

# Machine Learning - Intro

Aarti Singh, Eric Xing

Machine Learning 10-701/15-781  
Sept 10, 2012



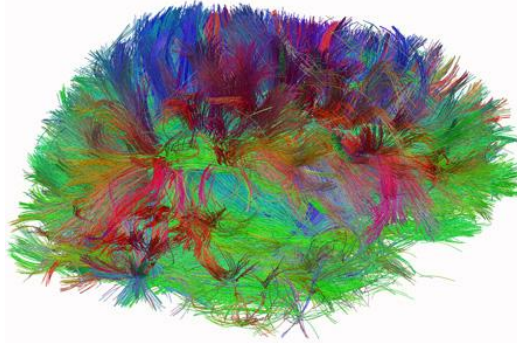
**MACHINE LEARNING** DEPARTMENT



# The Age of Big Data



CERN Collider  
 $320 \times 10^{12}$  bytes/second



Personal Connectome  
 $10^{18}$  bytes/human

facebook

1 billion  
messages/day

twitter

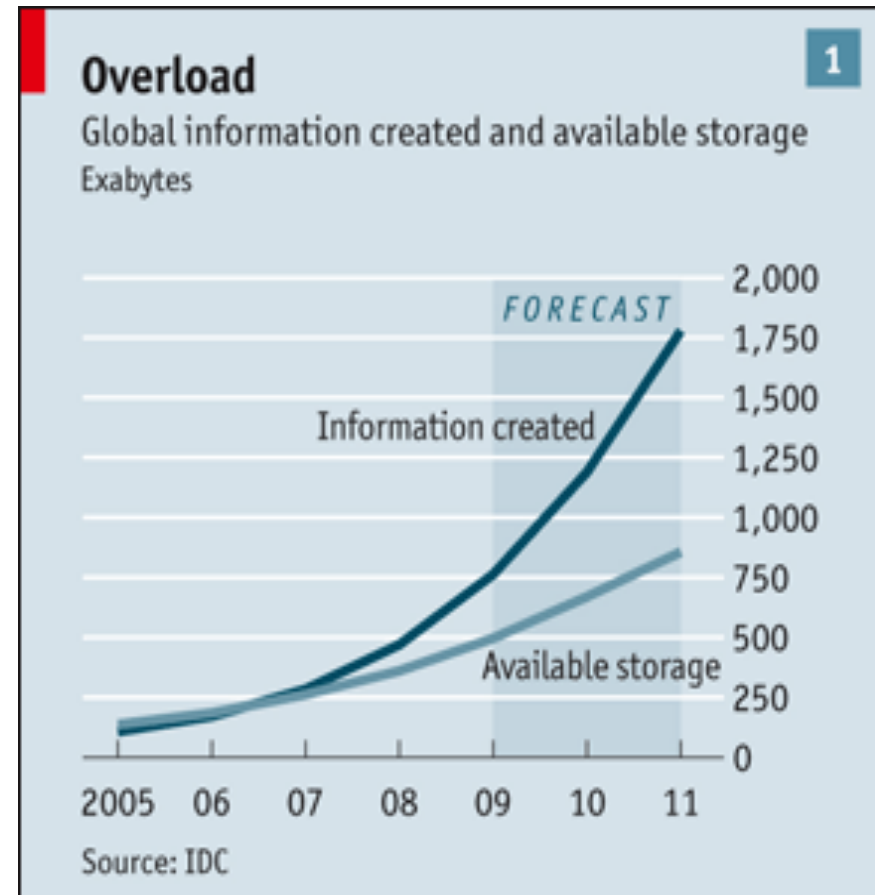
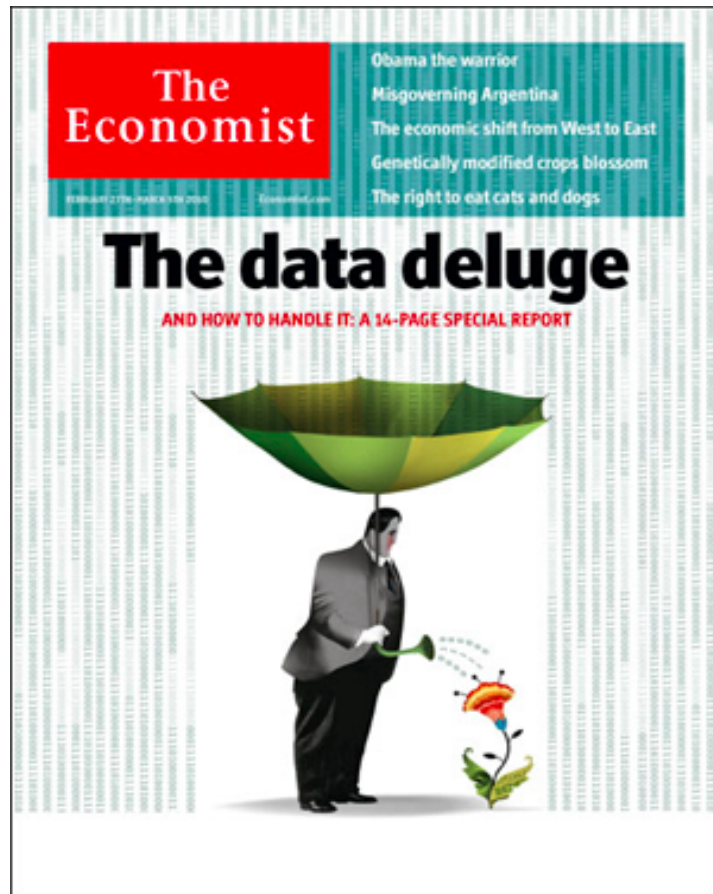


