General Principles of Human and Machine Learning

Lecture 7: Compression and resource constraints

David G. Nagy

https://hmc-lab.com/GPHML.html

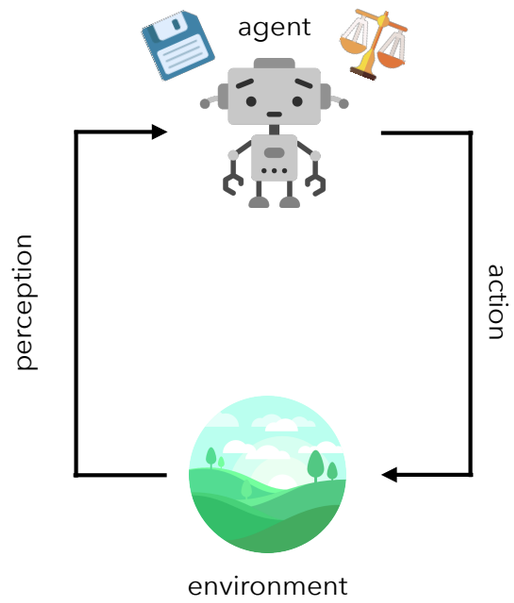
admin

- quiz
  - maximum score for quiz #2 is reduced to 16 from 18
  - from now on, pop quizzes may contain content from the lecture in the same week

Date	Remarks	Lecture	Tutorial	TA	Readings
Week 1:		Oct 15: Introduction (slides)	Oct 16 (slides)	Alex	Spicer & Sanborn (2019). What does the mind learn?
Week 2:		Oct 22: Origins of biological and artificial learning (slides)	Oct 23 (slides)	Turan	[1] Behaviorism [2] What is a perceptron? (Blog post)
Week 3:		Oct 29: Symbolic AI and Cognitive maps (slides)	Oct 30 (Quiz #1)	Alex	[1] Garnelo & Shanahan (2019) [2] Boorman et al., 2021
Week 4:		Nov 5: Introduction to RL (slides)	Nov 6 (slides)	Turan	Sutton & Barton (Ch. 1 & 2)
Week 5:		Nov 12: Advances in RL (slides)	Nov 13 (Quiz #2)	Turan	Neftci & Averbeck (2019)

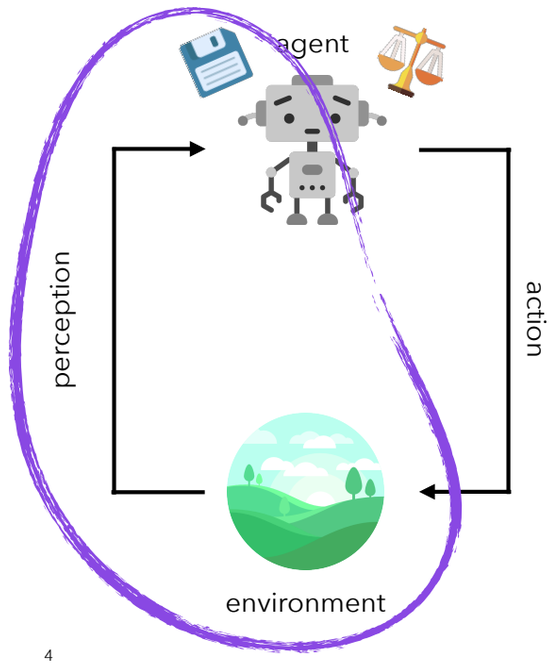
- halfway feedback form on course, please submit at
  - https://forms.gle/BWBHobeVZniJKuKdA

today

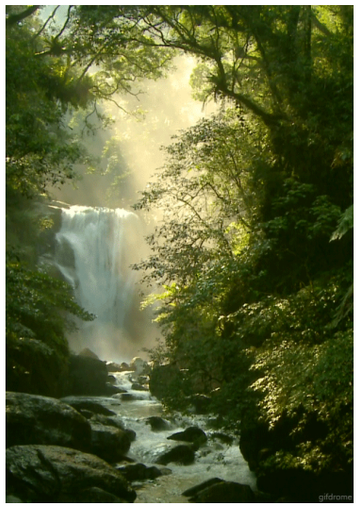


today

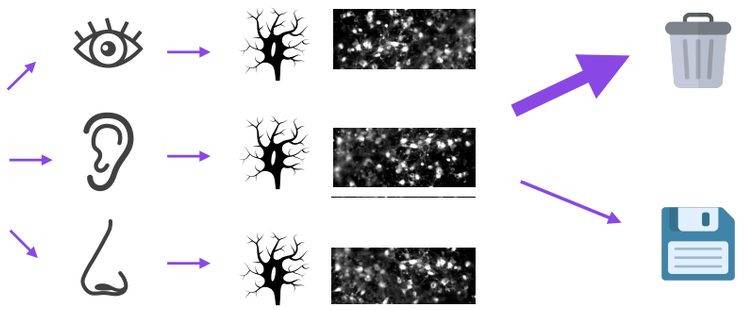
- focus on constraints on **memory**
  - lossless and lossy compression
  - generative compression
    - perception as bayesian inference
    - human memory distortions



computational problem of memory



environment

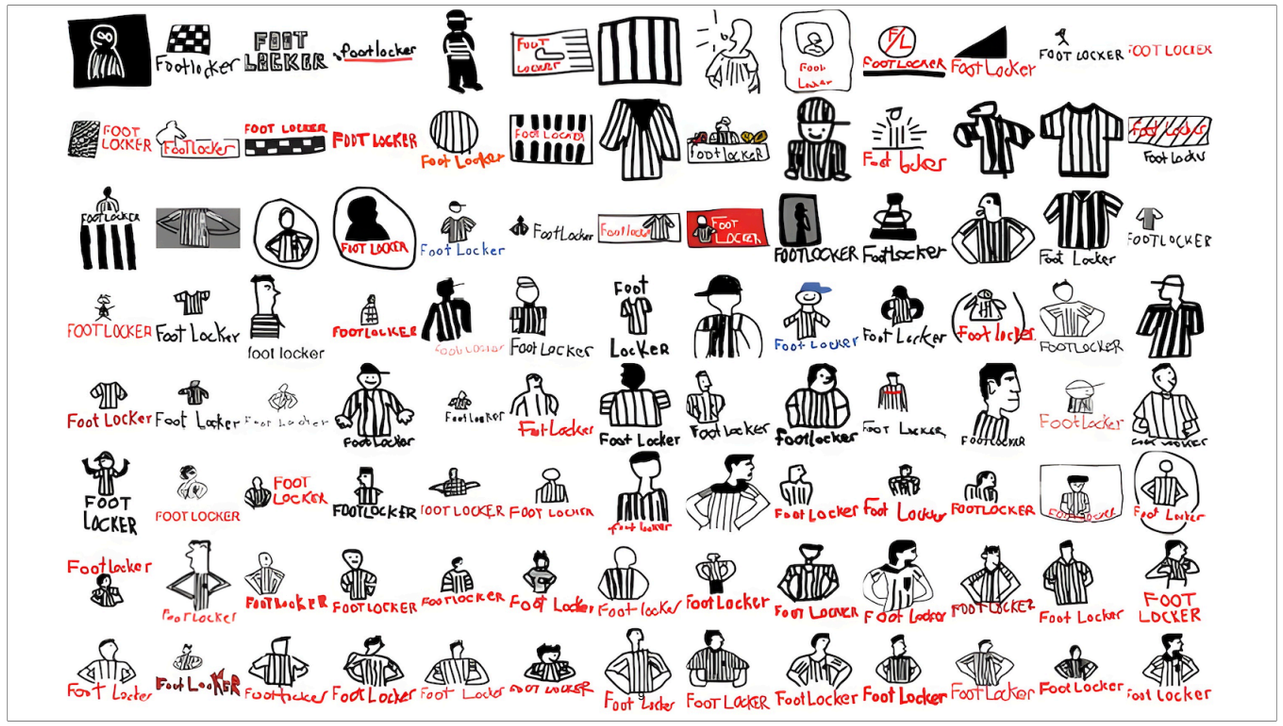


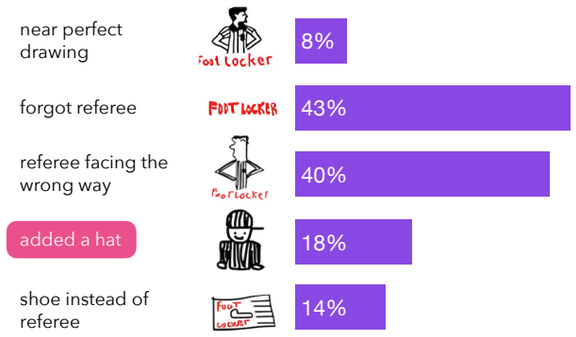
memory

(Nickerson & Adams, 1979)


adidas
BURGER KING
Domino's
ELEVEN
Foot Locker
Walmart
TARGET.
IKEA

Also happens for logos of well known companies. Maybe people just aren’t good at drawing?


Footlocker
FOOT LOCKER
Foot Locker
FOOTLOCKER
foot locker

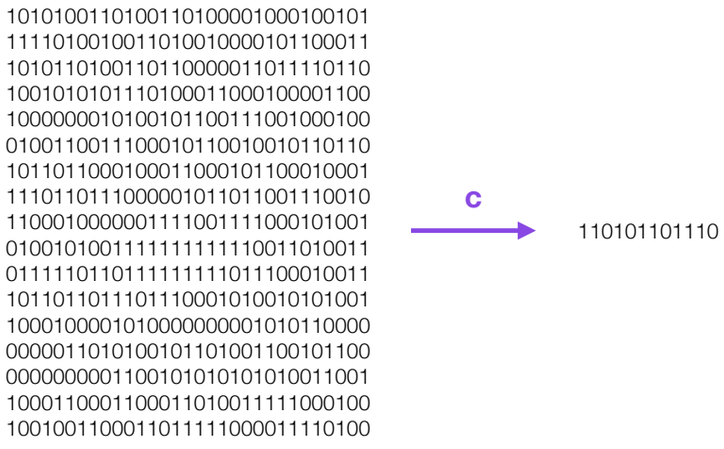

near perfect drawing
8%
forgot referee
43%
referee facing the wrong way
40%
added a hat
18%
shoe instead of referee
14%

recognition

Prasad & Bainbridge, 2023

- 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

compression



compression

I walked my four legged animal that barks on the day before today after the huge glowing ball of fire left the sky

compression

I walked my four legged animal that barks on the day before today after the huge glowing ball of fire left the sky

compression

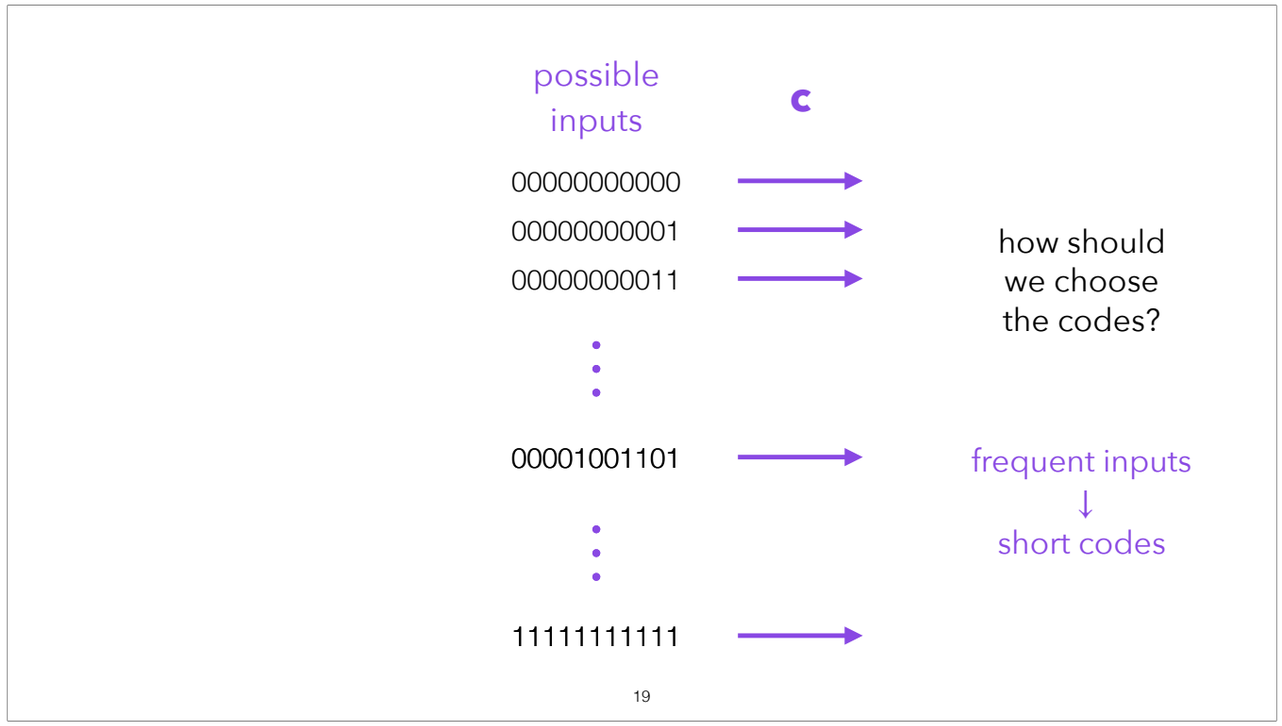
I walked my dog on the day before today after the huge glowing ball of fire left the sky

