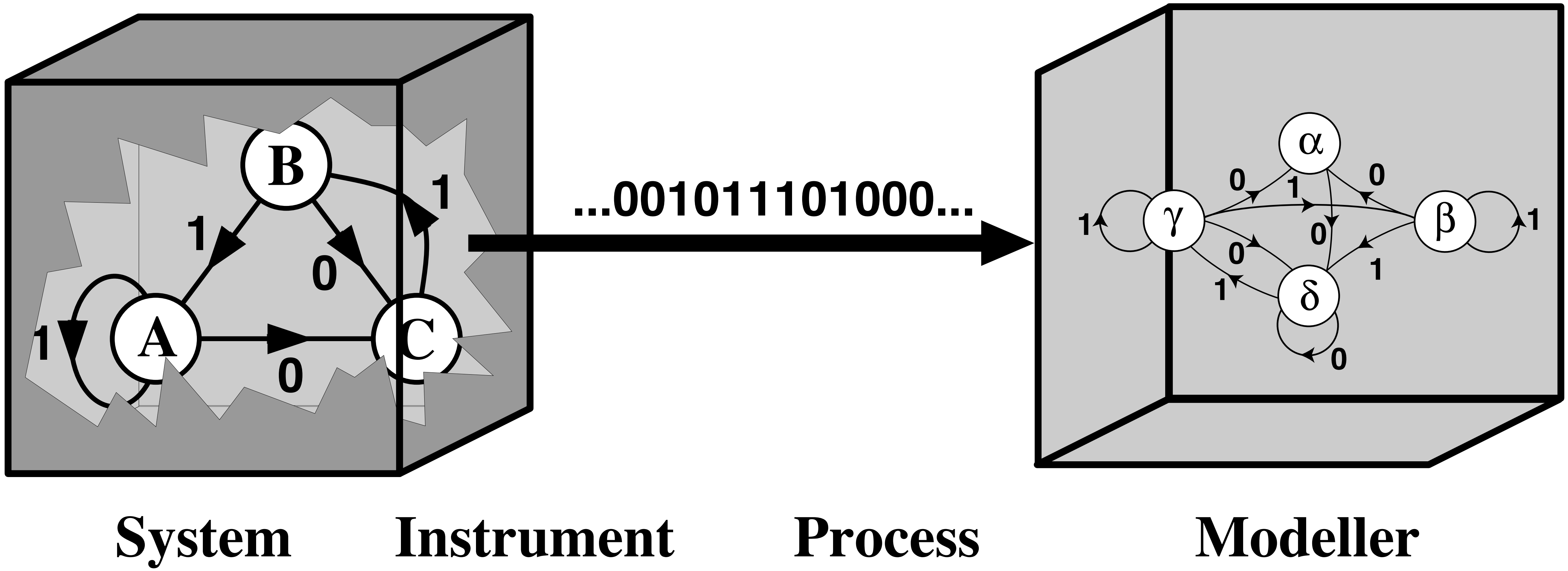
# The Learning Channel

# The Learning Channel:




Central questions:

What are the states?

What is the dynamic?

# The Learning Channel ...

## 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) describing the process.

# The Learning Channel ...

## The Prediction Game ...

### Process I:

# The Learning Channel ...

## The Prediction Game ...

### Process I:

Past: $\ldots 111111111111$

# The Learning Channel ...

## The Prediction Game ...

### Process I:

Past: $\ldots 111111111111$

Your prediction is?

# The Learning Channel ...

## The Prediction Game ...

### Process I:

Past: $\ldots 111111111111$

Your prediction is?

Future: $111111111111\ldots$

# The Learning Channel ...

## The Prediction Game ...

### Process I:

Past: $\ldots 111111111111$

Your prediction is?

Future: $111111111111\ldots$

Your model (states & dynamic) is?

## The Prediction Game ...

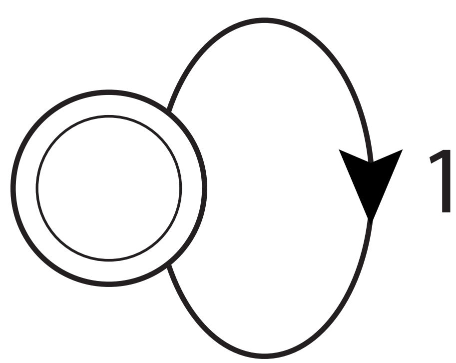
### Process I:

Past: $\ldots 111111111111$

Your prediction is?

Future: $111111111111\ldots$

Your model (states & dynamic) is?



# The Learning Channel ...

## The Prediction Game ...

### Process II:

# The Learning Channel ...

## The Prediction Game ...

### Process II:

Past: $\ldots 10110010001101110$

Complexity Lecture 2: Computational Mechanics (CSSS 2012); Jim Crutchfield

# The Learning Channel ...

## The Prediction Game ...

### Process II:

Past: $\ldots 10110010001101110$

Your prediction is?

# The Learning Channel ...

## The Prediction Game ...

### Process II:

Past: $\ldots 10111001000110110$

Your prediction is?

Analysis: All words of length L occur & equally often

## The Prediction Game ...

### Process II:

Past: $\ldots 1011001000110 1110$

Your prediction is?

Analysis: All words of length L occur & equally often

Future: Well, anything can happen, how about?

# The Learning Channel ...

## The Prediction Game ...

### Process II:

Past: $\ldots 1011100100011101110$

Your prediction is?

Analysis: All words of length L occur & equally often

Future: Well, anything can happen, how about?

$0101011101000110 1\ldots$

## The Prediction Game ...

### Process II:

Past: $\ldots 10111001000 1101110$

Your prediction is?

Analysis: All words of length L occur & equally often

Future: Well, anything can happen, how about?

$$0101011101000 1101\ldots$$

Your model is?

# The Learning Channel ...

## The Prediction Game ...

### Process II:

Past: $\ldots 10111001000110111 0$

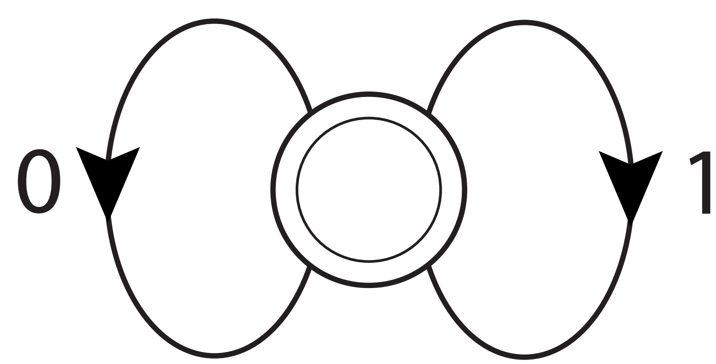
Your prediction is?

Analysis: All words of length L occur & equally often

Future: Well, anything can happen, how about?

$0101011101000110 1\ldots$

Your model is?



# The Learning Channel ...

## The Prediction Game ...

### Process III:

# The Learning Channel ...

## The Prediction Game ...

### Process III:

Past: $\ldots 1010101010101010$

# The Learning Channel ...

## The Prediction Game ...

### Process III:

Past: $\ldots 1010101010101010$

Your prediction is?

