

Machine Learning - Intro

Aarti Singh

Machine Learning 10-701/15-781
Sept 8, 2010

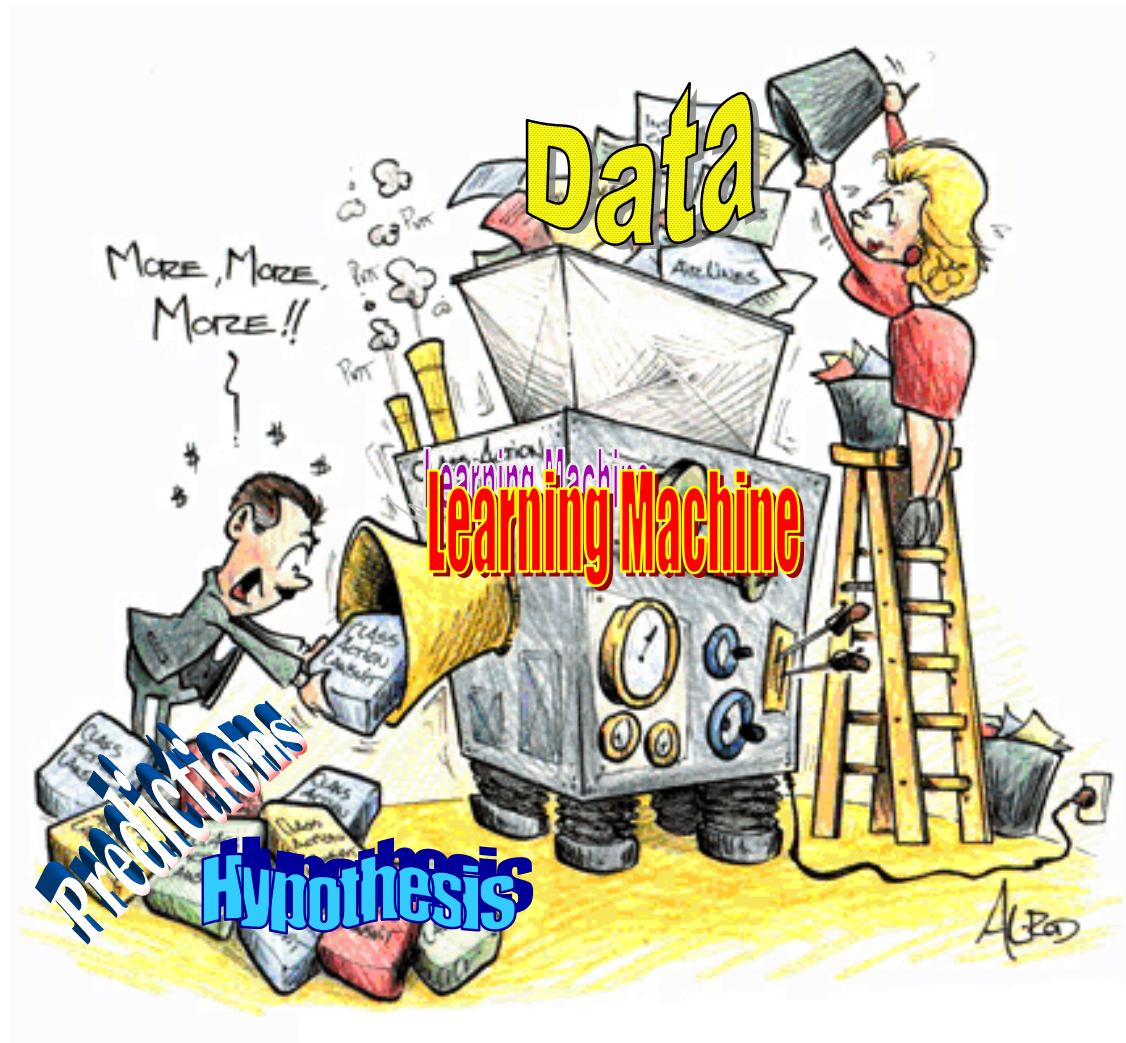


MACHINE LEARNING DEPARTMENT



What is Machine Learning?

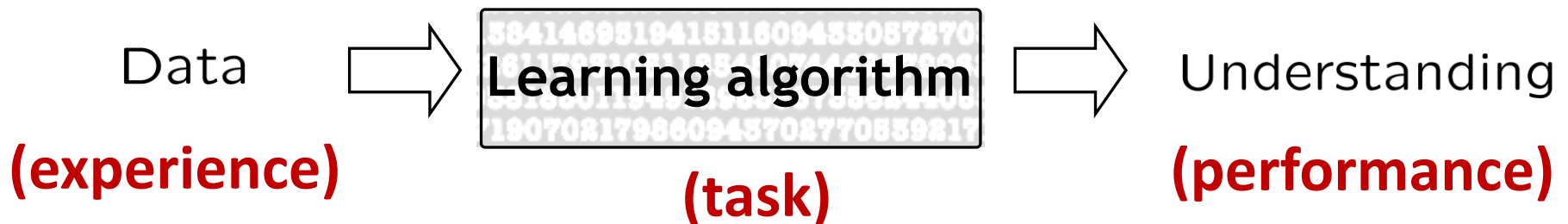
What is Machine Learning?



What is Machine Learning?

Study of algorithms that

- improve their performance
- at some task
- with experience

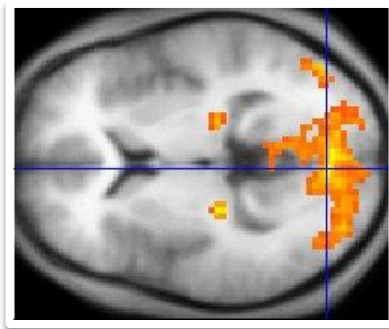


From Data to Understanding ...

Machine Learning in Action

Machine Learning in Action

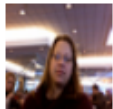
- Decoding thoughts from brain scans



Rob a bank ...

[Home](#) » [Health & Wellness](#)

Brain Scans: Are You a Criminal?



Published February 07, 2007 by:

[Andrea Okrentowich](#)

[View Profile](#) | [Follow](#) | [Add to Favorites](#)

More:

[Brain Scans](#)

[Brain Scan](#)

[Disposition](#)

[Defendant](#)

[Criminal Behavior](#)

