Patterns in Time

Foundations and Frontiers of Complex Systems
First Latin American Complex Systems Summer School
San Carlos de Bariloche, Argentina

James P. Crutchfield

Complexity Sciences Center, Director

Physics Department

University of California, Davis &

Santa Fe Institute

cse.ucdavis.edu/~chaos

5 December 2008

Two Lectures

- First Lecture
  - Why Must We Model?
  - Randomness & Beyond: Information Theory
- Second Lecture
  - Structure: Computational Mechanics
  - Learning: Rate Distortion Theory
  - The Future: Interactive Learning

Complex Systems Topics Covered

- Randomness & Its Origins
- Kinds of Information
- Representing Structure: Automata
- Intrinsic Computation
- Causal Inference

Main References

- [RURO]:

  J. P. Crutchfield & D. P. Feldman

  Regularities Unseen, Randomness Observed: Levels of Entropy Convergence

  http://cse.ucdavis.edu/~cmg/compmech/pubs/ruro.htm

- [CMPPSS]:

  C. R. Shalizi & J. P. Crutchfield

  Computational Mechanics: Pattern and Prediction, Structure and Simplicity

  http://cse.ucdavis.edu/~cmg/compmech/pubs/cmppss.htm

- [NCASO] Course on Natural Computation:

  http://cse.ucdavis.edu/~cmg/courses/ncaso/

  Lectures 11 - 16.

- These notes @ http://cse.ucdavis.edu/~chaos/chaos/talks.htm

Why Must We Model?

Three reasons!

Why We Must Model I

Nature spontaneously organizes

Emergent structures

Why We Must Model 2

- Engineered systems spontaneously organize
  - Internet route flapping
  - Power-law Internet self-organization
  - Financial markets crash
  - Power grids fail spectacularly
  - Social pattern formation on the web
  - ...

And so ... 1 + 2 = ...

- Problem:

  Emergent structures not given directly by the system coordinates, governing equations of motion, or design plan

- Consequence:

  Each needs its own explanatory basis

Why We Must Model 3

- Fundamental Mathematics: Intrinsic Randomness

- Nonlinear dynamical systems:

  Chaotic systems: Shannon entropy rate

  $h_\mu > 0$ [Kolomogorov 1958]

- Kolmogorov-Chaitin complexity of Data:

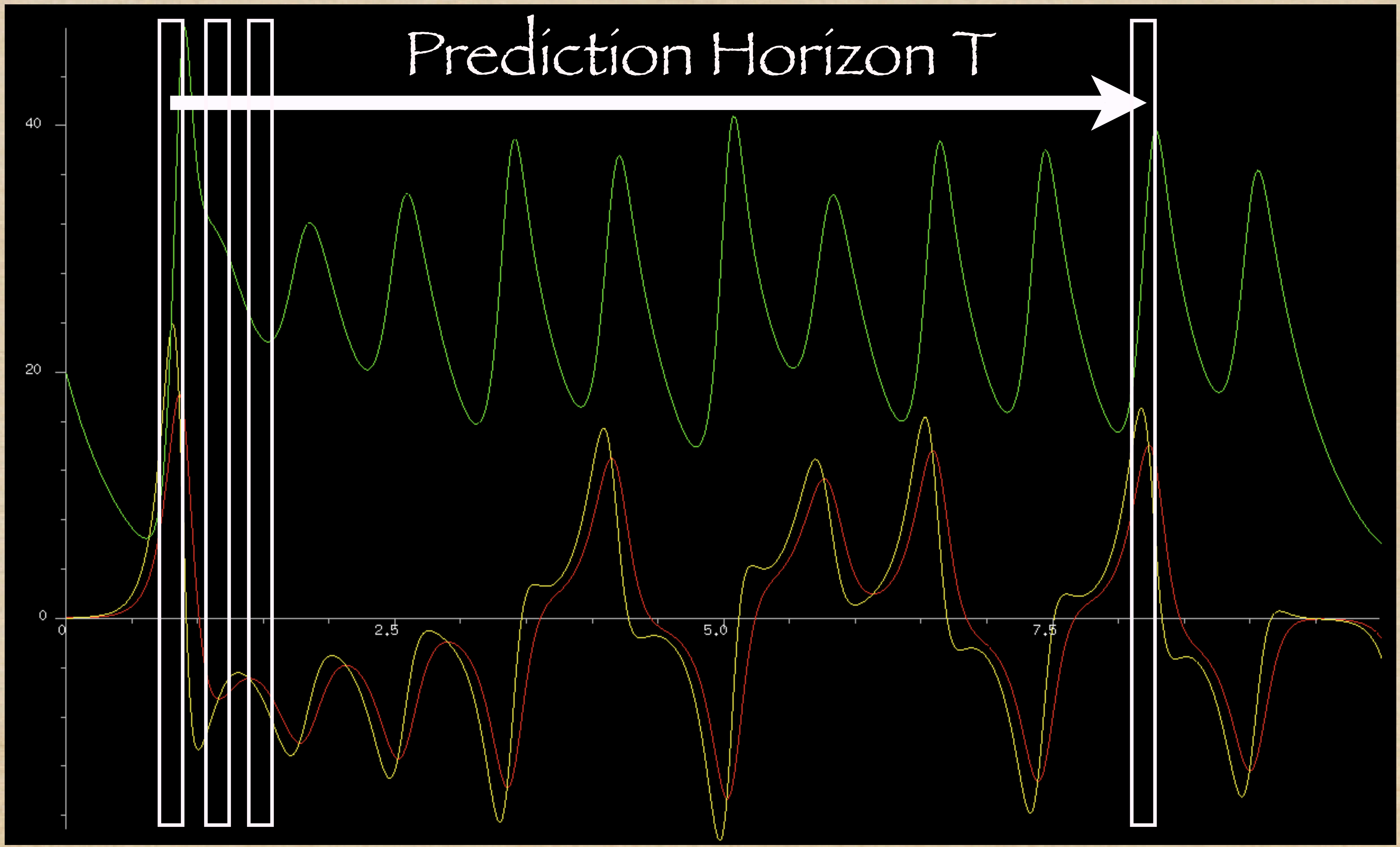
  |Shortest Turing Machine Program to Predict Data|

  [Kolomogorov 1963, Chaitin 1964]

- KC complexity = f(Shannon entropy rate):

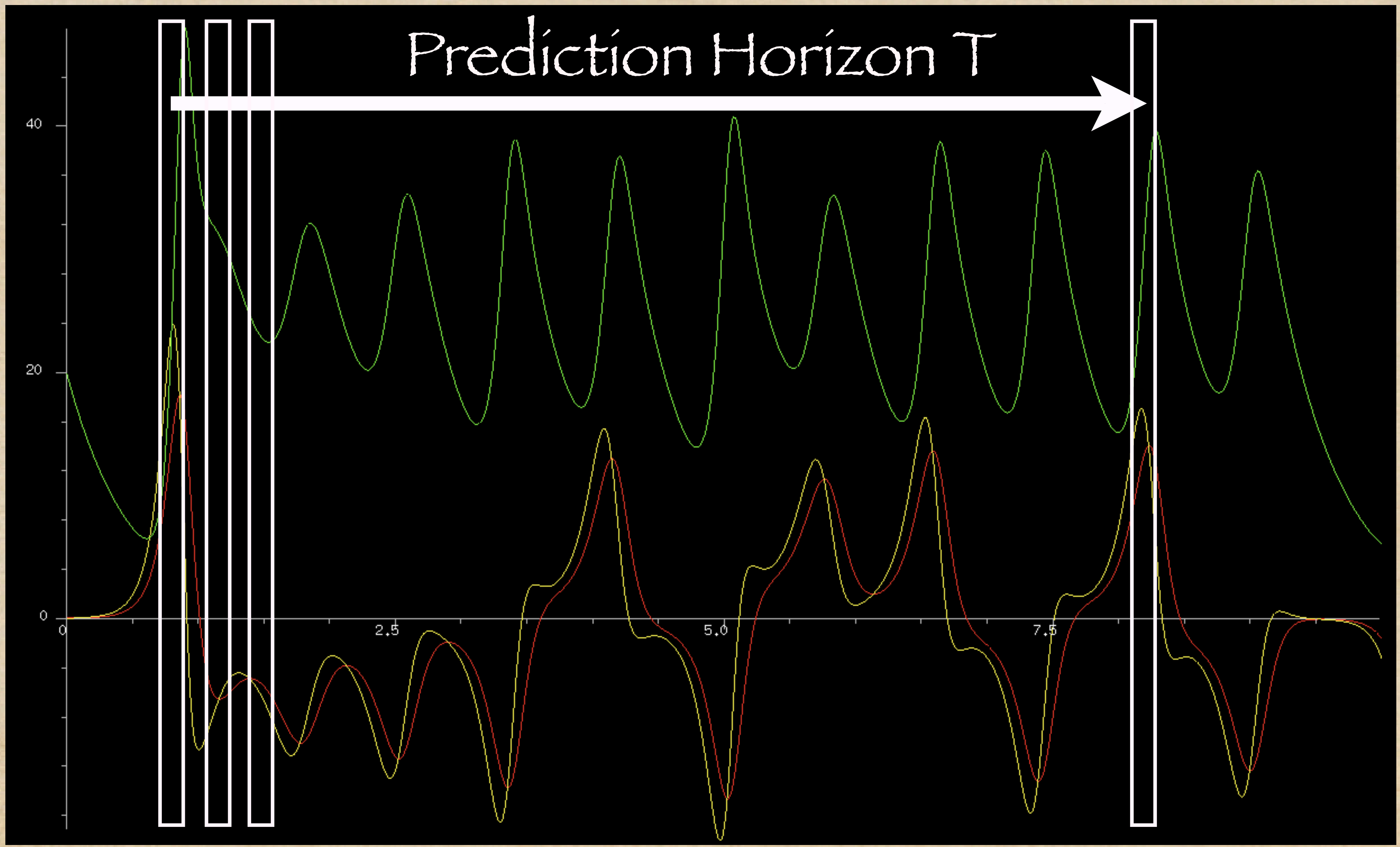
  $$|\mathrm{Program}| \propto e^{h_\mu |\mathrm{Data}|}$$ [Brudno 1978]

Exponential Increase in Prediction Resources



Exponential Increase in Prediction Resources

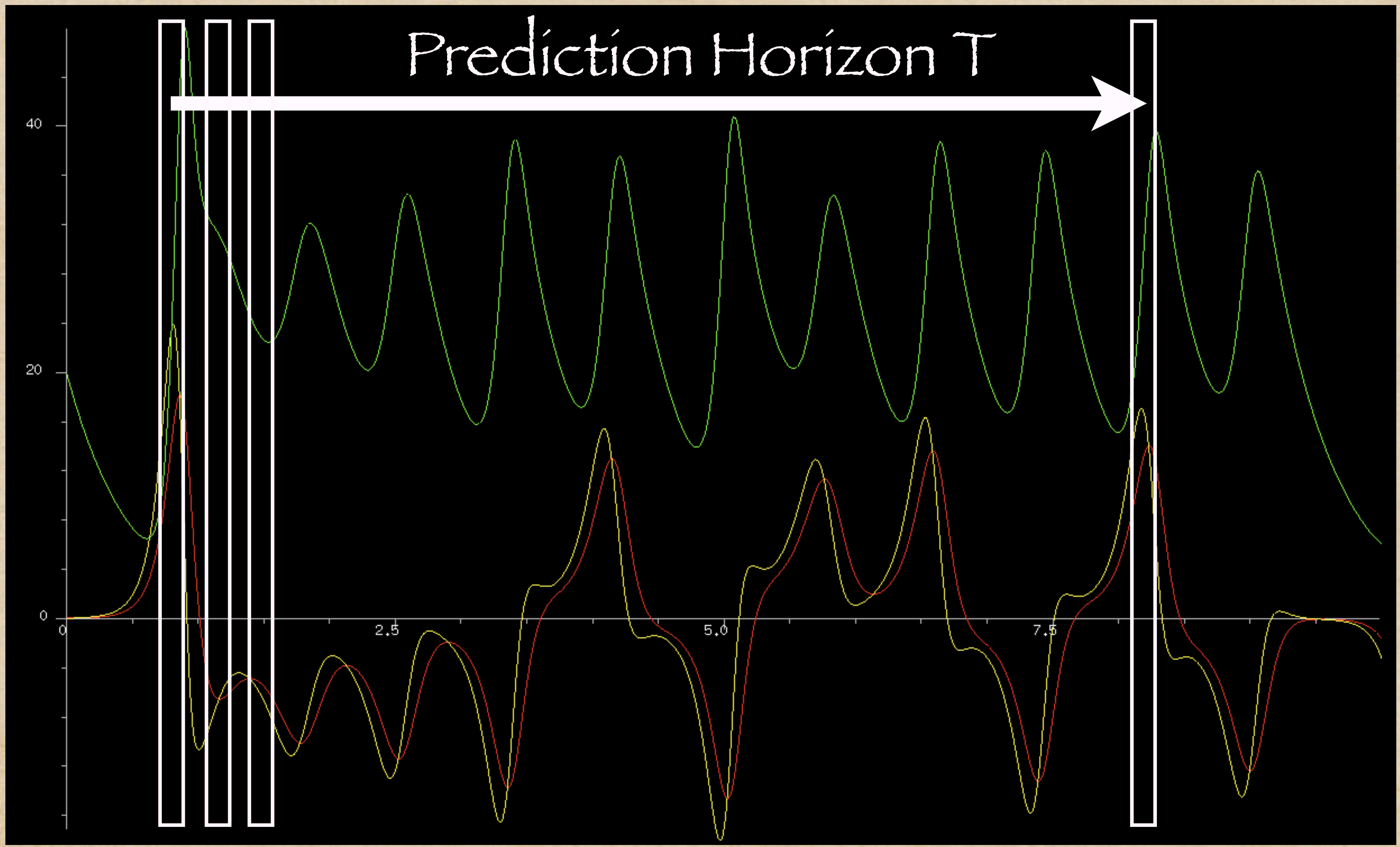
$$\text{Accuracy} \propto e^{-T}$$



Exponential Increase in Prediction Resources

$$|\text{Measurements}| \propto e^{T}$$

$$\text{Accuracy} \propto e^{-T}$$

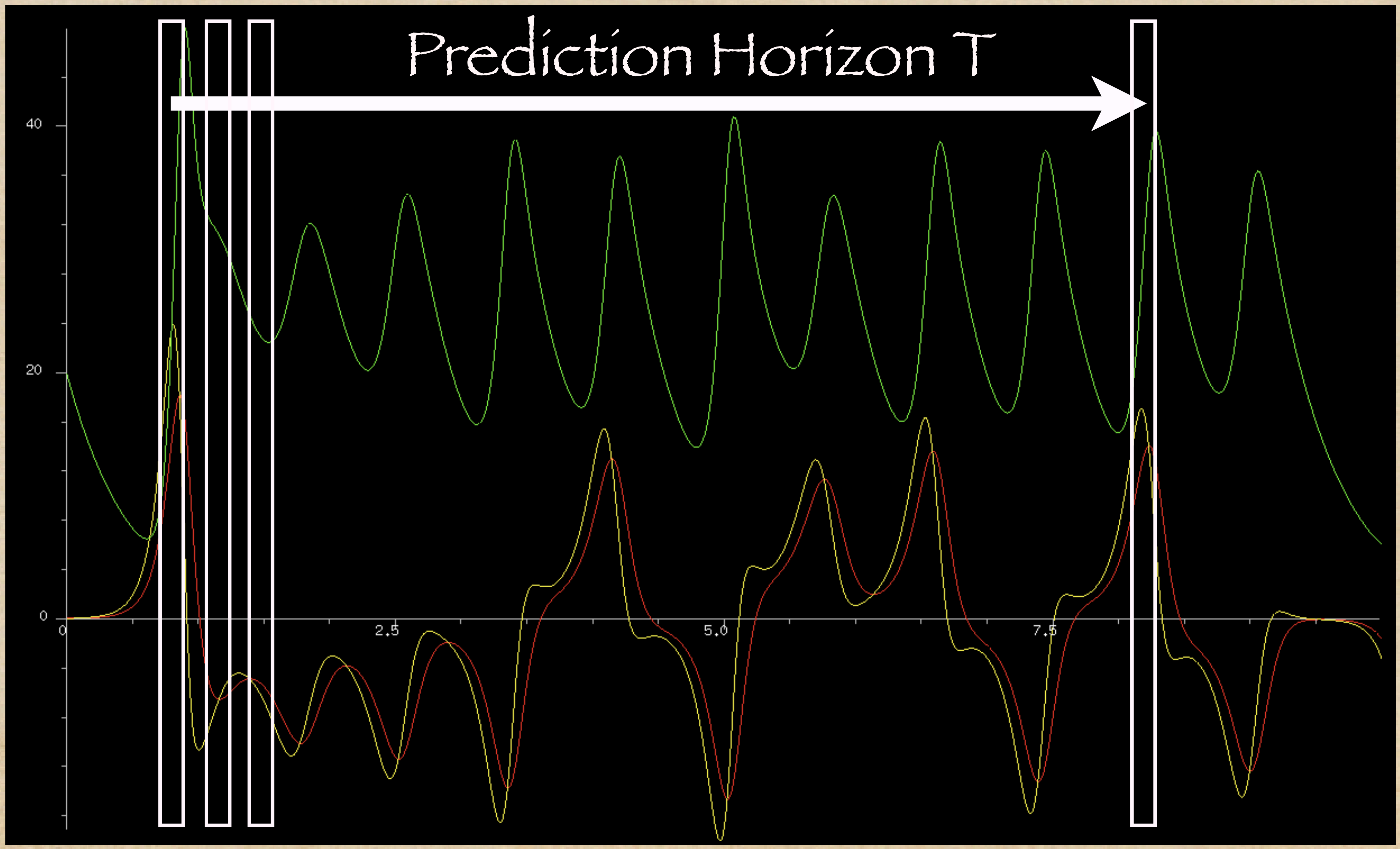


Exponential Increase in Prediction Resources

$$\text{Accuracy} \propto e^{-T}$$

$$|\text{Measurements}| \propto e^{T}$$

$$|\text{Compute time}| \propto e^{T}$$



Consequence

- No short cuts!
  - No closed-form solutions
  - Probabilistic description misses embedded order
  - No computational speed-ups
  - Must compute full trajectory

- Right representation is critical for reducing the prediction error as far as possible (but no further!)

Fundamental in Nonlinear Dynamics!

- Each nonlinear system requires its own representation
- Selecting balance between ascribing structure or noise to a measurement depends on representation
- Fundamental issue: (Theory of) Theory Building
- Subsidiary issue:

  Statistical fluctuations due to finite data sample

  (This is not about machine learning!)

Information Theory for Complex Systems

- Processes
- Information versus Entropy
- Communication Channels
- Processes as Channels
- First Pass: Measures of Complexity

Additional Reference

Thomas Cover & Joy Thomas

"Elements of Information Theory"

Second Edition (2006)

Stochastic Processes:

Chain of random variables:

$$\overleftrightarrow{S} \equiv \ldots S_{-2} S_{-1} S_0 S_1 S_2 \ldots$$

Random variable: S_t

Alphabet: $\mathcal{A}$

Realization:

$$\cdots s_{-2} s_{-1} s_0 s_1 s_2 \cdots \; ; \; s_t \in \mathcal{A}$$

Stochastic Processes:

Chain of random variables: $\overleftrightarrow{S} = \overleftarrow{S}_t \overrightarrow{S}_t$

Past: $\overleftarrow{S}_t = \ldots S_{t-2} S_{t-1} S_t$

Future: $\overrightarrow{S}_t = S_{t+1} S_{t+2} S_{t+3} \ldots$

L-Block: $S_t^L \equiv S_t S_{t+1} \ldots S_{t+L-1}$

Word: $s_t^L \equiv s_t s_{t+1} \ldots s_{t+L-1} \in \mathcal{A}^L$

Stochastic Processes ...

Process:

$$\Pr(\overleftrightarrow{S}) = \Pr(\ldots S_{-2} S_{-1} S_0 S_1 S_2 \ldots)$$

Sequence (or word) distributions:

$$\{\Pr(S_t^L) = \Pr(S_t S_{t+1} \ldots S_{t+L-1}) : S_t \in \mathcal{A}\}$$

Process:

$$\{\Pr(S_t^L) : \forall t, L\}$$

Consistency condition:

$$\Pr(S_t^{L-1}) = \sum_{S_{t+L-1}} \Pr(S_t^L)$$

Types of Stochastic Process:

Stationary process:

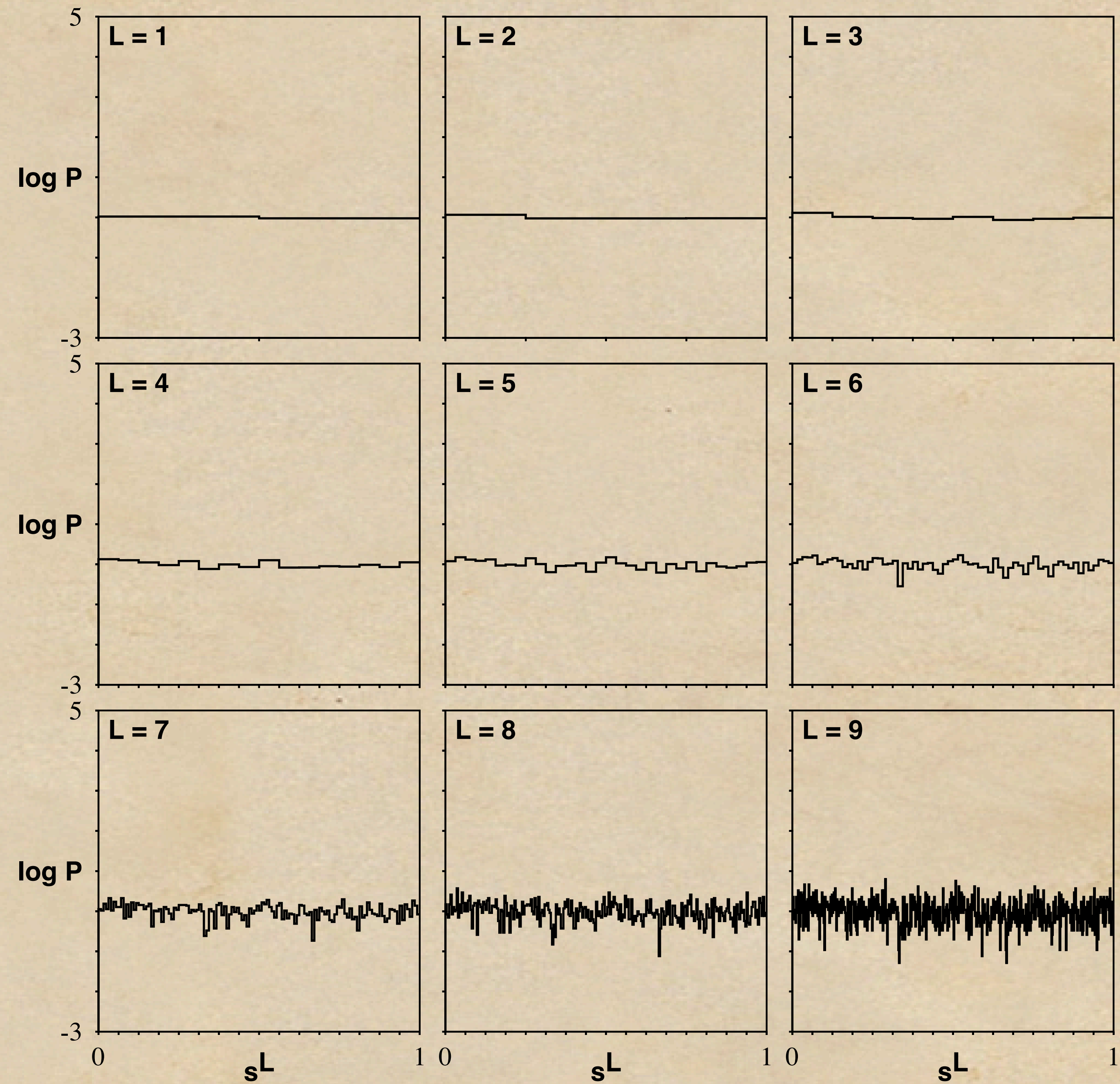
$$\Pr(S_t S_{t+1} \ldots S_{t+L-1}) = \Pr(S_0 S_1 \ldots S_{L-1})$$

Assume stationarity, unless otherwise noted.

Notationally: Drop time indices.

Example:
Fair Coin ...

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



Word as binary fraction:

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

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

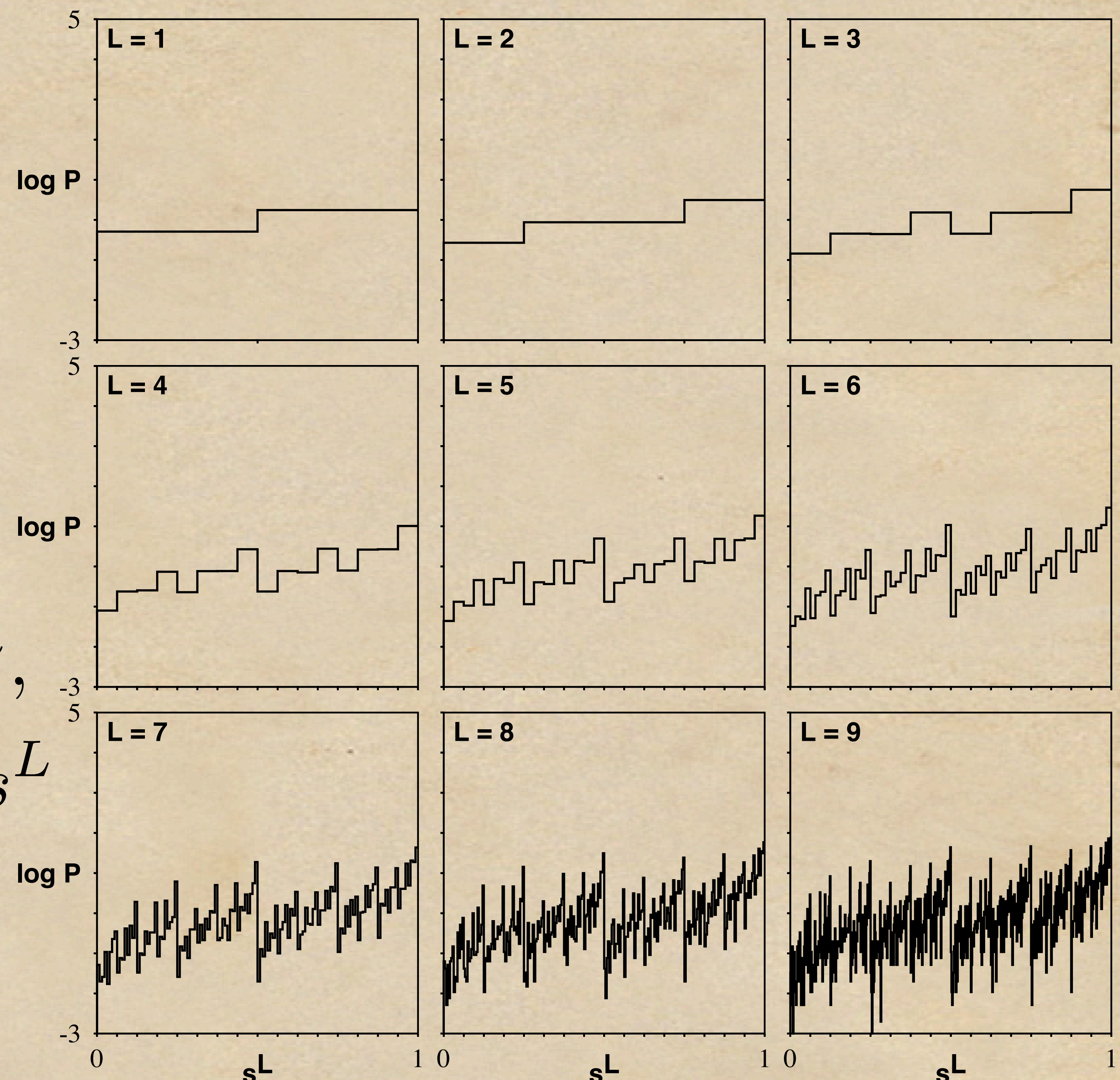
$$s^L \in [0, 1]$$

Example:
Biased Coin ...