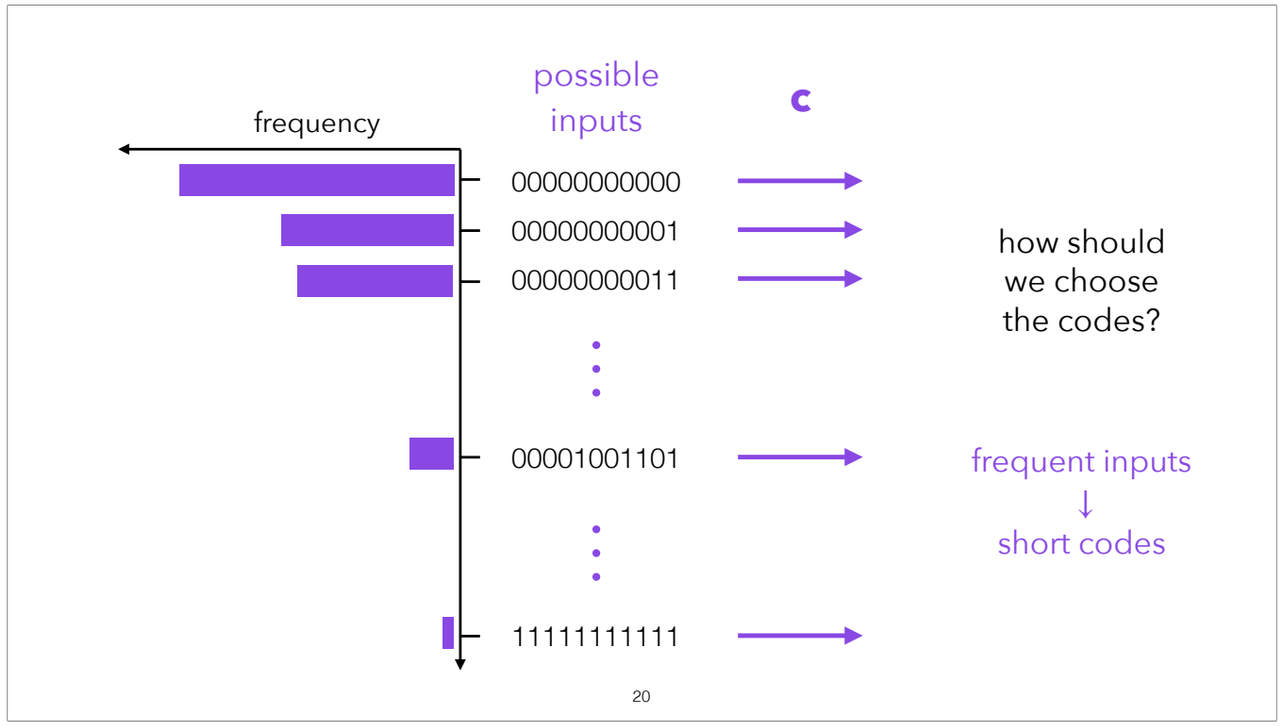
compression

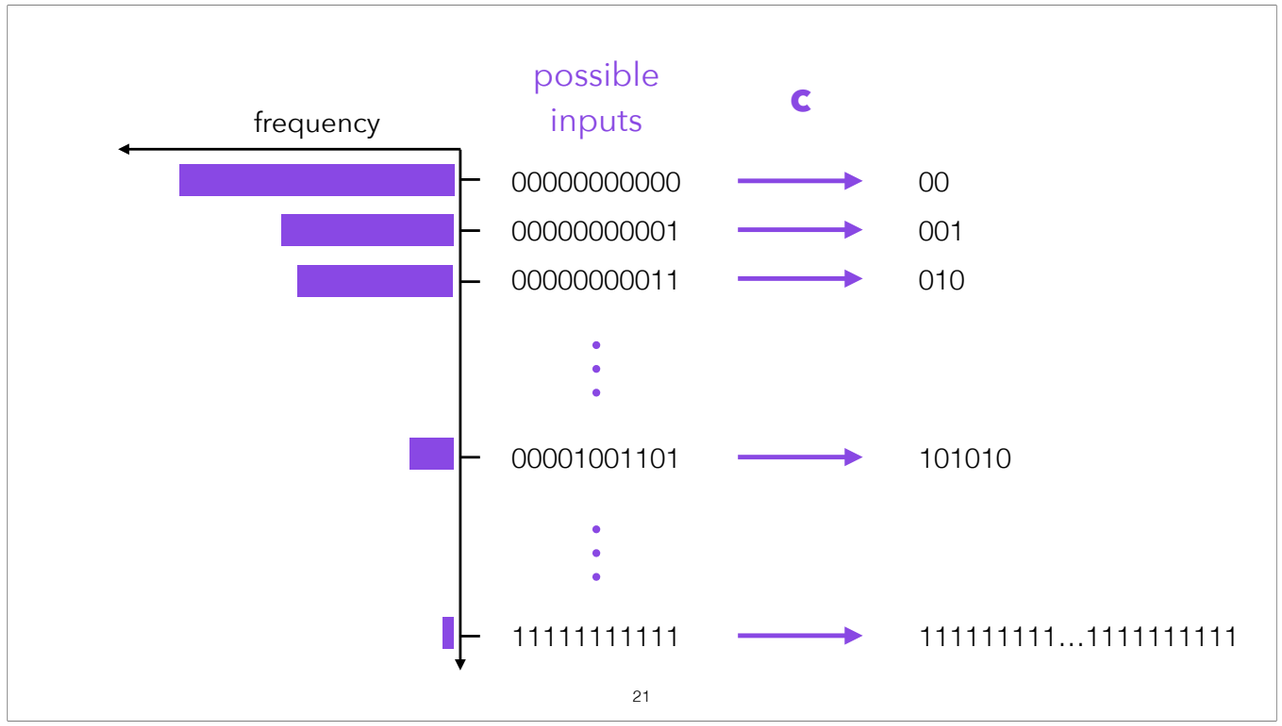
I walked my dog yesterday after the huge glowing ball of fire left the sky

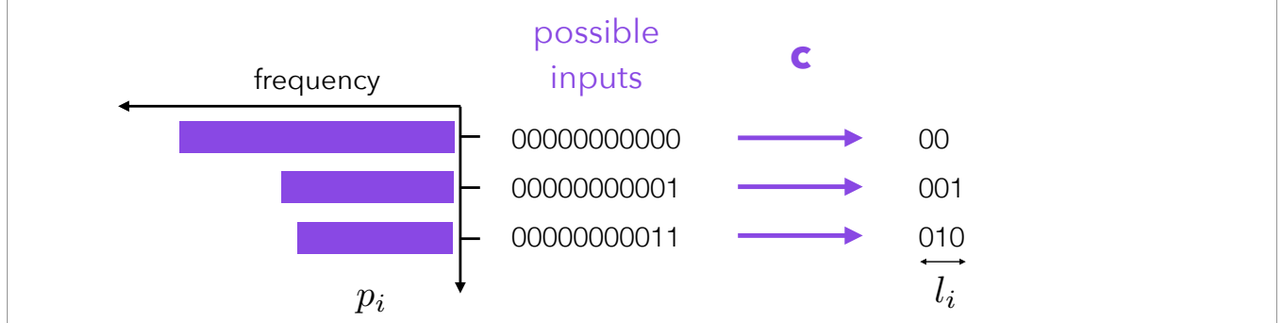
compression

I walked my dog yesterday after sunset


possible inputs
c
00000000000
00000000001
00000000011
00001001101
11111111111
how should we choose the codes?
frequent inputs
↓
short codes

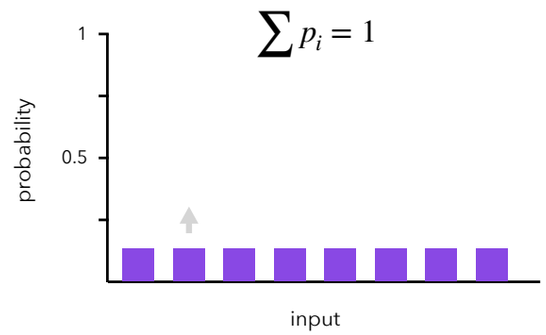

possible
inputs
C
frequency
00000000000
00000000001
00000000011
00001001101
11111111111
how should
we choose
the codes?
frequent inputs
↓
short codes


possible
inputs
C
frequency
00000000000
00000000001
00000000011
00001001101
11111111111
00
001
010
101010
111111111...1111111111

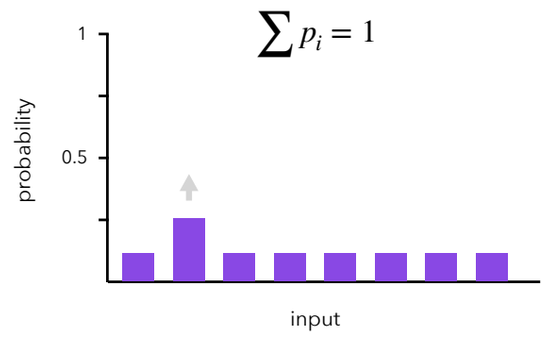


p_i is the probability or relative frequency of an input string
l_i is the length of the code word that is used for the i-th input

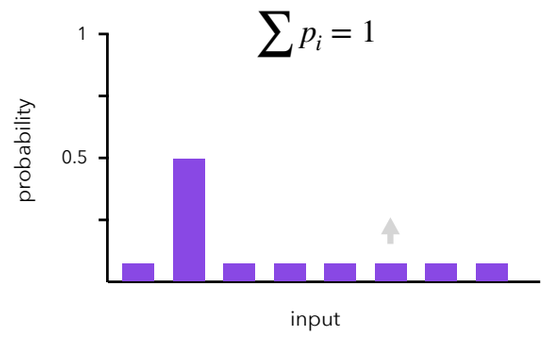
probability budget



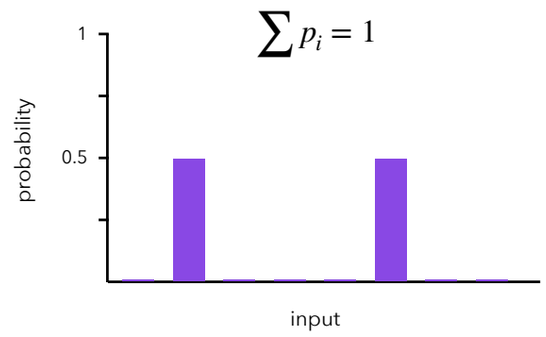
probability budget



probability budget



probability budget



codeword budget

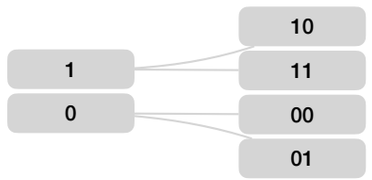
how many codewords of length l_i?

