

# A normative account of episodic memory

David G. Nagy, Balázs Török, Gergő Orbán



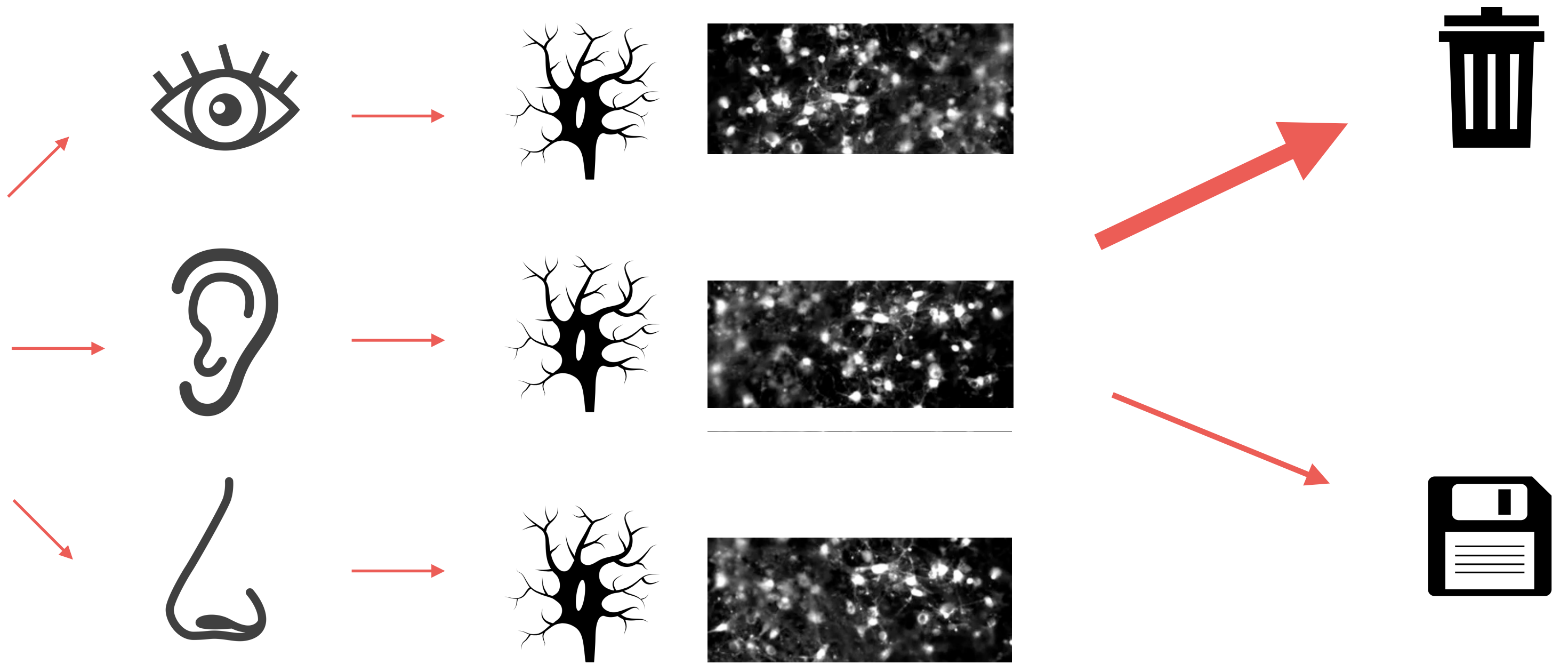
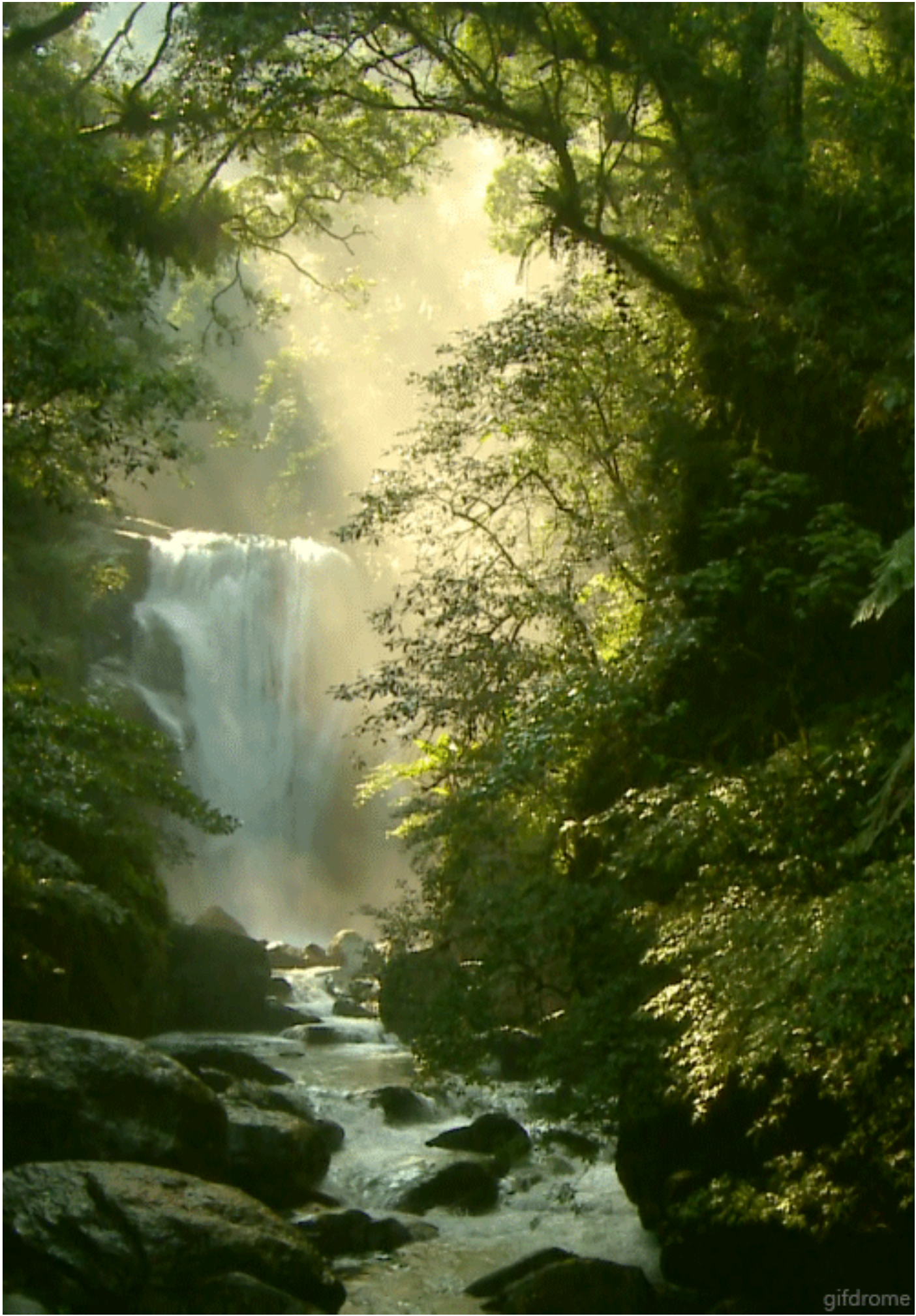
wigner csnl 

Dayan Lab

Max Planck Institute

Tübingen, 2022





l.

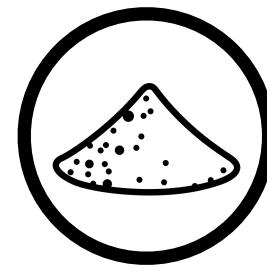




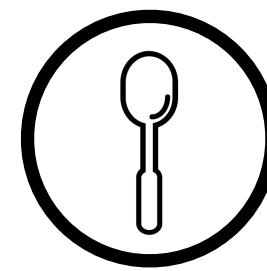
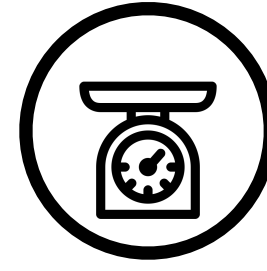




beans



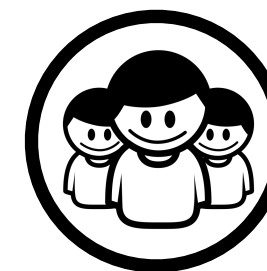
grind  
setting

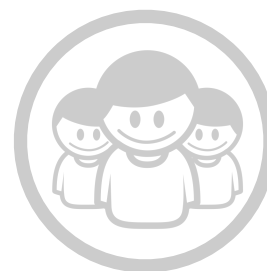
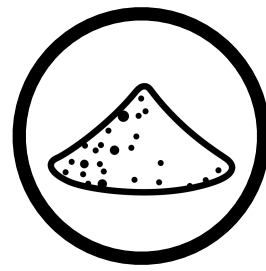


weather



background  
music

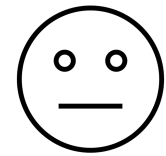




A

5

tap



shirt

18

no

rain

yes

0

B

7

tap



pyjamas

17

yes

sunny

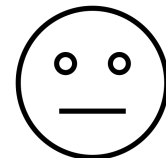
yes

1

B

8

bottled



pyjamas

19

no

sunny

yes

1

B

9

tap



shirt

18

yes

cloudy

no

4

B

10

tap



shirt

14

yes

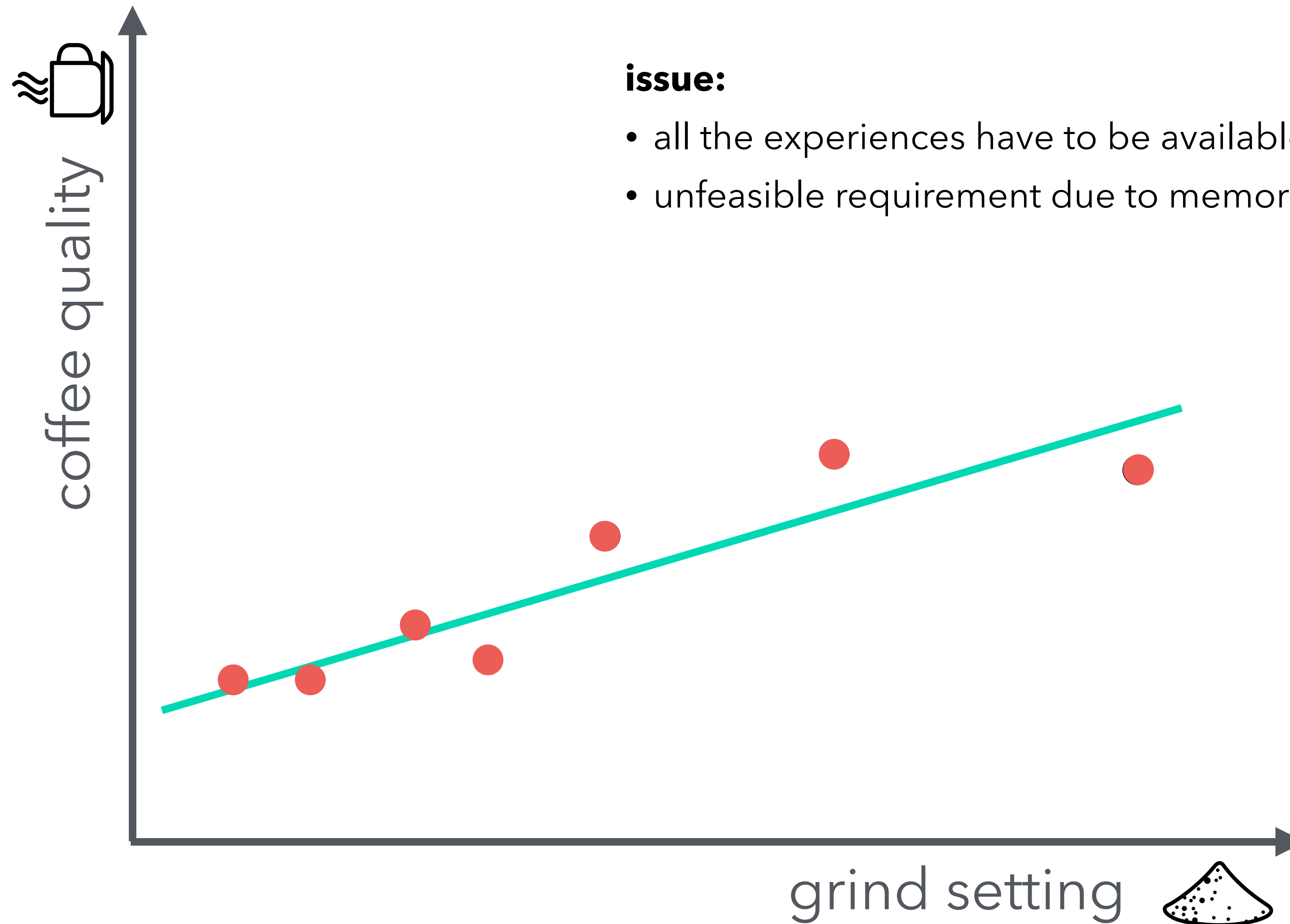
rain

no

1

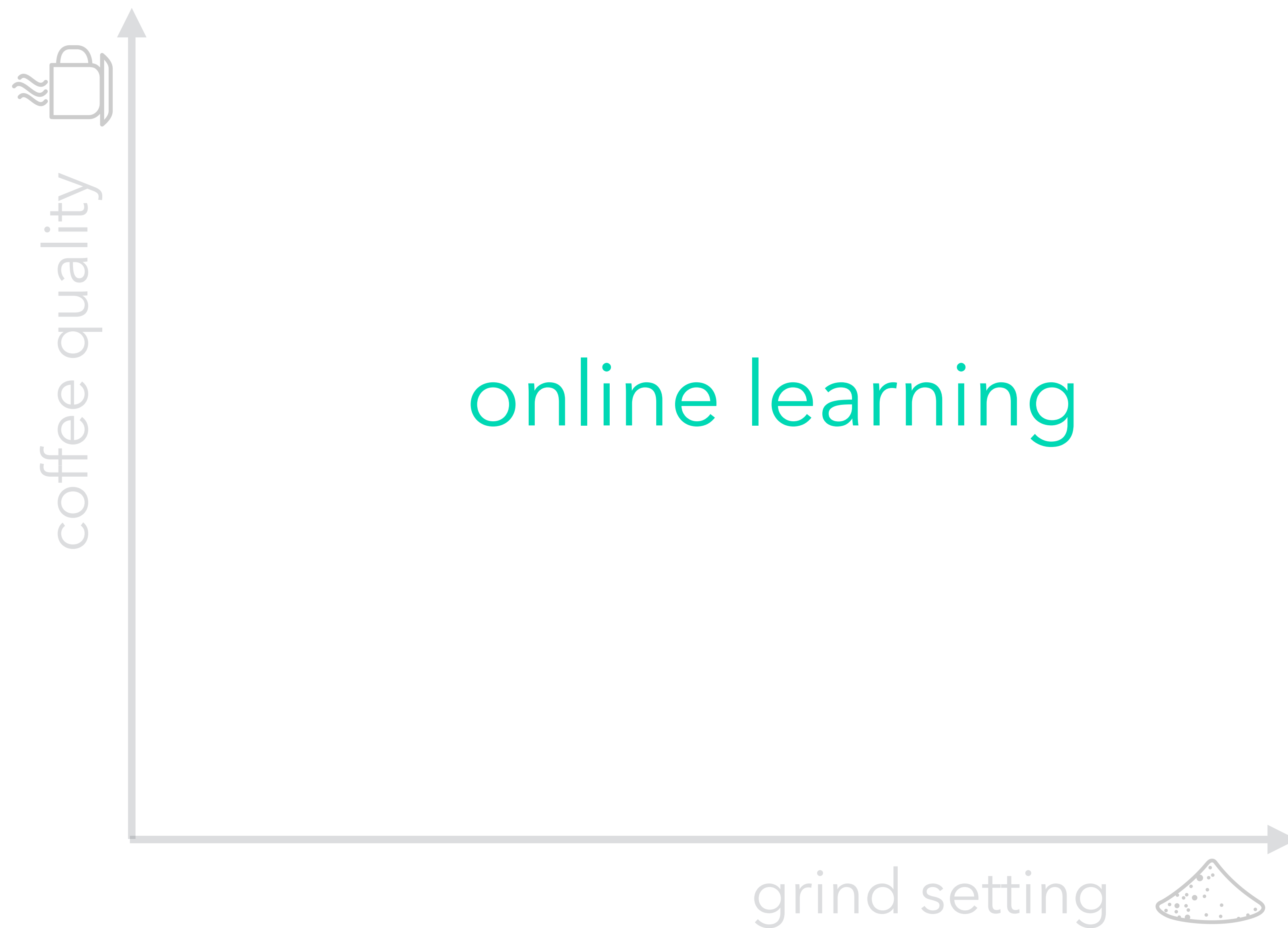


# batch learning












## issue:

- all the experiences have to be available at the same time
- unfeasible requirement due to memory constraints.











# online learning

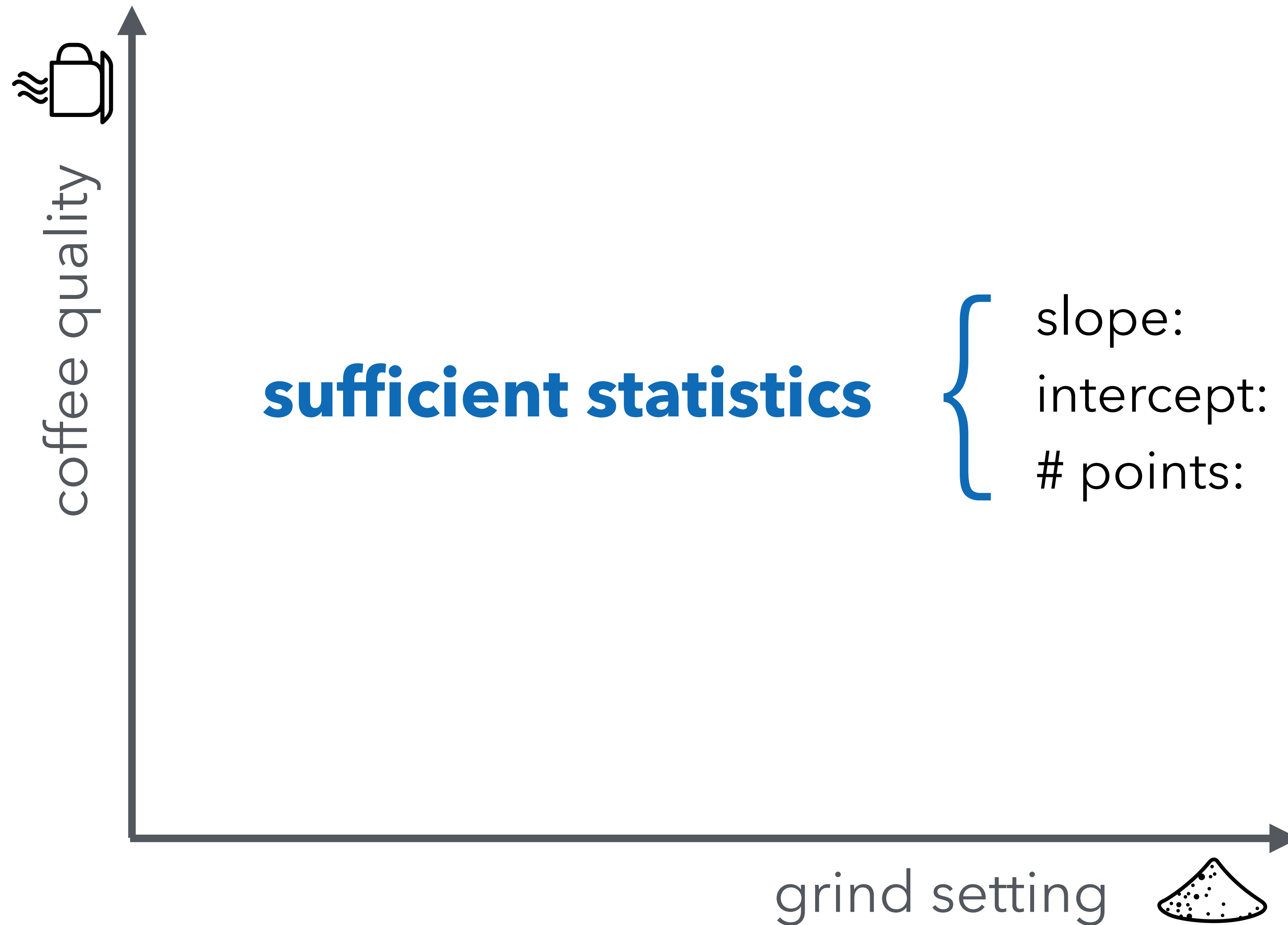
			
A	5	tap	
B	7	tap	
B	8	bottled	
B	9	tap	
B	10	tap	



**sufficient statistics**

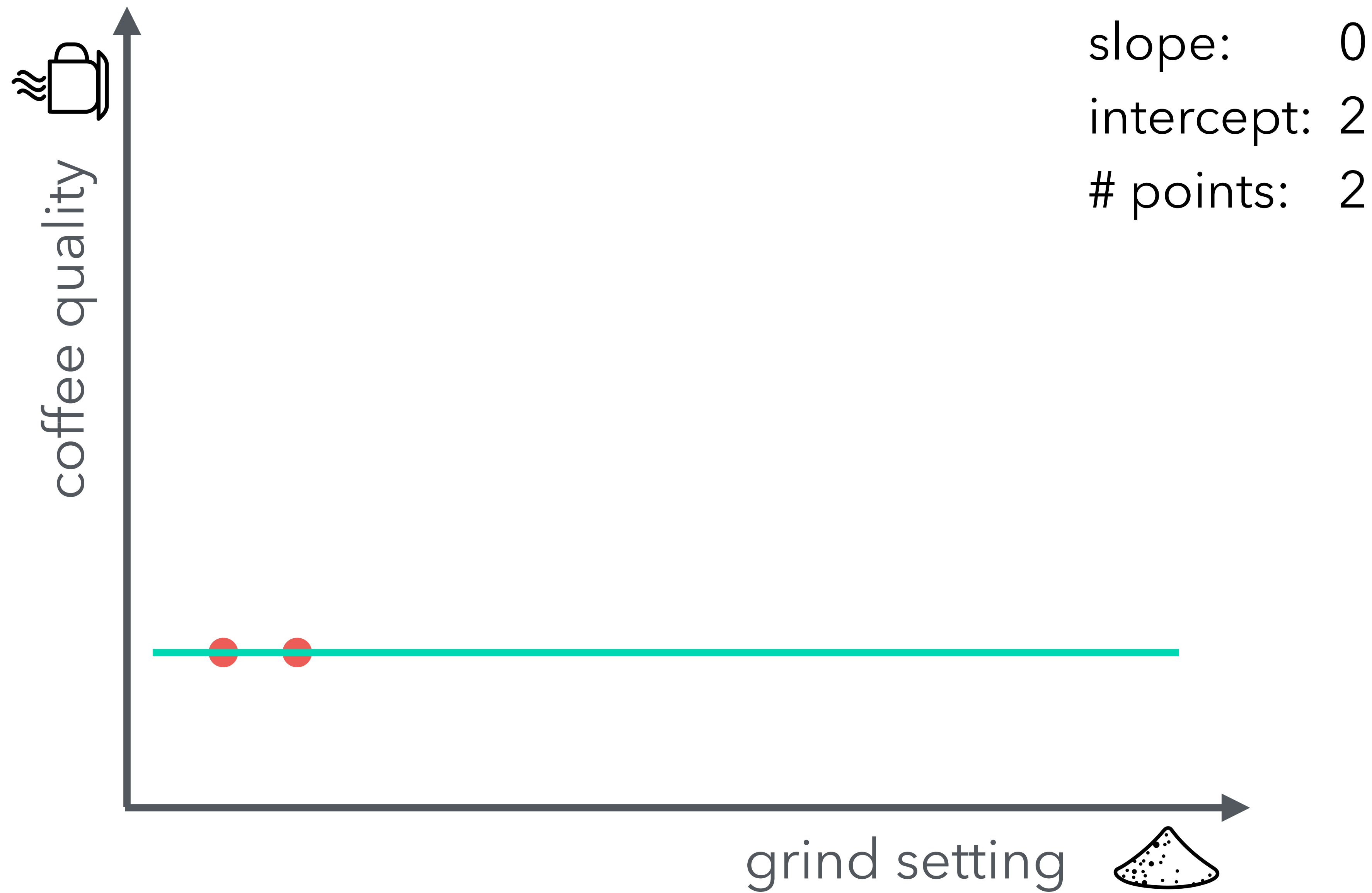
						
shirt	18	no	rain	yes	0	
pyjamas	17	yes	sunny	yes	1	
pyjamas	19	no	sunny	yes	1	
shirt	18	yes	cloudy	no	4	
shirt	14	yes	rain	no	1	

# online learning

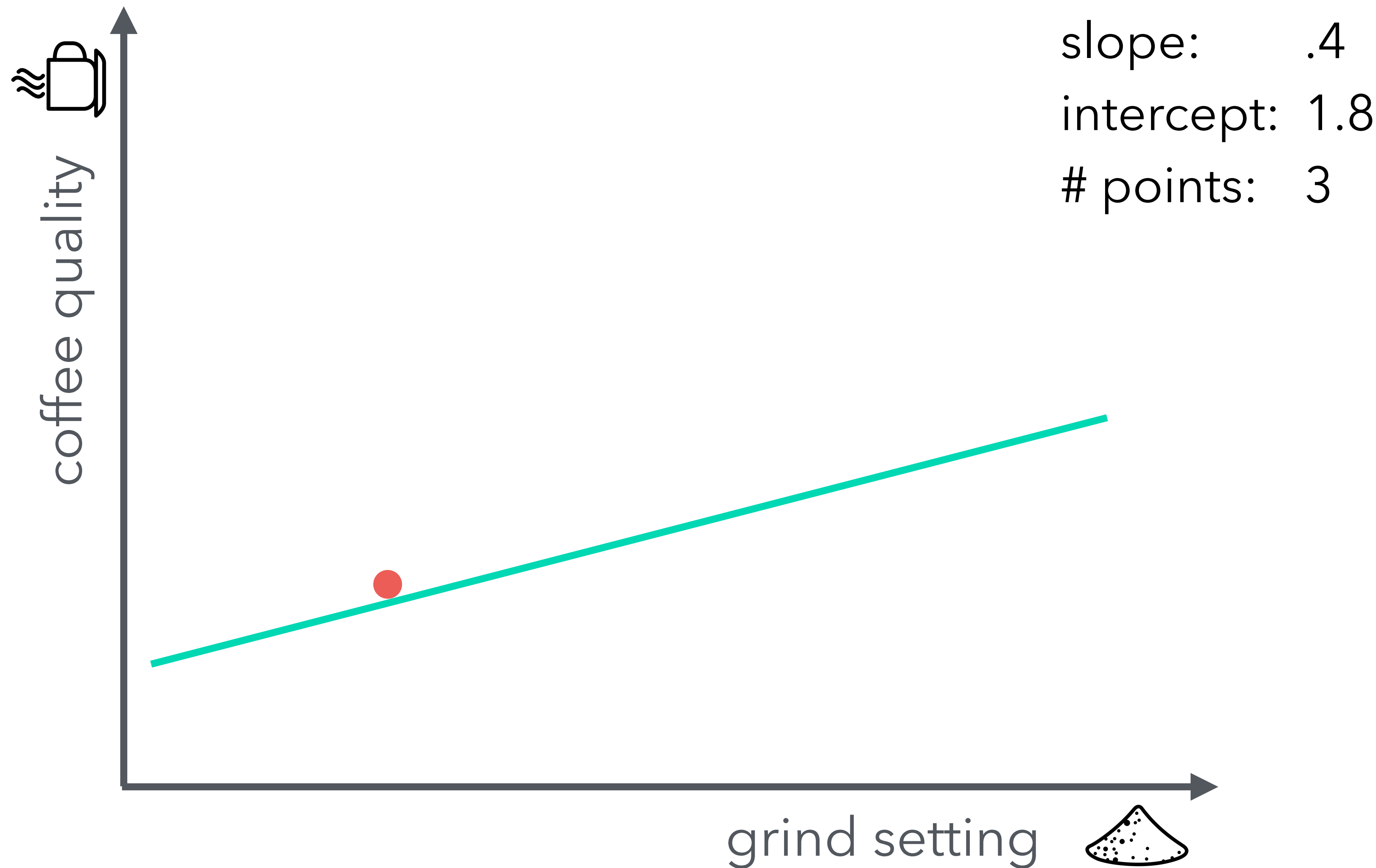




# online learning

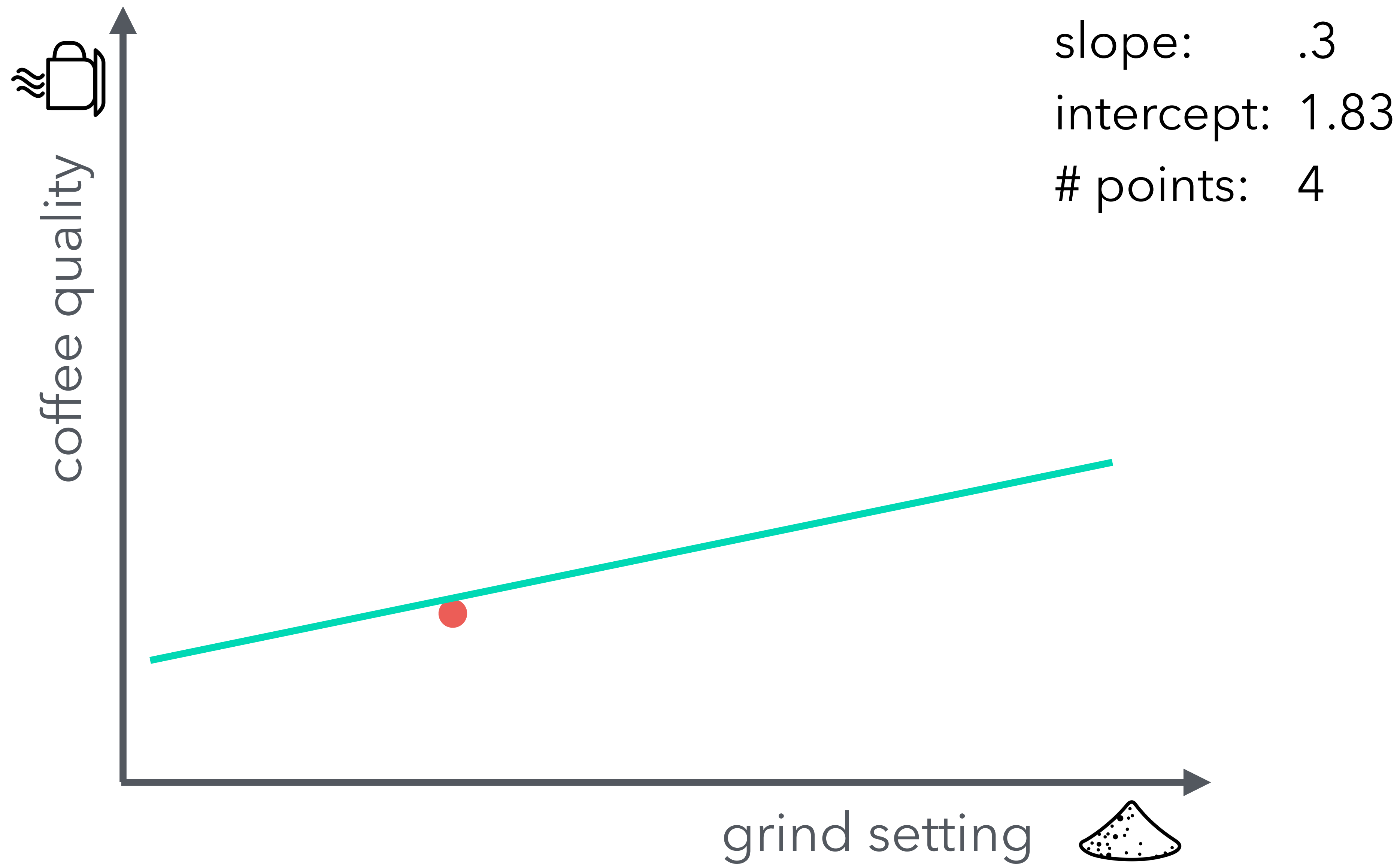


# online learning

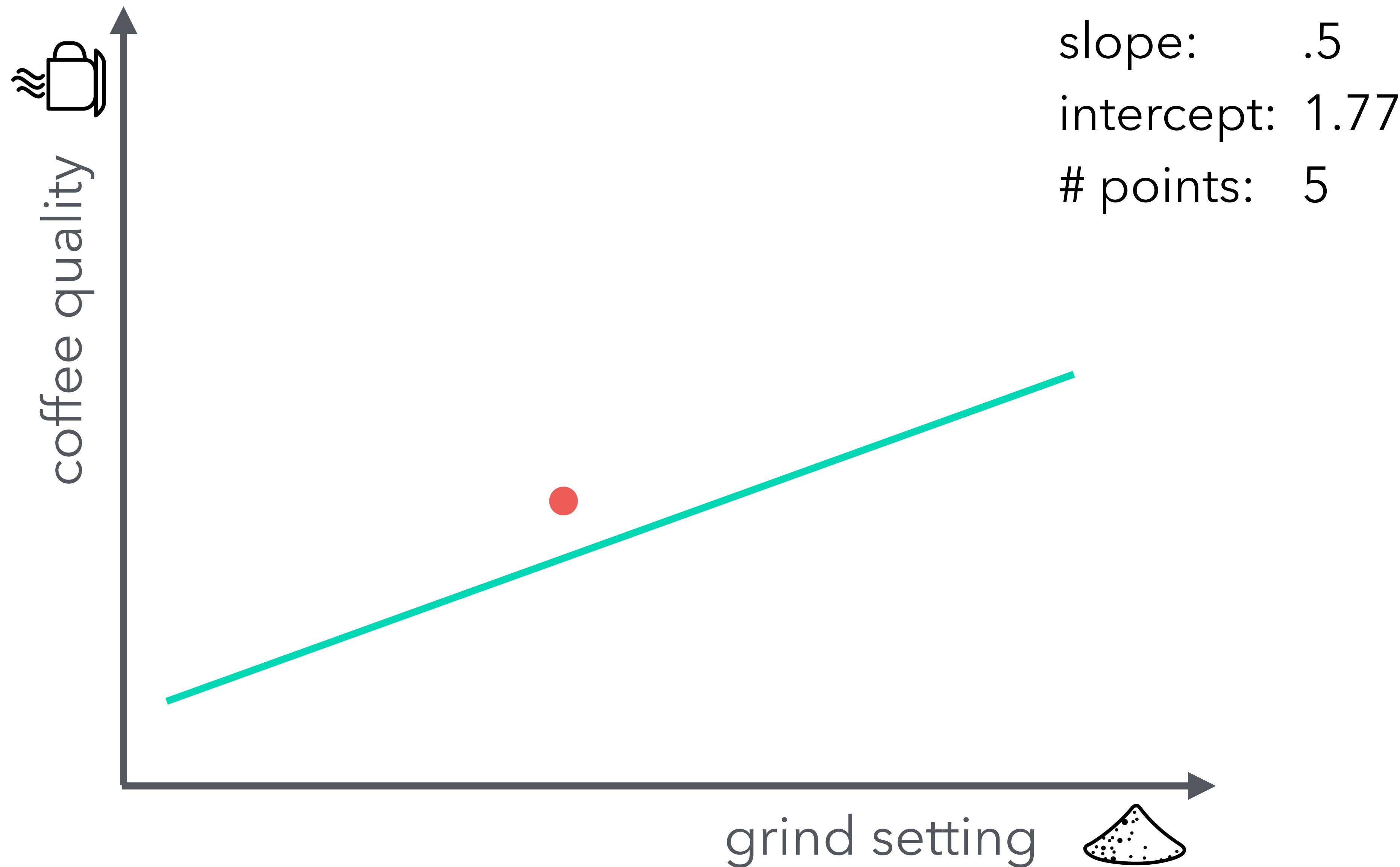




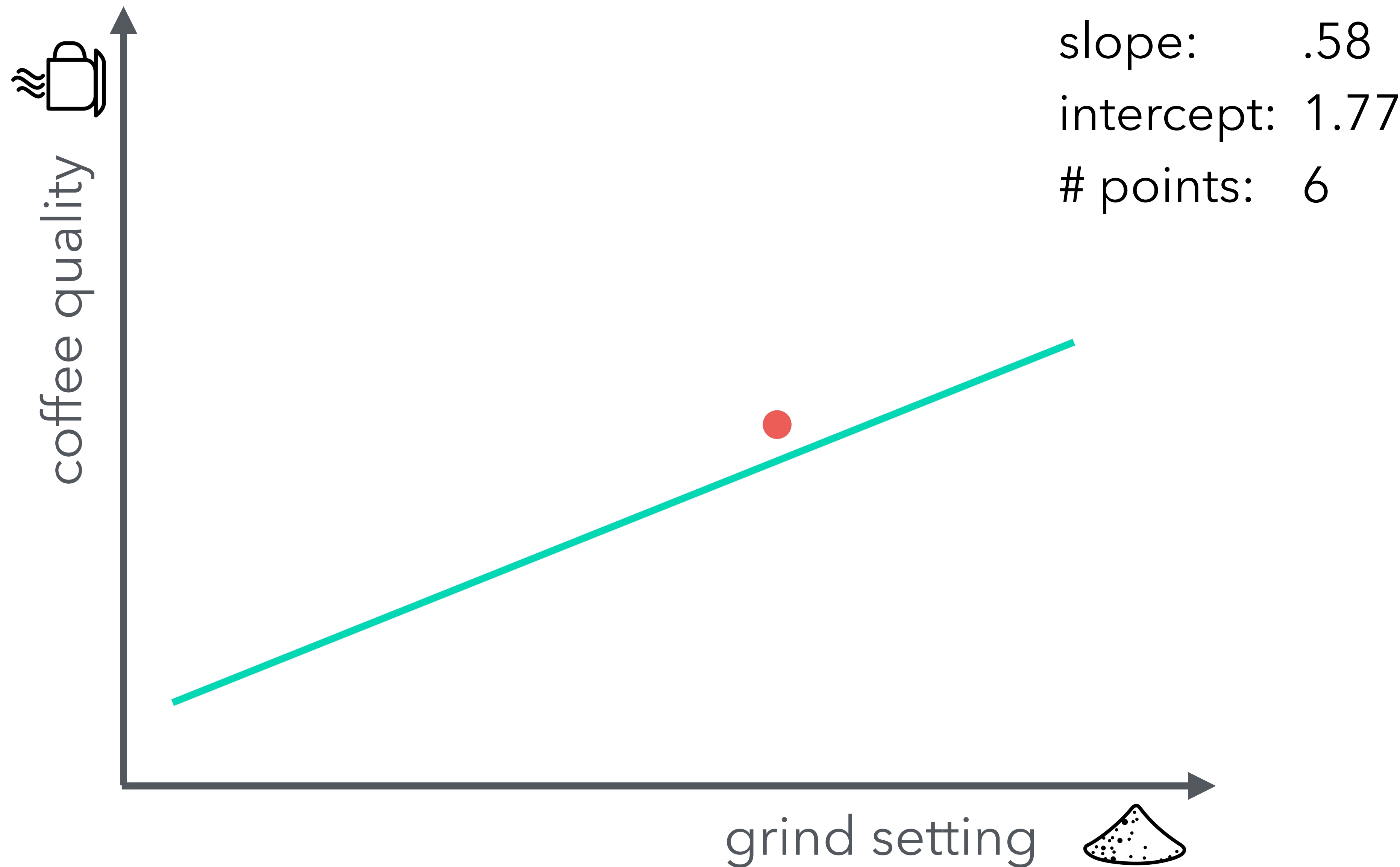
# online learning



# online learning

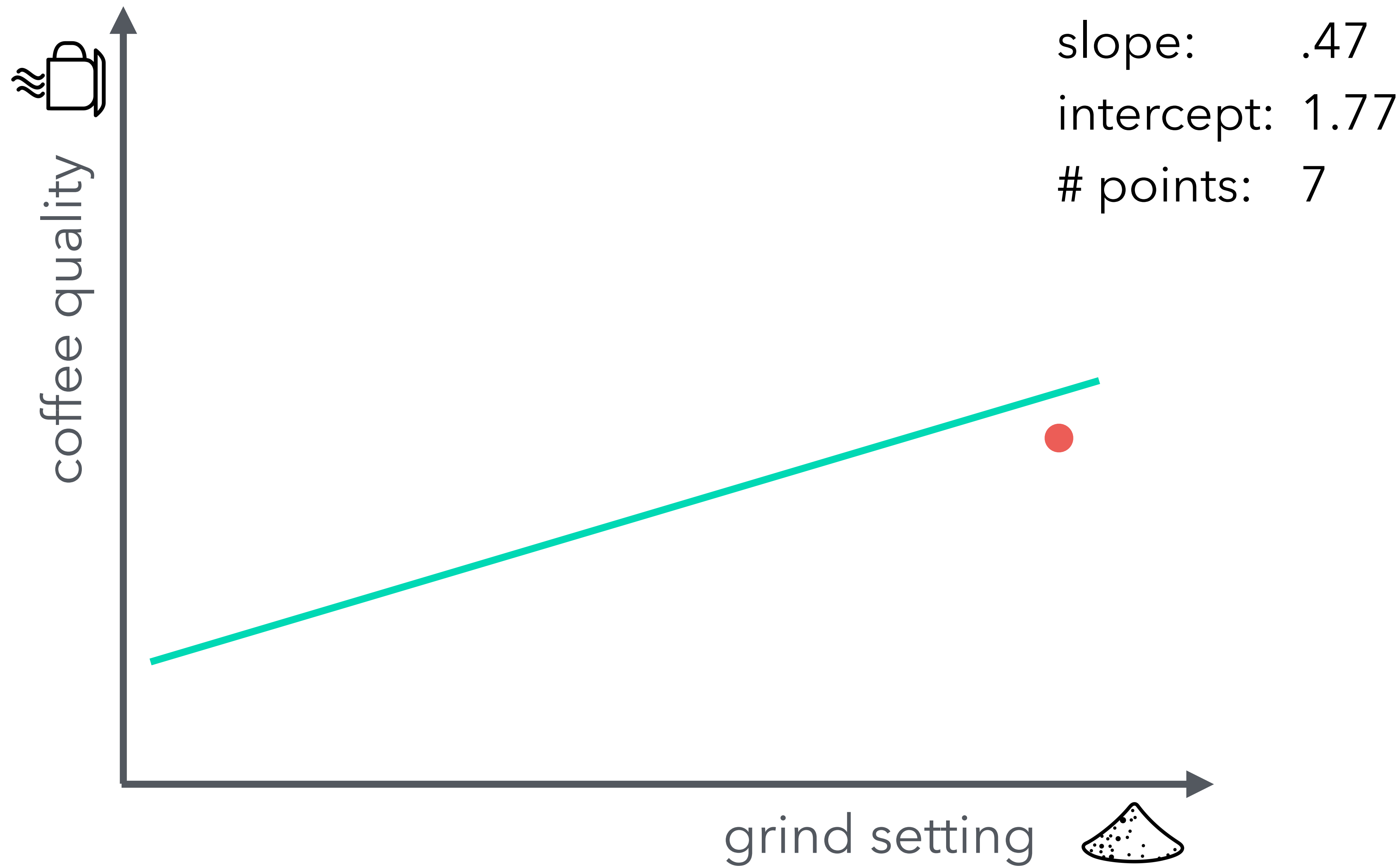


# online learning

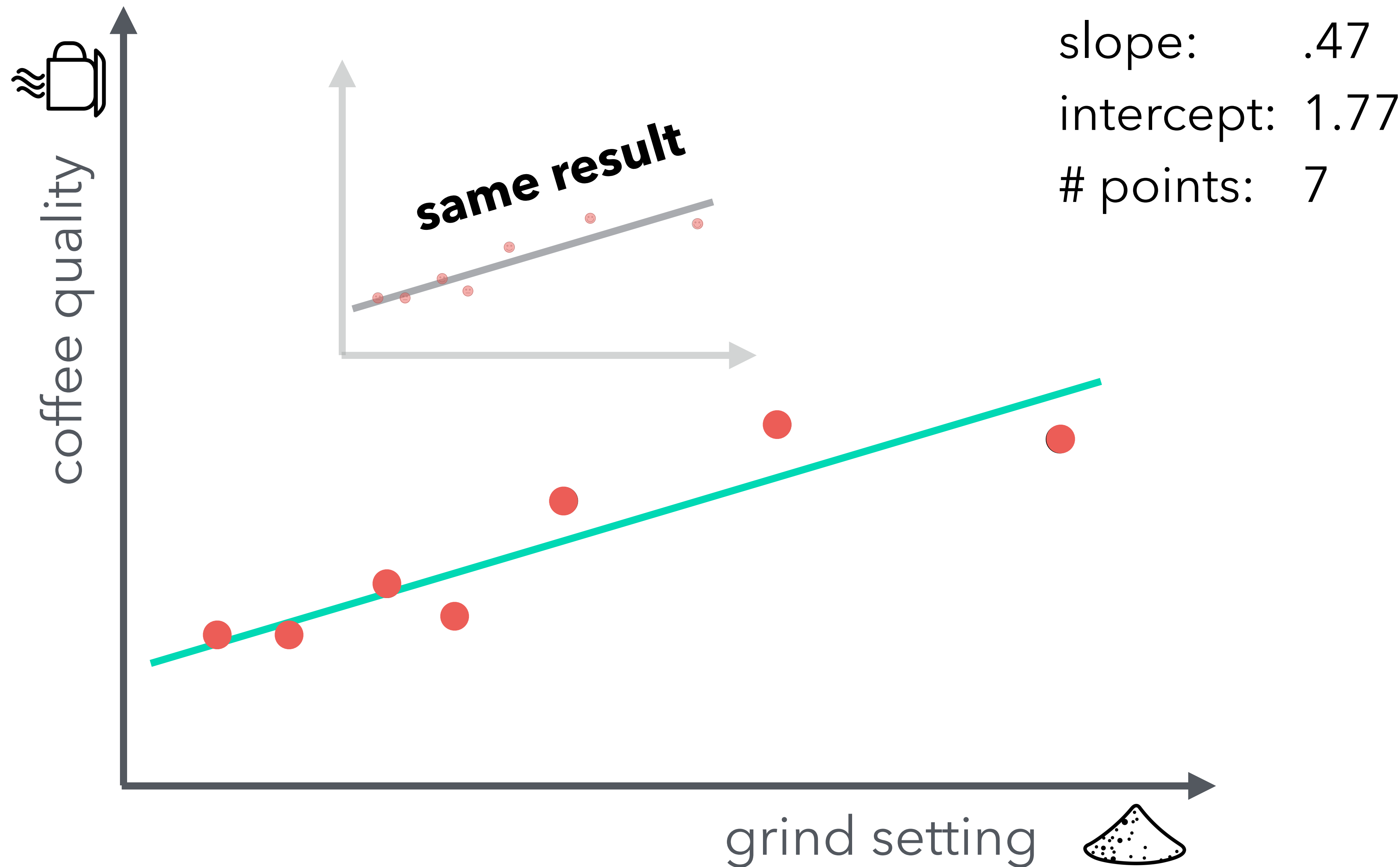




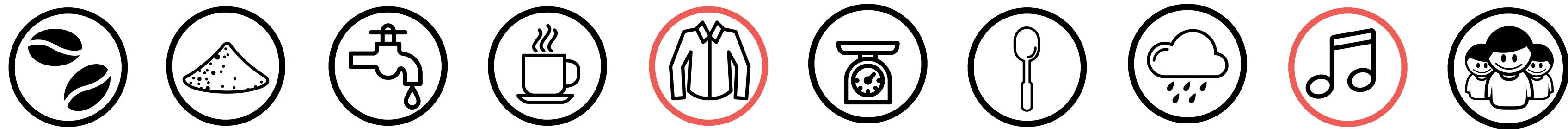
# online learning



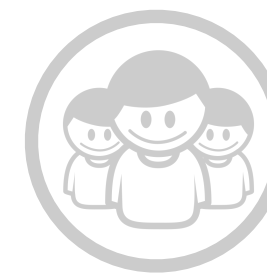
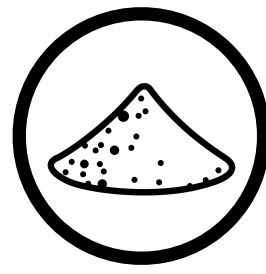
# online learning



what use is an episodic memory? <sup>[1]</sup>



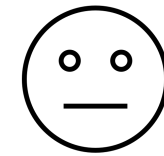
<sup>[1]</sup> Lengyel & Dayan, 2009



A

5

tap



shirt

18

no

rain

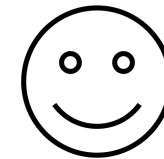
yes

0

B

7

tap



pyjamas

17

yes

sunny

yes

1

⋮

B

9

tap



shirt

18

yes

cloudy

no

4

B

7

tap



shirt

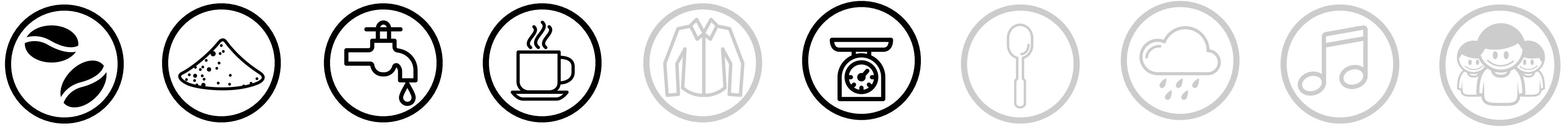
14

yes

rain

no

1



A      5      tap      ☹️      shirt      18      no      rain      yes      0

B      7      tap      😊      pyjamas      17

yes      sunny      yes      1

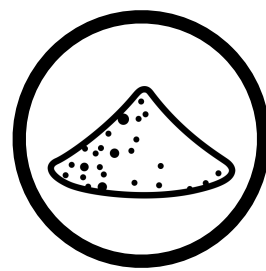
⋮

B      9      tap      😊      shirt      18      yes      cloudy      no      4

B      7      tap      ☹️      shirt      14

yes      rain      no      1

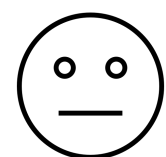




A

5

tap



?

?

?

?

?

?

B

7

tap



?

?

?

?

?

?

⋮

B

9

tap



?

?

?

?

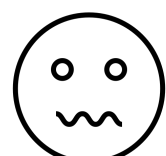
?

?

B

7

tap



?

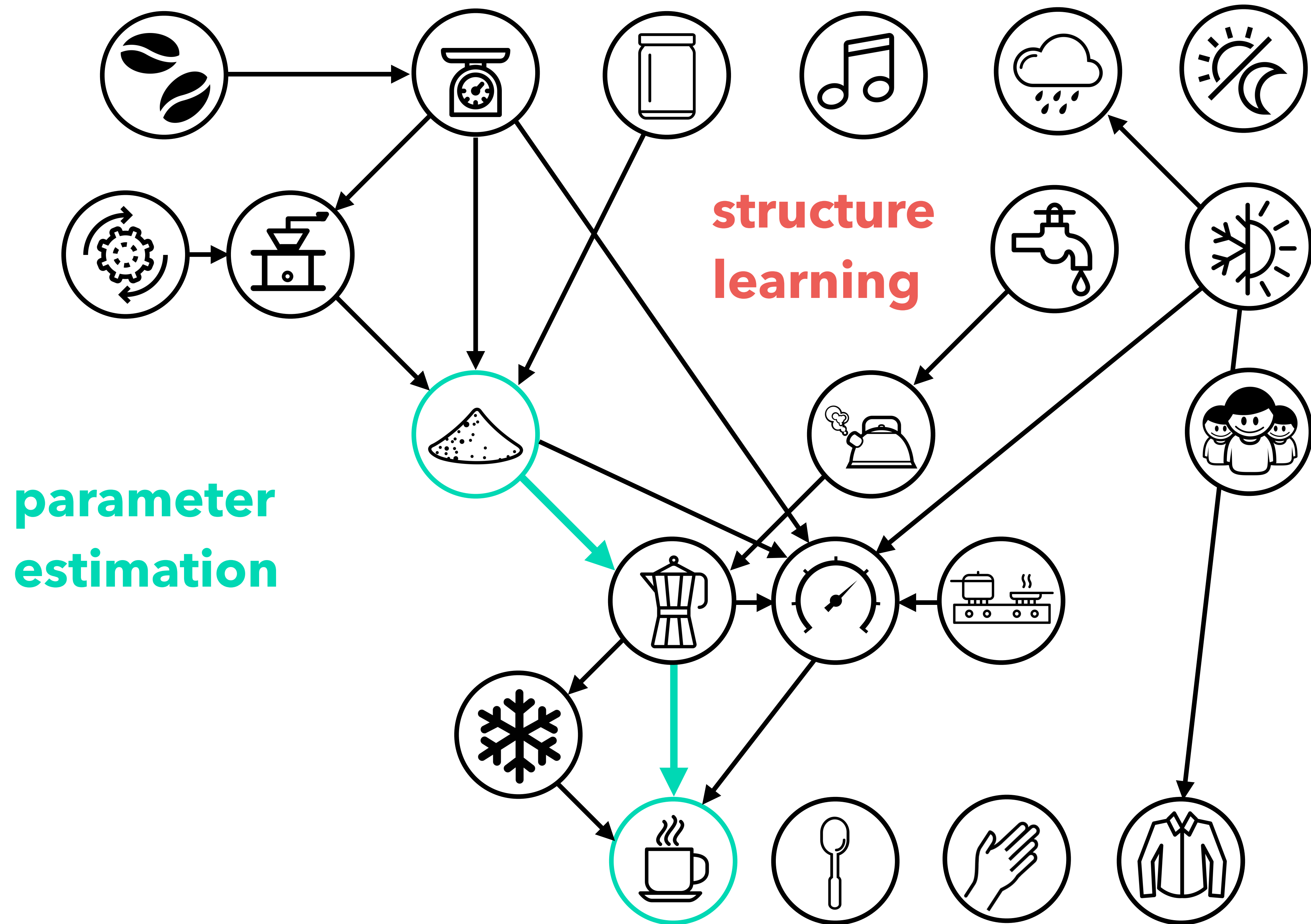
?

?

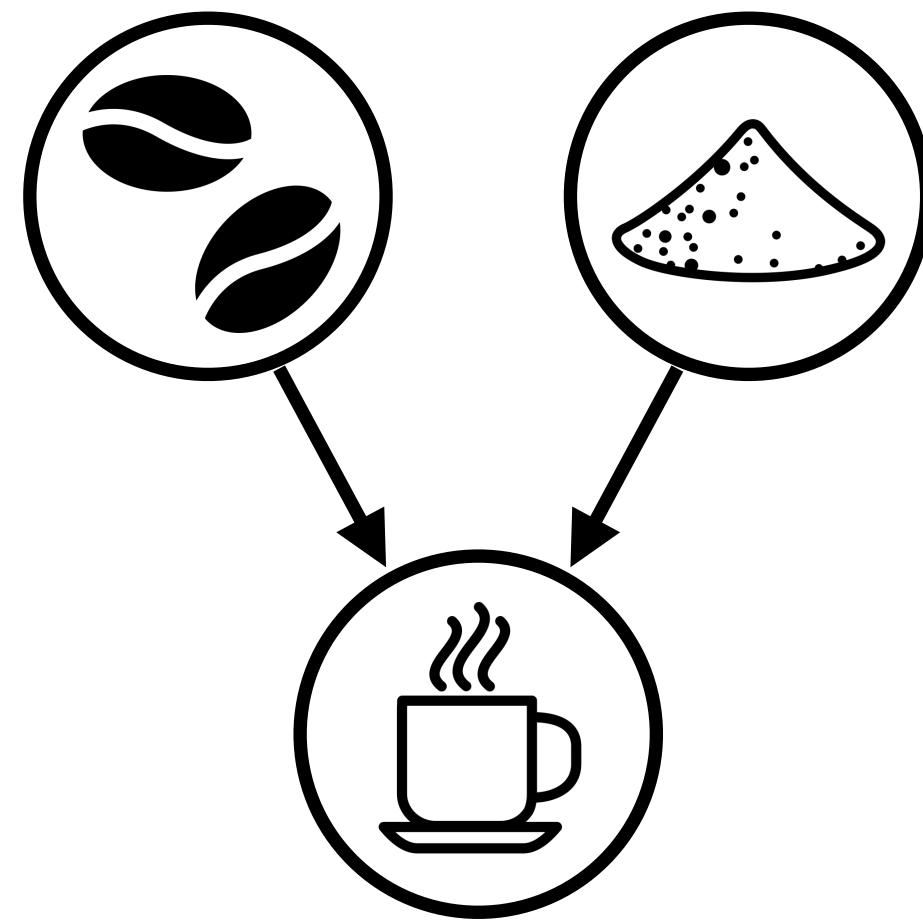
?

?

?

