Machine Learning - Intro

Aarti Singh

Machine Learning 10-315
Aug 31, 2020

Teaching team

Instructor:

Aarti

Admin:

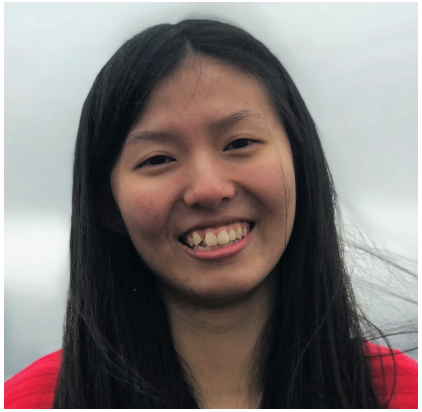
Mary

TAs:

Komal

Alex

Alden



Vicky

Logistics

Lectures: Mon, Wed 9:50-11:10 am Remote

Recitations: Fri 9:50-11:10 am Remote

Office hours:

Day	Time	Location	Staff
Mon	11:15 am-12:15 pm	Remote	Komal
Tues	5:00-6:00 pm	Remote	Vicky
Wed	4:00-5:00 pm	Remote	Aarti
Thurs	6:40-7:30 pm	DH 2315, Remote	Alden
Sat	1:20-2:10 pm	GHC 4401, Remote	Alex

Zoom links: Canvas

Lectures and recitations will be recorded. Strictly for your use only.

Breakout rooms and Office hours will NOT be recorded.

Logistics

In case of technical issues during lecture:

please try logging back in

if issues remains > 5 mins, I will send an email when resolved.

If not resolved, recorded lecture + extra office hours

Webpage: https://www.cs.cmu.edu/~aarti/Class/10315_Fall20

Syllabus, policies, schedule of lectures, recitations, office hours, slides, reading material, homeworks, …

Piazza: http://piazza.com/cmu/fall2020/10315

announcements, questions for Teaching team, discussion forum for students

Homework submission: Gradescope

Grades: Canvas

Expectations

- Remote sessions (lecture, recitation, breakout room, office hours)
  - Turn on video if possible, especially for breakout rooms and office hours
  - Keep yourself muted unless asking or responding to questions

  Interact!
  - Ask questions in class by raising hand or via Zoom chat
  - Respond to questions in class by raising hand or via Zoom chat
  - Anonymous polls on Zoom
  - Breakout rooms and jamboard

  Groups 1-10: Jamboard_1_10

  Groups 11-20: Jamboard_11_20

- In-person Office hours
  - Only attend OH you are assigned to
  - Follow CMU guidance on masks, distancing and physical space, etc.

Recitations

- Strongly recommended
  - Brush up pre-requisites
  - Hands-on exercises
  - Review material (difficult topics, clear misunderstandings, extra new topics, HW and exam solutions)
  - Ask questions

- 1st Probability Review - **FRIDAY**

  by Komal and Alex

  Fri Sept 4 9:50-11:10 am Remote

Grading

- Grading
  - 4 homework assignments (4 x 15% = 60%)
  - 5 QnAs (20%)
  - 1 midterm, 1 final (both in class): (10+10 = 20%)

- Late days
  - total 4 across homeworks and QnAs
  - No partial credit after late days
  - late days are for unforeseen situations (interviews, conference, etc.), do NOT include them in your plan

Homeworks & QnAs

- 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
    - list resources used (references, discussants) on top of submitted homework
- Homeworks are hard, start early ☺
- Due on gradescope at 11:59 pm

Waitlist + Audits + Pass/Fail

- Waitlist

  we'll let everyone in [as long as there is space in room]

  wait to see how many students drop

  keep attending lectures, recitations and office hours virtually and doing HW

- Audits and Pass/Fail

  Audits NOT allowed

  Pass/Fail allowed – talk to your academic advisor

About the course

- Machine Learning **Algorithms and Principles**
  - Classification: Naïve Bayes, Logistic Regression, Neural Networks, Support Vector Machines, k-NN, Decision Trees, Boosting
  - Regression: Linear regression, Kernel regression, Nonparametric regression
  - Unsupervised methods: Kernel density estimation, k-means and hierarchical clustering, PCA, nonlinear dimensionality reduction
  - Core concepts: Probability, Optimization, Theory, Model selection, overfitting, bias-variance tradeoffs …

