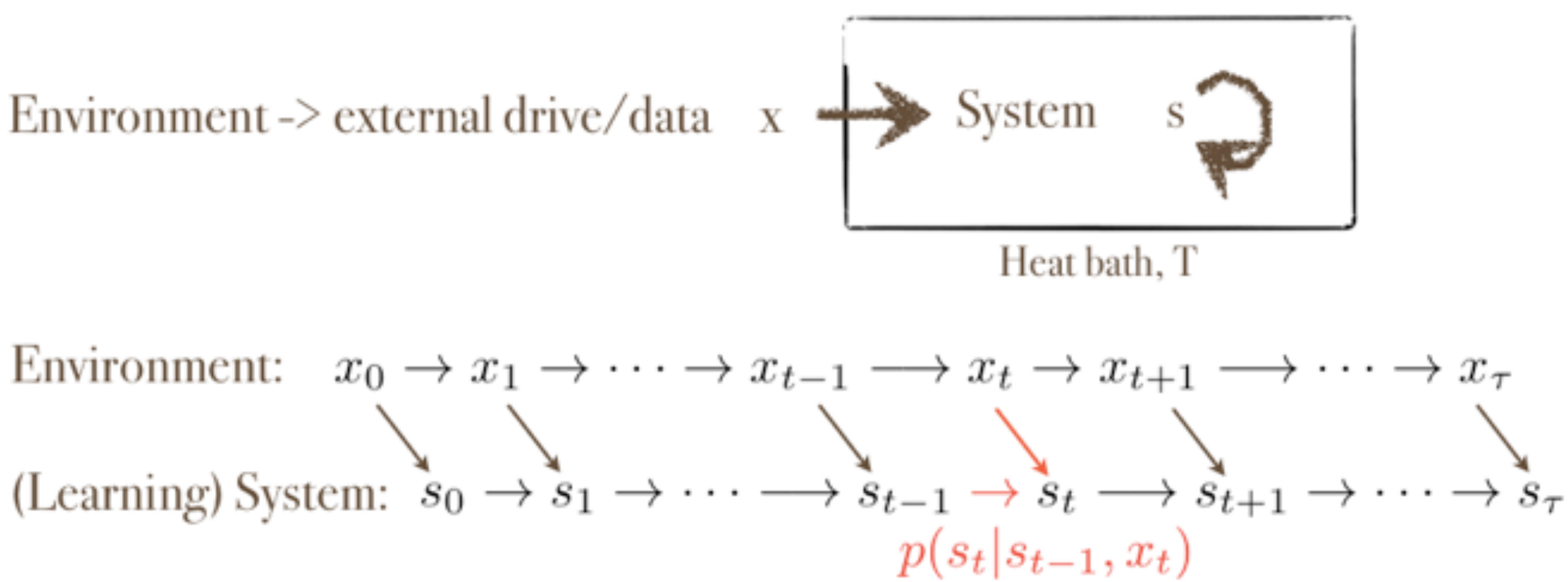
Cost and utility of data representation

Susanne Still
University of Hawaii at Manoa

Reminder

- Looking for a derivation of a prescription for learning from physics, in particular from physical limits of information processing.
- The first hint we have is: in stochastic driven systems, Dissipation is lower bound by nonpredictive information



$$\boxed{\beta\left(\langle W\rangle - \Delta F\right) \geq \sum_{t=0}^{\tau}\left(I[S_t; X_t] - I[S_t; X_{t+1}]\right)}$$

- Now we will look at generic data representation and ask about the limits to dissipation.

Data representation-abstract treatment

- Data (=description of state of some physical system): random variable X, with realizations x. Probability $p(x)$
- Representation of data: r.v. Y, with realizations y, probabilistic map $p(y|x)$
- Cost of data representation: some least effort is required
- Utility in data representation: some of the acquired knowledge could be utilized.
  - imagine some physical mechanism to extract work
  - maps X to a quantity Z that is relevant with respect to work extraction. In general: probabilistic map $p(z|x)$
- Hypothesis H = description of entire physical setup

Data representation-abstract treatment

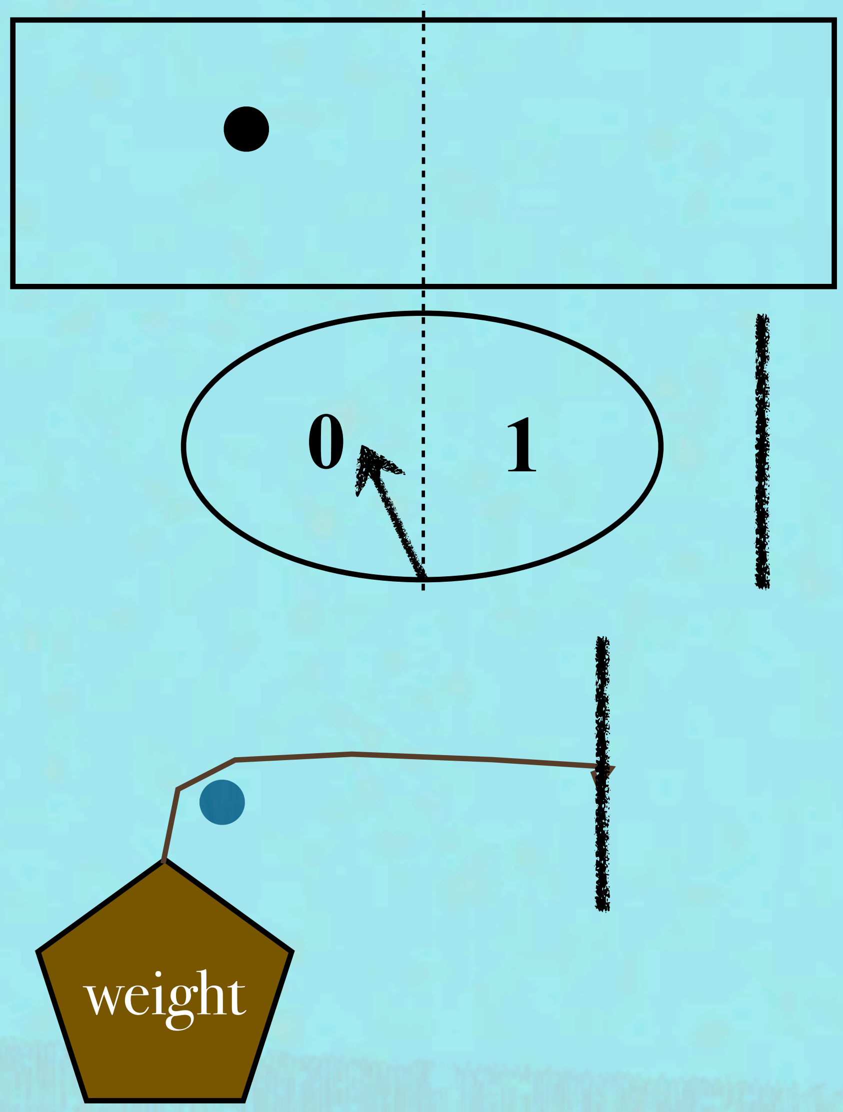
- Data (=description of state of some physical system): random variable X, with realizations x. Probability $p(x|H)$
- Representation of data: r.v. Y, with realizations y, probabilistic map $p(y|x, H)$
- Cost of data representation: some least effort is required
- Utility in data representation: some of the acquired knowledge could be utilized.
  - imagine some physical mechanism to extract work
  - maps X to a quantity Z that is relevant with respect to work extraction. In general: probabilistic map $p(z|x, H)$
- Hypothesis H = description of entire physical setup

