Applied Analytics and Predictive Modeling

Spring 2020

Lecture-13

Lydia Manikonda

manikl@rpi.edu

Rensselaer

Today’s agenda

- Final exam details
- Class statistics
- Quick review
- Celebrating graduates
- Project submission – deadline tonight

Final Exam

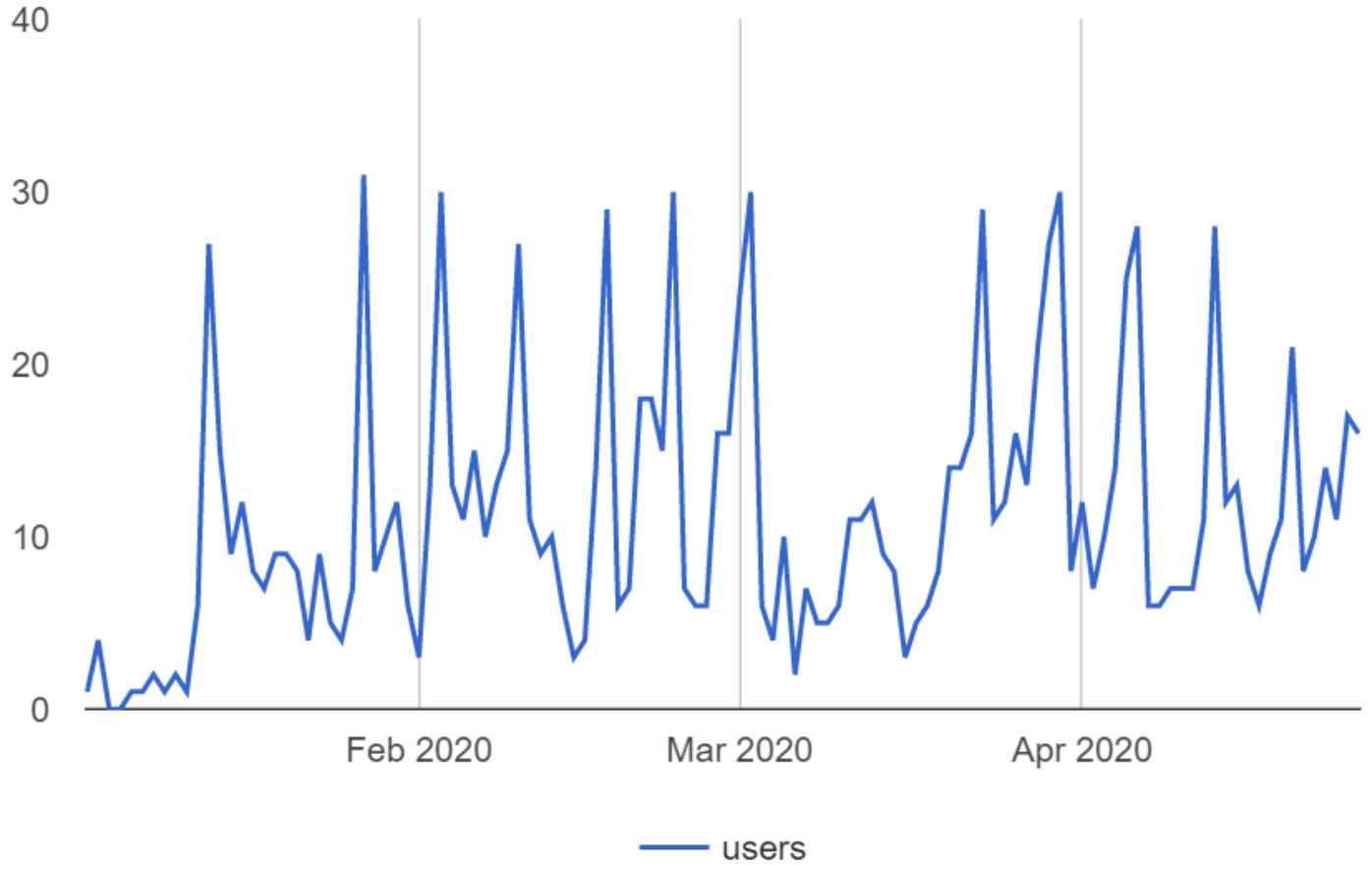
- I will distribute the final exam at 6 am 04/28/2020.
- Due on **05/01/2020 at 11:59 pm**. No late submissions are allowed. Multiple attempts are possible.
- Submission format: pdf ONLY
- Submission link on LMS: https://lms.rpi.edu/webapps/assignment/uploadAssignment?content_id=_132891_1&course_id=_5250_1&group_id=&mode=cpview

