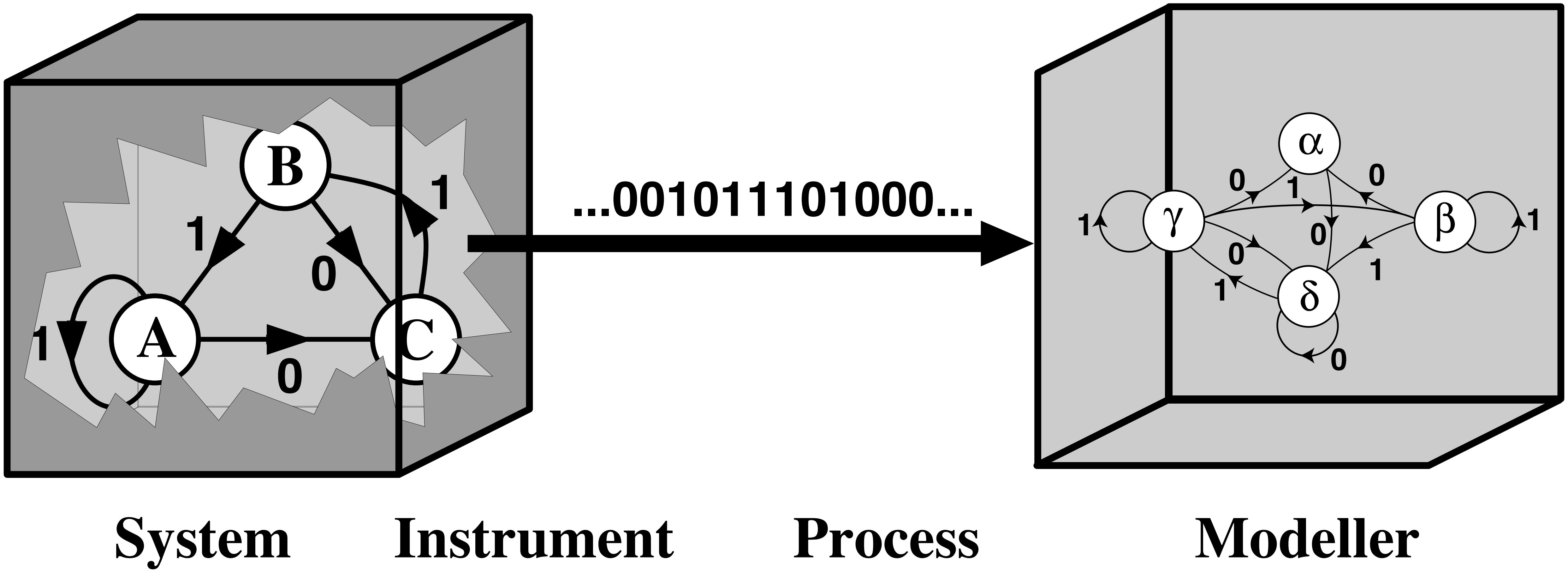
The Learning Channel

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 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 Learning Channel ...

The Prediction Game ...

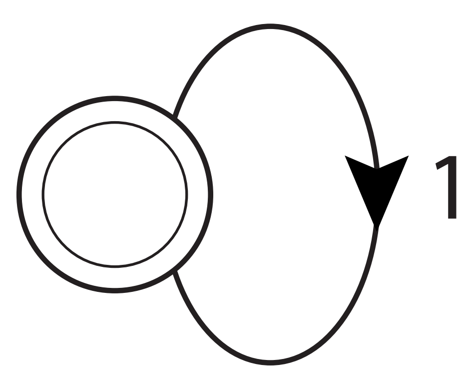
Process I:

Past: $\ldots 111111111111$

Your prediction is?

Future: $111111111111\ldots$

Your model (states & dynamic) is?



The Learning Channel ...

The Prediction Game ...

Process II:

The Prediction Game ...

Process II:

Past: $\ldots 10110010001101110$

The Learning Channel ...

The Prediction Game ...

Process II:

Past: $\ldots 10110010001101110$

Your prediction is?

The Learning Channel ...

The Prediction Game ...

Process II:

Past: $\ldots 10110010001101110$

Your prediction is?

Analysis: All words of length L occur & equally often

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?

The Learning Channel ...

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?

$$0101011101000110 1 \ldots$$

The Learning Channel ...

The Prediction Game ...

Process II:

Past: $\ldots 10111001000110110$

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 Learning Channel ...

The Prediction Game ...

Process II:

Past: $\ldots 10110010001101110$

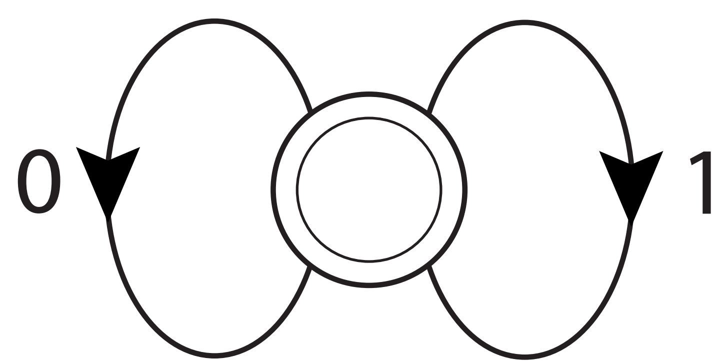
Your prediction is?

Analysis: All words of length L occur & equally often

Future: Well, anything can happen, how about?

$$0101011101001101\ldots$$

Your model is?



Complexity Lecture 2: Computational Mechanics (CSSS 2011); Jim Crutchfield

The Learning Channel ...

The Prediction Game ...

Process III:

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: $1010101010101010 \ldots$

The Learning Channel ...

The Prediction Game ...

Process III:

Past: $\ldots 1010101010101010$

Your prediction is?

Future: $1010101010101010\ldots$

Your model is?

The Learning Channel ...

The Prediction Game ...

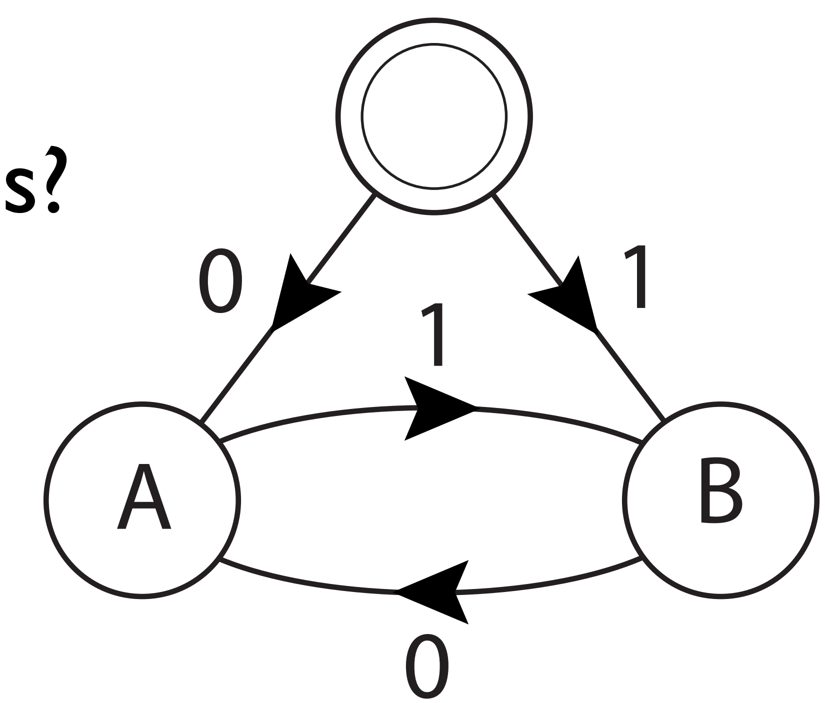
Process III:

