Thermodynamic efficiency implies predictive inference

Susanne Still
University of Hawaii at Mānoa

Information processing, learning and adaptation

- Natural phenomena occurring in the physical world
- Want a physics based explanation!

Information processing, learning and adaptation

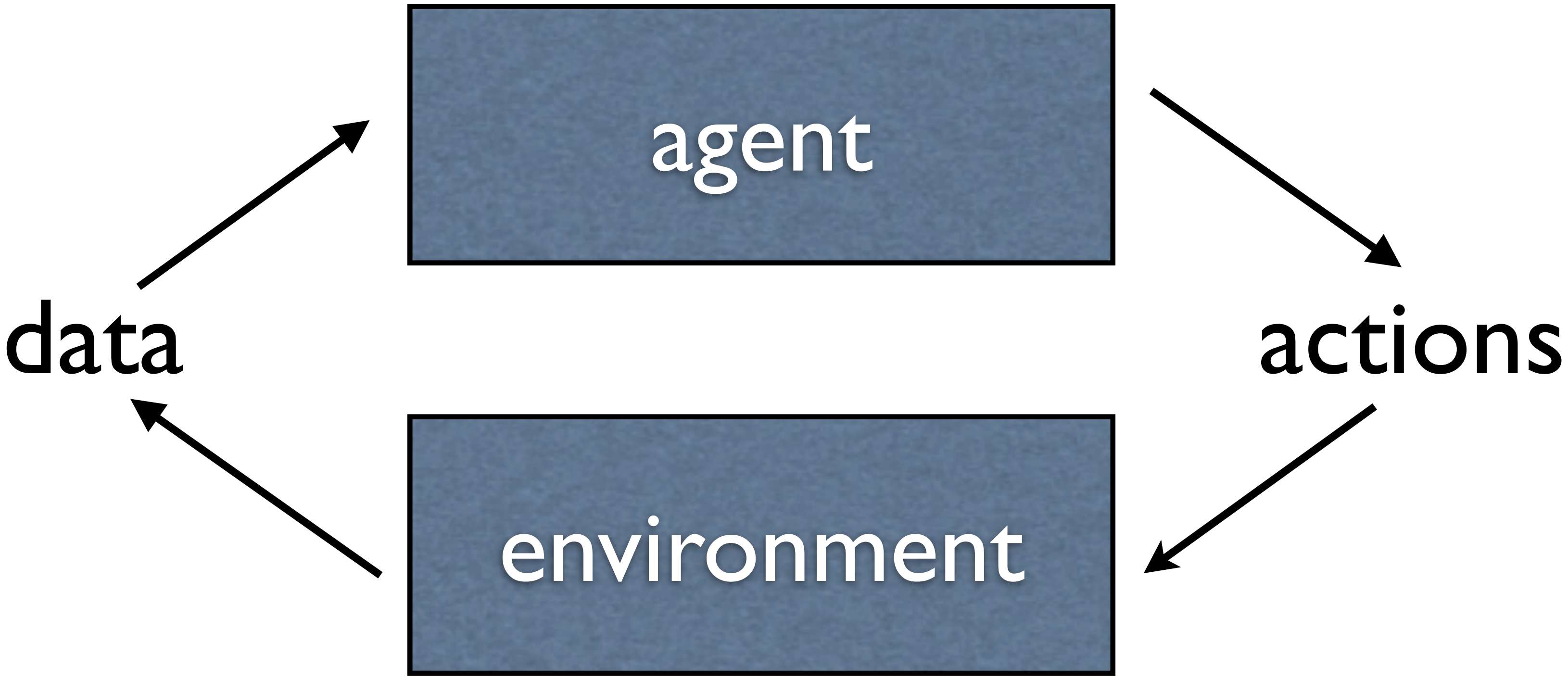
- This is something observers do! Classically, the observer is taken out of physics. Many physicists feel uncomfortable about trying to discover a theory of how the observer describes the world.
- Physical modeling of neural systems has lead to the construction of neural networks and deep learning, and to some degree also an explanation of why they work, but no complete, overarching theory.
- Machine learning has some solid theoretical foundations, e.g. in statistical learning theory. This is based on mathematical arguments, not on physics. Makes use of ad hoc measures.

Information processing, learning and adaptation

- We want a *physics based approach* in which:
  - complex behavior emerges from simple first principles
  - rules for learning and adaptation can be derived from those principles, instead of having to choose objective functions ad hoc

Problem description:

- Observers interact with their environment: sense, process information, and act. Thus often called ``agents”:



- Important: there are generic restrictions on agents. E. g.: Finite operating times, partial observability; partial control.

Abstract models of agents

Describe decision making and behavior as optimization

- What is being optimized? Often: some kind of utility function, under the constraint of a cost function
- Examples for utility:
  - rewards (reinforcement learning)
  - loss or risk function (decision theory)
  - error function (control theory)
  - payoff (game theory)
- Cost is often motivated by some philosophical argument, such as Ockham's razor, or statistical arguments. Cost = measure of complexity or capacity of the model class.

Big issue with this approach

- Utility function and cost function need to be specified.

  - It is not always *a priori* clear which utility to assign to actions/outcomes of actions.
  - There are many “reasonable” complexity measures

- If there is no guiding principle, then we end up with descriptive modeling, rather than an explanatory theory.

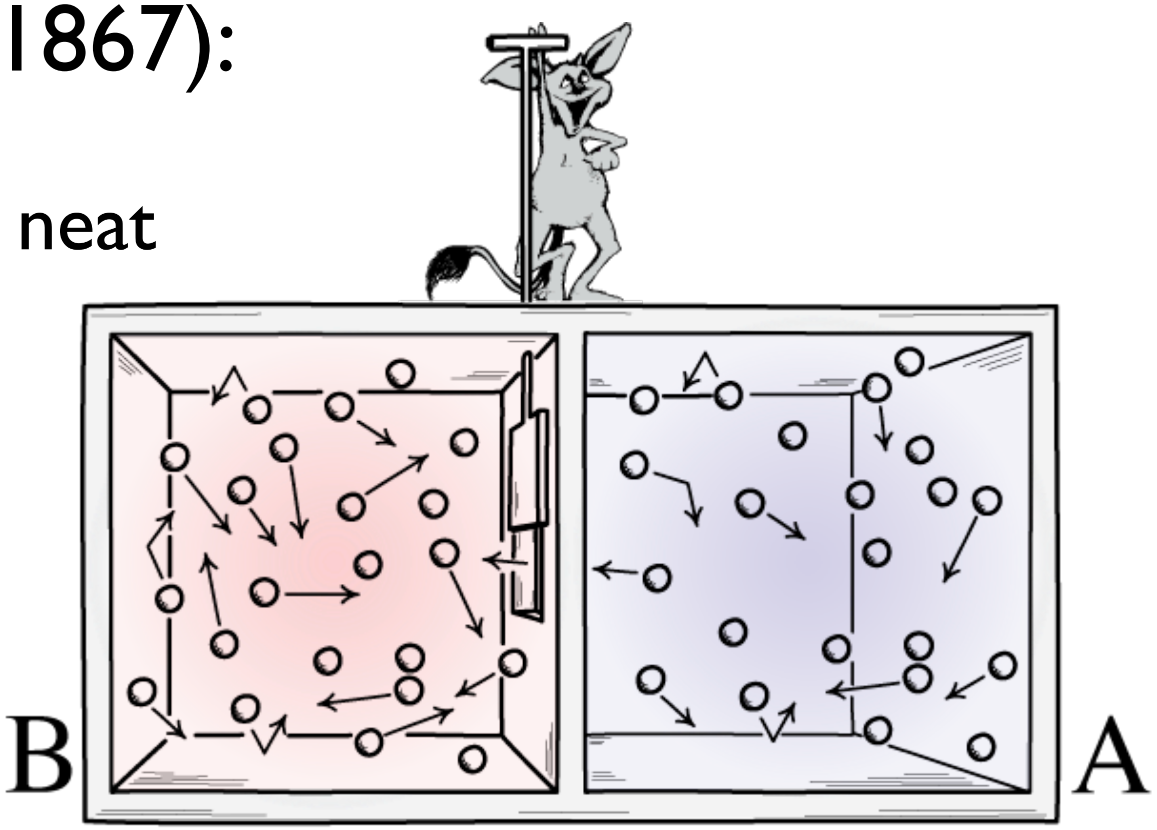
Take a different approach:

1) Investigate physical nature of information
2) Identify fundamental limits to information processing
3) Postulate one simple principle
4) Derive rules for learning strategies and derive concrete learning methods

Thermodynamic origins of information

- Discussions about the foundations of thermodynamics between Maxwell, Tait, Thompson, Clausius, and others...
- ...Maxwell's "demon" emerged (1867):

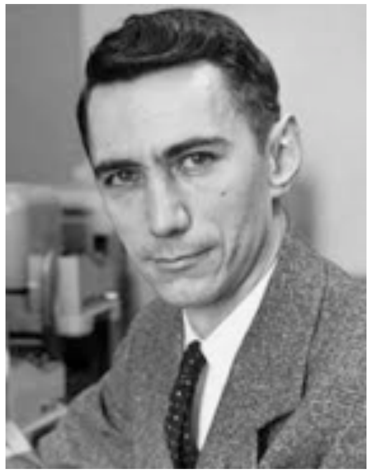
``very observant and neat fingered being"

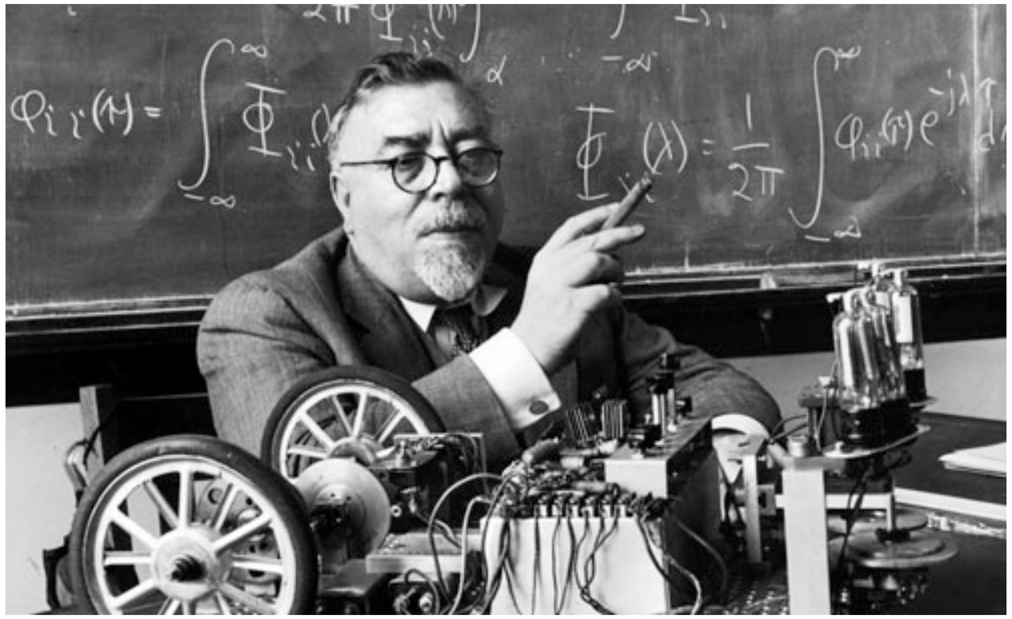


- ...and inspired many. (Smoluchowski 1924, Szilard 1929, Brillouin1951 Landauer 1961, Bennett, 1973, Zurek 1986, ...)

- Szilard’s 1929 work in particular outlined a physical foundation for information and information processing...

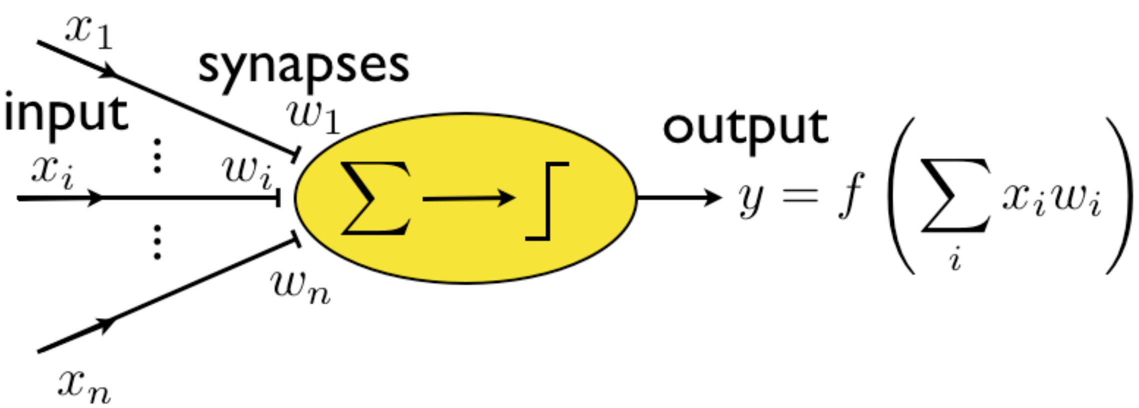
- ...inspiring more work attempting to go from an energy-based view of the nature to an information-based view (von Neumann, Wiener, Shannon, ...)





- ... and (among other things) specific studies of mathematical models of certain neurons and their connections...

- ..."first neural networks" emerged (McCulloch&Pitts 1943)


x_1
synapses
input
w_1
x_i
w_i
$\Sigma \longrightarrow \int$
output
$y = f\left(\sum_i x_i w_i\right)$
w_n
x_n

A Logical Calculus of Ideas Immanent in Nervous Activity

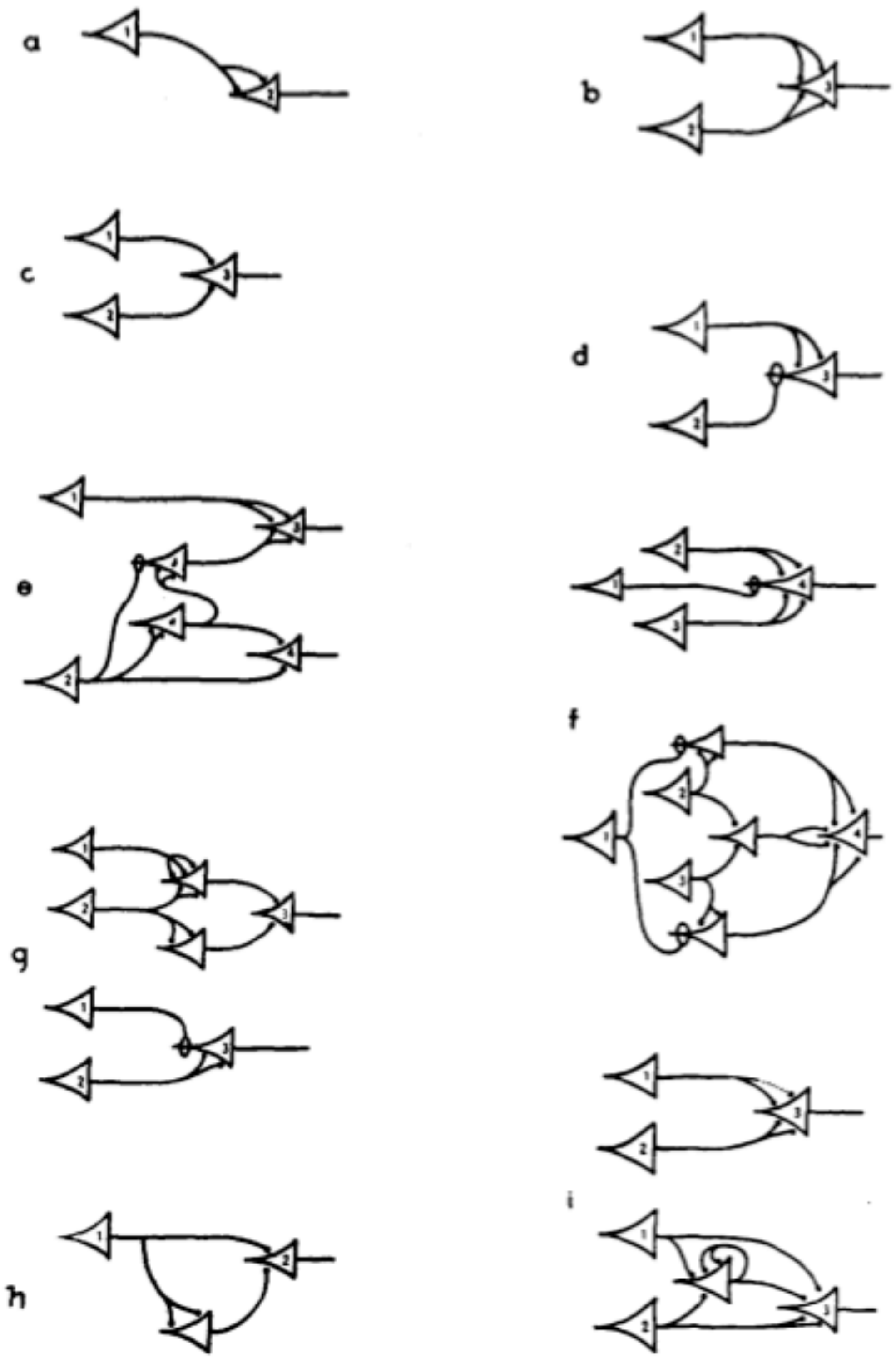


FIGURE 1

- (refined over the years, but some of the basic ideas still in use)

- ...and inspired first "learning machine", the Perceptron (Rosenblatt, 1957)

- Novikoff's perceptron convergence theorem (1962) inspired many...