MRI Scans as Courtroom Evidence

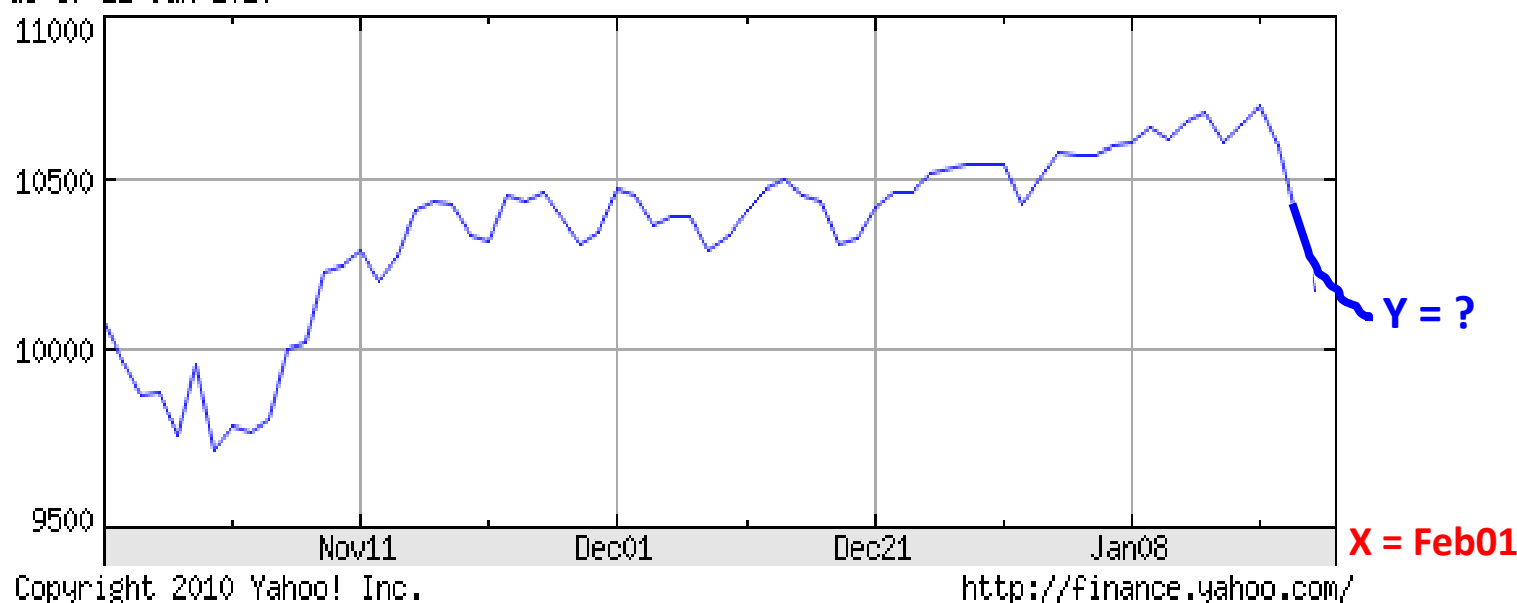
The average Joe's MRI scan can show a brain abnormality, do we proceed to check him into the nearest mental institution or prison? That would make about as much sense as trying to prove a defendant innocent of a violent



Machine Learning in Action

- Stock Market Prediction

DJ INDU AVERAGE (DOW JONES & CO
as of 22-Jan-2010



Machine Learning in Action

- Document classification



Sports
Science
News

Machine Learning in Action

- Spam filtering

Welcome to New Media Installation: Art that Learns

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.

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

=== 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



Spam/
Not spam

Machine Learning in Action

- Cars navigating on their own



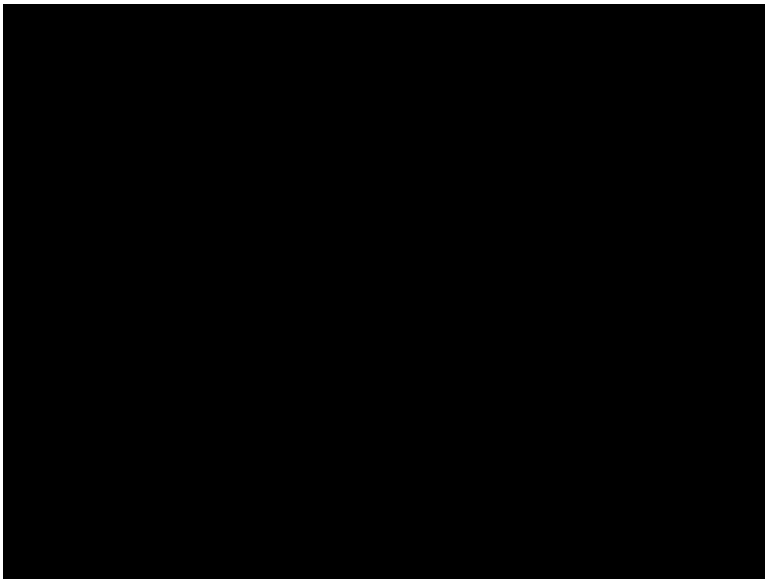
Boss, the self-driving SUV
1st place in the DARPA Urban
Challenge.

Photo courtesy of Tartan Racing.



Machine Learning in Action

- The **best** helicopter pilot is now a computer!
 - it runs a program that learns how to fly and make acrobatic maneuvers by itself!
 - no taped instructions, joysticks, or things like that ...



Machine Learning in Action

- Robot assistant?

[<http://stair.stanford.edu/>]



Machine Learning in Action

- Many, many more...

Speech recognition, Natural language processing

Computer vision

Web forensics

Medical outcomes analysis

Computational biology

Sensor networks

Social networks

...

Machine Learning in Action



ML students and postdocs at G-20 Pittsburgh Summit 2009

ML is trending!

- Wide applicability
- Very large-scale complex systems
 - Internet (billions of nodes), sensor network (new multi-modal sensing devices), genetics (human genome)
- Huge multi-dimensional data sets
 - 30,000 genes x 10,000 drugs x 100 species x ...
- Software too complex to write by hand
- Improved machine learning algorithms
- Improved data capture (Terabytes, Petabytes of data), networking, faster computers
- Demand for self-customization to user, environment


ML has a long way to go ...

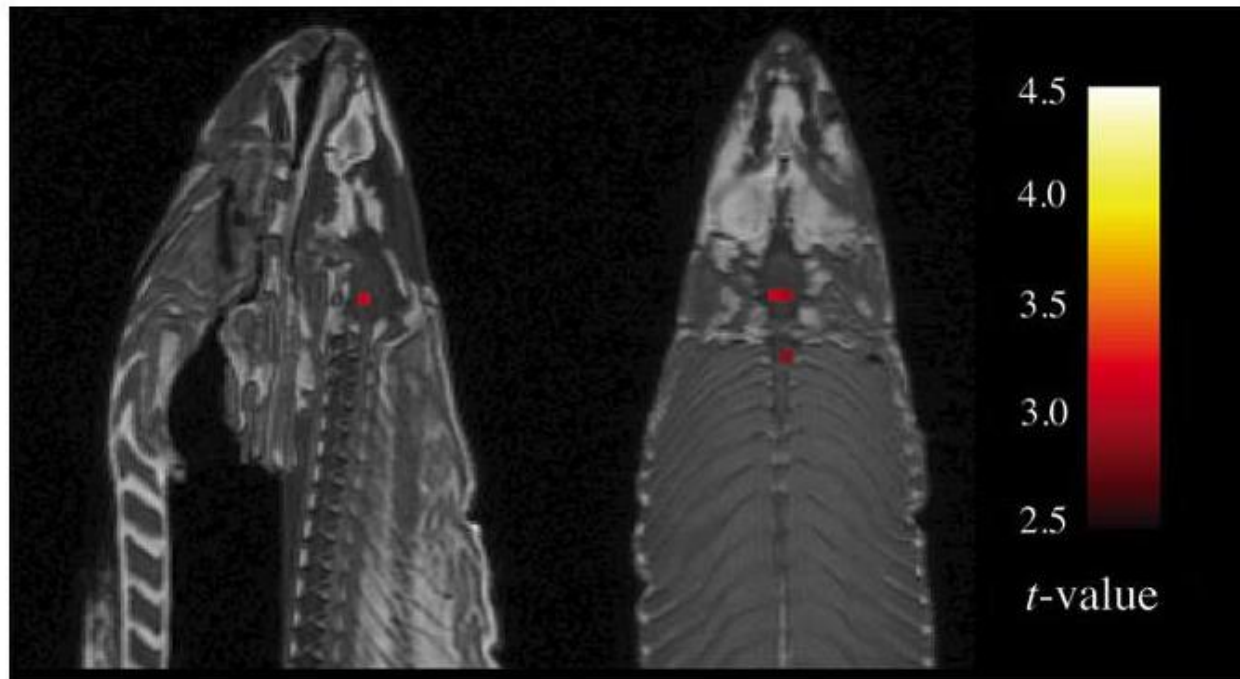
WIRED SCIENCE

NEWS FOR YOUR NEURONS



Scanning Dead Salmon in fMRI Machine Highlights Risk of Red Herrings

By Alexis Madrigal  September 18, 2009 | 5:37 pm | Categories: [Brains and Behavior](#)



ML has a long way to go ...