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 $\mathcal{A}$

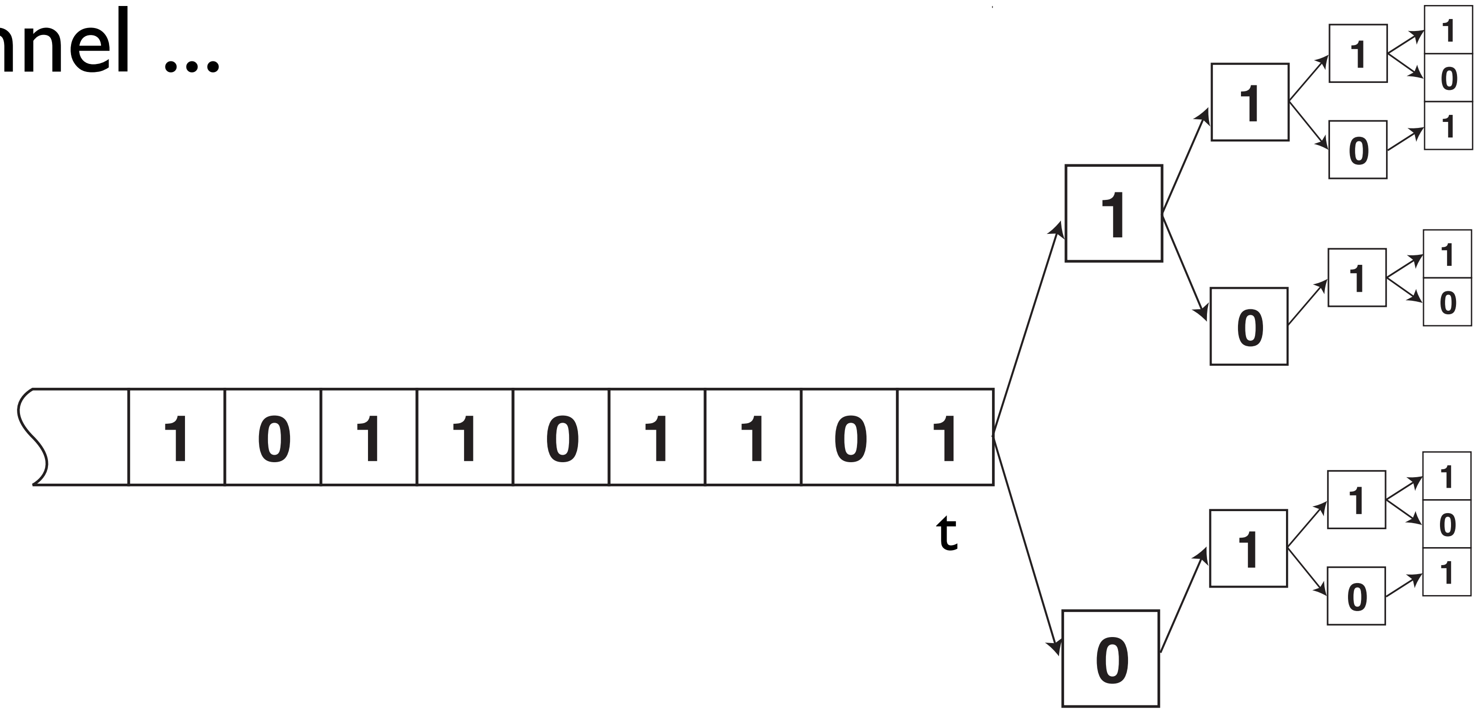
These symbols only indirectly reflect the hidden states

The Learning Channel ...

Effective States:

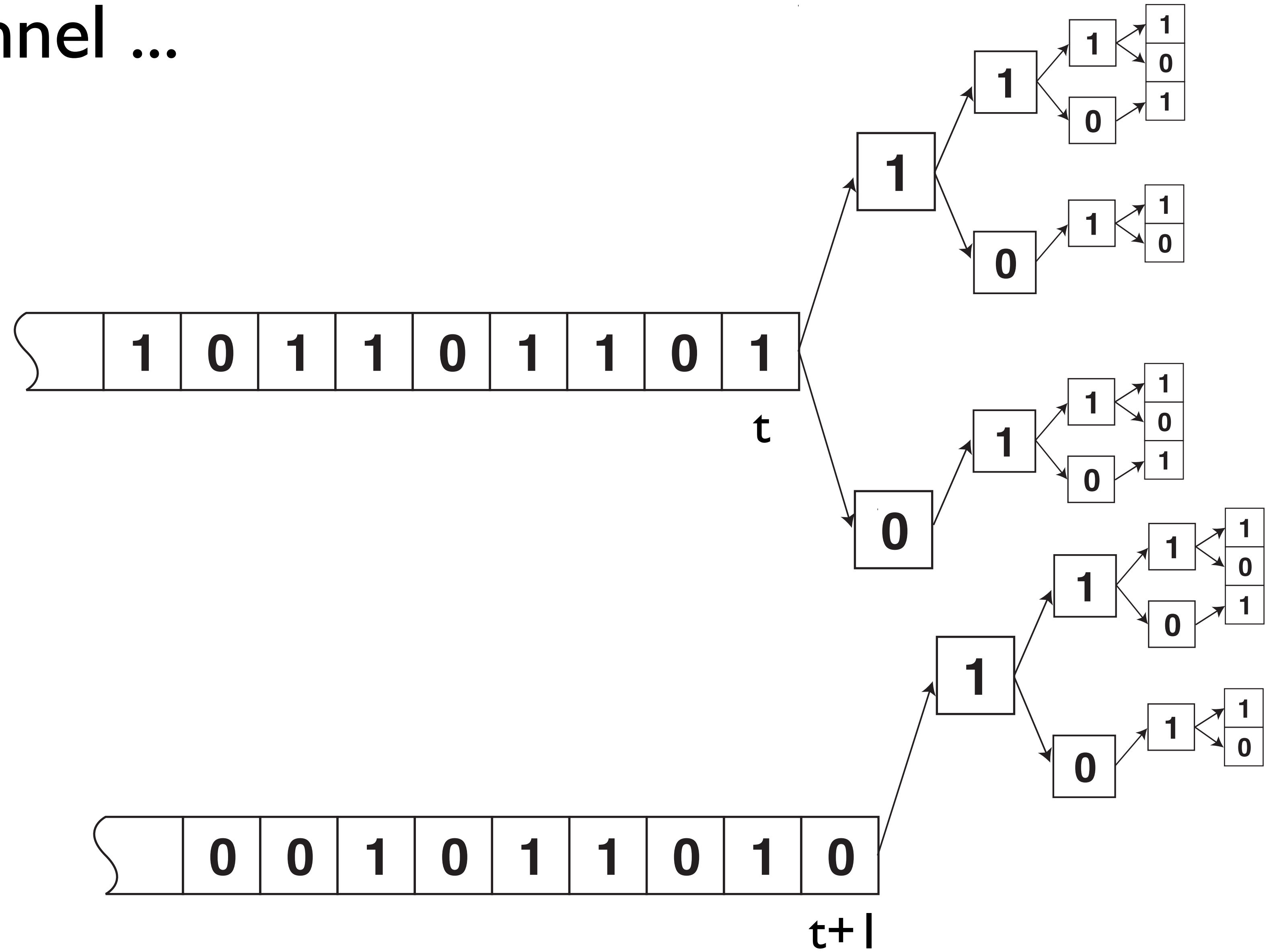
The Learning Channel ...

Effective States:



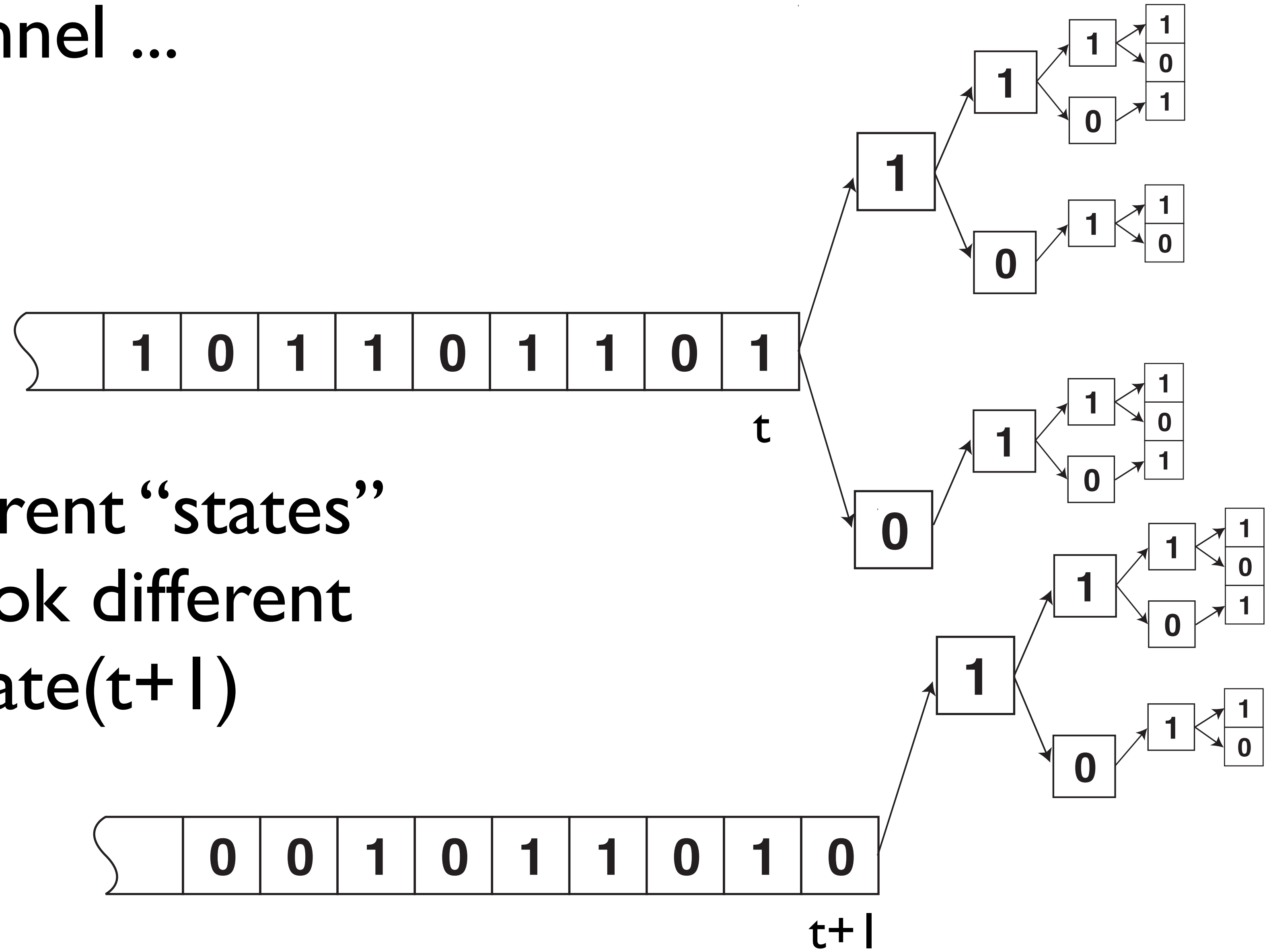
The Learning Channel ...

Effective States:



The Learning Channel ...

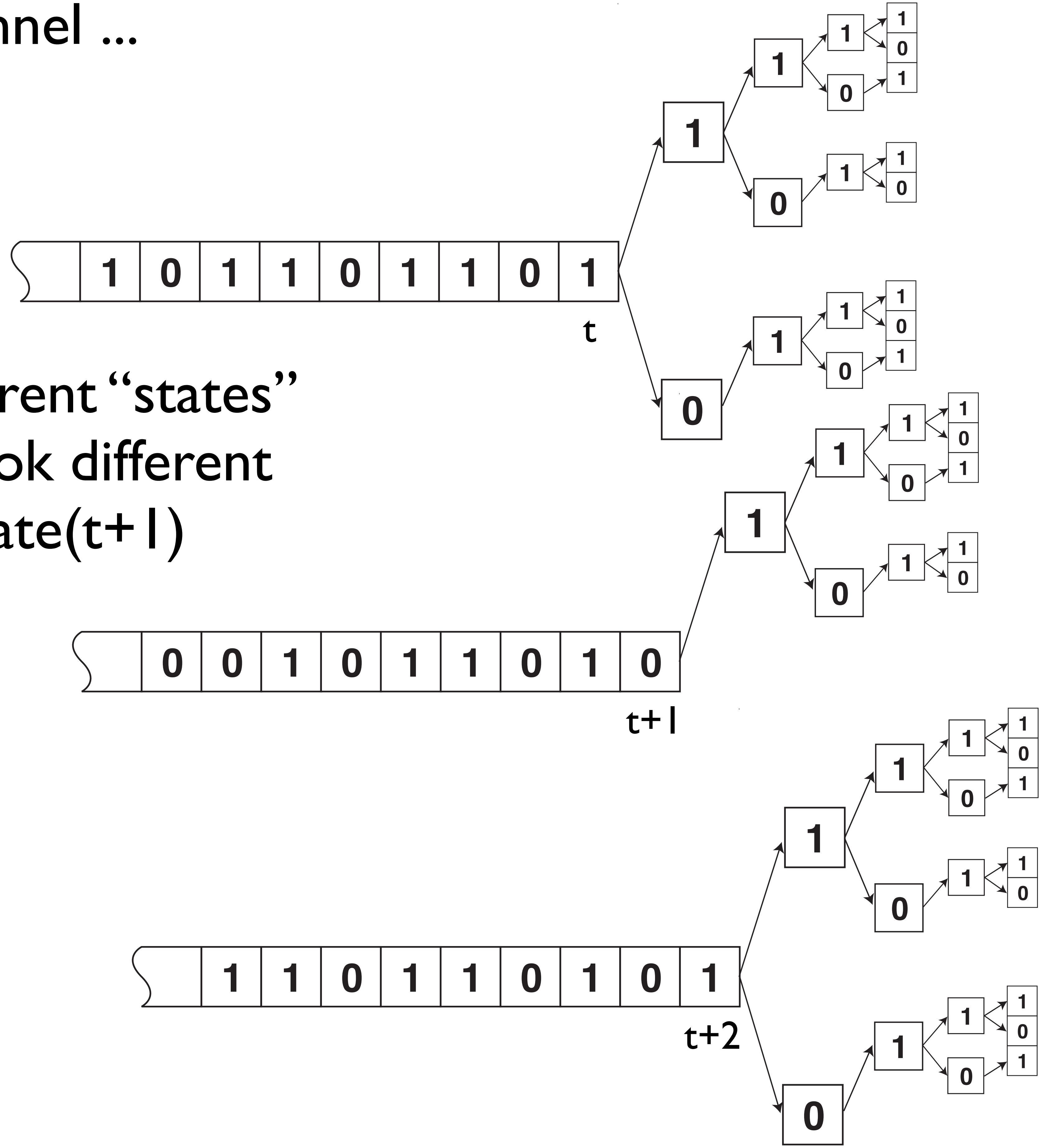
Effective States:



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

The Learning Channel ...

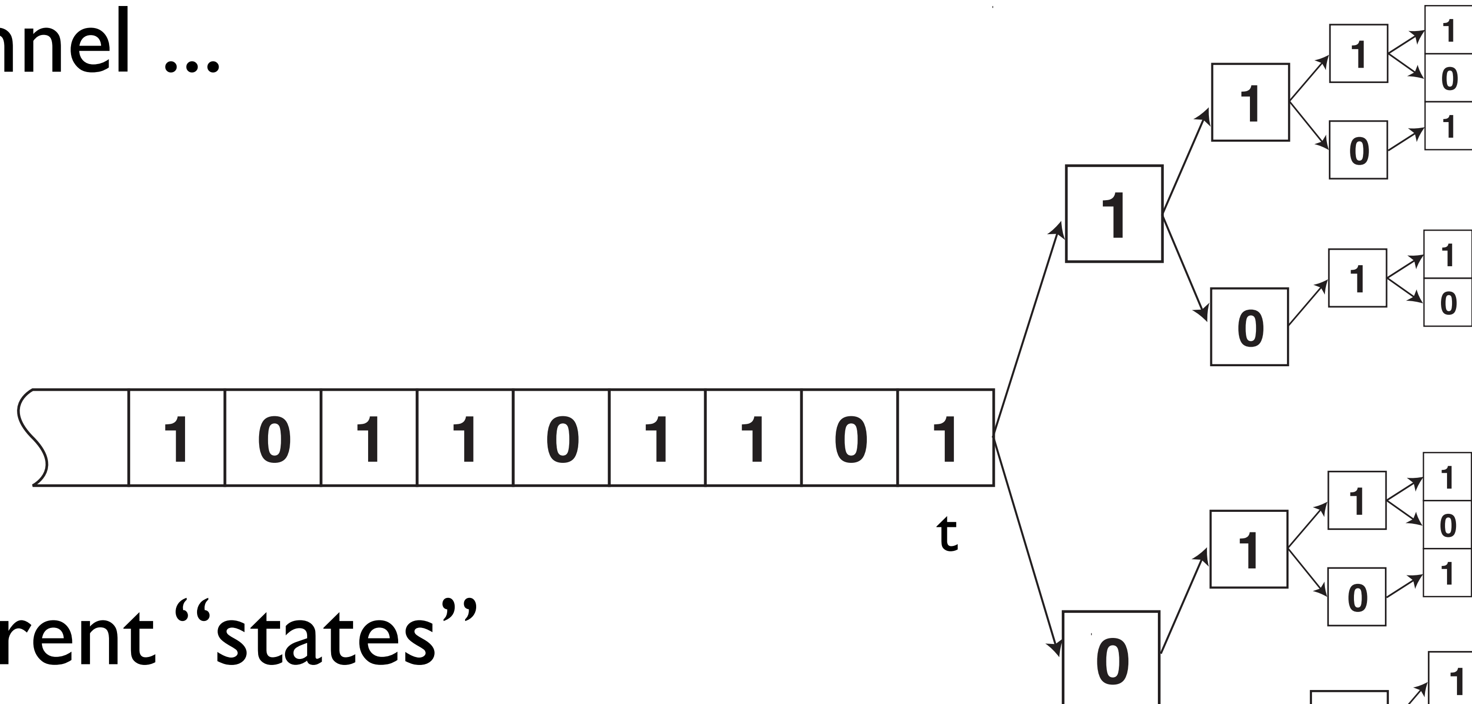
Effective States:



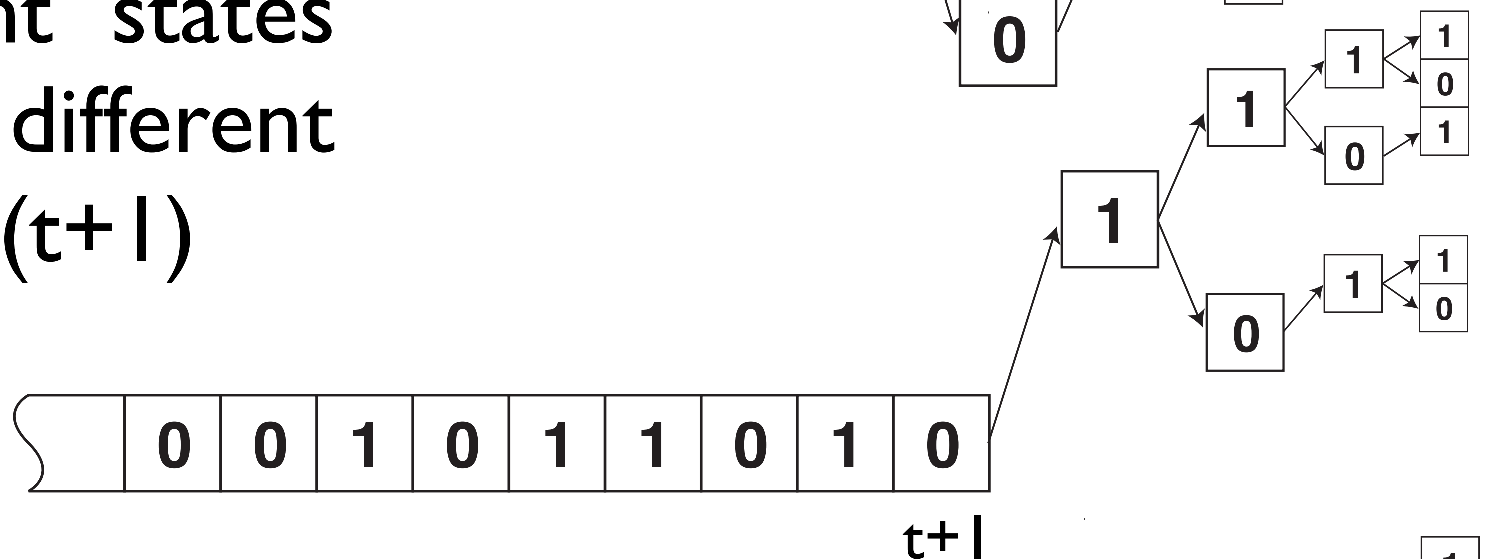
Process is in different "states"
when futures look different
State(t) $\sim$ State(t+1)

The Learning Channel ...

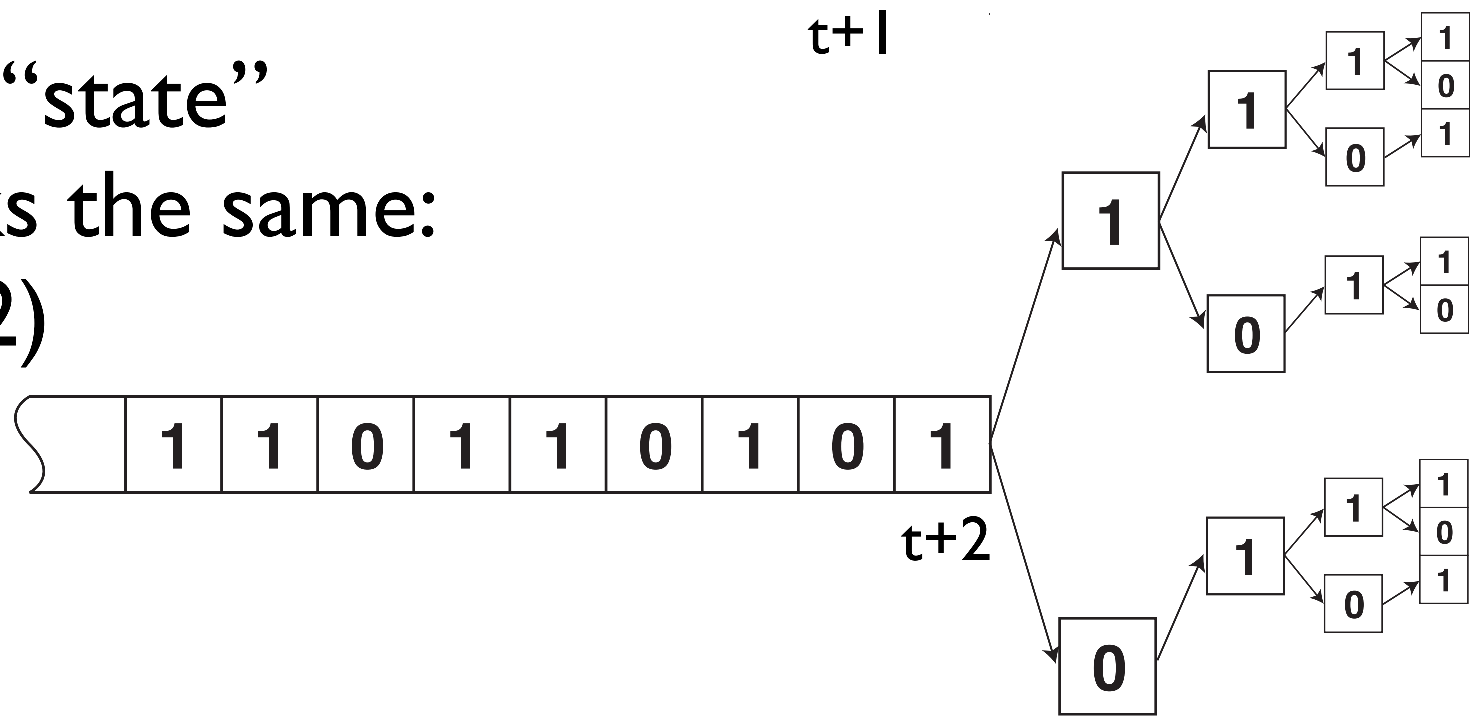
Effective States:



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 2011); Jim Crutchfield

The Learning Channel ...

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:

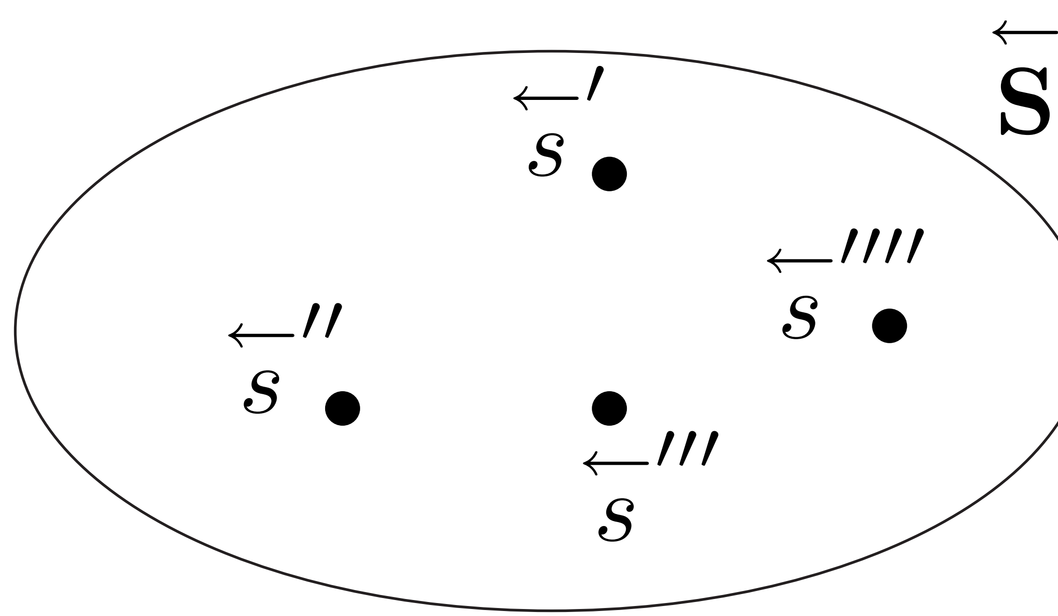
Predict as much about the future $\overrightarrow{S}$,

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

The Learning Channel ...