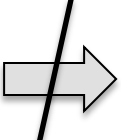
200 million  
tweets/day

“Every day, people create the equivalent of 2.5 **quintillion** bytes of data from sensors, mobile devices, online transactions, and social networks:

*The Huffington Post: Arnal Dayaratna: IBM Releases Big Data*

# The Age of Big Data




Data  Knowledge

# **From Data to Knowledge ...**



## A cartoon illustration depicting a 'Learning Machine'. A woman in a red dress stands on a wooden ladder, pouring a bucket of 'Data' (represented by papers with labels like 'Air Lines', 'Pork', and 'Pom') into the top of a large, boxy machine. The machine has a yellow funnel on its left side, from which a man in a suit is shouting 'MORE, MORE, MORE!!' while holding a block labeled 'Class Action Suits'. The machine itself has a clock face and several knobs. At the bottom, it outputs several blocks labeled 'Predictions' and 'Hypothesis'. The machine is plugged into a wall outlet. The text 'Learning Machine' is written across the front of the machine in a colorful, stylized font. The signature 'ALPOD' is in the bottom right corner.

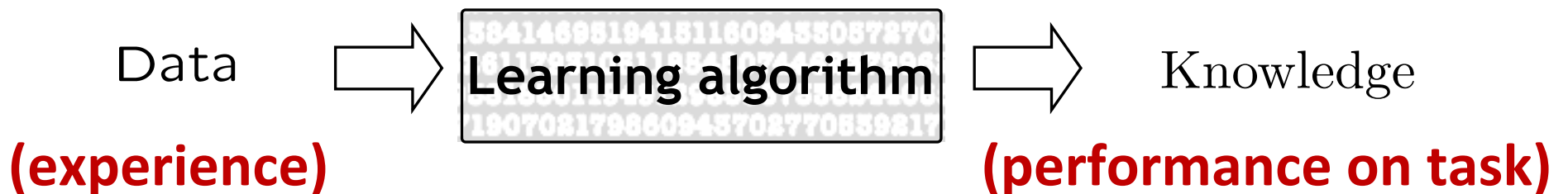


Learning algorithm

# What is Machine Learning?

Design and Analysis of algorithms that

- improve their performance
- at some task
- with experience

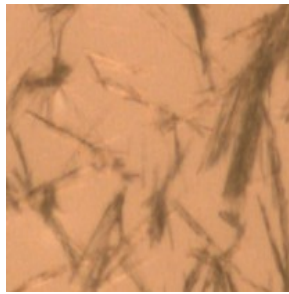


# Human learning

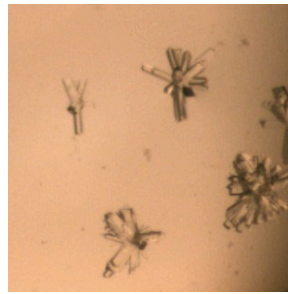
**Task:** Learning stage of protein crystallization



Crystal



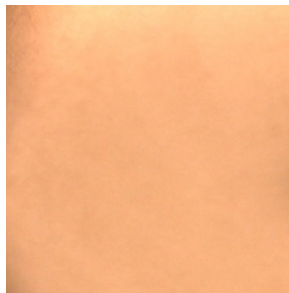
Needle



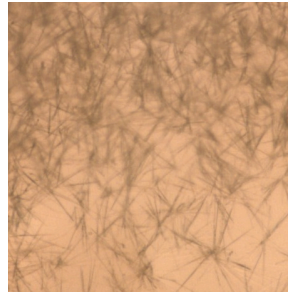
Tree



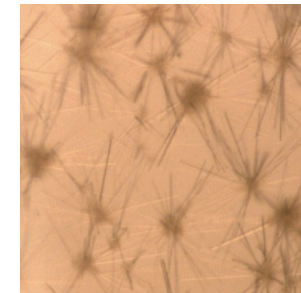
Tree



Empty



Needle



?

**Experience**

**Performance**

# Machine Learning in Action

# Machine Learning in Action

- Document classification



Sports  
Science  
News

# Machine Learning in Action

- Spam filtering

## Welcome to New Media Installation: Art that Learns

Hi everyone,

Welcome to New Media Installation: Art that Learns

The class will start tomorrow.

\*\*\*Make sure you attend the first class, even if you are on the Wait List.\*\*\*

The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: [10615-announce@cs.cmu.edu](mailto:10615-announce@cs.cmu.edu).