Sequence Distribution:

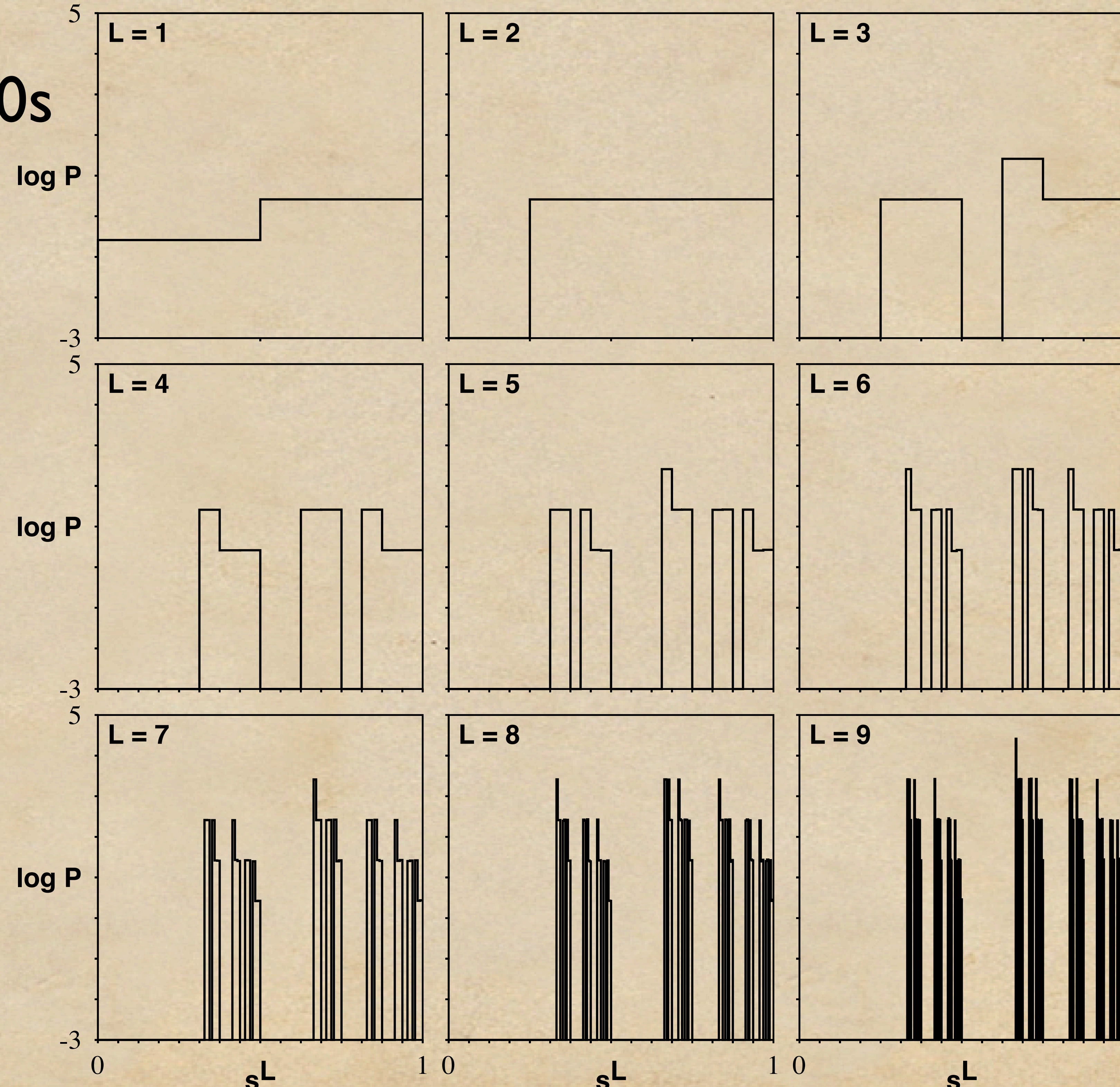
$$\Pr(s^L) = p^n (1-p)^{L-n},$$

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



Example: Golden Mean Process

No Consecutive 0s



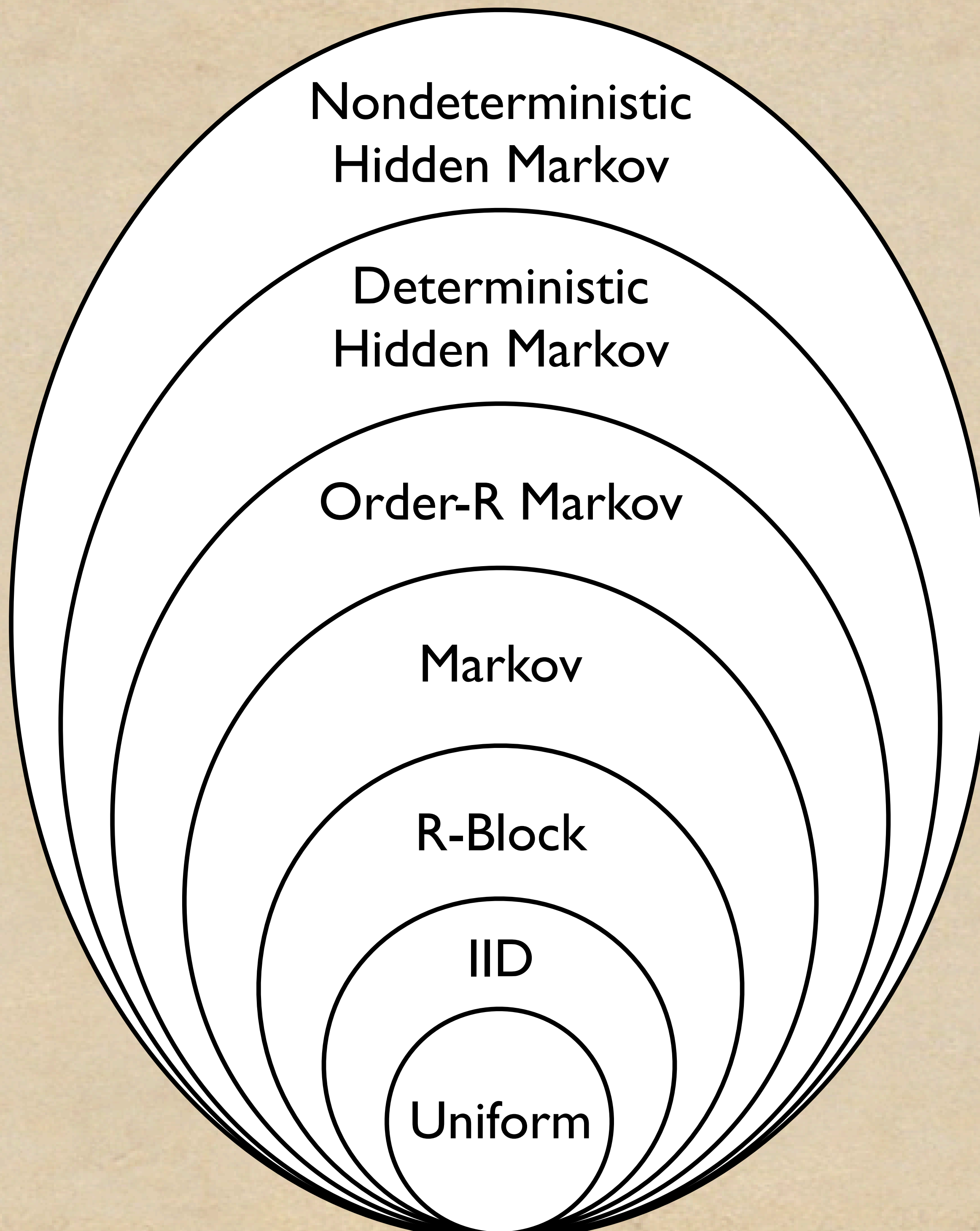
Models of Stochastic Processes ...

Two Lessons:

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

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

Classification of Discrete Stochastic Processes:



Sources of Information:

Apparent randomness:

Uncontrolled initial conditions

Actively generated: Deterministic chaos

Hidden regularity:

Ignorance of forces

Limited capacity to model structure

Issues:

What is information?

How do we measure unpredictability

How do we quantify structure?

Information $\neq$ Energy

History:

Boltzmann (19th Century):

Equilibrium in large-scale systems

Hartley-Shannon-Wiener (Early 20th):

Communication & Cryptography

Current threads (late 20th century):

Coding, Statistics, Dynamics, and Learning

Information as uncertainty and surprise:

Observe something unexpected: gain information

Bateson: "A difference that makes a difference"

Information as uncertainty and surprise ...

How to formalize?

Shannon's approach: Connection with Boltzmann's Entropy

A measure of surprise.

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

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

Completely unpredictable: Maximally surprised

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

How to measure?

Information = f(Pr(event))?

Random variables: X, Y; events $x, y \in \{1, 2, \ldots, k\}$

Distribution: $\Pr(X) = (p_1, \ldots, p_k)$

Shorthand: $X \sim p(x) \quad Y \sim p(y)$

Khinchin axioms for a measure of information:

Entropy: $H(X) = H(p_1, \ldots, p_k)$

(1) Maximum at equidistribution:

$$H(p_1, \ldots, p_k) \leq H\left(\tfrac{1}{k}, \ldots, \tfrac{1}{k}\right)$$

(2) Continuous function of distribution:

$$H(p_1, \ldots, p_k) \text{ versus } p_i$$

(3) Expansibility:

$$H(p_1, \ldots, p_k) = H(p_1, \ldots, p_k, p_{k+1} = 0)$$

(4) Additivity of independent systems:

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

Khinchin axioms for a measure of information ...

Theorem:

Get unique (up to a factor) functional form,

The Shannon entropy:

$$H(X) \propto -\sum_{i=1}^{k} p_i \log p_i$$

Shannon Entropy: $X \sim P$ $\quad x \in \mathcal{X} = \{1, 2, \ldots, k\}$

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

Note: $0 \log 0 = 0$

$$H(X) = \langle - \log_2 p(x) \rangle$$

Units:

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

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

Properties:

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

Examples: Binary random variable X

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

$H(X)$?

Binary entropy function:

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

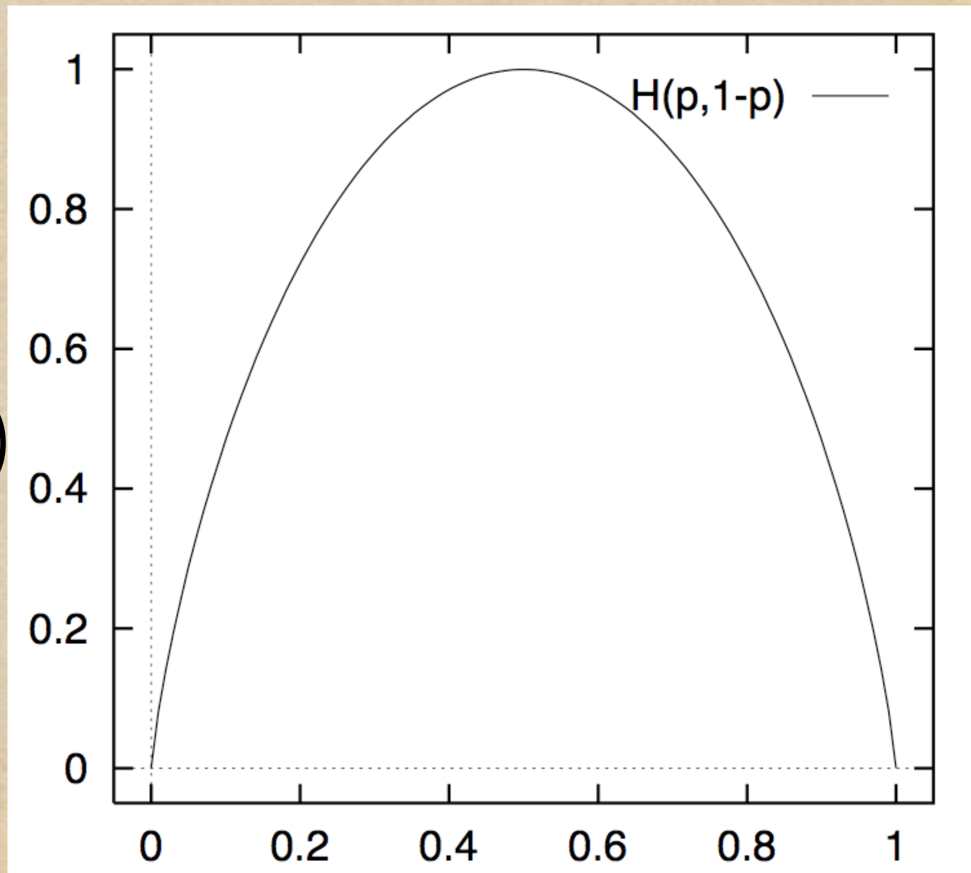
Fair coin: $p = \frac{1}{2}$

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

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

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

Note: $0 \cdot \log 0 = 0$



Example: 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)

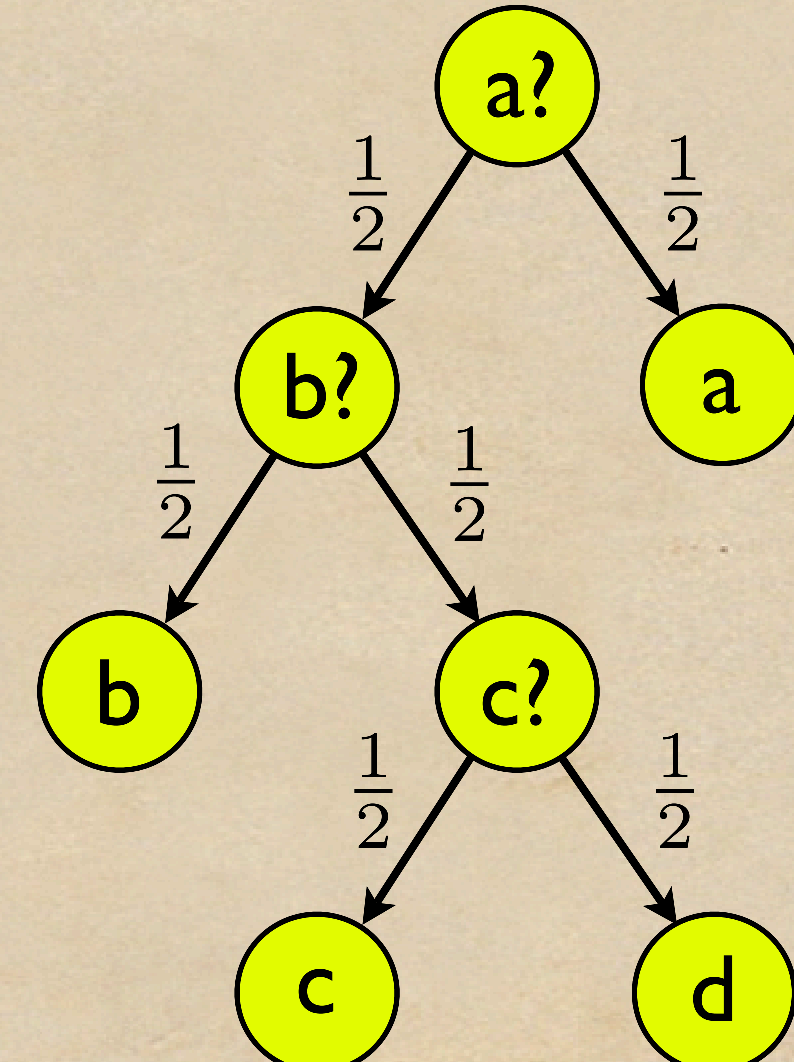
x = c? (only get this far a quarter of the time)

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

Theorem: Optimal way to ask questions.

Example: IID Process over four events ...

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



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

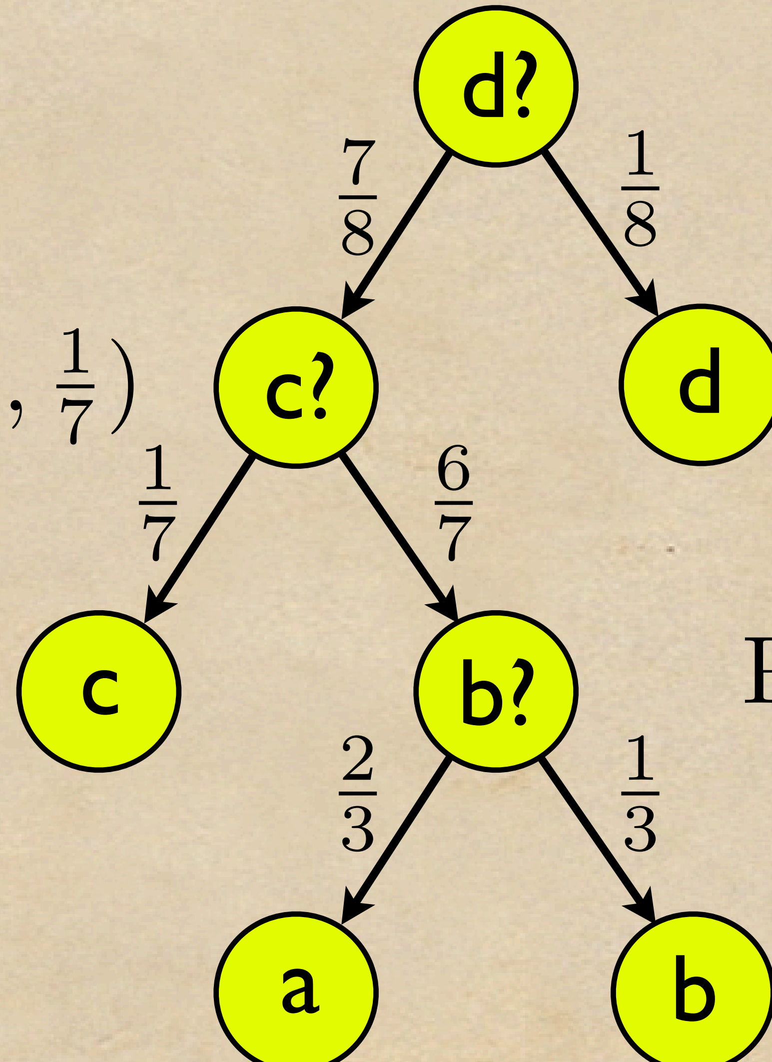
Example: IID Process over four events ...

Query in a different order:

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

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

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



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

Example: IID Process over four events

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

At each stage, ask question that is 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.

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

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: $H(X|Y) \neq H(Y|X)$

Relative Entropy of Two Distributions:

$$X \sim P \;\&\; Y \sim Q, \text{ over common } x \in \mathcal{X}$$

Relative Entropy:

$$\mathcal{D}(X||Y) = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)}$$

Note: $0 \log \frac{0}{q} = 0$

$$p \log \frac{p}{0} = \infty$$

Typically applied to: $Q : \; q(x) > 0 \;, \; \forall \; x \in \mathcal{X}$

Alternate use (notation):

$$\mathcal{D}(P||Q)$$

Relative Entropy of Two Distributions ...

Properties:

$$
\begin{aligned}
&(1)\ \mathcal{D}(X||Y) &\geq\ & 0 \\
&(2)\ \mathcal{D}(X||Y) &=\ & 0 \ \Leftrightarrow\ P = Q \\
&(3)\ \mathcal{D}(X||Y) &\neq\ & \mathcal{D}(Y||X)
\end{aligned}
$$

Also called:

Kullback-Leibler Distance

Information Gain: Number of bits of describing X as Y

Discrimination between X & Y

Not a distance: not symmetric, no triangle inequality

Common Information Between Two Random Variables:

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

Mutual Information:

$$I(X;Y) = \mathcal{D}(P(x,y)||P(x)P(y))$$

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

Properties:

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

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

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

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

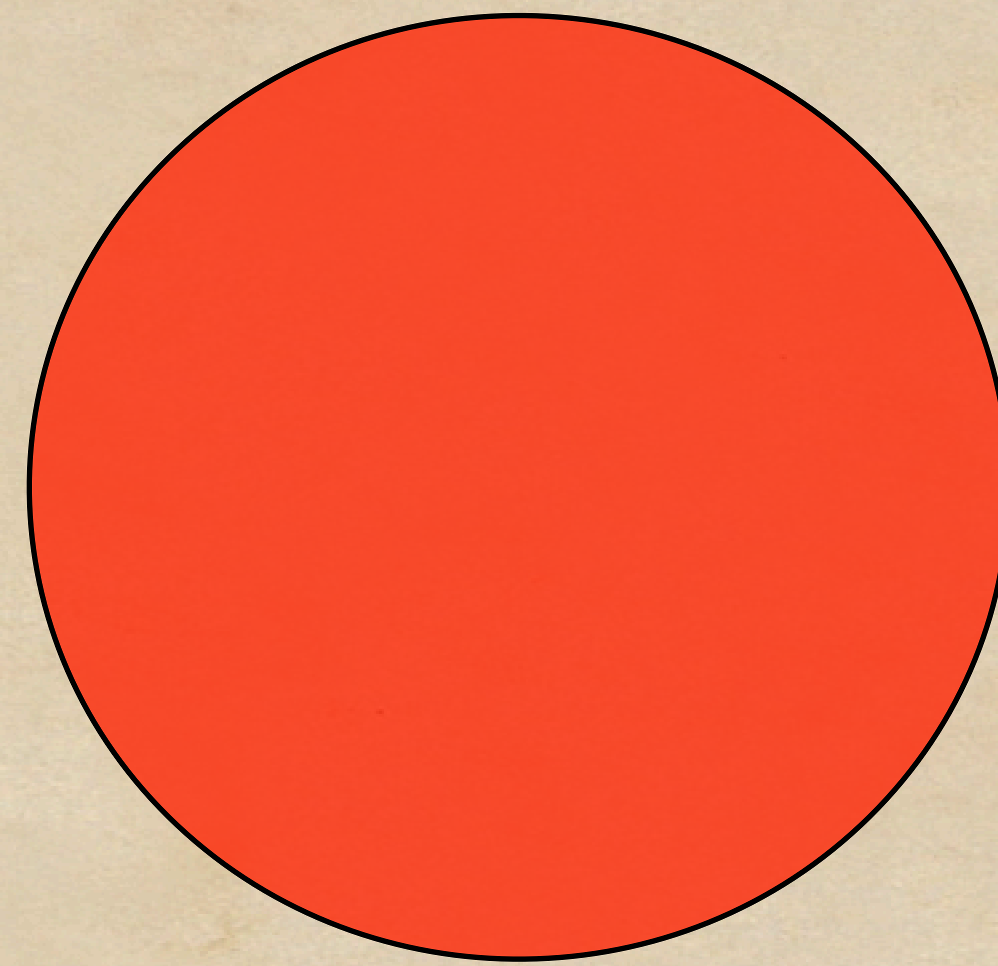
(5) $I(X;X) = H(X)$

Interpretation:

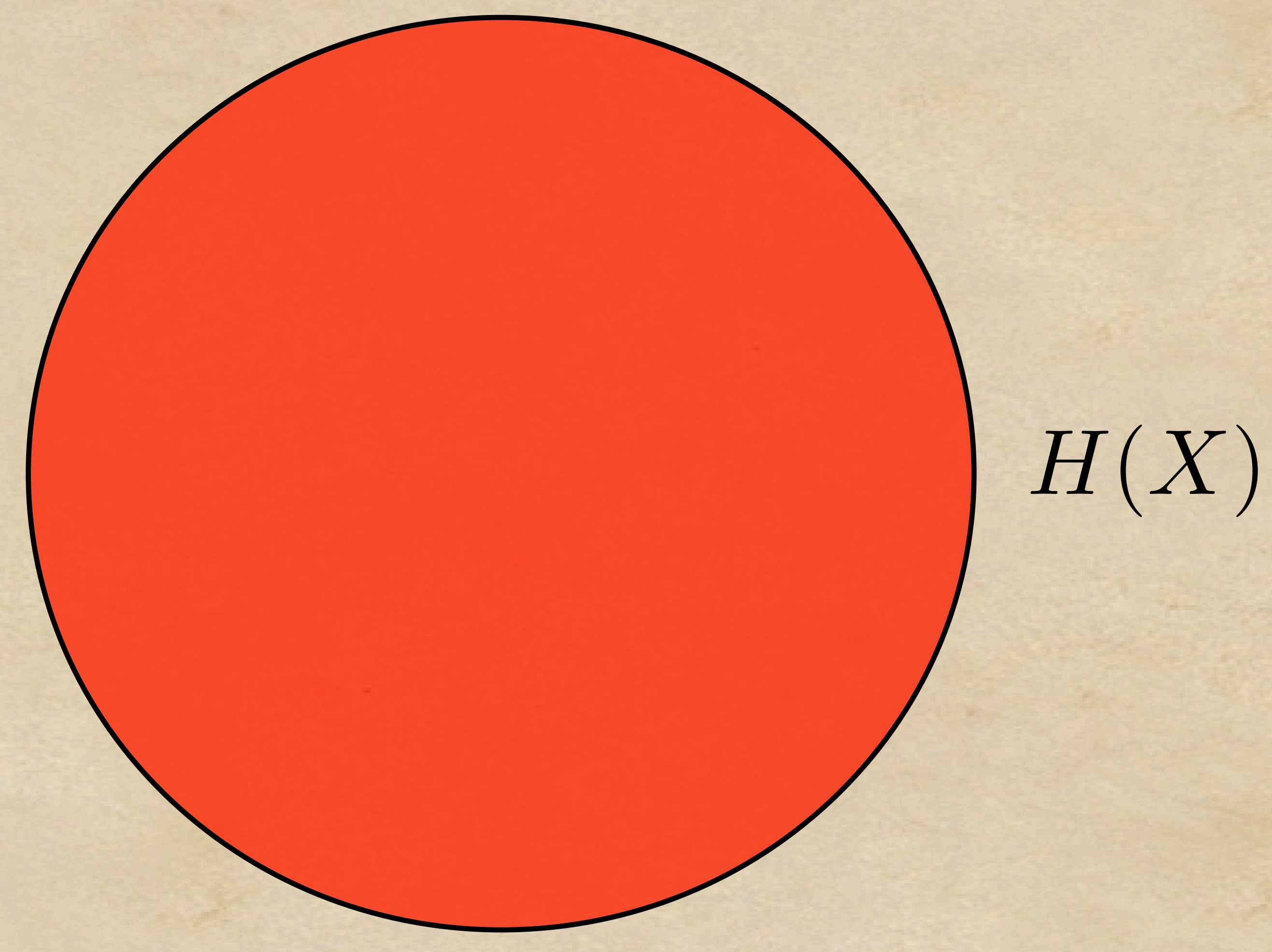
- Information one variable has about another
- Information shared between two variables
- Measure of dependence between two variables

Event Space Relationships of Information Quantifiers:

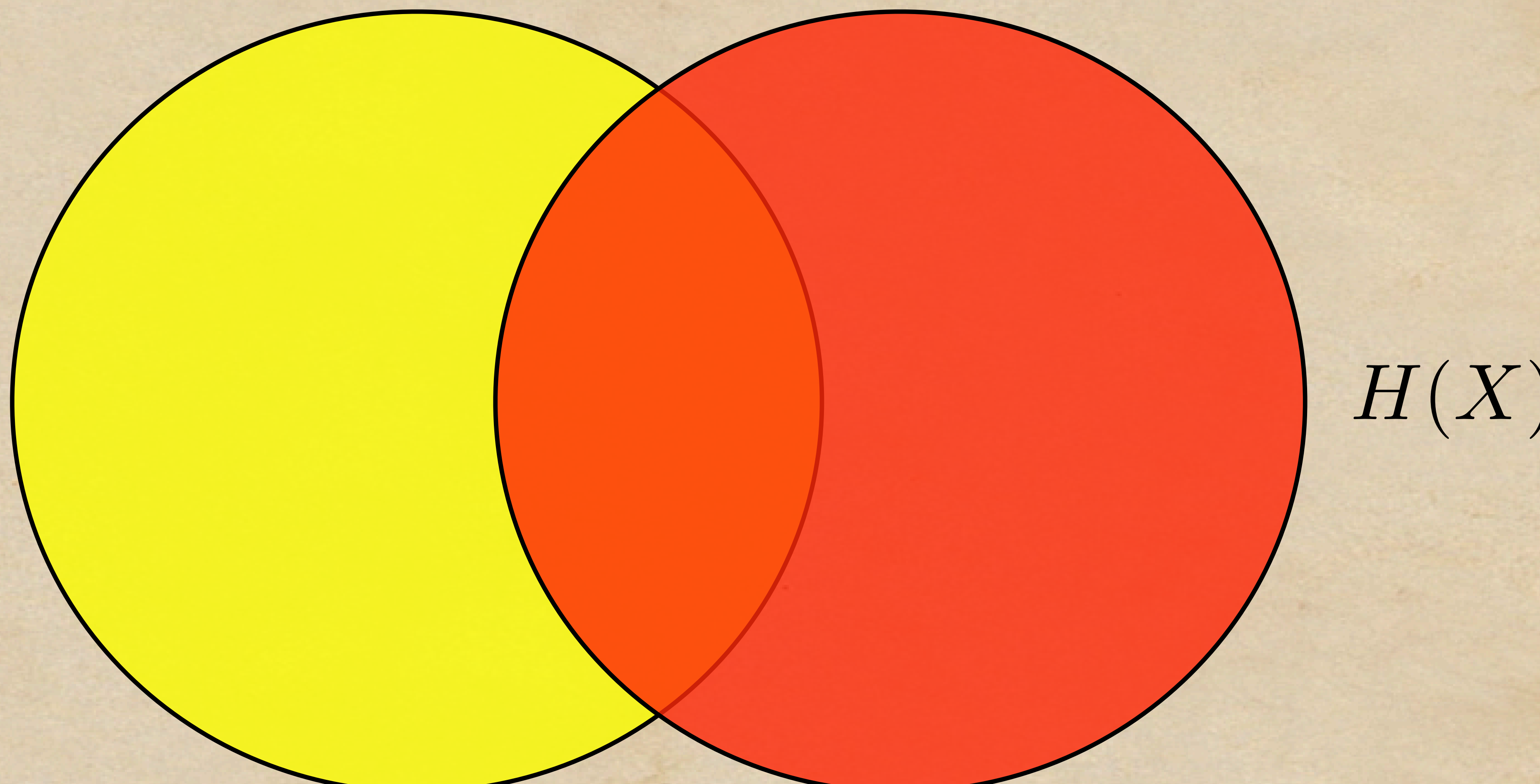
Event Space Relationships of Information Quantifiers:



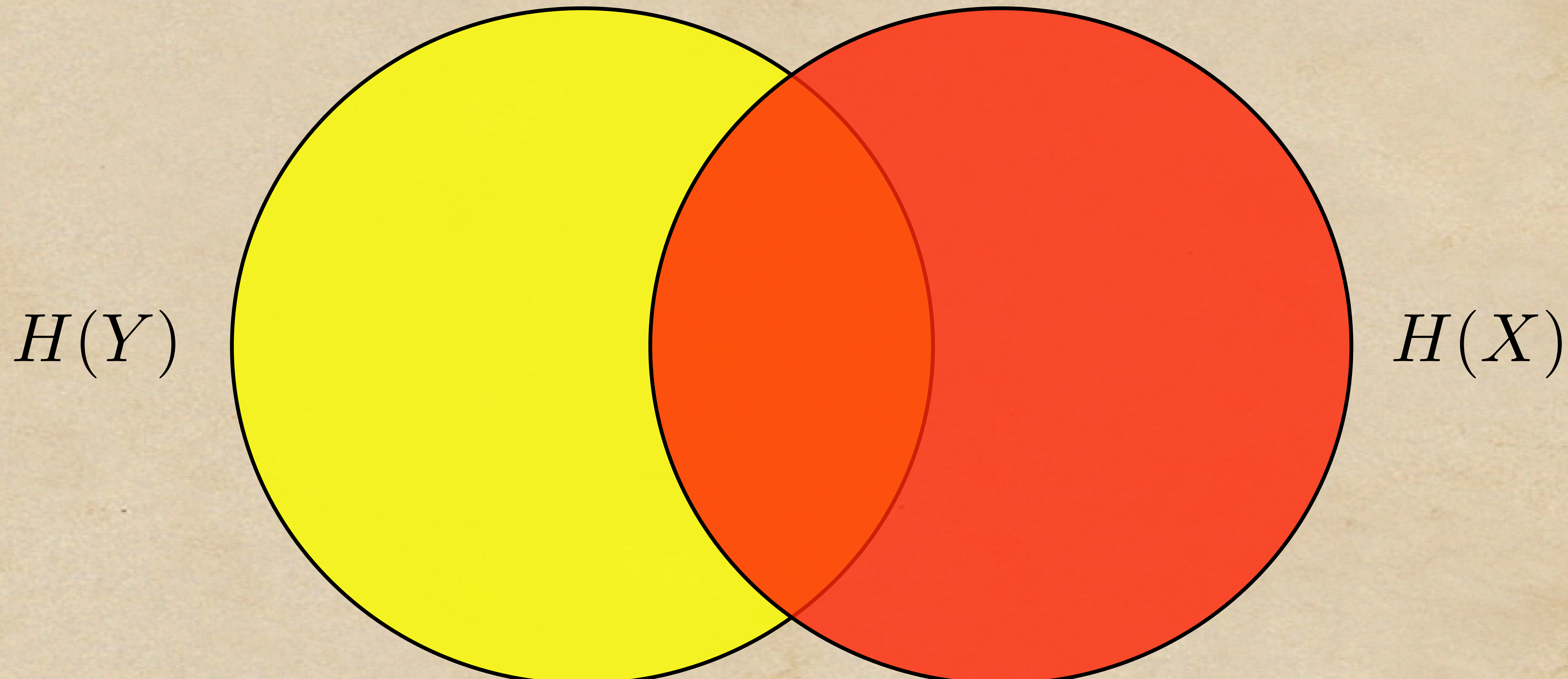
Event Space Relationships of Information Quantifiers:



