Probabilistic models of cognition

Karla Štěpánová

CIIRC CTU in Prague,

Robotics and machine perception

https://www.karlastepanova.cz (personal webpage)

https://imitrob.ciirc.cvut.cz (Imitation learning group)

Karla.Stepanova@cvut.cz

Computational cognitive modeling

➨ **Computational cognitive modelling**

= simulations of complex mental processes in different areas of cognition

Goal is to describe, understand, model and predict observed human behaviour

➨ **Cognition**

= mental process of knowing, including aspects such as awareness, perception, reasoning and judgement

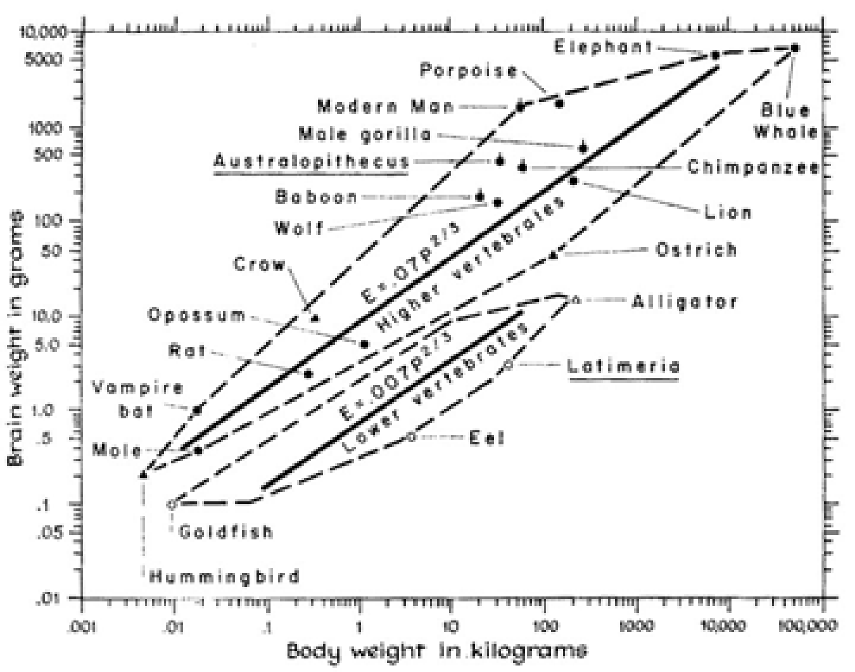
Latin word cognition: -co (intensive) + nosecere (to learn)

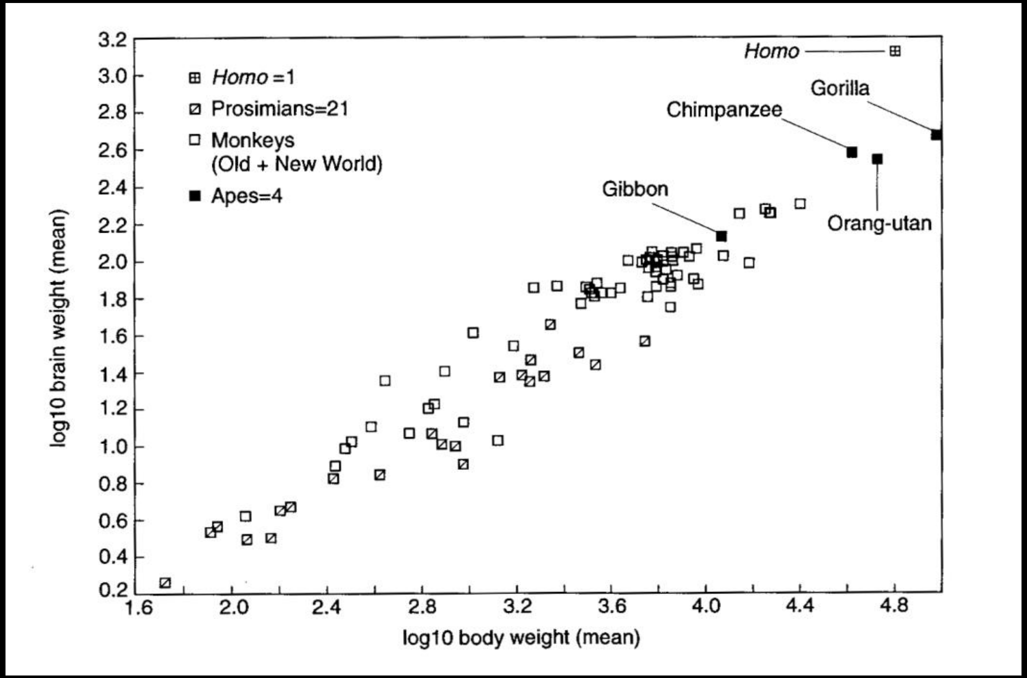
➨ **Modeling**

Data never speak for themselves, require a model to be understood and explained

Several alternative models -> compare -> quantitative evaluation and intellectual judgement

Motivation

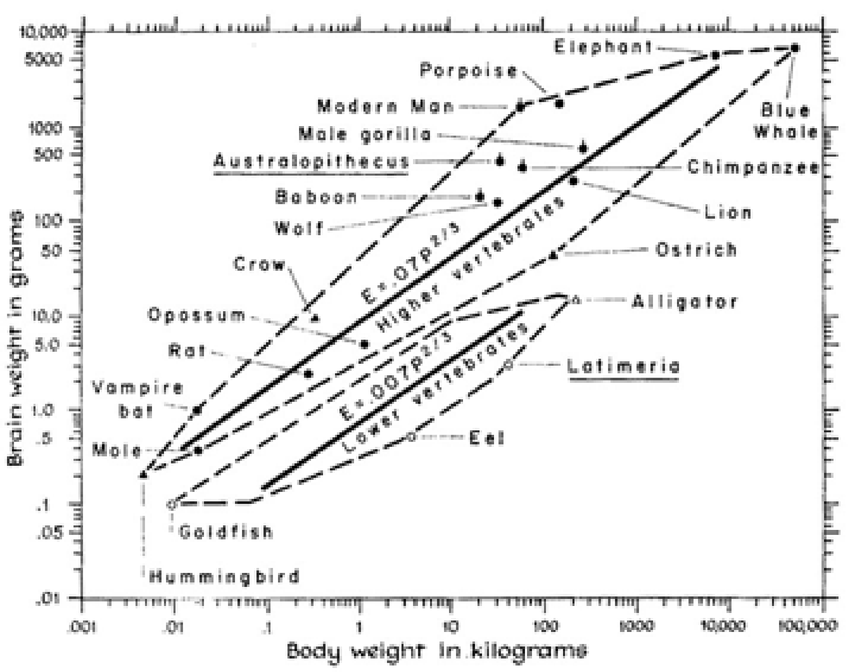


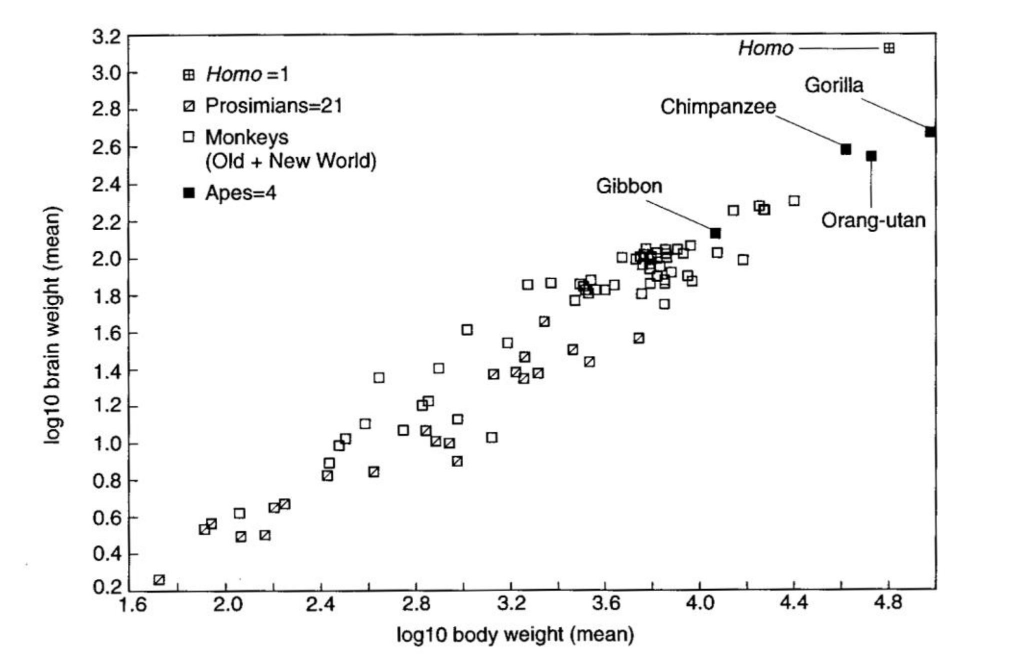


Brain-to-body mass ratio, Encephalization Quotient

Treeshrew (squirrel)

Motivation





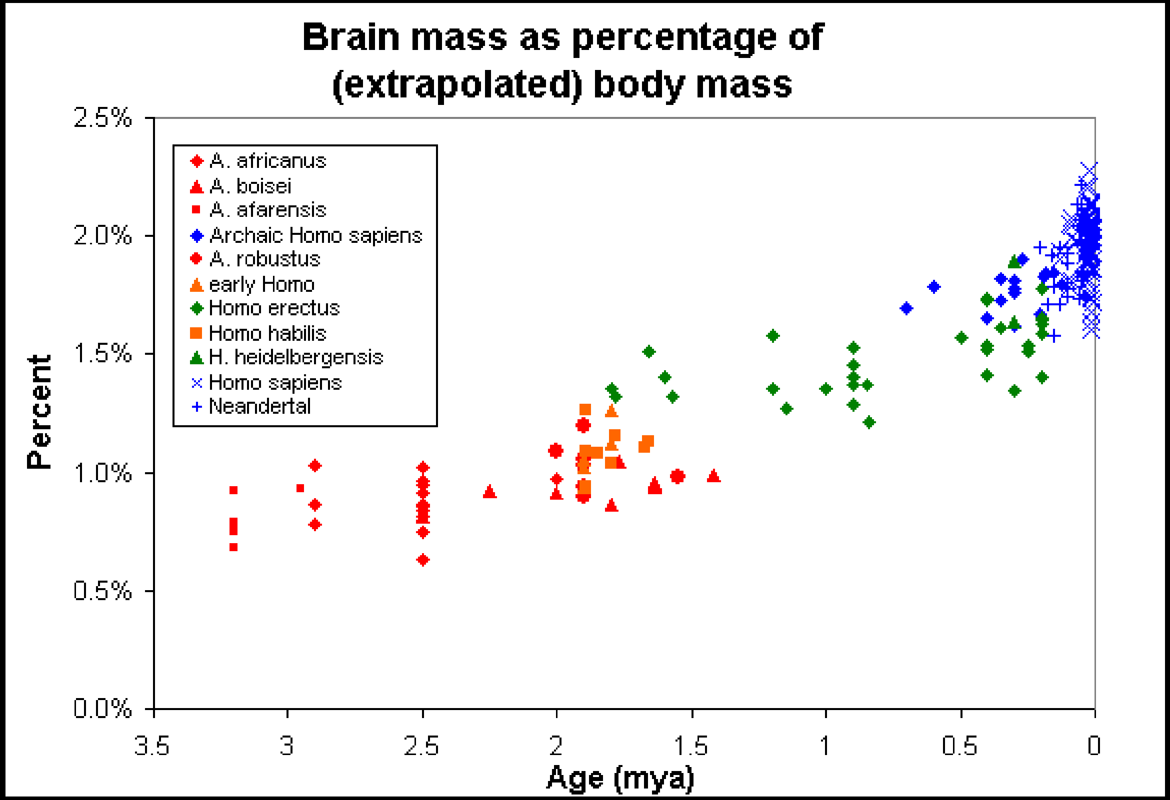
The environment and lifestyle shape how the brain looks like

- Ants – path integration
- Migrating birds – compass based on sky
- Honeybees – dance to show food, internal clock to compensate movement of the sun

Steven Pinker: How the mind works

Brain-to-body mass ratio, Encephalization Quotient

Motivation



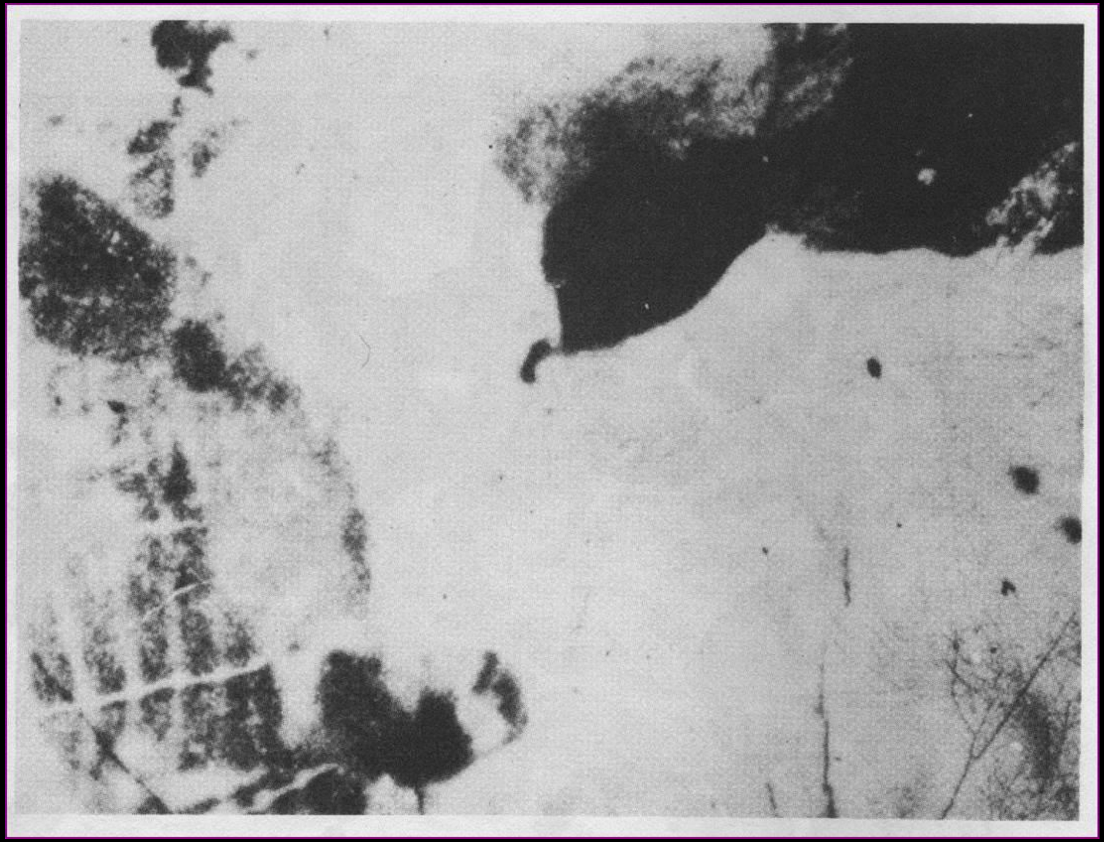
Language

Technology

Art, culture, high tech

Henneberg and de Miguel (2004): Variation in hominid brain size. How much is due to method?

Motivation



Motivation

Sotto, E. (2007). *When teaching becomes learning: A theory and practice of teaching*. Bloomsbury Publishing.

Motivation

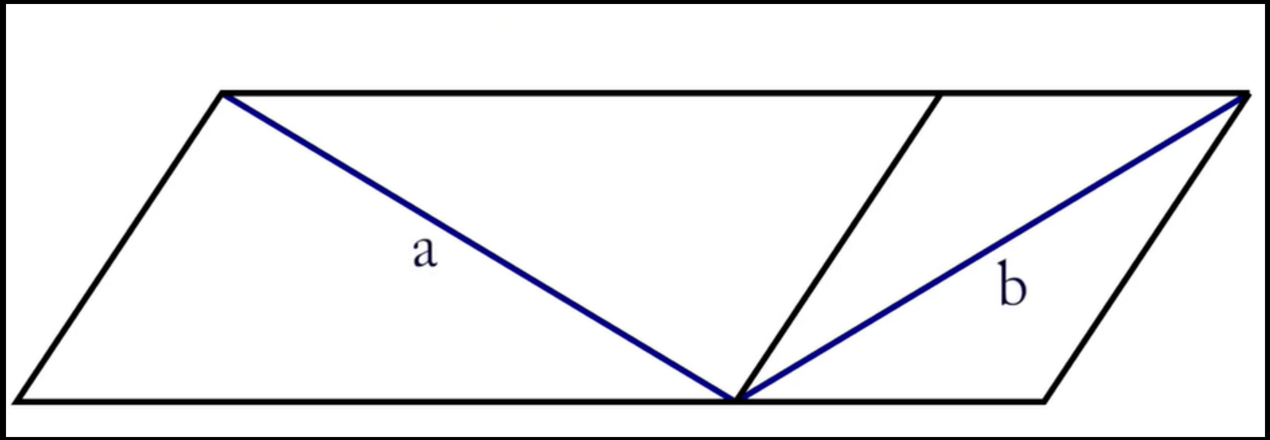
Sotto, E. (2007). *When teaching becomes learning: A theory and practice of teaching*. Bloomsbury Publishing.

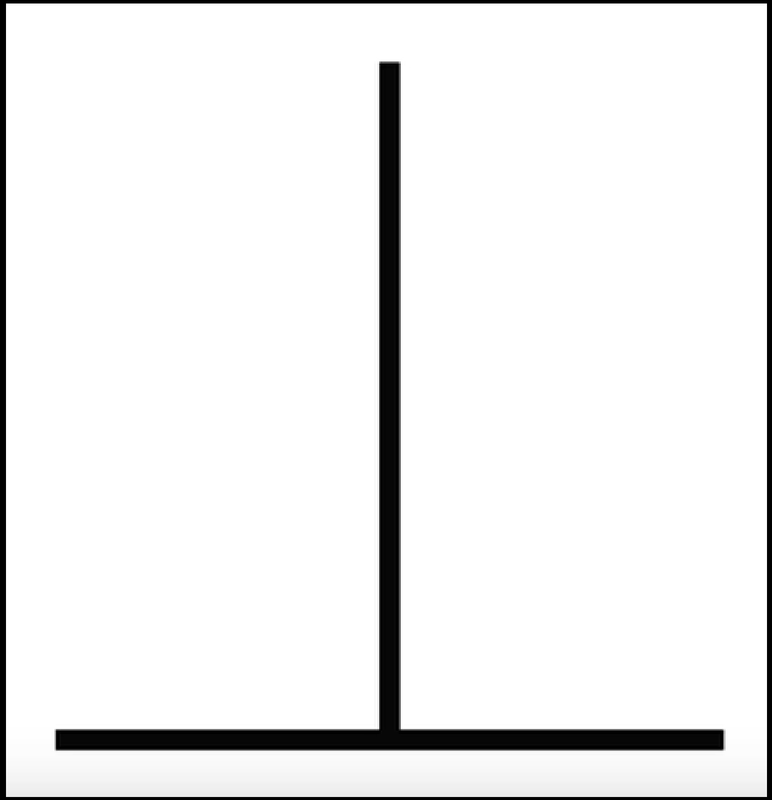
Motivation

Sotto, E. (2007). *When teaching becomes learning: A theory and practice of teaching*. Bloomsbury Publishing.

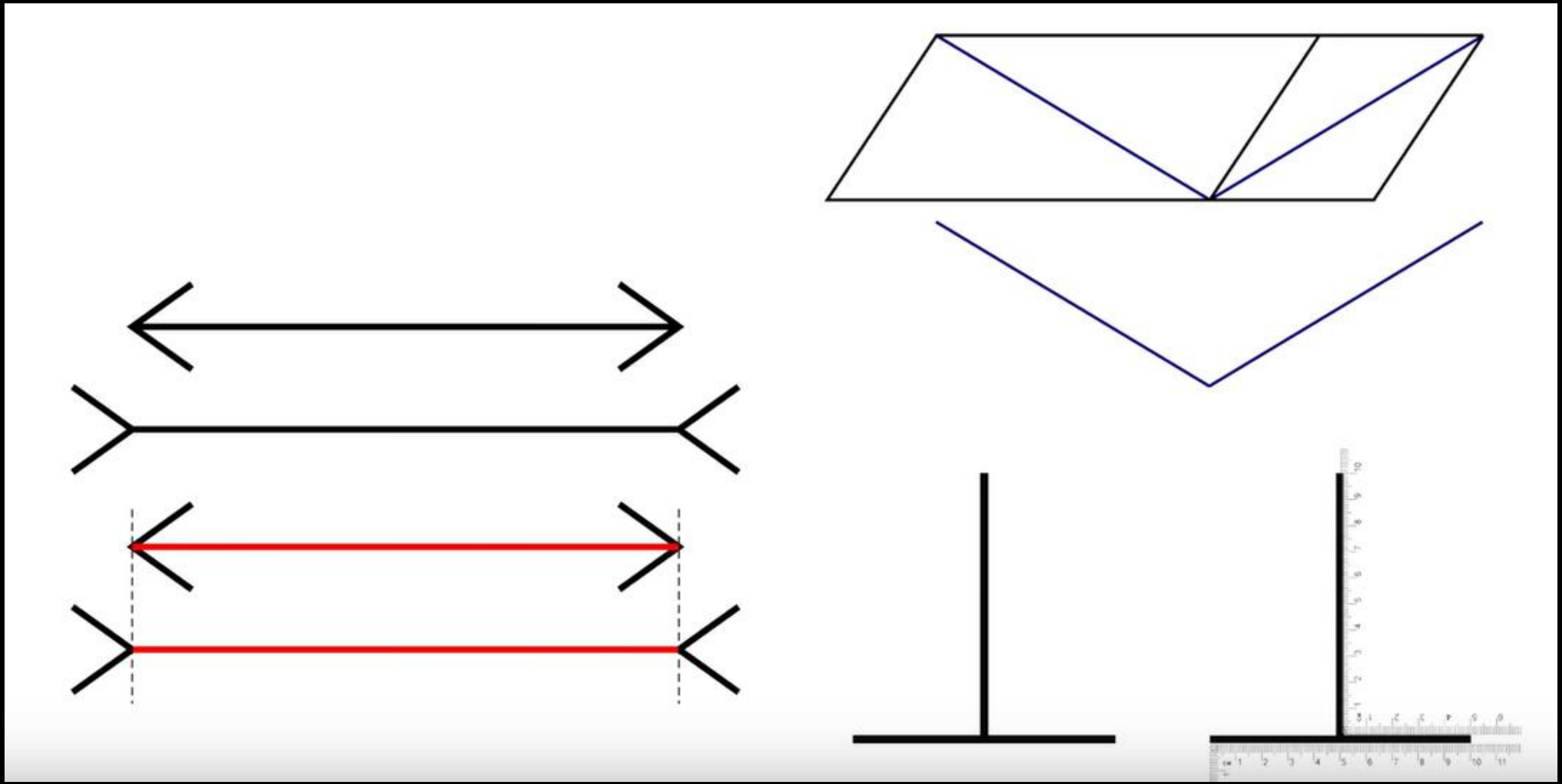
Motivation

Motivation



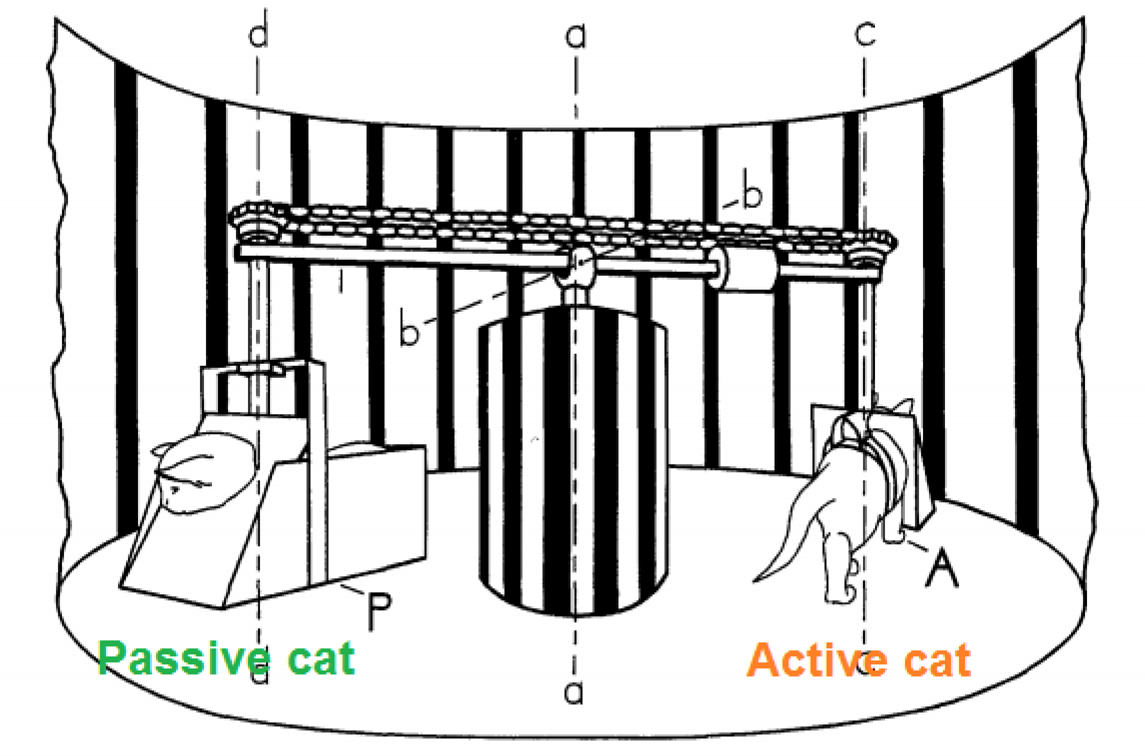


Motivation



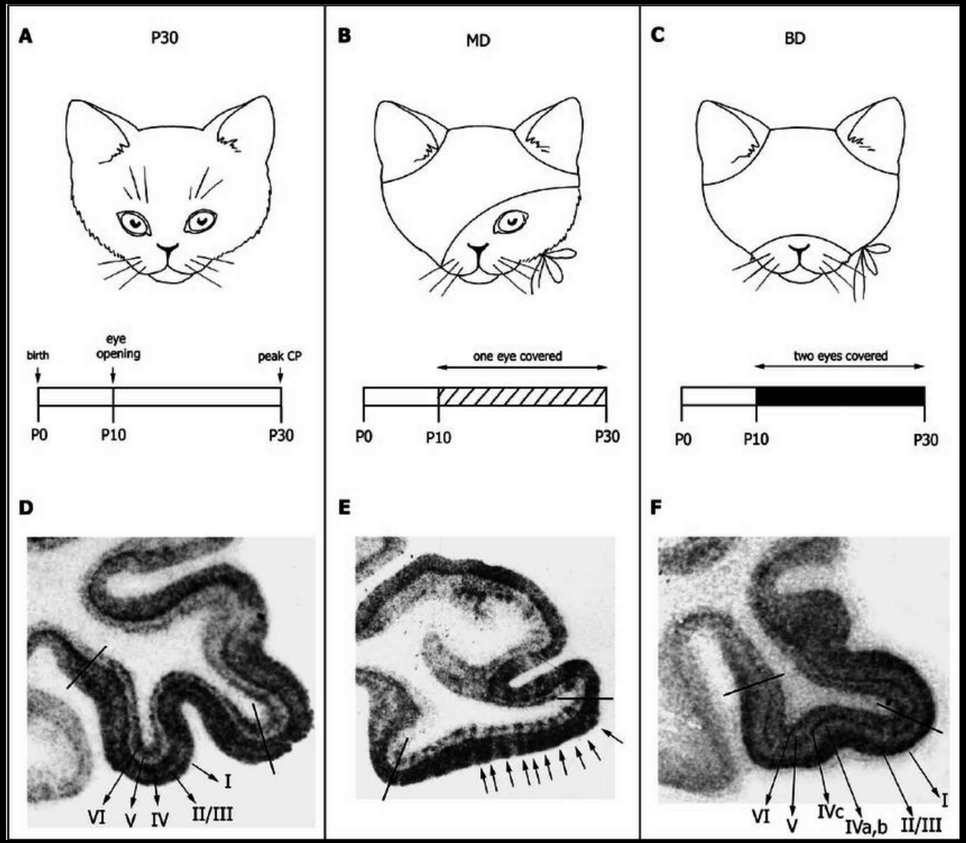
Based on where you live you are less/more prone to these illusions

Motivation



Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *Journal of comparative and physiological psychology, 56*(5), 872.

