Machine Learning and Data Analytics
ME 5013- Fall 2019

Lecture 01
Welcome!

UTSA
The University of Texas at San Antonio™

Adel Alaeddini, PhD
Associate Professor of Mechanical Engineering
Advanced Data Engineering Lab
adel.alaeddini@utsa.edu

Disclosure: some of the slides are adopted from Dr. Andrew Ng lecture notes on "Machine Learning" at Stanford university

Preliminaries

- **General Information**
  - Class Syllabus and Topics
  - Text and Other References
  - Grading
- **Introduction to Machine Learning and Data Analytics**
  - Definition
  - Examples
  - Machine Learning Algorithms
  - Hands-on (Python)

Class Team

Instructor

Adel Alaeddini, PhD
Associate Professor of Mechanical Engineering
The University of Texas at San Antonio
One UTSA Circle
San Antonio, TX 78249
Email: adel.alaeddini@utsa.edu
Website: https://www.alaeddini.org/
Tel: 210.458.8747

Teaching Assistant

Syed Hasib Akhter Faruqui
Ph.D. Candidate of Mechanical Engineering
The University of Texas at San Antonio
One UTSA Circle
San Antonio, TX 78249
Email: syed-hasib-akhter.faruqui@utsa.edu

General Information

Class information

Class type	Lecture
Instructor	Dr. Adel Alaeddini
Instructor's Office, contact info	EB 3.04.48 adel.alaeddini@utsa.edu (210) 458-8747
Instructor's Office Hours	TR 2:30-3:30 pm or by appointment
Meeting Time/Room	TR. 1:00-2:15pm./ MS 2.02.20

Grading

Class Evaluations	1%
Attendance	5%
In-class participation	10%
Homework (Typically 8)	10%
Mid-term (Typically 2)	15%
Quizzes (Typically 4)	10%
Project	25%
Final Examination	25%
Total	101%

Grading

- **Lectures:**
  - Students are expected to have the textbooks throughout the semester.
  - The exams may have open books & notes sections and exchange of book during the exams is not allowed.
  - Lecture notes, which include PowerPoint slides and photocopies from reference books, will be handed out in class and/or made available on-line
  - The PowerPoint slides have gaps that need to be filled out by students during lectures.
  - In-class discussions will be encouraged through various active learning exercises; therefore attendance is a key to success.

- **In-Class Participation:**
  - The course is based on active learning exercises and interactions.
  - Bi-Weekly assignments will be given and they must be completed prior to the Tuesday lecture of the following week.
  - I may ask unannounced questions during lectures from the weekly assignments, as well as the lecture of the day and the randomly selected student/s or all students and they will have to respond in 30-45 seconds. You are strongly recommended to study the material before class and stay awake during lectures.
  - Attendance plus in-class participation are 15% of the overall grade.

Grading

- **Attendance:**
  - Each student is expected to attend the lectures unless he/she has a valid excuse that has been communicated to the instructor in a timely manner.
  - If you miss 6 sessions, then you have to drop the course or directly fail
  - Tardiness beyond the first 5 minutes of the class will not be tolerated.
  - Leaving the class early is also not acceptable. Students need to talk with the instructor prior to class if they have a valid excuse.
  - **During class please turn off your cell phones, ipdas, and/or any other device that is not related to the subject matter**.

- **Homework:**
  - Individual and/or group-based homework assignments may be given
  - For group-based assignments each group will turn in one answer sheet.
  - If a group member does not contribute, then do not put his/her name on the answer sheet
  - There will be a 10% reduction per day for late submissions, effectively immediately past the time of the submission (such as Tuesday, 1:00 pm… after 1:00 pm, 10% penalty is in effect until 1:00pm of the next day).

- **Projects:**
  - There will be a written project.
  - Late submission of the project will result in losing the whole project points

Grading

- **Exams:**
  - Two midterm exams and a comprehensive final exam will be given.
  - Exams may consist of two parts: Part 1 Multiple Choice (Closed Books & Notes) and Part 2 Problems (Open Books and Notes).
  - The midterm exams will not be comprehensive.
  - Students will be responsible for all lecture material that has been covered for the examinations regardless if an absence was or was not excused.
  - Make-up exam is only given in case of a medical emergency.
  - Contacting the instructor, if possible, before missing the exam will definitely help in working out a solution for the student. The format of the make-up exams could completely be different than the regular tests.

- **Evaluations**
  - In order to facilitate an effective learning environment, you may be asked to anonymously evaluate the course content and my teaching style during the semester. I use feedback provided by students in course evaluations to improve my teaching and that is also used by the university as one factor in evaluating my class effectiveness. There will be 1% extra credit for completion of course surveys.
  - In addition, since some activities will require group work, students may also be asked to evaluate each other (i.e., peer evaluation) in their groups in order for me to give fair grades to individuals.

Grading

- **"Unauthorized" Collaboration:**
  - Although teamwork is encouraged for studying the course material and for some assignments, each student is required to do his/her own original thinking during exams. Do not tempt your neighbor into cheating the final letter grades.

- **Plagiarism:**
  - Copying or using someone else's work without giving credit (citation) intentionally or unintentionally is a violation of UTSA's Student Code of Conduct.
  - Academic dishonesty will be dealt with the University regulations.

References

Textbook(s) and/or required material:

- Pang-Ning Tan, Michael Steinbach, Anuj Karpante, Vipin Kumar (2019), Introduction to Data Mining. *Second Edition*.

Reference books

- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013), An Introduction to Statistical Learning: with Applications in R, Springer Texts in Statistics. (*Available online*)
- T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning. Second Edition, Springer, 2009. (*Available online*)

Topics Coverage

Topics Covered:

1. Parametric regression: Linear/polynomial regression
2. Regularization and cross validation
3. Nonparametric regression: Kernel regression, Gaussian processes
4. Classification basics: Loss functions, logistic regression, Decision trees
5. Support vector machines, convex optimization
6. Ensemble methods: Boosting, bagging, random forest
7. Dimension reduction: Principal component analysis
8. Clustering: K-mean, mixture models, EM algorithms
9. Graphical Models: Kalman filter, Bayesian networks
10. Deep learning: Neural networks, convolutional neural networks, recurrent neural networks
11. Reinforcement Learning: Markov Decision Process, Reinforcement Learning, Deep Reinforcement Learning

Software:

Python

*Chapters Coverage depend on the class progress

Motivating Examples

*Chapters Coverage depend on the class progress

Introduction

Machine Learning / Statistical Learning/ Data Mining

- Grew out of work in AI
- New capability for computers

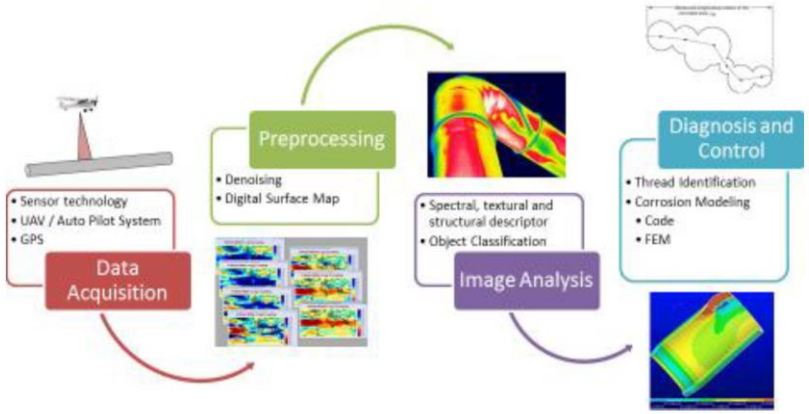
Applications:

- Database mining
  Large datasets from growth of automation/web.
  E.g., Web click data, medical records, biology, engineering
- Applications can't program by hand.
  E.g., Autonomous helicopter, handwriting recognition, most of Natural Language Processing (NLP), Computer Vision.
- Self-customizing programs
  E.g., Amazon, Netflix product recommendations
- Understanding human learning (brain, real AI).

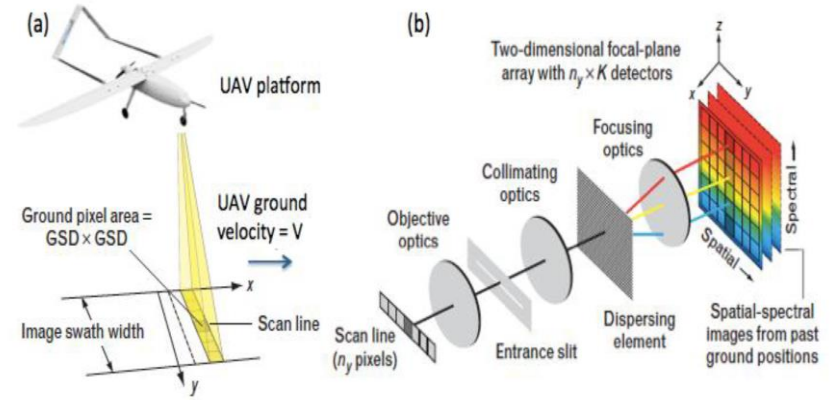
UAV based Pipeline Monitoring System

(1)

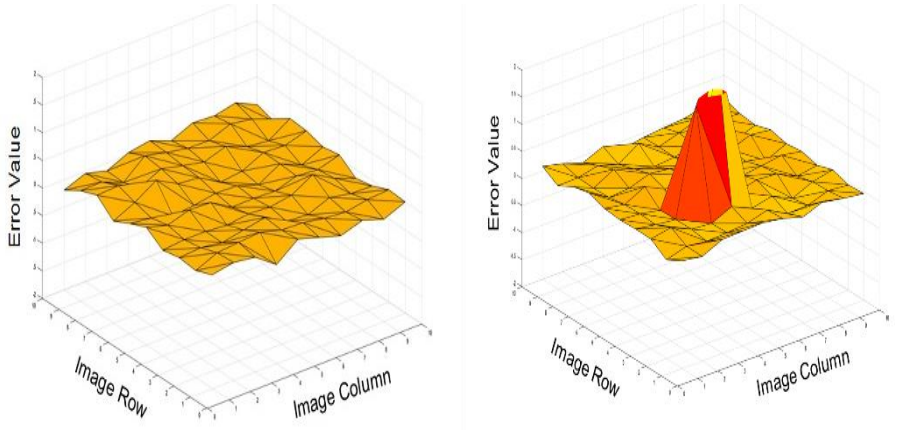
(2)


Preprocessing
• Denoising
• Digital Surface Map
• Sensor technology
• UAV / Auto Pilot System
• GPS
Data Acquisition
• Spectral, textural and structural descriptor
• Object Classification
Image Analysis
Diagnosis and Control
• Thread Identification
• Corrosion Modeling
• Code
• FEM

(3)

