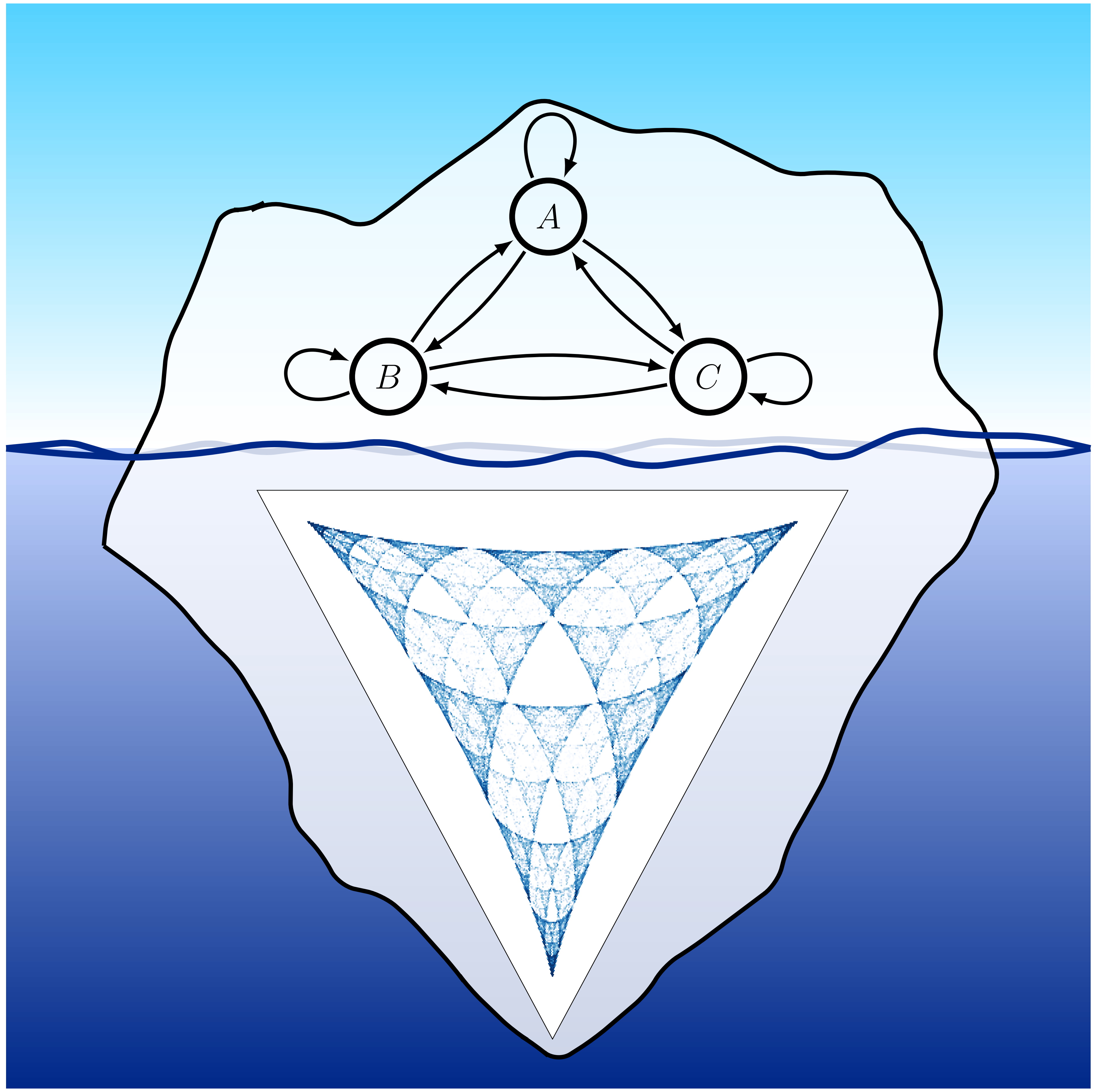
Finite and Infinite Models: Optimal Prediction of Hidden Markov Proceses

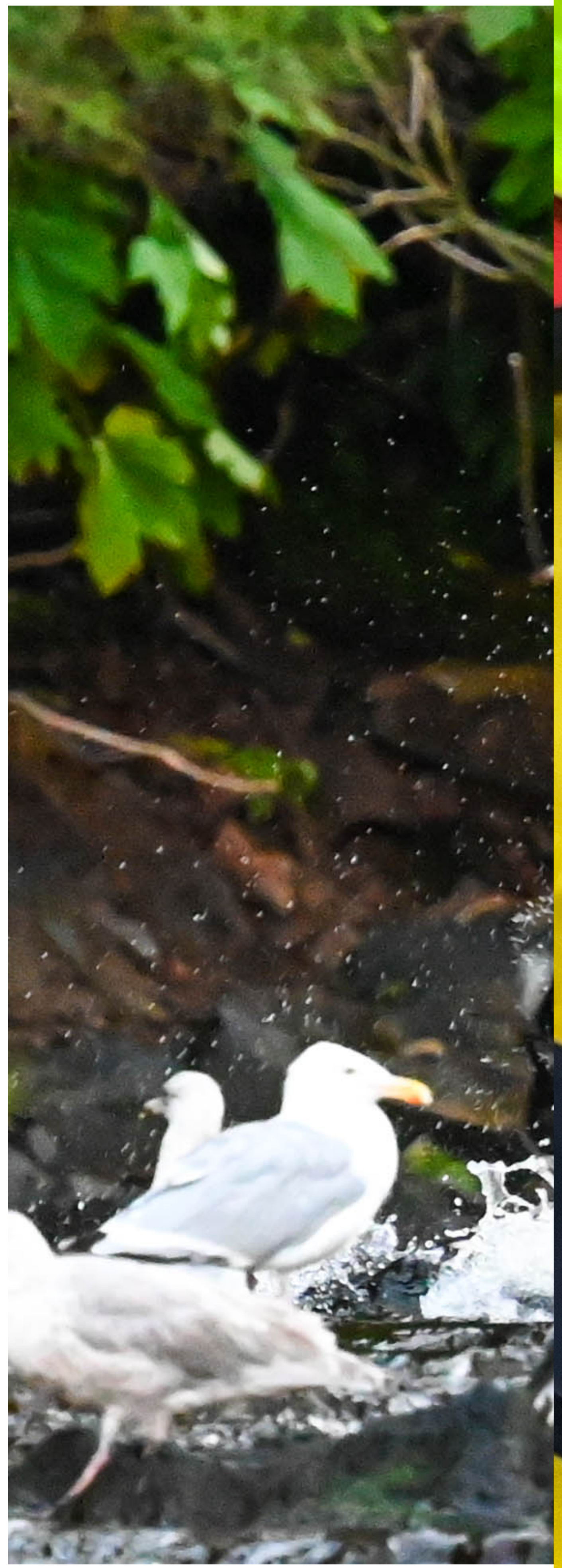


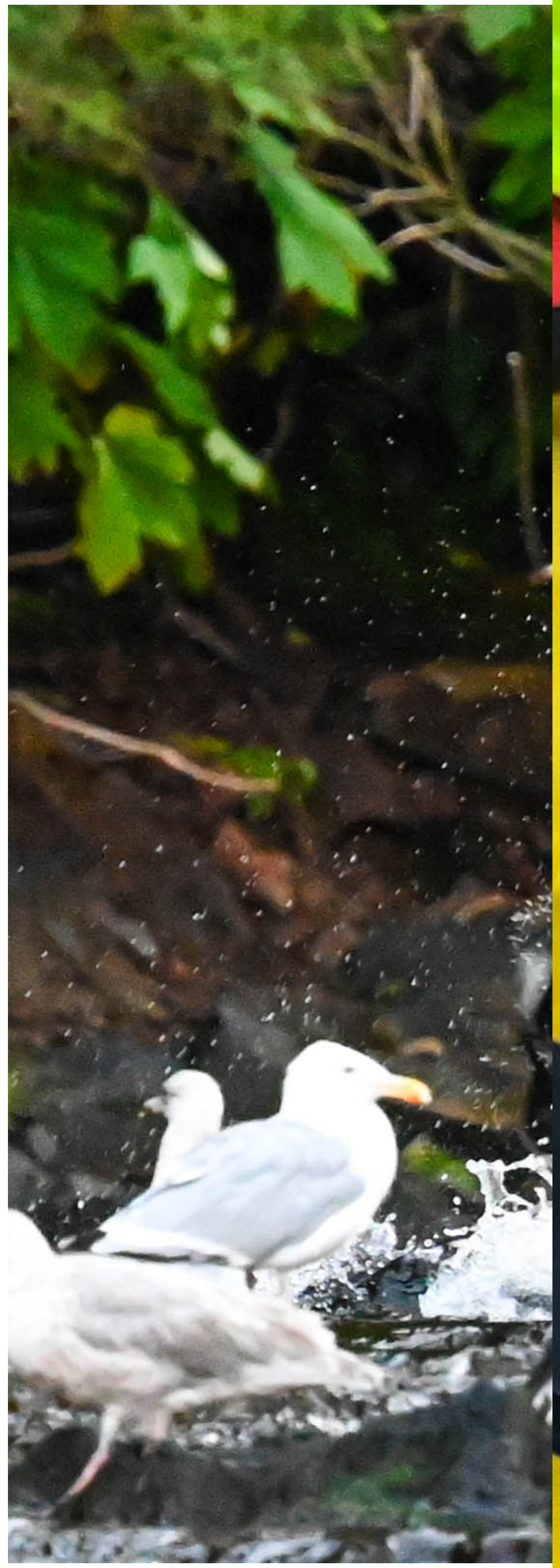
Alexandra M. Jurgens

INRIA Sud Ouest - Bordeaux

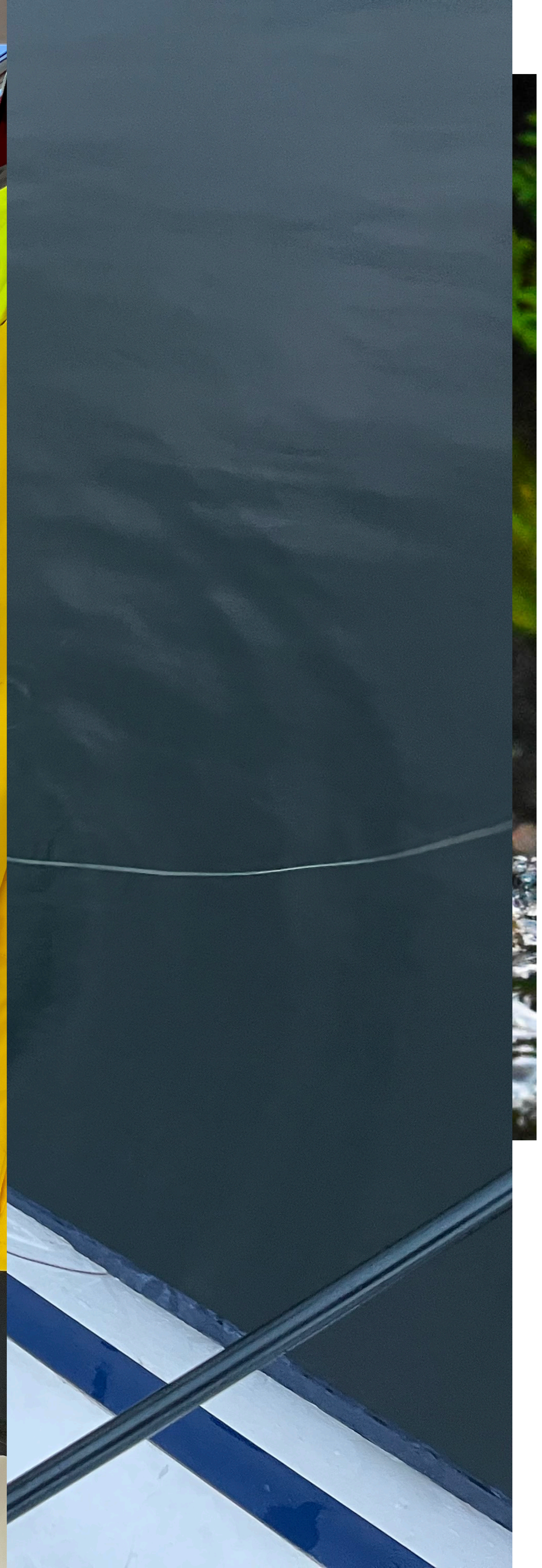
Information Theory as a Bridge Across the Geosciences and Modeling Sciences

11/09/2023






FLYING FISH
VICTORIA



review article

Simple mathematical models with very complicated dynamics

Robert M. May*

First-order difference equations arise in many contexts in the biological, economic and social sciences. Such equations, even though simple and deterministic, can exhibit a surprising array of dynamical behaviour, from stable points, to a bifurcating hierarchy of stable cycles, to apparently random fluctuations. There are consequently many fascinating problems, some concerned with delicate mathematical aspects of the fine structure of the trajectories, and some concerned with the practical implications and applications. This is an interpretive review of them.

Nature Vol. 261 June 10 1976 459

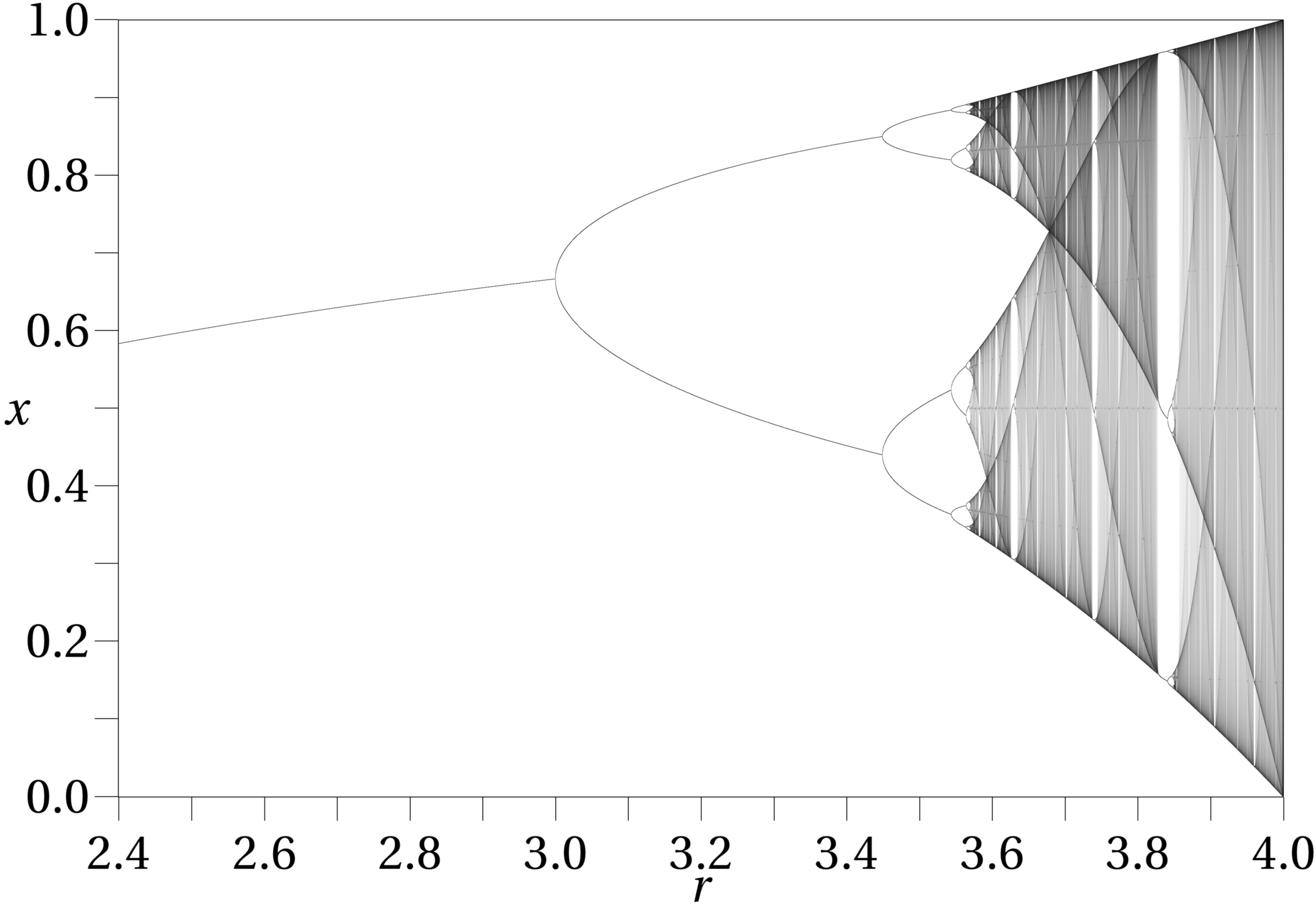
review article

Simple mathematical models with very complicated dynamics

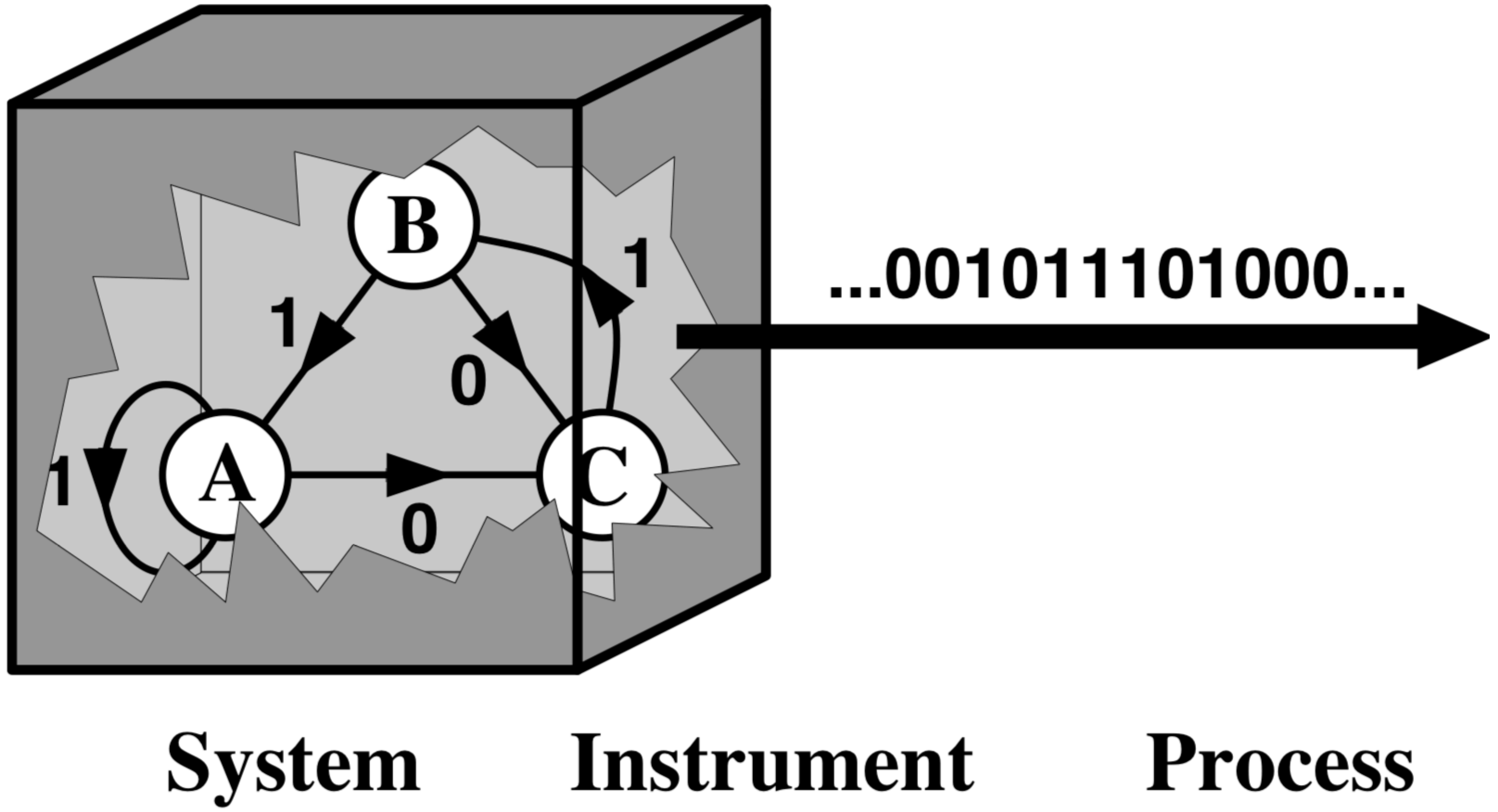
Robert M. May*

First-order difference equations arise in many contexts in the biological, economic and social sciences. Such equations, even though simple and deterministic, can exhibit a surprising array of dynamical behaviour, from stable points, to a bifurcating hierarchy of stable cycles, to apparently random fluctuations. There are consequently many fascinating problems, some concerned with delicate mathematical aspects of the fine structure of the trajectories, and some concerned with the practical implications and applications. This is an interpretive review of them.

$$x_{n+1} = rx_n\left(1 - x_n\right)$$



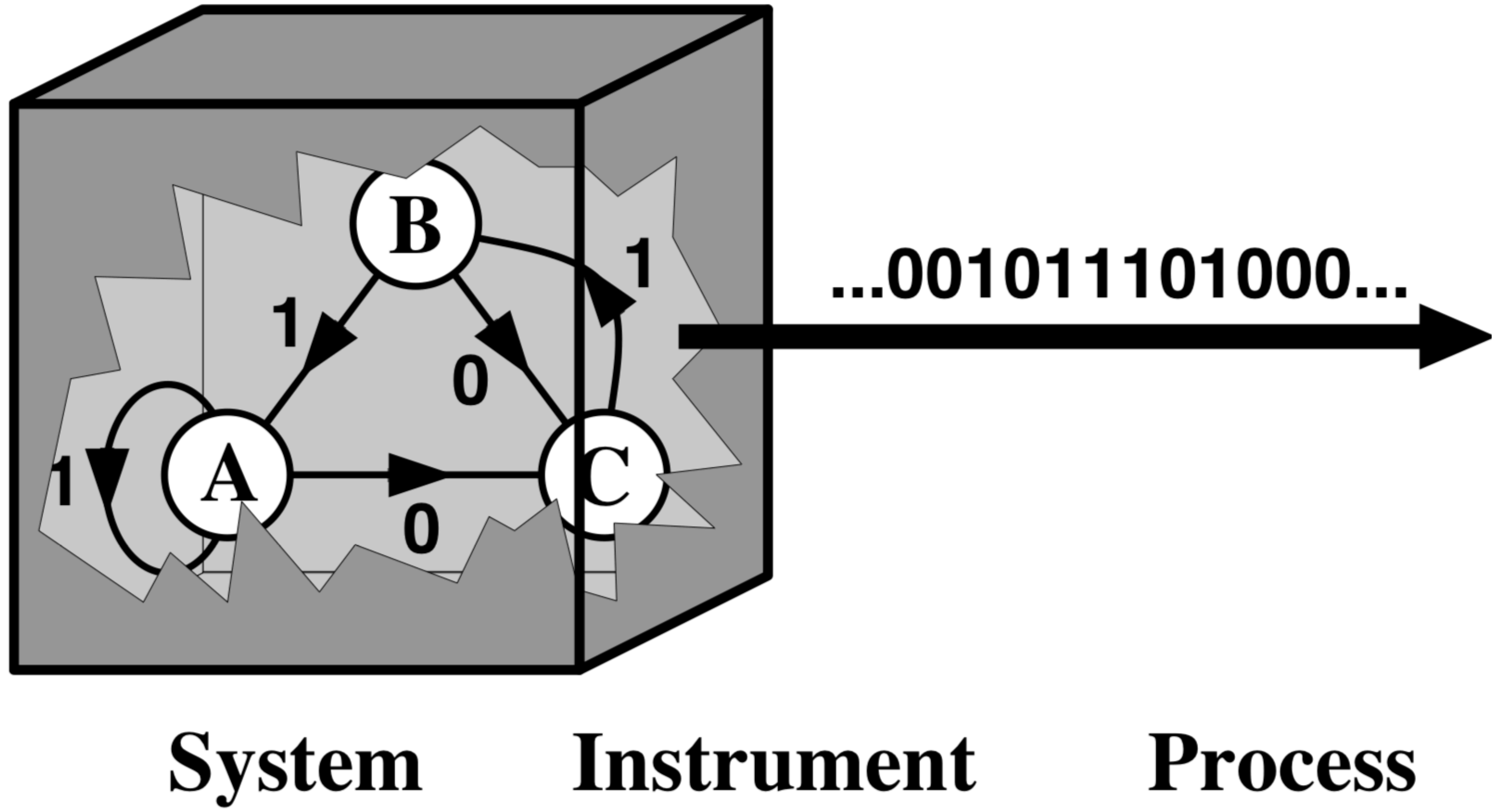
Object of Study: Processes



The random variables X_i may take on values in alphabet A:

$$\overleftrightarrow{X} = \ldots X_{-1} X_0 X_1 \ldots$$

Object of Study: Processes

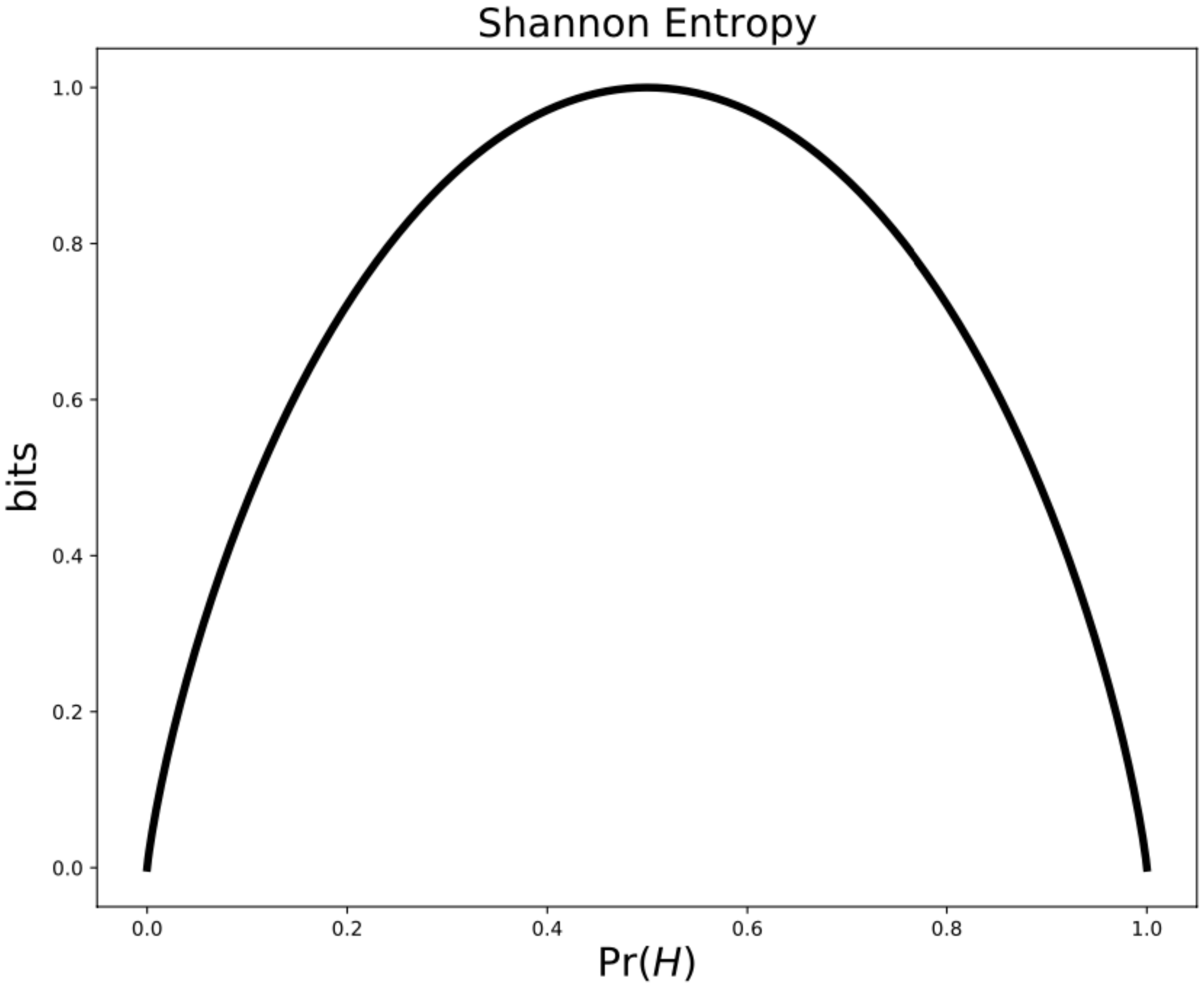


Let $A = \{0,1\}$. Then a *realization* of the process is written

$$\overleftrightarrow{x} = \ldots x_{-1} x_0 x_1 \ldots$$
$$= \ldots 011 \ldots$$

A process P is defined as the probability distribution over bi-infinite strings.

