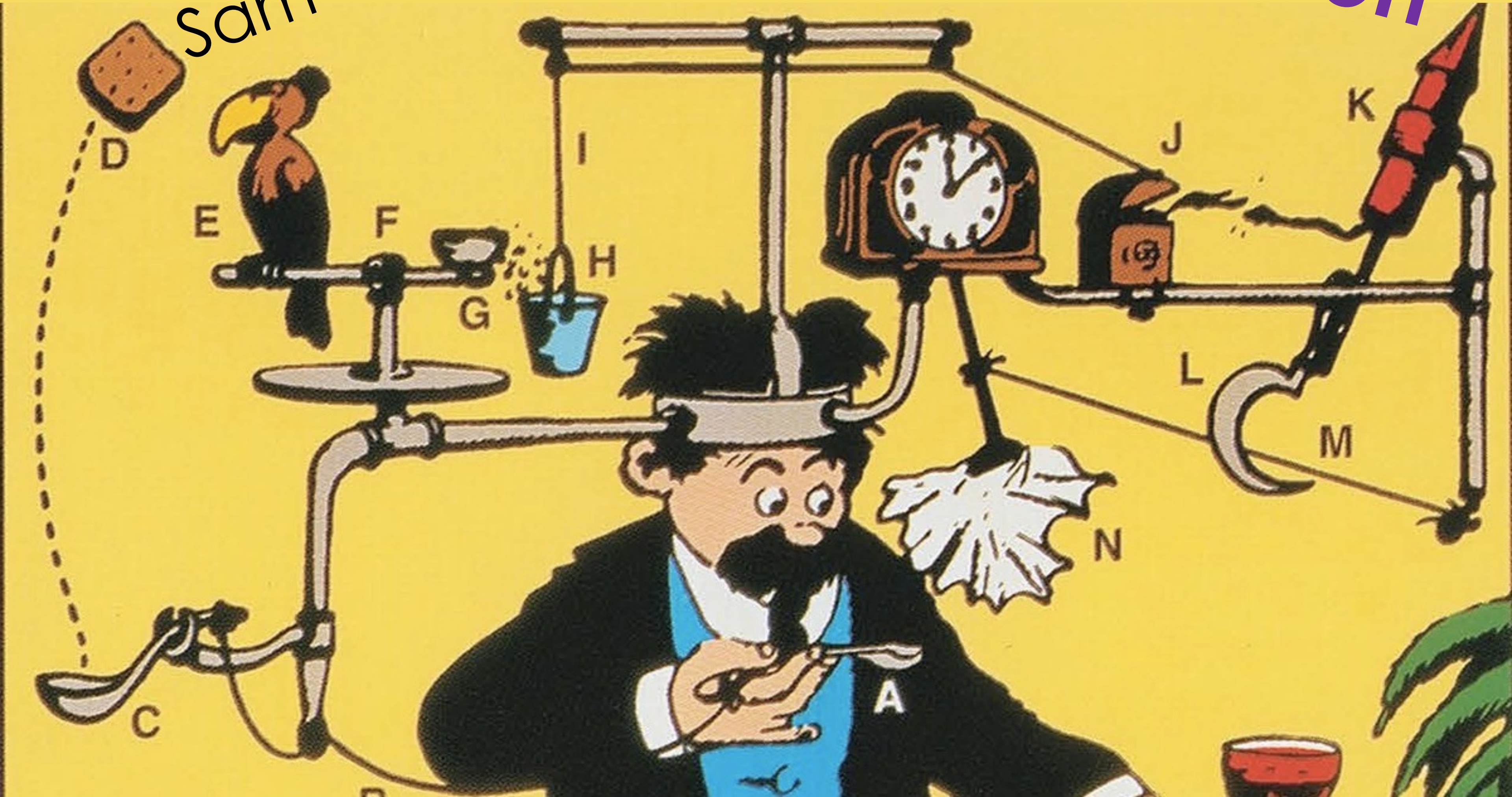
Policy Compression

Sam Gershman



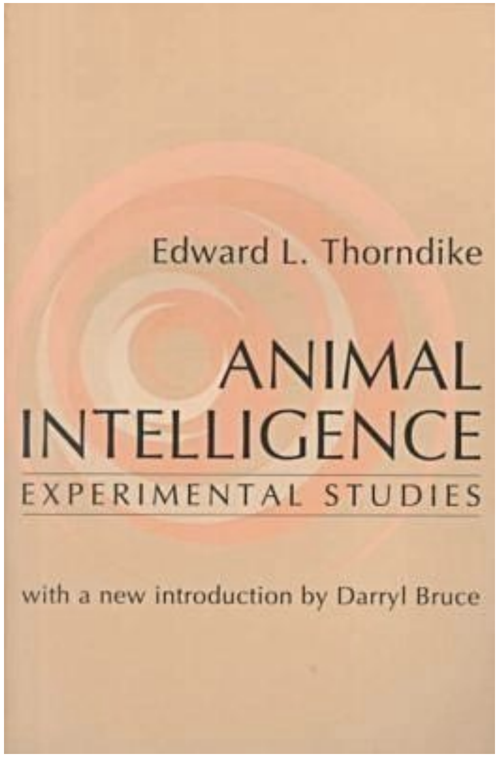
Acknowledgments

Lucy Lai

and big thanks to Anne Collins for sharing her data!

Two laws of learning

Edward Thorndike



Edward L. Thorndike

ANIMAL INTELLIGENCE

EXPERIMENTAL STUDIES

with a new introduction by Darryl Bruce

Two laws of learning

Edward Thorndike

Law of Effect: if you got rewarded for doing something, you will likely do it again.

Two laws of learning

Edward Thorndike

Law of Effect: if you got rewarded for doing something, you will likely do it again.

Law of Exercise: if you did something, you will likely do it again.

The puzzle of perseveration

- *Why does simply taking an action make it more likely that we will repeat in the future?*
- Reinforcement learning theory suggests that this is a bug. I will argue that it's a feature.

Standard normative analysis

$\pi(a|s)$ **policy**: probabilistic mapping from states to actions

V^π **value**: average reward under a policy

$\pi^* = \operatorname*{argmax}_\pi V^\pi$ optimal policy maximizes value

A new normative analysis

$\pi(a|s)$ **policy**: probabilistic mapping from states to actions

V^π **value**: average reward under a policy

$$\pi^* = \underset{\pi}{\operatorname{argmax}} V^\pi$$

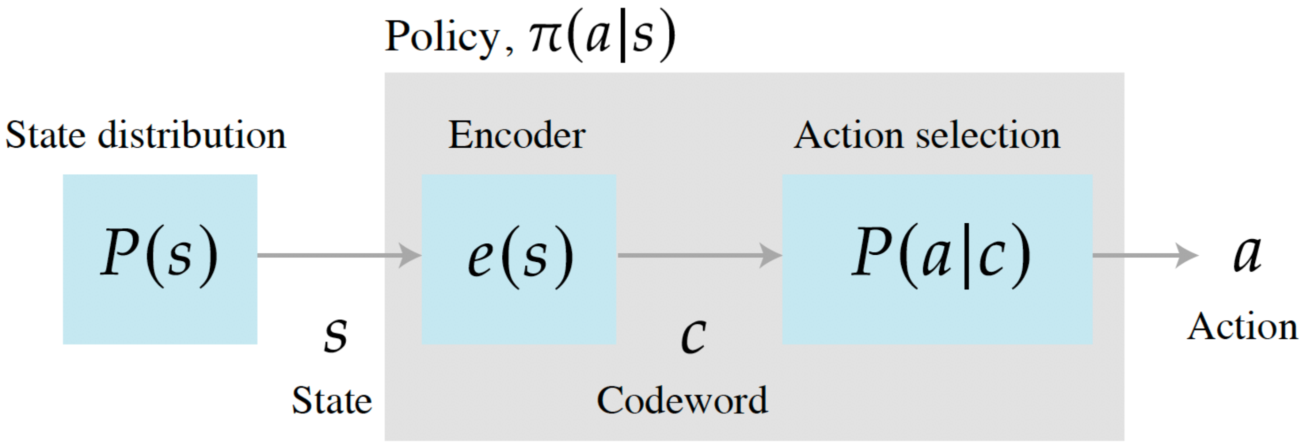
optimal policy maximizes value

$$\text{subject to } \underbrace{I^\pi(S;A)}_{\text{mutual information between states and actions}} = C$$

subject to a capacity constraint

= number of bits needed to encode the policy

Action selection as a capacity-limited channel



Solution: biased softmax policy

$$\pi^*(a|s) \propto \exp\left[\underbrace{\beta Q(s,a)}_{\text{expected reward}} + \overbrace{\log P^*(a)}^{\text{marginal action probability}}\right].$$

Solution: biased softmax policy

LAW OF EXERCISE

$$\pi^*(a|s) \propto \exp\left[\beta \underbrace{Q(s,a)}_{\text{expected reward}} + \overbrace{\log P^*(a)}^{\text{marginal action probability}}\right].$$

