Introduction to Machine Learning

Inas A. Yassine, PhD

Assoc. Prof. , Systems and Biomedical Engineering Department, Cairo University

inas.yassine@eng.cu.edu.eg

Outlines

- What is machine learning
  - By example
  - Definition
- Medical Informatics
  - Definition
  - Applications
    - Clinical decision making
    - Content based Medical Image Retrieval

What is Machine Learning

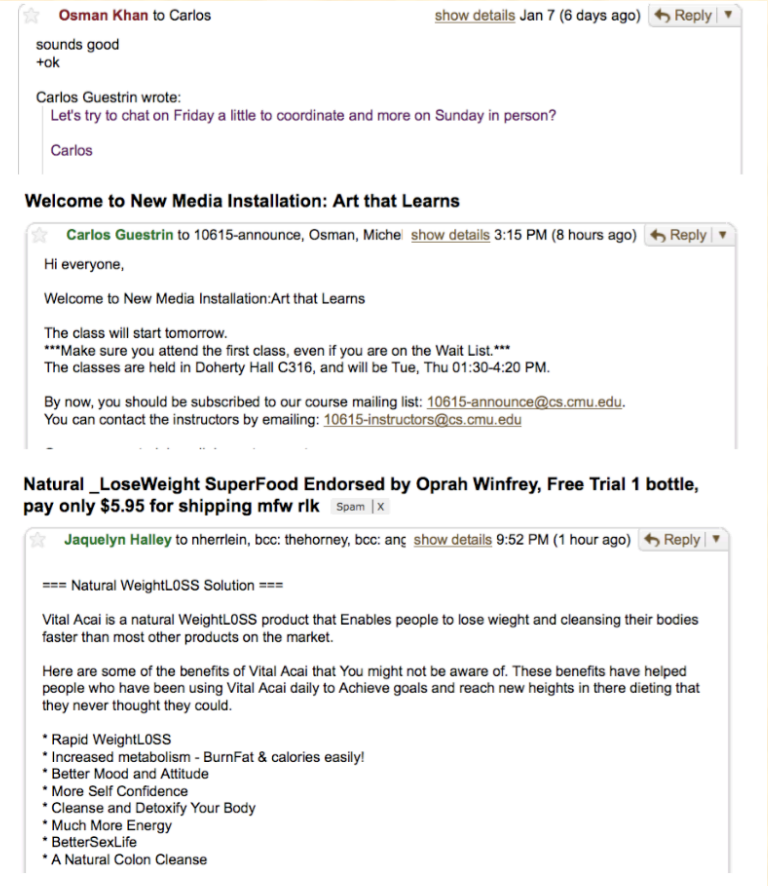
(By Example)

Classification

From Data to Discrete Classes

Spam Filtering

data



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, Miche 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.
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 Spam | x

