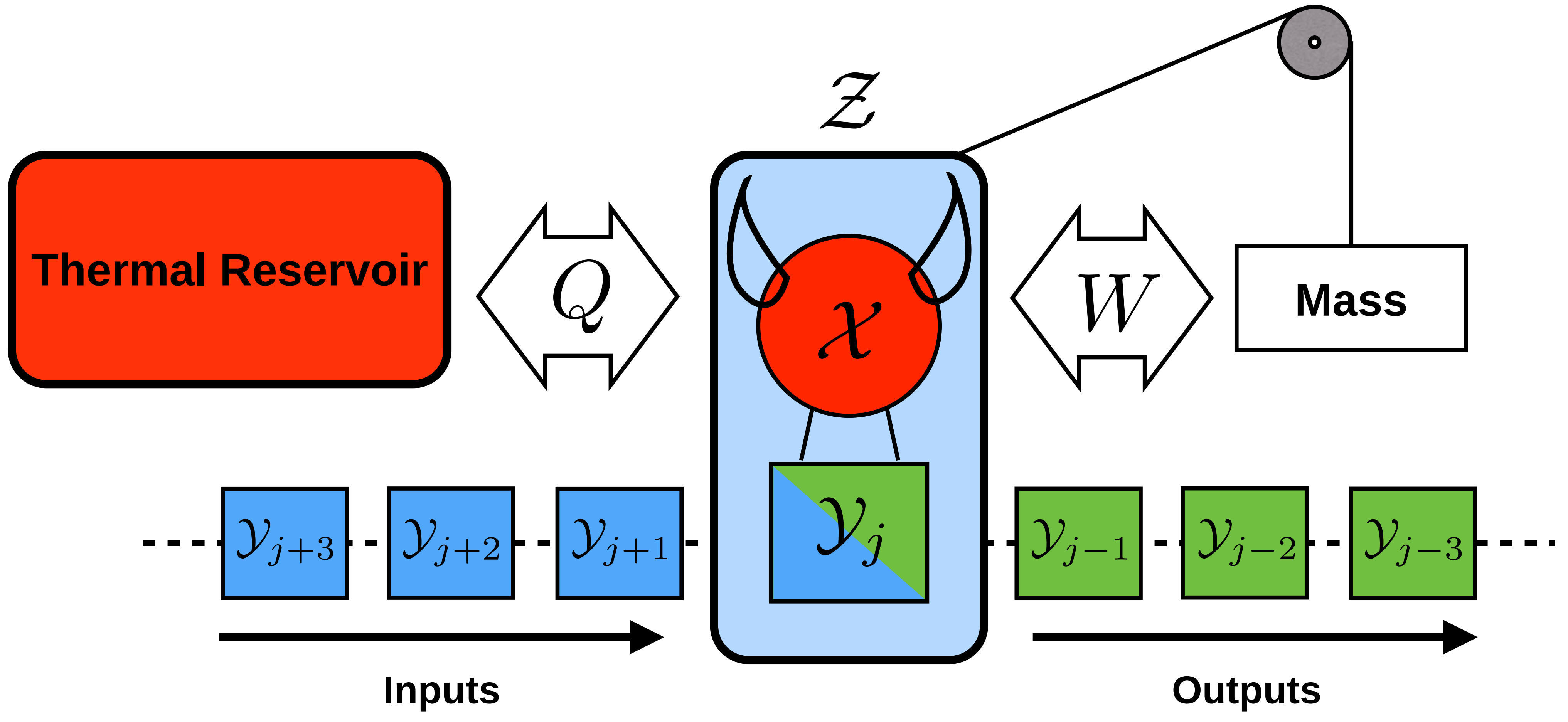
Thermodynamic Overfitting: Limits on Complexity in Thermodynamic Learning

Alec Boyd

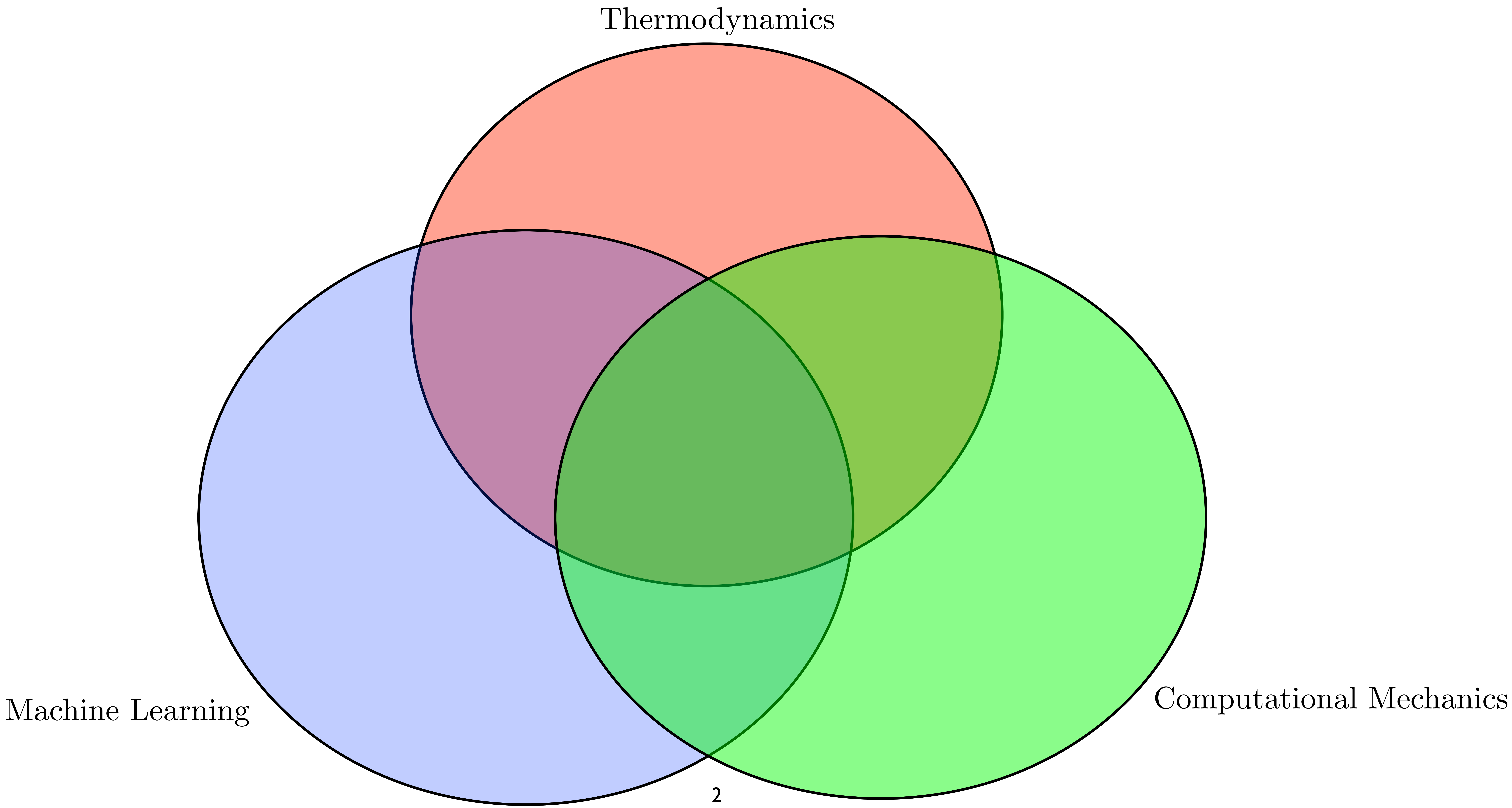
11 Sept, 2023

Information Theory as a Bridge Across the Geosciences and Modeling Sciences

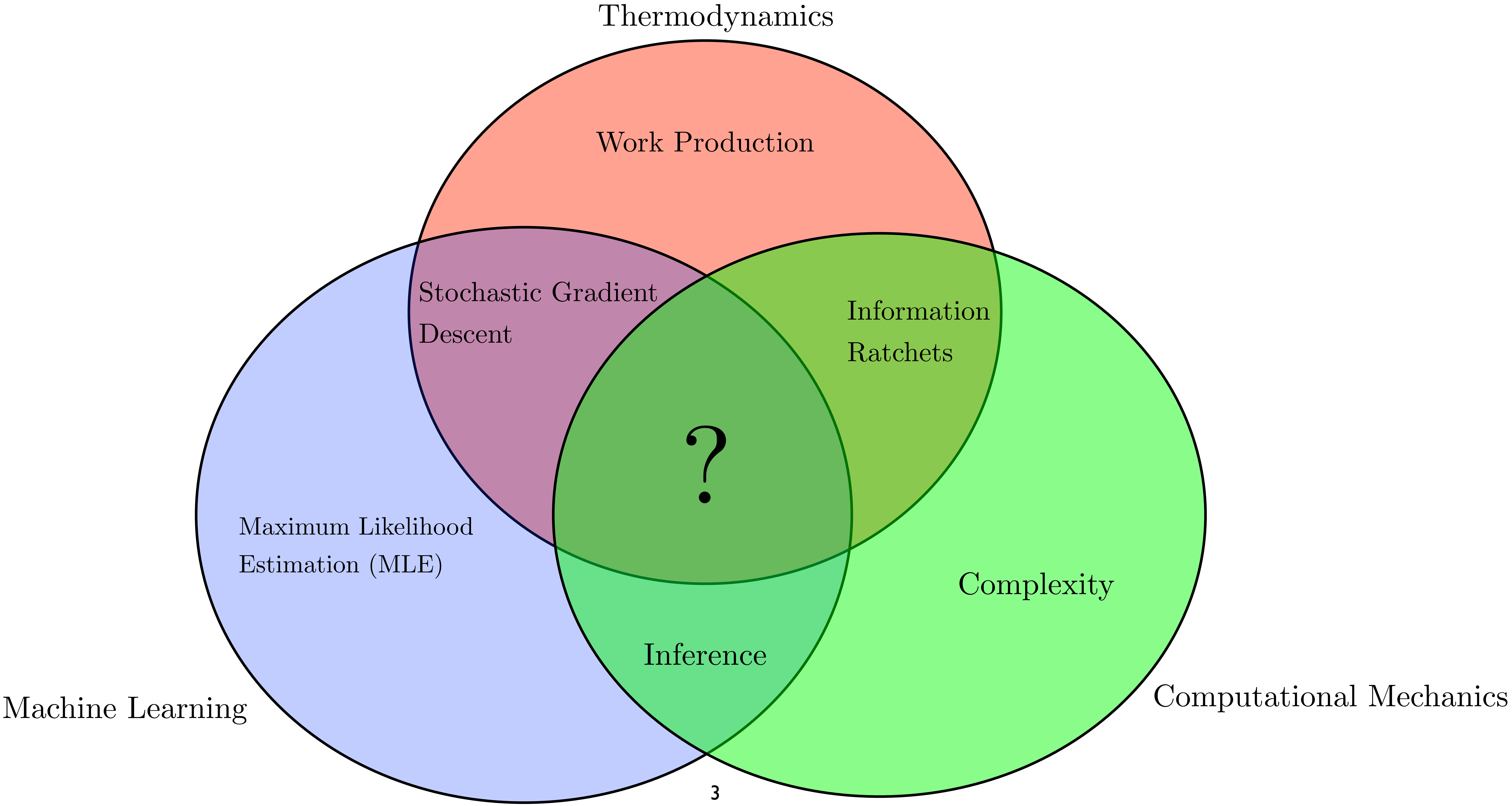


BITS
Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin
FQXi
FOUNDATIONAL QUESTIONS INSTITUTE
IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn

Big Picture


Thermodynamics
Machine Learning
Computational Mechanics

Big Picture


Thermodynamics
Work Production
Stochastic Gradient
Descent
Information
Ratchets
?
Maximum Likelihood
Estimation (MLE)
Complexity
Inference
Machine Learning
Computational Mechanics

Entropy and Information

Probability: $\Pr(X = x)$

In Physics

Gibbs Entropy: $S[X] \equiv -k_B \sum_x \Pr(X = x) \ln \Pr(X = x)$

$$= k_B \ln 2 H[X]$$

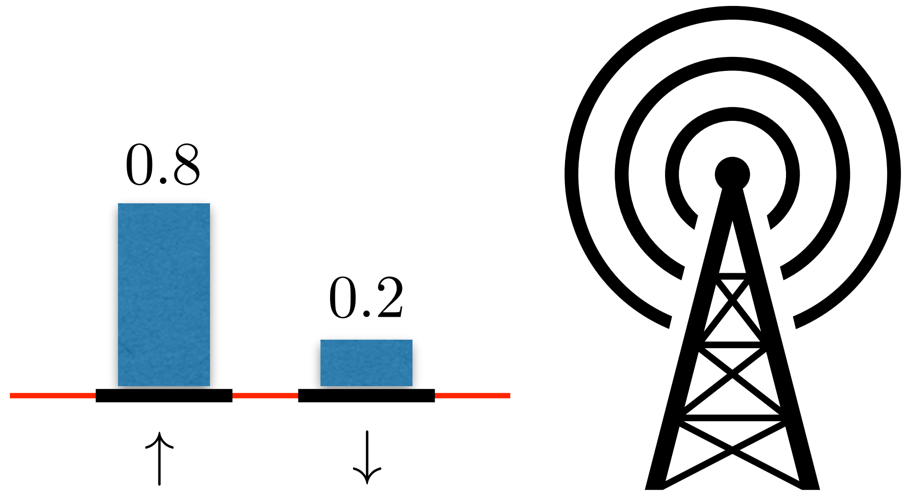
Determines change in energy of thermal reservoir

$$Q = T \Delta S_{\text{reservoir}}$$

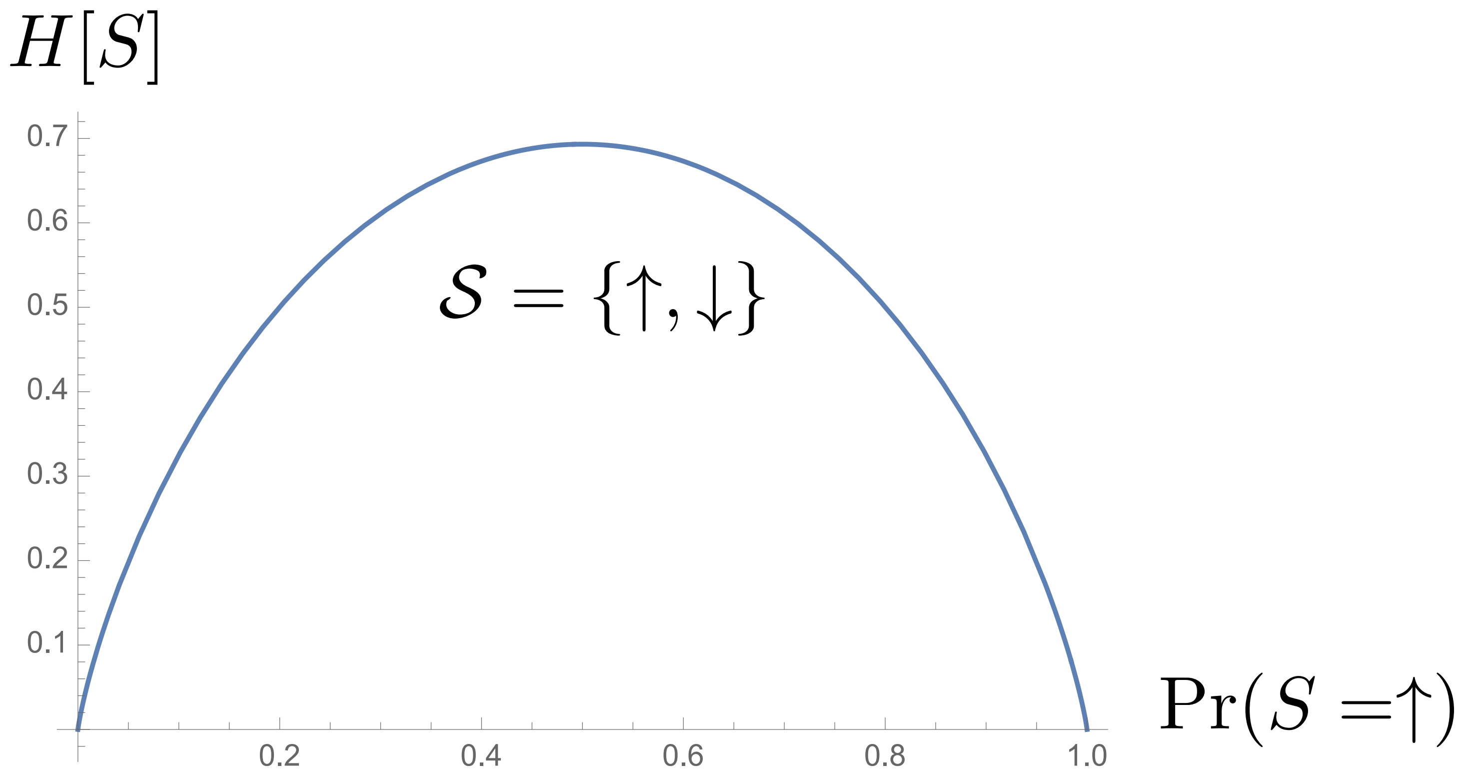
In Information Theory

Shannon Entropy: $H[X] \equiv -\sum_x \Pr(X = x) \log_2 \Pr(X = x)$

Determines average number of bits necessary to communicate distribution



$$\langle N_{\text{bits}} \rangle \geq H[X]$$



Entropy and Information

Probability: $\Pr(X = x)$

In Physics

Gibbs Entropy: $$S[X] \equiv -k_B \sum_x \Pr(X = x) \ln \Pr(X = x)$$
$$= k_B \ln 2 H[X]$$

Determines change in energy of thermal reservoir

$$Q = T \Delta S_{\text{reservoir}}$$

Express Shannon entropy in terms of Nats instead of Bits

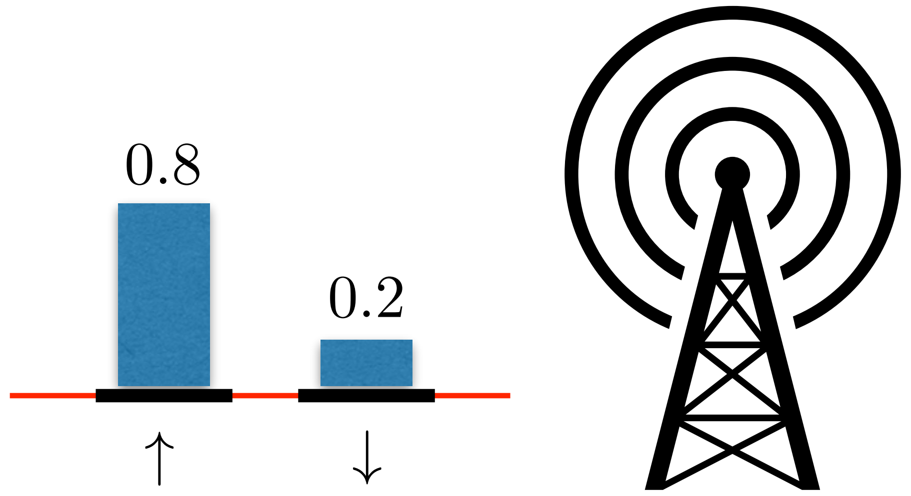
$$H[X] = -\sum_x \Pr(X = x) \ln \Pr(X = x)$$

$$S[X] = k_B H[X]$$

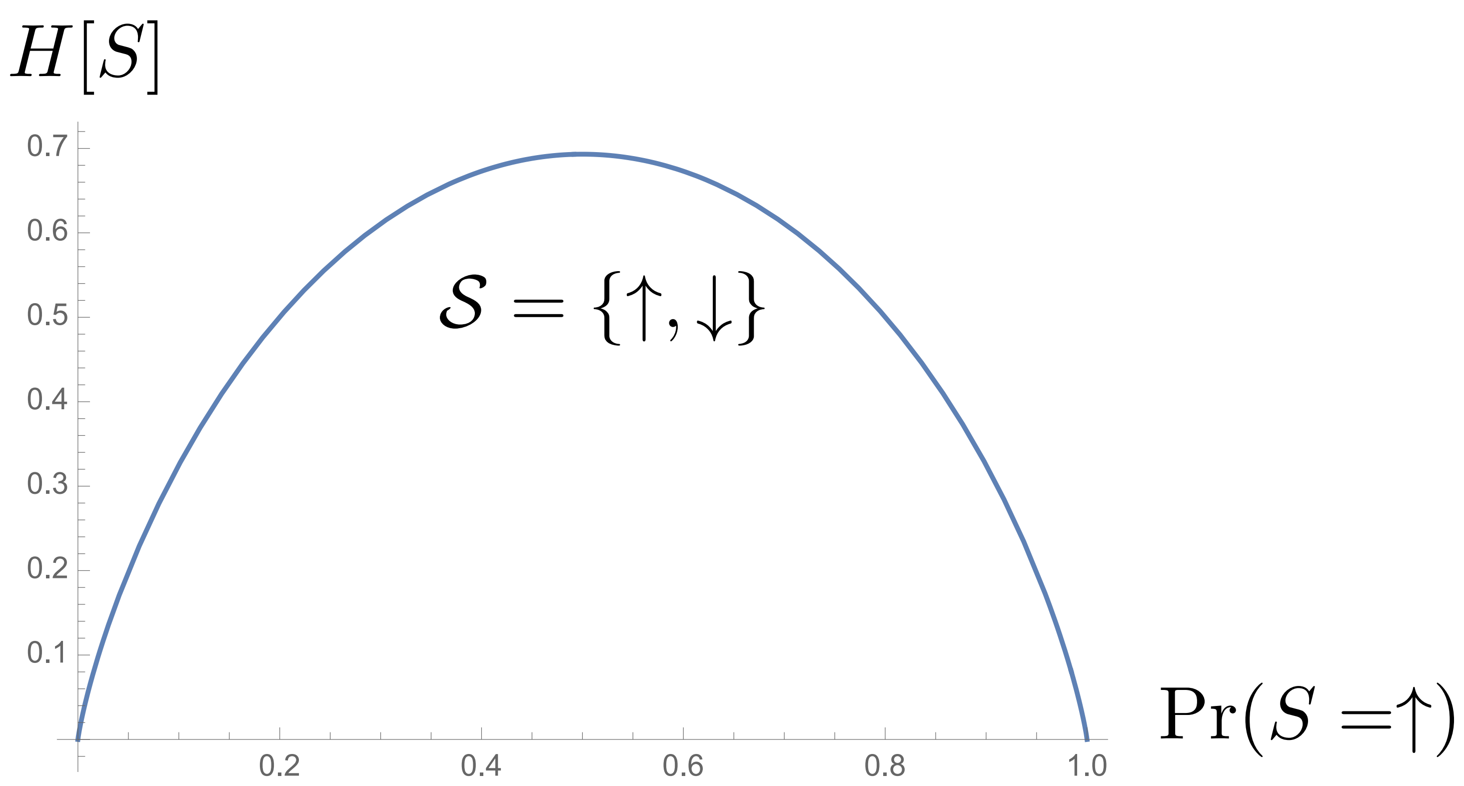
In Information Theory

Shannon Entropy: $$H[X] \equiv -\sum_x \Pr(X = x) \log_2 \Pr(X = x)$$

Determines average number of bits necessary to communicate distribution



$$\langle N_{\text{bits}} \rangle \geq H[X]$$



Thermodynamic (Informational) Principle of Organization

The Second Law of thermodynamics:

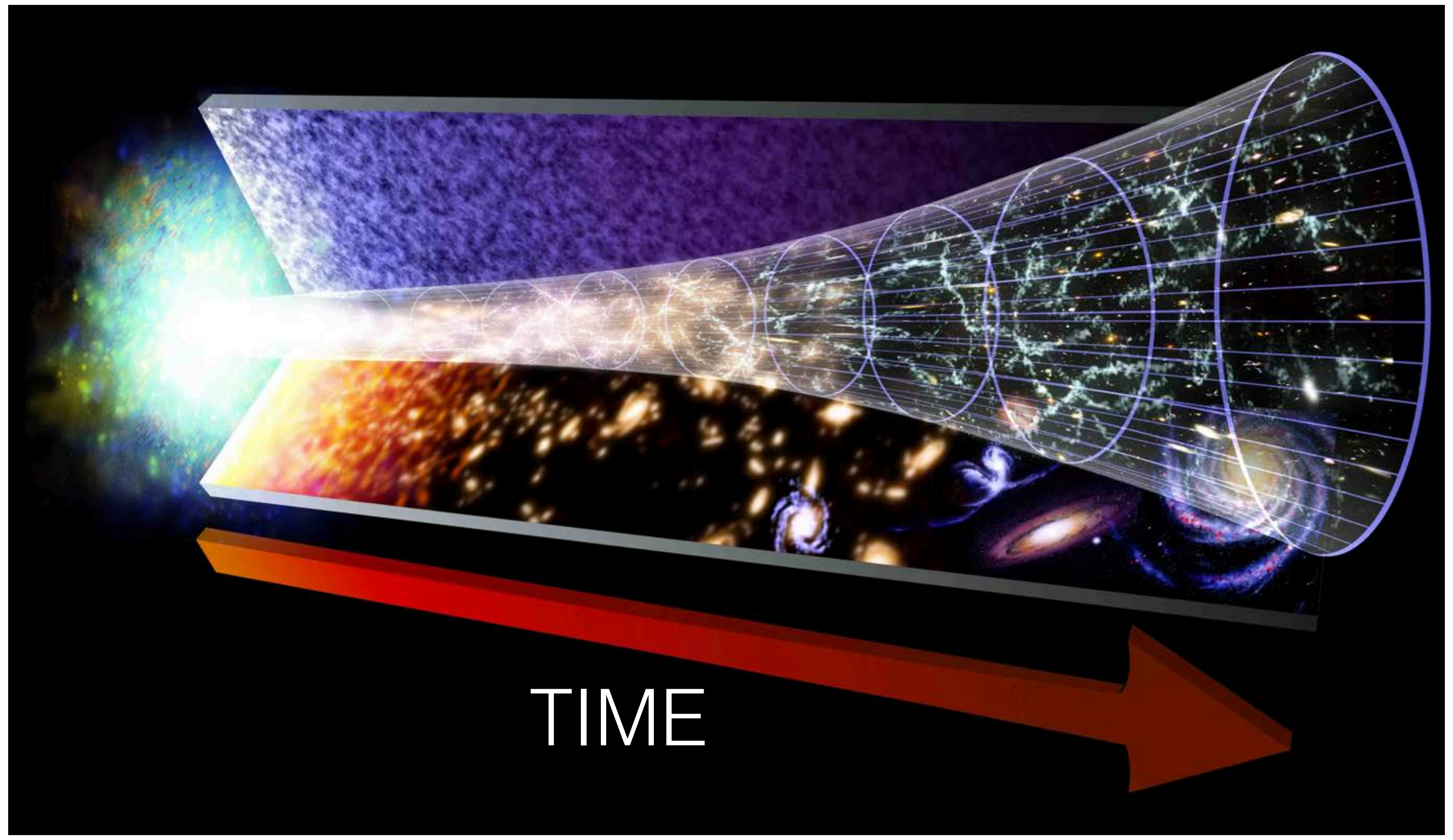
$$\Delta S_{\text{isolated system}} \geq 0$$

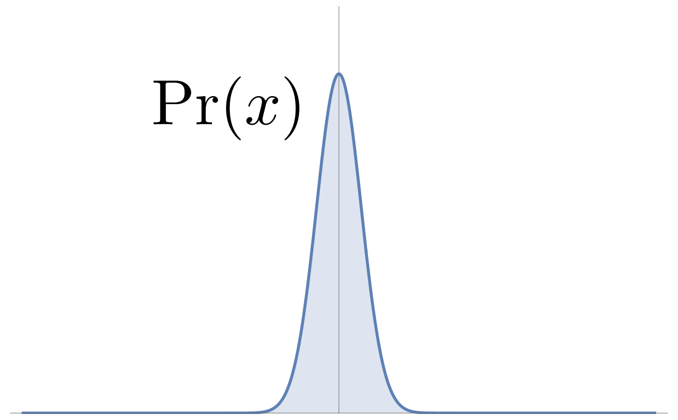
https://www.sciencefocus.com/science/how-high-must-you-sing-to-shatter-a-wine-glass/

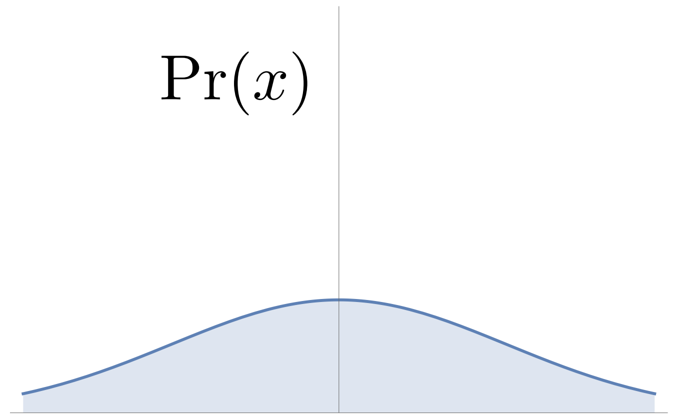
The universe is an isolated system

$$\Delta S_{\text{universe}} \geq 0$$

https://theconversation.com/how-could-the-big-bang-arise-from-nothing-171986



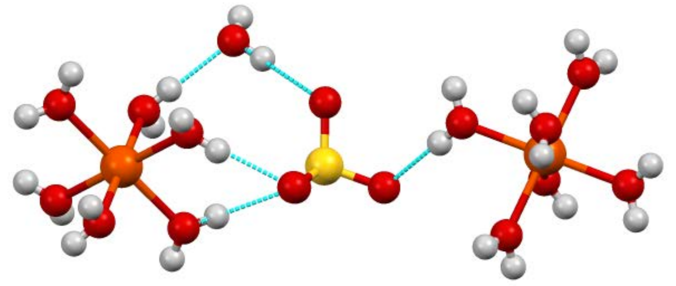




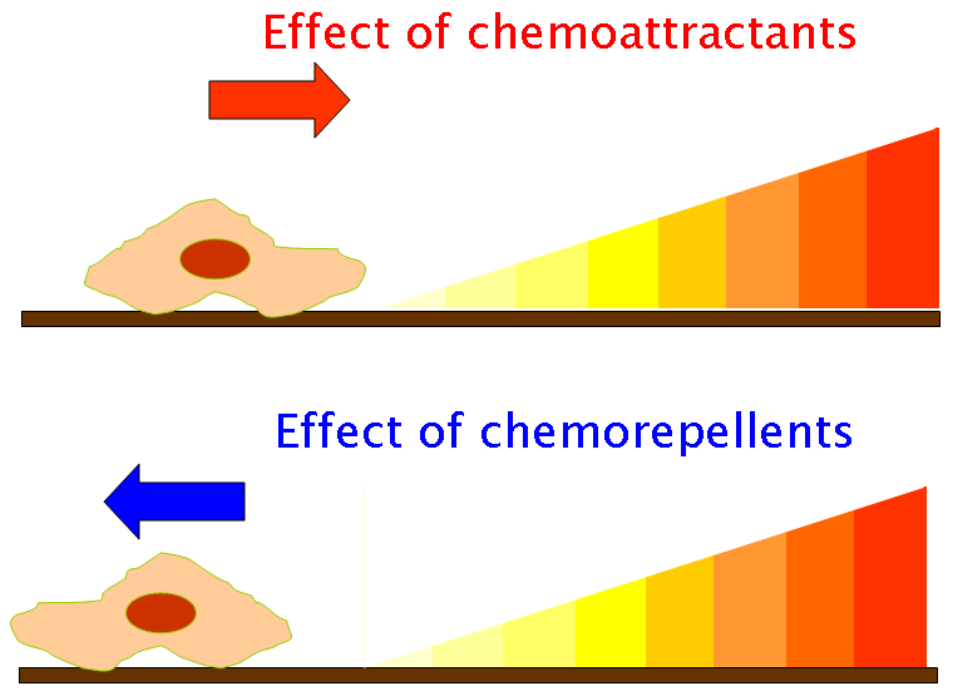
Motivating Questions

Does intelligence and complexity typically emerge in nonequilibrium systems?

Increasing Complexity and Intelligence



https://en.wikipedia.org/wiki/Properties_of_water



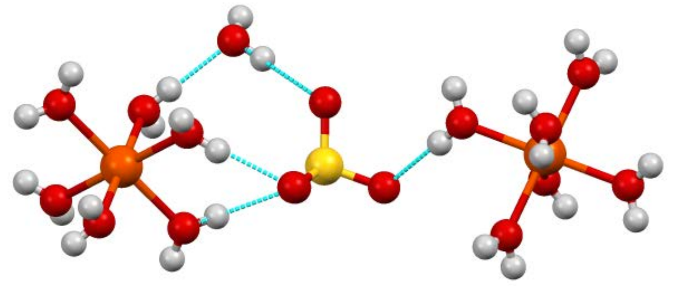
© Kohidai, L. 2008

https://en.wikipedia.org/wiki/Shanghai

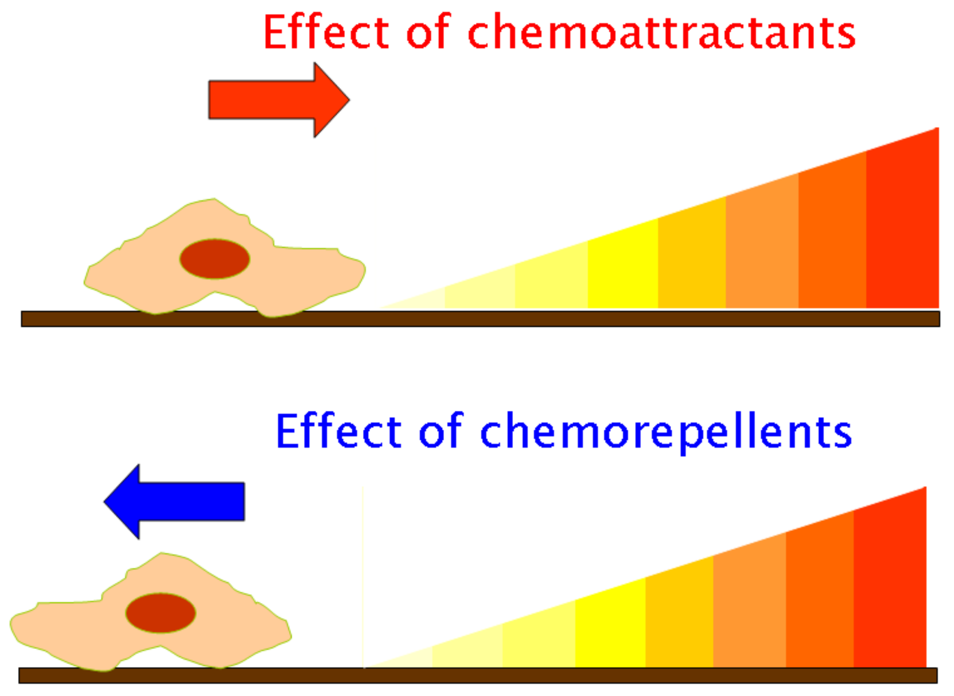
Motivating Questions

Does intelligence and complexity typically emerge in nonequilibrium systems?

Increasing Complexity and Intelligence



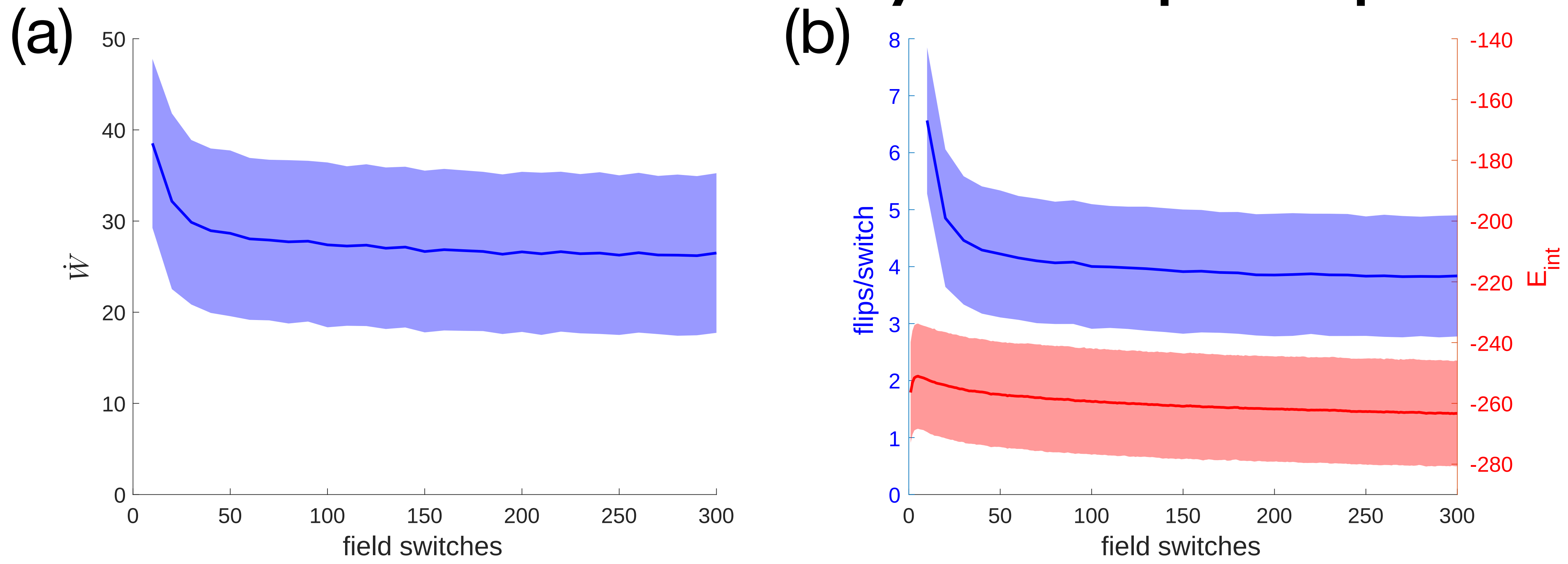
https://en.wikipedia.org/wiki/Properties_of_water



© Kohidai, L. 2008

https://en.wikipedia.org/wiki/Shanghai

Are there other thermodynamic principles of organization for nonequilibrium systems?



minimum work absorption

Self-organized novelty detection in driven spin glasses

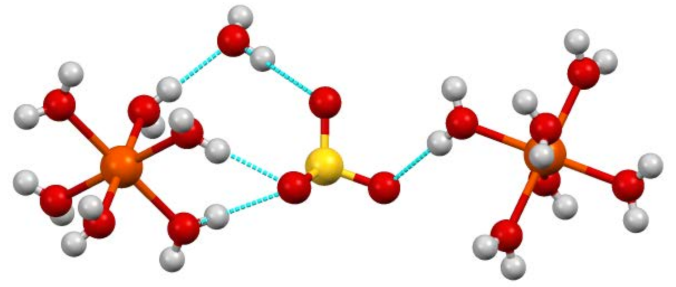
Jacob M. Gold *[1] and Jeremy L. England[2, 3]

arXiv:1911.07216v1 [nlin.AO] 17 Nov 2019

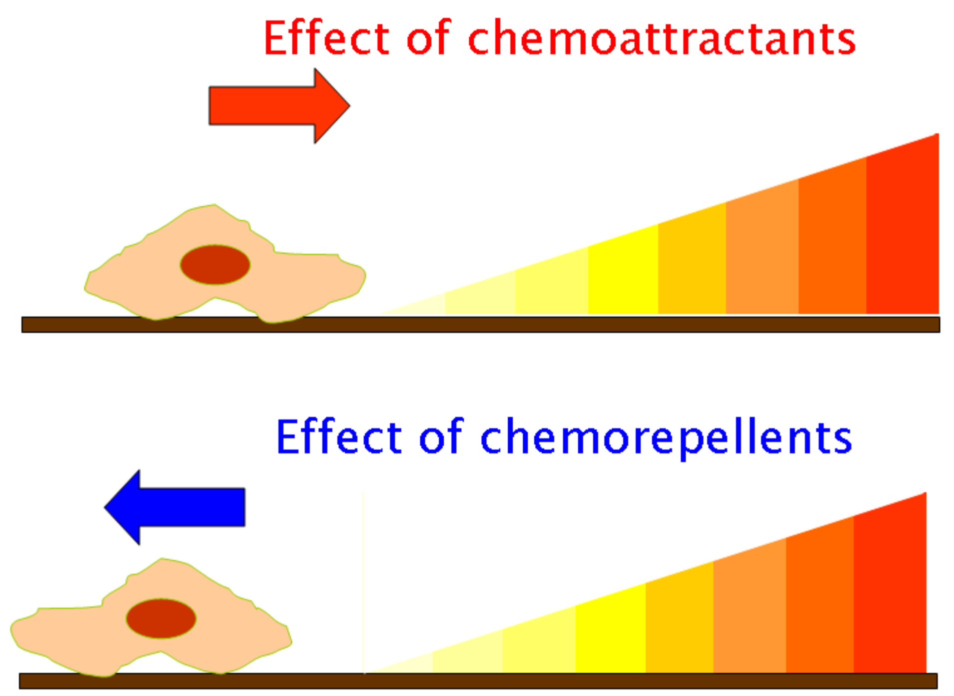
Motivating Questions

Does intelligence and complexity typically emerge in nonequilibrium systems?

Increasing Complexity and Intelligence



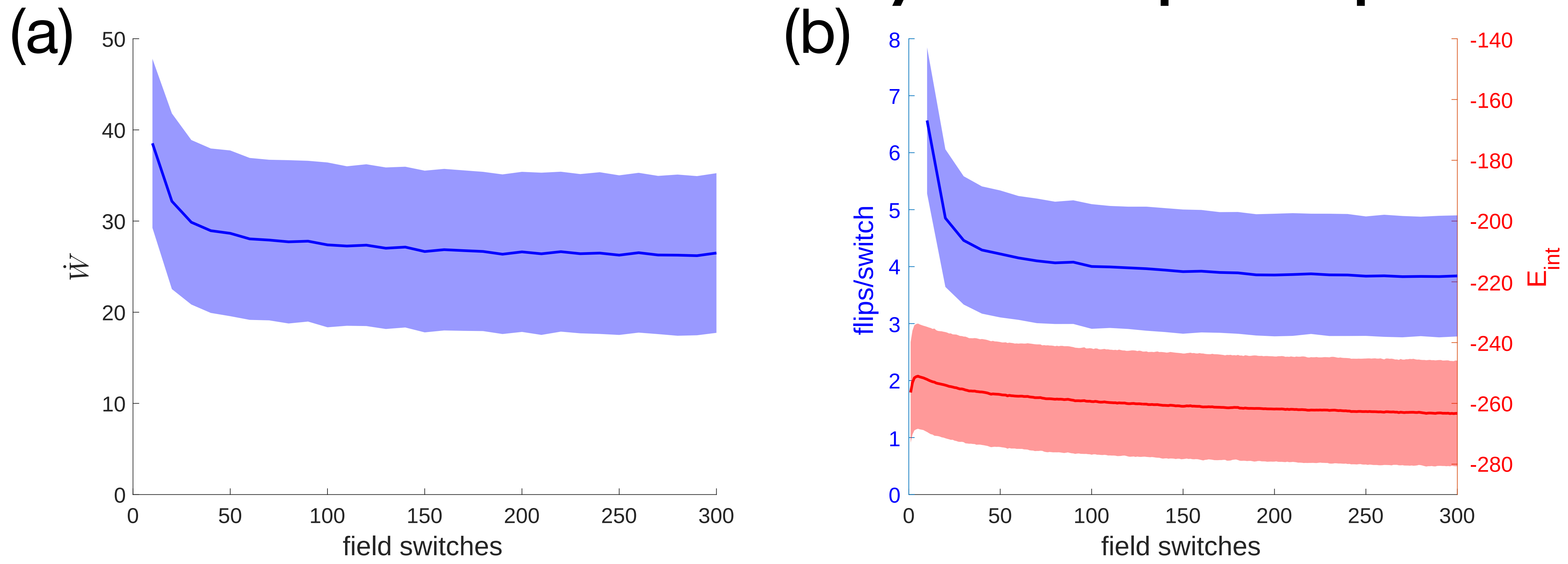
https://en.wikipedia.org/wiki/Properties_of_water



© Kohidai, L. 2008

https://en.wikipedia.org/wiki/Shanghai

Are there other thermodynamic principles of organization for nonequilibrium systems?



minimum work absorption

Self-organized novelty detection in driven spin glasses

Jacob M. Gold *[1] and Jeremy L. England[2, 3]

arXiv:1911.07216v1 [nlin.AO] 17 Nov 2019

- -First address "Complexity"
- -Then thermodynamics
- -Then learning

Complexity and Computational Mechanics

A pattern is a process

Random Variables: $Y_{a:b} = Y_a Y_{a+1} \cdots Y_{b-1}$

Probability of a realization $y_{a:b}$: $\Pr(Y_{a:b} = y_{a:b})$

Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

Complexity and Computational Mechanics

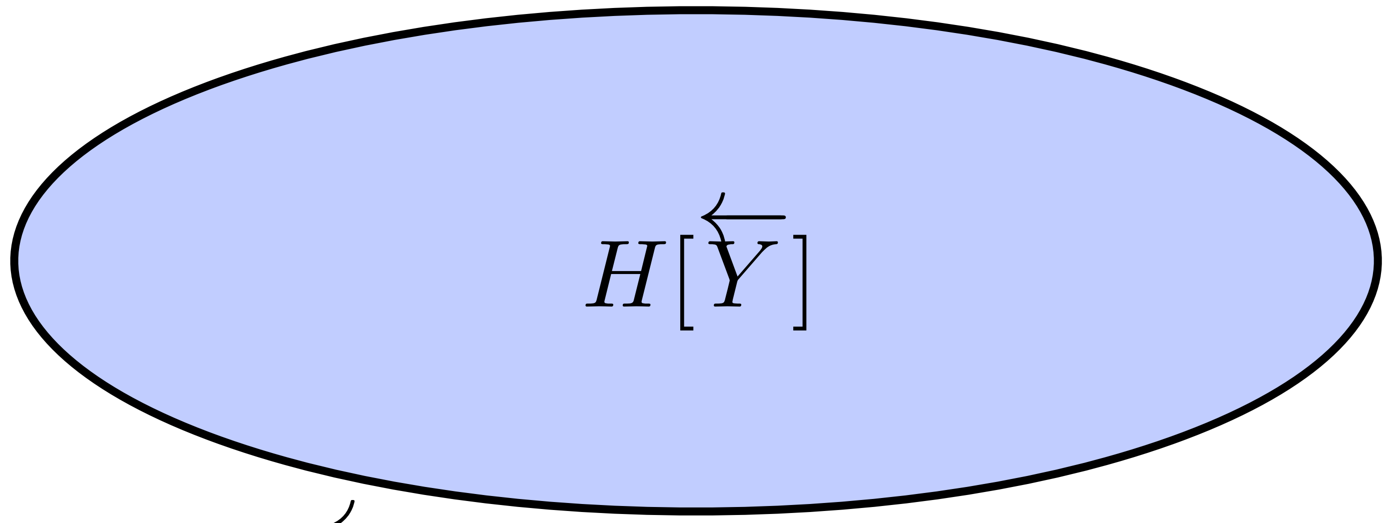
A pattern is a process

Random Variables: $Y_{a:b} = Y_a Y_{a+1} \cdots Y_{b-1}$

Probability of a realization $y_{a:b}$: $\Pr(Y_{a:b} = y_{a:b})$

Entropy of a process

$$H[Y_{a:b}] \equiv -\sum_{y_{a:b}} \Pr(Y_{a:b} = y_{a:b}) \ln \Pr(Y_{a:b} = y_{a:b})$$



$$\overleftarrow{Y} = Y_{-\infty:0} = \cdots Y_{-2} Y_{-1}$$

Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

Complexity and Computational Mechanics

A pattern is a process

Random Variables: $Y_{a:b} = Y_a Y_{a+1} \cdots Y_{b-1}$

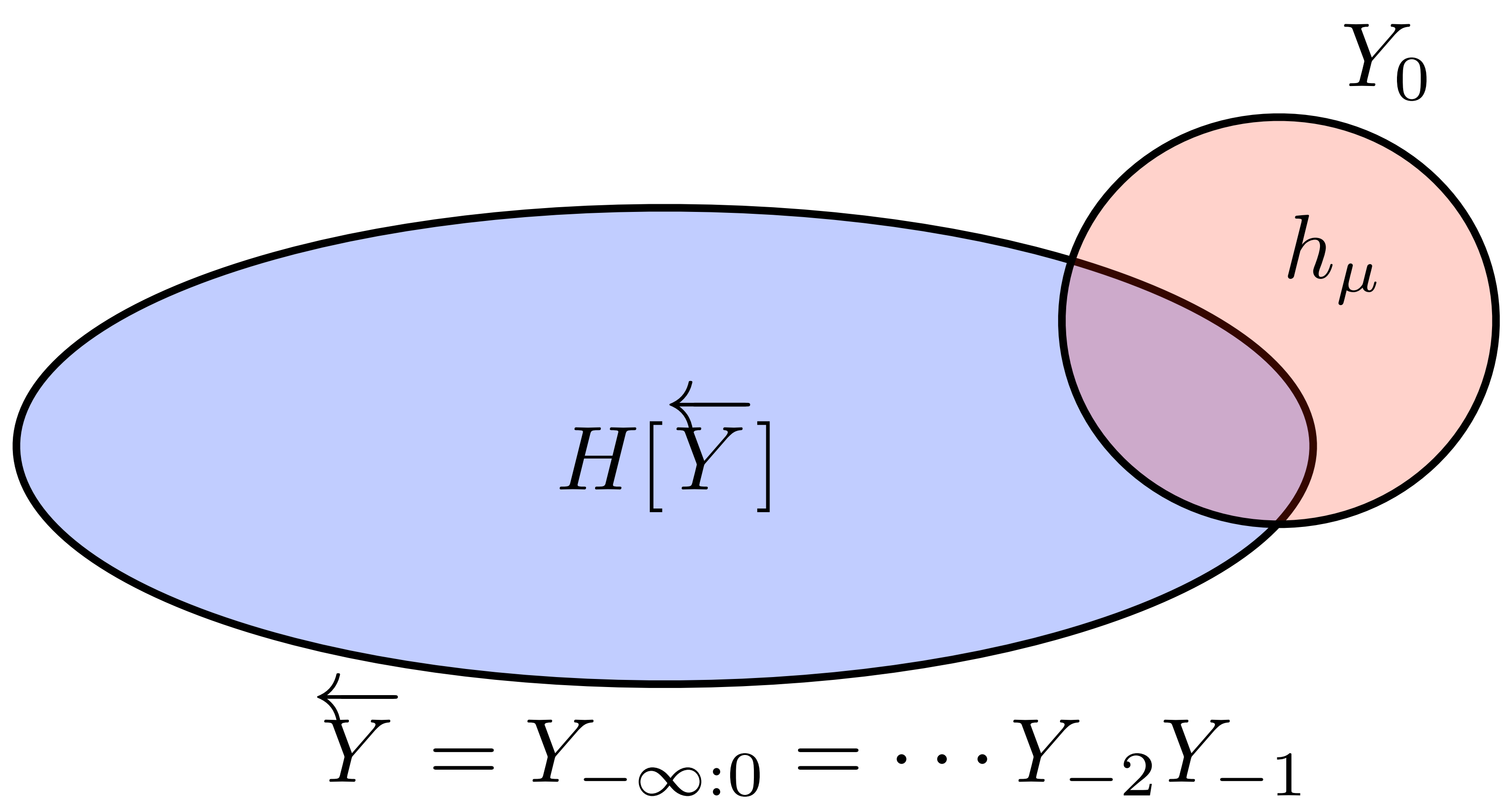
Probability of a realization $y_{a:b}$: $\Pr(Y_{a:b} = y_{a:b})$

Entropy of a process

$$H[Y_{a:b}] \equiv -\sum_{y_{a:b}} \Pr(Y_{a:b} = y_{a:b}) \ln \Pr(Y_{a:b} = y_{a:b})$$

Entropy rate of a process

$$h_\mu \equiv \lim_{L \to \infty} \frac{H[Y_{0:L}]}{L} = H[Y_0 | \overleftarrow{Y}_0]$$



Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

Complexity and Computational Mechanics

A pattern is a process

Random Variables: $Y_{a:b} = Y_a Y_{a+1} \cdots Y_{b-1}$

Probability of a realization $y_{a:b}$: $\Pr(Y_{a:b} = y_{a:b})$

Entropy of a process

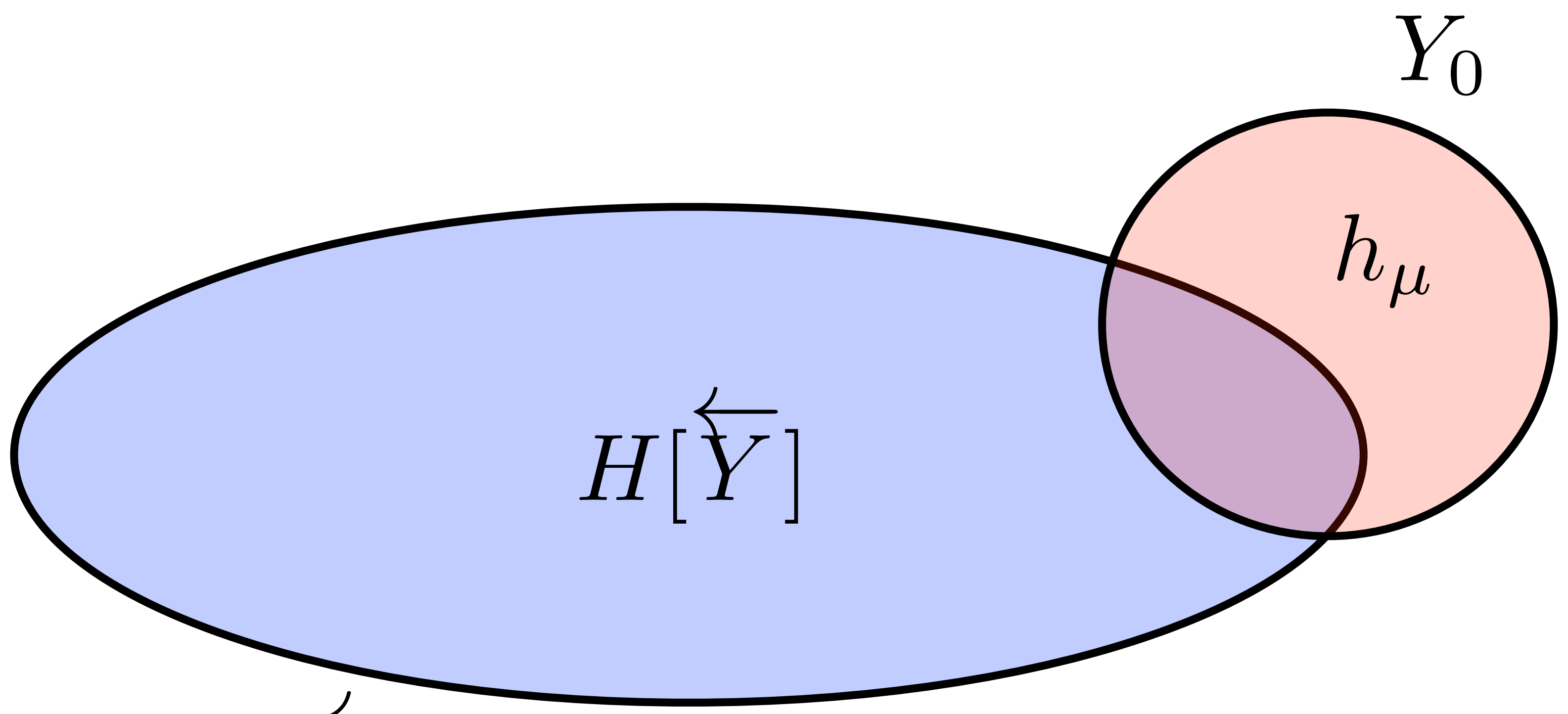
$$H[Y_{a:b}] \equiv -\sum_{y_{a:b}} \Pr(Y_{a:b} = y_{a:b}) \ln \Pr(Y_{a:b} = y_{a:b})$$

Entropy rate of a process

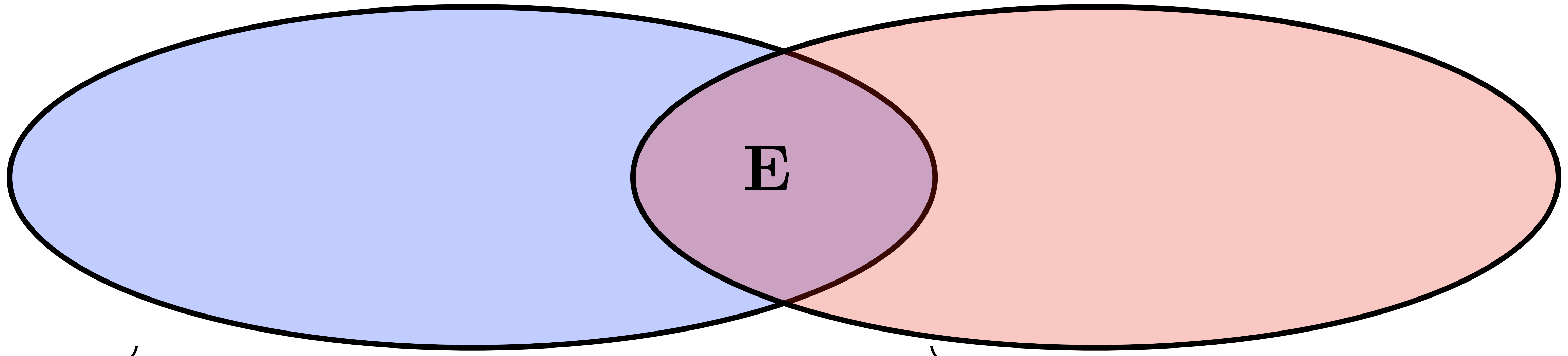
$$h_\mu \equiv \lim_{L\to\infty} \frac{H[Y_{0:L}]}{L} = H[Y_0|\overleftarrow{Y}_0]$$

Excess Entropy

$$\mathbf{E} \equiv H[\overleftarrow{Y}_0] + H[\overrightarrow{Y}_0] - H[\overleftarrow{Y}_0, \overrightarrow{Y}_0] \equiv I[\overleftarrow{Y}_0; \overrightarrow{Y}_0]$$



$$\overleftarrow{Y} = Y_{-\infty:0} = \cdots Y_{-2} Y_{-1}$$



$$\overleftarrow{Y} = Y_{-\infty:0} = \cdots Y_{-2} Y_{-1} \quad \overrightarrow{Y} = Y_{0:\infty} = Y_0 Y_1 \cdots$$

Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

Complexity and Computational Mechanics

A pattern is a process

Random Variables: $Y_{a:b} = Y_a Y_{a+1} \cdots Y_{b-1}$

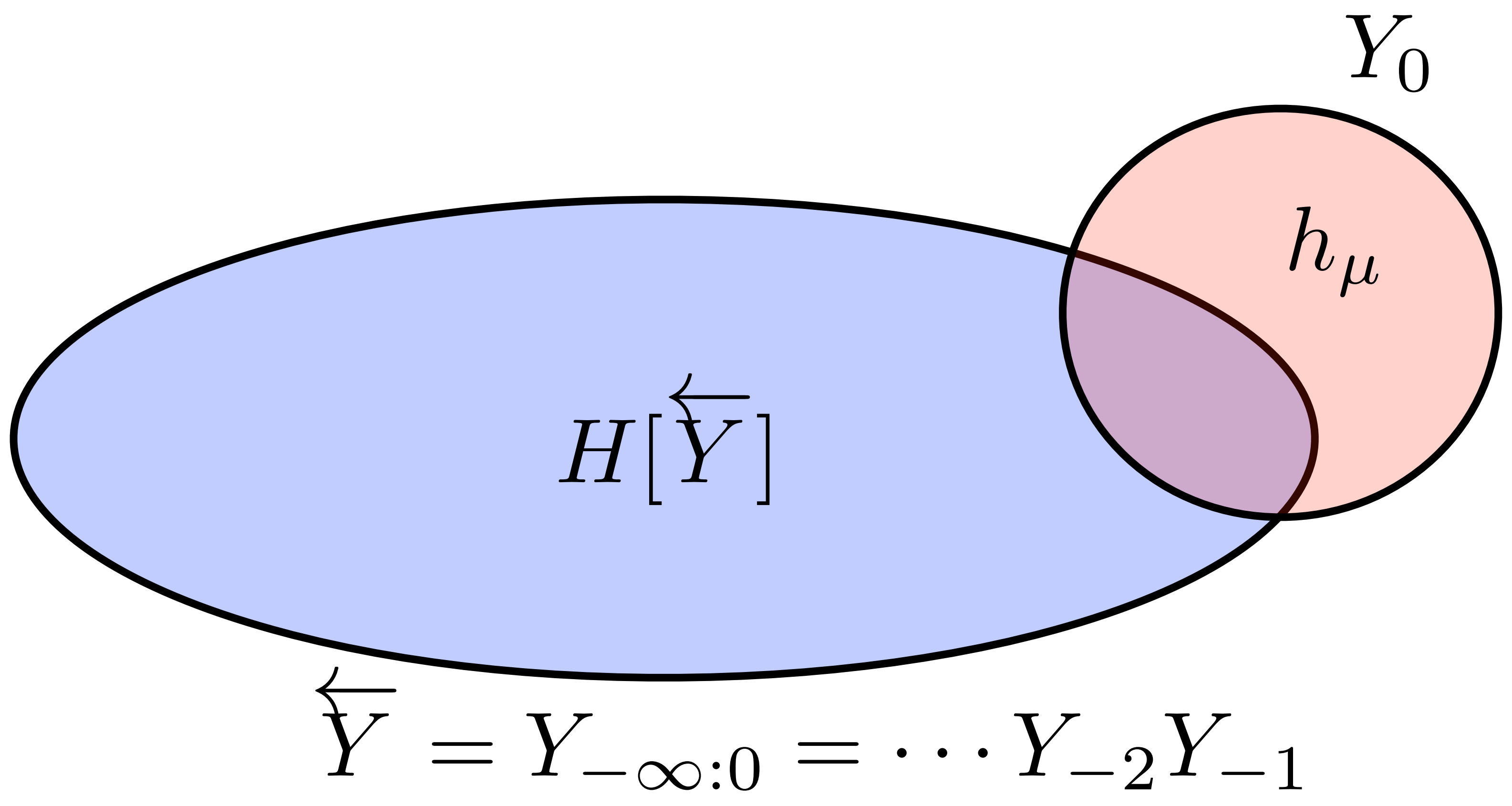
Probability of a realization $y_{a:b}$: $\Pr(Y_{a:b} = y_{a:b})$

Entropy of a process

$$H[Y_{a:b}] \equiv -\sum_{y_{a:b}} \Pr(Y_{a:b} = y_{a:b}) \ln \Pr(Y_{a:b} = y_{a:b})$$

Entropy rate of a process

$$h_\mu \equiv \lim_{L\to\infty} \frac{H[Y_{0:L}]}{L} = H[Y_0|\overleftarrow{Y}_0]$$

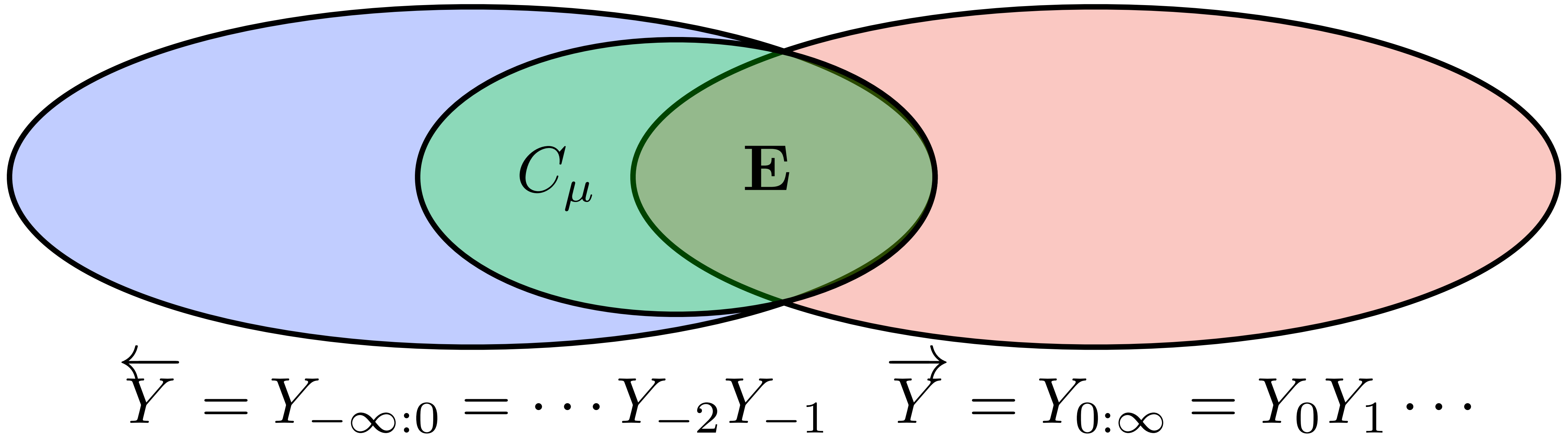


Excess Entropy

$$\mathbf{E} \equiv H[\overleftarrow{Y}_0] + H[\overrightarrow{Y}_0] - H[\overleftarrow{Y}_0, \overrightarrow{Y}_0] \equiv I[\overleftarrow{Y}_0; \overrightarrow{Y}_0]$$

Statistical Complexity

$$C_\mu = \min_{\epsilon \ni I[\epsilon(\overleftarrow{Y});\overrightarrow{Y}] = I[\overleftarrow{Y};\overrightarrow{Y}]} H[\epsilon(\overleftarrow{Y})]$$



Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

Complexity and Computational Mechanics

A pattern is a process

Random Variables: $Y_{a:b} = Y_a Y_{a+1} \cdots Y_{b-1}$

Probability of a realization $y_{a:b}$: $\Pr(Y_{a:b} = y_{a:b})$

Entropy of a process

$$H[Y_{a:b}] \equiv - \sum_{y_{a:b}} \Pr(Y_{a:b} = y_{a:b}) \ln \Pr(Y_{a:b} = y_{a:b})$$

Entropy rate of a process

$$h_\mu \equiv \lim_{L \to \infty} \frac{H[Y_{0:L}]}{L} = H[Y_0 | \overleftarrow{Y}_0]$$