- ... including Vapnik. Important insight: Empirical inductive inference implemented as minimization of empirical error is not consistent without further assumptions! There needs to be a restriction on the "complexity" of the function class, or its "capacity" for explaining data (Vapnik and Chervonenkis, 1971).

- ...This has shaped our modern view of machine learning as empirical inference with *finite* data (not error-free data).

- The *most ambitious version* of empirical, inductive inference: From empirical data to underlying laws.

  Experiences ⟹ Observer(s) ⟹ Laws

- Note the difference from deductive inference (e.g. ab initio calculations): From known laws to specific predictions.

  Laws ⟹ Explanations

- Note that machine learning often solves a *simpler problem*: From finite data to predictions. (don't need to know the underlying rules, just make good predictions; maybe this is what creatures do to survive?)

Usefulness of a model and its cost

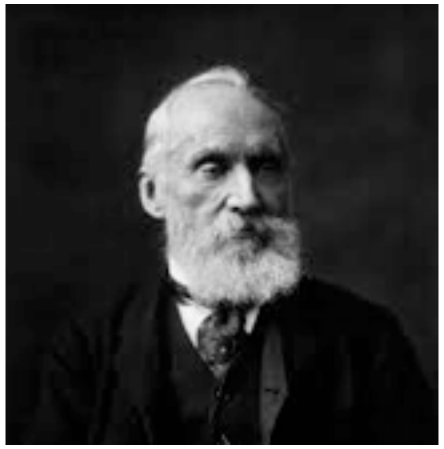
- This brings us back to the central question...

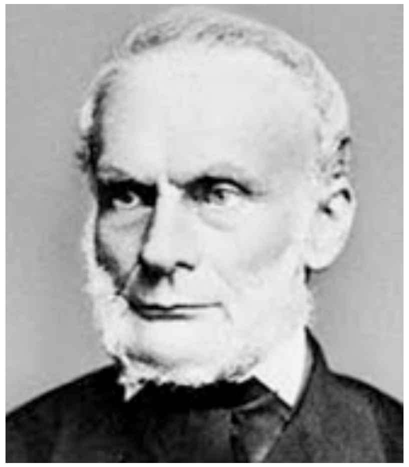
- there should be laws of nature (physical laws) from which rules for inductive inference can be derived -- in simple words: how should an observer represent the data in physical memory?

- Back to step 1) Investigate the physical nature of information

From Carnot's idealized heat engine to Szilard's information engine

- Kelvin, Clausius and others were inspired by work the french engineer Nicolas Léonard Sadi Carnot published in 1824.



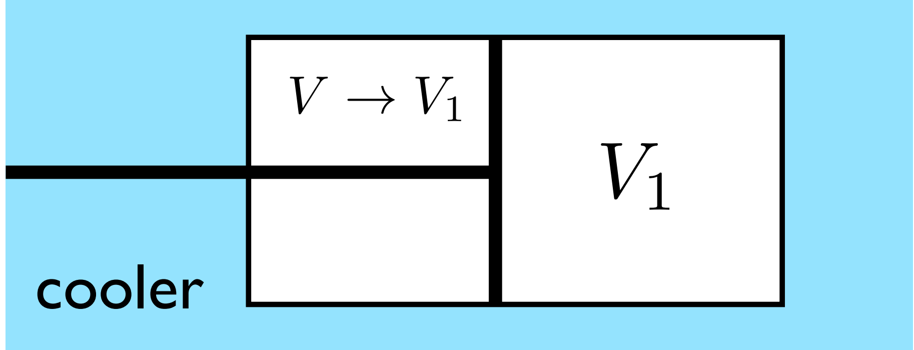


- We take a **shortcut,** Skipping Maxwell's „demon" (1867), and go straight from Carnot's *idealized heat engine* to Szilard's *idealized information engine*. (1929).

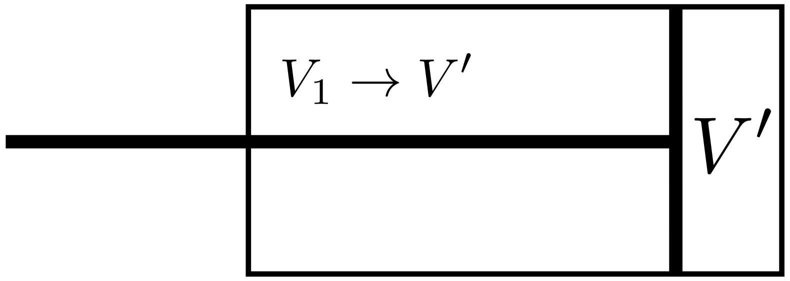
Carnot process

Work in

1) Isothermal compression from V to V_1 at low temperature T.

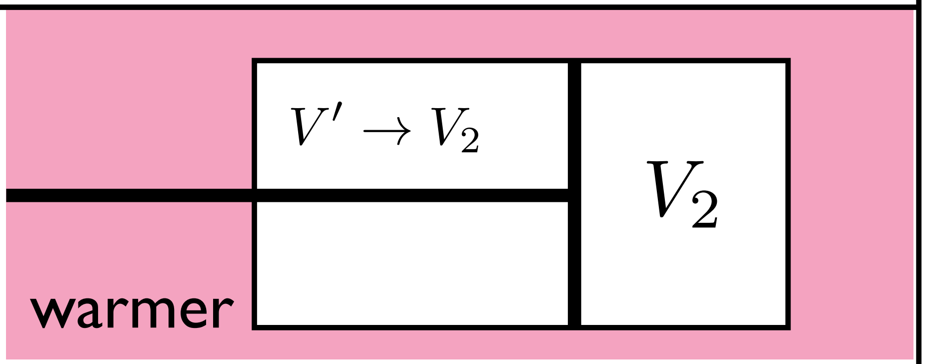


2) Isentropic compression to V'. Temperature of the gas changes from T to T'. $T \leq T'$

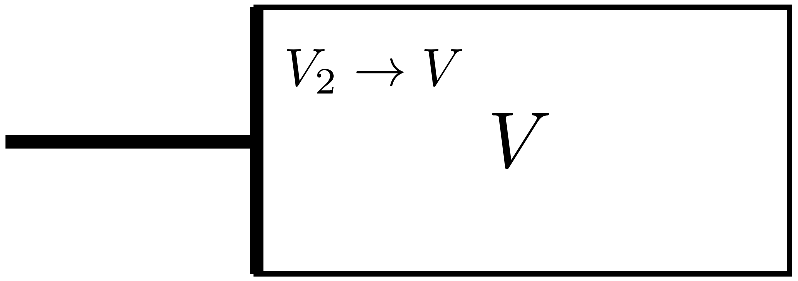


Work out

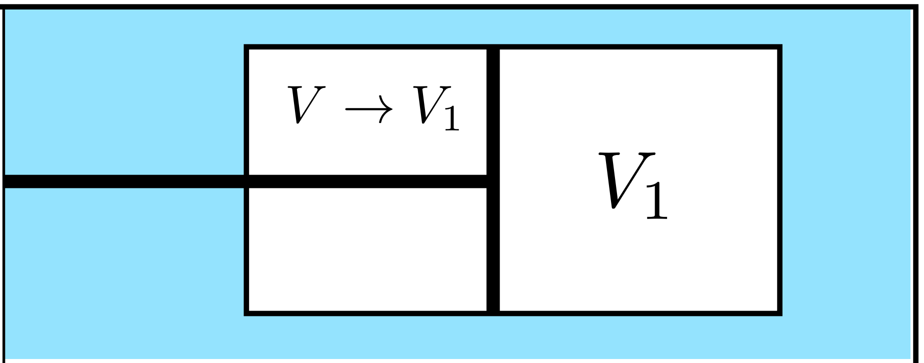
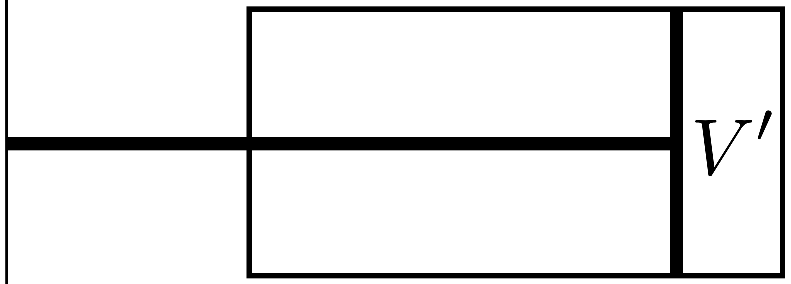
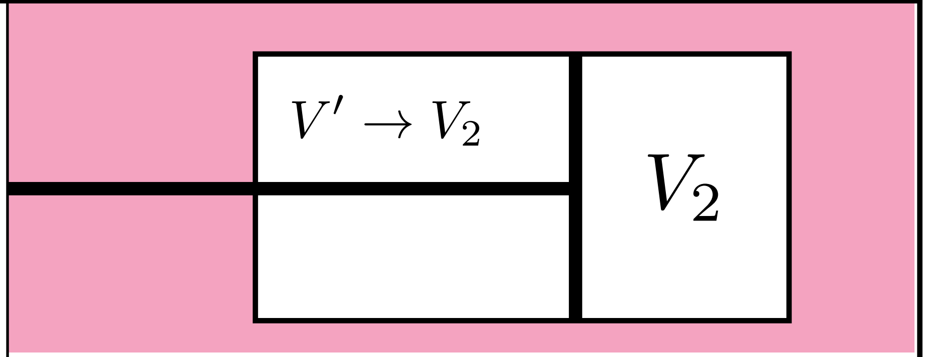
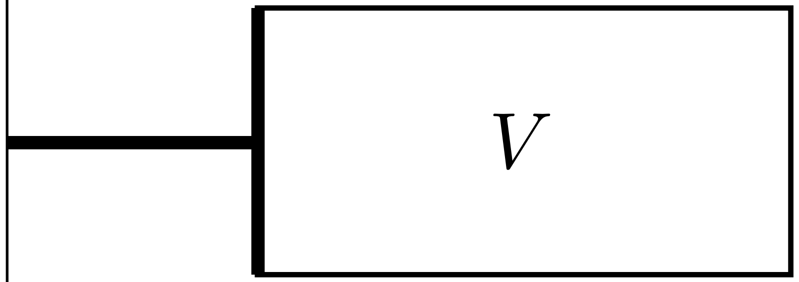
3) Isothermal expansion to V_2 at high temperature T'.



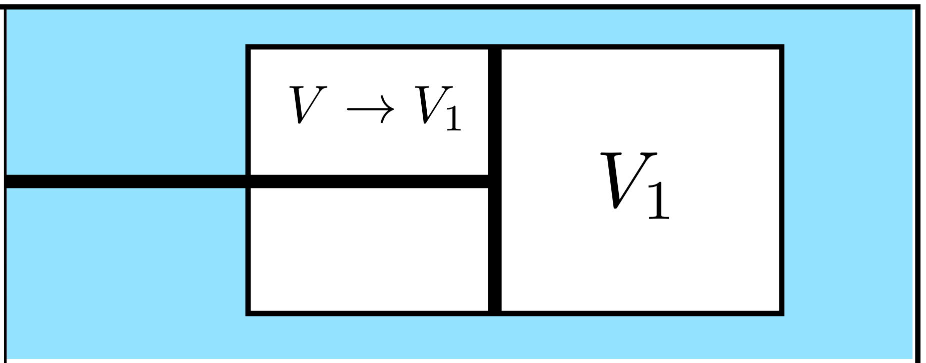
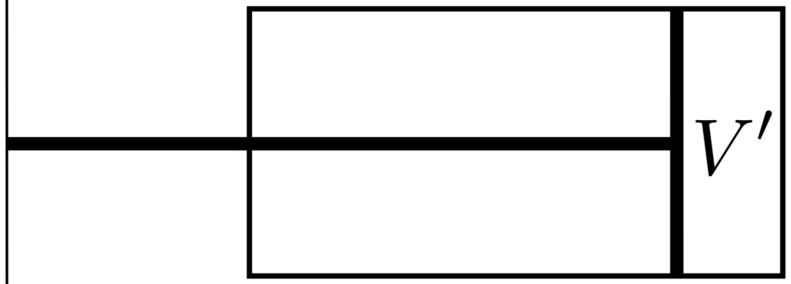
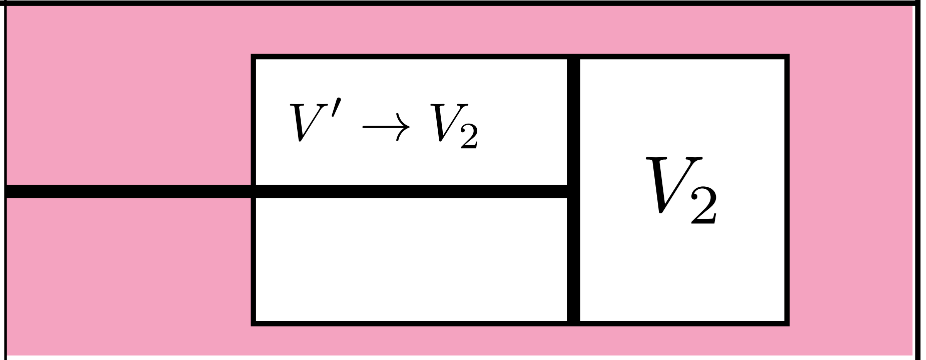
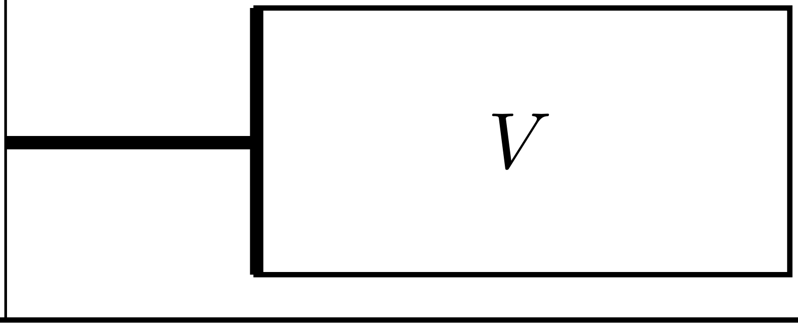
4) Isentropic expansion to V. Temperature change: T' to T