# The Learning Channel ...

## 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 Learning Channel ...

## The Prediction Game ...

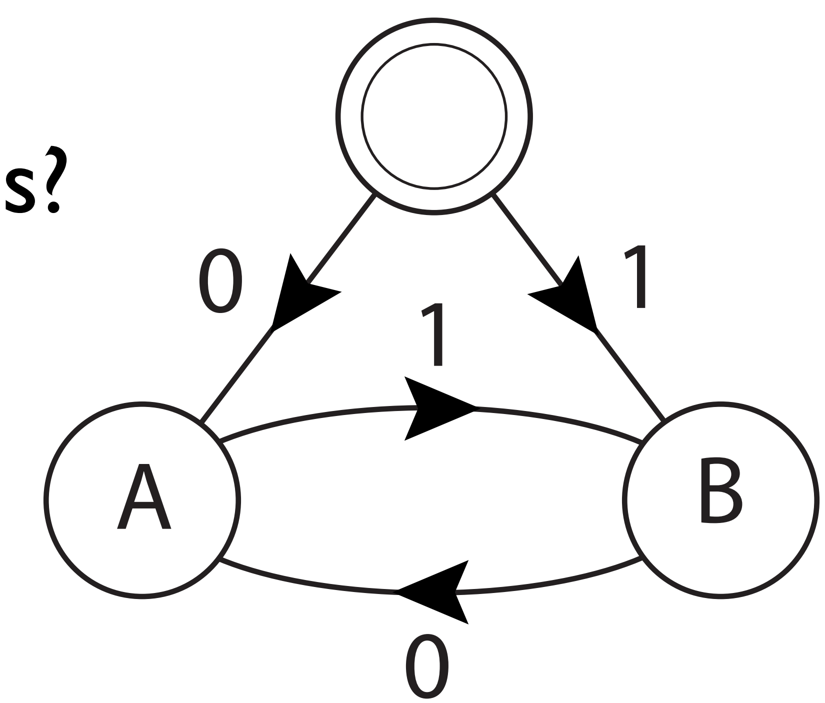
### Process III:

Past: $\ldots 1010101010101010$

Your prediction is?

Future: $1010101010101010\ldots$

Your model is?



# The Learning Channel ...

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 $\mathcal{A}$

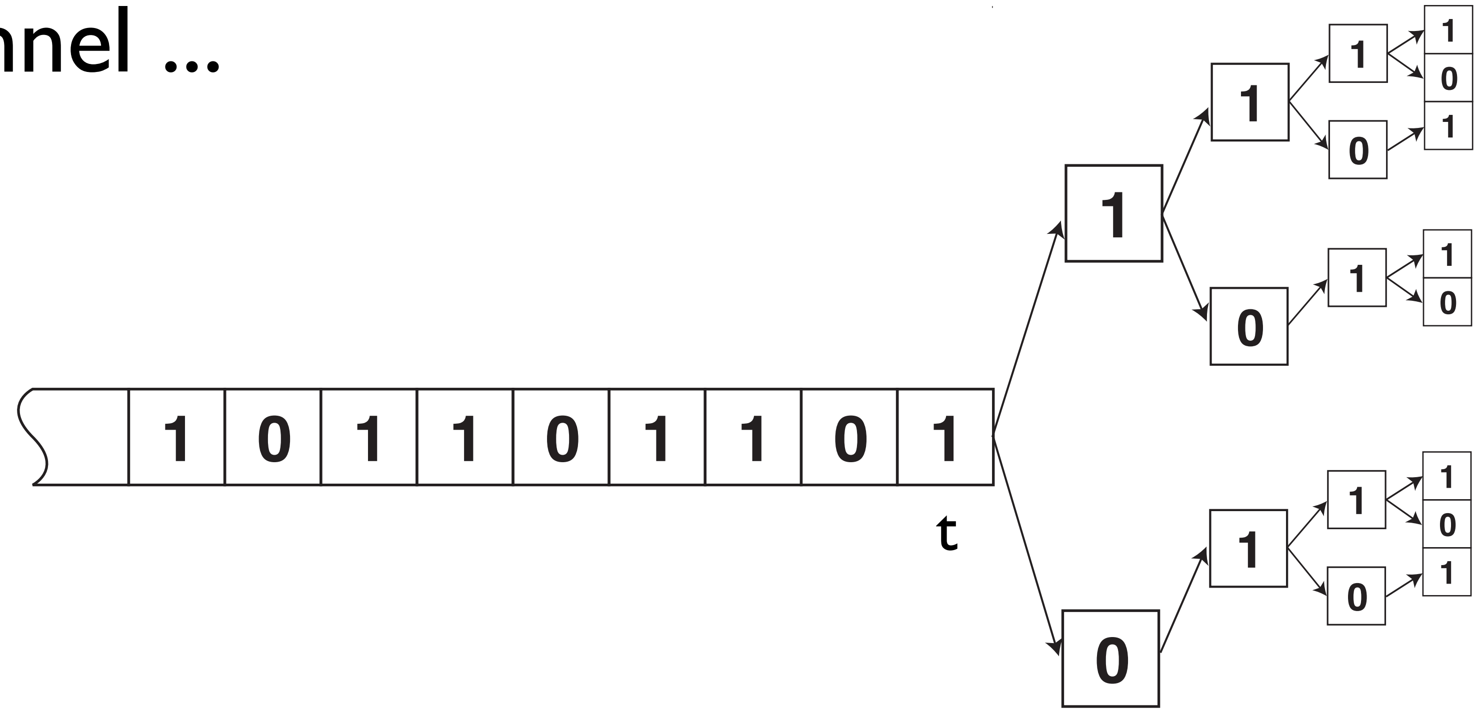
These symbols only indirectly reflect the hidden states

