Information in Complex Systems

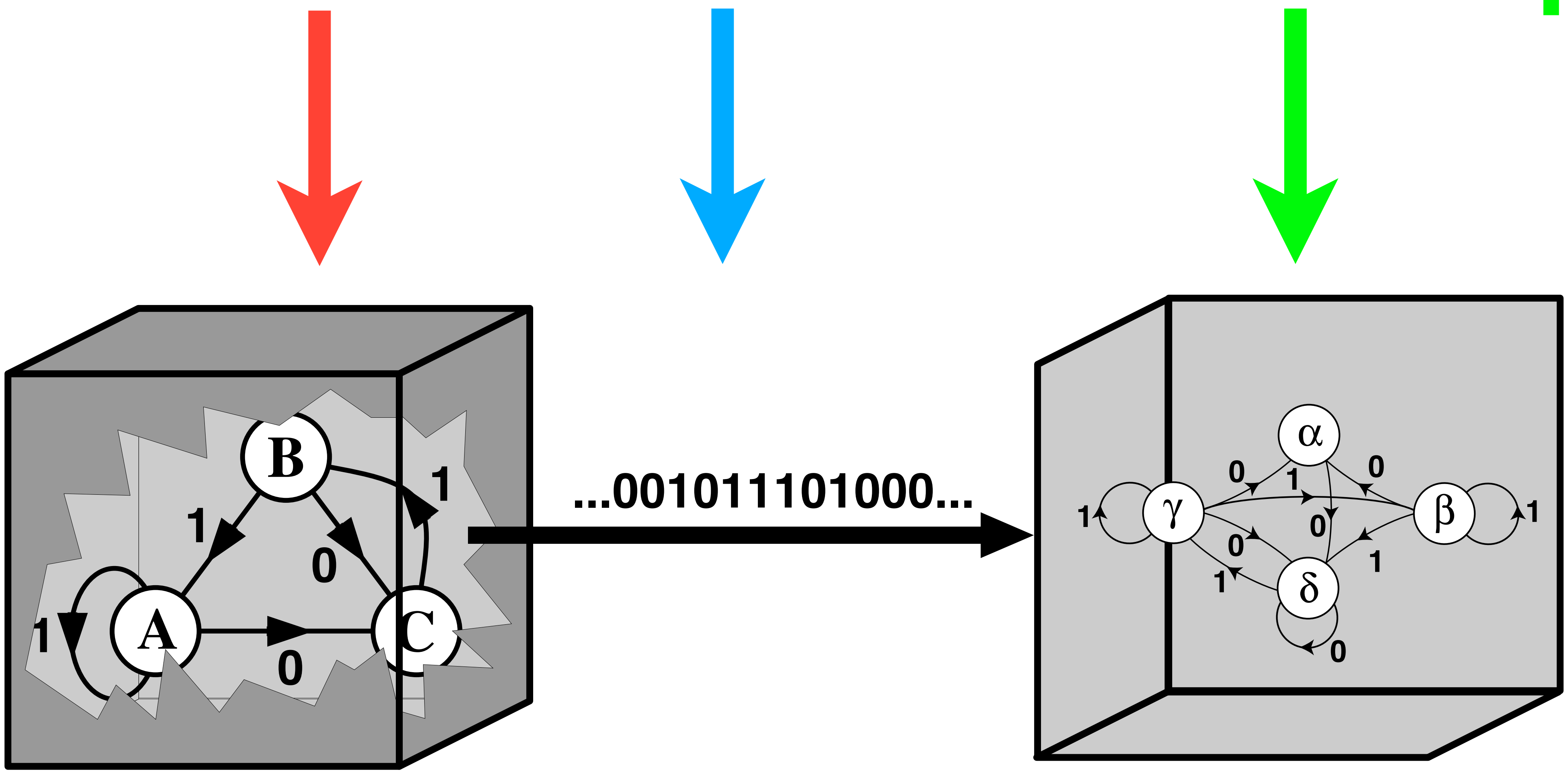
Jim Crutchfield
Complexity Sciences Center
Physics Department
University of California at Davis

Complex Systems Summer School
Institute of American Indian Arts
Santa Fe, New Mexico
20 June 2019

Main Question

Randomness versus Structure?

Now Next Finish Up



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

The Learning Channel

Information in Complex Systems

Now:

- Algorithmic Basis of Probability
- Information Theory
- Information Measures

Later:

- Measuring Structure
- Intrinsic Computation
- Optimal Models
- Physics of Information

Information in Complex Systems

References? For example:

Stanislaw Lem, *Chance and Order*, New Yorker **59** (1984) 88-98.

T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, Second Edition (2006) Chapters 1 - 7.

M. Li and P.M.B. Vitanyi, *An Introduction to Kolmogorov Complexity and its Applications*, Springer, New York (1993).

J. P. Crutchfield and D. P. Feldman, "Regularities Unseen, Randomness Observed: Levels of Entropy Convergence", CHAOS **13**:1 (2003) 25-54.

R. G. James, C. J. Ellison, and J. P. Crutchfield, "Anatomy of a Bit: Information in a Time Series Observation", CHAOS **21**:1 (2011) 037109.

J. P. Crutchfield, "Between Order and Chaos", Nature Physics **8** (January 2012) 17-24.

A. B. Boyd and J. P. Crutchfield, "Demon Dynamics: Deterministic Chaos, the Szilard Map, and the Intelligence of Thermodynamic Systems", Physical Review Letters 116 (2016) 190601.

See http://csc.ucdavis.edu/~cmg/

See online course: http://csc.ucdavis.edu/~chaos/courses/ncaso/

Processes and Their Models ...

Main questions now:

- How do we characterize measured processes?
  - Degrees of unpredictability & randomness?
  - Use probabilities?
  - What correlational structure is there?
- How do we build a model from the process itself?
- How much can we reconstruct about the hidden internal dynamics?

Algorithmic Basis of Probability

Kolmogorov-Chaitin Complexity Theory

The question:

Algorithmic foundation for probability?

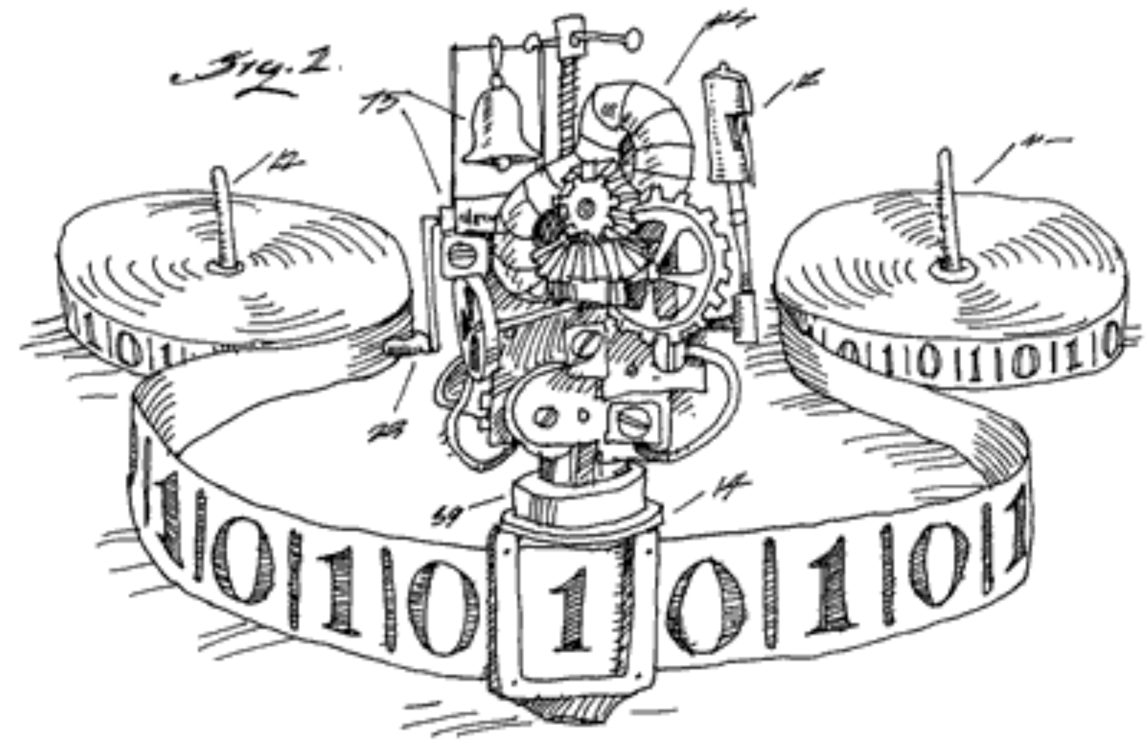
History:

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: Algorithmic 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
$\Rightarrow$ Calculate any input-output function

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

(Python interpreter w/ infinite memory.)

Kolmogorov-Chaitin Complexity of given object:
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 $K(x)$?

$$
\begin{array}{l}
00100100001111110110101010001000 \\
10000101101000110000100011010011 \\
00010011000110011000101000101110 \\
00000011011100000111001101000100
\end{array}
$$

π

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

(Bailey–Borwein–Plouffe 1997)

$$
\begin{array}{l}
10000010100011011111101110011100 \\
01101101001100010110010001010100 \\
00101100011011000110001110111000 \\
10110100010000111000111001110011
\end{array}
$$

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 ...

Sources of Information?

Apparent randomness:

- Uncontrolled initial conditions
- Actively generated: Deterministic chaos

Hidden regularity:

- Ignorance of forces
- Limited capacity to model structure

Information ...

Information as uncertainty and surprise ...

How to formalize?

Shannon's approach:

A measure of surprise.

Connection with Boltzmann's thermodynamic entropy

Self-information of an event $\propto -\log \Pr(\text{event})$.

Predictable: No surprise $-\log 1 = 0$

Completely unpredictable: Maximally surprised

$$-\log \frac{1}{\text{Number of Events}} = \log(\text{Number of Events})$$

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: $H(X) = [\text{bits}]$

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

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 ...

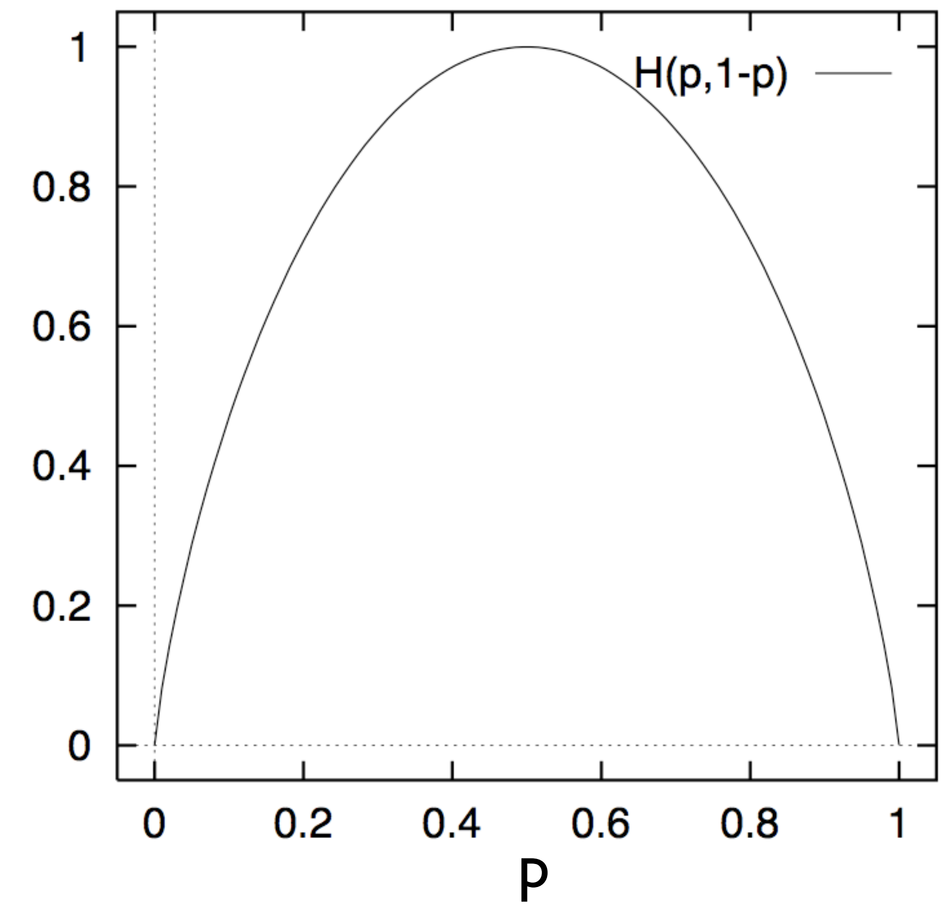
Example: Binary random variable X (Biased Coin)

$$\mathcal{X} = \{0, 1\} \qquad \Pr(1) = p \;\&\; \Pr(0) = 1 - p$$

$H(X)$?

