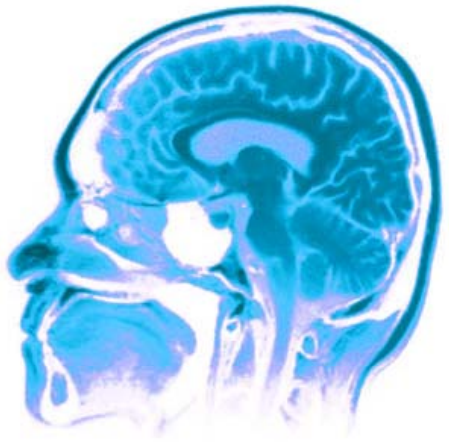
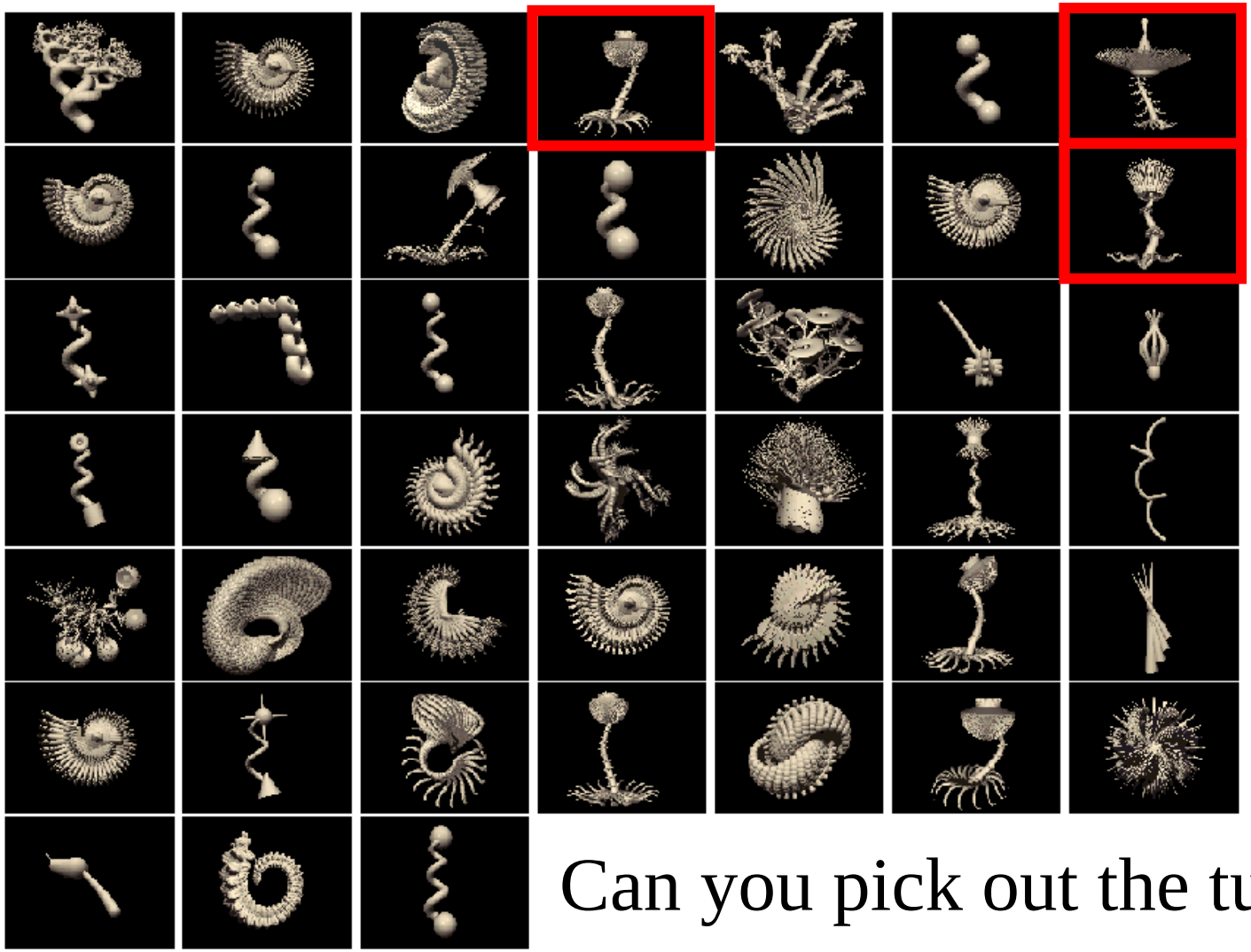
CPSC540



Introduction to Machine Learning

Nando de Freitas
2011



Can you pick out the tufas?

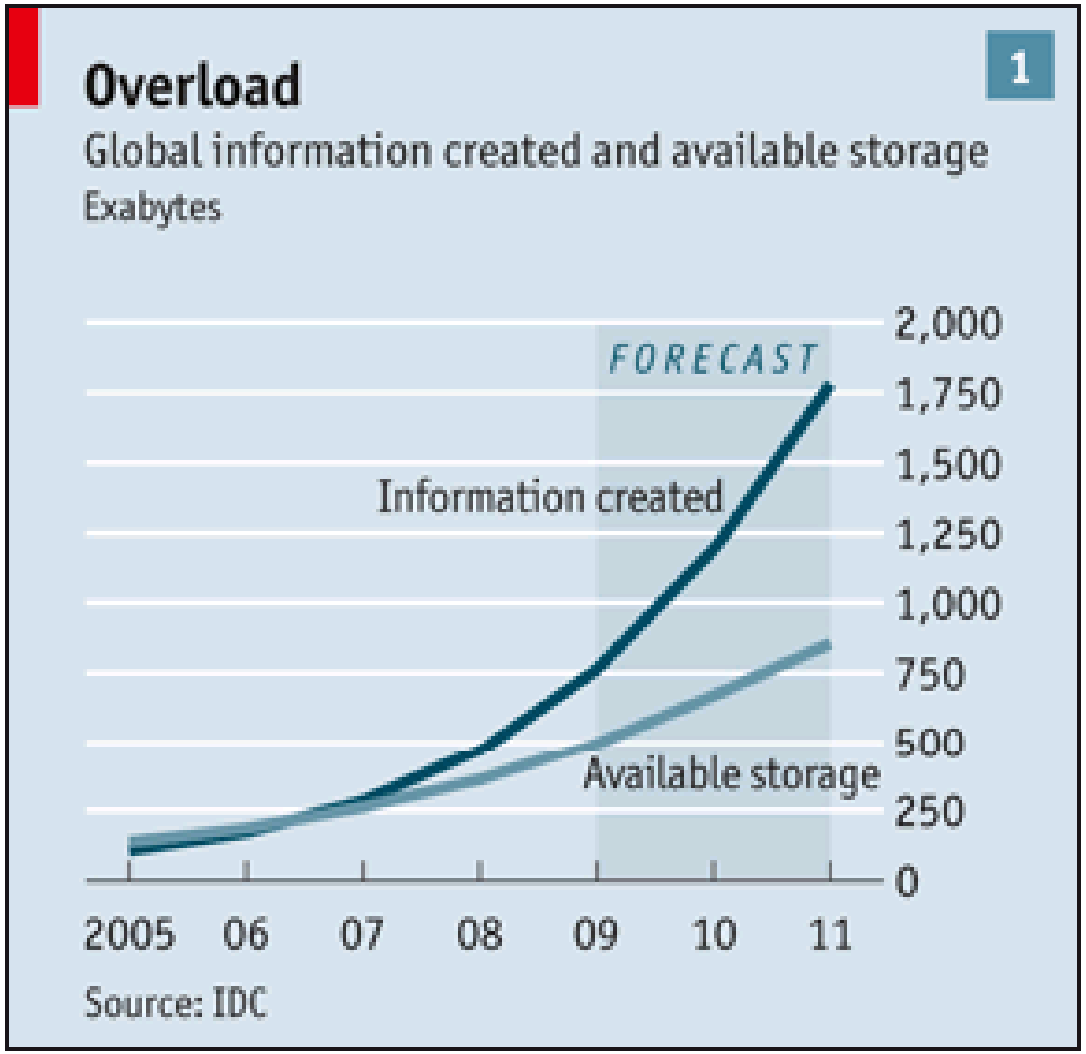
Josh Tenenbaum

Why Machine learning

- **Find patterns in data**
- **Make predictions**
- **Make decisions**
- **All at the same time**

- **Understand cognition**
- **Deal with the data deluge**
- **Build useful products, e.g. autonomous cars/robots, environmental anomaly detection, …**

Big data


Overload
1
Global information created and available storage
Exabytes
FORECAST
Information created
Available storage
2,000
1,750
1,500
1,250
1,000
750
500
250
0
2005
06
07
08
09
10
11
Source: IDC

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

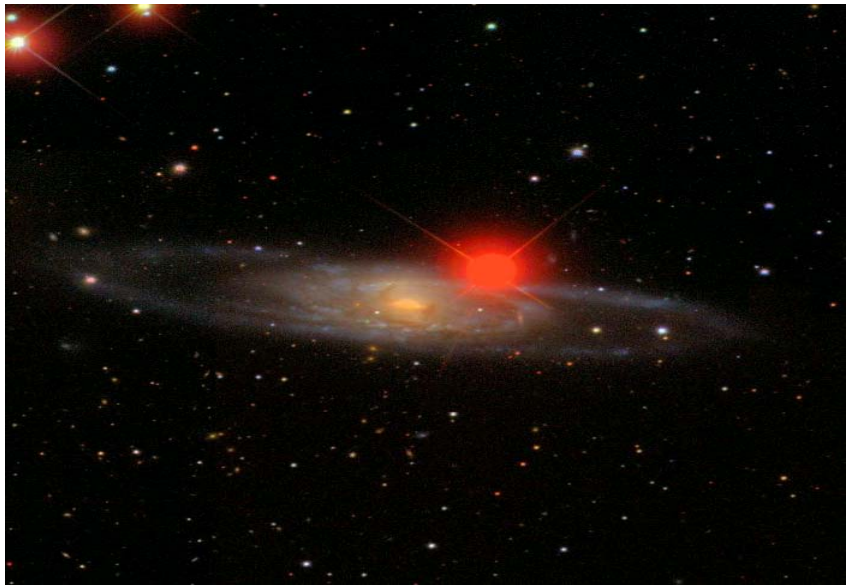
Source: *The Economist*

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

Wikipedia Current revisions only uncompressed ~112 GB (**896,000,000,000** bits)

Human brain ~**100, 000,000,000** neurons and ~**60,000, 000,000,000** synapses

Big data: Surveying the universe



"When the **Sloan Digital Sky Survey** started work in 2000, its telescope in New Mexico collected more data in its first few weeks than had been amassed in the entire history of astronomy.

Now, a decade later, its archive contains a whopping **140 terabytes** of information.

A successor, the **Large Synoptic Survey Telescope**, due to come on stream in Chile in 2016, will acquire that quantity of data every five days."

[*The Economist*, February 2010]

Big data: Financial markets

- Skyrocketing data volumes: 1.5 million messages/sec and growing
- About 70% of volume in US equity markets submitted electronically

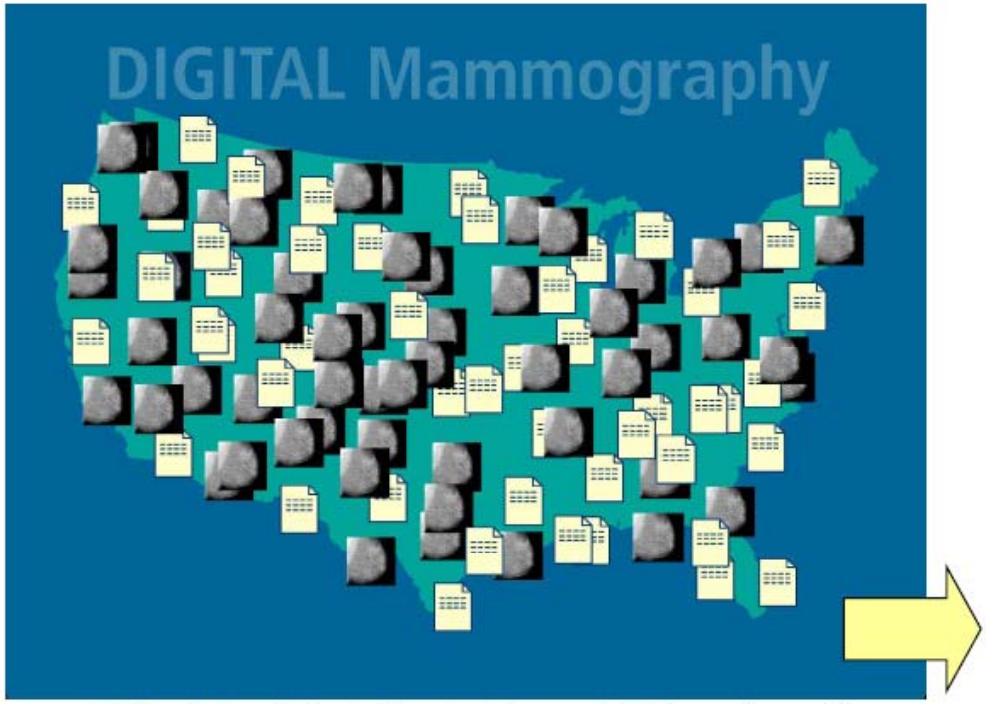
"*A 1-millisecond advantage in trading applications can be worth $100 million a year to a major brokerage.*"

-- The TABB Group

Courtesy of *Alan Wagner, UBC*

Big data: Medicine

National Digital Mammography Archive: a system designed to include a database growing by 28 PB per year according to IBM sources.



Highly Distributed and Massive Source

Use High Performance Networks, Hierarchical Storage and Indexing

Google

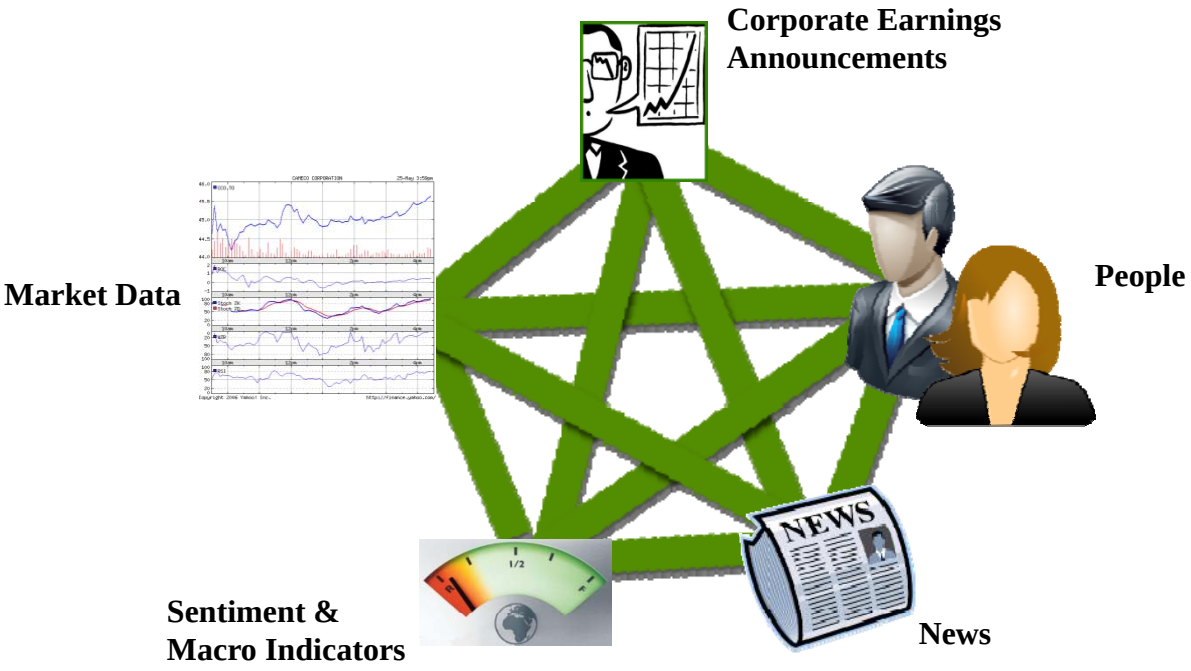
YouTube

- **Library of Congress** text database of ~20 TB
- **AT&T** 323 TB, 1.9 trillion phone call records.
- **World of Warcraft** utilizes 1.3 PB of storage to maintain its game.
- **Avatar** movie reported to have taken over 1 PB of local storage at *Weta Digital* for the rendering of the 3D CGI effects.
- **Google** processes ~24 PB of data per day.
- **YouTube:** 24 hours of video uploaded every minute. More video is uploaded in 60 days than all 3 major US networks created in 60 years. According to *cisco*, internet video will generate over 18 EB of traffic per month in 2013.