LAW OF EFFECT

Solution: biased softmax policy

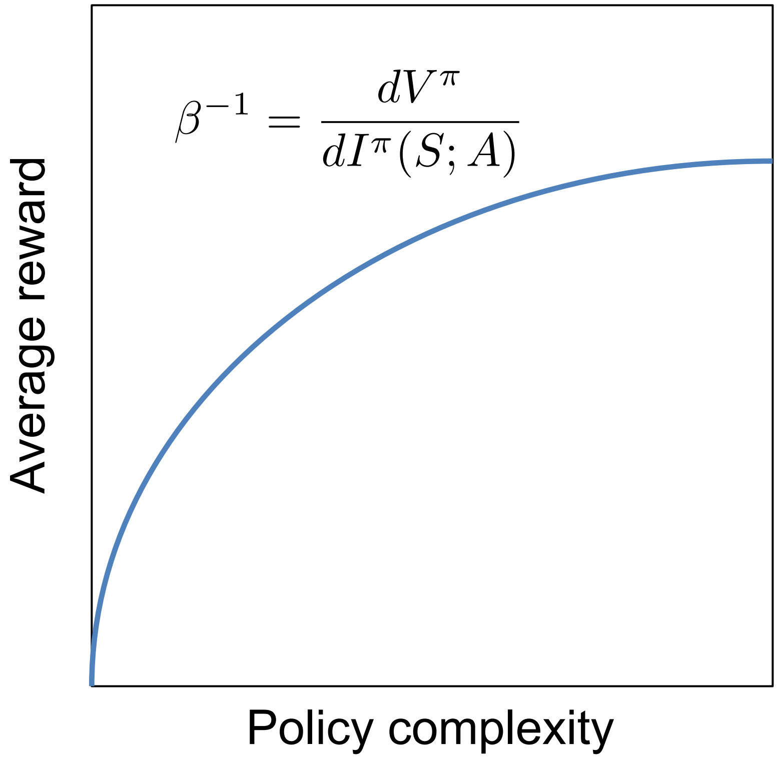
LAW OF EXERCISE

inverse temperature controls the balance

$$\pi^*(a|s) \propto \exp\left[\beta \underbrace{Q(s,a)}_{\text{expected reward}} + \overbrace{\log P^*(a)}^{\text{marginal action probability}}\right].$$

LAW OF EFFECT

The reward-complexity trade-off

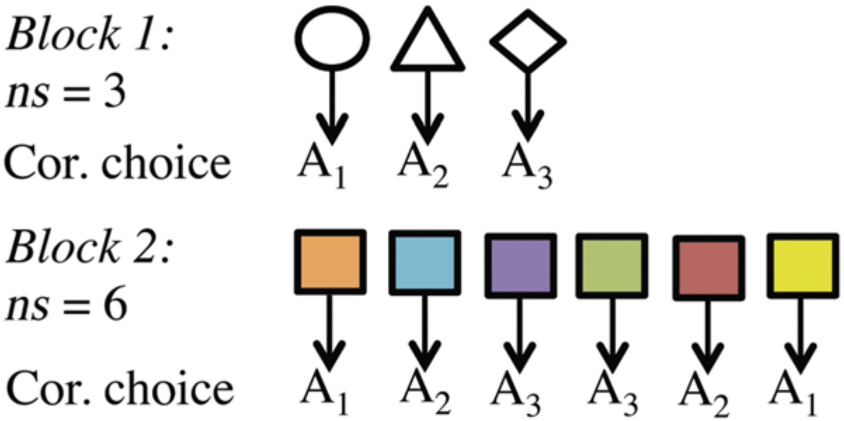


Reward-complexity curve is:

- Monotonic
- Concave

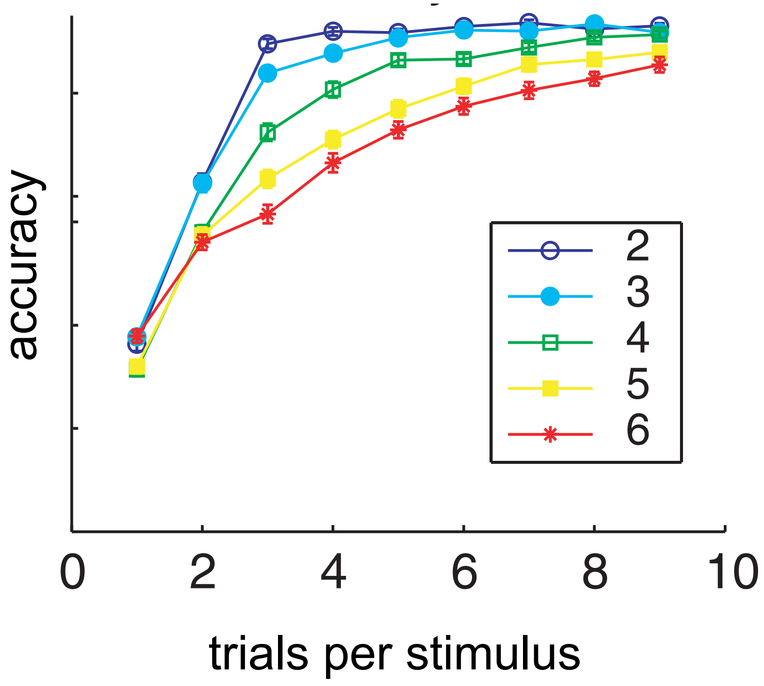
Slope corresponds to temperature (policy stochasticity), which decreases monotonically.

Evidence for capacity limit



Block 1:
ns = 3
Cor. choice A_1 A_2 A_3

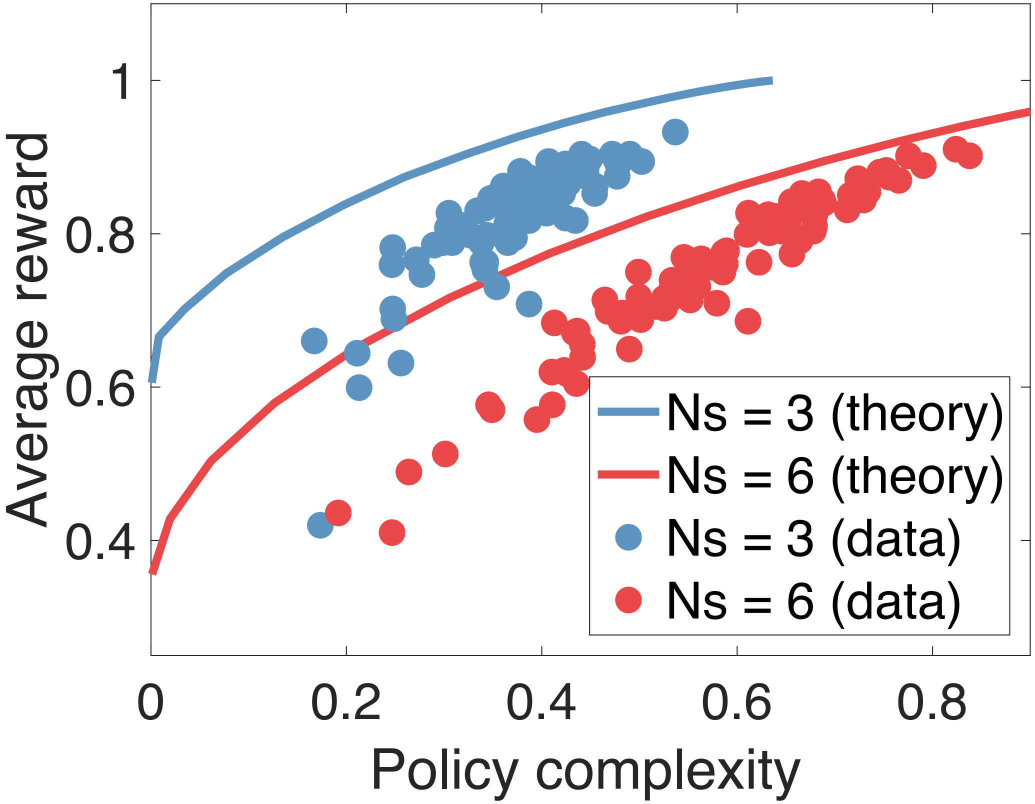
Block 2:
ns = 6
Cor. choice A_1 A_2 A_3 A_3 A_2 A_1



Accuracy in a contextual bandit task decreases with the number of stimuli (set size).

Collins & Frank (2012); Collins (2018)

Empirical trade-off



Human performance correlates strongly with the optimal curves, but falls systematically below them, particularly for individuals with low complexity policies.

Gershman (2020), *Cognition*

Quantitative test of the theory

overparametrized policy:

$$\pi(a|s) \propto \exp\left[\beta\hat{Q}(s,a) + \tau \log P(a)\right]$$

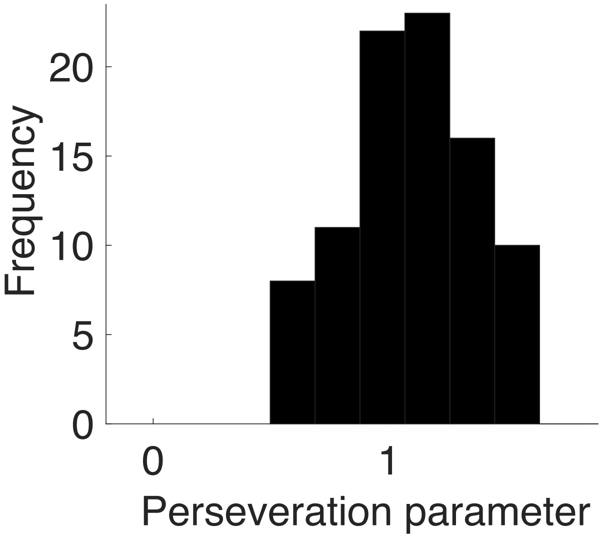
theory predicts this
should be close to 1

Quantitative test of the theory

overparametrized policy:

$$\pi(a|s) \propto \exp\left[\beta\hat{Q}(s,a) + \tau \log P(a)\right]$$

theory predicts this should be close to 1



Gershman (2020), *Cognition*

The trade-off in schizophrenia

- Studies have indicated that schizophrenia is associated with impaired exertion of cognitive and physical effort to earn reward.