- See **tentative** lecture schedule on webpage – MAY CHANGE
- Material: Class slides + Reading material

Recommended textbooks

- Textbooks (Recommended, not required):

  Pattern Recognition and Machine Learning, Christopher Bishop (available online)

  Machine Learning: A probabilistic perspective, Kevin Murphy

  (available online)

  Machine Learning, Tom Mitchell

  The elements of statistical learning: Data mining, inference and prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman

Pre-requisites

- **Assume mathematical maturity**
  - Basic Probability and Statistics

    Probability distributions – discrete and continuous, Mean, Variance, Conditional probabilities, Bayes rule, Central limit theorem…
  - Programming (python) and principles of computing
  - Multivariate Calculus

    Derivatives, integrals of multi-variate functions
  - Linear Algebra

    Matrix inversions, eigendecomposition, …

- **Tutorial videos**
  - Probability, Calculus, Functional Analysis, SVD

    https://www.youtube.com/channel/UC7gOYDYEgXG1yIH_rc2LgOw/playlists
  - Linear Algebra

    http://www.cs.cmu.edu/~zkolter/course/linalg/index.html

Related courses

- Related courses – Intro to ML algorithms and principles

  10-301 – Undergrad version for non-SCS majors

  10-601 – Masters version

  10-701 – PhD version

  10-715 – PhD students doing research in machine learning (hardest, most mathematical)

  Other related courses:

  10-606, 10-607 – Math background for ML

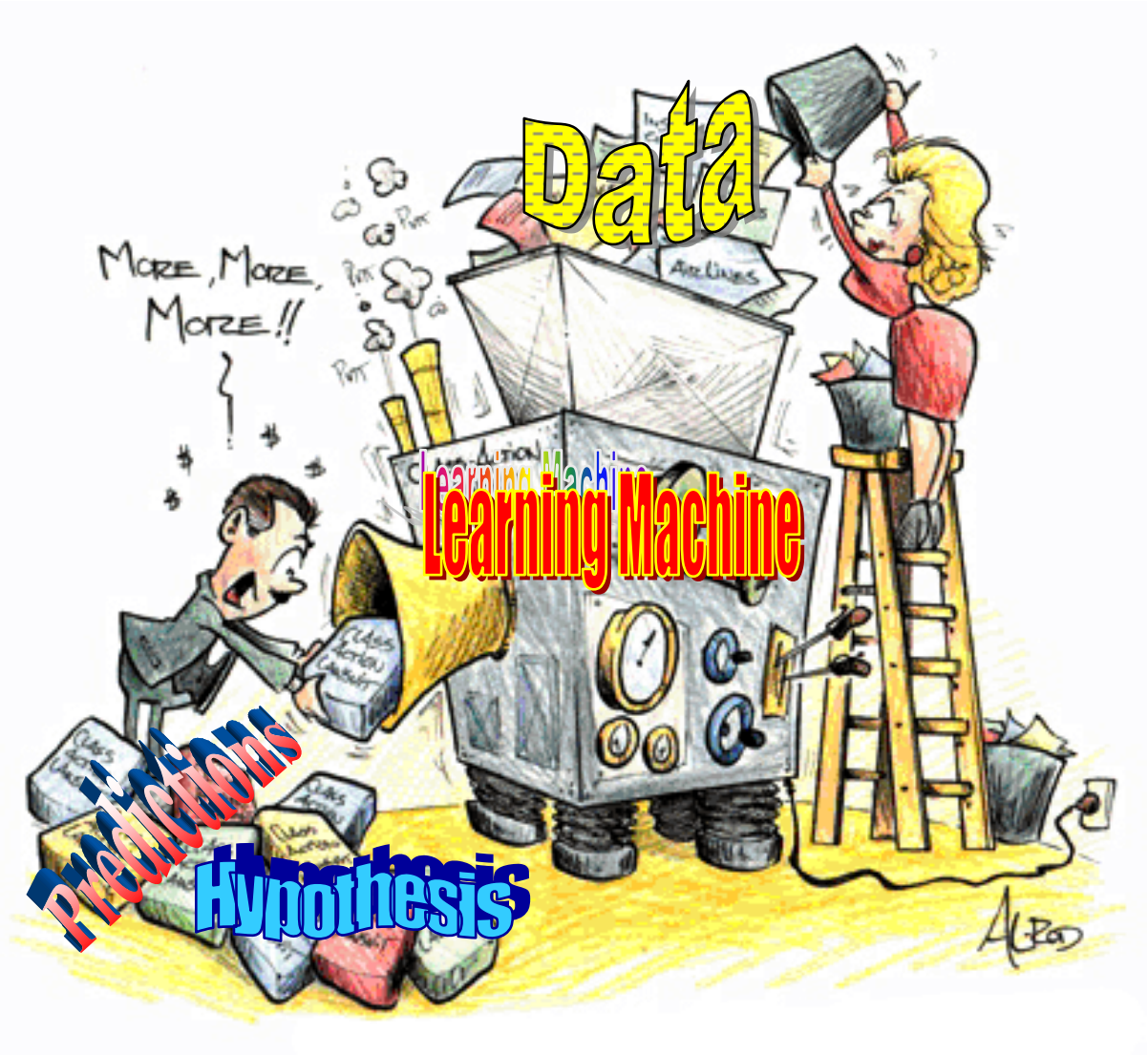
  10-605, 10-805 – Machine Learning with Large Datasets

