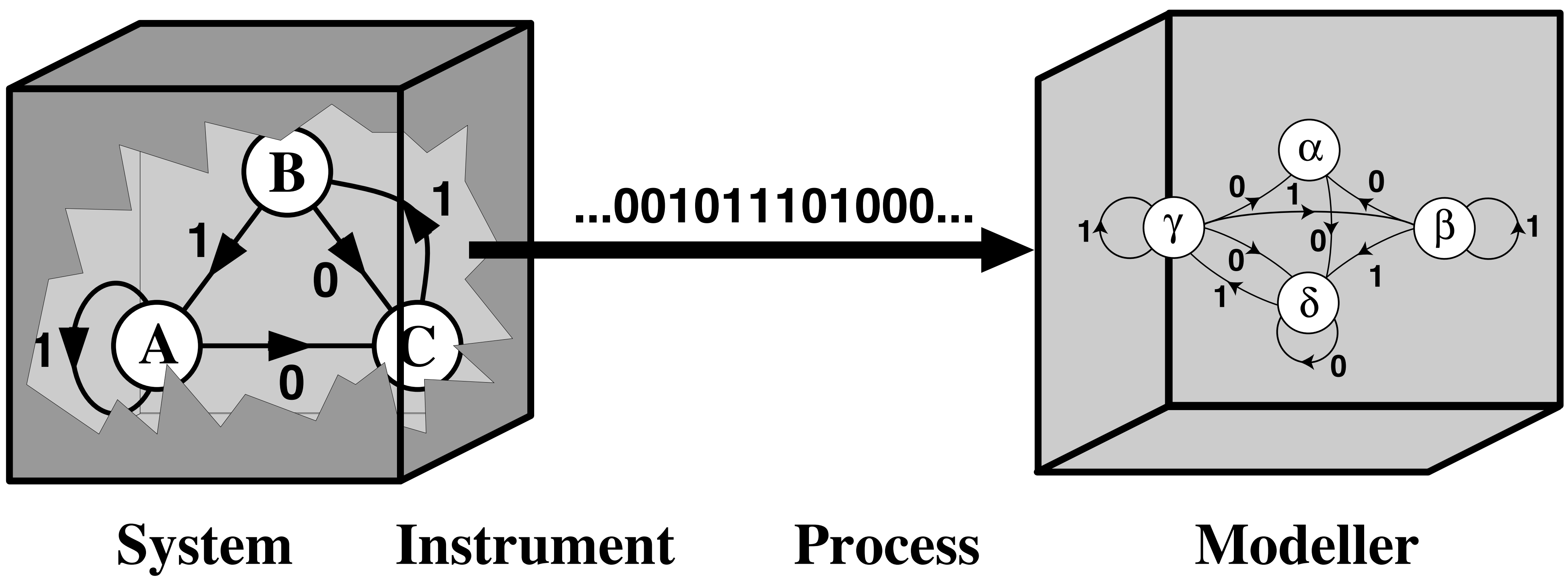
Brief Technical Review

Jim Crutchfield
Complexity Sciences Center
University of California, Davis

Workshop
Institute for Advanced Study
University of Amsterdam
6 May 2019
Amsterdam



The Learning Channel

Information in Complex Systems

- Algorithmic Basis of Probability
- Information Theory
- Information in Processes
- Memory in Processes
- Intrinsic Computation

Algorithmic Basis of Probability

Kolmogorov-Chaitin Complexity Theory

The question:

Algorithmic foundation for probability?

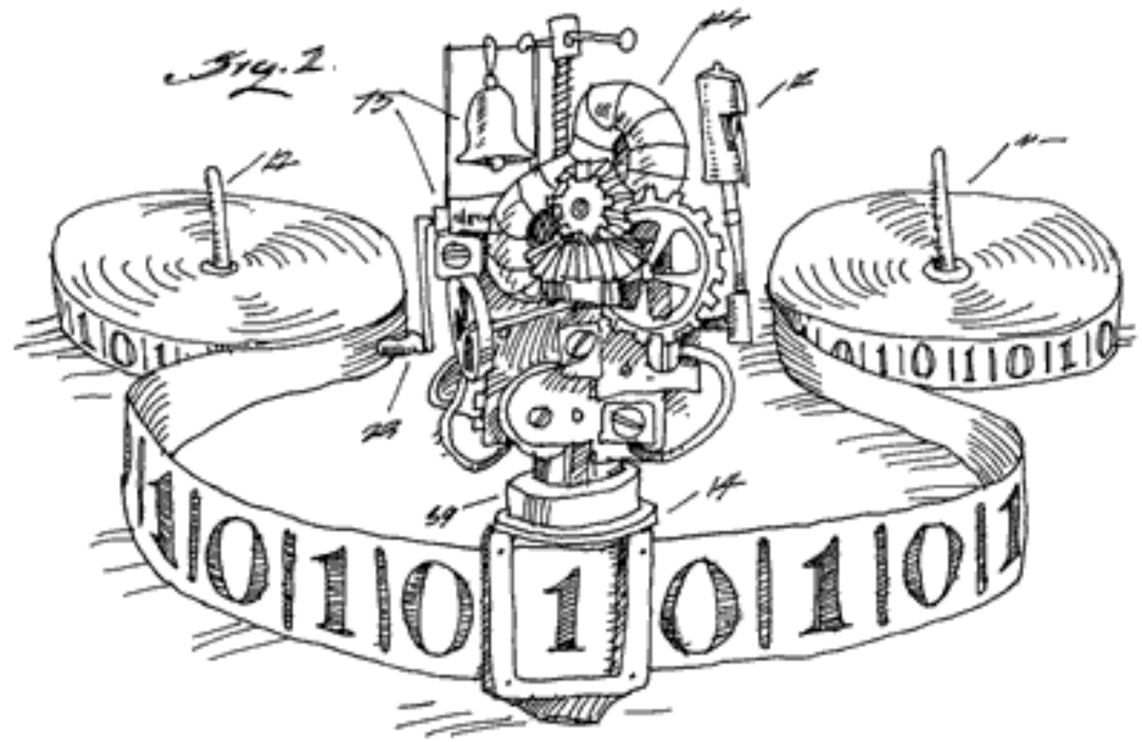
History:

1776: Treatise on probability theory (Laplace)
1920s: Frequency stability (von Mises)
1930s: Foundations of probability theory (Kolmogorov)
1940s: Information theory (Shannon … Szilard 1920s!)
1940s: Automata & computing theory (Turing)
1960s: KC Complexity Theory
(Kolomogorov, Chaitin, Solomonoff, ...)

Kolmogorov-Chaitin Complexity

Turing's machine (1937):

Finite-state controller +
Infinite read-write tape



Machine M:

Device to generate output $x = 1010111\ldots$ from program p:

$$M(p) = x$$

Kolmogorov-Chaitin Complexity

Universal Turing Machine: U

Sufficient states, control logic, and tape alphabet

$\Rrightarrow$ Calculate any input-output function

UTM programs generate output: $U(p) = x$

(Python interpreter w/ infinite memory.)