# The Learning Channel ...

## Effective States:

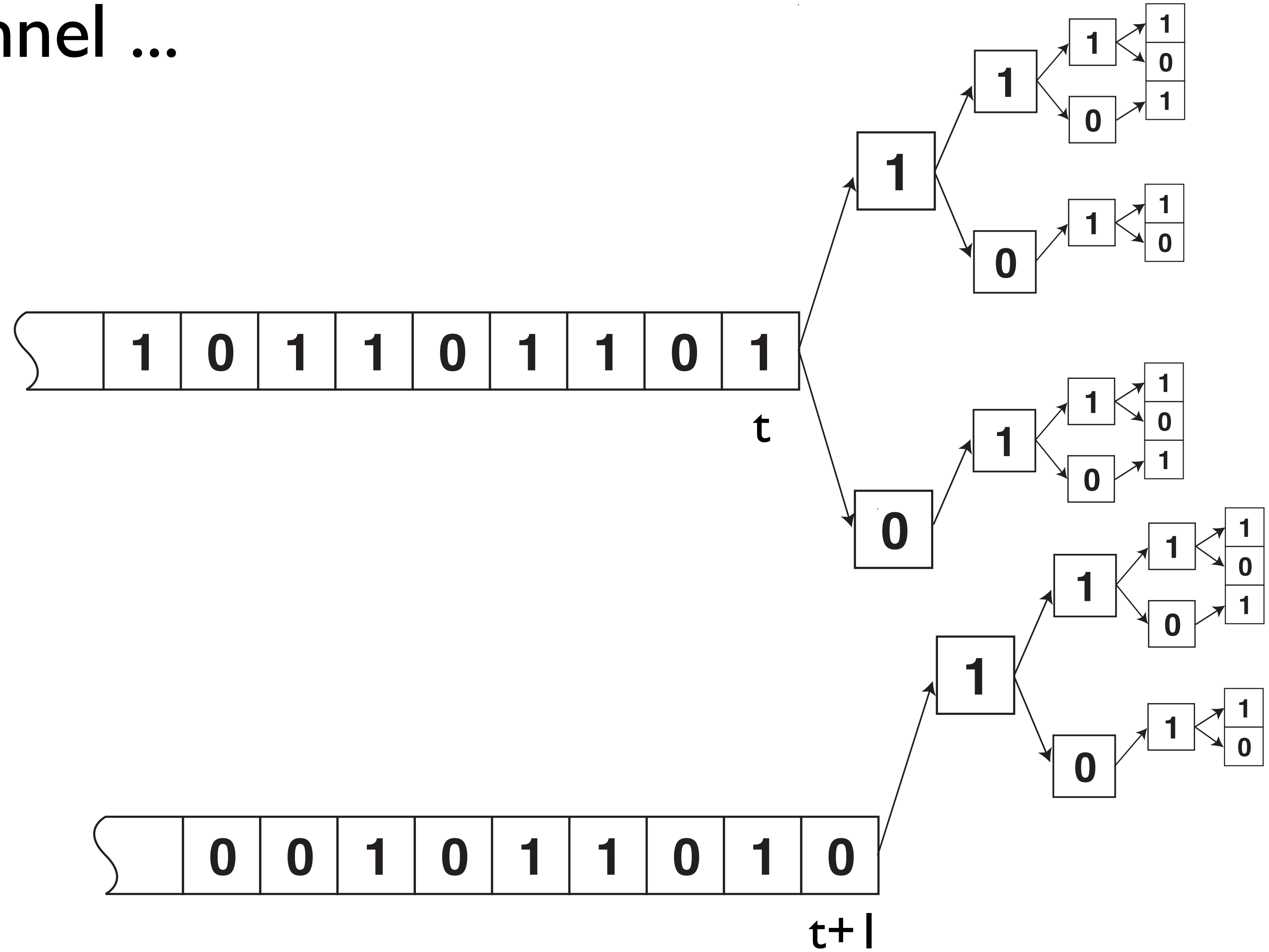
# The Learning Channel ...

## Effective States:



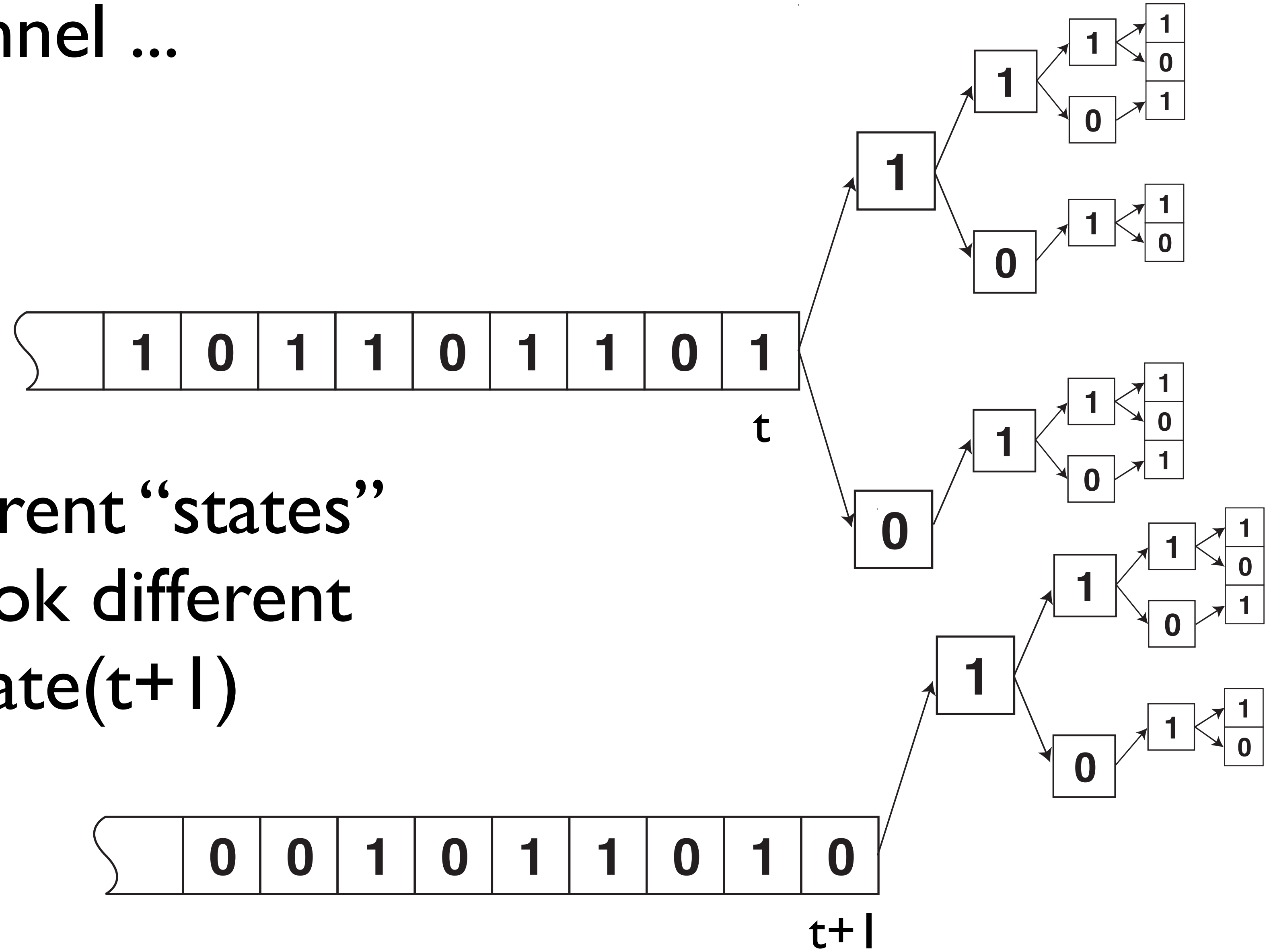
# The Learning Channel ...