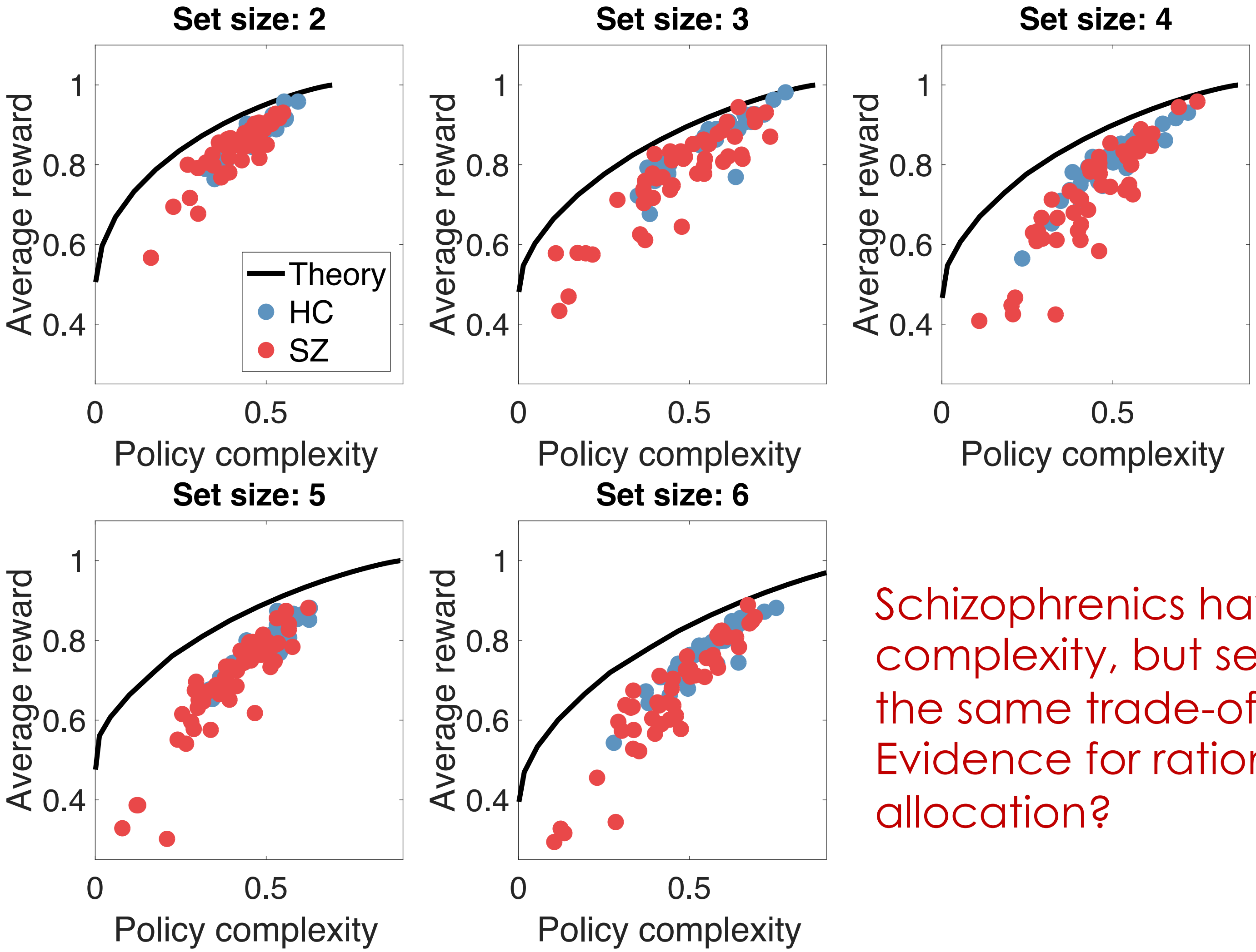
The trade-off in schizophrenia

- Studies have indicated that schizophrenia is associated with impaired exertion of cognitive and physical effort to earn reward.
- Clinically this can manifest as negative symptoms such as avolition (lack of motivation to do tasks) and anhedonia (lack of pleasure).

The trade-off in schizophrenia

- Studies have indicated that schizophrenia is associated with impaired exertion of cognitive and physical effort to earn reward.
- Clinically this can manifest as negative symptoms such as avolition (lack of motivation to do tasks) and anhedonia (lack of pleasure).
- Can we understand this dysfunction in terms of the reward-complexity trade-off?

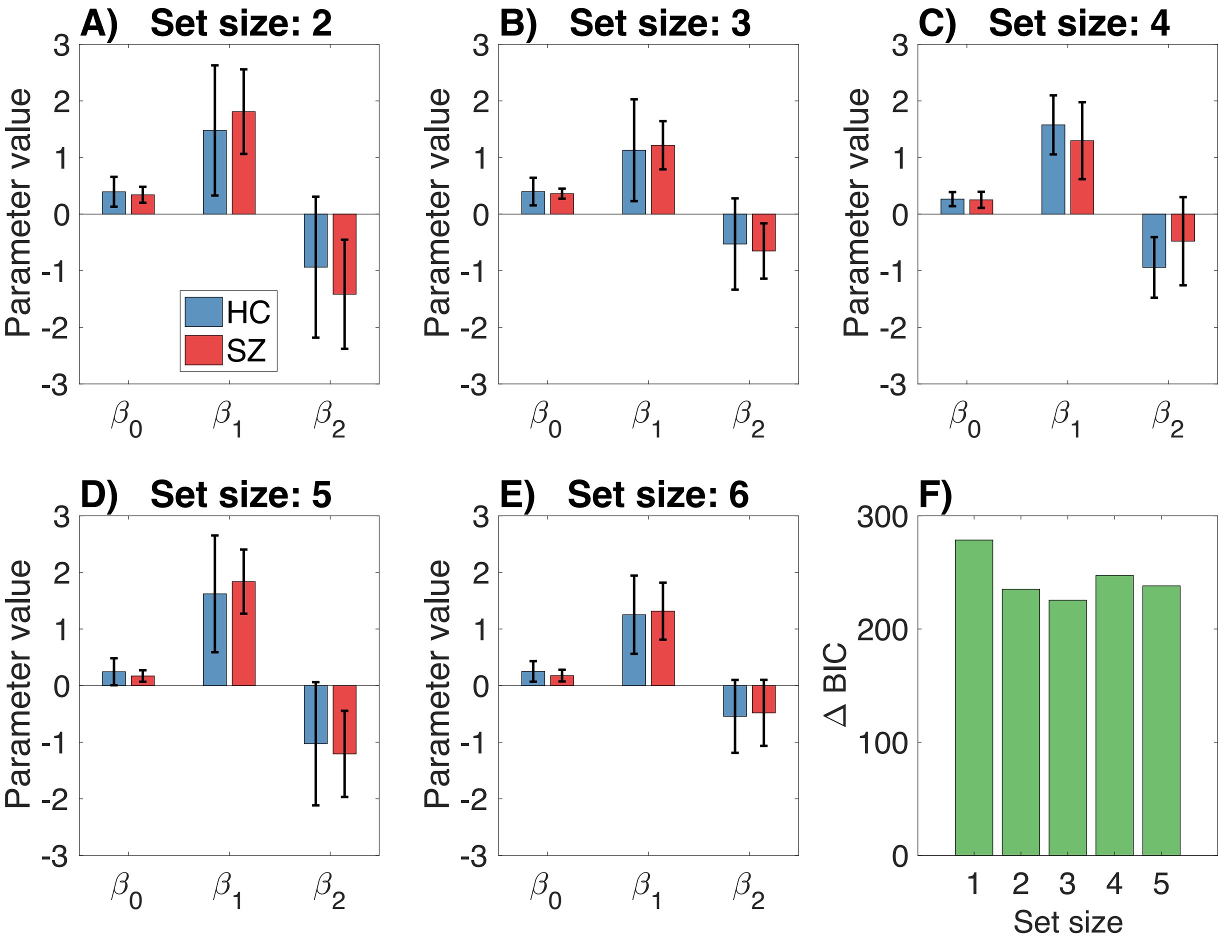
The trade-off in schizophrenia



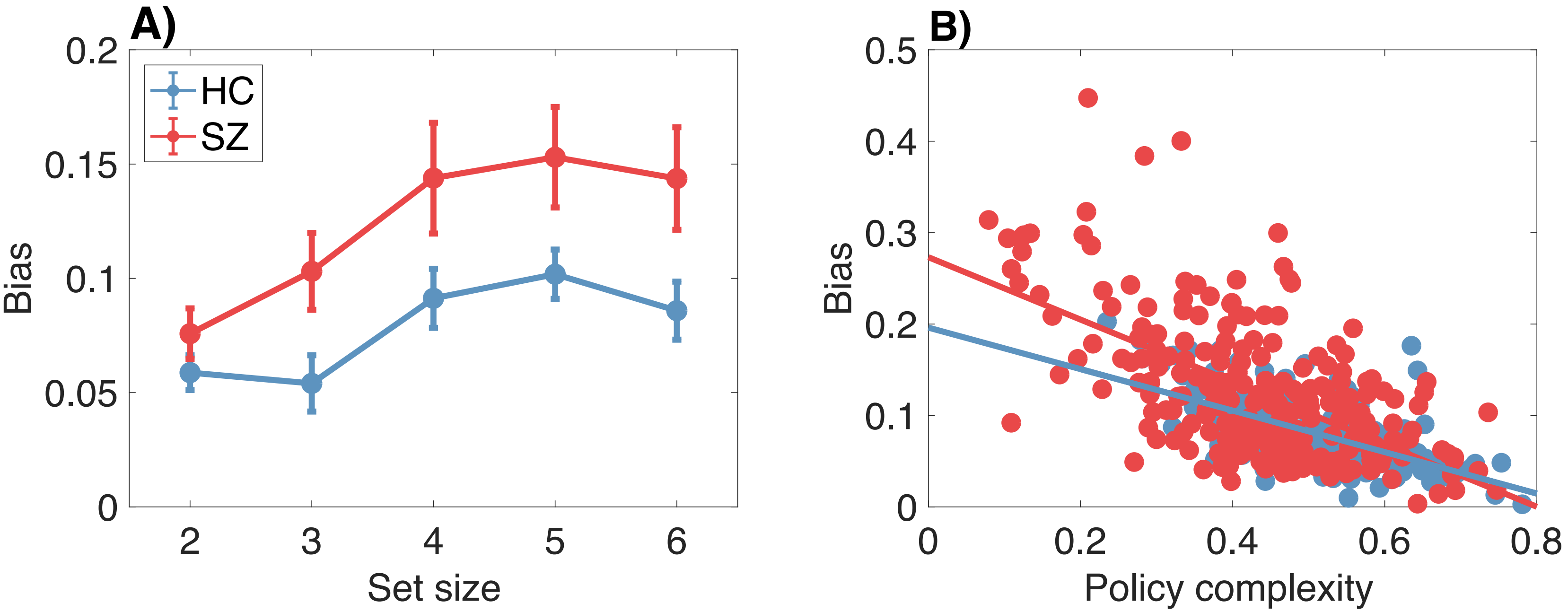
Schizophrenics have lower policy complexity, but seem to fall on the same trade-off function. Evidence for rational effort allocation?