  11-663 – Machine Learning in Practice (ML software)

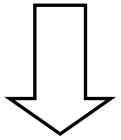
  10-702, 10-704, 10-707, 10-708, 10-709, 15-859(B) – related advanced topics

What is Machine Learning?

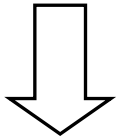
What is Machine Learning?



Data



Learning algorithm



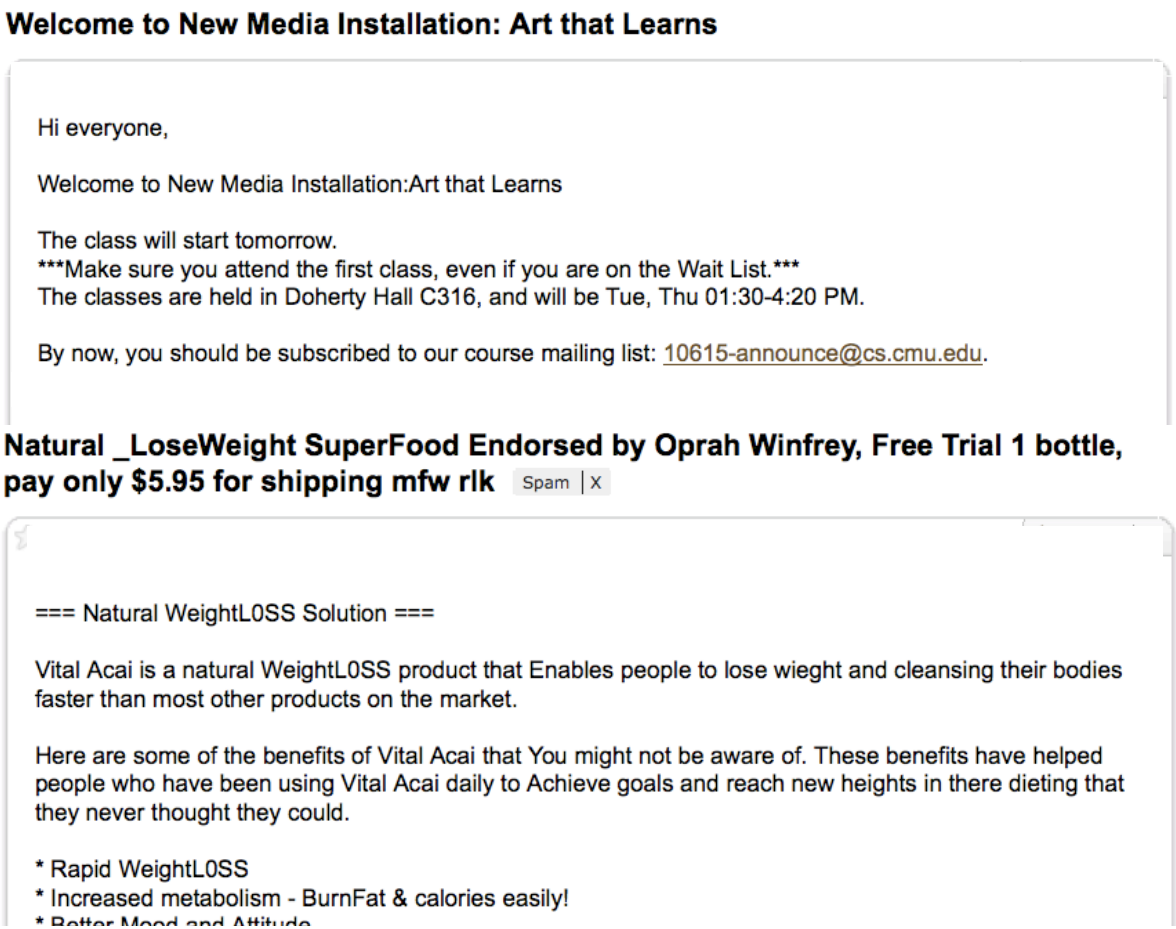
Knowledge

From Data to Knowledge …

Machine Learning in Action

Machine Learning in Action

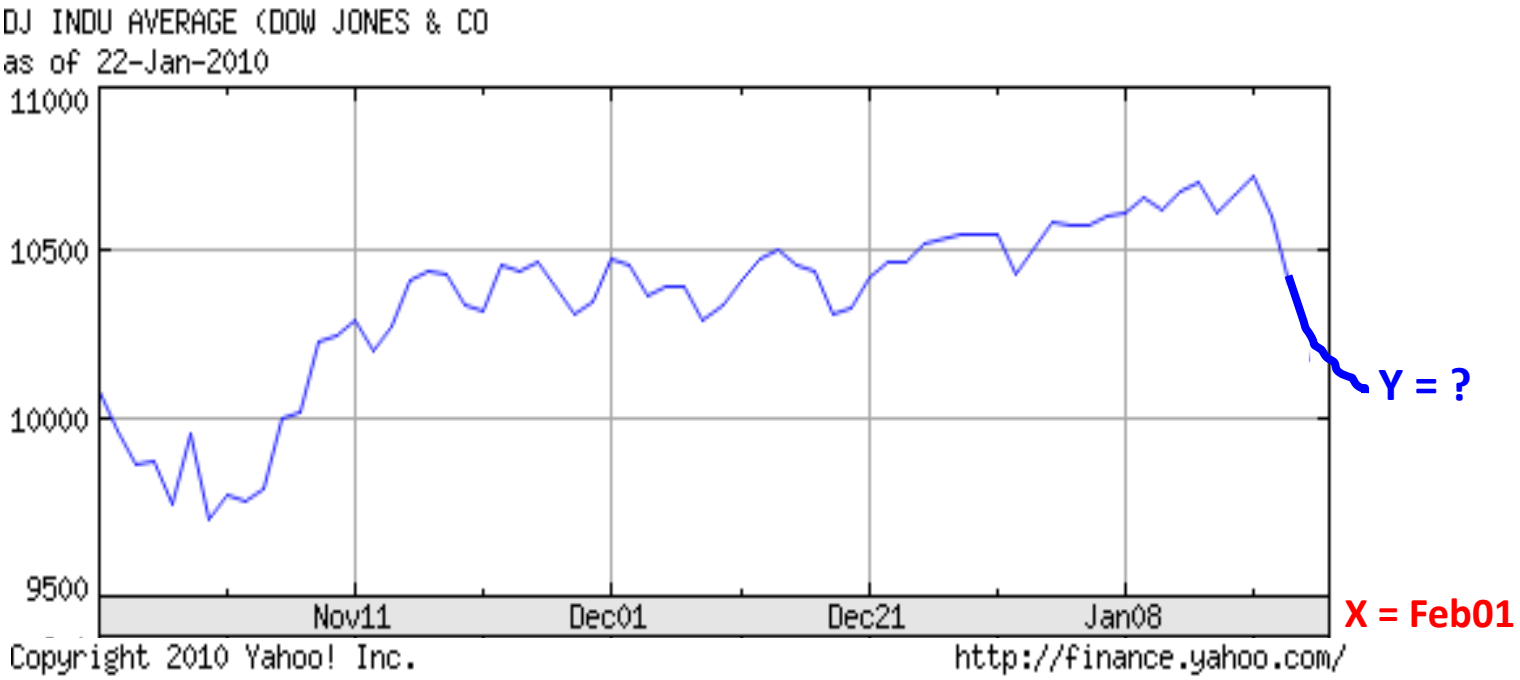
- Spam filtering



Spam/
Not spam

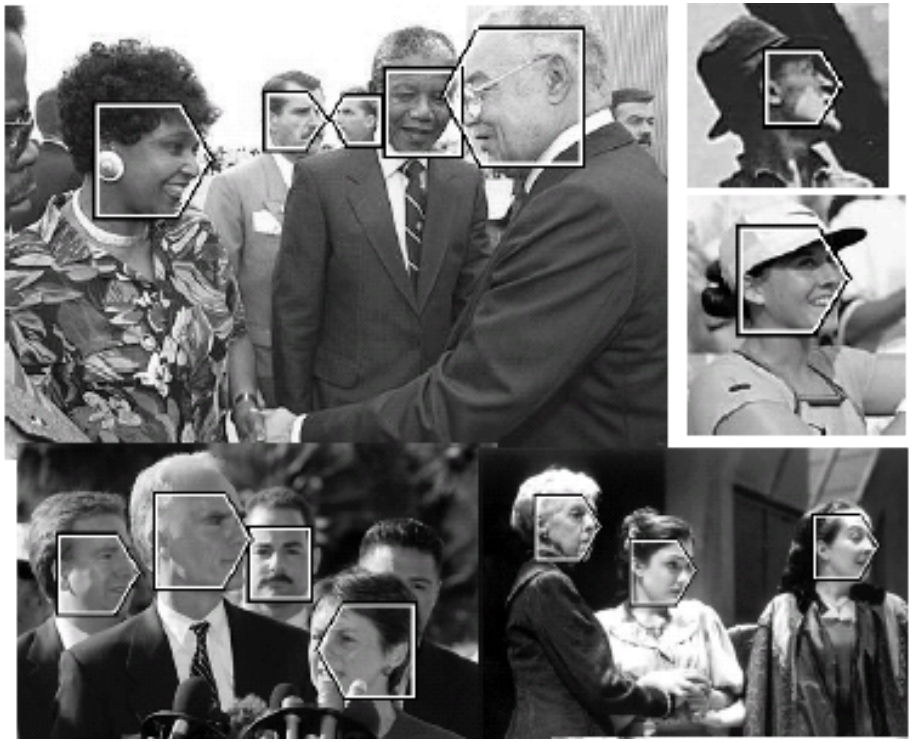