Speech Recognition gone Awry

What this course is about

- Covers a wide range of Machine Learning techniques
 - from basic to state-of-the-art
- You will learn about the methods you heard about:
 - Naïve Bayes, logistic regression, nearest-neighbor, decision trees, boosting, neural nets, overfitting, regularization, dimensionality reduction, PCA, error bounds, VC dimension, SVMs, kernels, margin bounds, K-means, EM, mixture models, semi-supervised learning, HMMs, graphical models, active learning, reinforcement learning...
- Covers algorithms, theory and applications
- **It's going to be fun and hard work 😊**

Machine Learning Tasks

Broad categories -

- **Supervised learning**

Classification, Regression

- **Unsupervised learning**

Density estimation, Clustering, Dimensionality reduction

- Semi-supervised learning
- Active learning
- Reinforcement learning
- Many more ...

Supervised Learning

Feature Space \mathcal{X}

Words in a document



Label Space \mathcal{Y}

“Sports”
“News”
“Science”
...



DJ INDU AVERAGE (DOW JONES & CO
as of 22-Jan-2010



Copyright 2010 Yahoo! Inc.

<http://finance.yahoo.com/>

Share Price
“\$ 24.50”



Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

Supervised Learning - Classification

Feature Space \mathcal{X}

Words in a document

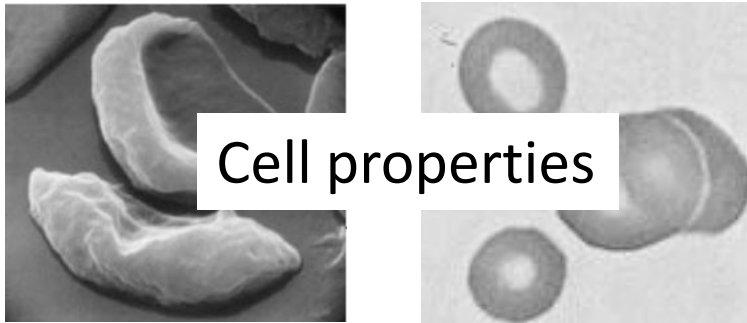


Label Space \mathcal{Y}

"Sports"
"News"
"Science"
...



Cell properties



"Anemic cell"
"Healthy cell"



Discrete Labels

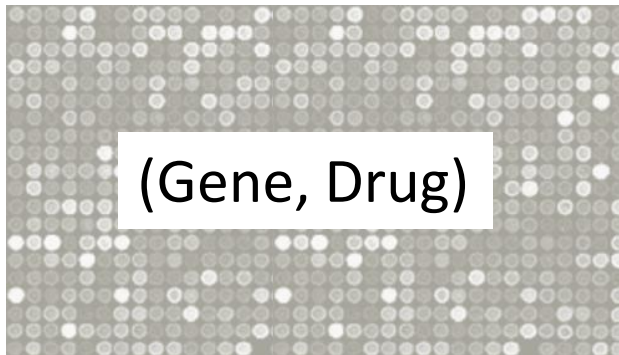
Supervised Learning - Regression

Feature Space \mathcal{X}

Label Space \mathcal{Y}



Share Price
"\$ 24.50"



Expression level
"0.01"

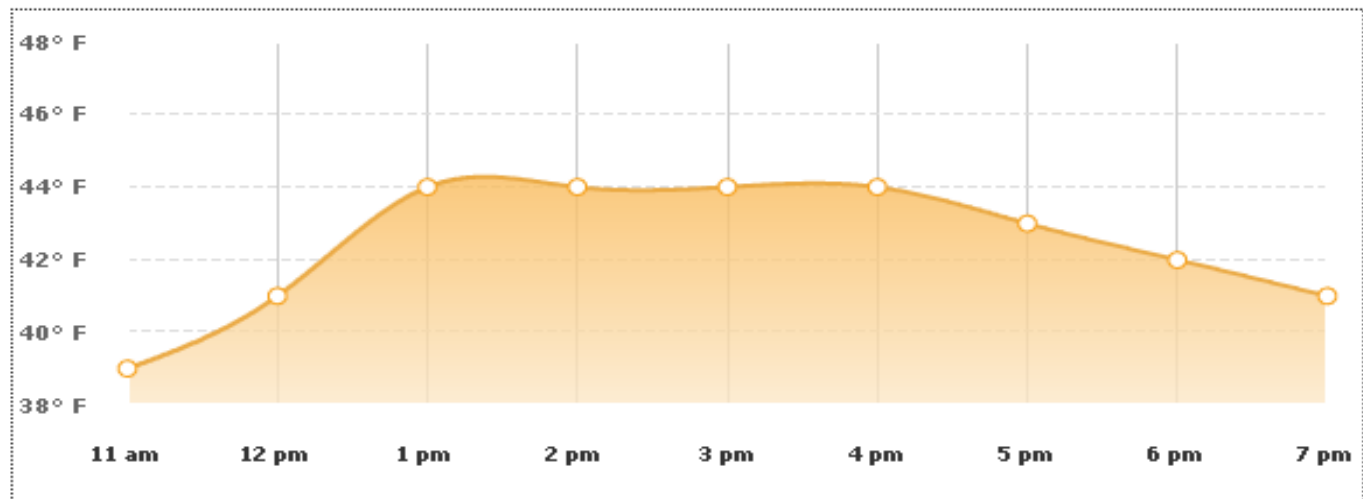
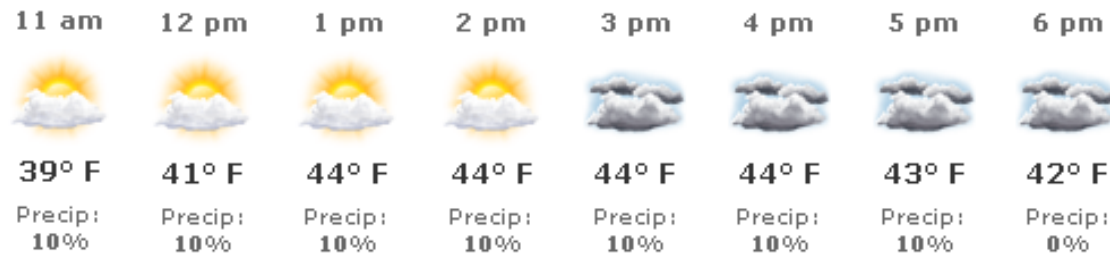
Continuous Labels

Supervised Learning problems

Features?

Labels?

Classification/Regression?



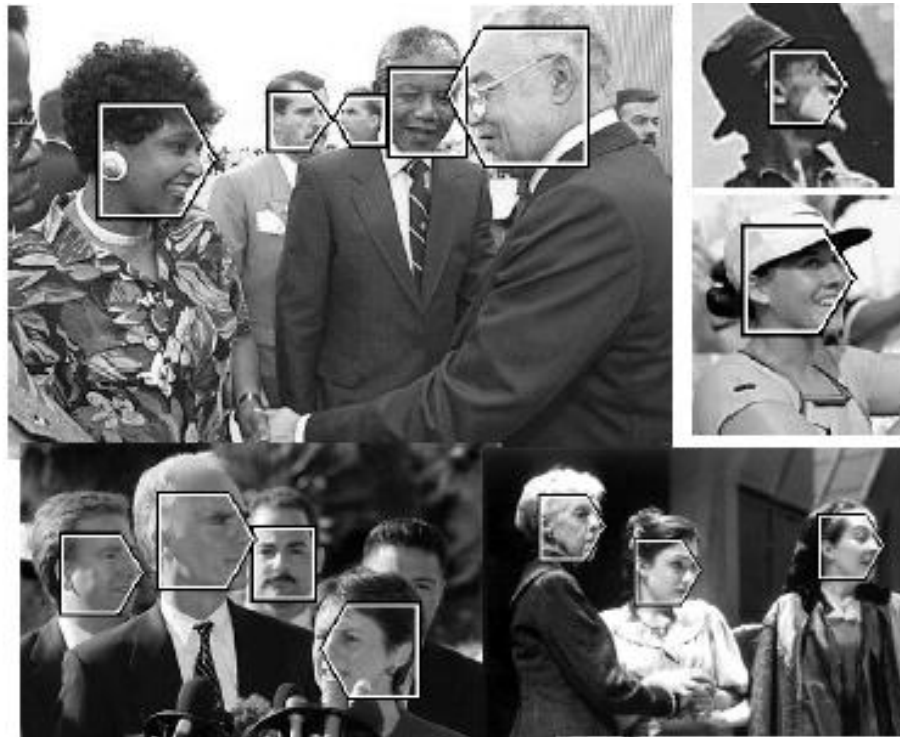
Temperature/Weather prediction

Supervised Learning problems

Features?