ML opportunities

Business

- Mining correlations, trends, spatio-temporal predictions.
- Efficient supply chain management.
- Opinion mining and sentiment analysis.
- Recommender systems.
- …



ML opportunities

Science

- Astronomy
- Biology
- Medicine
- Ecology
- Brain Science
- …

AQUATIC INFORMATICS™

Safety

- Crime stats
- Emergency response
- …

Government and institutional accountability

Big data: text

“Large” text dataset:

- **1,000,000 words in 1967**
- **1,000,000,000,000 words in 2006**

Success stories:

- **Speech recognition**
- **Machine translation**

What is the common thing that makes both of these work well?

- **Lots of labeled data**
- **Memorization is a good policy**

[Halevy, Norvig & Pereira, 2009]

Statistical machine translation

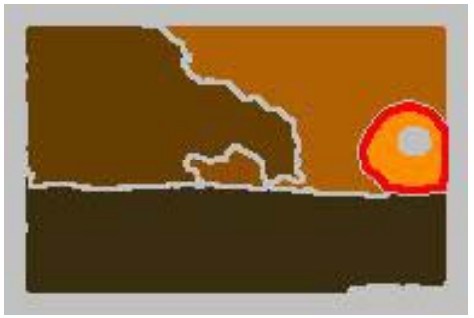
I love you **I love chocolate** **I am**

Yo te amo **Yo amo el chocolate** **Yo soy**

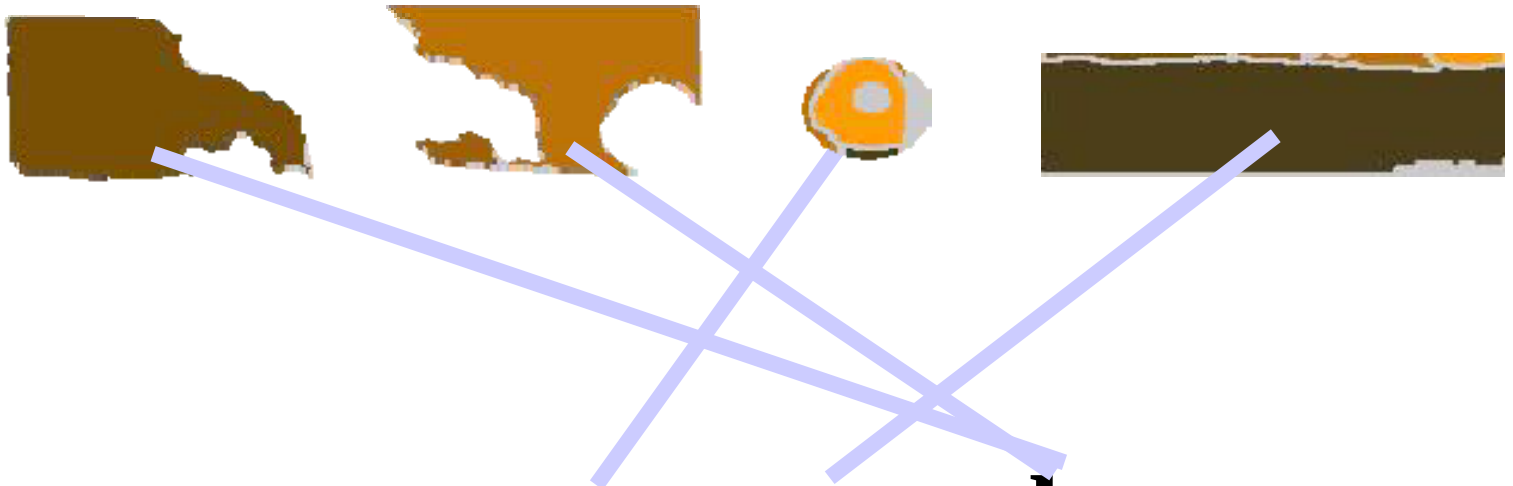
1. **Get many sentence pairs – easy.**
2. **Compute correspondences**
3. **Compute translation table: P(*Spanish*|*English*)**
4. **Repeat steps 2 and 3 till convergence**

Statistical machine translation

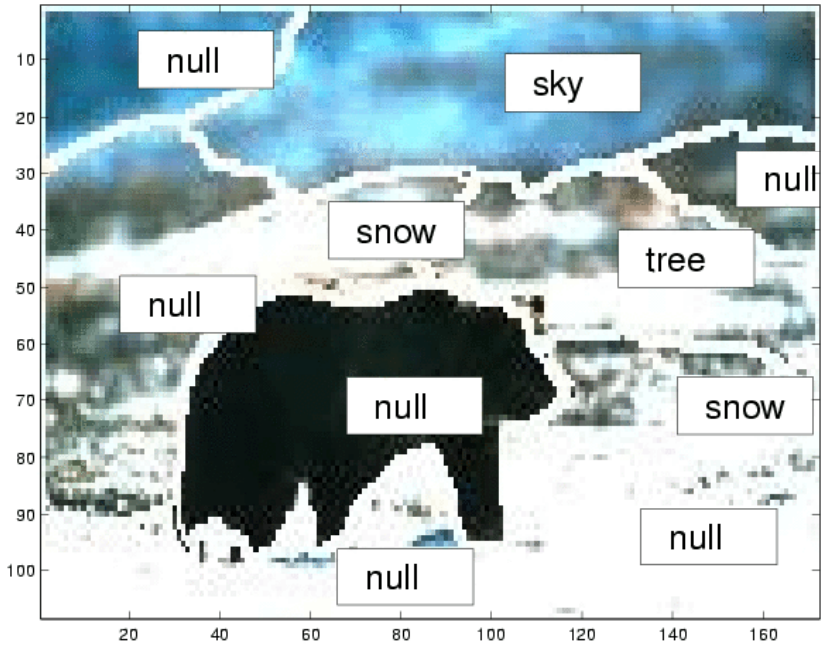
“Gorgeous red sea, sun and sky”

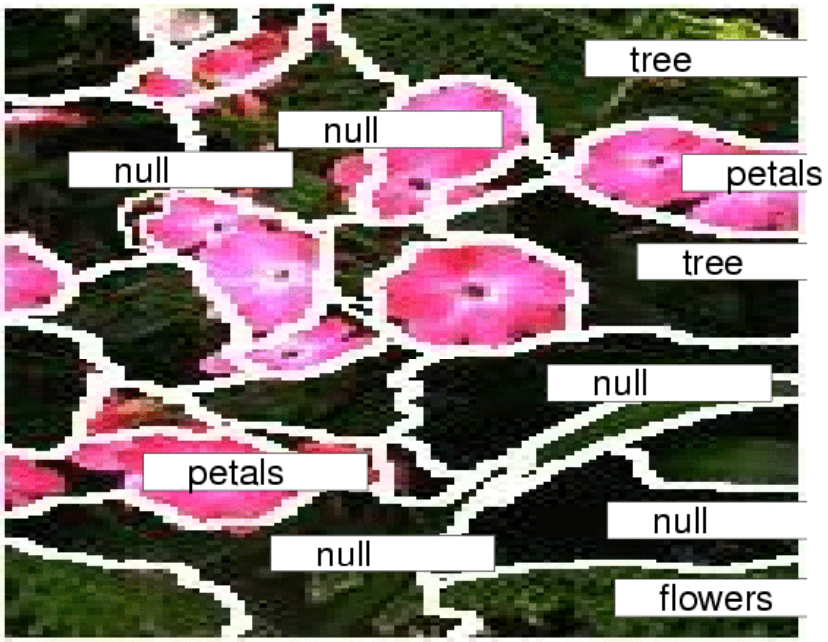


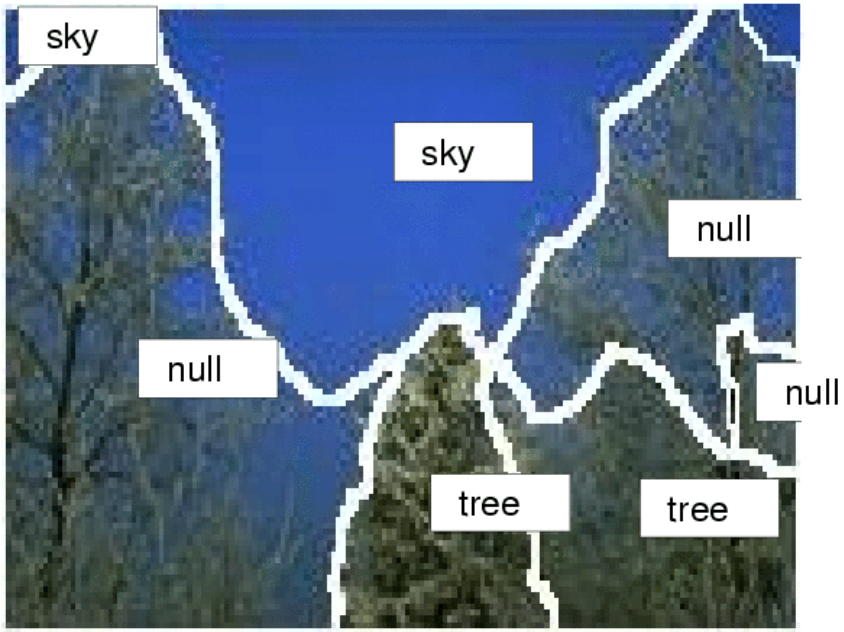
sun sea sky

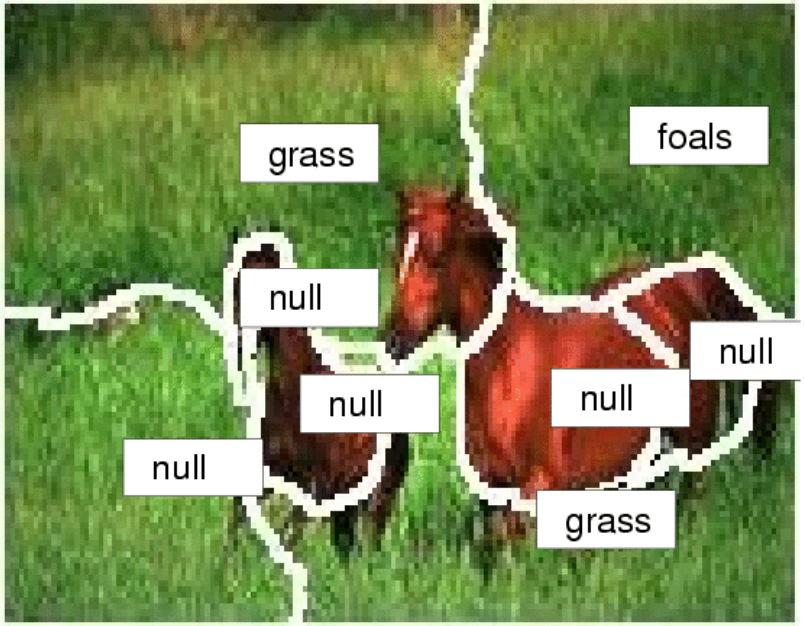


sun sea sky

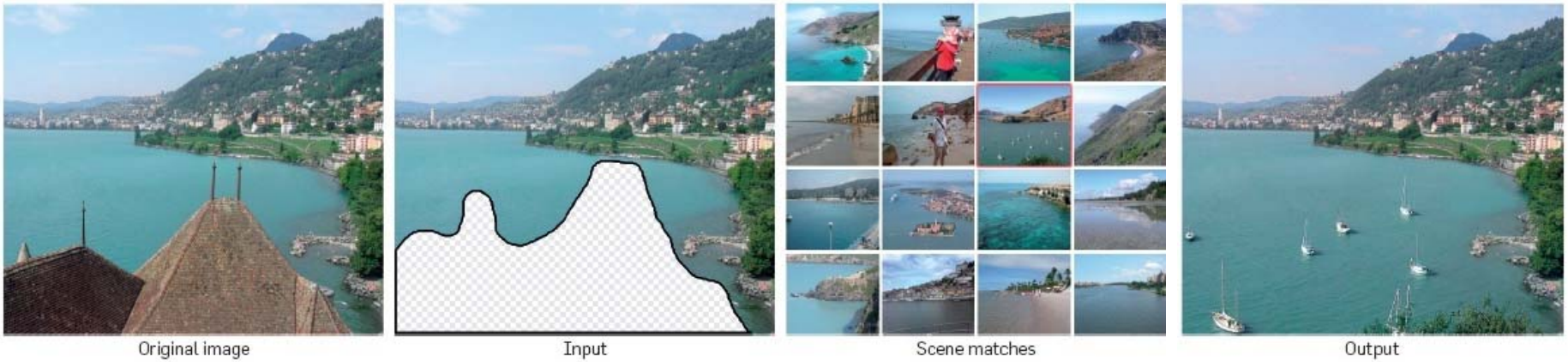

null
sky
null
snow
tree
null
null
snow
null
null
10
20
30
40
50
60
70
80
90
100
20
40
60
80
100
120
140
160


tree
null
null
petals
tree
null
petals
null
null
flowers


sky
sky
null
null
null
tree
tree


grass
foals
null
null
null
null
null
null
grass

Scene completion: more data is better



Given an input image with a missing region,
Efros uses matching scenes from a large
collection of photographs to complete the image

[Efros, 2008]

The semantic challenge

"We've already solved the sociological problem of building a network infrastructure that has encouraged hundreds of millions of authors to share a trillion pages of content.

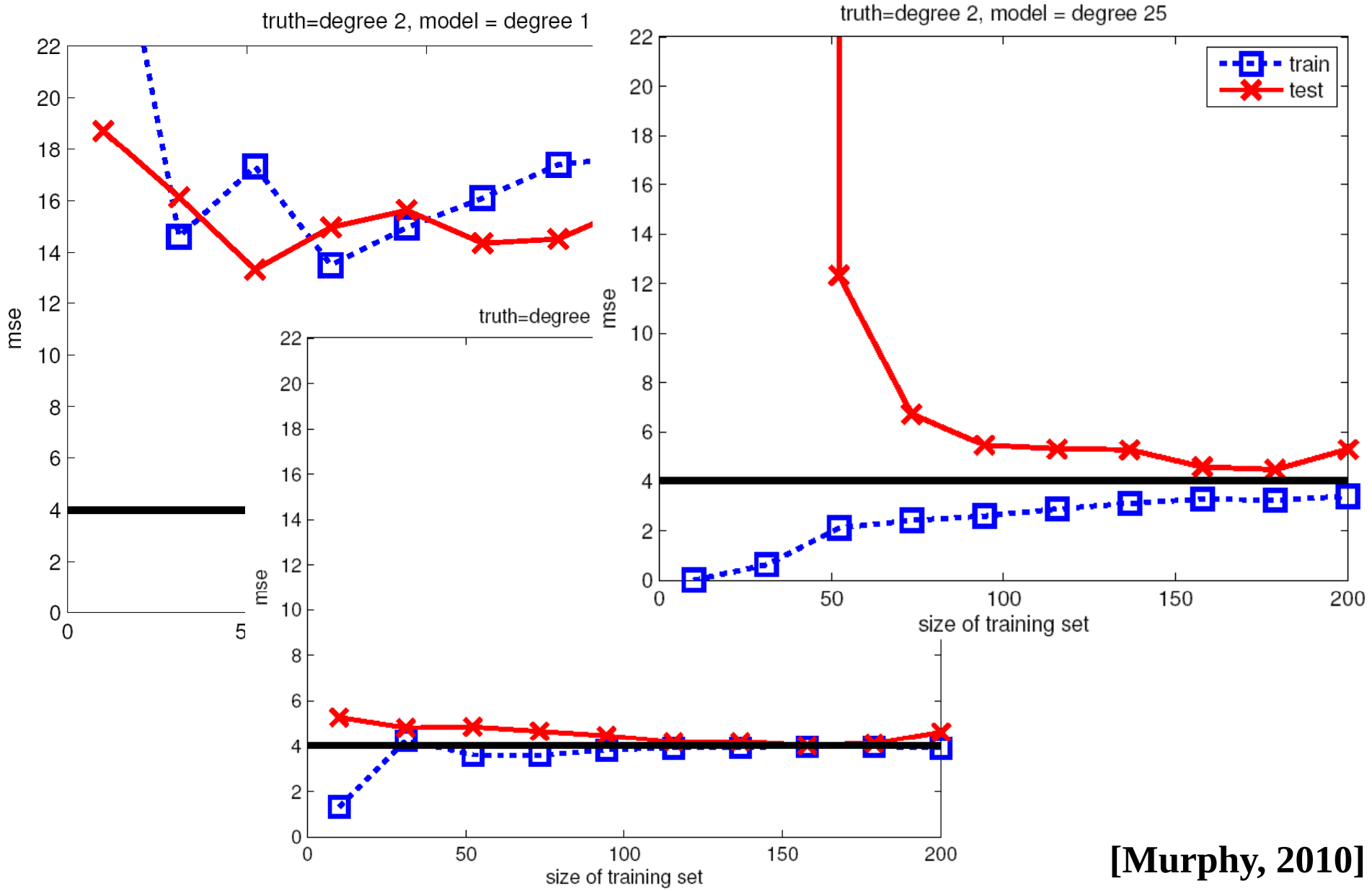
We've solved the technological problem of aggregating and indexing all this content.

