

# **CSE 546**

## **Machine Learning**

Instructor: Luke Zettlemoyer

TA: Lydia Chilton

Slides adapted from Pedro Domingos and Carlos Guestrin

# Logistics

- **Instructor:** Luke Zettlemoyer
  - Email: [lsz@cs](mailto:lsz@cs)
  - Office: CSE 658
  - Office hours: Tuesdays 11-12
- **TA:** Lydia Chilton
  - Email: [hmslydia@cs](mailto:hmslydia@cs)
  - Office hours: TBD
- **Web:** [www.cs.washington.edu/546](http://www.cs.washington.edu/546)

# Evaluation

- 3-4 homeworks (40% total)
- Midterm (25%)
  - Actually, 2/3 term
- Final mini-project (30%)
  - Approx. one month's work. Can incorporate your research! Or, could replicate paper, etc.
- Course participation (5%)
  - includes in class, message board, etc.

# Source Materials

Pattern Recognition and Machine Learning.  
Christopher Bishop, Springer, 2007

- Optional:
  - R. Duda, P. Hart & D. Stork, ***Pattern Classification*** (2<sup>nd</sup> ed.), Wiley (Required)
  - T. Mitchell, ***Machine Learning***, McGraw-Hill (Recommended)
  - Papers

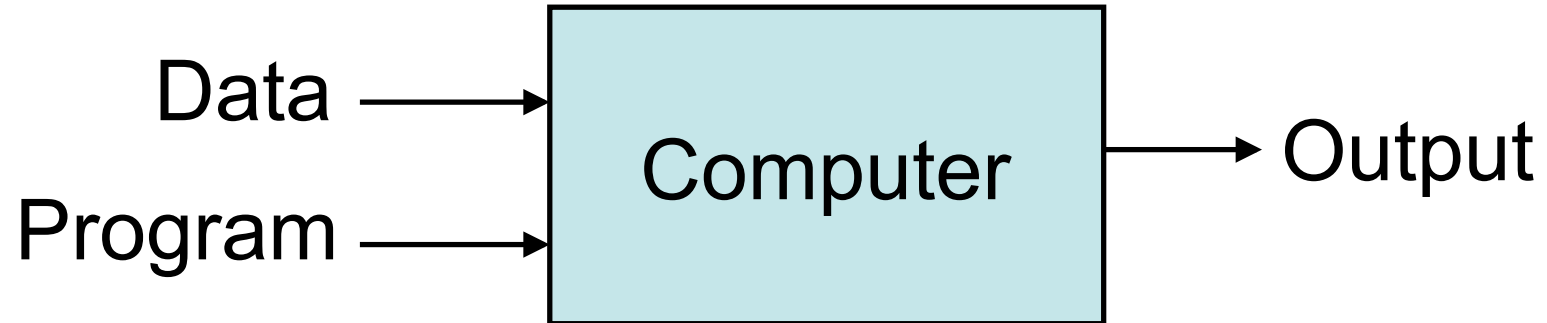
# A Few Quotes

- “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Chairman, Microsoft)
- “Machine learning is the next Internet” (Tony Tether, Director, DARPA)
- Machine learning is the hot new thing” (John Hennessy, President, Stanford)
- “Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, Dir. Research, Yahoo)
- “Machine learning is going to result in a real revolution” (Greg Papadopoulos, CTO, Sun)
- “Machine learning is today’s discontinuity” (Jerry Yang, CEO, Yahoo)

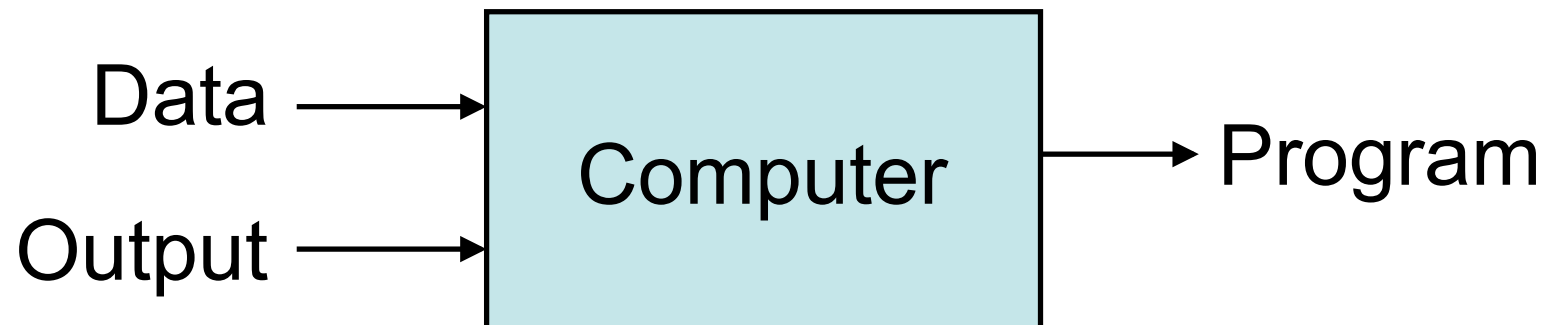
# So What Is Machine Learning?

- Automating automation
- Getting computers to program themselves
- Writing software is the bottleneck
- Let the data do the work instead!
- The future of Computer Science!!!

## Traditional Programming



## Machine Learning



# Magic?

**No, more like gardening**

- **Seeds** = Algorithms
- **Nutrients** = Data
- **Gardener** = You
- **Plants** = Programs





# What We Will Cover

- **Supervised learning**
  - Decision tree induction
  - Linear models for regression and classification
  - Instance-based learning
  - Bayesian learning
  - Neural networks
  - Support vector machines
  - Model ensembles
  - Learning theory
- **Unsupervised learning**
  - Clustering
  - Dimensionality reduction

# **What is Machine Learning ?**

## **(by examples)**

# **Classification**

**from data to discrete classes**

# Spam filtering

data

prediction

★ **Osman Khan** to Carlos [show details](#) Jan 7 (6 days ago) [Reply](#)

sounds good  
+ok

Carlos Guestrin wrote:  
Let's try to chat on Friday a little to coordinate and more on Sunday in person?