Class statistics

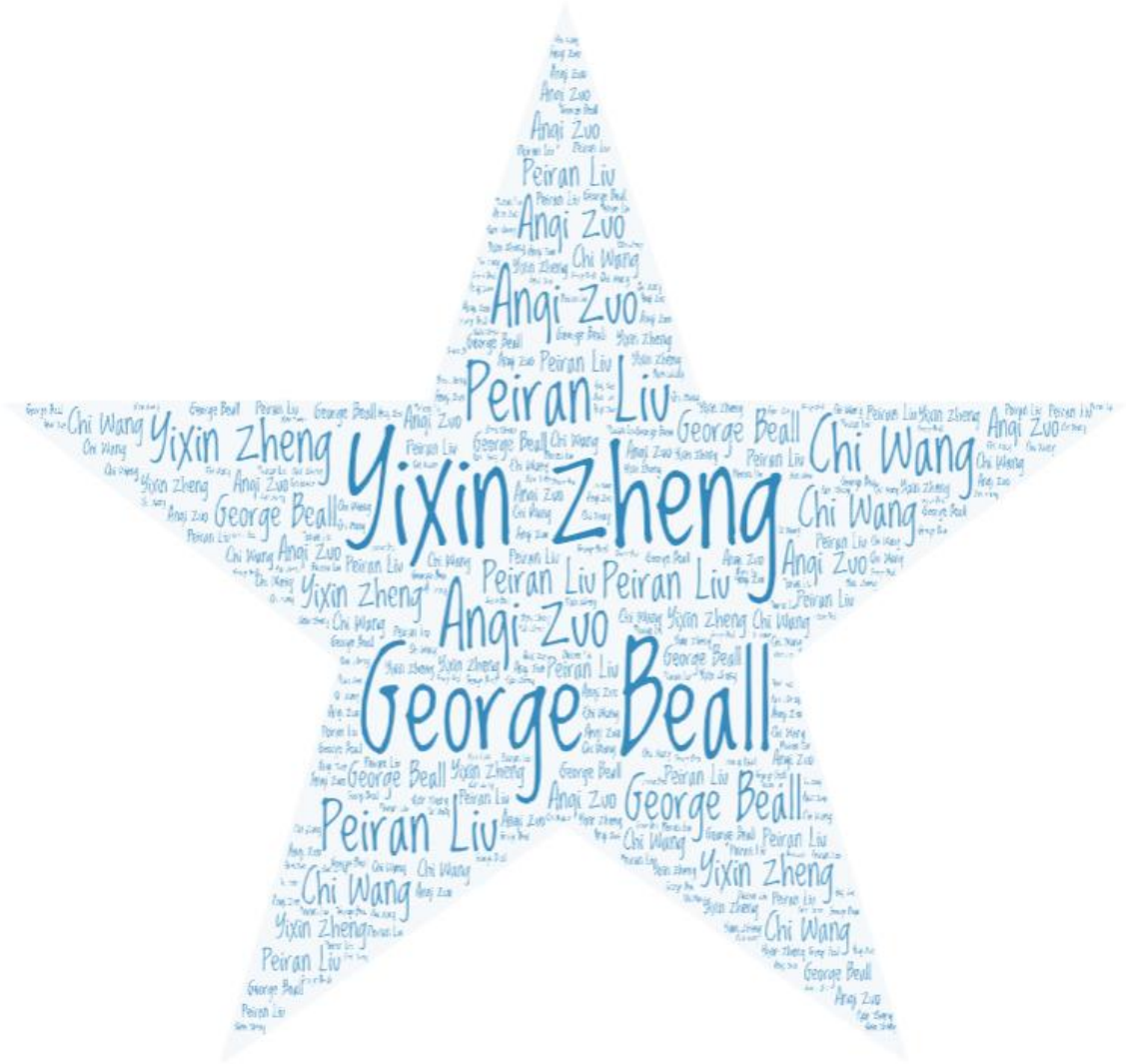
103	total posts*
253	total contributions**
5	un-credited contributions***
35	instructors' responses
6	students' responses
43 min	avg. response time

Most popular days to use Piazza

Mondays are the popular days followed by Fridays..