Information Theory



The *Shannon entropy* over a random variable is defined:

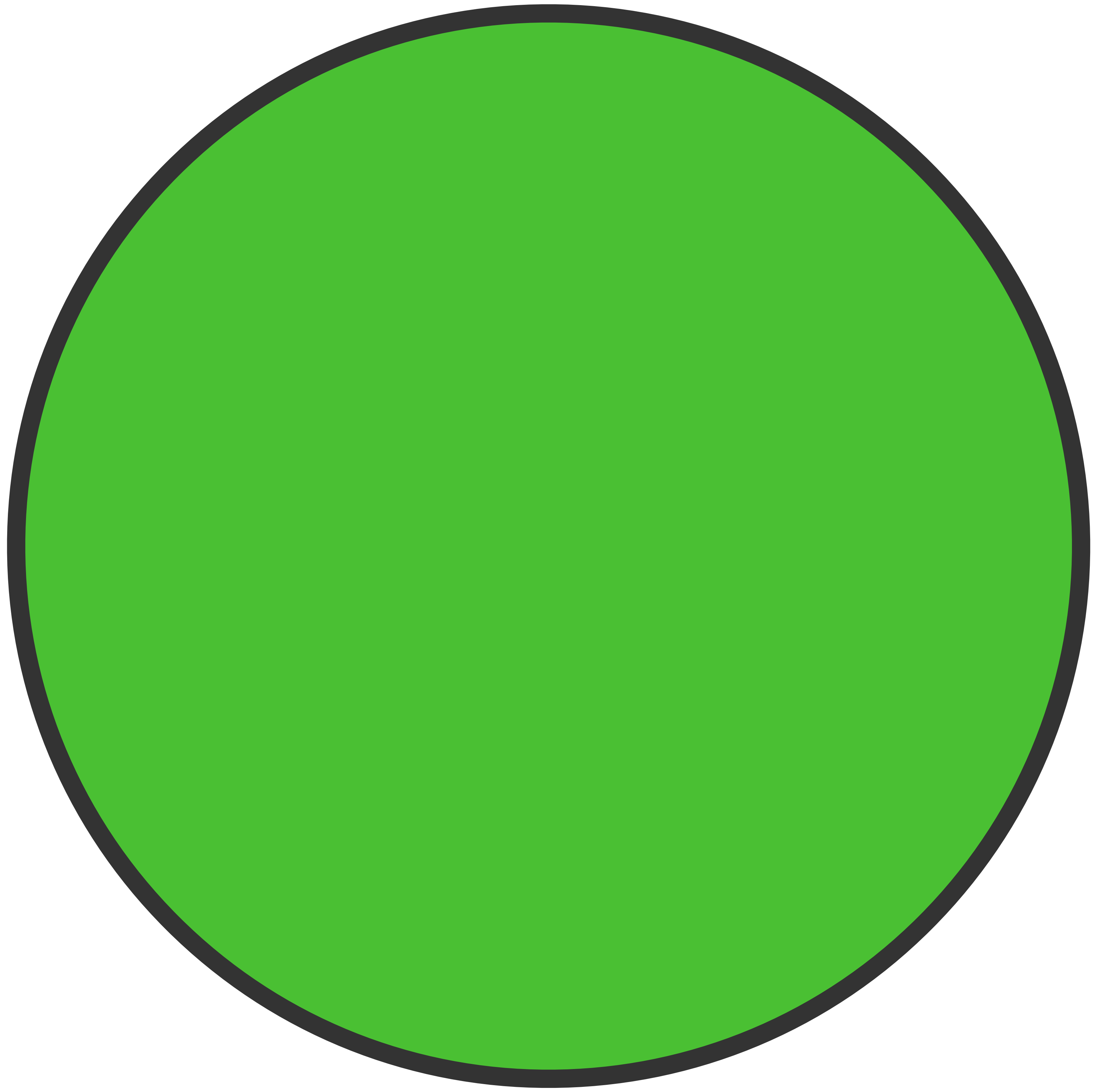
$$H[X] = -\sum_{x \in A} \Pr(X = x)\ \log_2 \Pr(X = x)$$

Claude E. Shannon. *A Mathematical Theory of Communication.* Bell System Technical Journal. **27** (3): 379–423. (1948)

Information Theory

$$H[X] = -\sum_{x \in A} \Pr(X = x)\, \log_2 \Pr(X = x)$$

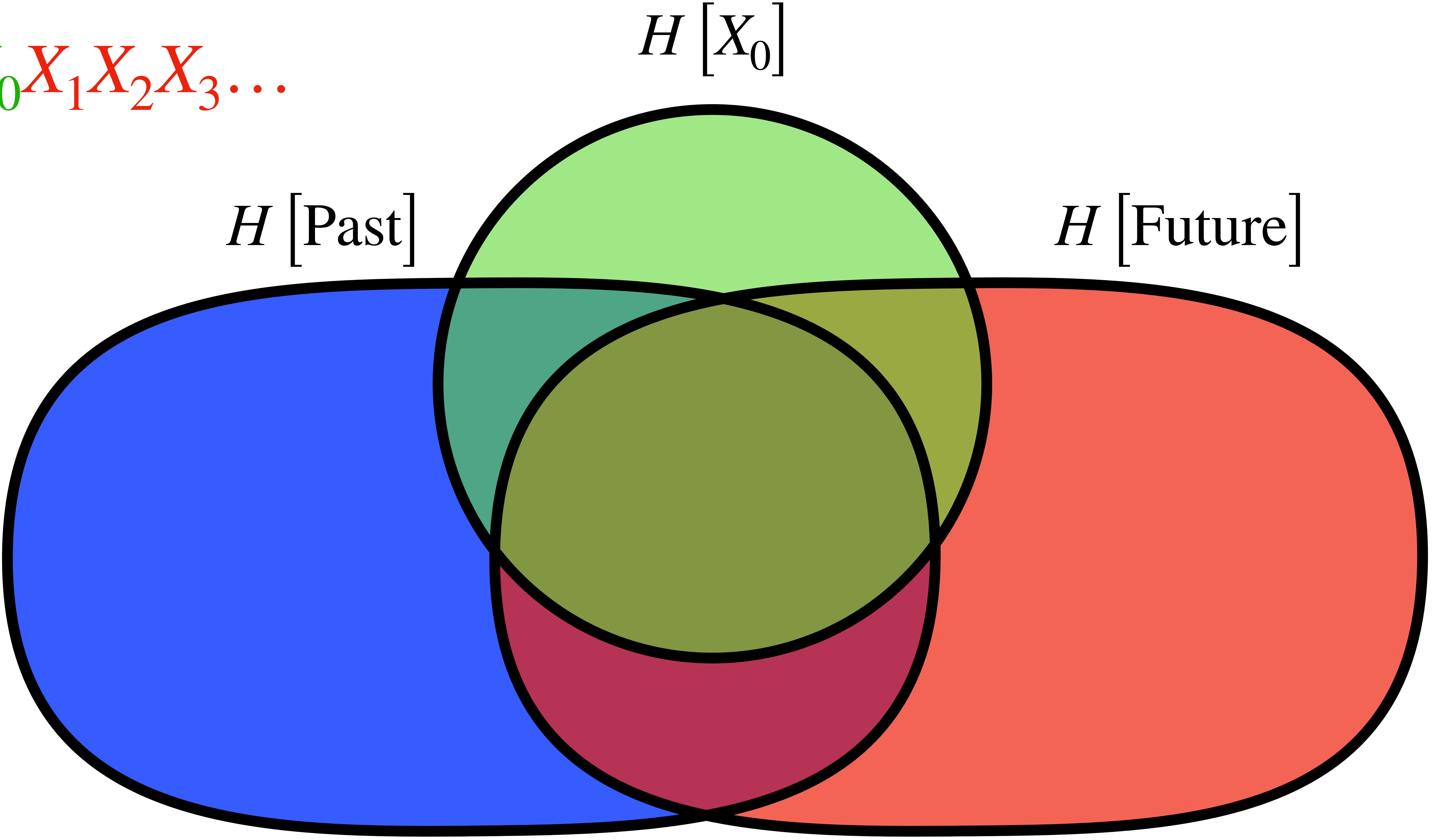
$H\left[X_0\right]$



Claude E. Shannon. *A Mathematical Theory of Communication.* Bell System Technical Journal. **27** (3): 379–423. (1948)

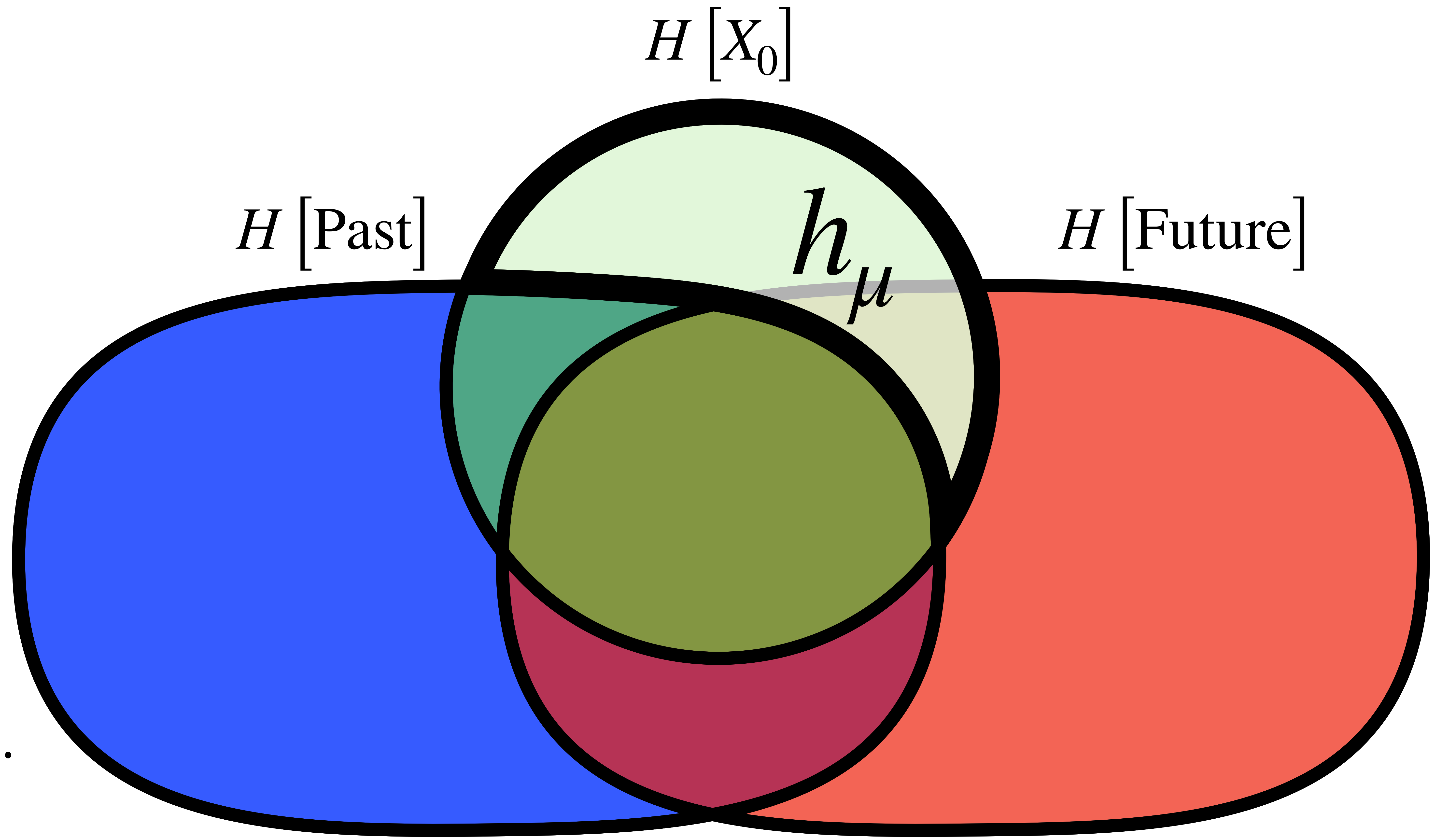
Information Theory

$$\overleftrightarrow{X} = \ldots X_{-3}X_{-2}X_{-1}X_0X_1X_2X_3\ldots$$



Entropy Rate

$$h_\mu = \lim_{L \to \infty} H\left[X_0 \mid \text{Past}\right]$$
$$= \lim_{L \to \infty} \left(H\left[X_0, \text{Past}\right] - H\left[\text{Past}\right] \right)$$

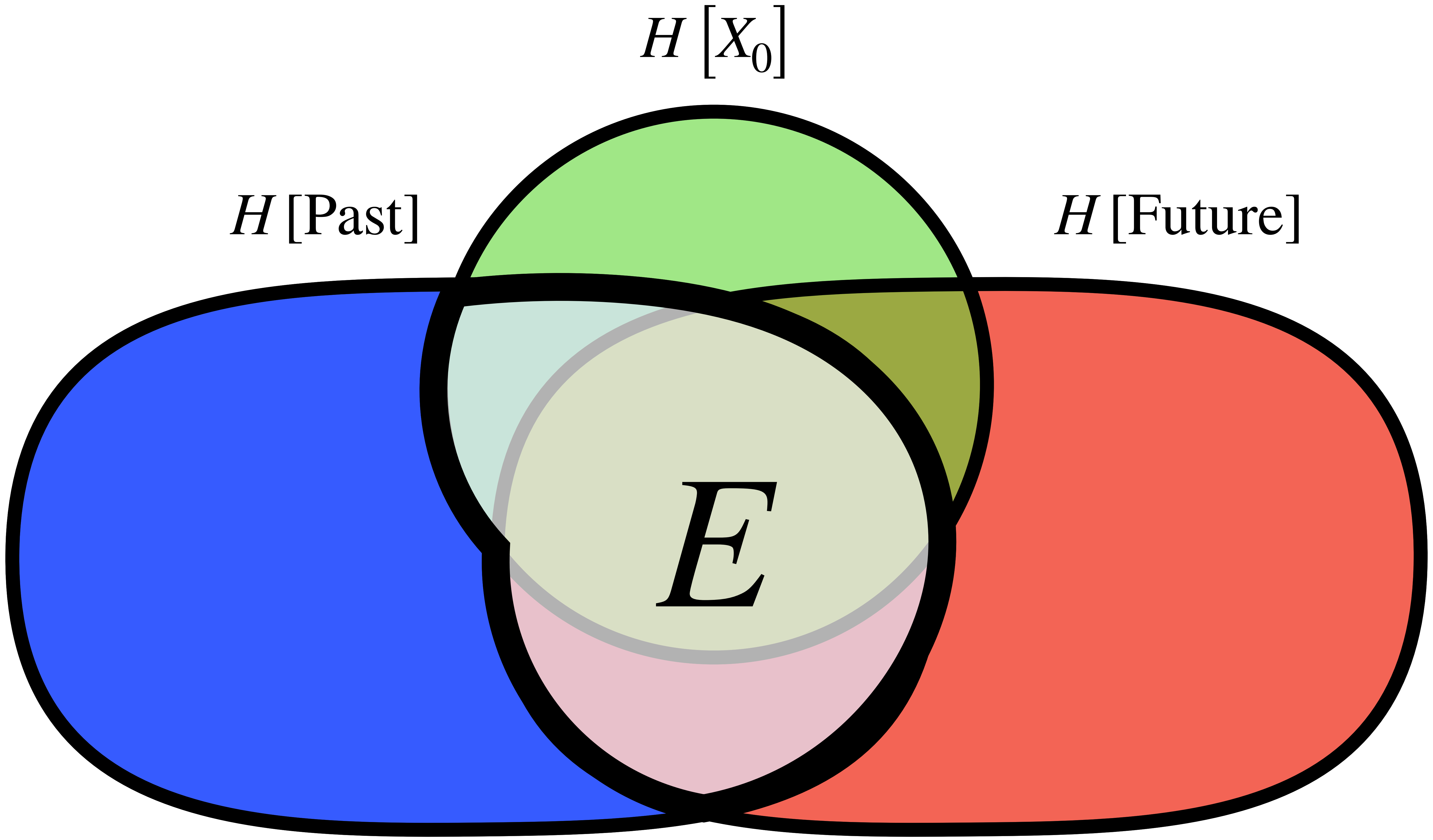


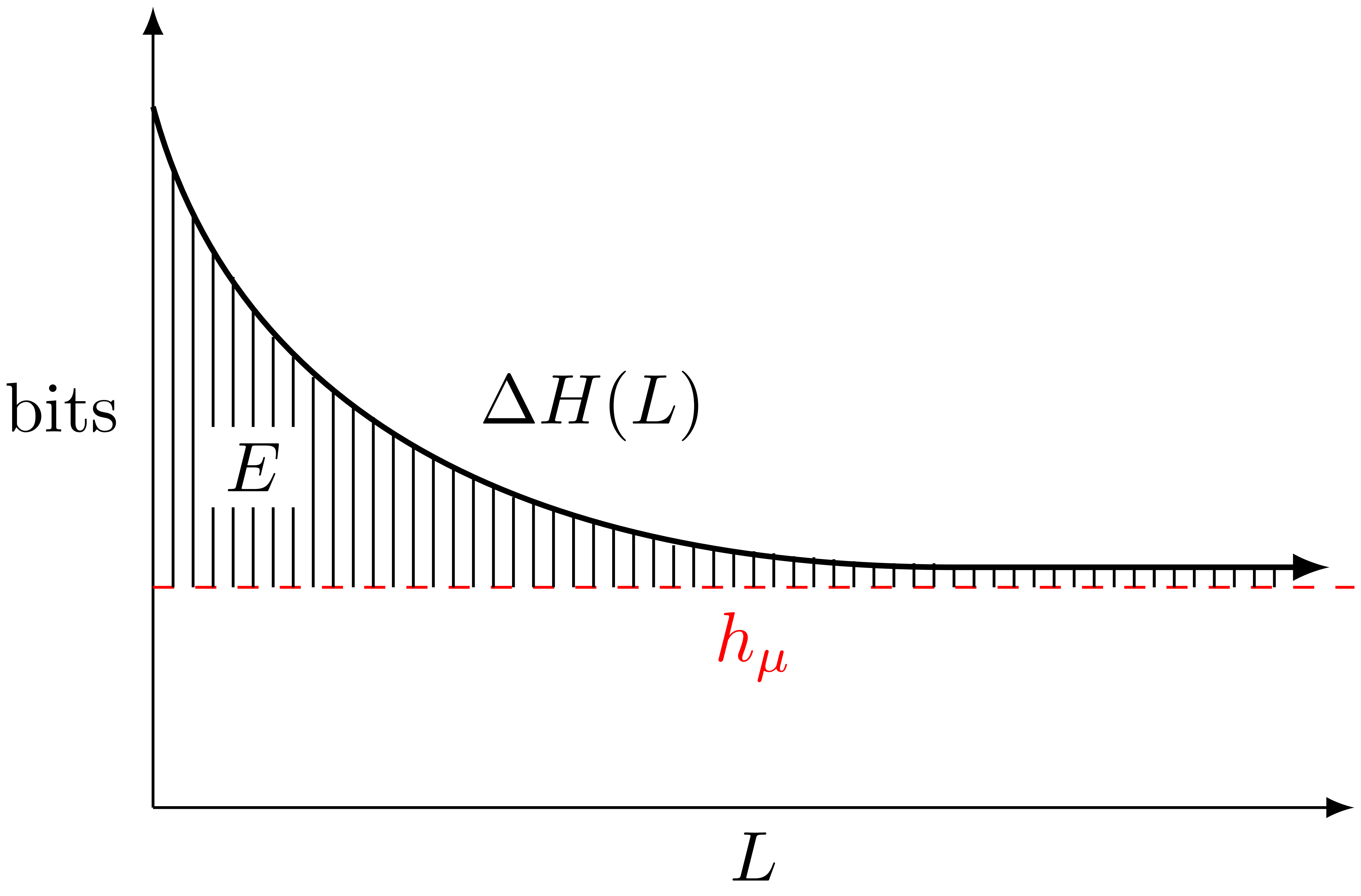
h_μ is the *irreducible randomness* of a process.

This is the limit of our predictive abilities.

James P. Crutchfield, David P. Feldman. *Regularities unseen, randomness observed: Levels of entropy convergence.* Chaos 1 March 2003; 13 (1): 25–54.

Excess Entropy

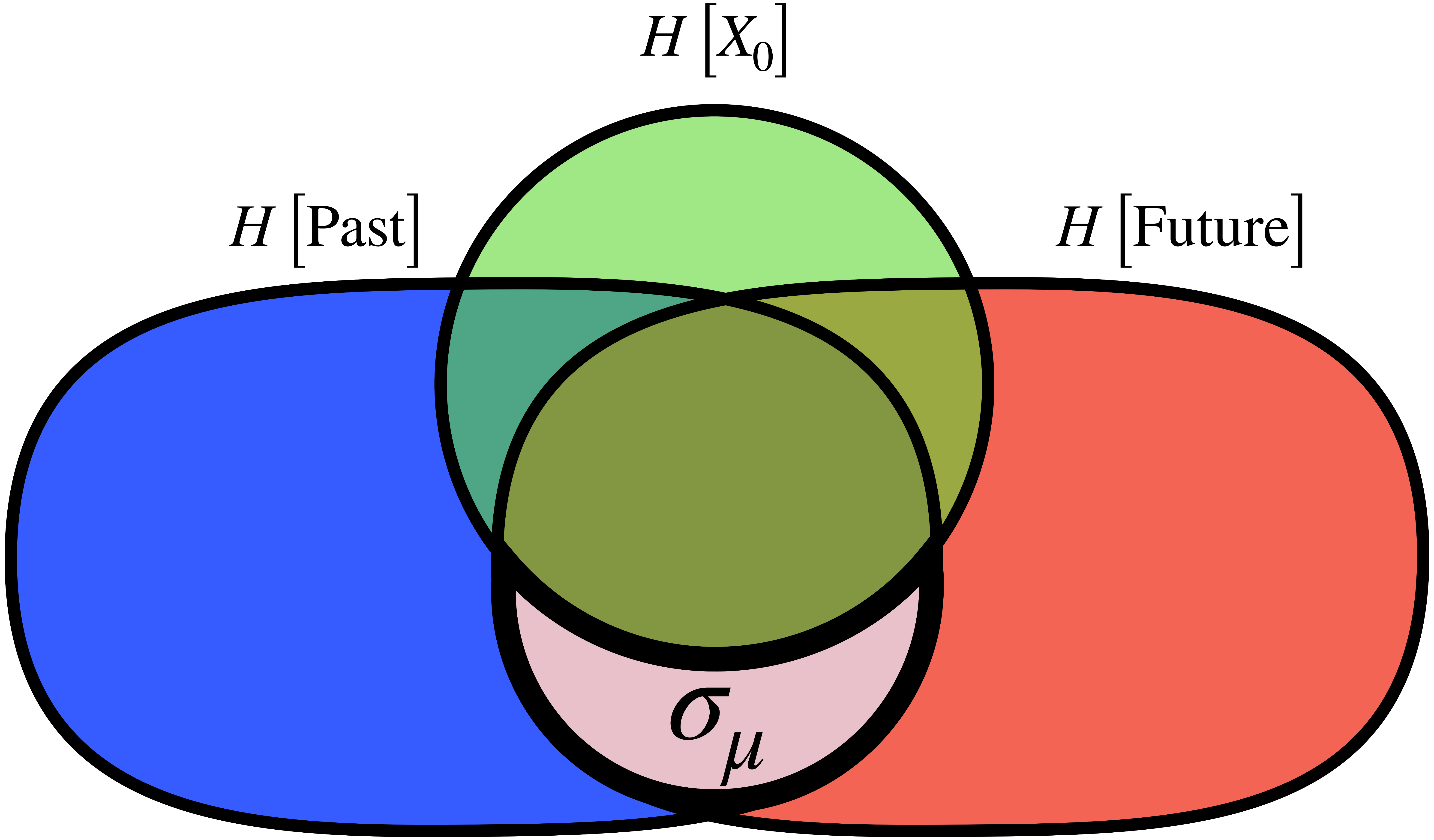




Excess entropy: information shared between the past and the present + future.

James P. Crutchfield, David P. Feldman. *Regularities unseen, randomness observed: Levels of entropy convergence.* Chaos 1 March 2003; 13 (1): 25–54.

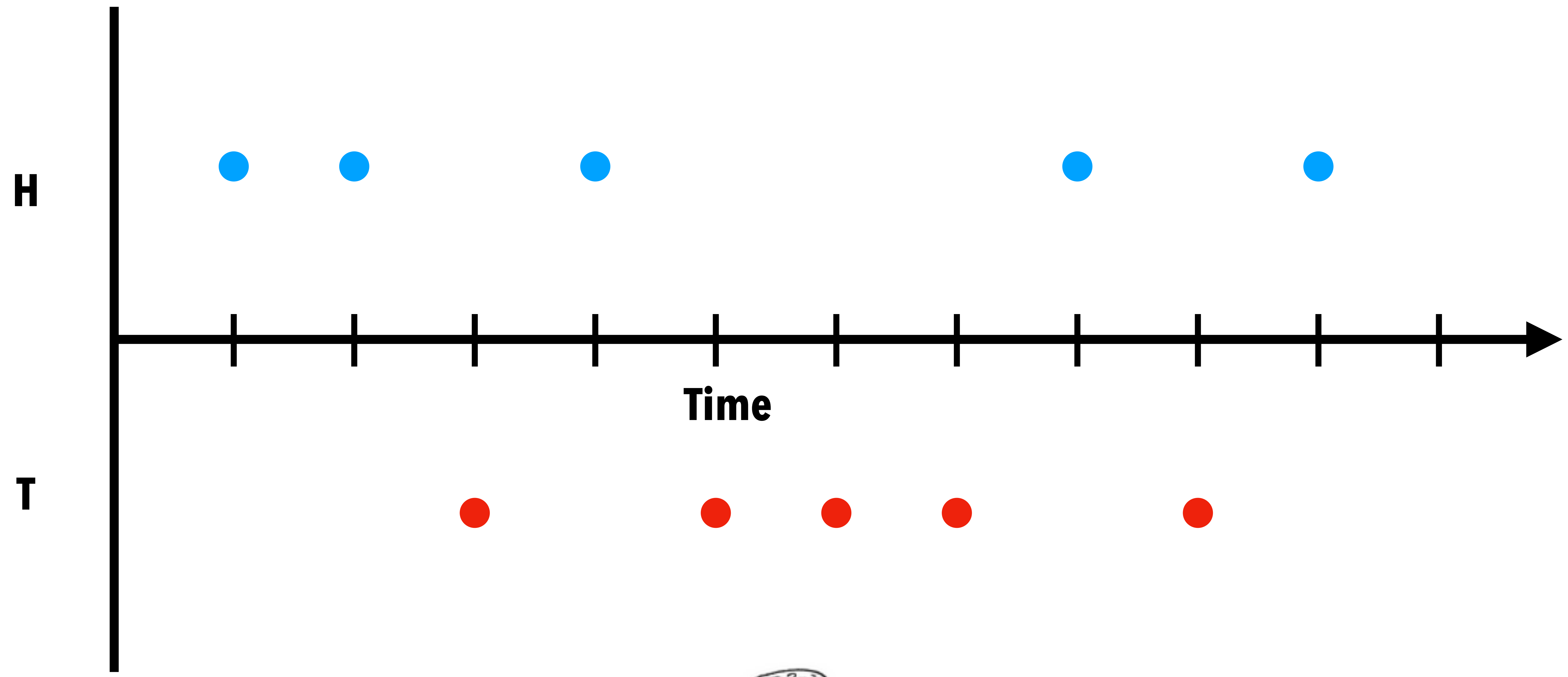
Elusive Information



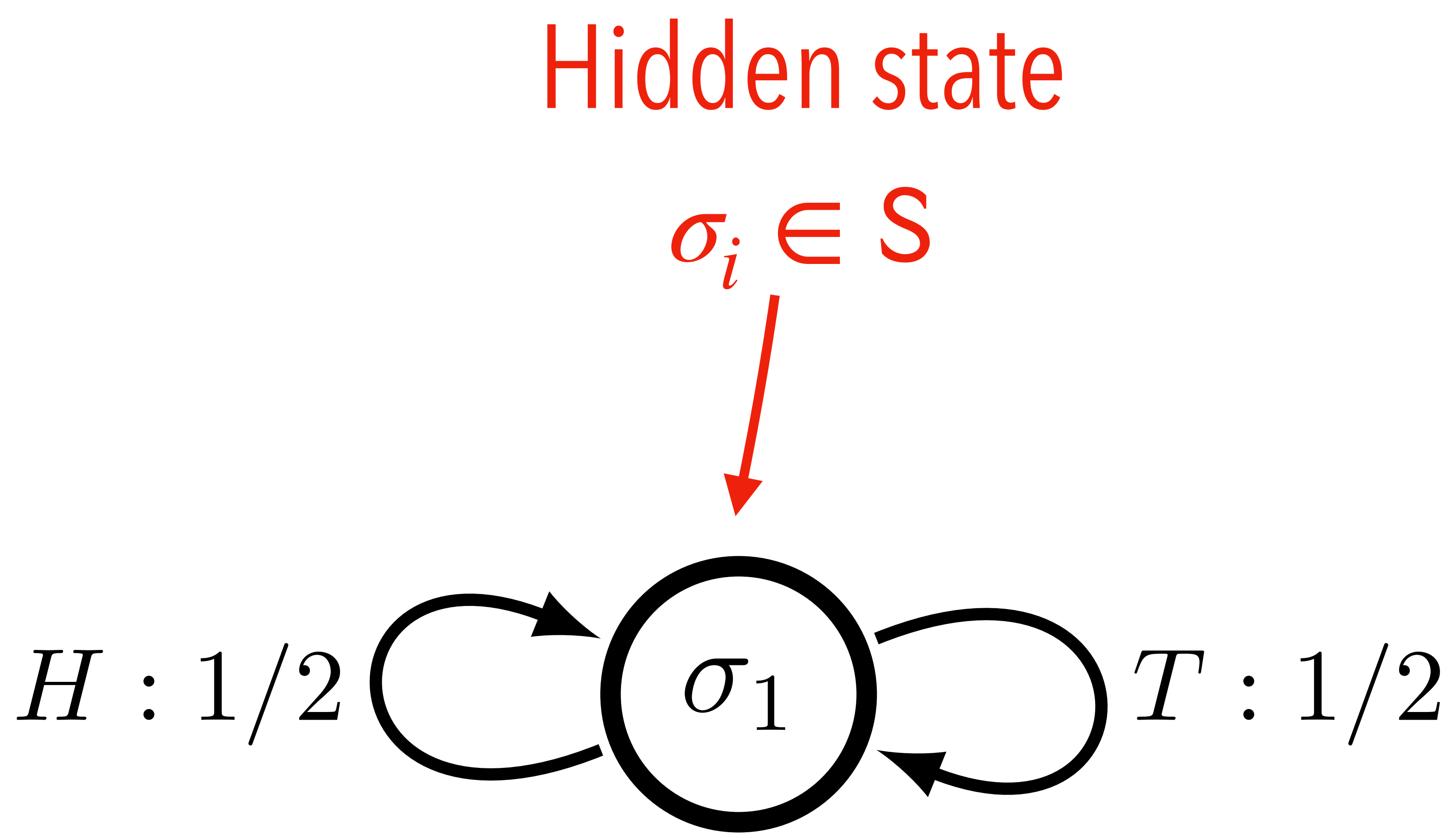
Ryan G. James, Christopher J. Ellison, James P. Crutchfield. *Anatomy of a bit: Information in a time series observation.* Chaos 21, 037109 (2011)

James P. Crutchfield, David P. Feldman. *Regularities unseen, randomness observed: Levels of entropy convergence.* Chaos 1 March 2003; 13 (1): 25–54.

