Probabilistic Learning Machines and the Information Revolution

Zoubin Ghahramani
Department of Engineering
University of Cambridge

What is intelligence?

What is learning?

Can we build computers and robots that learn?

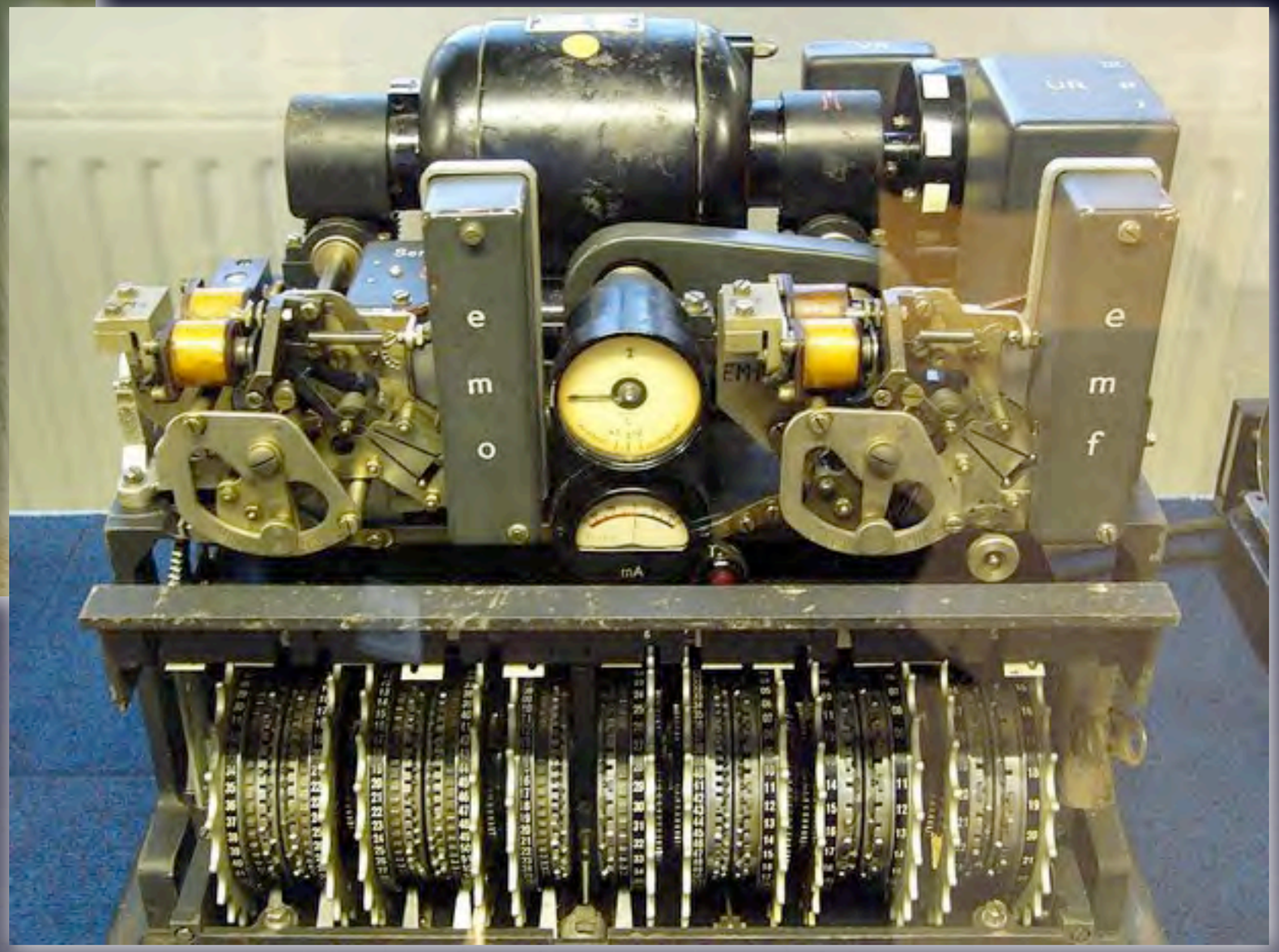
Will they ever be as intelligent...

or *more* intelligent...

than humans?

Alan Turing (1912-1954)


Zur Beachtung!


e
m
o
mA
e
m
f

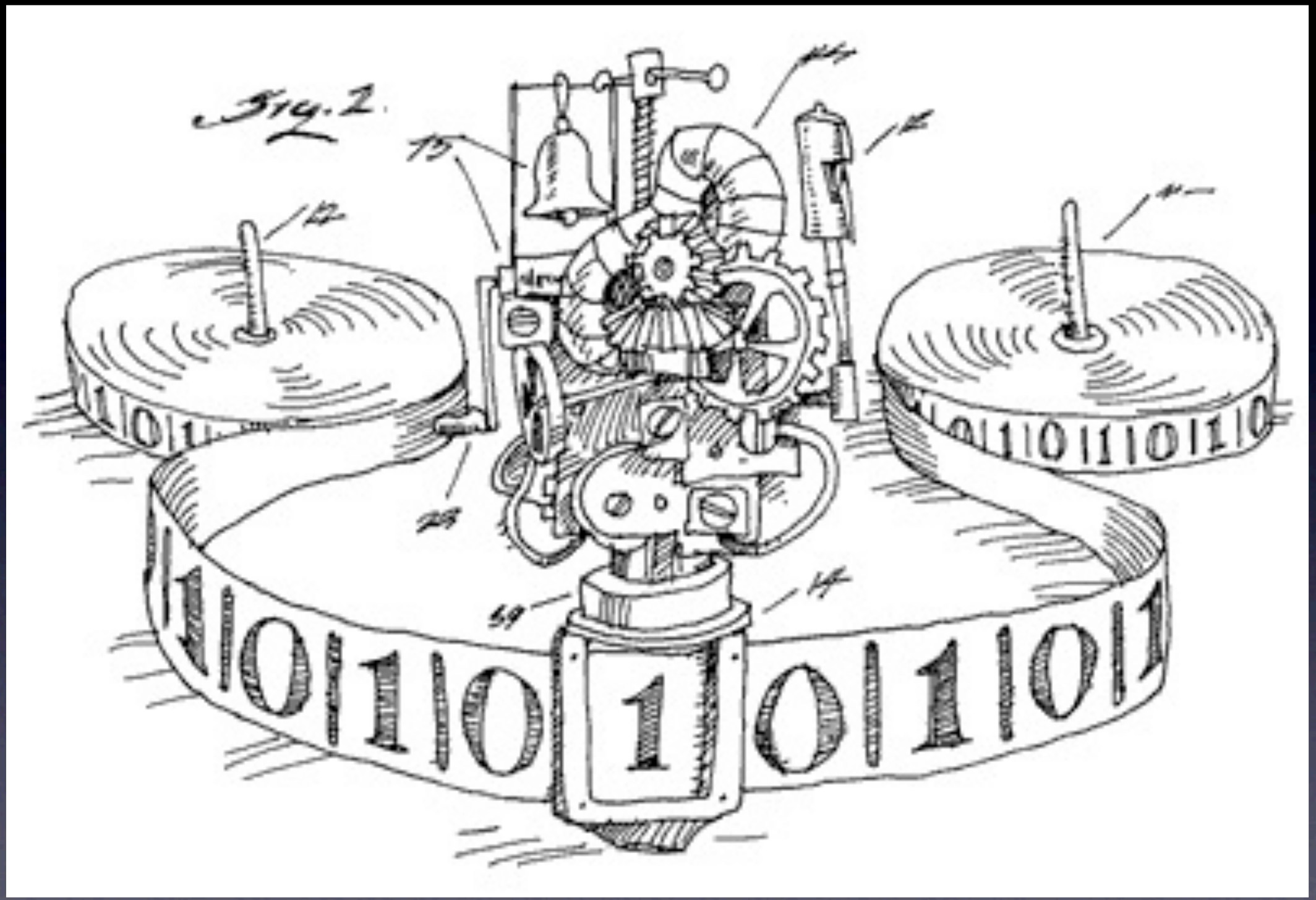
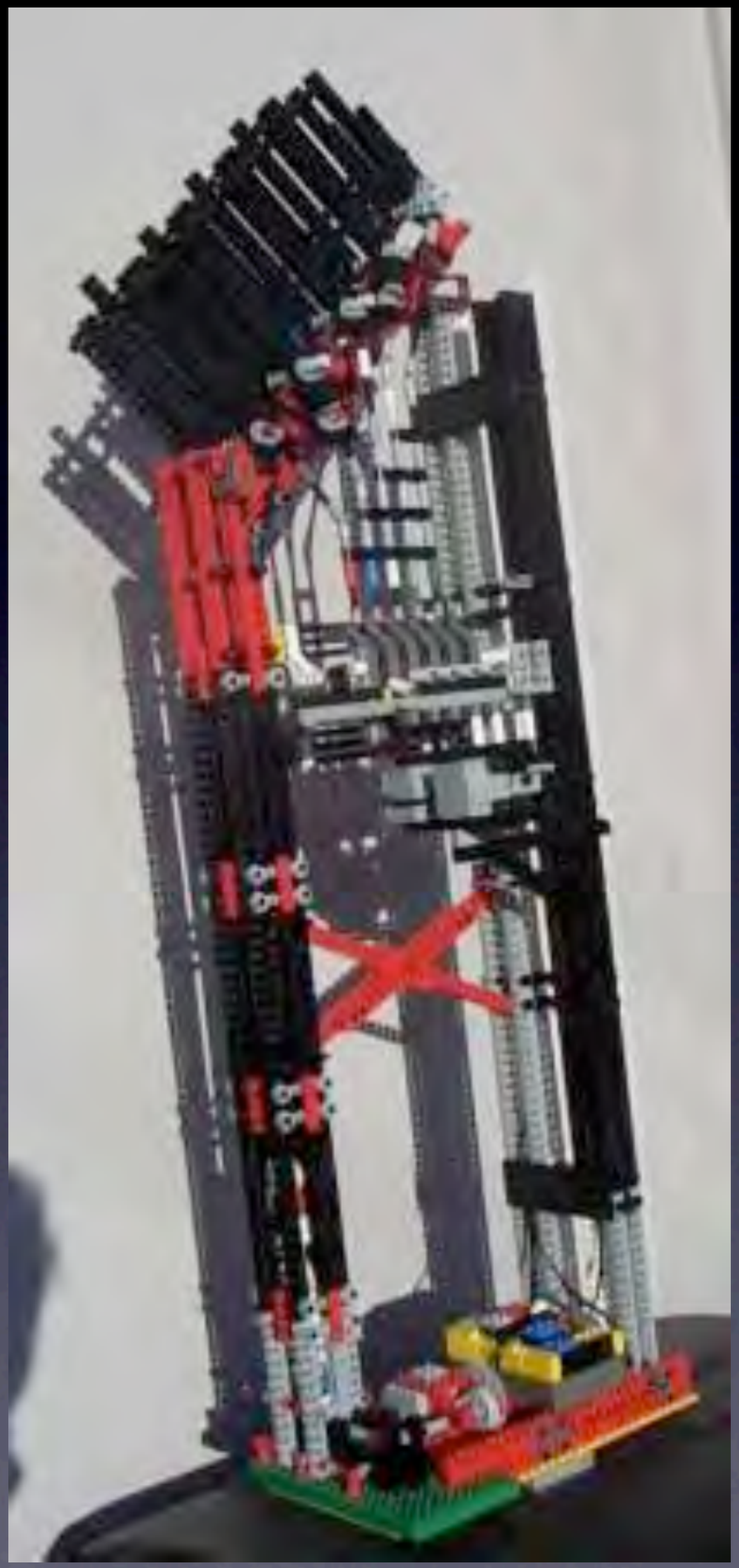


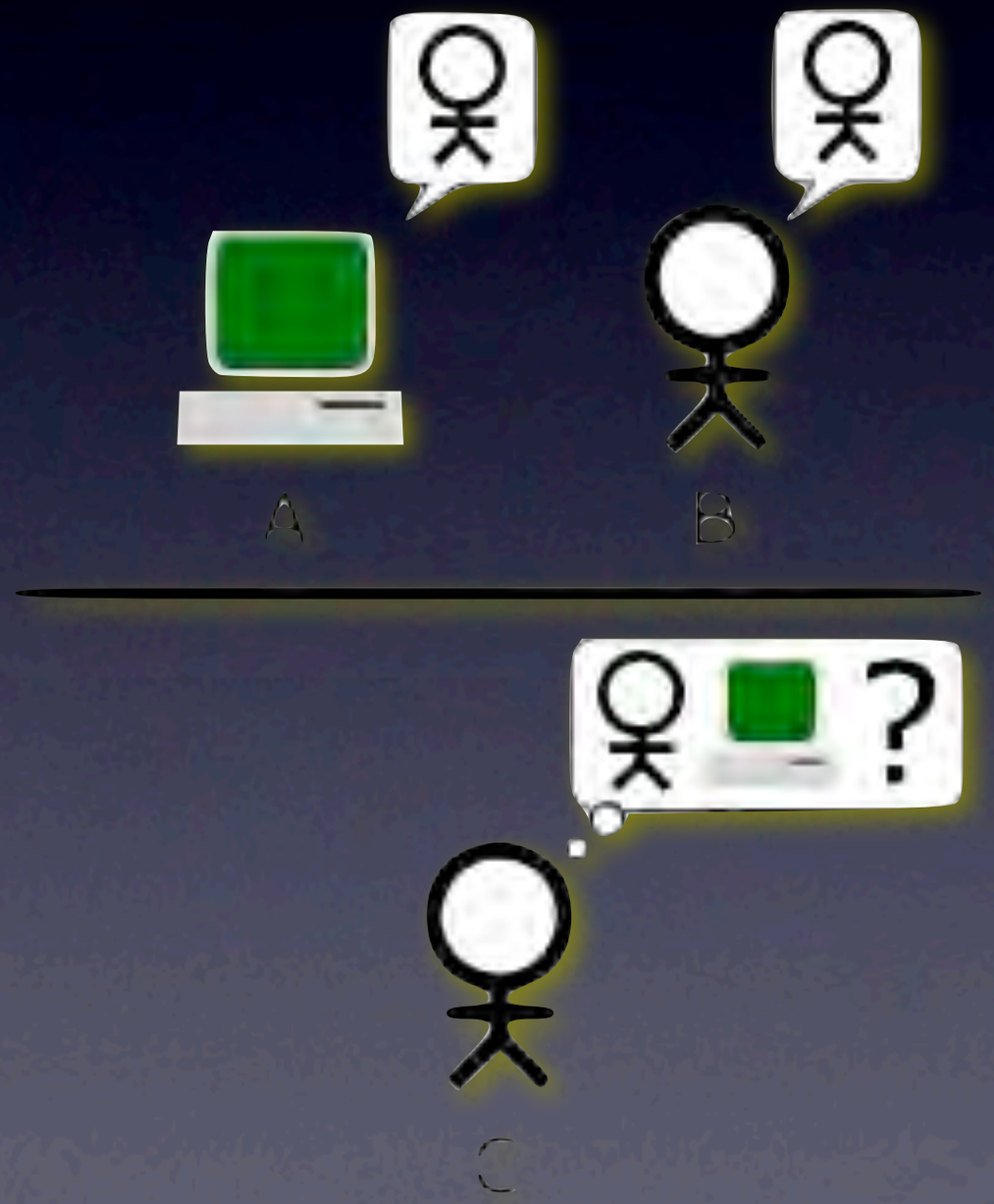
Fig. 2.
1 0 1
1 0 1 1 0 1 1 0 1 1 0
1 0 1 1 0 1 1 0 1 1 0 1