## Effective States:



# The Learning Channel ...

## Effective States:

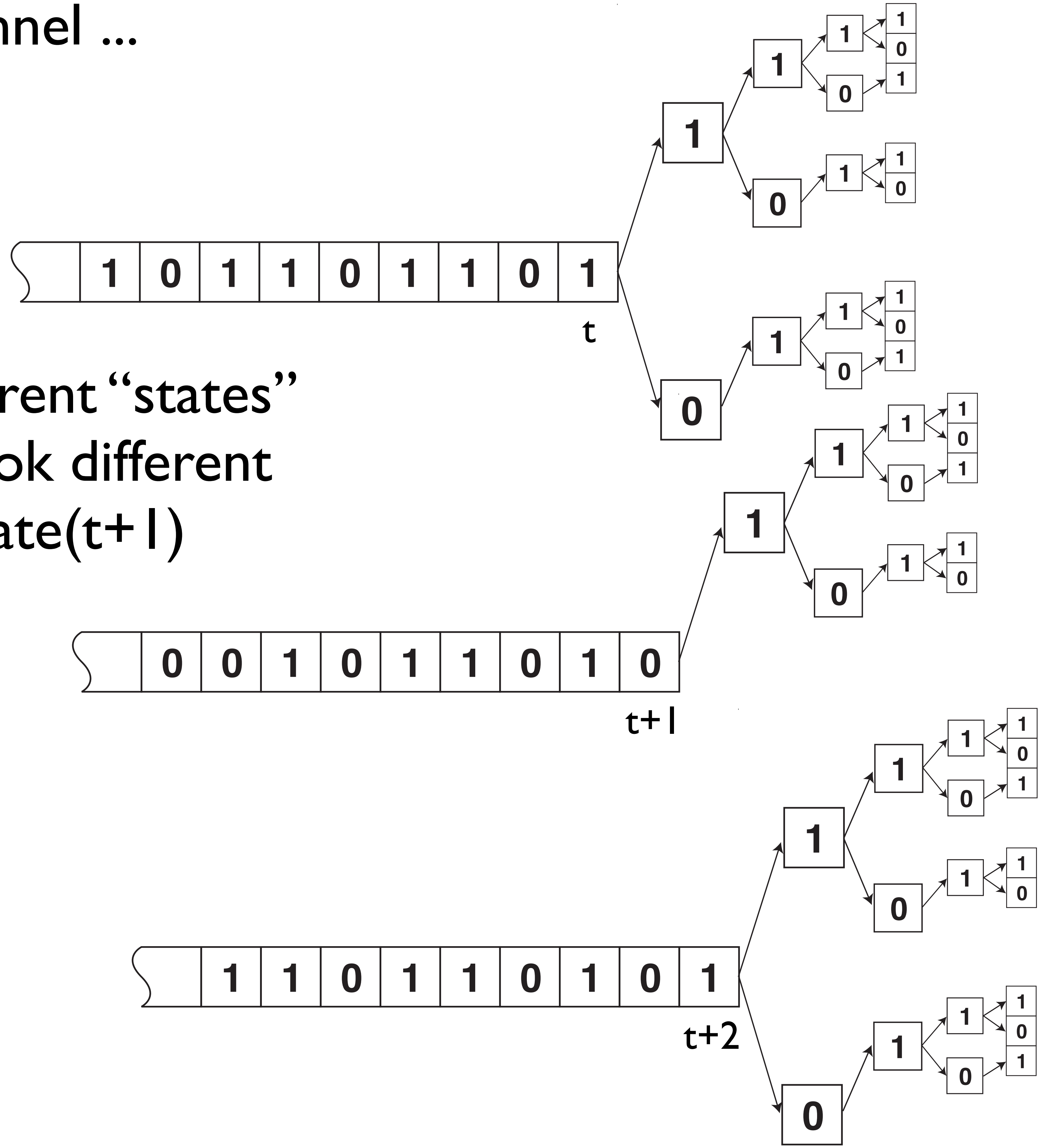



Process is in different "states"
when futures look different
State(t) $\nsim$ State(t+1)

Complexity Lecture 2: Computational Mechanics (CSSS 2012); Jim Crutchfield

# The Learning Channel ...

## Effective States:



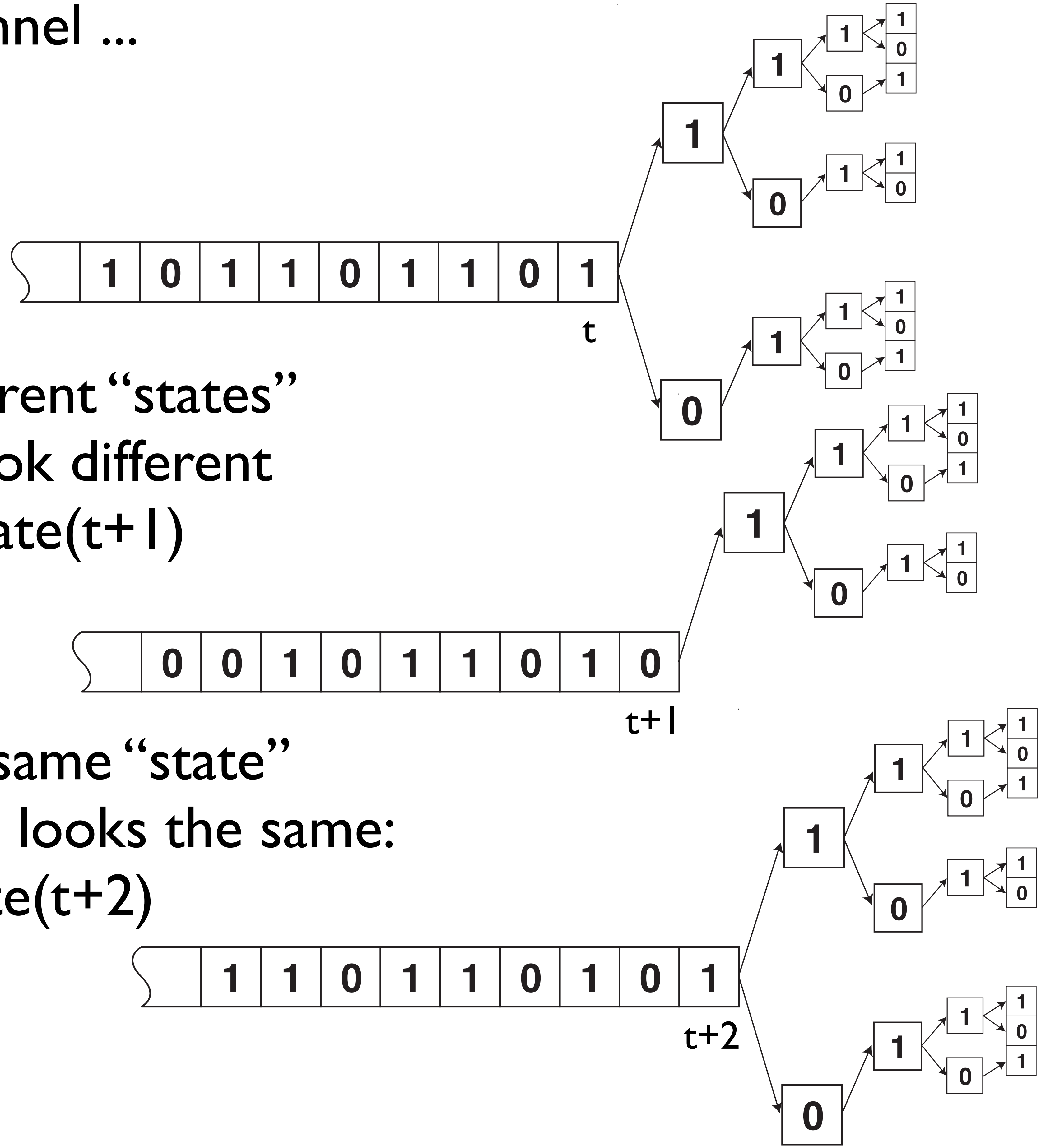

Process is in different "states"
when futures look different
State(t) $\sim$ State(t+1)

# The Learning Channel ...

Effective States:

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

Process is in the same "state"
when the future looks the same:
State(t) ~ State(t+2)




Complexity Lecture 2: Computational Mechanics (CSSS 2012); Jim Crutchfield

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.