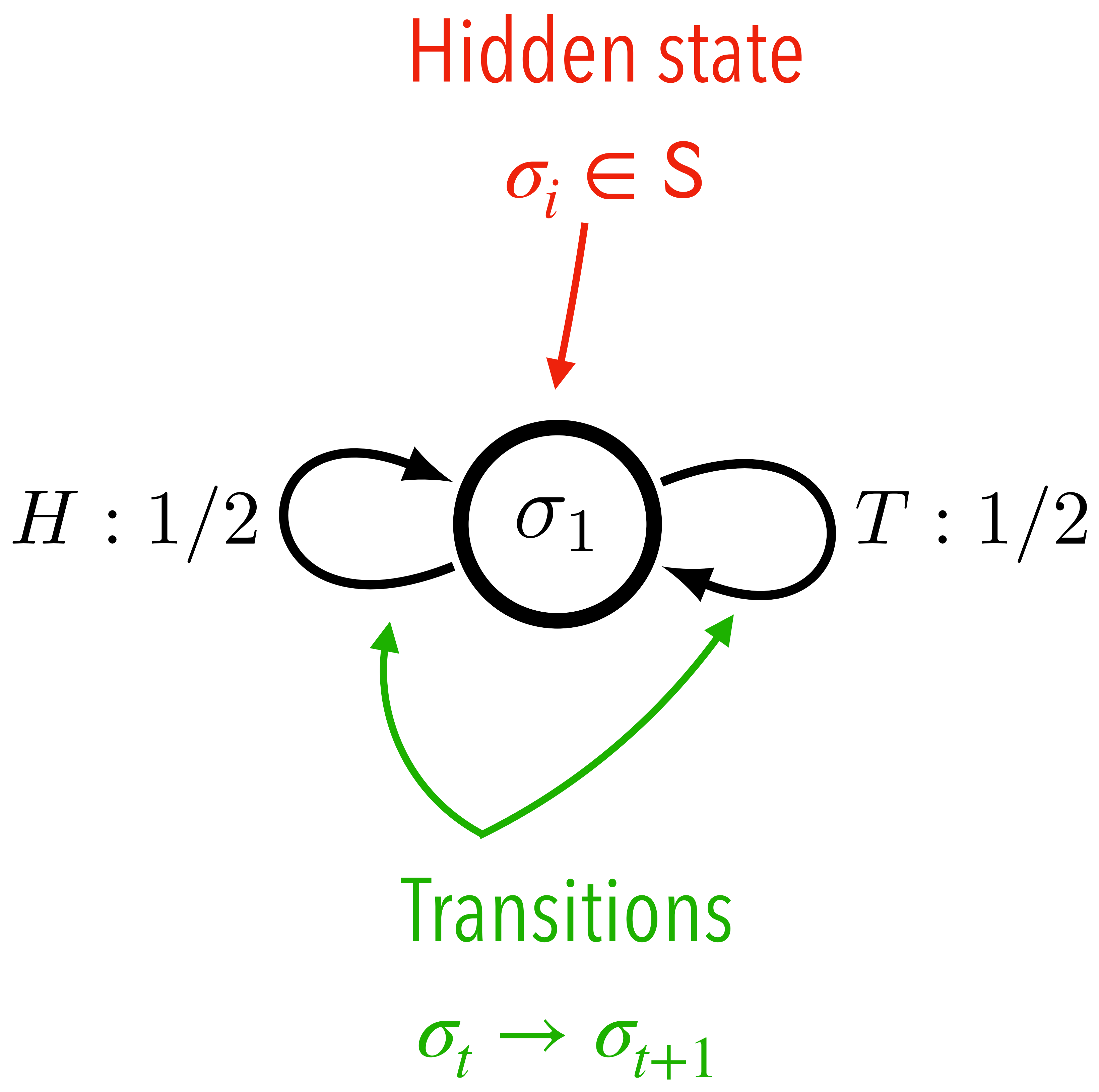
Models of Time Series



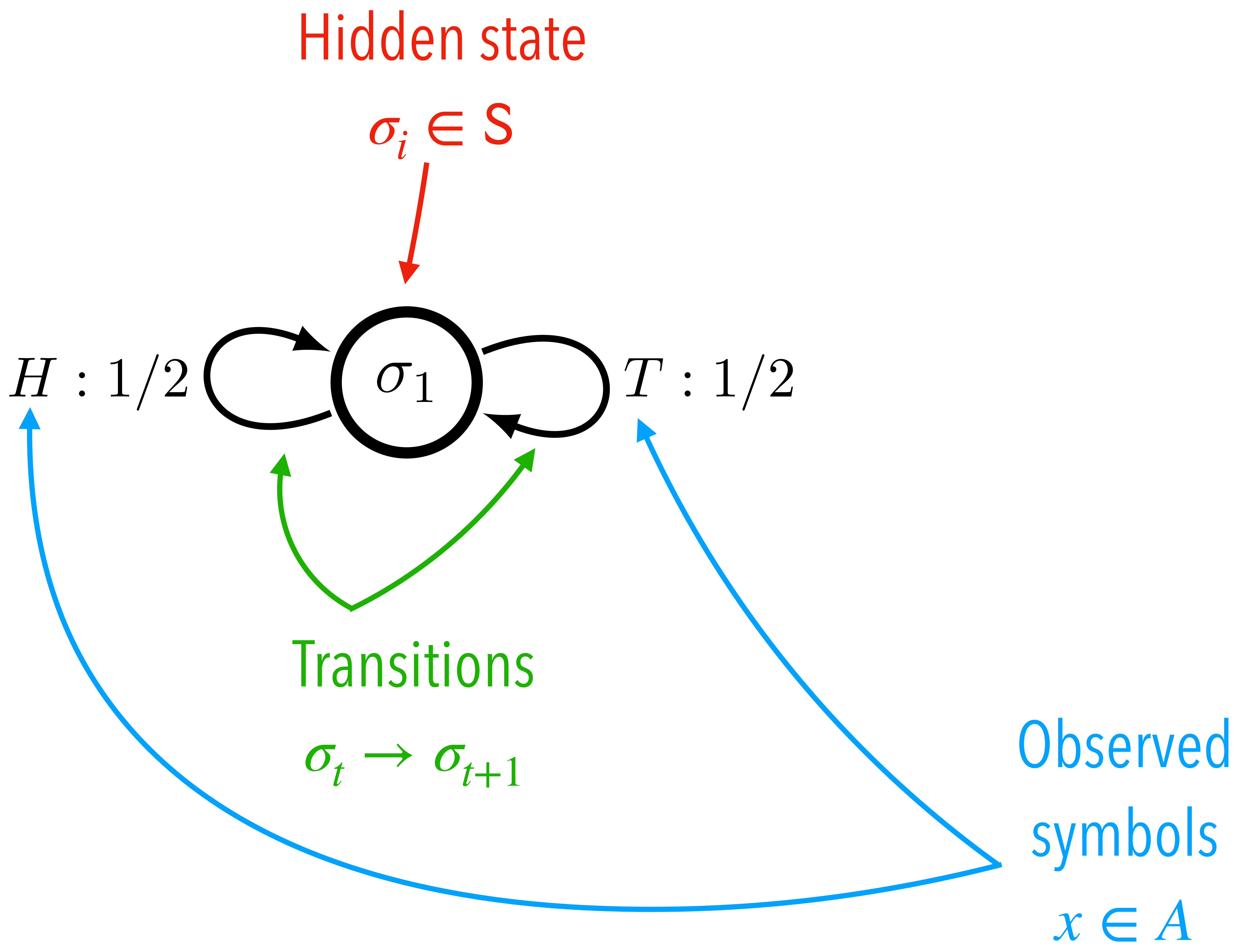
Models of Time Series



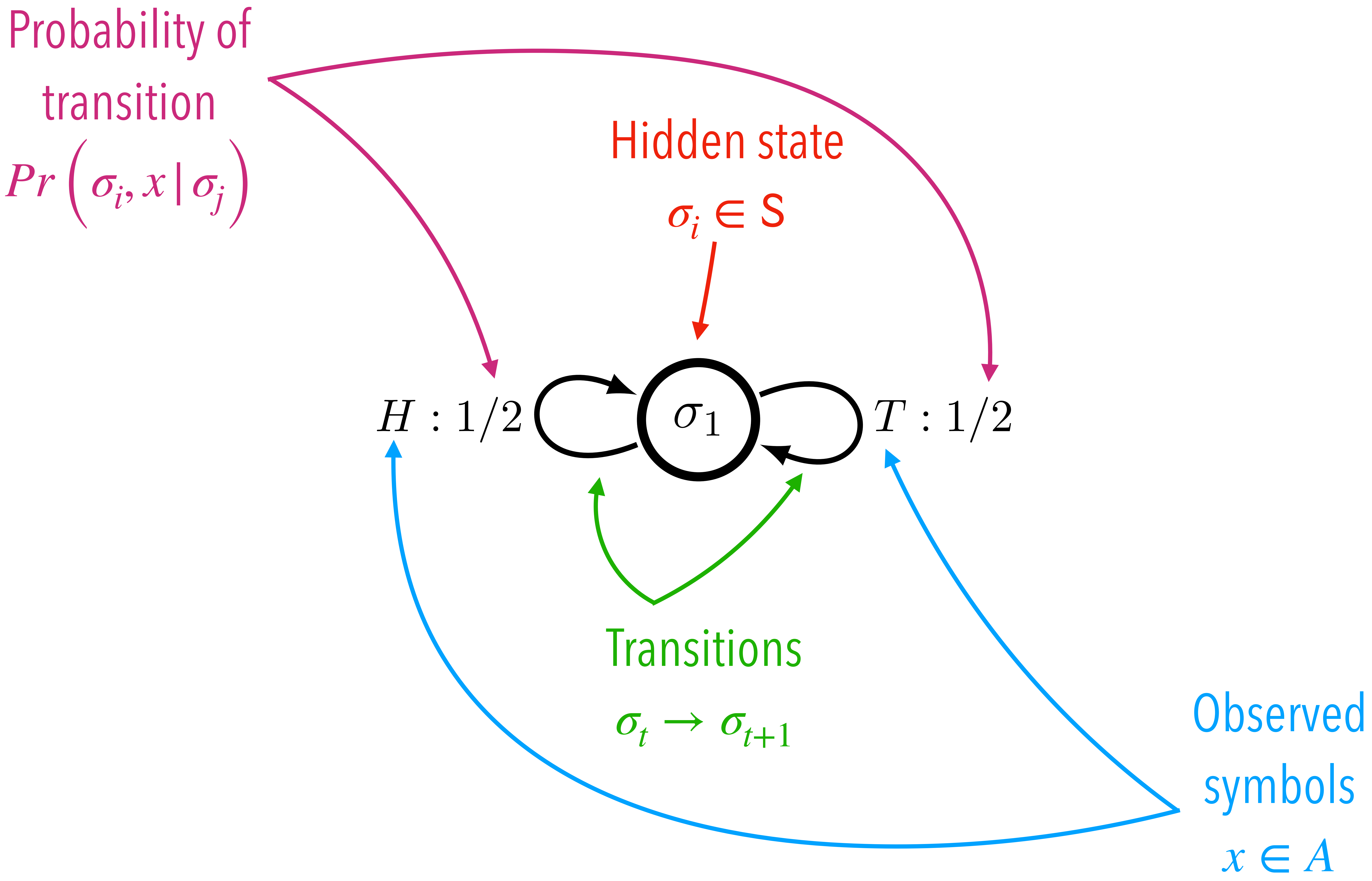
Models of Time Series



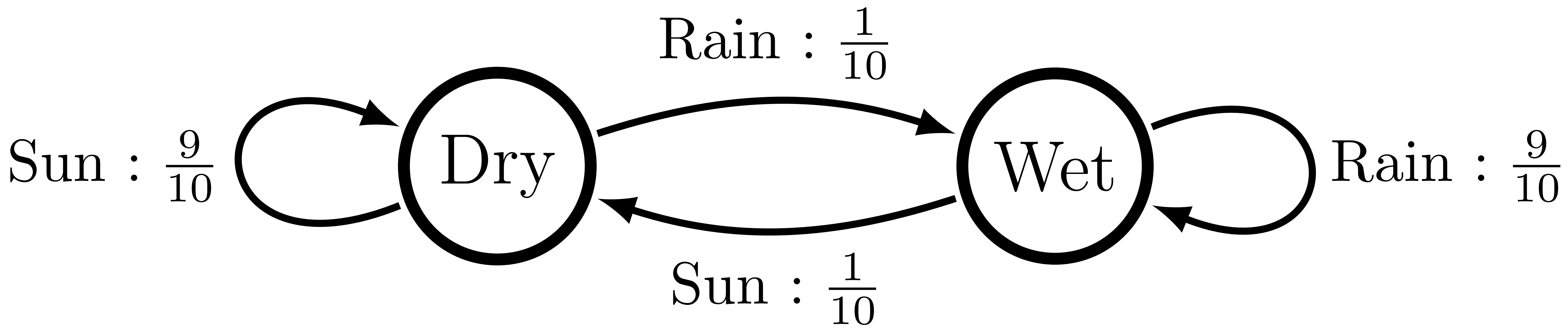
Models of Time Series



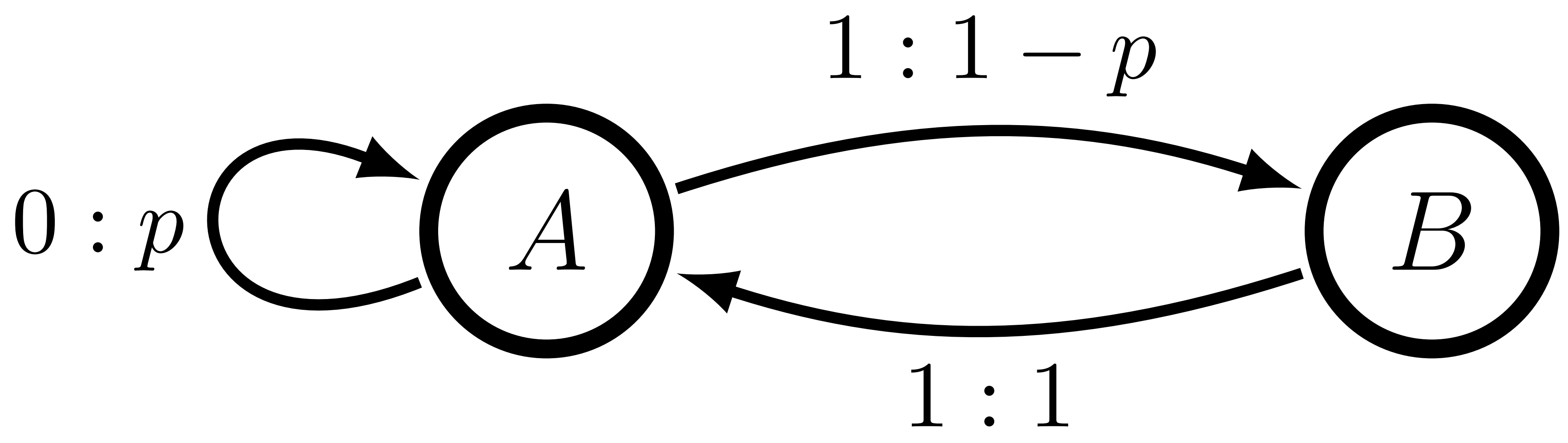
Models of Time Series



Hidden Markov Models

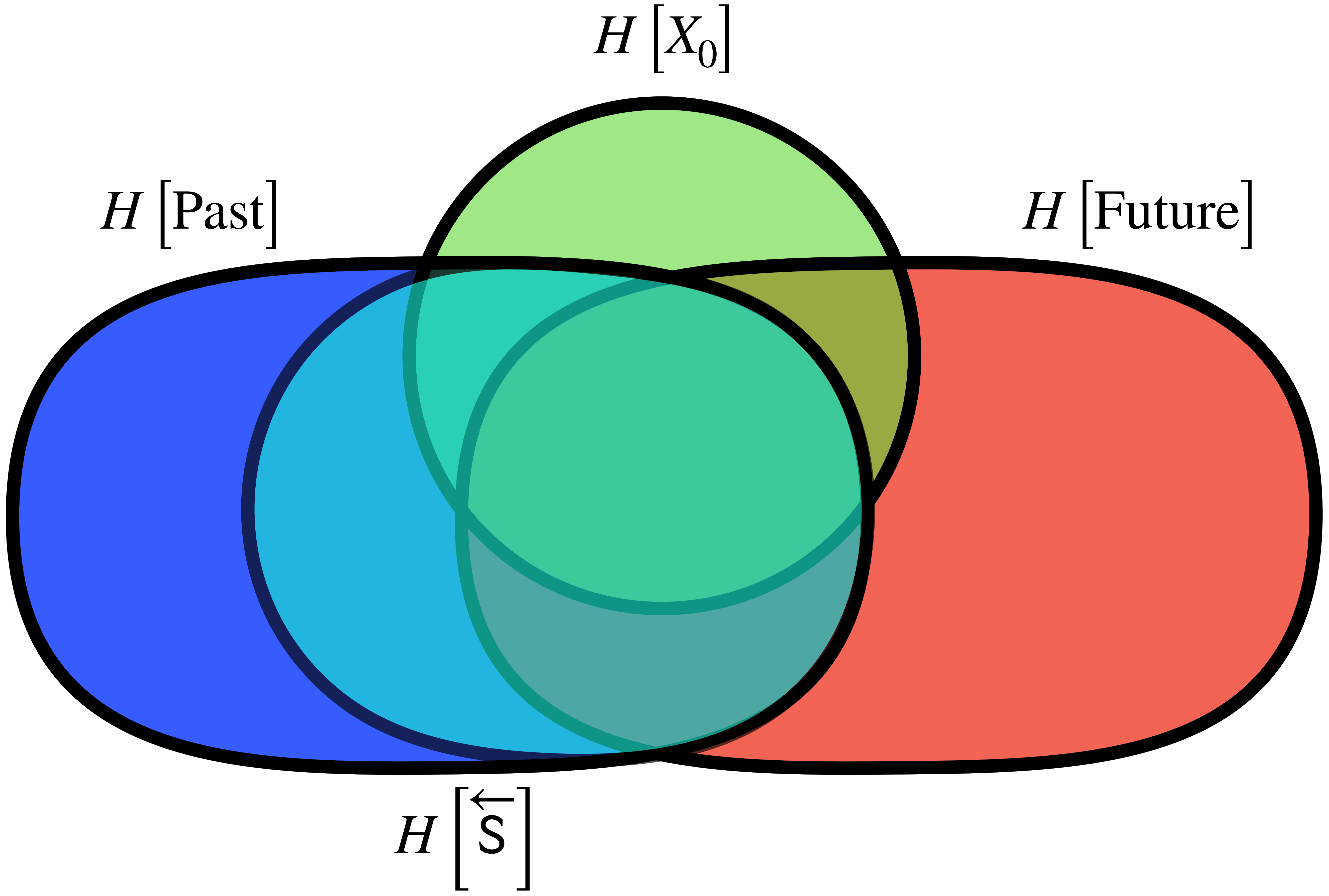


The Utility of (Good) Models



$$h_\mu = \sum_{\sigma \in S} \Pr(\sigma)\, H\left[X \mid \sigma\right]$$

The Utility of (Good) Models



- States are a function of the past
- Optimal predictor
- Minimal in size
- Unique

Ryan G. James, Christopher J. Ellison, James P. Crutchfield. *Anatomy of a bit: Information in a time series observation.* Chaos 21, 037109 (2011)

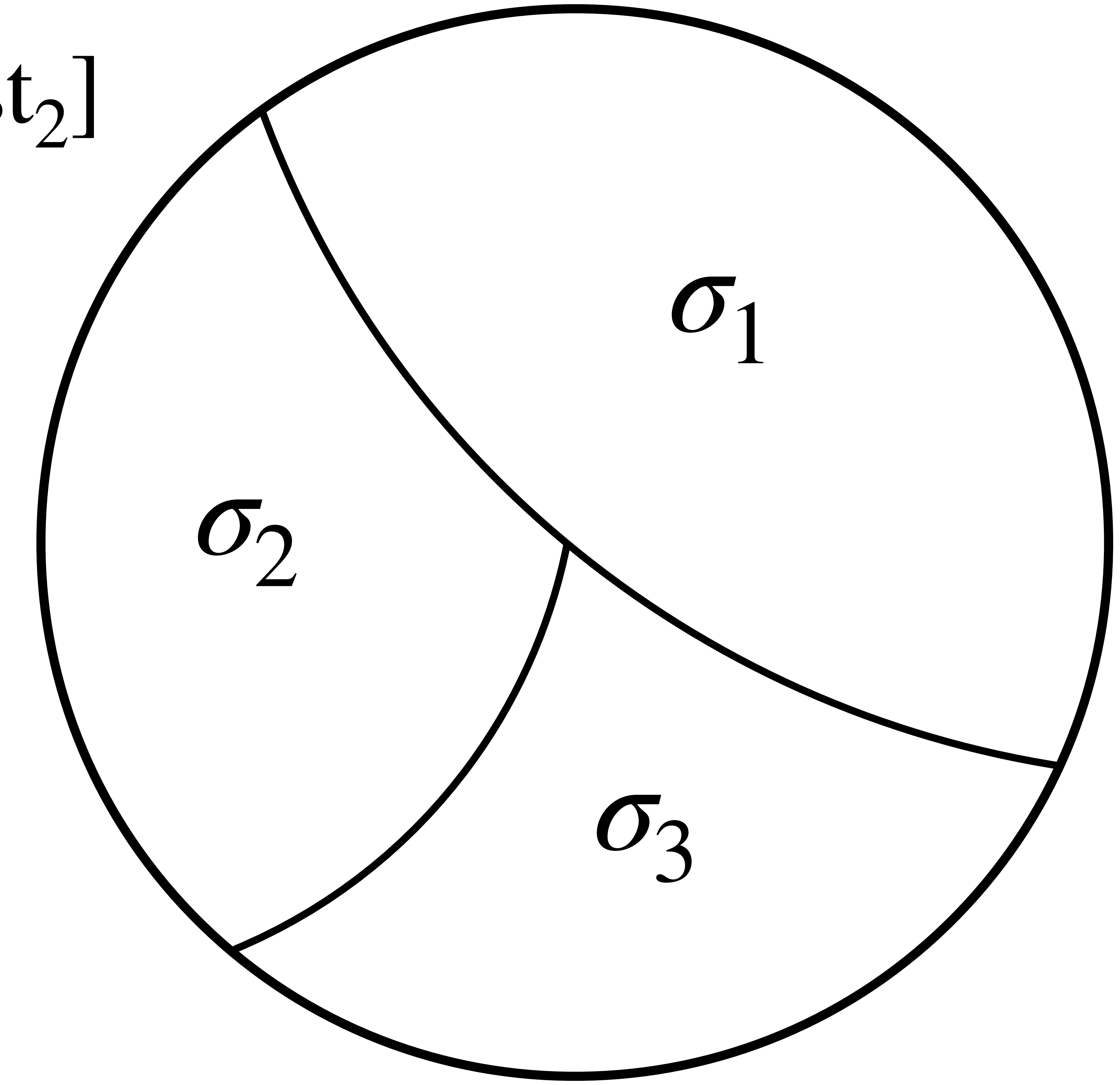
C. R. Shalizi and J. P. Crutchfield. "Computational Mechanics: Pattern and Prediction, Structure and Simplicity". In: J. Stat. Phys. 104 (2001), pp. 817–879.

What's "Good"?

$$\text{Past}_1 \sim \text{Past}_2 \iff \Pr[\text{Future} \mid \text{Past}_1] = \Pr[\text{Future} \mid \text{Past}_2]$$

Equivalence class partitions pasts into the *causal states*

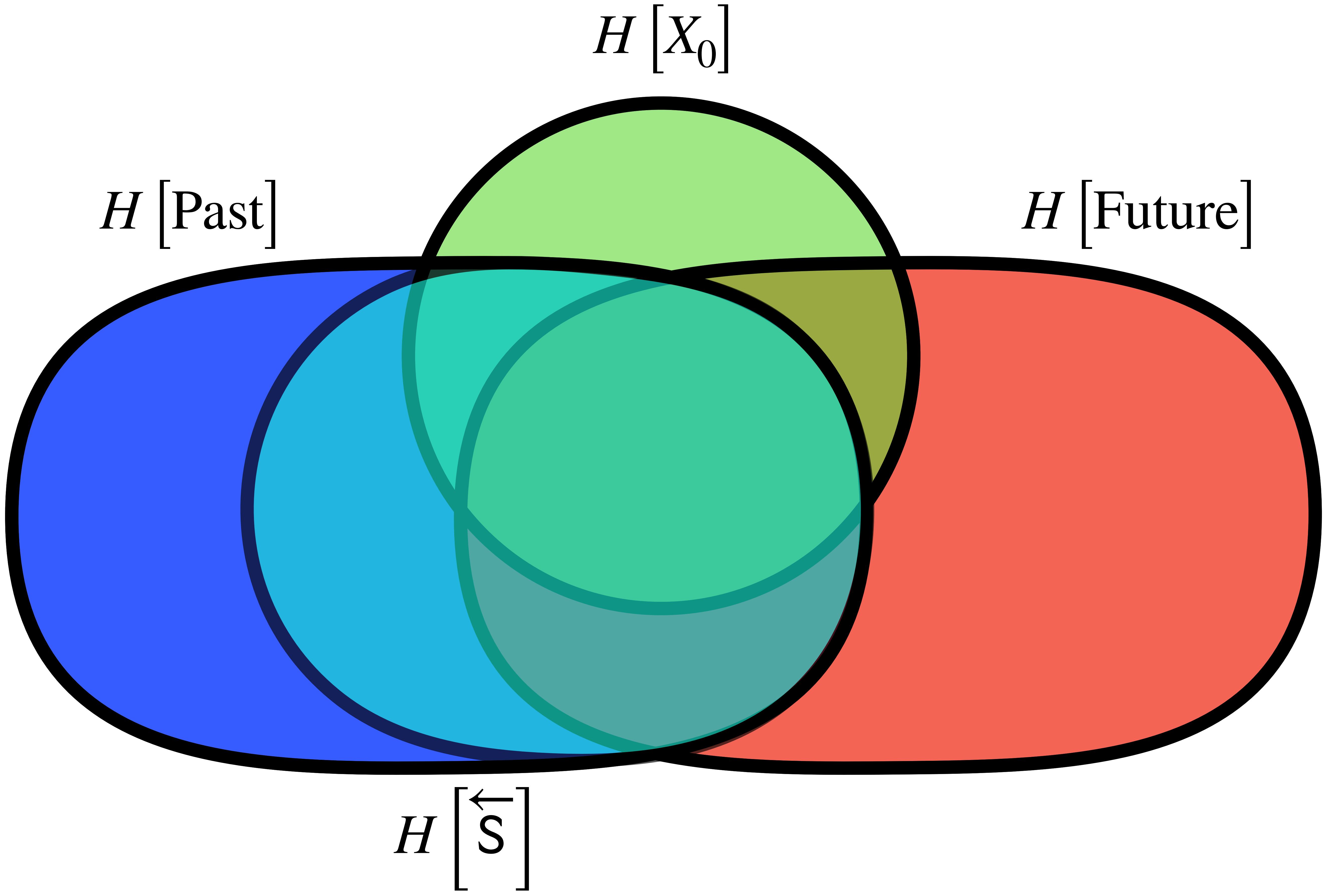
$$\mathsf{S} = \mathbf{H}/\sim .$$



Space of all
Pasts $\mathbf{H}$

James P. Crutchfield, K. Young. *Inferring statistical complexity*. Phys. Rev. Lett., 63 (2) (1989), pp. 105-108

ϵ-Machines are "Good"

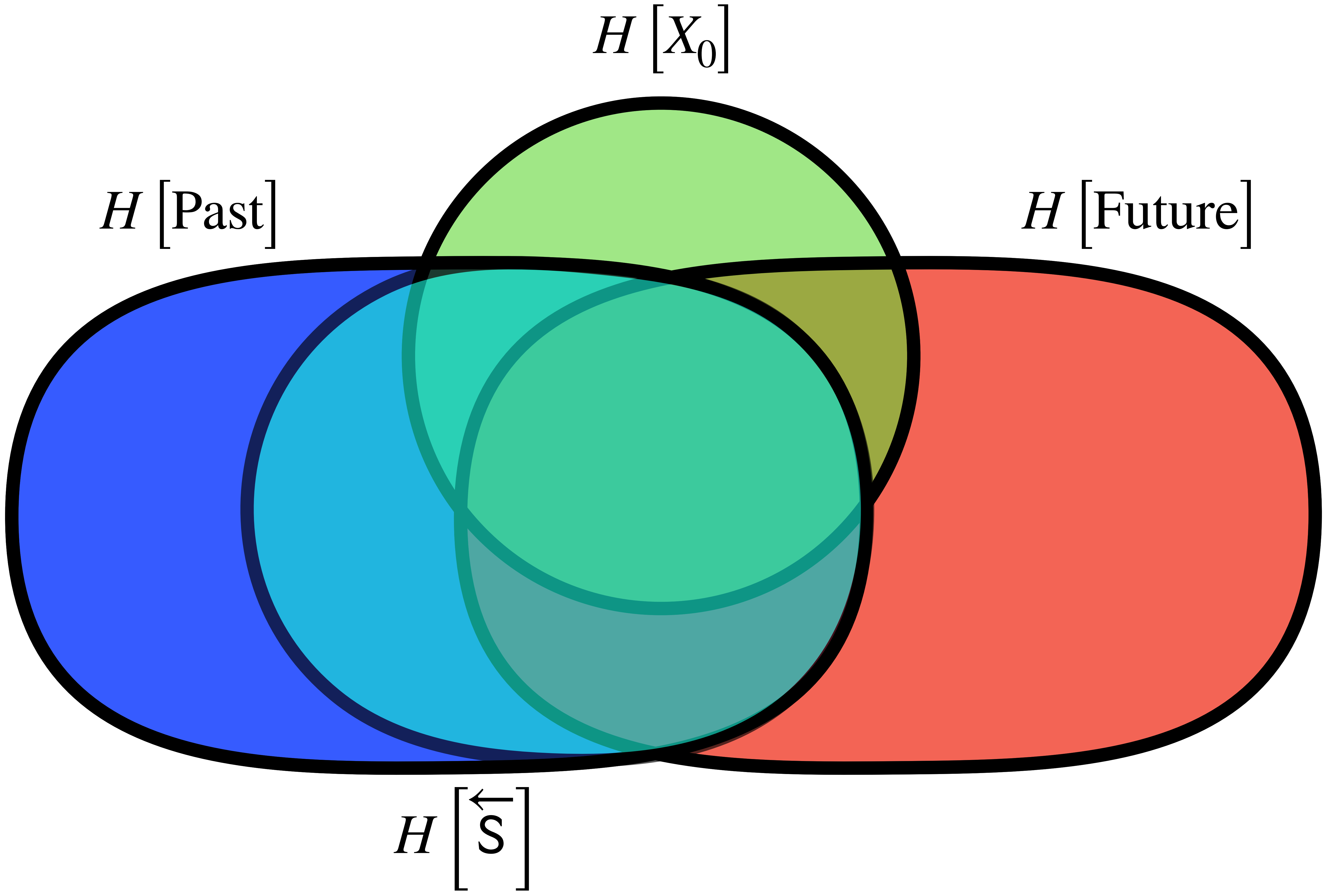


- States are a function of the past
- Optimal predictor
- Minimal in size
- Unique

➡ The ϵ machine

C. R. Shalizi and J. P. Crutchfield. "Computational Mechanics: Pattern and Prediction, Structure and Simplicity". In: J. Stat. Phys. 104 (2001), pp. 817–879.

ϵ-Machines are "Good"



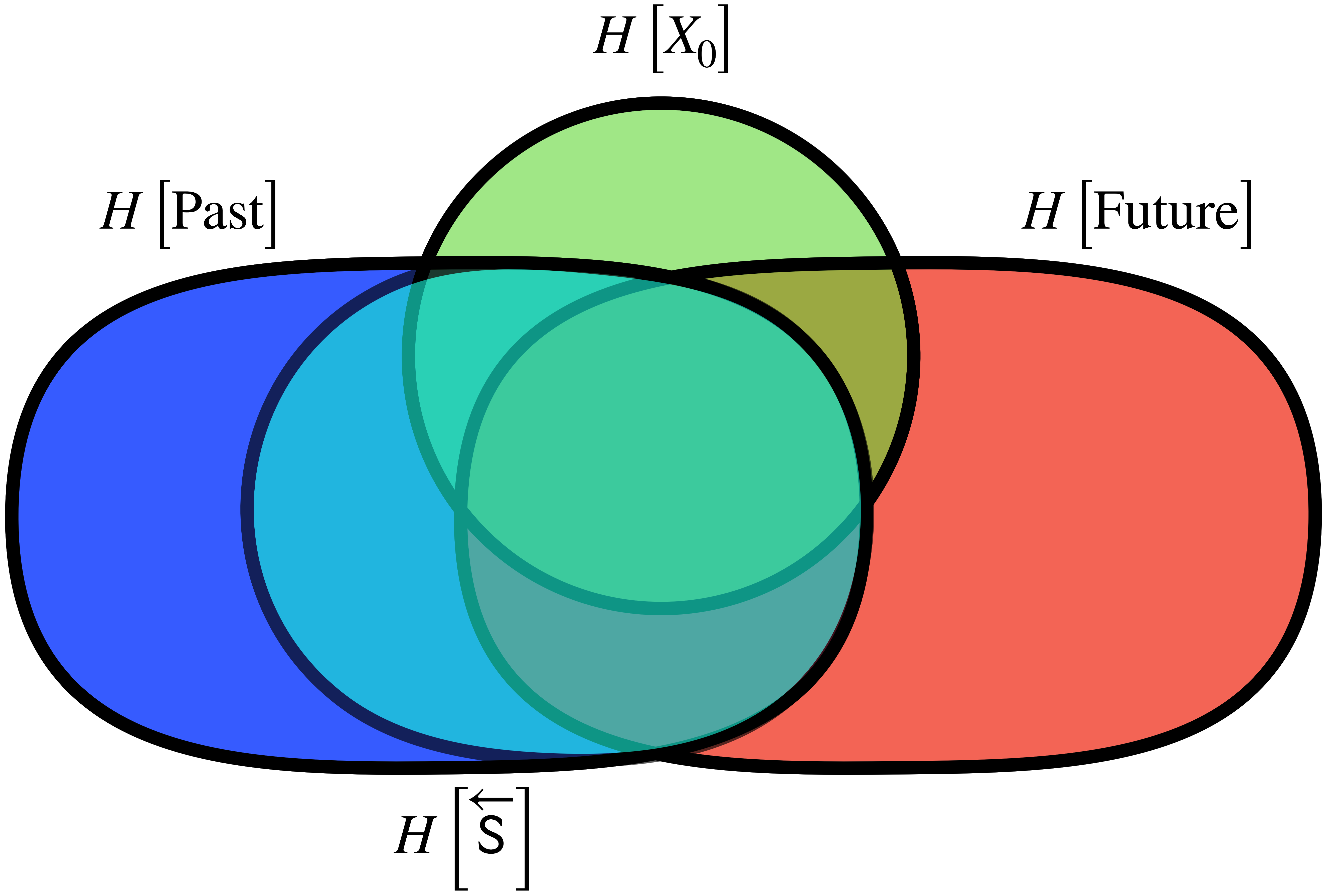
- States are a function of the past
- Optimal predictor
- Minimal in size
- Unique

➡ The ϵ machine

Mechanism?