Labels?

Classification/Regression?



Face Detection

Supervised Learning problems

Features?

Labels?

Classification/Regression?



Environmental Mapping

Supervised Learning problems

Features?

Labels?

Classification/Regression?



Robotic Control

Unsupervised Learning

Aka “learning without a teacher”

Feature Space \mathcal{X}

Words in a document

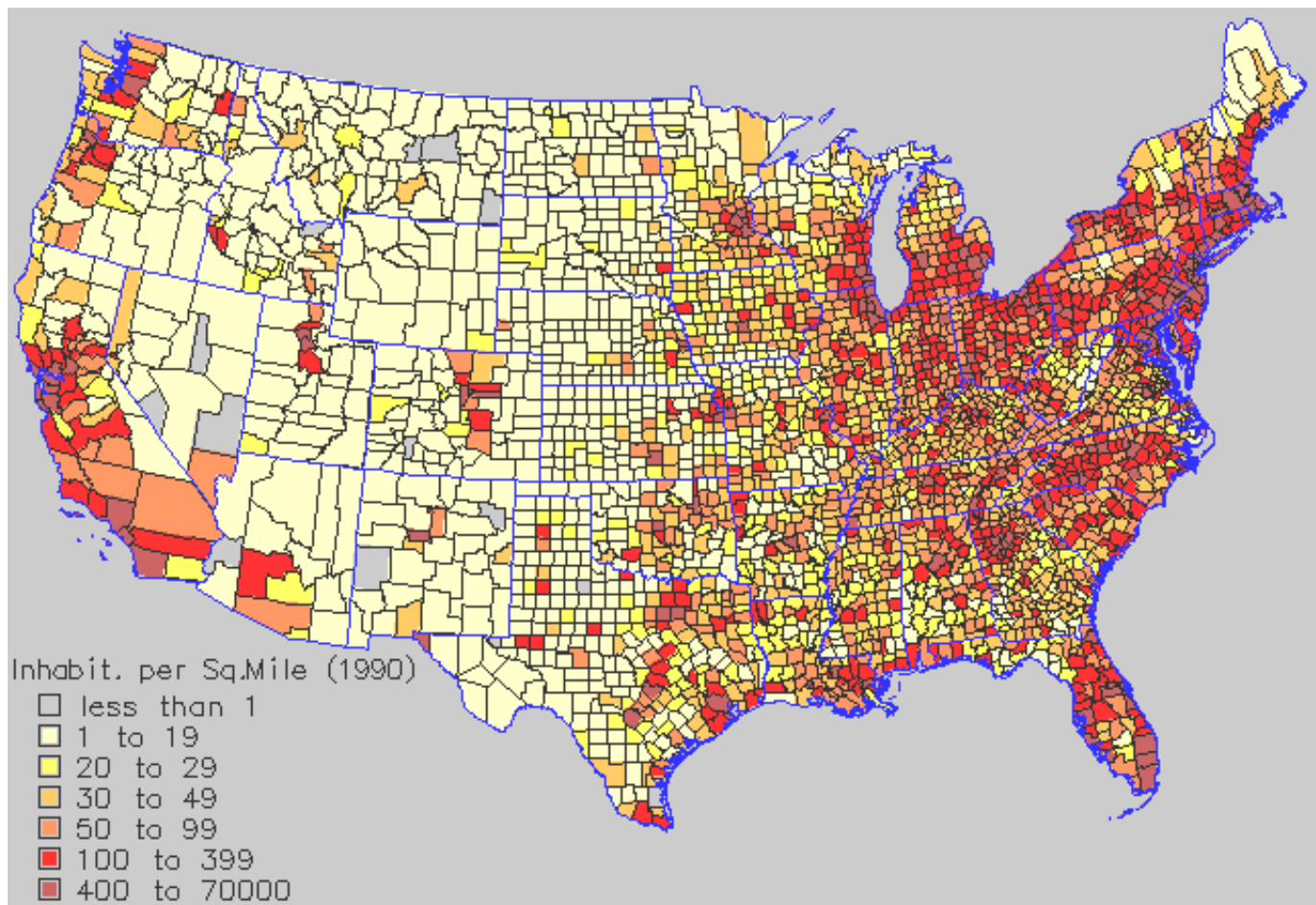


Word distribution
(Probability of a word)

Task: Given $X \in \mathcal{X}$, learn $f(X)$.

Unsupervised Learning – Density Estimation

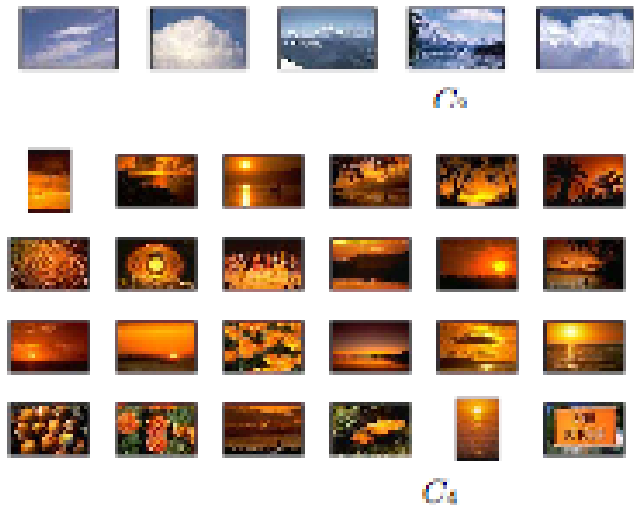
Population density




Unsupervised Learning – clustering

Group similar things e.g. images

[Goldberger et al.]



Unsupervised Learning – clustering web search results

web news images wikipedia blogs jobs more »Search[advanced preferences](#)

clusters sources sites

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)

Cluster **Human** contains 8 documents.