1

0

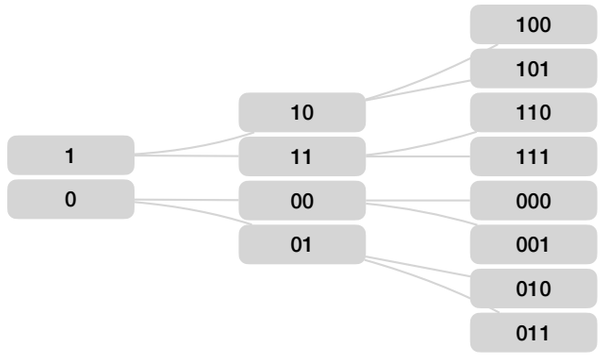
codeword budget

how many codewords of length l_i?



codeword budget

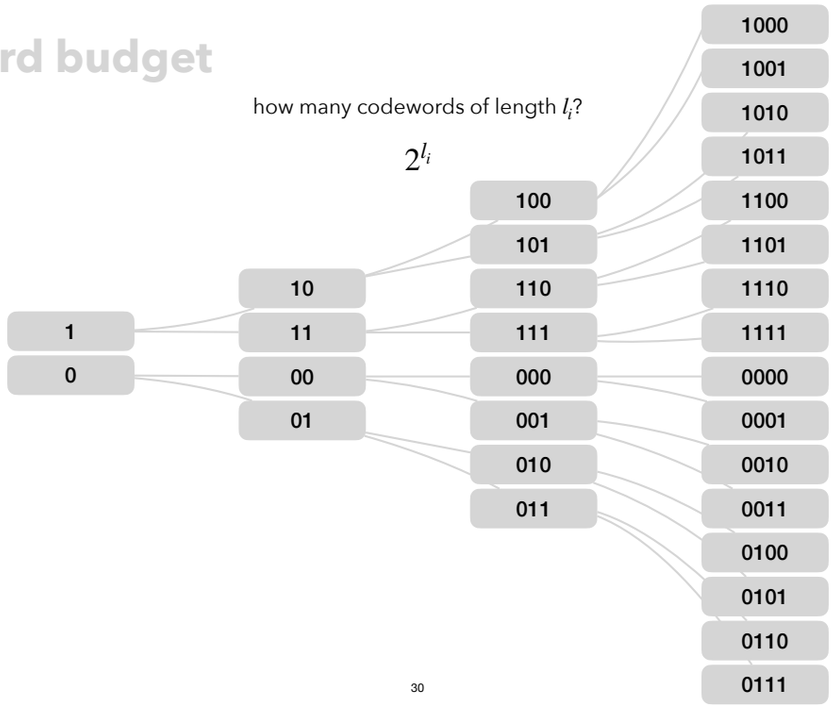
how many codewords of length l_i?



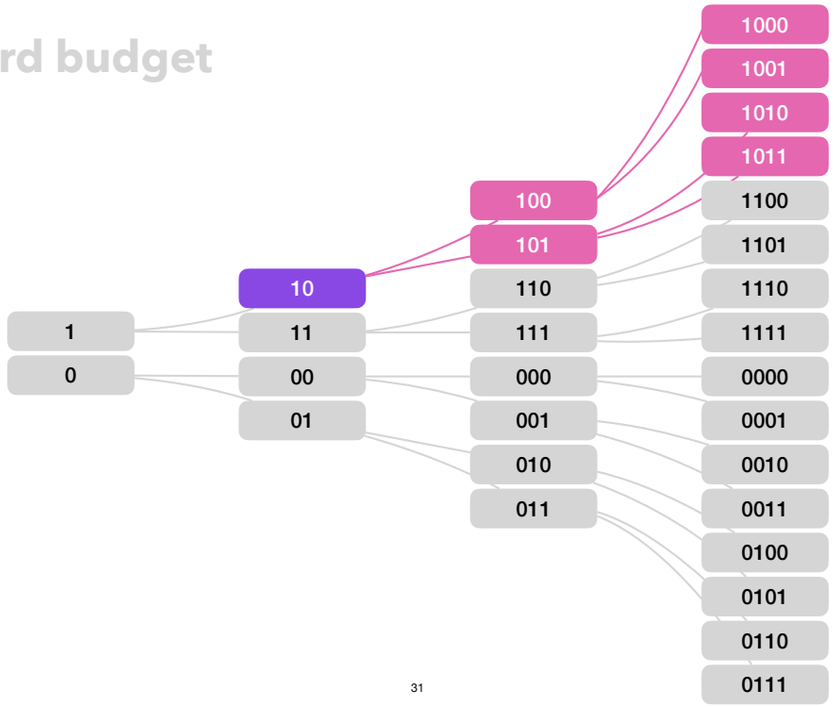
codeword budget

how many codewords of length l_i?

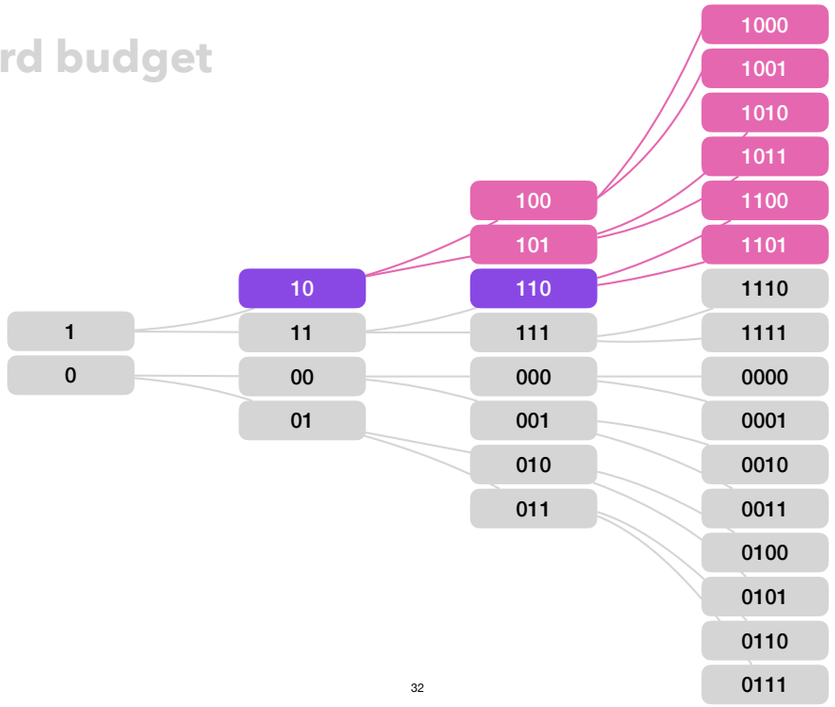
$$2^{l_i}$$



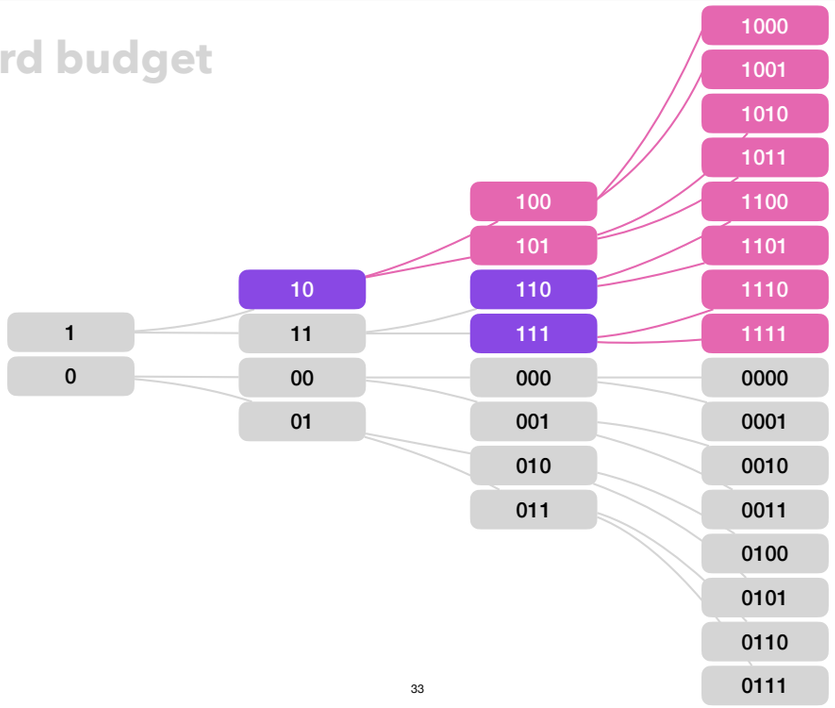
codeword budget



codeword budget

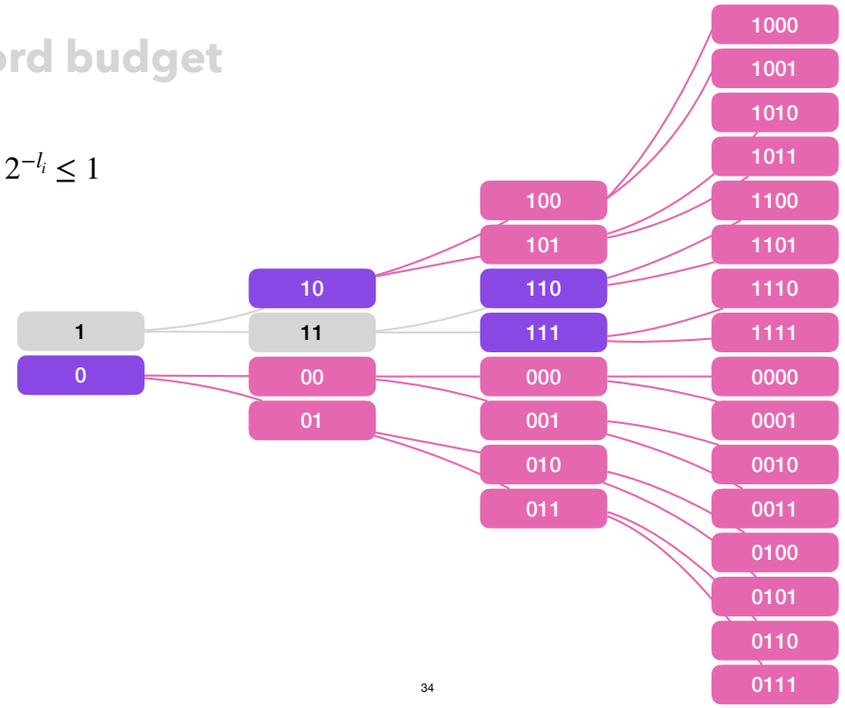


codeword budget



codeword budget

$$\sum 2^{-l_i} \leq 1$$



Kraft-McMillan inequality
proof: https://en.wikipedia.org/wiki/Kraft–McMillan_inequality#Proof_for_prefix_codes

example coding function

$$p_1 = 1/16 \;\rightarrow\; l_1 = 4$$

1000	1100	0000	0100
1001	1101	0001	0101
1010	1110	0010	0110
1011	1111	0011	0111

- the grid is a probability distribution with 16 equally likely outcomes
- each element of the grid is given a unique codeword

example coding function

$p_1 = 1/16 \rightarrow l_1 = 4$

1000	1100	0000	0100
1001	1101	0001	0101
1010	1110	0010	0110
1011	1111	0011	0111

example coding function

$p_1 = 1/16 \rightarrow l_1 = 4$

1000	1100	0000	0100
1001	1101	0001	0101
1010	1110	0010	0110
	1111	0011	0111

$p_3 = 1/8 \rightarrow l_3 = 3$

example coding function

$p_1 = 1/16 \rightarrow l_1 = 4$

1000	1100	0000	0100
1001	1101	0001	0101
101	1110	0010	0110
	1111	0011	0111

$p_3 = 1/8 \rightarrow l_3 = 3$

example coding function

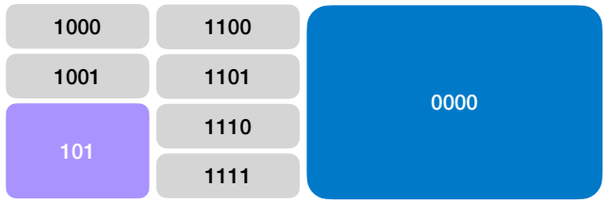
$p_1 = 1/16 \rightarrow l_1 = 4$

1000	1100	0000	0100
1001	1101	0001	0101
101	1110	0010	0110
	1111	0011	0111

$p_3 = 1/8 \rightarrow l_3 = 3$

example coding function

$p_1 = 1/16 \rightarrow l_1 = 4$

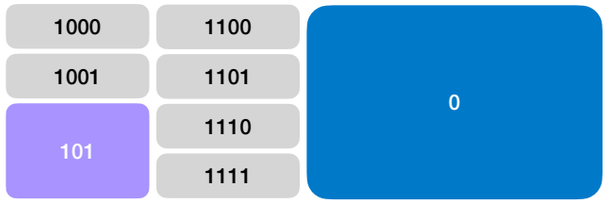


$p_3 = 1/8 \rightarrow l_3 = 3$

$p_9 = 1/2 \rightarrow l_9 = 1$

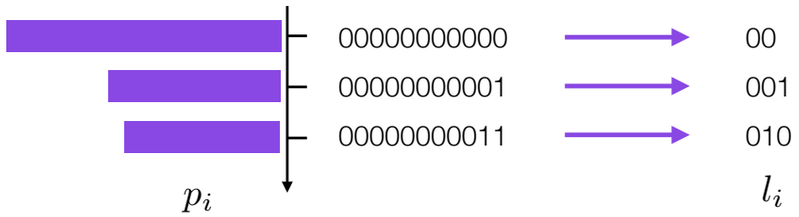
example coding function

$p_1 = 1/16 \rightarrow l_1 = 4$



$p_3 = 1/8 \rightarrow l_3 = 3$

$p_9 = 1/2 \rightarrow l_9 = 1$



$p_i = 2^{-l_i}$

probabilistic models

$l_i = -\log p_i$

possible codes (languages)

