Journal of Statistical Mechanics: Theory and Experiment > Volume 2013 > September 2013

Matteo Marsili *et al J. Stat. Mech.* (2013) P09003 doi:10.1088/1742-5468/2013/09/P09003

On sampling and modeling complex systems

Matteo Marsili[1], Iacopo Mastromatteo[2] and Yasser Roudi[3,4]

Or:

arXiv.org > physics > arXiv:1301.3622

Physics > Data Analysis, Statistics and Probability

On sampling and modeling complex systems

Matteo Marsili, Iacopo Mastromatteo, Yasser Roudi

(Submitted on 16 Jan 2013 (v1), last revised 1 Nov 2013 (this version, v4))

Science is a miracle

The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve. We should be grateful for it and hope it will remain valid also in future research and that it will extend, for the better of for the worse, to our pleasure, even though perhaps also to our bafflement, to wide branches of learning (E. P. Wigner 1960)

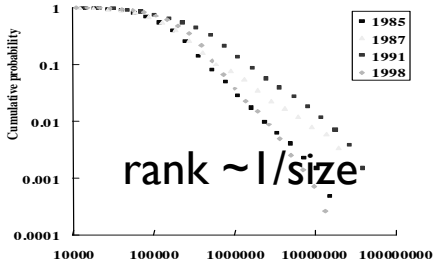
- Galaxies have millions of stars, a piece of material has 10^{32} molecules, ...
  Yet, we understand their behavior in terms of few relevant variables!
- Will this work for a cell (10^4 genes), the brain (10^7 neurons)
  an economy (10^6 individuals)... ?
- We build airplanes. Can we also cure cancer or avoid the next financial crisis?
- Even if the answer is (likely) no, what is the best we can do?
- How to find the relevant variables?

Facts and more questions

- Fact 1:
  Data deluge + advanced experimental techniques (e.g. sequencing)
  yet problems involve a huge number of variables (e.g. 10^4 genes)
  and prediction is hard (e.g. drug design)

- Fact 2:
  We observe "Criticality", as a statistical regularity,
  in a wide variety of different systems as cities,
  the brain, languages, economy/finance, biology.
  Why?

- Questions:
  Are there overarching organizing principles (e.g. SOC)?
  Can we exploit "criticality" (e.g. for model selection)?



(land prices in Japan
Kaizoji & Kaizoji 2006)

P. Bak How Nature Works (1996)
T. Mora & W. Bialek, J.Stat.Phys. (2011)
S. Ki Baek et al. N. J. Physics (2012)

Criticality: Zipf's law

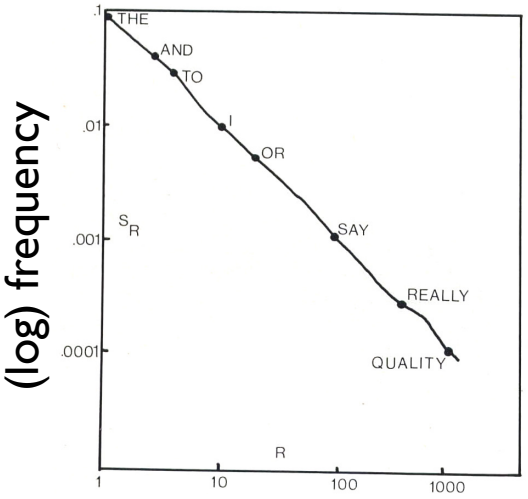
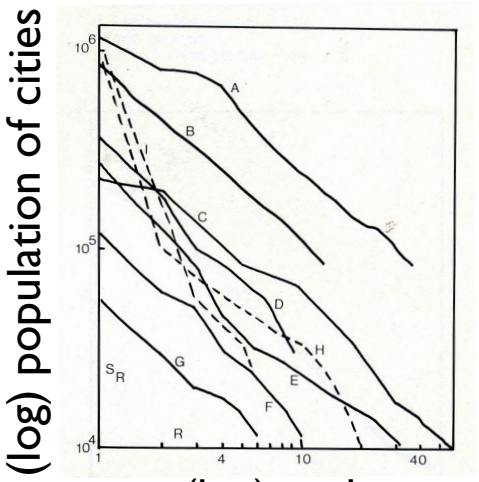


Figure 1 Frequency of word usage in English

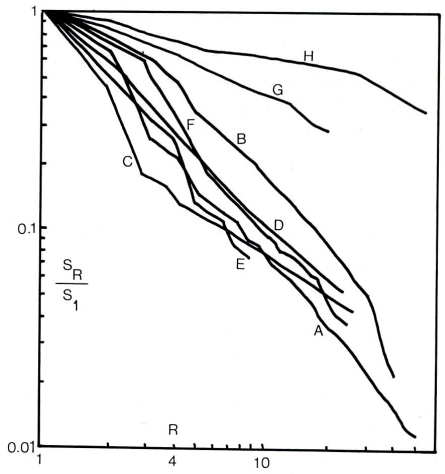
(log) rank

(G. Kirby 1985)



(log) rank

A	United States	B	China
C	West Germany	D	Spain
E	France	F	East Germany
G	Switzerland	H	United Kingdom
I	Mexico		

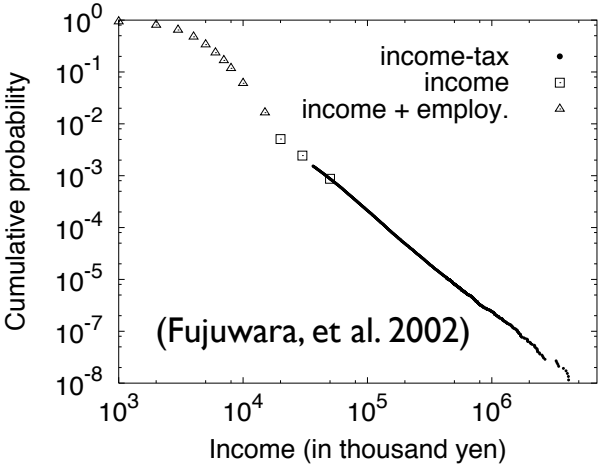


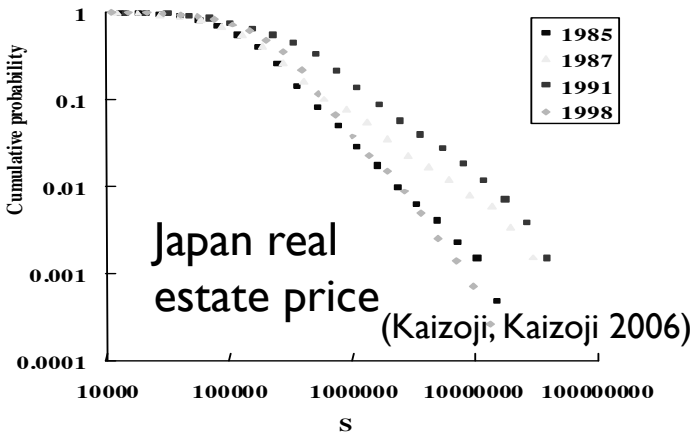
A	Populations of all countries
B	Number of ships built by all countries
C	Students at English universities
D	Building Societies by assets
E	Populations of World's religions
F	US insurance companies by staff
G	World languages
H	English public schools by students