- [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 categories that are based on visible traits (especially skin color, cranial or facial features and hair texture), and self-identity by culture and over time, and are often controversial for scientific as well as social and political reasons. History · More on [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 sailboat race (see **Race** and ethnicity in the United States Census, official definitions of "**race**" used by the US Census Bureau. Historical definitions of **race**; **Race** (bearing), the inner and outer rings of a rolling-element bearing. **RACE** Literature · Video games
[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 ...
[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
[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. 100, pp. 1-10, 1999.
[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 inherited characteristics.
[www.answers.com/topic/race-1](#) - [cache] - Live

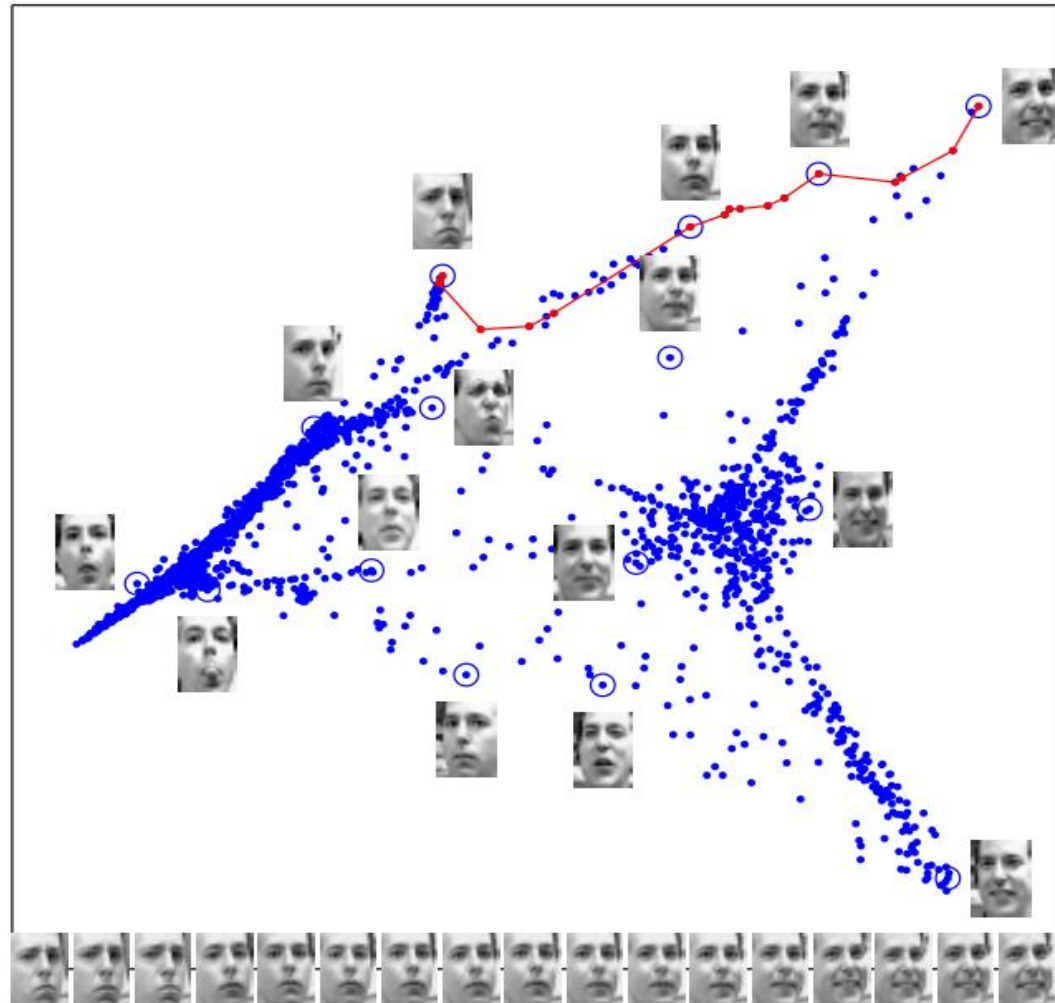
Unsupervised Learning - Embedding

Dimensionality Reduction

[Saul & Roweis '03]

Images have thousands or millions of pixels.

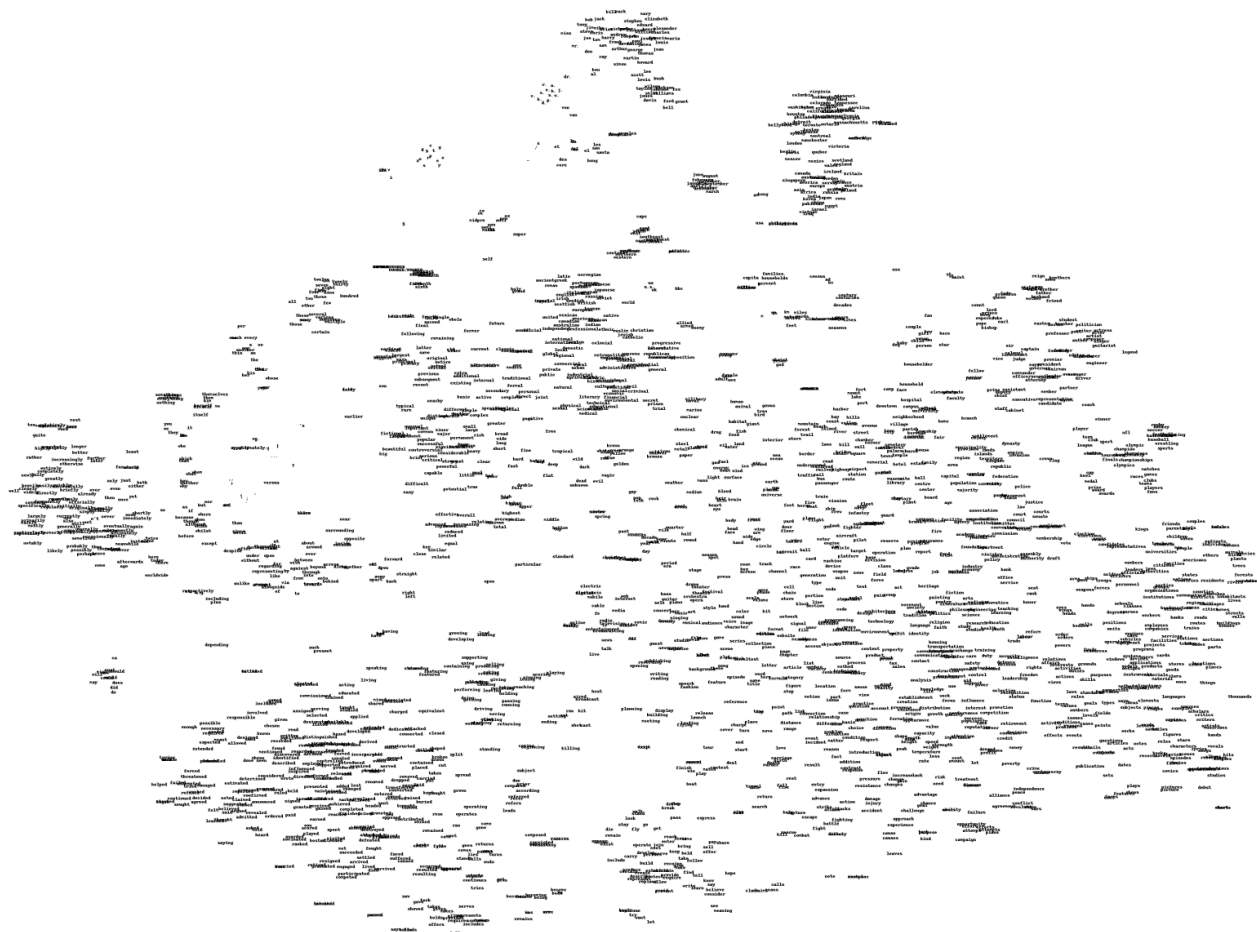
Can we give each image a coordinate, such that similar images are near each other?



