INFSCI 2915: Machine Learning Introduction

Mai Abdelhakim

Department of Informatics and Networked Systems

School of Computing and Information

610 IS Building

Spring 2018

Outline

- Logistics
- Course objectives
- What is Machine learning?
- Applications
- Supervised vs. unsupervised learning
- Regression vs. classification
- Python Introduction (separate slides)

Course

- Courseweb/Blackboard
  - Log in under **INFSCI 2915 Special Topics: Foundations**
  - Please check the website regularly
  - All changes and announcements will be there

- Class meeting: Monday, G31 Benedum Hall

Contact

- Instructor: Mai Abdelhakim
  - PhD from Michigan State University
- Contact me if you have any question or need to discuss anything
  - Email address: maia@pitt.edu
  - Office: 610 IS building
  - **Office Hours**: Thursday 3:00pm - 4:00pm & by appointment
- Graduate Student Assistant
  - Who: TBD
  - Email: TBD
  - **Office hours:** TBD

Course Requirements

- Participation
  - Class exercises
  - Discussion board:
    - Create threads/forums to ask questions about machine learning or python
    - Contribute to existing forums by commenting/answering questions
    - Check discussion board regularly
- Assignments
  - Submission on courseweb
  - Late assignments
    - will be penalized
    - not accepted more than one week late
- Midterm
- Final exam
- Final Project

Grading

- Assignment & Participation 40%
- Midterm 20%
- Final exam and final project: 40%

Grading Policy:

- Your work must be your own!
- No credits for vague answers

Course Objectives

- Explain concepts, process, and algorithms of machine learning
- Enables you to differentiate between different machine learning algorithms
- Assess the performance of learning algorithms
- Describe best practices in applying machine learning
- Apply machine learning algorithms with python

Textbook & References

- **Textbook:**
  An Introduction to Statistical Learning: with Applications in R , by James Gareth et al., 2013
  Available online: http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf
- **Introduction to Machine Learning with Python**, by Andreas Müller et al., 2016
- **Python Machine Learning**, by Sebastian Raschka, 2015
- **Pattern Recognition and Machine Learning,** by Christopher Bishop, 2006.
- **Elements of Statistical Learning**, by Trevor Hastie et al.
- **Pattern Classification,** Richard Duda et al.
- Additional reading may be posted

Prerequisites

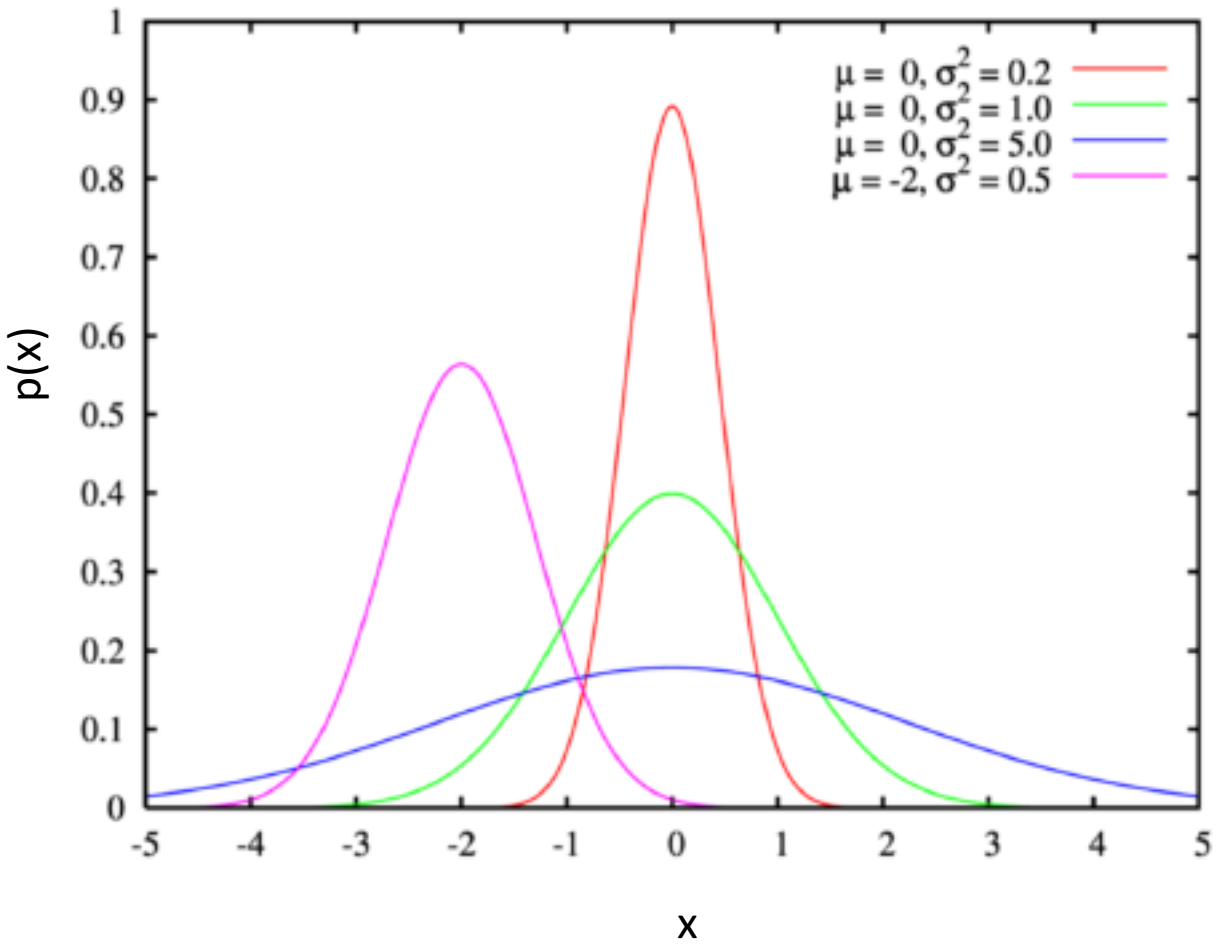
- Probability theory
- Statistics
- Calculus
- Linear algebra

Probability Theory - Basics

- **Random variable**: Uncertainty about the outcome
- **Probability**: measures the likelihood / frequency of occurrence of a random variable
  - Value between [0,1]
- **Expected value/mean:** average value of a random variable
- **Variance:** measures the deviation from the mean value

Probability Theory - Basics

- **Probability distribution**
  - **Gaussian/ Normal distribution** is most common
    - Fully characterized by mean and variance
- **Conditional probability**: Given that an event has occurred, what is the probability that another event will occur



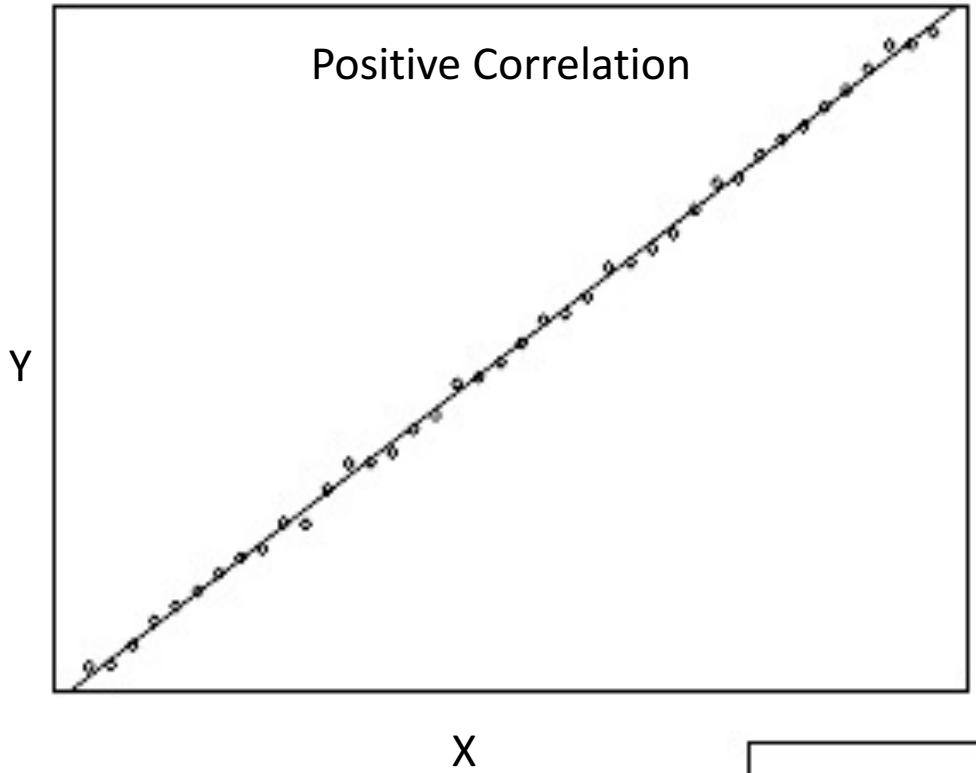
Normal distribution, mean μ and variance σ^2