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

=== 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

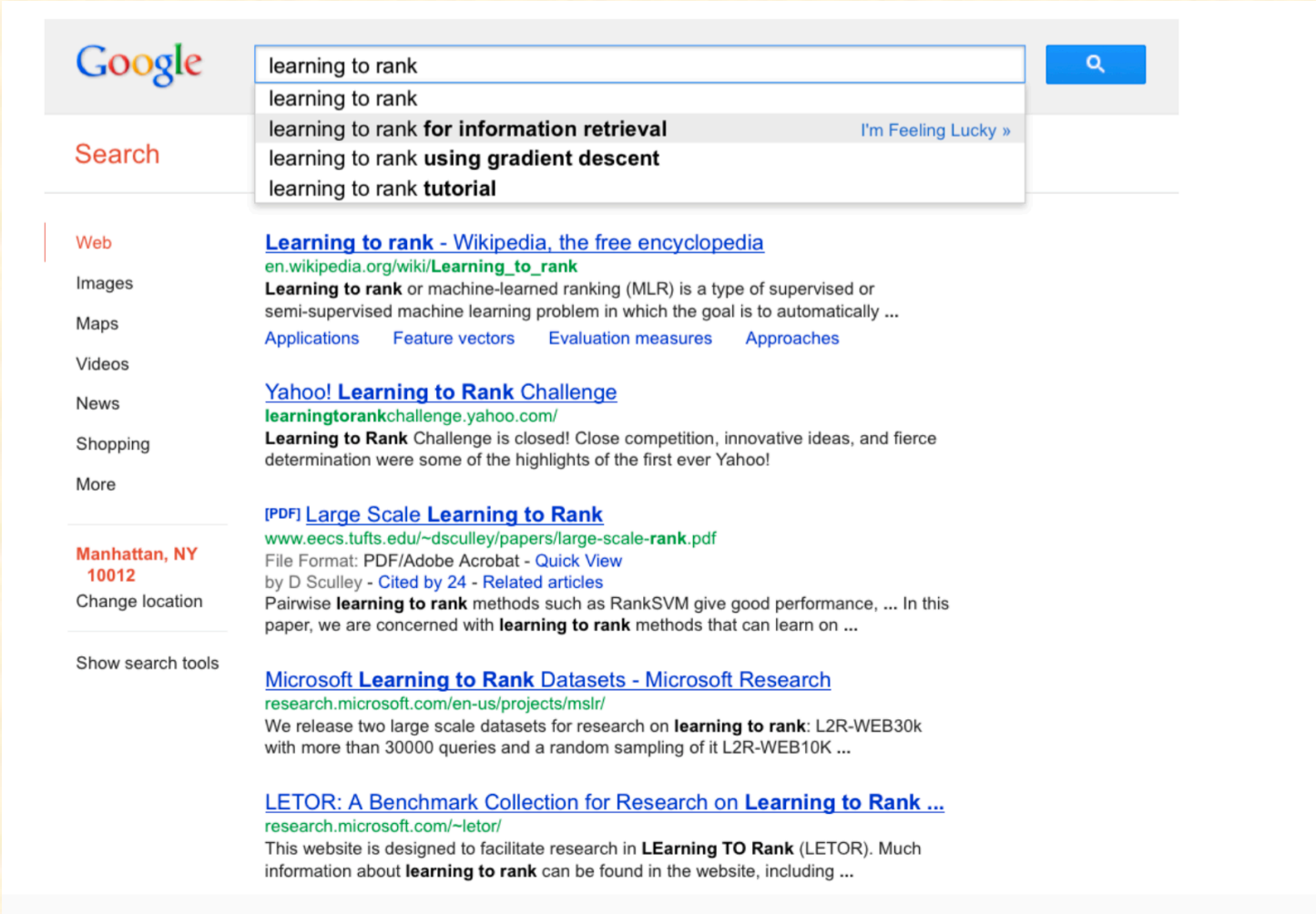
Temperature

72° F

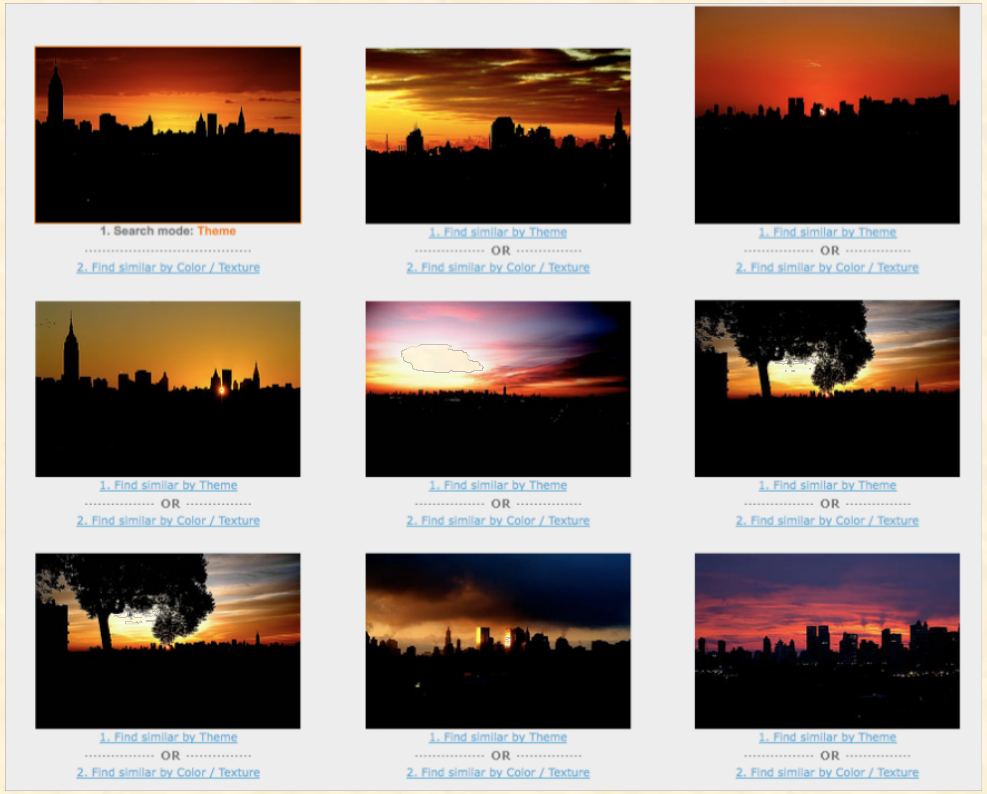
Ranking

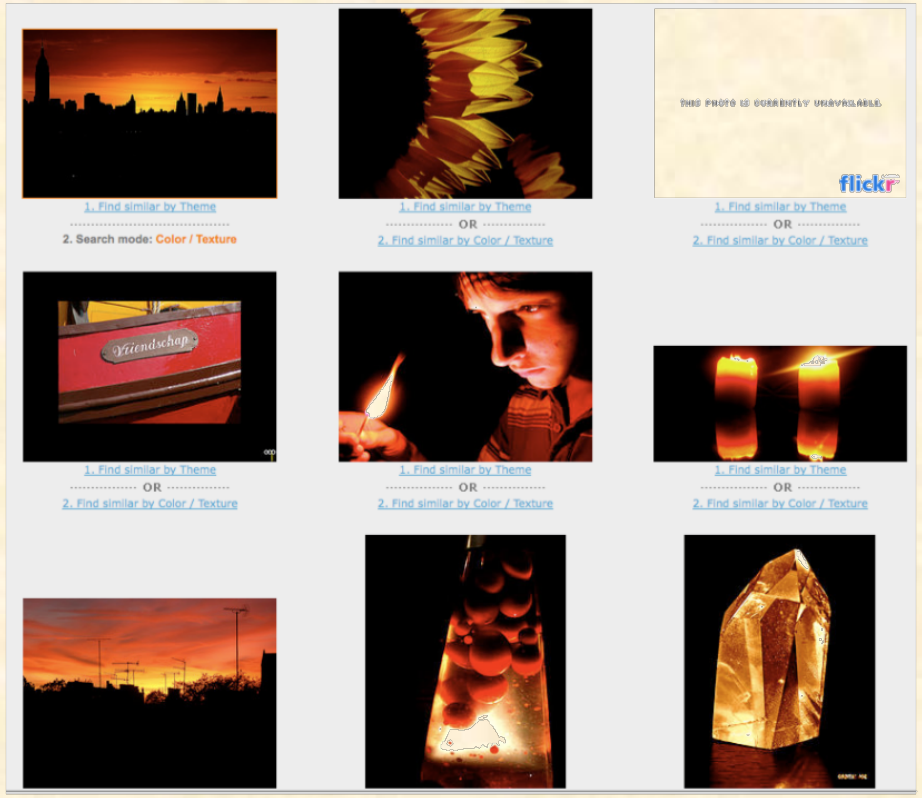
Comparing Items

Web Search


Google
learning to rank
learning to rank
learning to rank for information retrieval
learning to rank using gradient descent
learning to rank tutorial
I'm Feeling Lucky »
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
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 ...

