

# Computational Data Analysis

## Machine Learning

**Yao Xie, Ph.D.**

*Associate Professor*

Harold R. and Mary Anne Nash Early Career Professor  
H. Milton Stewart School of Industrial and Systems  
Engineering

Introduction

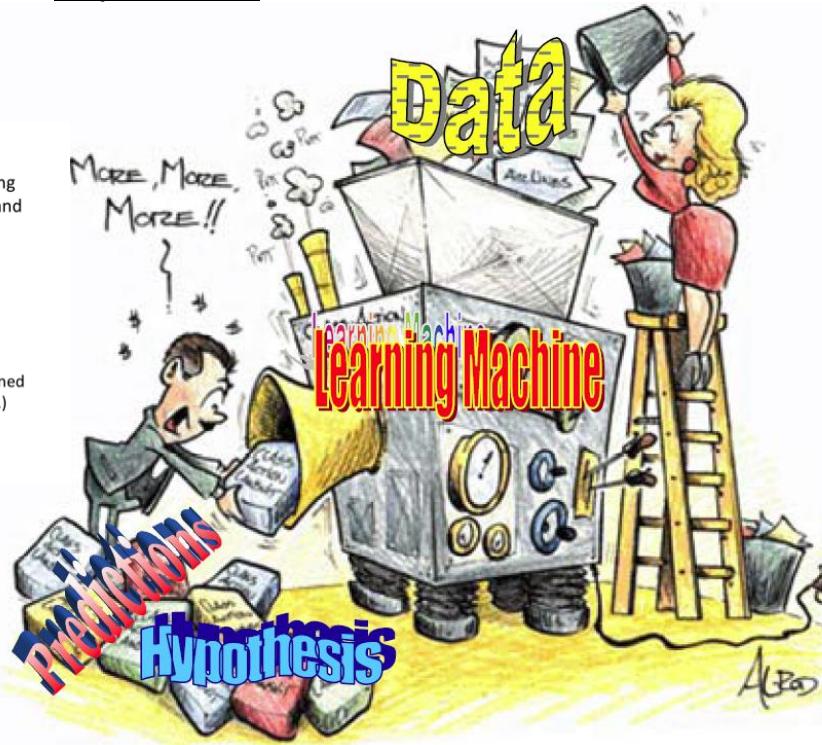


# What is machine learning (ML)

- Study of algorithms that improve their performance at some task with experience

## Syllabus

- Cover a number of most commonly used machine learning algorithms in a sufficient amount of details about math and theory.
- Organization
  - Unsupervised learning (data exploration)
    - Learning without labels or feedback
  - Supervised learning (predictive models)
    - Learning with labels, focusing on predictive performance
  - Advanced machine learning methods (nonlinearity, combined models, advanced statistical models for complex data, etc.)
    - Nonlinearity, complex data, real-world applications
- Basic optimization and math background