Carlos



## Welcome to New Media Installation: Art that Learns

★ **Carlos Guestrin** to 10615-announce, Osman, Michel [show details](#) 3:15 PM (8 hours ago) [Reply](#)

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).  
You can contact the instructors by emailing: [10615-instructors@cs.cmu.edu](mailto:10615-instructors@cs.cmu.edu)



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

★ **Jaquelyn Halley** to nherrlein, bcc: thehorney, bcc: anç [show details](#) 9:52 PM (1 hour ago) [Reply](#)

=== 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
- \* More Self Confidence
- \* Cleanse and Detoxify Your Body
- \* Much More Energy
- \* BetterSexLife
- \* A Natural Colon Cleanse



Spam  
VS  
Not Spam

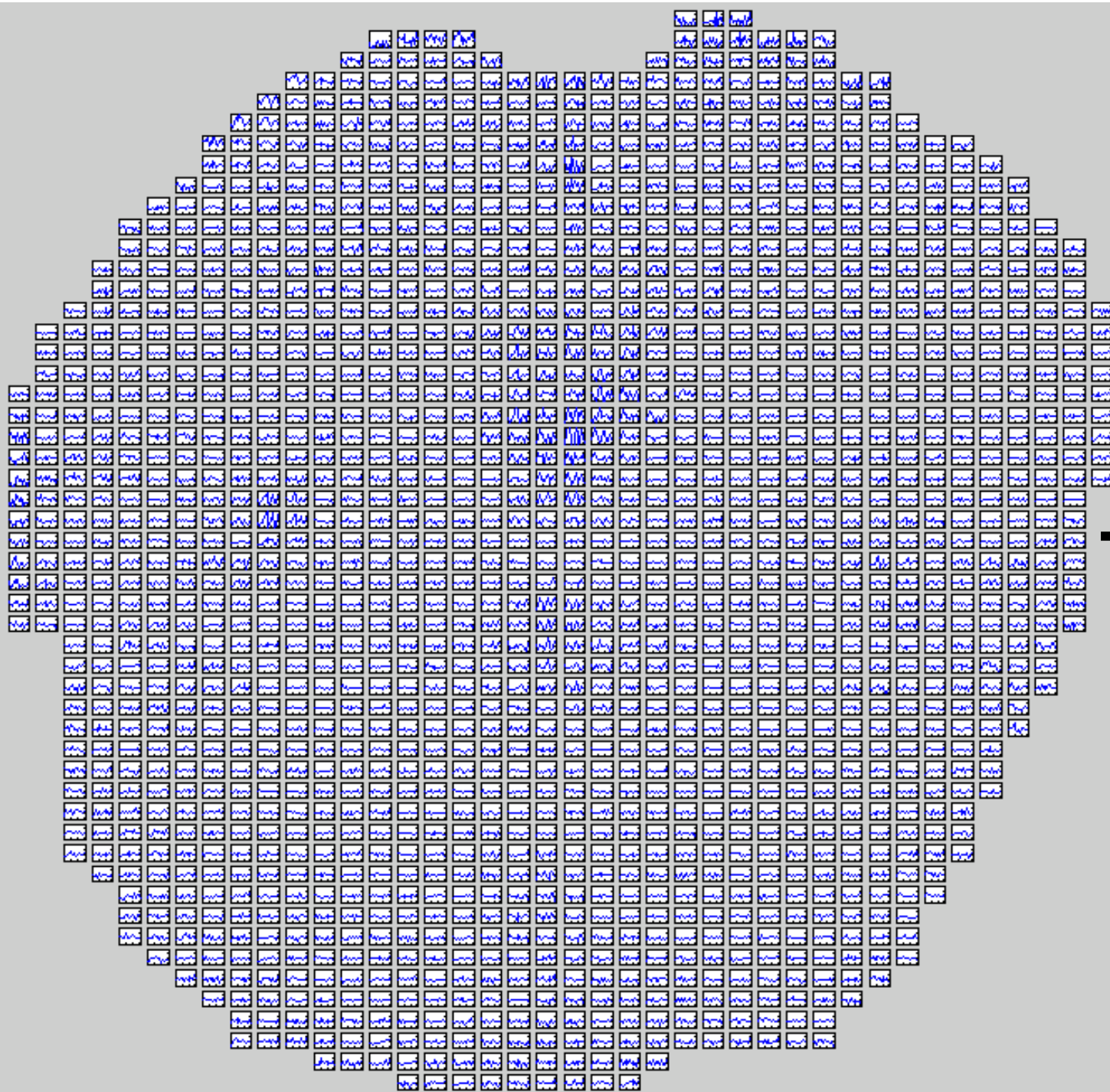
# Object detection

(Prof. H. Schneiderman)



Example training images  
for each orientation

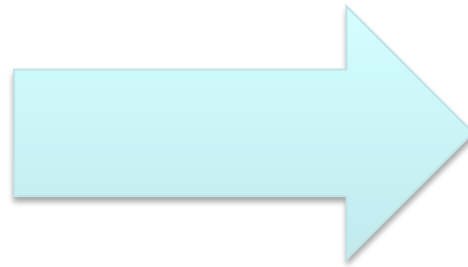




Reading  
a noun  
(vs verb)

[Rustandi et al.,  
2005]

# Weather prediction



# **Regression**

**predicting a numeric value**



# Stock market



# Weather prediction revisited

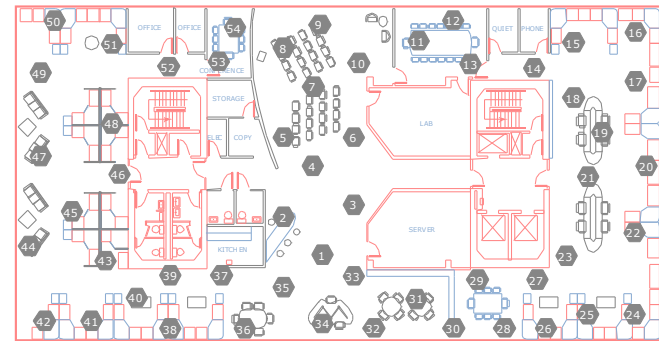


Temperature  
72° F

# Modeling sensor data



- Measure temperatures at some locations
- Predict temperatures throughout the environment





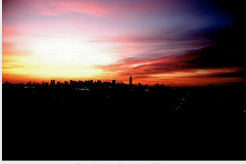

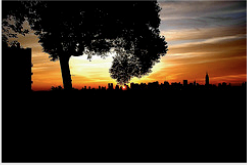
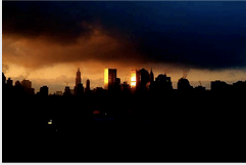
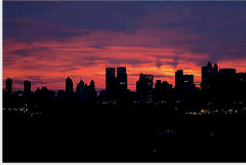