Kolmogorov-Chaitin Complexity:

Size of smallest program p that generates object x

$$K(x) = \min\{|p| : \; U(p) = x\}$$

Kolmogorov-Chaitin Complexity

Consider Python program:

```
def generate_x():
    print x
```

And so:

$$K(x) \leq |x| + \text{constant}$$

For most objects:

$$K(x) \approx |x|$$

Kolmogorov-Chaitin Complexity is not computable.

(Theorem: No program can calculate $K(x)$.)

Kolmogorov-Chaitin Complexity

Exercise! Which has high, which low $K(x)$?

00100100001111110110101010001000
10000101101000110000100011010011
00010011000110011000101000101110
00000011011100000111001101000100

π

Algorithm $\Rightarrow$ low $K(x)$

(Bailey–Borwein–Plouffe 1997)

10000010100011011111101110011100
01101101001100010110010001010100
00101100011011000110001110111000
10110100010000111000111001110011

Random

High $K(x)$

Kolmogorov-Chaitin Complexity

Lessons:

A random object is its own shortest description.

$K(x)$ maximized by random objects.

Probability of objects:

$$\Pr(x) \approx 2^{-K(x)}$$

Alternatives?

Computable?
Scientifically applicable?

Information

Information ...

Information as uncertainty and surprise:

Observe something unexpected:
Gain information

Bateson: "A difference that makes a difference"

Information ...

Shannon Entropy: $X \sim P$

$$x \in \mathcal{X} = \{1, 2, \ldots, k\}$$
$$P = \{\Pr(x = 1), \Pr(x = 2), \ldots\}$$

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

Note: $0 \log 0 = 0$

Units:

Log base 2: $H(X) = [\mathrm{bits}]$

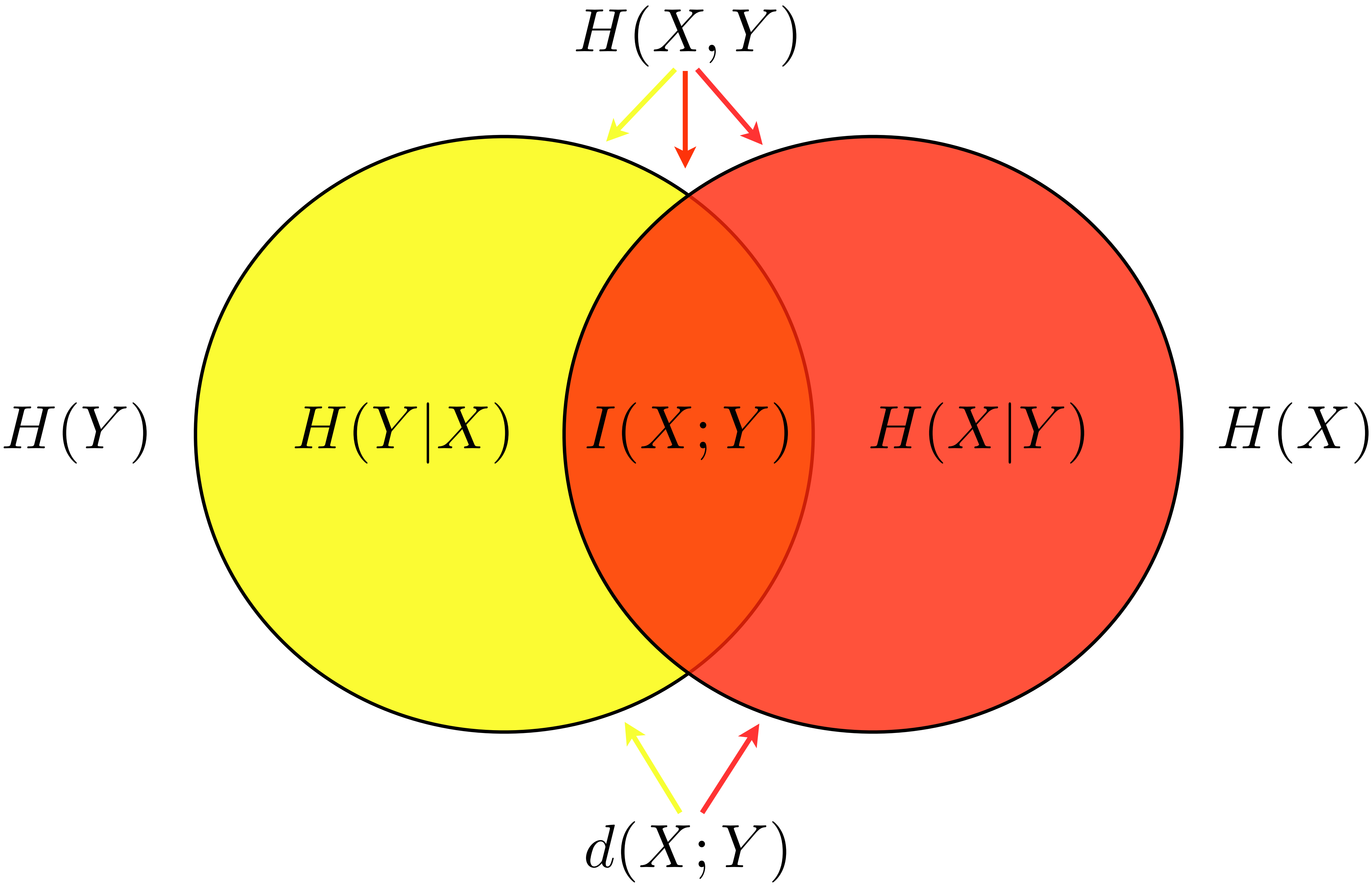
Natural log: $H(X) = [\mathrm{nats}]$

Properties:

1. Positivity: $H(X) \geq 0$
2. Predictive: $H(X) = 0 \Leftrightarrow p(x) = 1$ for one and only one x
3. Random: $H(X) = \log_2 k \Leftrightarrow p(x) = U(x) = 1/k$

Information ...

Event Space Relationships of Information Quantifiers:



Why information?

1. Accounts for any type of co-relation
   - Statistical correlation ~ linear only
   - Information measures nonlinear correlation
2. Broadly applicable:
   - Many systems don't have "energy", physical modeling precluded
   - Information defined: social, biological, engineering, ... systems
3. Comparable units across different systems:
   - Correlation: Meters v. volts v. dollars v. ergs v. ...
   - Information: bits.
4. Probability theory ~ Statistics ~ Information
5. Complex systems:
   - Emergent patterns!
   - We don't know these ahead of time

Information ...