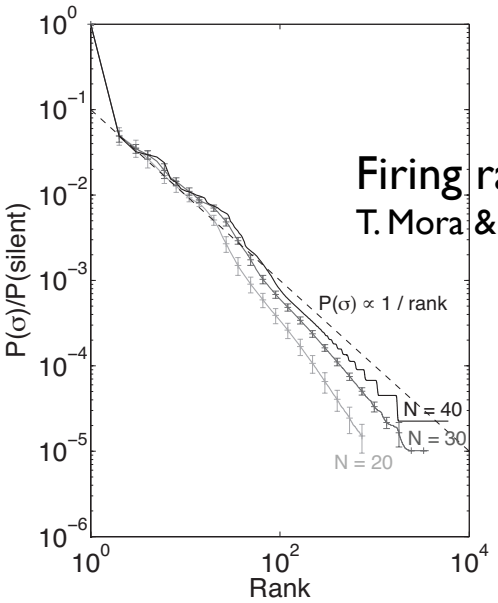
$$\text{rank} \propto \text{size}^{-1} \quad \Rightarrow \quad N(\text{size}) \sim \text{size}^{-2}$$

Proof: $$\int_{\text{size}}^{\infty} dk N(k) = \text{rank}(\text{size})$$

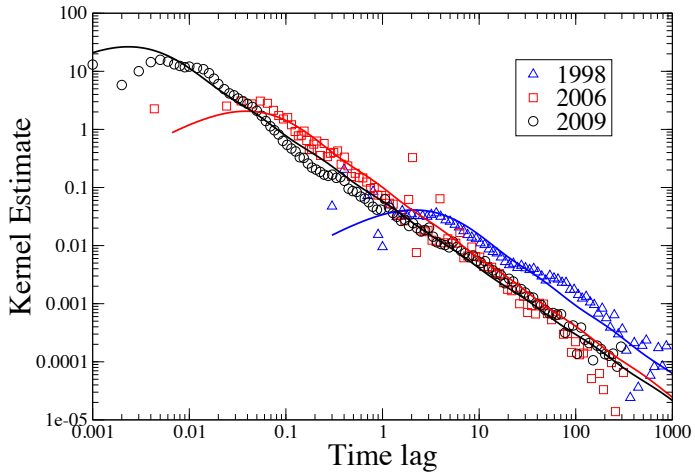
Criticality in economics and the brain







Firing rates of neurons in the retina
T. Mora & W. Bialek, J.Stat.Phys. (2011)



Financial markets as critical branching processes
(Hardiman, Bercot, Bouchaud 2013)

The only things all these phenomena have in common is:
They are samples of solutions to the same optimization problem

E.g. language: an efficient way to say complex things

- The problem our brain solves when speaking is "what is the next word?"
- The choice of the next word must depend on all other words before and after, depending on what you want to say
- If this choice were restricted to few words, you could not express complex concepts
- If the next world could be any possible out of 50000 words, speaking would be computationally hard
- The fact that the frequency of words follows Zipf's law implies that we strike a balance between these two extremes
- Note: Words are the relevant variables in language by definition (and among words there are more and less relevant ones)

Why do you live where you live?

- I chose to live where I live because my zip code can be nicely decomposed in primes: 34151 = 13 x 37 x 71
- Normal people choose where to live depending on their job, marriage, interests, etc
- Typically the zip code is not a relevant variable in this choice, whereas the city is.
- The distribution of city sizes contains information about how people choose where to live. It is individual choices that make the distribution informative
- Yet this choice depends on a host of unobserved variables

Complex system

= many degrees of freedom + function

- Complex systems are not random:
  - Individuals do not live in random cities
  - We do not choose words at random when speaking
  - Proteins are not random sequences of amino acids
  - ...
- Only part of what they do is accessible to us:

There are known knowns. These are things we know that we know.
There are known unknowns.
That is to say, there are things that we know we don't know.
But there are also unknown unknowns.
These are things we don't know we don't know.

