

**Machine Learning and  
Computational Statistics  
(DS-GA-1003 and CSCI-GA.2567)**

David Sontag  
New York University

Slides adapted from Luke Zettlemoyer, Vibhav Gogate, Pedro Domingos, and Carlos Guestrin

# Logistics

- **Class webpage:**
  - <http://cs.nyu.edu/~dsontag/courses/ml14/>
  - Sign up for mailing list!
- **Required lab (instructor: Yoni Halpern)**
  - Thursdays, 8:10-9pm in WWH 109
  - Optional Q&A session from 9:10-9:40pm
- **My office hours:**
  - Tuesdays 7:15-8:15pm
  - 715 Broadway, 12<sup>th</sup> floor, Room 1204

# Evaluation

- About 7 homeworks (45%)
  - Both theory and programming
  - See collaboration policy on class webpage
- Midterm exam (25%)
- Project (25%)
- Course participation (5%)

# Problem sets

- First assignment out tonight! Due 2/6.
- See problem set policy on course website
  - First try to solve the problems on your own
  - Then, can discuss with other classmates
  - Write-up solutions on your own
  - List names of anyone you talked to
- Graders:  
Akshay Kumar, Mick Jermsurawong



# Projects

- Be creative – think of new problems that you can tackle using machine learning
  - Scope: ~40 hours/person
- Logistics:
  - 2 students per group
  - Begins in March. Project proposal due week after midterm exam
  - Will still be problem sets during this period!
- Project advisers:
  - David Rosenberg, Kurt Miller, Alex Simma

# Prerequisites

## **MS in Data Science students:**

- Intro to Data Science (DS-GA-1001)
- Statistical and Mathematical Methods (DS-GA-1002)

## **MS in Computer Science students:**

- Fundamental Algorithms (CSCI-GA.1170)
- Mathematical Techniques for Computer Science Applications (CSCI-GA.1180)

# Background needed

- **Programming**
  - Python or Matlab recommended
- **Linear algebra**
  - Matrices, vectors, systems of linear equations
  - Eigenvectors, matrix rank
  - Singular value decomposition
- **Multivariable calculus**
  - Derivatives, integration, tangent planes
  - Optimization, Lagrange multipliers
- **Probability**
  - Random variables, independence, Bayes' rule, marginalization
  - Gaussian distribution

# Source Materials

**No textbook required. Readings will come from freely available online material.**

If you really want a book for an additional reference, this is a good option:

- K. Murphy, ***Machine Learning: a Probabilistic Perspective***, MIT Press, 2012

# **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

# Face recognition

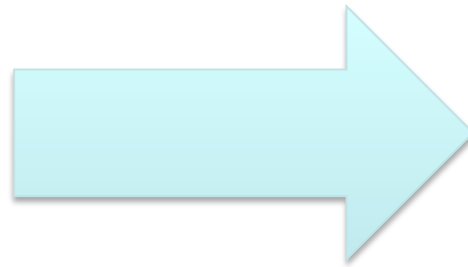


Example training images  
for each orientation





# Weather prediction



# Regression

**predicting a numeric value**

# Stock market



# Weather prediction revisited



Temperature  
72° F

# Ranking

comparing items

# Web search

The image shows a Google search interface. At the top left is the Google logo. A search bar contains the text "learning to rank". Below the search bar, a dropdown menu displays several suggestions: "learning to rank", "learning to rank for information retrieval" (with a link to "I'm Feeling Lucky »"), "learning to rank using gradient descent", and "learning to rank tutorial". To the right of the search bar is a blue search button with a magnifying glass icon. Below the search bar, the word "Search" is written in red. On the left side, there is a vertical navigation menu with links for "Web", "Images", "Maps", "Videos", "News", "Shopping", and "More". Below this menu, the location "Manhattan, NY 10012" is displayed, along with a "Change location" link and a "Show search tools" link. The main content area on the right displays search results. The first result is a Wikipedia entry titled "Learning to rank - Wikipedia, the free encyclopedia" with the URL "en.wikipedia.org/wiki/Learning\_to\_rank". The description states that "Learning to rank" or machine-learned ranking (MLR) is a type of supervised or semi-supervised machine learning problem. Below the description are links for "Applications", "Feature vectors", "Evaluation measures", and "Approaches". The second result is "Yahoo! Learning to Rank Challenge" with the URL "learningtorankchallenge.yahoo.com/". The description says the challenge is closed and mentions "close competition, innovative ideas, and fierce determination". The third result is a PDF titled "Large Scale Learning to Rank" from Tufts University, with the URL "www.eecs.tufts.edu/~dsculley/papers/large-scale-rank.pdf". The description mentions "Pairwise learning to rank methods such as RankSVM give good performance". The fourth result is "Microsoft Learning to Rank Datasets - Microsoft Research" with the URL "research.microsoft.com/en-us/projects/mslr/". The description says "We release two large scale datasets for research on learning to rank: L2R-WEB30k with more than 30000 queries and a random sampling of it L2R-WEB10K ...". The fifth result is "LETOR: A Benchmark Collection for Research on Learning to Rank ..." with the URL "research.microsoft.com/~letor/". The description says "This website is designed to facilitate research in LEarning TO Rank (LETOR). Much information about learning to rank can be found in the website, including ...".

Google

learning to rank

learning to rank

learning to rank for information retrieval [I'm Feeling Lucky »](#)

learning to rank using gradient descent

learning to rank tutorial

Search

Web

Images

Maps

Videos

News

Shopping

More

Manhattan, NY 10012

Change location

Show search tools

[Learning to rank - Wikipedia, the free encyclopedia](#)  
[en.wikipedia.org/wiki/Learning\\_to\\_rank](http://en.wikipedia.org/wiki/Learning_to_rank)  
**Learning to rank** or machine-learned ranking (MLR) is a type of supervised or semi-supervised machine learning problem in which the goal is to automatically ...  
[Applications](#) [Feature vectors](#) [Evaluation measures](#) [Approaches](#)

[Yahoo! Learning to Rank Challenge](#)  
[learningtorankchallenge.yahoo.com/](http://learningtorankchallenge.yahoo.com/)  
**Learning to Rank** Challenge is closed! Close competition, innovative ideas, and fierce determination were some of the highlights of the first ever Yahoo!

[\[PDF\] Large Scale Learning to Rank](#)  
[www.eecs.tufts.edu/~dsculley/papers/large-scale-rank.pdf](http://www.eecs.tufts.edu/~dsculley/papers/large-scale-rank.pdf)  
File Format: PDF/Adobe Acrobat - [Quick View](#)  
by D Sculley - [Cited by 24](#) - [Related articles](#)  
Pairwise **learning to rank** methods such as RankSVM give good performance, ... In this paper, we are concerned with **learning to rank** methods that can learn on ...

[Microsoft Learning to Rank Datasets - Microsoft Research](#)  
[research.microsoft.com/en-us/projects/mslr/](http://research.microsoft.com/en-us/projects/mslr/)  
We release two large scale datasets for research on **learning to rank**: L2R-WEB30k with more than 30000 queries and a random sampling of it L2R-WEB10K ...

[LETOR: A Benchmark Collection for Research on Learning to Rank ...](#)  
[research.microsoft.com/~letor/](http://research.microsoft.com/~letor/)  
This website is designed to facilitate research in **LEarning TO Rank** (LETOR). Much information about **learning to rank** can be found in the website, including ...