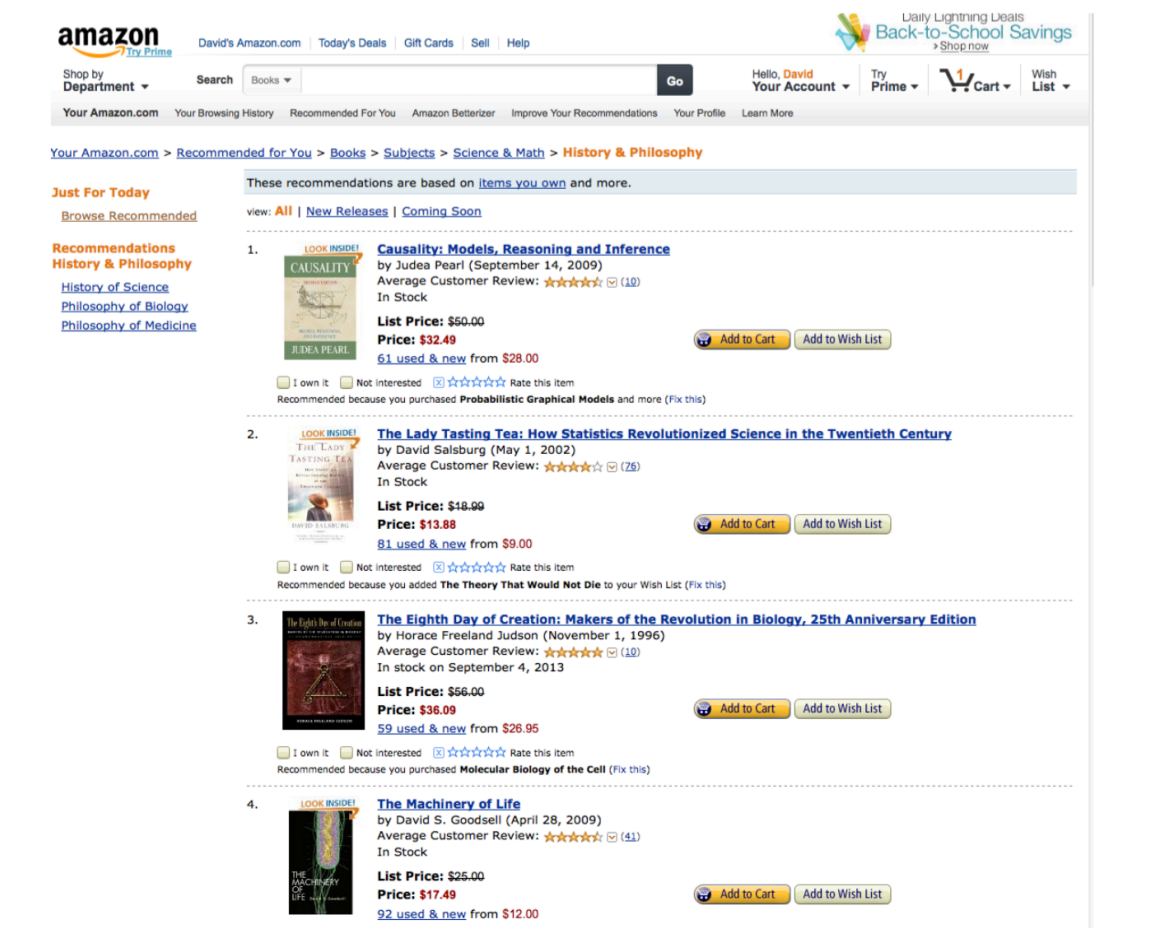
Given Image, Find Similar Images





Collaborative Filtering

Recommendation Systems



Recommendation Systems

Machine learning competition with a $1 million prize

Leaderboard

Display top 20 leaders.

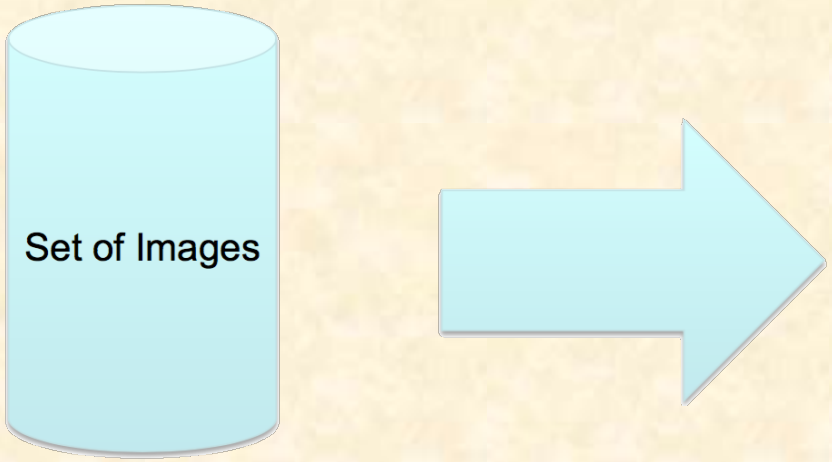
Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	The Ensemble	0.8553	10.10	2009-07-26 18:38:22
2	BellKor's Pragmatic Chaos	0.8554	10.09	2009-07-26 18:18:28
	Grand Prize - RMSE <= 0.8563			
3	Grand Prize Team	0.8571	9.91	2009-07-24 13:07:49
4	Opera Solutions and Vandelay United	0.8573	9.89	2009-07-25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

