Introduction to Machine Learning (CSCI-UA.0480-007)

David Sontag
New York University

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

Logistics

- **Class webpage:**
  - http://cs.nyu.edu/~dsontag/courses/ml16/
  - Sign up for Piazza!
- **Office hours:** TBD
- **Teaching assistant**:
  Kevin Jiao <jjiao@stern.nyu.edu>
- **Graders:**
  - Yijun Xiao <ryjxiao@nyu.edu>
  - Alexandre Sablayrolles
    <alexandre.sablayrolles@gmail.com>

Evaluation

- 6-7 homeworks (50%)
  - Both theory and programming
  - Collaboration policy:
    - 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
- Midterm exam (25%)
- Project (20%)
- Course participation (5%)

Projects

- Be creative – think of new problems that you can tackle using machine learning
  - Scope: ~40 hours/person

- Logistics:
  - 2-3 students per group
  - Begins mid-March. Project proposal due week after midterm exam
  - Will still be problem sets during this period!

Prerequisites

REQUIRED:

- **Basic algorithms** (CS 310)
  - Dynamic programming, algorithmic analysis
  - *Can be taken concurrently*

STRONGLY RECOMMENDED:

- **Linear algebra** (Math 140)
  - Matrices, vectors, systems of linear equations
  - Eigenvectors, matrix rank
  - Singular value decomposition
- **Multivariable calculus** (Math 123)
  - Derivatives, integration, tangent planes
  - Optimization, Lagrange multipliers
- **Good programming skills:** Python highly recommended

Source Materials

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

If you really want a book for an additional reference, these are OK options:

- C. Bishop, ***Pattern Recognition and Machine Learning***, Springer, 2007
- K. Murphy, ***Machine Learning: a Probabilistic Perspective***, MIT Press, 2012
- … may update this list throughout semester. I wouldn’t buy anything yet.

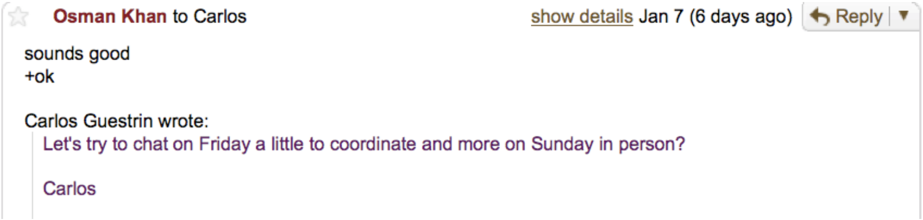
What is Machine Learning ?
(by examples)

Classification

from data to discrete classes

Spam filtering

data

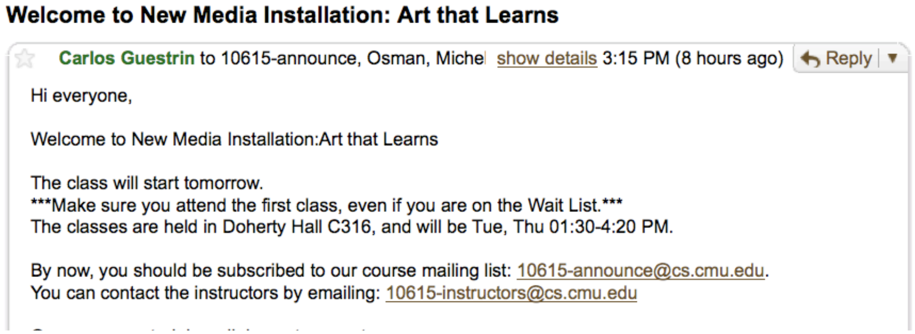


Osman Khan to Carlos show details Jan 7 (6 days ago)

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

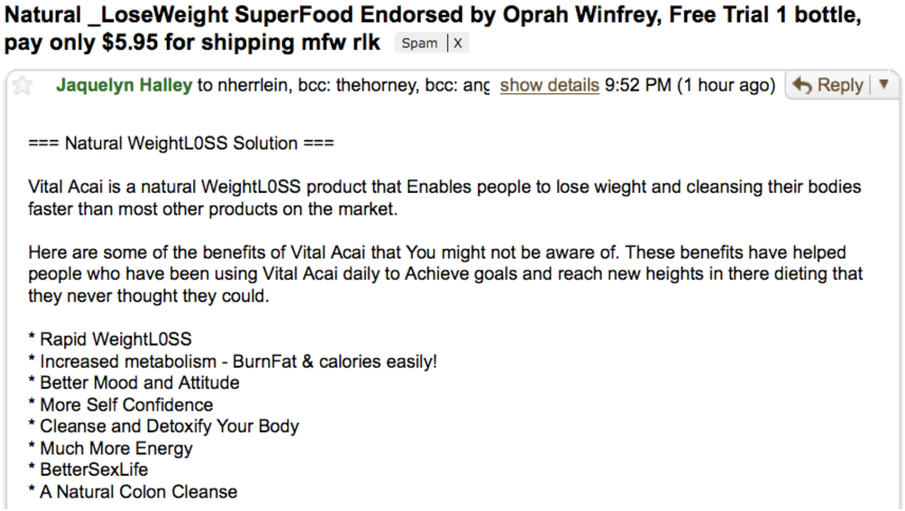
Carlos Guestrin to 10615-announce, Osman, Michel show details 3:15 PM (8 hours ago)

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.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu



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

Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: ang show details 9:52 PM (1 hour ago)

=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS 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 WeightL0SS
* 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