Machine Learning in Action

- Stock Market Prediction



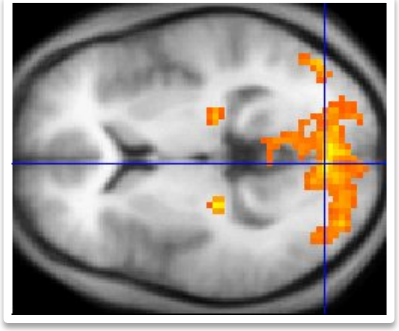
Machine Learning in Action

- Face detection



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

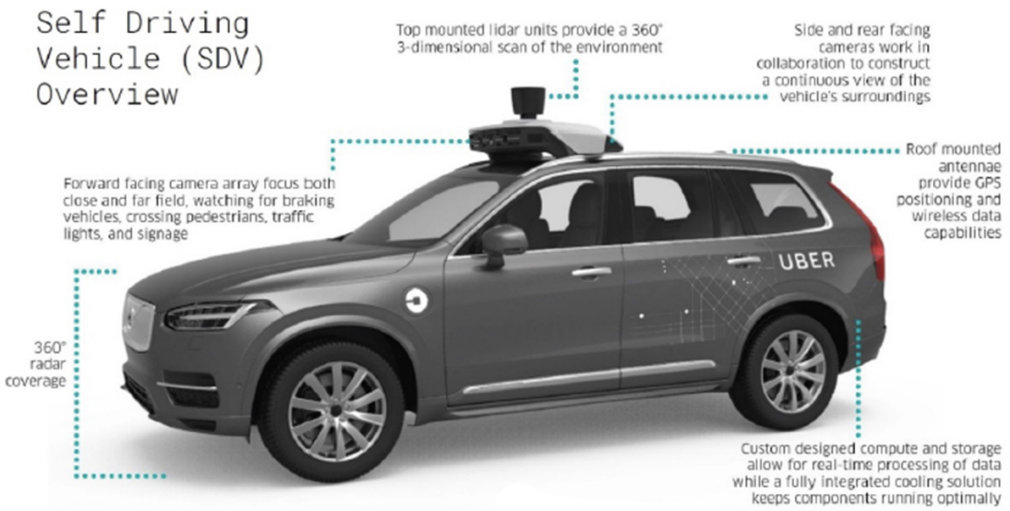
MRI Scans as Courtroom Evidence

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

- Self-driving Cars

Boss, the self-driving SUV
1st place in the DARPA Urban Challenge.
Photo: Tartan Racing.



UBER self-driving car

Photo: IEEE spectrum

Machine Learning in Action

Document classification

Speech recognition, Natural language processing