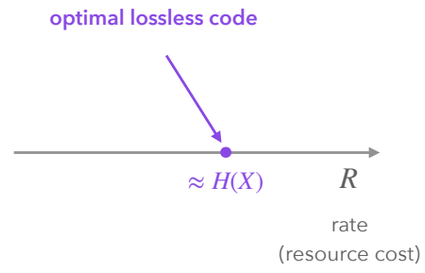
find a model that describes the world well

find a 'language' in which experiences can be described concisely

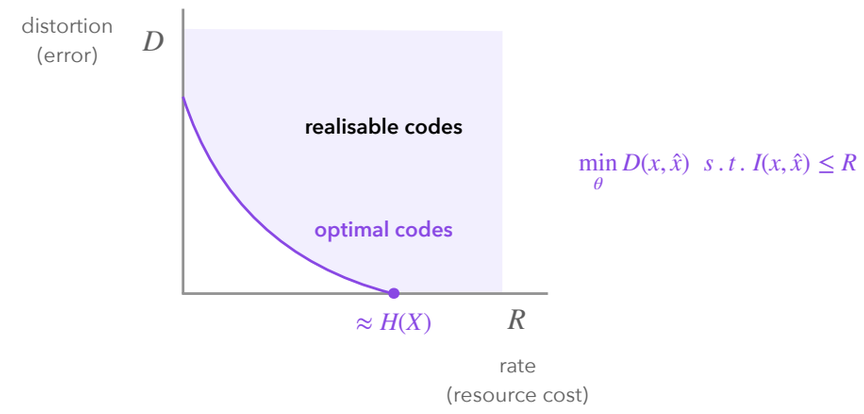
lossless compression

(source coding theorem)

$$H(X) = \mathbb{E}[-\log p_i] \leq \mathbb{E}[l_i] < \mathbb{E}[-\log p_i] + 1$$



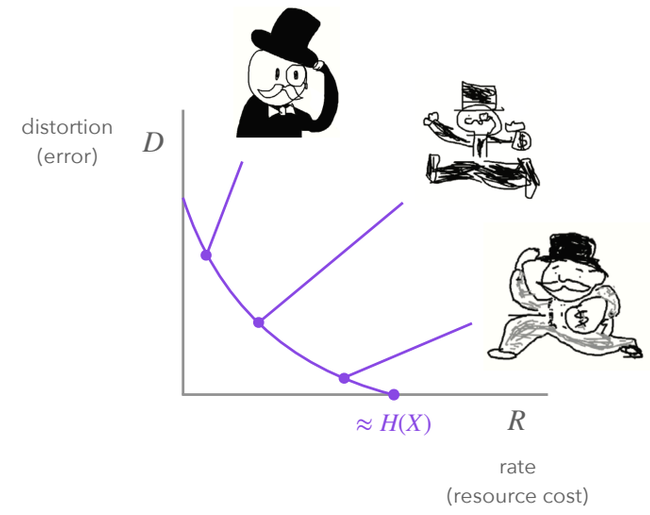
lossy compression



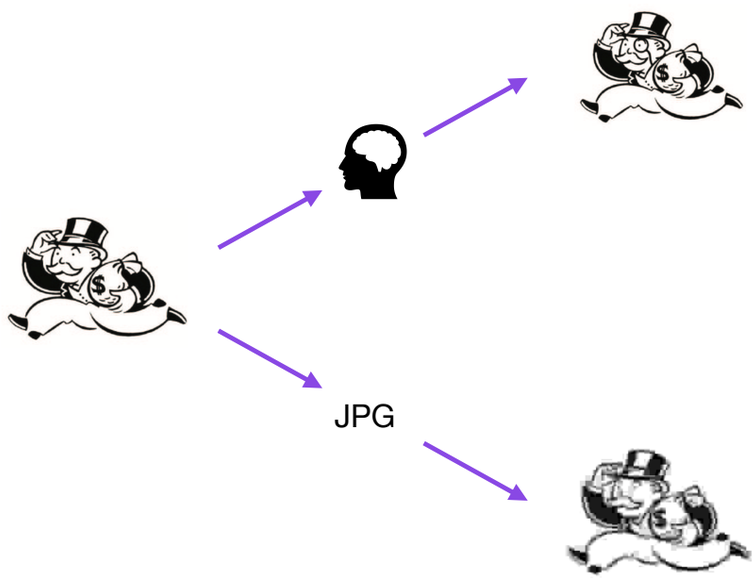
$$\min_{\theta} D(x, \hat{x}) \;\; s.t. \; I(x, \hat{x}) \leq R$$

I(A,B) is mutual information between A and B
this is the information that A contains about B (or that B contains about A)
I(A,B)=H(A)-H(A|B)=H(B)-H(B|A)
that is, the reduction in entropy about A if we learn the value of B
if we B is the memory trace for A, this tells us how noisy vs reliable the memory device is

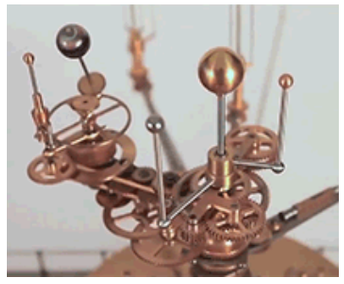
rate distortion trade-off



Prasad & Bainbridge, 2023


JPG

generative compression



$\hat{P}(x \mid \theta)$

generative model

$P(x)$

environment

a generative model is a probabilistic model of how the environment generates observations

generative compression

semantic memory

$\hat{P}(x \mid \theta)$

sensory experience

x

perception as inference

what are we interested in

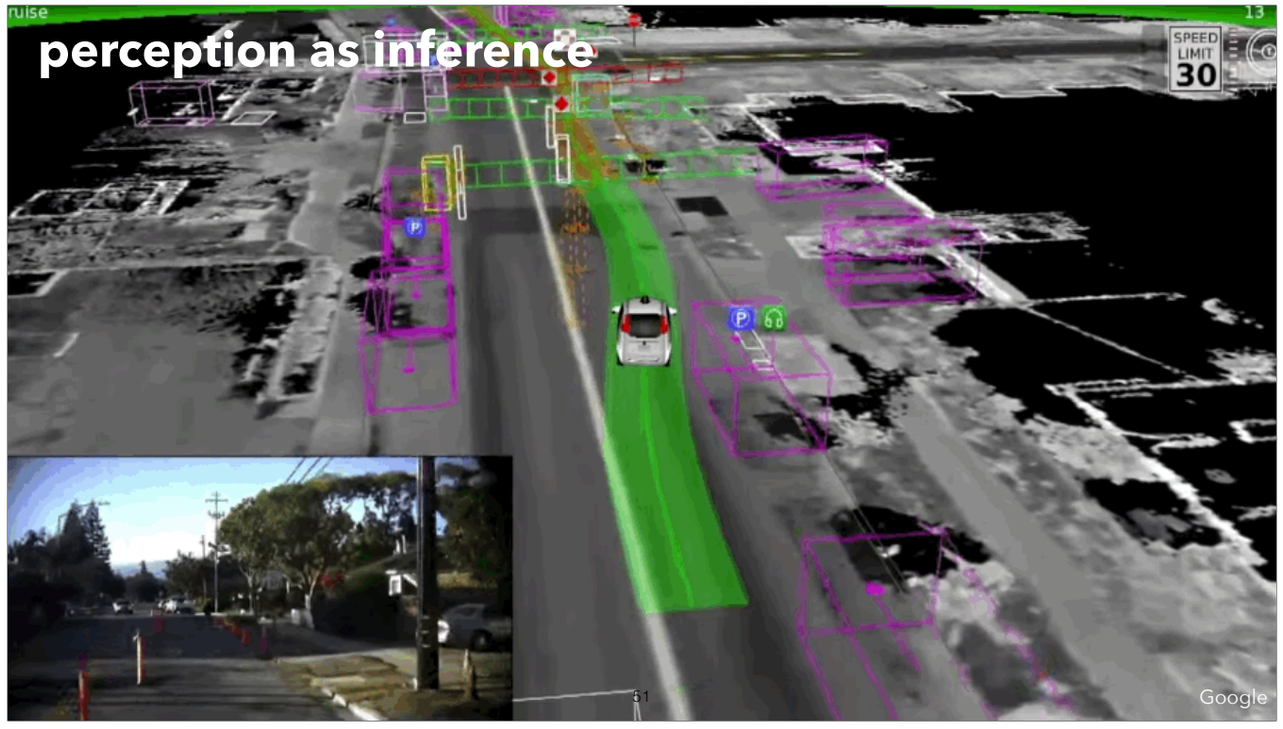
- what objects are around us
- how far
- who are around us
- what are they thinking
- what is going to happen

inference

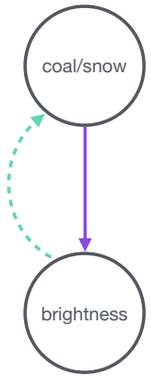
what can we observe

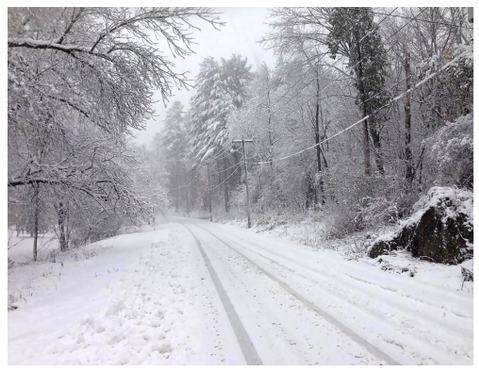
- incoming photons
- air vibrations
- temperature fluctuations
- certain molecules

perception as inference



perception as inference





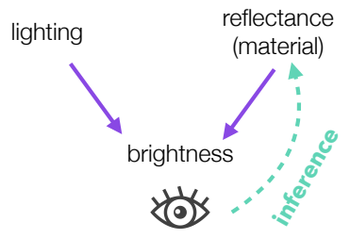
snow in the evening seems white

perception as inference

coal in the sun still seems black, even though more photons arrive from it to our eyes than snow in the dark

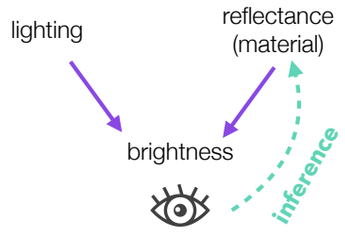
perception as inference

lighting x reflectance = brightness

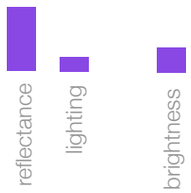


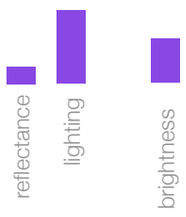
perception as inference

lighting x reflectance = brightness

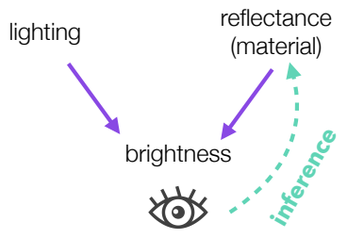


perception as inference



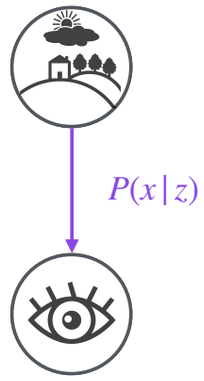


lighting x reflectance = brightness

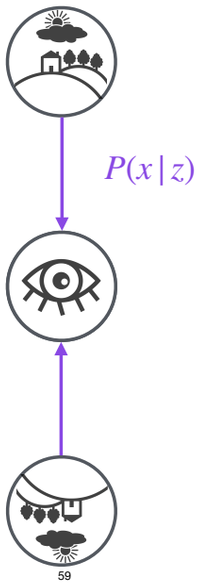


perception as inference

- generative direction
  - if the environment was in this state, what would I see?