Motivation



Directly affecting visual cortex development of the cat, Hubel + Wiesel

Brain

- "A human brain has about 10^{15} synapses (10^{11} neurons) which operate at about 10^2 per second implying about 10^{17} bit ops per second" J. Langford
  - …a transcription of 1 second of brain activity at the neural spike level would fill up about 4 000 ordinary 3 TB hard drive
- …and consumes 20% of body's oxygen (2% metabolism, aprox.1.3kg)

- Is it worth?

Kandel, E.R., Schwartz, J.H. and Jessell, T.M. eds., 2000. *Principles of neural science* (Vol. 4, pp. 1227-1246). New York: McGraw-hill.

Brain x LLMs

- “A human brain has about 10^{15} synapses which operate at about 10^{2} per second implying about 10^{17} bit ops per second” J. Langford
- …a transcription of 1 second of brain activity at the neural spike level would fill up about 40 000 ordinary 300 Gb hard drive
- …and consumes 20% of body’s oxygen (aprox.1.3kg)

Release	Model	Size	Paper
2019	GPT-2	1.5B	Language Models are Unsupervised Multitask Learners
2020	GPT-3	175B	Language Models are Few-Shot Learners
2021	Gopher	280B	Scaling Language Models: Methods, Analysis & Insights from Training Gopher
2022	PaLM	540B	PaLM: Scaling Language Modeling with Pathways
2022	Chinchilla	70B	Training Compute-Optimal Large Language Models
2022	OPT	175B	OPT: Open Pre-trained Transformer Language Models
2022	BLOOM	176B	BLOOM: A 176B-Parameter Open-Access Multilingual Language Model
2022	Galactica	120B	Galactica: A Large Language Model for Science
2023	LLaMA	65B	LLaMA: Open and Efficient Foundation Language Models

Some of the popular LLMs architectures. Image by Author

To make a particular example, it is known that LLaMA used a training dataset containing **1.4 trillion tokens** with a total size of **4.6 terabytes**!

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

2048 GPUs x $3.93 GPU per hour x 24 hours x 21 days =

4.05 million dollars

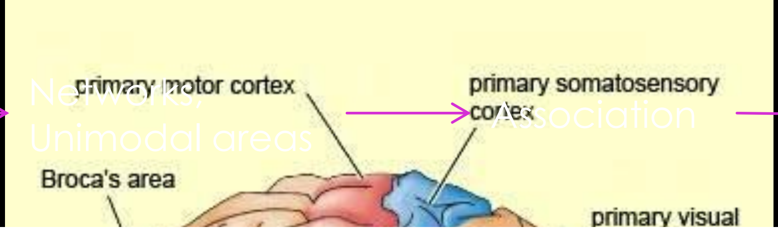
cost of training GPT-3, and the authors got **355 GPU-years and 4.6 million dollars.**

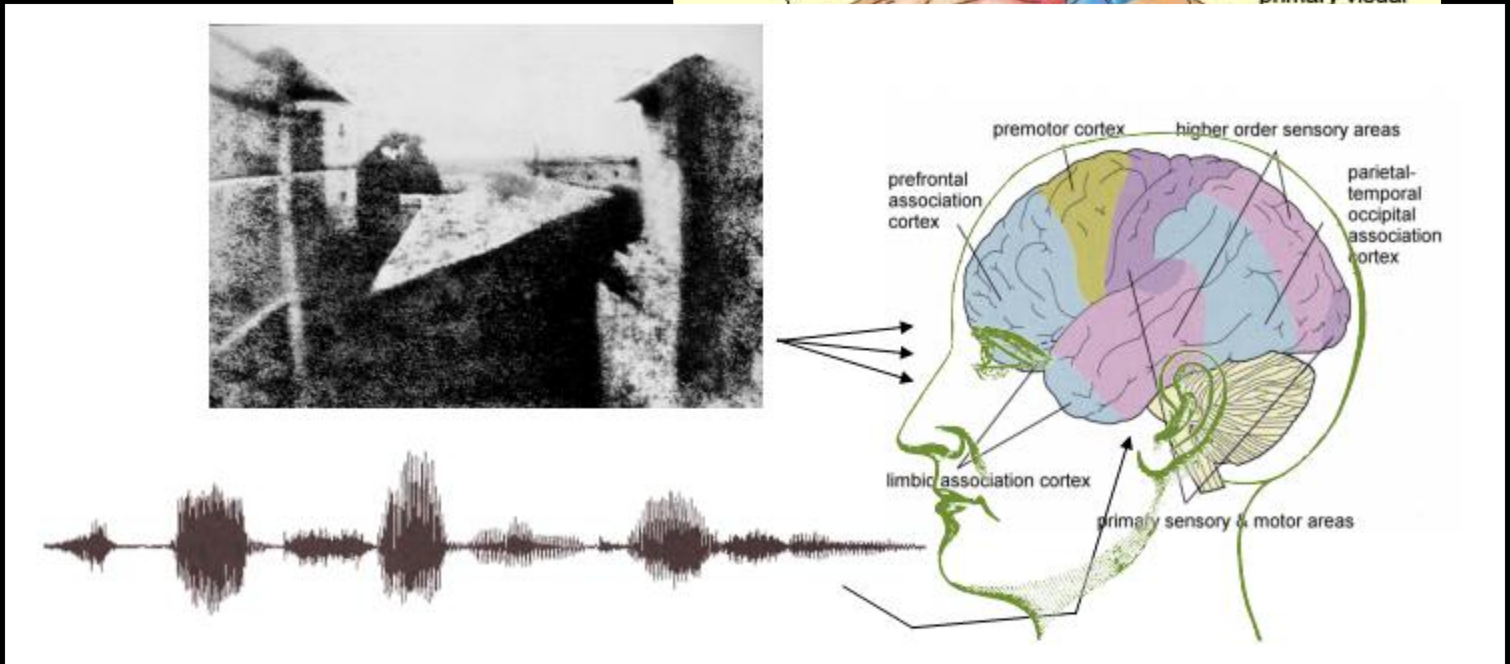
https://towardsdatascience.com/behind-the-millions-estimating-the-scale-of-large-language-models-97bd7287fb6b

Kandel, E.R., Schwartz, J.H. and Jessell, T.M. eds., 2000. *Principles of neural science* (Vol. 4, pp. 1227-1246). New York: McGraw-hill.

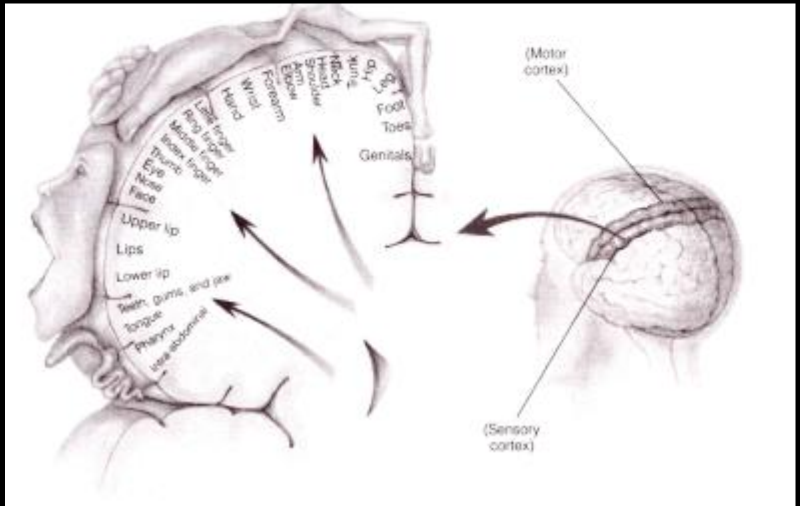
Brain - Information processing

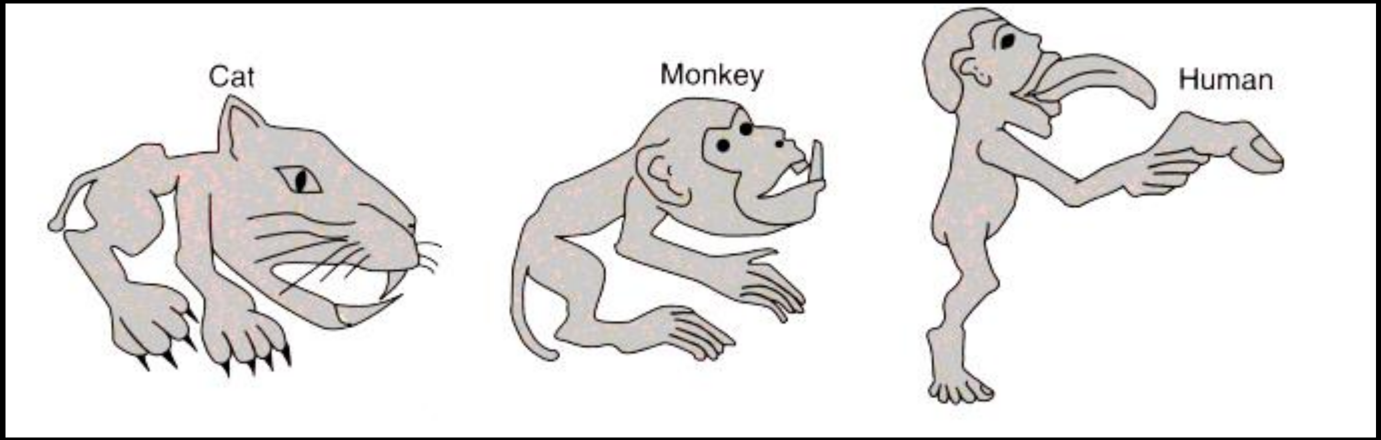
Single units → Networks, Unimodal areas → Association → Evolution in time, reasoning, induction



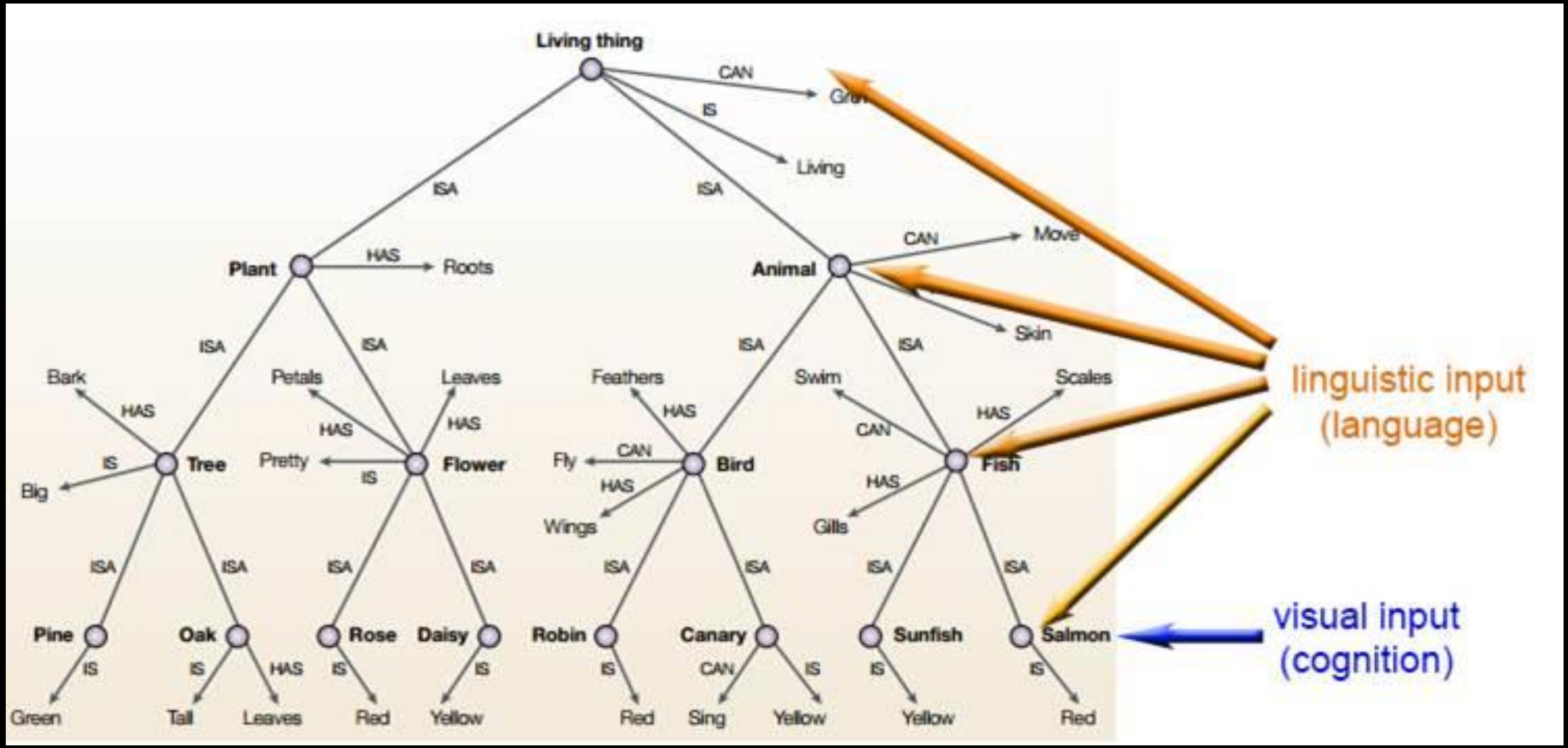


Brain – information processing





Brain - Data processing and representation



Cognitive models

- Traditional models of cognition
  - Connectionism — Adaptivity
  - Rule-based (Minski 1968, a priori rules) — Apriori knowledge
  - Parametric model-based — Adaptivity + apriori knowledge

Computational complexity

- Parametric model-based models: Parameters can capture variabilities and uncertainties in the data (prob.density distribution)
- Physical theory of mind: apriori knowledge + adaptivity + ability of computation in the real time

Cognitive models – probabilistic approach

- What makes people smart?
  - Memory?
  - Deduction?
  - Induction and intuition?
- How can we infer so much from so little evidence?

  - Making concepts from examples - few shot/one-shot learning
  - Prior knowledge

Bayesian approach

For any hypothesis *h* and data *d*,

Posterior probability

Likelihood

Prior probability

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')}$$

Sum over space of alternative hypotheses

Probabilistic cognitive models

1. The discovery of structural form (Kemp and Tenenbaum, 2008)
2. Optimal predictions in everyday cognition (Griffiths and Tenenbaum, 2006)
3. Markov Chain Monte Carlo with people (Sanborn and Griffiths, 2008)

Bayesian approach - structure

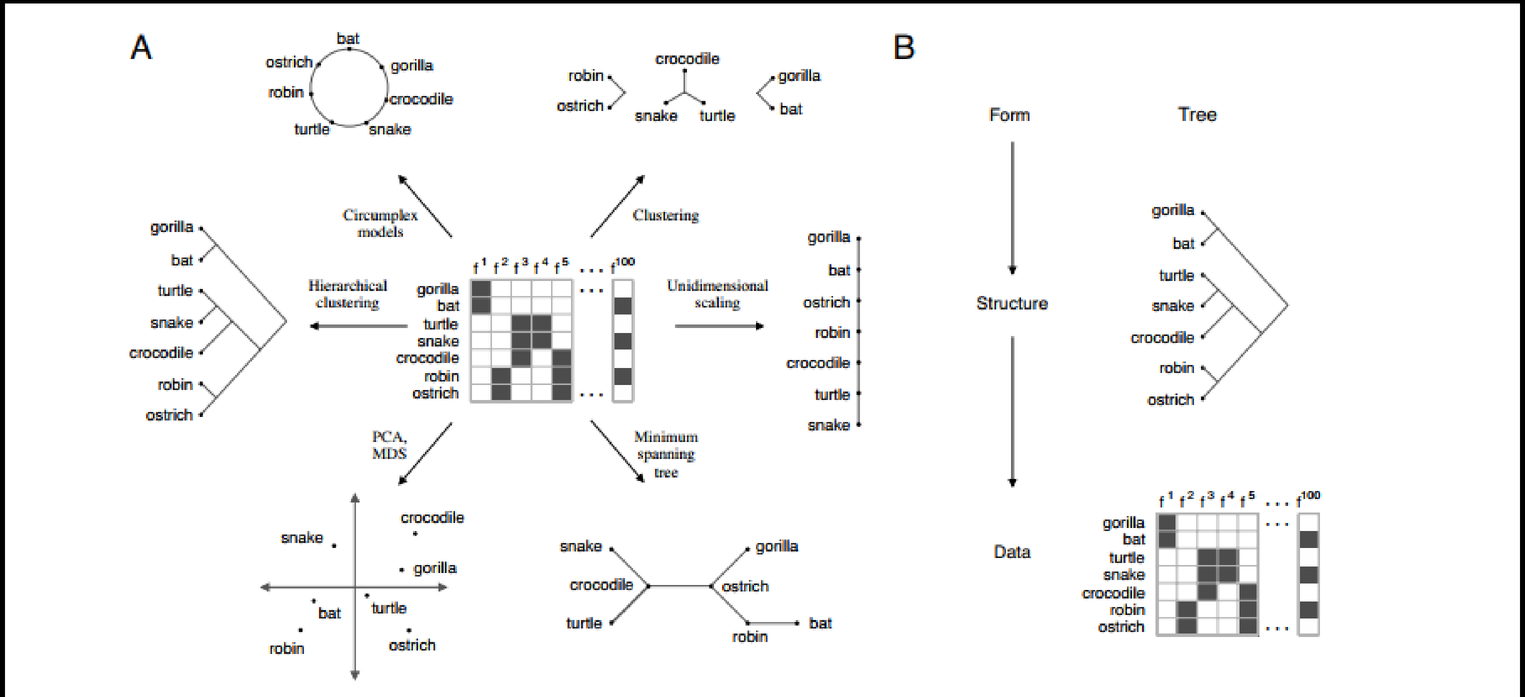
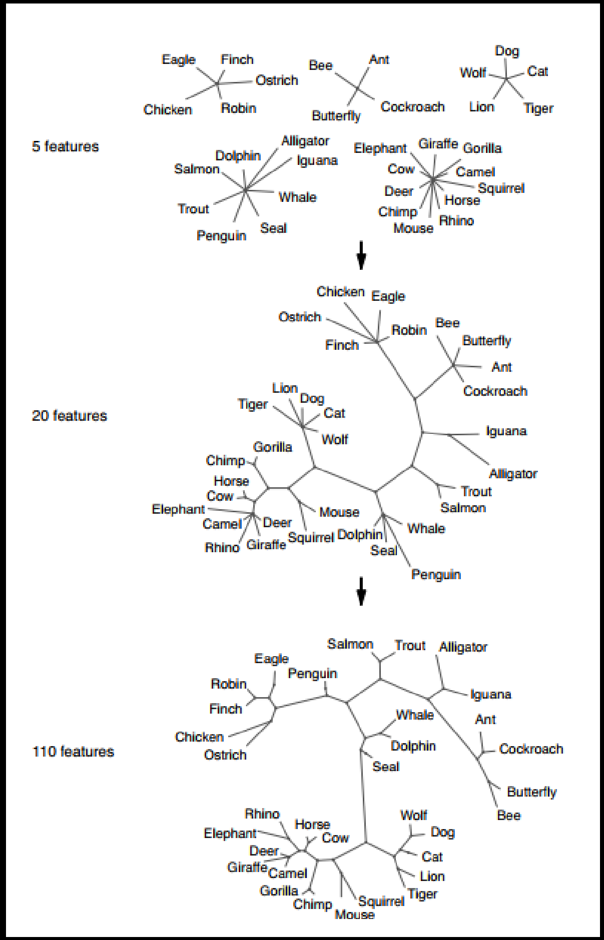


Fig. 1. Finding structure in data. (*A*) Standard structure learning algorithms search for representations of a single form that is fixed in advance. Shown here are methods that discover six different kinds of structures given a matrix of binary features. (*B*) A hierarchical model that discovers the form *F* and the structure *S* that best account for the data *D*. The model searches for the form and structure that jointly maximize $P(S, F|D) \propto P(D|S)P(S|F)P(F)$.



Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105*(31), 10687-10692.

Bayesian approach - structure

- Data D
- Structure S
- Form F

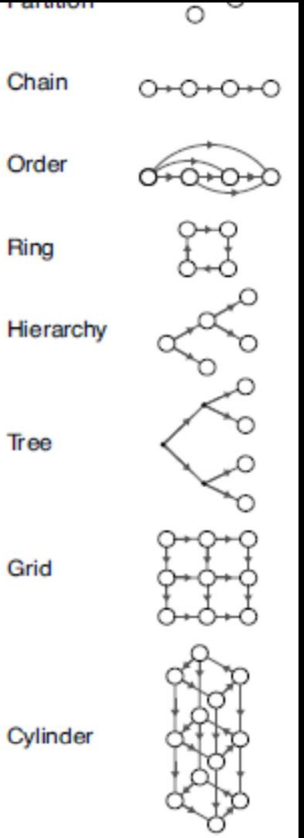
uniform prior

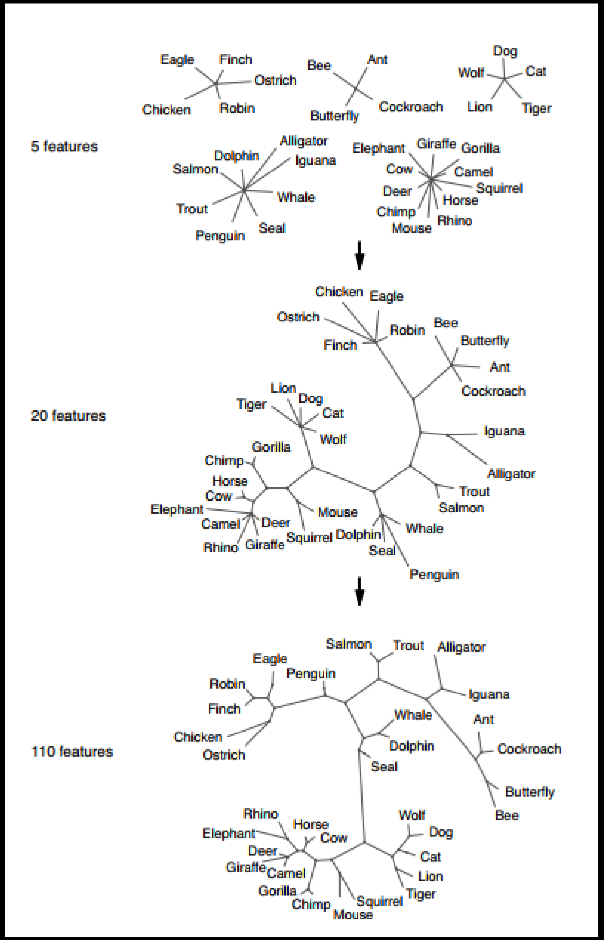
similarity metric

$$p(S,F \mid D) \propto p(F)\,p(S \mid F)\,p(D \mid S)$$

joint posterior over structure and form

higher when S has fewer clusters





Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105*(31), 10687-10692.

Bayesian approach – everyday life

Movie grosses: Imagine you hear about a movie that has taken in 10 million dollars at the box office, but don't know how long it has been running. What would you predict for the total amount of box office intake for that movie?

Poem lengths: If your friend read you her favorite line of poetry, and told you it was line 5 of a poem, what would you predict for the total length of the poem?

Life spans: Insurance agencies employ actuaries to make predictions about people's life spans—the age at which they will die—based upon demographic information. If you were assessing an insurance case for an 18-year-old man, what would you predict for his life span?

Reigns of pharaohs: If you opened a book about the history of ancient Egypt to a page listing the reigns of the pharaohs, and noticed that at 4000 BC a particular pharaoh had been ruling for 11 years, what would you predict for the total duration of his reign?

Lengths of marriages: A friend is telling you about an acquaintance whom you do not know. In passing, he happens to mention that this person has been married for 23 years. How long do you think this person's marriage will last?

Movie run times: If you made a surprise visit to a friend, and found that they had been watching a movie for 30 minutes, what would you predict for the length of the movie?

Terms of U.S. representatives: If you heard a member of the House of Representatives had served for 15 years, what would you predict his total term in the House would be?

Baking times for cakes: Imagine you are in somebody's kitchen and notice that a cake is in the oven. The timer shows that it has been baking for 35 minutes. What would you predict for the total amount of time the cake needs to bake?