Real Communication Theory:

How to compress a process:

Can't do better than $H(X)$

(Shannon's First Theorem)

How to communicate a process's data: $H(X) \leq \mathcal{C}$

Can transmit error-free at rates up to channel capacity

(Shannon's Second Theorem)

Both results give operational meaning to entropy.

Previously, entropy motivated as a measure of surprise.

George Carlin (1937-2008)

Information in Processes

Information in Processes ...

Entropy Growth for Stationary Stochastic Processes: $\Pr(\overleftrightarrow{S})$

Block Entropy:

$$H(L) = H(\Pr(s^L)) = -\sum_{s^L \in \mathcal{A}} \Pr(s^L) \log_2 \Pr(s^L)$$

Monotonic increasing: $H(L) \geq H(L-1)$

Adding a random variable cannot decrease entropy:

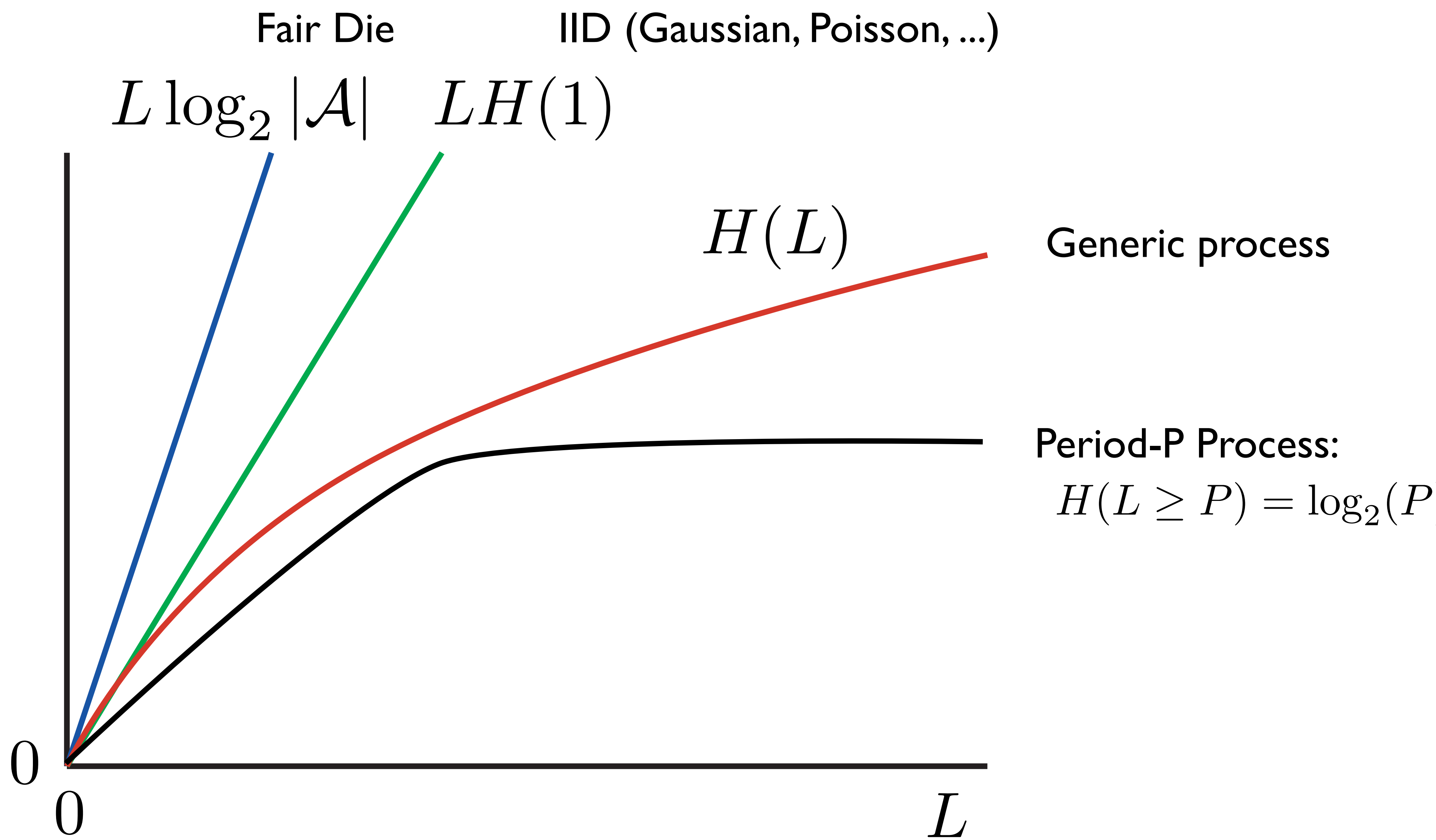
$$H(S_1, S_2, \ldots, S_L) \leq H(S_1, S_2, \ldots, S_L, S_{L+1})$$

No measurements, no information: $H(0) = 0$

Information in Processes ...

Entropy Growth for Stationary Stochastic Processes ...

Block Entropy ...



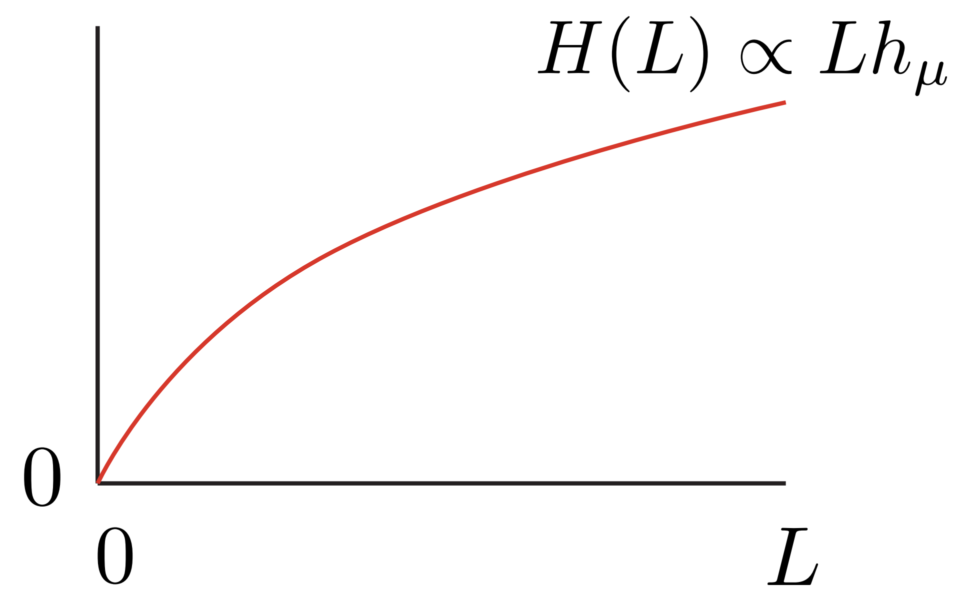
Information in Processes ...