Binary entropy function:

$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$



Fair coin: $p = \frac{1}{2}$

$$H(p) = 1 \text{ bit}$$

Completely biased coin: $p = 0 \ (\text{or } 1)$

$$H(p) = 0 \text{ bits}$$

Recall: $0 \cdot \log 0 = 0$

Information ...

Example: Independent, Identically Distributed (IID) Process over four events

$$\mathcal{X} = \{a, b, c, d\} \qquad \Pr(X) = (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8})$$

Entropy: $H(X) = \frac{7}{4}$ bits

Number of questions to identify the event?

x = a? (must always ask at least one question)

x = b? (this is necessary only half the time)

x = c? (only get this far a quarter of the time)

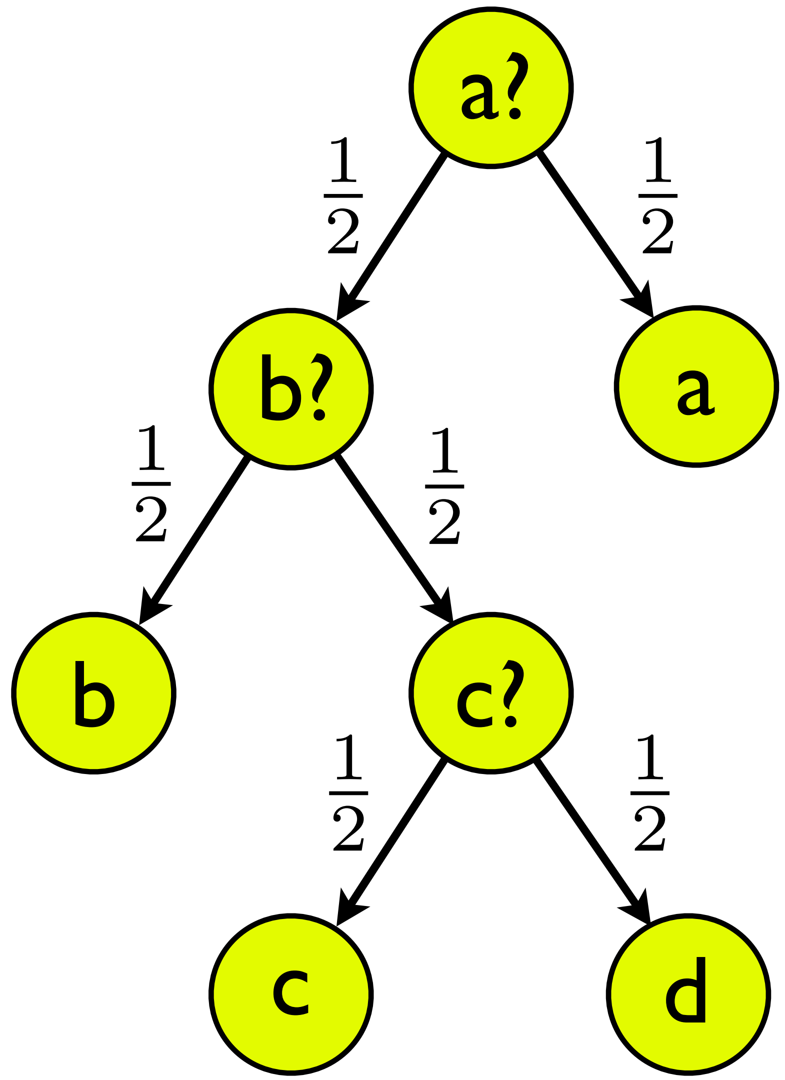
Average number: $1 \cdot 1 + 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} = 1.75$ questions

Interpretation? Optimal way to ask questions.

Information ...

Example: IID Process over four events ...

Average number: $1 \cdot 1 + 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} = 1.75$ questions



$$\Pr(X) = (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8})$$

Information ...

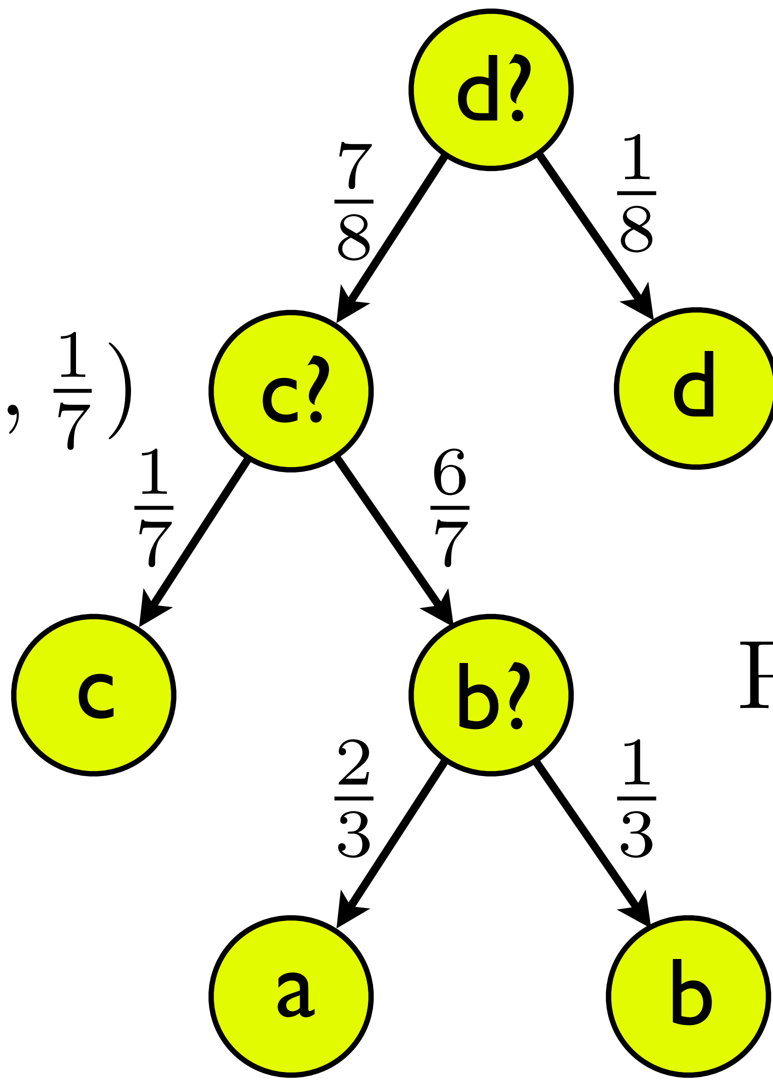
Example: IID Process over four events ...

Query in a different order:

Average number: $1 \cdot 1 + 1 \cdot \frac{7}{8} + 1 \cdot \frac{6}{7} \approx 2.7$ questions

$$\Pr(X) = (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8})$$

$$\Pr(a, b, c) = (\tfrac{4}{7}, \tfrac{2}{7}, \tfrac{1}{7})$$



$$\Pr(a, b) = (\tfrac{2}{3}, \tfrac{1}{3})$$

Information ...

Example: IID Process over four events

Entropy: $H(X) = \frac{7}{4}$ bits

At each stage, ask questions that are most informative.

Choose partitions of event space that give "most random" measurements.

Theorem:
Entropy gives the smallest number of questions to identify an event, on average.

Information ...

Interpretations of Shannon Entropy:

Observer's *degree of surprise* in outcome of a random variable

Uncertainty *in* random variable

Information required to *describe* random variable

A measure of *flatness* of a distribution

Information ...

Two random variables: $(X, Y) \sim p(x, y)$

Joint Entropy: Average uncertainty in X and Y occurring

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

Independent:

$$X \perp Y \Rightarrow H(X,Y) = H(X) + H(Y)$$

Conditional Entropy: Average uncertainty in X, knowing Y

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

$$H(X|Y) = H(X,Y) - H(Y)$$

Not symmetric: $H(X|Y) \neq H(Y|X)$

Common Information Between Two Random Variables:

$$X \sim p(x) \;\&\; Y \sim p(y)$$

$$(X, Y) \sim p(x, y)$$

Mutual Information:

$$I(X;Y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

Information ...

Mutual Information ...

Properties:

(1) $I(X;Y) \geq 0$

(2) $I(X;Y) = I(Y;X)$

(3) $I(X;Y) = H(X) - H(X|Y)$

(4) $I(X;Y) = H(X) + H(Y) - H(X,Y)$

(5) $I(X;X) = H(X)$

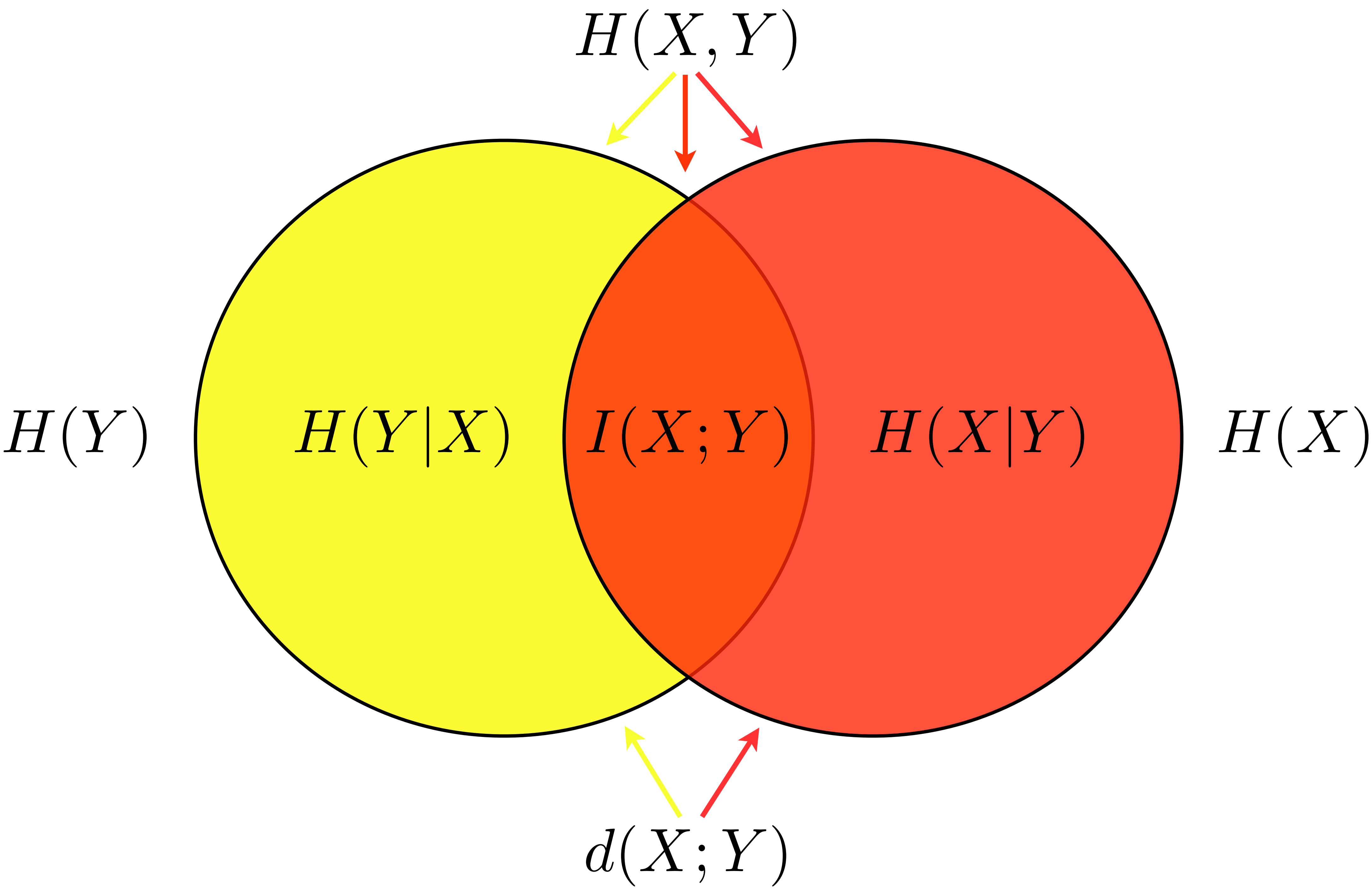
(6) $X \perp Y \Rightarrow I(X;Y) = 0$

Interpretations:

Information one variable has about another
Information shared between two variables
Measure of dependence between two variables

Information ...