Gershman & Lai (2021, *Computational Psychiatry*
data from Collins et al. (2014), *J Neuro*

Patients and controls lie on the same empirical trade-off curve



Quantifying bias



Bias increases with set size and decreases with policy complexity. The relationship with policy complexity is approximately the same in controls and patients.

Understanding bias

- We hypothesized that the bias might arise from suboptimal learning: people with excessively *high* learning rates will fall below the optimal trade-off curve.

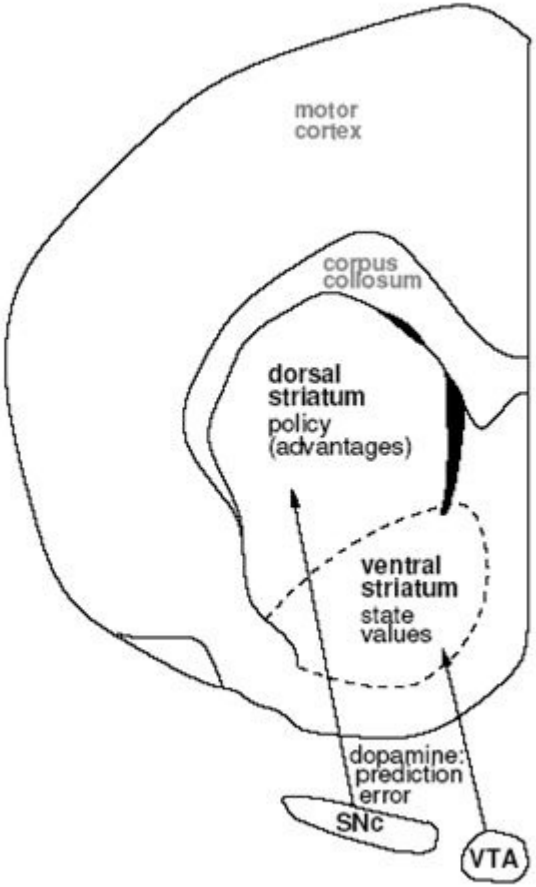
Understanding bias

- We hypothesized that the bias might arise from suboptimal learning: people with excessively *high* learning rates will fall below the optimal trade-off curve.
- We built a learning model that captures this idea, showing that it does a good job reproducing the reward-complexity curves.

Understanding bias

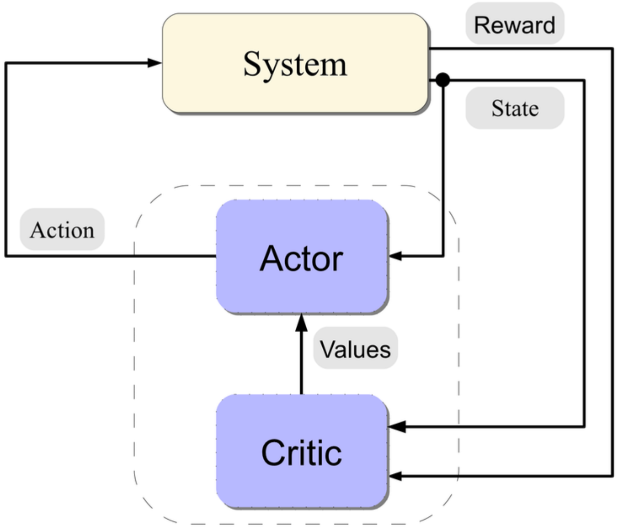
- We hypothesized that the bias might arise from suboptimal learning: people with excessively *high* learning rates will fall below the optimal trade-off curve.
- We built a learning model that captures this idea, showing that it does a good job reproducing the reward-complexity curves.
- People with schizophrenia had higher learning rates, as did subjects who showed a greater bias.

Actor-critic model

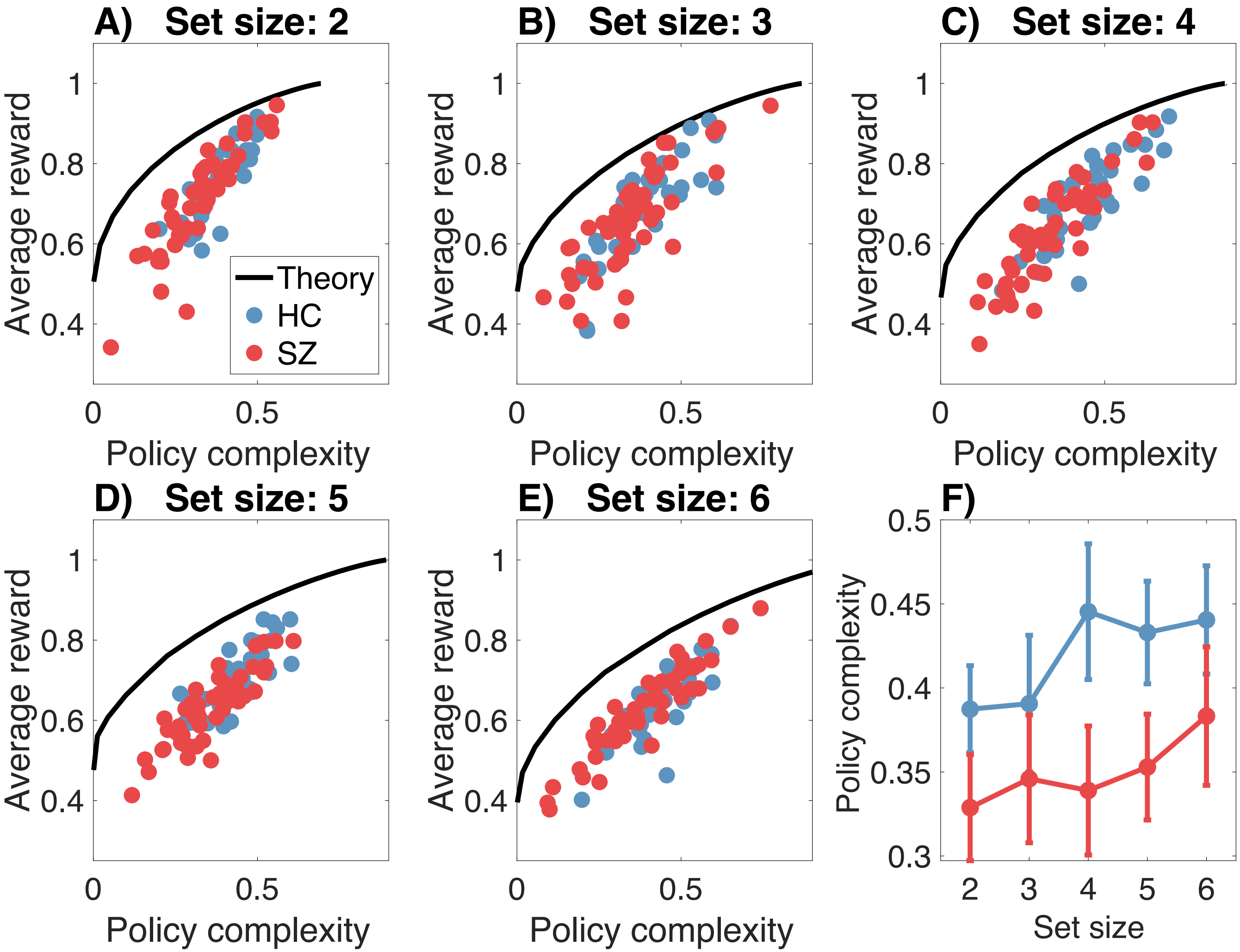


”Critic” (ventral striatum) learns state values.
“Actor” (dorsal striatum) learns the policy.

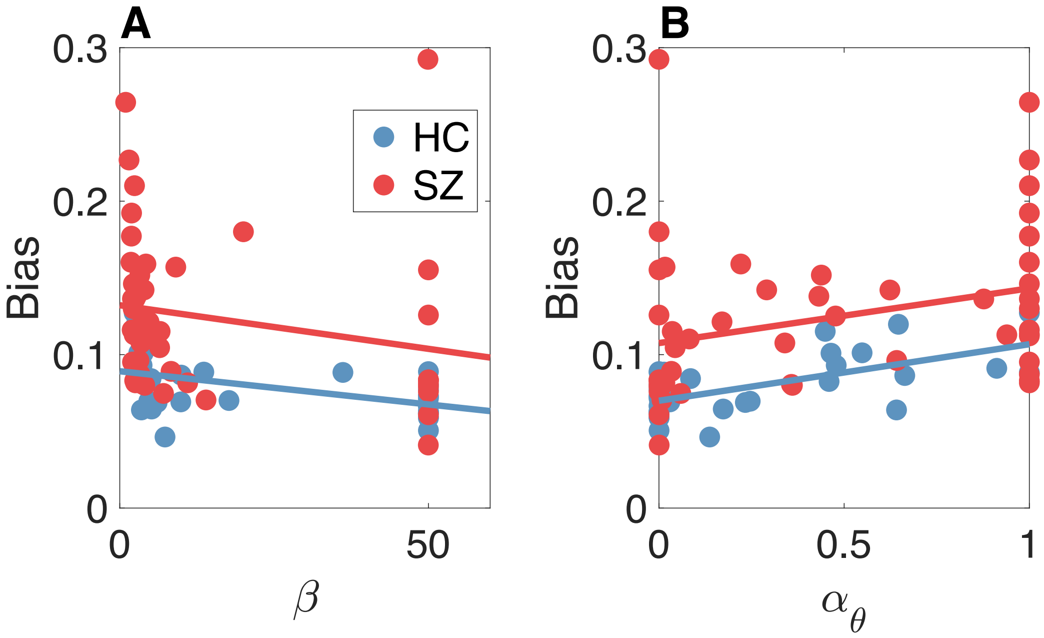
Importantly, prediction errors are penalized by policy complexity, driving the actor towards a low-complexity policy.