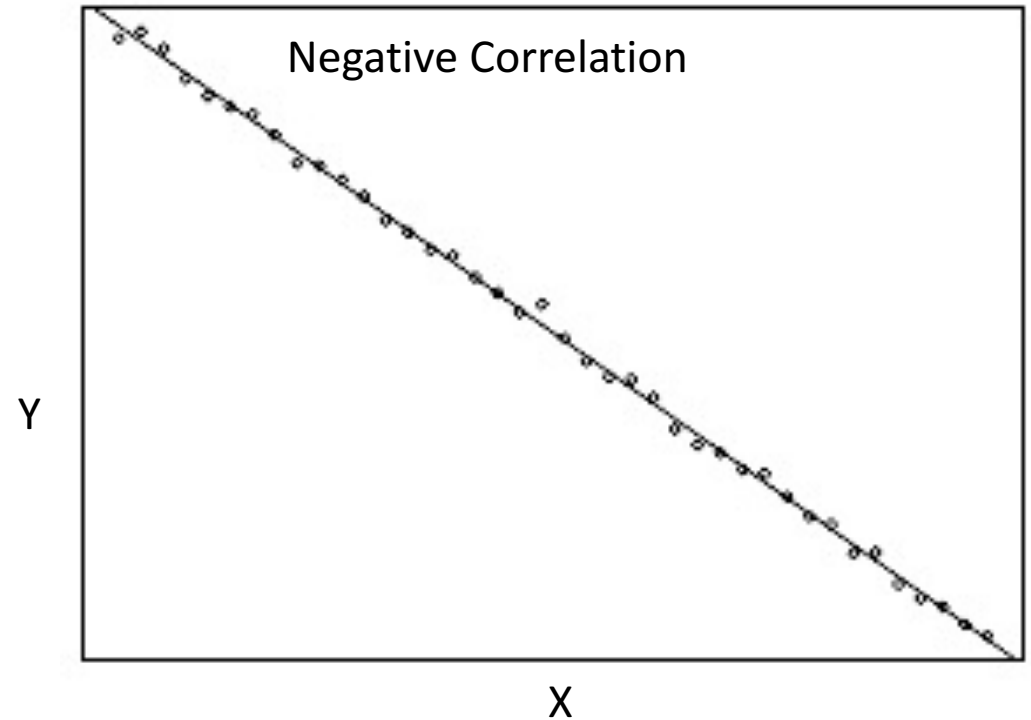
Probability Theory - Basics

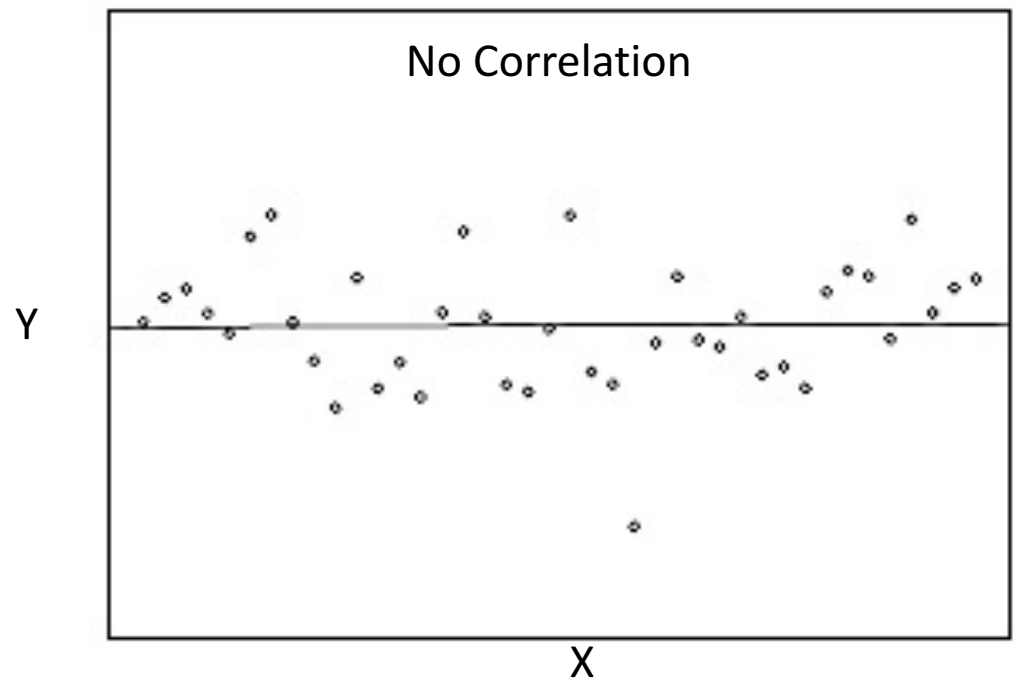
- **Correlation** between two variables describes how strong they are associated with each other
- Measured by covariance matrix or correlation coefficient
  - Covariance between two variables X, Y: COV(X,Y) = E[(X-mean(X))(Y-mean(Y)]
  - Correlation coefficient between X, Y ($\rho_{x,y}$) Is a value in the range of [-1,1]

$$\rho_{x,y} = \frac{COV(X,Y)}{\sqrt{VAR(X)}\sqrt{VAR(Y)}}$$

  - If $\rho_{x,y}$=0, then there is no correlation
  - If $\rho_{x,y}$=1, then there is a positive correlation
  - If $\rho_{x,y}$=-1, then there is a negative correlation


Positive Correlation
Y
X


Negative Correlation
Y
X


No Correlation
Y
X

What is Machine Learning?

- Field of study that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959).
  - Arthur Samuel, Stanford University, Pioneer in artificial intelligence & computer gaming

What is Machine Learning?

- Subfield of artificial intelligence
- Intersection of computer science and statistics
  - Statistics make conclusions from data, and estimate reliability of conclusions
  - Computer science: ability to solve problems, large-scale computing
- How can we build computer system that learn and improve with experience?
- Machine learns with respect to a particular **task T**, **performance metric P** and **experience E**, if the **performance** P on task T **improves with experience E**.

Tom Mitchell, The Discipline of Machine Learning

Why Machine Learning is Important

- Used across industries to improve efficiency, productivity, flexibility, safety and create new business models

- Huge impact on the economy

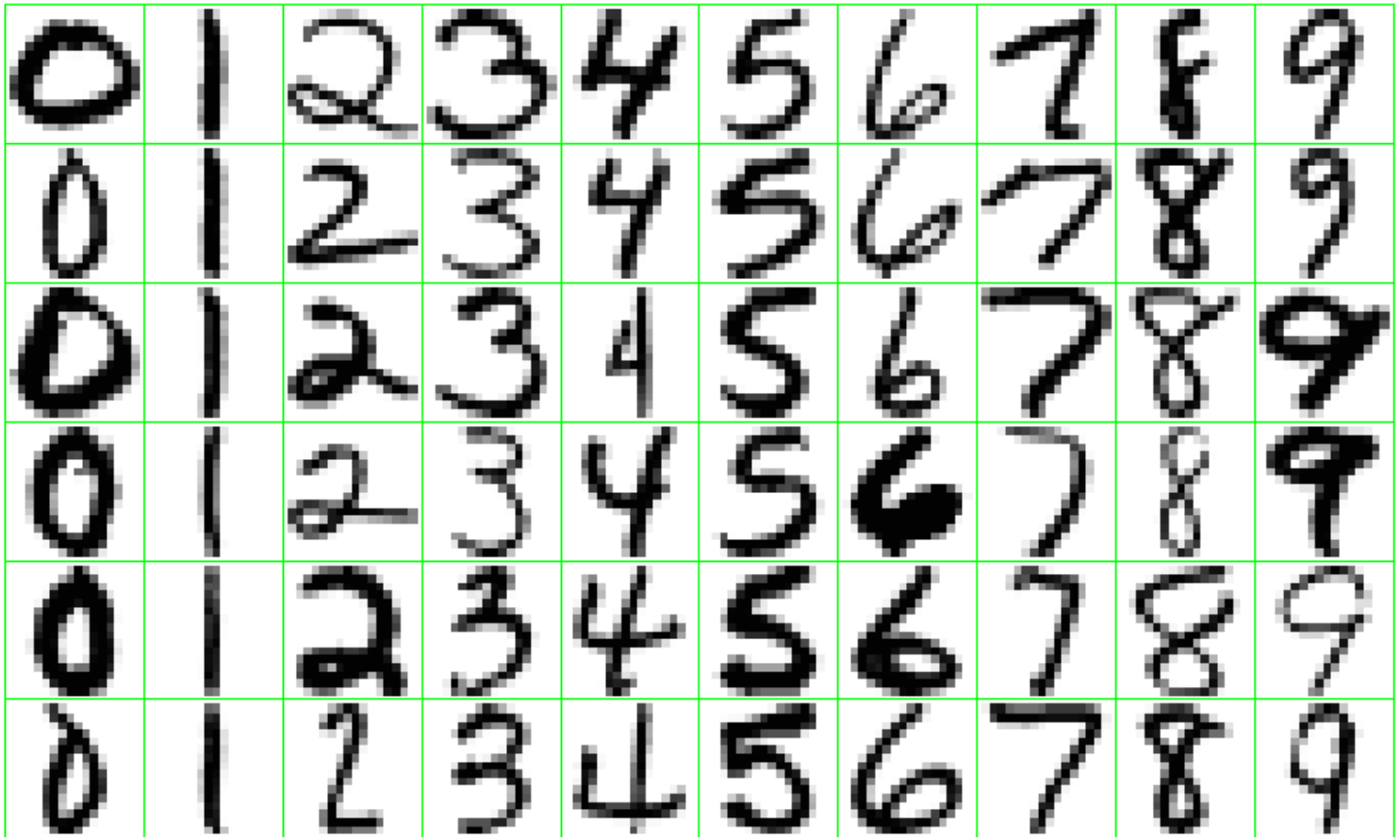
Applications

- Email Spam Detection
- Input features: relative frequency of most commonly occurring words, punctuation marks

	free	!	edu
Spam	0.52	0.51	0.01
Not spam	0.07	0.11	0.29

Applications – Cont.

- Computer vision: e.g., pattern recognition,
  - US post office: automatically sort letters containing handwritten addresses



MNIST database: large database of handwritten digits

Applications – Cont.