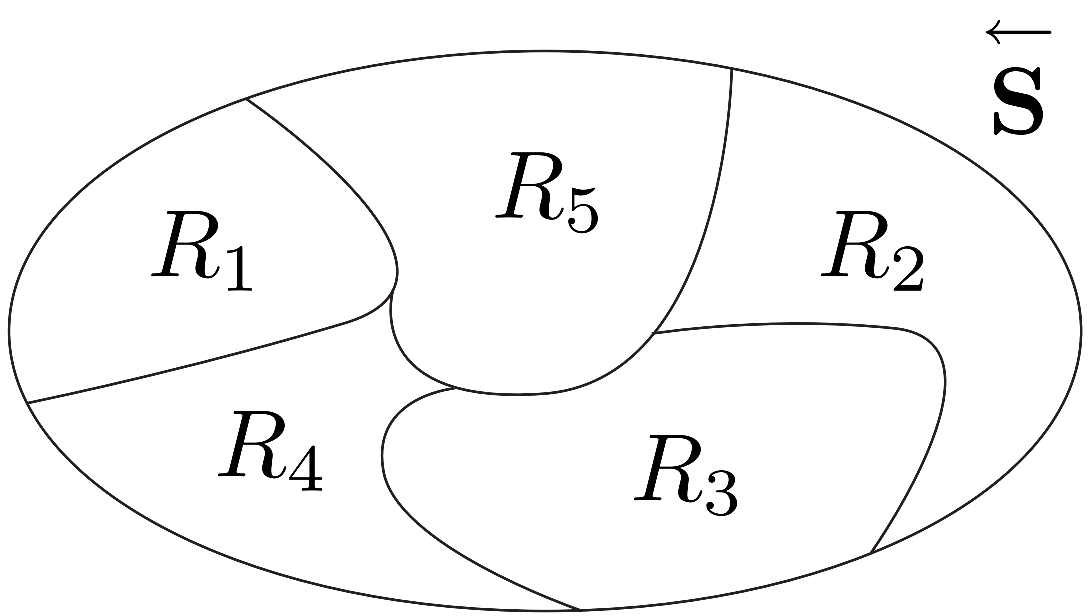
# The Learning Channel ...

## Space of Histories ...

Histories leading to the same predictions are equivalent.

Effective States = Partitions of History:

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



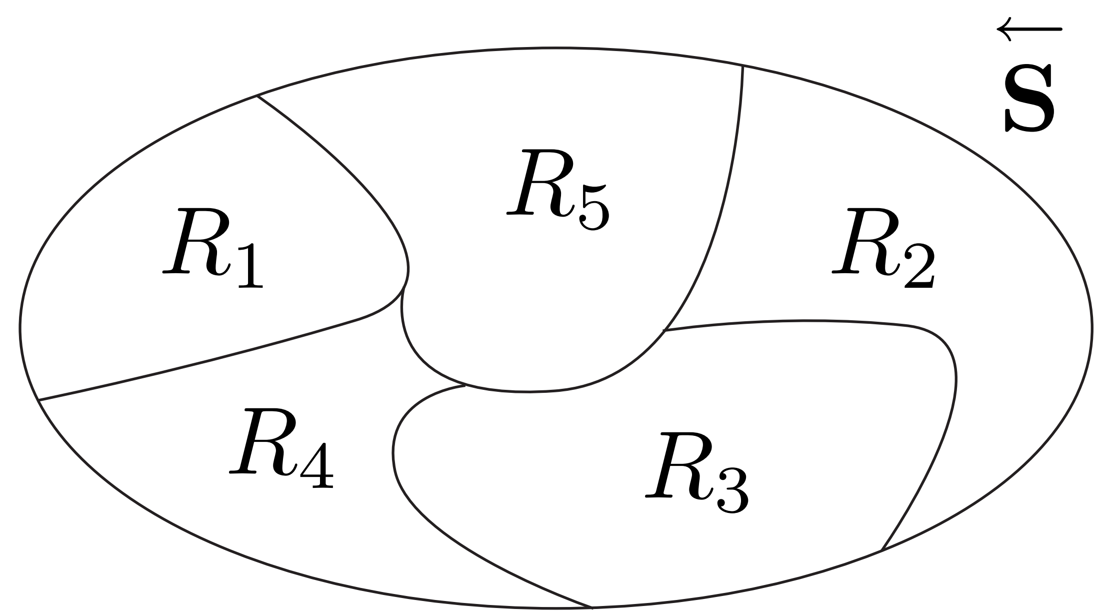
# The Learning Channel ...

## Space of Histories ...

Map from histories to partition elements:

$$\eta : \overleftarrow{\mathbf{S}} \to R$$

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




Random variable:

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

Distribution over Effective States:

$$\Pr(R = R_i) = \sum_{\overleftarrow{s} : \eta(\overleftarrow{s}) = R_i} \Pr(\overleftarrow{s})$$

# The Learning Channel ...

## How Effective are the Effective States?

Effective Prediction Error: Given a candidate partition $R$

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

Uncertainty about future given effective states

Effective Prediction Error Rate:

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

Entropy rate given effective states

# The Learning Channel ...

## 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 $R$.

# The Learning Channel ...

## 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:

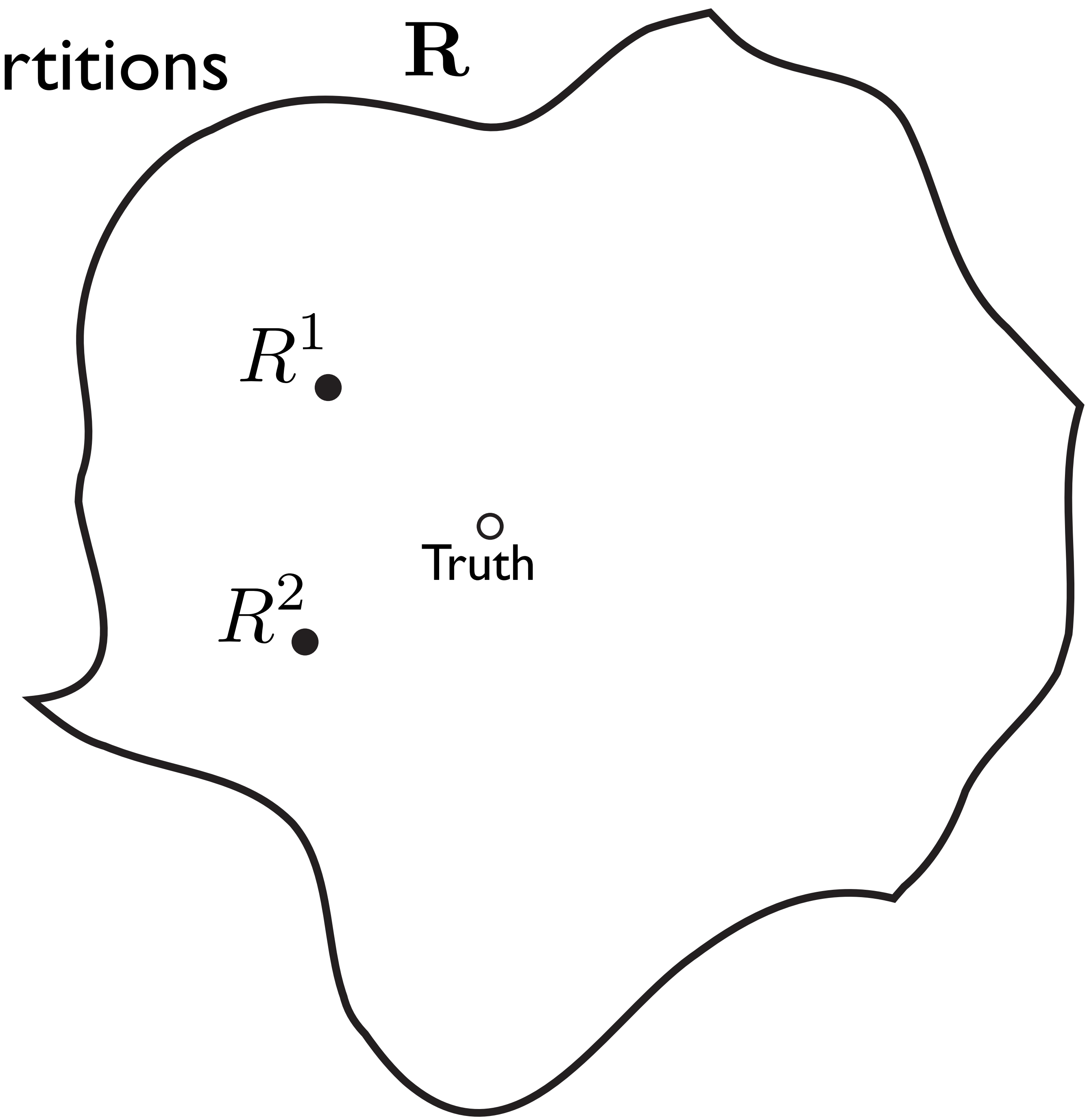
Can we find the smallest such set?