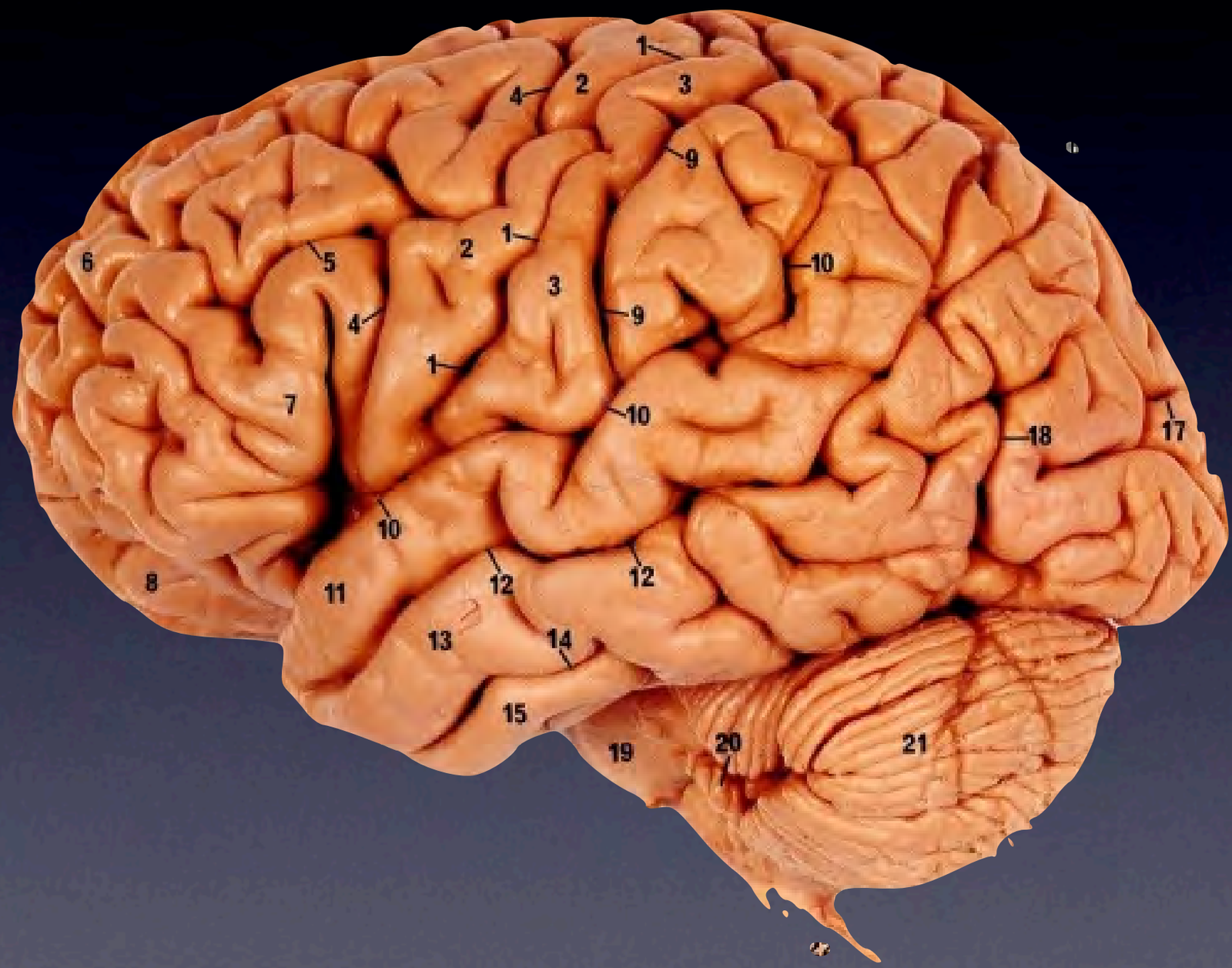

Mainframe

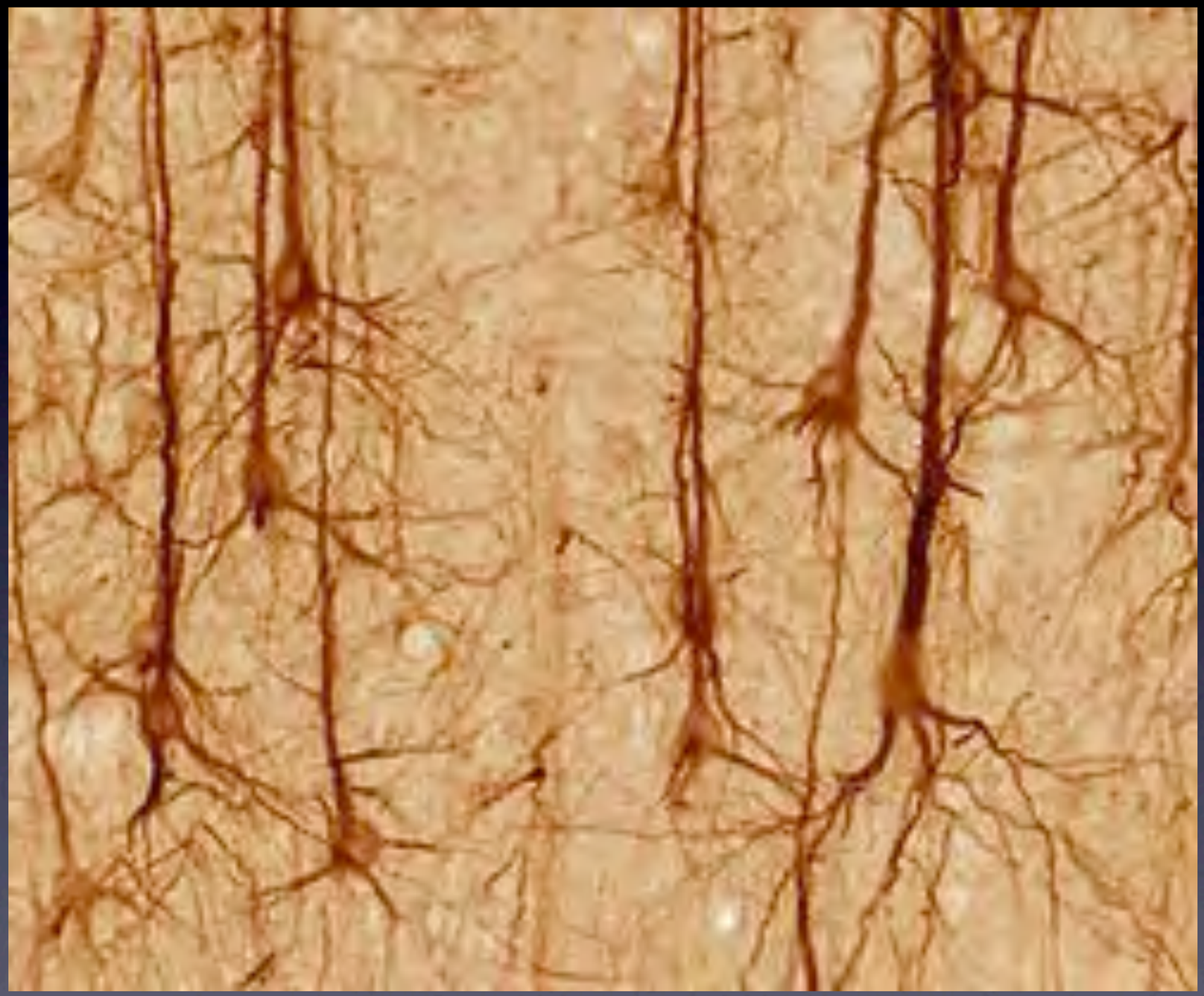
What is intelligence?

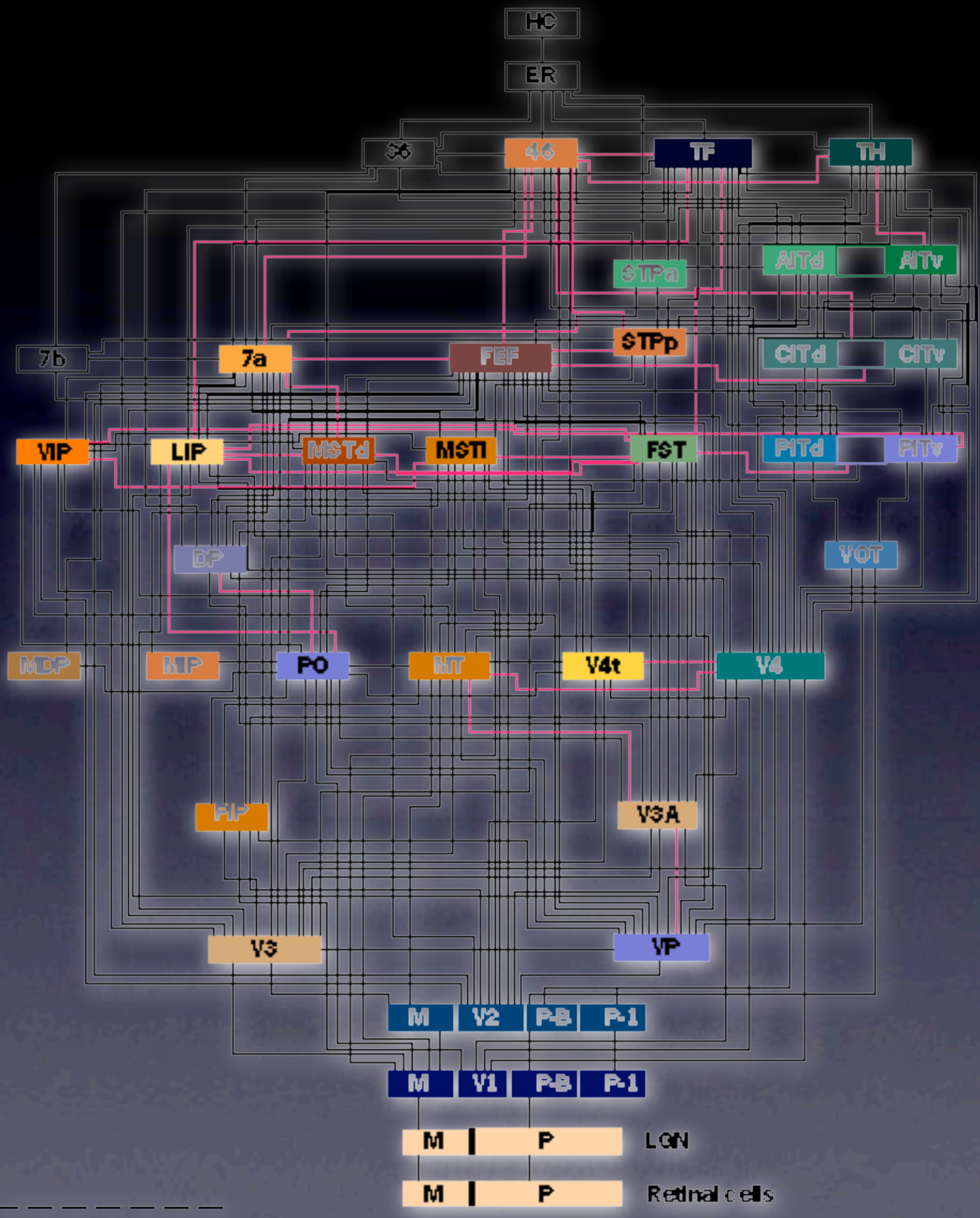
The Turing Test

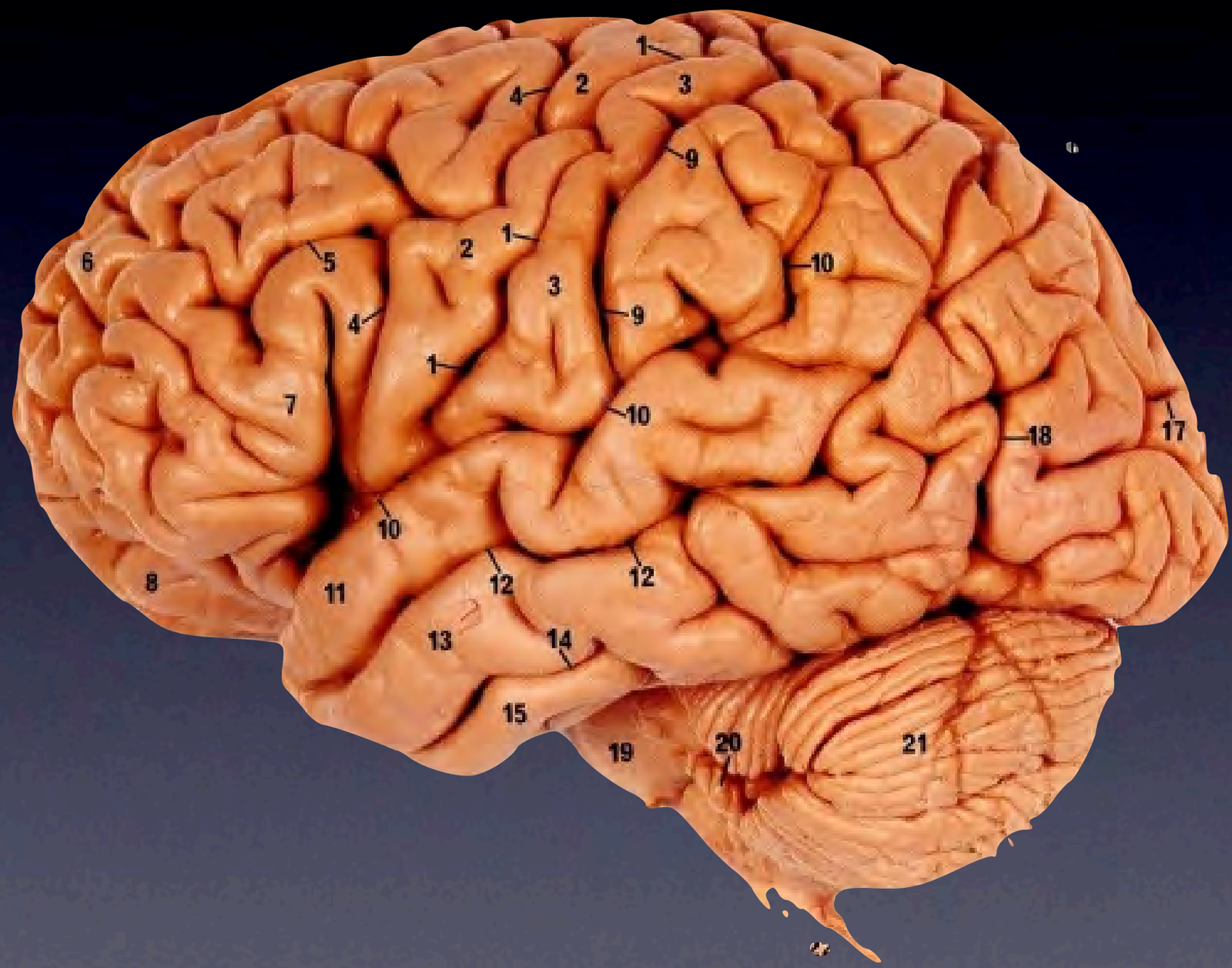


"On the Internet, nobody knows you're a dog."


1
2
3
4
9
1
2
3
4
5
6
9
10
1
7
10
18
17
8
10
11
12
12
13
14
15
19
20
21

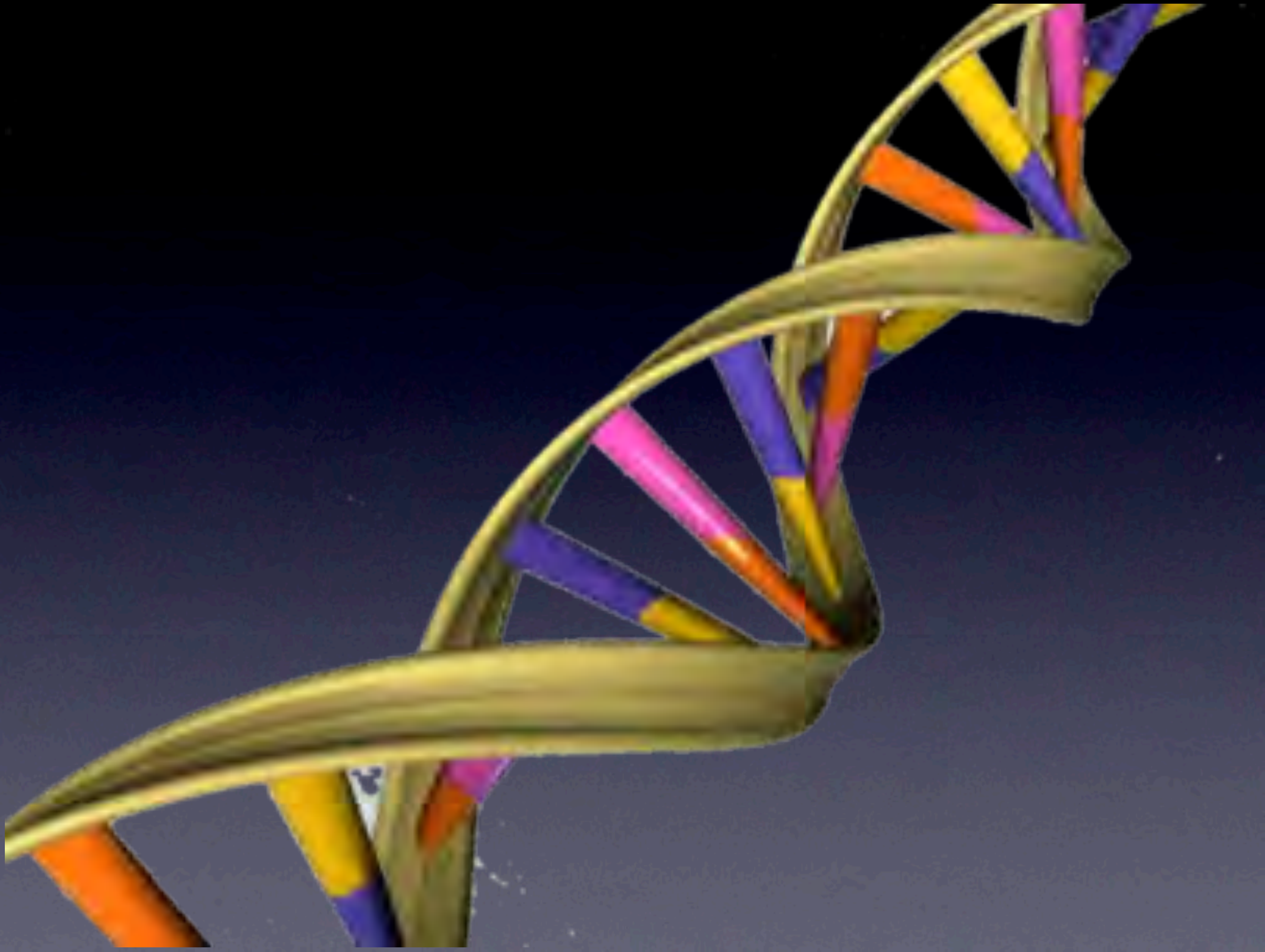



HC
ER
36
46
TF
TH
STPa
AITd
AITv
7b
7a
FEF
STPp
CITd
CITv
VIP
LIP
MSTd
MSTl
FST
PITd
PITv
DP
VOT
MDP
MIP
PO
MT
V4t
V4
PIP
V3A
V3
VP
M
V2
P-B
P-I
M
V1
P-B
P-I
M
P
LGN
M
P
Retinal cells


1
2
3
4
9
1
2
3
4
5
6
9
10
1
7
10
18
17
8
10
11
12
12
13
14
15
19
20
21

I O

yes no



=


USB

400 MB/h x 16 h/day x 365 days/year x 80 years

=

200,000 GB

400 MB/h x 16 h/day x 365 days/year x 80 years

=

200 TB

400 MB/h x 16 h/day x 365 days/year x 80 years

x 1% =

2 TB ?


1
2
3
4
9
6
5
2
1
10
3
4
9
1
10
7
18
17
10
8
11
12
12
13
14
15
19
20
21

Theories

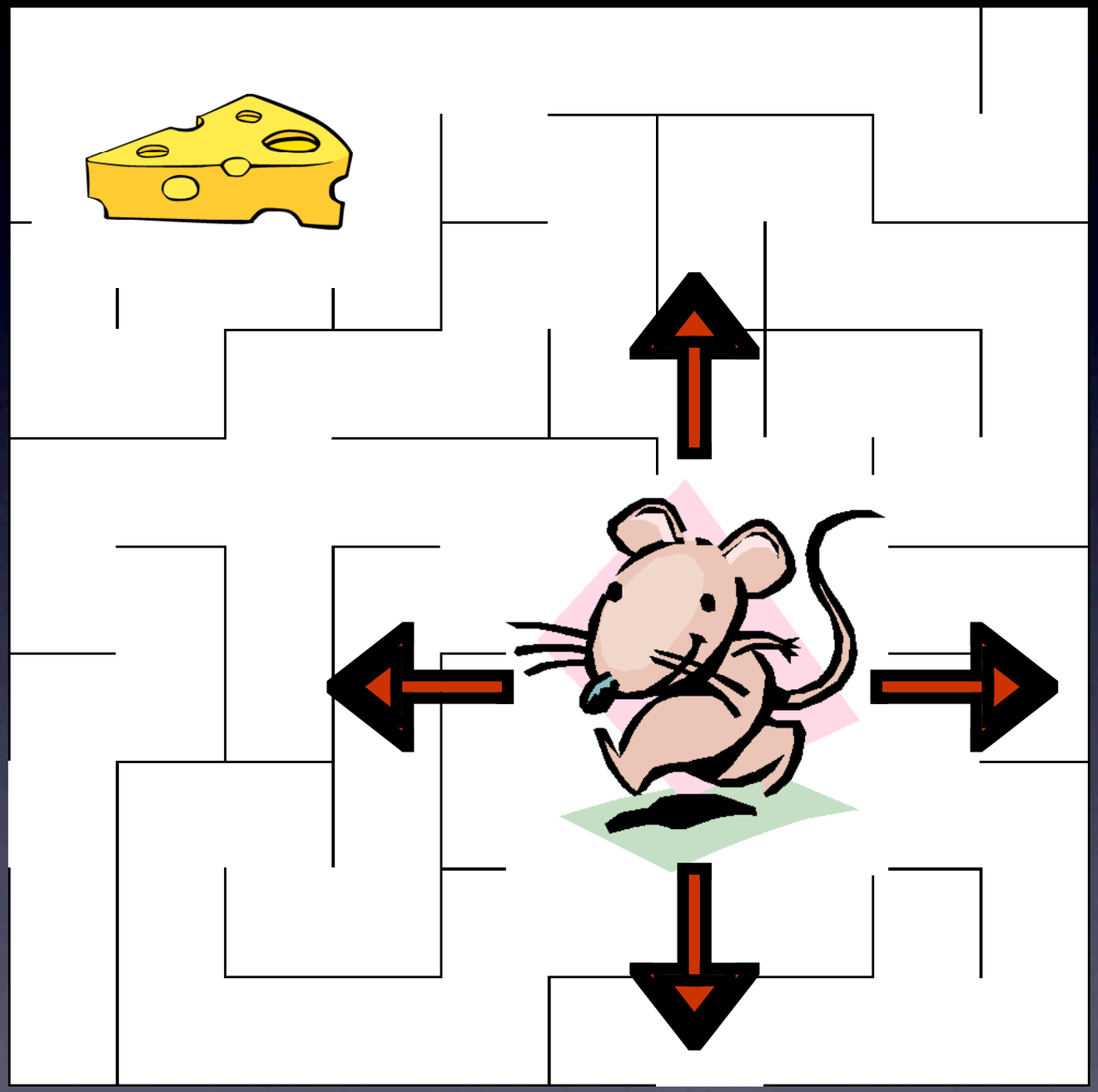
$$V^*(s) = R(s) + \max_a \gamma \sum_{s'} P(s'|s,a) V^*(s')$$