# Given image, find similar images

1. Search mode: Theme  
.....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

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

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

THIS PHOTO IS CURRENTLY UNAVAILABLE.  
flickr  
1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

1. Find similar by Theme  
..... OR .....  
2. Find similar by Color / Texture

# Collaborative Filtering



# Recommendation systems

The screenshot shows the Amazon.com interface with a navigation bar at the top. The main content area displays a list of four book recommendations under the heading 'Recommended for You'. Each recommendation includes a book cover, title, author, publication date, average customer review, and price information. The books are:

- Causality: Models, Reasoning and Inference** by Judea Pearl (September 14, 2009). Average Customer Review: 4.5 stars (10 reviews). List Price: \$50.00, Price: \$32.49. 61 used & new from \$28.00.
- The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century** by David Salsburg (May 1, 2002). Average Customer Review: 4.5 stars (26 reviews). List Price: \$18.99, Price: \$13.88. 81 used & new from \$9.00.
- The Eighth Day of Creation: Makers of the Revolution in Biology, 25th Anniversary Edition** by Horace Freeland Judson (November 1, 1996). Average Customer Review: 4.5 stars (10 reviews). List Price: \$56.00, Price: \$36.09. 59 used & new from \$26.95.
- The Machinery of Life** by David S. Goodsell (April 28, 2009). Average Customer Review: 4.5 stars (41 reviews). List Price: \$25.00, Price: \$17.49. 92 used & new from \$12.00.

Each item also features an 'Add to Cart' and 'Add to Wish List' button, and a 'Rate this item' section with a star rating and a 'Fix this' link.

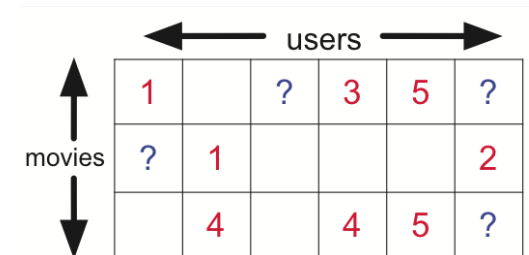
# Recommendation systems

Machine learning competition with a \$1 million prize

## Leaderboard

Display top  leaders.

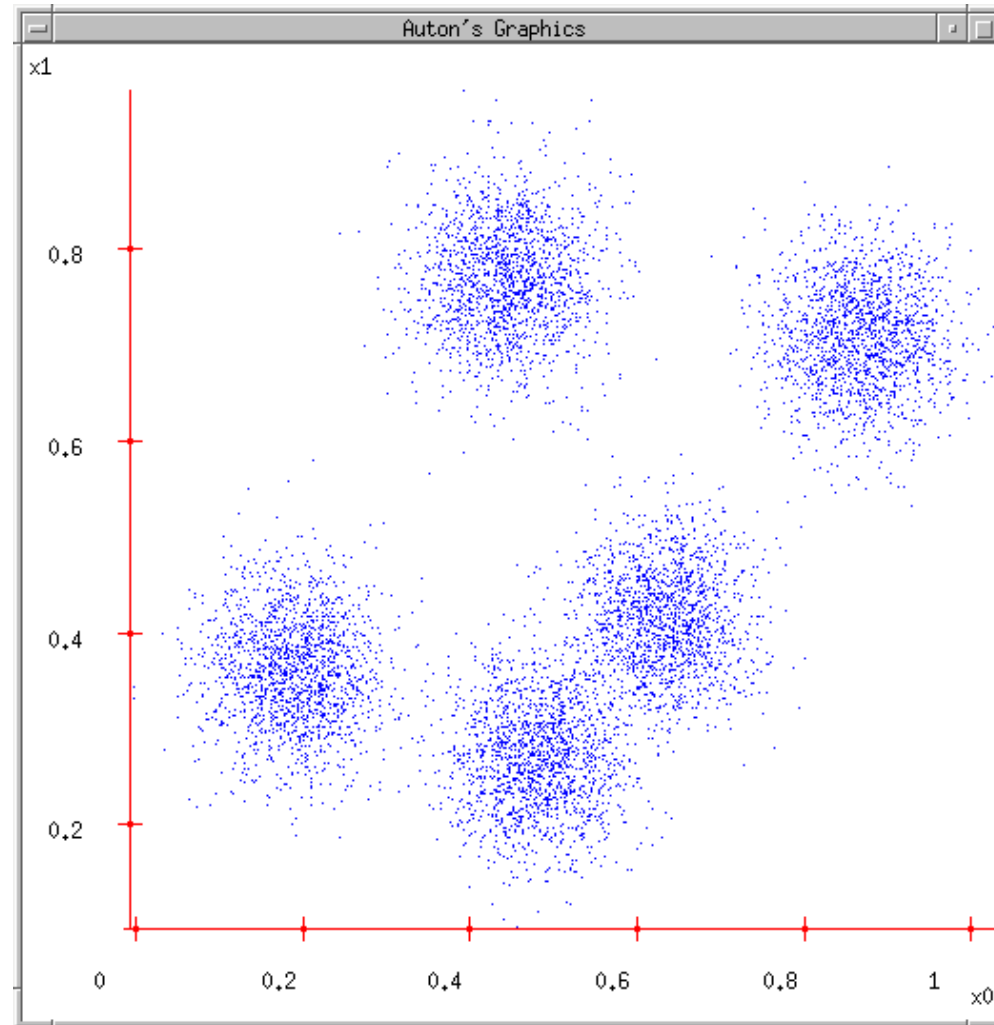
Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	<a href="#">The Ensemble</a>	0.8553	10.10	2009-07-26 18:38:22
2	<a href="#">BellKor's Pragmatic Chaos</a>	0.8554	10.09	2009-07-26 18:18:28
<b>Grand Prize - RMSE &lt;= 0.8563</b>				
3	<a href="#">Grand Prize Team</a>	0.8571	9.91	2009-07-24 13:07:49
4	<a href="#">Opera Solutions and Vandelay United</a>	0.8573	9.89	2009-07-25 20:05:52
5	<a href="#">Vandelay Industries!</a>	0.8579	9.83	2009-07-26 02:49:53
6	<a href="#">PragmaticTheory</a>	0.8582	9.80	2009-07-12 15:09:53
7	<a href="#">BellKor in BigChaos</a>	0.8590	9.71	2009-07-26 12:57:25
8	<a href="#">Dace</a>	0.8603	9.58	2009-07-24 17:18:43
9	<a href="#">Opera Solutions</a>	0.8611	9.49	2009-07-26 18:02:08
10	<a href="#">BellKor</a>	0.8612	9.48	2009-07-26 17:19:11
11	<a href="#">BioChaos</a>	0.8613	9.47	2009-06-23 23:06:52
12	<a href="#">Feeds2</a>	0.8613	9.47	2009-07-24 20:06:46
<b>Progress Prize 2008 - RMSE = 0.8616 - Winning Team: BellKor in BigChaos</b>				
13	<a href="#">xianqliang</a>	0.8633	9.26	2009-07-21 02:04:40
14	<a href="#">Gravity</a>	0.8634	9.25	2009-07-26 15:58:34
15	<a href="#">Ces</a>	0.8642	9.17	2009-07-25 17:42:38
16	<a href="#">Invisible Ideas</a>	0.8644	9.14	2009-07-20 03:26:12
17	<a href="#">Just a guy in a garage</a>	0.8650	9.08	2009-07-22 14:10:42
18	<a href="#">Craig Carmichael</a>	0.8656	9.02	2009-07-25 16:00:54
19	<a href="#">J Dennis Su</a>	0.8658	9.00	2009-03-11 09:41:54
20	<a href="#">acmehill</a>	0.8659	8.99	2009-04-16 06:29:35
<b>Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell</b>				
<b>Cinematch score on quiz subset - RMSE = 0.9514</b>				