# Common to industrial scale problems



13 million wikipedia pages



800 million users



6 billion photos



340 million tweets per day

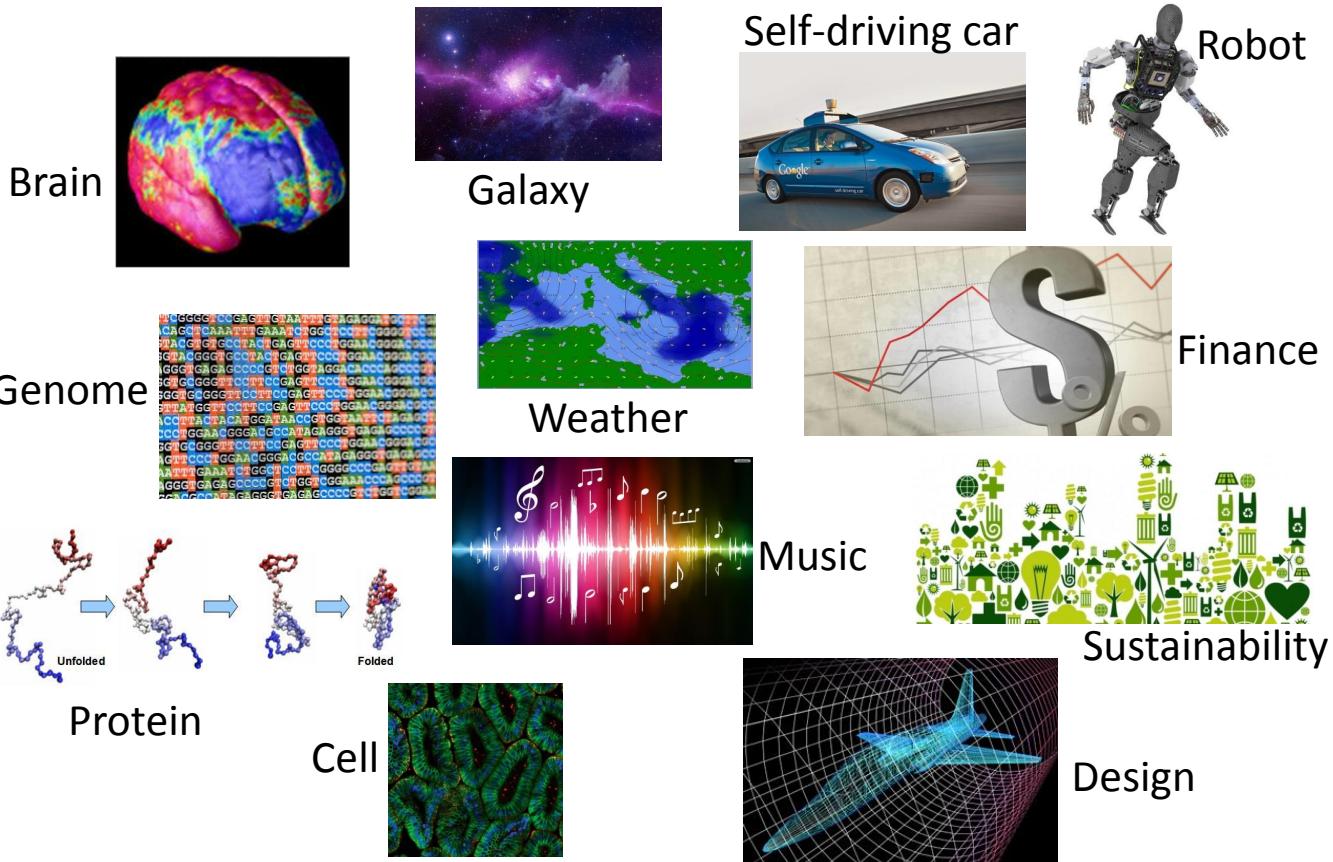


24 hours video uploaded per minutes



> 1 trillion webpages

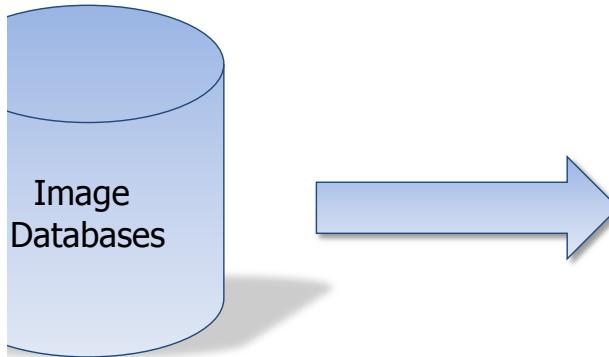
# Increasingly relevant to science problems



# Organizing Images

## Syllabus: unsupervised learning

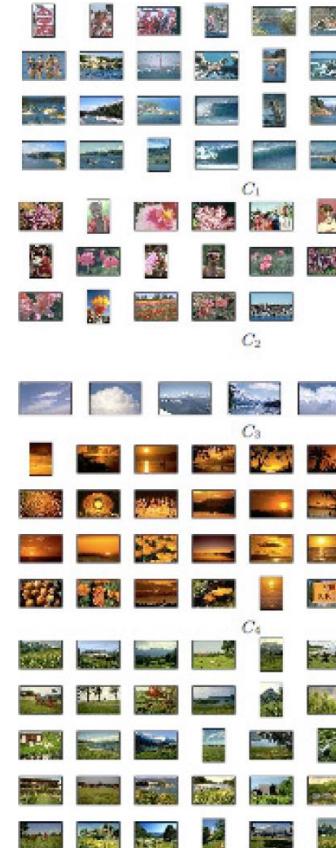
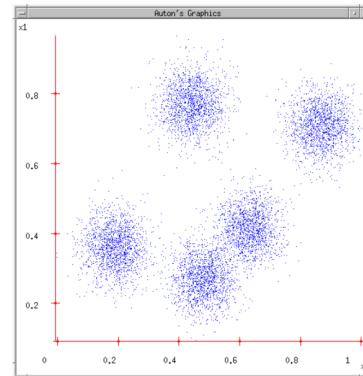
- Learning without labels or without optimizing for predictive task
  - Clustering vectorial data
    - Kmeans
    - Hierarchical clustering
  - Clustering networks
    - Spectral algorithm
  - Dimensionality reduction,
    - Principal component analysis
  - Dimensionality for manifold
    - Locally linear embedding
  - Density estimation
    - Feature selection
    - Novelty/abnormality detection