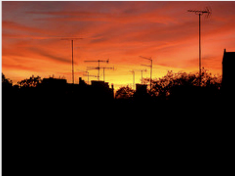


[Guestrin et al. '04]

# Similarity

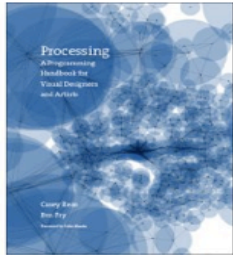
finding data

# Given image, find similar images

 <p>1. Search mode: Theme ..... 2. Find similar by Color / Texture</p>	 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>	 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>
 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>	 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>	 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>
 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>	 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>	 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>

 <p>1. Find similar by Theme ..... 2. Search mode: Color / Texture</p>	 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>	 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>
 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>	 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>	 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>
 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>	 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>	 <p>1. Find similar by Theme ..... OR ..... 2. Find similar by Color / Texture</p>

# Collaborative Filtering



[See larger image](#)

[Share your own customer images](#)

Publisher: [learn how customers can search inside this book.](#)

**Please tell the publisher:**



[I'd like to read this book on Kindle](#)

Don't have a Kindle? [Get yours here.](#)

## Processing: A Programming Handbook for Visual Designers and Artists (Hardcover)

by [Casey Reas](#) (Author), [Ben Fry](#) (Author), [John Maeda](#) (Foreword)

★★★★☆ (13 customer reviews)

Available from [these sellers.](#)

**31 new** from \$47.95 **8 used** from \$43.56

**Get Free Two-Day Shipping**

Get Free Two-Day Shipping for three months with a special extended free trial of Amazon Prime™. Add this eligible textbook to your cart to qualify. Sign up at checkout. [See details.](#)

## Related Education & Training Services in Pittsburgh [\(What's this?\)](#) | [Change location](#)

[Learn HTML Coding](#)

[www.FullSail.edu](#) - Earn Your Bachelor's Degree in Web Design and Development.

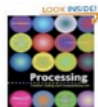
[Create Websites with HTML](#)

<http://www.unex.Berkeley.edu> - Learn HTML Online, Start Anytime! with UC Berkeley Extension

[Intensive XSLT Training](#)

[www.objectdatalabs.com/course10.asp](#) - OnSite or in NYC, LA, SFO, ORD, DC Will customize & train as few as 3

## Customers Who Bought This Item Also Bought



[Processing: Creative Coding and Computational A...](#) by Ira Greenberg

★★★★☆ (7) \$43.99



[Visualizing Data: Exploring and Explaining Data...](#) by Ben Fry

★★★★☆ (11) \$26.39



[Making Things Talk: Practical Methods for Conne...](#) by Tom Igoe

★★★★☆ (15) \$19.79



[Physical Computing: Sensing and Controlling the...](#) by Tom Igoe

★★★★☆ (20) \$19.00



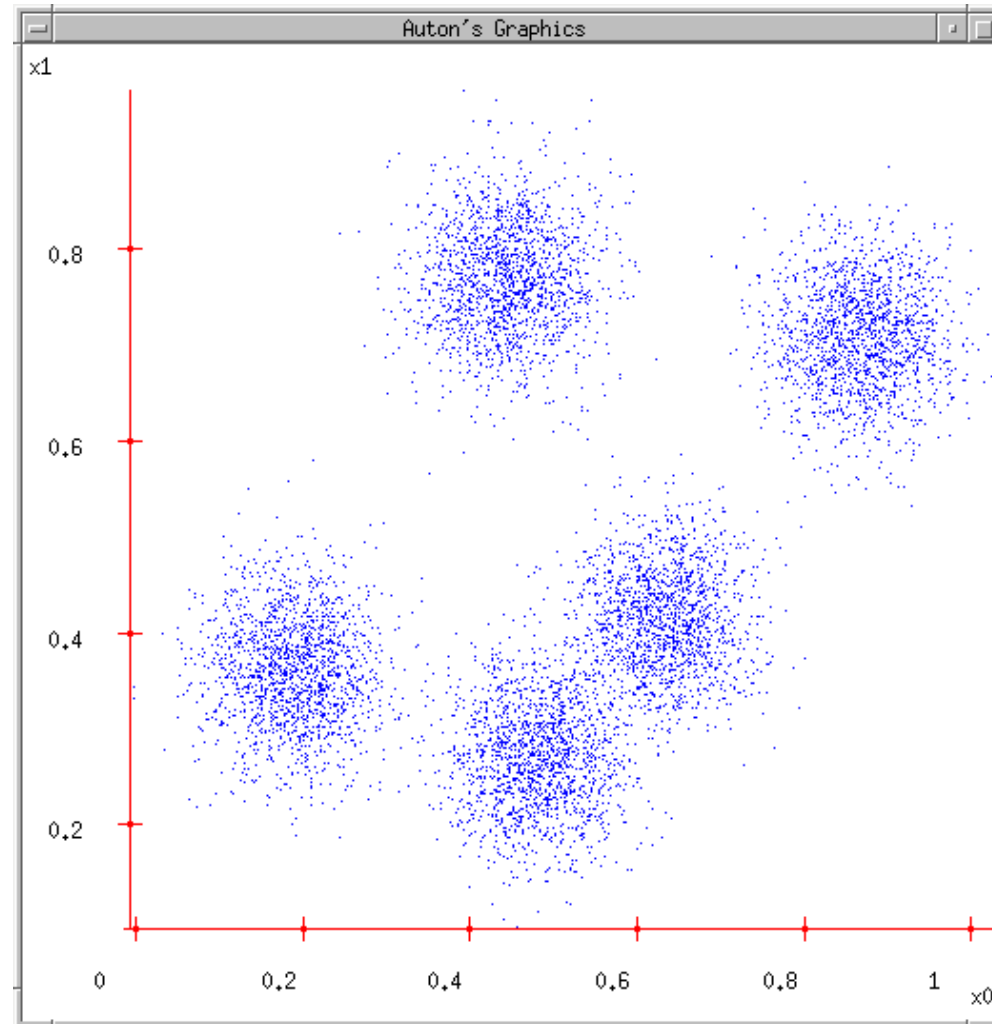
[Learning Processing: A Beginner's Guide to...](#) by Daniel Shiffman