**Natural \_LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only \$5.95 for shipping mfw rlk** Spam | X

=== Natural WeightLOSS Solution ===

Vital Acai is a natural WeightLOSS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

- \* Rapid WeightLOSS
- \* Increased metabolism - BurnFat & calories easily!
- \* Better Mood and Attitude

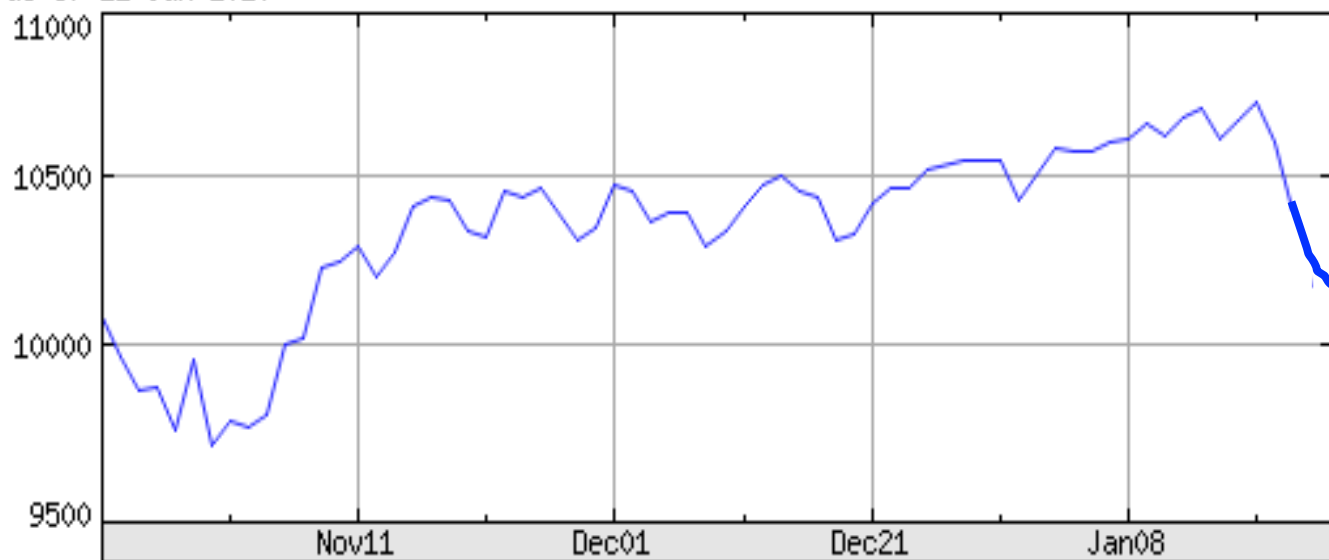


Spam/  
Not spam

# Machine Learning in Action

- Stock Market Prediction

DJ INDU AVERAGE (DOW JONES & CO)  
as of 22-Jan-2010



Copyright 2010 Yahoo! Inc.

<http://finance.yahoo.com/>

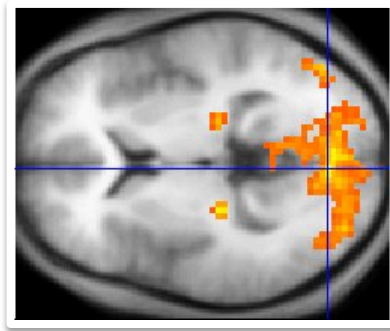
$Y = ?$

$X = \text{Feb01}$



# Machine Learning in Action

- Decoding thoughts from brain scans



Rob a bank ...

[Home](#) » [Health & Wellness](#)

## Brain Scans: Are You a Criminal?



Published February 07, 2007 by:

[Andrea Okrentowich](#)

[View Profile](#) | [Follow](#) | [Add to Favorites](#)

More:

[Brain Scans](#)

[Brain Scan](#)

[Disposition](#)

[Defendant](#)

[Criminal Behavior](#)