C. R. Shalizi and J. P. Crutchfield. "Computational Mechanics: Pattern and Prediction, Structure and Simplicity". In: J. Stat. Phys. 104 (2001), pp. 817–879.

ϵ-Machines are "Good"

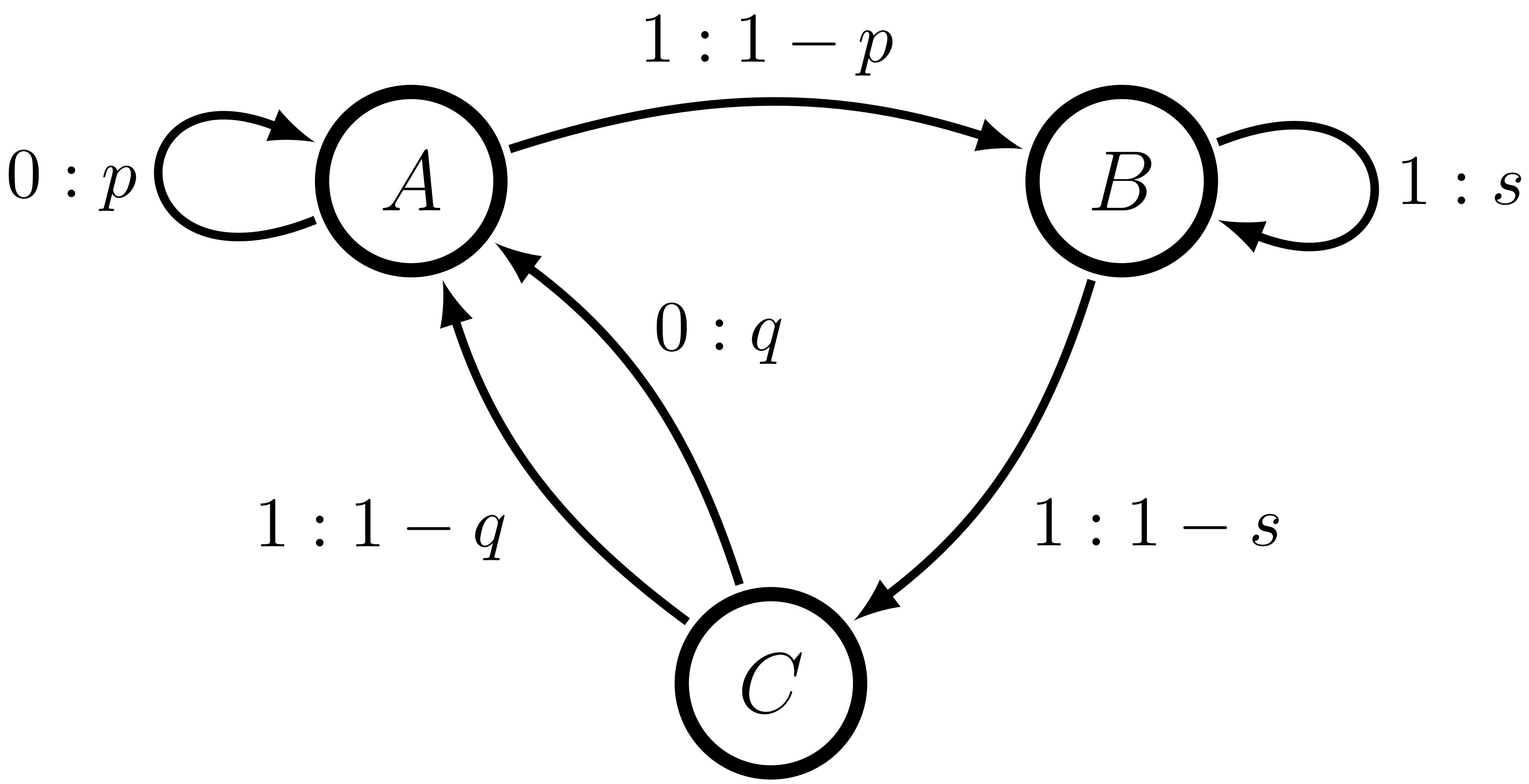


The number of bits required to store the ϵ -machine is called the *statistical complexity*:

$$C_\mu = H[\mathsf{S}]$$

James P. Crutchfield, K. Young. *Inferring statistical complexity*. Phys. Rev. Lett., 63 (2) (1989), pp. 105-108.
C. R. Shalizi and J. P. Crutchfield. "Computational Mechanics: Pattern and Prediction, Structure and Simplicity". In: J. Stat. Phys. 104 (2001), pp. 817–879.

Most HMMs are *Not* Optimal Predictors!



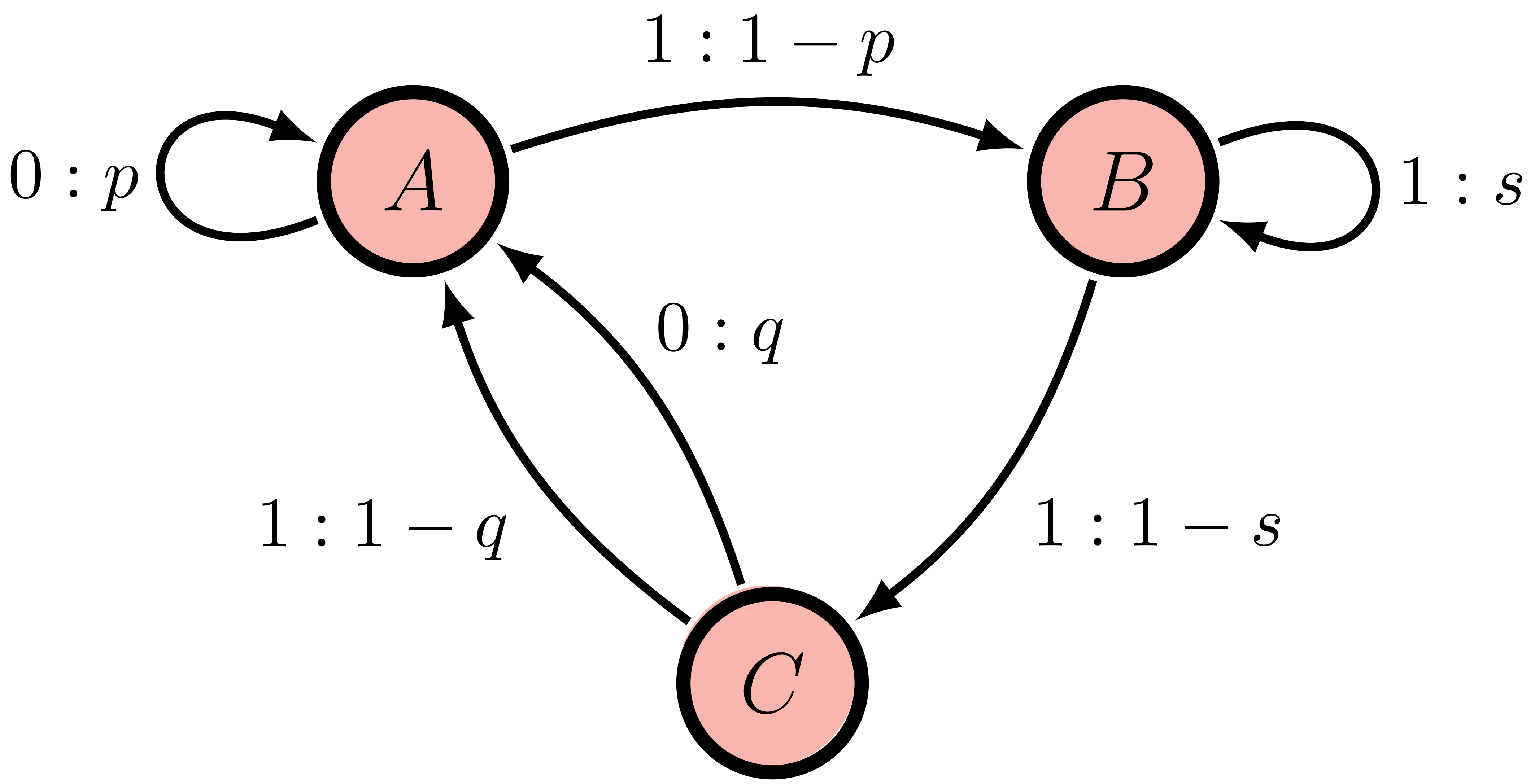
$$h_\mu \neq \sum_{\sigma \in \mathcal{S}} \Pr(\sigma)\, H\left[X \mid \sigma\right]$$

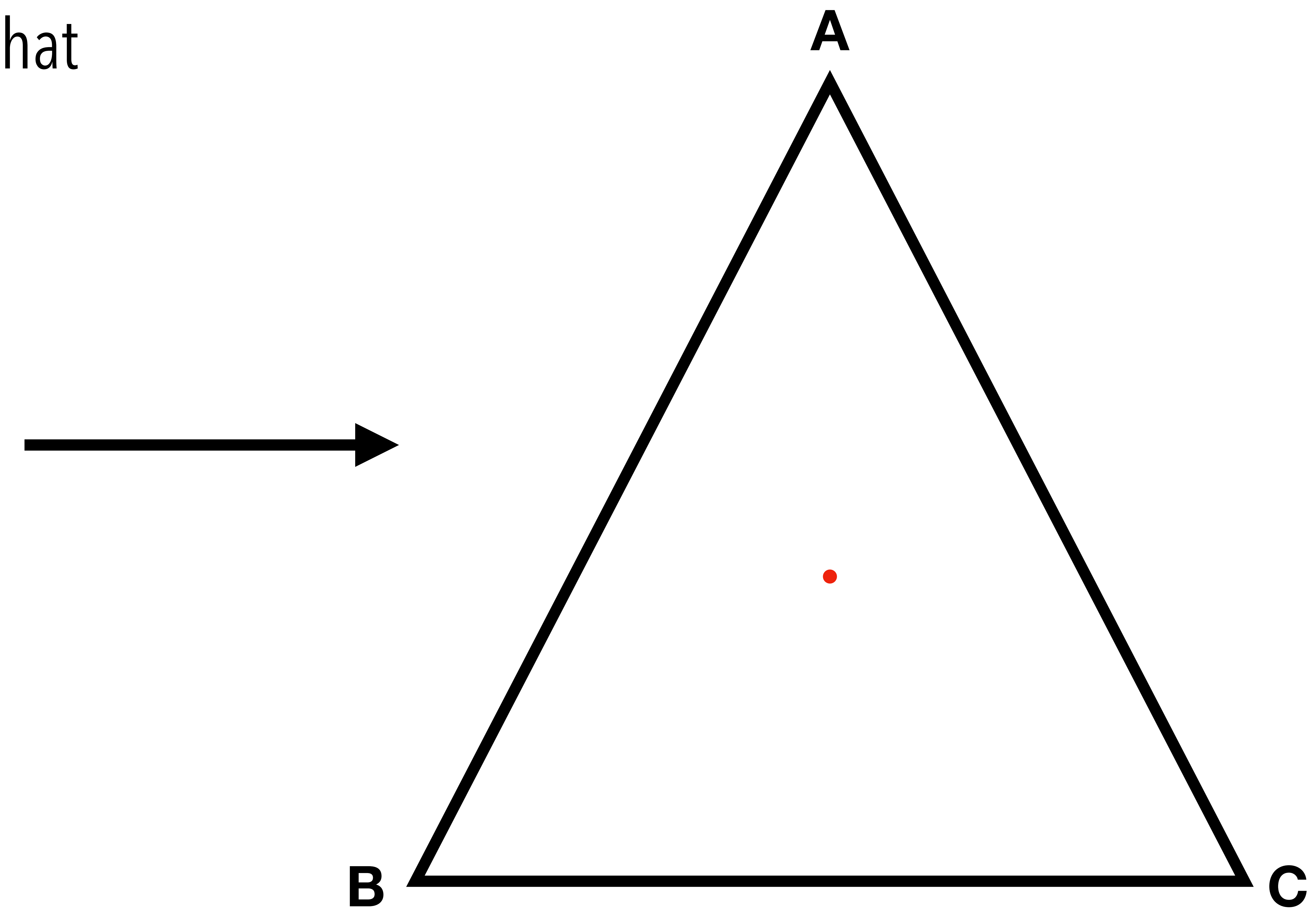
Problem Statement

Given an arbitrary finite-state hidden Markov model,
how to get ϵ-machine?

Observe and Update Game

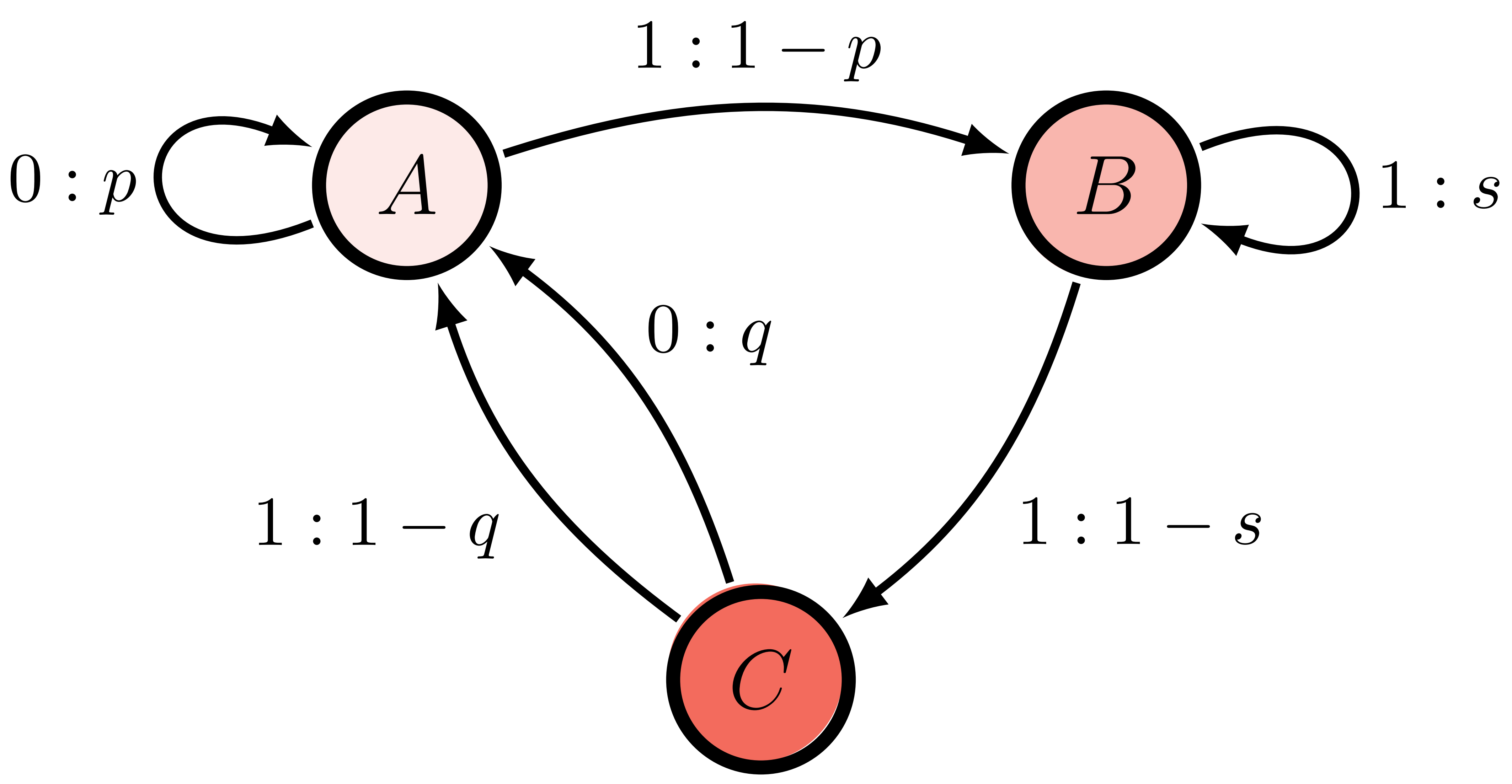
Take our "bad" HMM and track our belief over what state we're in:



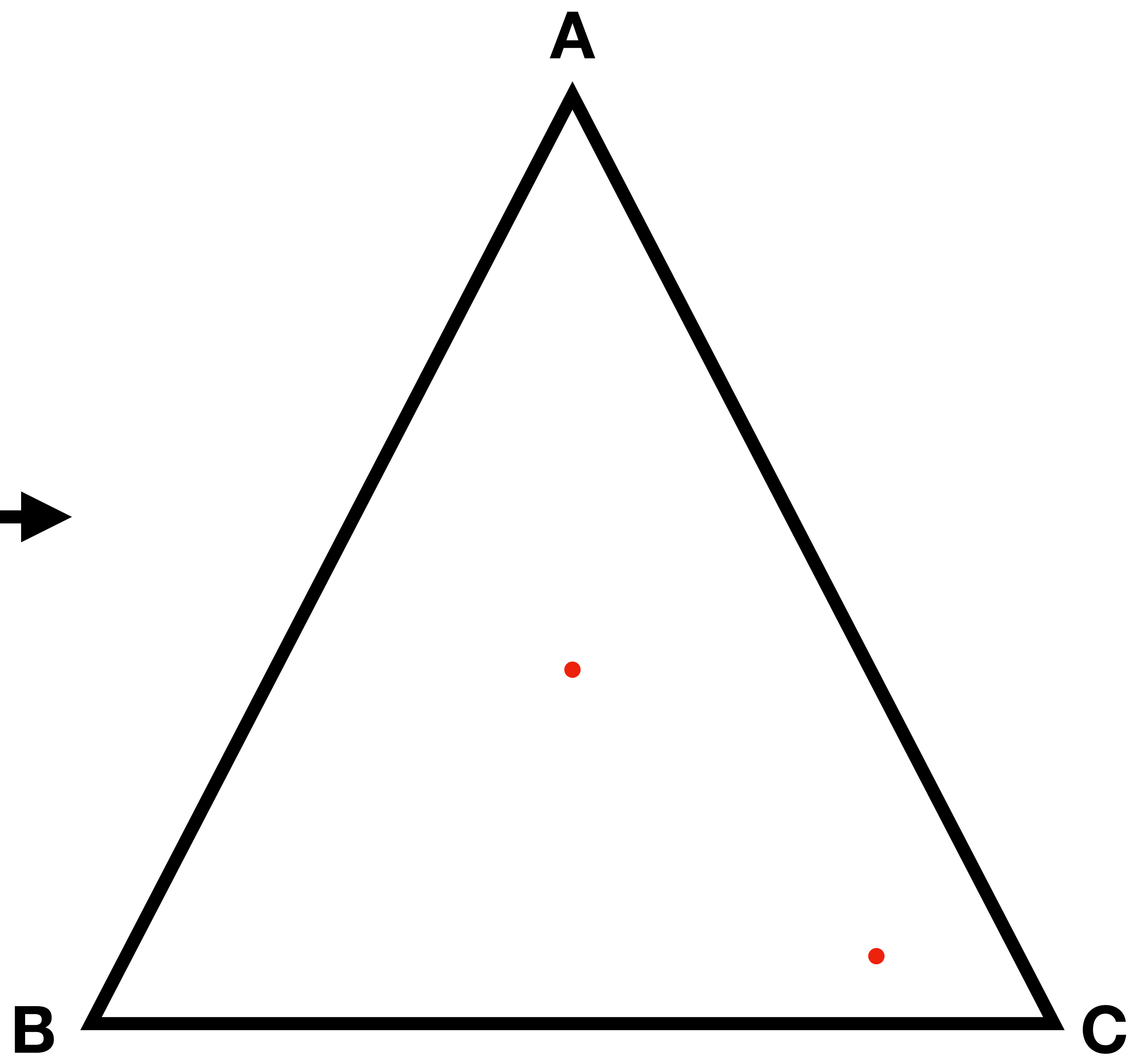


Observed sequence: λ

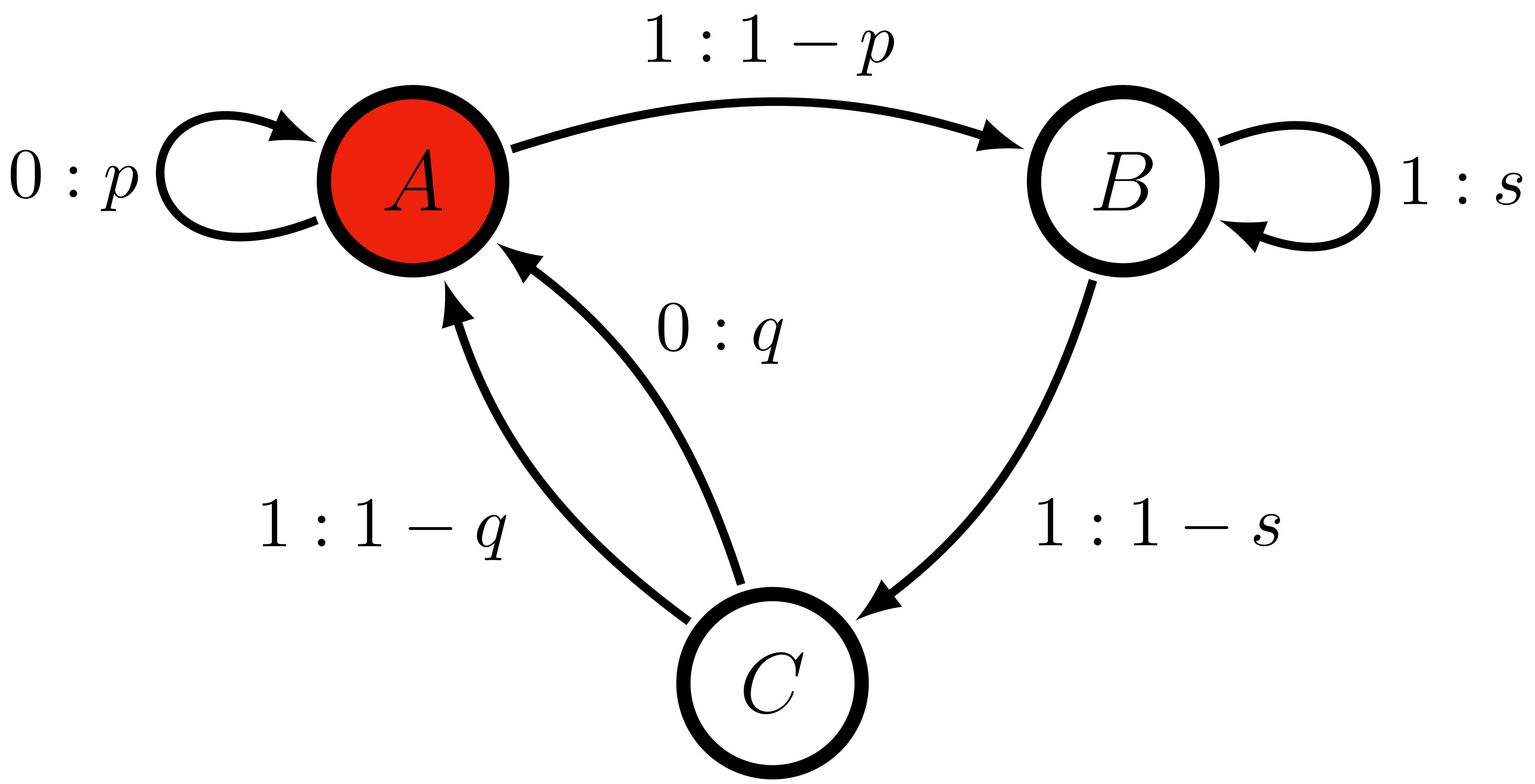
Observe and Update Game



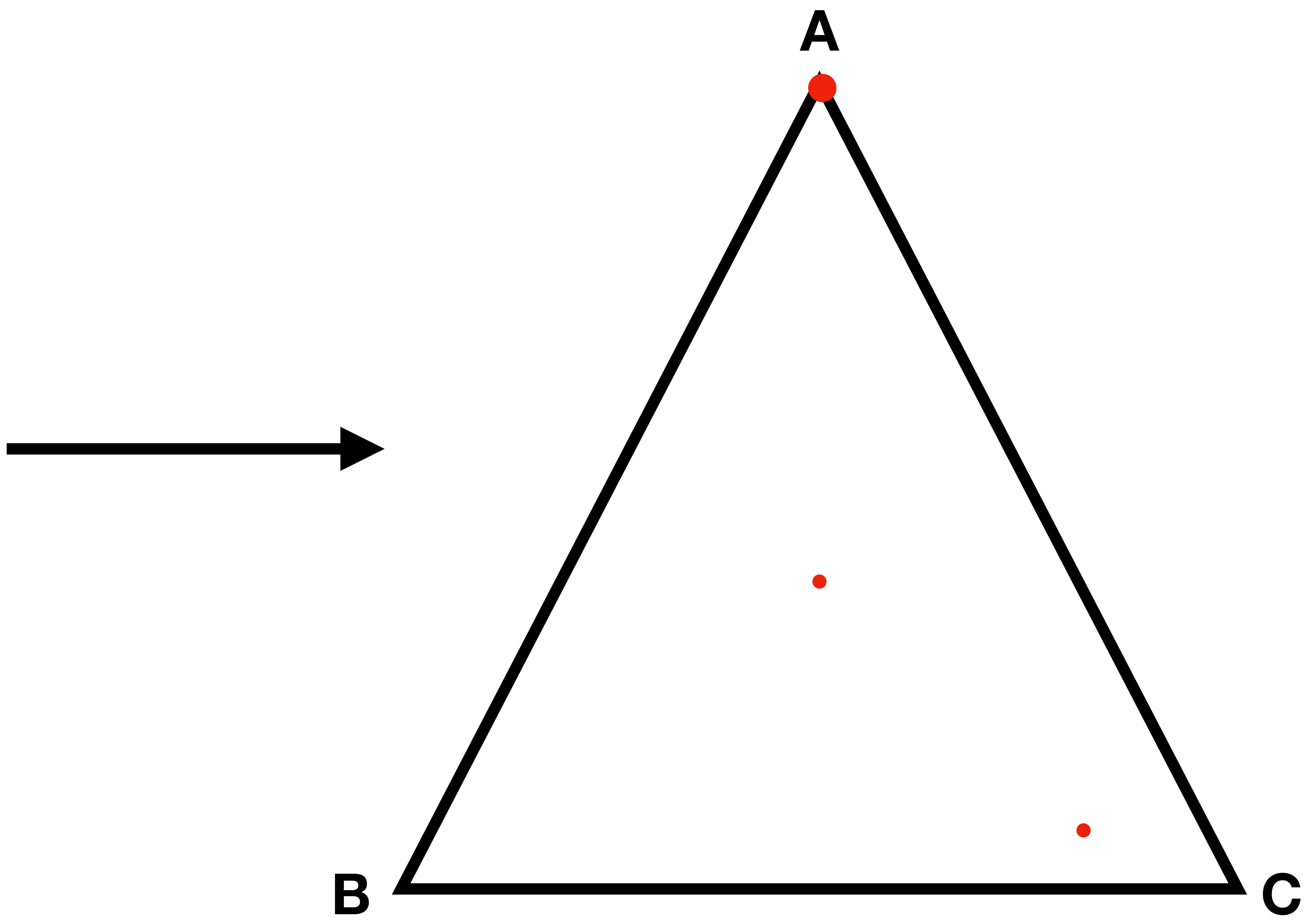
Observed sequence: $\lambda 1$



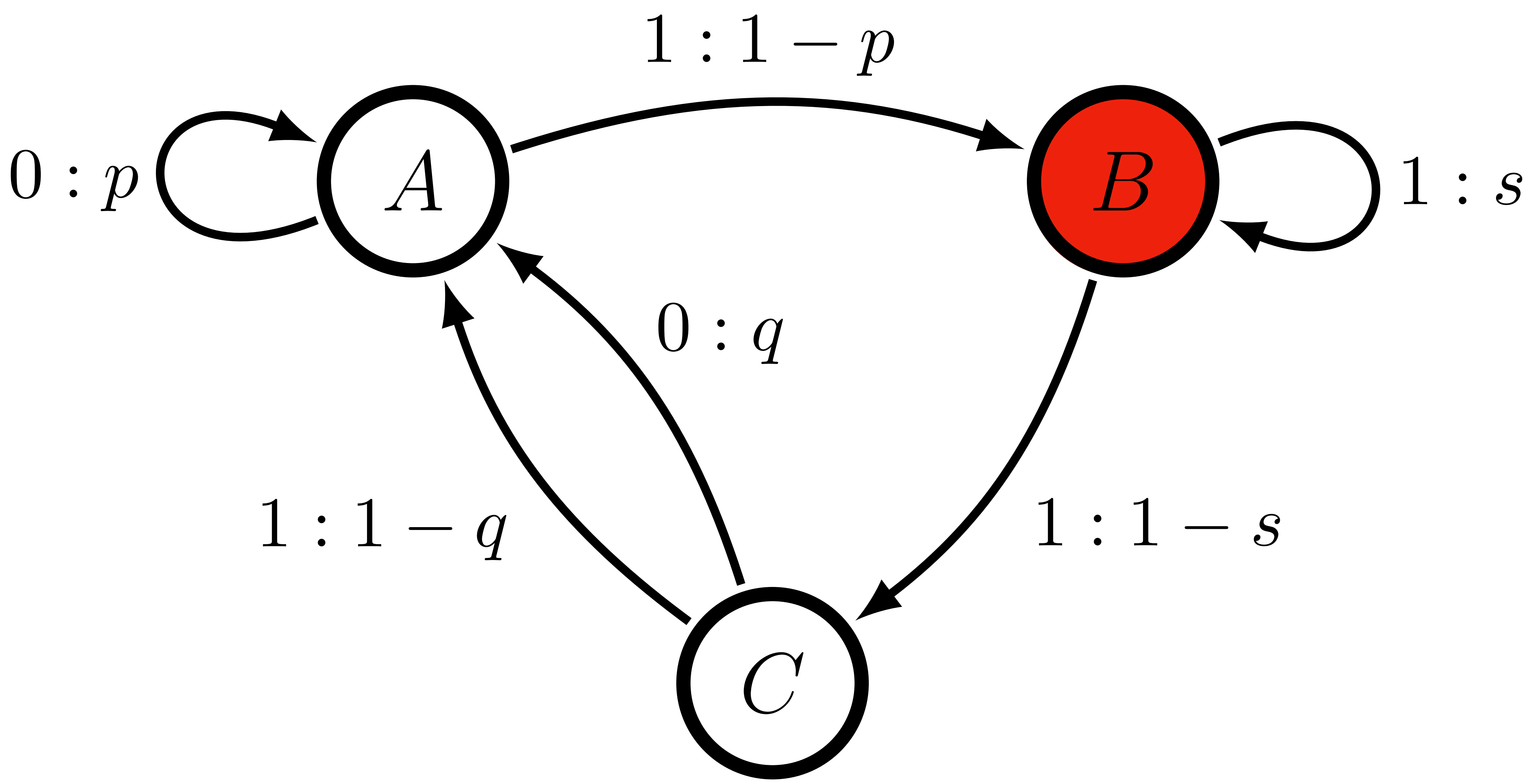
Observe and Update Game



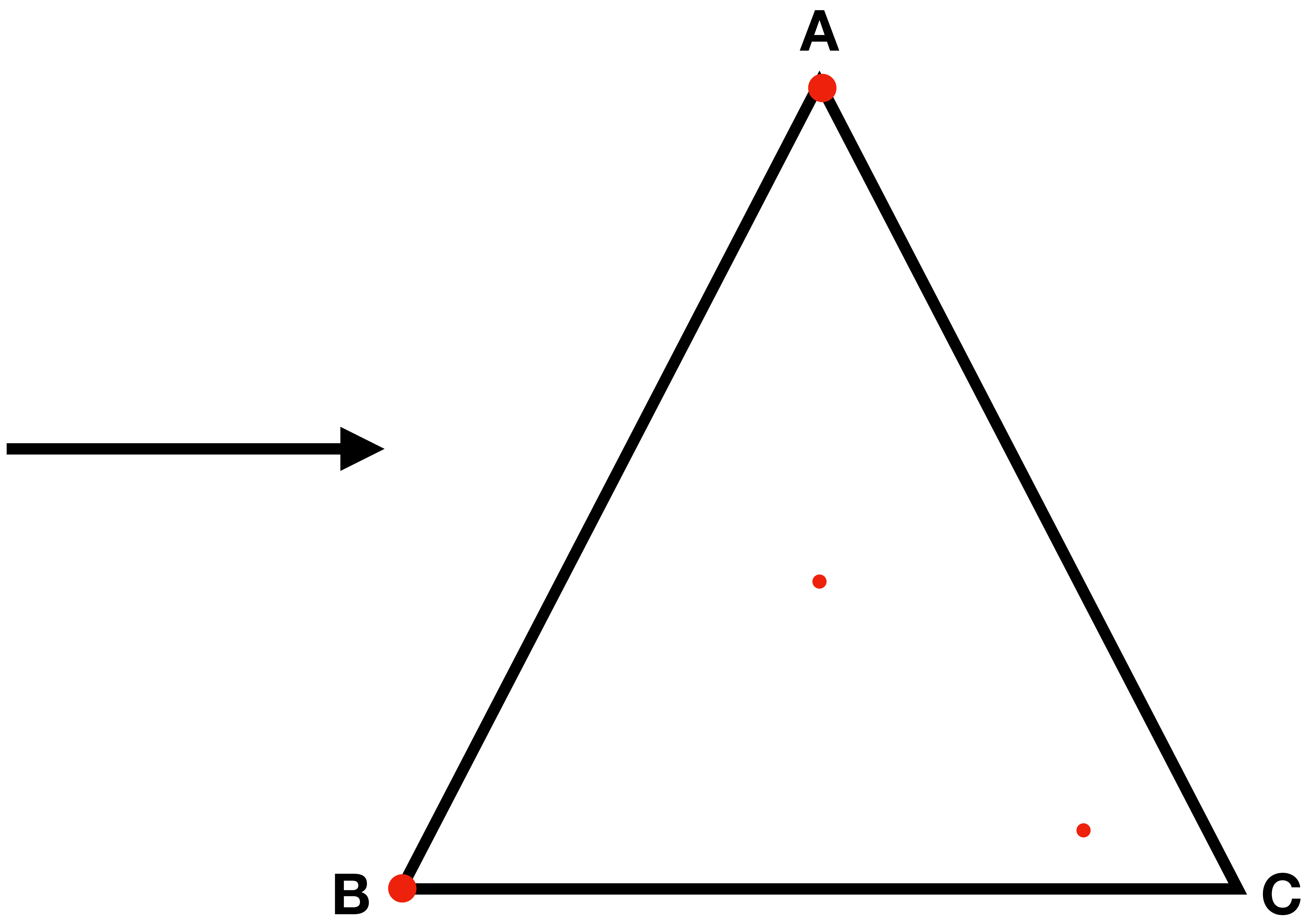
Observed sequence: $\lambda 10$



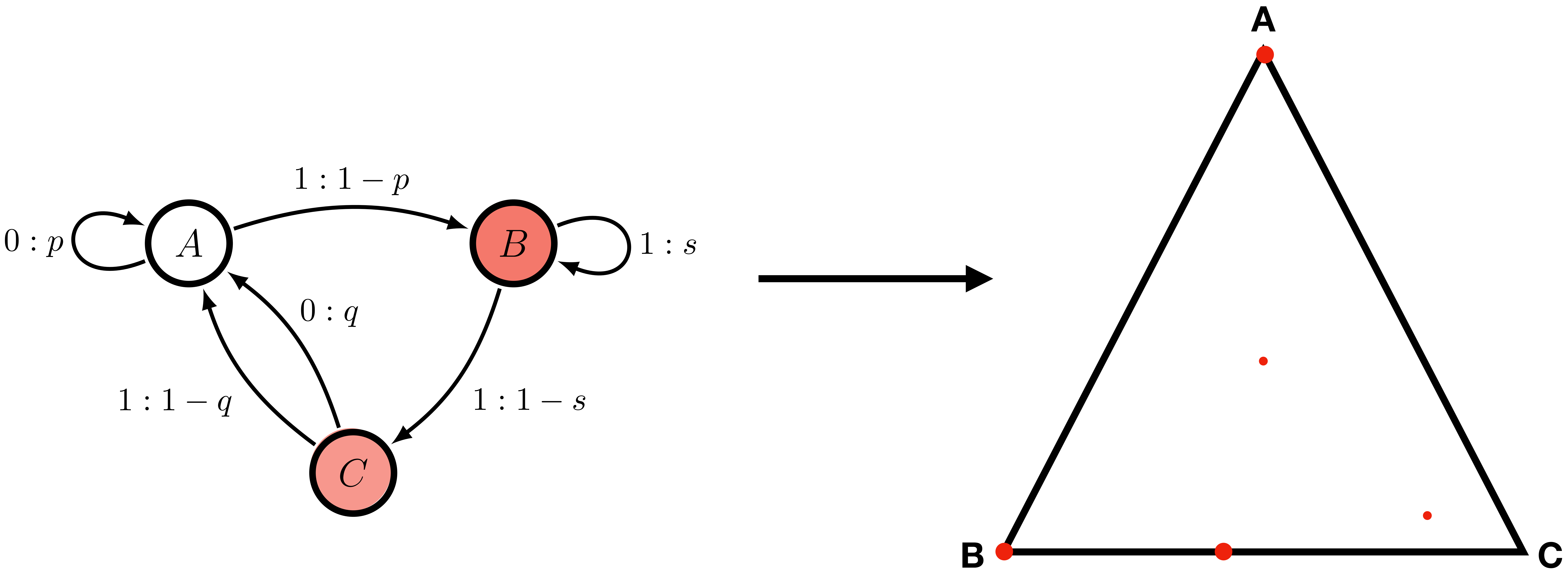
Observe and Update Game



Observed sequence: $\lambda 101$

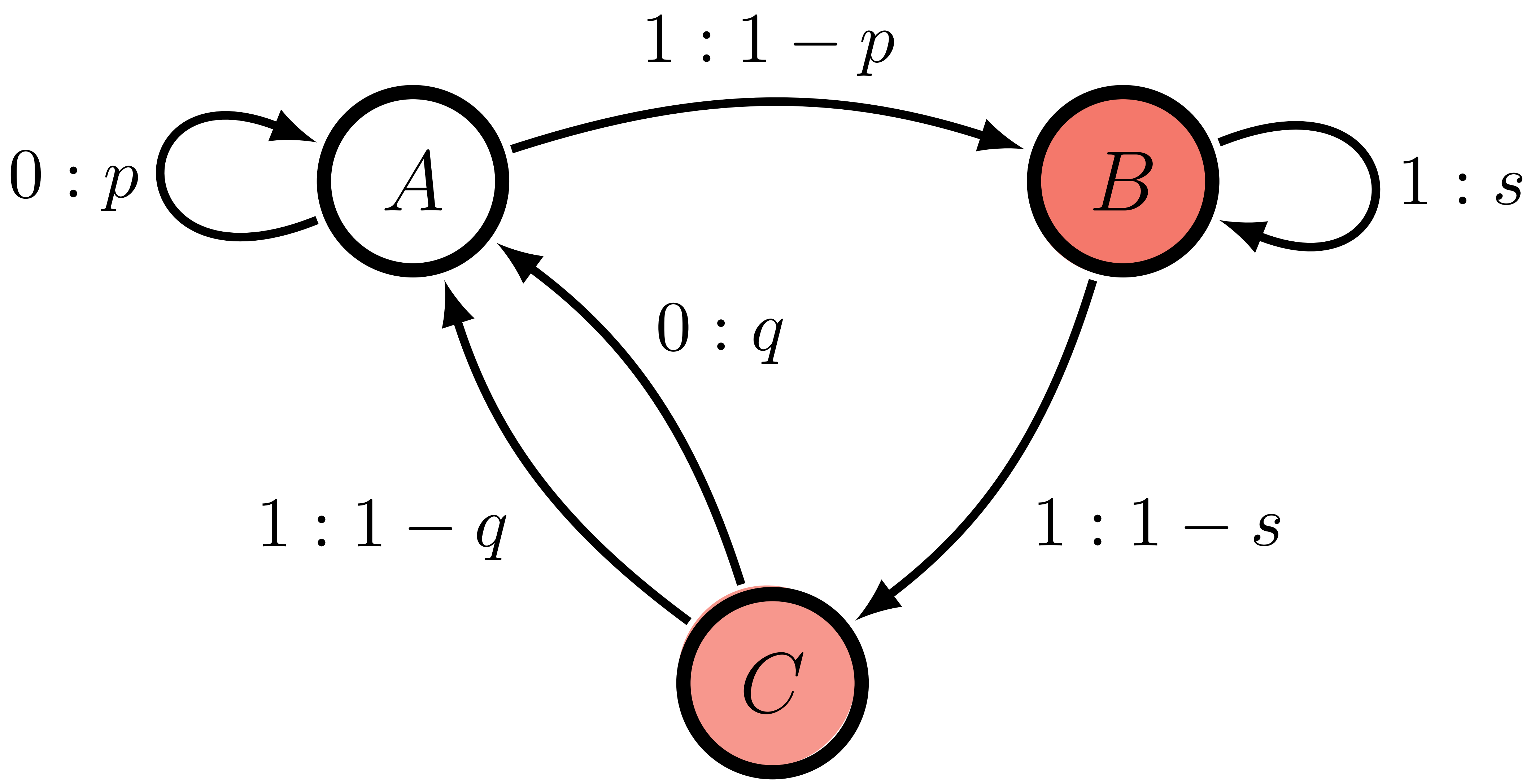


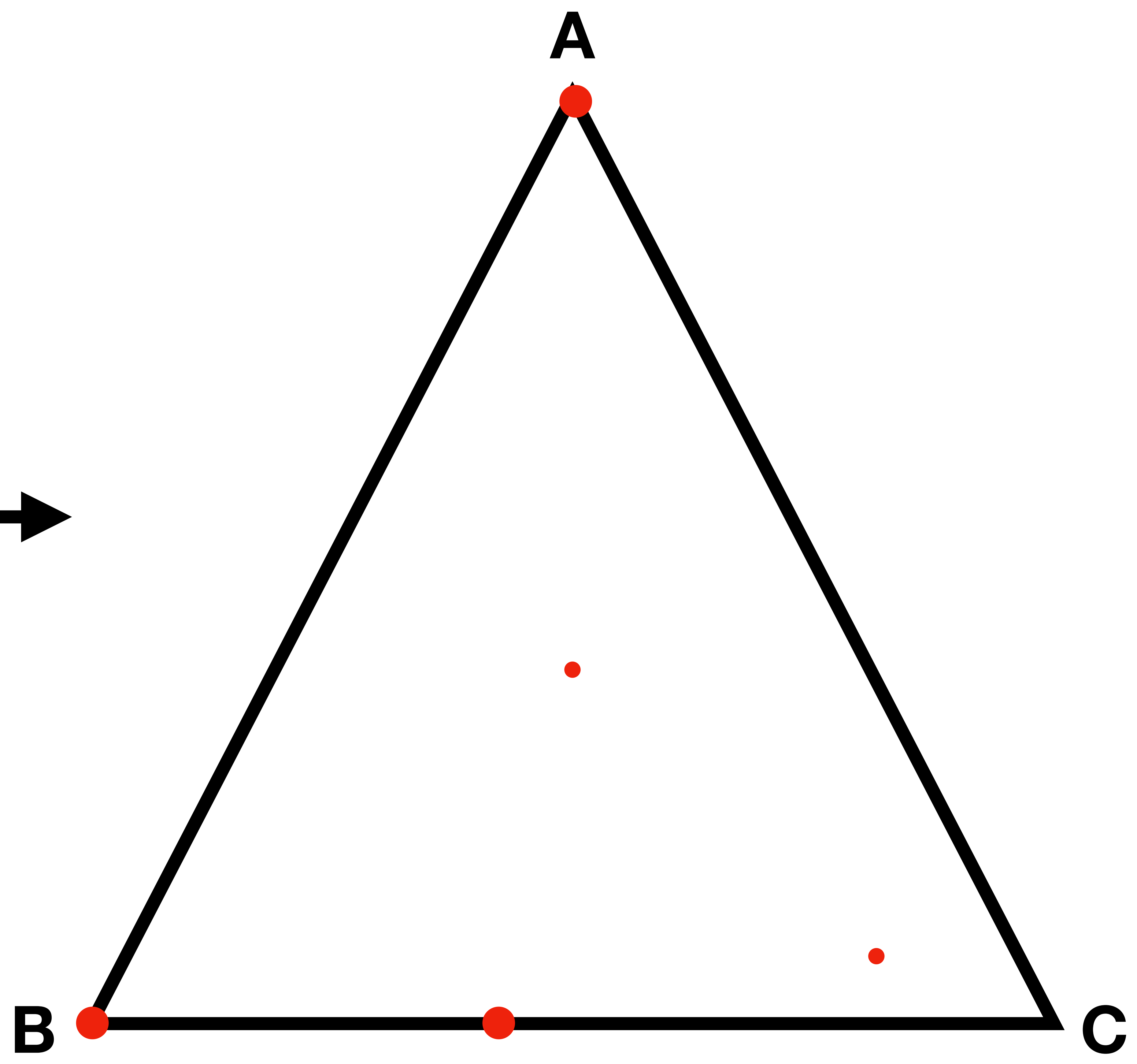
Observe and Update Game



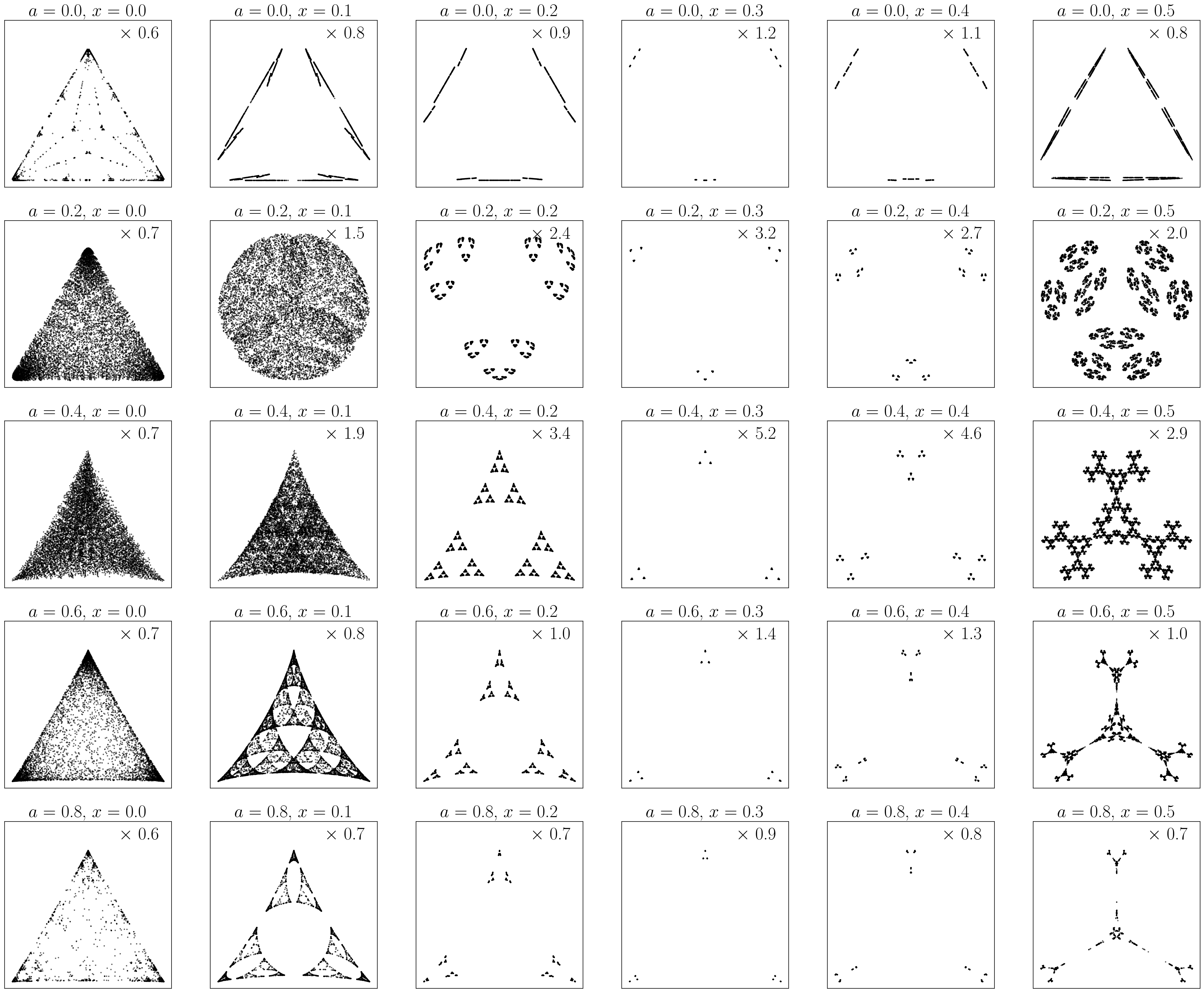
Observed sequence: $\lambda 1011$

Observe and Update Game

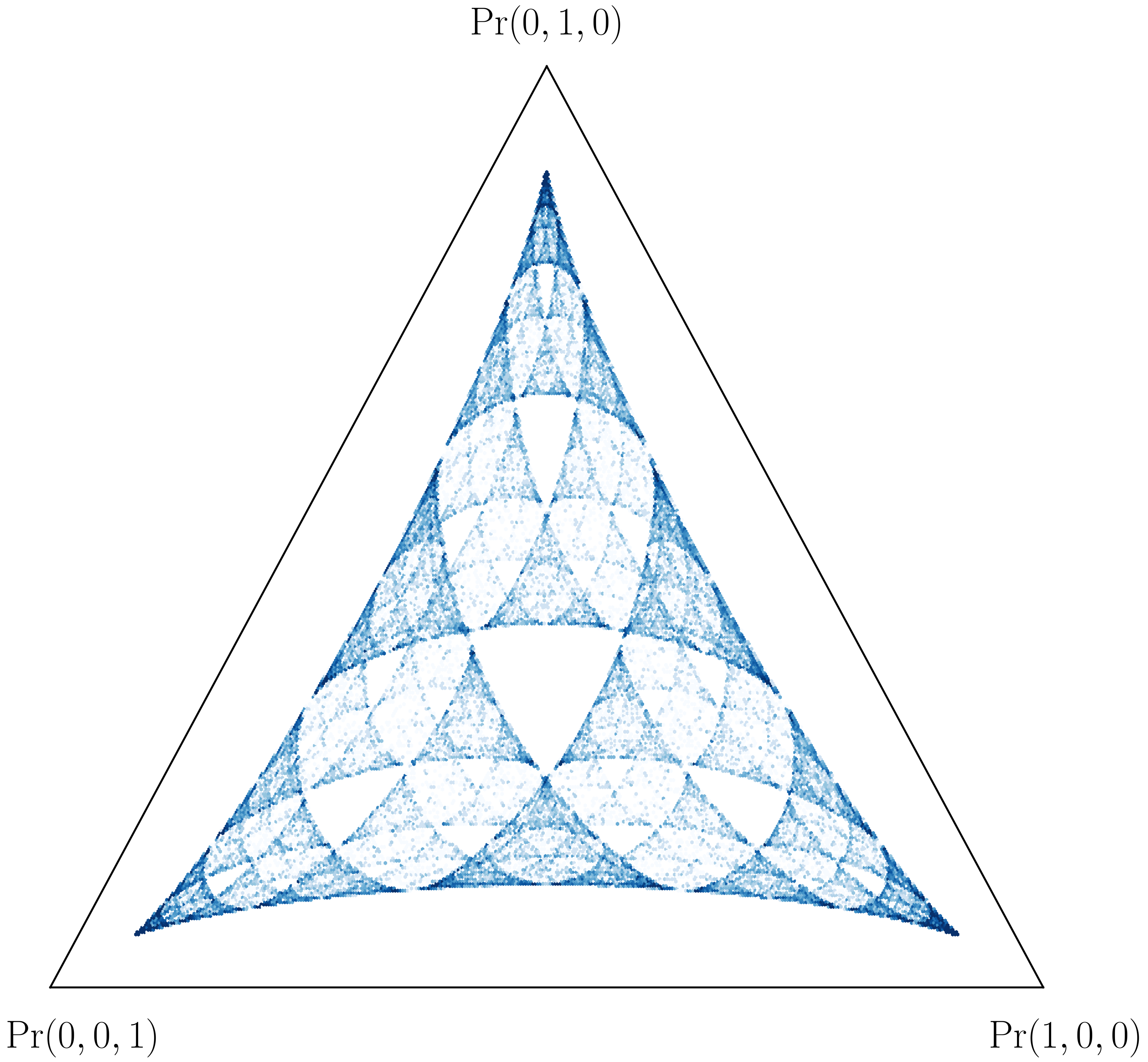




$\rightarrow$ Build up a set $R = \left\{ \eta : \eta(w) = \Pr\left(S_l \,|\, X_{0:l} = w, S_0 = \pi\right) \right\}$ of belief states.


$a = 0.0,\ x = 0.0$ × 0.6
$a = 0.0,\ x = 0.1$ × 0.8
$a = 0.0,\ x = 0.2$ × 0.9
$a = 0.0,\ x = 0.3$ × 1.2
$a = 0.0,\ x = 0.4$ × 1.1
$a = 0.0,\ x = 0.5$ × 0.8
$a = 0.2,\ x = 0.0$ × 0.7
$a = 0.2,\ x = 0.1$ × 1.5
$a = 0.2,\ x = 0.2$ × 2.4
$a = 0.2,\ x = 0.3$ × 3.2
$a = 0.2,\ x = 0.4$ × 2.7
$a = 0.2,\ x = 0.5$ × 2.0
$a = 0.4,\ x = 0.0$ × 0.7
$a = 0.4,\ x = 0.1$ × 1.9
$a = 0.4,\ x = 0.2$ × 3.4
$a = 0.4,\ x = 0.3$ × 5.2
$a = 0.4,\ x = 0.4$ × 4.6
$a = 0.4,\ x = 0.5$ × 2.9
$a = 0.6,\ x = 0.0$ × 0.7
$a = 0.6,\ x = 0.1$ × 0.8
$a = 0.6,\ x = 0.2$ × 1.0
$a = 0.6,\ x = 0.3$ × 1.4
$a = 0.6,\ x = 0.4$ × 1.3
$a = 0.6,\ x = 0.5$ × 1.0
$a = 0.8,\ x = 0.0$ × 0.6
$a = 0.8,\ x = 0.1$ × 0.7
$a = 0.8,\ x = 0.2$ × 0.7
$a = 0.8,\ x = 0.3$ × 0.9
$a = 0.8,\ x = 0.4$ × 0.8
$a = 0.8,\ x = 0.5$ × 0.7

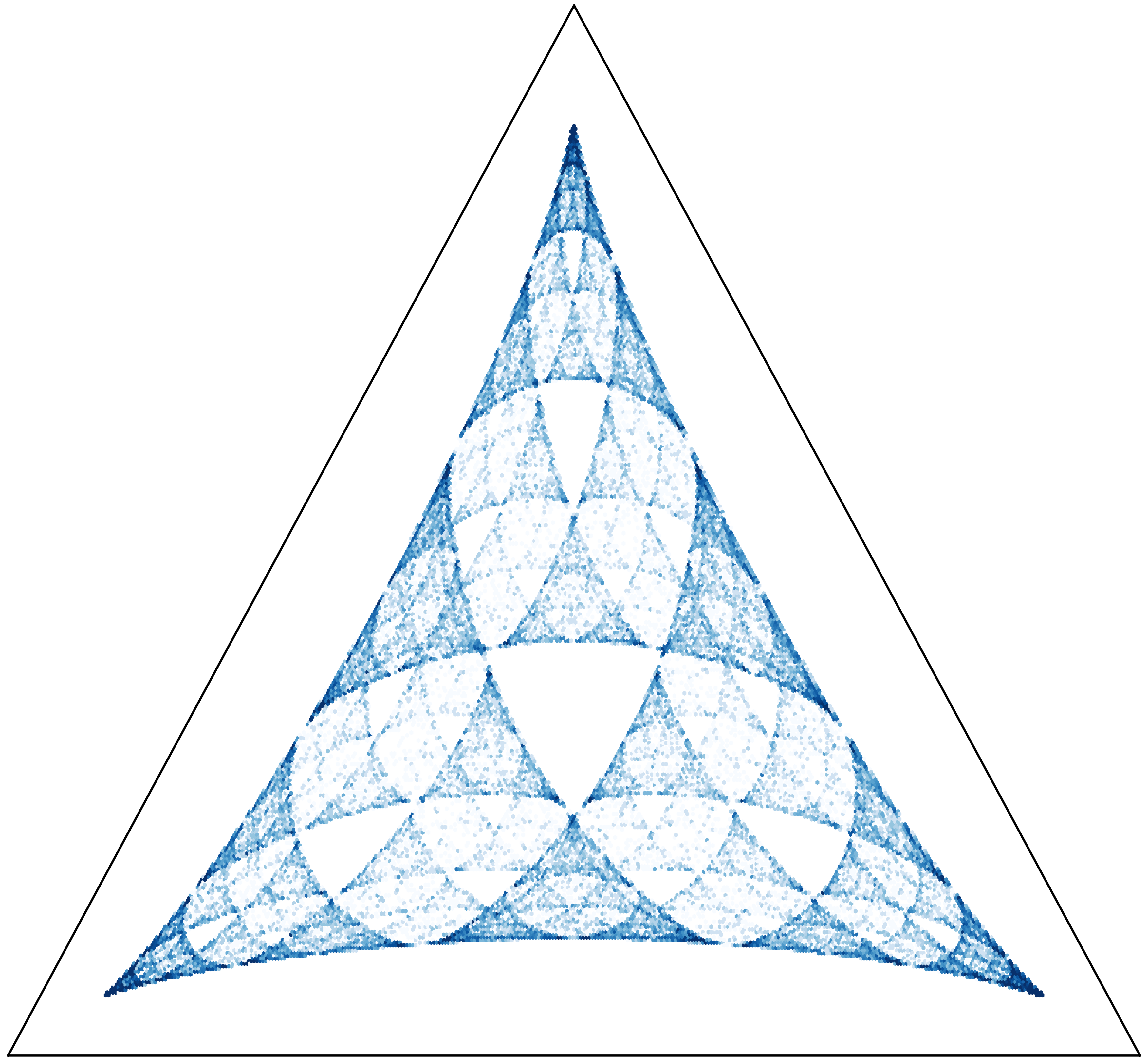
The Infinite State ϵ-Machine



The set of belief states $\boldsymbol{R}$ is the attractor of the "observe and update" stochastic dynamical system (known as an *iterated function system*).

Alex M. Jurgens, James P. Crutchfield. *Shannon Entropy Rate of Hidden Markov Processes*. J. Stat. Phys., 183 (2), 1-18, 2020.

The Infinite State ϵ-Machine

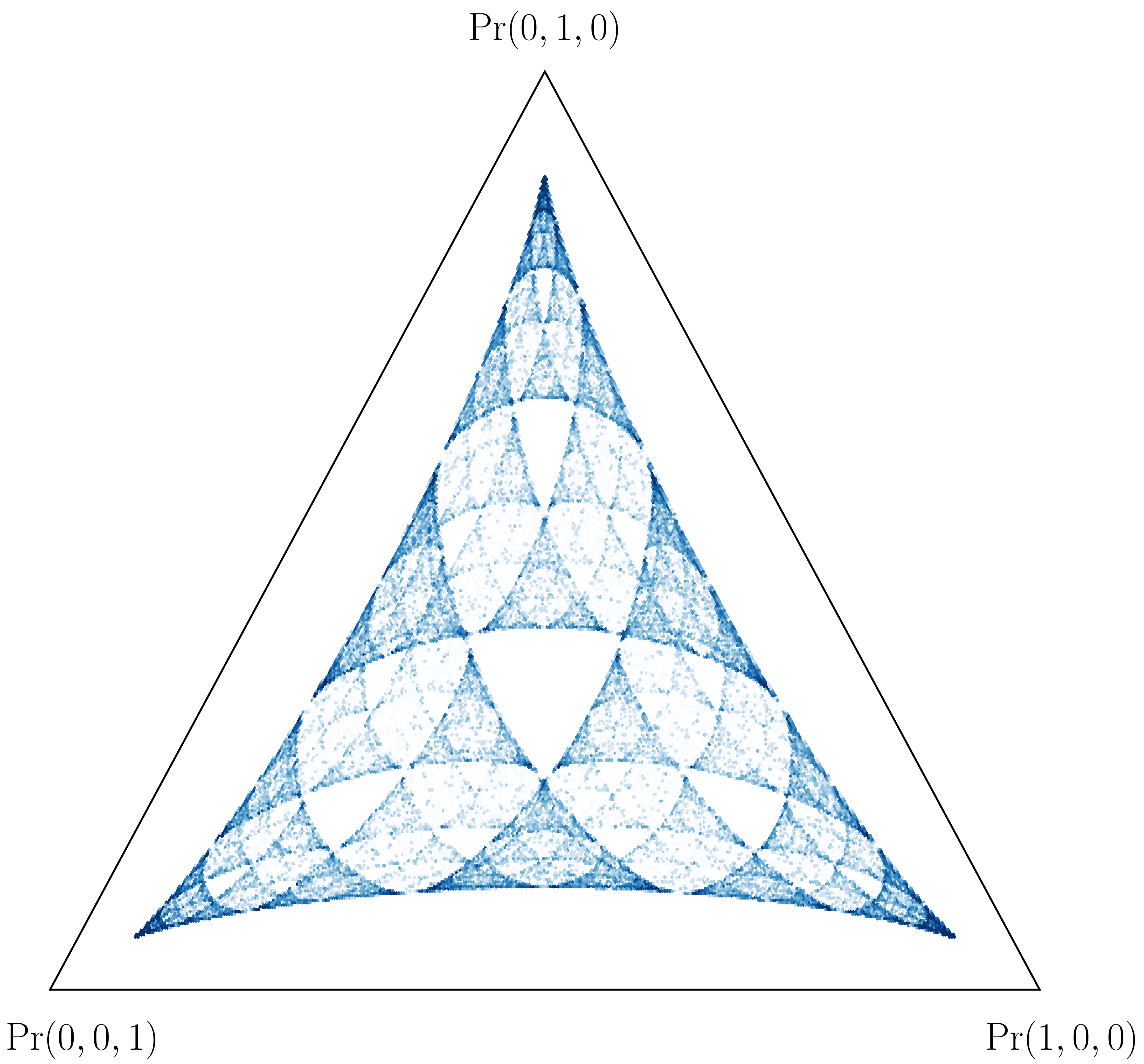


The set of belief states $\boldsymbol{R}$ is the attractor of the "observe and update" stochastic dynamical system (known as an *iterated function system*).

This attractor exists and is unique due to contractivity, but is generically fractal-like.

Alex M. Jurgens, James P. Crutchfield. *Shannon Entropy Rate of Hidden Markov Processes*. J. Stat. Phys., 183 (2), 1-18, 2020.