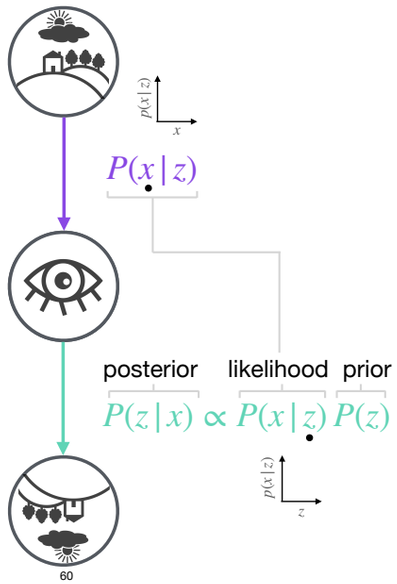


perception as inference

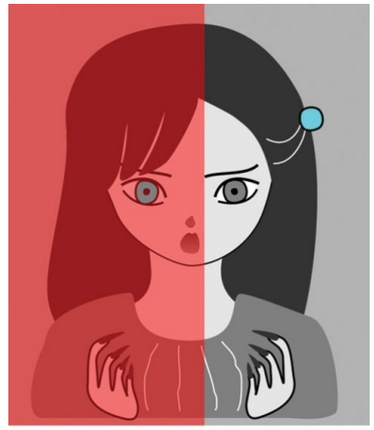


perception as inference

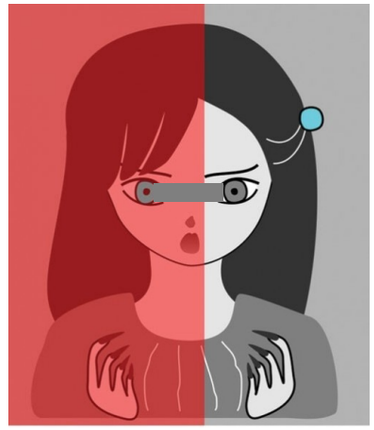
- inverse direction
  - if I see this, what state is the environment in?
    - inverting the generative model
    - “vision is inverse graphics”
    - Bayesian inference
    - recognition model



perception as inference

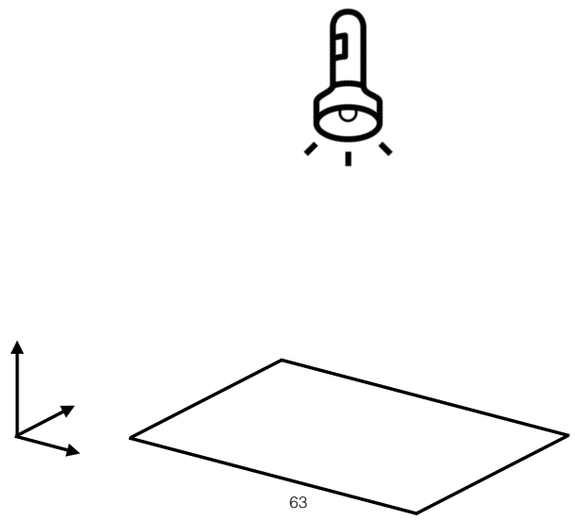


perception as inference



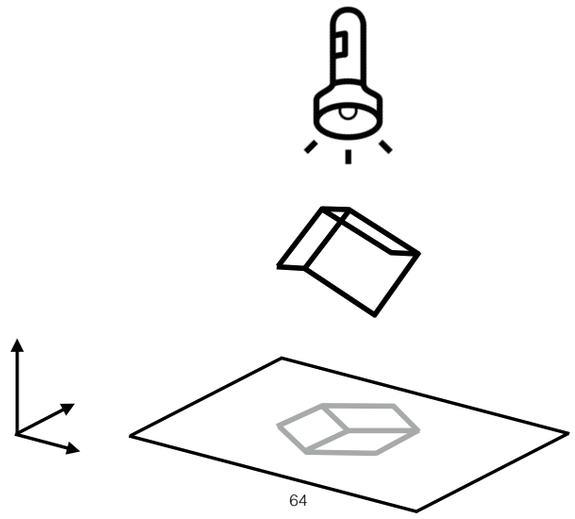
- what we see is not the data but an interpretation of the data
- ‘unconscious inference’ over latent variables

perception as inference



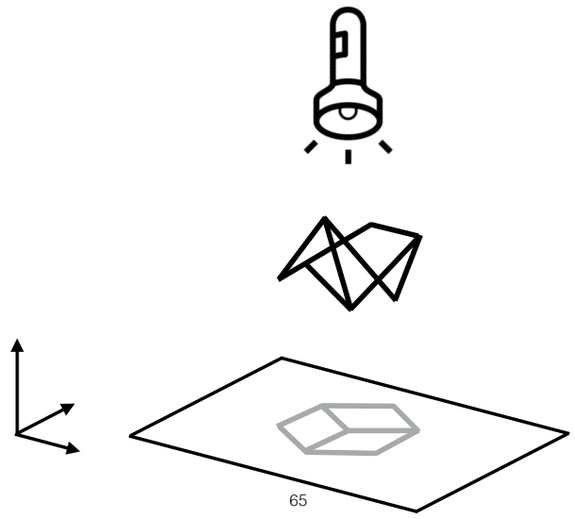
Kersten & Yuille, 2003

perception as inference



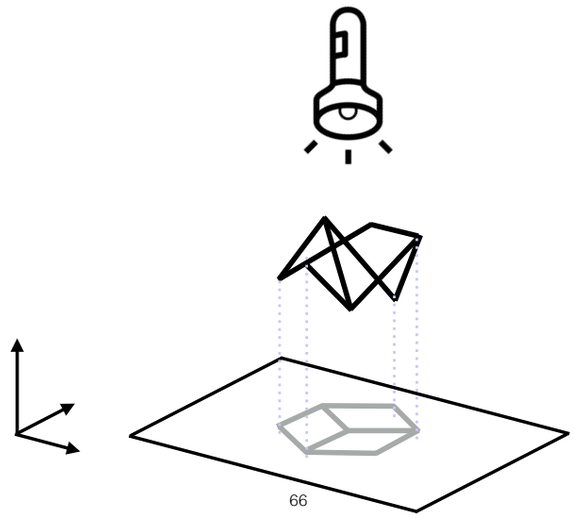
Kersten & Yuille, 2003

perception as inference



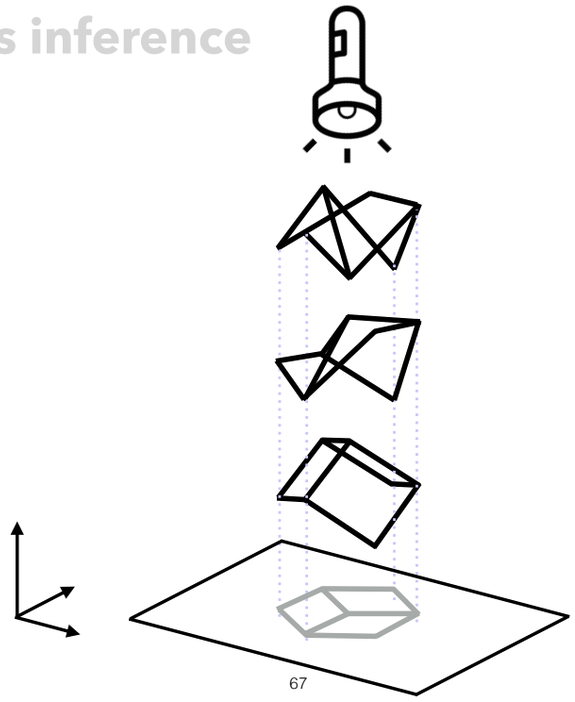
Kersten & Yuille, 2003

perception as inference



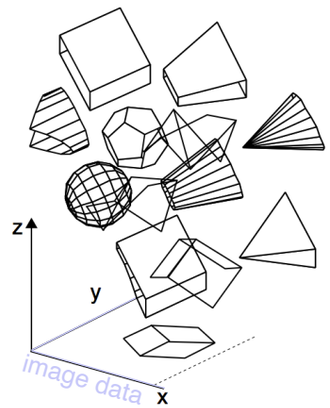
Kersten & Yuille, 2003

perception as inference



Kersten & Yuille, 2003

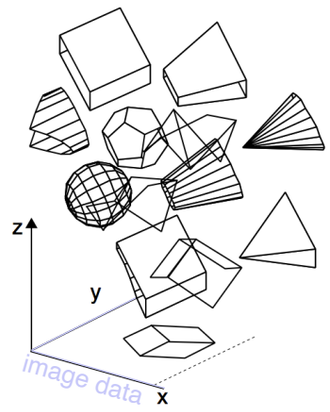
perception as inference



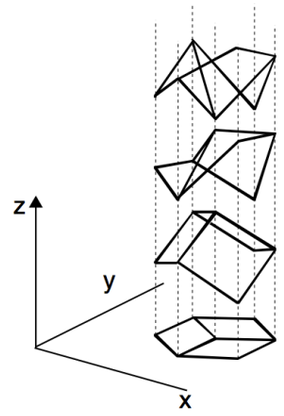
hypotheses sampled
from prior

Kersten & Yuille, 2003

perception as inference



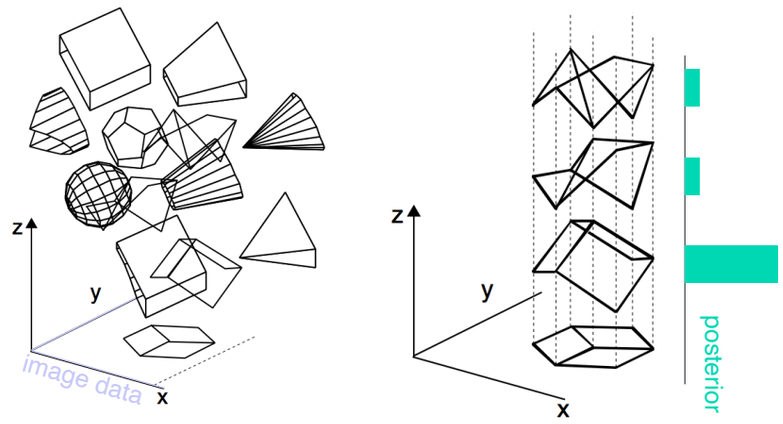
hypotheses sampled
from prior



hypotheses with
nonzero likelihood

Kersten & Yuille, 2003

perception as inference

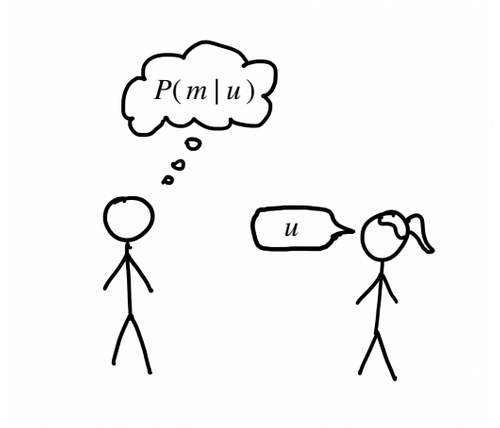


hypotheses sampled from prior

hypotheses with nonzero likelihood

Kersten & Yuille, 2003

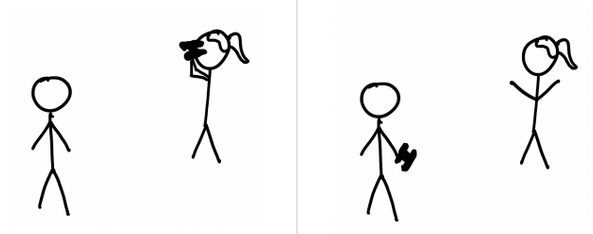
perception as inference



Goodman & Frank, 2016

perception as inference

"the girl saw the boy with the telescope"