Entropy Rates for Stationary Stochastic Processes:

Entropy per symbol is given by the Source Entropy Rate:

$$h_\mu = \lim_{L \to \infty} \frac{H(L)}{L}$$

(When limits exists.)



Interpretations:

Asymptotic growth rate of entropy
Irreducible randomness of process
Average description length (per symbol) of process

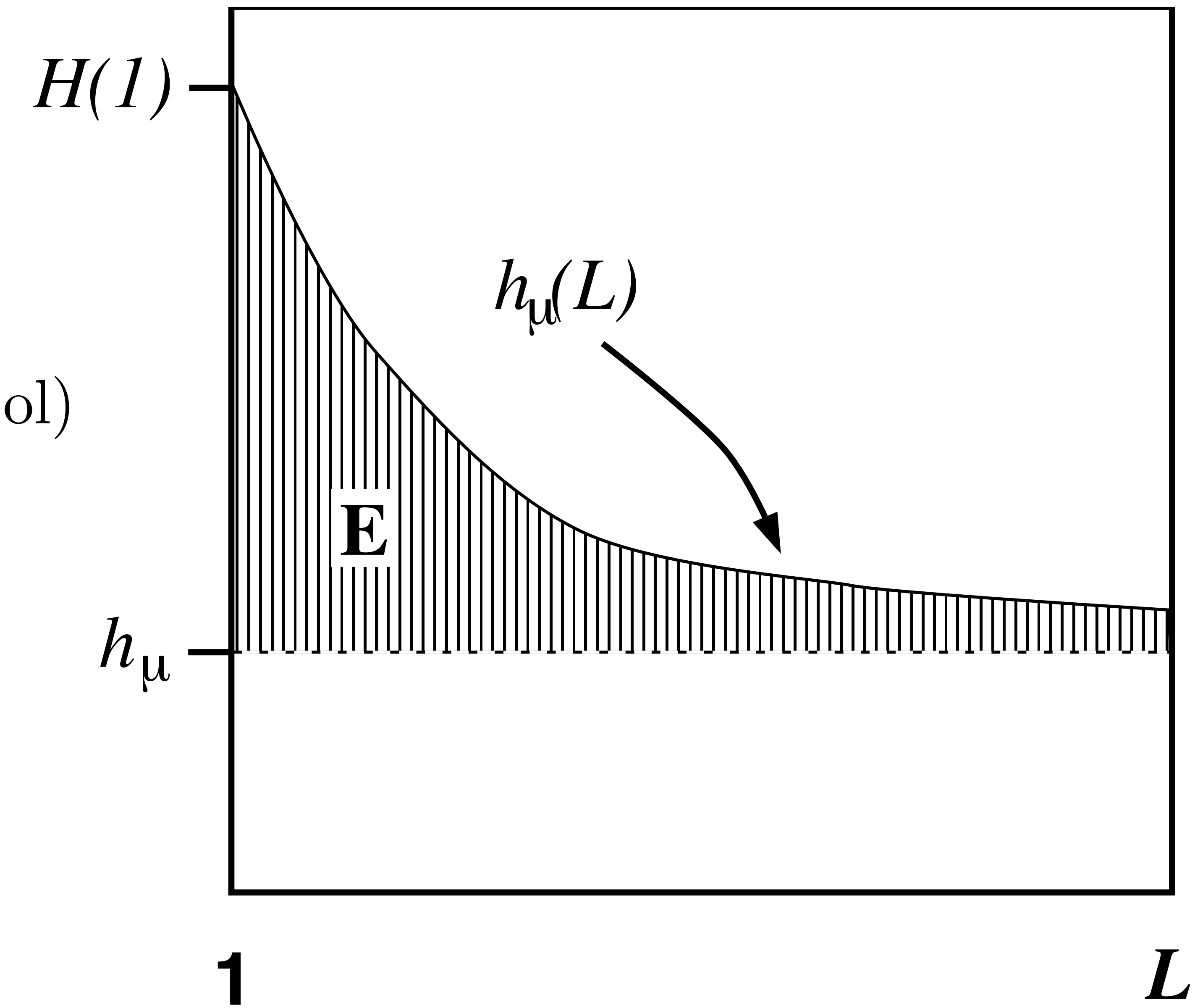
Memory in Processes ...

Excess Entropy:

As entropy convergence:

$$\mathbf{E} = \sum_{L=1}^{\infty} [h_\mu(L) - h_\mu]$$

$(\Delta L = 1 \text{ symbol})$



Properties:

(1) Units: $\mathbf{E} = [\text{bits}]$

(2) Positive: $\mathbf{E} \geq 0$

(3) Controls convergence to actual randomness.

(4) Slow convergence ⇔ Correlations at longer words.

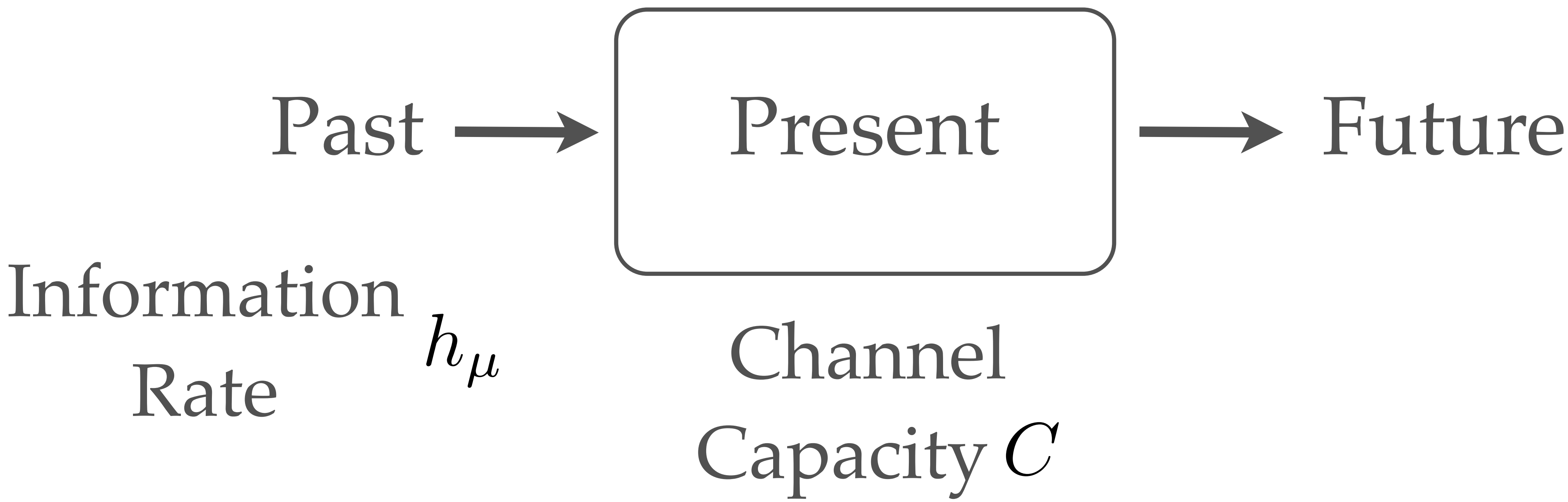
(5) Complementary to entropy rate.