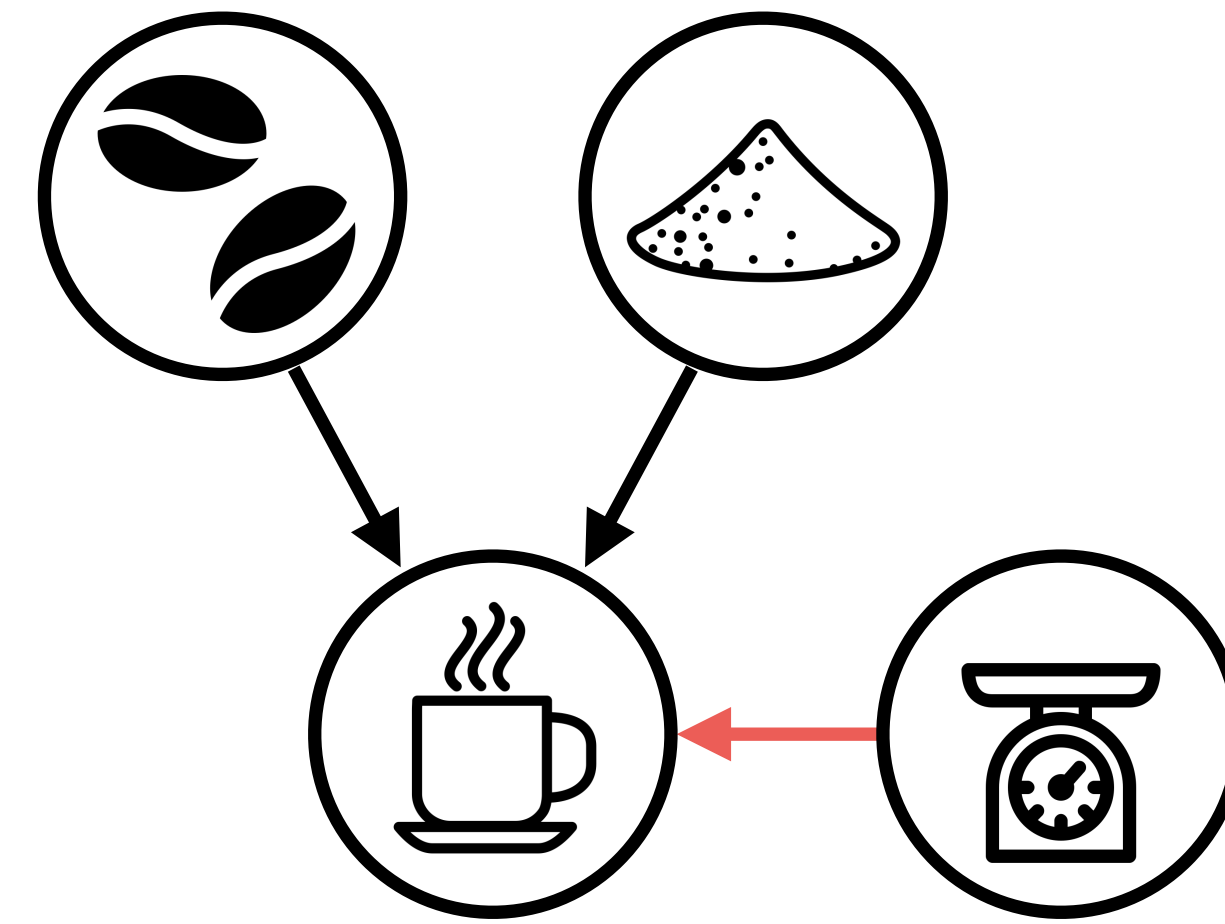


# structure learning



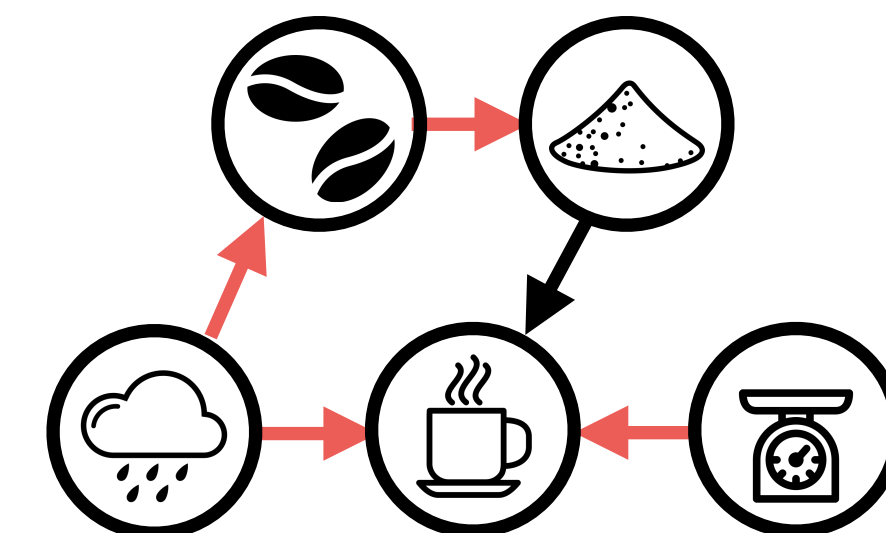
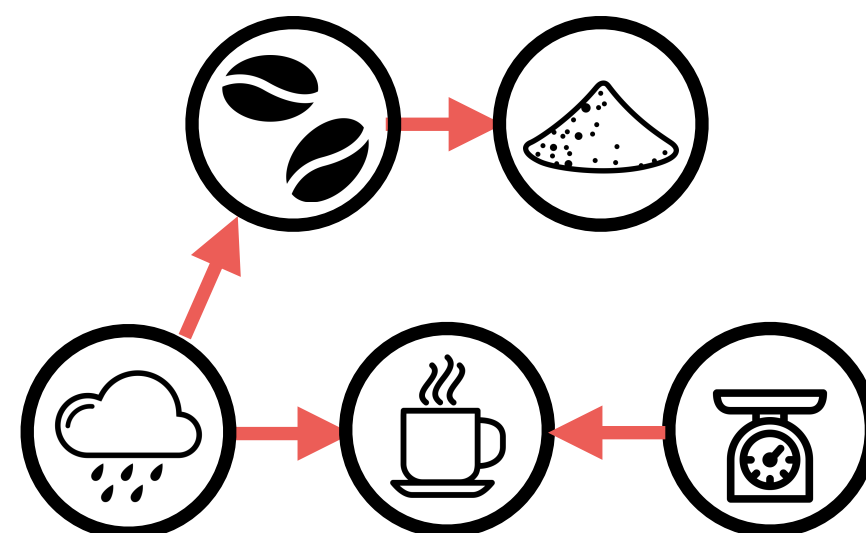
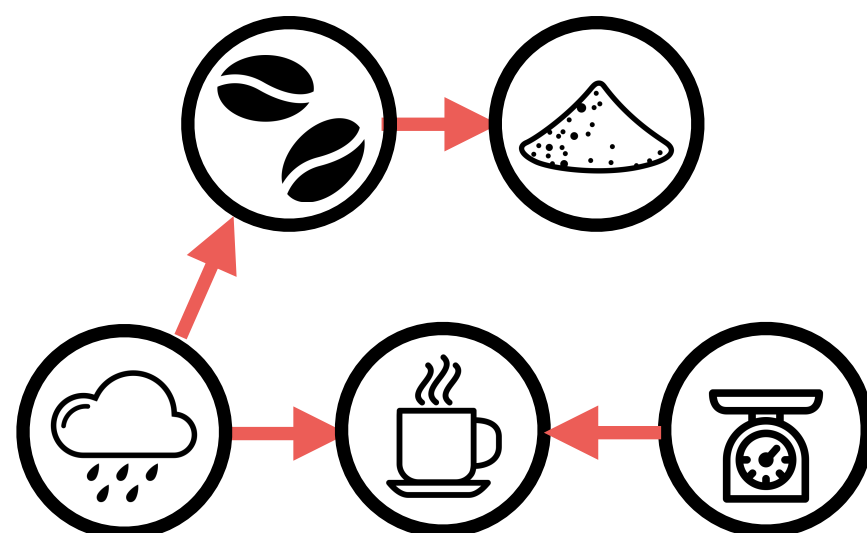
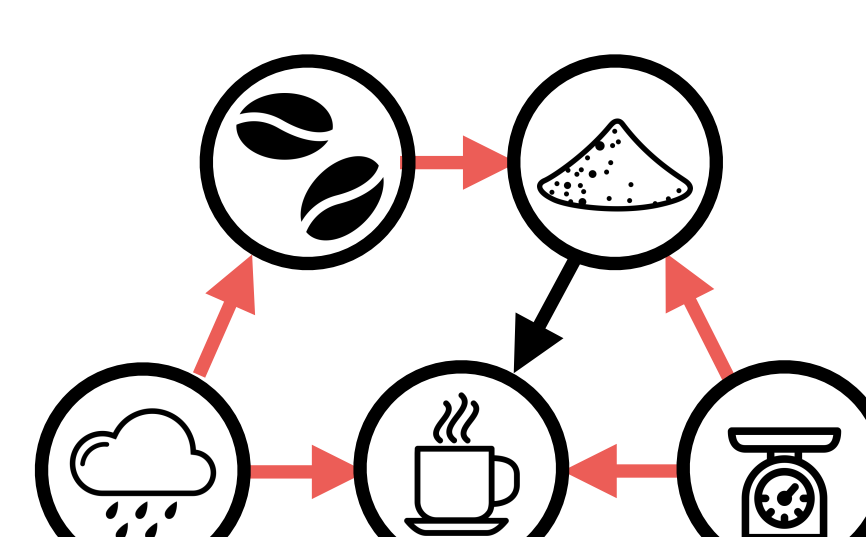
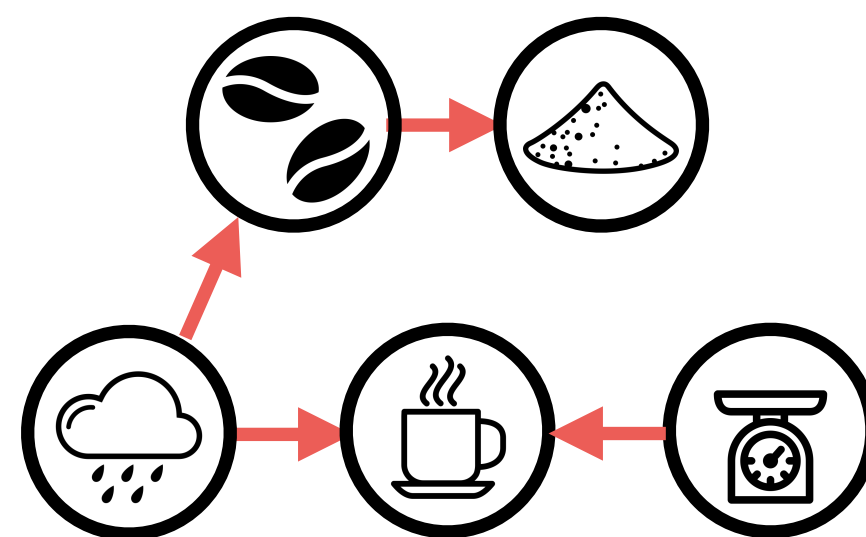
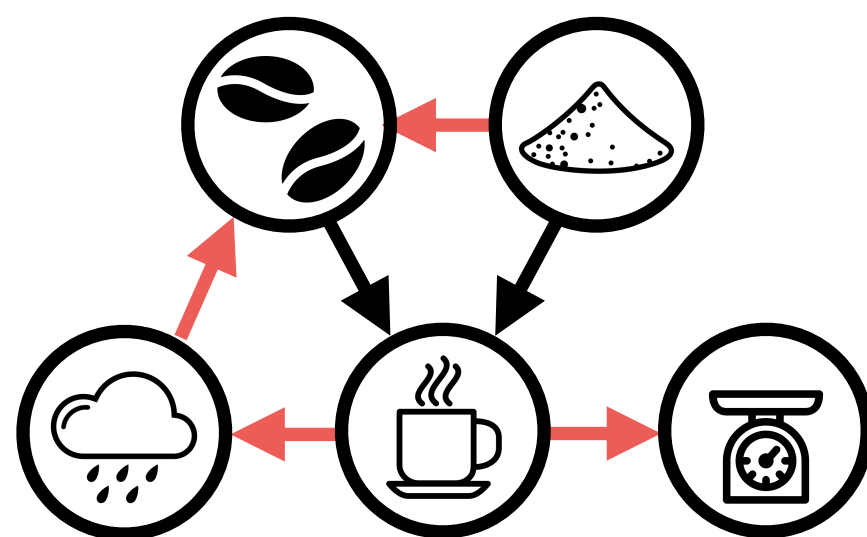
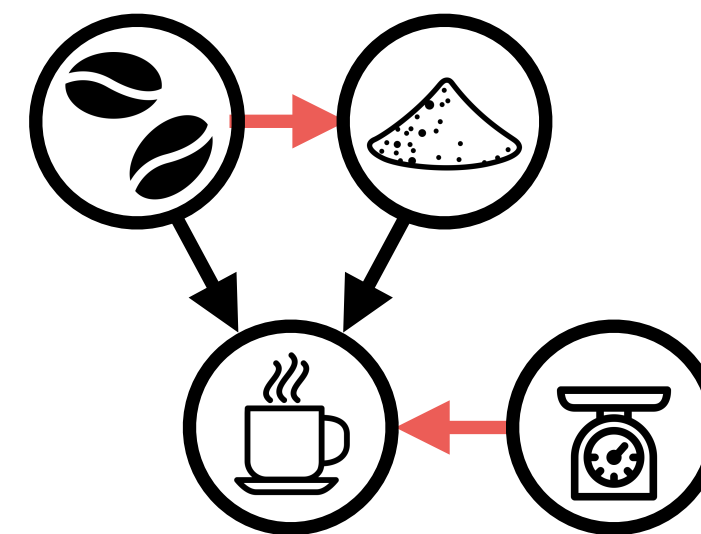
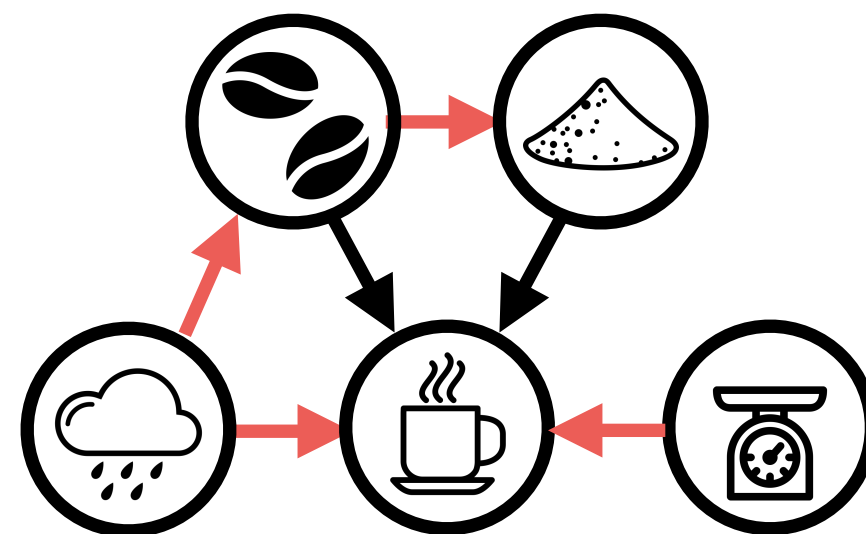
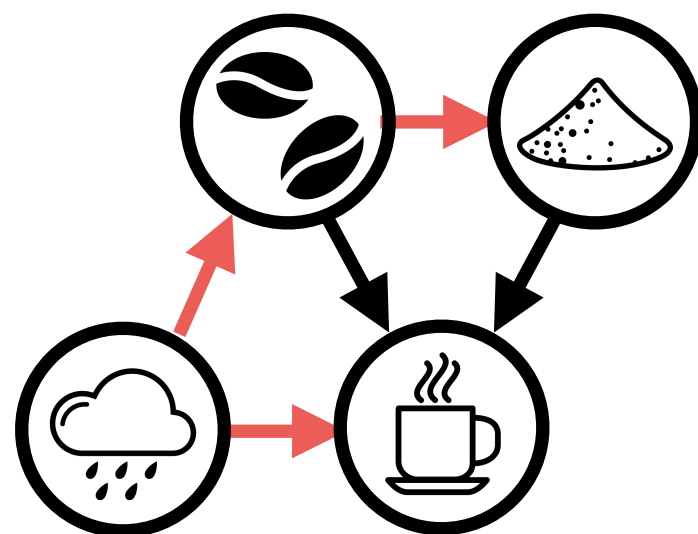
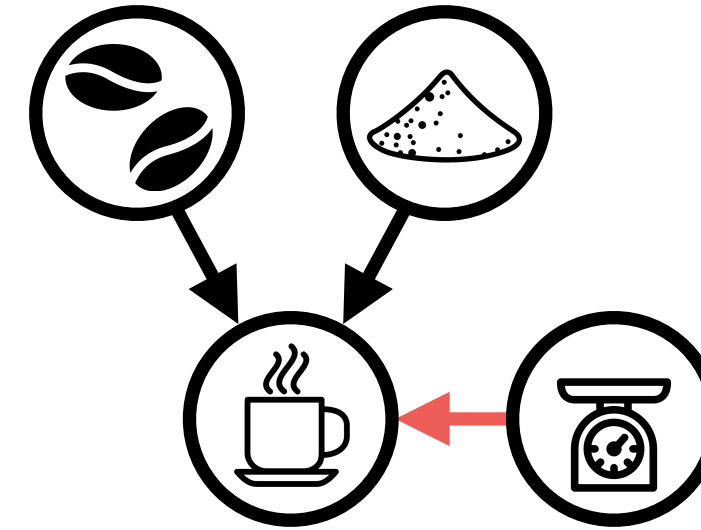
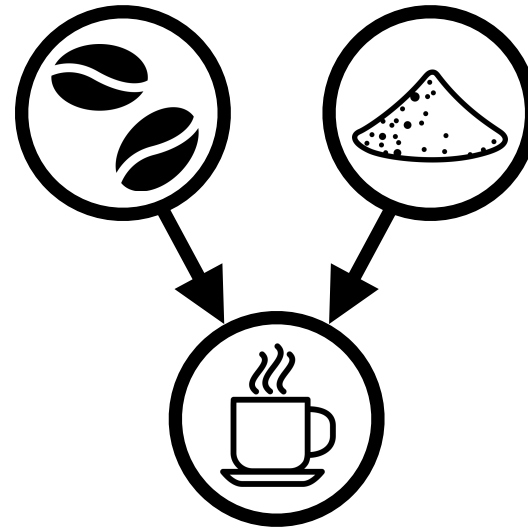
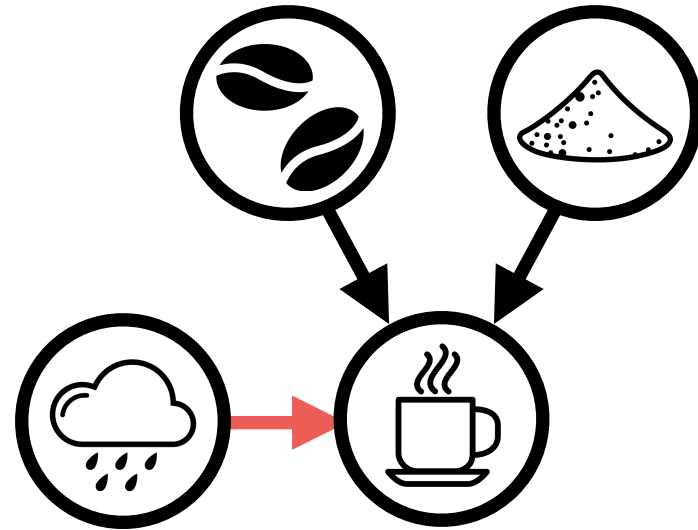
structure 1

**vs**



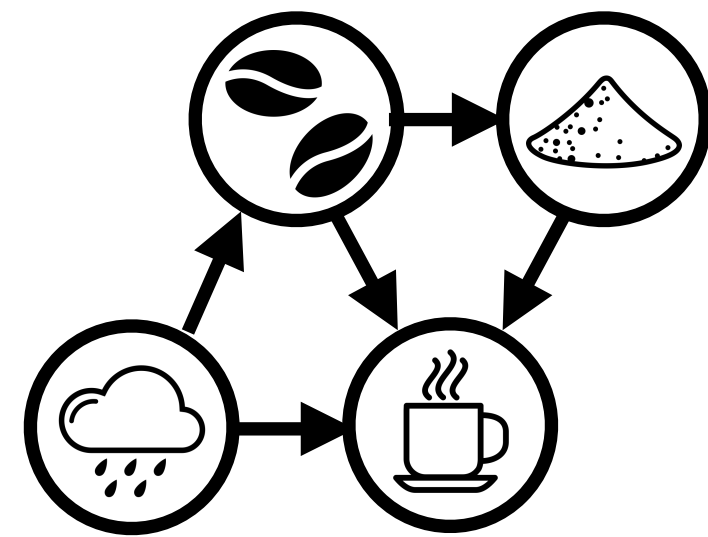
structure 2

- online structure learning
  - comparison of model structures
  - requires sufficient statistics for each candidate structure



# memory systems

- can't store sufficient statistic, can't store data either
- proposed approximate solution:



sufficient statistics  
for best structure

+

										
A	5	tap		shirt	18	no	rain	yes	0	
B	7	tap		pyjamas	17	yes	sunny	yes	1	
B	8	bottled		pyjamas	19	no	sunny	yes	1	

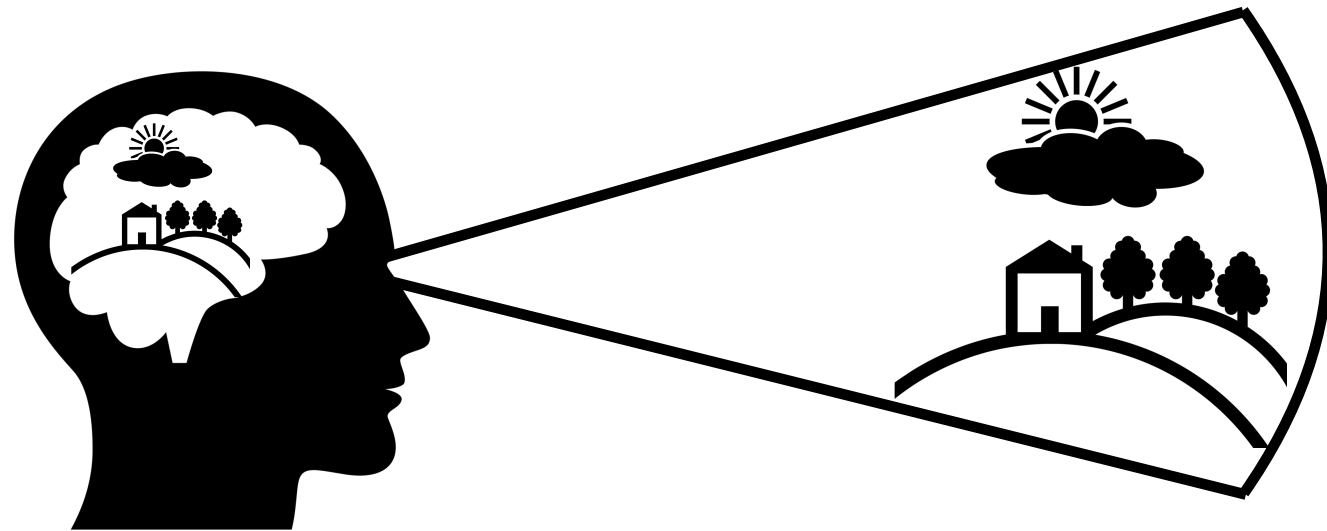
subset of episodes



# memory systems

## **semantic**

general knowledge about  
how the world works

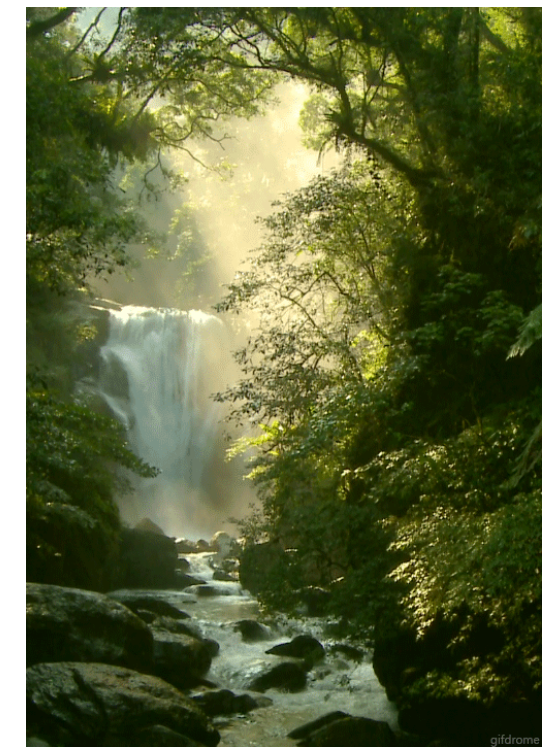


$$p(x, z, \theta \mid \mathcal{D})$$

a probabilistic model of the environment

## **episodic**

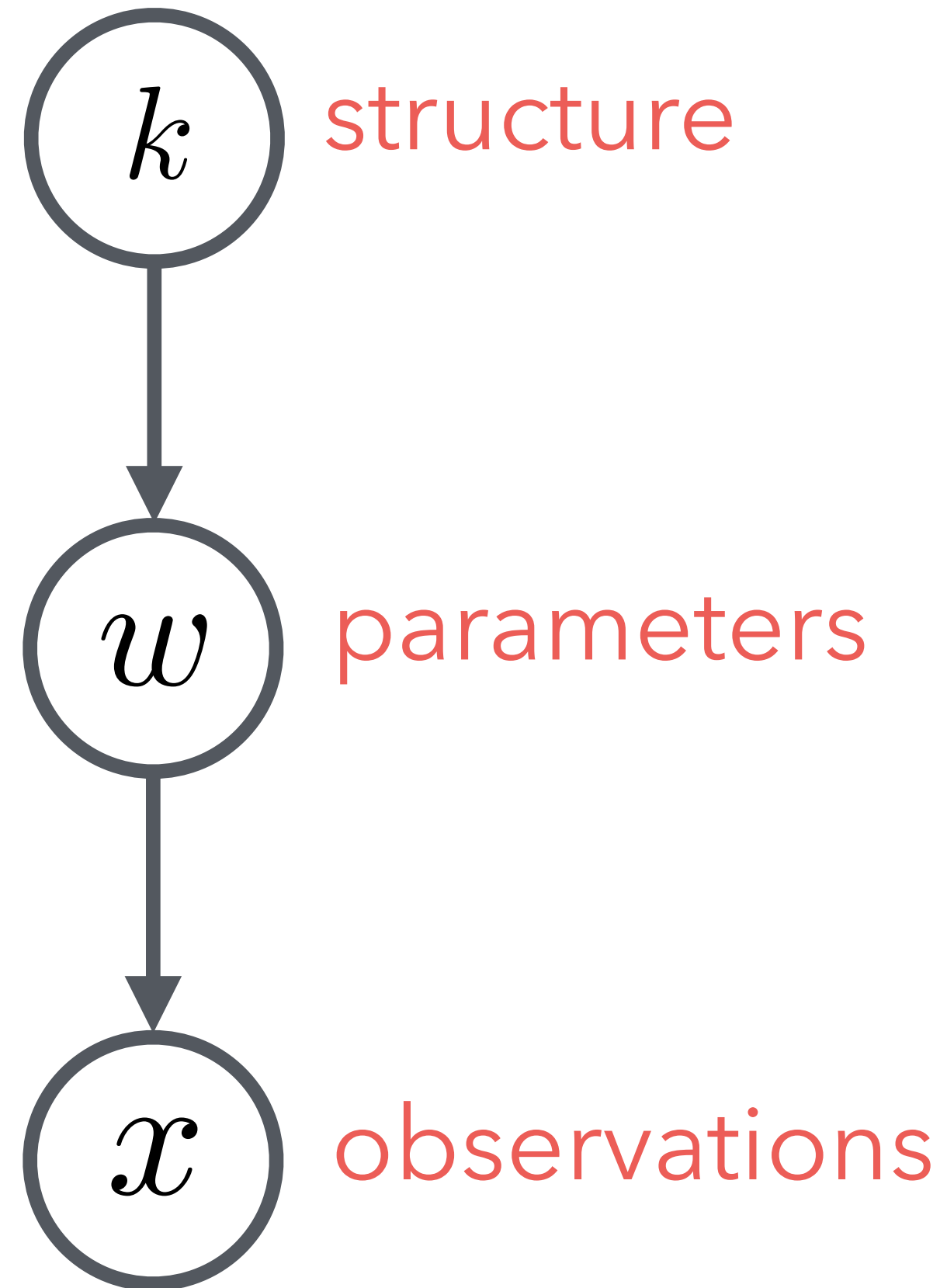
concrete experiences



$$\{x_t\} \subset \mathcal{D}$$

a subset of observations

# formalisation

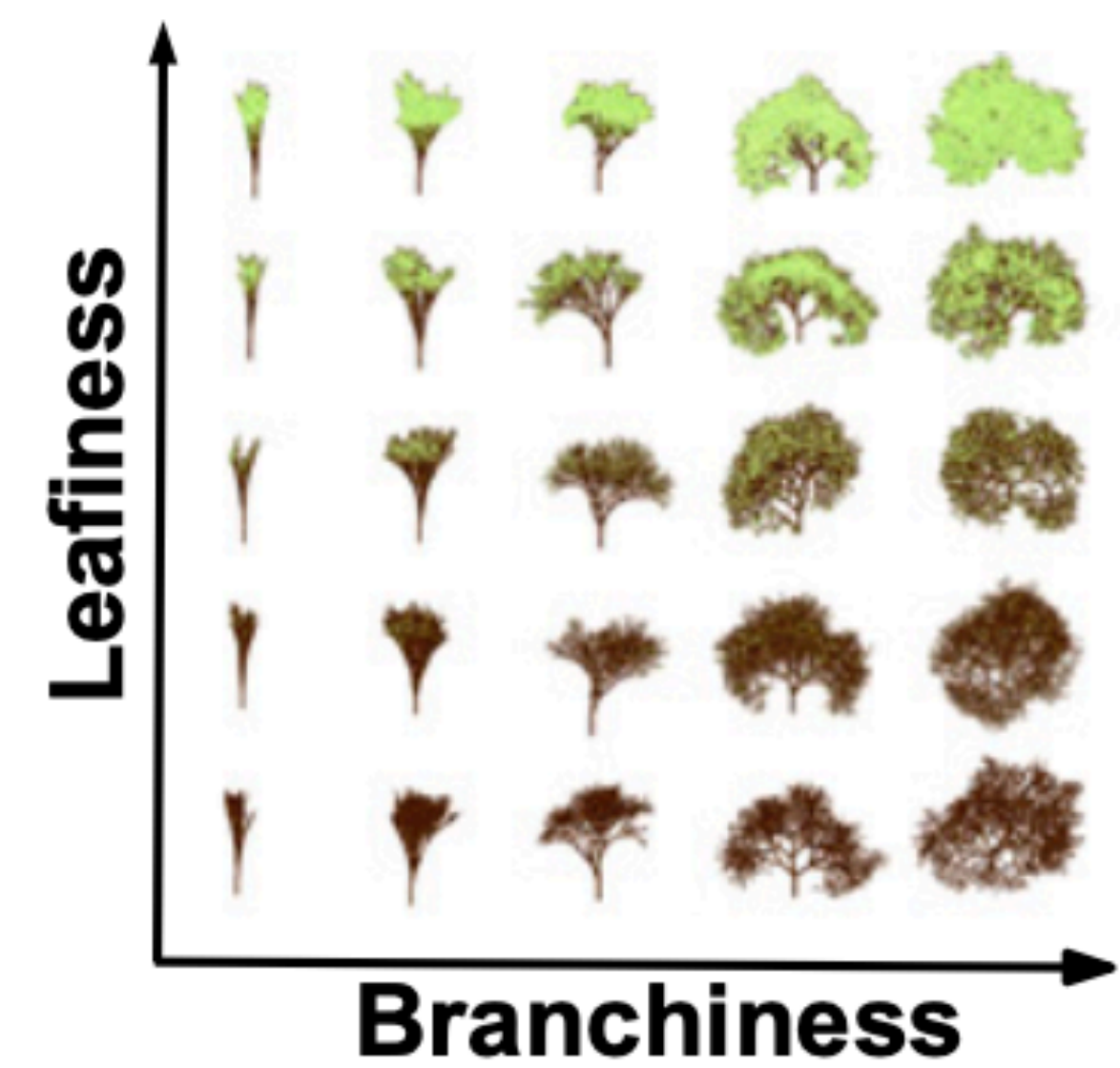
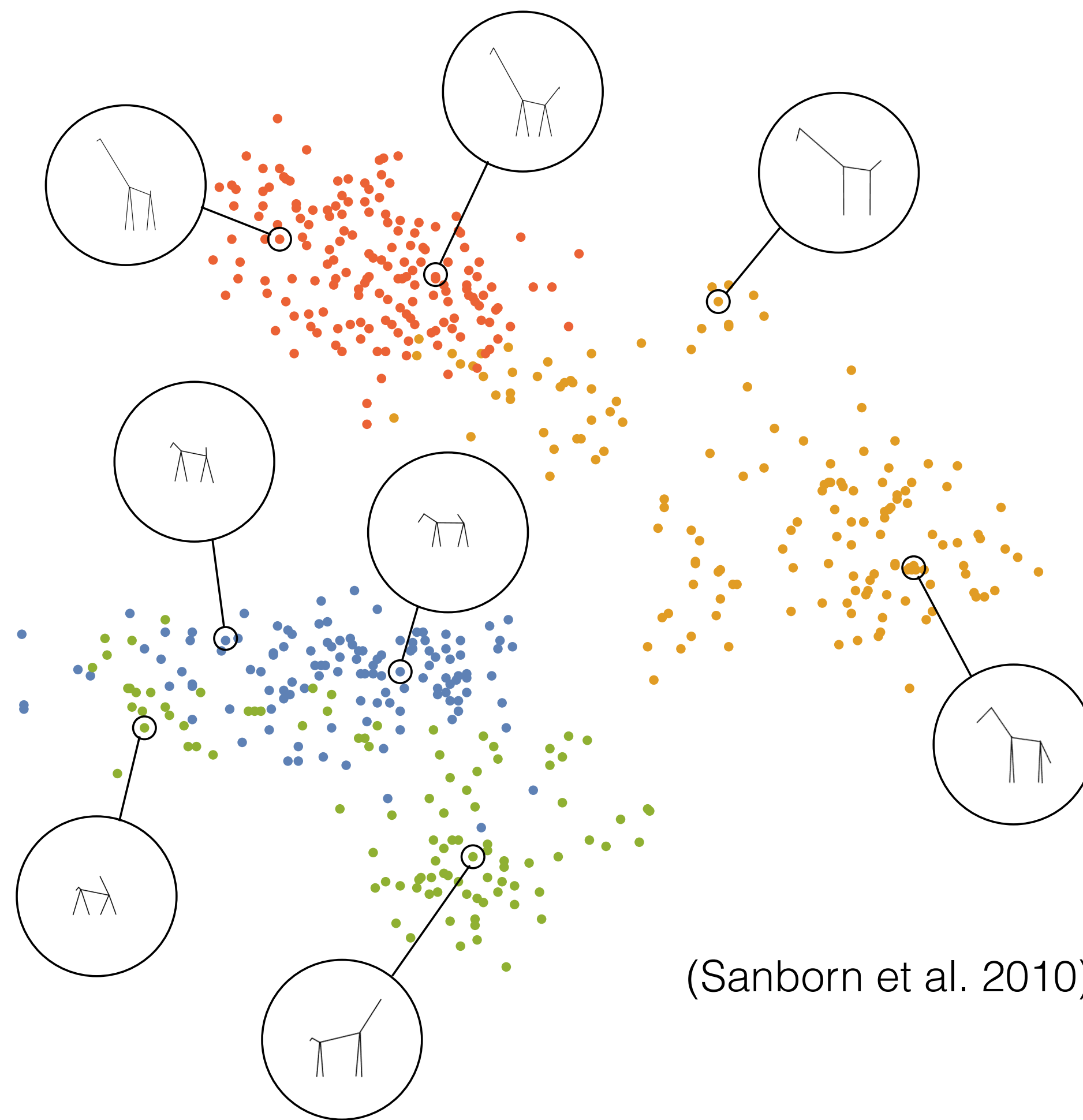


$$P(\text{structure}|\text{episodes})$$

$$P(\text{parameter}|\text{episodes}, \text{structure})$$

**sufficient statistics**

$$P(x|\text{episodes})$$

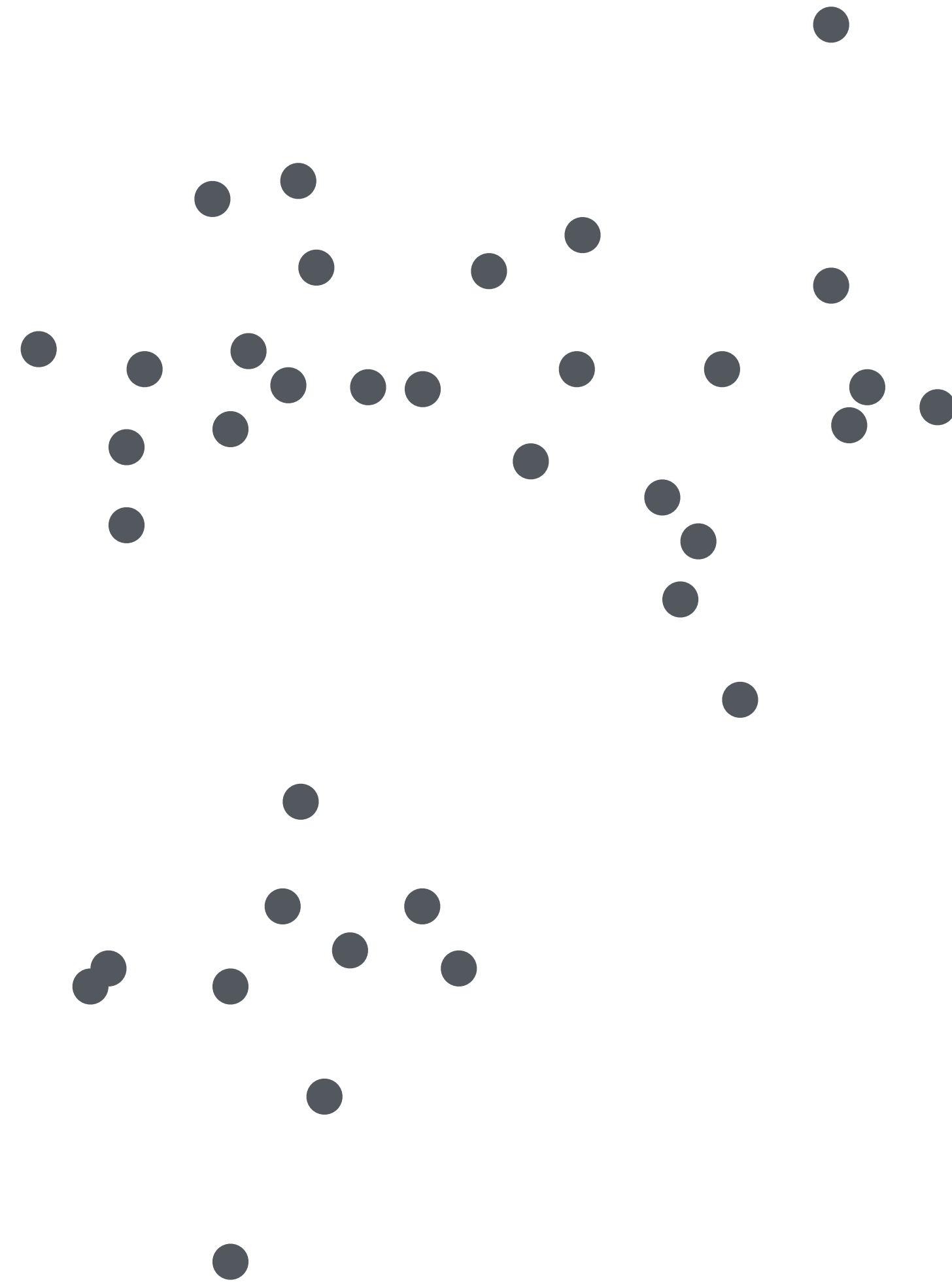


(Flesch et al. 2018)

# toy setting

## Mixture of Gaussians

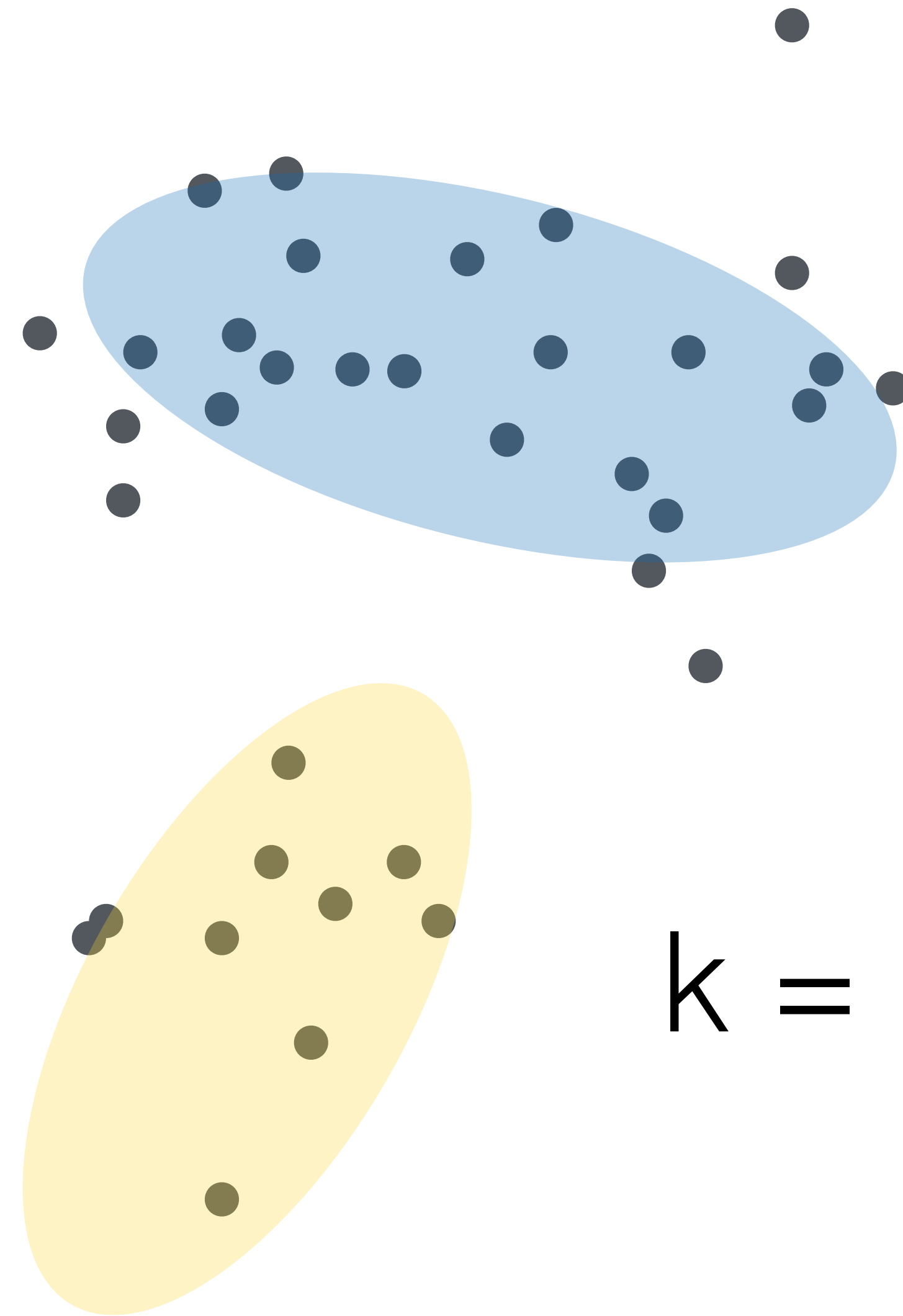
- known variance and mixing coefficients
- unknown number of components  
(**structure learning**)



# toy setting

Mixture of Gaussians

- known variance and mixing coefficients
- unknown number of components  
(**structure learning**)

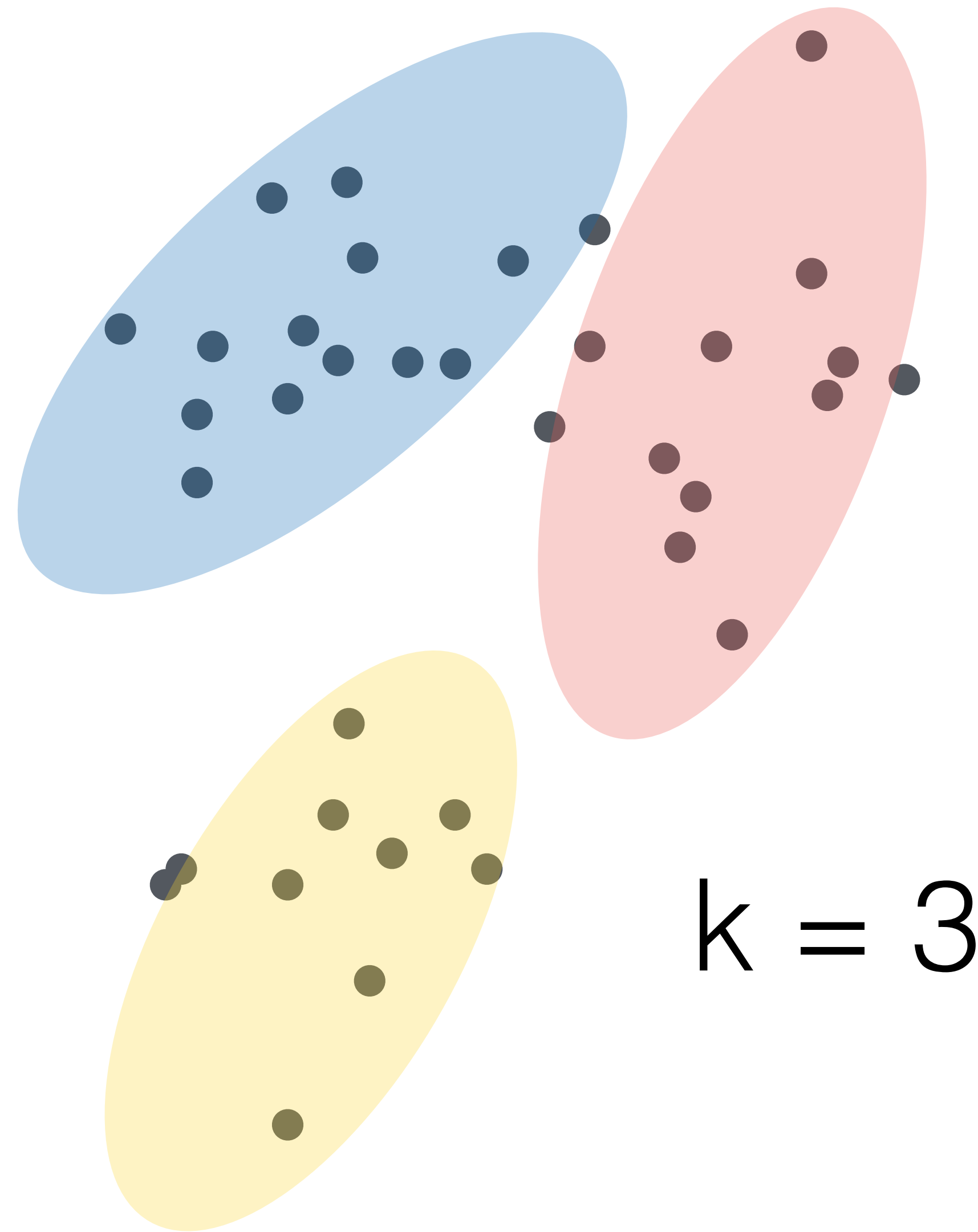




# toy setting

Mixture of Gaussians

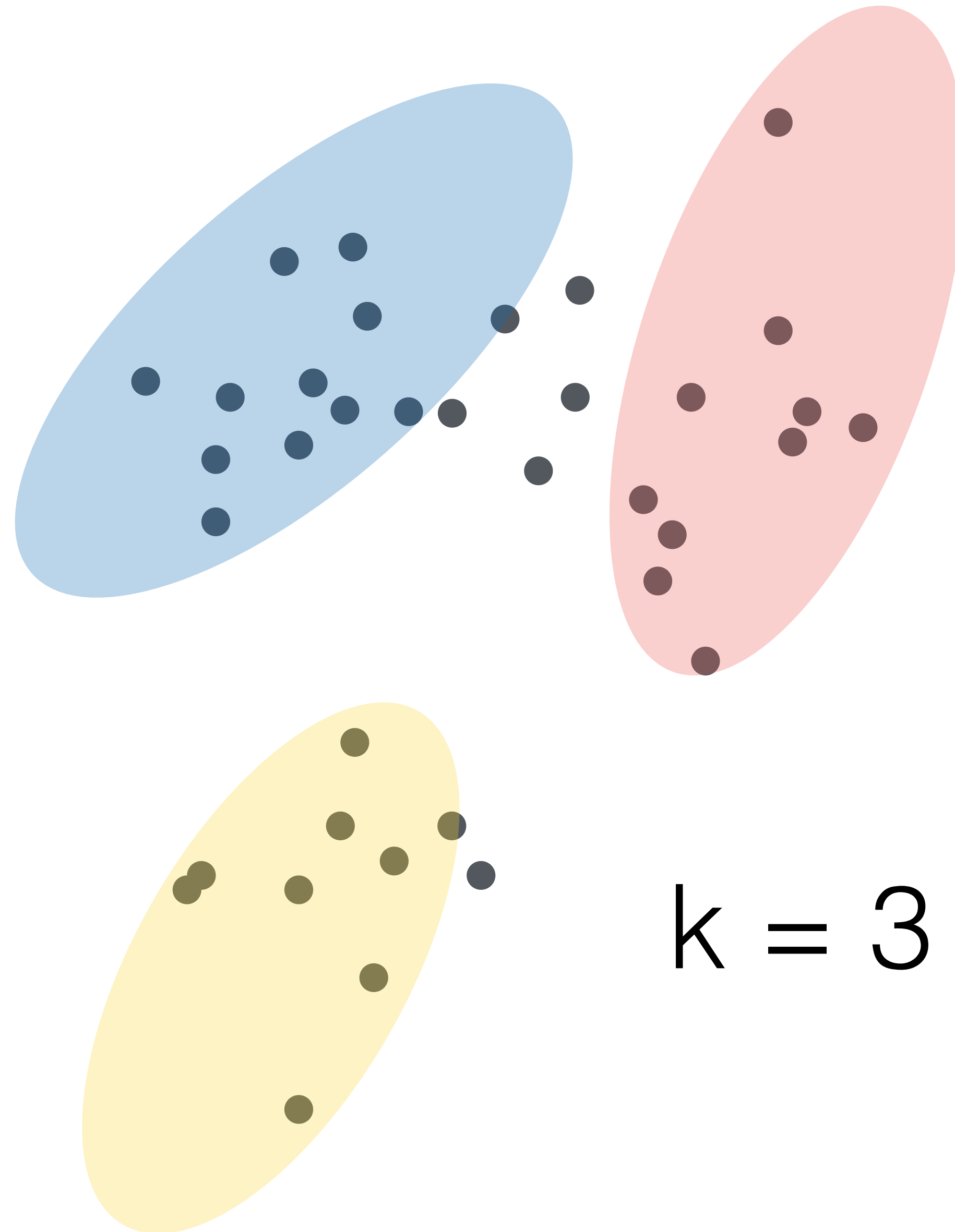
- known variance and mixing coefficients
- unknown number of components  
(**structure learning**)



# toy setting

## Mixture of Gaussians

- known variance and mixing coefficients
- unknown number of components (**structure learning**)
- unknown means (**parameter learning**)



unconstrained learner

$$P(x|\text{all episodes})$$

semantic learner

$$P(x|\text{sufficient statistics})$$

episodic learner

$$P(x|\text{sufficient statistics}) + \sum_{\text{EM}} \delta(\text{episode}_i)$$

unconstrained learner

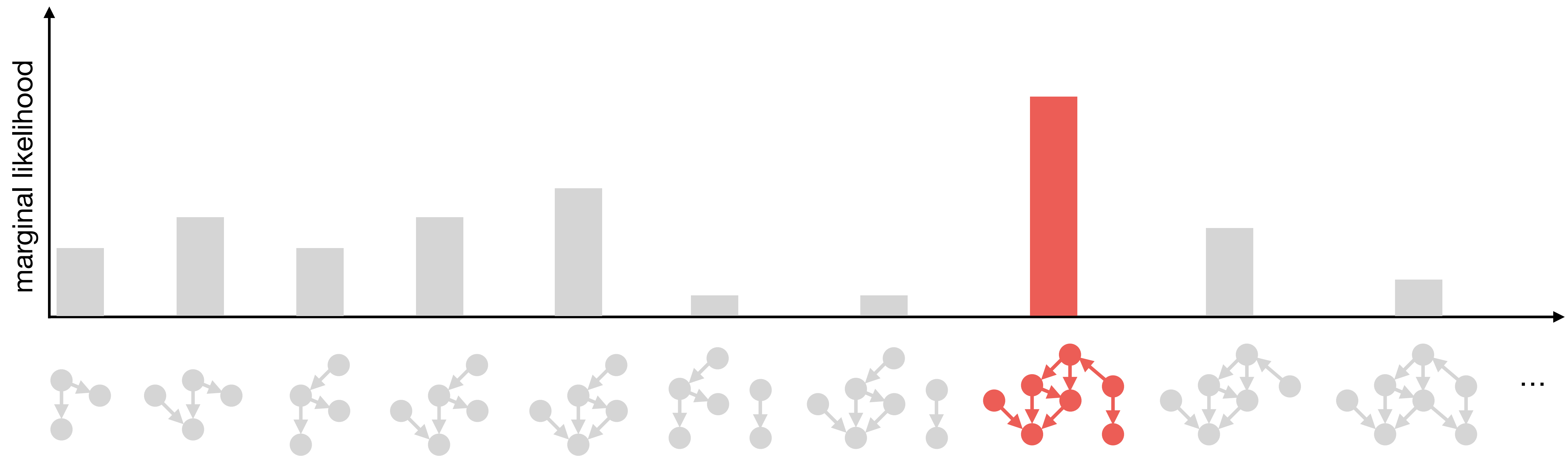
$$P(x|\text{all episodes})$$

semantic learner

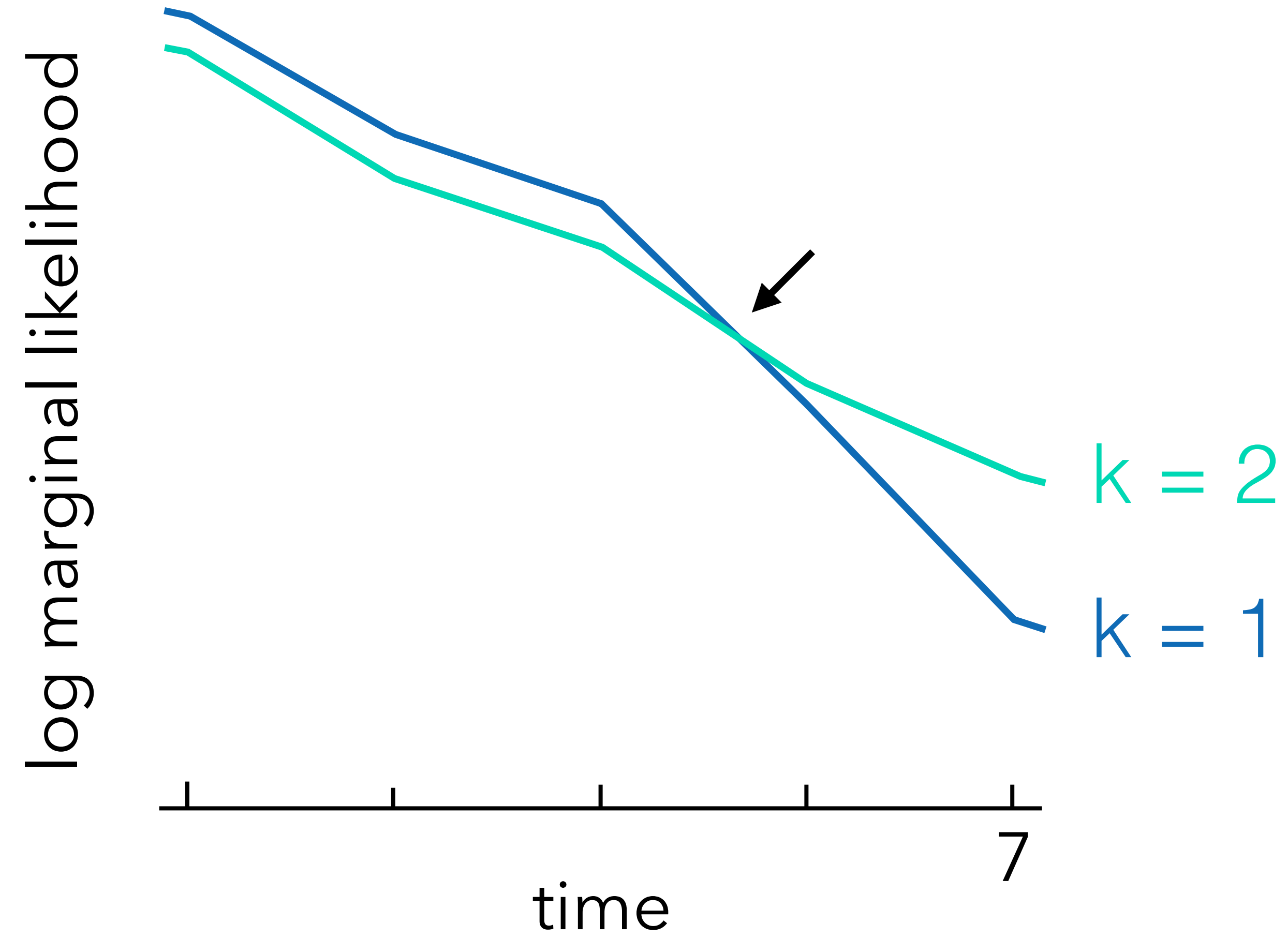
$$P(x|\text{sufficient statistics})$$

episodic learner

$$P(x|\text{sufficient statistics}) + \sum_{\text{EM}} \delta(\text{episode}_i)$$



# unconstrained learner





unconstrained learner

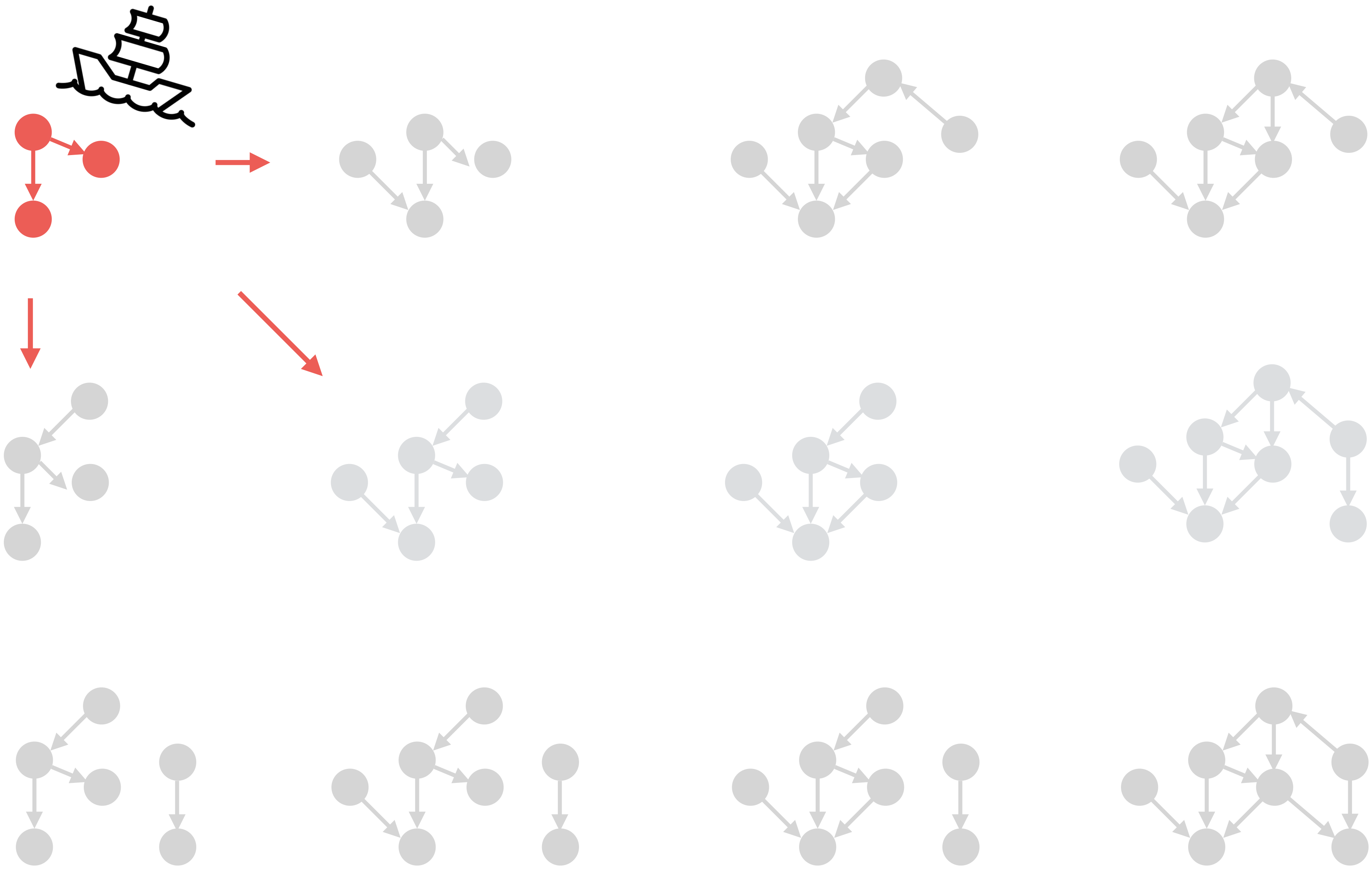
$$P(x|\text{all episodes})$$

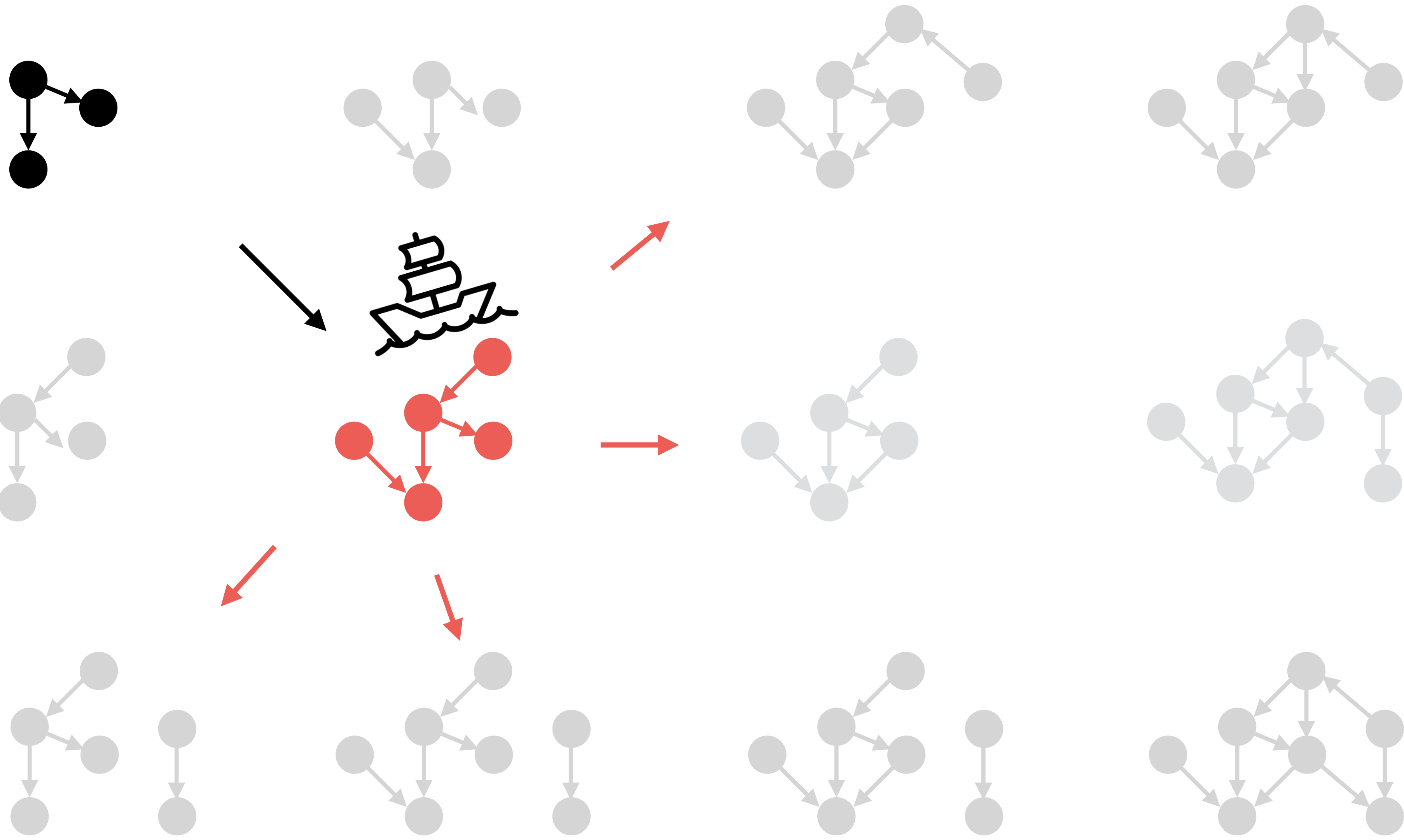
semantic learner

$$P(x|\text{sufficient statistics})$$

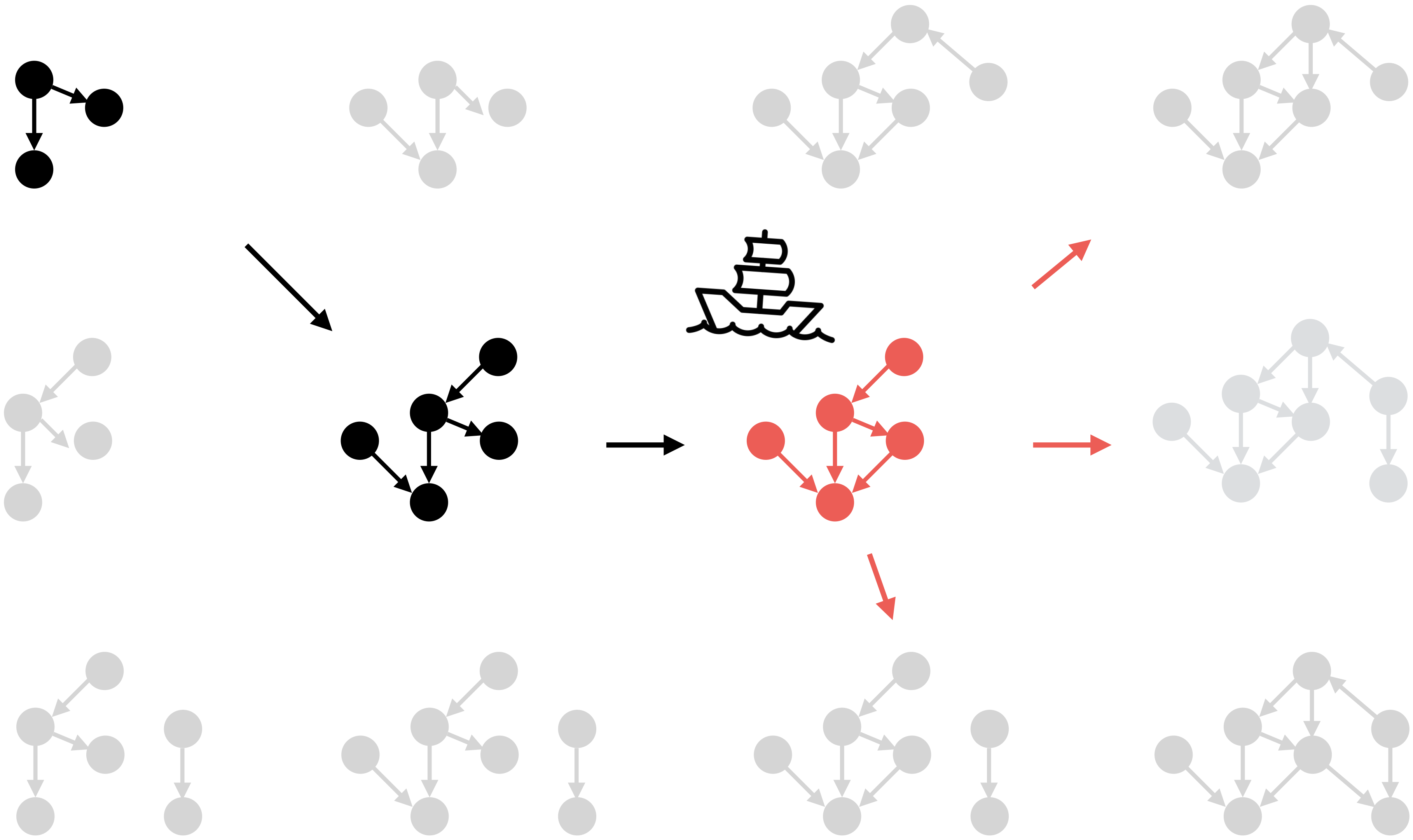
episodic learner

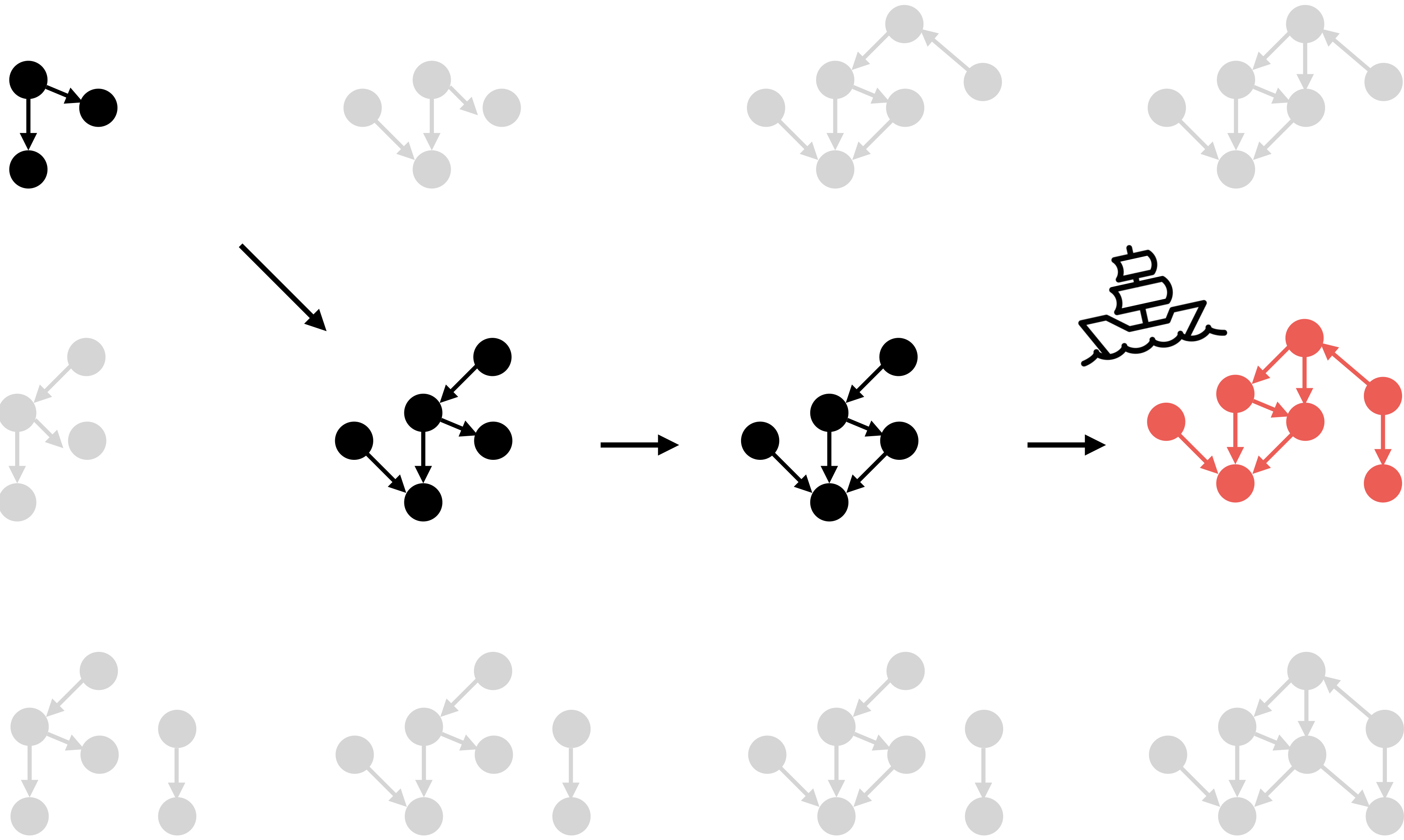
$$P(x|\text{sufficient statistics}) + \sum_{\text{EM}} \delta(\text{episode}_i)$$

