.

# Machine Learning and Data Mining

Nando de Freitas
September 8, 2004

## Lecture 1 - *Introduction*

**OBJECTIVE:** Understand the several ways in which the machine learning and data mining problems arise in practice.

◇ MACHINE LEARNING AND DATA MINING

*Machine Learning and Data Mining* are the processes of deriving abstractions of the real world from a set of observations. Data mining focuses on databases. The resulting abstractions (models) are useful for

1. Making decisions under uncertainty.
2. Predicting future events.
3. Classifying massive quantities of data quickly.
4. Finding patterns (clusters, hierarchies, abnormalities, associations) in the data.
5. Developing autonomous agents (robots, game agents and other programs).

## ◇ MACHINE LEARNING AND OTHER FIELDS

Machine learning is closely related to many disciplines of human endeavor. For example:

**Information Theory** :

- Compression: Models are compressed versions of the real world.
- Complexity: Suppose we want to transmit a message over a communication channel

$$\textit{Sender} \xrightarrow{data} \textit{Channel} \xrightarrow{data} \textit{Receiver}$$

To gain more efficiency, we can compress the data and send both the compressed data and the model to decompress the data.

$$\textit{Sender} \xrightarrow{data} \textit{Encoder} \xrightarrow[model]{comp.\ data} \textit{Channel} \xrightarrow[model]{comp.\ data} \textit{Decoder} \xrightarrow{data} \textit{Receiver}$$

There is a fundamental tradeoff between the amount of compression and the cost of transmitting the model. More complex models allow for more compression, but are expensive to transmit. Learners that balance these two costs tend to perform better in the future. That is, they *generalise* well.

**Probability Theory** :

- Modelling noise.
- Dealing with uncertainty: occlusion, missing data, synonymy and polisemy, unknown inputs.

SHOULD I HAVE THE SOUP? EVERYTHING ELSE IS SO EXPENSIVE, AND I DON'T KNOW WHO'S PAYING... ARE STATISTICIANS STINGY? I'VE NEVER GONE OUT WITH ONE BEFORE... THOUGH I ONCE KNEW A VERY GENEROUS ACCOUNTANT...
SHOULD I HAVE THE SOUP? 27 OUT OF THE 36 TIMES I'VE HAD IT, IT WAS PRETTY GOOD... BUT IS MONDAY THE REGULAR CHEF'S NIGHT OFF? AND WHAT IF ALL THE AIR MOLECULES IN THE ROOM SUDDENLY FLY UP TO THE CEILING?

**Statistics** :

- *Data Analysis and Visualisation*: gathering, display and summary of data.
- *Inference*: drawing statistical conclusions from specific data.

**Computer Science** :

- Theory.
- Database technology.
- Software engineering.
- Hardware.

**Optimisation** : Searching for optimal *parameters* and models in constrained and unconstrained settings is ubiquitous in machine learning.

**Philosophy** : The study of the nature of knowledge (epistemology) is central to machine learning. Understanding the learning process and the resulting abstractions is a

question of fundamental importance to human beings. At the onset of Western philosophy, Plato and Aristotle distinguished between "essential" and "accidental" properties of things. The Zen patriarch, Bodhidharma also tried to get to the essence of things by asking "what is that?" in a serious sense, of course.

**Other Branches of Science** :

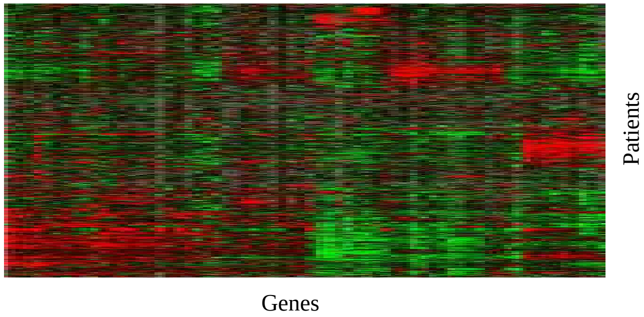
- Game theory.
- Econometrics.
- Cognitive science.
- Engineering.
- Psychology.
- Biology.