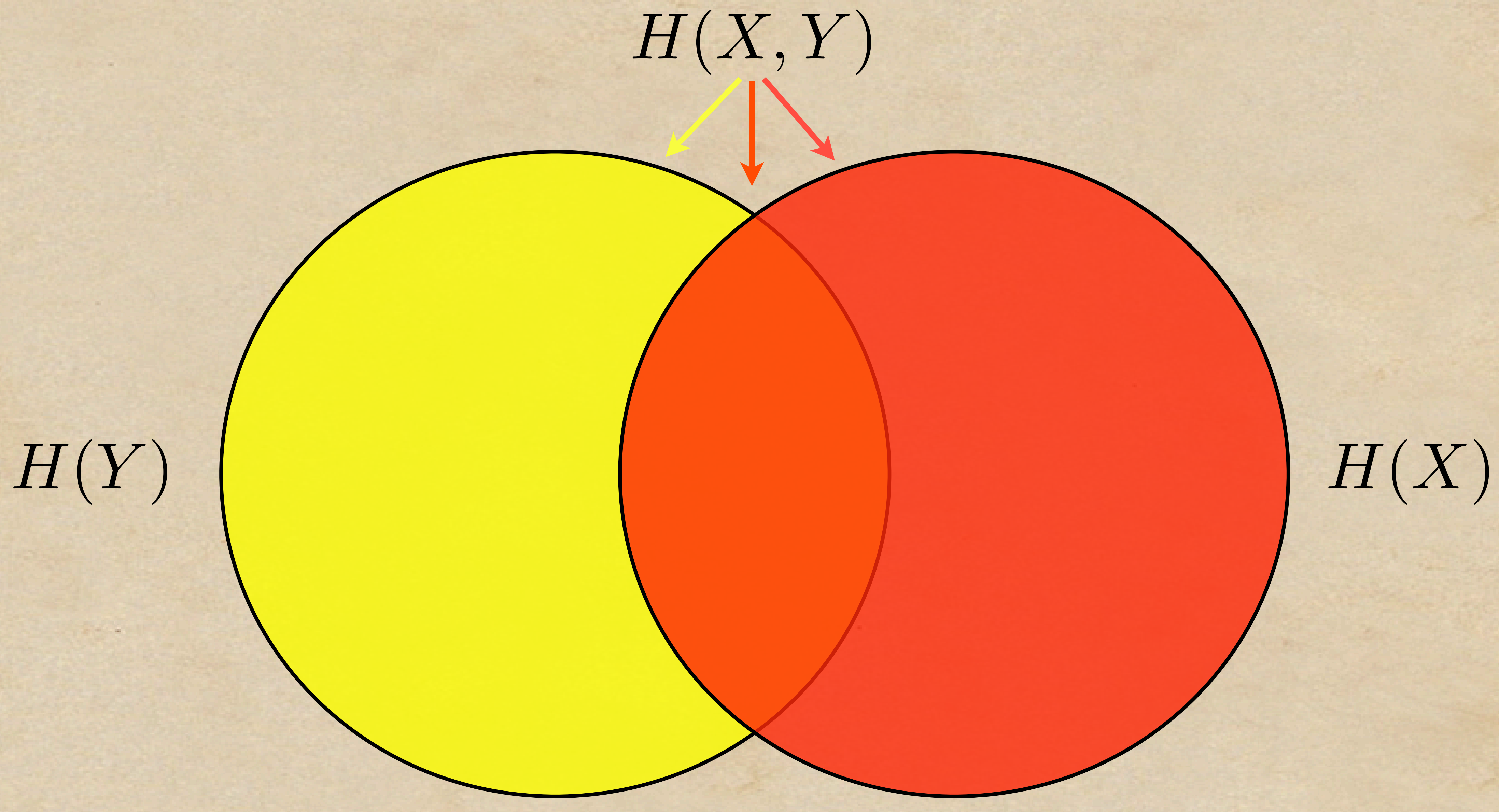
Event Space Relationships of Information Quantifiers:



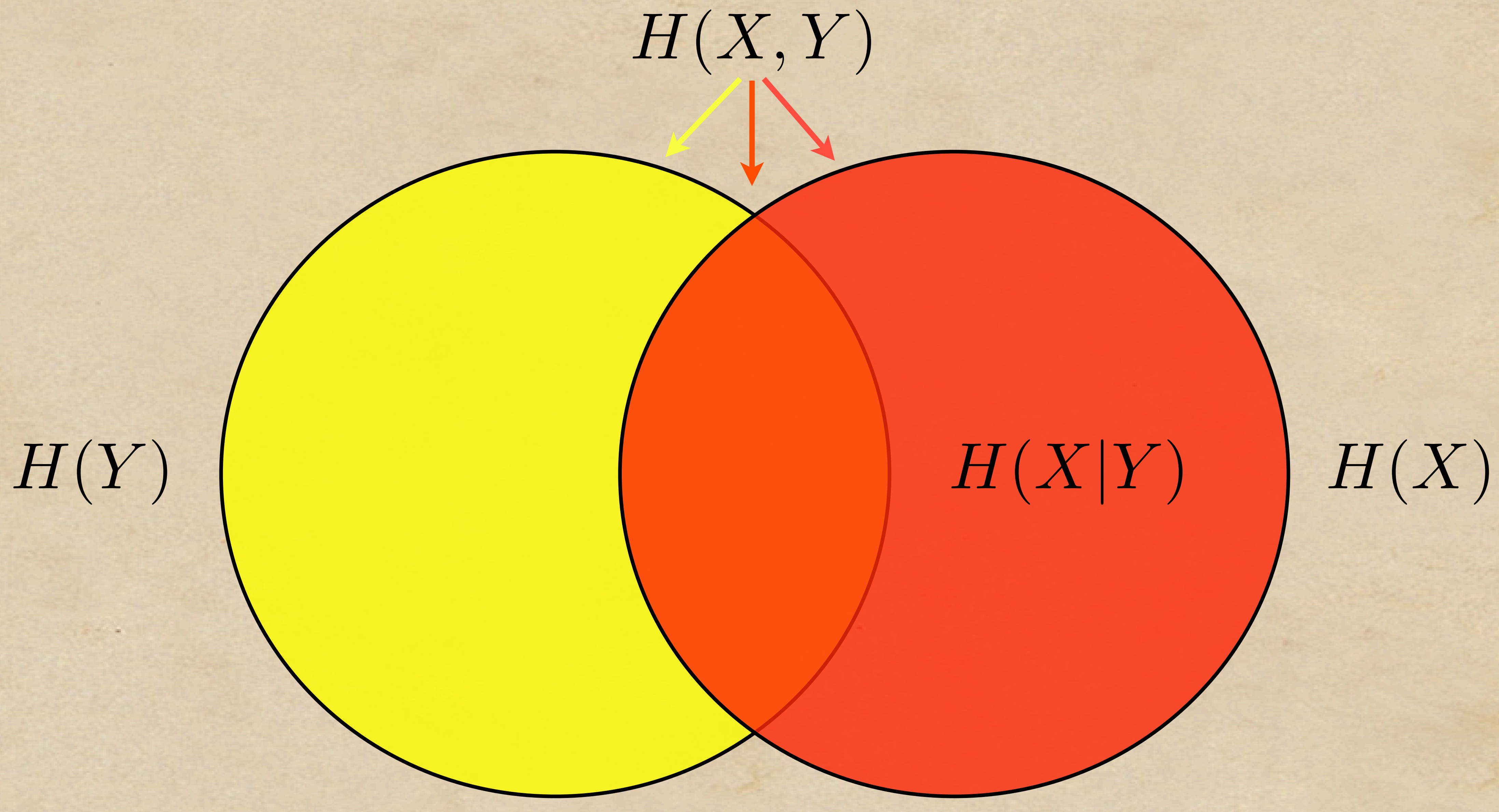
Event Space Relationships of Information Quantifiers:



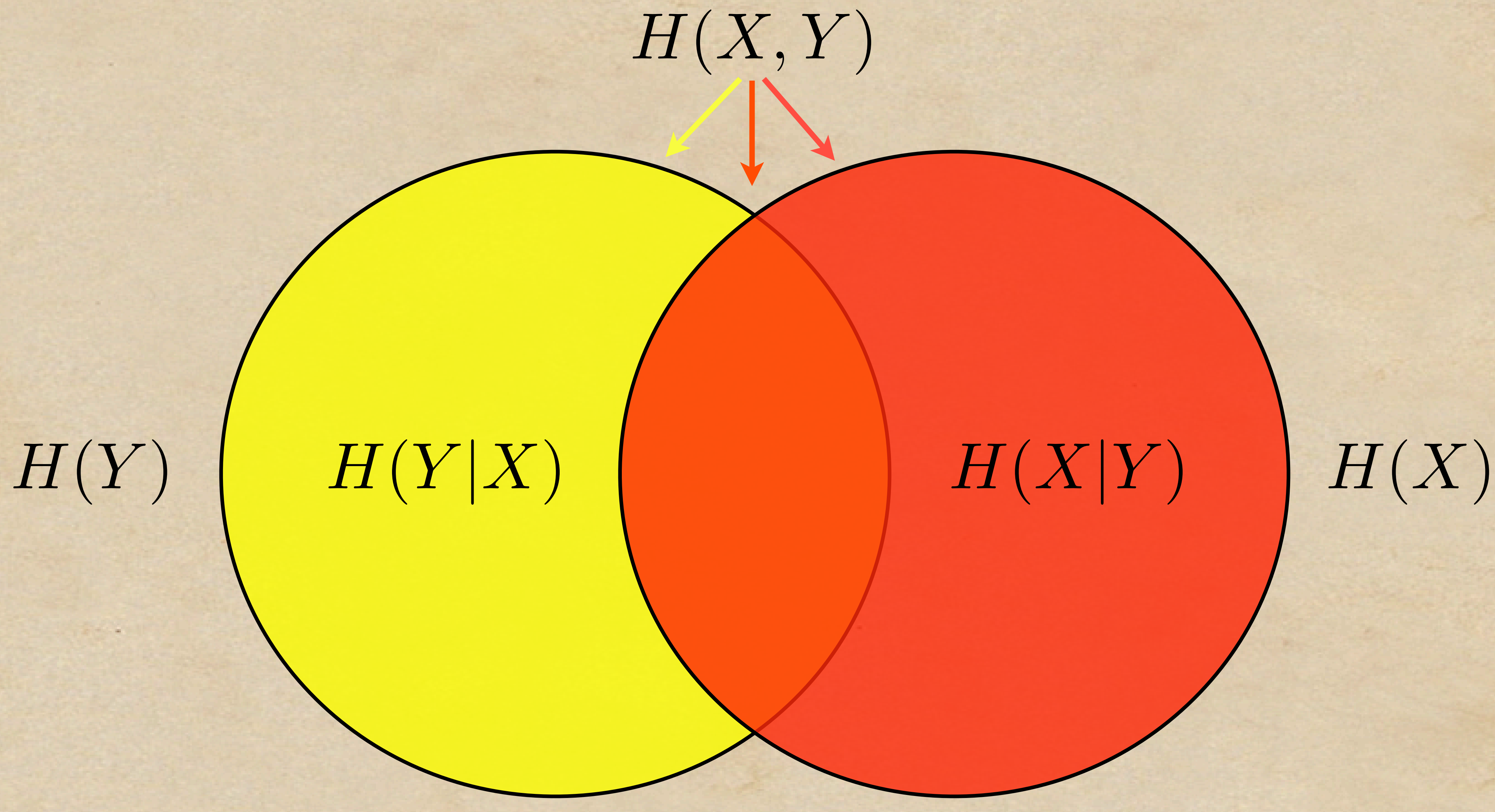
Event Space Relationships of Information Quantifiers:



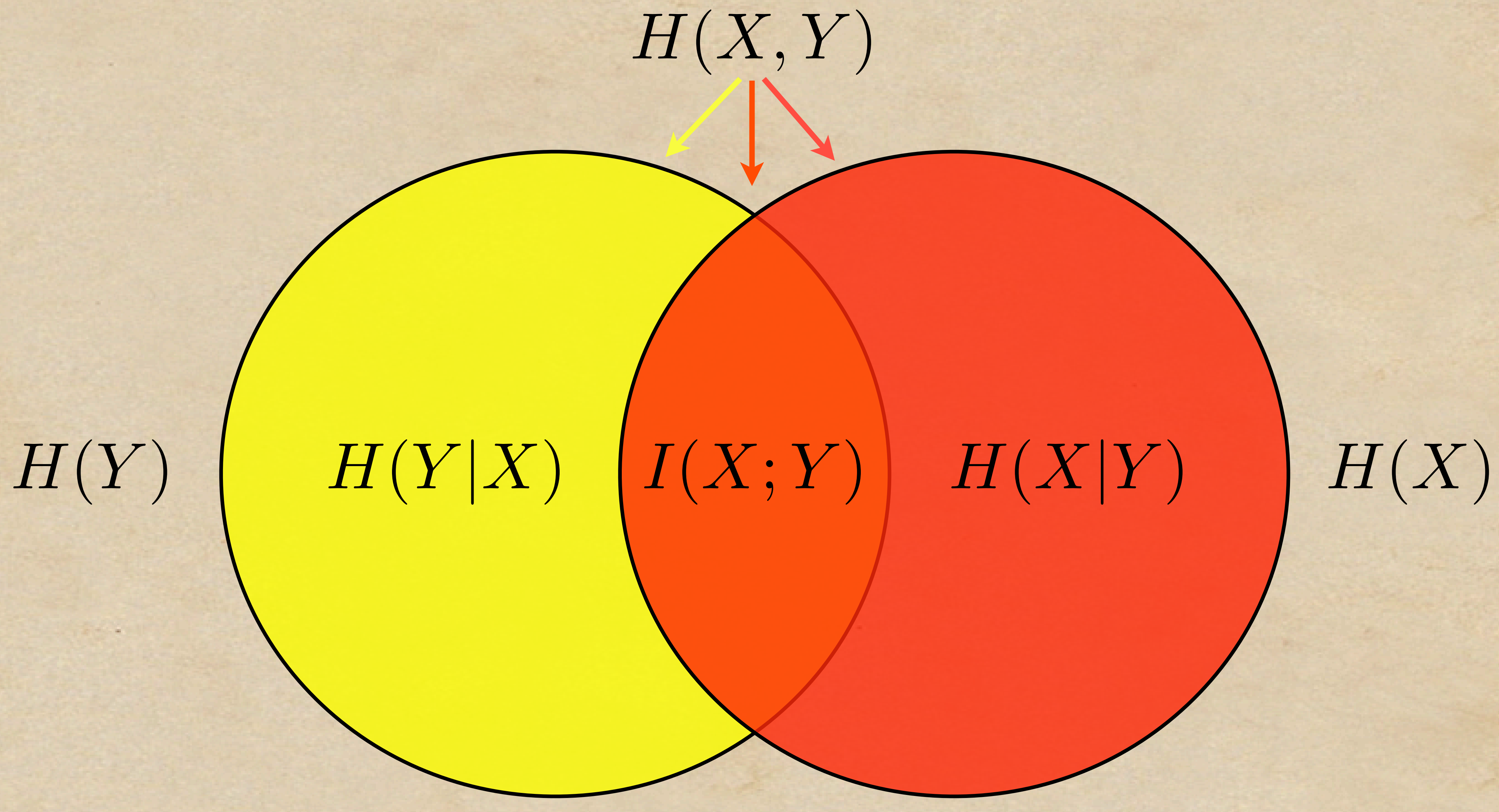
Event Space Relationships of Information Quantifiers:



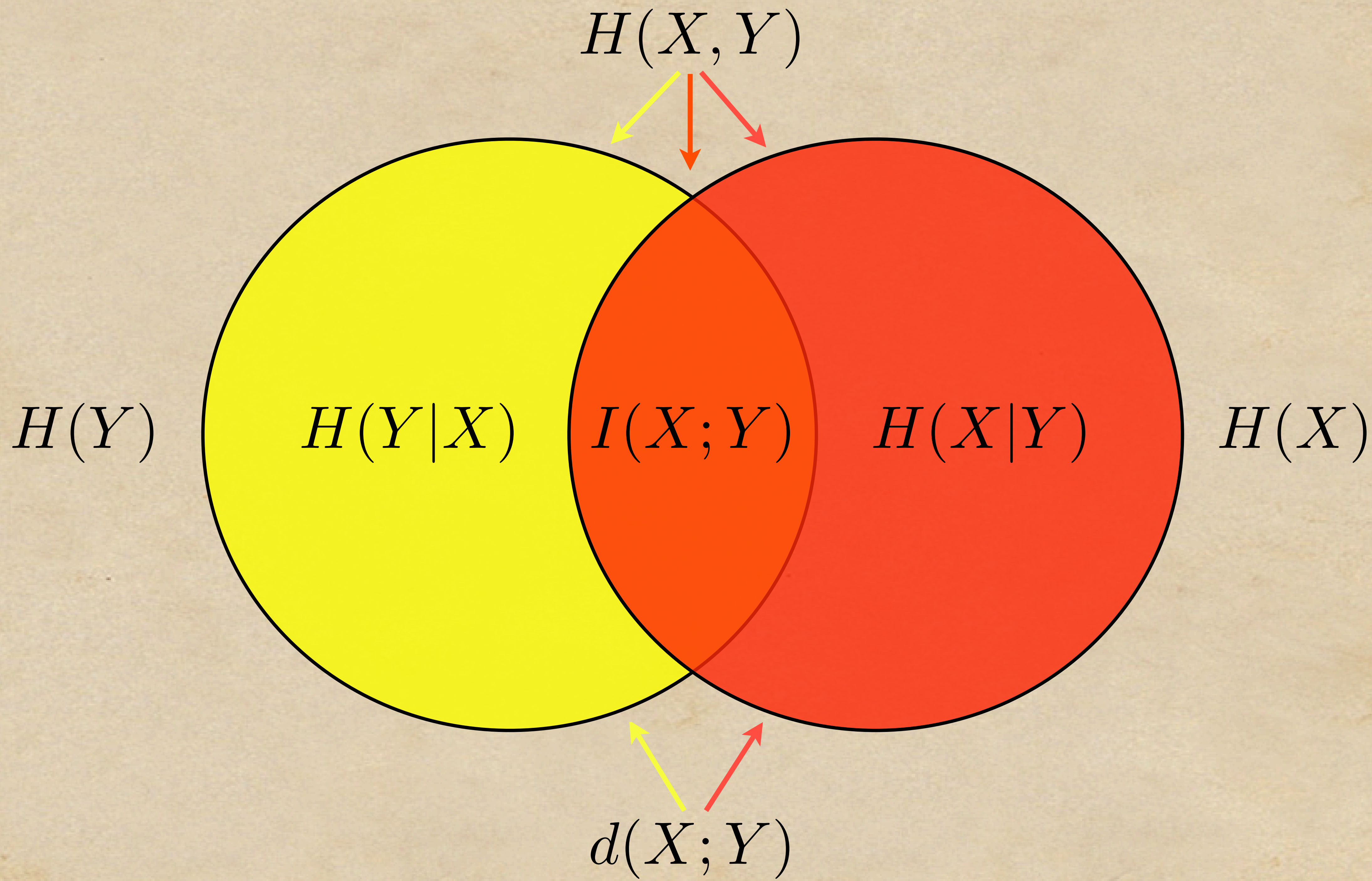
Event Space Relationships of Information Quantifiers:



Event Space Relationships of Information Quantifiers:



Event Space Relationships of Information Quantifiers:



Central Results of Information Theory

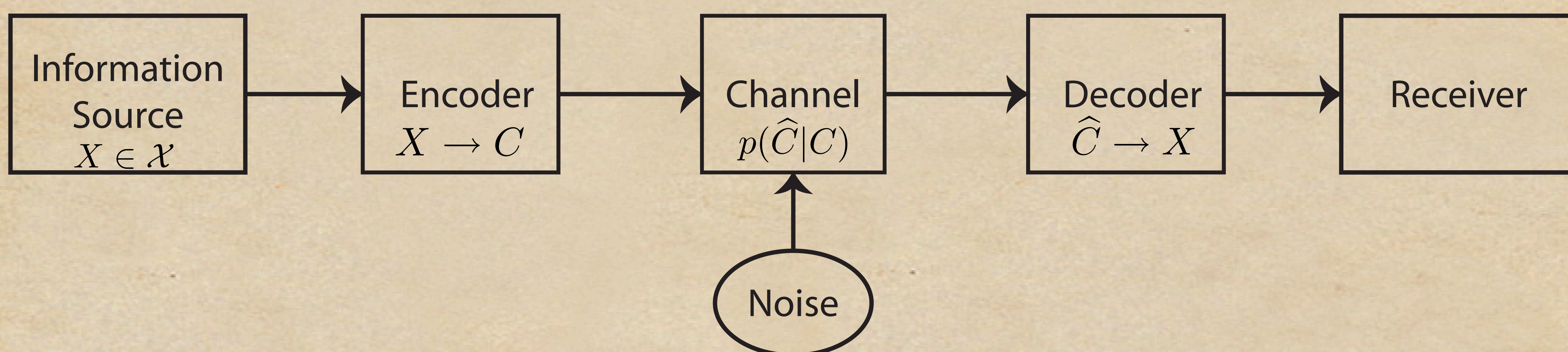
How to compress a process:
Can't do better than H(X)
(Shannon's First Theorem)

How to communicate a process's data:
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.

Communication channel

Messages	Codewords	Corrupted Codewords	Inferred Messages
$\ldots x_3 x_2 x_1$	$\ldots C(x_3)C(x_2)C(x_1)$	$\ldots \widehat{C}(x_3)\widehat{C}(x_2)\widehat{C}(x_1)$	$\ldots x_3 x_2 x_1$



Coding for Communication Channels

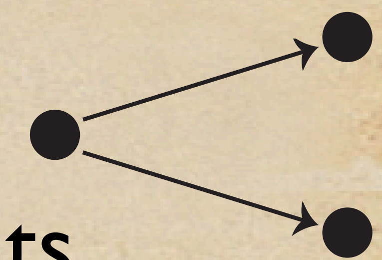
Kinds of channel:

- Phone line, ftp transfer, monologue, ...
- Dynamical system at time t and t+1
- Spin system at one site and another
- Measurement channel

Channel coding problem is to overcome errors:

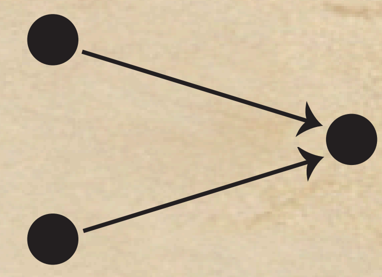
Equivocation:

Same input sequence leads to different outputs



Ambiguity:

Two different inputs lead to same output



Strategy:

Find channel inputs that are *least ambiguous* given distortion properties.

Codebook: Map information source onto those inputs.

Source $\sim P(x)$, but you use $Q(x)$.

Use incorrect probability model for code construction: $Q \neq P$

$$\langle l(x) \rangle = H(X) + \mathcal{D}(P||Q)$$

Data Compression Theorem (Shannon's First Theorem):

$$R(C) \geq H(X)$$

Cannot compress source (at code rate $R(C)$) below its entropy rate.

Operational meaning of entropy: A fundamental limit.

Discrete channel:

Input: $X \sim p(x)$

Output: $Y \sim p(y)$

Channel: $p(y|x)$

Memoryless channel:

$$p(y_t|x_t x_{t-1} \cdots) = p(y_t|x_t)$$

Channel Capacity:

$$\mathcal{C} = \max_{p(x)} \; I(X;Y)$$

Highest rate one can transmit over channel.

Channel Capacity:

$$\mathcal{C} = \max_{p(x)} \; I(X;Y)$$

Extremes of no communication:

No info to send: $H(X) = 0$

$$I(X;Y) = H(X) - H(X|Y) = 0 - 0 = 0$$

Complete distortion:

Output independent of input: $X \perp Y$

$$I(X;Y) = 0$$

Duality:

Compression removes redundancy for smallest description.

Transmission adds redundancy: compensate channel errors.

Channel Coding Theorem (Shannon's Second Theorem):

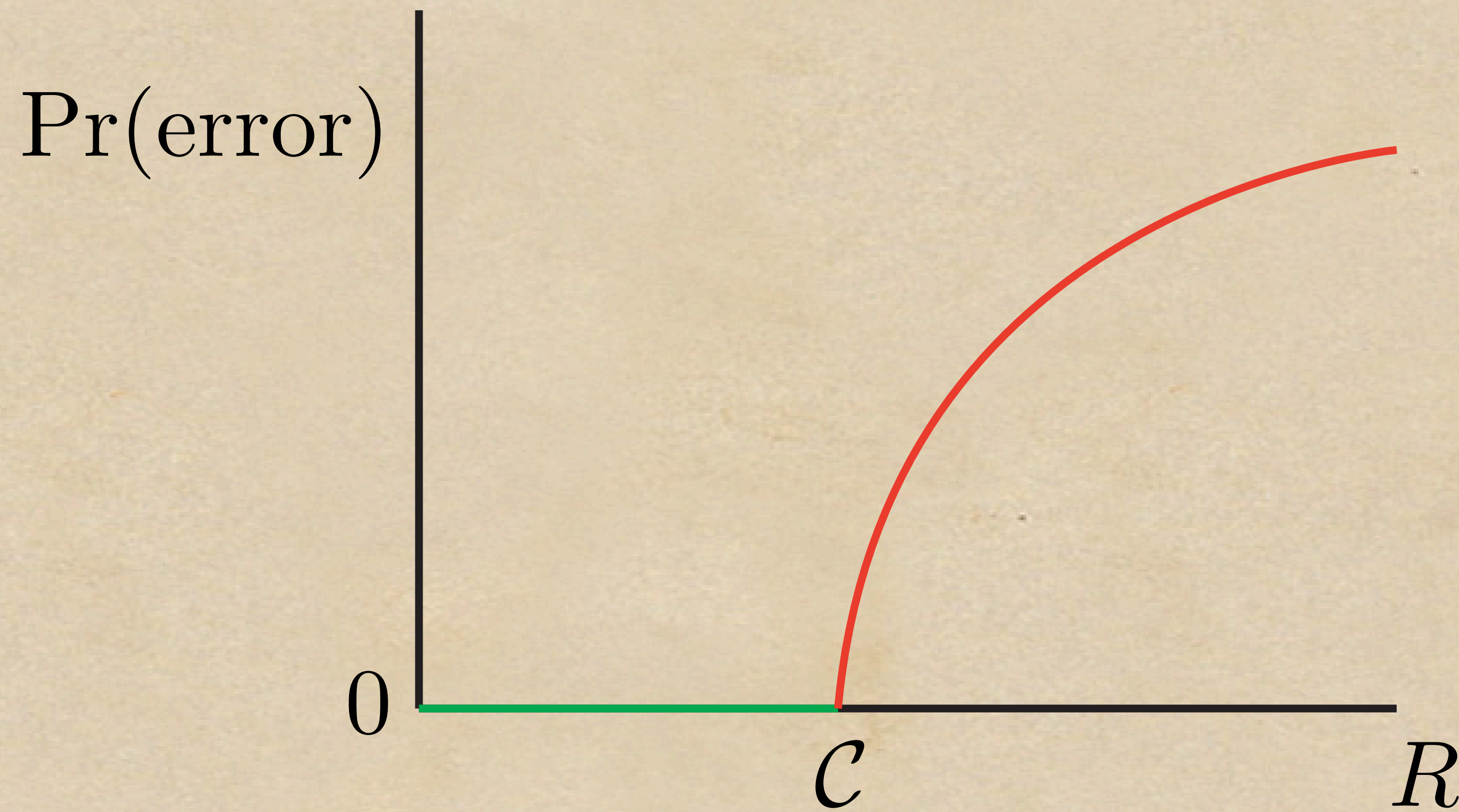
(1) Capacity is the maximum reliable transmission rate.
(2) Error-free codes exist if $R(C) < \mathcal{C}$.

Idea:

Use long block lengths.
In effect, have noisy channel with non-overlapping outputs.
Choose codewords (channel inputs) that
produce non-overlapping output sequences.

Channel Coding Theorem ...

What happens when transmitting above capacity, $R > \mathcal{C}$?



(Typical of complex systems?)

Information Theory for General 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:

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

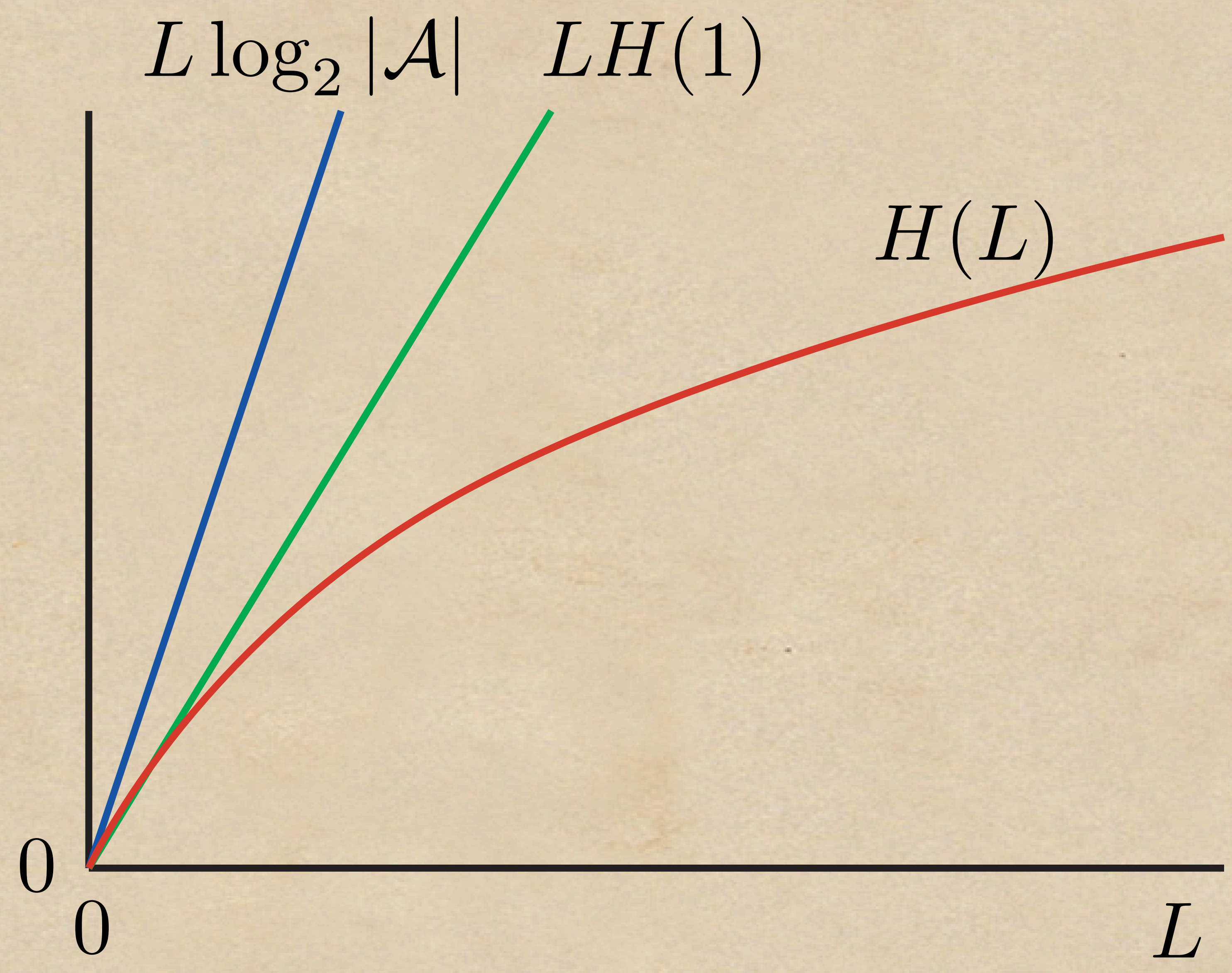
No measurements, no information: $H(0) = 0$

Bounds:

(1) Crude: $H(L) \leq L \log_2 |\mathcal{A}|$

(2) 1-block Markov: $H(L) \leq L H(1)$

Block Entropy ...