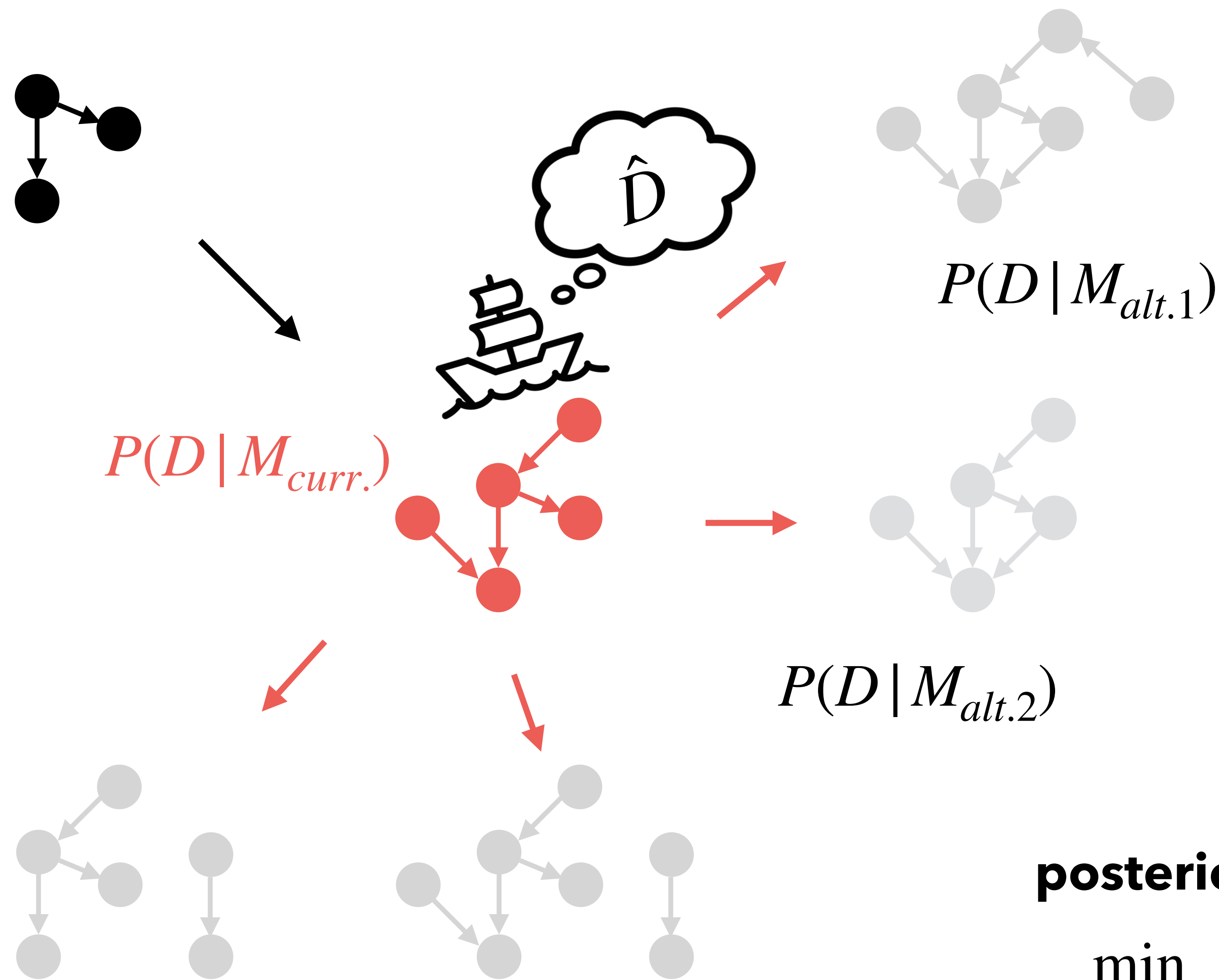




(Bramley et al. 2017, Nagy et al 2016)







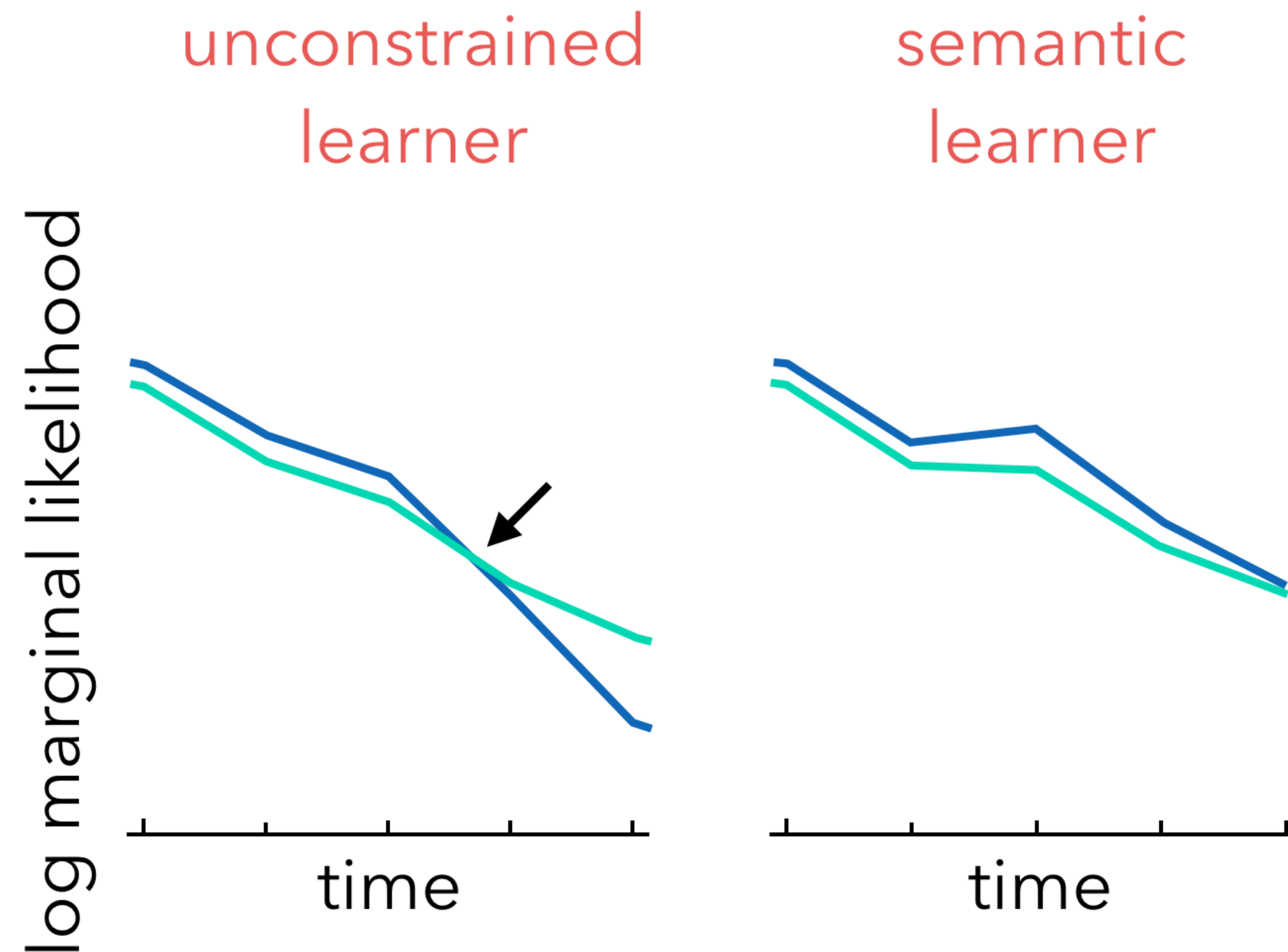
**generative replay**

**marginal likelihood estimate**

$$\mathbb{E}_{\hat{D} \sim P(D | M_{curr.})} [P(\hat{D} | M_i)]$$

**posterior reconstruction**

$$\min_{P(\theta | M_{new}, \eta)} KL[P(x | M_{old}, \hat{D}) || P(x | M_{new}, \eta)]$$



- When the episodes are converted into the posterior for the first model, there are **features of the data that it can't represent**.
- It is often these features that provide the evidence for alternative models
- Since we lose these features at every step, **evidence for alternative models can't accumulate**,
- which introduces a bias towards the current model



unconstrained learner

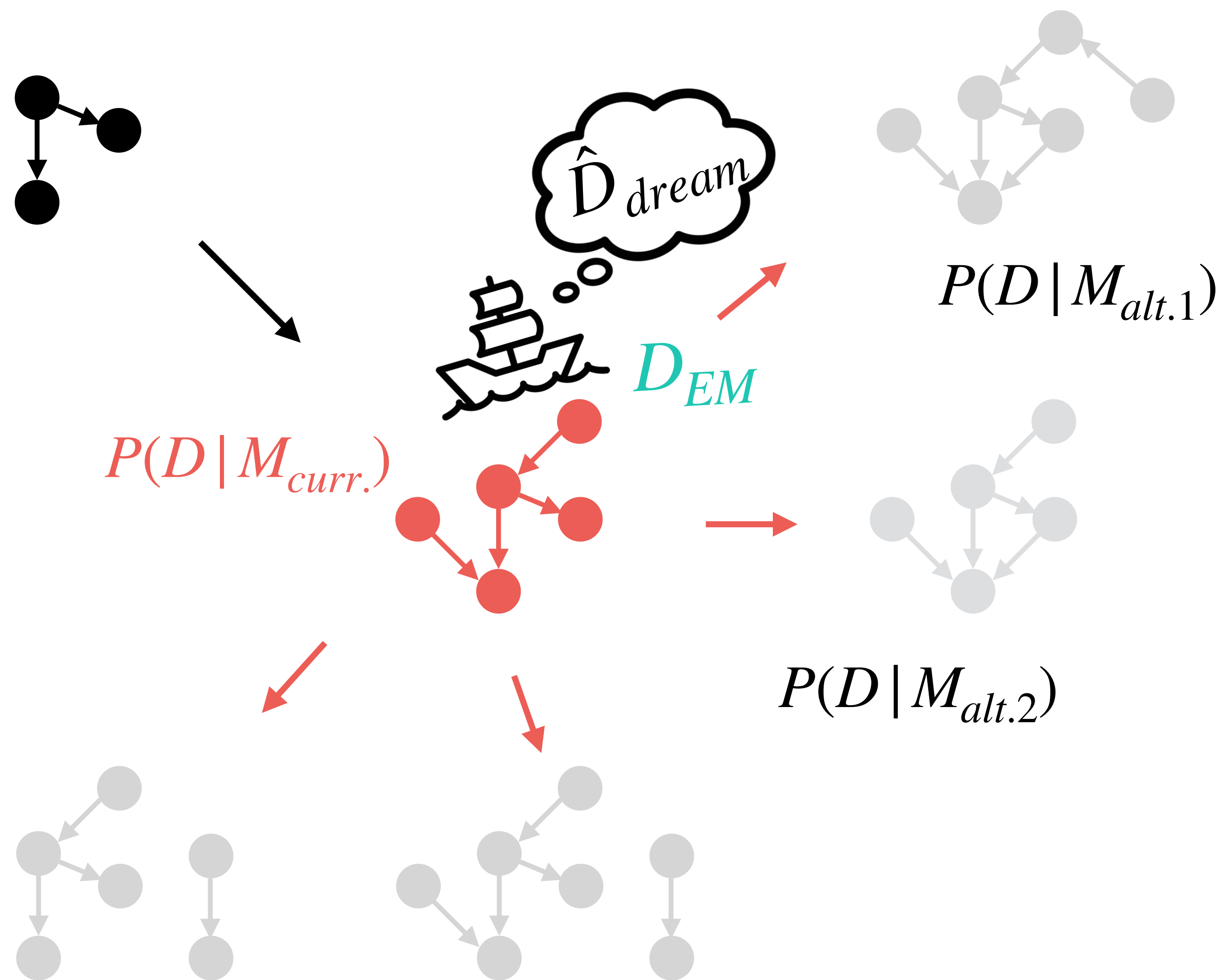
$$P(x|\text{all episodes})$$

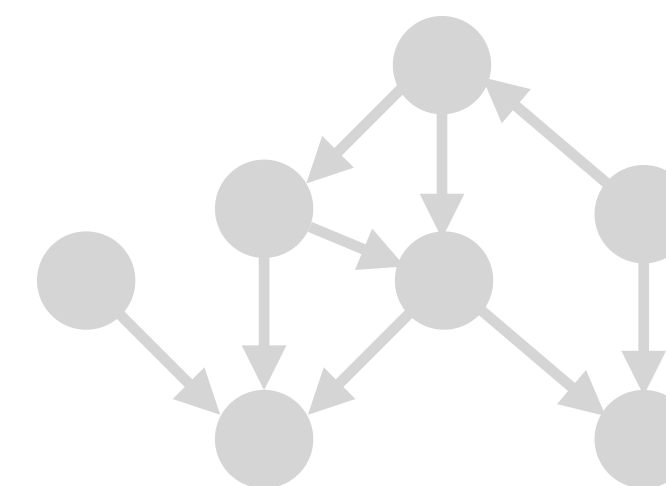
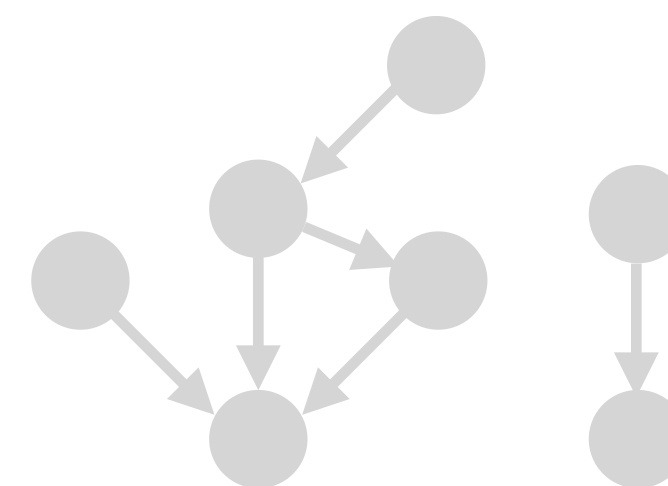
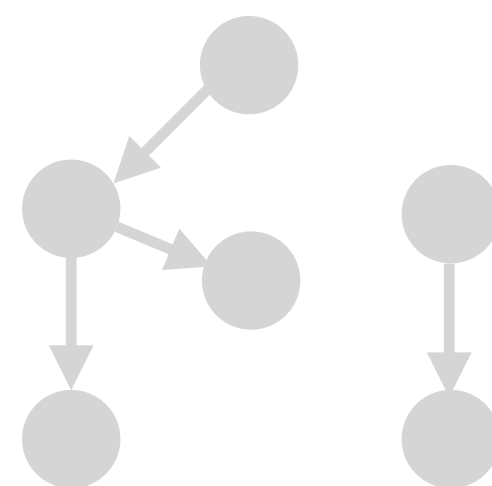
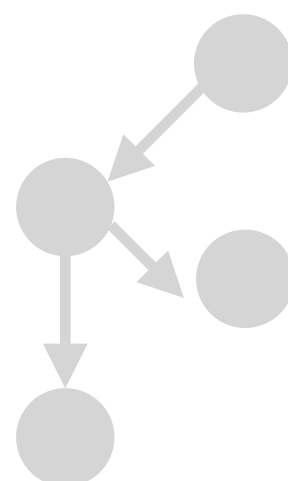
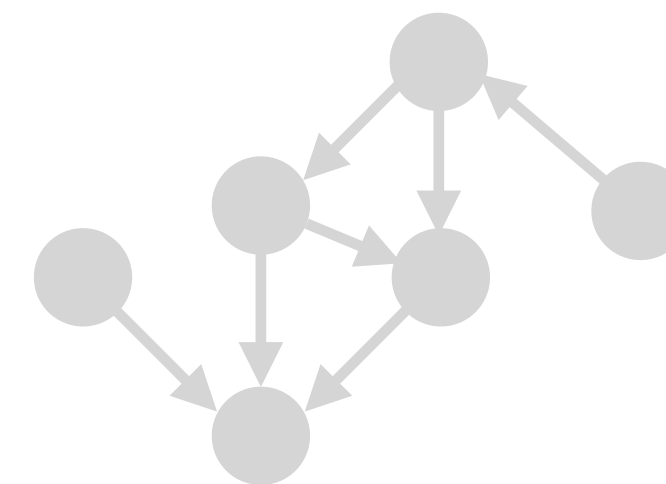
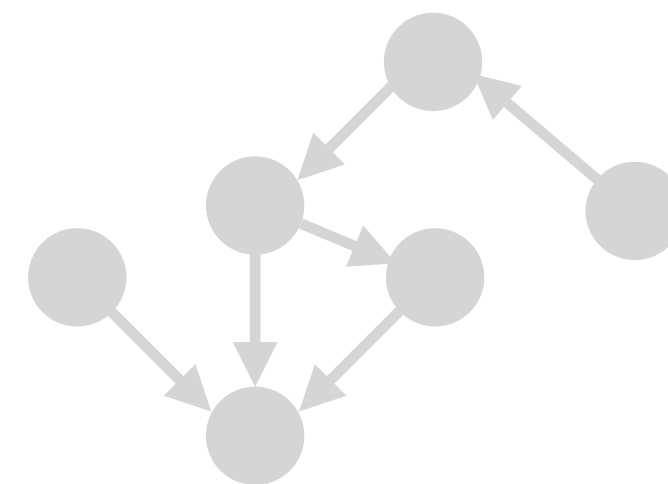
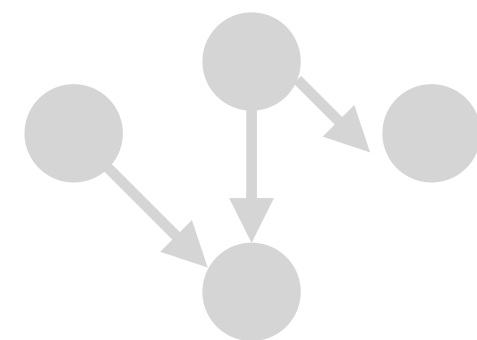
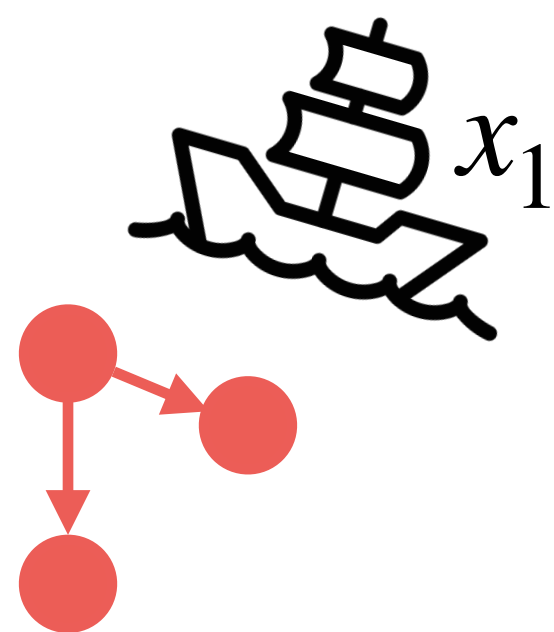
semantic learner

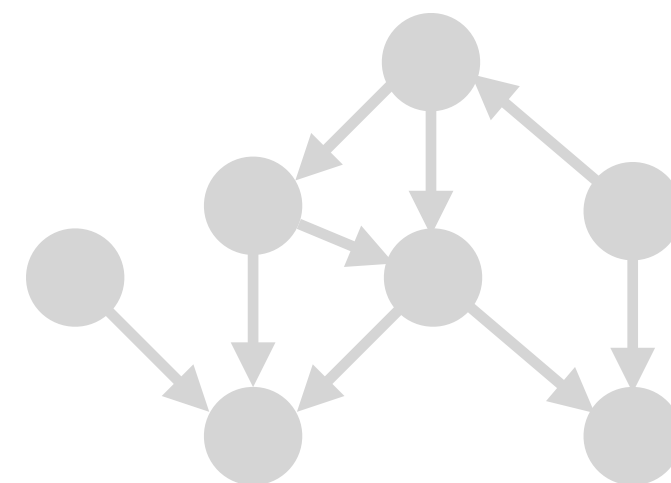
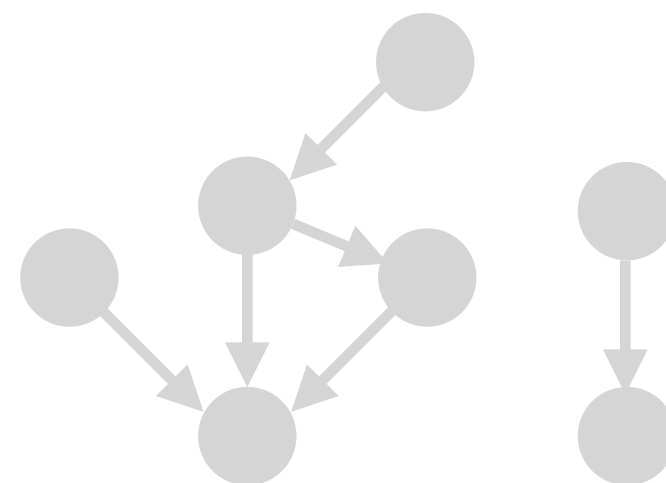
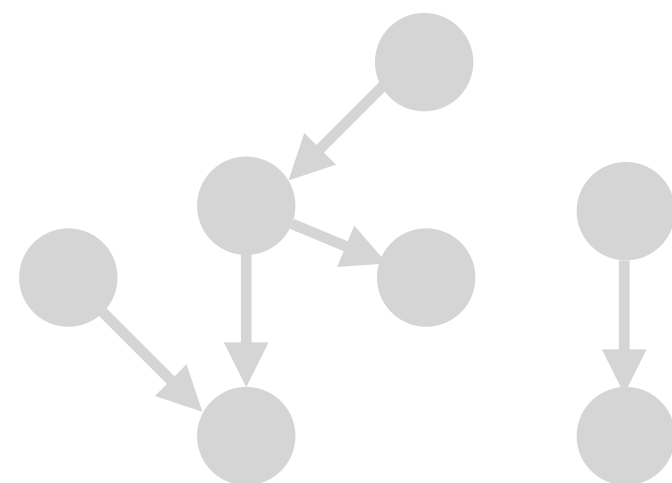
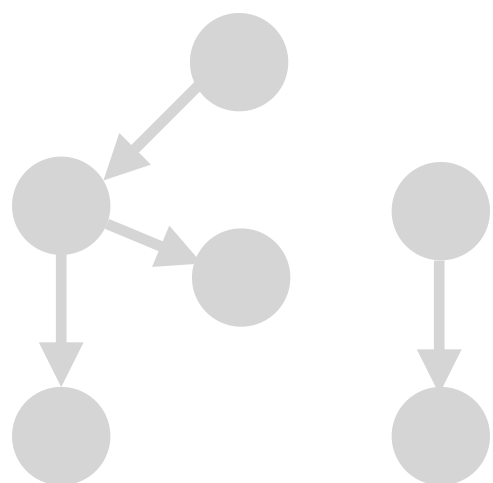
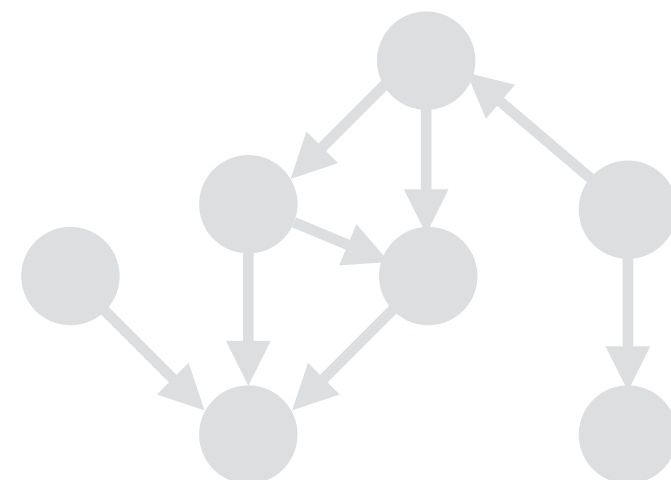
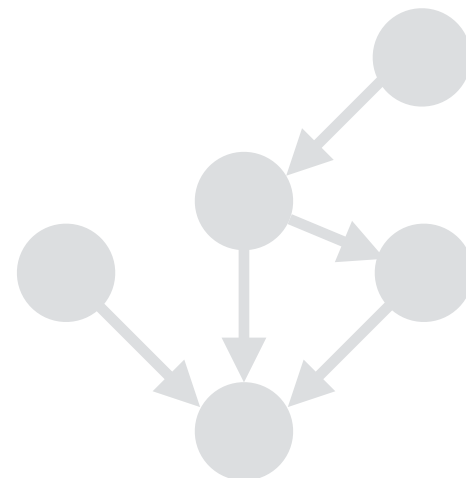
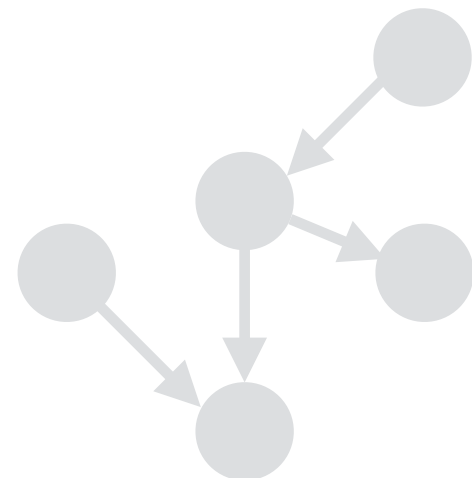
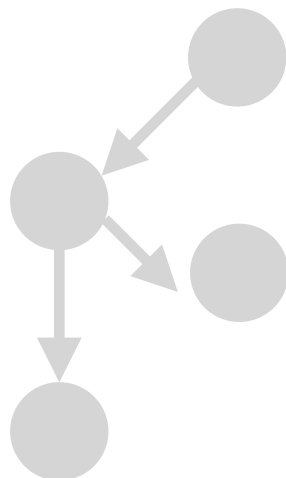
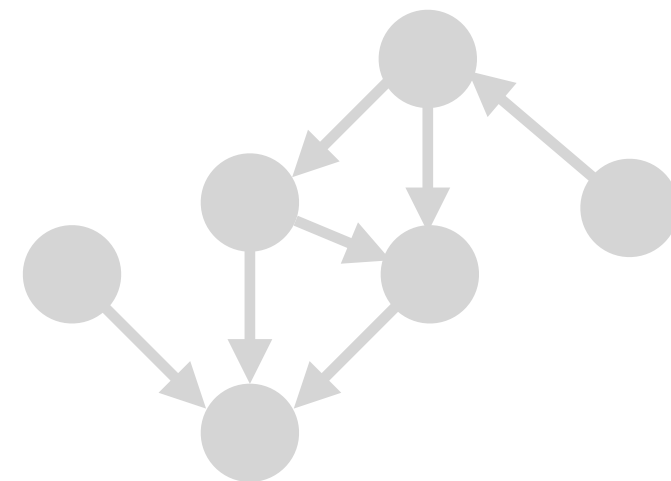
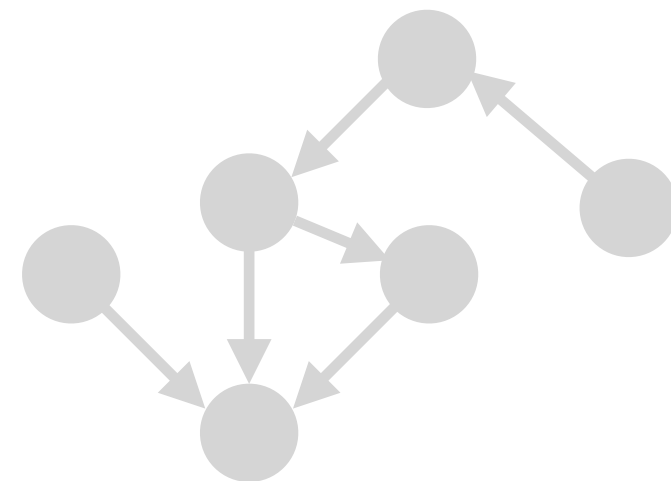
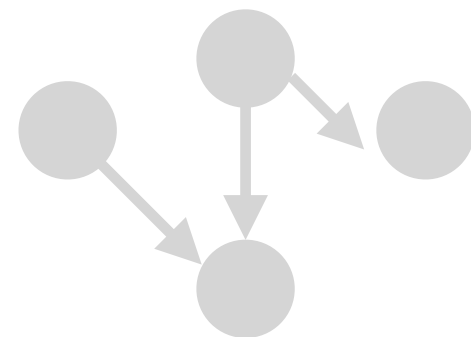
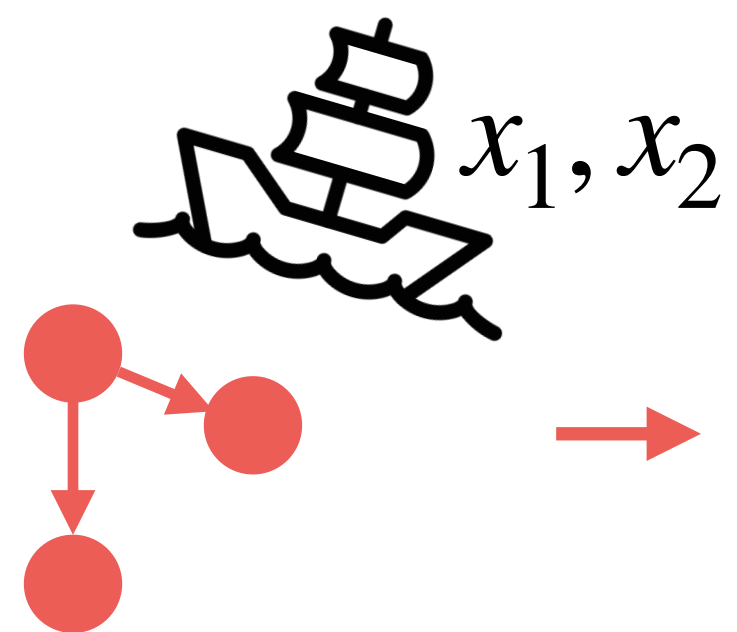
$$P(x|\text{sufficient statistics})$$

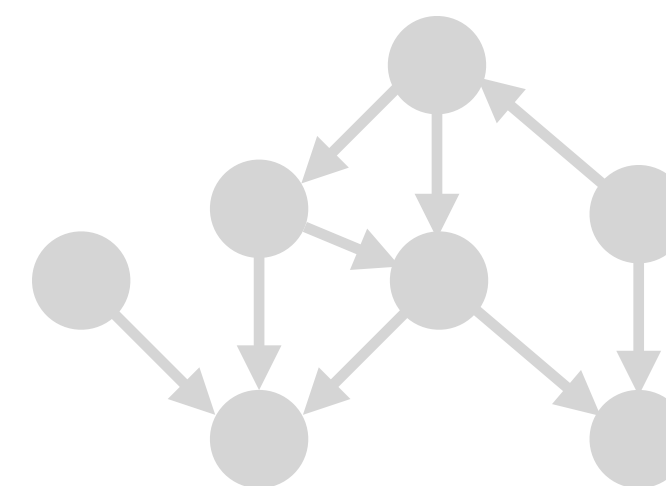
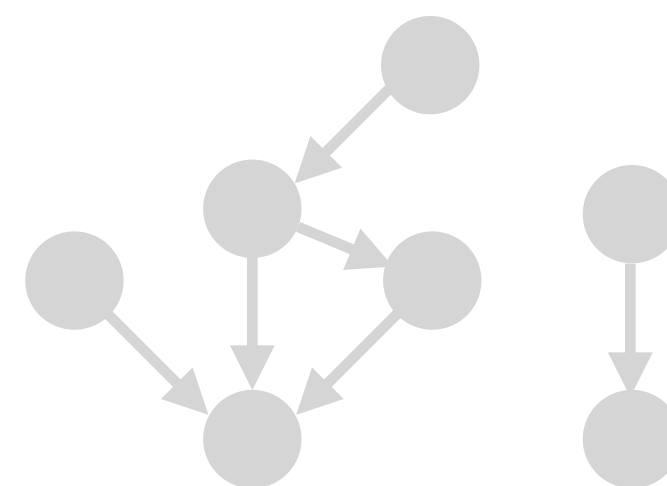
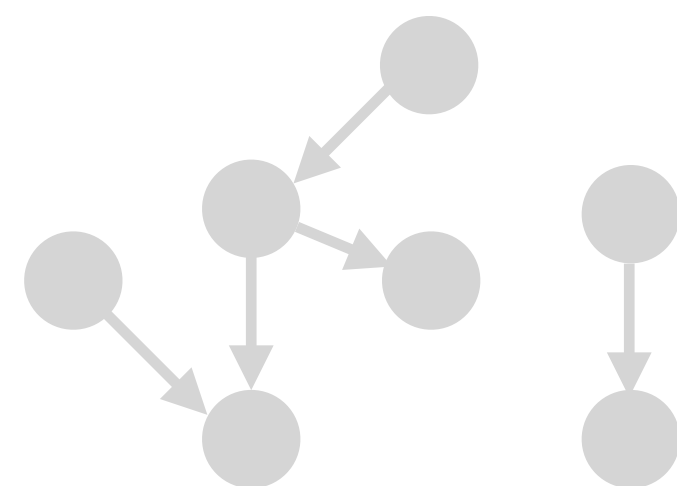
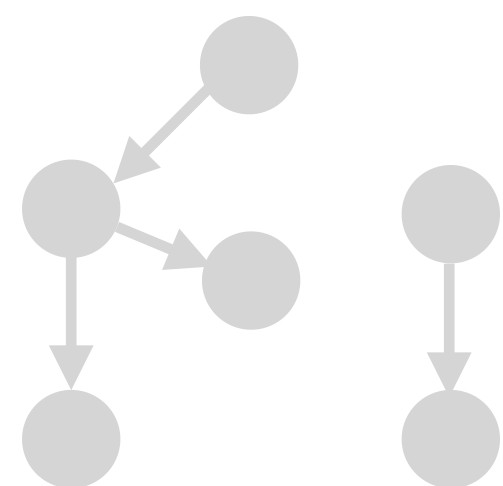
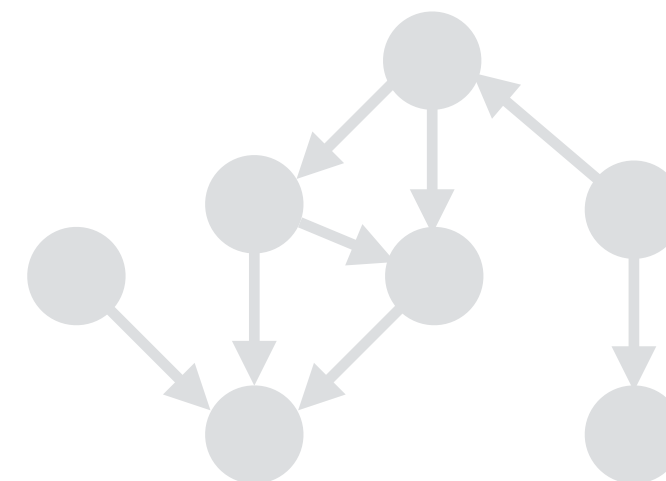
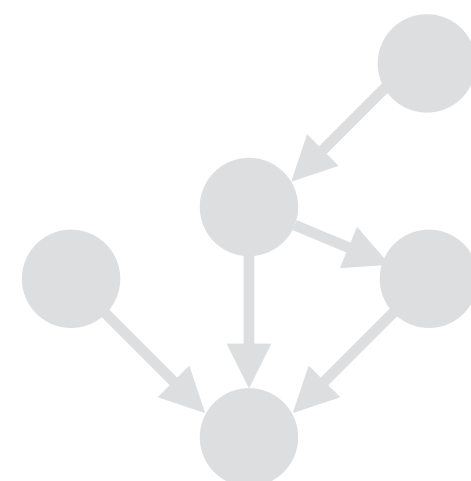
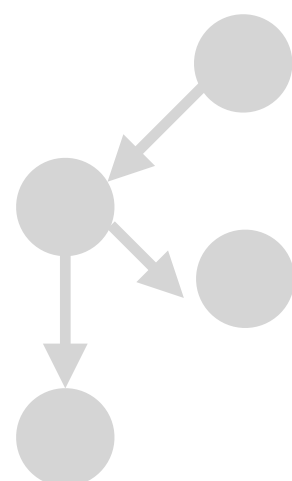
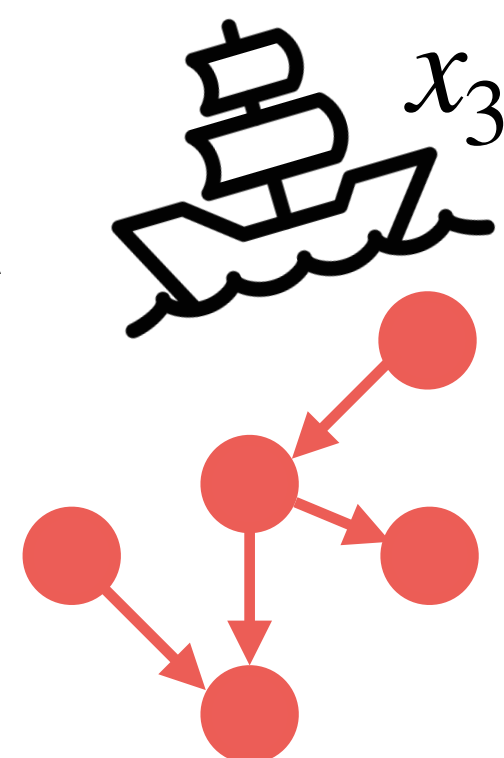
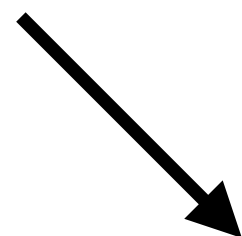
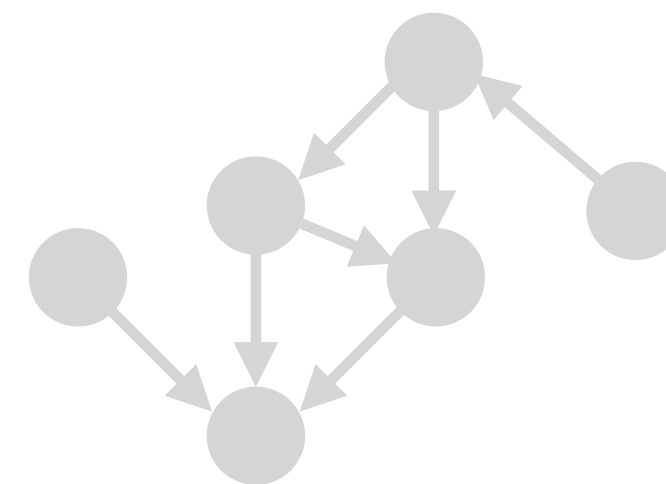
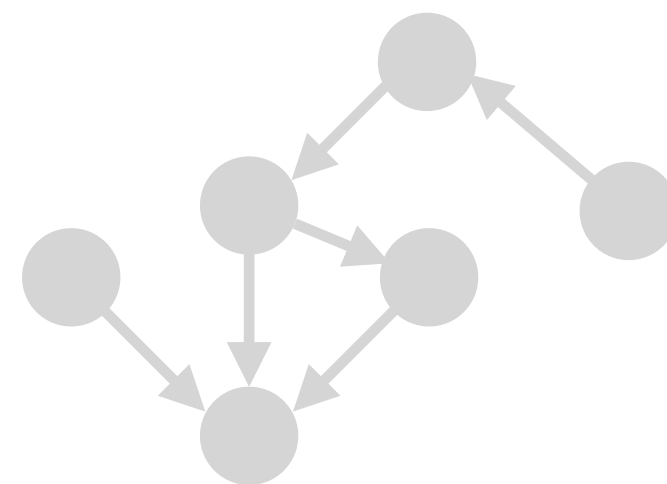
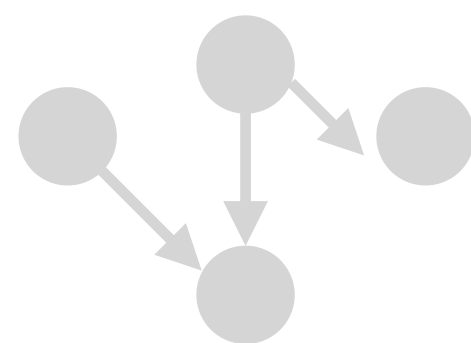
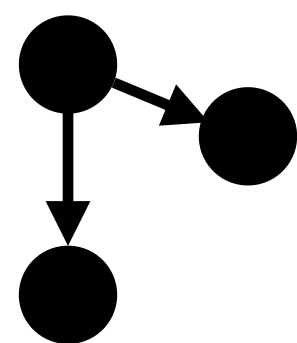
episodic learner

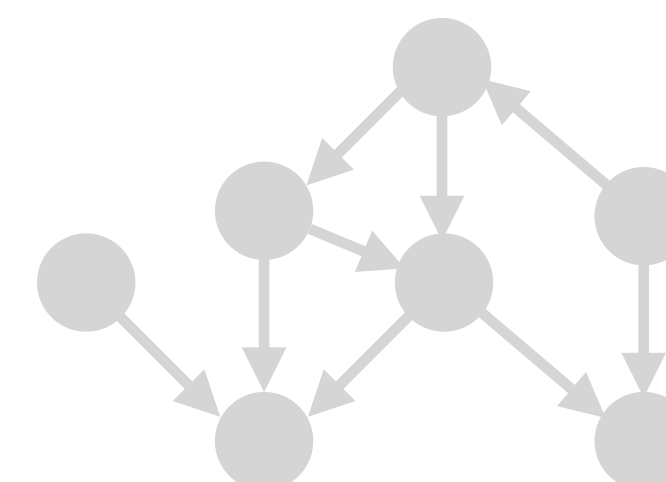
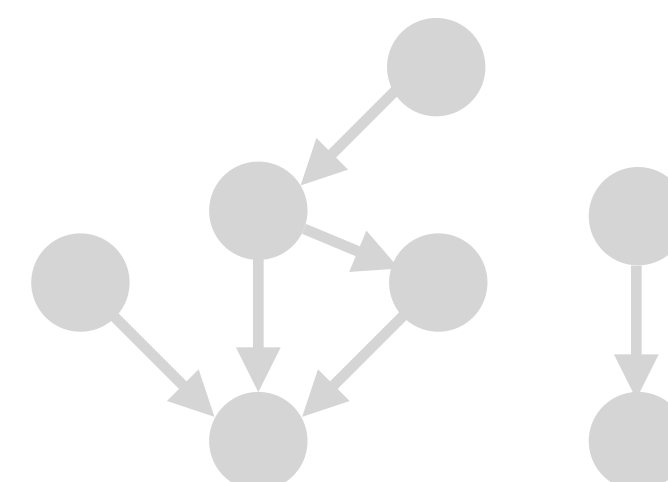
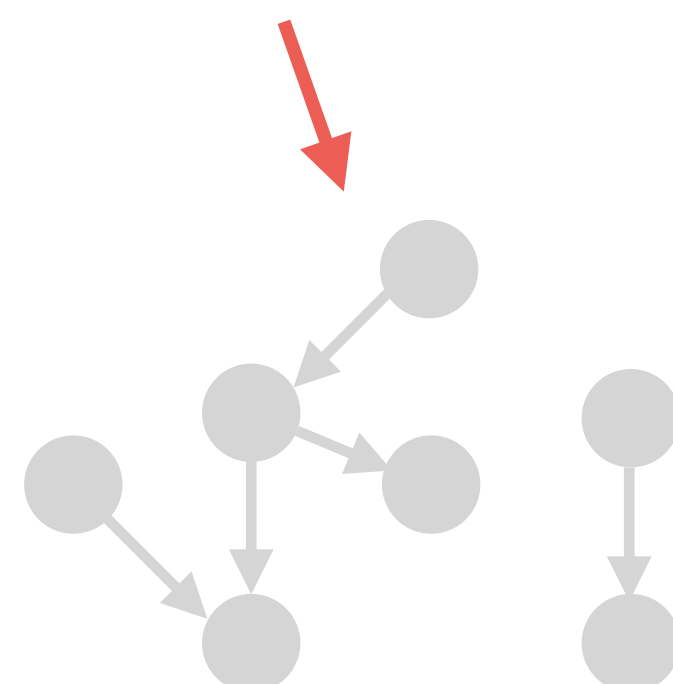
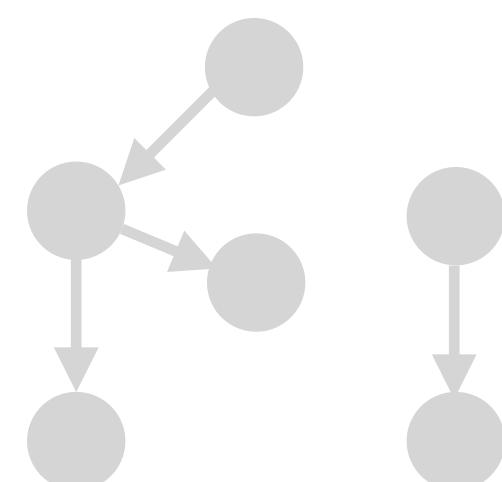
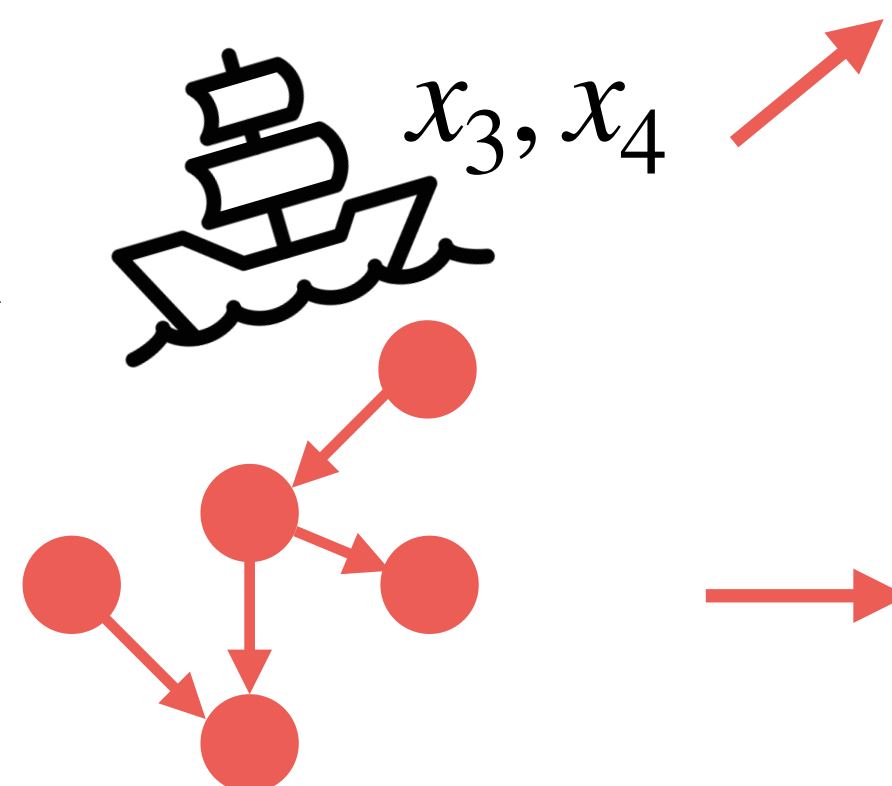
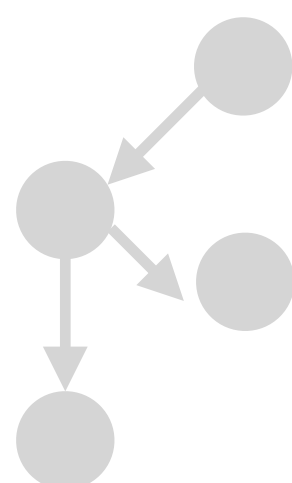
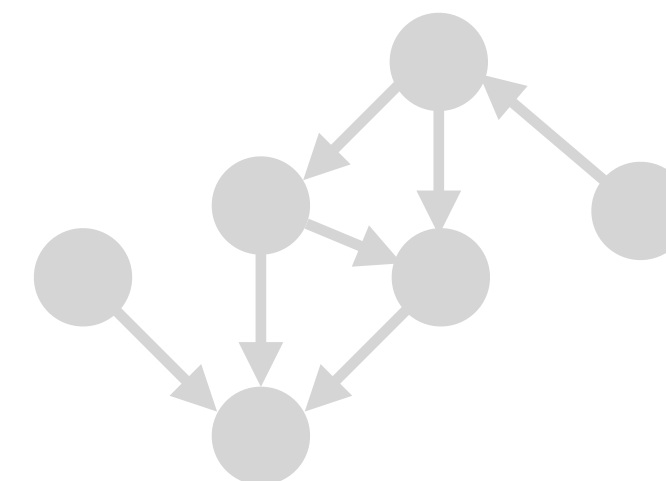
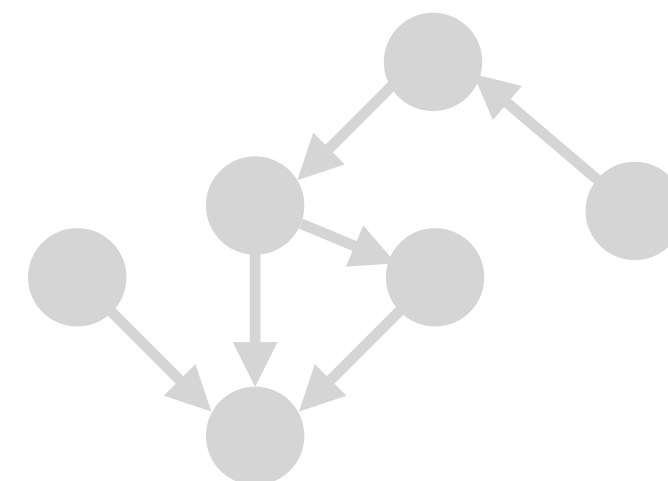
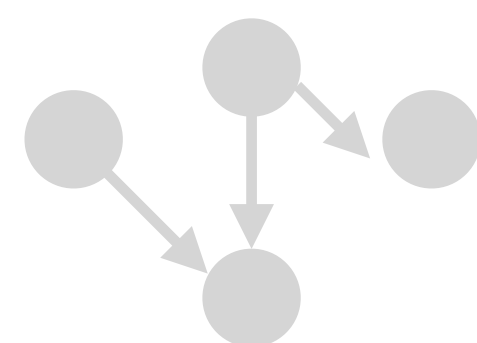
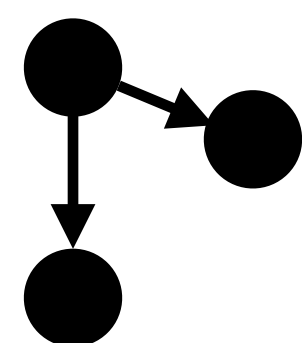
$$P(x|\text{sufficient statistics}) + \sum_{\text{EM}} \delta(\text{episode}_i)$$