- Digital personal assistants (such as apple's Siri)
  - Speech recognition system
    - Audio signal to output text

Where is IS building?

Applications – Cont.

- Robot systems: e.g. Autonomous driving – use real-time image recognition and video processing
- Learning capabilities make them more capable, flexible, and safer

Applications – Cont.

- Health applications:
  - Drug design and discovery, find tumors in medical images that are hard to detect
  - **Bio–surveillance** - **detect and track disease** (track emergency room admission reports, purchases over-the-counter medicines)

- Google's DeepMind and University of Oxford used machine learning to create a **lip-reading system**

- eCommerce**:**
  - Product recommendations (Netflix, amazon)
    - Netflix prize: http://www.netflixprize.com/

Tom Mitchell,
Andrew Ng, Machine Learning

Machine Learning in Industries

Healthcare

Diagnose disease, Predict personalized health outcomes

Automotive

Autonomous Driving,

Navigation

Finance

Identify fraudulent transactions, approve loans

Media

Personalized advertising

Manufacturing

Automation, predictive maintenance

Agriculture

Personalized crops to individual conditions

Network security

Detect and identify attacks/hacks

Machine Learning Algorithms

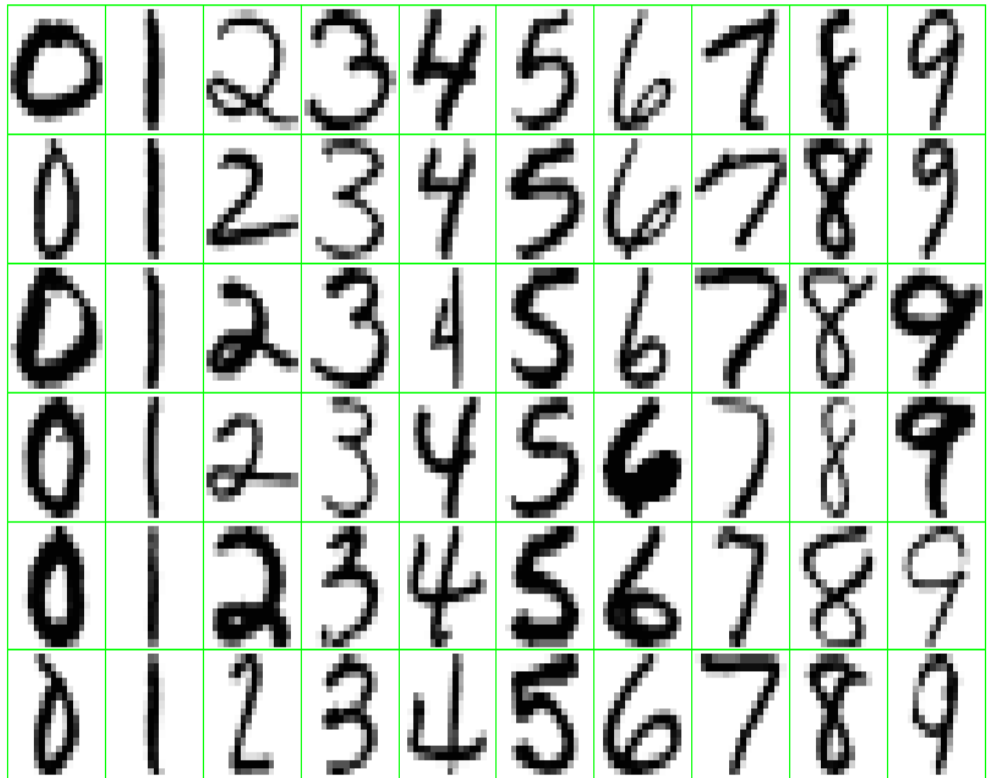
- **Supervised Learning**: Learn to predict from labeled data
- **Unsupervised learning**: Find structure in unlabeled data
- Others: Reinforcement learning

0----zero
1----one
provide 0 is zero,given answer

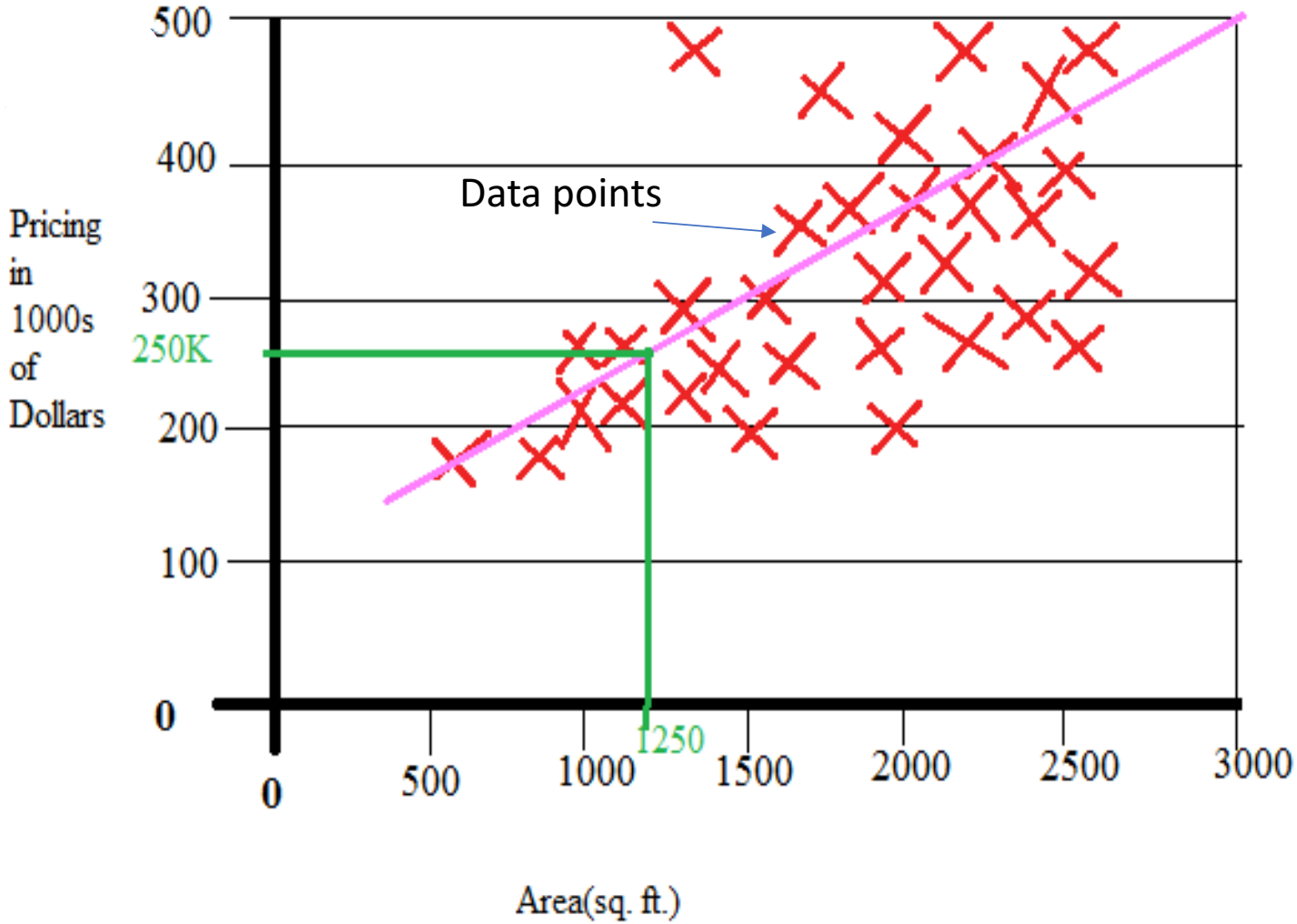
Supervised Learning Examples

Training data contains the labels (that needs to be predicted for new examples)