Memory in Processes ...

Excess Entropy ...

Mutual information between past and future: Process as channel

Process $\Pr(\overleftarrow{X}, \overrightarrow{X})$ communicates past $\overleftarrow{X}$ to future $\overrightarrow{X}$:

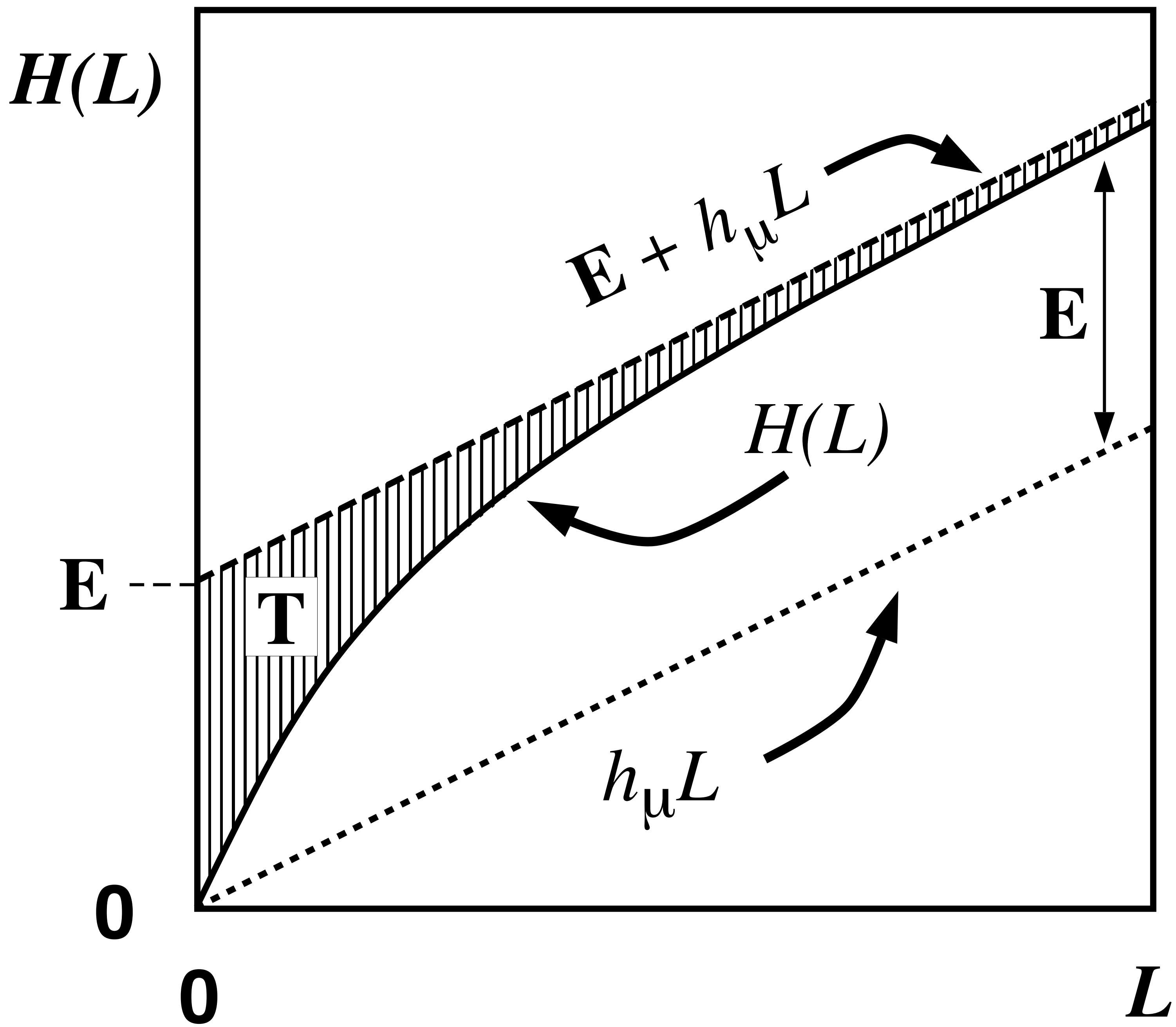


Excess Entropy as Channel Utilization:

$$\mathbf{E} = I[\overleftarrow{X}; \overrightarrow{X}]$$

Memory in Processes ...

Information-Entropy Roadmap for a Stochastic Process:



Memory in Processes ...

What is information?

Depends on the question!

Uncertainty, surprise, randomness, ...
Compressibility.
Transmission rate.
Memory, apparent stored information, ...
Synchronization.
Created and actively stored.
Created and forgotten.
Predictive information.
Predictable information.
...

Algorithmic Basis of Information

Kolmogorov-Chaitin Complexity versus Shannon Information

KC Complexity versus Shannon Information

Consider average KC Complexity of source $X_{0:\ell}$:

$$K(\ell) \equiv \langle K(x_{0:\ell}) \rangle_{\text{realizations}}$$

Recall Block Entropy:

$$H(\ell) \equiv H[\Pr(X_{0:\ell})]$$

Their growth rates equal the Shannon entropy rate:

$$h_\mu = \lim_{\ell \to \infty} \frac{H(\ell)}{\ell} = \lim_{\ell \to \infty} \frac{K(\ell)}{\ell}$$

KC Complexity of typical realizations from an information source grows proportional to the Shannon entropy rate [Brudno 1978].

KC Complexity versus Shannon Information

Again, KC Complexity is a measure of randomness, unpredictability, surprise, ...

As well as being a measure of the *deterministic* computing resources requires to *exactly* reproduce a given finite string.

KC Complexity and entropy rate maximized by IID processes.

