Introduction to Machine Learning (CSCI-UA.0480-002)

David Sontag
New York University

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

Logistics

- **Class webpage:**
  - http://cs.nyu.edu/~dsontag/courses/ml12/
  - Sign up for mailing list!

- **Office hours:**
  - Tuesday 3:30-4:30pm and by appointment.
  - 715 Broadway, 12th floor, Room 1204

- **Grader:** Jinglun Dong
  - Email: jinglundong@gmail.com

Evaluation

- About 7 homeworks (50%)
  - Both theory and programming
  - See collaboration policy on class webpage
- Midterm & final exam (45%)
- Course participation (5%)

Prerequisites

- **Basic algorithms** (CS 310)
  - Dynamic programming, algorithmic analysis
- **Linear algebra** (Math 140)
  - Matrices, vectors, systems of linear equations
  - Eigenvectors, matrix rank
  - Singular value decomposition
- **Multivariable calculus** (Math 123)
  - Derivatives, integration, tangent planes
  - Lagrange multipliers
- **Probability** (Math 233 or 235)

Source Materials

Optional textbooks:

- C. Bishop, ***Pattern Recognition and Machine Learning***, Springer, 2007
- K. Murphy, ***Machine Learning: a Probabilistic Perspective***, MIT Press, 2012

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, former 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, former Dir. Research, Yahoo)
- “Machine learning is going to result in a real revolution” (Greg Papadopoulos, former CTO, Sun)
- “Machine learning is today’s discontinuity” (Jerry Yang, former CEO, Yahoo)

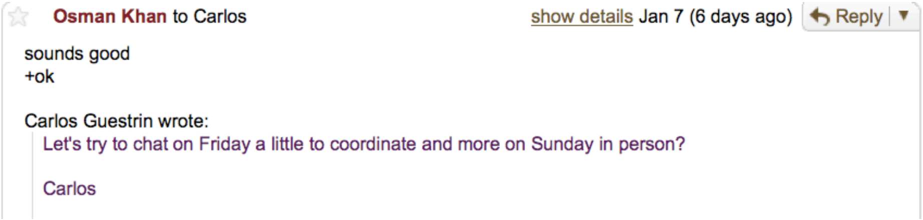
What is Machine Learning ?
(by examples)

Classification

from data to discrete classes

Spam filtering

data

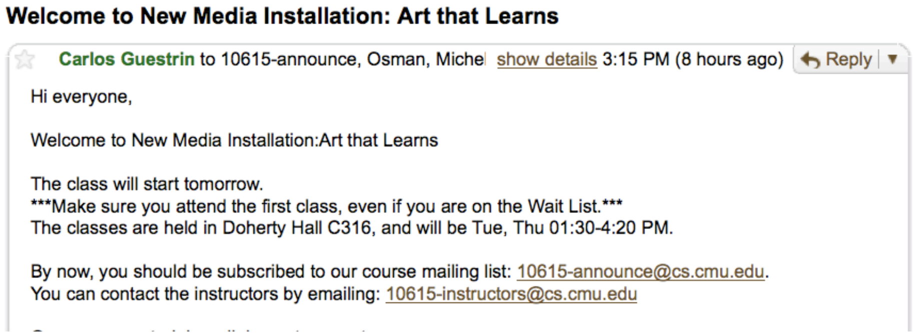


Osman Khan to Carlos — Jan 7 (6 days ago)

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

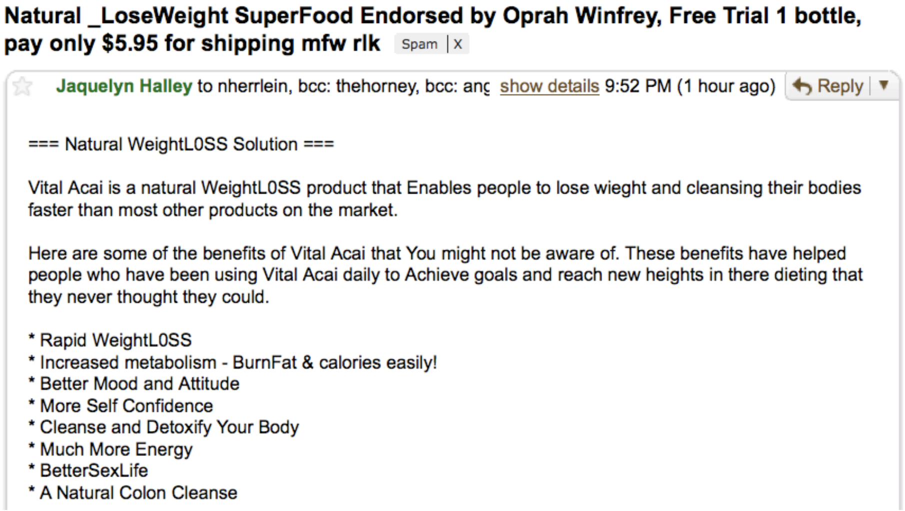
Carlos Guestrin to 10615-announce, Osman, Michel — 3:15 PM (8 hours ago)

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 — 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