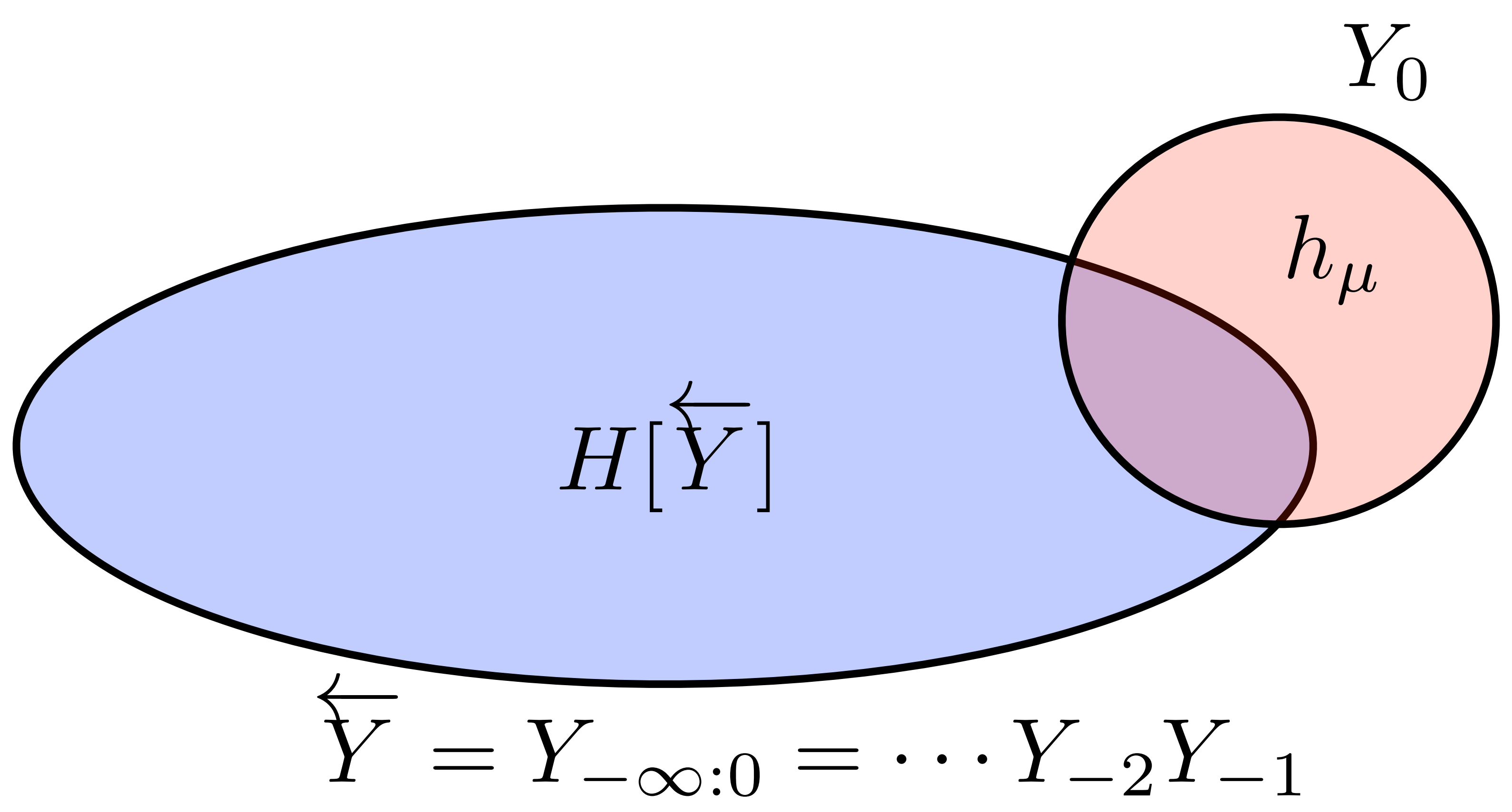
Excess Entropy

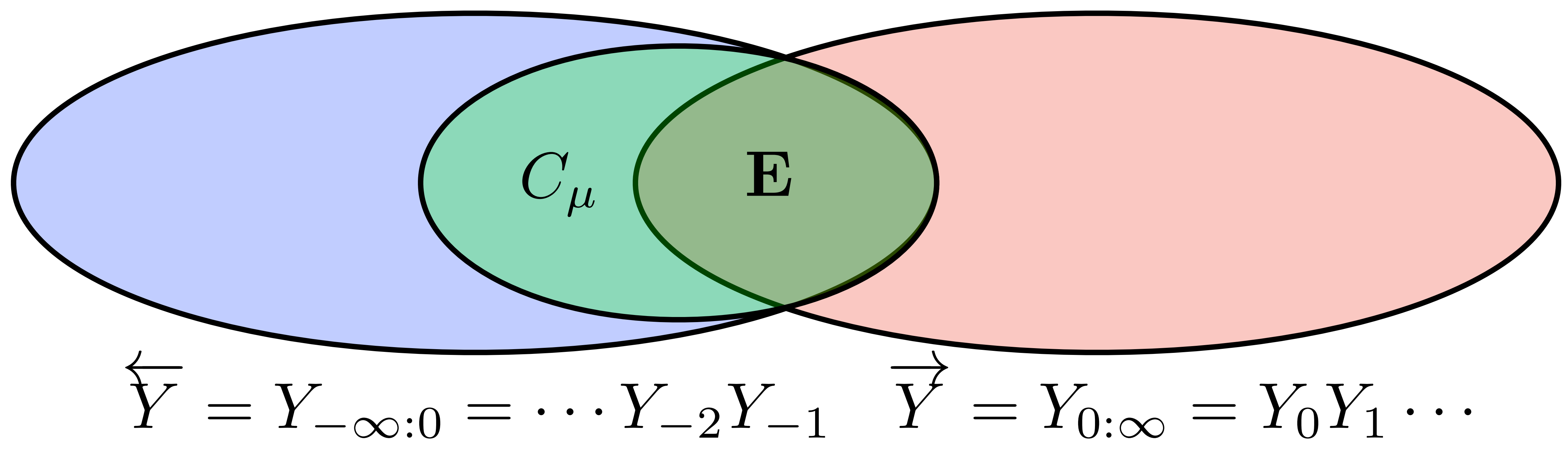
$$\mathbf{E} \equiv H[\overleftarrow{Y}_0] + H[\overrightarrow{Y}_0] - H[\overleftarrow{Y}_0, \overrightarrow{Y}_0] \equiv I[\overleftarrow{Y}_0 ; \overrightarrow{Y}_0]$$

Statistical Complexity

$$C_\mu = \min_{\epsilon \ni I[\epsilon(\overleftarrow{Y}); \overrightarrow{Y}] = I[\overleftarrow{Y}; \overrightarrow{Y}]} H[\epsilon(\overleftarrow{Y})]$$

Minimum sufficient statistic of the past about the future





Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

Complexity and Computational Mechanics

A pattern is a process

Random Variables: $Y_{a:b} = Y_a Y_{a+1} \cdots Y_{b-1}$

Probability of a realization $y_{a:b}$: $\Pr(Y_{a:b} = y_{a:b})$

Entropy of a process

$$H[Y_{a:b}] \equiv - \sum_{y_{a:b}} \Pr(Y_{a:b} = y_{a:b}) \ln \Pr(Y_{a:b} = y_{a:b})$$

Entropy rate of a process

$$h_\mu \equiv \lim_{L \to \infty} \frac{H[Y_{0:L}]}{L} = H[Y_0 | \overleftarrow{Y}_0]$$

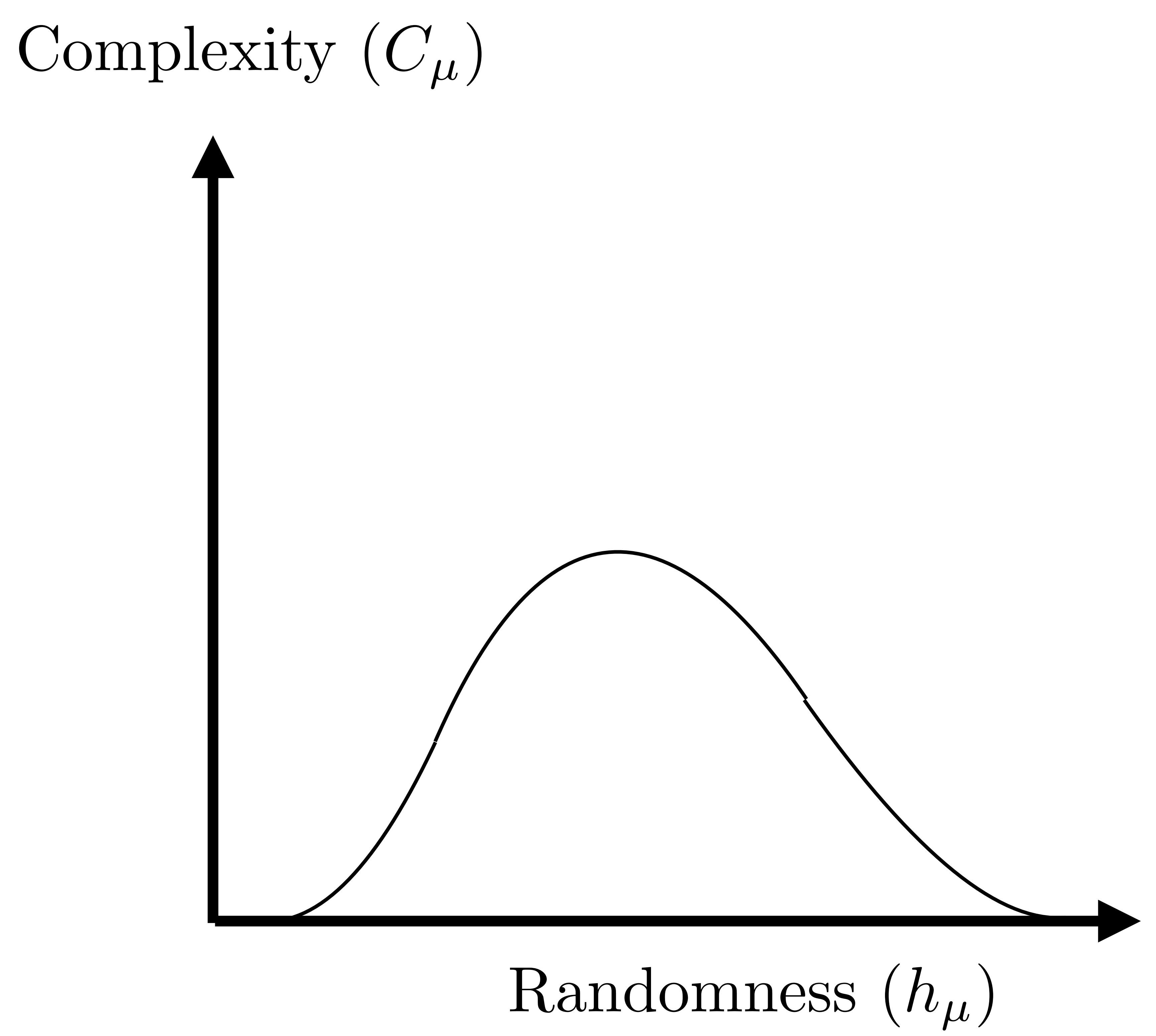
Excess Entropy

$$\mathbf{E} \equiv H[\overleftarrow{Y}_0] + H[\overrightarrow{Y}_0] - H[\overleftarrow{Y}_0, \overrightarrow{Y}_0] \equiv I[\overleftarrow{Y}_0; \overrightarrow{Y}_0]$$

Statistical Complexity

$$C_\mu = \min_{\epsilon \ni I[\epsilon(\overleftarrow{Y}); \overrightarrow{Y}] = I[\overleftarrow{Y}; \overrightarrow{Y}]} H[\epsilon(\overleftarrow{Y})]$$

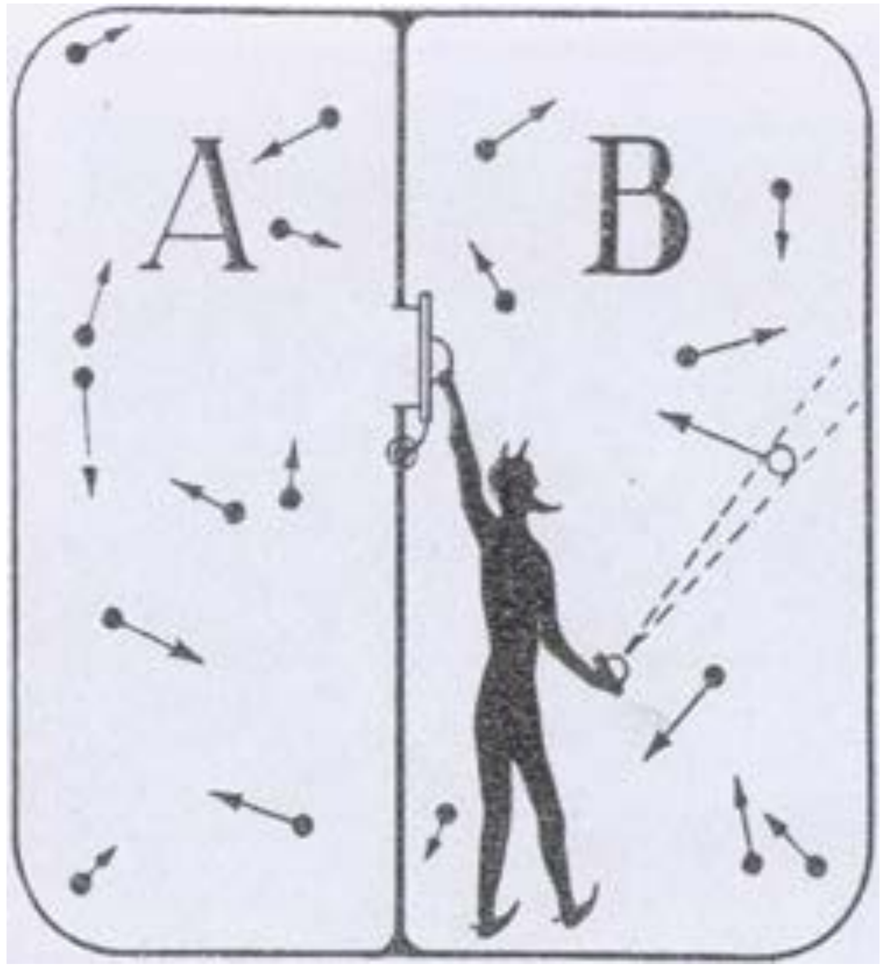
Minimum sufficient statistic of the past about the future



Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

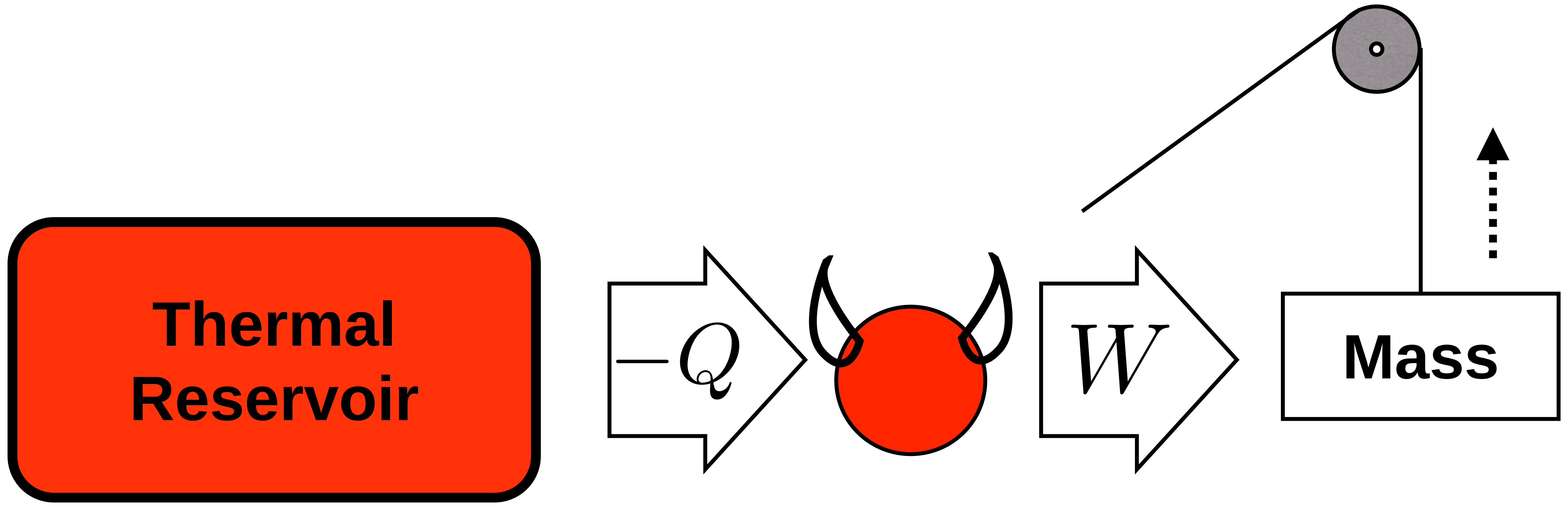
Harvesting Energy From Information

Maxwell's Demon (1867)



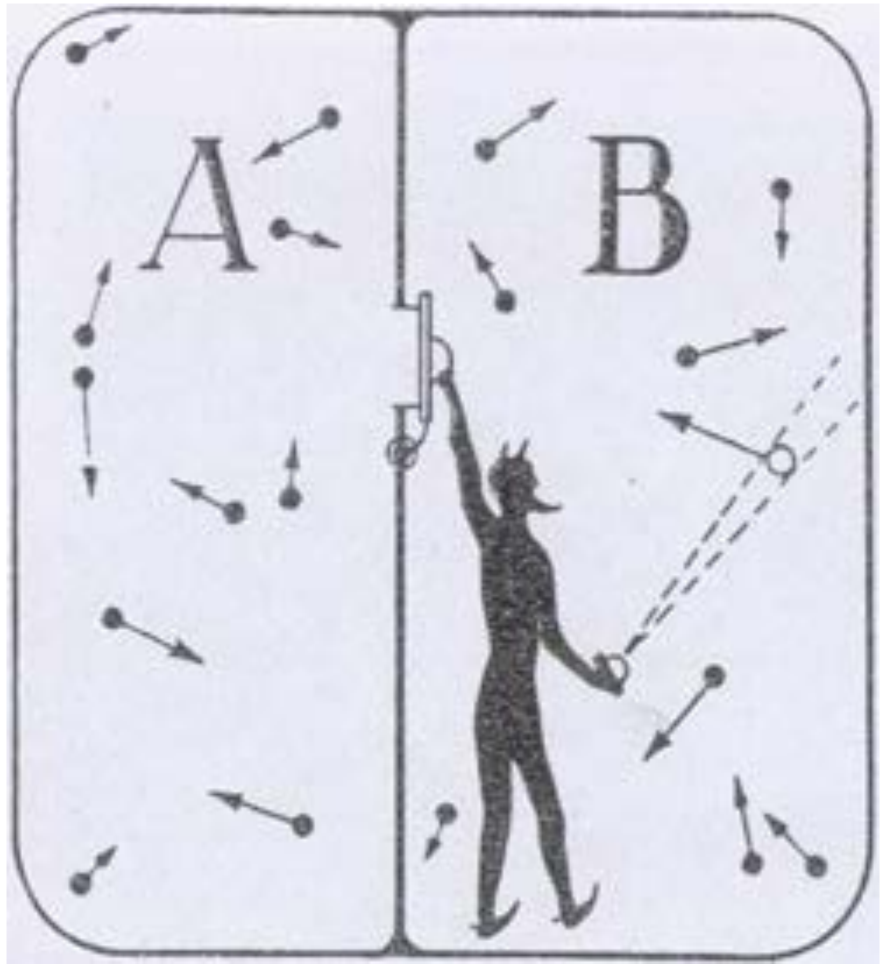
http://www.eoht.info/page/Maxwell's+demon

Feedback/Control



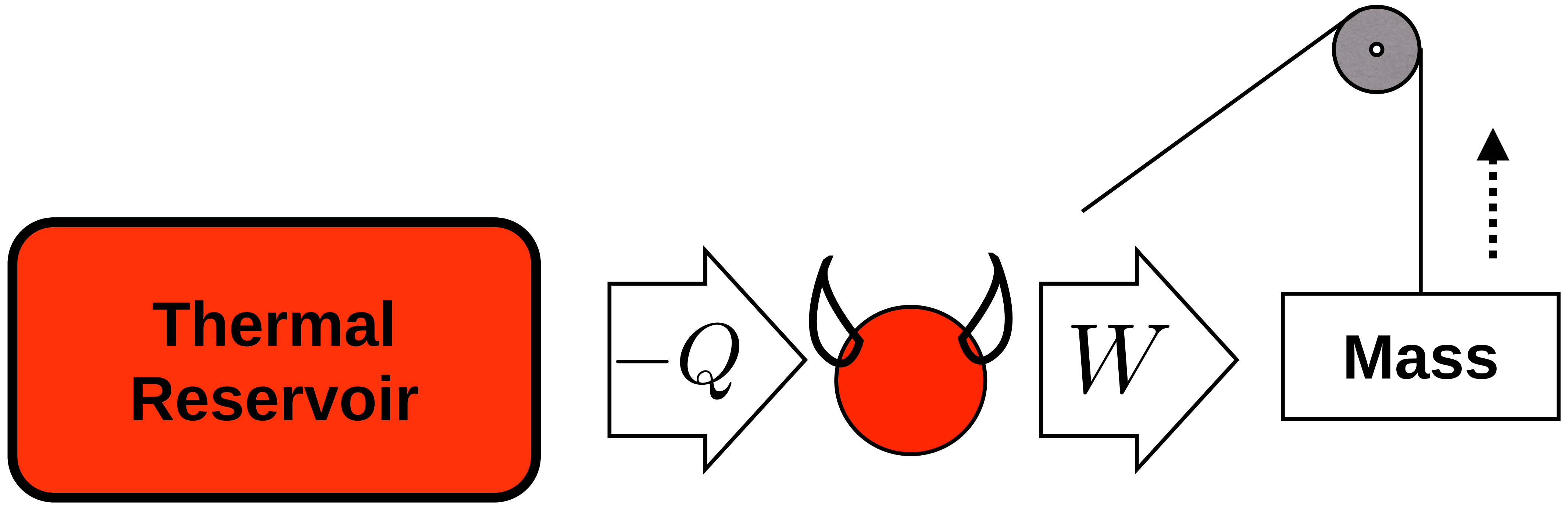
Harvesting Energy From Information

Maxwell's Demon (1867)



http://www.eoht.info/page/Maxwell's+demon

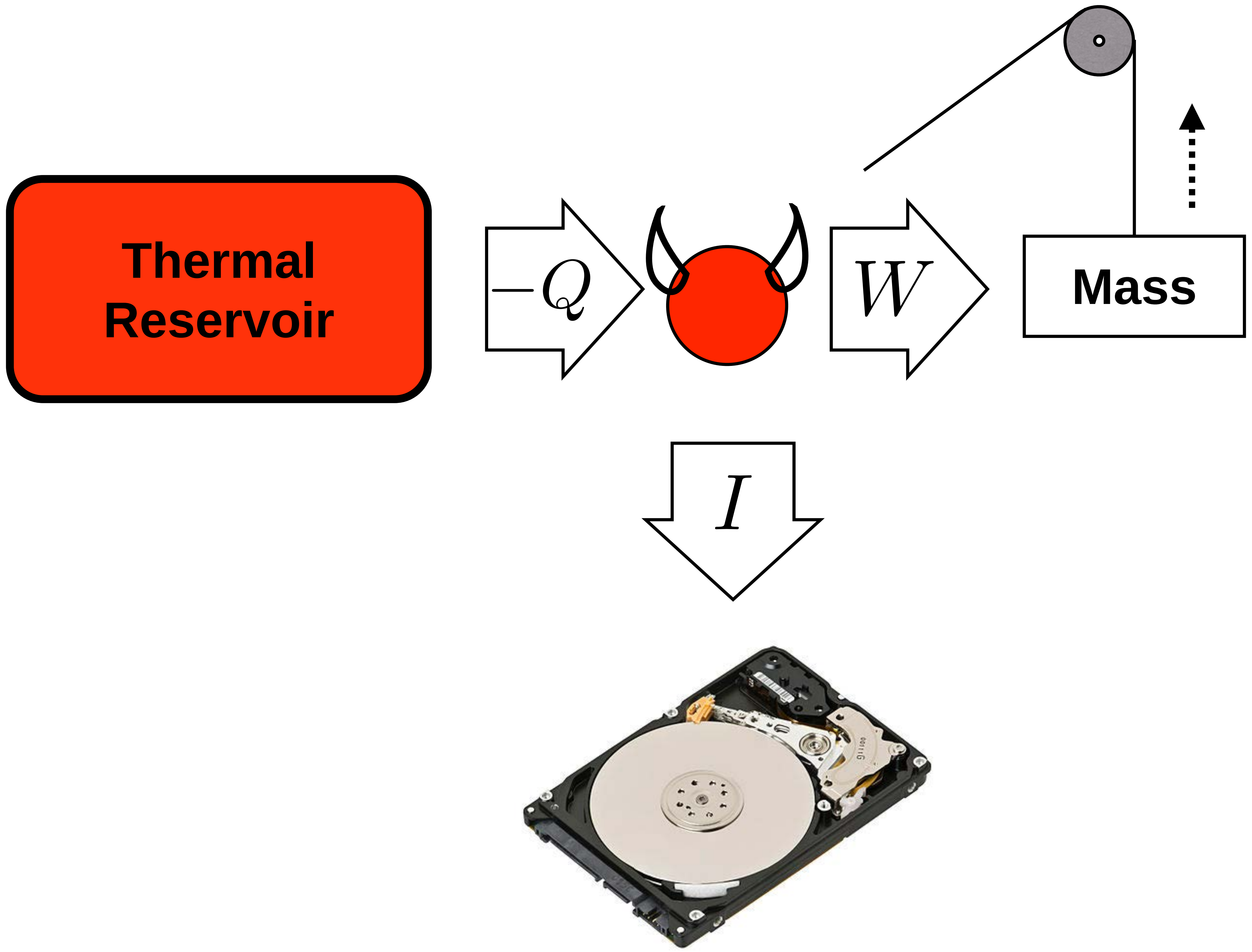
Feedback/Control



Impossible because of Landauer's Bound: $\langle W_{\text{erase}} \rangle \leq -k_B T \ln 2$

https://tinyurl.com/ErasingSim

Information is a Thermodynamic Fuel

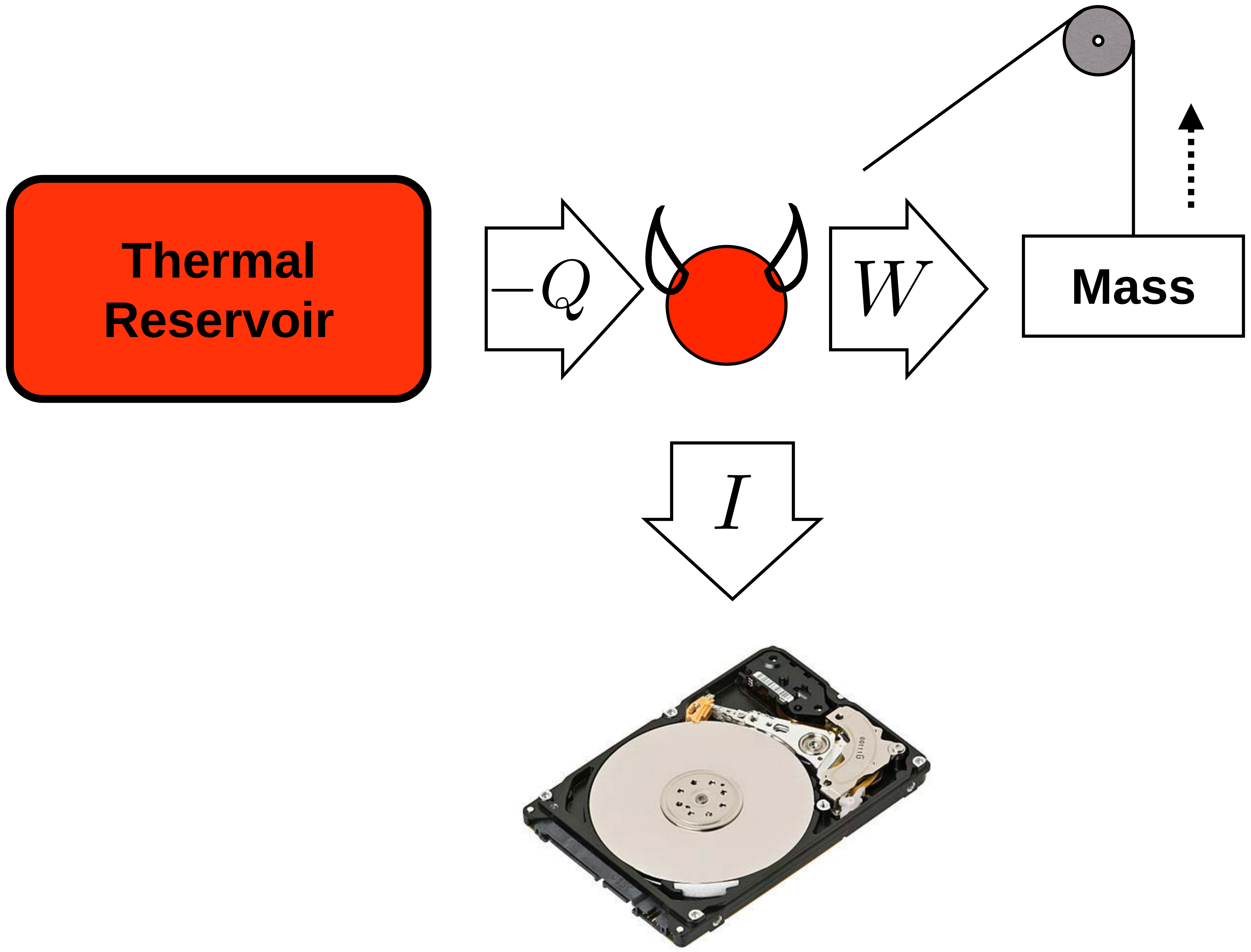


Instead of erasing: write to a hard drive

Second Law: $\Delta S^{\text{total}} = \langle Q \rangle / T + \Delta S_{\text{hard-drive}} \geq 0$

Energy Conservation: $\langle W \rangle = -\langle Q \rangle \leq T \Delta S_{\text{hard-drive}} = k_B T \Delta H_{\text{hard-drive}}$

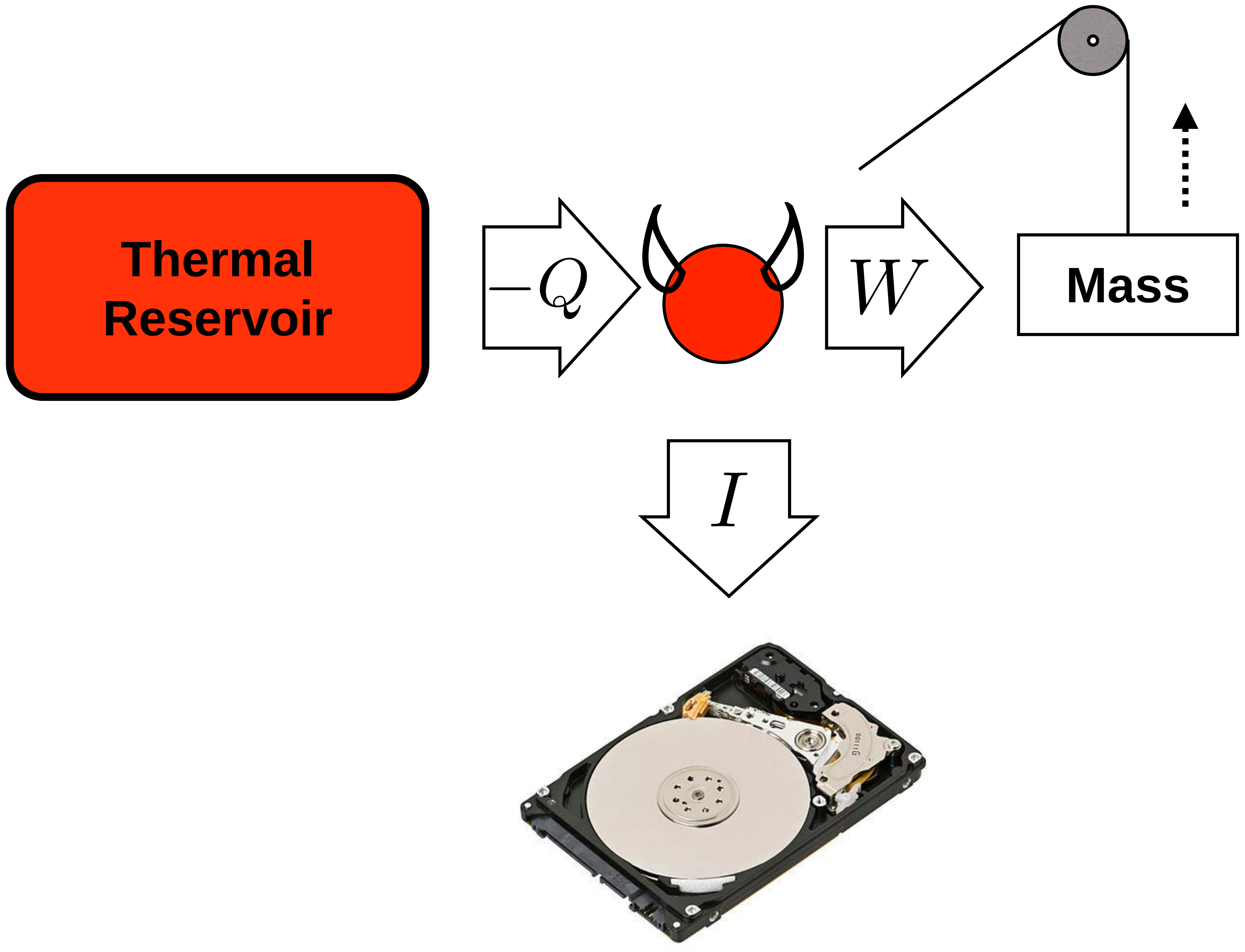
Information is a Thermodynamic Fuel

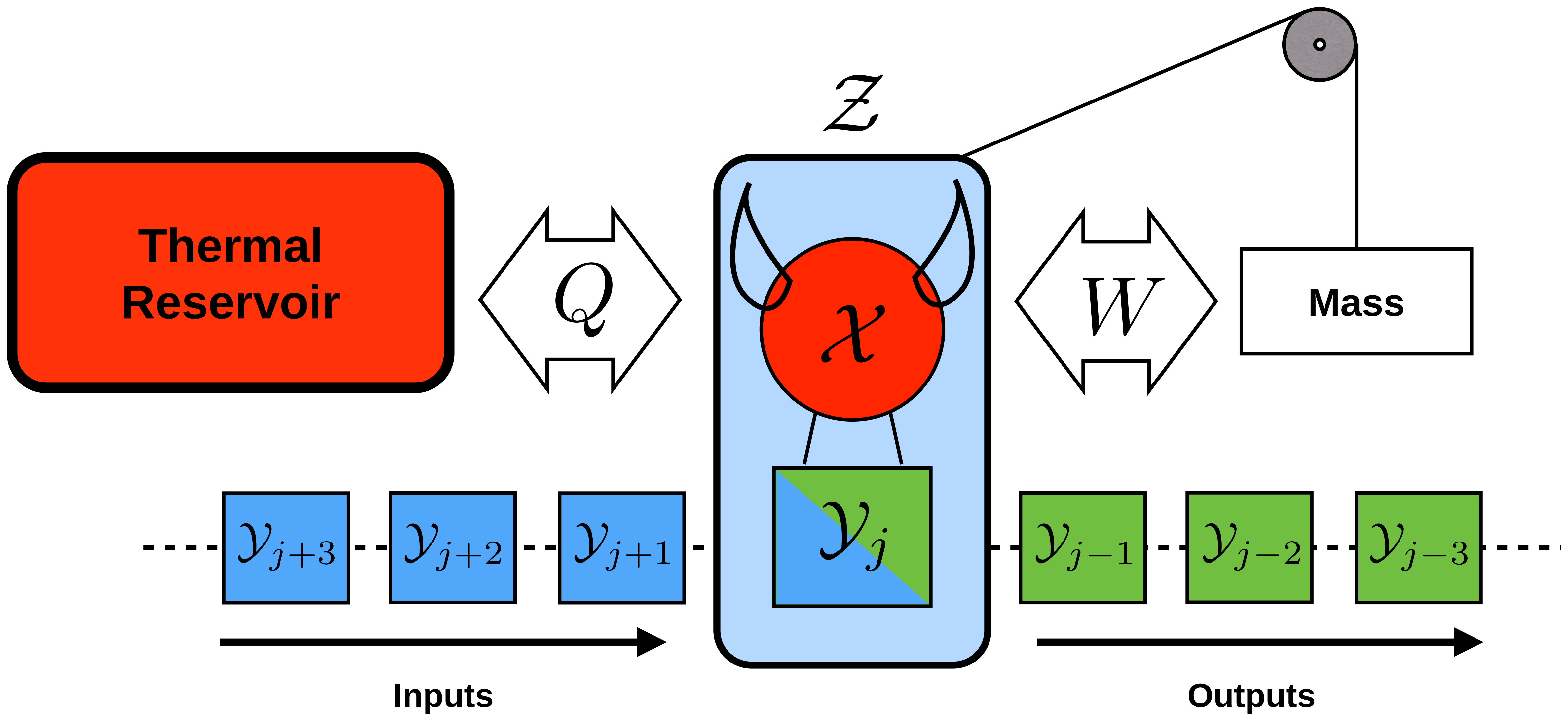


Generalized Landauer's bound: write to a hard drive to produce work

$$\langle W \rangle \leq k_B T \left(H[\mathrm{HD}_{final}] - H[\mathrm{HD}_{init}] \right)$$

Information is a Thermodynamic Fuel





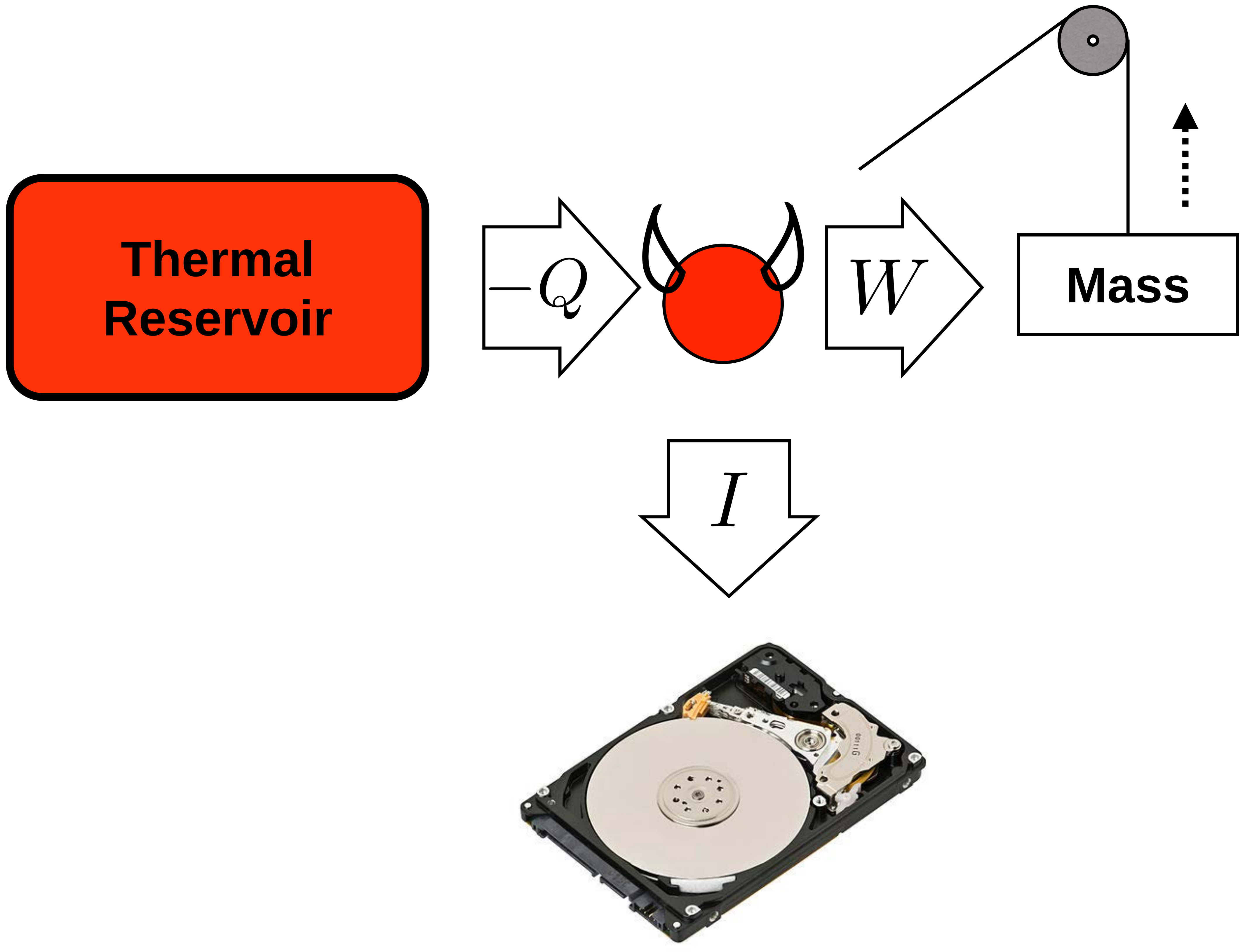
$$\text{IPSL: } \langle W \rangle_\infty \leq k_B T (h_\mu^{\text{out}} - h_\mu^{\text{in}})$$

Generalized Landauer's bound: write to a hard drive to produce work

$$\langle W \rangle \leq k_B T \left(H[\text{HD}_{final}] - H[\text{HD}_{init}] \right)$$

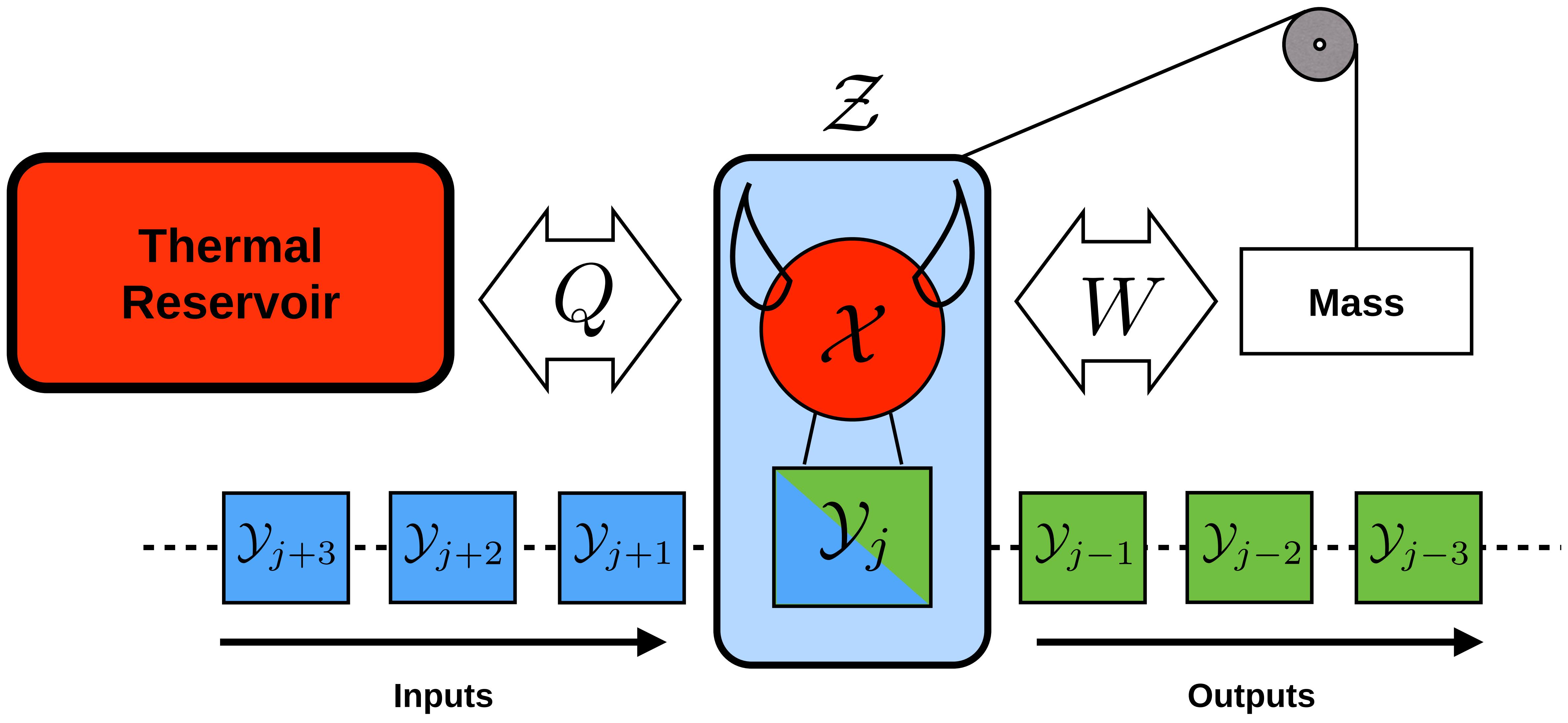
A. B. Boyd, D. Mandal, and J. P. Crutchfield. Identifying functional thermodynamics in autonomous Maxwellian ratchets. *New J. Physics*, **18**, 023049, (2016)

Information is a Thermodynamic Fuel



Generalized Landauer's bound: write to a hard drive to produce work

$$\langle W \rangle \leq k_B T \left(H[\mathrm{HD}_{final}] - H[\mathrm{HD}_{init}] \right)$$



$$\text{IPSL: } \langle W \rangle_\infty \leq k_B T (h_\mu^{\text{out}} - h_\mu^{\text{in}})$$

Thermodynamics to computational mechanics

A. B. Boyd, D. Mandal, and J. P. Crutchfield. Identifying functional thermodynamics in autonomous Maxwellian ratchets. *New J. Physics*, **18**, 023049, (2016)

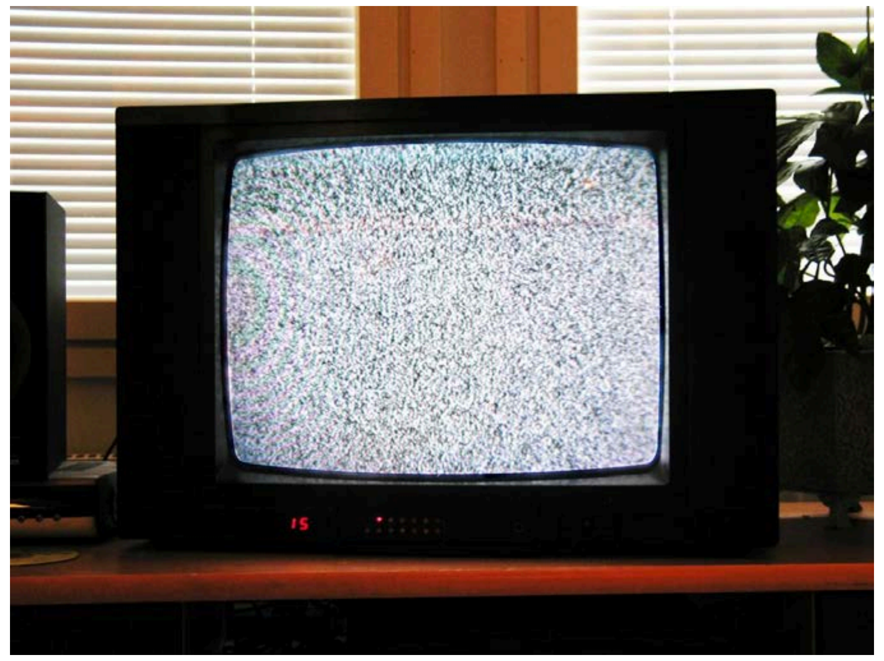
Information Fuels

Work production depends on what the hard drive stores

$$\langle W \rangle \leq k_B T \left(H[\mathrm{HD}_{final}] - H[\mathrm{HD}_{init}] \right)$$

Information Fuels

Work production depends on what the hard drive stores

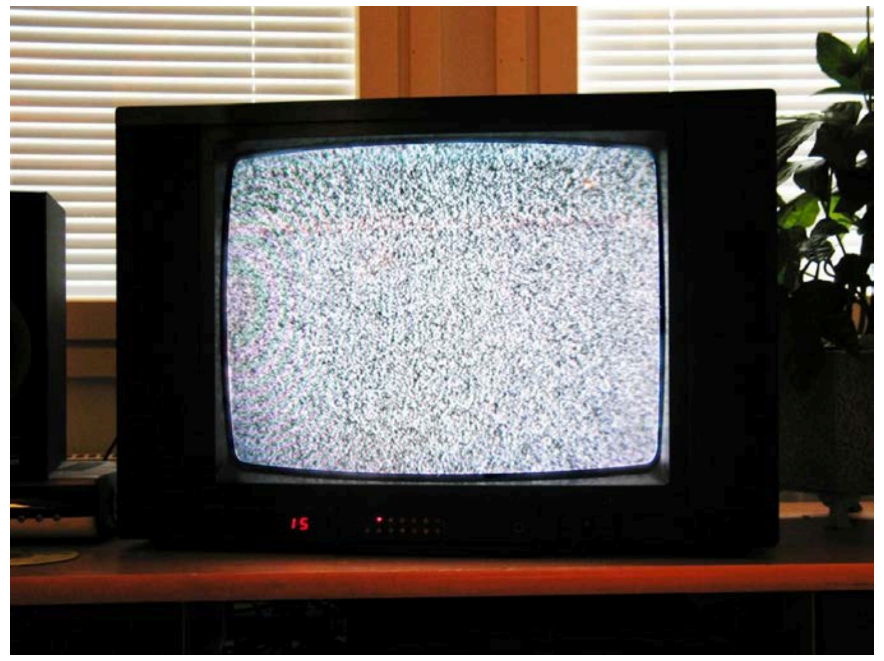


https://en.wikipedia.org/wiki/File:TV_noise.jpg

$$\langle W \rangle \leq k_B T \left(H[\mathrm{HD}_{final}] - H[\mathrm{HD}_{init}] \right)$$

Information Fuels

Work production depends on what the hard drive stores



https://en.wikipedia.org/wiki/File:TV_noise.jpg

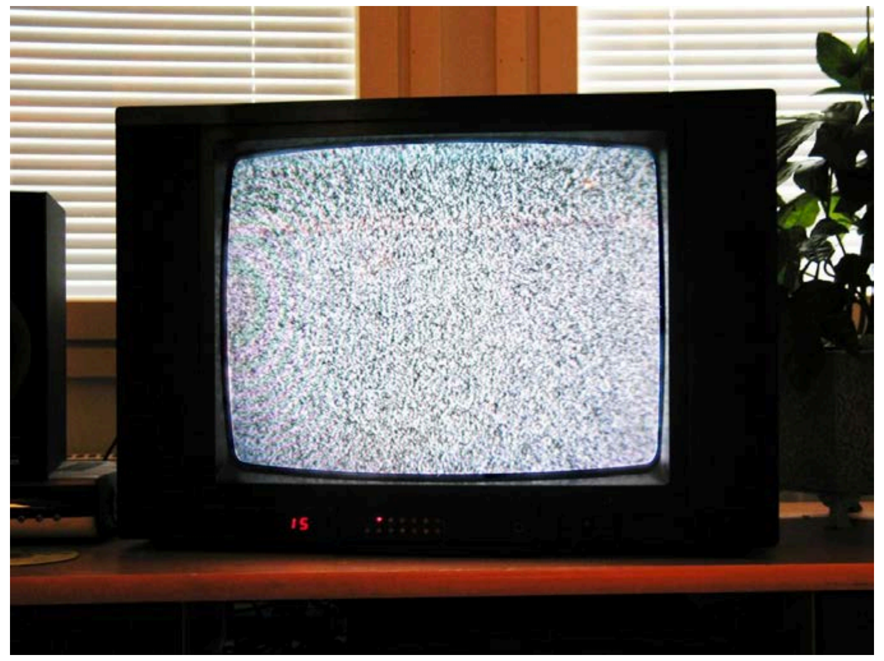
$$H[\text{white noise}] = N \ln 2$$

$$\langle W \rangle \leq 0$$

$$\langle W \rangle \leq k_B T \left(H[\text{HD}_{final}] - H[\text{HD}_{init}] \right)$$

Information Fuels

Work production depends on what the hard drive stores



https://en.wikipedia.org/wiki/File:TV_noise.jpg

$$H[\text{white noise}] = N \ln 2$$

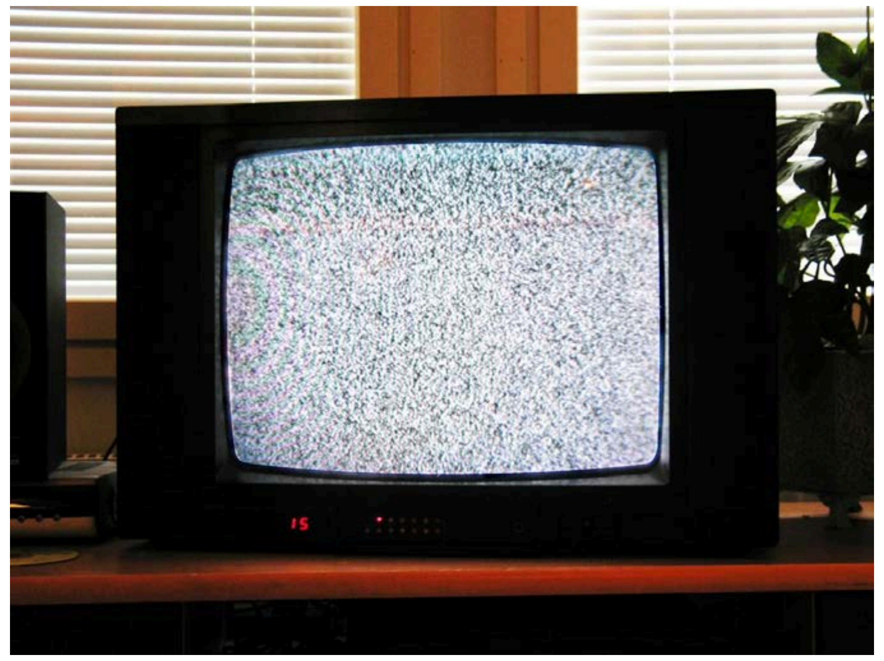
$$\langle W \rangle \leq 0$$

$$\langle W \rangle \leq k_B T \left(H[\text{HD}_{final}] - H[\text{HD}_{init}] \right)$$

000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000

Information Fuels

Work production depends on what the hard drive stores



https://en.wikipedia.org/wiki/File:TV_noise.jpg

$$H[\text{white noise}] = N \ln 2$$

$$\langle W \rangle \leq 0$$

$$\langle W \rangle \leq k_B T \left(H[\text{HD}_{final}] - H[\text{HD}_{init}] \right)$$

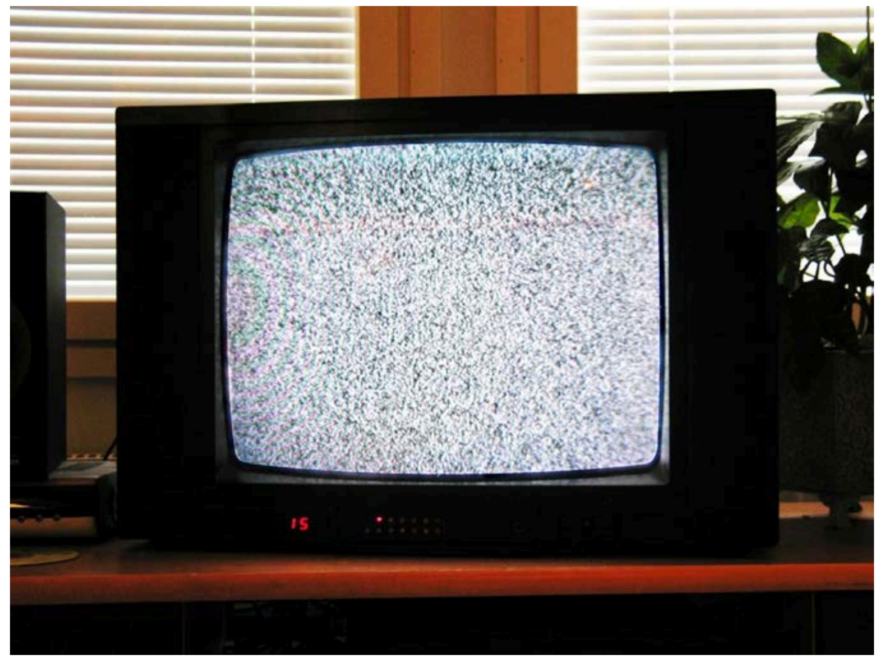
$$H[\text{all zeros}] = 0$$

$$\langle W \rangle \leq k_B T N \ln 2$$

0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000

Information Fuels

Work production depends on what the hard drive stores



https://en.wikipedia.org/wiki/File:TV_noise.jpg

$$H[\text{white noise}] = N \ln 2$$

$$\langle W \rangle \leq 0$$

$$\langle W \rangle \leq k_B T \left(H[\text{HD}_{final}] - H[\text{HD}_{init}] \right)$$

https://en.wikipedia.org/wiki/Harry_Potter_(character)

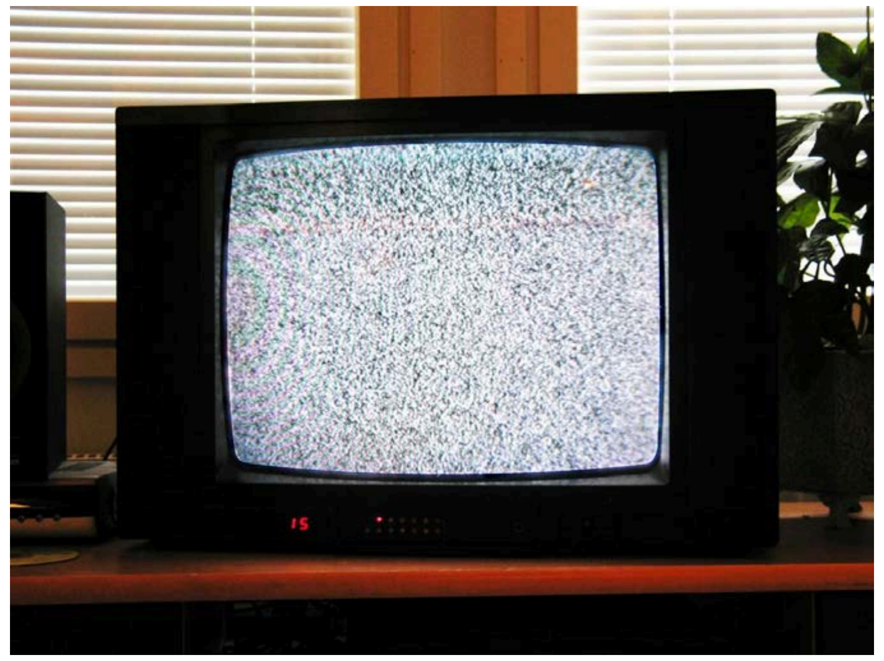
$$H[\text{all zeros}] = 0$$

$$\langle W \rangle \leq k_B T N \ln 2$$

000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000

Information Fuels

Work production depends on what the hard drive stores



https://en.wikipedia.org/wiki/File:TV_noise.jpg

$$H[\text{white noise}] = N \ln 2$$

$$\langle W \rangle \leq 0$$

$$\langle W \rangle \leq k_B T \left(H[\text{HD}_{final}] - H[\text{HD}_{init}] \right)$$

$$H[\text{all zeros}] = 0$$

$$\langle W \rangle \leq k_B T N \ln 2$$

000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000

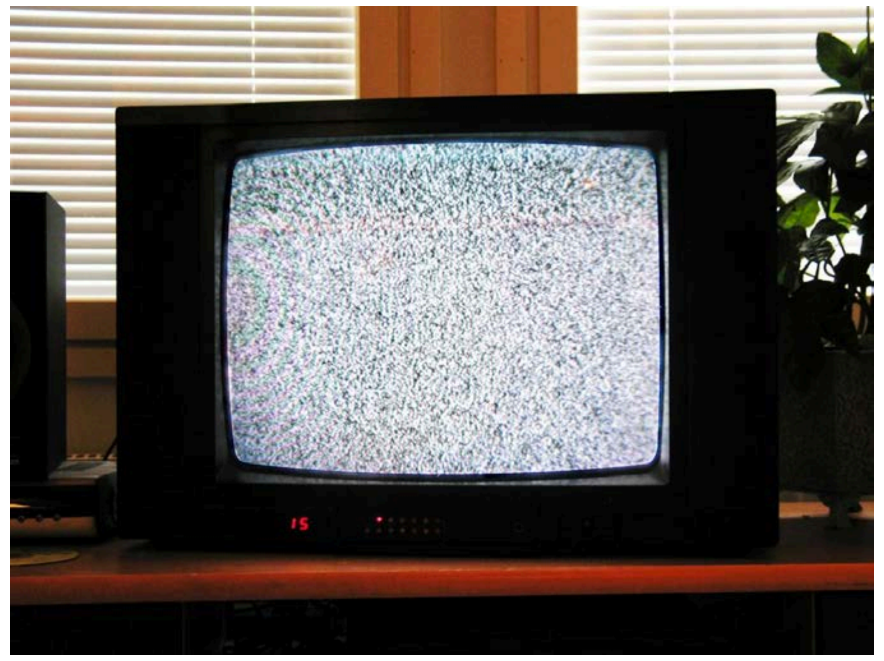
https://en.wikipedia.org/wiki/Harry_Potter_(character)

$$H[\text{Harry Potter}] = 0$$

$$\langle W \rangle \leq k_B T N \ln 2$$

Information Fuels

Work production depends on what the hard drive stores



https://en.wikipedia.org/wiki/File:TV_noise.jpg

$$H[\text{white noise}] = N \ln 2$$

$$\langle W \rangle \leq 0$$

$$\langle W \rangle \leq k_B T \left(H[\text{HD}_{final}] - H[\text{HD}_{init}] \right)$$

$$H[\text{all zeros}] = 0$$

$$\langle W \rangle \leq k_B T N \ln 2$$

0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000

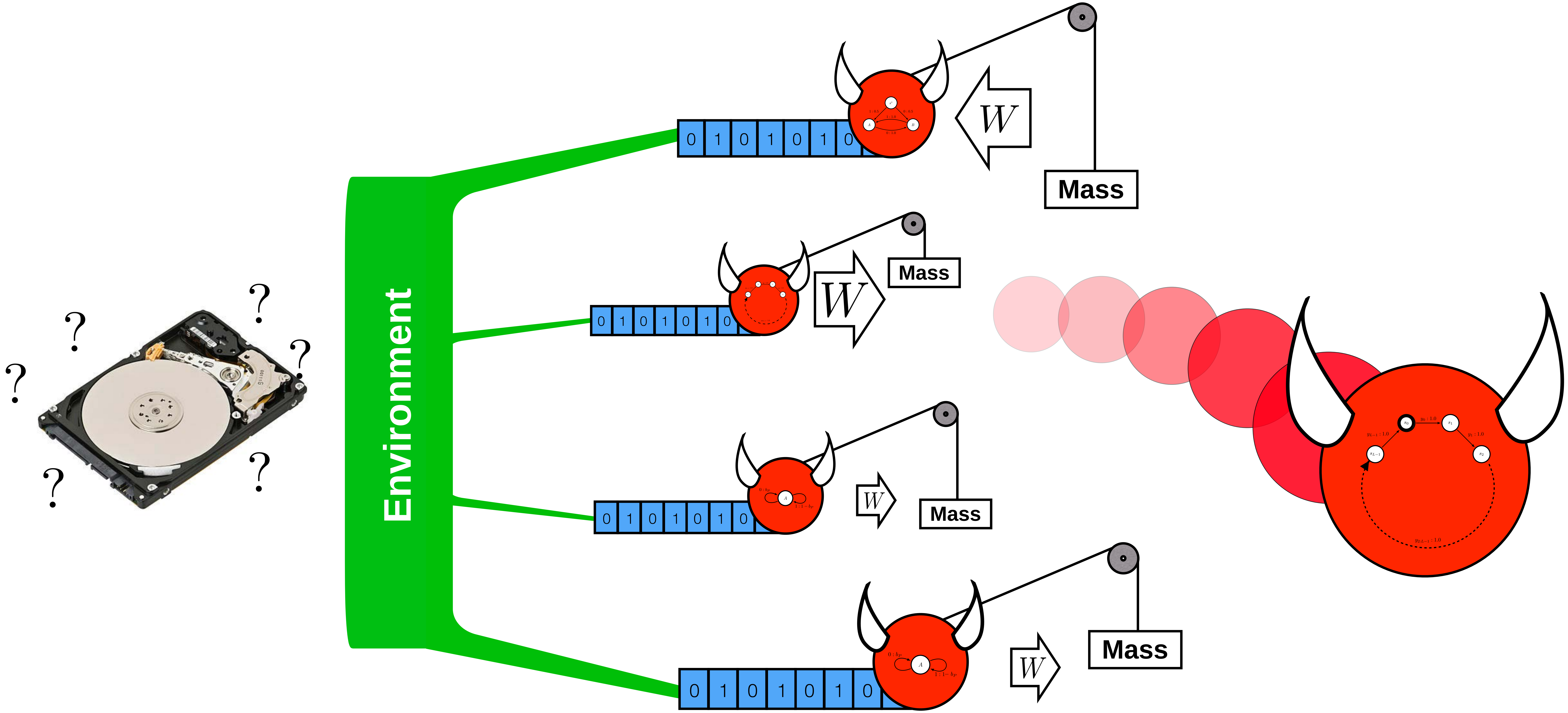
https://en.wikipedia.org/wiki/Harry_Potter_(character)

$$H[\text{Harry Potter}] = 0$$

$$\langle W \rangle \leq k_B T N \ln 2$$

Harry Potter may appear like white noise if one doesn't have memory, because it is more **complex** (requires memory for prediction).