★★★★☆ (7) \$44.95

# Clustering

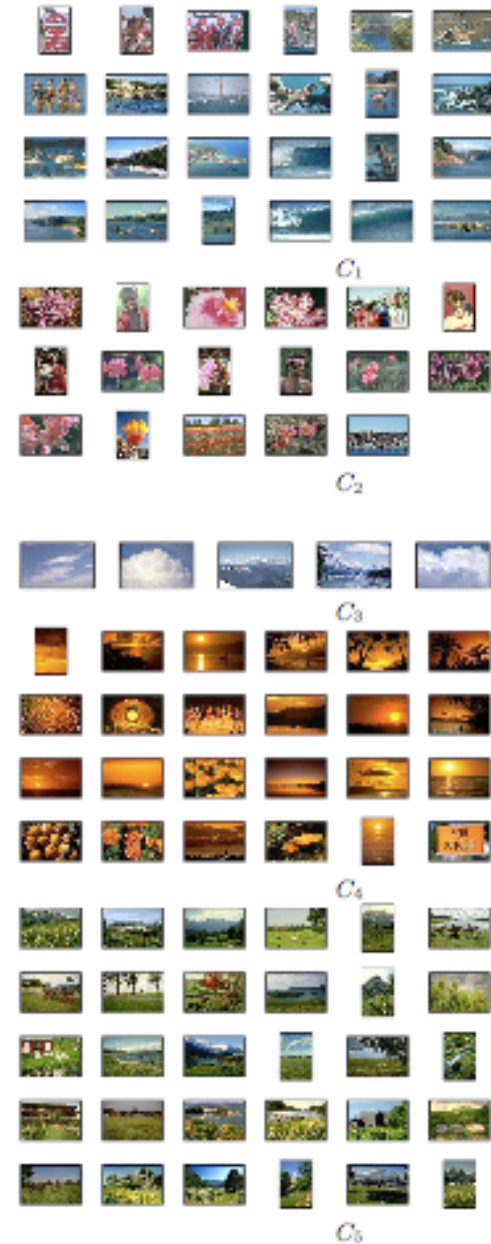
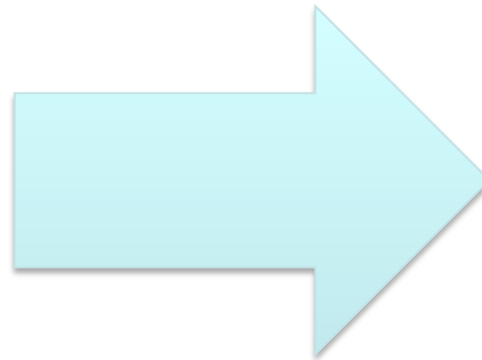
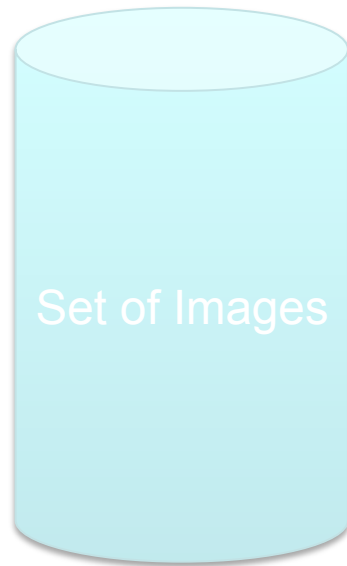
**discovering structure in data**

# Clustering Data: Group similar things





# Clustering images



[Goldberger et al.]

# Clustering web search results

Clusty [web](#) [news](#) [images](#) [wikipedia](#) [blogs](#) [jobs](#) [more »](#)

race   [advanced preferences](#)

clusters sources sites

All Results (238) remix

- Car (28)
  - Race cars (7)
  - Photos, Races Scheduled (5)
  - Game (4)
  - Track (3)
  - Nascar (2)
  - Equipment And Safety (2)
  - Other Topics (7)
- Photos (22)
- Game (14)
- Definition (13)
- Team (18)
- Human (8)**
  - Classification Of Human (2)
  - Statement, Evolved (2)
  - Other Topics (4)
- Weekend (8)
- Ethnicity And Race (7)
- Race for the Cure (8)
- Race Information (8)

[more](#) | [all clusters](#)

find in clusters:

Cluster Human contains 8 documents.

Search Results

- [Race \(classification of human beings\) - Wikipedia, the free ...](#)   

The term **race** or racial group usually refers to the concept of dividing **humans** into populations or groups on the basis of various sets of characteristics. The most widely used **human** racial categories are based on visible traits (especially skin color, cranial or facial features and hair texture), and self-identification. Conceptions of **race**, as well as specific ways of grouping **races**, vary by culture and over time, and are often controversial for scientific as well as social and political reasons. History · Modern debates · Political and ...  
[en.wikipedia.org/wiki/Race\\_\(classification\\_of\\_human\\_beings\)](#) - [cache] - Live, Ask
- [Race - Wikipedia, the free encyclopedia](#)   

General. **Racing** competitions The **Race** (yachting **race**), or La Course du millénaire, a no-rules round-the-world sailing event; **Race** (biology), classification of flora and fauna; **Race** (classification of **human** beings) **Race** and ethnicity in the United States Census, official definitions of "**race**" used by the US Census Bureau; **Race** and genetics, notion of racial classifications based on genetics. Historical definitions of **race**; **Race** (bearing), the inner and outer rings of a rolling-element bearing. **RACE** in molecular biology "Rapid ... General · Surnames · Television · Music · Literature · Video games  
[en.wikipedia.org/wiki/Race](#) - [cache] - Live, Ask
- [Publications | Human Rights Watch](#)   

The use of torture, unlawful rendition, secret prisons, unfair trials, ... Risks to Migrants, Refugees, and Asylum Seekers in Egypt and Israel ... In the run-up to the Beijing Olympics in August 2008, ...  
[www.hrw.org/background/usa/race](#) - [cache] - Ask
- [Amazon.com: Race: The Reality Of Human Differences: Vincent Sarich ...](#)   