Waiting times: If you were calling a telephone box office to book tickets and had been on hold for 3 minutes, what would you predict for the total time you would be on hold?

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science*, *17*(9), 767-773.

Bayesian approach – everyday life

Movie grosses: Imagine you hear about a movie that has taken in 10 million dollars at the box office, but don't know how long it has been running. What would you predict for the total amount of box office intake for that movie?

Poem lengths: If your friend read you her favorite line of poetry, and told you it was line 5 of a poem, what would you predict for the total length of the poem?

Life spans: Insurance agencies employ actuaries to make predictions about people's life spans—the age at which they will die—based upon demographic information. If you were assessing an insurance case for an 18-year-old man, what would you predict for his life span?

Reigns of pharaohs: If you opened a book about the history of ancient Egypt to a page listing the reigns of the pharaohs, and noticed that at 4000 BC a particular pharaoh had been ruling for 11 years, what would you predict for the total duration of his reign?

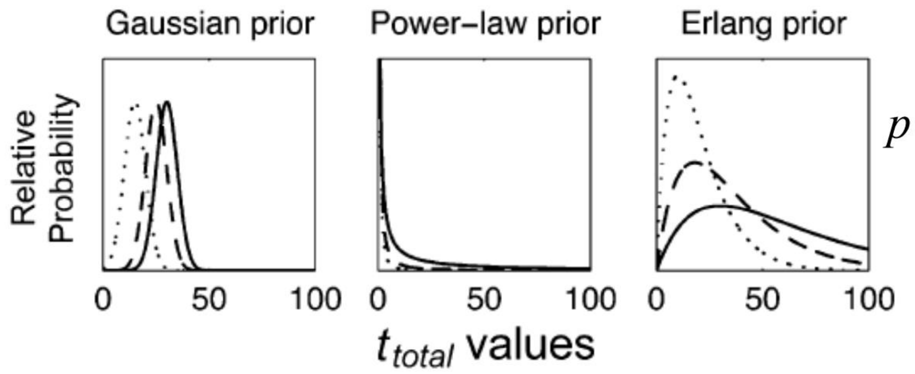
Lengths of marriages: A friend is telling you about an acquaintance whom you do not know. In passing, he happens to mention that this person has been married for 23 years. How long do you think this person's marriage will last?

Movie run times: If you made a surprise visit to a friend, and found that they had been watching a movie for 30 minutes, what would you predict for the length of the movie?

Terms of U.S. representatives: If you heard a member of the House of Representatives had served for 15 years, what would you predict his total term in the House would be?

Baking times for cakes: Imagine you are in somebody's kitchen and notice that a cake is in the oven. The timer shows that it has been baking for 35 minutes. What would you predict for the total amount of time the cake needs to bake?

Waiting times: If you were calling a telephone box office to book tickets and had been on hold for 3 minutes, what would you predict for the total time you would be on hold?



$$p\left(t_{\text{total}}\right) \propto t_{\text{total}}^{k-1} e^{-\frac{t_{\text{total}}}{\beta}}$$

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science*, *17*(9), *767-773.*

Bayesian approach – everyday life

Movie grosses: Imagine you hear about a movie that has taken in 10 million dollars at the box office, but don't know how long it has been running. What would you predict for the total amount of box office intake for that movie?

Poem lengths: If your friend read you her favorite line of poetry, and told you it was line 5 of a poem, what would you predict for the total length of the poem?

Life spans: Insurance agencies employ actuaries to make predictions about people's life spans—the age at which they will die—based upon demographic information. If you were assessing an insurance case for an 18-year-old man, what would you predict for his life span?

Reigns of pharaohs: If you opened a book about the history of ancient Egypt to a page listing the reigns of the pharaohs, and noticed that at 4000 BC a particular pharaoh had been ruling for 11 years, what would you predict for the total duration of his reign?

Lengths of marriages: A friend is telling you about an acquaintance whom you do not know. In passing, he happens to mention that this person has been married for 23 years. How long do you think this person's marriage will last?

Movie run times: If you made a surprise visit to a friend, and found that they had been watching a movie for 30 minutes, what would you predict for the length of the movie?

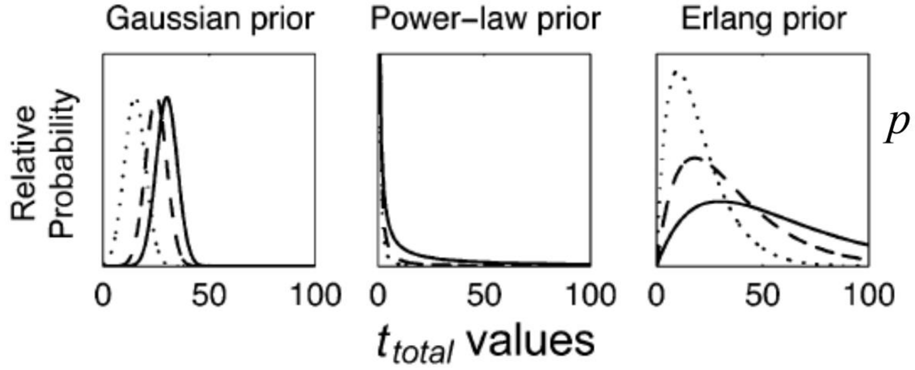
Terms of U.S. representatives: If you heard a member of the House of Representatives had served for 15 years, what would you predict his total term in the House would be?

Baking times for cakes: Imagine you are in somebody's kitchen and notice that a cake is in the oven. The timer shows that it has been baking for 35 minutes. What would you predict for the total amount of time the cake needs to bake?

Waiting times: If you were calling a telephone box office to book tickets and had been on hold for 3 minutes, what would you predict for the total time you would be on hold?

$$p(t_{total} \mid t) \propto p(t \mid t_{total})\, p(t_{total})$$

$$\frac{1}{t_{total}}$$



$$p(t_{total}) \propto t_{total}^{k-1} e^{-\frac{t_{total}}{\beta}}$$

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science*, *17*(9), 767-773.

Bayesian approach – everyday life

Movie grosses: Imagine you hear about a movie that has taken in 10 million dollars at the box office, but don't know how long it has been running. What would you predict for the total amount of box office intake for that movie? Power law $60 million

Poem lengths: If your friend read you her favorite line of poetry, and told you it was line 5 of a poem, what would you predict for the total length of the poem? Power law 17

Life spans: Insurance agencies employ actuaries to make predictions about people's life spans—the age at which they will die—based upon demographic information. If you were assessing an insurance case for an 18-year-old man, what would you predict for his life span? Gaussian 78 years

Reigns of pharaohs: If you opened a book about the history of ancient Egypt to a page listing the reigns of the pharaohs, and noticed that at 4000 BC a particular pharaoh had been ruling for 11 years, what would you predict for the total duration of his reign?

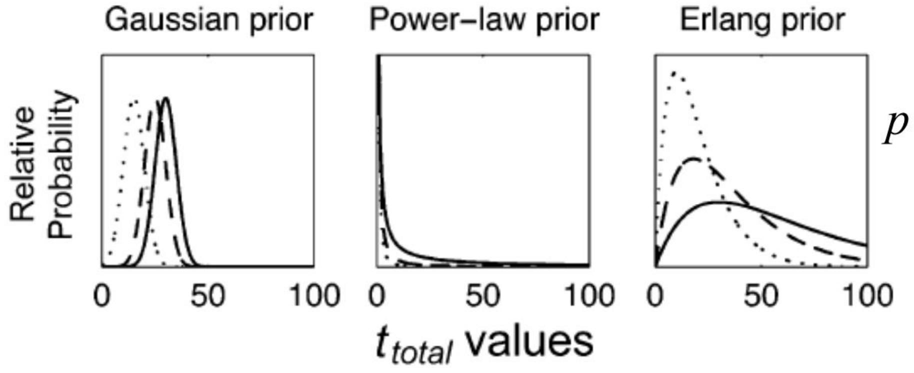
Lengths of marriages: A friend is telling you about an acquaintance whom you do not know. In passing, he happens to mention that this person has been married for 23 years. How long do you think this person's marriage will last?

Movie run times: If you made a surprise visit to a friend, and found that they had been watching a movie for 30 minutes, what would you predict for the length of the movie? Gaussian 55 mins

Terms of U.S. representatives: If you heard a member of the House of Representatives had served for 15 years, what would you predict his total term in the House would be? Erlang 11 years

Baking times for cakes: Imagine you are in somebody's kitchen and notice that a cake is in the oven. The timer shows that it has been baking for 35 minutes. What would you predict for the total amount of time the cake needs to bake?

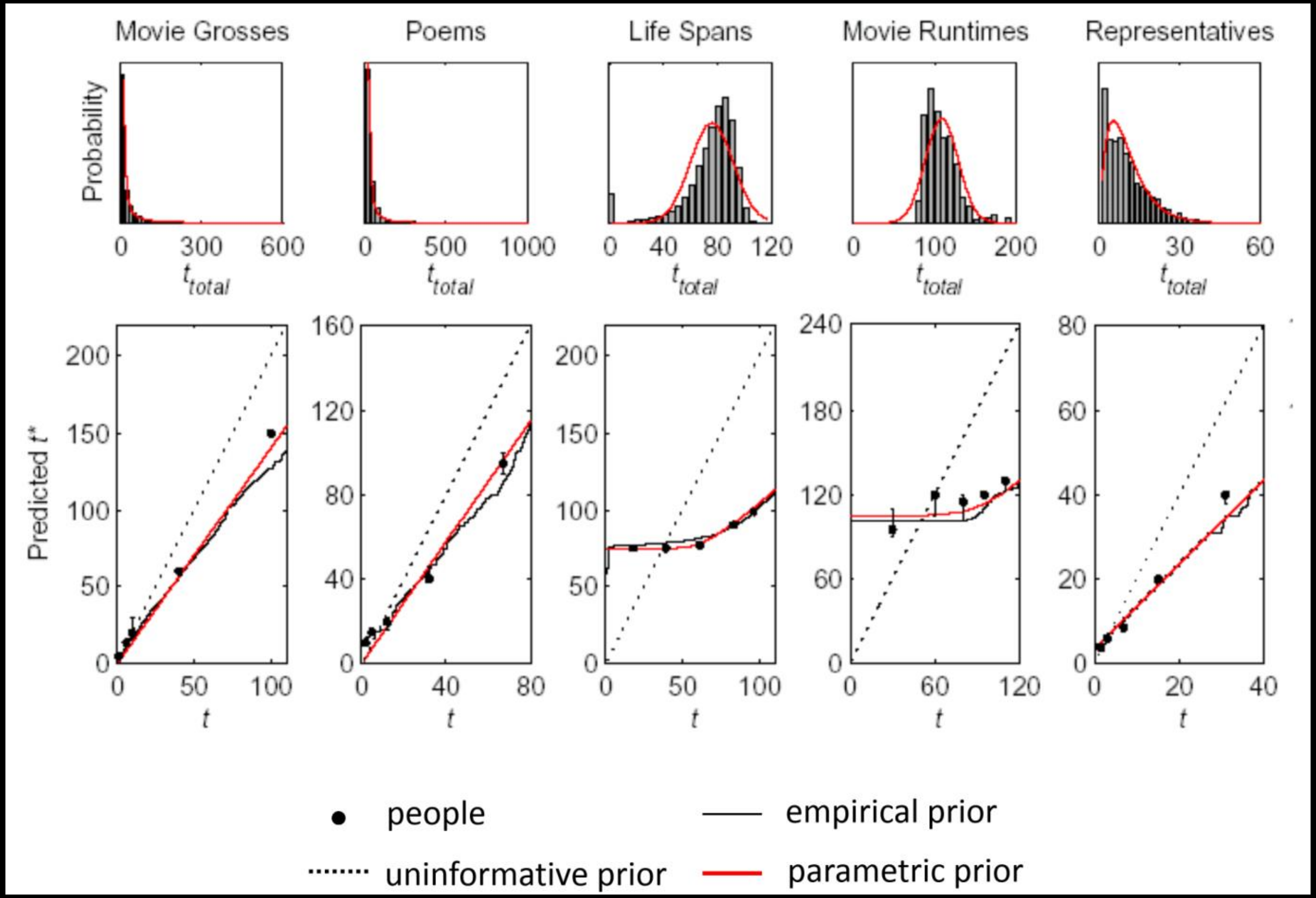
Waiting times: If you were calling a telephone box office to book tickets and had been on hold for 3 minutes, what would you predict for the total time you would be on hold?



$$p\left(t_{\text{total}}\right) \propto t_{\text{total}}^{k-1} e^{-\frac{t_{\text{total}}}{\beta}}$$

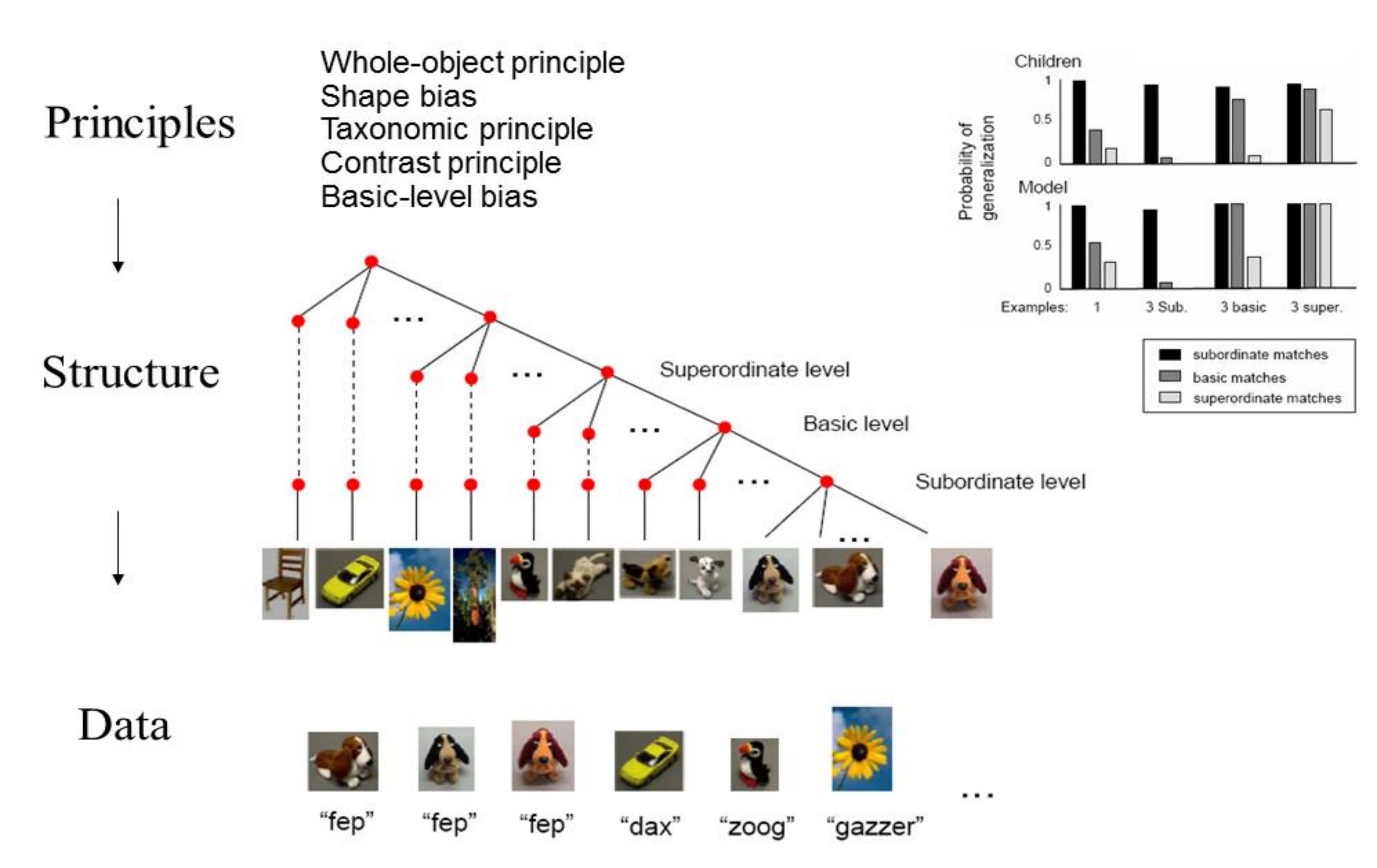
Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science*, *17*(9), 767-773.

Bayesian approach – everyday life



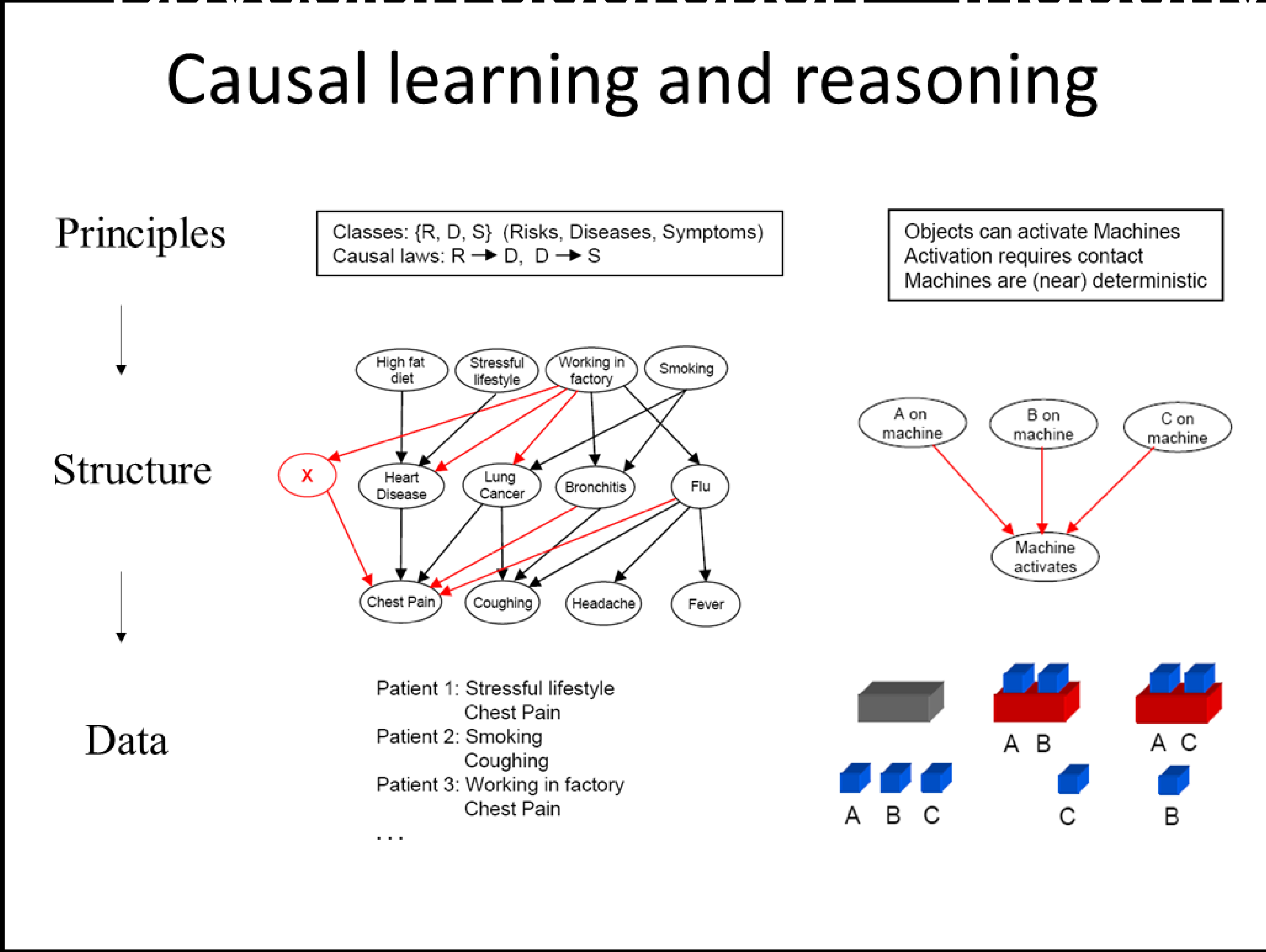
Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science*, *17*(9), 767-773.

Bayesian approach – inductive learning



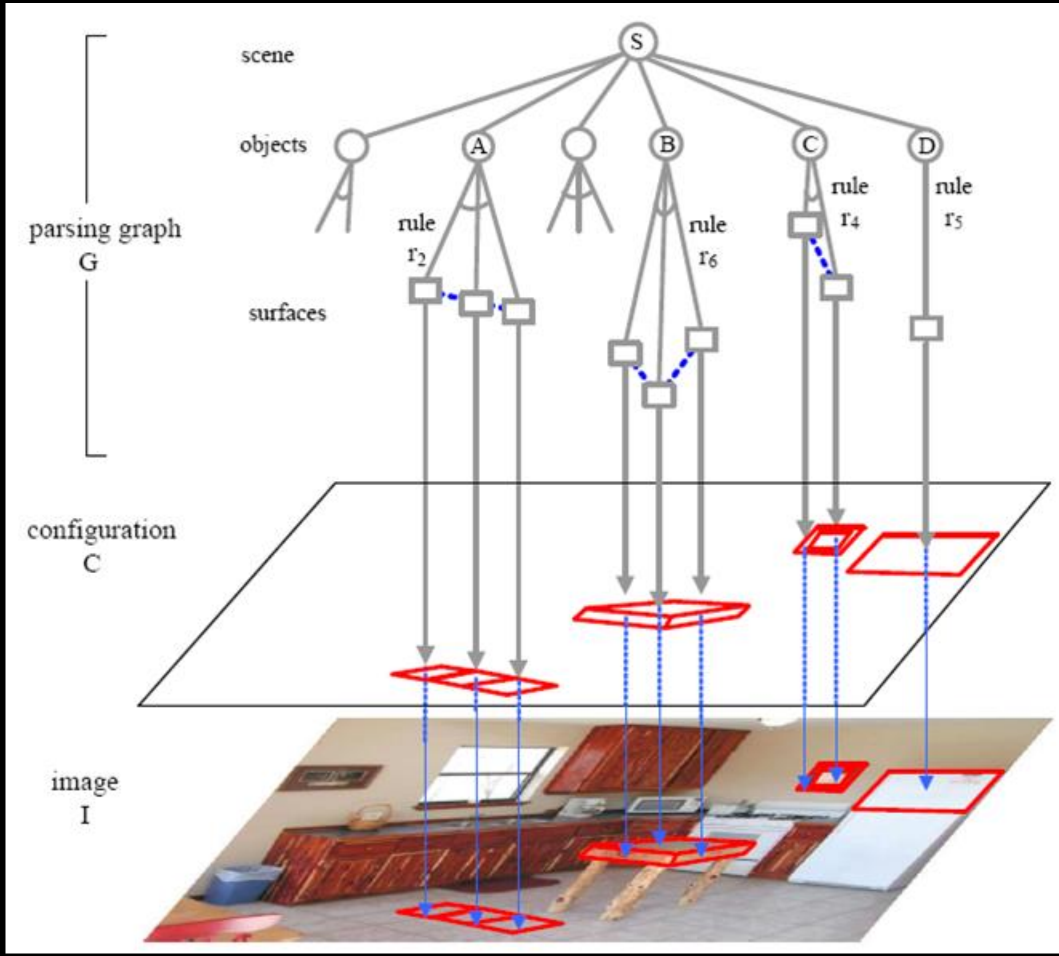
Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309-318.

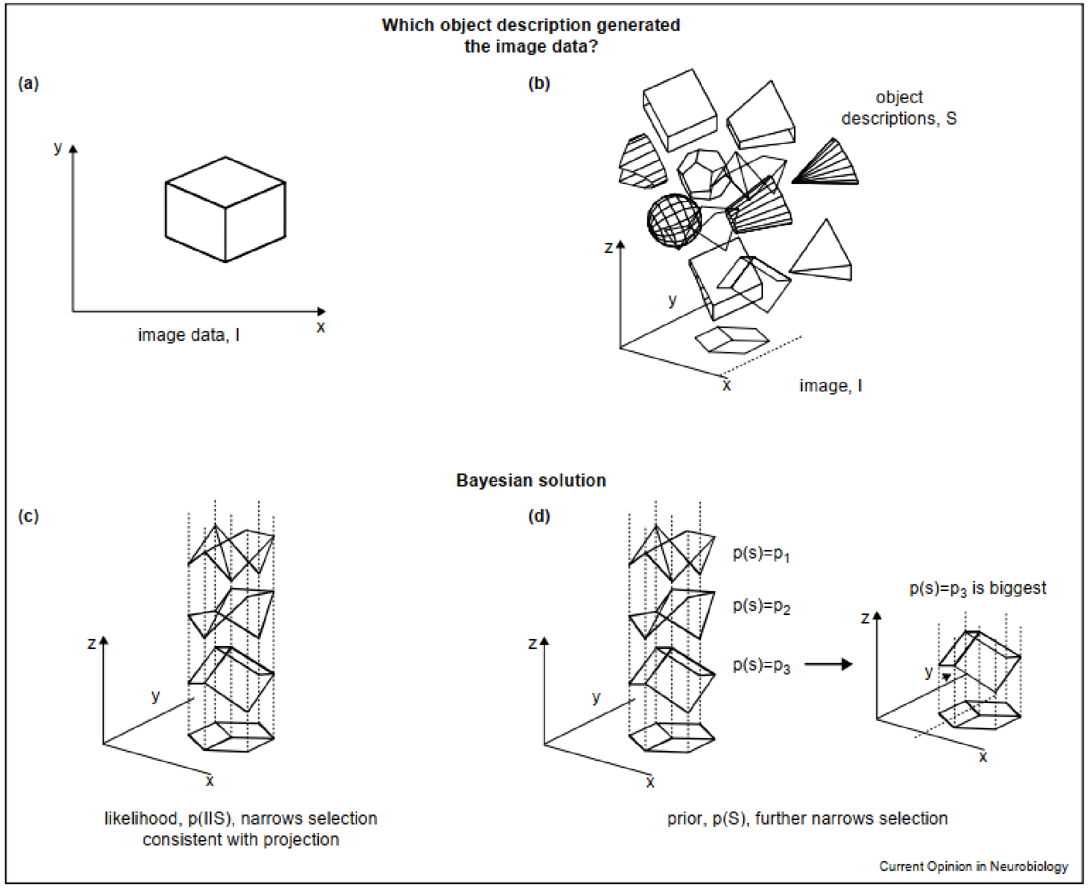
Bayesian approach – inductive learning



Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, *10*(7), 309-318.