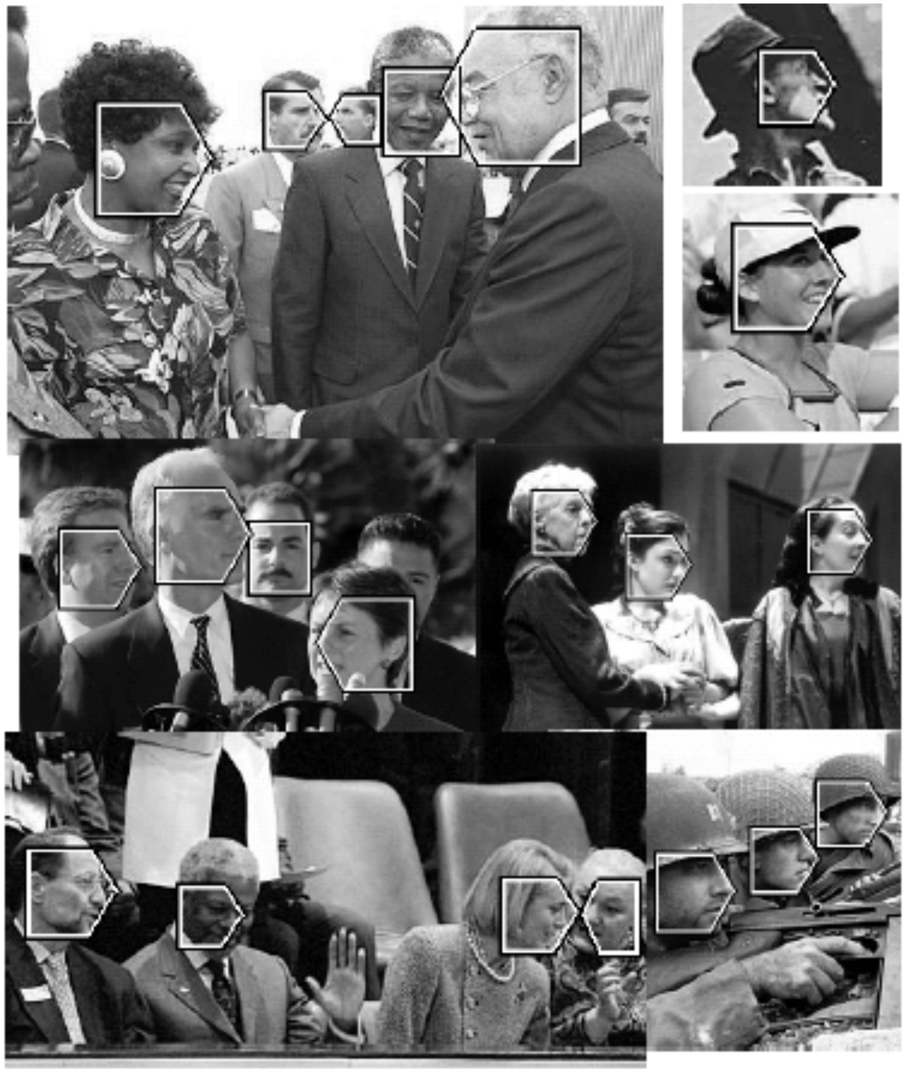
→

prediction

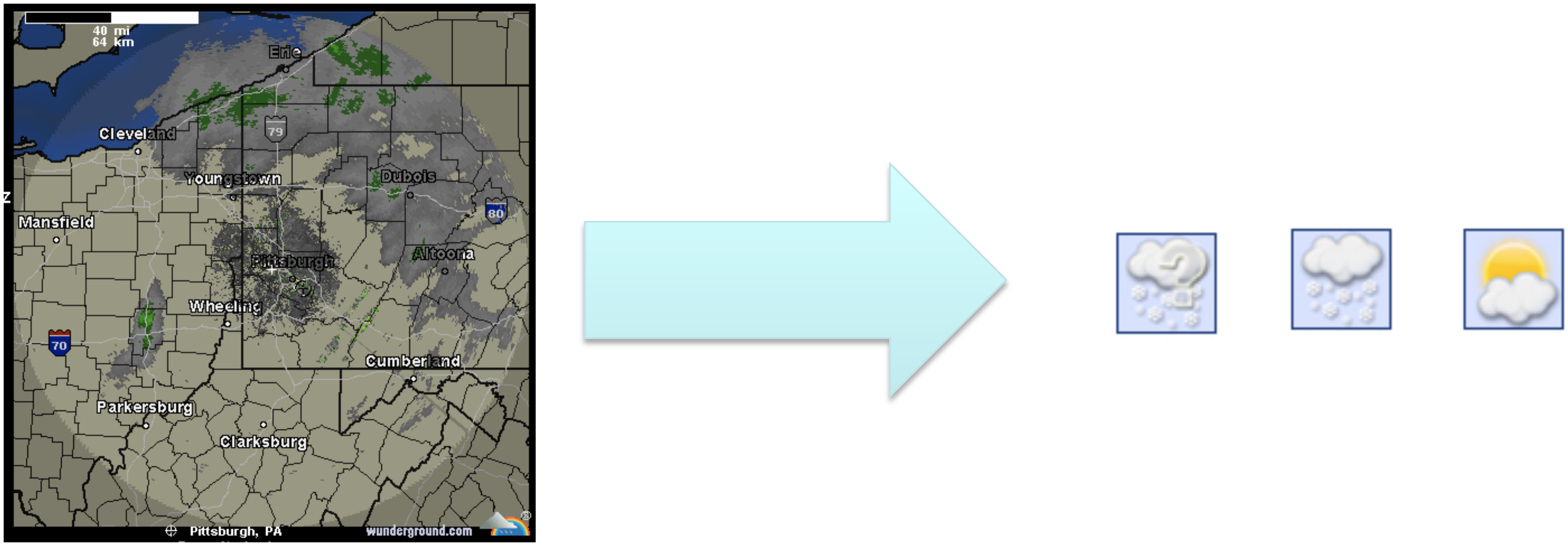
Spam
vs.
Not Spam

Object detection

Example training images for each orientation



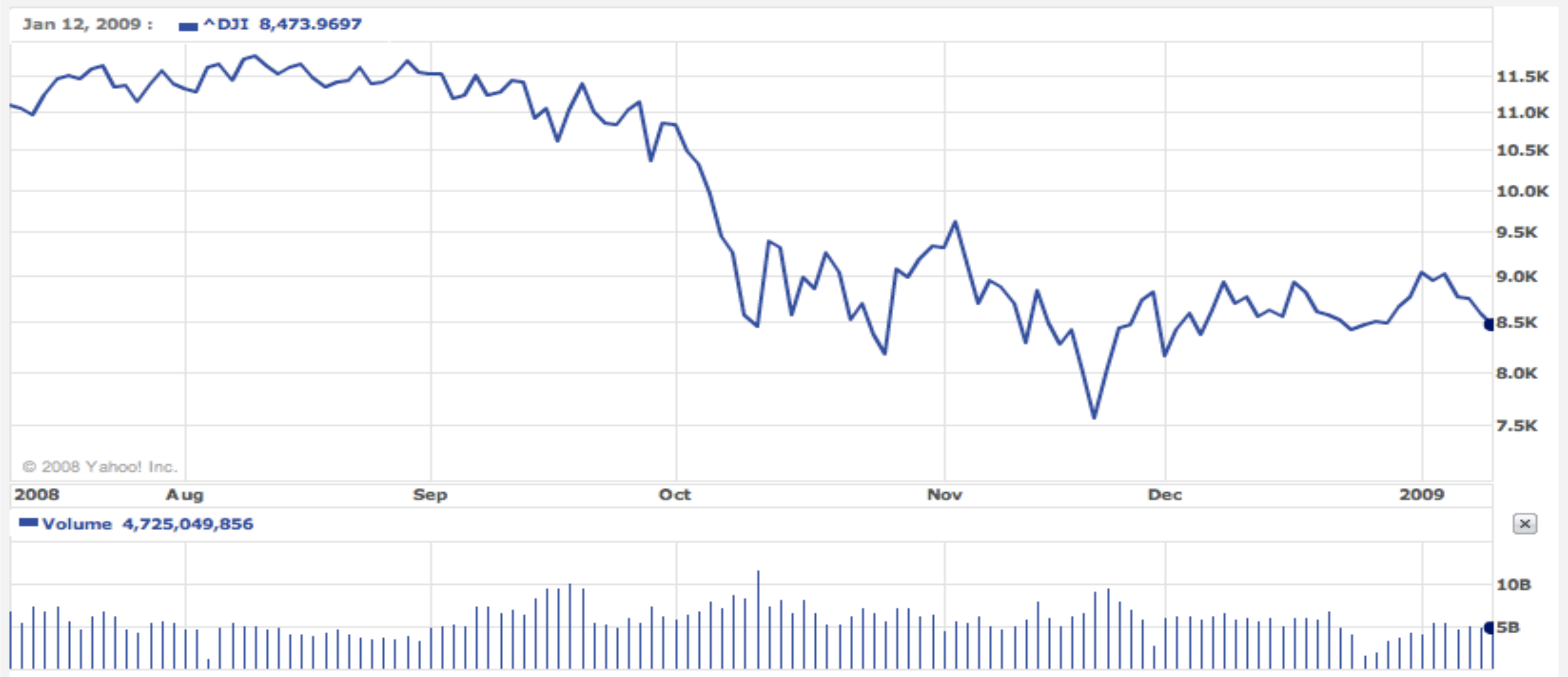
Weather prediction



Regression

predicting a numeric value

Stock market



Weather prediction revisted

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

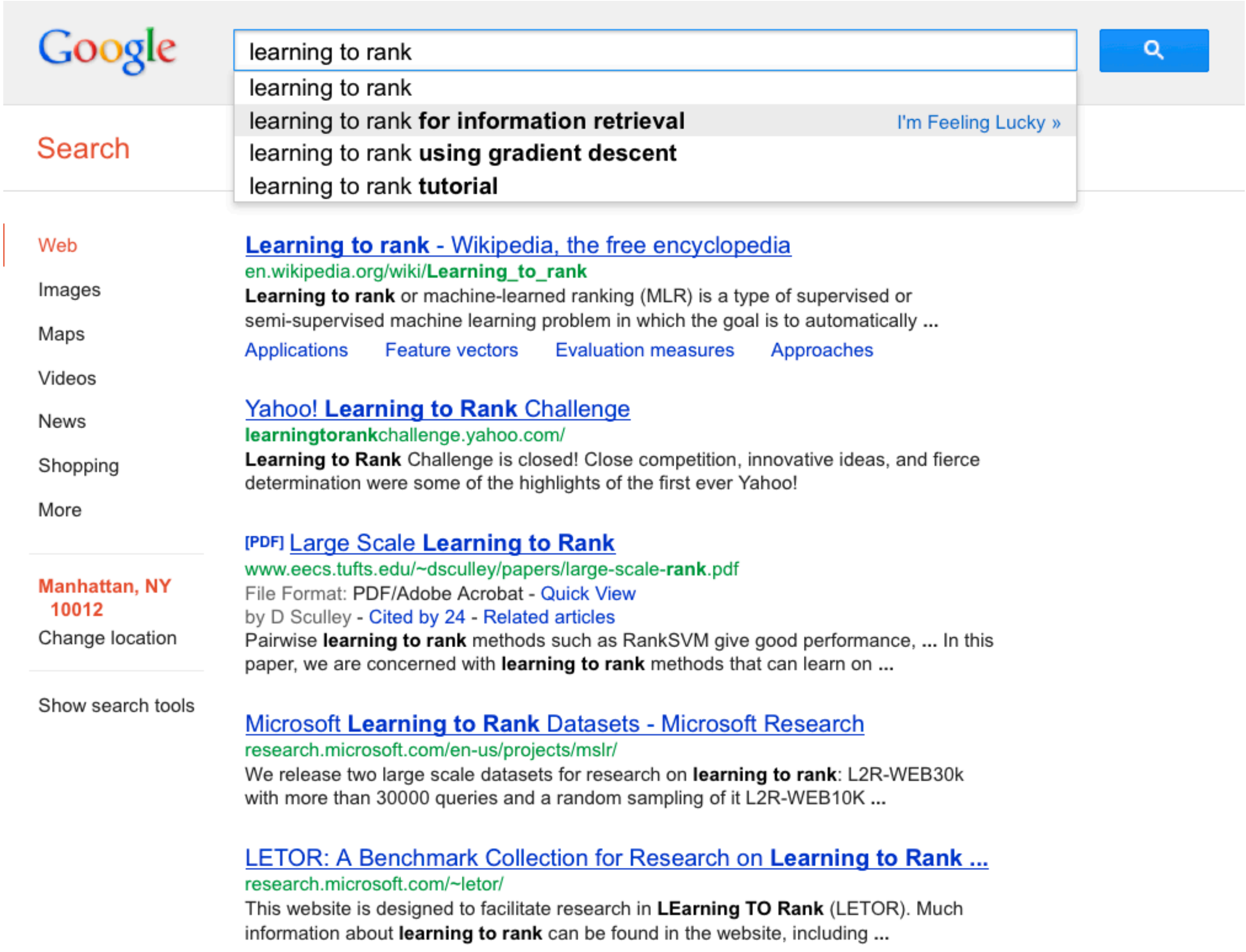
Temperature

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

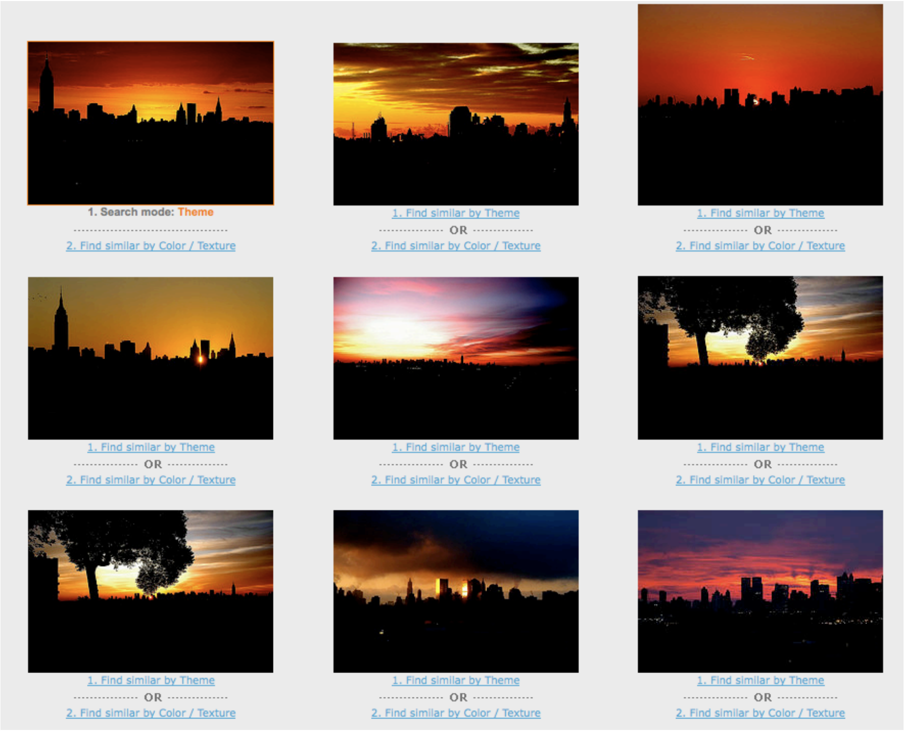
Yahoo! **Learning to Rank** Challenge