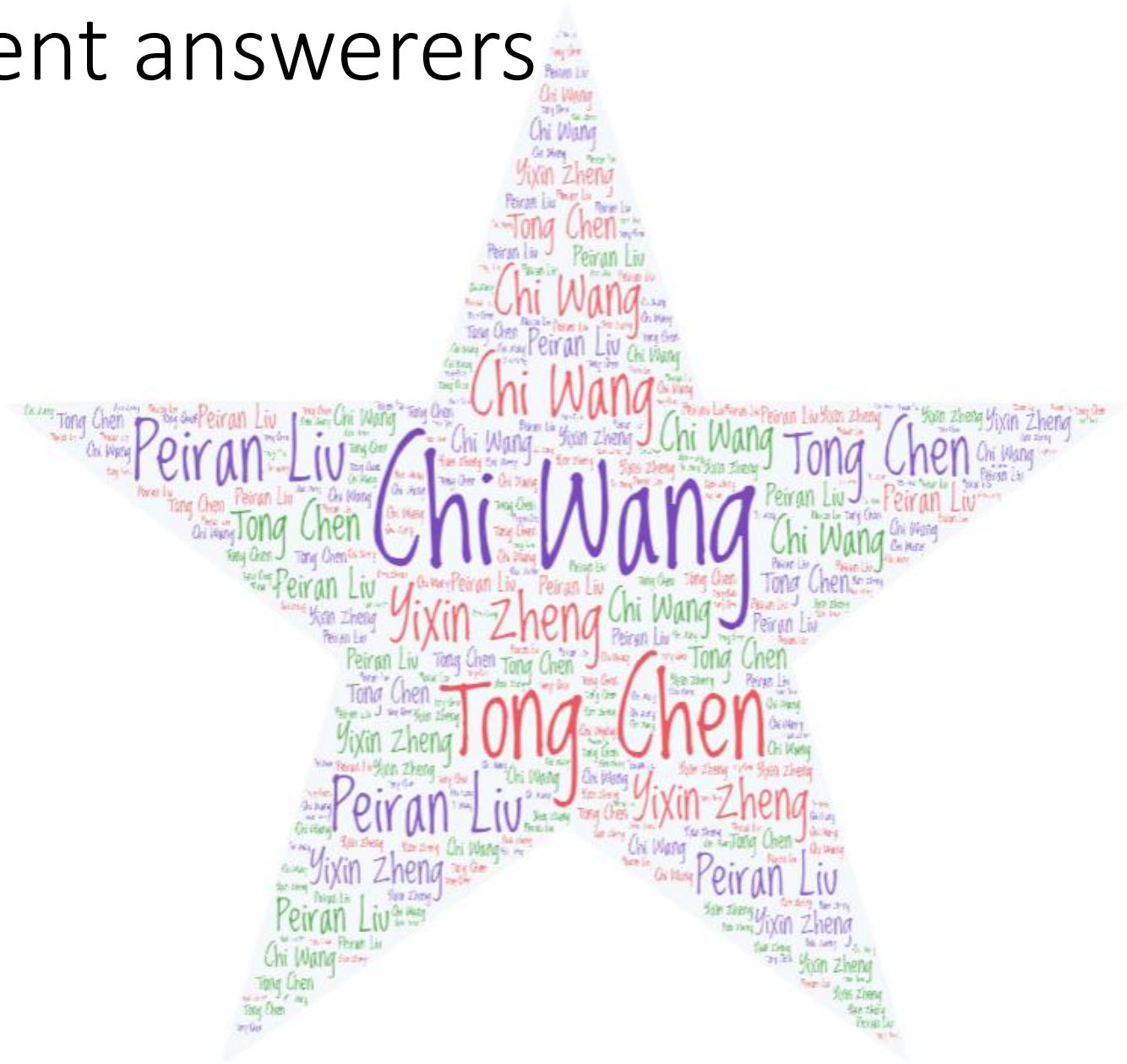


Top students using Piazza



Top student askers

Top student answerers



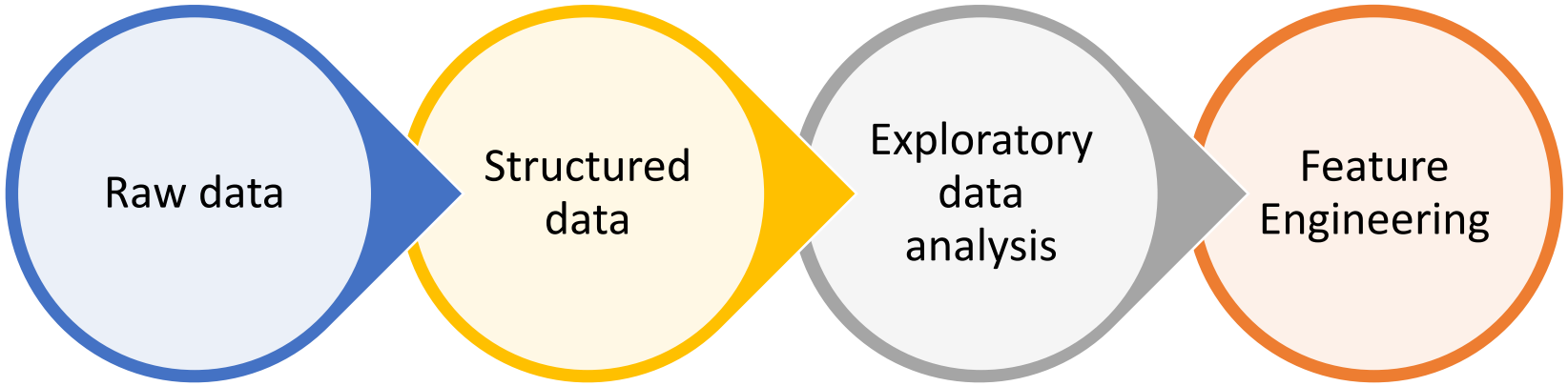
Top student listeners

Topics covered

- Started with Python fundamentals