j_1 j_2

Goodman & Frank, 2016

perception as inference

McGurk & MacDonald, 1976

generative compression

x

input

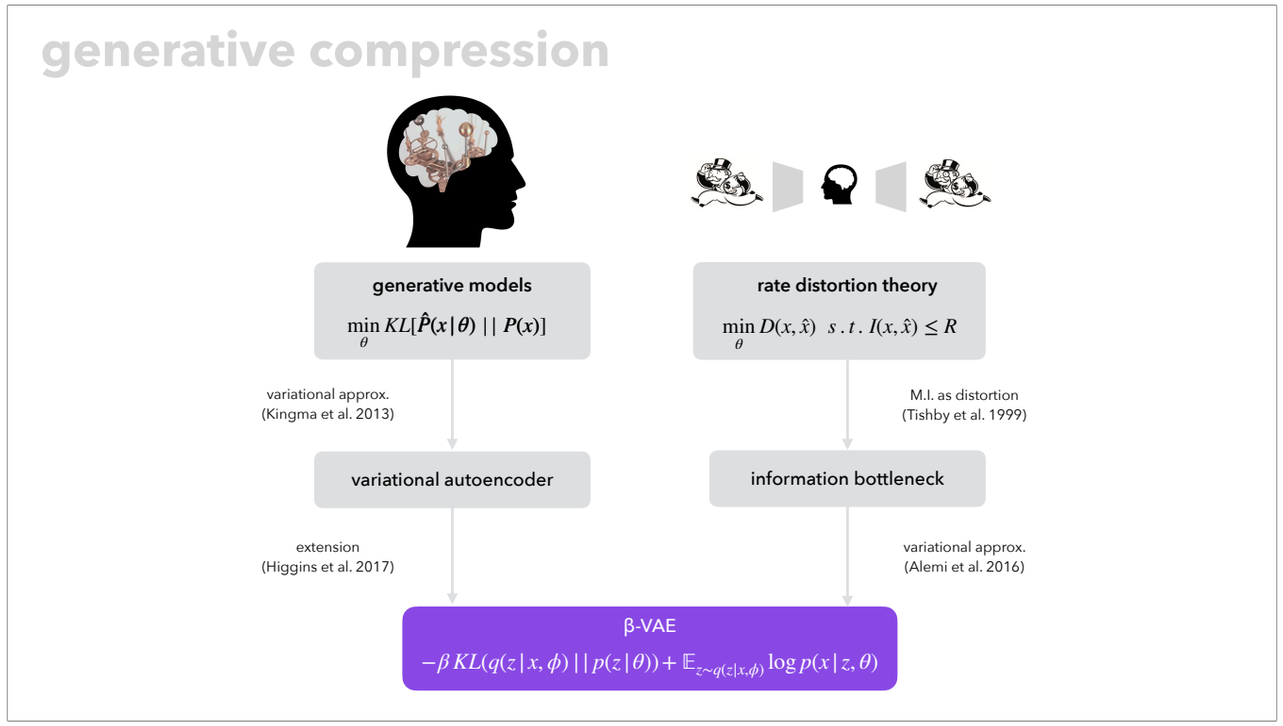
z

memory trace
(code)

$\hat{x}$

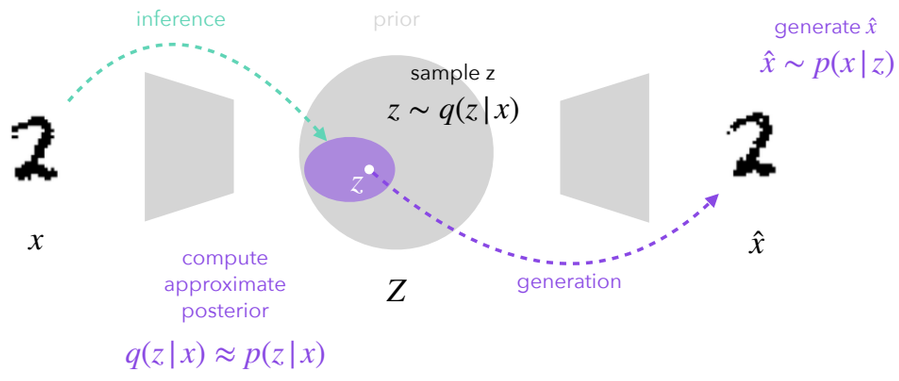
reconstruction

Kingma & Welling, 2014; Rezende et al, 2014



beta-VAE can be seen as both as a generative model + an inference method for approximately inverting it, and also as a lossy compression algorithm

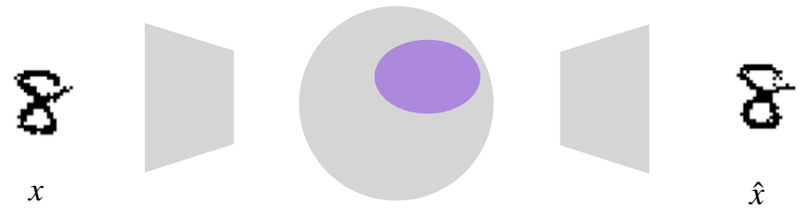
variational autoencoder



Kingma & Welling, 2014; Rezende et al, 2014

1. compute an approximate posterior based on sample x
2. sample a single z from this posterior
3. conditioning the generative network on this sample z, generate a reconstruction

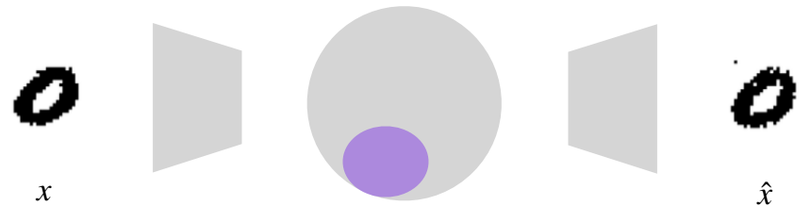
variational autoencoder



Kingma & Welling, 2014; Rezende et al, 2014

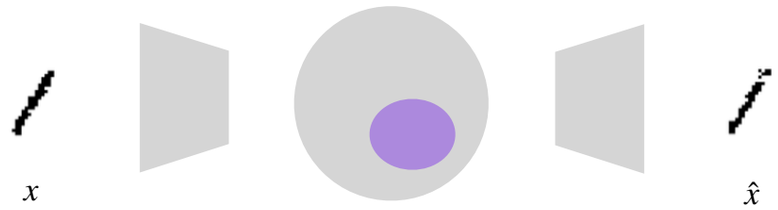
different x-es will correspond to different posteriors and therefore different reconstructions, but in case the posteriors overlap, the memory traces for different stimuli might be confused

variational autoencoder



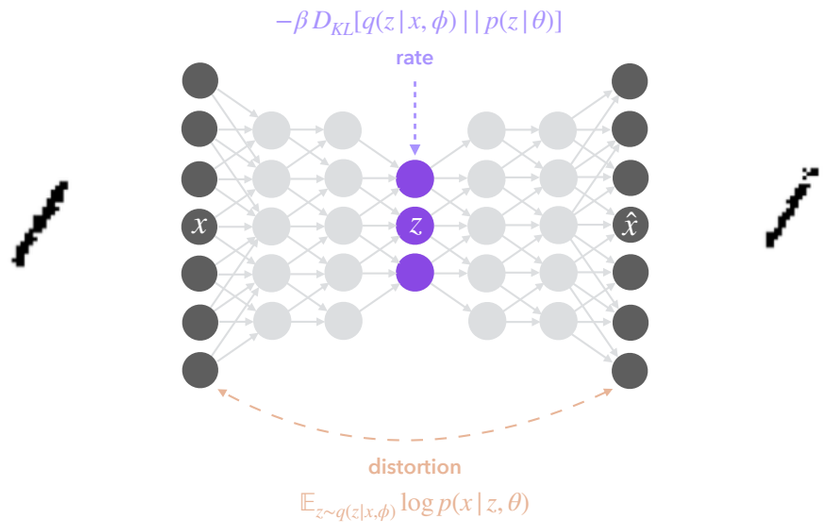
Kingma & Welling, 2014; Rezende et al, 2014

variational autoencoder



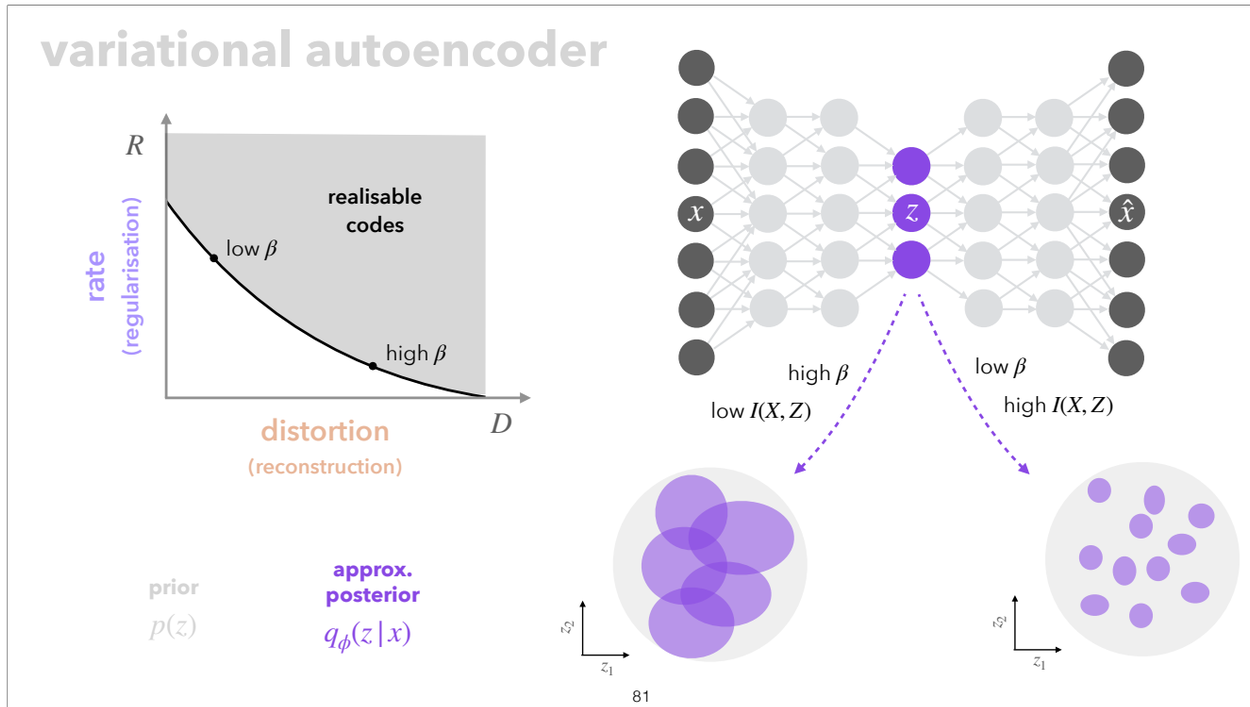
Kingma & Welling, 2014; Rezende et al, 2014

variational autoencoder



Kingma & Welling, 2014; Rezende et al, 2014

- in a VAE, both the inference method and the generative model are neural networks
- the objective function for training the beta-VAE consists of two terms, which in the compression view correspond to rate and distortion. In the generative modelling literature these are called ‘regularisation term’ and ‘reconstruction’ term respectively
- D_KL is KL-divergence, a form of (almost) distance between probability distributions



- intuitively, when KL term is weighted more strongly (using beta), posteriors have to resemble the wide and spherical prior more
- this makes it more likely that they overlap and makes the variance of the sampled z larger
- this corresponds to a lower rate (less information in the memory trace about the original input)

variational autoencoder

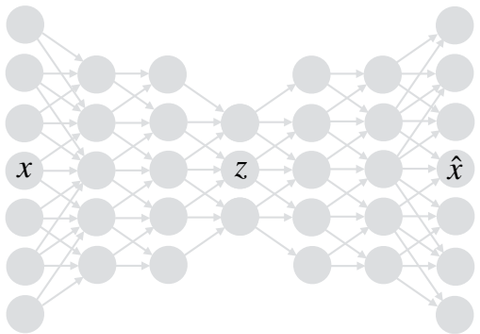
chess positions

sketches

word lists

```
butter
food
eat
sandwich
…
```

x



```
food
butter
bread
sandwich
…
```

$\sim P(\hat{x}\,|\,x)$

3000 chess games (FICS database)

quickdraw75k object pairs

(small subset of) wikipedia

$\mathscr{D}$

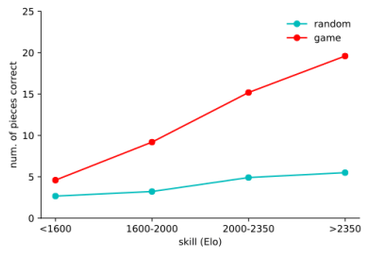
Nagy et al. 2020

Bates & Jacobs 2020,Hedayati et al.