Handwritten digit predictions

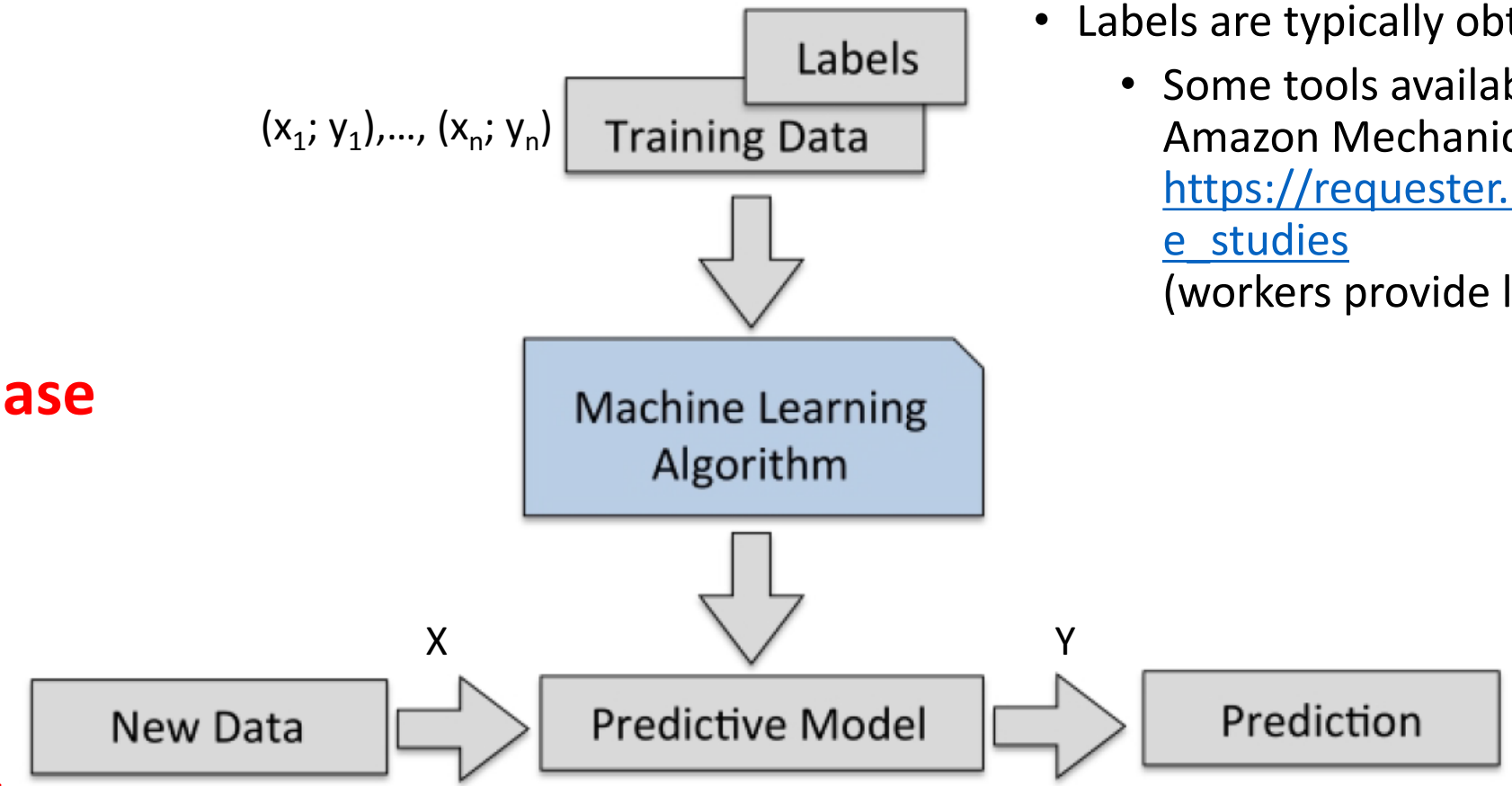


House price prediction



Supervised Learning

- Learn using **labeled data (correct answers are given in learning phase)**
- Then, make predictions of previously unseen data

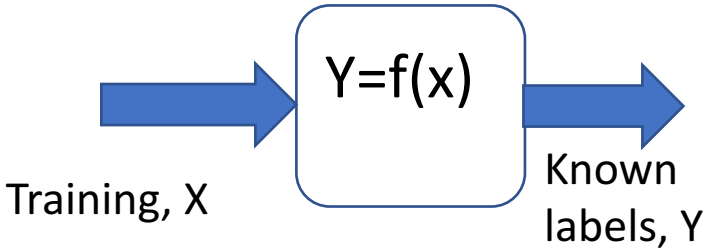


- Labels are typically obtained manually
  - Some tools available, e.g. Amazon Mechanical Turk: https://requester.mturk.com/case_studies (workers provide labels)

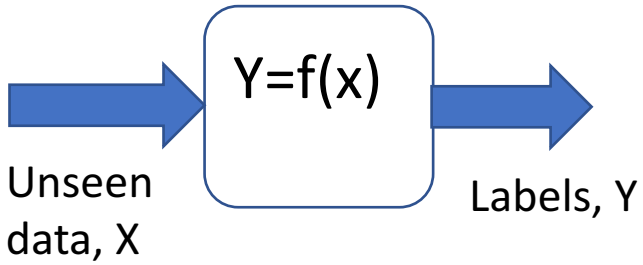
Learning phase

Prediction phase

Training Phase: using labelled examples (training), the model learns, i.e., obtain function f, where Y=f(x)



Prediction phase: trained model is used to predict labels for previously unseen data.
Estimate Y for new X



Variables

- Outcome measurement **Y**
  - Also called **label, target value, response, dependent variable**
- Input **features** vector of length P **(P features)**
  - Also called **predictors, inputs, independent variables**
  - Selection of features has a huge impact on the machine learning algorithms, and depends on the application
    - Example: pixel values, time, location, area
- We have **n training data** instances: also called **observations**, **data points**
  - Supervised learning: Input feature and output label pair: $(x_1; y_1),\ldots, (x_n; y_n)$, where $x_i=\begin{pmatrix} x_{i,1} \\ \cdot \\ x_{i,p} \end{pmatrix}$
  - Unsupervised learning: Input features only: $x_i=\begin{pmatrix} x_{i,1} \\ \cdot \\ x_{i,p} \end{pmatrix}$

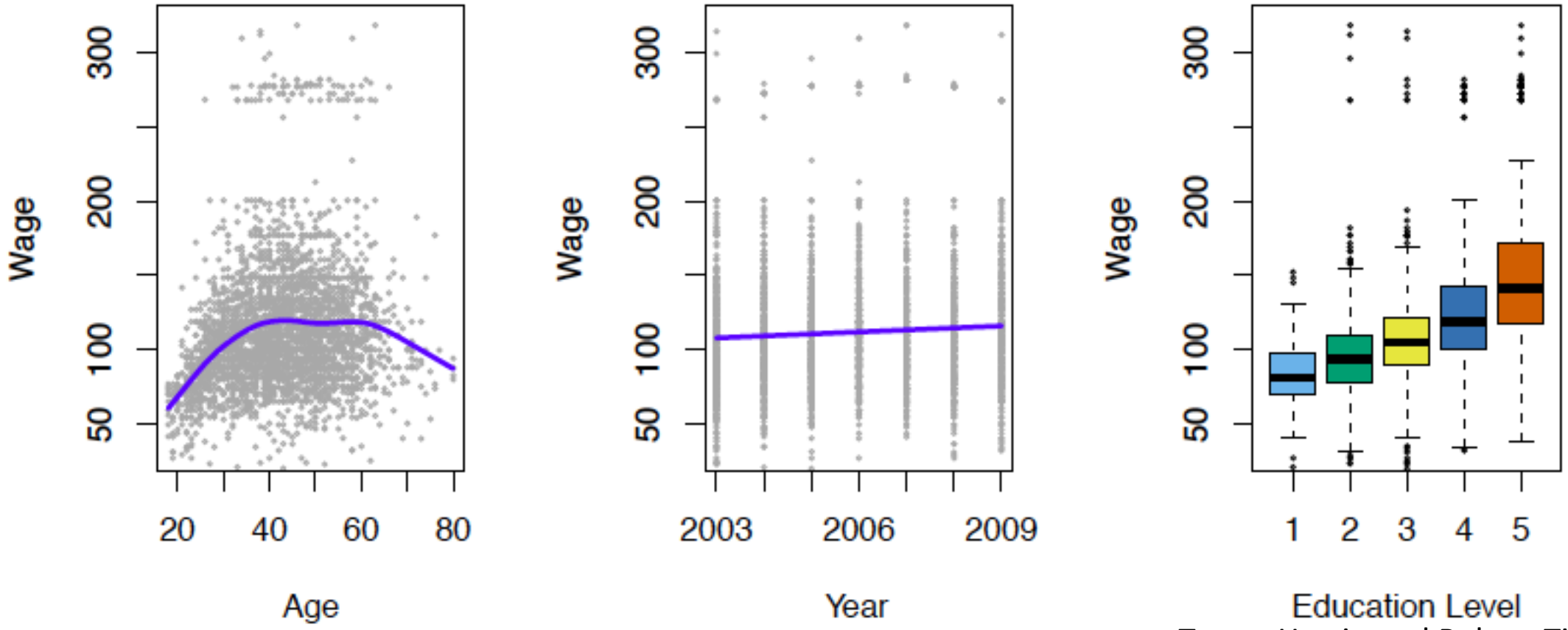
Supervised Learning

- Learn to predict target values from labeled data (Y available)

- Two types of problems
  - **Regression**: Target values (Y) are continuous/quantitative
    - E.g. price, wage, blood pressure
  - **Classification**: Target values (Y) are discrete/finite/qualitative
    - E.g. gender, digits 0-9, cancer type