Carnot process

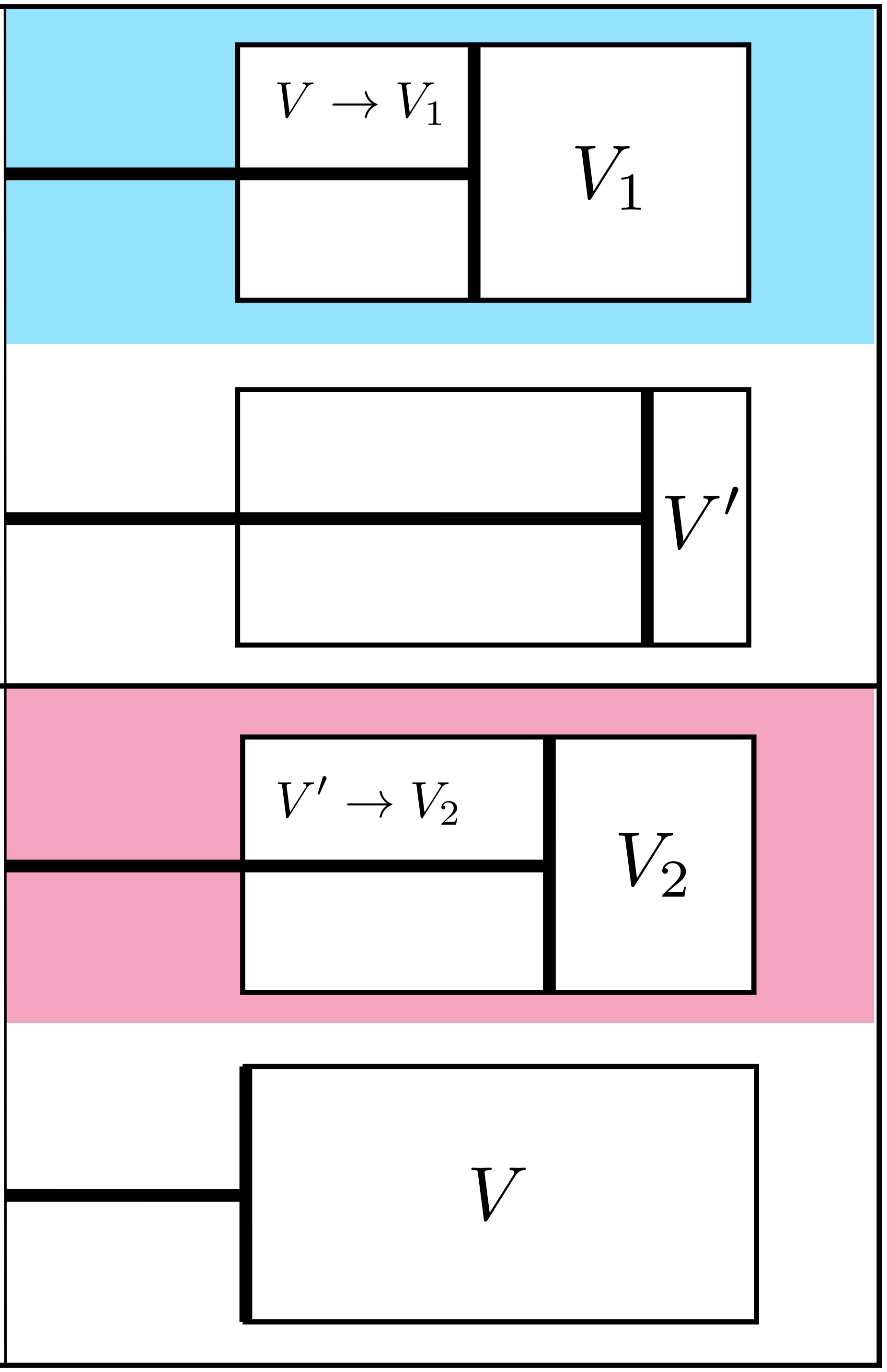
Work	**Heat**	
$W = NkT \ln\left[\frac{V}{V_1}\right]$	$= -Q$	
$W_{\text{compr}} = nC_{\text{v}} \int_{T}^{T'} dT$	$Q = 0$	
$-W' = NkT' \ln\left[\frac{V_2}{V'}\right]$	$= Q'$	
$W_{\text{expand}} = nC_{\text{v}} \int_{T'}^{T} dT$	$Q = 0$	

Compute total work out

Work	Heat	
$W = NkT \ln\left[\frac{V}{V_1}\right]$	$= -Q$	
$W_{\text{compr}} \xcancel{=} nC_{\text{v}} \int_{T}^{T'} dT$	$Q = 0$	
$-W' = NkT' \ln\left[\frac{V_2}{V'}\right]$	$= Q'$	
$W_{\text{expand}} \xcancel{=} nC_{\text{v}} \int_{T'}^{T} dT$	$Q = 0$	

Compute total work out

Work	**Heat**
$W = NkT \ln\left[\frac{V}{V_1}\right]$	$= -Q$
$-W' = NkT' \ln\left[\frac{V_2}{V'}\right]$	$= Q'$



Total work out =

$$-W' - W = Nk\left(T' \ln\left[\frac{V_2}{V'}\right] - T \ln\left[\frac{V}{V_1}\right]\right)$$

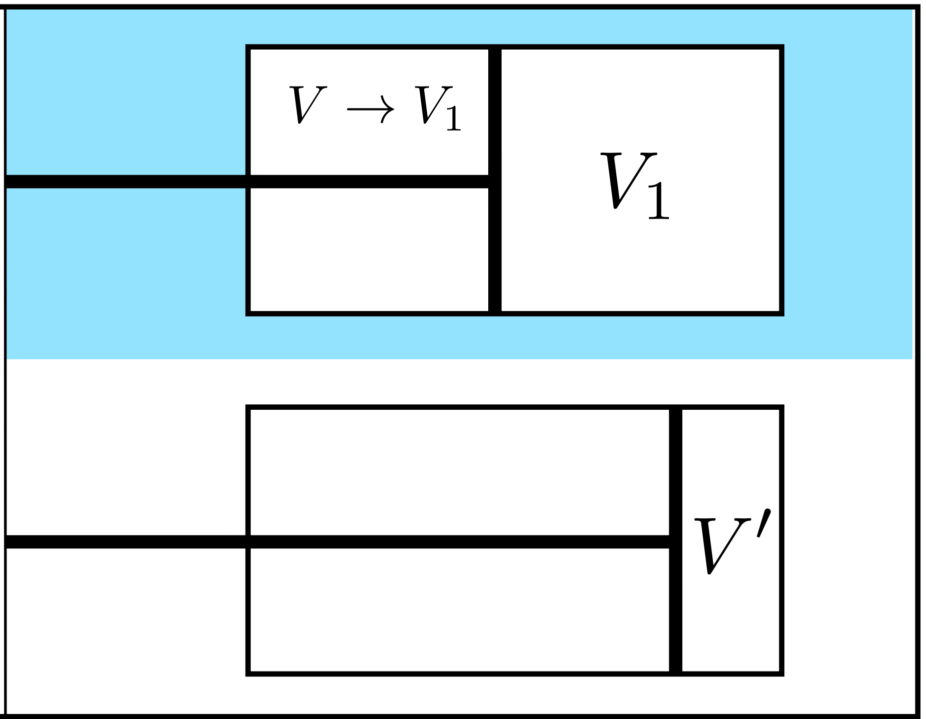
adiabatic equation implies: $\frac{V}{V_1} = \frac{V_2}{V'} \equiv \nu$

Total work out = $Nk(T' - T)\ln[\nu]$

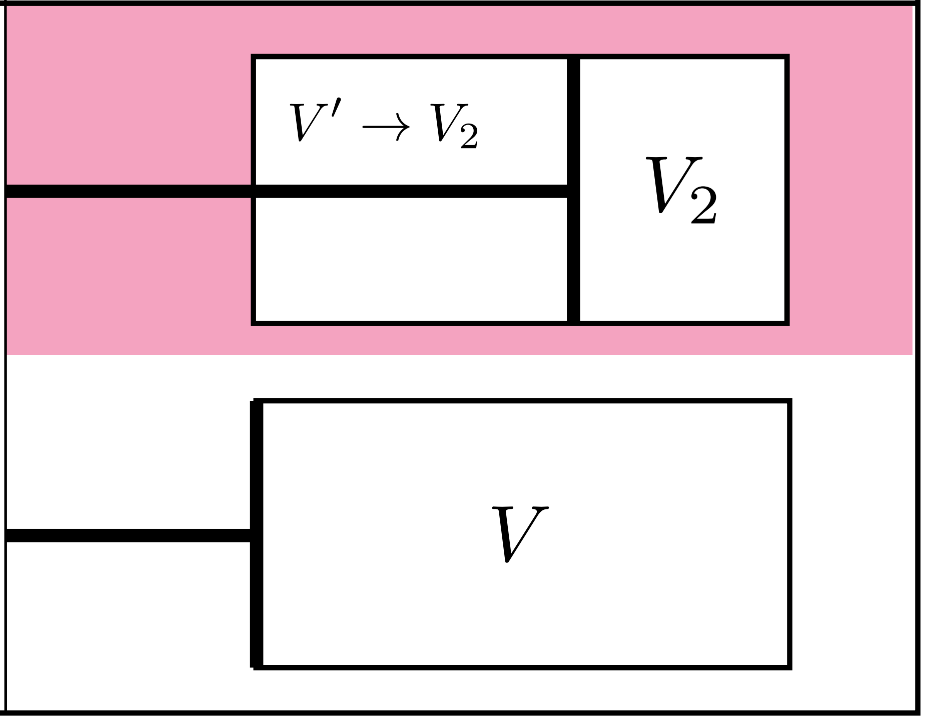
Carnot efficiency = $\dfrac{\textbf{total work out}}{\textbf{heat in at T'}}$

$$\eta_C := \frac{-W' - W}{Q'} = 1 - \frac{T}{T'}$$

$$W = NkT \ln[\nu] = -Q$$



$$-W' = NkT' \ln[\nu] = Q'$$



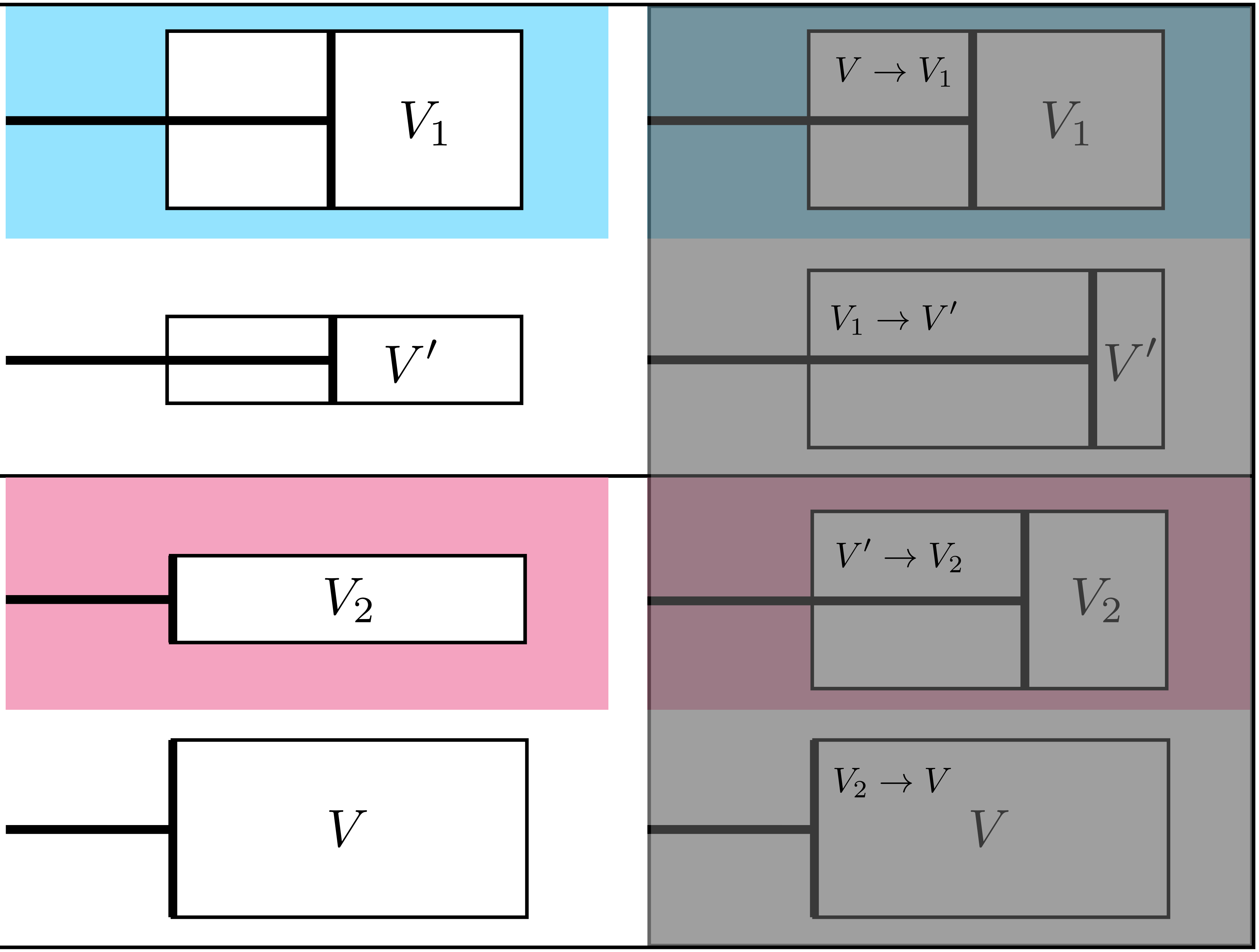
Carnot process equivalent

1) Isothermal compression V to V_1@ T.

2) Isentropic compression V_1 to V’. T to T’.

3) Isothermal expansion V’ to V_2@ T’.

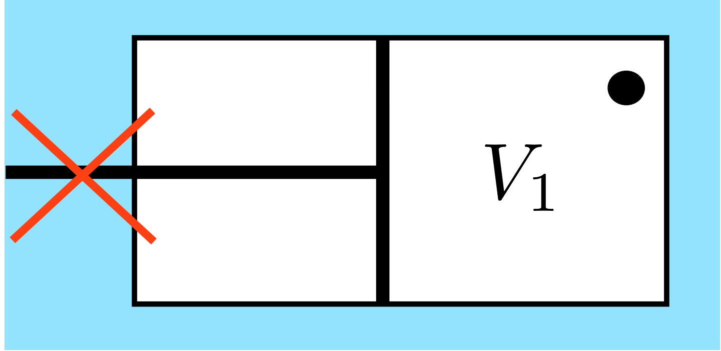
4) Isentropic expansion V_2 to V. T’ to T.



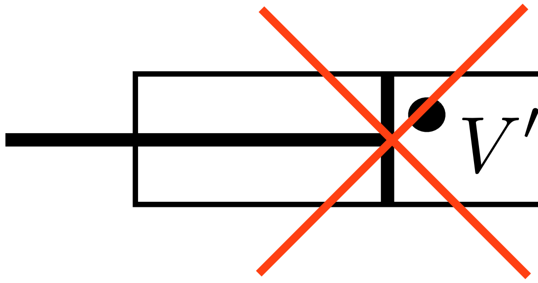
Carnot process to Szilard engine

One particle gas!

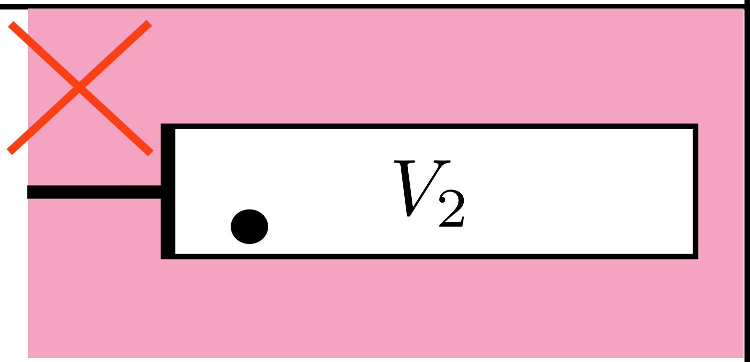
1) Isoth. comp. to V_1@ T.



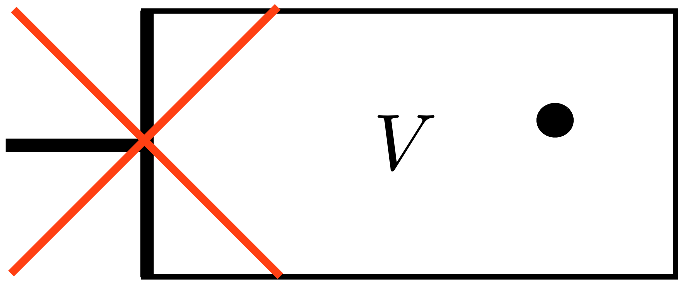
2) ~~Isentr. comp. to V'. T to T'.~~



3) Isoth. exp. to V_2@ T'.