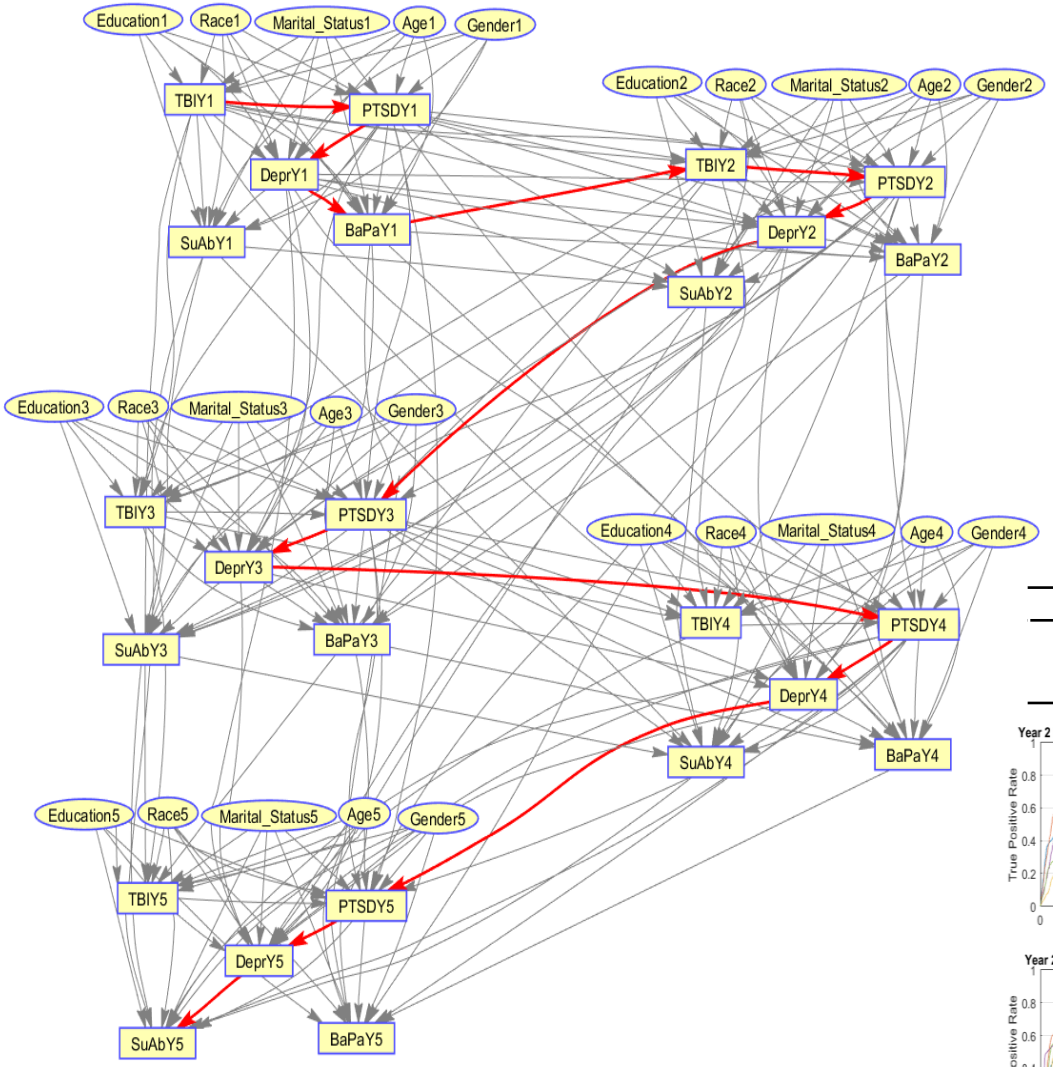

(a)
UAV platform
Ground pixel area = GSD × GSD
UAV ground velocity = V
Image swath width
Scan line
(b)
Two-dimensional focal-plane array with $n_y \times K$ detectors
Focusing optics
Collimating optics
Objective optics
Spectral
Spatial
Scan line (n_y pixels)
Entrance slit
Dispersing element
Spatial-spectral images from past ground positions

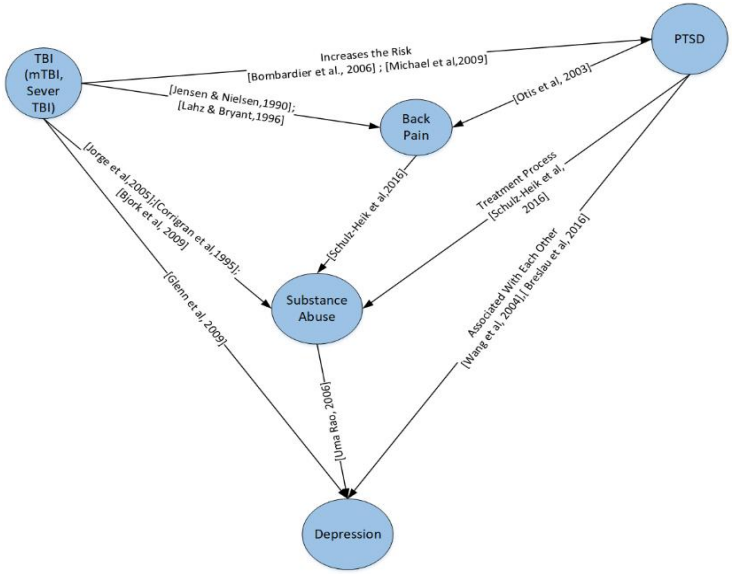
(4)


Error Value
Image Row
Image Column
Error Value
Image Row
Image Column

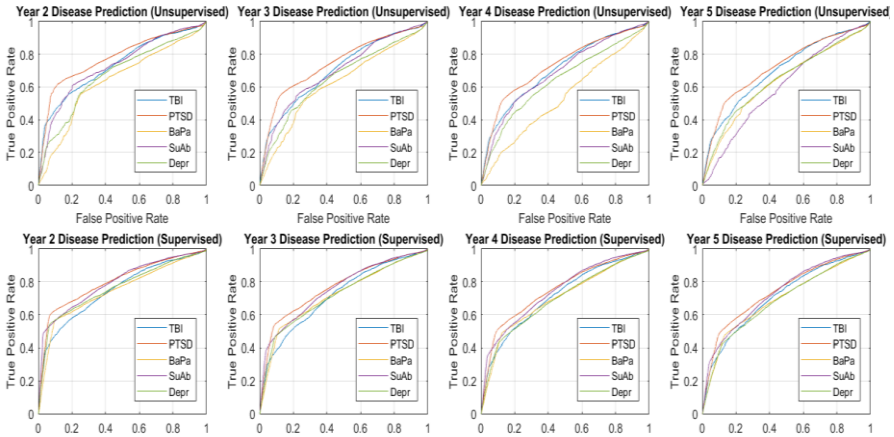
Predictive Modeling of Multiple Chronic Conditions Development

Unsupervised Multi-Level Temporal Bayesian Network (U-MTBN)





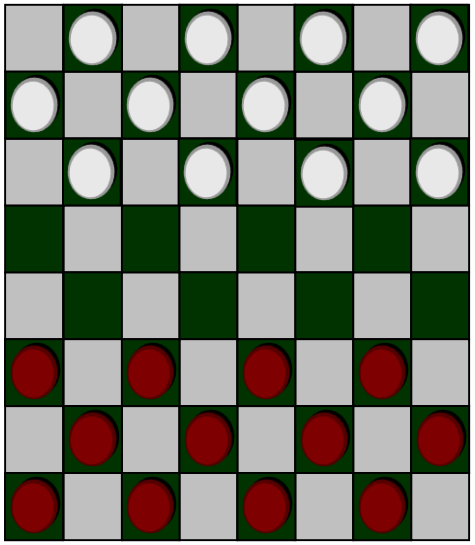
	Comparison method	Supervised
Unsupervised	Cosine similarity	80%
	Correlation coefficient	80%



Faruqui, S. H., Alaeddini, A., Jaramillo, C. A., & Pugh, M. J., Mining Major Transitions of Chronic Conditions in Patients with Multiple Chronic Conditions, *PLOS ONE*, 13.7 (2018):74-87.

Machine Learning definition

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.
- Tom Mitchell (1998). Well-posed Learning Problem: A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.



Exercise

"A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?

(a) Classifying emails as spam or not spam.
(b) Watching you label emails as spam or not spam.
(c) The number (or fraction) of emails correctly classified as spam/not spam.
(d) None of the above—this is not a machine learning problem.

Introduction

Statistics

- Is the study of the collection, analysis, interpretation, presentation, and organization of data.

Artificial Intelligence (AI)

- AI is the study of how to create intelligent agents.
- In practice, it is how to program a computer to behave and perform a task as an intelligent agent (say, a person) would. This does not have to involve learning or induction at all.

Machine Learning

- Field of study that gives computers the ability to learn without being explicitly programmed.
- Machine learning arose as a subfield of Artificial Intelligence
- Provide new capability to computers
- Machine learning has a greater emphasis on large scale applications and prediction accuracy

Statistical Learning

- Is a framework for machine learning drawing from the fields of statistics and functional analysis.
- Statistical learning arose as a subfield of Statistics.
- Statistical learning emphasizes models and their interpretability, and precision and uncertainty

Data Mining

- Is applied machine learning. It focuses more on the practical aspects of deploying machine learning algorithms on large datasets. It is very much similar to machine learning.