Data representation-abstract treatment

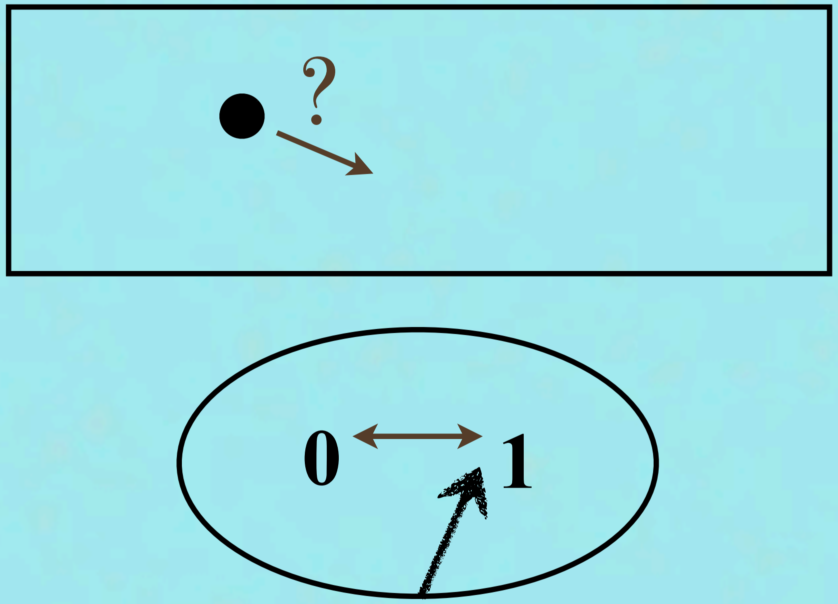
- Data (=description of state of some physical system): random variable X, with realizations x. Probability $p(x)$
- Representation of data: r.v. Y, with realizations y, probabilistic map $p(y|x)$
- Cost of data representation: some least effort is required
- Utility in data representation: some of the acquired knowledge could be utilized.
  - imagine some physical mechanism to extract work
  - maps X to a quantity Z that is relevant with respect to work extraction. In general: probabilistic map $p(z|x)$
- Hypothesis H = description of entire physical setup

A simple example: Szilard's engine

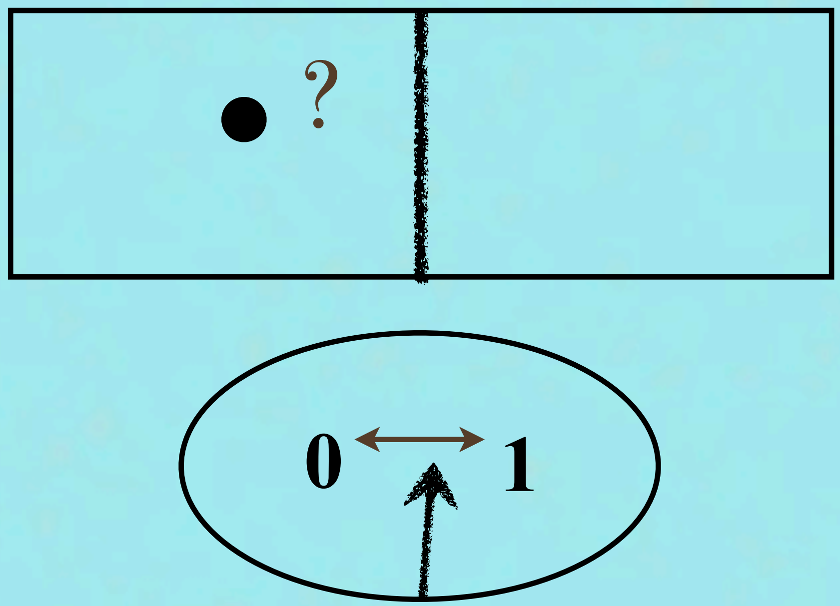
- One particle gas in box, in contact with heat bath
- Measurement device
- Apparatus to implement work extraction