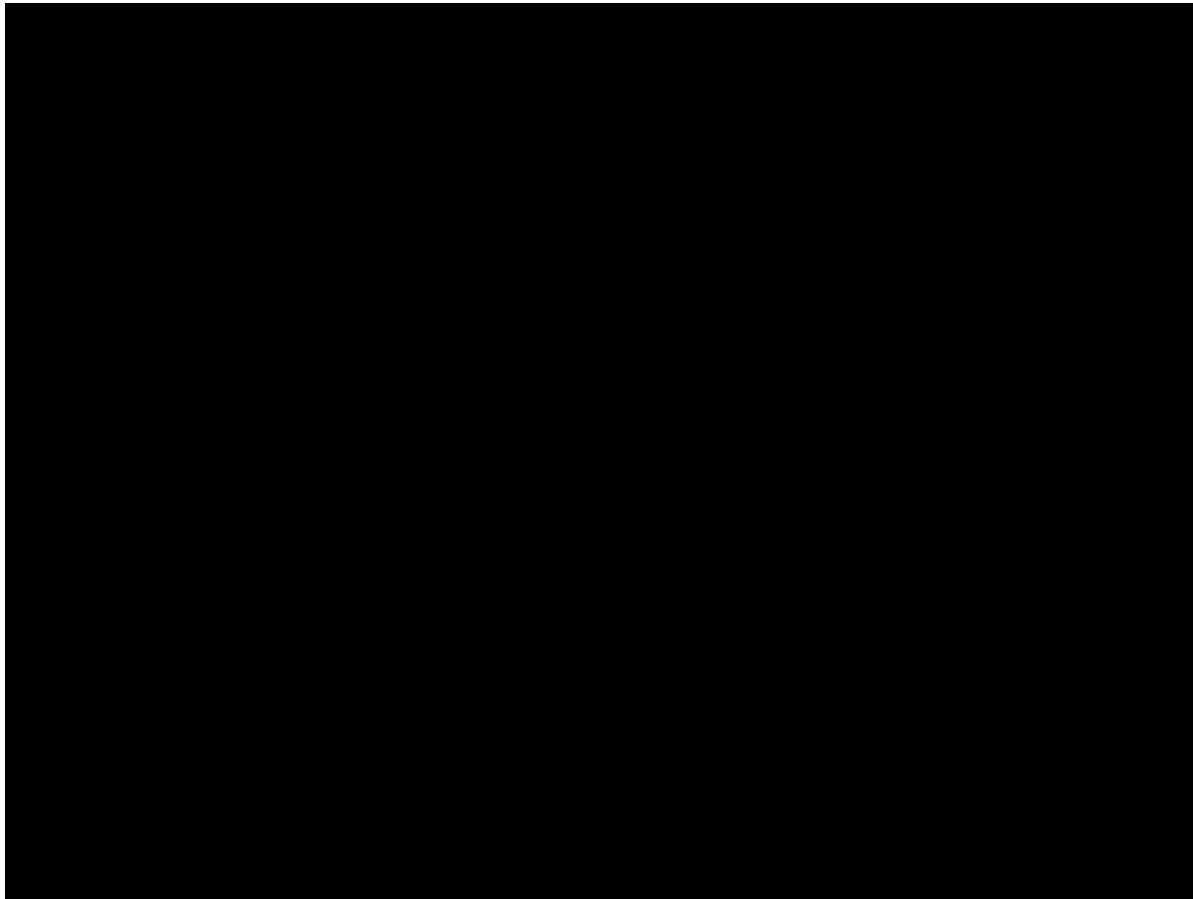
### MRI Scans as Courtroom Evidence

The average Joe's MRI scan can show a brain abnormality, do we proceed to check him into the nearest mental institution or prison? That would make about as much sense as trying to prove a defendant innocent of a violent



# Machine Learning in Action

- The **best** helicopter pilot is now a computer!



<http://heli.stanford.edu/>

# Machine Learning in Action

- Many, many more...

Speech recognition, Natural language processing

Computer vision

Robotics

Web forensics

Medical data analysis

Computational biology

Sensor networks

Social networks

...

# ML is trending!

- Wide applicability
- Very large-scale complex systems
  - Internet (billions of nodes), sensor network (new multi-modal sensing devices), genetics (human genome)
- Huge multi-dimensional data sets
  - 30,000 genes x 10,000 drugs x 100 species x ...
- Software too complex to write by hand
- Improved machine learning algorithms
- Improved data capture (Terabytes, Petabytes of data), networking, faster computers
- Demand for self-customization to user, environment

