

LECTURE 1

INTRODUCTION TO

WEB MINING

LEK HSIANG HUI

OUTLINE

Introduction to Web Mining

Supervised Learning

Unsupervised Learning

Web Scraping Demo

INTRODUCTION TO WEB MINING



Introduction
to Web
Mining

Supervised
Learning

Unsupervised
Learning

Web
Scraping
Demo

DATA



Companies are generating a LOT of data
e.g. sales, transaction, customer data, etc

MORE DATA



Data is produced
not only by the
companies but
also by others
about the
companies

DATA MINING

Since everything is computerized nowadays, data is now stored in digital form (e.g. databases)

From these databases, the purpose of data mining is to look for patterns so as to discover more insights to the raw data

- Raw Data → Patterns → Knowledge

This knowledge will be helpful for decision making

HOW TO MAKE DECISION MAKING?

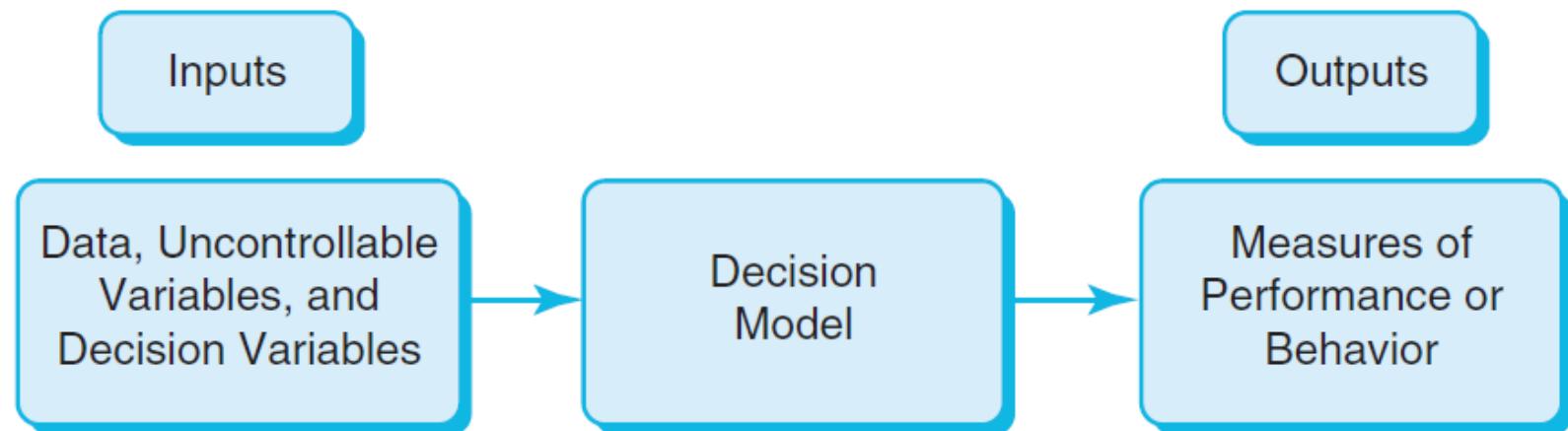
Decision Making

- Make **prediction** based on the data
 - E.g. Based on past experience (i.e. historical data) predict how many patients we are expecting today
- To do this, we build **decision models** using historical data
- The models can then allow us to make prediction of future data instances

WHAT IS A DECISION MODEL?

Decision Model is a model used to understand, analyze, or facilitate decision making

- Can be in the form of a mathematical formula or software



WEB MINING

Web mining is concerned with mining data from the web which can then be transformed into knowledge

2 main aspects we will focus on

- Mining Web Content
- Data mining techniques for handling web content
(regression, classification, clustering, recommender systems)

WHY THIS COURSE?

Most courses only focus on the predictive modeling aspect

- Assumes that all datasets can be downloaded (e.g. from Kaggle)

However:

- The datasets we need is usually not readily available
- Having the ability to mine your own data from the web is useful skill!

CLASSIC DECISION PROBLEMS

Regression (Supervised)

- Stock price prediction

Classification (Supervised)

- Weather forecast (sunny, rainy, cloudy, etc)

Clustering (Unsupervised)

- Group Weibo/Twitter users based on their interest

What's the difference
between regression &
classification?