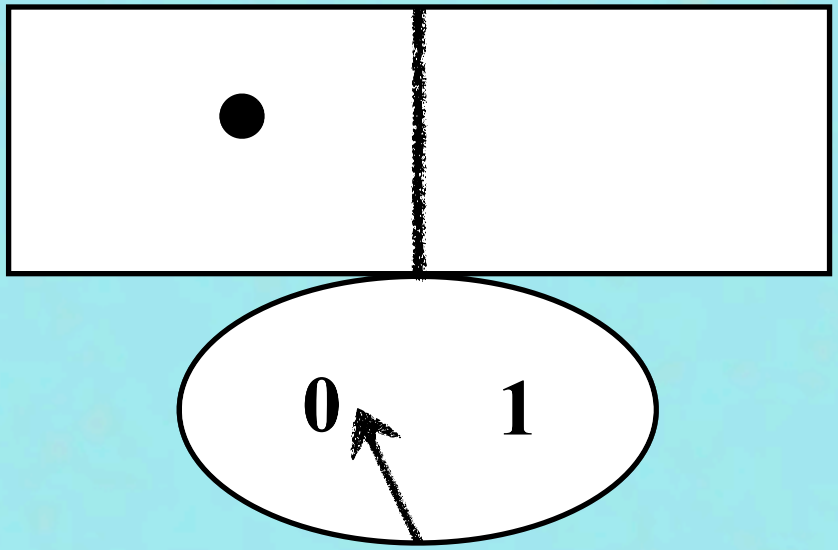
1) uncorrelated



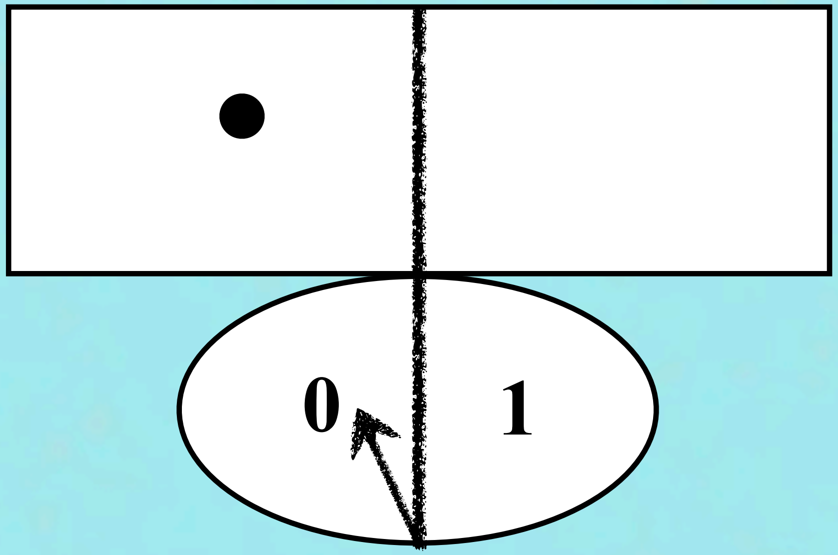
2) prepare measurement



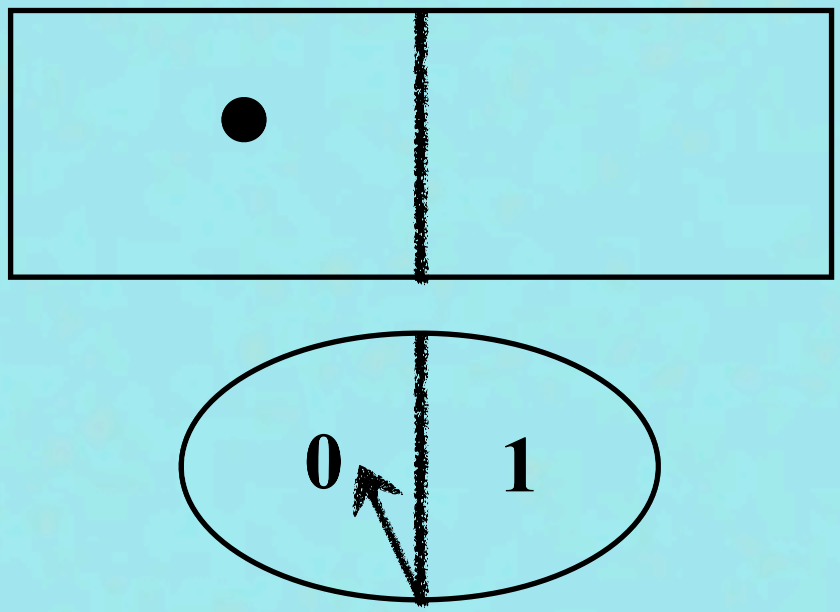
3) correlated



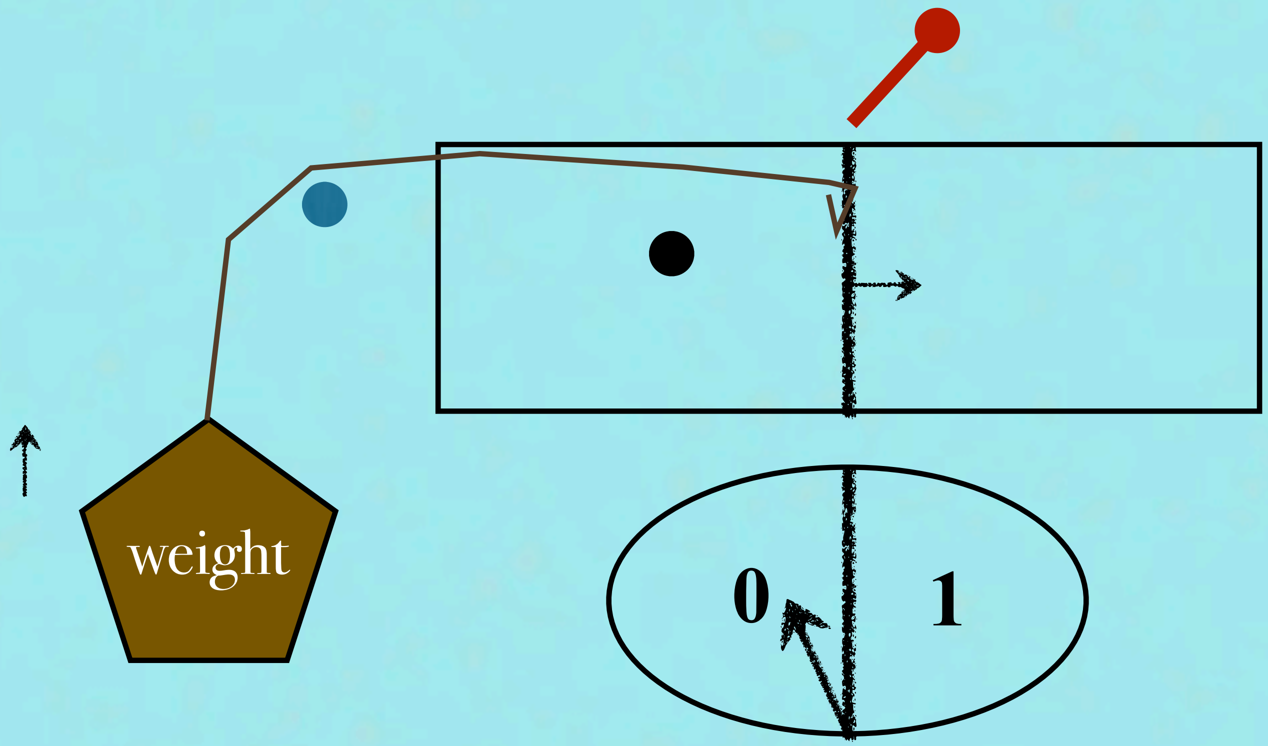
4) outcome stabilized



5) disconnect

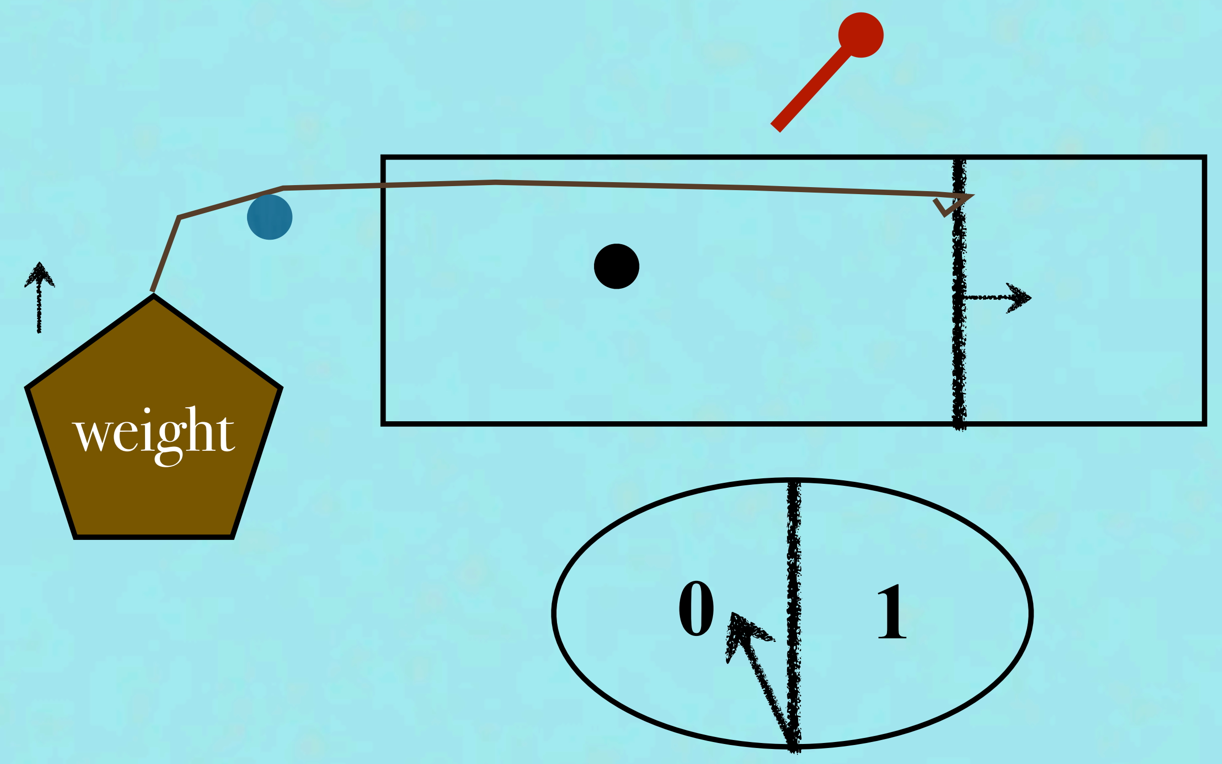


6) start work extraction protocol



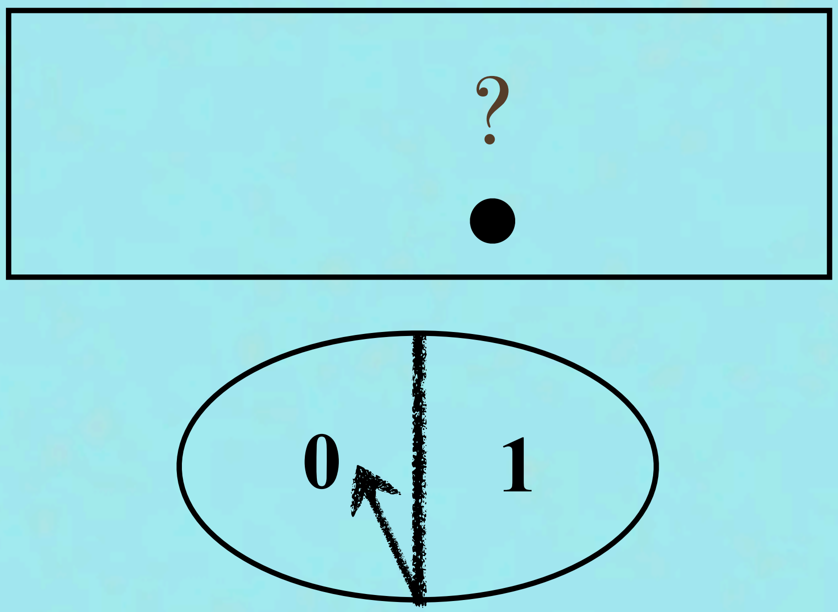
Instruction: “move away from side the pointer is on”

7) run work extraction protocol

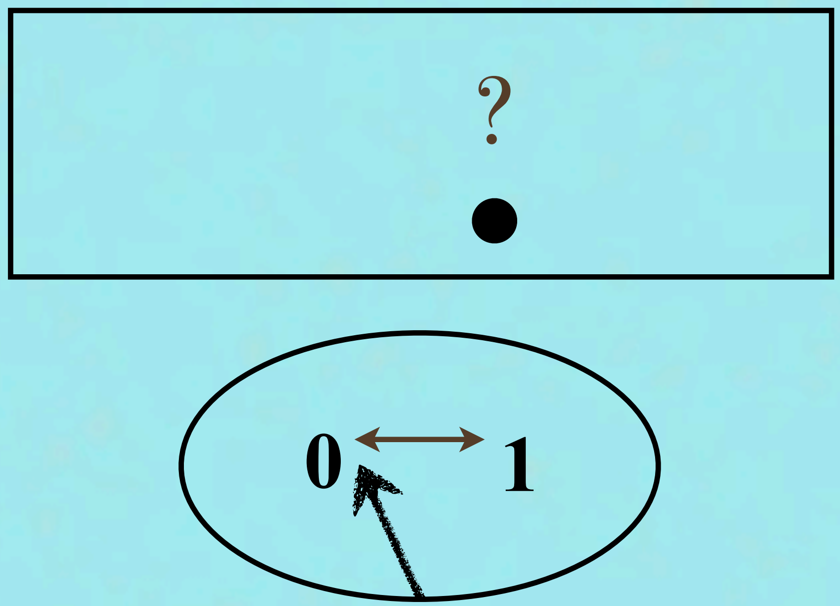


Instruction: “move away from side the pointer is on”