KC Complexity versus Statistical Complexity

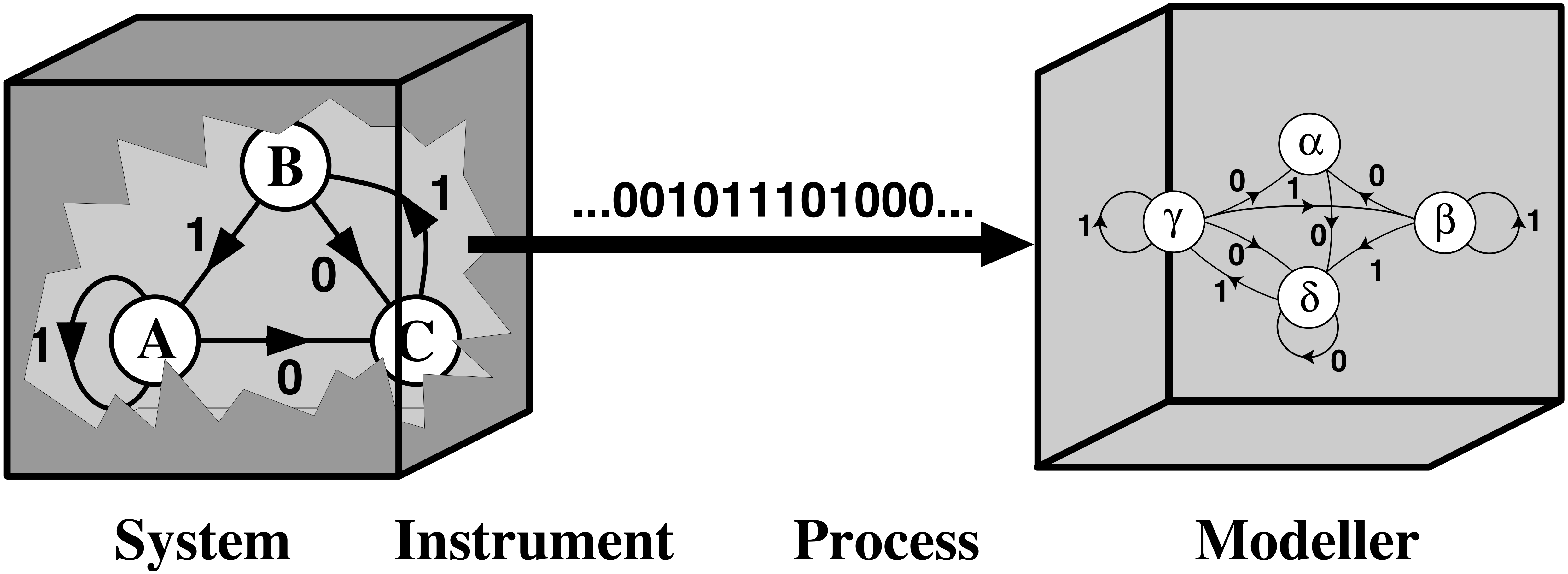
KC Complexity Theory:

- Great mathematics.
- Uncomputable.
- Not quantitative: constants of proportionality unknown

Quantitative sciences use Information Theory instead.

Intrinsic Computation

The Learning Channel:



Central questions:

What are the states?

What is the dynamic?

The Learning Channel ...

Causal States:

Causal State:

Set of pasts with same morph $\Pr(\overrightarrow{S} \mid \overleftarrow{s})$.

Set of histories that lead to same predictions.

Predictive equivalence relation:

$$\overleftarrow{s}' \sim \overleftarrow{s}'' \iff \Pr(\overrightarrow{S} \mid \overleftarrow{S} = \overleftarrow{s}') = \Pr(\overrightarrow{S} \mid \overleftarrow{S} = \overleftarrow{s}'')$$

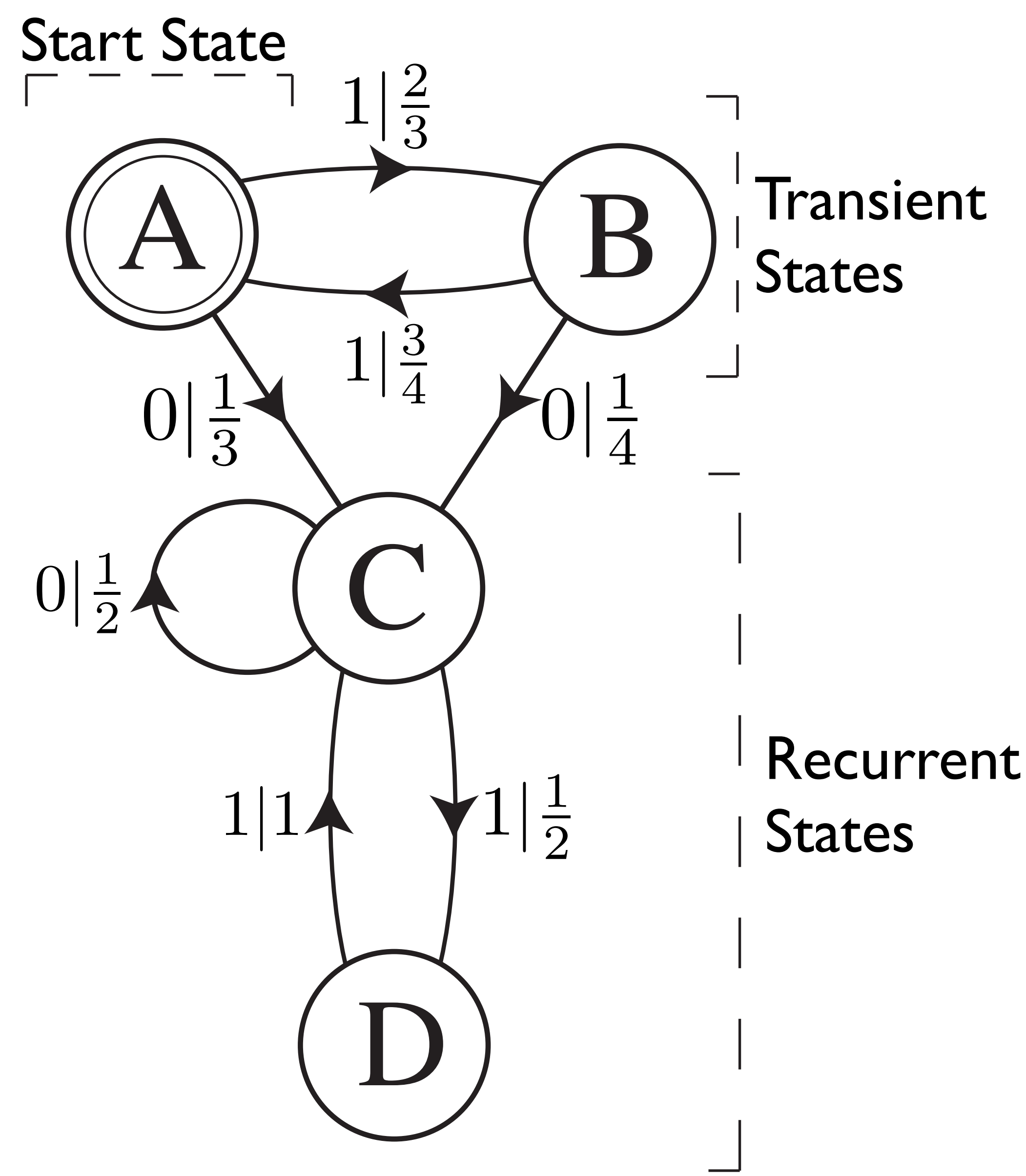
$$\overleftarrow{s}', \overleftarrow{s}'' \in \overleftarrow{\mathbf{S}}$$