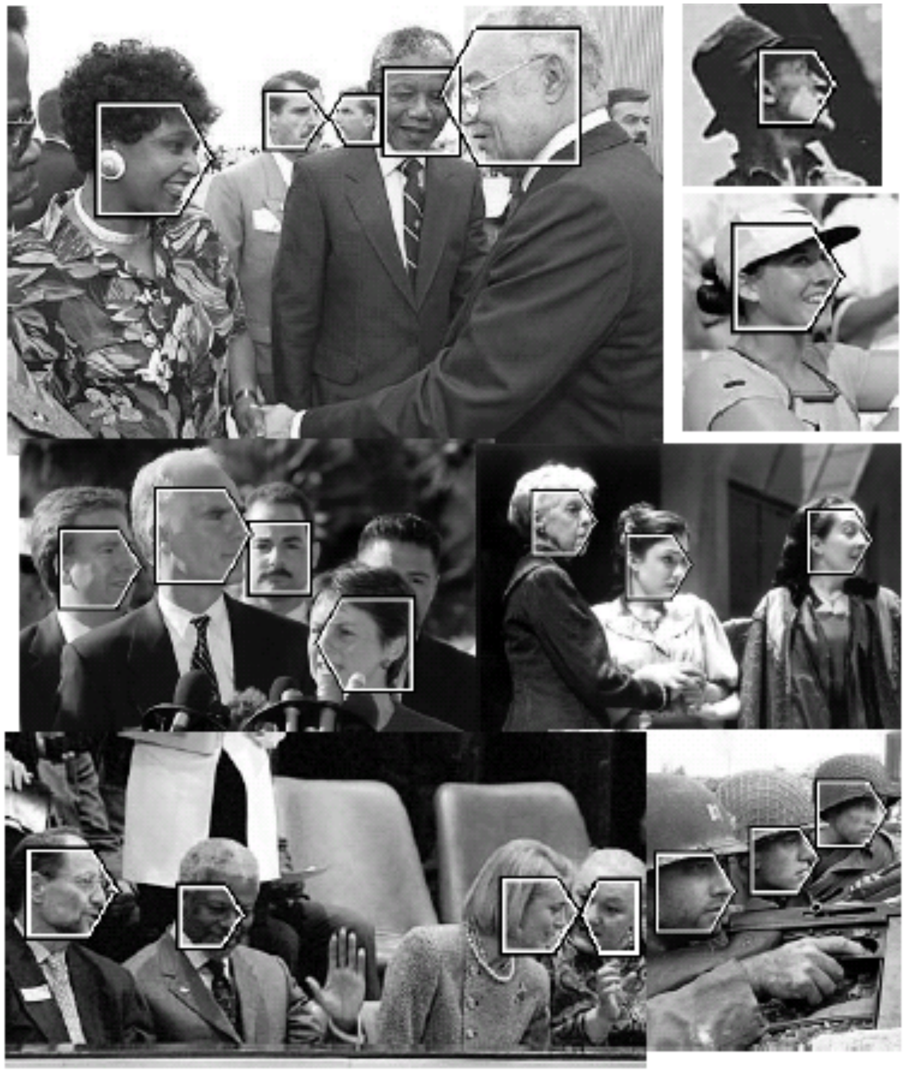
→

prediction

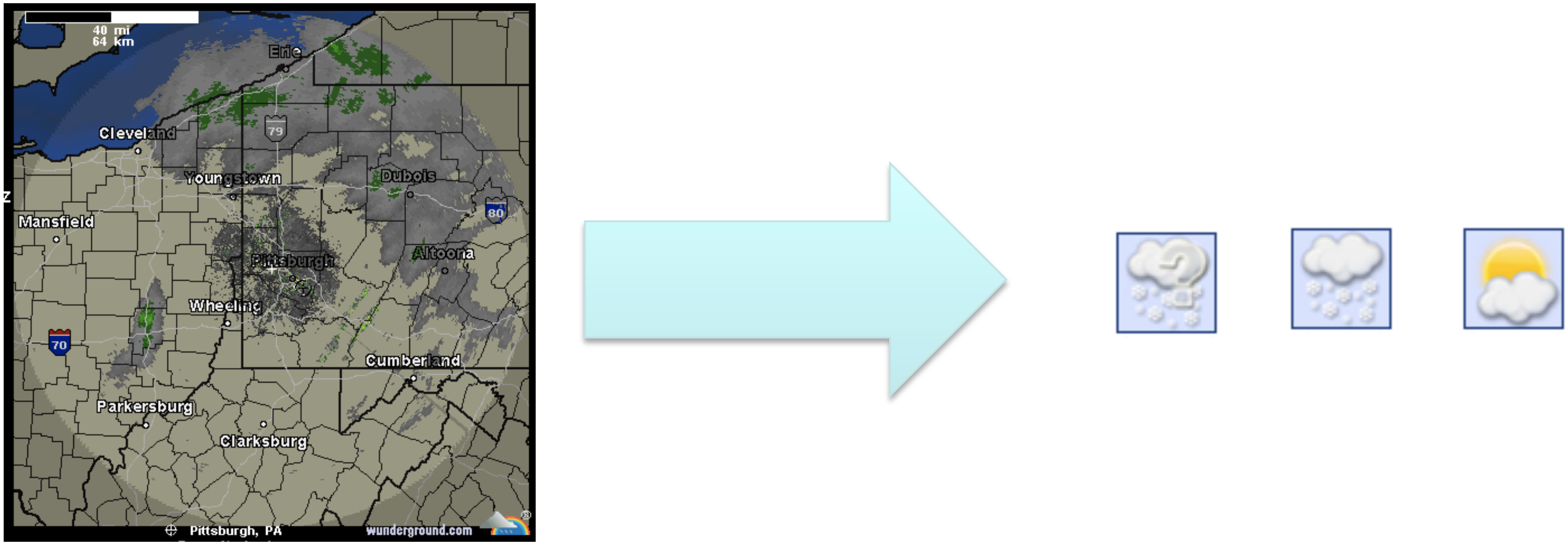
Spam
vs.
Not Spam

Face recognition

Example training images
for each orientation



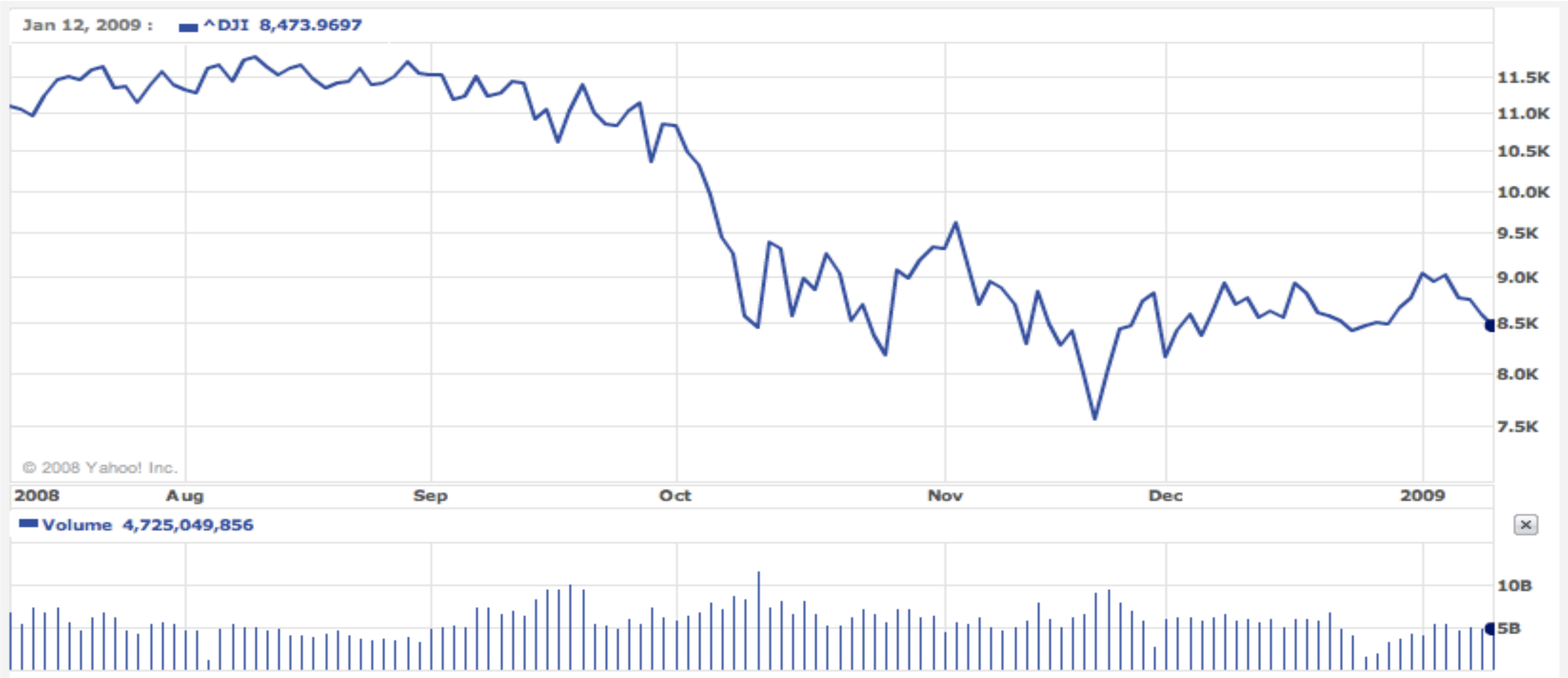
Weather prediction



Regression

predicting a numeric value

Stock market



Weather prediction revisited

40 mi
64 km
Erie
Cleveland
79
Youngstown
Dubois
80
Mansfield
Pittsburgh
Altoona
Wheeling
70
Cumberland
Parkersburg
Clarksburg
Pittsburgh, PA
wunderground.com