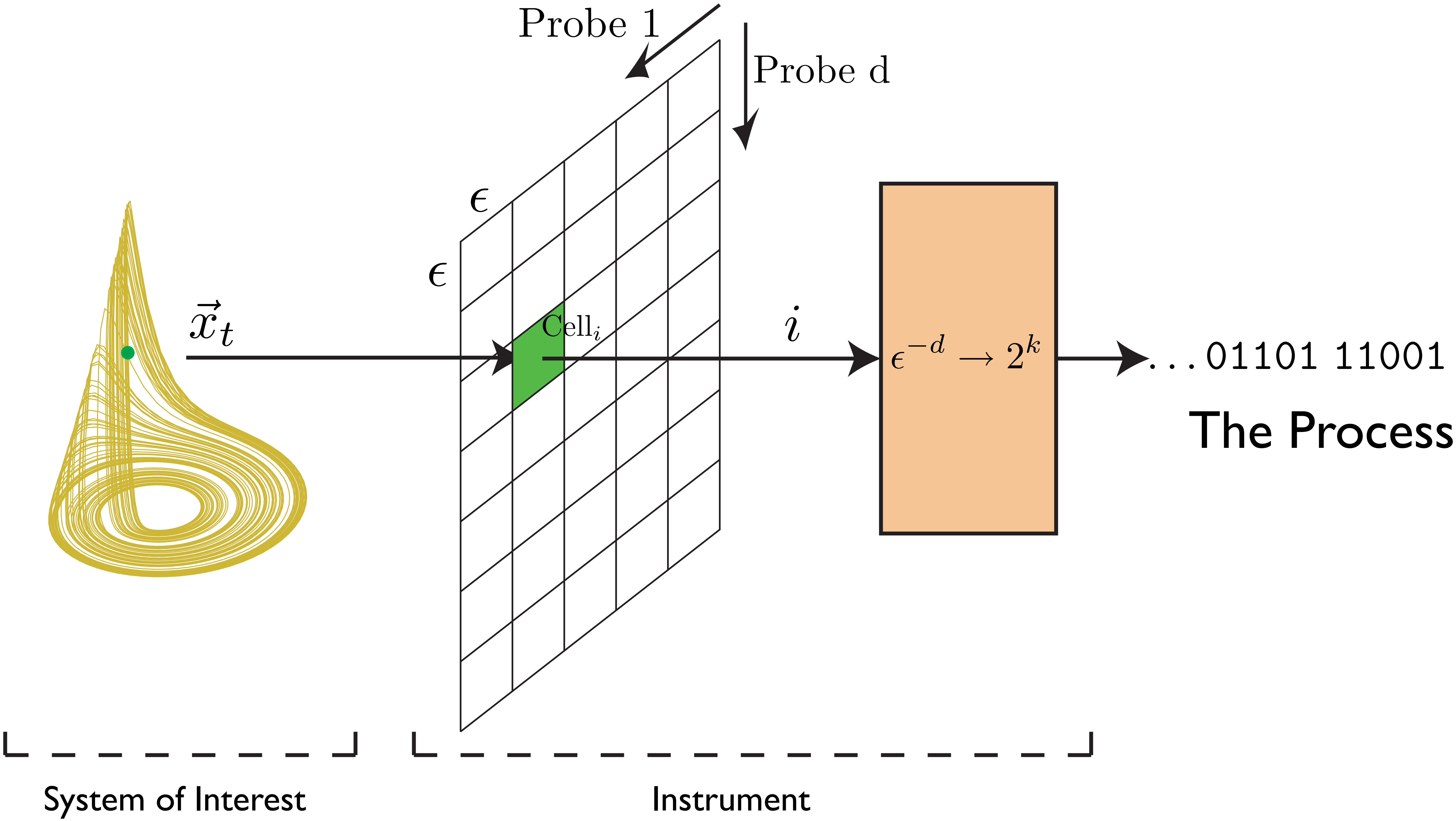
Event Space Relationships of Information Quantifiers:



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

Processes and Their Models



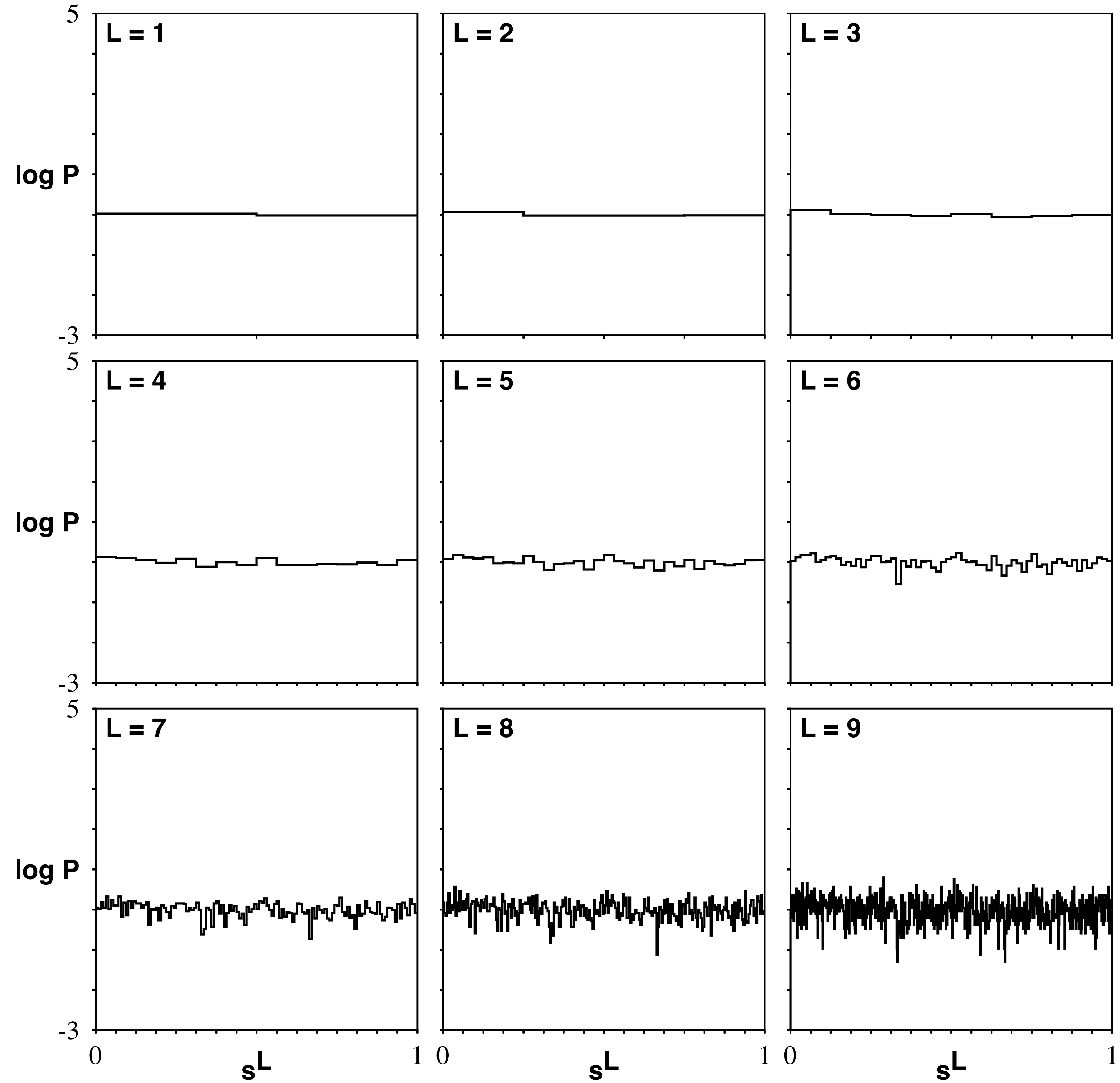
Measurement Channel

Processes and Their Models ...

Models of Stochastic Processes ...

Fair Coin ...

Sequence Distribution: $\Pr(v^L) = 2^{-L}$



Word as binary fraction:

$$s^L = s_1 s_2 \ldots s_L$$

$$\text{“}s^L\text{”} = \sum_{i=1}^{L} \frac{s_i}{2^i}$$

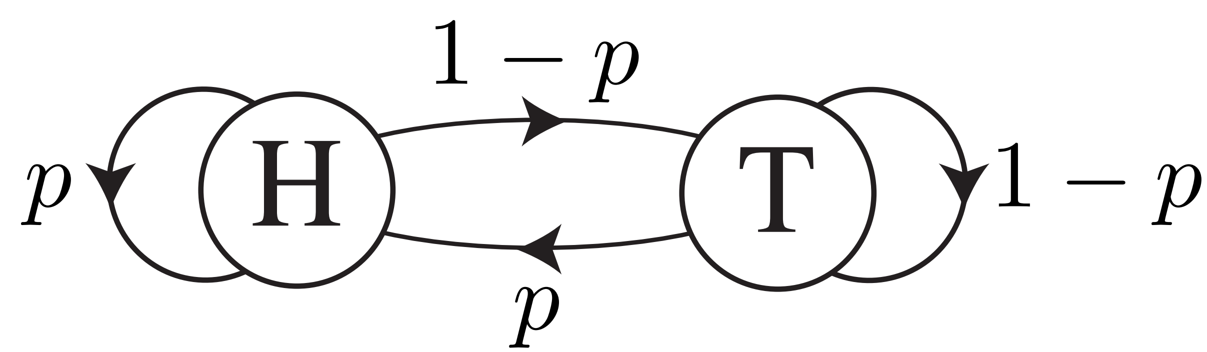
$$s^L \in [0, 1]$$

Processes and Their Models ...

Models of Stochastic Processes ...

Biased Coin: $\mathcal{A} = \{H, T\}$

$$T = \begin{pmatrix} p & 1-p \\ p & 1-p \end{pmatrix}$$



$$\Pr(H) = p$$

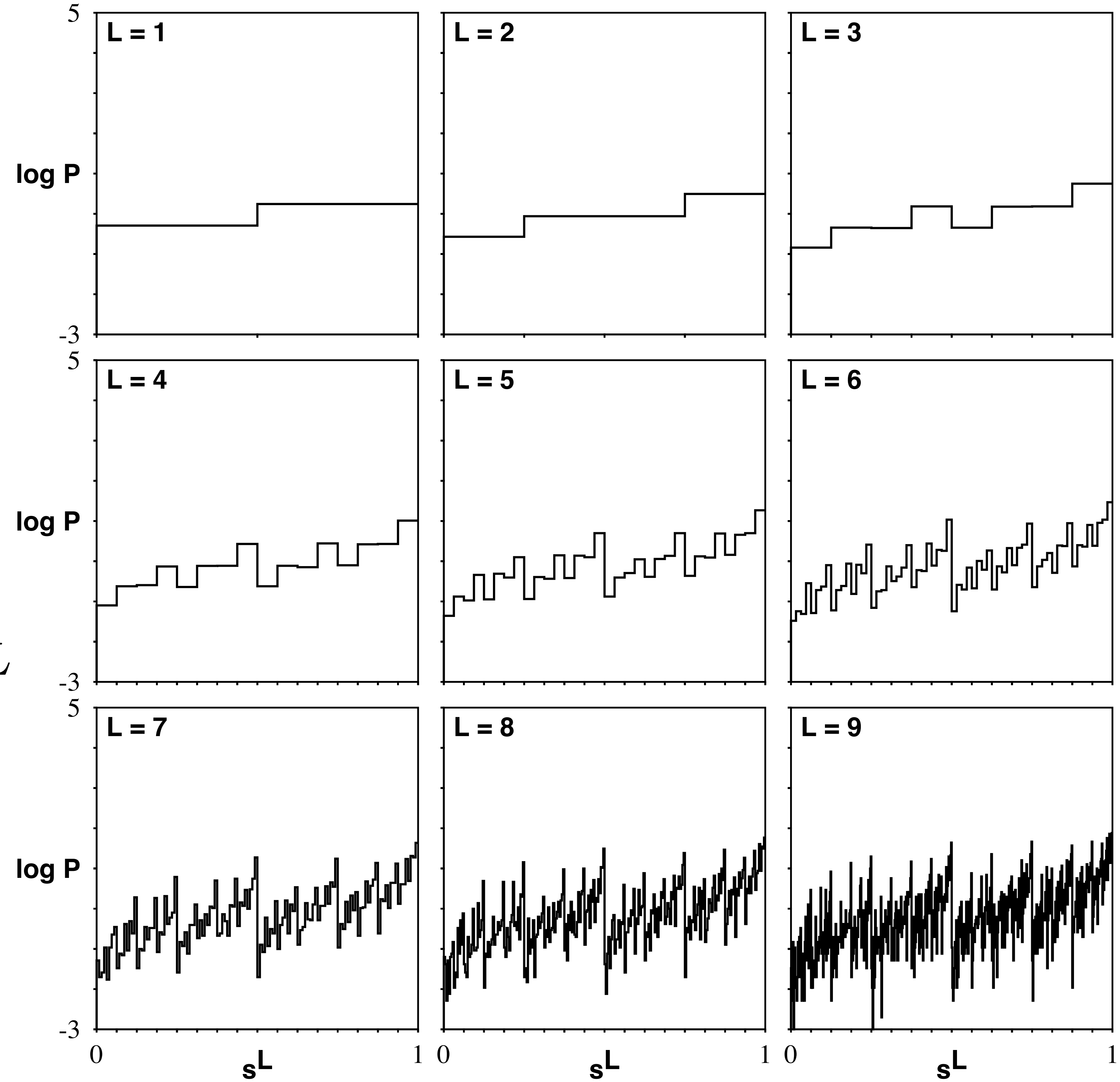
$$\Pr(T) = 1 - p$$

$$\pi = \Pr(p, 1-p)$$

Processes and Their Models ...