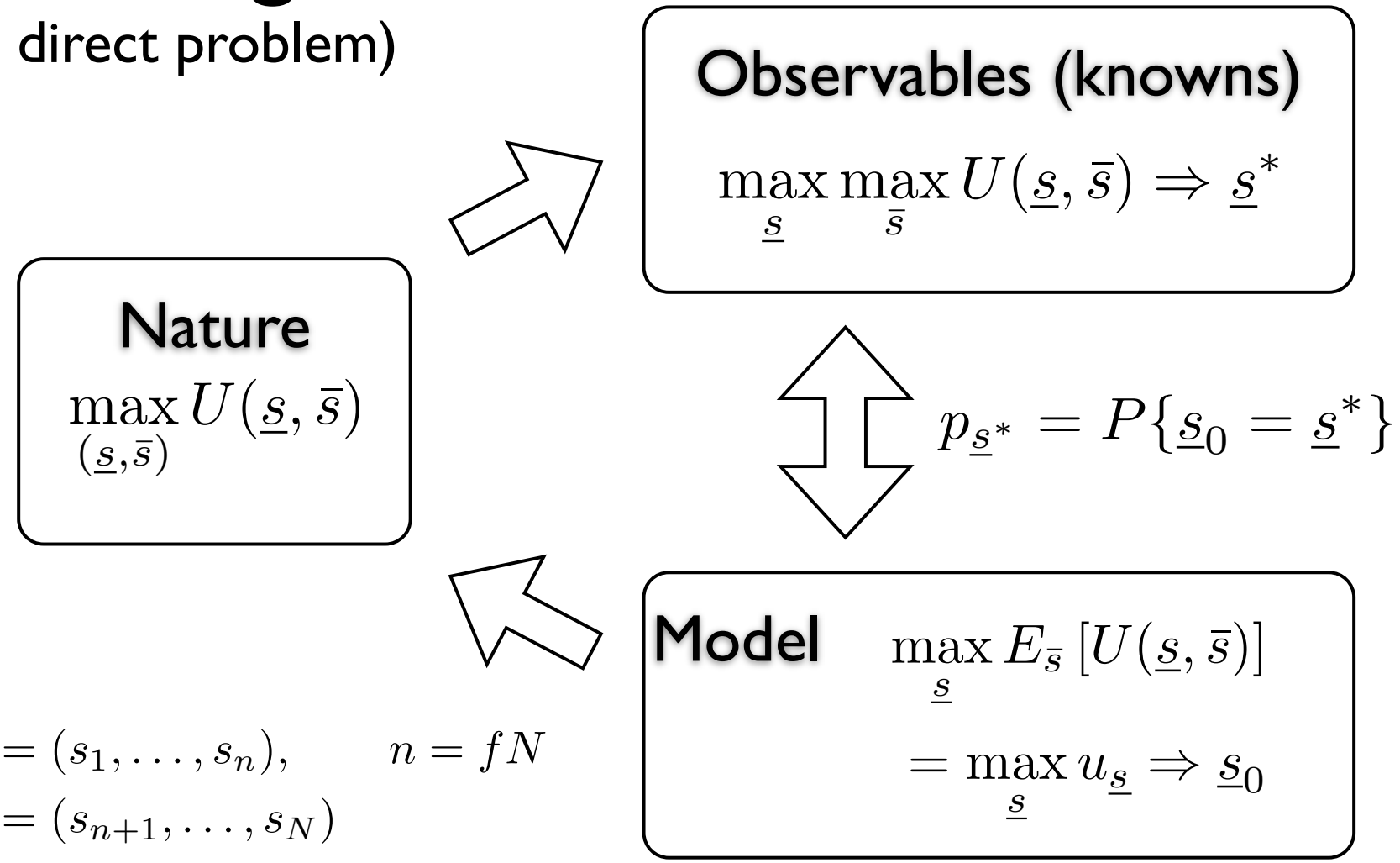
Donald Rumsfeld

Key issue:
what variables do we look at?

- If the variables we look at are irrelevant, we just get noise
- Relevant variables are those the system cares about.
- If the variables that we put in our **models** are relevant
  we can be predictive
- If the variables that we **sample** are relevant
  we can infer what the system is doing
- Relevance of the variables must reflect in the statics of the sample's frequency distribution
- Can we quantify this?
- Can we use this to find what the relevant variables are?

Modeling:

(the direct problem)



$\underline{s} = (s_1, \ldots, s_n), \qquad n = fN$

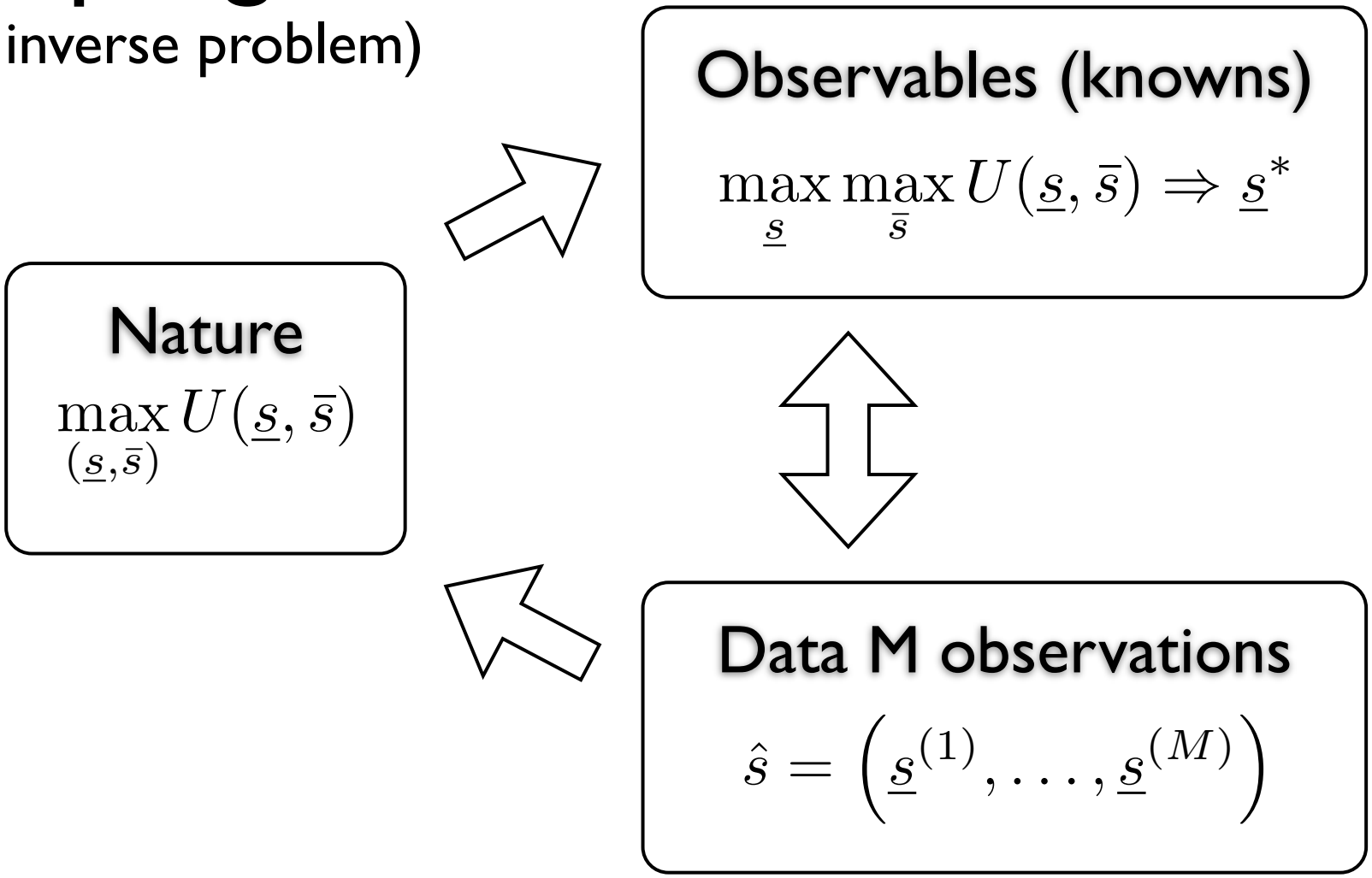
$\bar{s} = (s_{n+1}, \ldots, s_N)$

Q: How many? How relevant?

$$P\{\underline{s}^* = \underline{s}\} = \frac{1}{Z(\beta)} e^{\beta u_{\underline{s}}}, \qquad Z(\beta) = \sum_{\underline{s}} e^{\beta u_{\underline{s}}}$$

Sampling:

(the inverse problem)



Q: What can I say on $u_{\underline{s}} = E_{\bar{s}}[U(\underline{s},\bar{s})]$?