  - Data structures, loops, conditionals, etc.
- Python packages
  - Numpy, Pandas
  - Seaborn for visualization

Topics covered

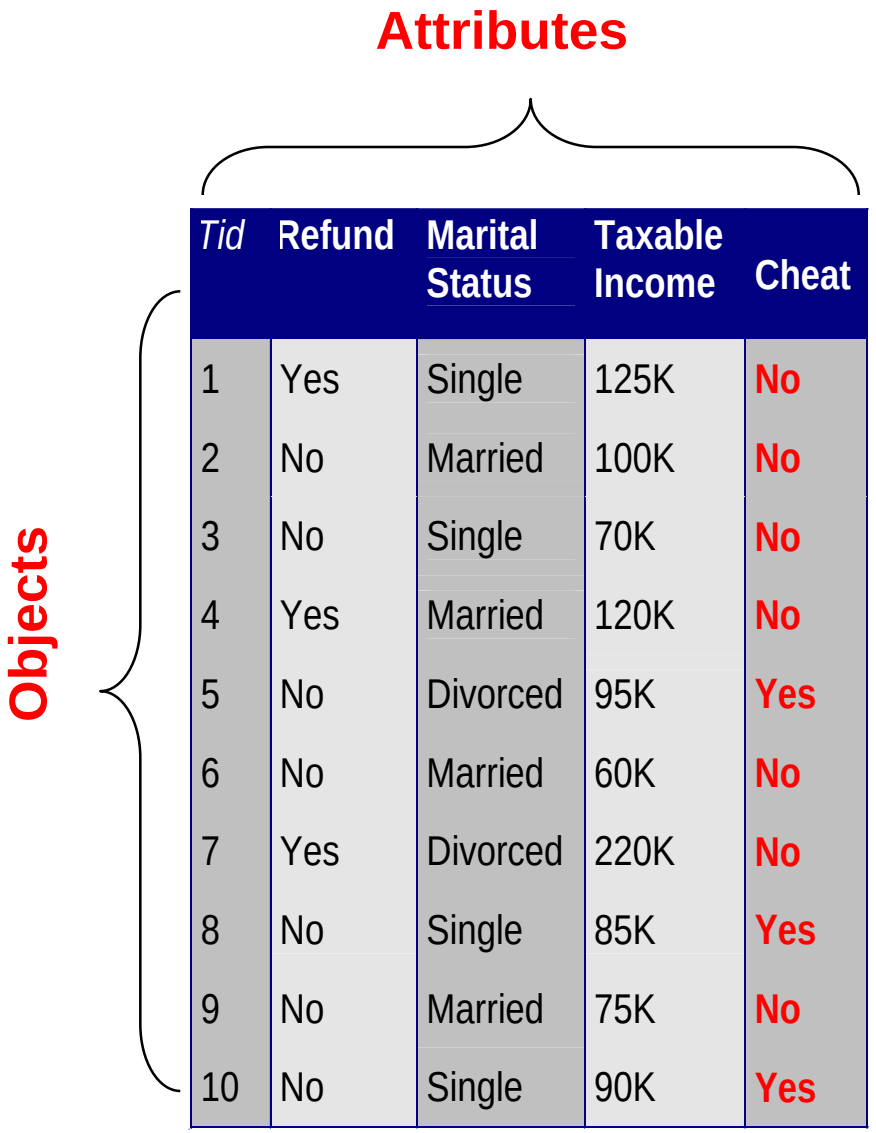
- Data preprocessing



What is data?