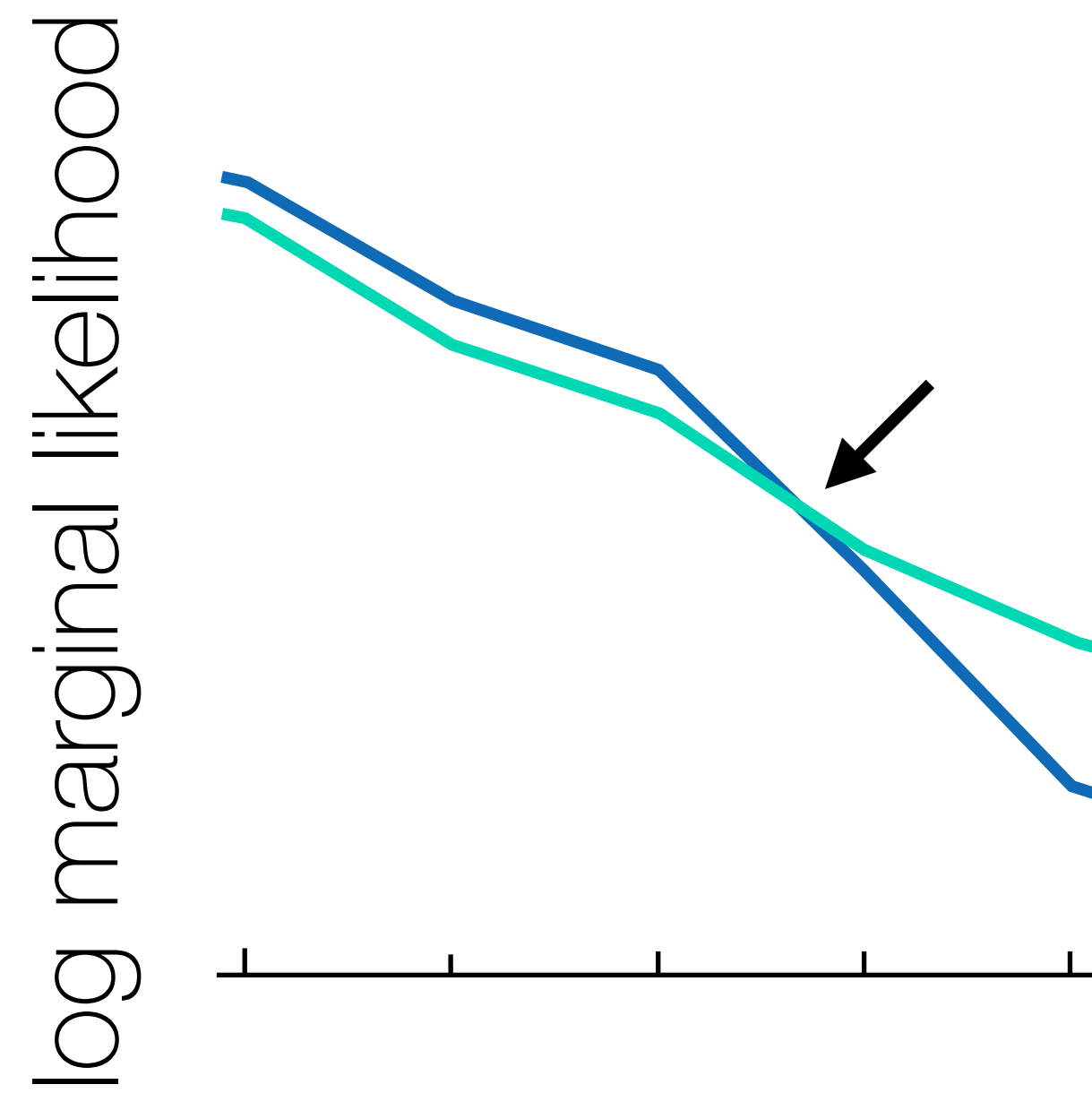




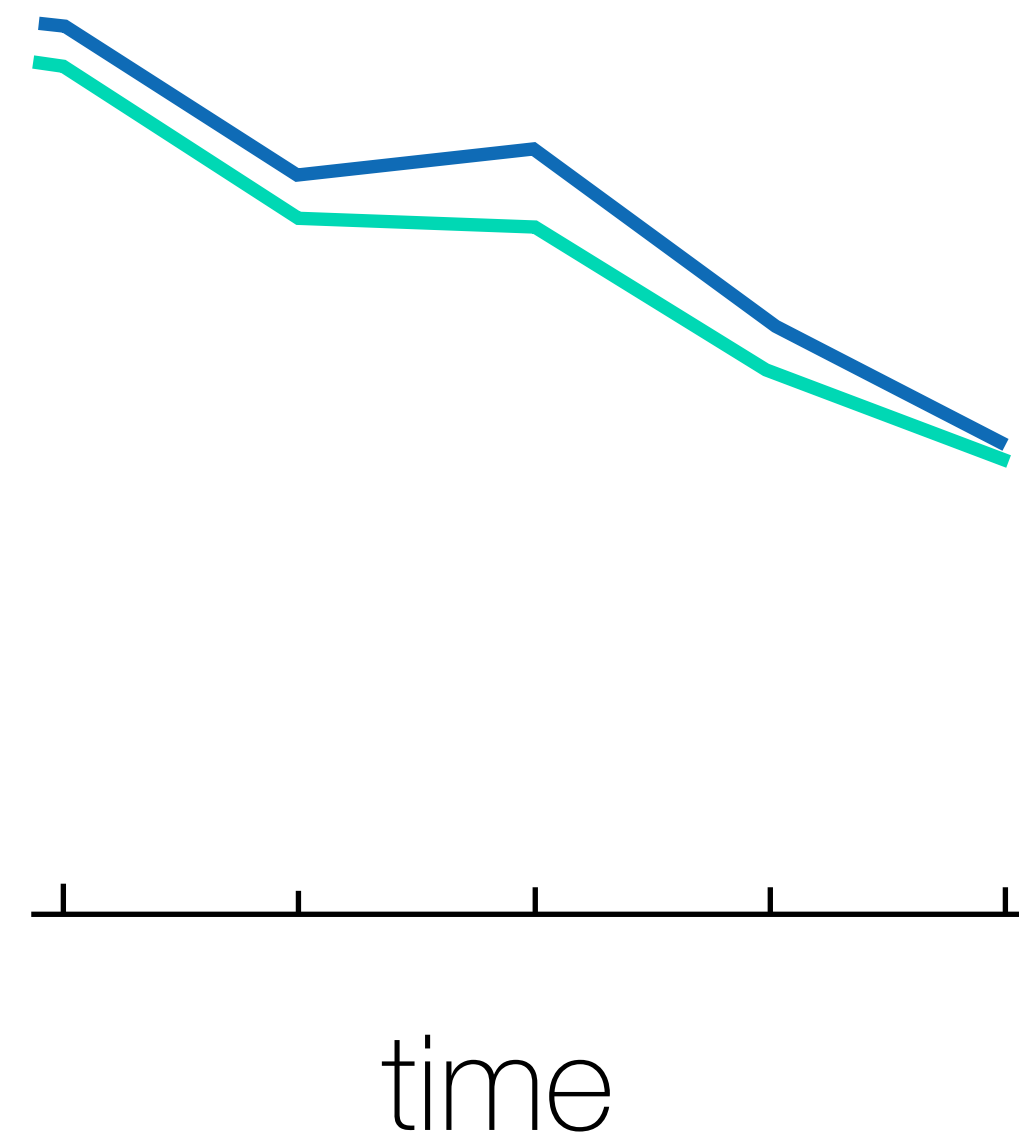




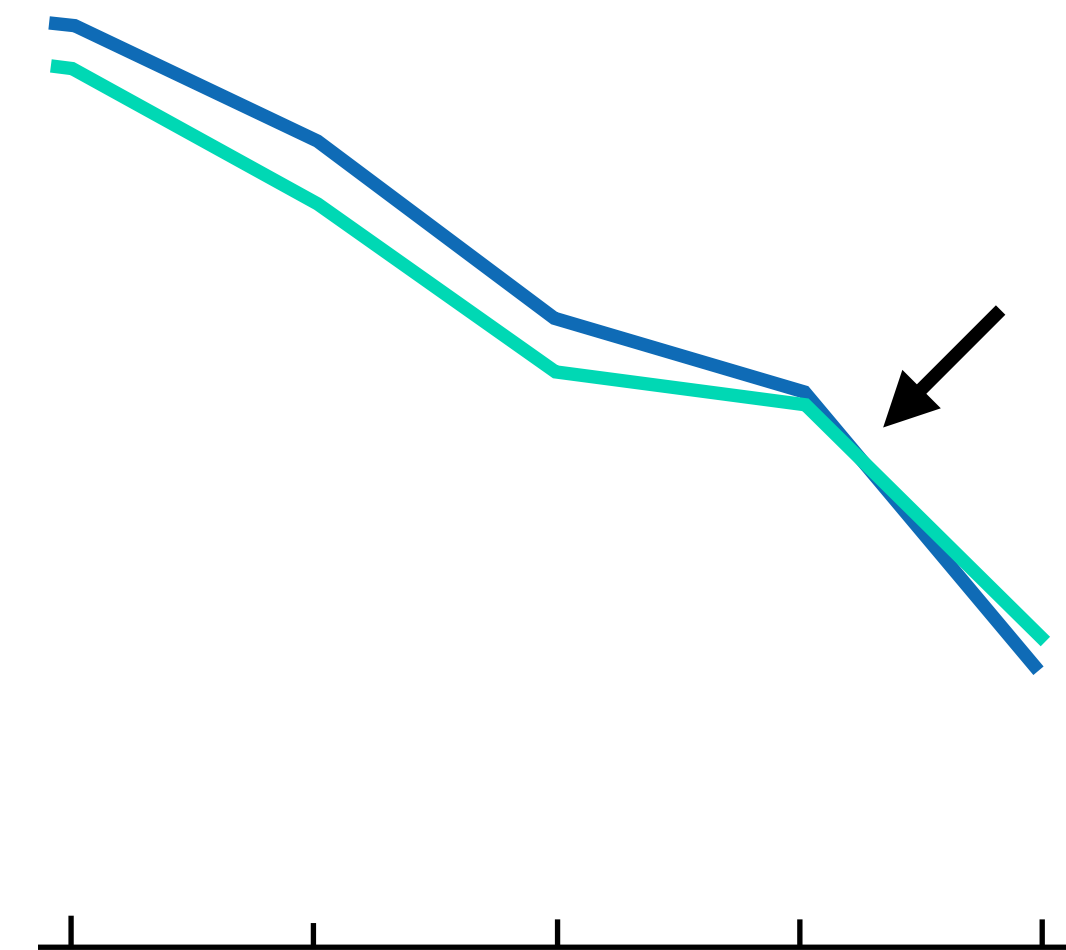
unconstrained  
learner



semantic  
learner



episodic  
learner

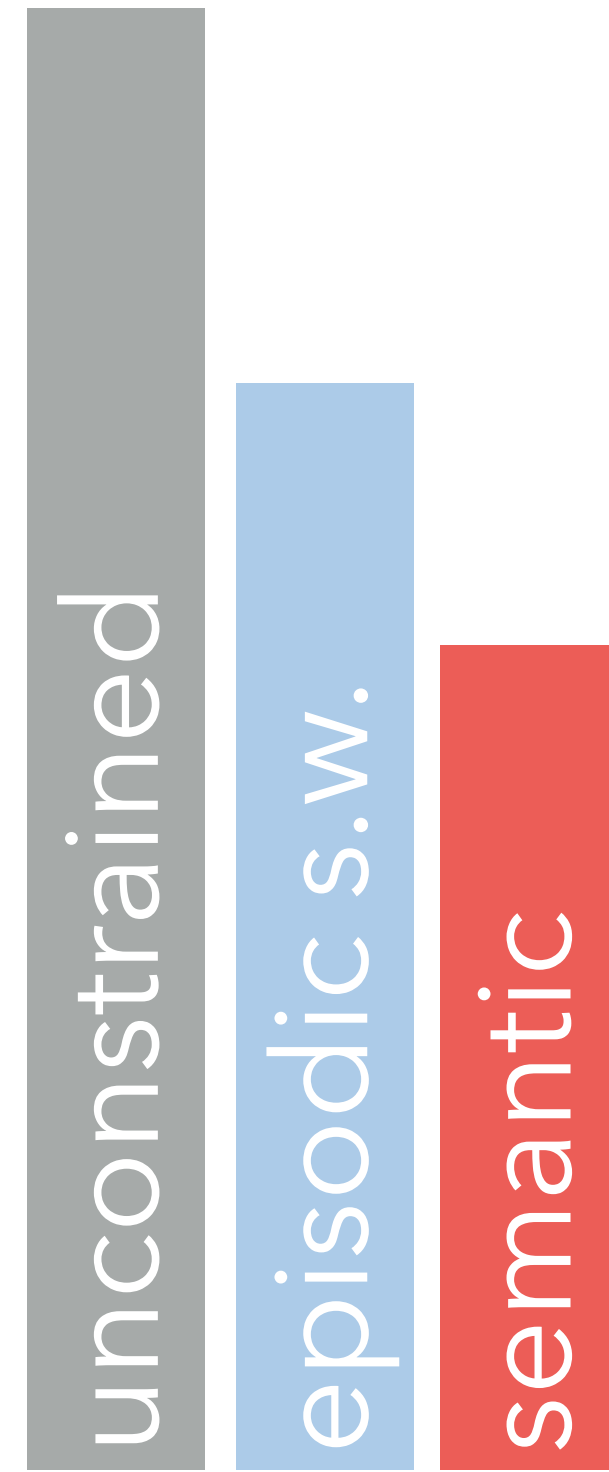




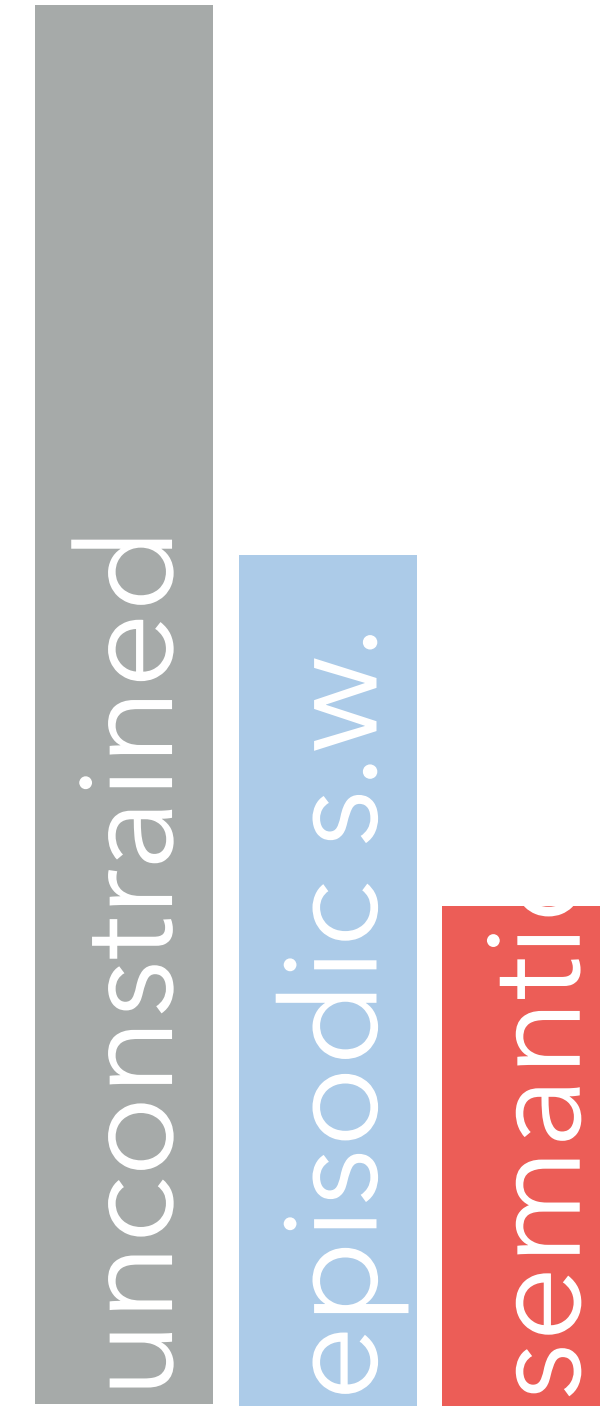
# results

correct structure discovered %

1.0  
0.8  
0.6  
0.4  
0.2  
0



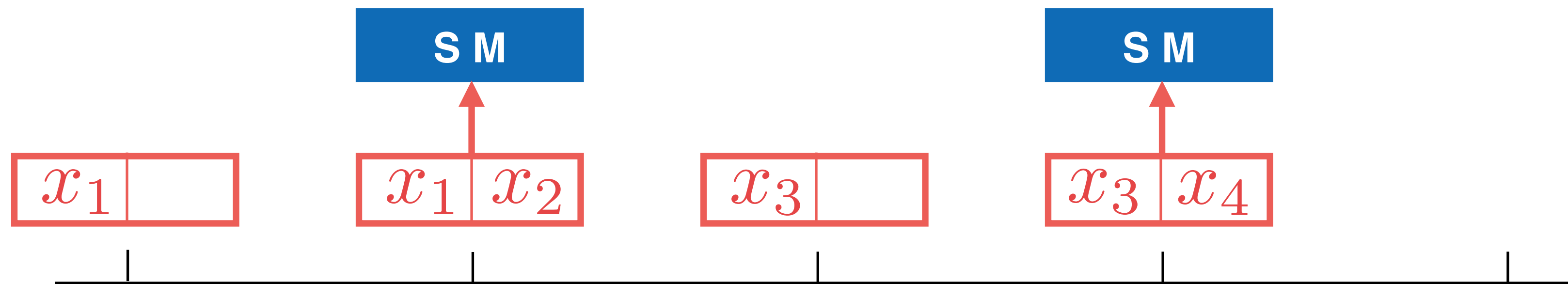
true k=2



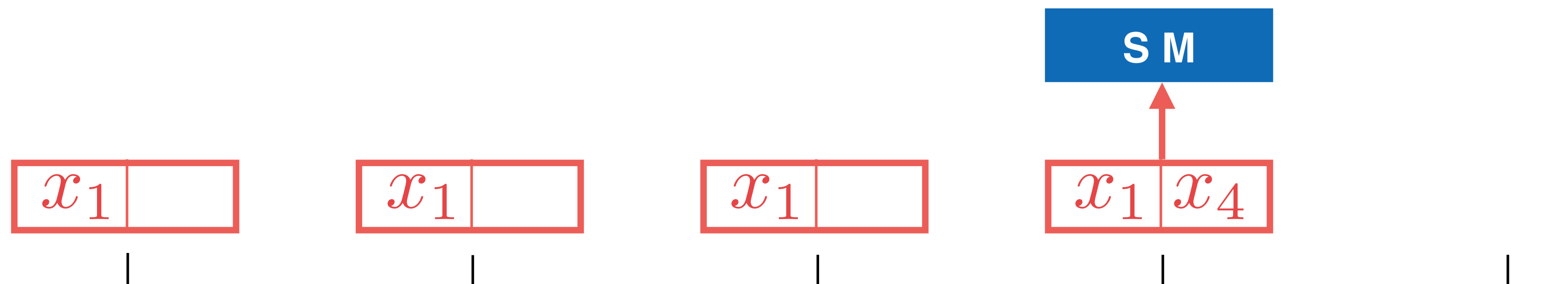
true k=3

# episodic learner

## sliding window



## selective memory



# selective memory

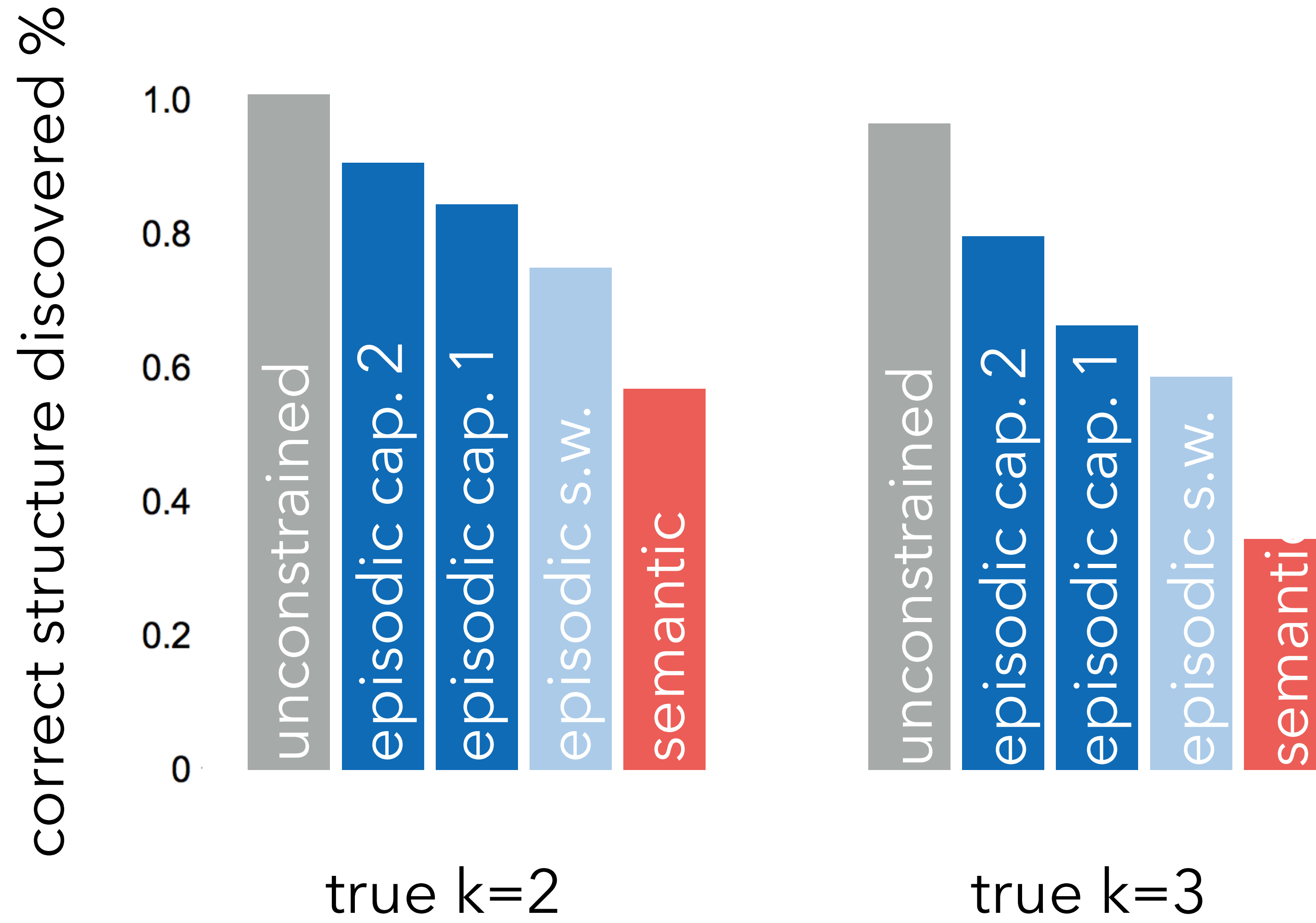
- what points are most informative about the model/parameters?
- points that can be easily summarised in a distribution should be stored in semantic memory, outliers should be stored separately

**surprise:**  $D_{KL}(\text{new posterior} || \text{posterior})$

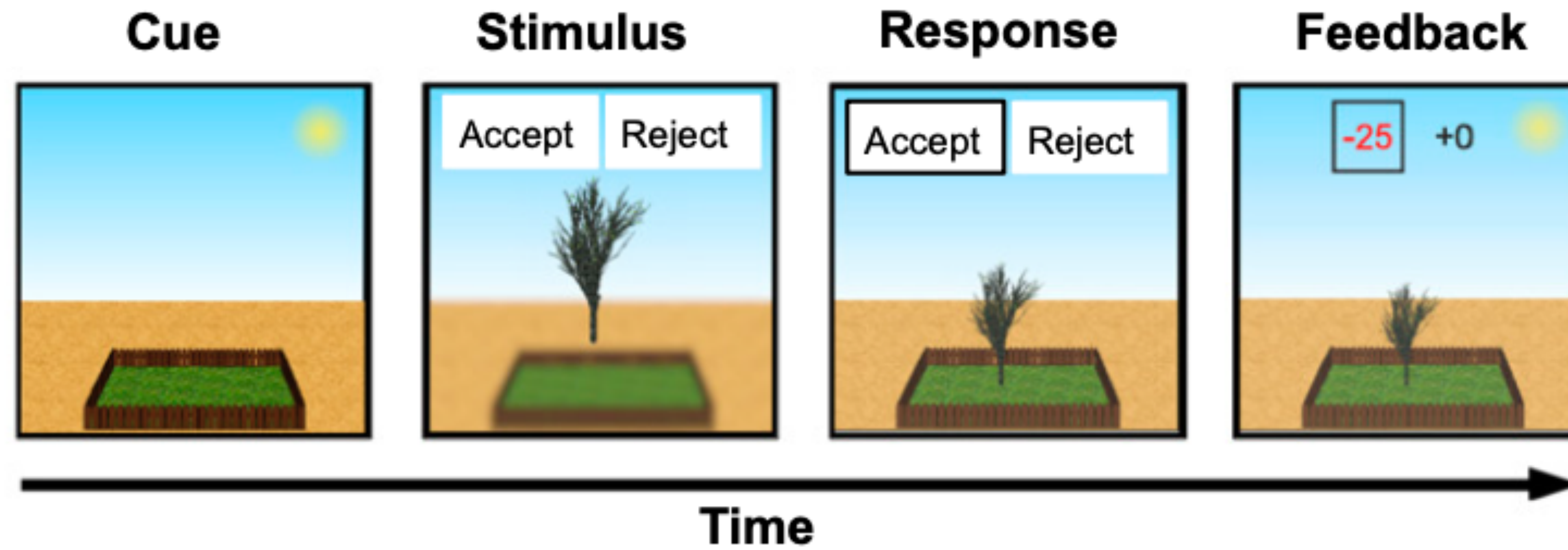
- a 10 hour drive on the highway may be stored as a summary
- while the first day at school should be remembered in detail



# results

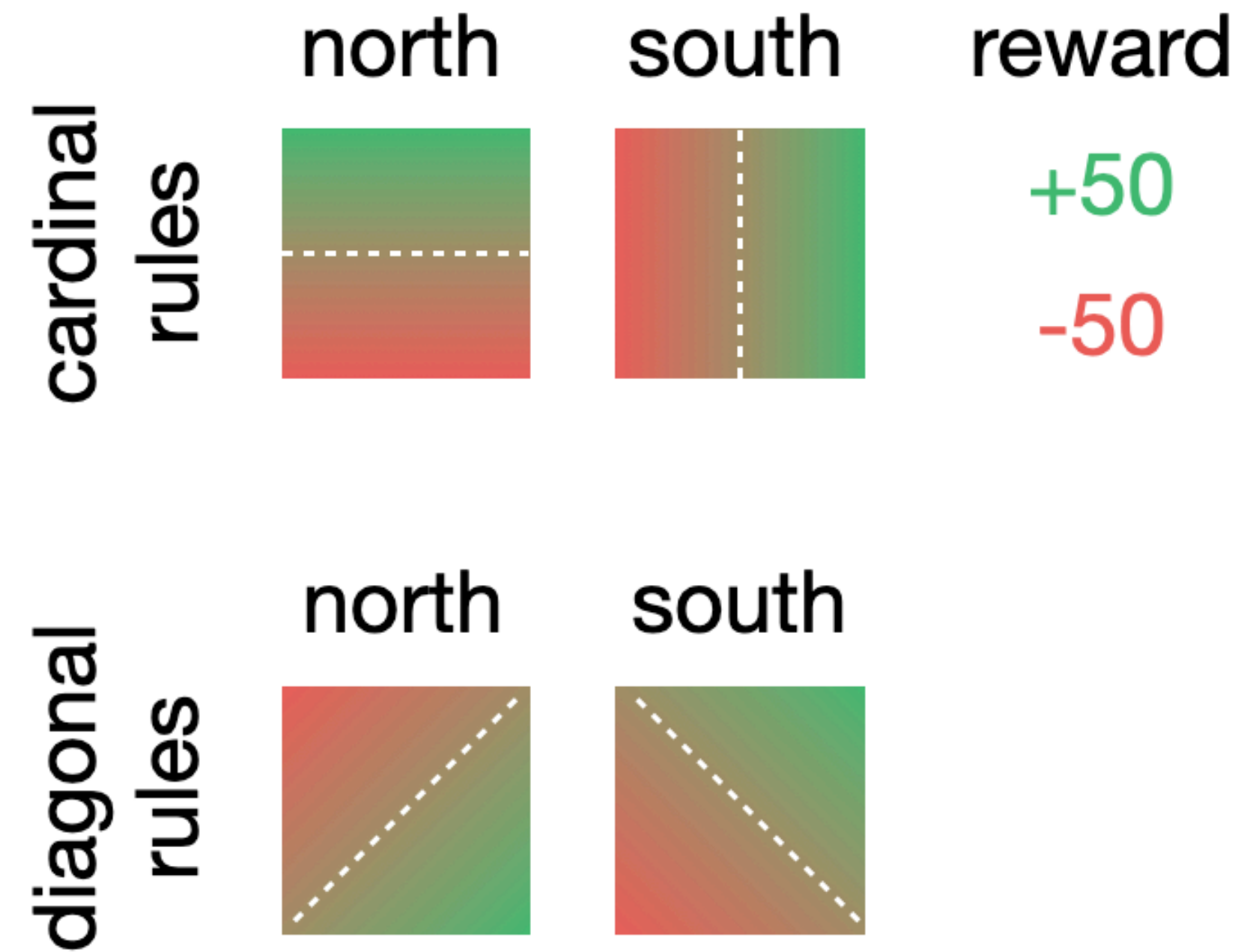
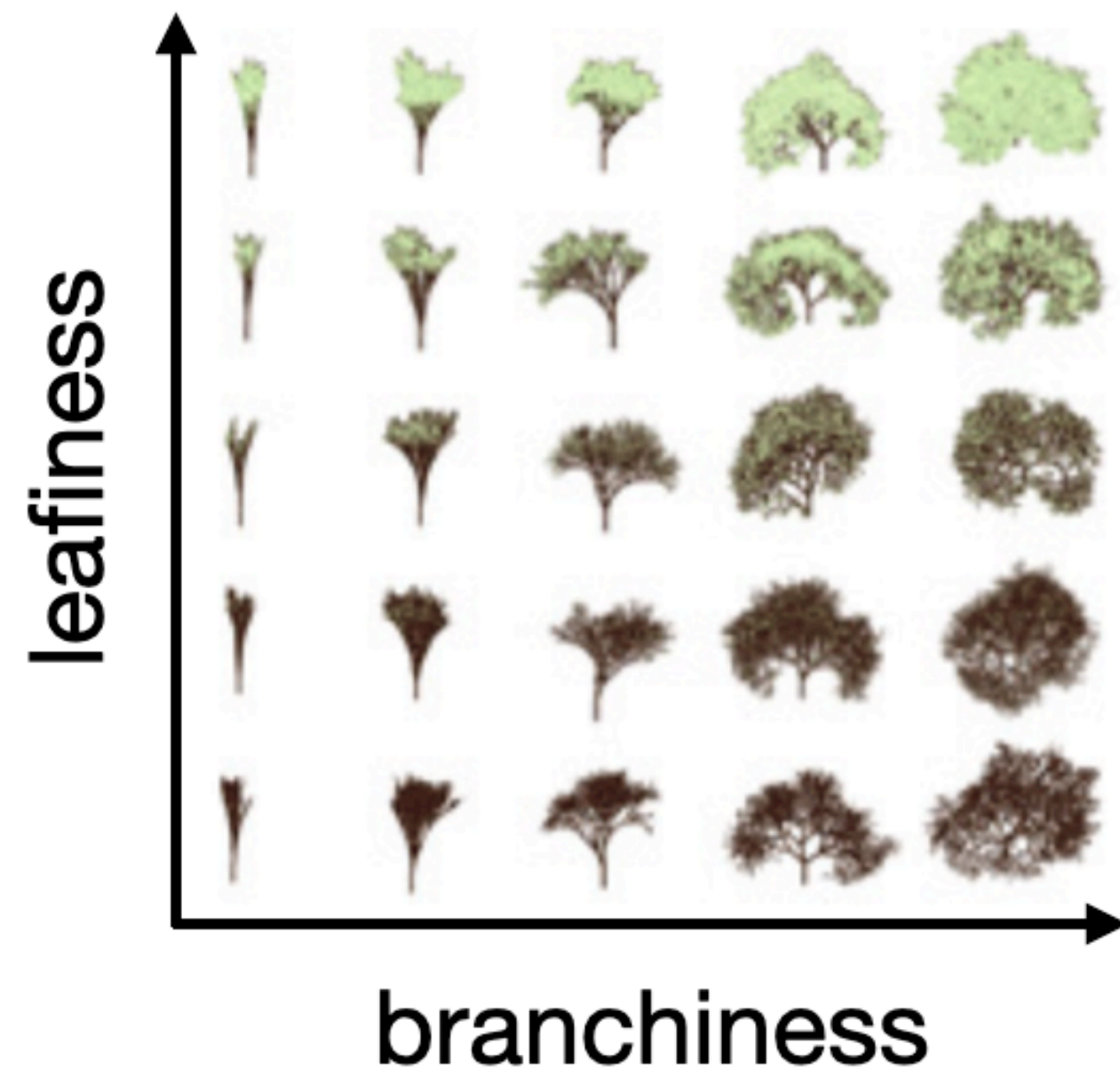


# tree task

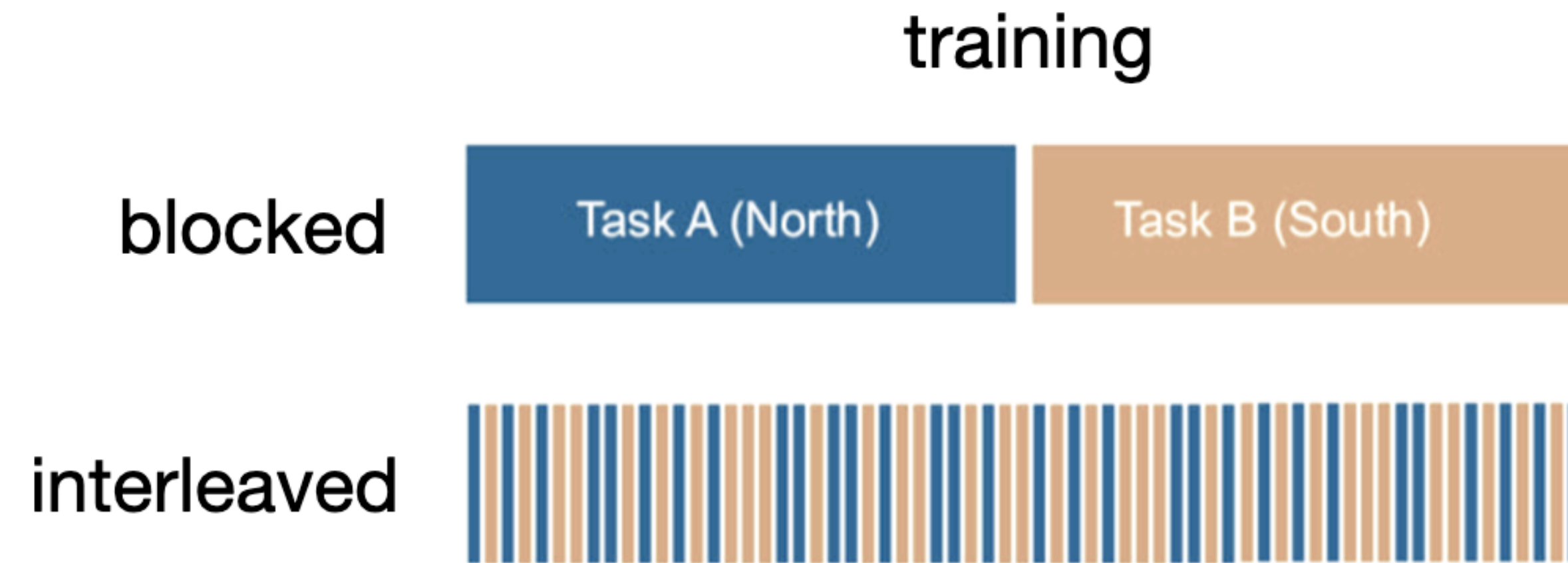




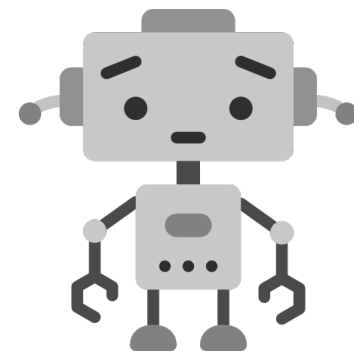
# tree task



# tree task



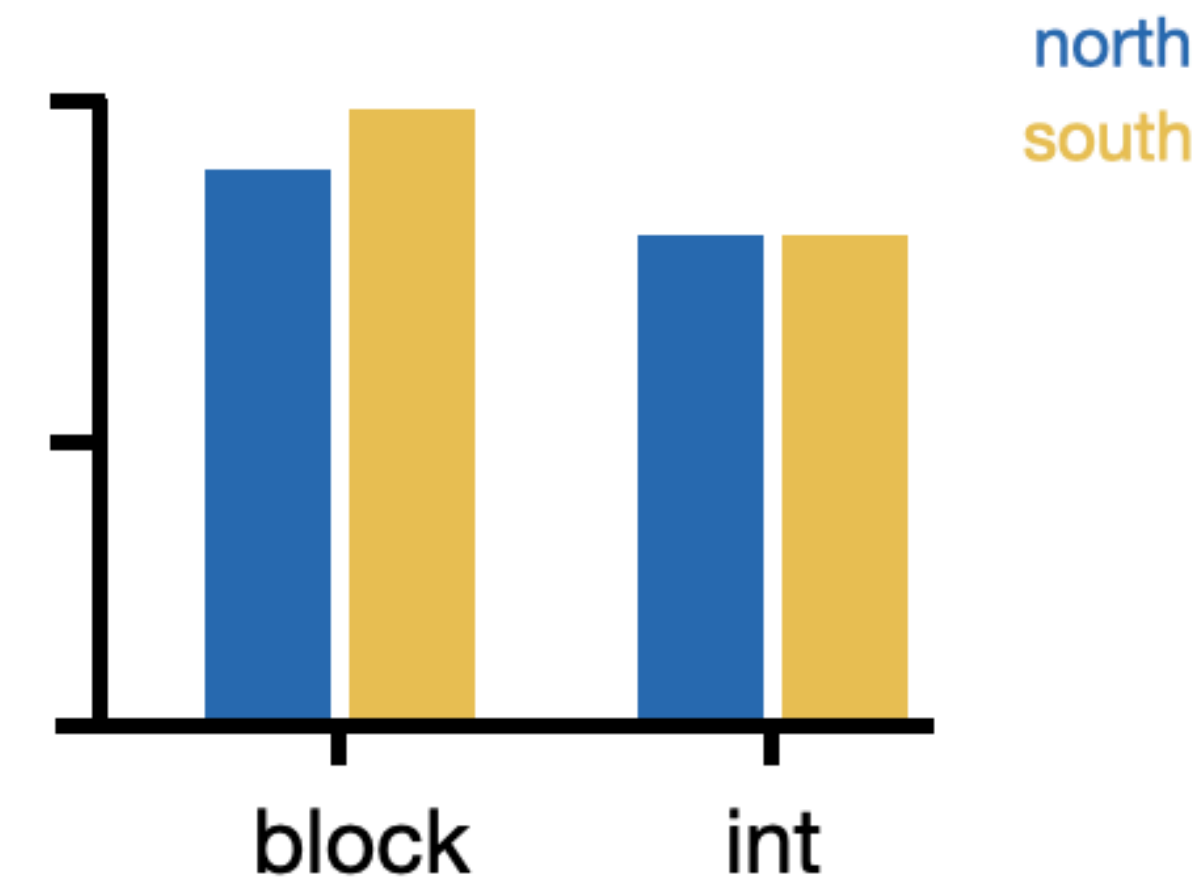
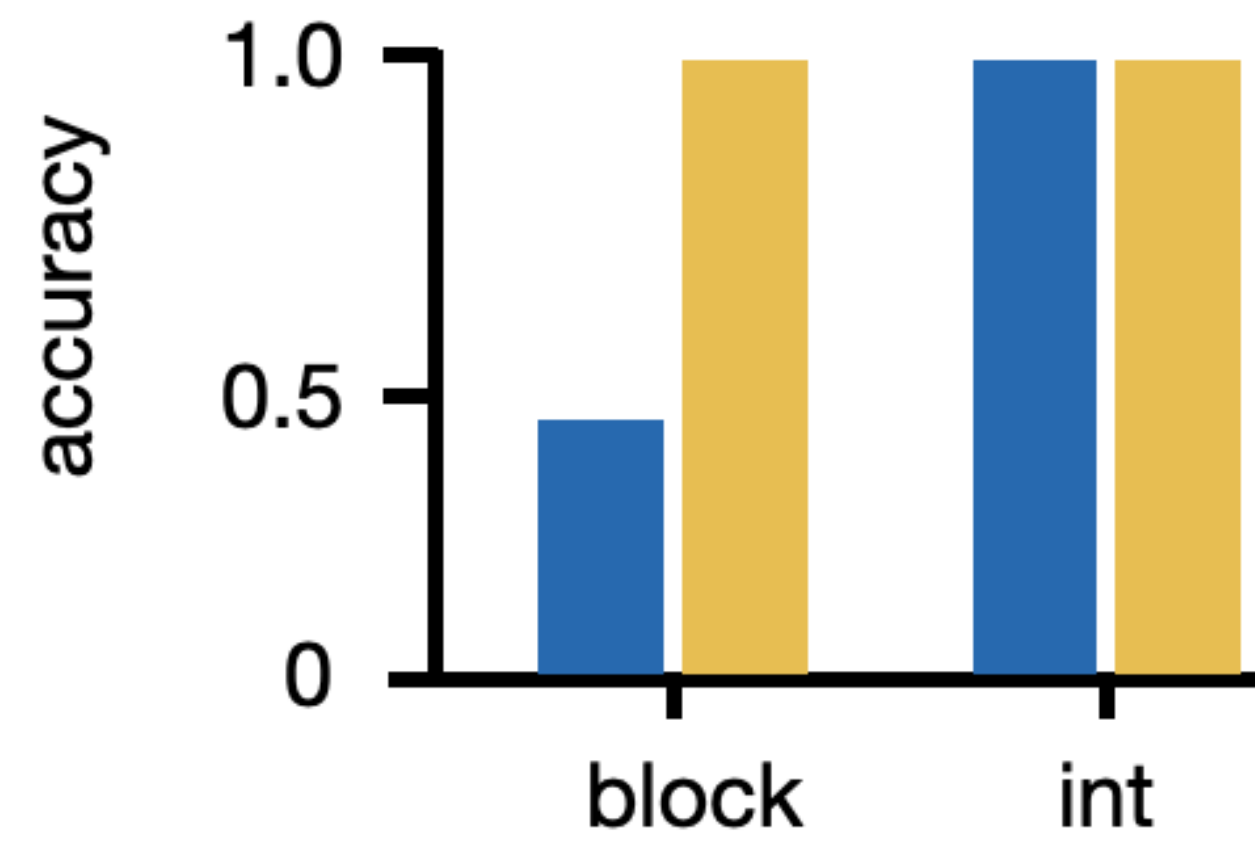




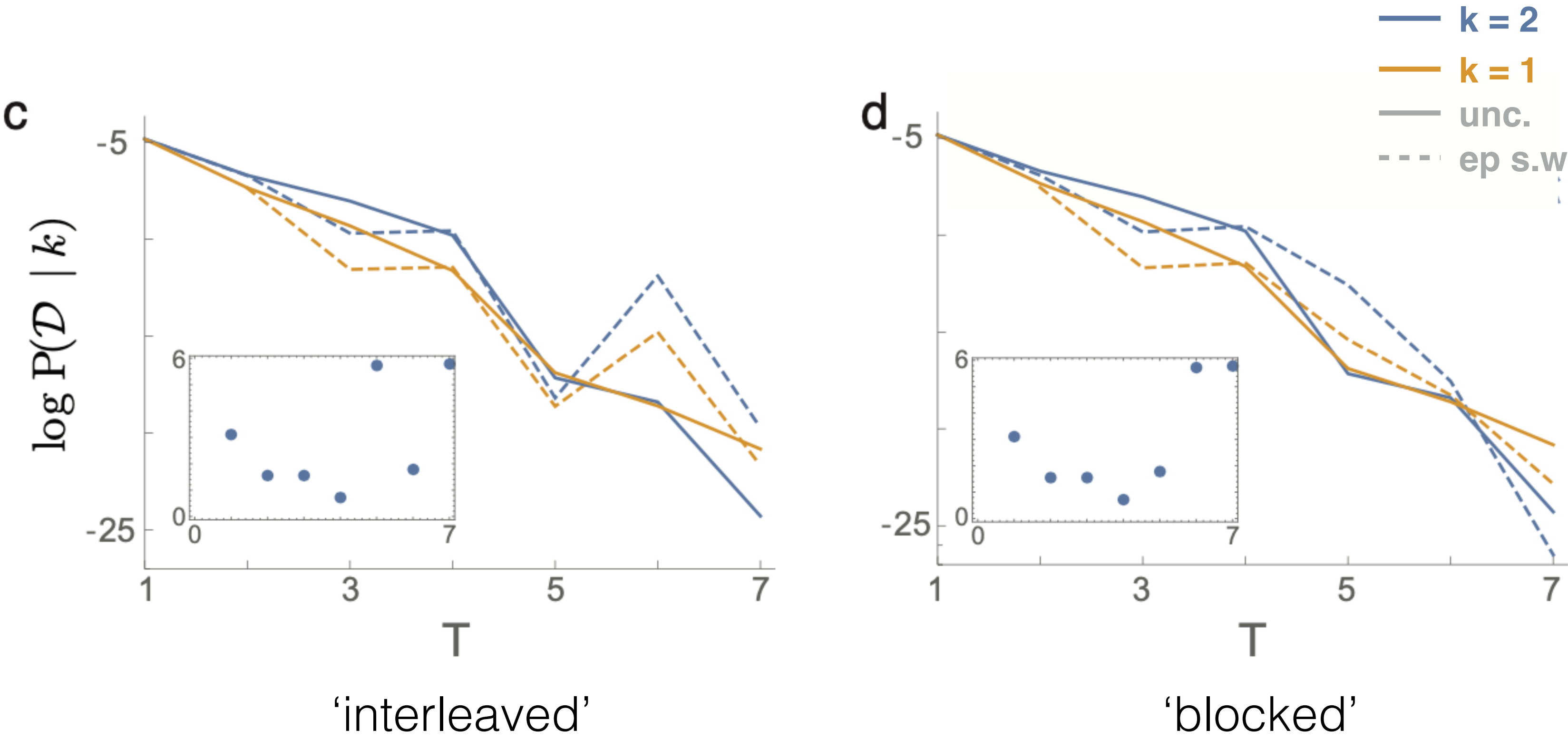
neural network

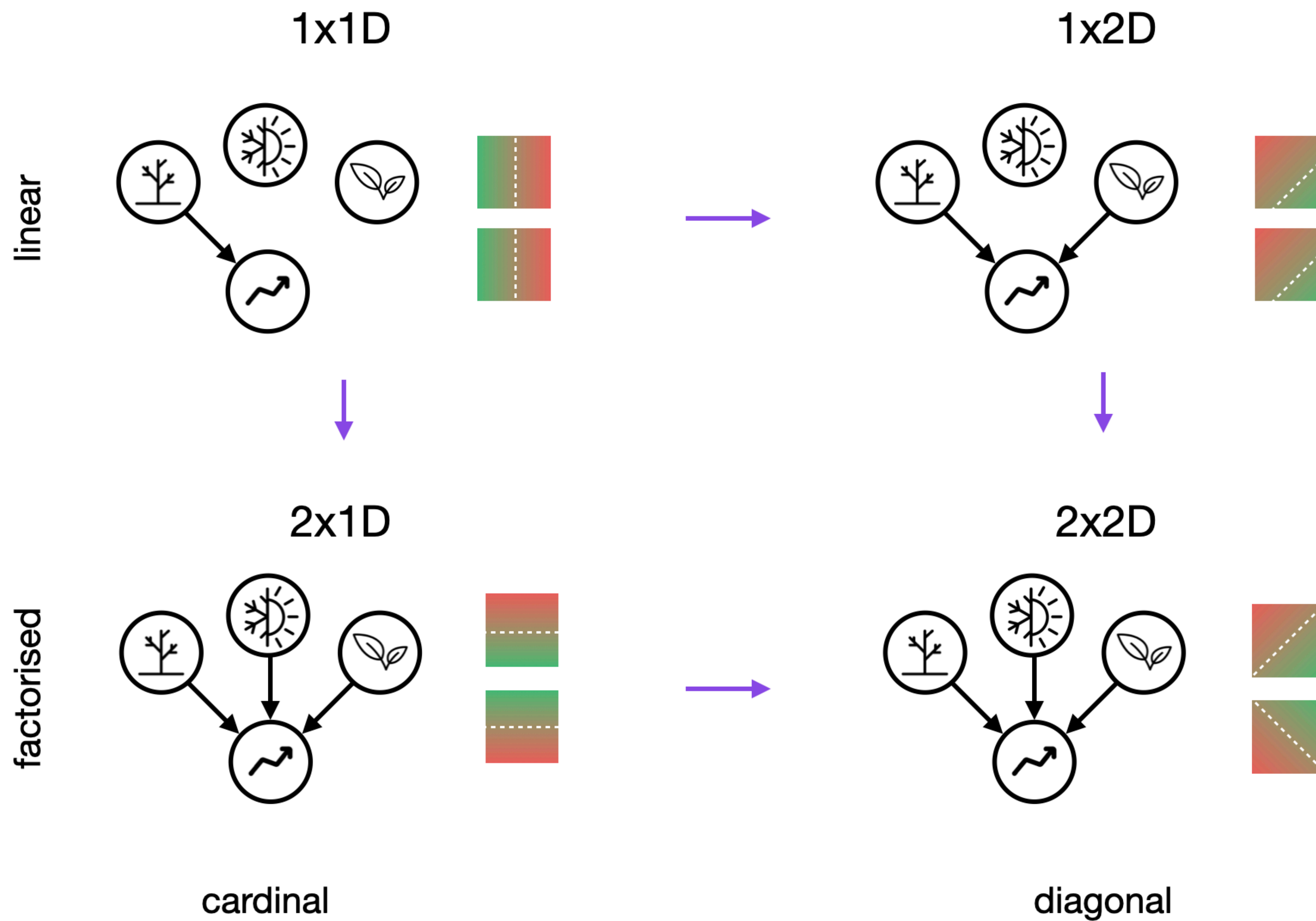


human



order effect

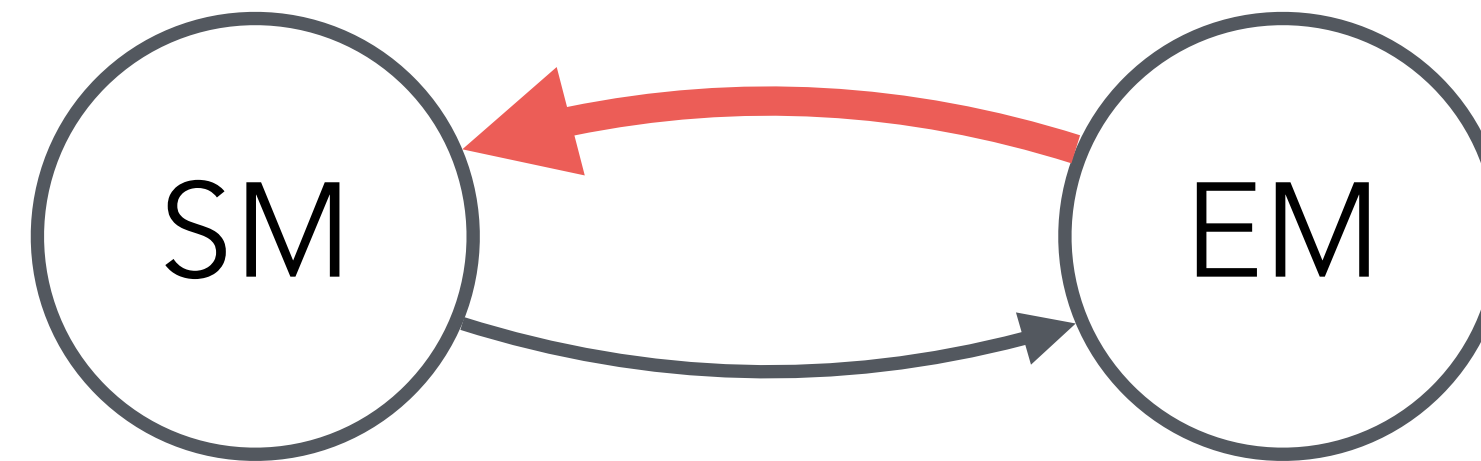




# memory systems

## **semantic**

general  
knowledge about  
how the world  
works



## **episodic**

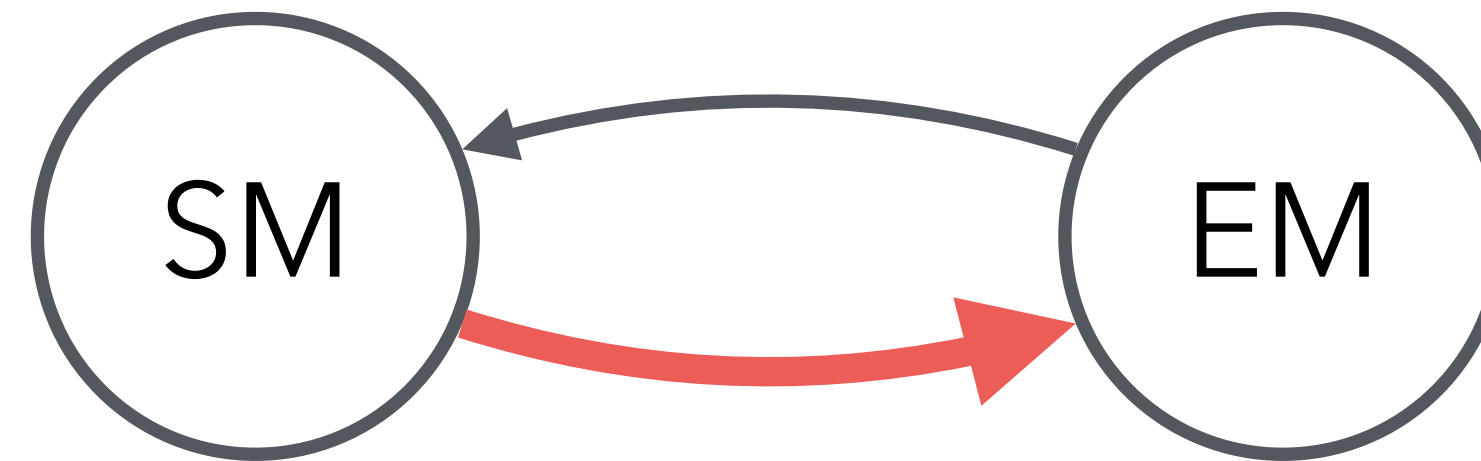
concrete  
experiences

I. why have an episodic memory?

# memory systems

## **semantic**

general  
knowledge about  
how the world  
works



## **episodic**

concrete  
experiences

I.

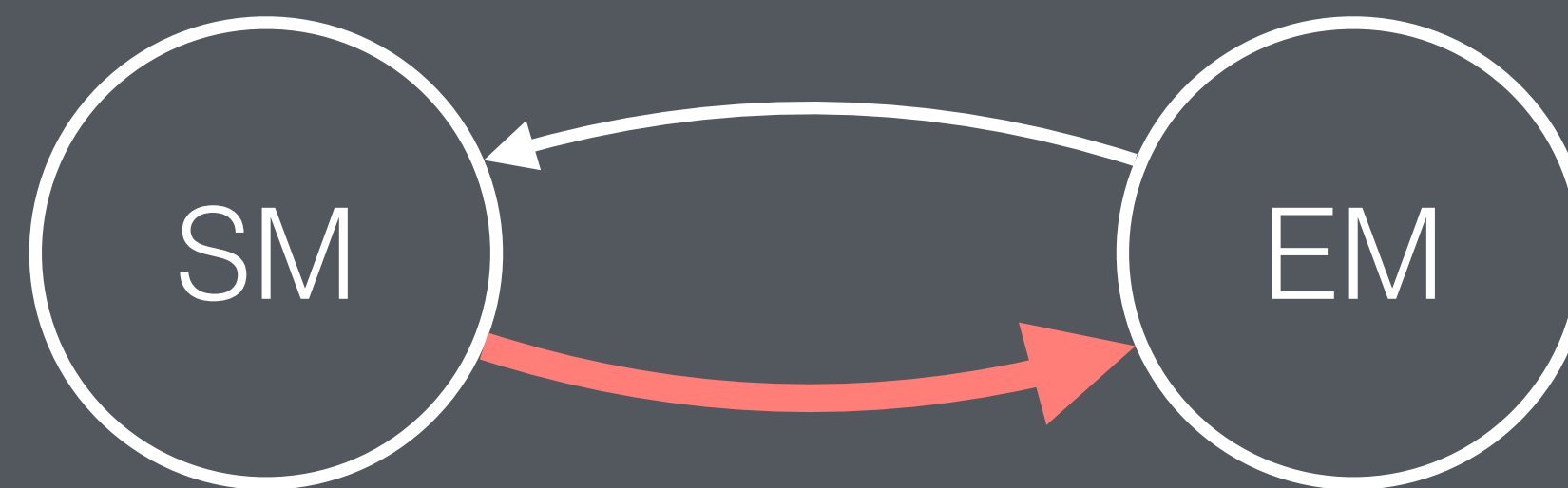
why have an episodic memory?

II.

compression of episodic memories

II.

semantic compression of episodes

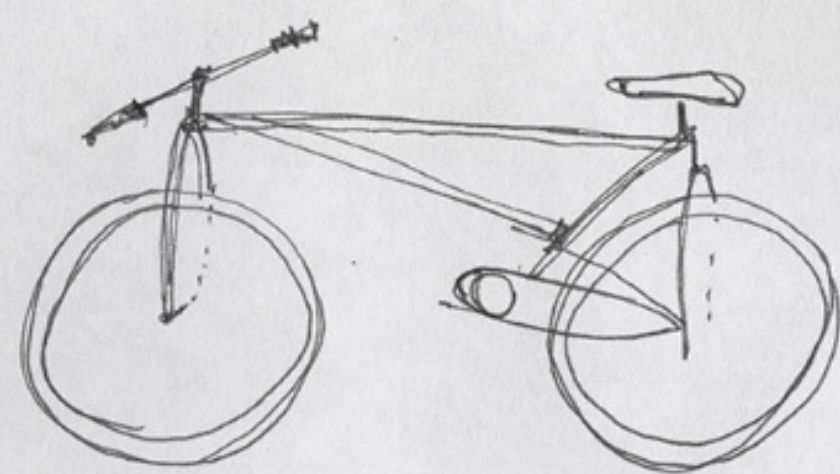










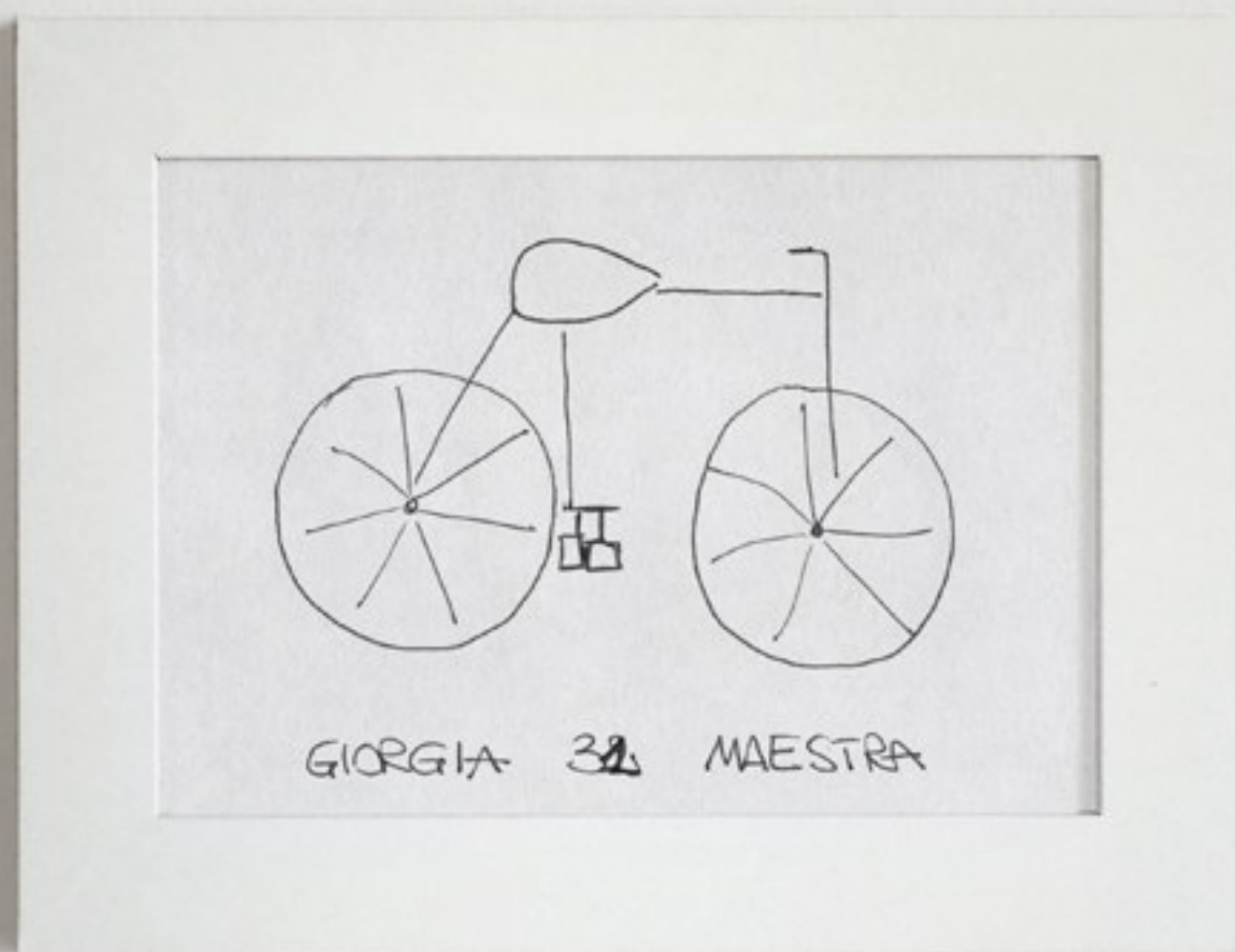


Anna 24 anni studente



Gianluca Gimini





Gianluca Gimini









Gianluca Gimini











Near perfect  
drawing



8%

Forgot the referee



43%

Referee facing  
the wrong way



40%

Drew a hat  
on the referee



18%

Drew a shoe  
instead of a referee

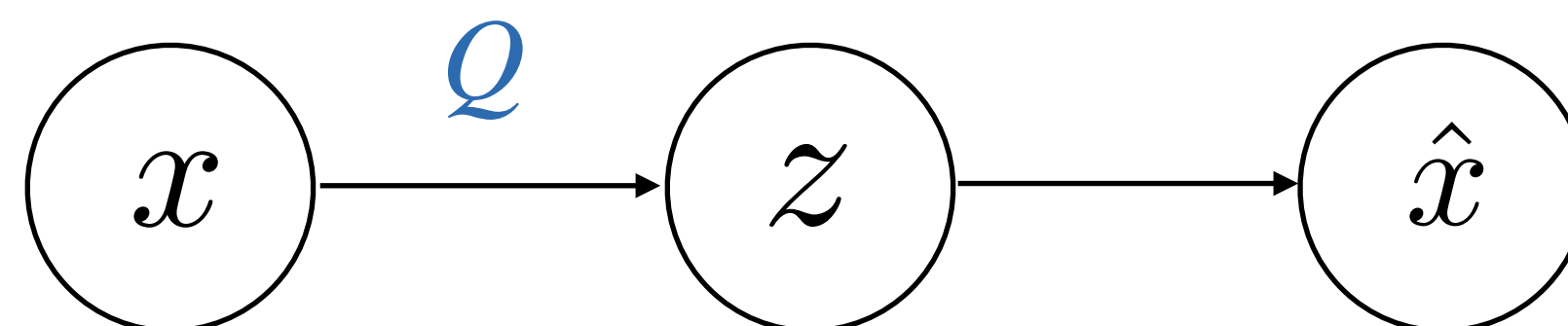


14%

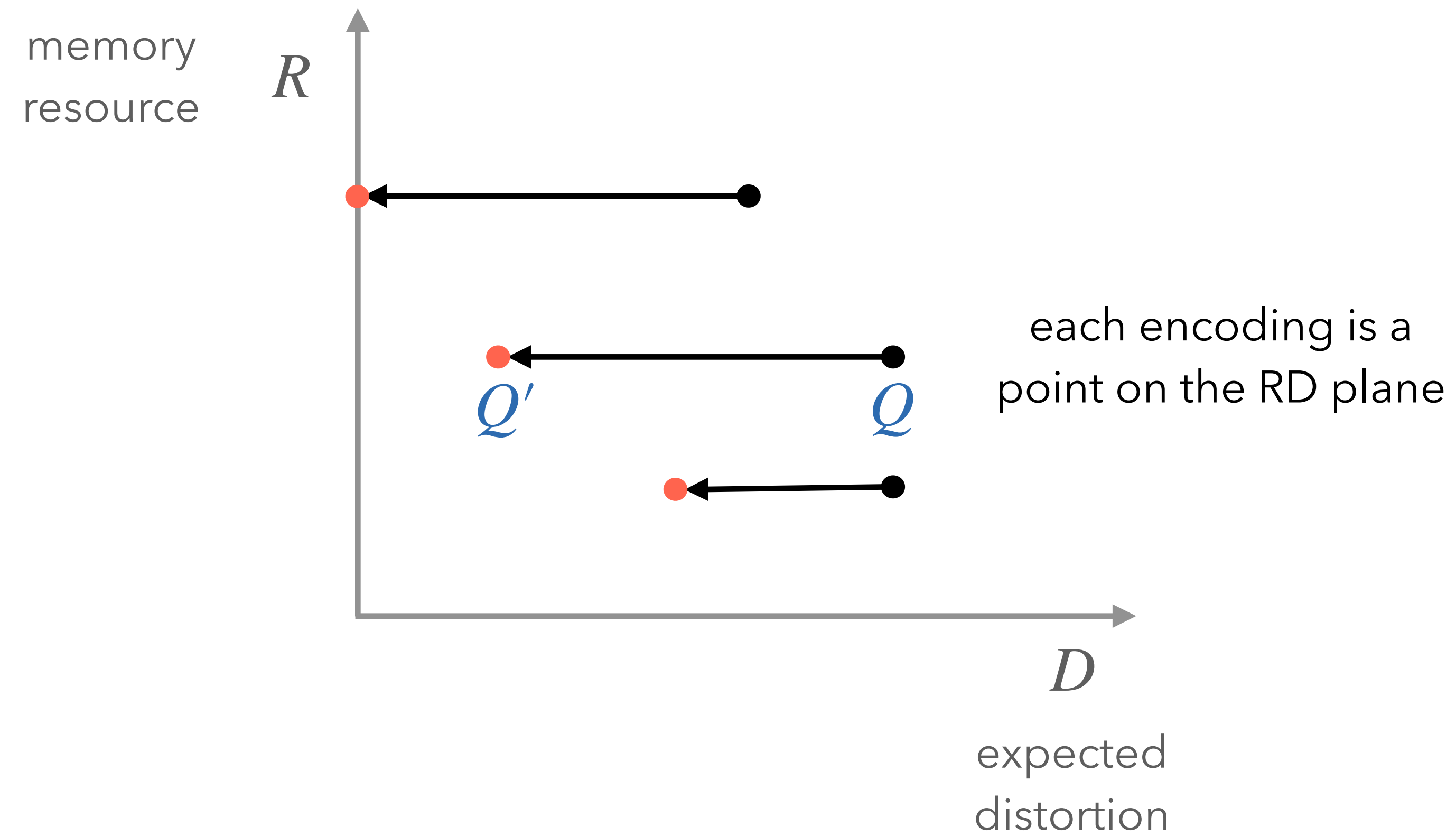
- is the lesson from this simply that human memory is poor?
- memory resources are certainly bounded
- but it is possible to do badly, do well or even optimally **in relation** to available resources