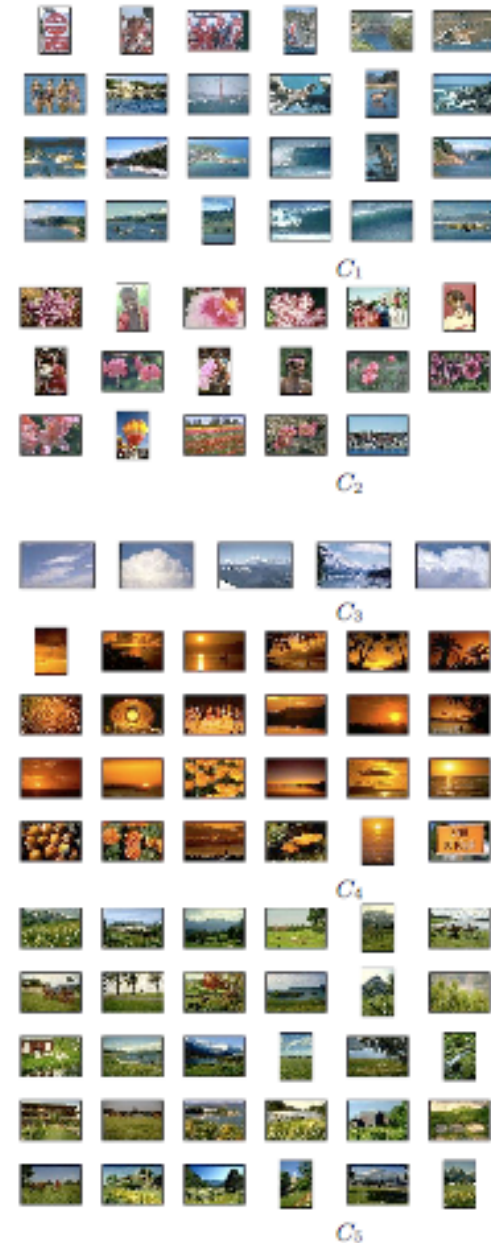
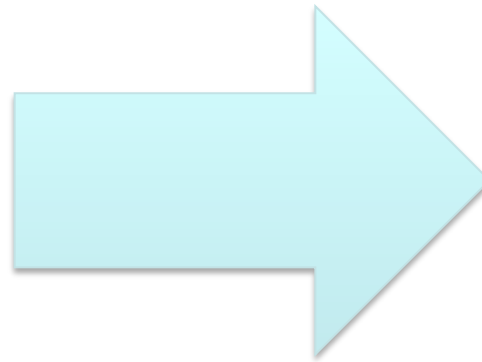
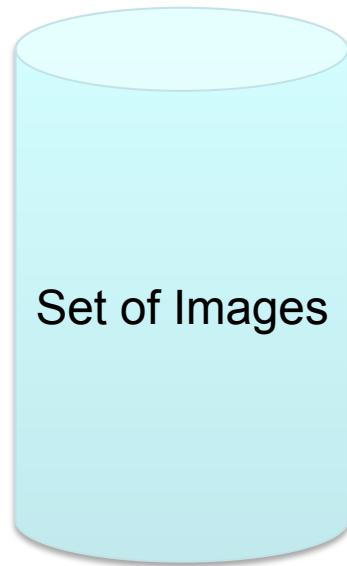
# Clustering

**discovering structure in data**

# Clustering Data: Group similar things



# Clustering images



[Goldberger et al.]

# Clustering web search results

web news images wikipedia blogs jobs more »

Clusty

race Search 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:  Find




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\\_\(classification\\_of\\_human\\_beings\)](http://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](http://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](http://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](http://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](http://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](http://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](http://www.dopefish.com) - [cache] - Open Directory