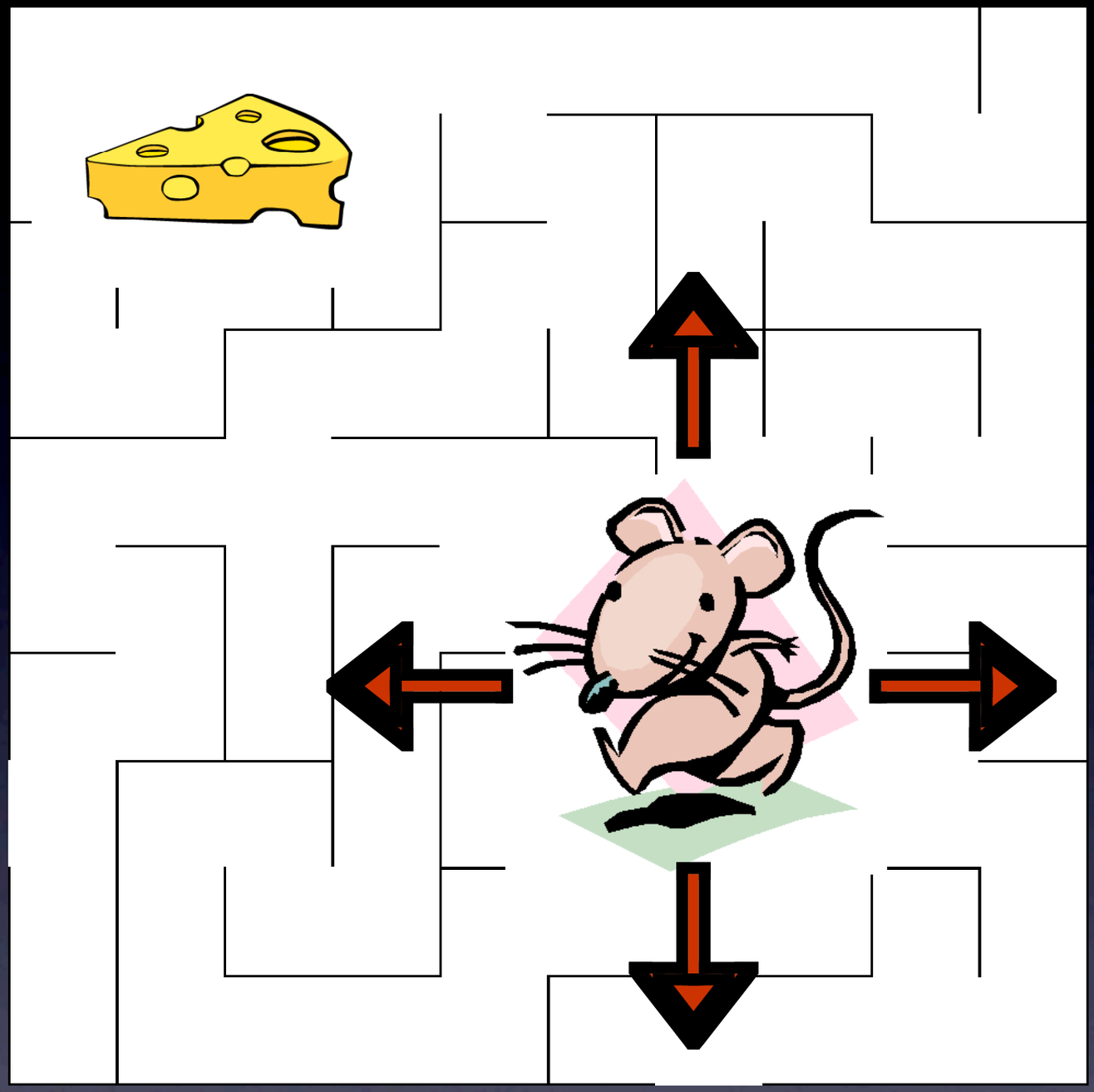
$$P(x|y) = \frac{P(y|x)P(x)}{\sum_{x'} P(y|x')P(x')}$$

$$V^*(s) = R(s) + \max_a \gamma \sum_{s'} P(s'|s,a) V^*(s')$$

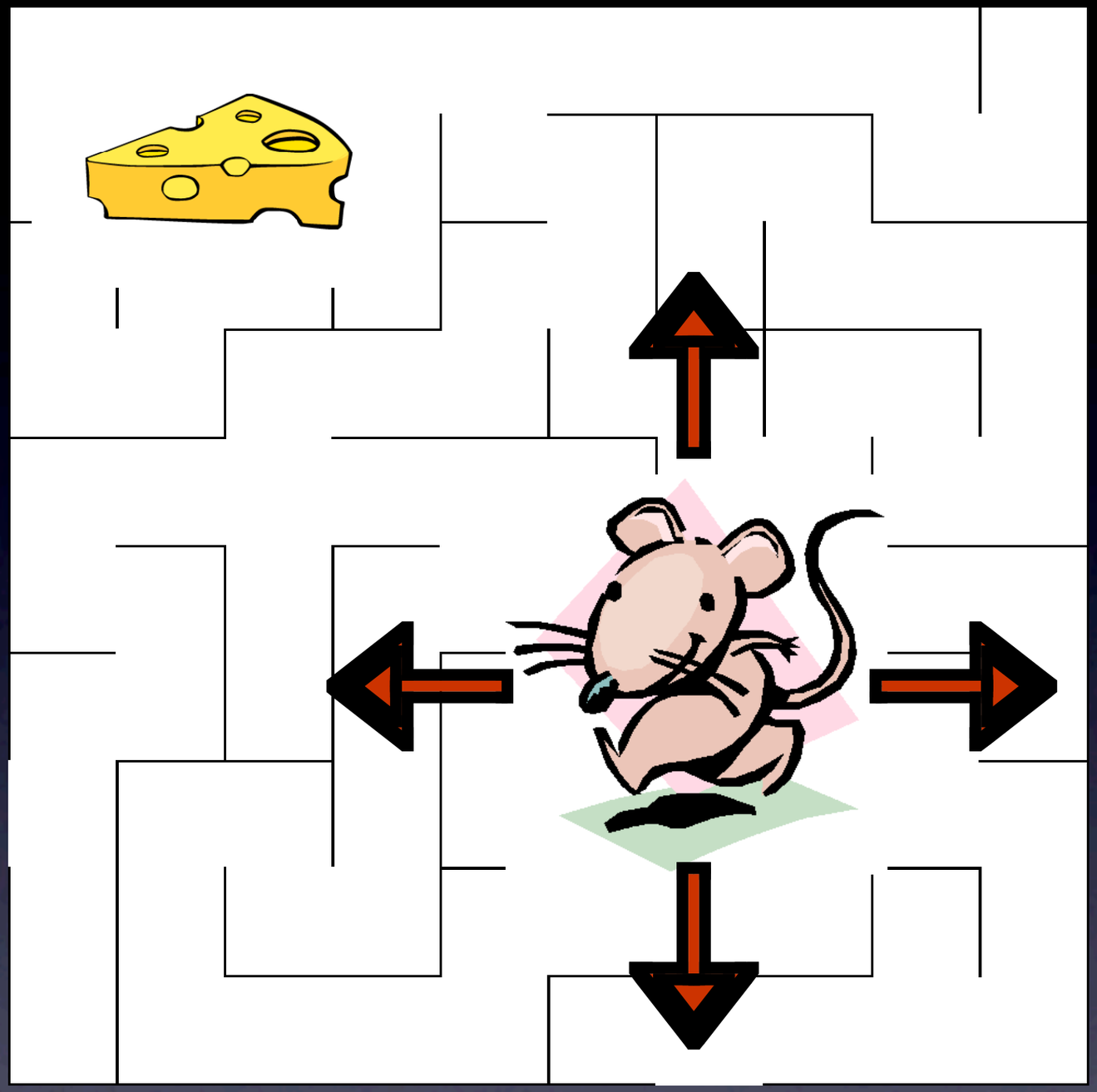
$$P(x|y) = \frac{P(y|x)P(x)}{\sum_{x'} P(y|x')P(x')}$$

Richard E. Bellman (1920-1984)

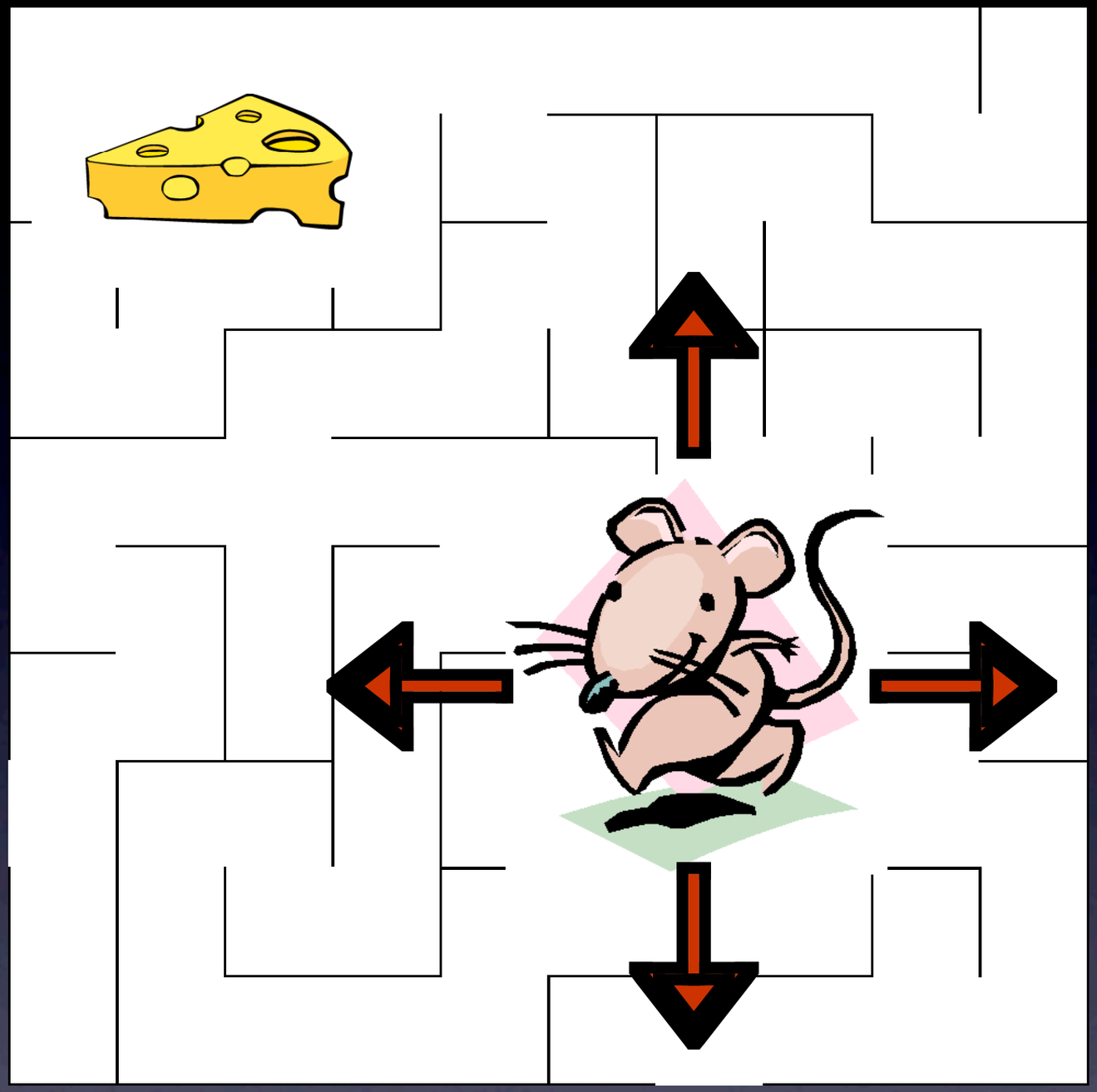




$$V^*(s) = R(s) + \max_a \gamma \sum_{s'} P(s'|s,a) V^*(s')$$

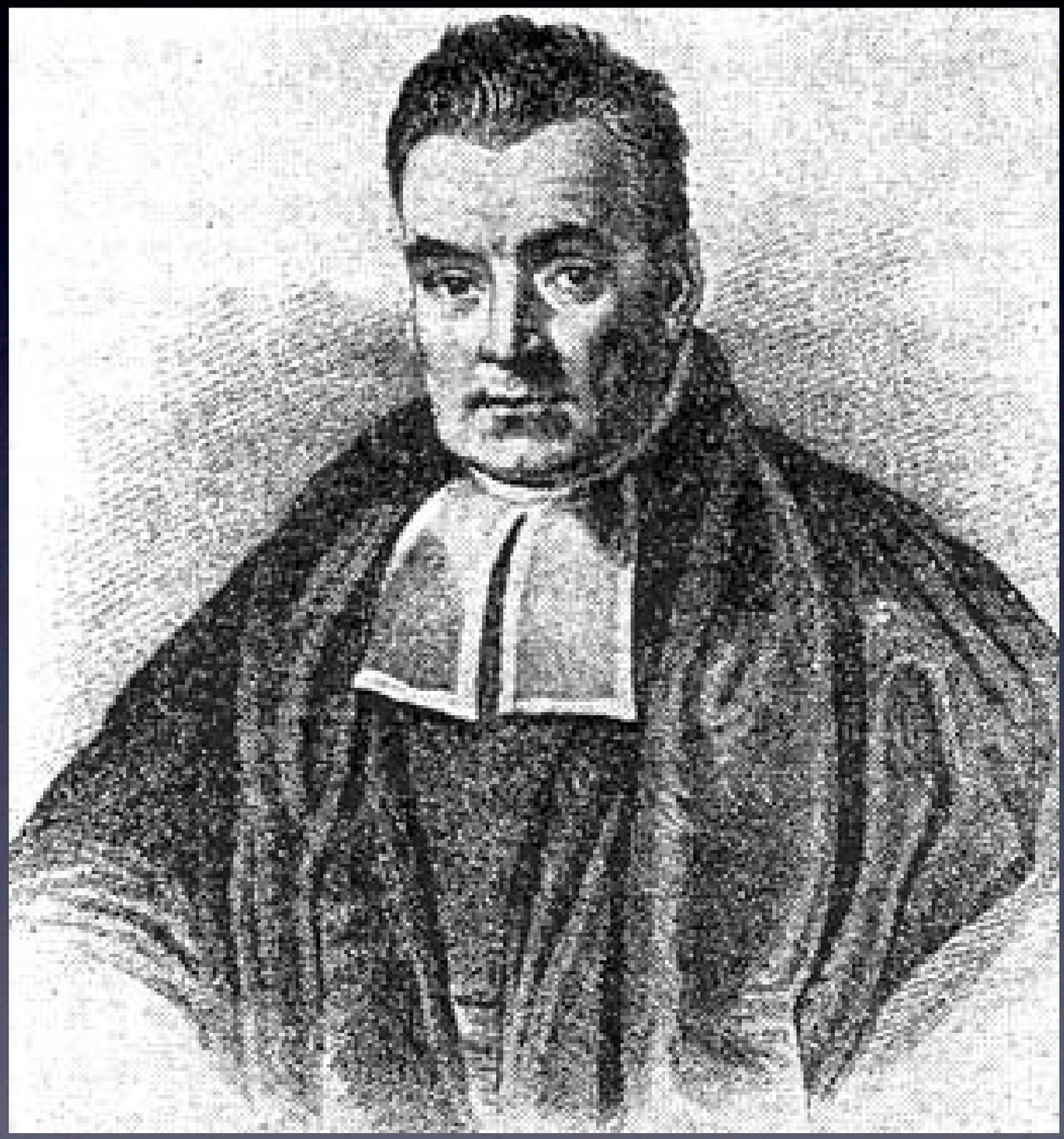


$$\text{value(state)} \quad = \quad \text{reward(state)} + \text{max}_{\text{actions}}\, \text{sum}_{\text{next states}}\, \text{prob(next state)}\, \text{value(next state)}$$

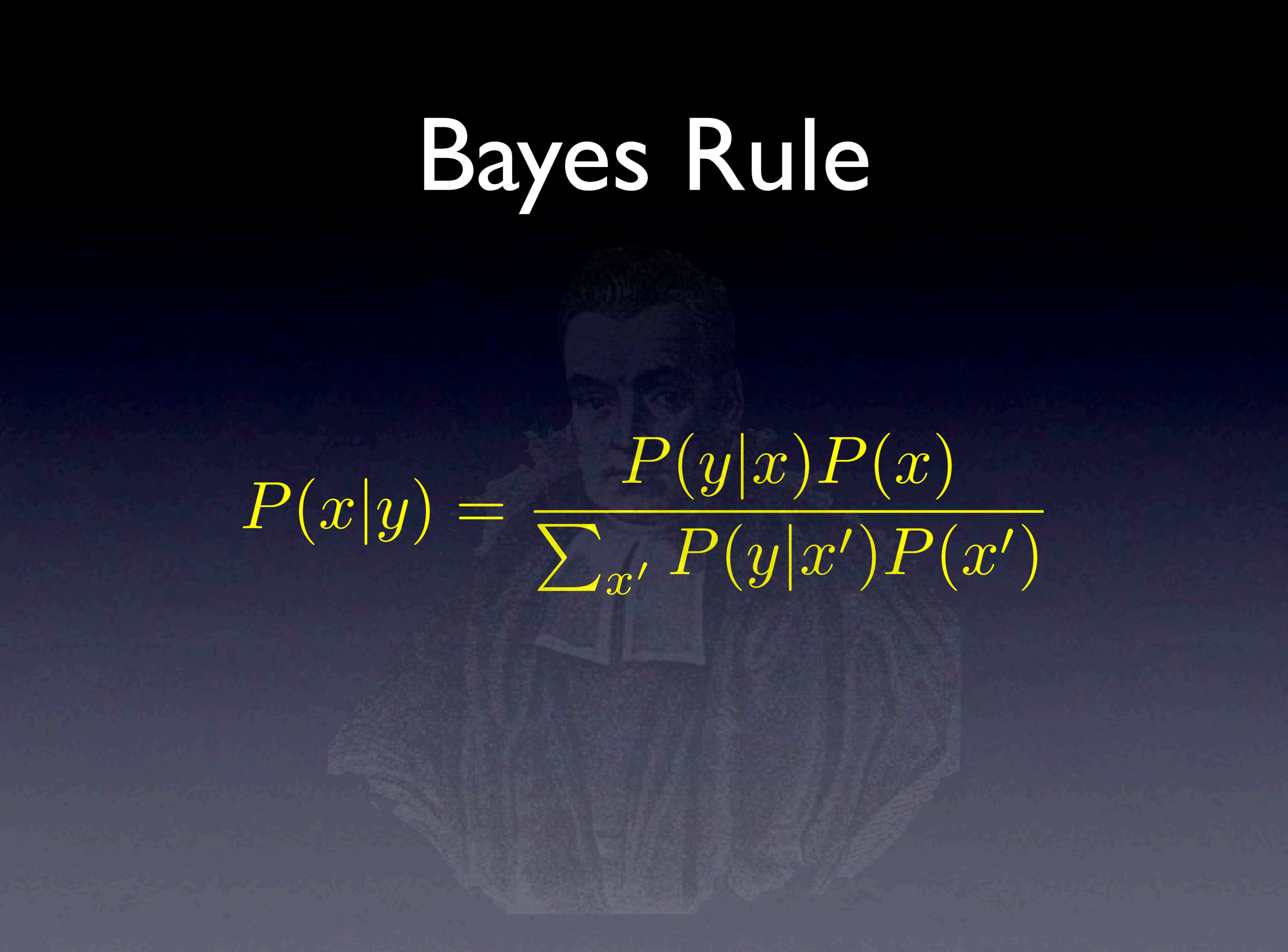


$$\text{value(state)} \quad = \quad \text{reward(state)} + \text{max}_{\text{actions}}\, \text{sum}_{\text{next states}}\, \text{prob(next state)}\, \text{value(next state)}$$