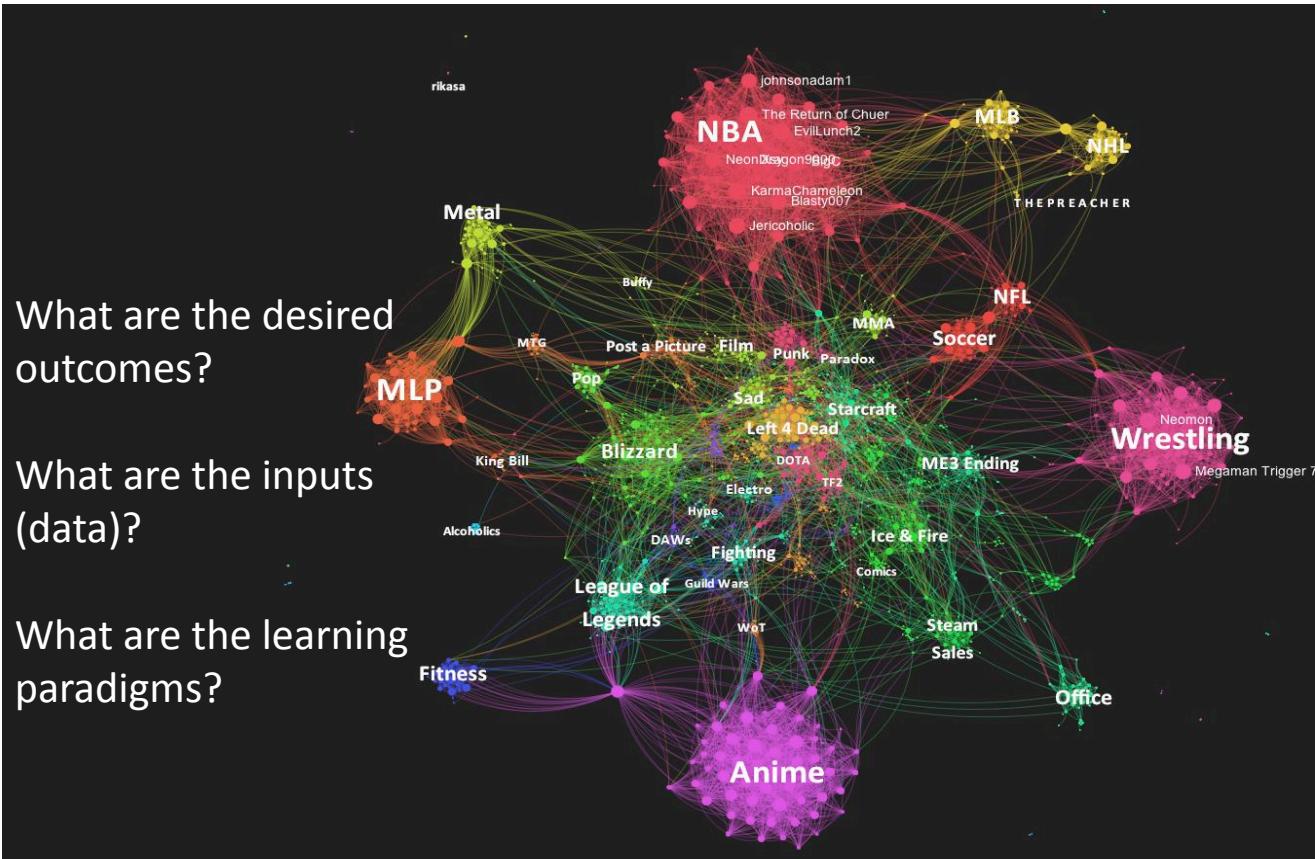
What are the desired outcomes?

What are the inputs (data)?