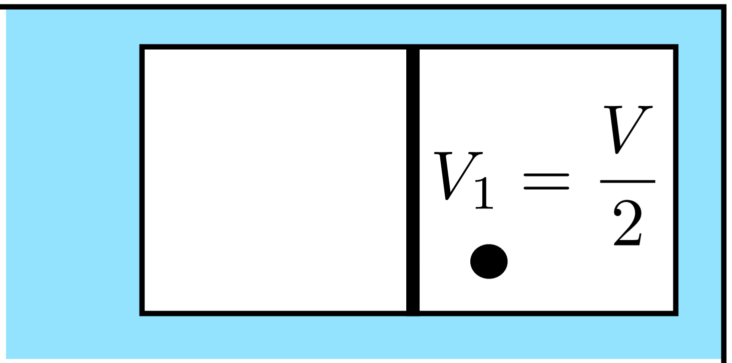
4) ~~Isentr. exp. to V. T' to T.~~



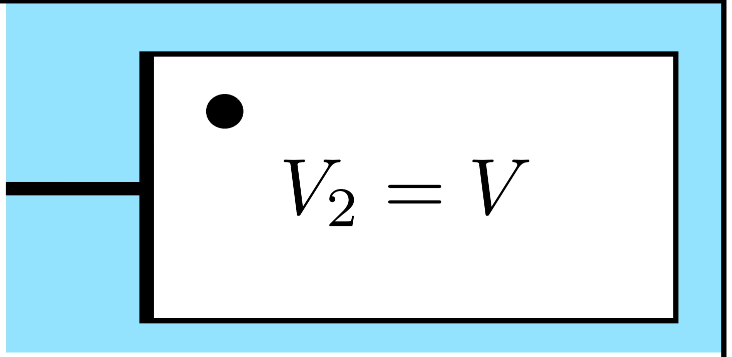
1) Insert wall

$$W = 0$$

no work



2) Isothermal expansion

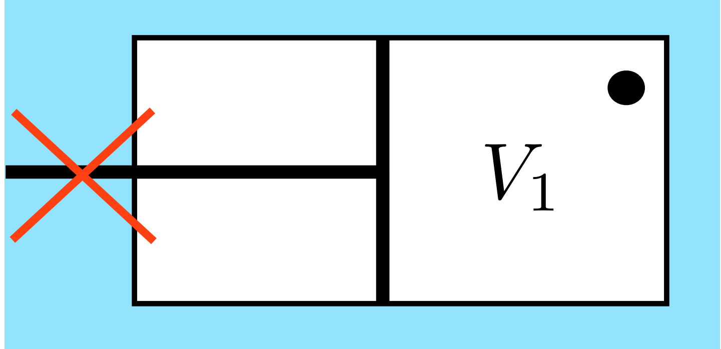


$$-W' = kT\ln[2]$$

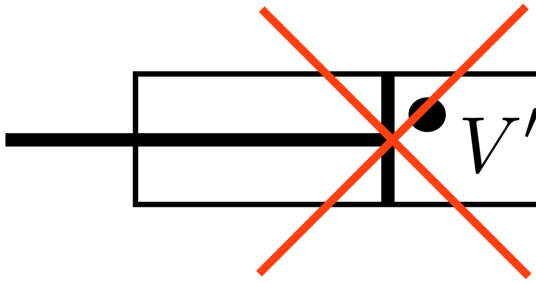
Carnot process to Szilard engine

One particle gas!

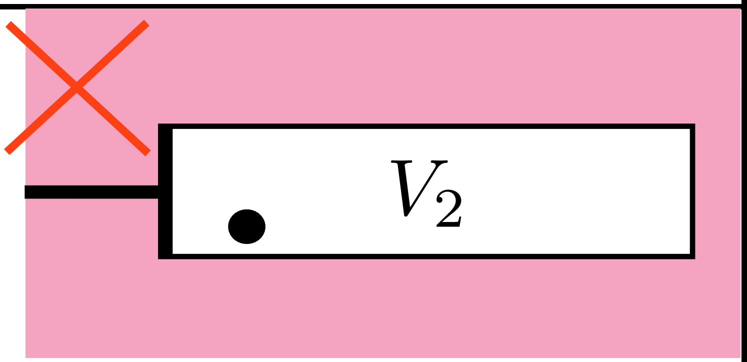
1) Isoth. comp. to V_1@ T.



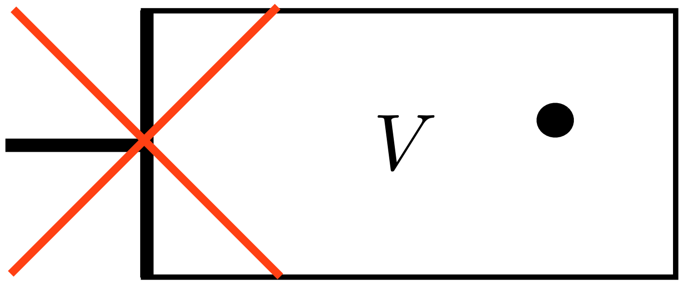
2) ~~Isentr. comp. to V'. T to T'.~~



3) Isoth. exp. to V_2@ T'.

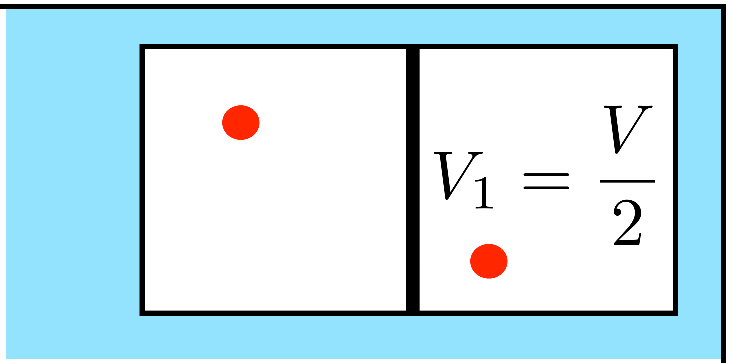


4) ~~Isentr. exp. to V. T' to T.~~



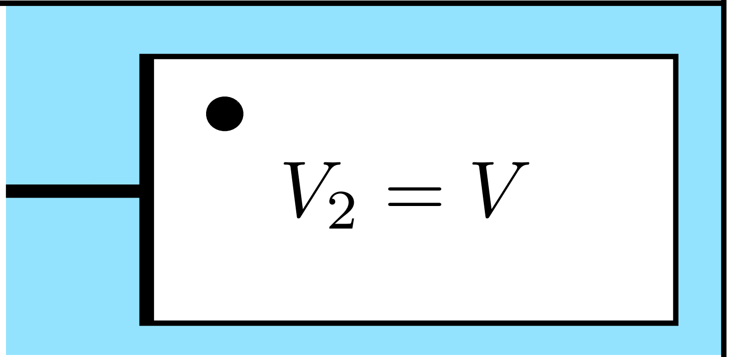
1) Insert wall

$$W = 0$$



?

2) Isothermal expansion

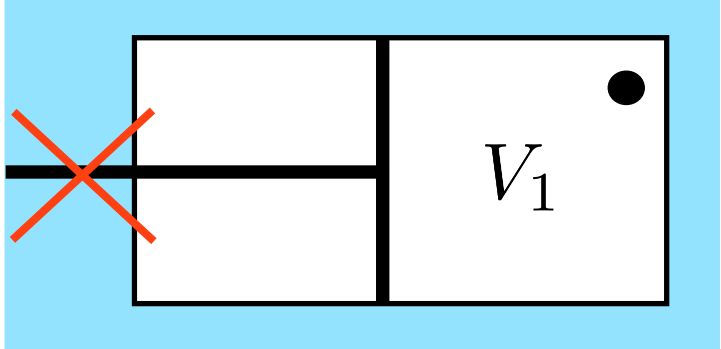


$$-W' = kT\ln[2]$$

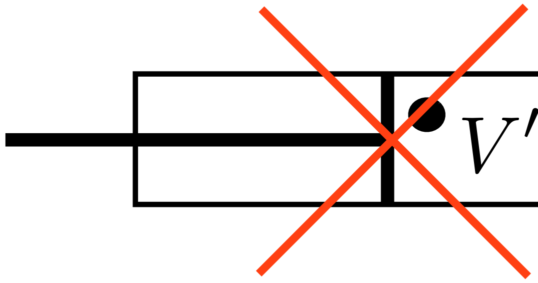
Carnot process to Szilard engine

One particle gas!

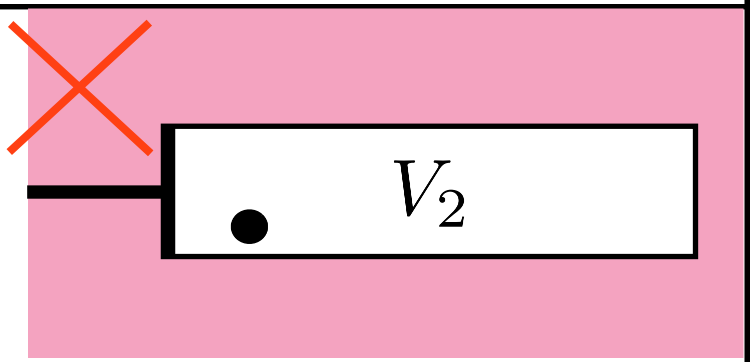
1) Isoth. comp. to V_1@ T.



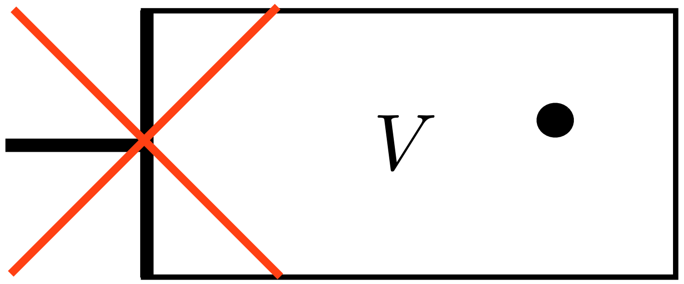
2) ~~Isentr. comp. to V'. T to T'.~~



3) Isoth. exp. to V_2@ T'.

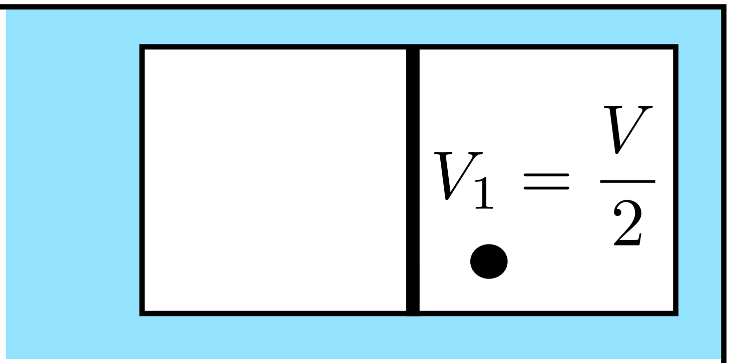


4) ~~Isentr. exp. to V. T' to T.~~



1) Insert wall

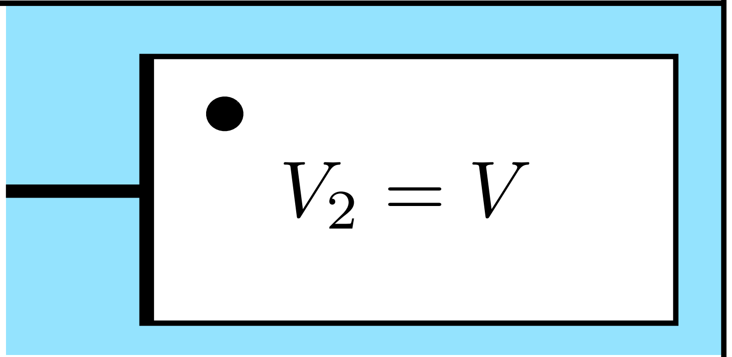
$W = 0$



2) **Must measure and remember particle location!**

must be dissipative

3) Isothermal expansion

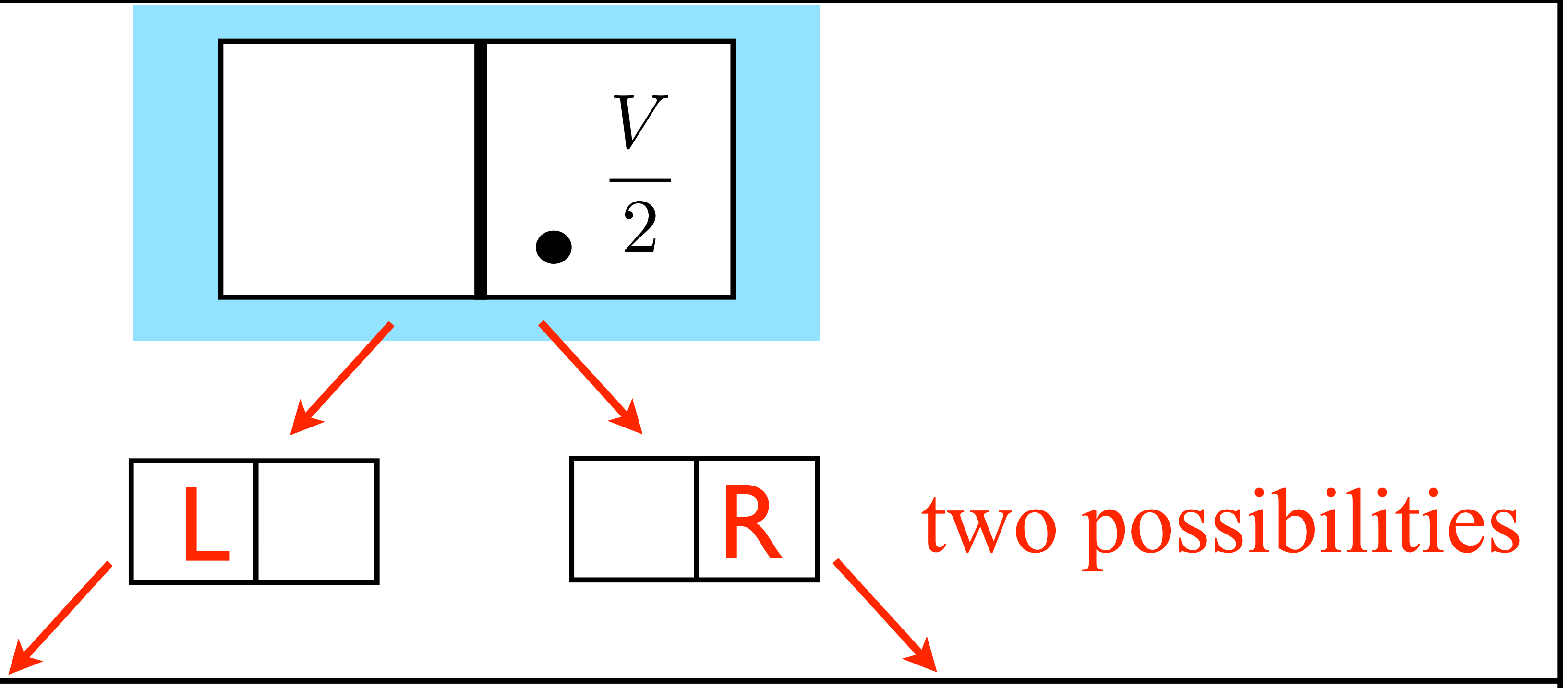


$-W' = kT \ln[2]$

Szilard engine

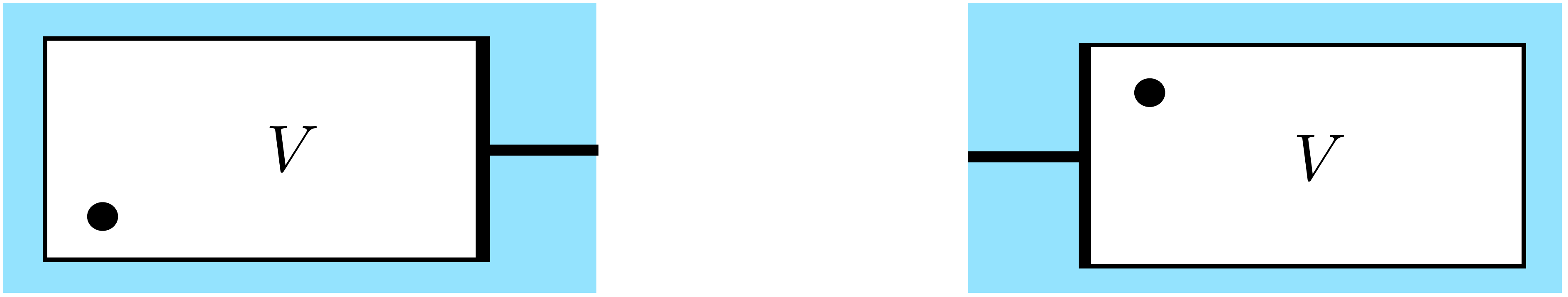
1) Insert wall

2) Measure and remember particle location.