But we're left with a scientific problem of interpreting the content"

Probability (*fact* given *evidence*) = ?

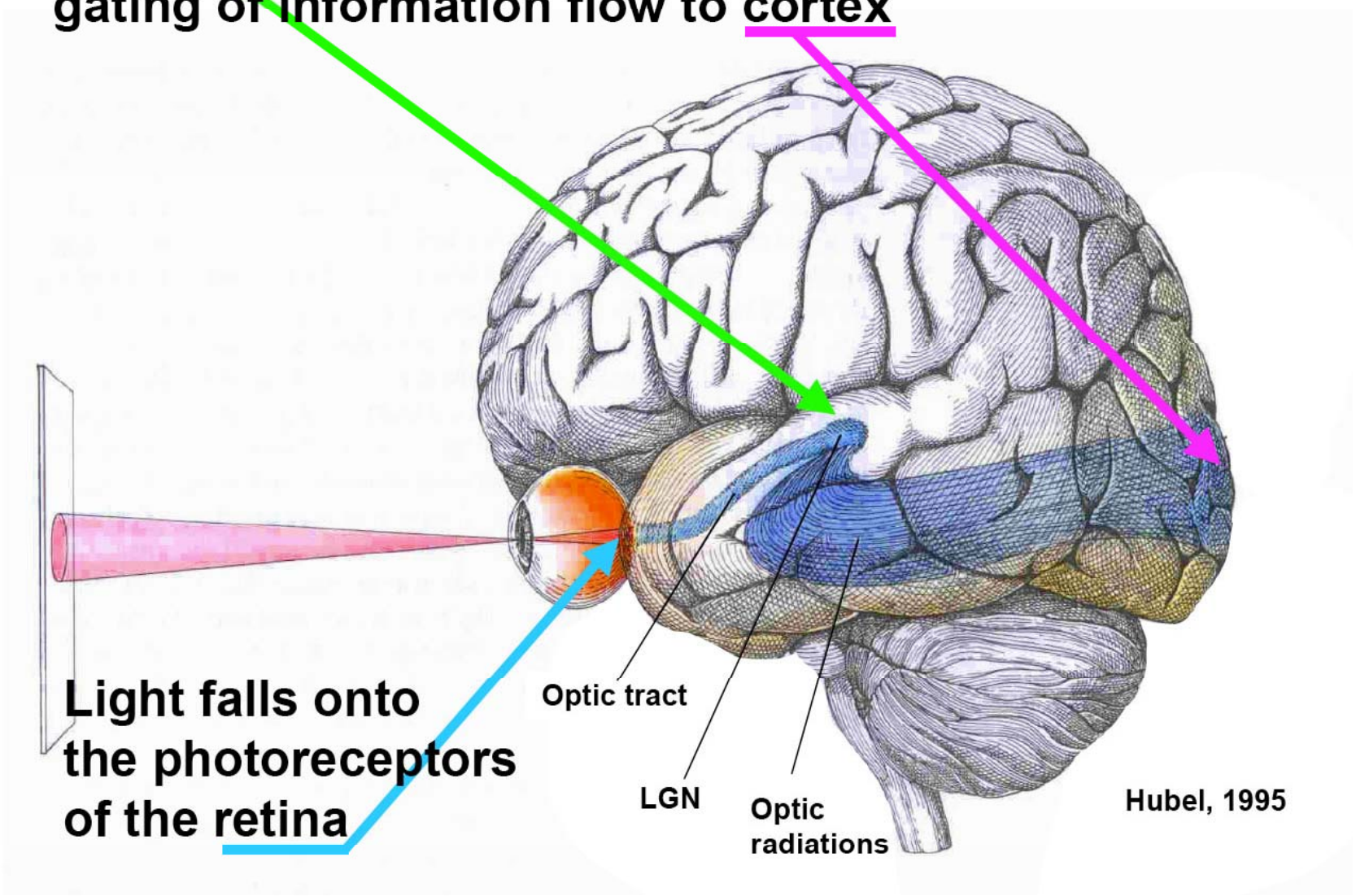
[Halevy, Norvig & Pereira, 2009]

Approximation, stats and optimization



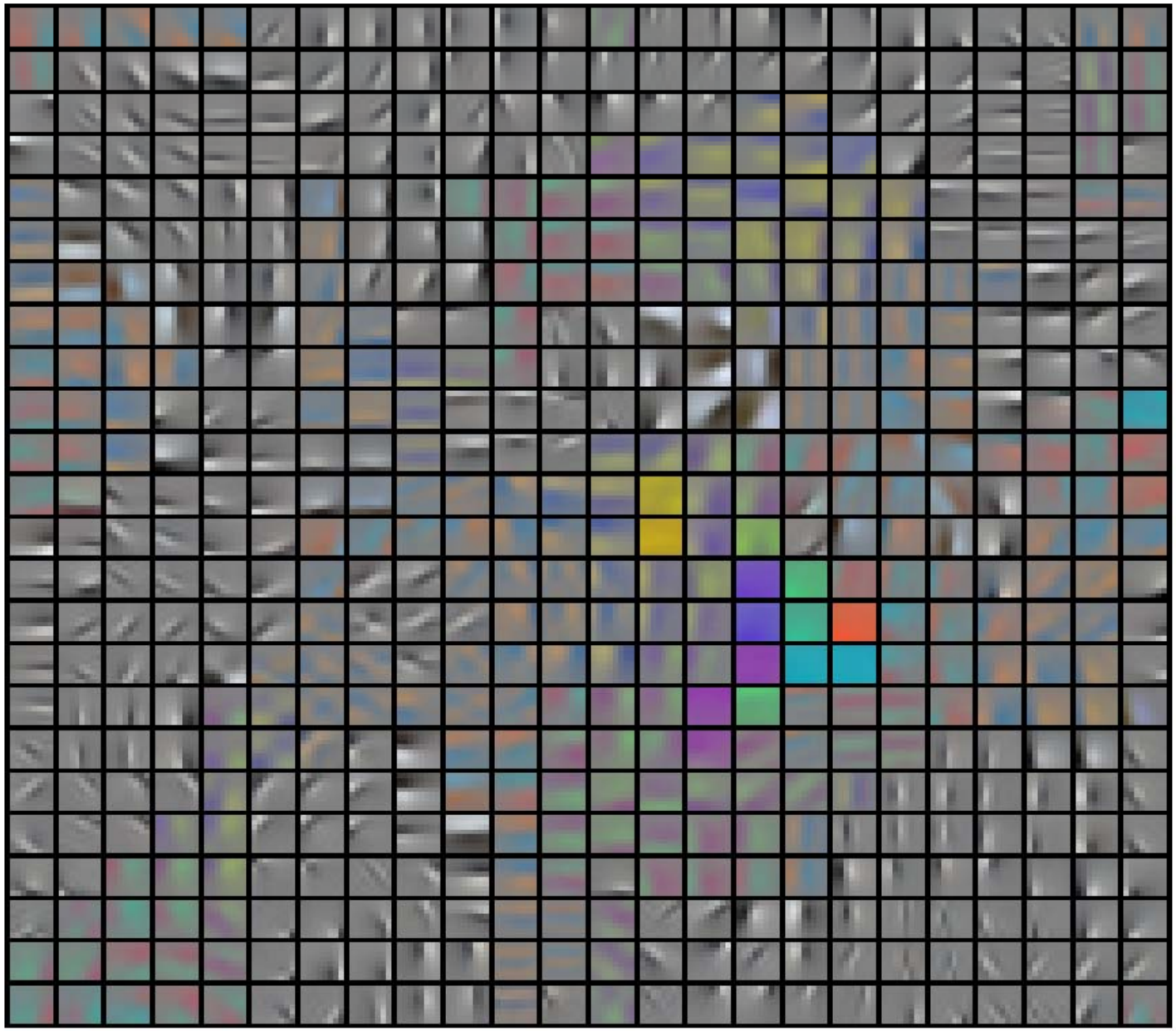
[Murphy, 2010]

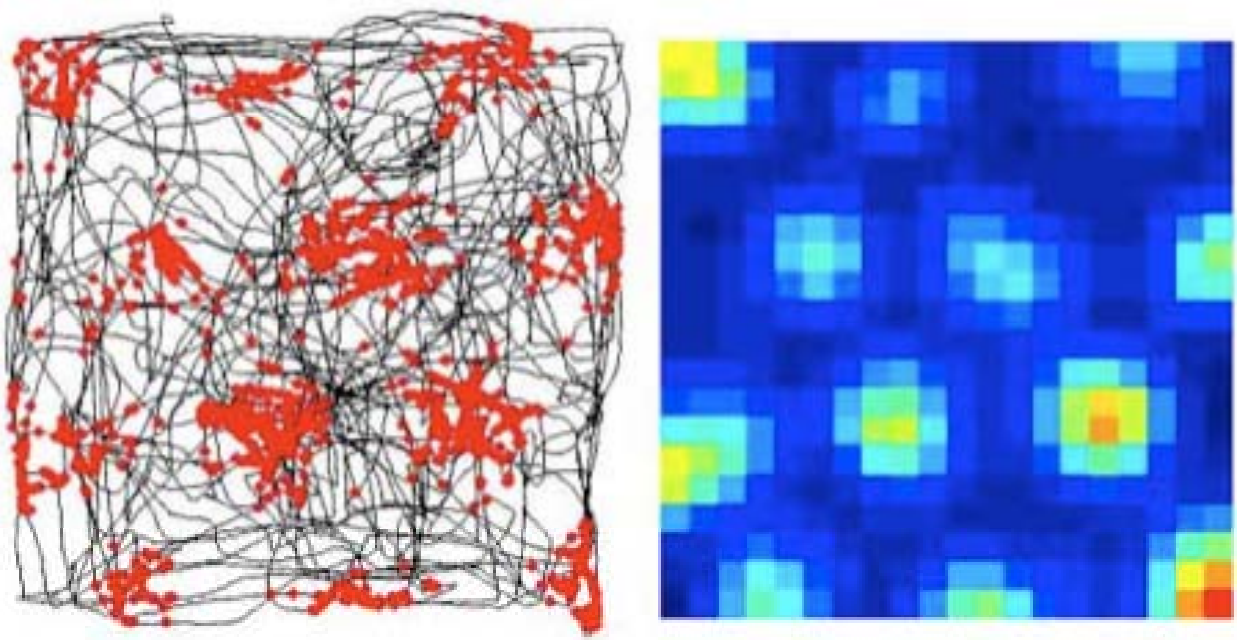
Thalamus (LGN) serves strategic role in gating of information flow to cortex



Light falls onto the photoreceptors of the retina

Topographic maps

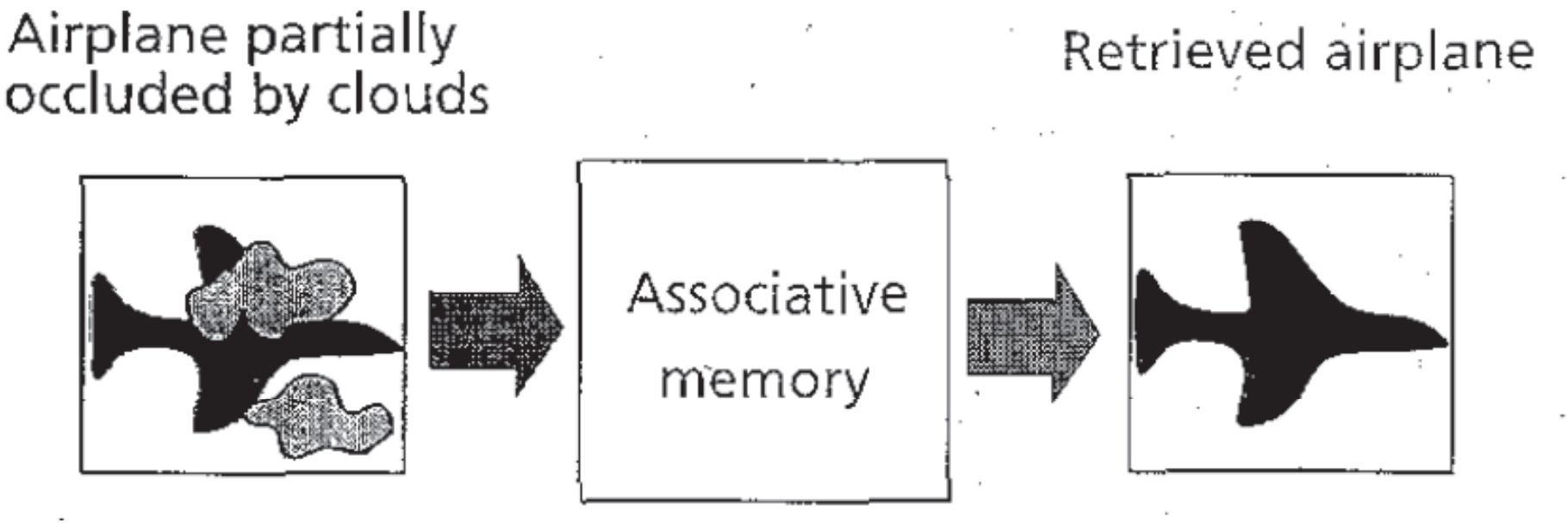




"the x and y coordinates correspond to the spatial location of a rat, which is running around freely inside a large box. The black lines in the left figure shows how this particular rat explored the box in a fairly haphazard manner. However, an electrode inserted in the rat's subcortex picks up a signal that is anything but chaotic: the responses of said neuron are given as red dots in the left figure, while the right figure gives the firing rate distribution (ranging from blue for silent and red for the peak rate of responding). Although the rat is running about randomly, this neuron is responding in a grid, seemingly coming on an off in response to the animal's spatial location."