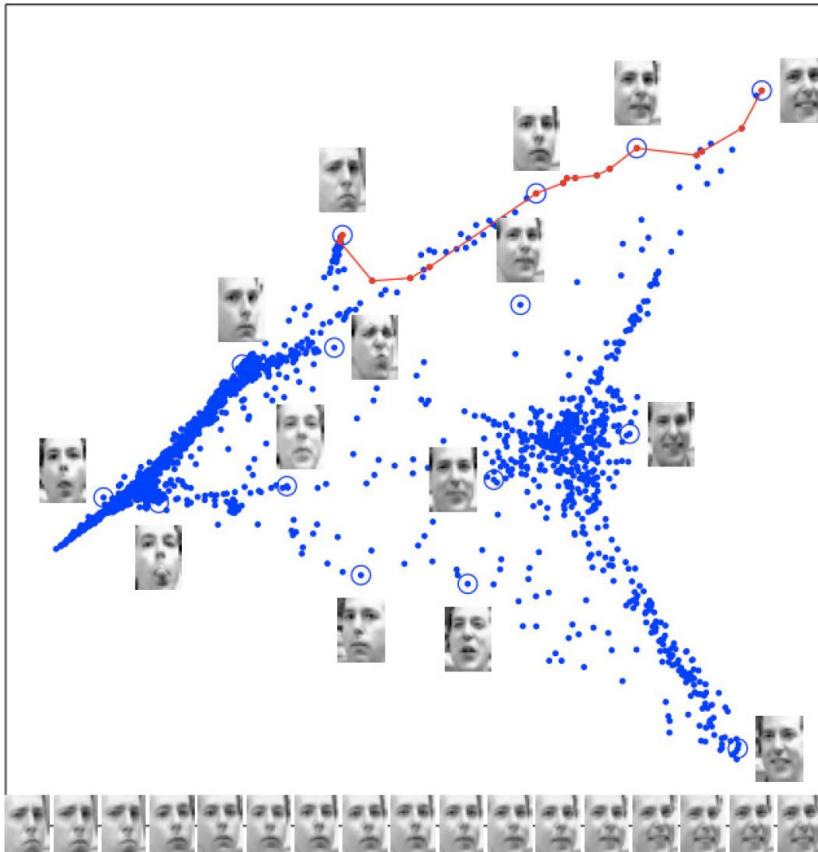
What are the learning paradigms?



# Find community in social networks



# Visualize Image Relations



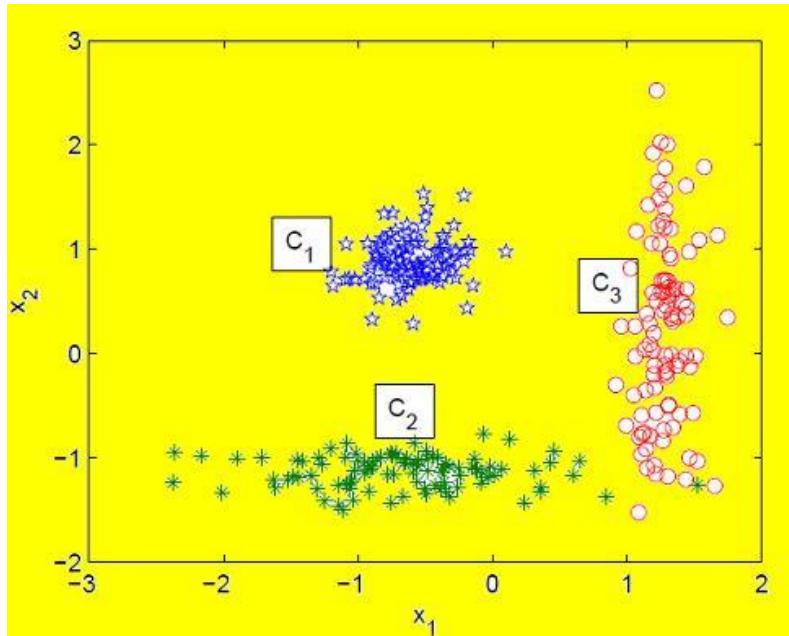
Each image has thousands or millions of pixels.