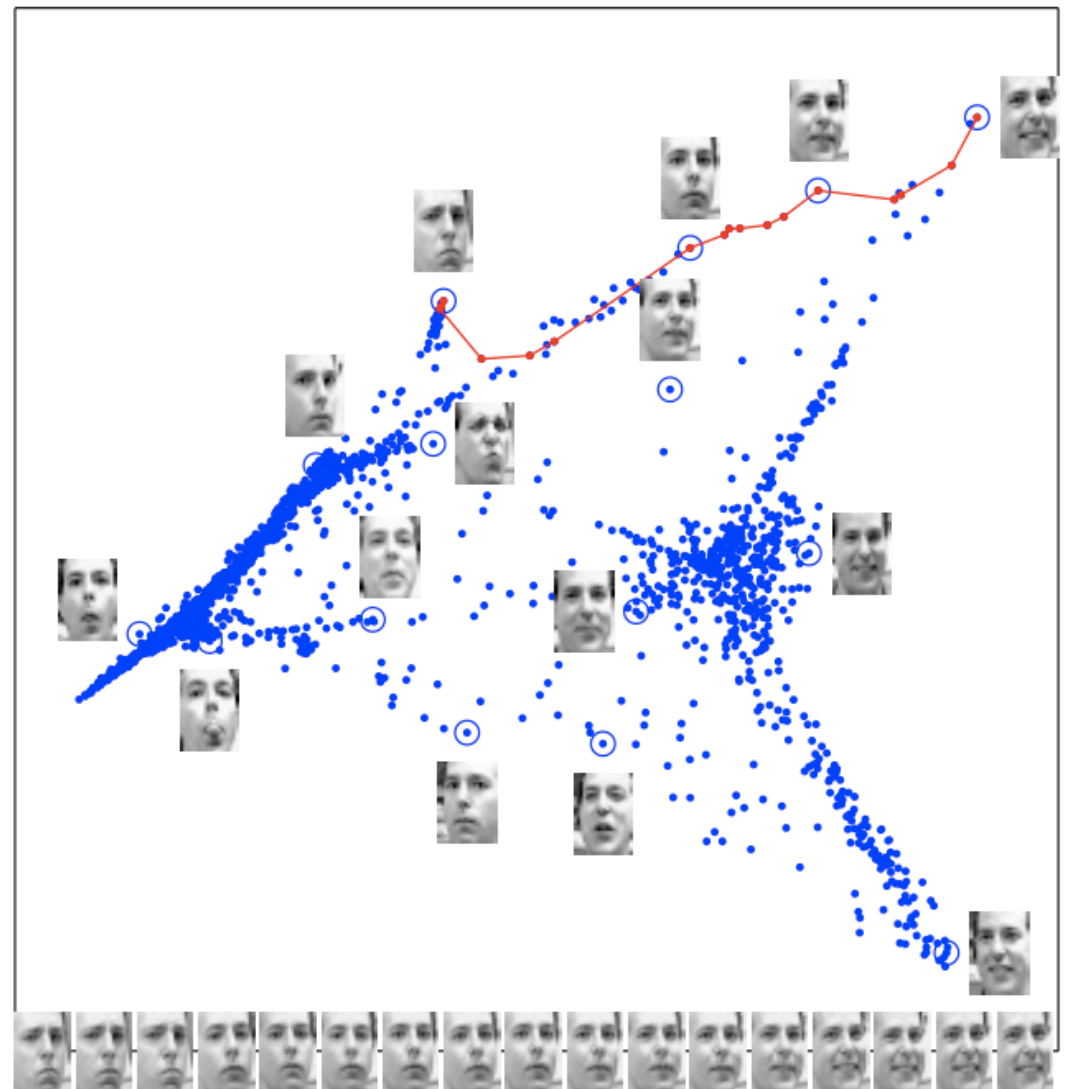
# **Embedding**

**visualizing data**



# Embedding images

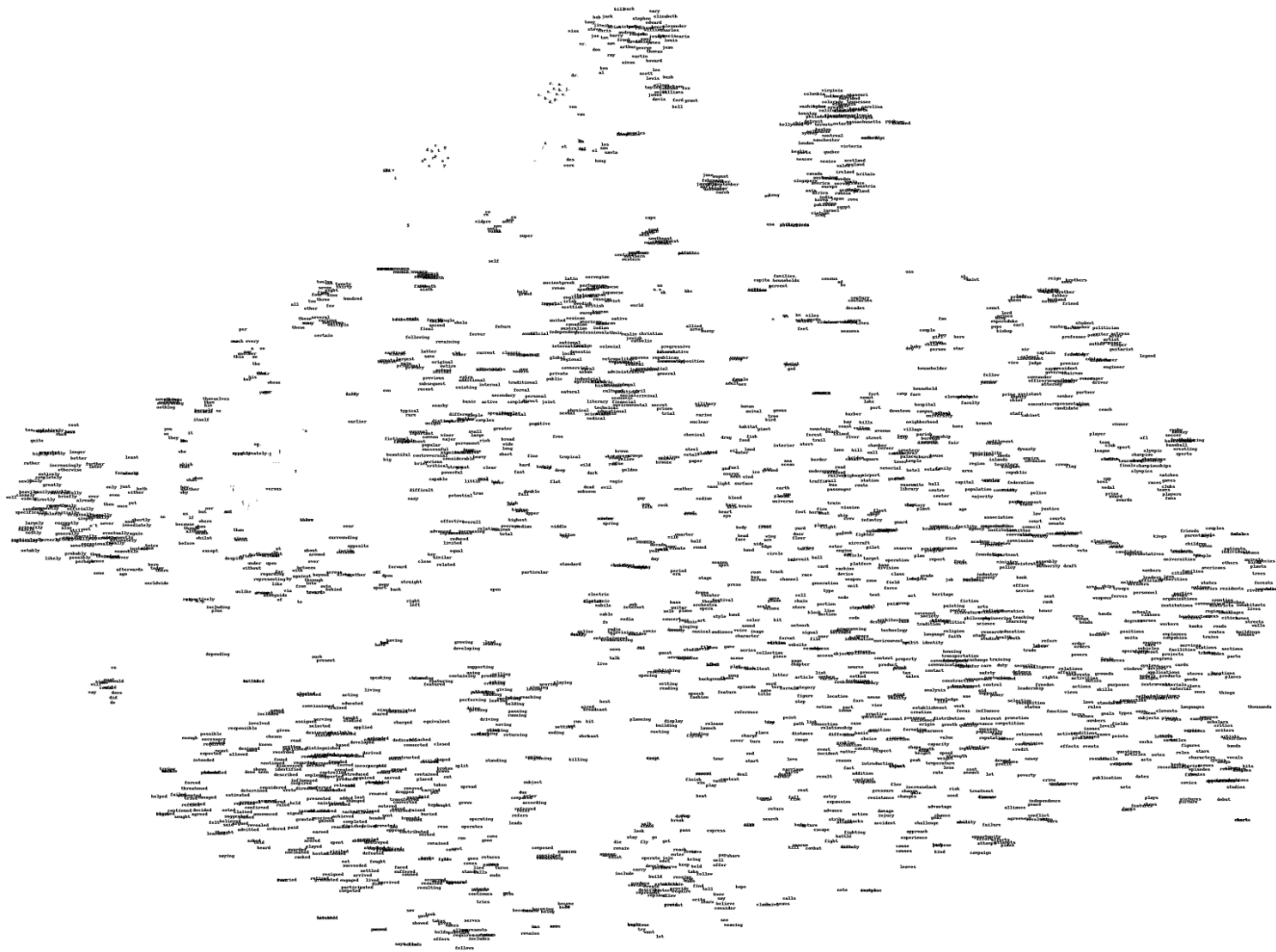
- Images have thousands or millions of pixels.
- Can we give each image a coordinate, such that similar images are near each other?



[Saul & Roweis '03]



# Embedding words



[Joseph Turian]

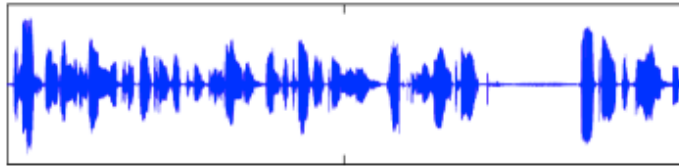
# Embedding words (zoom in)



# **Structured prediction**

**from data to discrete classes**

# Speech recognition



# Natural language processing

I need to hide a body  
noun, verb, preposition, ...



# 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
  - Big data
  - Improved machine learning algorithms
  - Faster computers
  - Good open-source software

# Course roadmap

- **First half of course: discriminative methods**
  - SVMs, kernel methods
  - Learning theory
  - Decision trees, boosting, deep learning
- **Second half of course: generative methods**
  - Graphical models, Gibbs sampling
  - Unsupervised learning, EM algorithm
  - Dimensionality reduction
  - LDA, topic models

# 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, Not Spam}
- **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)



# A Supervised Learning Problem

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

- Our goal is to find a function  $f : X \rightarrow Y$ 
  - $X = \{0,1\}^4$
  - $Y = \{0,1\}$
- **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?

# 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?

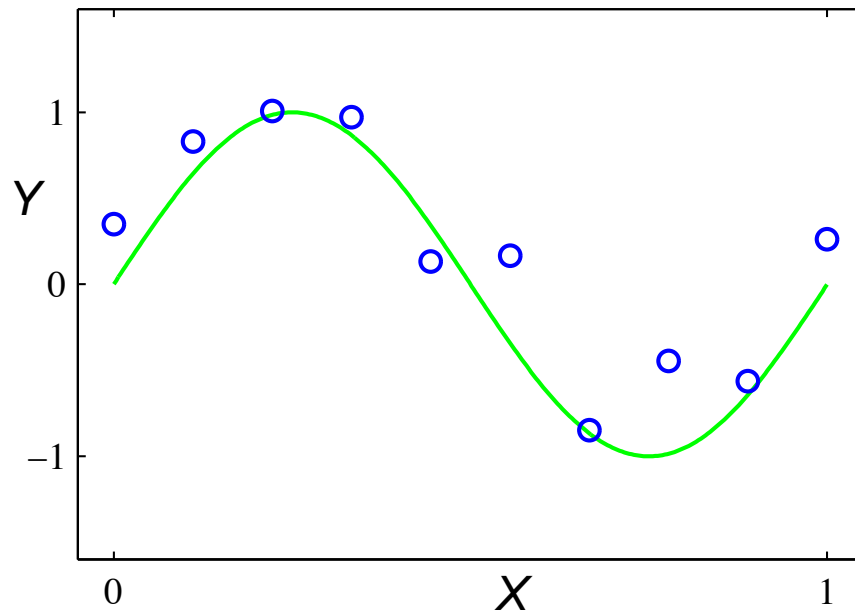
Rule	Counterexample
$\Rightarrow y$	1
$x_1 \Rightarrow y$	3
$x_2 \Rightarrow y$	2
$x_3 \Rightarrow y$	1
$x_4 \Rightarrow y$	7
$x_1 \wedge x_2 \Rightarrow y$	3
$x_1 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \Rightarrow y$	3
$x_2 \wedge x_4 \Rightarrow y$	3
$x_3 \wedge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

# Second example: Regression

Dataset: 10 (X,Y) points generated from a sin function, with noise



- Regression:

- $f : X \rightarrow Y$

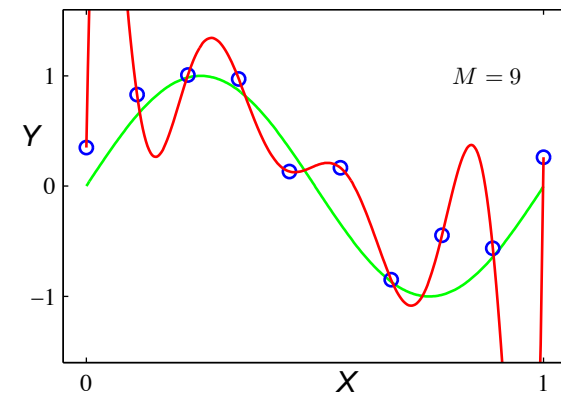
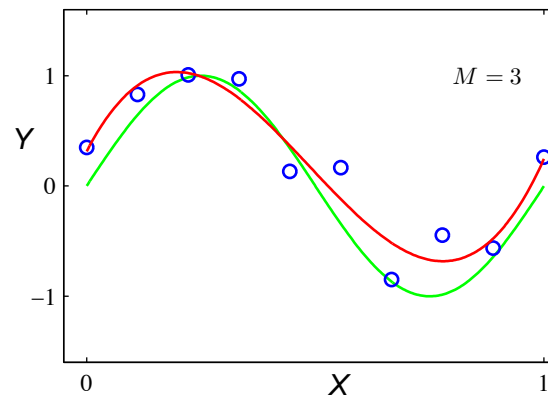
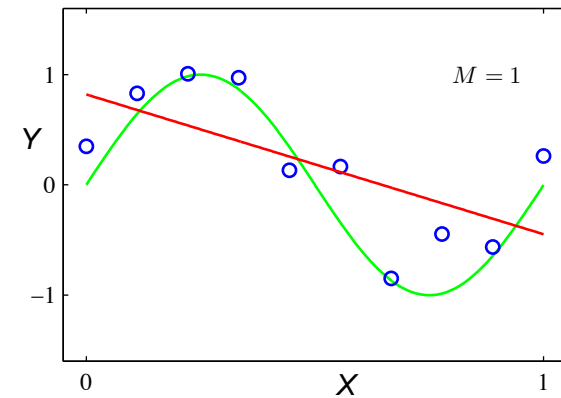
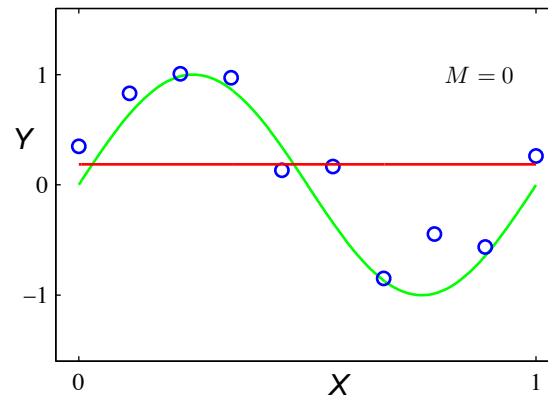
- $X = \mathfrak{R}$

- $Y = \mathfrak{R}$

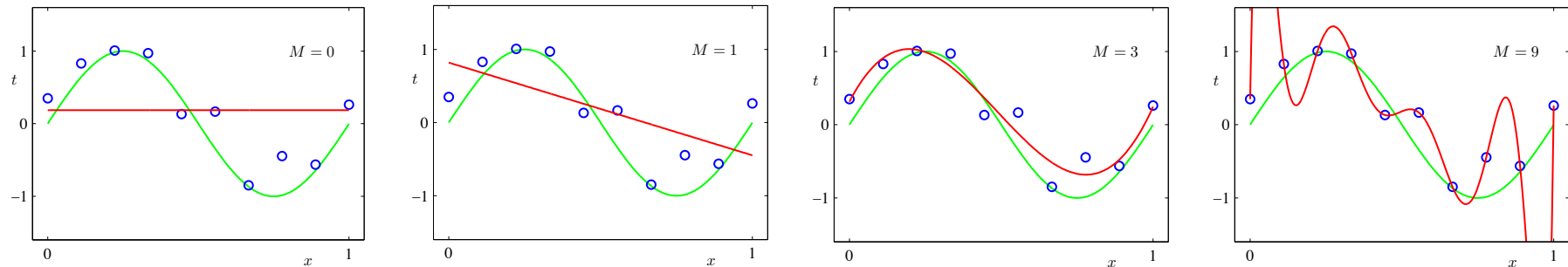
# Degree-M Polynomials

How about letting  $f$  be a degree  $M$  polynomial?

•Which one is **best**?



# Hypo. Space: Degree-N Polynomials



We measure error using a *loss function*  $L(y, \hat{y})$

For regression, a common choice is squared loss:

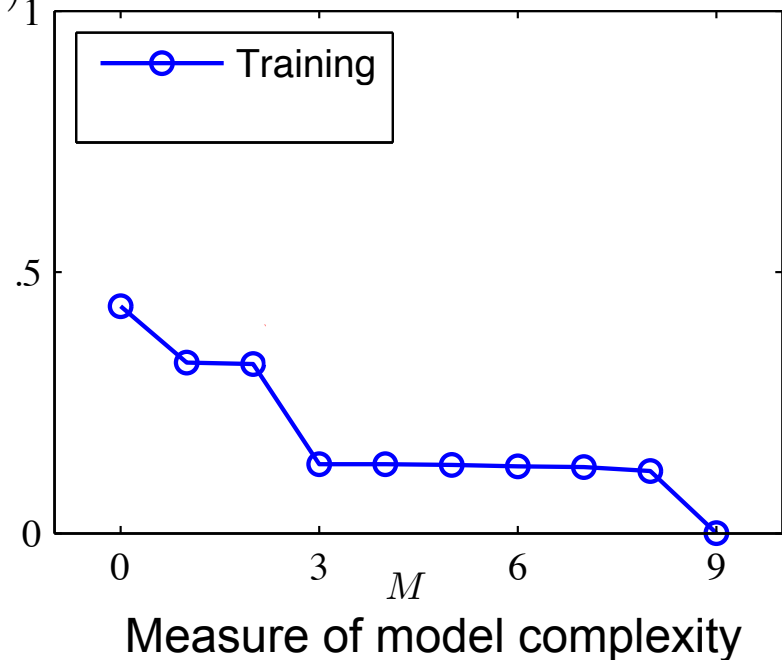
$$L(y_i, f(x_i)) = (y_i - f(x_i))^2$$

Squared error

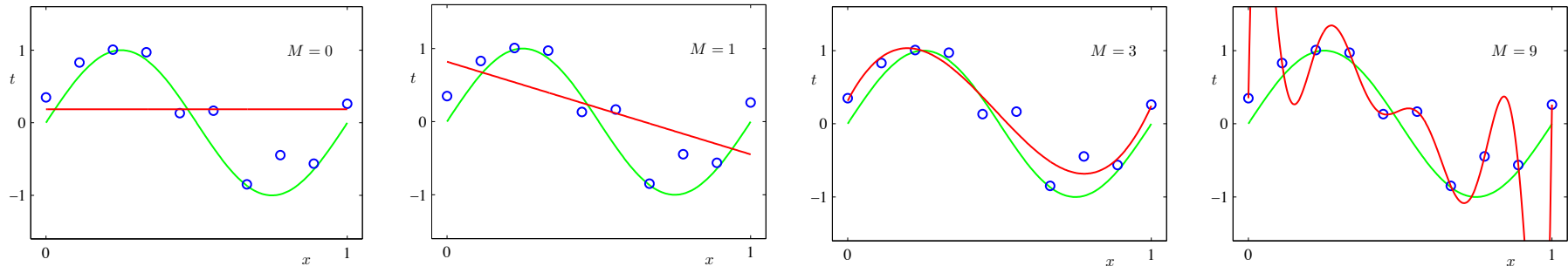
The *empirical loss* of the function  $f$  applied to the training data is then:

$$\frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) = \frac{1}{N} \sum_{i=1}^N (y_i - f(x_i))^2$$