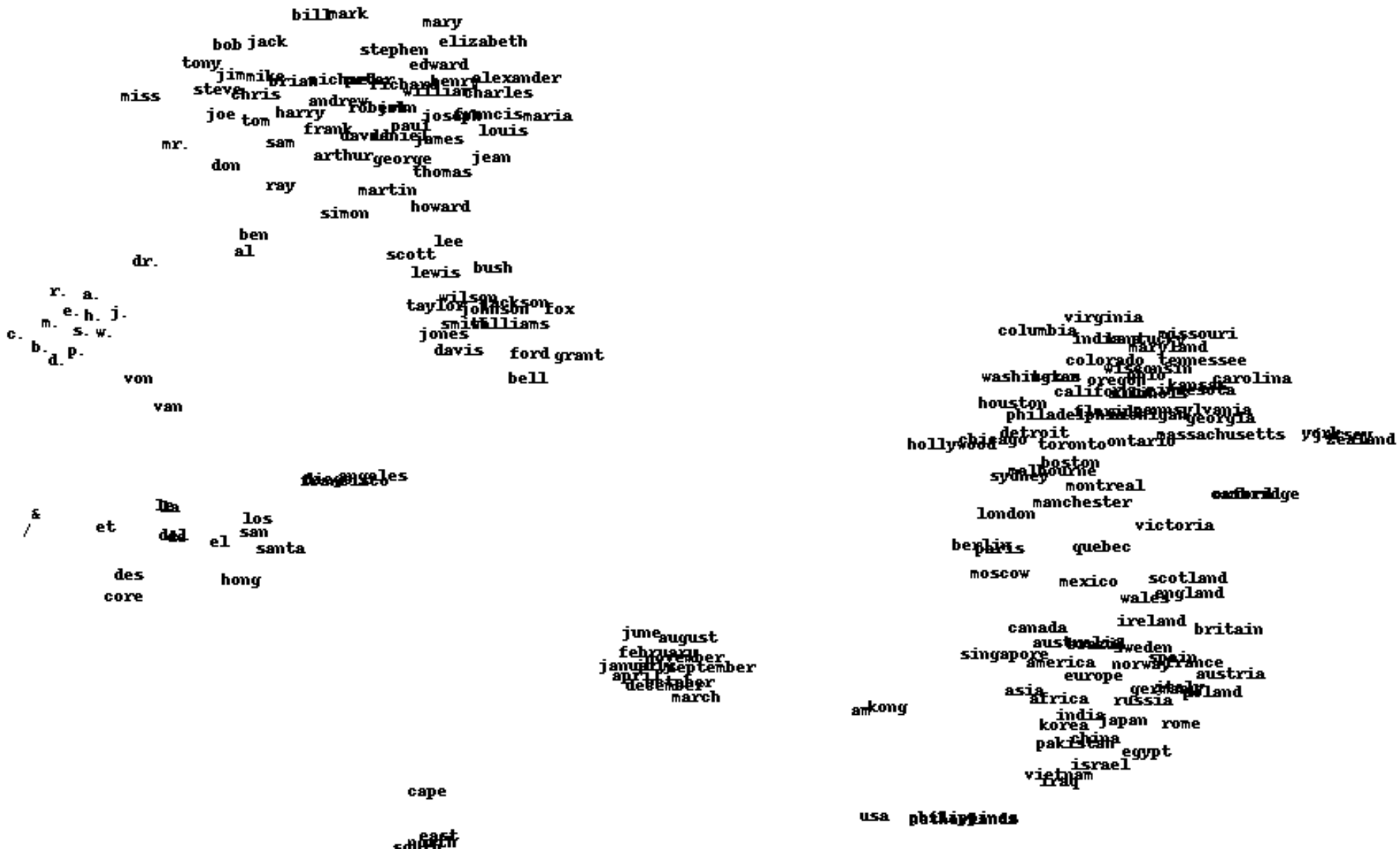
Unsupervised Learning - Embedding

Dimensionality Reduction - words

[Joseph Turian]



Unsupervised Learning - Embedding



Machine Learning Tasks

Broad categories -

- **Supervised learning**

Classification, Regression

- **Unsupervised learning**

Density estimation, Clustering, Dimensionality reduction

- Semi-supervised learning
- Active learning
- Reinforcement learning
- Many more ...

Machine Learning Class webpage

- <http://www.cs.cmu.edu/~aarti/Class/10701/index.html>

Auditing

- To satisfy the auditing requirement, you must either:
 - Do *two* homeworks, and get at least 75% of the points in each; or
 - Take the final, and get at least 50% of the points; or
 - Do a class project
 - Only need to submit project proposal and present poster, and get at least 80% points in the poster
- Please, send the instructors an email saying that you will be auditing the class and what you plan to do.

Prerequisites

- Probabilities
 - Distributions, densities, marginalization...
- Basic statistics
 - Moments, typical distributions, regression...
- Algorithms
 - Dynamic programming, basic data structures, complexity...
- Programming
 - Mostly your choice of language, but Matlab will be very useful
- We provide some background, but the class will be fast paced
- Ability to deal with “abstract mathematical concepts”

Recitations

- Strongly recommended
 - Brush up pre-requisites
 - Review material (difficult topics, clear misunderstandings, extra new topics)
 - Ask questions
- Basics of Probability
- Thursday, Sept 9, Tomorrow!
- NSH 3305



Rob Hall

Textbooks

- Recommended Textbook:
 - Pattern Recognition and Machine Learning; Chris Bishop
- Secondary Textbooks:
 - The Elements of Statistical Learning: Data Mining, Inference, and Prediction; Trevor Hastie, Robert Tibshirani, Jerome Friedman (see online link)
 - Machine Learning; Tom Mitchell
 - Information Theory, Inference, and Learning Algorithms; David MacKay

Grading

- 5 Homeworks (35%)
 - First one goes out next week (watch email)
 - Start early, Start early, Start early, Start early, Start early, Start early, Start early, Start early, Start early
- Final project (25%)
 - Details out around Sept. 30th
 - Projects done individually, or groups of two students
- Midterm (20%)
 - Wed., Oct 20 in class
- Final exam (20%)
 - TBD by registrar

Homeworks

- Homeworks are hard, start early 😊
- Due in the beginning of class
- 2 late days for the semester
- After late days are used up:
 - Half credit within 48 hours
 - Zero credit after 48 hours
- Atleast 4 homeworks **must be handed in**, even for zero credit
- Late homeworks handed in to Michelle Martin, GHC 8001

Homeworks

- Collaboration
 - You may **discuss** the questions
 - Each student writes their own answers
 - Each student must write their own code for the programming part
 - **Please don't search for answers on the web, Google, previous years' homeworks, etc.**
 - please ask us if you are not sure if you can use a particular reference

First Point of Contact for HWs

- To facilitate interaction, a TA will be assigned to each homework question – This will be your “first point of contact” for this question
 - But, you can always ask any of us

Communication Channel

- For e-mailing instructors, always use:
 - 10701-instructors@cs.cmu.edu
- For announcements, subscribe to:
 - 10701-announce@cs
 - <https://mailman.srv.cs.cmu.edu/mailman/listinfo/10701-announce>
- For discussions, use blackboard
 - <https://blackboard.andrew.cmu.edu/>

Your saviours - TAs



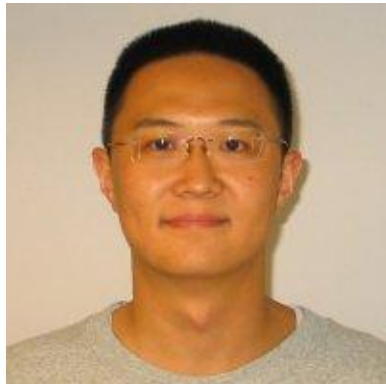
Leman Akoglu



Min Chi



Rob Hall



T. K. Huang



Jayant Krishnamurthy

**Great resources for
learning,
Interact with them!**

Leman's research interests



WIKIPEDIA
The Free Encyclopedia



Graph mining (large, time-varying graphs)

- Patterns and generators
 - What characteristics do "real" graphs exhibit?
 - Can we model a given graph to generate realistic graphs?
- Anomaly detection
 - Can we spot "suspicious" nodes?
 - Can we point "suspicious" events?
- Recommendations
 - How can we answer "who's-close to-whom" queries on disk-resident, time-varying graphs?
 - How do we recommend both "close" and "profitable" links?