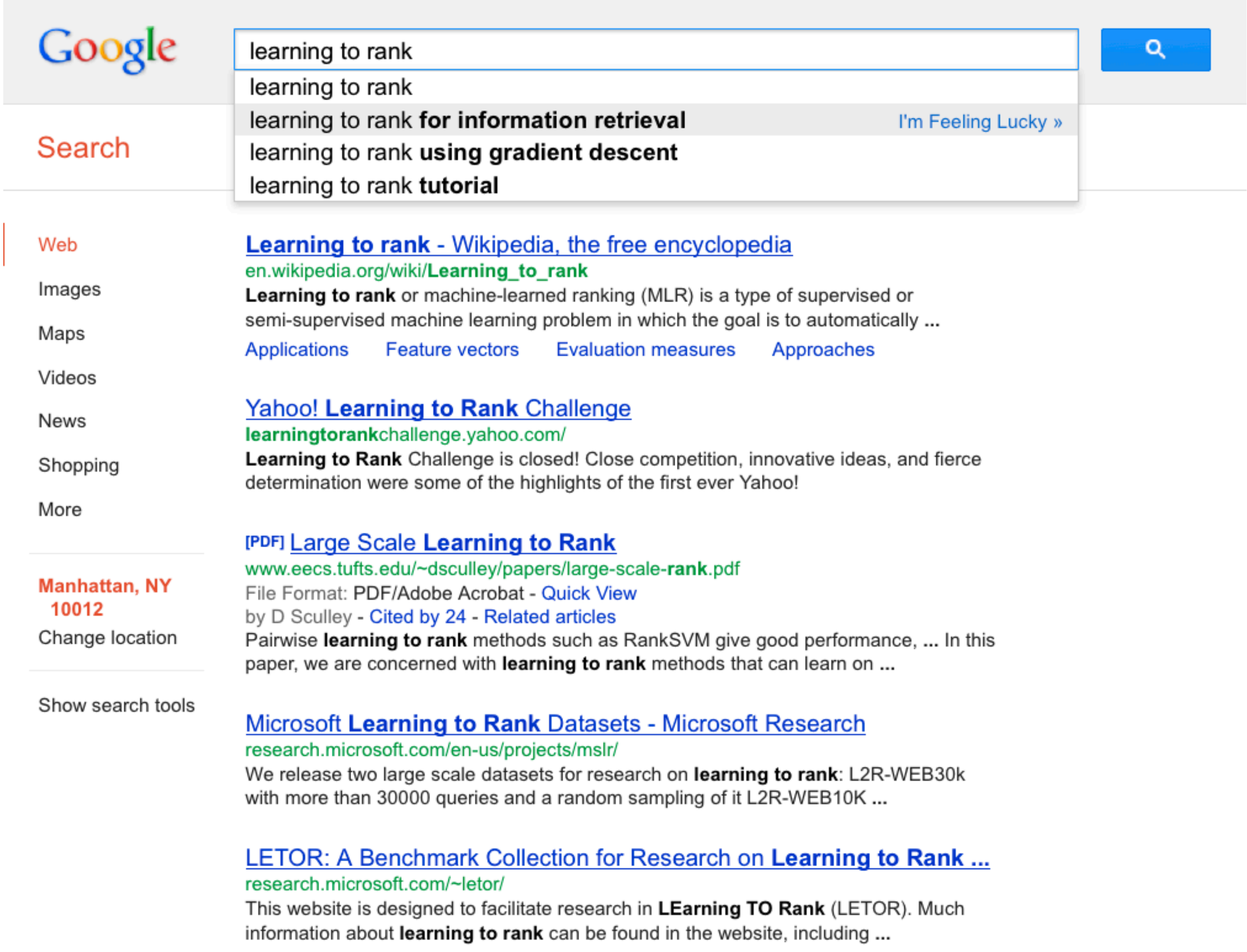
Temperature

72° F

Ranking

comparing items

Web search



learning to rank

learning to rank
learning to rank **for information retrieval**
learning to rank **using gradient descent**
learning to rank **tutorial**

Learning to rank - Wikipedia, the free encyclopedia
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

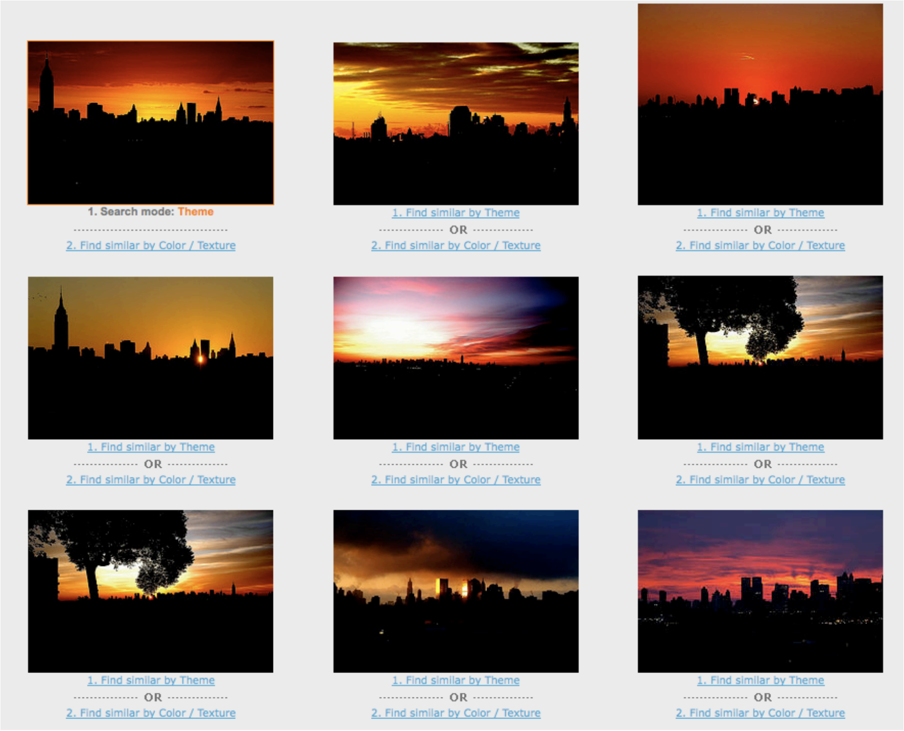
Yahoo! **Learning to Rank** Challenge
learningtorankchallenge.yahoo.com/
Learning to Rank Challenge is closed! Close competition, innovative ideas, and fierce determination were some of the highlights of the first ever Yahoo!

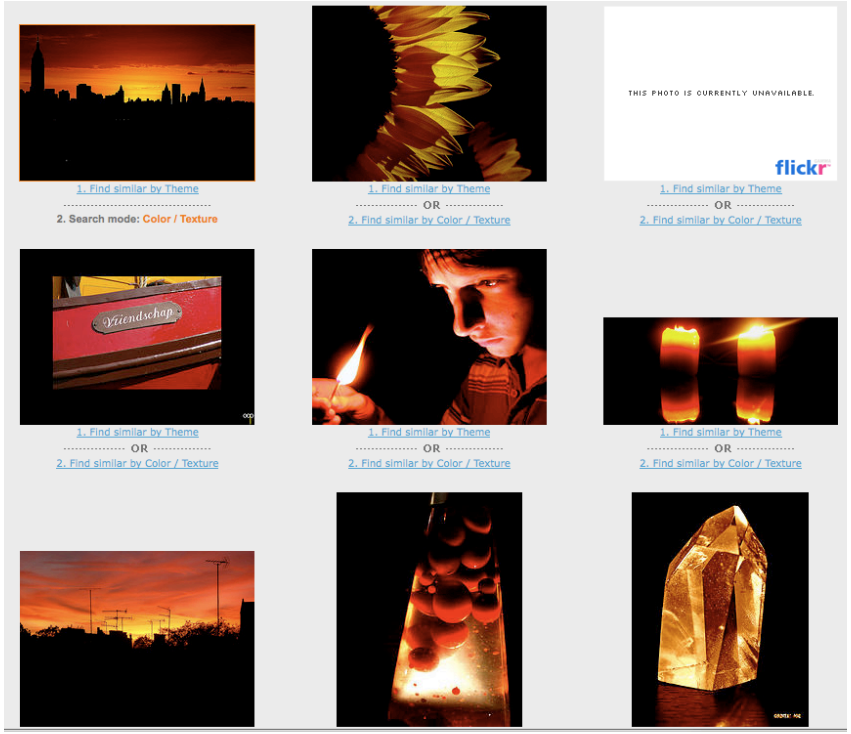
[PDF] Large Scale **Learning to Rank**
www.eecs.tufts.edu/~dsculley/papers/large-scale-**rank**.pdf
File Format: PDF/Adobe Acrobat - Quick View
by D Sculley - Cited by 24 - Related articles
Pairwise **learning to rank** methods such as RankSVM give good performance, ... In this paper, we are concerned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

Microsoft **Learning to Rank** Datasets - Microsoft Research
research.microsoft.com/en-us/projects/mslr/
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/
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