Computer vision

Robotics

Web forensics

Medical data analysis

Sensor networks

Social networks

Smart buildings

…

Machine Learning in Action

- How have you interacted with ML in your daily life so far?

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
  - 1.6 million images, 1000 object categories
  - 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

"Data scientist: The sexiest job of the 21st century"

(Harvard Business Review)

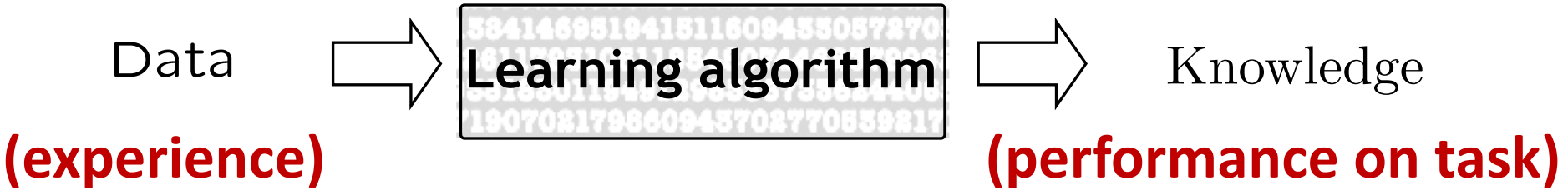
Enjoy!

- 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…