8) end work extraction protocol

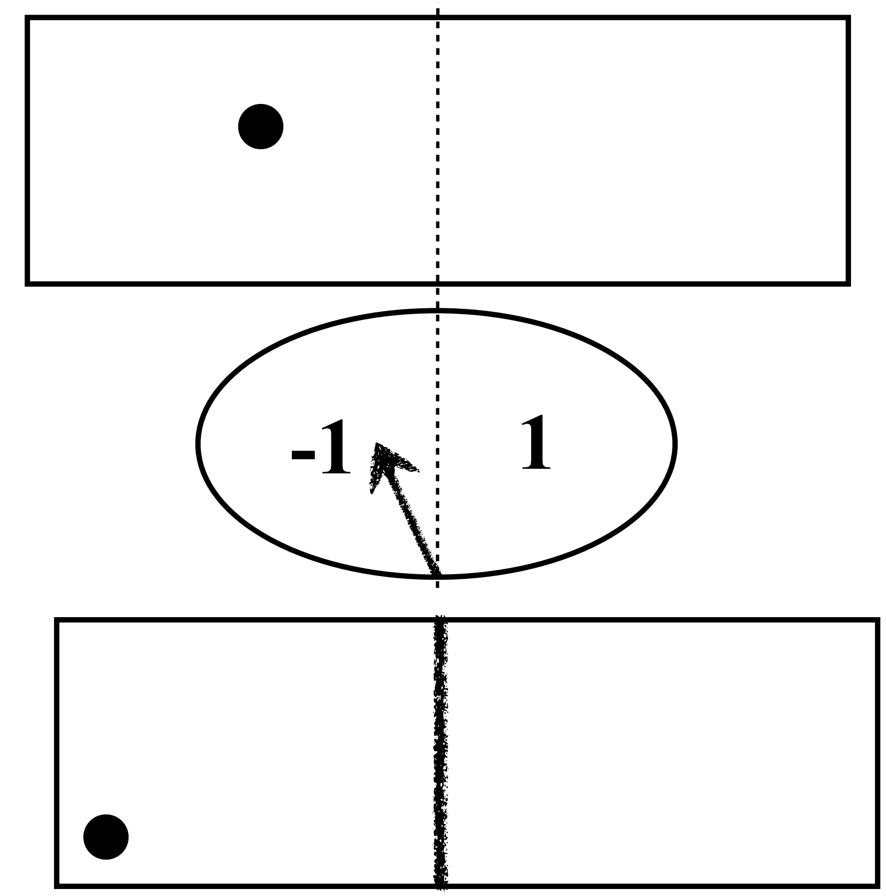


9) release "memory" of outcome

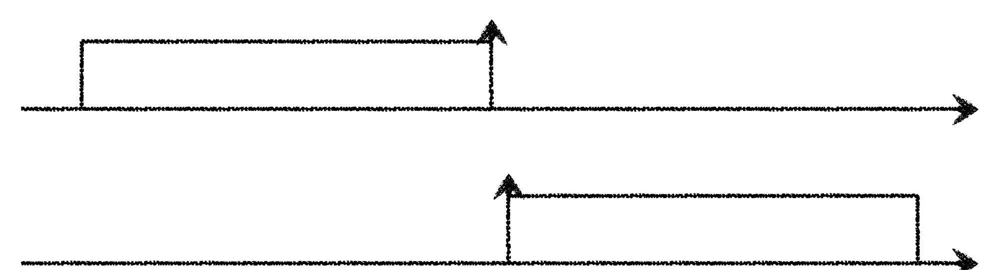


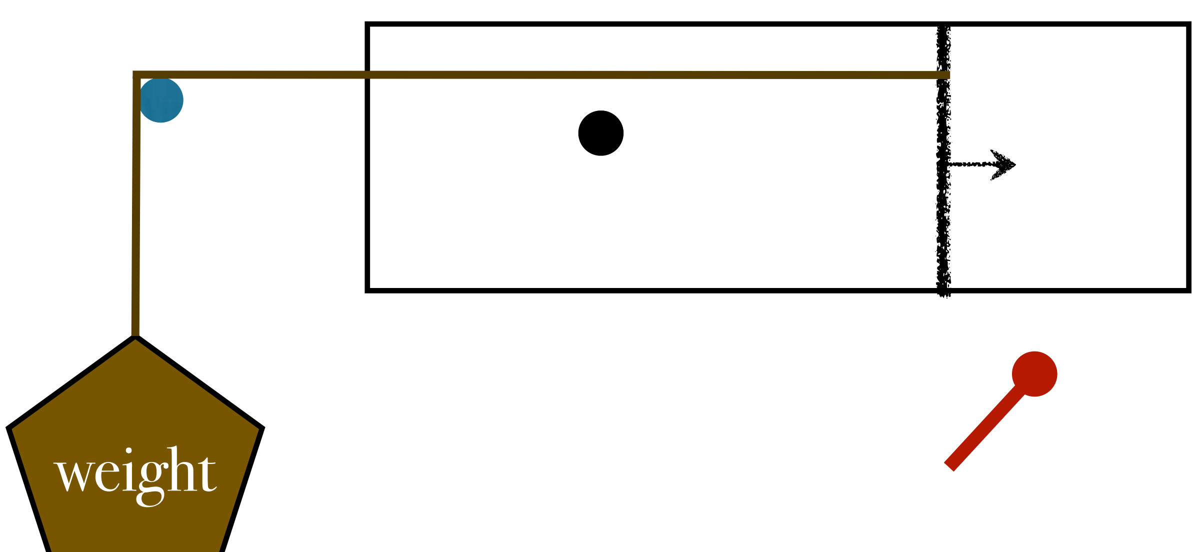
Data representation - simple example

- Szillard's "engine"
- Data x = description of physical system: 3-d position and velocity of the particle, p(x)
- Representation of data: y={-1,1} p(y|x) = potentially noisy measurement
- Relevant quantity z: position of the particle along the x-axis during work extraction protocol



p(z|x) =





- Hypothesis = description of physical setup: Volume and shape of box. Temperature of bath. Mechanism used to extract work.

Data representation-general treatment

- (Measurement and) representation of data
  - Change distribution of outcomes from average to the one reflecting the data, with a data dependent protocol

$$p(y) \xrightarrow{\Lambda(x)} p(y|x)$$

  - Free energy change:

$$\begin{aligned} \langle \Delta F_m \rangle &= \langle F[p(y|x)] \rangle_{p(x)} - F[p(y)] \\ &= \langle \Delta E_m \rangle - kTH[Y|X] + kTH[Y] \\ &= \langle W_m \rangle + \langle Q_m \rangle + kTI[X,Y] \end{aligned}$$

  - Second law:

$$\langle W_m \rangle \geq \langle \Delta F_m \rangle \;\; \Leftrightarrow \;\; -\langle Q_m \rangle \geq kTI[X,Y] \geq 0$$

least effort

dissipated heat

$$\boxed{W_{\text{in}} \geq \Delta F_{\text{LE}}} \qquad \boxed{Q_{\text{out}} \geq kTI[X,Y]}$$

Work extraction-general treatment

- Work extraction apparatus:

1. introduces relevant quantity, z, and determines p(z|x)
2. stabilizes outcome distribution:

$$p(z|y) = \sum_x p(z|x)p(x|y) \overset{\text{Bayes}}{=} \frac{1}{p(y)} \sum_x p(z|x)p(y|x)p(x)$$

4. extracts work:

$$p(z|y) \xrightarrow{\Lambda(y)} p(z)$$

$$\langle \Delta F_e \rangle = F[p(z)] - \langle F[p(z|y)] \rangle_{p(y)}$$
$$= \langle W_e \rangle + \langle Q_e \rangle - kTI[Y, Z]$$