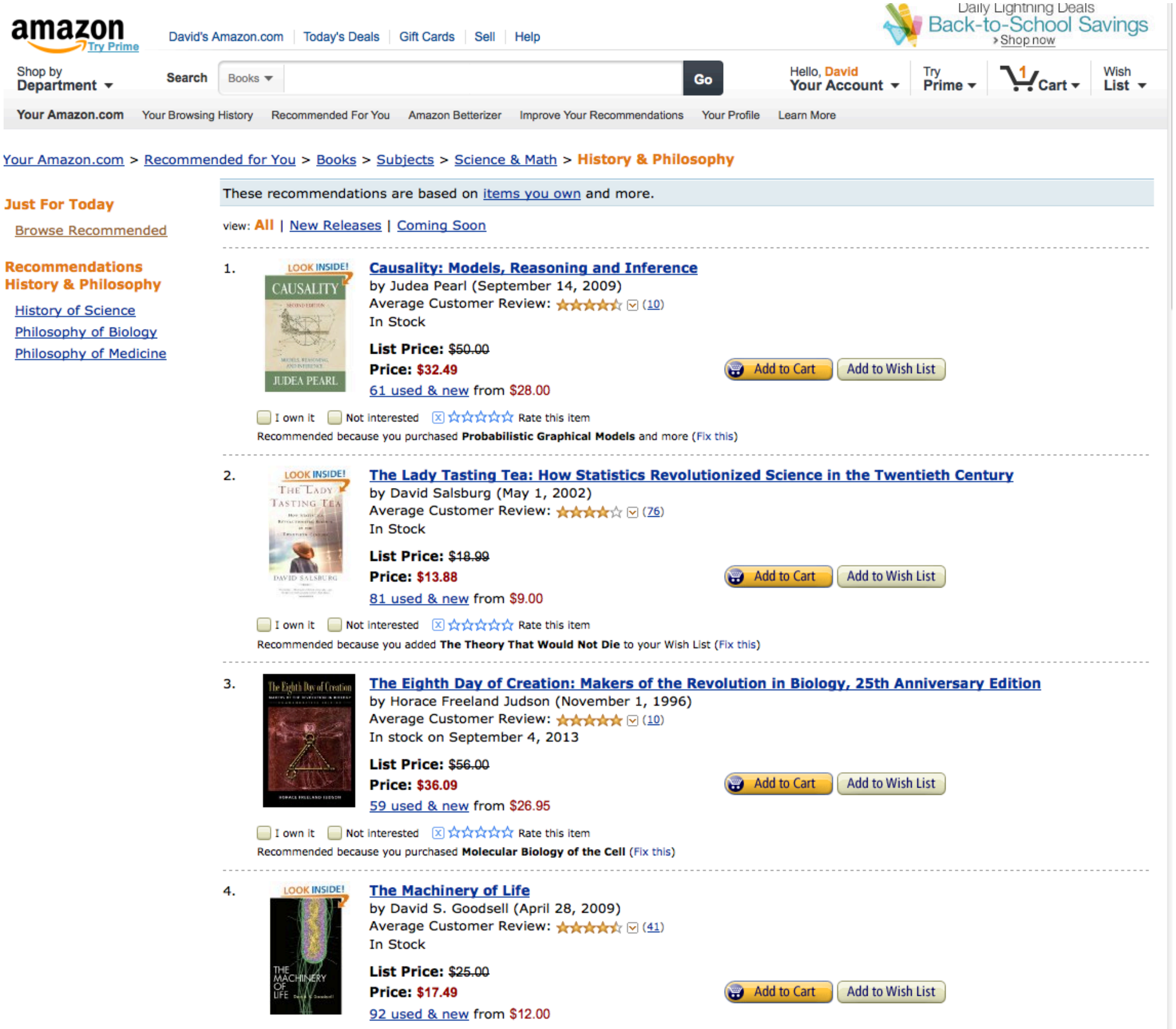

1. Search mode: Theme
2. Find similar by Color / Texture
1. Find similar by Theme
OR
2. Find similar by Color / Texture
1. Find similar by Theme
OR
2. Find similar by Color / Texture
1. Find similar by Theme
OR
2. Find similar by Color / Texture
1. Find similar by Theme
OR
2. Find similar by Color /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

http://www.tiltomo.com/

Collaborative Filtering

Recommendation systems



amazon Try Prime

David's Amazon.com | Today's Deals | Gift Cards | Sell | Help

Daily Lightning Deals
Back-to-School Savings
› Shop now

Shop by Department ▾ | Search | Books ▾ | Go | Hello, David Your Account ▾ | Try Prime ▾ | 1 Cart ▾ | Wish List ▾

Your Amazon.com | Your Browsing History | Recommended For You | Amazon Betterizer | Improve Your Recommendations | Your Profile | Learn More

Your Amazon.com > Recommended for You > Books > Subjects > Science & Math > **History & Philosophy**

Just For Today

Browse Recommended

Recommendations History & Philosophy

History of Science

Philosophy of Biology

Philosophy of Medicine

These recommendations are based on items you own and more.

view: **All** | New Releases | Coming Soon

1. **Causality: Models, Reasoning and Inference**
by Judea Pearl (September 14, 2009)
Average Customer Review: ★★★★½ (10)
In Stock
List Price: ~~$50.00~~
Price: $32.49
61 used & new from $28.00
Add to Cart | Add to Wish List
□ I own it □ Not interested ☒ ☆☆☆☆☆ Rate this item
Recommended because you purchased **Probabilistic Graphical Models** and more (Fix this)

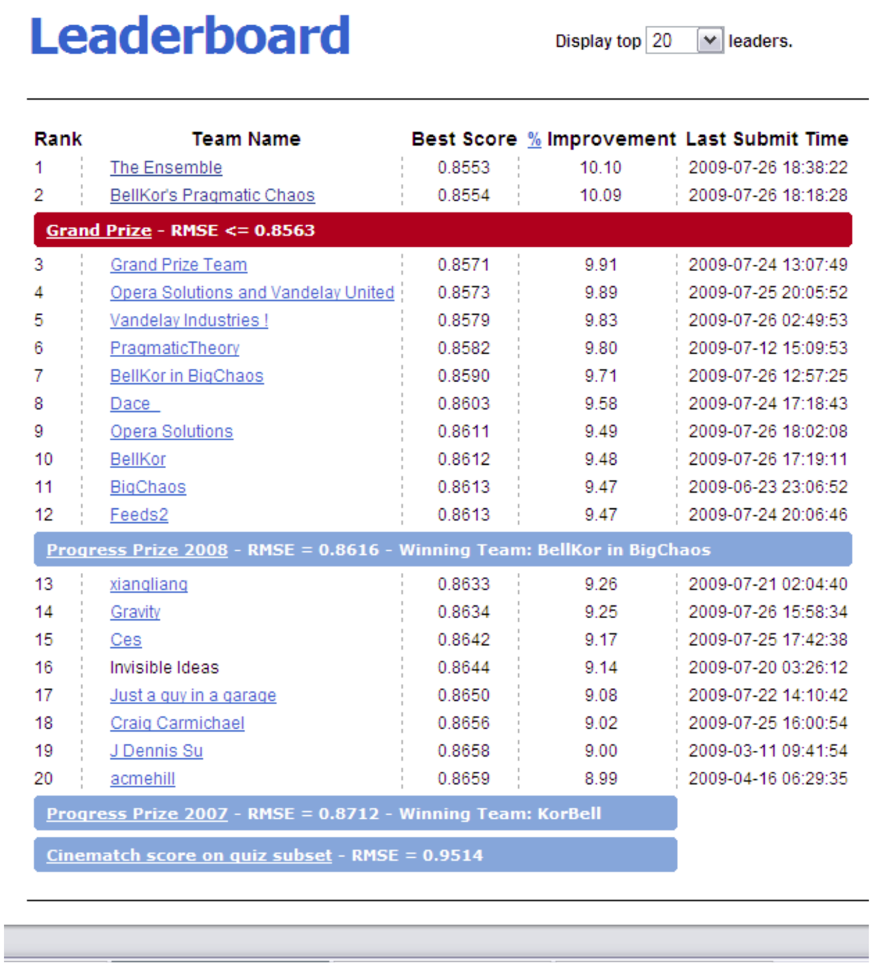
2. **The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century**
by David Salsburg (May 1, 2002)
Average Customer Review: ★★★★☆ (76)
In Stock
List Price: ~~$18.99~~
Price: $13.88
81 used & new from $9.00
Add to Cart | Add to Wish List
□ I own it □ Not interested ☒ ☆☆☆☆☆ Rate this item
Recommended because you added **The Theory That Would Not Die** to your Wish List (Fix this)

3. **The Eighth Day of Creation: Makers of the Revolution in Biology, 25th Anniversary Edition**
by Horace Freeland Judson (November 1, 1996)
Average Customer Review: ★★★★★ (10)
In stock on September 4, 2013
List Price: ~~$56.00~~
Price: $36.09
59 used & new from $26.95
Add to Cart | Add to Wish List
□ I own it □ Not interested ☒ ☆☆☆☆☆ Rate this item
Recommended because you purchased **Molecular Biology of the Cell** (Fix this)

4. **The Machinery of Life**
by David S. Goodsell (April 28, 2009)
Average Customer Review: ★★★★½ (41)
In Stock
List Price: ~~$25.00~~
Price: $17.49
92 used & new from $12.00
Add to Cart | Add to Wish List

Recommendation systems

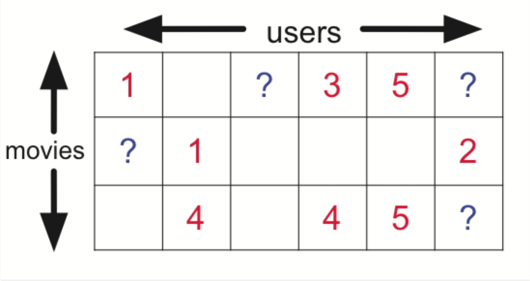
Machine learning competition with a $1 million prize



Leaderboard

Display top 20 leaders.