Models of Stochastic Processes ...

Biased Coin ...



Sequence Distribution:

$$\Pr(s^L) = p^n (1-p)^{L-n},$$

$$n = \text{Number } H\text{s in } s^L$$

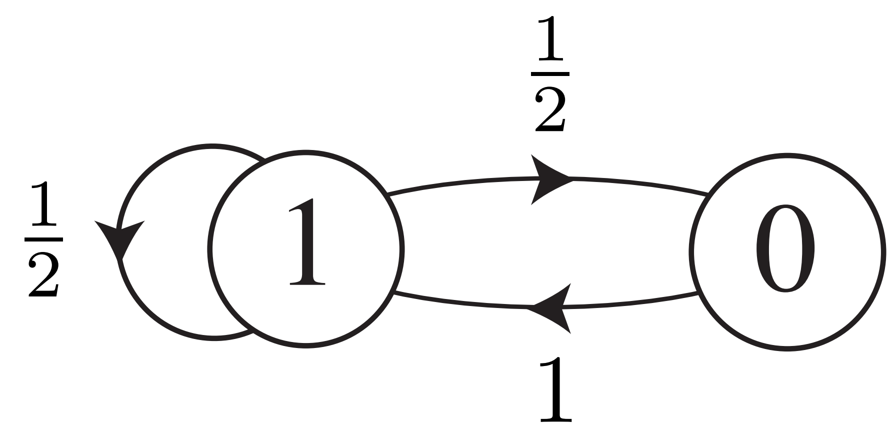
Processes and Their Models ...

Models of Stochastic Processes ...

Golden Mean Process = "No consecutive 0s"

Markov chain over 1-Blocks: $\mathcal{A} = \{0, 1\}$

$$T = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{pmatrix}$$



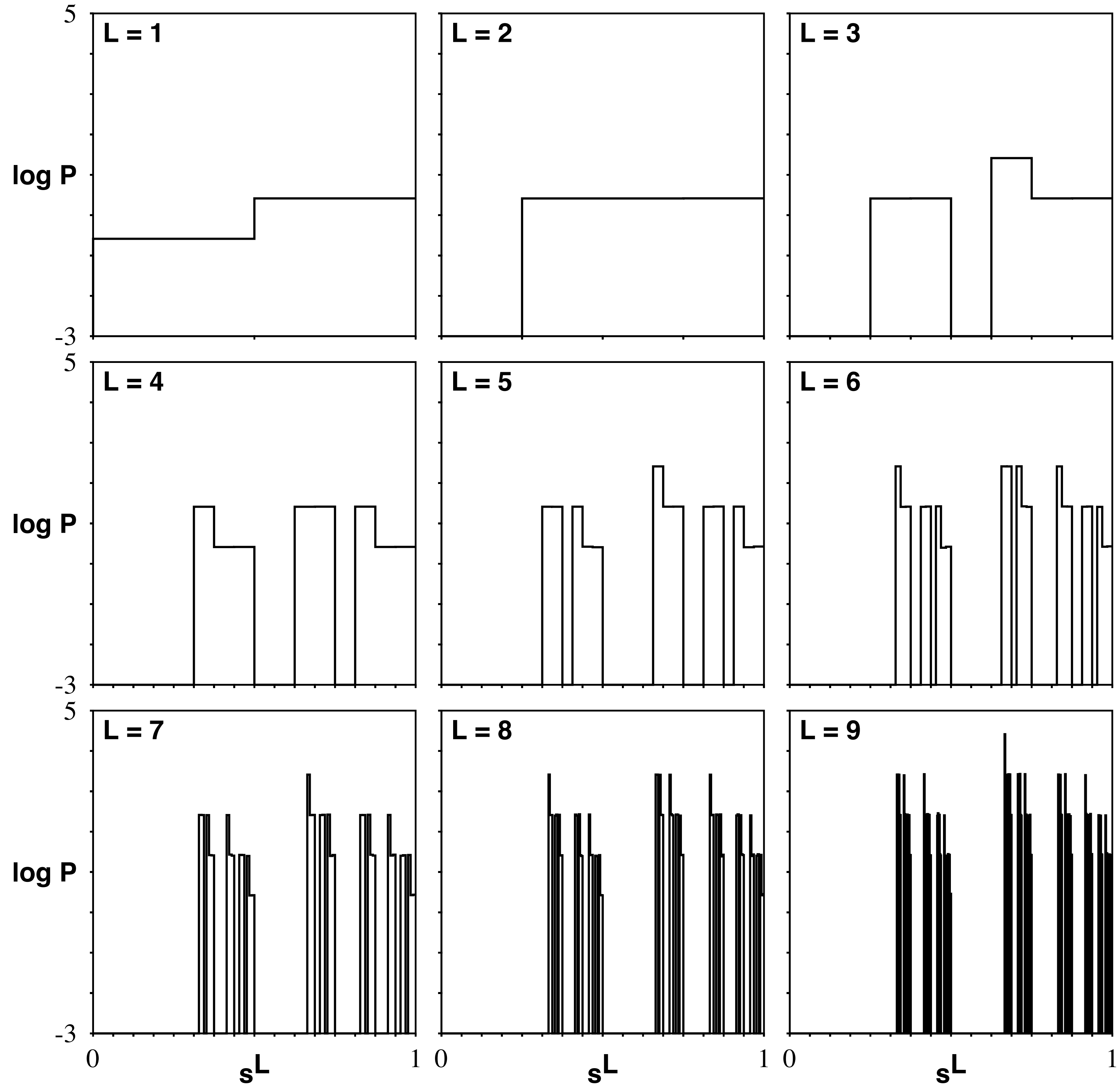
$$\pi = \Pr(V = 1, V = 0)$$
$$= \left(\tfrac{2}{3}, \tfrac{1}{3}\right)$$

As an order-1 Markov chain.
A minimal-order model of the GM Process.

Processes and Their Models ...

Models of Stochastic Processes ...

Golden Mean:



Models of Stochastic Processes ...

Two Lessons:

Structure in the behavior: $\mathrm{supp}\ \Pr(s^L)$

Structure in the distribution of behaviors: $\Pr(s^L)$

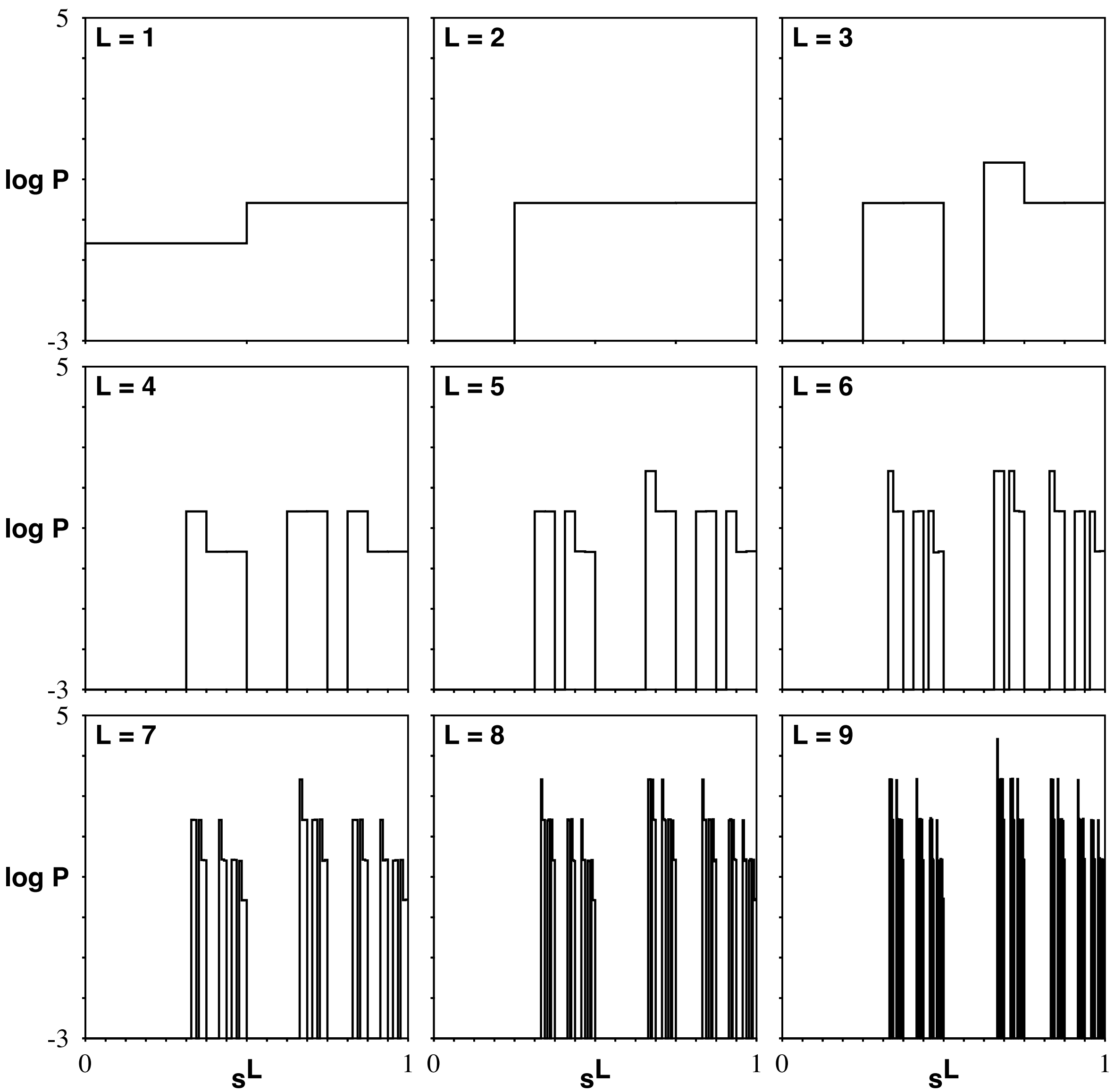
Processes and Their Models ...

Models of Stochastic Processes ...