- Second law: $\langle W_e \rangle - \langle \Delta F_e \rangle \ = \ -\langle Q_e \rangle + kTI[Y, Z] \geq 0$

$$-\langle W_e \rangle \leq -\langle \Delta F_e \rangle \ \Leftrightarrow \ \langle Q_e \rangle \leq kTI[Y, Z]$$

work potential ↗ ↖ absorbed heat

$$\boxed{W_{\text{out}} \leq \Delta F_{\text{WP}}} \qquad \boxed{Q_{\text{in}} \leq kTI[Y, Z]}$$

Closing the cycle-general treatment

- Reset measurement apparatus? Continuous measurements: current state y is not correlated with new data x. What is p(y)? Just the average:

$$p(y) = \sum_x p(y|x)p(x)$$

- Reset system? Work extraction mechanism is constructed to do this. Ends in p(z). Preparation step creates a particular decomposition:

$$p(z) = \sum_x p(z|x)p(x)$$

- Total average energy change in a cycle:

$$\langle \Delta E \rangle = \underbrace{\langle W_m \rangle}_{W_{\text{in}}} + \underbrace{\langle Q_m \rangle}_{-Q_{\text{out}}} + \underbrace{\langle W_e \rangle}_{-W_{\text{out}}} + \underbrace{\langle Q_e \rangle}_{Q_{\text{in}}} \stackrel{!}{=} 0$$

$$\Rightarrow \quad W_{\text{in}} - W_{\text{out}} = Q_{out} - Q_{in}$$ Dissipation

Dissipation-general treatment

- Dissipation is lower bound by information trade-off between total memory and relevant information:

$$\boxed{Q_{\mathrm{out}} - Q_{\mathrm{in}} \geq k\underbrace{T_1}_{\text{Measurement instrument}} I[X,Y] - k\underbrace{T_2}_{\text{Work medium}} I[Y,Z]}$$

- Minimize the bound over all data representations:

$$\min_{p(y|x)} \left(I[X,Y] - \frac{T_2}{T_1} I[Y,Z] \right)$$

This is the optimization performed by the *"Information Bottleneck" method (introduced by Tishby, Pereira and Bialek in 1999)*

Again - with example

Data representation-general treatment

- (Measurement and) representation of data
  - Change distribution of outcomes from average to the one reflecting the data, with a data dependent protocol

$$p(y) \xrightarrow{\Lambda(x)} p(y|x)$$

  - Free energy change:

$$\begin{aligned} \langle \Delta F_m \rangle &= \langle F[p(y|x)] \rangle_{p(x)} - F[p(y)] \\ &= \langle \Delta E_m \rangle - kTH[Y|X] + kTH[Y] \\ &= \langle W_m \rangle + \langle Q_m \rangle + kTI[X,Y] \end{aligned}$$

  - Second law:

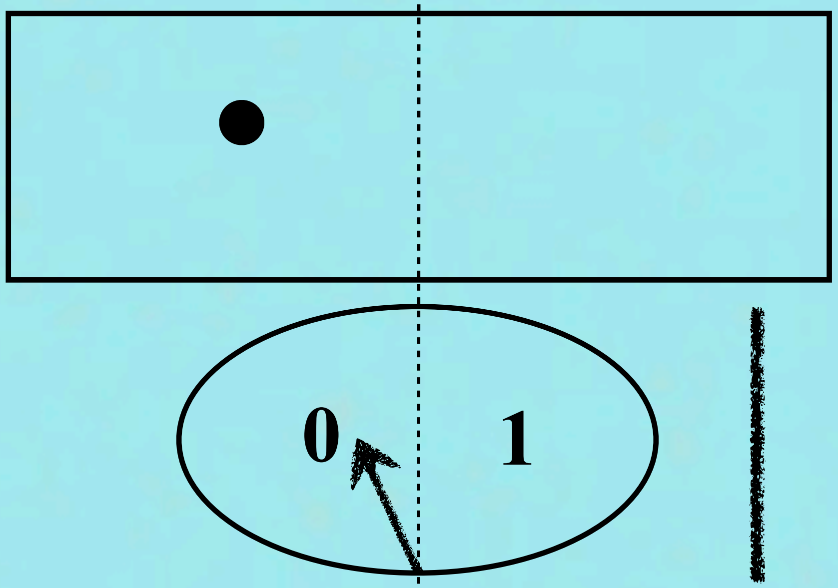
$$\langle W_m \rangle \geq \langle \Delta F_m \rangle \;\Leftrightarrow\; -\langle Q_m \rangle \geq kTI[X,Y] \geq 0$$

least effort

dissipated heat

$$\boxed{W_{\text{in}} \geq \Delta F_{\text{LE}}} \qquad \boxed{Q_{\text{out}} \geq kTI[X,Y]}$$

Nonequilibrium free energy change



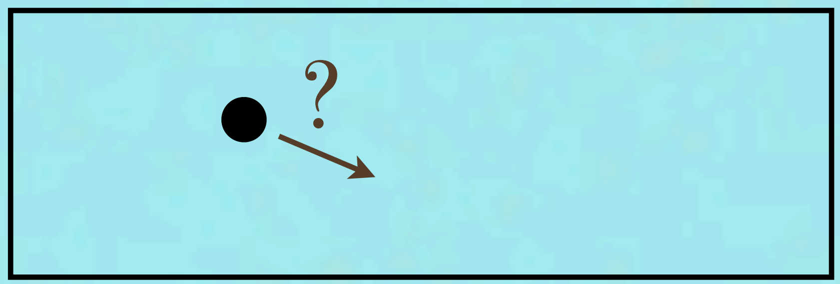
- System to be measured: x
- Measurement outcome: y
- Average nonequilibrium free energy:

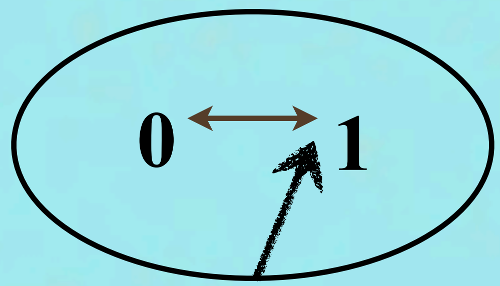
$$\langle F[p(y|x)]\rangle_{p(x)} = \langle E_x(y)\rangle_{p(x,y)} - H[Y|X]$$

- Look at the change in average nonequilibrium free energy!

1) uncorrelated

$$p(x, y) = p(x)p(y)$$





$$p(y) = \frac{1}{2}$$

$$p(y|x) = p(y)$$

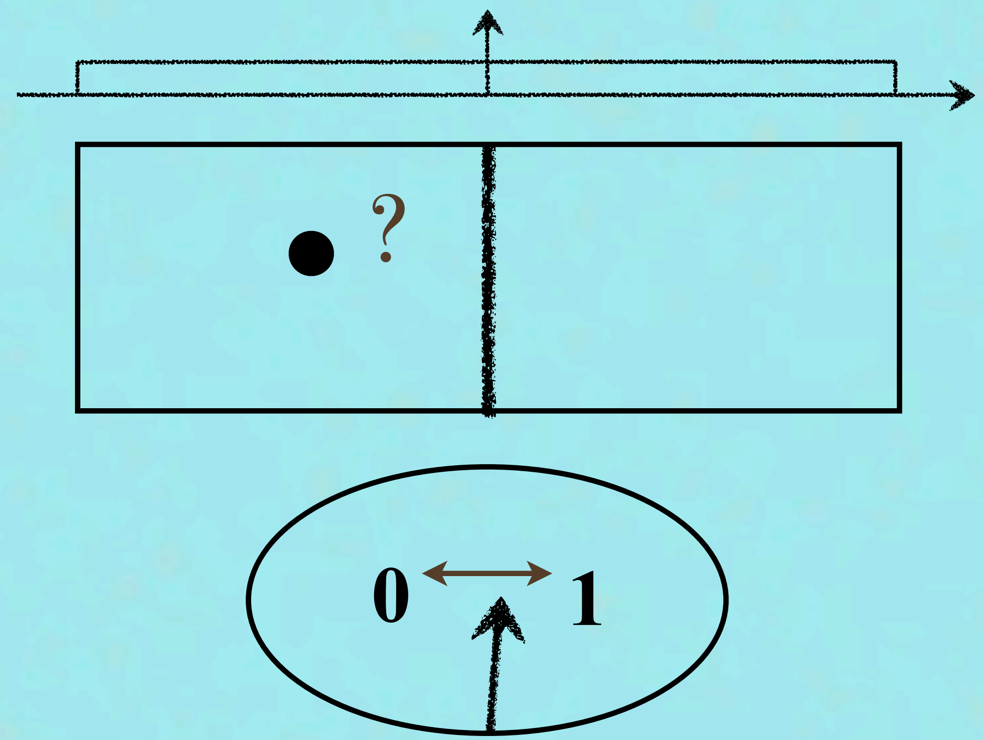
- Average nonequilibrium free energy:

$$\langle F[p(y|x)] \rangle_{p(x)} = F[p(y)]$$
$$\boxed{= \langle E(y) \rangle_{p(y)} - kTH[Y]}$$
$$= E_Y - kT \log(2)$$