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

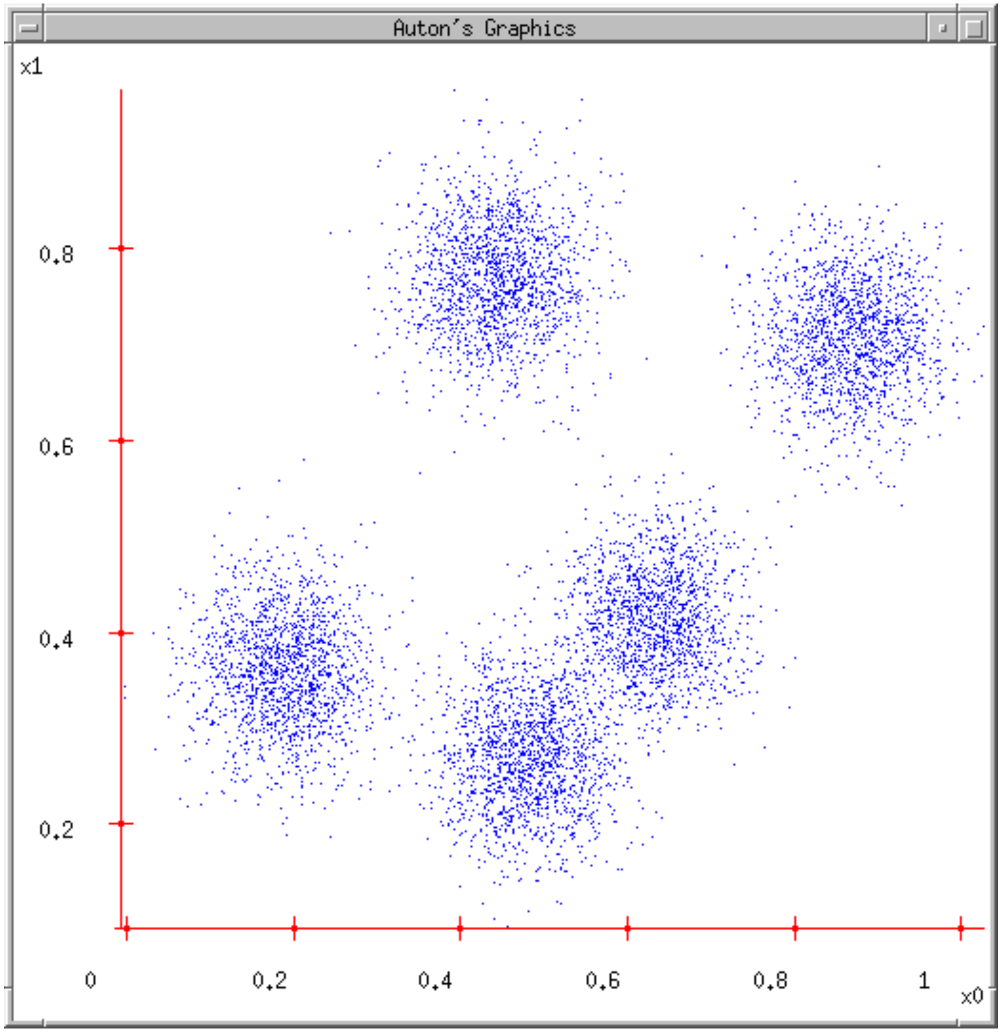


1		?	3	5	?
?	1				2
	4		4	5	?