learningtorankchallenge.yahoo.com/
Learning to Rank Challenge is closed! Close competition, innovative ideas, and fierce determination were some of the highlights of the first ever Yahoo!

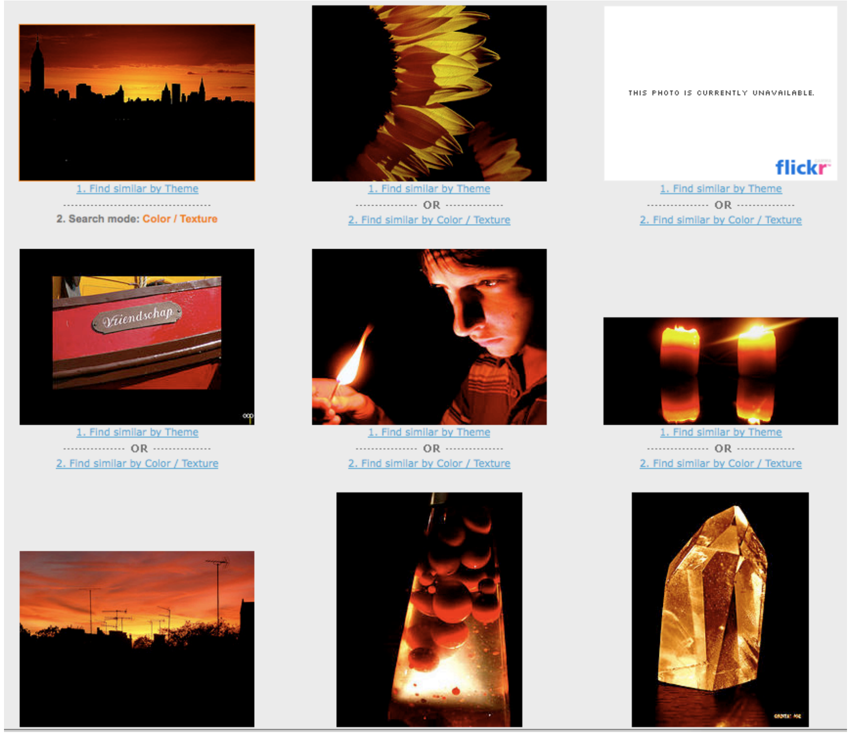
[PDF] Large Scale **Learning to Rank**
www.eecs.tufts.edu/~dsculley/papers/large-scale-**rank**.pdf
File Format: PDF/Adobe Acrobat - Quick View
by D Sculley - Cited by 24 - Related articles
Pairwise **learning to rank** methods such as RankSVM give good performance, ... In this paper, we are concerned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