Amazon.com: **Race: The Reality Of Human Differences: Vincent Sarich, Frank Miele: Books ...** From Publishers Weekly Sarich, a Berkeley emeritus anthropologist, and Miele, an editor ...  
[www.amazon.com/Race-Reality-Differences-Vincent-Sarich/dp/0813340861](#) - [cache] - Live
- [AAPA Statement on Biological Aspects of Race](#)   

AAPA Statement on Biological Aspects of **Race** ... Published in the American Journal of Physical Anthropology, vol. 101, pp 569-570, 1996 ... PREAMBLE As scientists who study **human** evolution and variation, ...  
[www.physanth.org/positions/race.html](#) - [cache] - Ask
- [race: Definition from Answers.com](#)   

**race** n. A local geographic or global **human** population distinguished as a more or less distinct group by genetically transmitted physical  
[www.answers.com/topic/race-1](#) - [cache] - Live
- [Dopefish.com](#)   

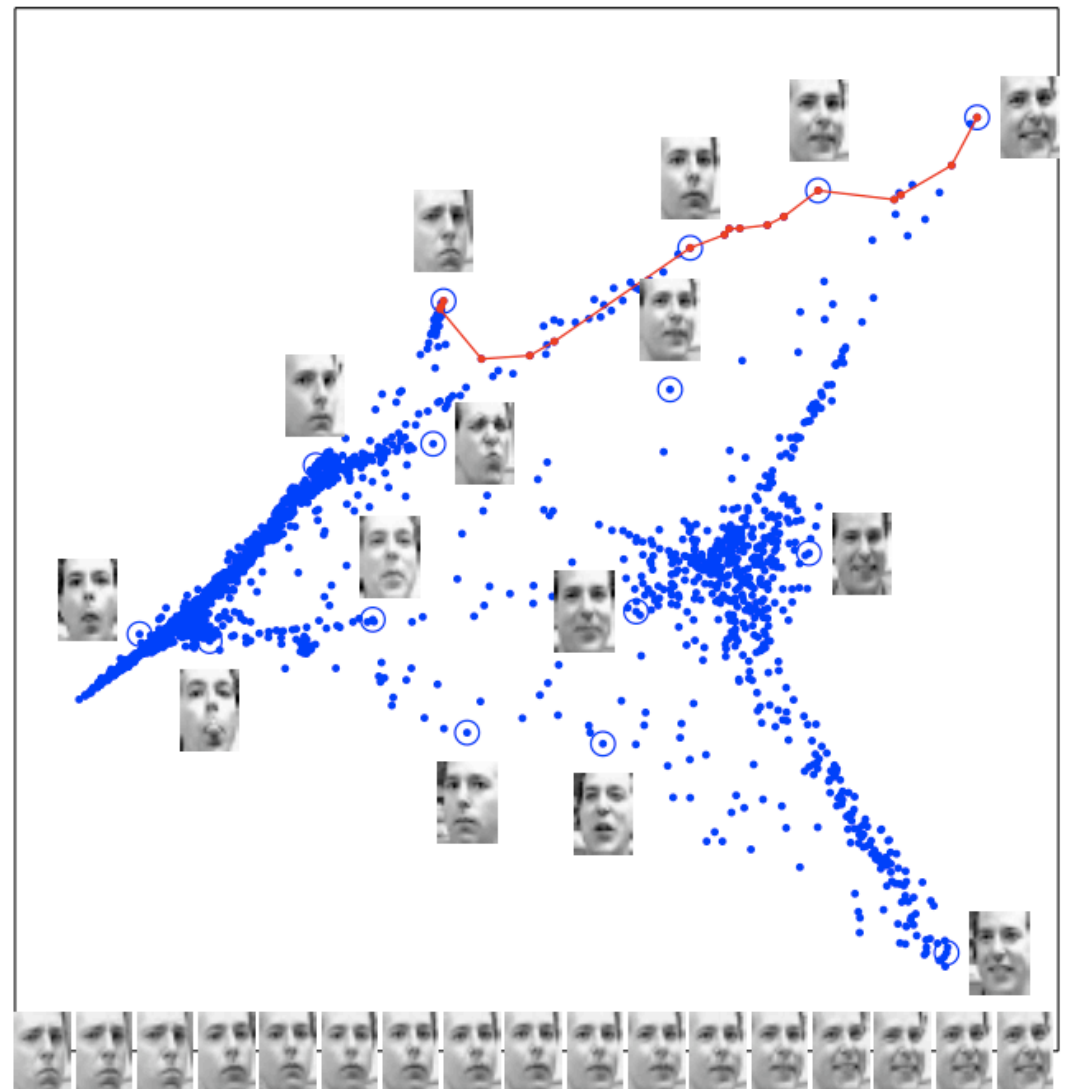
Site for newbies as well as experienced Dopefish followers, chronicling the birth of the Dopefish, its numerous appearances in several computer games, and its eventual take-over of the **human race**. Maintained by Mr. Dopefish himself, Joe Siegler of Apogee Software.  
[www.dopefish.com](#) - [cache] - Open Directory