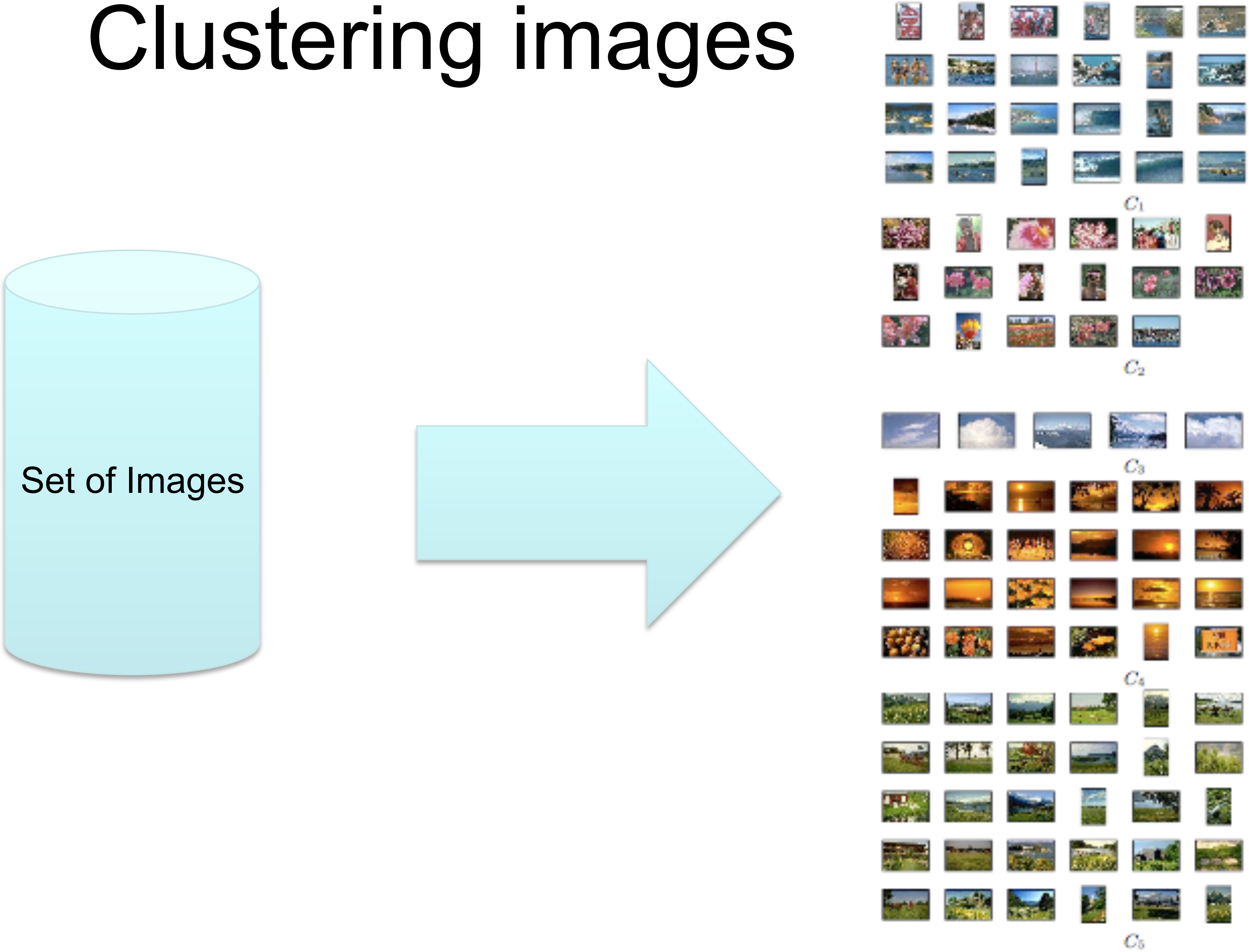
Clustering

discovering structure in data

Clustering Data: Group similar things



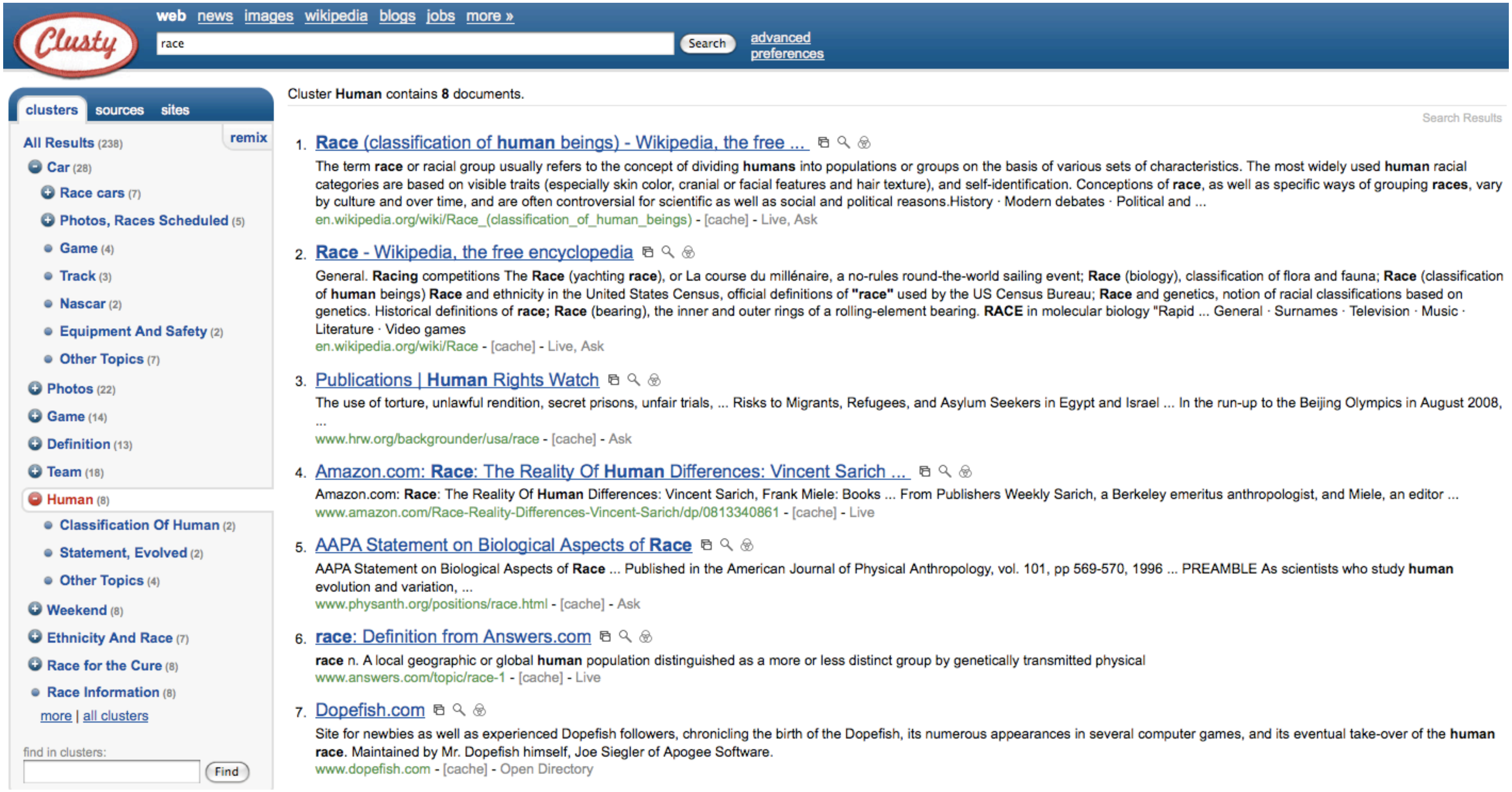
Clustering images



Set of Images

[Goldberger et al.]

Clustering web search results



- All Results (238)
  - 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

Cluster **Human** contains **8** documents.

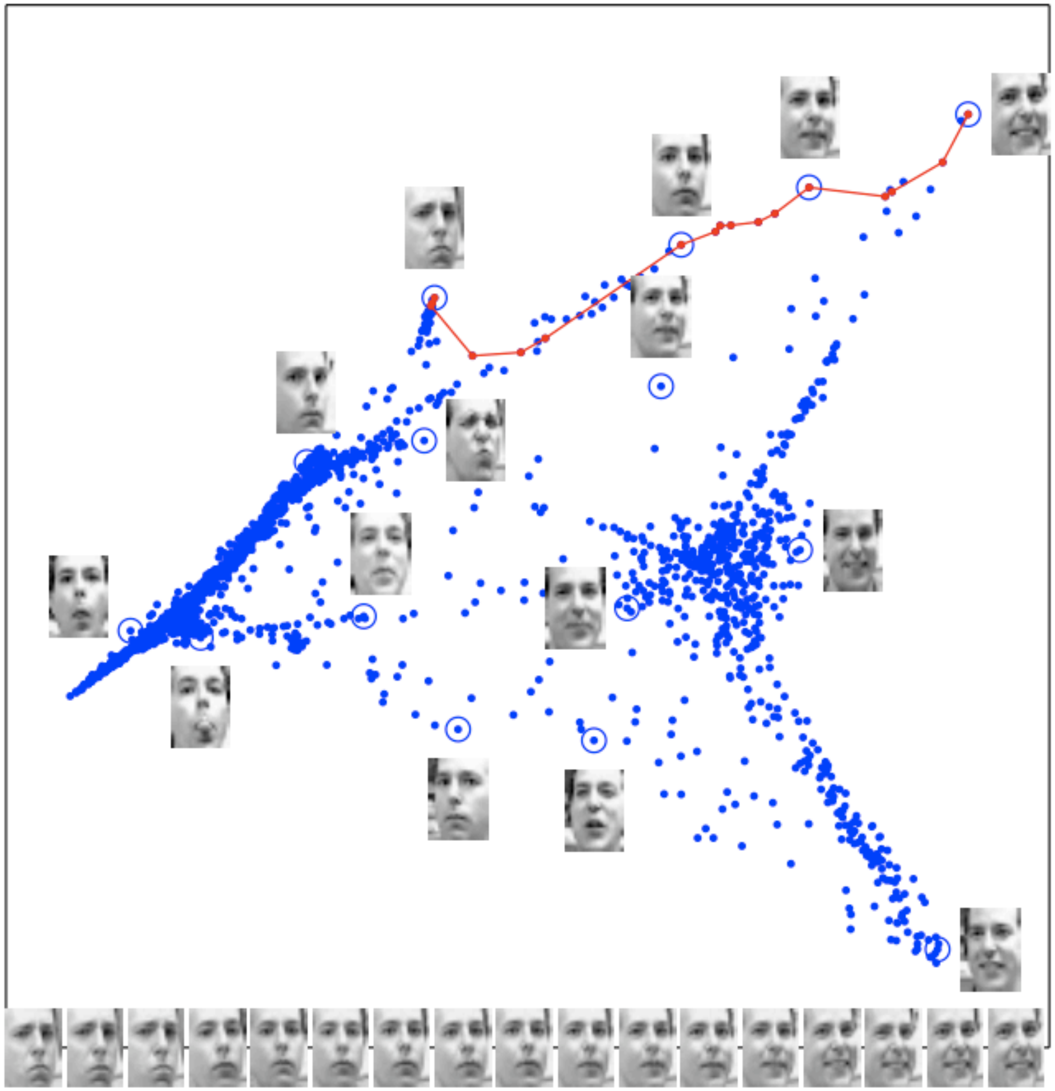
1. **Race** (classification of **human** beings) - Wikipedia, the free ...
   The term **race** or racial group usually refers to the concept of dividing **humans** into populations or groups on the basis of various sets of characteristics. The most widely used **human** racial categories are based on 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
2. **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
3. 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/backgrounder/usa/race - [cache] - Ask
4. 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
5. 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
6. **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
7. 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