two possibilities

3) Isothermal expansion

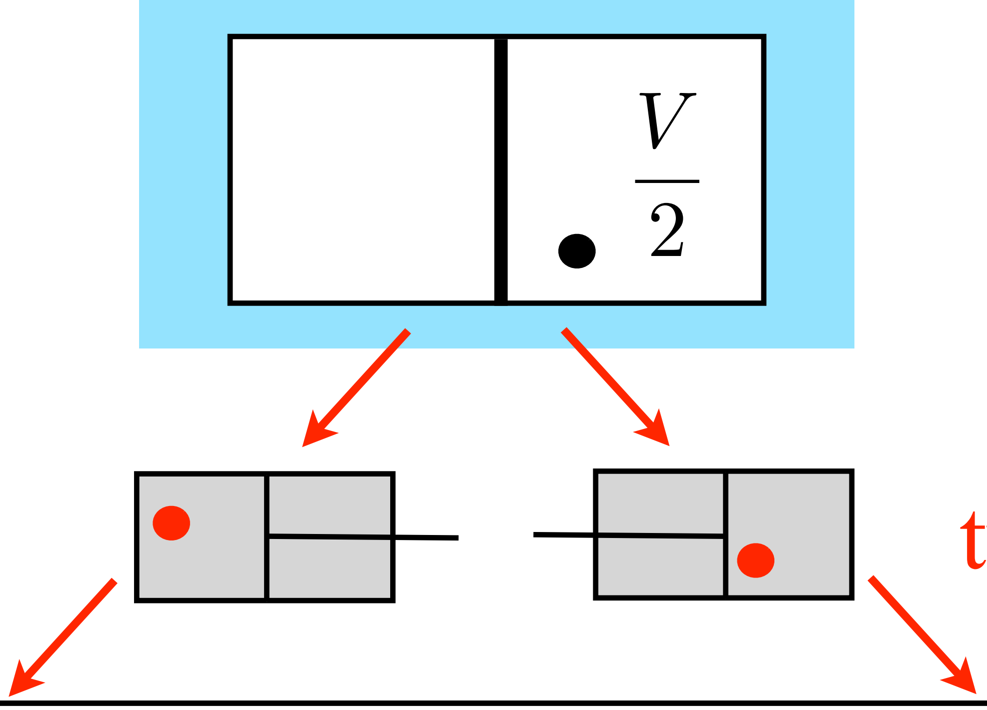


Simple, concrete implementation

1) Insert wall

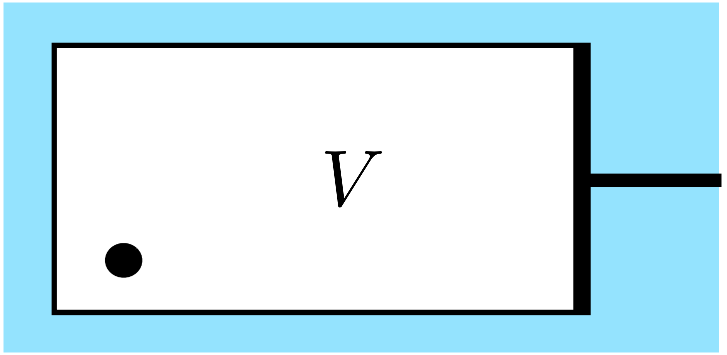
$$W = -Q = 0$$

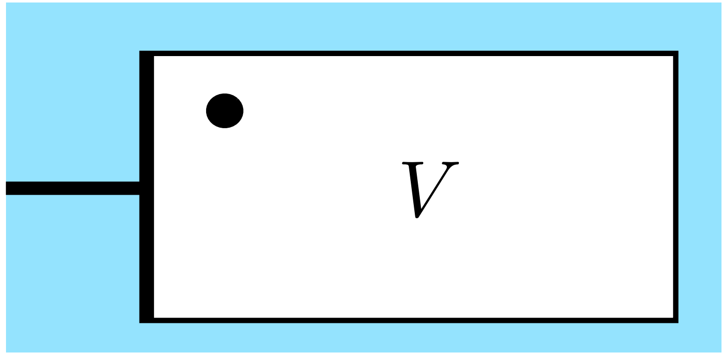
2) Measure and remember particle location: isothermal compression



two possibilities

3) Isothermal expansion





Close the cycle:

1) Insert partition

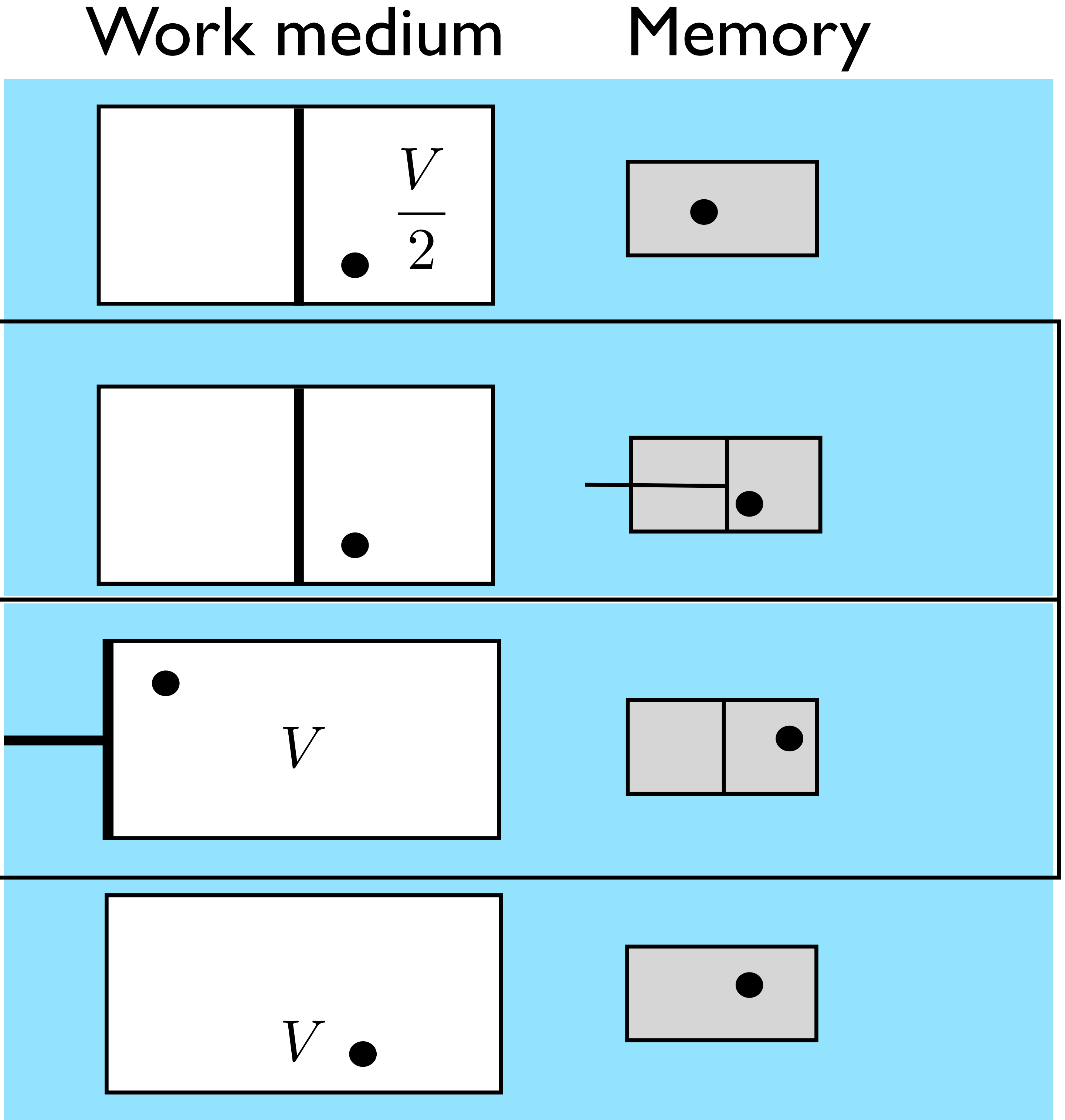
2) Measure and remember particle location: isotherm. comp. mem.

$W_M = -Q_M = kT \ln[2]$

3) Isothermal expansion of system

$-W' = Q' = kT \ln[2]$

4) Pull out partition



Szilard engine:

1) Insert partition

2) Measure and remember

Work into information

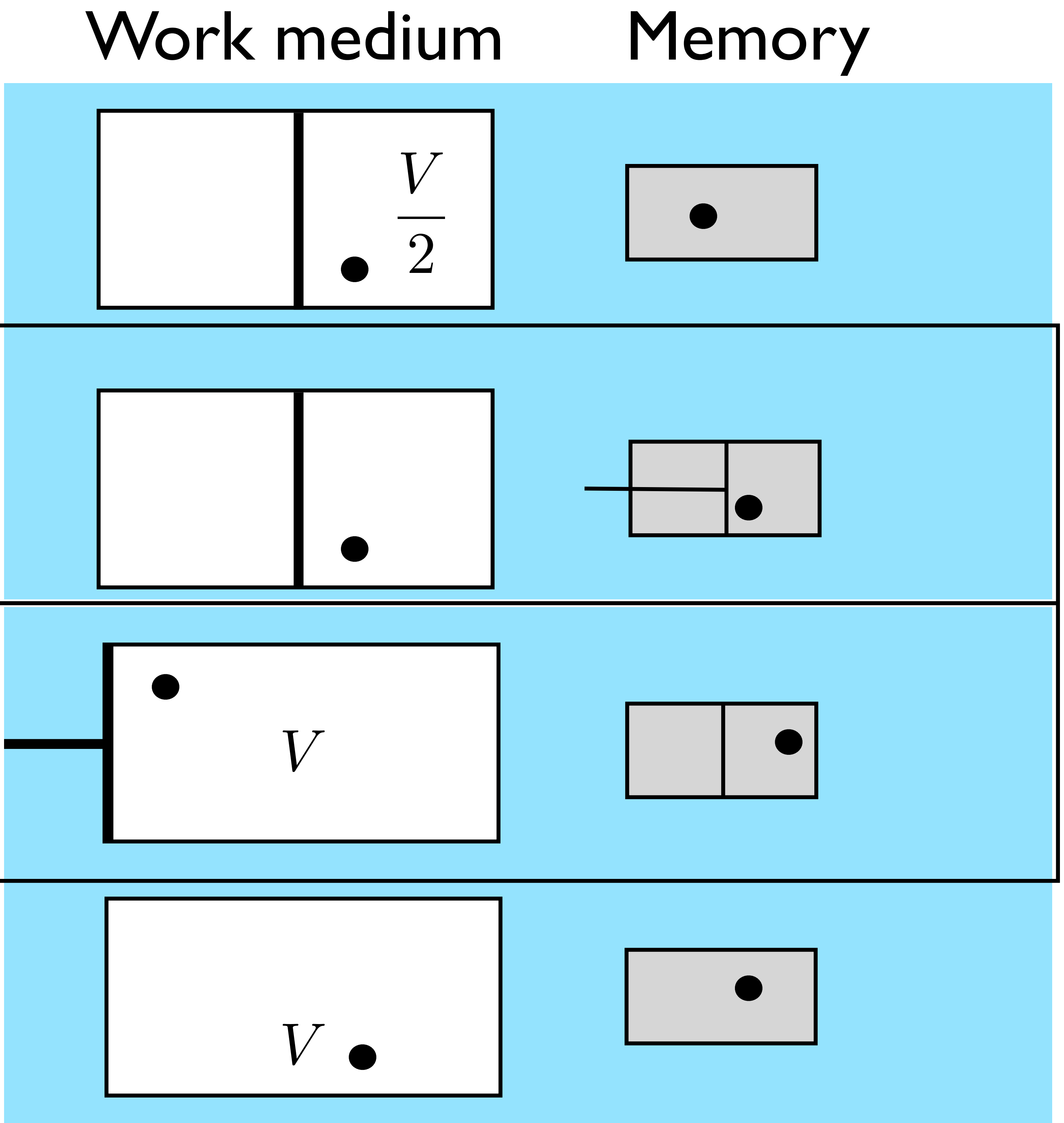
$$W_M = -Q_M = kT\ln[2]$$

3) Isothermal expansion

Information into work

$$-W' = Q' = kT\ln[2]$$

4) Pull out partition



No total work out: $-W' - W_M = 0$

Szilard engine:

1) Insert partition into system and pull out partition in memory

2) Measure and remember

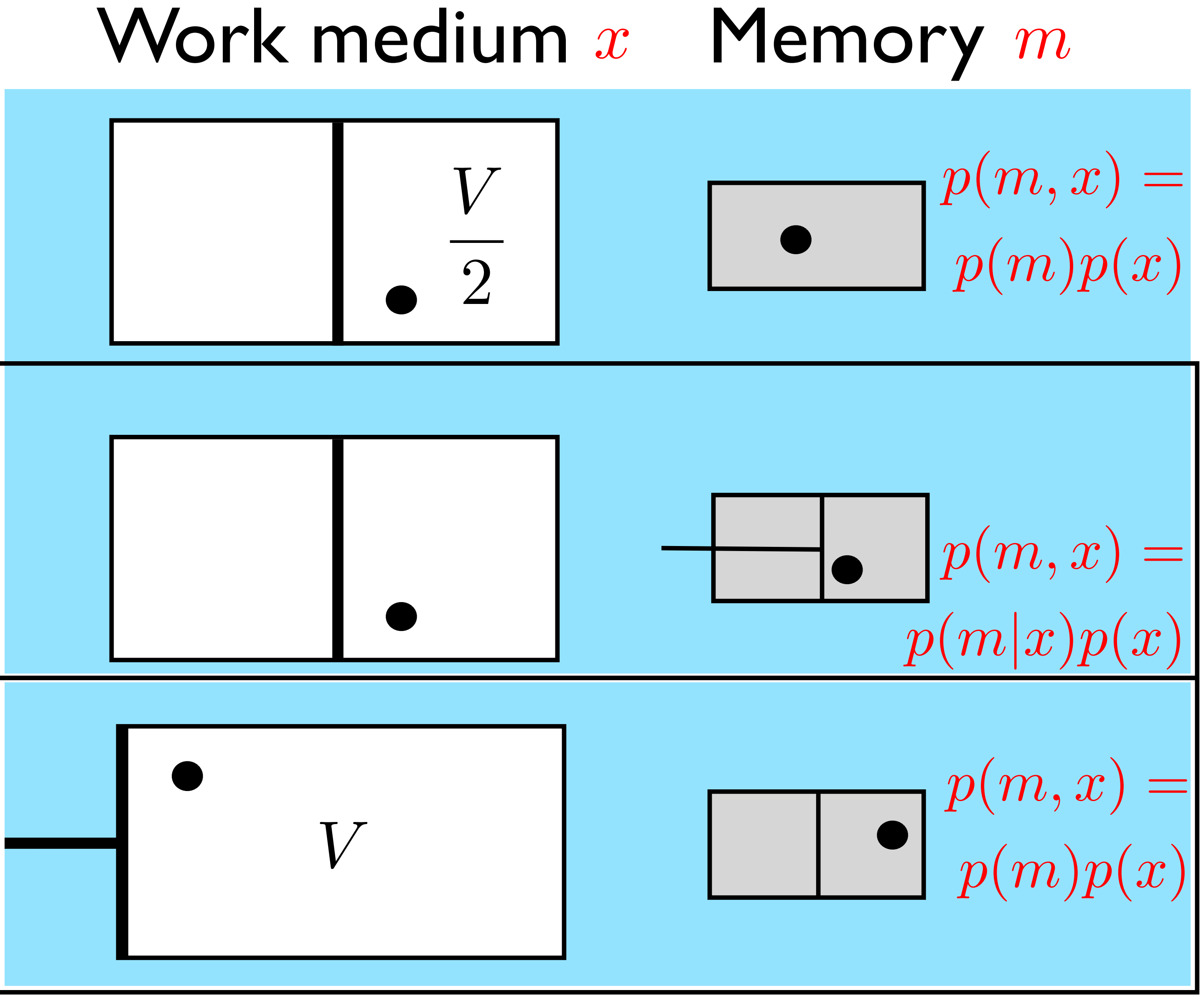
Work into information

$$W_M = -Q_M = kTI[m,x]$$

3) Isothermal expansion

Information into work

$$-W' = Q' = kTI[m,x]$$



Mutual information between the particle in the original box (x) and the particle in the memory (m) = entropy change (Gibbs-Shannon entropy)

