CPSC 340
Machine Learning & Data Mining

Nando de Freitas

http://www.cs.ubc.ca/~nando/340-2008

Acknowledgement

Many thanks to the following people for making available some of the slides, figures and videos used in these slides:

- Kevin Murphy (UBC)
- Kevin Leyton-Brown (UBC)
- Tom Griffiths (Berkeley)
- Josh Tenenbaum (MIT)
- Kobus Barnard (Arizona)
- All my awesome students at UBC

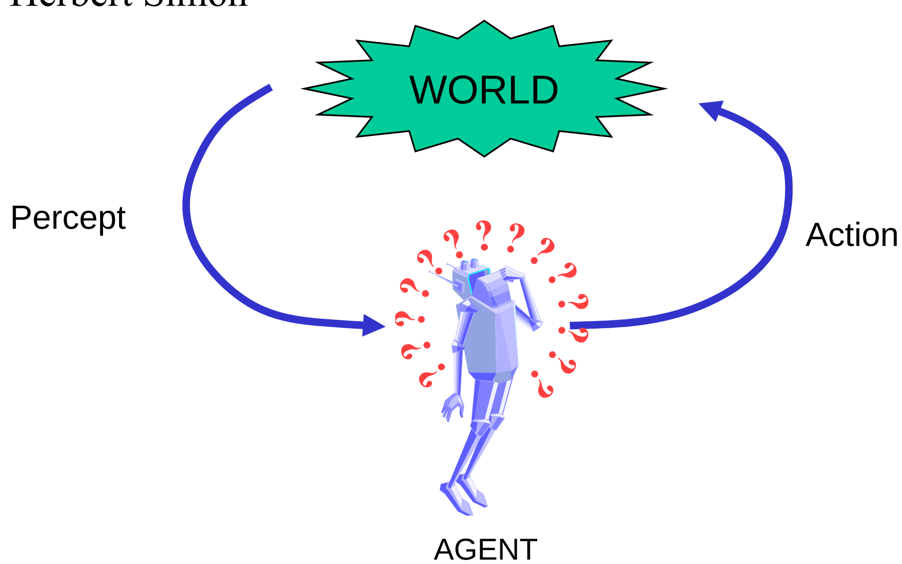
Introduction to machine learning

- What is machine learning?
- How is machine learning related to other fields?
- Machine learning applications
- Types of learning
  - Supervised learning
    - regression
    - classification
  - Unsupervised learning
    - clustering
    - data association
    - abnormality detection
    - dimensionality reduction
    - structure learning
  - Semi-supervised learning
  - Active learning
  - Reinforcement learning and control of partially observed Markov decision processes.

What is machine learning?

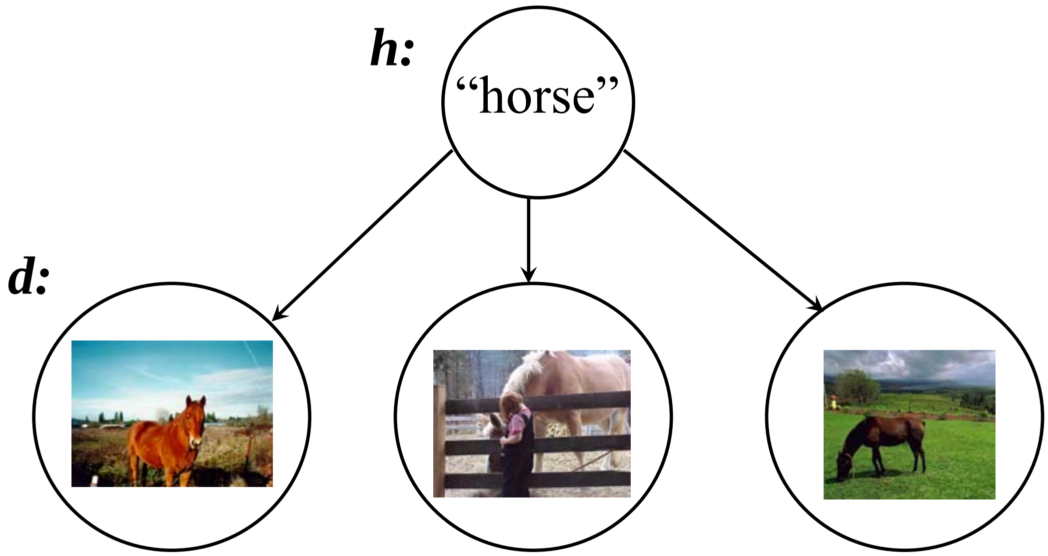
``*Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the task or tasks drawn from the same population more efficiently and more effectively the next time."*

-- Herbert Simon

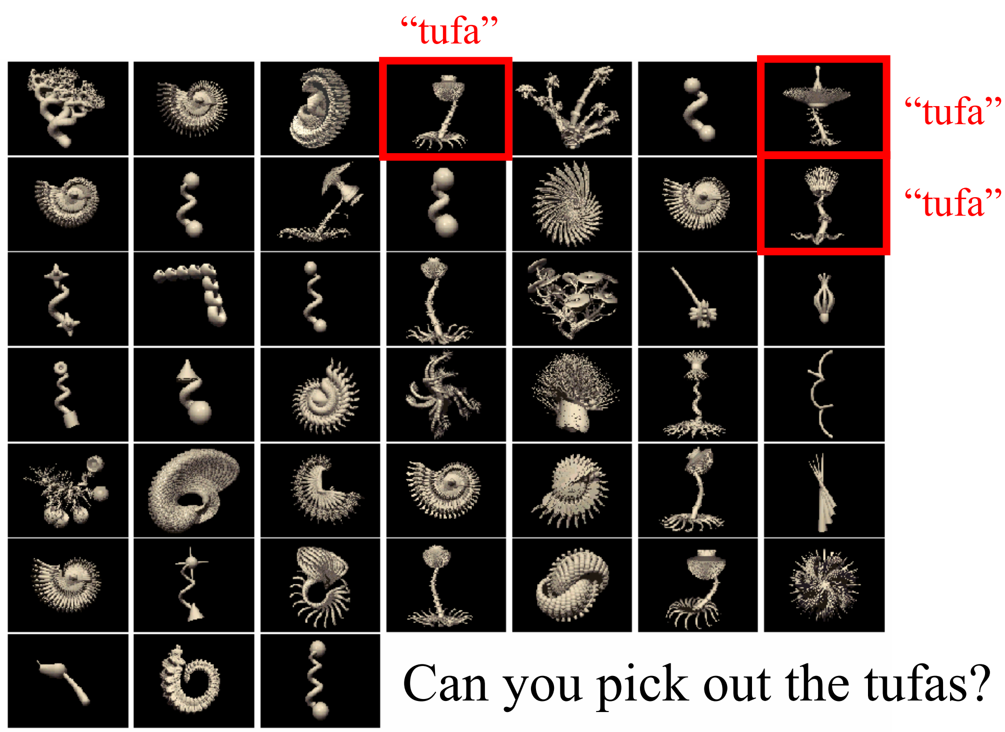


What is machine learning?

Machine learning is concerned with the process of constructing abstractions of the real world (concepts, functions, relations and ways of acting) automatically from observations.



Learning concepts and words



Can you pick out the tufas?

Source: *Josh Tenenbaum*

Why “Learn” ?

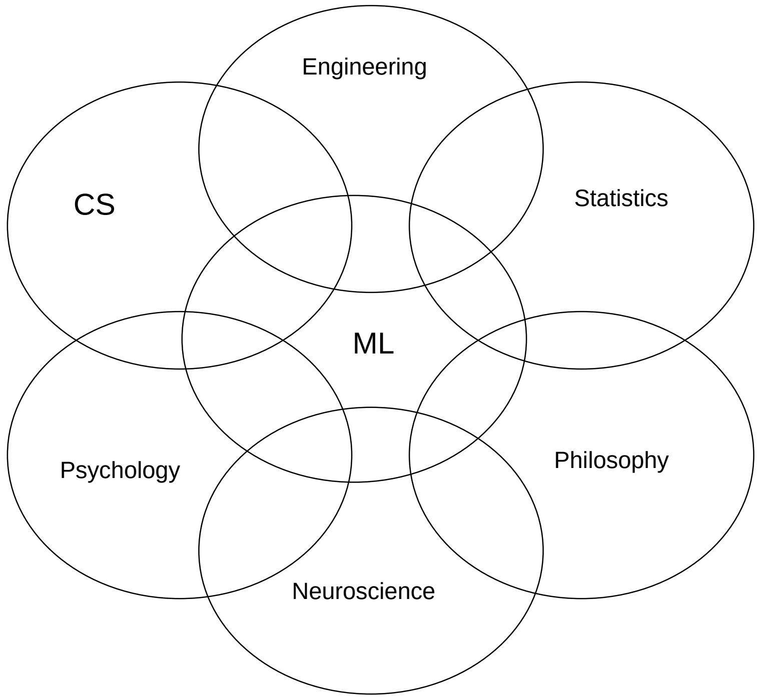
Learning is used when:

- Human expertise is absent (navigating on Mars)
- Humans are unable to explain their expertise (speech recognition, vision, language)
- Solution changes in time (routing on a computer network)
- Solution needs to be adapted to particular cases (user biometrics)
- The problem size is to vast for our limited reasoning capabilities (calculating webpage ranks)

Introduction to machine learning

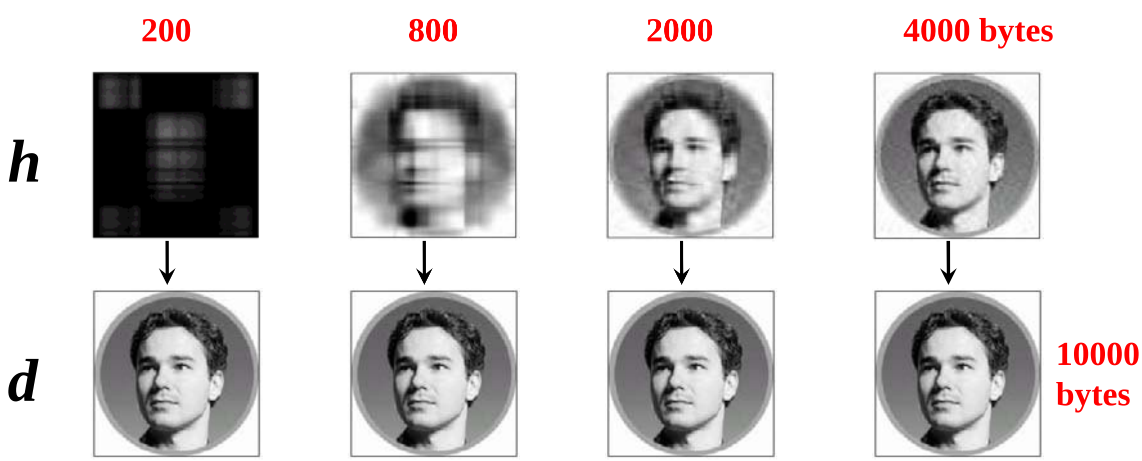
- What is machine learning?
- How is machine learning related to other fields?
- Machine learning applications
- Types of learning
  - Supervised learning
    - regression
    - classification
  - Unsupervised learning
    - clustering
    - data association
    - abnormality detection
    - dimensionality reduction
    - structure learning
  - Semi-supervised learning
  - Active learning
  - Reinforcement learning and control of partially observed Markov decision processes.

How is machine learning related to other fields?



Learning and information theory

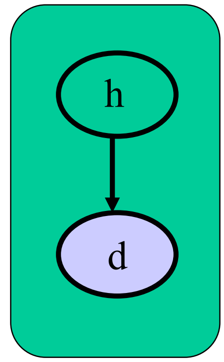
Data compression and transmission over a noisy channel provide some insight into the process of learning

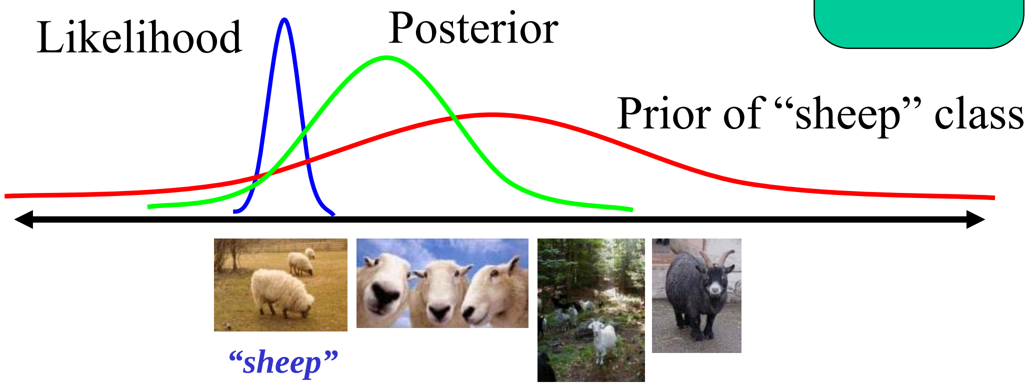


(i) Which compressions capture the essence of the image?
(ii) Which one is best to recognize the same subject in a different photo?

Learning and Bayesian inference

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





Speech recognition

P(words | sound) α P(sound | words) P(words)

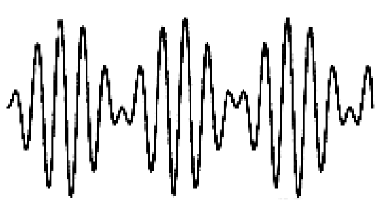
Final beliefs

Likelihood of data
eg mixture of Gaussians

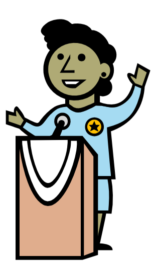
Language model
eg Markov model

Hidden Markov Model (HMM)

"Recognize speech"



"Wreck a nice beach"

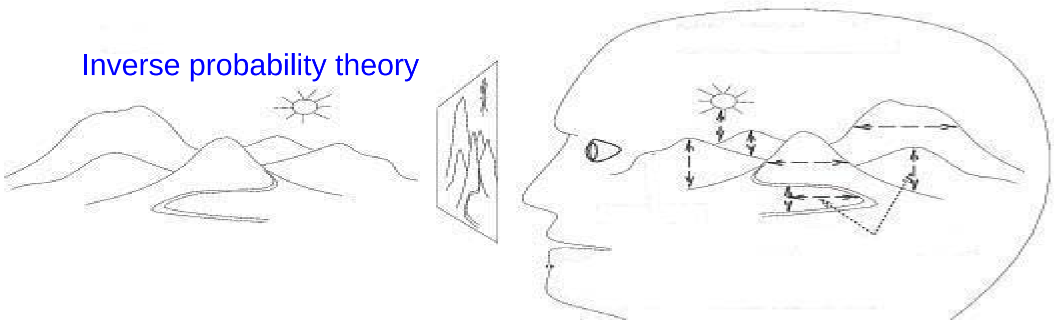


Vision as inverse graphics

p(world | image) α p(image |world) p(world)

Final beliefs — Likelihood of data — Initial beliefs

Inverse probability theory



World — Image — Beliefs about world

Learning, decision theory and control

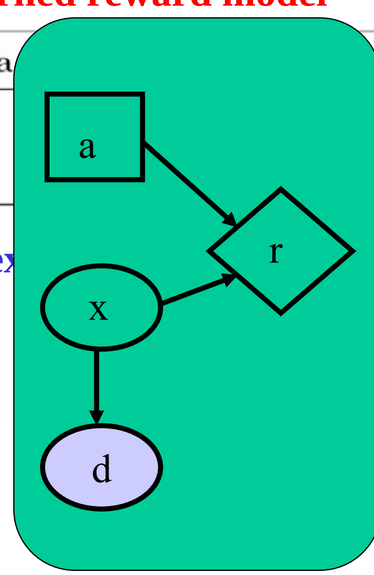
Utilitarian view: We need models to make the right decisions under uncertainty. Inference and decision making are intertwined.