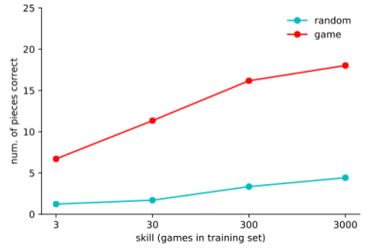
domain expertise

reconstruct state of chess board after 5 seconds viewing time

domain expertise



Gobet & Simon, 1995



Nagy et al. 2020

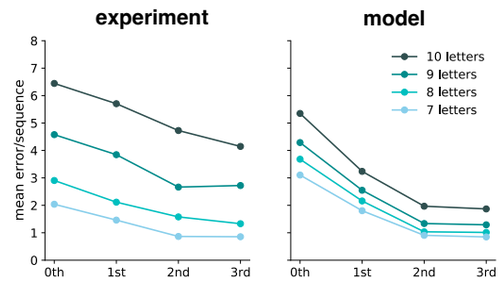
- accuracy increases with domain expertise, but only for configurations from actual chess games
- for randomly shuffled configurations, skill does not help much
- skill in model is amount of training of beta-VAE

domain congruence

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

- reconstruct words with statistics that are increasingly congruent with the statistics of english language
- very similar to chess example, but there we had two degrees of congruence - random and game

domain congruence

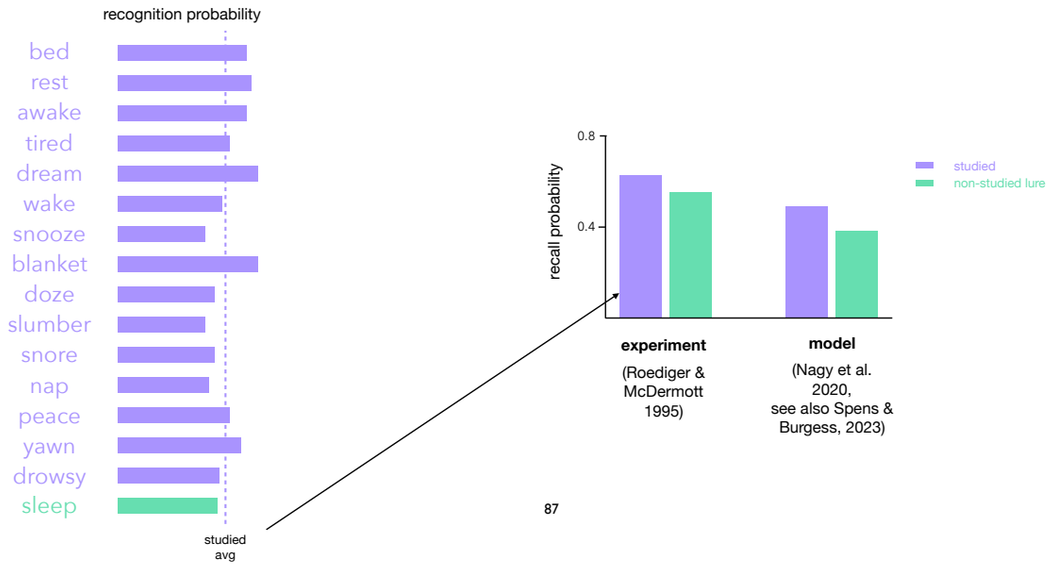


uniform	1st order	2nd order	3rd order	full word
RCIFODWVIL	TNEOOESHHE	HIRTOCLENO	BETEREASYS	PLANTATION
GKTODKPENF	INOLGGOLVN	DOVEECOFOF	CRAGETTERS	FLASHLIGHT
TZXKHAWCCF	PDOASLOTPP	SESERAICCG	TOWERSIBLE	UNCOMMONLY
NGORHQIYWB	AEOCAOIAON	AREDAGORTZ	DEEMEREANY	ALIENATION
BVNJSYZXUA	IRCRENFCTN	CUNSIGOSUR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

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

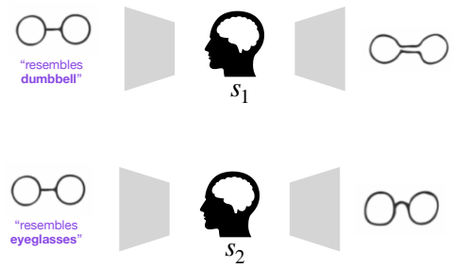
- accuracy increases with degree of congruence with english, and decreases with length of word

gist-based distortions



- lure is a semantically related word that was not in the list
- lures are falsely recognised with comparable probability to studied items
- effect of regenerating the list from stored latent representation

gist-based distortions



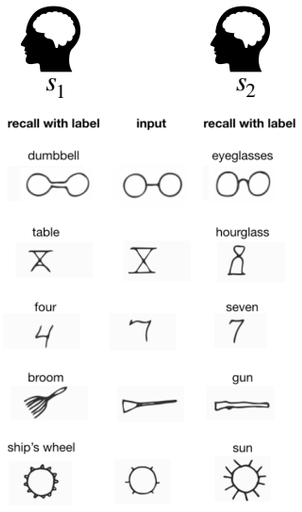
- subjects view ambiguous stimuli, with different labels for different subjects
- labels introduce label-specific distortions in recall

gist-based distortions

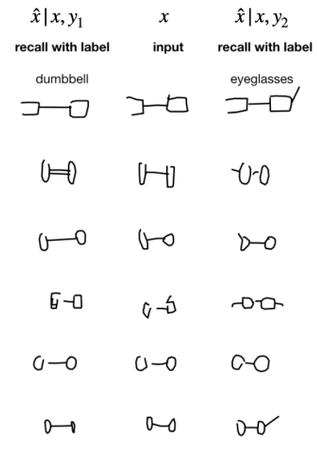
recall with label	input	recall with label
dumbbell		eyeglasses
table		hourglass
four		seven
broom		gun
ship's wheel		sun