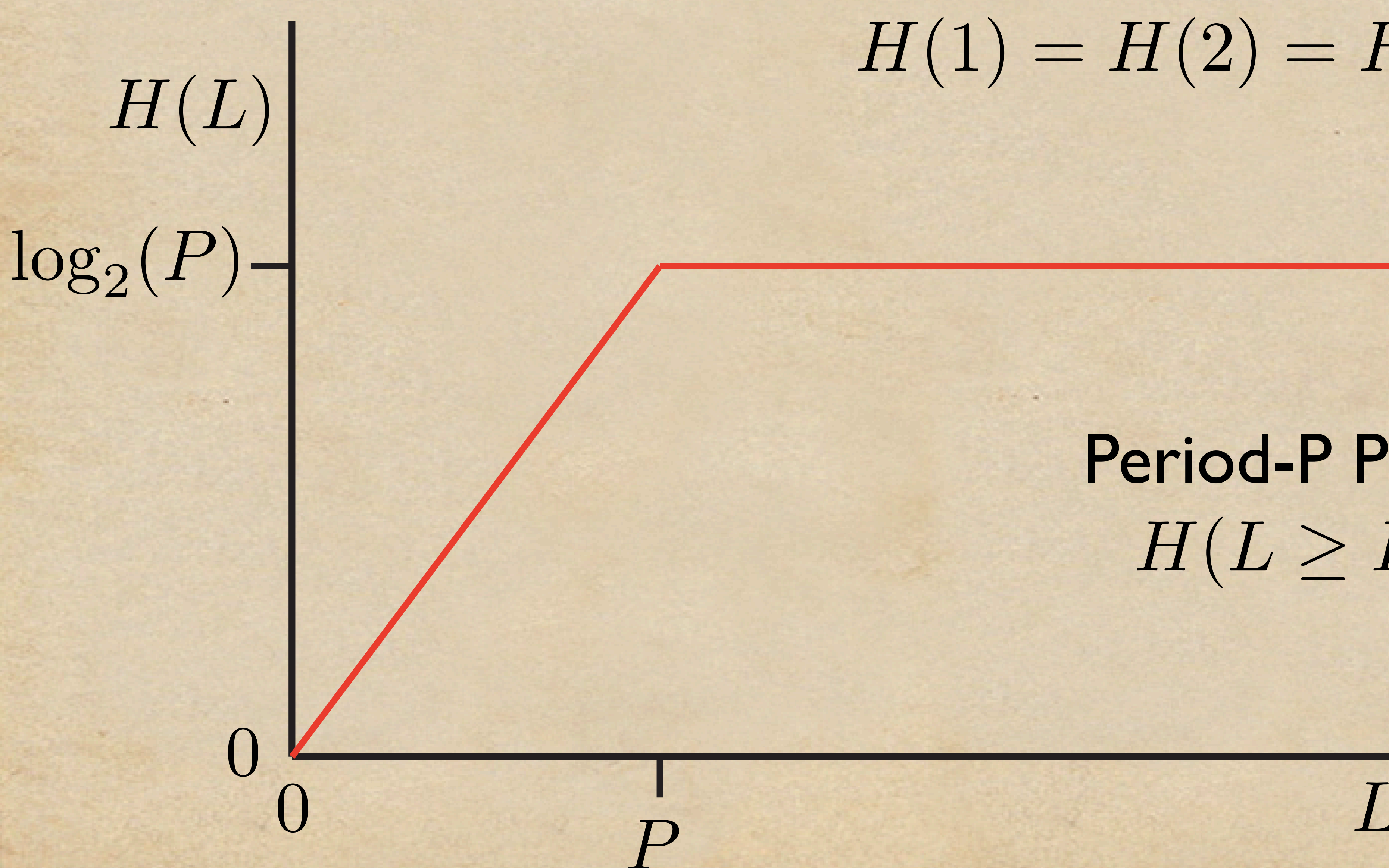


Block Entropy ...

Example: Period-2 Process

$$\Pr(0) = \Pr(1) = \frac{1}{2}$$
$$\Pr(01) = \Pr(10) = \frac{1}{2}$$
$$\Pr(101) = \Pr(010) = \frac{1}{2}$$
$$\Pr(s^L) = 0, \text{ otherwise}$$

$$H(1) = H(2) = H(L \geq 1) = 1$$



Period-P Process:

$$H(L \geq P) = \log_2(P)$$

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

Use: Typical sequences have probability: $\Pr(s^L) \approx 2^{-L \cdot h_\mu}$

(Shannon-MacMillian-Breiman Theorem)

Length-L Estimate of Entropy Rate:

$$\widehat{h}_\mu(L) = H(s_L | s_1 \cdots s_{L-1})$$

$$\widehat{h}_\mu(L) = H(L) - H(L-1)$$

Boundary conditions:

$\widehat{h}_\mu(0) = \log_2 |\mathcal{A}|$: no measurements, all things possible

$\widehat{h}_\mu(1) = H(1)$

Monotonic decreasing: $\widehat{h}_\mu(L) \leq \widehat{h}_\mu(L-1)$

Conditioning cannot increase entropy:

$$H(s_L | s_1 \cdots s_{L-1}) \leq H(s_L | s_2 \cdots s_{L-1}) = H(s_{L-1} | s_1 \cdots s_{L-2})$$

Entropy rate as conditional (predictive) entropy:

$$\widehat{h}_\mu = \lim_{L\to\infty} \widehat{h}_\mu(L) = \lim_{L\to\infty} H(s_0|\overleftarrow{s}^L) = H(s_0|\overleftarrow{s})$$

Interpretations:

Uncertainty in next measurement, given past

A measure of unpredictability

Asymptotic slope of block entropy

Alternate entropy rate definitions agree:

$\widehat{h}_\mu = h_\mu$

Memory in Processes

Motivation:

Previous: Measures of randomness
Block entropy $H(L)$
Entropy rate h_μ

Today's end point:
Measures of memory & information storage

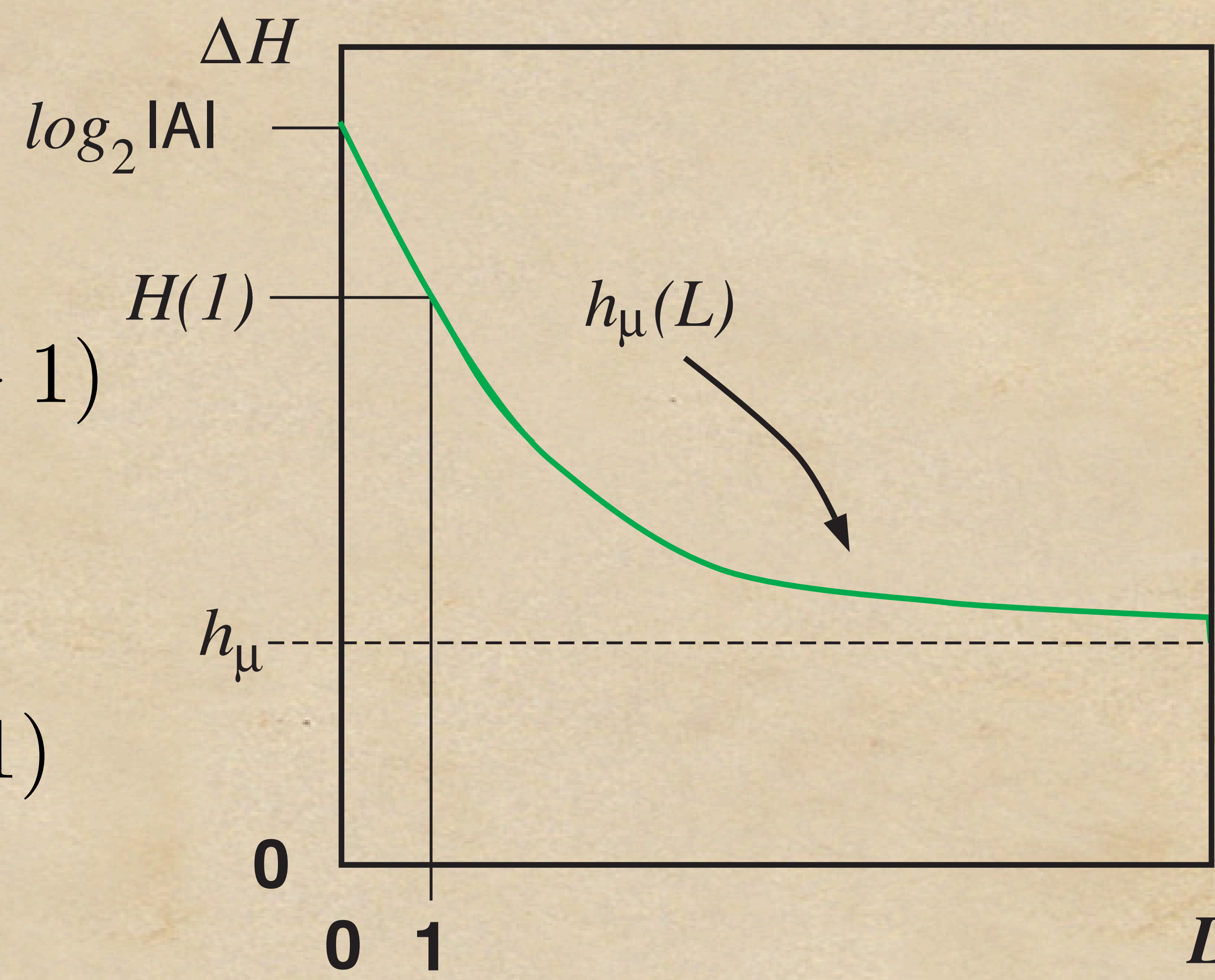
Big Picture:
Complementary properties of a source.
Need both: Measures of randomness *and* structure.

Entropy Convergence

Length-L entropy rate estimate:

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

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



Monotonic decreasing:

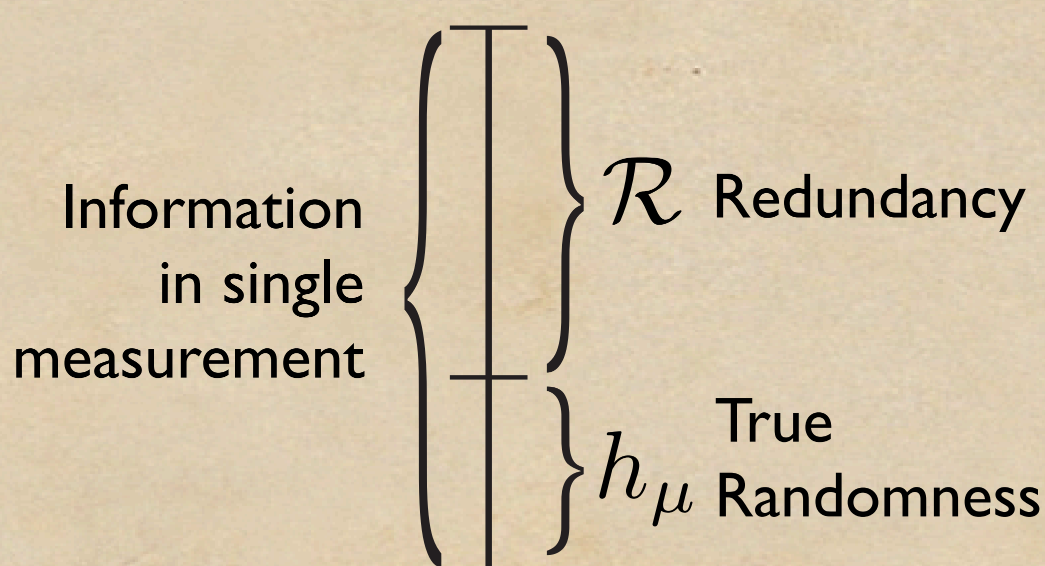
$$h_\mu(L) \leq h_\mu(L-1)$$

Process appears less random
as account for longer correlations

Redundancy in Processes

$$\mathcal{R} = \log_2 |\mathcal{A}| - h_\mu$$

Anatomy of Measurement:

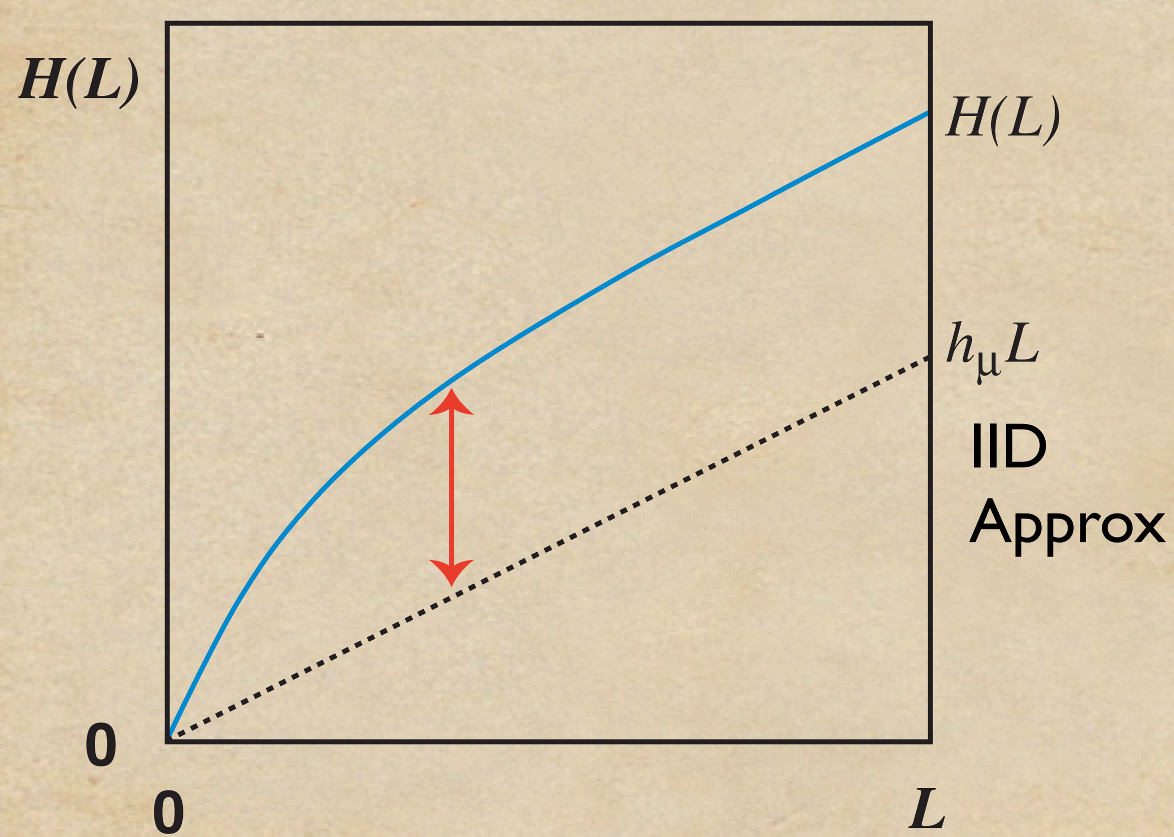


Redundancy in Processes ...

$$\mathcal{R} = \lim_{L\to\infty} \mathcal{D}(\Pr(s^L)||U(s^L))$$

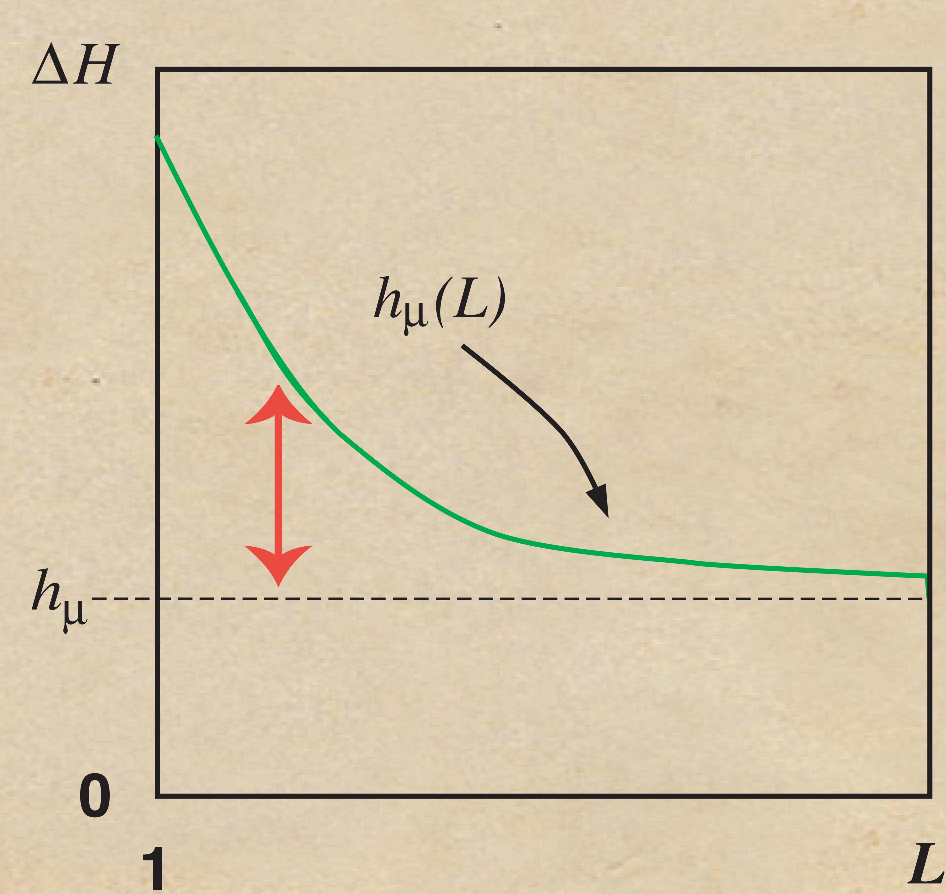
Redundancy in words:

$$\mathcal{R}(L) = H(L) - h_\mu L$$



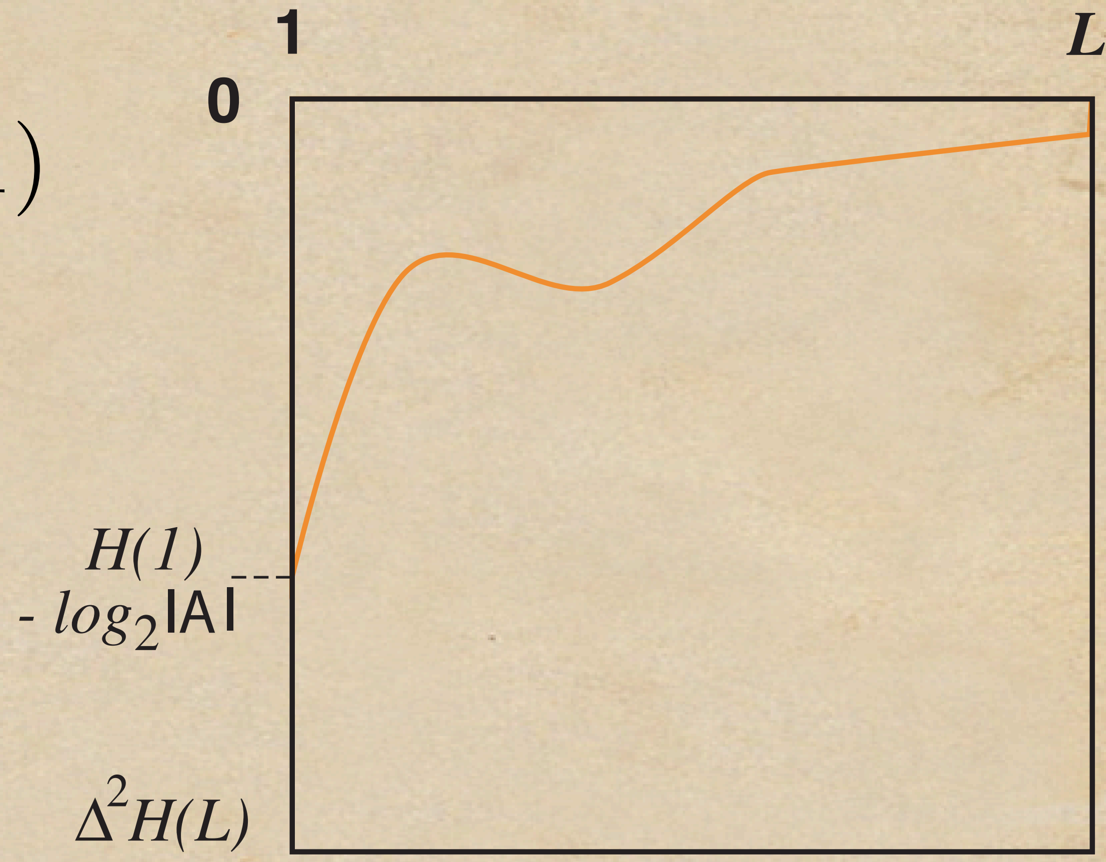
Redundancy per symbol:

$$r(L) = \mathcal{R}(L) - \mathcal{R}(L-1) = h_\mu(L) - h_\mu$$



Predictability Gain

$$\Delta^2 H(L) = h_\mu(L) - h_\mu(L-1)$$



Boundary condition:

$$\Delta^2 H(1) = H(1) - \log_2 |\mathcal{A}|$$

Rate at which unpredictability is lost

Properties:

(1) $H(L)$ Curvature: $\Delta^2 H(L) = H(L) - 2H(L-1) - H(L-2)$

(2) $H(L)$ Concavity: $\Delta^2 H(L) \leq 0$

(3) $|\Delta^2 H(L)| \gg 1 \Rightarrow$ Lth measurement significant

Entropy Hierarchy

So far have taken derivatives:

(1) Block entropy: $H(L)$

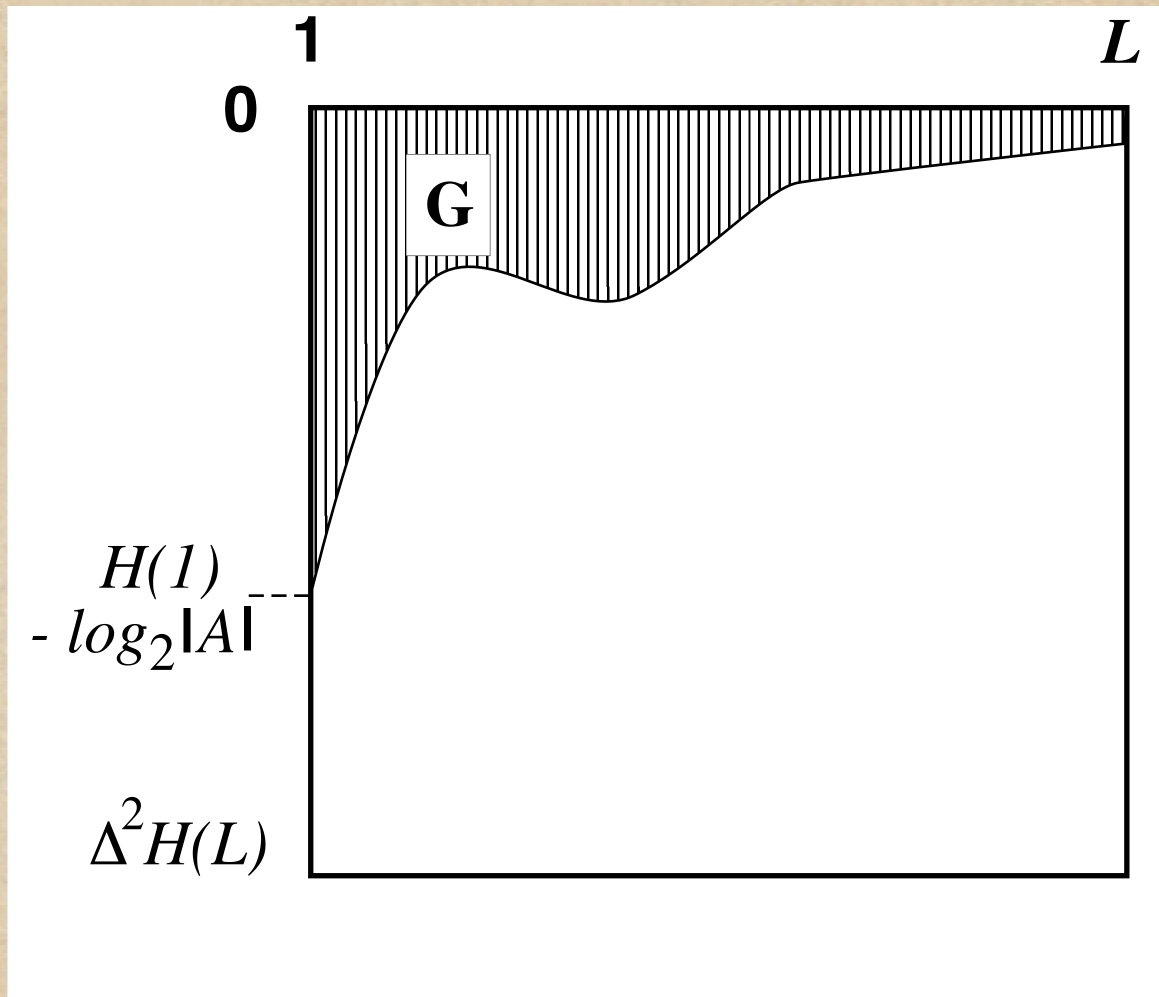
(2) Entropy rate: $h_\mu(L) = \Delta H(L)$

(3) Predictability gain: $\Delta h_\mu(L) = \Delta^2 H(L)$

Now take integrals!

Total Predictability:

$$\mathbf{G} = \sum_{L=1}^{\infty} \Delta^2 H(L)$$



Redundancy:

$$-\mathbf{G} = \mathcal{R} = \log_2 |\mathcal{A}| - h_\mu$$

Interpretation:

(1) Account for all correlations to see intrinsic randomness

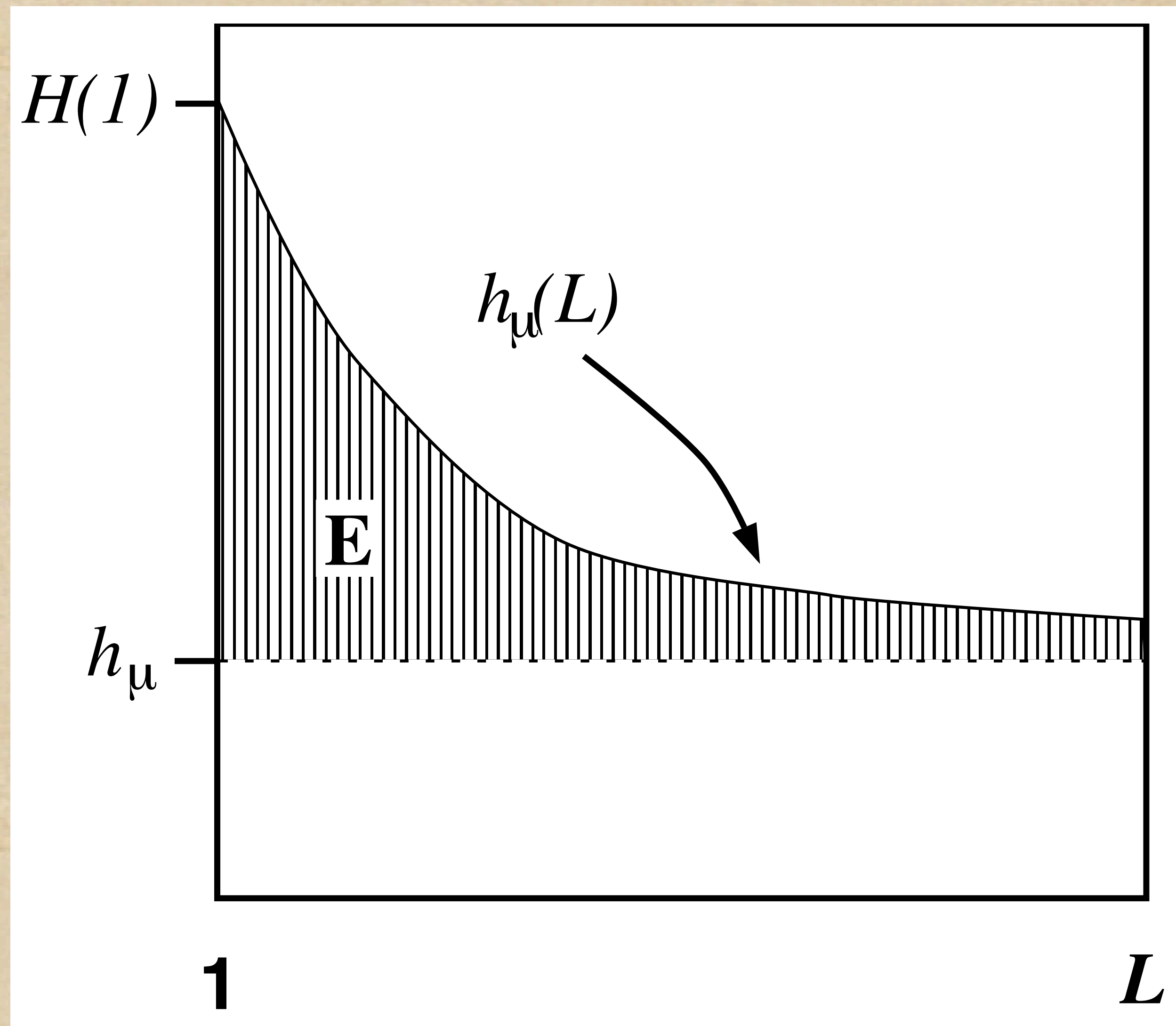
(2) Until that point, correlations appear as *excess randomness*

Excess Entropy:

As entropy convergence:

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

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



As intrinsic redundancy:

$$\mathbf{E} = \sum_{L=1}^{\infty} r(L)$$

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.

Excess Entropy ...