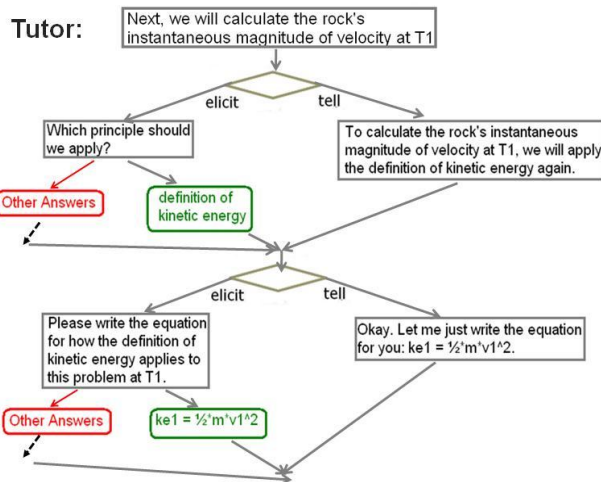
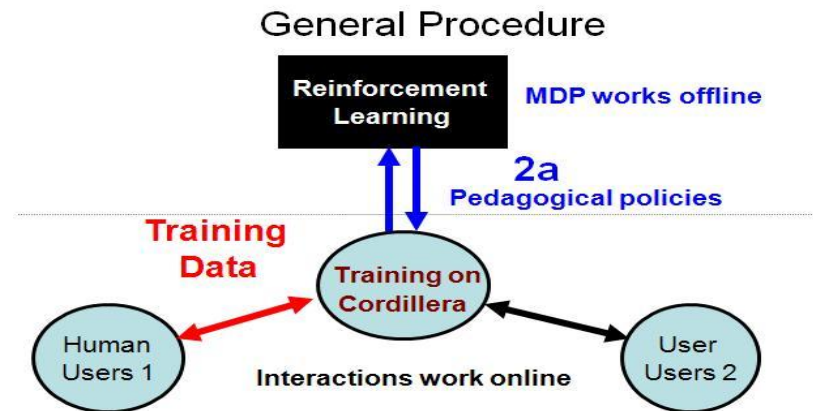
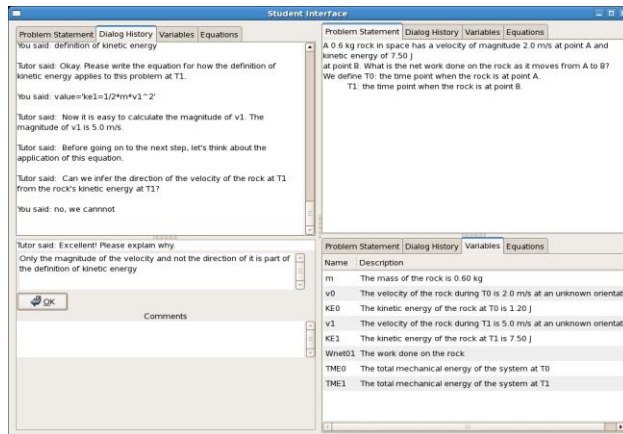


at&t

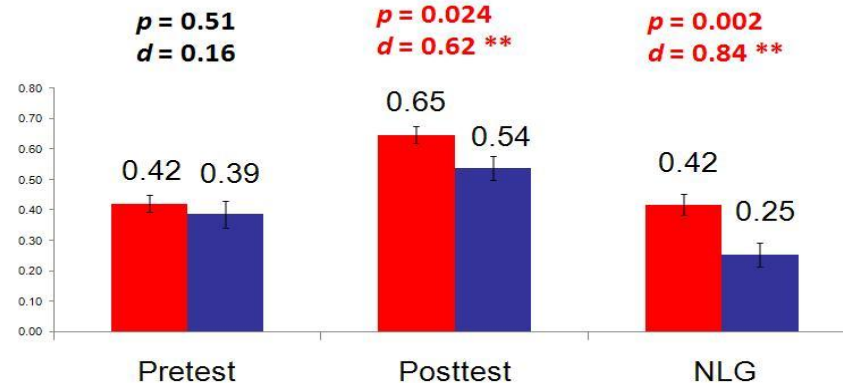


Applying Reinforcement Learning To Induce Pedagogical Strategies

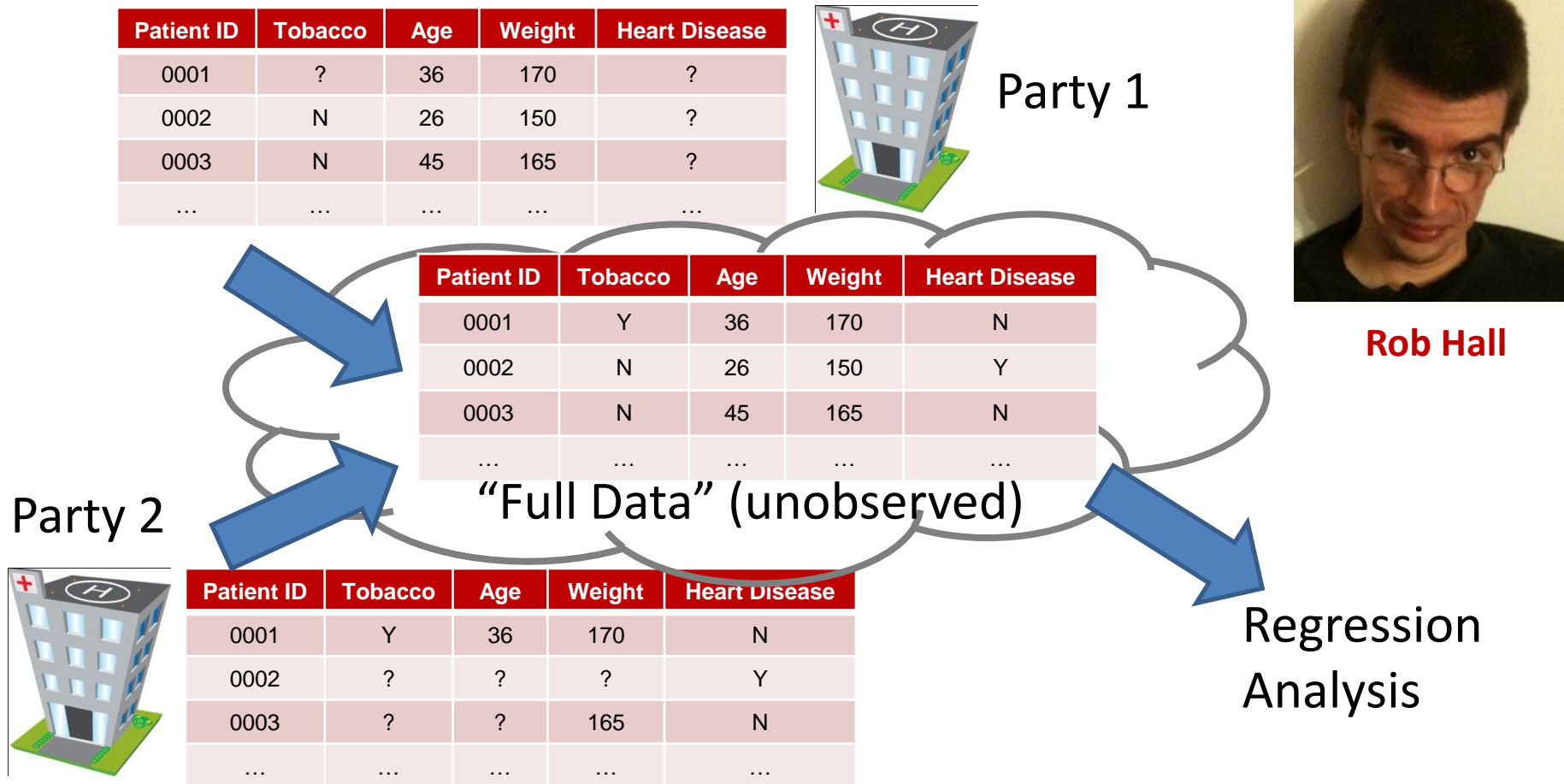
Min Chi, Machine Learning Department, Carnegie Mellon University