[Hafting et al 2005]

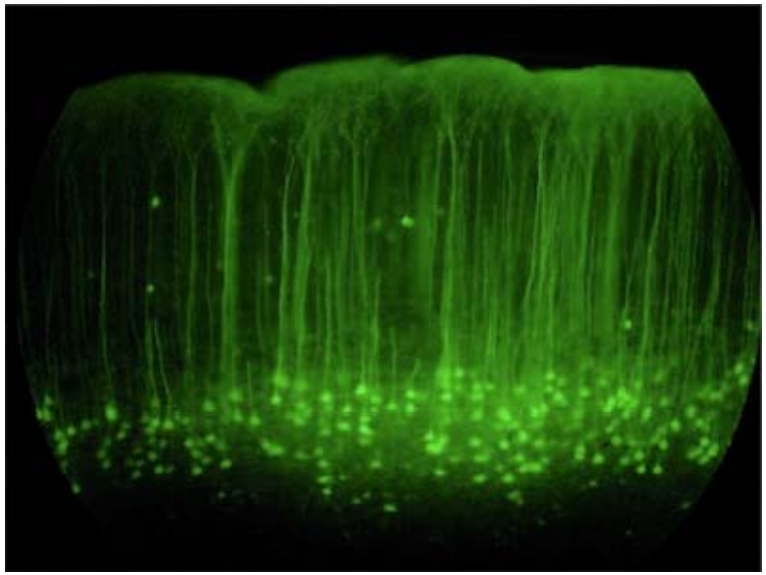
Associative memory

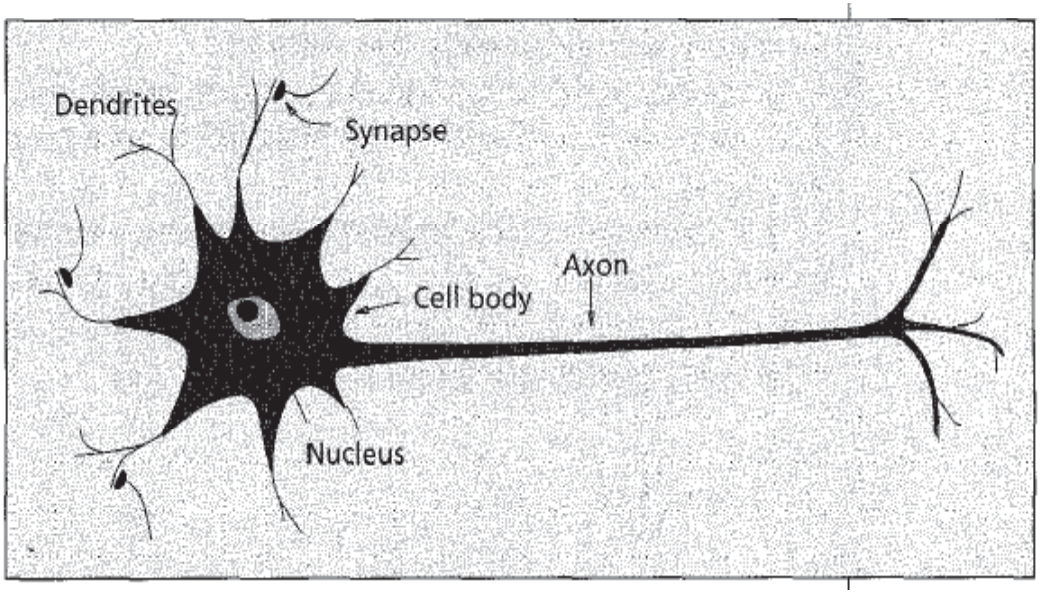


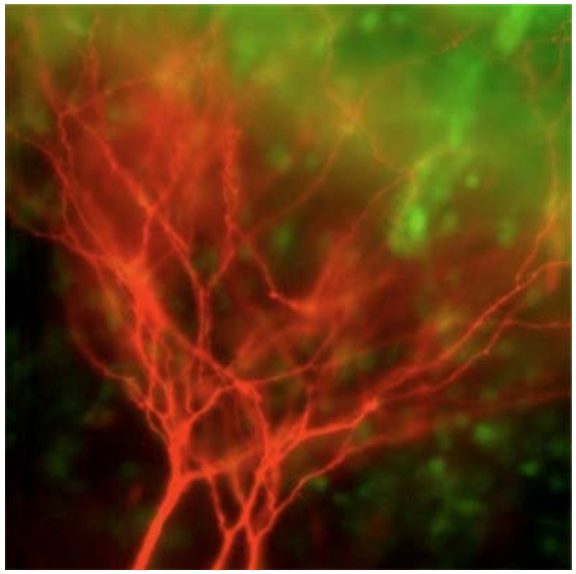
Example 2: Say the alphabet, …. backward

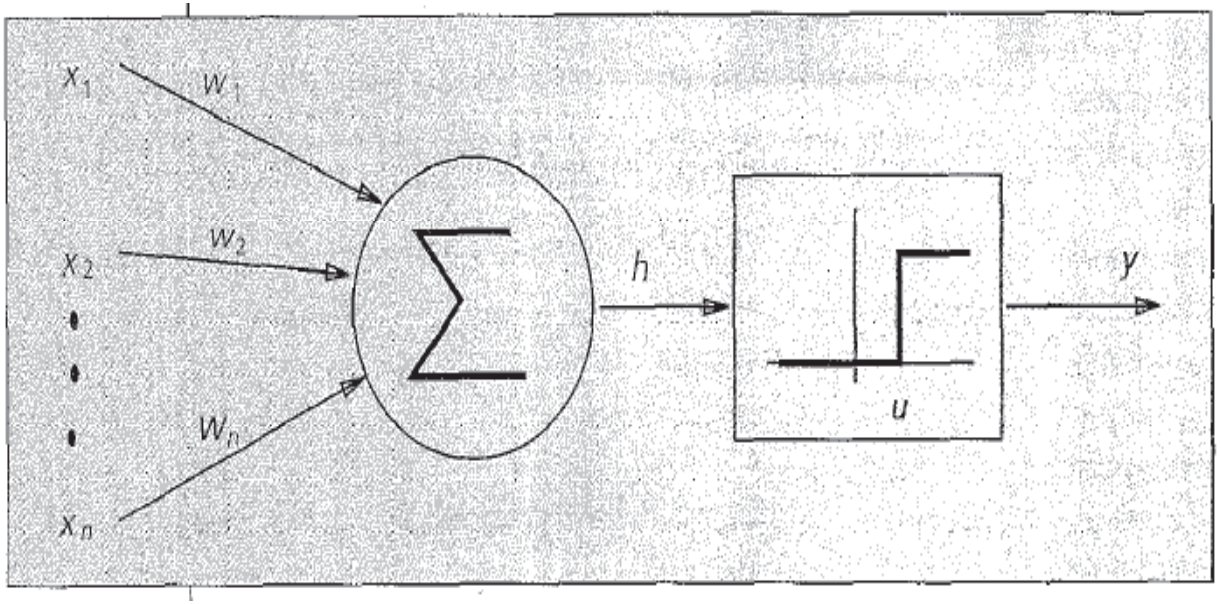
[Jain, Mao & Mohiuddin, 1996]

McCulloch-Pitts model of a neuron

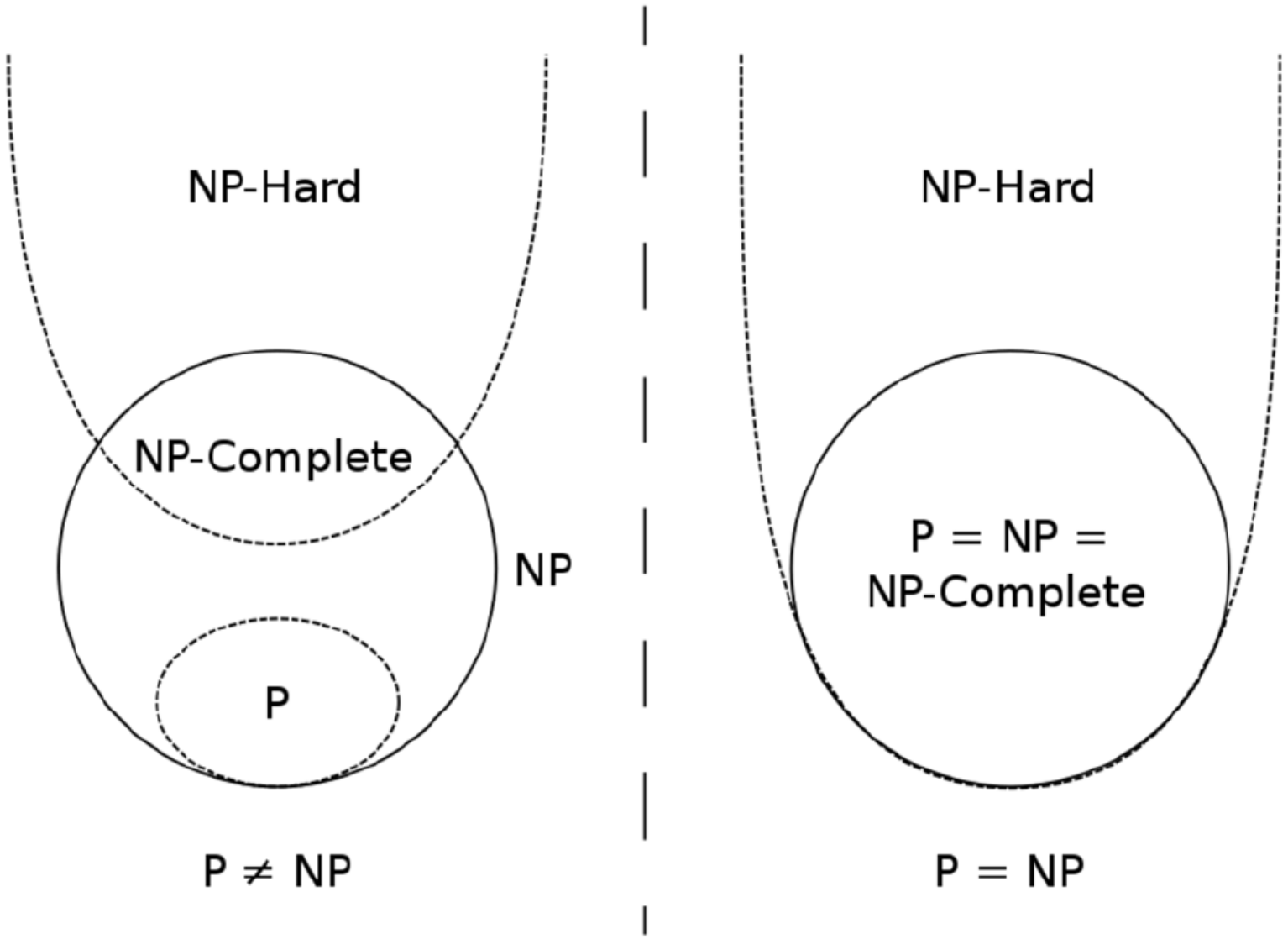






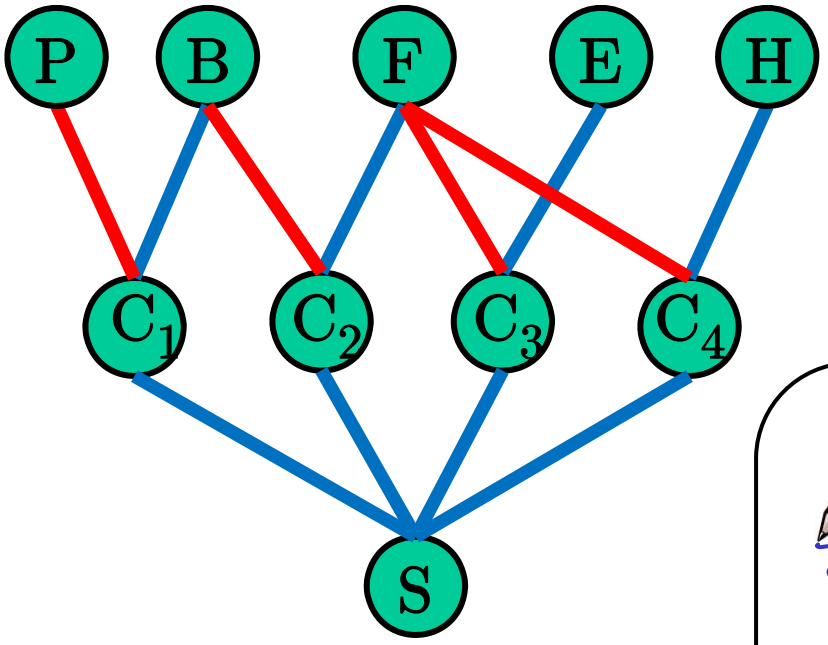


Complexity



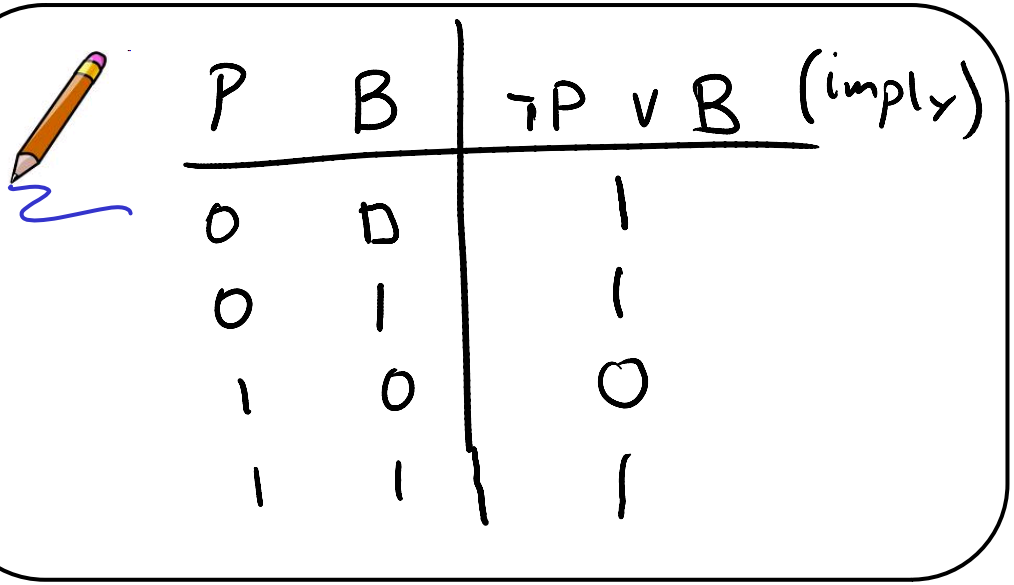
Re-visiting logic, NP and 2-SAT

Consider the CNF expression $S = C_1 \wedge \ldots \wedge C_m$, where each clause C_i is a disjunction of literals $x_{i,1} \vee \ldots \vee x_{i,k_i}$ defined on propositional variables. When each clause has two parents at most, the problem is known as 2-SAT.

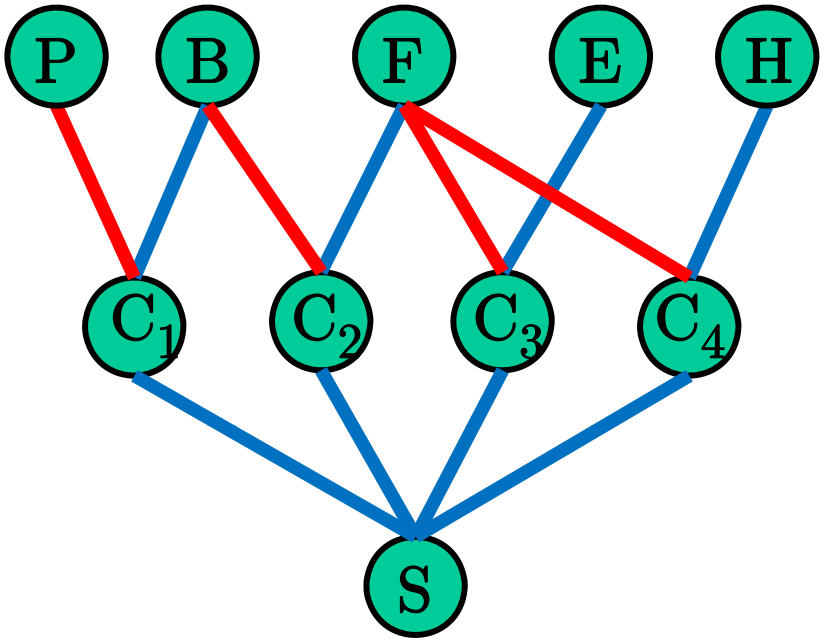


implication symbol

$$Parrot \implies Bird \qquad \neg P \vee B$$
$$Bird \implies Flies \qquad \neg B \vee F$$
$$Flies \implies Escapes \qquad \neg F \vee E$$
$$Flies \implies HasWings \qquad \neg F \vee H$$