Learned population model

$$\begin{cases} p(\mathbf{x} = healthy) = 0.9 \\ p(\mathbf{x} = cancer) = 0.1 \end{cases}$$

Learned reward model

	a	...tment
$\mathbf{x} = healthy$		
$\mathbf{x} = cancer$		

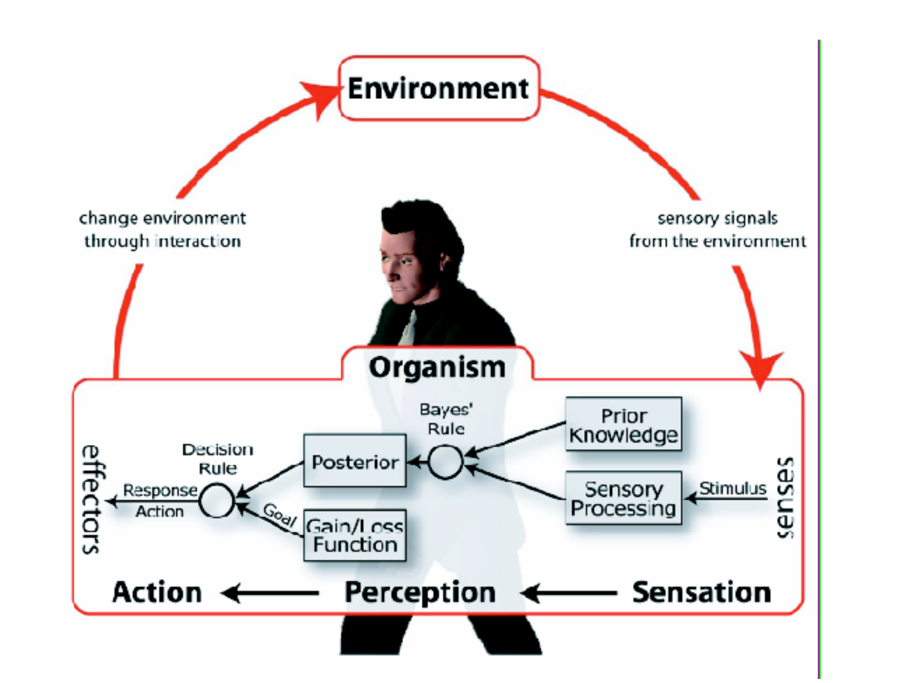


We choose the action that maximizes the ex

$$EU(\mathbf{a}) = \sum_{\mathbf{x} \in \{healthy, cancer\}} r(\mathbf{x}, \mathbf{a})\, p(\mathbf{x})$$

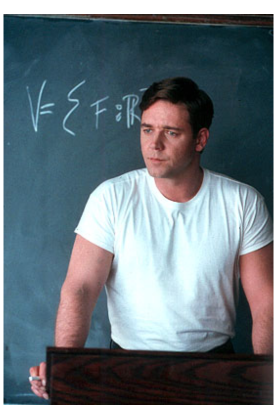
$$EU(\mathbf{a} = treatment) = -27.2$$
$$EU(\mathbf{a} = no\ treatment) = -10$$

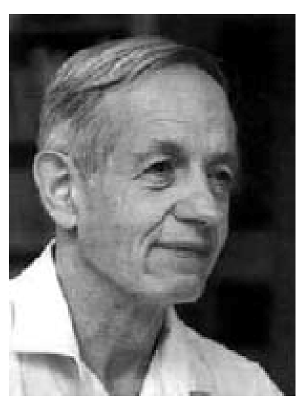
People as Bayesian reasoners



Learning and expected utility are related to game theory

- Learning opponents' policies
- Language acquisition, evolution and processing





Introduction to machine learning

- What is machine learning?
- How is machine learning related to other fields?
- Machine learning applications
- Types of learning
  - Supervised learning
    - regression
    - classification
  - Unsupervised learning
    - clustering
    - data association
    - abnormality detection
    - dimensionality reduction
    - structure learning
  - Semi-supervised learning
  - Active learning
  - Reinforcement learning and control of partially observed Markov decision processes.

Chess

- In 1996 and 1997, Gary Kasparov, the world chess grandmaster played two tournaments against Deep Blue, a program written by researchers at IBM

Source: *IBM Research*

Deep Blue's Results in the first tournament:

won 1 game, lost 3 and tied 1

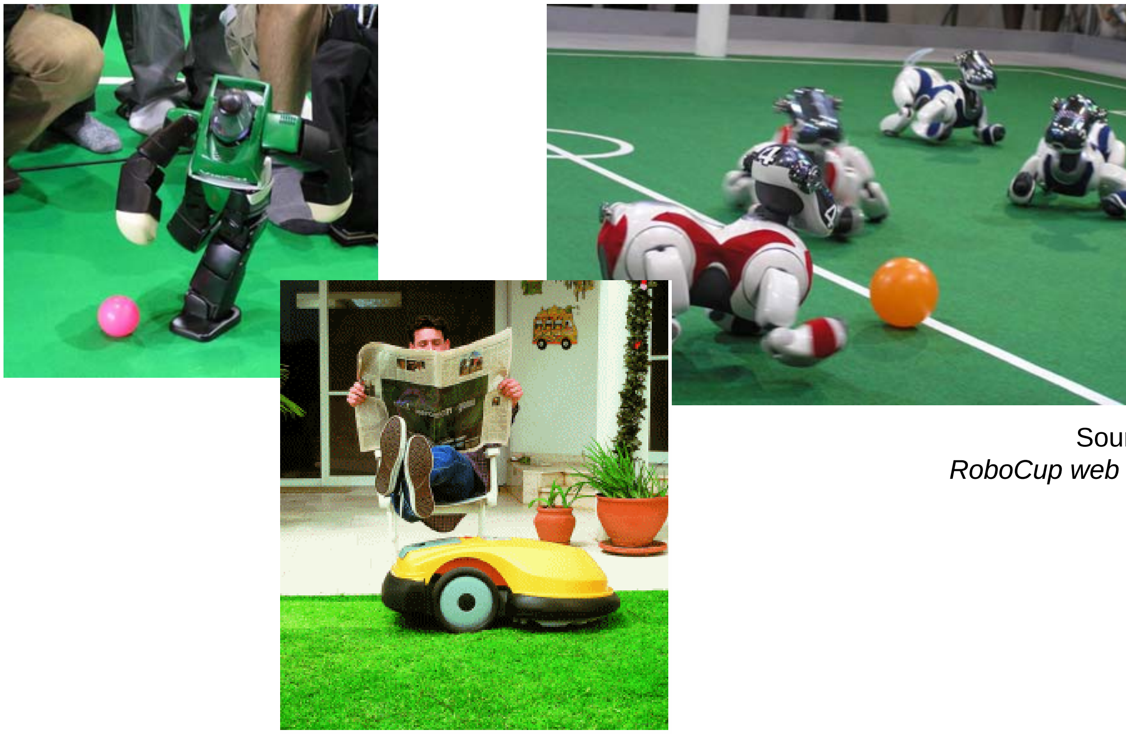
- first time a reigning world champion lost to a computer
- although Kasparov didn't see it that way…

Source: *CNN*

Deep Blue's Results in the second tournament:

– second tournament: won 3 games, lost 2, tied 1

Learning is essential to building autonomous robots

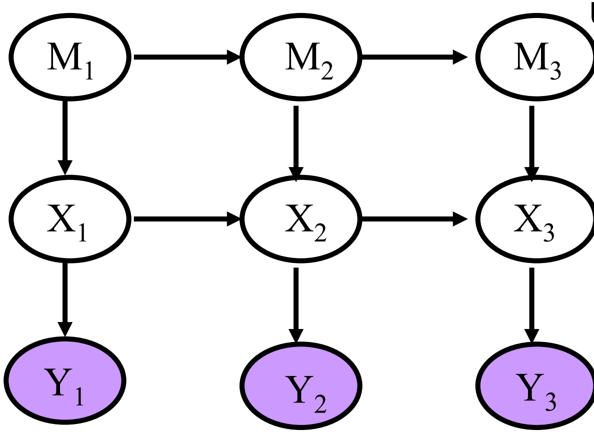


Source:
RoboCup web site

Autonomous robots and self-diagnosis

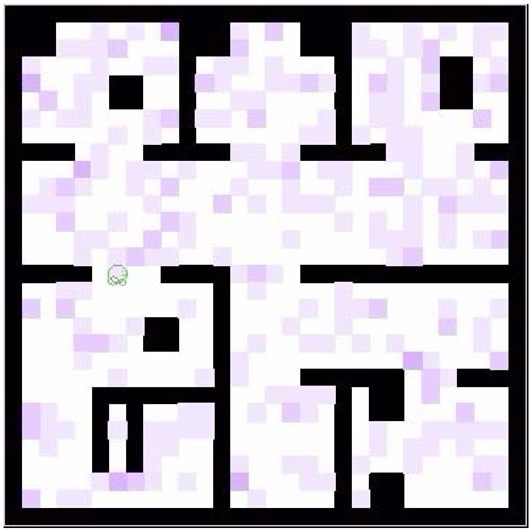
Unknown internal discrete state

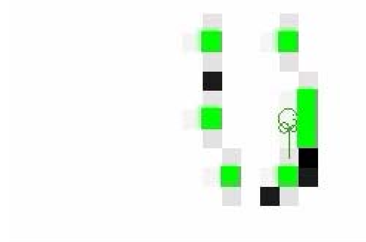
Unknown continuous signals



Sensor readings

Simultaneous localization and map learning

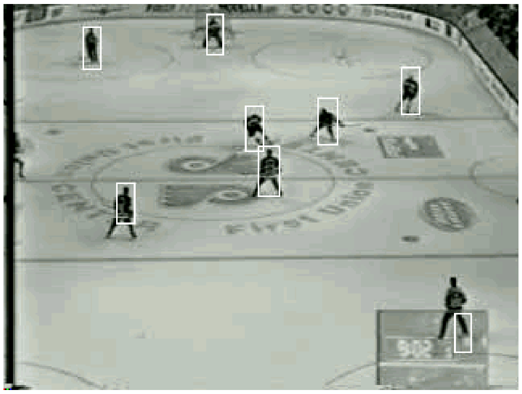


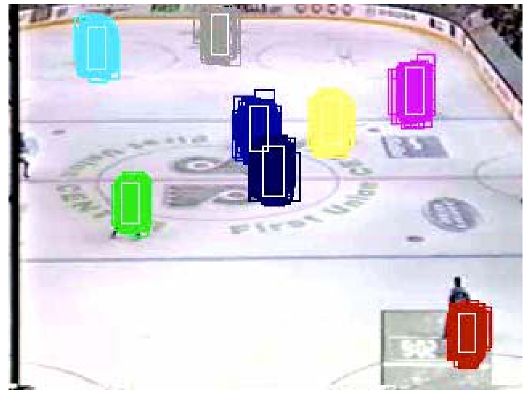


Robots that learn to drive

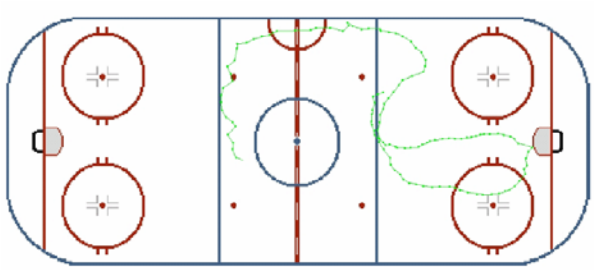
Source:
Sebastian Thrun

Tracking and activity recognition

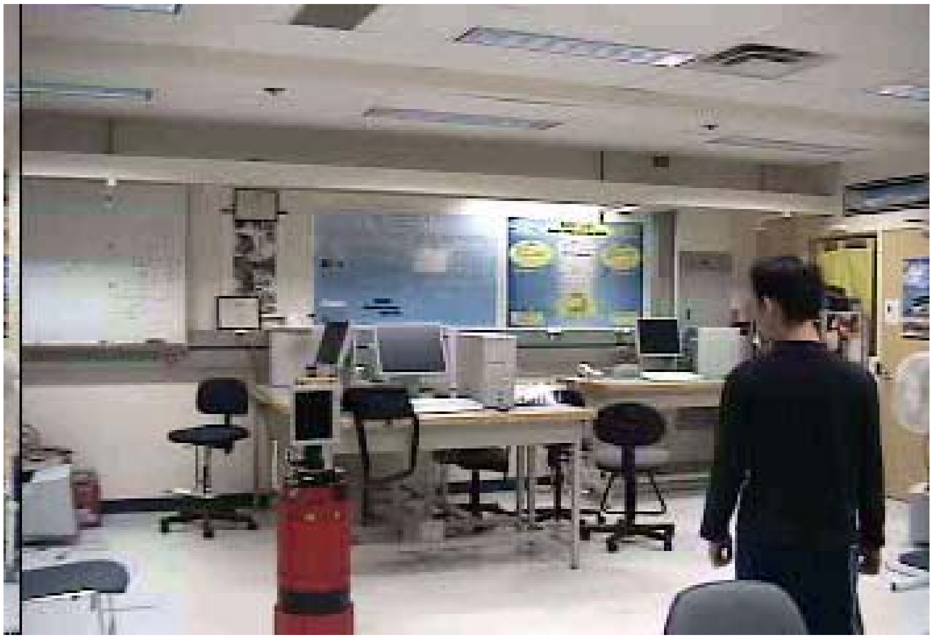




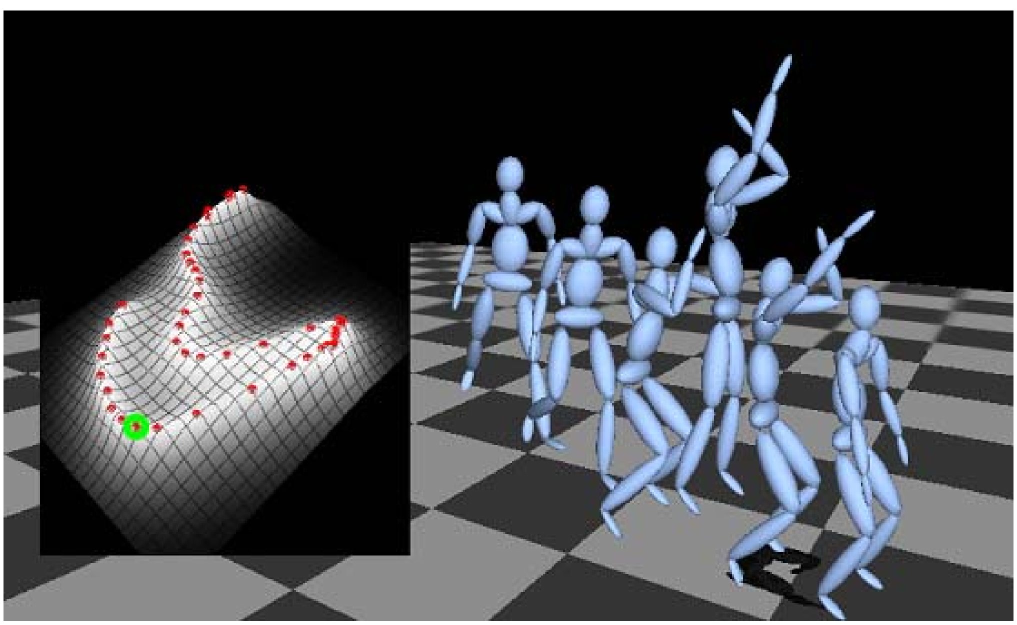
Data mining and games



Tracking robots



Animation and control



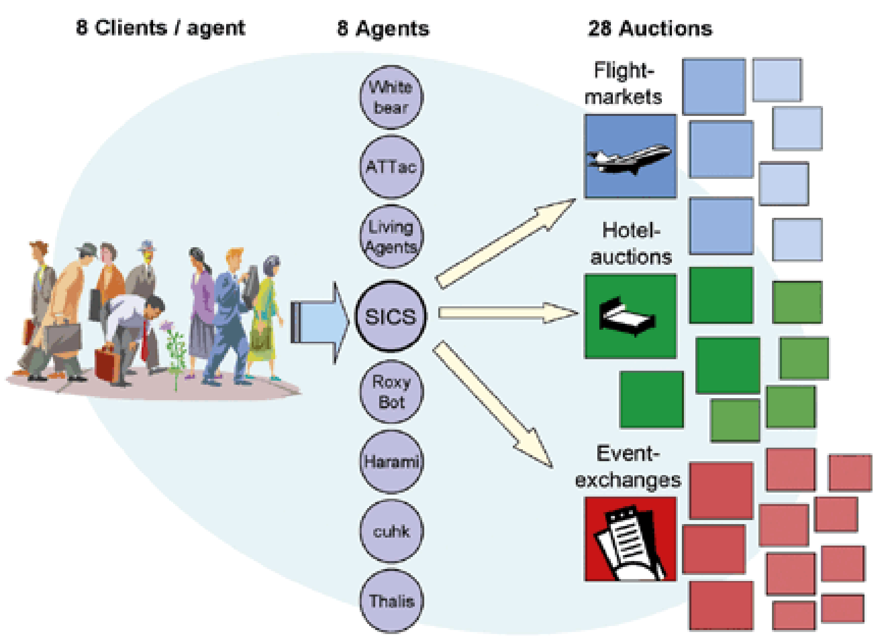
Source: *Aaron Hertzmann*

Learning agents that play poker

- In full 10-player games Poki is better than a typical low-limit casino player and wins consistently; however, not as good as most experts
- New programs being developed for the 2-player game are quite a bit better, and we believe they will very soon surpass all human players

Source: *The University of Alberta GAMES Group*

Learning web-bots



Source: *Swedish Institute of Computer Science*

“Universal Grammar”

↓ *P*(grammar | UG)

Grammar

↓ *P*(phrase structure | grammar)

Phrase structure

↓ *P*(utterance | phrase structure)

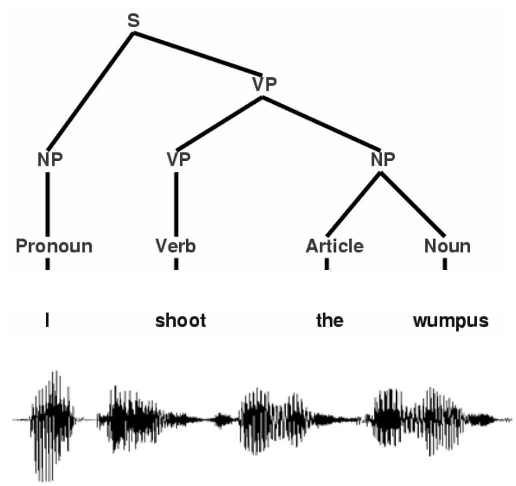
Utterance

↓ *P*(speech | utterance)

Speech signal

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

$$\begin{aligned} S &\rightarrow NP\ VP \\ NP &\rightarrow Det\,[Adj]\,Noun\,[RelClause] \\ RelClause &\rightarrow [Rel]\,NP\ V \\ VP &\rightarrow VP\ NP \\ VP &\rightarrow Verb \end{aligned}$$



Source: Tenenbaum

Natural language understanding

- P(meaning | words) α P(words | meaning) P(meaning)
- We do not yet know good ways to represent "meaning" (knowledge representation problem)
- Most current approaches involve "shallow parsing", where the meaning of a sentence can be represented by fields in a database, eg
  - "Microsoft acquired AOL for \$1M yesterday"
  - "Yahoo failed to avoid a hostile takeover from Google"

Buyer	Buyee	When	Price
MS	AOL	Yesterday	$1M
Google	Yahoo	?	?

Structure learning: Phylogenetic Tree Reconstruction (Nir Friedman et al.)