2) prepare measurement

induces Z

no change in Y

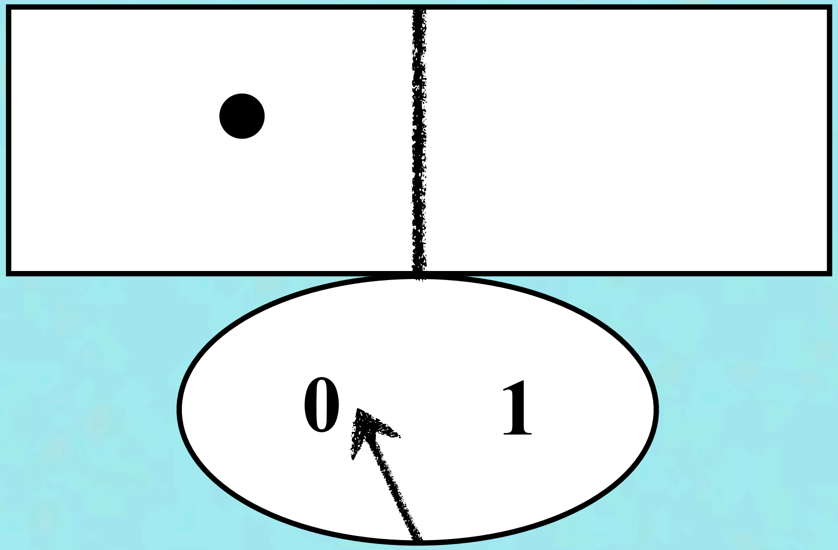


$$p(z) = \langle p(z|x) \rangle_{p(x)}$$
$$= \frac{1}{L}$$

$$p(y|x) = p(y)$$

$$\boxed{F[p(y)] = E_Y - kT \log(2)}$$

3) correlated



error free (for simplicity)

$$p(y|x) = \delta_{y,\theta(x)}$$

Free energy : $\boxed{\langle F[p(y|x)]\rangle_{p(x)} = \langle E_x(y)\rangle_{p(x,y)} - kTH[Y|X]}$

<u>**Free energy change:**</u>

least effort

$$\Delta F_{LE} = \langle F[p(y|x)]\rangle_{p(x)} - F[p(y)] \leq W_Y$$

$$= W_Y + Q_Y + kTI[X,Y]$$

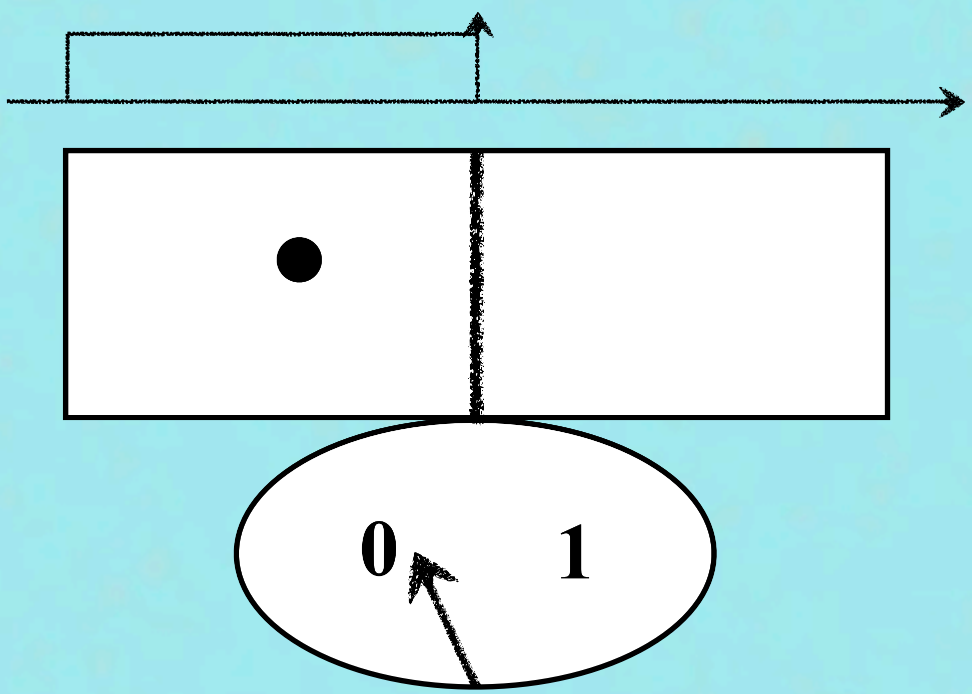
$$= W_Y + Q_Y + kT\log(2)$$

$$\Rightarrow Q_{\text{out}} = -Q_Y \geq kTI[X,Y]$$

$$= kT\log(2)$$

Dissipated heat is at least kT log(2)

3) correlated



$$\Delta E = 0 \ \Rightarrow W_Y = -Q_Y$$
$$\Leftrightarrow W_{\text{in}} = Q_{\text{out}}$$

least effort

$$\Delta F_{LE} = kTI[X,Y] = kT\log(2)$$

Work extraction-general treatment

- Work extraction apparatus:

  1. introduces relevant quantity, z, and determines p(z|x)
  2. stabilizes outcome distribution:

$$p(z|y) = \sum_x p(z|x)p(x|y) \overset{\text{Bayes}}{=} \frac{1}{p(y)} \sum_x p(z|x)p(y|x)p(x)$$

  4. extracts work:

$$p(z|y) \xrightarrow{\Lambda(y)} p(z)$$

$$\langle \Delta F_e \rangle = F[p(z)] - \langle F[p(z|y)] \rangle_{p(y)}$$
$$= \langle W_e \rangle + \langle Q_e \rangle - kTI[Y,Z]$$

- Second law: $\langle W_e \rangle - \langle \Delta F_e \rangle \;=\; -\langle Q_e \rangle + kTI[Y,Z] \geq 0$

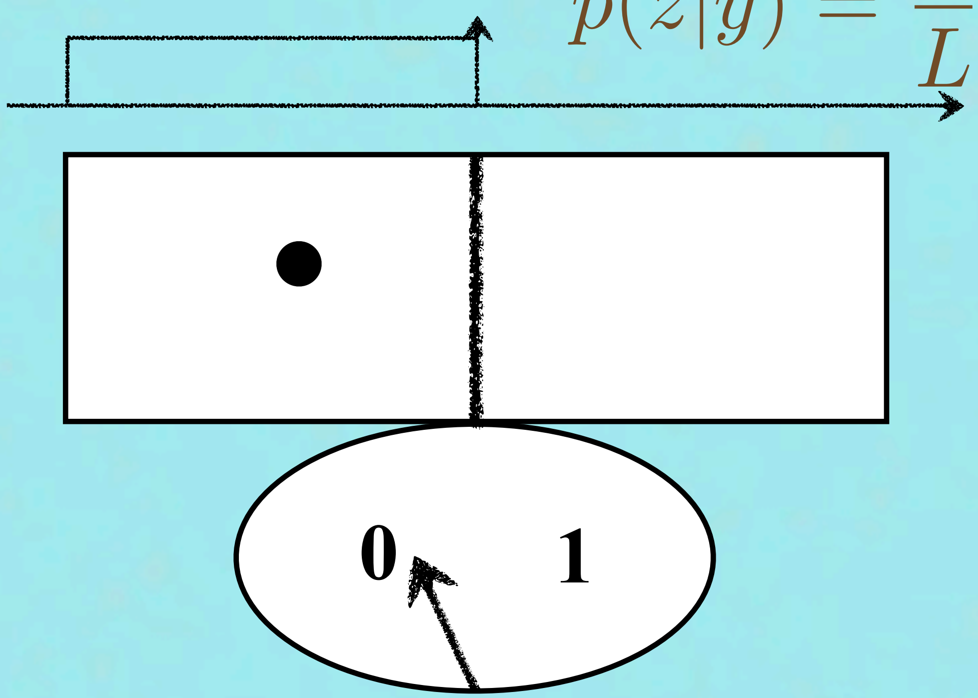
$$-\langle W_e \rangle \leq -\langle \Delta F_e \rangle \;\Leftrightarrow\; \langle Q_e \rangle \leq kTI[Y,Z]$$

work potential ↗ ↖ absorbed heat

$$\boxed{W_{\text{out}} \leq \Delta F_{\text{WP}}} \qquad \boxed{Q_{\text{in}} \leq kTI[Y,Z]}$$