◇ APPLICATION AREAS

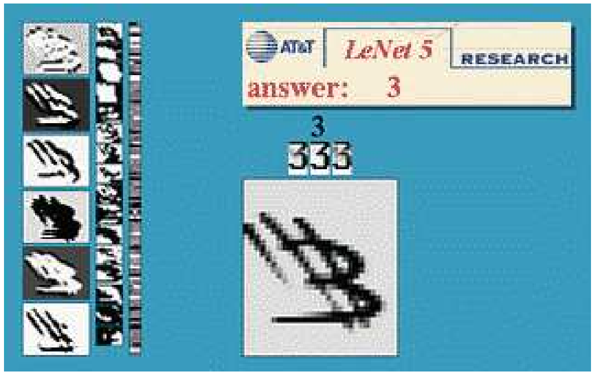
Machine learning and data mining play an important role in the following fields:

**Software** : Teaching the computer instead of programming it.

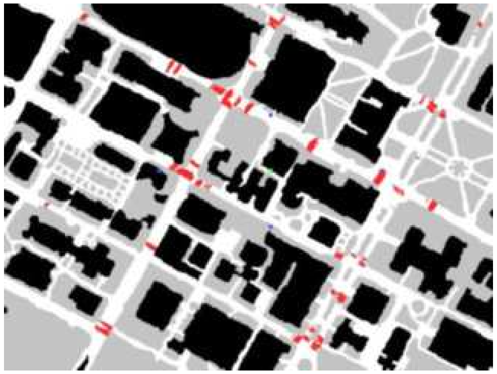
**Bioinformatics** : Sequence alignment, DNA micro-arrays, drug design, novelty detection.



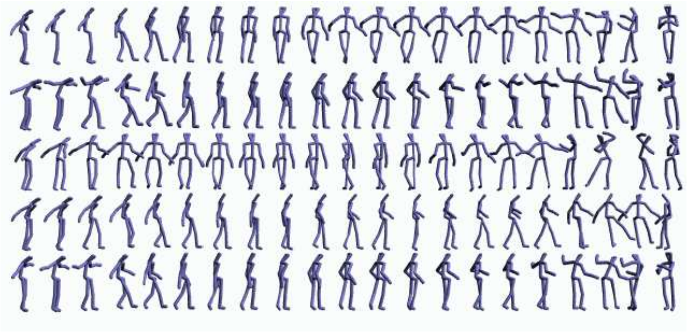
**Computer Vision** : Handwritten digit recognition (Le Cun), tracking, segmentation, object recognition.

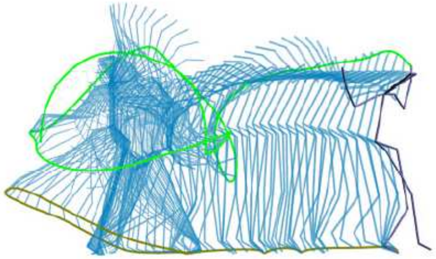


**Robotics** : State estimation, control, localisation and map building.



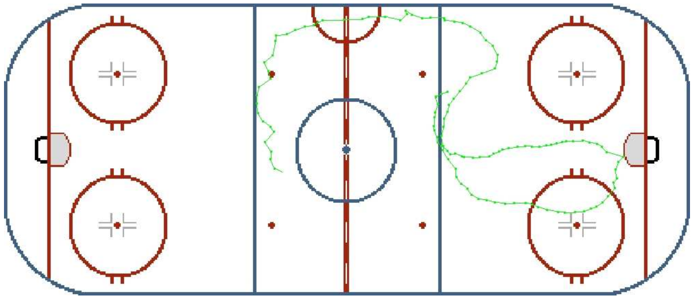
**Computer Graphics** : Automatic motion generation, realistic simulation. E.g., style machines by Brand and Hertzmann:





**Electronic Commerce** : Data mining, collaborative filtering, recommender systems, spam.

**Computer Games** : Intelligent agents and realistic games.

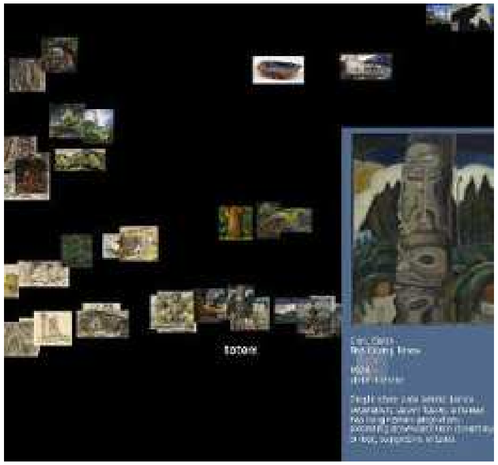


**Financial Analysis** : Options and derivatives, forex, portfolio allocation.

**Medical Sciences** : Epidemiology, diagnosis, prognosis, drug design.

**Speech** : Recognition, speaker identification.

**Multimedia** : Sound, video, text and image databases; multimedia translation, browsing, information retrieval (search engines).