- Collection of **data objects** and their **attributes**
- According to Tan et al.,

- An **attribute** is a property or characteristic of an object
  - Also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an **object**
  - Also known as tuple, record, point, case, sample, etc.

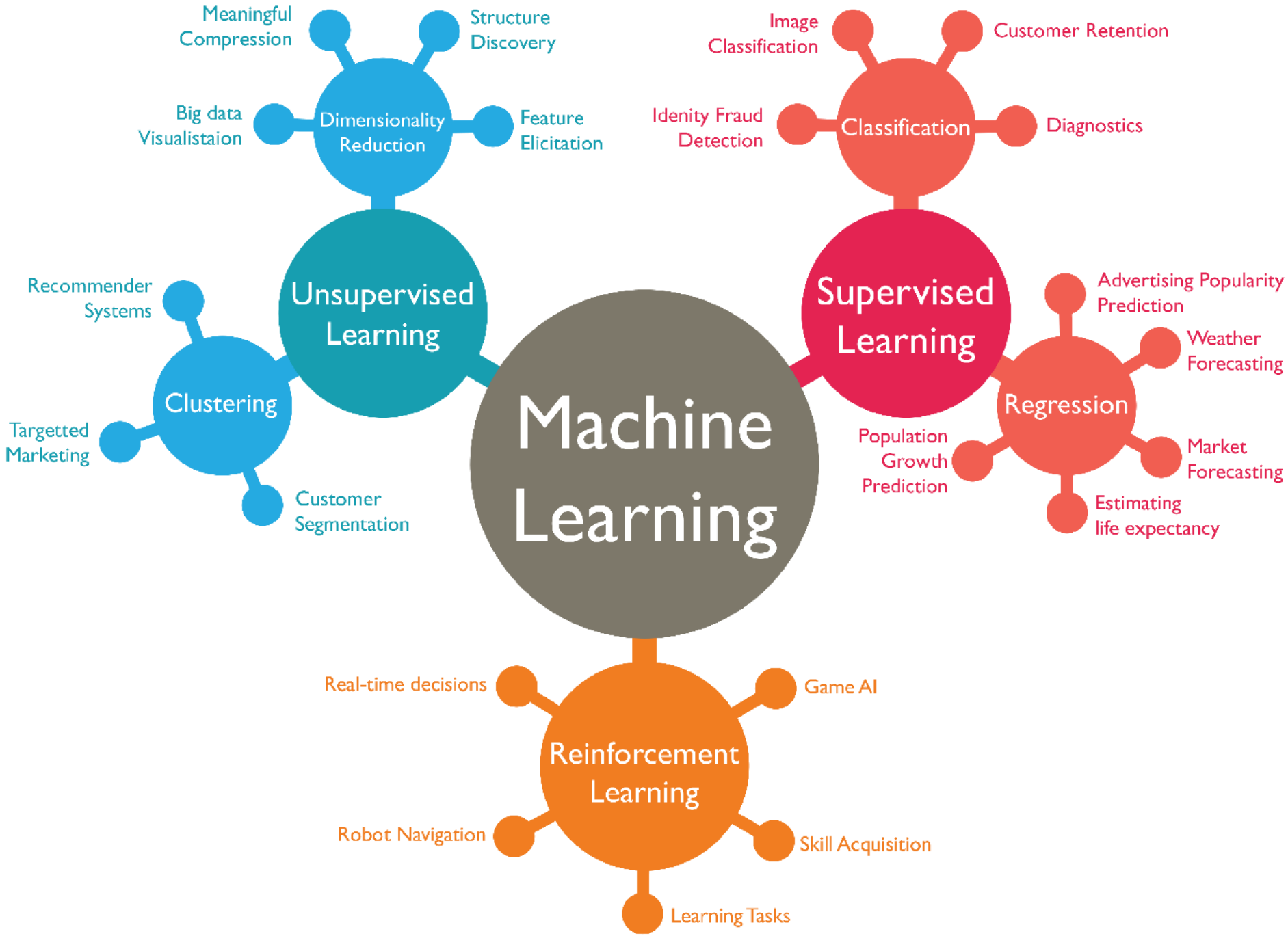


Types of Attributes

- Nominal
  - Examples: ID numbers, zip codes, eye color
- Ordinal
  - Examples: Rankings (expertise level on a scale of 1-10), grades, height {tall, medium, short}
- Interval
  - Examples: Calendar dates, temperature in Celsius or Fahrenheit
- Ratio
  - Examples: Temperature in Kelvin, length, time, counts

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

What is a model?