The ϵ-Machine ...

$$\text{Process} \Rightarrow \text{Predictive equivalence} \Rightarrow \epsilon - \text{Machine}$$

$$\Pr(\overleftrightarrow{S}) \Rightarrow \overleftarrow{\mathbf{S}} / \sim \Rightarrow \epsilon - \text{Machine}$$

$$\mathcal{M} = \left\{ \boldsymbol{\mathcal{S}}, \{T^{(s)}, s \in \mathcal{A}\} \right\}$$



Unique Start State:

$$\mathcal{S}_0 = [\lambda]$$

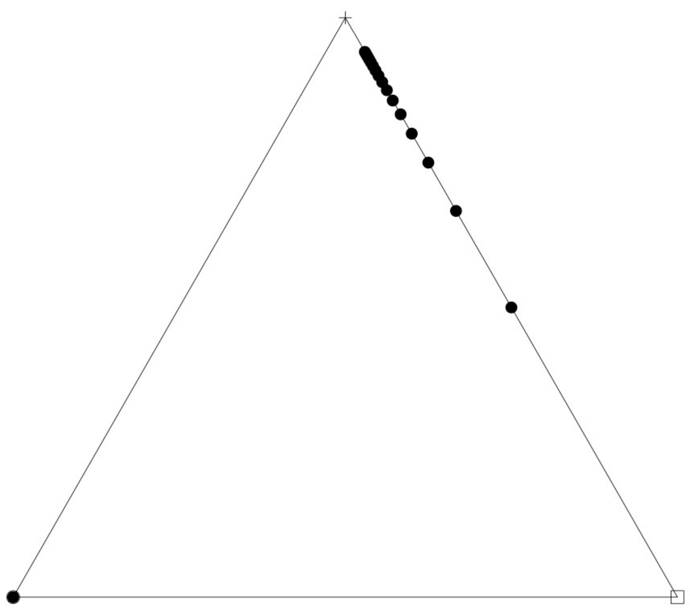
$$\Pr(\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2, \ldots) = (1, 0, 0, \ldots)$$

Transient States

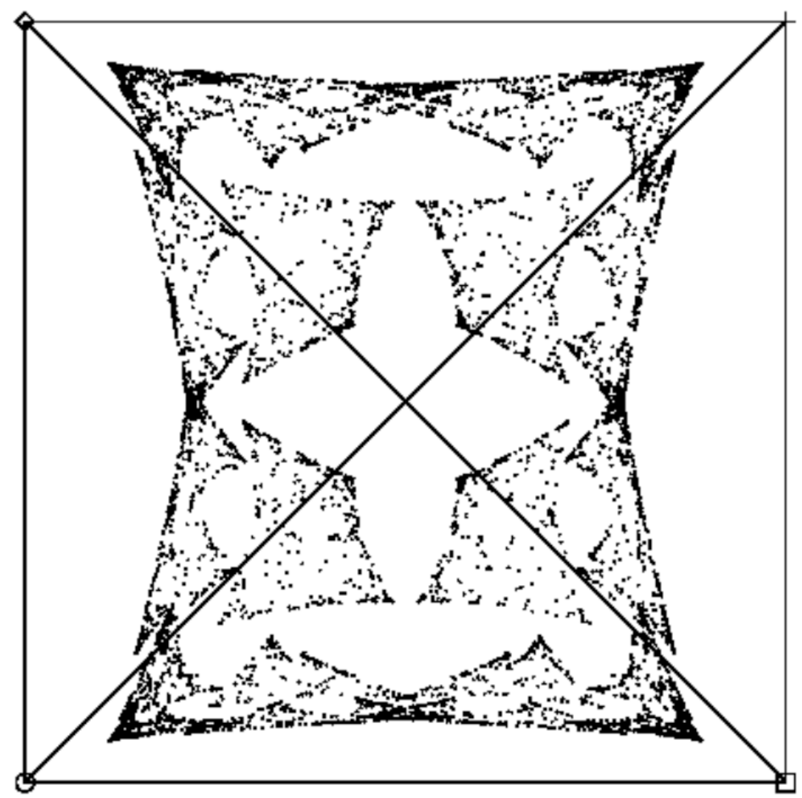
Recurrent States

The ϵ-Machine ...

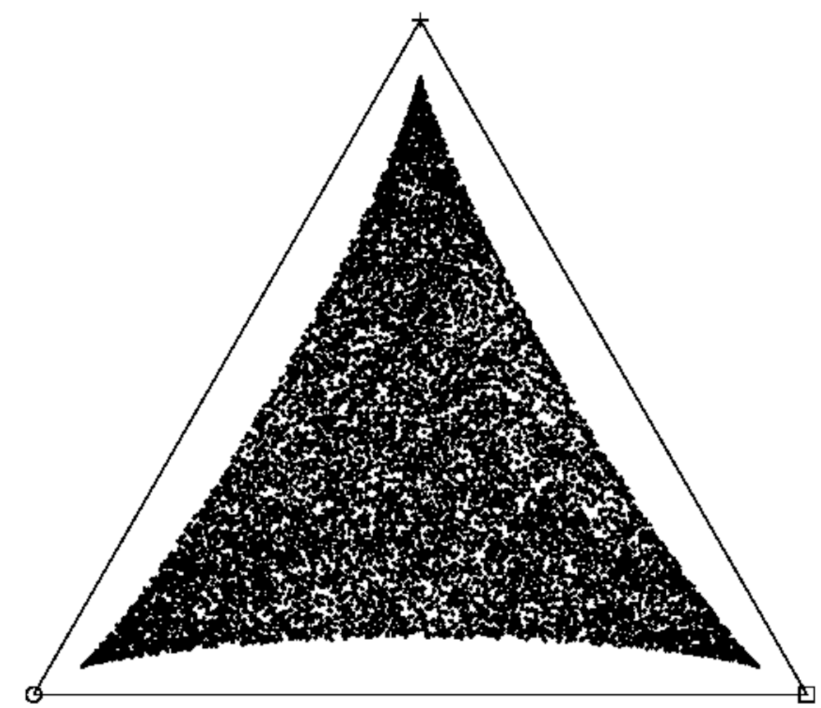
The ϵ-Machine of a Process ...



Denumerable
Causal States



Fractal



Continuous

The ϵ-Machine ...

Summary:

$\epsilon\mathrm{M}$:

(1) Optimal predictor: Lower prediction error than any rival.

(2) Minimal size: Smallest of the prescient rivals.

(3) Unique: Smallest, optimal, unifilar predictor is equivalent.

(4) Model of the process: Reproduces all of process's statistics.