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

# Feature selection

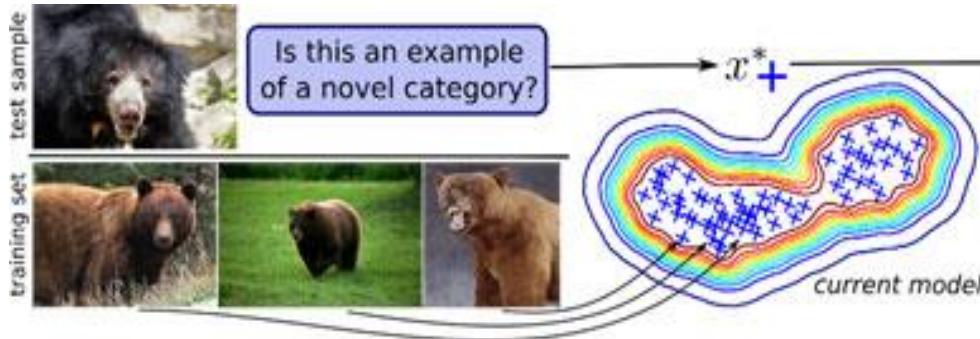


What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

# Novelty/abnormality detection



Find  
abnormal  
object



What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

# Image classification

- Learning with labels, focusing on predictive performance
  - Classifications
    - Nearest neighbor classifier
    - Naïve Bayes classifier
    - Logistic regression
    - Support vector machine
  - Combined classifiers
    - Boosting
  - Regressions
    - Ridge regression
    - Cross-validation



mite	container ship	motor scooter	leopard
black widow	lifeboat	go-kart	jaguar
cockroach	amphibian	moped	cheetah
tick	fireboat	bumper car	snow leopard
starfish	drilling platform	golfcart	Egyptian cat



convertible	agaric	dalmatian	squirrel monkey
grille	mushroom	grape	spider monkey
pickup	jelly fungus	elderberry	titi
beach wagon	gill fungus	ffordshire bullterrier	indri
fire engine	dead-man's-fingers	currant	howler monkey

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

# Face Detection



What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

# Weather Prediction



Predict

Numeric values:  
40 F  
Wind: NE at 14 km/h  
Humidity: 83%

Predict

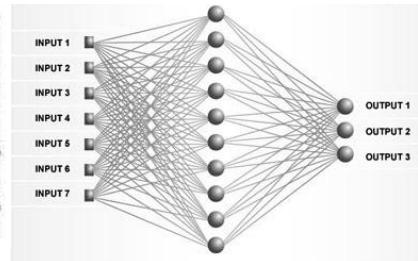
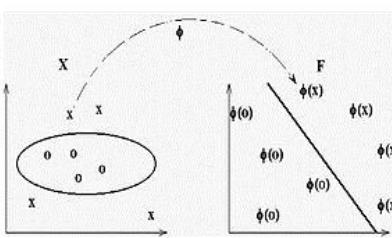
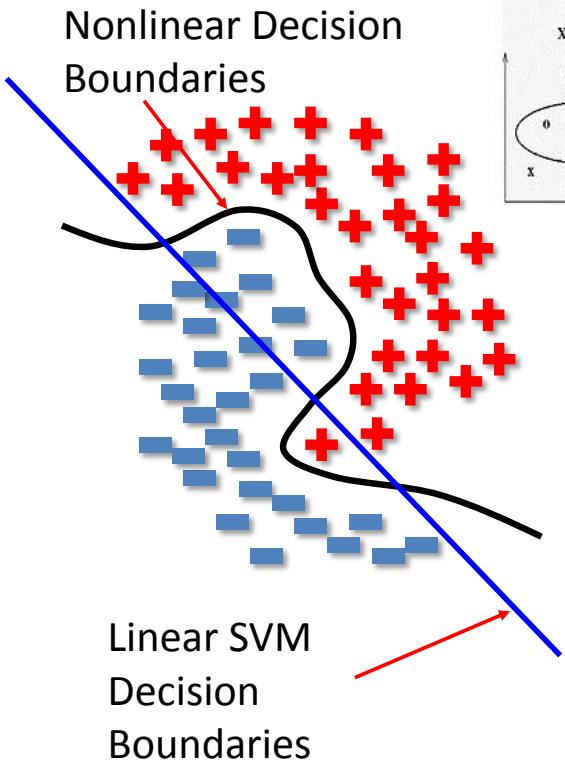
What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?



# Nonlinear classifier

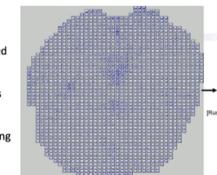


What are the desired outcomes?