- Mathematical representation of a real-world process.
- In other words, description of a system using mathematical concepts.

- Three different types of models can be built:
  - Supervised learning
  - Unsupervised learning
  - Semi-supervised learning

Definition of Classification

Given a collection of records (training set)

- Each record is by characterized by a tuple ($\boldsymbol{x}$,y), where $\boldsymbol{x}$ is the attribute set and y is the class label
  - # $\boldsymbol{x}$: attribute, predictor, independent variable, input
  - # y: class, response, dependent variable, output

Task:

- Learn a model that maps each attribute set $\boldsymbol{x}$ into one of the predefined class labels y

Example -- Classification tasks

Task	Attribute set, $\boldsymbol{x}$	Class label, y
Categorizing email messages	Features extracted from email message header and content	spam or non-spam
Identifying tumor cells	Features extracted from MRI scans	malignant or benign cells
Cataloging galaxies	Features extracted from telescope images	Elliptical, spiral, or irregular-shaped galaxies

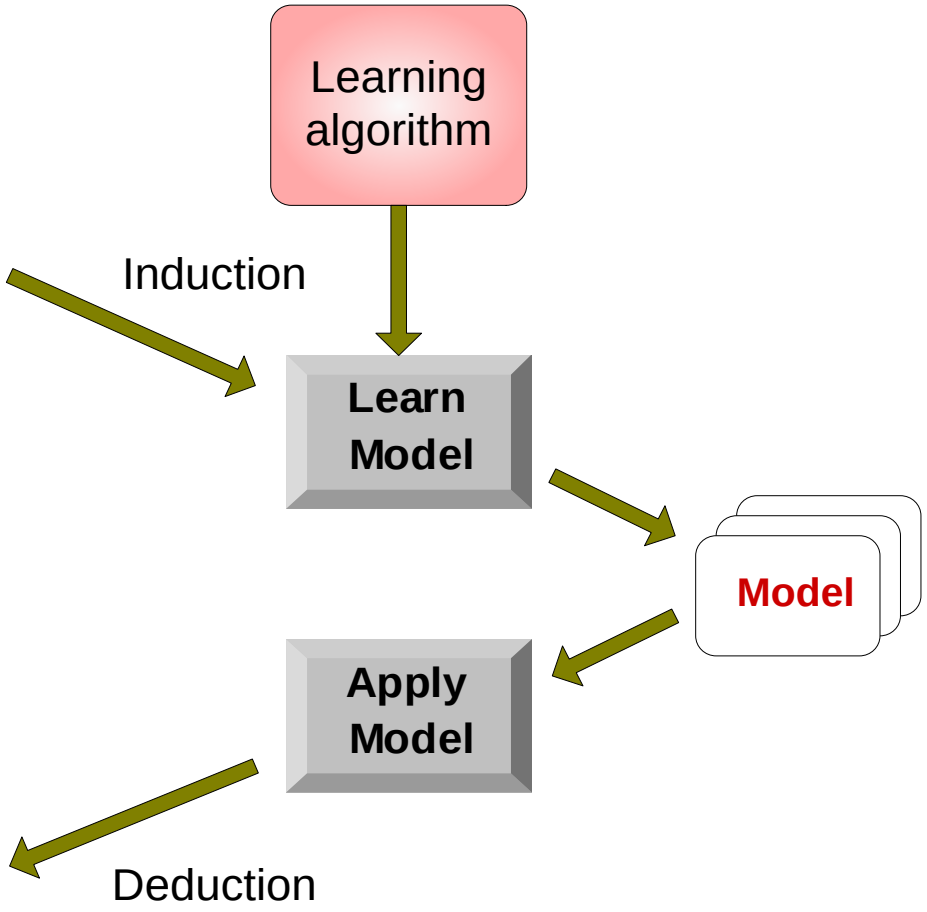
Classification model

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



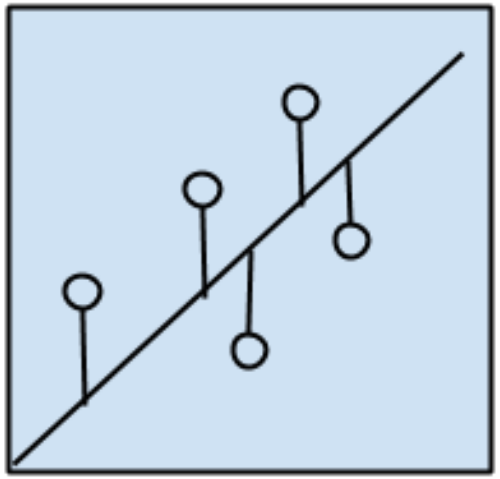
Classification Techniques

- Base Classifiers
  - Decision Tree based Methods
  - Rule-based Methods
  - Nearest-neighbor
  - Neural Networks
  - Deep Learning
  - Naïve Bayes and Bayesian Belief Networks
  - Support Vector Machines