Carmichael et al., 1932

gist-based distortions

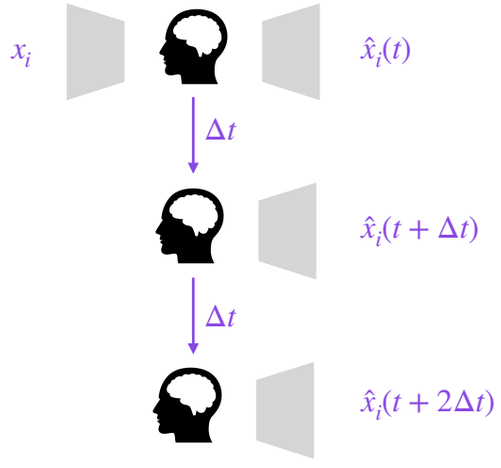


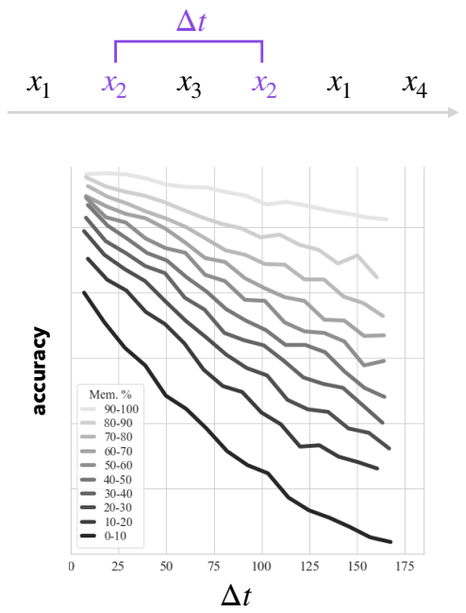
Carmichael et al., 1932



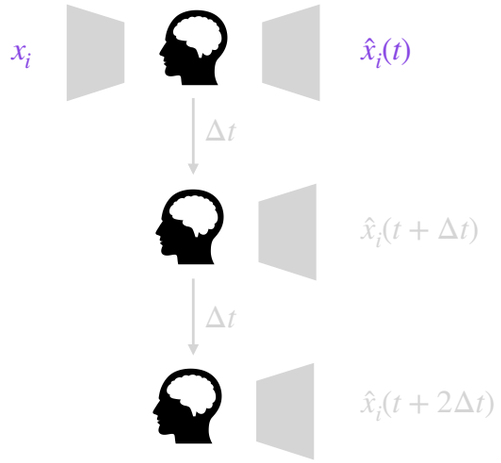
Nagy et al. 2020

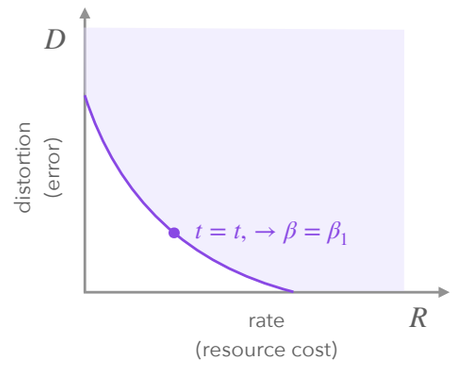
delayed recall



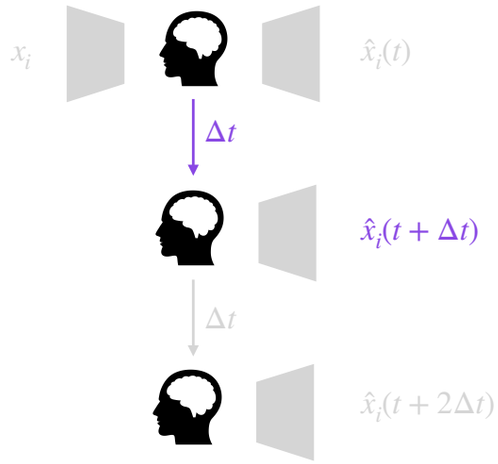


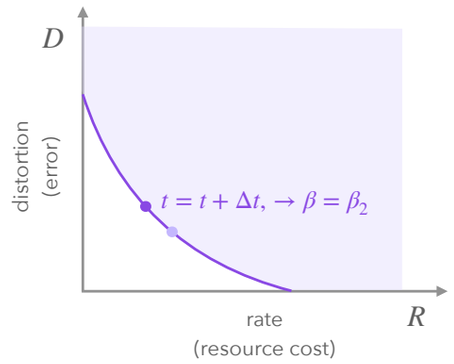
delayed recall



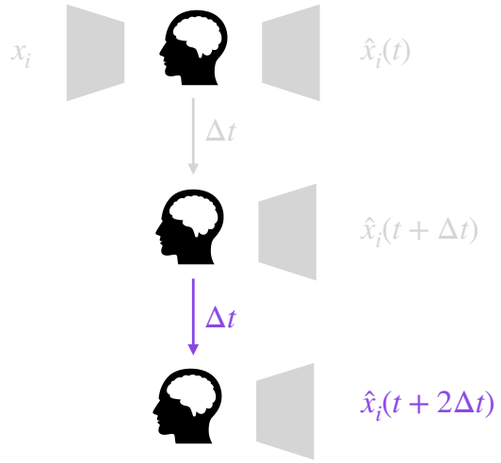


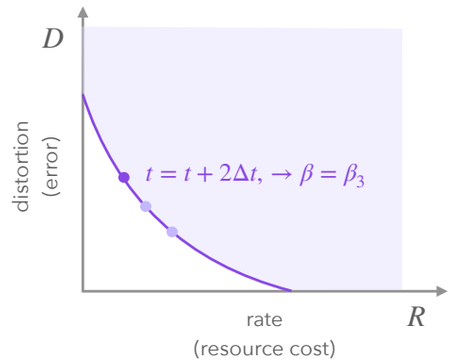
delayed recall



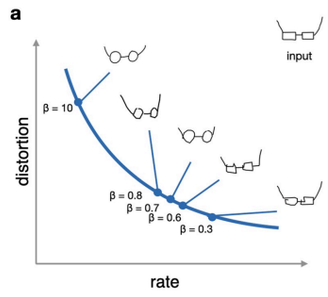


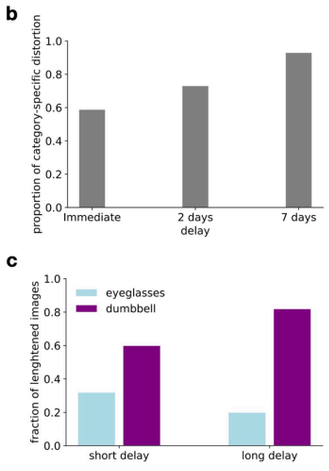
delayed recall

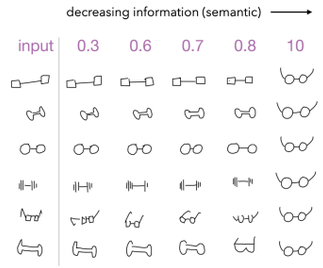




delayed recall

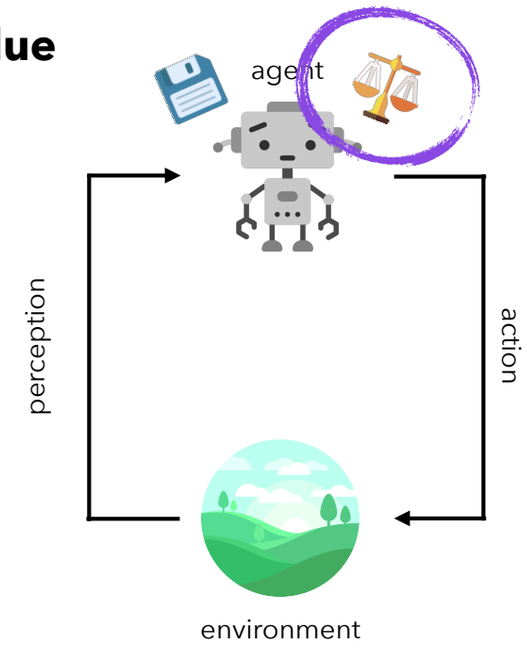




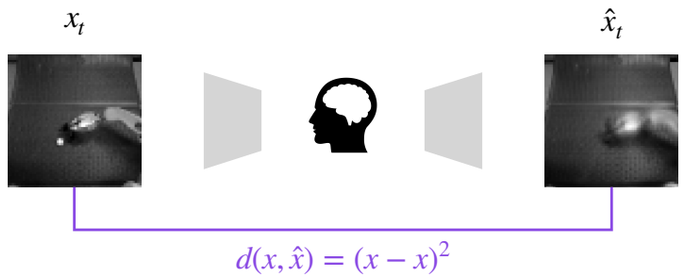


Nagy et al. 2020

goals, tasks, value



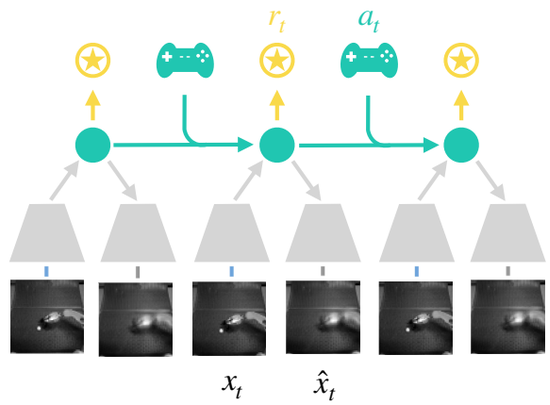
goals, tasks, value



see Bates & Jacobs, 2020

- in engineering contexts, typical distortion is MSE in pixel space
- for a robot that needs to manipulate a small white ball, removing this ball from the input leads to almost negligible reconstruction error
- need to overweight errors that are relevant to rewards

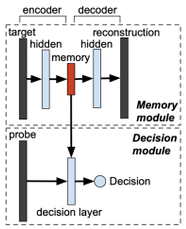
goals, tasks, value

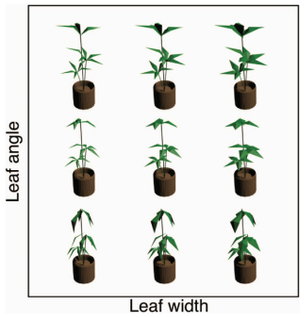


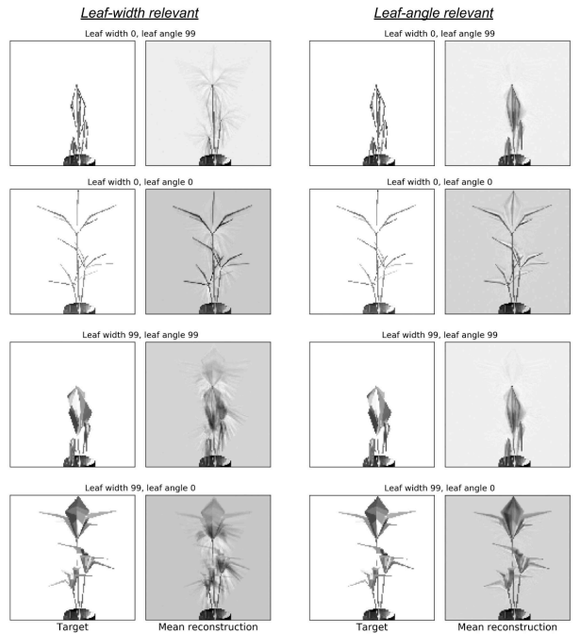
based on Hafner et al., 2023

- this can be incorporated in VAEs, for example this is the basis of the DREAMER v3 model that you've seen in lecture 5

goals, tasks, value

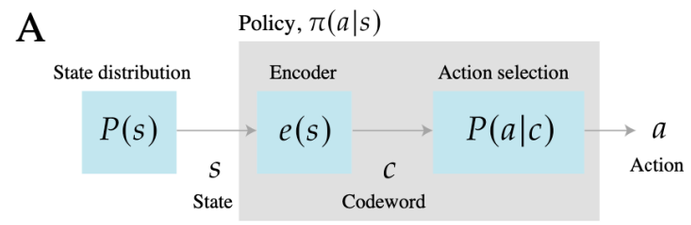






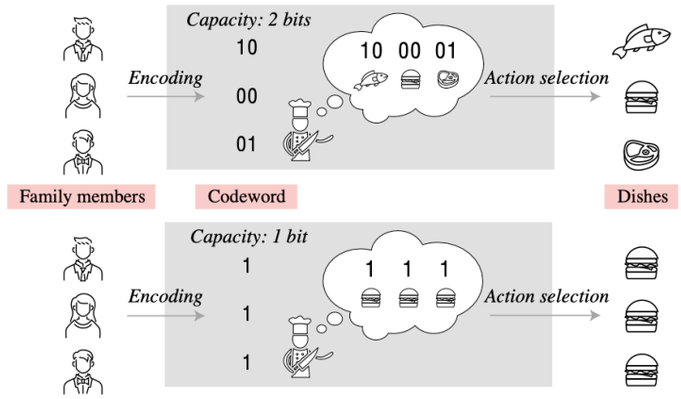
Bates & Jacobs, 2020

- humans are also sensitive to reward-relevance in memory accuracy



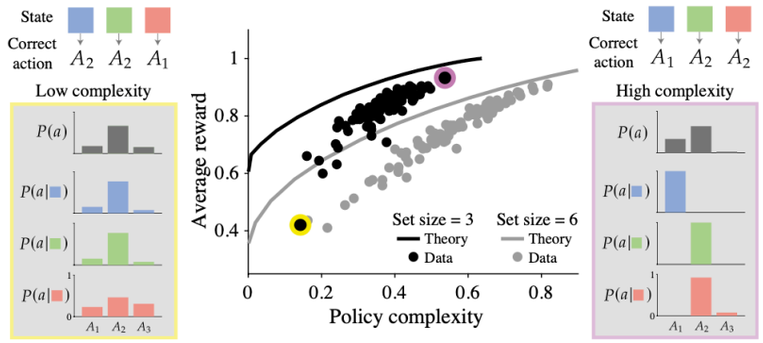
Lai & Gershman, 2021

goals, tasks, value

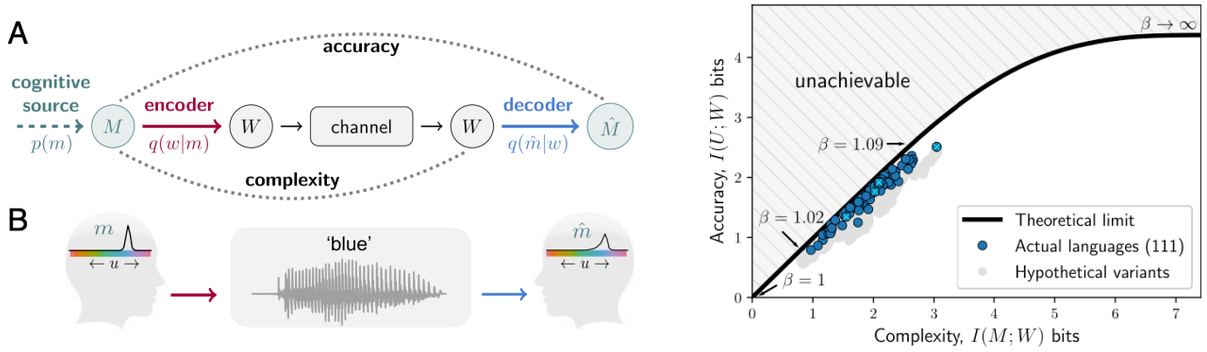


Lai & Gershman, 2021

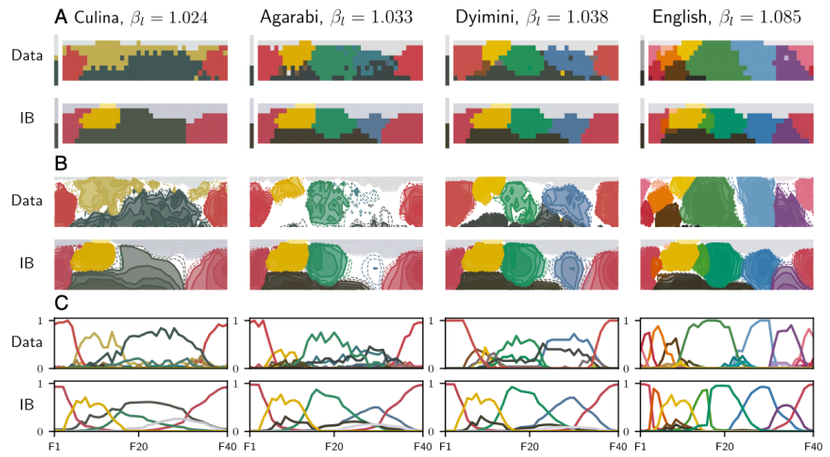
goals, tasks, value



Lai & Gershman, 2021



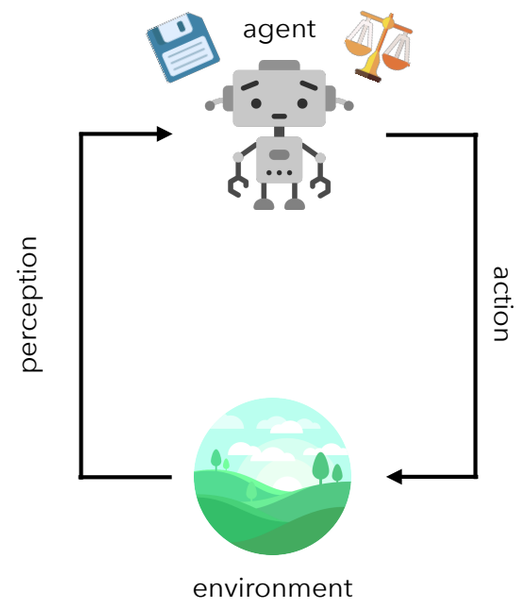
Zaslavsky et al 2018



Zaslavsky et al 2018

summary

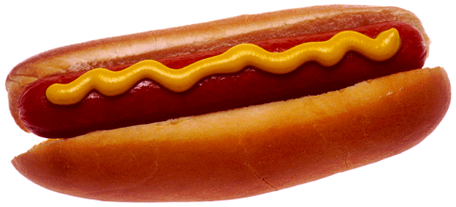
- focus on constraints on **memory**
  - lossless compression
  - lossy compression
  - generative compression
    - perception as bayesian inference
    - human memory distortions



can information capture all constraints?

xkcd

Next week: Concepts & Categories



Is a hotdog a sandwich?