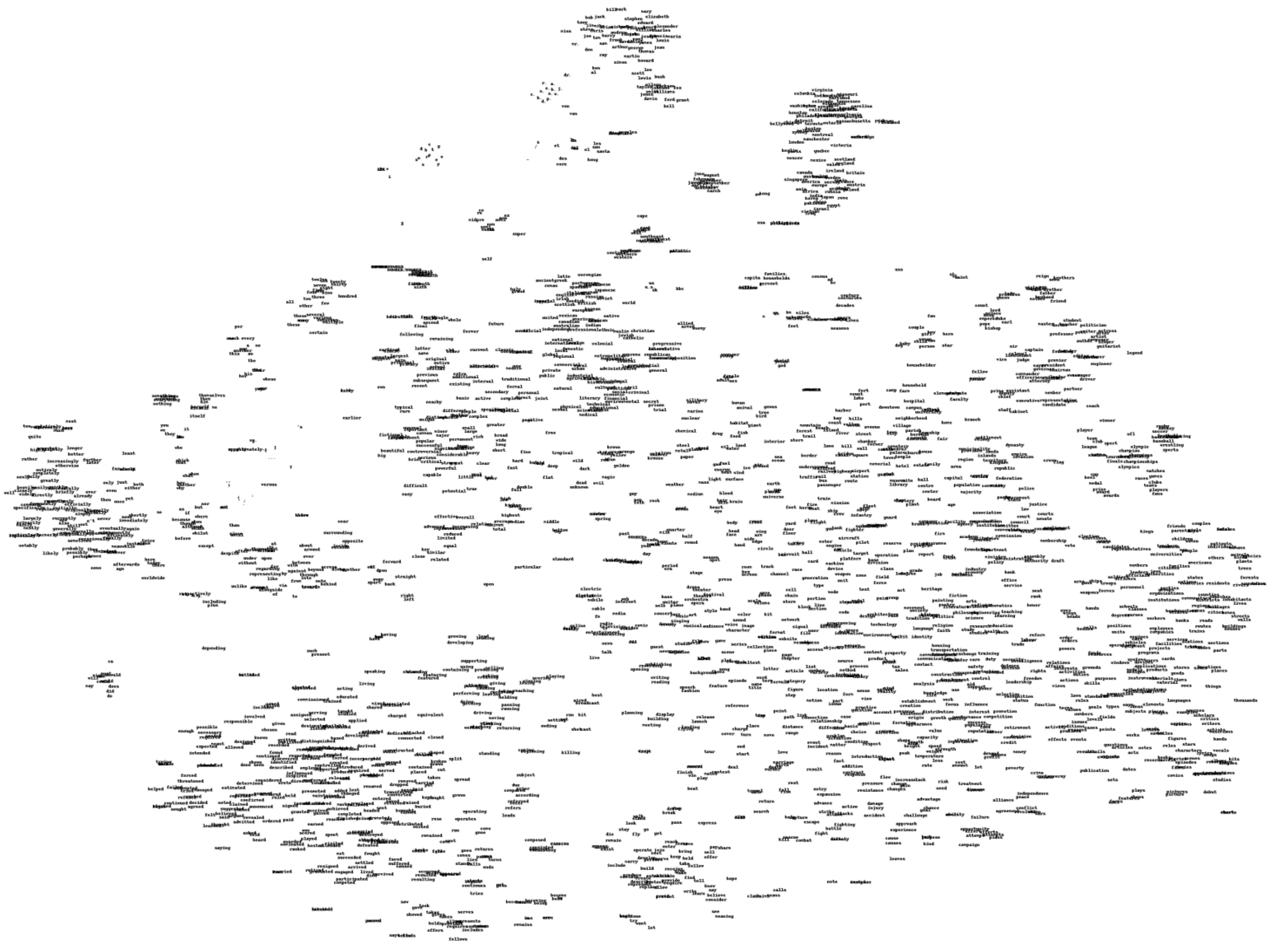
Embedding images

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



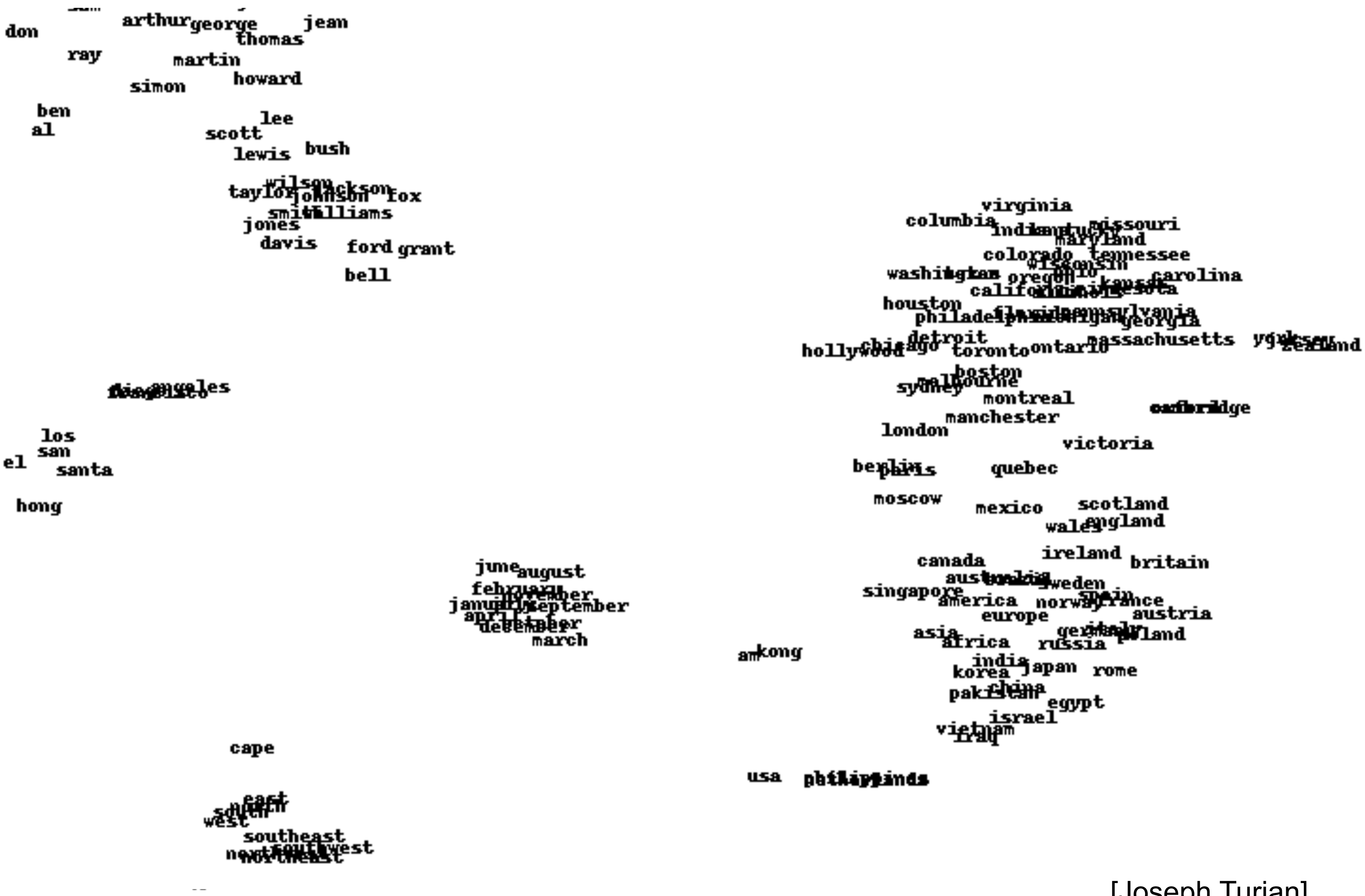
[Saul & Roweis '03]

Embedding words



[Joseph Turian]

Embedding words (zoom in)

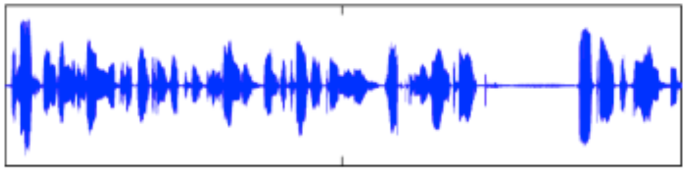


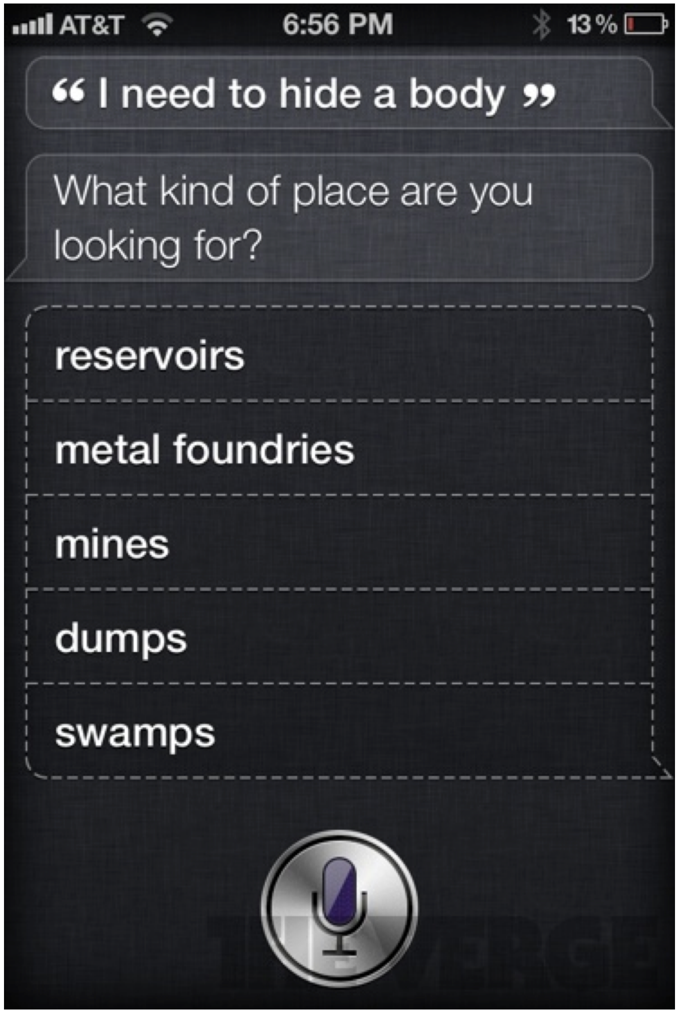
[Joseph Turian]