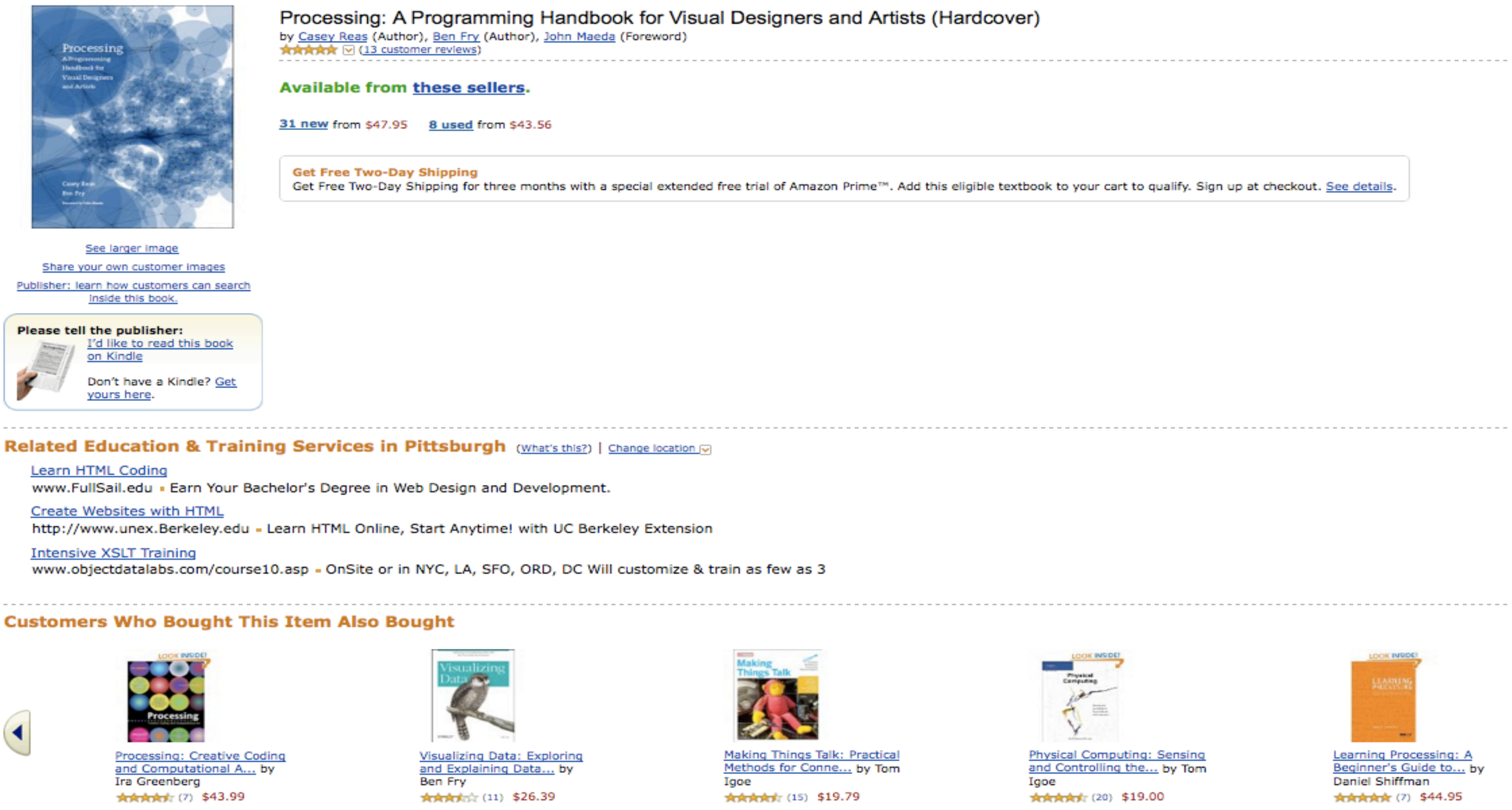

1. Search mode: Theme
2. Find similar by Color / Texture
1. Find similar by Theme
OR
2. Find similar by Color / Texture
1. Find similar by Theme
OR
2. Find similar by Color / Texture
1. Find similar by Theme
OR
2. Find similar by Color / Texture
1. Find similar by Theme
OR
2. Find similar by Color /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



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.

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.

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

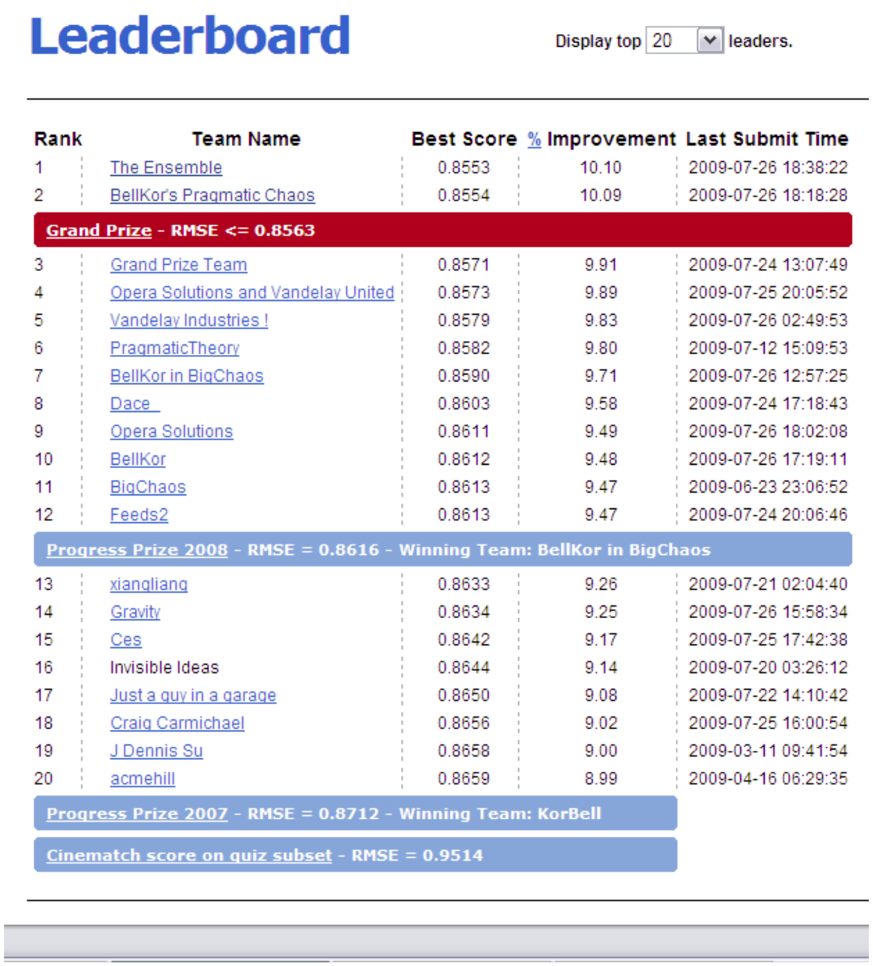
Making Things Talk: Practical Methods for Conne... by Tom Igoe ★★★★★ (15) $19.79

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

Recommendation systems

Machine learning competition with a $1 million prize



Leaderboard

Display top 20 leaders.

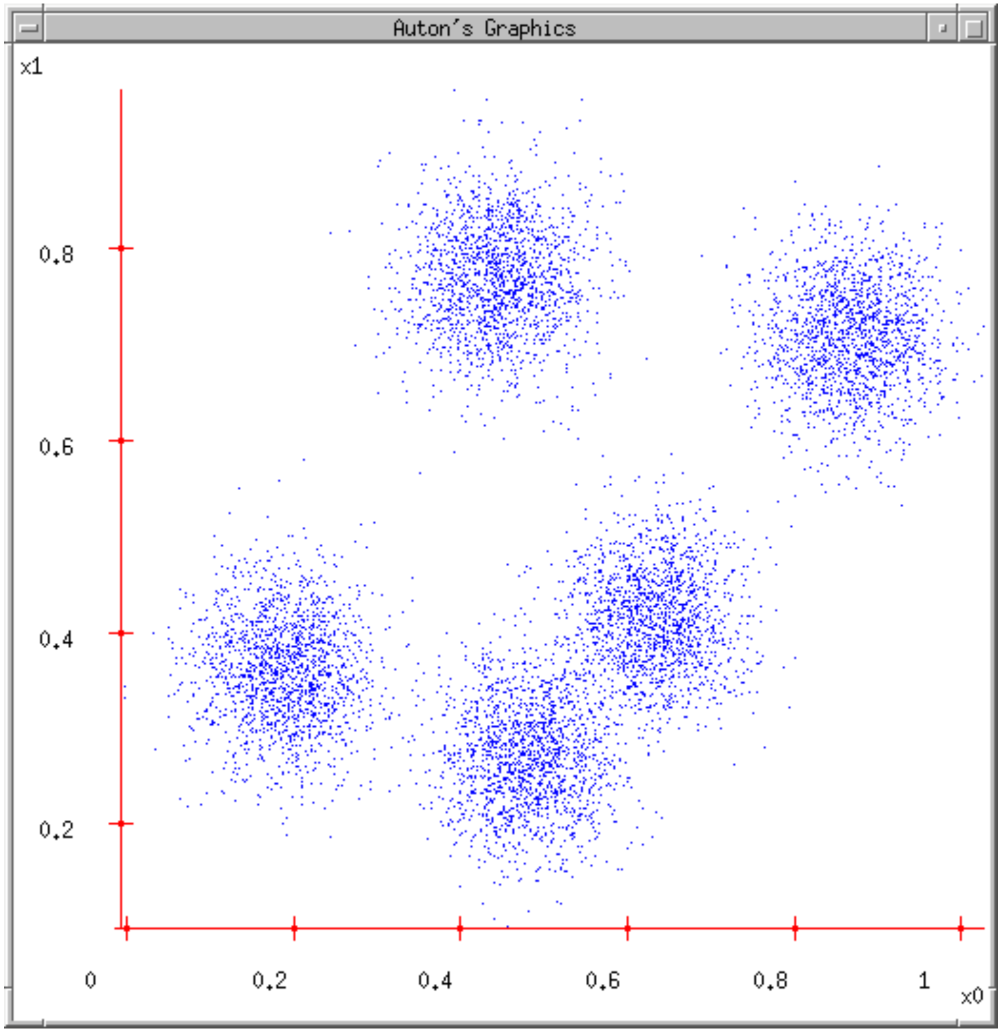
Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	The Ensemble	0.8553	10.10	2009-07-26 18:38:22
2	BellKor's Pragmatic Chaos	0.8554	10.09	2009-07-26 18:18:28
Grand Prize - RMSE <= 0.8563				
3	Grand Prize Team	0.8571	9.91	2009-07-24 13:07:49
4	Opera Solutions and Vandelay United	0.8573	9.89	2009-07-25 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				

	← users →					
↑ movies ↓	1		?	3	5	?
	?	1				2
		4		4	5	?