SUPERVISED LEARNING



Introduction
to Web
Mining

Supervised
Learning

Unsupervised
Learning

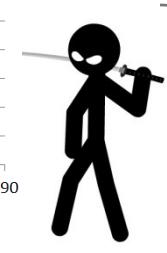
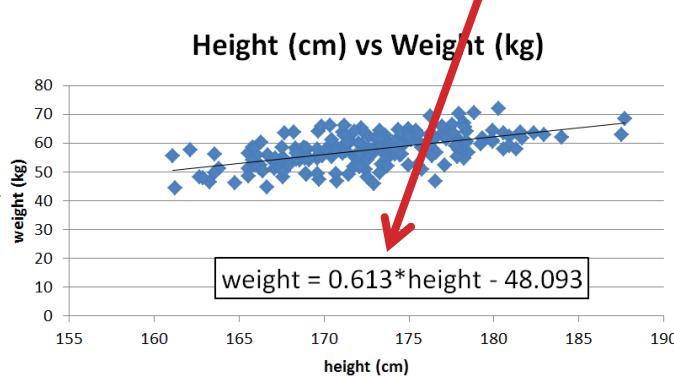
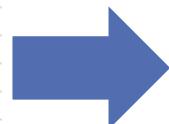
Web
Scraping
Demo

REGRESSION

Regression (Supervised)

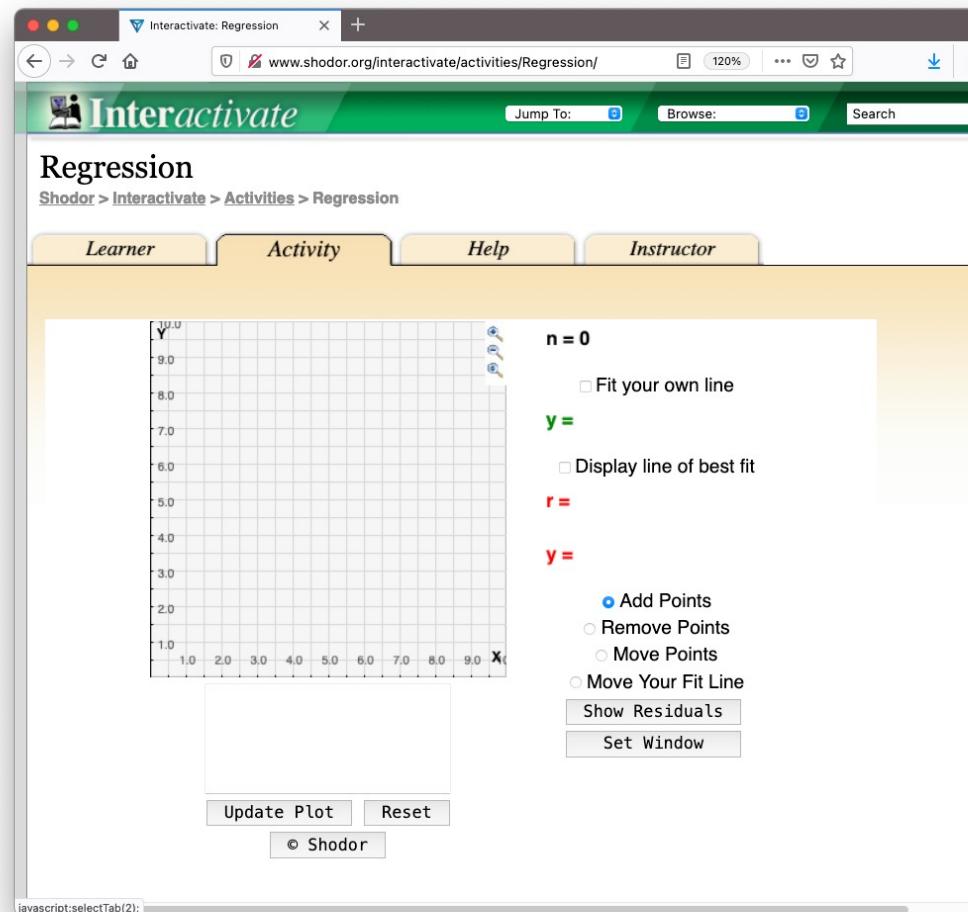
- Using existing data instances to learn a **model** for predicting subsequent instances
- Example:
 - Assume we have a list of human height and weight

Height(cm)	Weight(kg)
167.0812	51.25136008
181.6608	61.91077208
176.276	69.41318376
173.2788	64.56428528
172.1866	65.4533256
174.498	55.9278936
177.292	64.17873208
177.8254	61.89716432
172.466	50.97013304
169.6212	54.73494664
168.8846	57.8103004
171.7548	51.77299088
173.482	56.97569112
170.4848	55.54687632
173.4312	52.85719528



height = 170cm
weight = ?