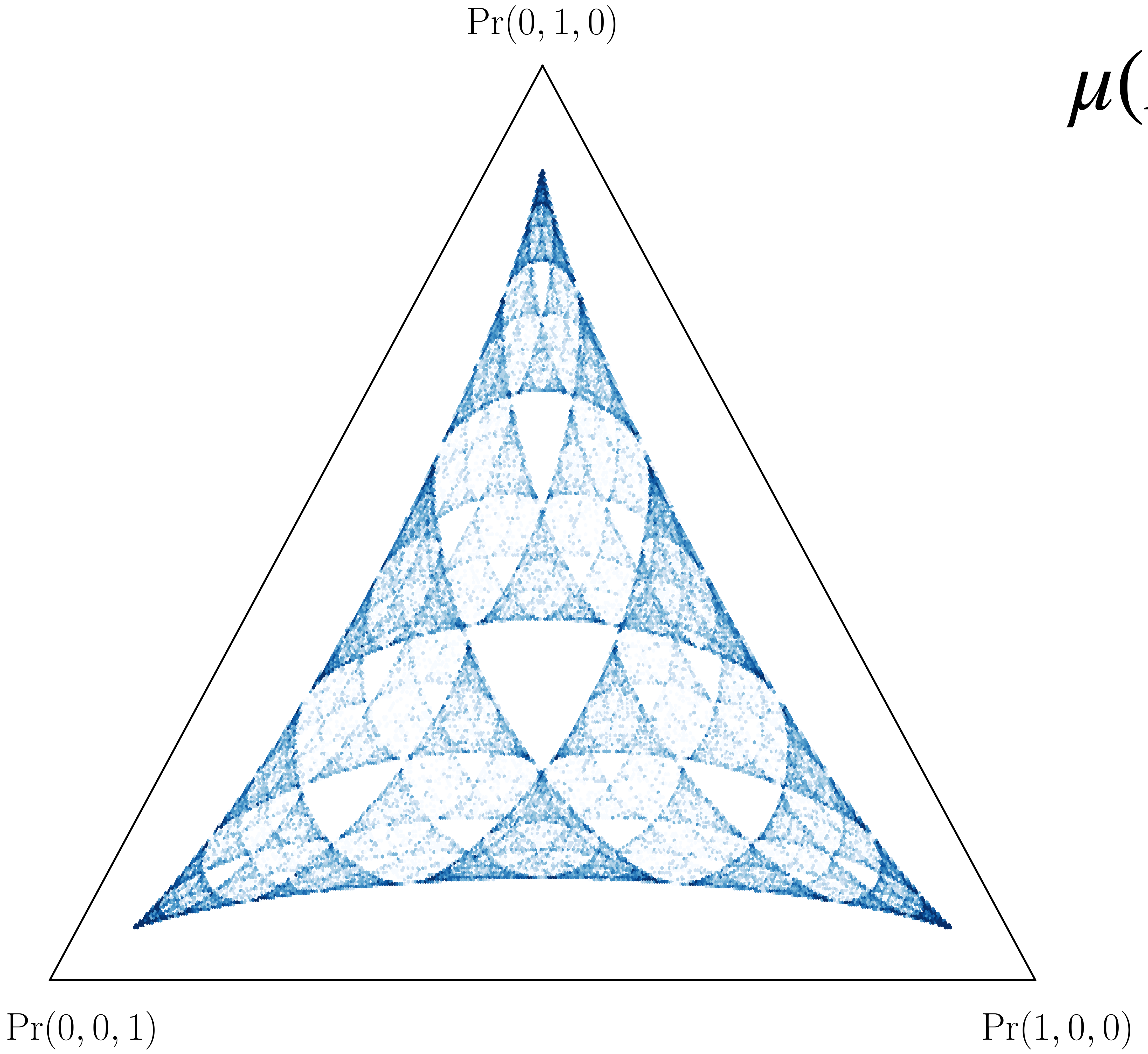
The Infinite State ϵ-Machine



For most *finitely generated* hidden Markov processes, the ϵ machine is an uncountably infinite state set + transitions between these states.

Alex M. Jurgens, James P. Crutchfield. *Shannon Entropy Rate of Hidden Markov Processes*. J. Stat. Phys., 183 (2), 1-18, 2020.

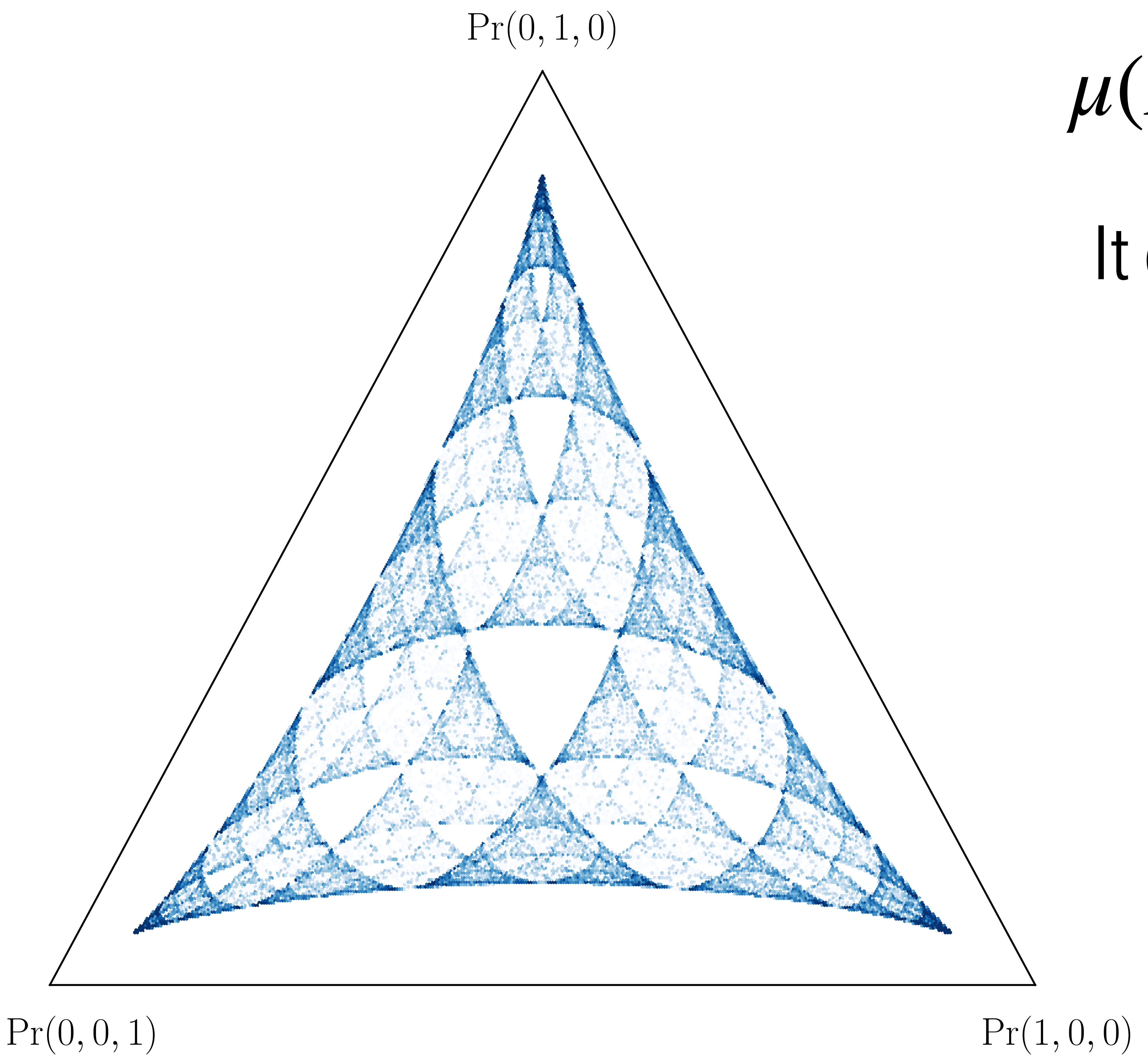
Calculating Shannon Entropy Rate



$\mu(R)$ is called the Blackwell measure.

D. Blackwell. *The entropy of functions of finite-state Markov chains.* 1957.

Calculating Shannon Entropy Rate



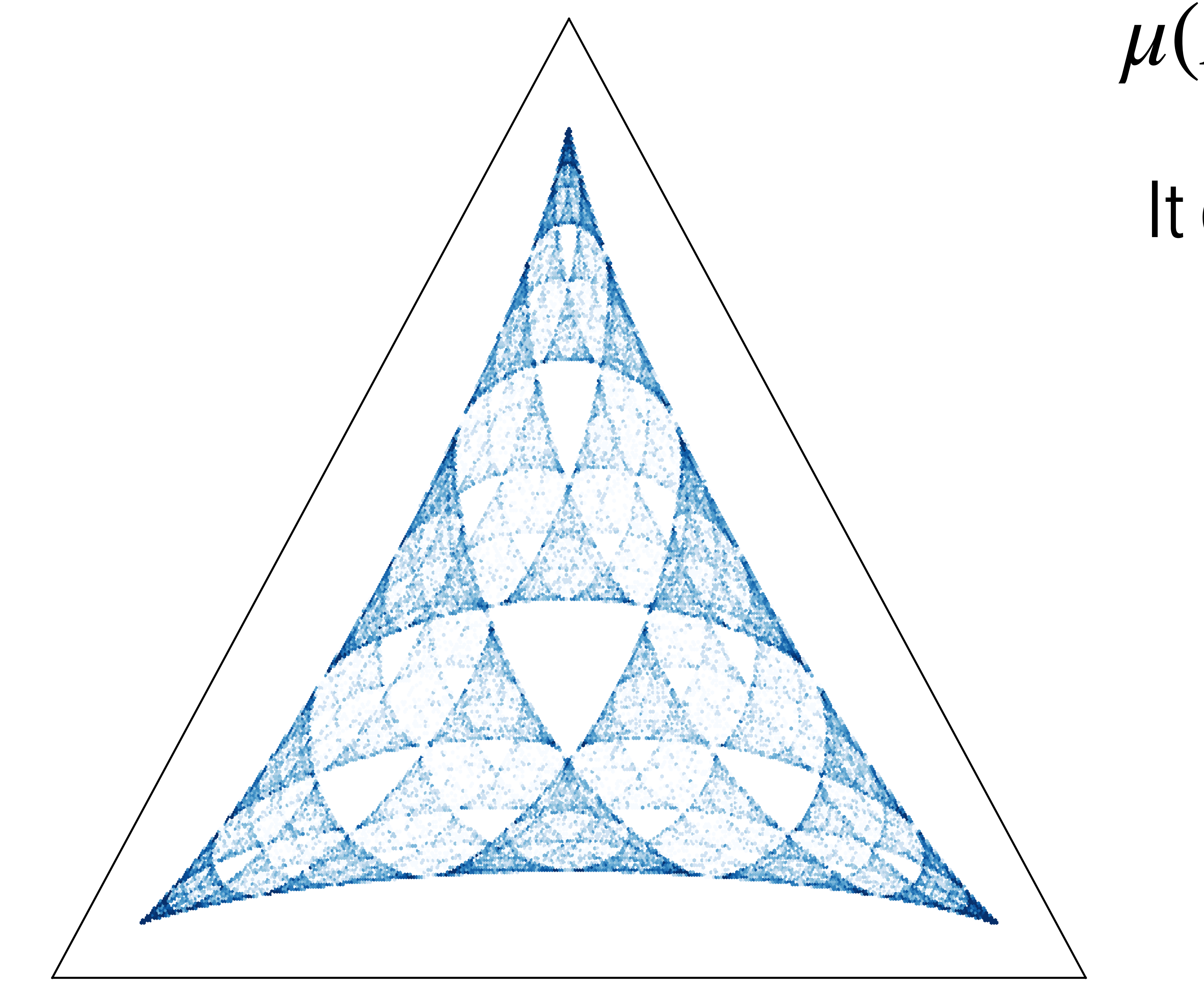
$\mu(R)$ is called the Blackwell measure.

It can be used to calculate the entropy rate:

$$h_\mu^B = \int_R \mathrm{d}\mu(\eta)\, H\left[X \,|\, \eta\right]$$

D. Blackwell. *The entropy of functions of finite-state Markov chains.* 1957.

Calculating Shannon Entropy Rate



$\mu(R)$ is called the Blackwell measure.

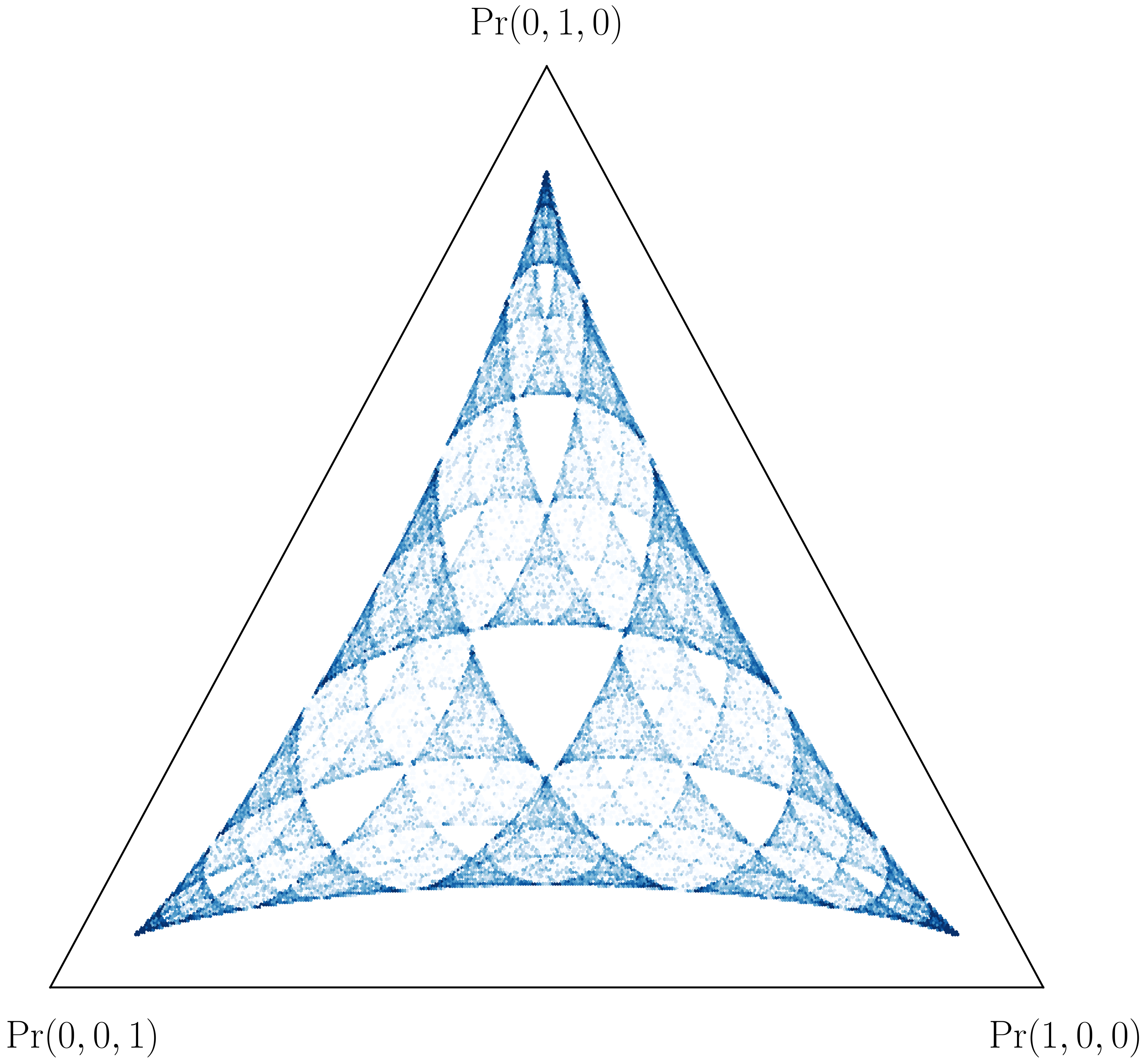
It can be used to calculate the entropy rate:

$$h_\mu^B = \int_R \mathrm{d}\mu(\eta)\, H\left[X \,|\, \eta\right]$$

$$\widehat{h_\mu^B} = \lim_{L\to\infty} \frac{1}{L} \sum_{t=0}^{L} H\left[X_t \,|\, \eta_t\right]$$

Alex M. Jurgens, James P. Crutchfield. *Shannon Entropy Rate of Hidden Markov Processes*. J. Stat. Phys., 183 (2), 1-18, 2020.
D. Blackwell. *The entropy of functions of finite-state Markov chains.* 1957.

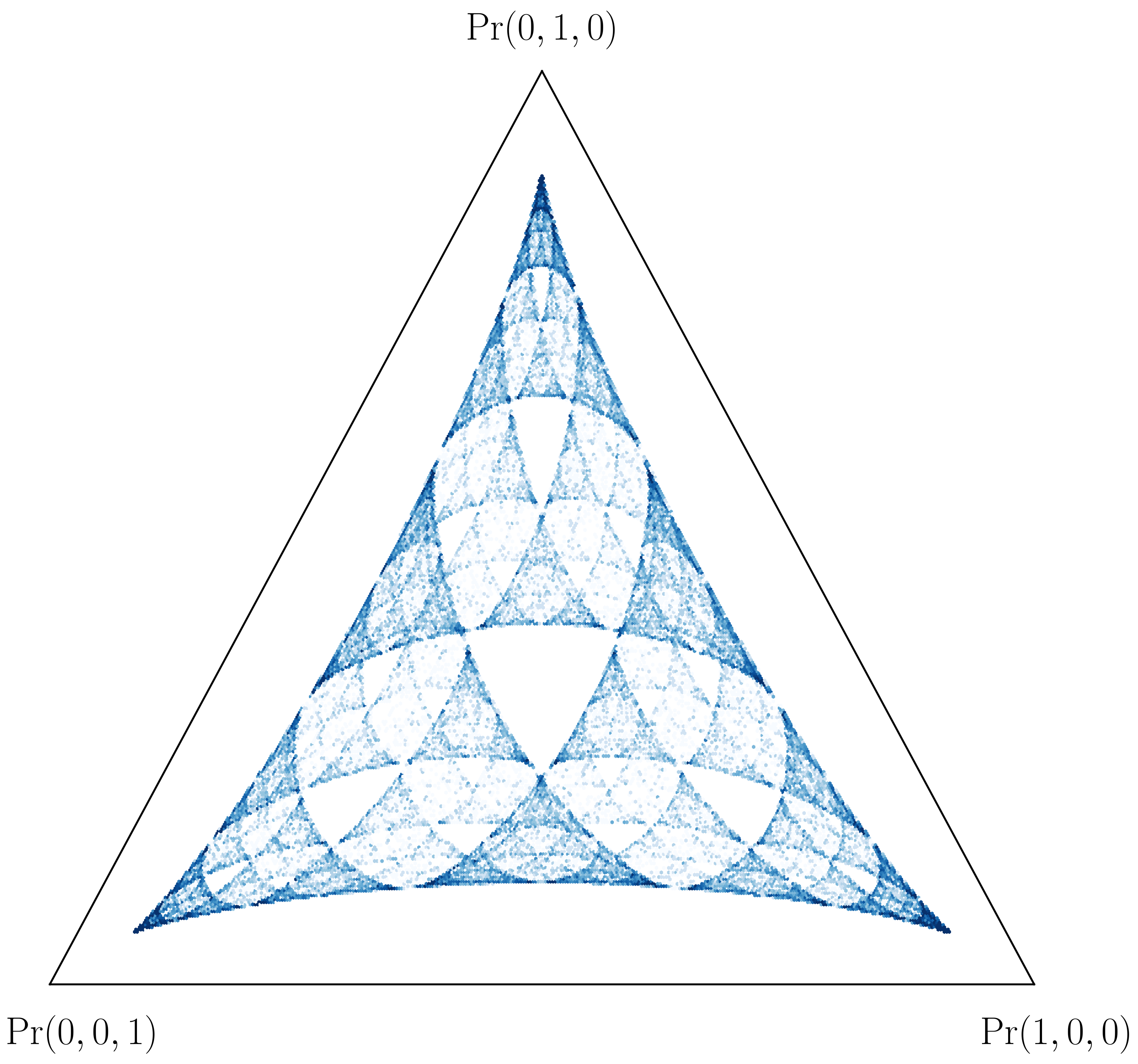
Structure: Statistical Complexity?



$$C_\mu \to \infty$$

Alexandra M. Jurgens, James P. Crutchfield. *Divergent Predictive Memory: The Statistical Complexity Dimension of Stationary, Ergodic Finite-State Hidden Markov Processes*. Chaos 31, 083114, 2021.
Alexandra M. Jurgens, James P. Crutchfield. *Ambiguity rate of hidden Markov processes*. Phys. Rev. E, 104 (2021)

Structure: Information Dimension



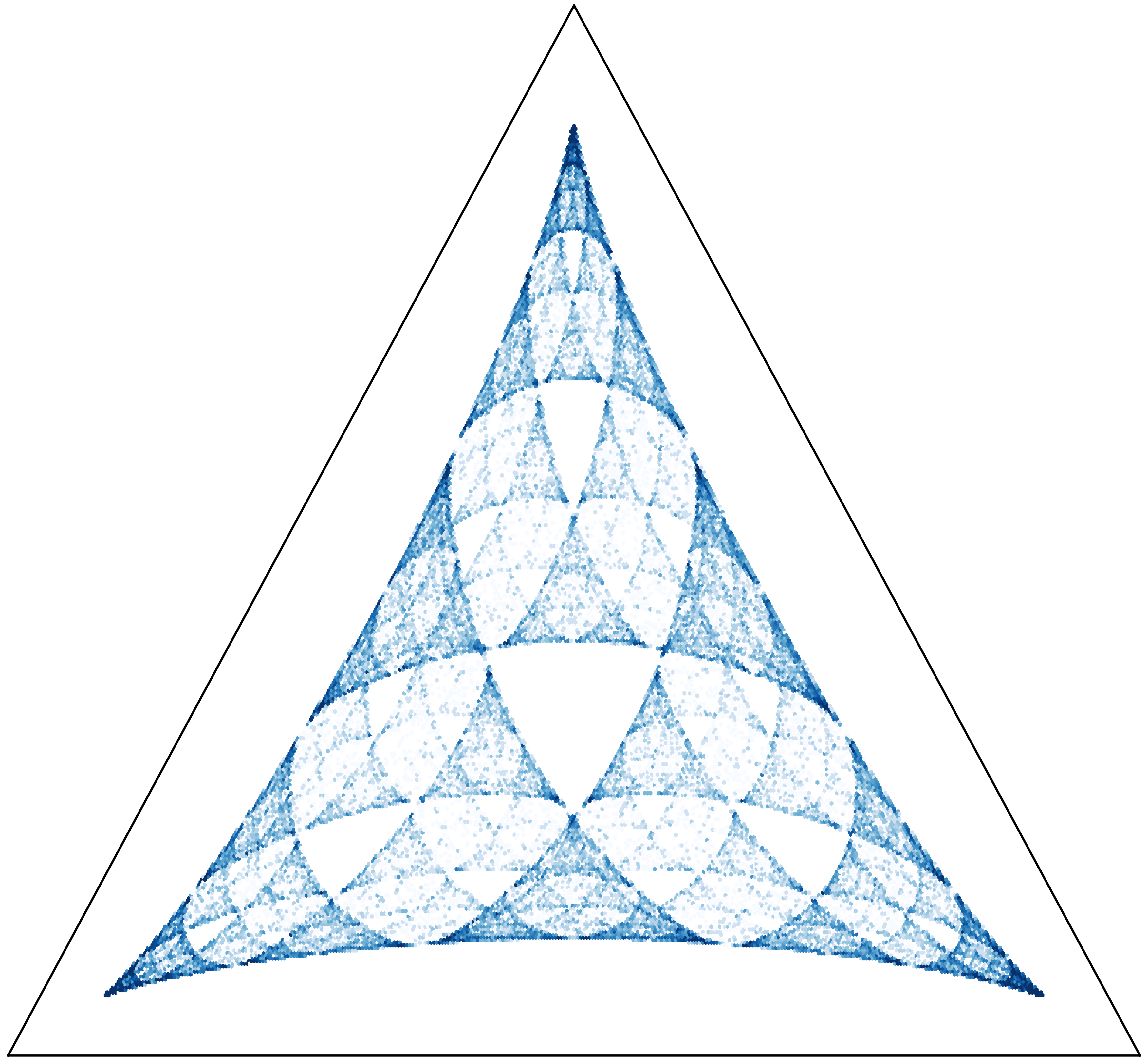
Information in a set

$$\dim_\mu(\mathsf{R}) \sim \frac{\Delta H\left[R_\epsilon\right]}{\Delta \ln \epsilon}$$

Scale of measurement

Alexandra M. Jurgens, James P. Crutchfield. *Divergent Predictive Memory: The Statistical Complexity Dimension of Stationary, Ergodic Finite-State Hidden Markov Processes*. Chaos 31, 083114, 2021.
Alexandra M. Jurgens, James P. Crutchfield. *Ambiguity rate of hidden Markov processes*. Phys. Rev. E, 104 (2021)

Statistical Complexity Dimension?

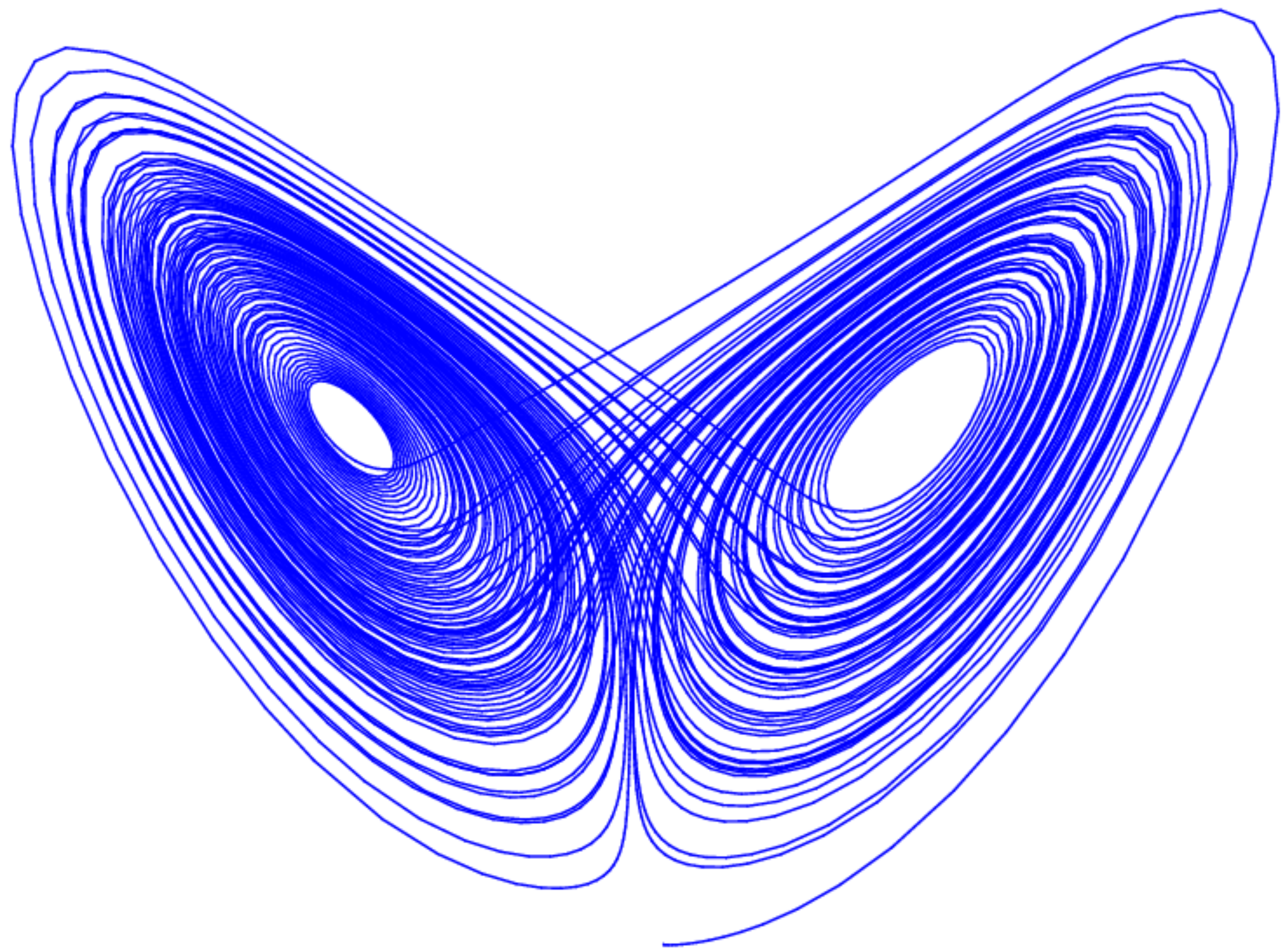


$$C_\mu \to \infty$$

$$\dim_\mu(R) \sim \frac{\Delta H\left[R_\epsilon\right]}{\Delta \ln \epsilon} = \frac{\Delta C_{\mu,\epsilon}}{\Delta \ln \epsilon}$$

Alexandra M. Jurgens, James P. Crutchfield. *Divergent Predictive Memory: The Statistical Complexity Dimension of Stationary, Ergodic Finite-State Hidden Markov Processes*. Chaos 31, 083114, 2021.
Alexandra M. Jurgens, James P. Crutchfield. *Ambiguity rate of hidden Markov processes*. Phys. Rev. E, 104 (2021)

Calculating Information Dimension



Calculate the Lyapunov spectrum:

$$\Gamma = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$$

$$\text{s.t.} \quad \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$$

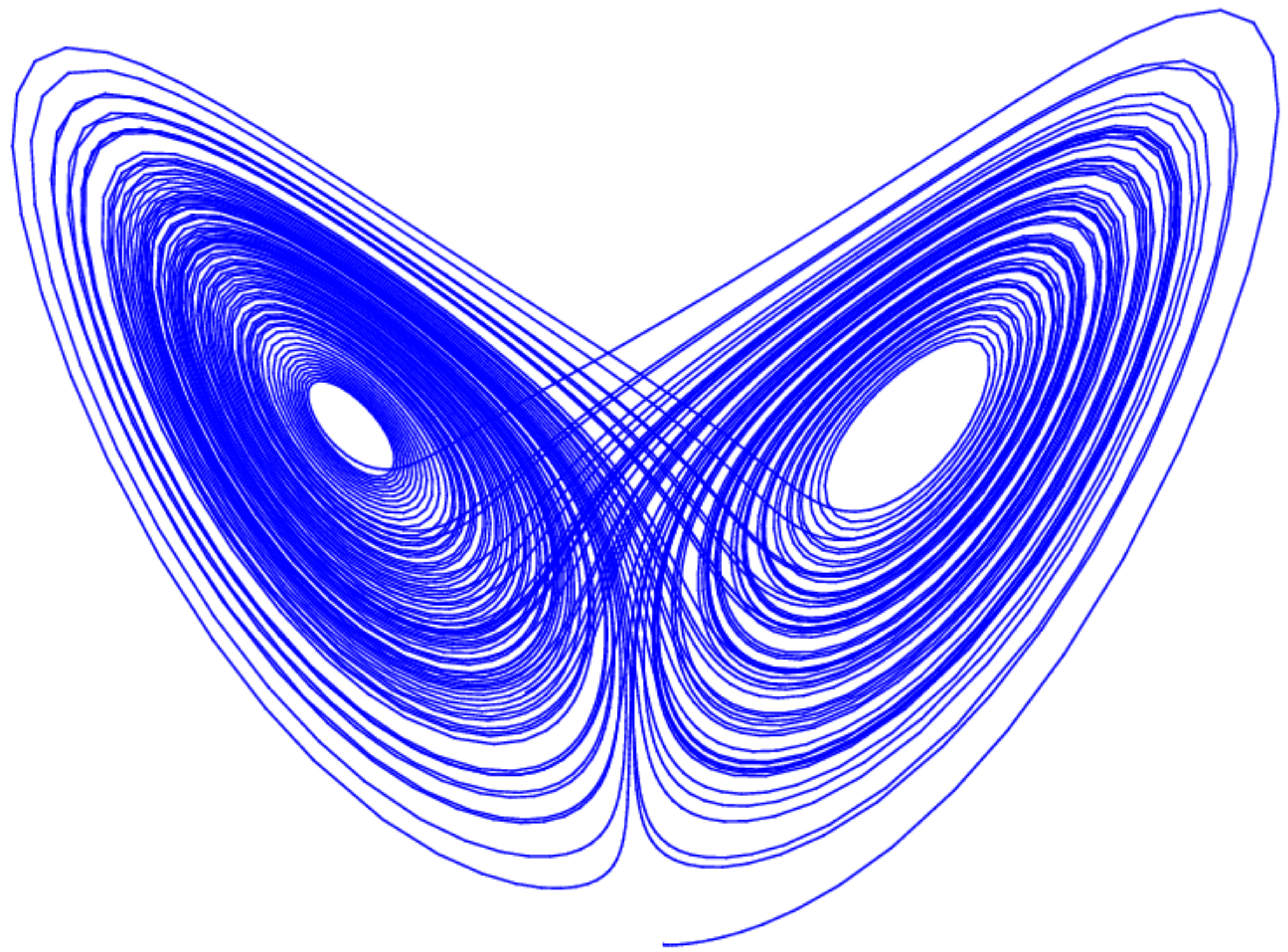
Kaplan–Yorke conjecture:

$$\dim_I = k + \frac{\sum_i^k \lambda_i}{|\lambda_{k+1}|}$$

Where k is the largest index for which the sum $\sum_i^k \lambda_i$ is positive.