Reverend Thomas Bayes (1702-1761)



Bayes Rule

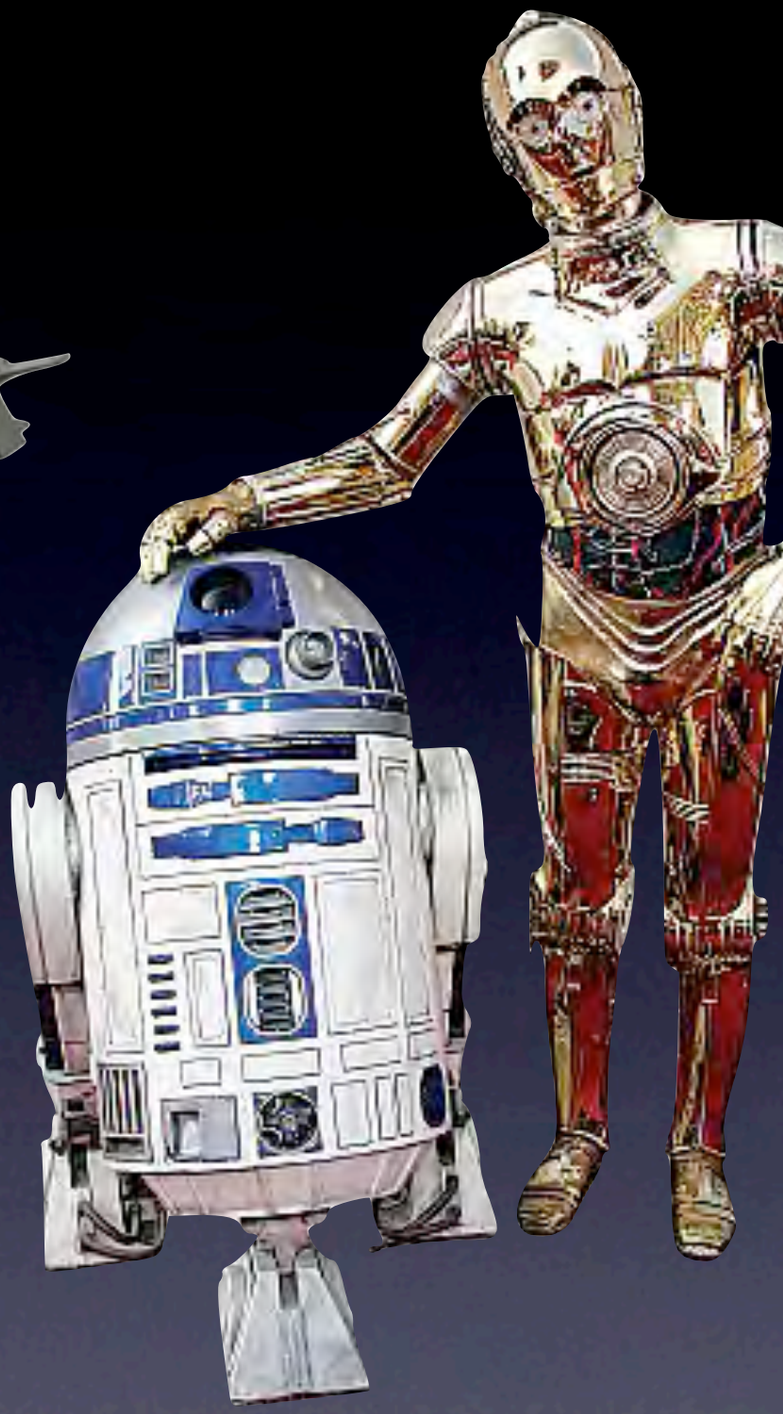


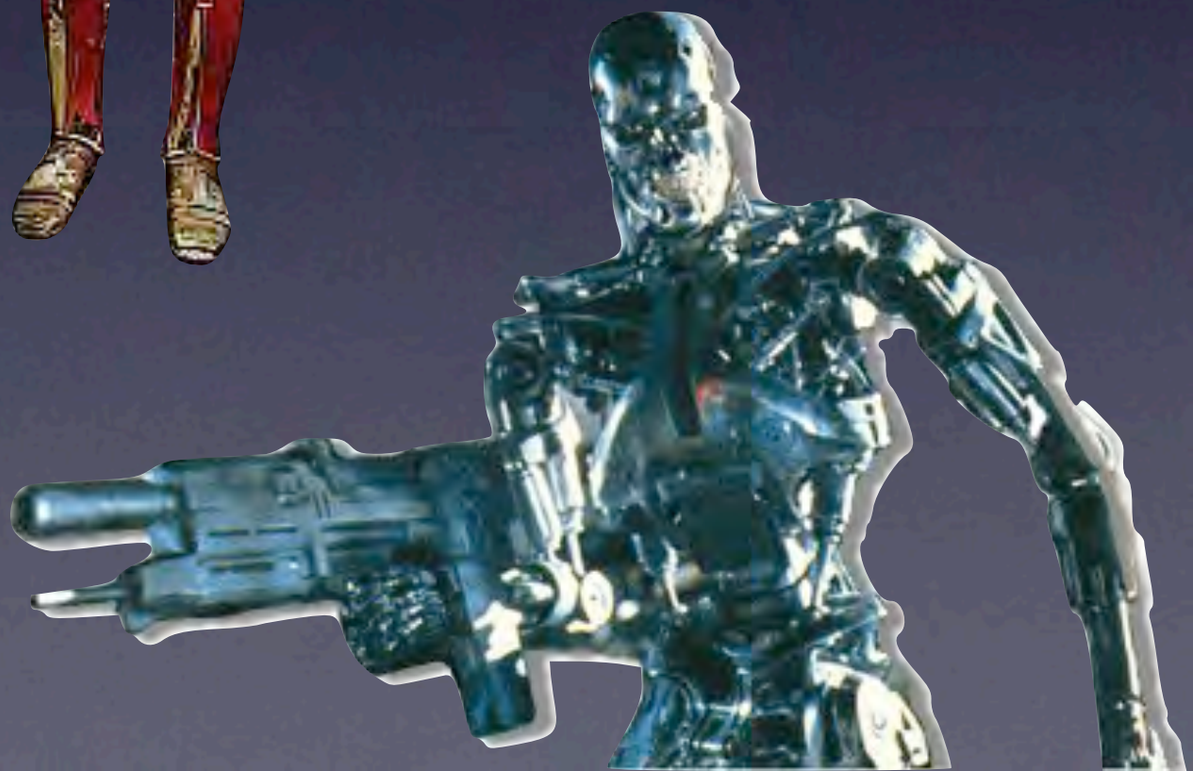
$$P(x|y) = \frac{P(y|x)P(x)}{\sum_{x'} P(y|x')P(x')}$$

Bayes Rule

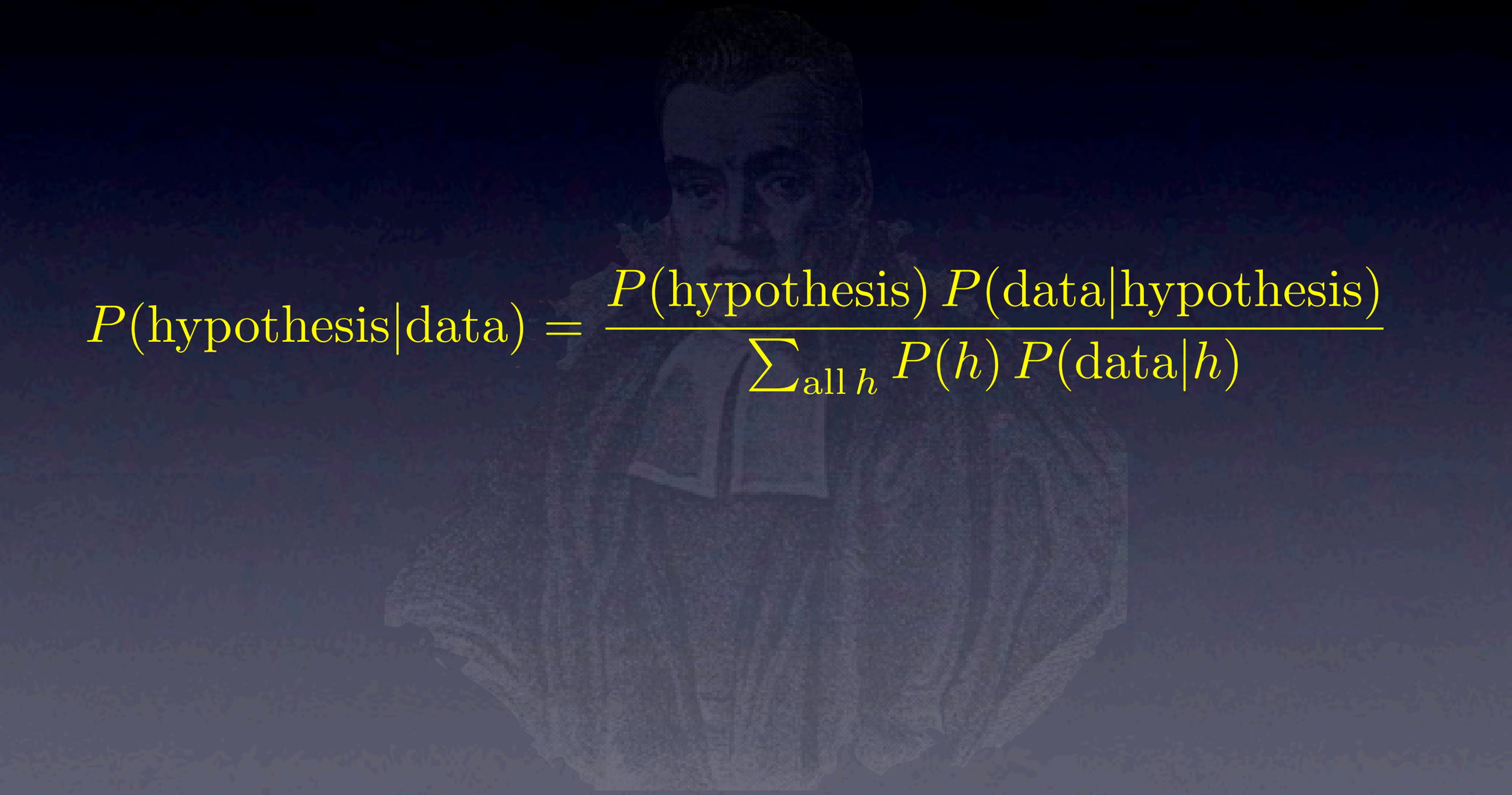
$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{hypothesis})\,P(\text{data}|\text{hypothesis})}{\sum_{\text{all } h} P(h)\,P(\text{data}|h)}$$


WALL·E



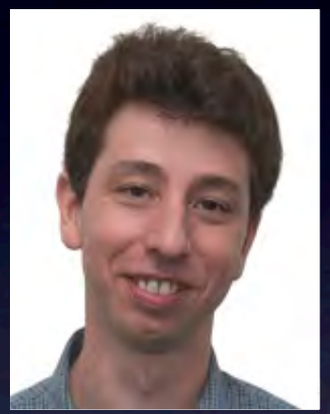


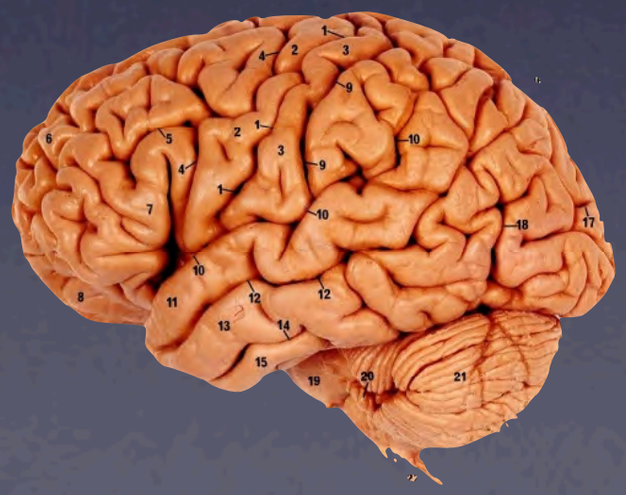
Bayes Rule

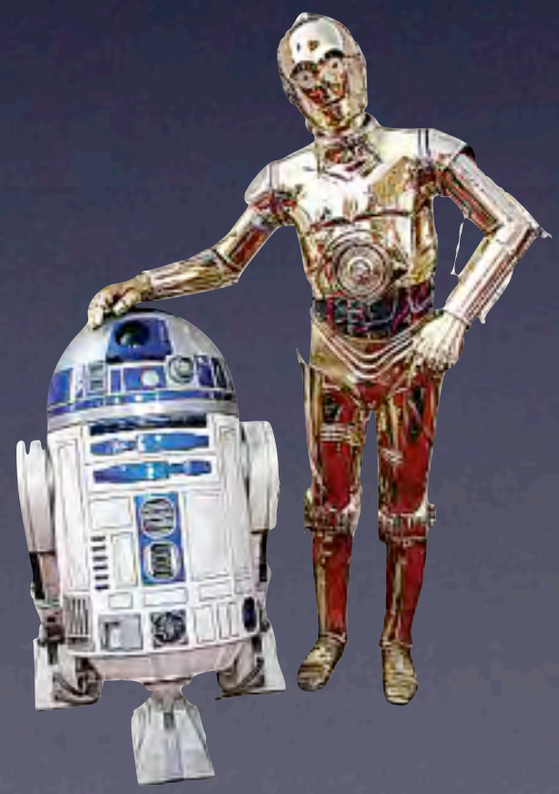


$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{hypothesis})\,P(\text{data}|\text{hypothesis})}{\sum_{\text{all } h} P(h)\,P(\text{data}|h)}$$

Computational and Biological Learning
Department of Engineering
University of Cambridge

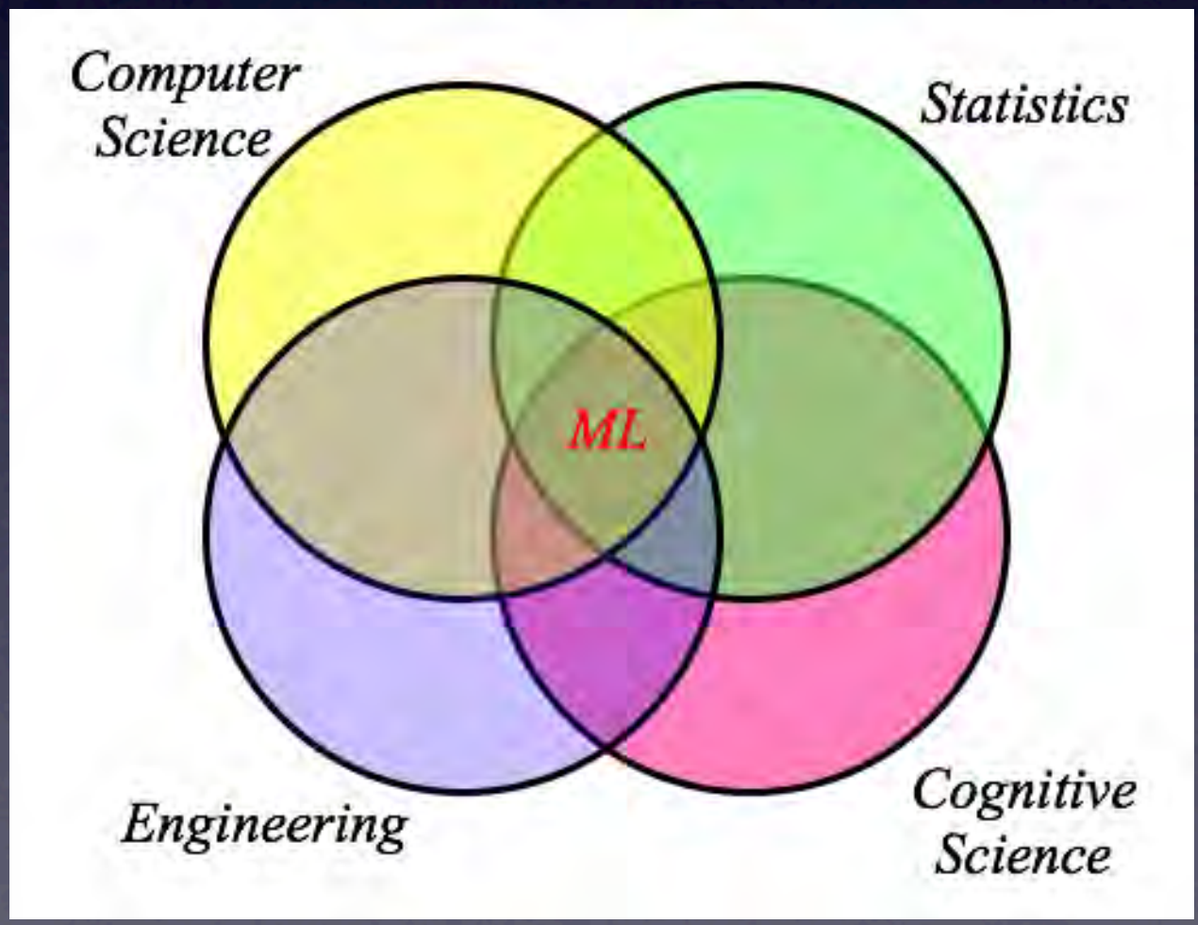







Machine Learning

Machine learning is an interdisciplinary field focusing on both the mathematical foundations and practical applications of systems that learn, reason and act.