**Are we there yet?**


# ML has a long way to go ...

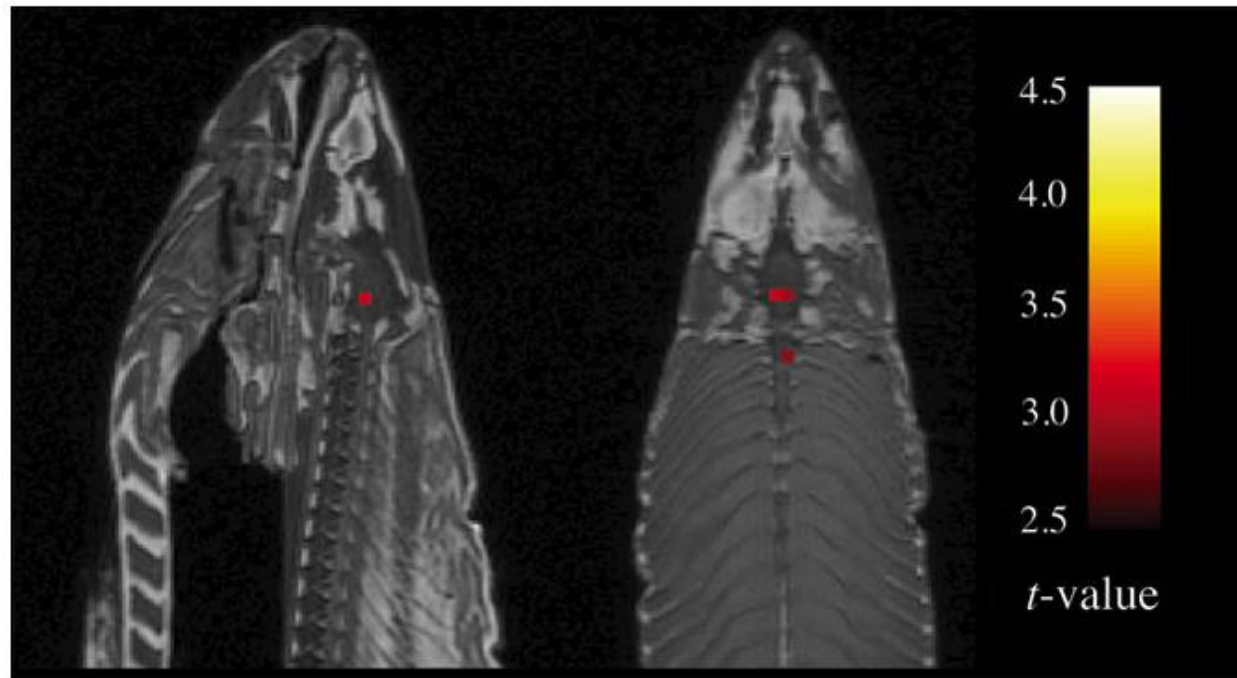
## WIRED SCIENCE

NEWS FOR YOUR NEURONS



### Scanning Dead Salmon in fMRI Machine Highlights Risk of Red Herrings

By Alexis Madrigal  September 18, 2009 | 5:37 pm | Categories: [Brains](#) and [Behavior](#)



# ML has a long way to go ...

## Speech Recognition gone Awry

[http://www.google.com/url?sa=t&source=web&cd=5&ved=0CCwQtwlwBA&url=http://video.google.com/videoplay?docid=-1123221217782777472&rct=j&q=bad%20speech%20recognition&ei=nvyGTN3kOMOAlAezu\\_HHDg&usg=AFQjCNHDTf0w6VudgJfbP3xAvDTFhbzWCQ&cad=rja](http://www.google.com/url?sa=t&source=web&cd=5&ved=0CCwQtwlwBA&url=http://video.google.com/videoplay?docid=-1123221217782777472&rct=j&q=bad%20speech%20recognition&ei=nvyGTN3kOMOAlAezu_HHDg&usg=AFQjCNHDTf0w6VudgJfbP3xAvDTFhbzWCQ&cad=rja)



# Machine Learning Tasks

Broad categories -

- **Supervised learning**

Classification, Regression

- **Unsupervised learning**

Density estimation, Clustering, Dimensionality reduction

- Semi-supervised learning
- Active learning
- Reinforcement learning
- Many more ...

# Supervised Learning

Feature Space  $\mathcal{X}$

Words in a document

Label Space  $\mathcal{Y}$

"Sports"  
"News"  
"Science"  
...

Discrete Labels  
**Classification**



DJ INDU AVERAGE (DOW JONES & CO  
as of 22-Jan-2010



Share Price  
"\$ 24.50"

Continuous Labels  
**Regression**



**Task:** Given  $X \in \mathcal{X}$ , predict  $Y \in \mathcal{Y}$ .

# Unsupervised Learning

Aka “learning without a teacher”