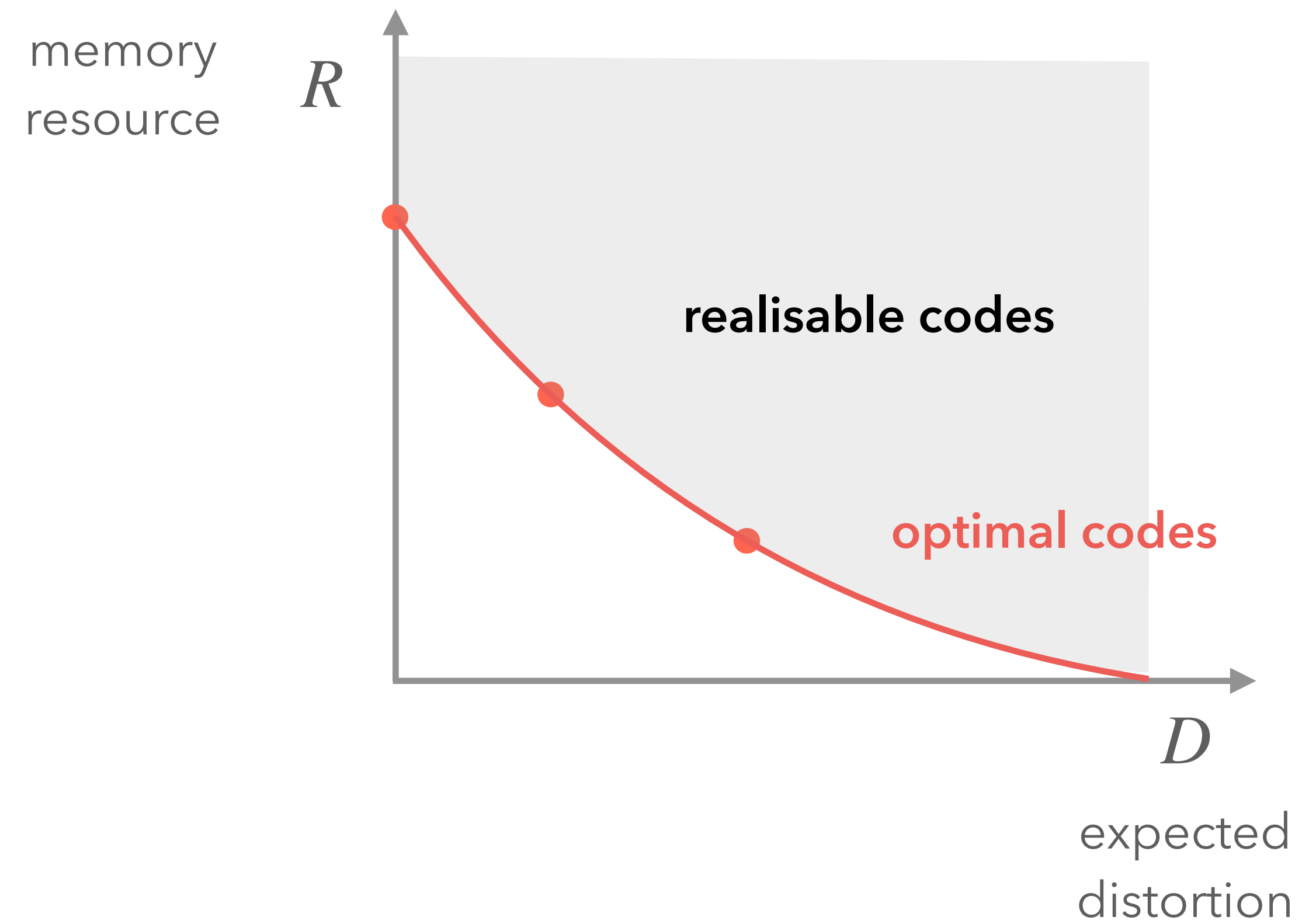
# optimal compression



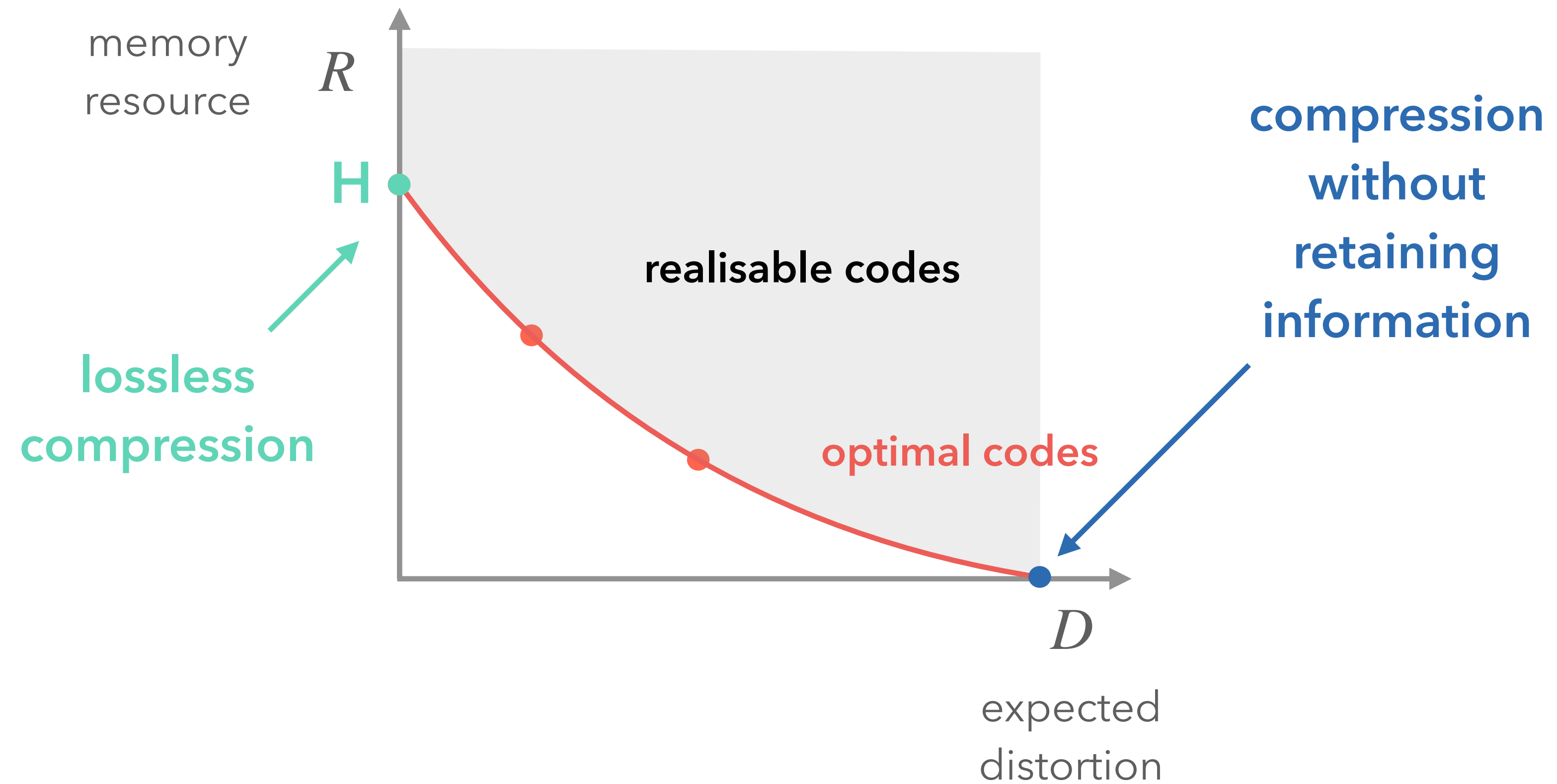
# optimal compression



# optimal compression



# optimal compression





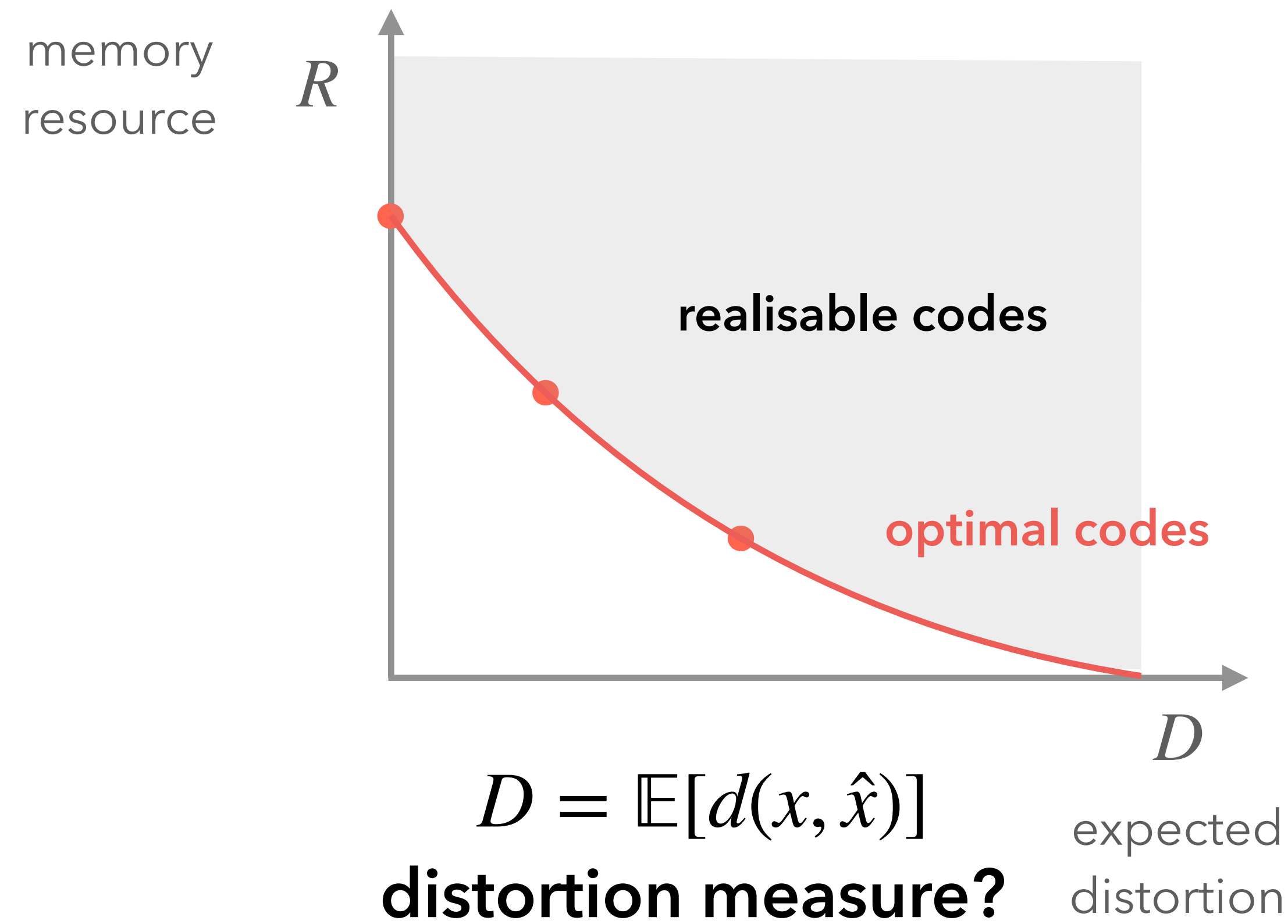
human



jpg



# optimal compression



||

**what changes in sensory input are relevant for the brain?**

# perception as inference

## what we observe

- incoming photons
- air vibrations
- temperature fluctuations
- certain molecules



## what we are interested in

- what objects are around us
- how far
- who are around us
- what are they thinking
- what is going to happen

# perception as inference

## observed variables

- incoming photons
- air vibrations
- temperature fluctuations
- certain molecules

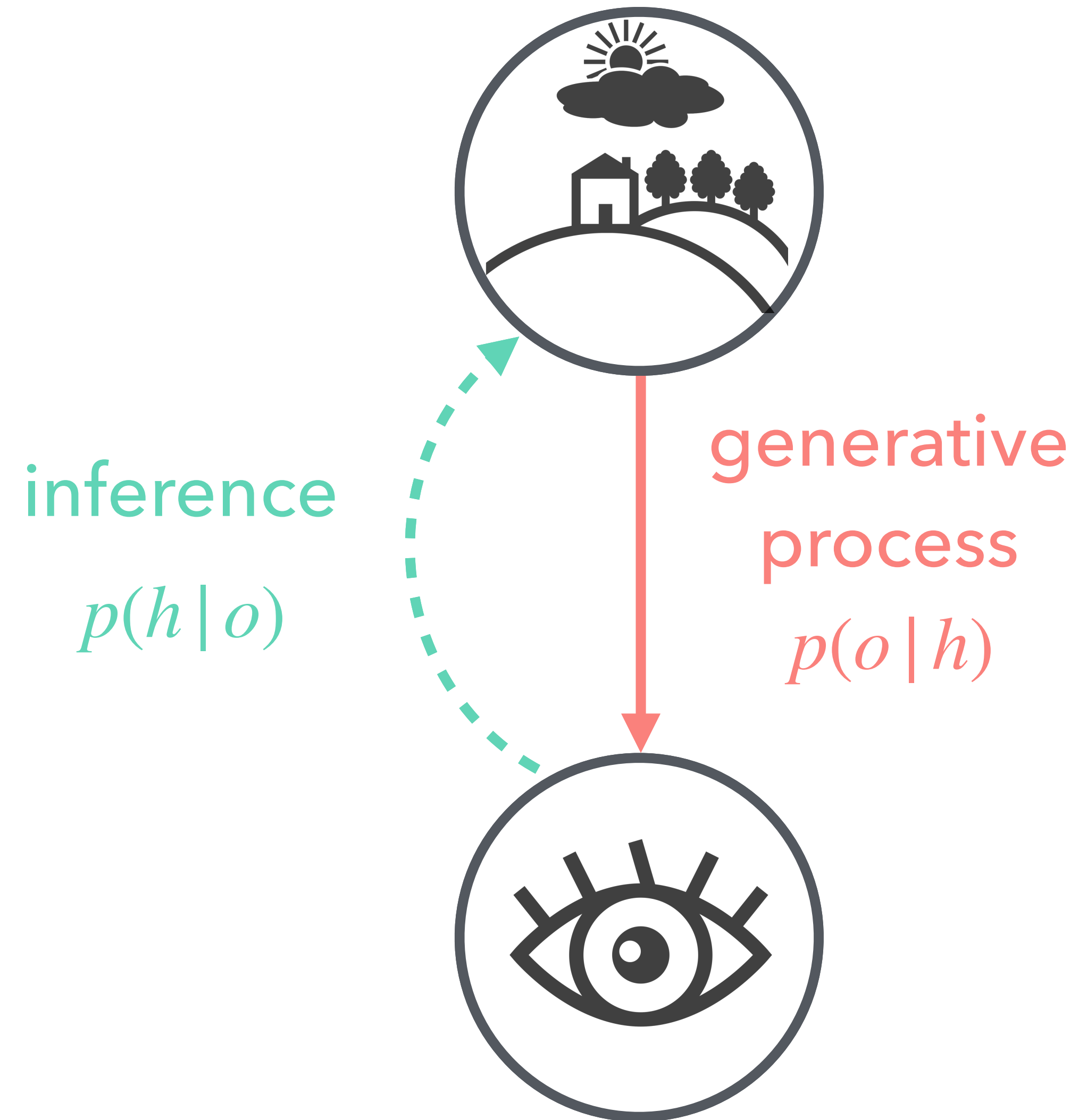


## latent variables

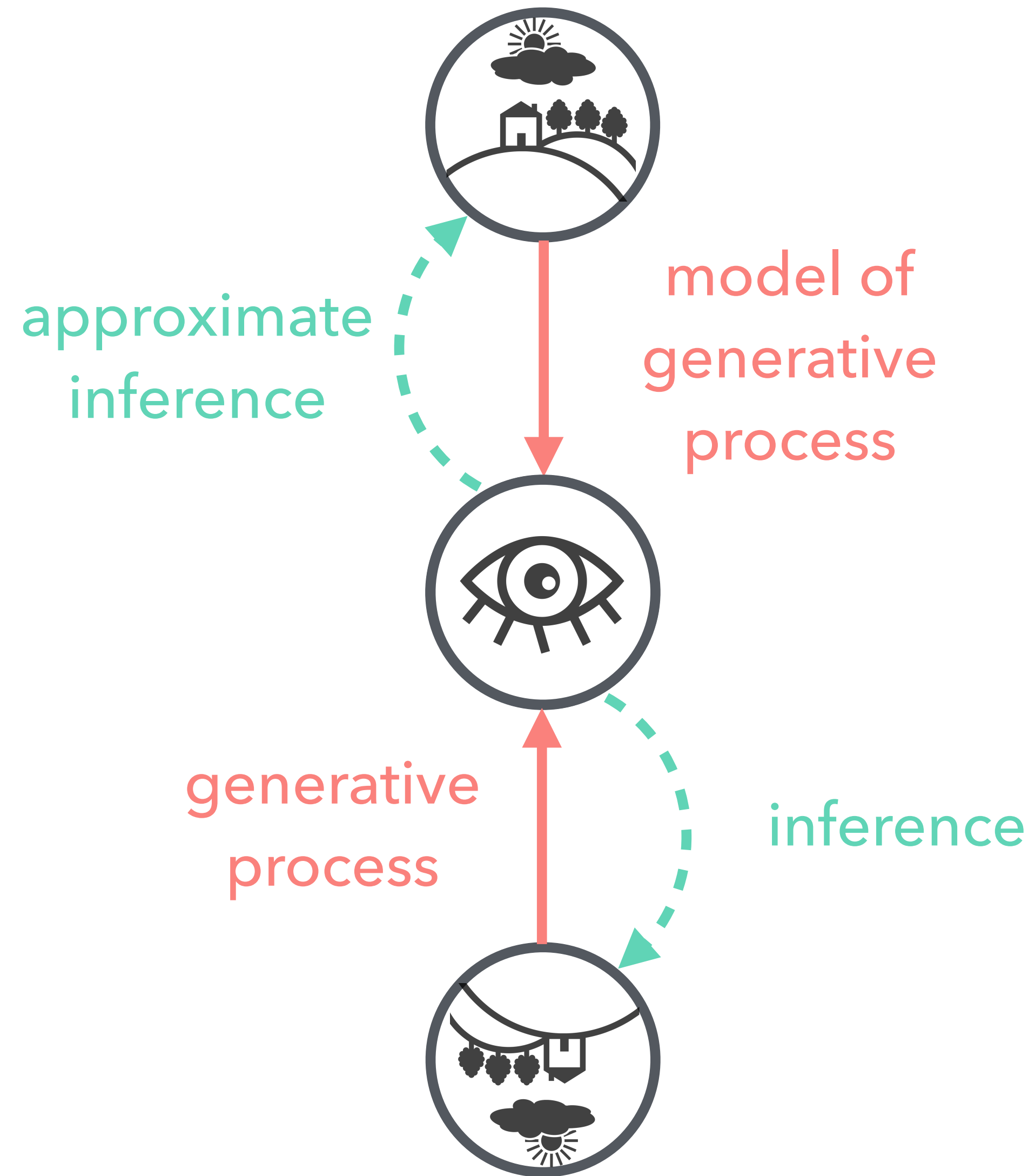
- what objects are around us
- how far
- who are around us
- what are they thinking
- what is going to happen



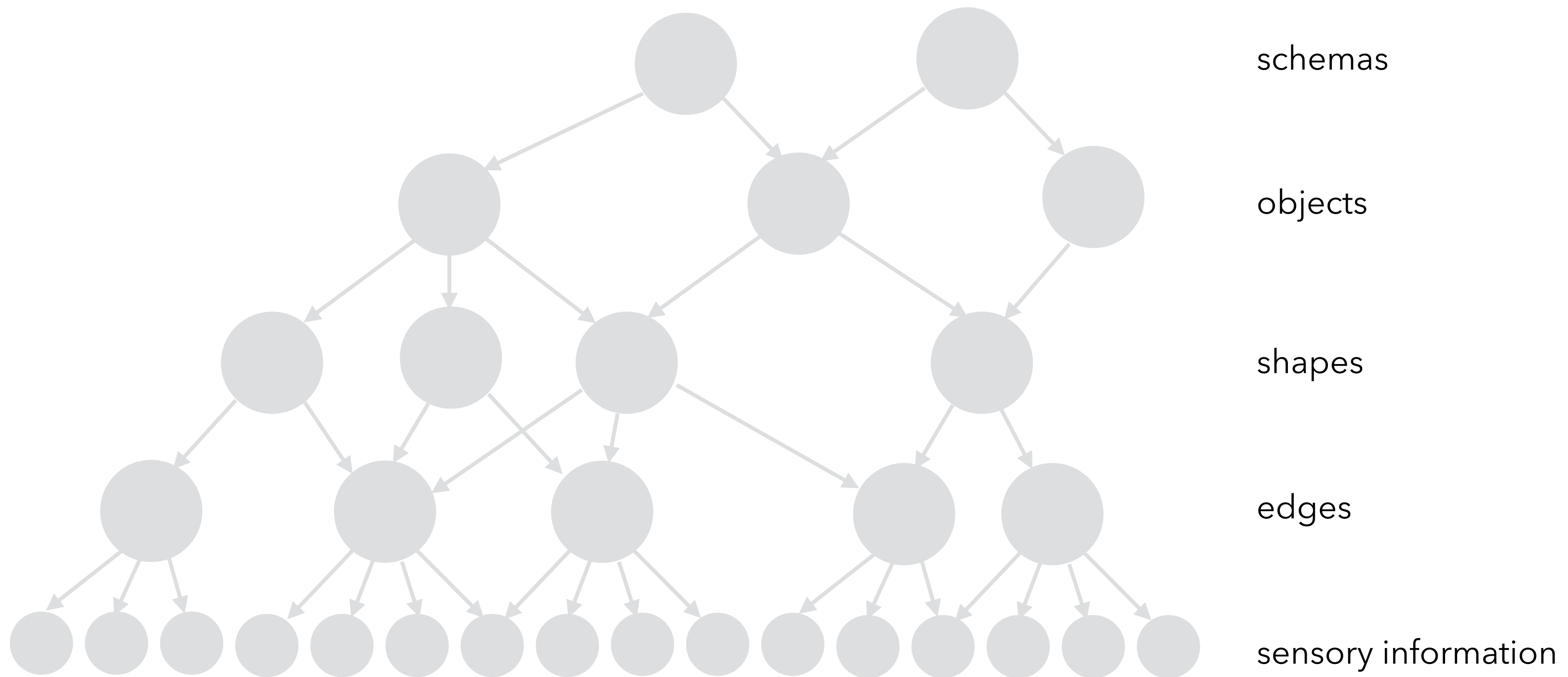
# perception as inference



# perception as inference



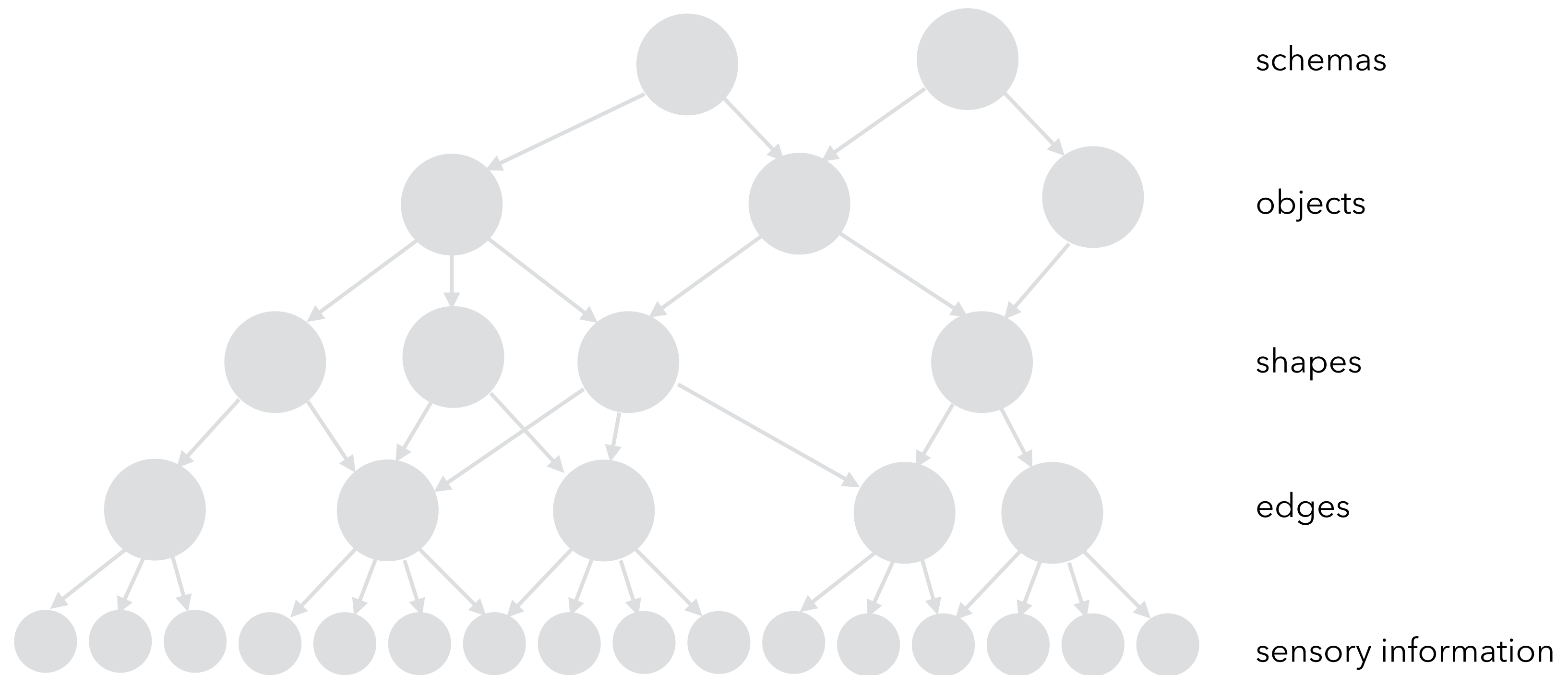
# perception as inference

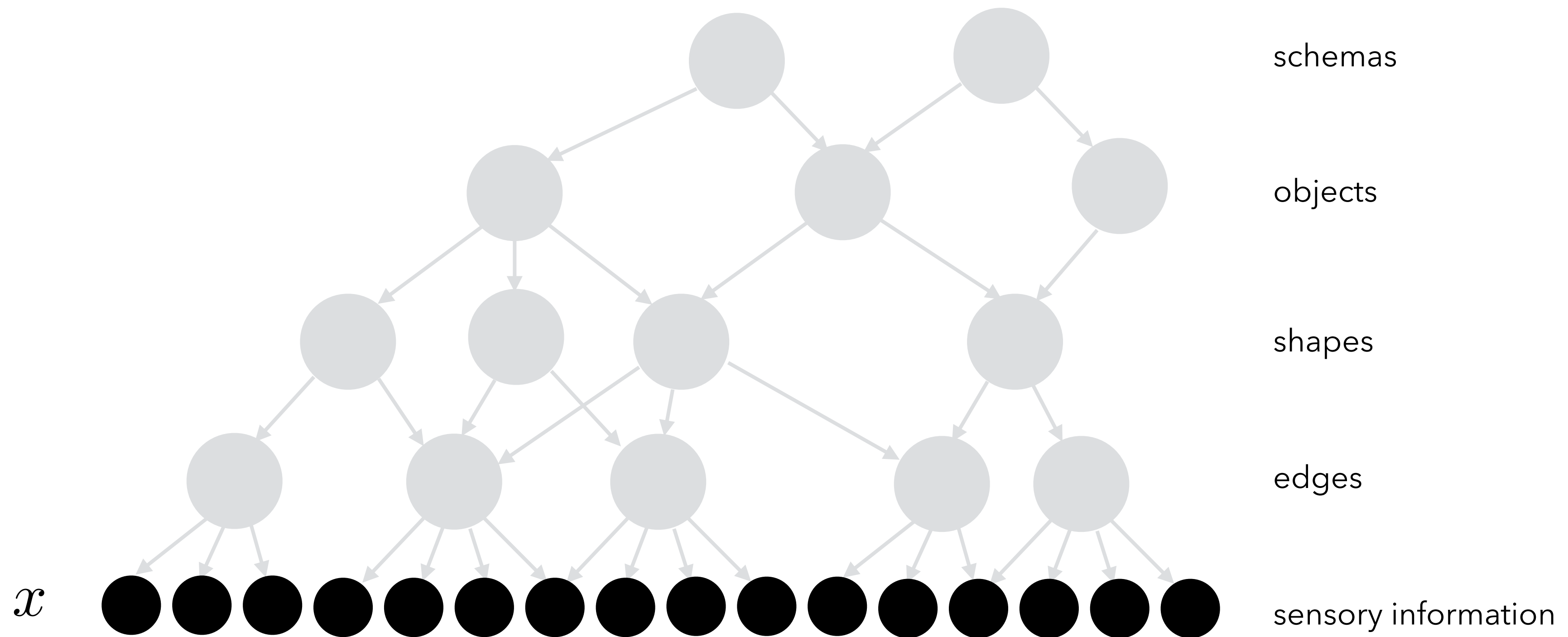


# **Semantic compression hypothesis**

Relevance in sensory input is defined by the latent variables of the internal generative model of the environment used for perception and decision making.

Memory approximates optimal lossy compression with the distortion metric defined by these latent variables.



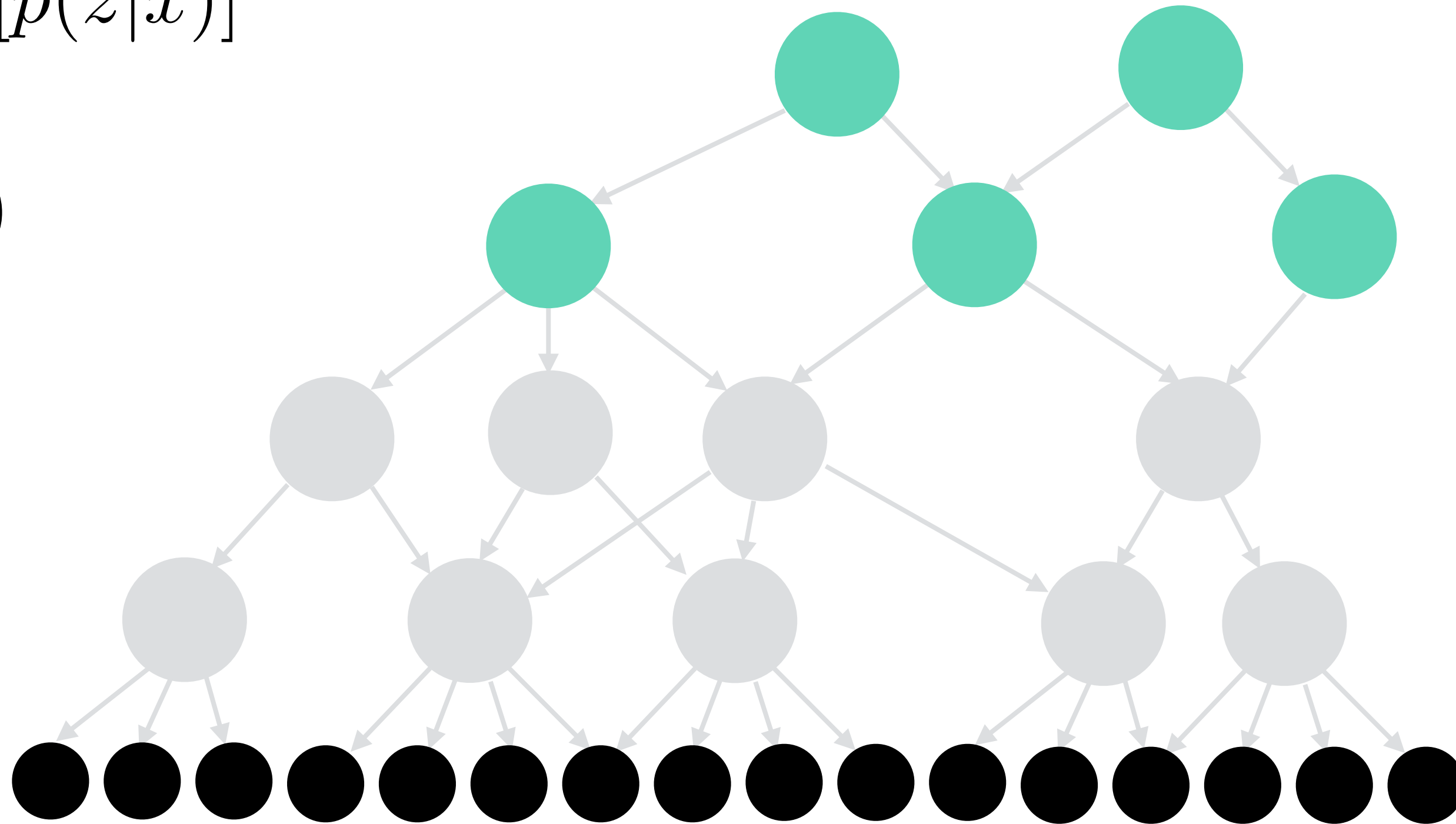


$$\hat{z} = \mathcal{E}[p(z|x)]$$

$$p(z|x)$$

inference

$x$



schemas

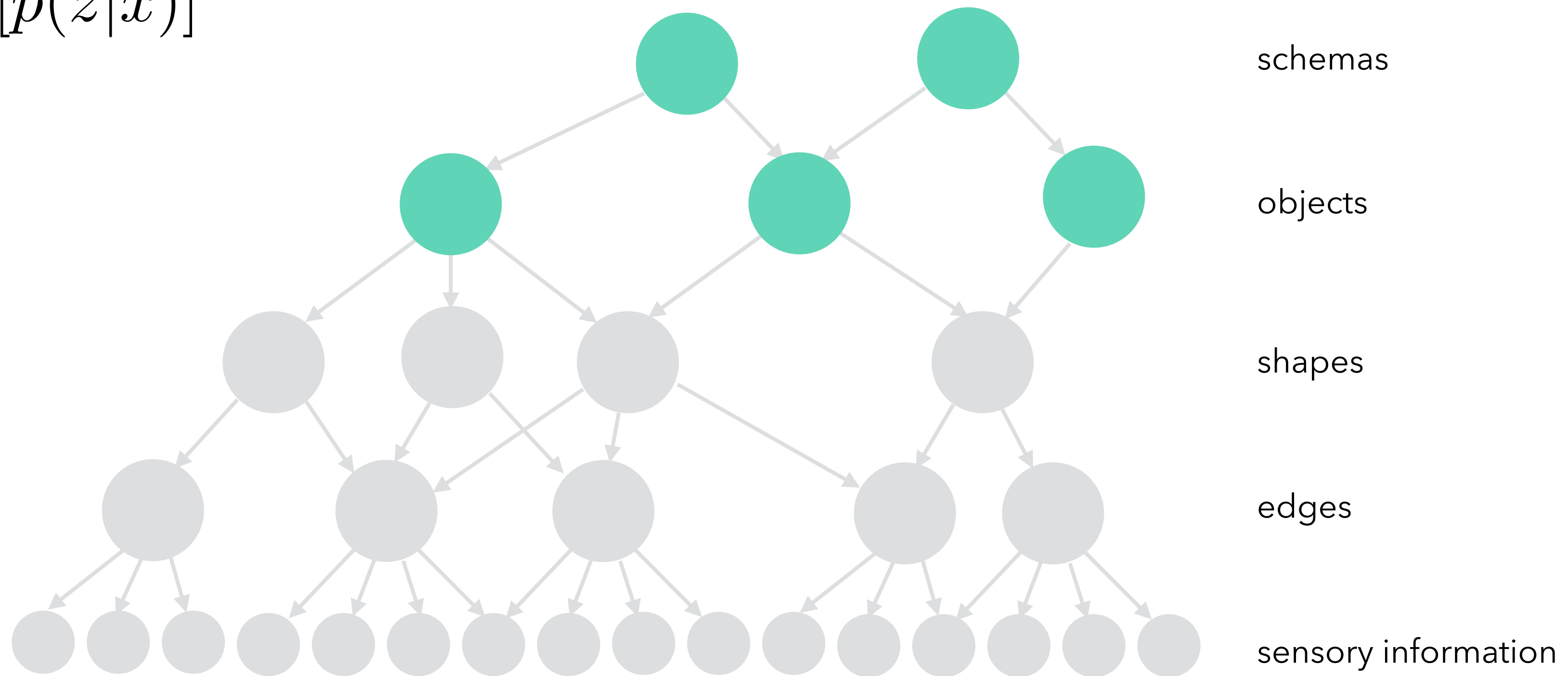
objects

shapes

edges

sensory information

$$\hat{z} = \mathcal{E}[p(z|x)]$$





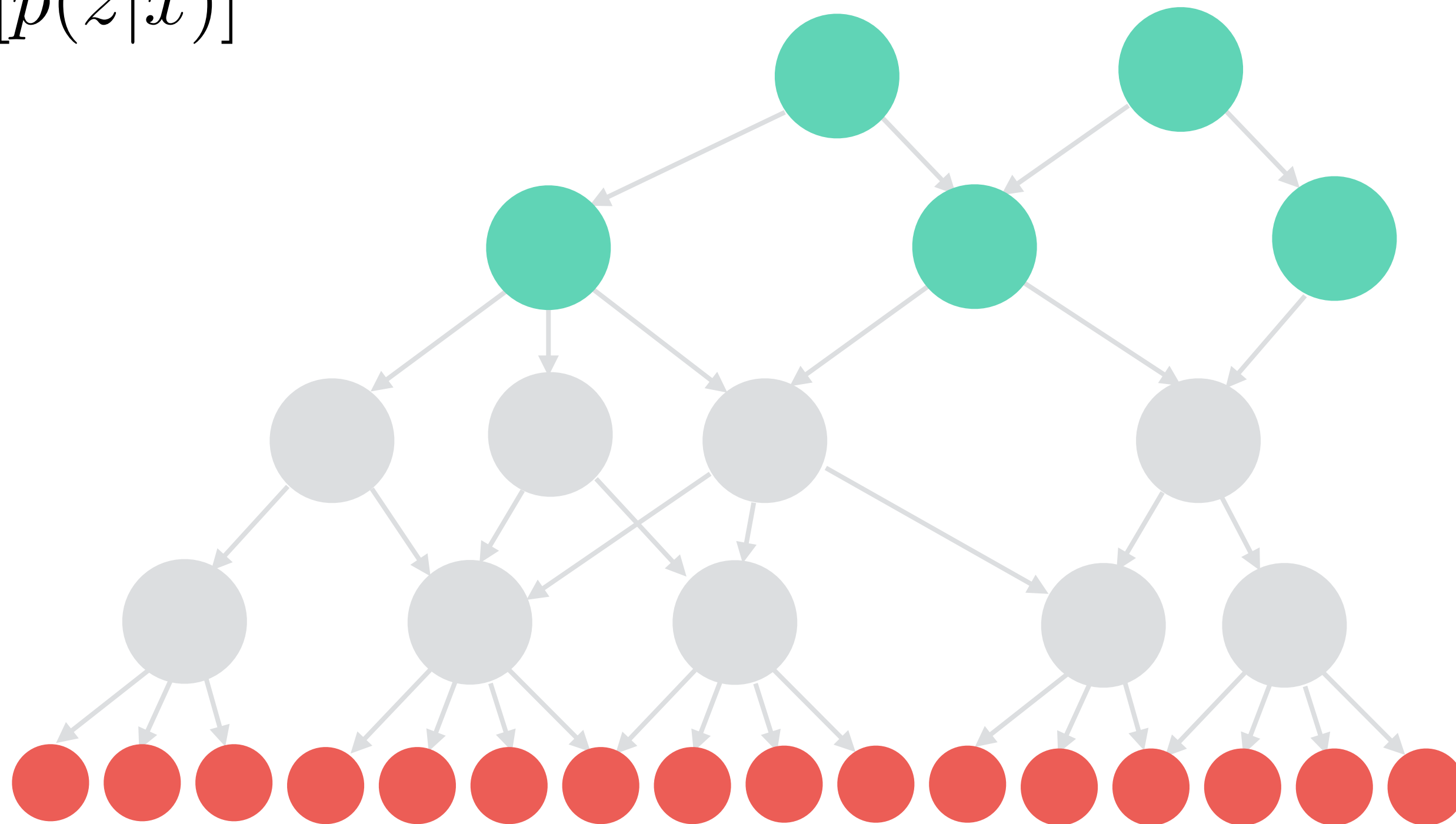
$$\hat{z} = \mathcal{E}[p(z|x)]$$



$$p(x|\hat{z})$$

generation

$$\hat{x}$$



schemas

objects

shapes

edges

sensory information

## generative models

Semantic  
Compression

variational approx.  
(Kingma et al. 2013)

VAE

extension  
(Higgins et al. 2017)

$\beta$ -VAE

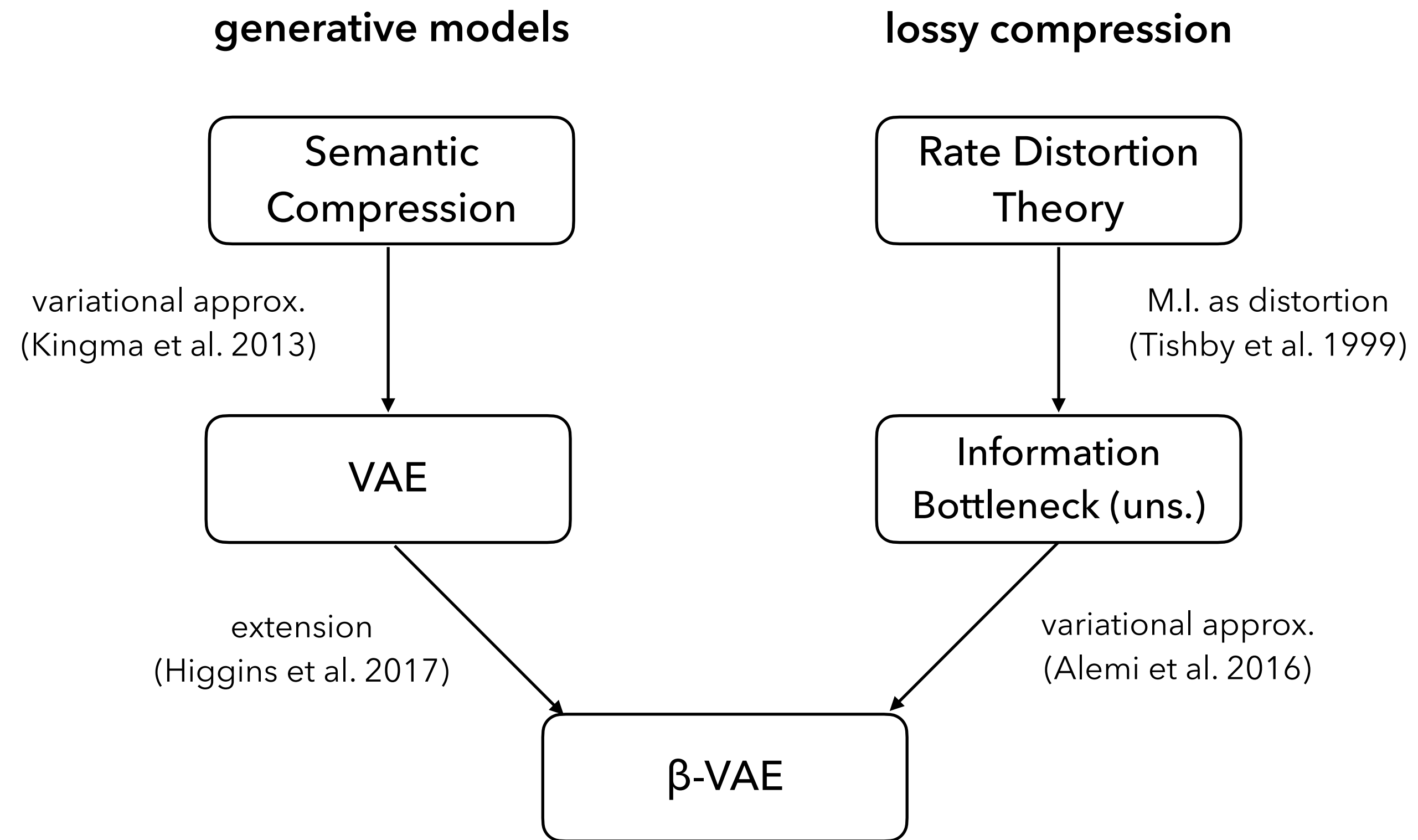
## lossy compression

Rate Distortion  
Theory

M.I. as distortion  
(Tishby et al. 1999)

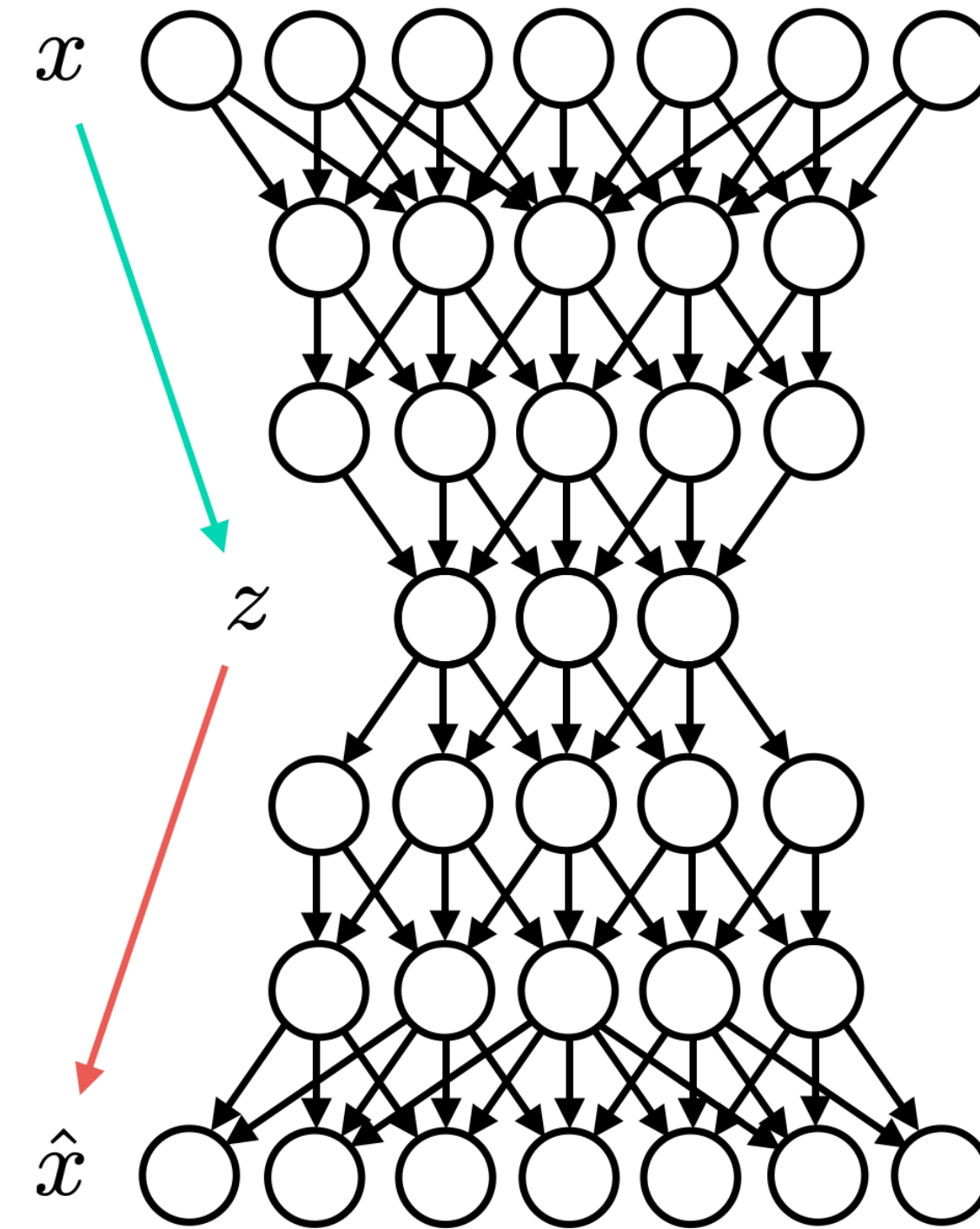
Information  
Bottleneck (uns.)

variational approx.  
(Alemi et al. 2016)



# beta-VAE

- unsupervised generative model
- approximate inference
  - variational bayes
  - neural networks as generic function approximators
- information theoretic interpretation as approximate lossy compression



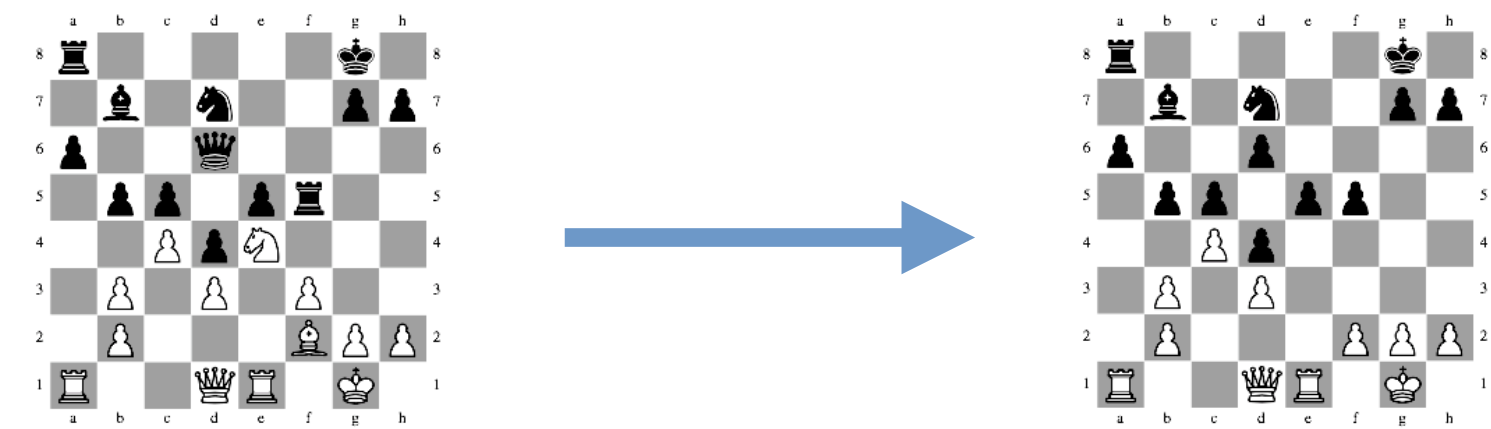
$$\mathcal{L}(\theta, \phi, x) = -\beta \underbrace{KL(q(z|x, \phi) || p(z|\theta))}_{\text{rate}} + \underbrace{\mathbb{E}_{z \sim q(z|x, \phi)} (\log p(x|z, \theta))}_{\text{distortion}}$$

# methods



**data**

**chess  
positions**



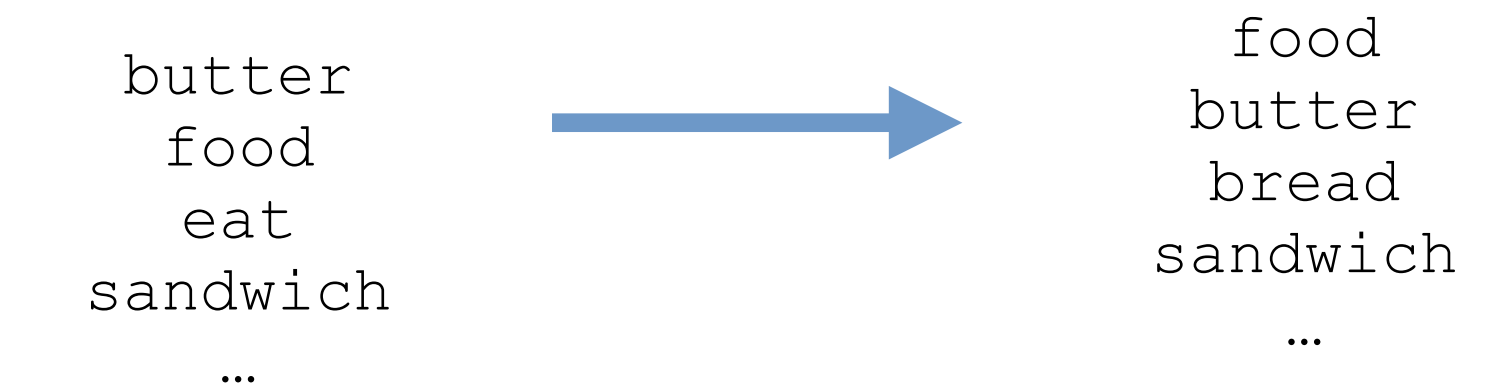
3000 chess games  
(FICS database)

**sketches**



quickdraw75k object  
pairs

**word lists**



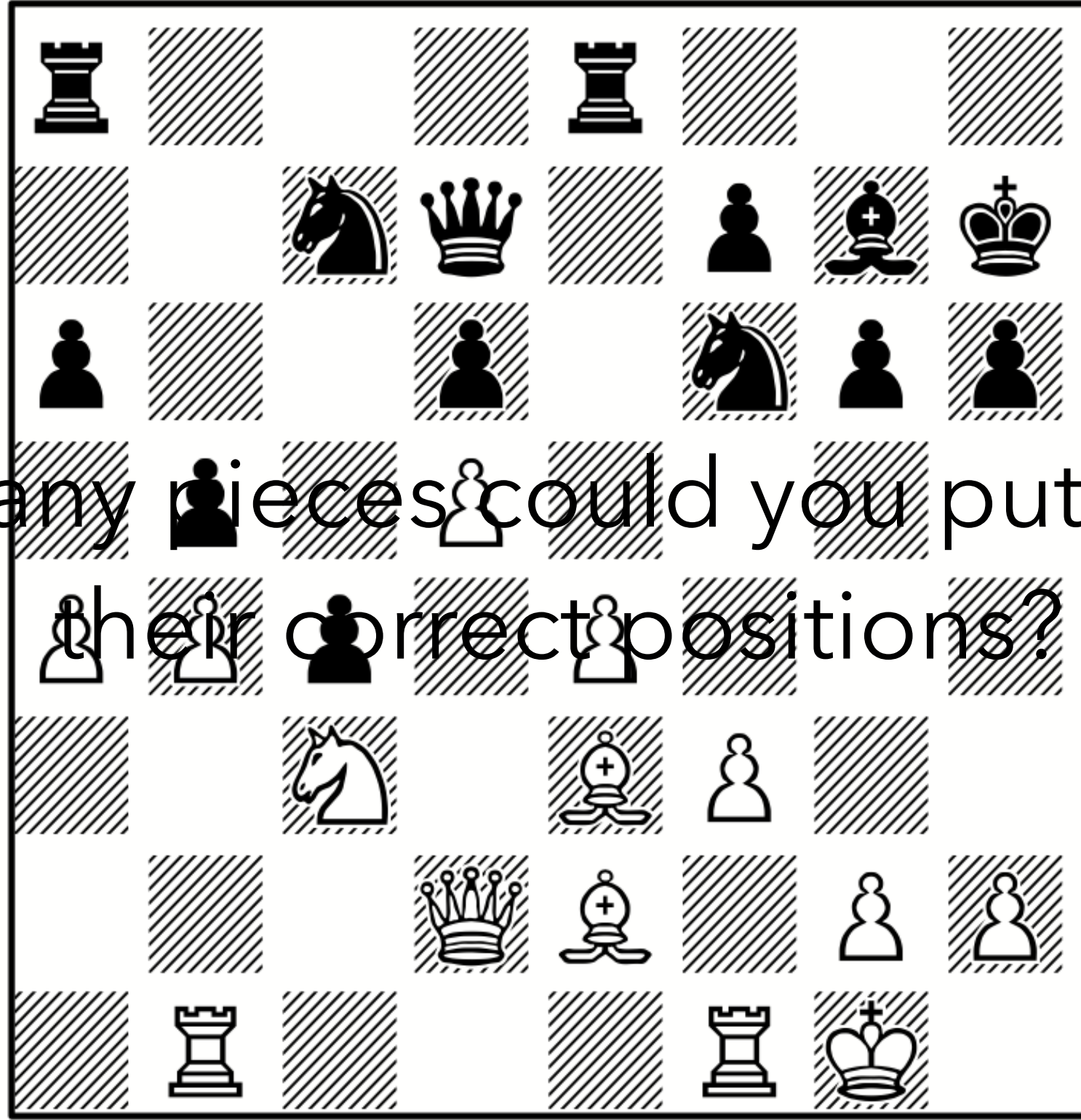
(small subset of)  
wikipedia

Consequence 1.

## **Domain expertise**

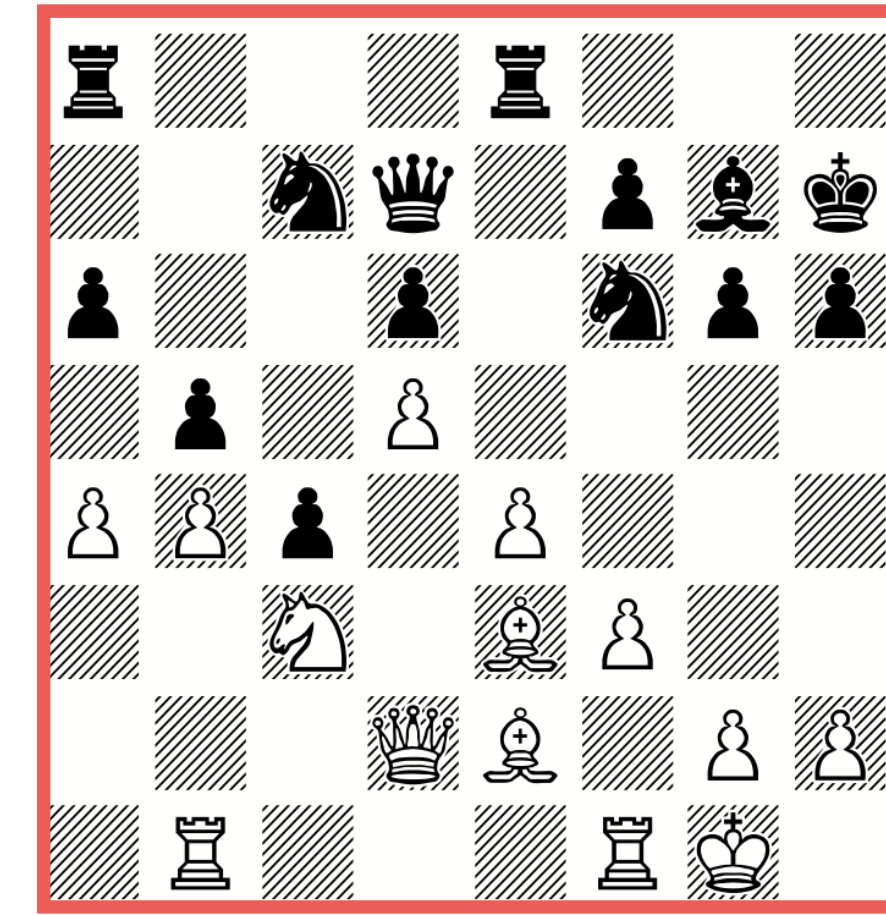
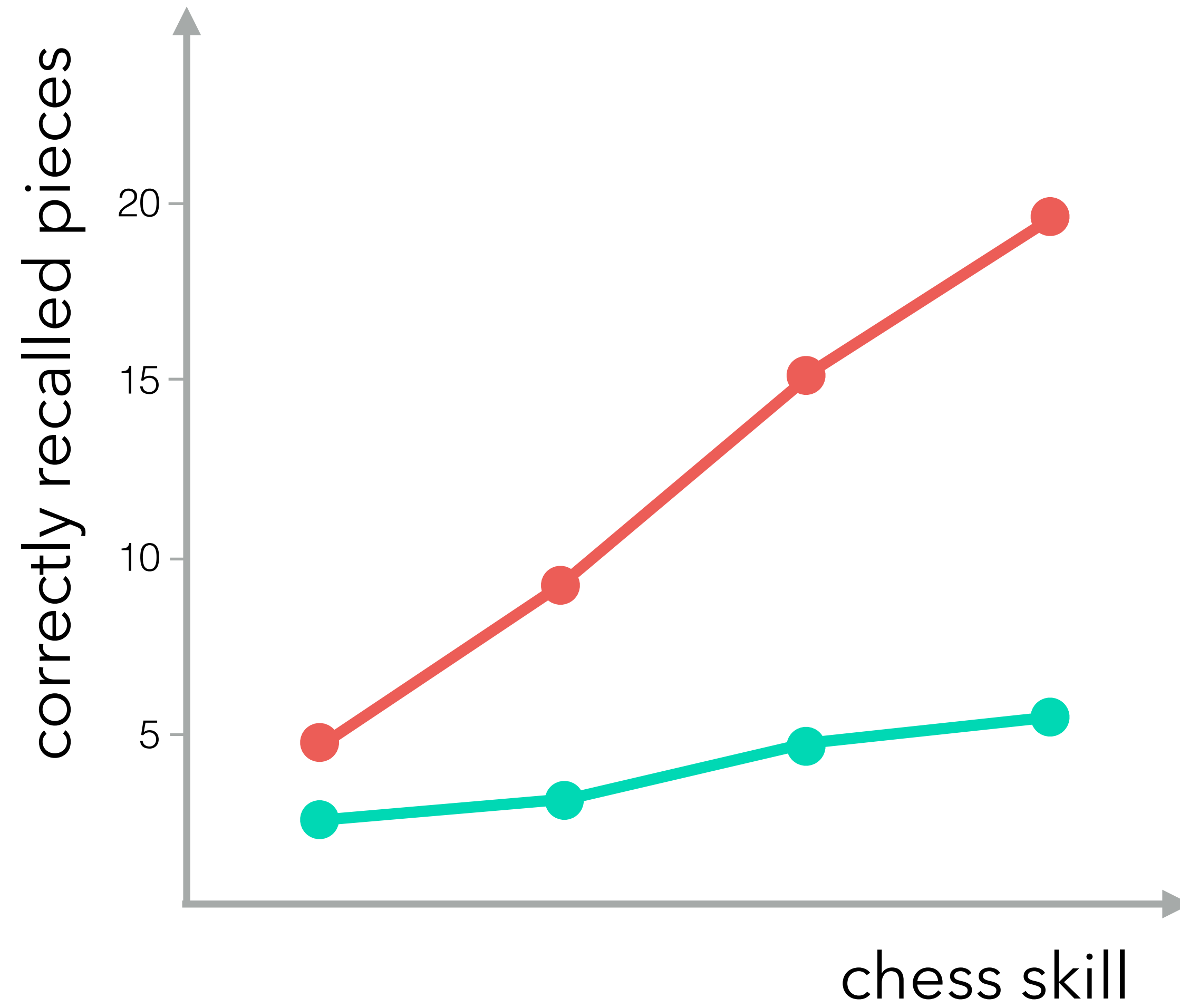
Since compression hinges on environmental statistics that were learned from observations, experience in a cognitive domain increases recall accuracy for observations congruent with this statistics.