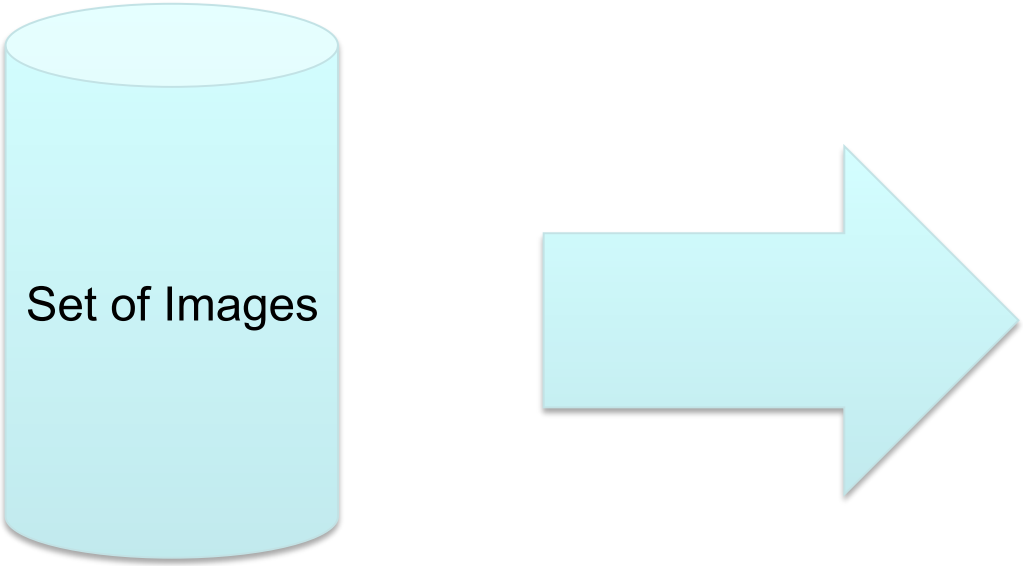
Clustering

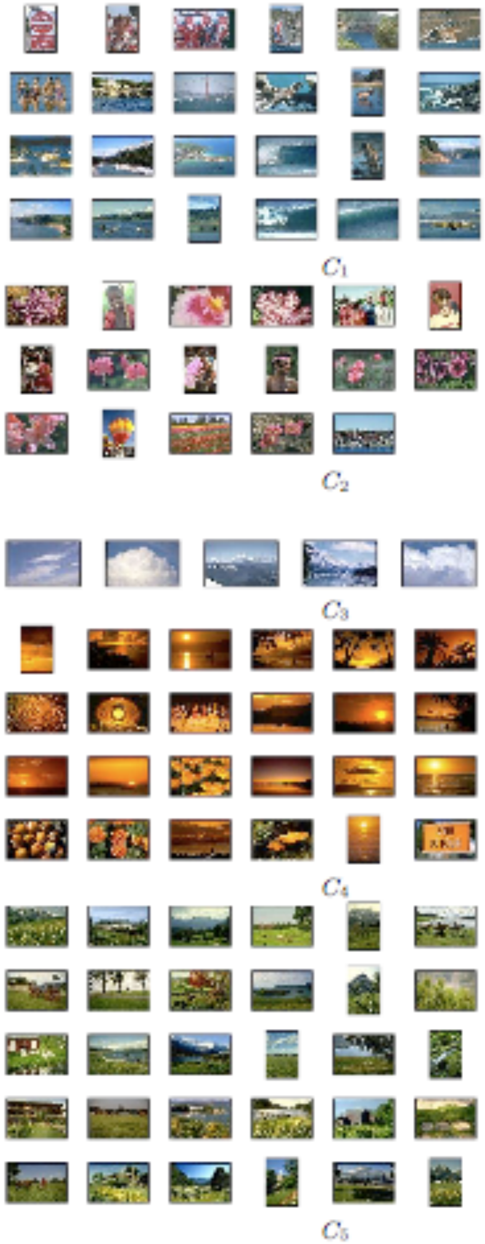
discovering structure in data

Clustering Data: Group similar things



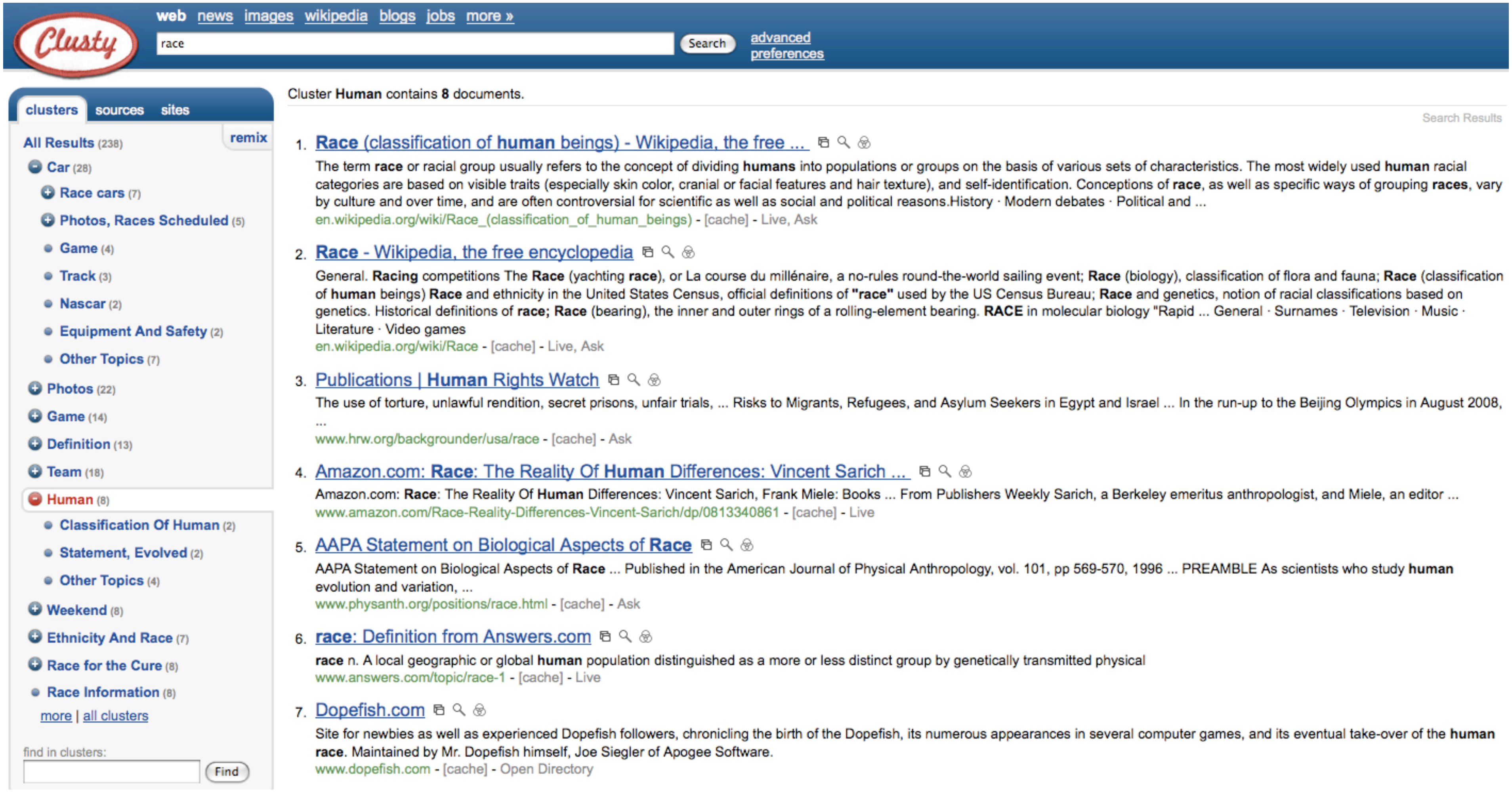
Clustering images





[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.