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

# Occam's Pool: The Space of Models

Model = Partition of History Space

Model Space $\mathbf{R}$ = Space of all partitions




Rival Models:

$$R_1, R_2 \in \mathbf{R}$$

## 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**:

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

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

# The Learning Channel ...

## Causal State Components

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:

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

Causal state map:

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

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

## Causal States ...

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

We have the effective states!

Now,

What is the dynamic?

# The Learning Channel ...

## Causal State Dynamic ...

### Causal-state Filtering:

$$\overleftrightarrow{s} = \ldots \; s_{-3} \quad s_{-2} \quad s_{-1} \quad s_0 \quad s_1 \quad s_2 \quad s_3 \; \ldots$$

$$\epsilon(\overleftrightarrow{s}) = \ldots \epsilon(\overleftarrow{s}_{-3})\epsilon(\overleftarrow{s}_{-2})\epsilon(\overleftarrow{s}_{-1})\epsilon(\overleftarrow{s}_{0})\epsilon(\overleftarrow{s}_{1})\epsilon(\overleftarrow{s}_{2})\epsilon(\overleftarrow{s}_{3}) \ldots$$

$$\overleftrightarrow{\mathcal{S}} = \ldots \; \mathcal{S}_{t=-3} \; \mathcal{S}_{t=-2} \; \mathcal{S}_{t=-1} \; \mathcal{S}_{t=0} \; \mathcal{S}_{t=1} \; \mathcal{S}_{t=2} \; \mathcal{S}_{t=3} \; \ldots$$

### Causal-state process:

$$\Pr(\overleftrightarrow{\mathcal{S}})$$

## Causal State Dynamic ...

Conditional transition probability:

$$\begin{aligned} T_{ij}^{(s)} &= \Pr(\mathcal{S}_j, s | \mathcal{S}_i) \\ &= \Pr\left(\mathcal{S} = \epsilon(\overleftarrow{s}\, s) | \mathcal{S} = \epsilon(\overleftarrow{s})\right) \end{aligned}$$

State-to-State Transitions:

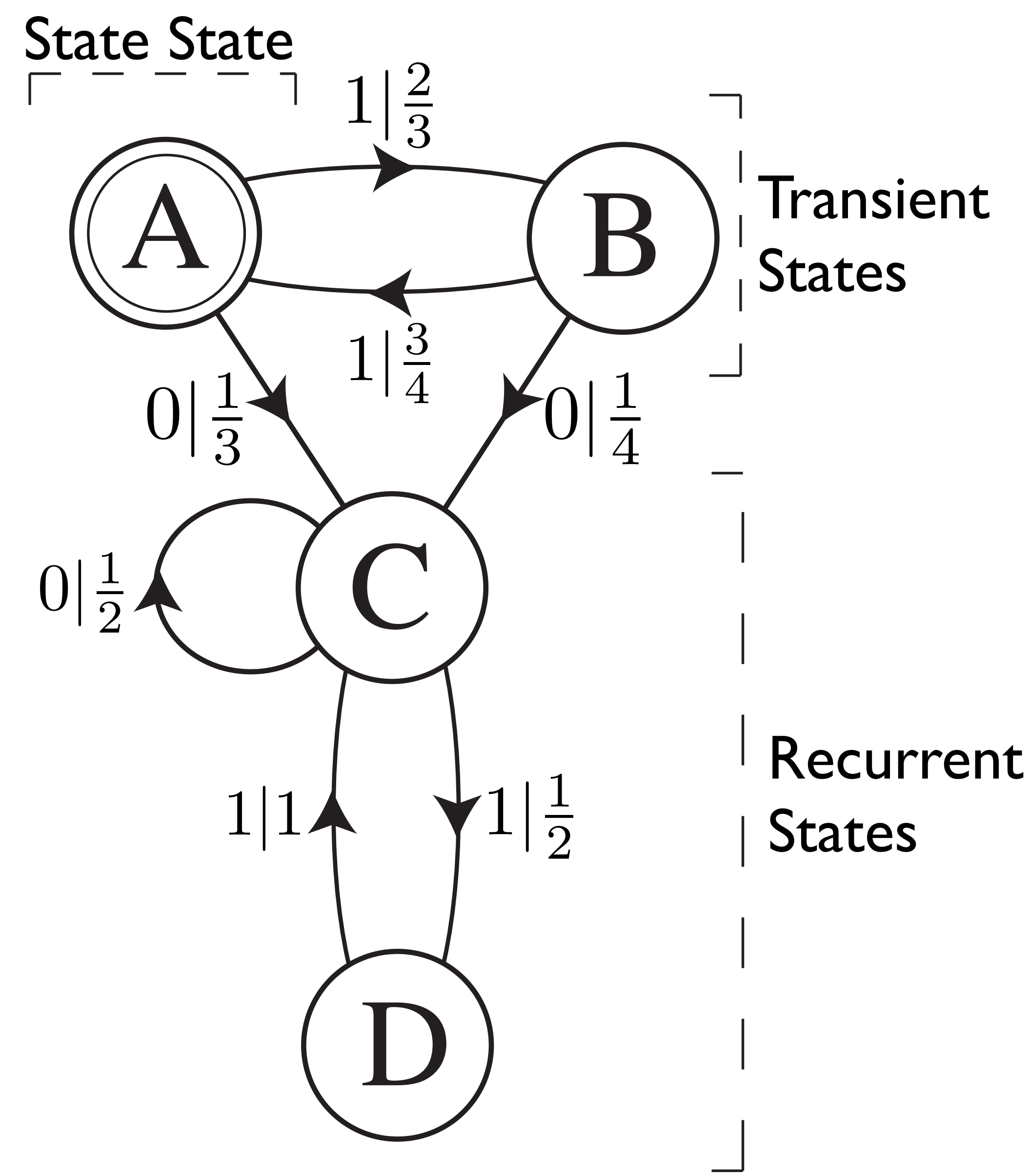
$$\{T_{ij}^{(s)} : s \in \mathcal{A},\ i, j = 0, 1, \ldots, |\boldsymbol{\mathcal{S}}|\}$$