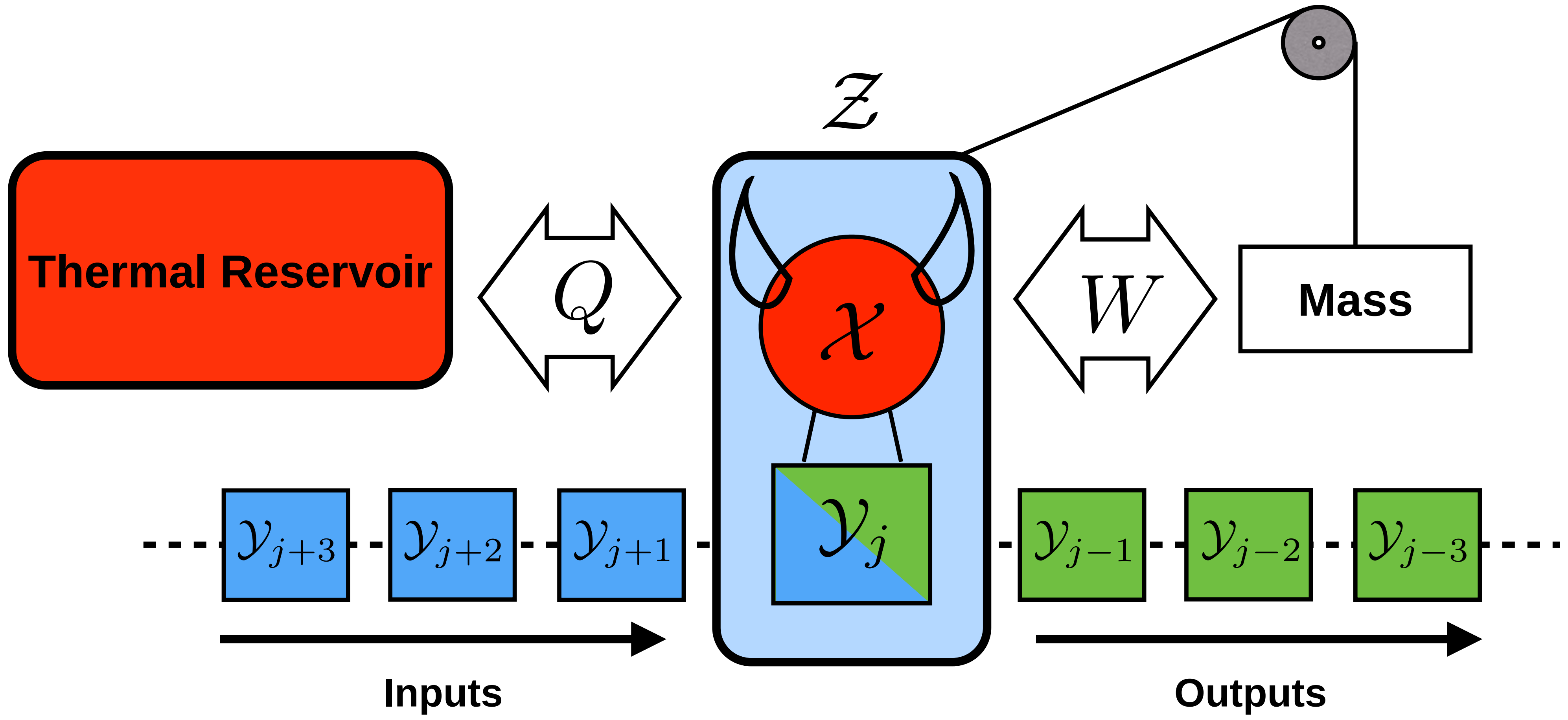
Challenge: Harvest Maximum Work



Thermodynamic Learning Through Maximum Work Production

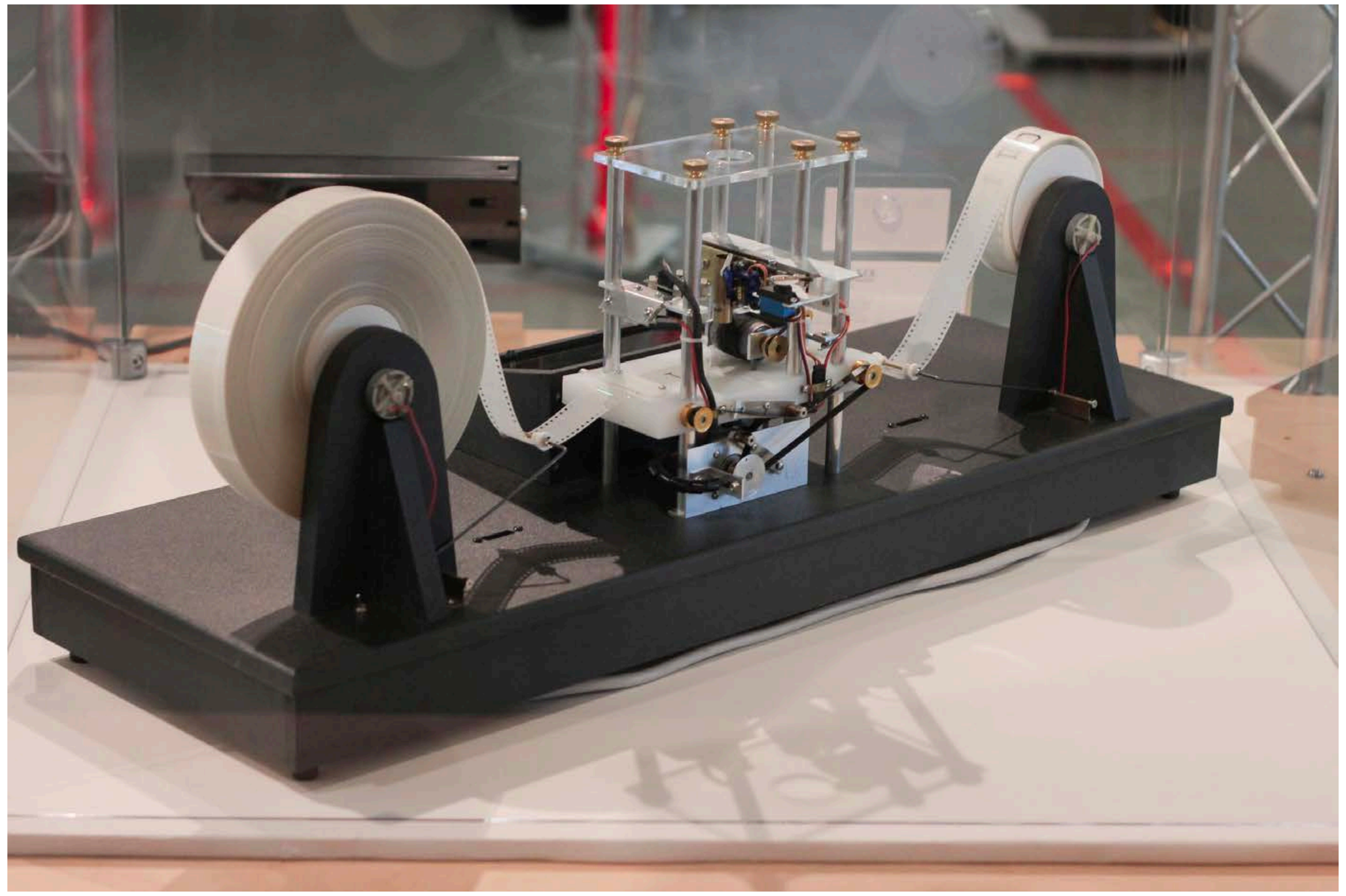
Modular Thermodynamic Ratchets

Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.



$\mathcal{X}$ = agent memory

$\mathcal{Y}_i$ = interaction bit at time i



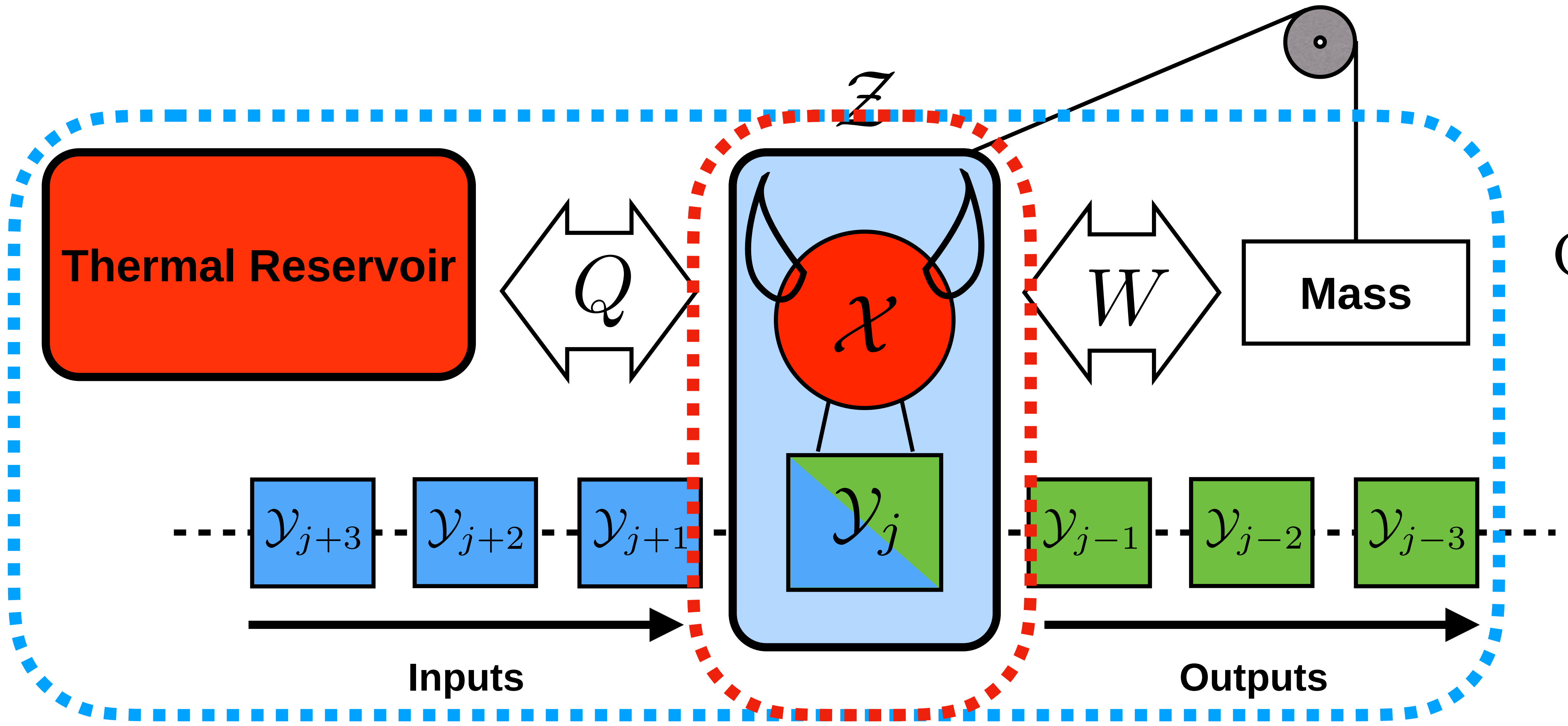
https://en.wikipedia.org/wiki/Turing_machine

Like a stochastic Turing machine

Operations are localized and "modular": $H(t) = H_{\mathcal{X},\mathcal{Y}_j}(t) + H_{\mathcal{Y}_{0:\infty} - \mathcal{Y}_j}$ $\rightarrow$ Landauer's bound applies locally:

$$\langle Q_i^{\text{local}} \rangle_{\min} = -k_B T \Delta H_{\mathcal{X}\mathcal{Y}_i}$$

Thermodynamics of Modularity



Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

$$\langle Q_i \rangle \geq \langle Q_i^{\text{local}} \rangle_{\text{min}} = -k_B T \Delta H_{\mathcal{X}\mathcal{Y}_i}$$

Global change in system entropy also includes past, future, and heat bath:

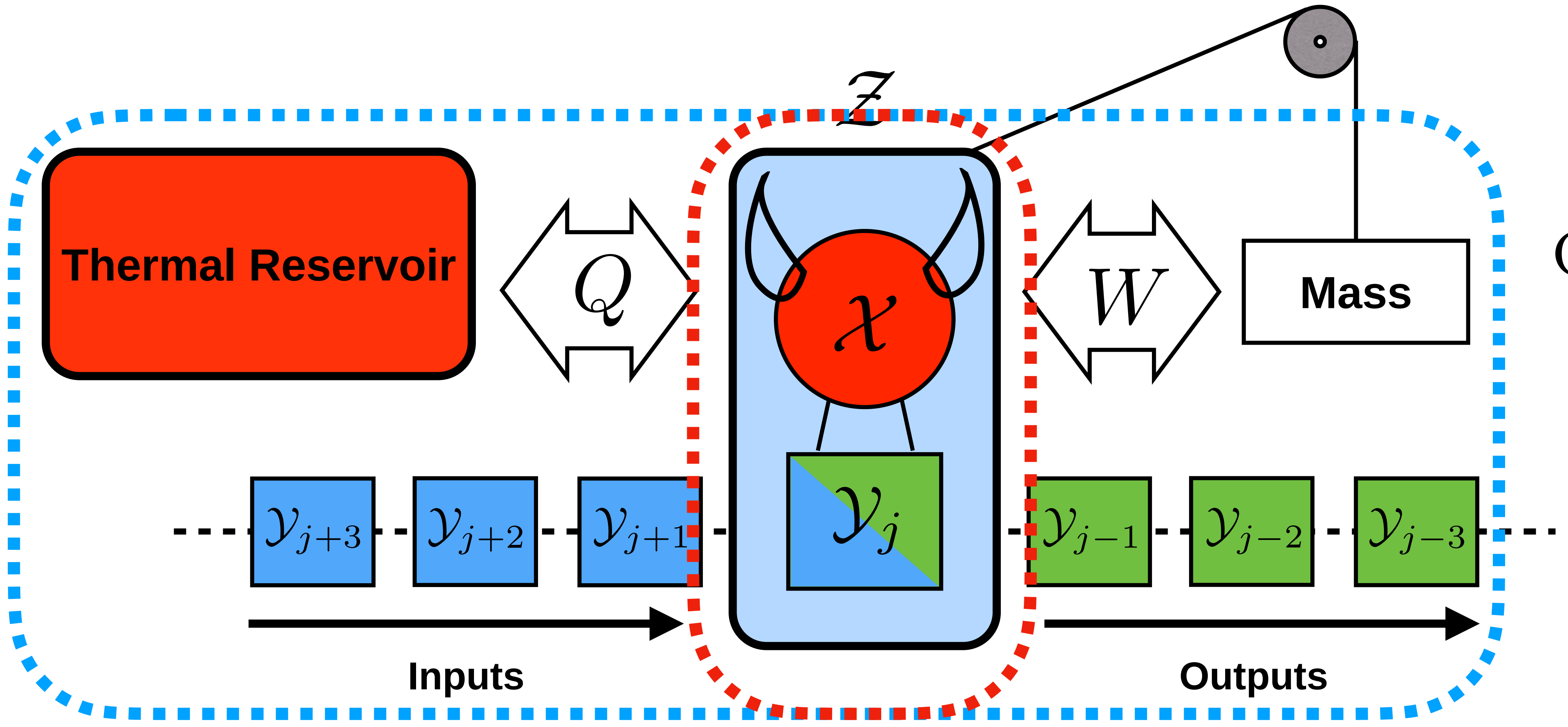
$$\langle \Delta S_i^{\text{total}} \rangle = \frac{\langle Q_i \rangle}{T} + k_B \Delta H_{\text{memory+bit string}}$$

$$\geq -k_B \Delta H_{XY_i} + k_B \Delta H_{\text{memory+bit string}}$$

$$= -k_B \Delta I[\text{memory+interaction bit; remaining bit string}]$$

$$\equiv \langle \Delta S_{\text{mod}}^{\text{total}} \rangle \qquad \textit{Modularity Dissipation}$$

Thermodynamics of Modularity



Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

$$\langle Q_i \rangle \geq \langle Q_i^{\text{local}} \rangle_{\min} = -k_B T \Delta H_{\mathcal{X}\mathcal{Y}_i}$$

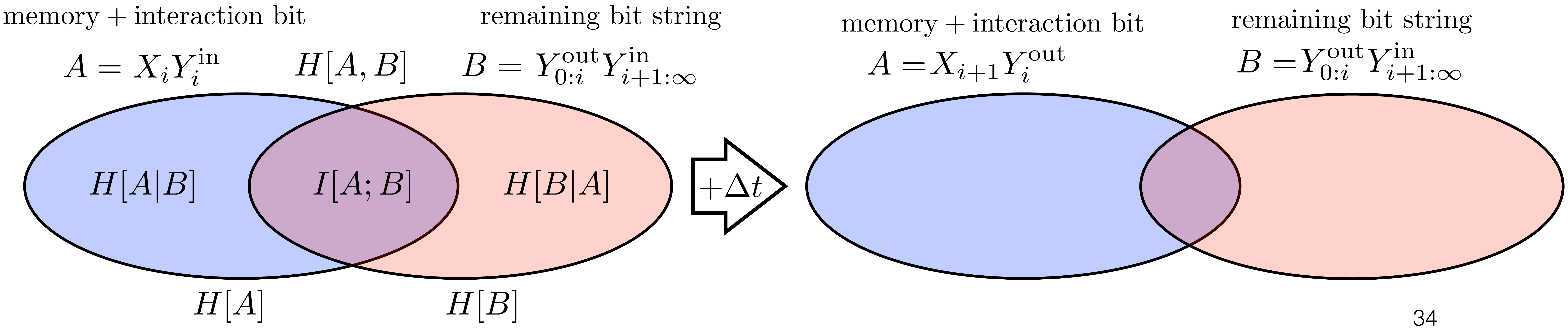
Global change in system entropy also includes past, future, and heat bath:

$$\langle \Delta S_i^{\text{total}} \rangle = \frac{\langle Q_i \rangle}{T} + k_B \Delta H_{\text{memory+bit string}}$$

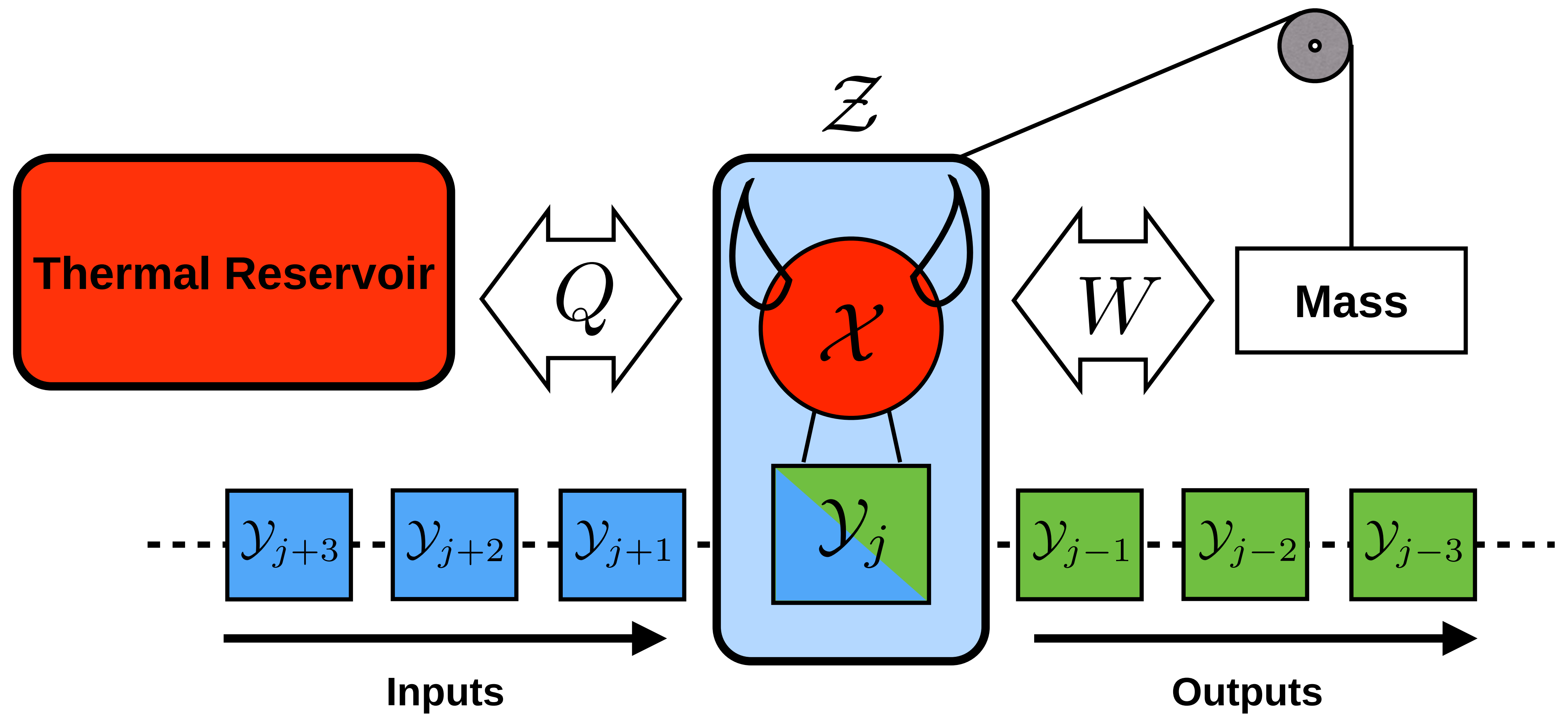
$$\geq -k_B \Delta H_{XY_i} + k_B \Delta H_{\text{memory+bit string}}$$

$$= -k_B \Delta I[\text{memory+interaction bit; remaining bit string}]$$

$$\equiv \langle \Delta S_{\text{mod}}^{\text{total}} \rangle \qquad \textit{Modularity Dissipation}$$

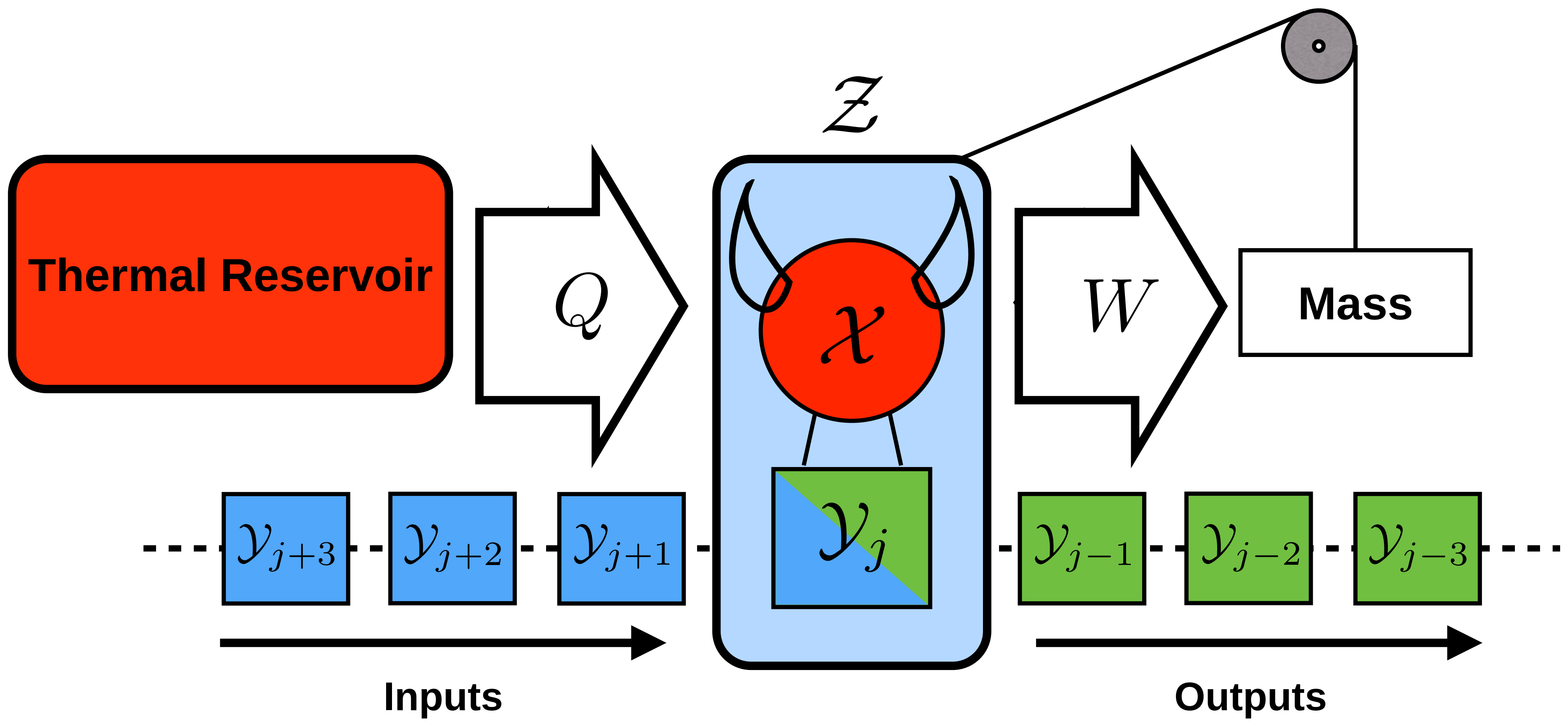


Harvesting Energy From Patterns



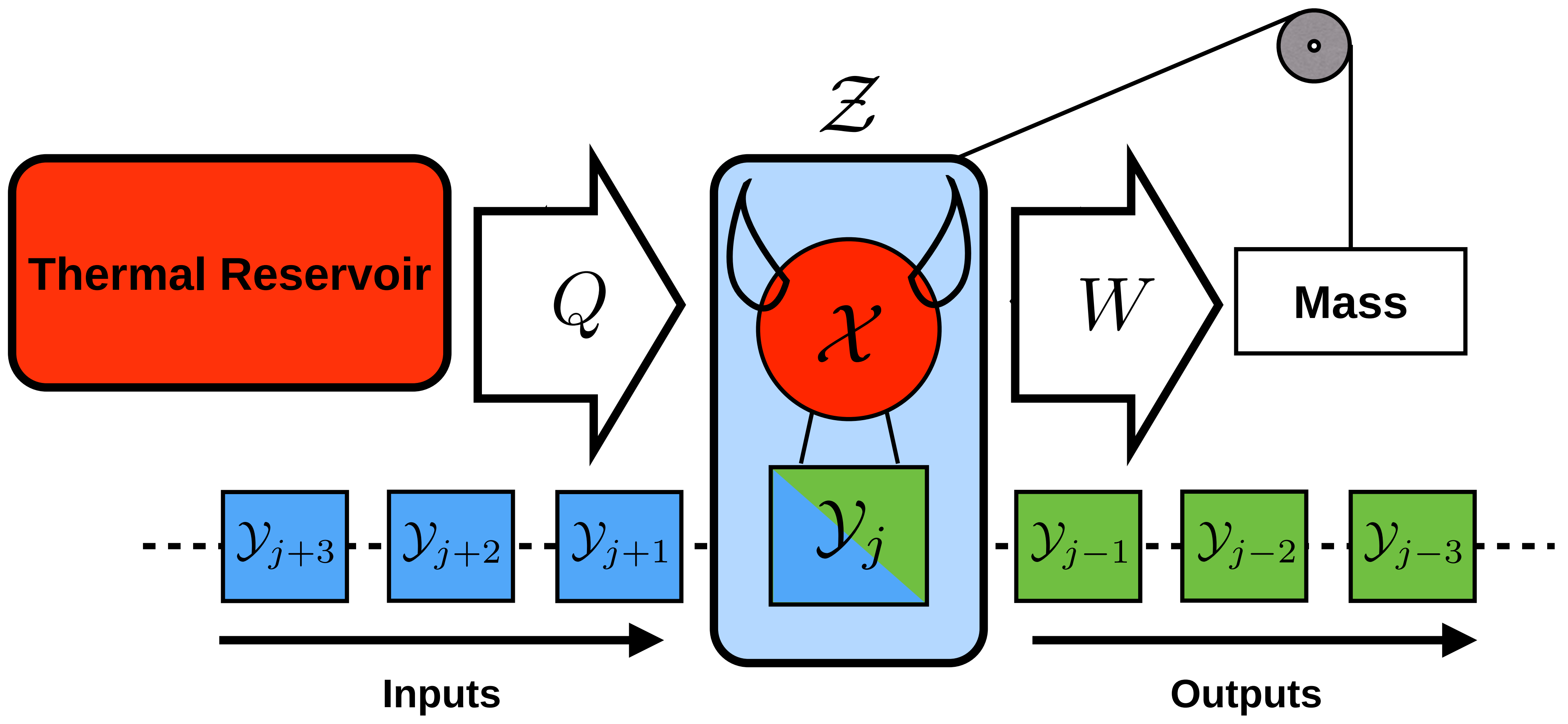
Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

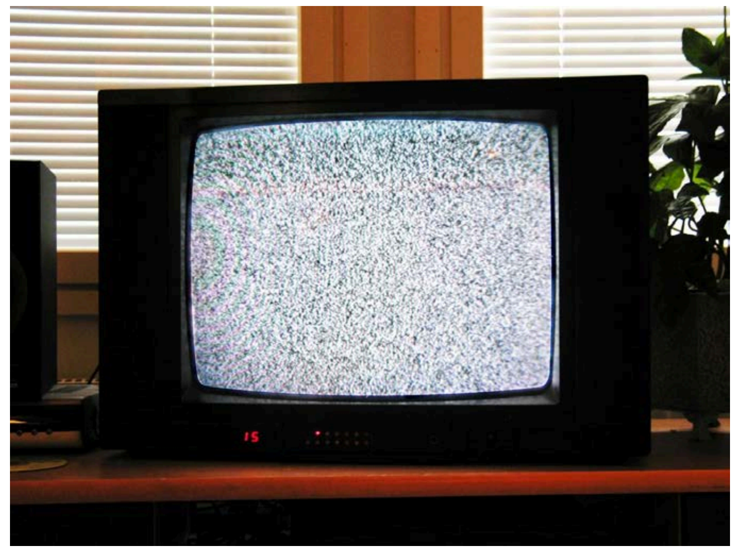
Harvesting Energy From Patterns



Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

Harvesting Energy From Patterns



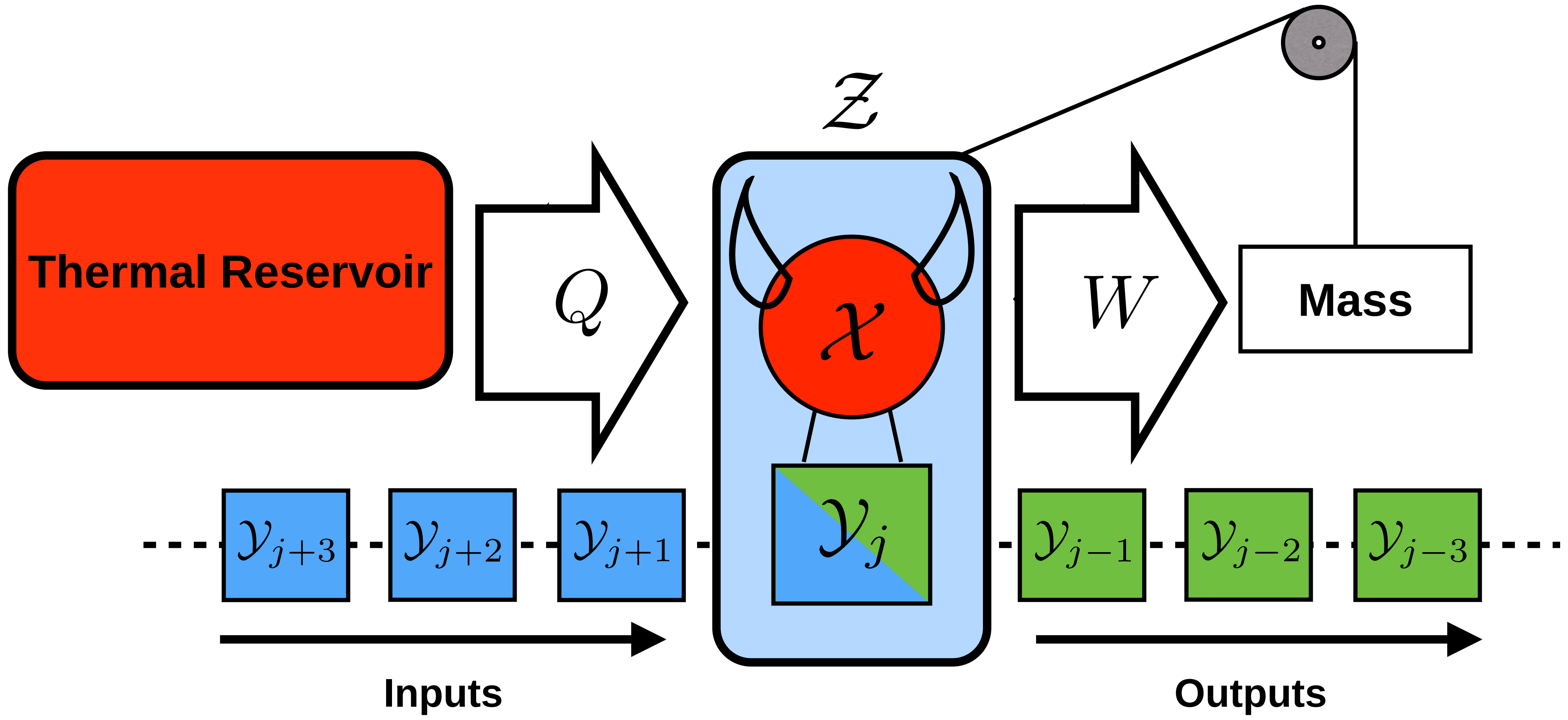


https://en.wikipedia.org/wiki/File:TV_noise.jpg

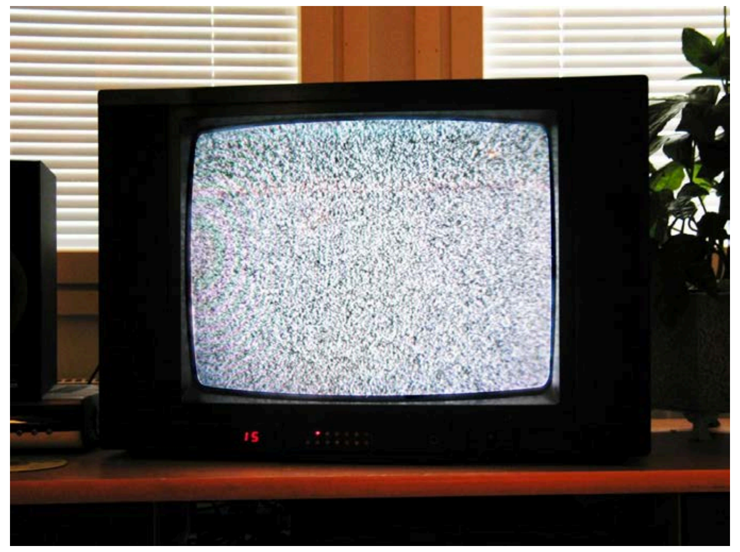
Outputs randomized and uncorrelated: $\Pr(Y^{\text{out}}_{a:b}) = \frac{1}{|\mathcal{Y}|^{b-a}}$

Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

Harvesting Energy From Patterns



https://en.wikipedia.org/wiki/Harry_Potter_(character)



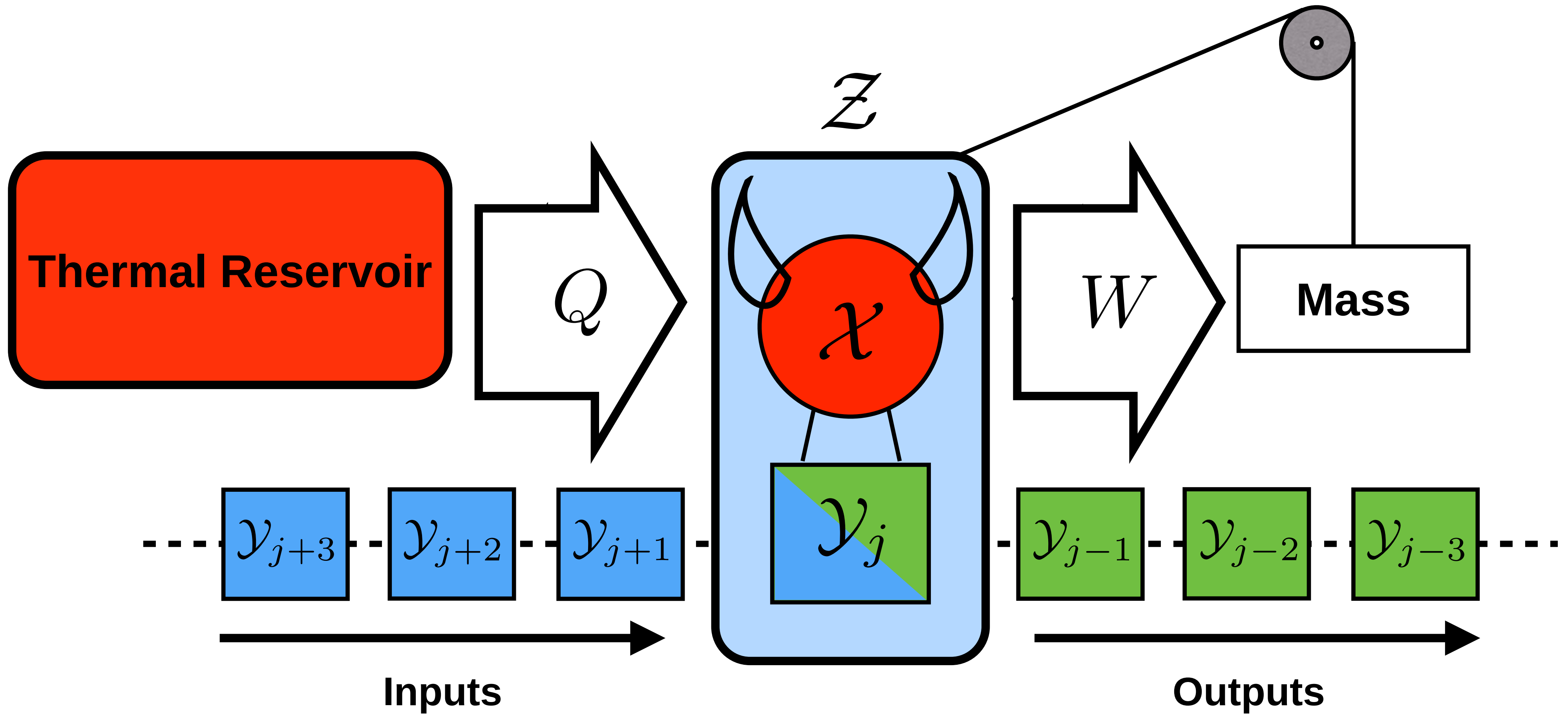
https://en.wikipedia.org/wiki/File:TV_noise.jpg

Outputs randomized and uncorrelated: $\Pr(Y^{\text{out}}_{a:b}) = \frac{1}{|\mathcal{Y}|^{b-a}}$

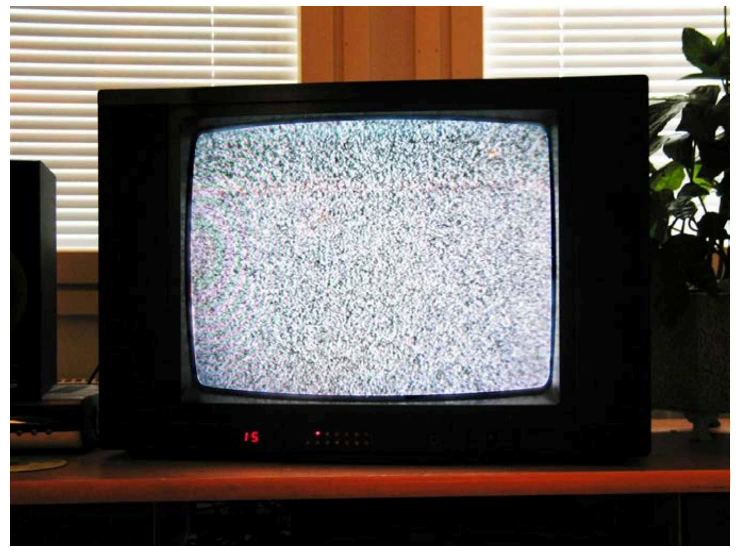
Input structured: $\Pr(Y^{\text{in}}_{a:b}) \neq \prod_{i=a}^{b-1} \Pr(Y^{\text{in}}_i)$

Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

Harvesting Energy From Patterns



https://en.wikipedia.org/wiki/Harry_Potter_(character)



https://en.wikipedia.org/wiki/File:TV_noise.jpg

Outputs randomized and uncorrelated: $\Pr(Y^{\text{out}}_{a:b}) = \dfrac{1}{|\mathcal{Y}|^{b-a}}$

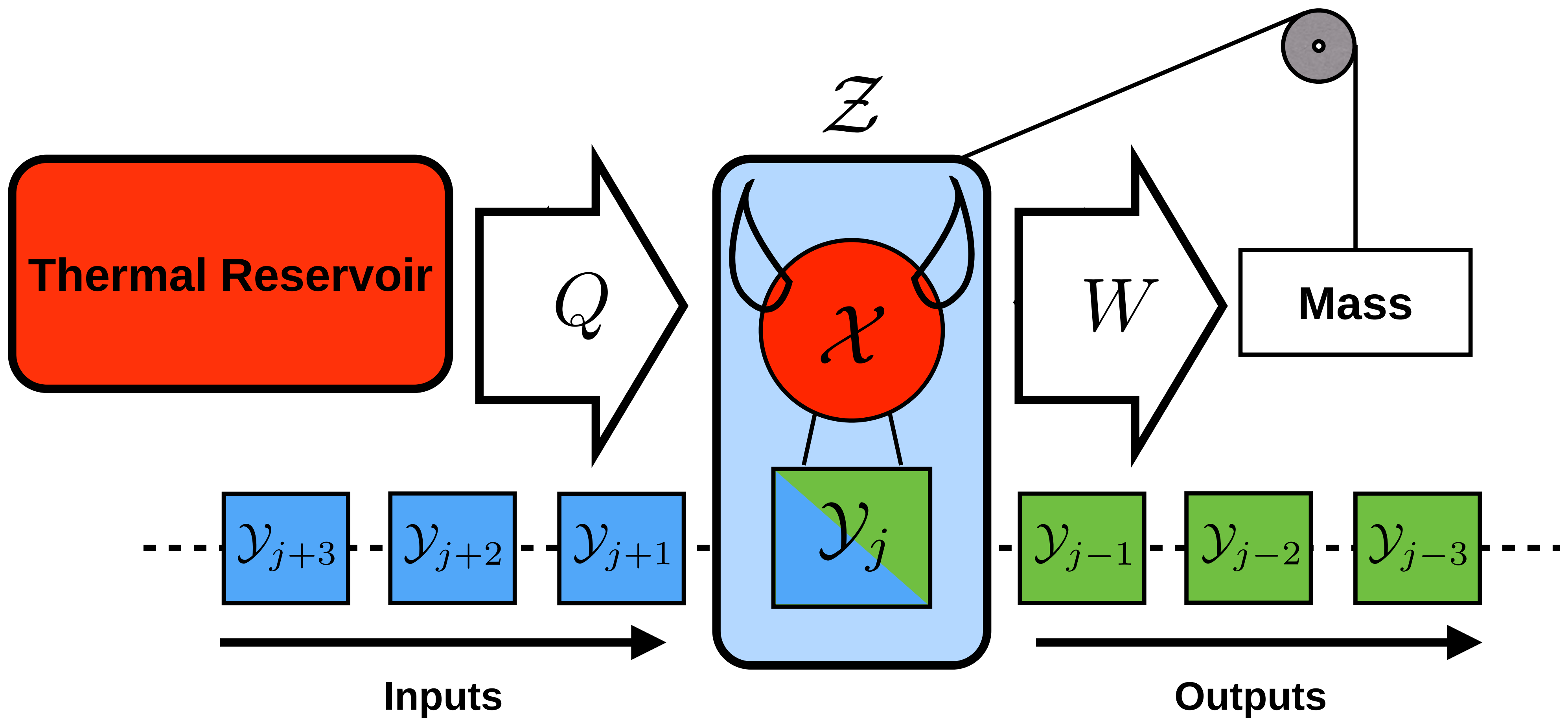
Input structured: $\Pr(Y^{\text{in}}_{a:b}) \neq \prod_{i=a}^{b-1} \Pr(Y^{\text{in}}_i)$

Remove dependence on outputs

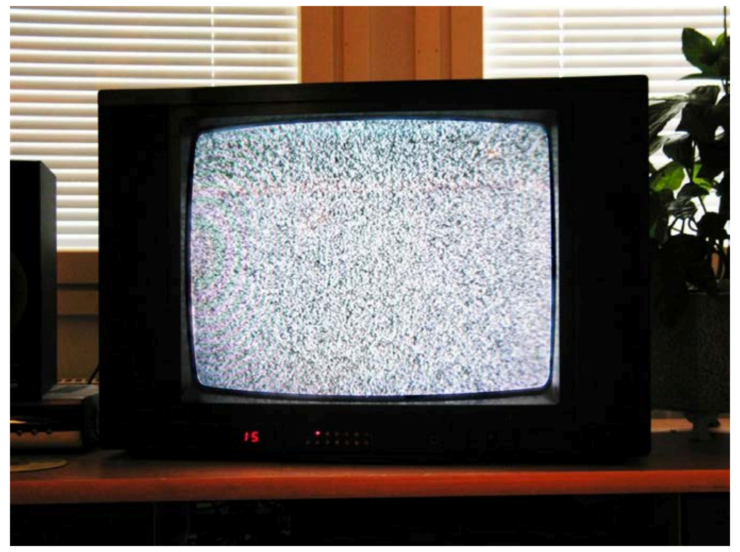
$$\langle \Delta S^{\text{total}}_{\text{mod}} \rangle = k_B (I[X_i Y^{\text{in}}_i ; \overrightarrow{Y}^{\text{in}}_{i+1}] - I[X_{i+1} ; \overrightarrow{Y}^{\text{in}}_{i+1}])$$

Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

Harvesting Energy From Patterns



https://en.wikipedia.org/wiki/Harry_Potter_(character)



https://en.wikipedia.org/wiki/File:TV_noise.jpg

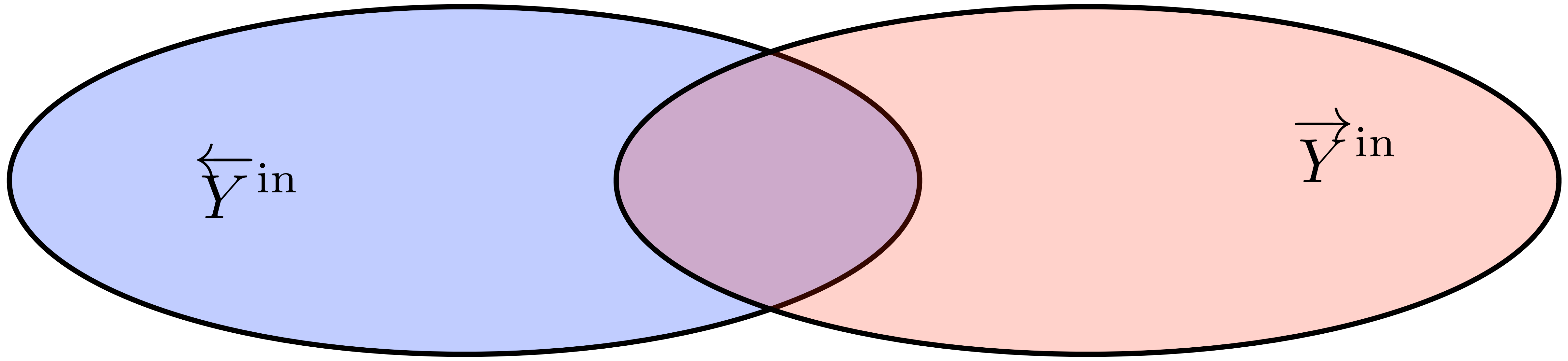
Outputs randomized and uncorrelated: $\Pr(Y^{\text{out}}_{a:b}) = \dfrac{1}{|\mathcal{Y}|^{b-a}}$

Input structured: $\Pr(Y^{\text{in}}_{a:b}) \neq \prod_{i=a}^{b-1} \Pr(Y^{\text{in}}_i)$

Remove dependence on outputs

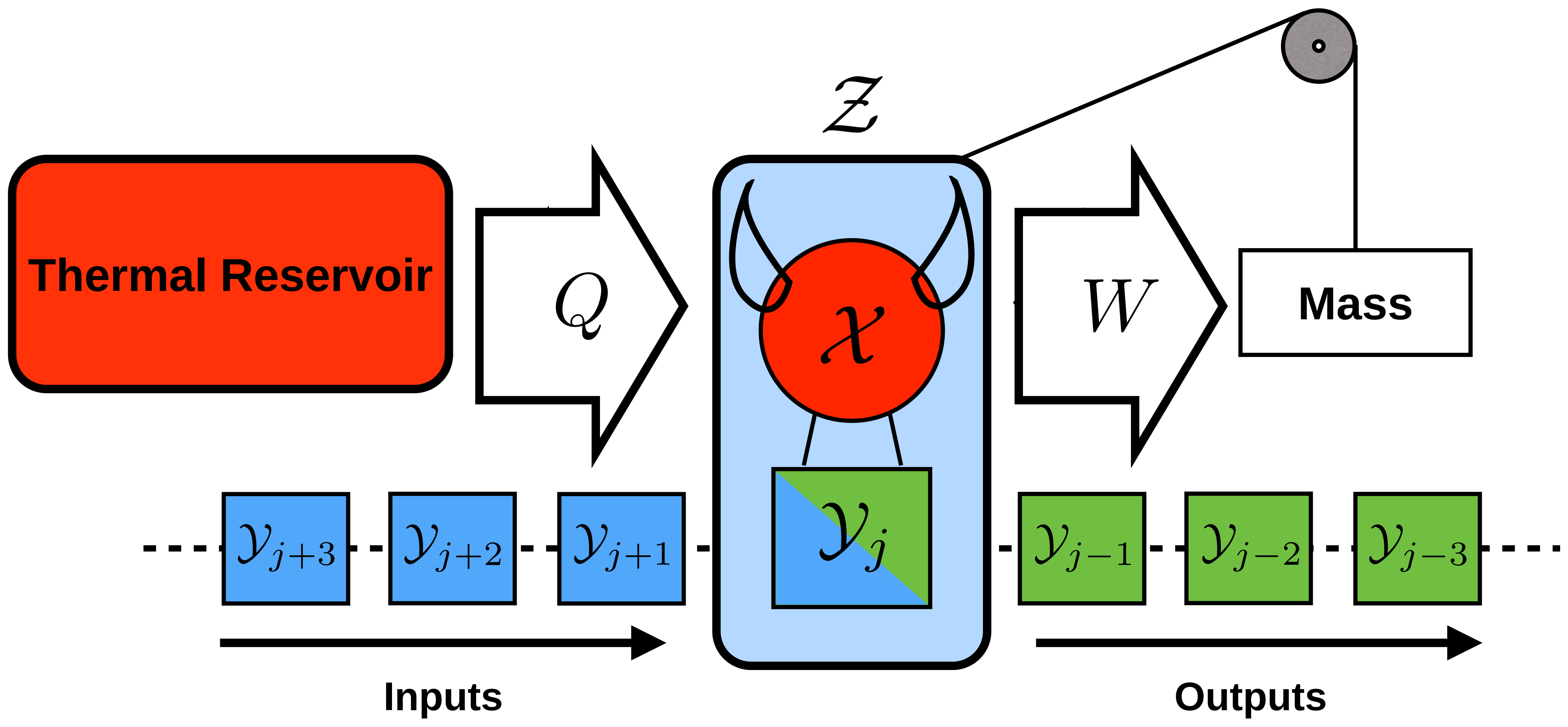
$$\langle \Delta S^{\text{total}}_{\text{mod}} \rangle = k_B (I[X_i Y^{\text{in}}_i ; \overrightarrow{Y}^{\text{in}}_{i+1}] - I[X_{i+1} ; \overrightarrow{Y}^{\text{in}}_{i+1}])$$

Of the many ways of erasing information, which is best?

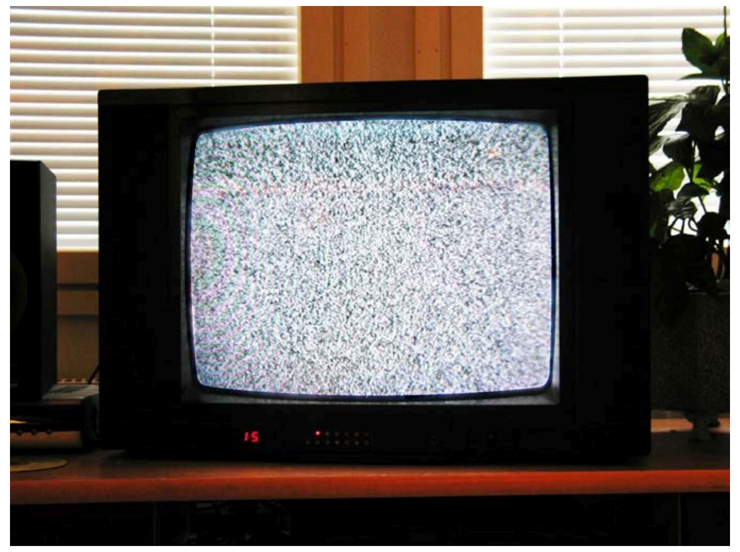


Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

Harvesting Energy From Patterns



https://en.wikipedia.org/wiki/Harry_Potter_(character)



https://en.wikipedia.org/wiki/File:TV_noise.jpg

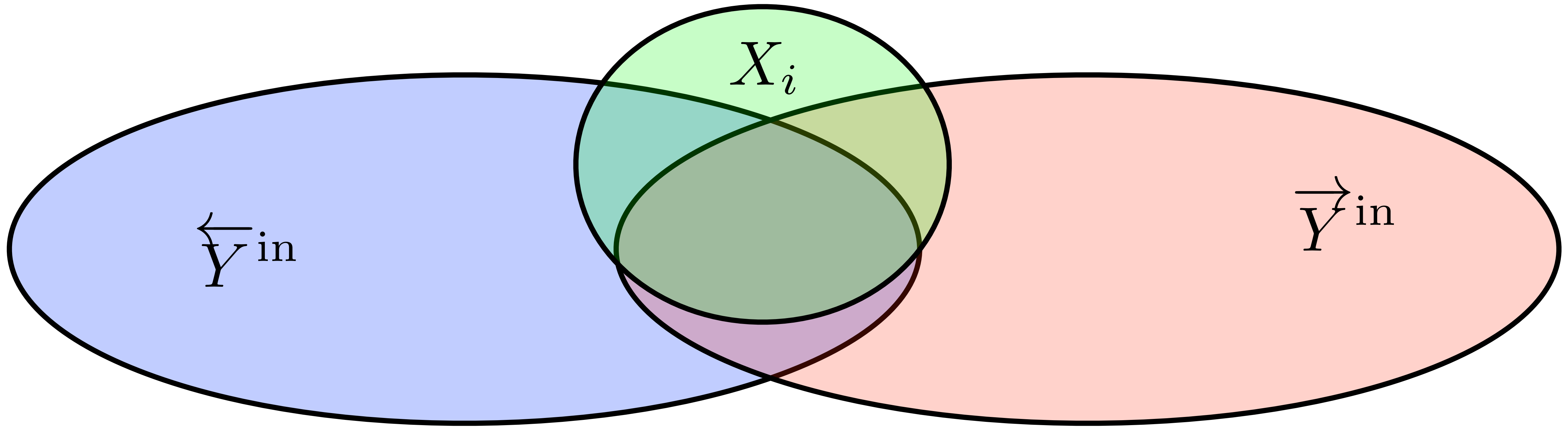
Outputs randomized and uncorrelated: $\Pr(Y^{\text{out}}_{a:b}) = \dfrac{1}{|\mathcal{Y}|^{b-a}}$

Input structured: $\Pr(Y^{\text{in}}_{a:b}) \neq \prod_{i=a}^{b-1} \Pr(Y^{\text{in}}_i)$

Remove dependence on outputs

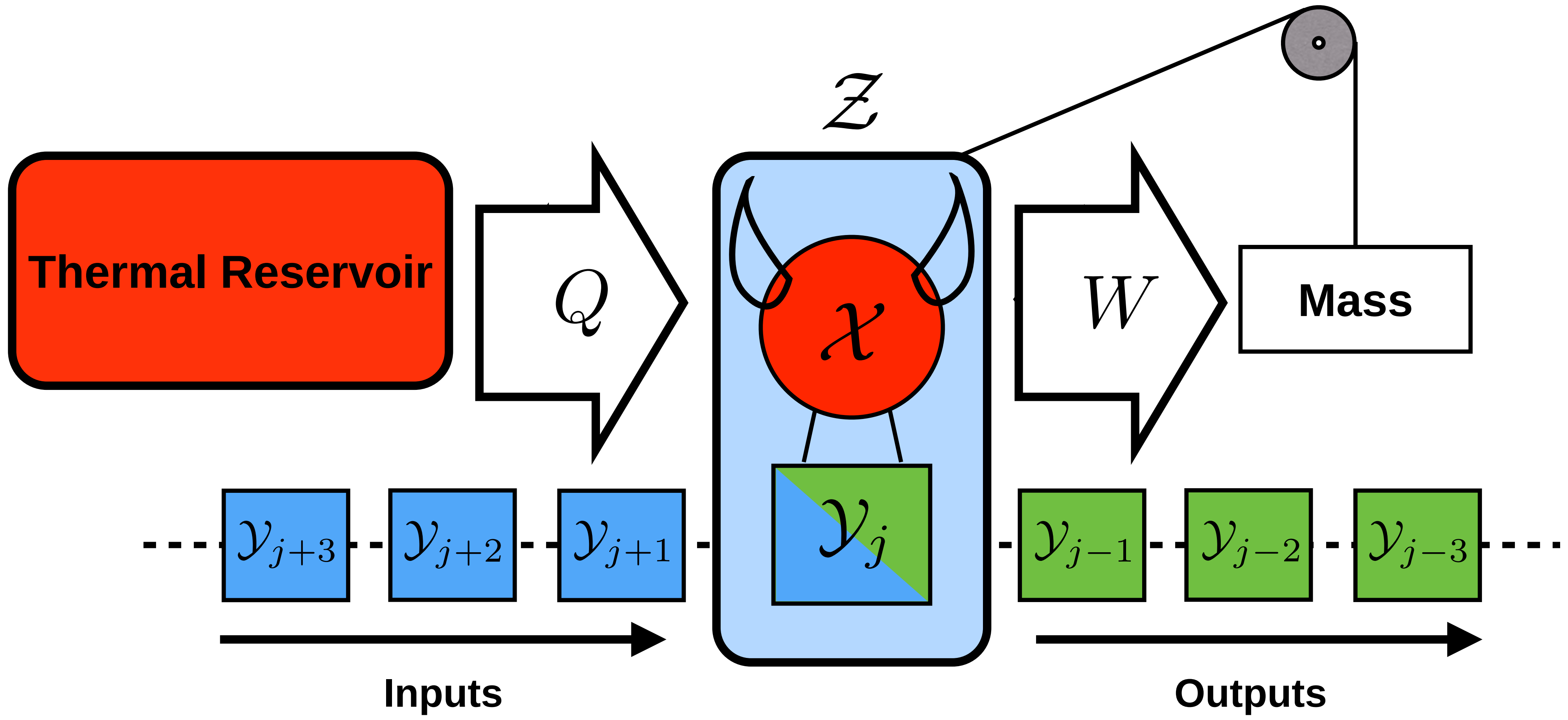
$$\langle \Delta S^{\text{total}}_{\text{mod}} \rangle = k_B(I[X_i Y_i^{\text{in}}; \overrightarrow{Y}^{\text{in}}_{i+1}] - I[X_{i+1}; \overrightarrow{Y}^{\text{in}}_{i+1}])$$

Of the many ways of erasing information, which is best?

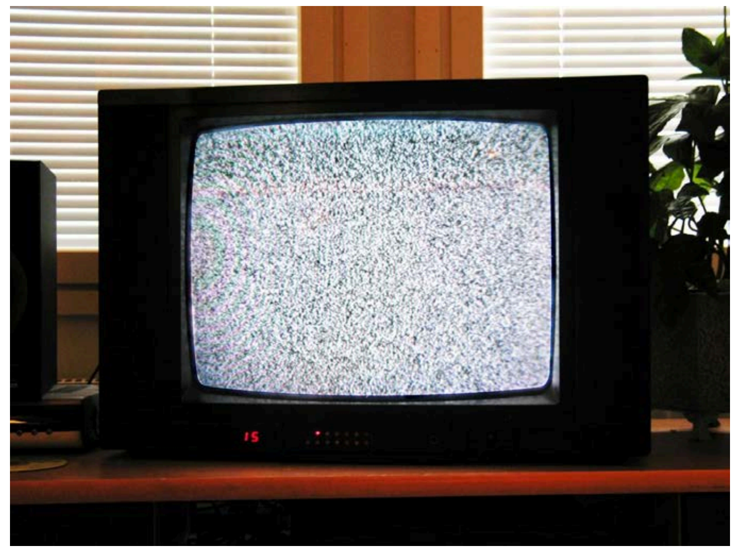


Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

Harvesting Energy From Patterns



https://en.wikipedia.org/wiki/Harry_Potter_(character)



https://en.wikipedia.org/wiki/File:TV_noise.jpg

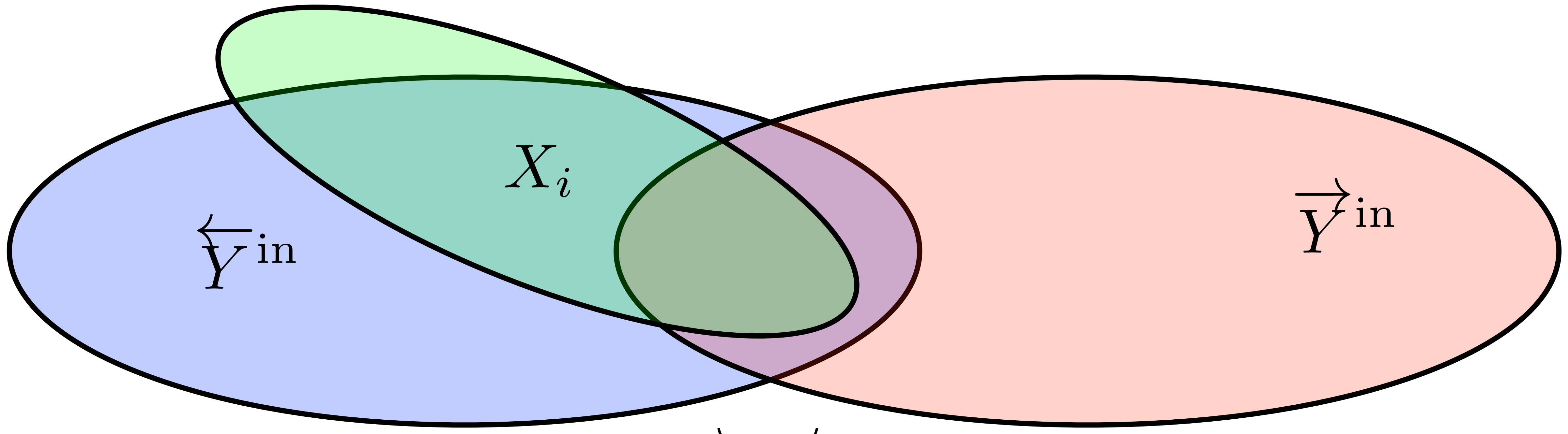
Outputs randomized and uncorrelated: $\Pr(Y_{a:b}^{\text{out}}) = \frac{1}{|\mathcal{Y}|^{b-a}}$

Input structured: $\Pr(Y_{a:b}^{\text{in}}) \neq \prod_{i=a}^{b-1} \Pr(Y_i^{\text{in}})$

Remove dependence on outputs

$$\langle \Delta S_{\text{mod}}^{\text{total}} \rangle = k_B (I[X_i Y_i^{\text{in}}; \overrightarrow{Y}_{i+1}^{\text{in}}] - I[X_{i+1}; \overrightarrow{Y}_{i+1}^{\text{in}}])$$

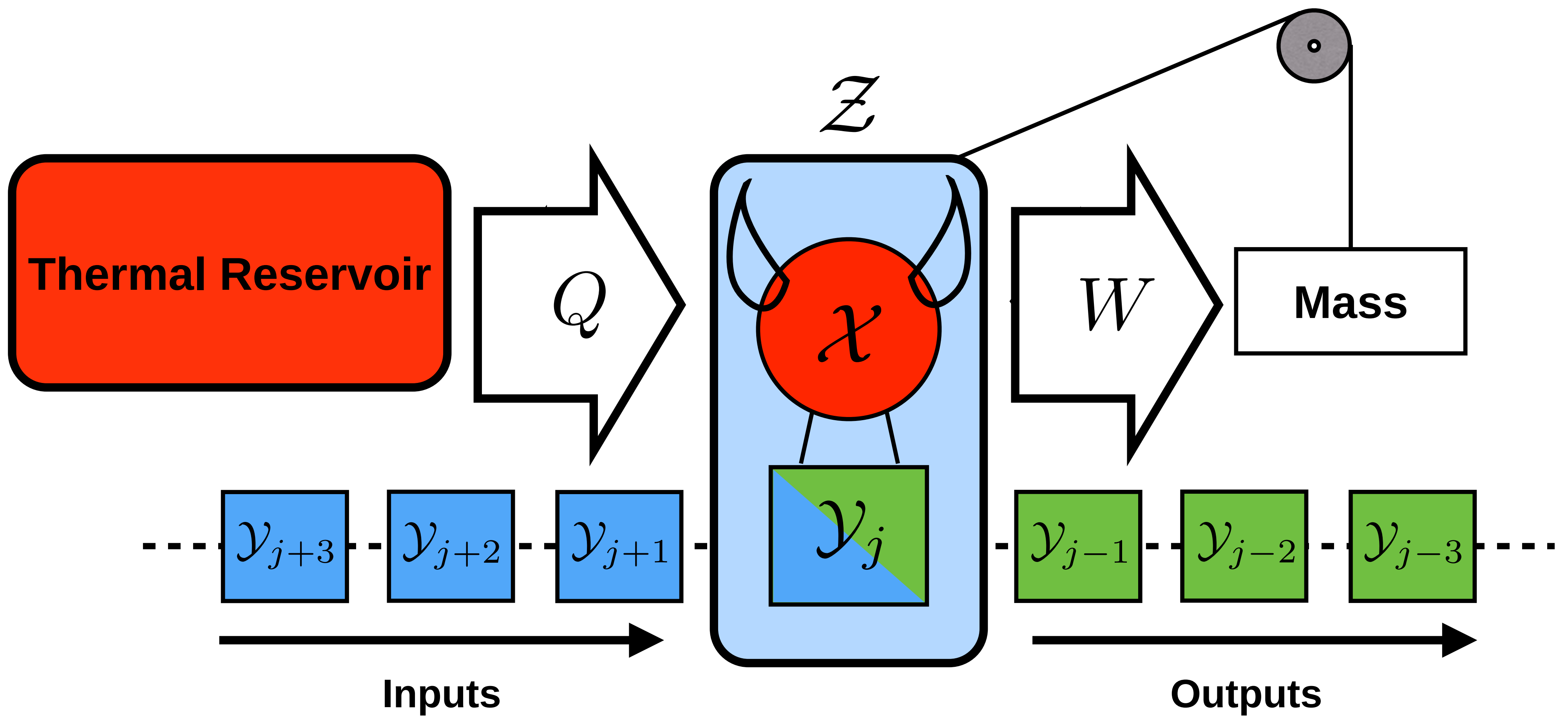
Of the many ways of erasing information, which is best?