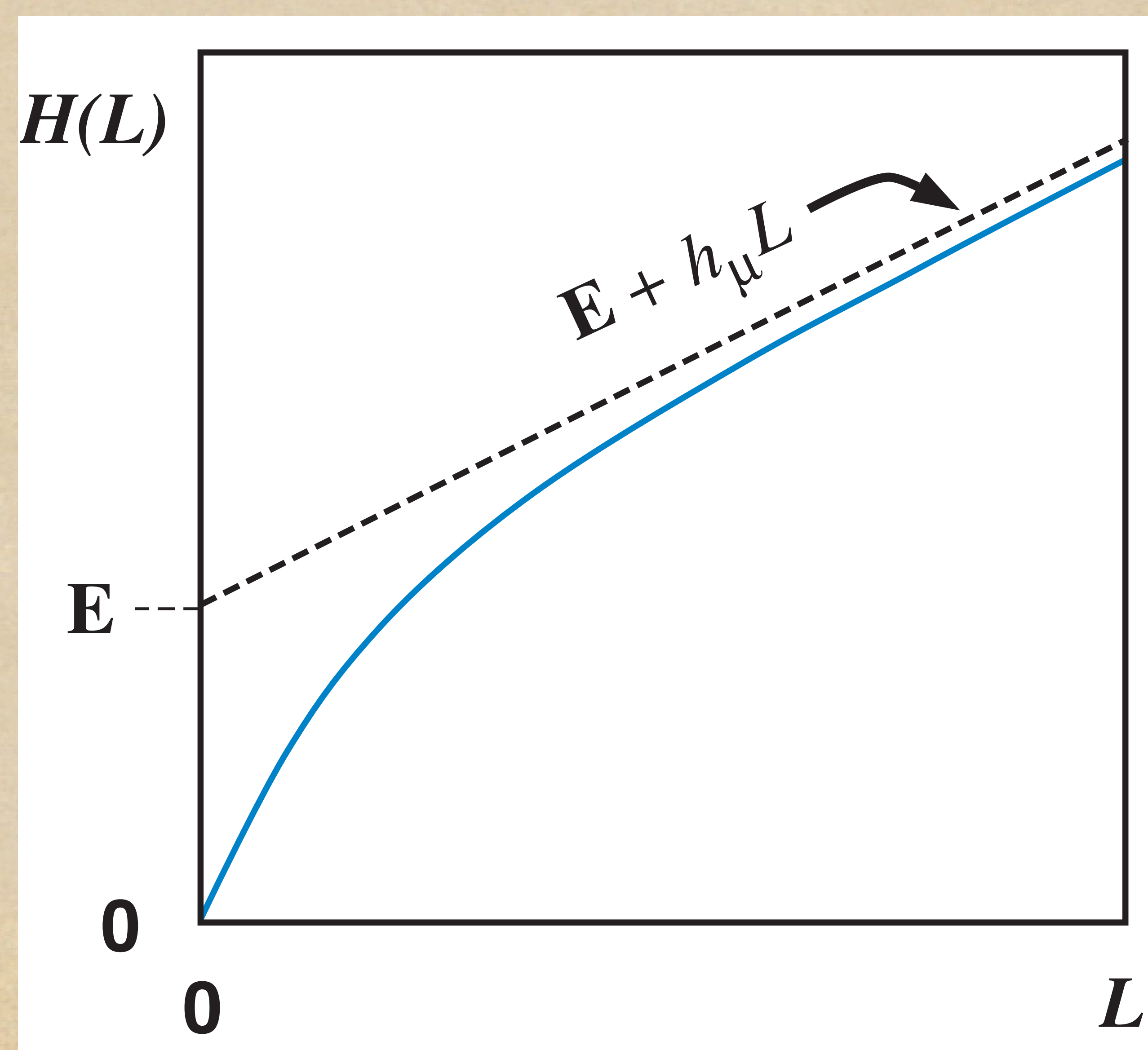
As asymptote of entropy growth:

$$\mathbf{E} = \lim_{L \to \infty} [H(L) - h_\mu L]$$

That is,

$$H(L) \propto \mathbf{E} + h_\mu L$$

Y-Intercept of
entropy growth

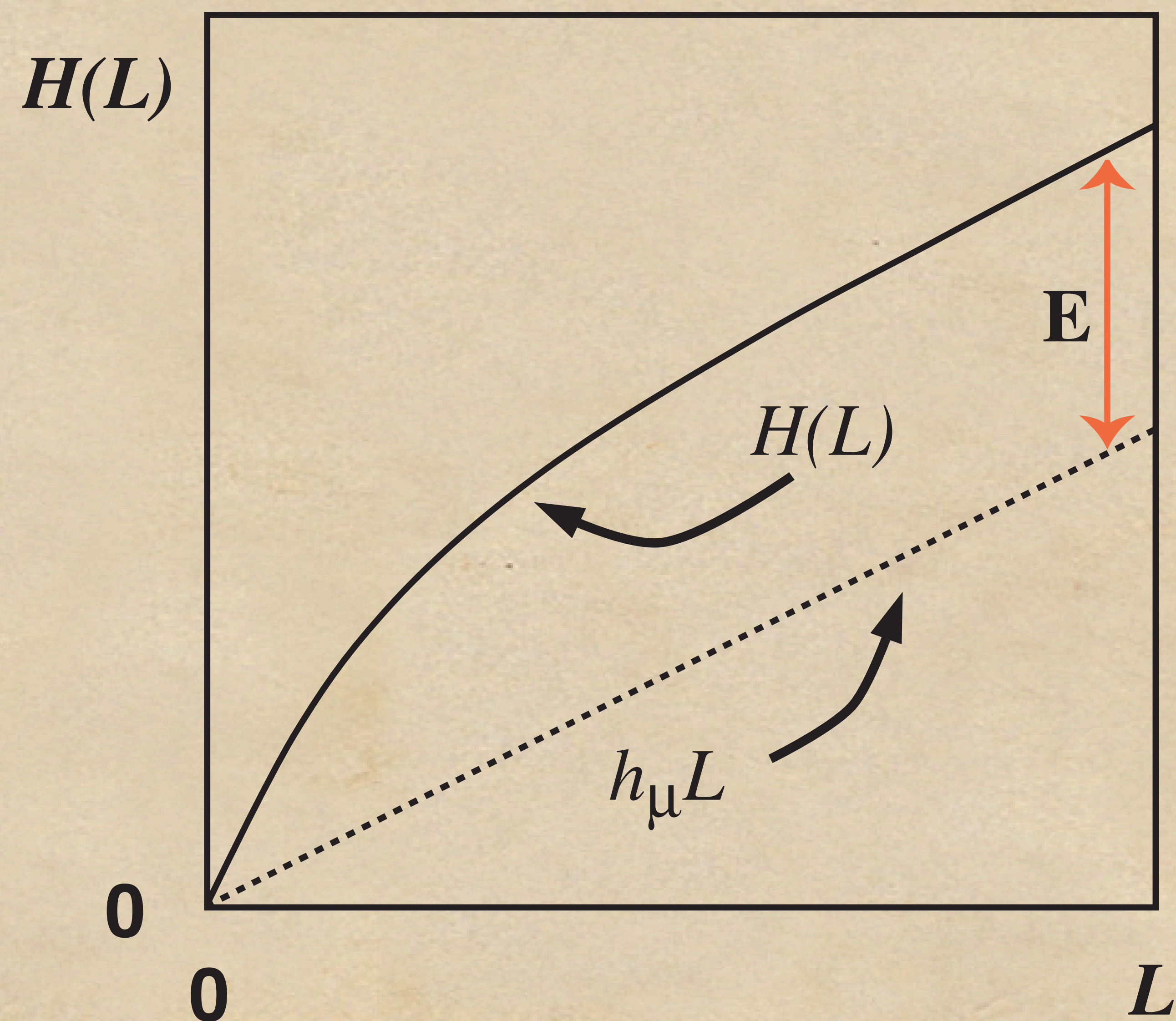


Excess Entropy ...

Cost of Amnesia:

Forget what you know:

Information needed to recover predicting with error ~h_μ



Cf. Memoryless Source: IID at same entropy rate

Excess Entropy ...

as Mutual information between past and future:

View process as a communication channel: Past to Future

$$\mathbf{E} = I(\overleftarrow{S}; \overrightarrow{S})$$

Property:

Symmetric in time

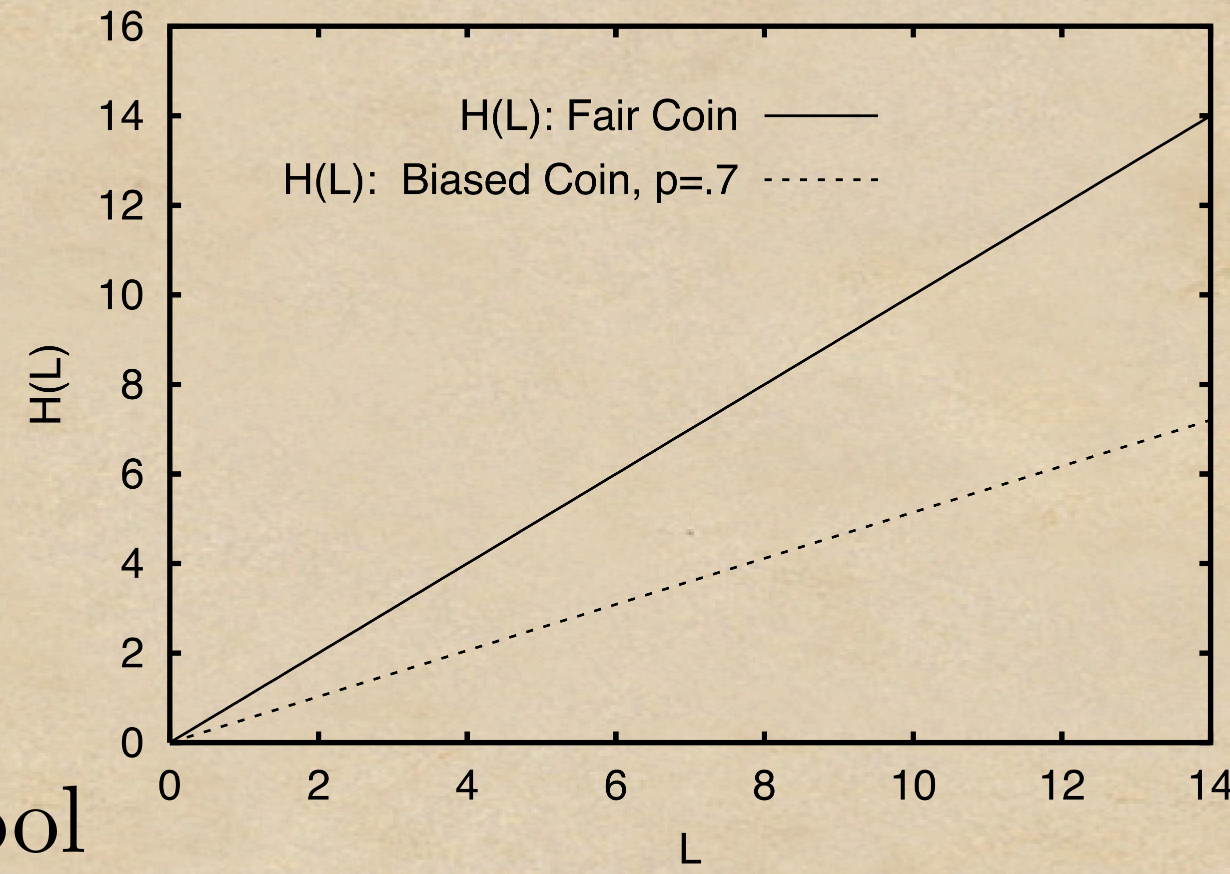
Interpretation:

Information that process communicates from past to future.

Reduction in uncertainty about the future, given the past.

Reduction in uncertainty about the past, given the future.

Examples of Excess Entropy:



Fair Coin:

$h_\mu = 1$ bit per symbol

$\mathbf{E} = 0$ bits

Biased Coin:

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

$\mathbf{E} = 0$ bits

Any IID Process:

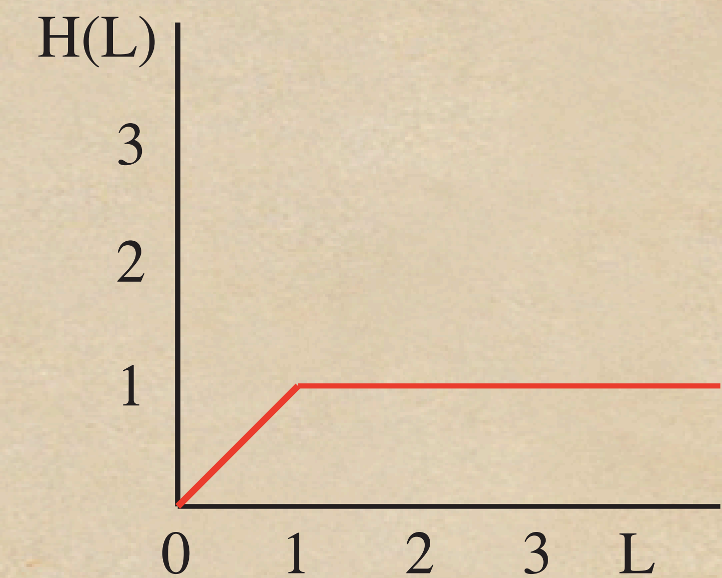
$h_\mu = H(X)$ bits per symbol

$\mathbf{E} = 0$ bits

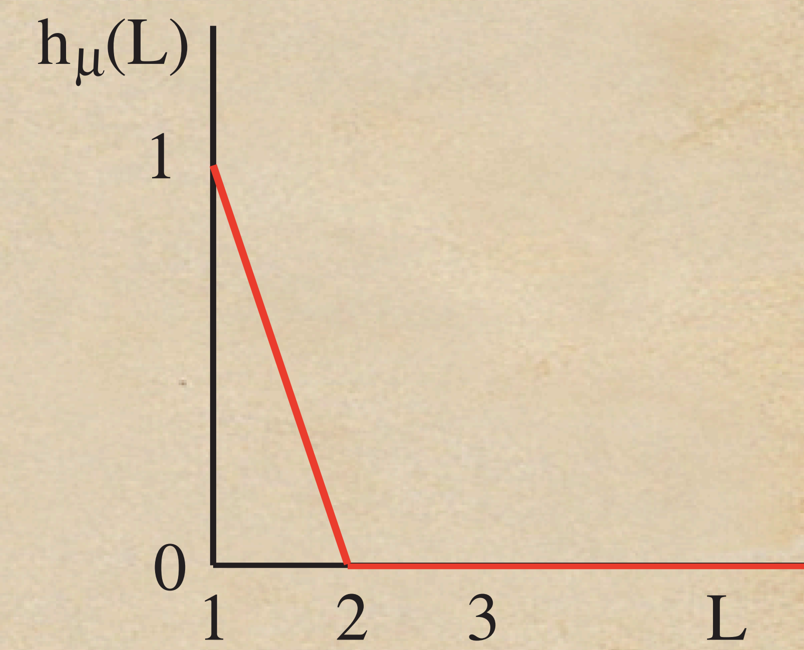
Examples of Excess Entropy ...

Period-2 Process: 010101010101

$$H(1) = 1$$
$$H(2) = 1$$
$$H(3) = 1$$



$$h_\mu(1) = 1$$
$$h_\mu(2) = 0$$
$$h_\mu(3) = 0$$



$$h_\mu = 0 \text{ bits per symbol}$$

$$\mathbf{E} = 1 \text{ bit}$$

Meaning:

1 bit of phase information

0-phase or 1-phase?

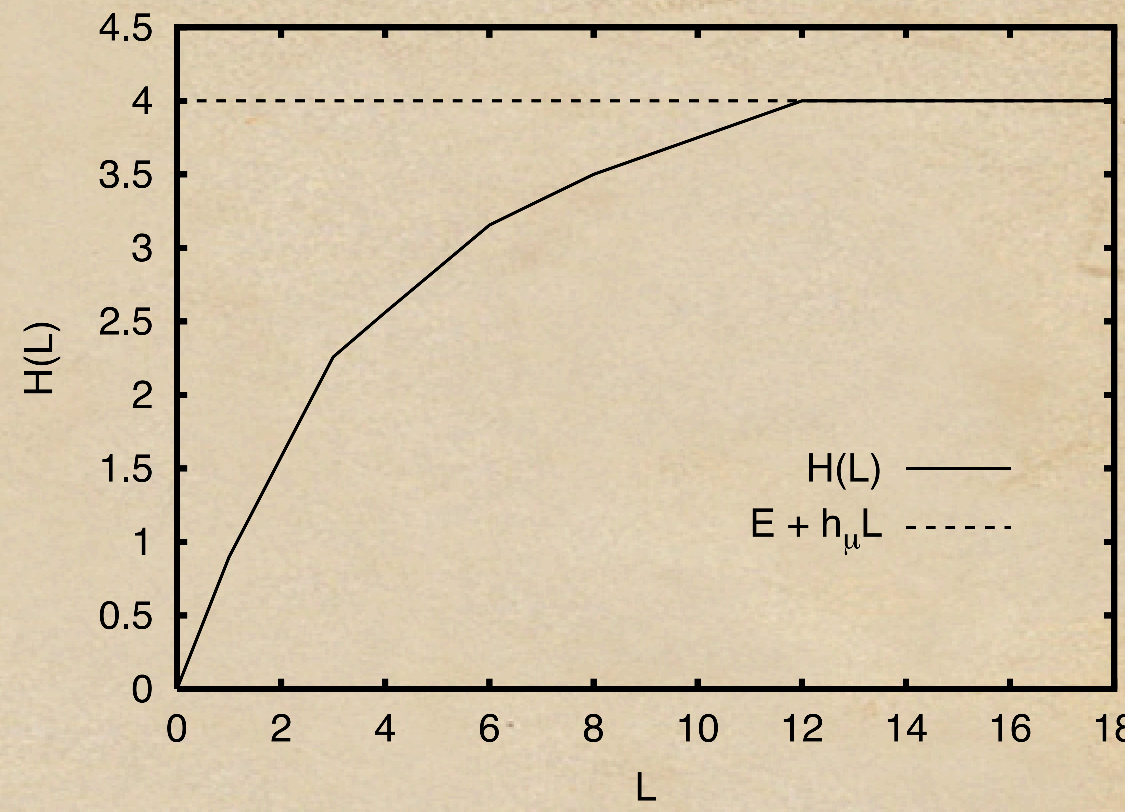
Examples of Excess Entropy ...

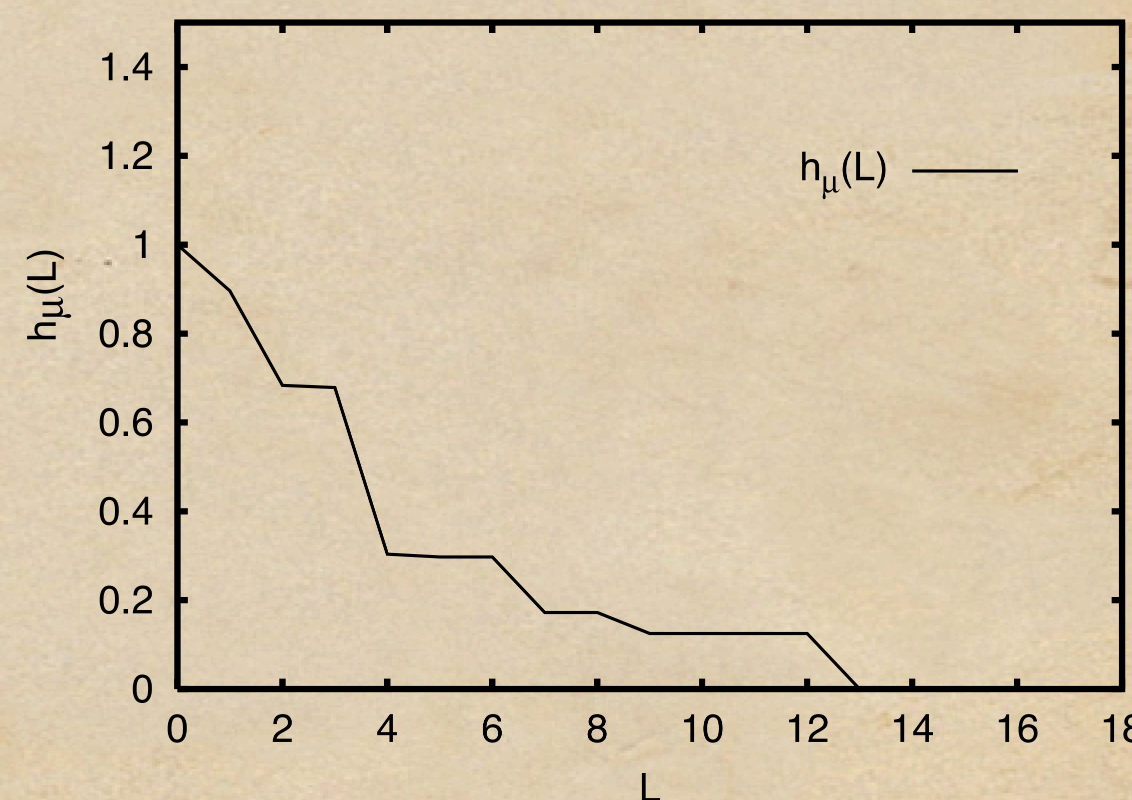
Period-16 Process:

$$(1010111011101110)^\infty$$

$$h_\mu = 0 \text{ bits per symbol}$$

$$\mathbf{E} = 4 \text{ bits}$$





Period-P Processes:

$$h_\mu = 0 \text{ bits per symbol}$$

$$\mathbf{E} = \log_2 P \text{ bits}$$

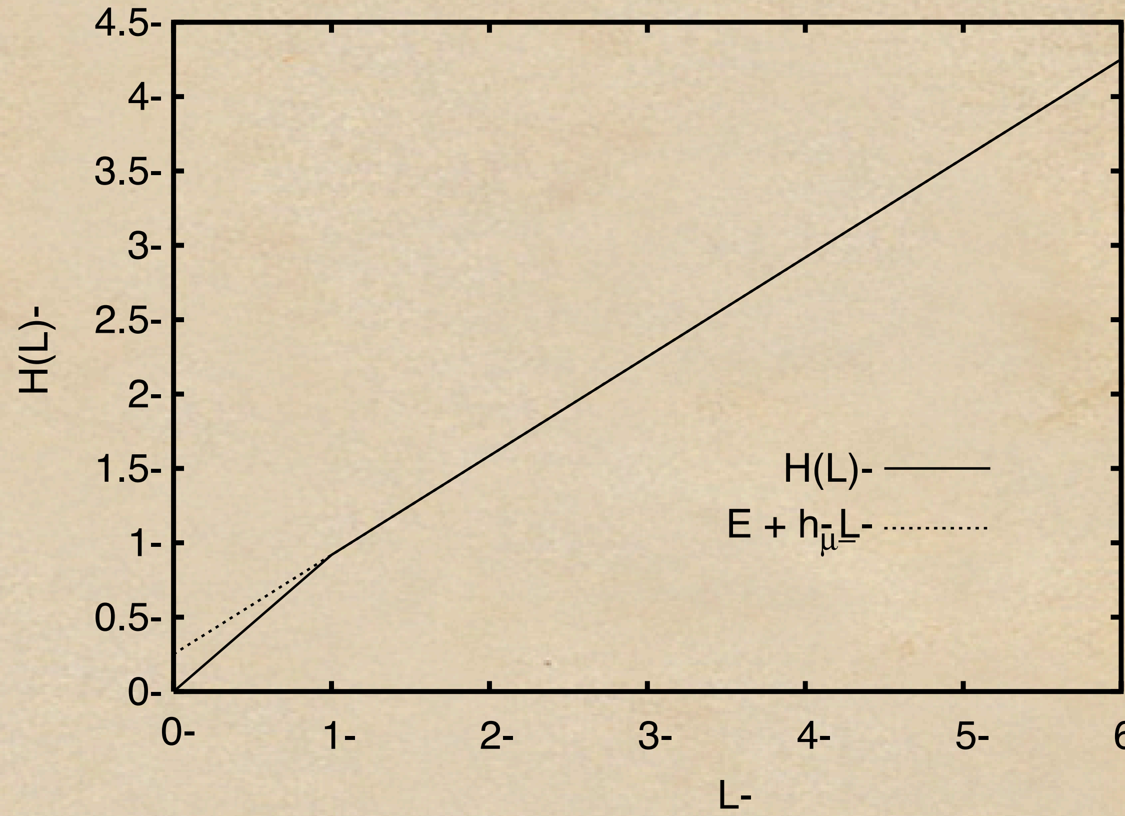
Entropy rate does not distinguish periodic processes

Examples of Excess Entropy ...

Golden Mean Process:

$$h_\mu = \tfrac{2}{3} \text{ bits per symbol}$$

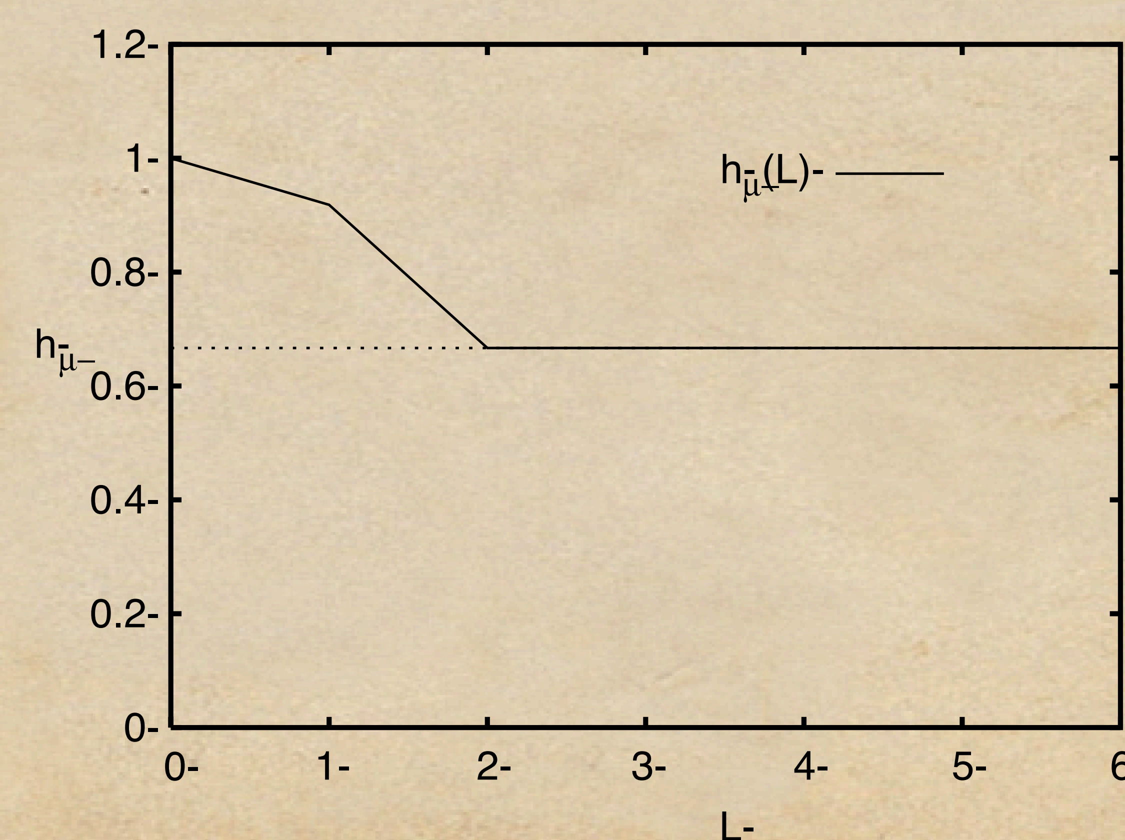
$$\mathbf{E} \approx 0.2516 \text{ bits}$$



R-Block Markov Chain:

$$\mathbf{E} = H(R) - R \cdot h_\mu$$

(E.g., 1D Ising Spin System)



Examples of Excess Entropy:

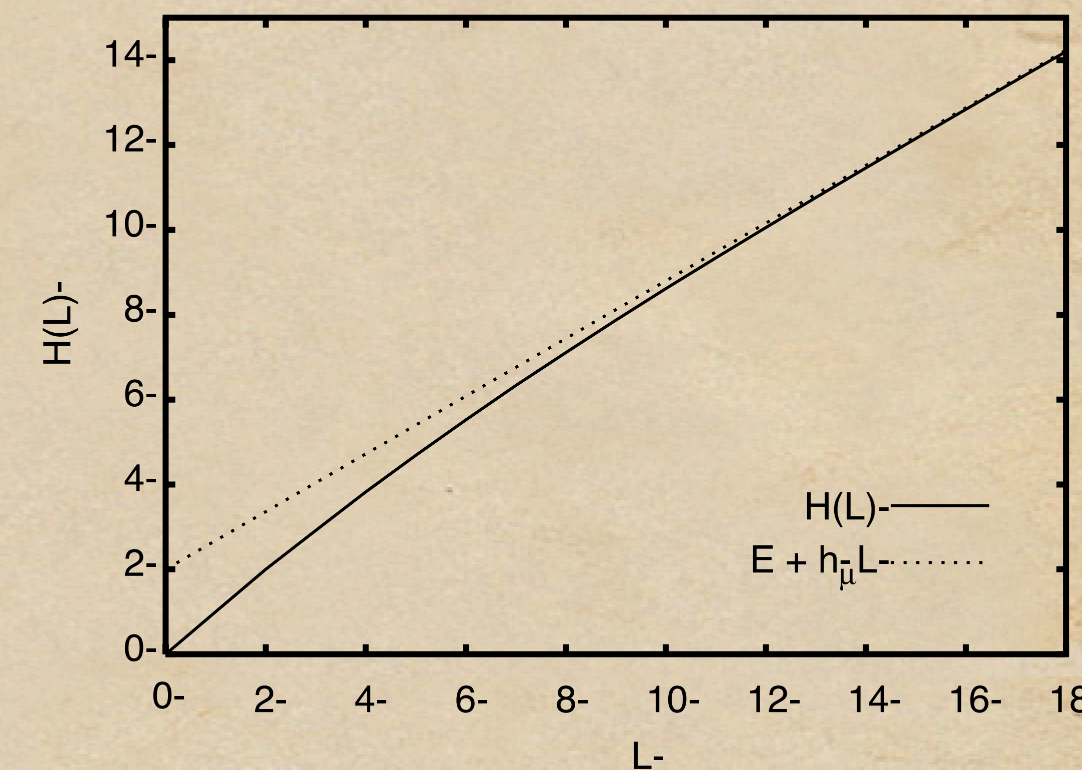
Finitary Processes: Exponential entropy convergence

Random-Random XOR (RRXOR) Process:

$$S_t = S_{t-1} \text{ XOR } S_{t-2}$$

$$h_\mu = \frac{2}{3} \text{ bits per symbol}$$

$$\mathbf{E} \approx 2.252 \text{ bits}$$

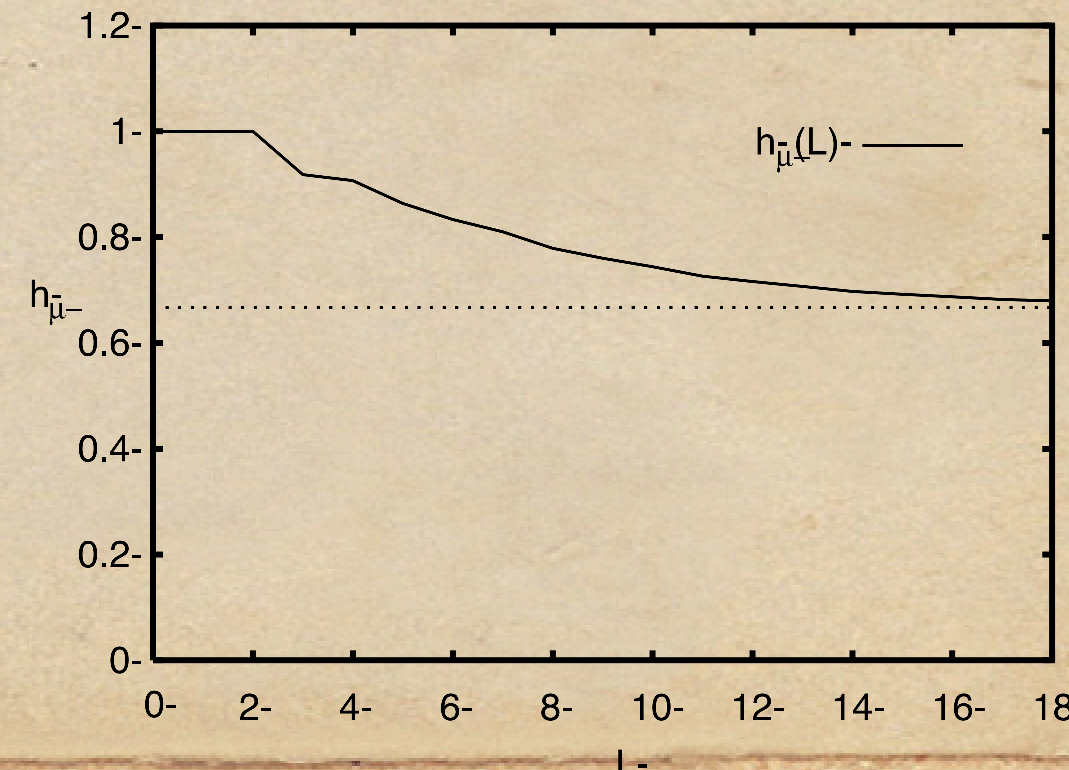


General finitary processes: Exponential convergence:

$$h_\mu(L) - h_\mu \approx 2^{-\gamma L}$$

$$\mathbf{E} = \frac{H(1) - h_\mu}{1 - 2^{-\gamma}}$$

$$\gamma \approx 0.30$$



Examples of Excess Entropy:

Infinitary Processes:

$$\mathbf{E} \rightarrow \infty$$

Excess entropy can diverge:

Slow entropy convergence

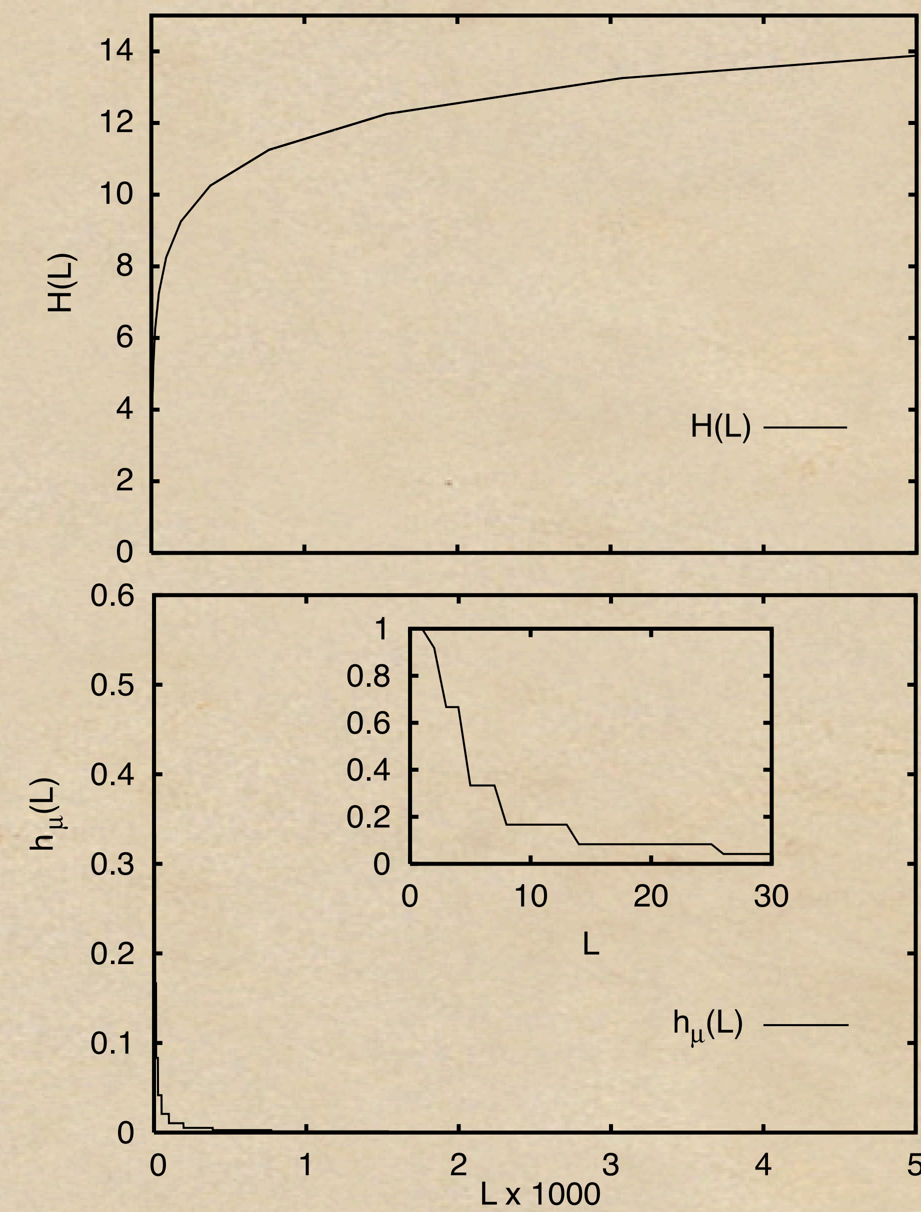
Long-range correlations

(e.g., at phase transitions)

Morse-Thue Process:

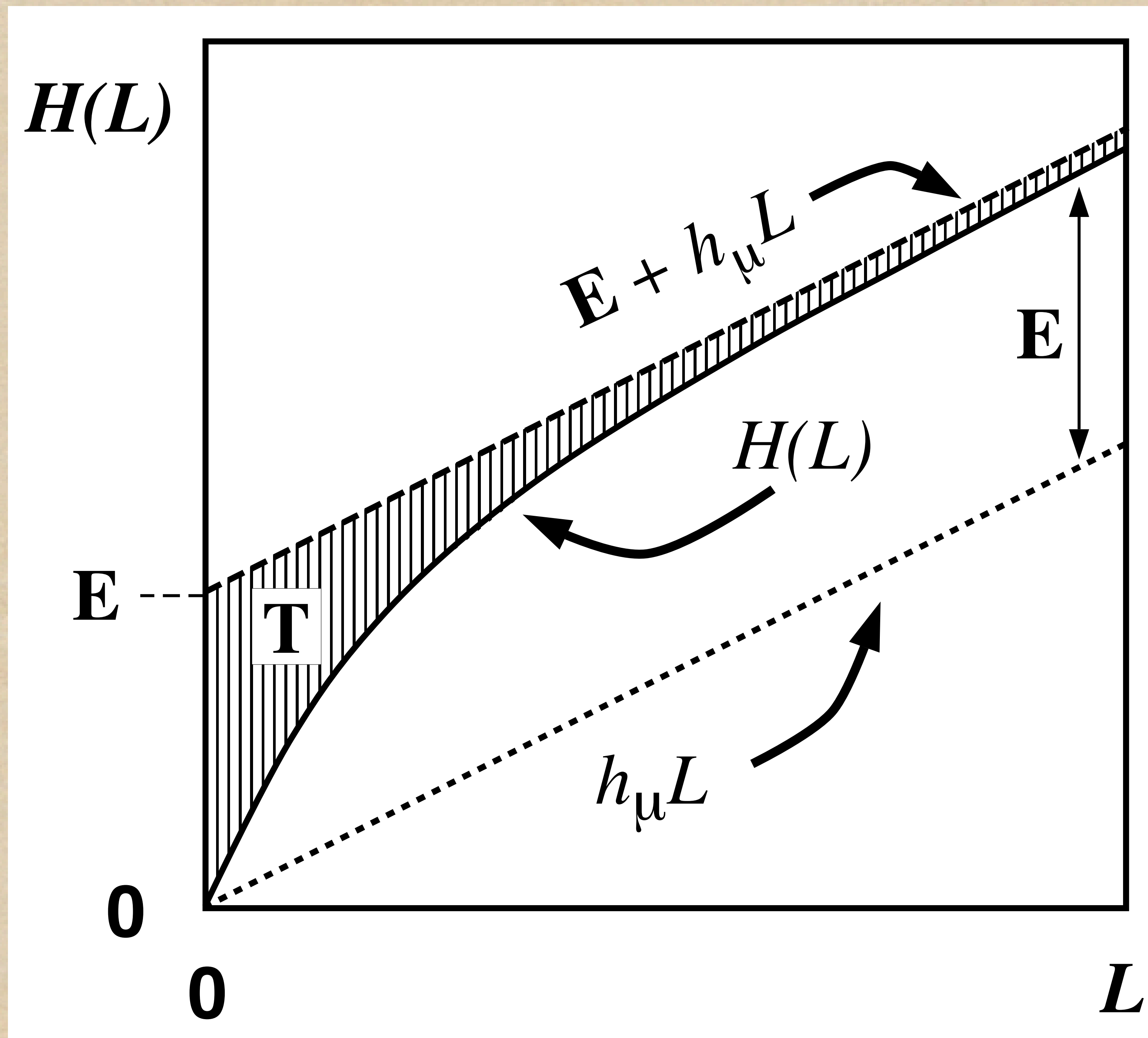
A context-free language

From Logistic map at onset of chaos



$$h_\mu = 0 \text{ bits per symbol}$$

Information-Entropy Roadmap for a Stochastic Process:



Calculus of the Entropy Hierarchy:

Via Discrete-Time Derivatives and Integrals

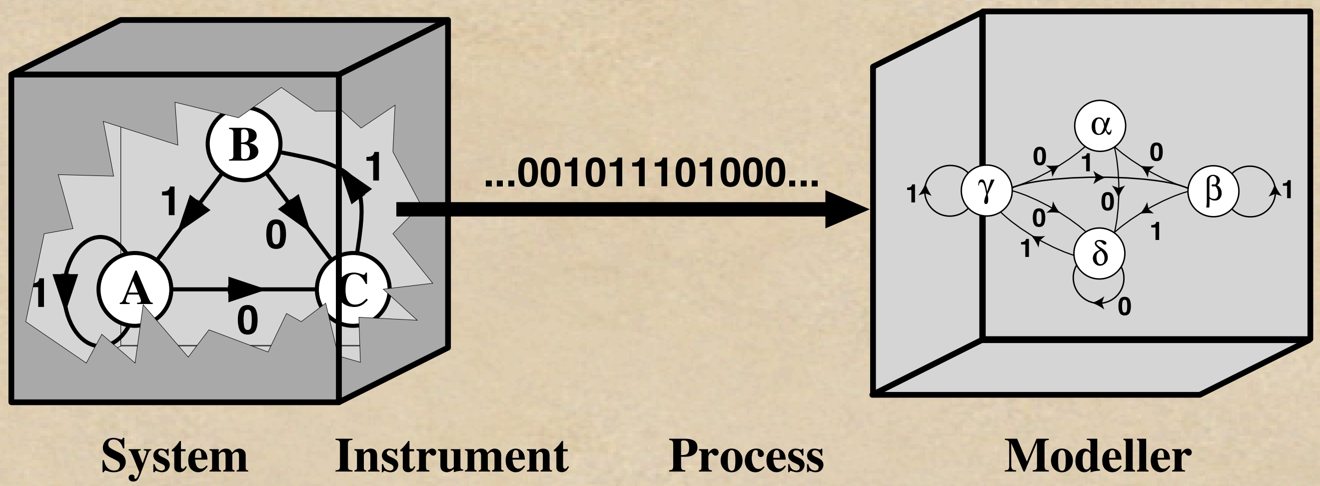
Level	Gain (Derivative)	Information (Integral)
0	Block Entropy $H(L)$	Transient Information $\mathbf{T} = \sum_{L=1}^{\infty} [\mathbf{E} + h_\mu L - H(L)]$
1	Entropy Rate Loss $h_\mu(L) = \Delta H(L)$	Excess Entropy $\mathbf{E} = \sum_{L=1}^{\infty} [h_\mu(L) - h_\mu]$
2	Predictability Gain $\Delta^2 H(L)$	Total Predictability (Redundancy) $\mathbf{G} = -\mathcal{R}$
...	...	...

What is information?

Depends on the question!

Uncertainty, surprise, randomness, ...
Compressibility.
Transmission rate.
Memory, apparent stored information, ...
Synchronization.
...

Project Pipeline:

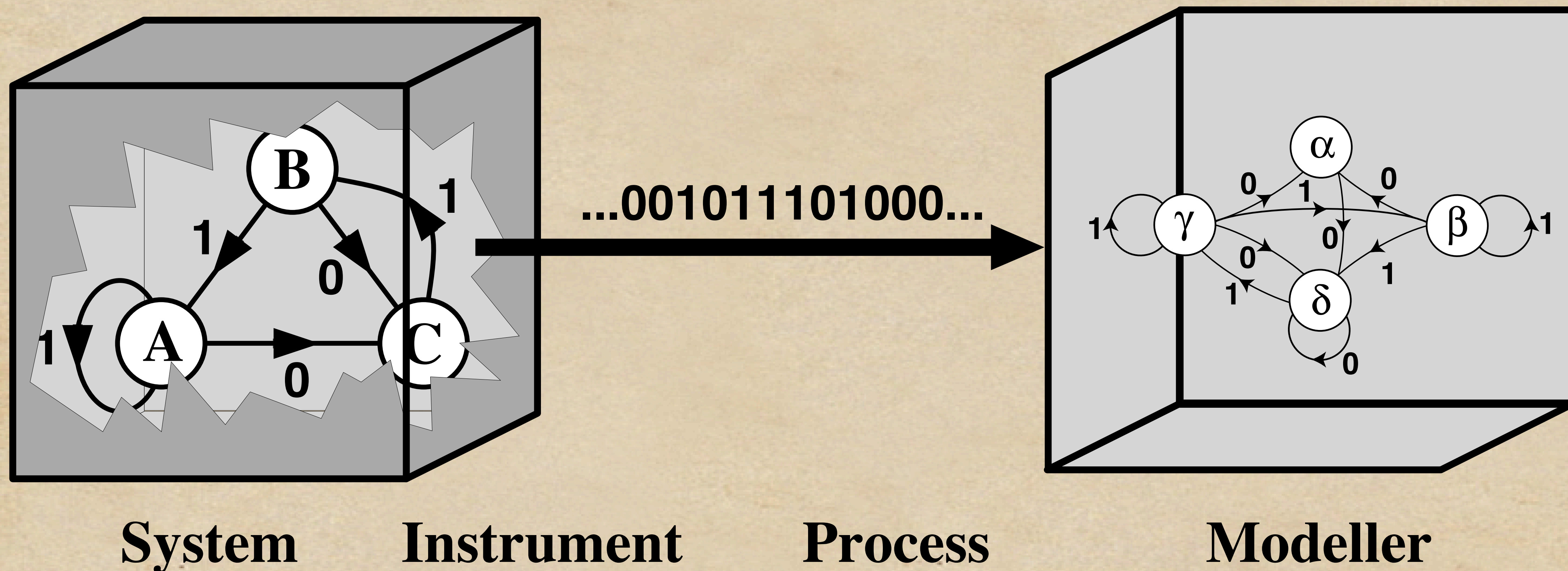


1. An information source: Calculate or estimate $\Pr(s^L)$
   a. Dynamical System:
      deterministic or stochastic
      low-dimensional or high-dimensional (spatial?)
   b. Design instrument (partition)
2. Information-theoretic analysis: $H(L)$
   a. How much information produced? h_μ
   b. How much stored information? $\mathbf{E}$
   c. How does observer synchronize? $\mathbf{T}$

Introduction to Computational Mechanics

- Learning as a Channel
- Prediction
- Causal Architecture
- Optimality
- Why We Must Model

The Learning Channel



Central questions:

What are the states?

What is the dynamic?

The Prediction Game

Rules:

1. I give you a data stream (an observed past sequence).

2. You predict its future.

3. You give a model (states & transitions) for the process.

The Prediction Game ...

Process I:

The Prediction Game ...

Process I:

Past: $\ldots 111111111111$

The Prediction Game ...

Process I:

Past: $\ldots 111111111111$

Your prediction is?

The Prediction Game ...

Process I:

Past: $\ldots 111111111111$

Your prediction is?

Future: $111111111111 \ldots$

The Prediction Game ...

Process I:

Past: $\ldots 111111111111$

Your prediction is?

Future: $111111111111\ldots$

Your model (states & dynamic) is?

The Prediction Game ...

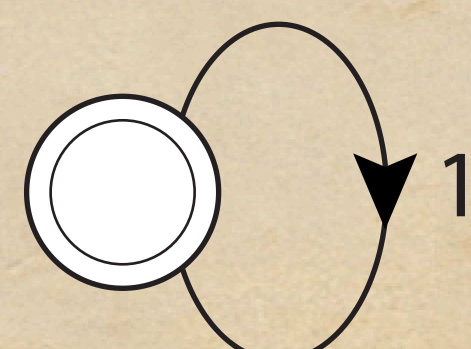
Process I:

Past: $\ldots 111111111111$

Your prediction is?

Future: $111111111111\ldots$

Your model (states & dynamic) is?



The Prediction Game ...

Process II:

The Prediction Game ...

Process II:

Past: $\ldots 10110010001101110$

The Prediction Game ...

Process II:

Past: $\ldots 10110010001101110$

Your prediction is?

The Prediction Game ...

Process II:

Past: $\ldots 10110010001101110$

Your prediction is?

Analysis: All words of length L occur & equally often

The Prediction Game ...

Process II:

Past: $\ldots 10110010001101110$

Your prediction is?

Analysis: All words of length L occur & equally often

Future: Well, anything can happen, how about?

The Prediction Game ...

Process II:

Past: $\ldots 10110010001101110$

Your prediction is?

Analysis: All words of length L occur & equally often

Future: Well, anything can happen, how about?

$01010111010001101\ldots$

The Prediction Game ...

Process II:

Past: $\ldots 10110010001101110$

Your prediction is?

Analysis: All words of length L occur & equally often

Future: Well, anything can happen, how about?

$$01010111010001101\ldots$$

Your model is?

The Prediction Game ...

Process II:

Past: $\ldots 101100100011011110$

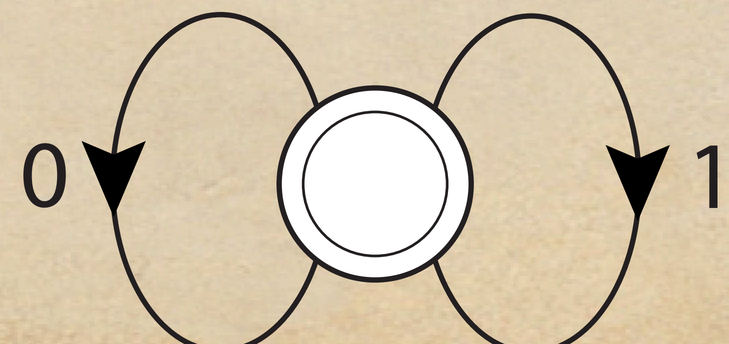
Your prediction is?

Analysis: All words of length L occur & equally often

Future: Well, anything can happen, how about?

$010101110100011101 \ldots$

Your model is?



The Prediction Game ...

Process III:

The Prediction Game ...

Process III:

Past: $\ldots 1010101010101010$

The Prediction Game ...

Process III:

Past: $\ldots 1010101010101010$

Your prediction is?

The Prediction Game ...

Process III:

Past: $\ldots 1010101010101010$

Your prediction is?

Future: $101010101010101 \ldots$

The Prediction Game ...

Process III:

Past: $\ldots 1010101010101010$

Your prediction is?

Future: $1010101010101010 \ldots$

Your model is?

The Prediction Game ...

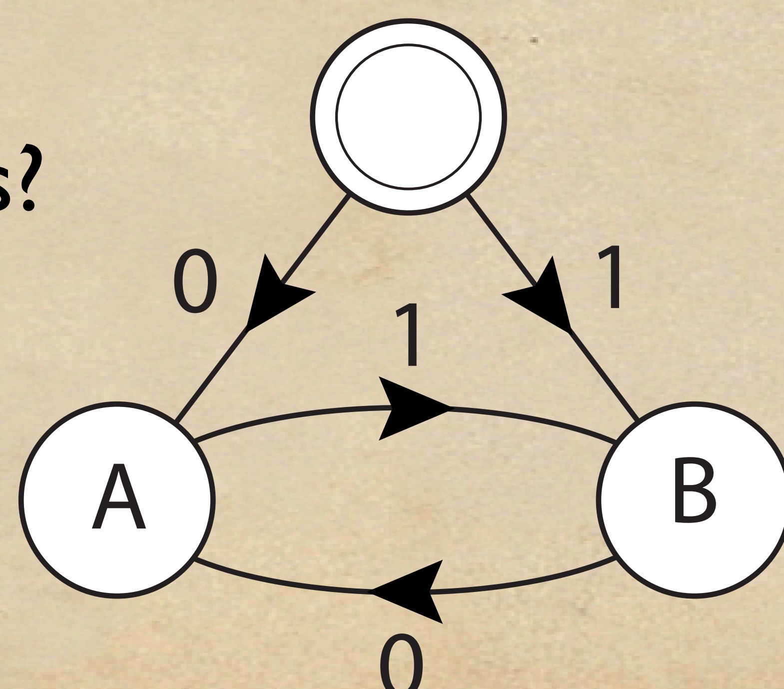
Process III:

Past: $\ldots 1010101010101010$

Your prediction is?

Future: $1010101010101010 \ldots$

Your model is?



Goal:

Predict the future $\overrightarrow{S}$

using information from the past $\overleftarrow{S}$

But what "information" to use?

We want to find the effective "states"

and the dynamic (state-to-state mapping)

How to define "states", if they are hidden?

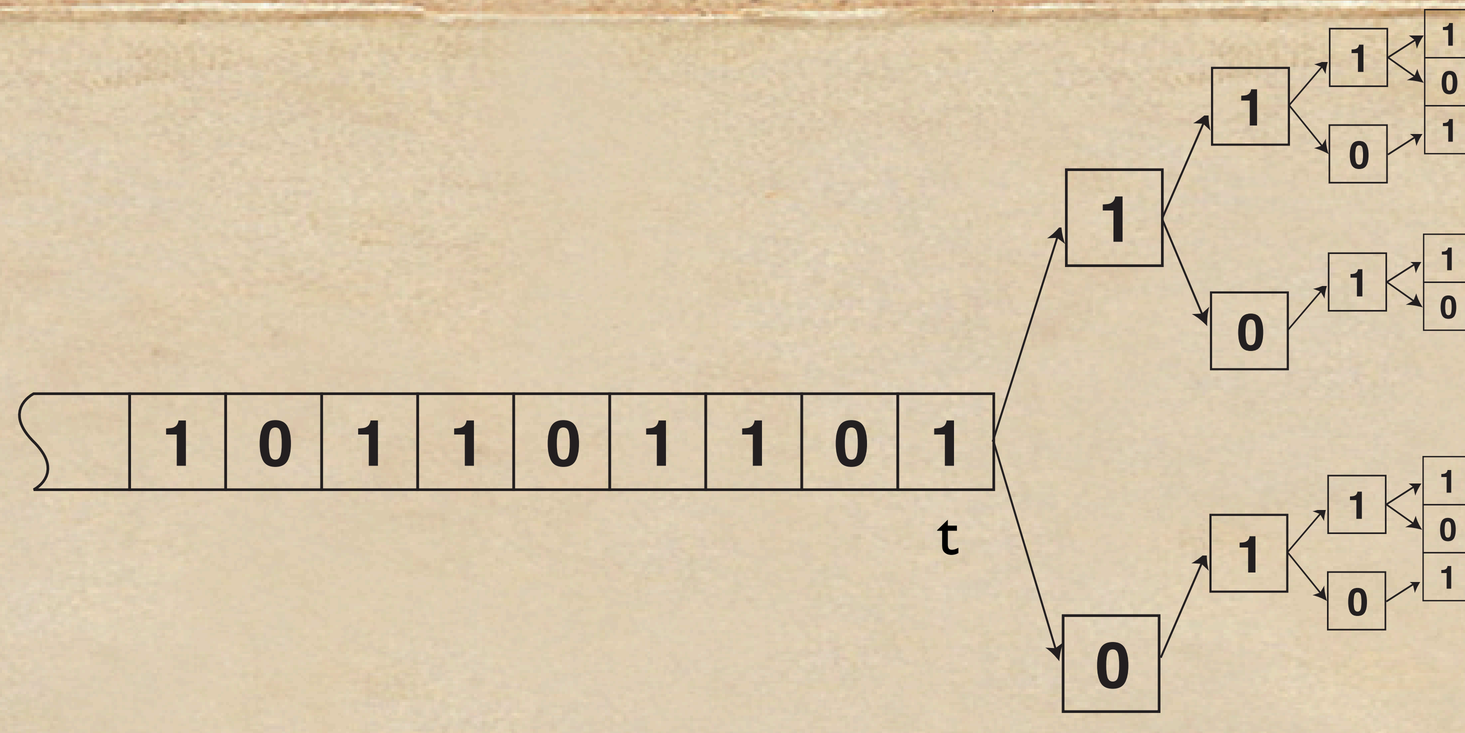
All we have are sequences of observations

Over some measurement alphabet $\mathcal{A}$

These symbols only indirectly reflect the hidden states

Effective States:

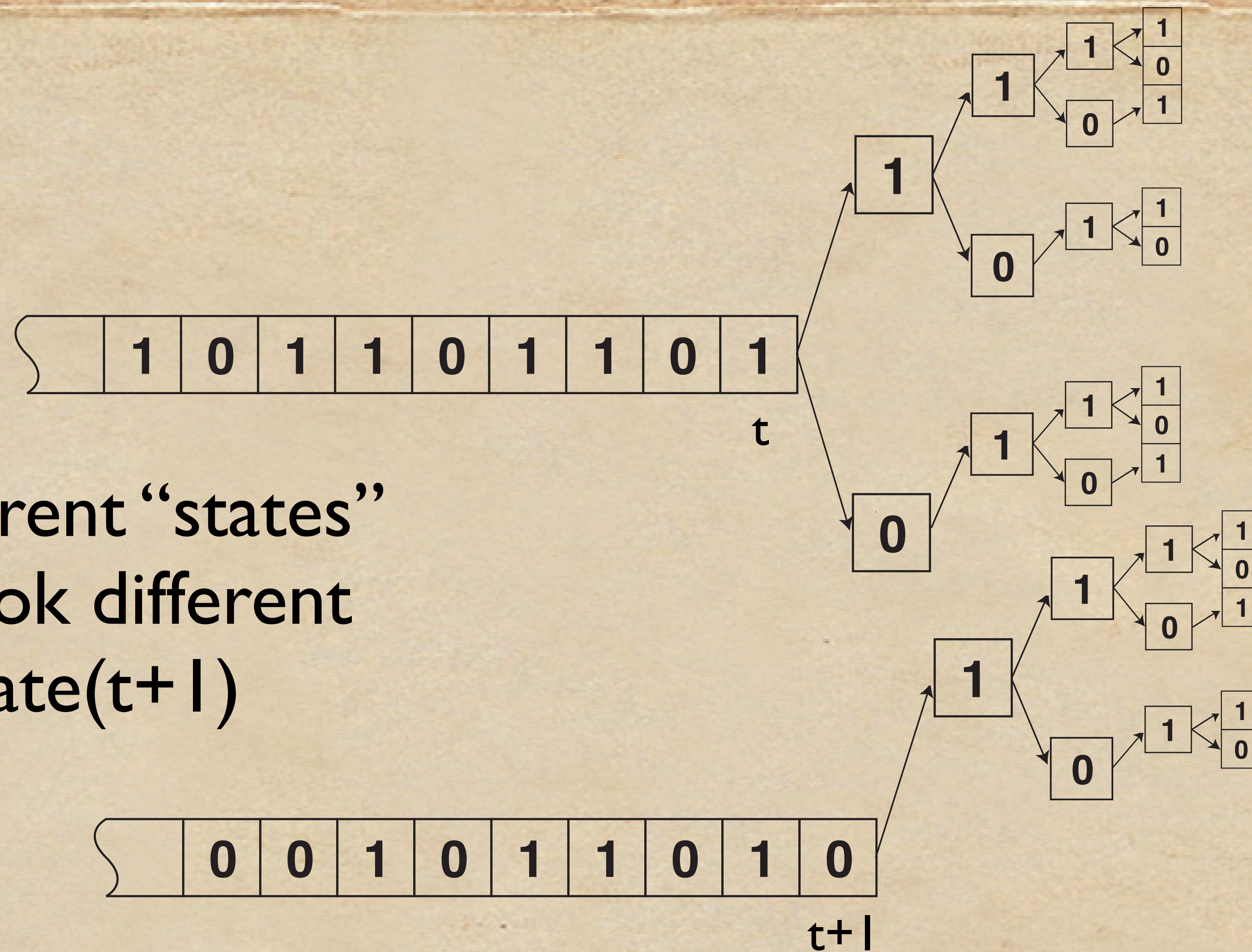
Effective States:



Effective States:

Effective States:

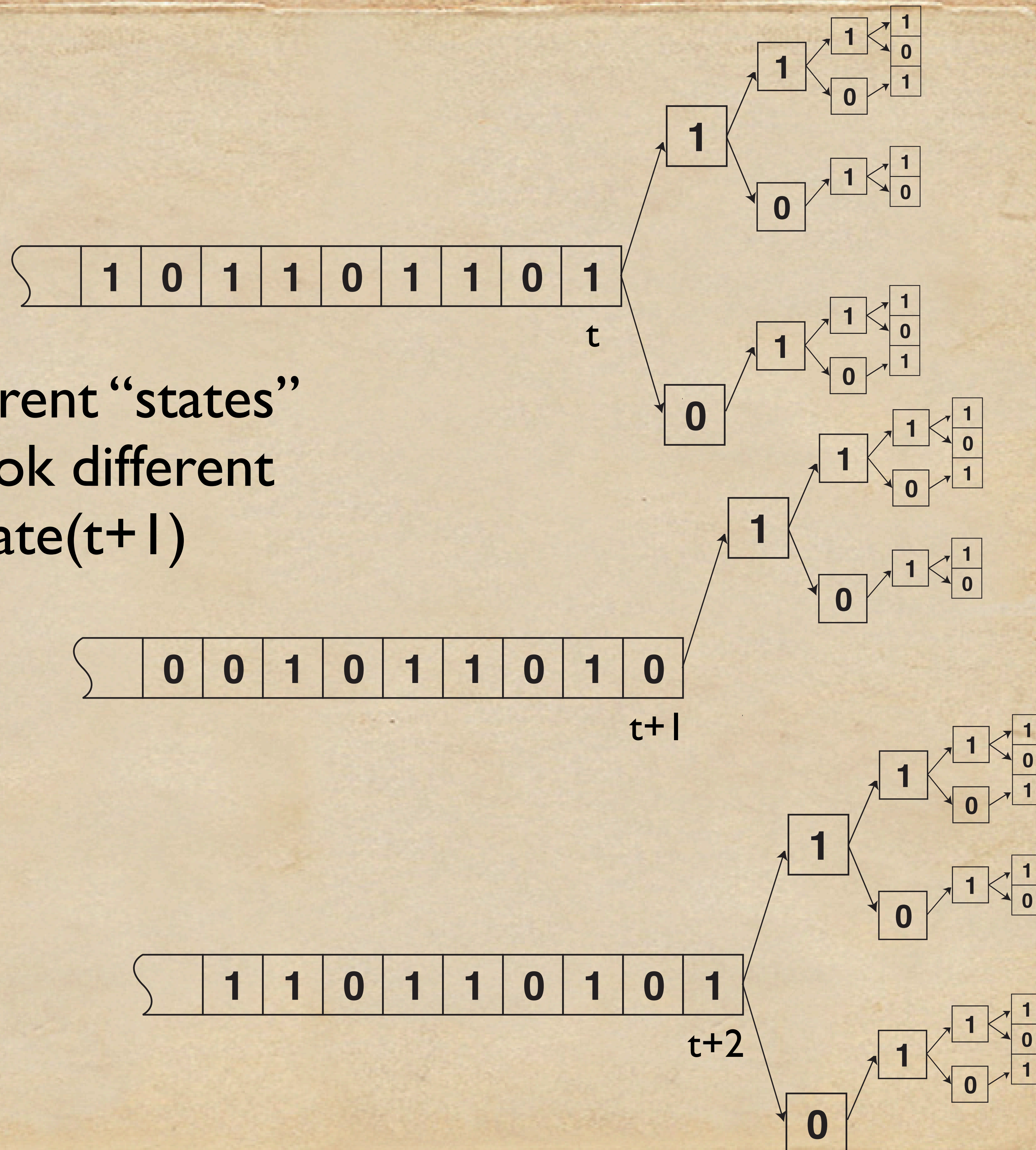
Process is in different "states"
when futures look different
State(t) $\nsim$ State(t+1)