Clustering

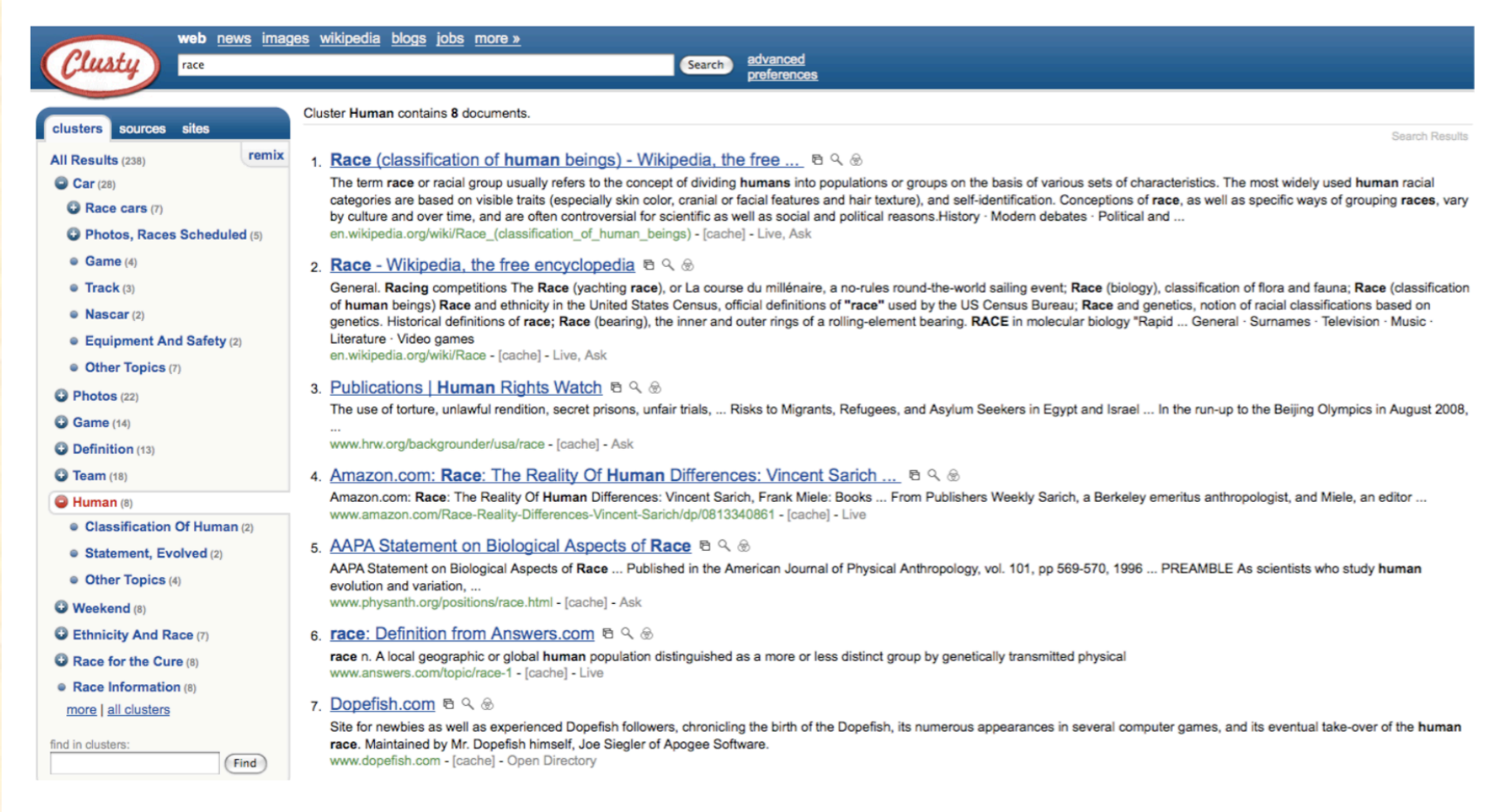
Discovering Structures in Data

Inas A. Yassine

Clustering Images



Clustering Web Search Results



What is Machine Learning

(Definition)

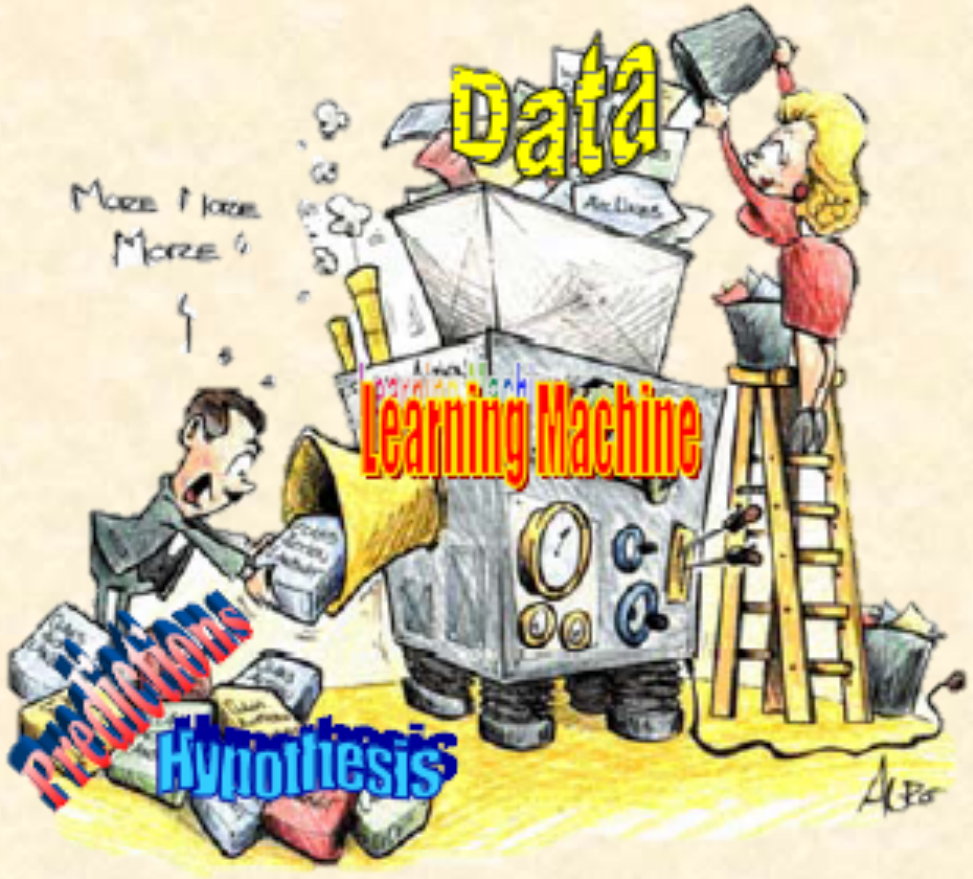
Why “Learn”?