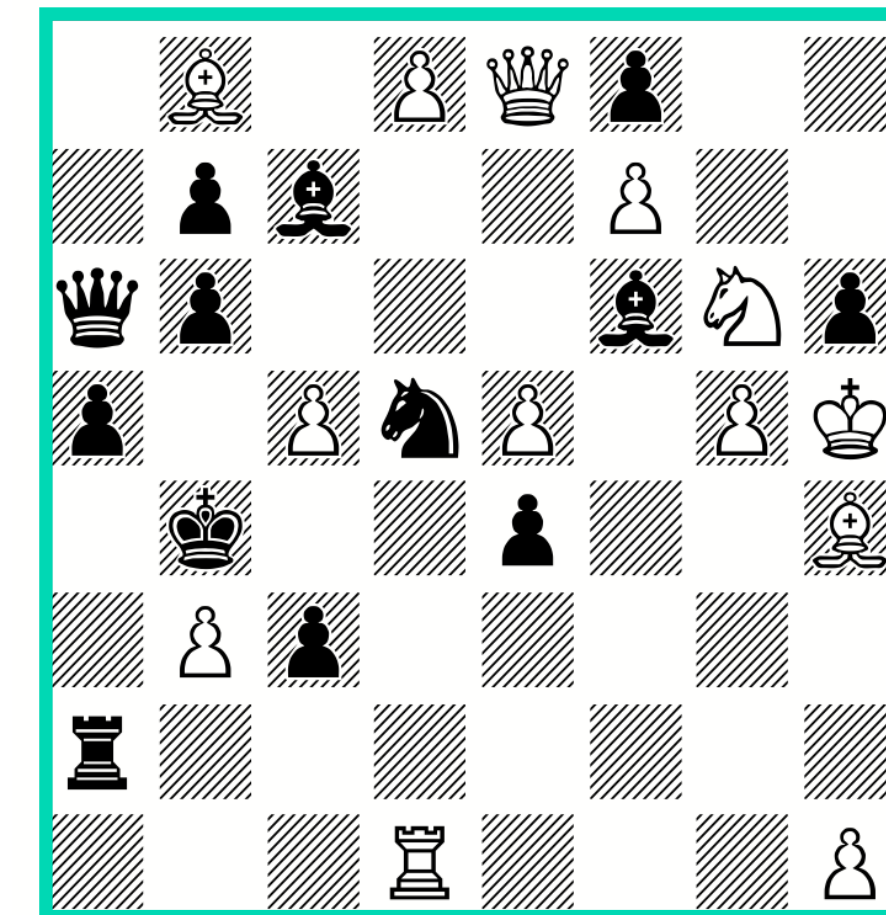


how many pieces could you put back to  
their correct positions?





game

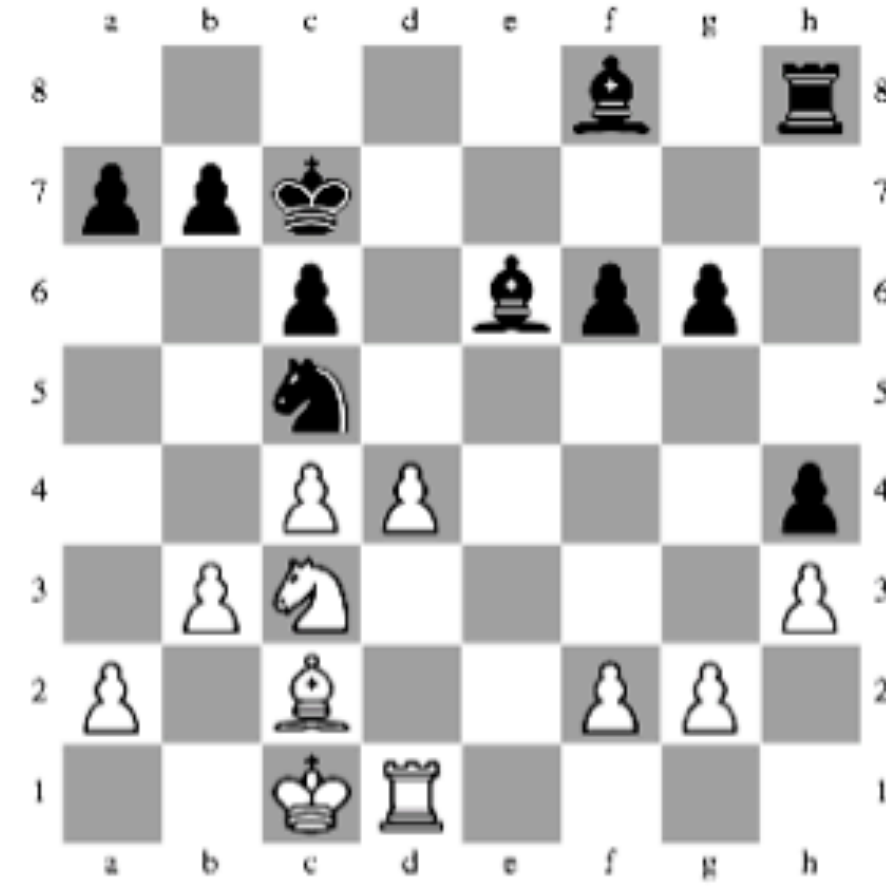


random

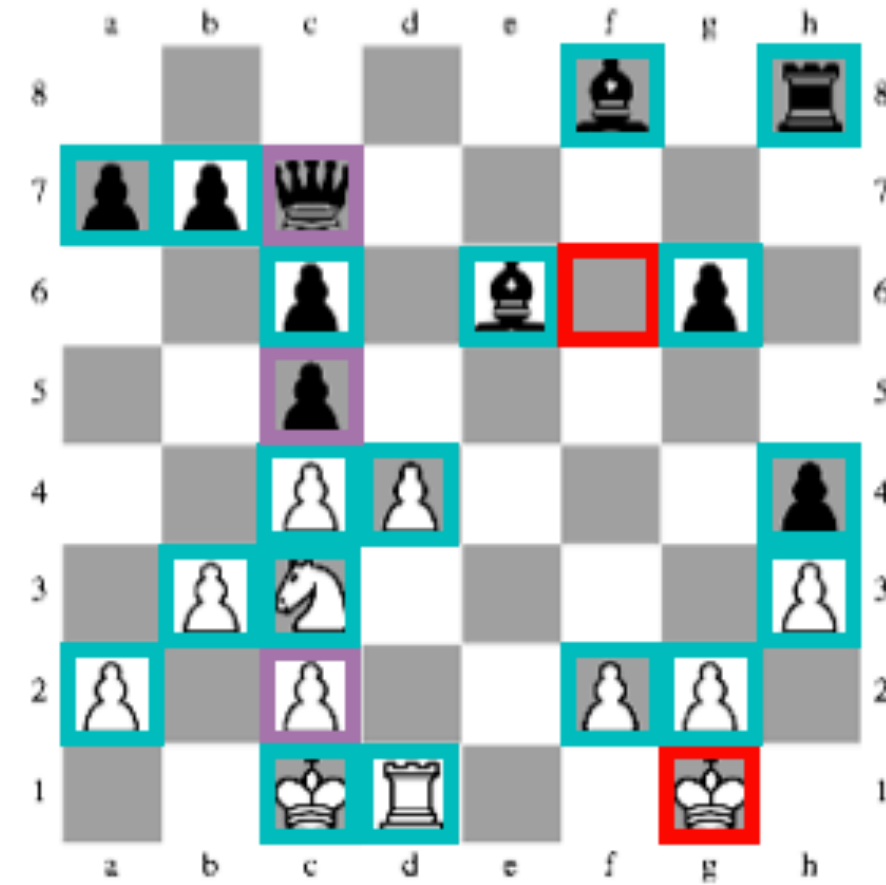
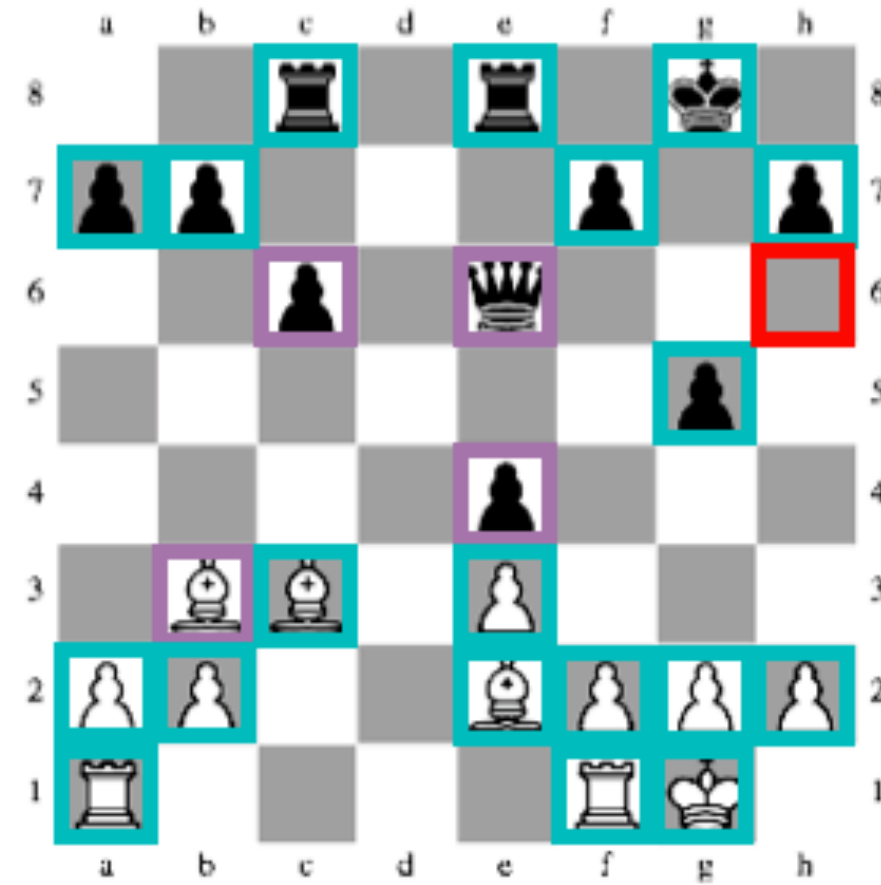
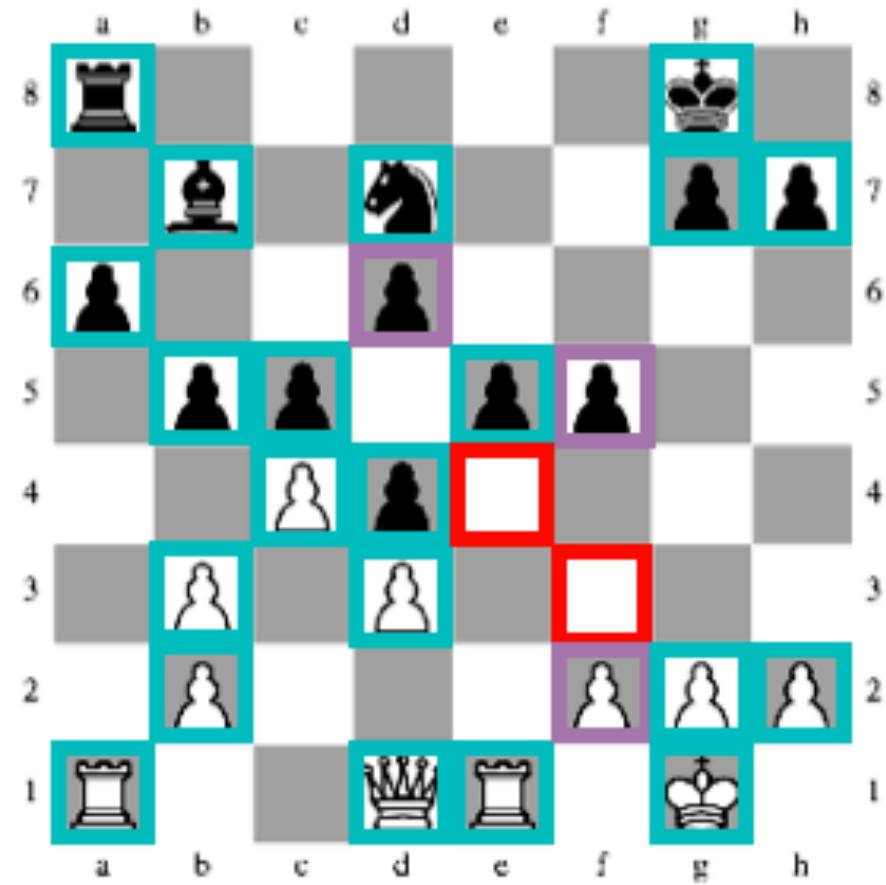


# reconstructions (game)

sample



reconstruction

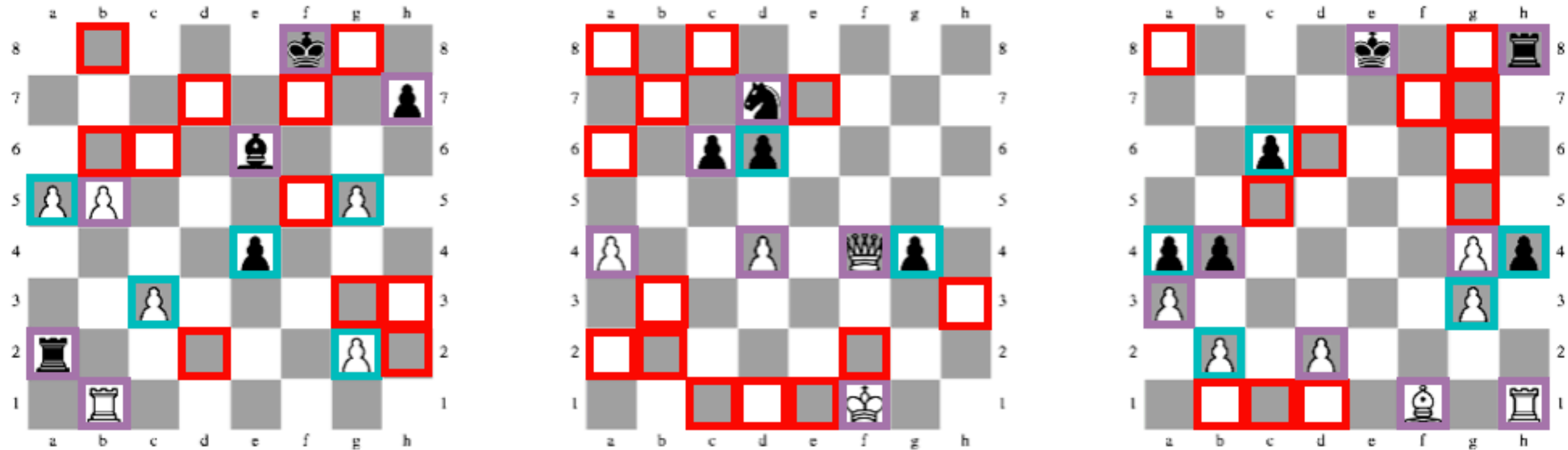


# reconstructions (random)

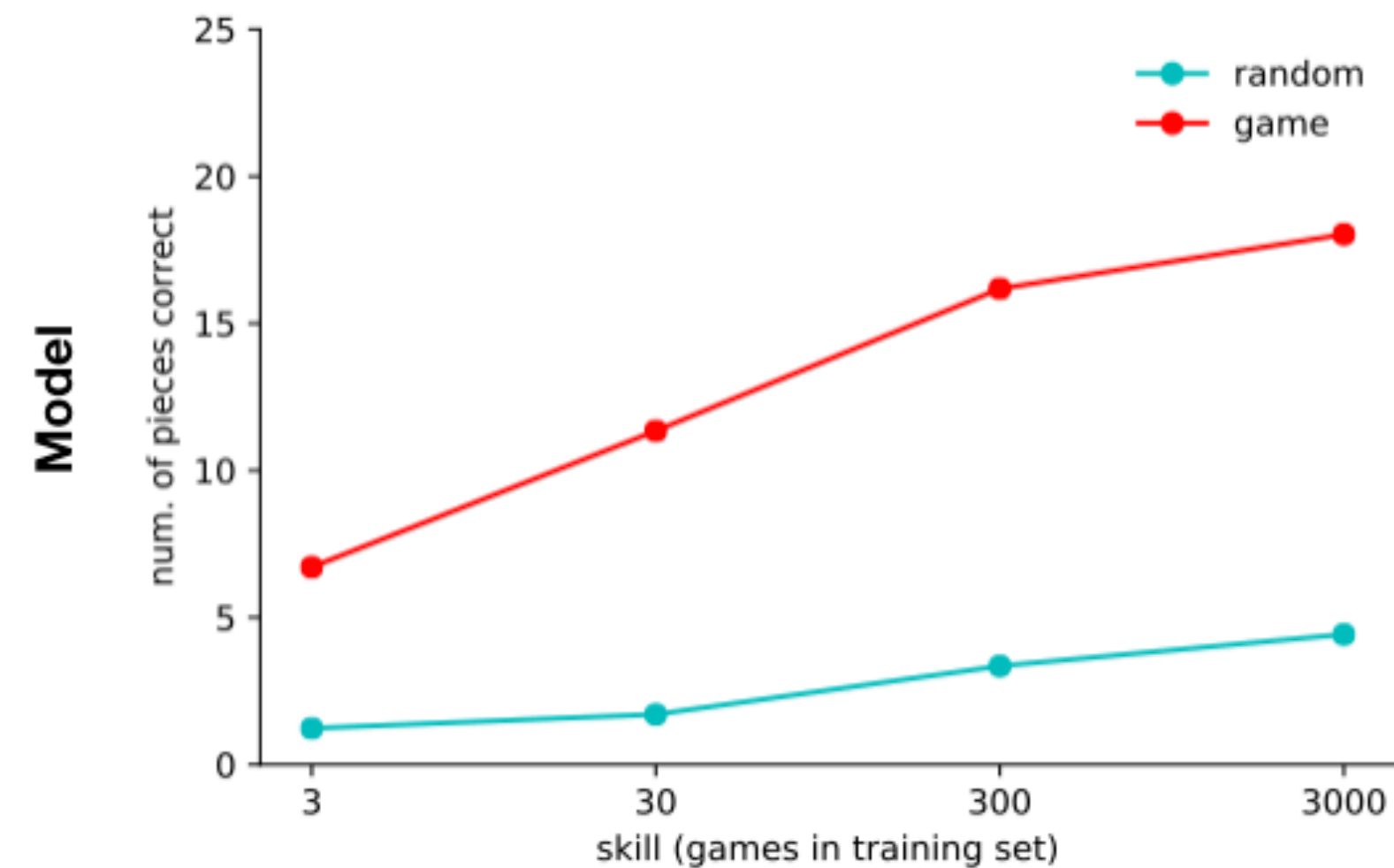
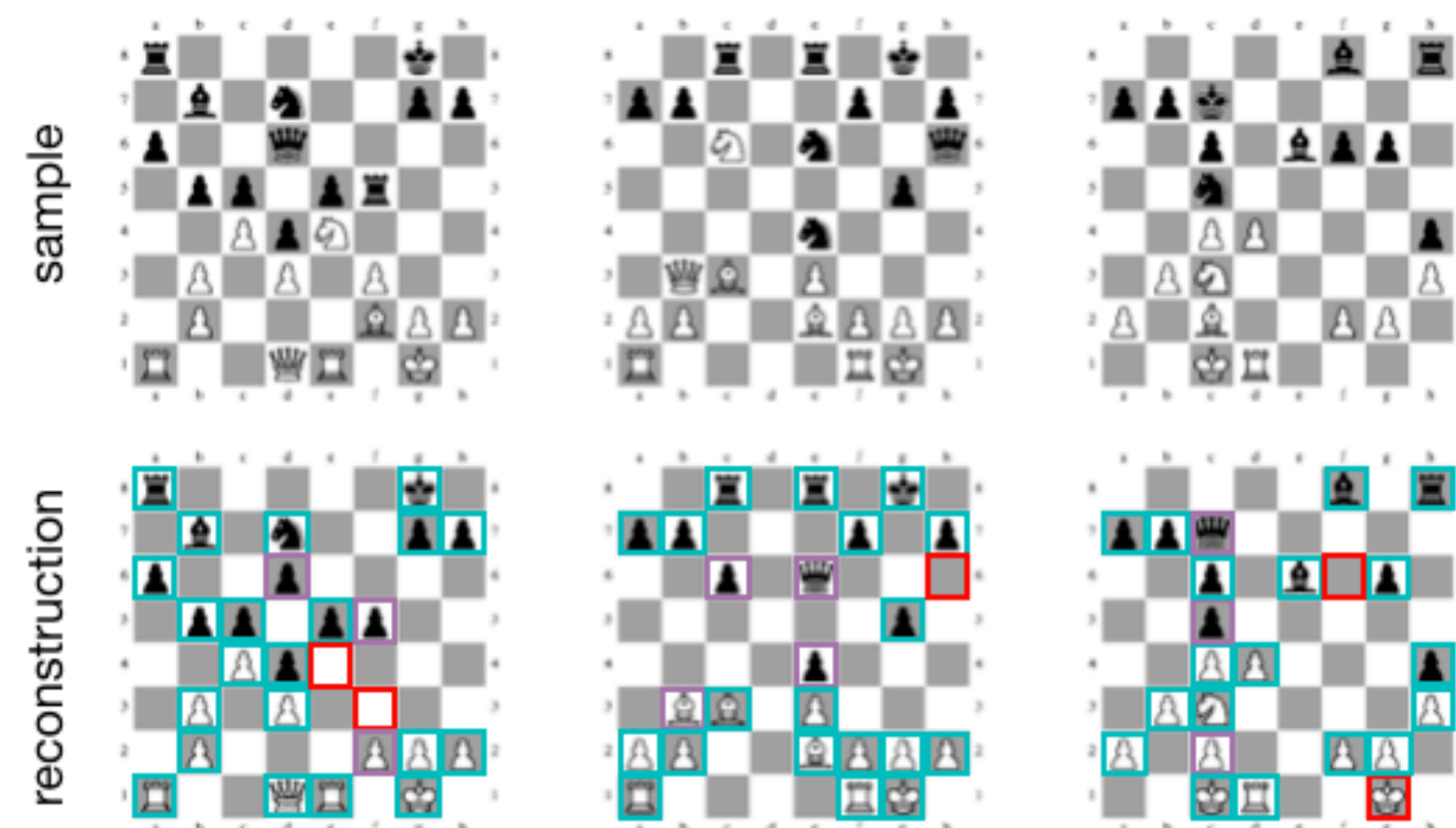
sample



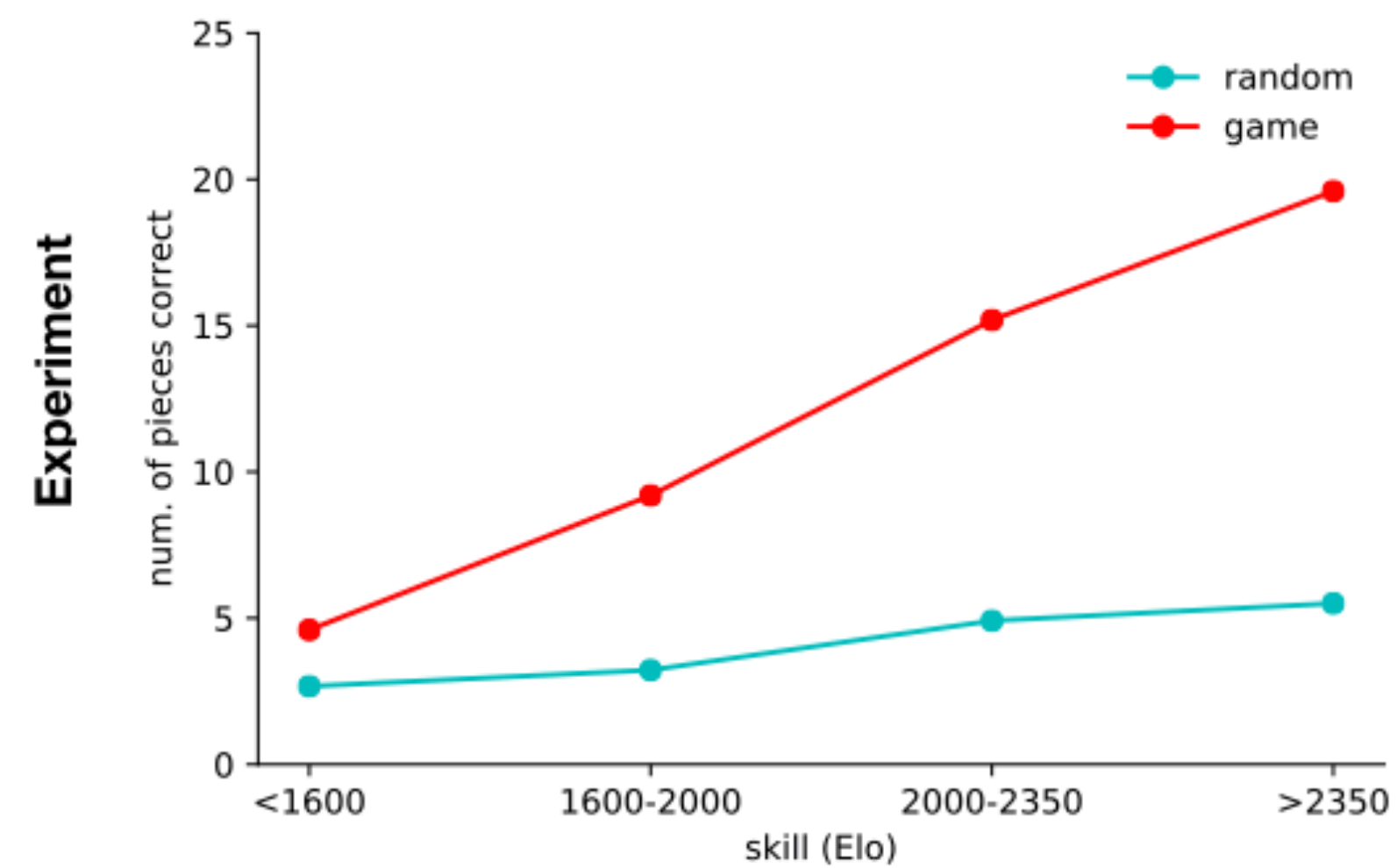
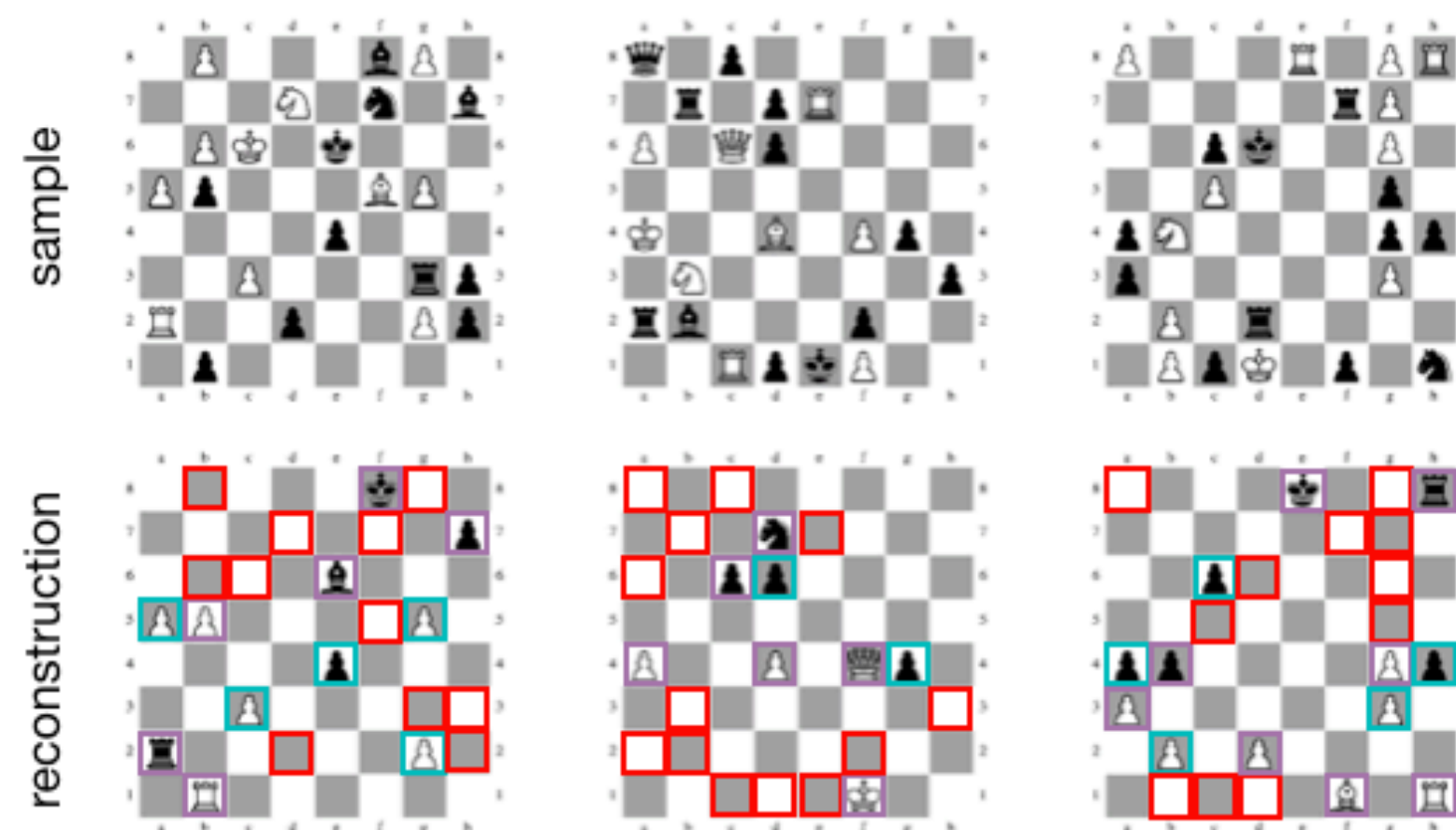
reconstruction



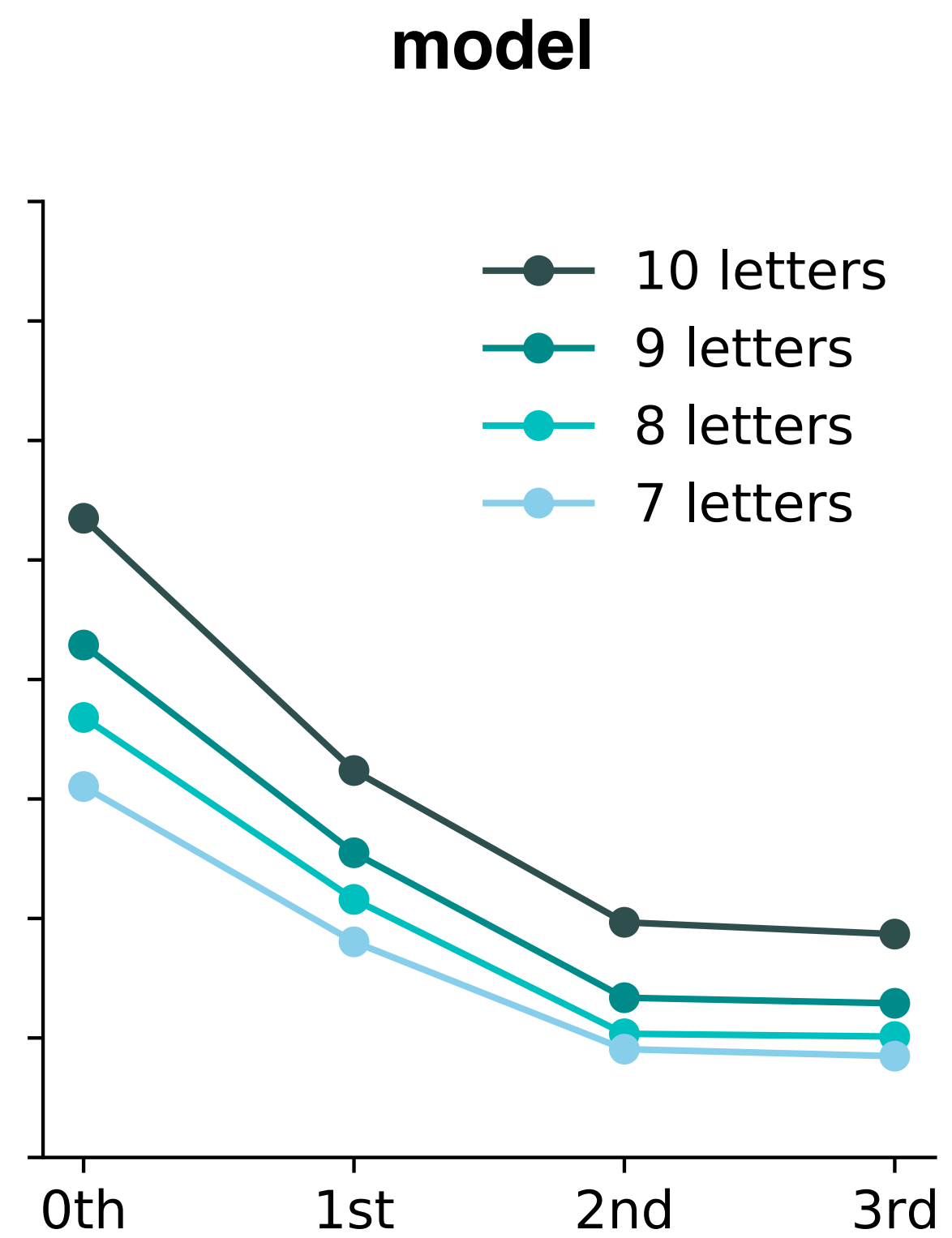
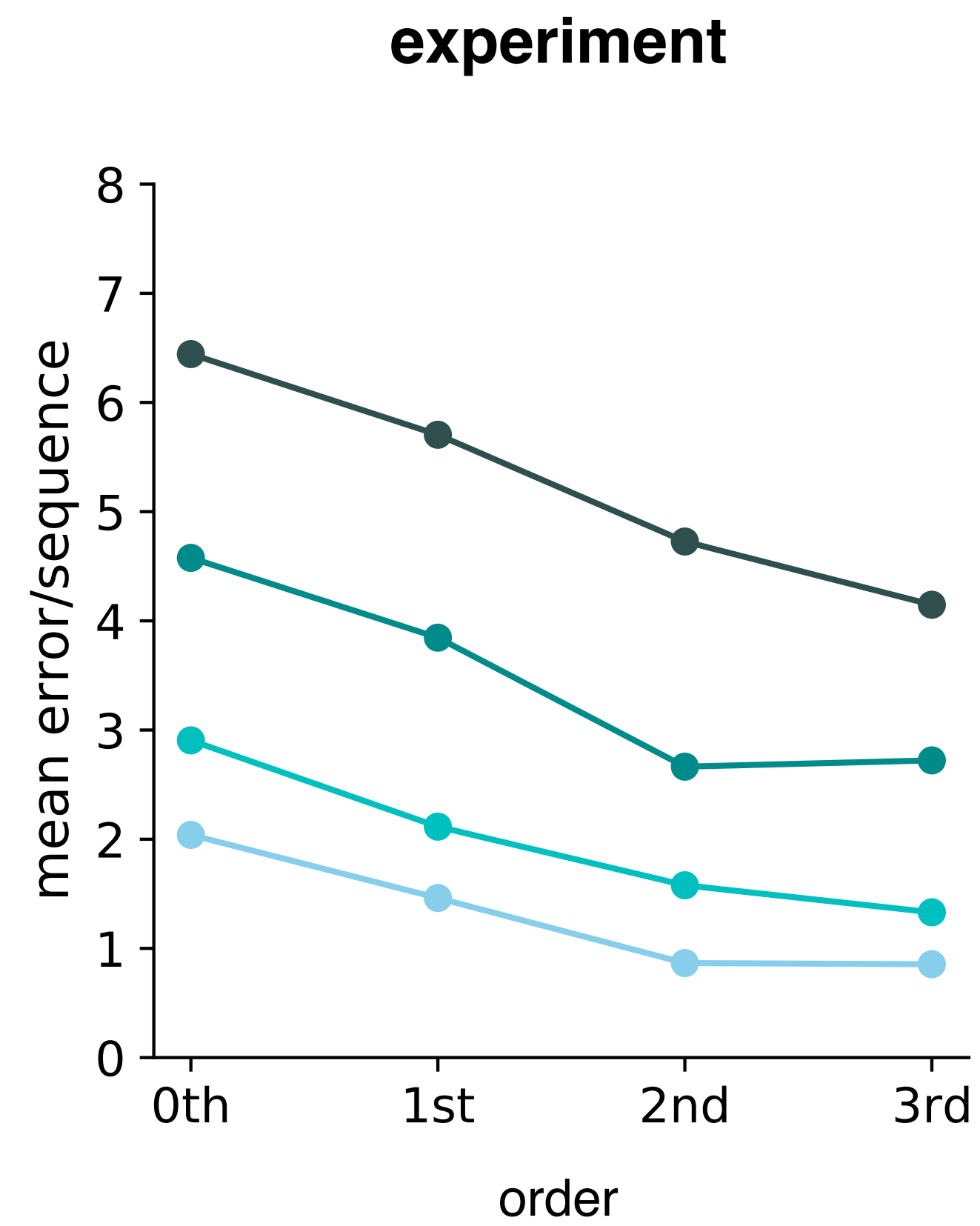
## game configurations



## random configurations



uniform	1st order	2nd order	3rd order	full word
RCIFODWVIL	TNEOOESHHE	HIRTOCLEN0	BETEREASYS	PLANTATION
GKTODKPENF	INOLGGOLVN	DOVEECOFOF	CRAGETTERS	FLASHLIGHT
TZXKHAWCCF	PDOASLOTPP	SESERAICCG	TOWERSIBLE	UNCOMMONLY
NGORHQIYWB	AEOCAOIAON	AREDAGORTZ	DEEMEREANY	ALIENATION
BVNJSYZXUA	IRC RENFCTN	CUNSIGOSUR	THERSERCHE	PICKPOCKET



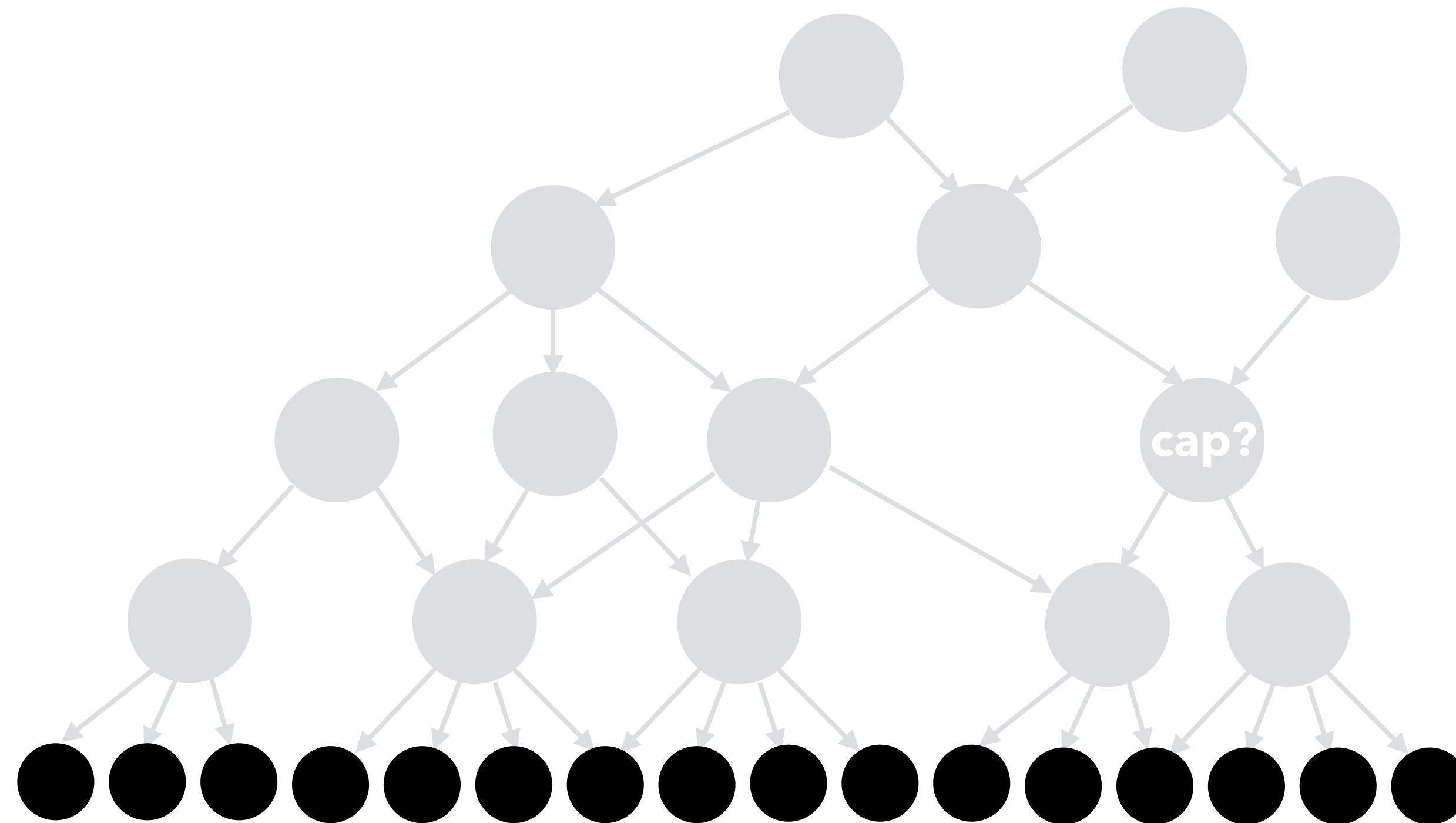
(experiment: Baddeley et al., 1971, model: Frater et al., 2022)

Consequence 2.

## **Gist-based errors**

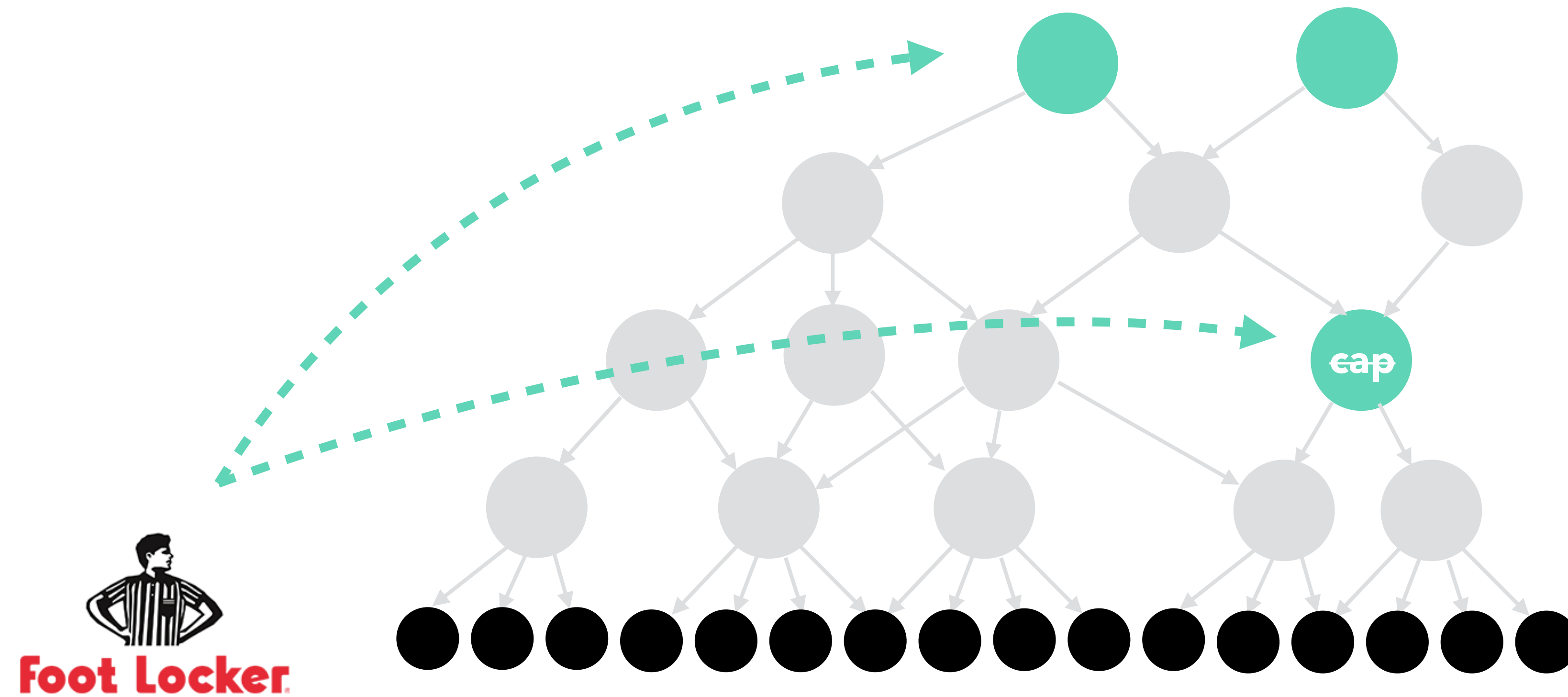
Features of the experience that were not stored in the memory trace will be sampled from the generative model, assigning values that could have been part of the observation with high probability.

# gist-based errors





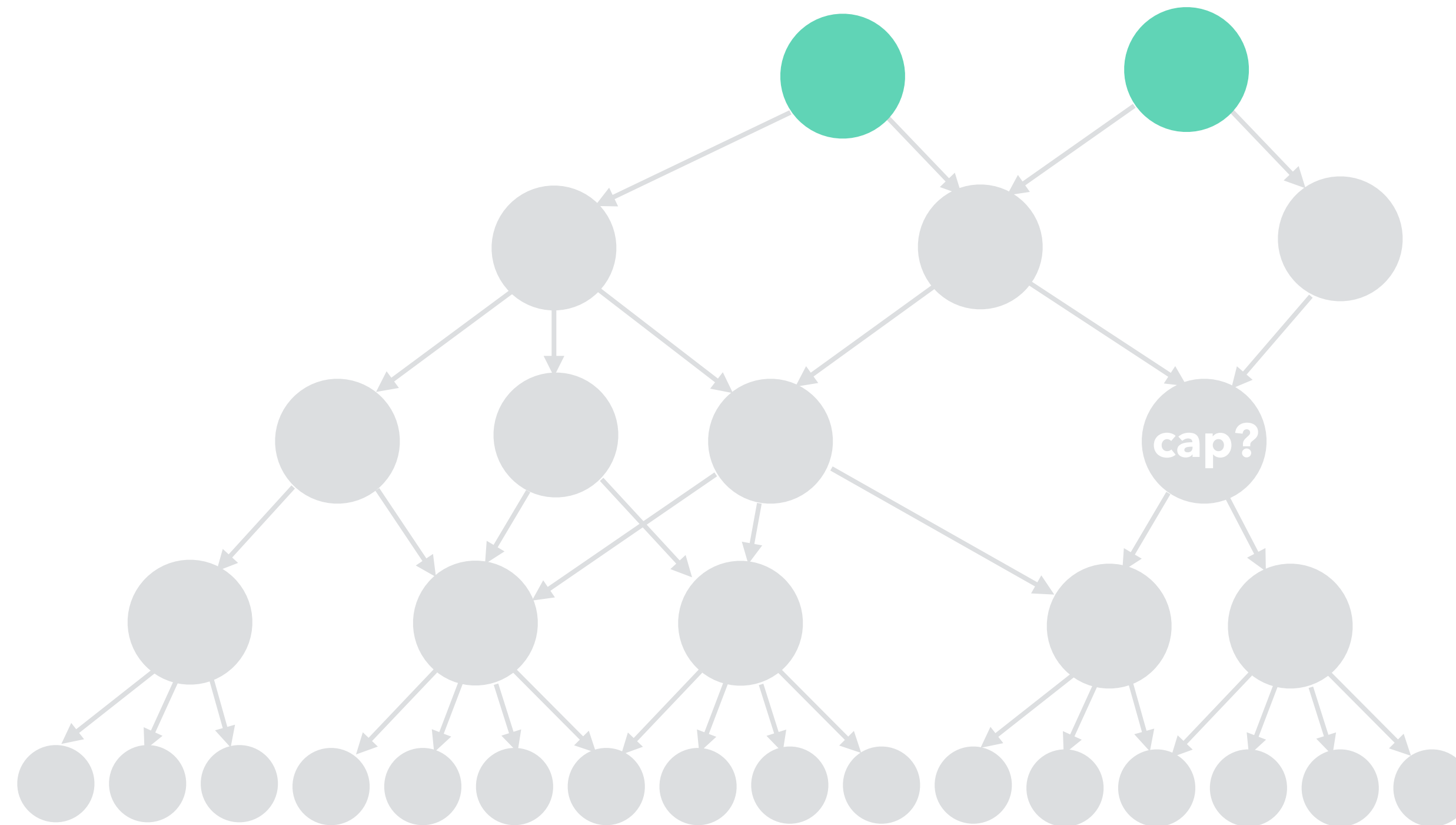
# gist-based errors





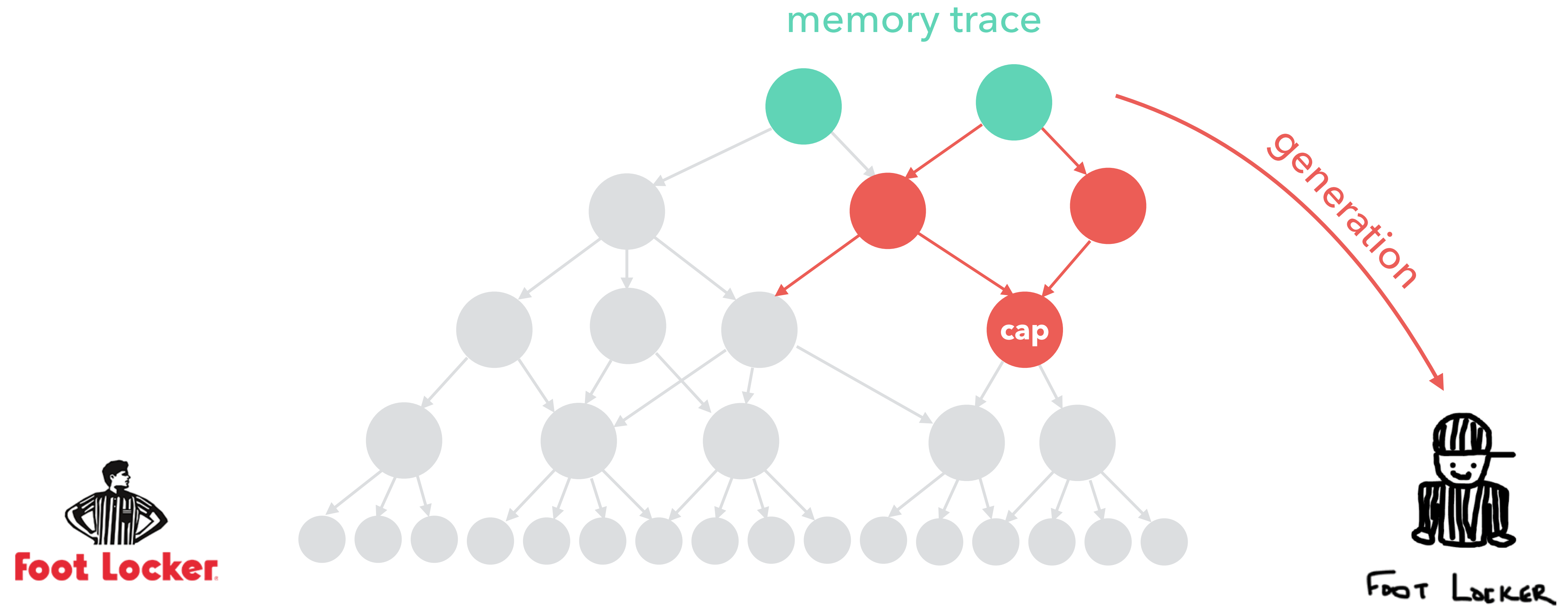
# gist-based errors

memory trace



**Foot Locker.**

# gist-based errors



**word lists**

drowsy

test

bed ?  
snore ?  
sleep ?  
aeroplane ?

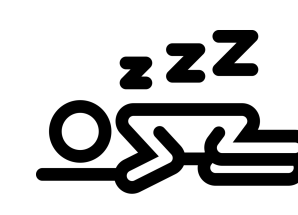
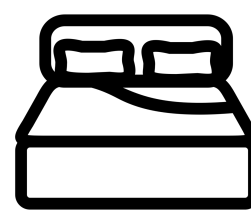
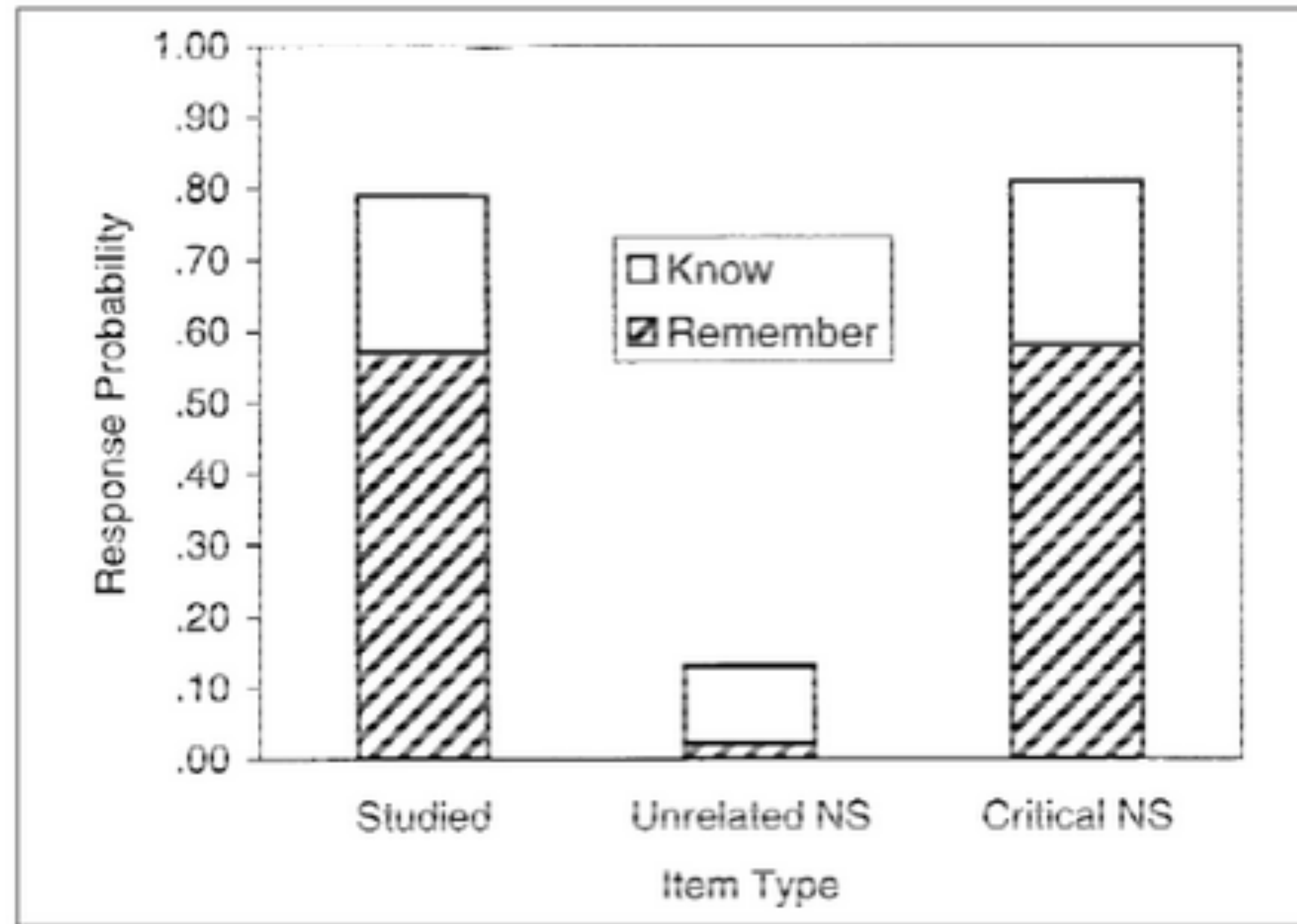
bed

snore

sleep

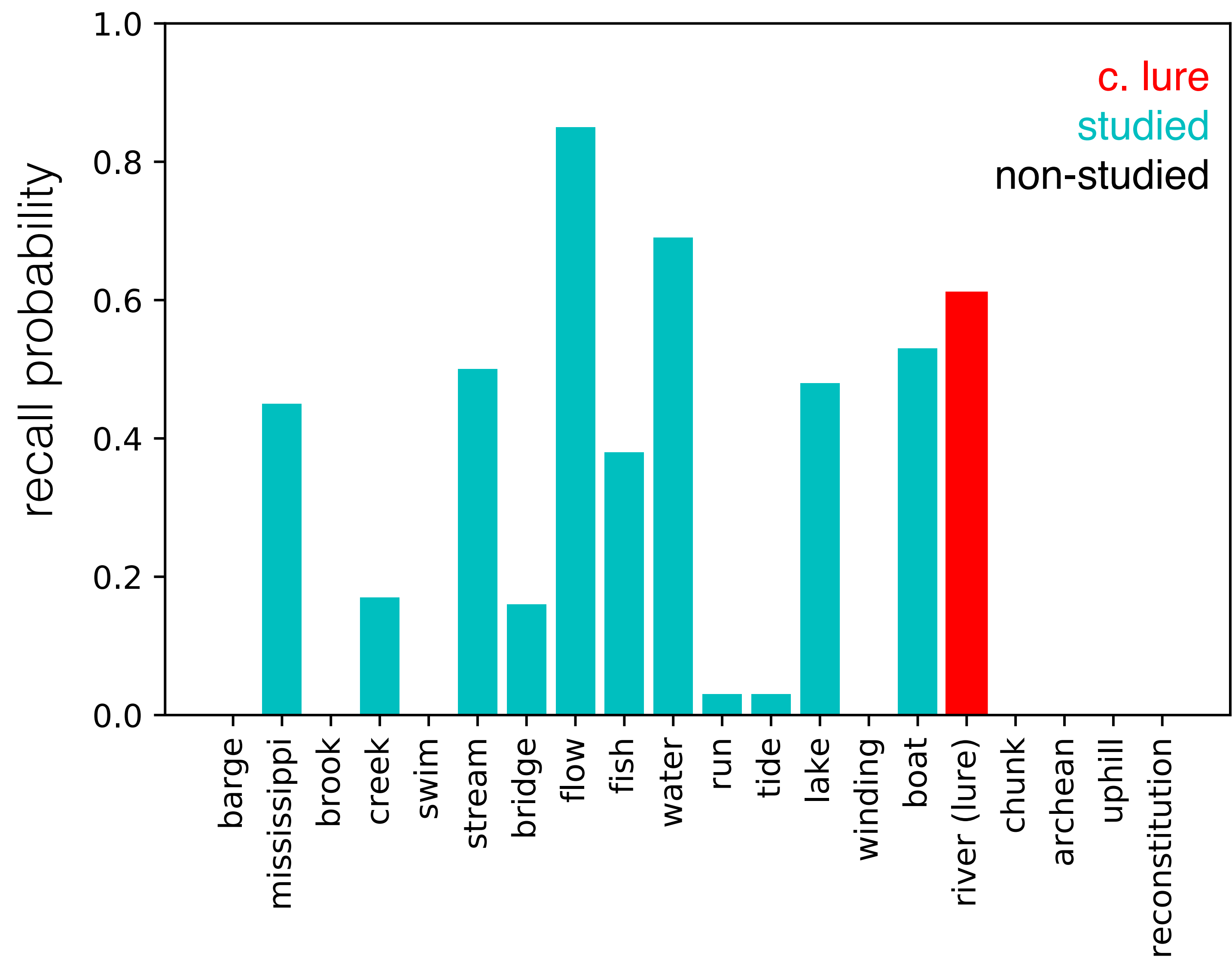
aeroplane

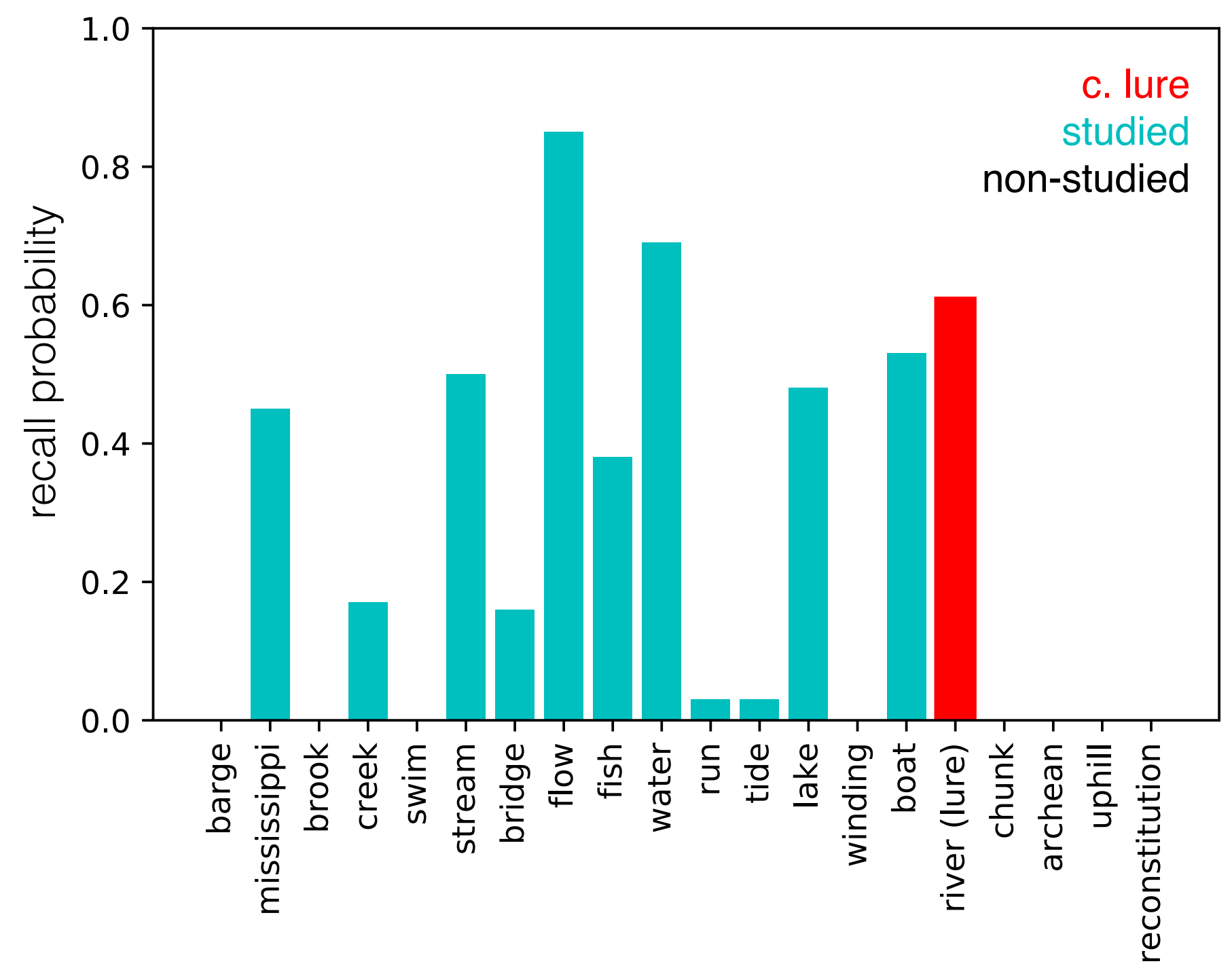
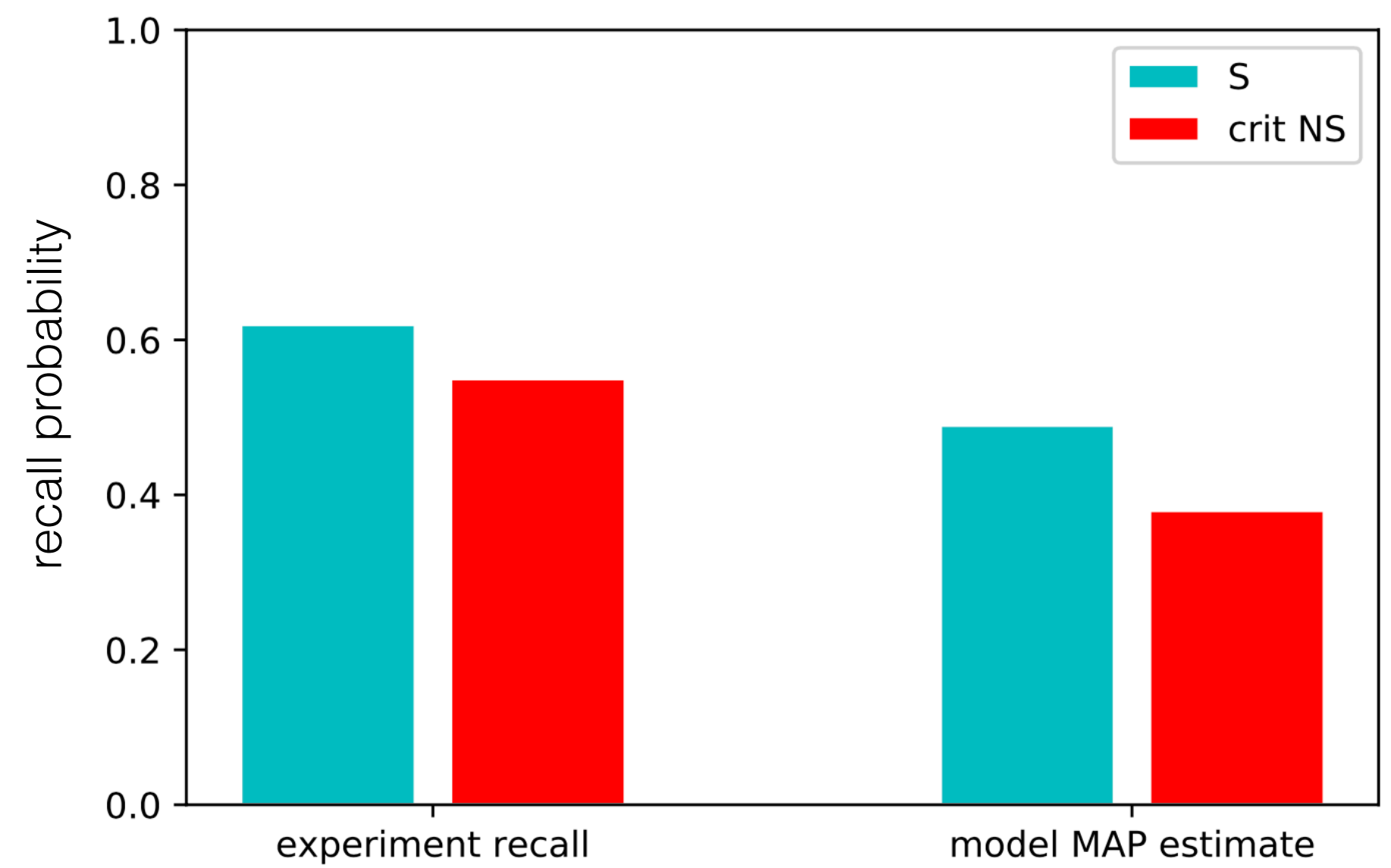
# DRM effect