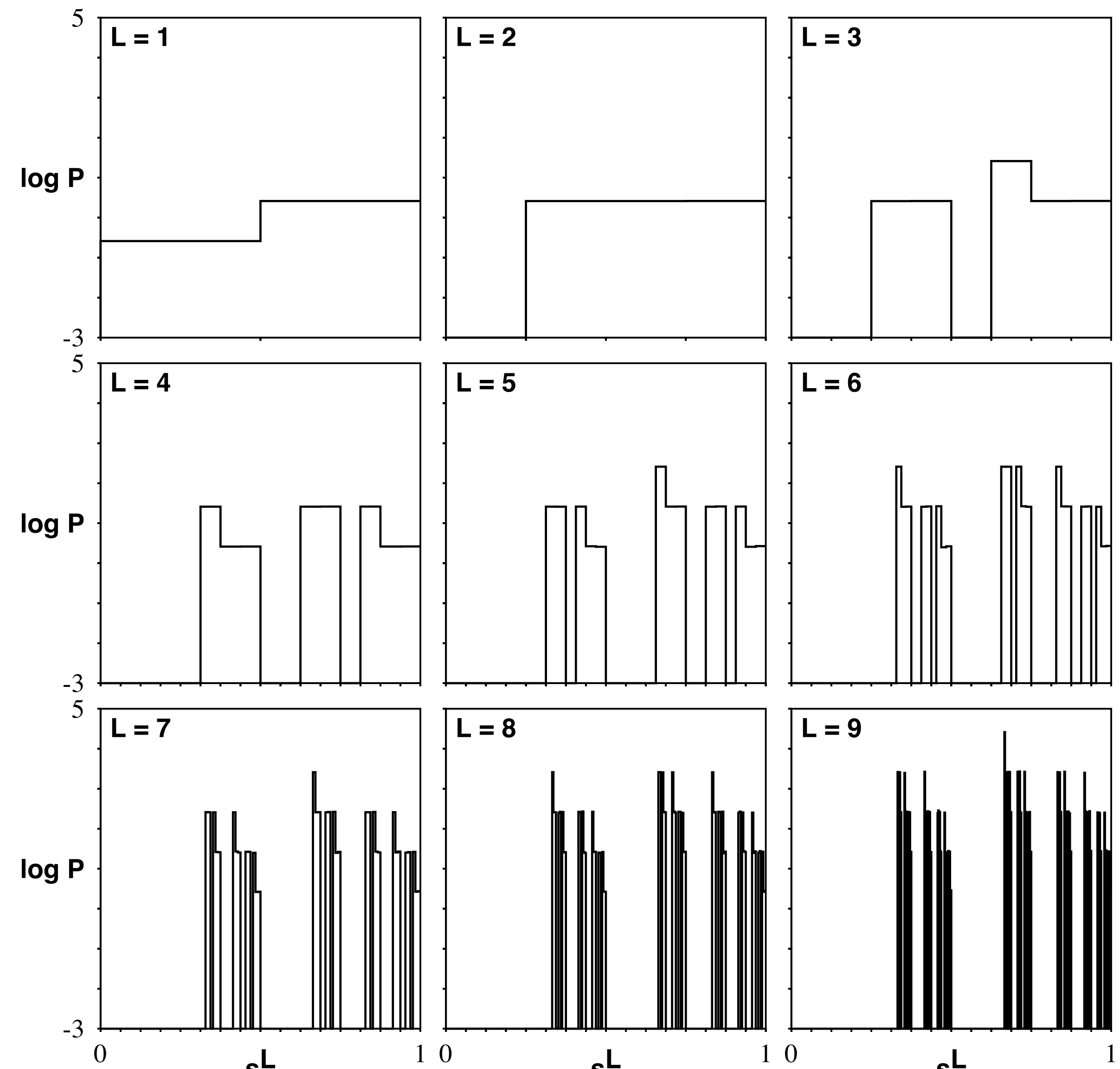
Golden Mean Process ... Sequence distributions:

Internal state sequences

(A = 1; B = 0)



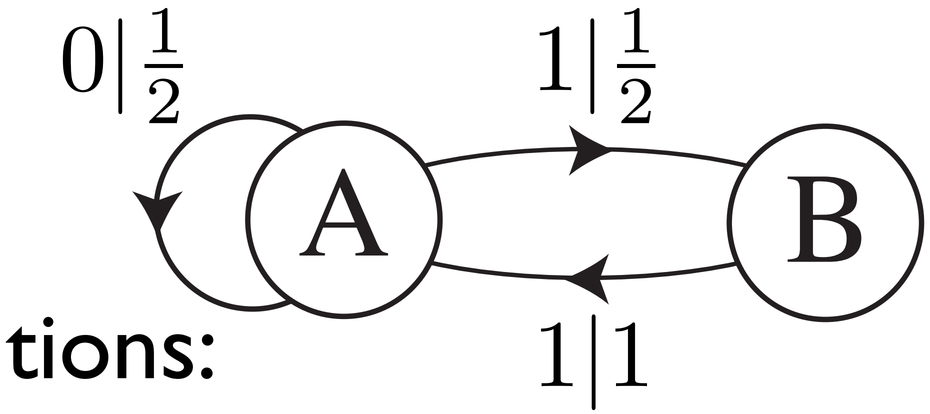
Observed sequences



Same!

Processes and Their Models ...

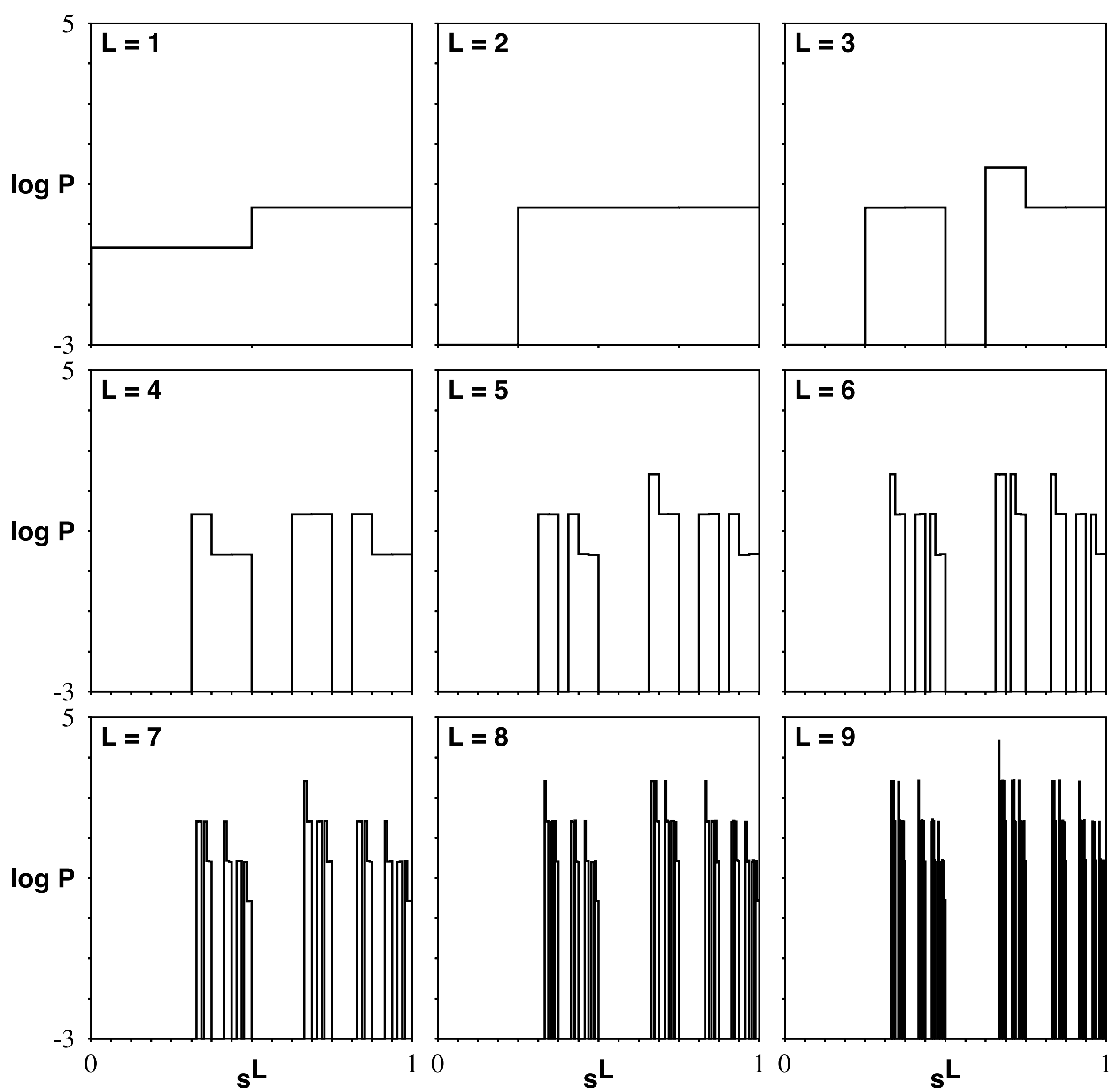
Models of Stochastic Processes ...



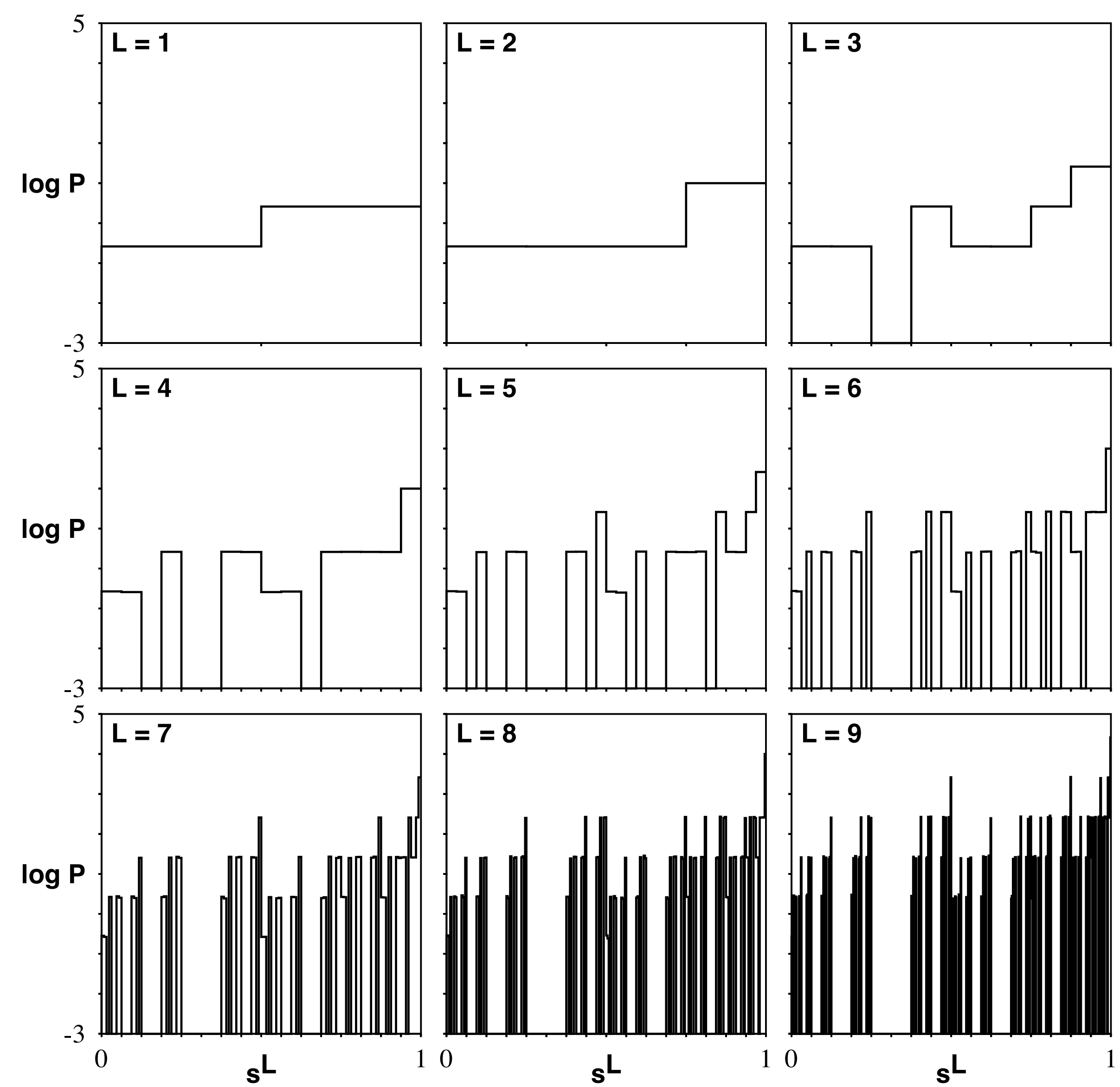
Even Process ... Sequence distributions:

Internal states (= GMP)

(A = 1; B = 0)



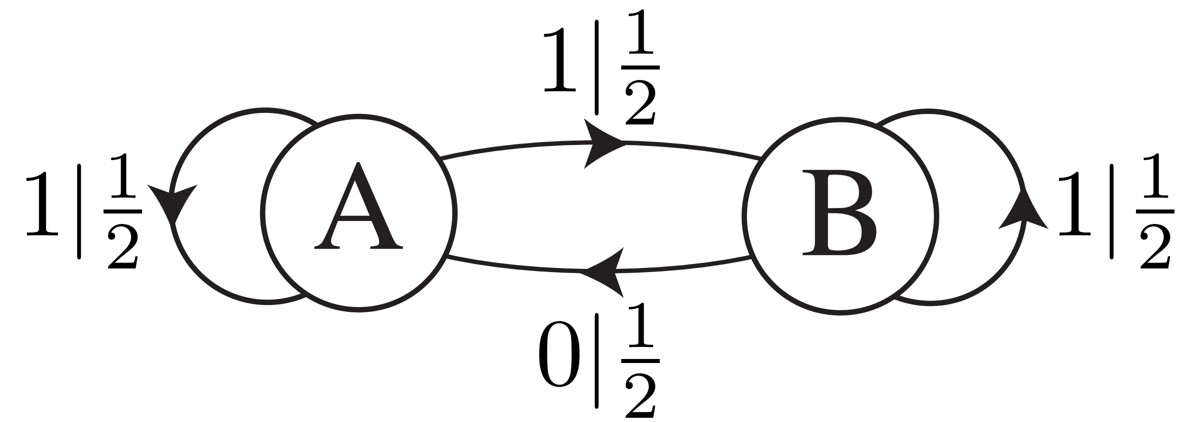
Observed sequences



Rather different!