Simulated curves



Parameter estimates



Inverse temperature is lower (more stochasticity) and actor learning rate is higher for schizophrenic subjects.

Deviation from the optimal trade-off curve can be explained partly by suboptimal learning.

Suboptimality

- Suboptimality is more pronounced in the schizophrenic group, which had higher actor learning rates that in turn produced greater bias.
- This fits with the theoretical observation that convergence of actor-critic algorithms depends on the actor learning much more slowly than the critic (Konda and Tsitsiklis, 2000)
- An actor that learns too fast can produce suboptimal behavior.

Summary

- Human decision making obeys a reward-complexity trade-off.

Summary

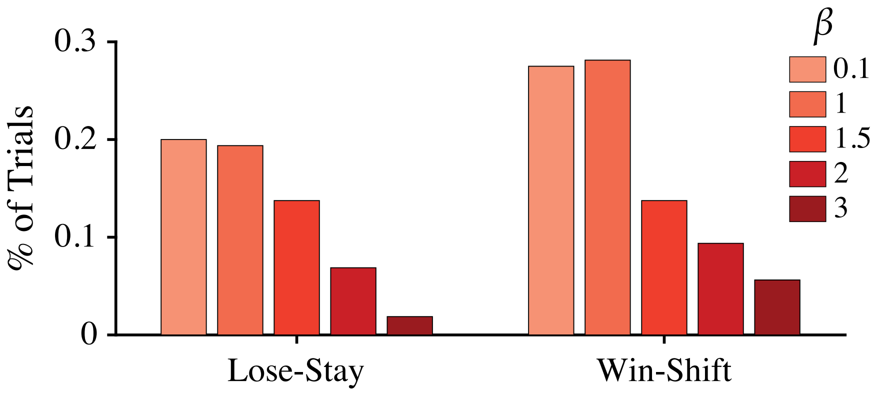
- Human decision making obeys a reward-complexity trade-off.
- Performance deviates systematically from the optimal trade-off curve, a phenomenon that can be explained by a learning model.

Summary

- Human decision making obeys a reward-complexity trade-off.
- Performance deviates systematically from the optimal trade-off curve, a phenomenon that can be explained by a learning model.
- Schizophrenic patients exhibit lower policy complexity and higher bias, possibly because of excessively high learning rates and low memory capacity.

Memory capacity and reversal learning

Schizophrenics exhibit higher rates of lose-stay and win-shift (Reddy et al., 2016). Our simulations show that this can arise from low capacity.



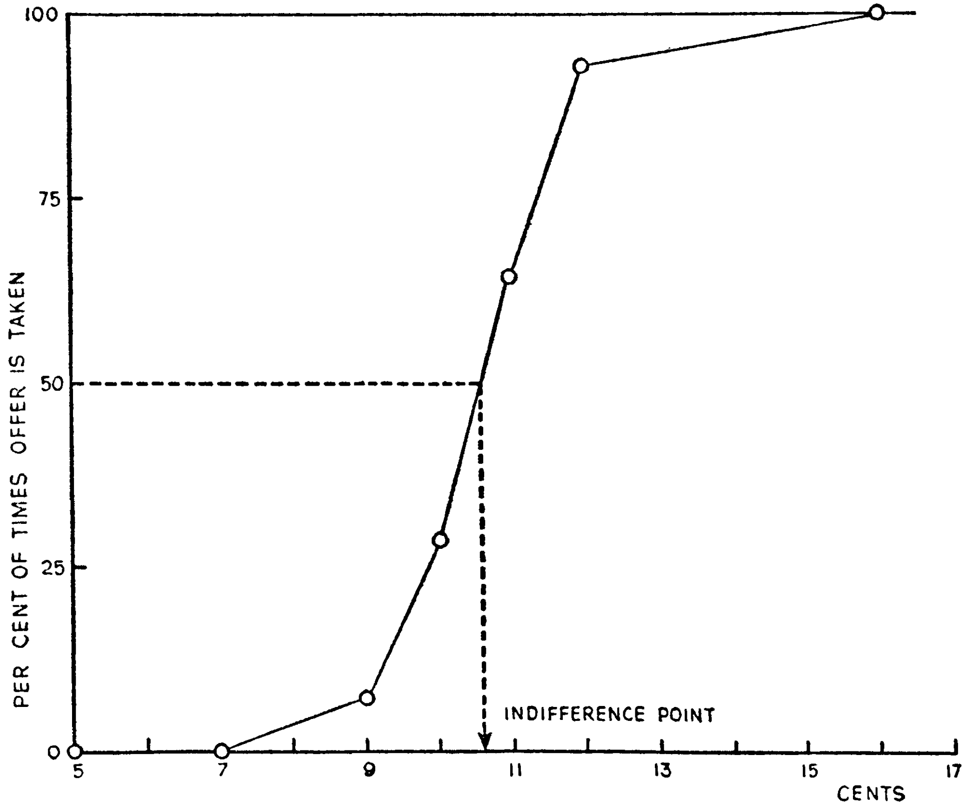
Lai & Gershman (2021), *Psychology of Learning and Motivation*

Beyond perseveration

- Policy compression implies several other interesting regularities.
  - Stochasticity
  - Relationship between response time and description length
  - Chunking

Stochasticity

Why are choices stochastic even when payoffs and probabilities are known?



Mosteller & Nogee (1951), *J Political Economy*

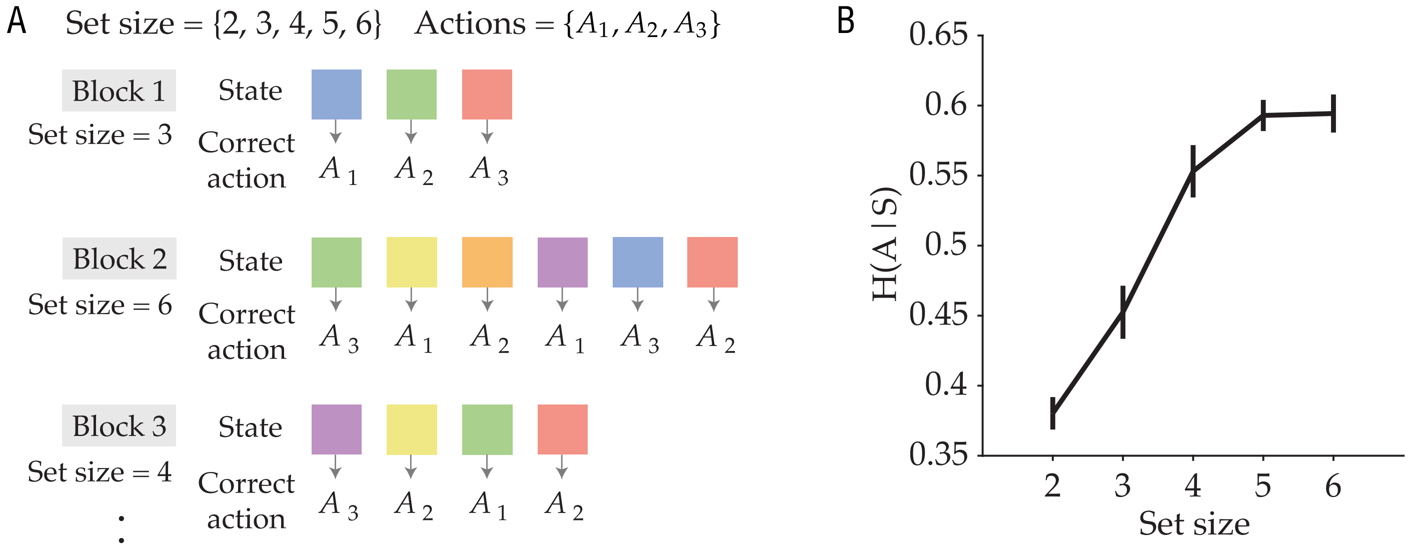
Stochasticity

$$\pi^*(a|s) \propto \exp\left[\beta Q(s,a) + \log P^*(a)\right].$$

Optimal policy under resource constraints is stochastic and equivalent to the softmax policy.

When the number of states is larger (higher cognitive load), stochasticity should be higher because the same resource pool is being shared across states.

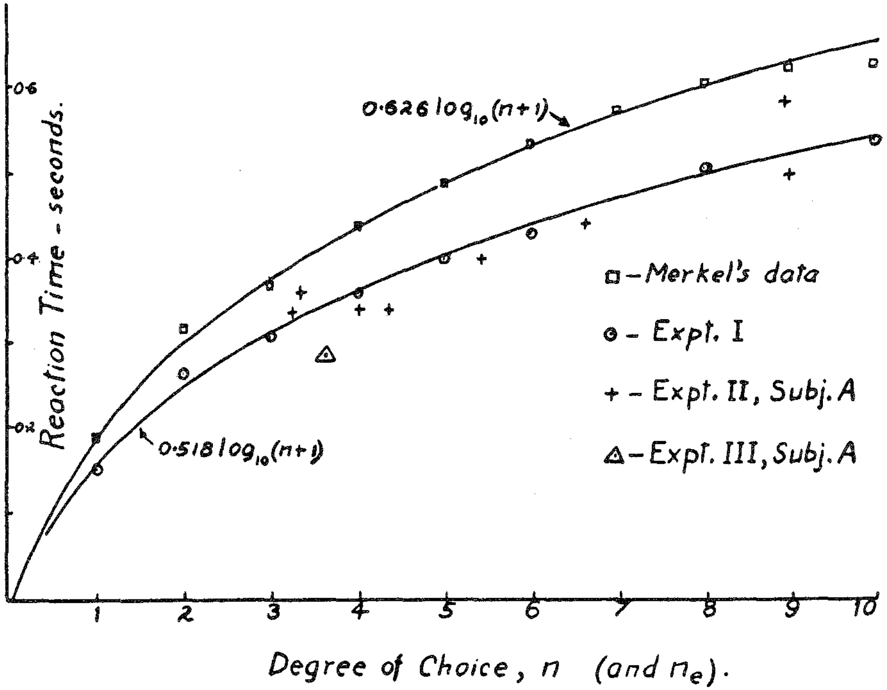
Stochasticity



Choice variability increases with set size.