Data Analytics

- The science of examining raw data with the purpose of drawing conclusions about that information.

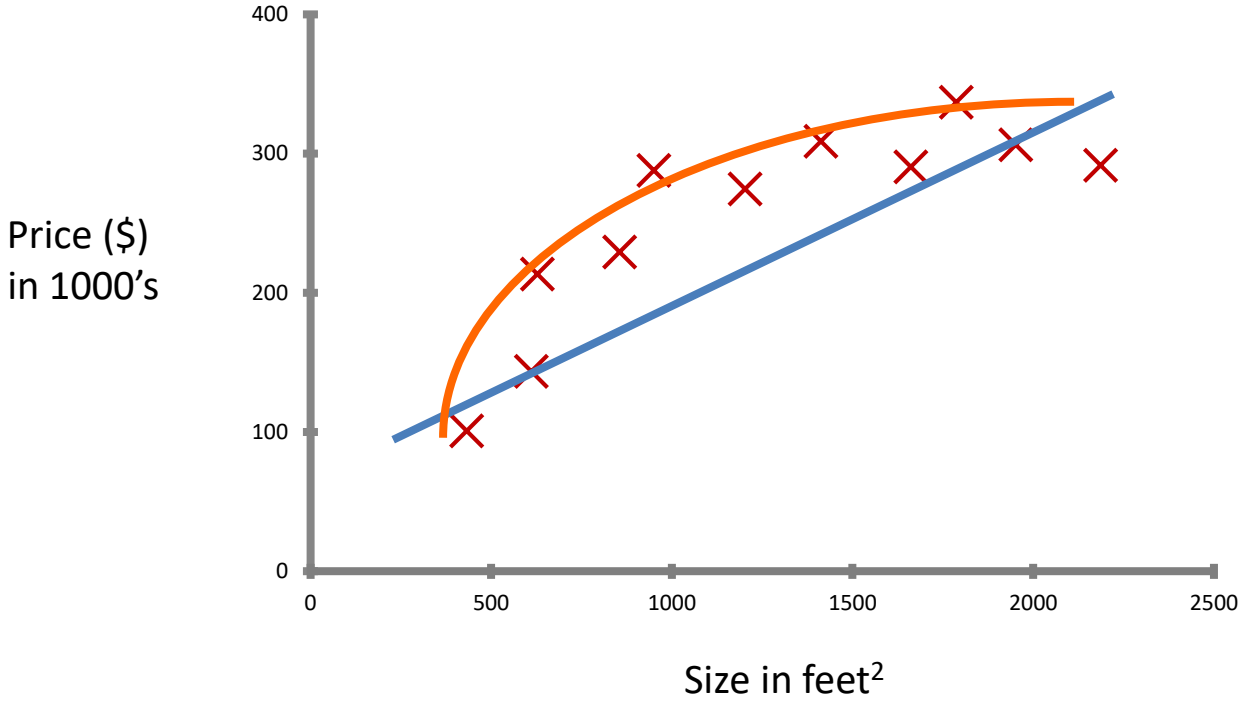
Supervised Learning

Machine learning algorithms:

-Supervised learning

-Unsupervised learning

Example: Housing price prediction.