Learning Gains

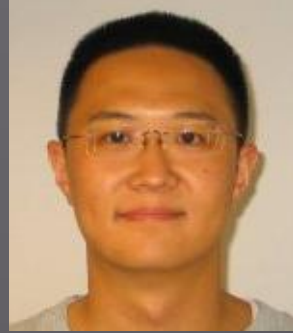


- Several parties have data on a common set of entities, but each party's data is incomplete:



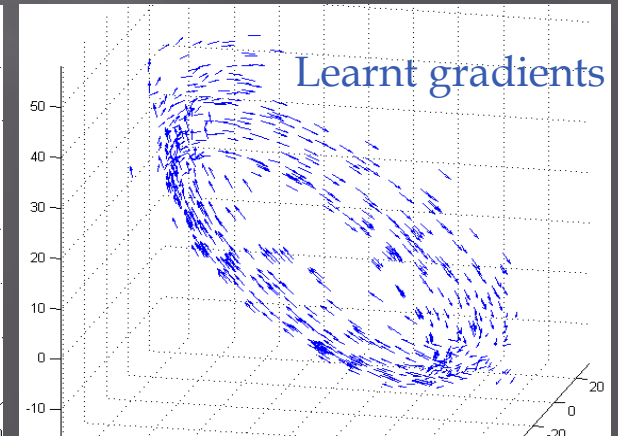
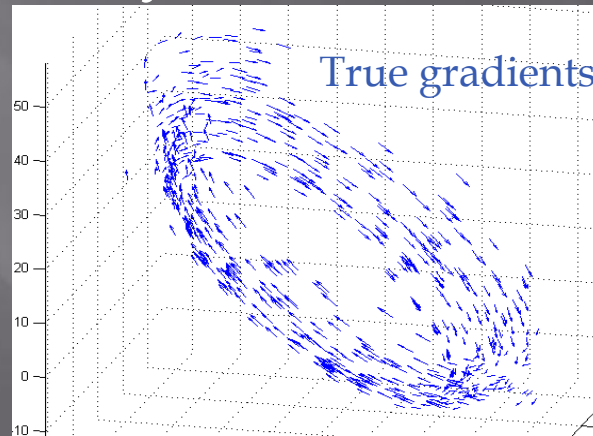
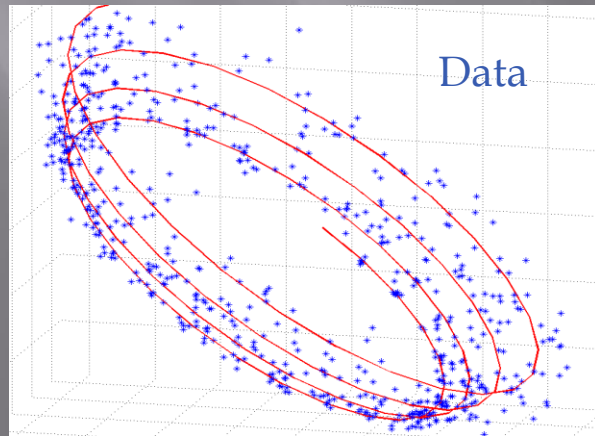
- Each party's data is private, and the parties are unwilling to share their data.
- We do regression on the unknown, full data matrix, without requiring the parties to reveal their private data.

Learning Dynamic Models from Non-sequenced Data



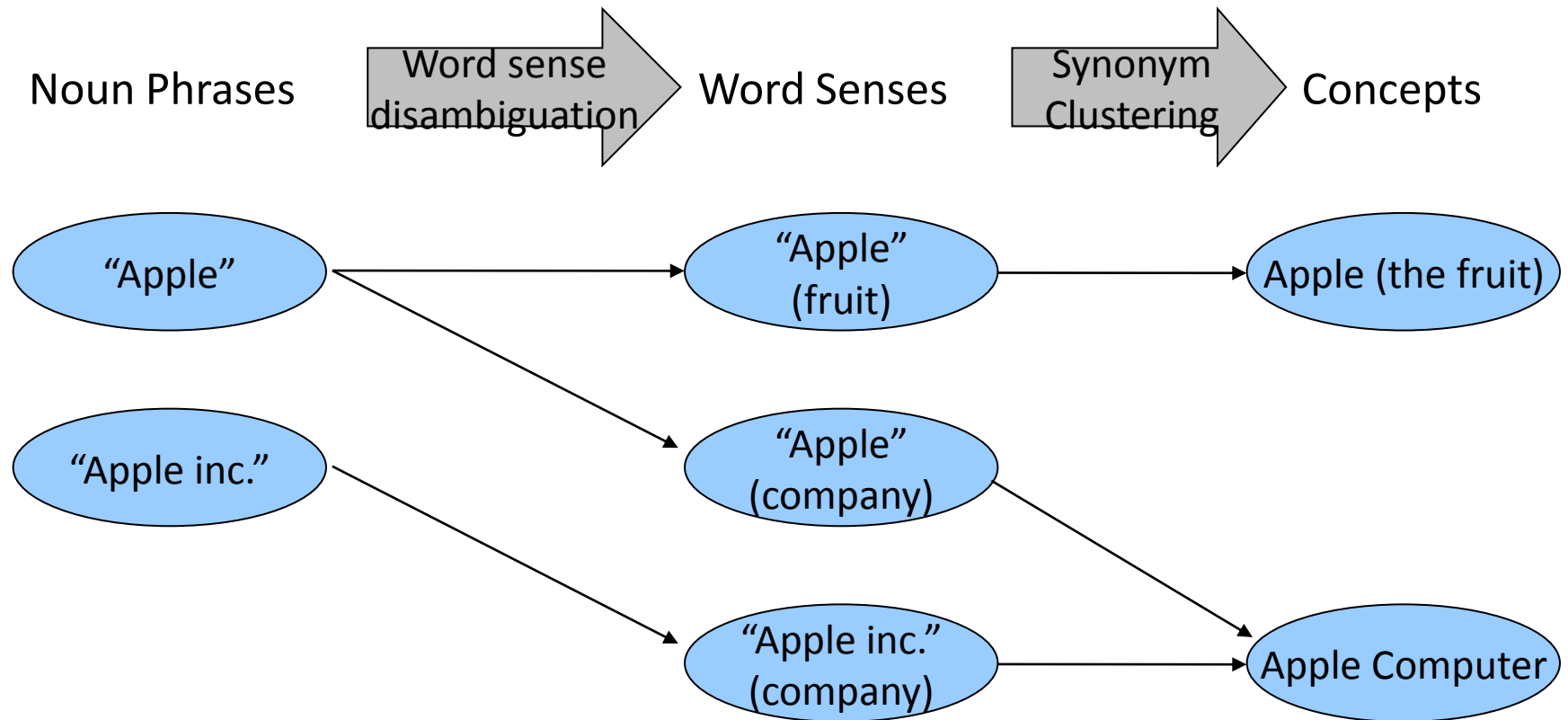
T.K. Huang

- ▣ Dynamic models are useful for analyzing time-evolving data, e.g., speech, video, robot movement
- ▣ Usual assumption: observations are **time-stamped**
- ▣ But sometimes “time” is NOT easily available:
 - Galaxy evolution (many static snapshots)
 - Chronic disease, e.g., Alzheimer’s (tracking patients is expensive)
 - Destructive measurement of biological processes
- ▣ How can we learn dynamic models from such data?



Synonym Resolution for Read the Web

Jayant Krishnamurthy



Your saviour

- Administrative Assistant



Michelle Martin

Late homeworks, administrative issues (registering, dropping, converting to audit ...)

Enjoy!

- ML is becoming ubiquitous in science, engineering and beyond
- This class should give you the basic foundation for applying ML and developing new methods
- The fun begins...