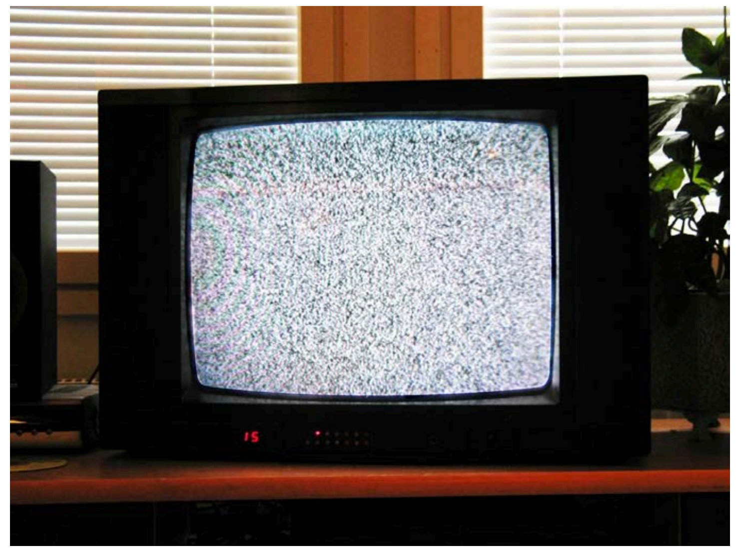
Nonanticipatory: $I[X_i; \overrightarrow{Y}^{\text{in}} | \overleftarrow{Y}^{\text{in}}] = 0$

Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

Harvesting Energy From Patterns



https://en.wikipedia.org/wiki/Harry_Potter_(character)



https://en.wikipedia.org/wiki/File:TV_noise.jpg

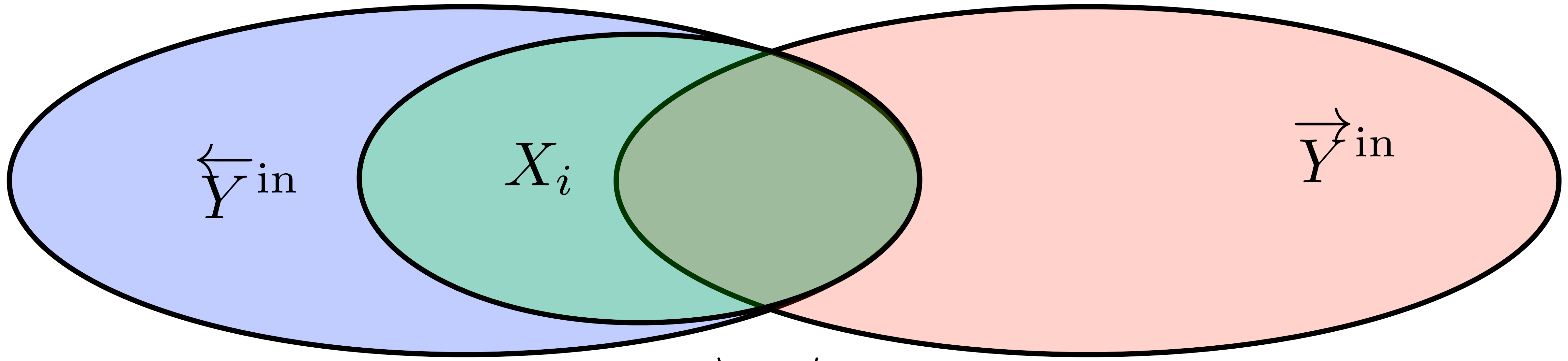
Outputs randomized and uncorrelated: $\Pr(Y^{\text{out}}_{a:b}) = \frac{1}{|\mathcal{Y}|^{b-a}}$

Input structured: $\Pr(Y^{\text{in}}_{a:b}) \neq \prod_{i=a}^{b-1} \Pr(Y^{\text{in}}_i)$

Remove dependence on outputs

$$\langle \Delta S^{\text{total}}_{\text{mod}} \rangle = k_B (I[X_i Y^{\text{in}}_i ; \overrightarrow{Y}^{\text{in}}_{i+1}] - I[X_{i+1} ; \overrightarrow{Y}^{\text{in}}_{i+1}])$$

Of the many ways of erasing information, which is best?

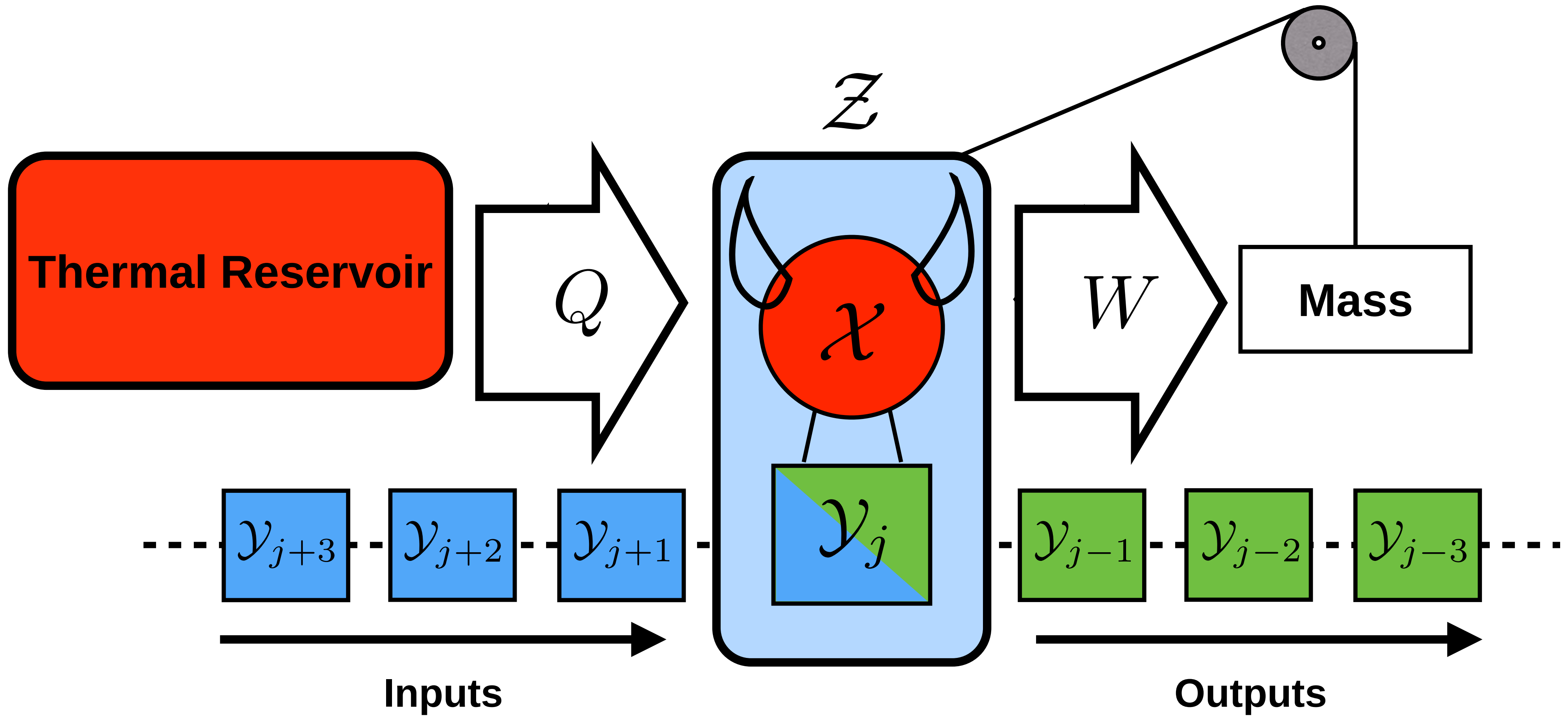


Nonanticipatory: $I[X_i ; \overrightarrow{Y}^{\text{in}} | \overleftarrow{Y}^{\text{in}}] = 0$

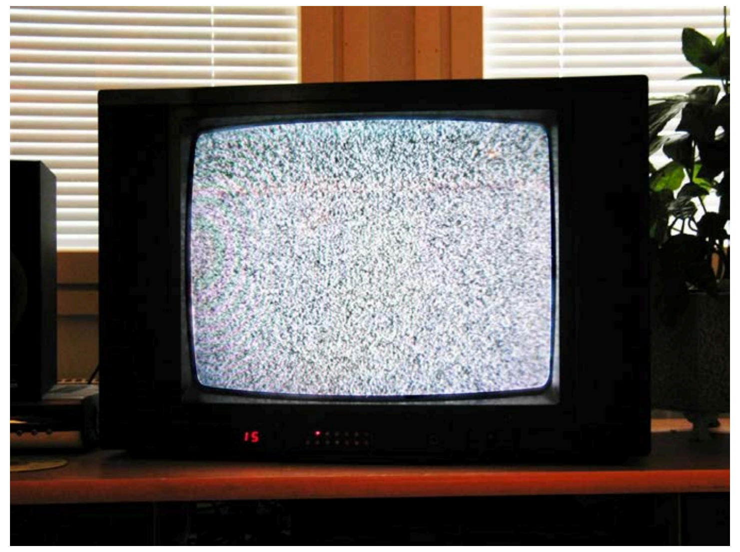
Thermodynamically "best": $\langle \Delta S^{\text{total}}_{\text{mod}} \rangle = 0$

Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

Harvesting Energy From Patterns



https://en.wikipedia.org/wiki/Harry_Potter_(character)



https://en.wikipedia.org/wiki/File:TV_noise.jpg

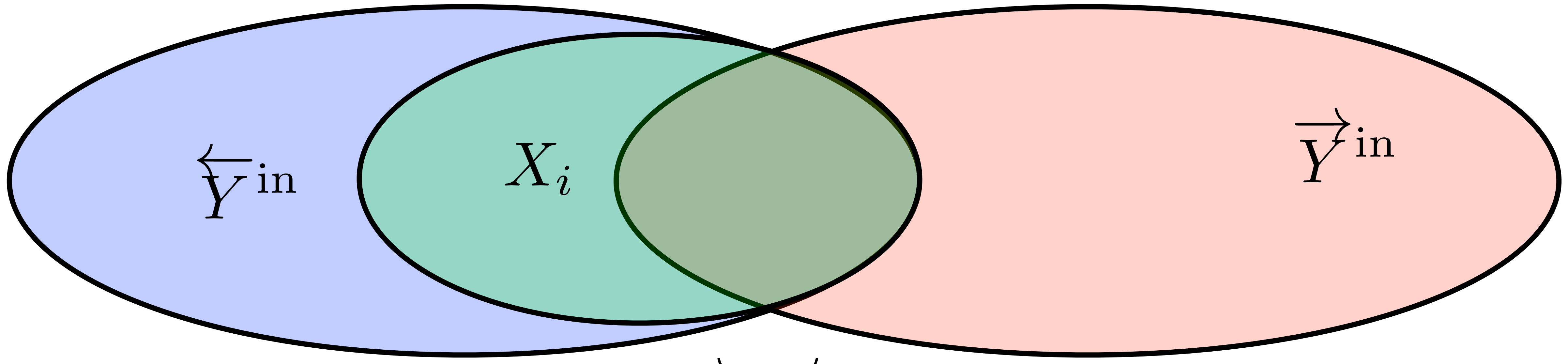
Outputs randomized and uncorrelated: $\Pr(Y^{\text{out}}_{a:b}) = \frac{1}{|\mathcal{Y}|^{b-a}}$

Input structured: $\Pr(Y^{\text{in}}_{a:b}) \neq \prod_{i=a}^{b-1} \Pr(Y^{\text{in}}_i)$

Remove dependence on outputs

$$\langle \Delta S^{\text{total}}_{\text{mod}} \rangle = k_B (I[X_i Y^{\text{in}}_i ; \overrightarrow{Y}^{\text{in}}_{i+1}] - I[X_{i+1} ; \overrightarrow{Y}^{\text{in}}_{i+1}])$$

Of the many ways of erasing information, which is best?



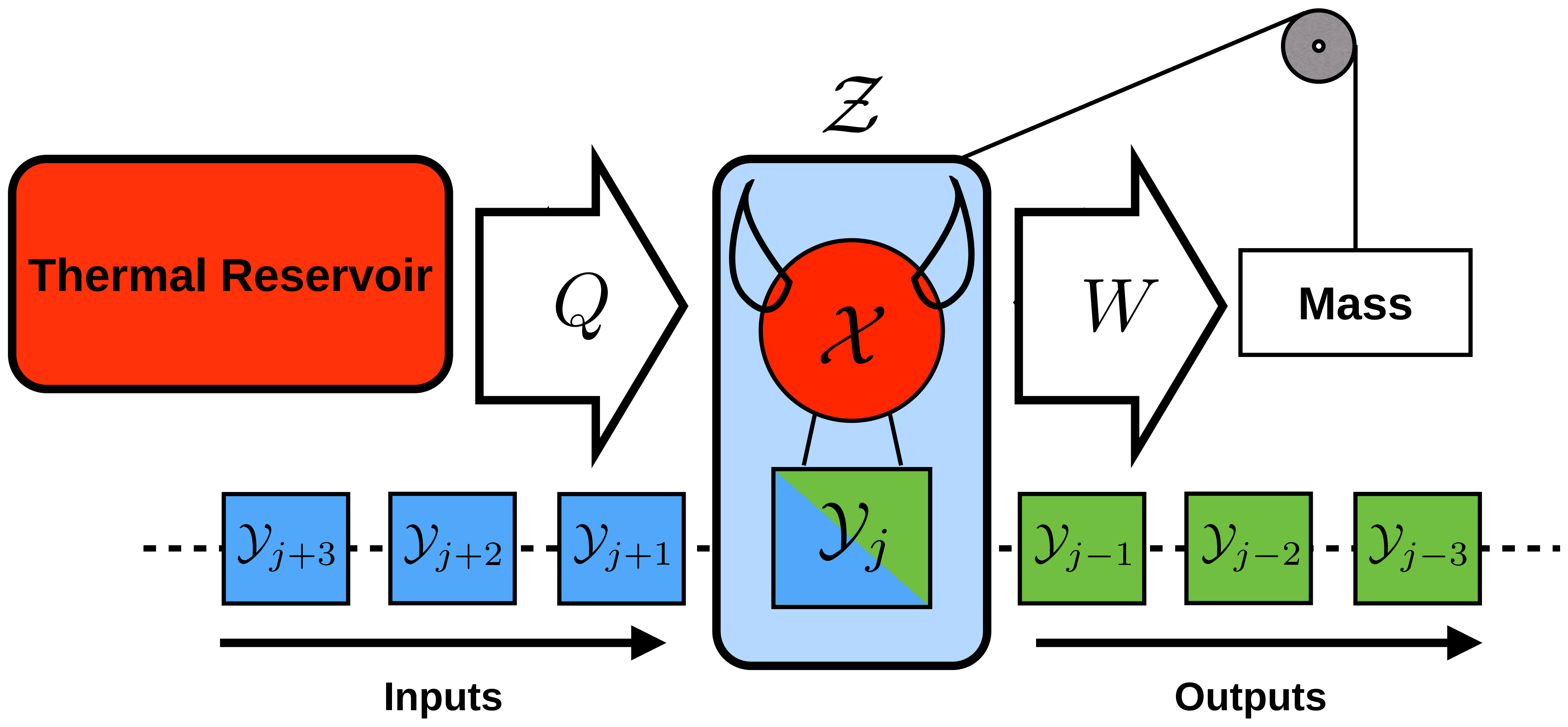
Nonanticipatory: $I[X_i ; \overrightarrow{Y}^{\text{in}} | \overleftarrow{Y}^{\text{in}}] = 0$

Thermodynamically "best": $\langle \Delta S^{\text{total}}_{\text{mod}} \rangle = 0$

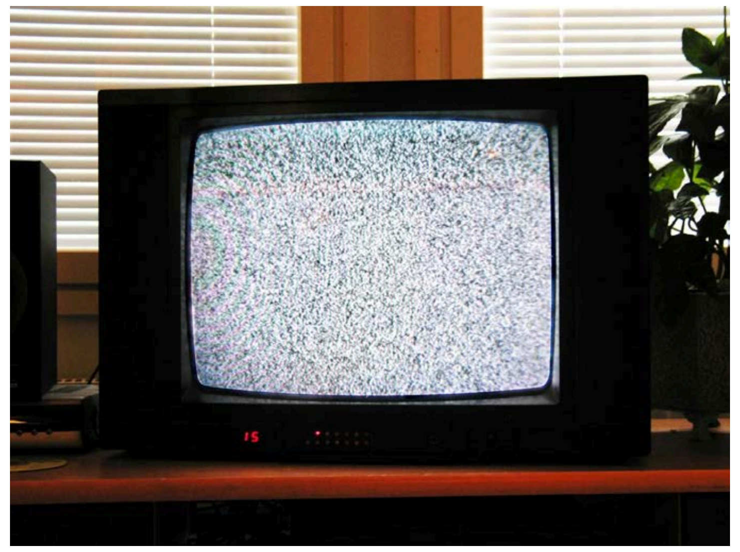
Implies $I[\overleftarrow{Y}^{\text{in}} ; \overrightarrow{Y}^{\text{in}} | X_i] = 0$

Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

Harvesting Energy From Patterns



https://en.wikipedia.org/wiki/Harry_Potter_(character)



https://en.wikipedia.org/wiki/File:TV_noise.jpg

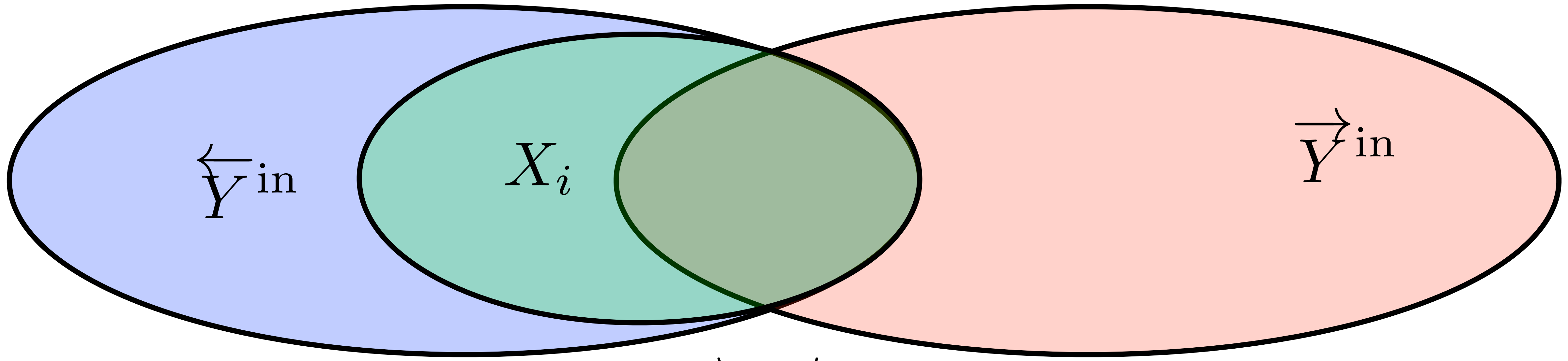
Outputs randomized and uncorrelated: $\Pr(Y^{\text{out}}_{a:b}) = \frac{1}{|\mathcal{Y}|^{b-a}}$

Input structured: $\Pr(Y^{\text{in}}_{a:b}) \neq \prod_{i=a}^{b-1} \Pr(Y^{\text{in}}_i)$

Remove dependence on outputs

$$\langle \Delta S^{\text{total}}_{\text{mod}} \rangle = k_B (I[X_i Y^{\text{in}}_i ; \overrightarrow{Y}^{\text{in}}_{i+1}] - I[X_{i+1}; \overrightarrow{Y}^{\text{in}}_{i+1}])$$

Of the many ways of erasing information, which is best?



Nonanticipatory: $I[X_i; \overrightarrow{Y}^{\text{in}} | \overleftarrow{Y}^{\text{in}}] = 0$

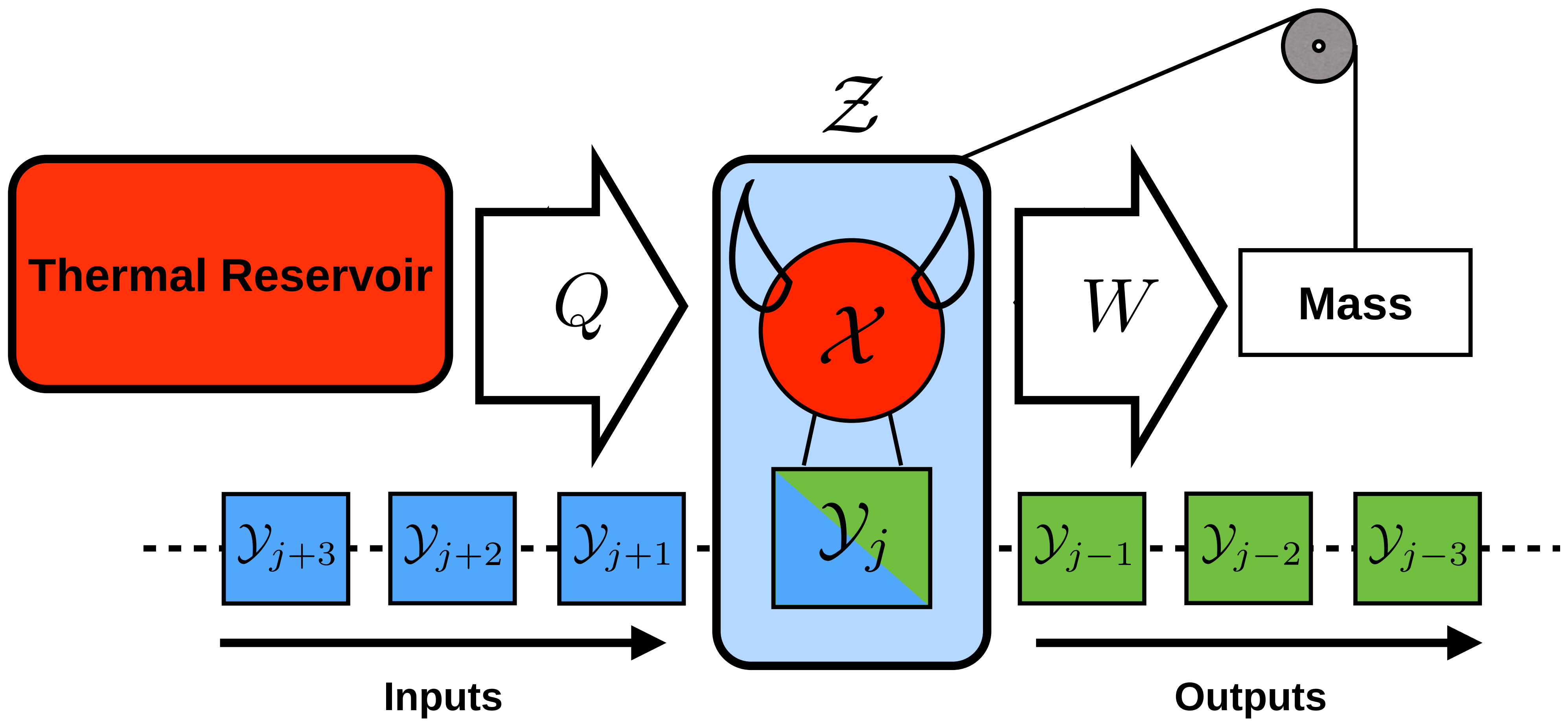
Thermodynamically "best": $\langle \Delta S^{\text{total}}_{\text{mod}} \rangle = 0$

Implies $I[\overleftarrow{Y}^{\text{in}}; \overrightarrow{Y}^{\text{in}} | X_i] = 0$

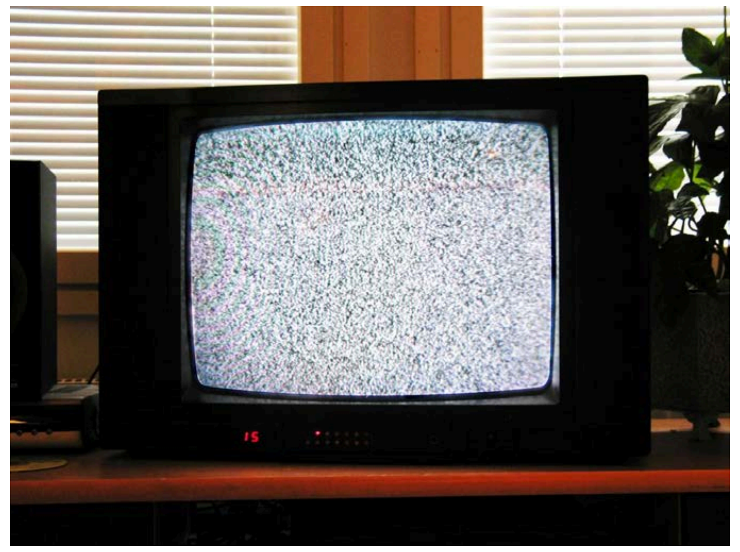
Minimum entropy production implies hidden state is predictive

Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

Harvesting Energy From Patterns



https://en.wikipedia.org/wiki/Harry_Potter_(character)



https://en.wikipedia.org/wiki/File:TV_noise.jpg

Memory is sufficient statistic $\overleftarrow{Y}^{\text{in}} \to \overrightarrow{Y}^{\text{in}}$

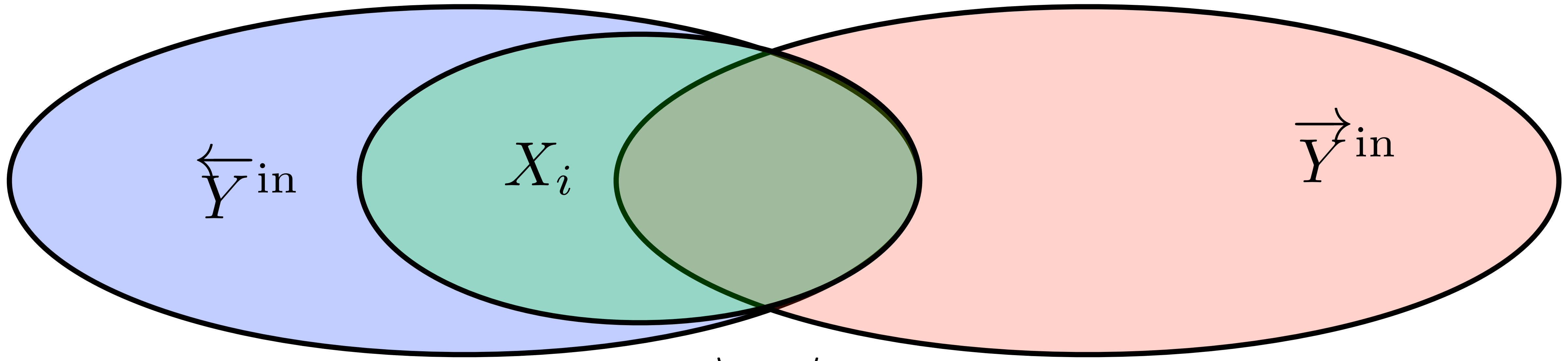
Outputs randomized and uncorrelated: $\Pr(Y^{\text{out}}_{a:b}) = \dfrac{1}{|\mathcal{Y}|^{b-a}}$

Input structured: $\Pr(Y^{\text{in}}_{a:b}) \neq \prod_{i=a}^{b-1} \Pr(Y^{\text{in}}_i)$

Remove dependence on outputs

$$\langle \Delta S^{\text{total}}_{\text{mod}} \rangle = k_B (I[X_i Y^{\text{in}}_i ; \overrightarrow{Y}^{\text{in}}_{i+1}] - I[X_{i+1}; \overrightarrow{Y}^{\text{in}}_{i+1}])$$

Of the many ways of erasing information, which is best?



Nonanticipatory: $I[X_i; \overrightarrow{Y}^{\text{in}} | \overleftarrow{Y}^{\text{in}}] = 0$

Thermodynamically "best": $\langle \Delta S^{\text{total}}_{\text{mod}} \rangle = 0$

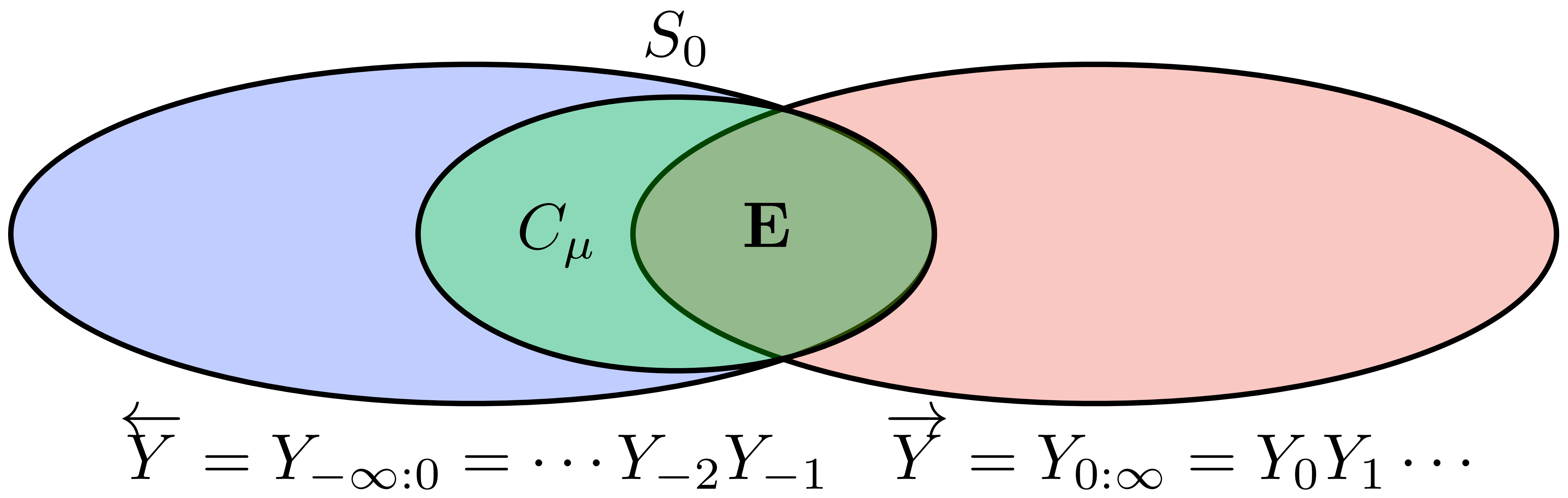
Implies $I[\overleftarrow{Y}^{\text{in}}; \overrightarrow{Y}^{\text{in}} | X_i] = 0$

Minimum entropy production implies hidden state is predictive

Boyd, Mandal, and Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

The Epsilon Machine

Statistical Complexity $C_\mu = \min_{\epsilon \ni I[\epsilon(\overleftarrow{Y});\overrightarrow{Y}]=I[\overleftarrow{Y};\overrightarrow{Y}]} H[\epsilon(\overleftarrow{Y})]$

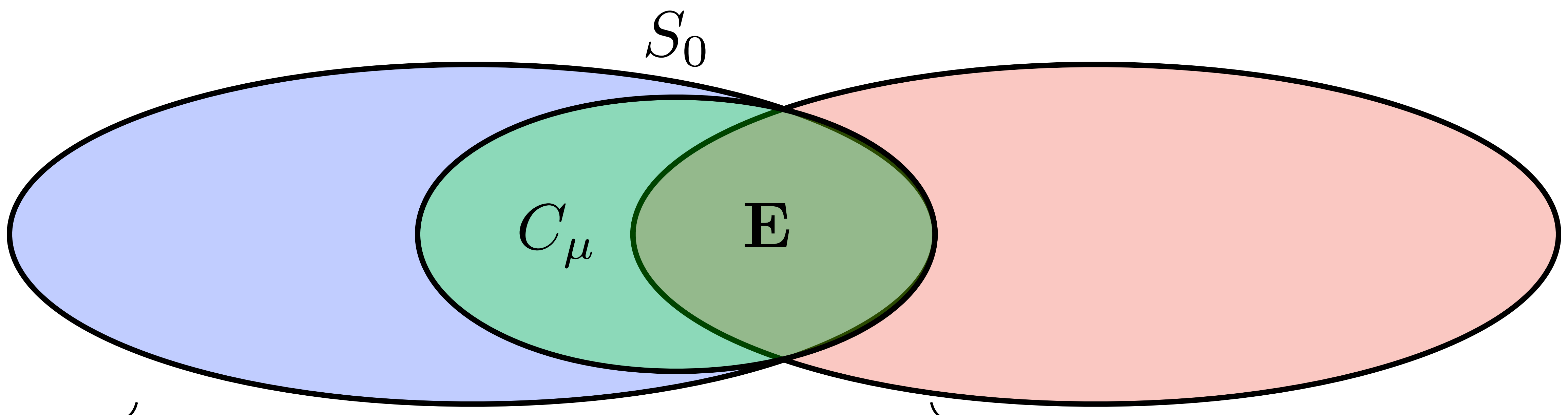


Epsilon-Machines:

Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

The Epsilon Machine

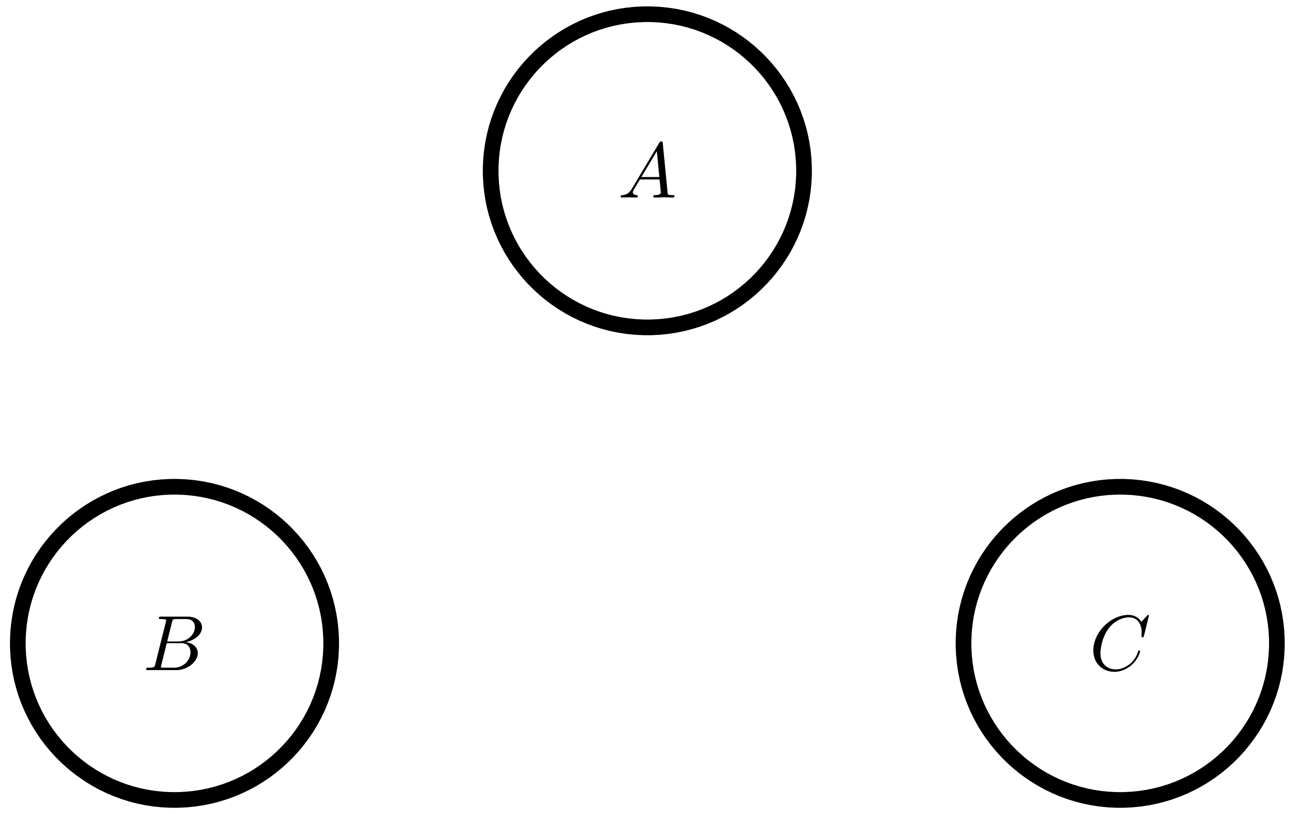
Statistical Complexity $C_\mu = \min_{\epsilon \ni I[\epsilon(\overleftarrow{Y});\overrightarrow{Y}] = I[\overleftarrow{Y};\overrightarrow{Y}]} H[\epsilon(\overleftarrow{Y})]$



$\overleftarrow{Y} = Y_{-\infty:0} = \cdots Y_{-2}Y_{-1}$ $\overrightarrow{Y} = Y_{0:\infty} = Y_0 Y_1 \cdots$

Epsilon-Machines:

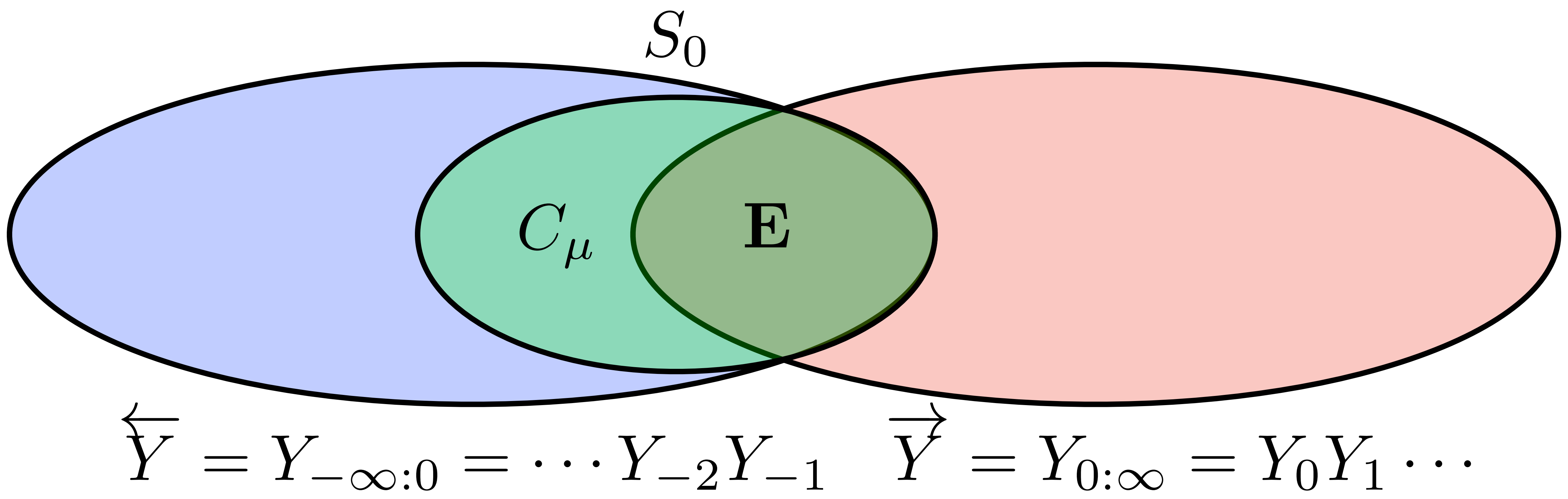
-Causal states $S_i = \epsilon(\overleftarrow{Y}_i)$



Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

The Epsilon Machine

Statistical Complexity $C_\mu = \min_{\epsilon \ni I[\epsilon(\overleftarrow{Y});\overrightarrow{Y}] = I[\overleftarrow{Y};\overrightarrow{Y}]} H[\epsilon(\overleftarrow{Y})]$

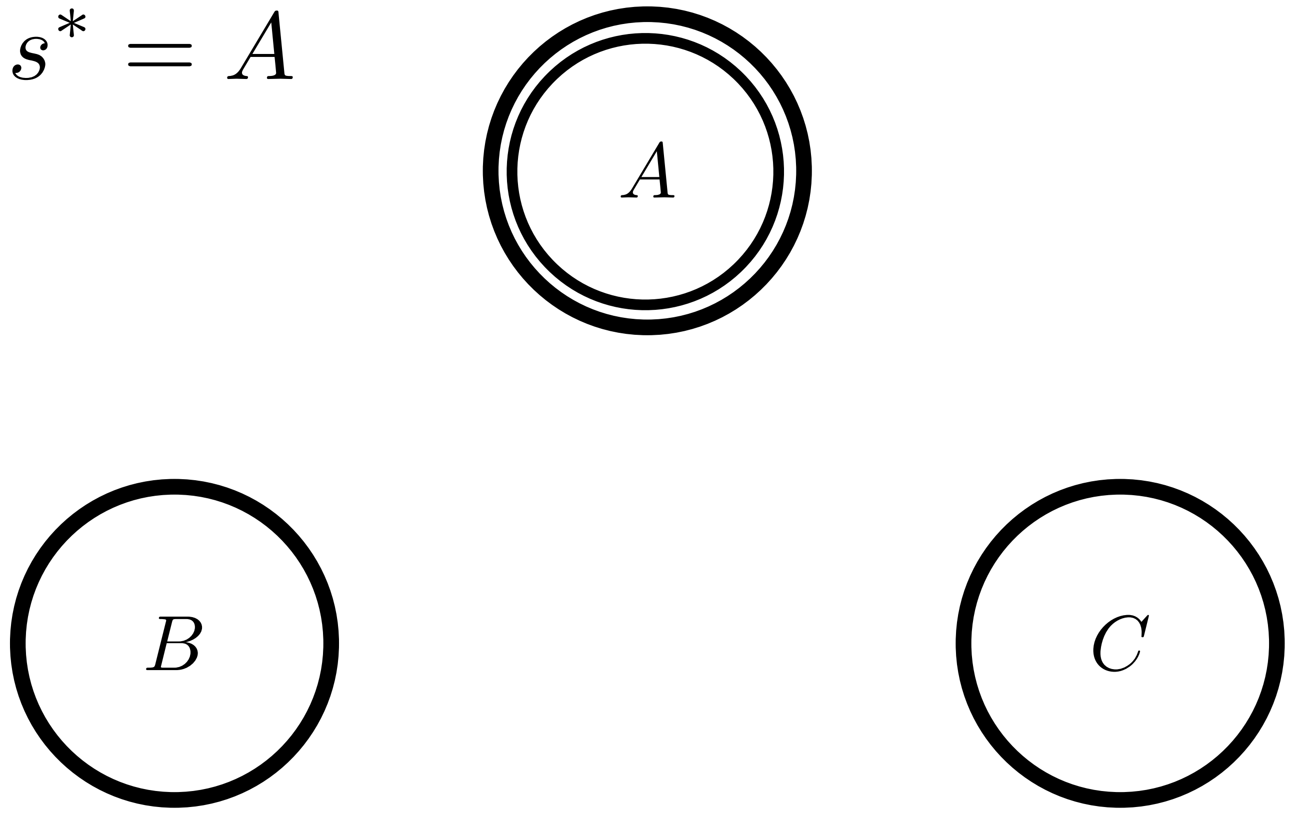


Epsilon-Machines:

-Causal states $S_i = \epsilon(\overleftarrow{Y}_i)$

-Start state $\Pr(S_0 = s^*) = 1$

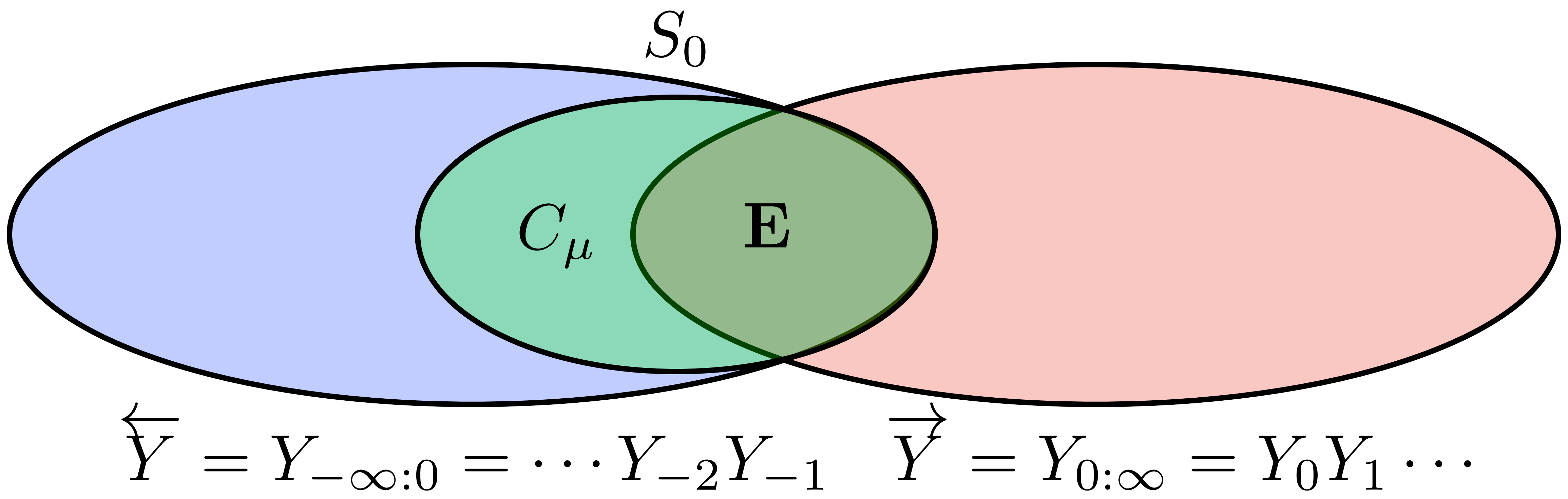
$s^* = A$



Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

The Epsilon Machine

Statistical Complexity $C_\mu = \min_{\epsilon \ni I[\epsilon(\overleftarrow{Y});\overrightarrow{Y}]=I[\overleftarrow{Y};\overrightarrow{Y}]} H[\epsilon(\overleftarrow{Y})]$

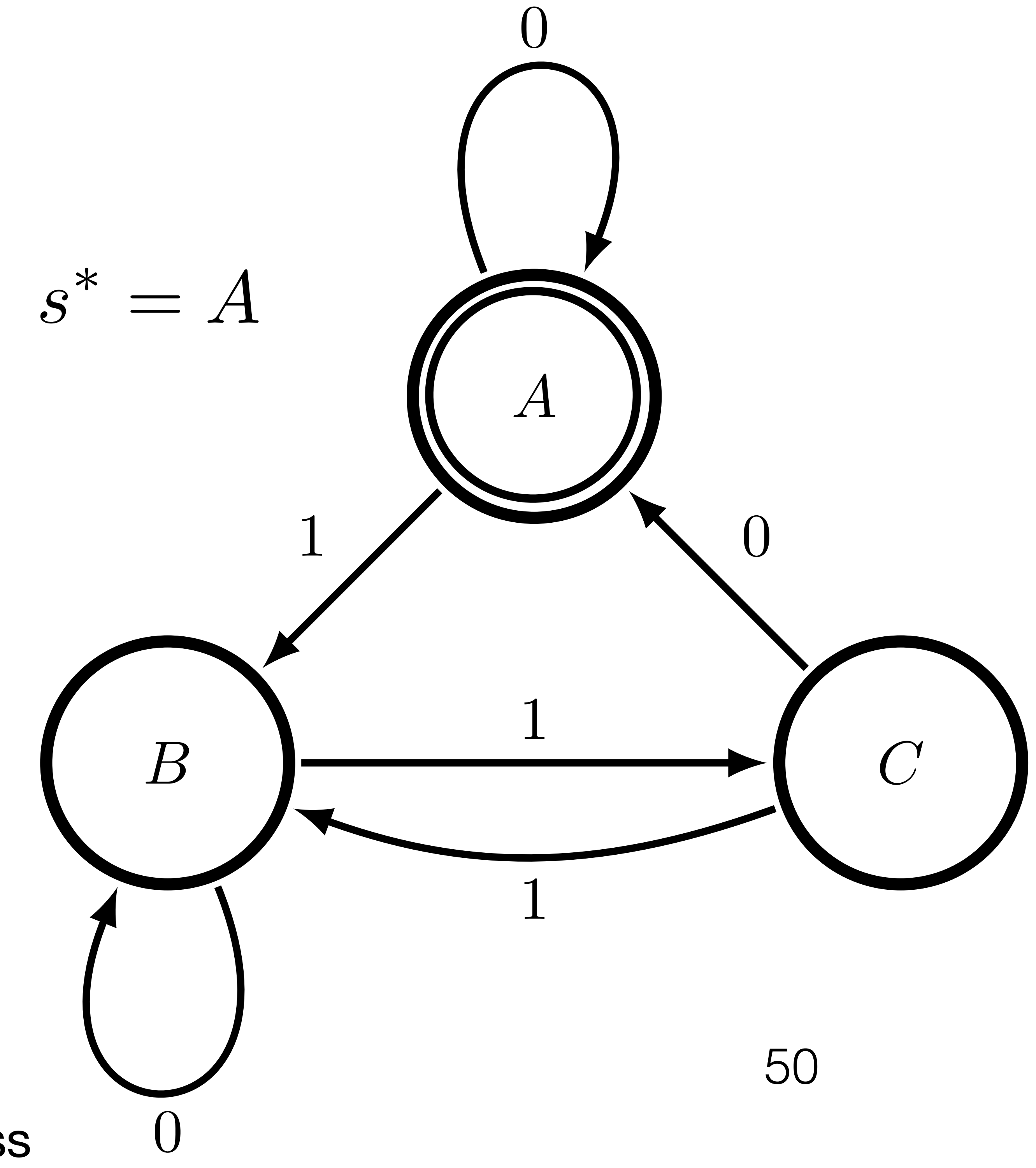


Epsilon-Machines:

-Causal states $S_i = \epsilon(\overleftarrow{Y}_i)$

-Start state $\Pr(S_0 = s^*) = 1$

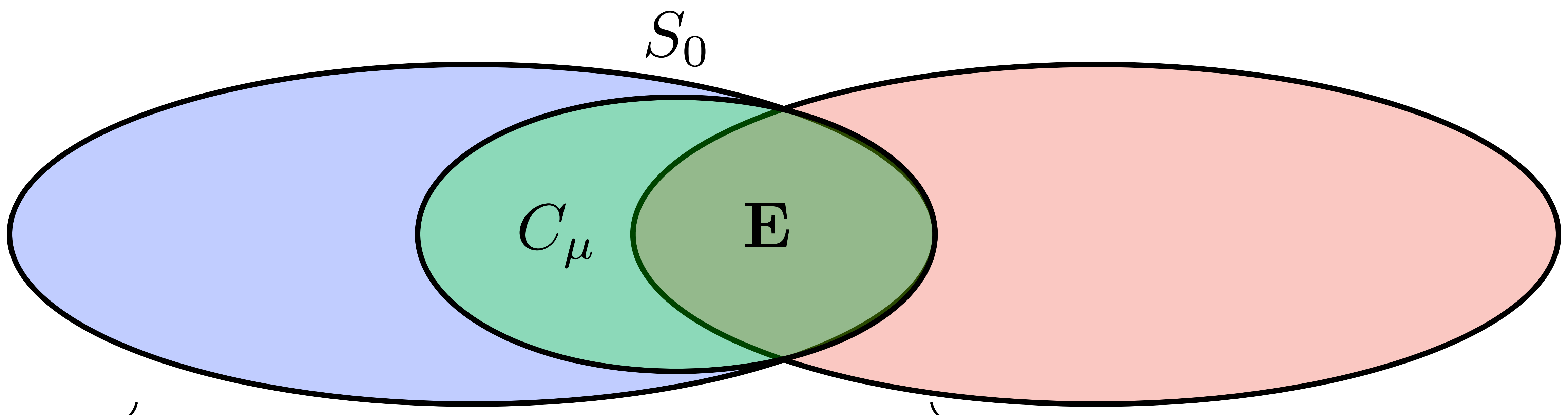
-Symbol-labeled transitions $S_{i+1} = \epsilon(S_i, Y_i)$
(Topology)



Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

The Epsilon Machine

Statistical Complexity $C_\mu = \min_{\epsilon \ni I[\epsilon(\overleftarrow{Y});\overrightarrow{Y}] = I[\overleftarrow{Y};\overrightarrow{Y}]} H[\epsilon(\overleftarrow{Y})]$



$\overleftarrow{Y} = Y_{-\infty:0} = \cdots Y_{-2}Y_{-1}$ $\overrightarrow{Y} = Y_{0:\infty} = Y_0 Y_1 \cdots$

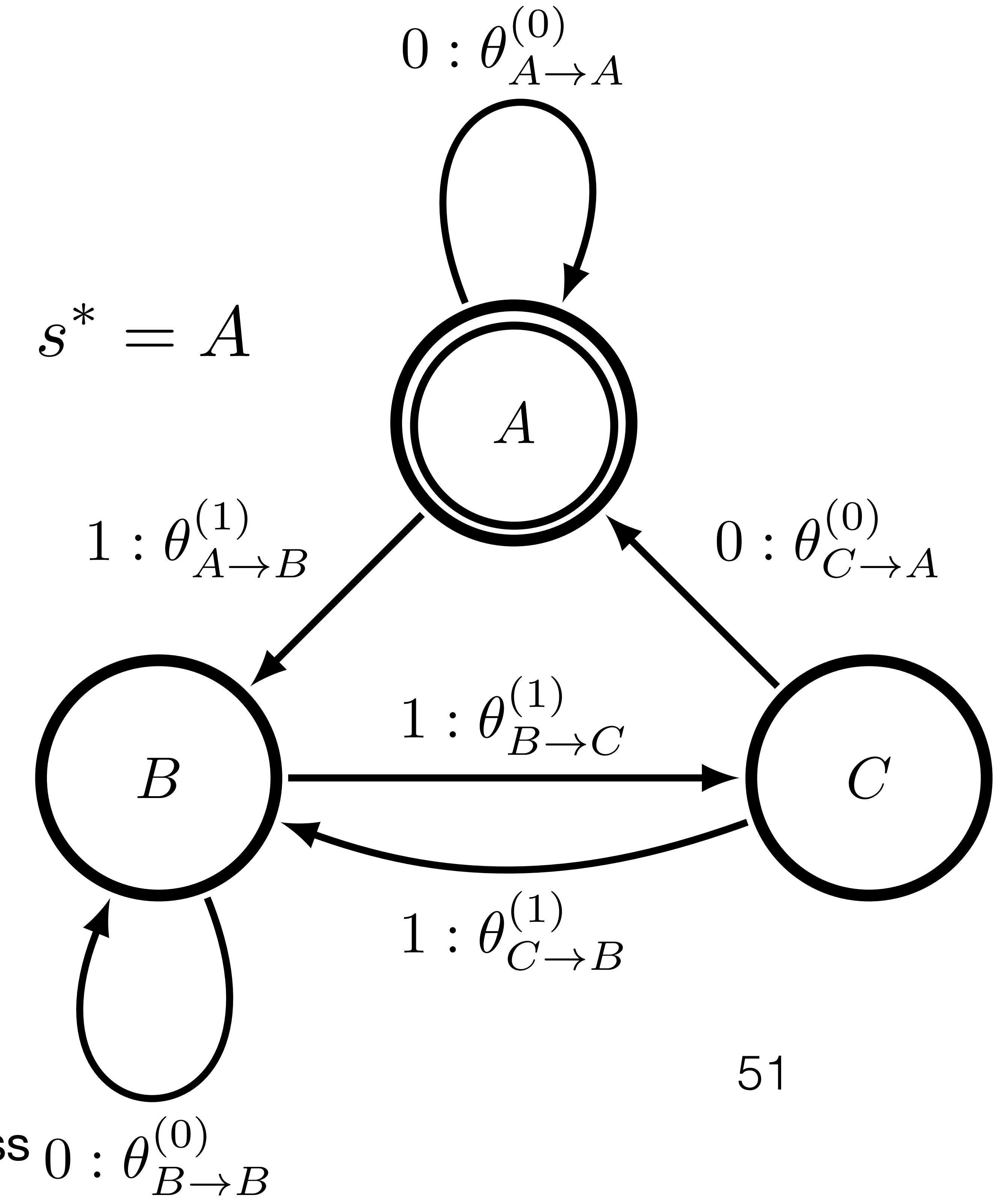
Epsilon-Machines:

-Causal states $S_i = \epsilon(\overleftarrow{Y}_i)$

-Start state $\Pr(S_0 = s^*) = 1$

-Symbol-labeled transitions $S_{i+1} = \epsilon(S_i, Y_i)$
(Topology)

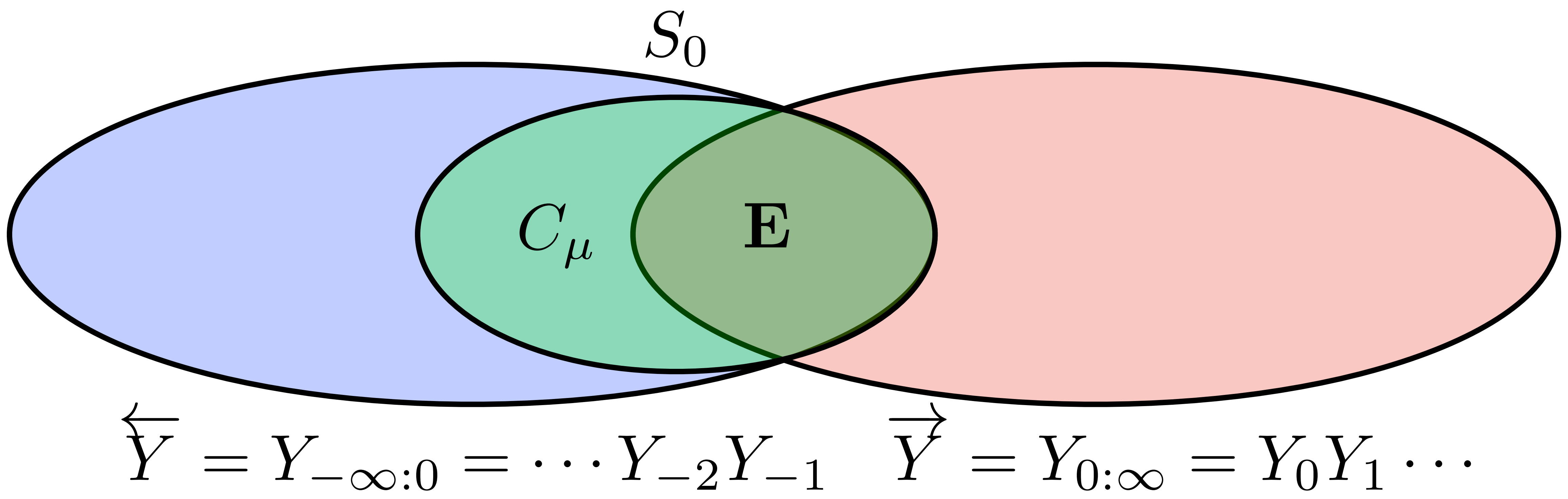
-Transition probabilities $\theta^{(y)}_{s \to s'} \equiv \Pr(S_{i+1} = s', Y_i = y | S_i = s)$



Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

The Epsilon Machine

Statistical Complexity $C_\mu = \min_{\epsilon \ni I[\epsilon(\overleftarrow{Y});\overrightarrow{Y}] = I[\overleftarrow{Y};\overrightarrow{Y}]} H[\epsilon(\overleftarrow{Y})]$



Epsilon-Machines:

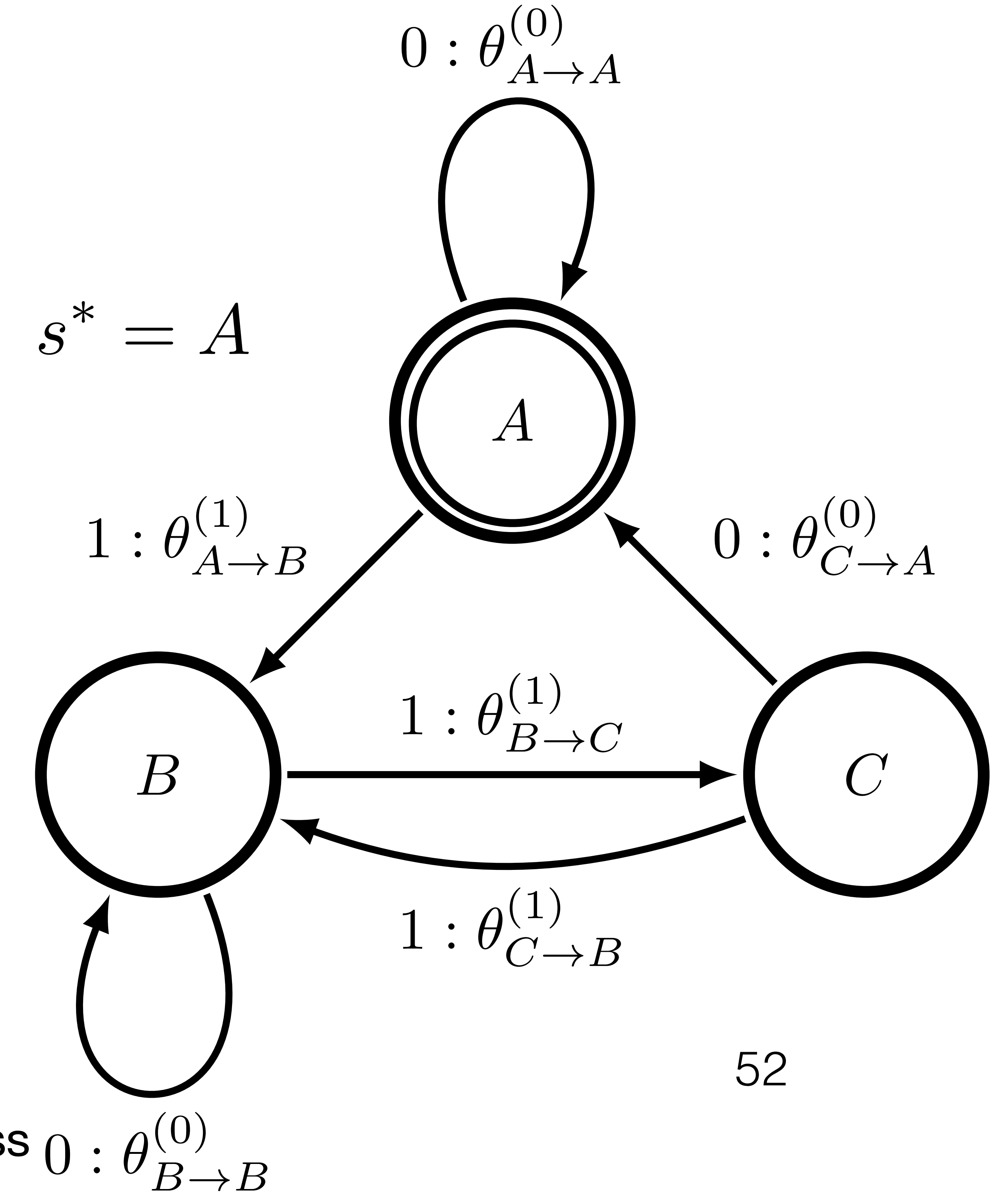
-Causal states $S_i = \epsilon(\overleftarrow{Y}_i)$

-Start state $\Pr(S_0 = s^*) = 1$

-Symbol-labeled transitions $S_{i+1} = \epsilon(S_i, Y_i)$
(Topology)

-Transition probabilities $\theta^{(y)}_{s \to s'} \equiv \Pr(S_{i+1} = s', Y_i = y | S_i = s)$

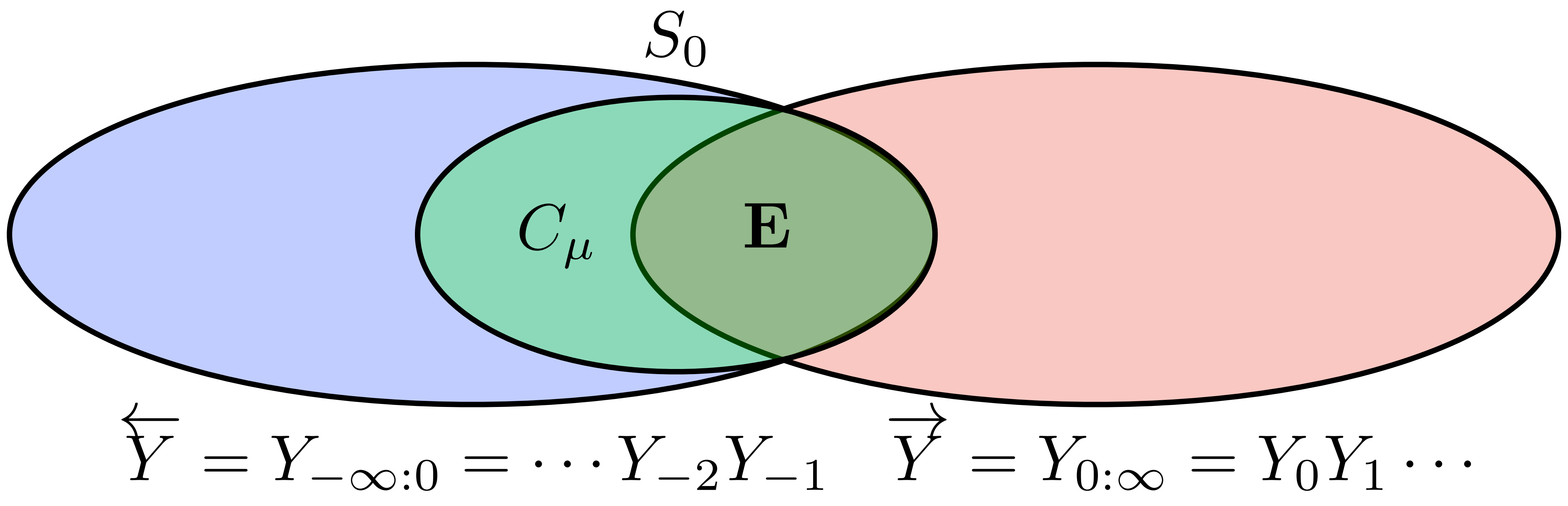
-Total Model Parameters $\theta = \{\mathcal{S}, \mathcal{Y}, s^*, \epsilon, \{\theta^{(y)}_{s \to s'}\}_{y \in \mathcal{Y}, s, s' \in \mathcal{S}}\}$



Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

The Epsilon Machine

Statistical Complexity $C_\mu = \min_{\epsilon \ni I[\epsilon(\overleftarrow{Y});\overrightarrow{Y}] = I[\overleftarrow{Y};\overrightarrow{Y}]} H[\epsilon(\overleftarrow{Y})]$



Epsilon-Machines:

-Causal states $S_i = \epsilon(\overleftarrow{Y}_i)$

-Start state $\Pr(S_0 = s^*) = 1$

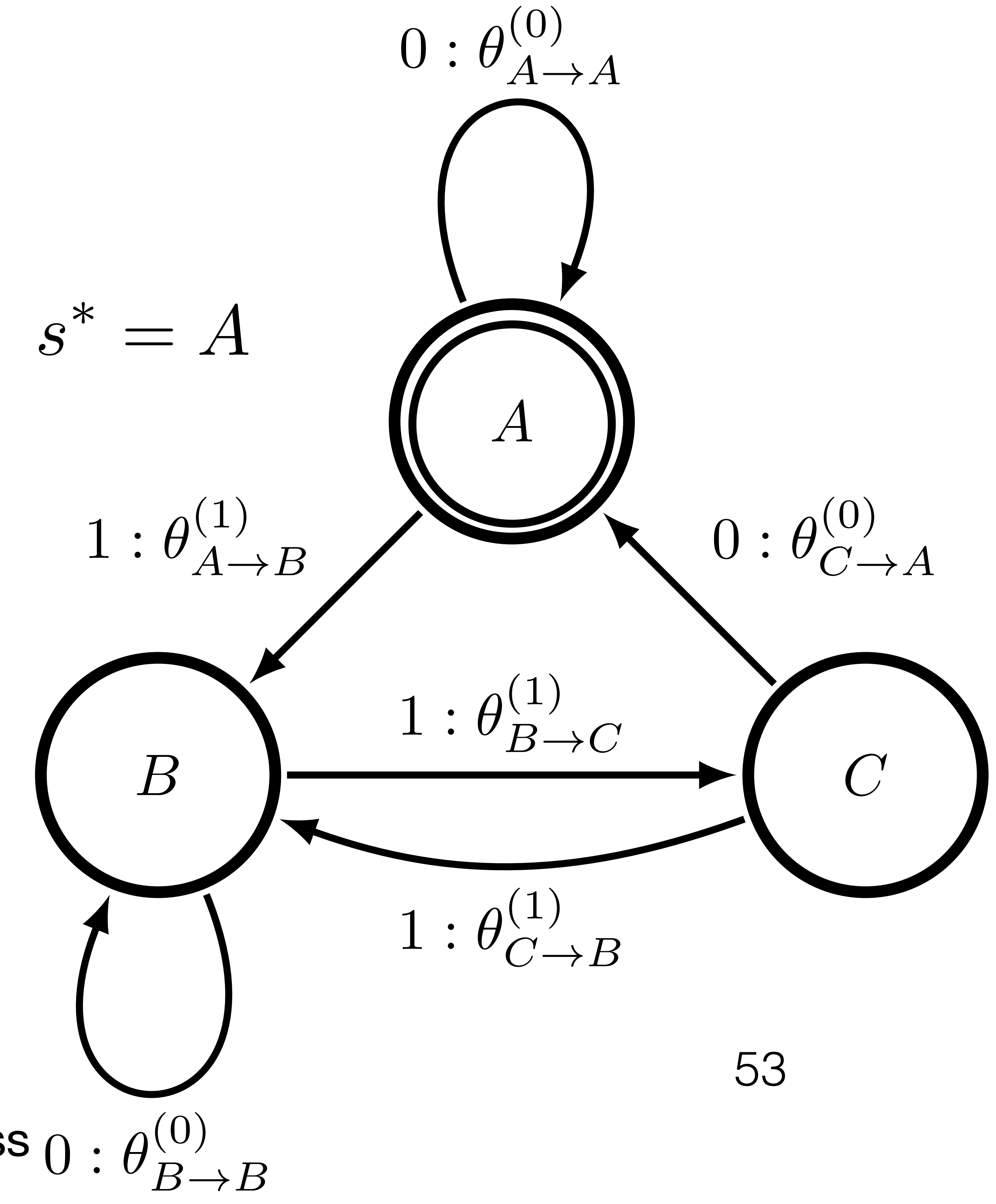
-Symbol-labeled transitions $S_{i+1} = \epsilon(S_i, Y_i)$
(Topology)

-Transition probabilities $\theta^{(y)}_{s \to s'} \equiv \Pr(S_{i+1} = s', Y_i = y | S_i = s)$

-Total Model Parameters $\theta = \{\mathcal{S}, \mathcal{Y}, s^*, \epsilon, \{\theta^{(y)}_{s\to s'}\}_{y\in\mathcal{Y}, s, s' \in \mathcal{S}}\}$

-Word probabilities

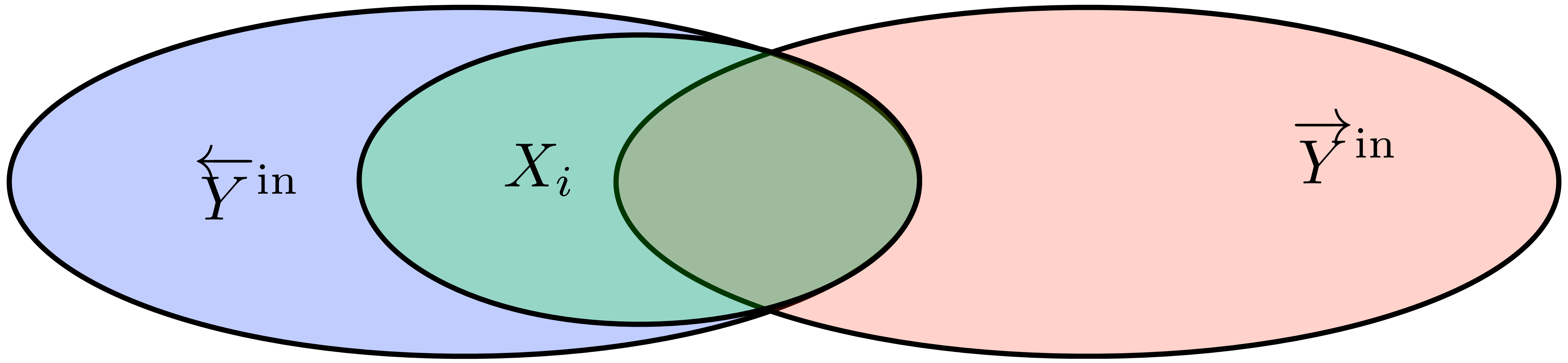
$$\Pr(Y^\theta_{0:L} = y_{0:L}) \equiv \Pr(Y_{0:L} = y_{0:L} | \Theta = \theta) = \prod_{i=0}^{L-1} \theta^{(y_i)}_{\epsilon(y_{0:i}) \to \epsilon(y_{i+1})}.$$



Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

Thermodynamics Implies Computational Mechanics

$\langle \Delta S^{\text{total}} \rangle = 0 \rightarrow$

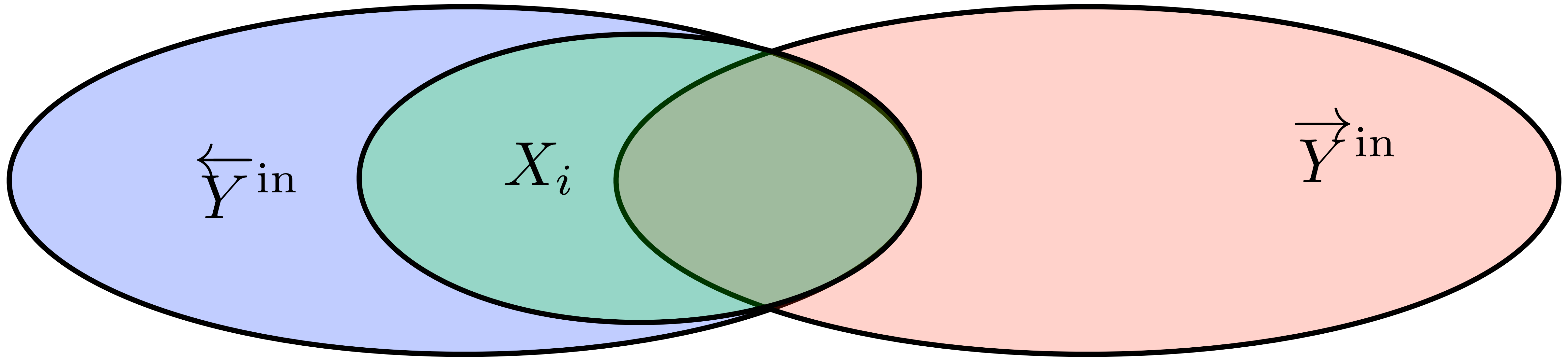


Thermodynamic efficiency requires that we predictively model our inputs.