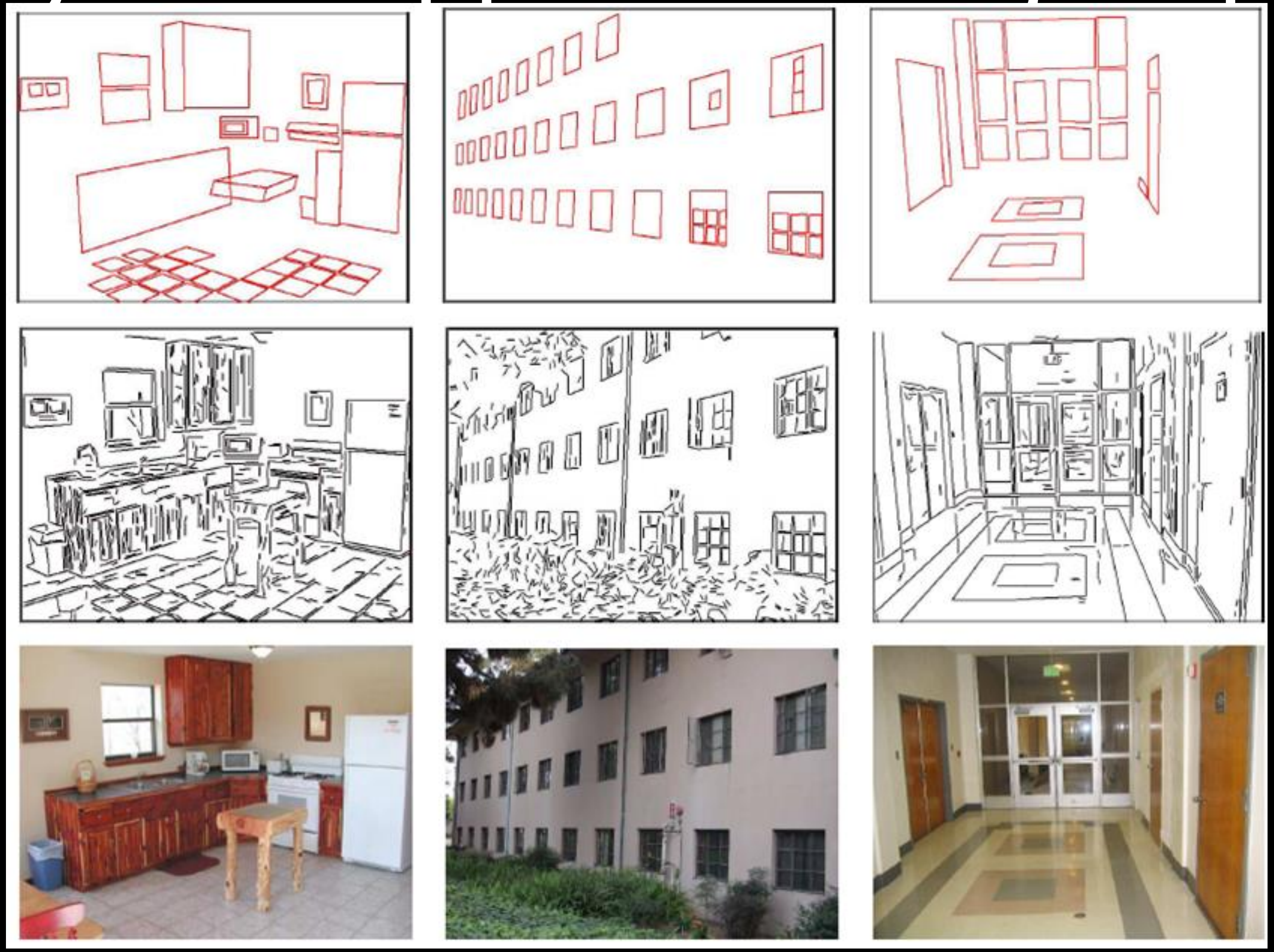
Bayesian approach – object perception



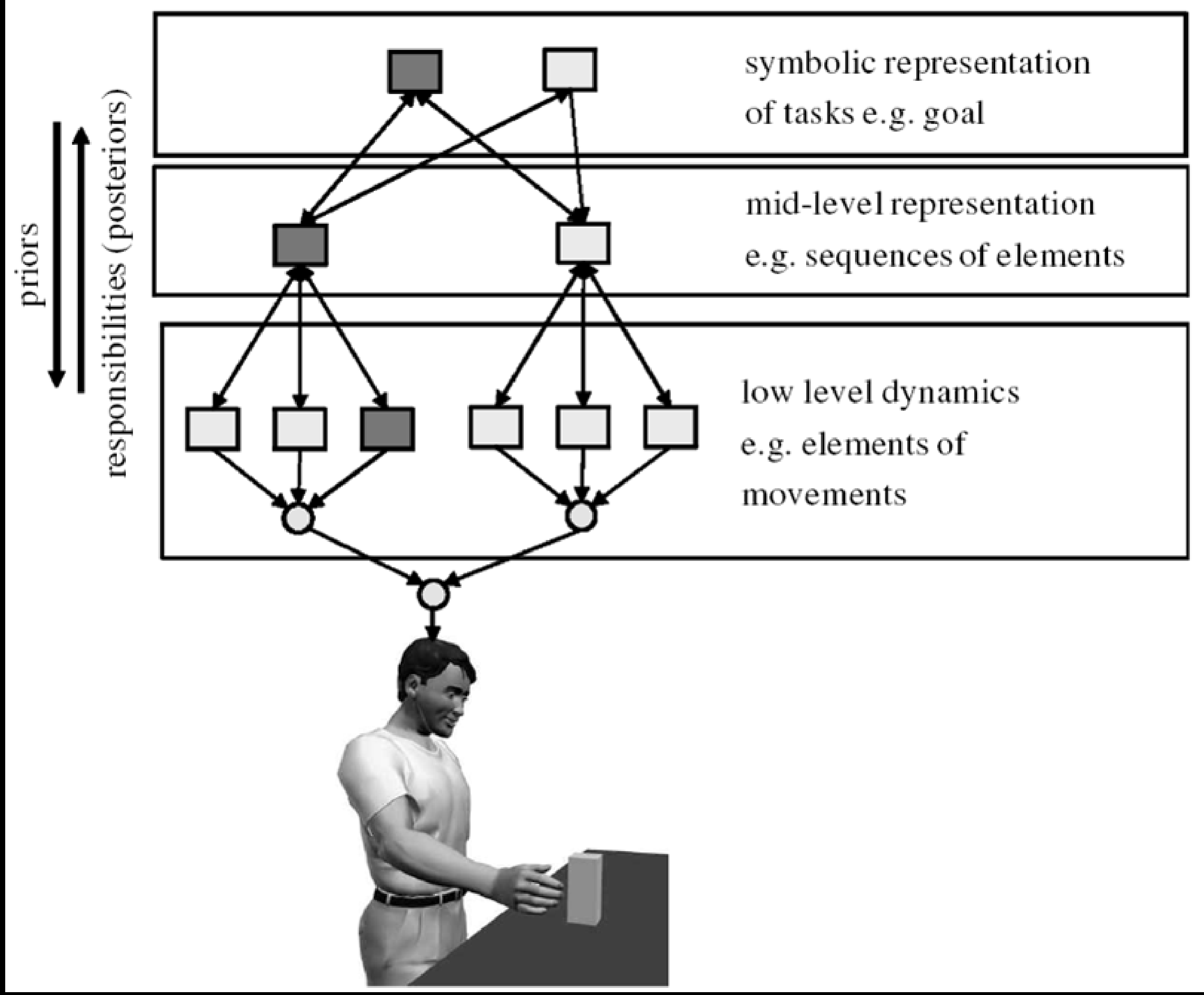


Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current opinion in neurobiology*, *13*(2), 150-158.

Bayesian approach – object perception



Bayesian approach – motorics



Bayesian approach – language

Universal Grammar

↓ P(grammar | UG)

Grammar

↓ P(phrase structure | grammar)

Phrase structure

↓ P(utterance | phrase structure)

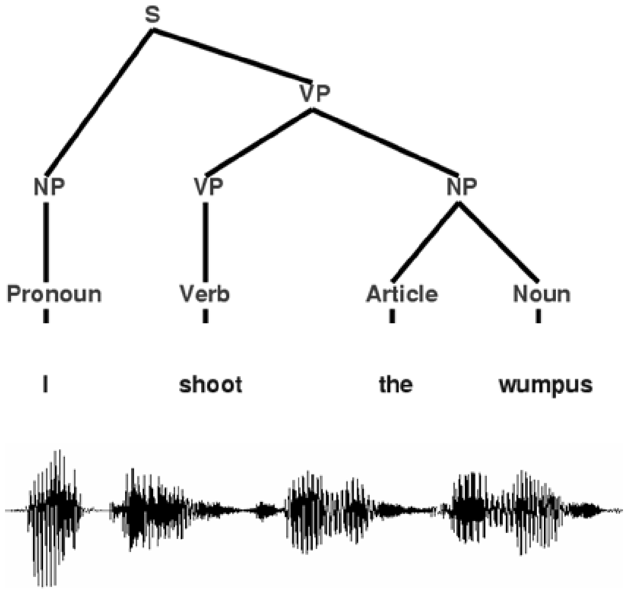
Utterance

↓ P(speech | utterance)

Speech signal

Hierarchical phrase structure grammars (e.g., CFG, HPSG, TAG)

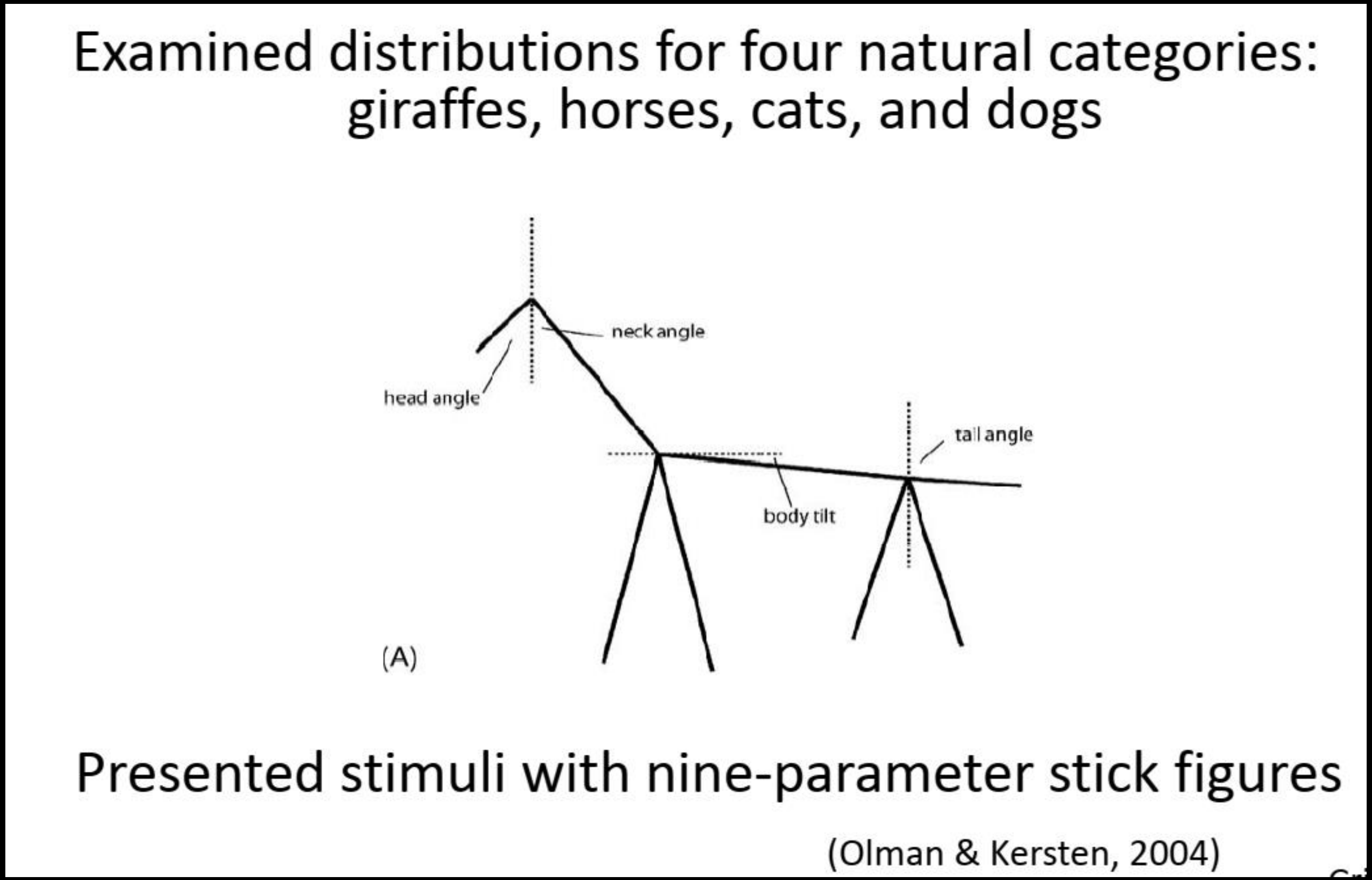
$$S \rightarrow NP\ VP$$
$$NP \rightarrow Det\,[Adj]\,Noun\,[RelClause]$$
$$RelClause \rightarrow [Rel]\ NP\ V$$
$$VP \rightarrow VP\ NP$$
$$VP \rightarrow Verb$$



Cognitive models

- People learn by modifying their belifefs about hypotheses
- How do people learn probability distributions?
- Markov Chain Monte Carlo: Markov chain that has the target distribution as stationary distribution
- Initialize with any state, guaranteed to converge after many iterations

Cognitive models



Sanborn, A., & Griffiths, T. L. (2008). Markov chain Monte Carlo with people. In *Advances in neural information processing systems* (pp. 1265-1272).

Cognitive models

A Bayesian analysis of the task

$$h_1 : x_1 \text{ is from } p(x|c); x_2 \text{ is from } g(x)$$

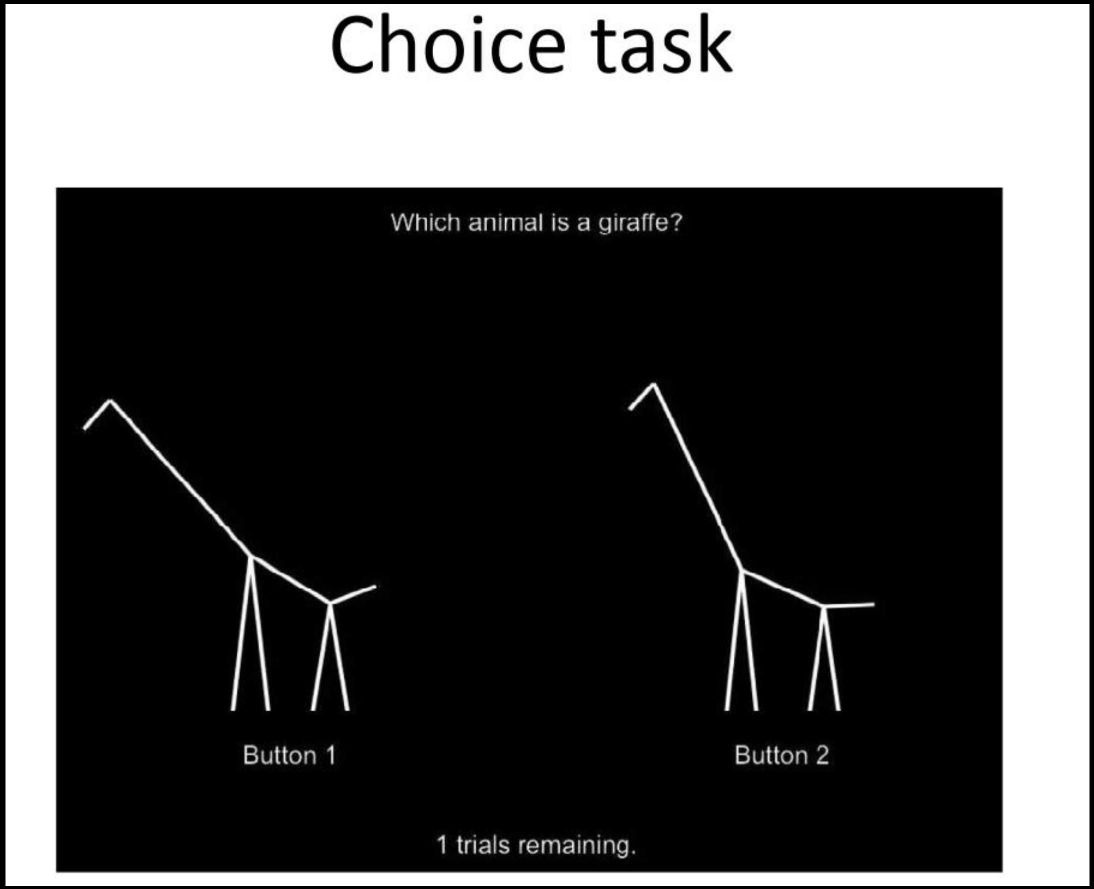
$$h_2 : x_2 \text{ is from } p(x|c); x_1 \text{ is from } g(x)$$

$$p(h_1|x_1, x_2) = \frac{p(x_1|c)g(x_2)p(h_1)}{p(x_1|c)g(x_2)p(h_1) + p(x_2|c)g(x_1)p(h_2)}$$

Assume: $p(h_1) = p(h_2)$
$g(x_1) = g(x_2)$

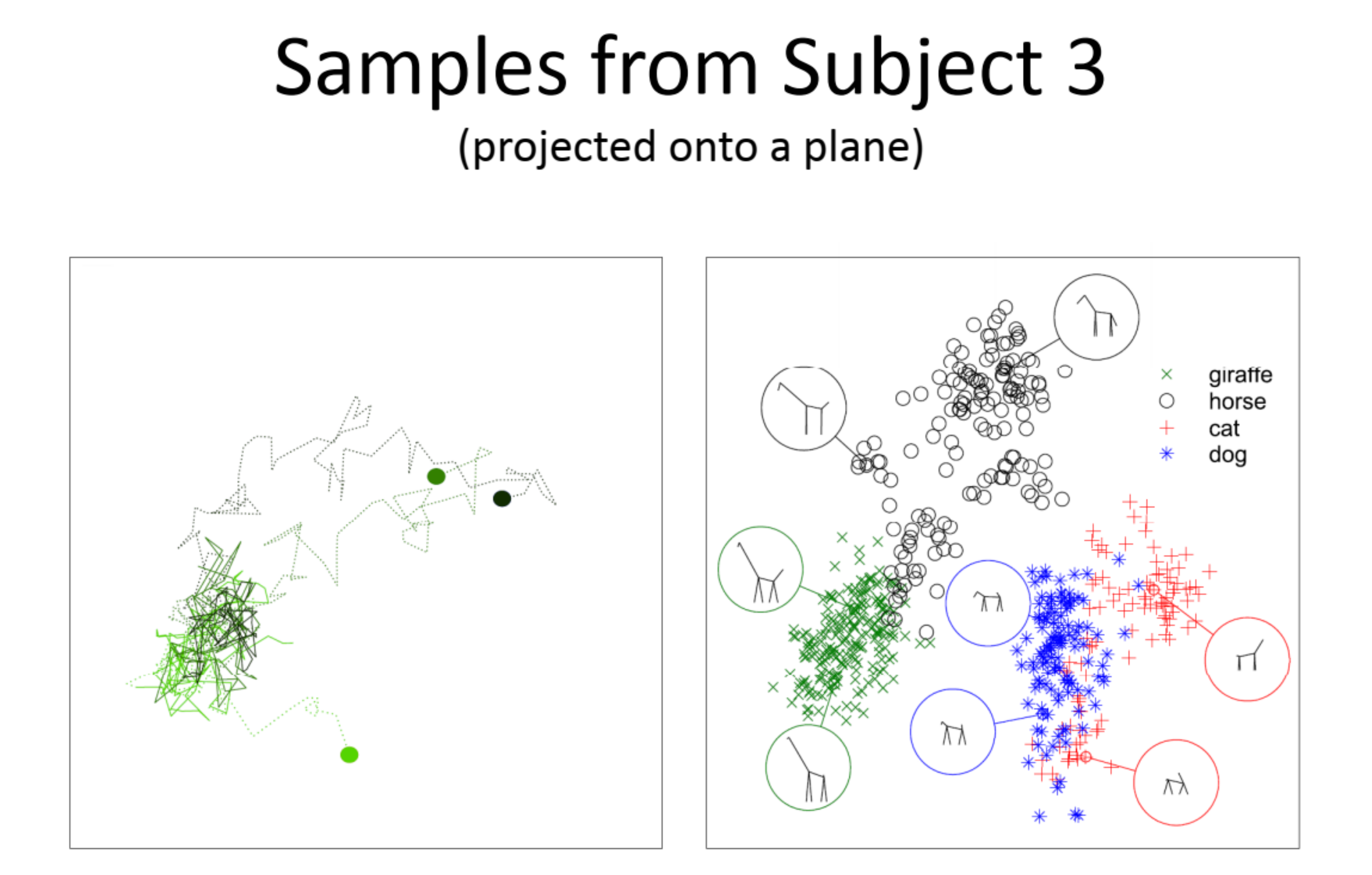
Sanborn, A., & Griffiths, T. L. (2008). Markov chain Monte Carlo with people. In *Advances in neural information processing systems* (pp. 1265-1272).

Cognitive models



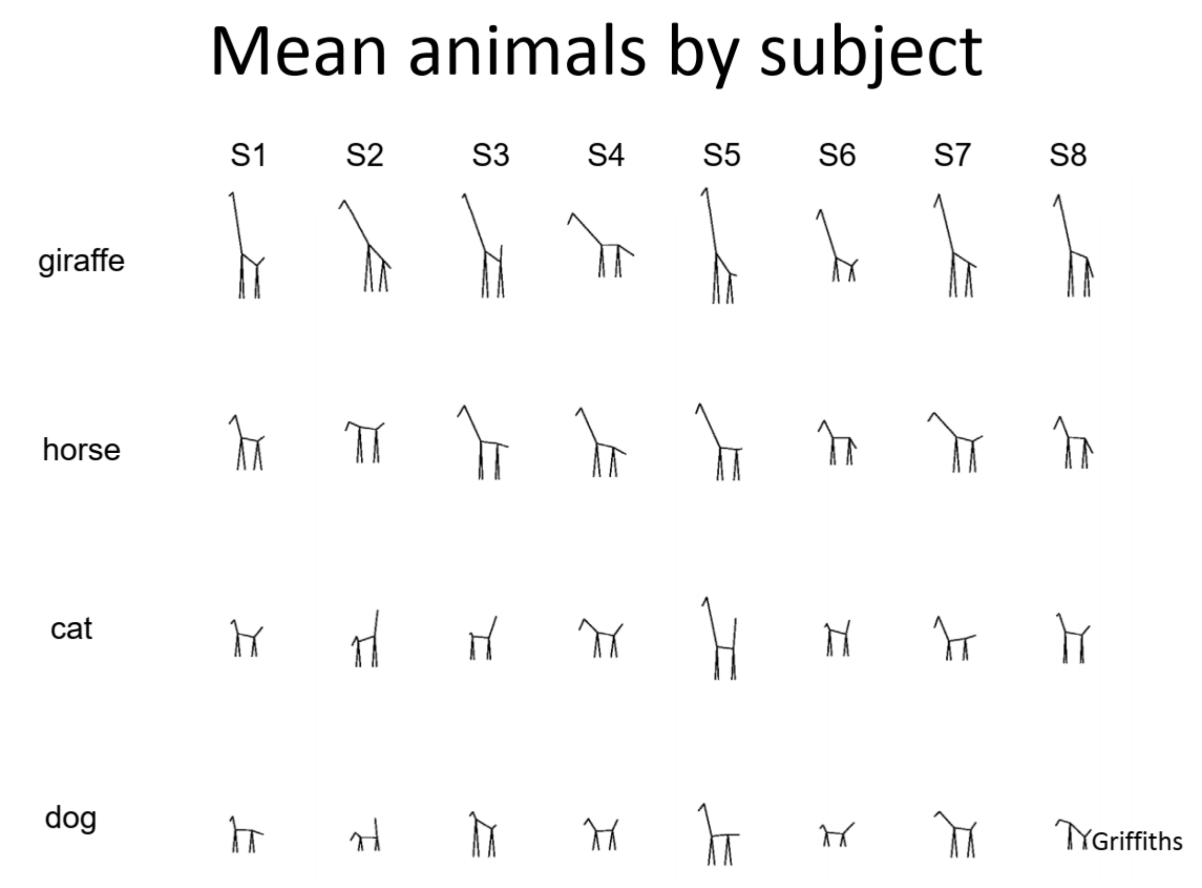
Sanborn, A., & Griffiths, T. L. (2008). Markov chain Monte Carlo with people. In *Advances in neural information processing systems* (pp. 1265-1272).

Cognitive models



Sanborn, A., & Griffiths, T. L. (2008). Markov chain Monte Carlo with people. In *Advances in neural information processing systems* (pp. 1265-1272).

Cognitive models



Sanborn, A., & Griffiths, T. L. (2008). Markov chain Monte Carlo with people. In *Advances in neural information processing systems* (pp. 1265-1272).

Cognitive models

Metropolis-Hastings algorithm

(Metropolis et al., 1953; Hastings, 1970)

Step 1: propose a state (we assume symmetrically)

$$Q(x^{(t+1)}|x^{(t)}) = Q(x^{(t)})|x^{(t+1)})$$

Step 2: decide whether to accept, with probability

$$A(x^{(t+1)}, x^{(t)}) = \min\left(1, \frac{p(x^{(t+1)})}{p(x^{(t)})}\right)$$

Metropolis acceptance function

$$A(x^{(t+1)}, x^{(t)}) = \frac{p(x^{(t+1)})}{p(x^{(t+1)}) + p(x^{(t)})}$$

Barker acceptance function

Sanborn, A., & Griffiths, T. L. (2008). Markov chain Monte Carlo with people. In *Advances in neural information processing systems* (pp. 1265-1272).