Re-visiting logic, NP and 2-SAT

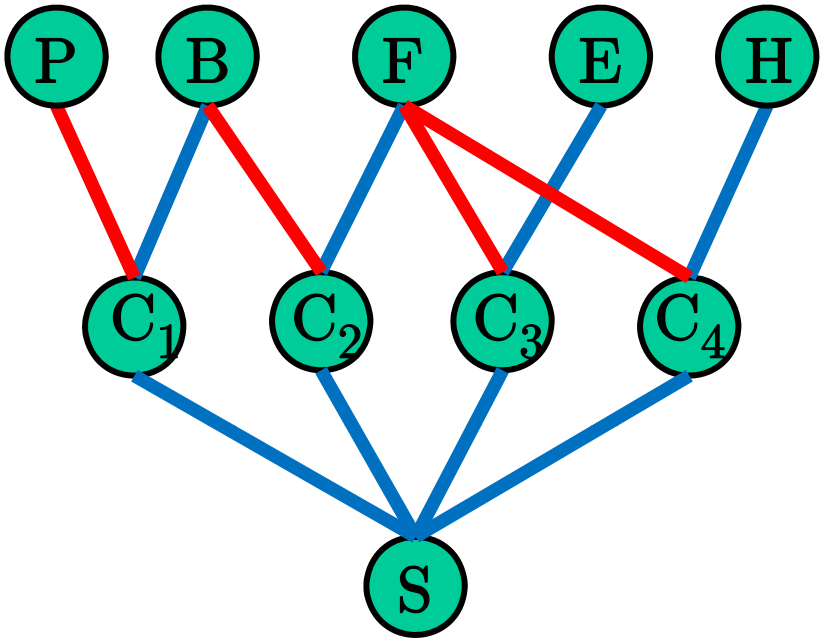


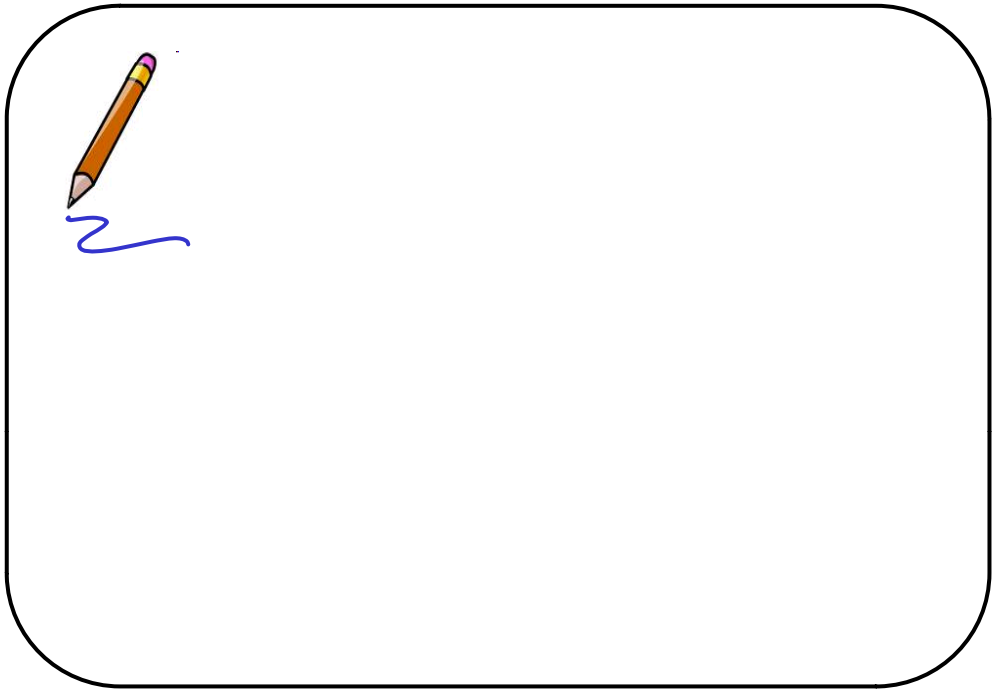
$$Parrot \implies Bird \qquad C_1 = \neg P \vee B$$
$$Bird \implies Flies \qquad C_2 = \neg B \vee F$$
$$Flies \implies Escapes \qquad C_3 = \neg F \vee E$$
$$Flies \implies HasWings \qquad C_4 = \neg F \vee H$$

$$S = C_1 \wedge C_2 \wedge C_3 \wedge C_4$$

1. **Verification: Does (P=1,B=1,F=1,E=1,H=1), i.e. (11111), satisfy this 2-SAT problem?**
2. **Verification: Does (10111) satisfy it?**
3. **Maximization: What is the maximum number of clauses that can be satisfied?**
4. **What is the number of possible assignments to (PBFEH)?**
5. **Counting: How many assignments satisfy this 2-SAT example?**

Logic, NP, 2-SAT and Monte Carlo

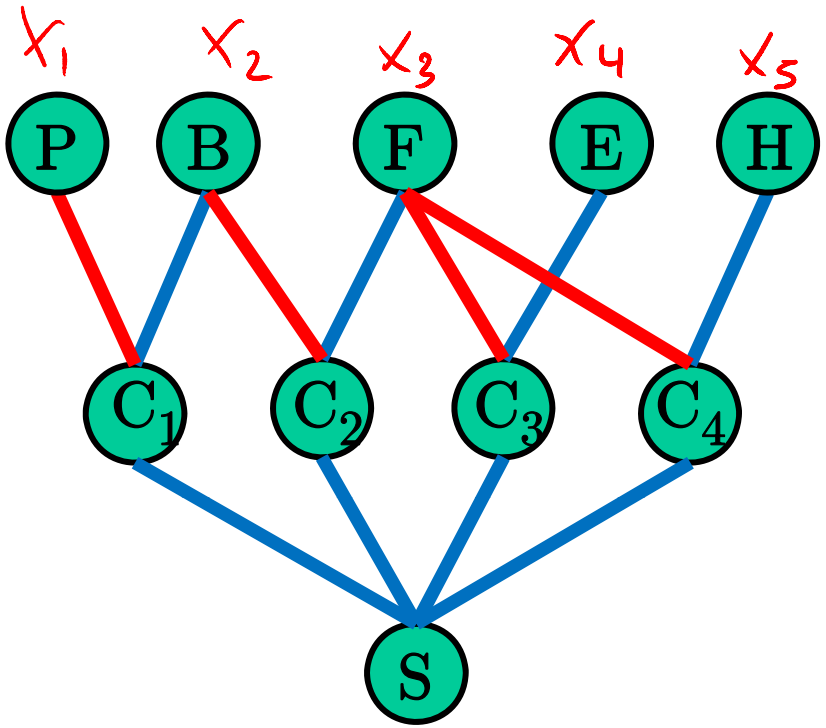




Counting: How many assignments satisfy this 2-SAT example
Approximate answer: Use the Monte Carlo Method.

i. **Sample P, B, F, E and H by flipping a coin for each variable N times.**
ii. **For each sample of (PBFEH), check for satisfiability.**
iii. **The probability of satisfiability, P(S=1), is approximated as the number of satisfying samples divided by N.**
iv. **The expected number number of satisfiable samples n = P(S=1) 2^5.**

From max-2-SAT to Energy



$\neg P \vee B \longrightarrow \theta_1 P(1-B)$

$\neg B \vee F \longrightarrow \theta_2 B(1-F)$

$\neg F \vee E \longrightarrow \theta_3 F(1-E)$

$\neg F \vee H \longrightarrow \theta_4 F(1-H)$ +

Introduce the notation

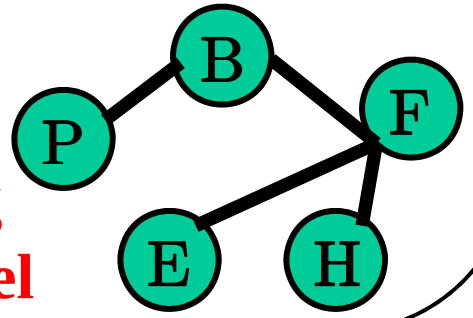
$P = X_1, B = X_2, F = X_3, E = X_4, H = X_5.$

The total energy of the system is: ϵ

$$E = \theta_1 P + \theta_2 B + (\theta_3 + \theta_4) F - \theta_1 PB - \theta_2 BF - \theta_3 FE - \theta_4 FH$$

Assume some clauses are harder to satisfy than others. Introduce a weight (θ) to measure this.

To obtain the Energy of the system of binary variables, use X for a negated propositional variable and 1-X otherwise. Then sum over all clauses.

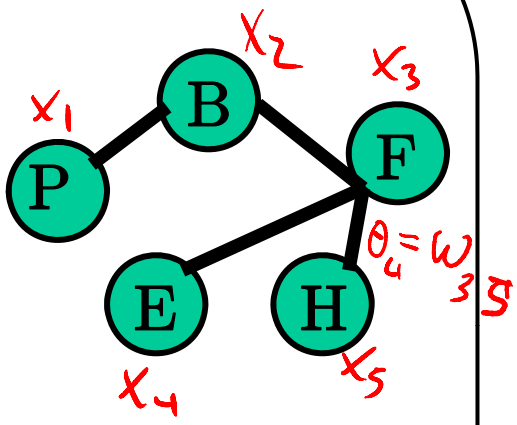


Ising model

From max-2-SAT to Energy

$$E = \theta_1 P + \theta_2 B + (\theta_3 + \theta_4) F - \theta_1 PB - \theta_2 BF - \theta_3 FE - \theta_4 FH$$

Let $P = x_1, B = x_2, F = x_3, E = x_4$ and $H = x_5$



The energy can be written as:

$$E = -\sum_{i=1}^{5} b_i x_i - \sum_{i=1}^{5} \sum_{j>i} x_i w_{ij} x_j$$

In our case:

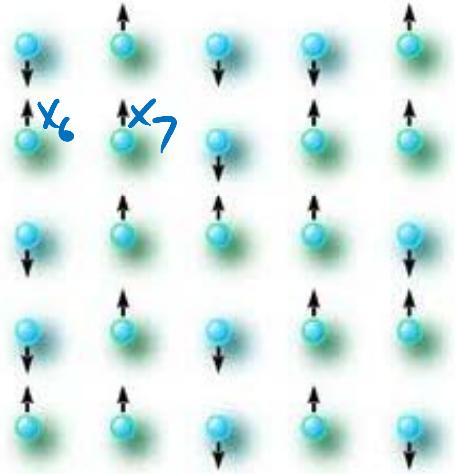
$b_1 = -\theta_1$ | $w_{12} = \theta_1$

$b_2 = -\theta_2$ | $w_{23} = \theta_2$

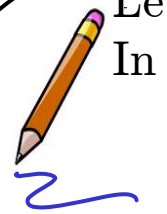
$b_3 = -\theta_3 - \theta_4$ | $w_{34} = \theta_3$

$b_4 = 0$ | $w_{35} =$

$b_5 = 0$ | $w_{45} =$



From max-2-SAT to Energy to Probability



Let us look at the energy of a few configurations, assuming all the $\theta_i = 1$. In this case the energy is simply:

$$E(x_1, x_2, \ldots, x_5) = x_1 + x_2 + 2x_3 - x_1x_2 - x_2x_3 - x_3x_4 - x_3x_5$$

What is the lowest energy? When is it attained?
What is the maximum energy?
What should the most probable configuration be?

x_1	x_2	x_3	x_4	x_5	E
1	1	1	1	1	0
0	1	1	1	1	0
1	0	1	1	1	1
1	1	1	0	0	2
0	0	1	1	1	0

$$P(x_1 x_2 x_3 x_4 x_5) = \frac{e^{-E(x_1 \cdots x_5)}}{Z}$$

$$Z = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_5} e^{-E(x_1, \cdots x_5)}$$

Boltzmann distribution

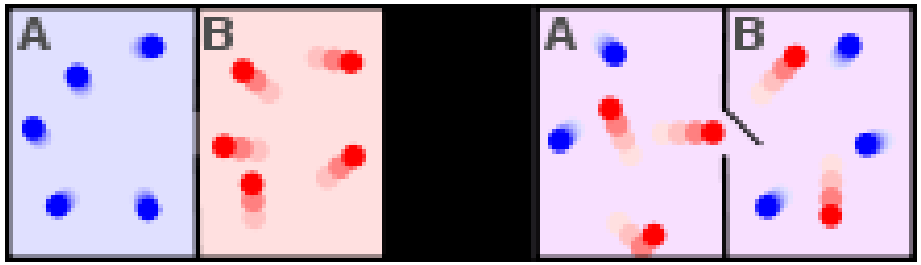
Ising models and the 2nd law of thermodynamics

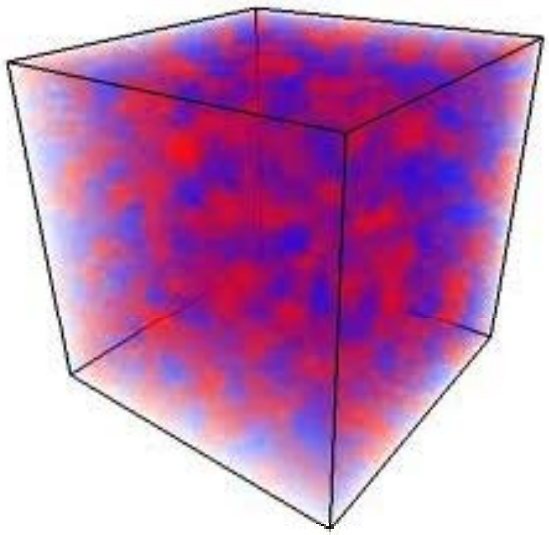
- **The Ising model describes many physical phenomena:**

  - "*The Ising model can be reinterpreted as a statistical model for the motion of atoms. A coarse model is to make space-time a lattice and imagine that each position either contains an atom or it doesn't.*" **Wikipedia Ising Model page.**

  - "*The original motivation for the model was the phenomenon of magnetism.*"

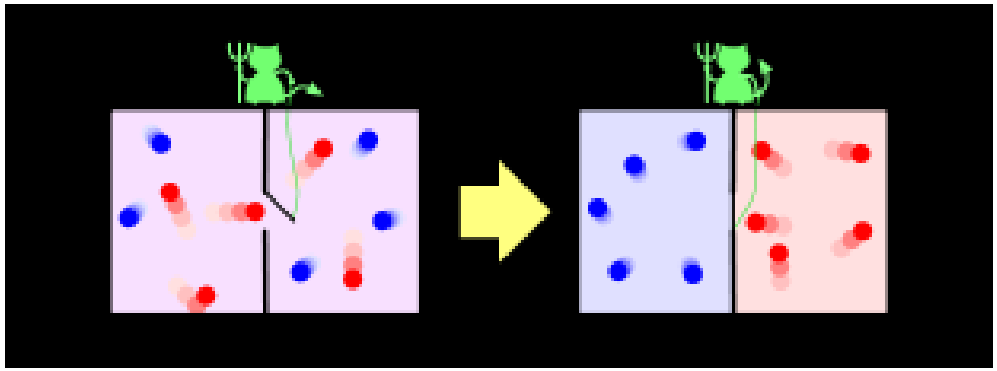
- **Second law of thermodynamics and stability.**





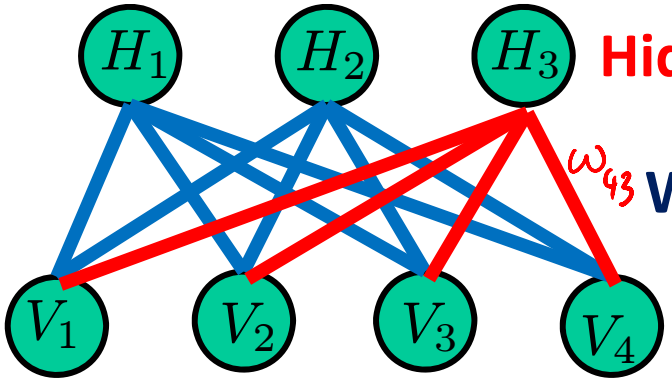
On information and energy – Maxwell's Demon

In this thought experiment, "*an imaginary container is divided into two parts by an insulated wall, with a door that can be opened and closed by what came to be called "Maxwell's Demon". The hypothetical demon is only able to let the "hot" molecules of gas flow through to a favored side of the chamber, causing that side to appear to spontaneously heat up while the other side cools down.*"



- **Does this violate the 2nd law?**
- **What is the relation of information and energy?**

Restricted Boltzmann Machines



Hidden variables (features)

Weighted connections form images

Visible (*e.g.* 4-pixel image)

A joint configuration $(\mathbf{v}, \mathbf{h})$ of the binary visible and hidden units has an energy given by the following RBM model:

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i \in pixels} b_i v_i - \sum_{j \in features} b_j h_j - \sum_{i,j} v_i w_{ij} h_j$$

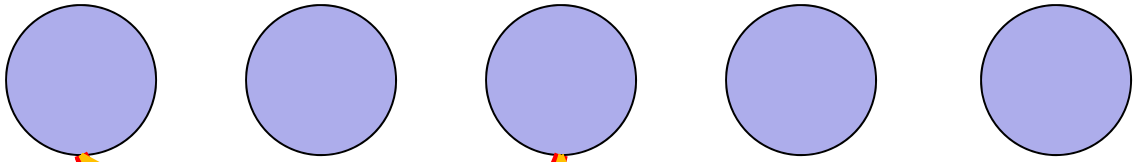
And hence a Boltzmann probability:

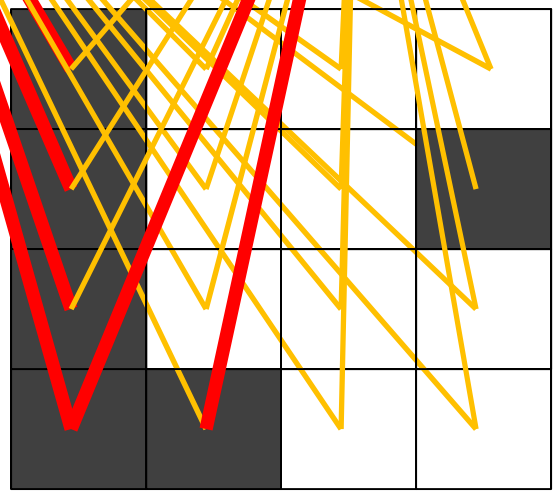
$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v}, \mathbf{h})}$$

Distributed representation

Feature vector **1** **0** **1** **0** **0**

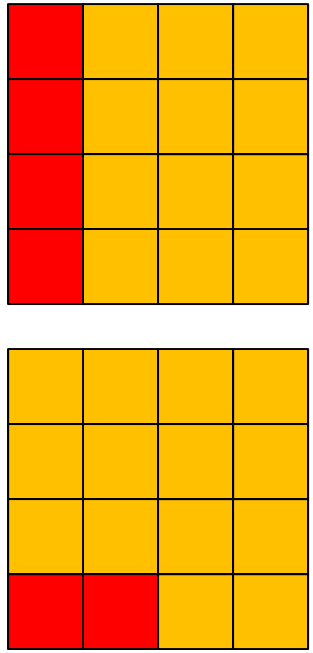
Hidden units





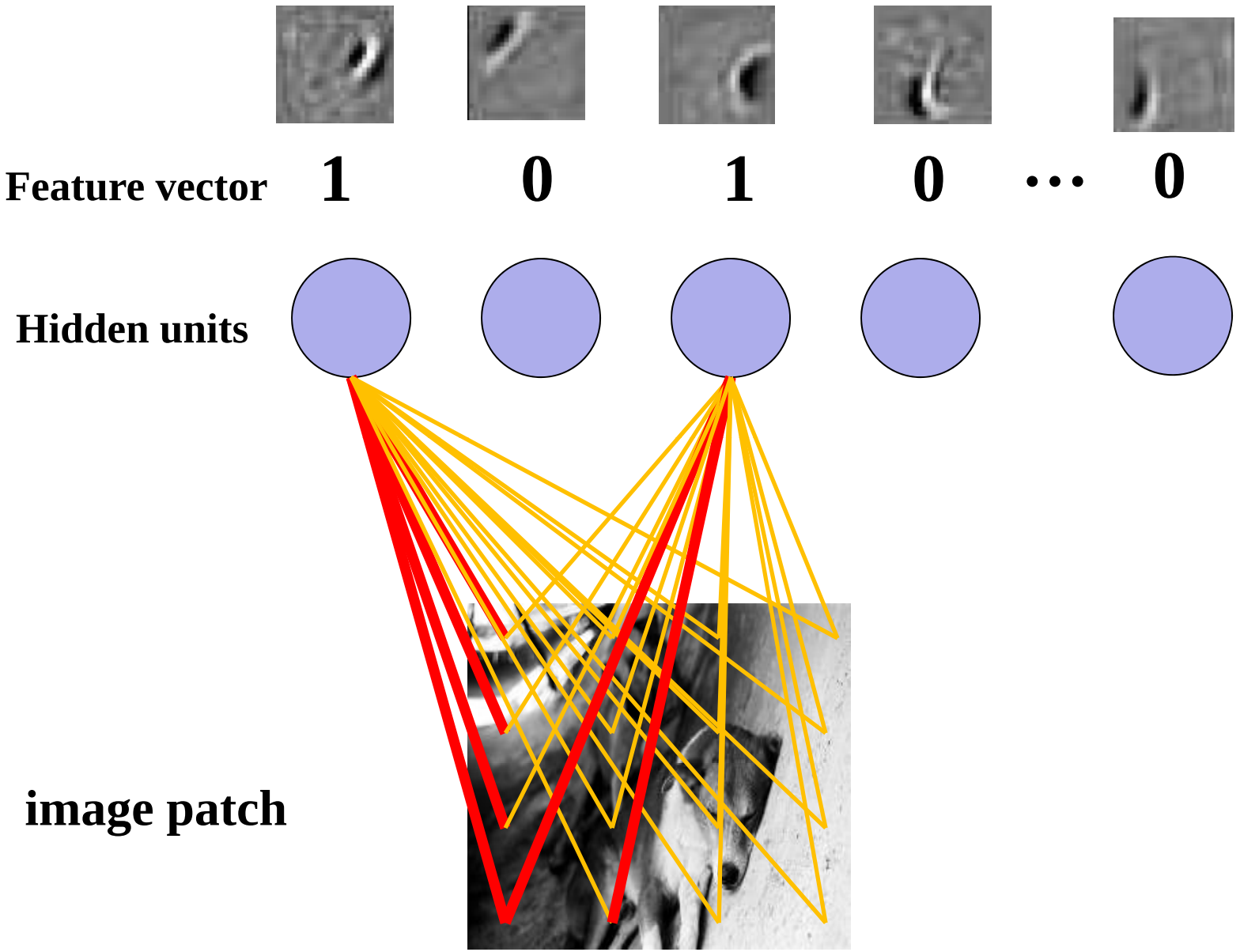
4x4 image patch

Learned weights

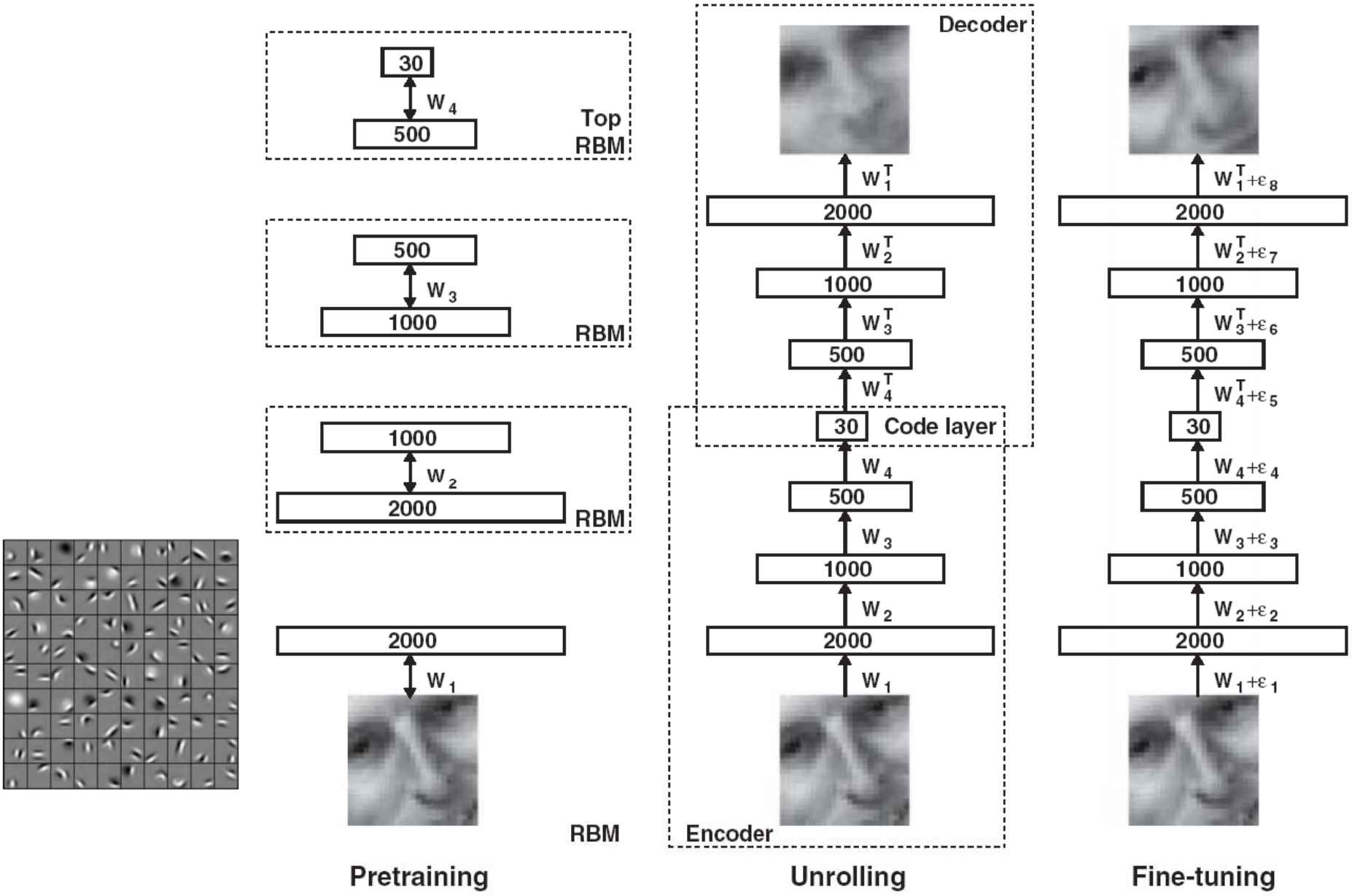


Insight: We're assuming edges occur often in nature, but dots don't

We learn the regular structures in the world

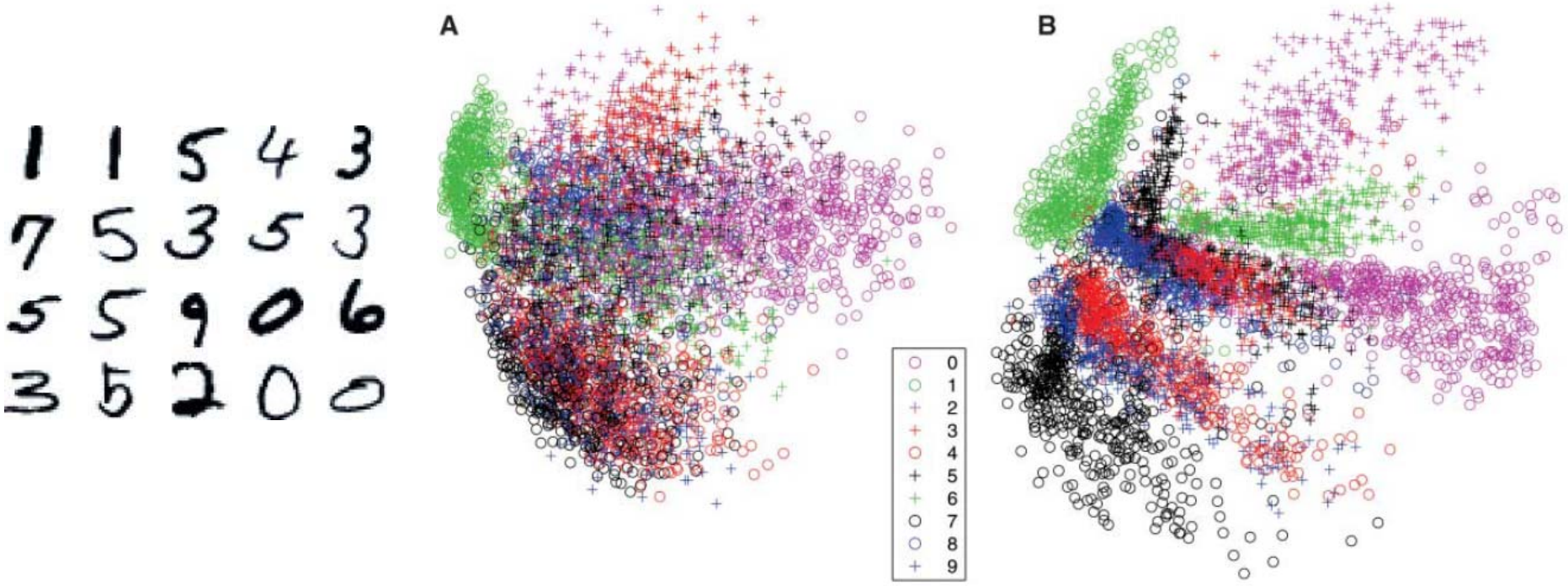

Feature vector
1
0
1
0
…
0
Hidden units
image patch

Deep learning (Hinton and collaborators)



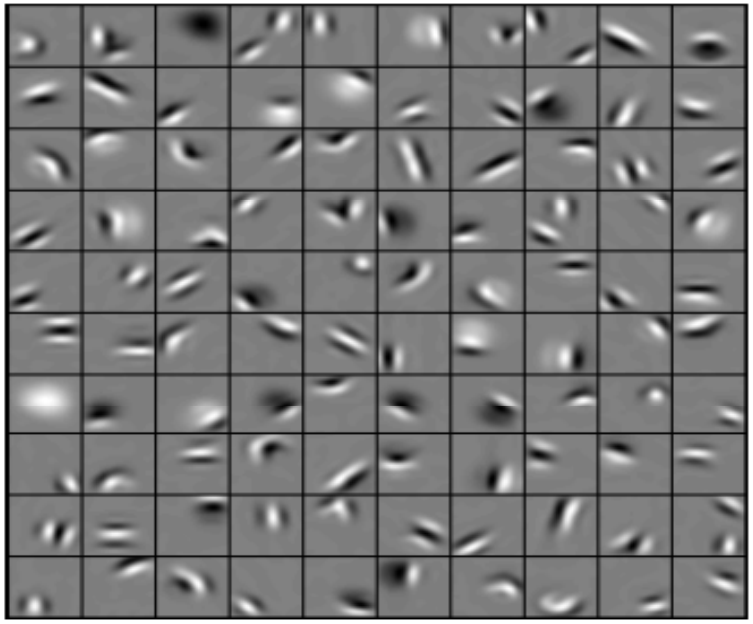
Encoding digits

(A) The two-dimensional codes for 500 digits of each class produced by taking the first two principal components of all 60,000 training images.
(B) The two-dimensional codes found by a 784-1000-500-250-2 autoencoder.



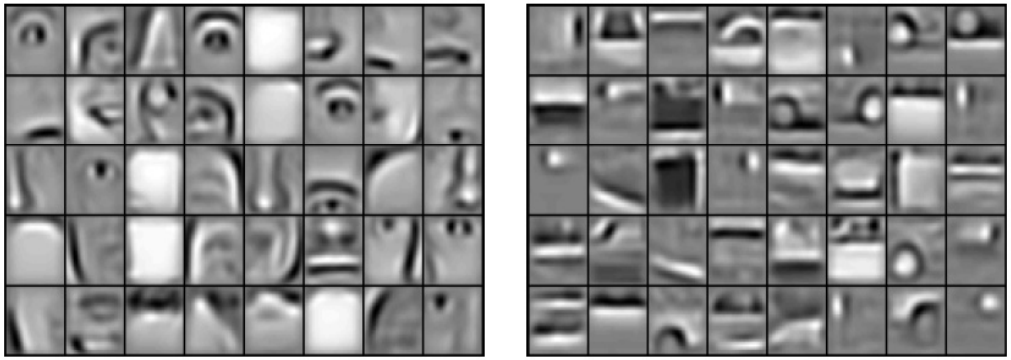
These 2-dimensional embeddings of images of digits enable us to make predictions (classification)

Layer 1

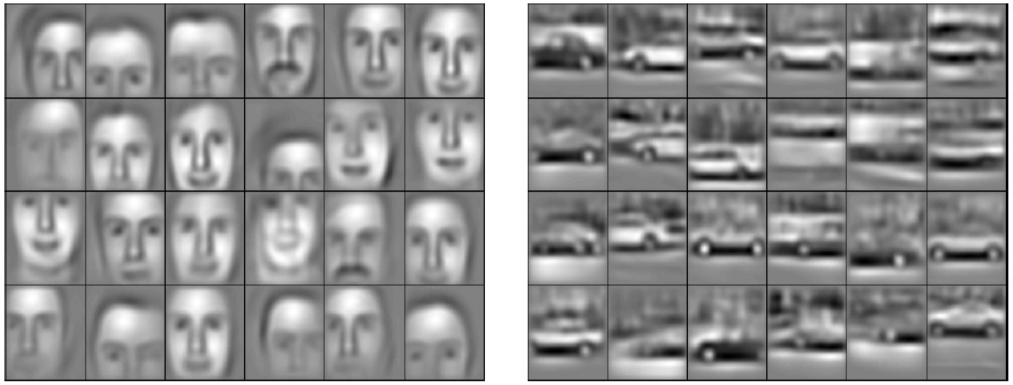


faces cars

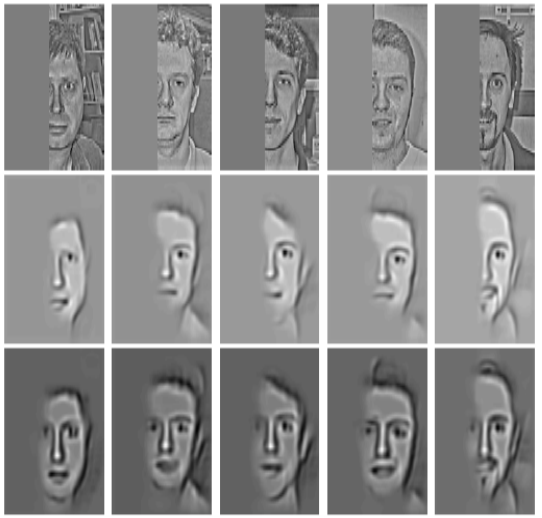
Layer 2



Layer 3



Completing scenes



[Honglak Lee et al 2009]

In the binary case where $v \in \{0,1\}^D$ and $h \in \{0,1\}^K$ the energy function can be expressed as:

$$E(v,h,W) = -\sum_{i=1}^{D}\sum_{j=1}^{K} v_i W_{ij} h_j - \sum_{i=1}^{D} v_i b_i - \sum_{j=1}^{K} h_j b_j.$$

The probabilities of each node can be easily obtained.

$$p(v_i = 1|h,W) = sigmoid\left(\sum_{j=1}^{K} W_{ij} h_j + b_i\right)$$

$$p(h_j = 1|v,W) = sigmoid\left(\sum_{i=1}^{D} W_{ij} v_i + b_j\right),$$

where $sigmoid(a) = \frac{1}{1+exp(-a)}$. The model is therefore easy to sample: One simply flips K coins for the hidden units and D coins for the visible units.