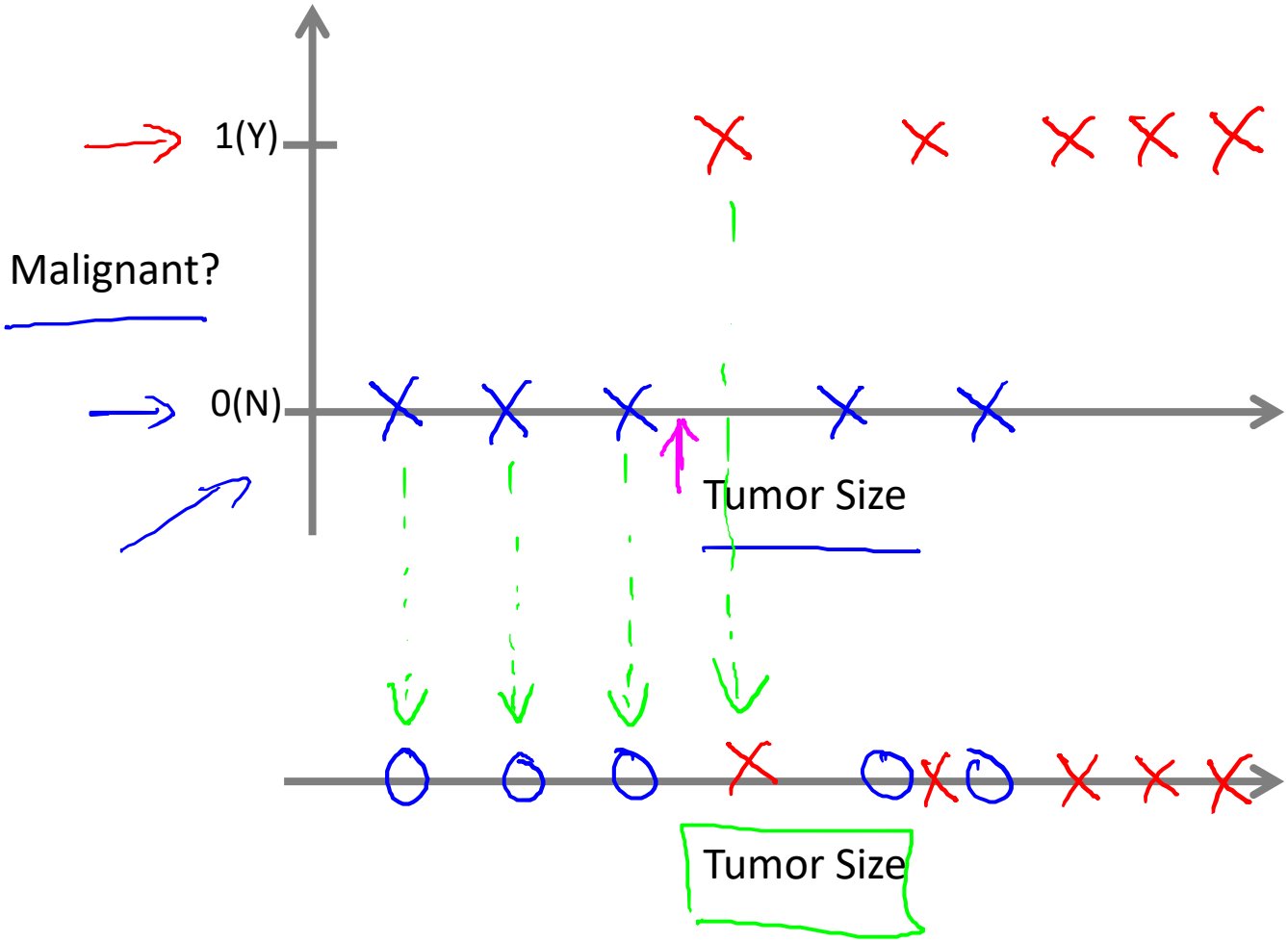


Supervised Learning

"right answers" given

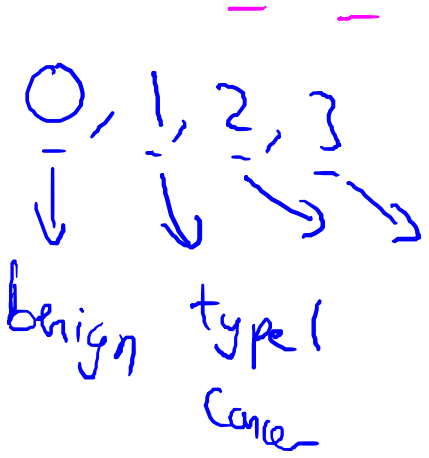
Regression: Predict continuous valued output (price)

Breast cancer (malignant, benign)

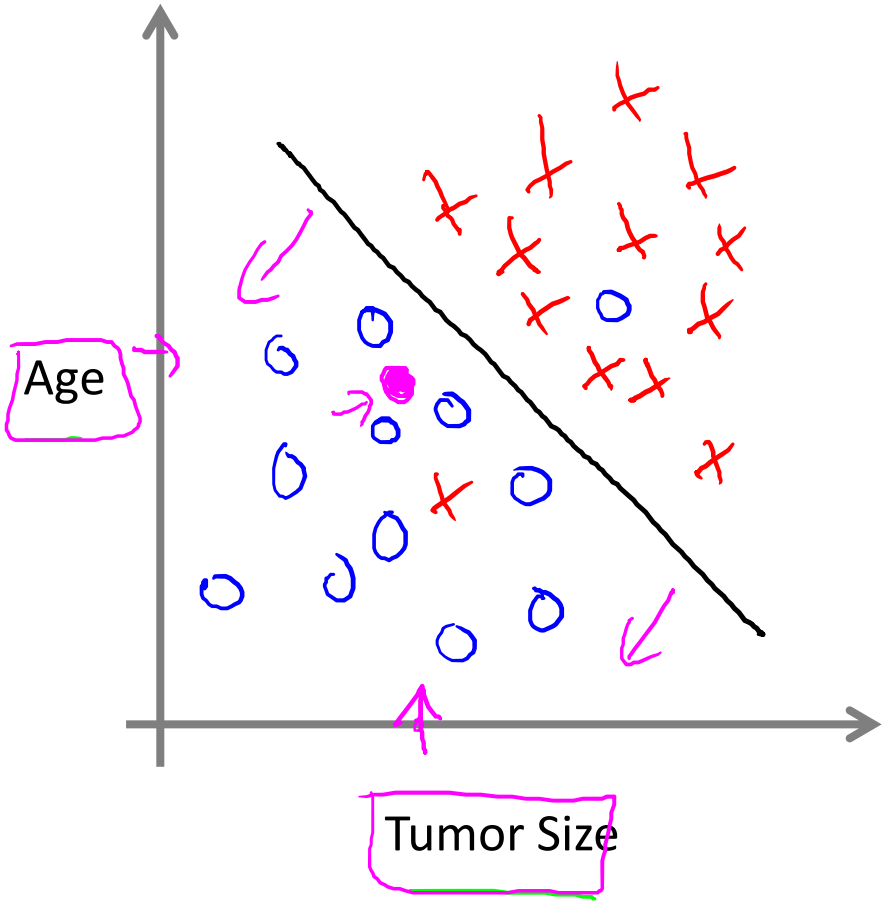


Classification

Discrete valued output (0 or 1)



Supervised Learning



- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape

...

Example

You're running a company, and you want to develop learning algorithms to address each of two problems.

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

Should you treat these as classification or as regression problems?

(a) Treat both as classification problems.

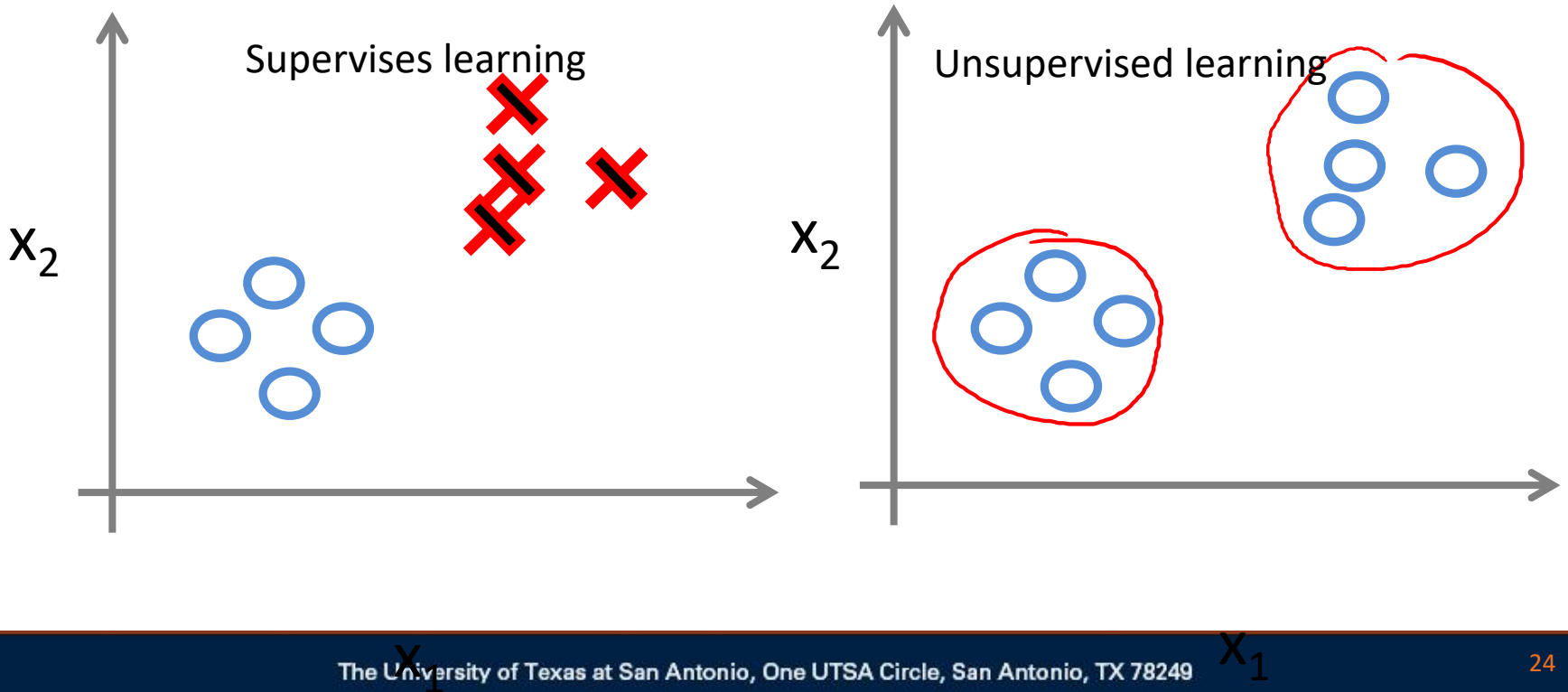
(b) Treat problem 1 as a classification problem, problem 2 as a regression problem.