Processes and Their Models ...

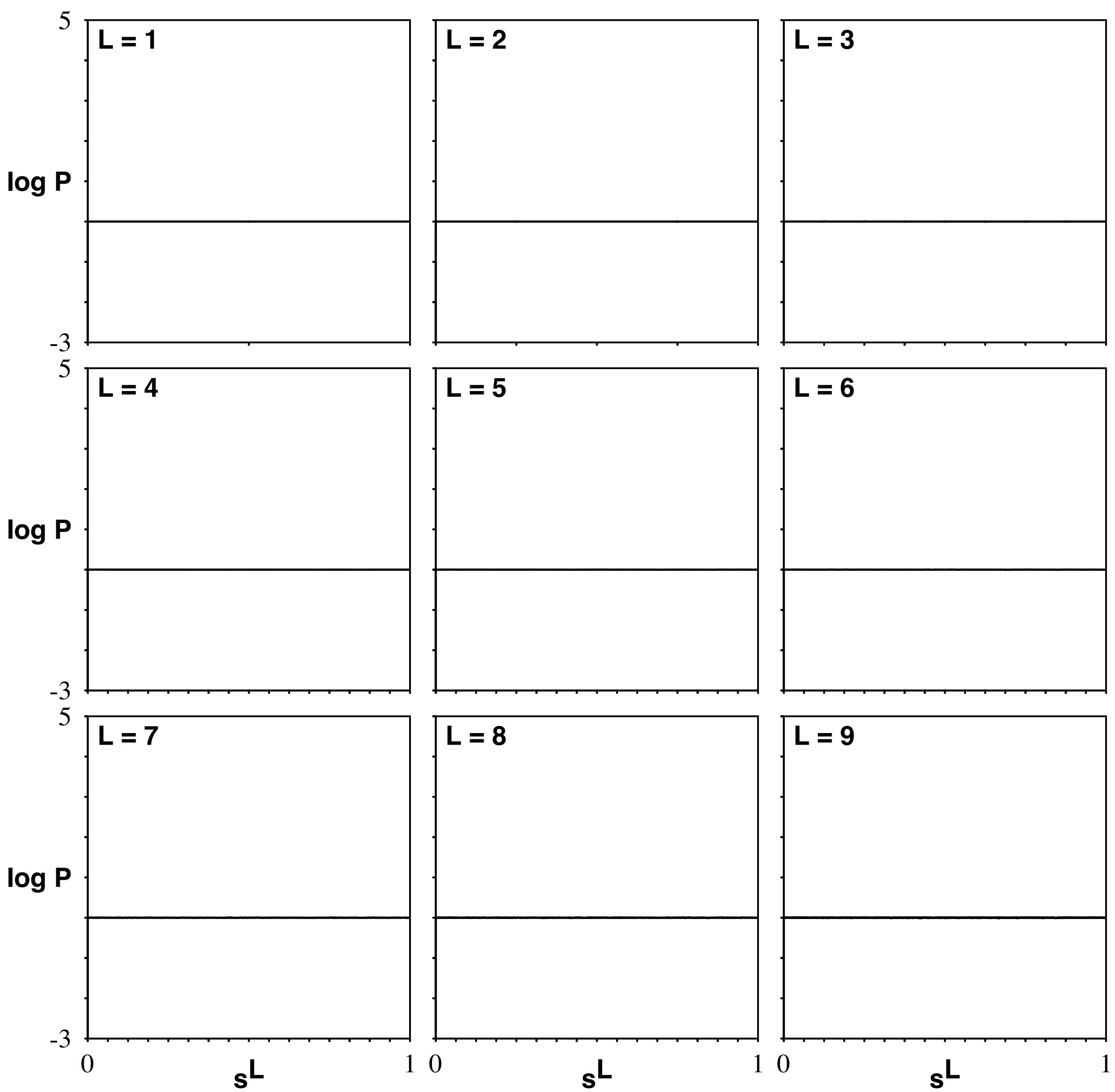
Models of Stochastic Processes ...



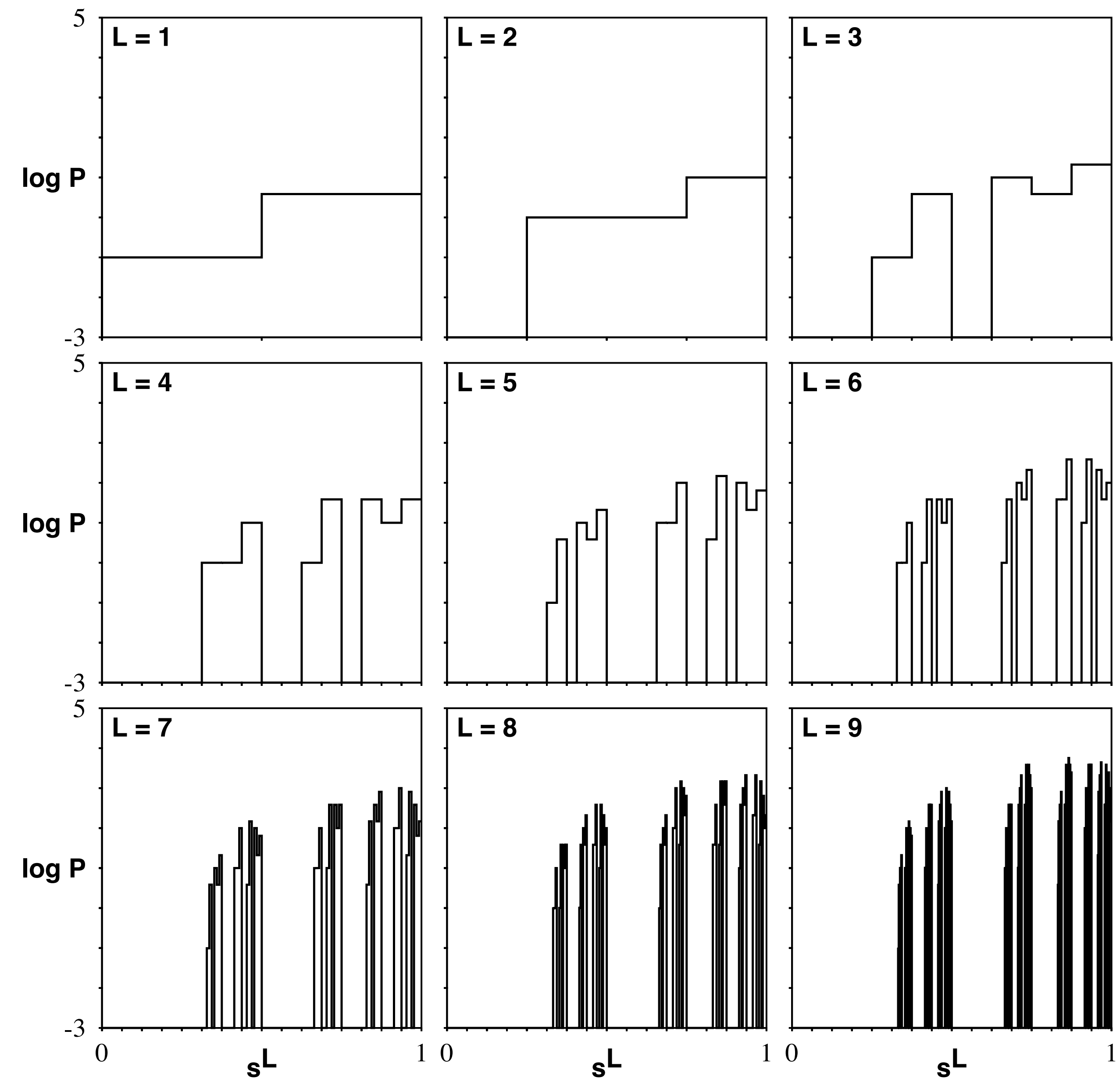
Simple Nonunifilar Source ...

Internal states (= Fair coin)

(A = 1; B = 0)



Observed sequences



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:

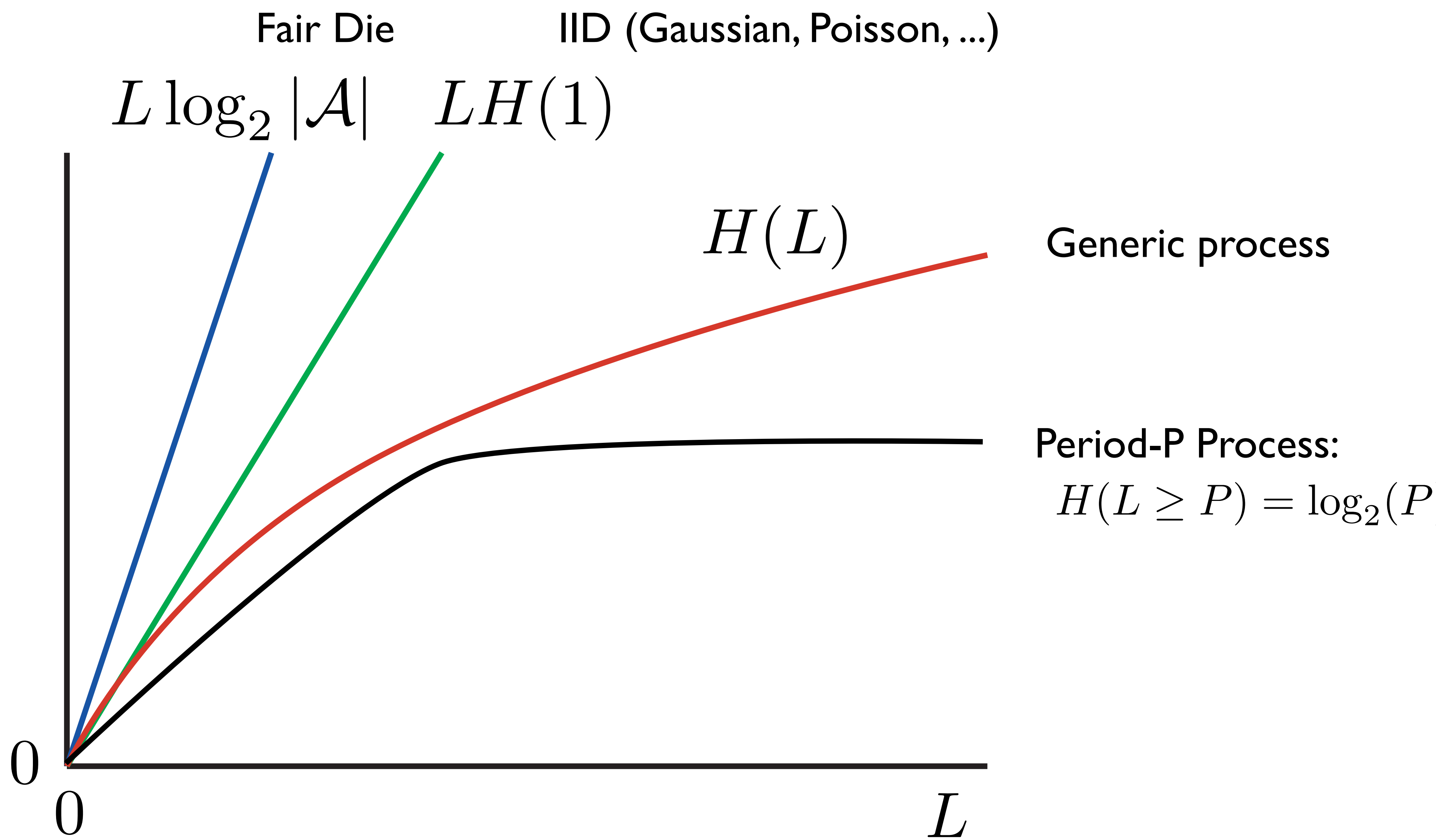
$$H(S_1, S_2, \ldots, S_L) \leq H(S_1, S_2, \ldots, S_L, S_{L+1})$$

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

Information in Processes ...