# The $\epsilon$-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:

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

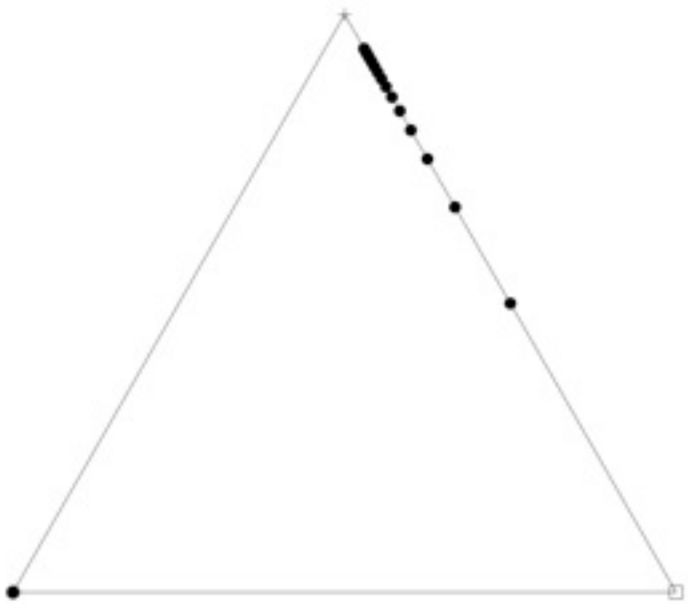
$$\Pr(\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2, \ldots) = (1, 0, 0, \ldots)$$

Transient States

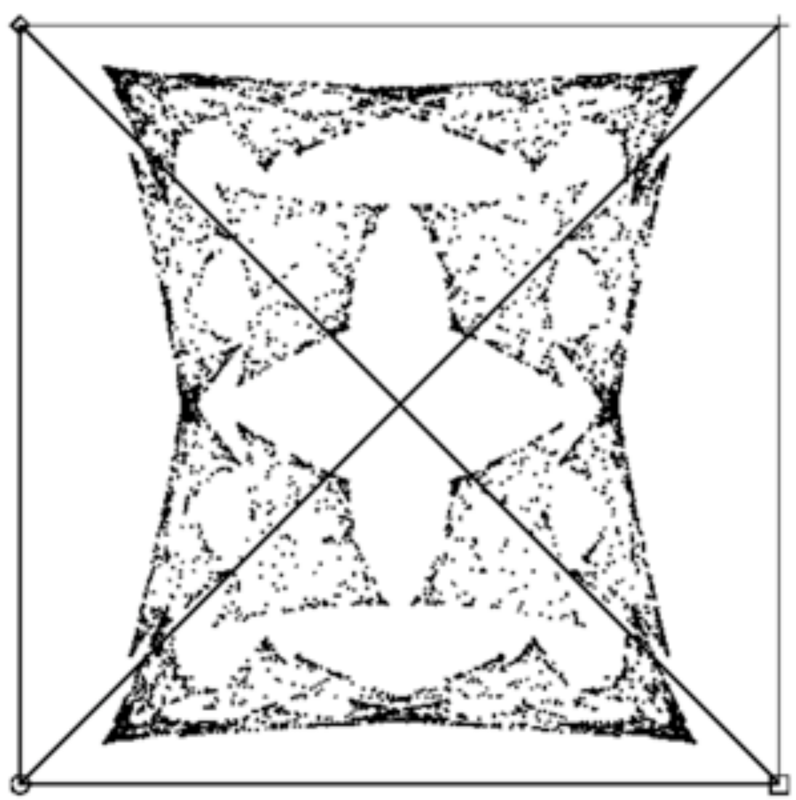
Recurrent States

# The Learning Channel ...

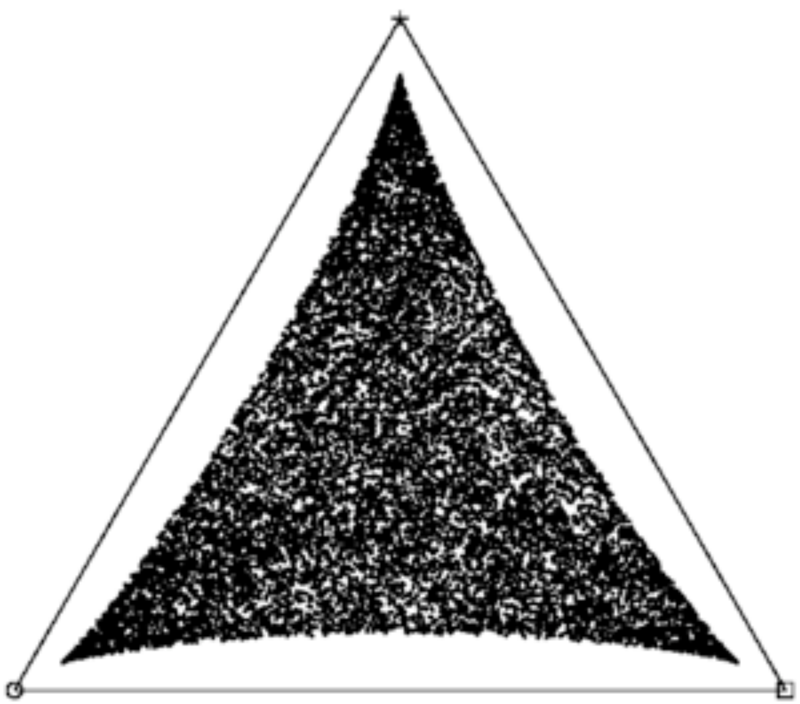
## The $\epsilon$-Machine of a Process ...



Denumerable Causal States

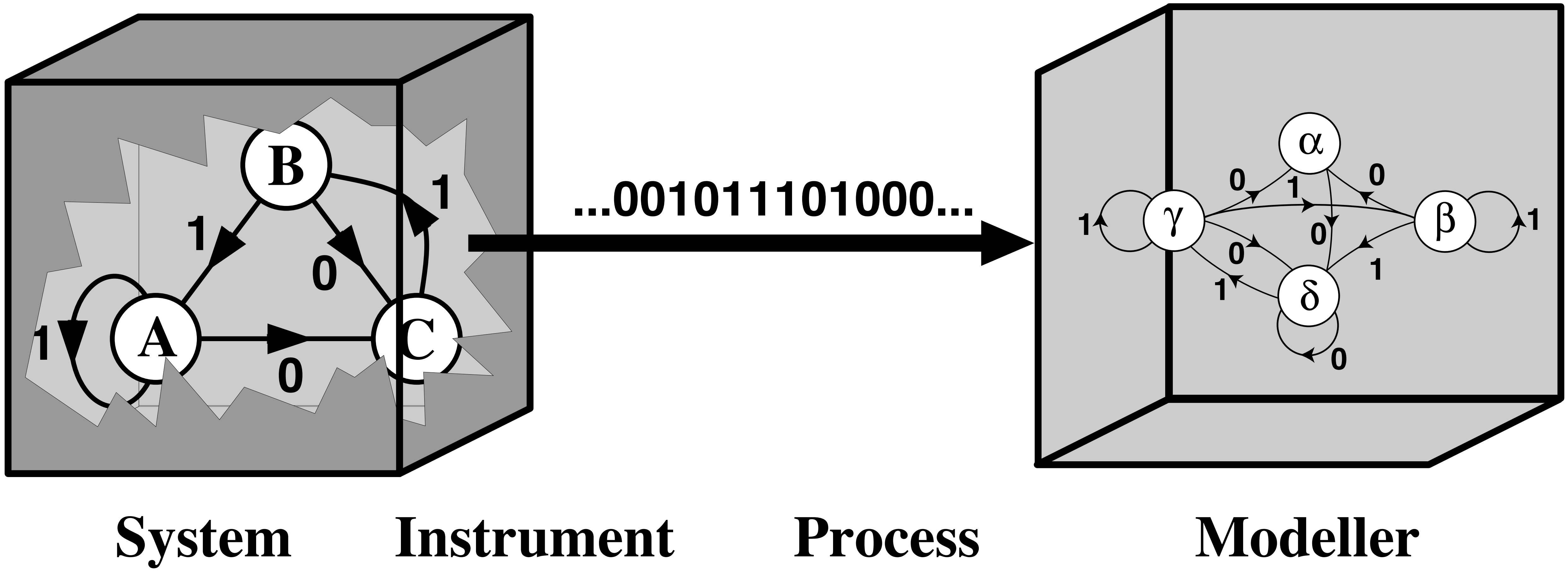


Fractal



Continuous

# The Learning Channel:




Central questions:

What are the states? Causal States

What is the dynamic? The $\epsilon$-Machine

# The $\epsilon$-Machine ...

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

$\epsilon$-Machine reproduces the process's word distribution:

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

$$s^L = s_1 s_2 \ldots s_L \qquad \mathcal{S}(t=0) = \mathcal{S}_0$$

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

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

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

# The $\epsilon$-Machine ...

Causal shielding:

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.

# The $\epsilon$-Machine ...

$\epsilon$Ms are Unifilar: $(\mathcal{S}_t, s) \rightarrow$ unique $\mathcal{S}_{t+1}$

(in automata theory, "deterministic")

That is:

(1) $\mathcal{S}_i \in \boldsymbol{\mathcal{S}},\; s \in \mathcal{A}$, at most one $\mathcal{S}_j \in \boldsymbol{\mathcal{S}}$:

$$\overleftarrow{s} \in \mathcal{S}_i \Rightarrow \overleftarrow{s} s \in \mathcal{S}_j$$

(2) If there is a next causal state $j$:

$$\mathcal{S}_{k \neq j} \in \boldsymbol{\mathcal{S}} \Rightarrow T_{ik}^{(s)} = 0$$

(3) If there is not:

$$T_{ij}^{(s)} = 0$$

# The $\epsilon$-Machine ...

## Unifiliarity ...

Consequence:

Unifilarity: 1-1 map between state-sequences & symbol-sequences.

Entropy rate expression requires this 1-1 mapping.

Can use $\epsilon\mathrm{M}$ to calculate entropy rate $h_\mu$ .
(Any unifilar presentation will do.)

## The $\epsilon$-Machine ...

$\epsilon$Ms 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:

$$H\left[\overrightarrow{S}^L | \mathcal{S}\right] = H\left[\overrightarrow{S}^L | \overleftarrow{s} \in \mathcal{S}\right]$$

(Causal equiv. rel'n)

$$= H\left[\overrightarrow{S}^L | \overleftarrow{s}\right]$$

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

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

(Data processing inequality)

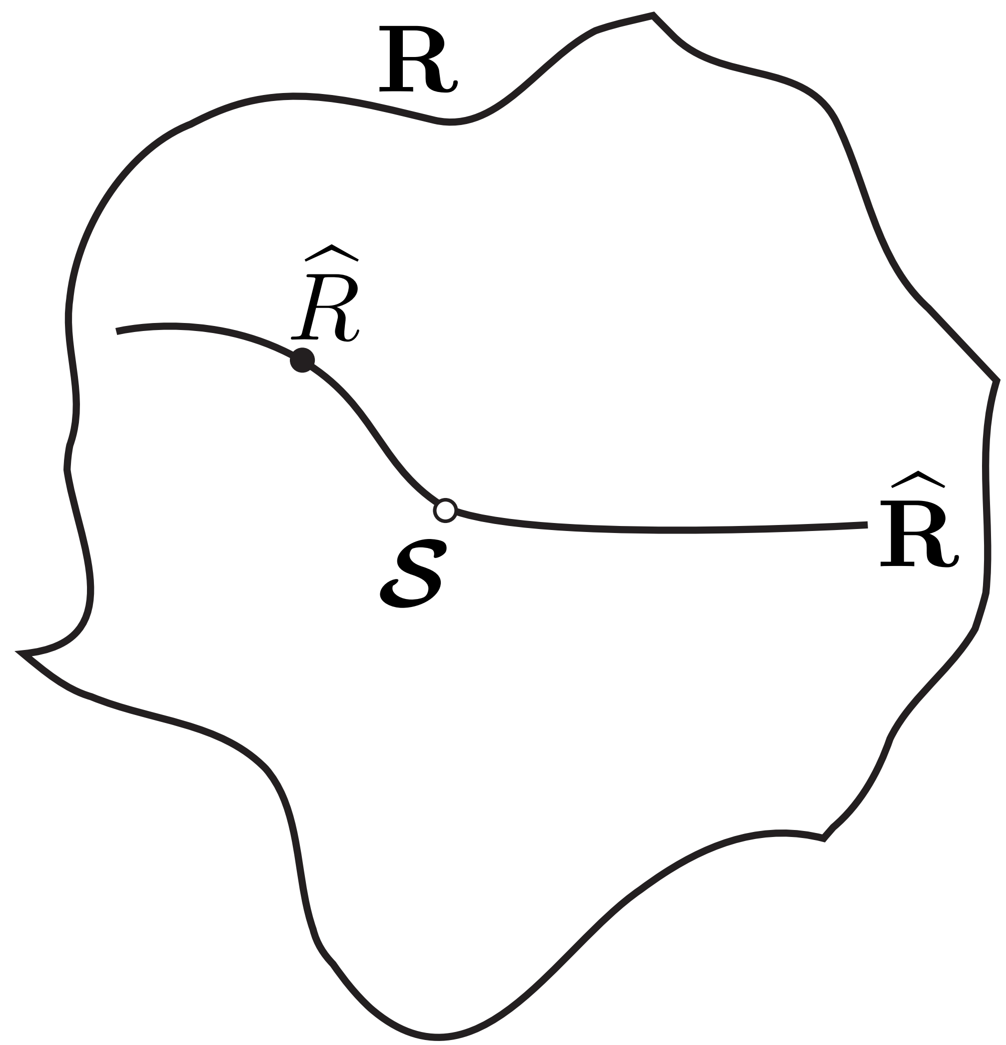
# The $\epsilon$-Machine ...

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.)

# The $\epsilon$-Machine ...

## 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 $C_\mu \leq H[\widehat{R}]$ $\square$

## The $\epsilon$-Machine ...

### Minimal Statistical Complexity ...

Consequence:

(1) $C_\mu$ measures historical information process stores.

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

Remarks:

(1) Causal states contain every difference (in past) that makes a difference (to future) (Bateson "information")

(2) Causal states are sufficient statistics for the future.

# The $\epsilon$-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.

# The $\epsilon$-Machine ...

Dynamical system's intrinsic computation:

(1) How much of past does process store?

(2) In what architecture is that information stored?

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

## The $\epsilon$-Machine ...

Dynamical system's intrinsic computation:

(1) How much of past does process store?

$$C_\mu$$

(2) In what architecture is that information stored?

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

# The $\epsilon$-Machine ...

Dynamical system'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?

# The $\epsilon$-Machine ...

Dynamical system'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?

$$h_\mu$$