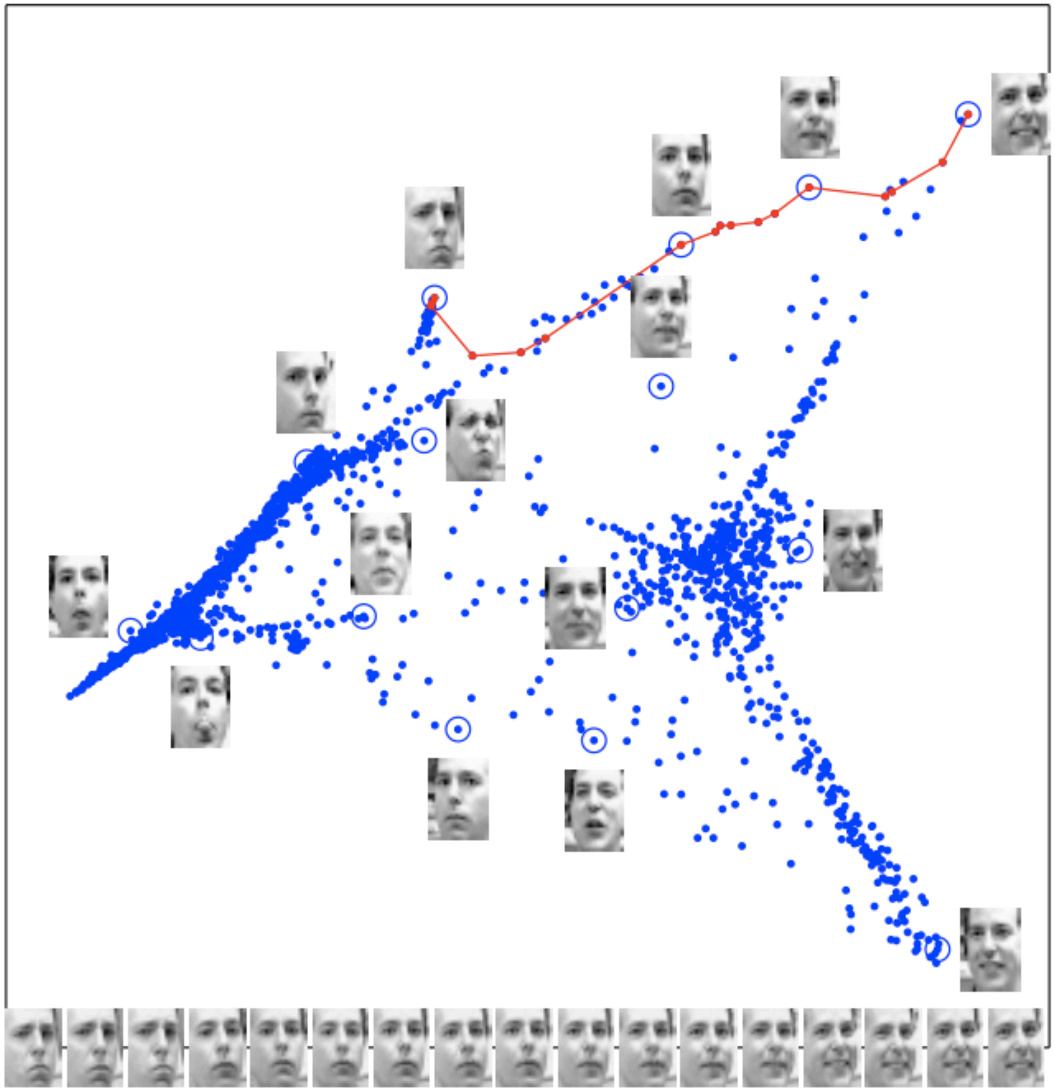
1. **Race** (classification of **human** beings) - Wikipedia, the free ...
   The term **race** or racial group usually refers to the concept of dividing **humans** into populations or groups on the basis of various sets of characteristics. The most widely used **human** racial categories are based on 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