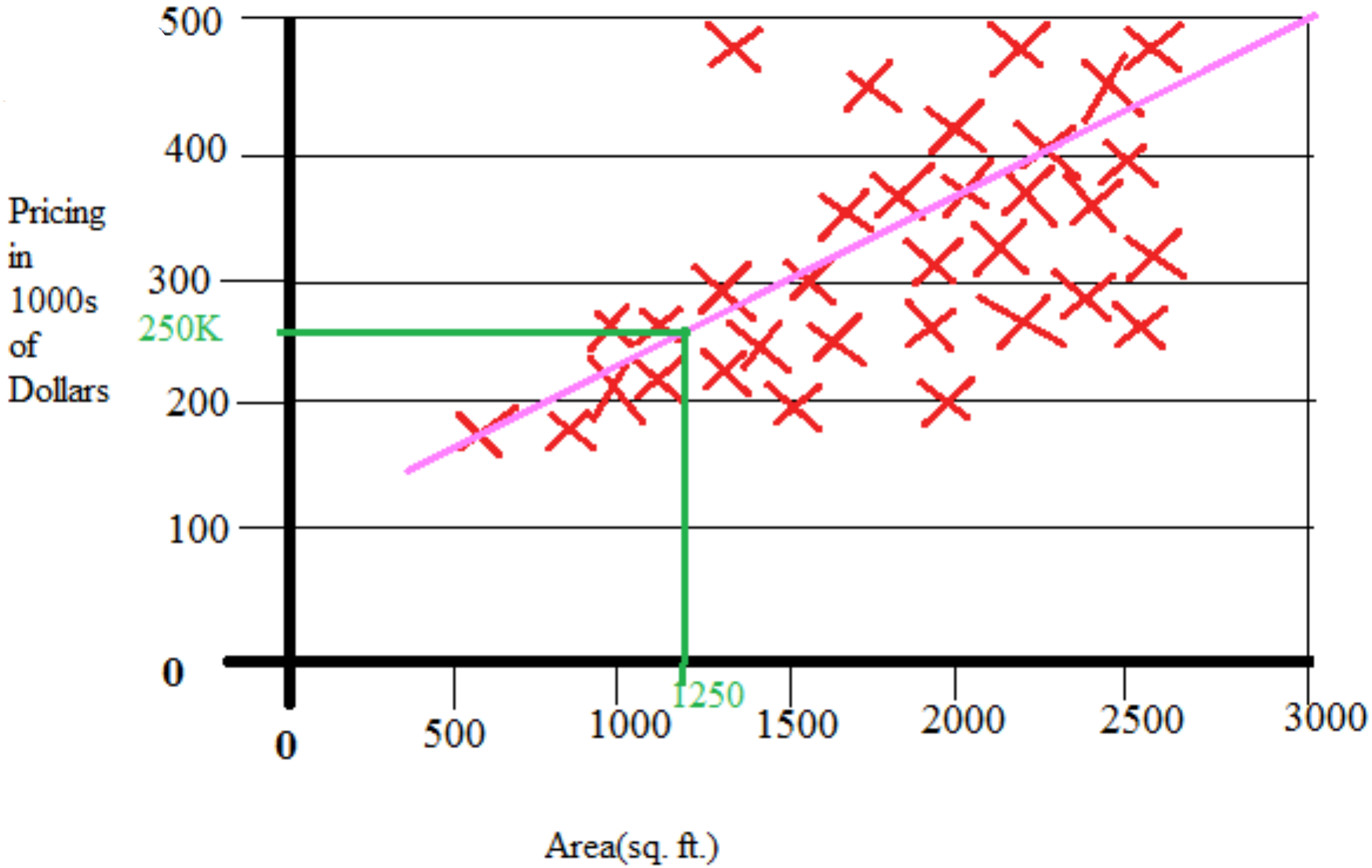
Supervised Learning – Regression Example

- Income survey data from the central Atlantic region-USA.
  - Label (Y): wage Features (X): Age, Year & Education level
  - What is the association between Y and X



Supervised Learning – Regression Example

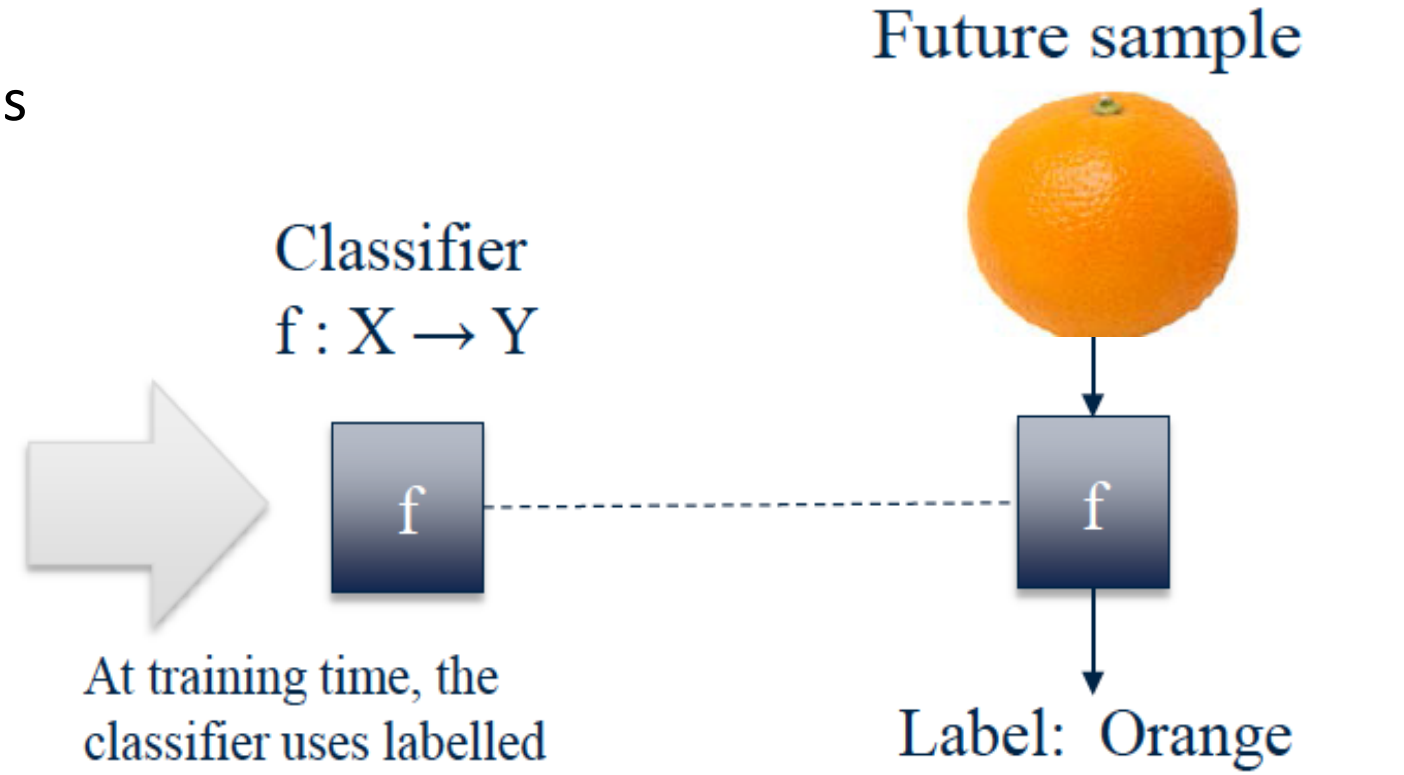
Predicting house price



Supervised Learning – Classification Example

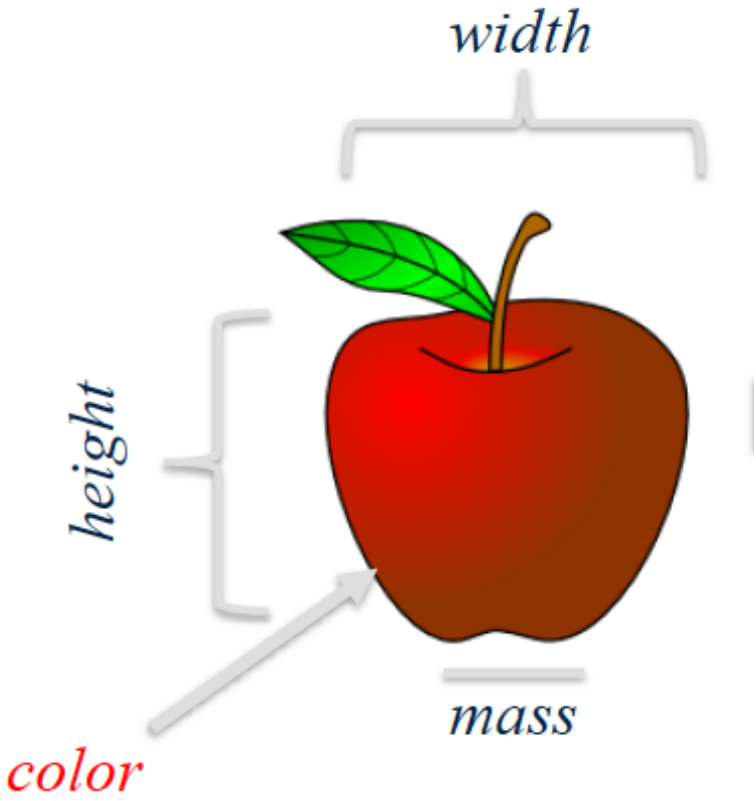
Fruit dataset: Apples, lemon, oranges

X Sample		Y Target Value (Label)	
	x_1	Apple	y_1
	x_2	Lemon	y_2
	x_3	Apple	y_3
	x_4	Orange	y_4



Supervised Learning – Classification Example Feature Representation

How to represent an observation?