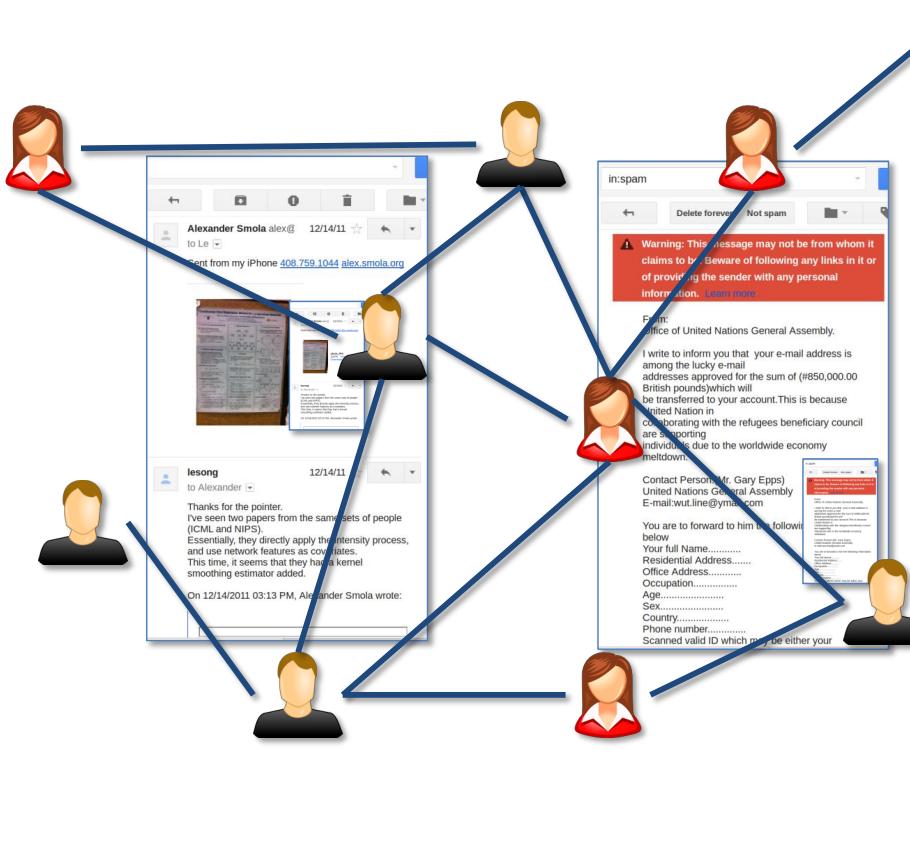
What are the inputs (data)?

What are the learning paradigms?

Understanding brain activity



# Spam Filtering



## What are the desired outcomes?

## What are the inputs (data)?

## What are the learning paradigms?

# Handwritten digit recognition/text annotation

Inter-character dependency

*The unexpected  
variabilities  
Embarrass*

What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

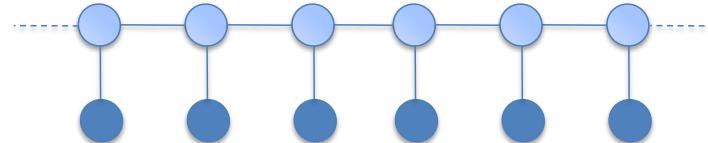
Inter-word dependency

Aoccdrnig to a sudty at Cmabrigde  
Uinervtisy, it deosn't mttaer in waht  
oredr the ltteers in a wrod are, the  
olny iprmoetnt tihng is taht the frist  
and lsat ltteer be at the rghit pclae.  
The rset can be a ttoal mses and you  
can stil raed it wouthit a porblm.  
Tihs is bcuseae the huamn mnid  
deos not raed ervey lteter by istlef,  
but the wrod as a wlohe.

# Speech recognition

Models

Hidden Markov Models



Text

“Machine Learning is the preferred method for speech recognition ...”

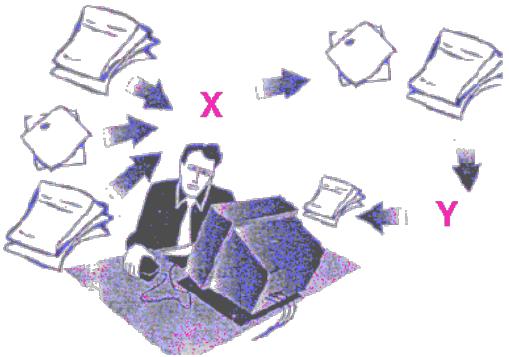


Audio signals



# Organizing documents

- Reading, digesting, and categorizing a vast text database is too much for human!