Input: Biological sequences

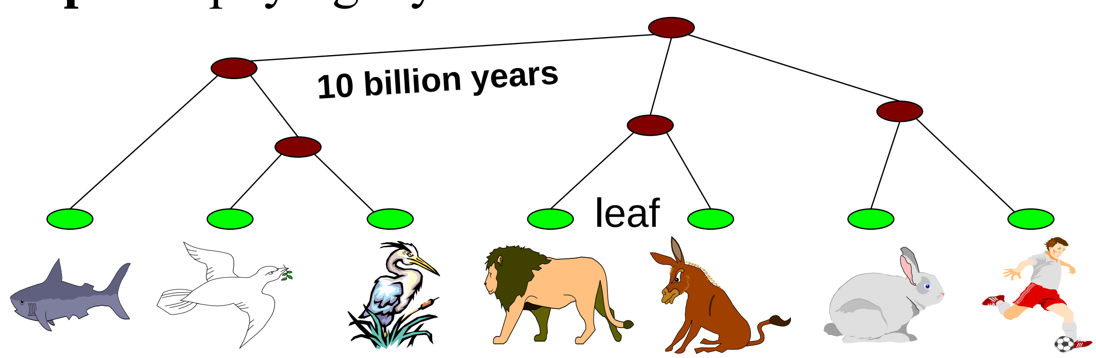
Human CGTTGC…

Chimp CCTAGG…

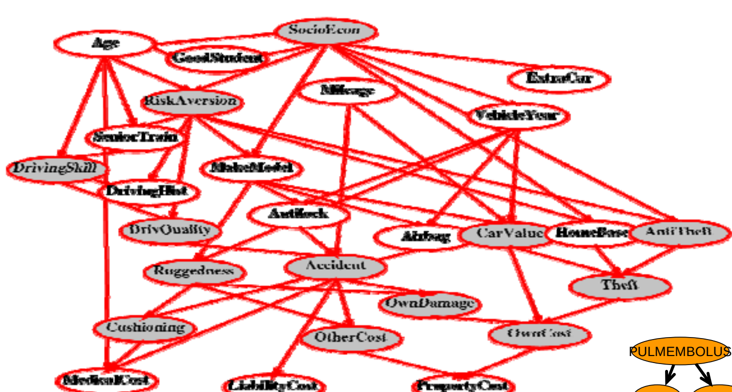
Orang CGAACG…

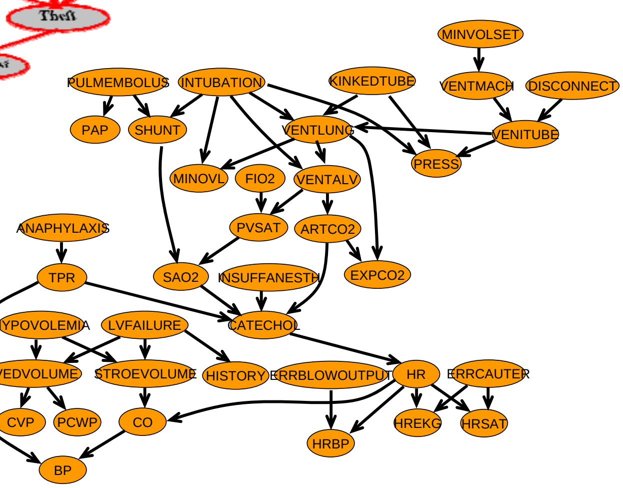
….

Output: a phylogeny



Learning probabilistic graphical models





Learning to fly

Source: *Andrew Ng*

Introduction to machine learning

- What is machine learning?
- How is machine learning related to other fields?
- Machine learning applications
- Types of learning
  - Supervised learning
    - classification
    - regression
  - Unsupervised learning
    - clustering
    - data association
    - abnormality detection
    - dimensionality reduction
    - structure learning
  - Semi-supervised learning
  - Active learning
  - Reinforcement learning and control of partially observed Markov decision processes.

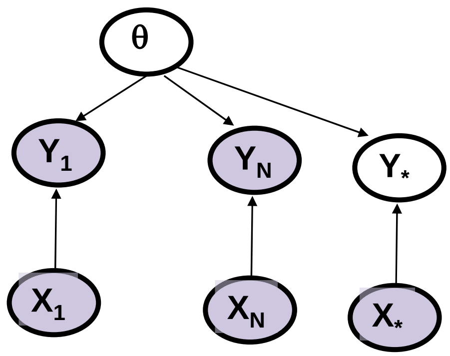
Supervised learning as Optimization

Training: For data *x*, teacher provides labels *y*. We optimize to infer the most probable model given the training data *D*=(*x,y*) and prior preferences

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} \log P(D|\theta) + \log P(\theta)$$

Testing: We predict the label of a new point

$$P(y_*|x_*, D) \approx P(y_*|x_*, \widehat{\theta}_{MAP})$$



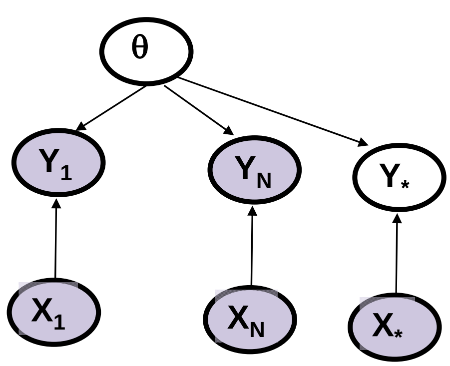
Supervised learning as Bayesian inference

Training: For data *x*, teacher provides labels *y*. We apply Bayes rule to infer the complete model given the training data *D*=(*x,y*) and prior

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

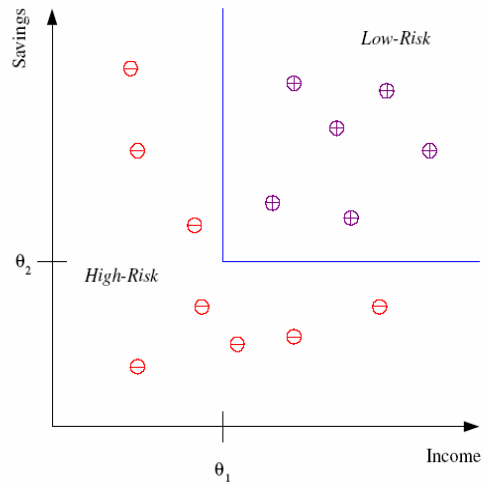
Testing: We predict the label of a new point

$$P(y_*|x_*, D) = \int P(y_*|x_*, \theta)P(\theta|D)d\theta$$



Classification

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their *income* and *savings*

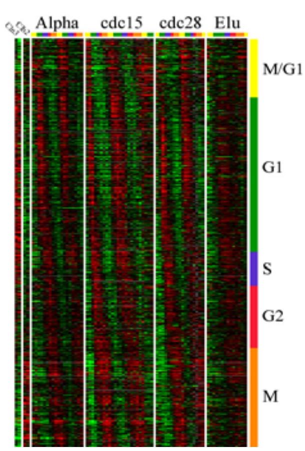


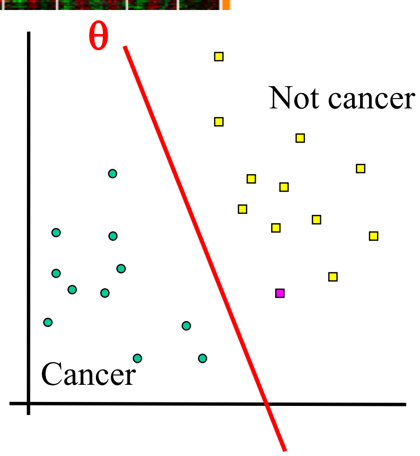
Discriminant: IF *income* > θ_1 AND *savings* > θ_2

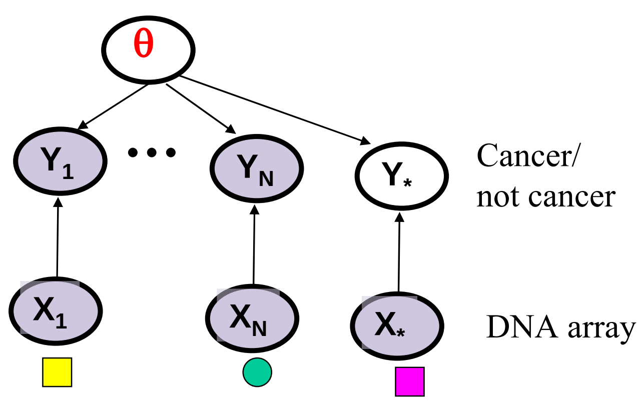
THEN low-risk ELSE high-risk

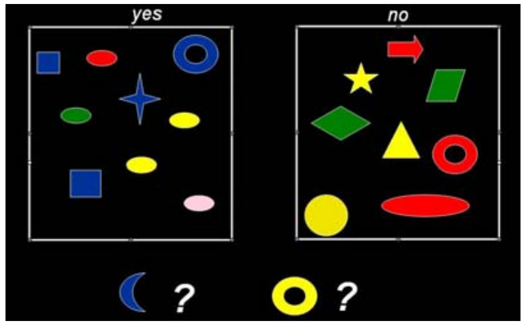
Input data is two dimensional, output is binary {0,1}

Classification of DNA arrays









p features (attributes)

Training set:
X: n × p
y: n × 1

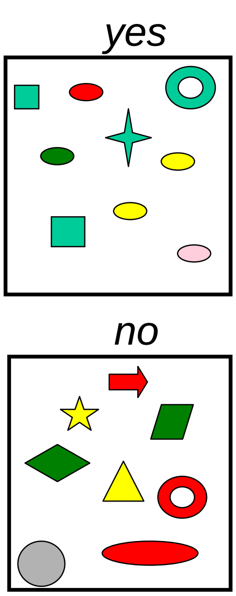
n cases

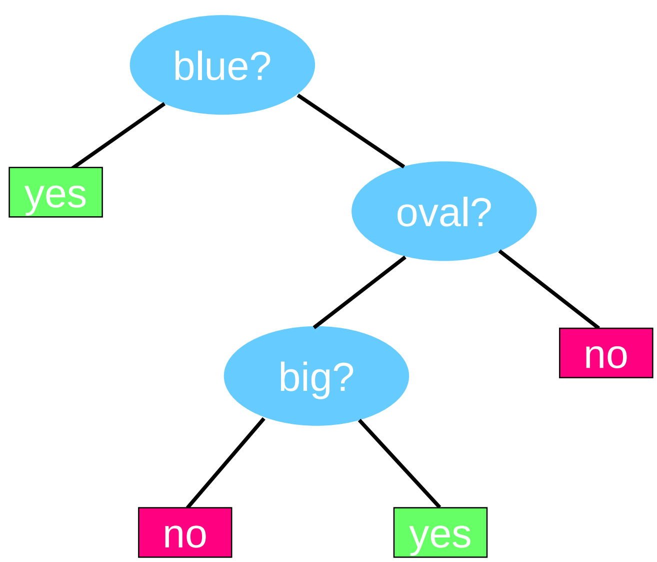
Color	Shape	Size	Label
Blue	Square	Small	Yes
Red	Ellipse	Small	Yes
Red	Ellipse	Large	No

Test set

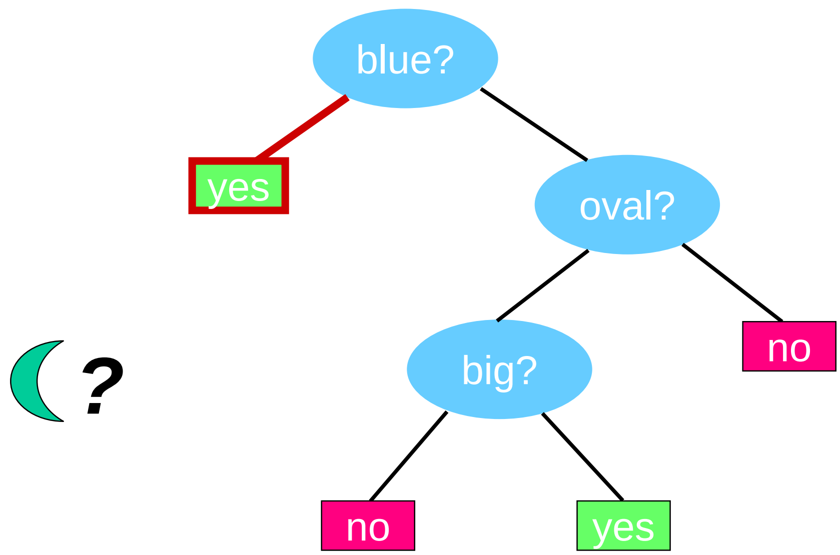
Blue	Crescent	Small	?
Yellow	Ring	Small	?

Hypothesis (decision tree)