Contrastive divergence learning

1. Sample hidden units $\widetilde{h_n}$ from $p(h|v_n, W^{(t)})$.
2. Sample imaginary data $\widetilde{v_n}$ from $p(v|\widetilde{h_n}, W^{(t)})$.
3. Sample hidden units again $\widetilde{\widetilde{h_n}}$ from $p(h|\widetilde{v_n}, W^{(t)})$.
4. Update the parameters:

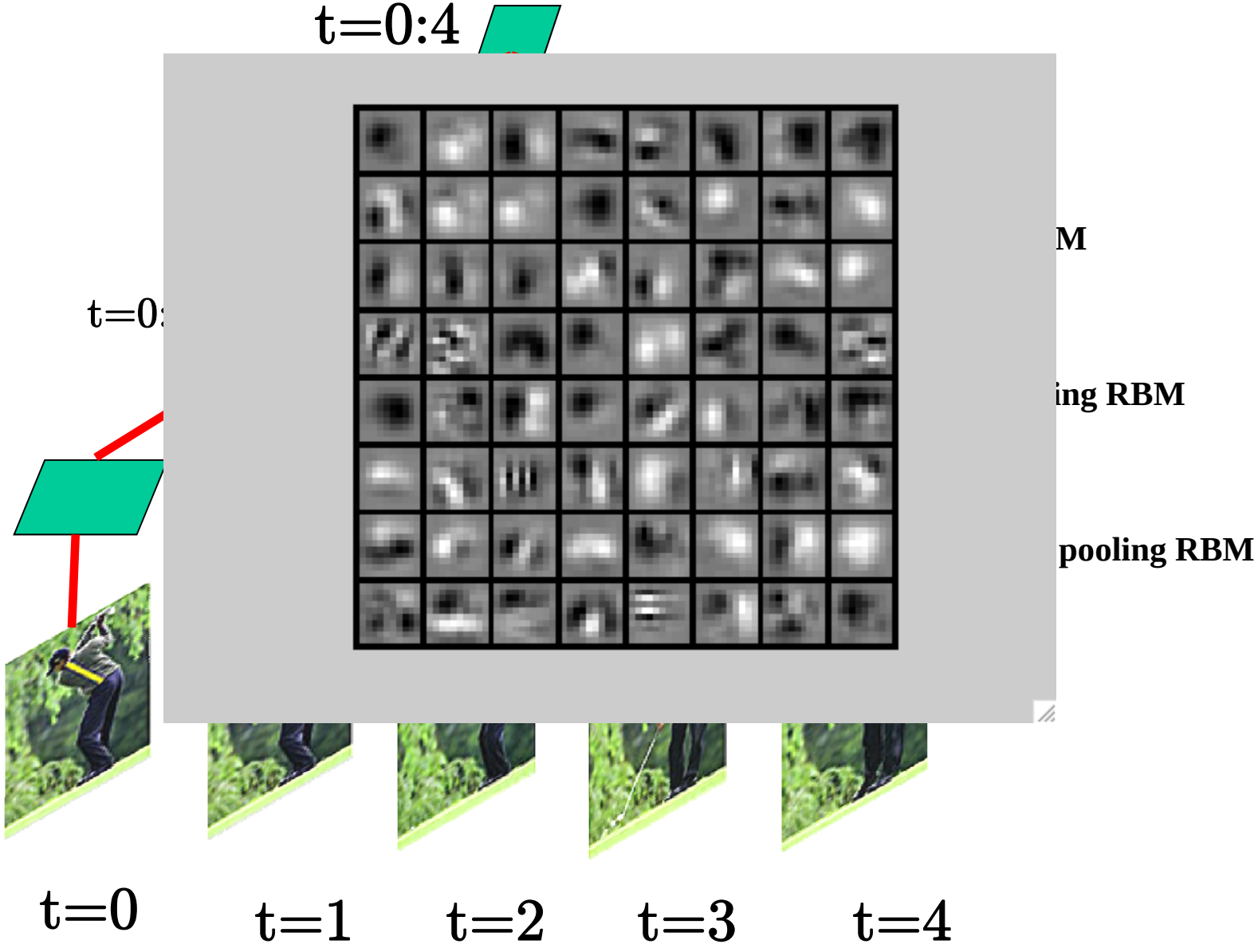
Real data

Confabulation

$$W_{dk}^{(t+1)} = W_{dk}^{(t)} + \eta^{(t)}\left[\frac{1}{N}\sum_{n=1}^{N} v_{dn}\widetilde{h_{kn}} - \frac{1}{N}\sum_{n=1}^{N} \widetilde{v_{dn}}\widetilde{\widetilde{h_{kn}}}\right]$$

5. Increase t to $t+1$ and go to step 2.

Hierarchical spatio-temporal feature learning



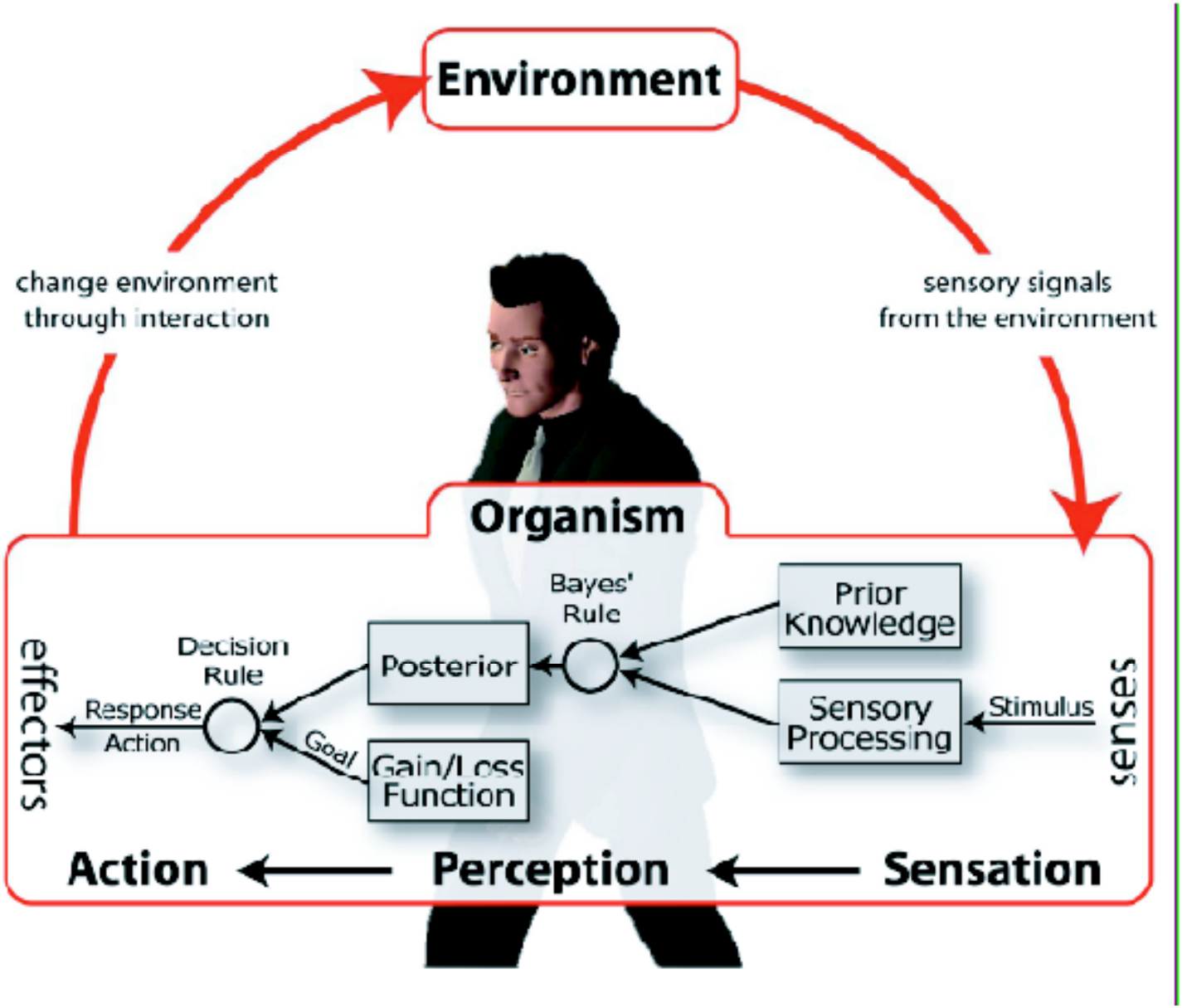
Hierarchical spatio-temporal feature learning

Observed gaze sequence

Model predictions

Change blindness

People as Bayesian reasoners

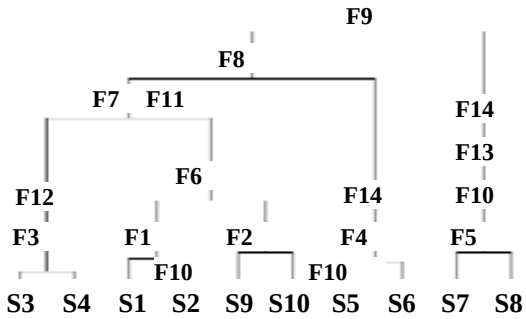


Theory

- **Species organized in taxonomic tree structure**
- **Feature *i* generated by mutation process with rate λ_i**

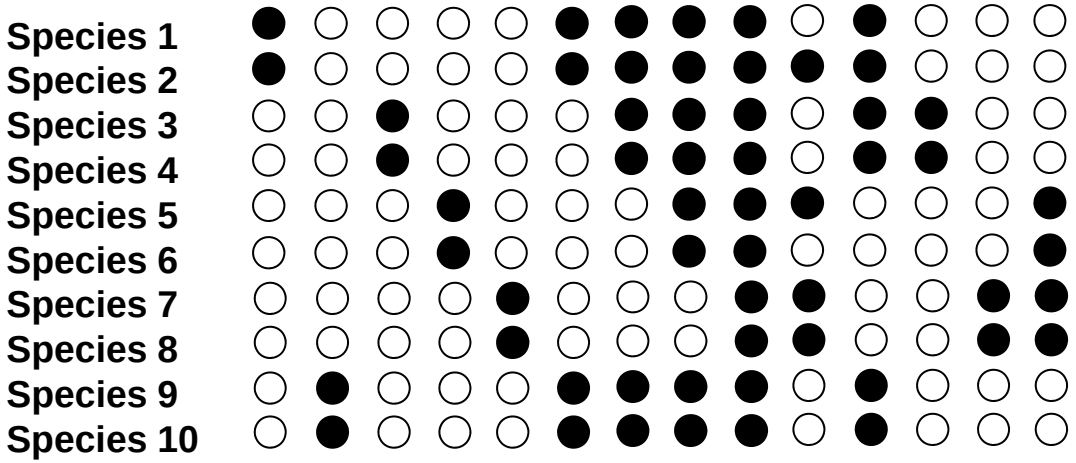
Domain Structure

$p(S|T)$



$p(D|S)$

Data



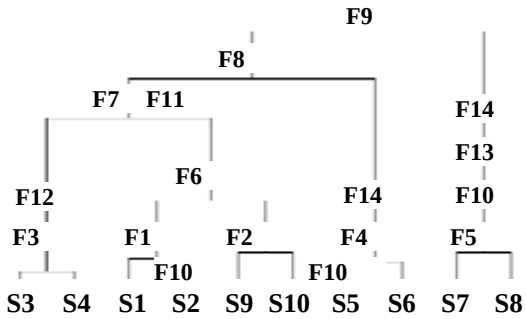
Josh Tenenbaum

Theory

- **Species organized in taxonomic tree structure**
- **Feature *i* generated by mutation process with rate λ_i**

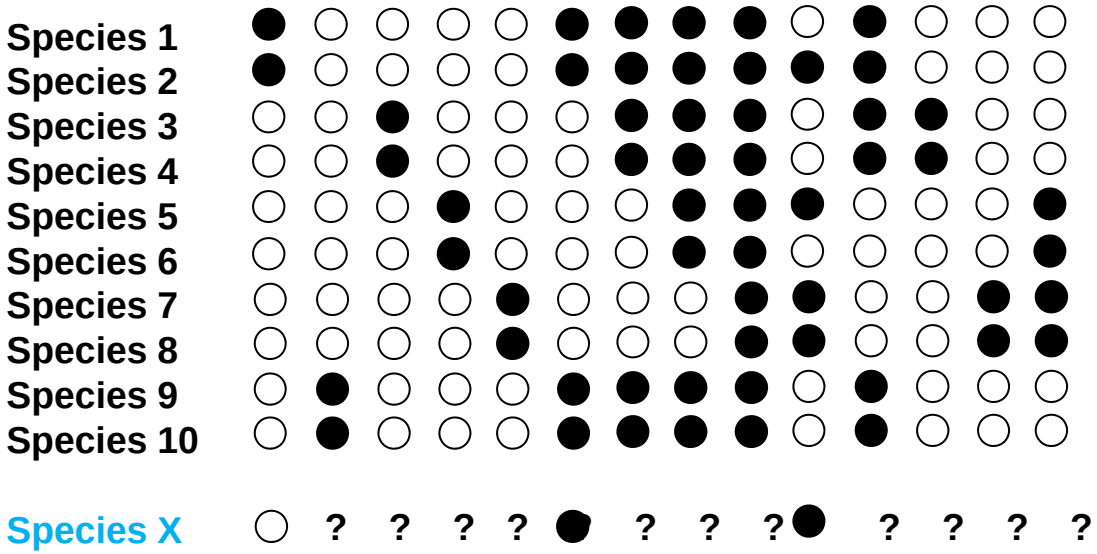
Domain Structure

$p(S|T)$



$p(D|S)$

Data



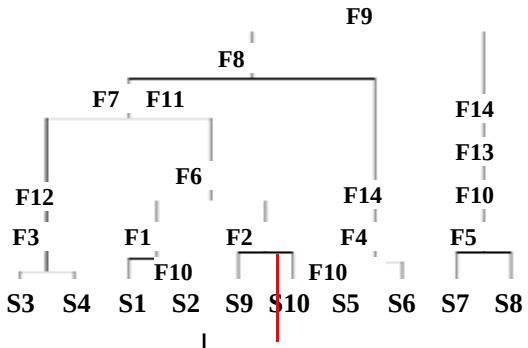
Josh Tenenbaum

Theory

- **Species organized in taxonomic tree structure**
- **Feature *i* generated by mutation process with rate λ_i**