(5) Causal shielding: Renders process's future independent of past.

Measures of Intrinsic Computation ...

A complex process's intrinsic computation:

(1) How much of past does process store?

$$C_\mu$$

(2) In what architecture is that information stored?

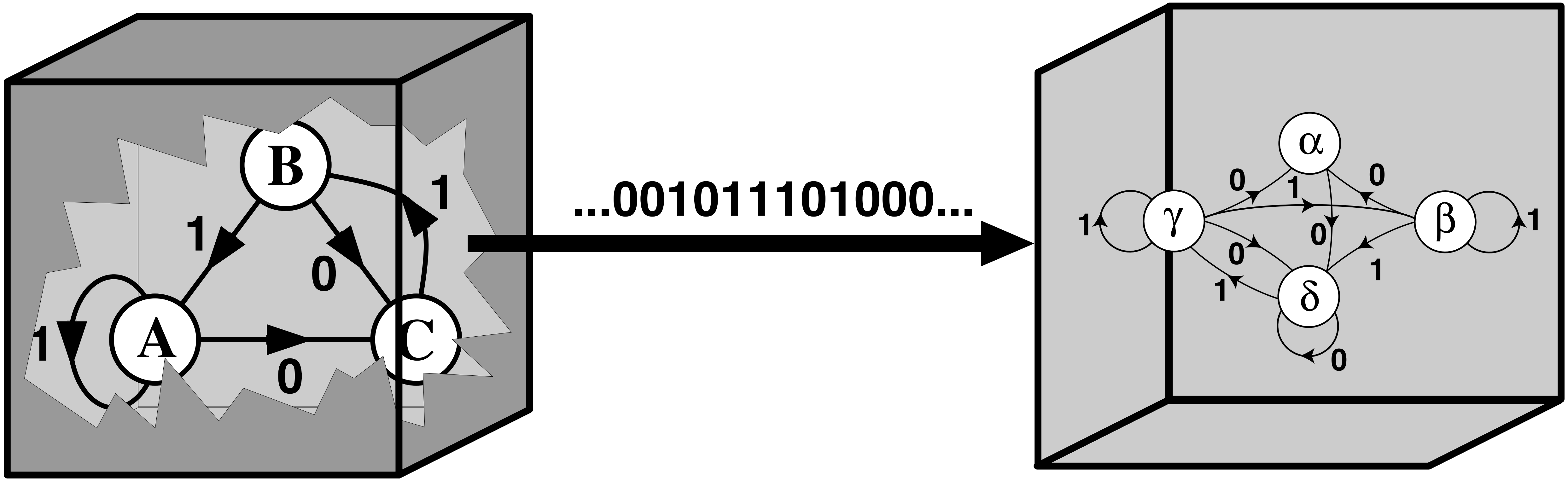
$$\left\{ \boldsymbol{\mathcal{S}}, \{T^{(s)}, s \in \mathcal{A}\} \right\}$$

(3) How is stored information used to produce future behavior?

$$h_\mu$$

Intrinsic Computation ...

Analysis narrative:



System **Instrument** **Process** **Modeller**

Forms of Chaos:
- Deterministic sources of novelty
- Mechanisms that produce unpredictability
- Sensitive dependence on initial condition
- Sensitive dependence on parameter

Measurement Theory:
- Partitions
- Optimal Instrument:

$$\max_{\{\mathcal{P}\}} h_\mu$$

$$\min_{\{\mathcal{P}\}} C_\mu$$

How random?

$$\lambda, H(L), h_\mu$$

How structured?

$$C_\mu, \mathbf{E}, \mathbf{T}, \mathbf{G}, \mathcal{R}$$

Universal model:
- $\epsilon - \mathrm{Machine}$
- Pattern defined
- Causal Architecture
- Intrinsic Computation

Intrinsic Computation ...

A system is unpredictable
if it has positive entropy rate: $h_\mu > 0$

A system is complex
if it has positive structural complexity measures: $C_\mu > 0$

A system is emergent
if its structural complexity measures increase over time:

$$C_\mu(t') > C_\mu(t), \text{ if } t' > t$$

A system is hidden
if its crypticity is positive: $\chi = C_\mu - \mathbf{E} > 0$

Algorithmic Basis of Information ...

Kolmogorov-Chaitin Complexity versus Statistical Complexity

KC Complexity versus Statistical Complexity

We saw that:
KC complexity of typical realizations from an information source grows proportional to the Shannon entropy rate:

$$K(x) \propto h_\mu |x|$$

Thus, KC complexity is a measure of randomness.

KC Complexity versus Statistical Complexity

What's the relationship to Statistical Complexity C_μ?

Since randomness drives Kolmogorov-Chaitin complexity, let's discount for generating randomness:

Programs consist of model m and data d (random part unexplained by m).

Sophistication of object:

$$S_k(x) = \min\{|m| : p = m + d \text{ and } |p| - K(x) \leq k\}$$

Also, uncomputable.

KC Complexity versus Statistical Complexity

Consider the average sophistication:

$$S(\ell) = \langle S_0(x_{0:\ell}) \rangle$$

It is statistical complexity:

$$C_\mu \propto_{\ell \gg 1} S(\ell)$$

Since program = model + data:

$$K(\ell) = S(\ell) + \langle |d| \rangle_{x_{0:\ell}}$$

We have:

$$K(\ell) \approx C_\mu + h_\mu \ell$$

Since a process has a structure, as ℓ gets large,
with probability 1 each possible $x_{0:\ell}$ has the same model.

KC Complexity versus Statistical Complexity

Recall the Block Entropy

$$H(\ell) \approx C_\mu + h_\mu \ell$$

Similar scaling.

$K(\ell)$ versus $H(\ell)$:

$\mathbf{E}$ quantifies the amount of information observed as ℓ gets large, whereas C_μ quantifies how much information it takes to predict as ℓ gets large.

KC Complexity versus Statistical Complexity

Kolmogorov-Chaitin Theory versus Computational Mechanics

First, ε−machine describes distribution over a system's behaviors, including individual realizations.

Second, one can exactly calculate the Shannon entropy rate for a system's behaviors.

Third, the computational model is a probabilistic UTM:
a Bernoulli-Turing Machine.

KC Complexity versus Statistical Complexity

Computational Mechanics was introduced to be a calculable, quantitative version of KC Complexity Theory.

Constructive! For finite eMs, all complexity/information measures
- can be calculated in closed form.
- $O(1)$ computational complexity.

So, much computational complexity in KC Theory and in Information Theory obviated.

“To know how to criticize is good,

to know how to create is better.”

Henri Poincaré, “Les Définitions en Mathématiques”, L’Eliseignement des Mathématiques **6** (1904) 255-283.

—, **Mathematical Definitions in Education**, Georges Carré, Paris (1904) Part II. Ch. 2 p. 129.

Thanks!