Space of Histories:

$$\overleftarrow{\mathbf{S}} = \mathcal{A}^{\mathbb{Z}^-} = \{\ldots s_{-3} s_{-2} s_{-1} : s_i \in \mathcal{A}, i = \ldots, -3, -2, -1\}$$



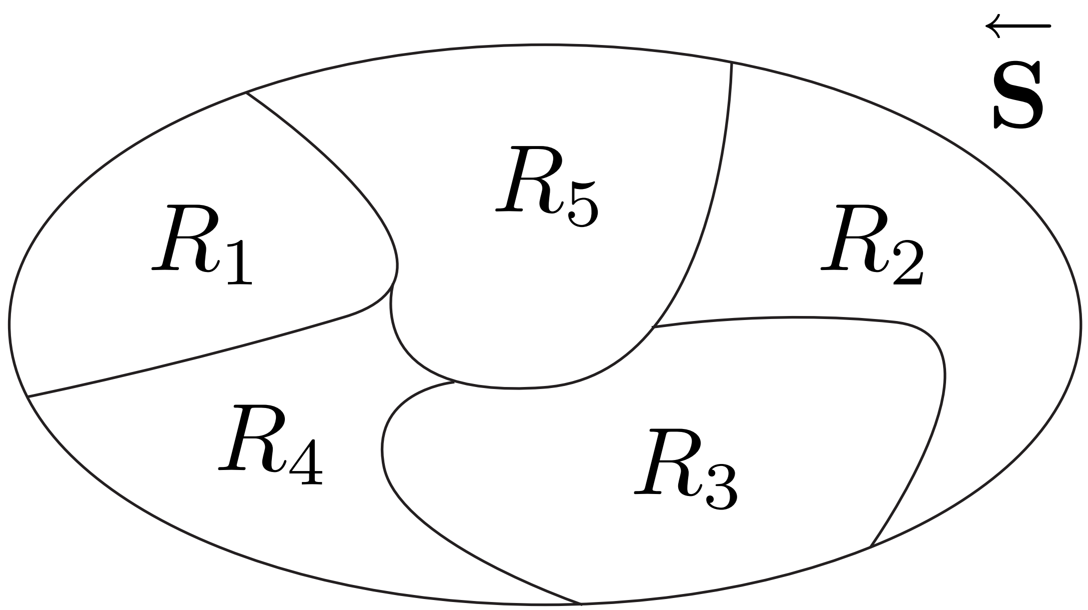
The Learning Channel ...

Space of Histories ...

Histories leading to the same predictions are equivalent.

Effective States = Partitions of History:

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



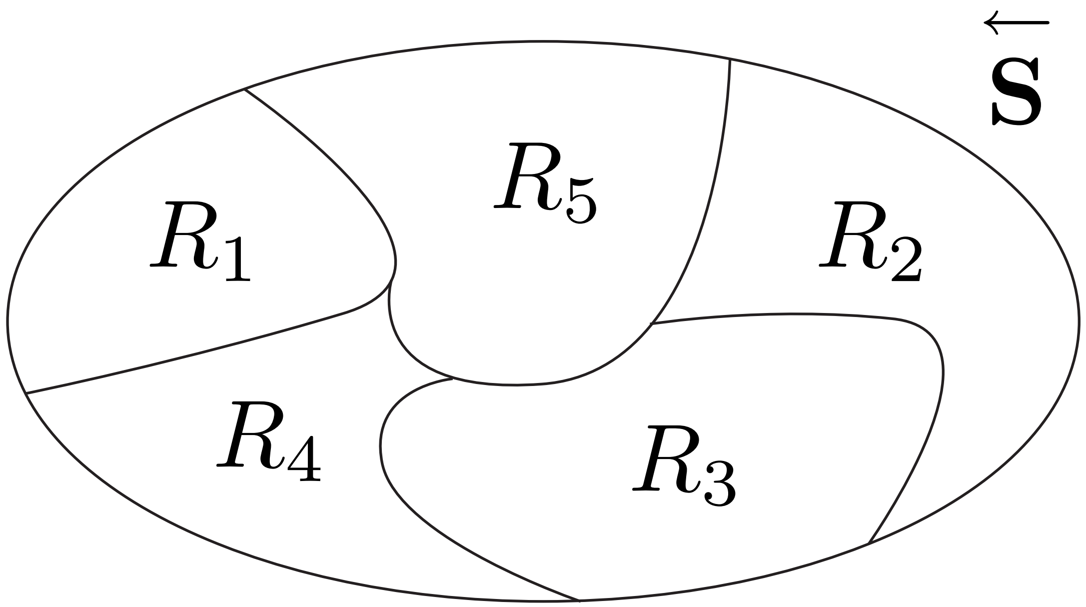
The Learning Channel ...

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:

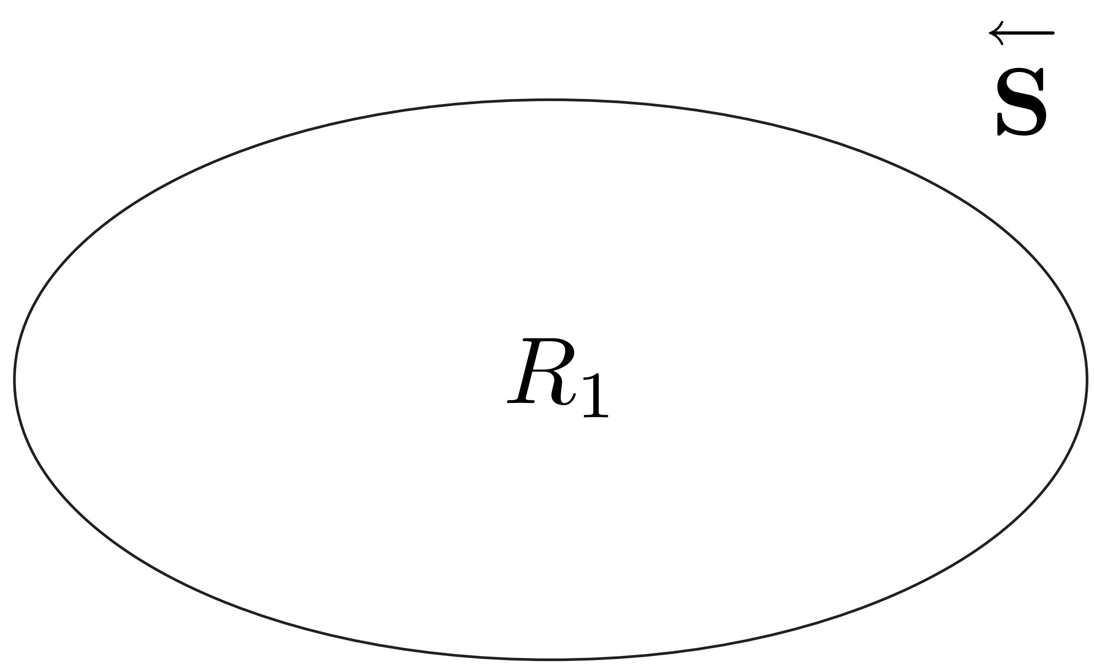
$$\Pr(R = R_i) = \sum_{\overleftarrow{s} : \eta(\overleftarrow{s}) = R_i} \Pr(\overleftarrow{s})$$

The Learning Channel ...

Space of Histories ...