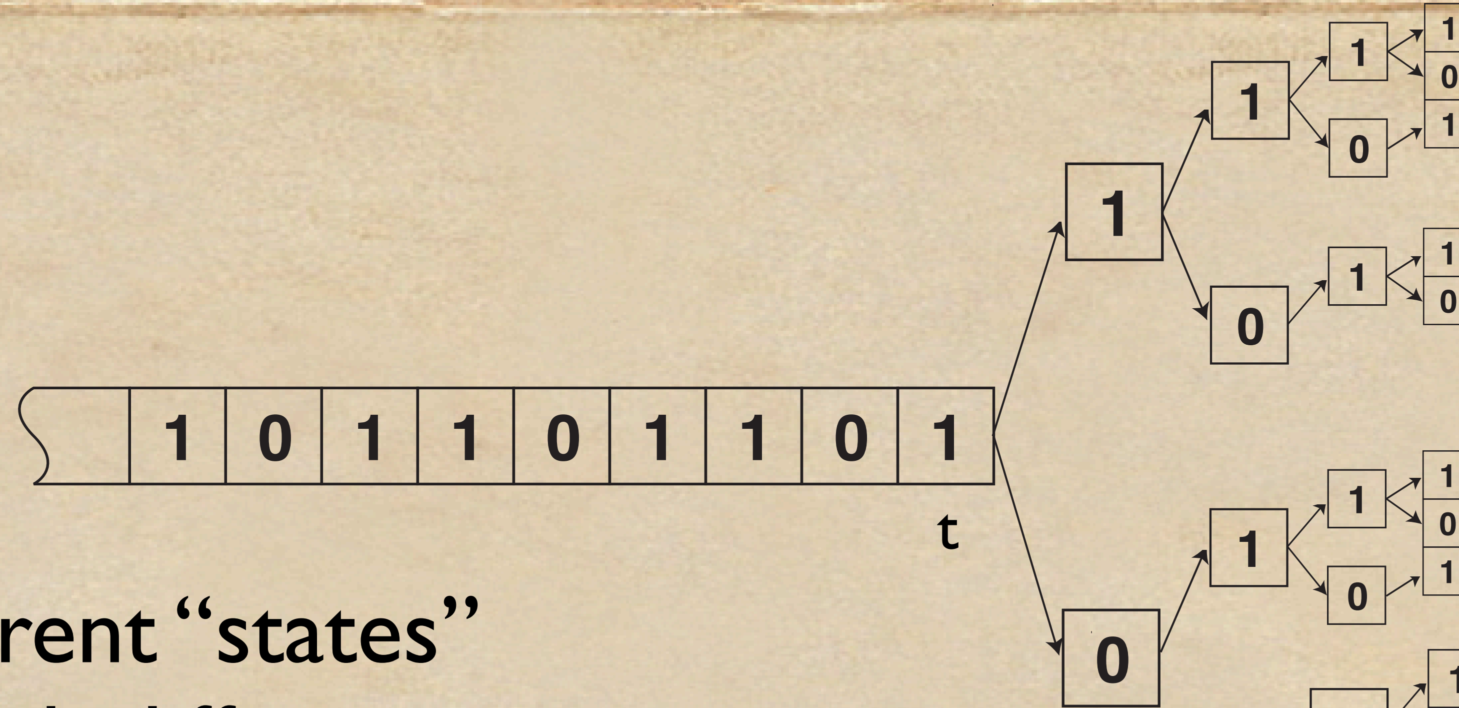
Effective States:

Process is in different "states" when futures look different

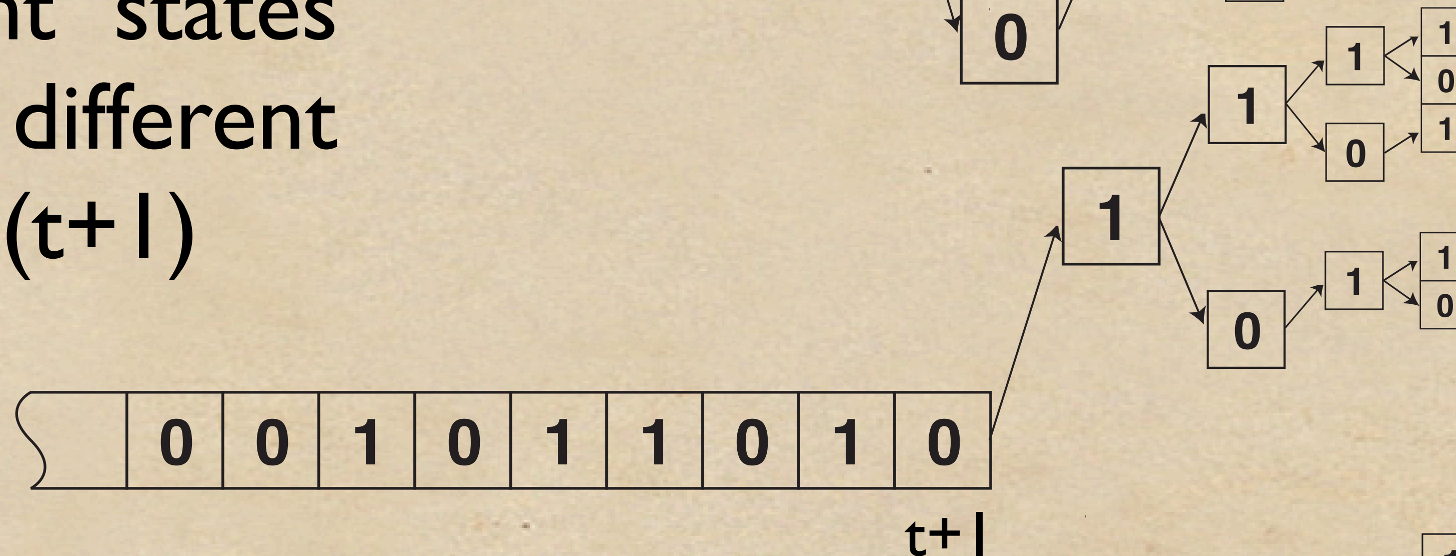
State(t) $\nsim$ State(t+1)



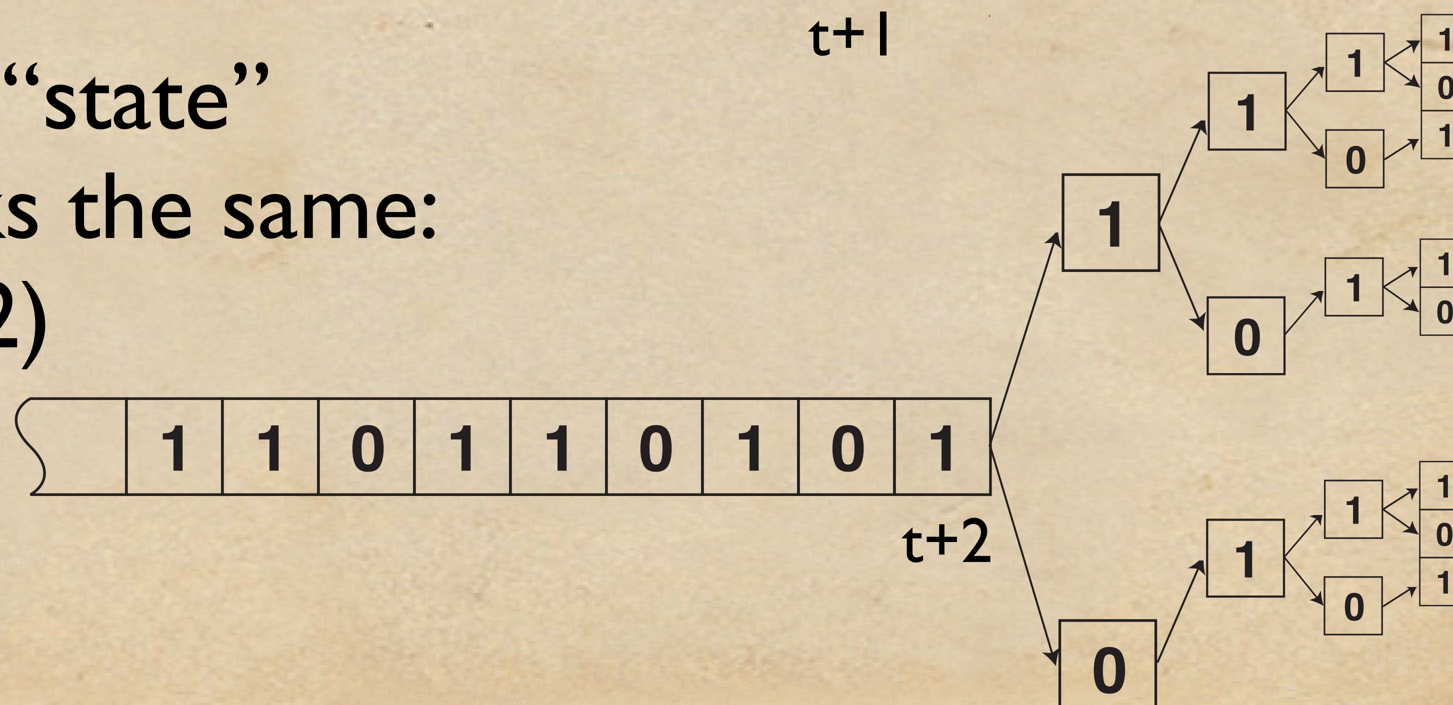
Effective States:



Process is in different "states"
when futures look different
$\text{State}(t) \nsim \text{State}(t+1)$



Process is in the same "state"
when the future looks the same:
$\text{State}(t) \sim \text{State}(t+2)$



Effective for what?

Prediction!

Find states that are effective for prediction.

What are the "predictive states" in the measurements?

Simple, but key observation:

Histories leading to the same predictions are equivalent.

What are these predictions?
How to group histories?

Effective for what?

What's a prediction?

A mapping from the past to the future.

Process $\Pr(\overleftrightarrow{S})$: $\overleftrightarrow{S} = \overleftarrow{S}\overrightarrow{S}$

Future: $\overrightarrow{S}^L$

Particular past: $\overleftarrow{s}$

Future Morph: $\Pr(\overrightarrow{S}^L \mid \overleftarrow{s})$ (the most general mapping)

Refined goal:

Predict as much about the future $\overrightarrow{S}$,

using as little of the past $\overleftarrow{S}$ as possible.

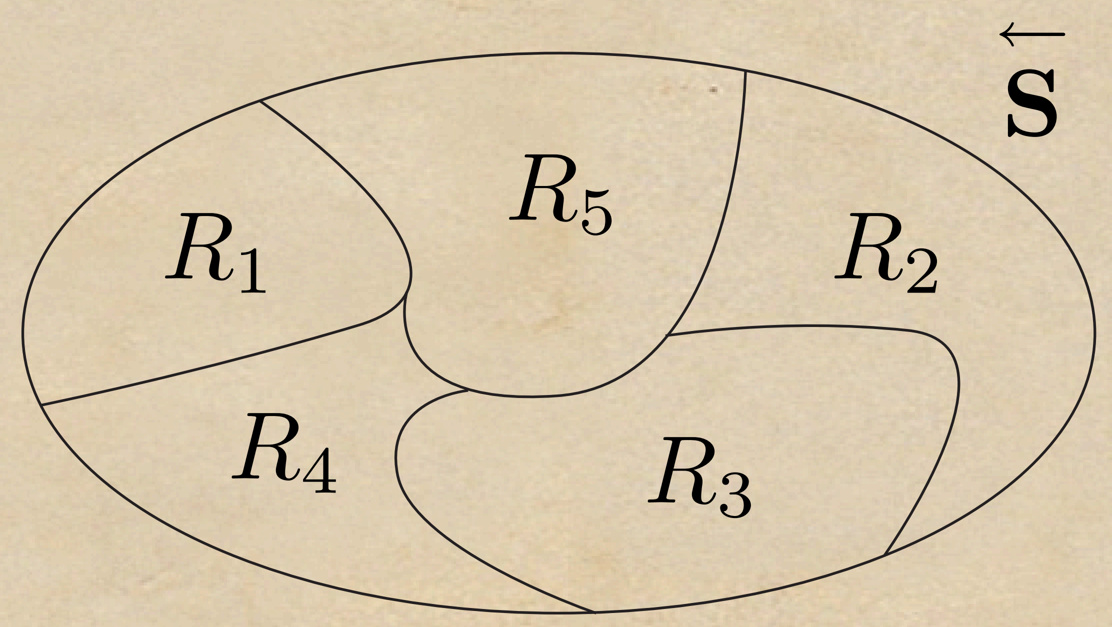
Grouping Predictively Equivalent Histories

Space of Histories:

Histories leading to the same predictions are equivalent.

Effective States = Partitions of History:

$$R = \{R_i : R_i \cap R_j = \emptyset, \overleftarrow{\mathbf{S}} = \bigcup_i R_i\}$$



How Effective are the Effective States?

Effective Prediction Error:

$$H[\overrightarrow{S}^L | R]$$

Uncertainty about future given effective states

Effective Prediction Error Rate:

$$h_\mu(R) = \lim_{L \to \infty} \frac{H[\overrightarrow{S}^L | R]}{L}$$

Entropy rate given effective states

Bounds:

$$h_\mu(R) \leq \log_2 |\mathcal{A}|$$

$$h_\mu(R_\emptyset) = \log_2 |\mathcal{A}|$$

How Effective are the Effective States ...

Prescience:

$$\Pi(R) = \log_2 |\mathcal{A}| - h_\mu(R)$$

Bad model says nothing about process:

$$\Pi(R_\emptyset) = 0$$

Upper bounded by Total Predictability:

$$\Pi(R) \leq |\mathbf{G}|$$

Find states R such that $h_\mu(R) = h_\mu$.

How Effective are the Effective States?

Statistical Complexity of the Effective States:

$$C_\mu(R) = H[R] = H(\Pr(R))$$

Interpretations:

Uncertainty in state.

Shannon information one gains when told effective state.

Model "size" $\propto \log_2(\text{number of states})$

Historical memory used by R.

Limits on Prediction:

Effective Prediction Error: $H[\overrightarrow{S}^L | R]$

$$H[\overrightarrow{S}^L | R] = H[\overrightarrow{S}^L | \eta(\overleftarrow{S})]$$
$$\geq H[\overrightarrow{S}^L \mid \overleftarrow{S}]$$

(Data Processing Inequality)

Models can do no better than to use histories.

That is, $h_\mu(R) \geq h_\mu$.

Goals Restated:

Question 1:

Can we find effective states that give good predictions?

$$H[\overrightarrow{S}^L | R] = H[\overrightarrow{S}^L \mid \overleftarrow{S}]$$

or

$$h_\mu(R) = h_\mu$$

Question 2:

Can we find the smallest such set?

$$\min C_\mu(R)$$

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:

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

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

Causal States ...

Causal State = Pasts with same morph: $\Pr(\overrightarrow{S} \mid \overleftarrow{s})$

$$\mathcal{S} = \{\overleftarrow{s}' : \overleftarrow{s}' \sim \overleftarrow{s}\}$$

Set of causal states:

$$\boldsymbol{\mathcal{S}} = \overleftarrow{\mathbf{S}} / \sim \quad = \{\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2, \ldots\}$$

Partition of histories:

$$\overleftarrow{\mathbf{S}} = \bigcup_i \mathcal{S}_i$$

$$\mathcal{S}_i \bigcap \mathcal{S}_j = \emptyset, i \neq j$$

Causal States ...

Causal state map:

$$\epsilon : \overleftarrow{\mathbf{S}} \rightarrow \boldsymbol{\mathcal{S}}$$

$$\epsilon(\overleftarrow{s}) = \{\overleftarrow{s}' : \overleftarrow{s}' \sim \overleftarrow{s}\}$$

Random variable:

$$\mathcal{S} = \epsilon(\overleftarrow{S})$$

Causal States ...

We've answered the first part of the modeling goal:

We have the effective states!

Now,

What is the dynamic?

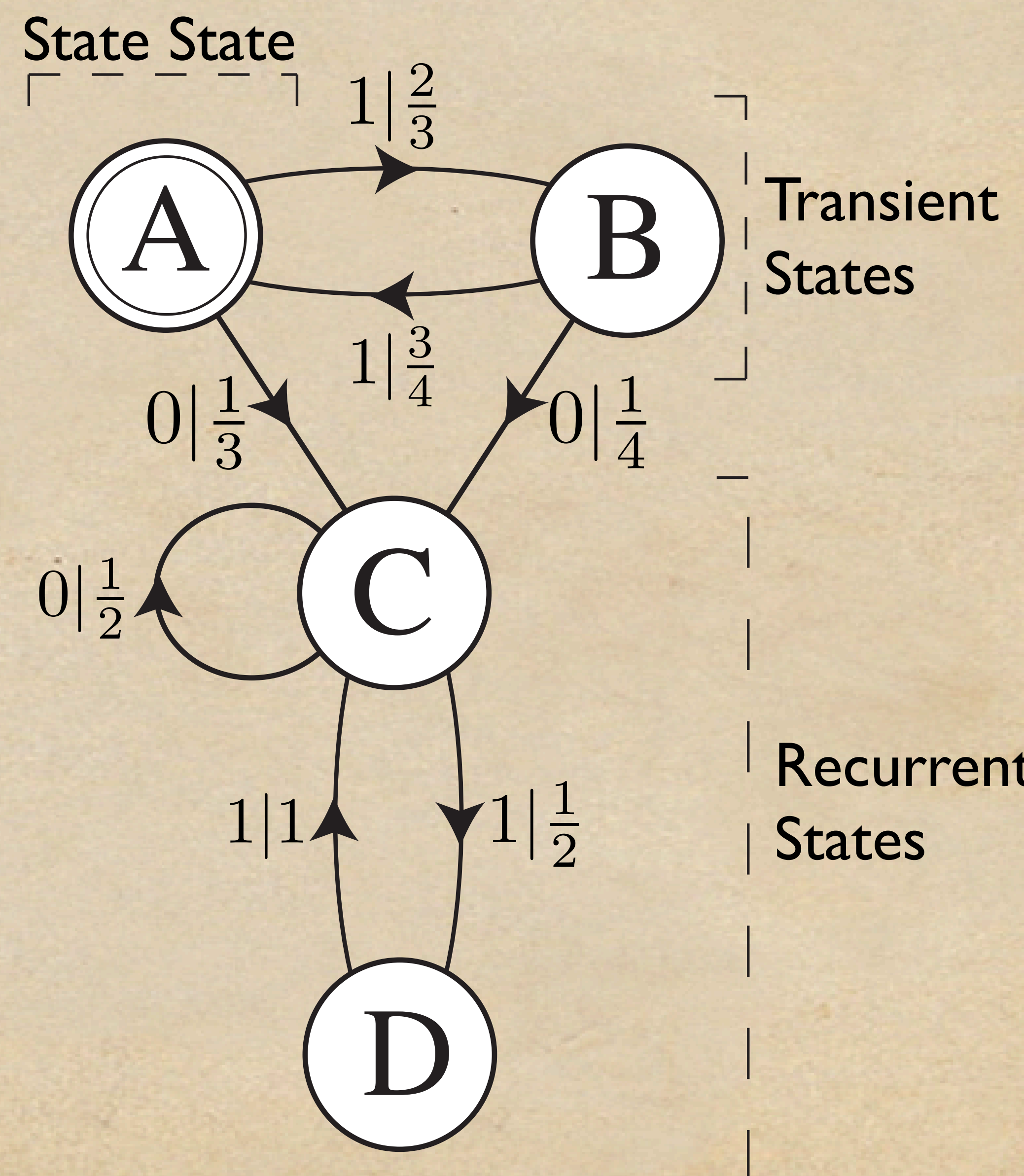
Not enough time ... see [CMPPSS] or [NCASO].

The ϵ-Machine of a Process ...

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

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

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



Unique Start State:

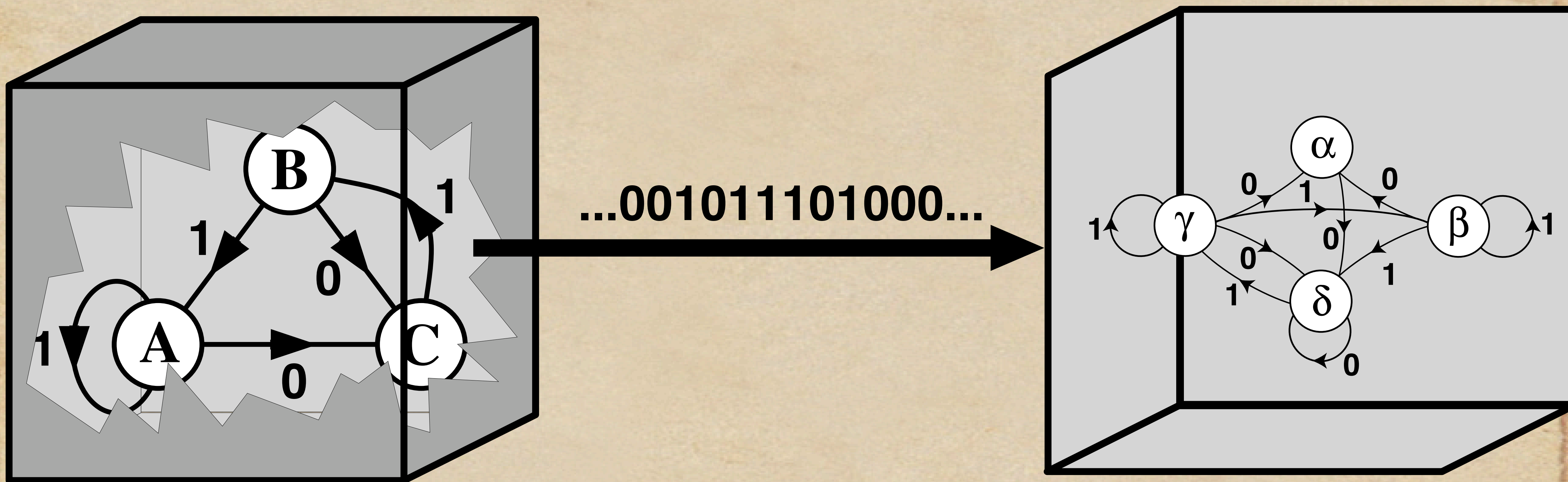
$$\mathcal{S}_0 = [\lambda]$$

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

Transient States

Recurrent States

The Learning Channel



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

Central questions:

What are the states? Causal States

What is the dynamic? The ϵ-Machine

A Model of a Process $\Pr(\overleftrightarrow{S})$:

ϵ-Machine reproduces the process's word distribution:

$$\Pr(s^1), \Pr(s^2), \Pr(s^3), \ldots$$

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

$$\begin{aligned} \Pr(s^L) &= \Pr(\mathcal{S}_0)\Pr(\mathcal{S}_0 \rightarrow_{s=s_1} \mathcal{S}(1))\Pr(\mathcal{S}(1) \rightarrow_{s=s_2} \mathcal{S}(2)) \\ &\quad \cdots \Pr(\mathcal{S}(L-1) \rightarrow_{s=s_L} \mathcal{S}(L)) \end{aligned}$$

Initially, $\Pr(\mathcal{S}_0) = 1$.

$$\Pr(s^L) = \prod_{l=1}^{L} T^{(s_l)}_{i=\epsilon(s^{l-1}), j=\epsilon(s^l)}$$

Past and Future are Independent given Causal State:

Process: $\Pr(\overleftrightarrow{S}) = \Pr(\overleftarrow{S}\overrightarrow{S})$

$$\Pr(\overleftarrow{S}\overrightarrow{S} \,|\mathcal{S}) = \Pr(\overleftarrow{S} \,|\mathcal{S}) \Pr(\overrightarrow{S} \,|\mathcal{S})$$

Causal states shield past & future from each other.

Similar to states of a Markov chain, but for hidden processes.

ϵMs are Optimal Predictors:

Compared to any rival effective states R:

$$H\left[\overrightarrow{S}^L | R\right] \geq H\left[\overrightarrow{S}^L | \mathcal{S}\right]$$

Proof sketch:
$$H\left[\overrightarrow{S}^L | \mathcal{S}\right] = H\left[\overrightarrow{S}^L | \overleftarrow{s} \in \mathcal{S}\right]$$
$$= H\left[\overrightarrow{S}^L | \overleftarrow{s}\right]$$
$$\leq H\left[\overrightarrow{S}^L | R\right]$$

$$R = \eta(\overleftarrow{s})$$

(Data processing inequality)

ϵMs are Optimal Predictors ...

Corollary:

$$h_\mu(\mathcal{S}) = h_\mu$$

Proof:

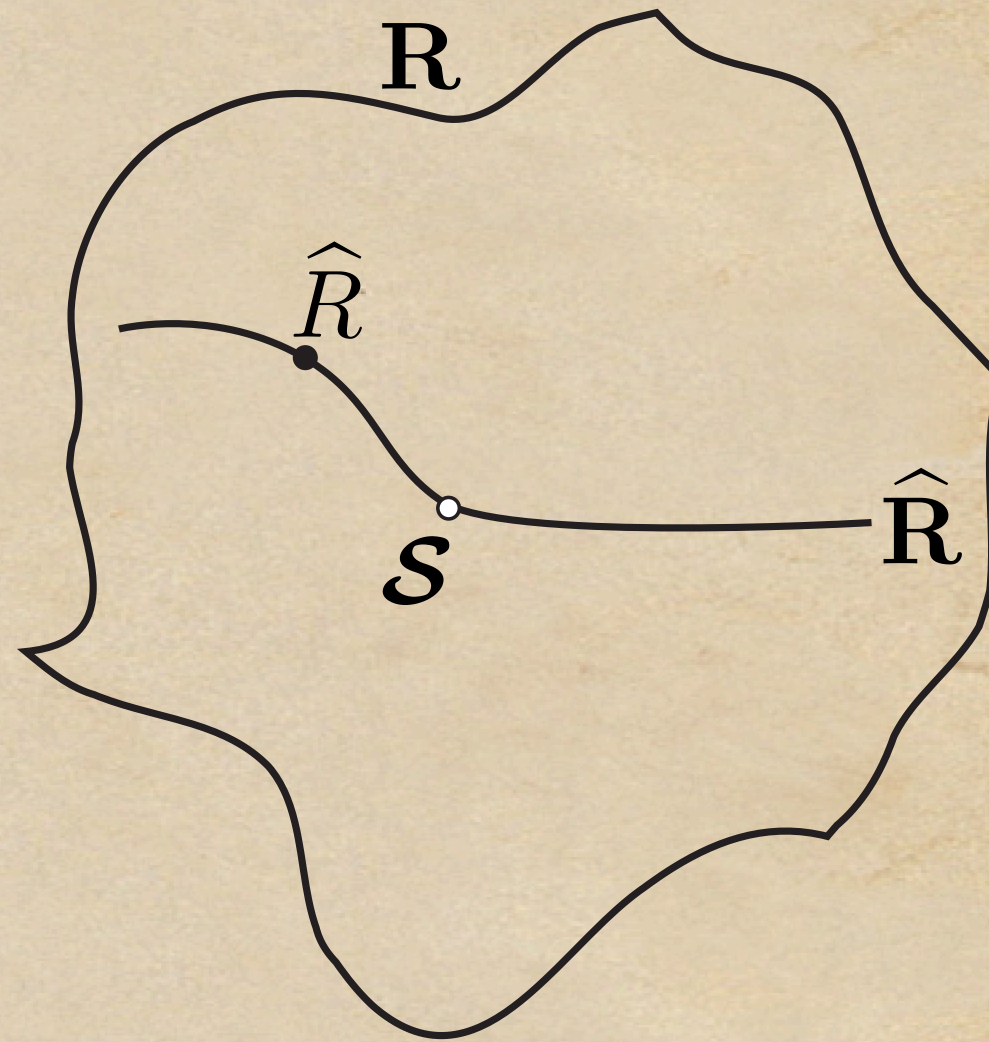
$$h_\mu(\mathcal{S}) = \lim_{L \to \infty} L^{-1} H[\overrightarrow{S}^L | \mathcal{S}] = \lim_{L \to \infty} L^{-1} H[\overrightarrow{S}^L | \overleftarrow{s}] = h_\mu$$

Prescient Rivals $\widehat{\mathbf{R}}$:

Alternative models that are optimal predictors

$\widehat{R} \in \widehat{\mathbf{R}}$

$$H[\overrightarrow{S}^L | \widehat{R}] = H[\overrightarrow{S}^L | \mathcal{S}]$$



(Prescient rivals are sufficient statistics for process's future.)

Minimal Statistical Complexity:

For all prescient rivals, $\epsilon\mathrm{M}$ is the smallest:

$$C_\mu(\widehat{R}) \geq C_\mu(\mathcal{S})$$

Proof sketch:

(1) Prescient rivals are refinements, so

$$\exists g : \mathcal{S} = g(\widehat{R})$$

(2) But

$$H[f(X)] \leq H[X] \Rightarrow H[\mathcal{S}] = H[g(\widehat{R})] \leq H[\widehat{R}]$$

(3) So $C_\mu \leq H[\widehat{R}]$

Minimal Statistical Complexity ...

Consequence:

(1) C_μ measures historical information process stores.

(2) This would not be true, if not minimal representation.

$\epsilon\mathrm{M}$ Summary:

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

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

(3) Unique: Smallest optimal predictor is equivalent.

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

(5) Causal Shielding:

States renders process's future independent of its past.

(6) "Deterministic": Sequence ~ one internal state path

Measures of Structural Complexity

Info. Theory Measures		Interpretation
Entropy Rate	h_μ	Intrinsic Randomness
Excess Entropy	$\mathbf{E}$	Info: Past to Future
Predictability Gain	$\mathbf{G}$	Redundancy
Transient Information	$\mathbf{T}$	Synchronization

How related to statistical complexity?

How to get from $\epsilon\mathrm{M}$?

Measures from the $\epsilon\mathrm{M}$:

Entropy Rate given $\epsilon\mathrm{M}$:

$$h_\mu(\boldsymbol{\mathcal{S}}) = - \sum_{\mathcal{S} \in \boldsymbol{\mathcal{S}}} \Pr(\mathcal{S}) \sum_{s \in \mathcal{A}, \mathcal{S}' \in \boldsymbol{\mathcal{S}}} T^{(s)}_{\mathcal{S}\mathcal{S}'} \log_2 T^{(s)}_{\mathcal{S}\mathcal{S}'}$$

where $\Pr(\mathcal{S})$ is casual-state asymptotic probability.

Possible only due to $\epsilon\mathrm{M}$ determinism/unifiliarity!

1-1 mapping between measurement sequences
& internal paths.

Entropy rate ...

Lesson: Need $\epsilon\mathrm{M}$ to calculate entropy rate.

Typical case example: Nontrivial, infinite $\epsilon\mathrm{M}$.

Curious:

Even to estimate a process's intrinsic randomness,

need to infer its structure.

Statistical Complexity of a Process:

$$C_\mu(\boldsymbol{\mathcal{S}}) = - \sum_{\mathcal{S} \in \boldsymbol{\mathcal{S}}} \Pr(\mathcal{S}) \log_2 \Pr(\mathcal{S})$$

where $\Pr(\mathcal{S})$ is casual-state asymptotic probability.

Meaning:

Shannon information in the causal states.

Amount of historical information a process stores.

Amount of structure in a process.

Bound on Excess Entropy:

$$\mathbf{E} \leq C_\mu$$

Proof sketch:

(1) $\mathbf{E} = I[\overrightarrow{S}; \overleftarrow{S}] = H[\overrightarrow{S}] - H[\overrightarrow{S} \mid \overleftarrow{S}]$

(2) Causal States: $H[\overrightarrow{S} \mid \overleftarrow{S}] = H[\overrightarrow{S} \mid \mathcal{S}]$

(3) $\mathbf{E} = H[\overrightarrow{S}] - H[\overrightarrow{S} \mid \mathcal{S}]$

$$= I[\overrightarrow{S}; \mathcal{S}]$$

$$= H[\mathcal{S}] - H[\mathcal{S} \mid \overrightarrow{S}]$$

$$\leq H[\mathcal{S}] = C_\mu$$

Bound on Excess Entropy ...

Consequence:

Possible for $\mathbf{E} \to 0$ when $C_\mu \gg 1$. (Cryptographic limit)

Excess entropy is *not* the process's stored information.

$\mathbf{E}$ is the *apparent* information,
as revealed in *measurement sequences*.

Statistical complexity *is* stored information.

Bound on Excess Entropy ...

Executive Summary:

C_μ is the amount of information the process requires

to communicate

$\mathbf{E}$ bits of information from the past to the future.

Bound on Excess Entropy: $\mathbf{E} \leq C_\mu$

Consequence:

The inequality is Why We Must Model.

Cannot simply use sequences as states.

There is internal structure not expressed by this.

Dynamical system's intrinsic computation:

(1) How much of past does process store?

(2) In what architecture is that information stored?

(3) How does stored information produce future behavior?

The Point

How nature is structured

is how nature computes.

Structure or Noise?