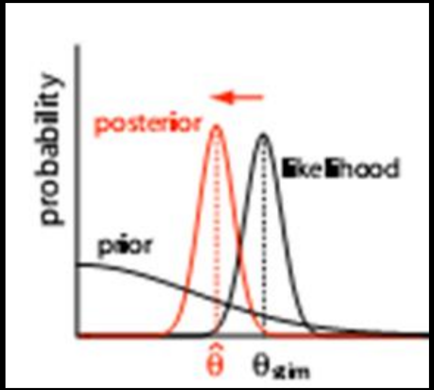
Cognitive models

- Probabilistic models can guide the design of experiments to measure psychological variables
- Markov Chain Monte Carlo can be used to sample from subjective probability distributions
  - Category distributions (Metropolis-Hastings)
  - Prior distributions (Gibbs sampling)
- Effective for exploring large stimulus spaces, with distributions on a small part of the space

Sanborn, A., & Griffiths, T. L. (2008). Markov chain Monte Carlo with people. In *Advances in neural information processing systems* (pp. 1265-1272).

Priors and posteriors

- Prior knowledge about the world can be used to interpret data in situation of uncertainty.
- Prediction: the more uncertain the data, the more the prior should influence the interpretation.
- The priors should reflect the statistics of the sensory world



Coin flipping example

Comparing two simple hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D:	HHTHT		
H_1, H_2:	"fair coin", "always heads"		
$P(D\|H_1)$ =	$1/2^5$	$P(H_1)$ =	999/1000
$P(D\|H_2)$ =	0	$P(H_2)$ =	1/1000

$$P(H_1|D) \,/\, P(H_2|D) = \text{infinity}$$

Coin flipping example

Comparing two simple hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

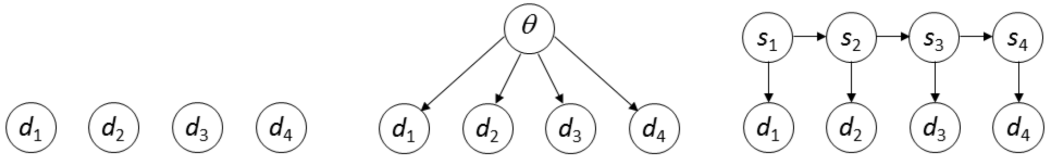
D:	HHHHH		
H_1, H_2:	"fair coin", "always heads"		
$P(D\|H_1) =$	$1/2^5$	$P(H_1) =$	999/1000
$P(D\|H_2) =$	1	$P(H_2) =$	1/1000

$$P(H_1|D) \,/\, P(H_2|D) \approx 30$$

Coin flipping example

Model selection

- Assume hypothesis space of possible models:



Fair coin: $P(\mathrm{H}) = 0.5$

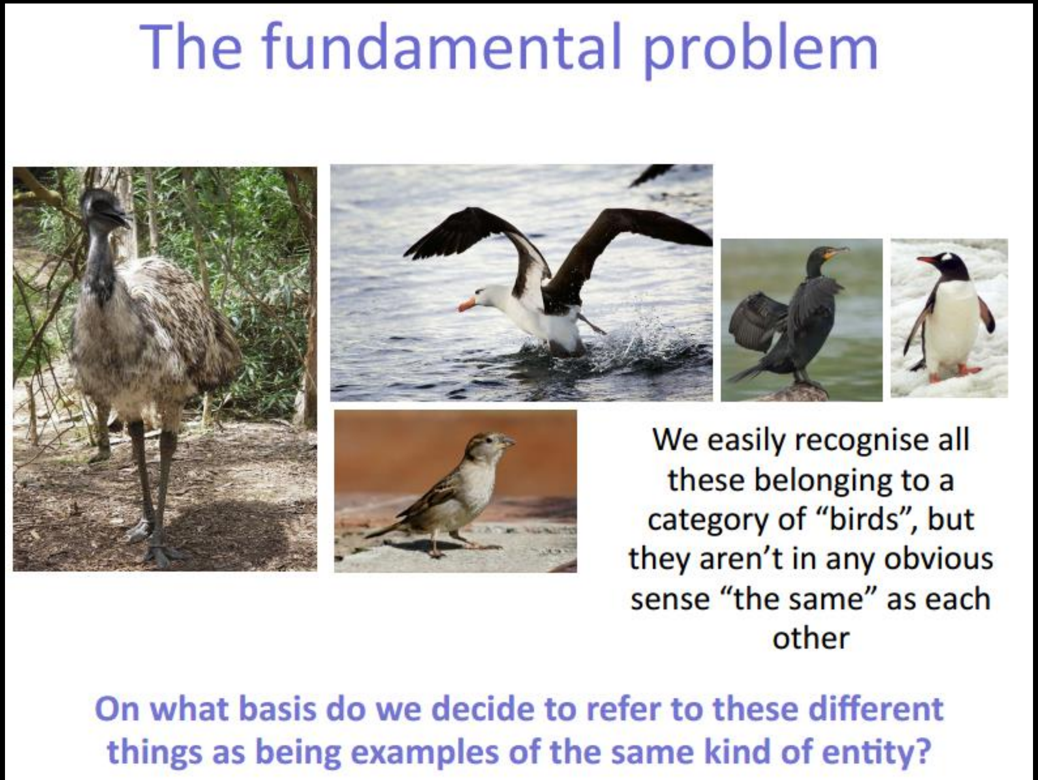
$P(\mathrm{H}) = \theta$

Hidden Markov model:
$s_i \in$ {Fair coin, Trick coin}

- Which model generated the data?
  – requires summing out hidden variables
  – requires some form of Occam's razor to trade off complexity with fit to the data.

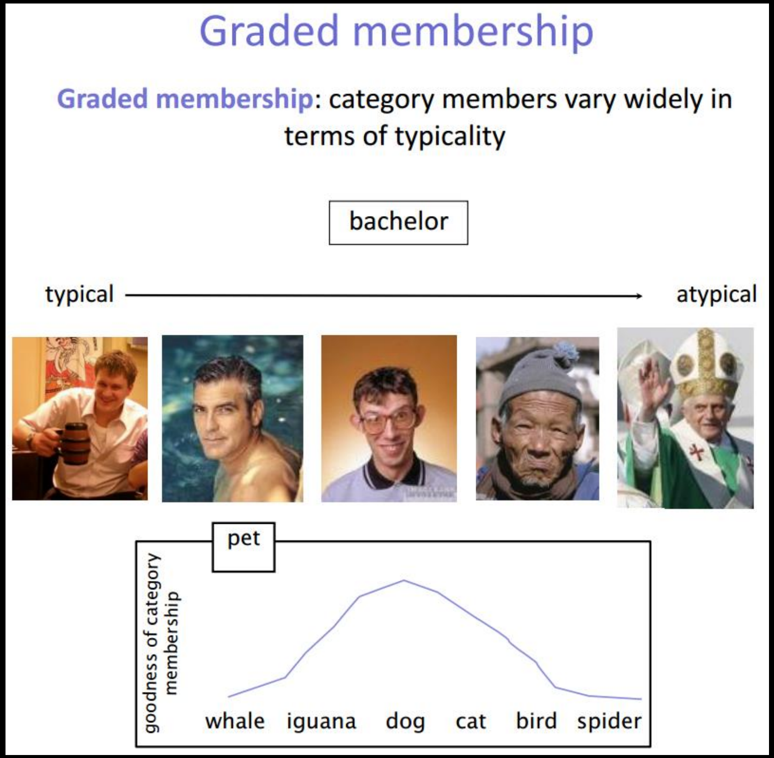
Clustering - Gaussian distribution

- Classical categorization – necessity and suficiency

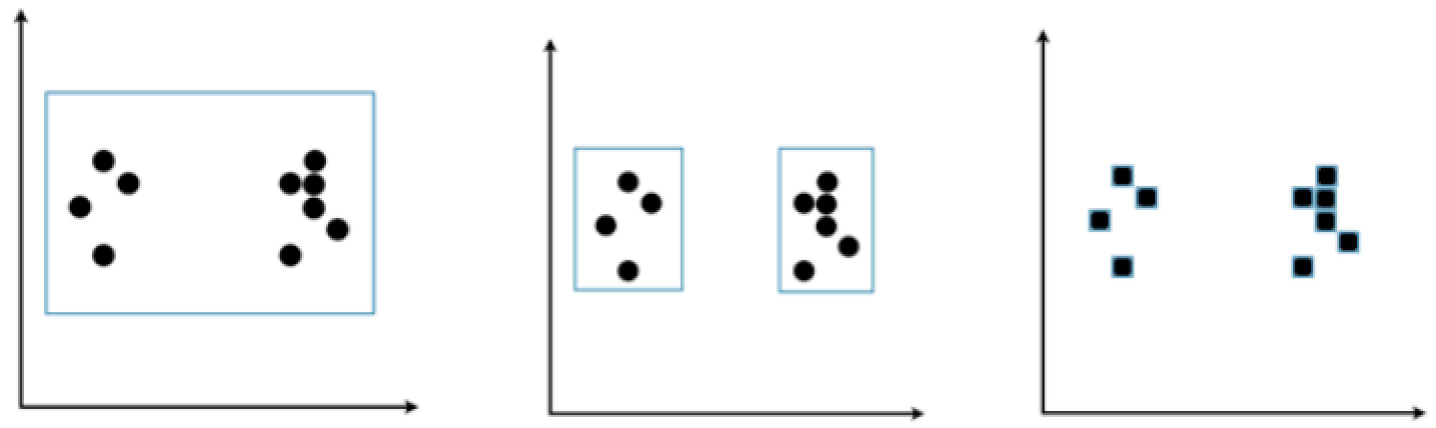


Clustering - Gaussian distribution

- Classical categorization – necessity and sufficiency
- Graded membership - likelihood



Gaussian distribution



How do we evaluate between these hypotheses?

Gaussian distribution

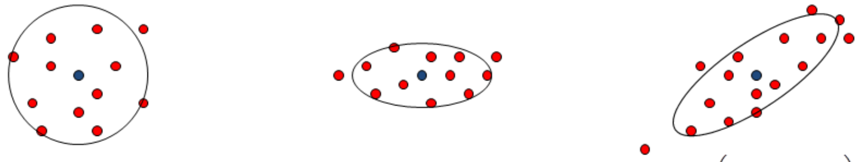
Multivariate Gaussians

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(x - \boxed{\mu})^2 / 2\boxed{\sigma^2}\}$$

mean variance/covariance matrix

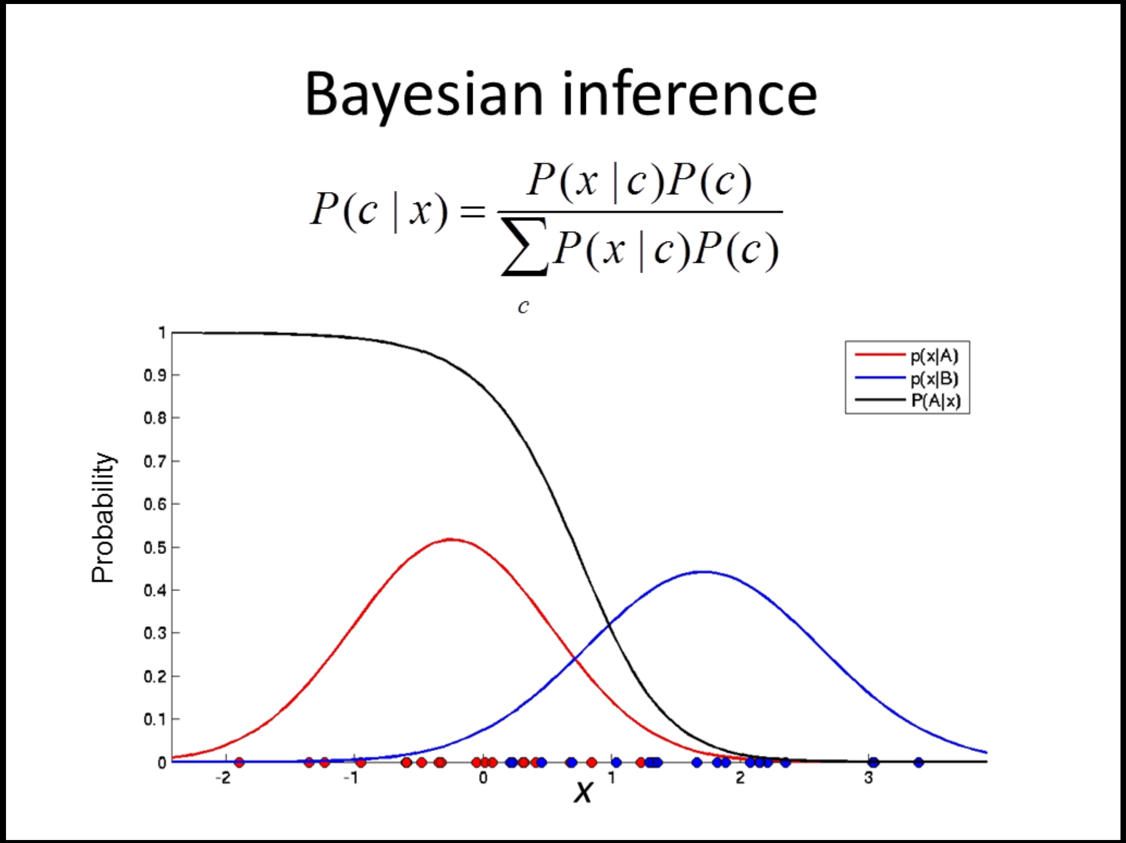
$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\{-(x - \boxed{\mu})^T \boxed{\Sigma}^{-1} (x - \mu)/2\}$$

quadratic form

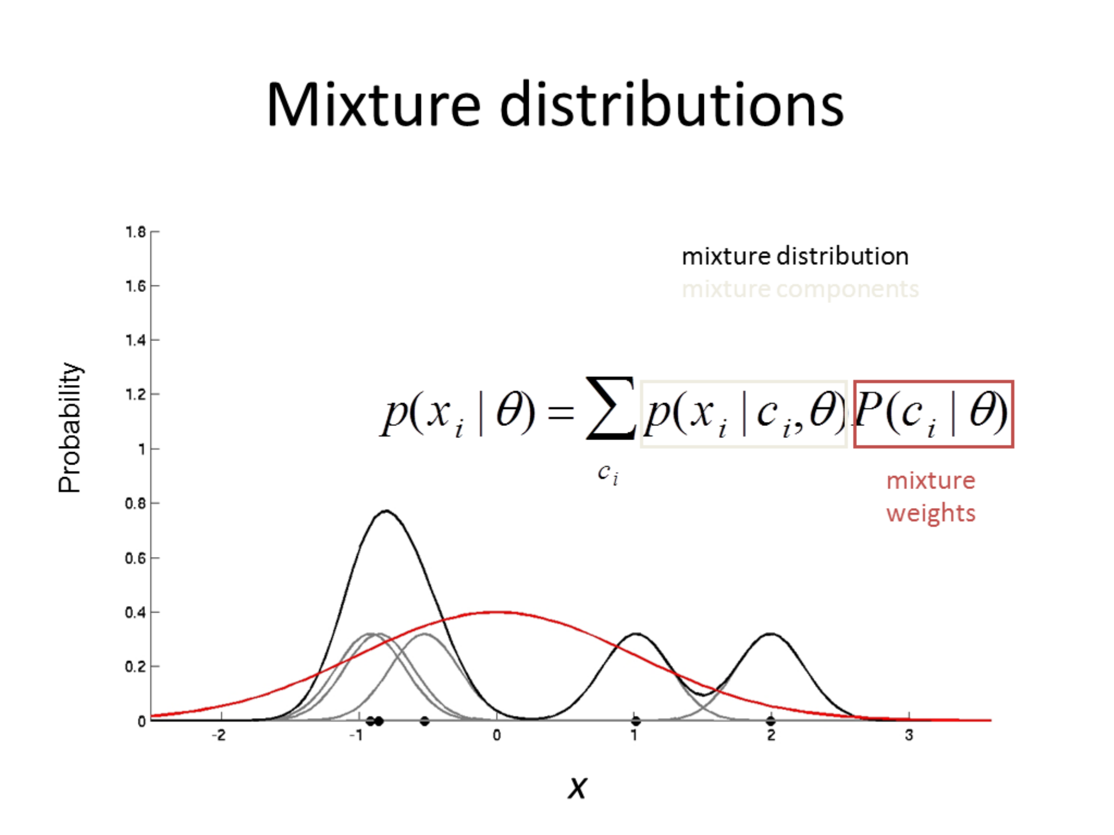


$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 0.25 \end{pmatrix}$ $\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$

Gaussian distribution



Gaussian distribution



Gaussian distribution

1, E-step: estimation of all probabilities $f_k(\mathbf{x}_i)$:

$$f_k(\mathbf{x}_i) = \frac{r_k l_k(\mathbf{x}_i|\mathbf{m}_k, \mathbf{S}_k)}{\sum_{k'=1}^{K} r_{k'} l(\mathbf{x}_i|\mathbf{\Theta}_{k'})}$$

2, M-step: choose the parameters which maximizes log-likelihood when the probabilities $f_k(\mathbf{xi})$ are known:

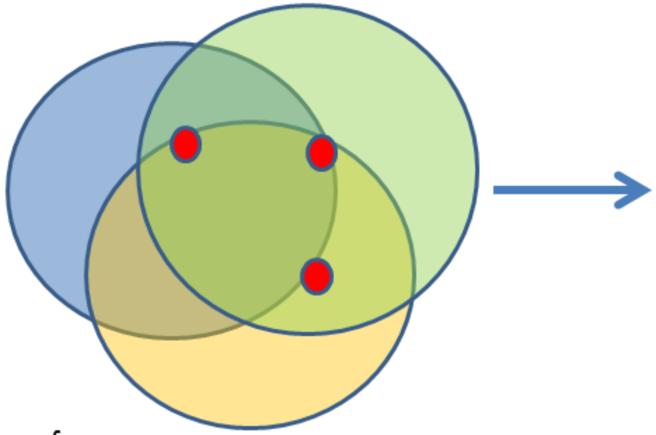
$$r_k = \frac{1}{N}\sum_{i=1}^{N} f_k(\mathbf{x}_i)$$

$$\mathbf{m}_k = \frac{\sum_{i=1}^{N} f_k(\mathbf{x}_i)\mathbf{x}_i}{\sum_{j=1}^{N} f_k(\mathbf{x}_j)}$$

$$\mathbf{S}_k = \frac{\sum_{i=1}^{N} f_k(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T}{\sum_{j=1}^{N} f_k(\mathbf{x}_j)}$$

Clustering - Gaussian distribution

- Dynamic creation of the relationships between internal representations and the world

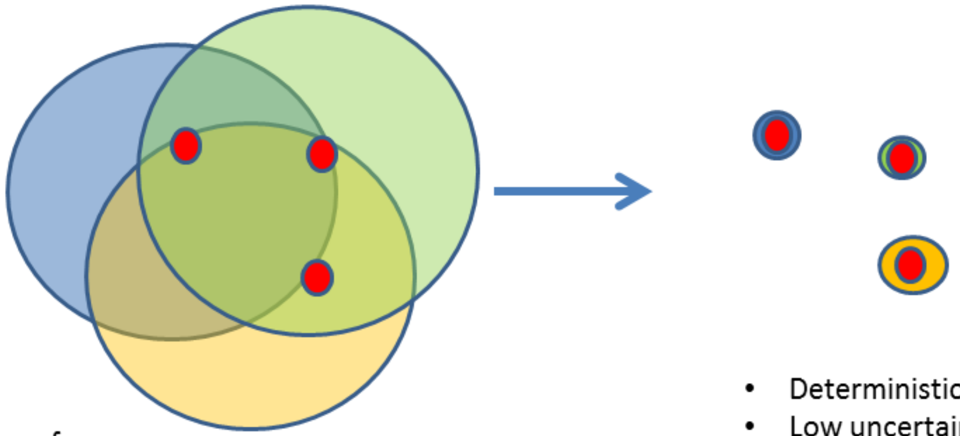


- Fuzzy forms
- Class membership with high fuzziness
- A priori models with very uncertain parameters

- Deterministic concepts
- Low uncertainity about class membership
- Models with fixed parameter values

Clustering - Gaussian distribution

- Dynamic creation of the relationships between internal representations and the world



- Fuzzy forms
- Class membership with high fuzziness
- A priori models with very uncertain parameters

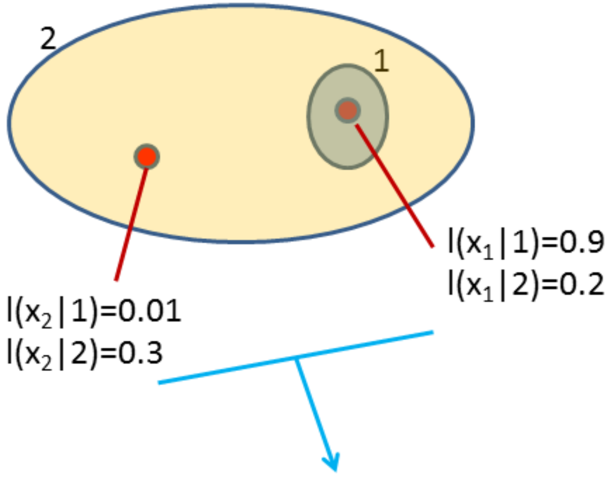
- Deterministic concepts
- Low uncertainity about class membership
- Models with fixed parameter values

- Heterohierarchical structure–many interative loops which include different levels of processing
- In each moment, many concepts (agents, objects) compete for their evidence

Clustering - Gaussian distribution

- Asociation(segmentation) Θ array of input data x with objects= division of inputs to subsets which are related to the given objects

l(n|k) – partial similarity of the point ***n*** with model ***k***
f(k|n) – membership of point ***n*** to model ***k***



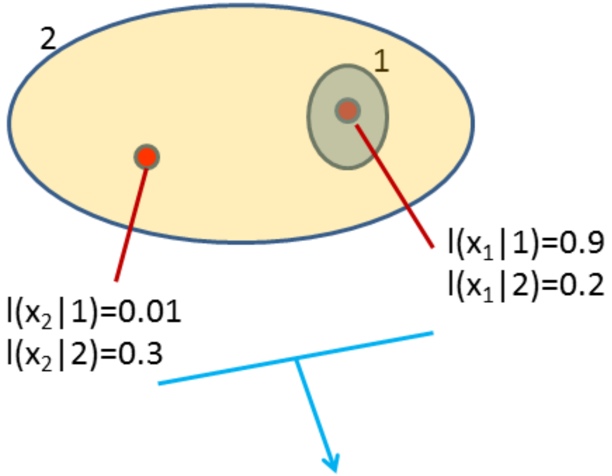
$f(1|x_1)=0.9/(0.9+0.01)=1$
$f(2|x_1)=0.3/(0.3+0.2)=0.6$

Clustering - Gaussian distribution

- Asociation(segmentation) Θ array of input data x with objects= division of inputs to subsets which are related to the given objects

l(n|k) – partial similarity of the point ***n*** with model ***k***

f(k|n) – membership of point ***n*** to model ***k***



$f(1|x_1)=0.9/(0.9+0.01)=1$

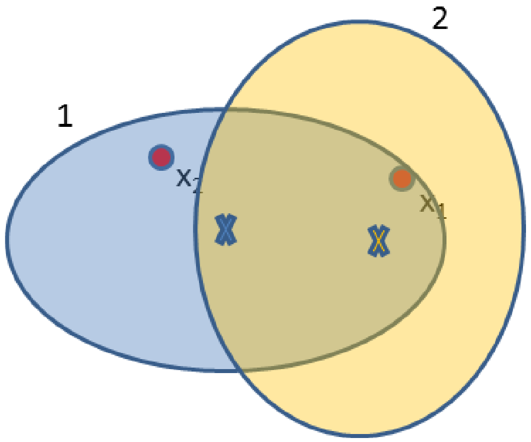
$f(2|x_1)=0.3/(0.3+0.2)=0.6$

Maximalization of complete conditional log-fuzzy similarity:

$$AZ\text{-}LL=\max_{Sk}\Sigma_n \ln\,[\Sigma_k\, f(k|n)]$$

Clustering - Gaussian distribution

1. Initialization of parameters (a priori knowledge)



Clustering - Gaussian distribution

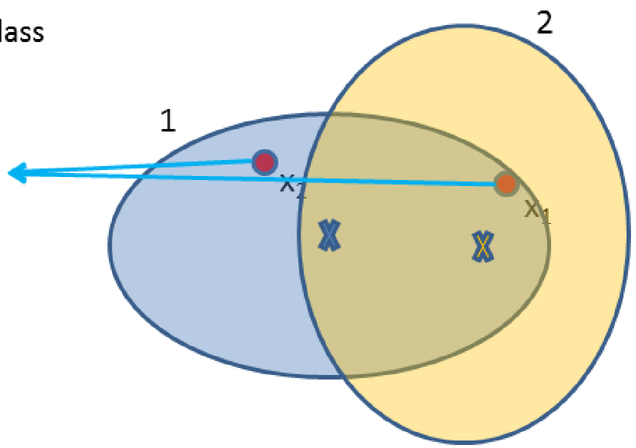
1. Initialization of parameters (a priori knowledge)

2. E – step: compute similarities l(n|k) and class memberships f(k|n)

$l(x_2|1)=0.1$
$l(x_2|2)=0.2$
$l(x_1|1)=0.3$
$l(x_1|2)=0.3$

$f(1|x_1)$, $f(1|x_2)$,
$f(2|x_1)$,$f(2|x_2)$

$$l_j(\vec{x}_i|\vec{m}_j,\vec{S}_j) = (2\pi)^{-d/2}\vec{S}_j^{-1/2}\exp[-0.5(\vec{x}_i-\vec{m}_j)^T \vec{S}_j^{-1}(\vec{x}_i-\vec{m}_j)] \quad (5)$$



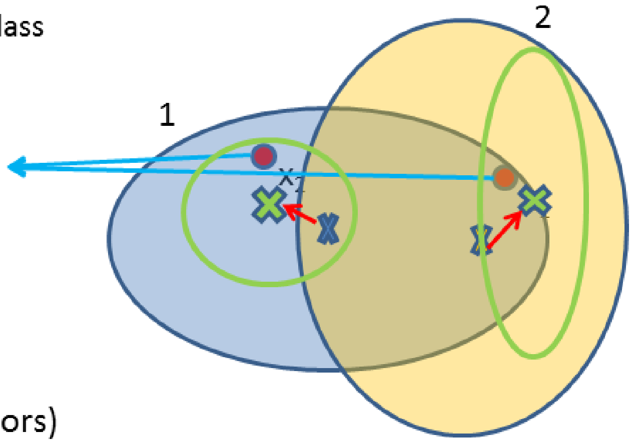
Clustering - Gaussian distribution

1. Initialization of parameters (a priori knowledge)

2. E – step: compute similarities $l(n|k)$ and class memberships $f(k|n)$

$$\left.\begin{aligned} l(x_2|1)&=0.1\\ l(x_2|2)&=0.2\\ l(x_1|1)&=0.3\\ l(x_1|2)&=0.3 \end{aligned}\right\} \quad \begin{aligned} &f(1|x_1), f(1|x_2),\\ &f(2|x_1), f(2|x_2) \end{aligned}$$

$$l_j(\vec{x}_i|\vec{m}_j, \vec{S}_j) = (2\pi)^{-d/2}\vec{S}_j^{\,-1/2}\exp[-0.5(\vec{x}_i-\vec{m}_j)^T \vec{S}_j^{\,-1}(\vec{x}_i-\vec{m}_j)] \quad (5)$$