Boyd, Alexander B., Dibyendu Mandal, and James P. Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

Thermodynamics Implies Computational Mechanics

$\langle \Delta S^{\text{total}} \rangle = 0 \rightarrow$



Thermodynamic efficiency requires that we predictively model our inputs.

Boyd, Alexander B., Dibyendu Mandal, and James P. Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

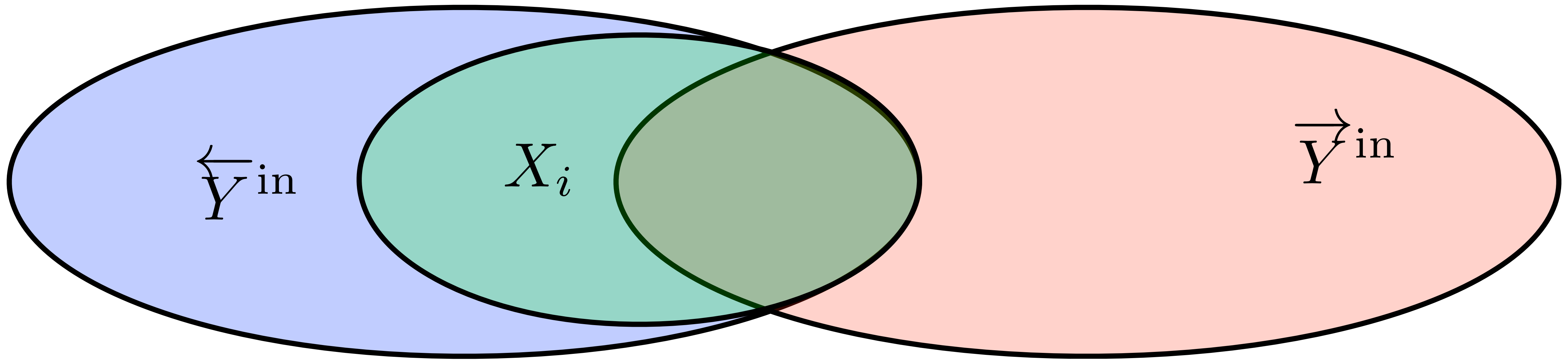
Principle of Requisite Complexity:

Information in memory must exceed predictive complexity of environment

$$H[X_i] \geq H[S_i] \equiv C_\mu^+$$

Thermodynamics implies Computational Mechanics

$\langle \Delta S^{\text{total}} \rangle = 0 \rightarrow$



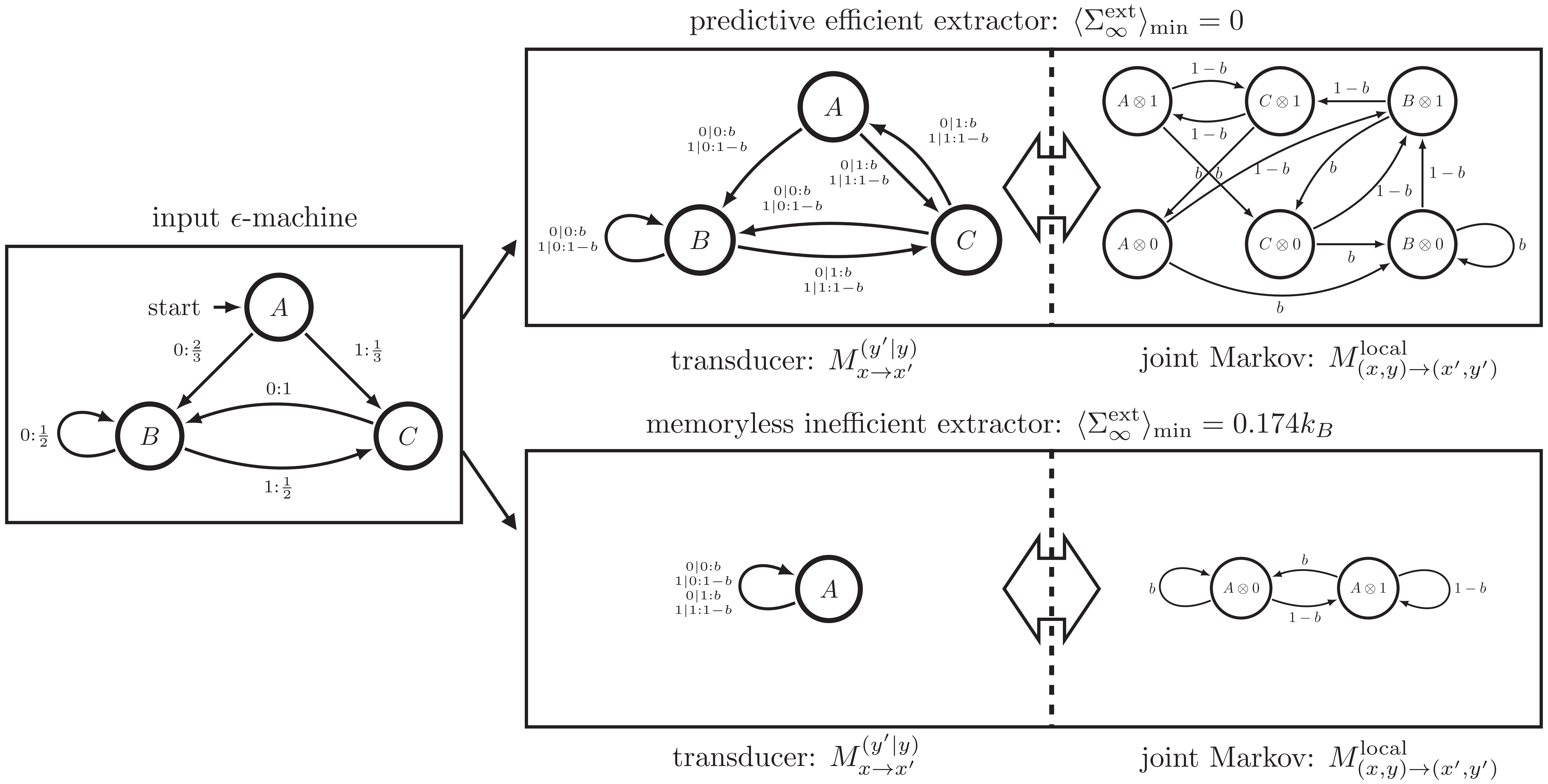
Thermodynamic efficiency requires that we predictively model our inputs.

Boyd, Alexander B., Dibyendu Mandal, and James P. Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

Principle of Requisite Complexity:

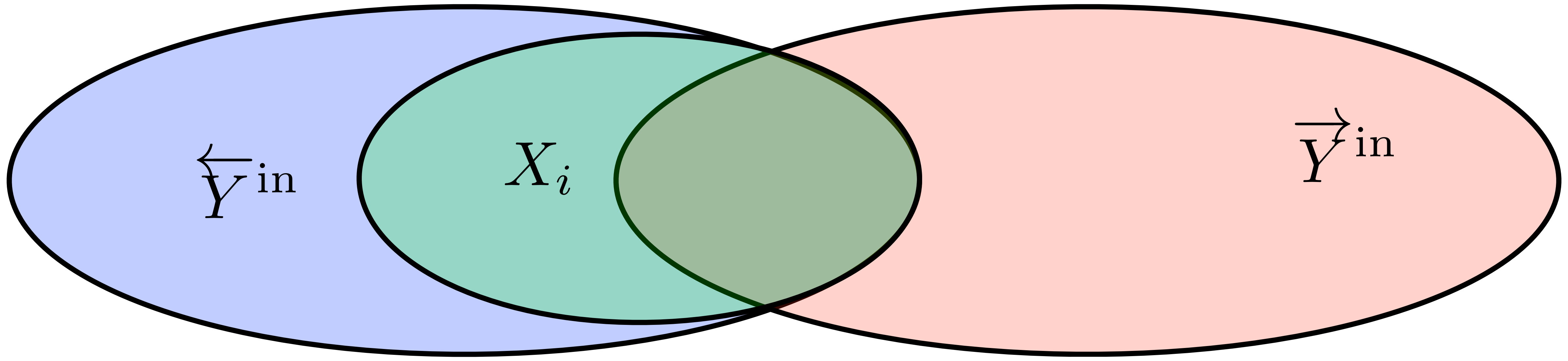
Information in memory must exceed predictive complexity of environment

$$H[X_i] \geq H[S_i] \equiv C_\mu^+$$



Thermodynamics implies Computational Mechanics

$\langle \Delta S^{\text{total}} \rangle = 0 \rightarrow$



Thermodynamic efficiency requires that we predictively model our inputs.

Boyd, Alexander B., Dibyendu Mandal, and James P. Crutchfield. "Thermodynamics of modularity: Structural costs beyond the Landauer bound." *Physical Review X* 8.3 (2018): 031036.

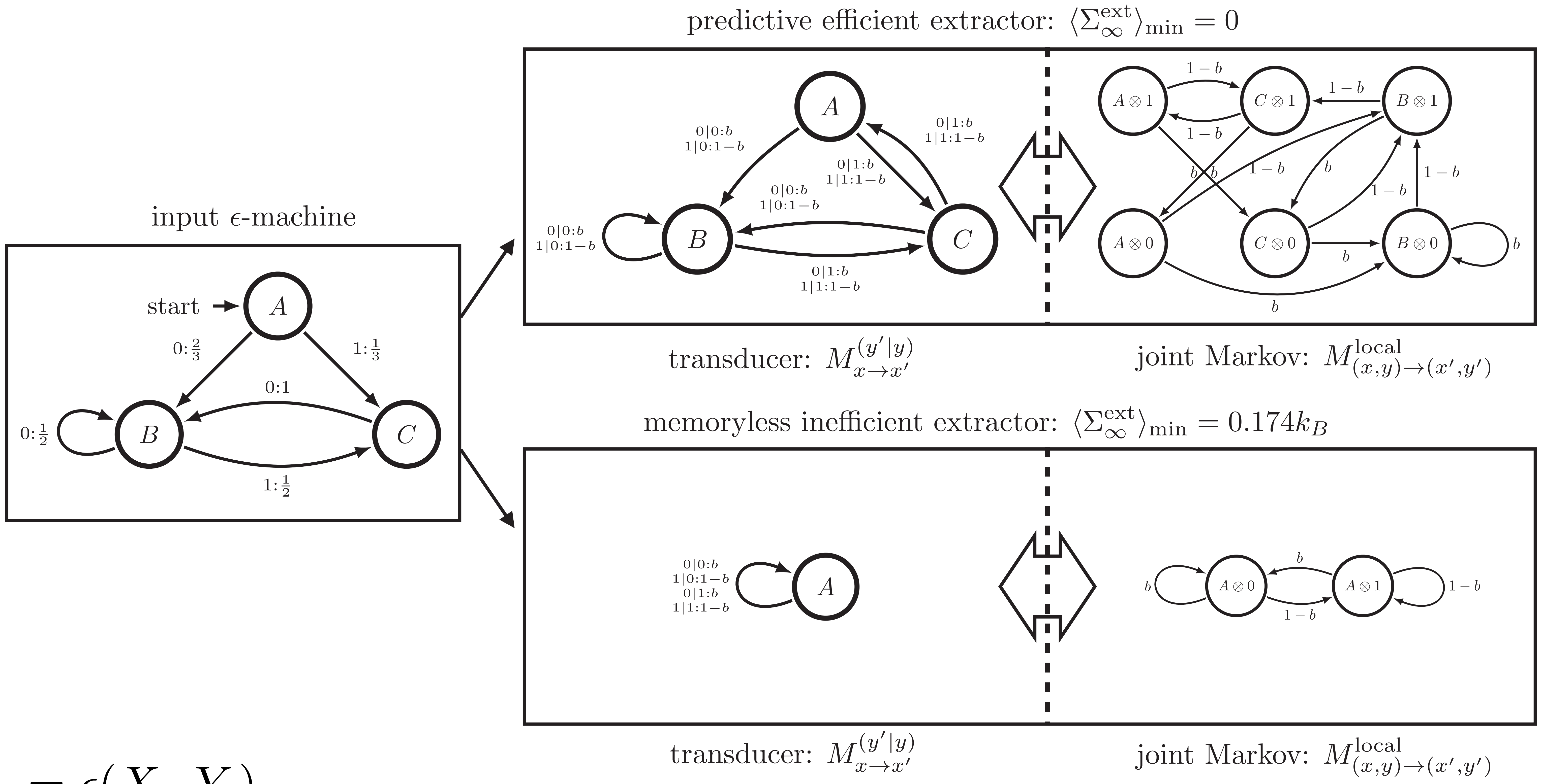
Principle of Requisite Complexity:

Information in memory must exceed predictive complexity of environment

$$H[X_i] \geq H[S_i] \equiv C_\mu^+$$

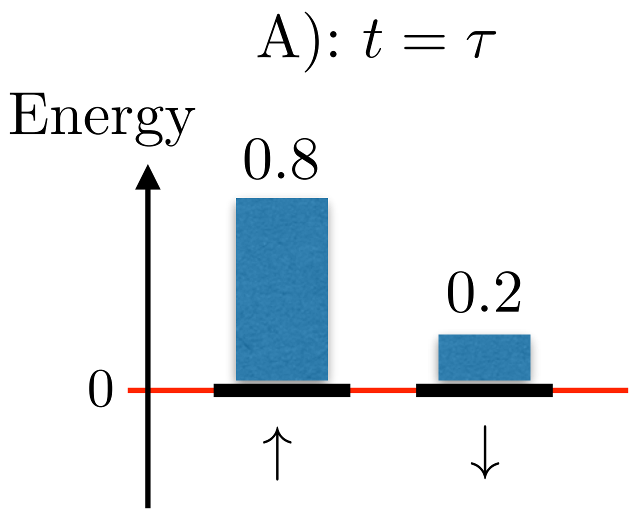
Thermodynamics determines minimal Topology and dynamics of hidden states

$$S_{i+1} = \epsilon(S_i, Y_i) \rightarrow X_{i+1} = \epsilon(X_i, Y_i)$$

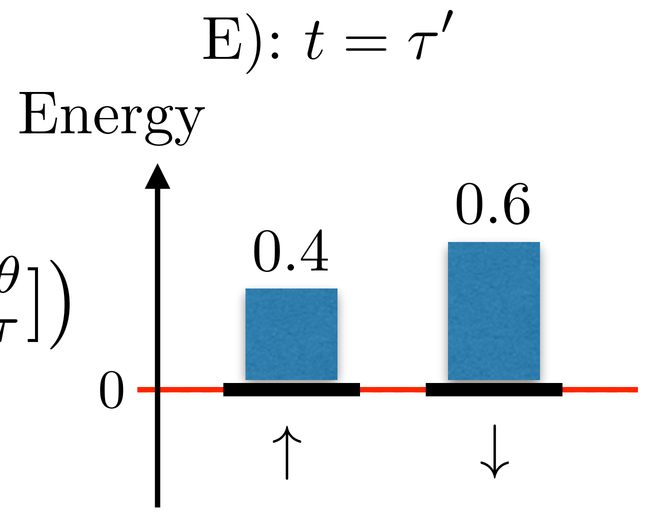


Efficient Information Engines

Energy landscape control must reflect estimated probabilities



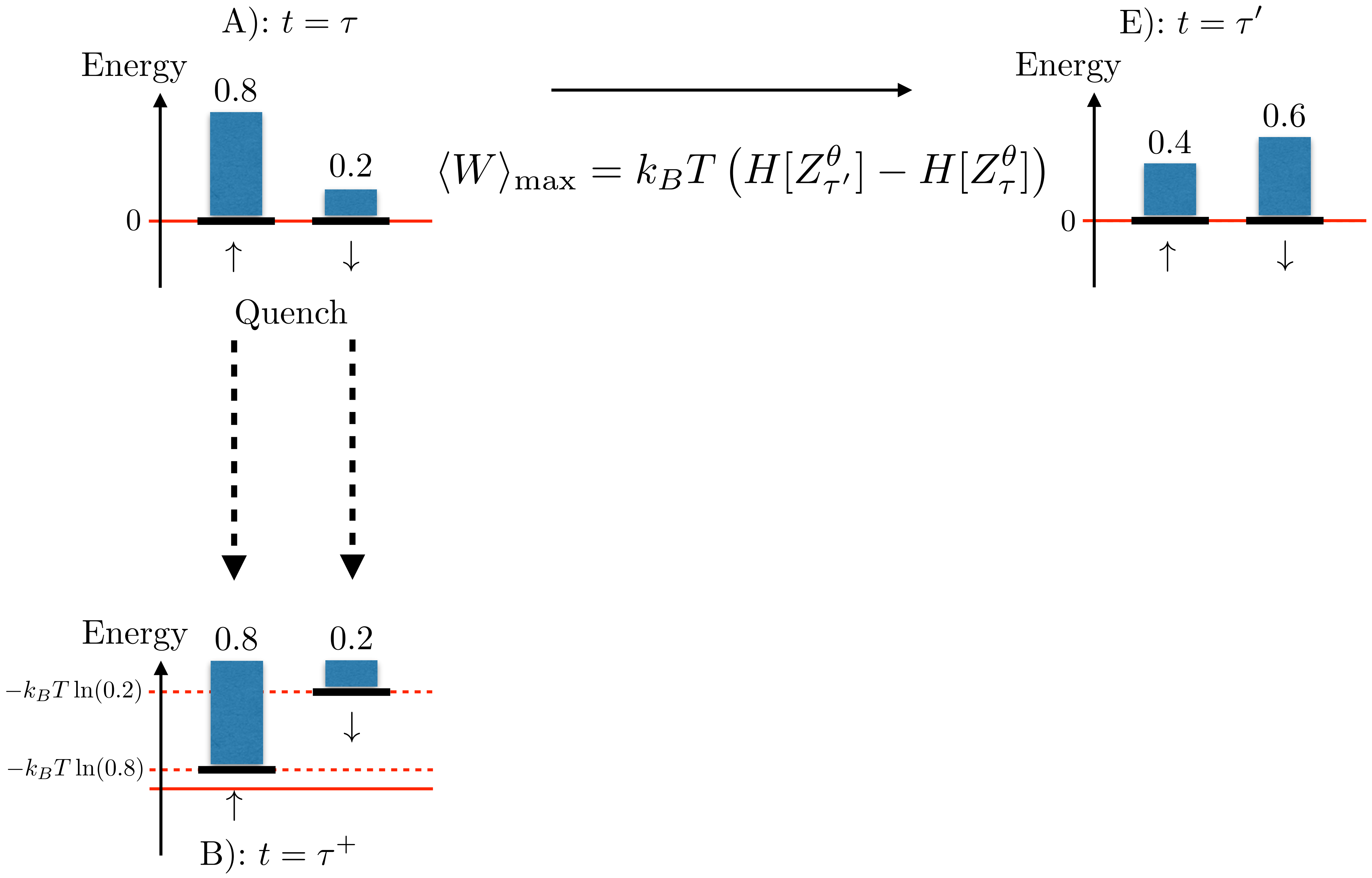
$$\langle W \rangle_{\max} = k_B T \left(H[Z^{\theta}_{\tau'}] - H[Z^{\theta}_{\tau}] \right)$$



Boyd, Alexander B., James P. Crutchfield, and Mile Gu. "Thermodynamic machine learning through maximum work production." *New Journal of Physics* 24.8 (2022): 083040.

Efficient Information Engines

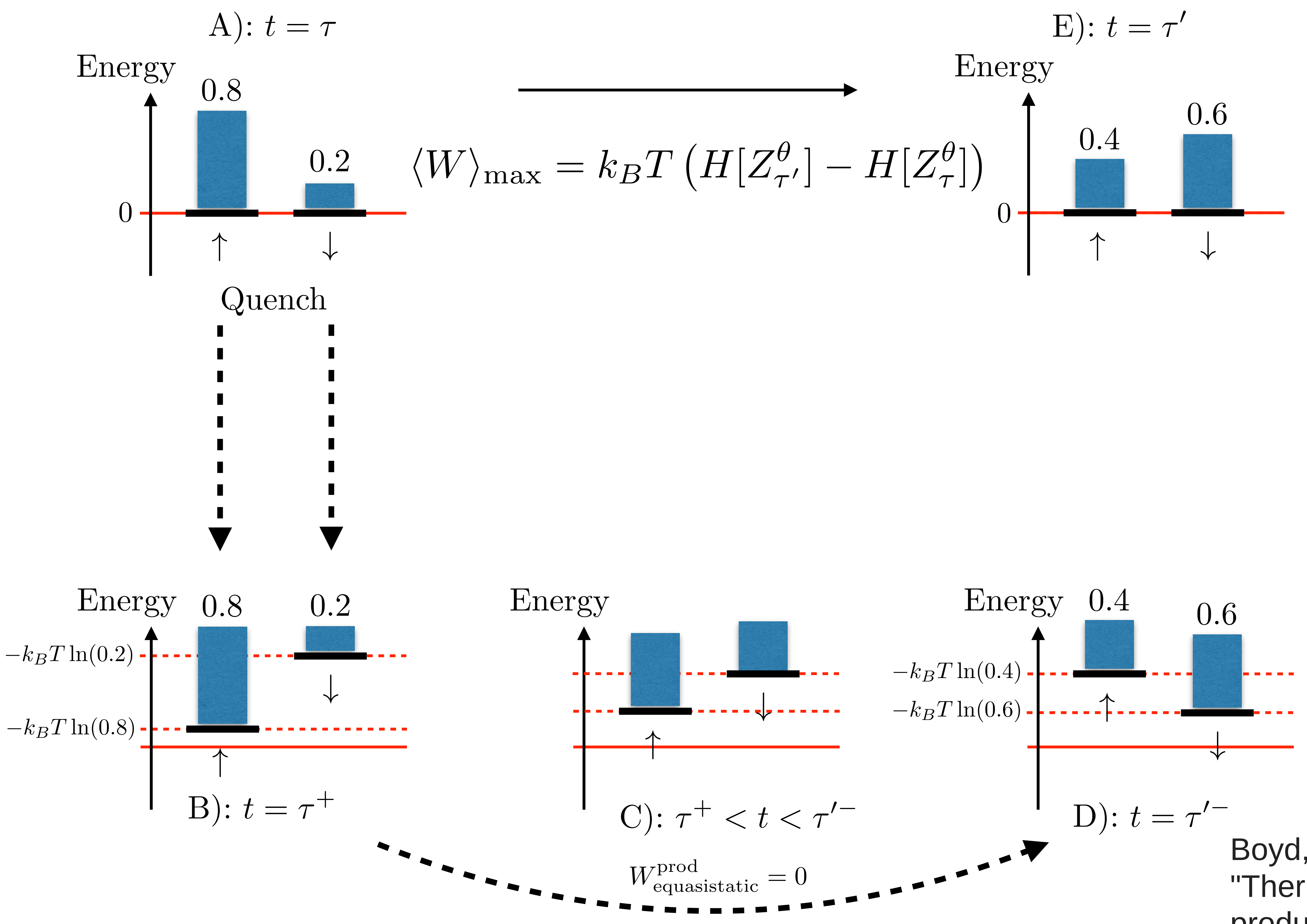
Energy landscape control must reflect estimated probabilities



Boyd, Alexander B., James P. Crutchfield, and Mile Gu. "Thermodynamic machine learning through maximum work production." *New Journal of Physics* 24.8 (2022): 083040.

Efficient Information Engines

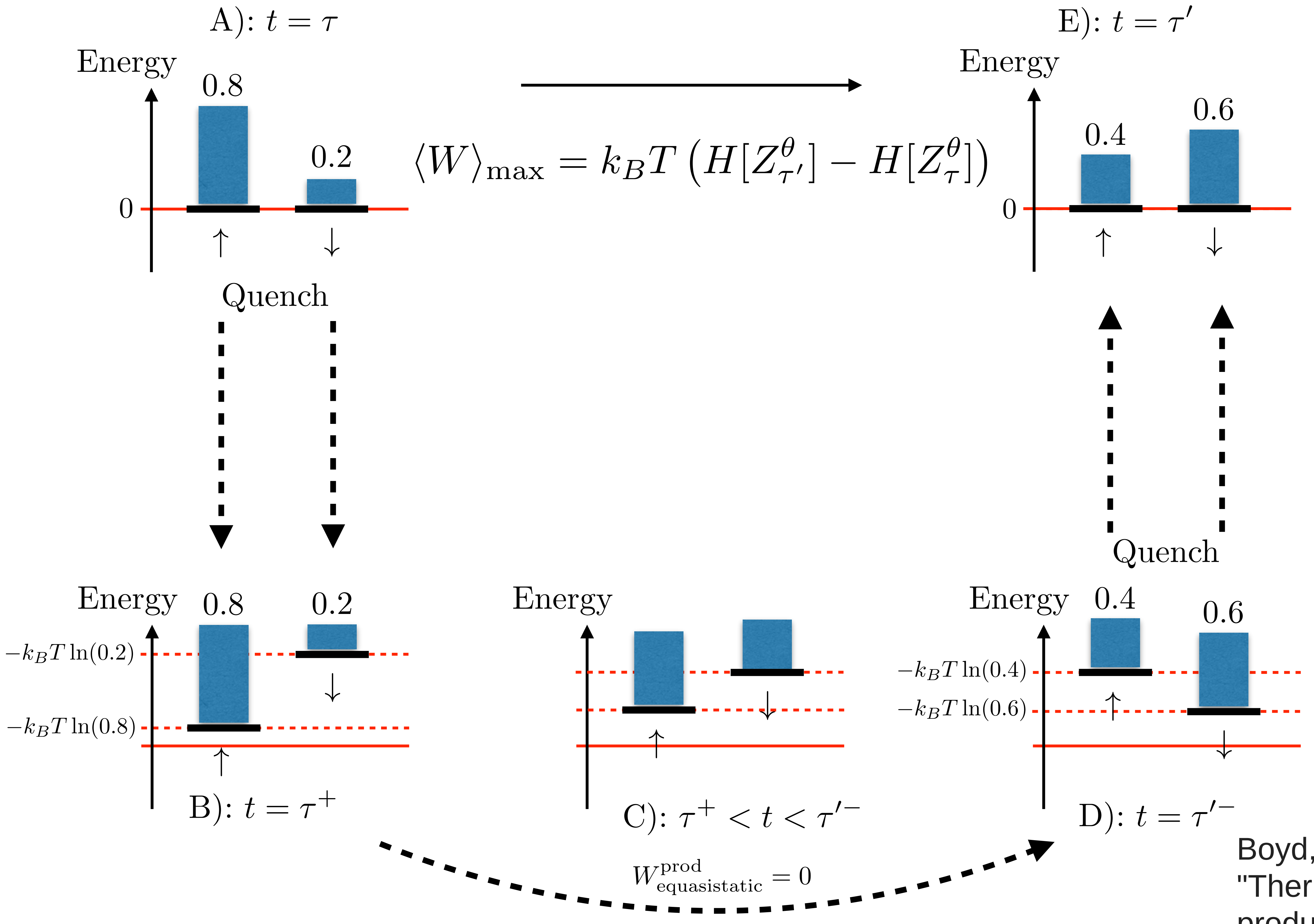
Energy landscape control must reflect estimated probabilities



Boyd, Alexander B., James P. Crutchfield, and Mile Gu. "Thermodynamic machine learning through maximum work production." *New Journal of Physics* 24.8 (2022): 083040.

Efficient Information Engines

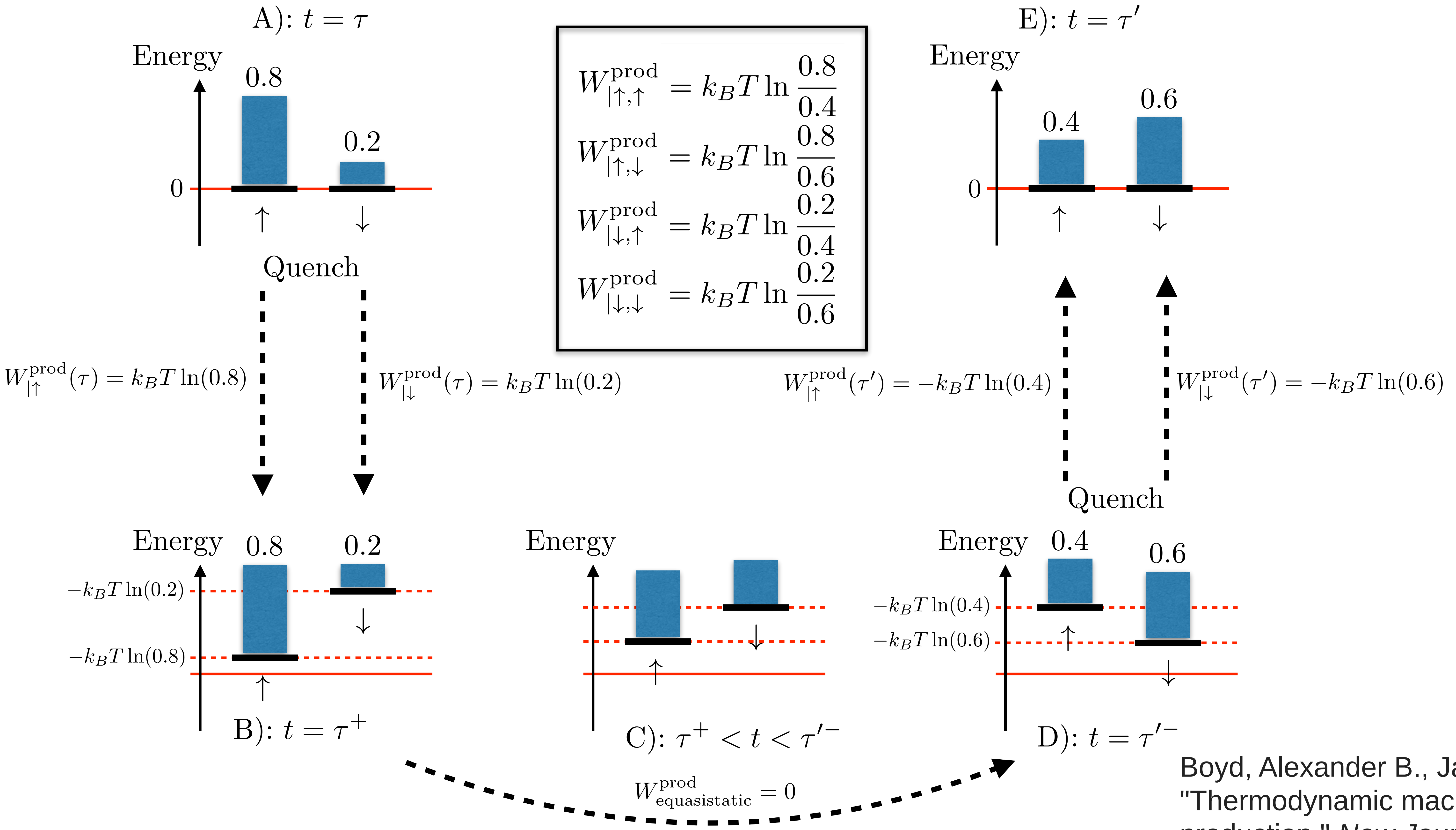
Energy landscape control must reflect estimated probabilities



Boyd, Alexander B., James P. Crutchfield, and Mile Gu. "Thermodynamic machine learning through maximum work production." *New Journal of Physics* 24.8 (2022): 083040.

Efficient Information Engines

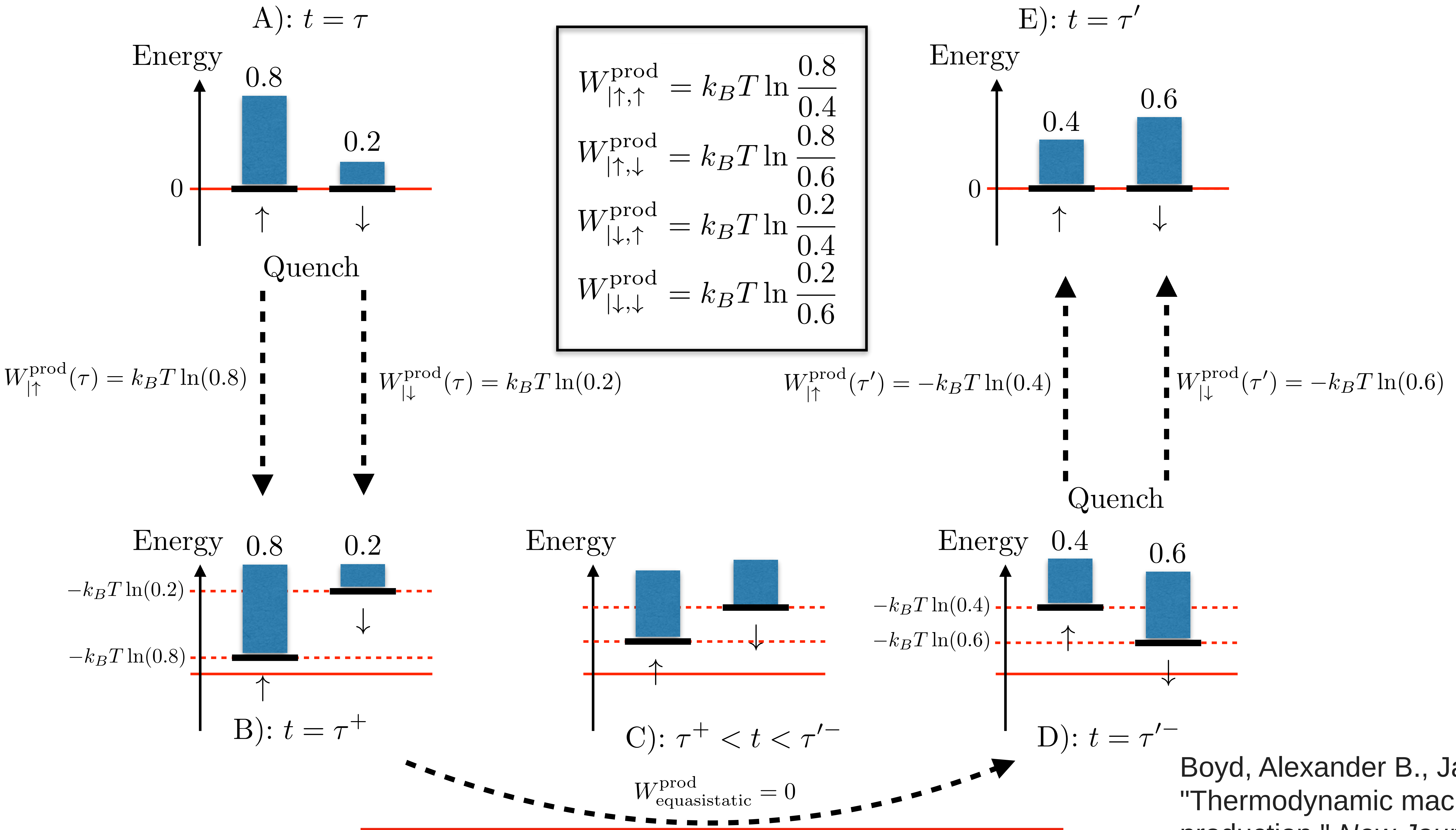
Energy landscape control must reflect estimated probabilities



Boyd, Alexander B., James P. Crutchfield, and Mile Gu. "Thermodynamic machine learning through maximum work production." *New Journal of Physics* 24.8 (2022): 083040.

Efficient Information Engines

Energy landscape control must reflect estimated probabilities.
Engine design and energies determined by model.



Efficient engines: $$\langle W^{\theta}_{z_\tau \to z_{\tau'}} \rangle = k_B T \ln \frac{\Pr(Z^{\theta}_{\tau} = z_\tau)}{\Pr(Z^{\theta'}_{\tau} = z_{\tau'})}$$

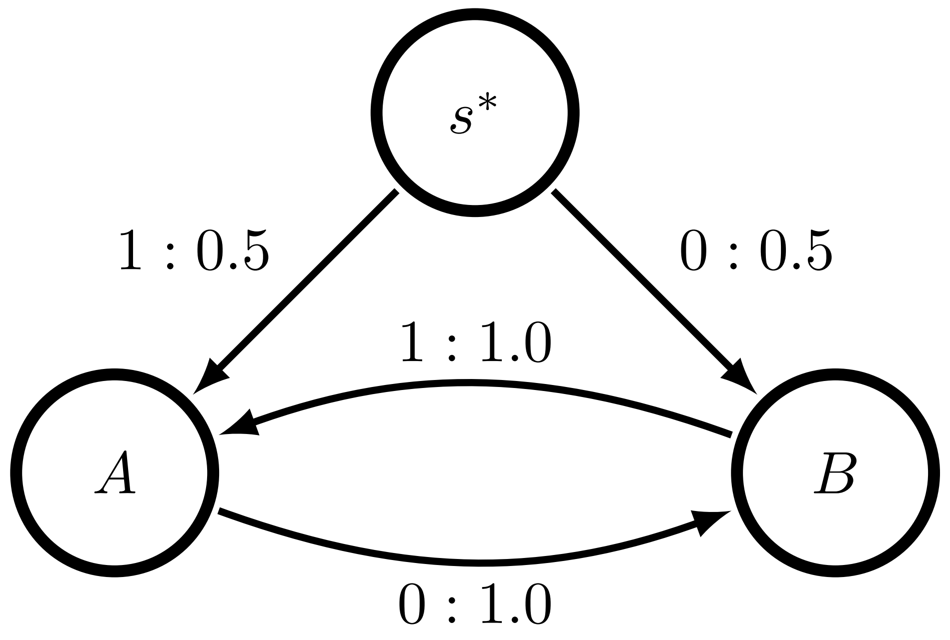
Boyd, Alexander B., James P. Crutchfield, and Mile Gu. "Thermodynamic machine learning through maximum work production." *New Journal of Physics* 24.8 (2022): 083040.

Efficient Engine — Model Equivalence

Estimated Process

$$\Pr(Y^{\theta}_{0:\infty} = 101010\cdots) = 0.5$$
$$\Pr(Y^{\theta}_{0:\infty} = 010101\cdots) = 0.5$$

Estimated ϵ-machine (model)



$$\Pr(Y^{\theta}_{0:L} = y_{0:L}) \equiv \sum_{s_{0:L+1}} \delta_{s_0,s^*} \prod_{i=0}^{L-1} \theta^{(y_i)}_{s_i \to s_{i+1}}$$

$$\theta^{(y)}_{s \to s'} = \Pr(S_{i+1} = s', Y_i^P = y | S_i = s)$$

Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

Boyd, Alexander B., James P. Crutchfield, and Mile Gu. "Thermodynamic machine learning through maximum work production." *New Journal of Physics* 24.8 (2022): 083040.

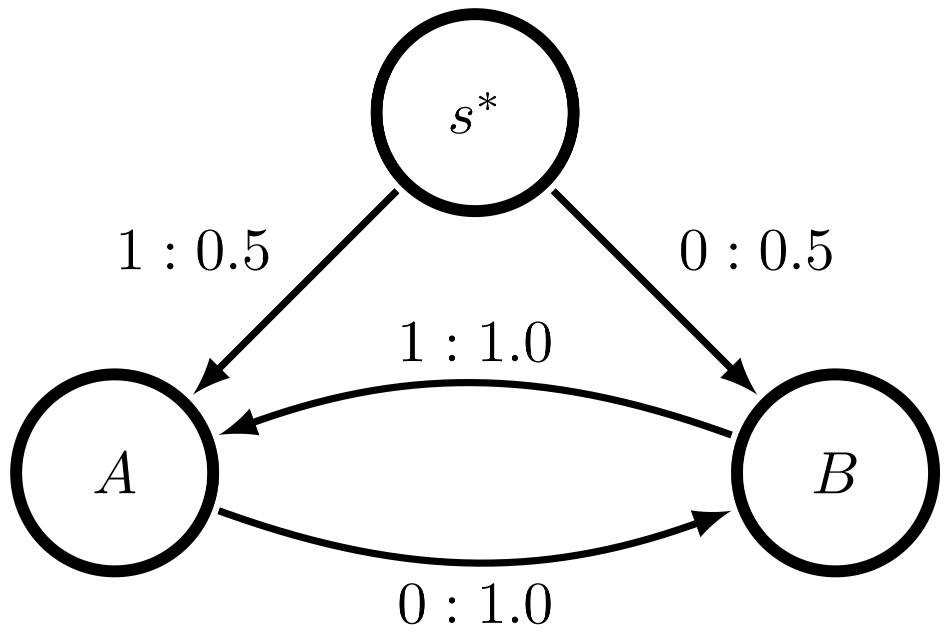
Efficient Engine — Model Equivalence

Estimated Process

$$\Pr(Y^{\theta}_{0:\infty} = 101010\cdots) = 0.5$$
$$\Pr(Y^{\theta}_{0:\infty} = 010101\cdots) = 0.5$$

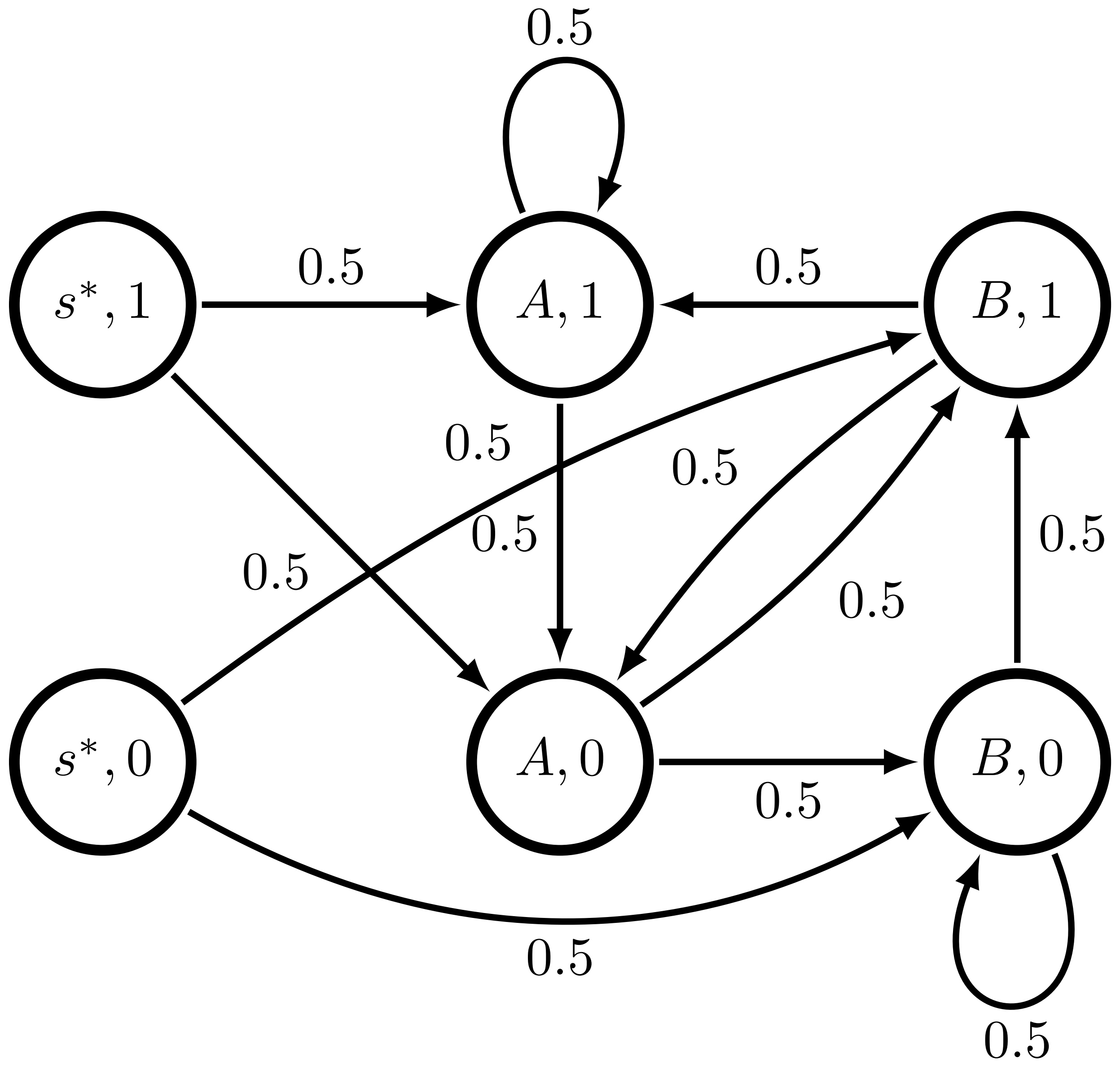
$$\Pr(Y^{\theta}_{0:L} = y_{0:L}) \equiv \sum_{s_{0:L+1}} \delta_{s_0,s^*} \prod_{i=0}^{L-1} \boxed{\theta^{(y_i)}_{s_i \to s_{i+1}}}$$

Estimated ϵ-machine (model)



$$\boxed{\theta^{(y)}_{s \to s'}} = \Pr(S_{i+1} = s', Y^P_i = y | S_i = s)$$

Efficient Agent



$$M_{xy \to x'y'} = \frac{1}{|\mathcal{Y}|} \times \begin{cases} \delta_{x', \epsilon(x,y)} & \text{if } \sum_{x'} \boxed{\theta^{(y)}_{x \to x'}} \neq 0 \\ \delta_{x',x} & \text{else.} \end{cases}$$

$$\Pr(Y^{\theta}_i = y_i, X^{\theta}_i = s_i) = \sum_{y_{0:i}, s_{0:i}, s_{i+1}} \delta_{s_0,s^*} \prod_{j=0}^{i} \boxed{\theta^{(y_j)}_{s_j \to s_{j+1}}}$$

Crutchfield and Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos, (2003)

Boyd, Alexander B., James P. Crutchfield, and Mile Gu. "Thermodynamic machine learning through maximum work production." *New Journal of Physics* 24.8 (2022): 083040.

Thermodynamic Learning: Maximum Work Production

For a single input string, efficient agents have a simple expression for work production

$$\begin{aligned}\langle W^{\theta}(y_{0:L})\rangle &= k_B T(\ln \Pr(Y^{\theta}_{0:L} = y_{0:L}) + L \ln |\mathcal{Y}|) \\ &= k_B T \ell(\theta | y_{0:L}) + k_B T L \ln |\mathcal{Y}|\end{aligned}$$

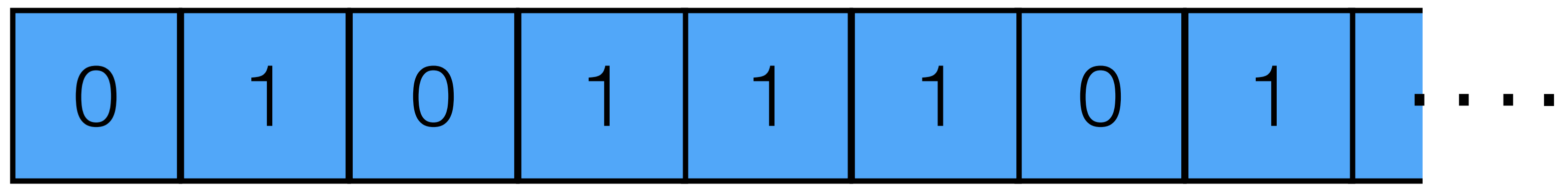
Maximizing log-likelihood of model from data also maximizes work production.

Boyd, Alexander B., James P. Crutchfield, and Mile Gu. "Thermodynamic machine learning through maximum work production." *New Journal of Physics* 24.8 (2022): 083040.

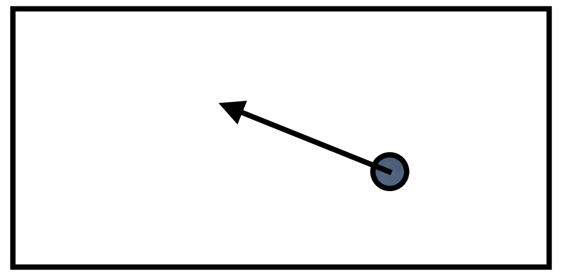
Maximum Likelihood Estimation and Maximum Work Production are equivalent

Machine Learning vs. Information Thermodynamics

1) Data

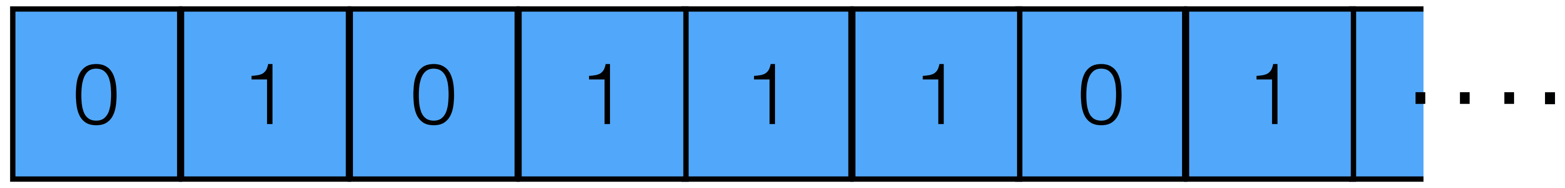


1) Physical system

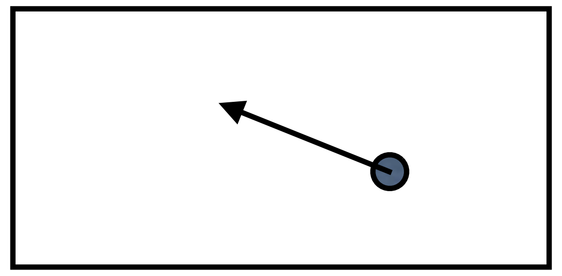


Machine Learning vs. Information Thermodynamics

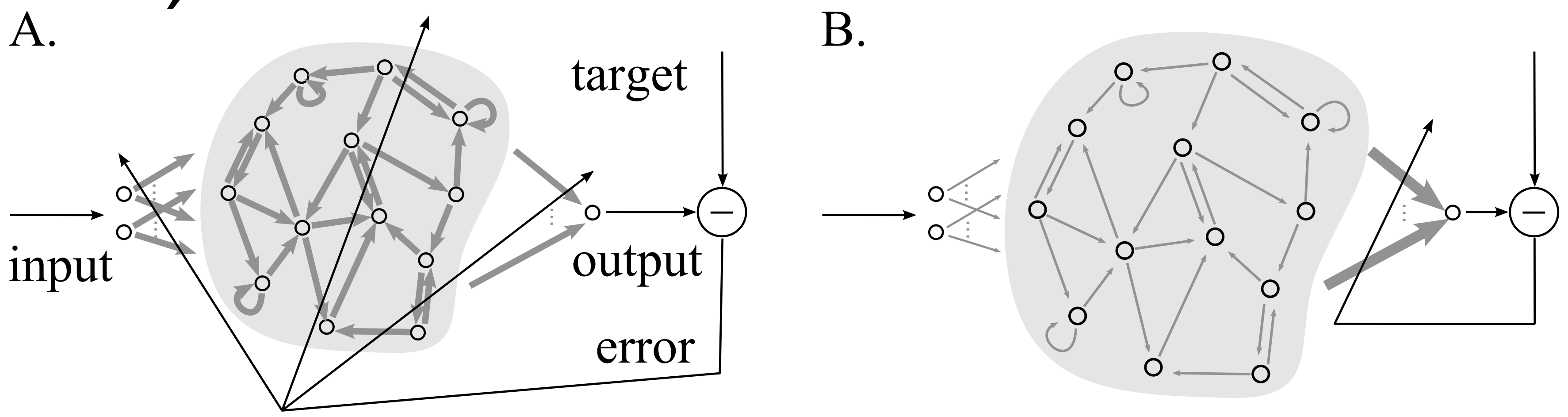
1) Data



1) Physical system

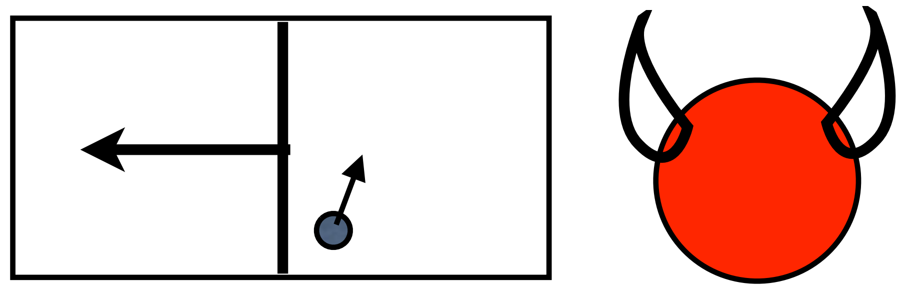


2) Model



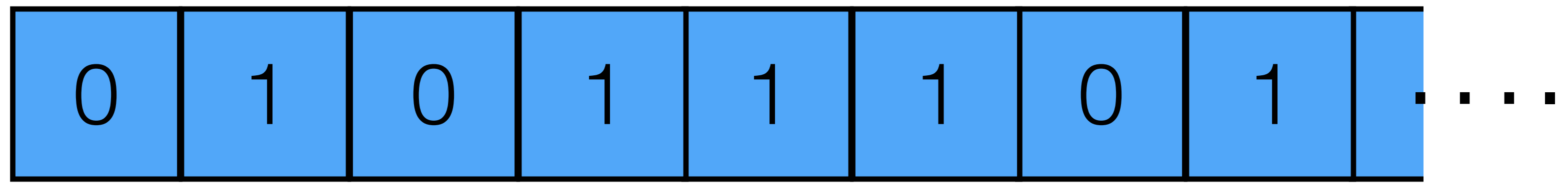
Mantas and Jaeger, Comp. Sci. Rev., (2009)

2) Demon/Agent/Controller

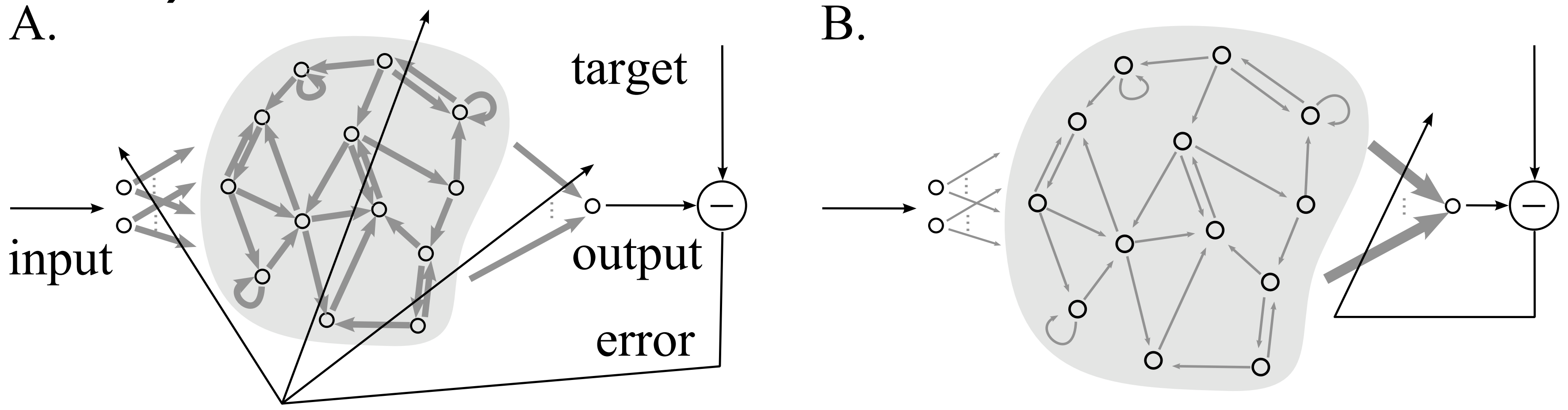


Machine Learning vs. Information Thermodynamics

1) Data



2) Model

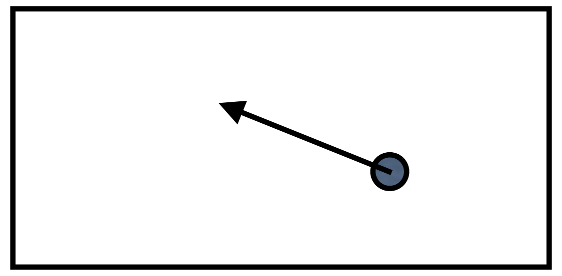


Mantas and Jaeger, Comp. Sci. Rev., (2009)

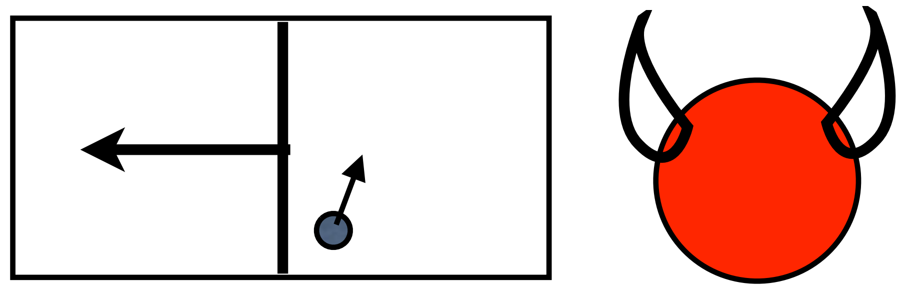
3) Performance measure

$$\ell(\theta|\vec{z}) = \sum_{i}^{N} \ln \Pr(Z = z_i|\Theta = \theta)$$

1) Physical system

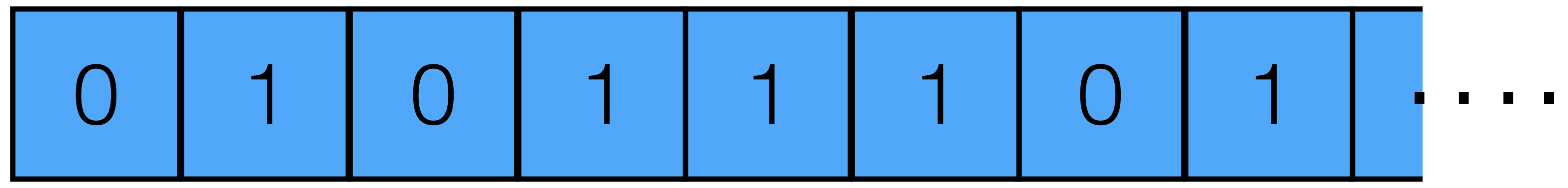


2) Demon/Agent/Controller

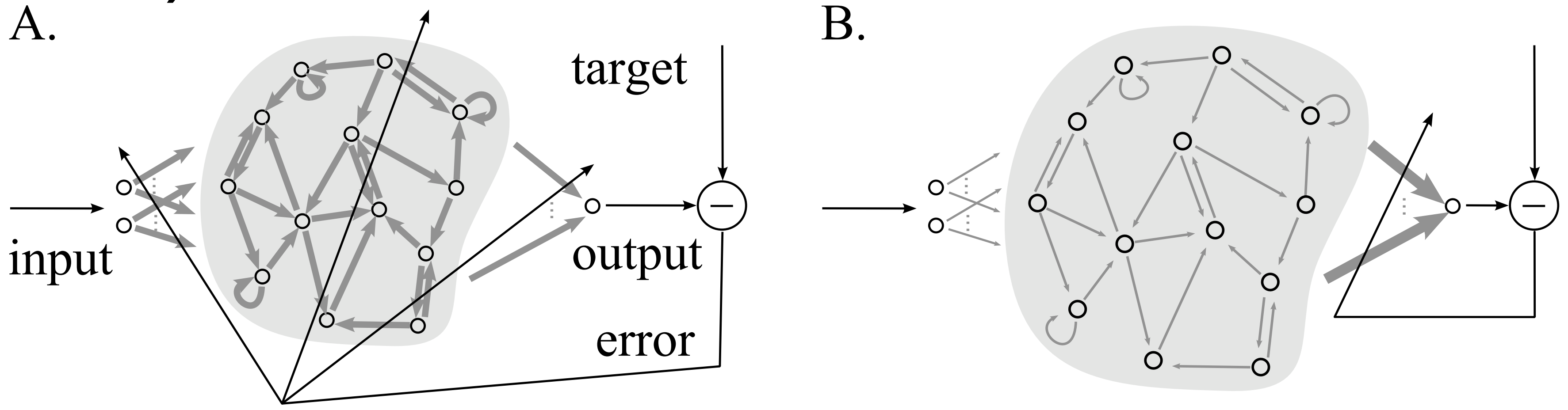


Machine Learning vs. Information Thermodynamics

1) Data



2) Model

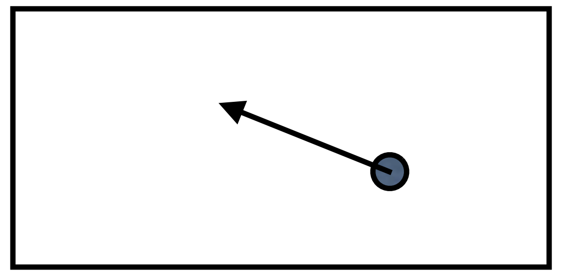


Mantas and Jaeger, Comp. Sci. Rev., (2009)

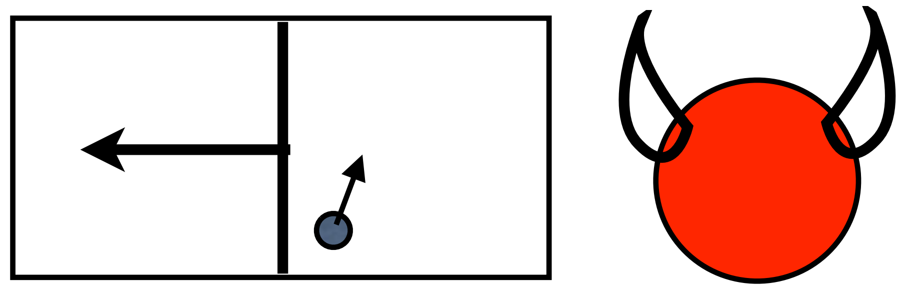
3) Performance measure

$$\ell(\theta|\vec{z}) = \sum_{i}^{N} \ln \Pr(Z = z_i|\Theta = \theta)$$

1) Physical system



2) Demon/Agent/Controller

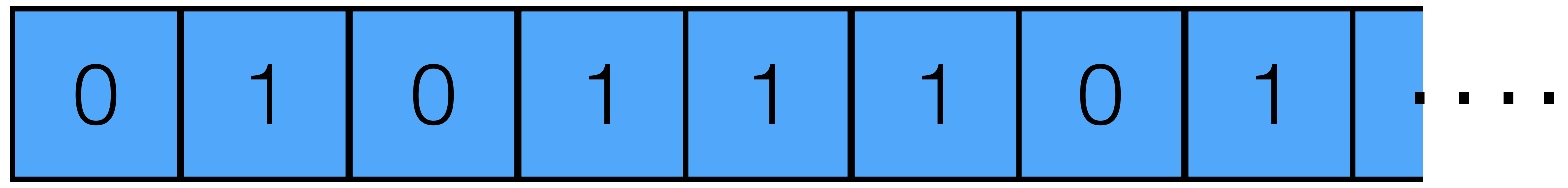


3) Work production

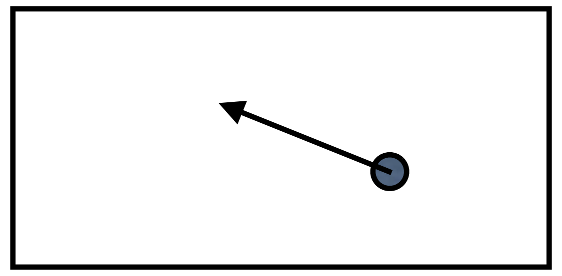
$$\langle W^{\theta}_{|y_{0:L}} \rangle = k_B T \ell(\theta|y_{0:L}) + k_B T L \ln |\mathcal{Y}|$$

Machine Learning vs. Information Thermodynamics

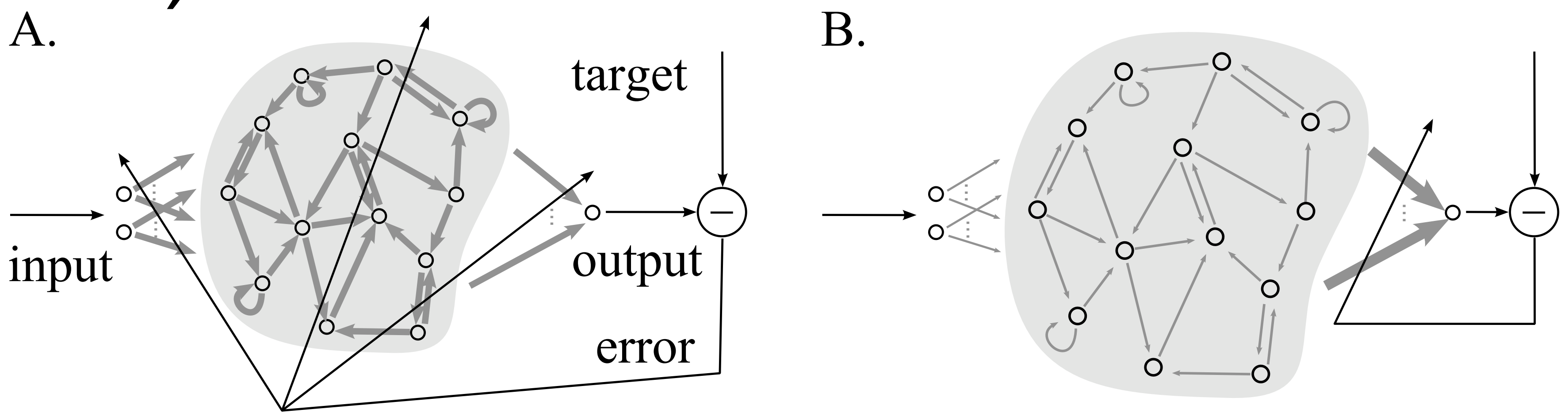
1) Data



1) Physical system

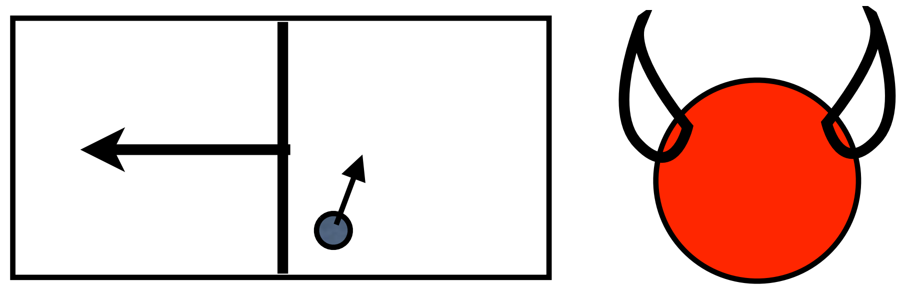


2) Model



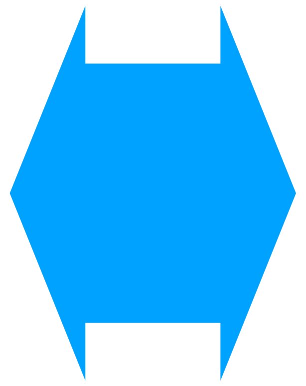
Mantas and Jaeger, Comp. Sci. Rev., (2009)

2) Demon/Agent/Controller



3) Performance measure

$$\ell(\theta|\vec{z}) = \sum_i^N \ln \Pr(Z = z_i|\Theta = \theta)$$



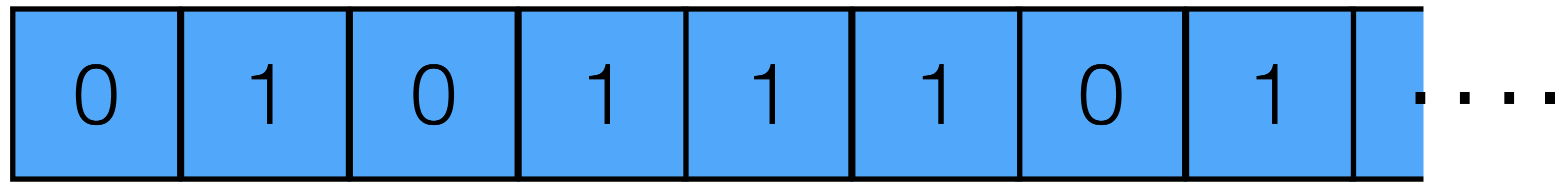
3) Work production

$$\langle W^{\theta}_{|y_{0:L}} \rangle = k_B T \ell(\theta|y_{0:L}) + k_B T L \ln |\mathcal{Y}|$$

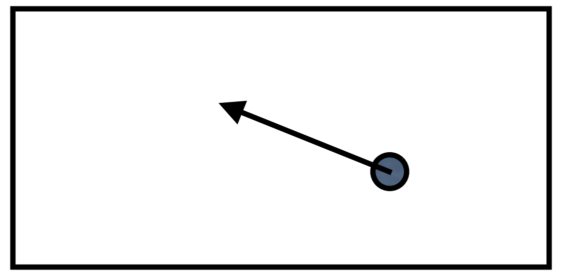
Maximum work production and Maximum Likelihood Estimation are equivalent.