What is Machine Learning?

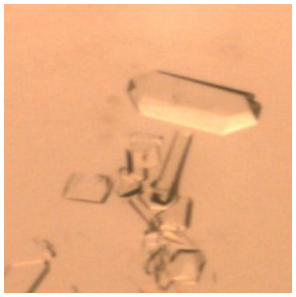
Design and Analysis of algorithms that

- improve their performance
- at some task
- with experience

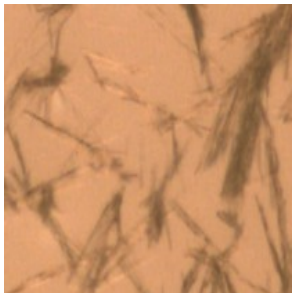


Human learning

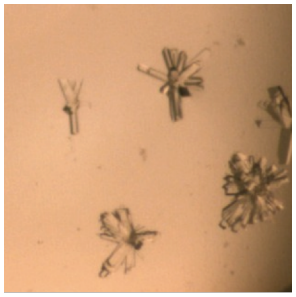
Task: Learning stage of protein crystallization



Crystal

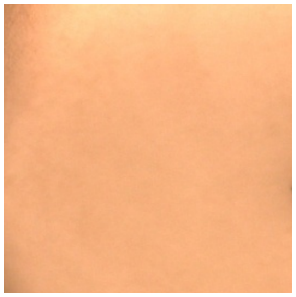


Needle

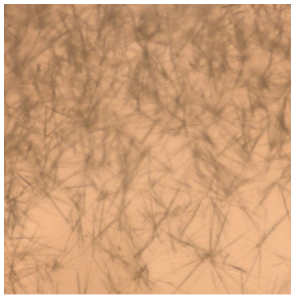


Tree

Tree



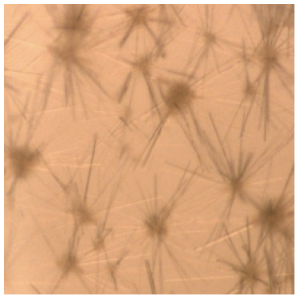
Empty



Needle

Experience

- Predict the label of the test image?