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- There is no need to “learn” to calculate payroll
- It is about seeking a predictive and/or executable understanding of natural/

  artificial subjects, phenomena, or activities from …
- Learning is used when:
  - Human expertise does not exist (navigating on Mars)
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (routing on a computer network)
  - Solution needs to be adapted to particular cases (user biometrics)

What is Machine Learning(ML)?

- Arthur Samuel: “It is a science that gives the computer the ability to learn without the need to write a program“.



What We Talk About When We Talk About“Learning”

- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Example in retail: Customer transactions to consumer behavior:

  People who bought “Da Vinci Code” also bought “The Five People You Meet in Heaven” (www.amazon.com)

- Build a model that is *a good and useful approximation* to the data.

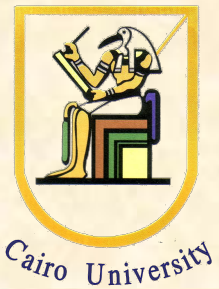
Data Mining

Is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data

Applications:

- Retail: Market basket analysis, Customer relationship management (CRM)
- Finance: Credit scoring, fraud detection
- Manufacturing: Optimization, troubleshooting
- Medicine: Medical diagnosis
- Telecommunications: Quality of service optimization
- Bioinformatics: Motifs, alignment
- Web mining: Search engines
- …

Supervised Learning : find f

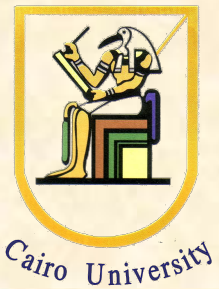


- Given: Training set $\{(x_i, y_i) \mid i = 1 \ldots N\}$
- Find: A good approximation to $f : X \rightarrow Y$

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 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 $f : X \rightarrow Y$
  - $X = \{0,1\}^4$
  - $Y = \{0,1\}$
- Question1:How should we pick the *hypothesis space*, the set of possible functions f ?
- Question2: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

UnSupervised Learning : find *groups*

- Given: Training set $\{(x_i) \mid i = 1 \ldots N\}$
- Find: Learning “what normally happens”
  →No output
- Clustering: Grouping similar instances
  - Customer segmentation in CRM
  - Image compression: Color quantization
  - Bioinformatics: Learning motifs

Reinforcement Learning

- Topics:
  - Policies: what actions(Reward, punishement) should an agent take in a particular situation
  - Utility estimation: how good is a state (→used by policy)
- No supervised output but delayed reward
- Credit assignment problem (what was responsible for the outcome)
- Applications:
  - Game playing
  - Robot in a maze
  - Multiple agents, partial observability, ...

Medical Informatics

Medical informatics

Medical informatics is the application of computers, communications and information technology and systems to all fields of medicine *

- medical care
- medical education
- medical research.

**MF Collen, MEDINFO '80, Tokyo*

What is medical informatics?

Medical informatics is the branch of science concerned with the use of computers and communication technology to acquire, store, analyze, communicate, and display medical information and knowledge to facilitate understanding and improve the accuracy, timeliness, and reliability of decision-making.*

**Warner, Sorenson and Bouhaddou, Knowledge Engineering in Health Informatics, 1997*

Clinical decision making

- Making sound clinical decisions requires:
  - right information, right time, right format
- Clinicians face a surplus of information
  - ***ambiguous, incomplete, or poorly organized***
- Rising tide of information
  - Expanding knowledge sources
    - 40K new biomedical articles per month
    - Publicly accessible online health info
    - Hundreds of pictures per scan for one patient

Clinical decision making: What is the problem?

- Man is an imperfect data processor
  - We are sensitive to the ***quantity*** and ***organization*** of information
- Army officers and pilots commit 'fatal errors' when given too many, too few, or poorly organized data
- The same is true for clinicians who 'watch' for events
- Clinicians are particularly susceptible to errors of ***omission***

Clinical decision making: What is the problem?