(c) Treat problem 1 as a regression problem, problem 2 as a classification problem.

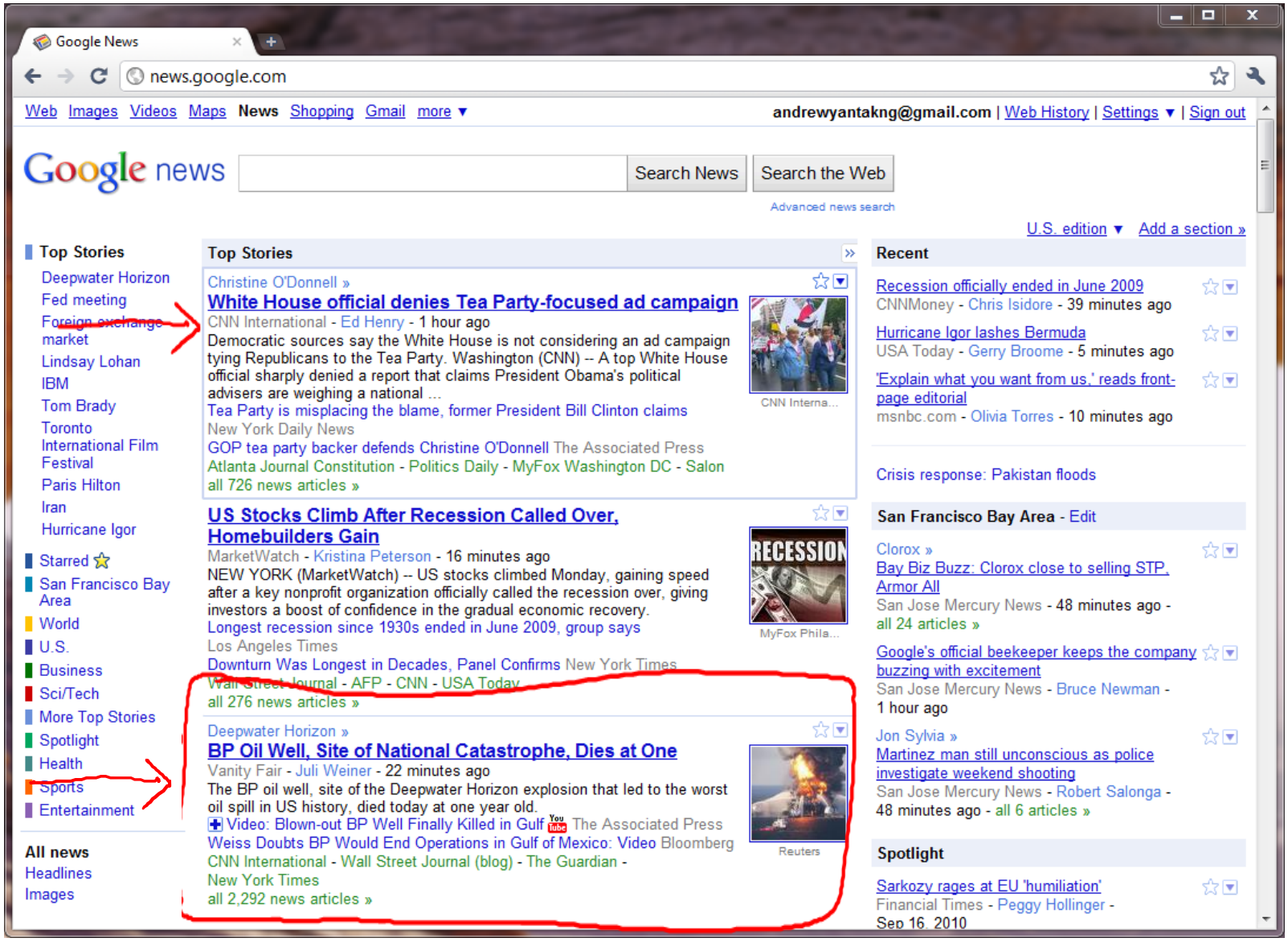
(d) Treat both as regression problems.

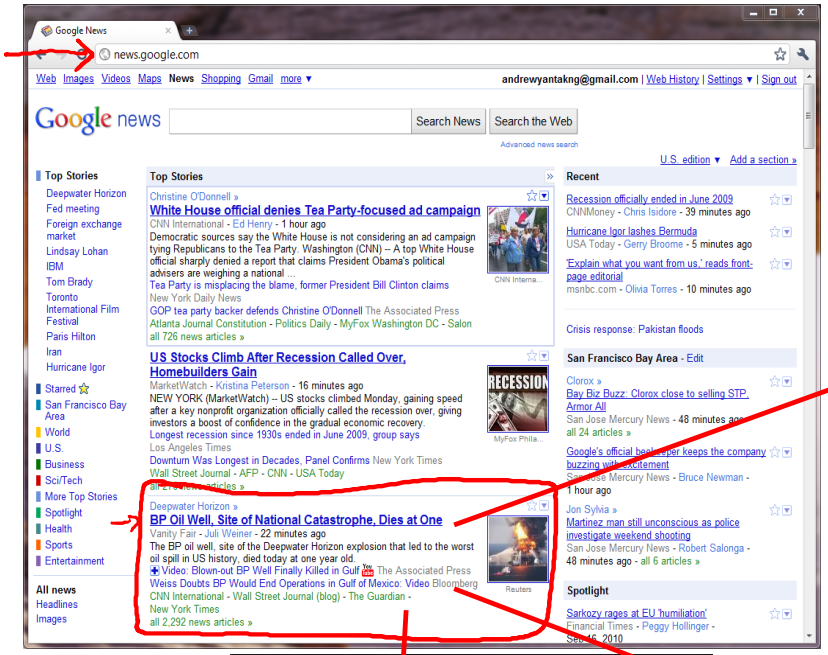
Unsupervised Learning

- Unsupervised learning allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables.
- We can derive this structure by clustering the data based on relationships among the variables in the data.
- With unsupervised learning there is no feedback based on the prediction results.