An Information Revolution

- We are in an era of abundant data
  - **Society:** the web, social networks, mobile networks, government, digital archives
  - **Science:** large-scale scientific experiments, biomedical data, climate data, scientific literature
  - **Business:** e-commerce, electronic trading, advertising, personalisation

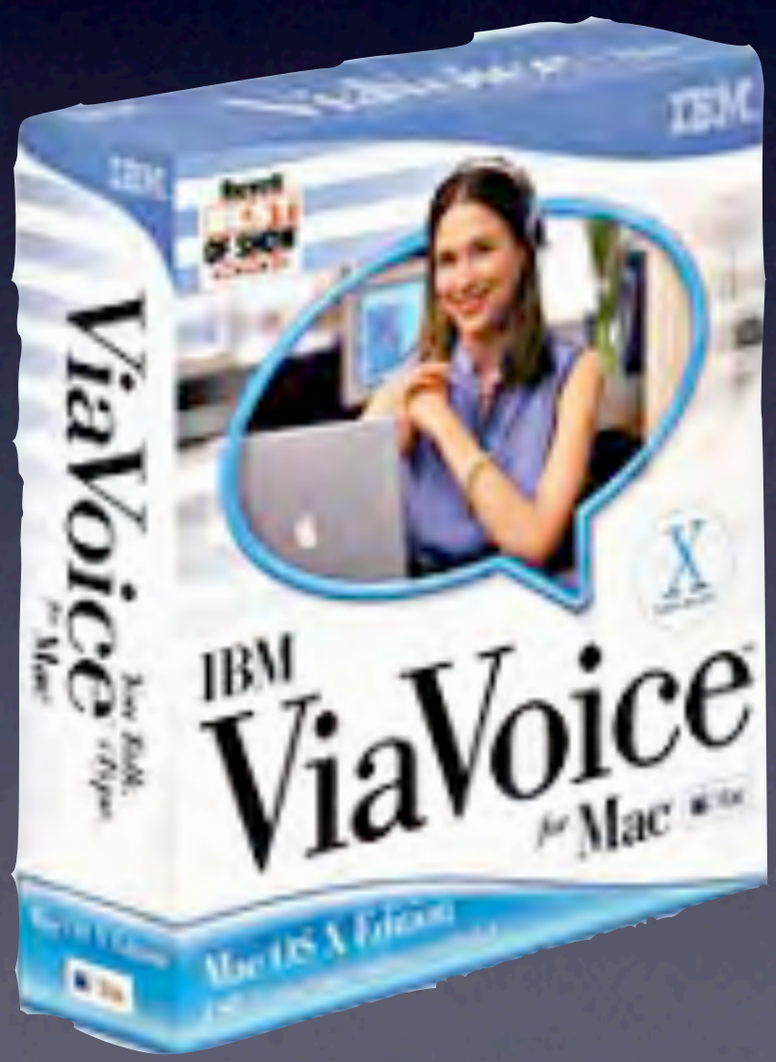
"Big Data"

Last year, there were 600 million people on Facebook who shared 360 billion items. There were 152 million blogs. 175 million Twitter users sent 25 billion tweets. "Big Data" is the latest in a long series of buzz words dealing with the vast volumes of information sweeping into the business world via the internet and other sources. Enterprises globally stored more than 7 exabytes of new data on disk drives in 2010, while consumers stored more than 6 exabytes of new data on devices such as PCs and notebooks. An exabyte is one billion gigabytes.

"Analytics Wave" *The Times of India*, Sept 27th, 2011

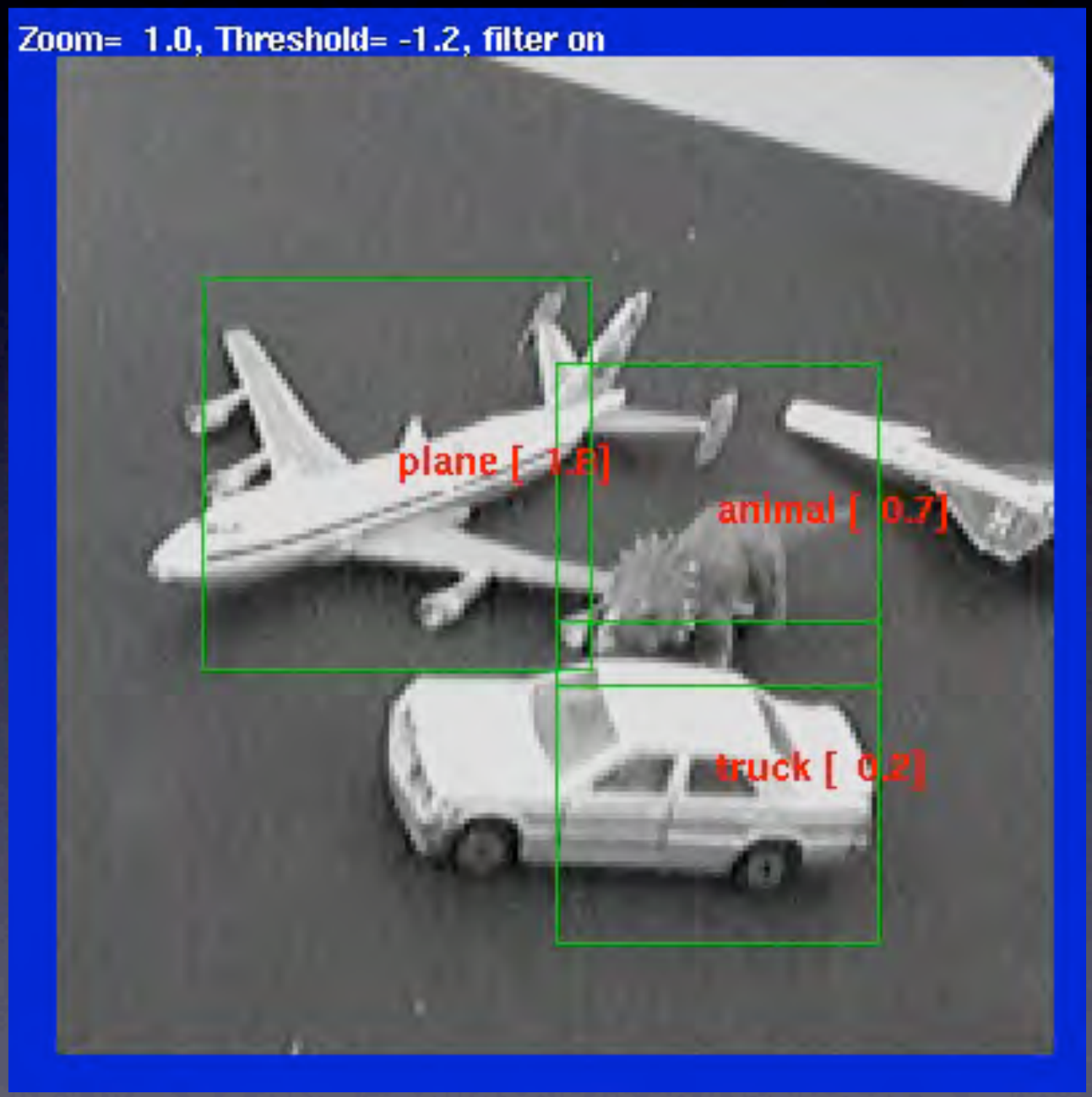
- We need tools for modelling, searching, visualising, and understanding large data sets

Automatic Speech Recognition

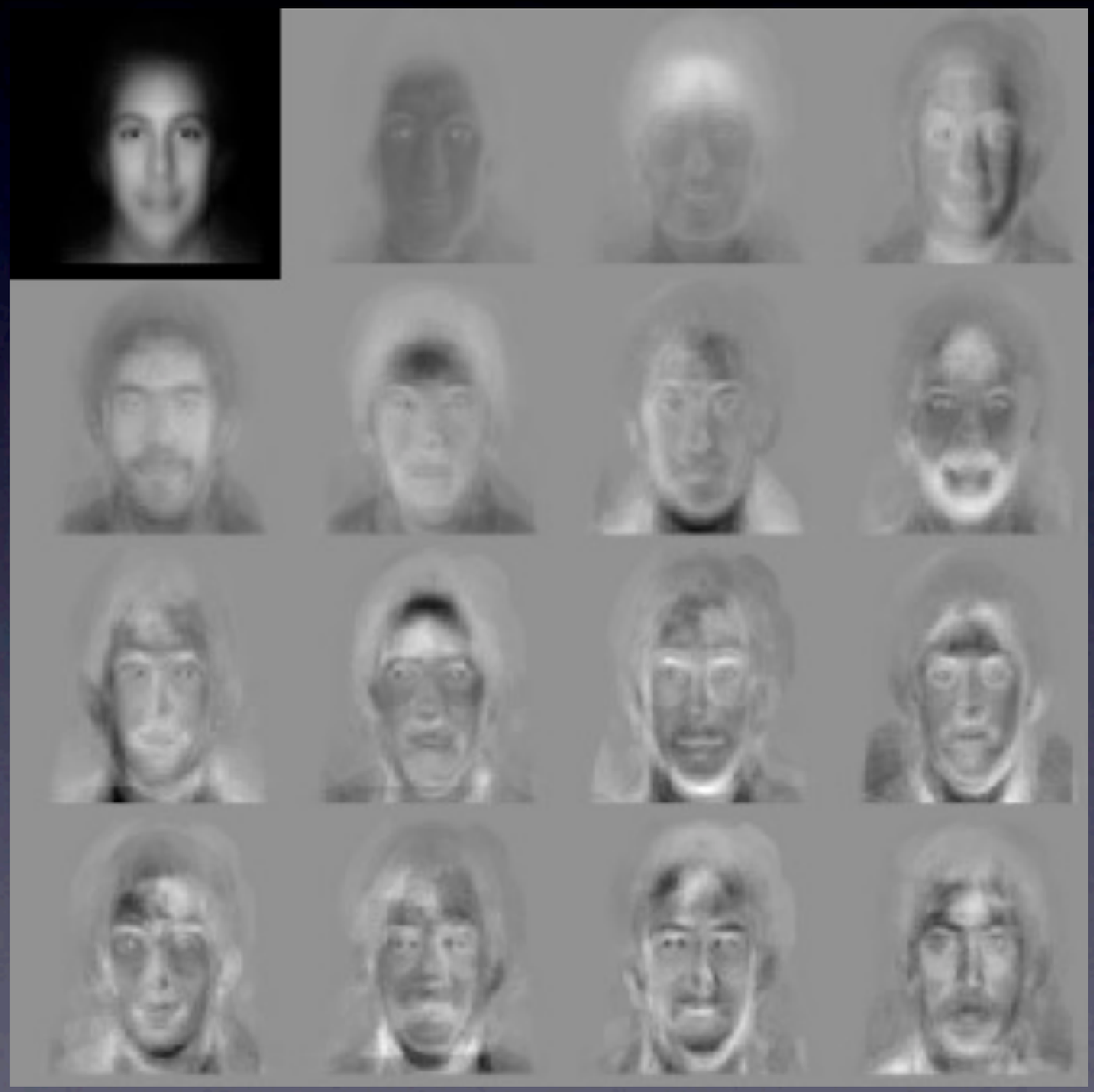


SAY WHAT
YOU WANT

Object Recognition

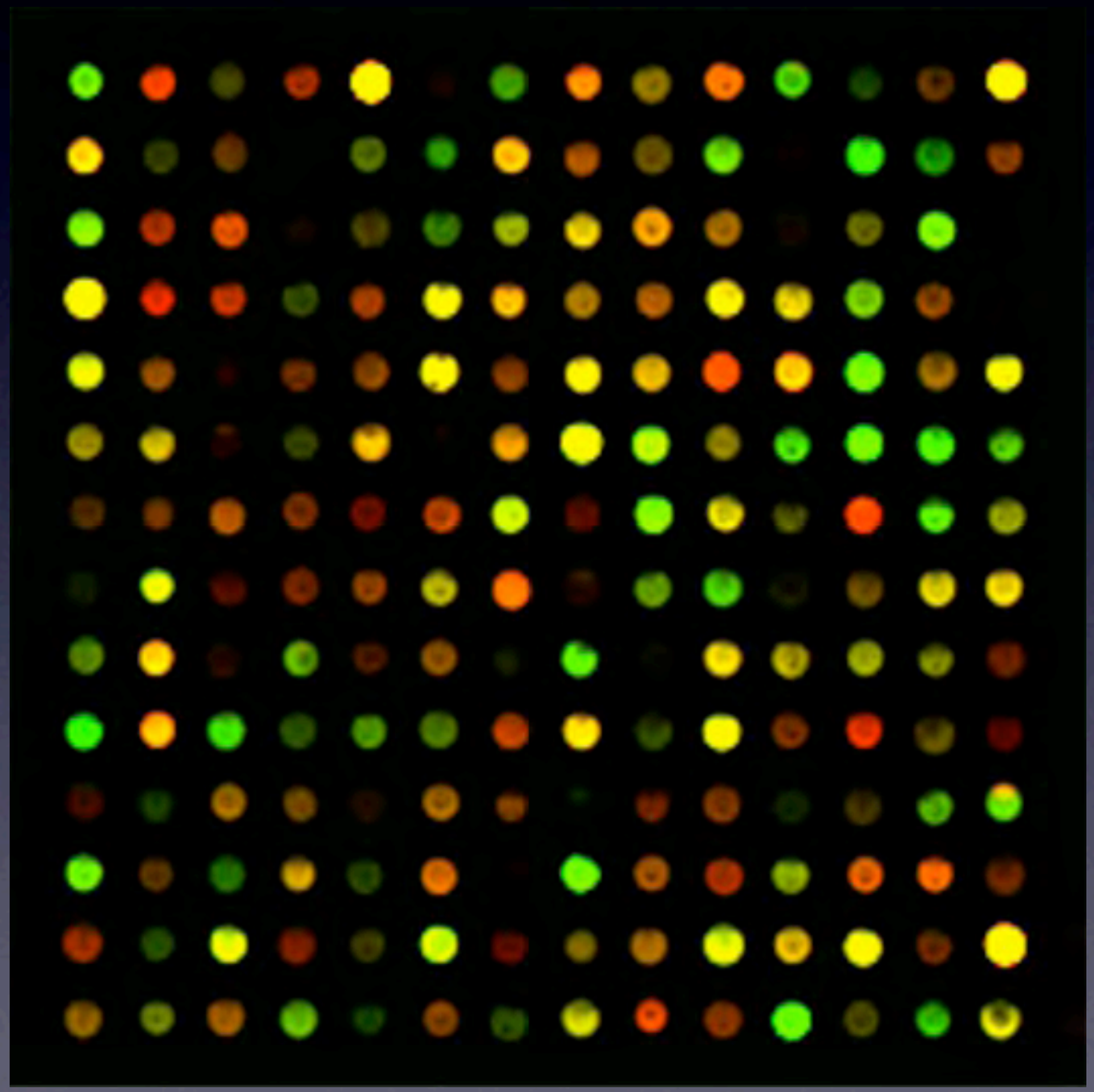


Face Detection and Recognition



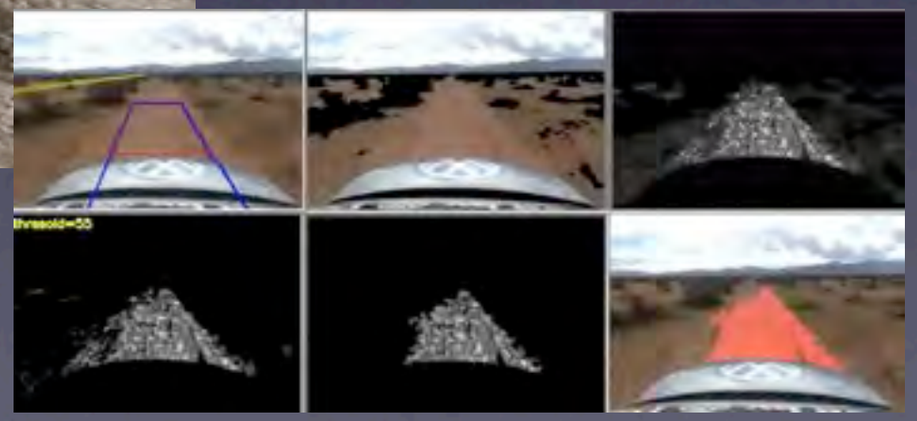
Bioinformatics

e.g. Modelling Gene Expression

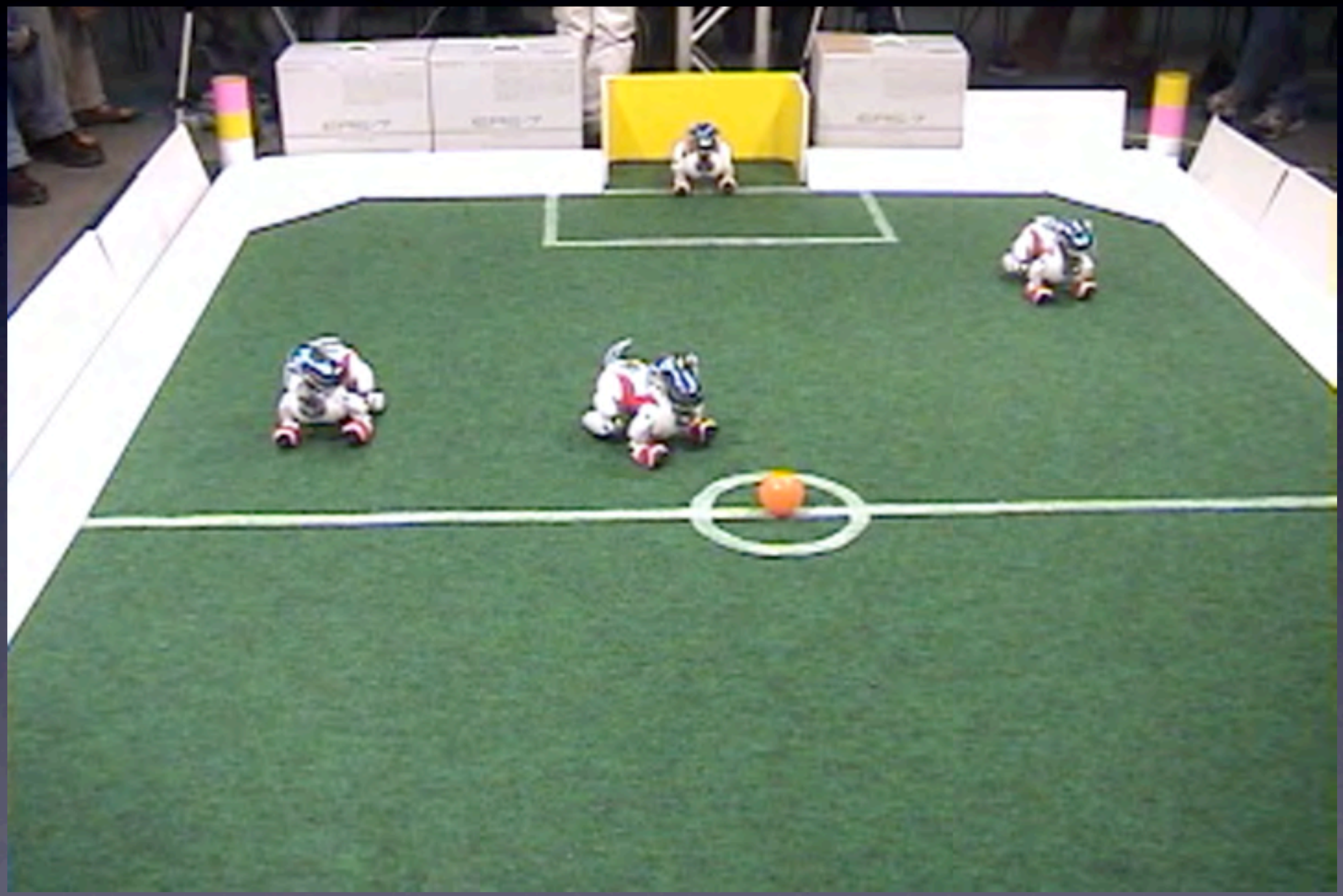


Recommending Movies, Books...