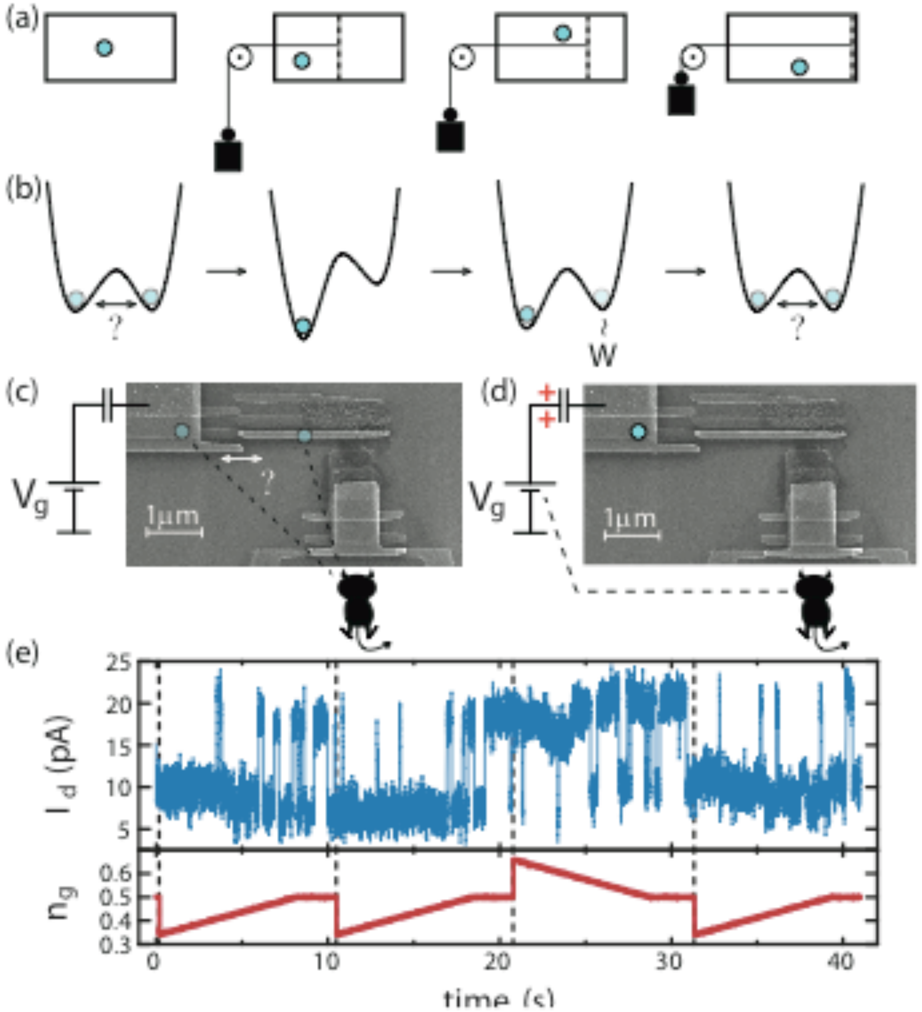
$$I[m,x] = H[m] + H[x] - H[m,x]$$

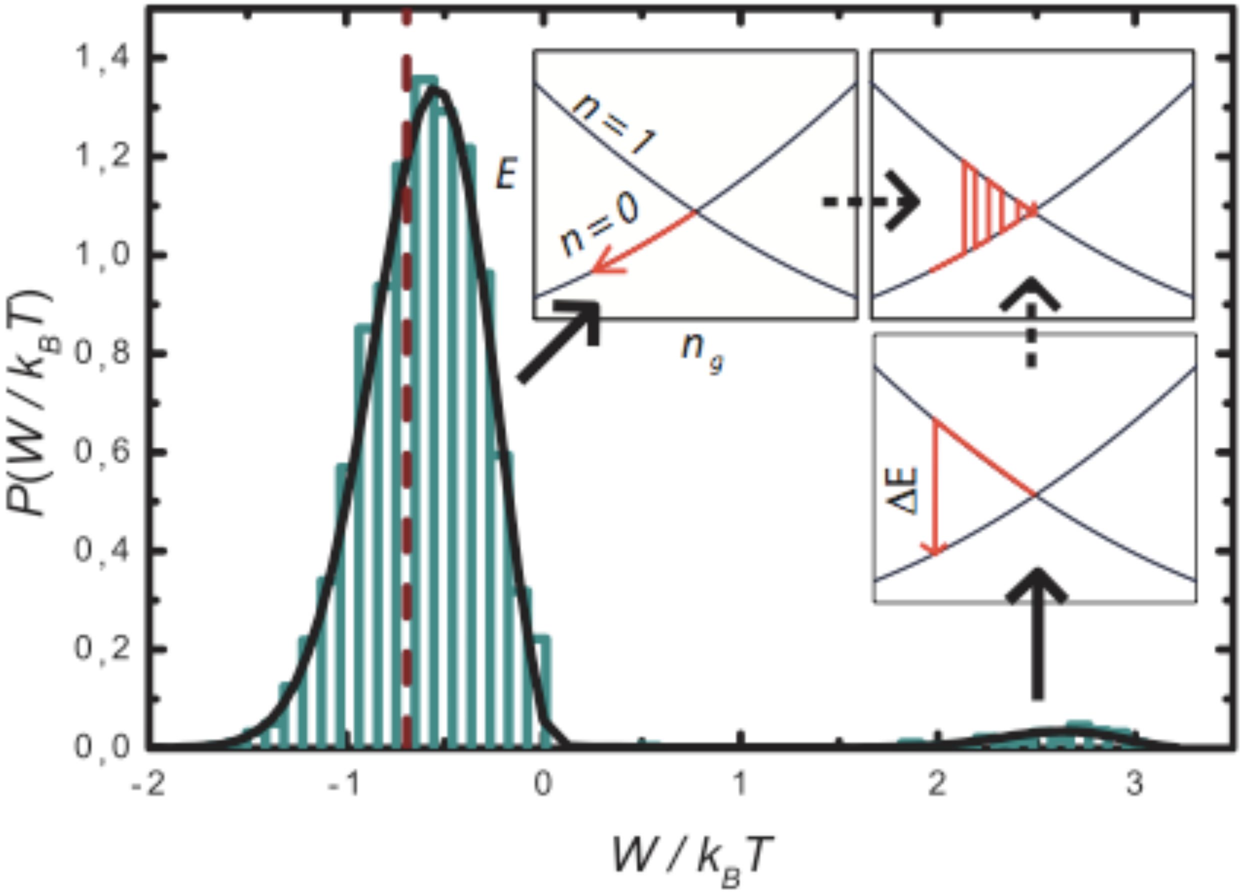
Experimental verification

- By now has been done with a variety of different systems. One example: single electron box (PNAS 2014):

Experimental realization of a Szilard engine with a single electron

J. V. Koski *,[1] V. F. Maisi,[1,2] J. P. Pekola †,[1] and D. V. Averin[3]

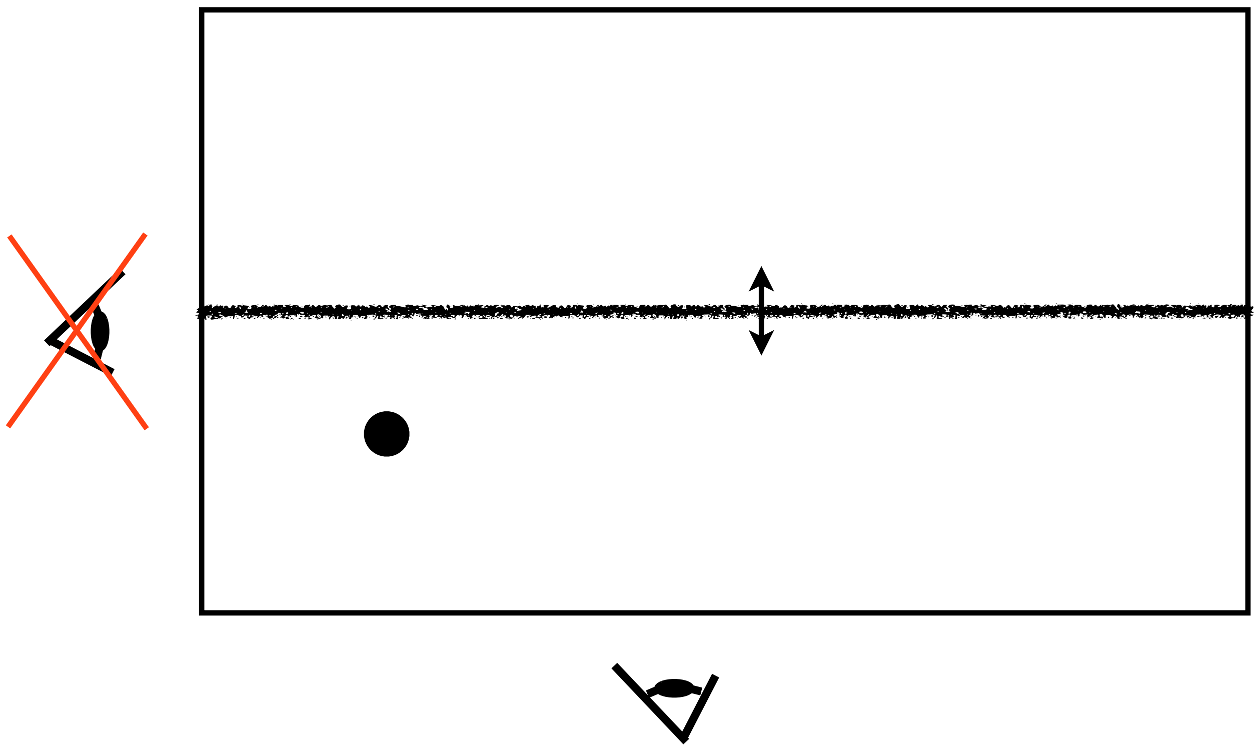




Assumptions in this discourse

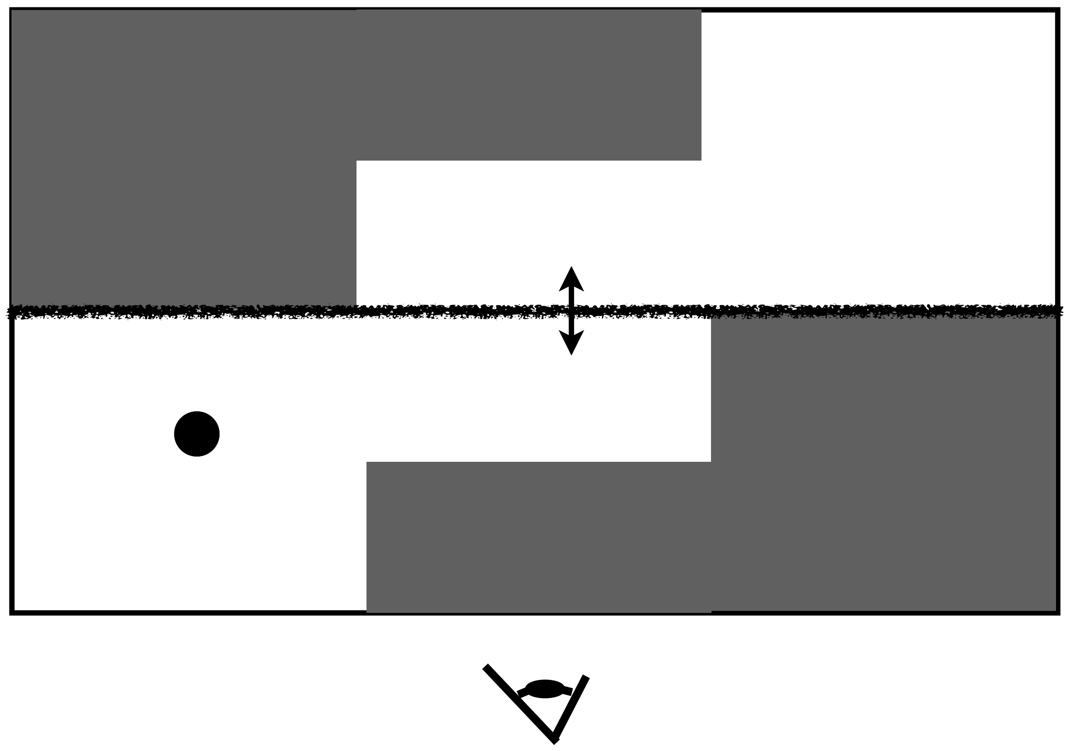
Idealized systems	Real world learning systems
• Can measure the relevant quantity	• Encounter partially observable environments **our focus today!**
• Can afford to move arbitrarily slowly	• Run at finite rates
• Have complete control	• Have limited control
➡ Can achieve the *ultimate* information-to-work conversion limit	➡ Are there tighter bounds?

Example: Modified Szilard box



- Partition moves along y-axis
- But: observer can measure only x-position

can’t extract any work (on average)

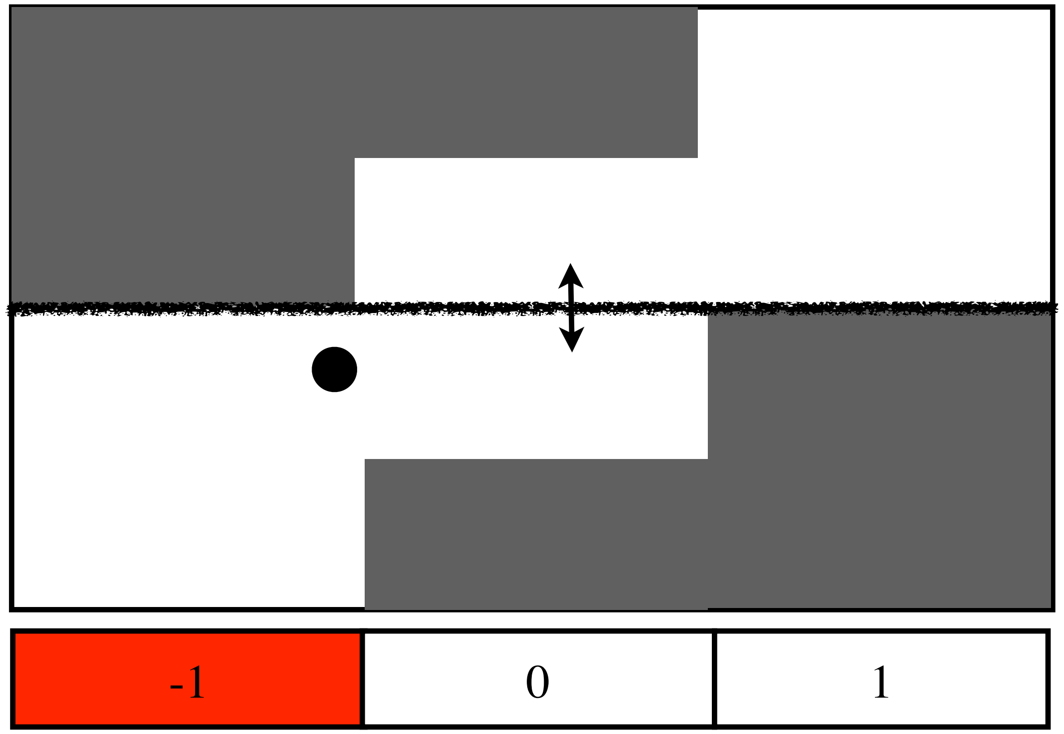


- Particle in box with excluded regions

➡ correlations

can extract work

How much work?


memory: m =

- Note: Fully informative memory costs more than it can yield!
- Captures I[m,x] = ln(3) => costs at least kT ln(3)
- Yields at most $\frac{2}{3}$ kT ln(2)
- Dissipation (= work lost) over cycle can, on average, be no less than kT (ln(3) - $\frac{2}{3}$ ln(2)) > 0 ... ultimate bound unachievable!