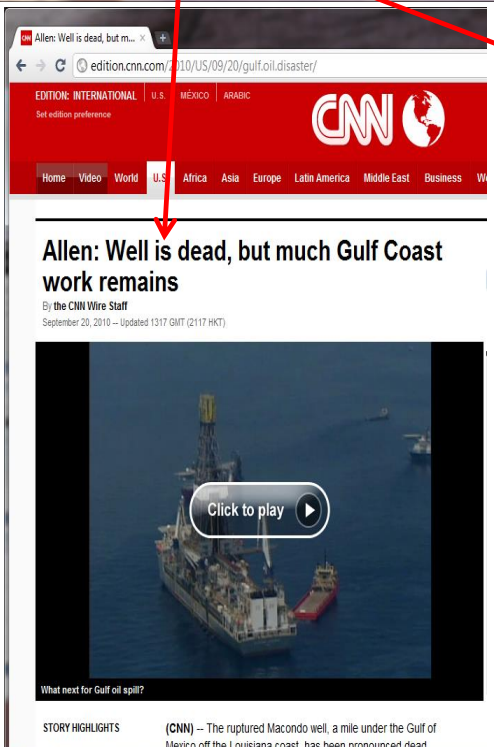


Example


Google News
news.google.com
Web Images Videos Maps News Shopping Gmail more ▼
andrewyantakng@gmail.com | Web History | Settings ▼ | Sign out
Google news
Search News
Search the Web
Advanced news search
U.S. edition ▼ Add a section »
Top Stories
Deepwater Horizon
Fed meeting
Foreign exchange market
Lindsay Lohan
IBM
Tom Brady
Toronto International Film Festival
Paris Hilton
Iran
Hurricane Igor
Starred
San Francisco Bay Area
World
U.S.
Business
Sci/Tech
More Top Stories
Spotlight
Health
Sports
Entertainment
All news
Headlines
Images
Top Stories
Christine O'Donnell »
White House official denies Tea Party-focused ad campaign
CNN International - Ed Henry - 1 hour ago
Democratic sources say the White House is not considering an ad campaign tying Republicans to the Tea Party. Washington (CNN) -- A top White House official sharply denied a report that claims President Obama's political advisers are weighing a national ...
Tea Party is misplacing the blame, former President Bill Clinton claims New York Daily News
GOP tea party backer defends Christine O'Donnell The Associated Press
Atlanta Journal Constitution - Politics Daily - MyFox Washington DC - Salon
all 726 news articles »
CNN Interna...
US Stocks Climb After Recession Called Over, Homebuilders Gain
MarketWatch - Kristina Peterson - 16 minutes ago
NEW YORK (MarketWatch) -- US stocks climbed Monday, gaining speed after a key nonprofit organization officially called the recession over, giving investors a boost of confidence in the gradual economic recovery.
Longest recession since 1930s ended in June 2009, group says Los Angeles Times
Downturn Was Longest in Decades, Panel Confirms New York Times
Wall Street Journal - AFP - CNN - USA Today
all 276 news articles »
RECESSION
MyFox Phila...
Deepwater Horizon »
BP Oil Well, Site of National Catastrophe, Dies at One
Vanity Fair - Juli Weiner - 22 minutes ago
The BP oil well, site of the Deepwater Horizon explosion that led to the worst oil spill in US history, died today at one year old.
Video: Blown-out BP Well Finally Killed in Gulf The Associated Press
Weiss Doubts BP Would End Operations in Gulf of Mexico: Video Bloomberg
CNN International - Wall Street Journal (blog) - The Guardian - New York Times
all 2,292 news articles »
Reuters
Recent
Recession officially ended in June 2009
CNNMoney - Chris Isidore - 39 minutes ago
Hurricane Igor lashes Bermuda
USA Today - Gerry Broome - 5 minutes ago
'Explain what you want from us,' reads front-page editorial
msnbc.com - Olivia Torres - 10 minutes ago
Crisis response: Pakistan floods
San Francisco Bay Area - Edit
Clorox »
Bay Biz Buzz: Clorox close to selling STP, Armor All
San Jose Mercury News - 48 minutes ago - all 24 articles »
Google's official beekeeper keeps the company buzzing with excitement
San Jose Mercury News - Bruce Newman - 1 hour ago
Jon Sylvia »
Martinez man still unconscious as police investigate weekend shooting
San Jose Mercury News - Robert Salonga - 48 minutes ago - all 6 articles »
Spotlight
Sarkozy rages at EU 'humiliation'
Financial Times - Peggy Hollinger - Sep 16, 2010

