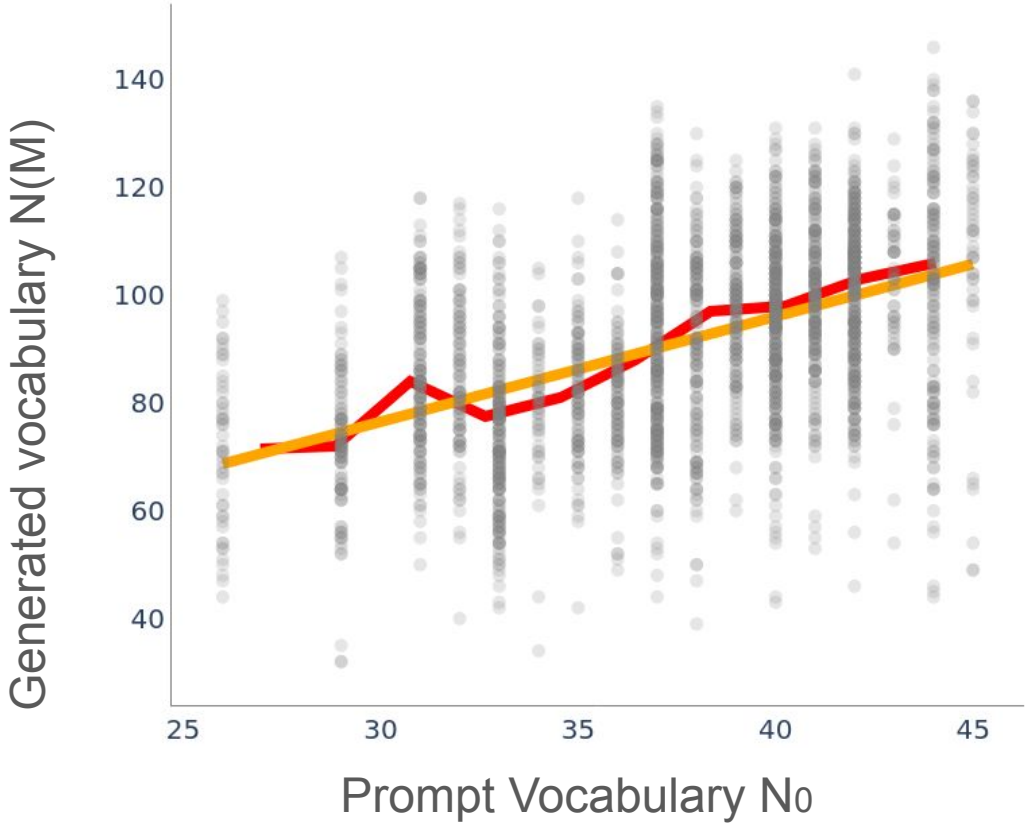
From OpenAI Developer Forum for GPT3

Use Case	Temperature
Code Generation	0.2
Creative Writing	0.7
Chatbot Responses	0.5
Code Comment Generation	0.3
Data Analysis Scripting	0.2
Exploratory Code Writing	0.6


Number of different words, N
0
50
100
150
200
0
50
100
150
200
Document size, M
hot GPT
cold GPT


Vocabulary Variance
2
100
5
2
10
5
2
1
5
2
0.1
5
2
0.01
5
GPT
hot GPT
Poisson scaling ~N
Taylor's law ~N2
10
2
3
4
5
6
7
8
9
100
2
Average vocabulary N

“Diversity generates diversity”



Component systems:

An exciting sandbox for statistical physicists

Simple models can support real-word applications

A two-way exchange: applications drive modelling, and modelling informs applications

Statistics of shared components in complex component systems
A Mazzolini, M Gherardi, M Caselle, M Cosentino Lagomarsino, M Osella
Physical Review X 8 (2), 021023. 2018

Zipf and Heaps laws from dependency structures in component systems
A Mazzolini, J Grilli, E de Lazzari, M Osella, M Cosentino Lagomarsino, M Gherardi
Phys Rev E 98, 012315. 2018

Heaps' law, statistics of shared components and temporal component distribution from a sample-space-reducing mechanism
A Mazzolini, A Colliva, M Caselle, M Osella
Phys Rev E 98, 98, 052139. 2018

Emergent statistical laws in single-cell transcriptomic data
S Lazzardi, F Valle, A Mazzolini, A Scialdone, M Caselle, M Osella
Phys Rev E 107, 04403. 2023

Diversity generates diversity: from genomes to large language models
A Mazzolini, F Valle, L Agasso, M Caselle, M Cosentino Lagomarsino, M Osella
In preparation

FUNDING:
PRIN 2022 - COD. 2022PY8MHN - GeCoS: Genomic Component Systems