When is M large enough?

What do samples (typically) look like when M is small?

Where is the information in the sample?

- Problem: what is the optimized function $u_{\underline{s}}$?
- Sample of M observations $\hat{s} = \left(\underline{s}^{(1)}, \ldots, \underline{s}^{(M)} \right)$
- $K_{\underline{s}} = \sum_{1=1}^{M} \delta_{\underline{s}^{(i)}, \underline{s}}$ gives information on $u_{\underline{s}}$

$$u_{\underline{s}} \approx c + \beta^{-1} \log K_{\underline{s}}$$

- The information contained in the sample is H[K]

How much information?
E.g. find Mr X in Slovenia

- Slovenia has M people, need $\log_2 M$ bits to find Mr X
- If you knew the size K_X of the city where X lives then you'd need $\log_2 [K_X\, N(K_X)]$ bits
- If you knew which city s_X X lives in, then you'd need $\log_2 K_X$ bits
- If all individuals live in the same city K_X=M then you don't gain any information either way
- If each individual lives in a different city (K_X=1)
  you don't gain anything if you know K_X
  you know everything if you know s_X
- Information gain depends on N(K) and the amount of information is given by H[K]

Information gain and entropy

$$H[K] = -\sum_k \frac{kN(k)}{M} \log_2 \frac{kN(k)}{M}$$

$$H[\underline{s}] = -\sum_k \frac{kN(k)}{M} \log_2 \frac{k}{M}$$

$$H[K] = H[\underline{s}] = 0$$

$$H[K] = 0, \quad H[\underline{s}] = \log_2 M$$

What is the most informative N(k) for $0 < H[s] < \log_2 M$?

Maximally informative samples (upper bound)

$$N(k): \quad \max_{\{N(k)\}} H[K]$$

$$\text{s.t.} \quad H[\underline{s}] = H_0$$

$$\sum_k kN(k) = M$$

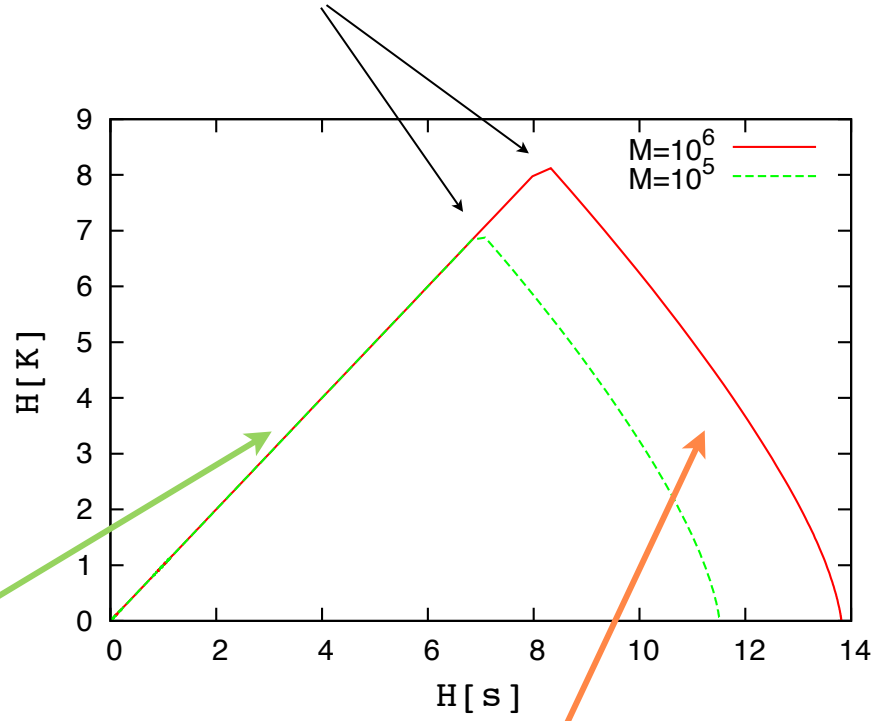
Data processing inequality:

$$H[\underline{s}] - H[K] = \sum_k \frac{kN(k)}{M} \log N(k)$$

$$\geq 0$$

$$N(k) = 1 \quad \sim \forall k$$

Zipf: $\mu = 2$



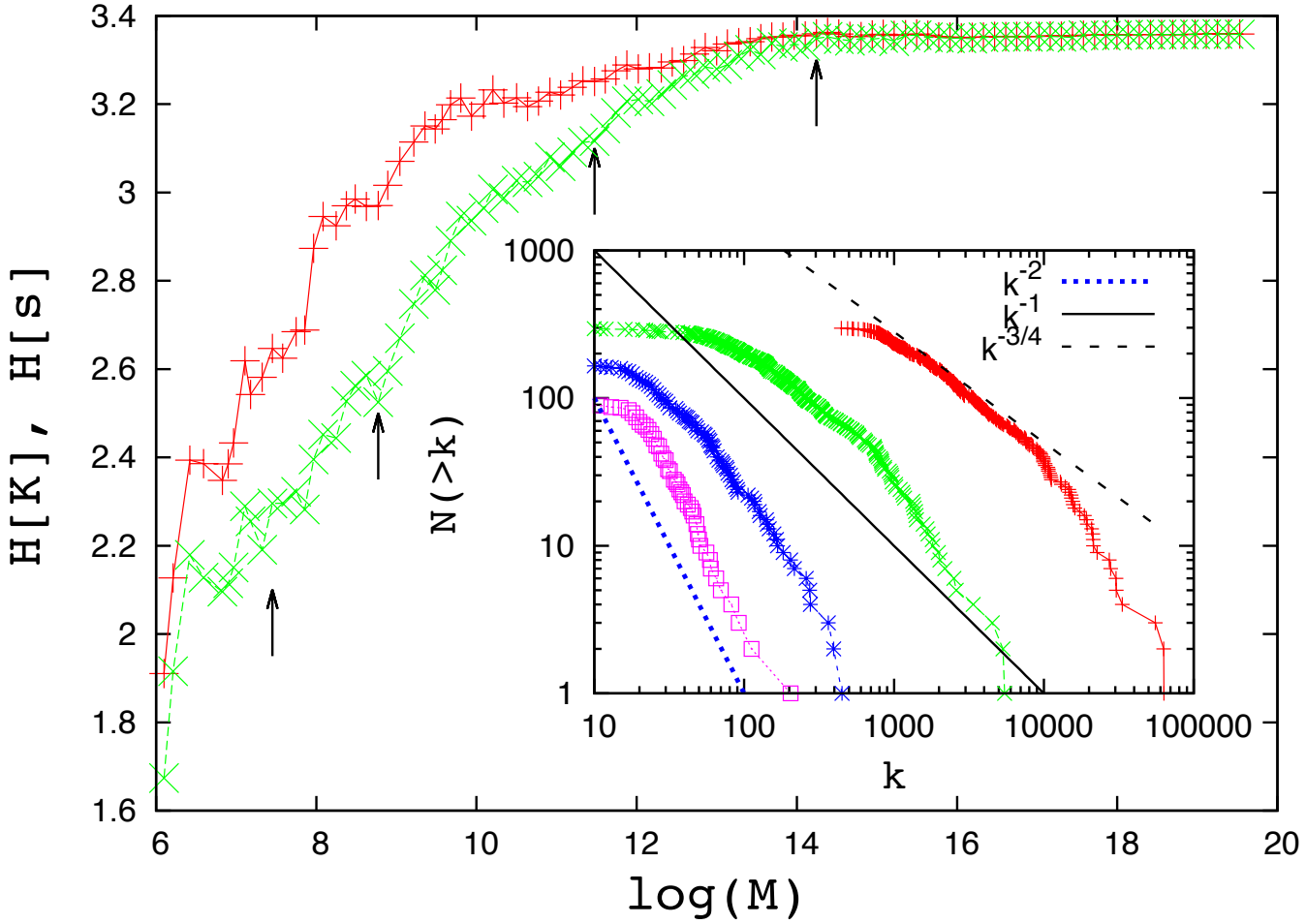
$$N(k) \sim k^{-\mu}$$

Applications/examples

- City size distribution
- Best classification of financial stocks
- Keywords in the "Origin of the Species"
- Finding relevant positions in proteins

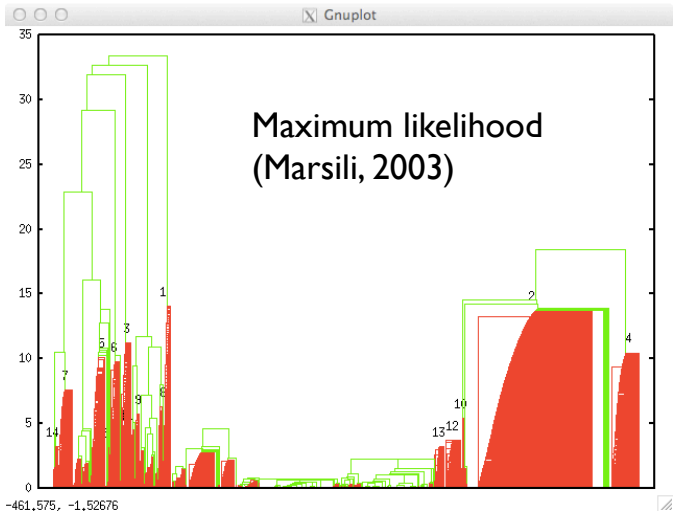
Subsampling city distribution

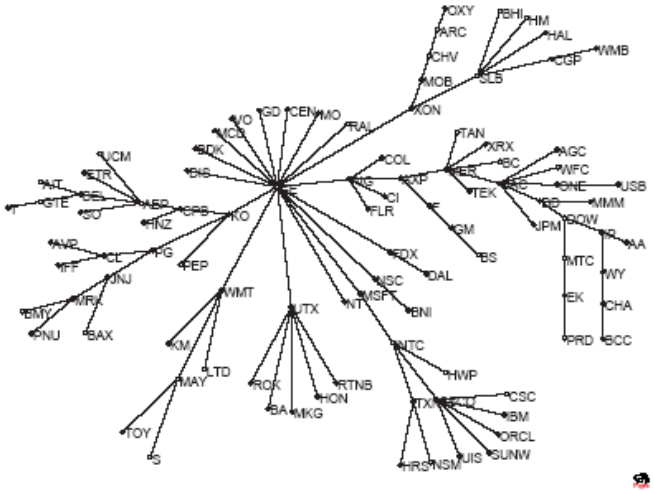
(IPUM database http://usa.ipums.org)



Finding relevant variables 1:
Classifying 4000 NYSE stocks

- Time series for M=4000 stocks,
  daily returns (1 Jan 1990 - 30 Apr 1999)
- $\underline{s}^{(i)}$ = label of stock i in hierarchical data clustering with N clusters
- Which method?