3) correlated

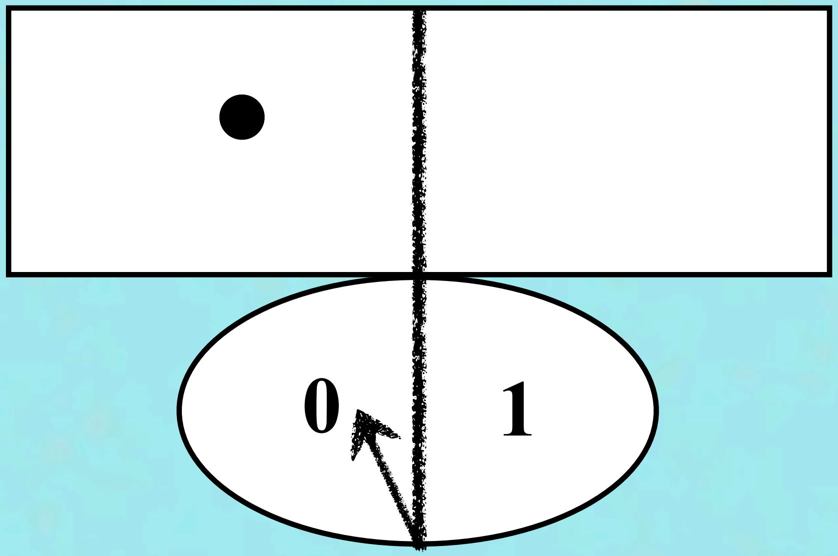
$$p(z|y) = \frac{2}{L}\left(y\,\theta(z) + (1-y)\theta(-z) \right)$$



work potential

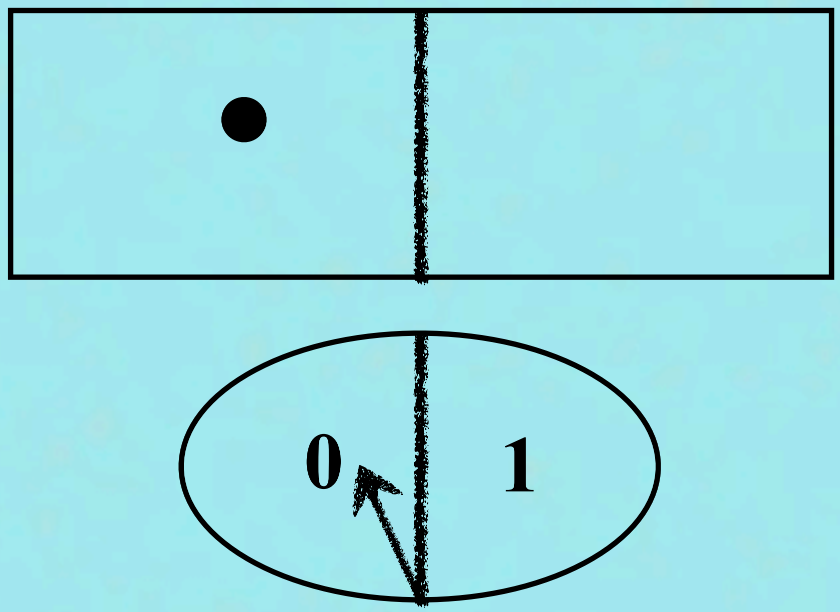
$$\begin{aligned}\Delta F_{WP} &= \langle F[p(z|y)] \rangle_{p(y)} - F[p(z)] \\ &= \Delta E + kT\, I[Y, Z] \\ &= kT \log(2)\end{aligned}$$