- Just completed: “In Principle”
- Now: “Practically”

Passive Learning

- Problem: Experiment to Learn World Model
  - The world behaves: $\overleftrightarrow{X} = \underset{\text{past}}{\overleftarrow{X}} \; \underset{\text{future}}{\overrightarrow{X}}$
  - Agent learns model of the world: States $\mathcal{R}$

Passively Learning a Model

- Pattern discovery:

  Learn the world's hidden states $\Pr(\mathcal{R}|\overleftarrow{X})$

- Causal shielding:

$$\Pr(\overleftarrow{X}\overrightarrow{X}) = \Pr(\overleftarrow{X}|\mathcal{R})\Pr(\overrightarrow{X}|\mathcal{R})$$

- Dynamics of learning:

  Search in the space of models: $\mathcal{R} \in \mathcal{M}$

Passively Learning a Model

- Causal shielding objective function

$$\min_{\Pr(\mathcal{R}|\overleftarrow{X})} \left(I[\overleftarrow{X};\mathcal{R}] + \beta I[\overleftarrow{X};\overrightarrow{X}\,|\mathcal{R}] \right)$$

Model: Map from histories to states	Info states contain about histories	Reduce info history has about future

$$\beta \sim 1/T$$

Passively Learning a Model

- Optimal states $\Pr(\mathcal{R}|\overleftarrow{X})$ are Gibbs states:

$$\Pr_{\text{opt}}(\mathcal{R}|\overleftarrow{X}) = \frac{\Pr(\mathcal{R})}{Z(\overleftarrow{X},\beta)} e^{-\beta E(\mathcal{R},\overleftarrow{X})}$$

where

$$E(\mathcal{R},\overleftarrow{X}) = \mathcal{D}\left(\Pr(\overrightarrow{X} \mid \overleftarrow{X})||\Pr(\overrightarrow{X}|\mathcal{R})\right)$$

$$\Pr(\overrightarrow{X}|\mathcal{R}) = \frac{1}{\Pr(\mathcal{R})}\sum_{\overleftarrow{X}}\Pr(\overrightarrow{X} \mid \overleftarrow{X})\Pr(\mathcal{R}|\overleftarrow{X})\Pr(\overleftarrow{X})$$

$$\Pr(\mathcal{R}) = \sum_{\overleftarrow{X}}\Pr(\mathcal{R}|\overleftarrow{X})\Pr(\overleftarrow{X})$$

Passively Learning a Model

- Solve these equations self-consistently

  (Analytical in special cases; numerical generally)

- Parametrized family of models:

  $$R_\beta \colon \Pr(\mathcal{R} | \overleftarrow{X})$$

- Structure or Noise?

  β trades-off model size against prediction error

What Do Solutions Mean?

Causal Models

- Causal architecture given by ϵ-Machine M:
    - Optimal predictor:

$$h_\mu(M) \leq h_\mu(\mathcal{R})$$

    - Minimal size (within optimal predictors $\widehat{\mathcal{R}}$):

$$C_\mu(M) \leq C_\mu(\widehat{\mathcal{R}})$$

    - Unique (within min, opt predictors)

JPC & K. Young, Inferring Statistical Complexity, Physical Review Letters 63 (1989) 105-108.
C. R. Shalizi & JPC, Journal Statistical Physics 104 (2001) 817-879.

Passively Learning a Model

- Theorem: Low-temperature limit

$$\beta \to \infty$$

Recover ϵ-Machine:

$$R_\beta \to M$$

- Conclusion:

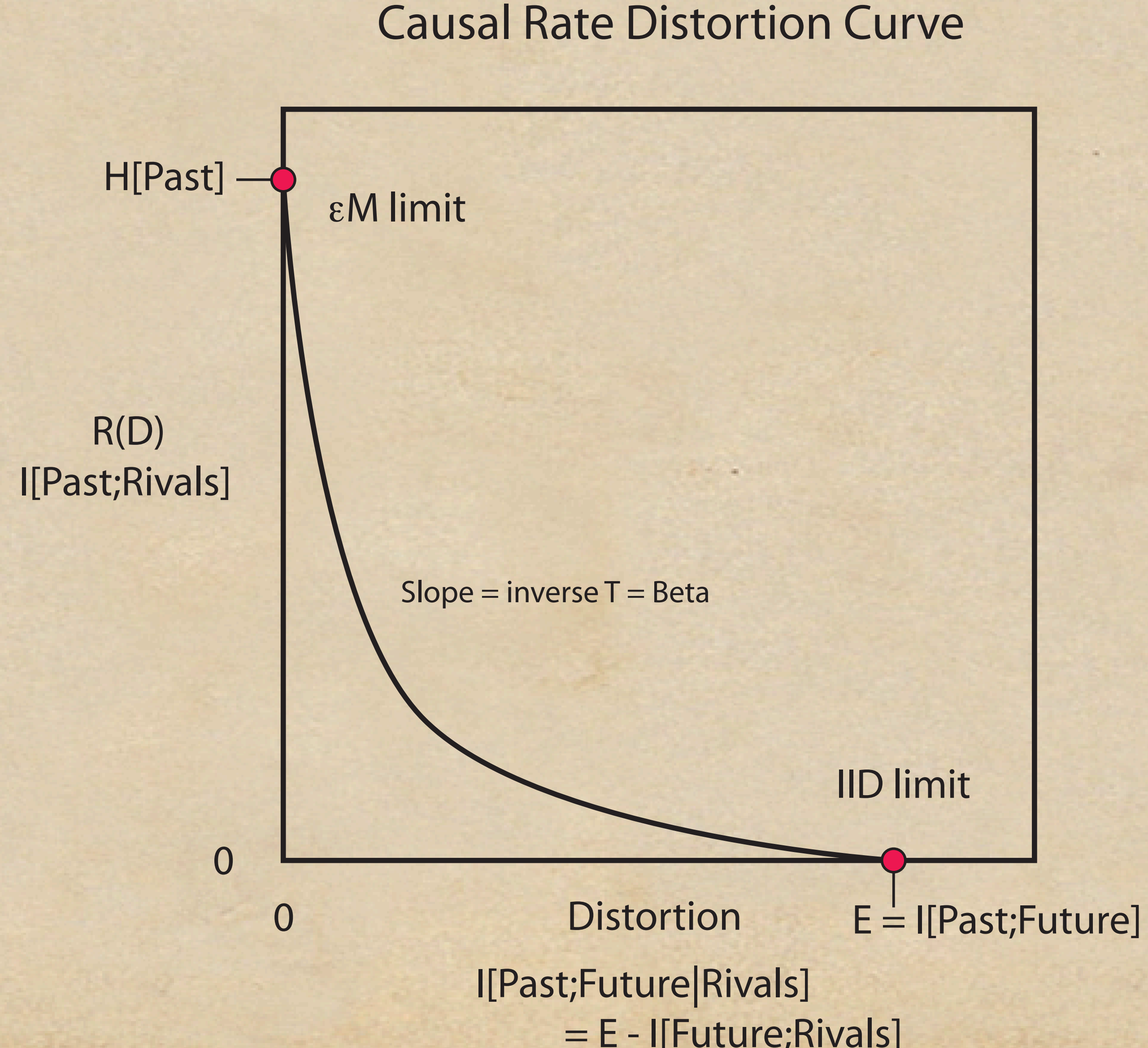
At given prediction error

R_β is best causal approximate.

Passively Learning a Model

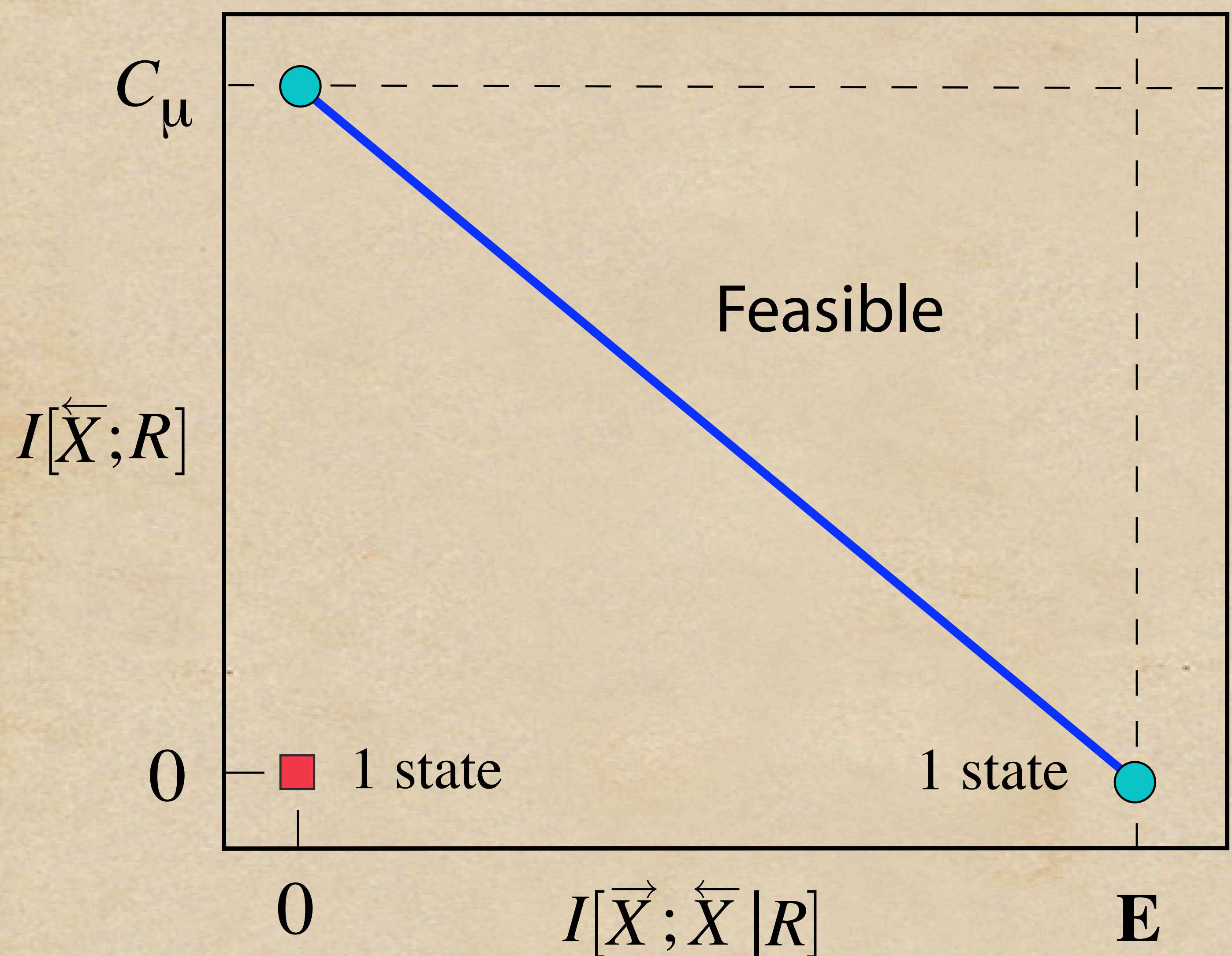
Optimally balance structure & error
At each level β of approximation

In theory



Passively Learning a Model

Analytical cases



Predictively Reversible:

$$\mathrm{P}(\overrightarrow{x} \mid \overleftarrow{x}) = \delta_{\overrightarrow{x},f(\overleftarrow{x})}$$

(e.g., periodic)

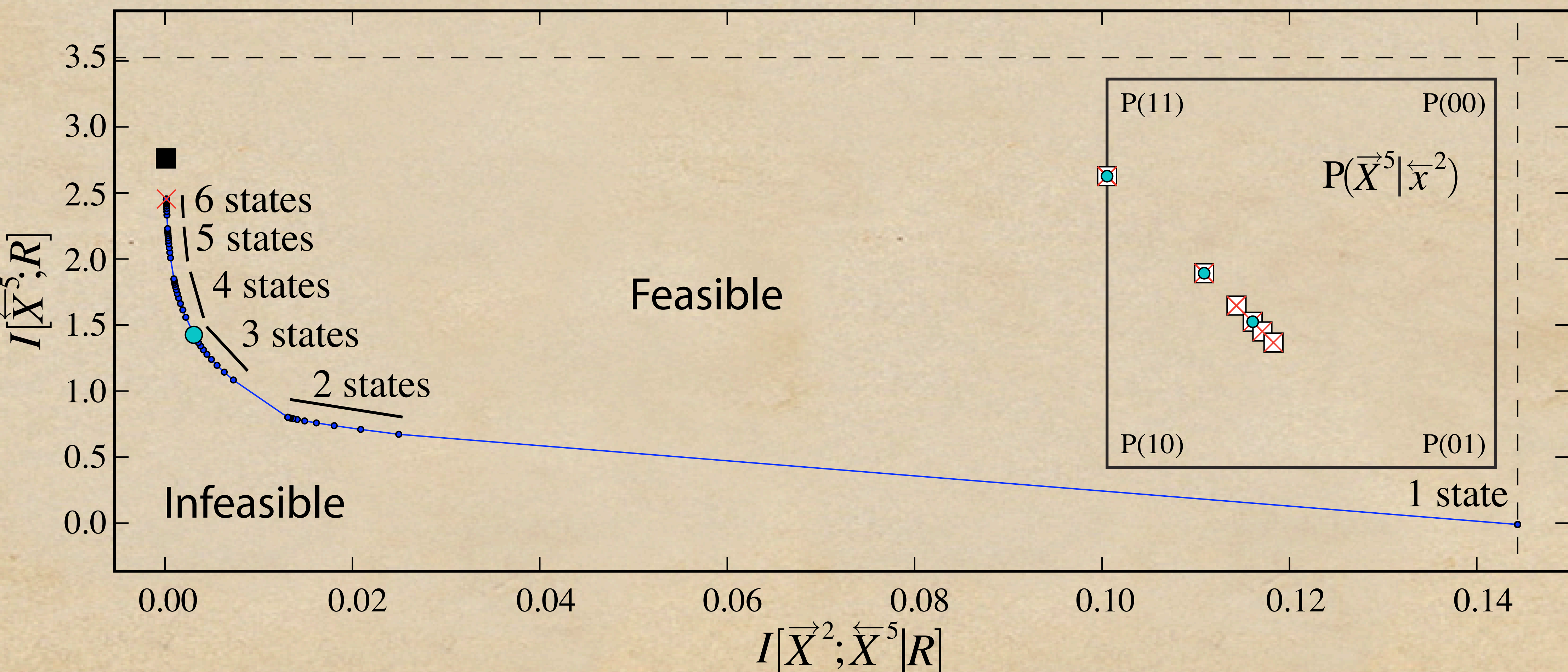
All IID processes:

$$\mathrm{P}(\overrightarrow{x} \mid \overleftarrow{x}) = \mathrm{P}(\overrightarrow{x})$$

Passively Learning a Model

Optimally balance structure & error
At each level β of approximation

In practice: Learn an oo-state world (SNS: simple nondeterminstic source)



Passively Learning a Model

- Causal compressibility: Shape of RD curve
  - Benefit of choosing smaller model for loss in predictability
  - Deviation from straight-line RD curve
- IID: No
- Predictively reversible: No
- SNS: Yes

Passively Learning a Model
(Conclusion)

- Causal shielding principle leads to
  - ϵ-Machine
  - Family of best approximations to ϵ-M

Passively Learning a Model
(related work)

- Computational mechanics:

  Focus on causal organization [Crutchfield et al 1989]
- Information bottleneck:

  Focus on prediction [Bialek et al 1997]
- Deterministic annealing:

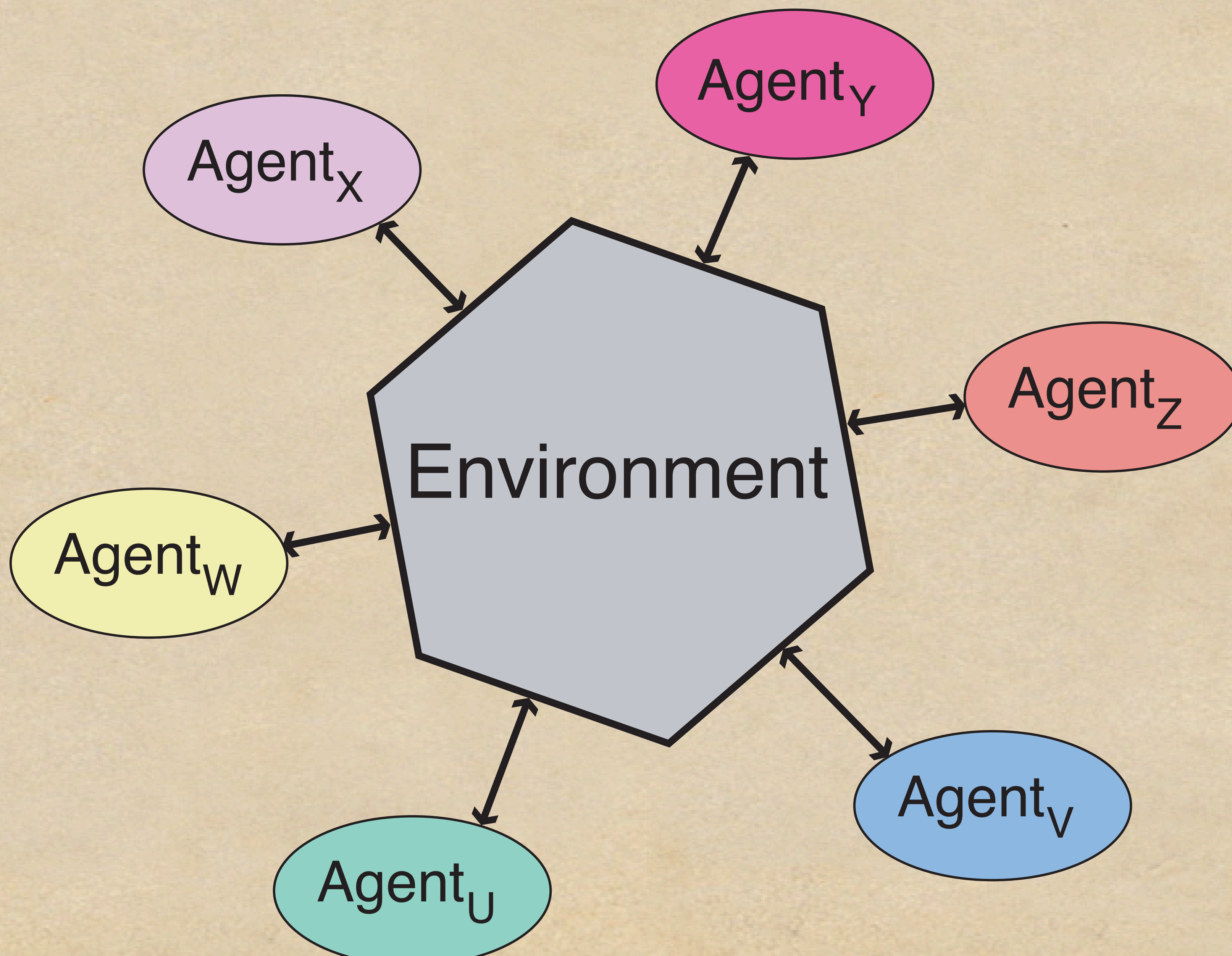
  Maximum entropy [Rose et al 1990]
- Goal-directed information processing:

  [Sanchez-Montañes et al 2004]

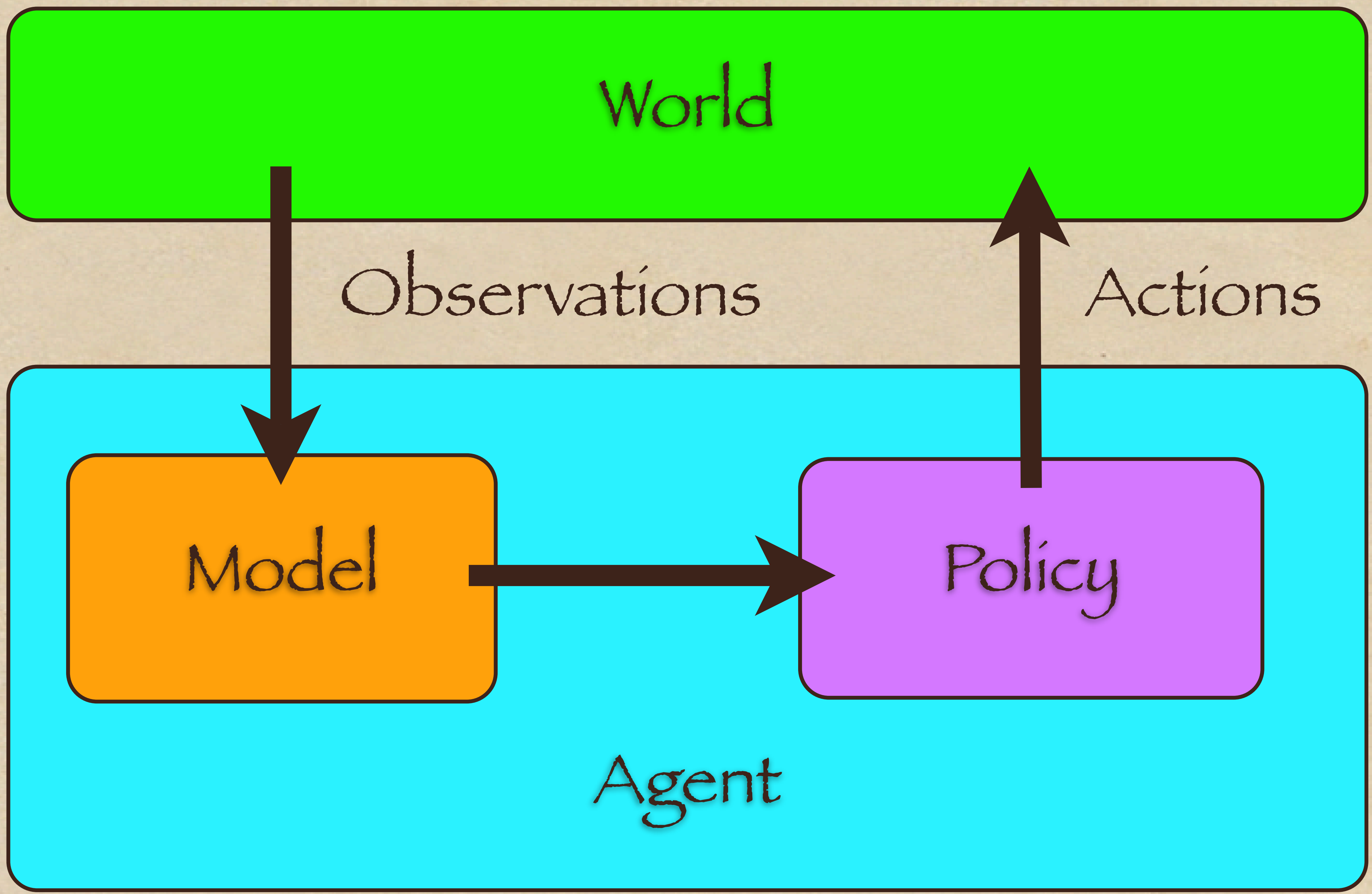
The Future?

Future:

Fundamental in Designing Multiagent Systems



The Feedback Loop



Knowledge + Action

- A central challenge:

  Actions change the world

    and so

    its statistics,

    and

    what is knowable.

Approaches

- Modeling:
  - Statistical inference
- Strategizing:
  - Game theory
- Adapting:
  - Reinforcement learning
- Group behavior:
  - Population dynamics (evolution & ecology)
- ...

Approaches: Sticking points

- Modeling:
  - Statistical inference: static, batch mode
- Strategizing:
  - Game theory: equilibria, no transients
- Adapting:
  - Reinforcement learning: a priori design, brittle
- Group behavior:
  - Population dynamics (evolution & ecology): individuals have no structure (don't learn)
- Where are the basic principles?

Interactive Learning

(Susanne Still, Chris Ellison, & JPC)

- Problem: Experiment to Learn World Model
  - The world behaves: $\overleftrightarrow{X} = \underset{\text{past}}{\overleftarrow{X}} \; \underset{\text{future}}{\overrightarrow{X}}$
  - Agent learns model of the world: States $\mathcal{R}$
  - Agent take actions $\mathcal{A}$
  - Those actions affect the world
  - Now the world is different!
- How to close the feedback loop?

arxiv.org: [0708.0654](0708.0654) [physics.gen-ph] & [0708.1580](0708.1580)[cs.IT]

Interactive Learning

- Decision: Using model, take actions
- Policy: $\Pr(\mathcal{A}|\overleftarrow{X})$ (or from $\mathcal{R}$)
- Experimentation objective function

$$\max_{\Pr(\mathcal{R}|\overleftarrow{X}),\Pr(\mathcal{A}|\overleftarrow{X})} \Big(I[\{\mathcal{R},\mathcal{A}\};\overrightarrow{X}] - \lambda I[\mathcal{R};\overleftarrow{X}] - \mu I[\mathcal{A};\overleftarrow{X}] \Big)$$

Model: Map from histories to states

Policy: Map from histories to actions

Info states/actions contain about futures

Info states contain about histories

Info actions contain about histories

Interactive Learning: Results

- Optimal model: Recover causal architecture
- Optimal policies
- Causally equivalent policies
- Curiosity: Take informative actions
- Control: Make world easier to model
- Balance of exploitation and control

arxiv.org: 0708.0654 [physics.gen-ph] & 0708.1580[cs.IT]

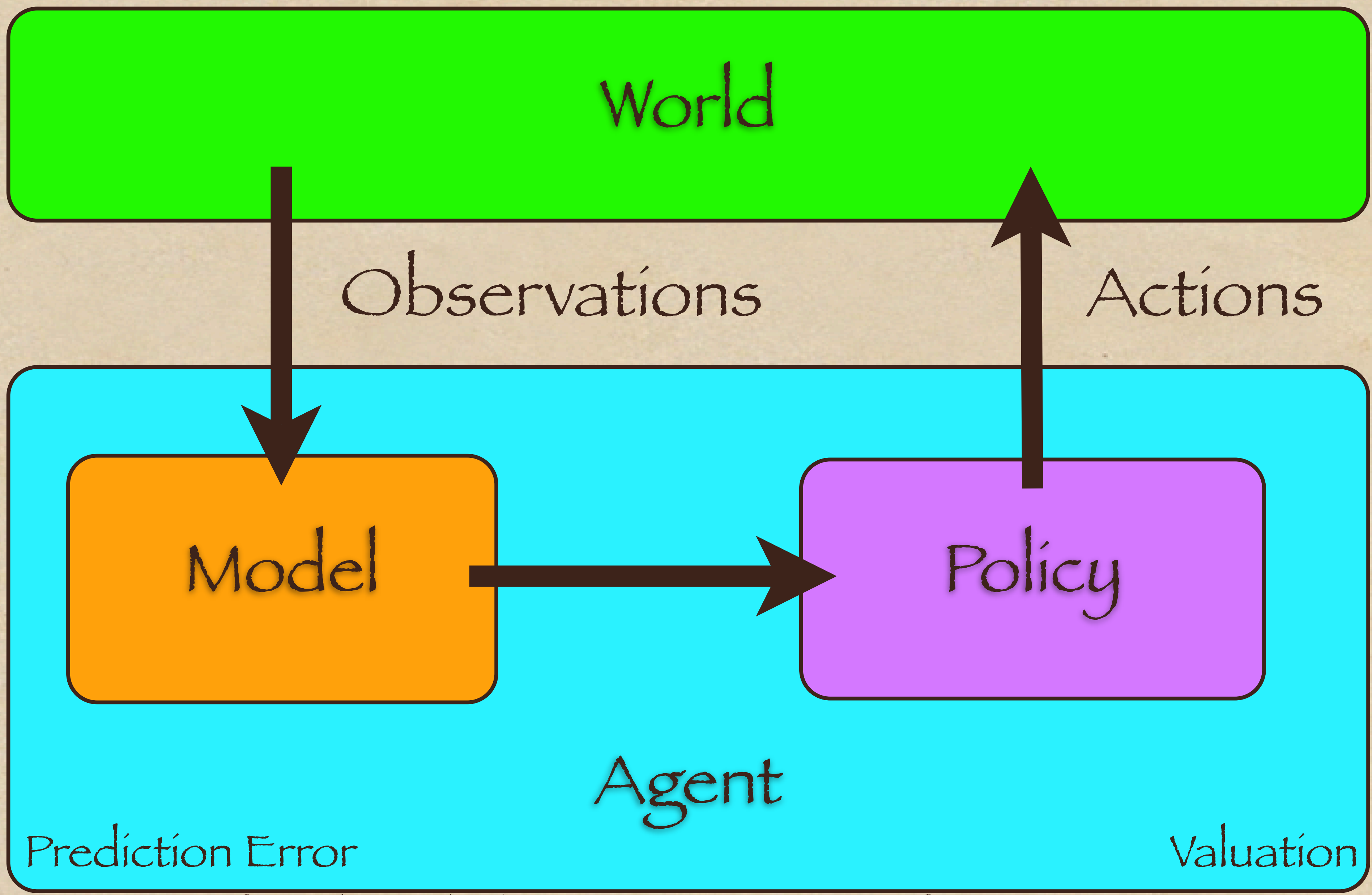
Connections

- Note iLearning subsumes:
  - Causal modeling
  - Game theory
  - Equilibrium economics
  - Reinforcement learning

Knowledge + Action

- Deviation from knowledge:
  - Model evaluation (e.g., prediction error)
- Valuation: Commitment of resources to action
  - Policy evaluation (e.g., average reward)
- How? Augment objective function
  - Change relative weighting of Lagrange multipliers: λ & μ
  - Add new terms: e.g., ...
- Examples:
  - "Science": Need accurate knowledge, at expense of producing it
  - "Politics": Need world to behave, independent of knowledge or cost
  - These are positions on causal rate-distortion curve

The Feedback Loop



Deviation from knowledge   Commitment of resources to actions

Main message

- Closing the loop:

  How interaction changes the world &

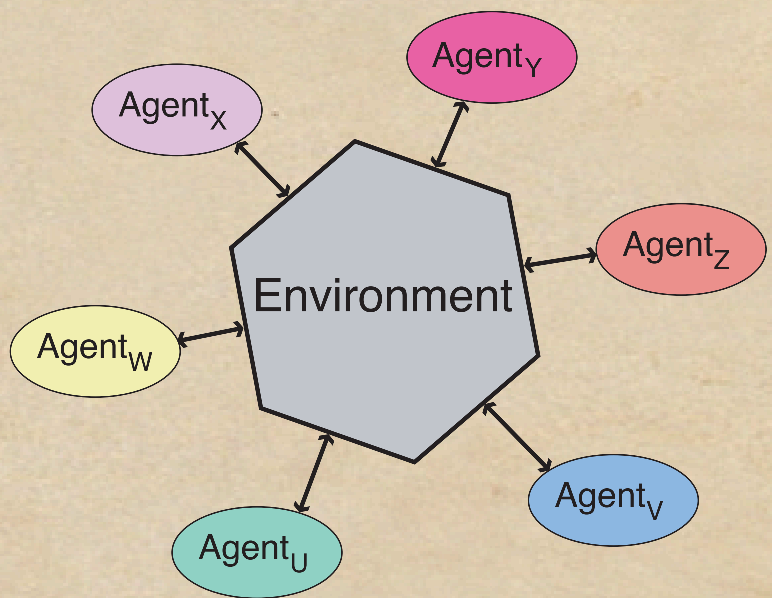
  how one adapts to those changes
- Theoretical foundations (& algorithms) for closing the feedback loop are now available.

Conclusion

- Basic principles follow from
  - Computational mechanics
  - Stat physics/Info theory (rate distortion)
- Balance structure & noise
- Balance exploitation & exploration
- Balance exploitation & control
- Challenge: Fold in risk

Prospects

- Collective Cognition:
  - Pattern discovery
  - Interactive learning
  - Adaptation dynamics
  - Emergent policy design
  - Multiagent dynamical systems



Summary

Why Build Models?

Kinds of Information

Causal Modeling

Intrinsic Computation

Balancing Noise & Structure

Interactive Learning

Multiagent Dynamical Systems

Thanks!