Machine Learning vs. Information Thermodynamics

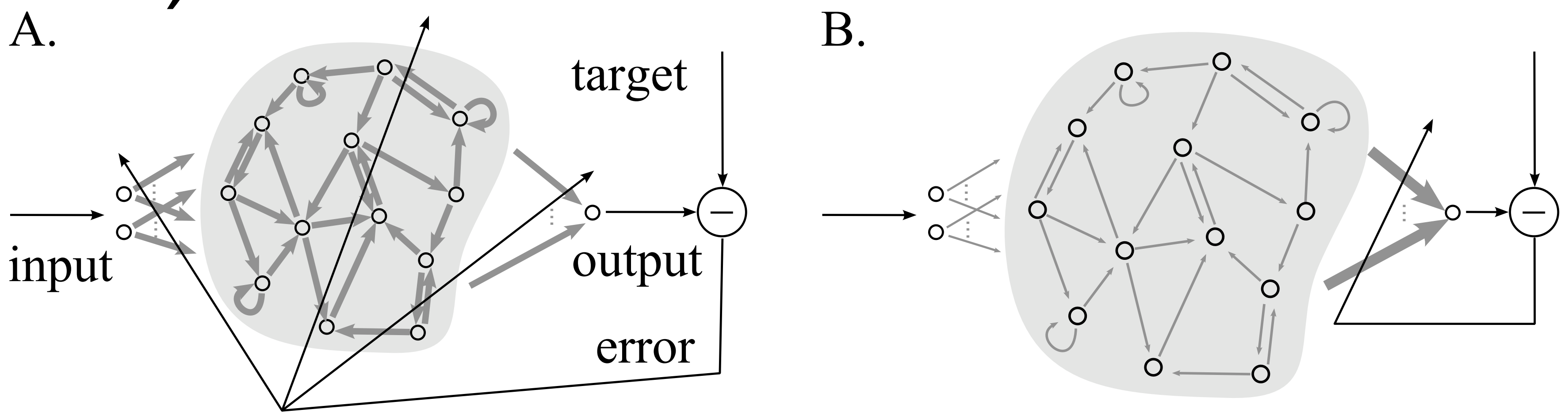
1) Data



1) Physical system

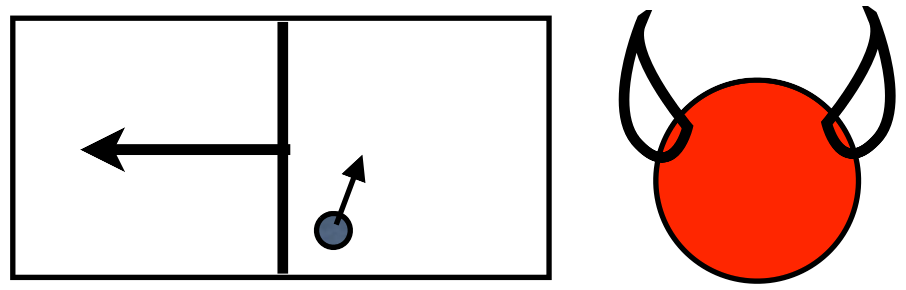


2) Model



Mantas and Jaeger, Comp. Sci. Rev., (2009)

2) Demon/Agent/Controller



3) Performance measure

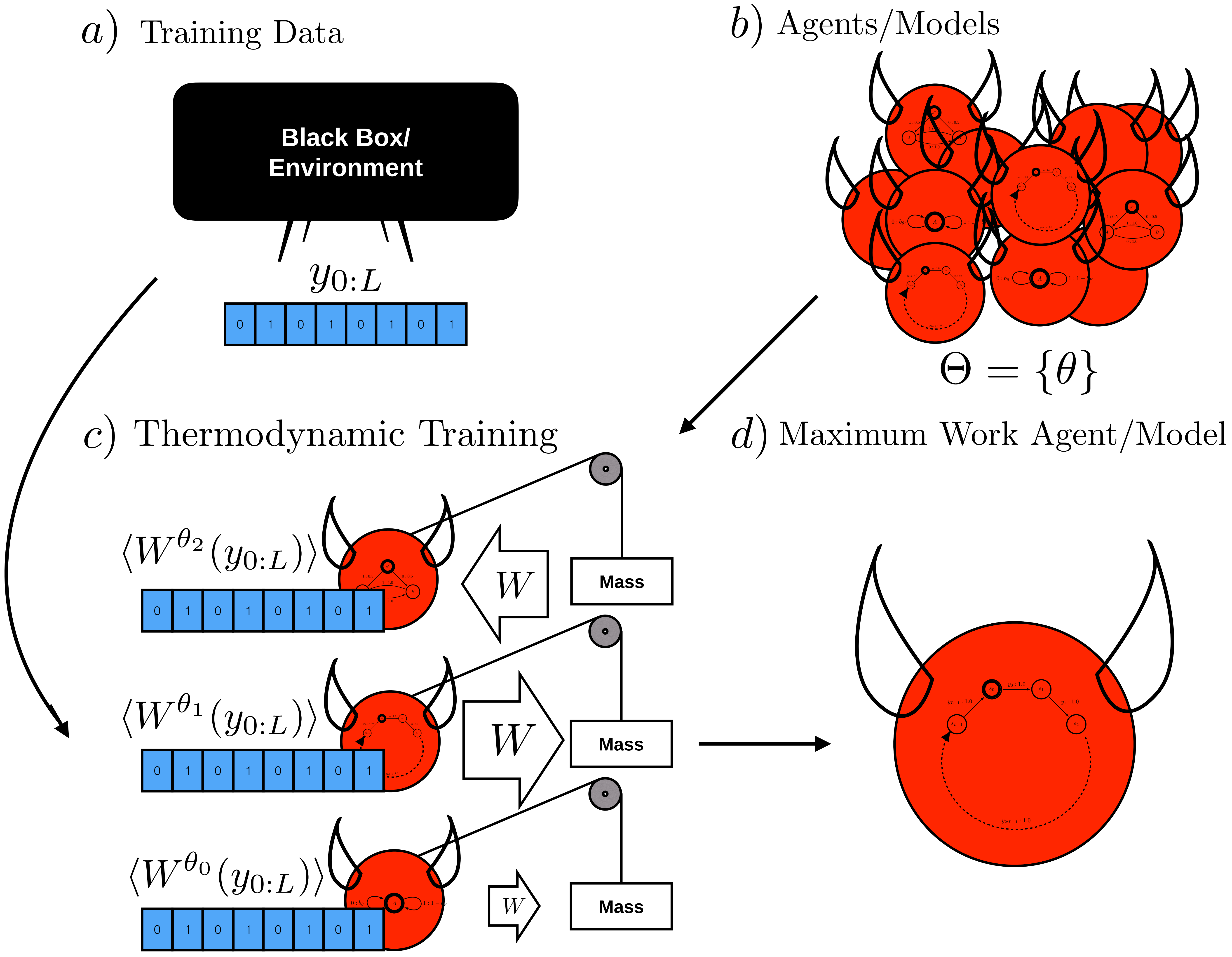
$$\ell(\theta|\vec{z}) = \sum_{i}^{N} \ln \Pr(Z = z_i | \Theta = \theta)$$

3) Work production

$$\langle W^{\theta}_{|y_{0:L}} \rangle = k_B T \ell(\theta | y_{0:L}) + k_B T L \ln |\mathcal{Y}|$$

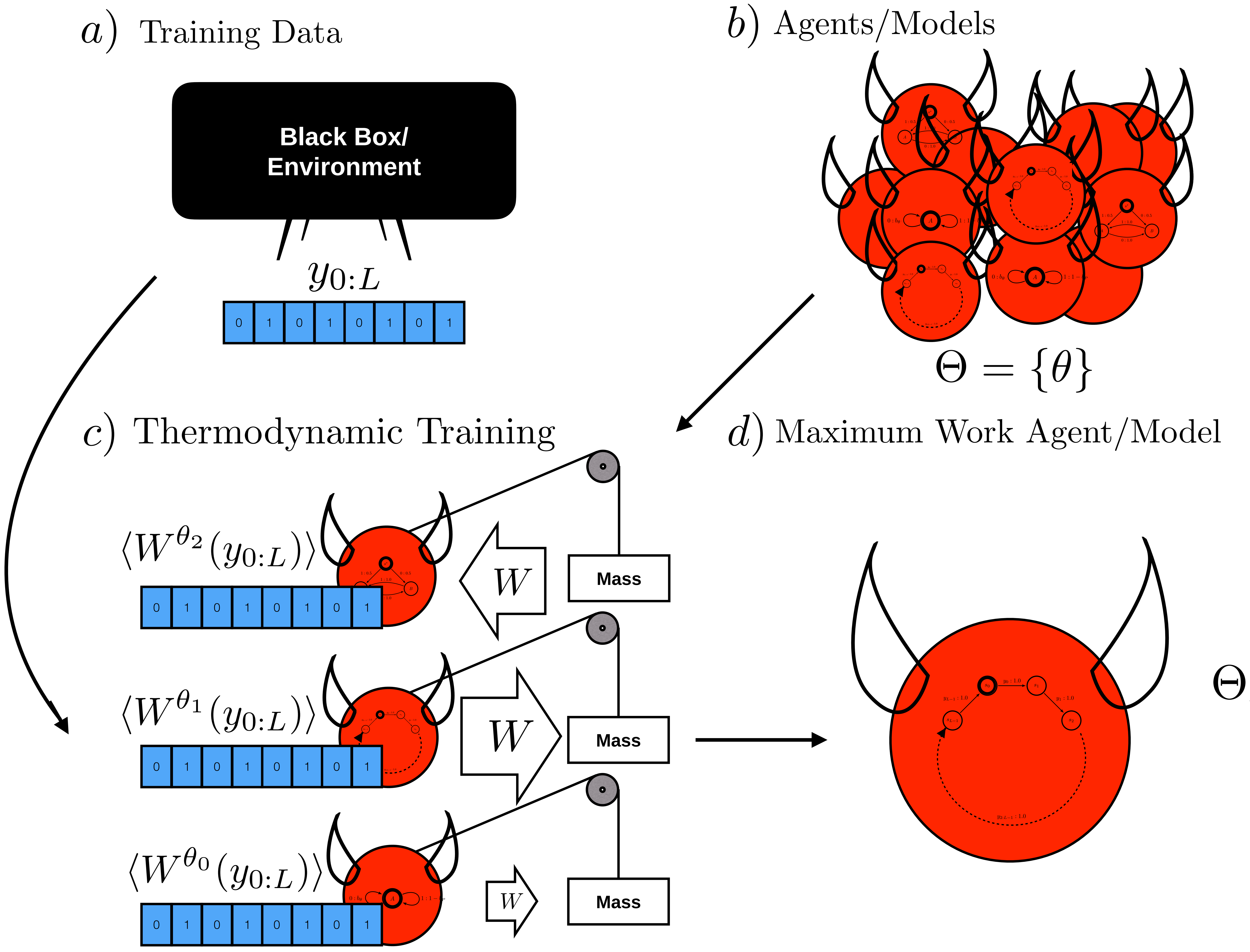
Maximum work production and Maximum Likelihood Estimation are equivalent. (When using epsilon machine models.)
What happens if we maximize work production?

Thermodynamic Training With Different Size Memories



Memory size: $n \equiv |\mathcal{X}|$

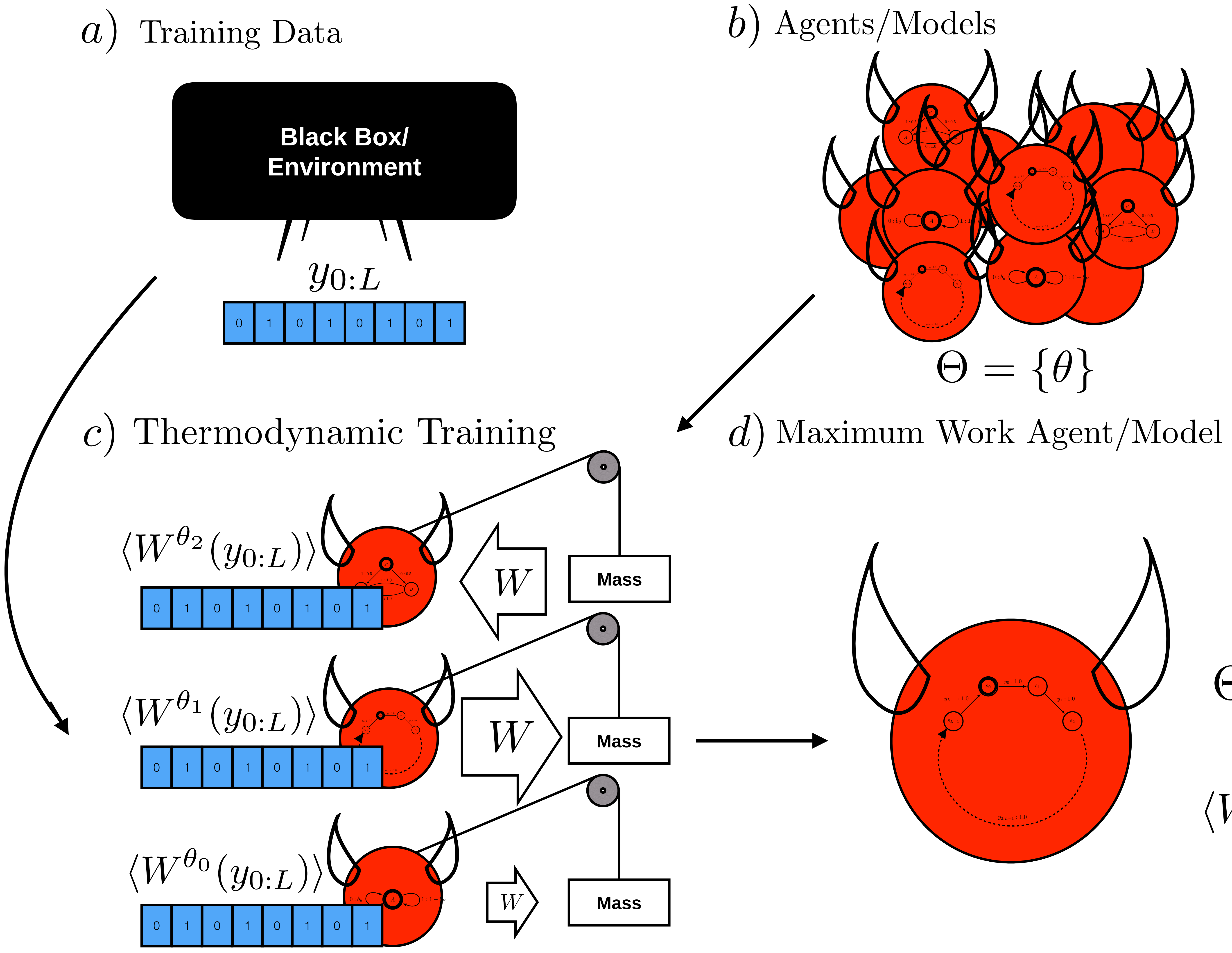
Thermodynamic Training With Different Size Memories



Memory size: $n \equiv |\mathcal{X}|$

$$\Theta_n^{\max}(y_{0:L}) = \underset{\theta \in \{\text{n state models}\}}{\operatorname{argmax}} \langle W^{\theta}(y_{0:L})\rangle$$

Thermodynamic Training With Different Size Memories

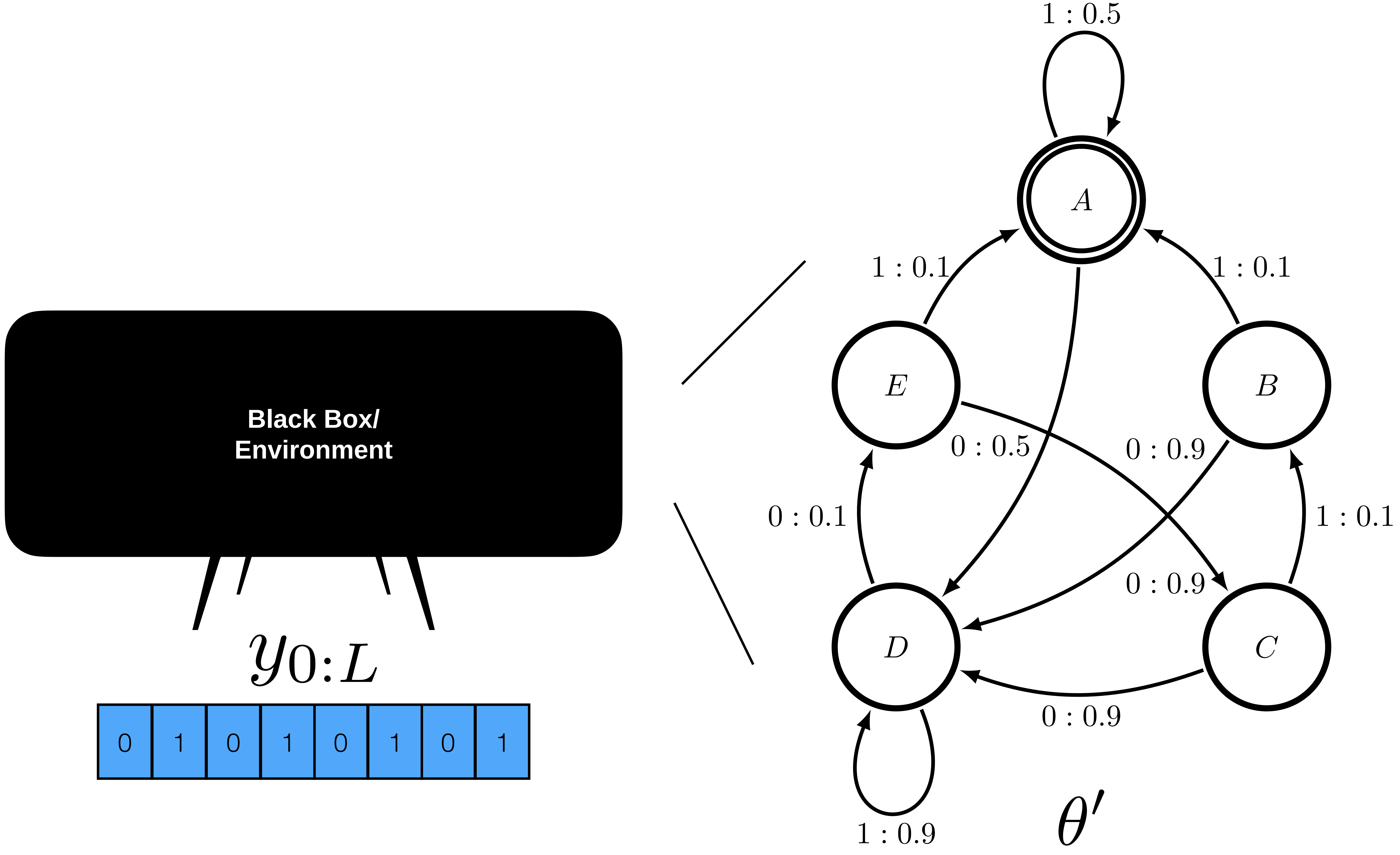


Memory size: $n \equiv |\mathcal{X}|$

$$\Theta_n^{\max}(y_{0:L}) = \underset{\theta \in \{\text{n state models}\}}{\operatorname{argmax}} \langle W^{\theta}(y_{0:L}) \rangle$$

$$\langle W_n^{\max}(y_{0:L}) \rangle = \max_{\theta \in \{\text{n state models}\}} \langle W^{\theta}(y_{0:L}) \rangle$$

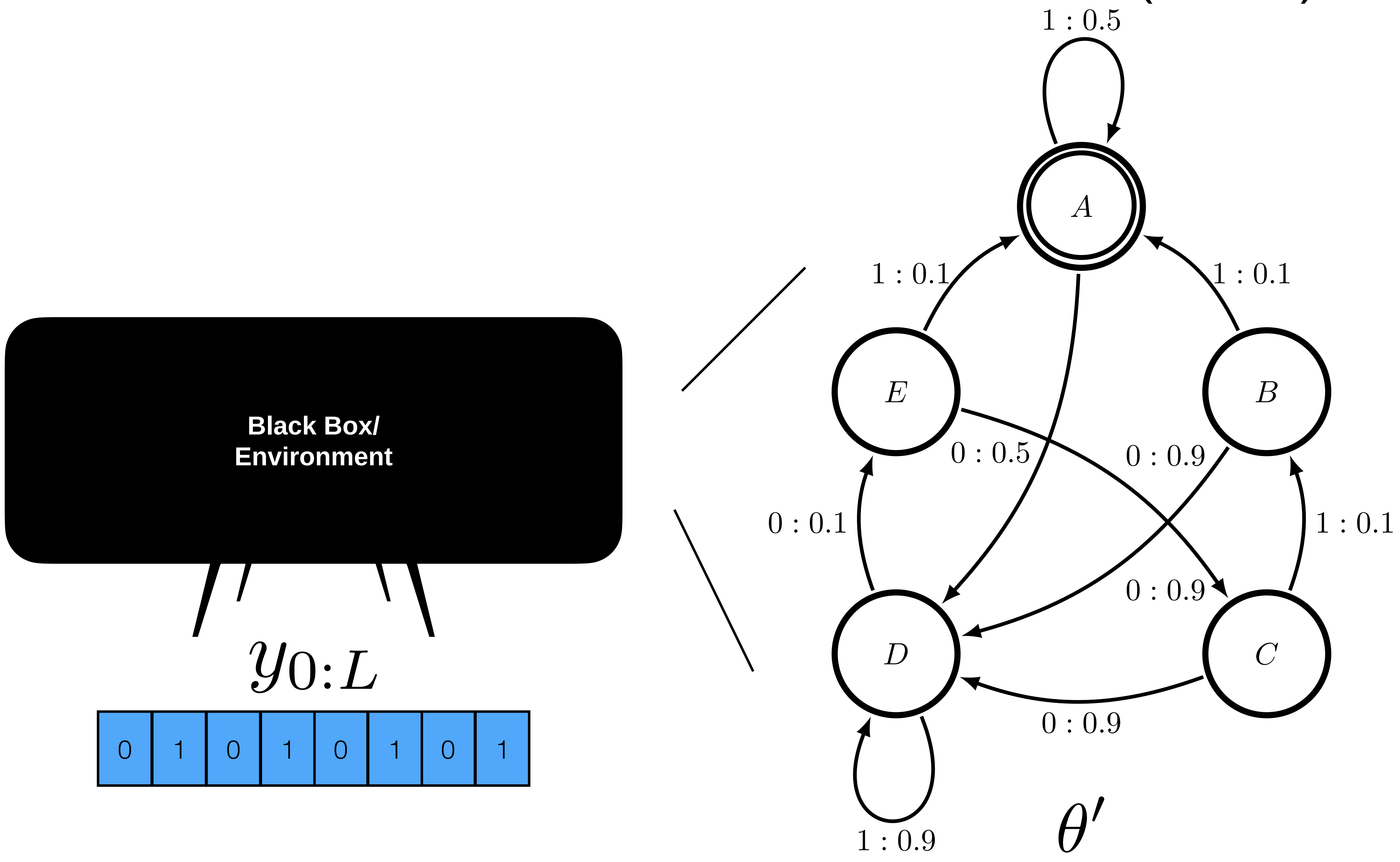
What's in the (Black) Box??



Entropy rate: $h_\mu^{\theta'} \equiv \lim_{L \to \infty} \frac{H[Y_{0:L}^{\theta'}]}{L}$

$$\Pr(Y_{0:L}^{\theta'}) = \Pr(Y_{0:L})$$

What's in the (Black) Box??



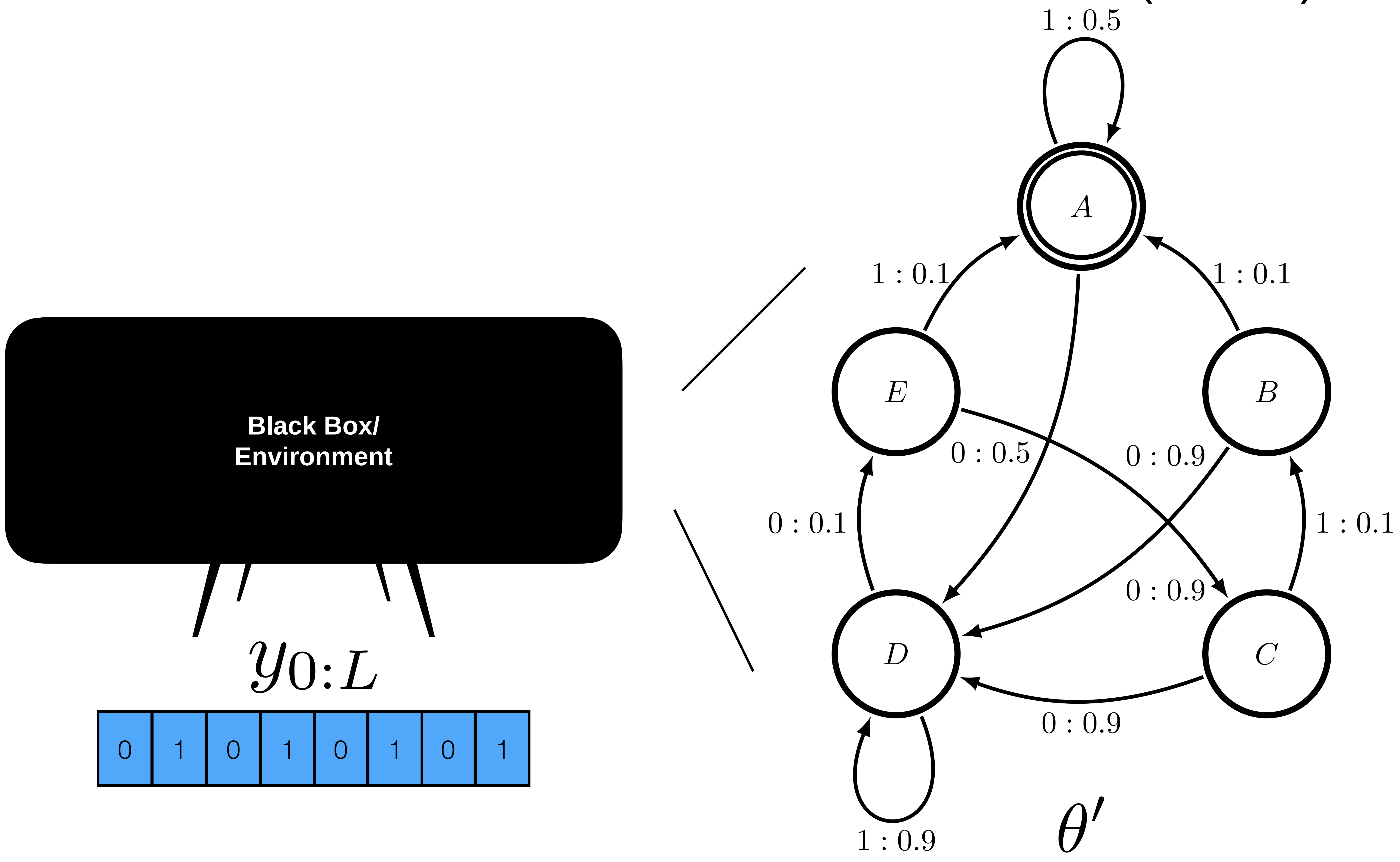
$$\Pr(Y^{\theta'}_{0:L}) = \Pr(Y_{0:L})$$

Entropy rate: $h^{\theta'}_{\mu} \equiv \lim_{L\to\infty} \frac{H[Y^{\theta'}_{0:L}]}{L}$

Asymptotic limit on work rate production:

$$\langle W \rangle_{\infty} \leq k_B T \left(\ln 2 - h^{\theta'}_{\mu} \right)$$
$$\equiv \langle W^{\theta'} \rangle_{\infty}$$

What's in the (Black) Box??



$$\Pr(Y^{\theta'}_{0:L}) = \Pr(Y_{0:L})$$

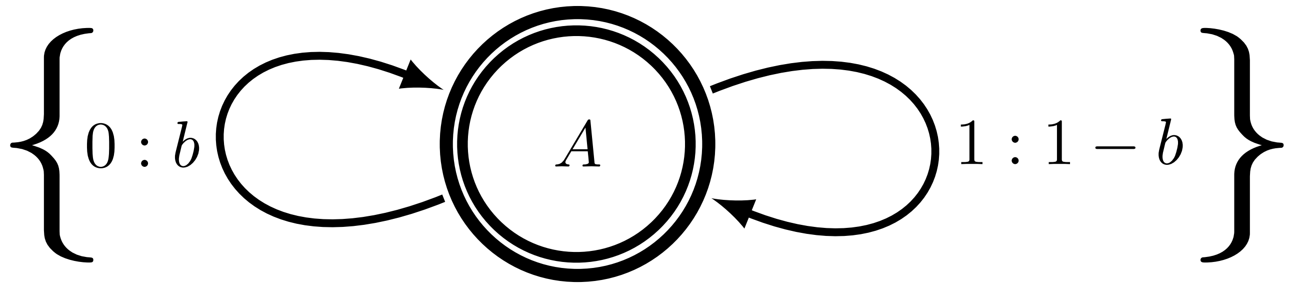
Entropy rate: $h^{\theta'}_{\mu} \equiv \lim_{L\to\infty} \frac{H[Y^{\theta'}_{0:L}]}{L}$

Asymptotic limit on work rate production:

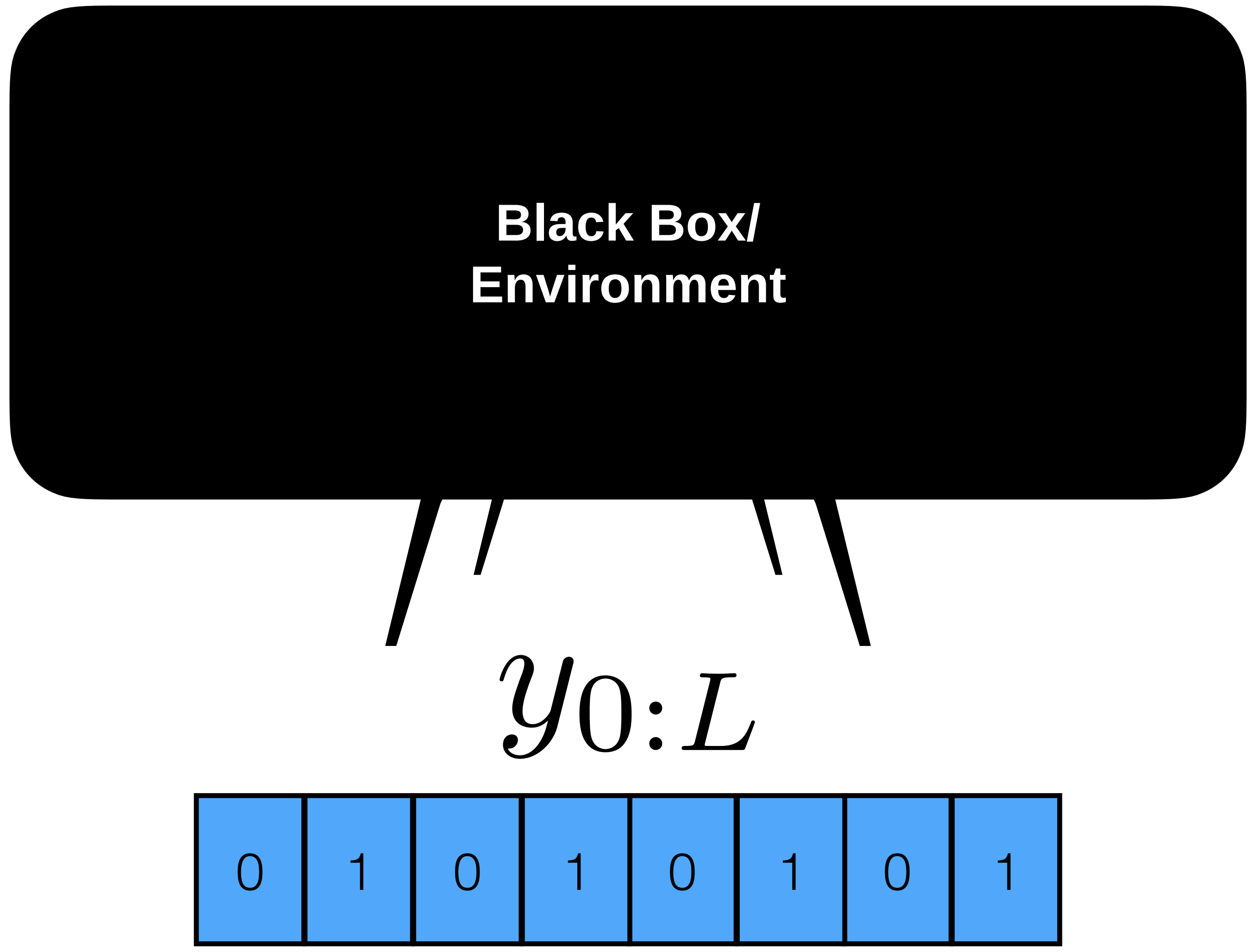
$$\langle W \rangle_{\infty} \leq k_B T \left(\ln 2 - h^{\theta'}_{\mu}\right)$$
$$\equiv \langle W^{\theta'} \rangle_{\infty}$$

Requisite Complexity:
5 causal states requires 5 memory states in the information engine.

1-State Machines

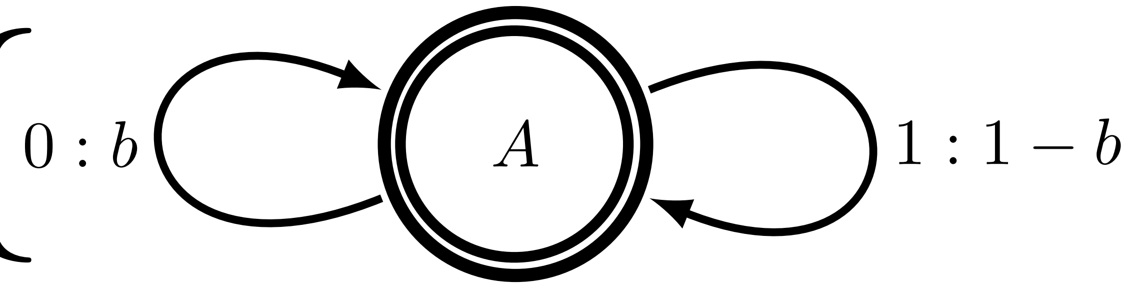
Possible Models: $\theta \in$ 

θ'

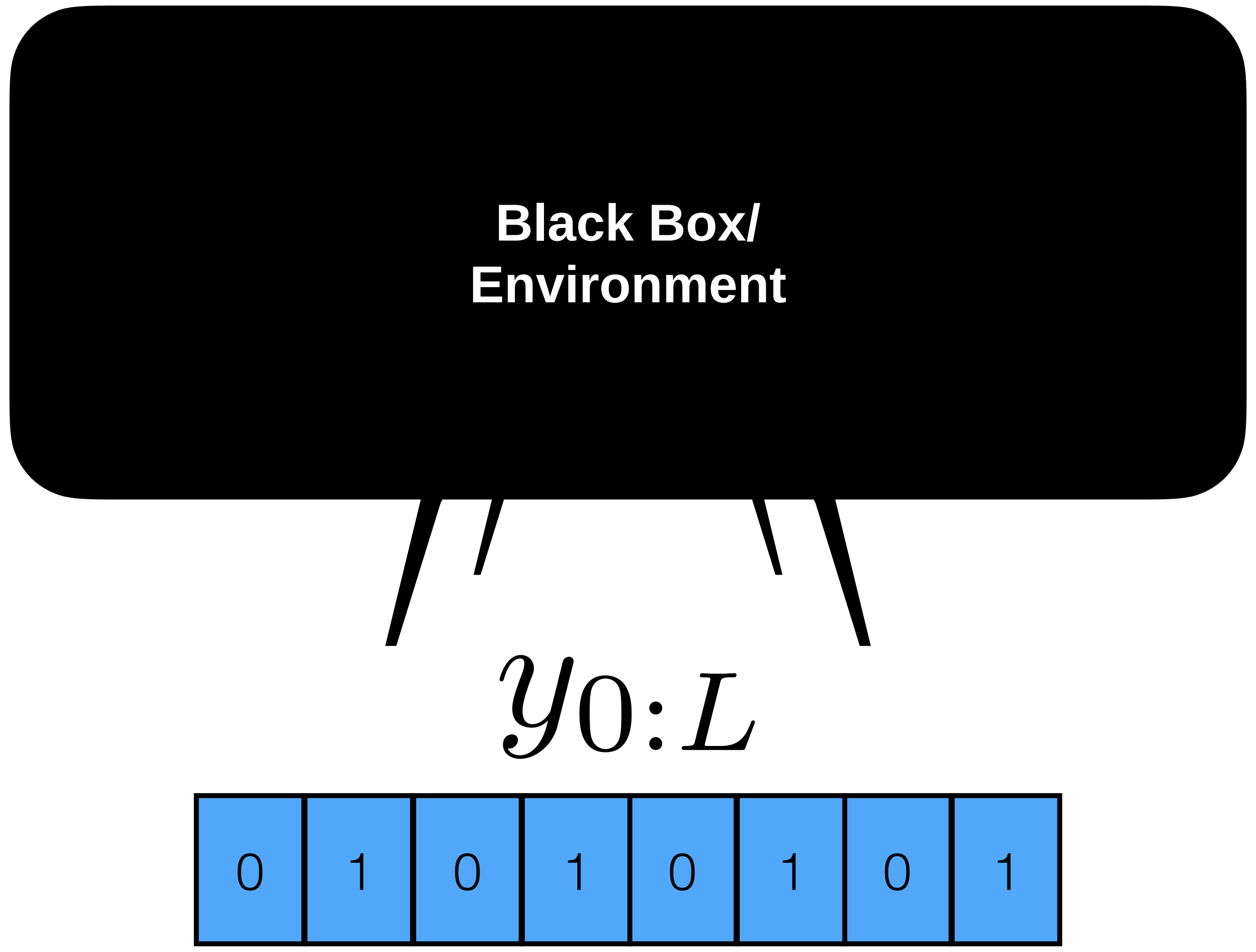


Inverse temperature: $\beta = \dfrac{1}{k_B T}$

1-State Machines

Possible Models: $\theta \in \Big\{$  $\Big\}$

θ'

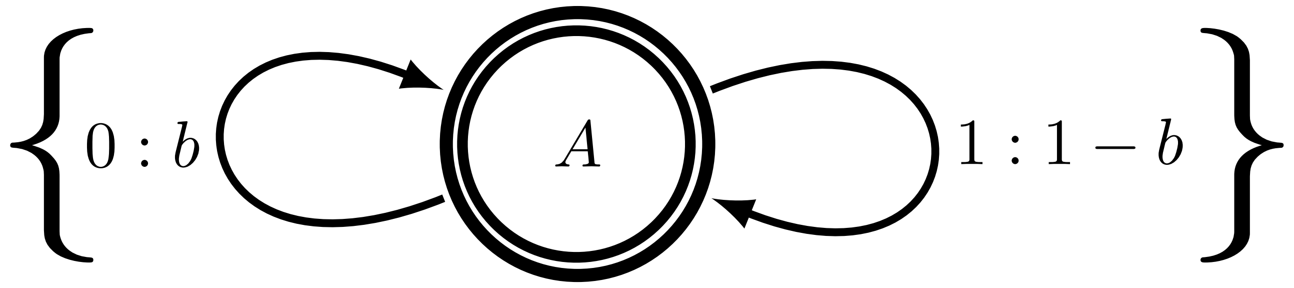


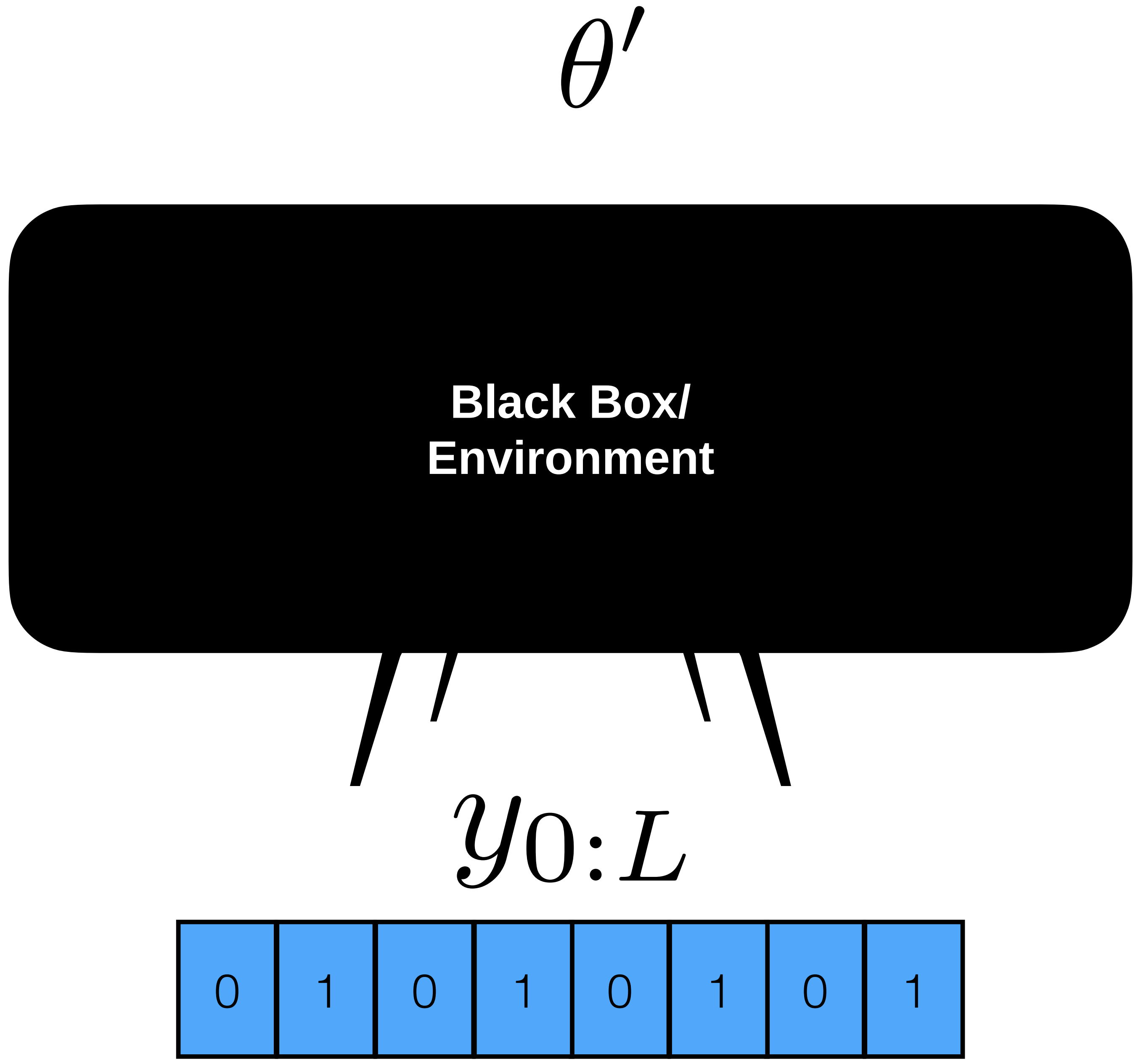
$$\beta \langle W^{\theta}(y_{0:L}) \rangle = L \ln |\mathcal{Y}| + \ln \Pr(Y^{\theta}_{0:L} = y_{0:L})$$

Inverse temperature: $\beta = \dfrac{1}{k_B T}$

1-State Machines

Possible Models: $\theta \in$

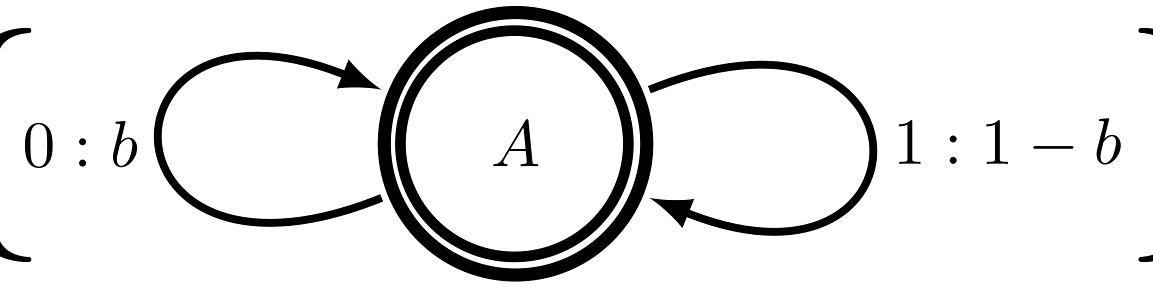




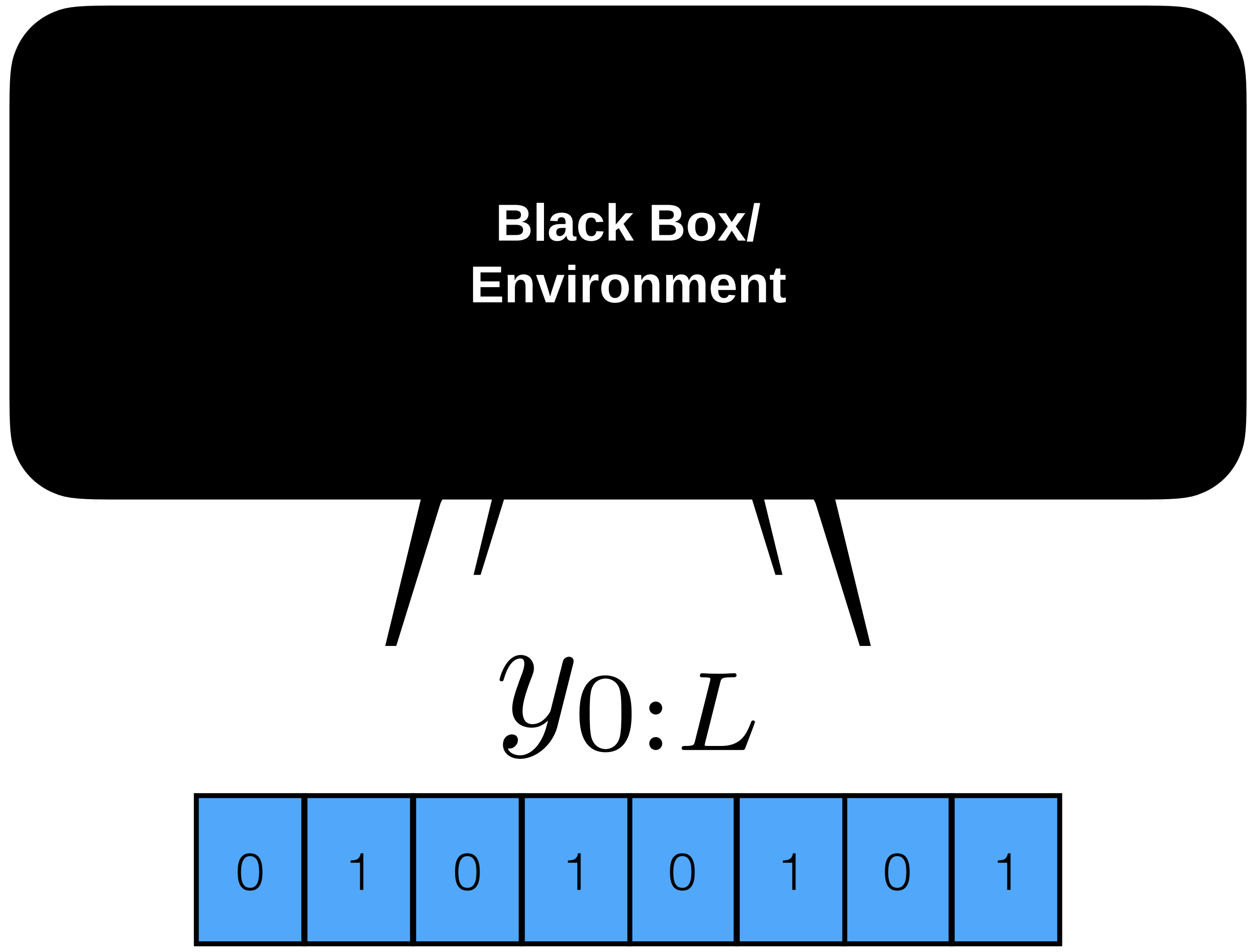
$$\beta \langle W^{\theta}(y_{0:L}) \rangle = L \ln |\mathcal{Y}| + \ln \Pr(Y^{\theta}_{0:L} = y_{0:L})$$
$$= L \ln |\mathcal{Y}| + \sum_{i=0}^{L-1} \ln \theta^{(y_i)}_{\epsilon(y_{0:i}) \to \epsilon(y_{0:i+1})}$$

Inverse temperature: $\beta = \frac{1}{k_B T}$

1-State Machines

Possible Models: $\theta \in \Big\{$  $\Big\}$

θ'



$$\beta \langle W^{\theta}(y_{0:L}) \rangle = L \ln |\mathcal{Y}| + \ln \Pr(Y^{\theta}_{0:L} = y_{0:L})$$

$$= L \ln |\mathcal{Y}| + \sum_{i=0}^{L-1} \ln \theta^{(y_i)}_{\epsilon(y_{0:i}) \to \epsilon(y_{0:i+1})}$$

$$= L \ln |\mathcal{Y}| + \sum_{s,y} N^{(y)}_s \ln \theta^{(y)}_{s \to \epsilon(s,y)}$$

$N^{(y)}_s \equiv$ number of times causal state s receives input y

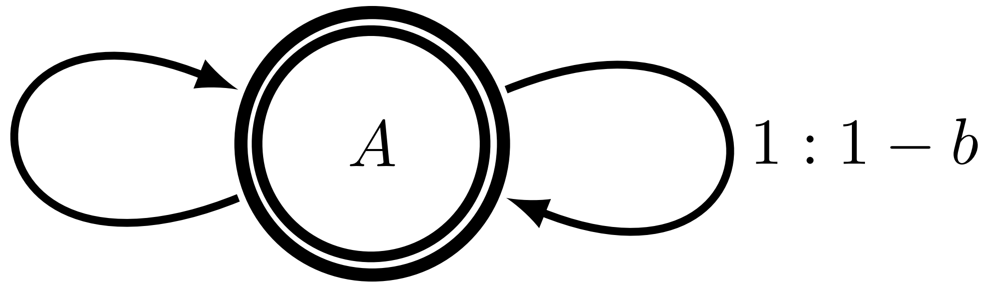
PHYSICAL REVIEW E **89**, 042119 (2014)

Bayesian structural inference for hidden processes

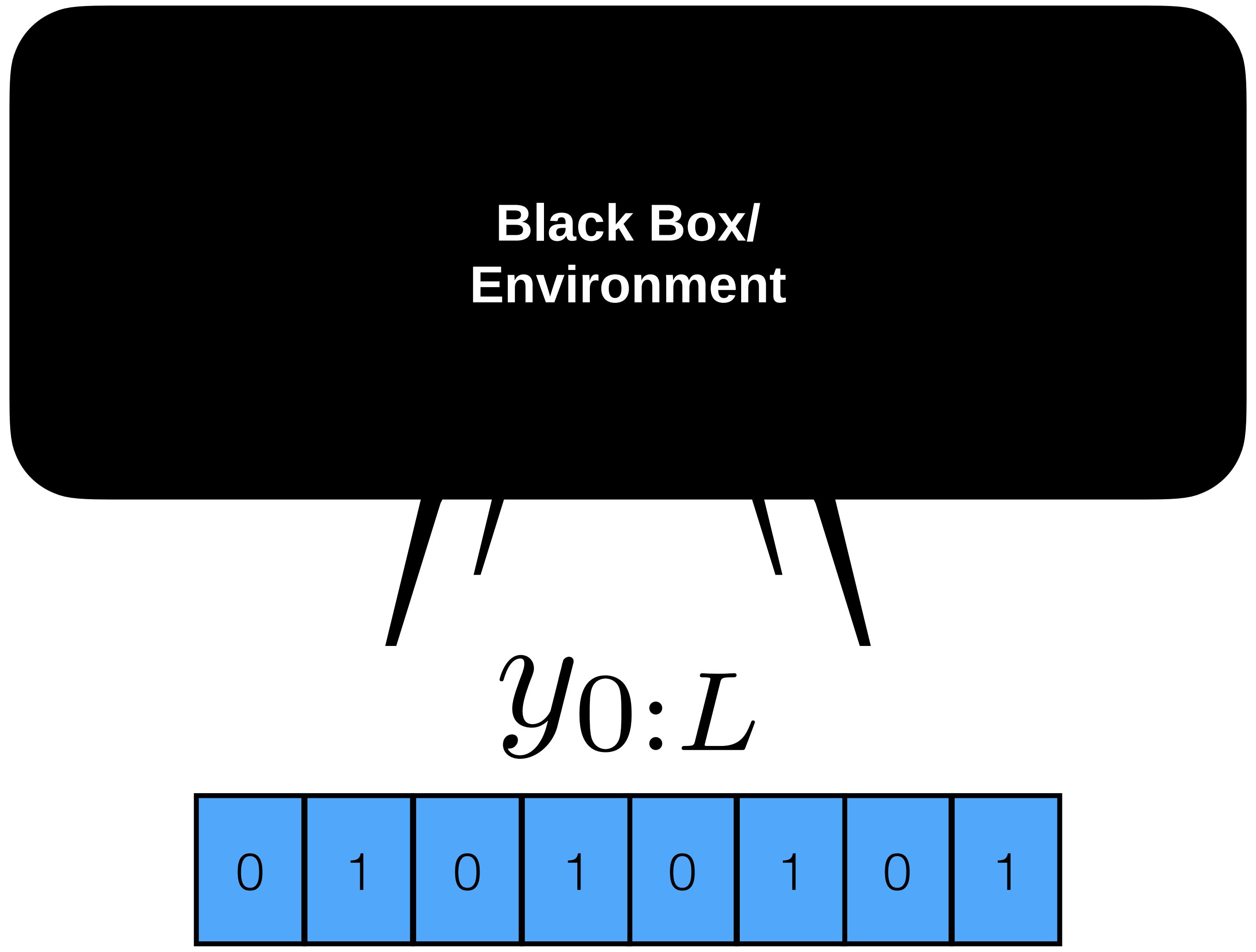
Christopher C. Strelioff[1,*] and James P. Crutchfield[1,2,†]

Inverse temperature: $\beta = \dfrac{1}{k_B T}$

1-State Machines

Possible Models: $\theta \in \Bigg\{$  $\Bigg\}$

$$\theta'$$



$$\begin{aligned}
\beta \langle W^{\theta}(y_{0:L}) \rangle &= L \ln |\mathcal{Y}| + \ln \Pr(Y^{\theta}_{0:L} = y_{0:L}) \\
&= L \ln |\mathcal{Y}| + \sum_{i=0}^{L-1} \ln \theta^{(y_i)}_{\epsilon(y_{0:i}) \to \epsilon(y_{0:i+1})} \\
&= L \ln |\mathcal{Y}| + \sum_{s,y} N^{(y)}_s \ln \theta^{(y)}_{s \to \epsilon(s,y)}
\end{aligned}$$

$$N^{(y)}_s \equiv \text{number of times causal state } s \text{ receives input } y$$

$$\theta^{(y)}_{s \to \epsilon(s,y)} = \frac{N^{(y)}_s}{\sum_y N^{(y)}_s} = \text{fraction of times causal state } s \text{ receives input } y$$

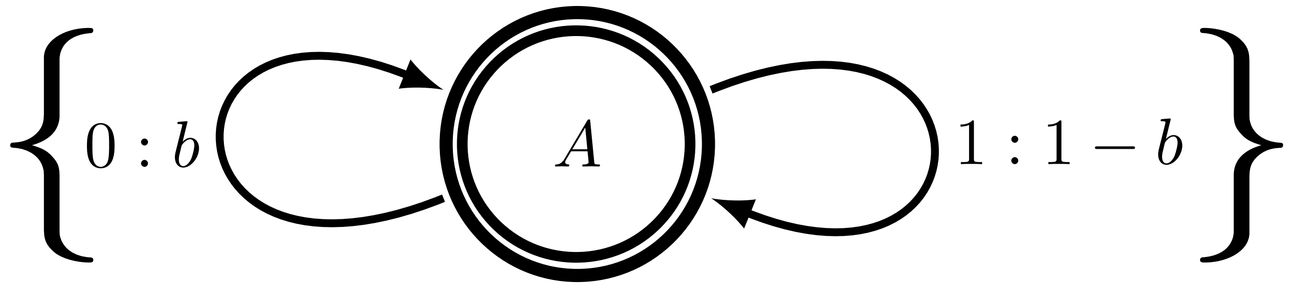
Inverse temperature: $\beta = \dfrac{1}{k_B T}$

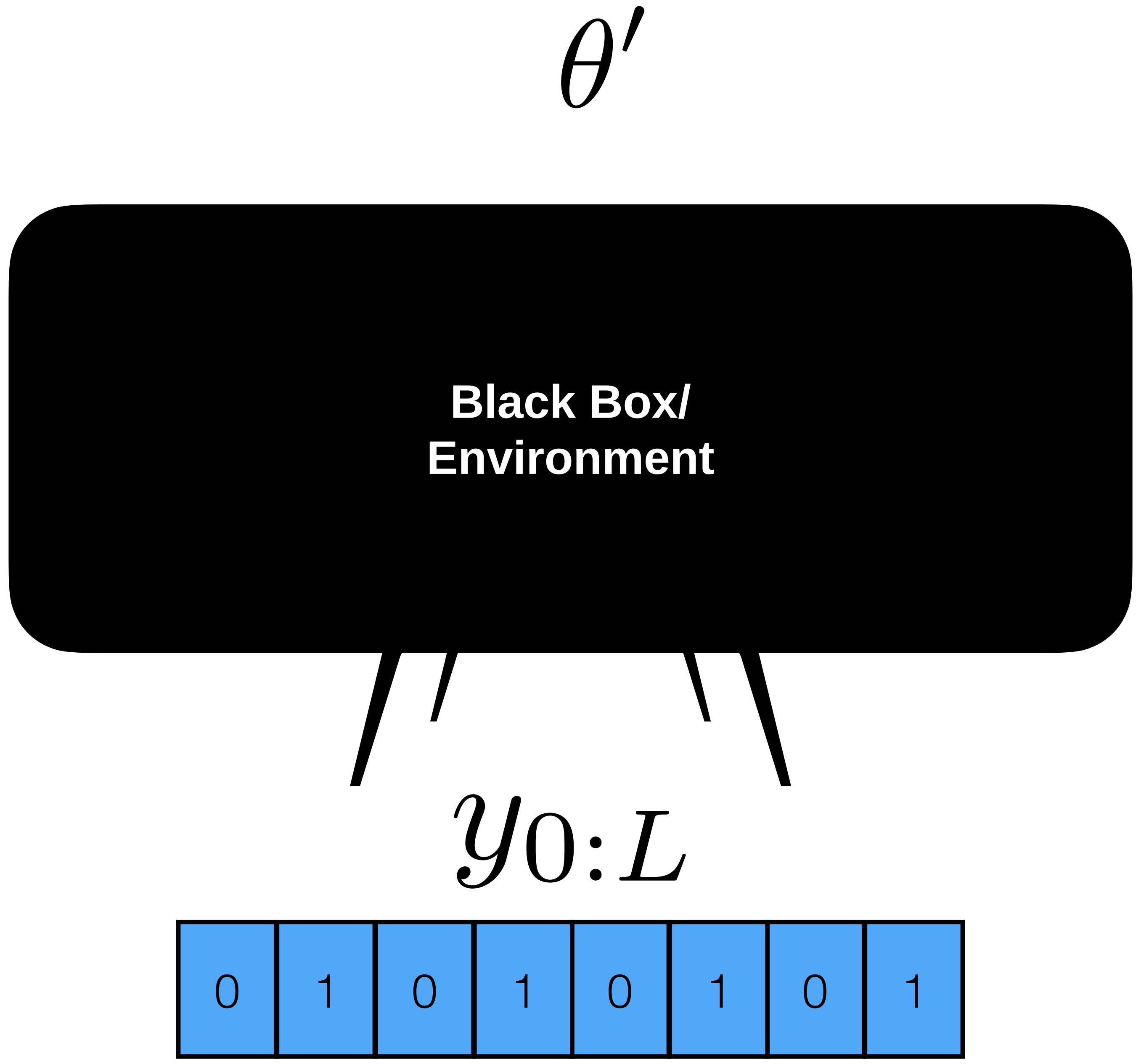
PHYSICAL REVIEW E **89**, 042119 (2014)