- Ensemble Classifiers
  - Boosting, Bagging, Random Forests

Regression Algorithms

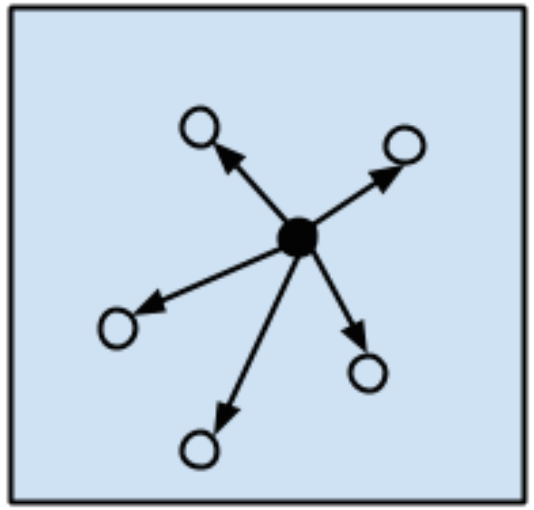
- Modeling the relationship between variables that are iteratively refined using a measure of error.
- Most popular regression algorithms are:
  - Ordinary least squares regression
  - Linear regression
  - Logistic regression
  - Multivariate adaptive regression splines
  - …



Regression Algorithms

Instance-based algorithms

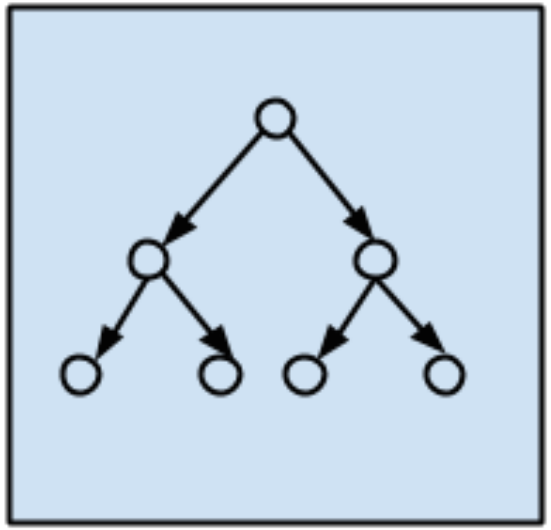
- This model is a decision problem with instances of training data that are deemed important or required to the model.
- Focus is put on the representation of the stored instances and similarity measures used between instances.
- Most popular instance-based algorithms are:
  - K-Nearest Neighbor (KNN)
  - Support Vector Machines (SVM)
  - Learning Vector Quantization
  - Self-Organizing Maps
  - …



Instance-based Algorithms

Decision Tree-based algorithms

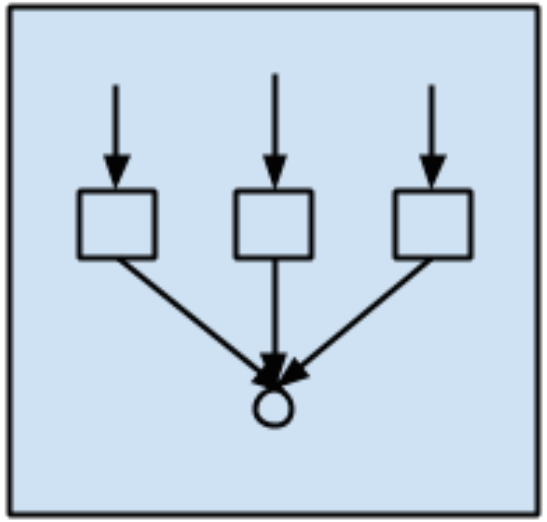
- These methods construct a model of decisions based on the actual values of attributes in the data.
- These decisions built are in the form of a tree.
- Most popular algorithms are:
  - Classification and Regression Tree
  - Conditional Decision Trees
  - ID3
  - C4.5 and C5.0
  - …



Decision Tree Algorithms

Ensemble Algorithms

- These are the models composed of multiple weaker models that are independently trained and the predictions are combined to make the overall prediction.
- Some of the popular algorithms are:
  - Boosting
  - Boostrapped Aggregation
  - AdaBoost
  - Gradient Boosting Machines
  - Random Forest
  - …