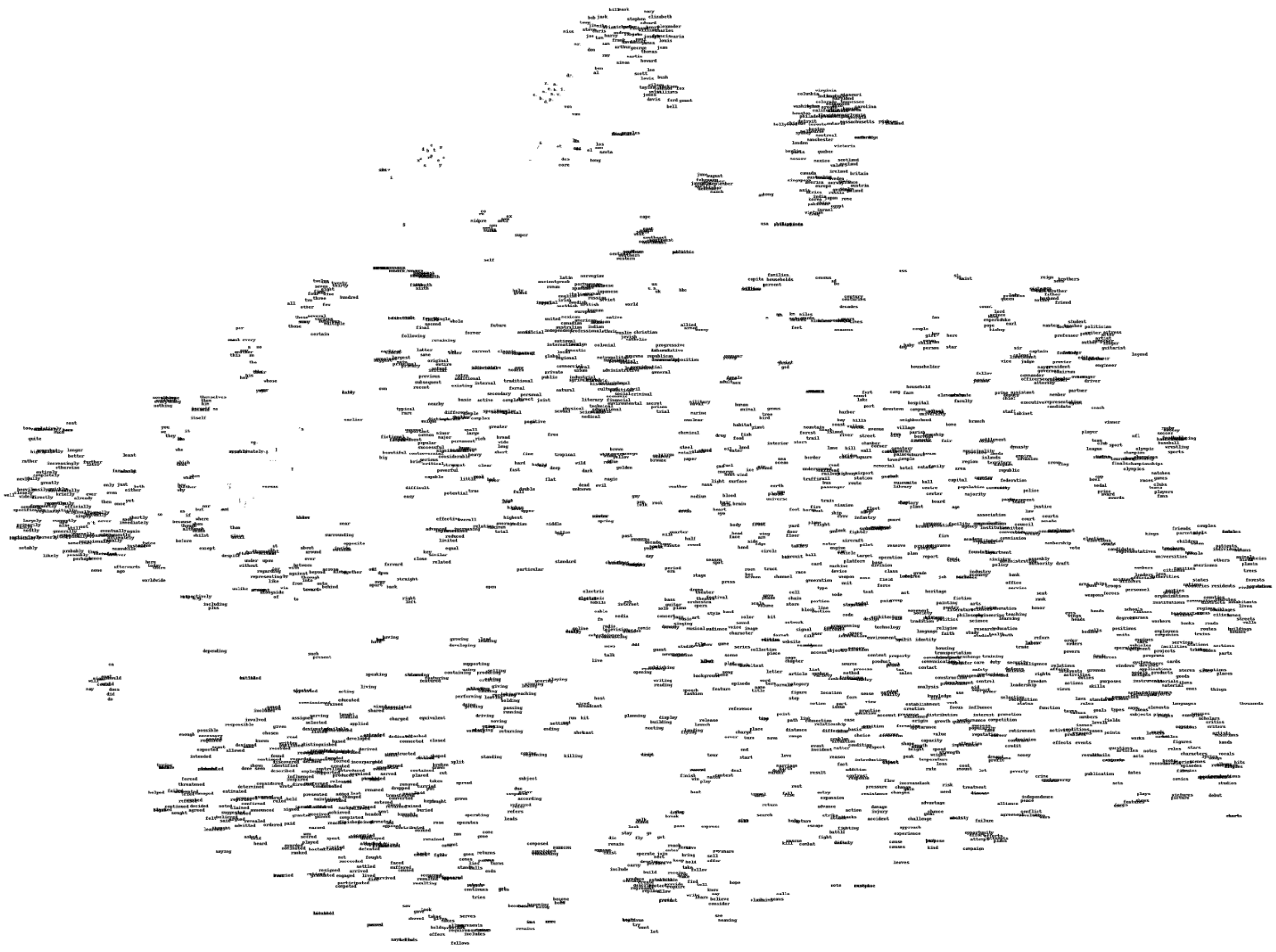
Embedding images

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



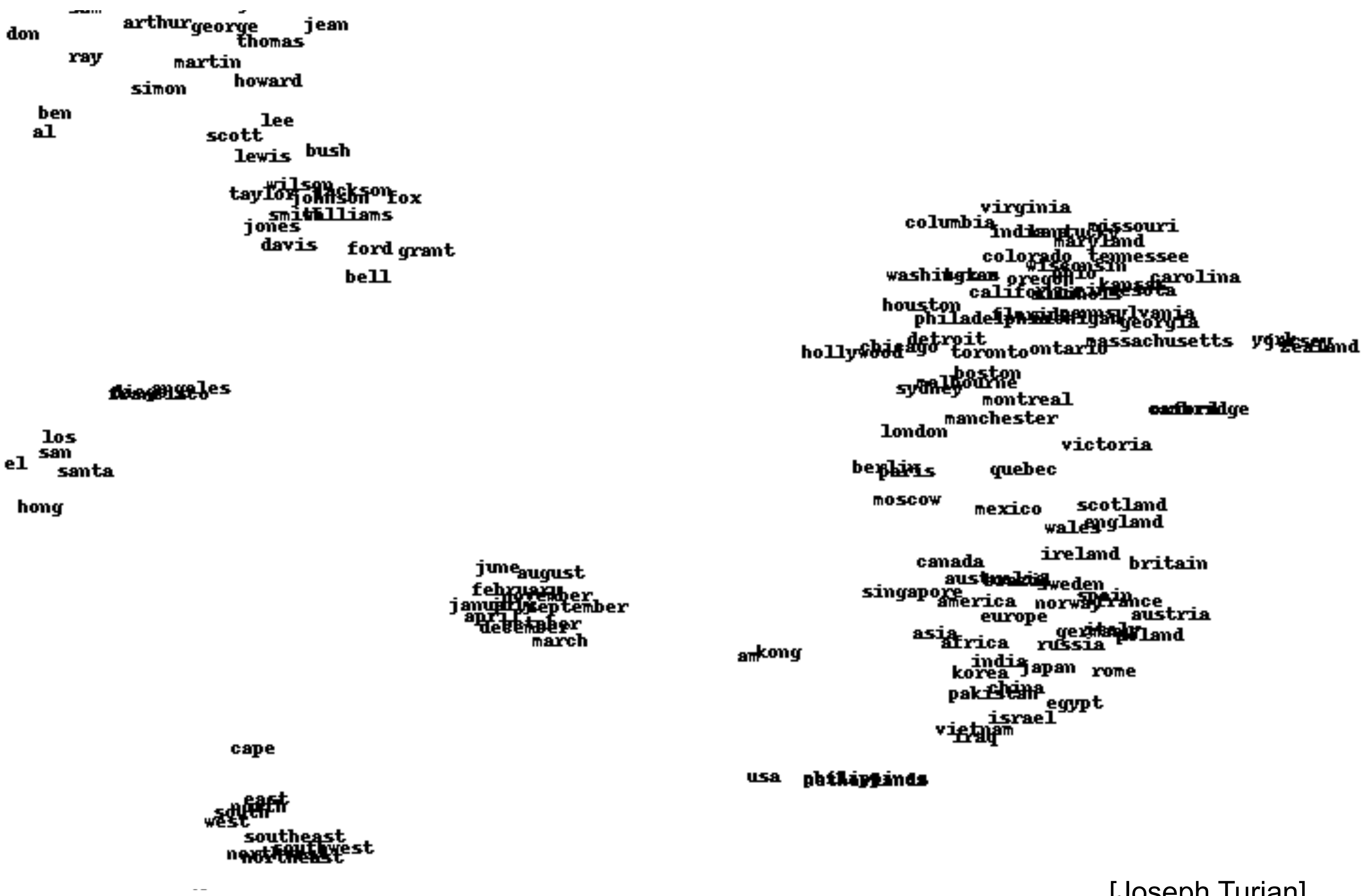
[Saul & Roweis ‘03]

Embedding words



[Joseph Turian]

Embedding words (zoom in)

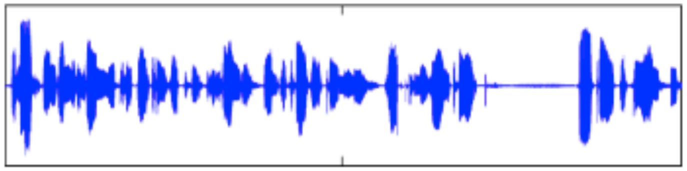


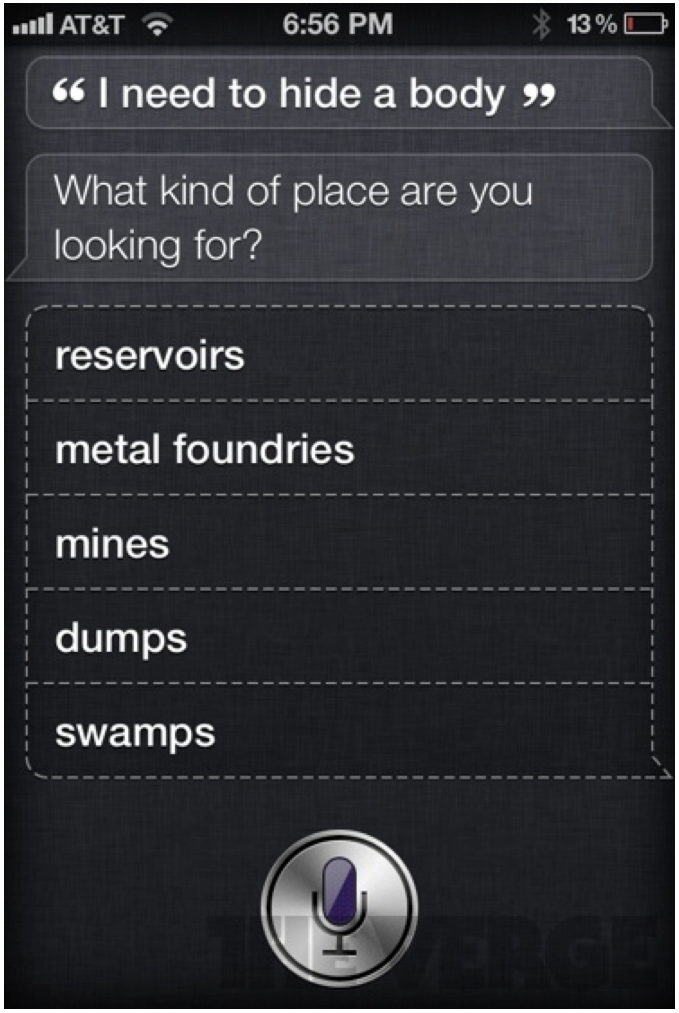
[Joseph Turian]

Structured prediction

from data to discrete classes

Speech recognition





“ I need to hide a body ”

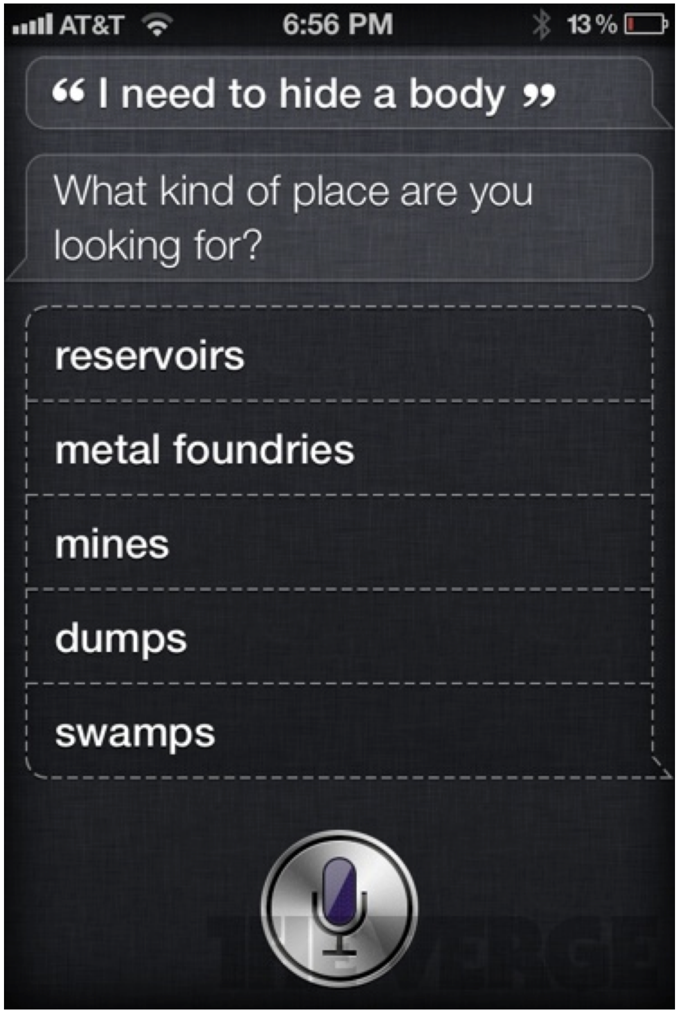
What kind of place are you looking for?