Learning curve



# Hypo. Space: Degree-N Polynomials



We measure error using a *loss function*  $L(y, \hat{y})$

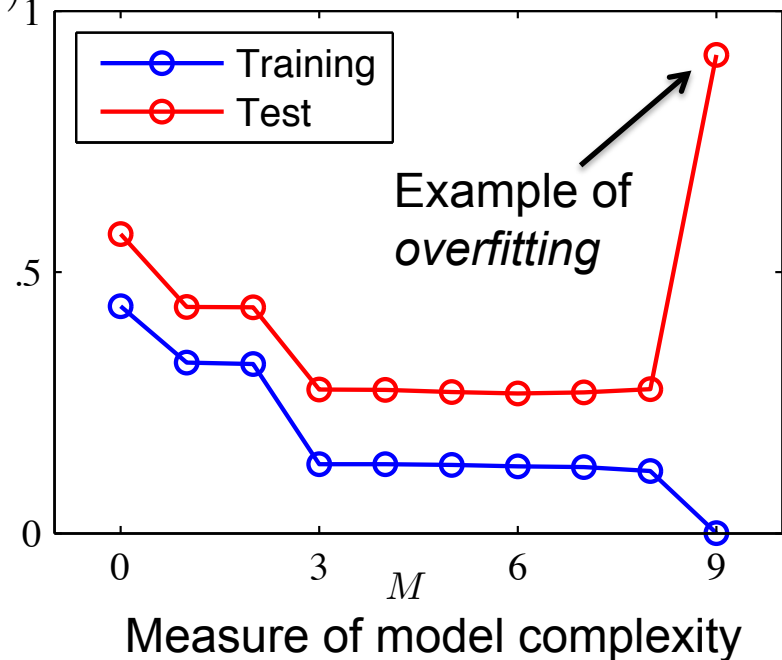
For regression, a common choice is squared loss:

$$L(y_i, f(x_i)) = (y_i - f(x_i))^2 \quad \text{Squared error}$$

The *empirical loss* of the function  $f$  applied to the training data is then:

$$\frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i)) = \frac{1}{N} \sum_{i=1}^N (y_i - f(x_i))^2$$

Learning curve



# Occam's Razor Principle

- William of **Occam**: Monk living in the 14<sup>th</sup> century
- Principle of parsimony:

“One should not increase, beyond what is necessary, the number of entities required to explain anything”

- When **many** solutions are available for a given problem, we should select the **simplest** one
- But what do we mean by **simple**?
- We will use **prior knowledge** of the problem to solve to define what is a simple solution

*Example of a prior: smoothness*

[Samy Bengio]



# Key Issues in Machine Learning

- How do we choose a hypothesis space?
  - Often we use **prior knowledge** to guide this choice
- How can we gauge the accuracy of a hypothesis on unseen data?
  - **Occam's razor**: use the *simplest* hypothesis consistent with data! This will help us avoid overfitting.
  - **Learning theory** will help us quantify our ability to **generalize** as a function of the amount of training data and the hypothesis space
- How do we find the best hypothesis?
  - This is an **algorithmic** question, the main topic of computer science
- How to model applications as machine learning problems?  
(engineering challenge)

# Binary classification

- **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. ...



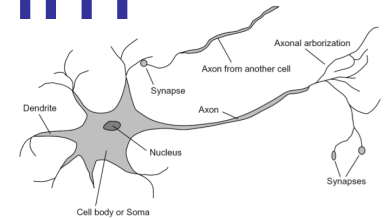
TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

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.

# The perceptron algorithm



- 1957: Perceptron algorithm invented by Rosenblatt

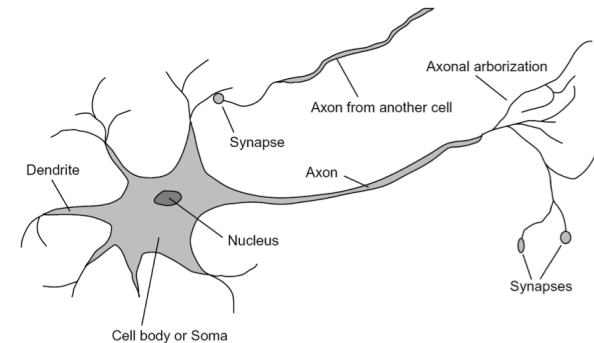
Wikipedia: “A handsome bachelor, he drove a classic MGA sports... for several years taught an interdisciplinary undergraduate honors course entitled "Theory of Brain Mechanisms" that drew students equally from Cornell's Engineering and Liberal Arts colleges...this course was a melange of ideas .. experimental brain surgery on epileptic patients while conscious, experiments on .. the visual cortex of cats, ... analog and digital electronic circuits that modeled various details of neuronal behavior (i.e. the perceptron itself, as a machine).”

- Built on work of Hebb (1949); also developed by Widrow-Hoff (1960)
- 1960: Perceptron Mark 1 Computer – hardware implementation
- 1969: Minsky & Papert book shows perceptrons limited to *linearly separable* data, and Rosenblatt dies in boating accident
- 1970's: Learning methods for two-layer neural networks

[William Cohen]

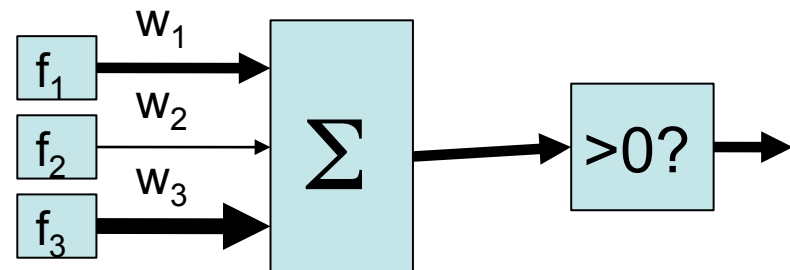
# Linear Classifiers

- Inputs are **feature values**
- Each feature has a **weight**
- Sum is the **activation**



$$\text{activation}_w(x) = \sum_i w_i \cdot f_i(x) = w \cdot f(x)$$

- If the activation is:
  - Positive, output *class 1*
  - Negative, output *class 2*



# Example: Spam

- Imagine 3 features (spam is “positive” class):
  - free (number of occurrences of “free”)
  - money (occurrences of “money”)
  - BIAS (intercept, always has value 1)

$$w \cdot f(x)$$

$$\sum_i w_i \cdot f_i(x)$$

$x$	$f(x)$	$w$	
“free money”	BIAS : 1 free : 1 money : 1 ...	BIAS : -3 free : 4 money : 2 ...	$(1)(-3) +$ $(1)(4) +$ $(1)(2) +$ ... = 3

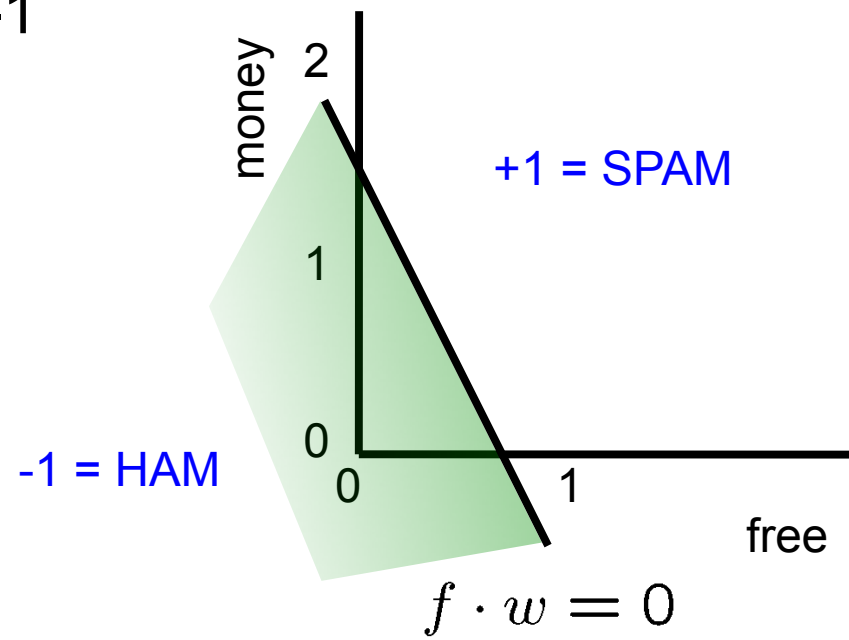
$w \cdot f(x) > 0 \rightarrow$  SPAM!!!