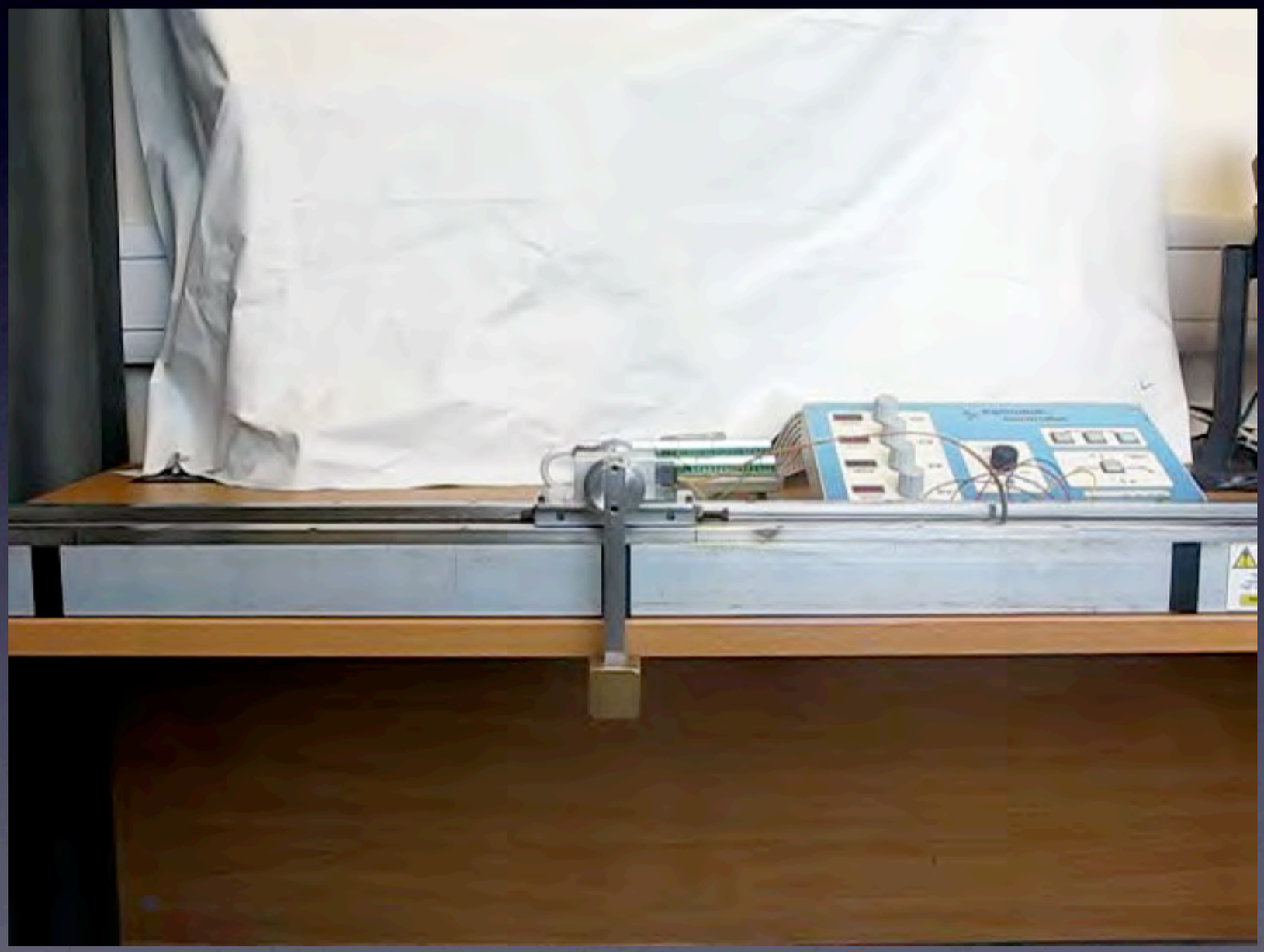
Autonomous Cars



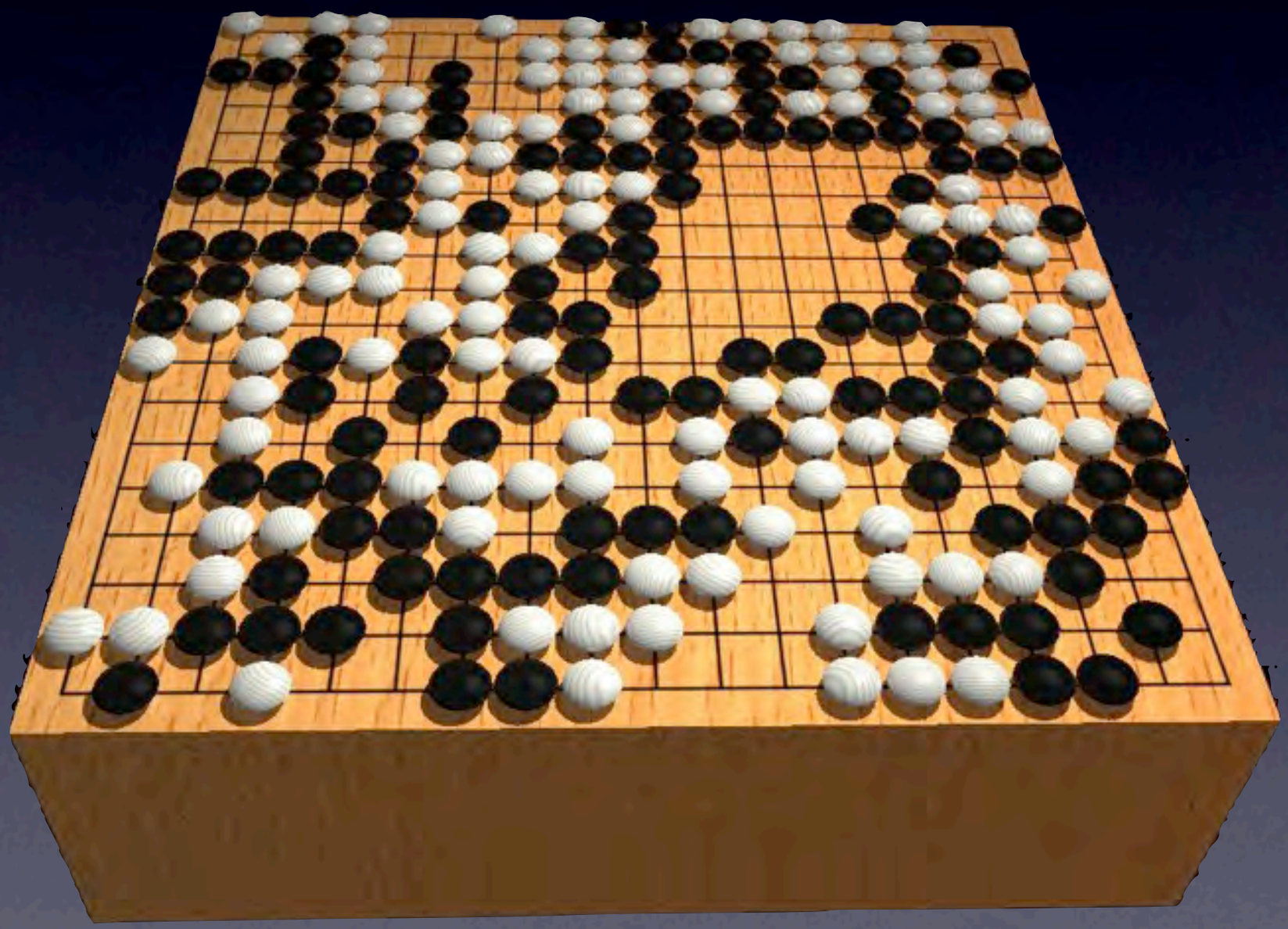
Robotic Football



Learning to balance

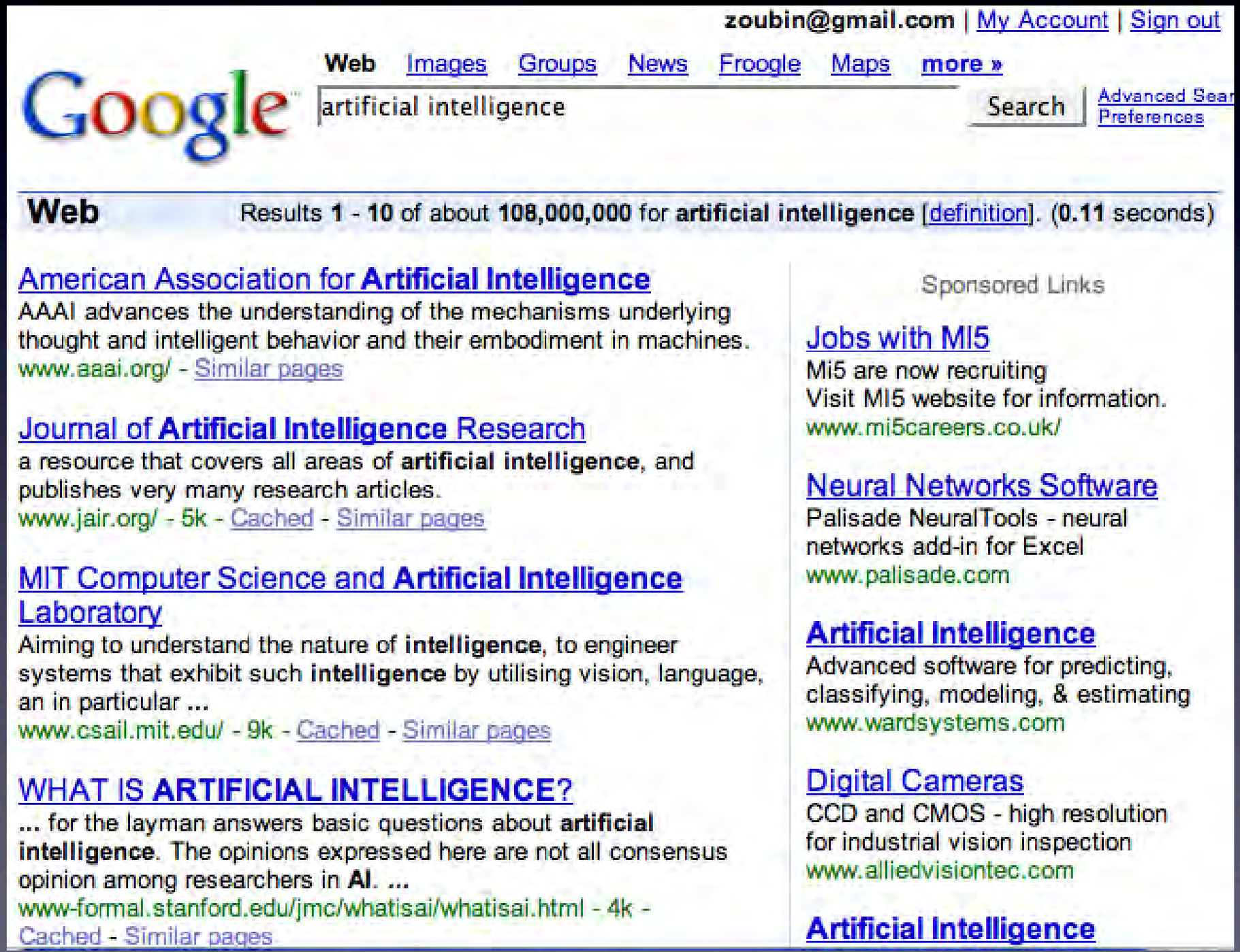


Computer Games

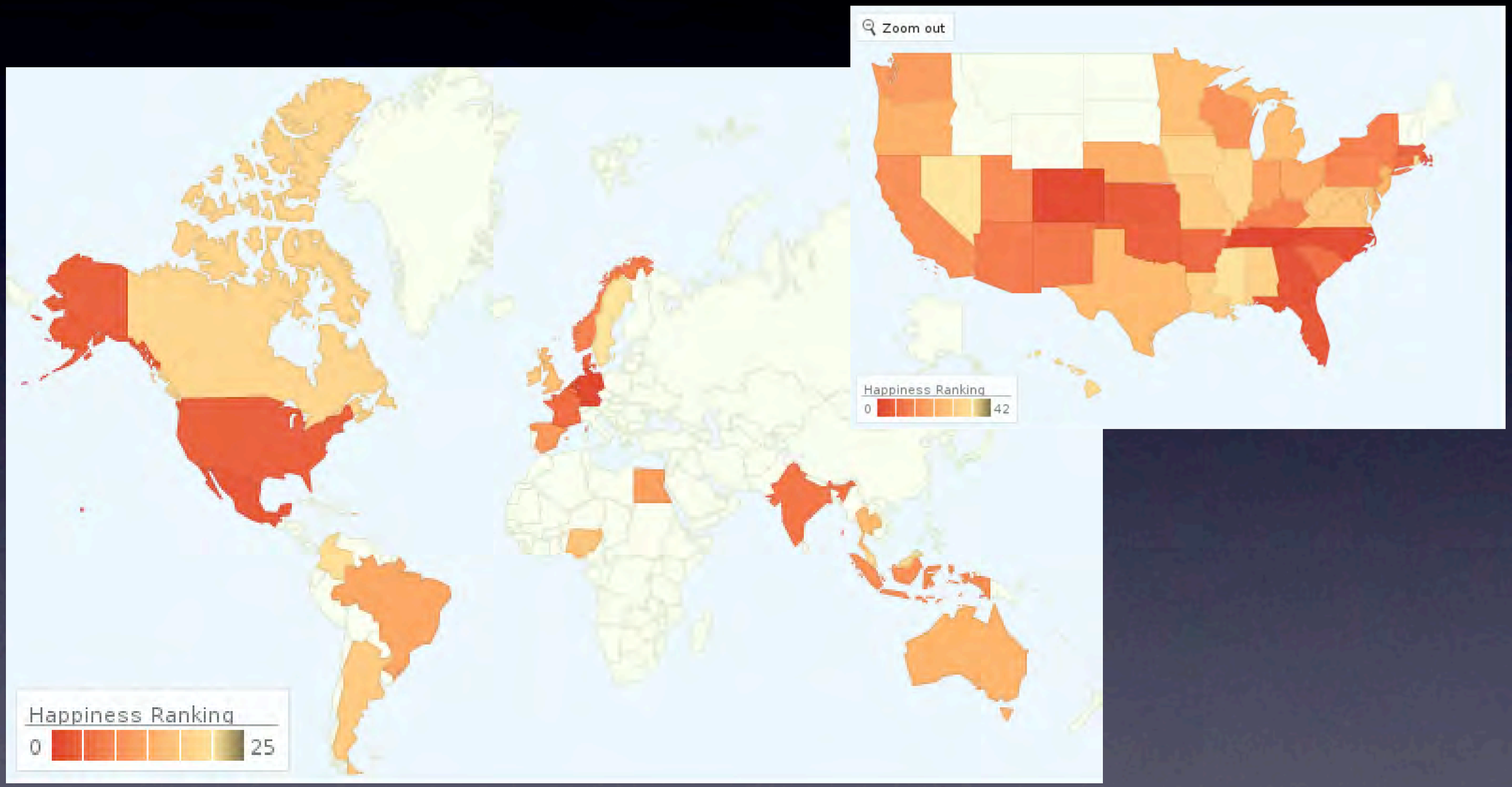


Financial Prediction

Web Search


zoubin@gmail.com | My Account | Sign out
Google
Web Images Groups News Froogle Maps more »
artificial intelligence
Search
Advanced Sear
Preferences
Web
Results 1 - 10 of about 108,000,000 for artificial intelligence [definition]. (0.11 seconds)
American Association for Artificial Intelligence
AAAI advances the understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines.
www.aaai.org/ - Similar pages
Journal of Artificial Intelligence Research
a resource that covers all areas of artificial intelligence, and publishes very many research articles.
www.jair.org/ - 5k - Cached - Similar pages
MIT Computer Science and Artificial Intelligence Laboratory
Aiming to understand the nature of intelligence, to engineer systems that exhibit such intelligence by utilising vision, language, an in particular ...
www.csail.mit.edu/ - 9k - Cached - Similar pages
WHAT IS ARTIFICIAL INTELLIGENCE?
... for the layman answers basic questions about artificial intelligence. The opinions expressed here are not all consensus opinion among researchers in AI. ...
www-formal.stanford.edu/jmc/whatisai/whatisai.html - 4k - Cached - Similar pages
Sponsored Links
Jobs with MI5
Mi5 are now recruiting
Visit MI5 website for information.
www.mi5careers.co.uk/
Neural Networks Software
Palisade NeuralTools - neural networks add-in for Excel
www.palisade.com
Artificial Intelligence
Advanced software for predicting, classifying, modeling, & estimating
www.wardsystems.com
Digital Cameras
CCD and CMOS - high resolution for industrial vision inspection
www.alliedvisiontec.com
Artificial Intelligence

Twitter Sentiment Modelling



Map of the world with countries ordered by relative rate of happy to sad tweets

Bayesian Nonparametric Machine Learning

Bayesian Machine Learning

Everything follows from two simple rules:

Sum rule: $P(x) = \sum_y P(x, y)$

Product rule: $P(x, y) = P(x)P(y|x)$

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

$P(\mathcal{D}\|\theta)$	likelihood of θ
$P(\theta)$	prior probability of θ
$P(\theta\|\mathcal{D})$	posterior of θ given $\mathcal{D}$

Prediction:

$$P(x|\mathcal{D}, m) = \int P(x|\theta, \mathcal{D}, m)P(\theta|\mathcal{D}, m)d\theta$$

Model Comparison:

$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}$$

$$P(\mathcal{D}|m) = \int P(\mathcal{D}|\theta, m)P(\theta|m)\, d\theta$$

Parametric vs Nonparametric Models

- *Parametric models* assume some finite set of parameters θ. Given the parameters, future predictions, x, are independent of the observed data, $\mathcal{D}$:

$$P(x|\theta, \mathcal{D}) = P(x|\theta)$$

  therefore θ capture everything there is to know about the data.

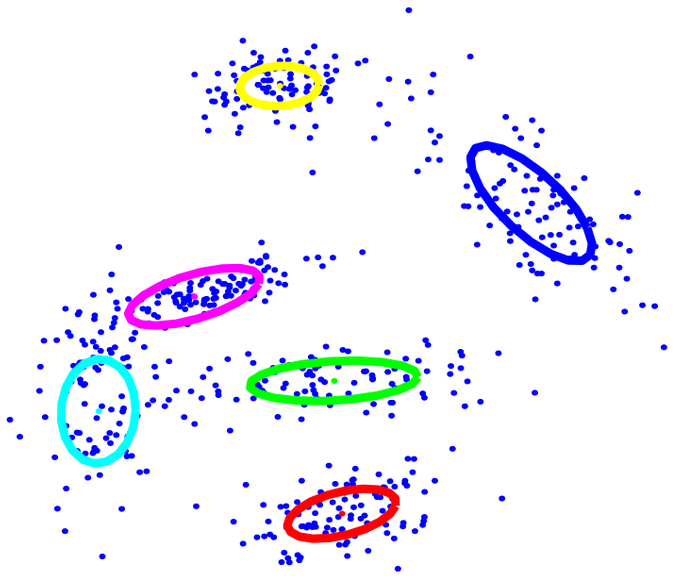
- So the complexity of the model is bounded even if the amount of data is unbounded. This makes them not very flexible.