Data from Collins et al. (2014), *J Neuro*
Lai & Gershman (2021), *Psychology of Learning and Motivation*

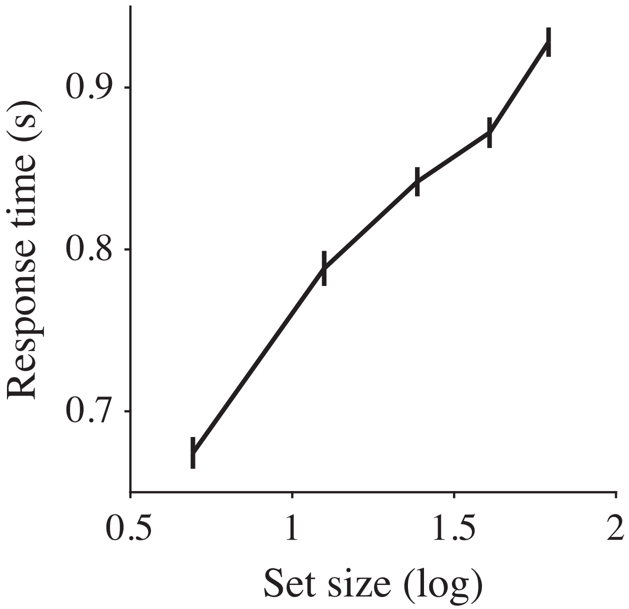
Response time



Hick's Law: response time increases logarithmically with the number of options

Hick (1952)

Response time



Hick's law in the Collins contextual bandit task.

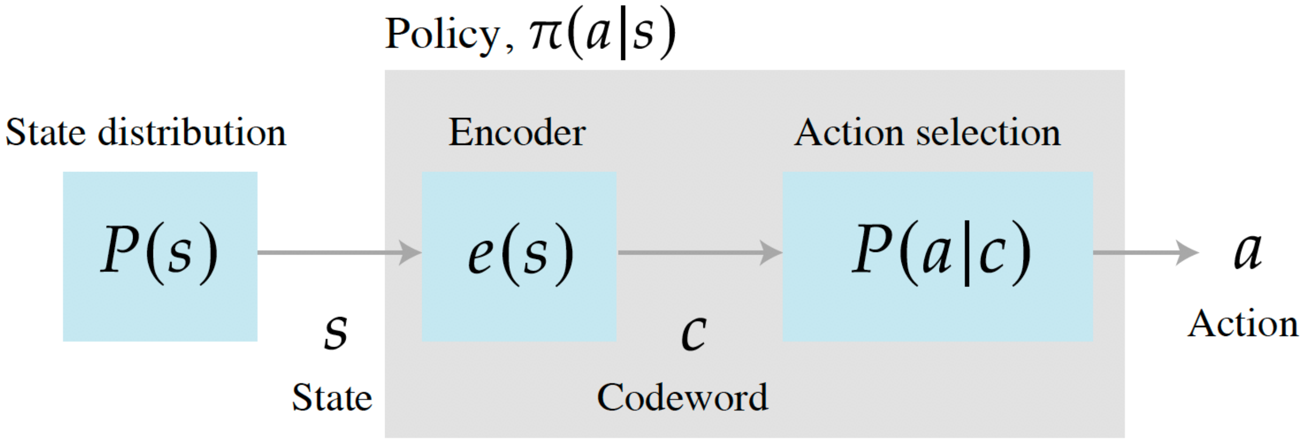
Data from Collins et al. (2014), *J Neuro*
Lai & Gershman (2021), *Psychology of Learning and Motivation*

Response time

Where does Hick’s law come from?

Response time

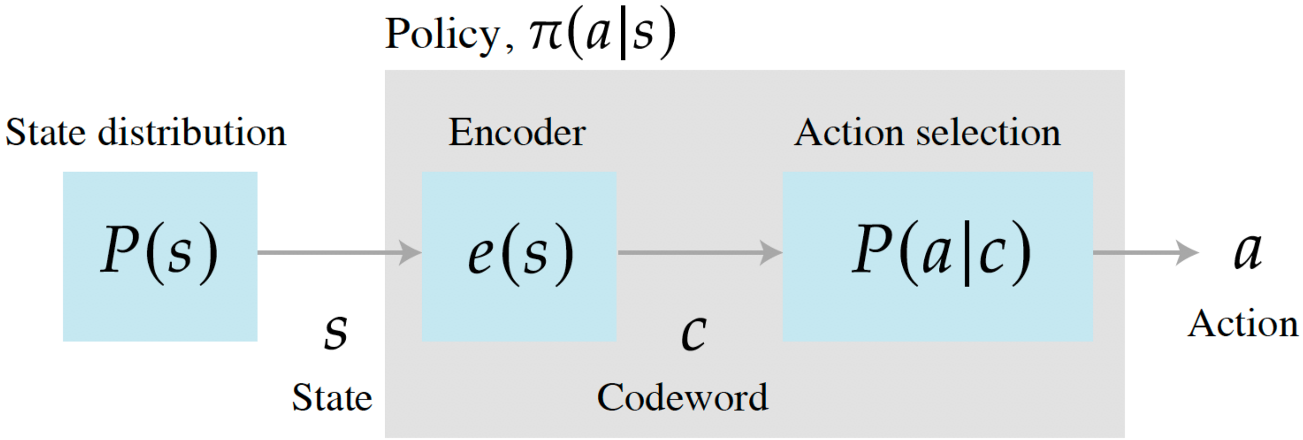
Where does Hick's law come from?



If each codeword is mapped deterministically to an action, the optimal encoder maps states to a binary codeword with length –log P(s).

Response time

Where does Hick's law come from?

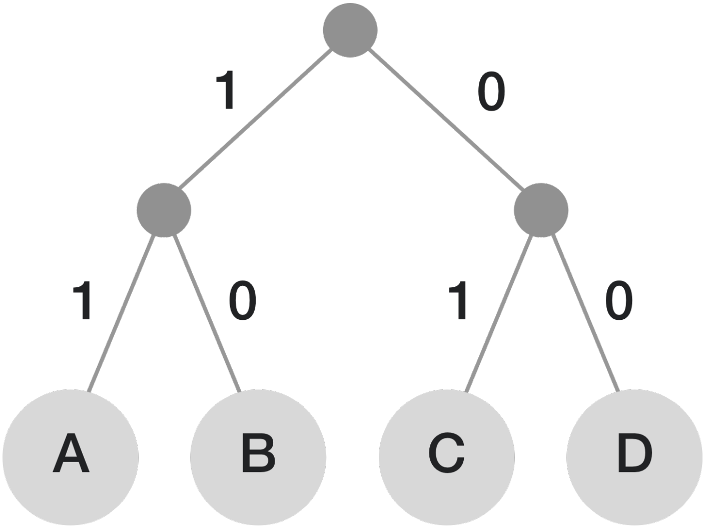


If each codeword is mapped deterministically to an action, the optimal encoder maps states to a binary codeword with length –log P(s).

If all states are equally probable, the expected code length is log N, where N is the # of states.

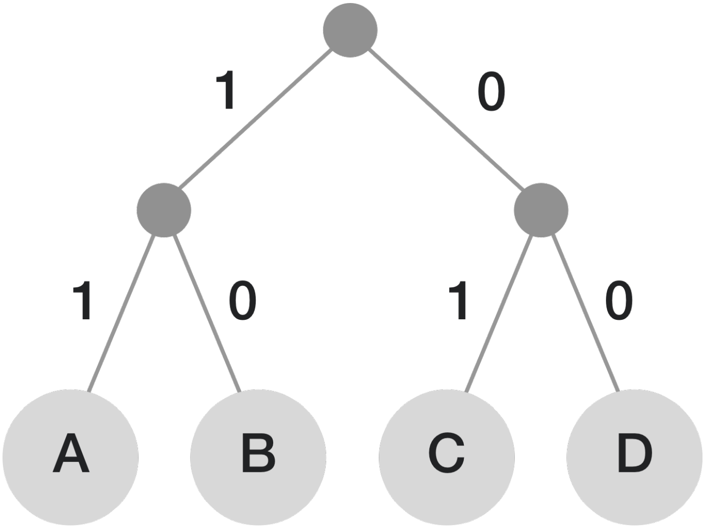
Response time

Huffman decoding: each codeword defines a sequence of yes/no questions that traverse a tree terminating in the input symbol.



Response time

Huffman decoding: each codeword defines a sequence of yes/no questions that traverse a tree terminating in the input symbol.



Response time = time it takes to traverse the tree = linear in code length

Response time

- If RT ~ code length, then responses should be slower for longer code lengths even when the set size is held fixed (see Hyman, 1953).