- We want:

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$260,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

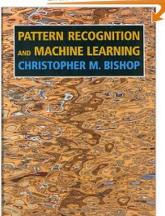
What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

# Product Recommendation

[Click to LOOK INSIDE!](#)



**Pattern Recognition and Machine Learning (Information Science and Statistics) [Hardcover]**  
Christopher M. Bishop (Author)

4.5★☆☆☆ (60 customer reviews) [Like](#) (74)

List Price: \$94.95  
Price: **\$67.98** & this item ships for FREE with Super Saver Shipping. [Details](#)  
You Save: \$26.97 (28%)  
[Special Offers Available](#)

**In Stock.**  
Ships from and sold by Amazon.com. Gift-wrap available.

Want it delivered Monday, January 9? Order it in the next 21 hours and 41 minutes, and choose One-Day Shipping at checkout. [Details](#)

42 new from \$67.98 23 used from \$69.97

FREE Two-Day Shipping for Students. [Learn more](#)

**Formats**      **Amazon Price**      **New from**      **Used from**

Hardcover	\$67.98	\$67.98	\$69.97
-----------	---------	---------	---------

**Book Trade-In**  
Sell Back Your Copy for \$56.97  
Whether you're new on Amazon for \$67.98 or somewhere else, you can sell it back through our Book Trade-in Program at the current price of \$56.97.

New Price	\$67.98
Trade-in Price	\$56.97
Price after Trade-in	\$11.01

[Share](#) [Email](#) [Facebook](#) [Twitter](#)

**Frequently Bought Together**



**Price For All Three: \$191.05**

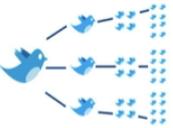
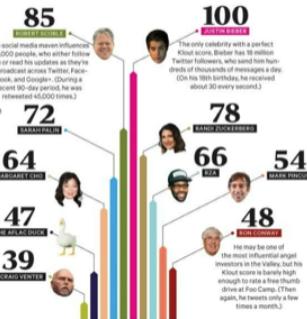
[Add all three to Cart](#) [Add all three to Wish List](#)  
[Show availability and shipping details](#)

- This item: Pattern Recognition and Machine Learning (Information Science and Statistics) by Christopher M. Bishop Hardcover \$67.98
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) by Trevor Hastie Hardcover \$63.79
- Machine Learning: An Algorithmic Perspective (Chapman & Hall/Crc Machine Learning & Pattern Recognition) by Stephen Marsland Hardcover \$59.28

**Customers Who Bought This Item Also Bought**

Machine Learning: An Algorithmic Perspective... by Stephen Marsland 4.5★☆☆☆ (16) \$59.28	Probabilistic Graphical Models: Principles and T... by Daphne Koller 4.5★☆☆☆ (6) \$72.68	Data Mining: Practical Machine Learning Tools a... by Ian H. Witten 4.5★☆☆☆ (17) \$44.07	The Elements of Statistical Learning: Data Minin... by Trevor Hastie 4.5★☆☆☆ (45) \$63.79	Pattern Classification (2nd Edition) by Richard O. Duda 4.5★☆☆☆ (32) \$91.41

What are the desired outcomes?



What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?

What are the inputs (data)?

What are the learning paradigms?

# Robot Control

- Now cars can find their own ways!



What are the desired outcomes?

What are the inputs (data)?

What are the learning paradigms?



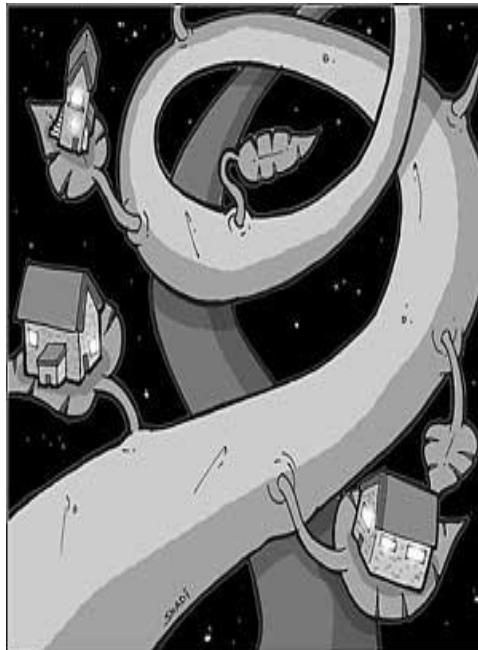
# Basics/Prerequisites

- Probabilities
  - Distributions, densities, marginalization, conditioning
- Statistics
  - Mean, variance, maximum likelihood estimation
- Linear algebra
  - Vector, matrix, multiplication, inversion, eigen-decomposition
- Algorithms and Programming
  - Matlab, Basic data structures, computational complexity
- Convex optimization
  - Basics will be covered during lecture

# Machine learning for apartment hunting

- Suppose you are to move to Atlanta
- And you want to find the **most reasonably priced** apartment satisfying your **needs**:

square-ft., # of bedroom, distance to campus ...