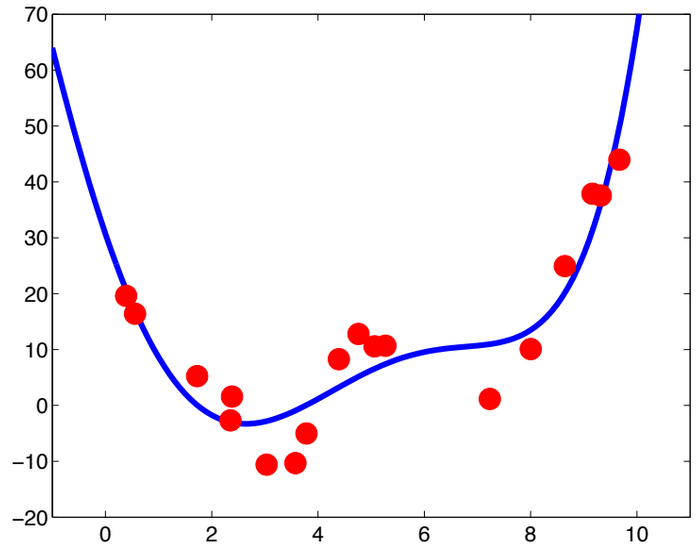
- *Non-parametric models* assume that the data distribution cannot be defined in terms of such a finite set of parameters. But they can often be defined by assuming an *infinite dimensional* θ. Usually we think of θ as a *function*.

- The amount of information that θ can capture about the data $\mathcal{D}$ can grow as the amount of data grows. This makes them more flexible.

Why nonparametrics?

- flexibility
- better predictive performance
- more realistic





Overview of nonparametric models and uses

Bayesian nonparametrics has many uses.

Some modelling goals and *examples* of associated nonparametric Bayesian models:

Modelling goal	**Example process**
Distributions on functions	Gaussian process
Distributions on distributions	Dirichlet process
	Polya Tree
Clustering	Chinese restaurant process
	Pitman-Yor process
Hierarchical clustering	Dirichlet diffusion tree
	Kingman's coalescent
Sparse binary matrices	Indian buffet processes
Survival analysis	Beta processes
Distributions on measures	Completely random measures
...	...

Gaussian and Dirichlet Processes

- Gaussian processes define a distribution on functions

$$f \sim \mathsf{GP}(\cdot|\mu, c)$$

where μ is the mean function and c is the covariance function.
We can think of GPs as "infinite-dimensional" Gaussians

- Dirichlet processes define a distribution on distributions (a measure on measures)

$$G \sim \mathsf{DP}(\cdot|G_0, \alpha)$$

where $\alpha > 0$ is a scaling parameter, and G_0 is the base measure.
We can think of DPs as "infinite-dimensional" Dirichlet distributions.

Note that both f and G are infinite dimensional objects.

Bayesian nonparametrics applied to models of other structured objects:

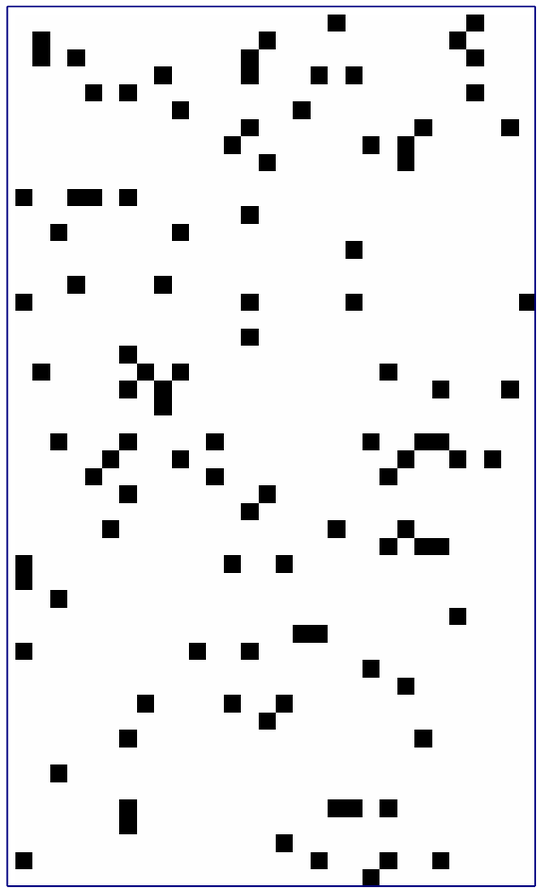
- Sparse Matrices
- Deep Sparse Graphical Models
- Hierarchies
- Covariances
- Network Structured Regression

From finite to infinite binary matrices

$z_{nk} = 1$ means object n has feature k:

$$z_{nk} \sim \text{Bernoulli}(\theta_k)$$

$$\theta_k \sim \text{Beta}(\alpha/K, 1)$$



- Note that $P(z_{nk} = 1|\alpha) = E(\theta_k) = \frac{\alpha/K}{\alpha/K+1}$, so as K grows larger the matrix gets sparser.
- So if $\mathbf{Z}$ is $N \times K$, the expected number of nonzero entries is $N\alpha/(1 + \alpha/K) < N\alpha$.
- Even in the $K \to \infty$ limit, the matrix is expected to have a finite number of non-zero entries.

Nonparametric Binary Matrix Factorization

genes $\times$ patients

users $\times$ movies

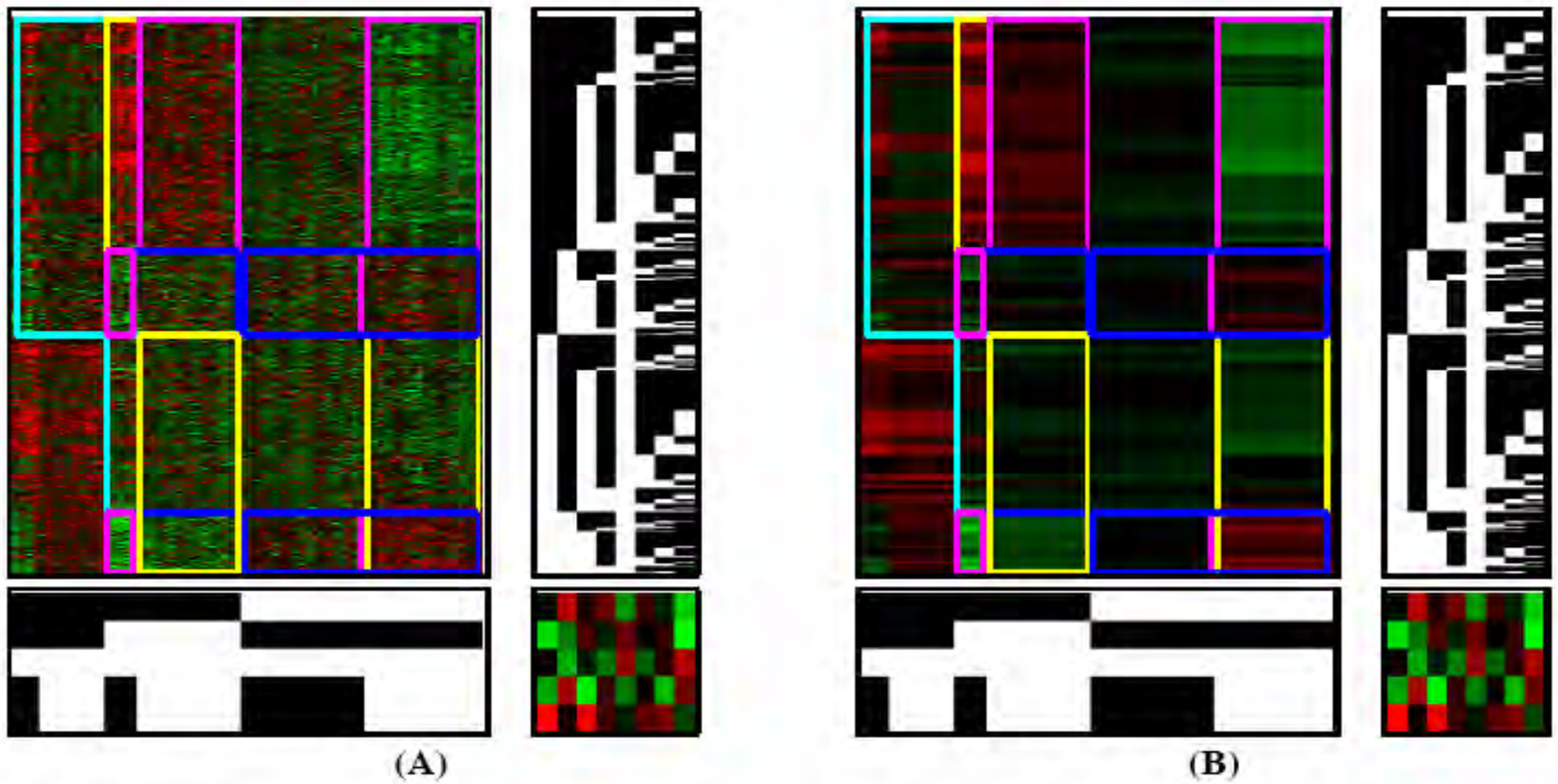
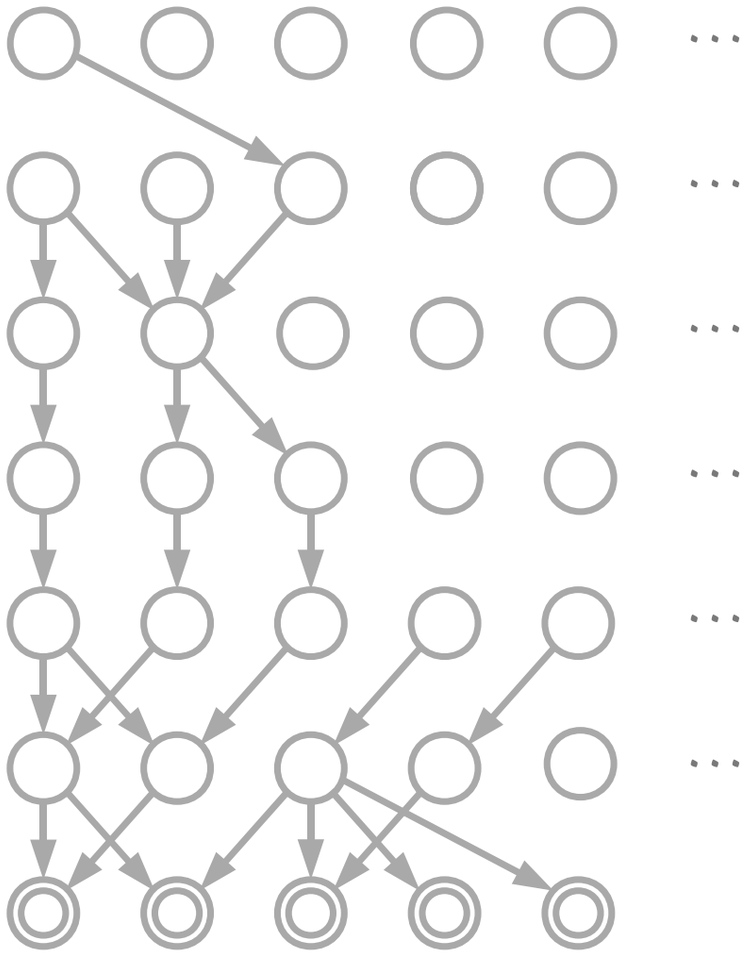
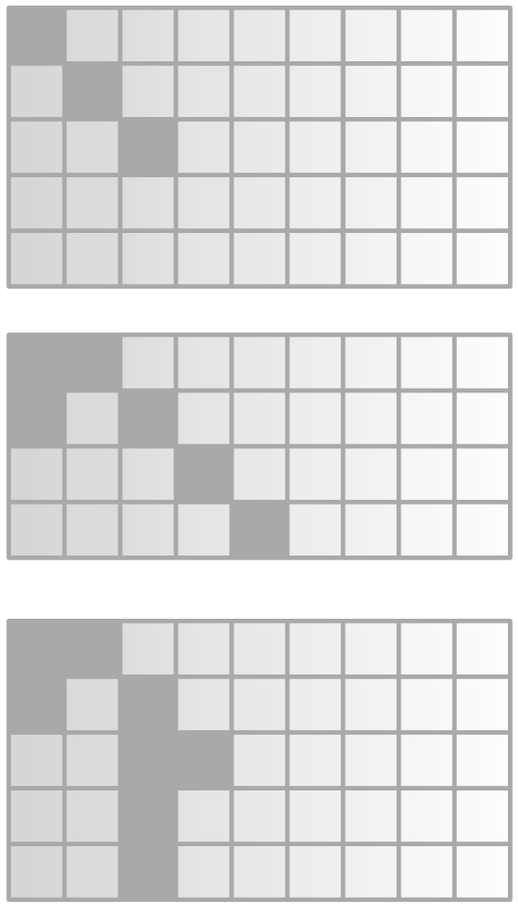


Figure 5: Gene expression results. **(A)** The top-left is $\mathbf{X}$ sorted according to contiguous features in the final $\mathbf{U}$ and $\mathbf{V}$ in the Markov chain. The bottom-left is $\mathbf{V}^\top$ and the top-right is $\mathbf{U}$. The bottom-right is $\mathbf{W}$. **(B)** The same as **(A)**, but the expected value of $\mathbf{X}$, $\hat{\mathbf{X}} = \mathbf{UWV}^\top$. We have hilighted regions that have both u_{ik} and v_{jl} on. For clarity, we have only shown the (at most) two largest contiguous regions for each feature pair.

Meeds et al (2007) Modeling Dyadic Data with Binary Latent Factors.

Learning Structure of Deep Sparse Graphical Models



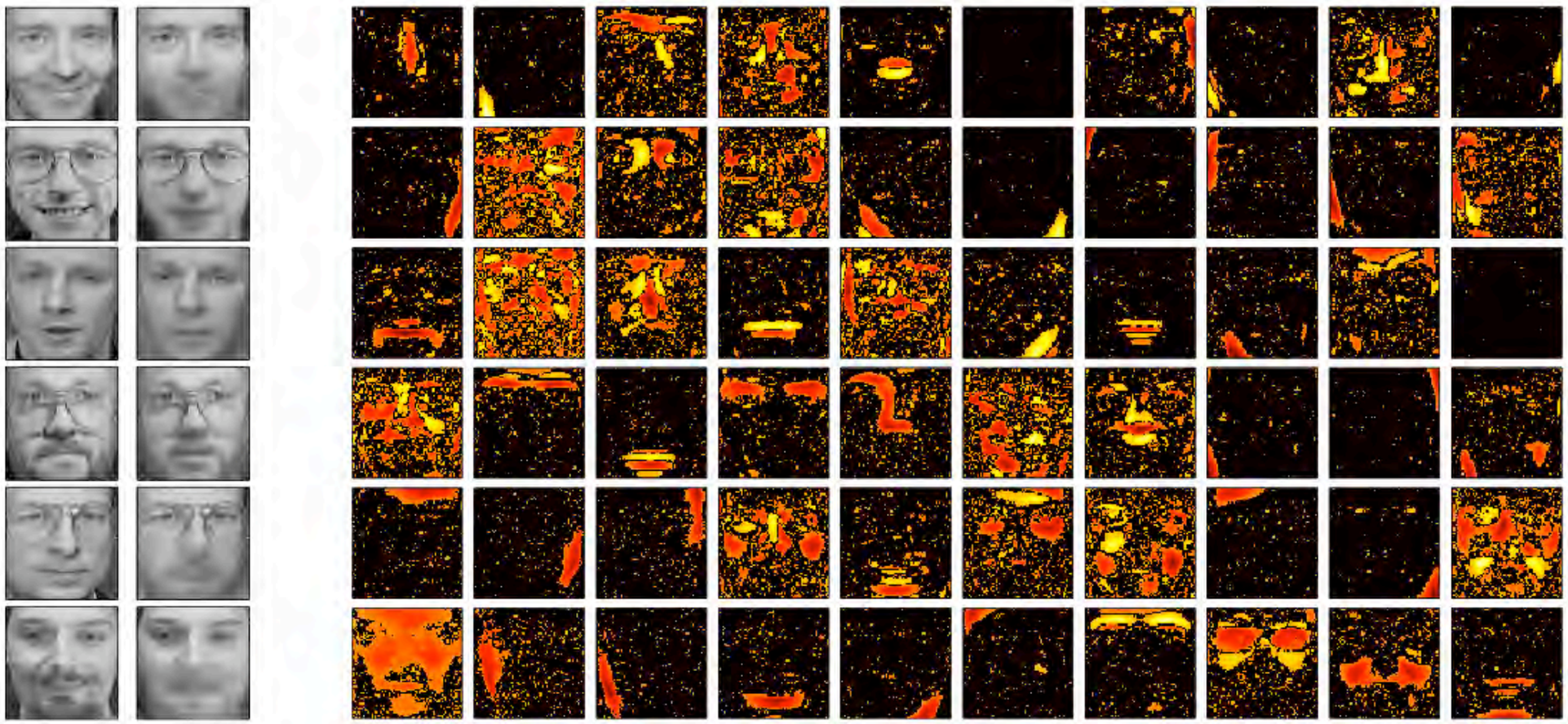


(w/ Ryan P. Adams, Hanna Wallach, 2010)

Learning Structure of Deep Sparse Graphical Models

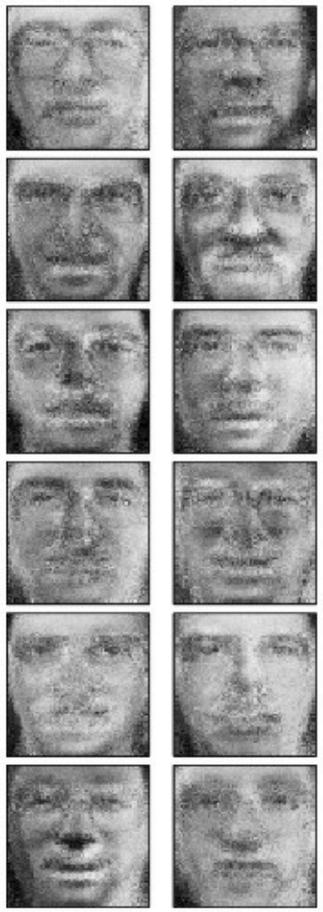
Olivetti Faces: 350 + 50 images of 40 faces (64×64)
Inferred: 3 hidden layers, 70 units per layer.