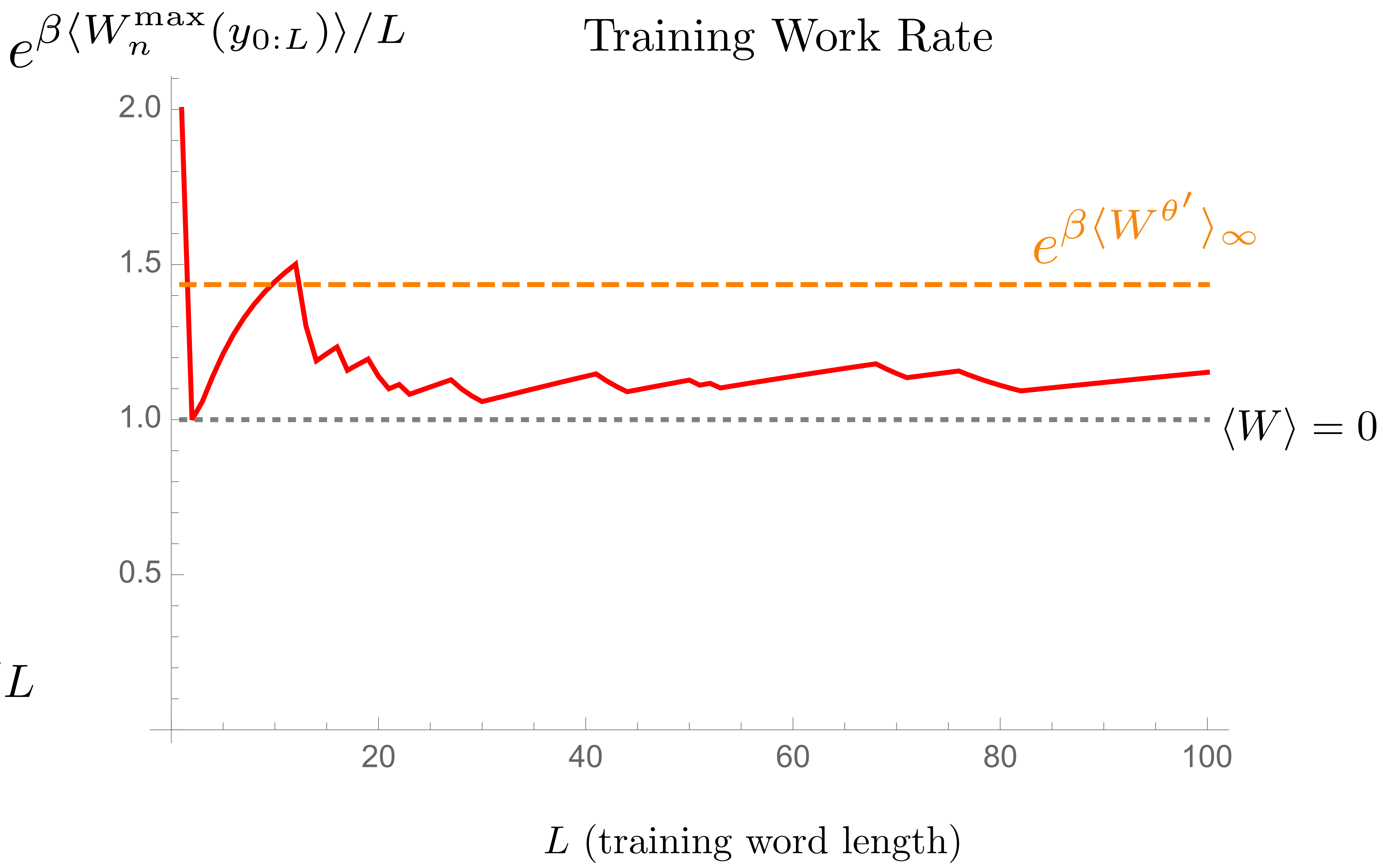
Bayesian structural inference for hidden processes

Christopher C. Strelioff[1,*] and James P. Crutchfield[1,2,†]

1-State Machines

Possible Models: $\theta \in$ 



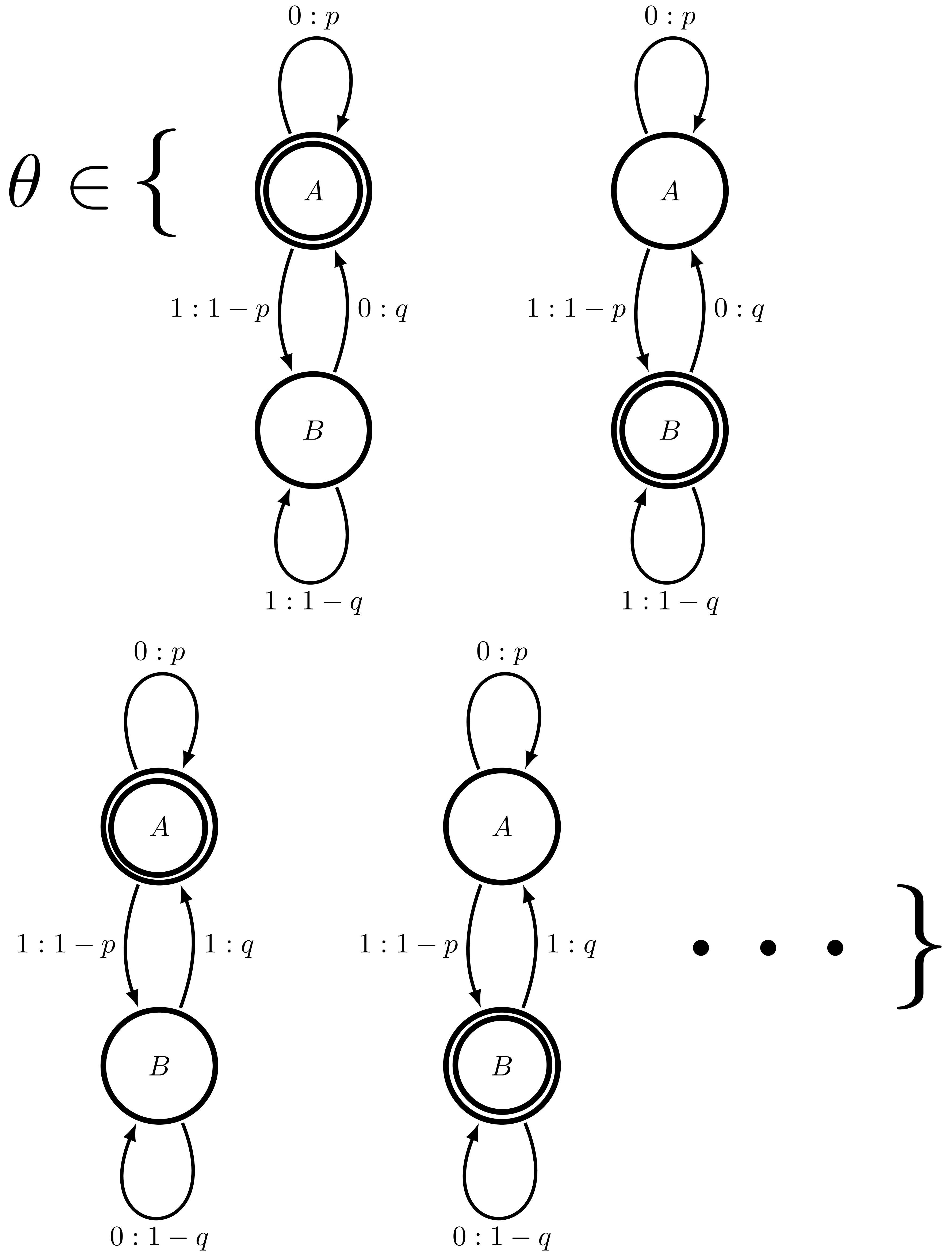


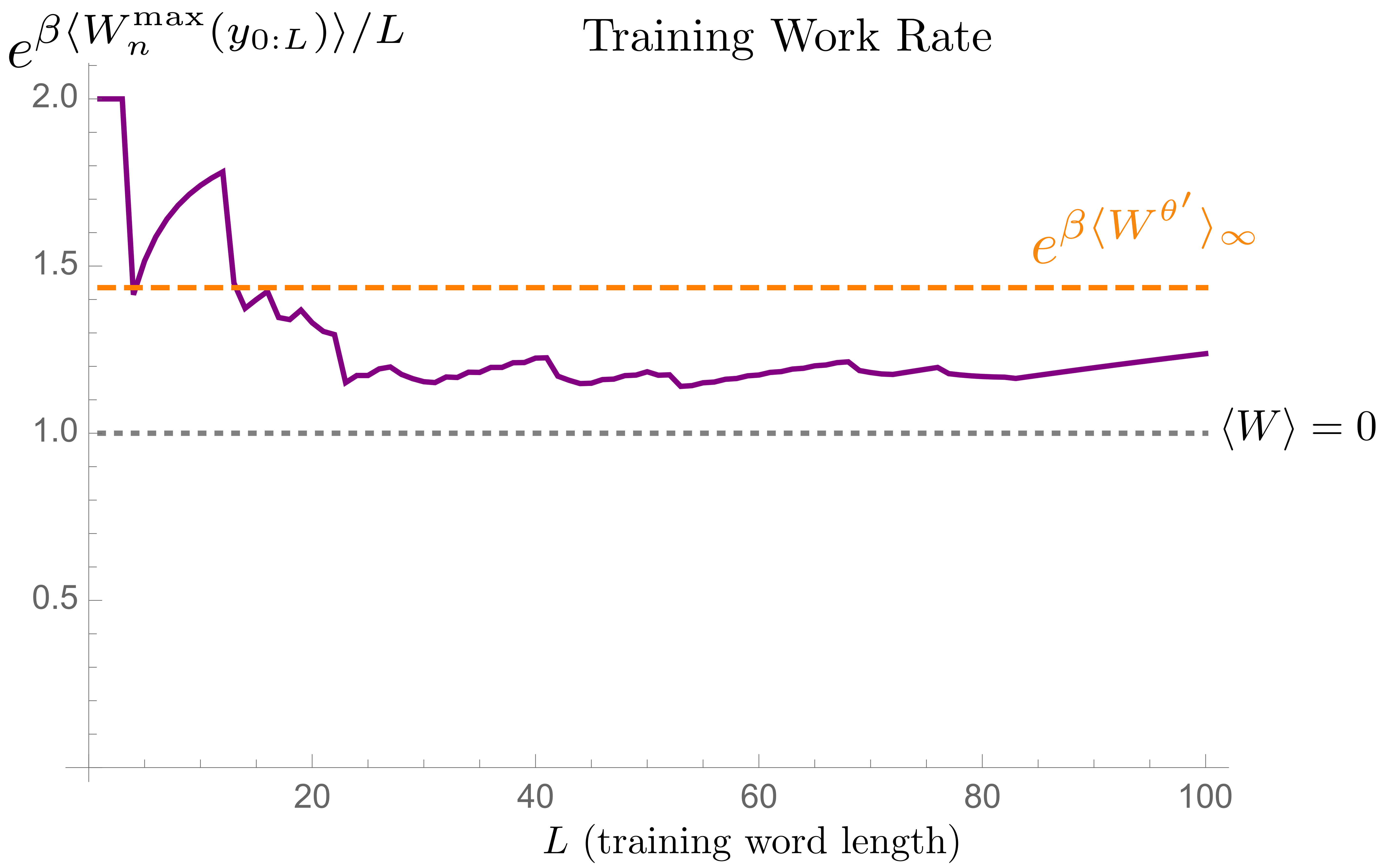
Finite time work rate: $\langle W_n^{\max}(y_{0:L}) \rangle / L$

Inverse temperature: $\beta = \dfrac{1}{k_B T}$

2-State Machines

Possible Models:

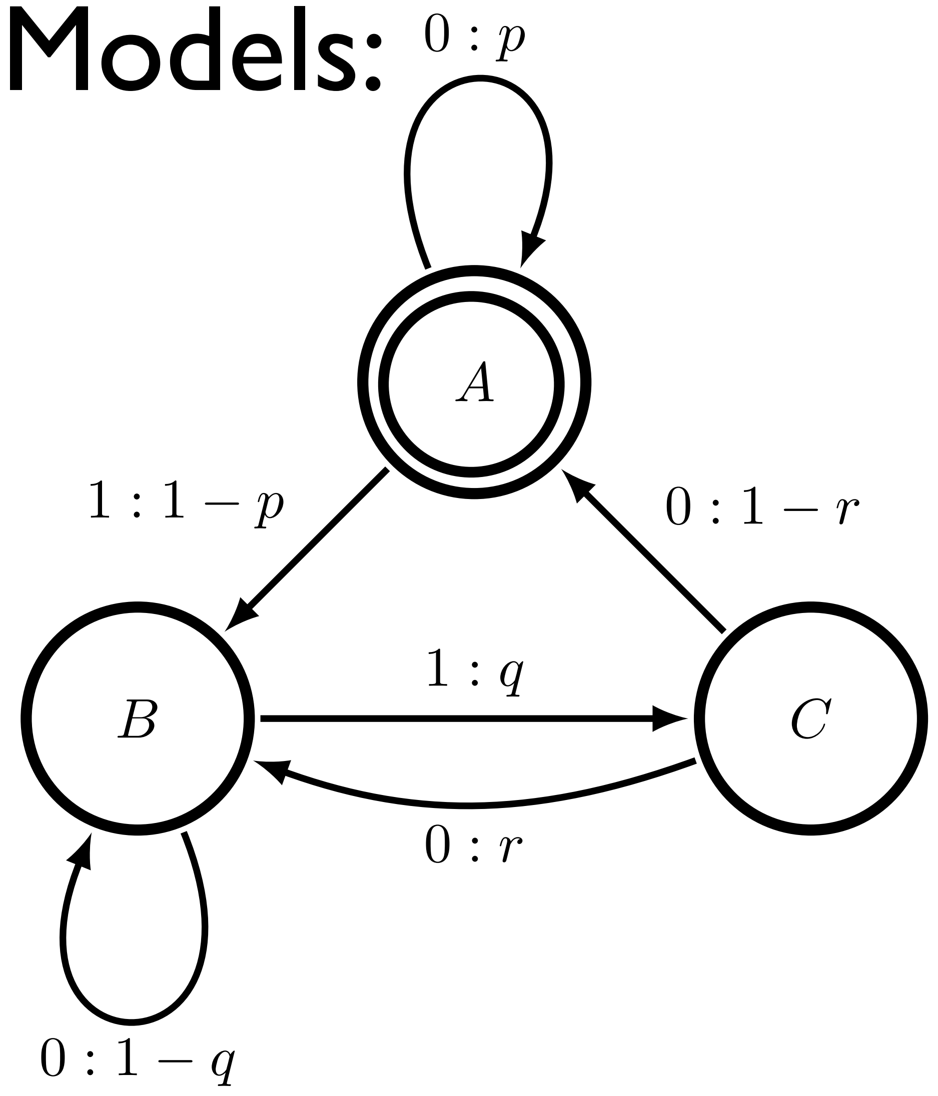


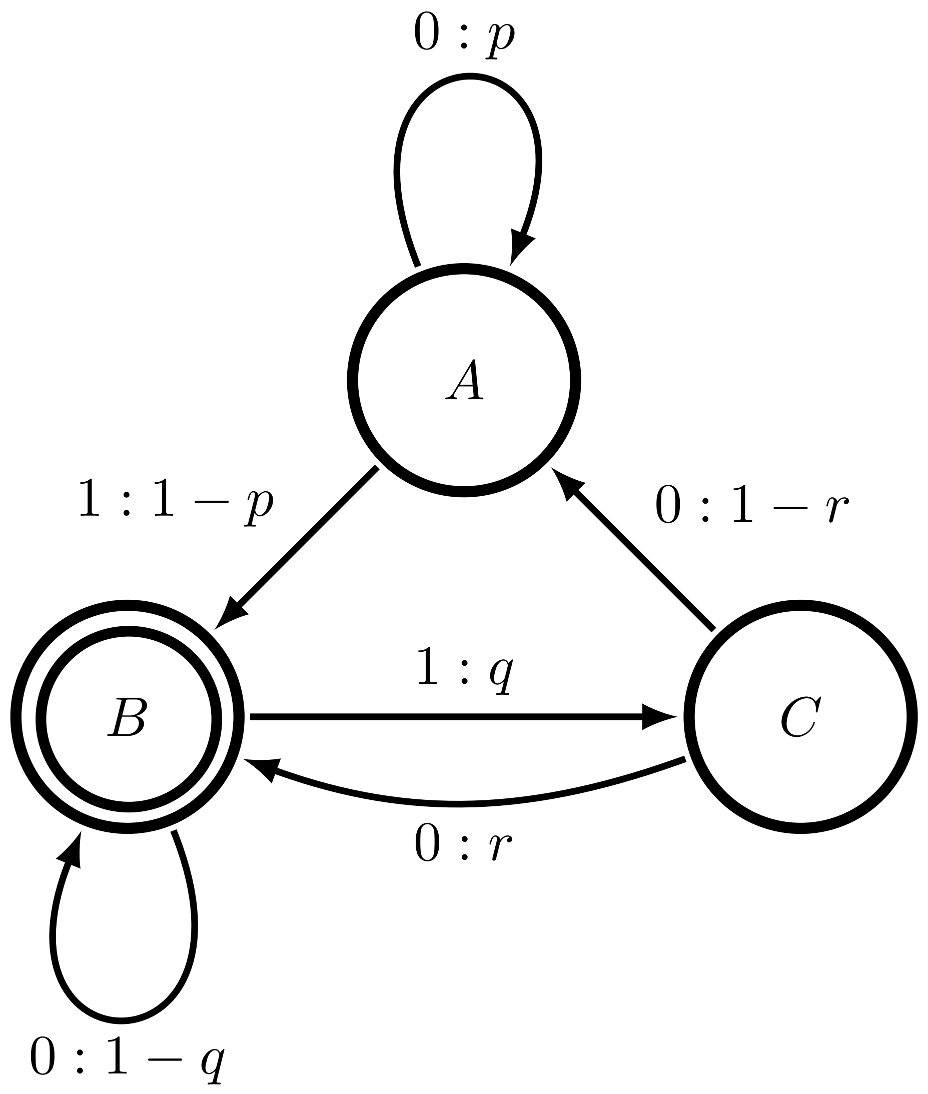


3-State Machines

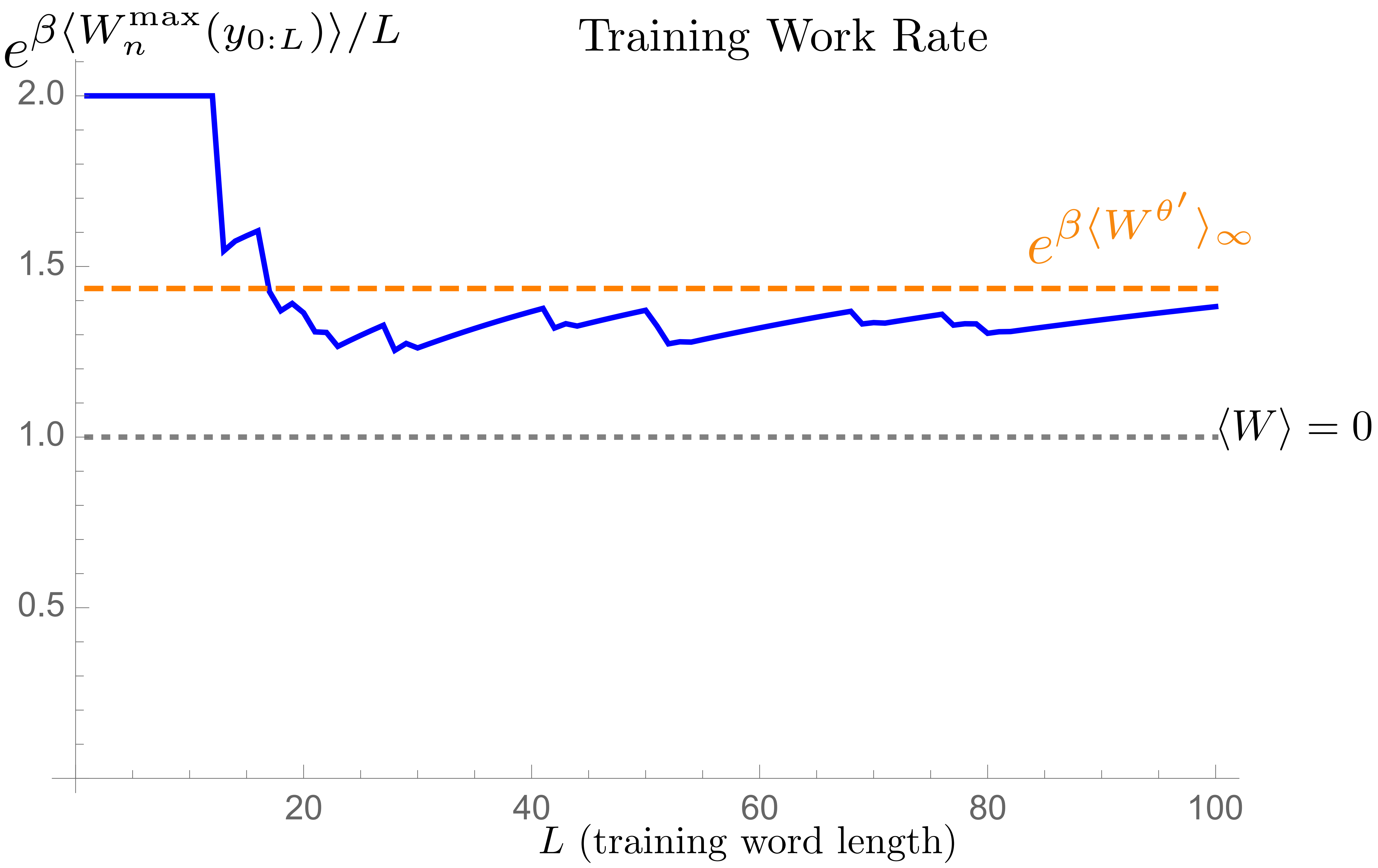
Possible Models:

$\theta \in \{$



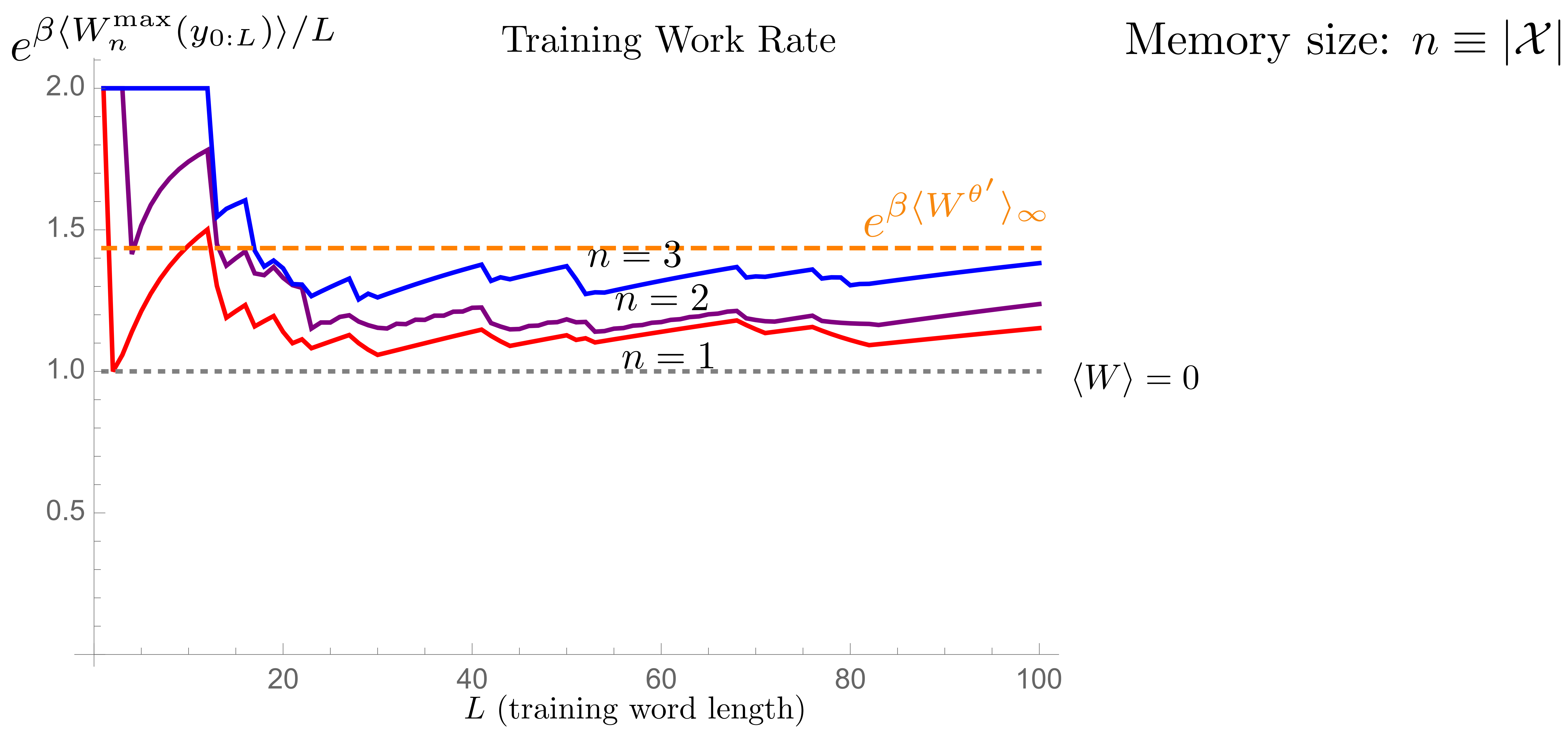


$\cdots \}$



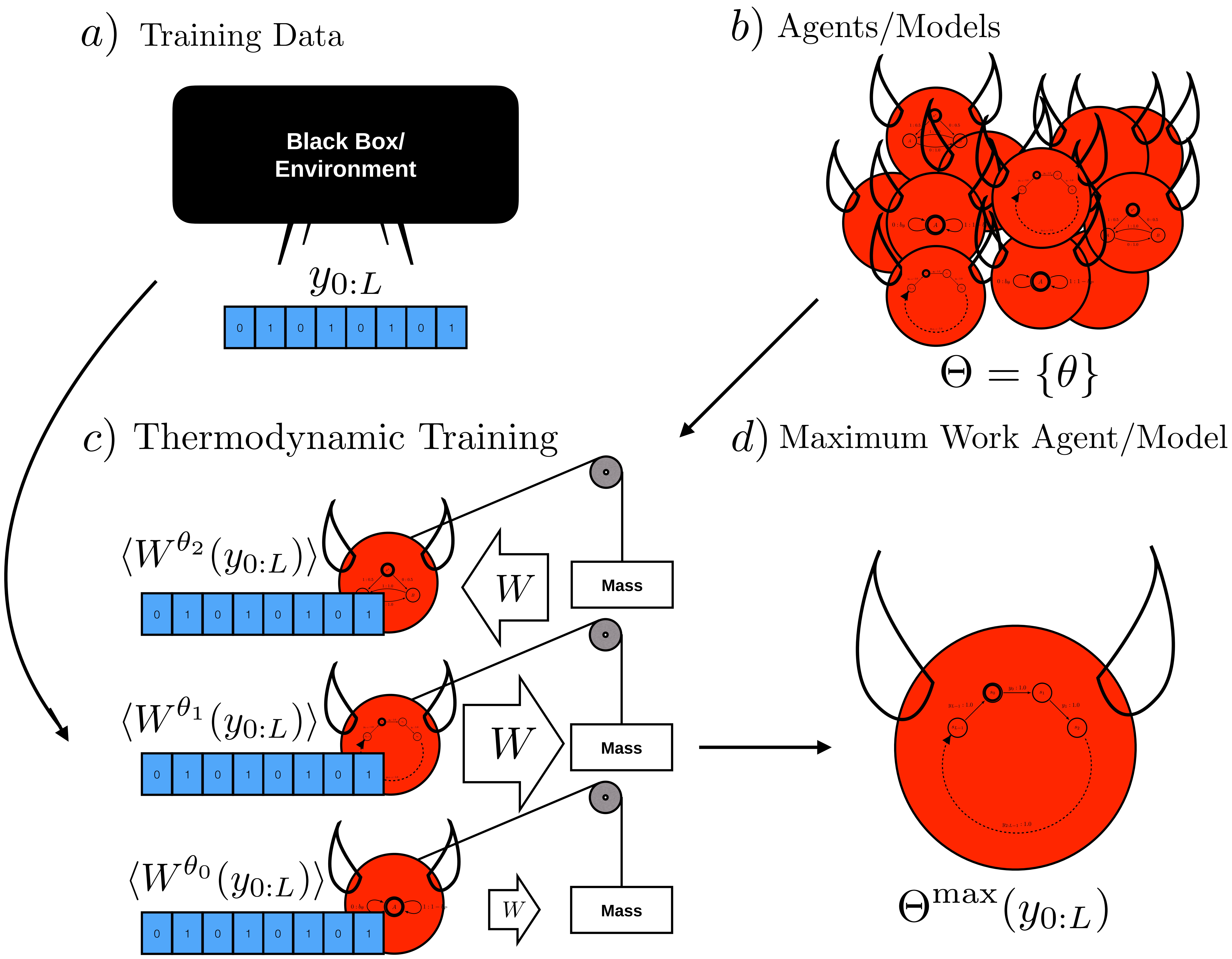
Benefit of Engine Memory

More complex engines can harvest more energy

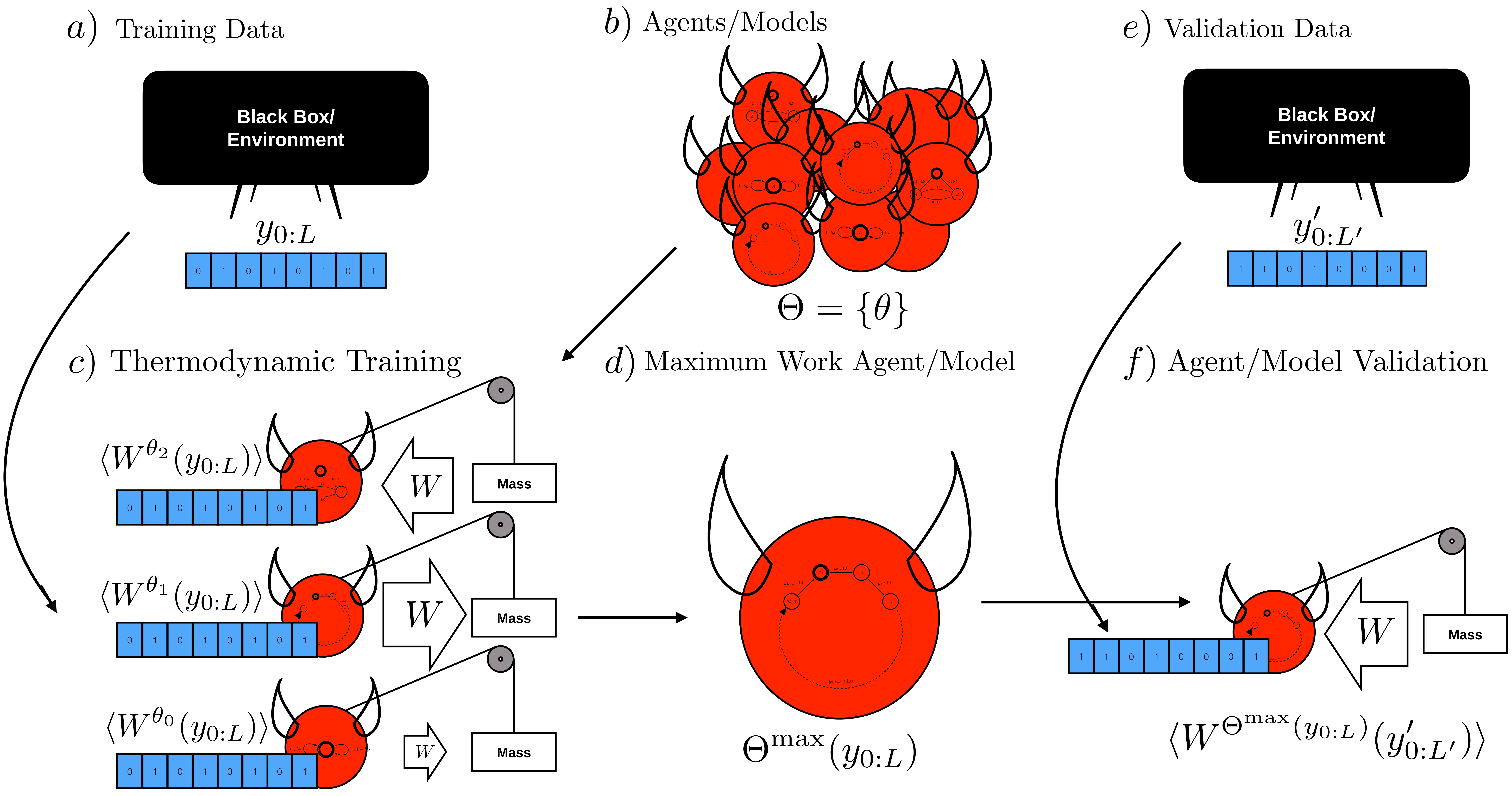


How well did the engine learn the pattern?

Thermodynamic Validation

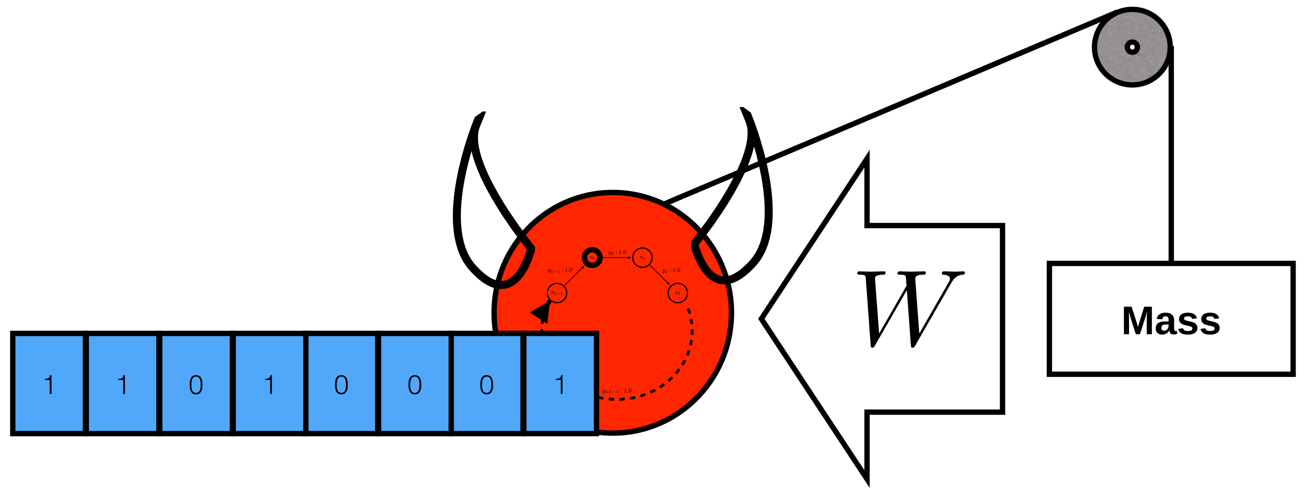


Thermodynamic Validation



Average Work and Dissipation

$f)$ Agent/Model Validation



$$\langle W^{\Theta^{\max}(y_{0:L})}(y'_{0:L'})\rangle$$

Average over input probabilities:

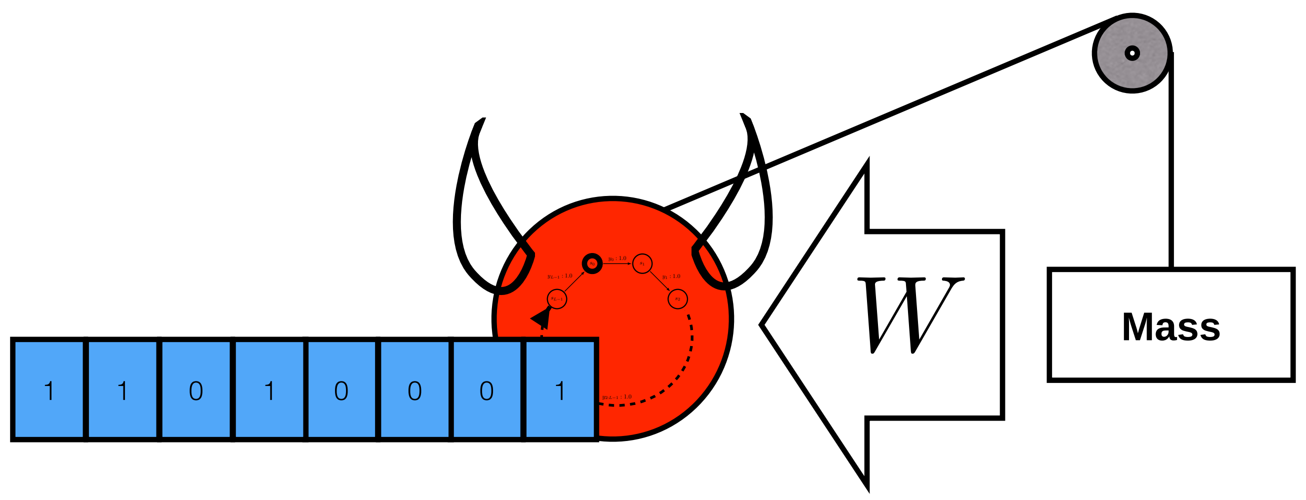
$$\frac{\langle W^\theta \rangle_{0:L}}{k_B T} = \sum_{y_{0:L}} \Pr(Y^{\theta'}_{0:L} = y_{0:L}) \frac{\langle W^\theta(y_{0:L})\rangle}{k_B T}$$

Asymptotic work rate as a validation measure:

$$\langle W^\theta \rangle_\infty \equiv \lim_{L \to \infty} \frac{\langle W^\theta \rangle_{0:L}}{L}$$

Average Work and Dissipation

f) Agent/Model Validation



$$\langle W^{\Theta^{\max}(y_{0:L})}(y'_{0:L'}) \rangle$$

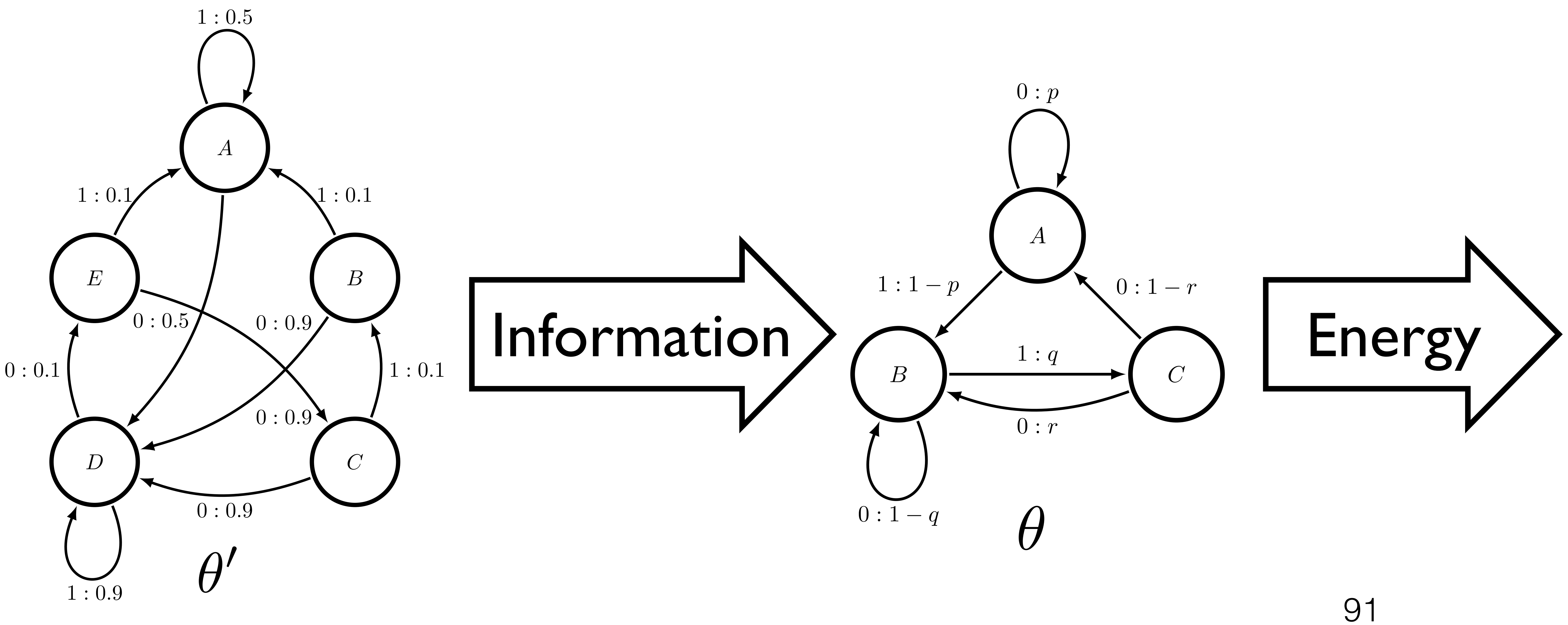
Average over input probabilities:

$$\frac{\langle W^{\theta} \rangle_{0:L}}{k_B T} = \sum_{y_{0:L}} \Pr(Y^{\theta'}_{0:L} = y_{0:L}) \frac{\langle W^{\theta}(y_{0:L}) \rangle}{k_B T}$$

Asymptotic work rate as a validation measure:

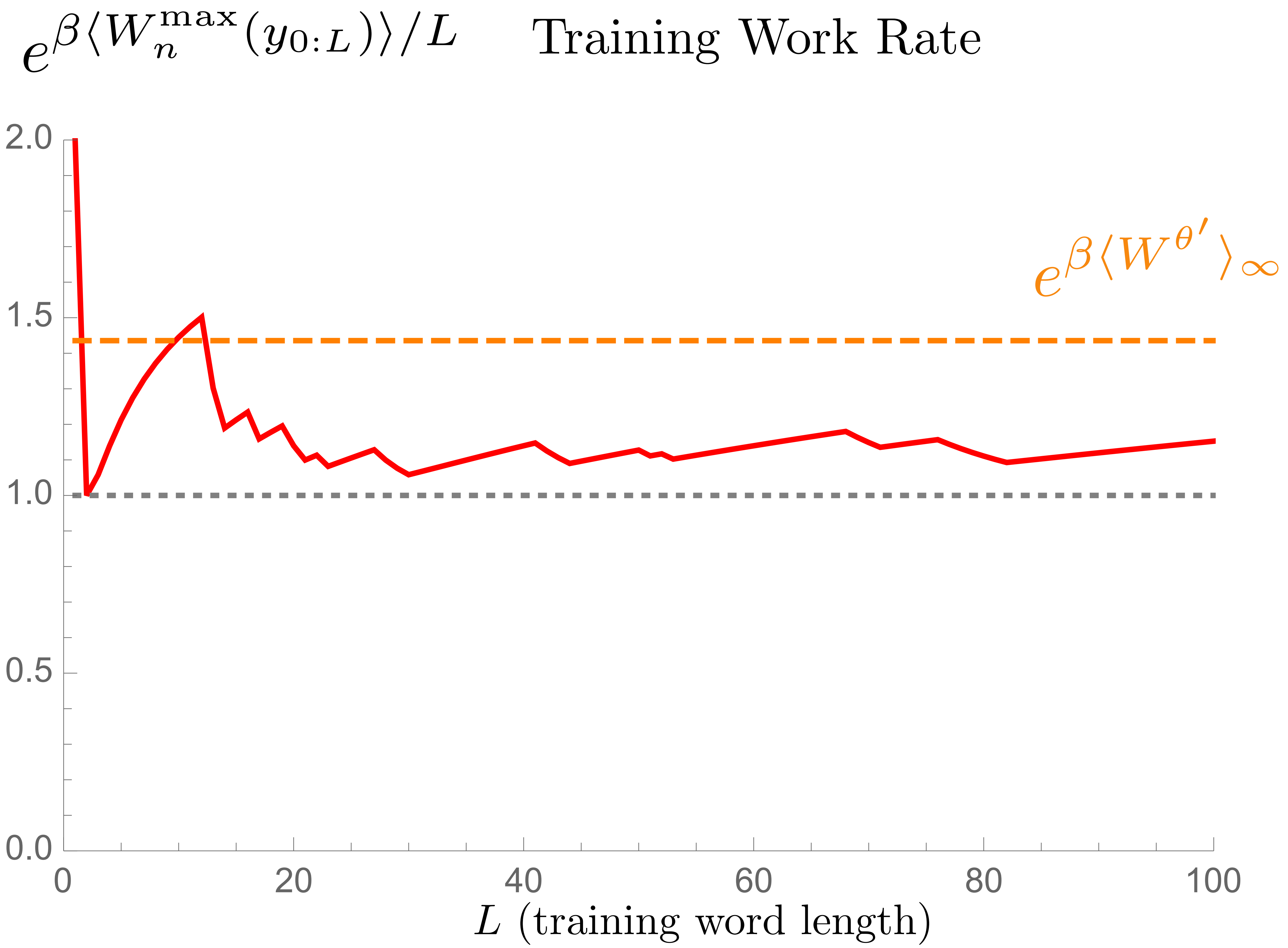
$$\langle W^{\theta} \rangle_{\infty} \equiv \lim_{L \to \infty} \frac{\langle W^{\theta} \rangle_{0:L}}{L}$$

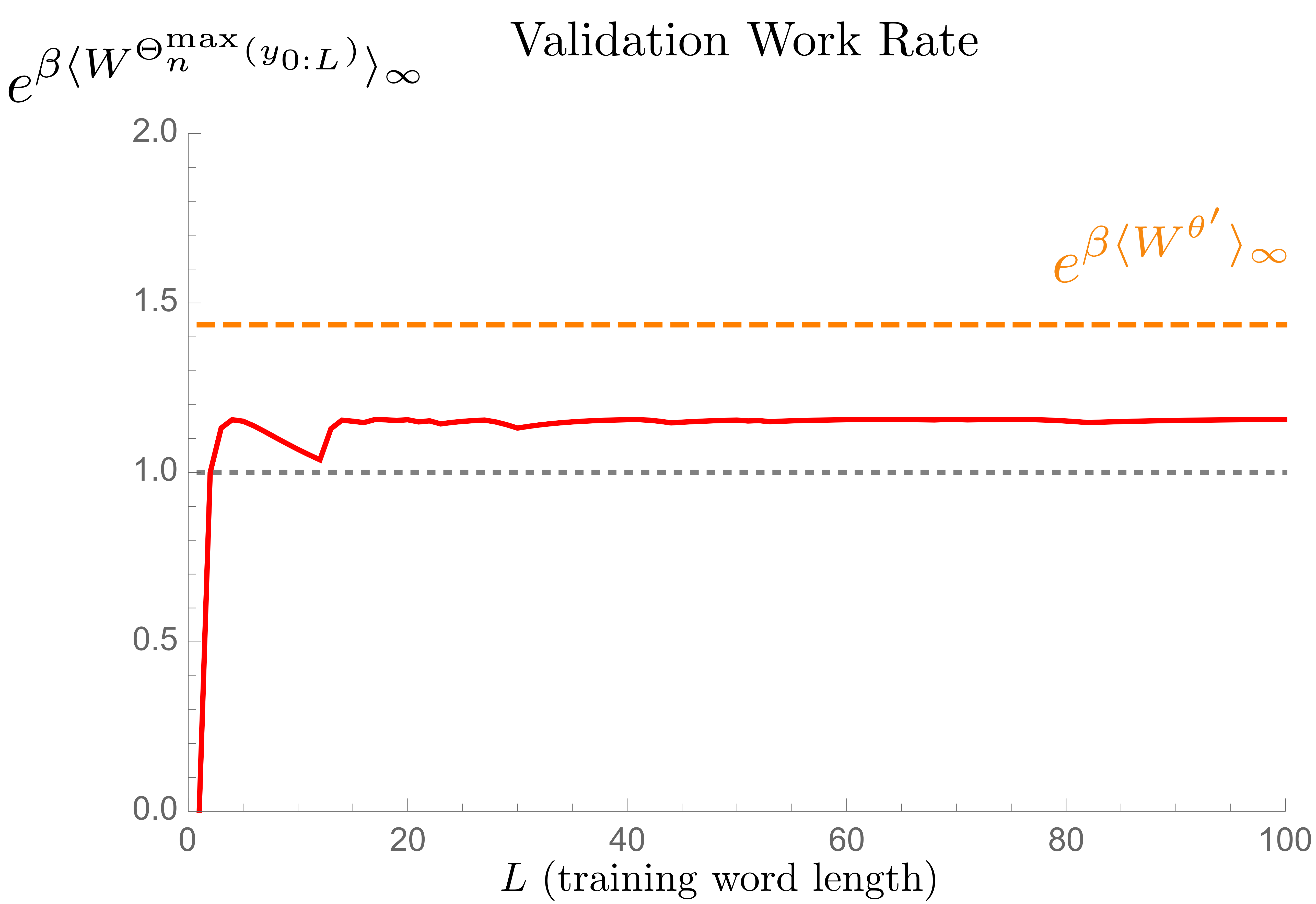
Determines long-term effectiveness of engine on input



1-State Validation

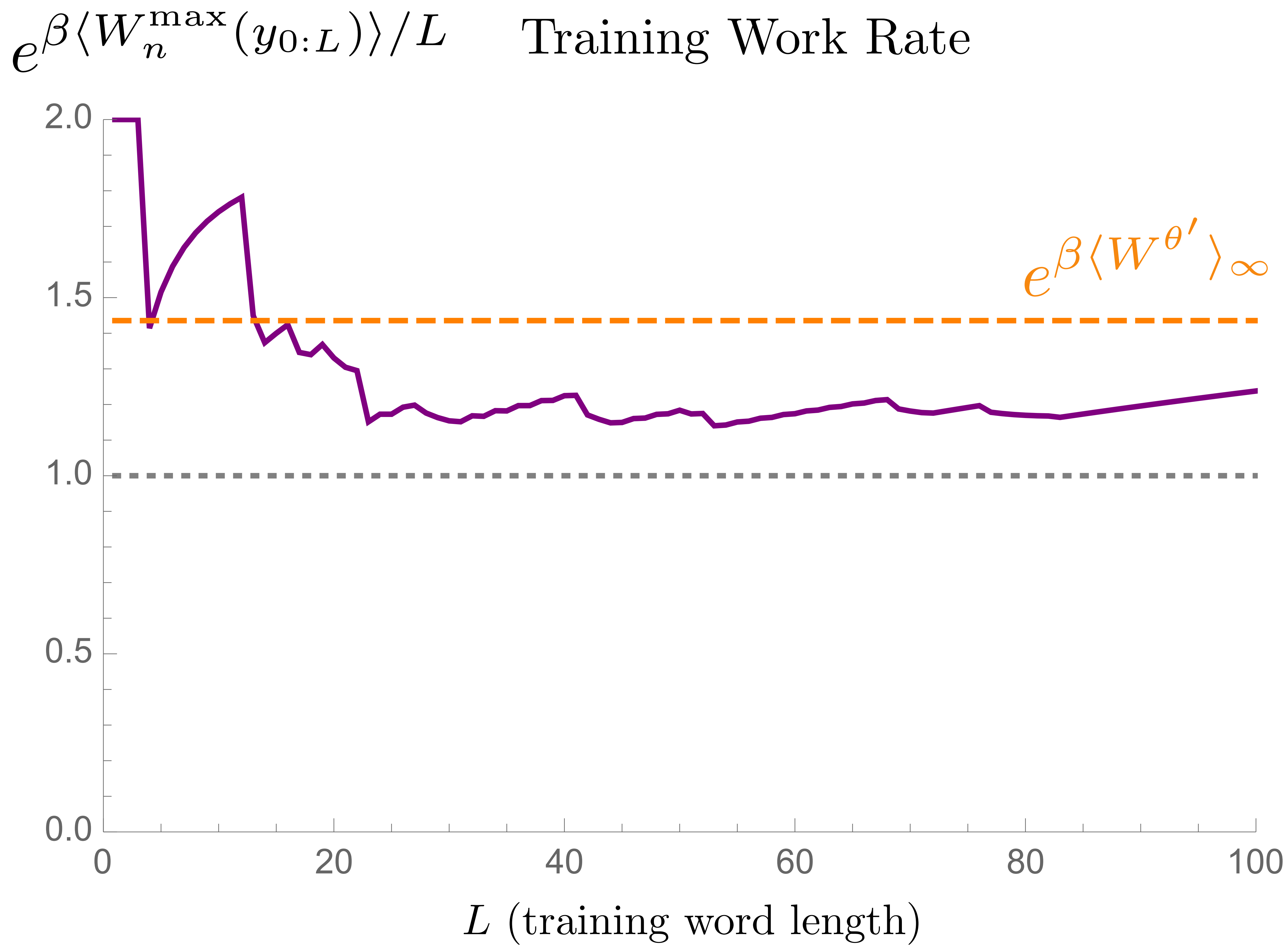
Excessively high training work rates lead to divergent in validation.
This is **thermodynamic overfitting**!

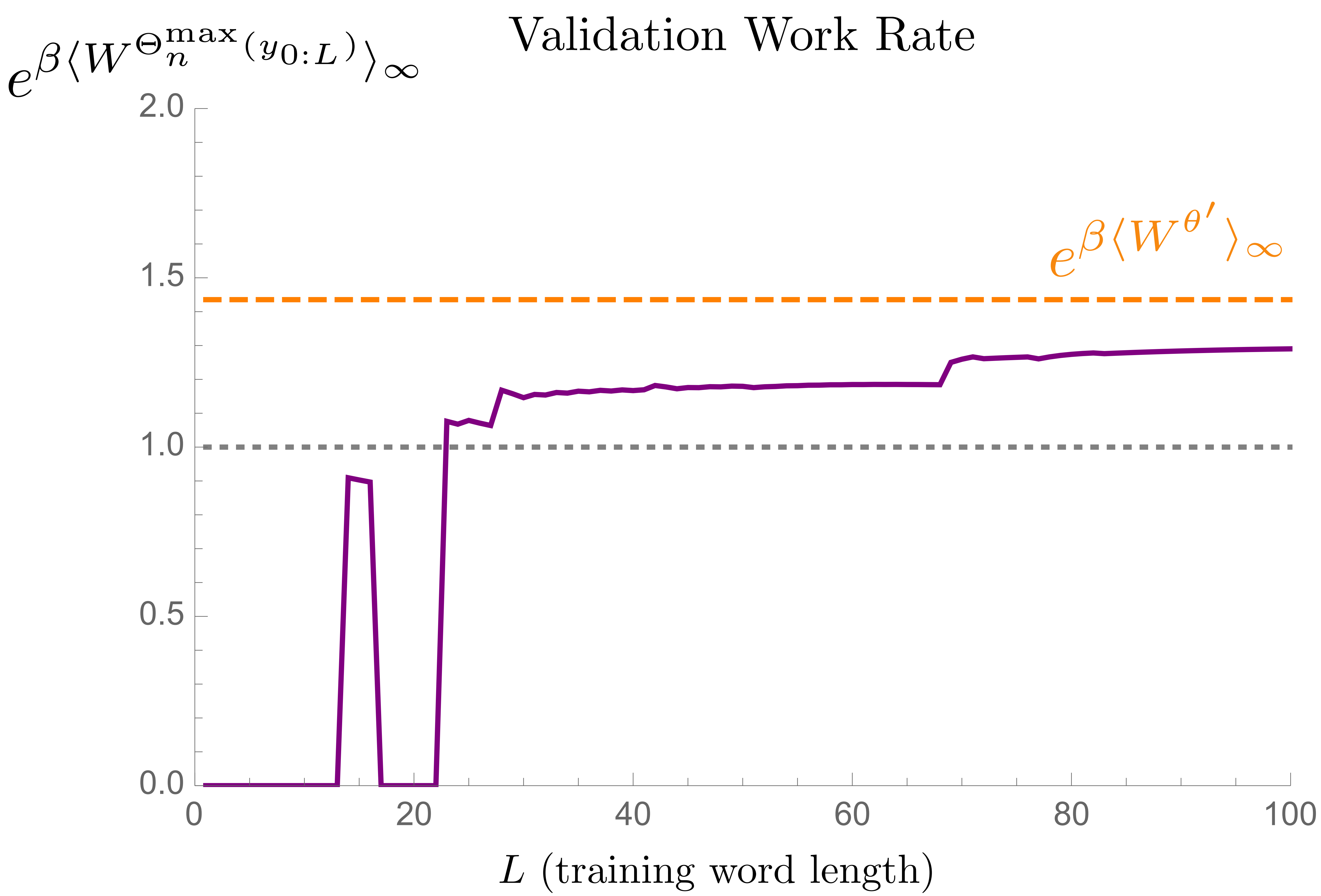




2-State Validation

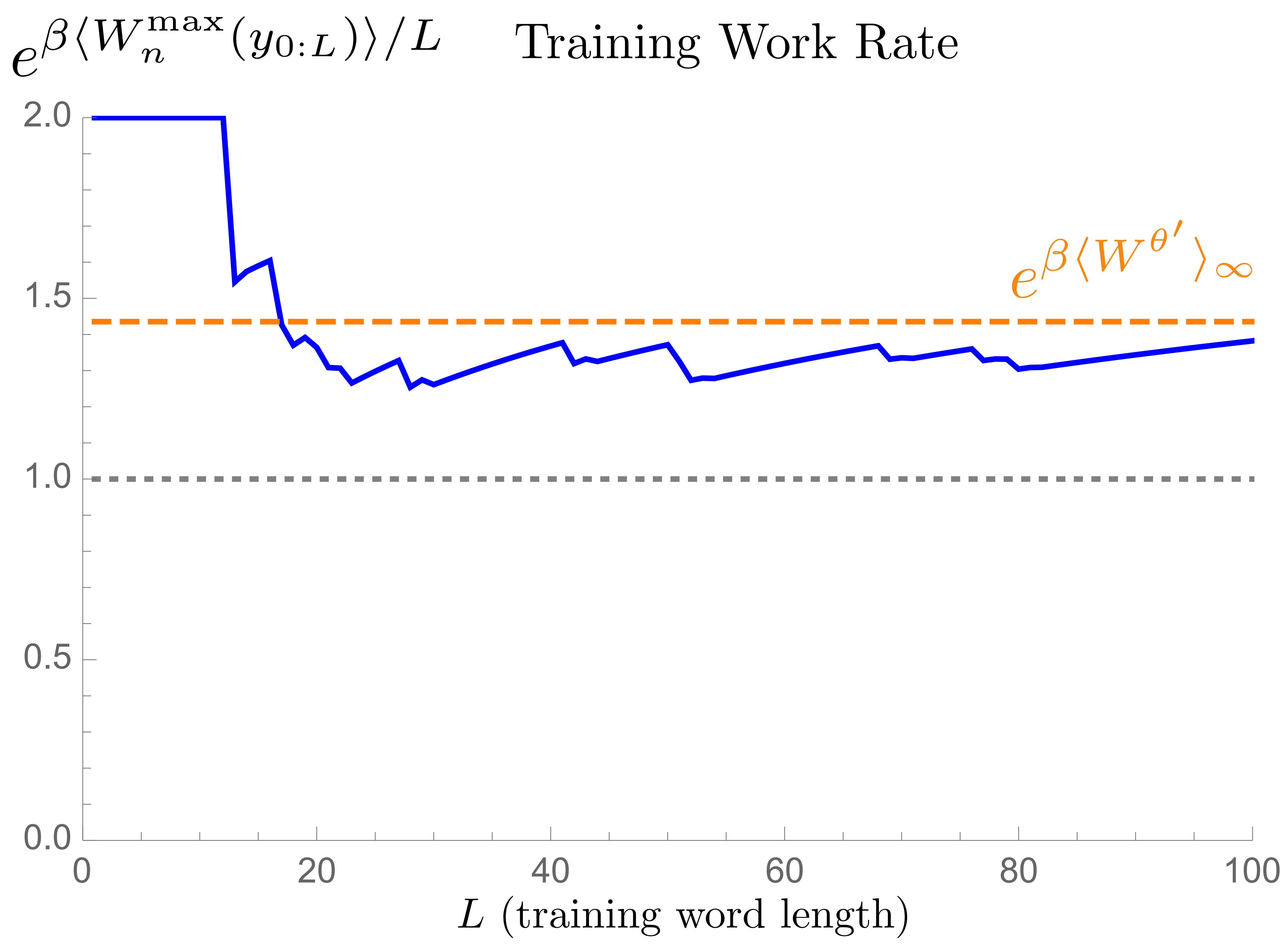
More frequent overfitting.

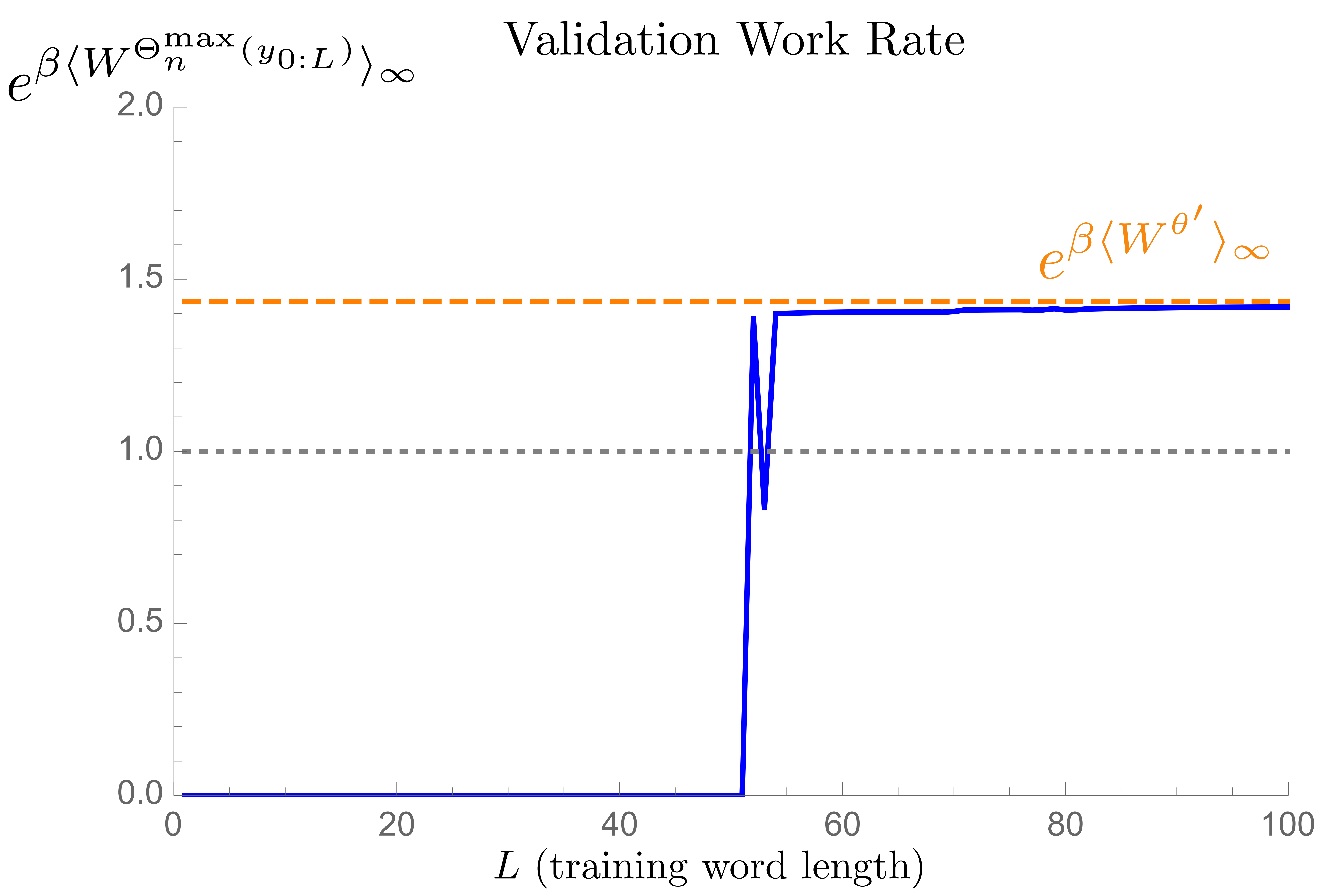




3-State Validation

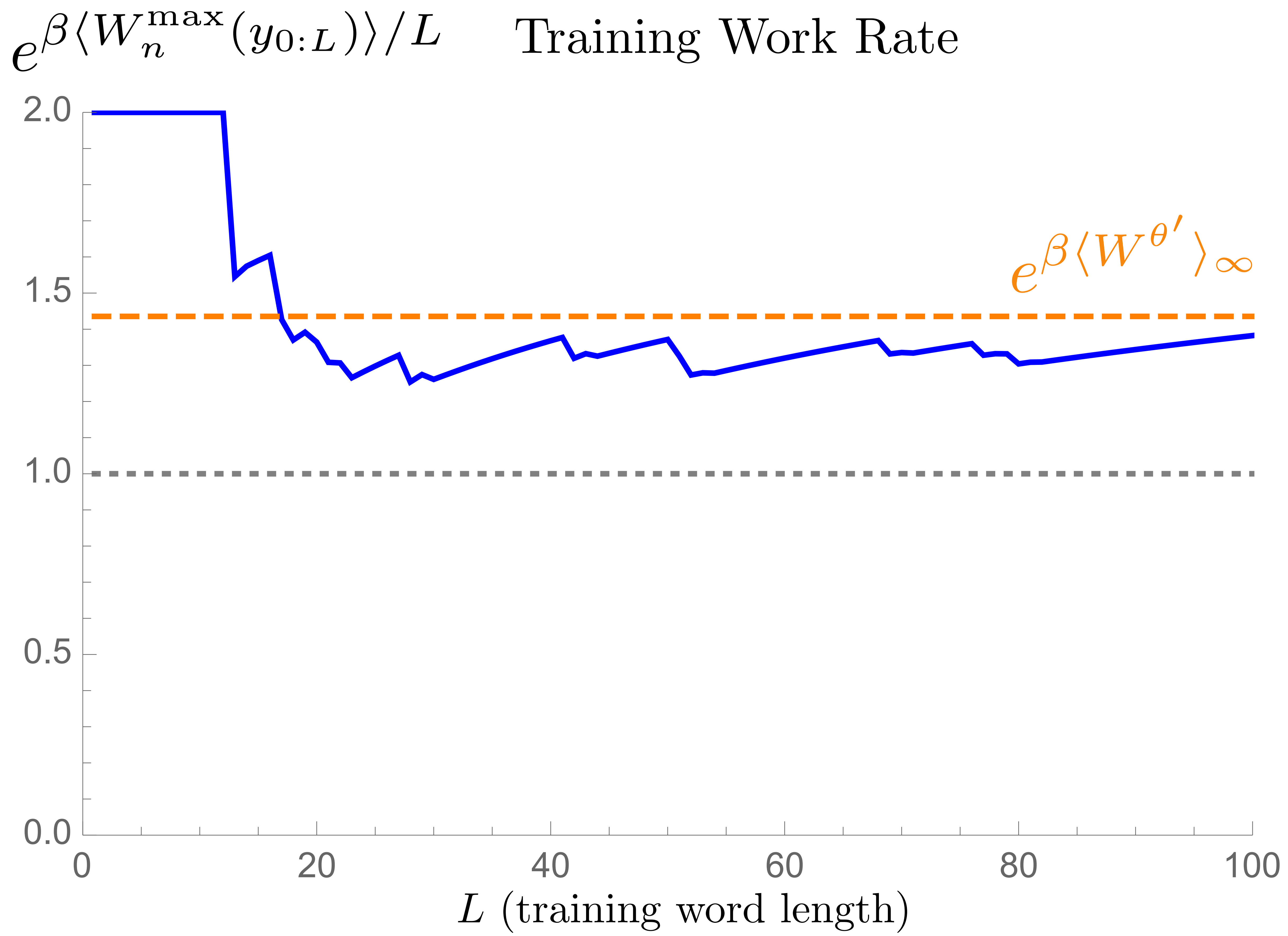
Even more frequent overfitting.