# **Embedding**

**visualizing data**

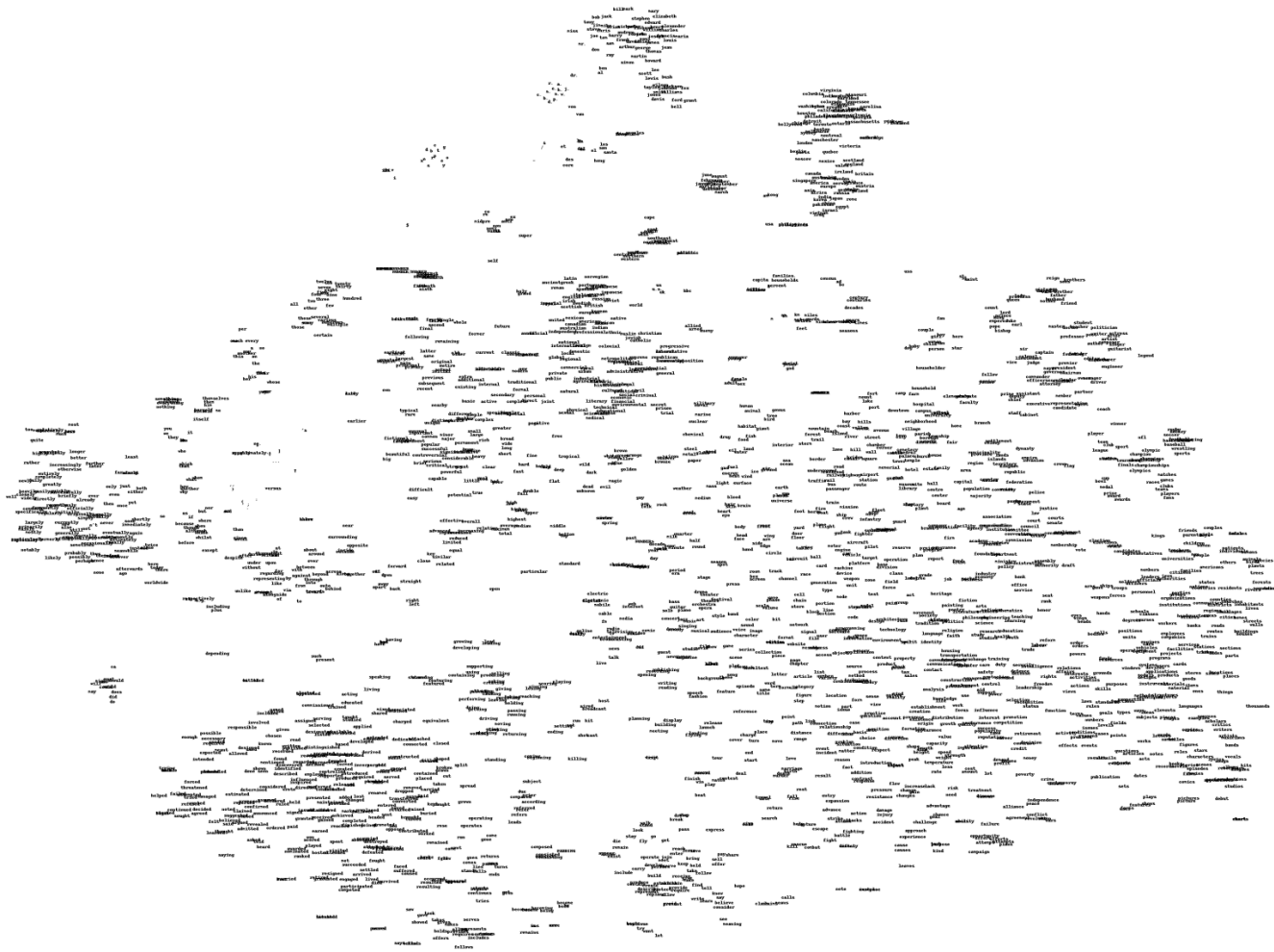
# Embedding images

- Images have thousands or millions of pixels.
- Can we give each image a coordinate, such that similar images are near each other?

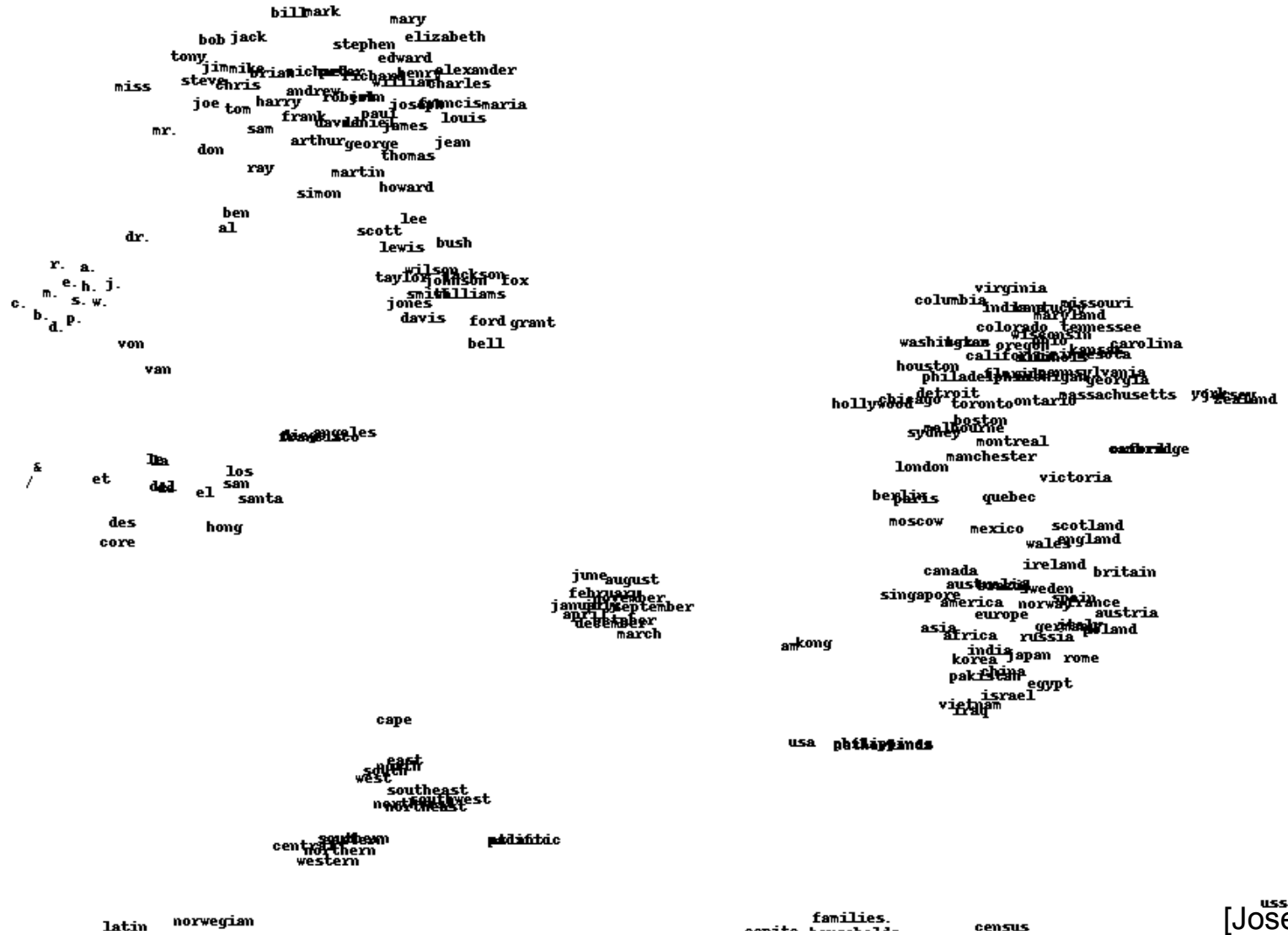


[Saul & Roweis '03]