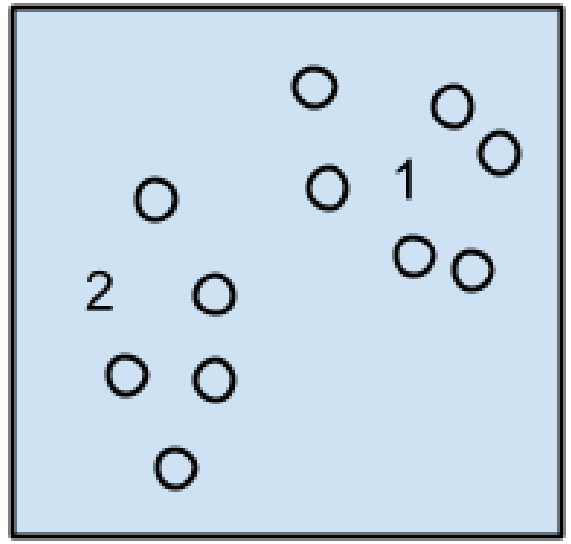


Ensemble Algorithms

Unsupervised Learning

Clustering Algorithms

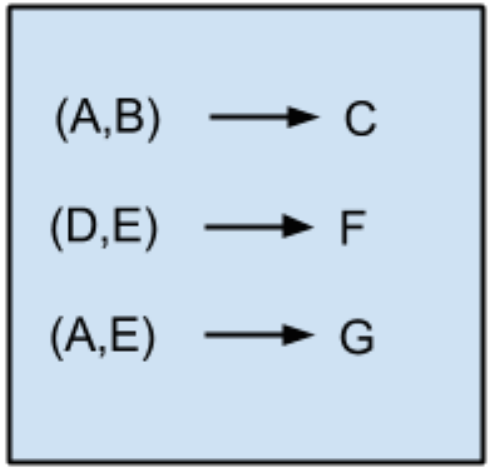
- These algorithms utilize the inherent structures in the data to organize them into various groups.
- Main goal is to find clusters that have high intra similarity and high inter similarity distances.
- Most popular clustering algorithms are:
  - K-Means
  - K-Medoids
  - Expectation Maximization
  - Hierarchical Clustering
  - …



Clustering Algorithms

Association Rule Learning Algorithms

- These methods extract rules that best explain the observed relationships between variables in the data
- Most popular algorithms are:
  - Apriori
  - Eclat
  - FP-growth
  - …



Association Rule Learning Algorithms

Congratulations Graduates!!

- Zhiyi Lin
- Ram Kissondial
- Minerva Opare-Addo
- James Lee
- George Beall
- Shiyun Liu

Vectorstock.com

To all the graduates

- Chi Wang says –

“You'll have a future full of many great achievements and accolades. Keep up the good work. Congrats!”

Zhiyi Lin

- Bingxin Liu says – “Hope you’ll always find yourself as happy and full of big, crazy dreams as you are today!”

- “Congratulations! You are a very smart student and a kind group member! I am very enjoyable to cooperate with you in the group project! Hope you have a promising future!”

- Tong Chen says – “Congratulations on your graduation and best wishes for your next adventure and so happy to share in the excitement of your graduation, and so very proud of you, too. 祝你们毕业快乐， 前程似锦！”

George Beall

- Bingxin Liu says – “So happy to be your partner, congratulations on your graduation and best wishes for your next adventure! ”

- Yue Wang says – “I had a good time working in the same group with you! You are really nice and interesting! Enjoy your life and good luck !”

- Tong Chen says – “Congratulations on your graduation and best wishes for your next adventure and so happy to share in the excitement of your graduation, and so very proud of you, too. 祝你们毕业快乐， 前程似锦！ ”

James Lee

- TJ says – “Congratulations, James! Hope you all the best in the future!”

- “Congratulations on your graduation and thanks for your work during the whole semester! Good luck with your future career at EY and hope the stock market will recover soon so that you can earn more money.”

Shiyun Liu

- Tong Chen says – “Congratulations on your graduation and best wishes for your next adventure and so happy to share in the excitement of your graduation, and so very proud of you, too. 祝你们毕业快乐，前程似锦！”

TJ

- Yueqi Wang says – “Congrats on your second program!!!”

Deadlines at a glance

- **Final project report (1 pdf per team): 04/27/2020 11:59 PM**
- **Final exam (Individual & NO collaboration): 05/01/2020 11:59 PM**

- Exam-1 resubmission: 05/01/2020 11:59 PM (Only if you are resubmitting)

All these above submissions should be made via LMS.