4) outcome stabilized



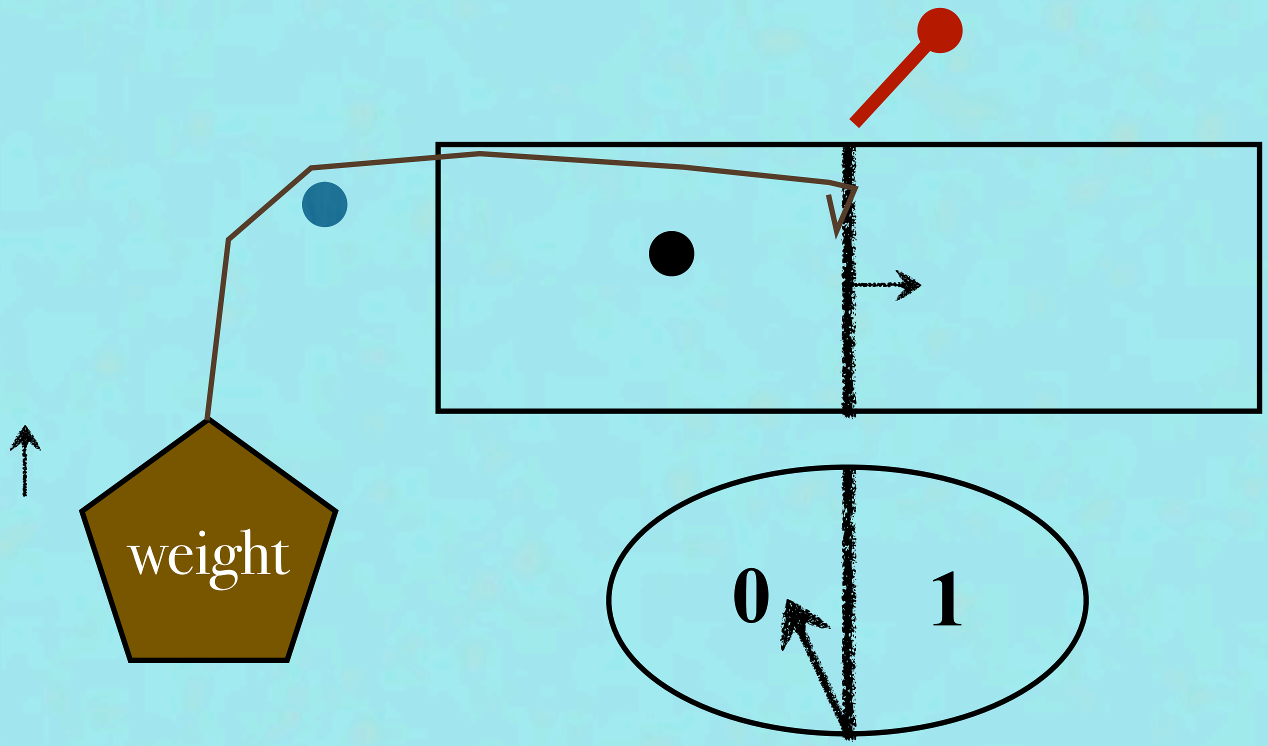
no change

5) disconnect



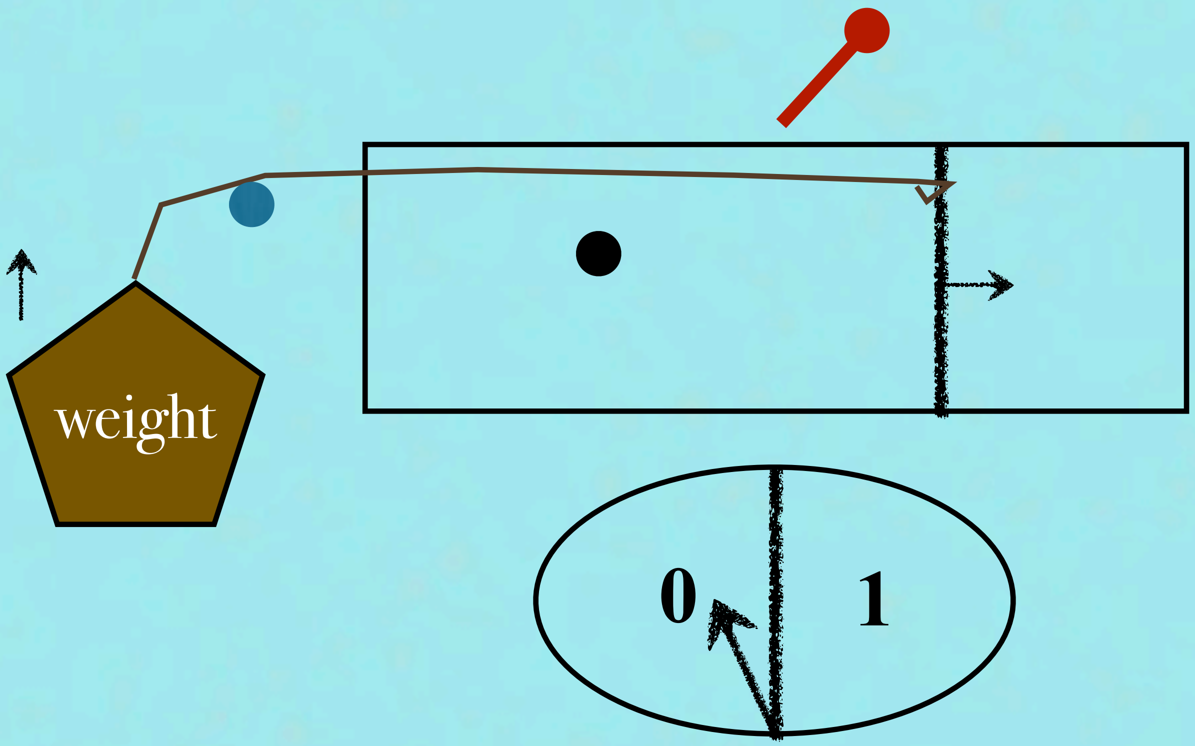
no change

6) start work extraction protocol



Instruction: “move away from side the pointer is on”

7) run work extraction protocol



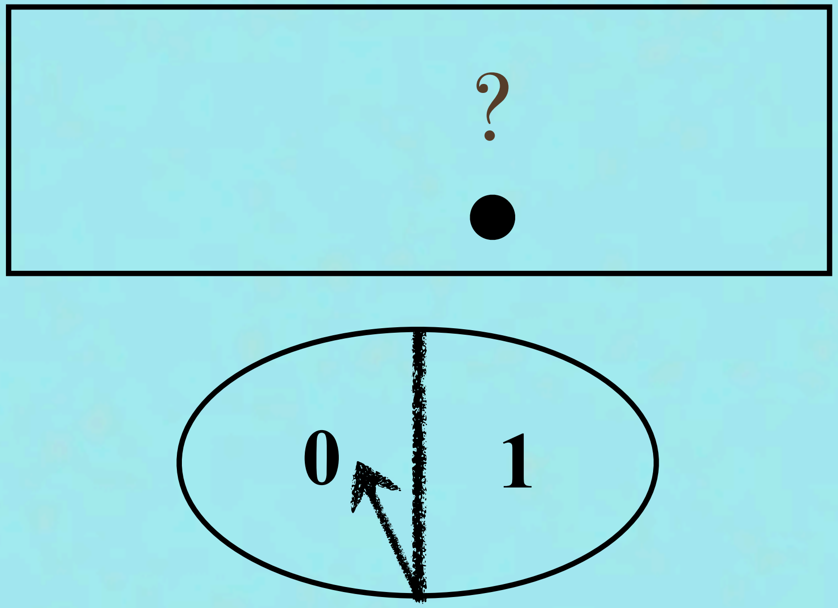
$$\langle F[p(z|y)] \rangle_{p(y)} \longrightarrow F[p(y)]$$

$$\Delta F = F[p(z)] - \langle F[p(z|y)] \rangle_{p(y)} = W_e + Q_e - kT I[Y, Z]$$

Extracted work: $W_{\text{out}} = -W_e \leq \Delta F_{WP} = kT \log(2)$

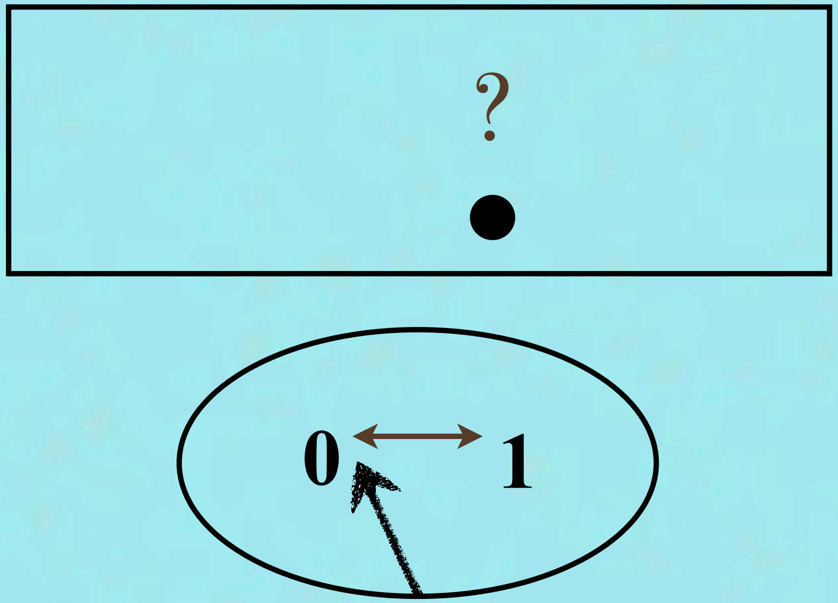
Absorbed heat: $Q_{\text{in}} = Q_e \leq kT\, I[Y, Z] = kT \log(2)$

8) end work extraction protocol



$$p(y) = \langle p(y|x) \rangle_{p(x)} = \frac{1}{2}$$

9) release "memory" of outcome



$$p(y) = \frac{1}{2}$$

no free energy change

Closing the cycle-general treatment

- Reset measurement apparatus? Continuous measurements: current state y is not correlated with new data x. What is p(y)? Just the average:

$$p(y) = \sum_x p(y|x)p(x)$$

- Reset system? Work extraction mechanism is constructed to do this. Ends in p(z). Preparation step creates a particular decomposition:

$$p(z) = \sum_x p(z|x)p(x)$$

- Total average energy change in a cycle:

$$\langle \Delta E \rangle = \underbrace{\langle W_m \rangle}_{W_{\text{in}}} + \underbrace{\langle Q_m \rangle}_{-Q_{\text{out}}} + \underbrace{\langle W_e \rangle}_{-W_{\text{out}}} + \underbrace{\langle Q_e \rangle}_{Q_{\text{in}}} \stackrel{!}{=} 0$$

$$\Rightarrow \quad W_{\text{in}} - W_{\text{out}} = Q_{out} - Q_{in}$$ Dissipation

Altogether, we have for our example

$$W_{\mathrm{in}} - W_{\mathrm{out}} = Q_{out} - Q_{in} \geq kTI[X,Y] - kTI[Y,Z] = kT\log(2) - kT\log(2) = 0$$

- The data representation we perform in the "standard" Szillard engine setup, namely say if the particle was on the left or right side of the partition, is optimal in terms of minimizing dissipation!

Dissipation-general treatment

- Dissipation is lower bound by information trade-off between total memory and relevant information:

$$\boxed{Q_{\mathrm{out}} - Q_{\mathrm{in}} \geq k\underbrace{T_1}_{\text{Measurement instrument}} I[X,Y] - k\underbrace{T_2}_{\text{Work medium}} I[Y,Z]}$$

- Minimize the bound over all data representations:

$$\min_{p(y|x)} \left(I[X,Y] - \frac{T_2}{T_1} I[Y,Z] \right)$$

This is the optimization performed by the *"Information Bottleneck" method (introduced by Tishby, Pereira and Bialek in 1999)*