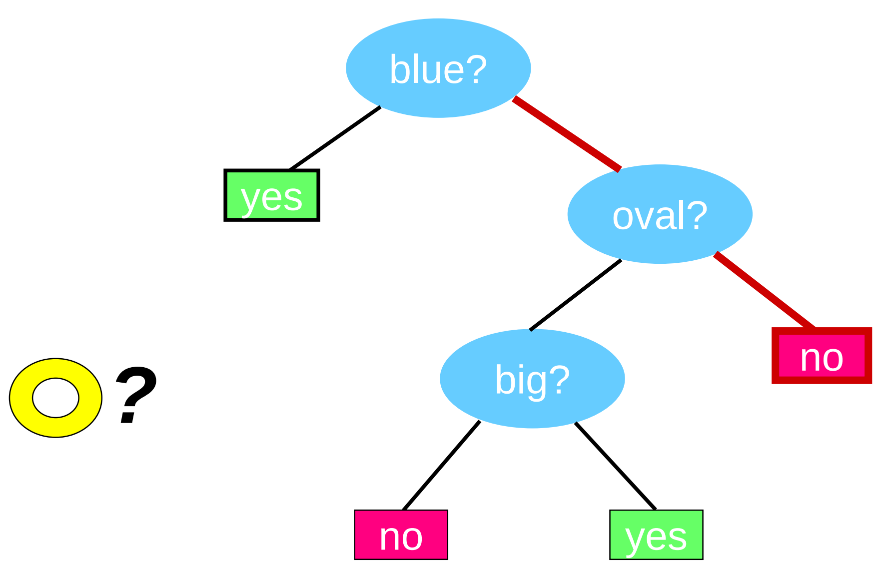




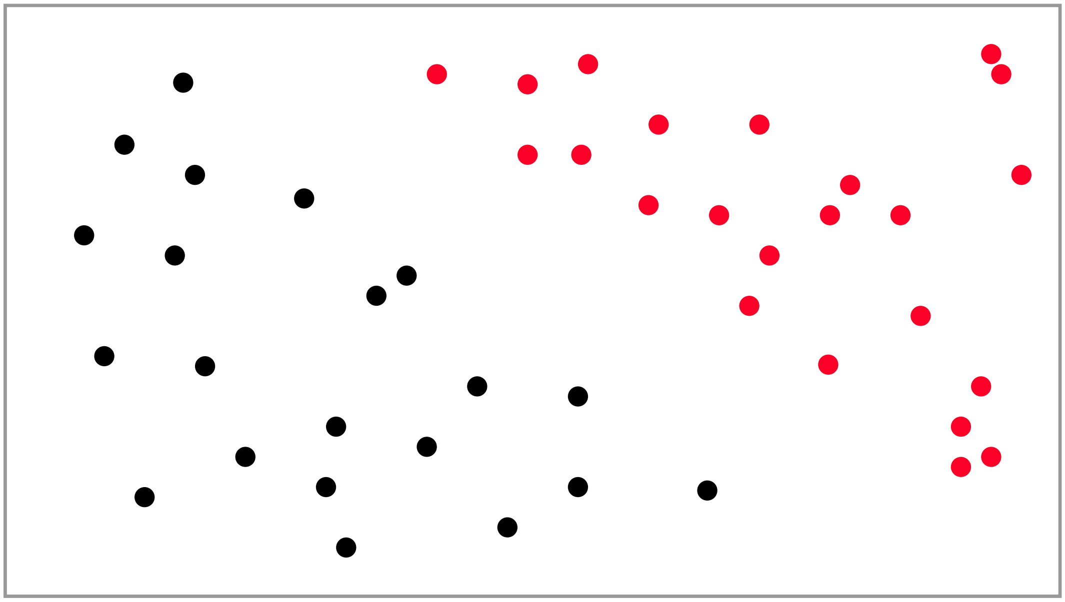
Decision Tree



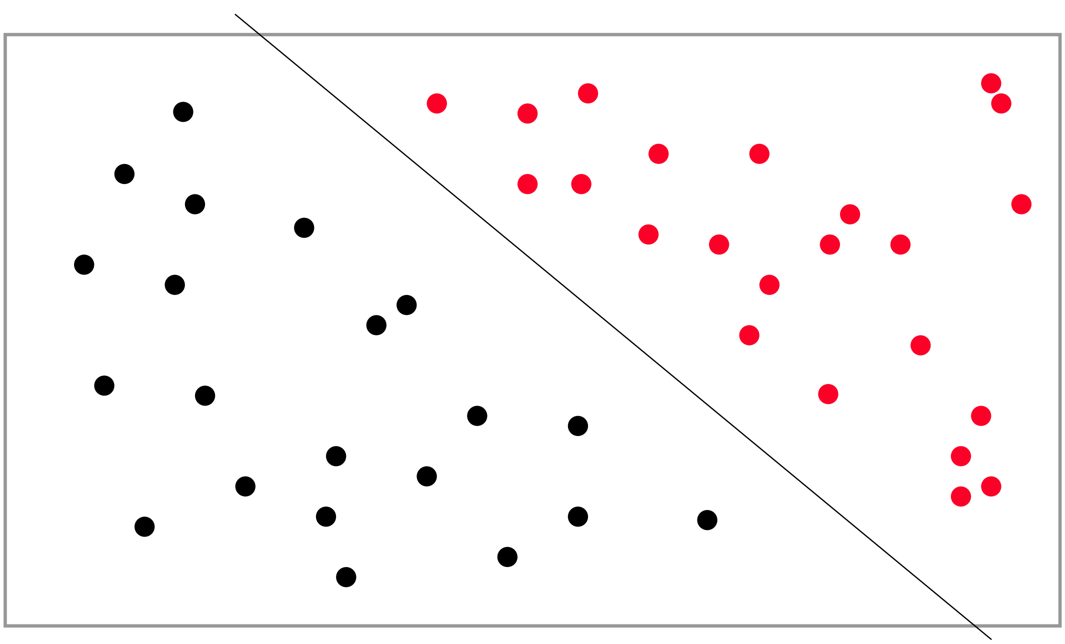
Decision Tree



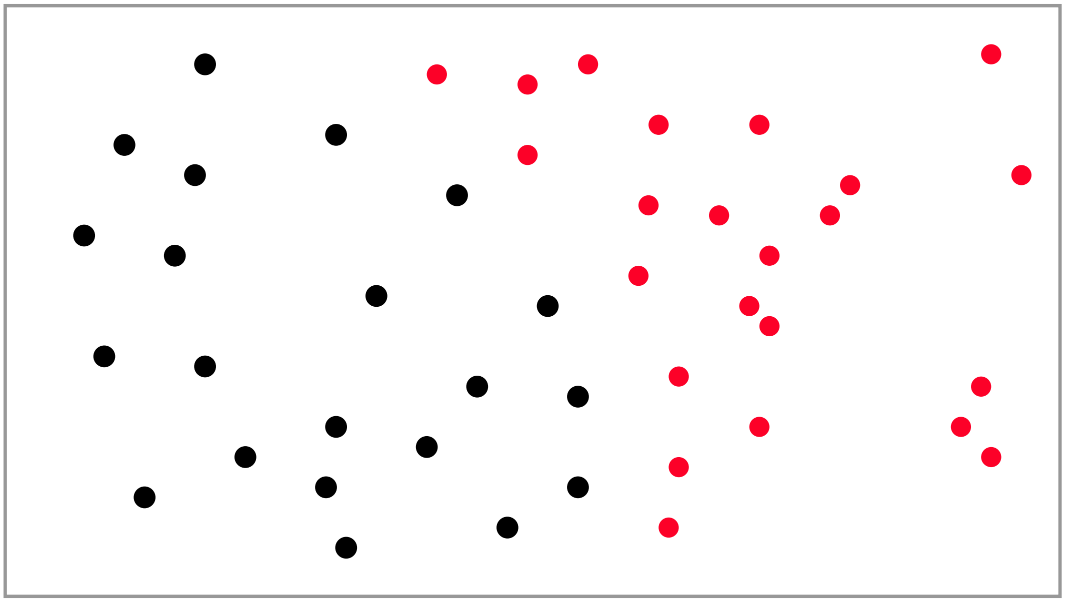
What's the right hypothesis?



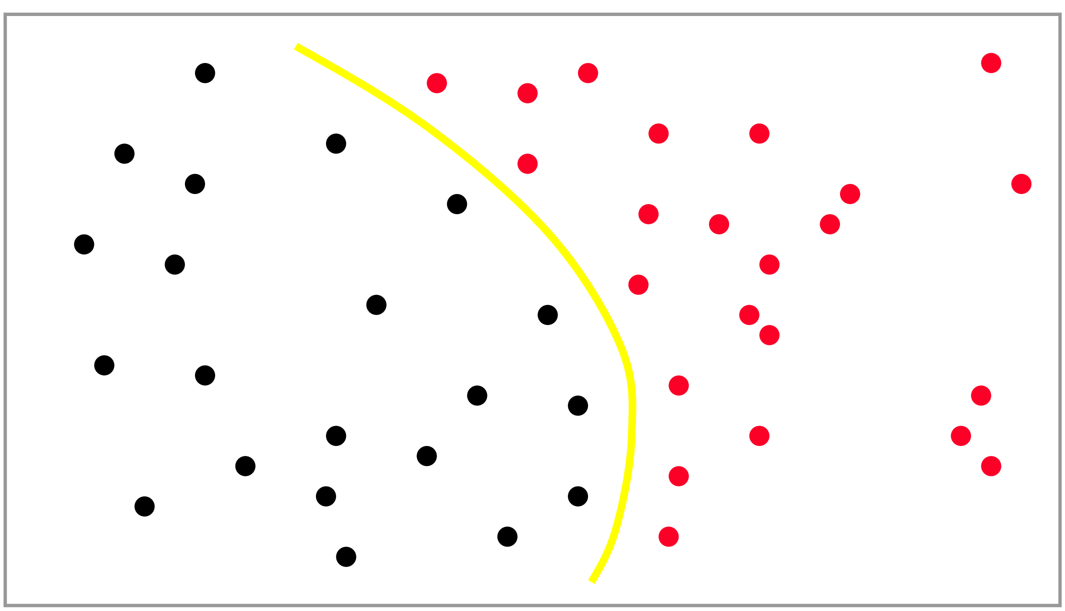
What's the right hypothesis?



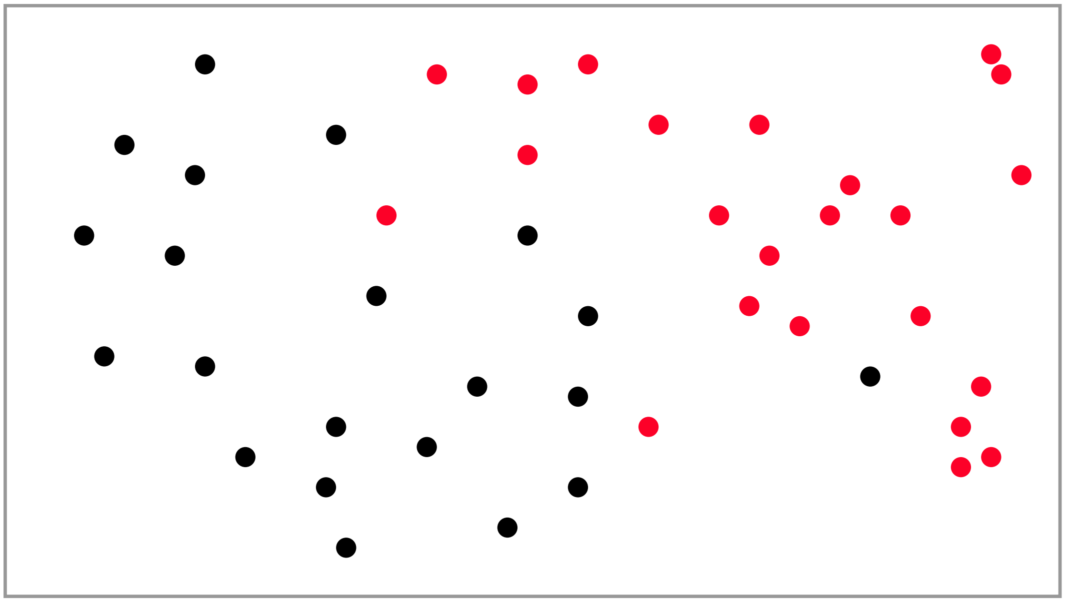
How about now?



How about now?

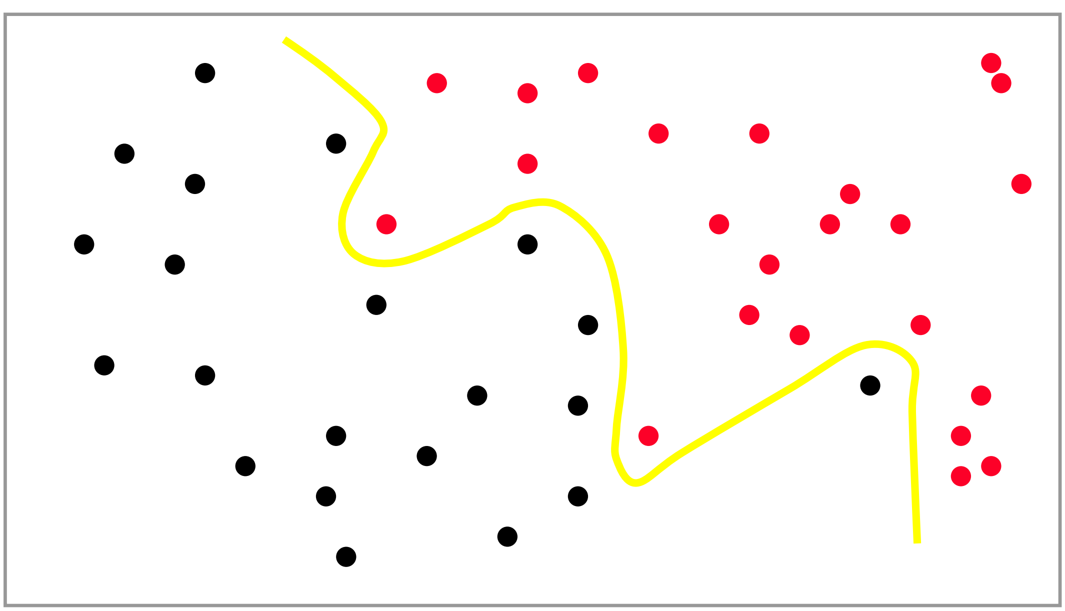


Noisy/ mislabeled data



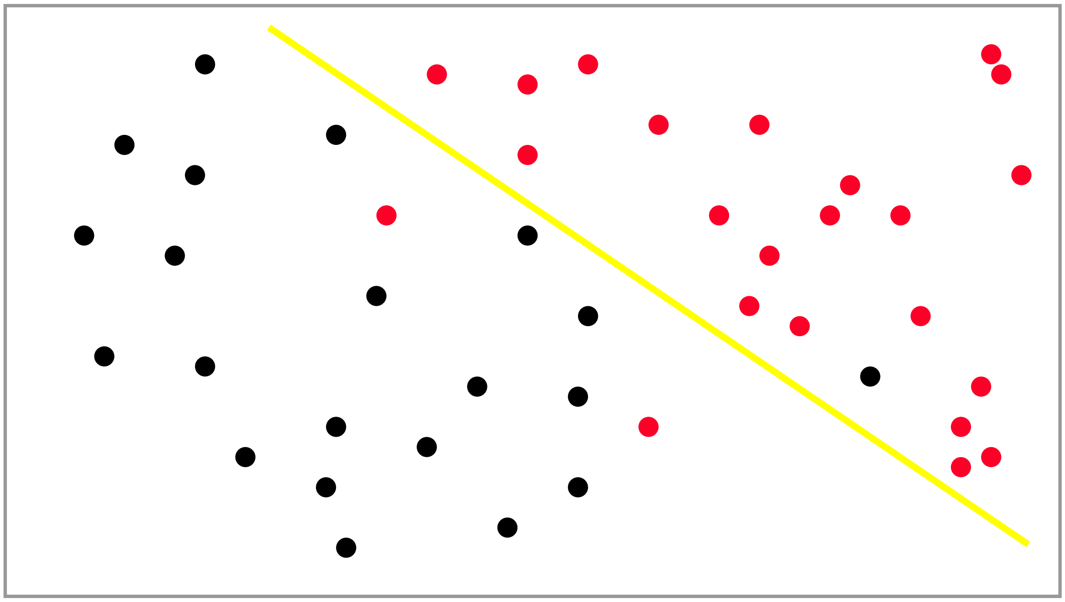
Overfitting

- Memorizes irrelevant details of training set

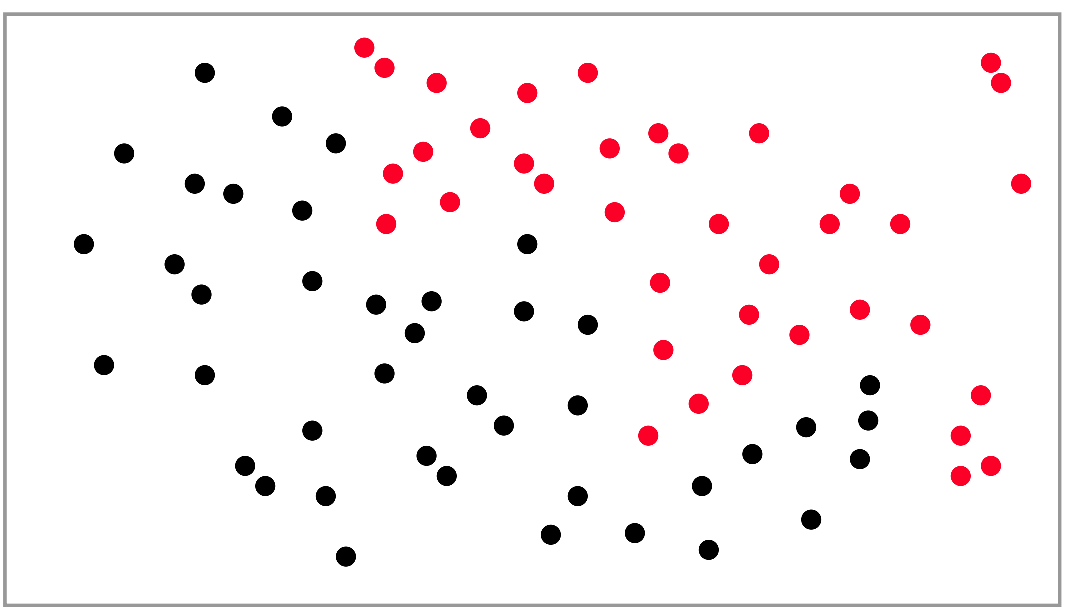


Underfitting

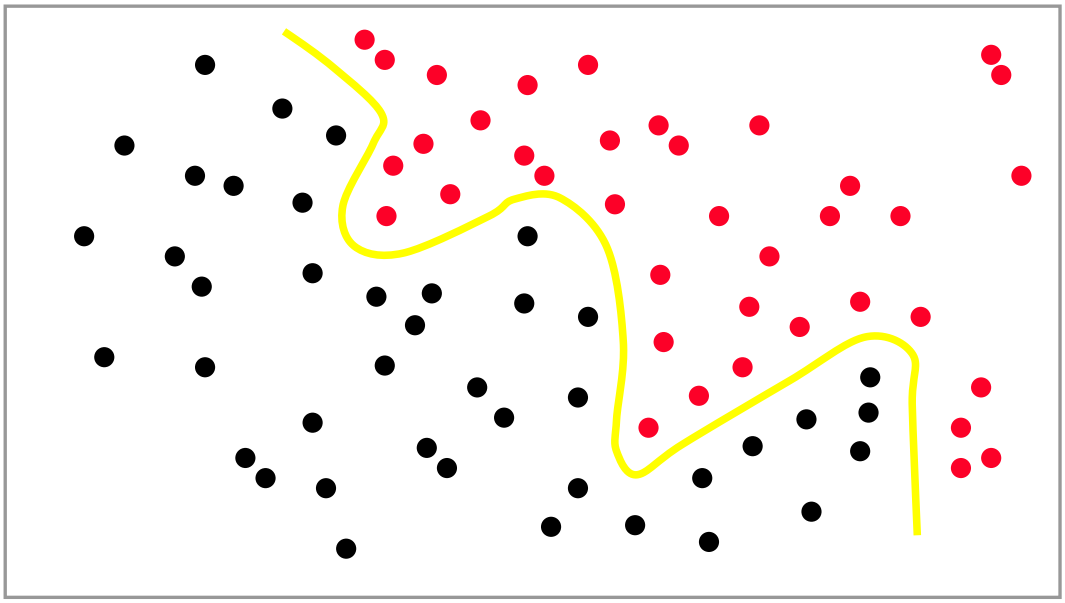
- Ignores essential details of training set



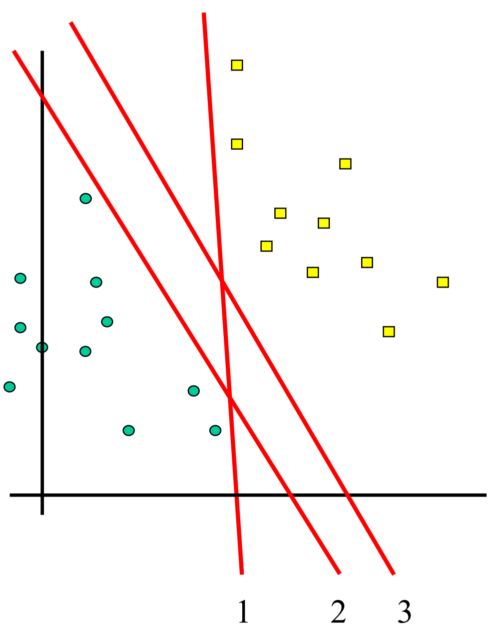
Now we're given a larger data set

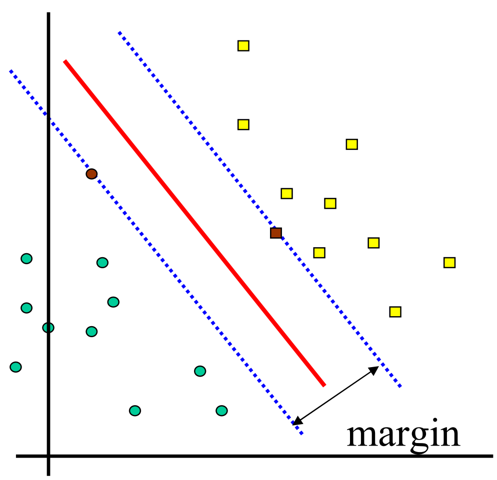


Now more complex hypothesis is ok



Which linear hypothesis is better?