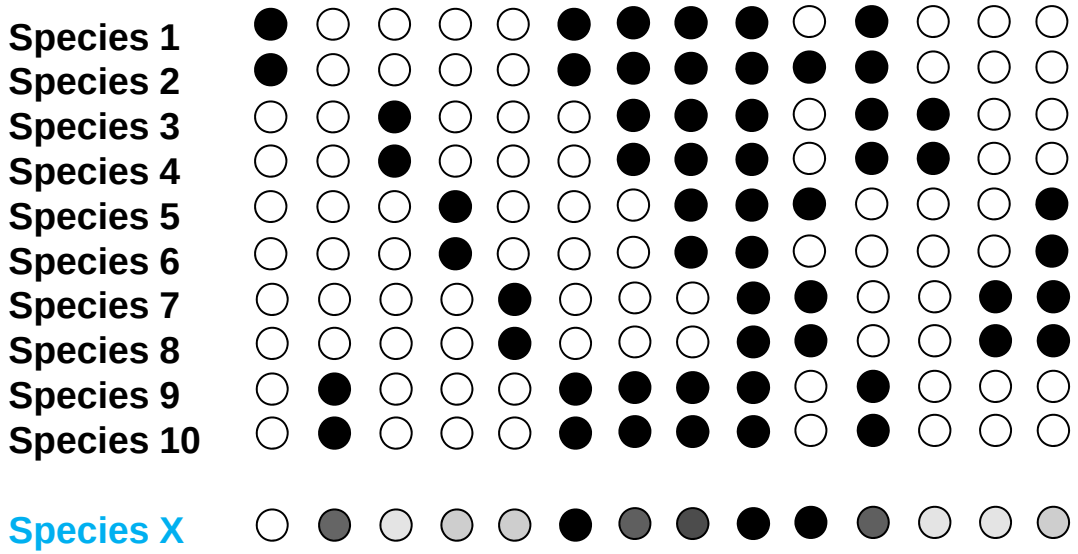
Domain Structure

$p(S|T)$



$p(D|S)$ SX

Data



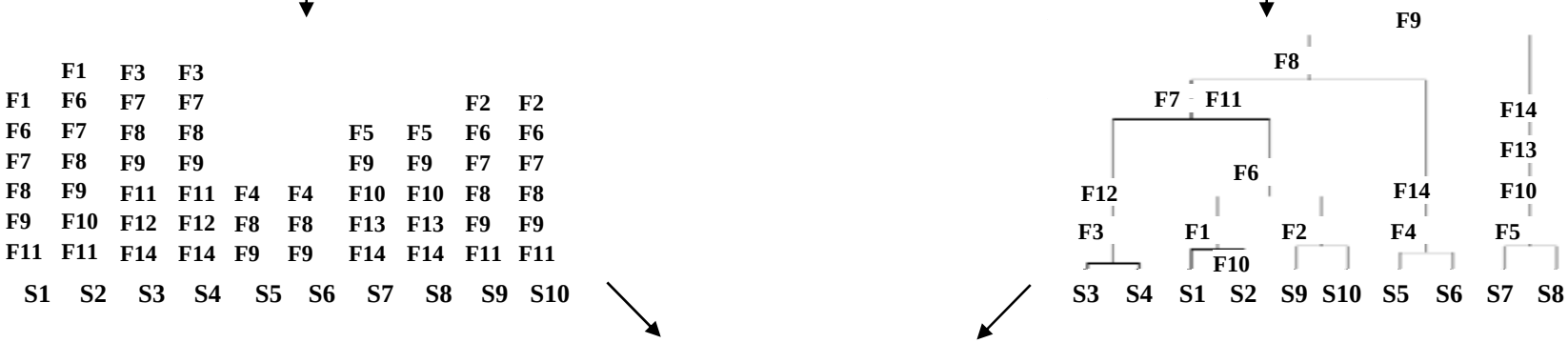
Josh Tenenbaum

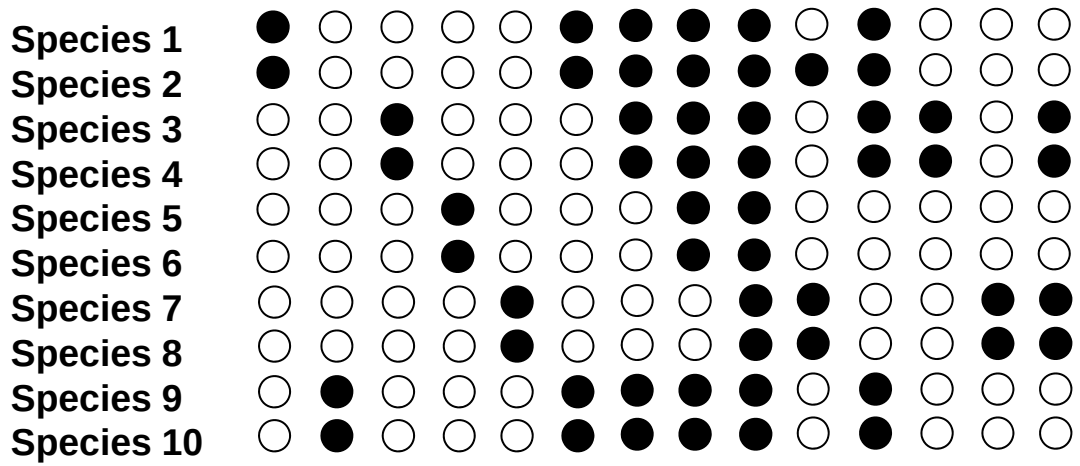
T0:

- **No organizational structure for species.**
- **Features distributed independently over species.**

T1:

- **Species organized in taxonomic tree structure.**
- **Features distributed via stochastic mutation process.**





Data

Features

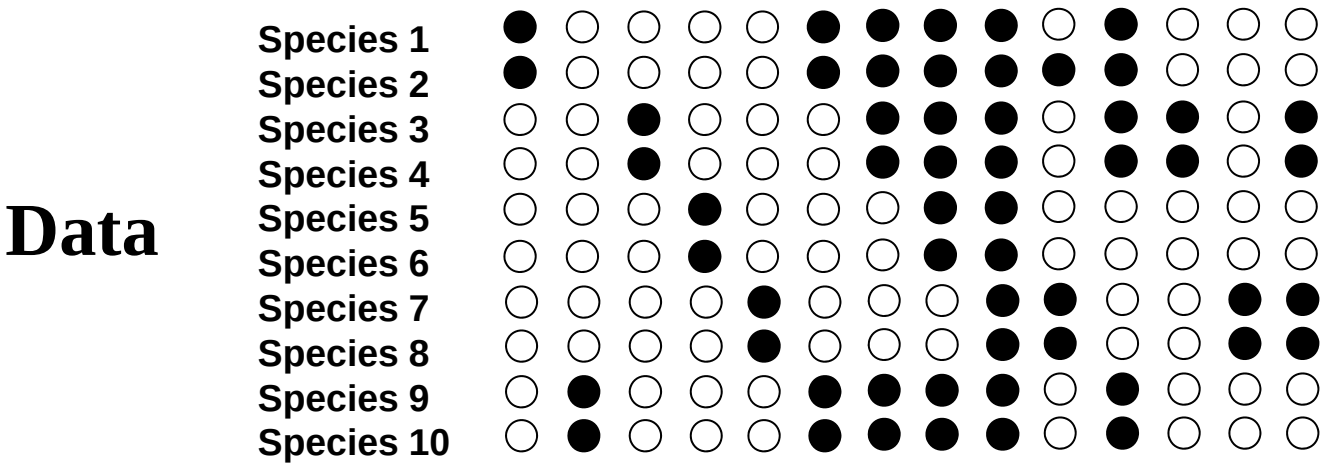
Josh Tenenbaum

T0: p(Data|T0) ~ 1.83 x 10^{-41}

- **No organizational structure for species.**
- **Features distributed independently over species.**

T1: p(Data|T1) ~ 2.42 x 10^{-32}

- **Species organized in taxonomic tree structure.**
- **Features distributed via stochastic mutation process.**

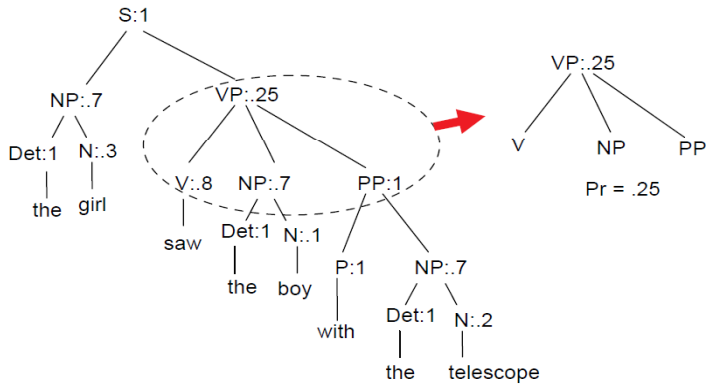


Josh Tenenbaum

(a)

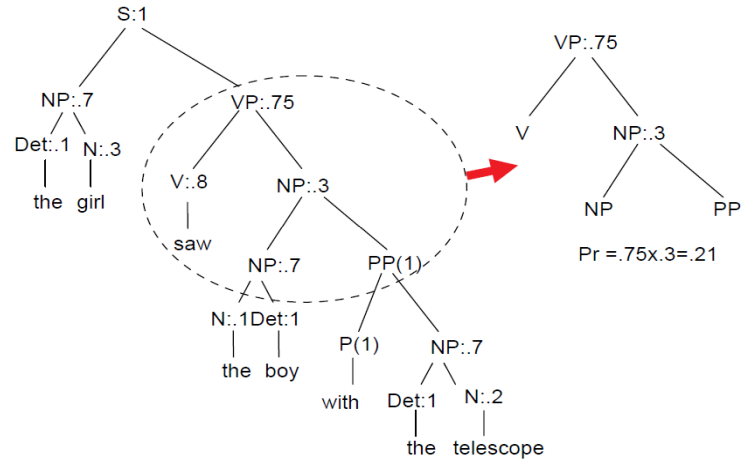
S→ NP VP	(1)	V→ saw	(.8)	N→ cat	(.1)
VP → V NP	(.75)	V → prodded	(.2	Det → the	(1)
VP → V NP PP	(.25)	N → telescope	(.2)	P → with	(1)
NP → Det Noun	(.7)	N → stick	(.3)		
NP → NP PP	(.3)	N → girl	(.3)		
PP → P NP	(1)	N → boy	(.1)		

(b)



Pr(tree) = 1 x .7 x 1 x .3 x .25 x .8 x .7 x 1 x .1 x 1 x 1 x .7 x 1 x .2 ≈ 0.00041

(c)



Pr(tree) = 1 x .7 x 1 x .3 x .75 x .8 x .3 x .7 x 1 x .1 x 1 x 1 x .7 x 1 x .2 ≈ 0.00037

(Chater & Manning, 2006)

“Universal Grammar”

↓ ***P*(grammar | UG)**

Grammar

↓ ***P*(phrase structure | grammar)**

Phrase structure

↓ ***P*(utterance | phrase structure)**

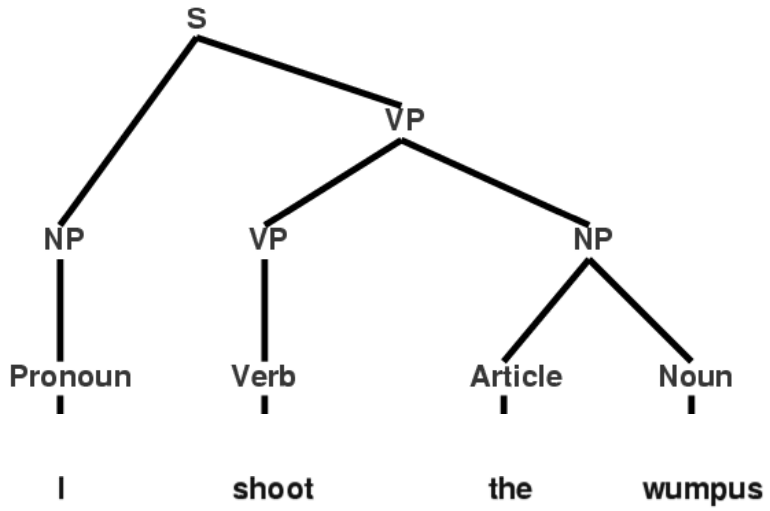
Utterance

↓ ***P*(speech | utterance)**

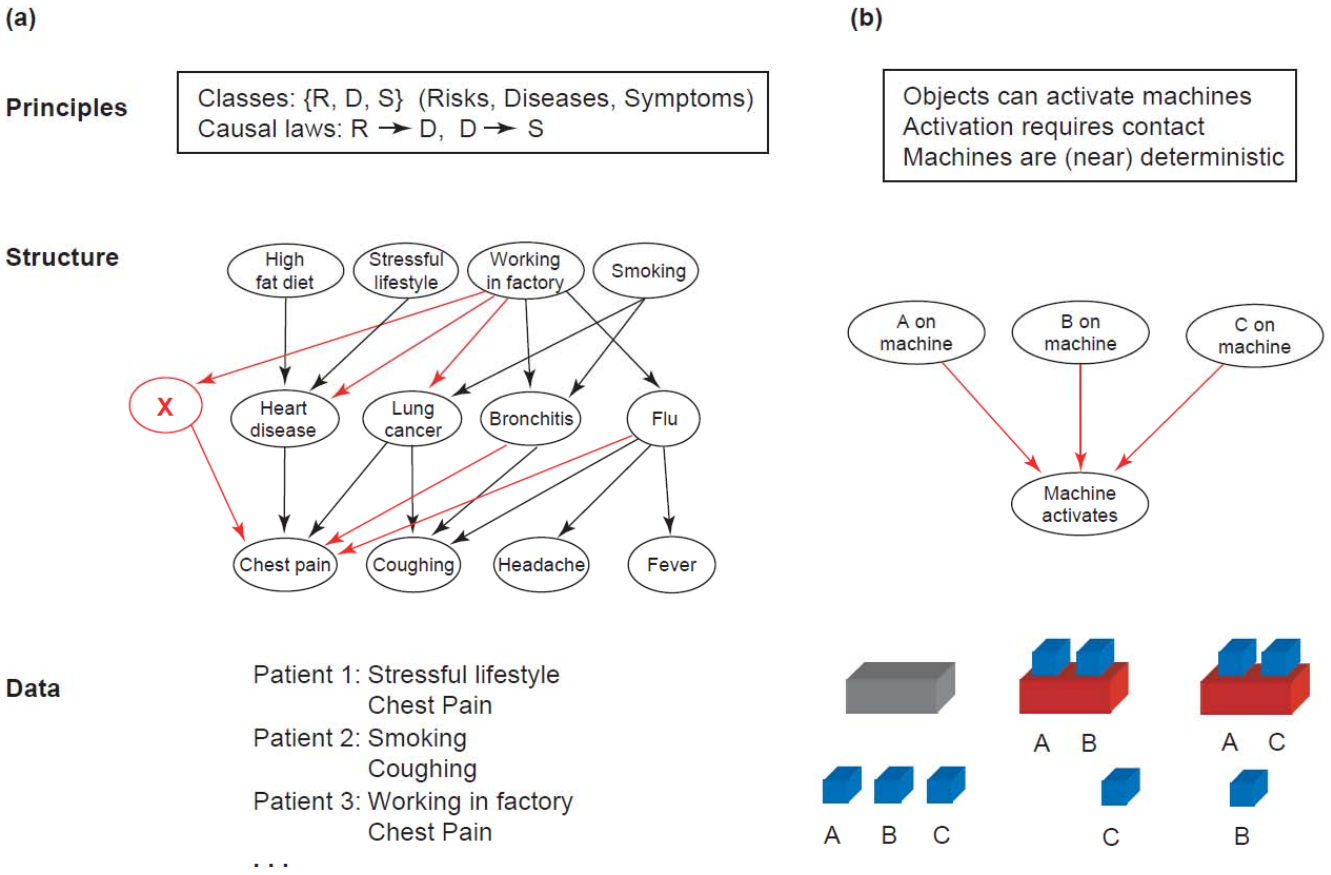
Speech signal

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

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



Josh Tenenbaum



Josh Tenenbaum