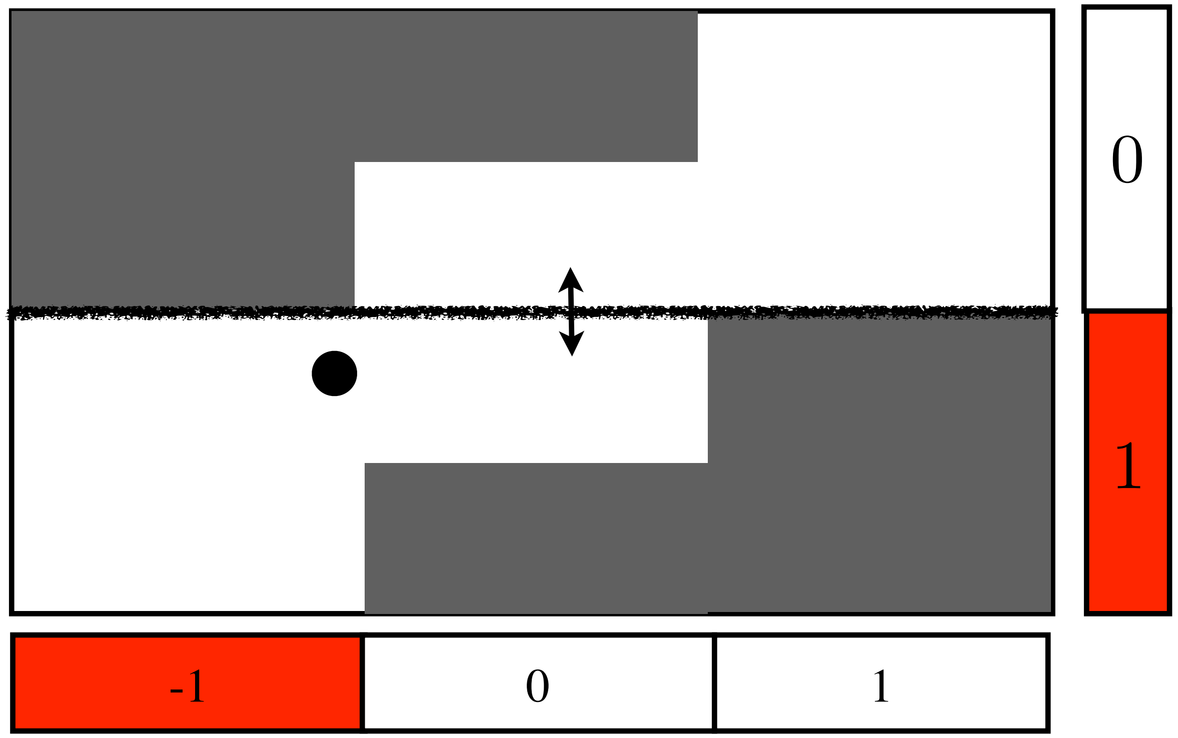
If the ultimate bound of zero is unachievable, we need to ask:

- Is there a *tighter, more meaningful, bound* on dissipation than zero?

Equivalent: Are partially observable information engines subject to a tighter bound on dissipation than that pointed out by Szilard and Landauer?

If so, could this bound inform how observers ought to process information efficiently?

Bound on dissipation?

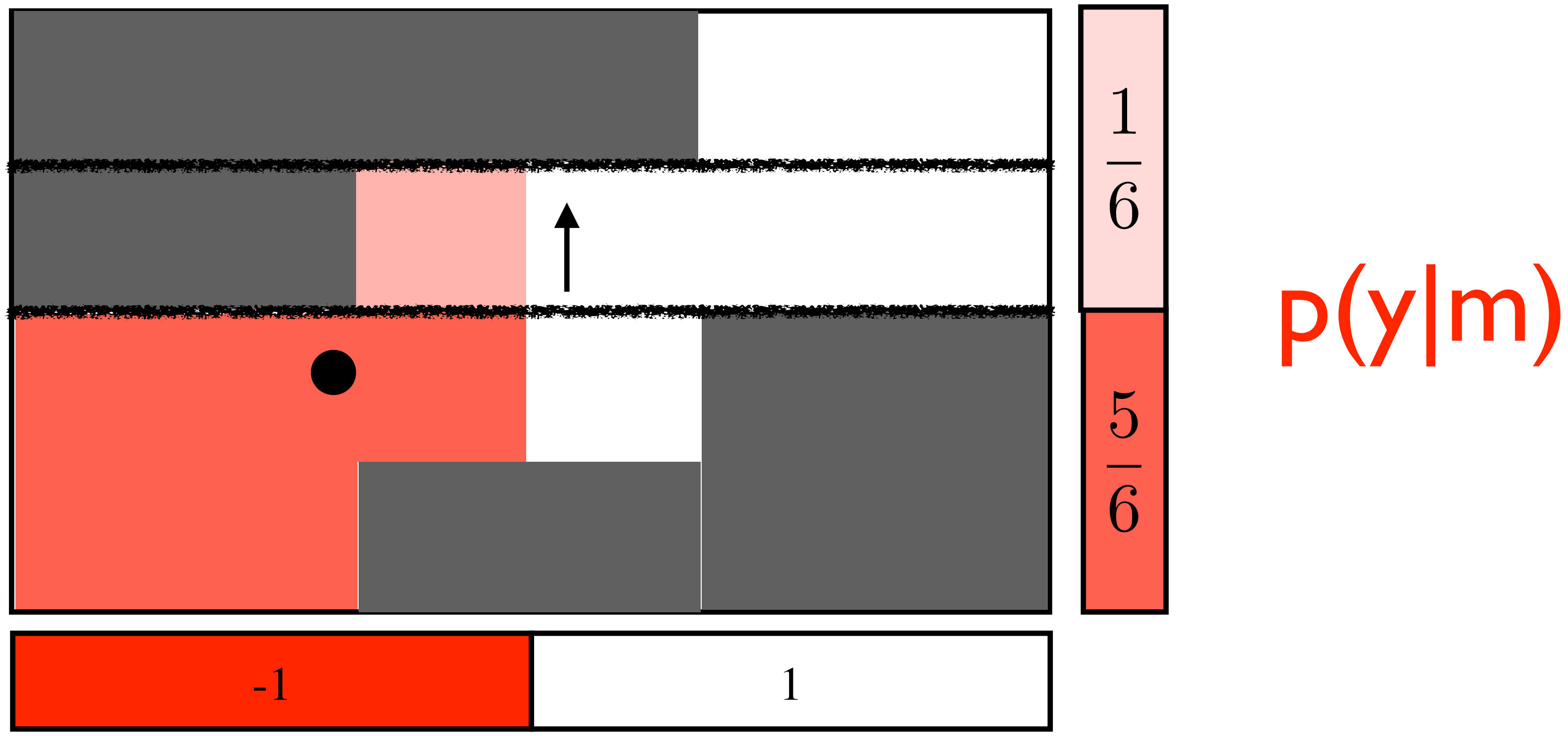


p(y|m)

memory: m =

- Important insight: Observer can not turn all information about the x position into work. Observer has to use memory to make an inference about the y position.Only relevant information (about the y position of the particle) can be turned into work!
- This memory indeed captures $I[m,y] = \frac{2}{3}\ln(2)$ relevant info.

Can we reach zero dissipation with a different memory?



memory: m =

- Lower the costs by making a less informative memory?
- Captures I[m,x] = ln(2) => costs at least kT ln(2)
- Yields at most kT $(\frac{5}{6}\ln(5) - \ln(3))$ = kT I[m,y]
- Dissipation at least kT $(\ln(2)+\ln(3)-\frac{5}{6}\ln(5)) > 0$

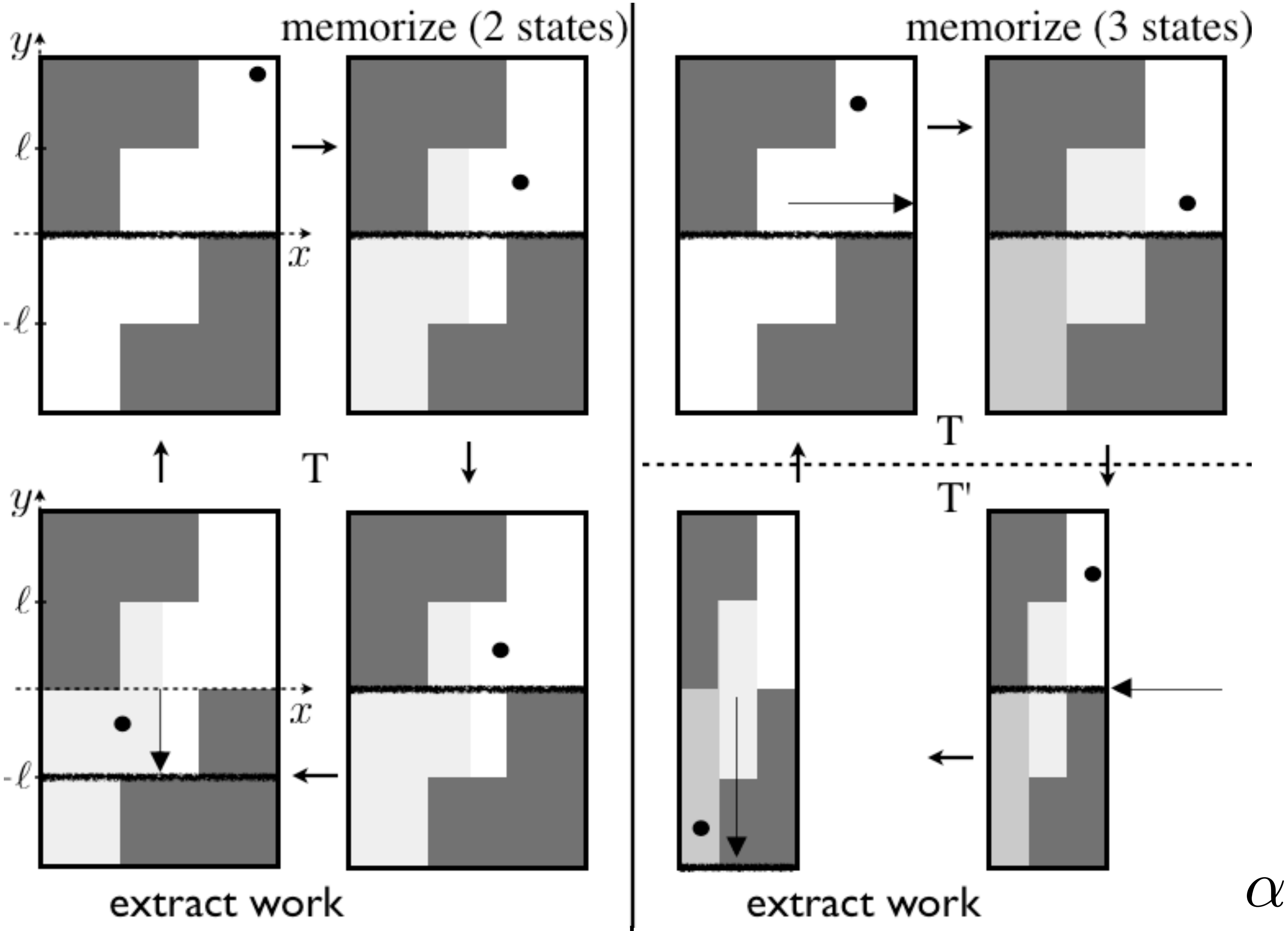
General bound

We will show that for generalized partially observable information engines:

- Relevant information is tighter bound on the work yield than total memorized information.
- Dissipation is bound by a trade off between total memory and relevant information.
- For isothermal information engines, dissipation is lower bound by irrelevant information.

Generalized information engines can run the memory forming step at a different temperature than the work extraction step.

Example: a “Szilard-Carnot” process



$$\alpha = \frac{T'}{T}$$

- Cost: $kT\ln(2)$ | $kT\ln(3)$
- Gain: $kT'\left(\frac{5}{6}\ln(5) - \ln(3)\right)$ | $\frac{2}{3}kT'\ln(2)$

$$\boxed{\alpha^* = \frac{\ln(3) - \ln(2)}{\ln(3) + 2\ln(2)/3 - 5\ln(5)/6} \simeq 1.847}$$

Illustrative simulations

- Modified Szilard box, two-state memory
  robshaw.net:8000/movie2/
- Modified Szilard box, three-state memory
  robshaw.net:8000/movie3/
- Modified Szilard box, three-state memory,
  with lateral adiabatic compression
  robshaw.net:8000/movie4/

- Carnot cycle animation: robshaw.net:8000/carnot/
- Szilard box animation: robshaw.net:8000/movie1/

General analysis

- Uses concepts that appear in “far-from-equilibrium” thermodynamics

  (An emerging area with much progress over the last 2 decades, pioneered by C. Jarzynski’s 1997 work relation...

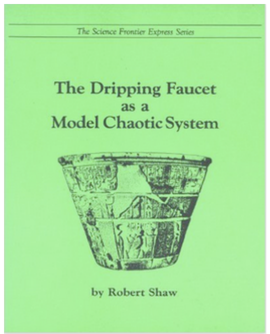
$$\left\langle e^{-\beta W} \right\rangle = e^{-\beta \Delta F}$$

...and G. Crooks’ 1998 detailed fluctuation theorem relating time reversal to dissipation)

Core concept used here: *nonequilibrium free energy* associated with a nonequilibrium distribution p:

$$\boxed{F = \langle E \rangle_p + kT \langle \log(p) \rangle_p}$$

- Main insight: there is additional free energy out of equilibrium:


The Science Frontier Express Series

The Dripping Faucet as a Model Chaotic System

by Robert Shaw

$$\boxed{F_{\mathrm{add}} = kT D[p_t \| p_{\mathrm{eq}}]}$$

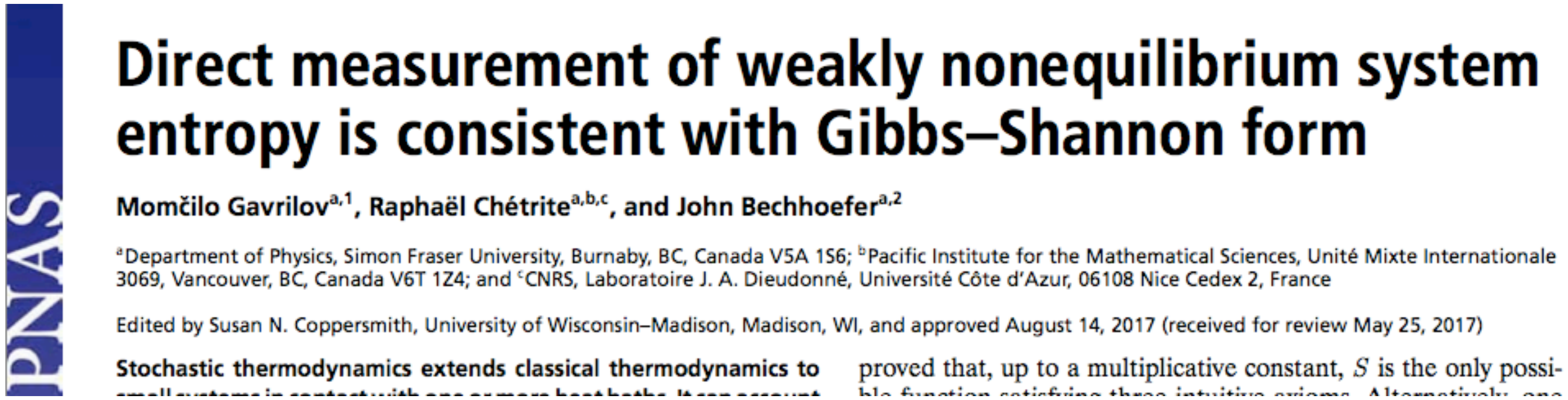
(e.g.: R. Shaw: The dripping faucet (1981), Takara, Hasegawa, Driebe, Phys. Lett. A(2010))

relative entropy:

$$D[p\|q] = \left\langle \ln\left[\frac{p}{q}\right]\right\rangle_p$$

- Then: Free energy = corresponding equilibrium free energy + additional free energy: $F = F_{\mathrm{eq}} + F_{\mathrm{add}}$