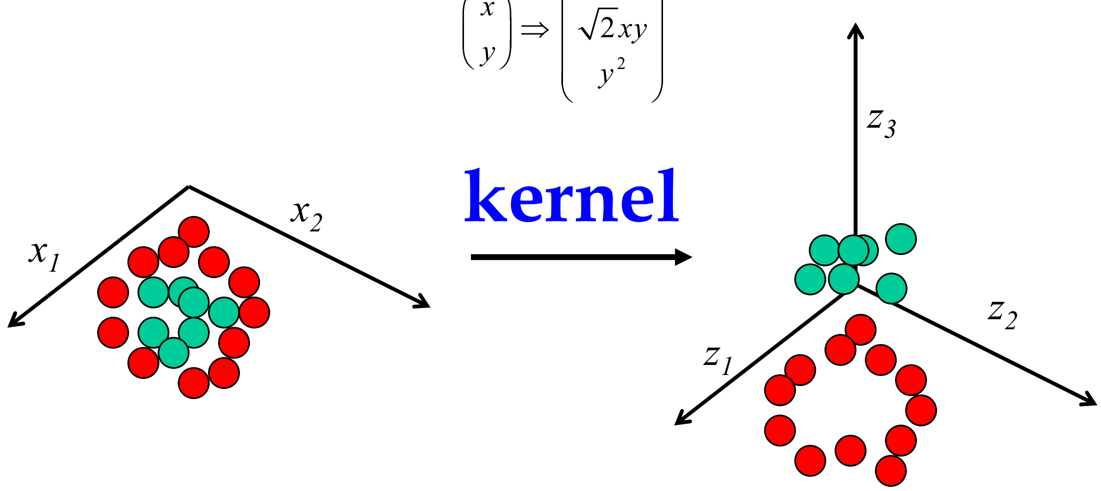


No free lunch theorem

- Unless you know something about the distribution of problems your learning algorithm will encounter, ***any hypothesis that agrees with all your data is as good as any other***. Learning is an ill-posed problem.
- You have to make *assumptions* about the underlying future.
- These assumptions are implicit in the choice of hypothesis space (and maybe the algorithm).
- Hence learning is inductive, not deductive.

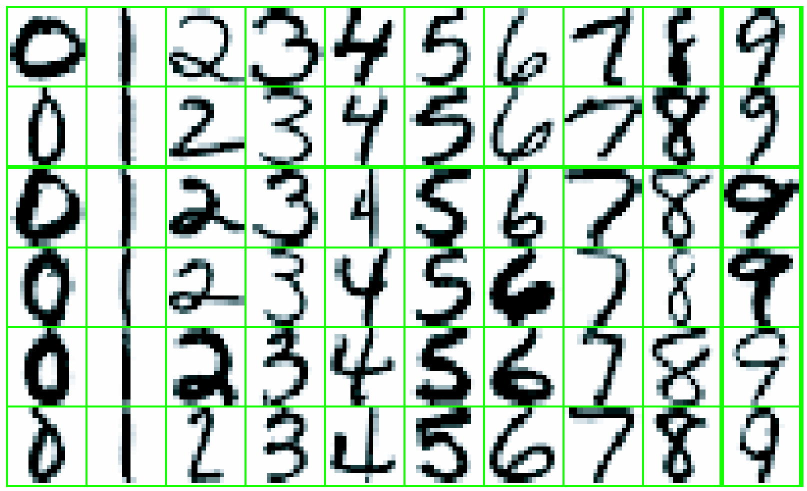
Building nonlinear classifiers: finding the right feature transformations or kernels

$$\begin{pmatrix} x \\ y \end{pmatrix} \Rightarrow \begin{pmatrix} x^2 \\ \sqrt{2}xy \\ y^2 \end{pmatrix}$$



Kernel implicitly maps from 2D to 3D, making problem linearly separable

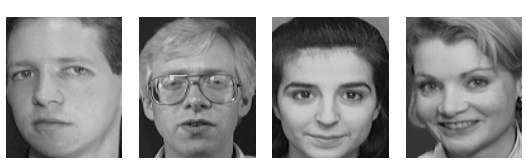
Example: Handwritten digit recognition for postal codes



Example: Face Recognition

Training examples of a person

Test images



AT&T Laboratories, Cambridge UK
http://www.uk.research.att.com/facedatabase.html

Linear regression

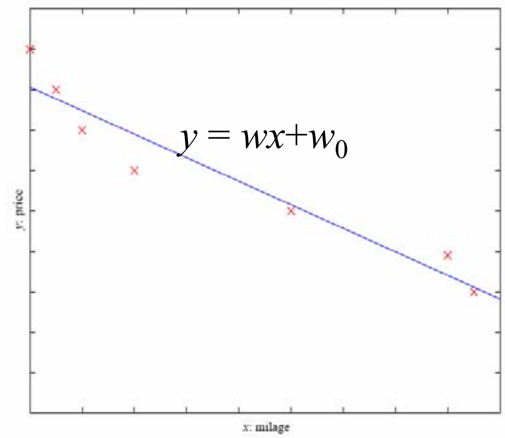
- Example: Price of a used car
- x : car attributes

  y : price

  $y = g(x,\theta)$

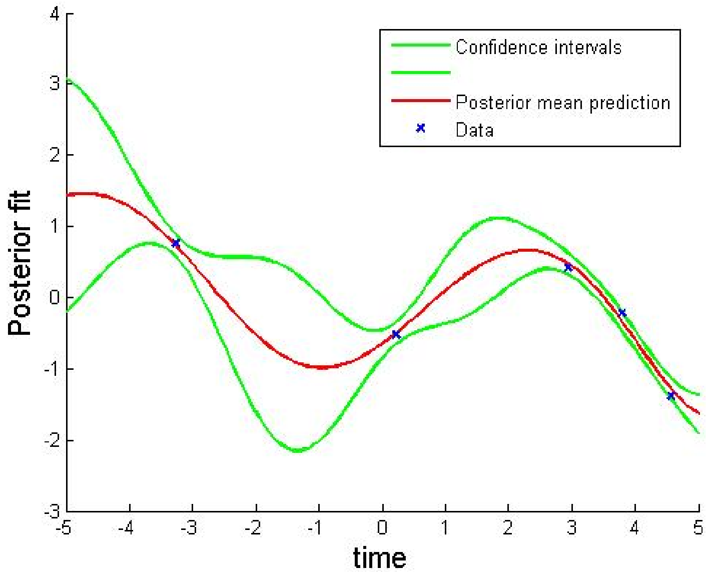
  g ( ) model,

  $\theta = (w,w_0)$ parameters (slope and intercept)



Regression is like classification except the output is a real-valued scalar

Nonlinear regression



Useful for:
- Prediction
- Control
- Compression
- Outlier detection
- Knowledge extraction

Introduction to machine learning

- What is machine learning?
- How is machine learning related to other fields?
- Machine learning applications
- Types of learning
  - Supervised learning
    - classification
    - regression
  - Unsupervised learning
    - clustering
    - data association
    - abnormality detection
    - dimensionality reduction
    - structure learning
  - Semi-supervised learning
  - Active learning
  - Reinforcement learning and control of partially observed Markov decision processes.