# Embedding words



# Embedding words (zoom in)



# **Reinforcement Learning**

**training by feedback**

# Learning to act

- Reinforcement learning
- An agent
  - Makes sensor observations
  - Must select action
  - Receives rewards
    - positive for “good” states
    - negative for “bad” states

## **Robot Motor Skill Coordination with EM-based Reinforcement Learning**

**Petar Kormushev, Sylvain Calinon,  
and Darwin G. Caldwell**

**Italian Institute of Technology**



# Growth of Machine Learning

- Machine learning is preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - Computational biology
  - Sensor networks
  - ...
- This trend is accelerating
  - Improved machine learning algorithms
  - Improved data capture, networking, faster computers
  - Software too complex to write by hand
  - New sensors / IO devices
  - Demand for self-customization to user, environment

# Supervised Learning: find $f$

- **Given:** Training set  $\{(x_i, y_i) \mid i = 1 \dots n\}$
- **Find:** A good approximation to  $f : X \rightarrow Y$

**Examples:** what are  $X$  and  $Y$ ?

- **Spam Detection**
  - Map email to {Spam,Ham}
- **Digit recognition**
  - Map pixels to {0,1,2,3,4,5,6,7,8,9}
- **Stock Prediction**
  - Map new, historic prices, etc. to  $\mathfrak{R}$  (the real numbers)

# Example: Spam Filter

- **Input:** email
- **Output:** spam/ham
- **Setup:**
  - Get a large collection of example emails, each labeled “spam” or “ham”
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future emails
- **Features:** The attributes used to make the ham / spam decision
  - Words: FREE!
  - Text Patterns: \$dd, CAPS
  - Non-text: SenderInContacts
  - ...



Dear Sir.

First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. ...



TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES  
FOR ONLY \$99



Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

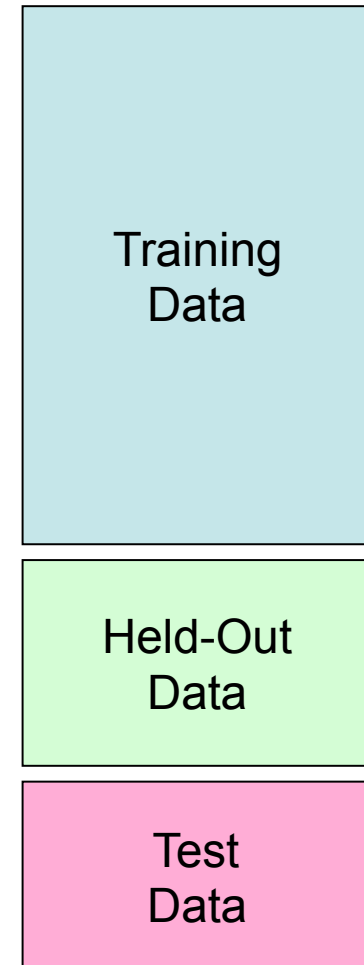
# Example: Digit Recognition

- **Input:** images / pixel grids
- **Output:** a digit 0-9
- **Setup:**
  - Get a large collection of example images, each labeled with a digit
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future digit images
- **Features:** The attributes used to make the digit decision
  - Pixels: (6,8)=ON
  - Shape Patterns: NumComponents, AspectRatio, NumLoops
  - ...



# Important Concepts

- **Data:** labeled instances, e.g. emails marked spam/ham
  - Training set
  - Held out set (sometimes call Validation set)
  - Test set
- **Features:** attribute-value pairs which characterize each  $x$
- **Experimentation cycle**
  - Select a hypothesis  $f$  to best match training set
  - (Tune hyperparameters on held-out set)
  - Compute accuracy of test set
  - Very important: never “peek” at the test set!
- **Evaluation**
  - Accuracy: fraction of instances predicted correctly
- **Overfitting and generalization**
  - Want a classifier which does well on *test* data
  - Overfitting: fitting the training data very closely, but not generalizing well
  - We’ ll investigate overfitting and generalization formally in a few lectures



# A Supervised Learning Problem

- Consider a simple, Boolean dataset:
  - $f : X \rightarrow Y$
  - $X = \{0,1\}^4$
  - $Y = \{0,1\}$
- **Question 1:** How should we pick the *hypothesis space*, the set of possible functions  $f$ ?
- **Question 2:** How do we find the best  $f$  in the hypothesis space?

Dataset:

Example	$x_1$	$x_2$	$x_3$	$x_4$	$y$
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

# Most General Hypothesis Space

Consider all possible boolean functions over four input features!

- $2^{16}$  possible hypotheses
- $2^9$  are consistent with our dataset
- How do we choose the best one?

$x_1$	$x_2$	$x_3$	$x_4$	$y$
0	0	0	0	?
0	0	0	1	?
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	?
1	0	0	0	?
1	0	0	1	1
1	0	1	0	?
1	0	1	1	?
1	1	0	0	0
1	1	0	1	?
1	1	1	0	?
1	1	1	1	?

Dataset:

Example	$x_1$	$x_2$	$x_3$	$x_4$	$y$
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

# A Restricted Hypothesis Space

Consider all conjunctive boolean functions.

- 16 possible hypotheses
- None are consistent with our dataset
- How do we choose the best one?