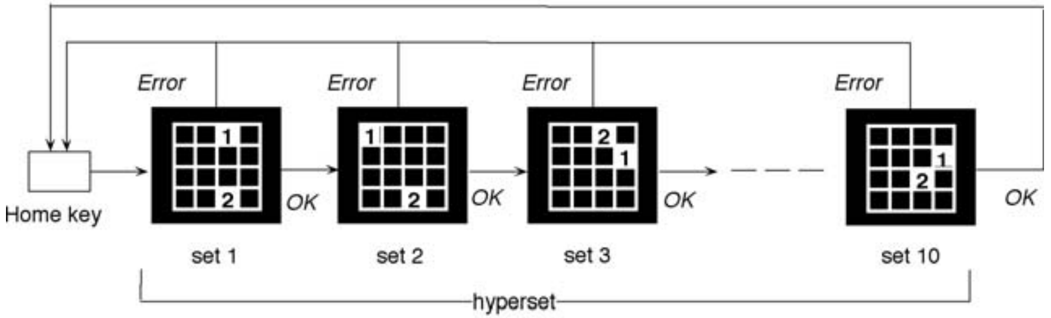
Response time

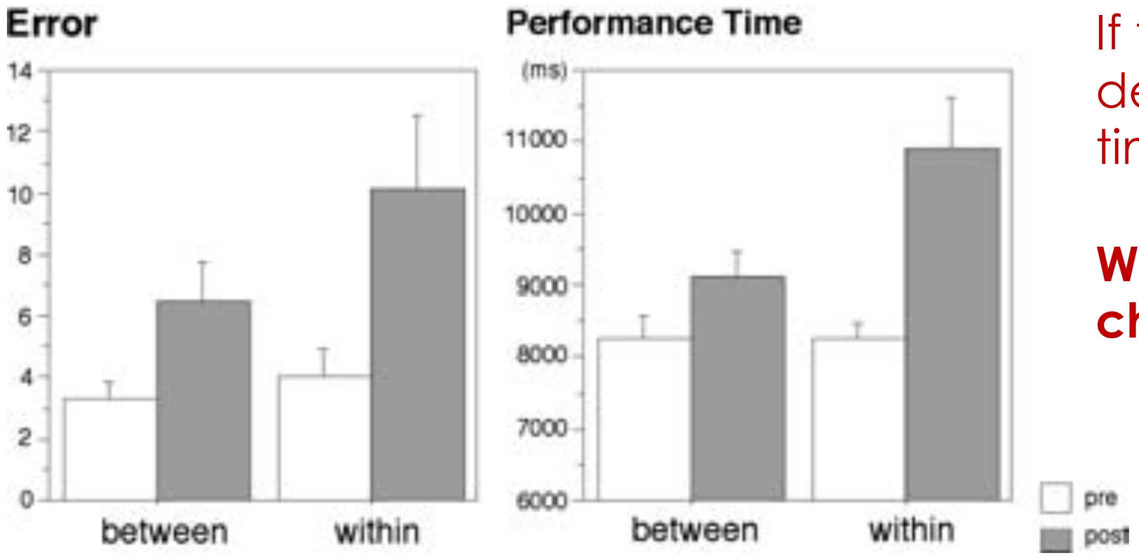
- If RT ~ code length, then responses should be slower for longer code lengths even when the set size is held fixed (see Hyman, 1953).
- Slope of the set size function decreases with practice (e.g., Hale, 1968), consistent with optimal compression arising from learning the probabilities.

Response time

- If RT ~ code length, then responses should be slower for longer code lengths even when the set size is held fixed (see Hyman, 1953).
- Slope of the set size function decreases with practice (e.g., Hale, 1968), consistent with optimal compression arising from learning the probabilities.
- Set size function flattens out for very large set sizes (e.g., Seibel, 1968), consistent with the existence of a capacity limit (hard cap on # of bits).

Action chunking





Repetition of action sequences causes actions to be spontaneously organized into chunks, making action selection faster.

If the sequences are shuffled to destroy the chunks, response time and error rate increase.

Why do we form action chunks?

Sakai et al. (2003)

Action chunking

- The complexity of an action policy depends on the degree to which it is state-dependent: we can think of policy complexity as quantifying the amount of memory that must be devoted to the state information when selecting actions.

Action chunking

- The complexity of an action policy depends on the degree to which it is state-dependent: we can think of policy complexity as quantifying the amount of memory that must be devoted to the state information when selecting actions.
- As capacity is reduced, the preference for multistep action chunks should increase, while the action execution time should decrease.

Conclusions

- Any resource-limited agent must compress their policies.

Conclusions

- Any resource-limited agent must compress their policies.
- This leaves measurable traces in patterns of perseveration, stochasticity, response time, and chunking.

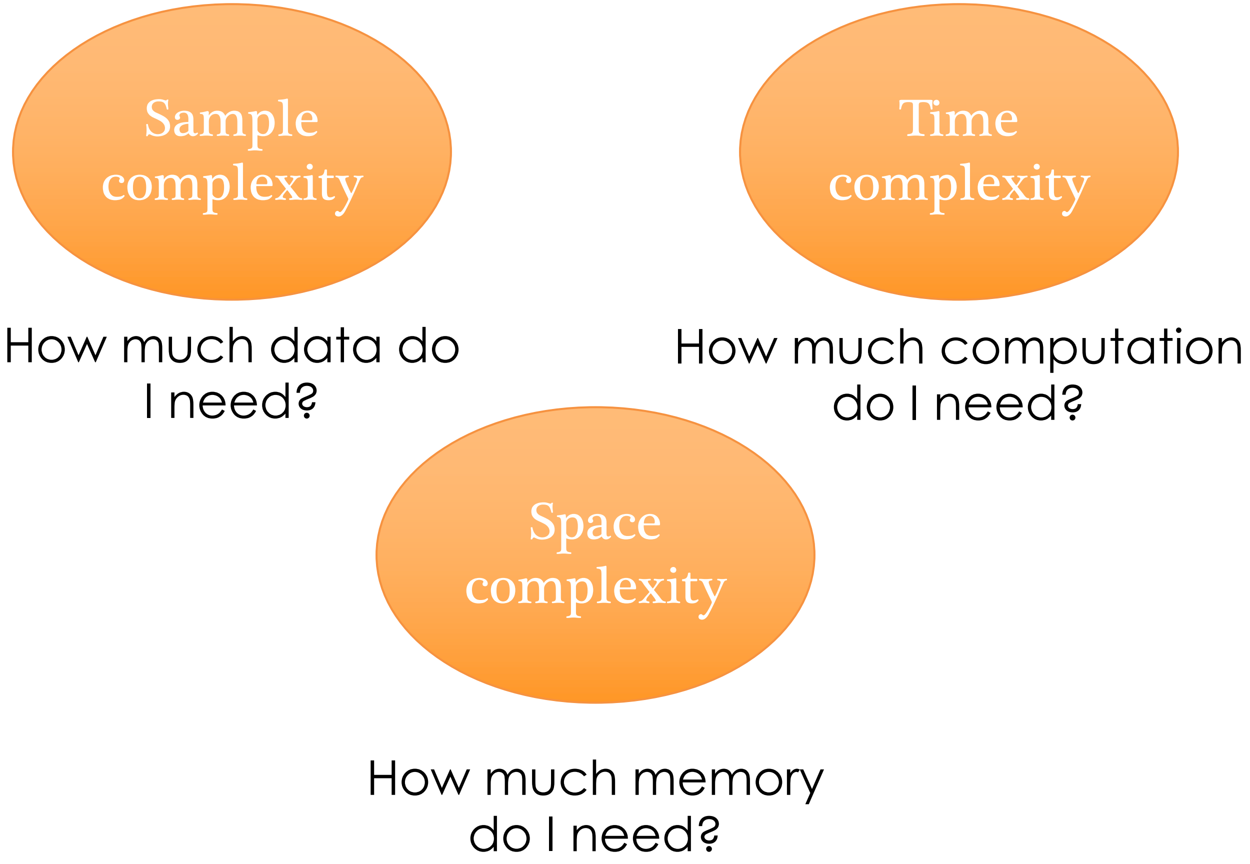
Conclusions

- Any resource-limited agent must compress their policies.
- This leaves measurable traces in patterns of perseveration, stochasticity, response time, and chunking.
- Policy compression offers a new way to formalize some aspects of mental illness.

Conclusions

- Any resource-limited agent must compress their policies.
- This leaves measurable traces in patterns of perseveration, stochasticity, response time, and chunking.
- Policy compression offers a new way to formalize some aspects of mental illness.
- By viewing optimality as a *frontier* rather than as a *point*, we can understand both individual heterogeneity and deviations from the optimal frontier in patients.

Linking notions of complexity



Sample complexity

How much data do I need?