Entropy Growth for Stationary Stochastic Processes ...

Block Entropy ...

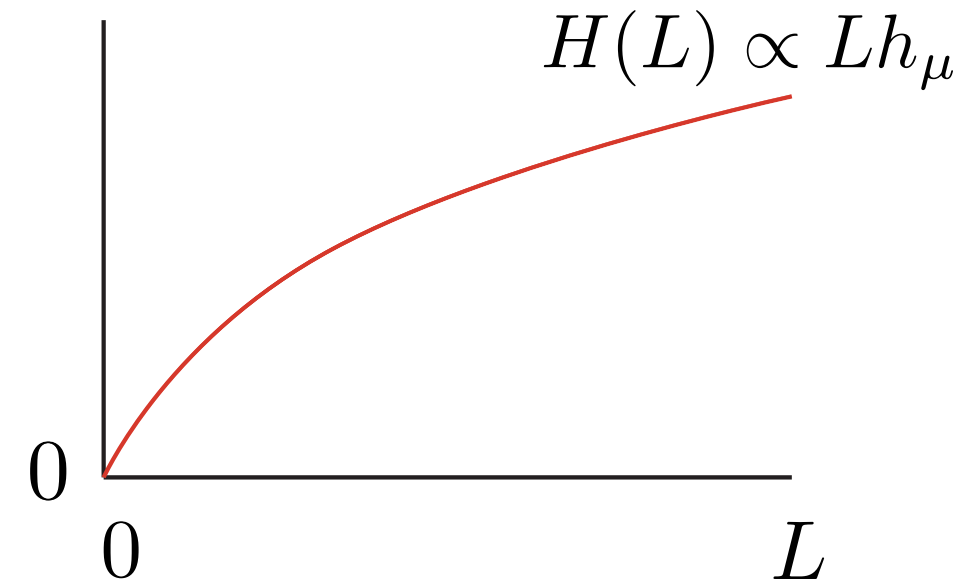


Entropy Rates for Stationary Stochastic Processes:

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

Information in Processes ...

Entropy Convergence:

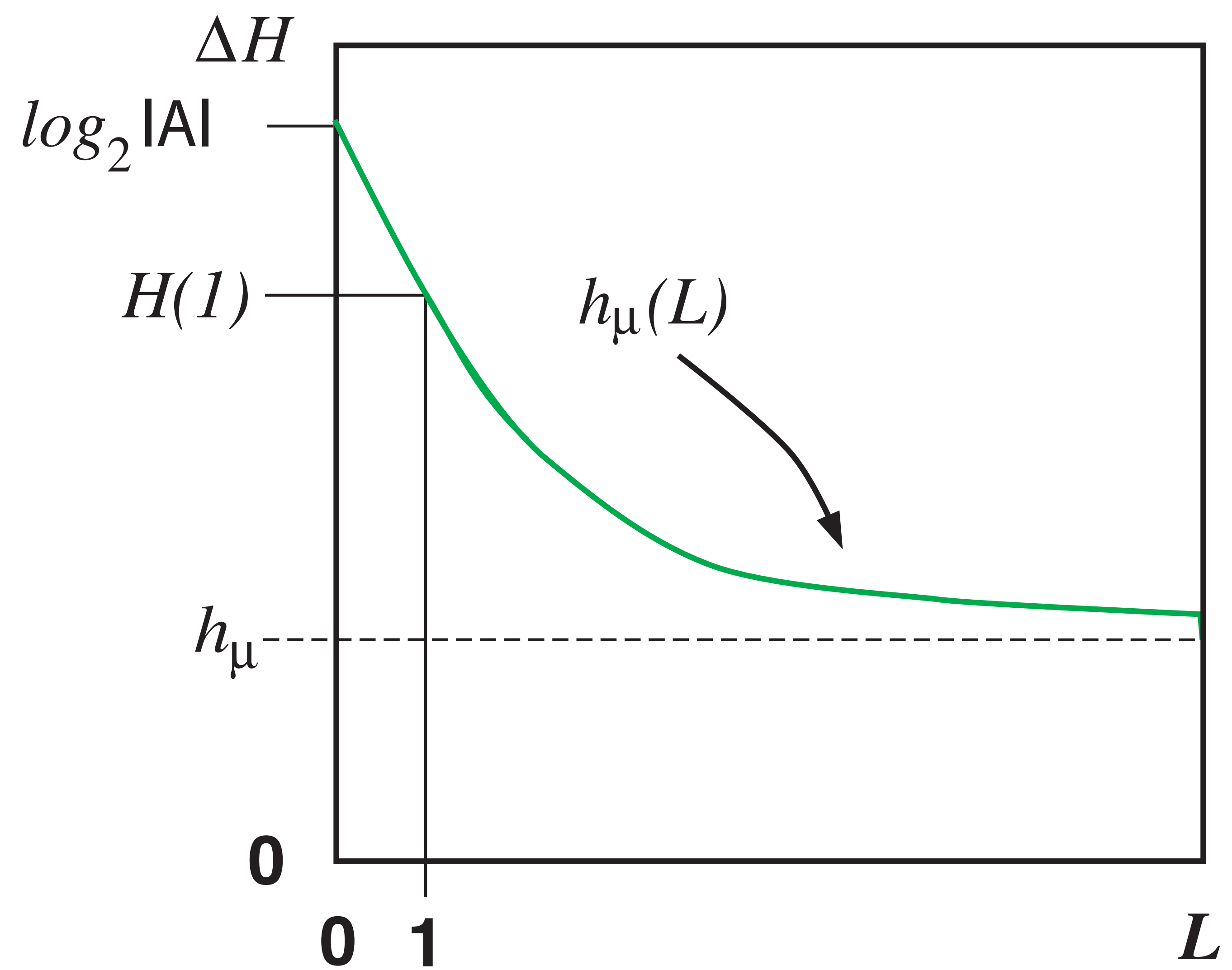
Length-L entropy rate estimate:

$$h_\mu(L) = H(L) - H(L-1)$$

$$h_\mu(L) = \Delta H(L)$$

Monotonic decreasing:

$$h_\mu(L) \leq h_\mu(L-1)$$



Process appears less random
as account for longer correlations

Memory in Processes

Information in Processes ...

Motivation:

Previous: Measures of randomness of information source

Block entropy $H(L)$

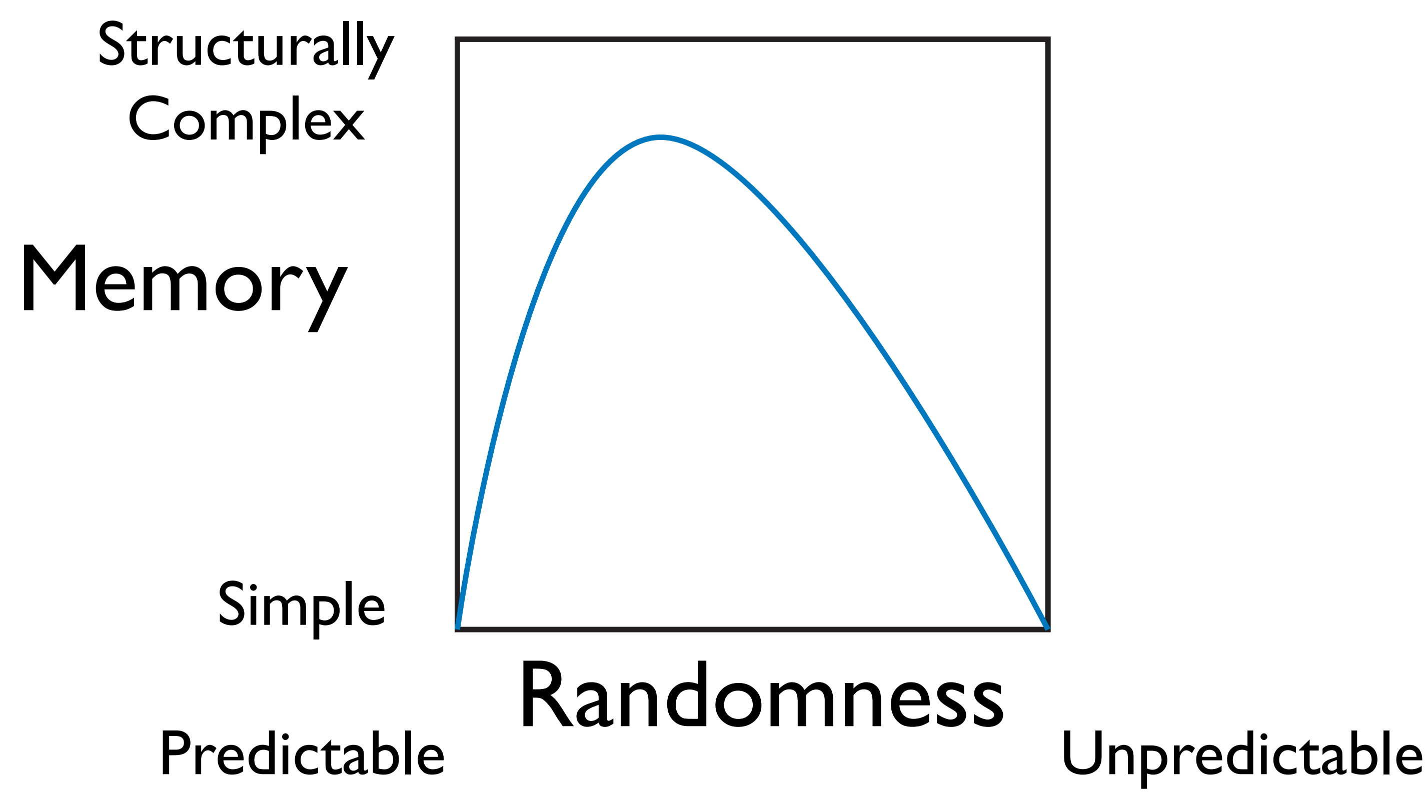
Entropy rate h_μ

Current target point:

Measures of memory & information storage

Big Picture:

Complementary.



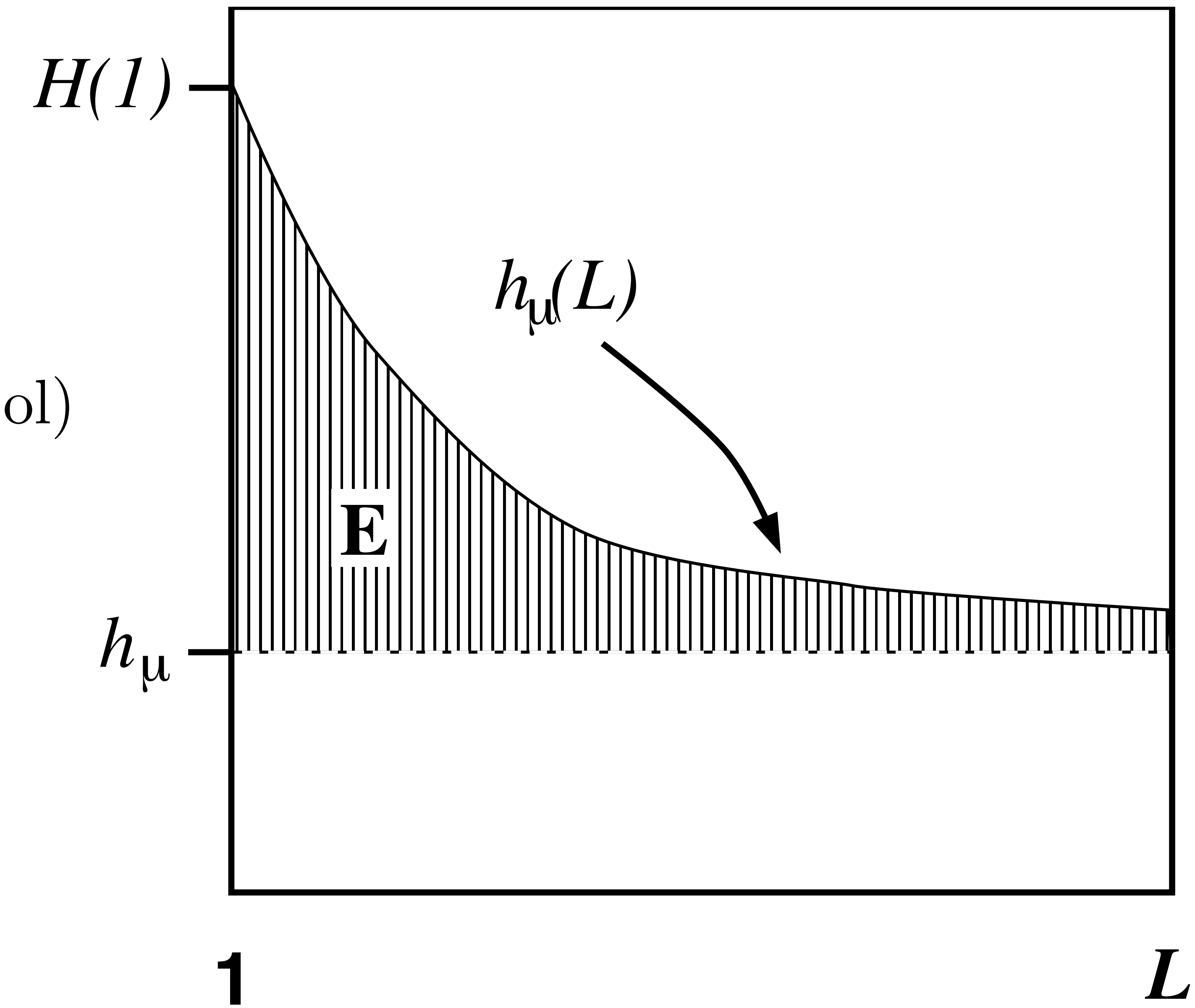
Memory in Processes ...