Null Model:

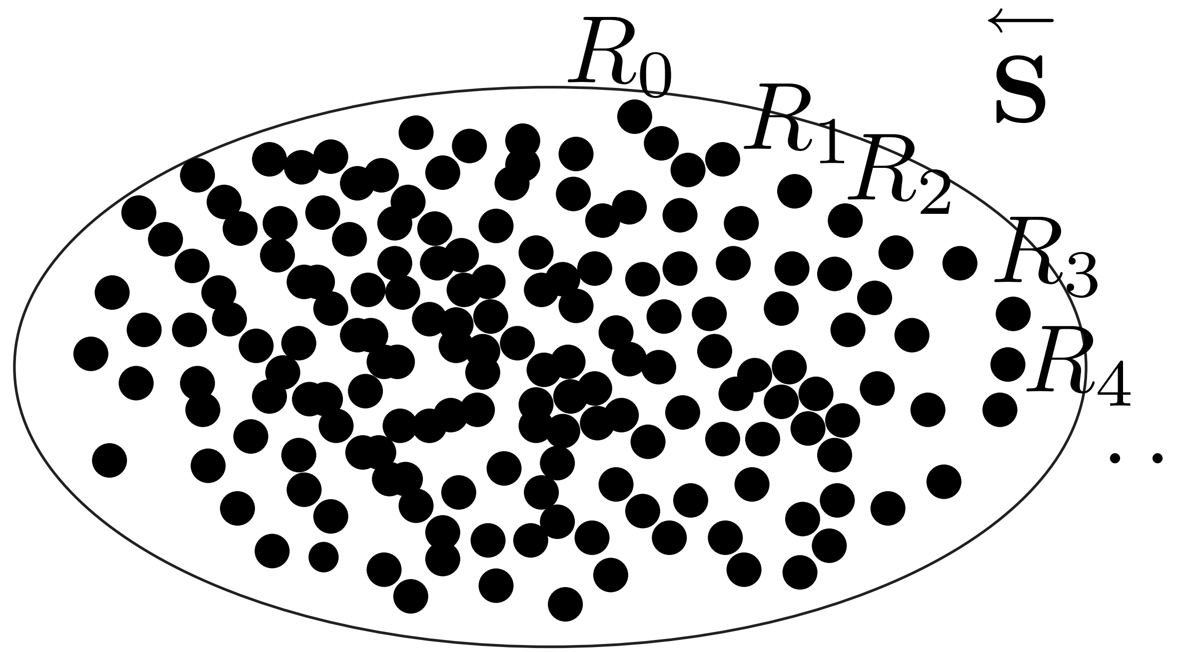
$$R_1 = \{R_1 : R_1 = \mathcal{A}^{\mathbb{Z}^-}\}$$



The Learning Channel ...
Space of Histories ...

Every-History-Is-Precious Model:

$$R_\infty = \{R_i : R_i \in \mathcal{A}^{\mathbb{Z}^-}\}$$



Each past is a state: $R_i = \overleftarrow{x}$

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:

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

Entropy rate given effective states

The Learning Channel ...

How Effective are the Effective States?

Effective Prediction Error ...

Bounds:

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

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

The Learning Channel ...

How Effective are the Effective States?

Effective Prediction Error ...

Limits on Prediction:

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

$$\geq H[\overrightarrow{S}^L \mid \overleftarrow{S}]$$

(Data Processing Inequality)

Models can do no better than to use histories.

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

In particular, $h_\mu(R = \overleftarrow{S}) = h_\mu$

The Learning Channel ...

How Effective are the Effective States ...

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

Solution: $h_\mu(R_\infty)$ ... rather verbose!

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?

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

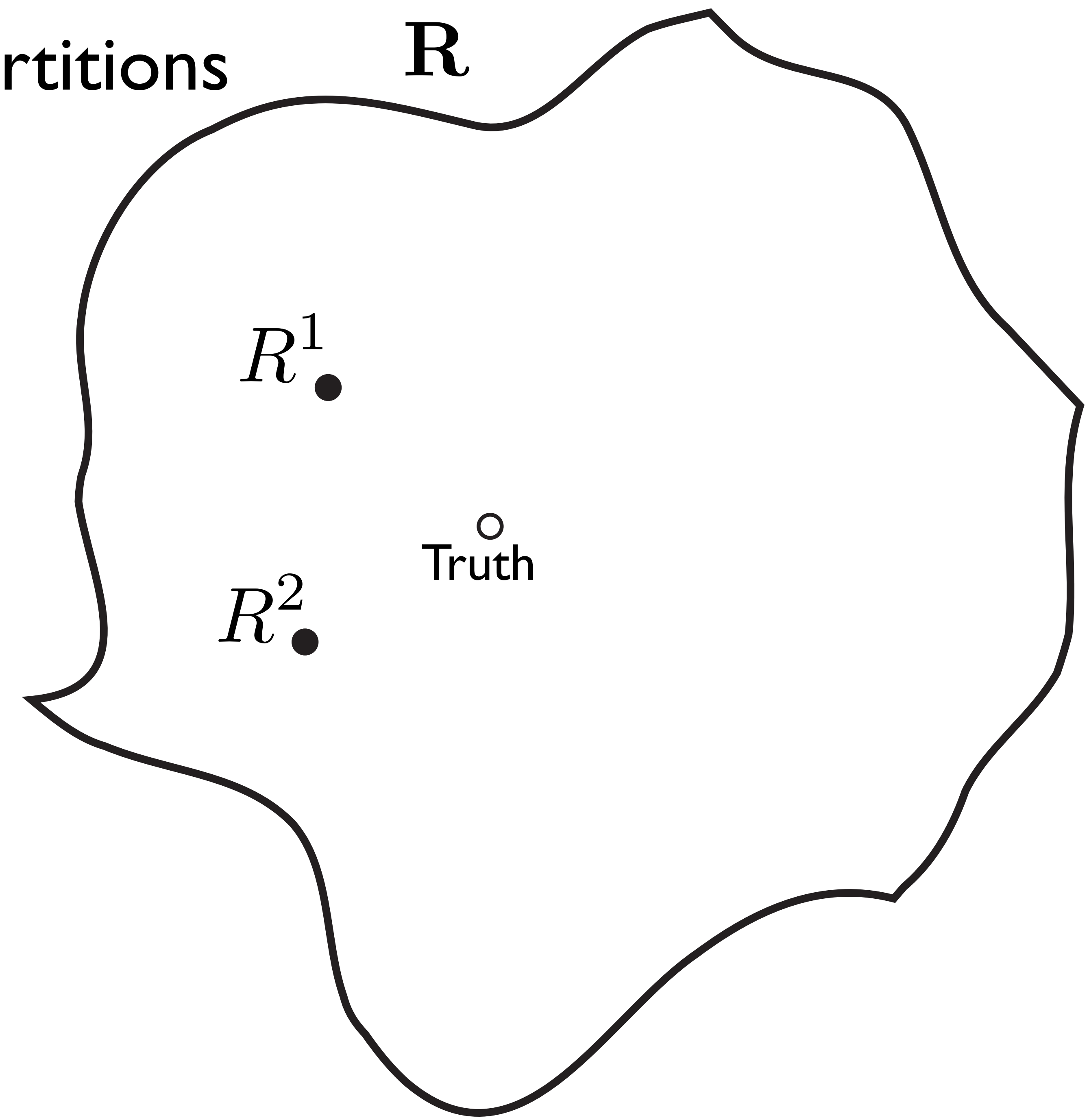
Occam's Pool: The Space of Models

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:

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

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

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:

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

Partition of histories:

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

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

Causal State Components ...

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 ...

Causal state morph:

$$\Pr\left(\overrightarrow{S}^L | \mathcal{S}\right)$$

$$L = 1, 2, \ldots, \ \forall s^L, \overleftarrow{s}$$

$$\Pr\left(\overrightarrow{S}^L = s^L | \mathcal{S} = \epsilon(\overleftarrow{s})\right) = \Pr\left(\overrightarrow{S}^L = s^L | \overleftarrow{s}\right)$$

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:

Have history:

$$\overleftarrow{s}' = \ldots s_{-3} s_{-2} s_{-1}$$

And so in state $\mathcal{S}_i = \epsilon(\overleftarrow{s}')$

Observe symbol: $s \in \mathcal{A}$

Have a new history:

$$\overleftarrow{s}'' = \overleftarrow{s}' s$$

$$\overleftarrow{s}'' = \ldots s_{-2} s_{-1} s$$

Now in state $\mathcal{S}_j = \epsilon(\overleftarrow{s}'')$

Transition: $\mathcal{S}_i \rightarrow_s \mathcal{S}_j$

Causal State Dynamic ...