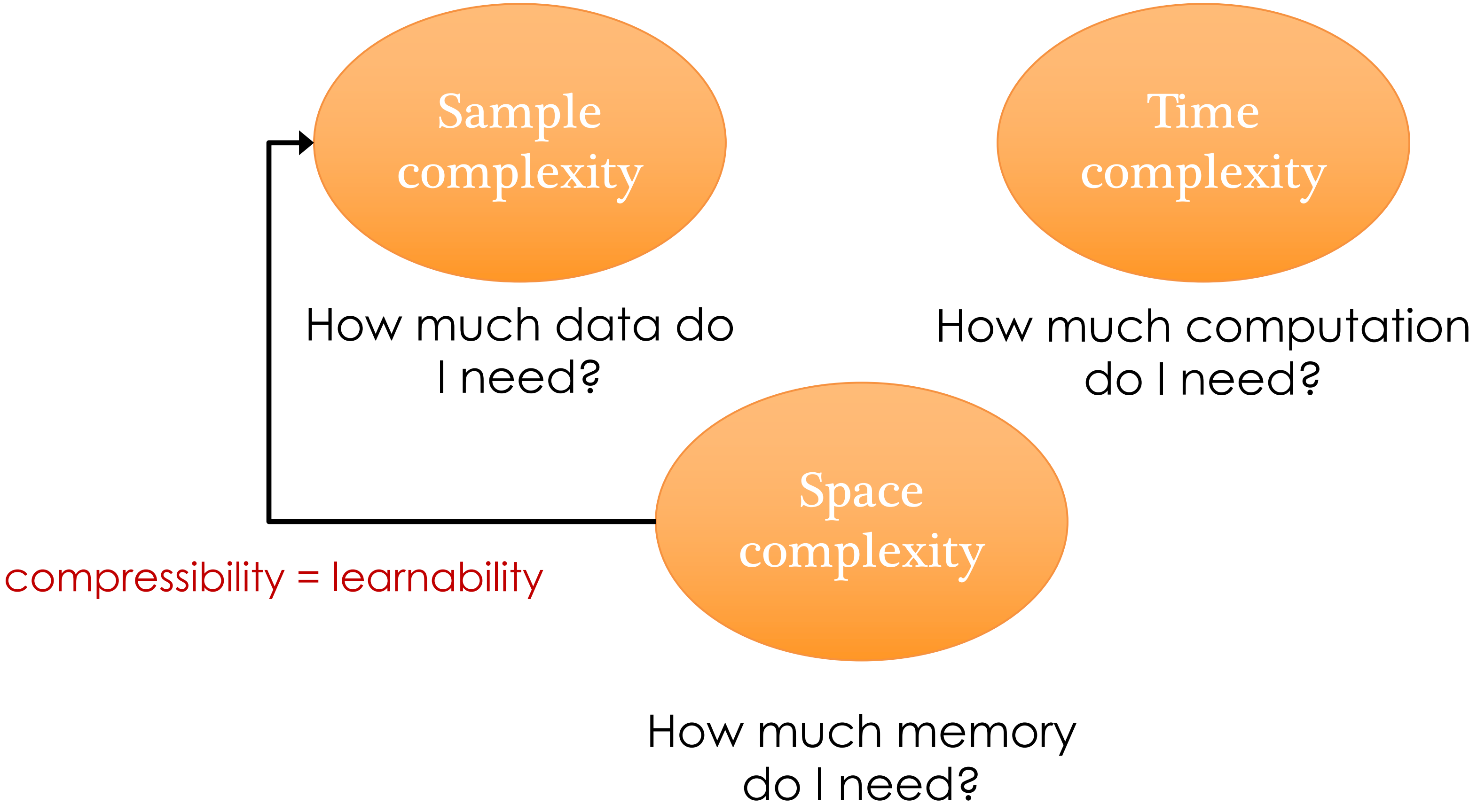
Time complexity

How much computation do I need?

Space complexity

How much memory do I need?

Linking notions of complexity



Sample complexity

How much data do I need?

Time complexity

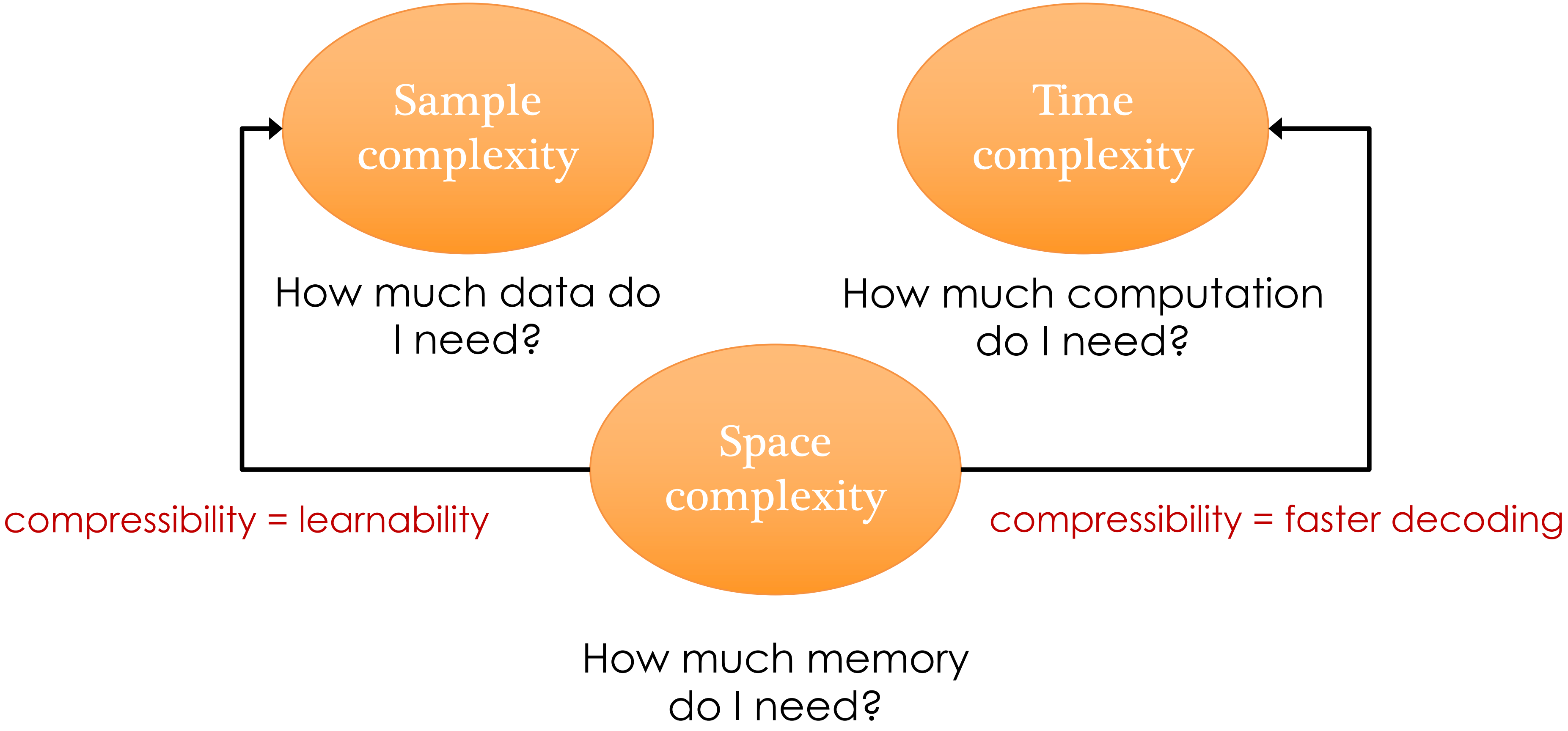
How much computation do I need?

Space complexity

How much memory do I need?

compressibility = learnability

Linking notions of complexity



Sample complexity

How much data do I need?

Time complexity

How much computation do I need?

Space complexity

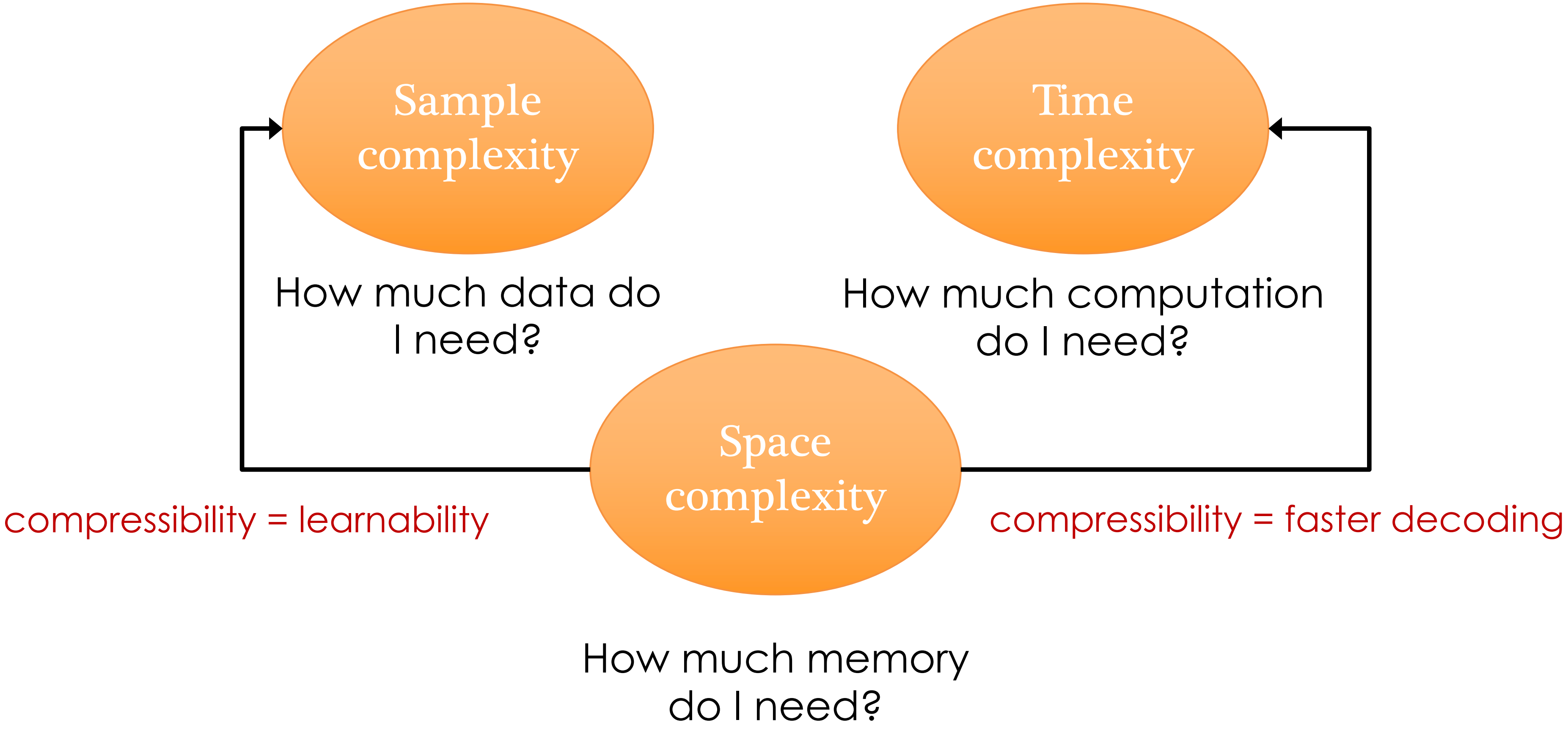
How much memory do I need?

compressibility = learnability

compressibility = faster decoding

Linking notions of complexity

These links suggest a transdiagnostic perspective on some clinical phenotypes



Sample complexity

How much data do I need?

Time complexity

How much computation do I need?

Space complexity

How much memory do I need?

compressibility = learnability

compressibility = faster decoding