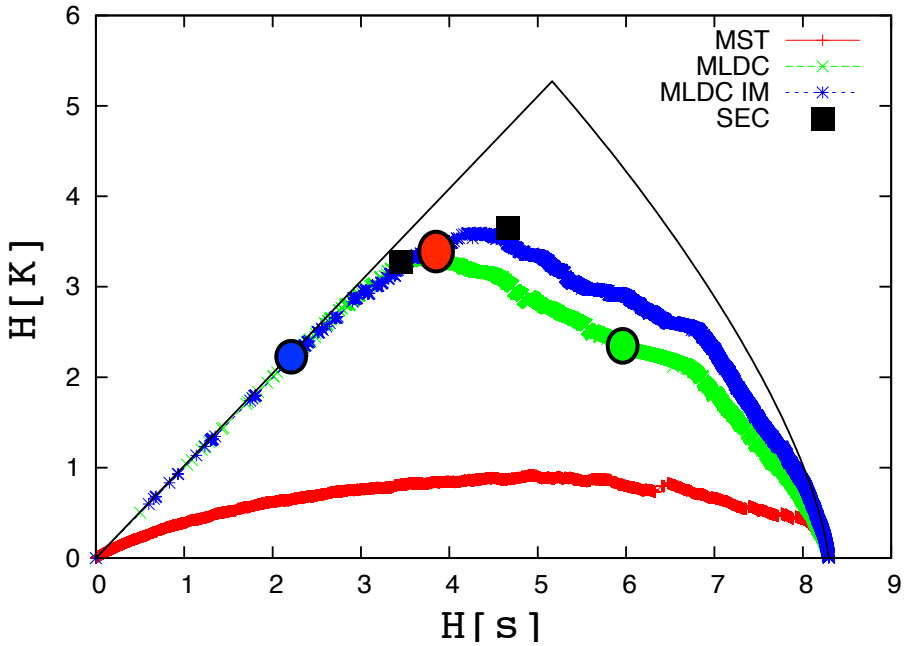


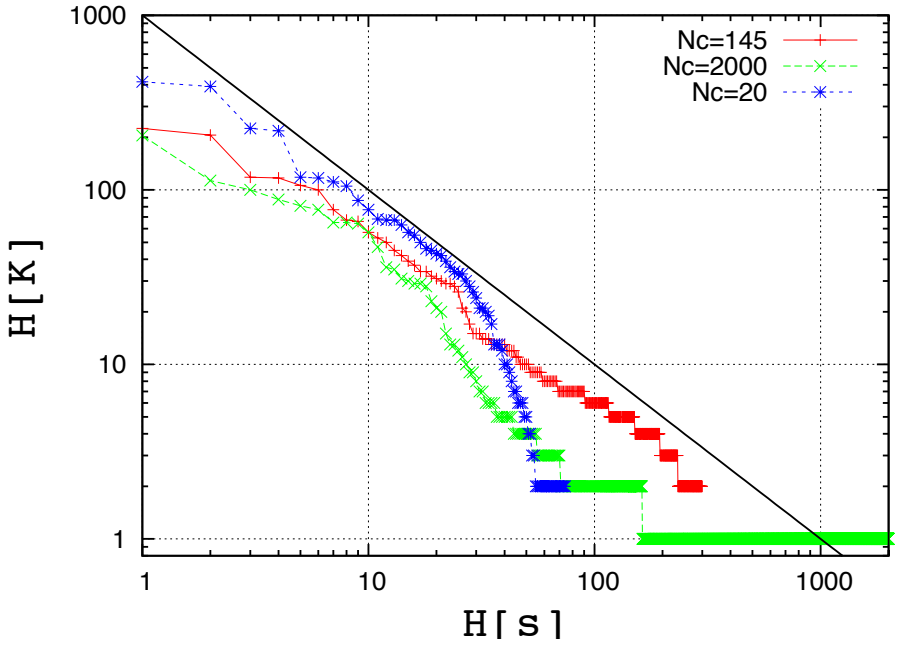


Minimal Spanning Tree (MST)
(Bonanno et. al. 2004, Tumminello et al. 2006)

H[K] can be used to score clustering methods

Data: $x_i(t)$ = (log)return of stock i=1,...,4000 in day t =1/1/90 - 30/4/99





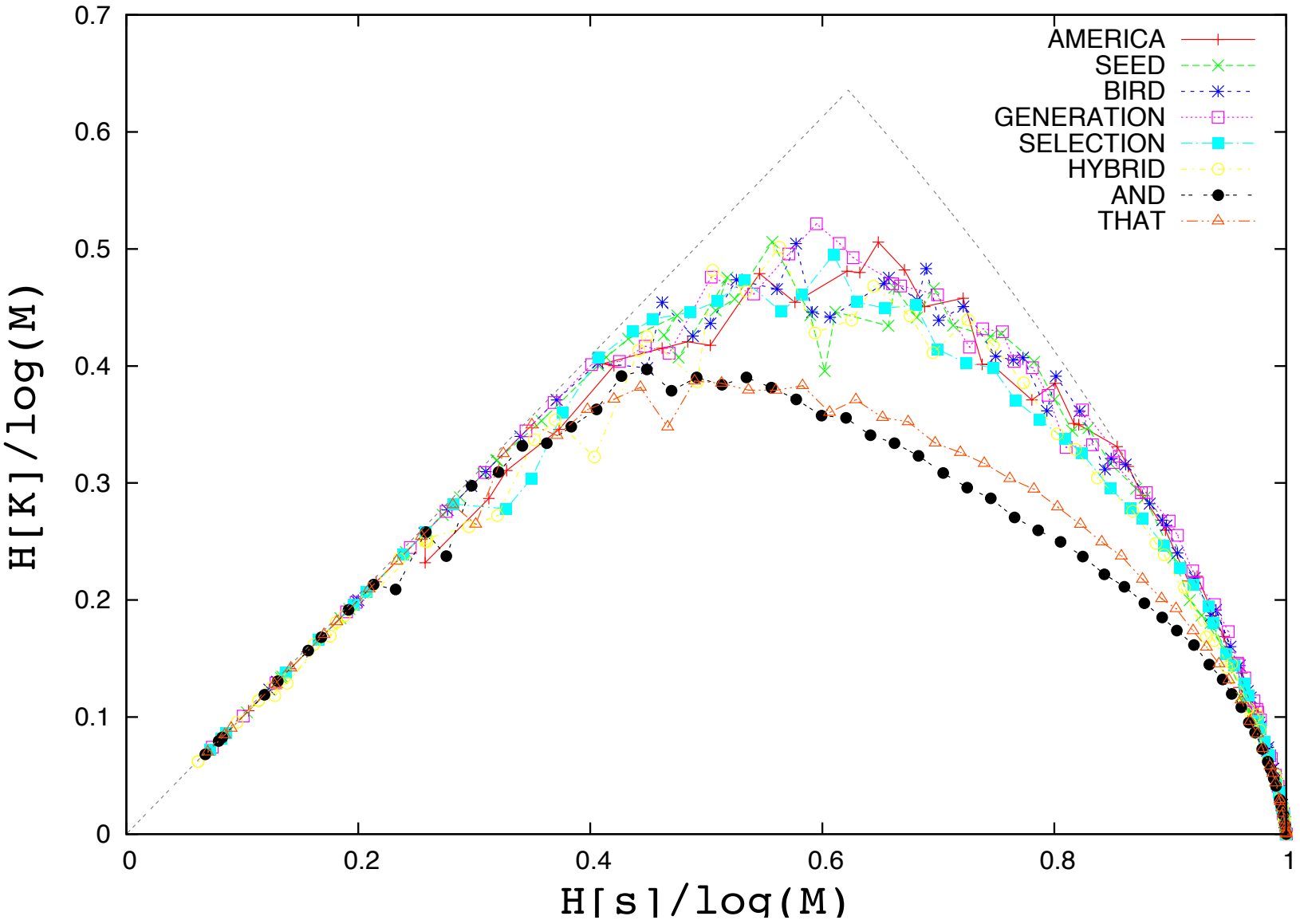
MST = Minimal Spanning Tree
MLDC = Maximum Likelihood Data Clustering
MLDC IM = MLDC on internal modes
SEC = US Security Exchange Commission classification