Causal-state filtering:

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

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

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

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 ϵ-Machine of a Process:

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

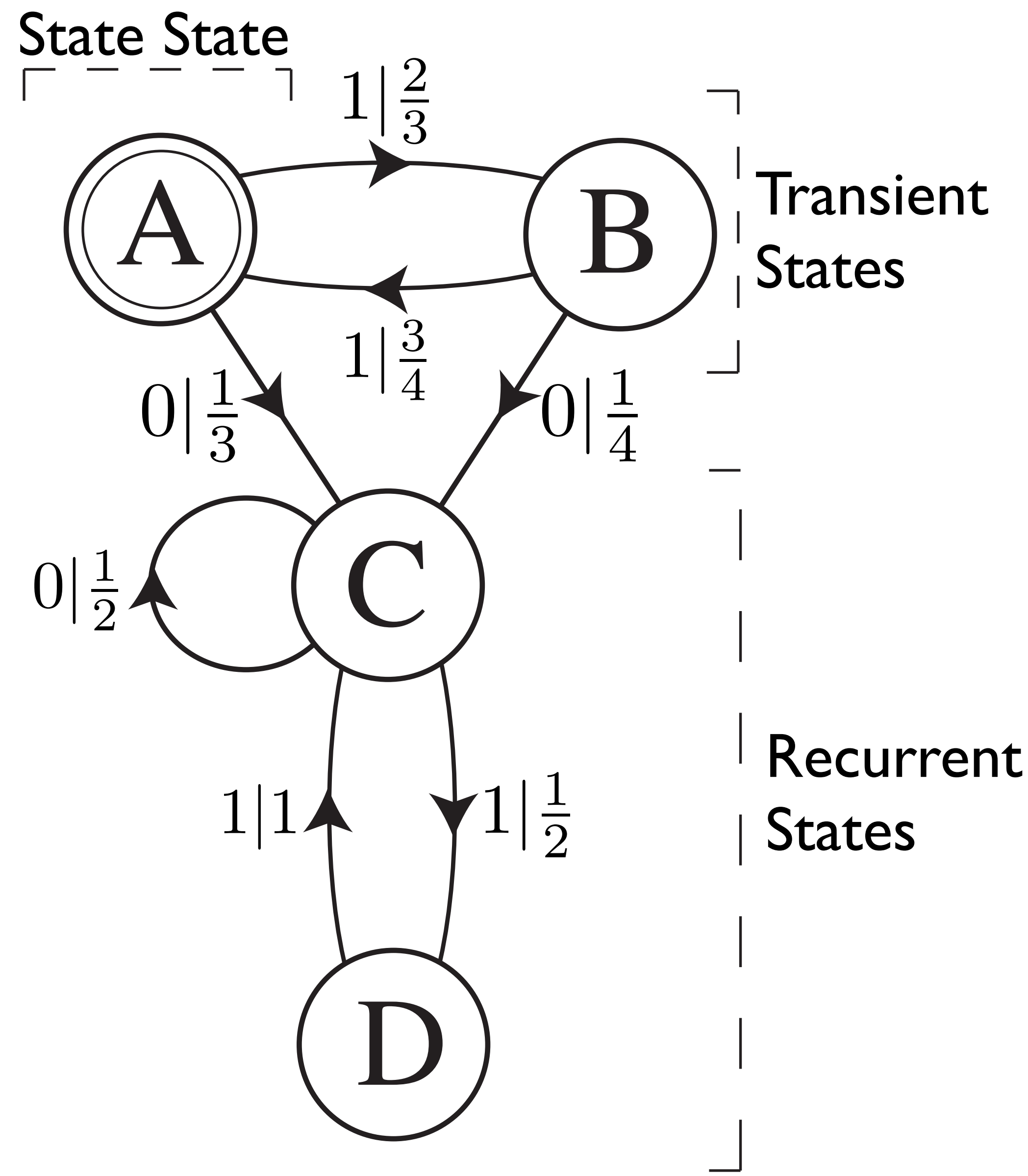
A type of hidden Markov model

The Learning Channel ...

The ϵ-Machine of a Process ...

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

For example ...



Complexity Lecture 2: Computational Mechanics (CSSS 2011); Jim Crutchfield

The Learning Channel ...

The ϵ-Machine ...

Unique Start State: Condition of total ignorance

Null symbol: λ

No measurements made:

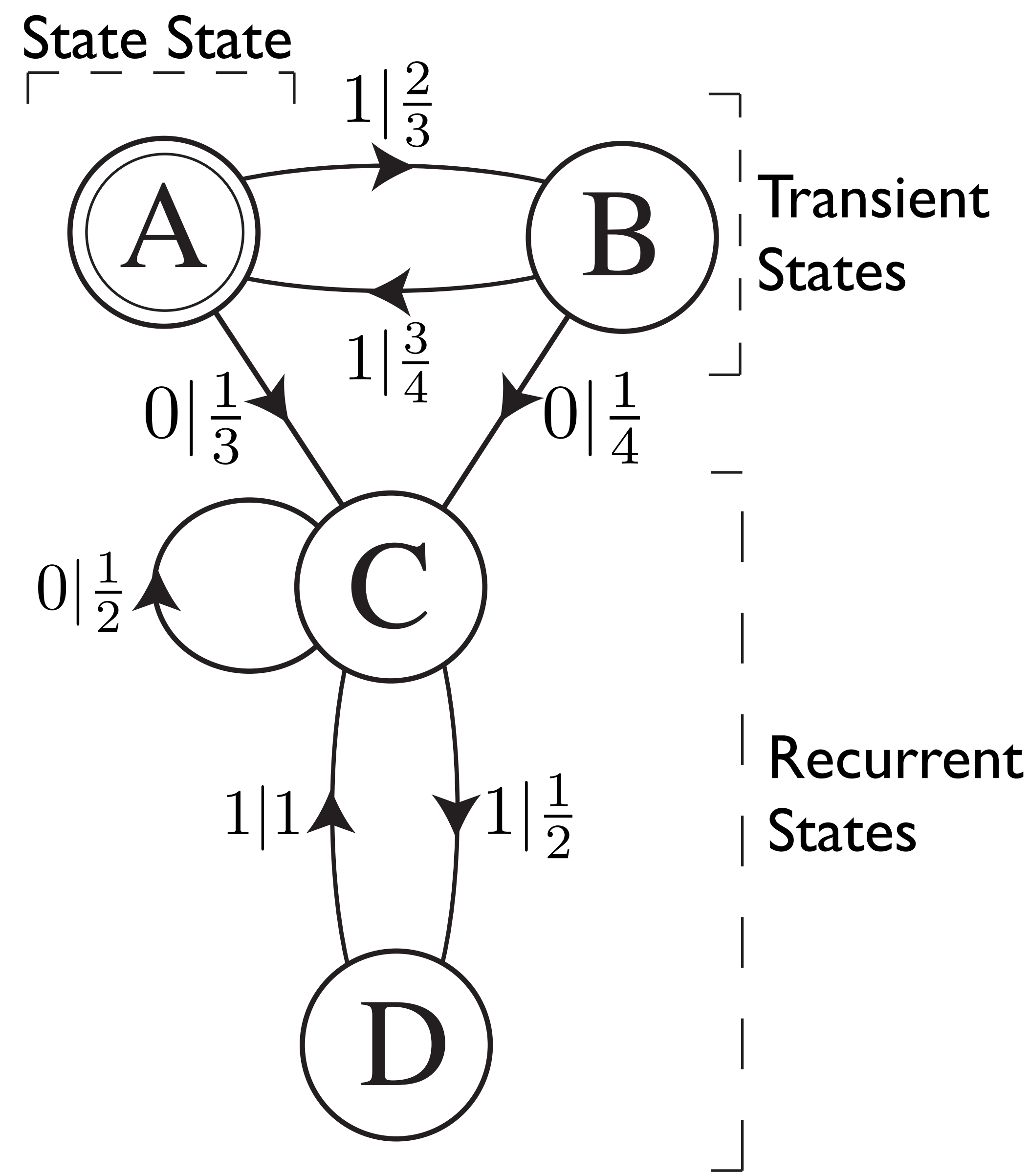
$$\overleftarrow{s} = \lambda$$

Start state:

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

Start state distribution:

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



The ϵ-Machine ...