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.

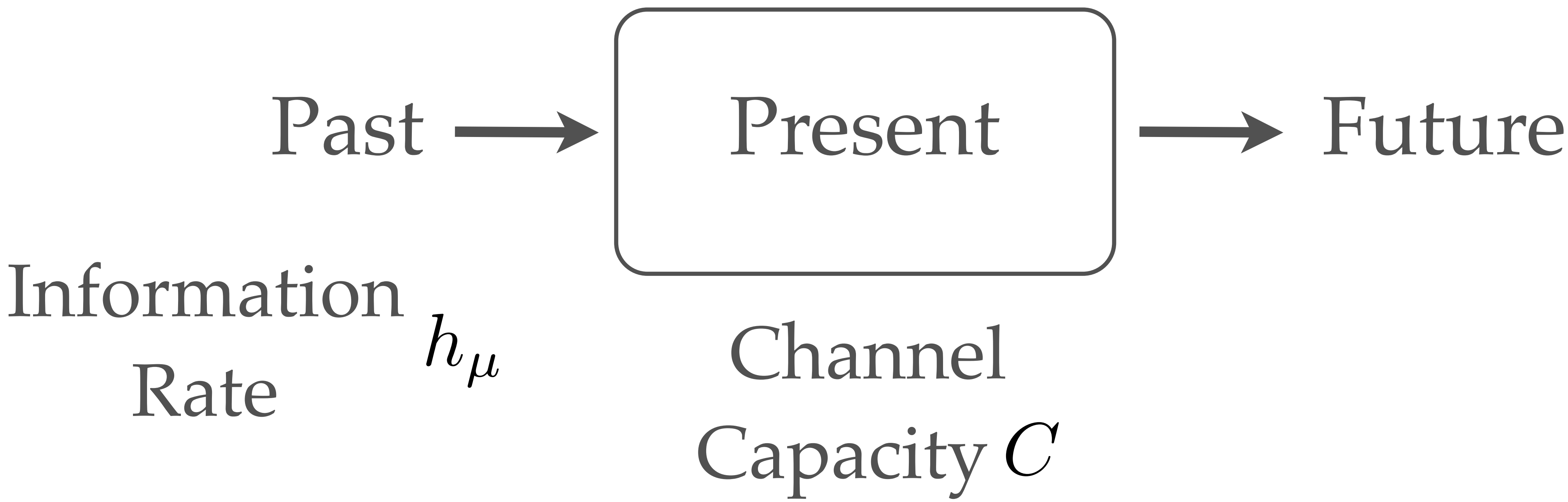
(5) Complementary to entropy rate.

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 $\overrightarrow{X}$:

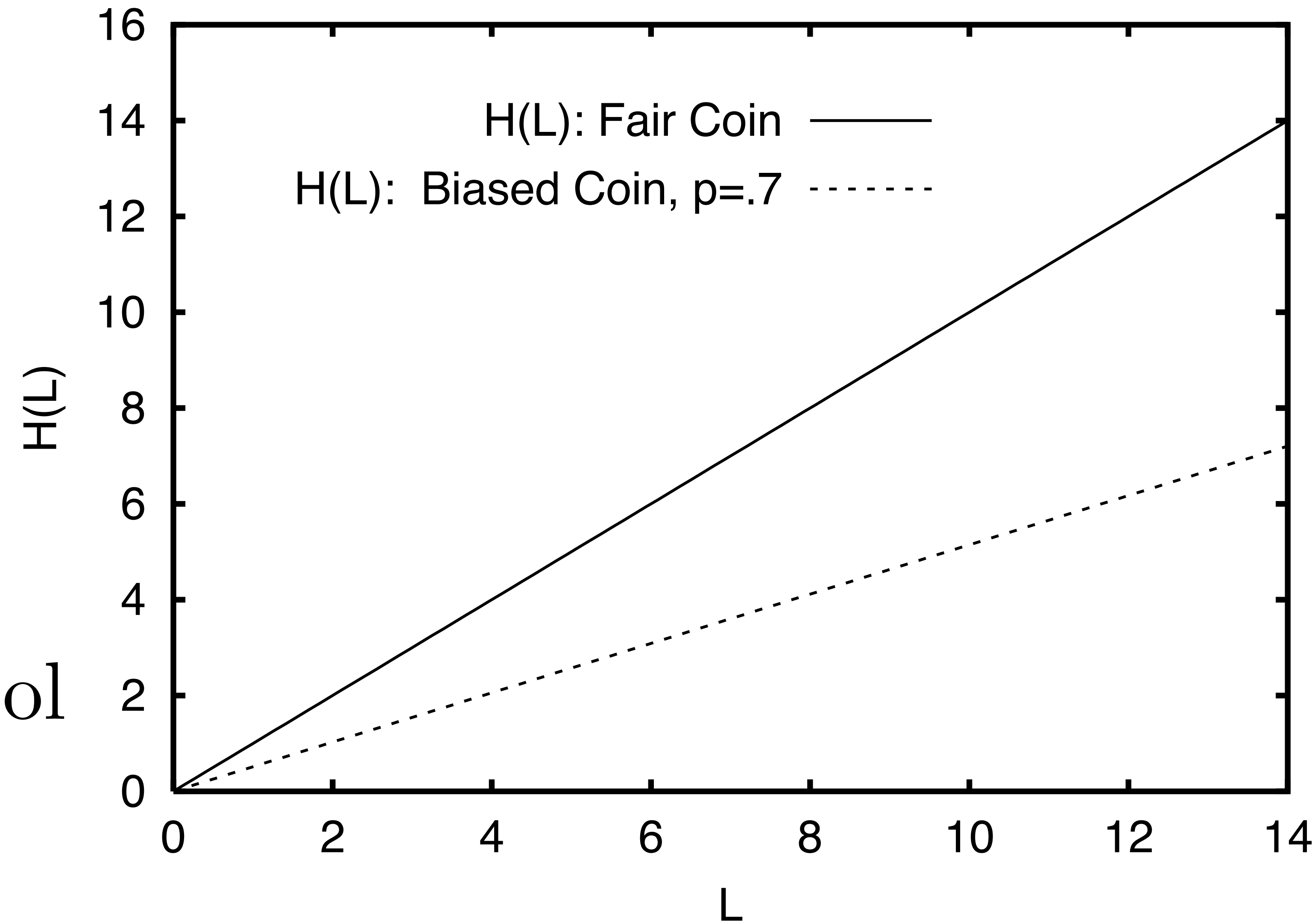


Excess Entropy as Channel Utilization:

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

Memory in Processes ...

Examples of Excess Entropy:



Fair Coin:

$$h_\mu = 1 \text{ bit per symbol}$$

$$\mathbf{E} = 0 \text{ bits}$$

Biased Coin:

$$h_\mu = H(p) \text{ bits per symbol}$$

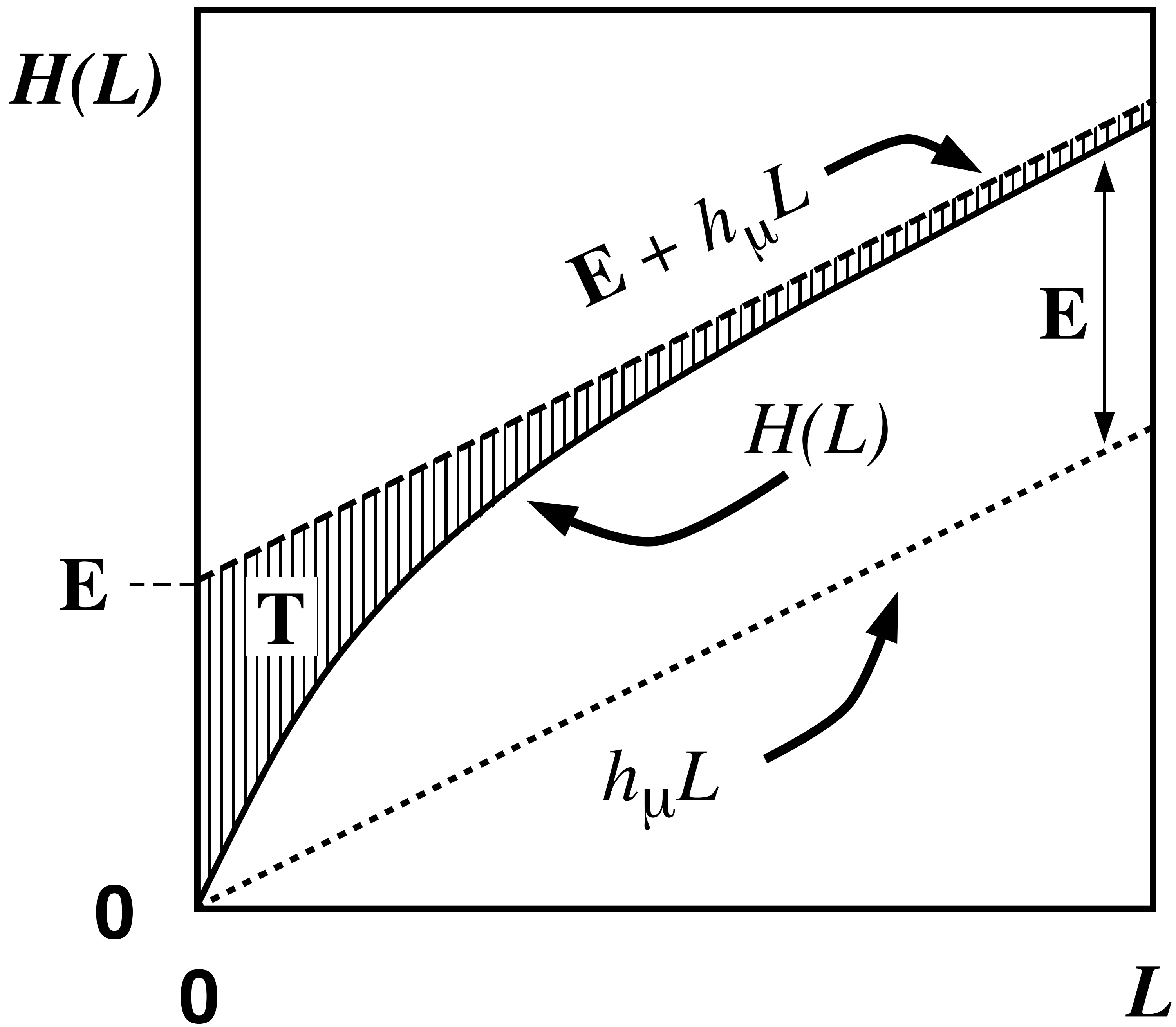
$$\mathbf{E} = 0 \text{ bits}$$

Any IID Process:

$$h_\mu = H(X) \text{ bits per symbol}$$

$$\mathbf{E} = 0 \text{ bits}$$

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.
...

Algorithmic Basis of Information

Kolmogorov-Chaitin Complexity versus Shannon Information

KC Complexity versus Shannon Information

Consider average KC Complexity of source:

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

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.

Information in Complex Systems

Done:
- Algorithmic Basis of Probability
- Information Theory
- Information Measures

Next:
- Measuring Structure
- Intrinsic Computation
- Optimal Models
- Physics of Information

See online course:
http://csc.ucdavis.edu/~chaos/courses/ncaso/