- Humans are “non-perfectable” data processors
  - Better performance requires more time to process
  - Irony
    - Clinicians increasingly face productivity expectations
    - Clinicians face increasing administrative tasks

What is medical imaging (MI)?

The study of medical imaging is concerned with the interaction of all forms of radiation with tissue

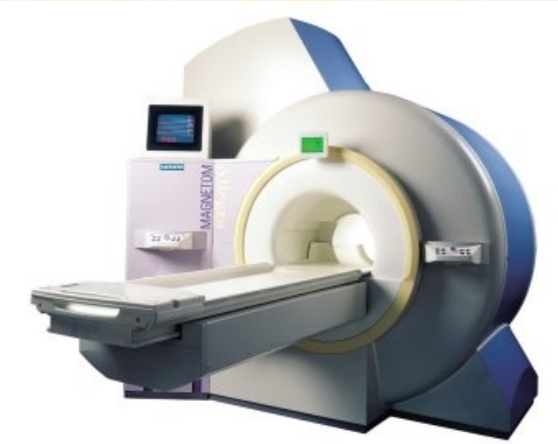
and

the development of appropriate technology to extract clinically useful information (usually displayed in an image format) from observation of this technology.

Sources of Images:

- **Structural/anatomical information (CT, MRI, US) - within each elemental volume, tissue-differentiating properties are measured.**
- **Information about function (PET, SPECT, fMRI).**

The imaging “chain”



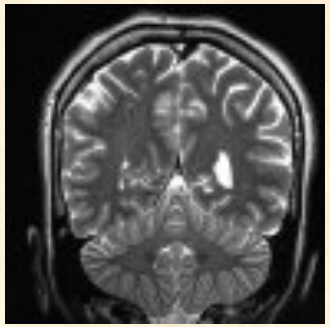
Signal acquisition

Filtering

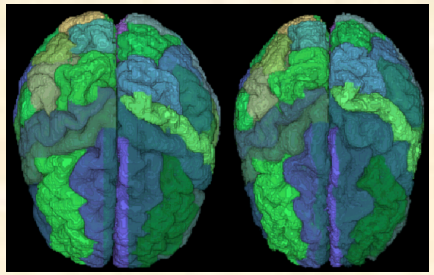
Raw data

“Raw data”

Reconstruction



Processing



Analysis

123……
2346……..
65789……
6578……..

Quantitative output

Computer-Aided Diagnosis

- **Computed Aided Diagnosis (CAD)** is diagnosis made by a radiologist *when the output of computerized image analysis methods has been incorporated into his or her medical decision-making process*.

- CAD may be interpreted broadly to incorporate both
  - the detection of the abnormality task and
  - the classification task: likelihood that the abnormality represents a malignancy

Motivation for CAD systems

The amount of image data acquired during a CT scan is becoming overwhelming for human vision and the overload of image data for interpretation may result in oversight errors.

Computed Aided Diagnosis for:

- **Breast Cancer**
  - Mammograhy Images
- **Lung Cancer**
  - A thoracic CT scan generates about 240 section images for radiologists to interpret.

Content Based Medical Image Retrieval

Inas A. Yassine

Content-based medical image retrieval (CBMS) systems

Definition of Content-based Image Retrieval:

Content-based image retrieval is a technique for retrieving images on the basis of automatically derived image features such as texture and shape.

Applications of Content-based Image Retrieval:

- **Teaching**
- **Research**
- **Diagnosis**
- **PACS and Electronic Patient Records**

Diagram of a CBIR

Image Database

Feature Extraction

Image Features
[D1, D2,…Dn]