Living area (ft <sup>2</sup> )	# bedroom	Rent (\$)
230	1	600
506	2	1000
433	2	1100
109	1	500
...		
150	1	?
270	1.5	?

# Linear Regression Model

- Assume  $y$  is a linear function of  $x$  (features) plus noise  $\epsilon$

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n + \epsilon$$

where  $\epsilon$  is an error model as Gaussian  $N(0, \sigma^2)$

Probability

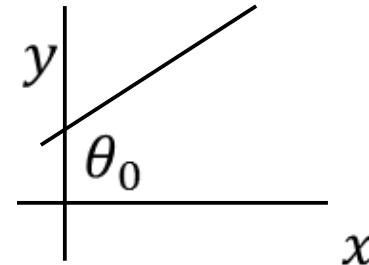
- Let  $\theta = (\theta_0, \theta_1, \dots, \theta_n)^\top$ , and augment data by one dimension

Linear algebra

$$x \leftarrow (1, x)^\top$$

Then  $y = \theta^\top x + \epsilon$

Linear algebra



# Least mean square method

- Given  $m$  data points, find  $\theta$  that minimizes the mean square error

$$\hat{\theta} = \operatorname{argmin}_{\theta} L(\theta) = \frac{1}{m} \sum_{i=1}^m (y^i - \theta^\top x^i)^2$$

Optimization

Statistics

- Set gradient to 0 and find parameter

Optimization

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{m} \sum_{i=1}^m (y^i - \theta^\top x^i) x^i = 0$$

Linear algebra

$$\Leftrightarrow -\frac{2}{m} \sum_{i=1}^m y^i x^i + \frac{2}{m} \sum_{i=1}^m x^i x^{i\top} \theta = 0$$

Statistics

Statistics

# Matrix version of the gradient

- Define  $X = (x^1, x^2, \dots, x^m), y = (y^1, y^2, \dots, y^m)^\top$ , gradient becomes

Linear algebra → 
$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{m} Xy + \frac{2}{m} XX^\top \theta$$

Linear algebra → 
$$\Rightarrow \hat{\theta} = (XX^\top)^{-1}Xy$$

Algorithms  
Programming

- Matrix inversion in  $\hat{\theta} = (XX^\top)^{-1}Xy$  **expensive** to compute

- Gradient descent

$$\hat{\theta}^{t+1} \leftarrow \hat{\theta}^t + \frac{\alpha}{m} \sum_i^m (y^i - \hat{\theta}^{t\top} x^i) x^i$$

Optimization

# Probabilistic Interpretation of LMS

- Assume  $y$  is a linear in  $x$  plus noise  $\epsilon$

$$y = \theta^\top x + \epsilon$$

- Assume  $\epsilon$  follows a Gaussian  $N(0, \sigma)$

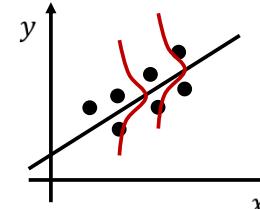
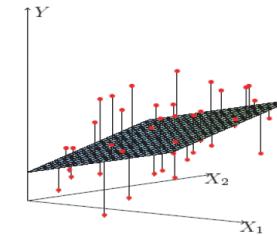
$$p(y^i | x^i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^i - \theta^\top x^i)^2}{2\sigma^2}\right)$$

- By independence assumption, likelihood is

$L(\theta)$

$$= \prod_i^m p(y^i | x^i; \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^m \exp\left(-\frac{\sum_i^m (y^i - \theta^\top x^i)^2}{2\sigma^2}\right)$$

Probability



# Probabilistic Interpretation of LMS, cont.

- Hence the log-likelihood is:

$$\log L(\theta) = m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_i^m (y^i - \theta^\top x^i)^2$$

- LMS is equivalent to MLE of  $\theta$  !

$$LMS: \frac{1}{m} \sum_i^m (y^i - \theta^\top x^i)^2$$

Statistics

- How to make it work in real data?

Algorithms  
Programming