- 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
  - 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 $\{(x_i, y_i) \mid i = 1 \ldots n\}$
- 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)

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 virture of its nature as being utterly confidencial 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

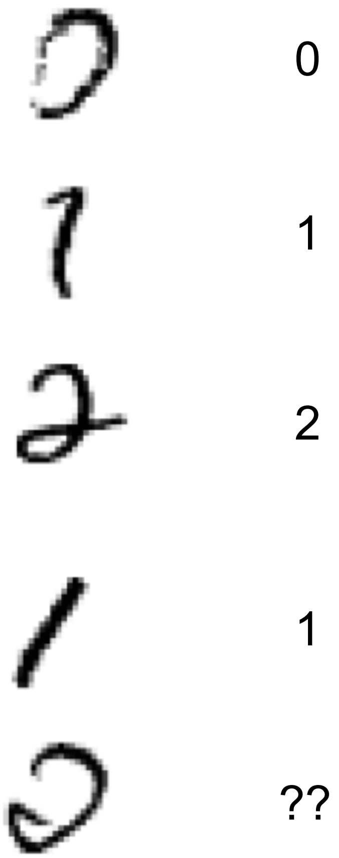
✓

Ok, Iknow 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

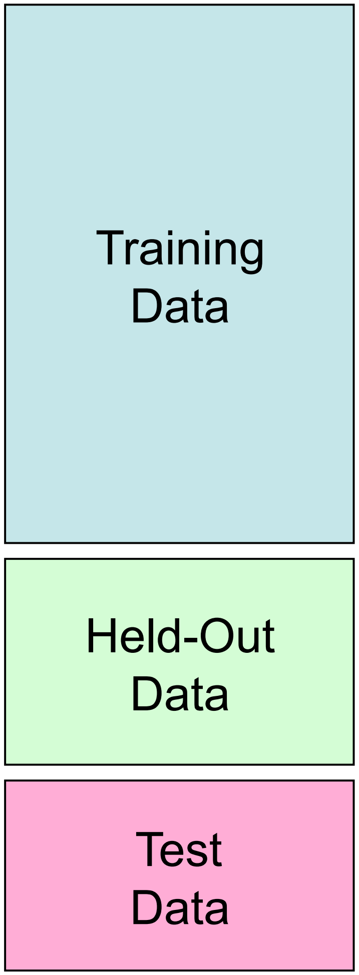
- Input: images / pixel grids
- Output: a digit 0-9
- Setup:
  - Get a large collection of example images, each labeled with a digit
  - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new, future digit images

- 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 or *validation* 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



Training Data

Held-Out Data

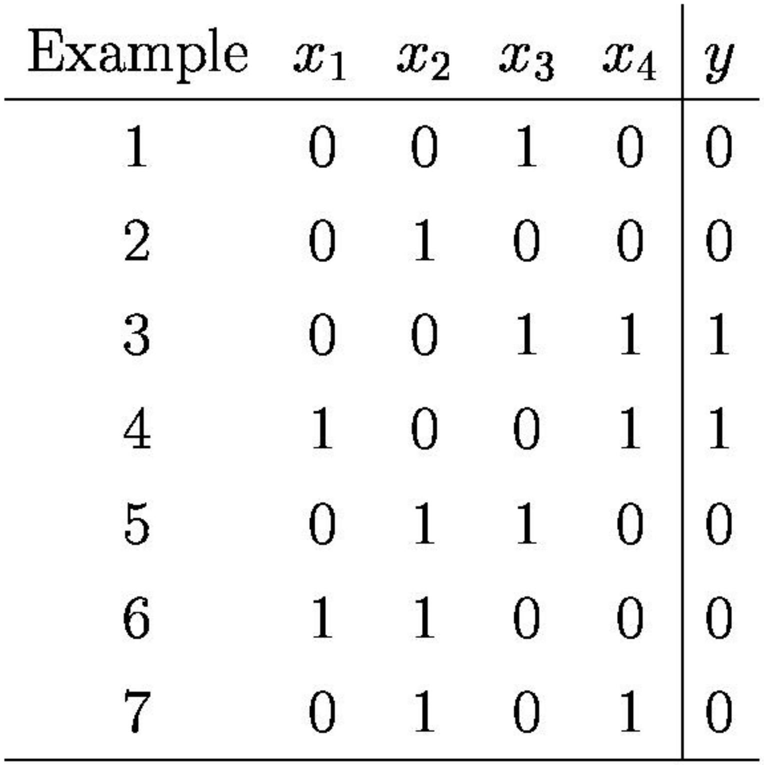
Test Data

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:

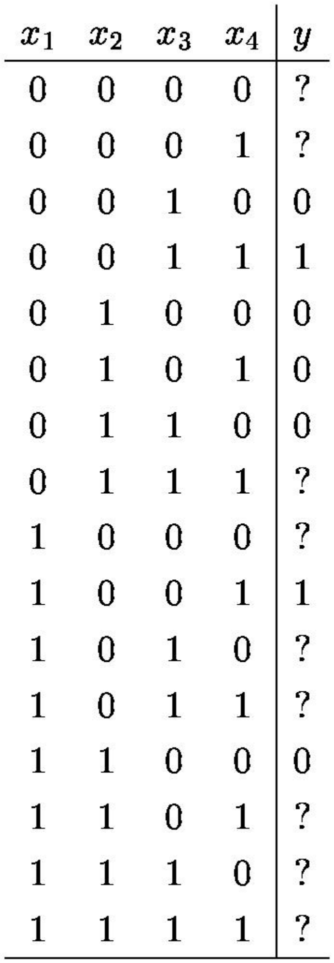


Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Most General Hypothesis Space

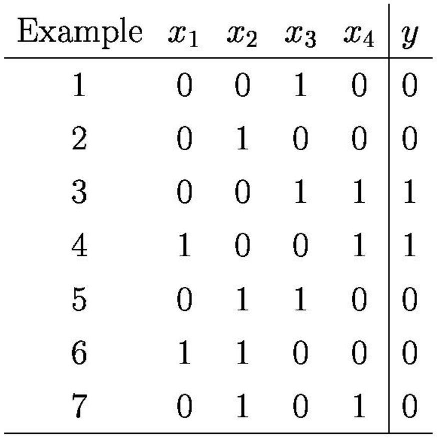
Consider all possible boolean functions over four input features!

- 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

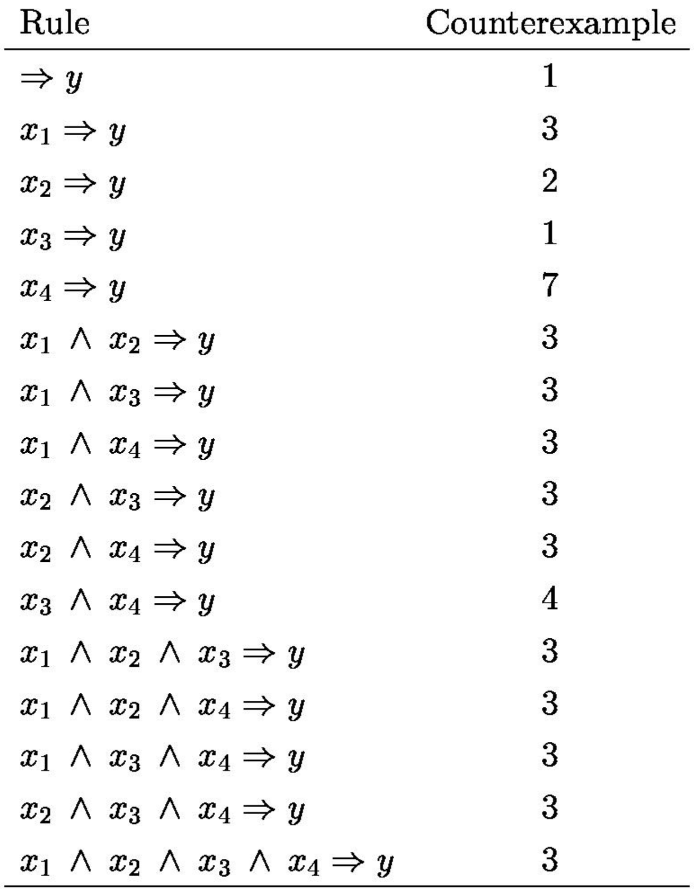
A Restricted Hypothesis Space

Consider all conjunctive boolean functions.

- 16 possible hypotheses

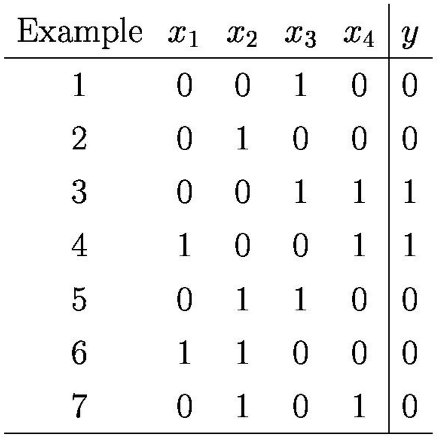
- None are consistent with our dataset

- 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