# Binary Decision Rule

- In the space of feature vectors
  - Examples are points
  - Weight vector and bias define a hyperplane
  - One side corresponds to  $Y=+1$
  - Other corresponds to  $Y=-1$

$w$

BIAS	:	-3
free	:	4
money	:	2
...	:	



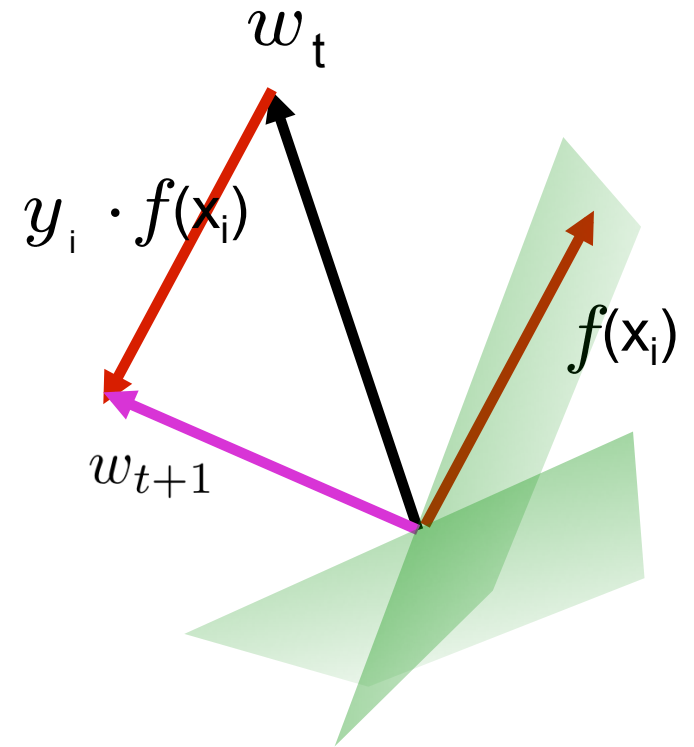
# The perceptron algorithm

- Start with weight vector =  $\vec{0}$
- For each training instance  $(x_i, y_i)$ :
  - Classify with current weights

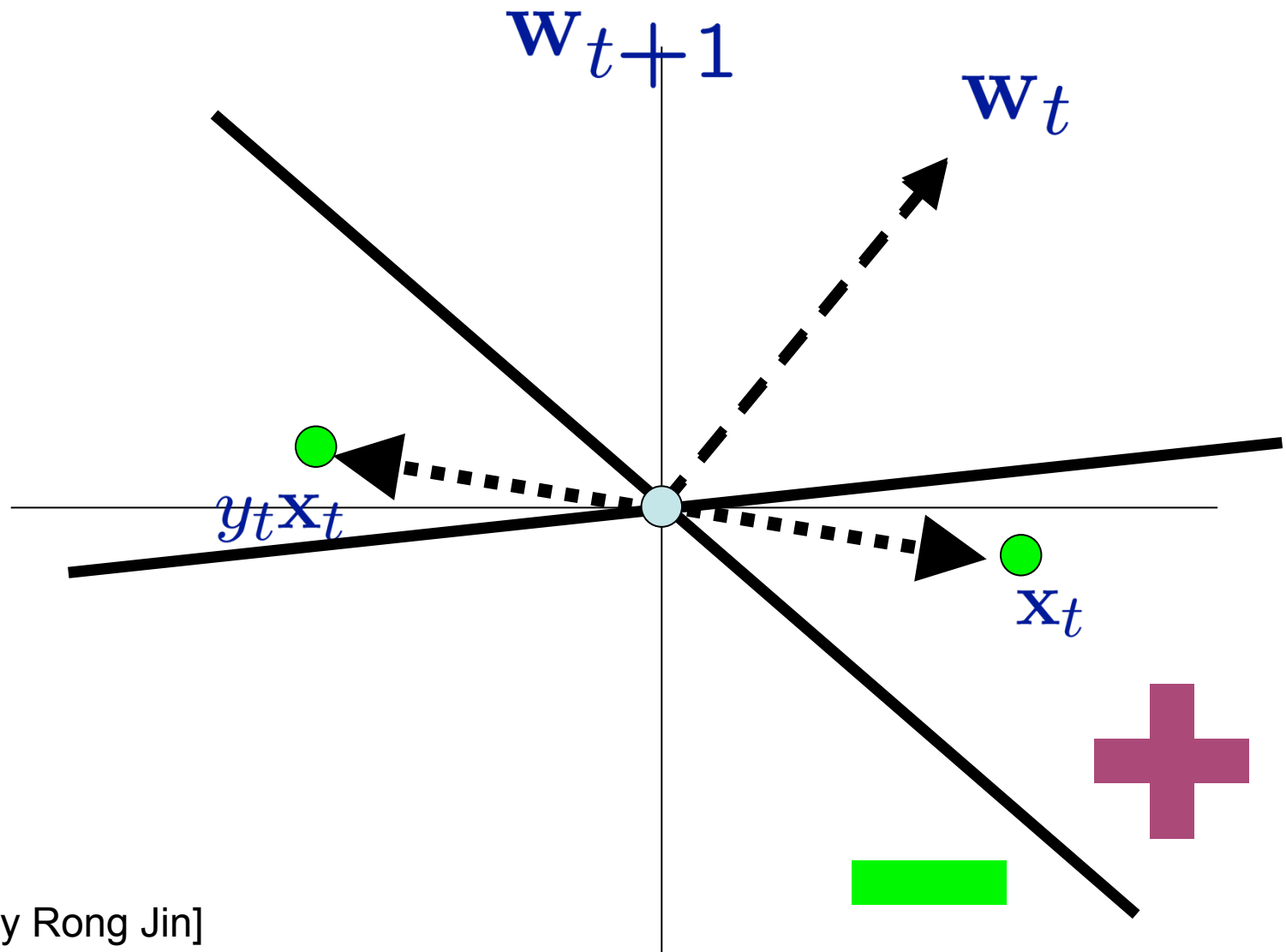
$$y = \begin{cases} +1 & \text{if } w \cdot f(x_i) \geq 0 \\ -1 & \text{if } w \cdot f(x_i) < 0 \end{cases}$$

- If correct (i.e.,  $y=y_i$ ), no change!
- If wrong: update

$$w = w + y_i f(x_i)$$



# Geometrical Interpretation



[Slide by Rong Jin]



## What questions should we ask about a learning algorithm?

- What is the perceptron algorithm's running time?
- If a weight vector with small training error exists, will perceptron find it?
- How well does the resulting classifier generalize to unseen data?