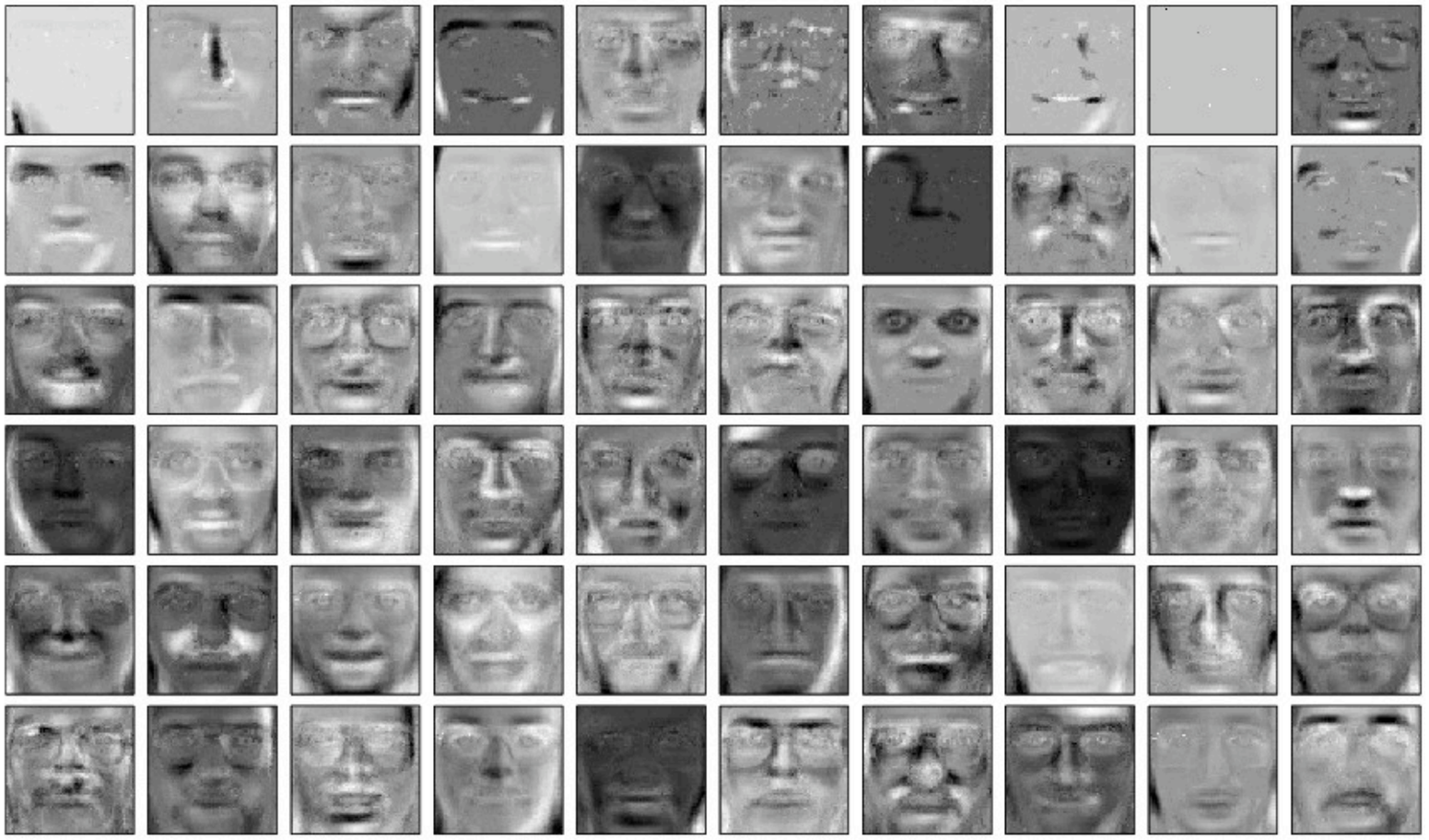
Reconstructions and Features:



Learning Structure of Deep Sparse Graphical Models

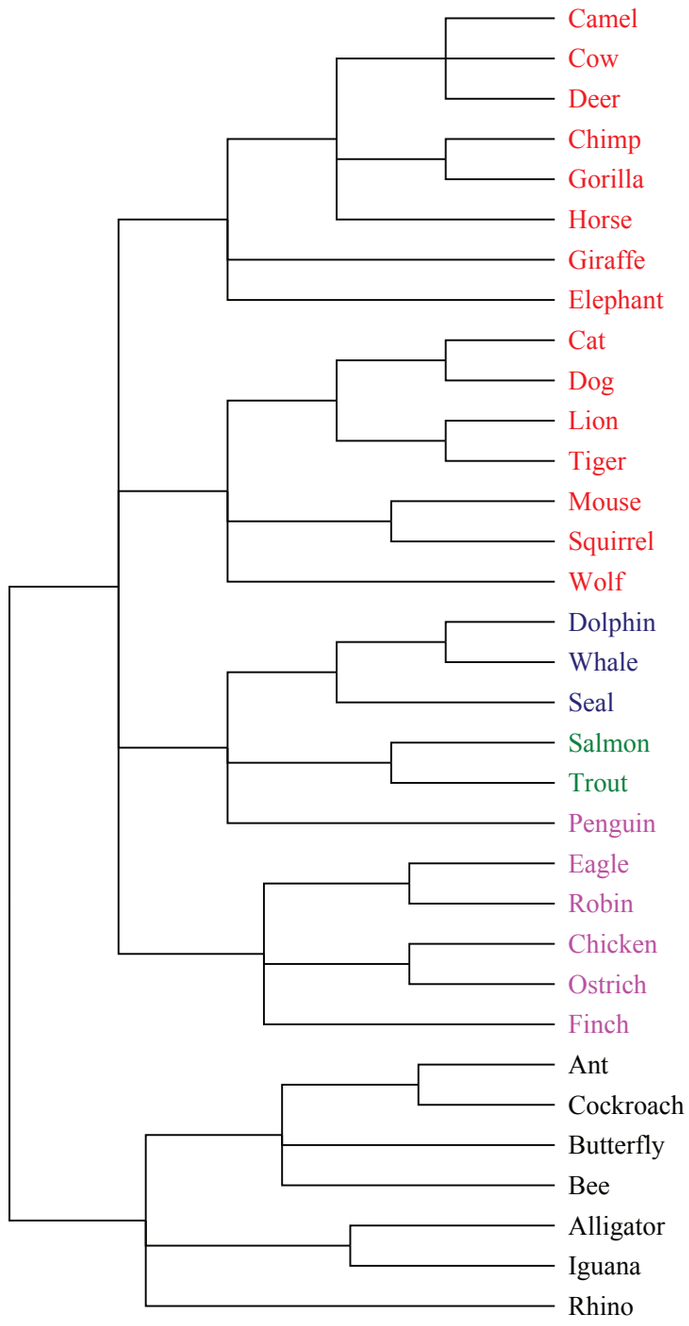
Fantasies and Activations:





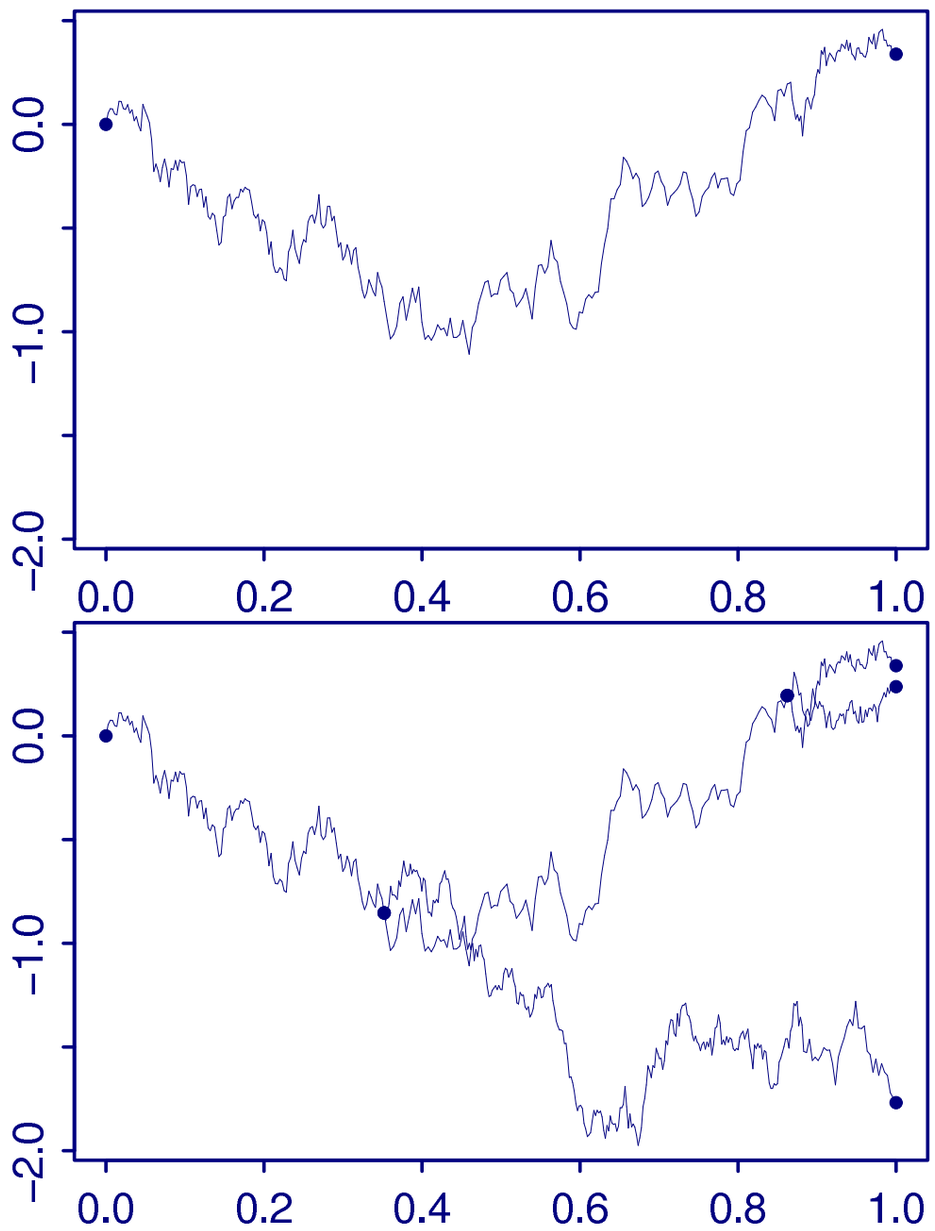
Hierarchies

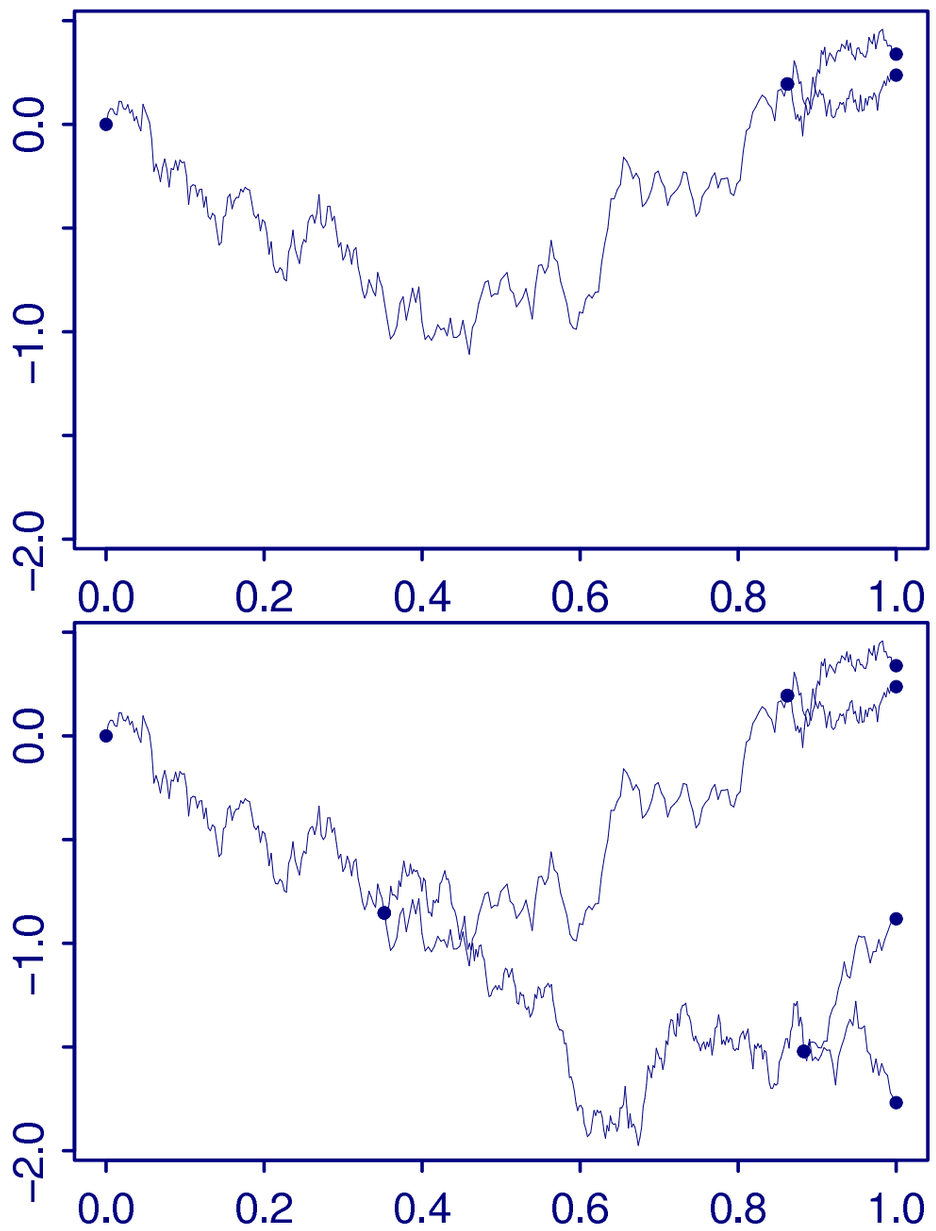
- true hierarchies
- parameter tying
- visualisation and interpretability



Dirichlet Diffusion Trees (DDT)

Generating from a DDT:





Figures from Neal, 2001.

Pitman-Yor Diffusion Trees

Generalises a DDT, but at a branch point, the probability of following each branch is given by a Pitman-Yor process:

$$P(\text{following branch } k) = \frac{b_k - \alpha}{m + \theta},$$
$$P(\text{diverging}) = \frac{\theta + \alpha K}{m + \theta},$$

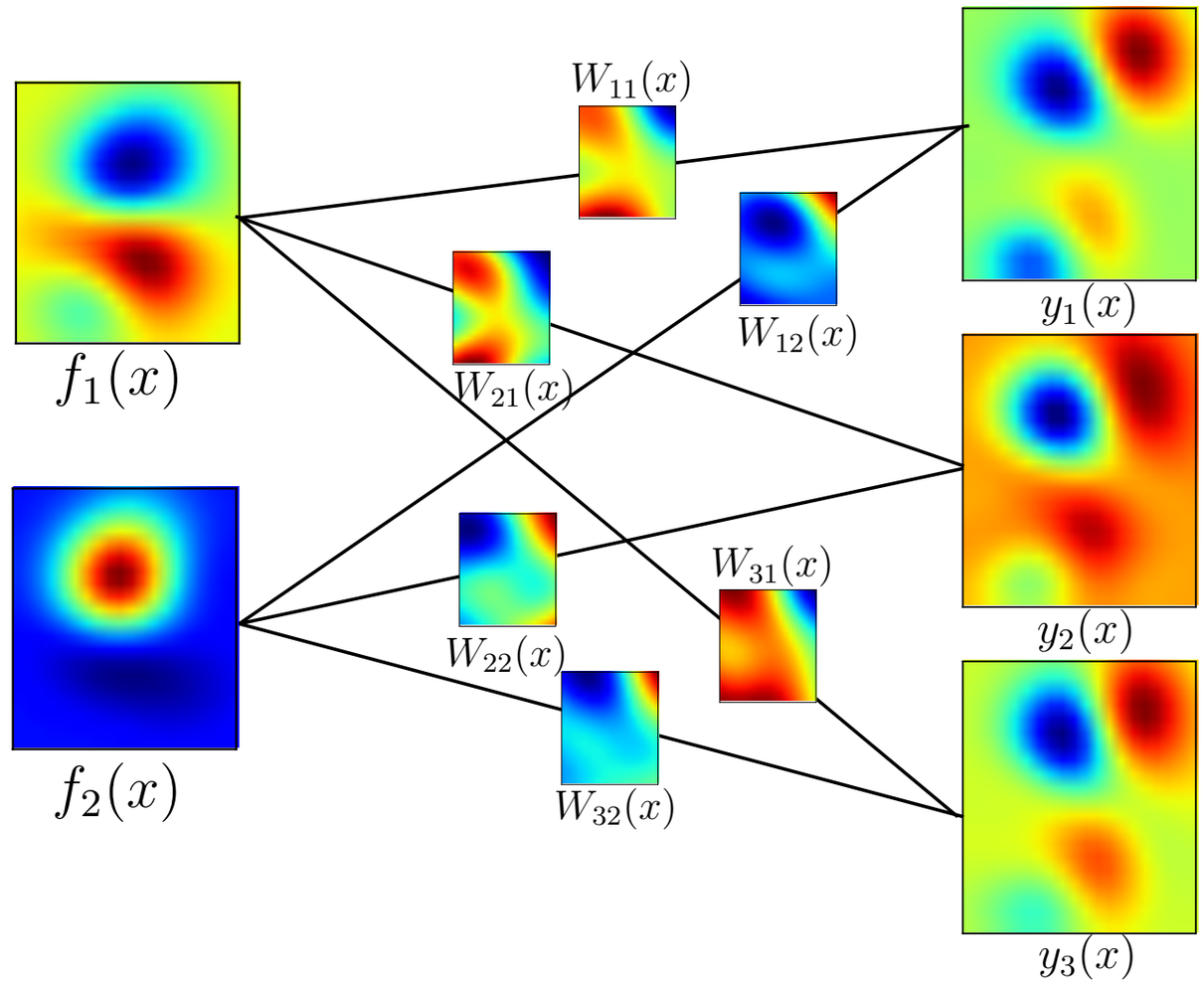
to maintain exchangeability the probability of diverging also has to change.

- naturally extends DDTs to arbitrary non-binary branching
- infinitely exchangeable over data
- prior over structure is the most general Markovian consistent and exchangeable distribution over trees

(w/ Knowles 2011)

Gaussian process regression networks

A model for multivariate regression which combines structural properties of Bayesian neural networks with the nonparametric flexibility of Gaussian processes



$$\mathbf{y}(x) = W(x)[\mathbf{f}(x) + \sigma_f \epsilon] + \sigma_y \mathbf{z}$$

(w/ David Knowles, Andrew Wilson, 2011)

Summary

- Probabilistic modelling and Bayesian inference are two sides of the same coin
- Bayesian machine learning treats learning as a probabilistic inference problem
- Bayesian methods work well when the models are flexible enough to capture relevant properties of the data
- This motivates non-parametric Bayesian methods, e.g.:
  - Indian buffet processes for sparse matrices and latent feature modelling
  - Pitman-Yor diffusion trees for hierarchical clustering
  - Wishart processes for covariance modelling
  - Gaussian process regression networks for multi-output regression