?

Performance

Tasks, Experience, Performance

Tasks, Experience, Performance

Machine Learning Tasks

Broad categories -

- **Supervised learning**

  Classification, Regression

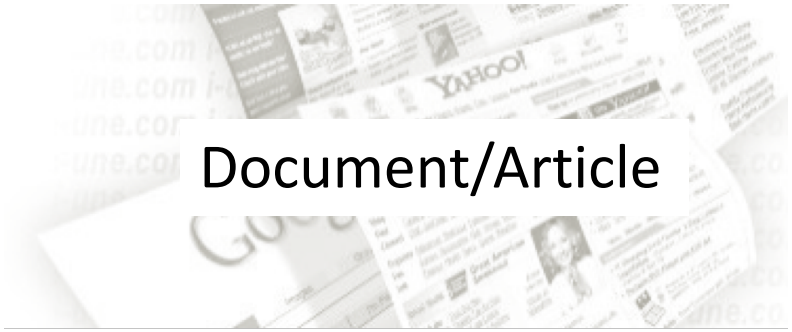
- **Unsupervised learning**

  Density estimation, Clustering, Dimensionality reduction

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

Supervised Learning

Input $X \in \mathcal{X}$



Document/Article

Label $Y \in \mathcal{Y}$

"Sports"
"News"
"Science"
…

Discrete Labels
Classification

Market information

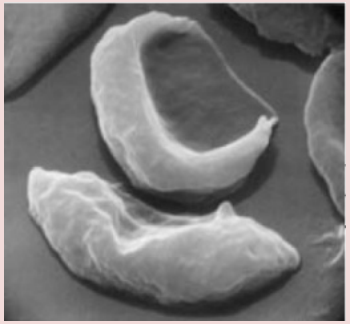
Share Price
"$ 24.50"

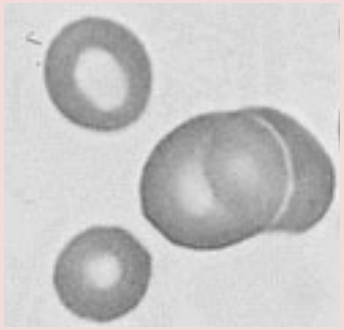
Continuous Labels
Regression

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

$\equiv$ Construct **prediction rule** $f : \mathcal{X} \to \mathcal{Y}$

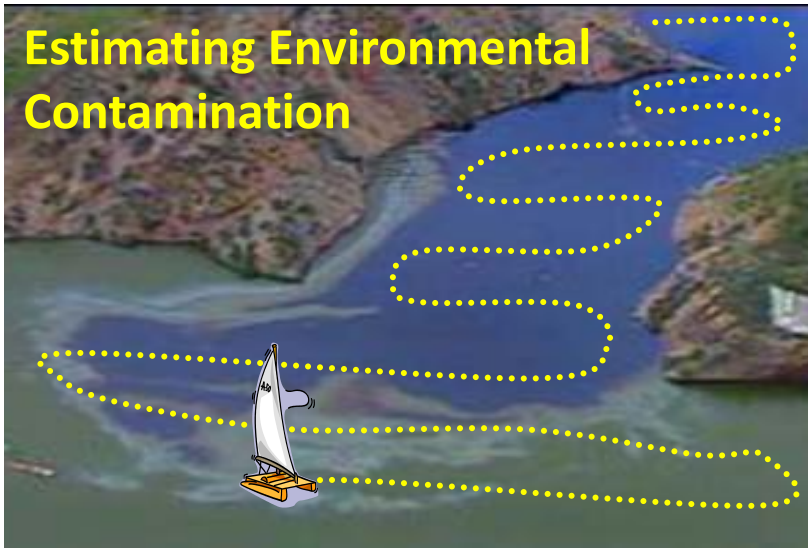
Classification or Regression?