Finding relevant variables II:
Keywords in text

- Text = (w_1,w_2, w_3, ... , w_L) in blocks of B words

- Montemurro, Zanette (2009): relevant words are those whose frequency distribution in blocks differs most from the random distribution.
- K_s=number of times w occurs in block s=1,..,L/B
- Words with larger H[K] are the most relevant (those that are chosen for specific reasons)

The Origin of the Species



Finding relevant variables III:
Choosing relevant positions in proteins

- Protein: amino-acid sequence $\vec{s} = (s_1, \ldots, s_N)$
- Function (e.g. response regulator receptor) is related to sequence (e.g. structure/contacts, active sites, etc)
- Data: Families of homologous proteins in PFAM database. Same function different organisms, different sequences $\vec{s}^{(1)} \ldots \vec{s}^{(M)}$

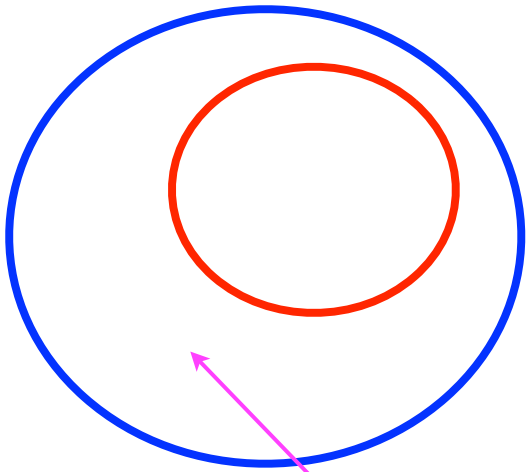
$$\vec{s}^{(i)} = \left(\underline{s}^{(i)}, \bar{s}^{(i)}\right)$$

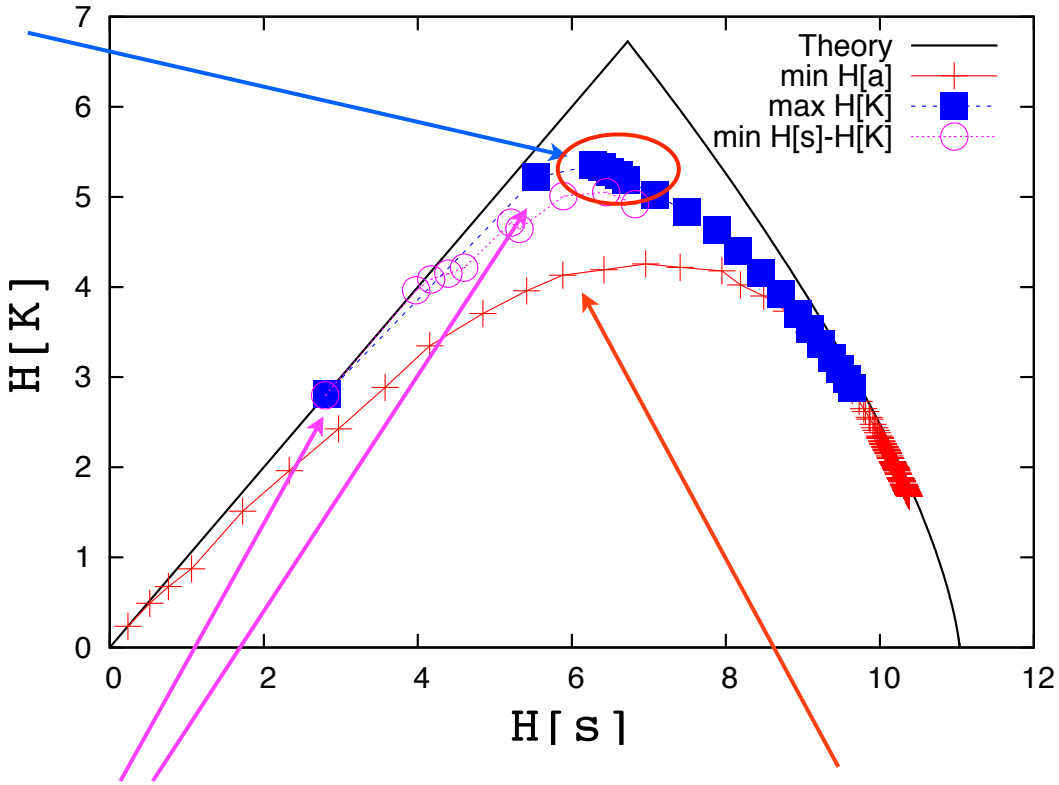
- How to find relevant variables?

1. subsequence of n most conserved amino-acids
2. subsequence that maximizes H[K]