- Experimental verification (Bechhoefer Lab, 2017): With colloidal particle in laser trap


PNAS

Direct measurement of weakly nonequilibrium system entropy is consistent with Gibbs–Shannon form

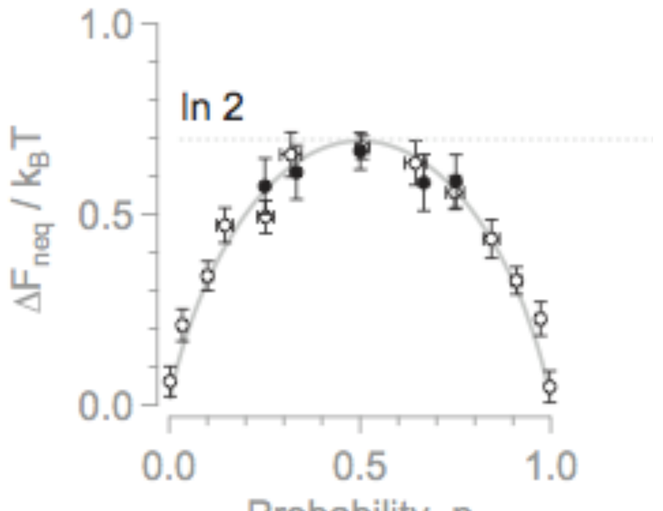
Momčilo Gavrilov[a,1], Raphaël Chétrite[a,b,c], and John Bechhoefer[a,2]

[a]Department of Physics, Simon Fraser University, Burnaby, BC, Canada V5A 1S6; [b]Pacific Institute for the Mathematical Sciences, Unité Mixte Internationale 3069, Vancouver, BC, Canada V6T 1Z4; and [c]CNRS, Laboratoire J. A. Dieudonné, Université Côte d'Azur, 06108 Nice Cedex 2, France

Edited by Susan N. Coppersmith, University of Wisconsin–Madison, Madison, WI, and approved August 14, 2017 (received for review May 25, 2017)

Stochastic thermodynamics extends classical thermodynamics to

proved that, up to a multiplicative constant, S is the only possi-



Partially observable generalized information engines

- System state (z) can be decomposed in two ways:
  - observables (x) and everything else ($\bar{x}$) OR
  - controlables (y) and everything else ($\bar{y}$)
- Memory (m): constructed from x; used to infer y.
- Free energy = Average energy - kT Entropy (Shannon entropy)
- Need to look at free energy change ΔF of the joint system-observer engine state (the random variables z and m):

$$\Delta F = \Delta E - kT\Delta H \underset{\uparrow}{=} W + Q - kT\Delta H$$

$$\boxed{\text{First law: } \Delta E = W + Q}$$

Partially observable generalized information engines

- Joint free energy change: $\Delta F = W + Q - kT\Delta H$
- Second law: $W - \Delta F = -Q + kT\Delta H \geq 0$
- Memory making step: $-Q_M \geq -kT\Delta H_M$

entropy decrease compensated by heat dissipation

- Work extraction step: $Q_E \leq kT'\Delta H_E$

absorbed heat compensated by entropy increase

- Now we need to compute the entropy changes:
- Notation: Entropy $H[p(z)] = -\left\langle \ln\left[p(z)\right]\right\rangle_{p(z)} \equiv H[Z]$
  Joint entropy $H[p(m,z)] = -\left\langle \ln\left[p(m,z)\right]\right\rangle_{p(m,z)} \equiv H[M,Z]$
  Conditional entropy $H[p(z|m)] = -\left\langle \ln\left[p(z|m)\right]\right\rangle_{p(m,z)} \equiv H[Z|M]$

$$H[p(m,z)] = H[(p(z|m)p(m)] = H[p(z|m)] + H[p(m)]$$

$$\Leftrightarrow \boxed{H[M,Z] = H[Z|M] + H[M]}$$

- Memory making step
  joint distribution BEFORE: $p(m)p(z)$
  (system and memory uncorrelated)
  joint distribution AFTER: $p(m|x)p(z)$
  (memory constructed only from observables)
  Entropy change: $\Delta H_M = H[M|X] - H[M] = -I[M,X]$

Entropy decreases by amount of information captured in memory

$$\boxed{-Q_M \geq kTI[M,X]}$$

Dissipated heat no less than information captured

This is known (Landauer / Szillard)

- Work extraction step
  BEFORE: $p(\bar{y}|y,m)p(y|m)p(m)$
  AFTER: $p(\bar{y}|y,m)p(y)p(m)$
  (no correlations left to exploit)
  Entropy change: $\Delta H_E = H[Y] - H[Y|M] = I[M,Y]$

Entropy increases by amount of inferred information utilized

$$\boxed{Q_E \leq kT'I[M,Y]}$$

Absorbed heat no more than **relevant** information

Lower bound on dissipation

- Memory making step $-Q_M \geq kTI[M,X] = kTI_{\rm mem}$ (Landauer)

- Work extraction step $Q_E \leq kT'I[M,Y] = kT'I_{\rm rel}$ (new)

- Dissipation $$\boxed{-Q \geq k\left(TI_{\rm mem} - T'I_{\rm rel}\right)}$$

(Still PRL, 2020)

- Isothermal engine (T' = T):

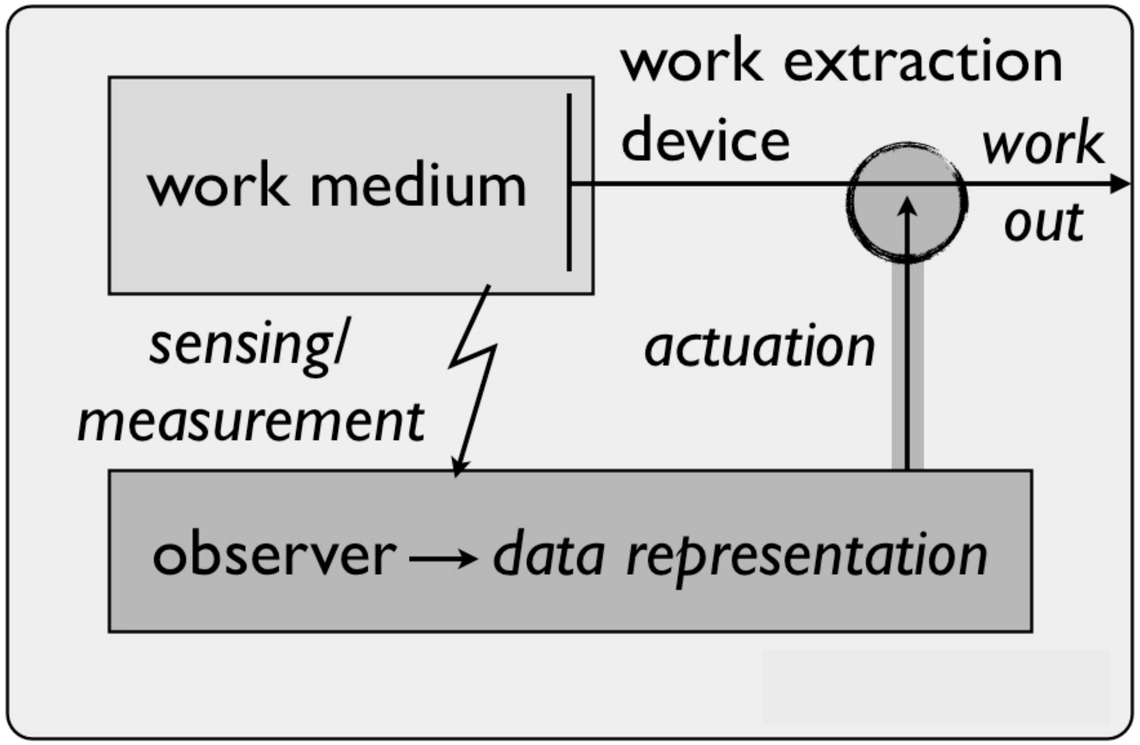
$$-Q \geq kTI_{\rm irrel}$$

$$I_{\rm irrel} = I_{\rm mem} - I_{\rm rel}$$

Dissipation is controlled by how much **irrelevant** information is captured!

From information engines to strategies for data representation and to machine learning

- Information engine:
  - acquire and process information (using energy)
  - use information to extract work



- Simplest model for a ``proto-agent".

Postulate for observers

- Choose a data representation strategy (in the form of a mapping from data to memory) that would allow for minimal wasted energy.
- The energy that is actually dissipated depends on the specific implementation and the environmental context.
- Minimize the bound on dissipation!
- (Do not minimize actual dissipation at all times, just make it possible that energy efficiency could be high whenever needed.)

Data representation strategy from minimizing lowest achievable dissipation

- General bound on dissipation

$$\boxed{-Q \geq k\,(T I_{\mathrm{mem}} - T' I_{\mathrm{rel}})}$$

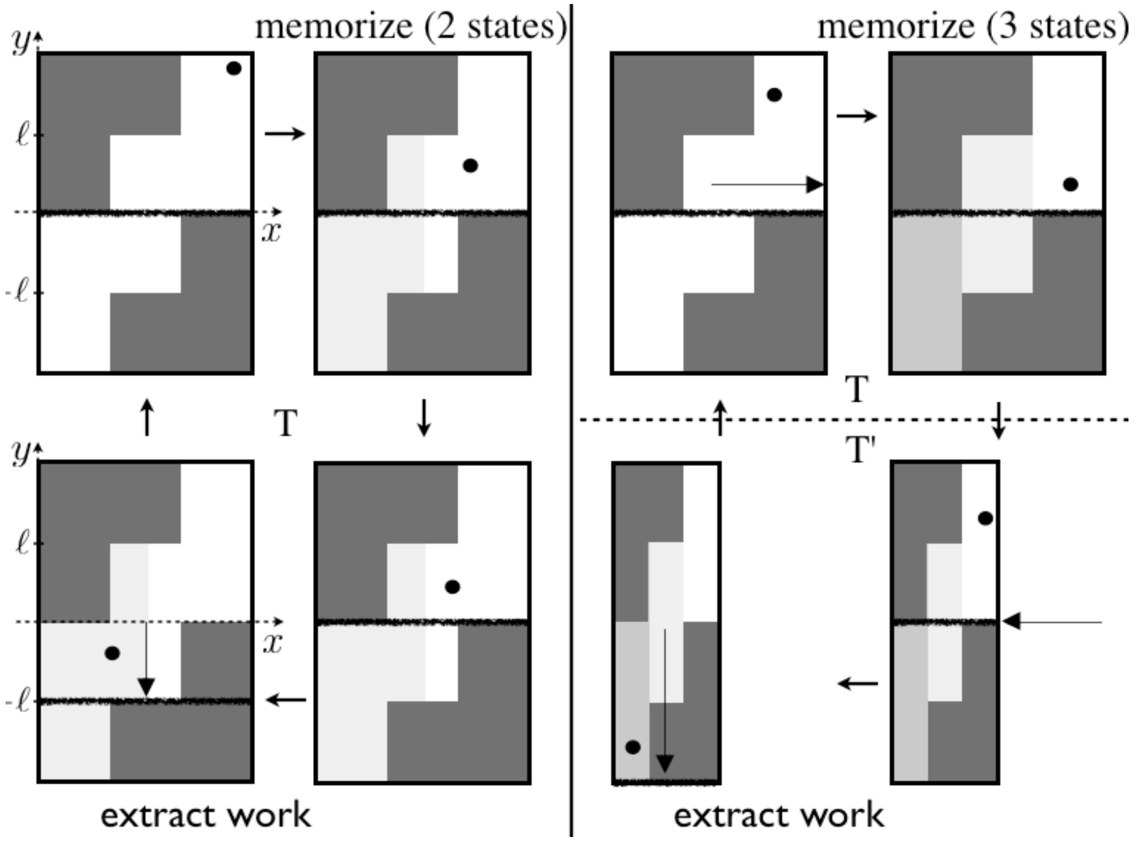
- minimization (over all possible stochastic maps from data to memory) is the same as (subject to normalization of p(m|x))

$$\min_{p(m|x)} \left(I[M,X] - \alpha I[M,Y]\right) \qquad \text{with} \qquad \alpha = \frac{T'}{T}$$

- => "Information Bottleneck" method (Tishby, Pereira, Bialek,1999)
- Solutions must satisfy $p(m|x) = \dfrac{p(m)}{Z(x,\alpha)} e^{-\alpha D[p(y|x) \| p(y|m)]}$

Efficiency

- Work in = dissipated heat = kT I_{mem}
- Work out = absorbed heat = kT’ I_{rel}
- Efficiency: $\dfrac{\textbf{total work out}}{\textbf{heat in at T'}}$



$$\eta = 1 - \frac{T}{T'}\frac{I_{\text{mem}}}{I_{\text{rel}}} = \eta_C - \frac{T}{T'}\frac{I_{\text{irrel}}}{I_{\text{rel}}}$$

Carnot efficiency is reduced in proportion to ratio of irrelevant to relevant information

- Therefore: Information Bottleneck method provides a strategy to make achieving maximum efficiency possible for an observer.

Generalized Information Bottleneck Framework

- Dynamical and interactive learning. S. Still (2009) EPL
  - Contains interesting special cases, for example Crutchfield's ``computational mechanics'' S. Still, J. P. Crutchfield and C. J. Ellison (2010) CHAOS; S. Still (2014) Entropy
  - Can be applied to reinforcement learning. S. Still and D. Precup (2012) Theory in Biosciences
    - Optimal behavior strategies emerge that ***balance*** control (exploitation) with exploration.
- Quantum generalization of IB A. Grimsmo and S. Still (2016) Phys. Rev. A
- Learning from finite data S. Still and W. Bialek (2004) Neural Comp.

Boltzmann machine

- Input patterns drive this neural network out of its parameter dependent equilibrium state, p, to a non-equilibrium state, q.
- The associated additional free energy, $D[q\|p]$ is dissipated during the relaxation process involved in predicting labels on new patterns.
- Those parameters are found that minimize $D[p\|q]$, thereby minimizing a lower bound on the average dissipation encountered during prediction.

Core ingredients of information theory

(S. Still, WITMSE, 2014)

- Shannon’s **rate** distortion curve directly follows from (Zipf’s) ``principle of minimum effort”, for cases where a distortion function, d, (equiv. utility) *is given.*

- A minimum effort coding strategy is then achieved by precisely

$$\min_{p(m|x)} \left(I[m,x] - \lambda \left\langle d(m,x) \right\rangle \right)$$

- Shannon’s **channel capacity** is the maximum work potential that can be achieved with a given channel.

Conclusions

- Making a data representation that minimizes the smallest possible dissipation that a partially observable information engine can achieve requires the use of predictive inference in the following sense:
- Keep only that part of the available information which is relevant to the task at hand.
- This is concretely solved by the Information Bottleneck method, which can be derived from the following (mild) postulate:
- An observer has no reason to represent data in a way that forces more dissipation than absolutely necessary.

Outlook

- Can we use the same reasoning to find other general learning strategies?
- General idea: Observers use those rules that **allow** them to come as close as possible to physical limits on information processing (whenever they actually need to).
- Other important limitations on learners to investigate: partial control, finite time operation.
- Other important physical limits on computation to explore: speed, accuracy, robustness.
- Extend treatment from average quantities to worst-case scenarios (e.g. single shot thermodynamics)

Thanks

- Funded by the **Foundational Questions Institute** (in part with the Fetzer Franklin Fund)
- Robert S. Shaw (movies; patient support and inspiration)
- Collaborators: W. Bialek, G. Crooks, A. Grimsmo, M. Marsili, D. Precup, D. Sivak, E. Stopnitzky, L. Altenberg, T. Bell, J. Crutchfield, T. Ouldridge