## ◇ TYPES OF LEARNING

### Supervised Learning

We are given input-output *training* data $\{x_{1:N}, y_{1:N}\}$, where $x_{1:N} \triangleq (x_1, x_2, \ldots, x_N)$. That is, we have a teacher that tell us the outcome $y$ for each input $x$. Learning involves adapting the model so that its predictions $\widehat{y}$ are close to $y$. To achieve this we need to introduce a $loss function$ that tells us how close $\widehat{y}$ is to $y$. Where does the loss function come from?

$$x \longrightarrow Model \longrightarrow \widehat{y}$$

After learning the model, we can apply it to novel inputs and study its response. If the predictions are accurate we have reason to believe the model is correct. We can exploit this during training by splitting the dataset into a training set and a test set. We learn the model with the training set and validate it with the test set. This is an example of a model selection technique knowns as *cross-validation*.

What are the advantages and disadvantages of this technique?

In the literature, inputs are also known as predictors, explanatory variables or covariates, while outputs are often referred to as responses or variates.

### Unsupervised Learning

Here, there is not teacher. The learner must identify structures and patterns in the data. Many times, there is no single correct answer. Examples of this include image segmentation and data clustering.

### Semi-supervised Learning

It's a mix of supervised and unsupervised learning.

### Reinforcement Learning

Here, the learner is given a reward for an action performed in a particular environment. Human cognitive tasks as well as “simple” motor tasks like balancing while walking seem to make use of this learning paradigm. RL, therefore, is likely to play an important role in graphics and computer games in the future.

**Active Learning**

$$World \xrightarrow{data} Passive\ Learner \longrightarrow Model$$

$$World \overset{data}{\underset{query}{\rightleftarrows}} Active\ Learner \longrightarrow Model$$

Active learners query the environment. Queries include questions and requests to carry out experiments. As an analogy, I like to think of good students as active learners! But, how do we select queries optimally? That is, what questions should we ask? What is the price of asking a question?

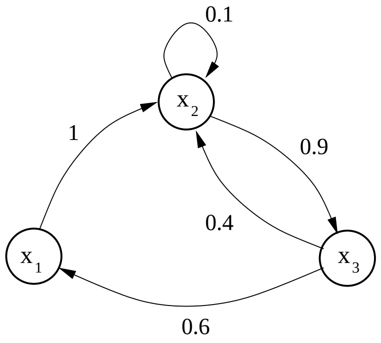
Active learning plays an important role when establishing *causal* relationships.

# Lecture 2 - *Google’s PageRank: Why math helps*

**OBJECTIVE:** Motivate linear algebra and probability as important and necessary tools for understanding large datasets. We also describe the algorithm at the core of the Google search engine.

◇ PAGERANK

Consider the following mini-web of 3 pages (the data):



The nodes are the webpages and the arrows are links. The numbers are the normalised number of links. We can re-write this directed graph as a **transition matrix**:

★

$T$ is a **stochastic matrix**: its columns add up to 1, so that

$$T_{i,j} = P(x_j | x_i)$$

$$\sum_j T_{i,j} = 1$$

In information retrieval, we want to know the "relevance" of each webpage. That is, we want to compute the probability of each webpage: $p(x_i)$ for $i = 1, 2, 3$.

Let's start with a random guess $\pi = (0.5, 0.2, 0.3)$ and "crawl the web" (multiply by $T$ several times). After, say $N = 100$,

iterations we get:

$$\pi^T T^N = (0.2, 0.4, 0.4)$$

We soon notice that no matter what initial $\pi$ we choose, we always converge to $p = (0.2, 0.4, 0.4)$. So

★

$$p^T T =$$

★

The distribution $p$ is a measure of the relevance of each page. Google uses this. But will this work always? When does it fail?

★

The **Perron-Frobenius Theorem** tell us that for any starting point, the chain will converge to the invariant distribution $p$, as long as $T$ is a stochastic transition matrix that obeys the following properties:

1. **Irreducibility**: For any state of the Markov chain, there is a positive probability of visiting all other states. That is, the matrix $T$ cannot be reduced to separate

smaller matrices, which is also the same as stating that the transition graph is connected.

2. **Aperiodicity**: The chain should not get trapped in cycles.

Google's strategy is to add am matrix of uniform noise $E$ to $T$:

$$L = T + \epsilon E$$

where $\epsilon$ is a small number. $L$ is then normalised. This ensures irreducibility.
How quickly does this algorithm converge? What determines the rate of convergence? Again matrix algebra and spectral theory provide the answers:

⋆