Feature Space  $\mathcal{X}$

Words in a document

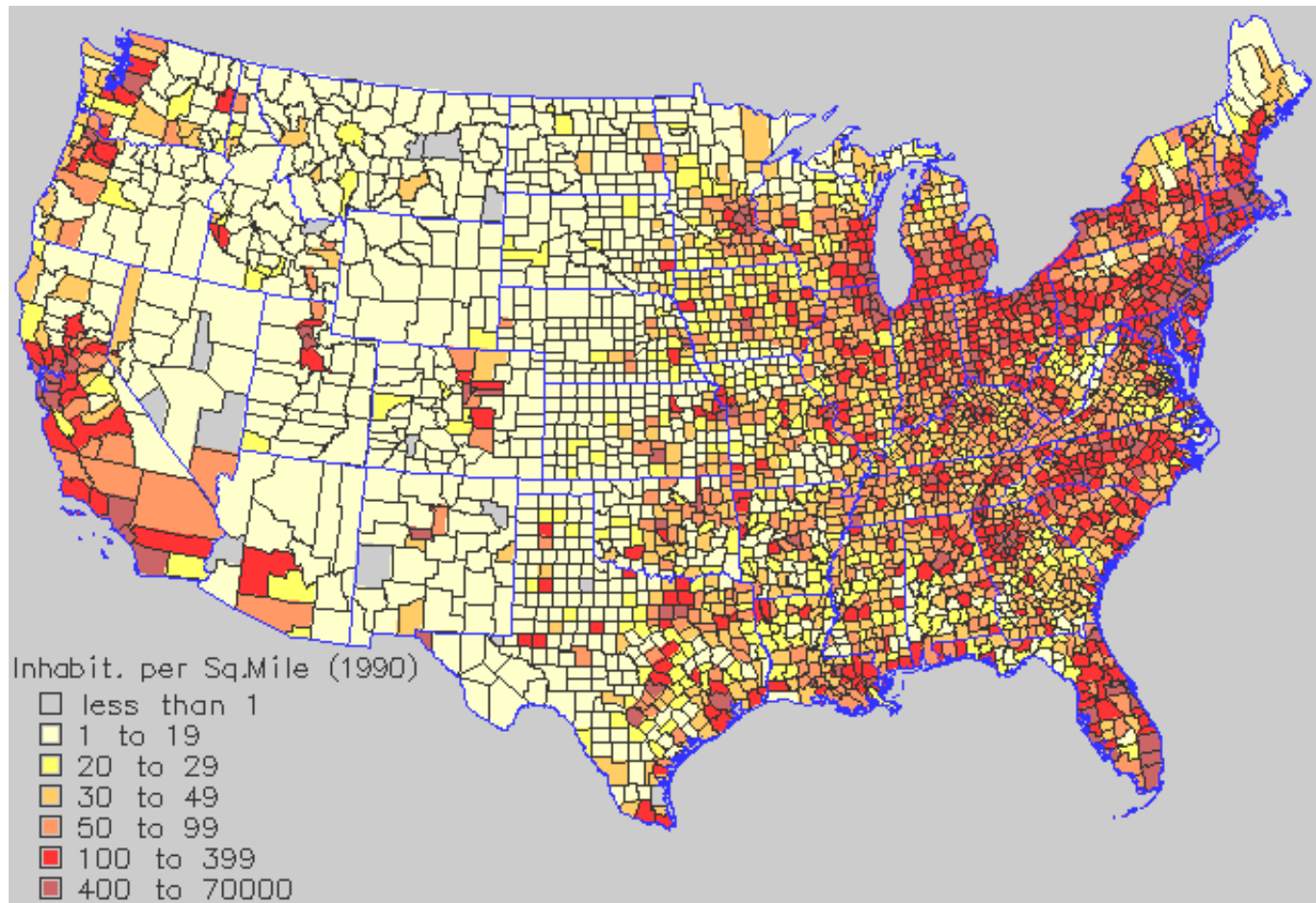


Word distribution  
(Probability of a word)

**Task:** Given  $X \in \mathcal{X}$ , learn  $f(X)$ .

# Unsupervised Learning

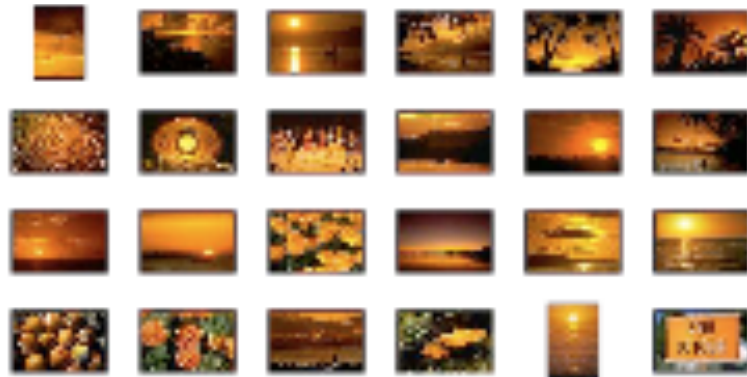
## Density/Distribution Estimation



# Unsupervised Learning

**Clustering** - Group similar things e.g. images

[Goldberger et al.]