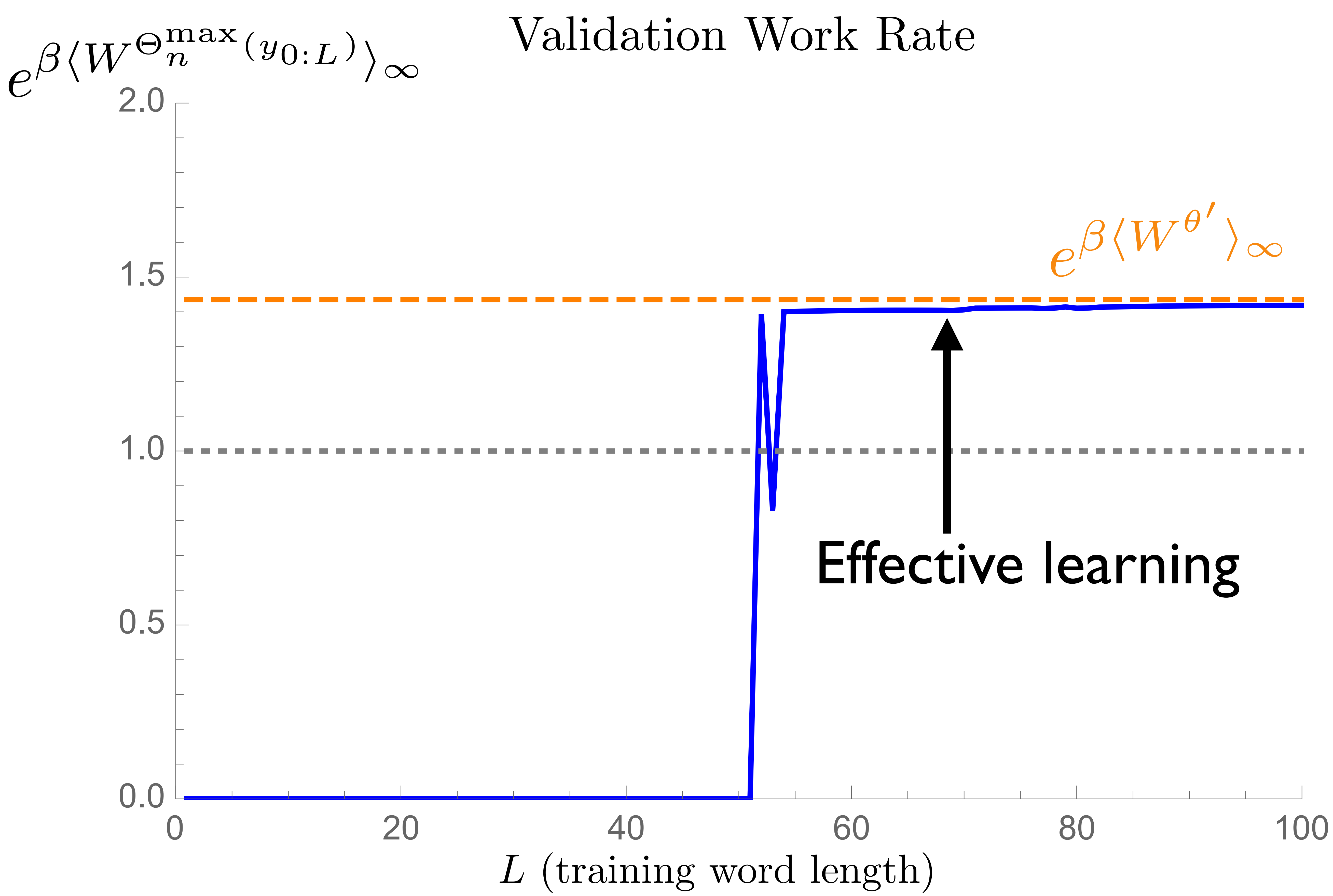




3-State Validation

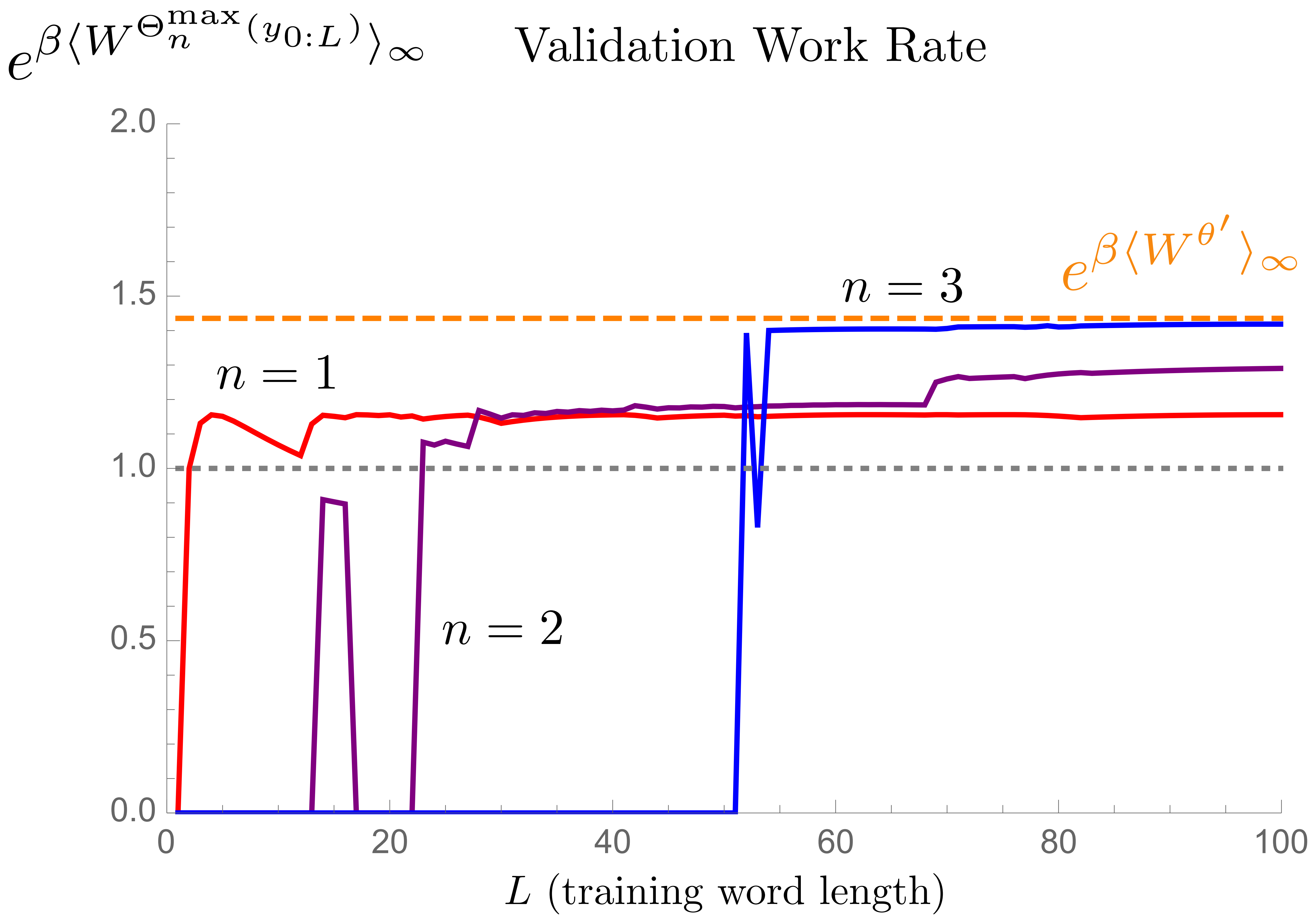
Even more frequent overfitting.





Training Vs Validation

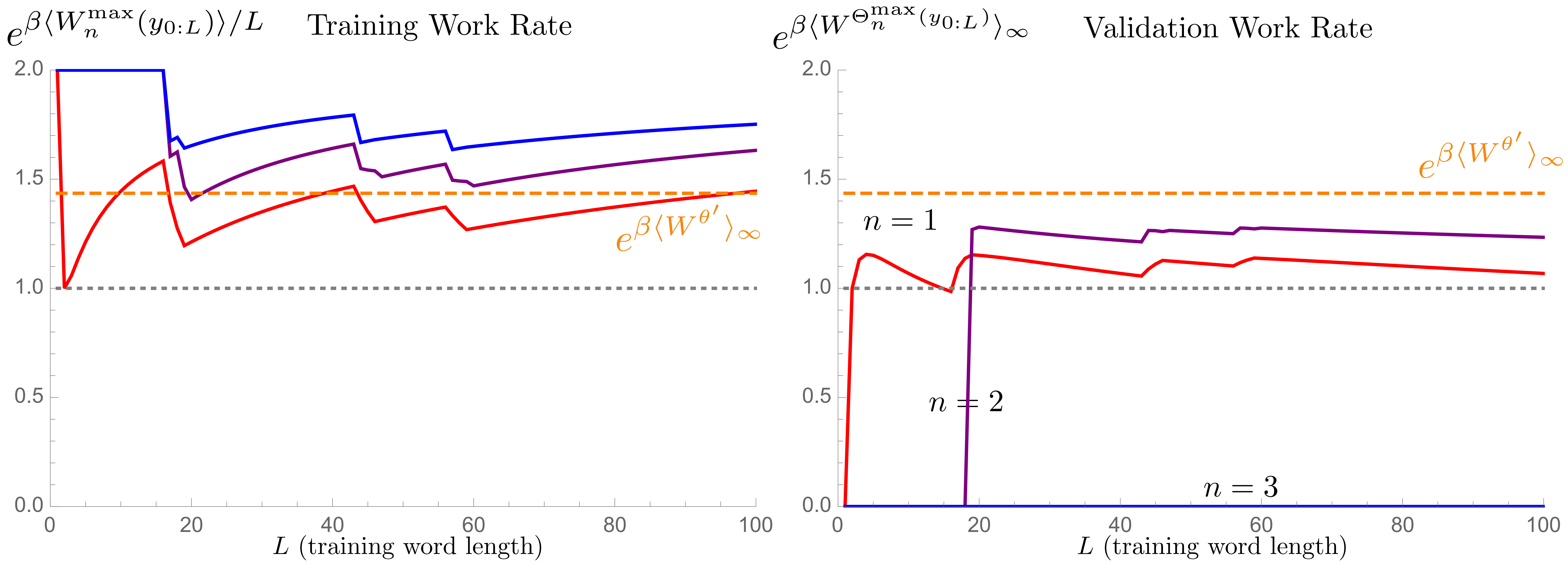
Memory offers a universal benefit in training, but often leads to divergent dissipation in validation.



Benefit of memory can only be found for large training sets.

Training Vs Validation

Memory offers a universal benefit in training, but often leads to divergent dissipation in validation.

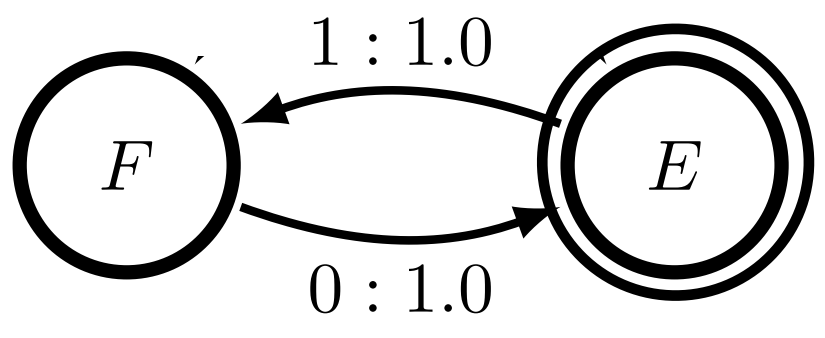


Benefit of memory can only be found for large training sets.

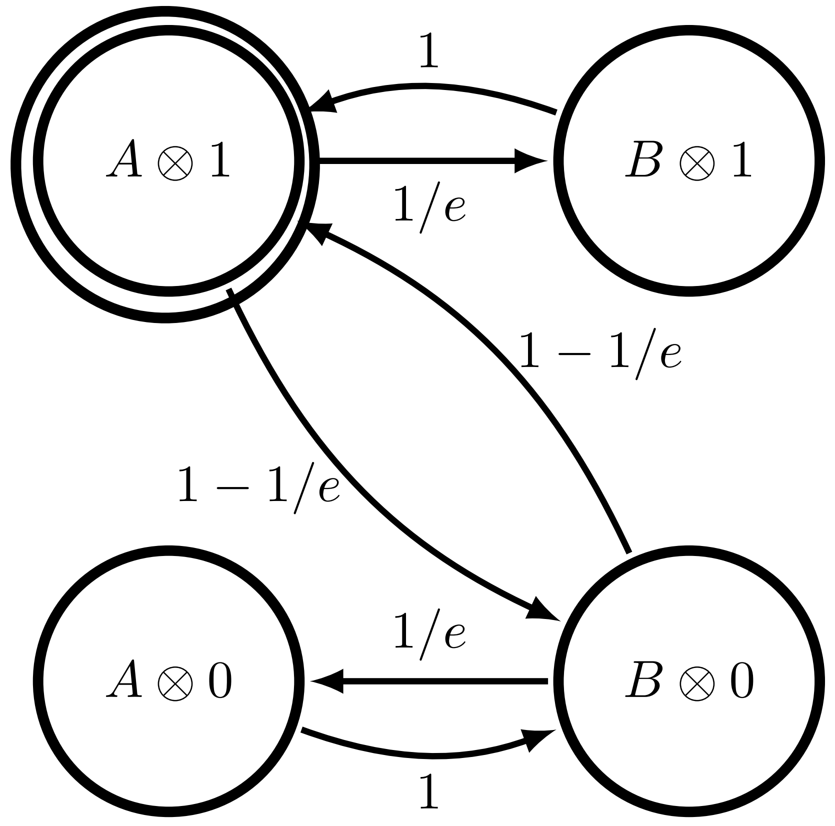
Correlation Powered Information Engines

Alexander B. Boyd, Dibyendu Mandal, and James P. Crutchfield
Phys. Rev. E **95**, 012152 (2017)

Period 2 Input



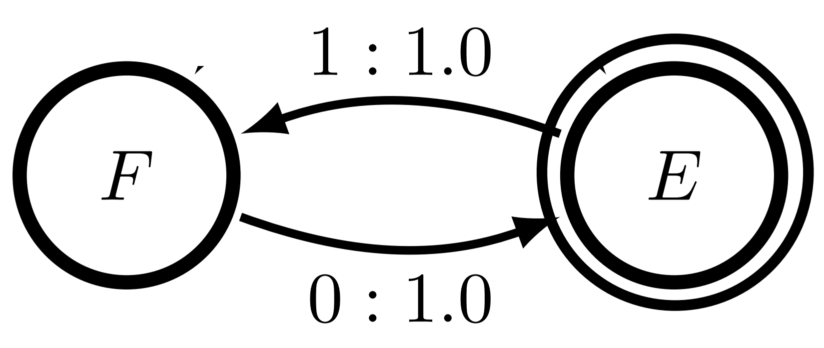
Candidate Ratchet



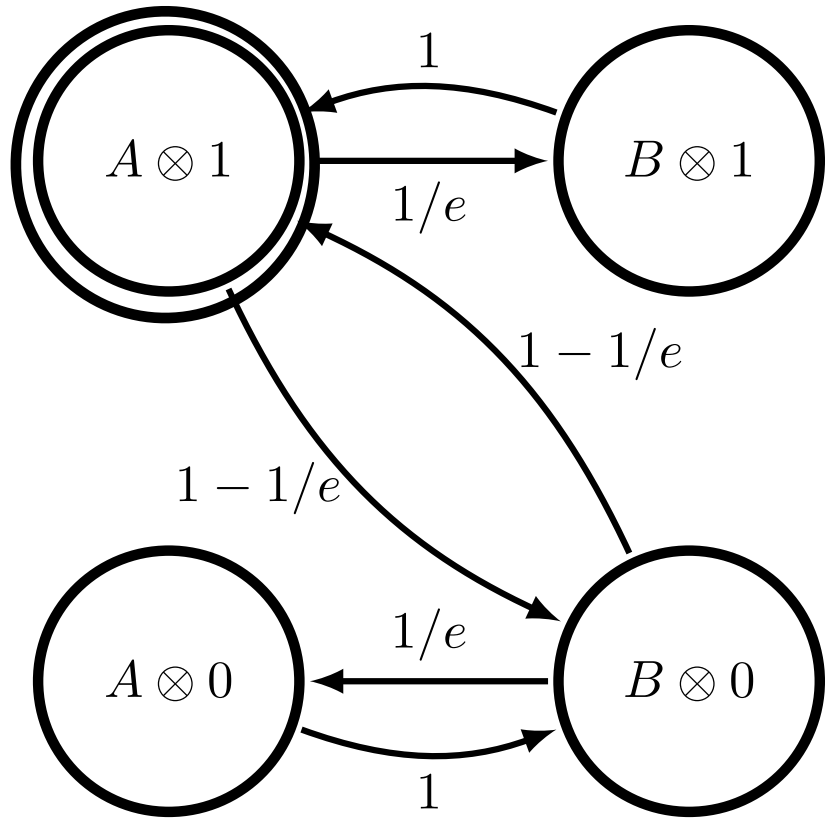
Correlation Powered Information Engines

Alexander B. Boyd, Dibyendu Mandal, and James P. Crutchfield
Phys. Rev. E **95**, 012152 (2017)

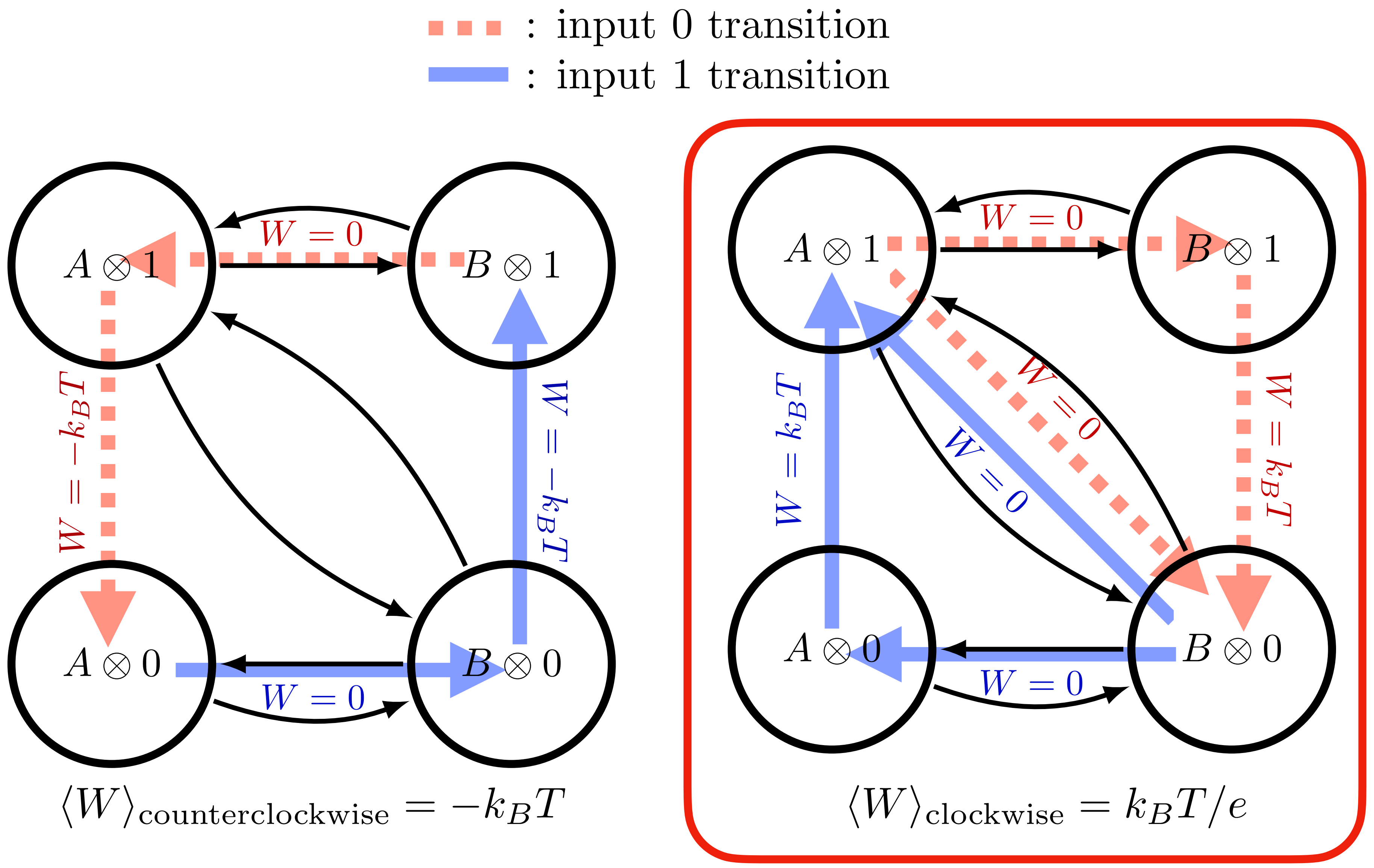
Period 2 Input



Candidate Ratchet



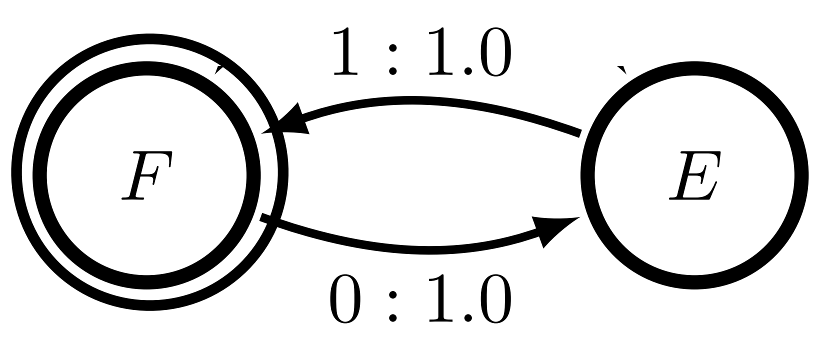
Ratchet Work Production



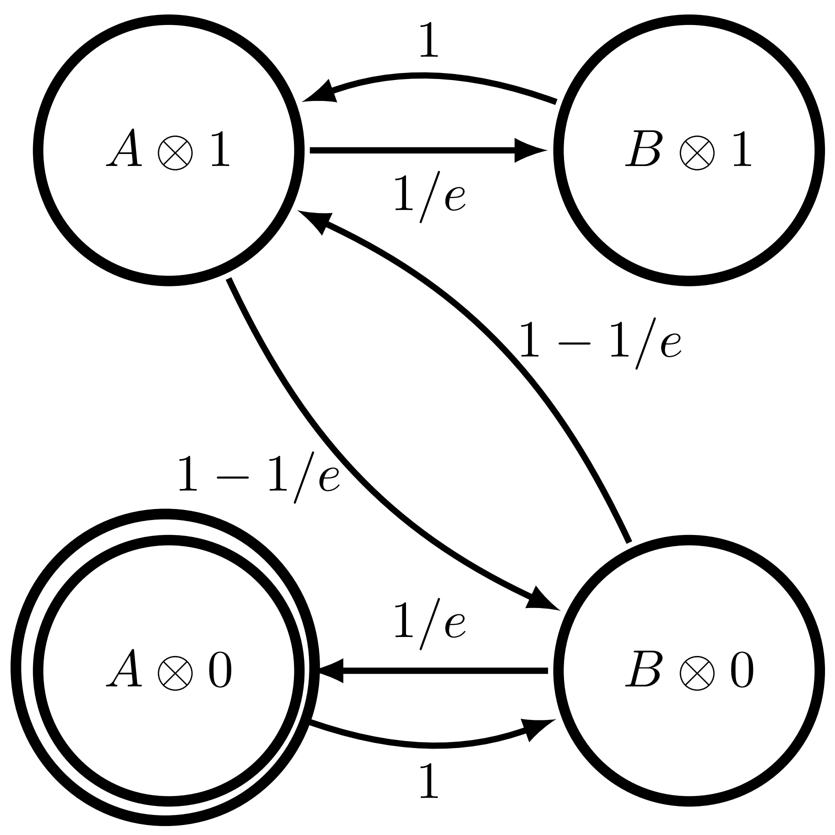
Correlation Powered Information Engines

Alexander B. Boyd, Dibyendu Mandal, and James P. Crutchfield
Phys. Rev. E **95**, 012152 (2017)

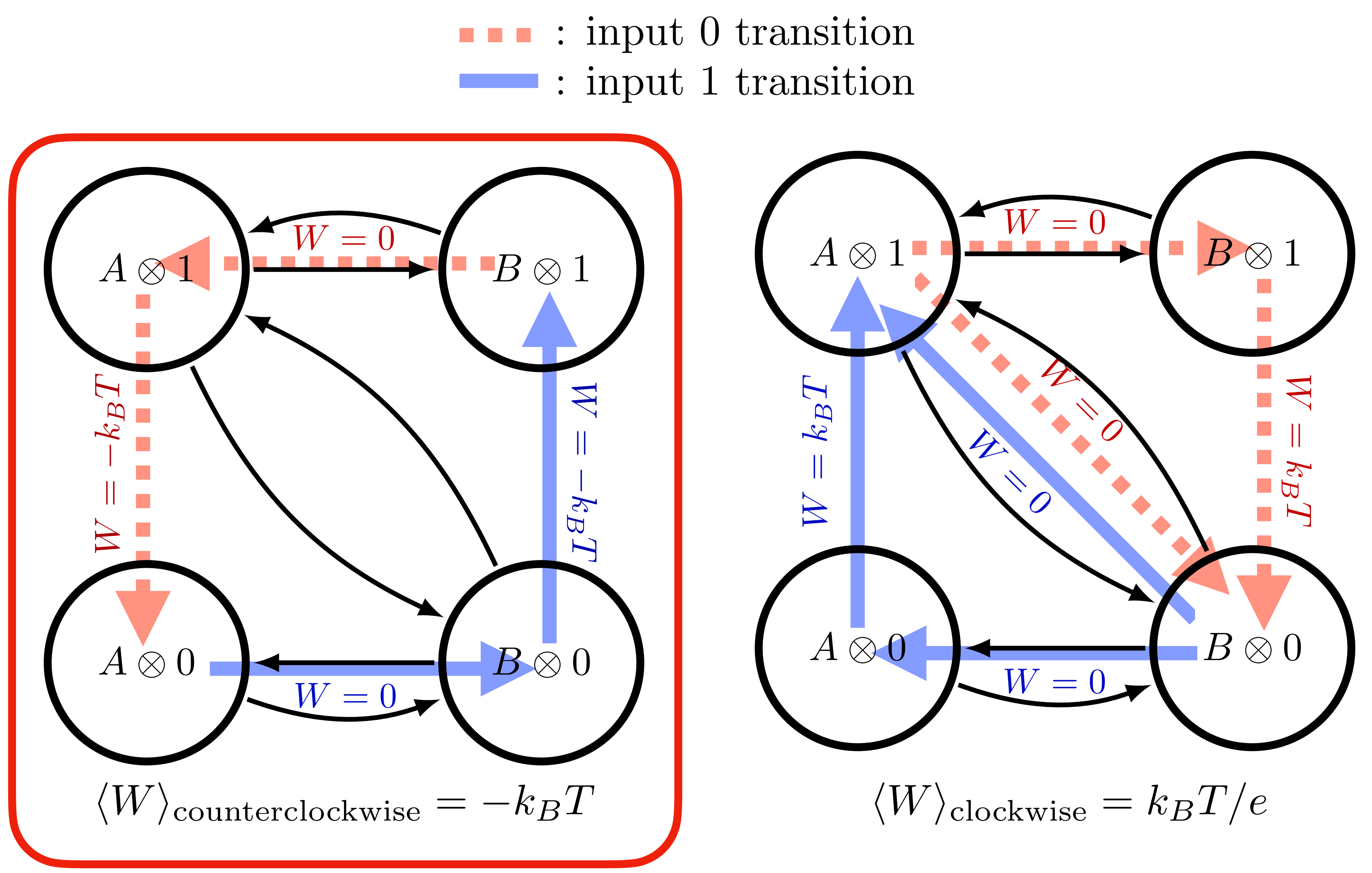
Period 2 Input



Candidate Ratchet



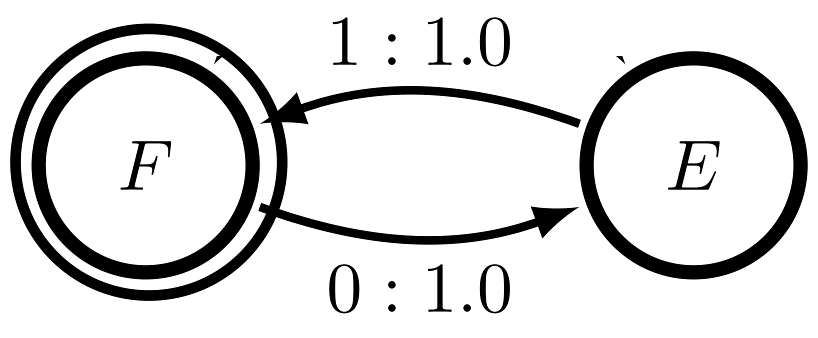
Ratchet Work Production



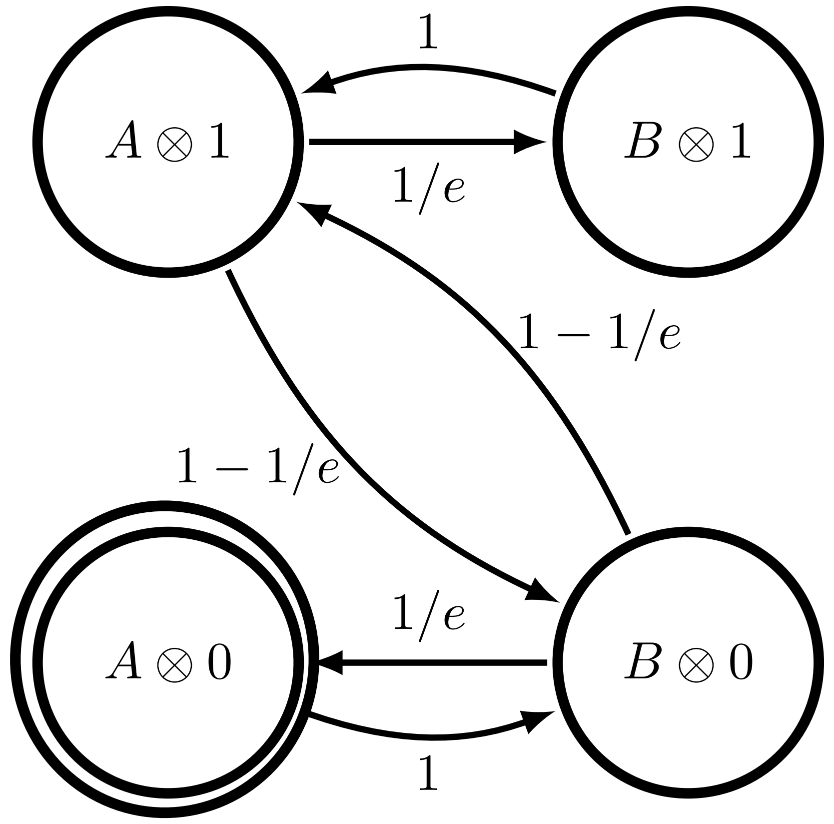
Correlation Powered Information Engines

Alexander B. Boyd, Dibyendu Mandal, and James P. Crutchfield
Phys. Rev. E **95**, 012152 (2017)

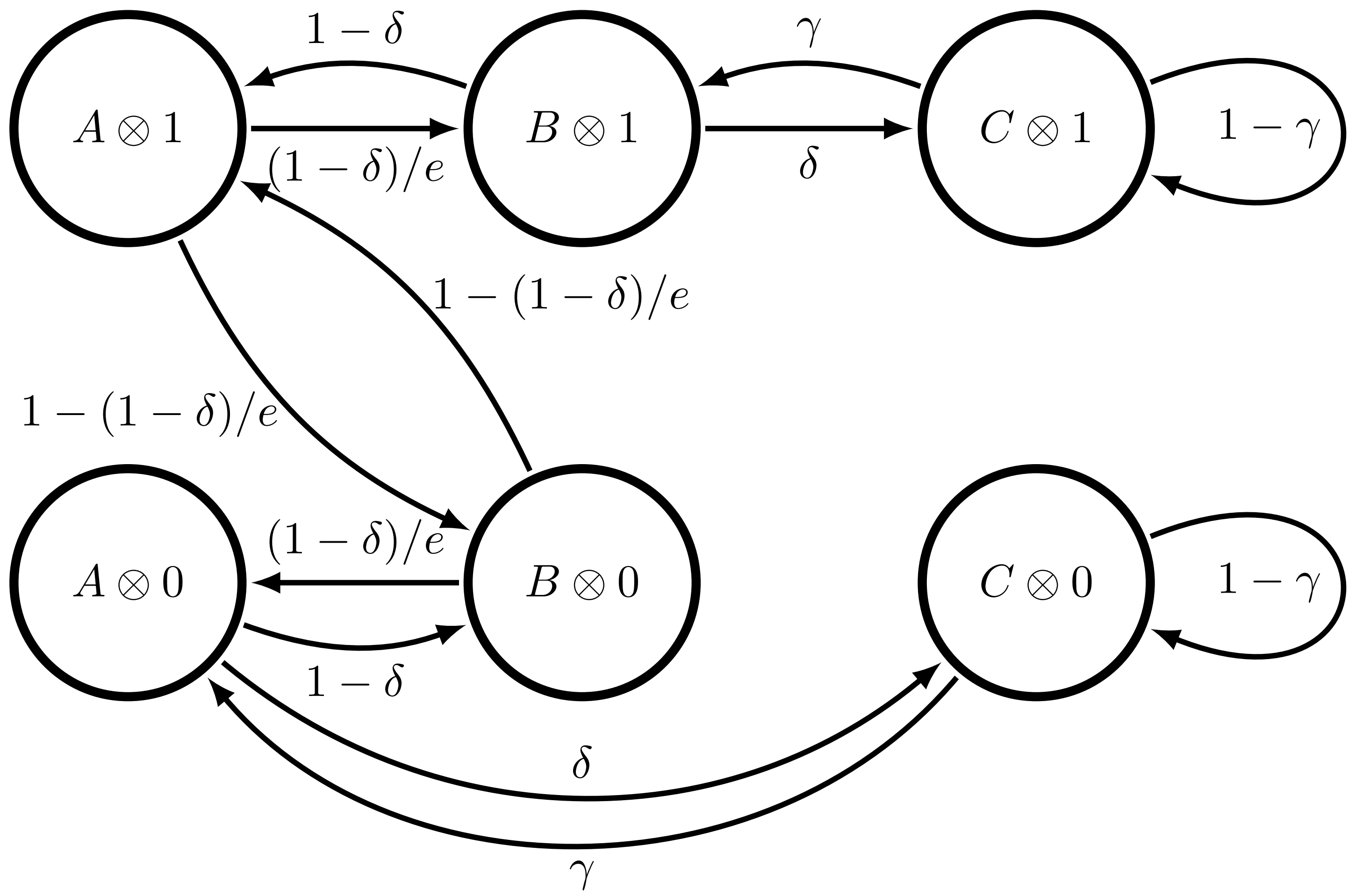
Period 2 Input



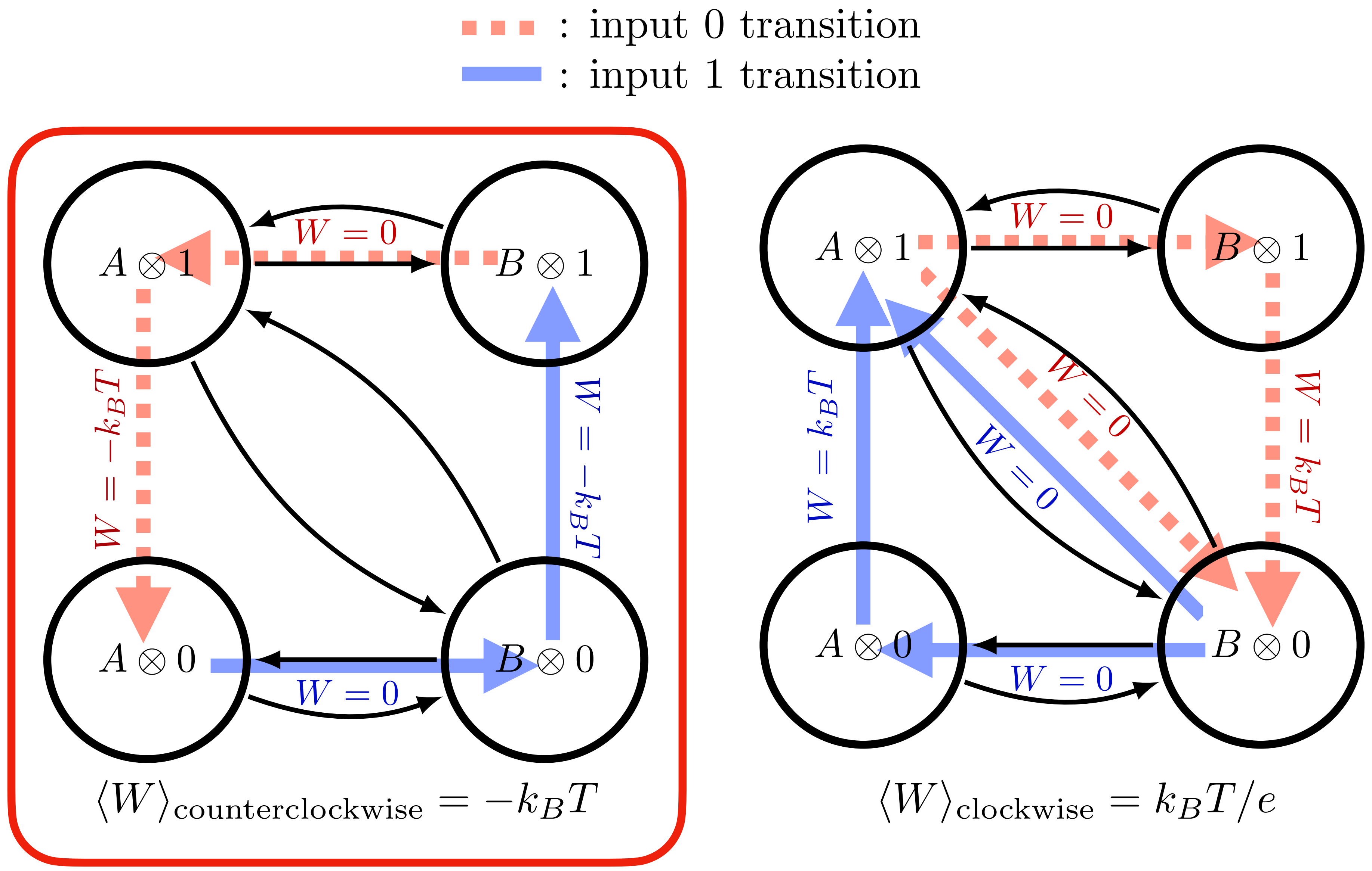
Candidate Ratchet



Must include synchronization mechanism



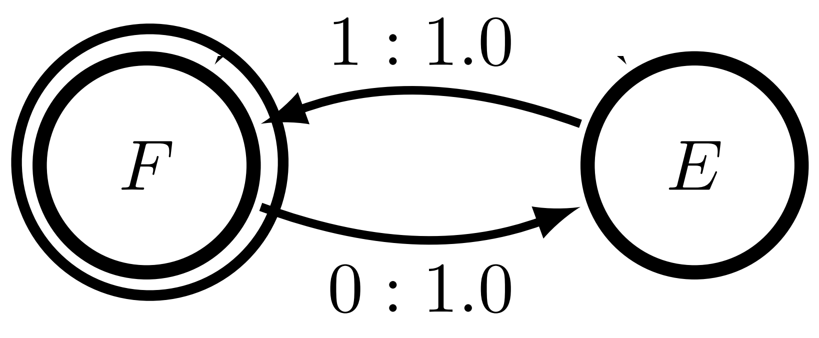
Ratchet Work Production



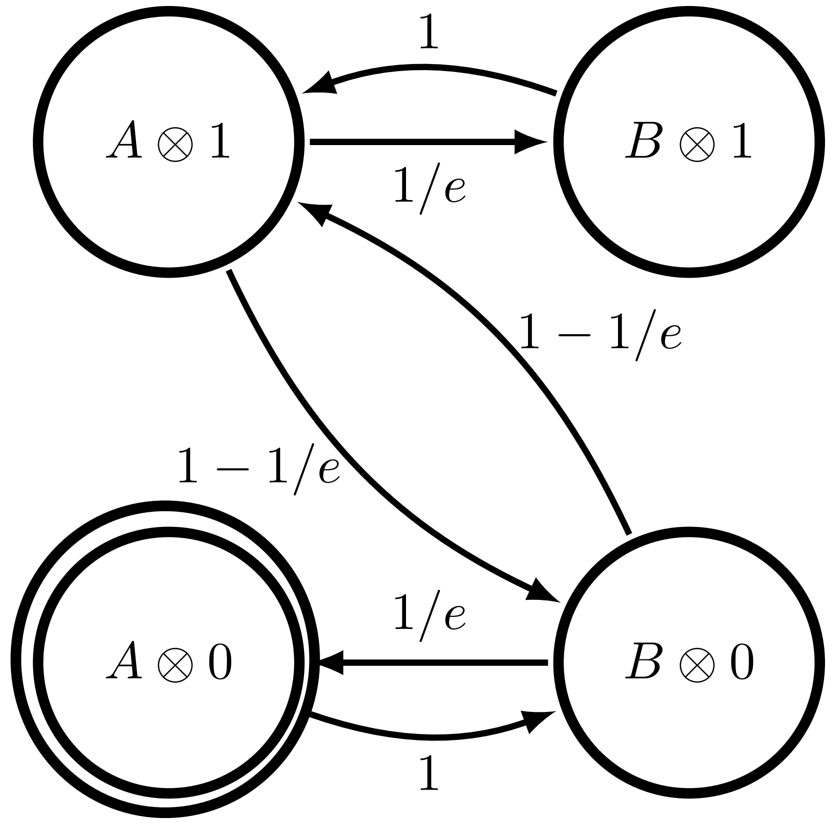
Correlation Powered Information Engines

Alexander B. Boyd, Dibyendu Mandal, and James P. Crutchfield
Phys. Rev. E **95**, 012152 (2017)

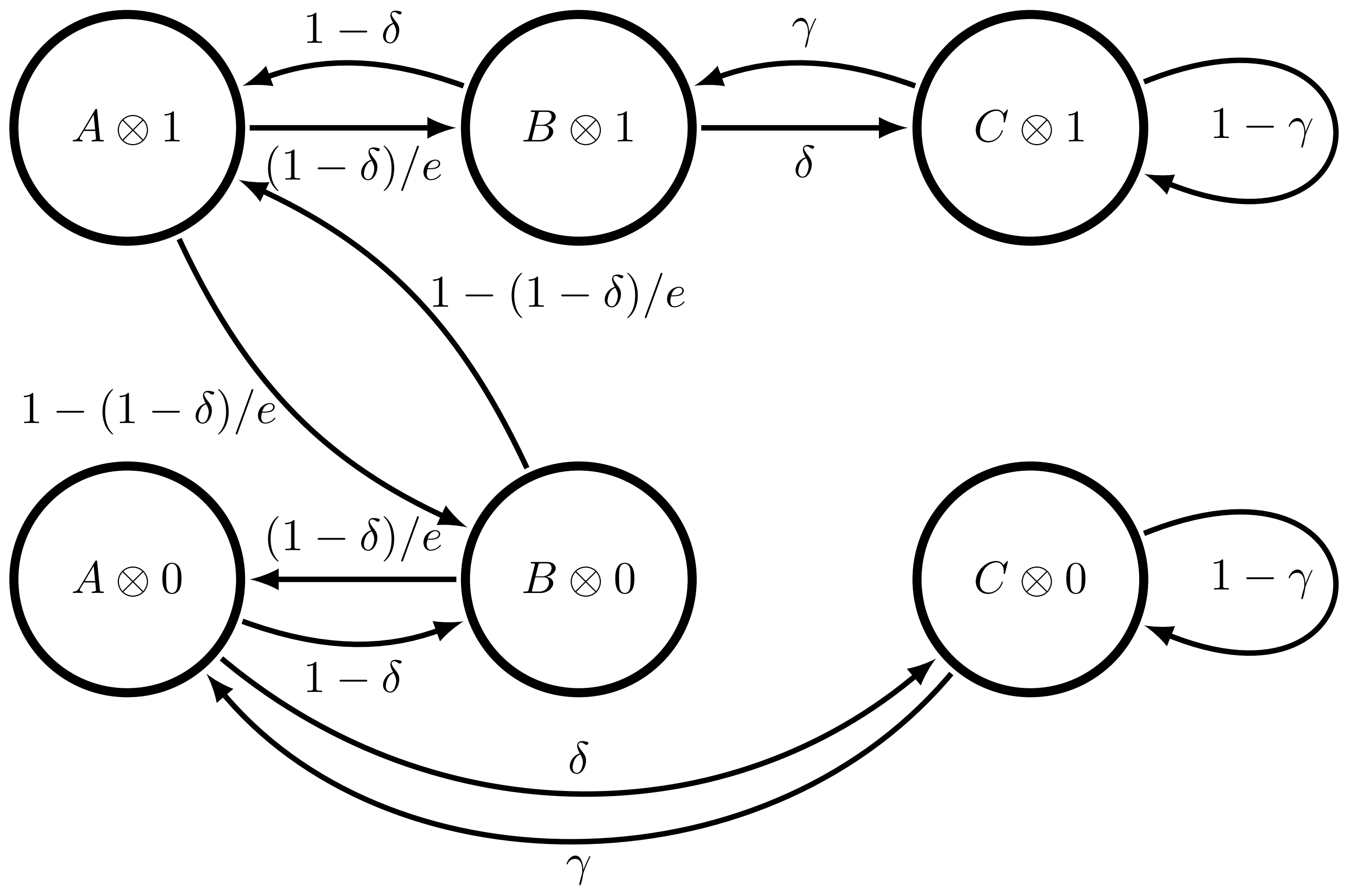
Period 2 Input



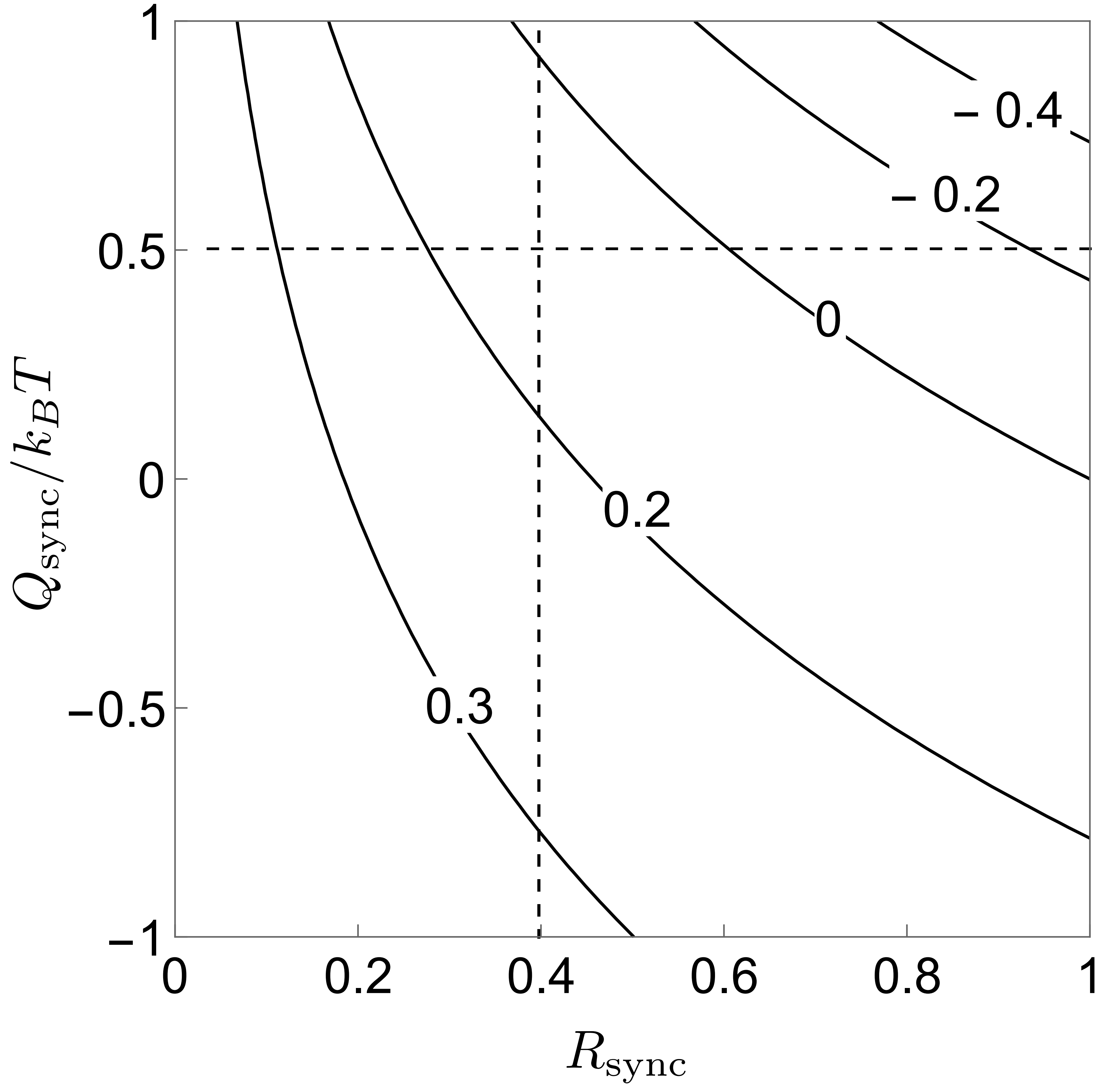
Candidate Ratchet



Must include synchronization mechanism



Synchronization Costs Work



Regularization

In machine learning:

input: $\vec{x}$

output: $\hat{W}\vec{x}$

target: $\vec{y}$

Regularization

In machine learning:

input: $\vec{x}$

output: $\hat{W}\vec{x}$

target: $\vec{y}$

unregularized: $\hat{W}^{\max} = \text{argmin}_W ||\hat{W}\vec{x} - \vec{y}||^2$

$= (\vec{x}^T\vec{x})^{-1}\vec{x}^T\vec{y}$ by linear regression

Regularization

In machine learning:

input: $\vec{x}$

output: $\hat{W}\vec{x}$

target: $\vec{y}$

unregularized: $\hat{W}^{\max} = \text{argmin}_W ||\hat{W}\vec{x} - \vec{y}||^2$

$= (\vec{x}^T\vec{x})^{-1}\vec{x}^T\vec{y}$ by linear regression

adding a cost to weight matrix regularizes: $\hat{W}^{\max} = \underset{W}{\text{argmin}} \left(||\hat{W}\vec{x} - \vec{y}||^2 + \lambda ||\hat{W}||^2 \right)$

Regularization

In machine learning:

input: $\vec{x}$

output: $\hat{W}\vec{x}$

target: $\vec{y}$

unregularized: $\hat{W}^{\max} = \mathrm{argmin}_W ||\hat{W}\vec{x} - \vec{y}||^2$

$= (\vec{x}^T\vec{x})^{-1}\vec{x}^T\vec{y}$ by linear regression

adding a cost to weight matrix regularizes: $\hat{W}^{\max} = \underset{W}{\mathrm{argmin}} \left(||\hat{W}\vec{x} - \vec{y}||^2 + \lambda ||\hat{W}||^2 \right)$

In thermodynamic learning:

Add a cost for synchronizing? $\theta^{\max}(y_{0:L}) = \underset{\theta}{\mathrm{argmax}} \left(W^{\theta}(y_{0:L}) + W^{\theta}_{\mathrm{sync}} \right)$?

Regularization

In machine learning:

input: $\vec{x}$

output: $\hat{W}\vec{x}$

target: $\vec{y}$

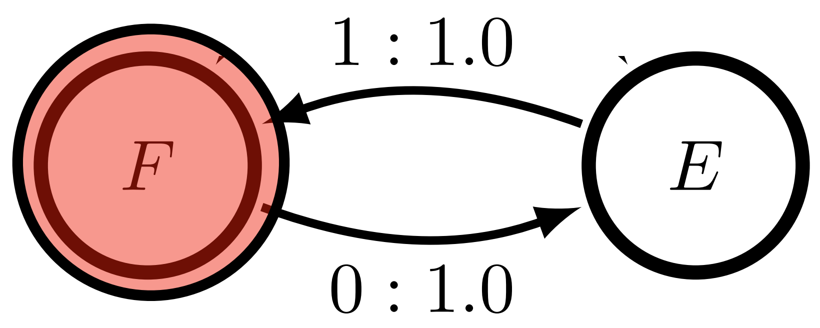
unregularized: $\hat{W}^{\max} = \mathrm{argmin}_W ||\hat{W}\vec{x} - \vec{y}||^2$

$= (\vec{x}^T\vec{x})^{-1}\vec{x}^T\vec{y}$ by linear regression

adding a cost to weight matrix regularizes: $\hat{W}^{\max} = \underset{W}{\mathrm{argmin}} \left(||\hat{W}\vec{x} - \vec{y}||^2 + \lambda ||\hat{W}||^2 \right)$

In thermodynamic learning:
Add a cost for synchronizing? $\theta^{\max}(y_{0:L}) = \underset{\theta}{\mathrm{argmax}} \left(W^\theta(y_{0:L}) + W^\theta_{\mathrm{sync}} \right)?$

Similar strategy: initialize in uniform memory distribution, such that cost of synchronization is incurred in operation



$$\beta \langle W^\theta(y_{0:L}) \rangle = L \ln |\mathcal{Y}| + \ln \Pr(Y^\theta_{0:L} = y_{0:L} | S_0 = s^*)$$

Regularization

In machine learning:

input: $\vec{x}$

output: $\hat{W}\vec{x}$

target: $\vec{y}$

unregularized: $\hat{W}^{\max} = \text{argmin}_W ||\hat{W}\vec{x} - \vec{y}||^2$

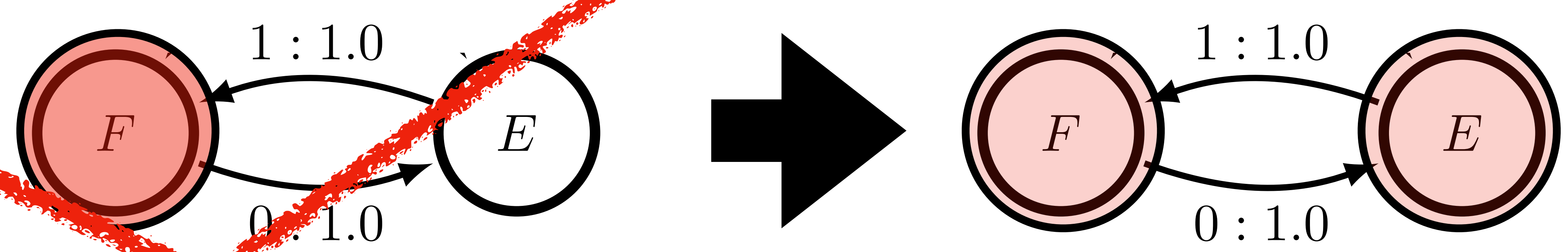
$= (\vec{x}^T\vec{x})^{-1}\vec{x}^T\vec{y}$ by linear regression

adding a cost to weight matrix regularizes: $\hat{W}^{\max} = \underset{W}{\text{argmin}} \left(||\hat{W}\vec{x} - \vec{y}||^2 + \lambda ||\hat{W}||^2 \right)$

In thermodynamic learning:

Add a cost for synchronizing? $\theta^{\max}(y_{0:L}) = \underset{\theta}{\text{argmax}} \left(W^\theta(y_{0:L}) + W^\theta_{\text{sync}} \right)$?

Similar strategy: initialize in uniform memory distribution, such that cost of synchronization is incurred in operation

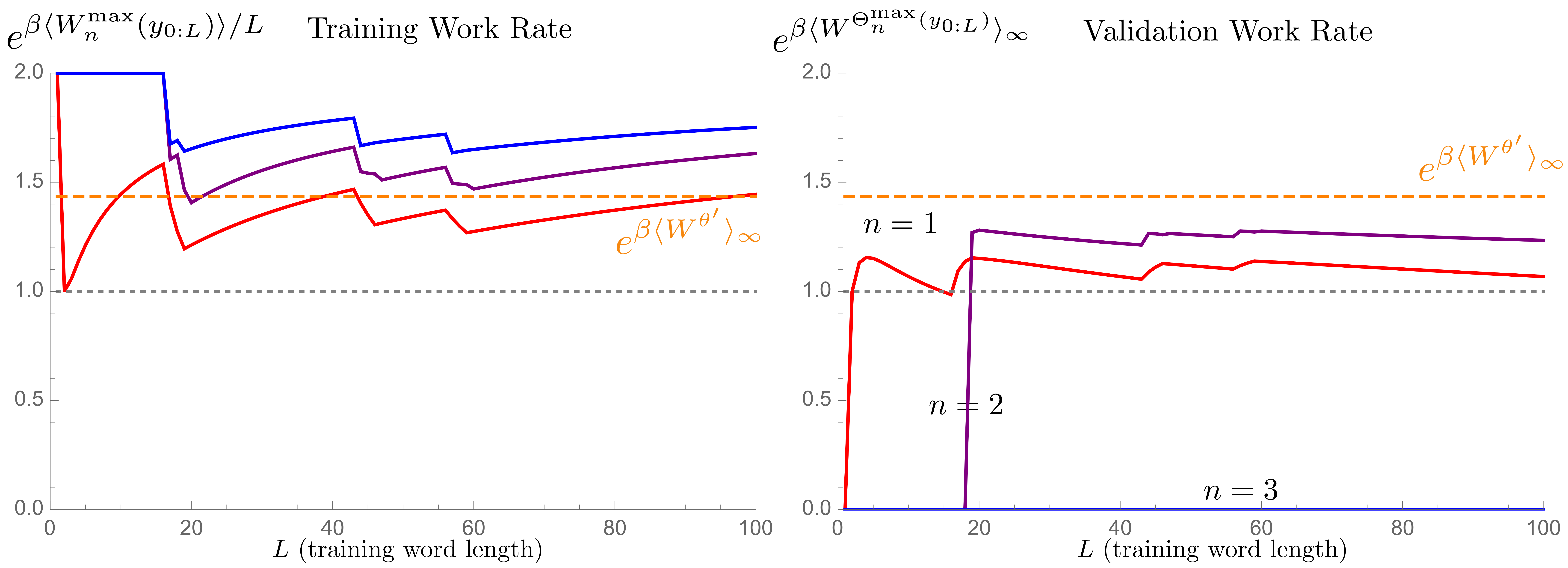


$\beta\langle W^\theta(y_{0:L})\rangle = L \ln |\mathcal{Y}| + \ln \Pr(Y^\theta_{0:L} = y_{0:L} | S_0 = s^*)$

$\beta\langle W^\theta(y_{0:L})\rangle = L \ln |\mathcal{Y}| + \sum_s \frac{1}{|\mathcal{S}|} \ln \Pr(Y^\theta_{0:L} = y_{0:L} | S_0 = s)$

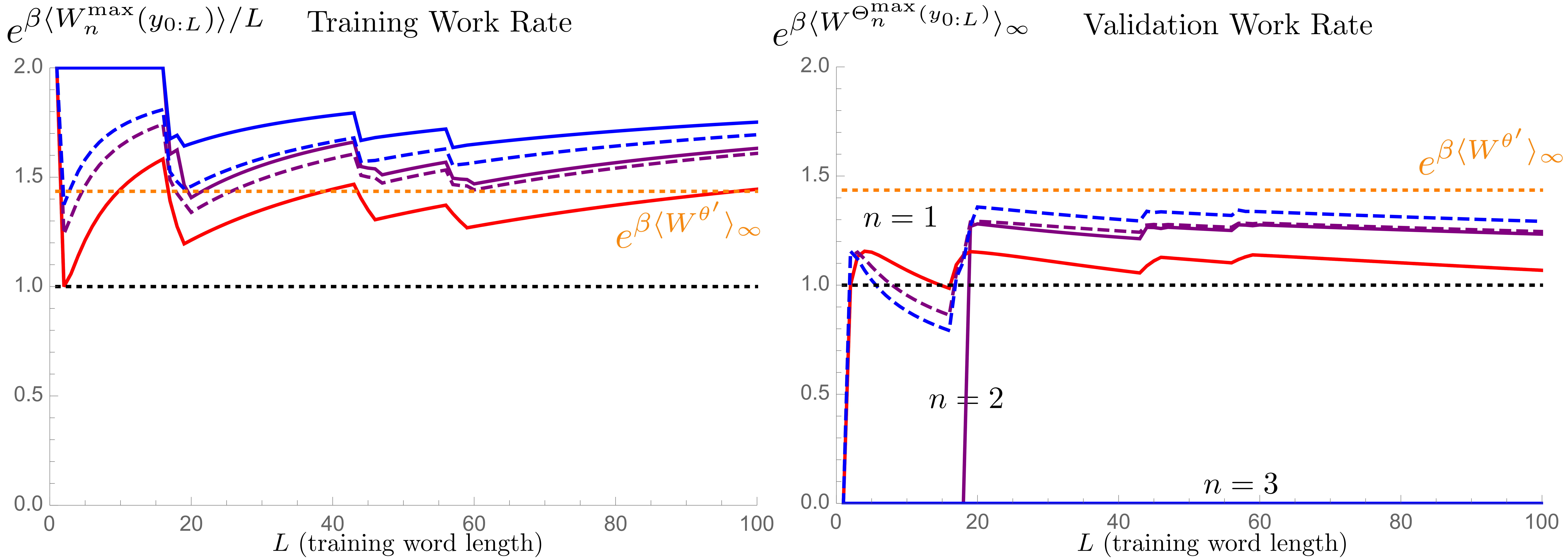
Regularization Through Synchronization

Without synchronization, maximum work agents badly overfit.



Regularization Through Synchronization

Without synchronization, maximum work agents badly overfit.



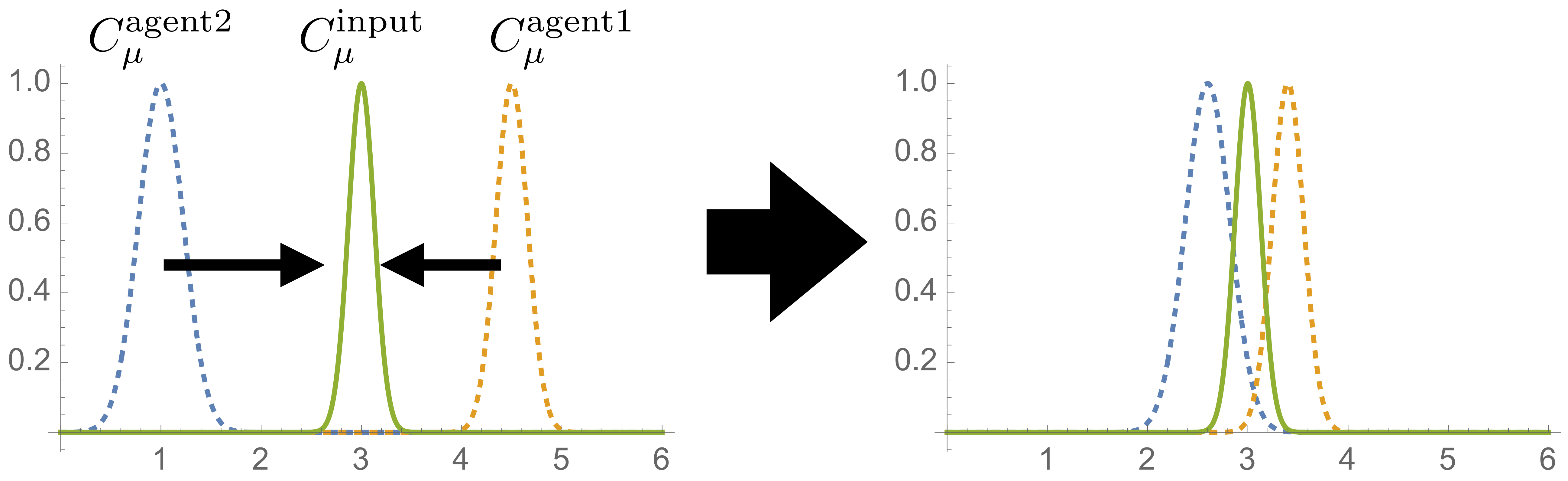
Synchronization mechanisms reduce training work rate, and seriously limits divergent dissipation.

Conclusion

Maximizing work production leads to learning complex models of patterns.

Overfitting to small data leads to divergent dissipation, which is more common for more complex information engines.

There is both a thermodynamic benefit and thermodynamic cost to complexity.



Learning can be partially regularized by requiring that engines autonomously synchronize.

A. B. Boyd, J. P. Crutchfield, and M. Gu. Thermodynamic Overfitting: Limits on Complexity in Thermodynamic Learning. *(Forthcoming)*

Acknowledgements

Collaborators
-James P. Crutchfield (UC Davis)
-Dibyendu Mandal
-Mile Gu (Nanyang Technological University)
-Felix Binder (Trinity College Dublin)

Funding Sources
-Foundational Questions Institute FQXi-RFPIPW-1910
-Templeton World Charity Foundation, Power of Information Independent Research Fellowship, TWCF0337, TWCF0560
-Army Research Office, W911NF-12-1-0288 and W911NF-13-1-0390

IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn

SCIENCE TO SHAPE THE FUTURE OF THE ARMY
ARMY RESEARCH OFFICE
1951

TEMPLETON WORLD
CHARITY FOUNDATION

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

FQXi
FOUNDATIONAL QUESTIONS INSTITUTE

BITS