Query Image

Similarity Retrieval

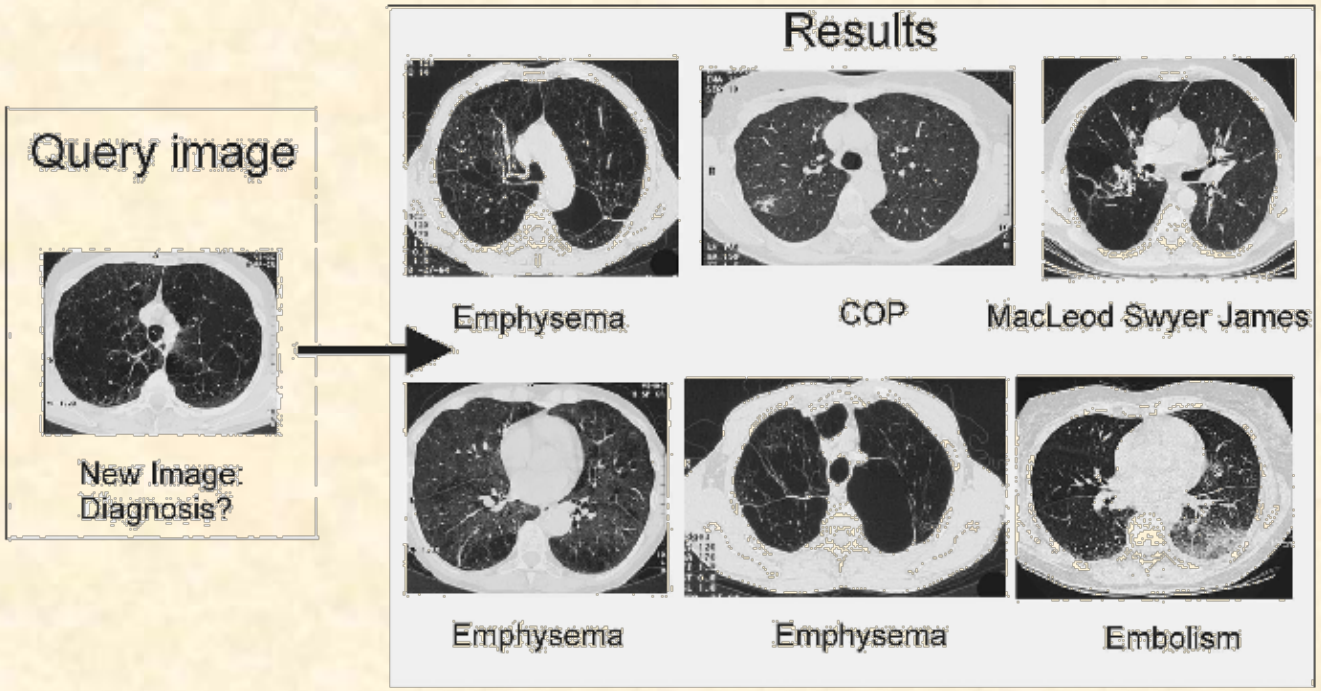
Feedback Algorithm

User Evaluation

http://viper.unige.ch/~muellerh/demoCLEFmed/index.php

CBIR as a Diagnosis Aid

An image retrieval system can help when the diagnosis depends strongly on direct visual properties of images in the context of evidence-based medicine or case-based reasoning



CBIR as a Teaching Tool

An image retrieval system will allow students/teachers to browse available data themselves in an easy and straightforward fashion by clicking on "show me similar images".

Advantages:

- stimulate self-learning and a comparison of similar cases
- find optimal cases for teaching

Teaching files:

- Casimage: http://www.casimage.com
- myPACS: http://www.mypacs.net

CBIR as a Research Tool

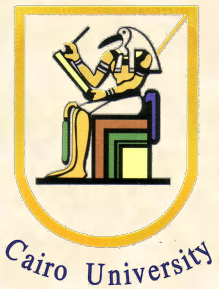


Image retrieval systems can be used:

- to complement text-based retrieval methods
- for visual knowledge management where by the images and associated textual data can be analyzed together
    - multimedia data mining can be applied to learn the unknown links between visual features and diagnosis or other patient information
- for quality control to find images that might have been misclassified

In Summary …

- Machine Learning is a Knowledge Discovery Process
- 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 due to
  - Big data
  - Improved machine learning algorithms – Faster computers
  - Good open-source software
- Data Mining Tasks
  - classification, clustering, regression , …

Course Content

- Introduction: Machine Learning for health informatics
- Linear prediction: Regression
- Learning strategies for basic classification algorithms: Linear classification, logistic regression
- Probabilistic models (Bayes Theory, Decision trees)
- Learning strategies for more advanced classification algorithms (SVM, …)
- Dimensionality Reduction (Principal component analysis)
- Subspace Clustering
- Neural Networks
- Reinforcement Learning
- Recommender Systems
- Overfitting, underfitting, and the generalization gap
- Machine Learning Algorithms testing and evaluation

Grading Criteria

Item	Course Grade Contribution
Final Exam	70%
Final Project	8%
Assignments/ Homework	15%
Midterm Exam	7%

Cheating Description

- **What is cheating?**
  - Sharing code: by copying, retyping, **looking at**, or supplying a file
  - Describing: verbal description of code from one person to another.
  - Coaching: helping your friend to write a line by line code
  - Searching the Web for solutions
  - Copying code from a previous course or online solution
  - You are only allowed to use code we supply
- What is NOT cheating?
  - Explaining how to use systems or tools
  - Helping others with high-level design issues
- Penalty for cheating
  - Zero in Assignments/ Exam

In Class/ Assignment General Code

- Electronic communications: forbidden
  - No email, instant messaging, cell phone calls, etc
- Presence in lectures: highly recommended
- No recordings of ANY KIND
- No side talks

.

Thank You ...