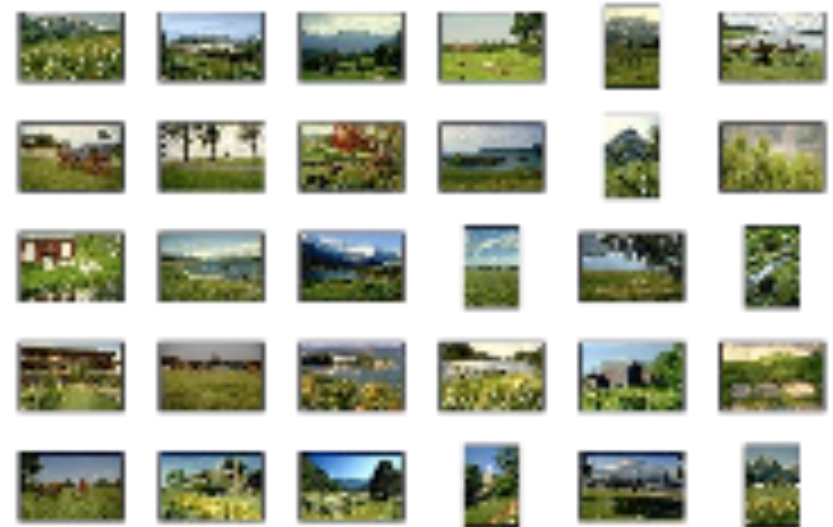
$C_4$



$C_2$



$C_3$



$C_5$

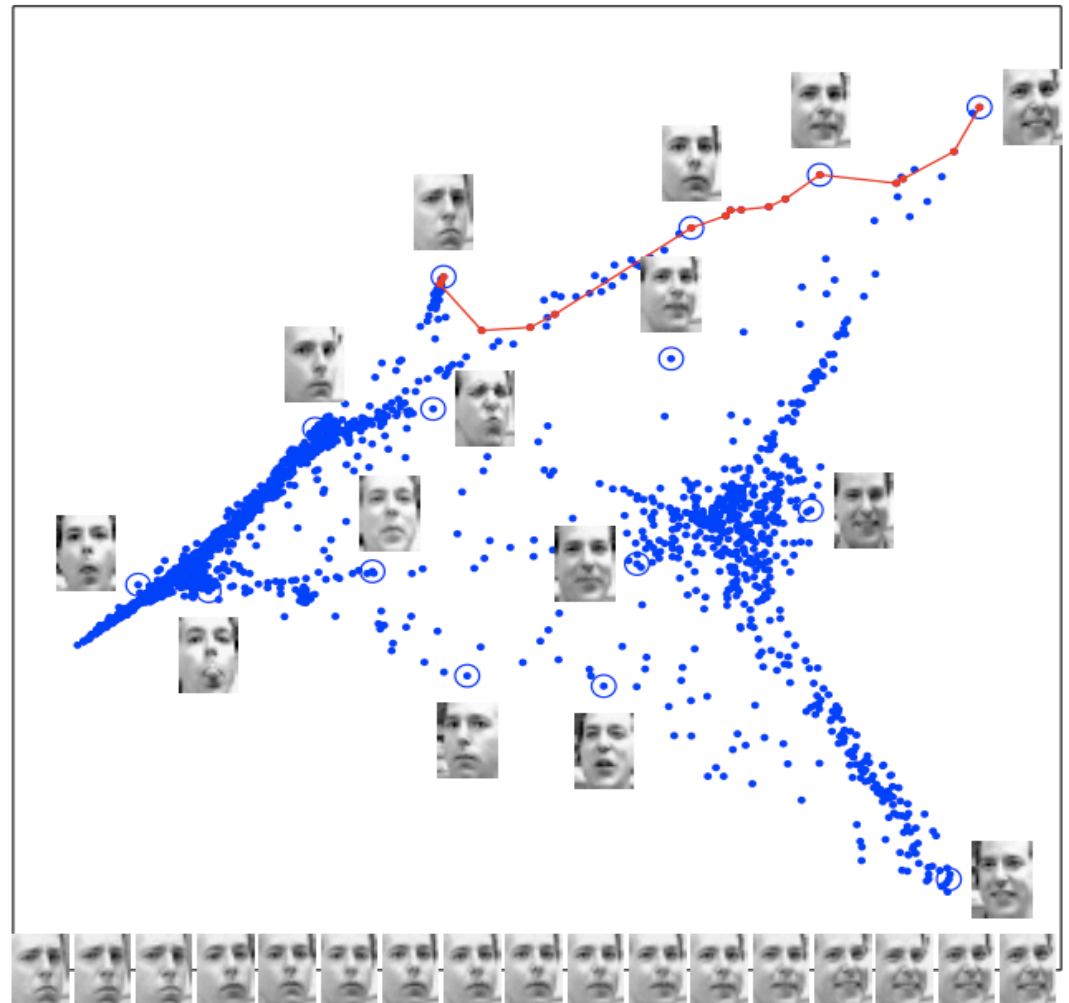
# Unsupervised Learning

## Dimensionality Reduction/Embedding

[Saul & Roweis '03]

Images have thousands or millions of pixels.

Can we give each image a coordinate, such that similar images are near each other?



# Machine Learning Tasks

Broad categories -

- **Supervised learning**

Classification, Regression

- **Unsupervised learning**

Density estimation, Clustering, Dimensionality reduction

- Semi-supervised learning
- Active learning
- Reinforcement learning
- Many more ...



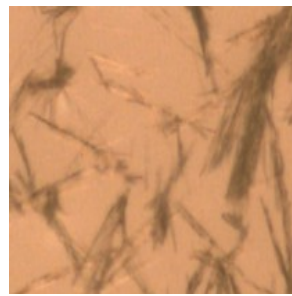
# **Key Issues in Machine Learning**

# Training Data vs. Test Data

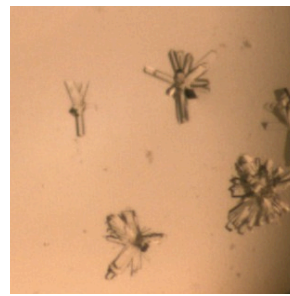
**Task:** Learning stage of protein crystallization