P. Frederickson, J. Kaplan, E. Yorke, J. Yorke. (1983). *The Lyapunov Dimension of Strange Attractors. J. Diff. Eqs.* 49 (2): 185–207.

Calculating Information Dimension



For the Lorenz attractor with $\sigma = 10$, $r = 28$, $b = 8/3$:

$$\Gamma = \{0.90563, 0, -14.57219\}$$

So the Lyapunov dimension is:

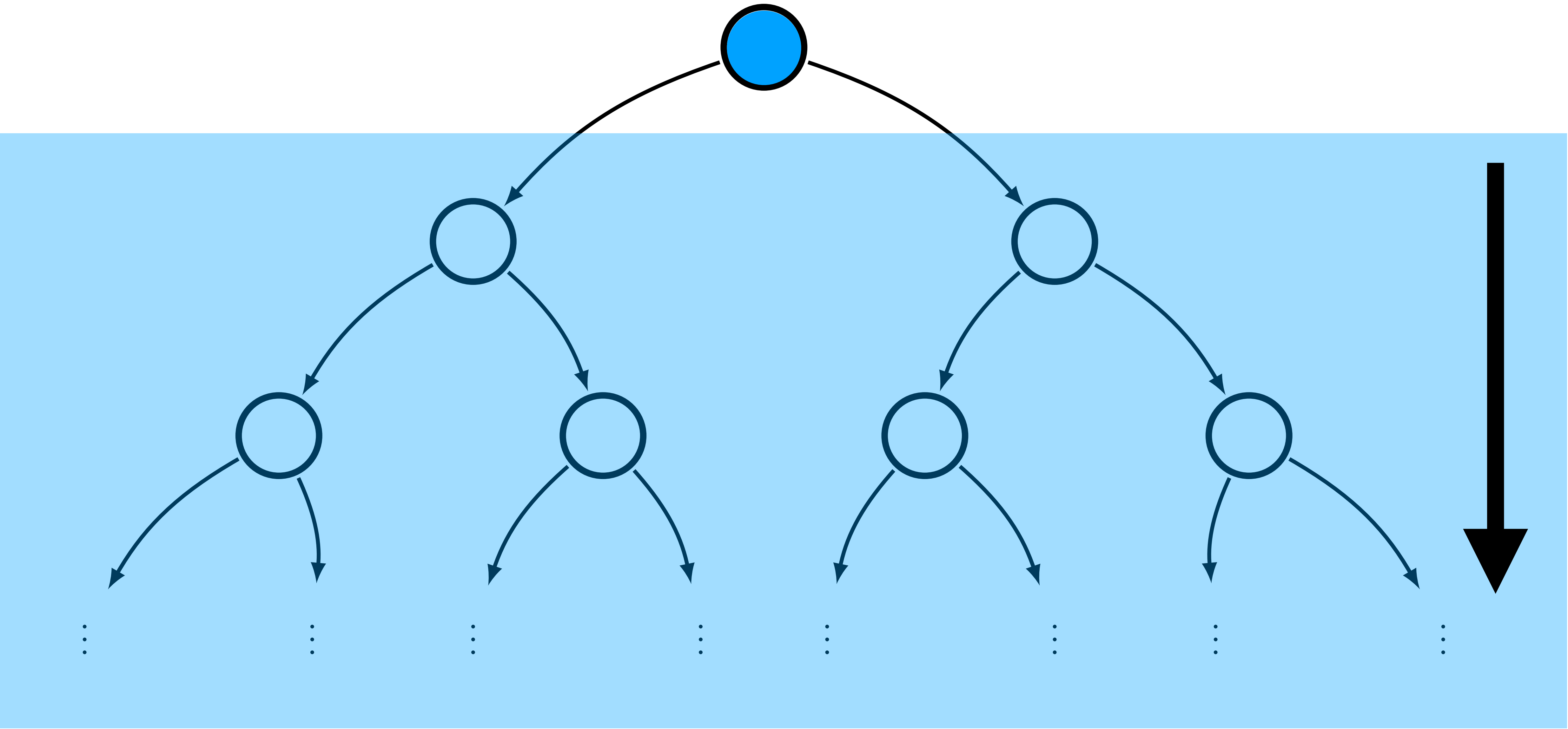
$$\dim_I = 2.06215$$

P. Frederickson, J. Kaplan, E. Yorke, J. Yorke. (1983). *The Lyapunov Dimension of Strange Attractors. J. Diff. Eqs.* 49 (2): 185–207.

New Quantity: The Ambiguity Rate

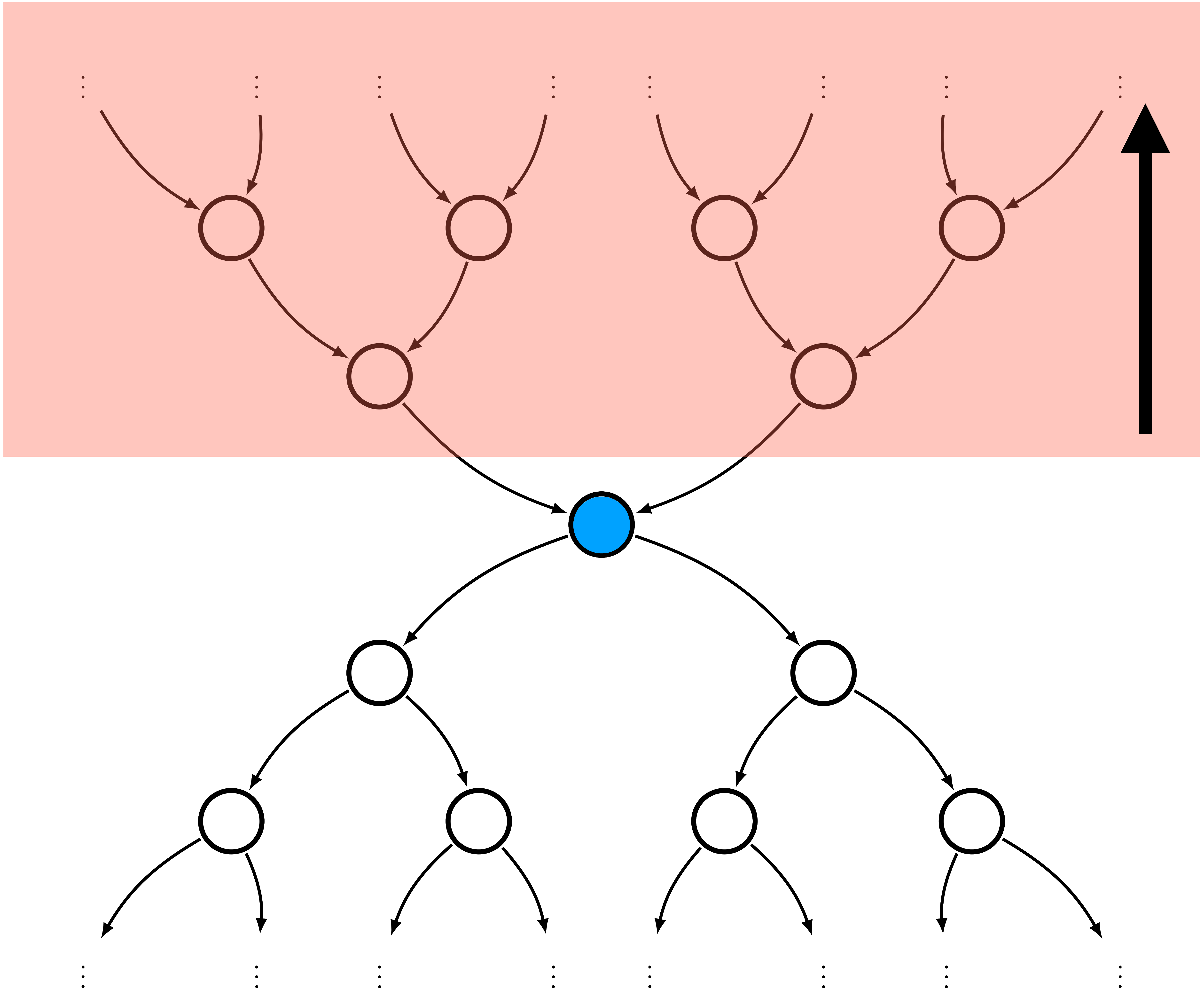
Entropy rate measures uncertainty in the next symbol given the present.

$$h_\mu = H\left[X_0, S_1 \mid S_0\right]$$



Alexandra M. Jurgens, James P. Crutchfield. *Ambiguity rate of hidden Markov processes*. Phys. Rev. E, 104 (2021)

New Quantity: The Ambiguity Rate



Entropy rate measures uncertainty in the next symbol given the present.

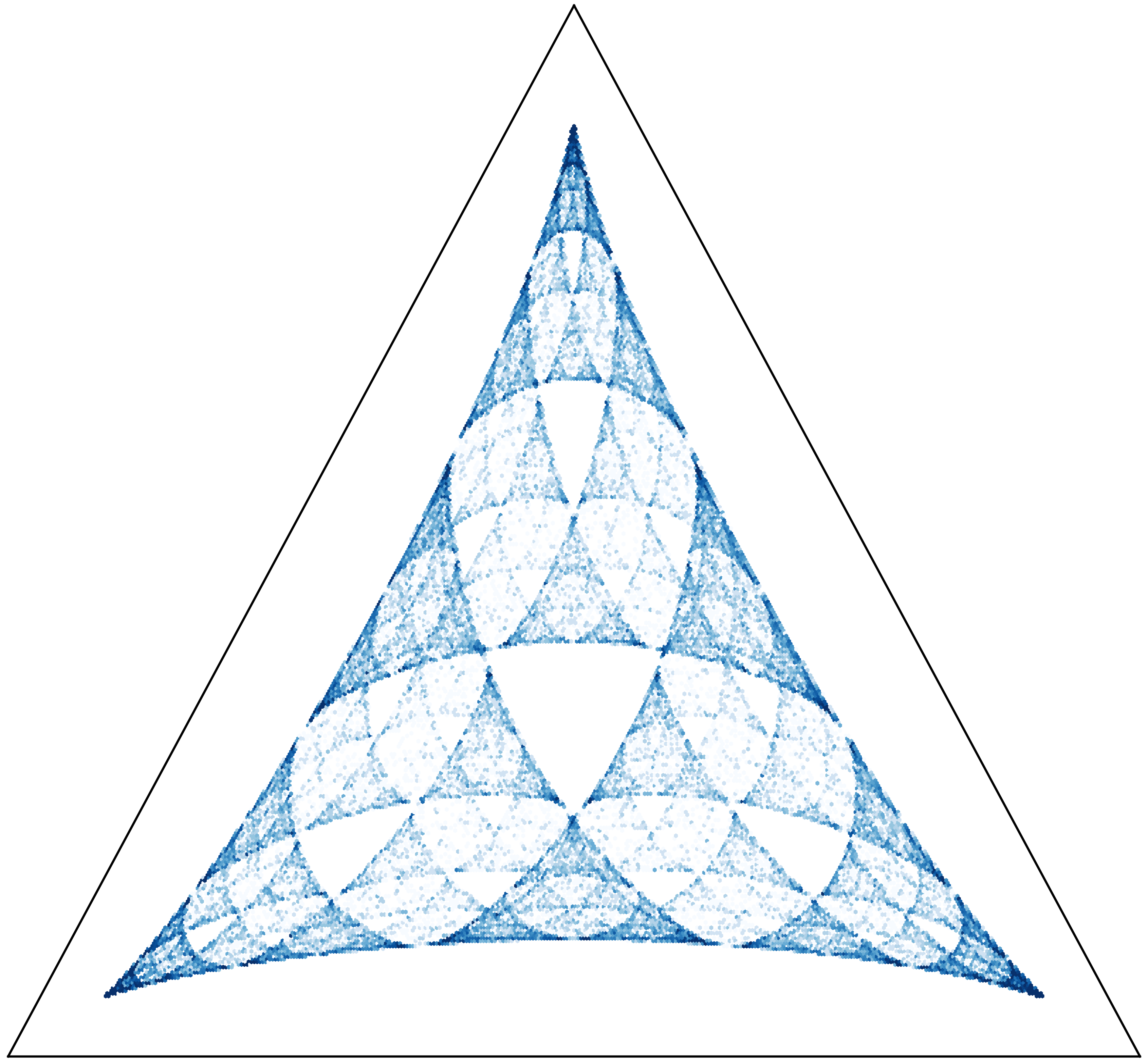
$$h_\mu = H\left[X_0, S_1 \mid S_0\right]$$

Ambiguity rate measures uncertainty *in prior symbol given the present*.

$$h_a = H\left[X_{-1}, \mathrm{S}_{-1} \mid \mathrm{S}_0\right]$$

Alexandra M. Jurgens, James P. Crutchfield. *Ambiguity rate of hidden Markov processes*. Phys. Rev. E, 104 (2021)

Statistical Complexity Dimension

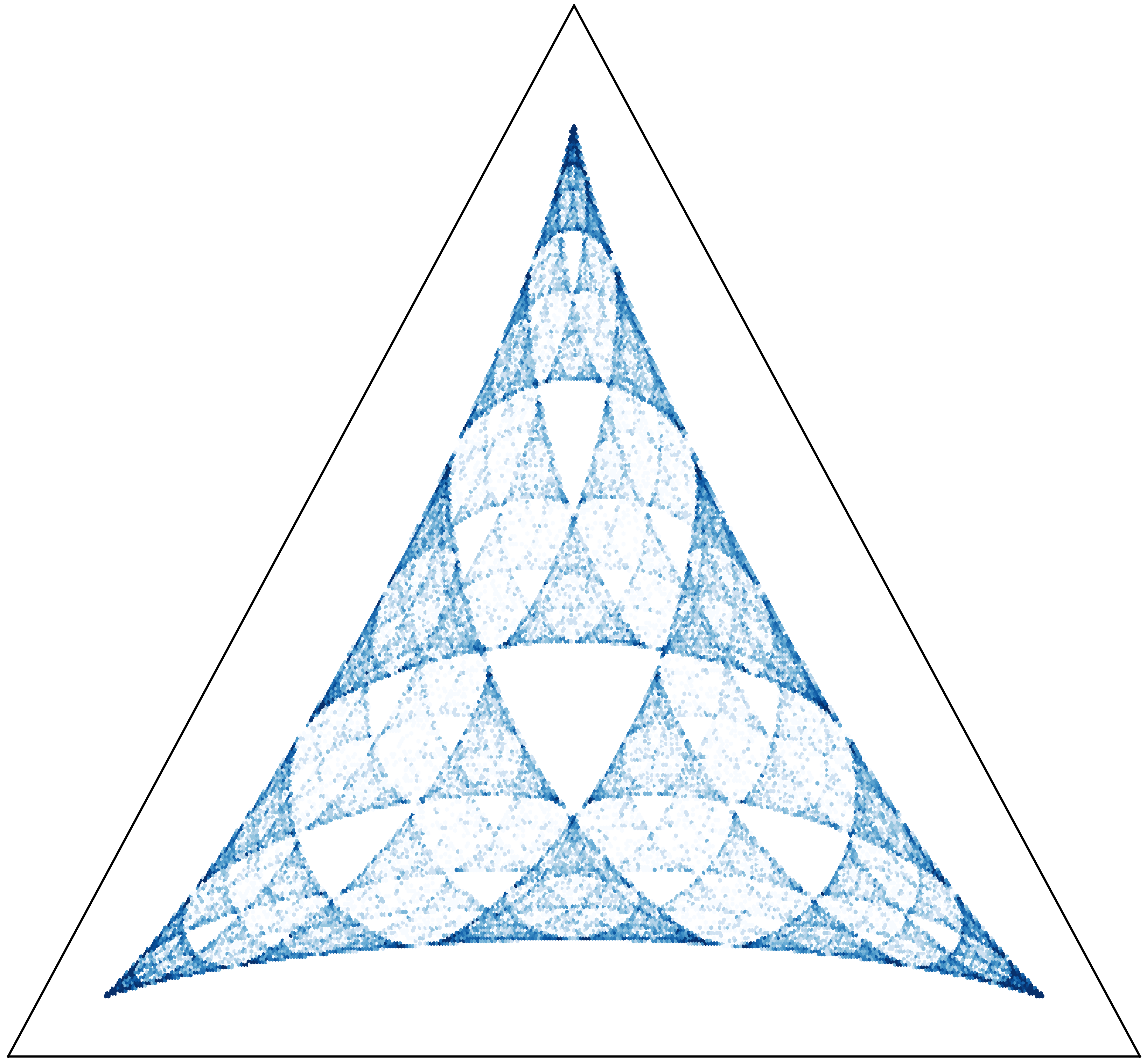


$$\dim_\mu(R) = k + \frac{h_\mu - h_a + \sum_i^k \lambda_i}{|\lambda_{k+1}|}$$

Alexandra M. Jurgens, James P. Crutchfield. *Divergent Predictive Memory: The Statistical Complexity Dimension of Stationary, Ergodic Finite-State Hidden Markov Processes*. Chaos 31, 083114, 2021.

Alexandra M. Jurgens, James P. Crutchfield. *Ambiguity rate of hidden Markov processes*. Phys. Rev. E, 104 (2021)

Statistical Complexity Dimension



Entropy rate

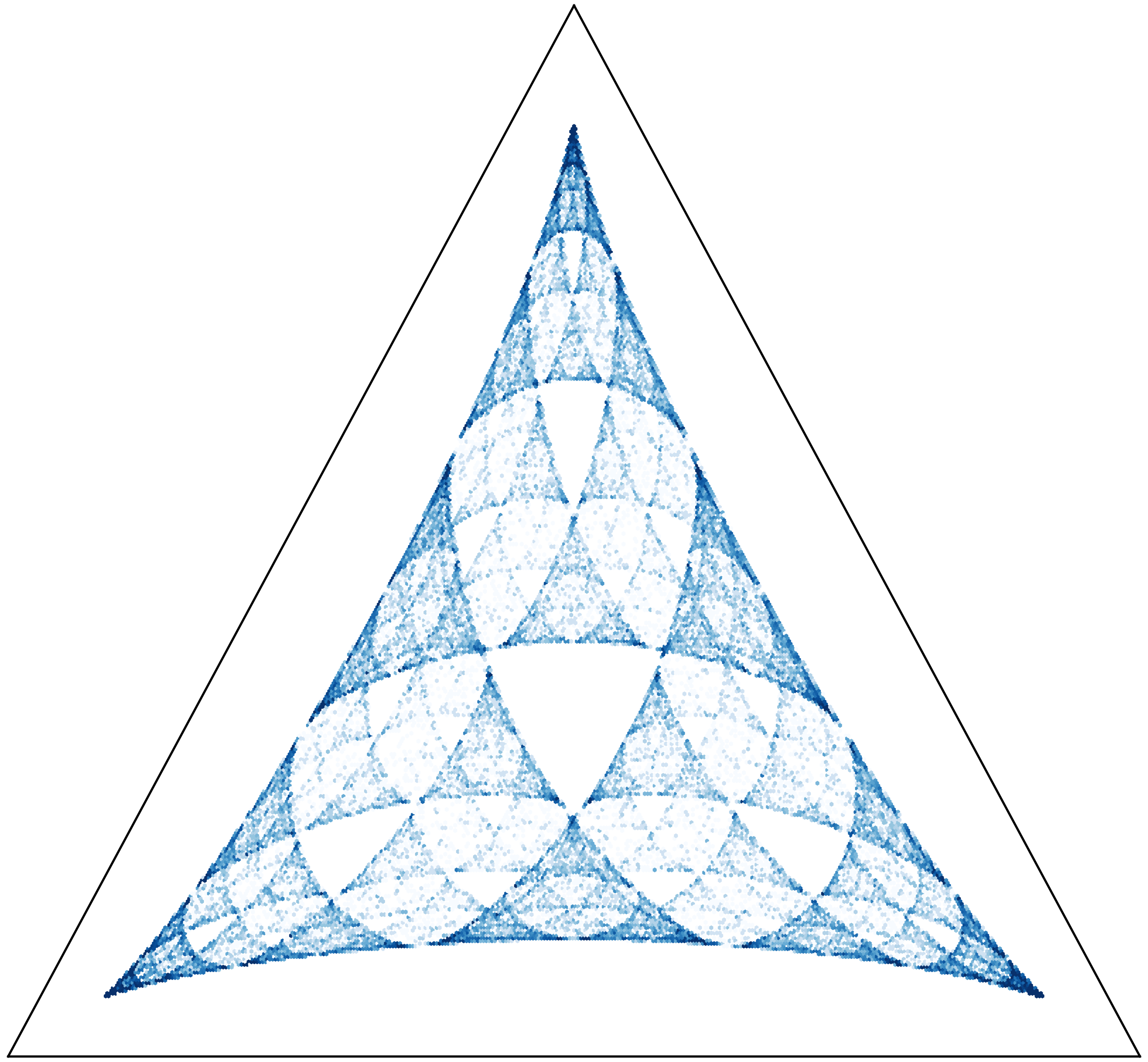
Ambiguity rate

$$\dim_\mu(R) = k + \frac{h_\mu - h_a + \sum_i^k \lambda_i}{|\lambda_{k+1}|}$$

Alexandra M. Jurgens, James P. Crutchfield. *Divergent Predictive Memory: The Statistical Complexity Dimension of Stationary, Ergodic Finite-State Hidden Markov Processes*. Chaos 31, 083114, 2021.

Alexandra M. Jurgens, James P. Crutchfield. *Ambiguity rate of hidden Markov processes*. Phys. Rev. E, 104 (2021)

Statistical Complexity Dimension



$$C_\mu \to \infty$$

$$\dim_\mu(R) = \frac{\Delta C_{\mu,\epsilon}}{\Delta \ln \epsilon} = k + \frac{h_\mu - h_a + \sum_i^k \lambda_i}{|\lambda_{k+1}|}$$

→ Can calculate randomness and structure, now for infinite states!

Alexandra M. Jurgens, James P. Crutchfield. *Divergent Predictive Memory: The Statistical Complexity Dimension of Stationary, Ergodic Finite-State Hidden Markov Processes*. Chaos 31, 083114, 2021.

Alexandra M. Jurgens, James P. Crutchfield. *Ambiguity rate of hidden Markov processes*. Phys. Rev. E, 104 (2021)

Growth Rate of Predictive Models

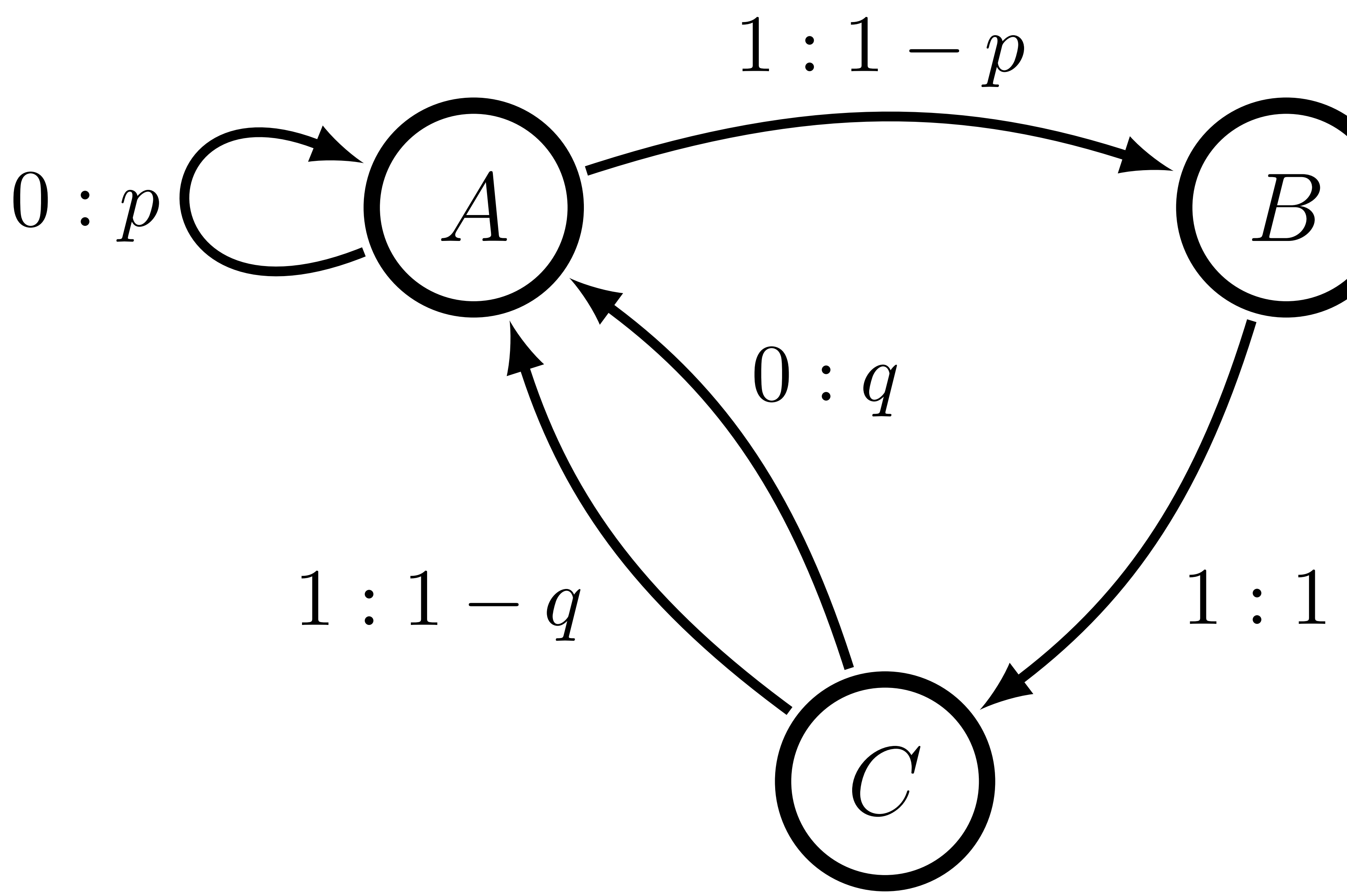
$$\Delta H\left[\text{Predictive states}\right] = h_\mu - h_a$$

Entropy rate

Ambiguity rate

Alexandra M. Jurgens, James P. Crutchfield. *Divergent Predictive Memory: The Statistical Complexity Dimension of Stationary, Ergodic Finite-State Hidden Markov Processes*. Chaos 31, 083114, 2021.
Alexandra M. Jurgens, James P. Crutchfield. *Ambiguity rate of hidden Markov processes*. Phys. Rev. E, 104 (2021)

Growth Rate of Predictive Models



For a finite state model:

$$h_\mu - h_a = 0$$

Alexandra M. Jurgens, James P. Crutchfield. *Divergent Predictive Memory: The Statistical Complexity Dimension of Stationary, Ergodic Finite-State Hidden Markov Processes*. Chaos 31, 083114, 2021.