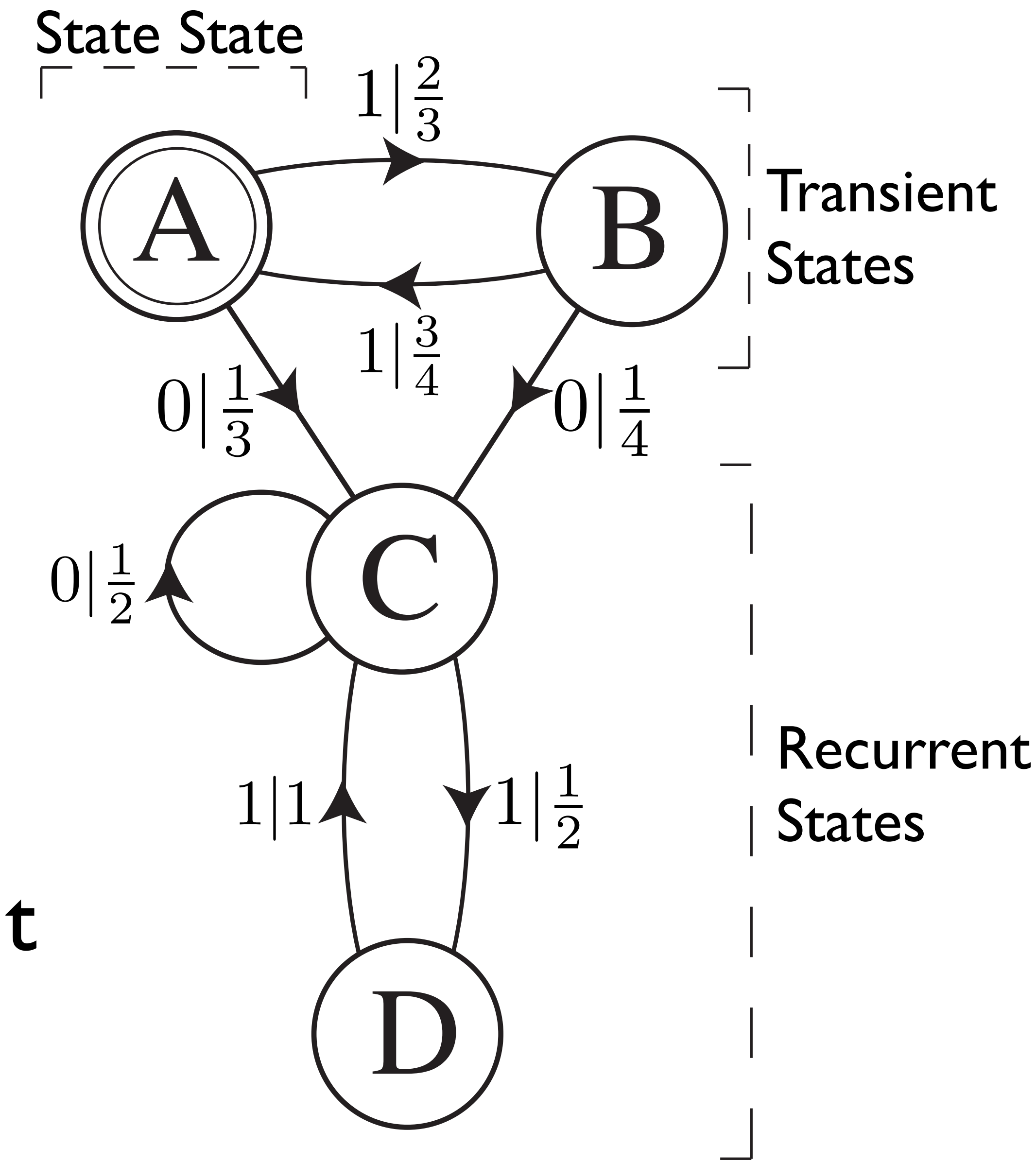
Transient States:

How one comes to know process's recurrent state

Recurrent States:

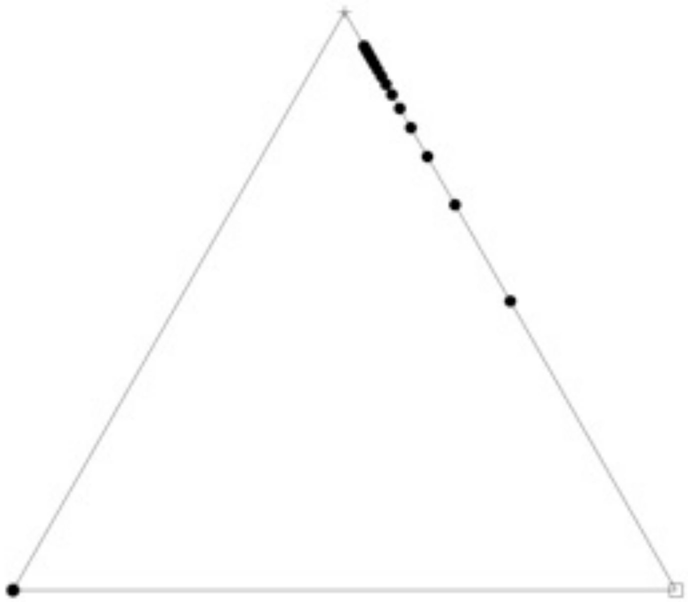
Stationary process:

Only one recurrent component

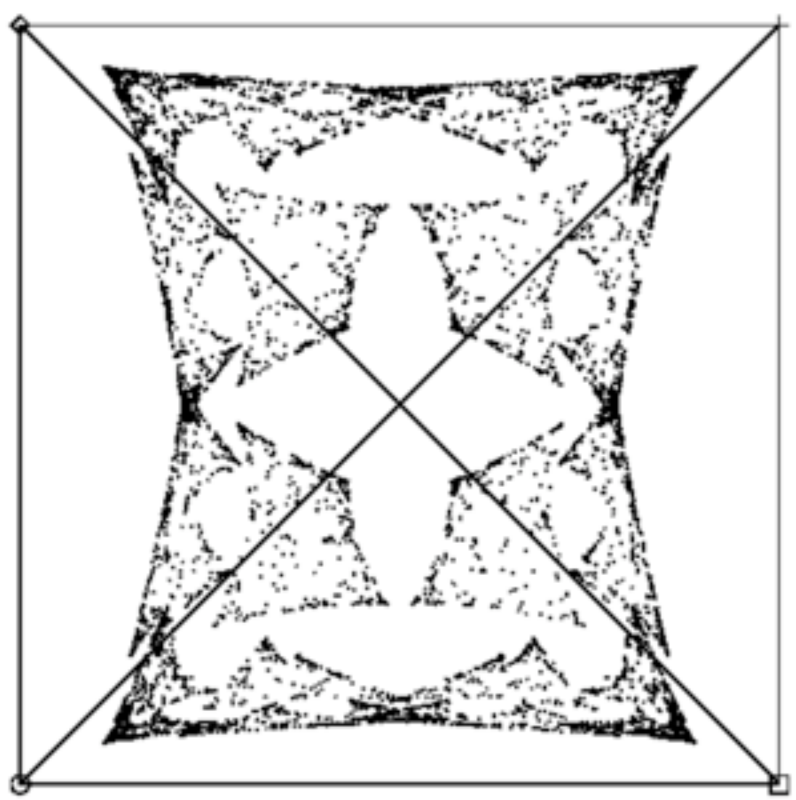


The Learning Channel ...

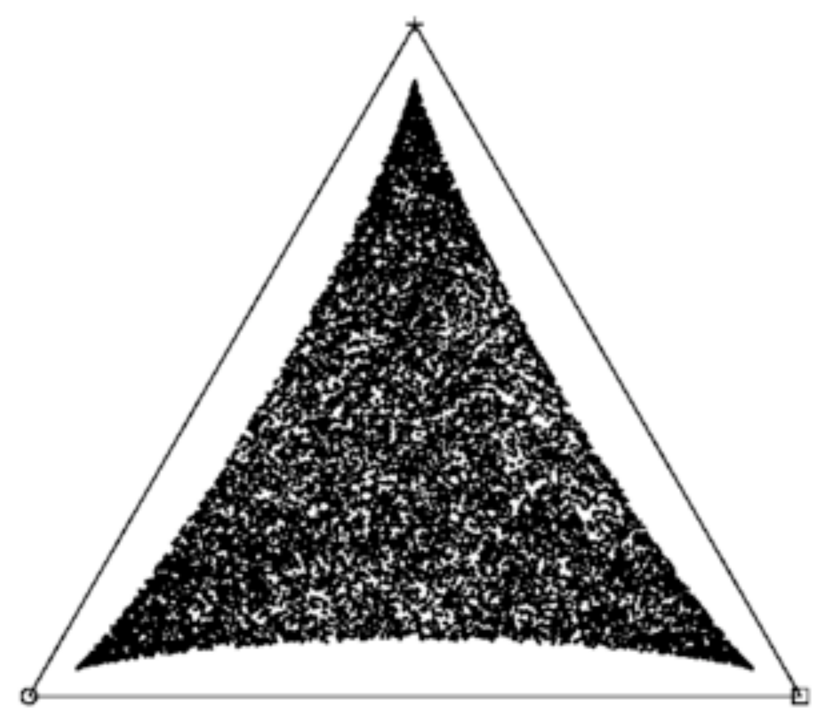
The ϵ-Machine of a Process ...



Denumerable Causal States



Fractal



Continuous

Complexity Lecture 2: Computational Mechanics (CSSS 2011); Jim Crutchfield