Feature representation X

mass	width	height	color_score
162	7.5	7.1	0.83

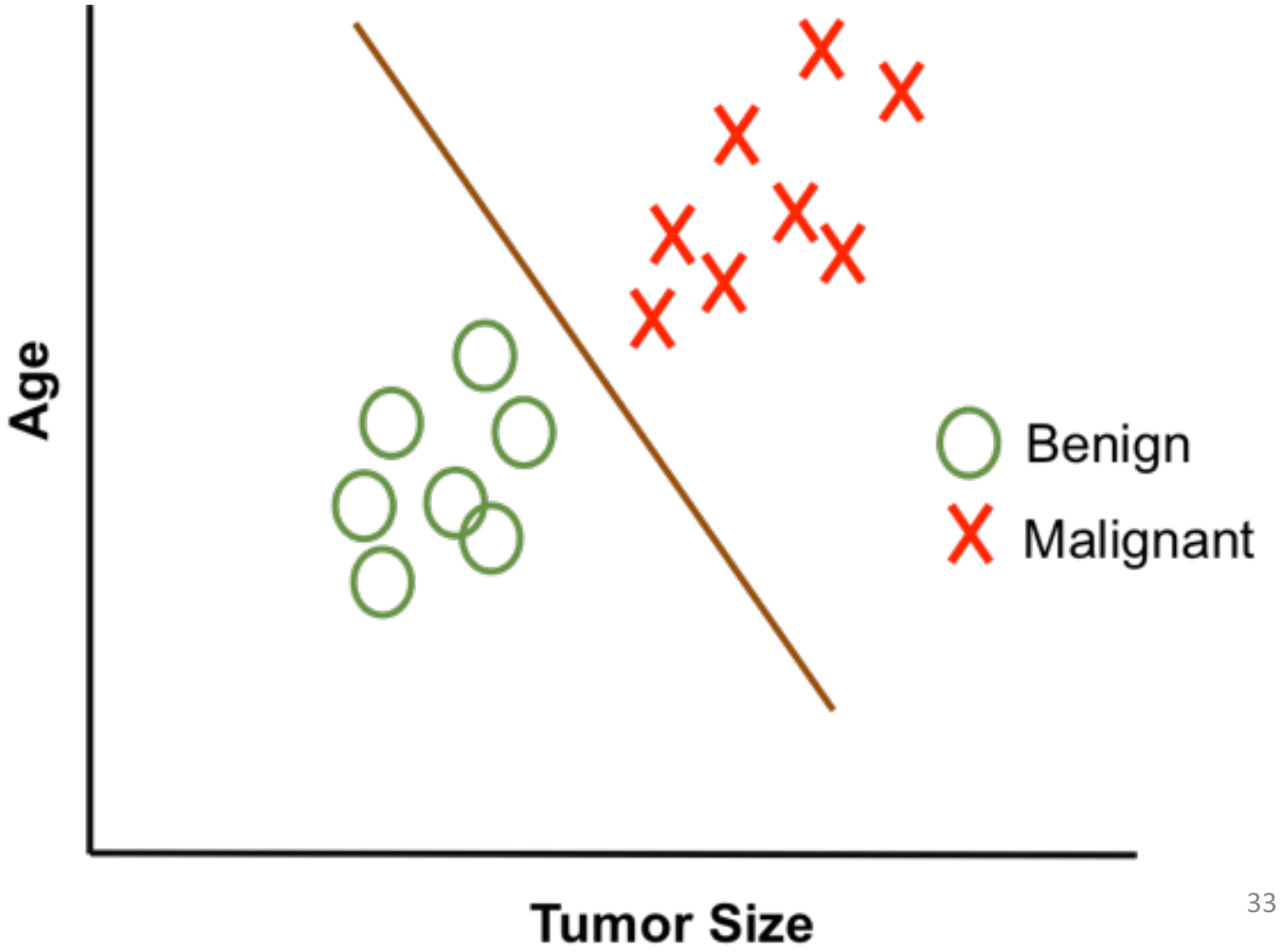
Classifier

Predicted class (apple)

Supervised Learning – Classification Example

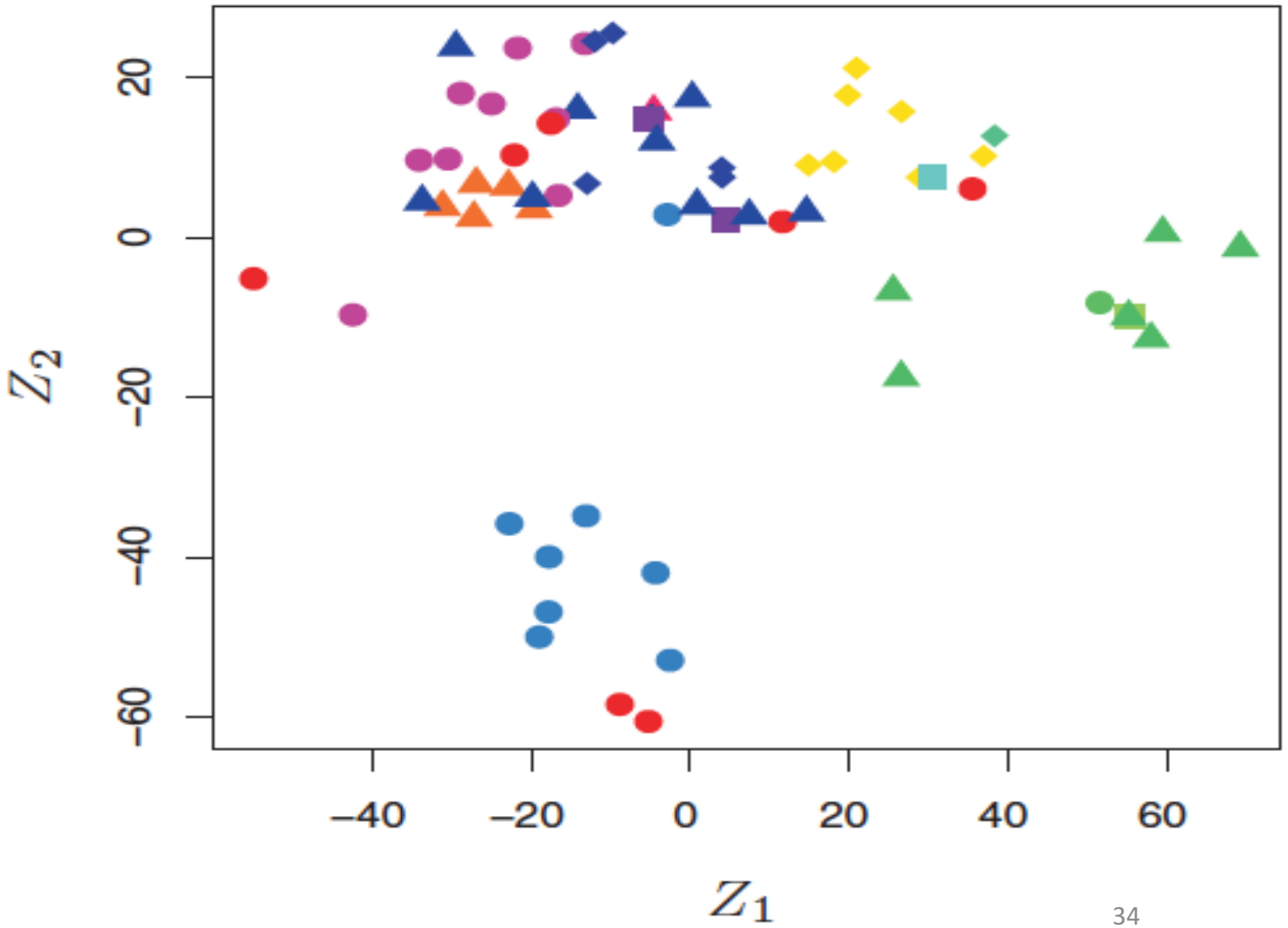
- Cancer classification example.
- Binary classification
  - Benign or Malignant cancer
  - Features: tumor size, age

Figure shows feature space with classification label



Supervised Learning – Classification Example

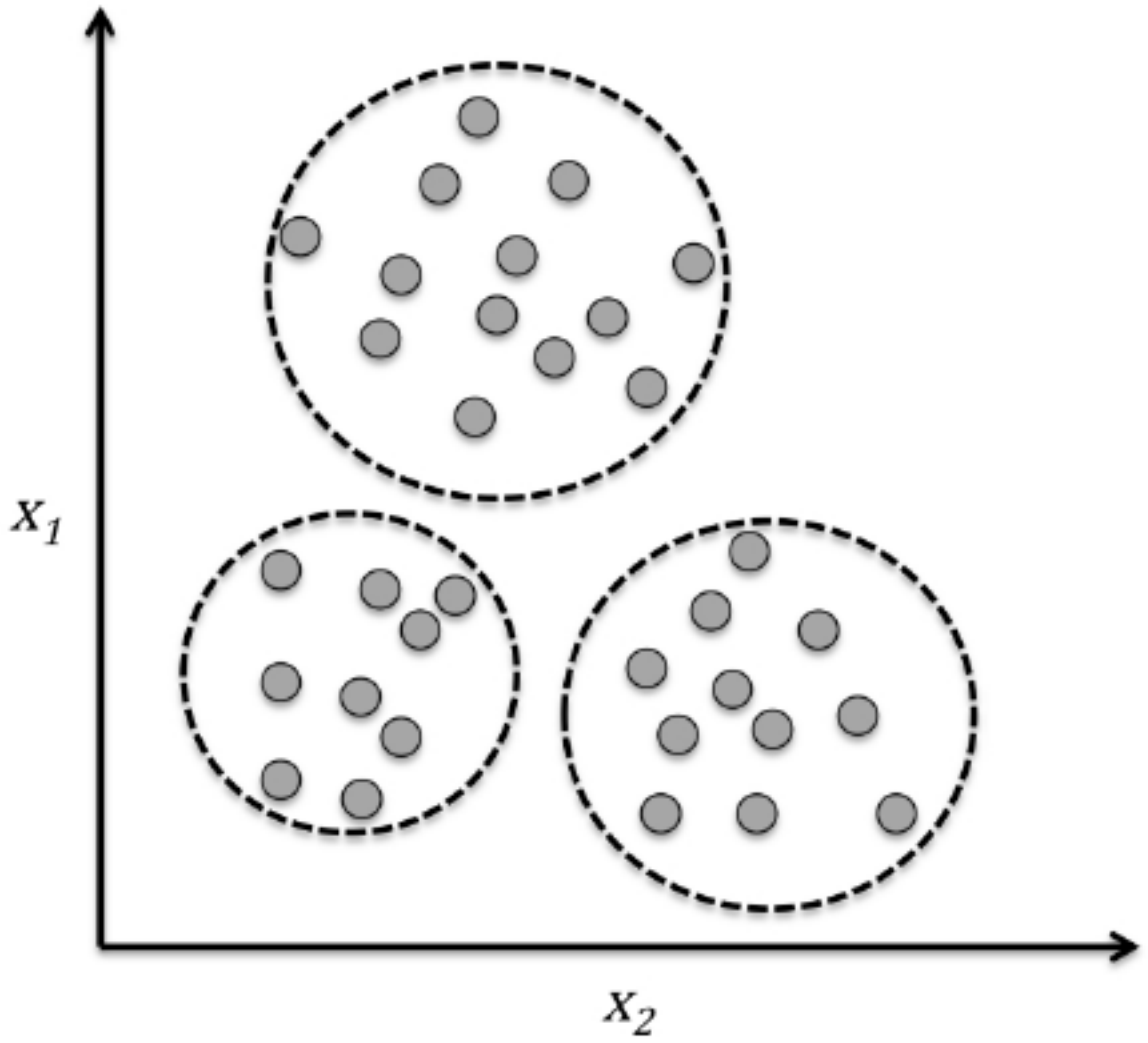
- Gene expression measurement for different cancer cell lines **classify cancer class**
  - From NCI60 dataset - National Cancer Institute
  - Using two features (2 principle components)



Gareth James et al.