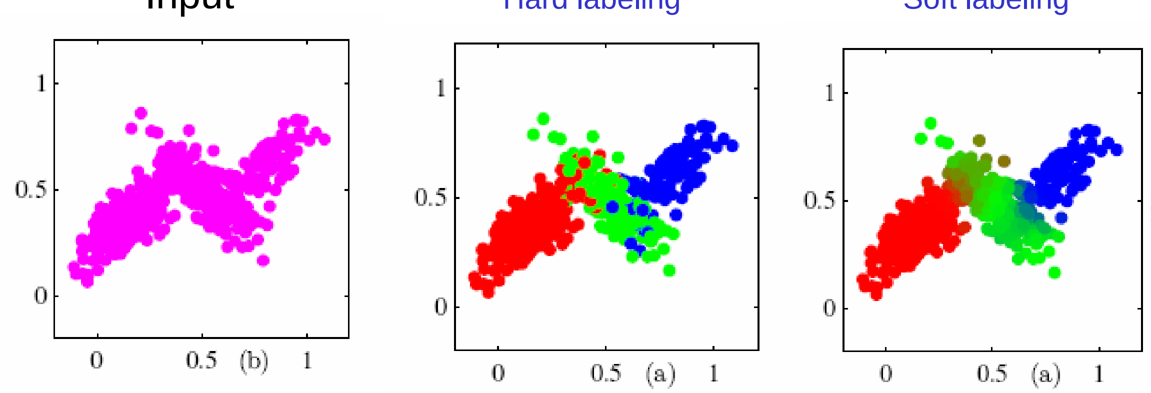
Clustering

Desired output

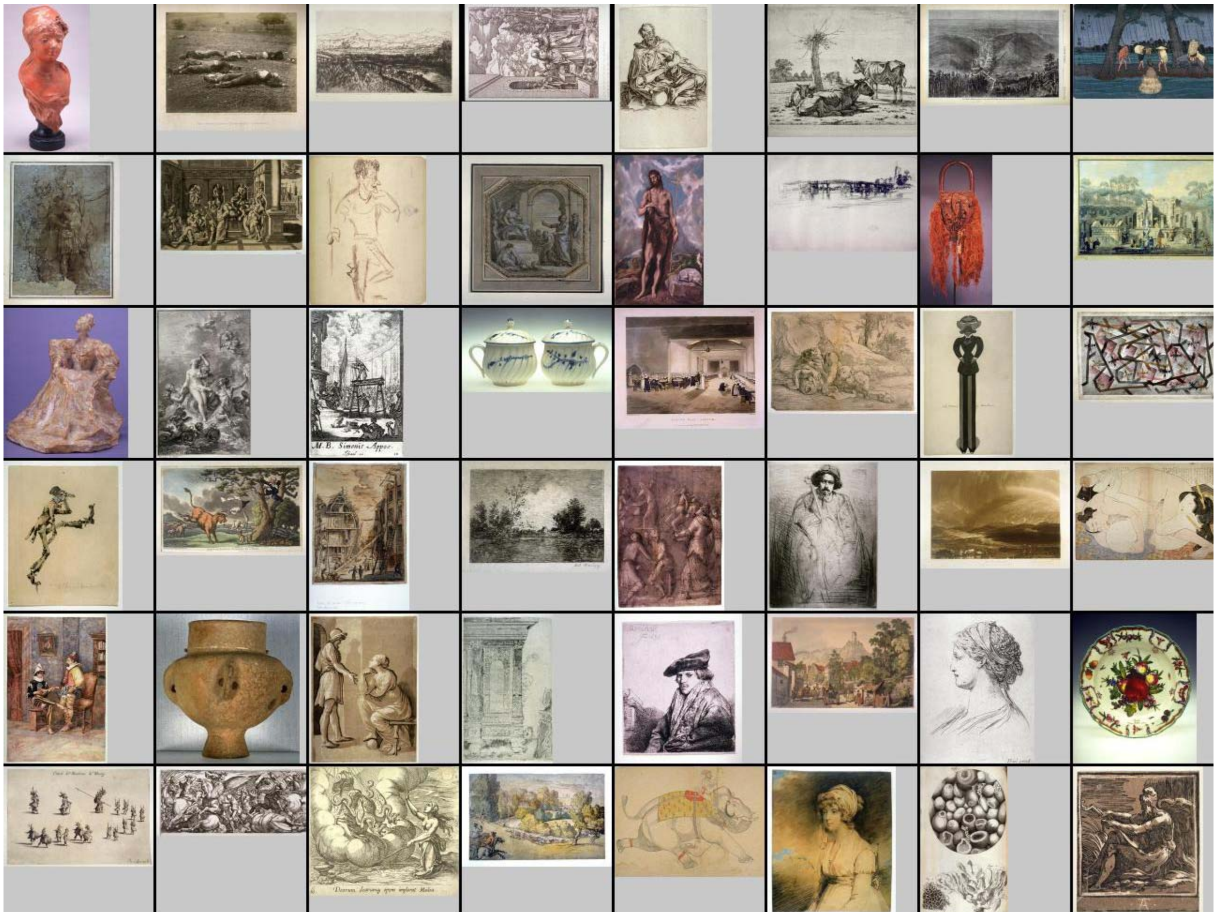
Input

Hard labeling

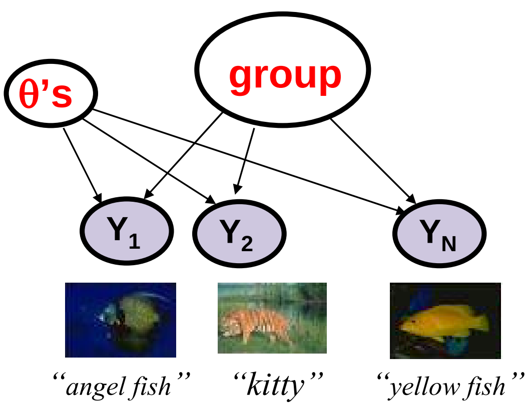
Soft labeling

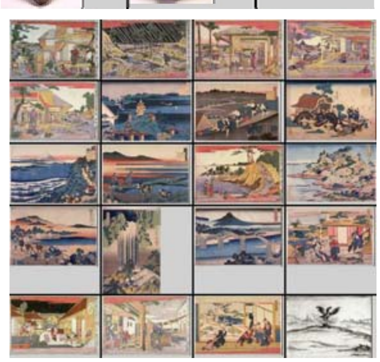


K=3 is the number of clusters, here chosen by hand

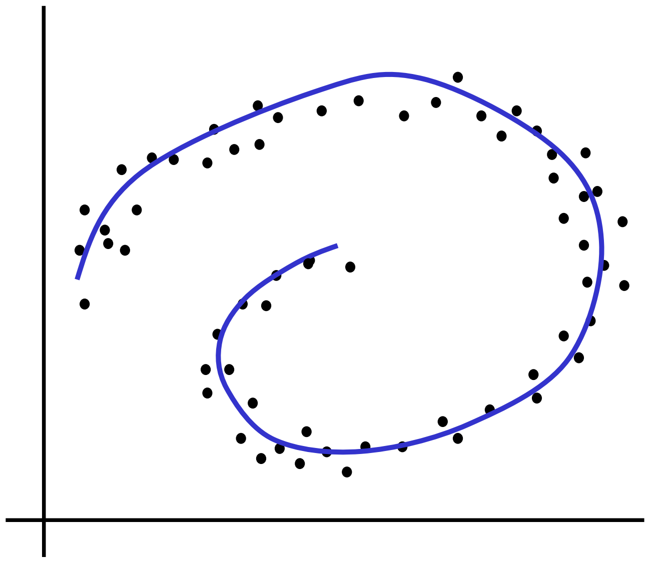


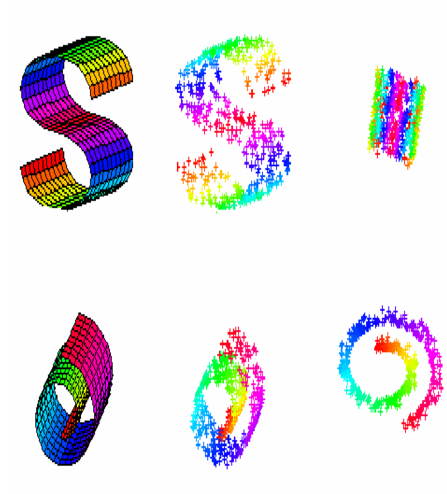
Clustering

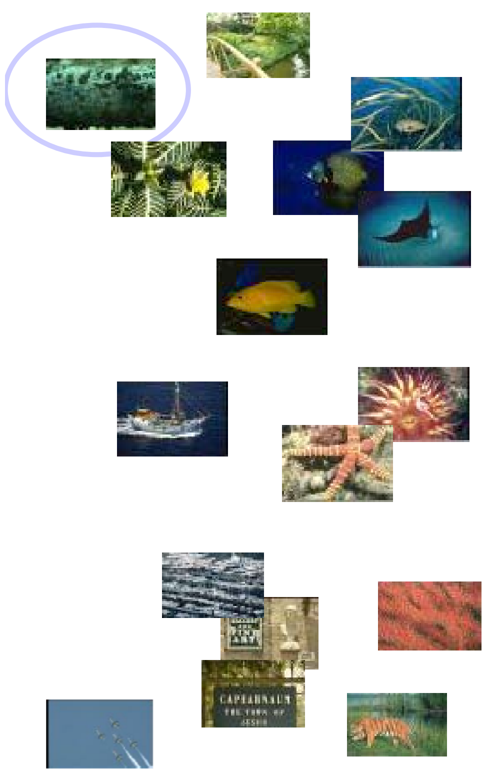


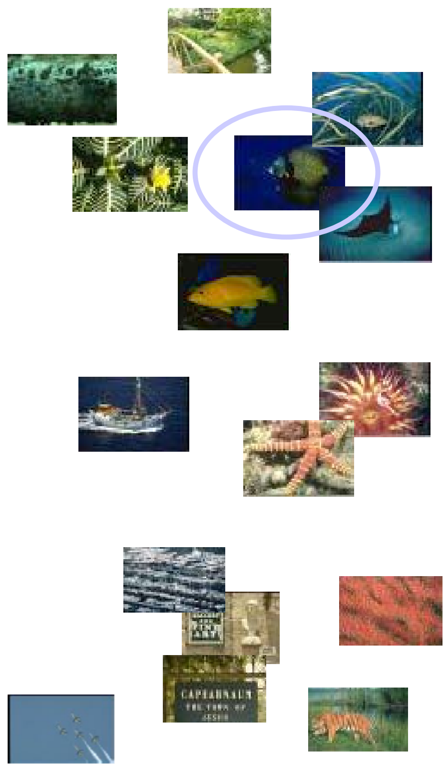


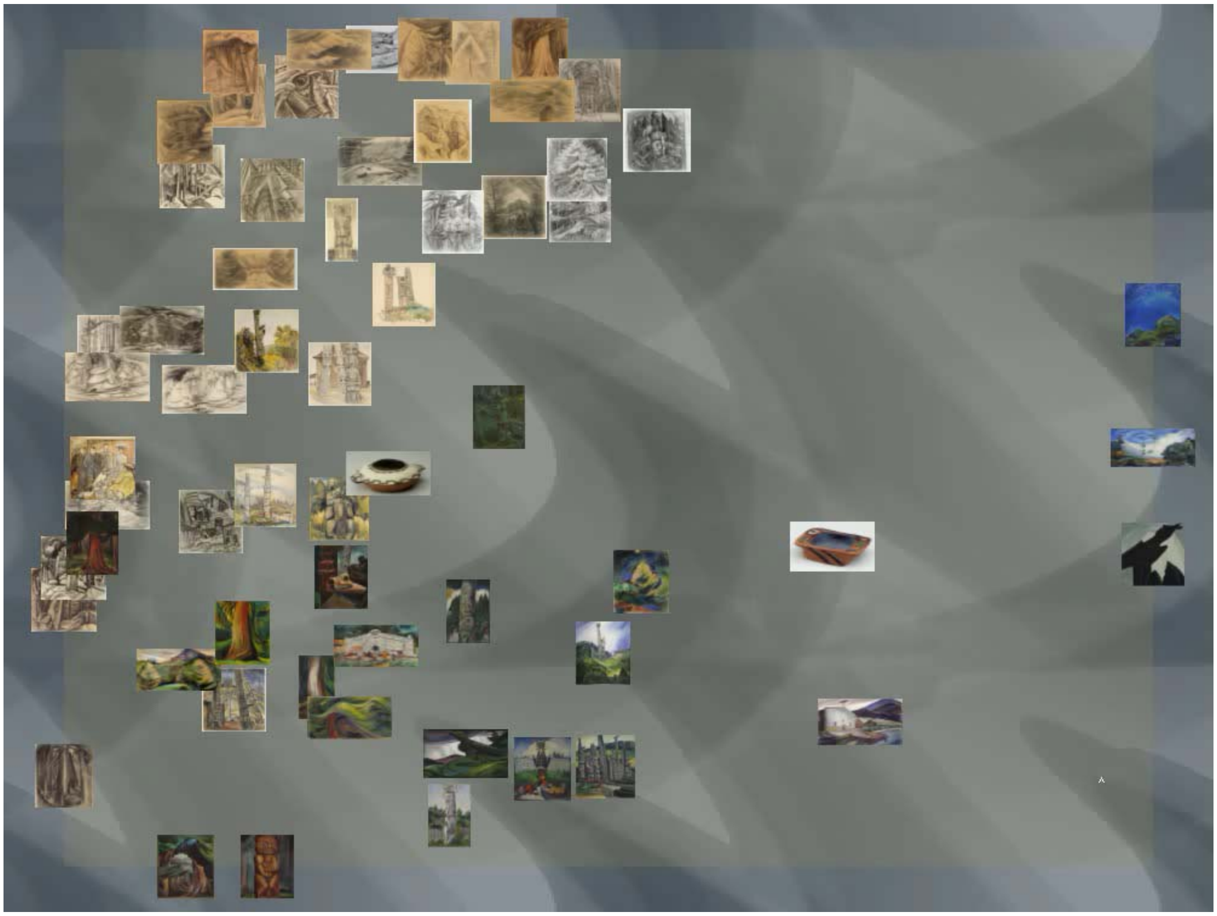
Discovering nonlinear manifolds











Query: “river tiger”
(Even though the words never occur together)

Retrieved items:

TIGER CAT WATER GRASS TIGER CAT WATER GRASS TIGER CAT GRASS TREES

TIGER CAT WATER GRASS TIGER CAT GRASS FOREST TIGER CAT WATER GRASS

Query: “water sky cloud”

Retrieved items:


1068
SUN CLOUDS
WATER SKY
58


1090
SUN CLOUDS
WATER SKY
53


10061
PLANE SKY
CLOUDS WATER
14

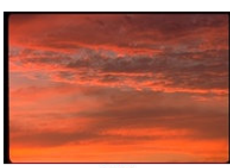

106064
icebergs WATER
SKY CLOUDS
11


106069
iceberg WATER
SKY CLOUDS
47


118011
WATER HARBOR
SKY CLOUDS
36

Query: “water sky cloud  ”

Retrieved items:


1066
CLOUDS glow
SKY SUN


1037
SUN SEA
WAVES SKY

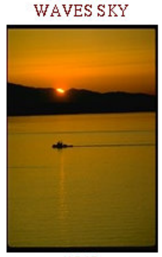

1027
SUN SEA
WAVES SKY


1083
SUN WATER
WAVES CLOUDS


1063
SUN SEA
SKY WAVES


1064
SUN CLOUDS
bay SKY


1038
SUN SEA
WAVES BIRDS

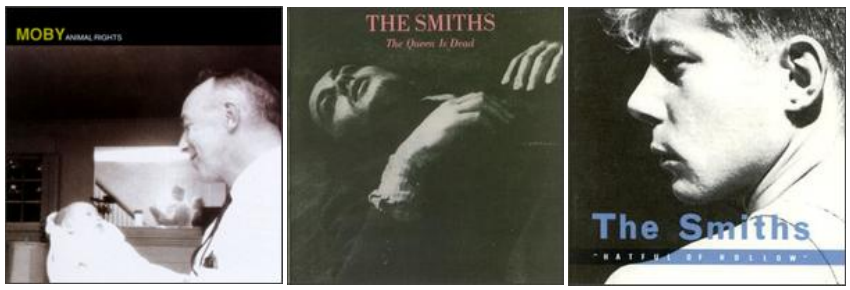

1040
SUN SEA
BOAT LAND


1028
SUN SEA
WAVES SKY

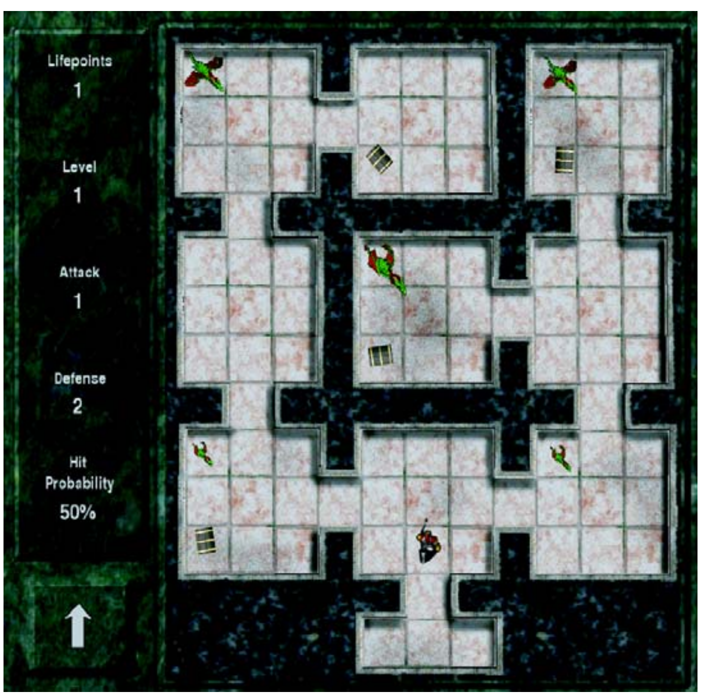

1015
SUN TREE
PLAIN SKY

Input image

Image matches



Touring: Online computer game



Query: Game state

Retrieved: Action

Input poem

The Waste Land (excerpt)

T S Eliot

For Ezra Pound,
il miglior fabbro.

I. The Burial of the Dead

April is the cruelest month, breeding
Lilacs out of the dead land, mixing
Memory and desire, stirring
Dull roots with spring rain.
Winter kept us warm, covering
Earth in forgetful snow, feeding
A little life with dried tubers.
Summer surprised us, coming over the Starnbergersee
With a shower of rain; we stopped in the colonnade
And went on in sunlight, into the Hofgarten,
And drank coffee, and talked for an hour.
Bin gar keine Russin, stamm' aus Litauen, echt deutsch.
And when we were children, staying at the arch-duke's,
My cousin's, he took me out on a sled,
And I was frightened. He said, Marie,
Marie, hold on tight. And down we went.
In the mountains, there you feel free.
I read, much of the night, and go south in winter.
…

Closest song match

One Hundred Years (excerpt)

The Cure

It doesn't matter if we all die
Ambition in the back of a black car
In a high building there is so much to do
Going home time
A story on the radio
Something small falls out of your mouth
And we laugh
A prayer for something better
Please love me
Meet my mother
But the fear takes hold
Have we got everything?
She struggles to get away
The pain
And the creeping feeling
A little black haired girl
Waiting for Saturday
The death of her father pushing her
Pushing her white face into the mirror
Aching inside me
…

- **Music**

Auto-illustration

Text Passage (Moby Dick)

"The large importance attached to the harpooneer's vocation is evinced by the fact, that originally in the old Dutch Fishery, two centuries and more ago, the command of a whale-ship …"

Query

large importance attached fact old dutch century more command whale ship was person was divided officer word means fat cutter time made days was general vessel whale hunting concern british title old dutch ...

Retrieved Images

PRINT NAVAL BATTLE JAPANESE SHIP CHINESE BEING SHIP WATER

PRINT SHIP SURROUNDED ICE SEVERAL SHIP SEEN WHALE OTHER CURRIER

PRINT ATTACK WAGON ROAD FOREST CALLOT

PRINT WAR FRIGATE UNITEDSTATE ENGLISH SHIP AMERICANSHIP CURRIER

PRINT SMALL BOAT APPROACHING BLOWING WHALE SHIP MOUNTAIN BACKGROUND CURRIER

PLAY BOAT PRINT KUNISADA

PRINT MEN SMALL MOUNTAIN HAS COME SEVERAL SMALL FOREGROUND POLITICAL

PRINT WHITE HOUSE GROUNDS BACKGROUND POLITICAL TYPE INDIAN ARMS TREE

Auto-annotation

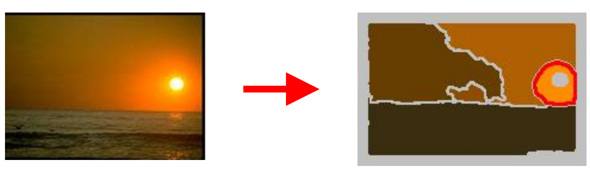
Associated Words

KUSATSU SERIES STATION TOKAIDO GOJUSANTSUGI PRINT HIROSHIGE

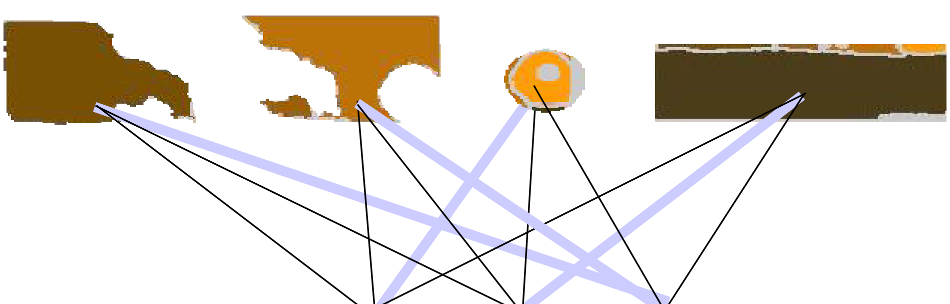
Predicted Words (rank order)

tokaido print hiroshige object artifact series ordering gojusantsugi station facility arrangement minakuchi

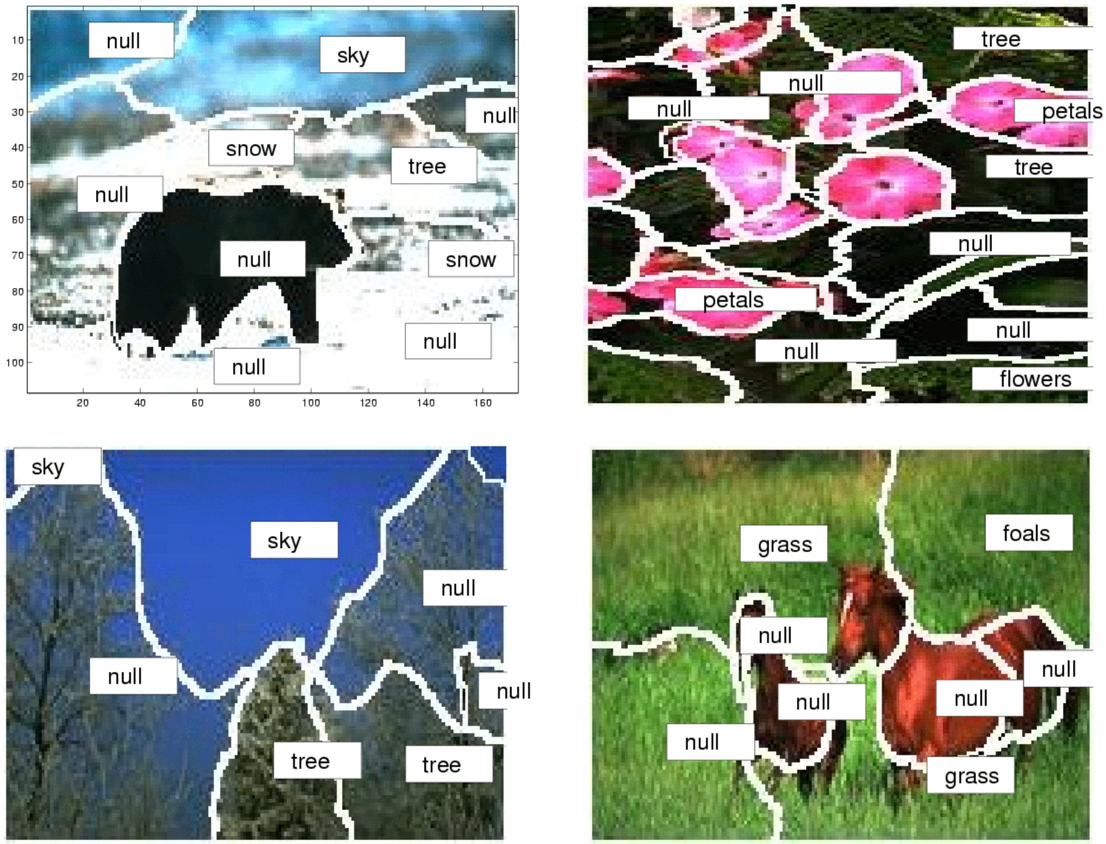
Translation and data association



“sun sea sky”



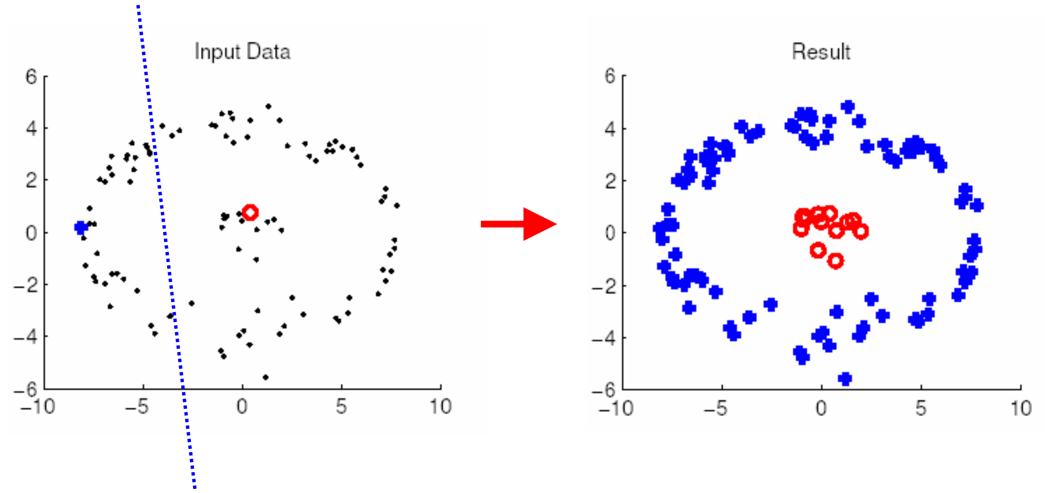
“sun sea sky”

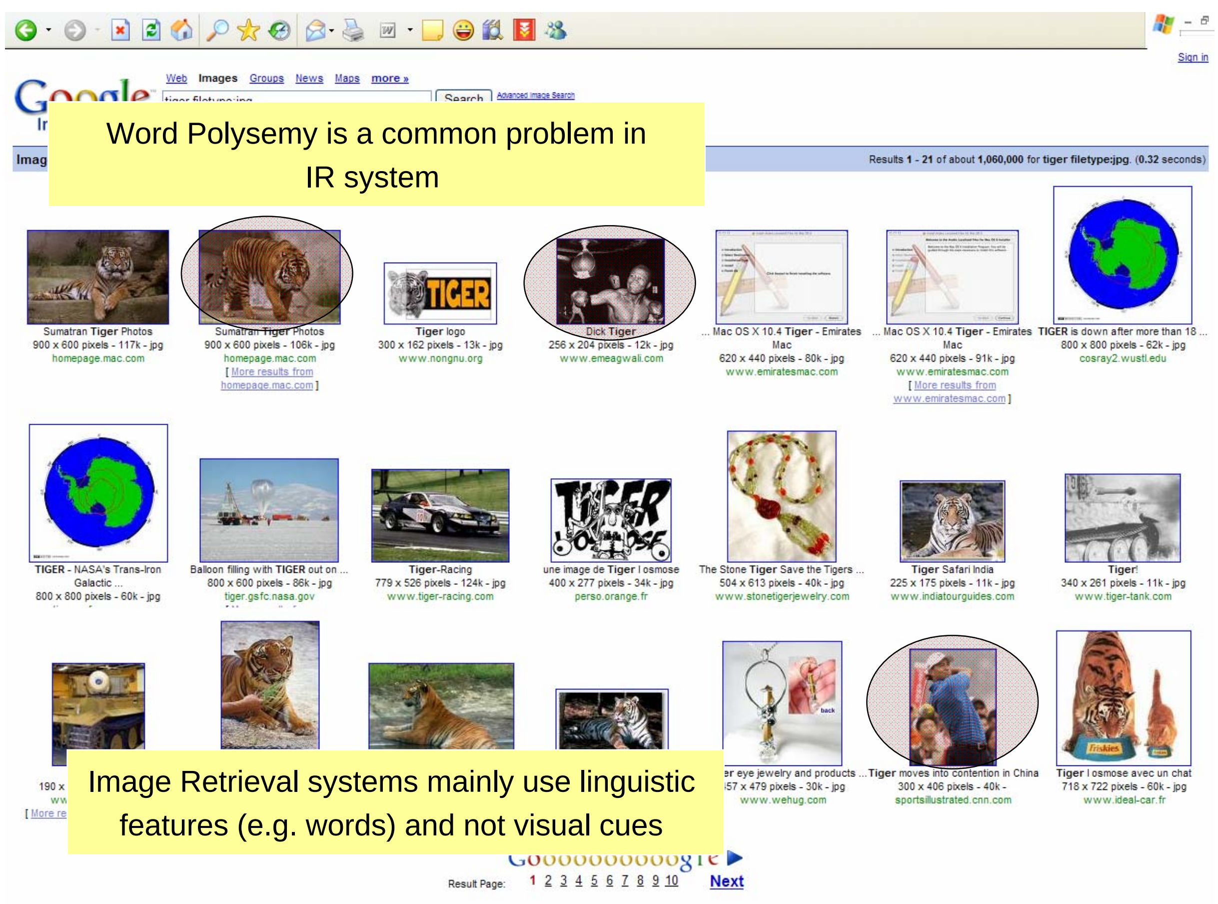


Introduction to machine learning

- What is machine learning?
- How is machine learning related to other fields?
- Machine learning applications
- Types of learning
  - Supervised learning
    - classification
    - regression
  - Unsupervised learning
    - clustering
    - data association
    - abnormality detection
    - dimensionality reduction
    - structure learning
  - Semi-supervised learning
  - Active learning
  - Reinforcement learning and control of partially observed Markov decision processes.