Structured prediction

from data to discrete classes

Speech recognition





“ I need to hide a body ”

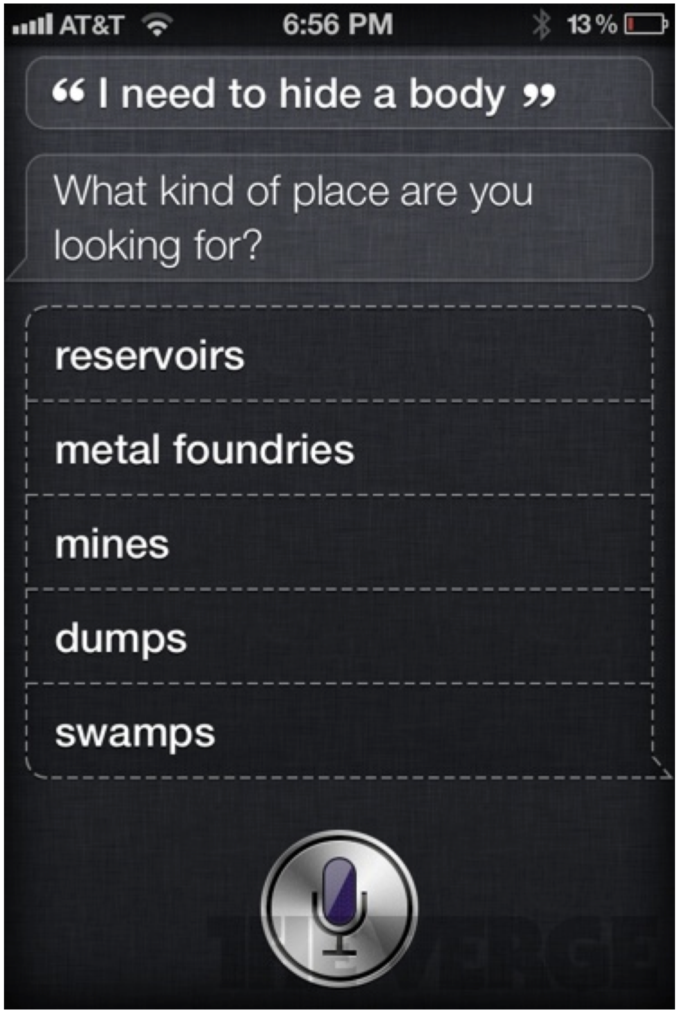
What kind of place are you looking for?

- reservoirs
- metal foundries
- mines
- dumps
- swamps

Natural language processing

I need to hide a body

noun, verb, preposition, …



“ I need to hide a body ”

What kind of place are you looking for?

reservoirs

metal foundries

mines

dumps

swamps

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: supervised learning**
  - SVMs, kernel methods
  - Learning theory
  - Decision trees, boosting, deep learning

- **Second half of course: data science**
  - Unsupervised learning, EM algorithm
  - Dimensionality reduction
  - Topic models

Supervised Learning: find f

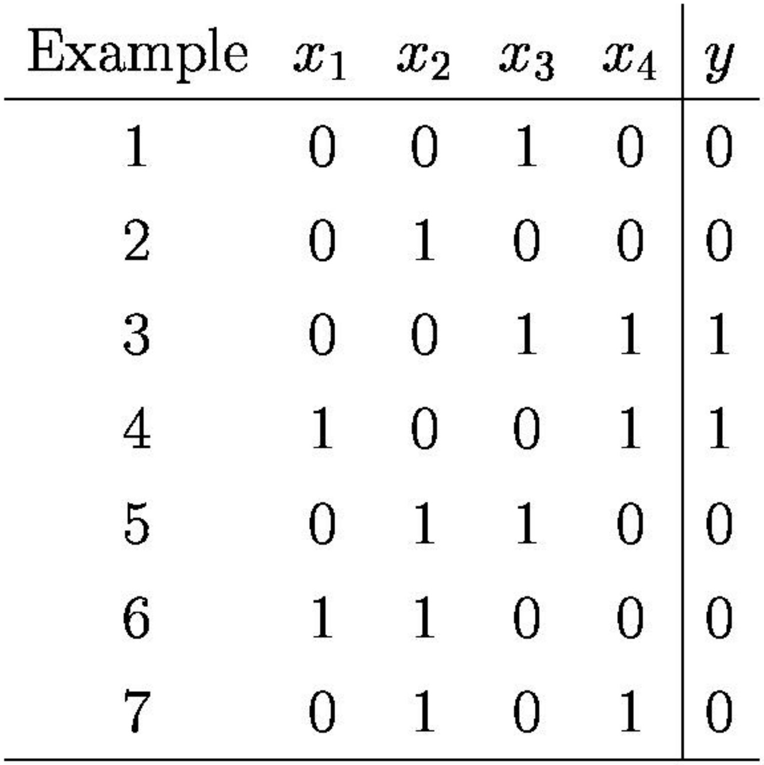
- Given: Training set $\{(x_i, y_i) \mid i = 1 \ldots N\}$
- Find: A good approximation to $f : X \rightarrow Y$

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 $\Re$ (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

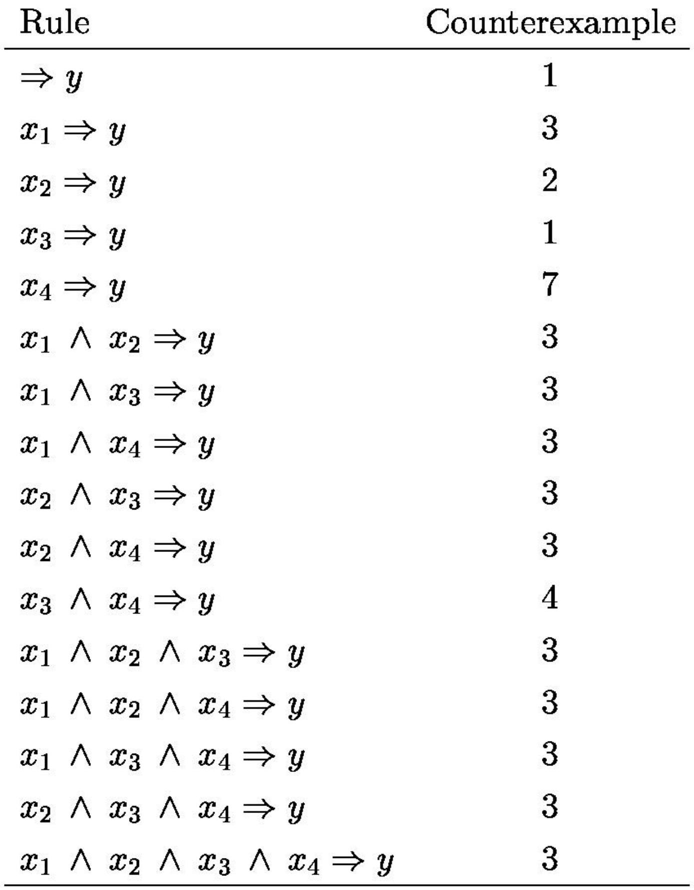
A Restricted Hypothesis Space

Consider all conjunctive boolean functions.

- 16 possible hypotheses

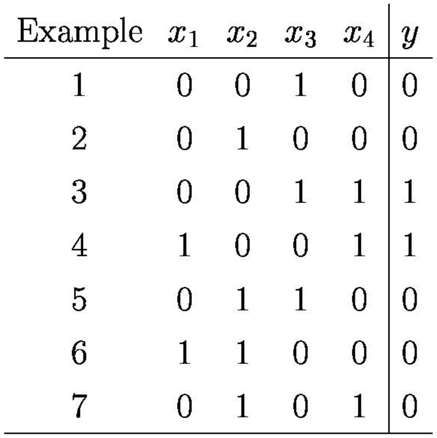
- None are consistent with our dataset

- 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

Occam's Razor Principle

- William of Occam: Monk living in the 14th 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)