Unsupervised Learning

- No labels
- Arrange data into clusters (similar groups)
- Difficult to evaluate

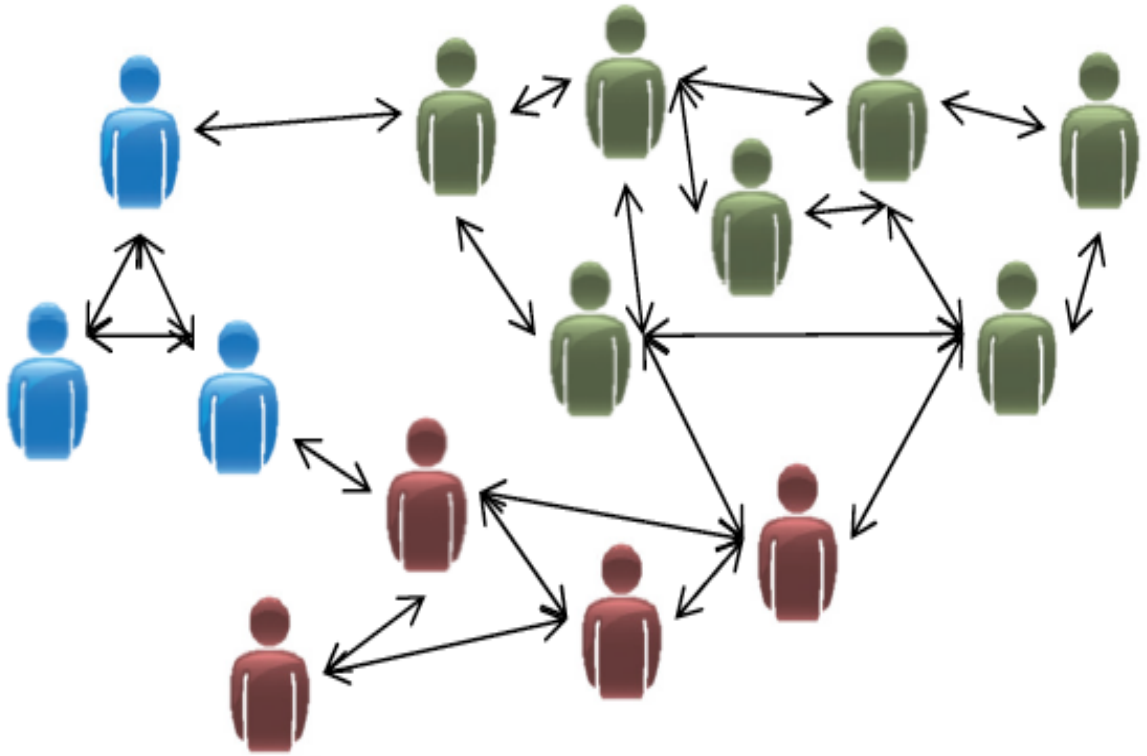


Unsupervised Learning

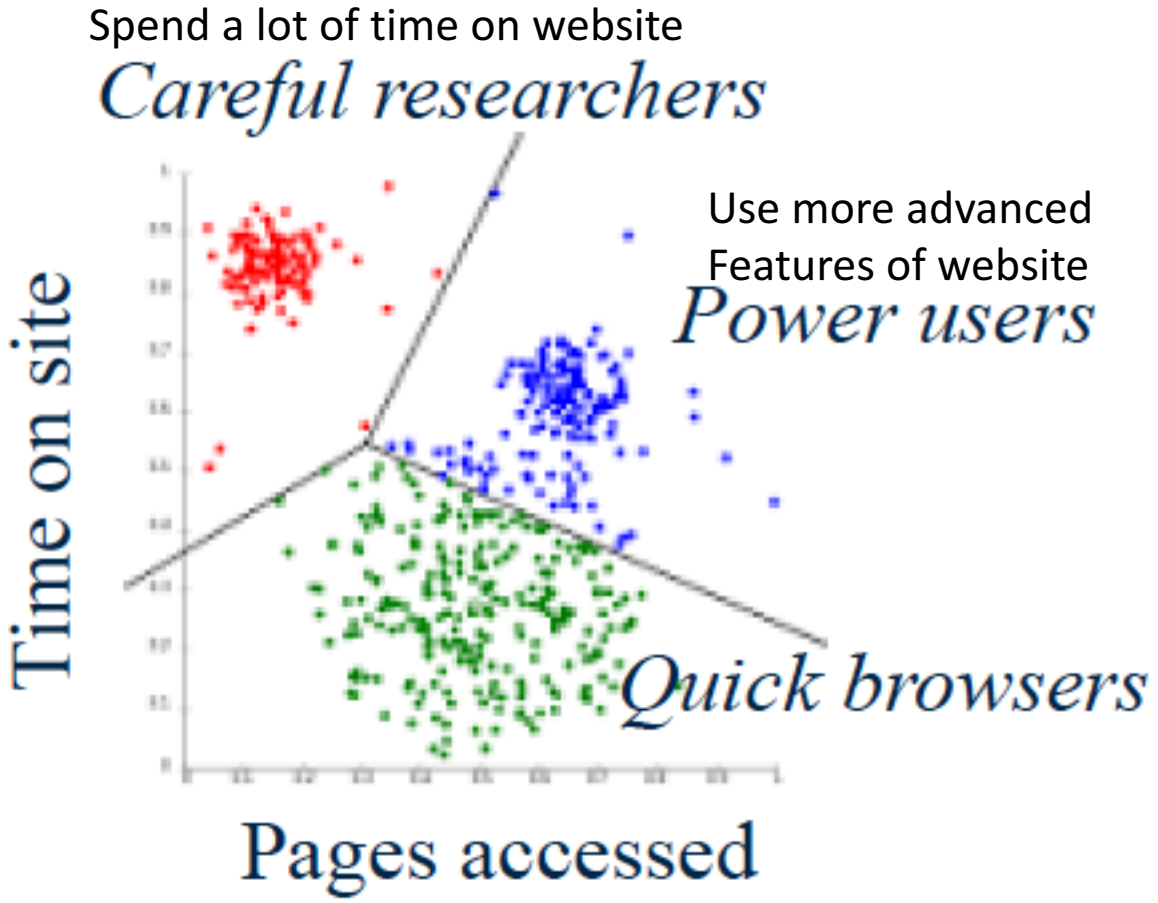
- Training samples are unlabeled
- Objective: find similarities/groups