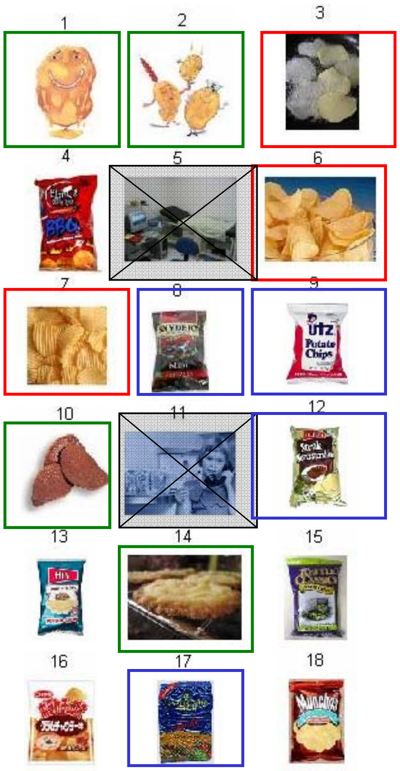
Semi-supervised learning



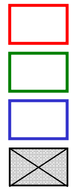


Word Polysemy is a common problem in IR system

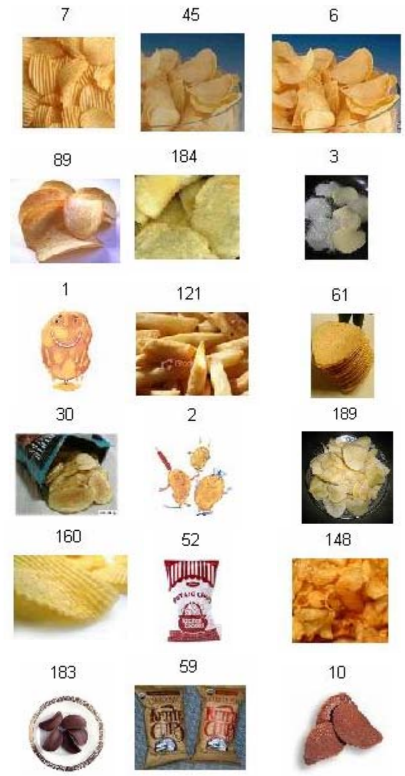
Image Retrieval systems mainly use linguistic features (e.g. words) and not visual cues before



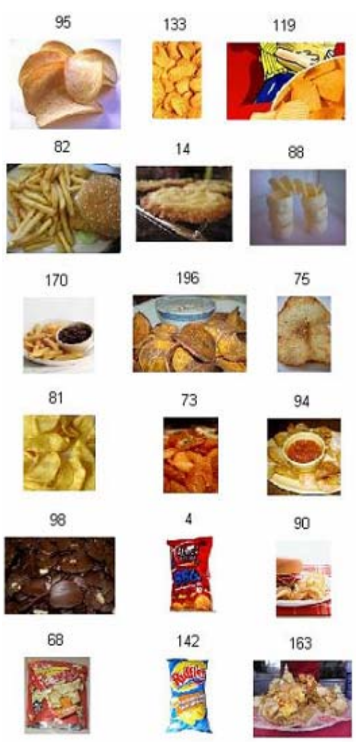
legend



after

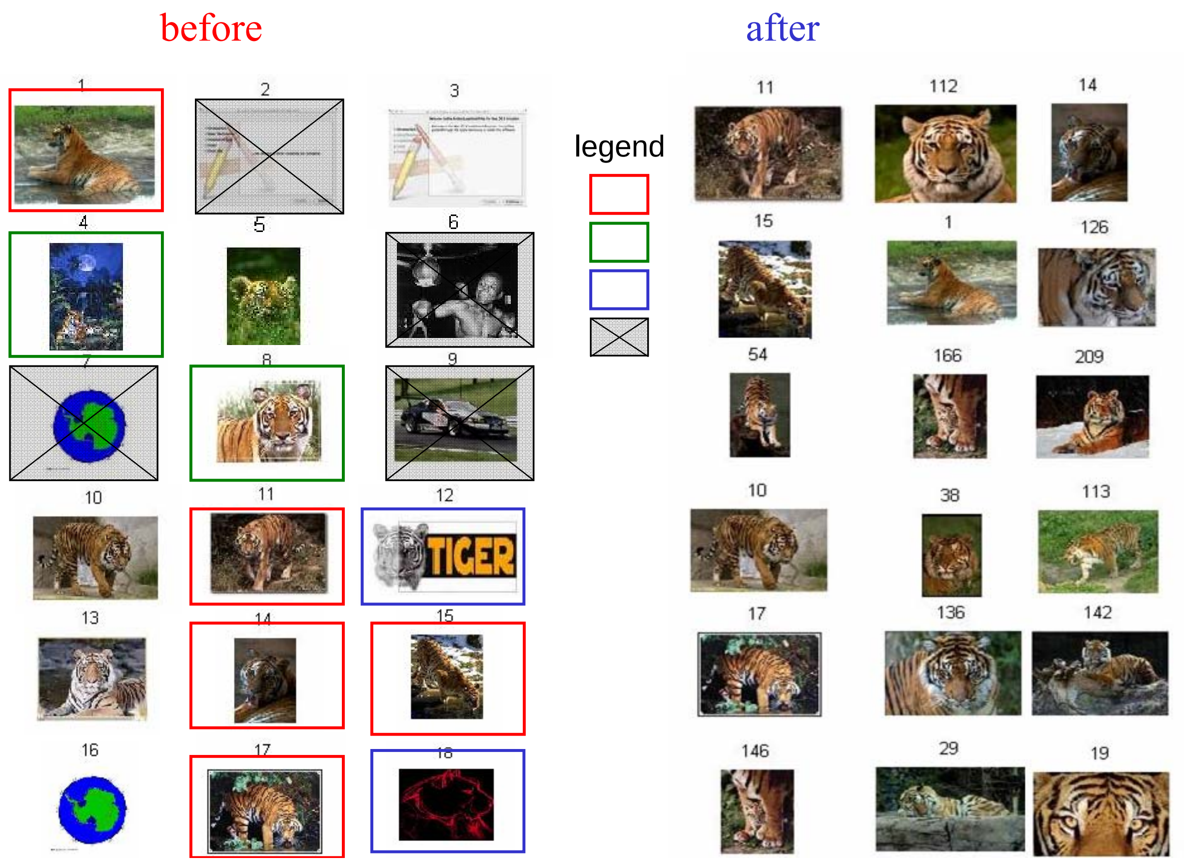


Page 2



Page 3

Page 11



Introduction to machine learning

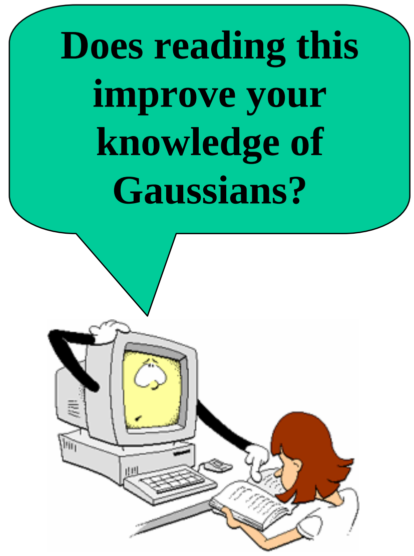
- What is machine learning?
- How is machine learning related to other fields?
- Machine learning applications
- Types of learning
  – Supervised learning
    - classification
    - regression
  – Unsupervised learning
    - clustering
    - data association
    - abnormality detection
    - dimensionality reduction
    - structure learning
  – Semi-supervised learning
  – Active learning
  – Reinforcement learning and control of partially observed Markov decision processes.

Active learning

• Active learning is a principled way of integrating decision theory with traditional statistical methods for learning models from data.

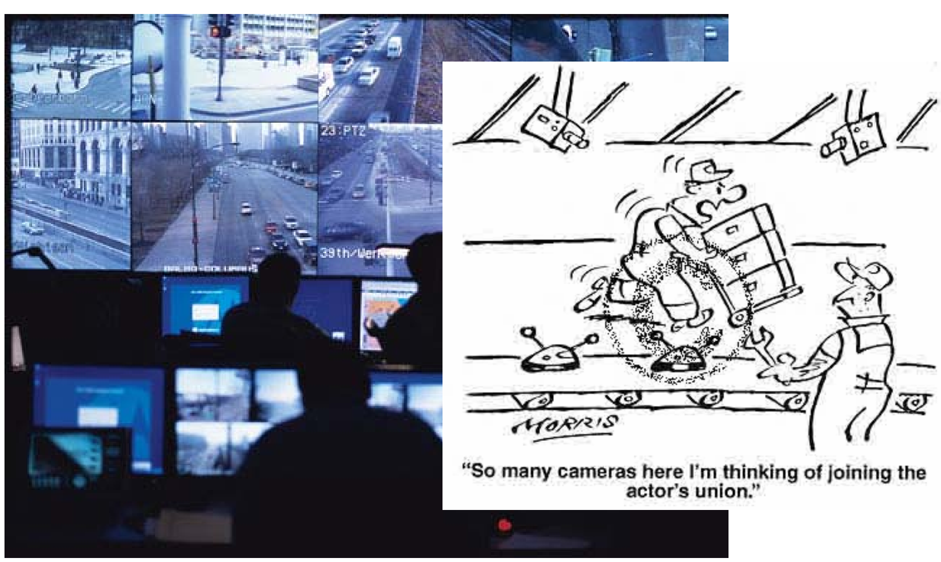
• In active learning, the machine can query the environment. That is, it can ask questions.

• Decision theory leads to optimal strategies for choosing when and what questions to ask in order to gather the best possible data. Good data is often better than a lot of data.



Active learning and surveillance

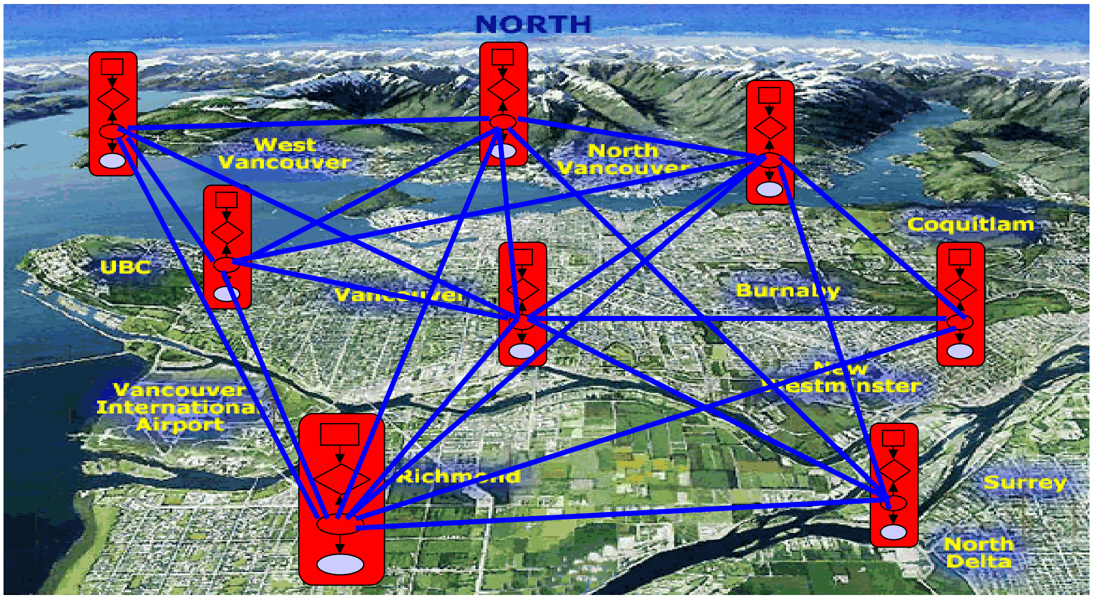
In network with thousands of cameras, which camera views should be presented to the human operator?



"So many cameras here I'm thinking of joining the actor's union."

Active learning and sensor networks

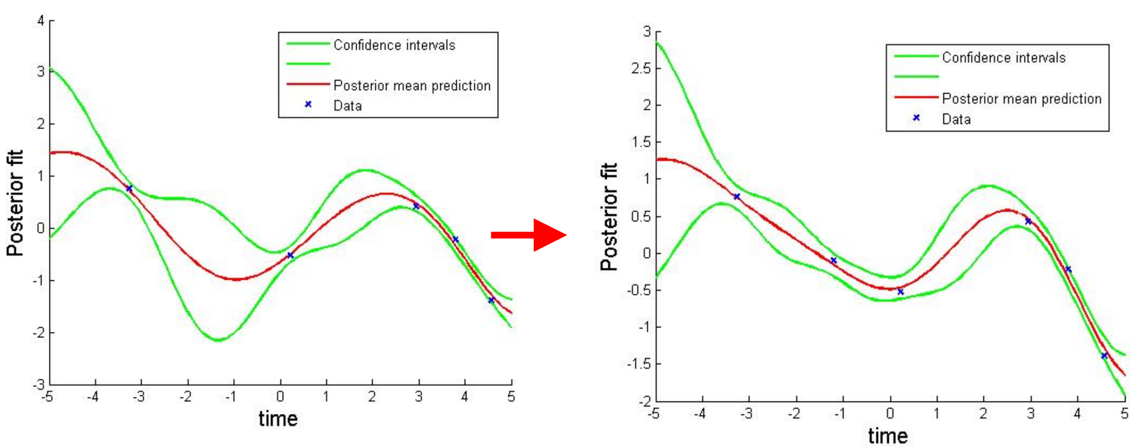
How do we optimally choose among a subset of sensors in order to obtain the best understanding of the world while minimizing resource expenditure (power, bandwidth, distractions)?