Conserved variables are not the only relevant relevant ones

most relevant variables

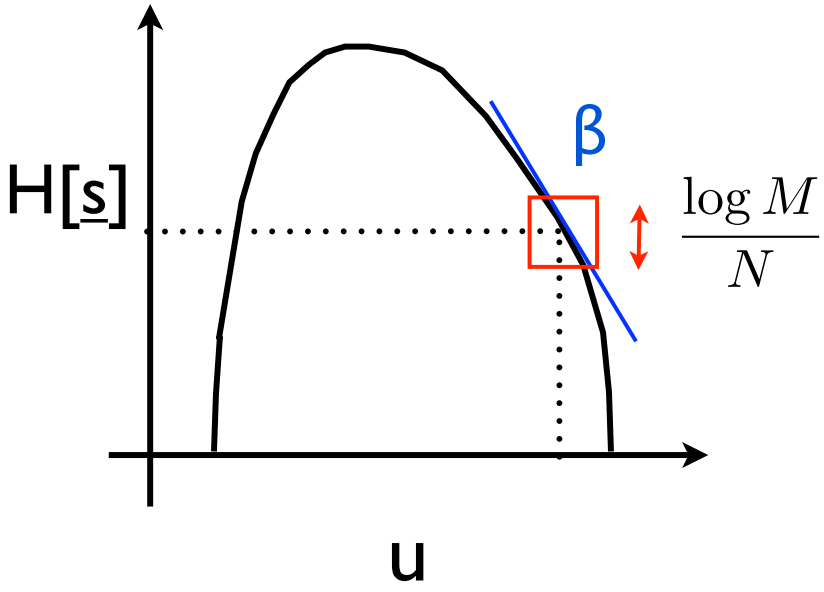


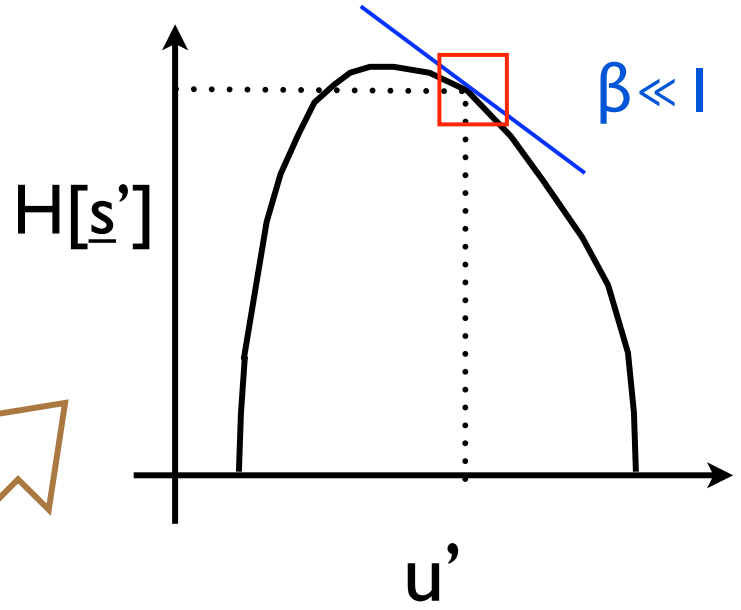


least conserved variables

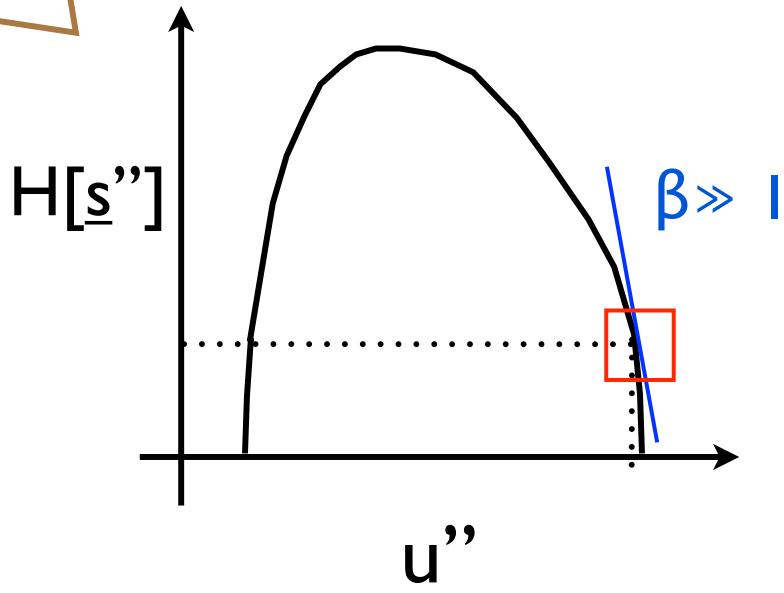
most conserved variables

Optimal sampling and experiment design:





?



What experiment would you do?
Which variables would you choose?

Summary

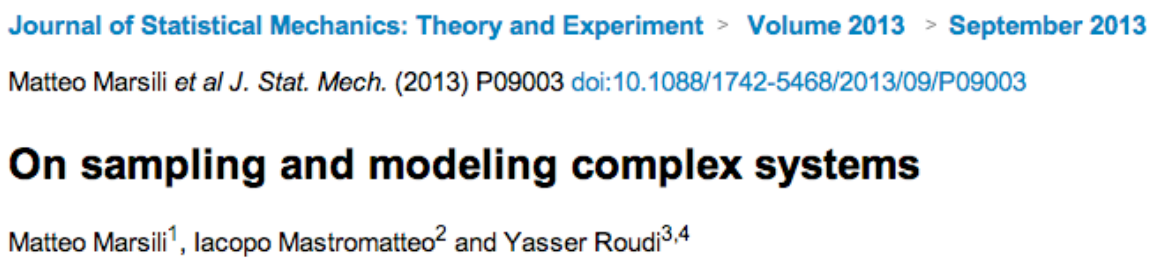
- Models may be predictive only when known variables are relevant
- Relevant variables are those for which samples "look critical" (i.e. most informative samples in the under-sampling regime are power laws)
- Zipf's law separates the under-sampling from well sampled regimes
- H[K] vs H[s] plot can be useful
  - to find relevant variables, keywords
  - to score clustering methods
  - ...
- Model free method

A numerical recipe

- Compute the number K_s of times you see observation s in your data
- Compute the number N(k) of observations that occur k times
- Compute and plot

$$H[K] = -\sum_k \frac{kN(k)}{M} \log \frac{kN(k)}{M}$$

$$H[s] = -\sum_k \frac{kN(k)}{M} \log \frac{k}{M}$$



Let's make this link more precise

- Variables $\vec{s} = (\underbrace{s_1, \ldots, s_n}_{\underline{s}}, \underbrace{s_{n+1}, \ldots, s_N}_{\overline{s}}), \qquad s_i = \pm 1, \; N \gg 1$

  $\underline{s}$ knowns $\overline{s}$ unknowns

- Function $U(\vec{s}) = u_{\underline{s}} + v_{\overline{s}|\underline{s}}, \qquad \left\langle v_{\overline{s}|\underline{s}} \right\rangle = 0$

  model unknown function

- Observable behavior $\underline{s}^* = \arg\max_{\underline{s}} \left[u_{\underline{s}} + \max_{\overline{s}} v_{\overline{s}|\underline{s}} \right]$

- Model's prediction $\underline{s}_0 = \arg\max_{\underline{s}} \left[u_{\underline{s}} \right]$
- Modeling: When is $\underline{s}_0 = \underline{s}^*$?
- Sampling: What can I learn from $\hat{s} = \left(\underline{s}^{(1)}, \ldots, \underline{s}^{(M)} \right)$?