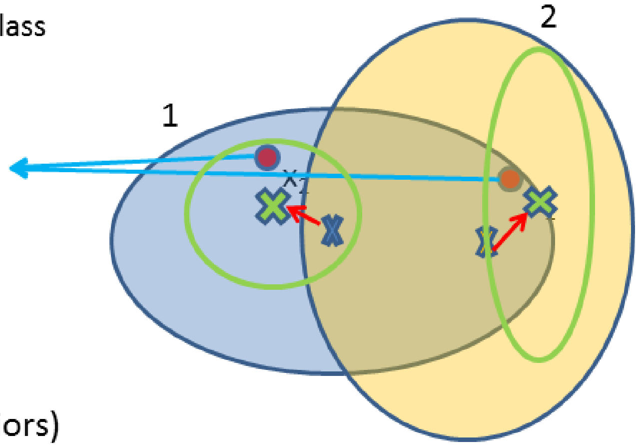
3. M-step:
- $d\mathbf{S}_k/dt$ (means, covariances, priors)
- $\mathbf{S}_k(t+dt)=\mathbf{S}_k(t)+ d\mathbf{S}_k/dt$

Clustering - Gaussian distribution

1. Initialization of parameters (a priori knowledge)

2. E – step: compute similarities l(n|k) and class memberships f(k|n)

$$\left.\begin{array}{l} l(x_2|1)=0.1 \\ l(x_2|2)=0.2 \\ l(x_1|1)=0.3 \\ l(x_1|2)=0.3 \end{array}\right\} \quad \begin{array}{l} f(1|x_1), f(1|x_2), \\ f(2|x_1), f(2|x_2) \end{array}$$



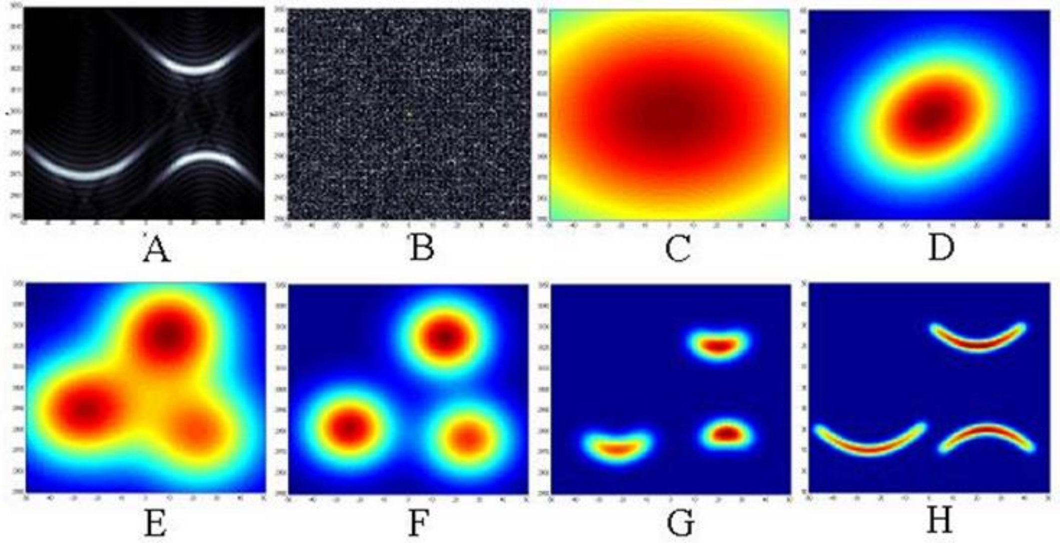
$$l_j(\vec{x}_i|\vec{m}_j,\vec{S}_j) = (2\pi)^{-d/2}\vec{S}_j^{-1/2}\exp[-0.5(\vec{x}_i-\vec{m}_j)^T \vec{S}_j^{-1}(\vec{x}_i-\vec{m}_j)] \quad (5)$$

3. M-step:
   - $d\mathbf{S}_k/dt$ (means, covariances, priors)
   - $\mathbf{S}_k(t+dt)=\mathbf{S}_k(t)+ d\mathbf{S}_k/dt$

4. LL(t)-LL(t-dt) < threshold ?

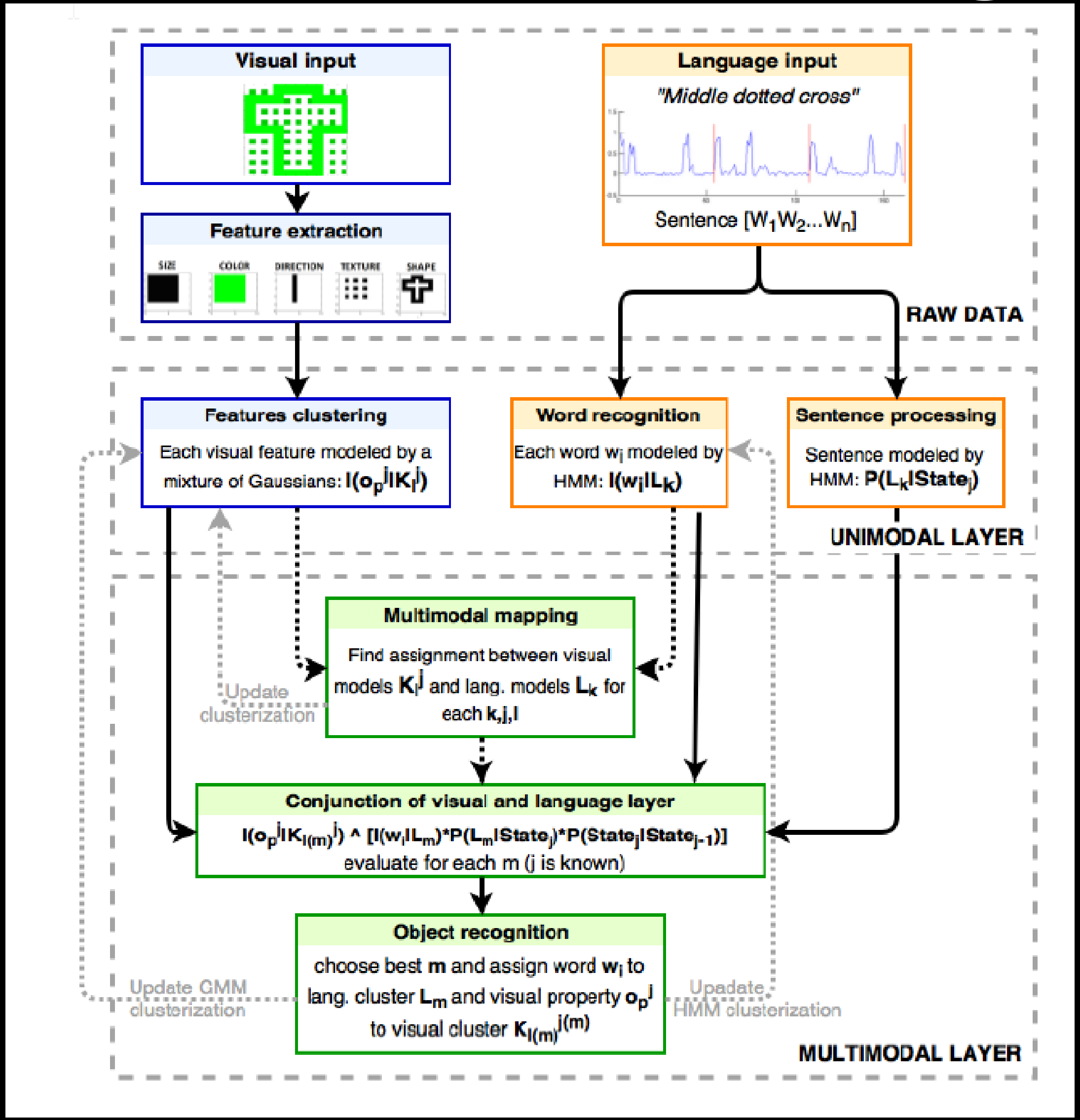
$$LL(\vec{\theta}) = \sum_{i=1}^{n}\ln\left(\sum_{j=1}^{K} r_j l_j(\vec{x}_i|\vec{m}_j,\vec{S}_j)\right)$$