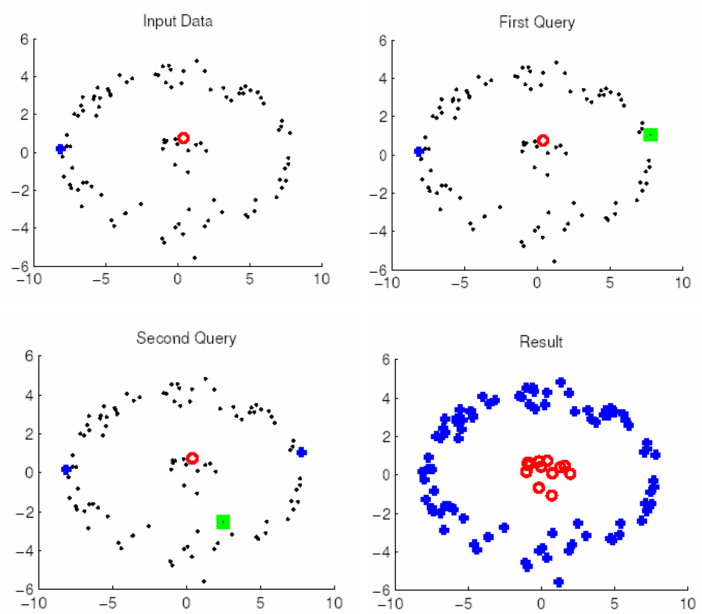
Nonlinear regression

Useful for predicting:

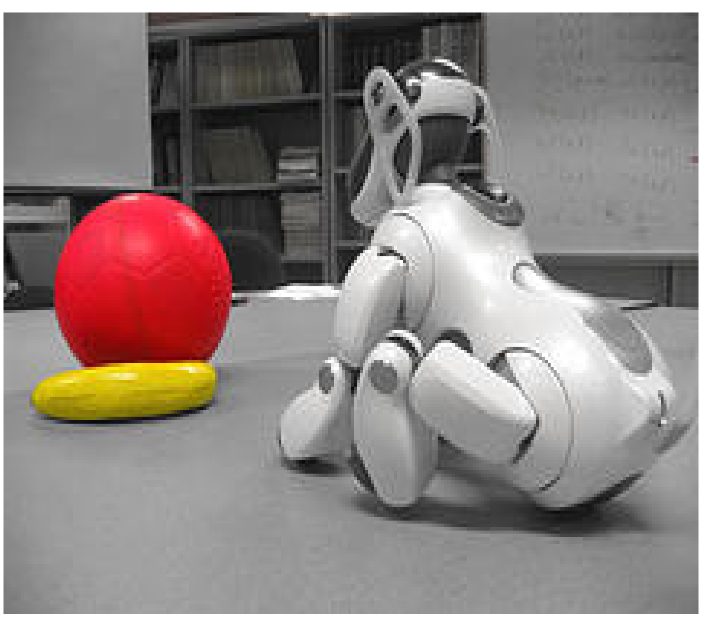
- House prices
- Drug dosages
- Chemical processes
- Spatial variables
- Output of control action

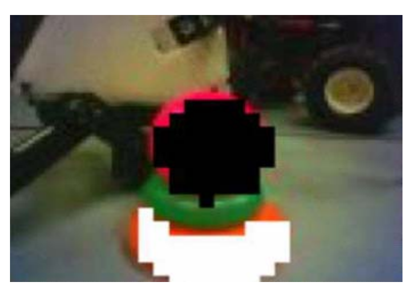


Active learning example



Interactive robots



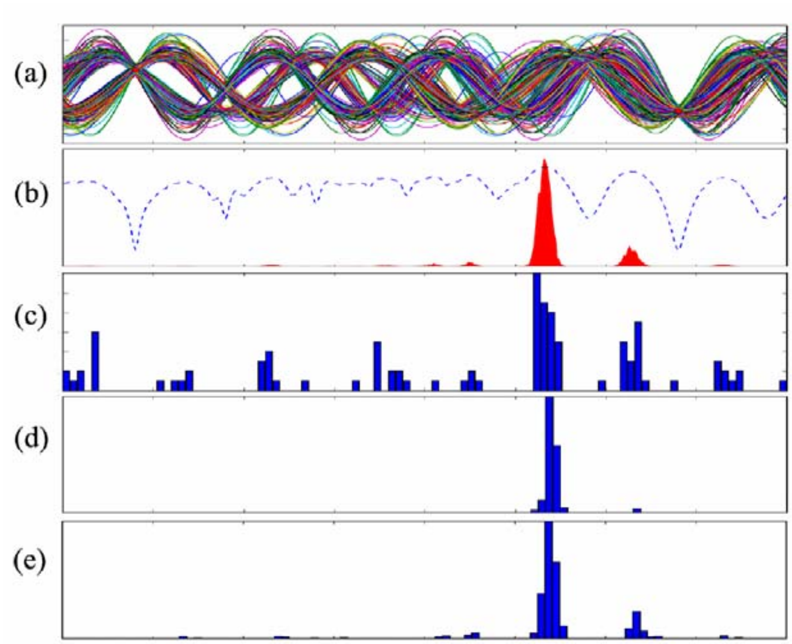


Experimental design

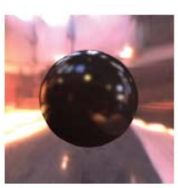
$$y = f(x; A, \omega, \rho) = A \sin\left(2\pi[(d\omega) + \rho]\right)$$

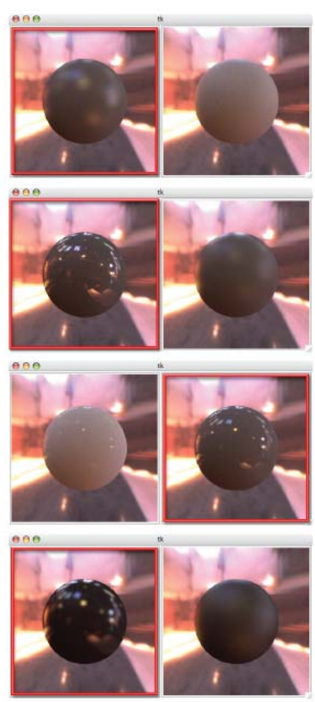
Choose experiments (covariates x in this case) so as to learn the most about the model parameters (amplitude, frequency and phase).

It also applies to model selection and validation of theories.

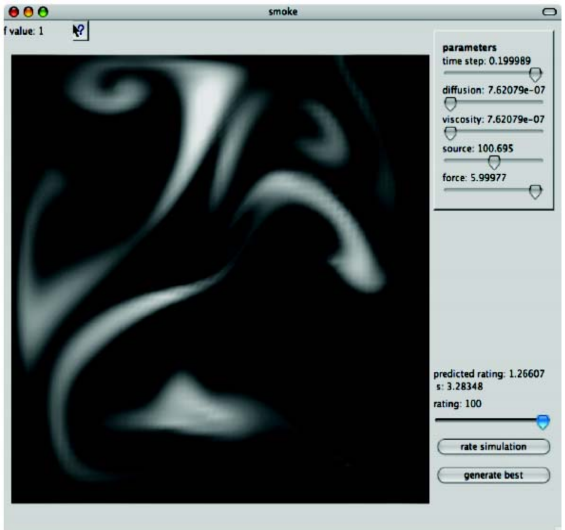


Intelligent user interfaces





Other Active Learning Problems



- Which sites should a crawler visit?
- Which tests to conduct in active diagnosis?
- What is a good animated walk?
- Interactive video search.
- Relevance feedback systems.
- Optimizing spatial and temporal allocation of sensors. How do we adapt to target maneuvers?
- Learning opponent's strategies.

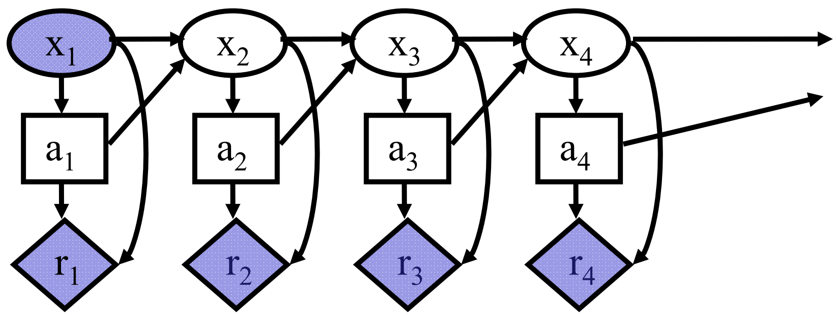
Introduction to machine learning

- What is machine learning?
- How is machine learning related to other fields?
- Machine learning applications
- Types of learning
  - Supervised learning
    - classification
    - regression
  - Unsupervised learning
    - clustering
    - data association
    - abnormality detection
    - dimensionality reduction
    - structure learning
  - Semi-supervised learning
  - Active learning
  - Reinforcement learning and control of partially observed Markov decision processes.

Partially Observed Markov Decision Processes (POMDPs)

During learning: we can estimate the **transition** p(x_t|x_t-1,a_t-1) and **reward** r(a,x) models by, say observing a human expert.

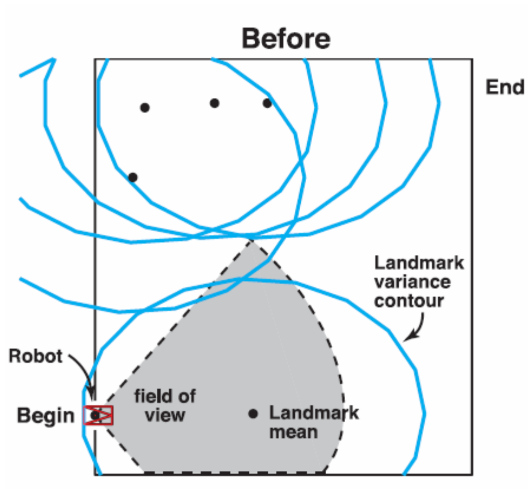
During planning: we learn the best sequence of actions (**policy**) so as to maximize the discounted sum of expected rewards.



Impressive applications (e.g. Ng et al)

Monte Carlo Planning: Problem statement

- The robot, with a rough prior map, has to accomplish a series of tasks in an environment (e.g. reaching End).
- It chooses a parameterized path $\pi^*(\boldsymbol{\theta})$ so as to learn the most about its own pose and the location of navigation landmarks (posterior map), while accomplishing tasks.



POMDP Formulation

- State (robot and landmark locations) $\mathbf{x}_t$
- Observations $\mathbf{y}_t$
- Actions (PID regulation about policy path) $\mathbf{u}_t \sim \pi(\boldsymbol{\theta})$
- Transition model $p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{u}_t)$
- Observation model $p(\mathbf{y}_t|\mathbf{x}_t)$
- Cost function: Average Mean Square Error

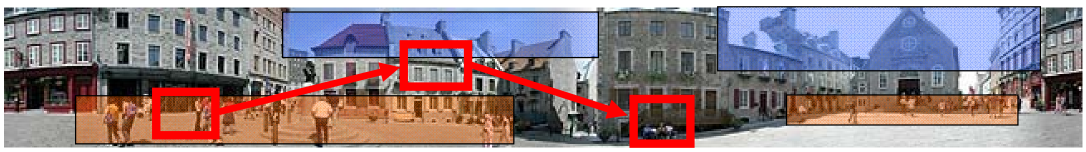
$$C^{\pi}_{AMSE} = \mathbb{E}_{p(\mathbf{x}_T, \mathbf{y}_{1:T}|\boldsymbol{\pi})}\left[(\widehat{\mathbf{x}}_T - \mathbf{x}_T)(\widehat{\mathbf{x}}_T - \mathbf{x}_T)^T\right]$$

$$\widehat{x} = \mathbb{E}_{p(\mathbf{x}_T|\mathbf{y}_{1:T}, \boldsymbol{\pi})}[\mathbf{x}_t]$$

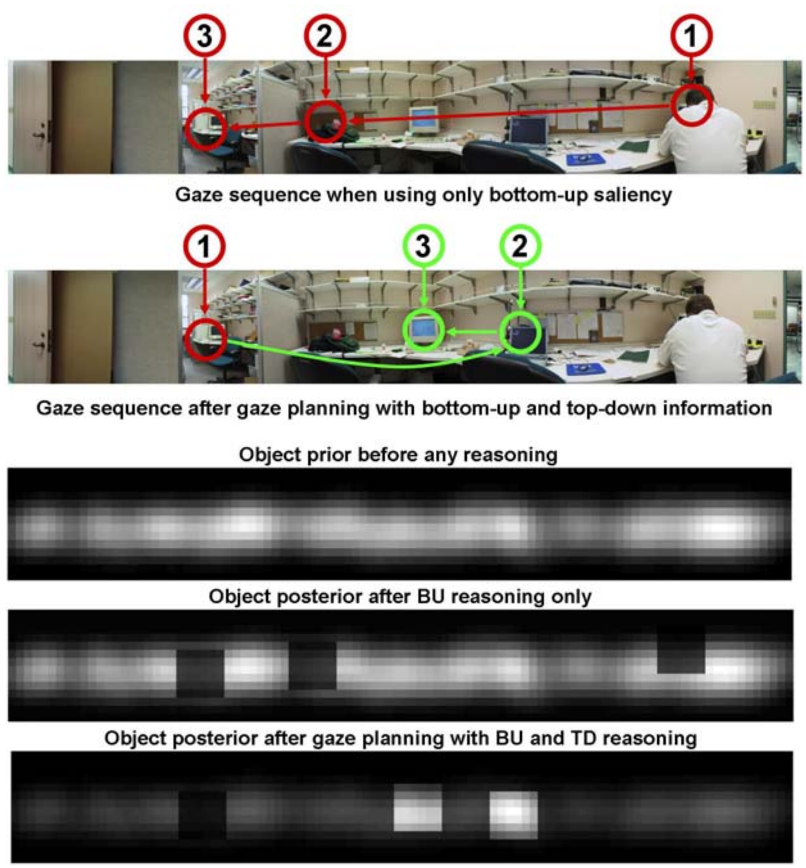
Target-directed attention

Assume, e.g. you have prior over where to find **people** and **sky**

Use Bayesian theory and maximum expected utility principle to combine this prior with bottom-up saliency maps and object likelihood models to obtain a more informative posterior.



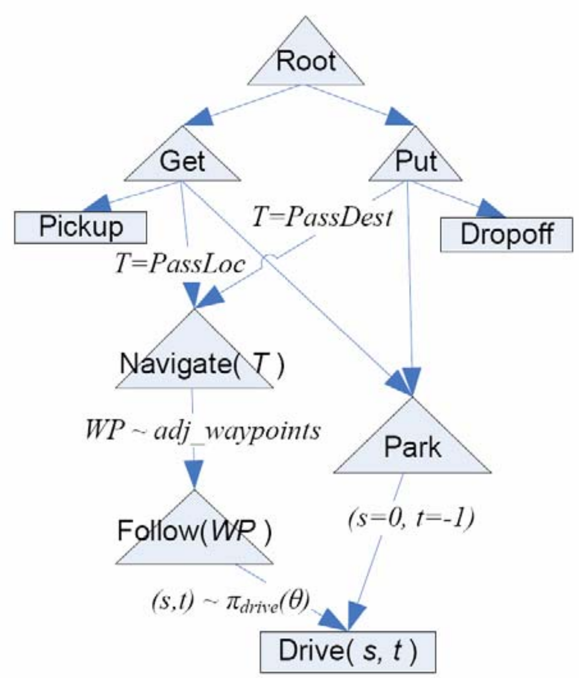
Qualitative results



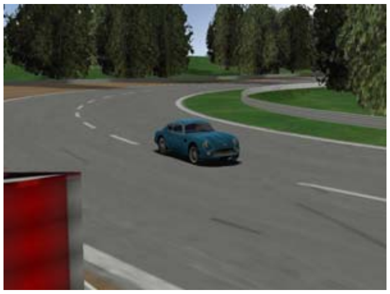
The games industry, rich in sophisticated large-scale simulators, is a great environment for the design and study of automatic decision making systems.

Hierarchical policy example

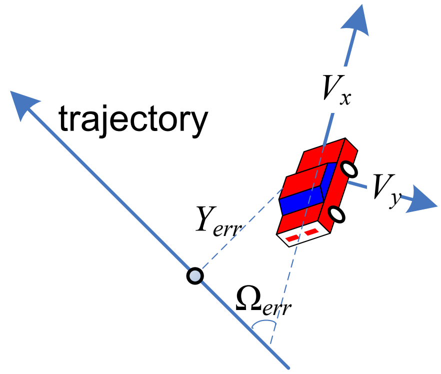
- High-level model-based learning for deciding when to navigate, park, pickup and dropoff passengers.
- Mid-level active path learning fo navigating a topological map.
- Low-level active policy optimizer to learn control of continuous non-linear vehicle dynamics.

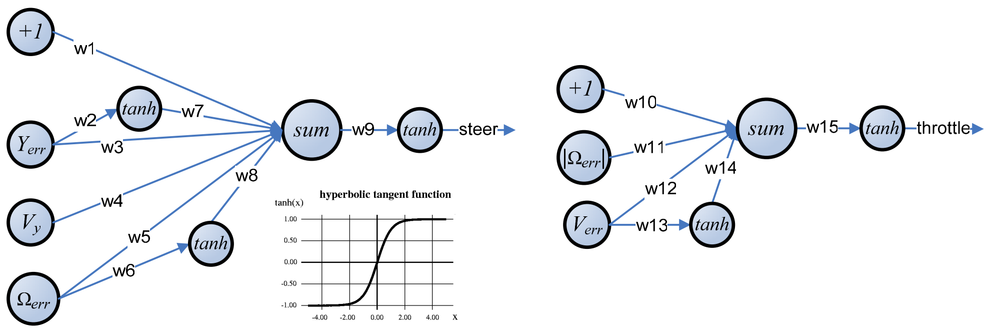


Low-Level: Trajectory following



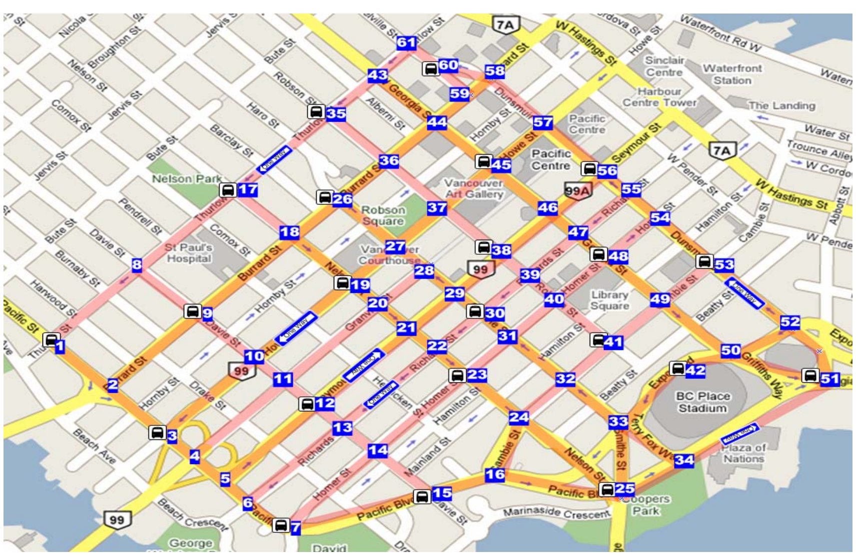
TORCS: 3D game engine that implements complex vehicle dynamics complete with manual and automatic transmission, engine, clutch, tire, and suspension models.





Active Path Finding in Middle Level

- Mid-level *Navigate* policy generates sequence of waypoints on a topological map to navigate from a location to a destination. *V(θ)* value function represents the path length from the current node, to the target.



Hierarchical systems apply to many robot tasks – key to build large systems

We used TORCS: A 3D game engine that implements complex vehicle dynamics complete with manual and automatic transmission, engine, clutch, tire, and suspension models.