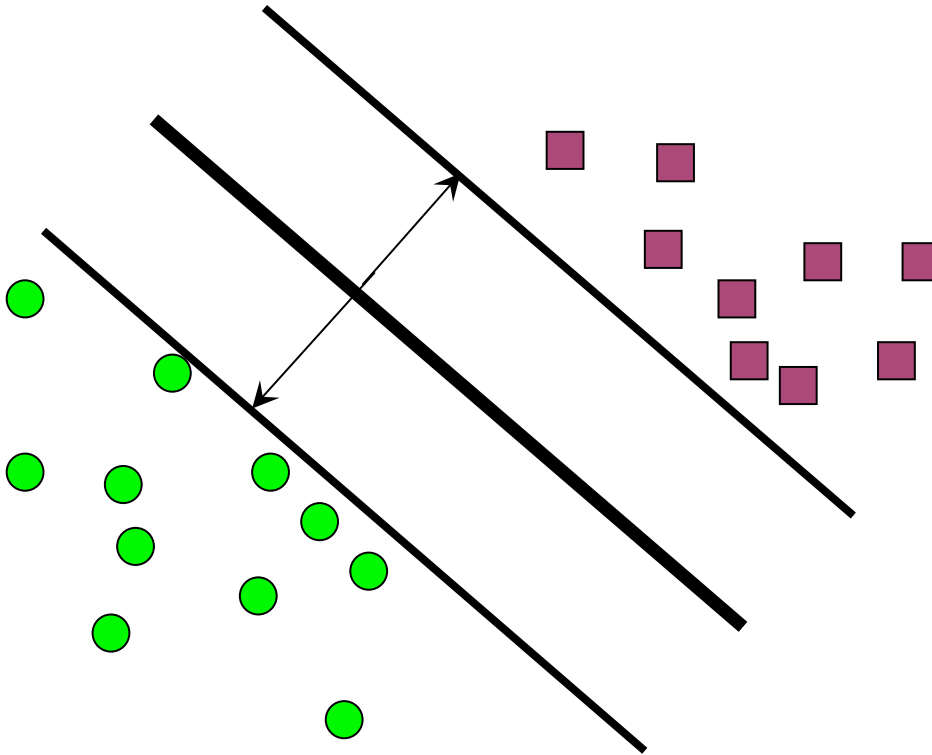
# Linearly Separable

$\exists w$  such that  $\forall t$

$$y_t(w \cdot x_t) \geq \gamma > 0$$



Called the *functional margin*  
with respect to the training set



Equivalently, for  $y_t = +1$ ,

$$w \cdot x_t \geq \gamma$$

and for  $y_t = -1$ ,

$$w \cdot x_t \leq -\gamma$$

# Mistake Bound for Perceptron

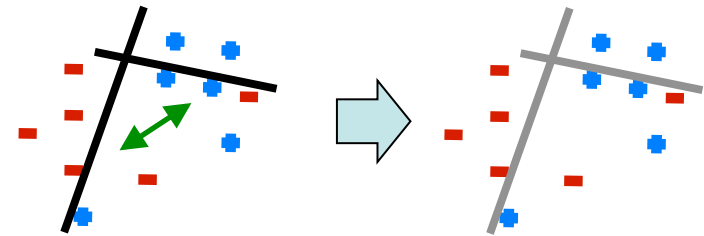
- Assume the data set  $D$  is linearly separable with *geometric* margin  $\gamma$ , i.e.,

$$\exists w^* \text{ s.t. } \|w^*\|_2 = 1 \text{ and } \forall t, y_t(w \cdot x_t) \geq \gamma$$

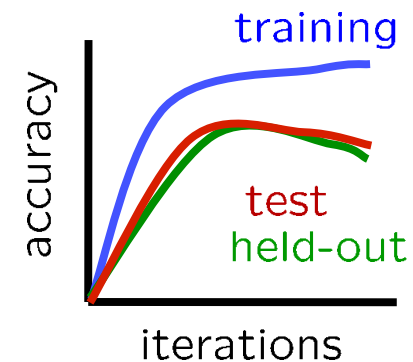
- Assume  $\|x_t\|_2 \leq R, \forall t$
- Theorem: The maximum number of mistakes made by the perceptron algorithm is bounded by  $R^2/\gamma^2$

# Problems with the perceptron algorithm

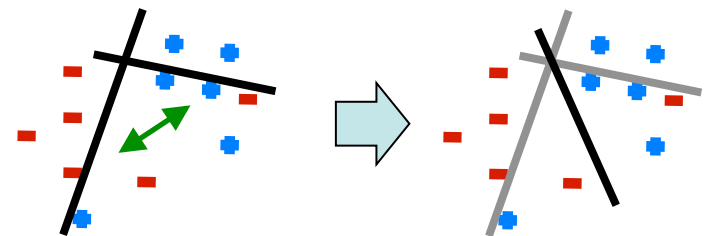
- If the data isn't linearly separable, no guarantees of convergence or training accuracy



- Even if the training data is linearly separable, perceptron can overfit

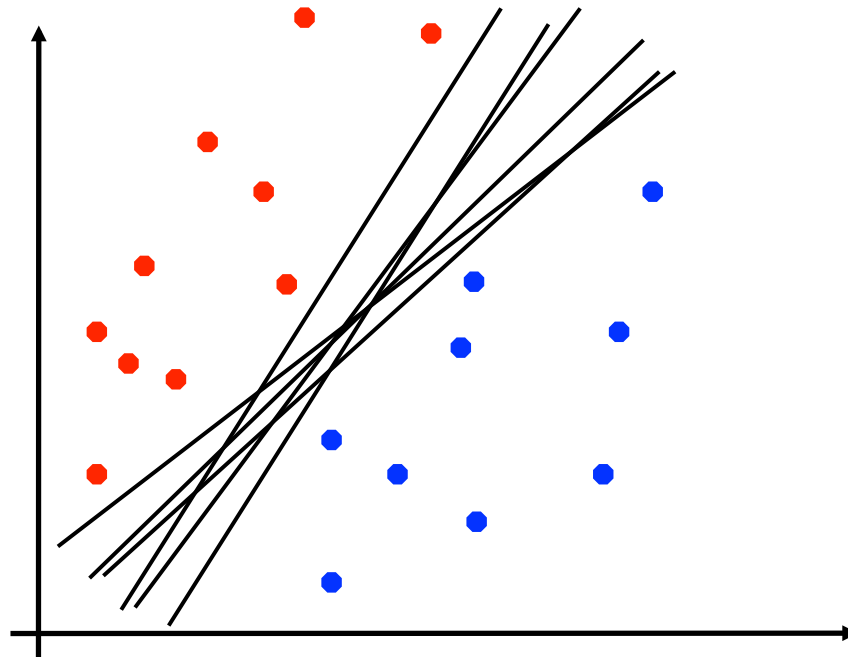


- **Averaged** perceptron is an algorithmic modification that helps with both issues
  - Averages the weight vectors across all iterations



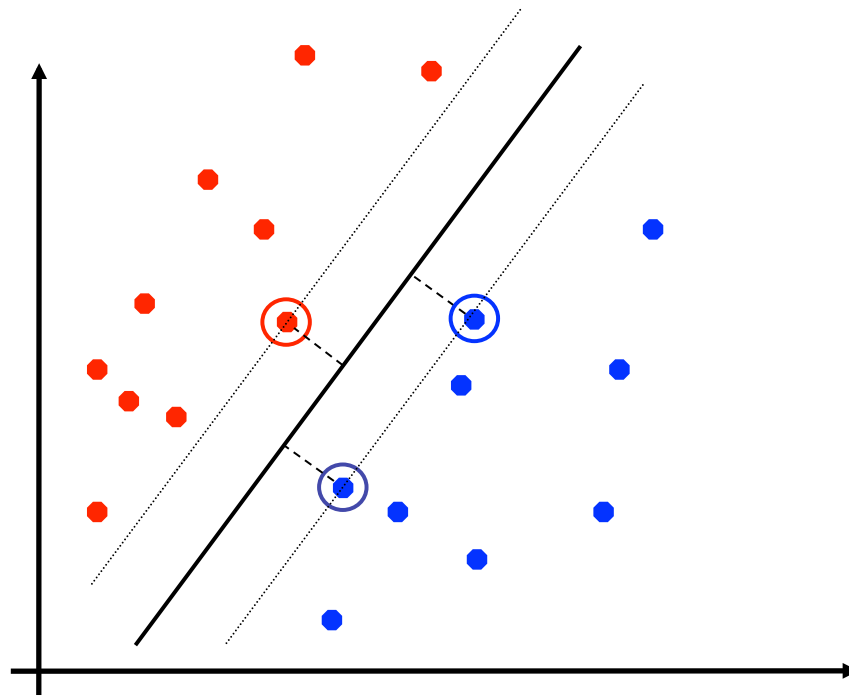
# Linear Separators

- Which of these linear separators is optimal?



# Next week: Support Vector Machines

- SVMs (Vapnik, 1990's) choose the linear separator with the **largest margin**



- Good according to intuition, theory, practice