Unsupervised Learning

- Clustering analysis
- Finding groups of similar users



Social network analysis

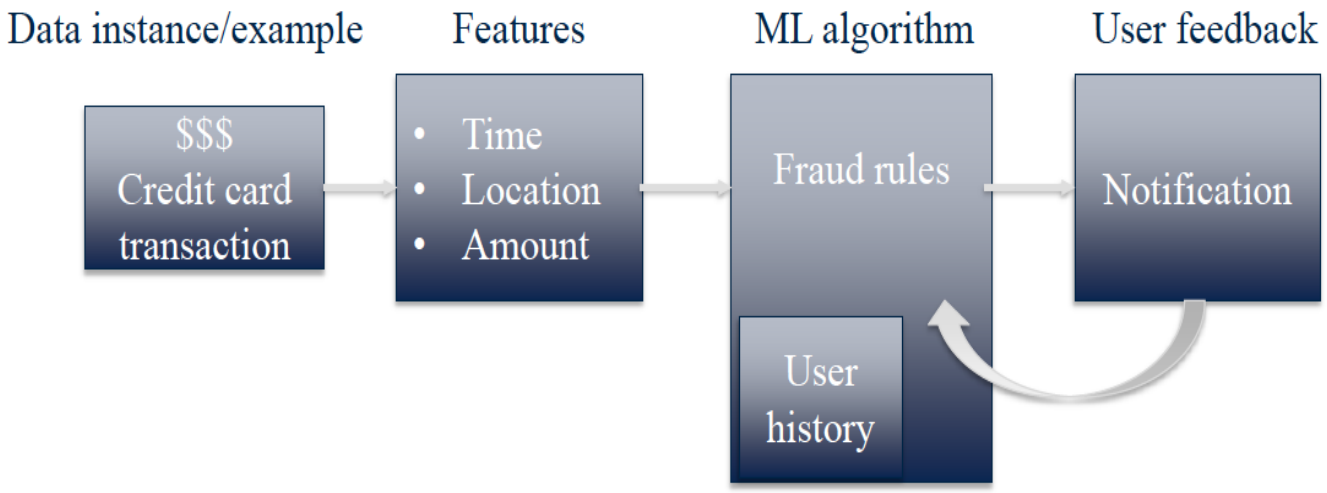


e-commerce example: Tailor website for each group

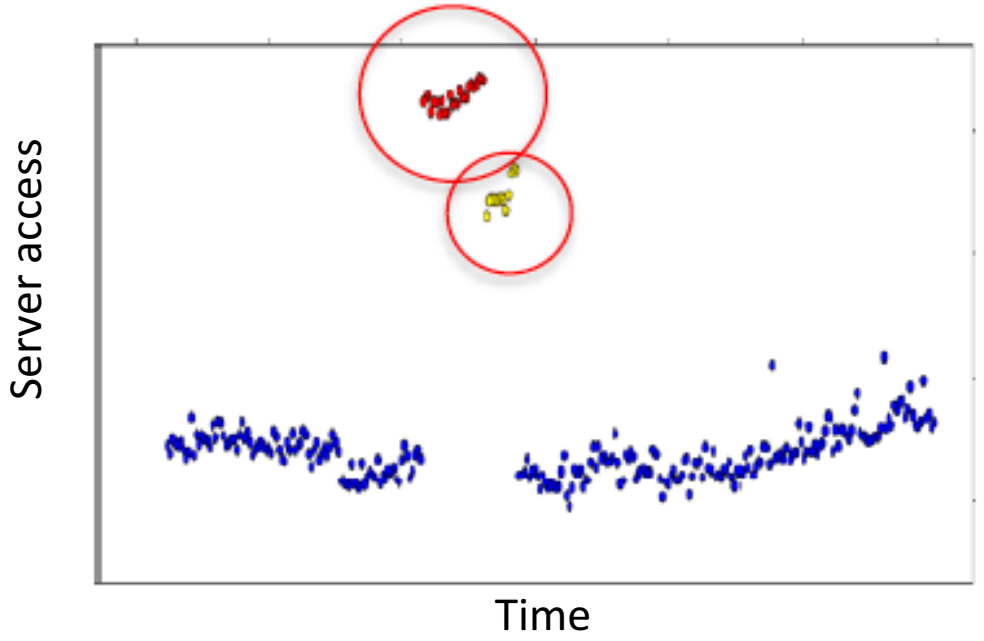
Unsupervised Learning

- Detecting abnormal patterns for security

Credit card fraud detection

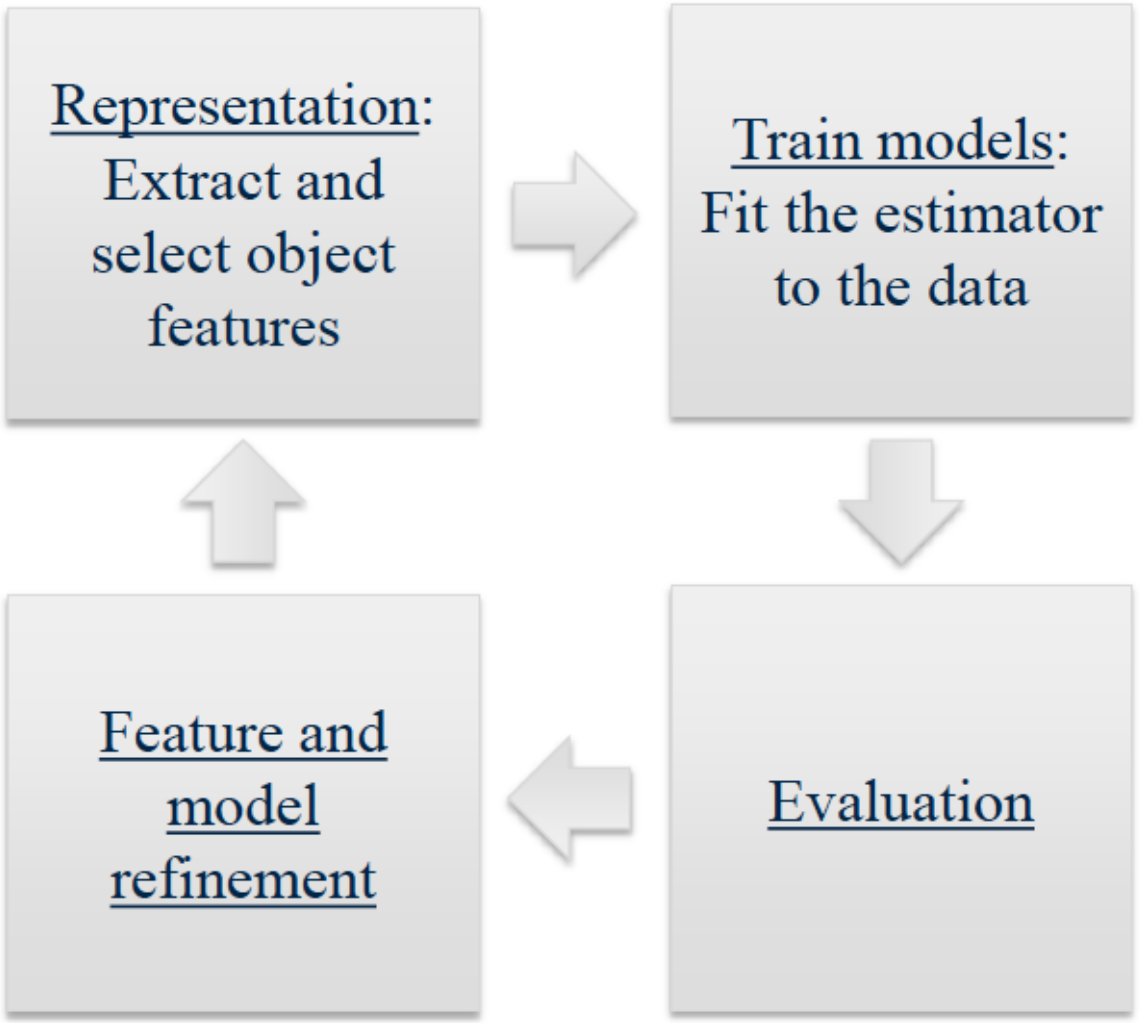


Abnormal server access pattern



Kevyn Collins-Thompson, Applied Machine Learning

Machine Learning Process



Search for model and features that results in high accuracy

Performance is function of:
What data?
What features?
What learning model?
…

Kevyn Collins-Thompson

Data

- Massive amounts of data are available and can be used to train machine learning models
  - online click streams
  - voice and video
  - mobile locations
  - sensors readings
- *Internet of Things* facilitates data collection
- Machine learning performance heavily depends on the data sets used to train the algorithms

Data Sets

- Many public sources
- Public Data Sets from Amazon
  http://aws.amazon.com/datasets?_encoding=UTF8&jiveRedirect=1
- HealthData.gov
  https://www.healthdata.gov/search/type/dataset
- Stanford Large Network Dataset Collection
  http://snap.stanford.edu/data/
- Machine learning competitions: https://www.kaggle.com/competitions
- More (check discussion board/courseweb)

Inspect Data

- Inspect your data
- Missing information
- Wrong readings
  - Correct or discard

	fruit_label	fruit_name	fruit_subtype	mass	width	height	color_score
0	1	apple	granny_smith	192	8.4	7.3	0.55
1	1	apple	granny_smith	180	8.0	6.8	0.59
2	1	apple	granny_smith	176	7.4	7.2	192
3	2	mandarin	mandarin	86	6.2	4.7	0.80
4	2	mandarin	mandarin	84	6.0	4.6	0.79
5	2	mandarin	apple	80	5.8	4.3	0.77
6	2	mandarin	mandarin	80	5.9	4.3	0.81
7	2	mandarin	mandarin	76	5.8	4.0	0.81
8	1	apple	braeburn	178	7.1	7.8	0.92
9	1	apple	braeburn		7.4	7.0	0.89
10	1	apple	braeburn		6.9	7.3	0.93
11	1	apple	braeburn		7.1	7.6	0.92
12	1	apple	braeburn		7.0	7.1	0.88
13	1	apple	golden_delicious		7.3	7.7	0.70
14	1	apple	golden_delicious	152	7.6	7.3	0.69

Course Outline (1) – Subject to Change

- Week 1: Introduction to machine learning, python introduction
- Week 3: Performance tradeoffs, KNN classification,
- Week 4: Linear regression single feature and multiple features,
- Week 5: polynomial regression, regularization
- Week 6-7: Classification
- Week 8: Midterm (26 Feb.)
- Week 9: Spring break

Course Outline (2) – Subject to Change

- Week 10: Cross validation, project proposal due (12 Mar.)
- Week 11-12: Support Vector Machines, Decision trees, Ensembles methods
- Week 13: Neural networks, Dimensionality reduction
- Week 14: Unsupervised learning, ethical considerations
- Week 15: Projects presentations (16 Apr.)
- Week 16: Final Exam (23. Apr.)

Project

- Team: 2-3 members per team
- March 12 - Projects proposal due: In your proposal should include
  - Title and team members
  - Description of the system/problem
  - Explain how machine learning will be used to solve your problem, and your overall approach
  - Mention the type and source of data you will use
  - Include the main responsibilities of each team member in the project
- April 9: Project report due
  - Comprehensive description of the problem, related work, data set, solution, and analysis/evaluation
- April 16: Project presentations

Software

- Python: (https://www.python.org/doc/)
  - **Python basics: A Whirlwind Tour of Python**, by Jake VanderPlas (available online)
- Installation: Anaconda (Recommended)
  https://www.continuum.io/downloads
  Choose **Python 3**
- Scikit-learn (http://scikit-learn.org/)