# model reconstruction





## Subject B

drawing

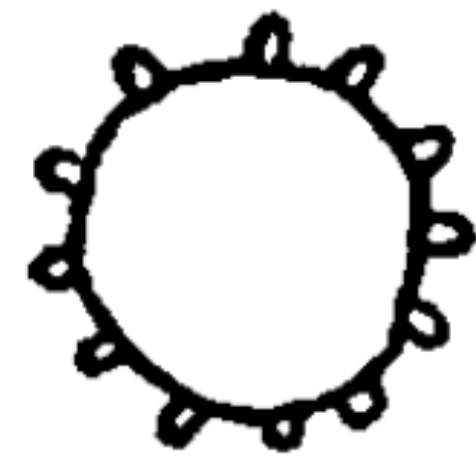
label

stimulus

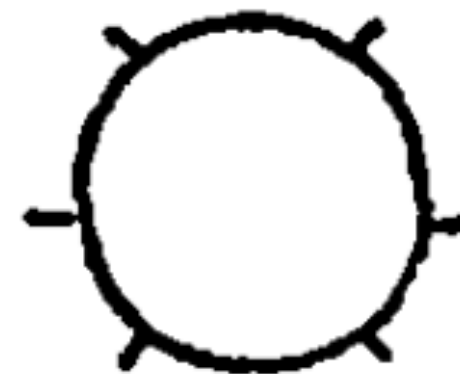
label

## Subject A

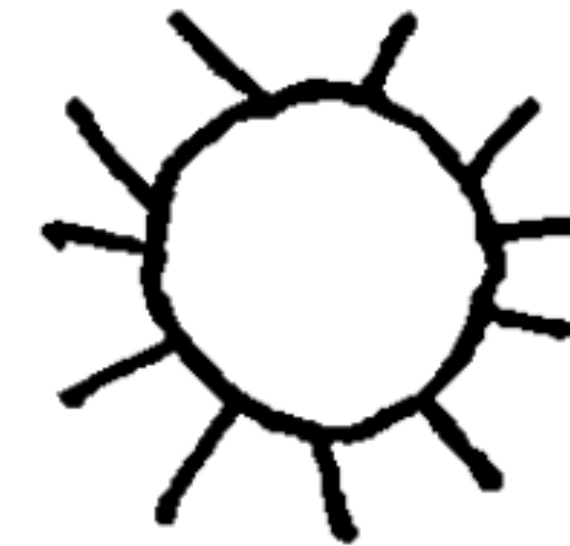
drawing



ship's  
wheel



sun



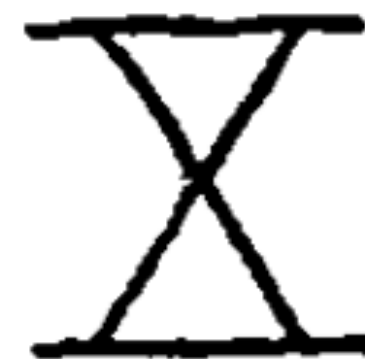
curtains in  
a window



diamond in  
a rectangle





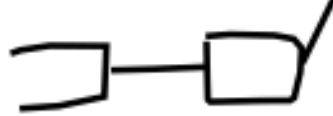
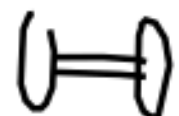
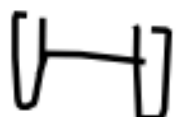
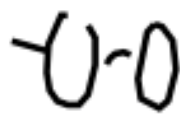
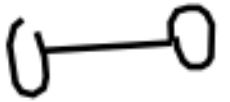
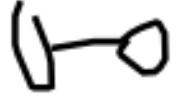
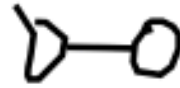
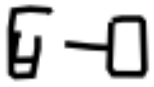
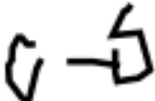


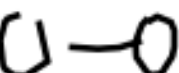
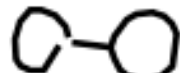


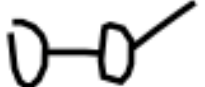
hourglass


















table



# Model

recall with label	input	recall with label
dumbbell 		eyeglasses 
		
		
		
		
		

# Experiment

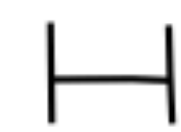
recall with label	input	recall with label
dumbbell 		eyeglasses 
table 		hourglass 
four 		seven 
broom 		gun 
ship's wheel 		sun 

**input**                      **recall with label**

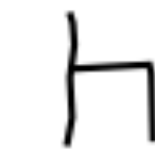
chair



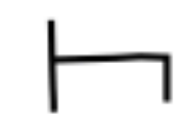
bed



chair



bed



wheel



fan



moon



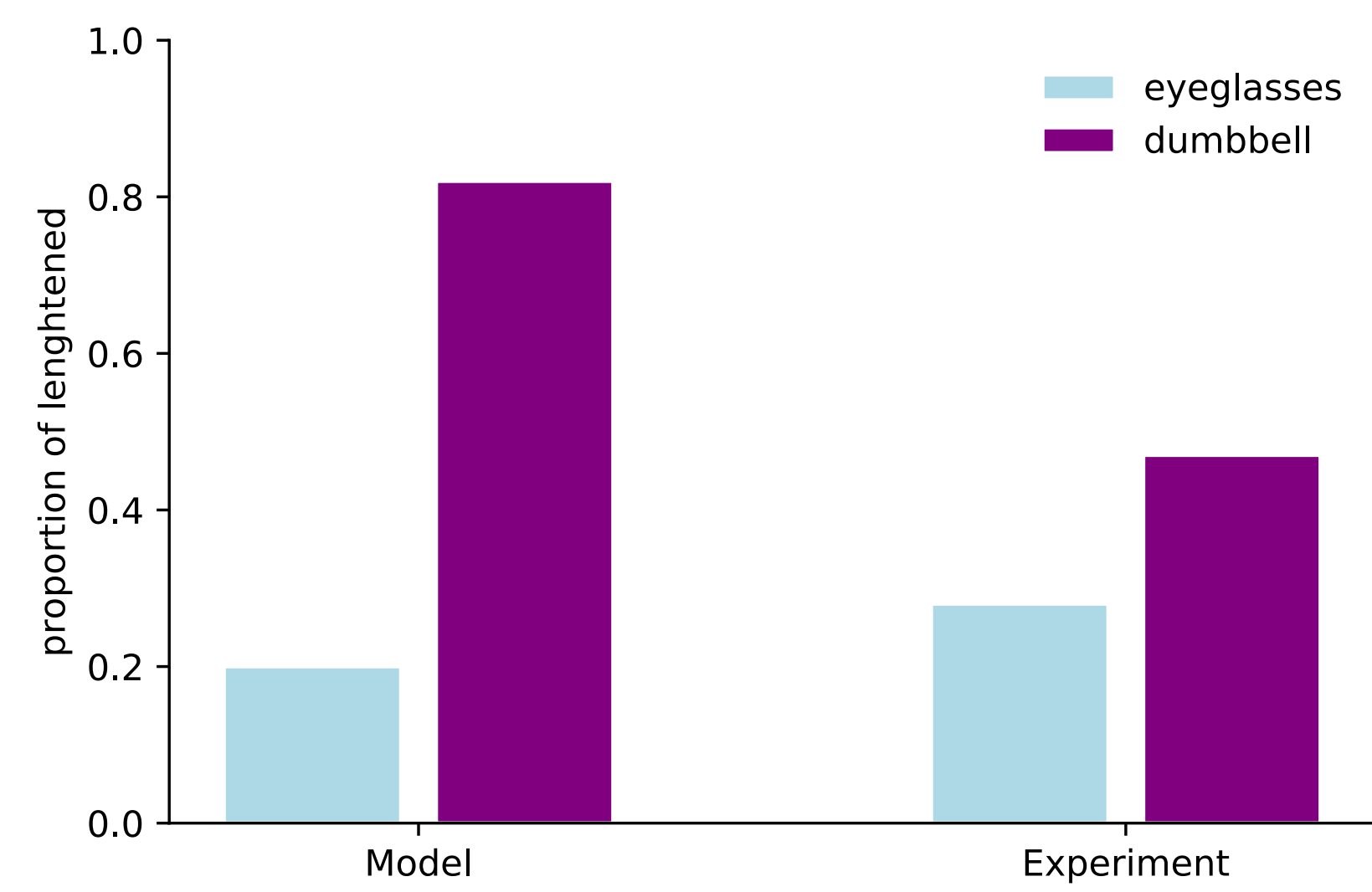
banana



pizza



wheel

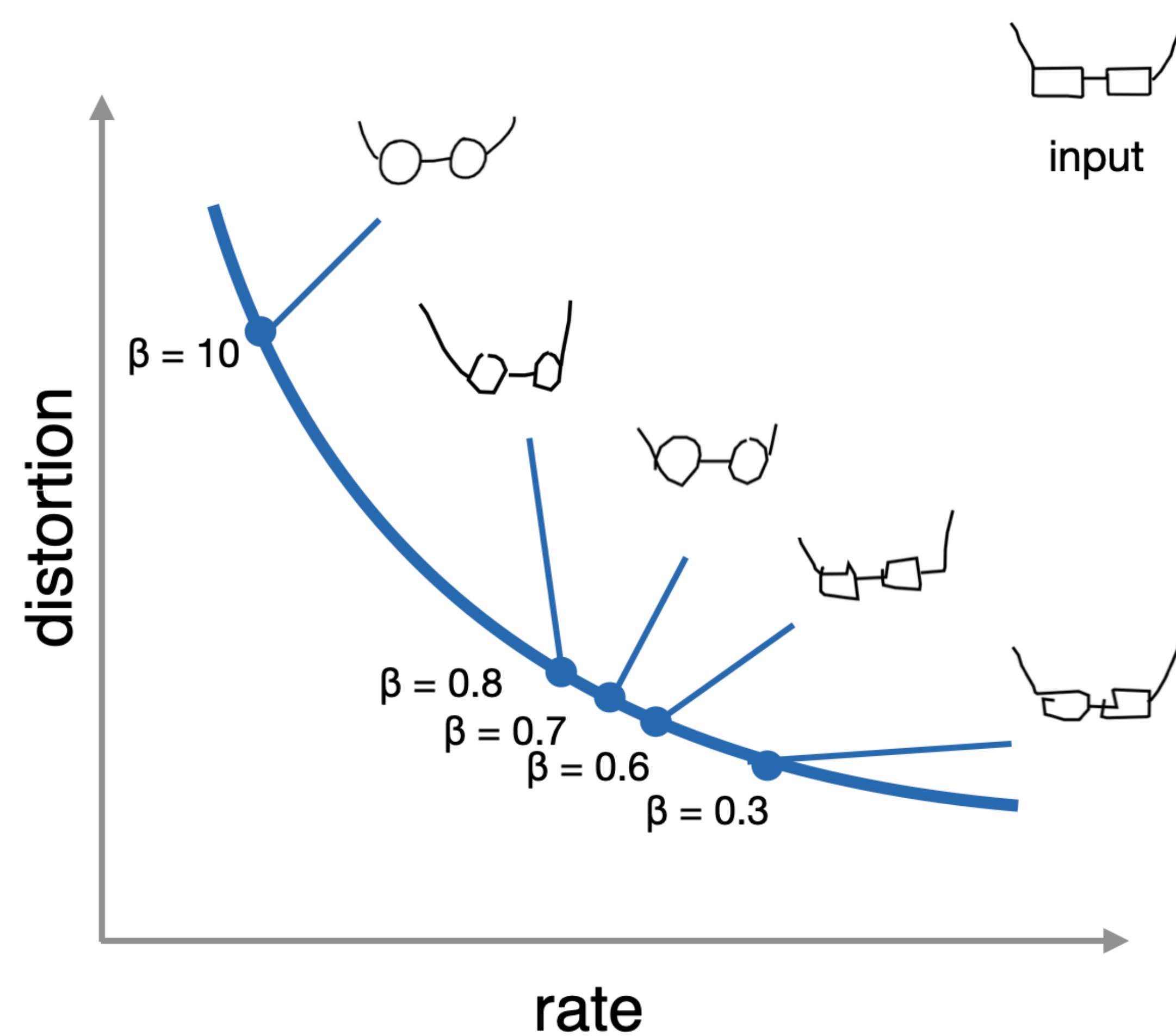


Consequence 3.

## **Rate distortion tradeoff**

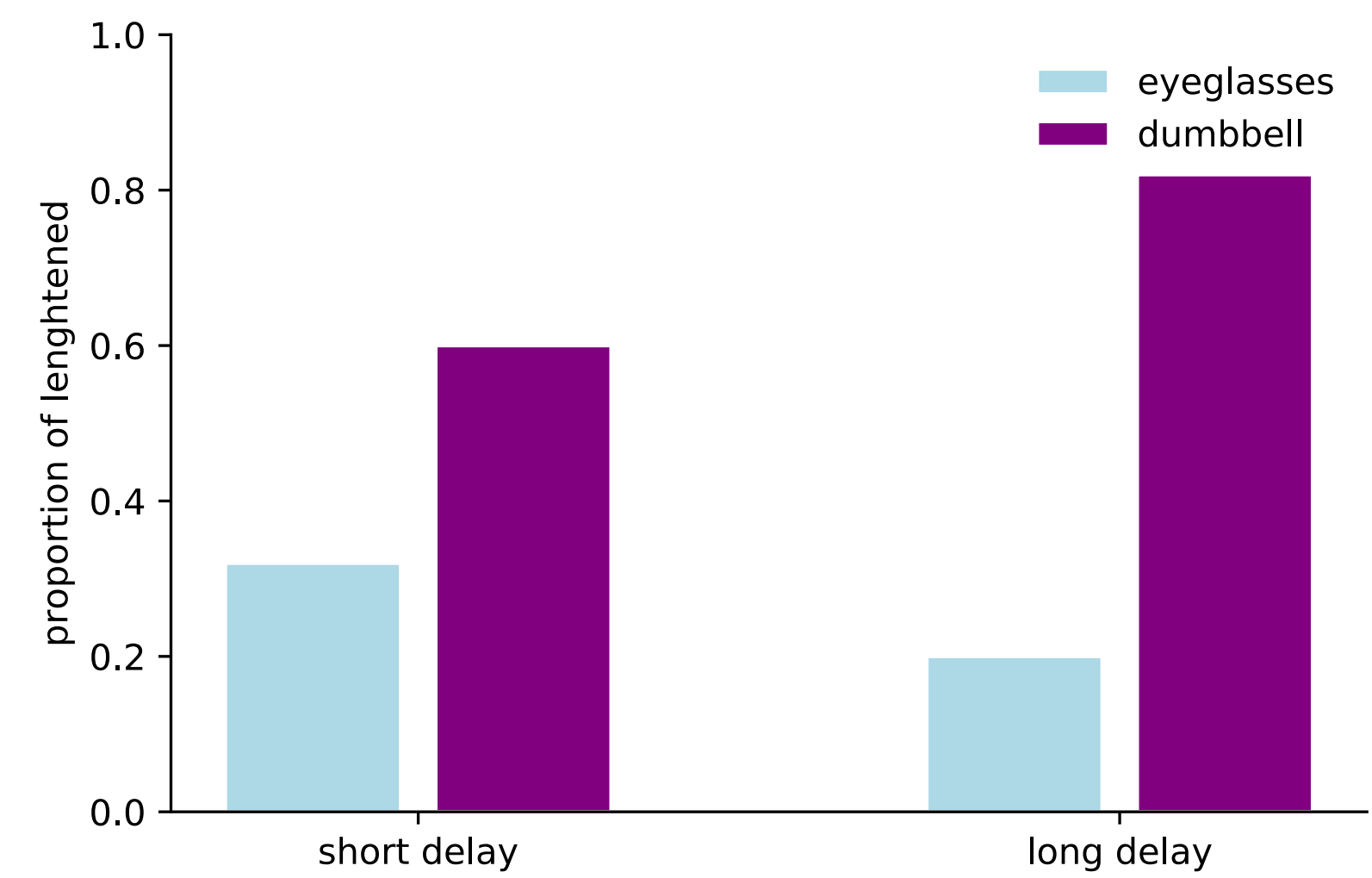
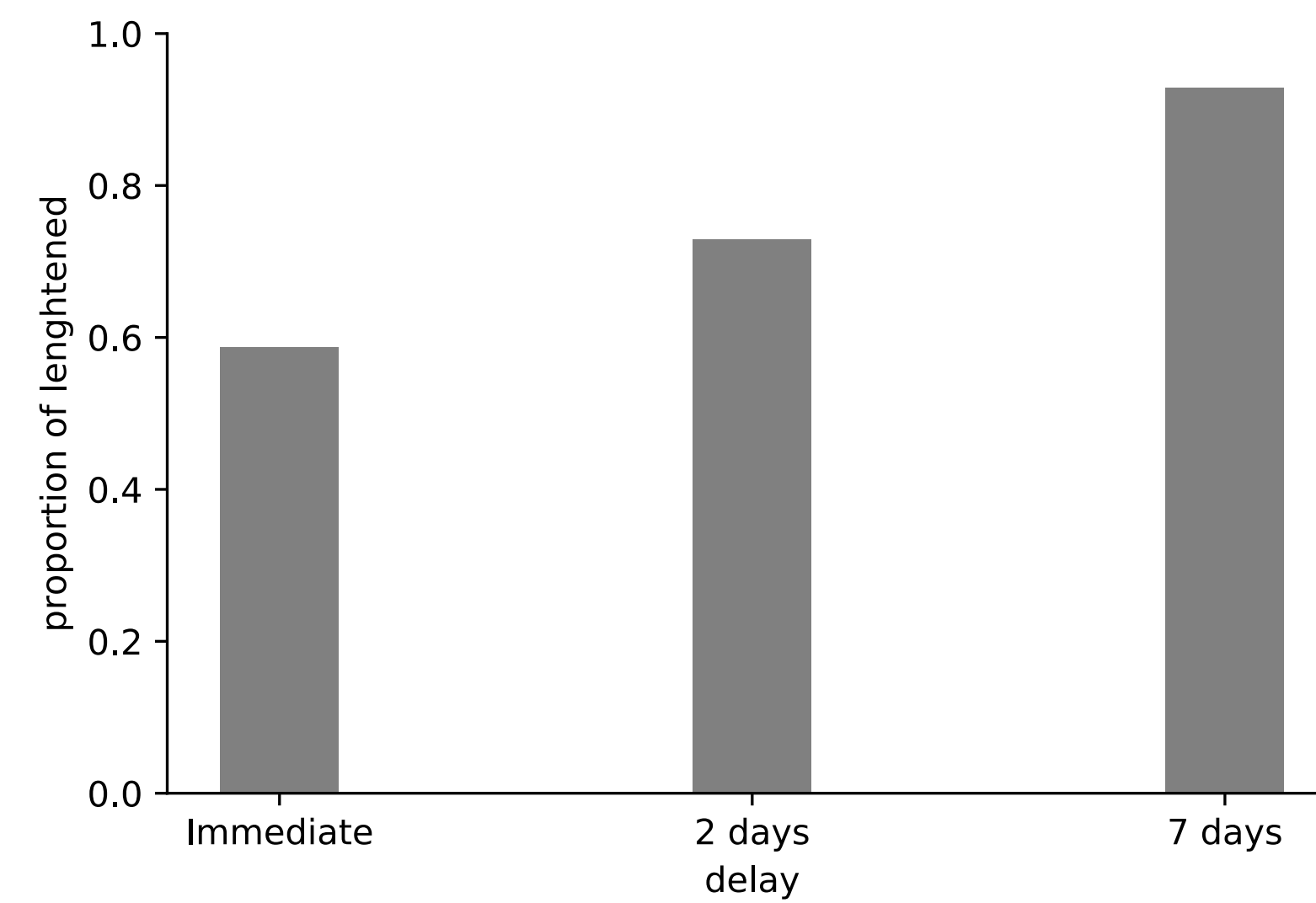
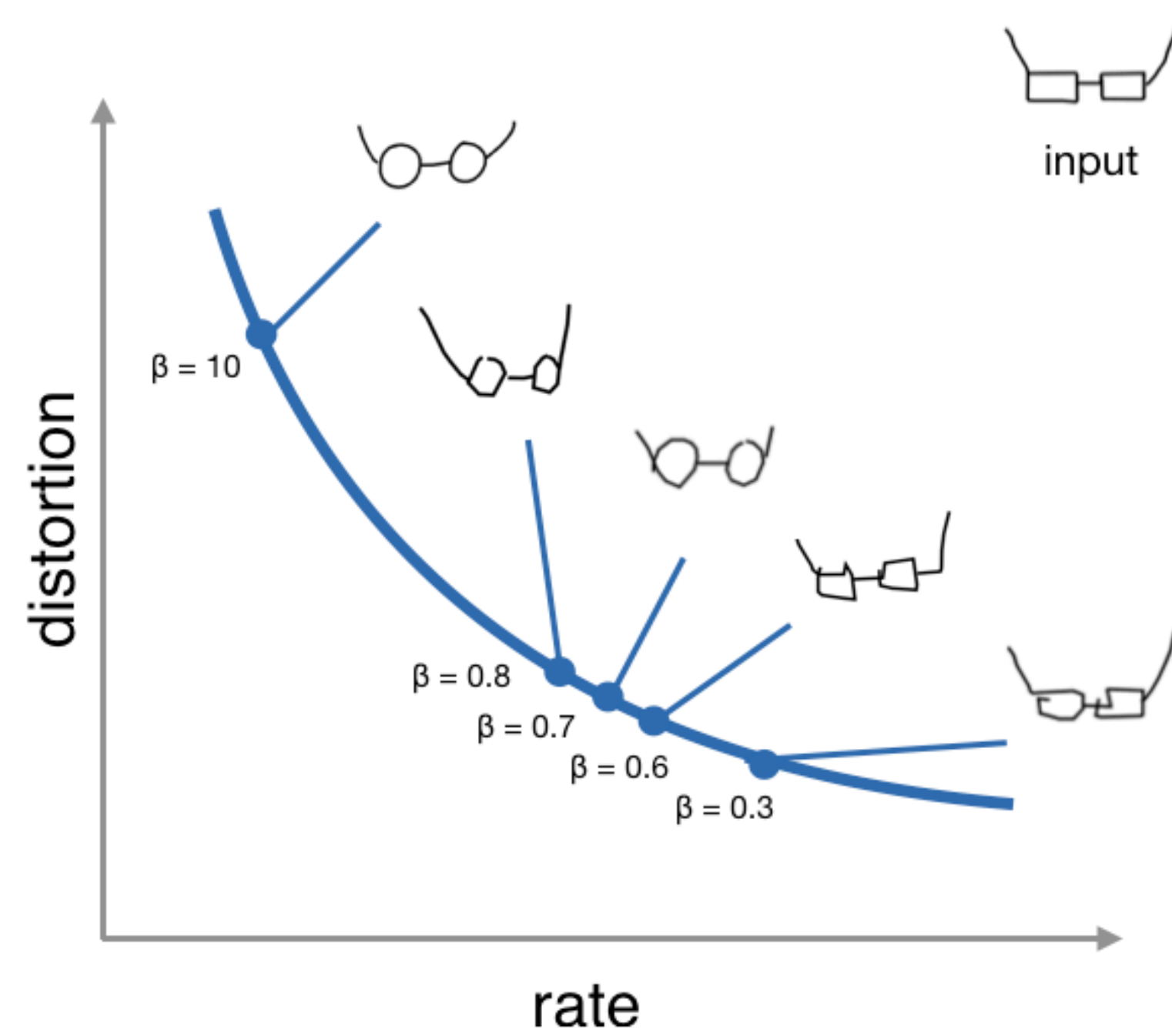
Available memory resources are unlikely to be constant as a function of time. Information theory provides a principled way of discarding information so that memories degrade gracefully.

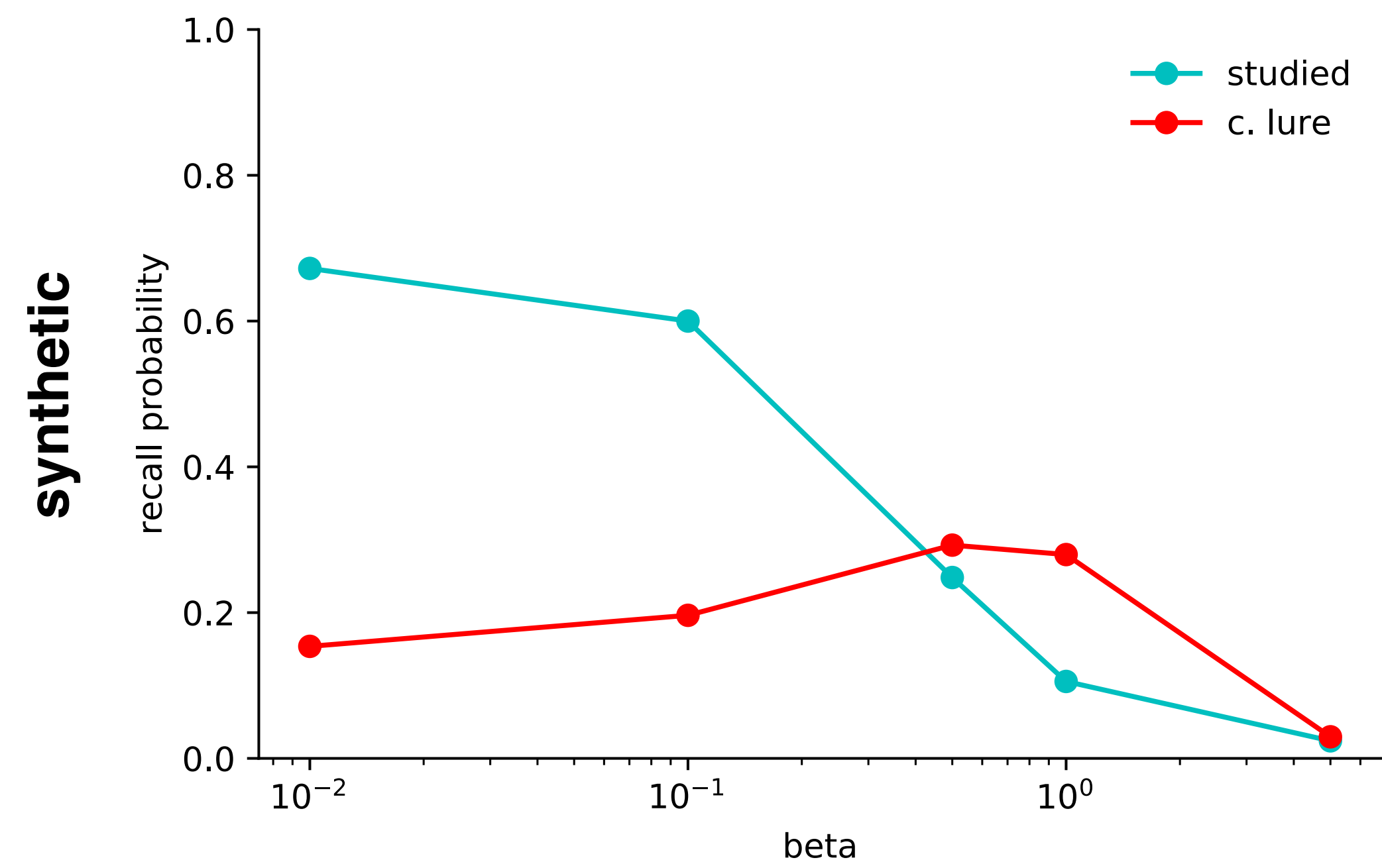
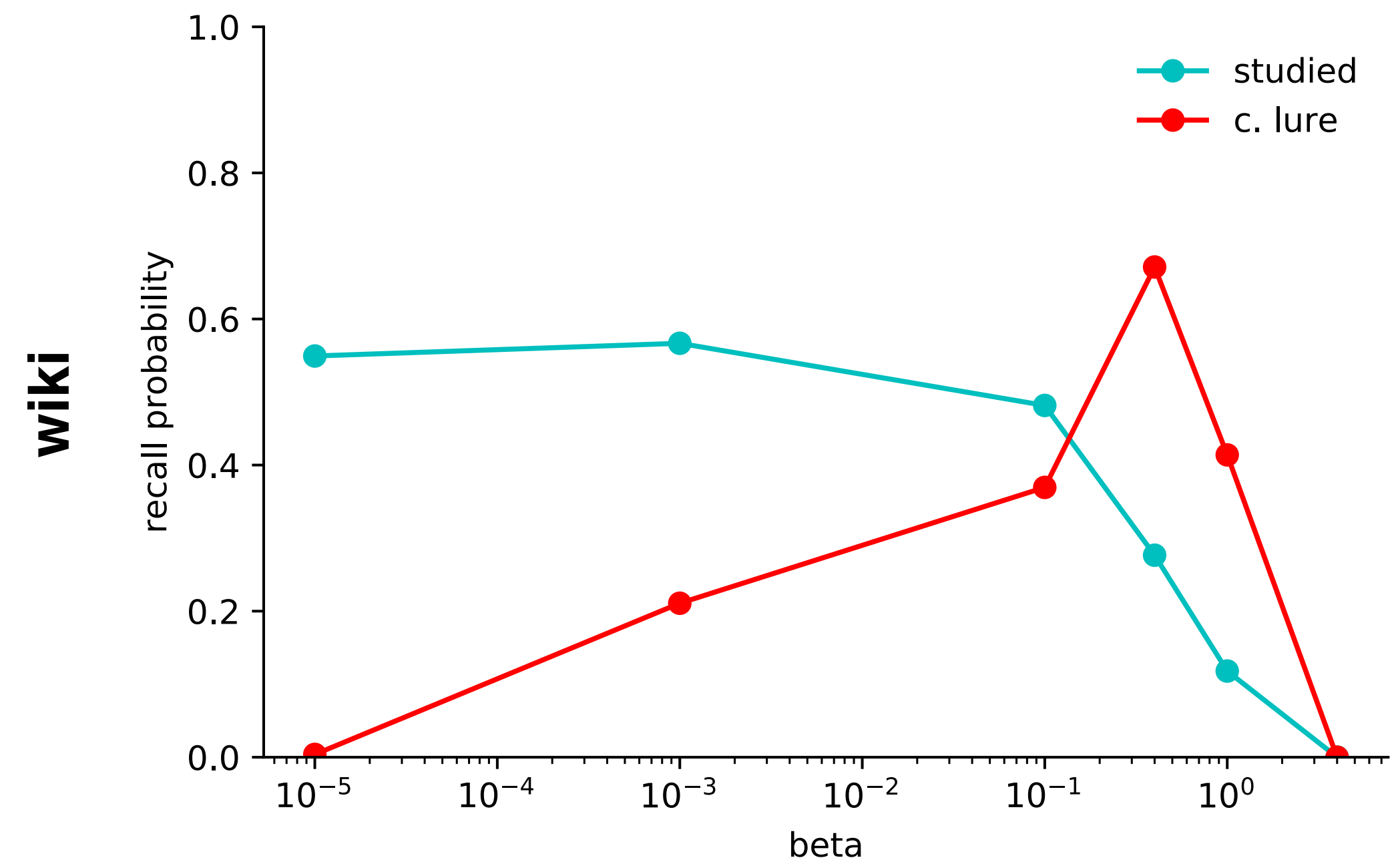


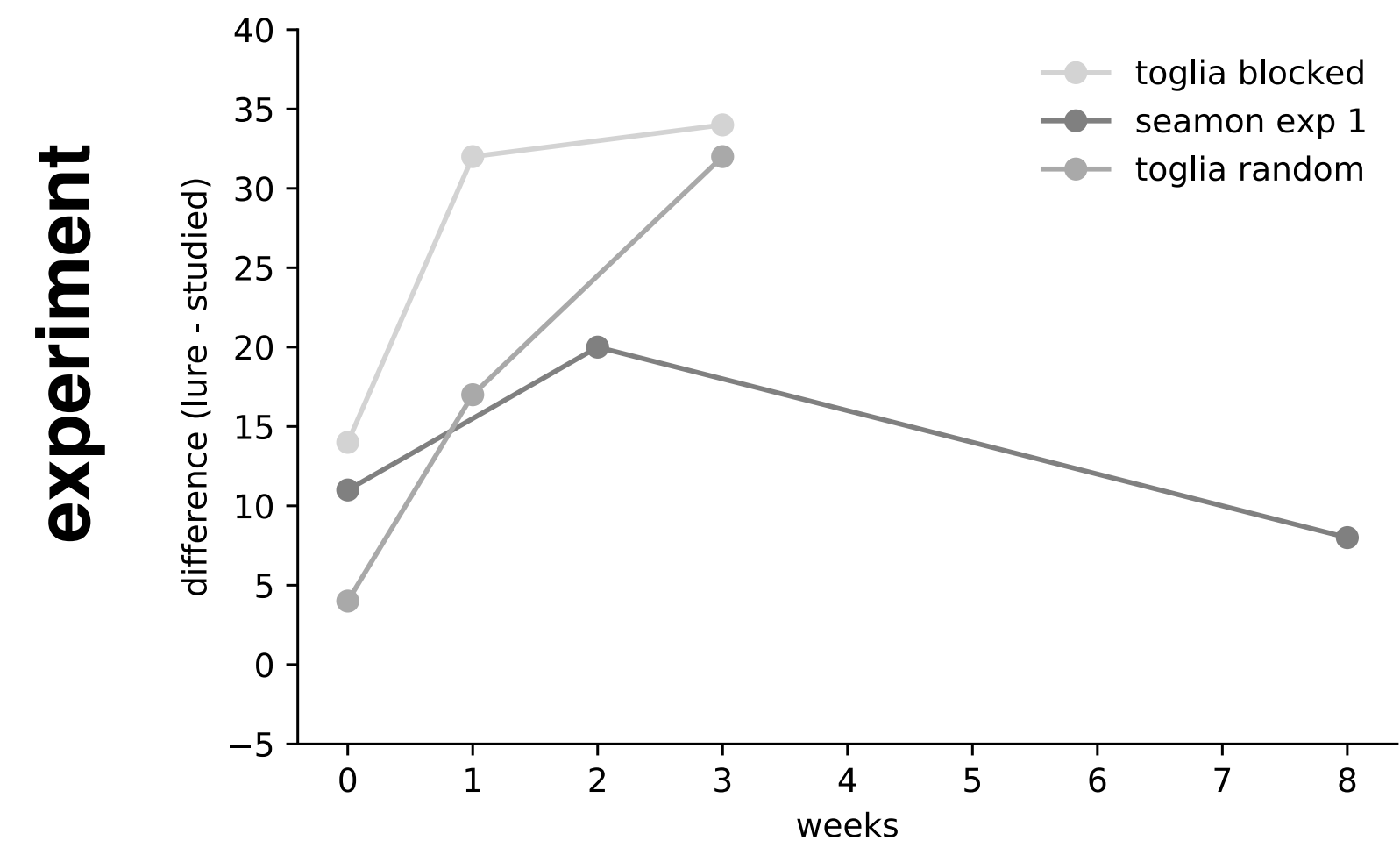
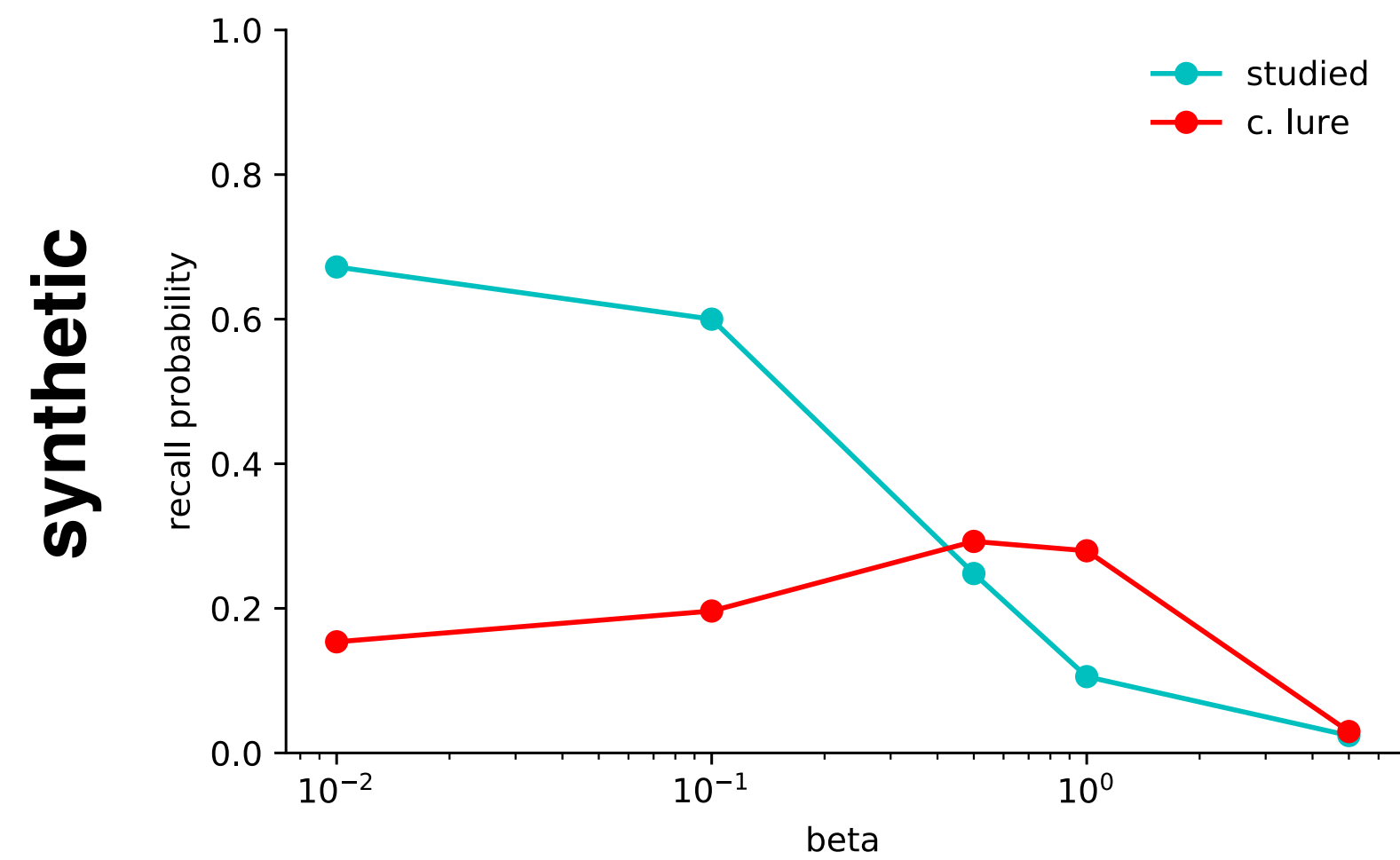
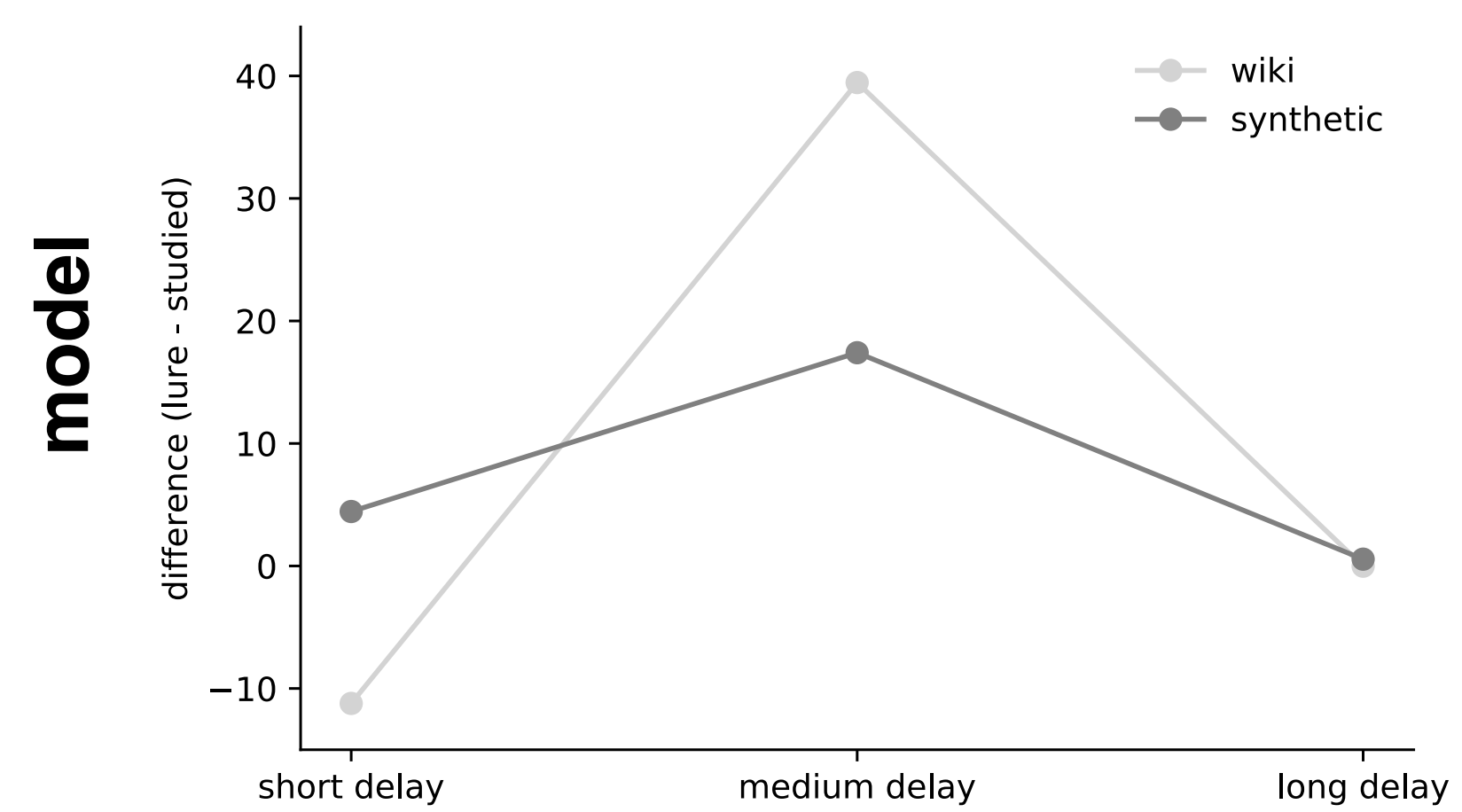
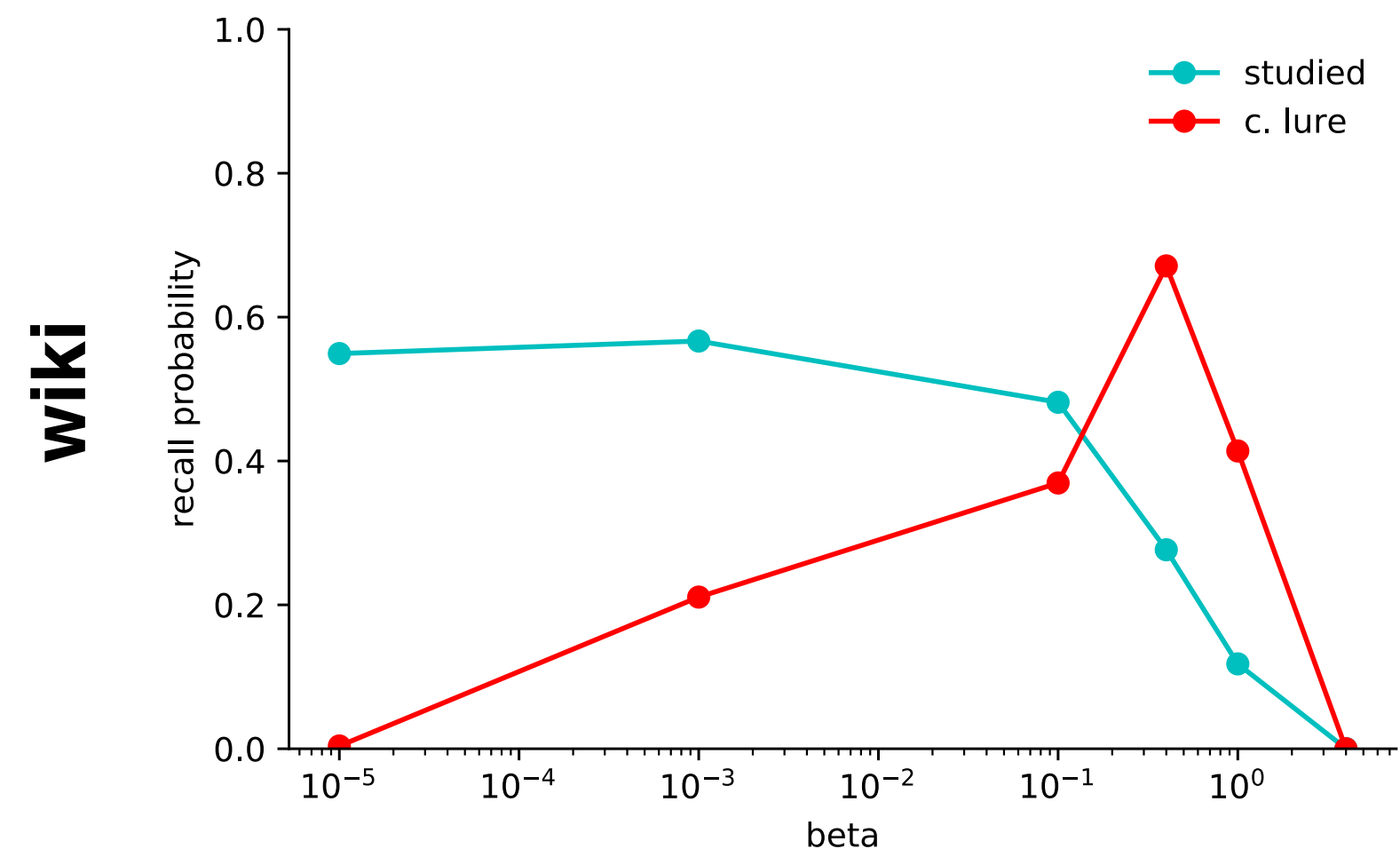


← less information  
(semantic,  
gist-like)

more information  
(episodic,  
verbatim-like) →







## studied

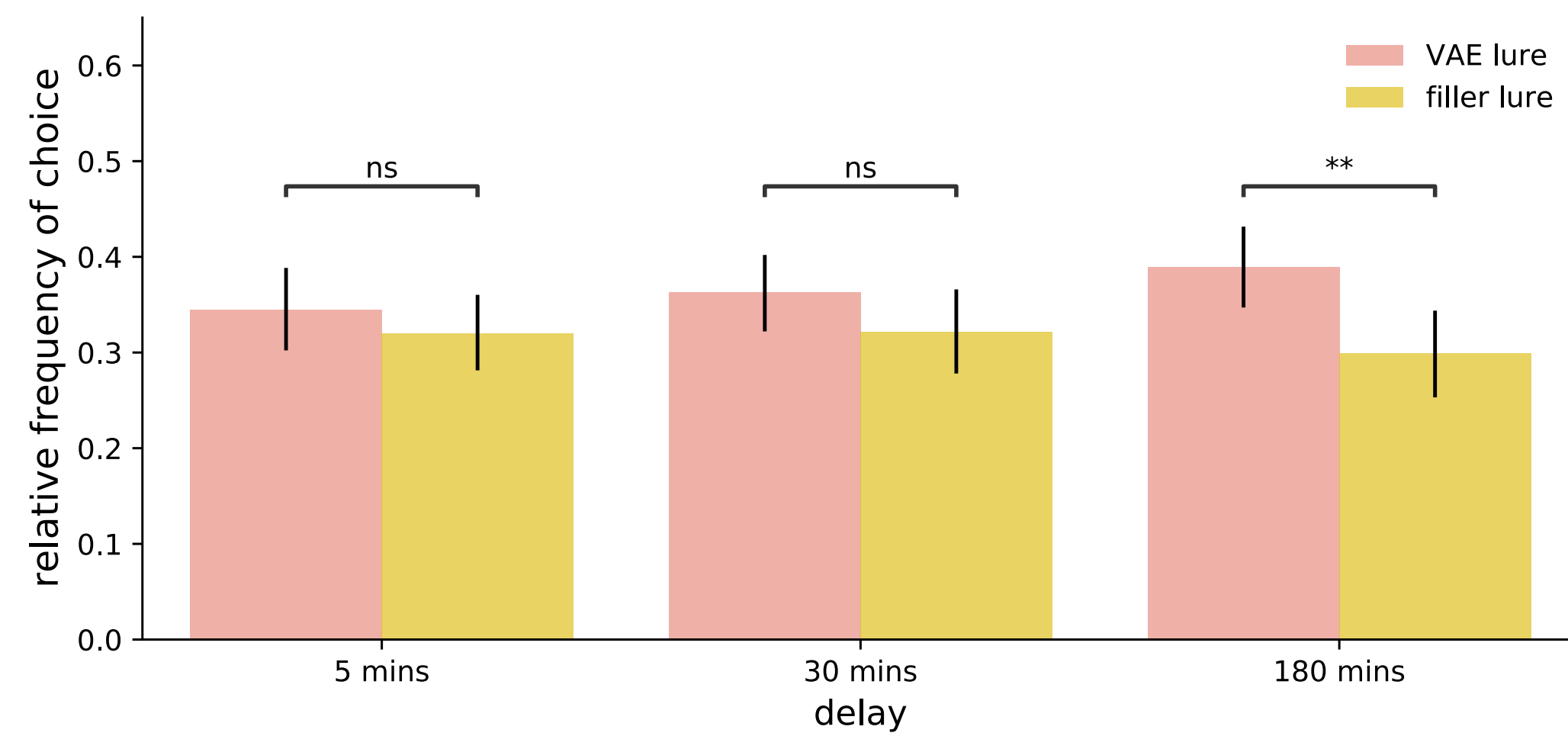
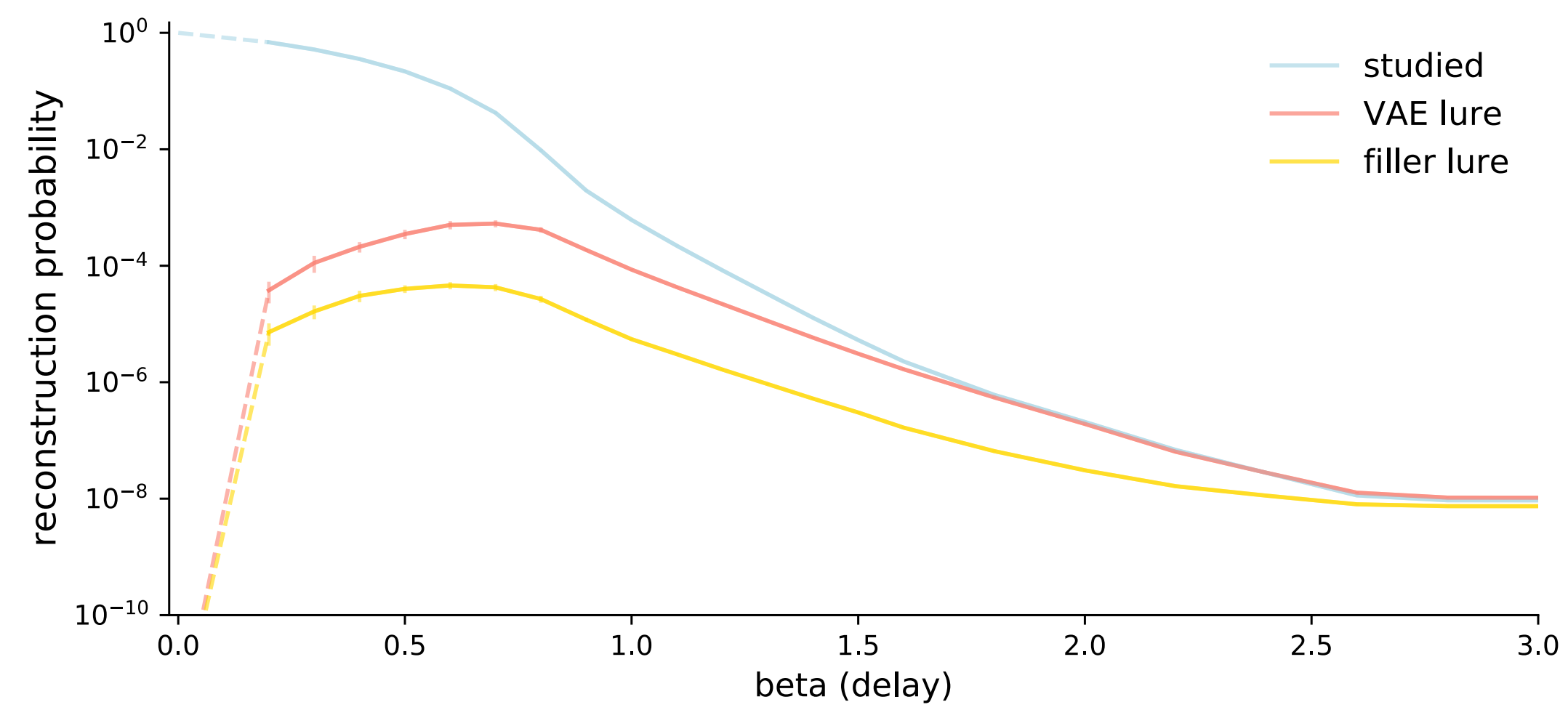
gonsing  
chignin  
emingli  
resibir  
briough  
dingran  
stingdo  
penitfu  
conchin  
plentri

## VAE lure

genring  
chiniin  
emiaghi  
resiuur  
broouth  
dingdon  
shinado  
penstqu  
cancrin  
plebari

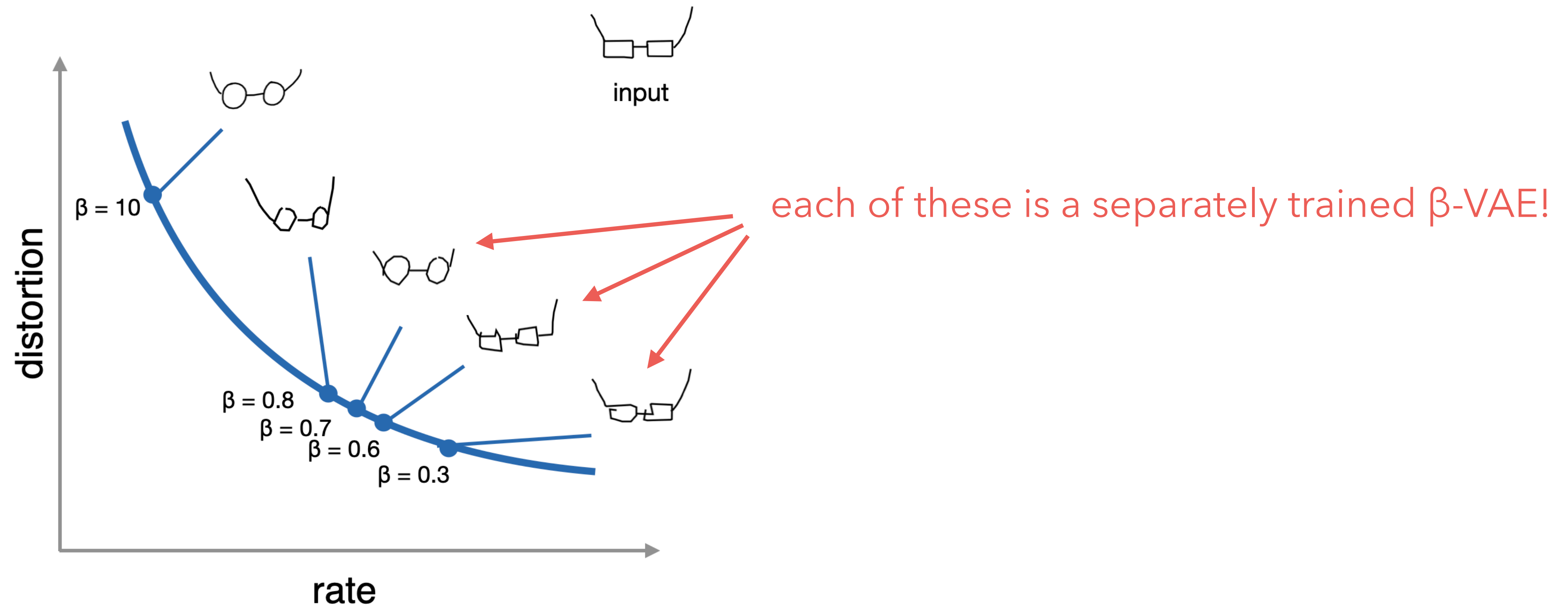
## filler lure

goesrng  
chegntn  
ersngli  
rxaibir  
brisdgh  
doneran  
stongeo  
pesinfu  
comshin  
pnentii

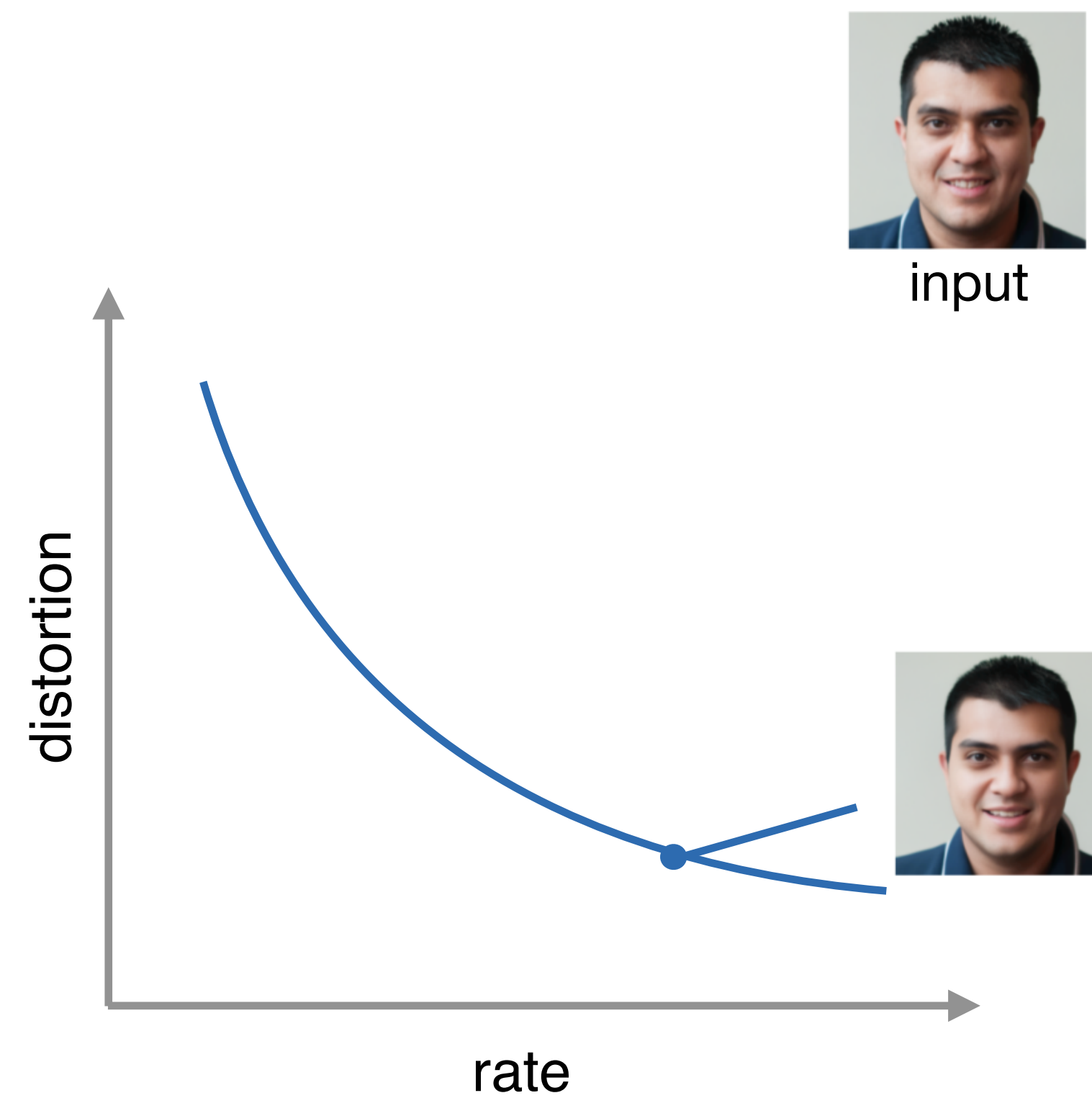
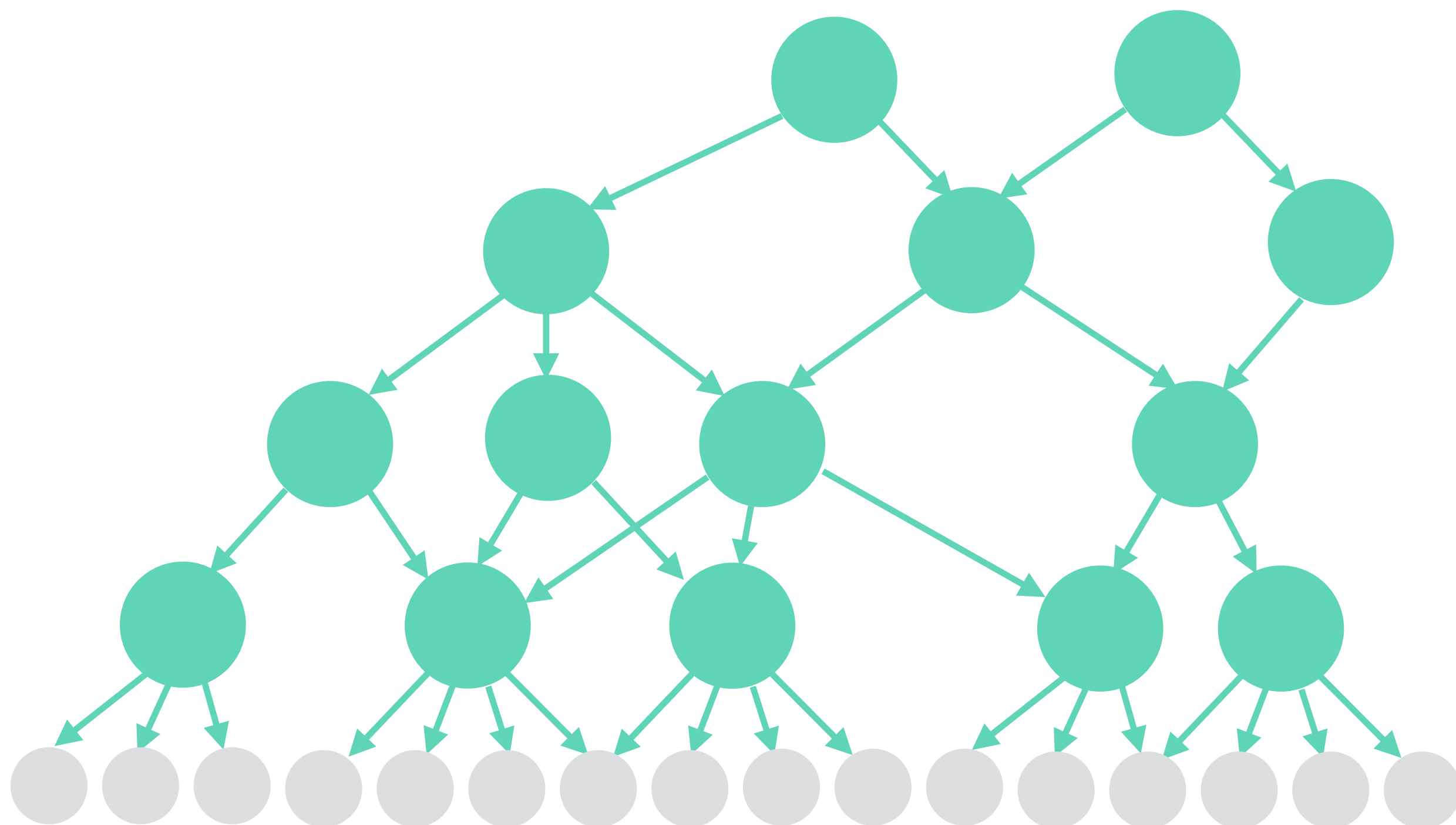