Rule	Counterexample
$\Rightarrow y$	1
$x_1 \Rightarrow y$	3
$x_2 \Rightarrow y$	2
$x_3 \Rightarrow y$	1
$x_4 \Rightarrow y$	7
$x_1 \wedge x_2 \Rightarrow y$	3
$x_1 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \Rightarrow y$	3
$x_2 \wedge x_4 \Rightarrow y$	3
$x_3 \wedge x_4 \Rightarrow y$	4
$x_1 \wedge x_2 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3

Dataset:

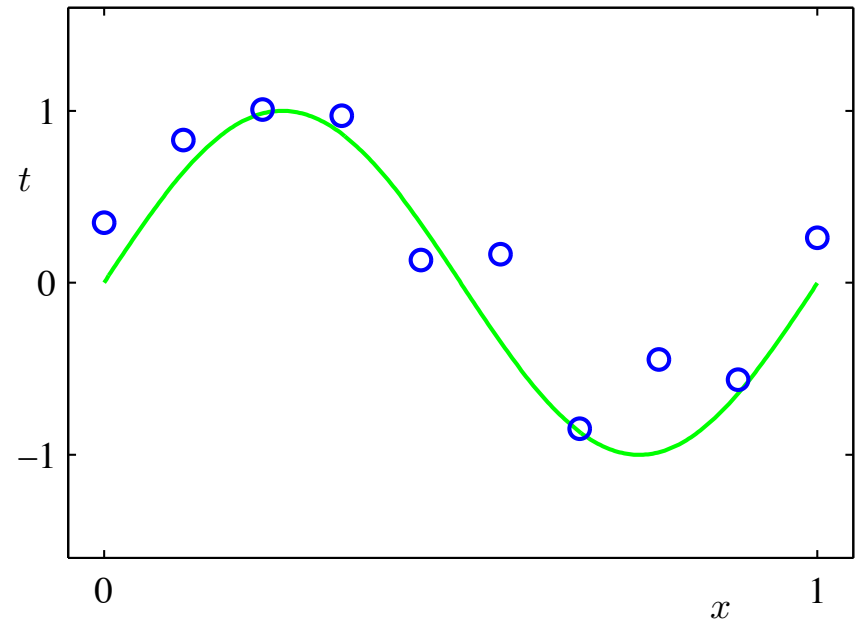
Example	$x_1$	$x_2$	$x_3$	$x_4$	$y$
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0



# Another Sup. Learning Problem

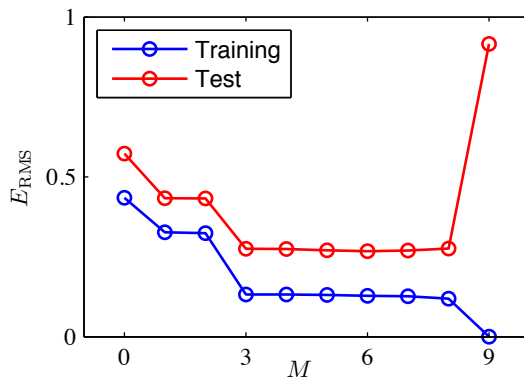
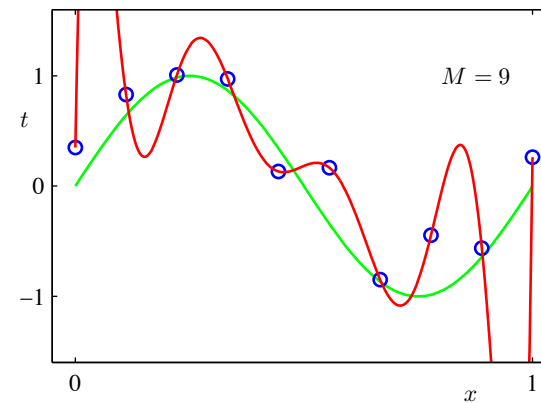
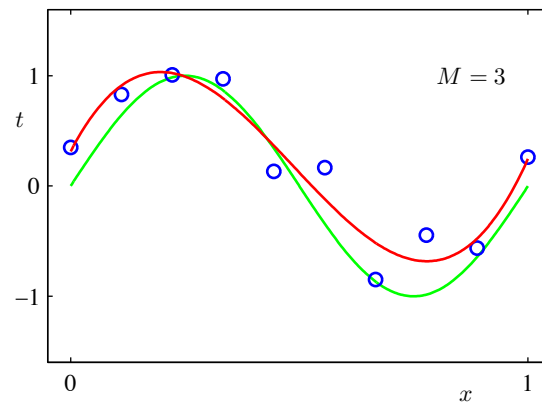
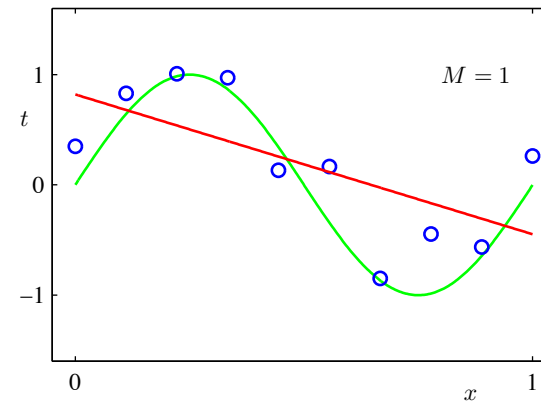
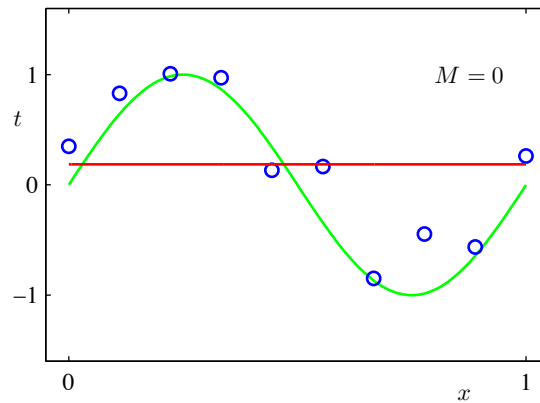
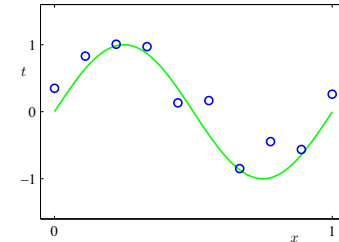
- Consider a simple, regression dataset:
  - $f : X \rightarrow Y$
  - $X = \mathfrak{R}$
  - $Y = \mathfrak{R}$
- **Question 1:** How should we pick the *hypothesis space*, the set of possible functions  $f$ ?
- **Question 2:** How do we find the best  $f$  in the hypothesis space?

Dataset: 10 points generated from a sin function, with noise



# Hypo. Space: Degree-N Polynomials

- Infinitely many hypotheses
- None / Infinitely many are consistent with our dataset
- How do we choose the best one?



# Key Issues in Machine Learning

- What are good hypothesis spaces?
- How to find the best hypothesis? (algorithms / complexity)
- How to optimize for accuracy of unseen testing data? (avoid overfitting, etc.)
- Can we have confidence in results? How much data is needed?
- How to model applications as machine learning problems? (engineering challenge)