Clustering - Gaussian distribution

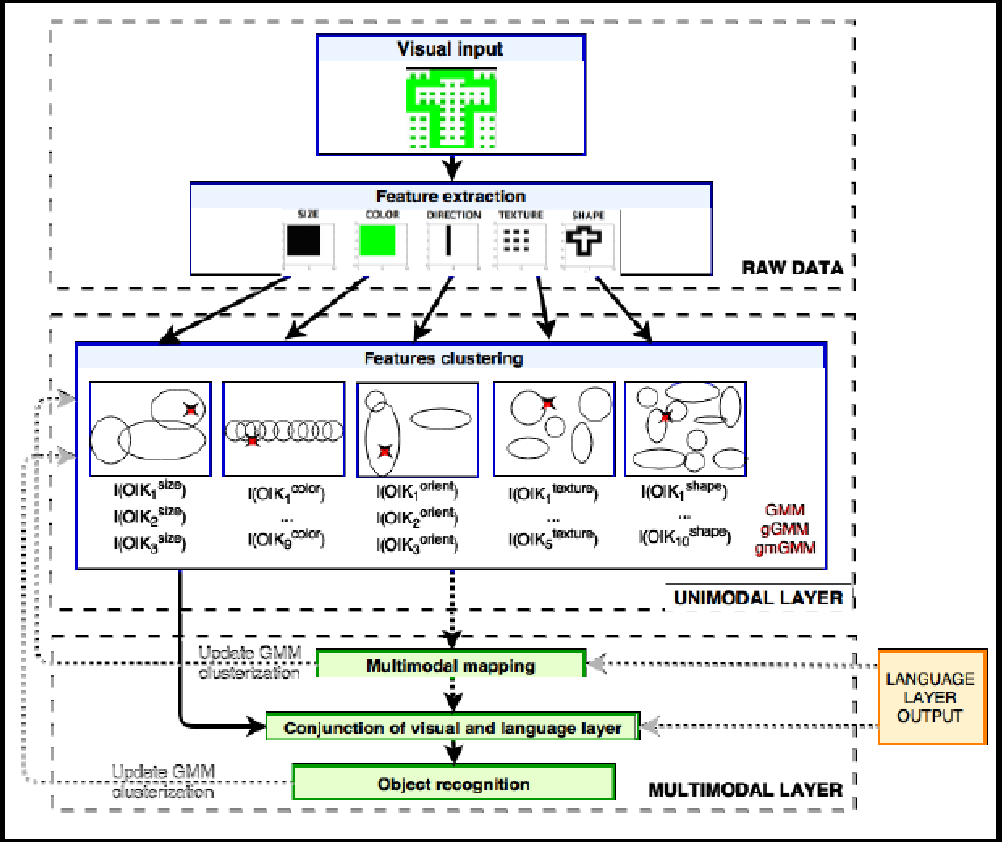


Paralels - unsupervised clustering
- mixture models
- EM algorithm

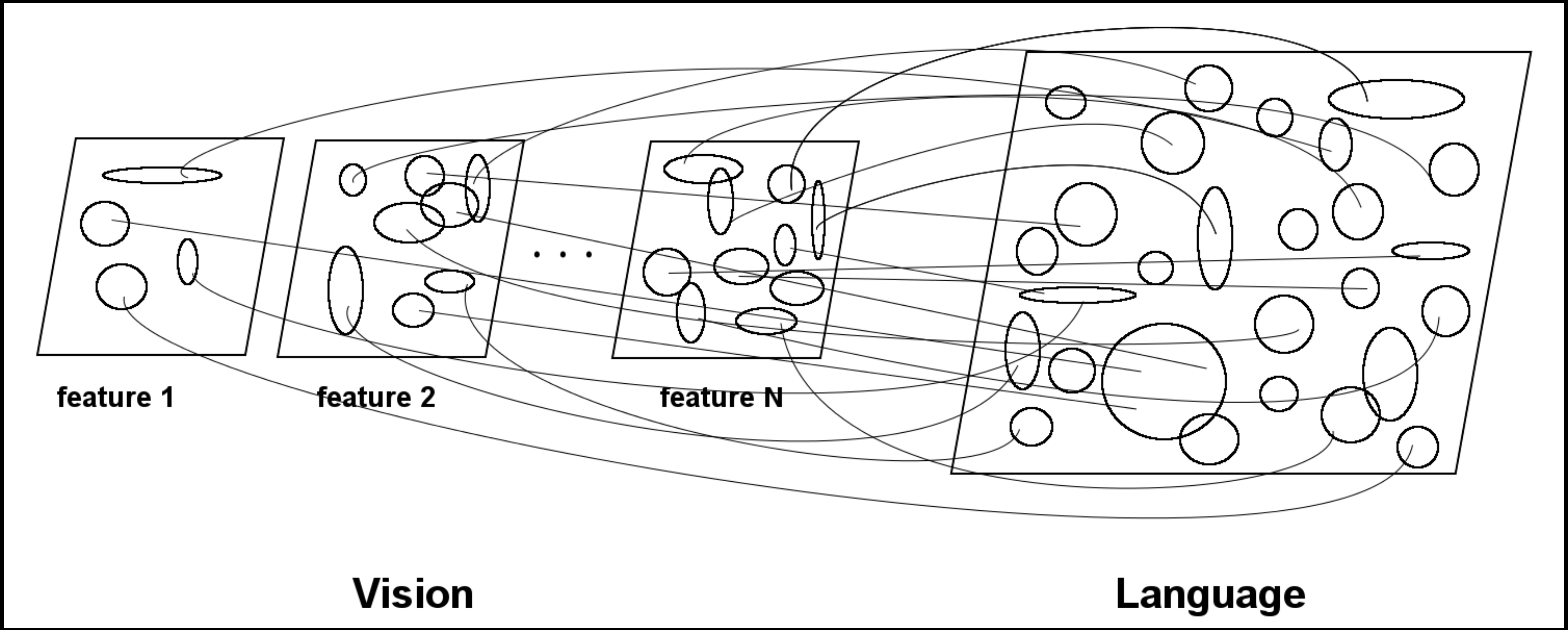
Cognitive architecture - mapping



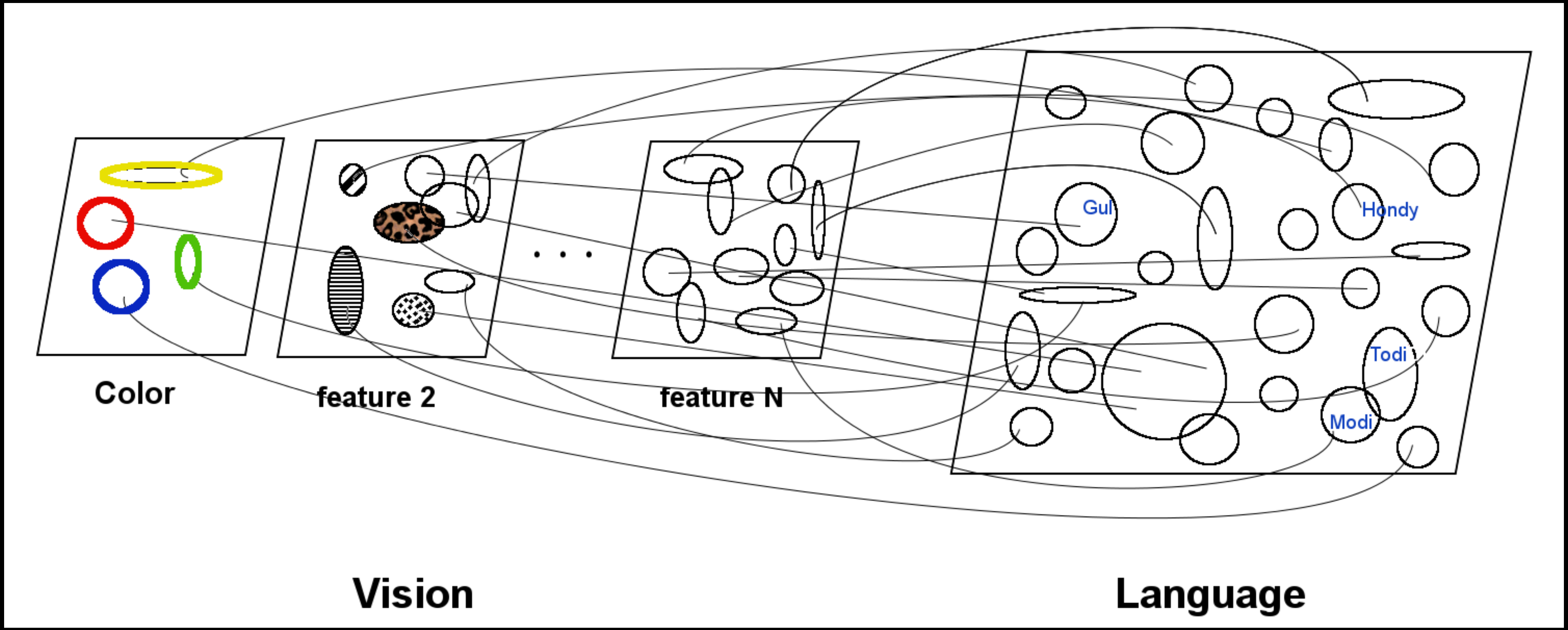
Cognitive architecture - mapping



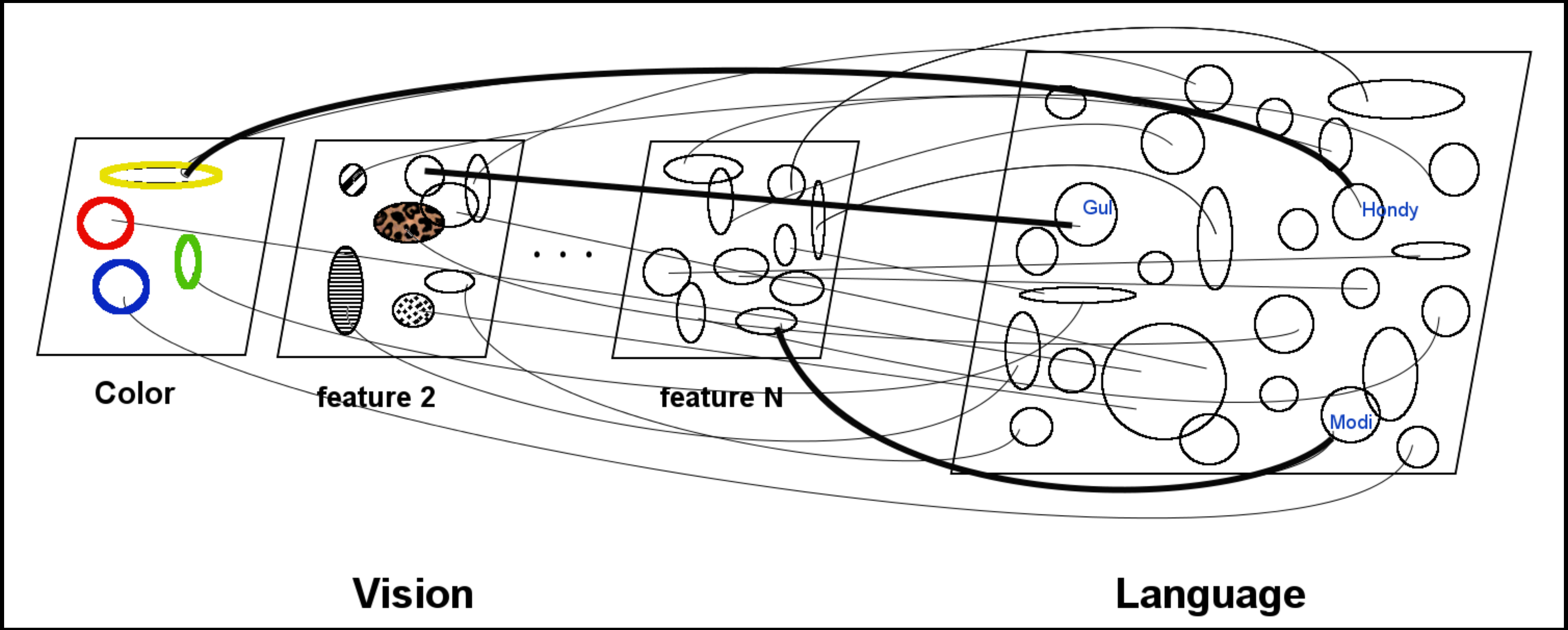
Cognitive architecture - mapping



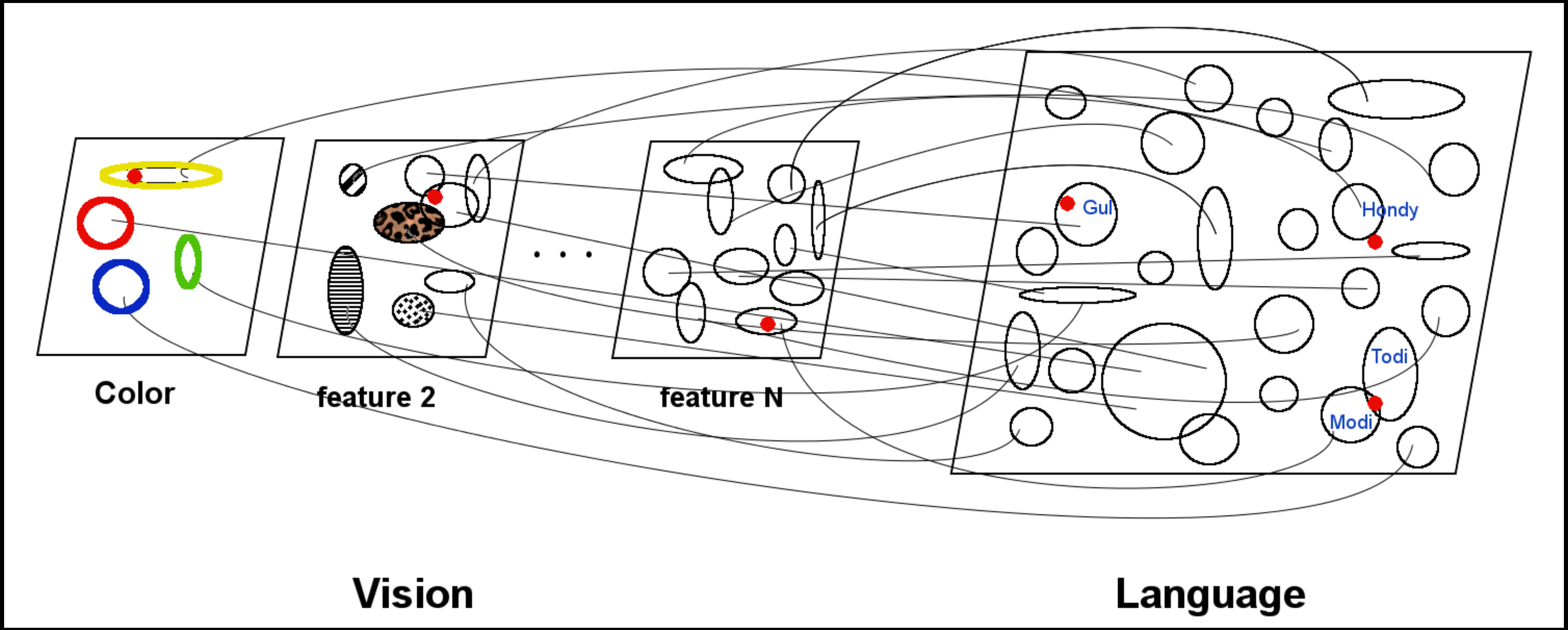
Cognitive architecture - mapping



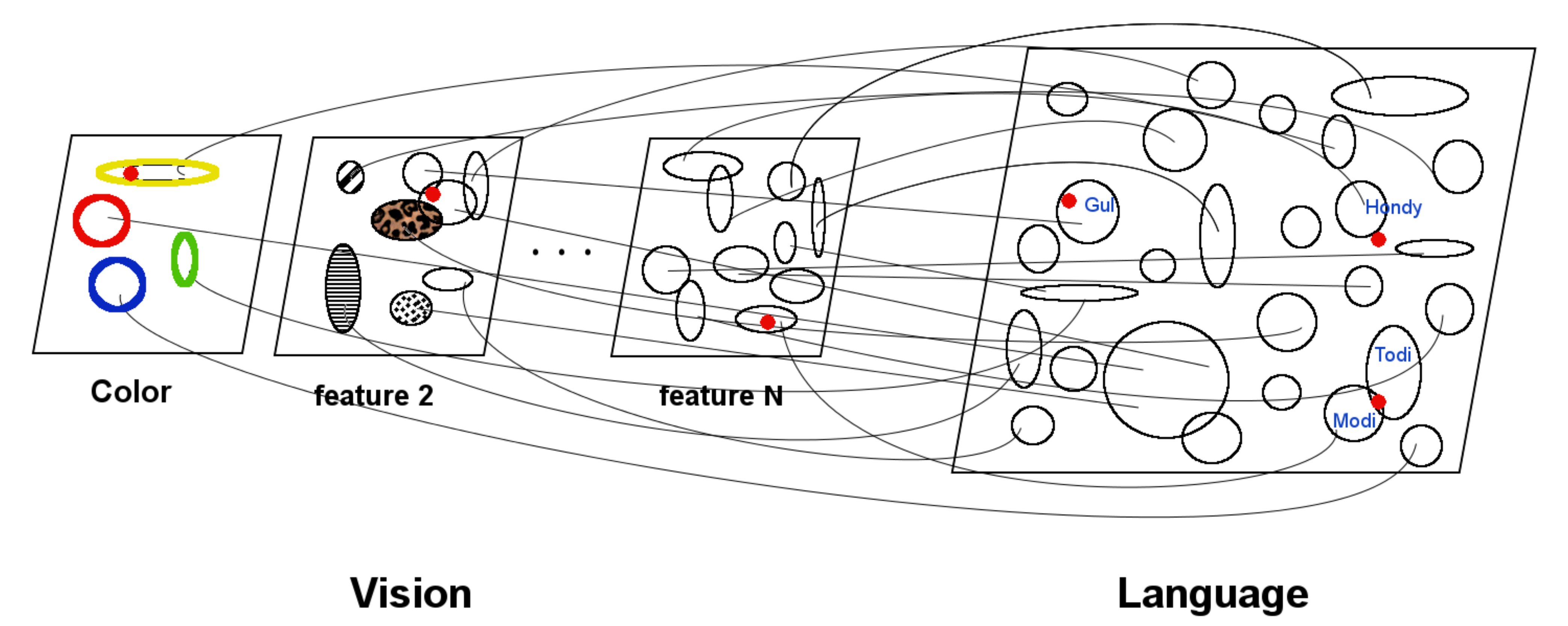
Cognitive architecture - mapping



Cognitive architecture - mapping

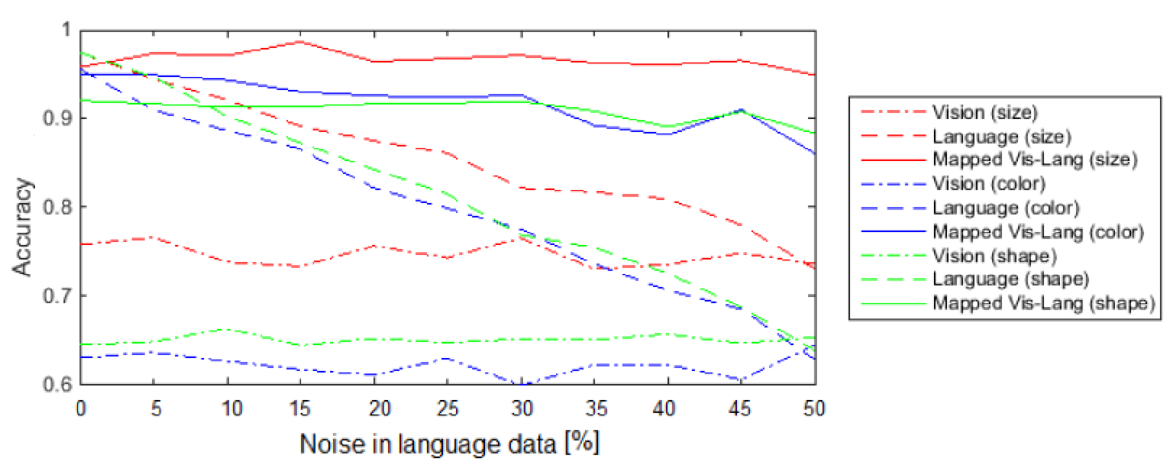


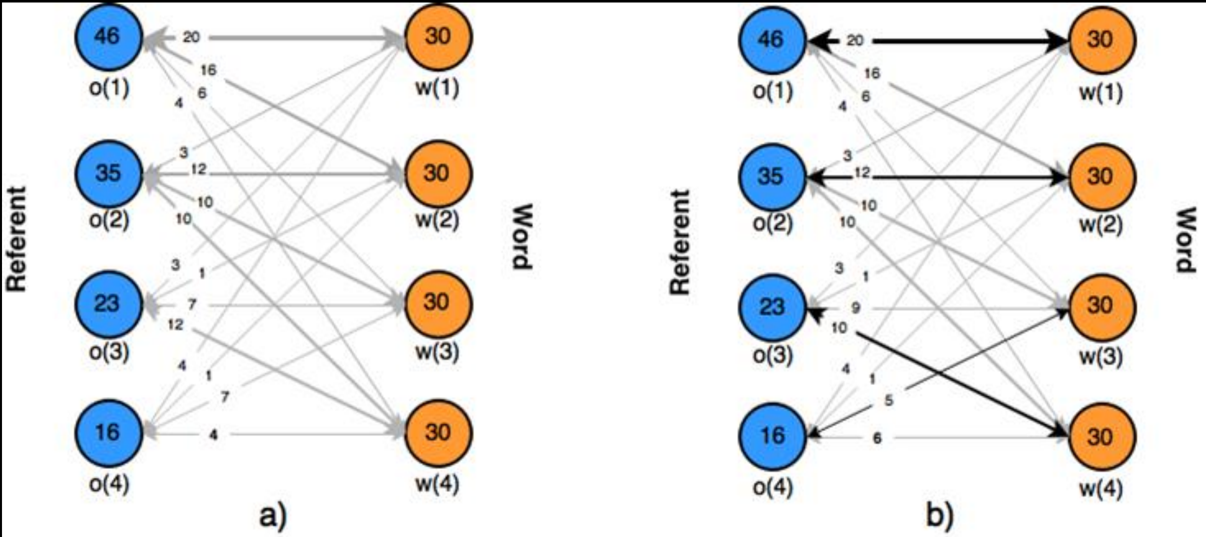
Cognitive architecture - mapping



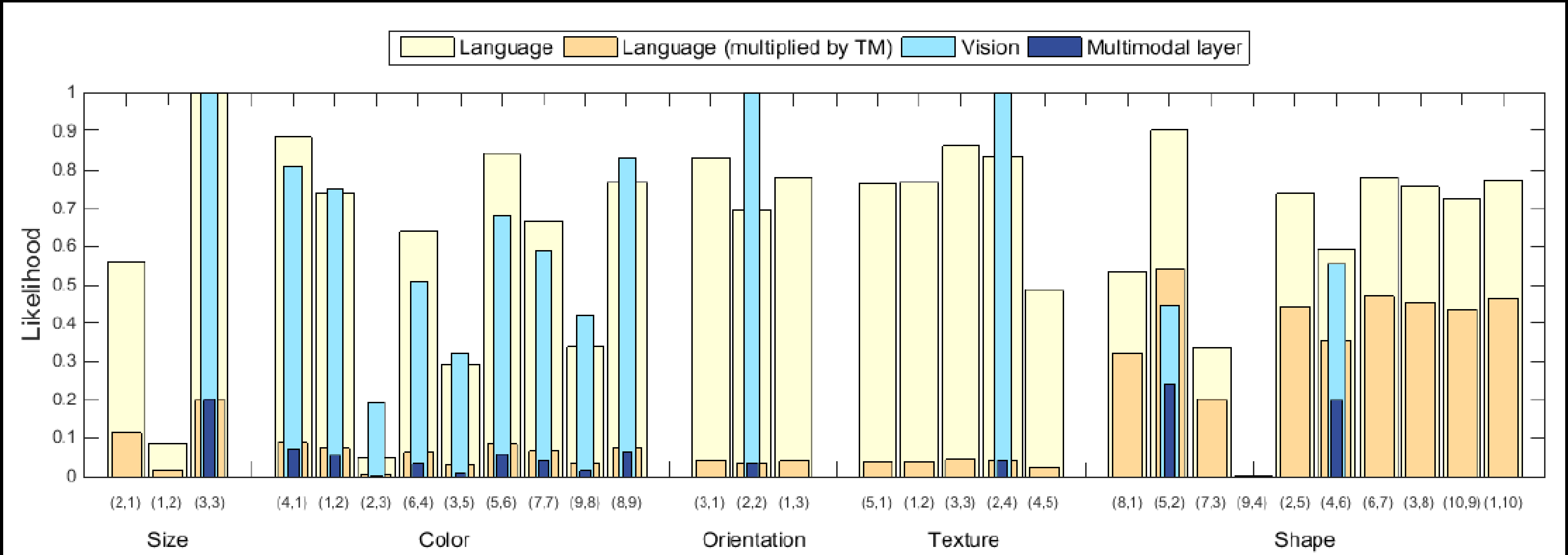
Cognitive architecture - mapping

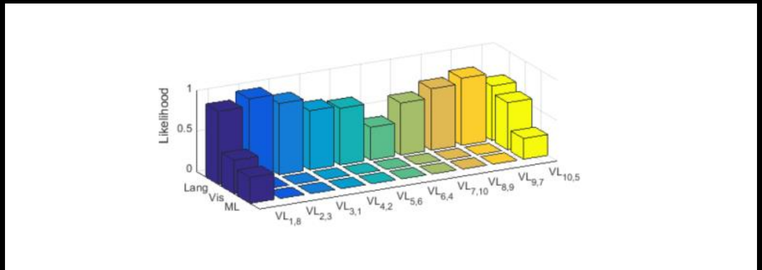
$$\phi(i,k) = p(i,k)\log_2 \frac{p(i,k)}{p(i)p(k)} \qquad p(i) = \frac{1}{N}\sum_{n=1}^{N} \lim f_i(\vec{x_n}) \qquad p(i,k) = \frac{1}{N}\sum_{n=1}^{N} \lim f_i(\vec{x_n}) f_k(\vec{x_n})$$



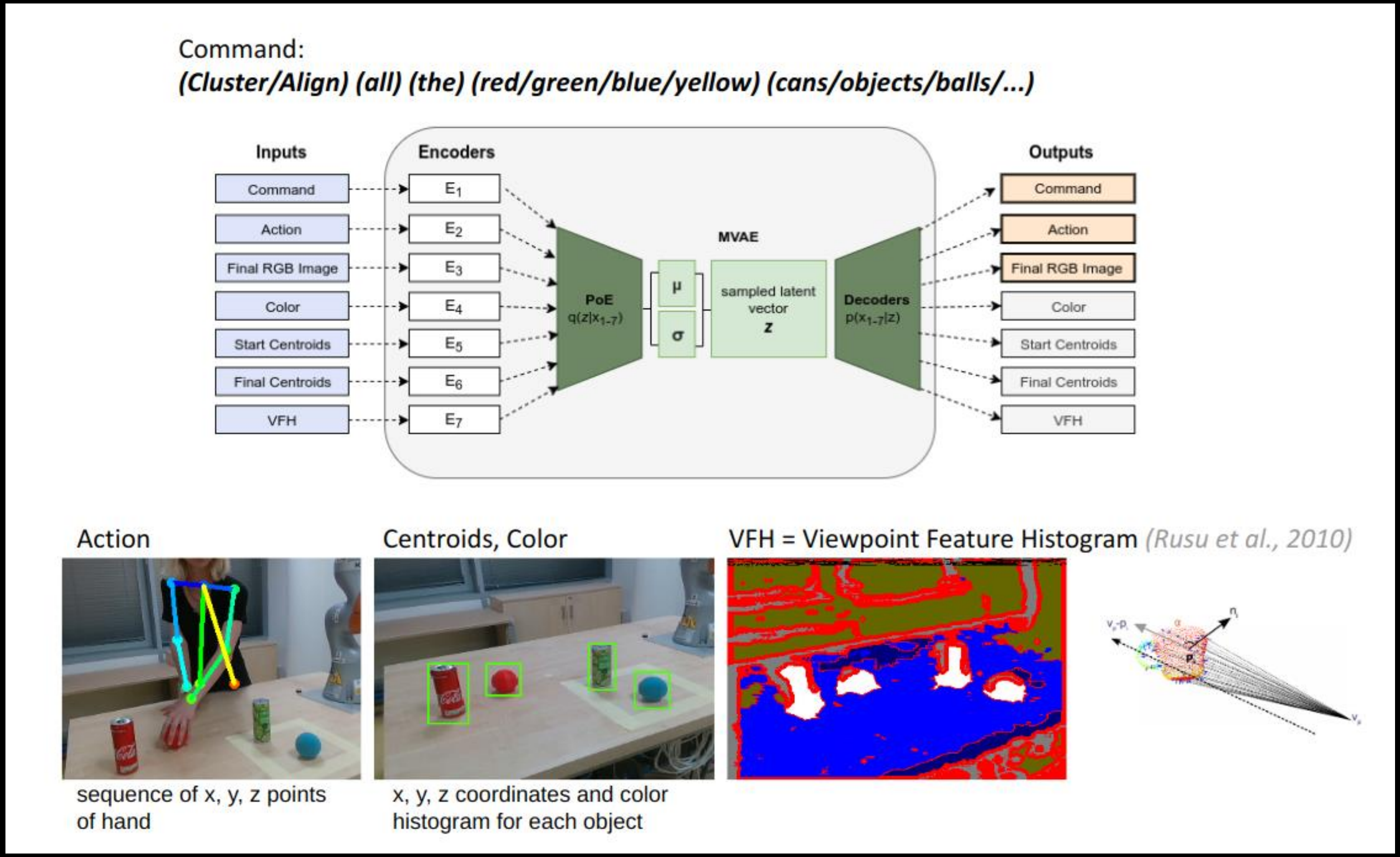


Cognitive architecture - mapping





Multimodal autoencoders and conditioning

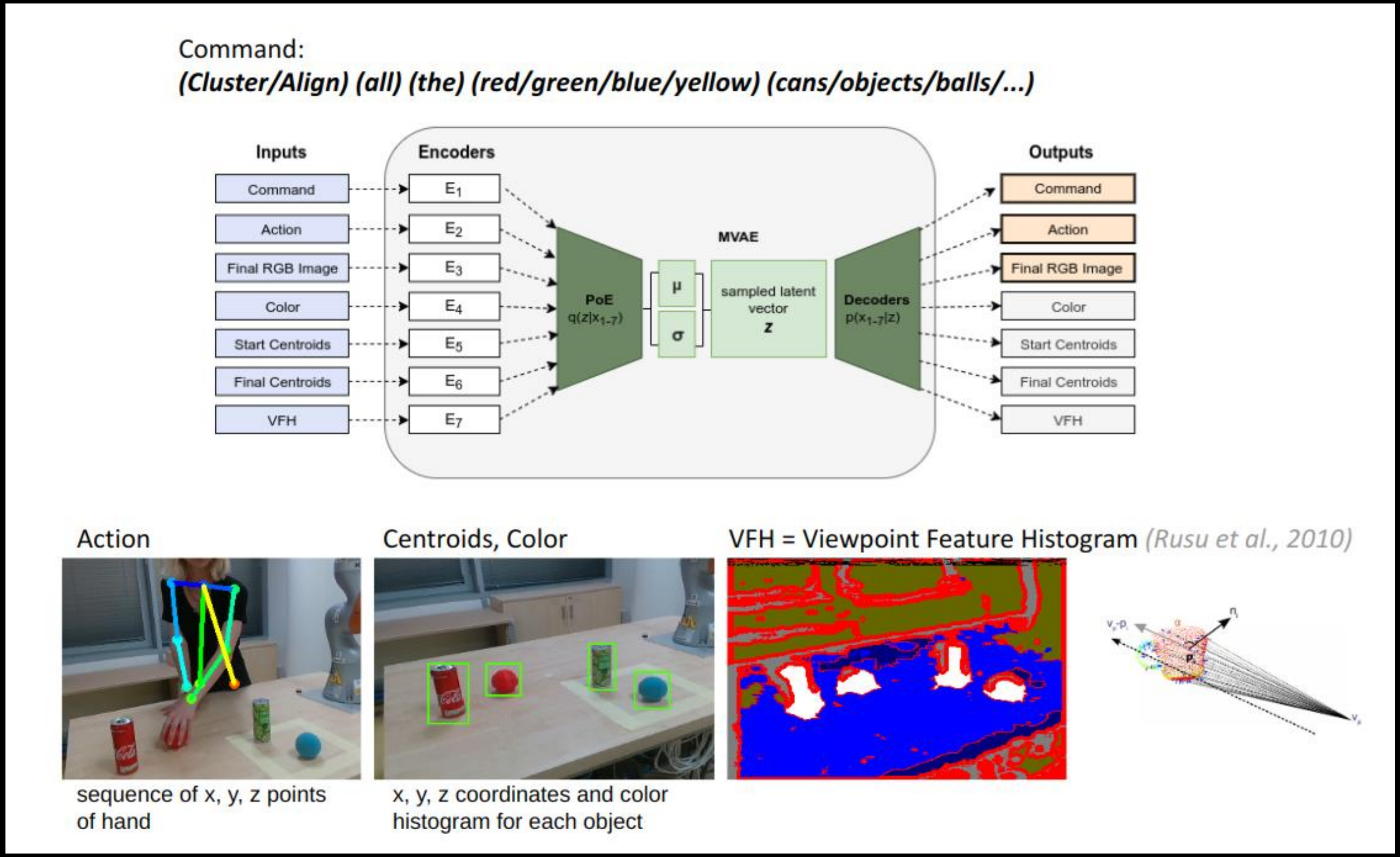


Observable variable x sampled from the latent space (mean, variance)

Expectation of a random variable	$E_x[f(x)] = \int x f(x) dx$
Chain rule of probability	$P(x,y) = P(x|y)P(y)$
Bayes' Theorem	$P(x \mid y) = \frac{P(y|x)P(x)}{P(y)}$

Similar concepts in diffusion models

Multimodal autoencoders and conditioning



Similar concepts in diffusion models

Observable variable x sampled from the latent space (mean, variance)

Expectation of a random variable	$E_x[f(x)] = \int x f(x) dx$
Chain rule of probability	$P(x, y) = P(x\|y)P(y)$
Bayes' Theorem	$P(x \mid y) = \frac{P(y\|x)P(x)}{P(y)}$

Kullback-Leiber divergence measures distance between two probability distributions

$$D_{KL}(P\|Q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

Multimodal autoencoders and conditioning

We can define the likelihood of our data as the marginalization over the joint probability with respect to the latent variable

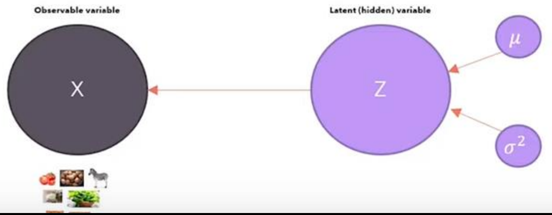
$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}$$

Is **intractable** because we would need to evaluate this integral over all latent variables Z.

… or we can use the Chain rule of probability

$$p(\boldsymbol{x}) = \frac{p(\boldsymbol{x}, \boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})}$$

We **don't** have a ground truth $p(\boldsymbol{z}|\boldsymbol{x})$
… which is also what we're trying to find!



Intractable problem = a problem that can be solved in theory (e.g. given large but finite resources, especially time), but for which in practice any solution takes too many resources to be useful, is known as an intractable problem.

Observable variable x sampled from the latent space (mean, variance)

Expectation of a random variable $\quad E_x[f(x)] = \int x f(x) dx$

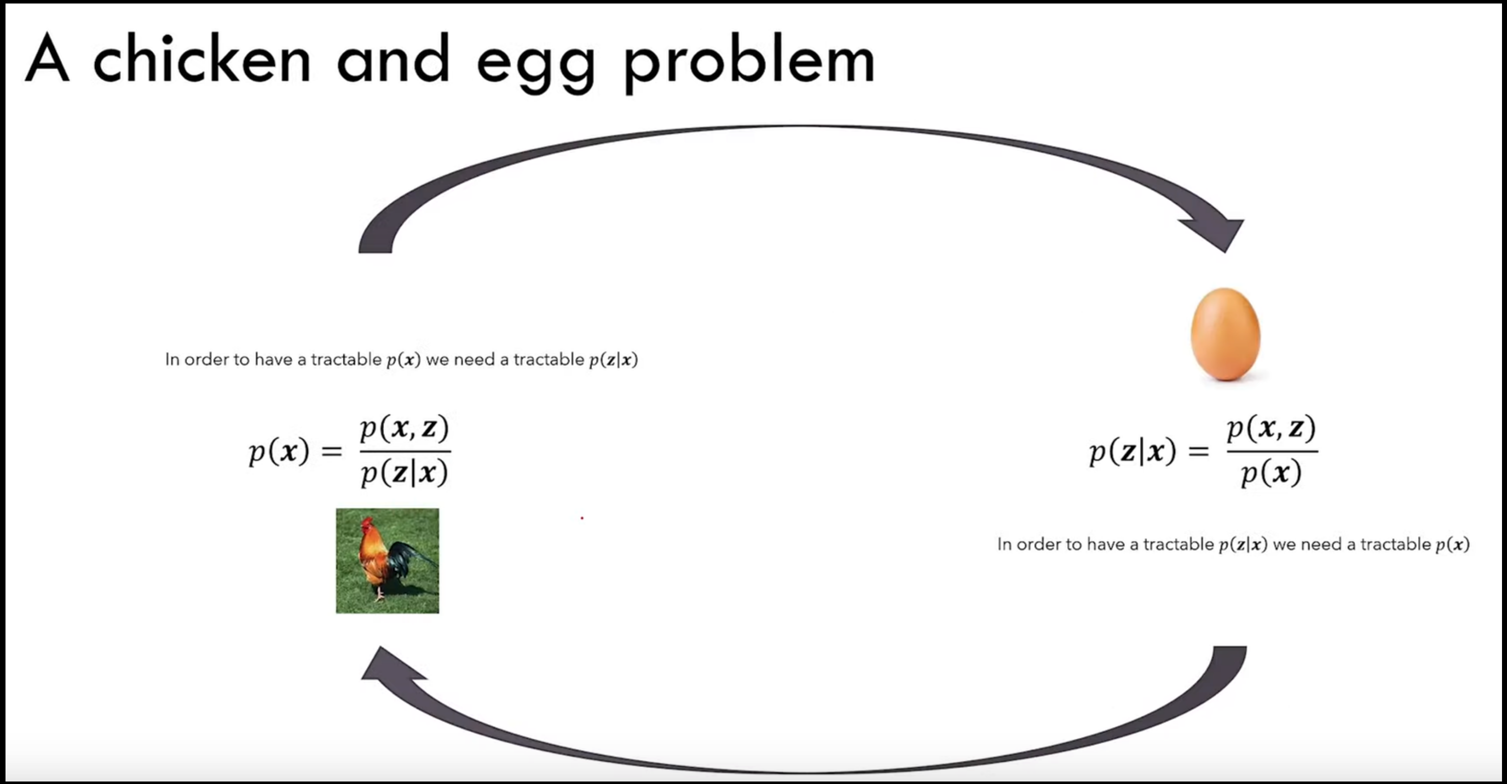
Chain rule of probability $\quad P(x, y) = P(x|y)P(y)$

Bayes' Theorem $\quad P(x \mid y) = \frac{P(y|x)P(x)}{P(y)}$

Kullback-Leiber divergence measures distance between two probability distributions

$$D_{KL}(P \| Q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

Multimodal autoencoders and conditioning



-> Need to approximate

Observable variable x sampled from the latent space (mean, variance)

Expectation of a random variable	$E_x[f(x)] = \int x f(x) dx$
Chain rule of probability	$P(x, y) = P(x\|y)P(y)$
Bayes' Theorem	$P(x \mid y) = \frac{P(y\|x)P(x)}{P(y)}$

Kullback-Leiber divergence measures distance between two probability distributions

$$D_{KL}(P\|Q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

Multimodal autoencoders and conditioning

$$\log p_\theta(\boldsymbol{x}) = \log p_\theta(\boldsymbol{x})$$

$$= \log p_\theta(\boldsymbol{x}) \boxed{\int q_\varphi(\mathbf{z}|\boldsymbol{x})d\mathbf{z}}$$ Multiply by 1

$$= \int \log p_\theta(\boldsymbol{x})\, q_\varphi(\mathbf{z}|\boldsymbol{x})d\mathbf{z}$$ Bring inside the integral

$$= E_{q_\varphi(\mathbf{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})]$$ Definition of expectation

$$= E_{q_\varphi(\mathbf{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x},\mathbf{z})}{p_\theta(\mathbf{z}|\boldsymbol{x})}\right]$$ Apply the equation $p_\theta(\boldsymbol{x}) = \frac{p_\theta(\boldsymbol{x},\mathbf{z})}{p_\theta(\mathbf{z}|\boldsymbol{x})}$

$$= E_{q_\varphi(\mathbf{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x},\mathbf{z})q_\varphi(\mathbf{z}|\boldsymbol{x})}{p_\theta(\mathbf{z}|\boldsymbol{x})q_\varphi(\mathbf{z}|\boldsymbol{x})}\right]$$ Multiply by 1

$$= E_{q_\varphi(\mathbf{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x},\mathbf{z})}{q_\varphi(\mathbf{z}|\boldsymbol{x})}\right] + E_{q_\varphi(\mathbf{z}|\boldsymbol{x})}\left[\log \frac{q_\varphi(\mathbf{z}|\boldsymbol{x})}{p_\theta(\mathbf{z}|\boldsymbol{x})}\right]$$ Split the expectation

$$= E_{q_\varphi(\mathbf{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x},\mathbf{z})}{q_\varphi(\mathbf{z}|\boldsymbol{x})}\right] + \underbrace{D_{KL}\left(q_\varphi(\mathbf{z}|\boldsymbol{x}) \| p_\theta(\mathbf{z}|\boldsymbol{x})\right)}_{\geq 0}$$ Definition of KL divergence

Observable variable x sampled from the latent space (mean, variance)

Expectation of a random variable	$E_x[f(x)] = \int x f(x)dx$
Chain rule of probability	$P(x,y) = P(x\|y)P(y)$
Bayes' Theorem	$P(x \mid y) = \frac{P(y\|x)P(x)}{P(y)}$

Kullback-Leiber divergence measures distance between two probability distributions $D_{KL}(P\|Q) = \int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$

Multimodal autoencoders and conditioning

$$\log p_\theta(\boldsymbol{x}) = \underbrace{E_{q_\varphi(\mathbf{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x},\mathbf{z})}{q_\varphi(\mathbf{z}|\boldsymbol{x})}\right]}_{\textbf{ELBO}} + \underbrace{D_{KL}\left(q_\varphi(\mathbf{z}|\boldsymbol{x}) \| p_\theta(\mathbf{z}|\boldsymbol{x})\right)}_{\geq 0}$$

ELBO = Evidence Lower Bound

MAX MAX

$$\underline{\log p_\theta(\boldsymbol{x})} \geq \underbrace{E_{q_\varphi(\mathbf{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x},\mathbf{z})}{q_\varphi(\mathbf{z}|\boldsymbol{x})}\right]}_{\text{ELBO}} = E_{q_\varphi(\mathbf{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x}|\mathbf{z})p(\mathbf{z})}{q_\varphi(\mathbf{z}|\boldsymbol{x})}\right]$$ Chain rule of probability

$$= E_{q_\varphi(\mathbf{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\mathbf{z})] + E_{q_\varphi(\mathbf{z}|\boldsymbol{x})}\left[\log \frac{p(\mathbf{z})}{q_\varphi(\mathbf{z}|\boldsymbol{x})}\right]$$ Split the expectation

$$= \underline{E_{q_\varphi(\mathbf{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\mathbf{z})] - D_{KL}\left(q_\varphi(\mathbf{z}|\boldsymbol{x}) \| p_\theta(\mathbf{z})\right)}$$ Definition of KL divergence

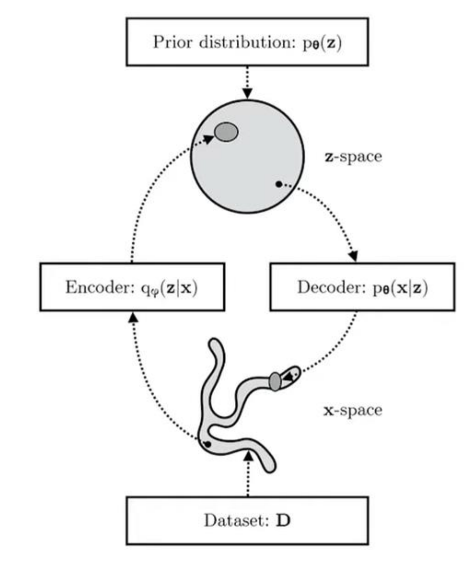
MAX MIN

Maximizing the ELBO means:

1. Maximizing the first term: maximizing the reconstruction likelihood of the decoder
2. Minimizing the second term: minimizing the distance between the learned distribution and the prior belief we have over the latent variable.