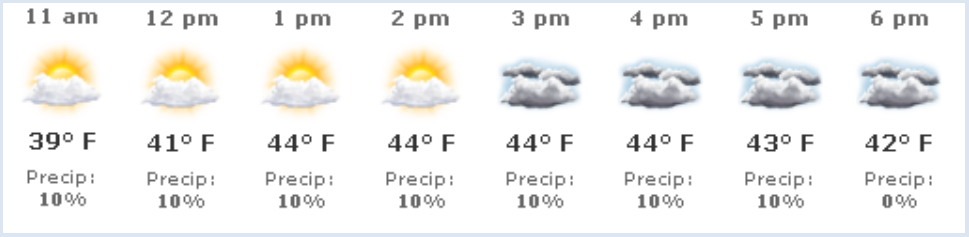




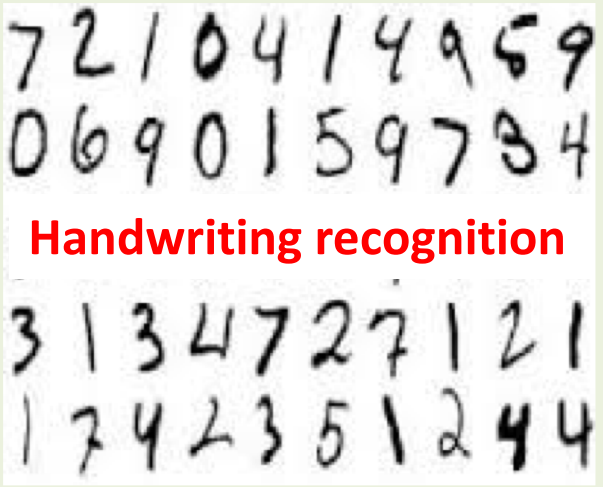
Medical Diagnosis

“Anemic”
“Healthy”





Weather prediction



Handwriting recognition

Unsupervised Learning

Aka "learning without a teacher"

Input $X \in \mathcal{X}$

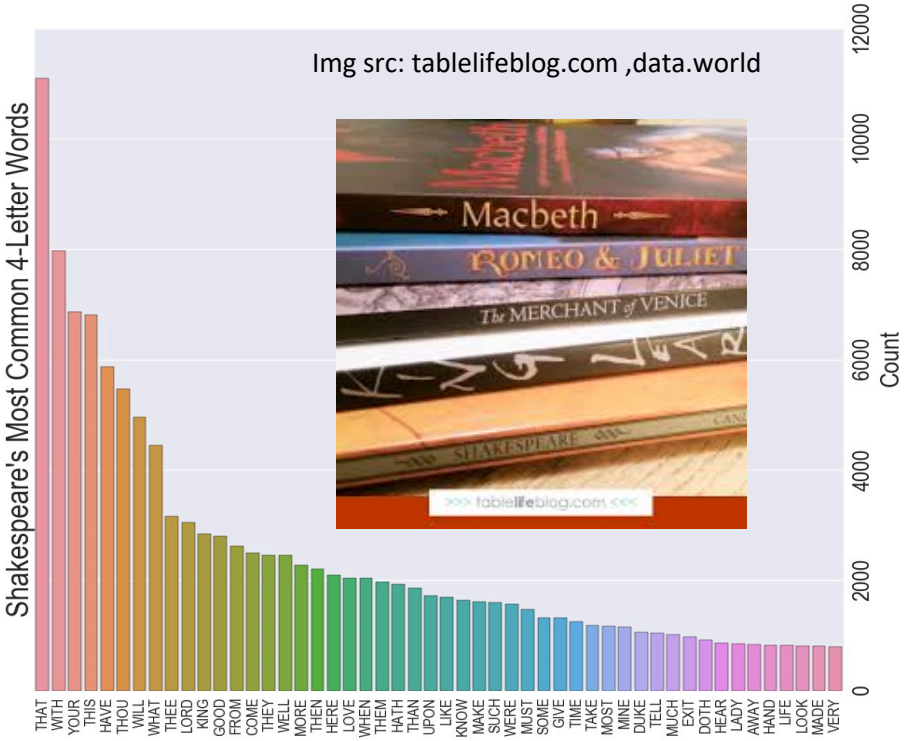
Word distribution
(Probability of a word)

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

Unsupervised Learning

Learning a Distribution

Bias of a coin



Distribution of words in text

- What other distribution would be interesting to learn?