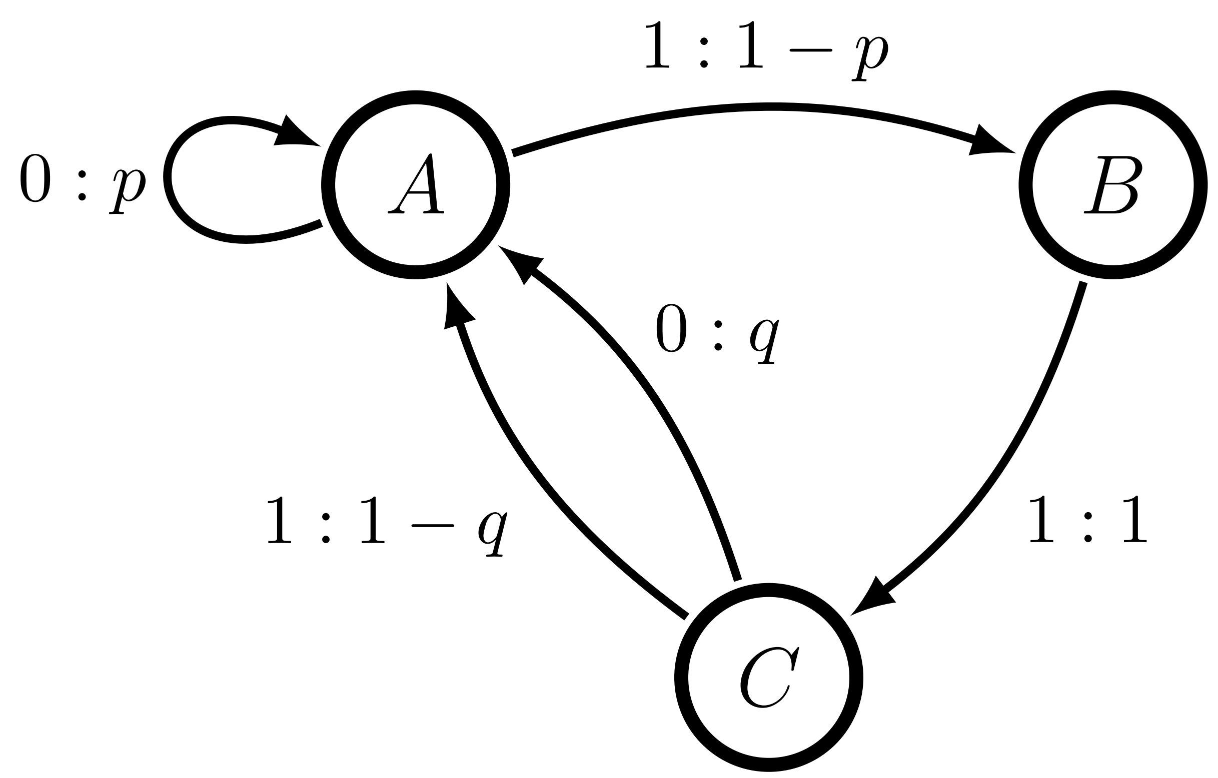
Alexandra M. Jurgens, James P. Crutchfield. *Ambiguity rate of hidden Markov processes*. Phys. Rev. E, 104 (2021)

Growth Rate of Predictive Models

$$\Delta H \left[\text{Predictive states}\right] = h_\mu - h_a$$

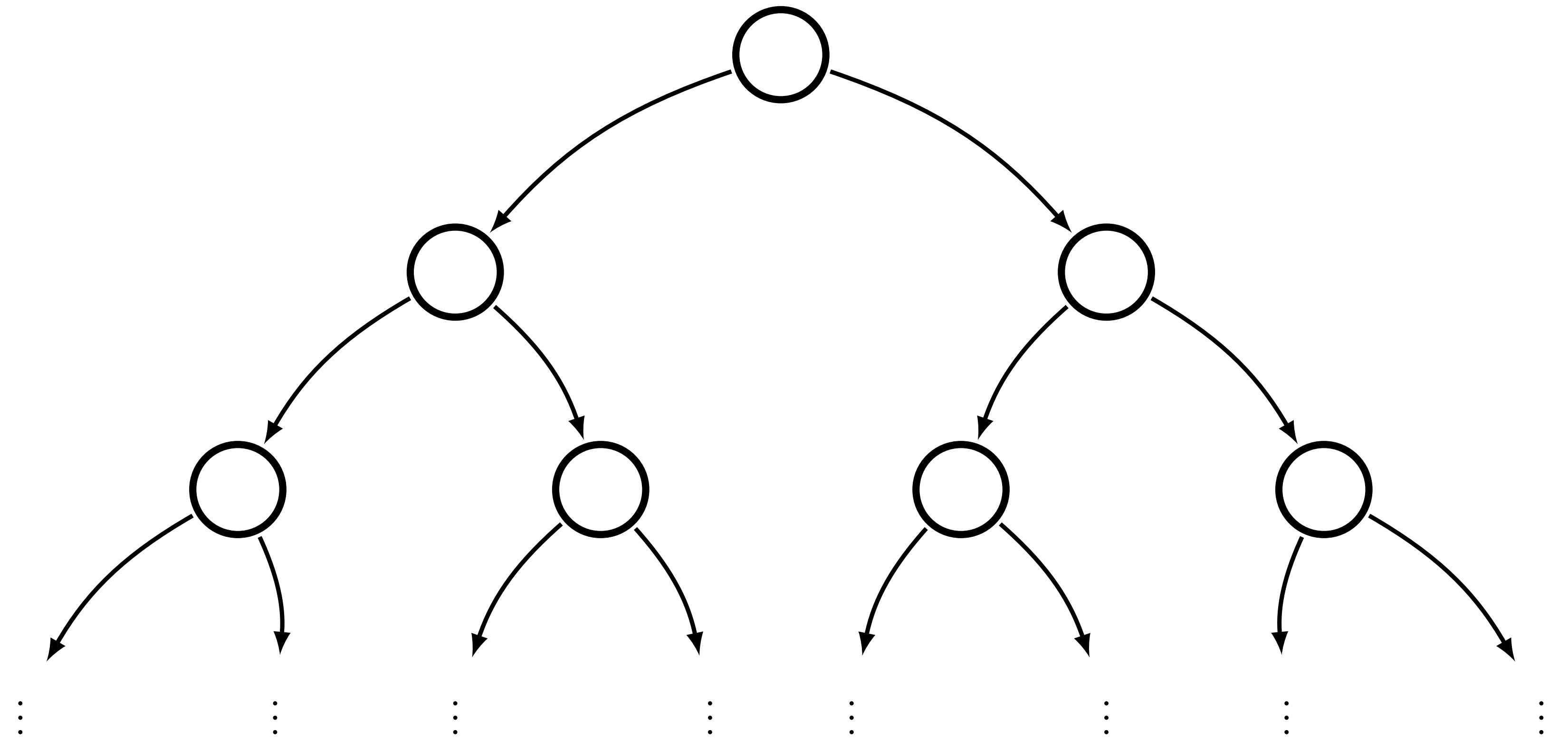
Finite states

$$h_\mu - h_a = 0$$



"Every history counts"

$$h_\mu - h_a = h_\mu$$



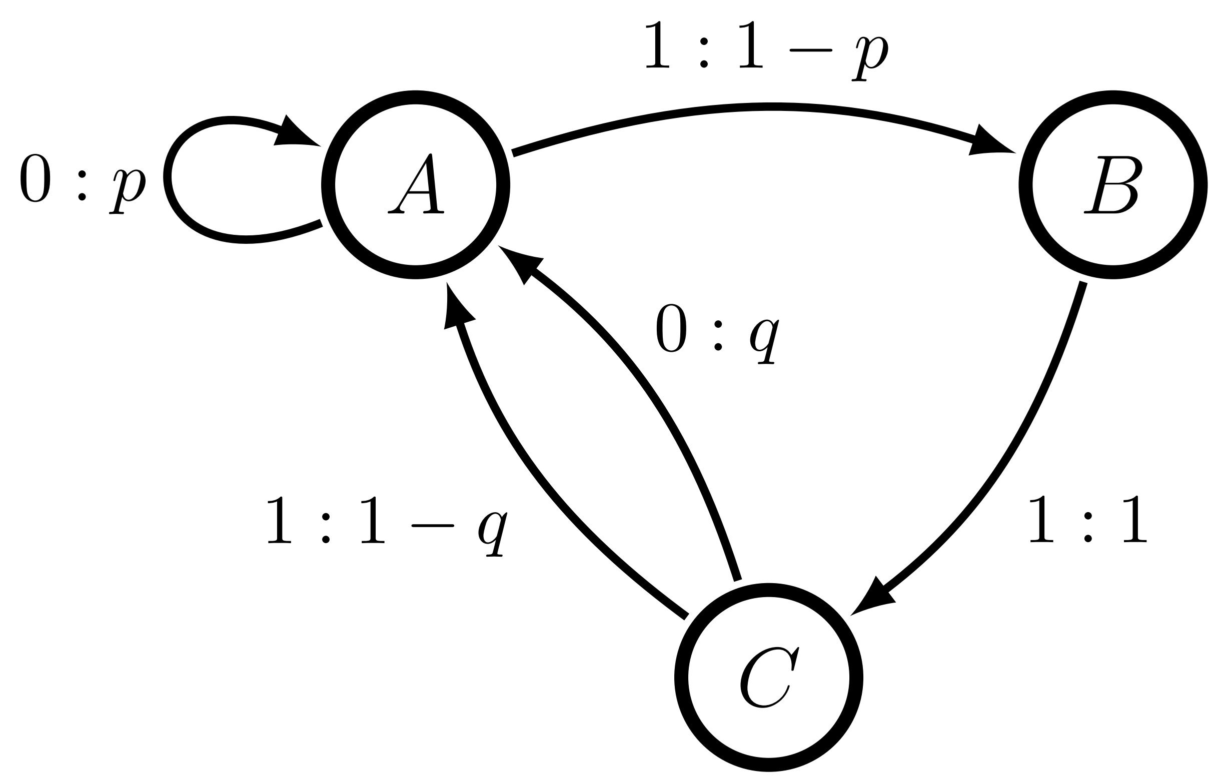
Alexandra M. Jurgens, James P. Crutchfield. *Divergent Predictive Memory: The Statistical Complexity Dimension of Stationary, Ergodic Finite-State Hidden Markov Processes*. Chaos 31, 083114, 2021.
Alexandra M. Jurgens, James P. Crutchfield. *Ambiguity rate of hidden Markov processes*. Phys. Rev. E, 104 (2021)

Growth Rate of Predictive Models

$$\Delta H\left[\text{Predictive states}\right] = h_\mu - h_a$$

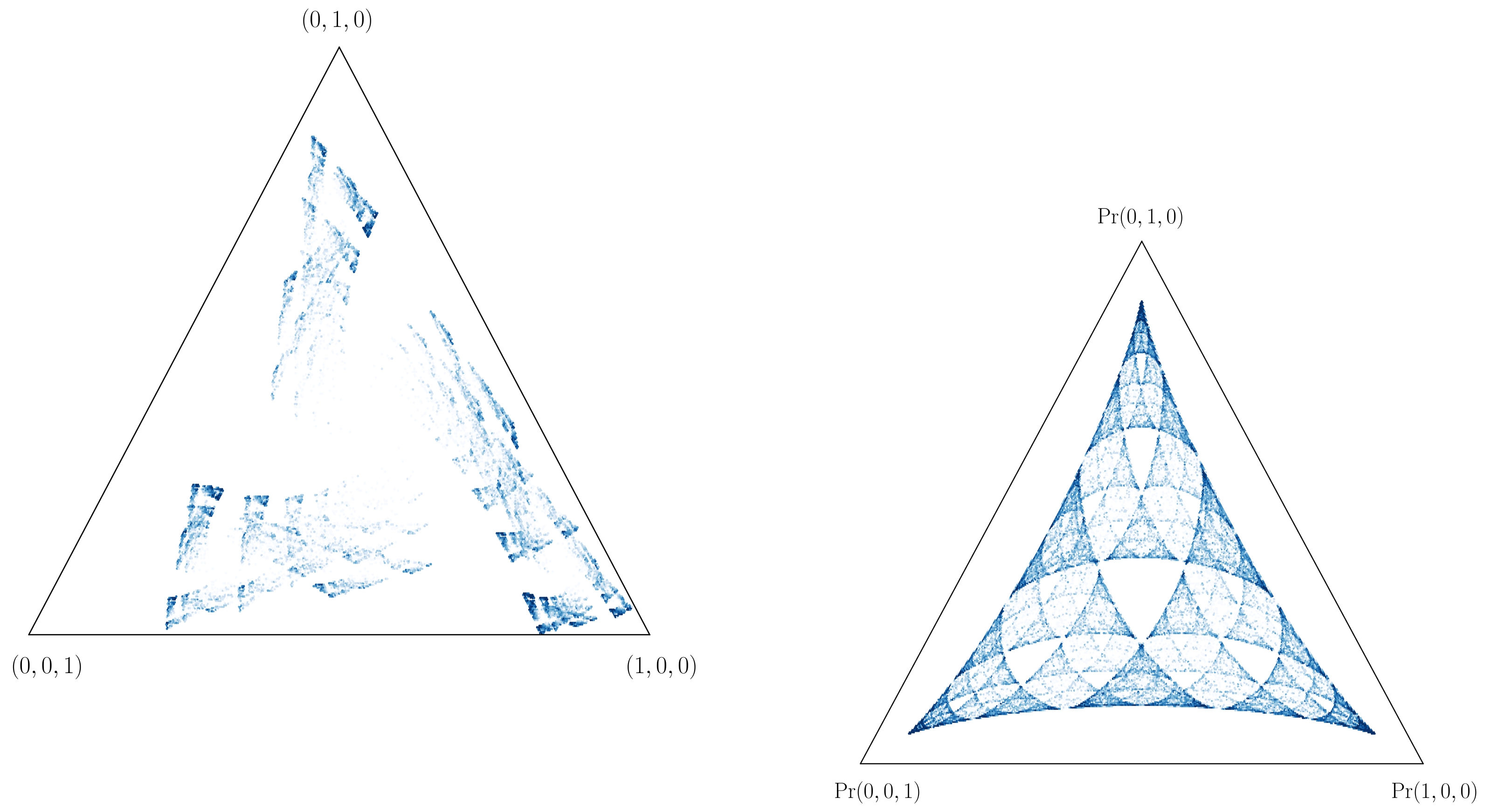
Finite states

$$h_\mu - h_a = 0$$



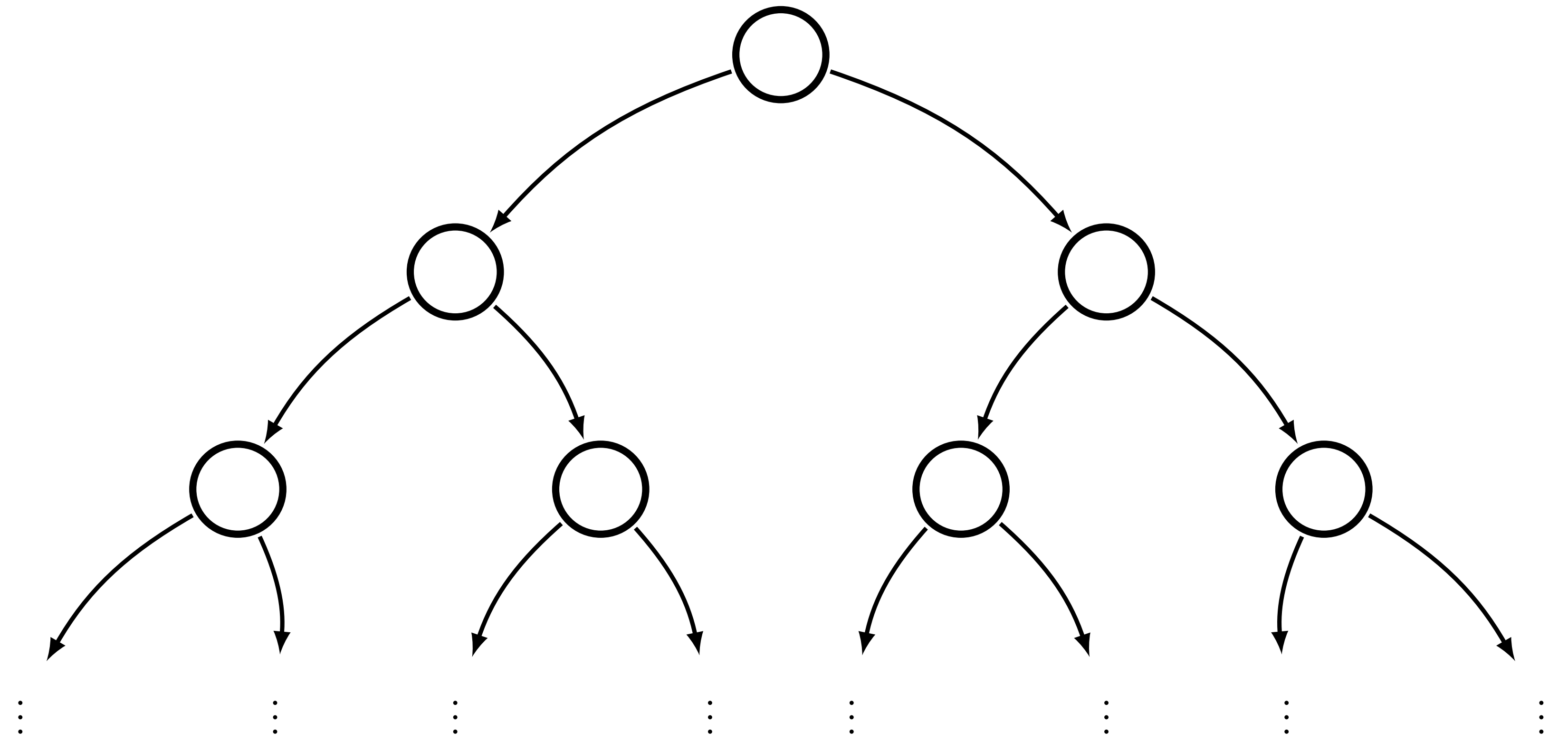
In general:

$$h_\mu > h_\mu - h_a > 0$$



"Every history counts"

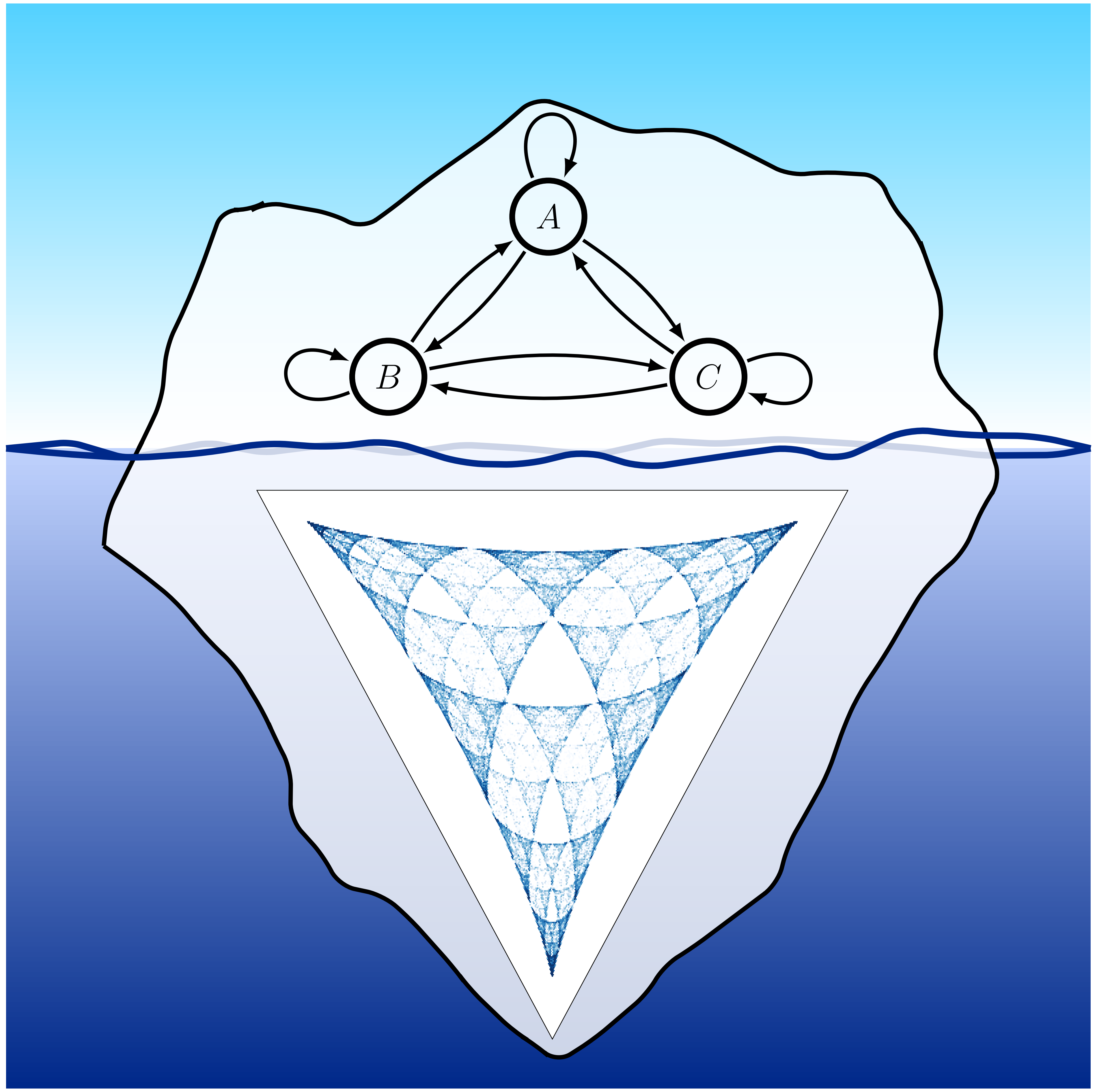
$$h_\mu - h_a = h_\mu$$



Alexandra M. Jurgens, James P. Crutchfield. *Divergent Predictive Memory: The Statistical Complexity Dimension of Stationary, Ergodic Finite-State Hidden Markov Processes*. Chaos 31, 083114, 2021.

Alexandra M. Jurgens, James P. Crutchfield. *Ambiguity rate of hidden Markov processes*. Phys. Rev. E, 104 (2021)

Acknowledgements + Questions



Thank you to my PhD advisor, Jim Crutchfield, and all those at the Complexity Sciences Center, especially Alec Boyd, Adam Rupe, Sam Loomis and for helpful discussions.

Thank you as well to my curent group leader at INRIA Bordeaux, Nicolas Brodu.

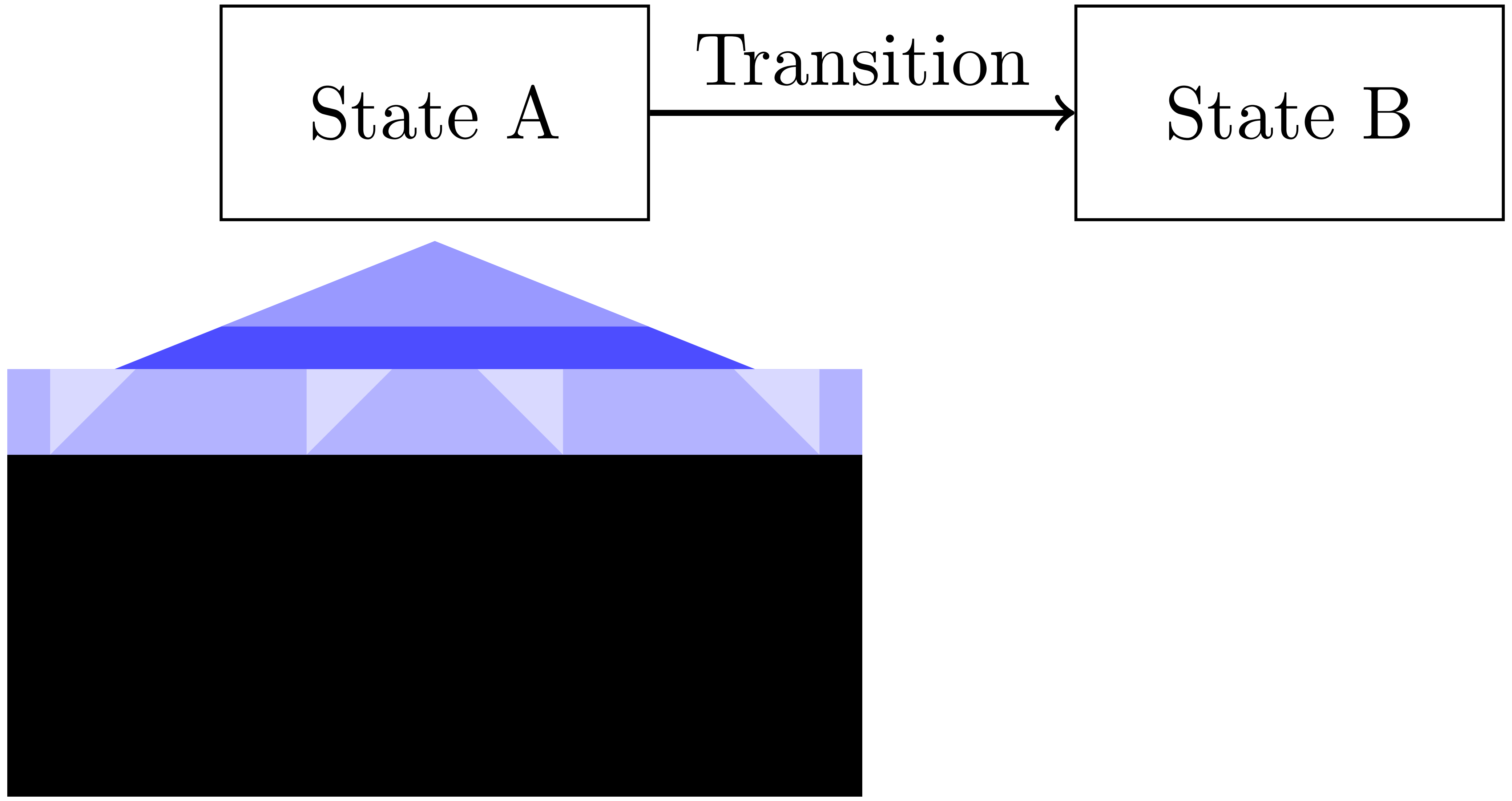
Contact:

Website: https://csc.ucdavis.edu/~ajurgens/
Email: alexandra.jurgens@inria.fr
Twitter: @amjurgens

Acknowledgements + Questions

Create tikz code depicting an iceberg, but overlay a finite state machine on the above water portion and a black rectangle on the underwater portion. Include code to replace the rectangle by an inserted image.

Certainly! Below is the TikZ code to depict an iceberg with a finite state machine (FSM) overlay on the above water portion and a black rectangle on the underwater portion. Additionally, it includes the code to replace the black rectangle with an inserted image.



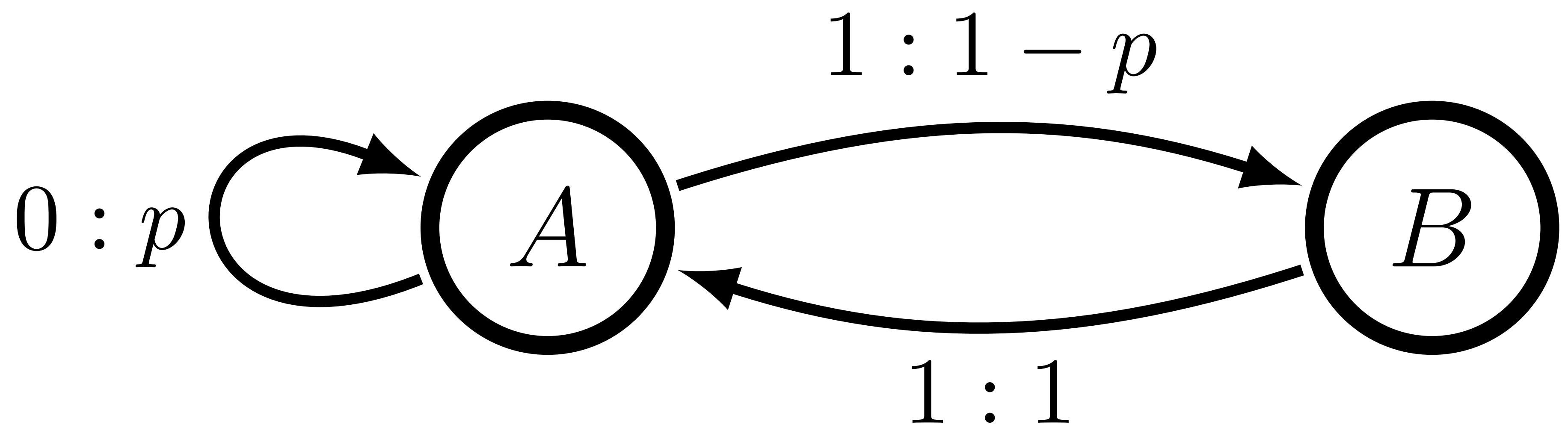
Thank you to my PhD advisor, Jim Crutchfield, and all those at the Complexity Sciences Center, especially Alec Boyd, Adam Rupe, Sam Loomis and for helpful discussions.

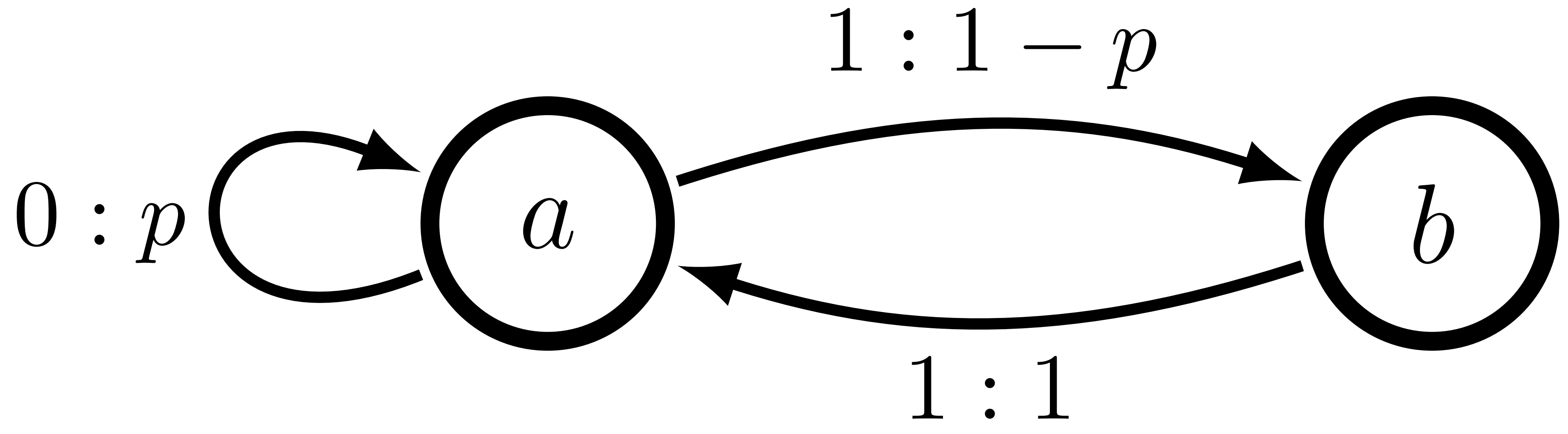
Thank you as well to my curent group leader at INRIA Bordeaux, Nicolas Brodu.

Contact:

Website: https://csc.ucdavis.edu/~ajurgens/
Email: alexandra.jurgens@inria.fr
Twitter: @amjurgens

The Utility of (Good) Models



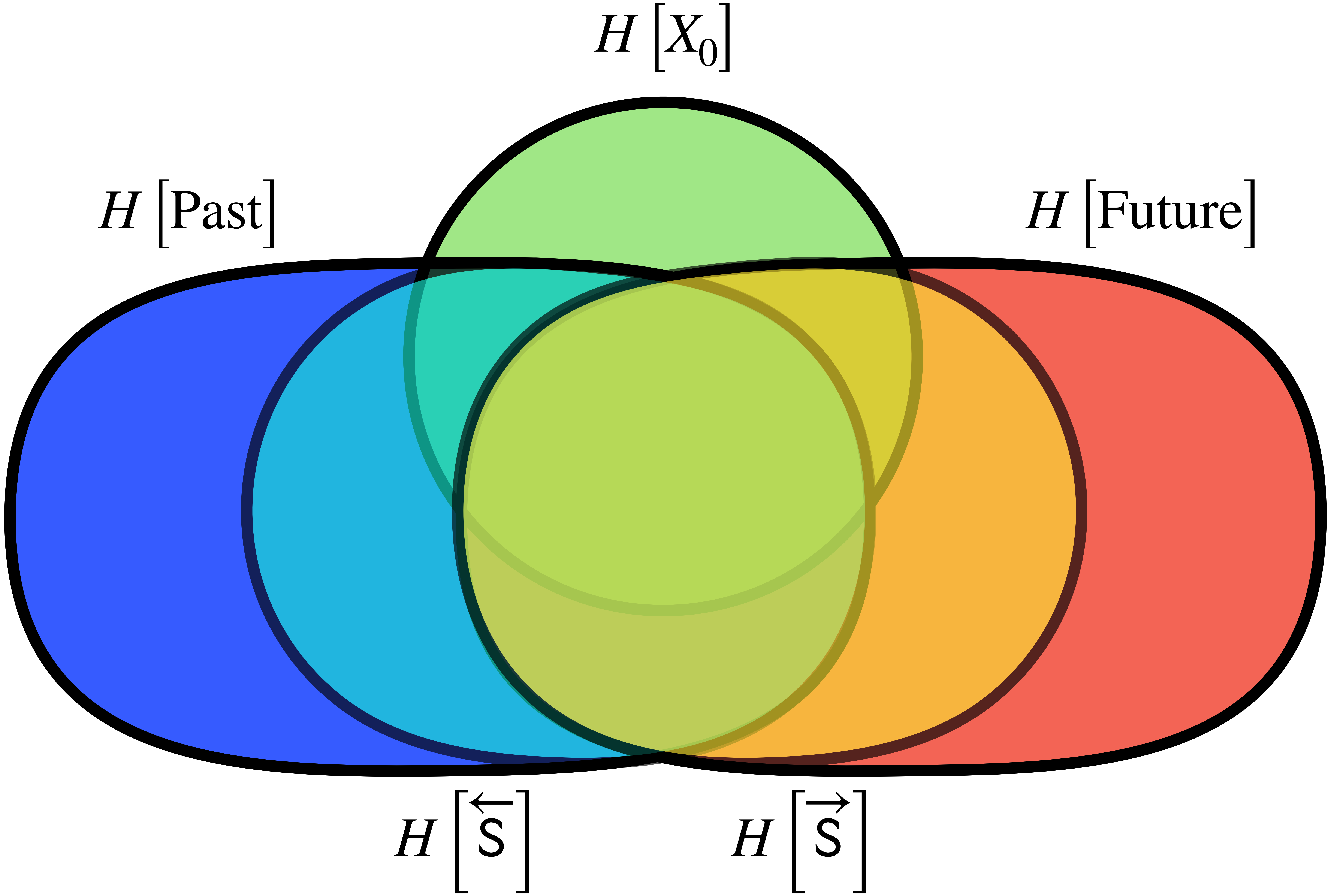


$$h_\mu = \sum_{\sigma \in \mathcal{S}} \Pr(\sigma)\, H\left[X \mid \sigma\right]$$

$$E = I\left[\overrightarrow{\mathcal{S}} ; \overleftarrow{\mathcal{S}}\right]$$

$$\sigma_\mu = I\left[\overrightarrow{\mathcal{S}} ; \overleftarrow{\mathcal{S}} \mid X_0\right]$$

The Utility of (Good) Models



Ryan G. James, Christopher J. Ellison, James P. Crutchfield. *Anatomy of a bit: Information in a time series observation.* Chaos 21, 037109 (2011)

James P. Crutchfield, David P. Feldman. *Regularities unseen, randomness observed: Levels of entropy convergence.* Chaos 1 March 2003; 13 (1): 25–54.