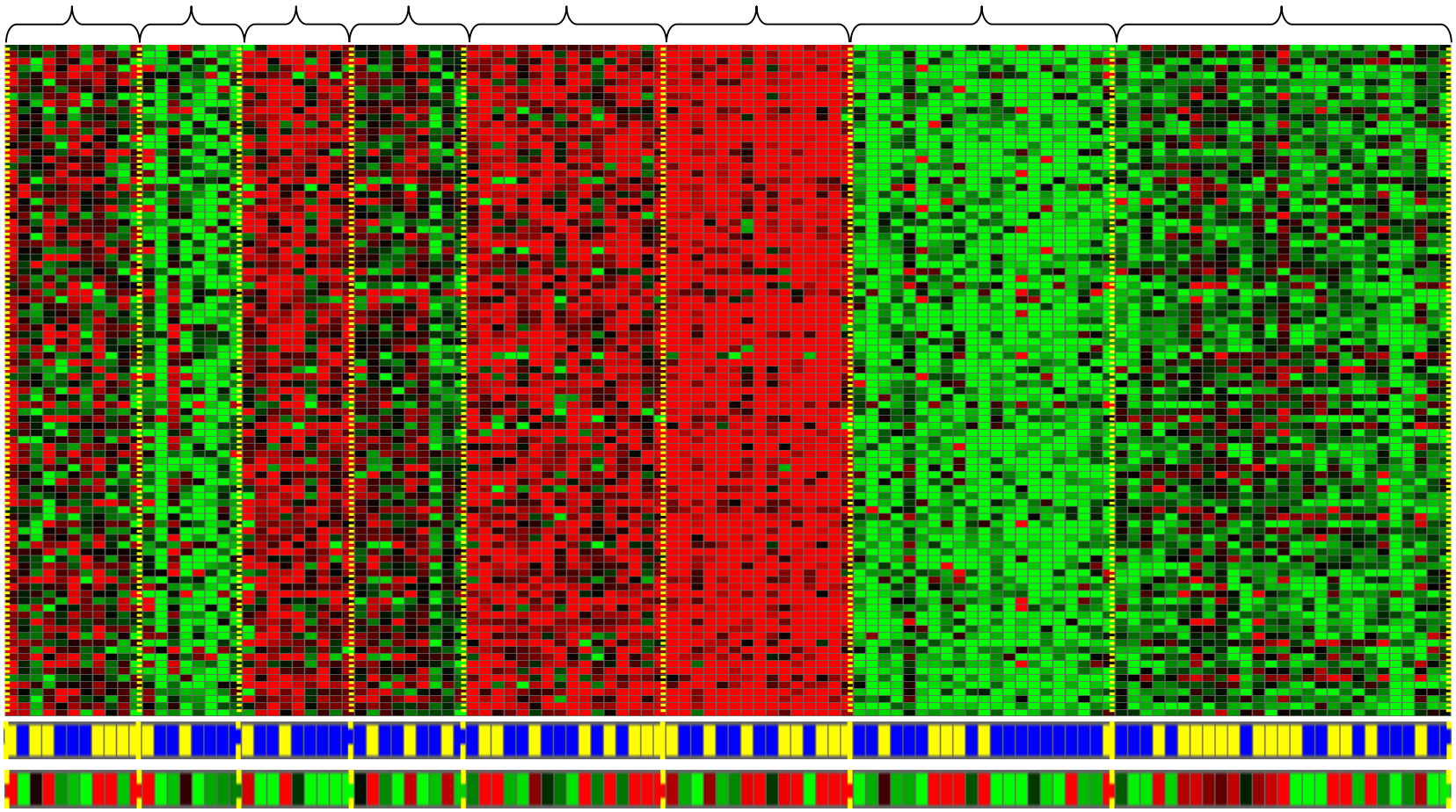

Google News
news.google.com
Google news
Search News
Search the Web
Top Stories
Recent
US Stocks Climb After Recession Called Over, Homebuilders Gain
BP Oil Well, Site of National Catastrophe, Dies at One
Spotlight


BP Kills Macondo, But Its ...
blogs.wsj.com/source/2010/09/20/bp-kills-macondo-but-its-legacy-lives-on/
THE WALL STREET JOURNAL.
THE SOURCE
BP Kills Macondo, But Its Legacy Lives On
Article
Comments (2)
About The Source
Most Recent

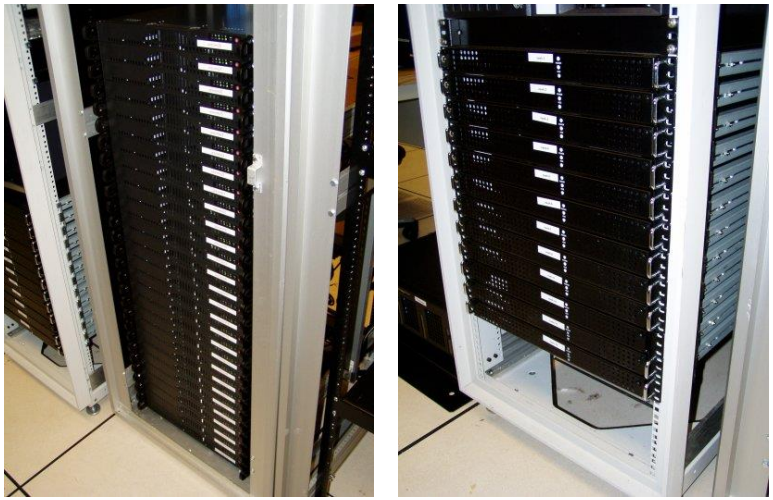

Allen: Well is dead, but m...
edition.cnn.com/2010/US/09/20/gulf.oil.disaster/
EDITION: INTERNATIONAL
CNN
Allen: Well is dead, but much Gulf Coast work remains
By the CNN Wire Staff
September 20, 2010 -- Updated 1317 GMT (2117 HKT)
Click to play
What next for Gulf oil spill?
STORY HIGHLIGHTS
(CNN) -- The ruptured Macondo well, a mile under the Gulf of


BP oil spill cost hits nearl...
www.guardian.co.uk/environment/2010/sep/20/bp-oil-spill-dee
guardian.co.uk
News | Sport | Comment | Culture | Business | Money | Life & style
Business | BP
BP oil spill cost hits nearly $10bn
BP has set up a $20bn compensation fund after the Deepwater Horizon disaster, which has so far paid out 19,000 claims totalling more than $240m
Julia Kollewe
guardian.co.uk, Monday 20 September 2010 08.33 BST
Article history
BP's costs for the Deepwater Horizon disaster have hit $10bn. Photograph: Hol/Reuters

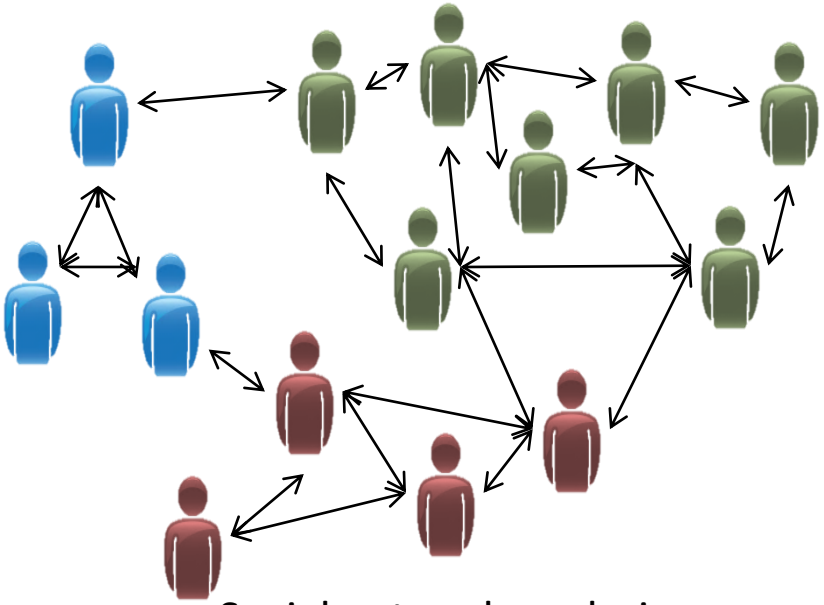
Example



[Source: Daphne Koller]



Organize computing clusters



Social network analysis

Market segmentation

Astronomical data analysis

Exercise

Of the following examples, which would you address using an unsupervised learning algorithm? (Check all that apply.)

(a) Given email labeled as spam/not spam, learn a spam filter.

(b) Given a set of news articles found on the web, group them into set of articles about the same story.

(c) Given a database of customer data, automatically discover market segments and group customers into different market segments.

(d) Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.