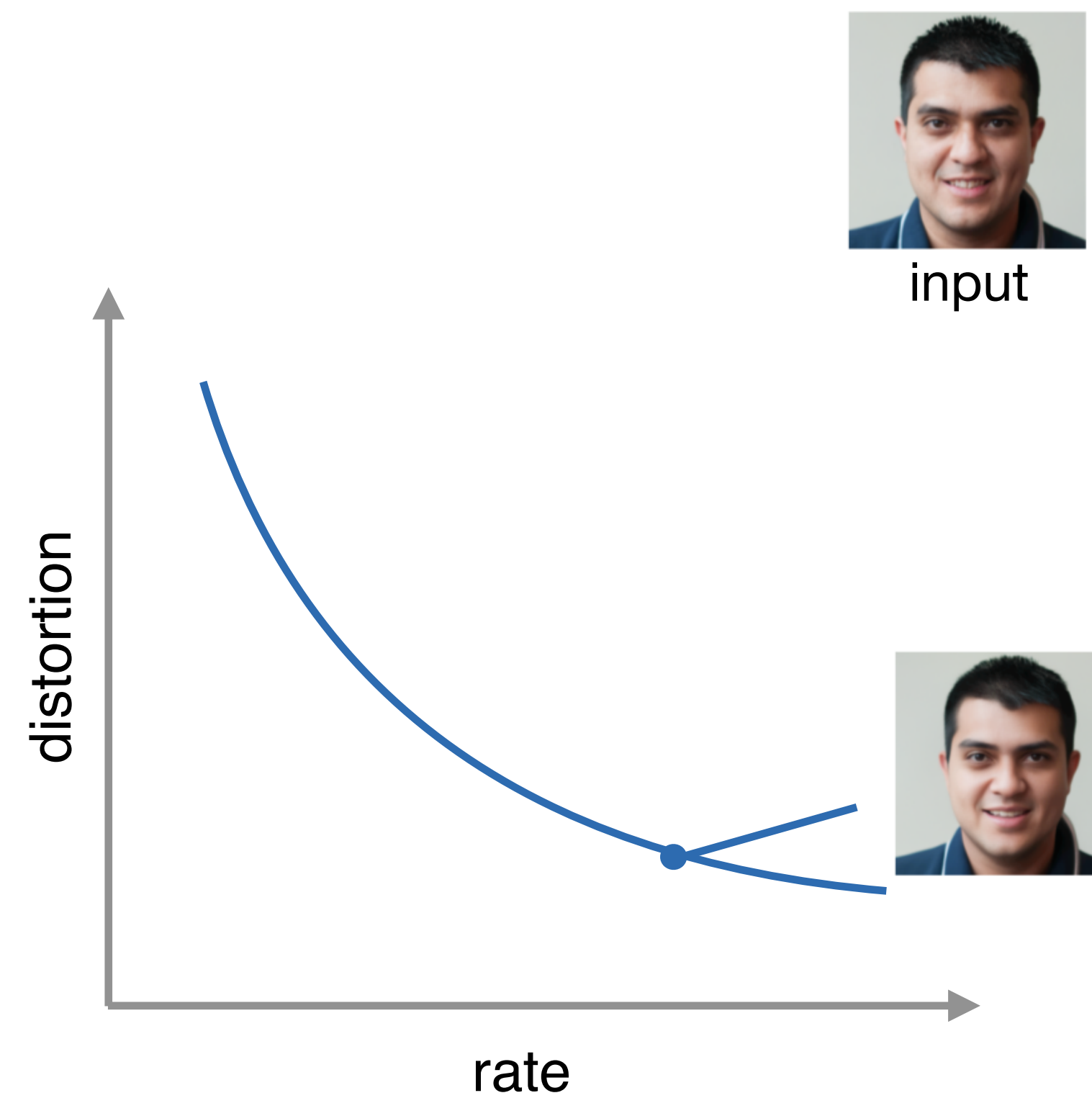
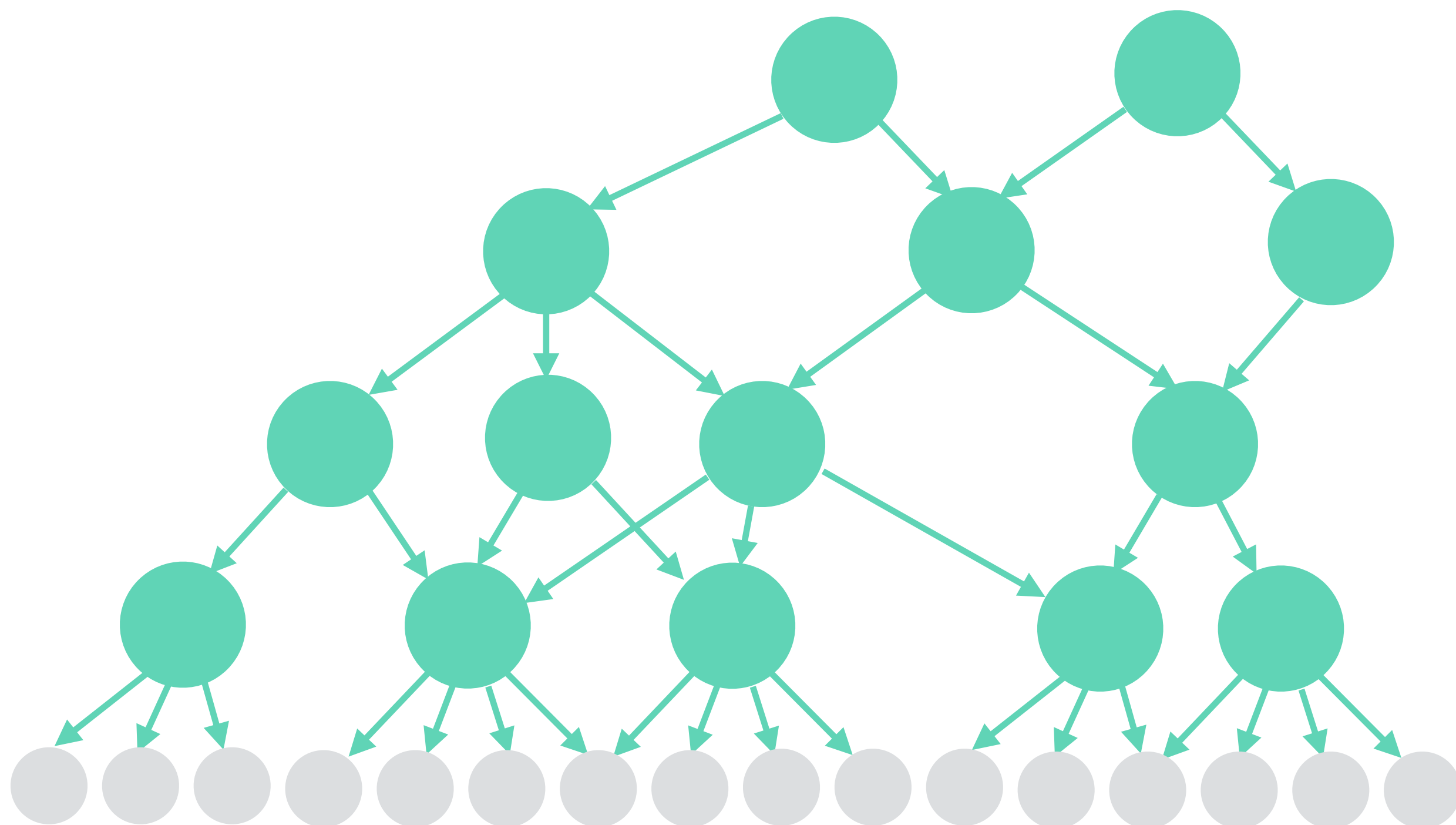
# variable-rate compression



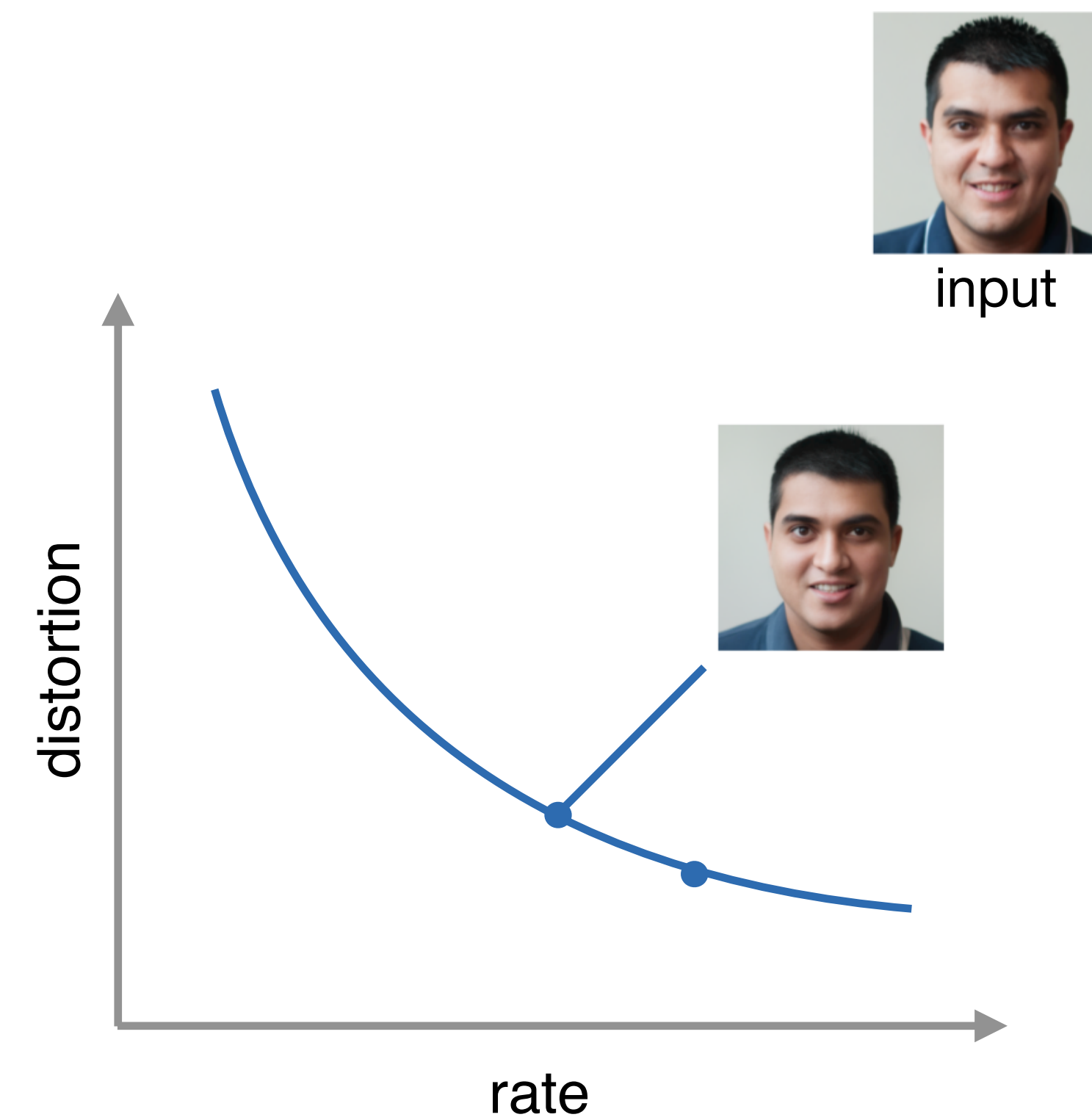
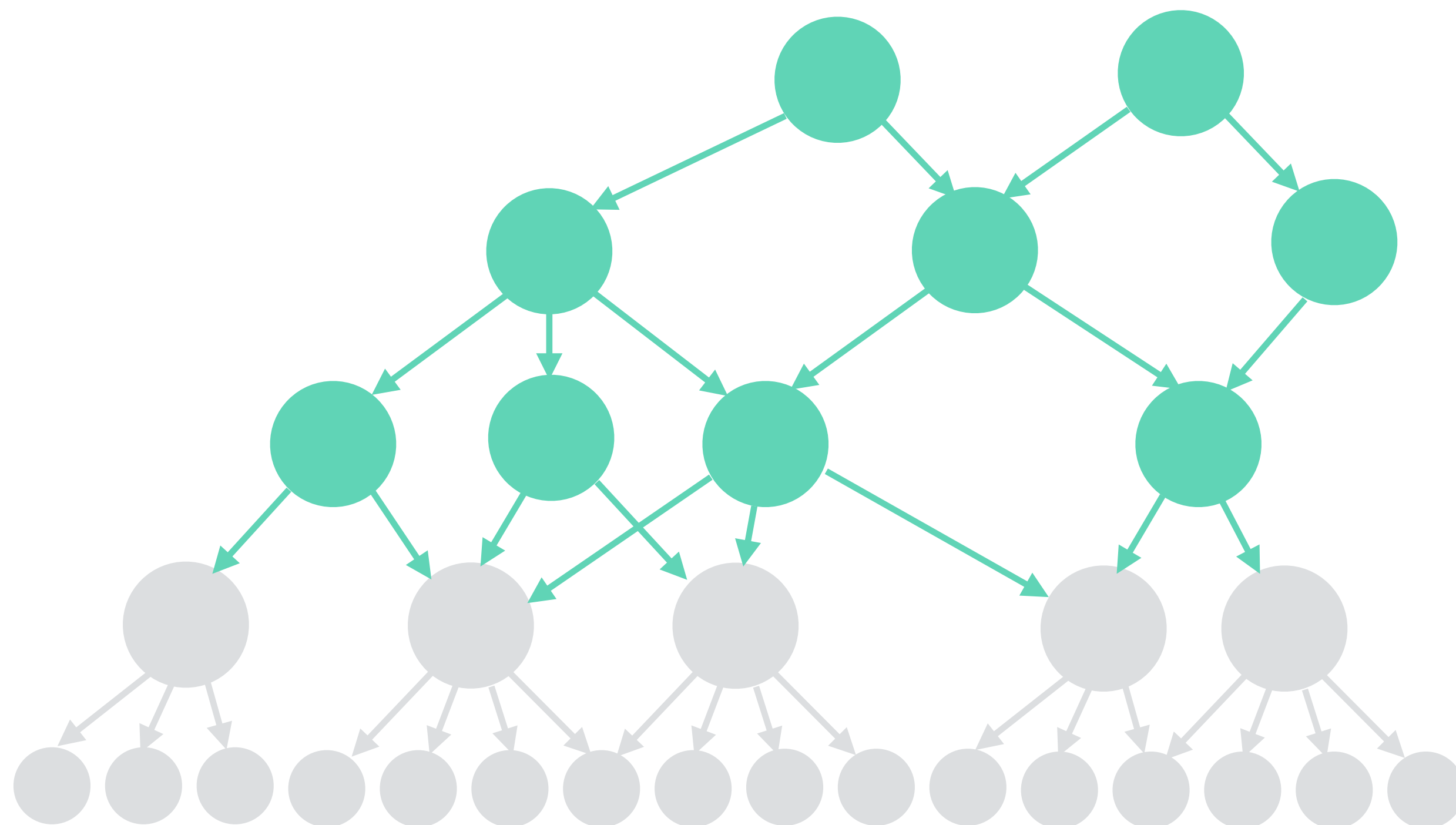
# hierarchical compression



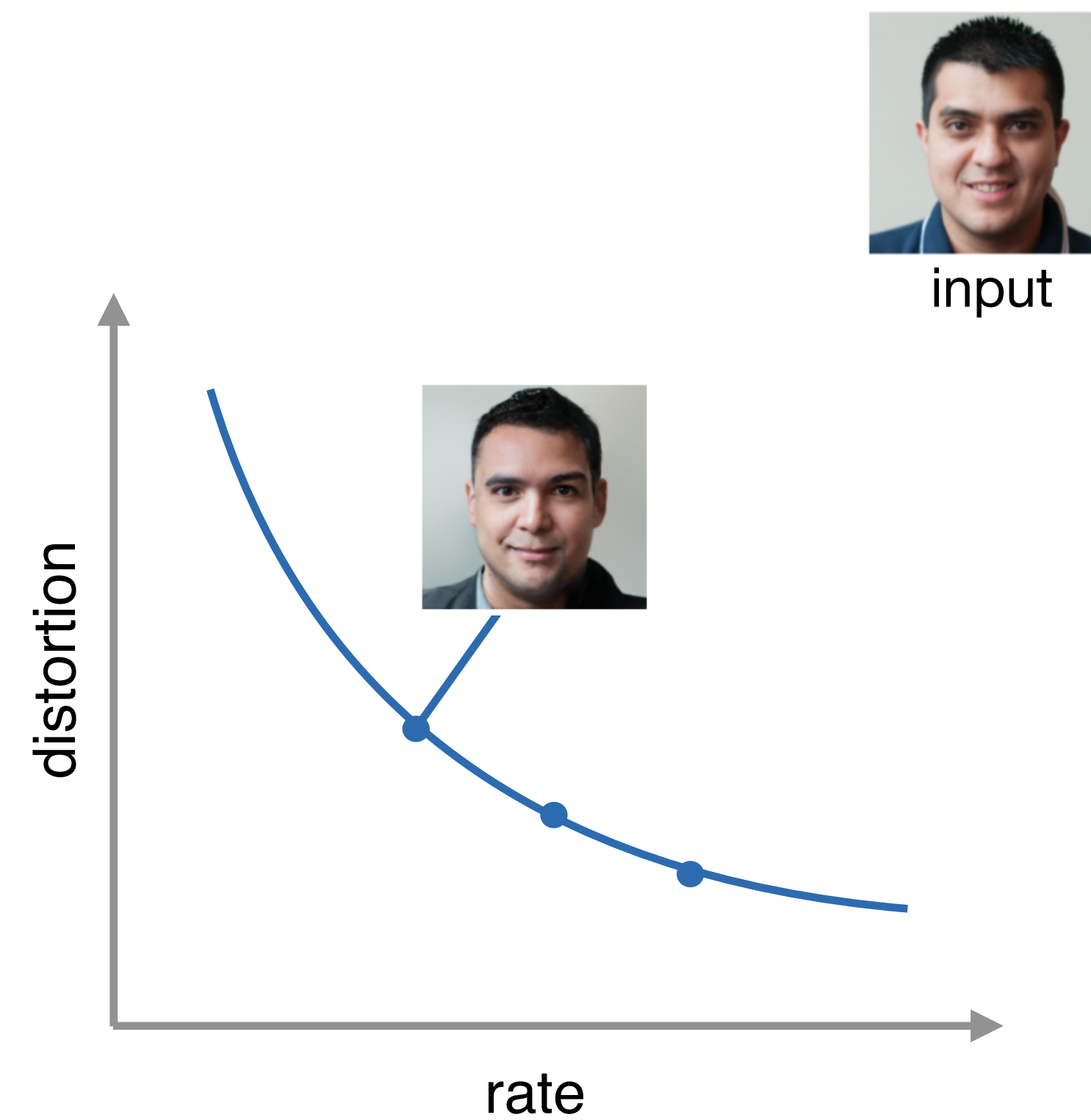
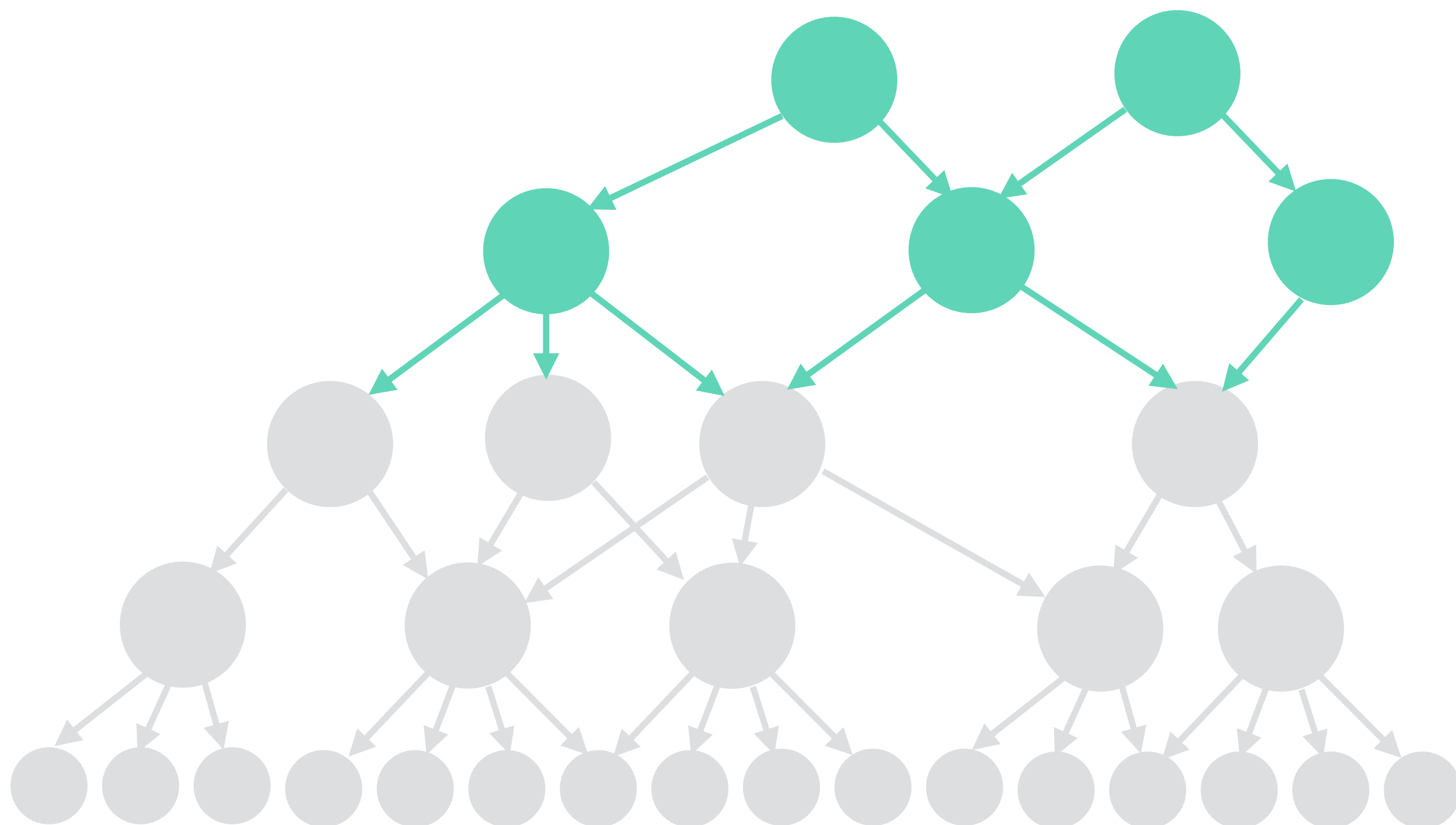
# hierarchical compression



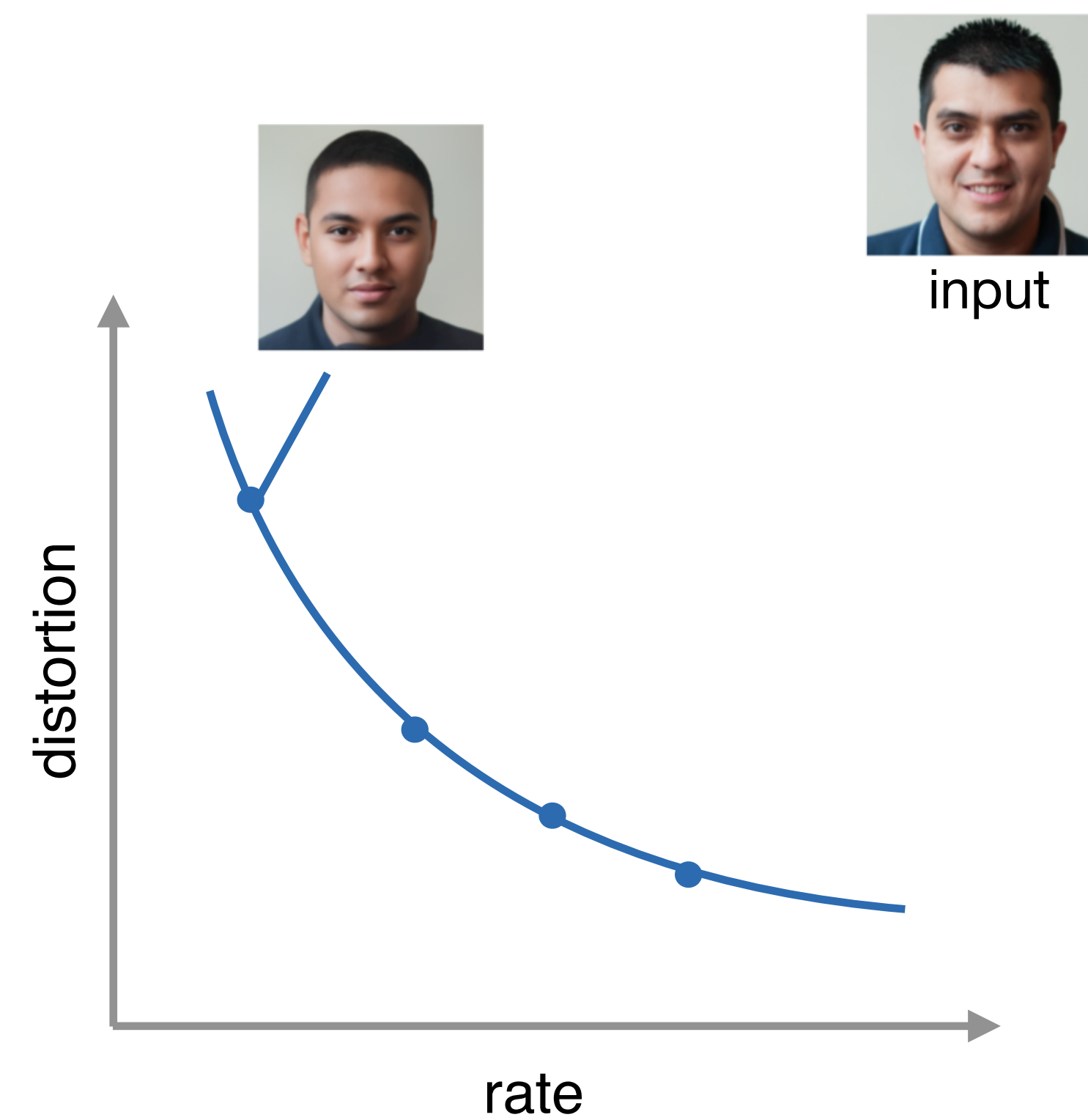
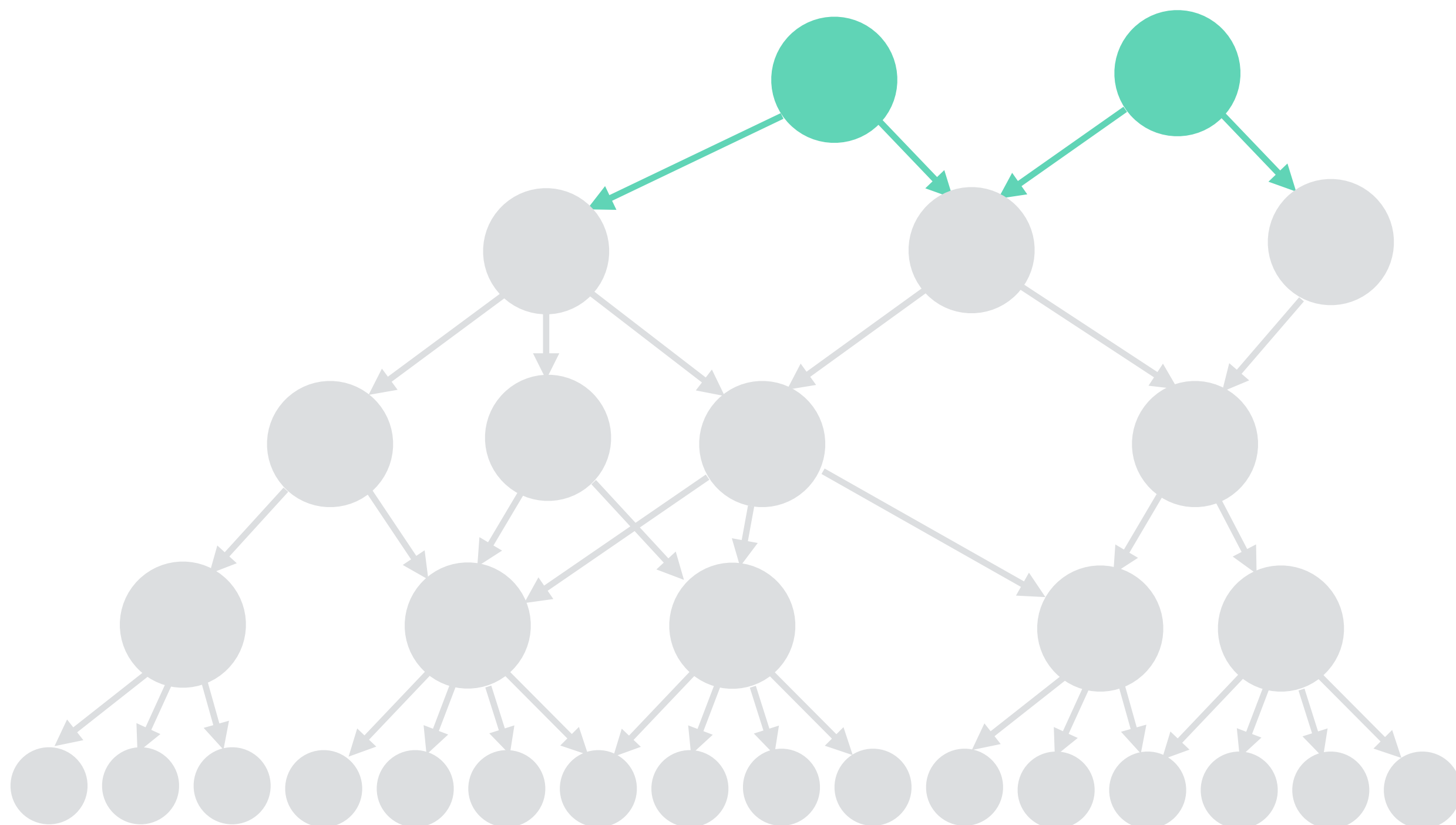
# hierarchical compression



# hierarchical compression

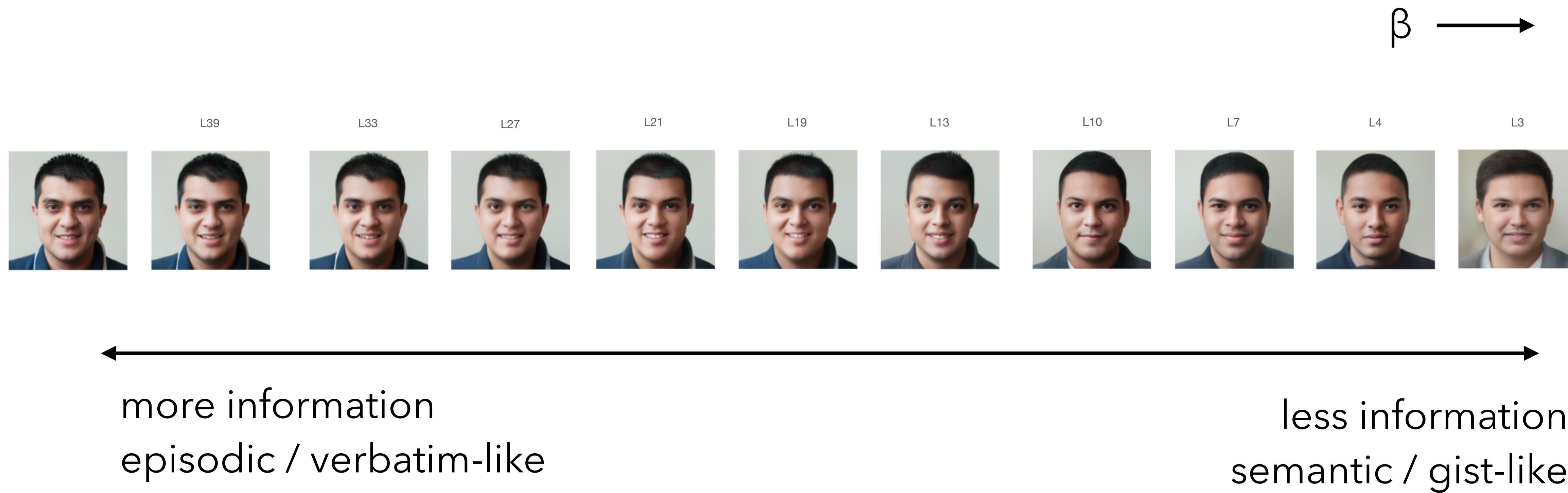


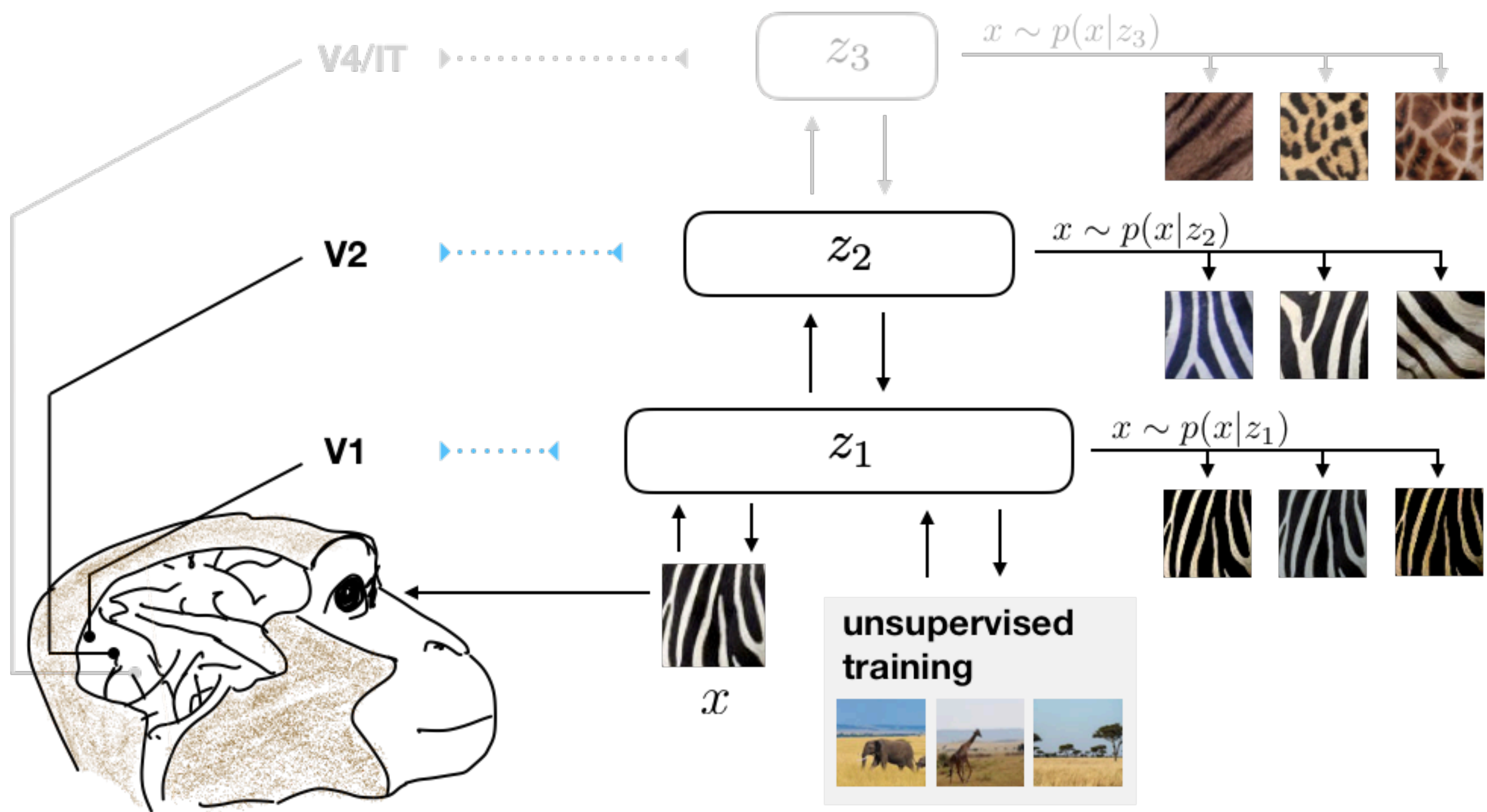
# hierarchical compression



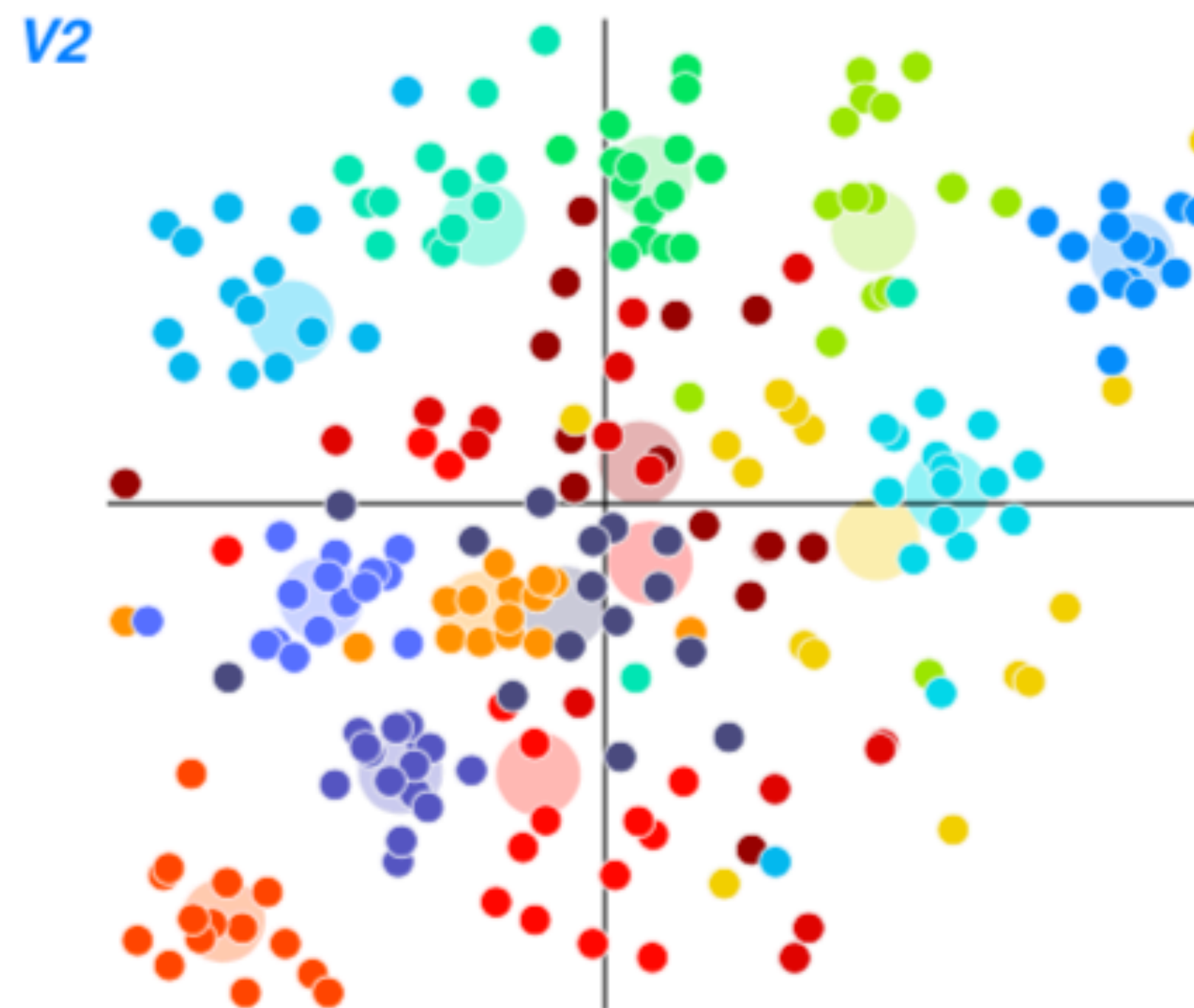
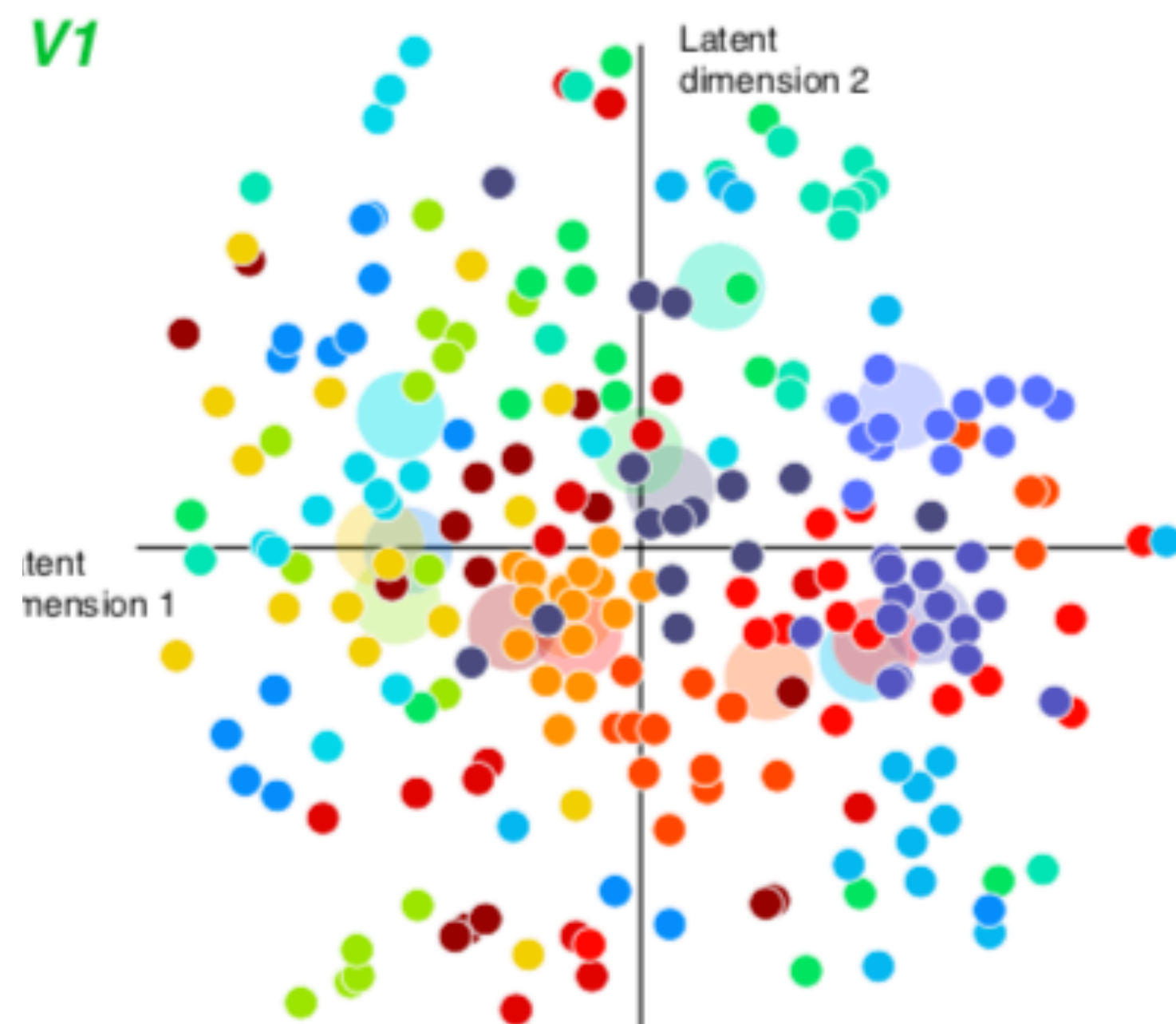
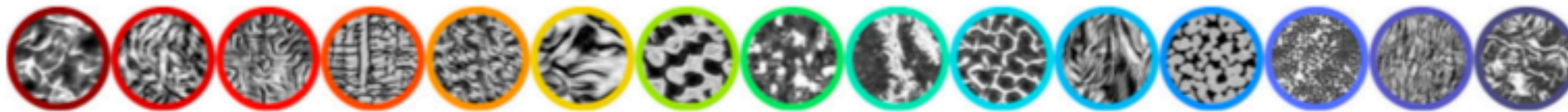


# hierarchical compression

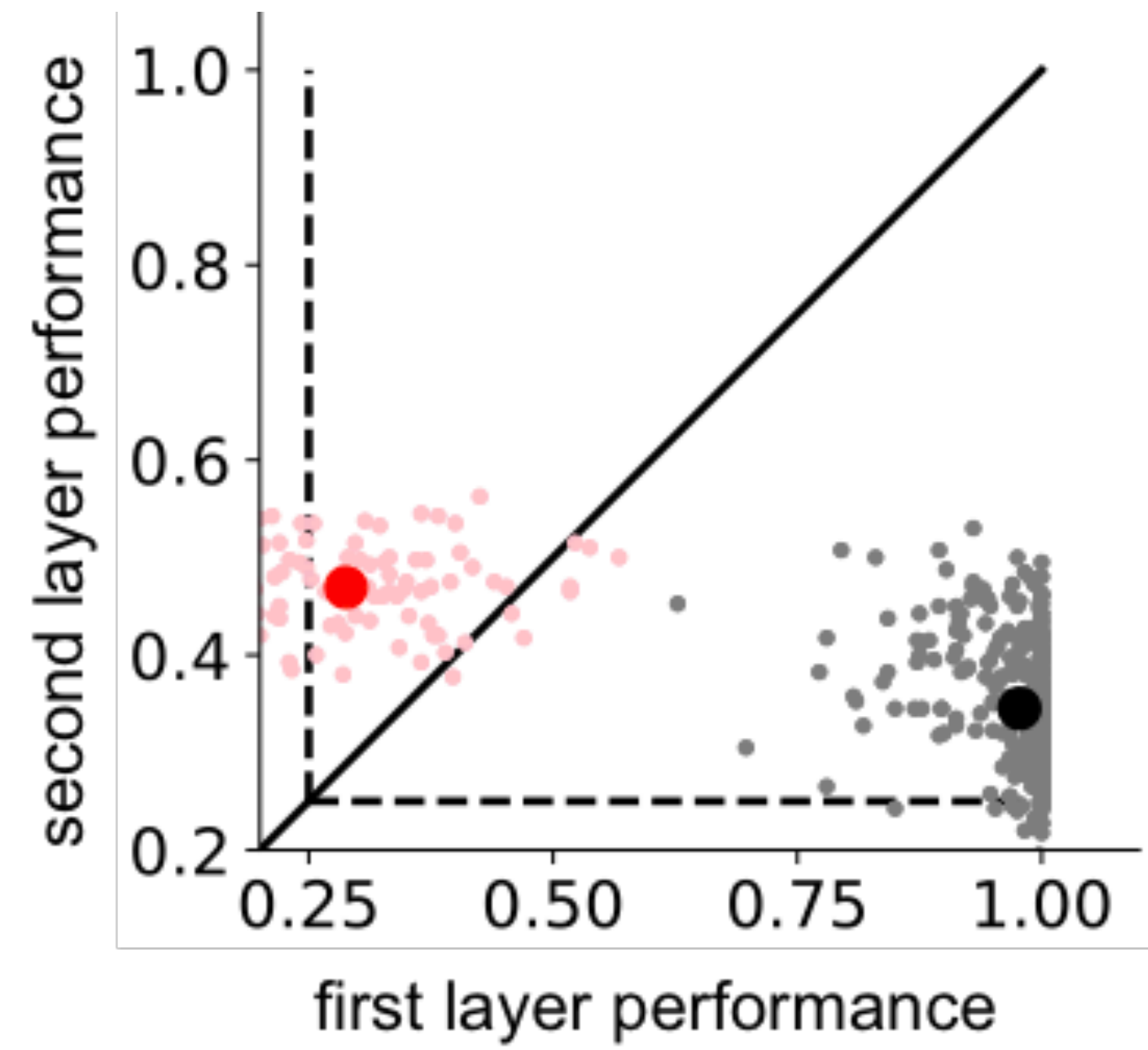




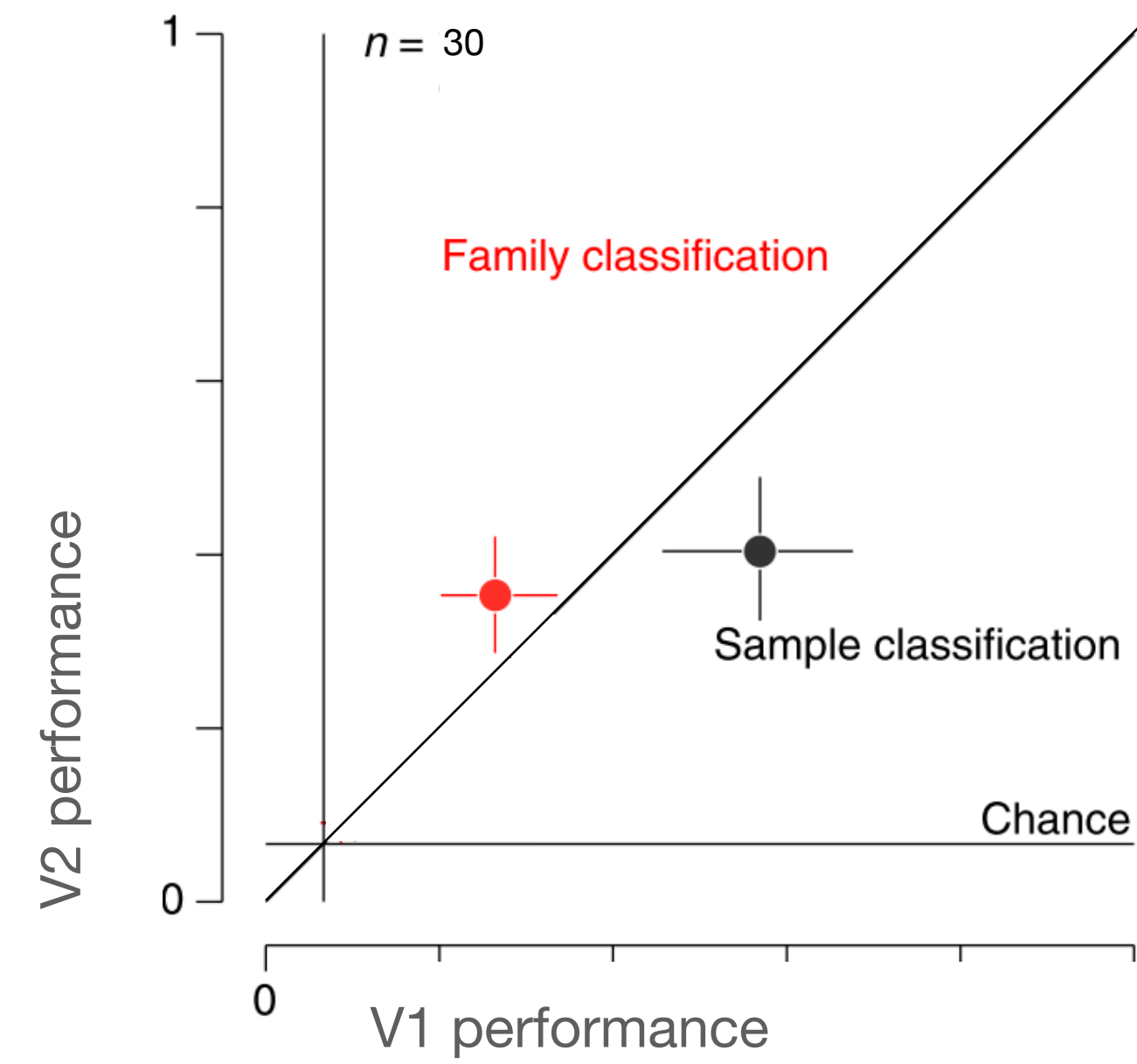




**model**

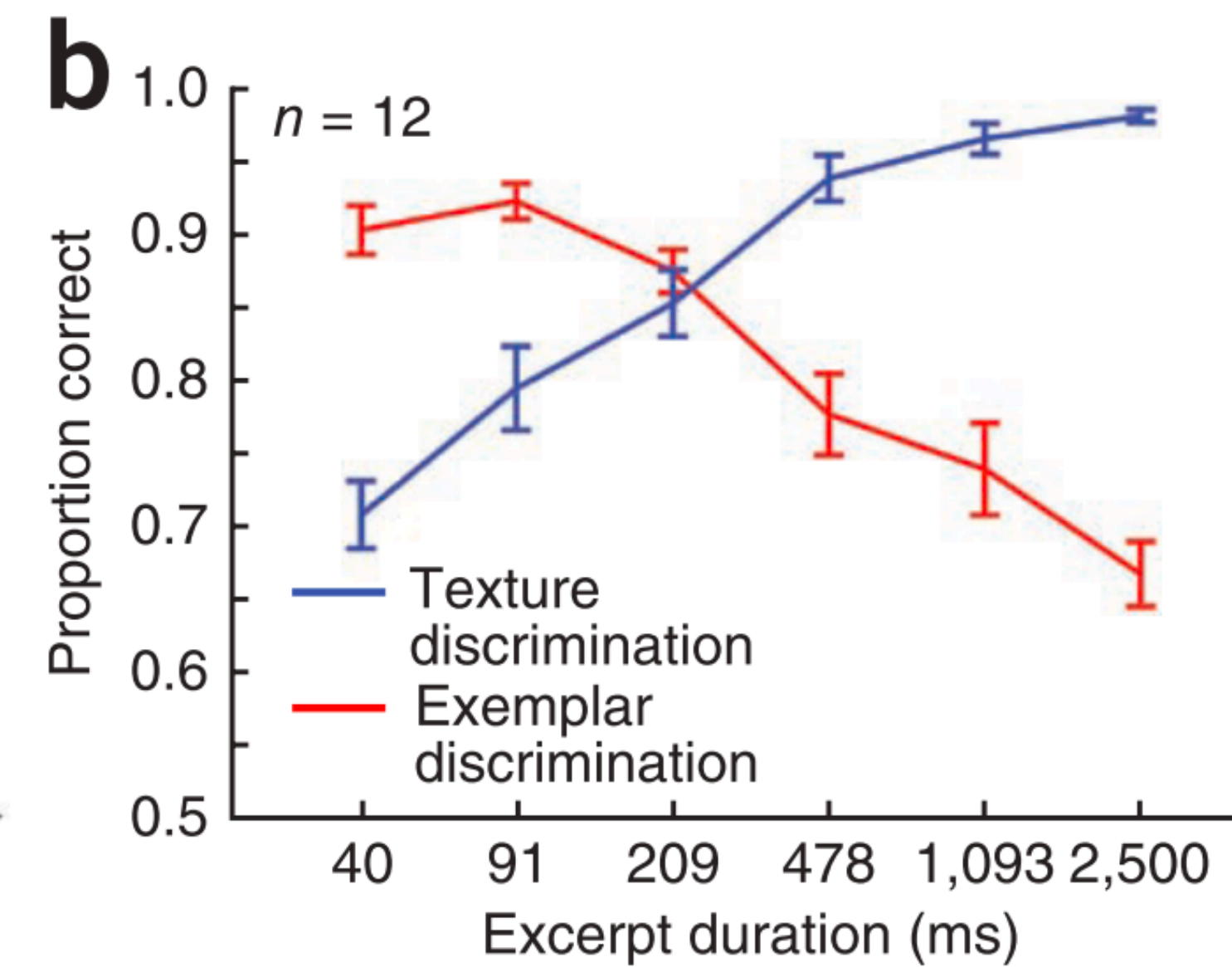
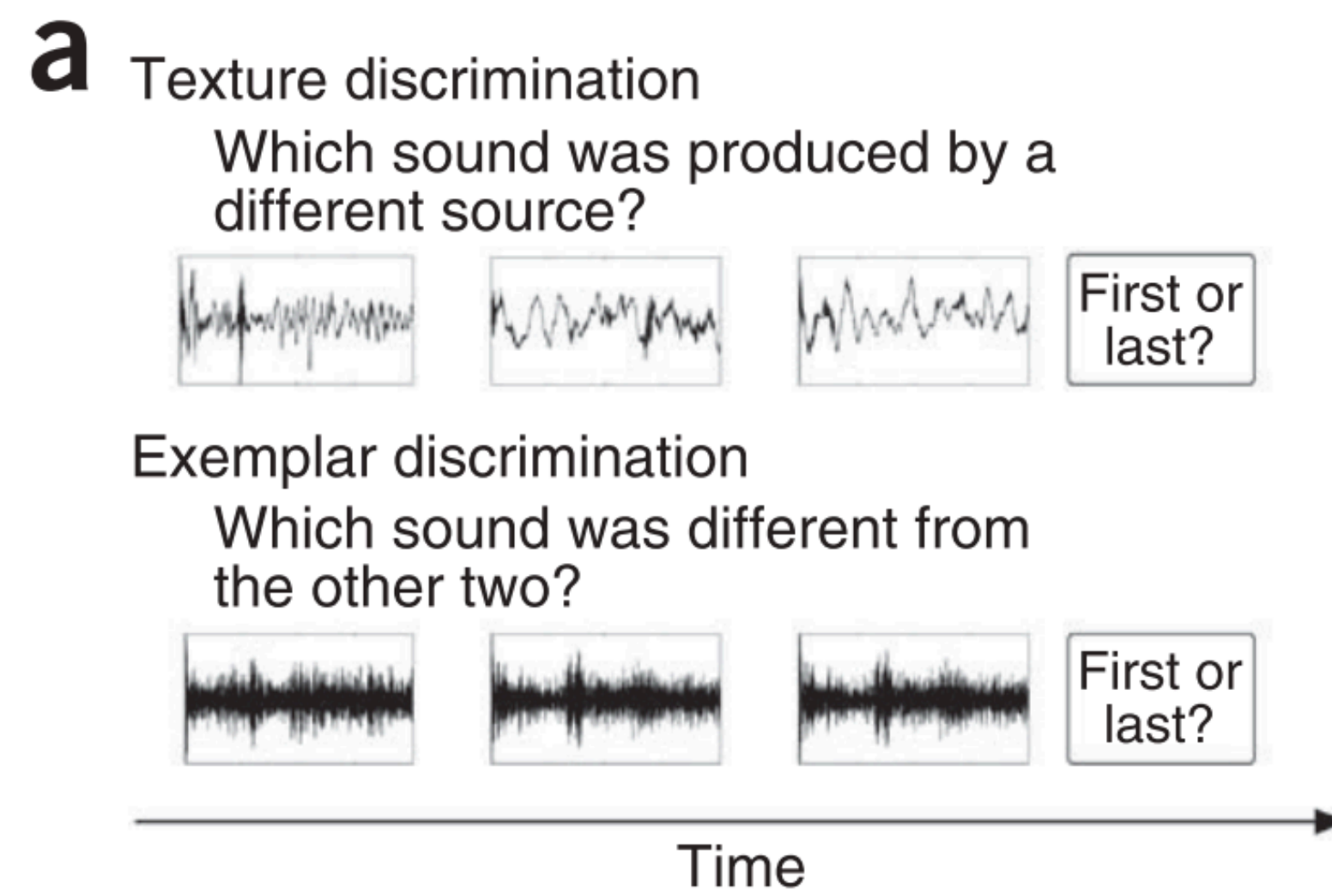
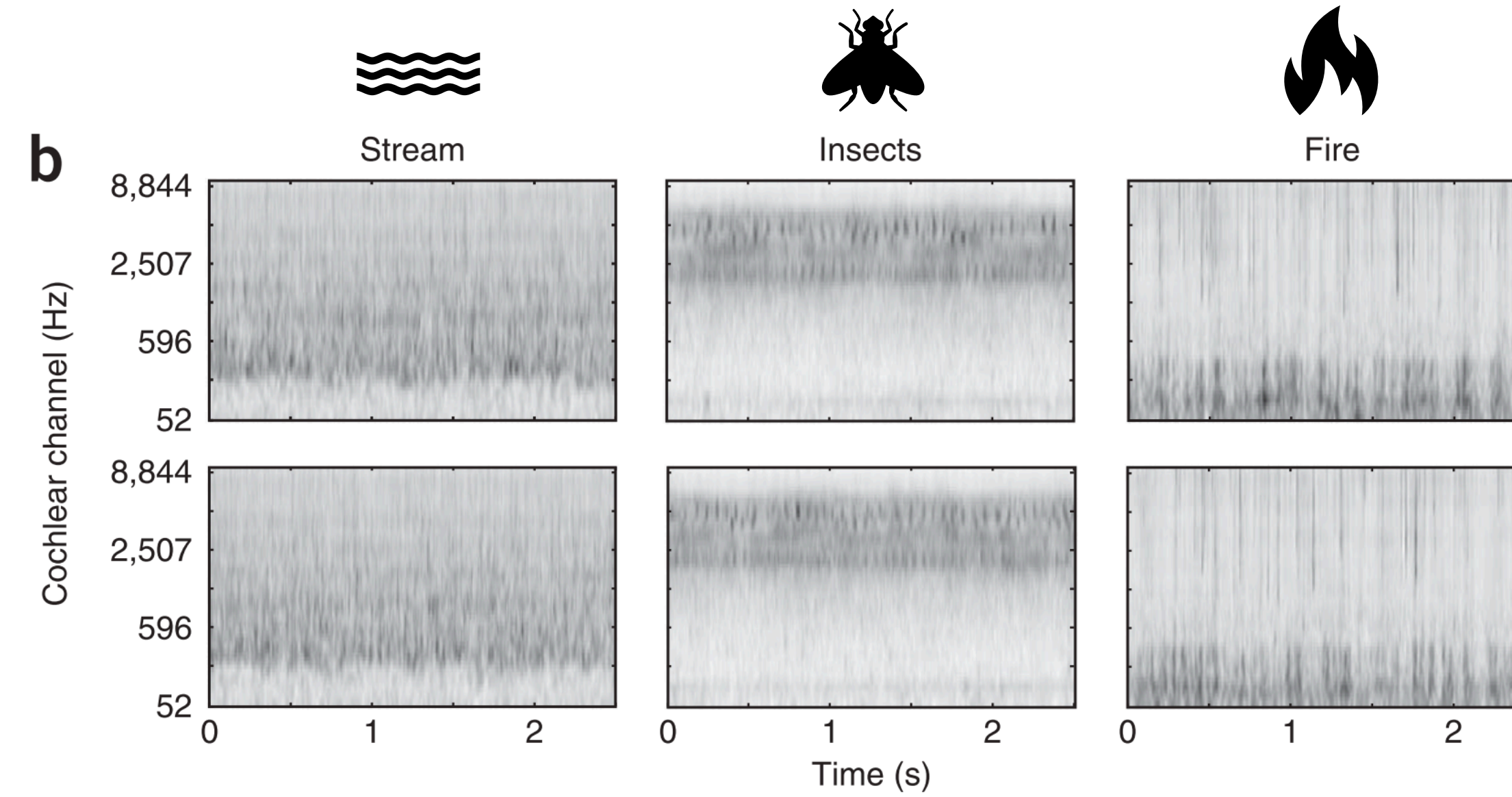


**experiment**



(experiment: Ziemba et al, 2016, model: Banyai et al, 2019)





# conclusions

We've argued that

- episodic memory can mitigate the problem of continual learning in case of uncertainty over model structure
- and that semantic memory can be used to compress episodes,
  - which can be formalised in the framework of lossy compression
  - and provide a normative, unifying explanation of a large variety of memory errors.
- Furthermore, we have proposed hierarchical generative models as a solution to variable-rate compression within a single model



outstanding questions

# acknowledgements



Gergő Orbán



Balázs Török



Csenge Fráter



Máté Büki



Mihály Bányai



Timo Flesch



Andrew Saxe



Chris Summerfield