Crystal



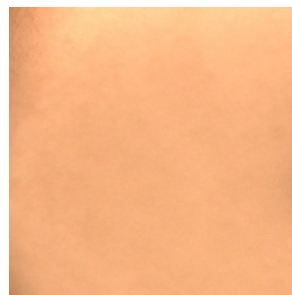
Needle



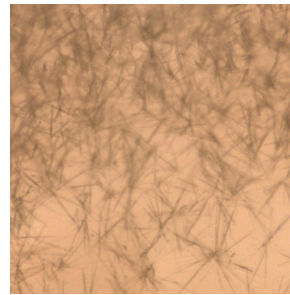
Tree



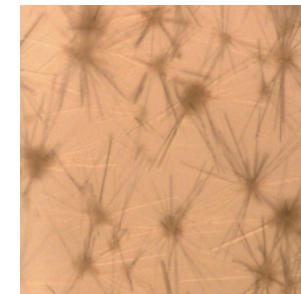
Tree



Empty



Needle

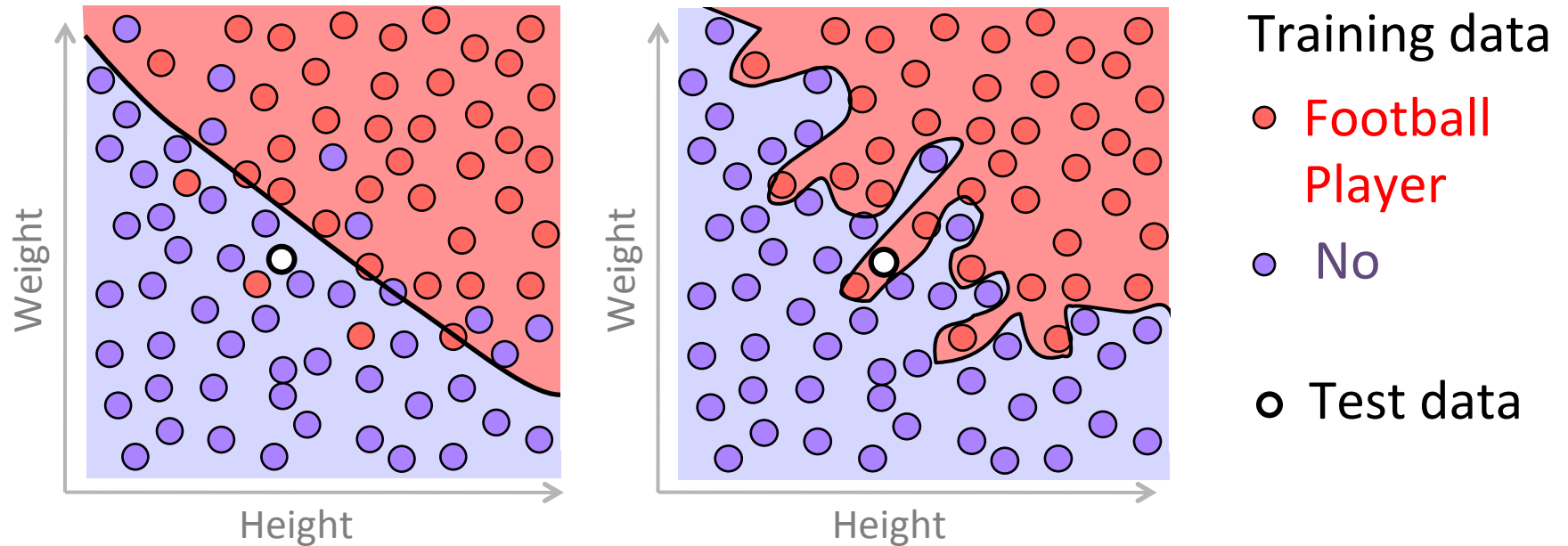


?

**Experience**

**Performance**

# Training Data vs. Test Data



- A good machine learning algorithm
  - Does not **overfit** training data
  - **Generalizes** well to test data

# Performance Measure

For a random test data  $X$ , measure of closeness between true label  $Y$  and prediction  $f(X)$

Binary Classification  $\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}}$  **0/1 loss**

Regression  $\text{loss}(Y, f(X)) = (f(X) - Y)^2$  **square loss**

Density Estimation  $\text{loss}(f(X)) = -\log(\mathbb{P}_f(X))$  **Negative log likelihood loss**

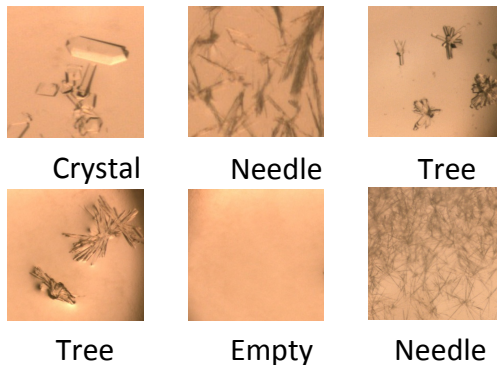
# Machine Learning vs. Optimization

Ideal goal: Construct **prediction rule**  $f : \mathcal{X} \rightarrow \mathcal{Y}$   
that works well for any test data point  $(X, Y) \sim P_{XY}$

Simply an optimization problem:  $\min_f \mathbb{E}_{XY} [\text{loss}(Y, f(X))]$

**BUT... Optimization depends on unknown  $P_{XY}$  !**

Training data (experience) provides a glimpse of  $P_{XY}$



$$\{(X^{(j)}, Y^{(j)})\}_{j=1}^n \sim P_{XY}$$

# Enjoy!

- ML is becoming ubiquitous in science, engineering and beyond
- This class should give you the basic foundation for applying ML and developing new methods
- The fun begins...