Observable variable x sampled from the latent space (mean, variance)

- We want z-space to be multivariate gaussian
- Learning to improve the reconstruction quality of x given the z space



Kingma, D.P. and Welling, M., 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4), pp.307-392.

Multimodal autoencoders and conditioning

How to maximize something that has a stochastic variable inside? (ELBO)

Observable variable x sampled from the latent space (mean, variance)

- We want z-space to be multivariate gaussian
- Learning to improve the reconstruction quality of x given the z space

$$L(\theta, \varphi, \boldsymbol{x}) = E_{q_\varphi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\varphi(\boldsymbol{z}|\boldsymbol{x})}\right] = E_{q_\varphi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}\left(q_\varphi(\boldsymbol{z}|\boldsymbol{x}) \| p_\theta(\boldsymbol{z})\right)$$

- When we have a function we want to **maximize**, we usually take the gradient and adjust the weights of the model so that they move **along** the gradient direction.
- When we have a function we want to **minimize**, we usually take the gradient, and adjust the weights of the model so that they move **against** the gradient direction.

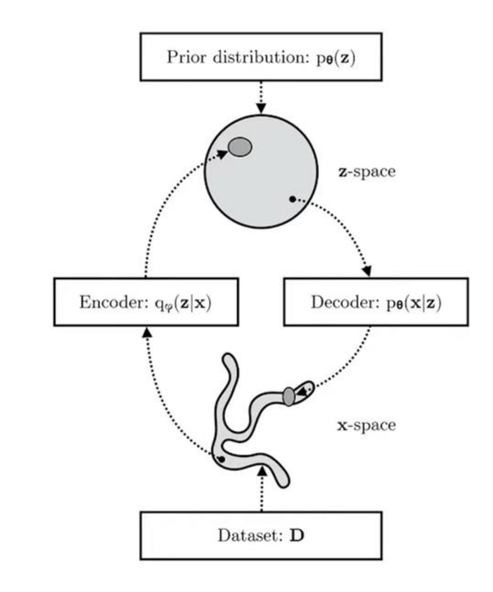
Stochastic Gradient Descent

When used to minimize the above function, a standard (or "batch") gradient descent method would perform the following iterations:

$$w := w - \eta \nabla Q(w) = w - \frac{\eta}{n} \sum_{i=1}^{n} \nabla Q_i(w),$$

where η is a step size (sometimes called the *learning rate* in machine learning).

https://www.youtube.com/watch?v=iwEzwTTalbg



Kingma, D.P. and Welling, M., 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, *12*(4), pp.307-392.

Multimodal autoencoders and conditioning

How to maximize something that has a stochastic variable inside? (ELBO)

+ Cannot run backpropagation on stochastic variable – need to sample from z space…reparametrization trick

$$L(\theta, \varphi, \boldsymbol{x}) = E_{q_\varphi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\varphi(\boldsymbol{z}|\boldsymbol{x})}\right] = E_{q_\varphi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{KL}\left(q_\varphi(\boldsymbol{z}|\boldsymbol{x}) \| p_\theta(\boldsymbol{z})\right)$$

- When we have a function we want to **maximize**, we usually take the gradient and adjust the weights of the model so that they move **along** the gradient direction.
- When we have a function we want to **minimize**, we usually take the gradient, and adjust the weights of the model so that they move **against** the gradient direction.

Stochastic Gradient Descent

When used to minimize the above function, a standard (or "batch") gradient descent method would perform the following iterations:

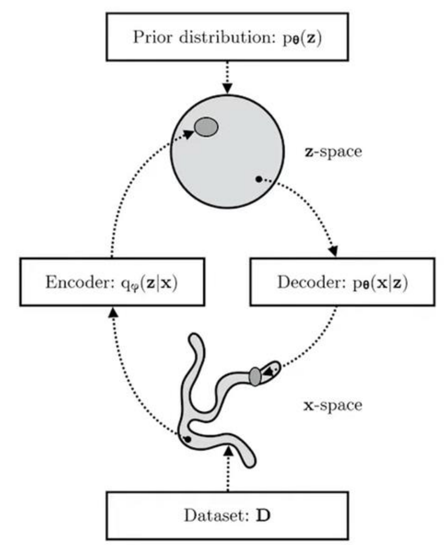
$$w := w - \eta \nabla Q(w) = w - \frac{\eta}{n} \sum_{i=1}^{n} \nabla Q_i(w),$$

where η is a step size (sometimes called the *learning rate* in machine learning).

https://www.youtube.com/watch?v=iwEzwTTalbg

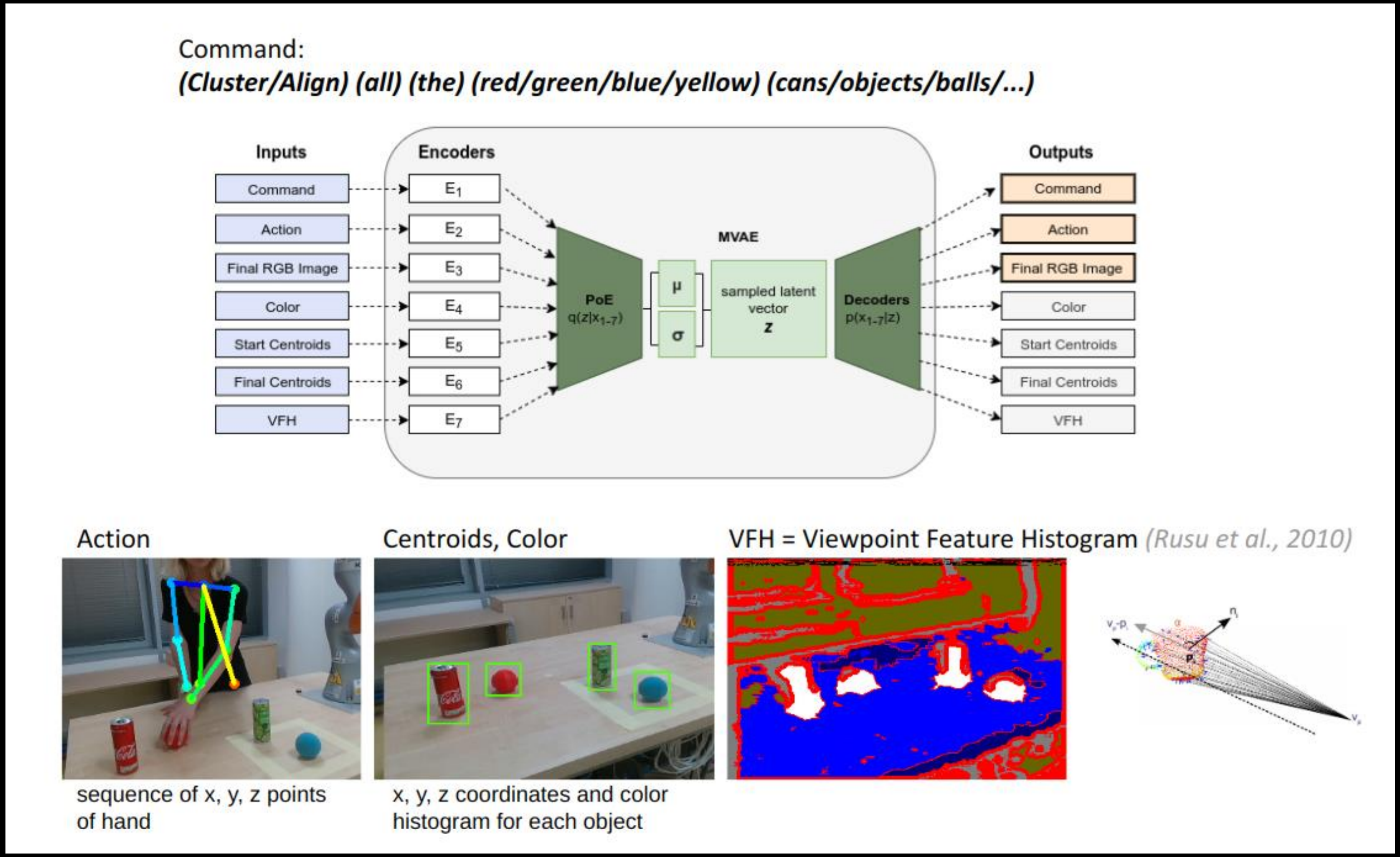
Observable variable x sampled from the latent space (mean, variance)

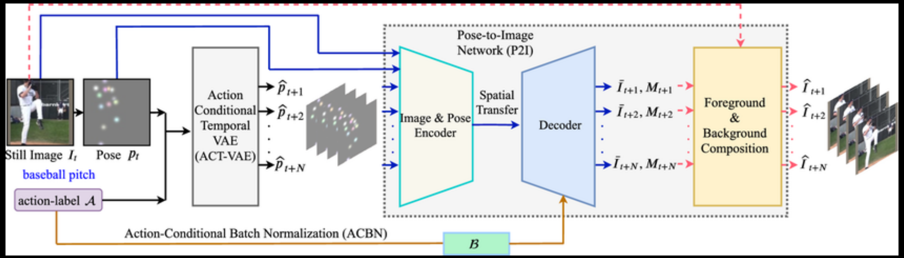
- We want z-space to be multivariate gaussian
- Learning to improve the reconstruction quality of x given the z space



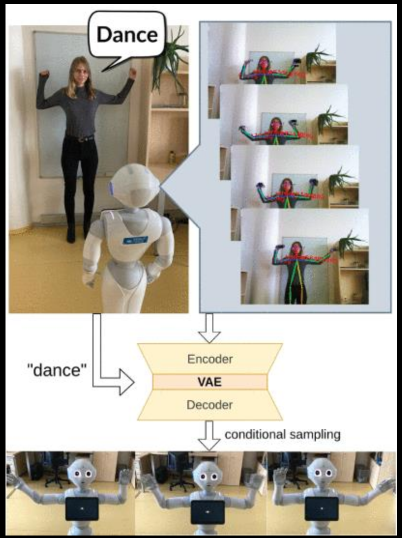
Kingma, D.P. and Welling, M., 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4), pp.307-392.

Multimodal autoencoders and conditioning





Xu, Xiaogang, et al. "Conditional temporal variational autoencoder for action video prediction." *International Journal of Computer Vision* 131.10 (2023): 2699-2722.



G. Sejnova and K. Stepanova, "Feedback-Driven Incremental Imitation Learning Using Sequential VAE," *2022 IEEE International Conference on Development and Learning (ICDL)*, London, United Kingdom, 2022, pp. 238-243.

Mapping gestures to robot action given the context of the situation

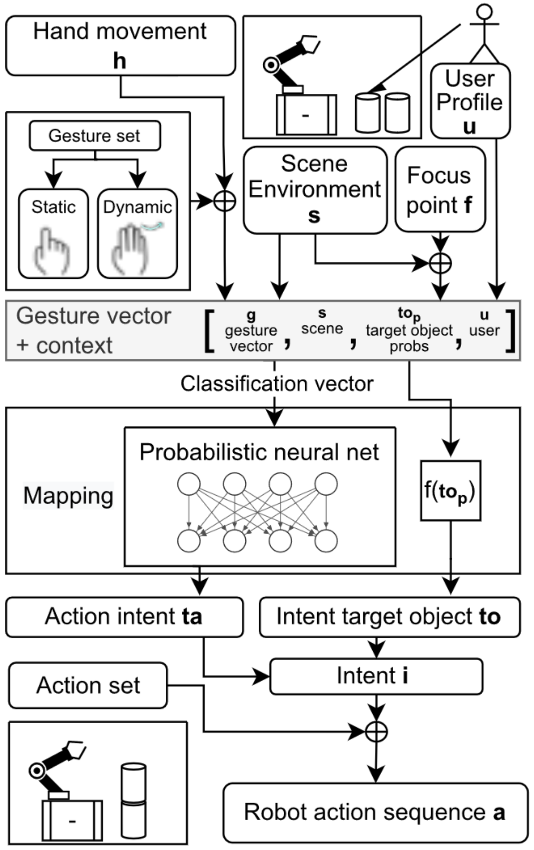


Fig. 1: Proposed system diagram displays the whole pipeline from hand observations all the way to robotic actions.

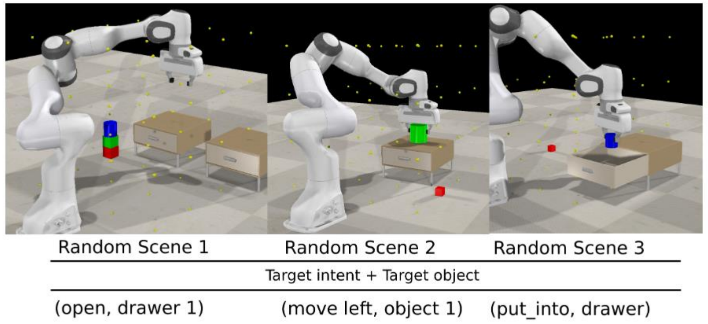


Fig. 5: Random scene example generation.

Vanc, Petr, Jan Kristof Behrens, and Karla Stepanova. "Context-aware robot control using gesture episodes." *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

Mapping language and gestures

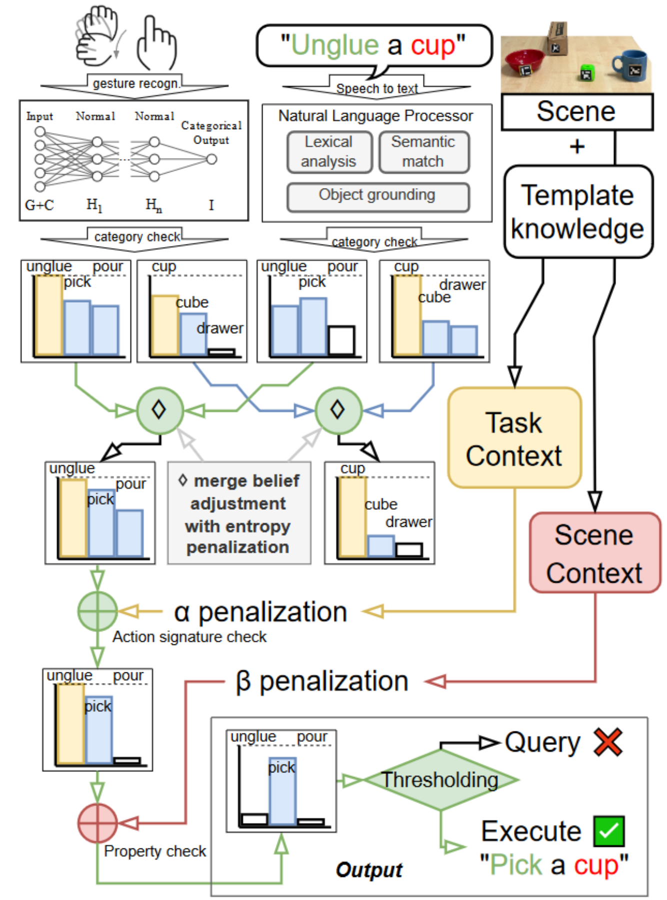


Fig. 2. Diagram of the proposed model for the case of two modalities (hand gestures and natural language) specifying action with one parameter (target object). Heard sentence "Unglue a cup" is correctly resolved into "Pick a cup" based on a fusion of data from both modalities and task and scene context".

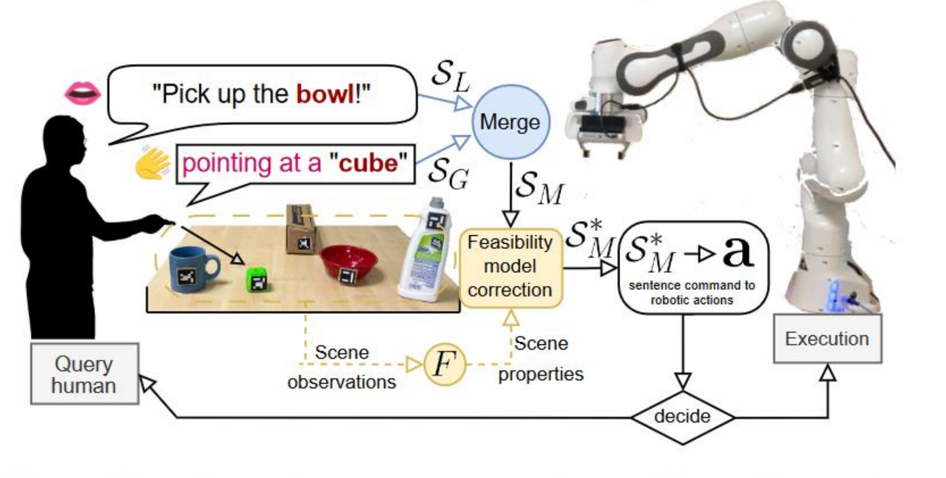
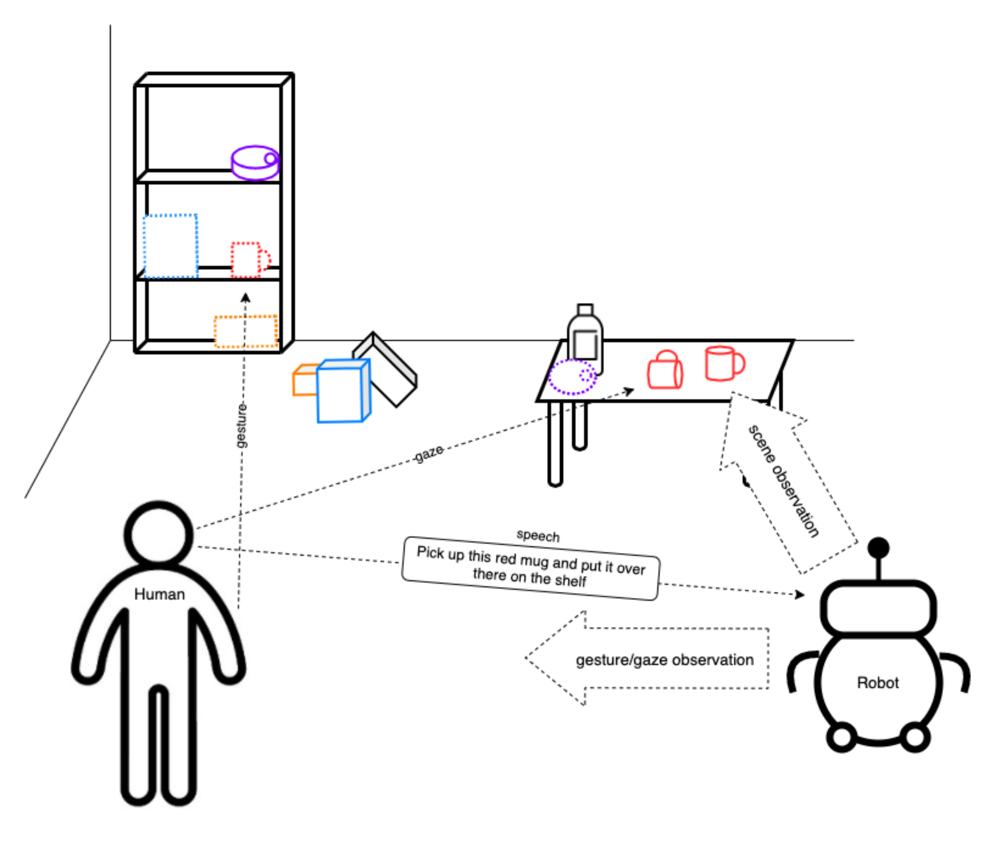
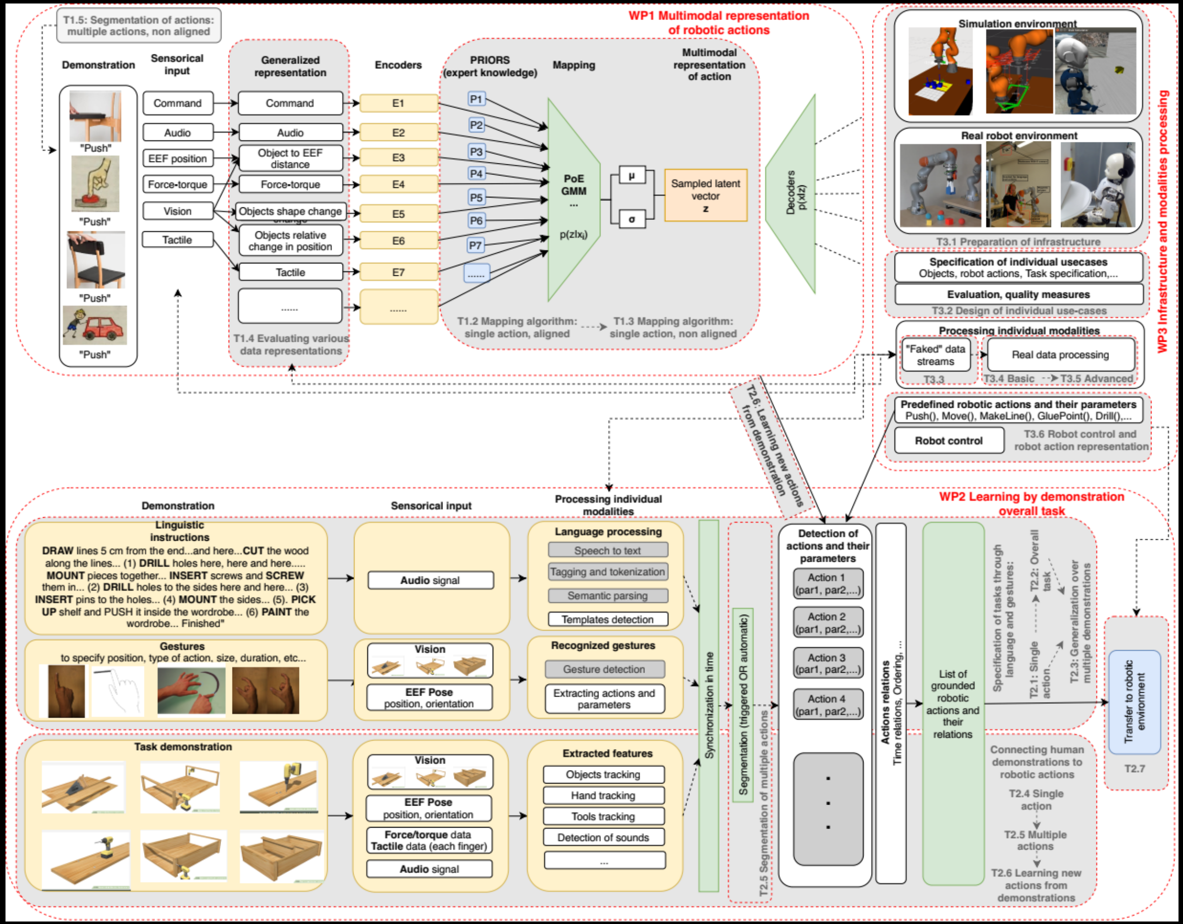


Fig. 1. Human-Robot Interaction experimental setup. The user's speech is captured by the microphone and the hand is captured by a hand detection device (e.g. Leap Motion Controller [8]).

Vanc, Petr, Radoslav Skoviera, and Karla Stepanova. "Tell and show: Combining multiple modalities to communicate manipulation tasks to a robot." *arXiv preprint arXiv:2404.01702* (2024).

iChores and Mirracle project





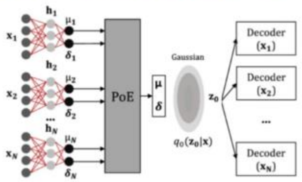
Imitrob team

http://imitrob.ciirc.cvut.cz

Contact:
Karla Stepanova
karla.stepanova@cvut.cz

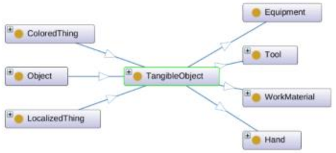
Multimodal Learning

- multimodal Variational Autoencoders
- probabilistic sensor fusion



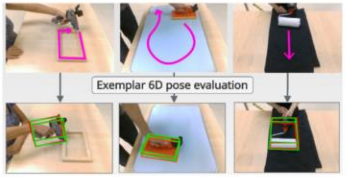
Ontologies

- knowledge representation
- learning and inference



Object Tracking

- visual segmentation
- 6DOF object detection
- trajectory motion tracking



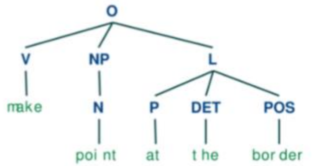
Applications

collaborative HRI workspace

Pepper for edutainment

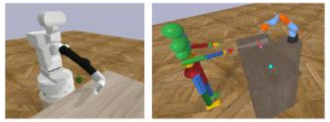
Language Understanding

- controlling robots with natural language commands
- word-sense disambiguation
- robots providing feedback



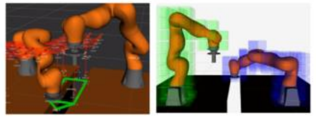
Simulators and VR

- learning from demonstration
- procedural task generation
- modular reinforcement learning
- virtual reality



Motion Planning and Scheduling

- path planning
- robot-robot collaboration
- human-robot collabotation



Gesture Recognition

- context-aware robot control using gestures
- reading human intent with multi-type gesture sentences