REGRESSION DEMO



<http://www.shodor.org/interactivate/activities/Regression/>

REGRESSION EXAMPLE

Dataset:

- <https://www.kaggle.com/mohansacharya/graduate-admissions>
- Prediction of Graduate Admissions from an Indian perspective
- 500 instances
- Admission_Predict_Ver1.1.csv

GRADUATE ADMISSION

A	B	C	D	E	F	G	H	I	
1	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
2	1	337	118	4	4.5	4.5	9.65	1	0.92
3	2	324	107	4	4	4.5	8.87	1	0.76
4	3	316	104	3	3	3.5	8	1	0.72
5	4	322	110	3	3.5	2.5	8.67	1	0.8
6	5	314	103	2	2	3	8.21	0	0.65
7	6	330	115	5	4.5	3	9.34	1	0.9
8	7	321	109	3	3	4	8.2	1	0.75
9	8	308	101	2	3	4	7.9	0	0.68
10	9	302	102	1	2	1.5	8	0	0.5
11	10	323	108	3	3.5	3	8.6	0	0.45
12	11	325	106	3	3.5	4	8.4	1	0.52
13	12	327	111	4	4	4.5	9	1	0.84

Identifier (unique for all instance) → Not useful for modeling

Undergraduate GPA
(out of 10)

Research Experience
(0 or 1)

Chance of Admin
(0 to 1)

GRADUATE ADMISSION

A	B	C	D	E	F	G	H	I	
1	1	337	118	4	4.5	4.5	9.65	1	0.92
2	2	324	107	4	4	4.5	8.87	1	0.76
3	3	316	104	3	3	3.5	8	1	0.72
4	4	322	110	3	3.5	2.5	8.67	1	0.8
5	5	314	103	2	2	3	8.21	0	0.65
6	6	330	115	5	4.5	3	9.34	1	0.9
7	7	321	109	3	3	4	8.2	1	0.75
8	8	308	101	2	3	4	7.9	0	0.68
9	9	302	102	1	2	1.5	8	0	0.5
10	10	323	108	3	3.5	3	8.6	0	0.45
11	11	325	106	3	3.5	4	8.4	1	0.52
12	12	327	111	4	4	4.5	9	1	0.84

GRE Scores
(out of 340)

TOEFL Score
(out of 120)

Uni. Rating
(out of 5)

Statement of Purpose
(SOP) & Letter of
Recommendation (LOR)
Recommendation Strength
(out of 5)

TYPES OF DATA

Target/Response
Dependent Variable

	A	B	C	D	E	F	G	H	I
1	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
2	1	337	118	4	4.5	4.5	9.65	1	0.92
3	2	324	107	4	4	4.5	8.87	1	0.76
4	3	316	104	3	3	3.5	8	1	0.72
5	4	322	110	3	3.5	2.5	8.67	1	0.8
6	5	314	103	2	2	3	8.21	0	0.65
7	6	330	115	5	4.5	3	9.34	1	0.9
8	7	321	109	3	3	4	8.2	1	0.75
9	8	308	101	2	3	4	7.9	0	0.68
10	9	302	102	1	2	1.5	8	0	0.5
11	10	323	108	3	3.5	3	8.6	0	0.45
12	11	325	106	3	3.5	4	8.4	1	0.52
13	12	327	111	4	4	4.5	9	1	0.84

Predictors
Independent Variables

TYPES OF DATA

Datasets with values for both predictors & response are also known as labeled data
(also known as training data)

A	B	C	D	E	F	G	H	I	
1	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
2	1	337	118	4	4.5	4.5	9.65	1	0.92
3	2	324	107	4	4	4.5	8.87	1	0.76
4	3	316	104	3	3	3.5	8	1	0.72
5	4	322	110	3	3.5	2.5	8.67	1	0.8
6	5	314	103	2	2	3	8.21	0	0.65
7	6	330	115	5	4.5	3	9.34	1	0.9
8	7	321	109	3	3	4	8.2	1	0.75
9	8	308	101	2	3	4	7.9	0	0.68
10	9	302	102	1	2	1.5	8	0	0.5
11	10	323	108	3	3.5	3	8.6	0	0.45
12	11	325	106	3	3.5	4	8.4	1	0.52
13	12	327	111	4	4	4.5	9	1	0.84

TYPES OF DATA

Datasets with only values for predictors are also known as unlabeled data (also known as testing data)

A	B	C	D	E	F	G	H	I	
1	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
2	401	304	100	2	3.5	3	8.22	0	?
3	402	315	105	2	3	3	8.34	0	?
4	403	324	109	3	3.5	3	8.94	1	?
5	404	330	116	4	4	3.5	9.23	1	?
6	405	311	101	3	2	2.5	7.64	1	?
7	406	302	99	3	2.5	3	7.45	0	?
8	407	322	103	4	3	2.5	8.02	1	?
9	408	298	100	3	2.5	4	7.95	1	?
10	409	297	101	3	2	4	7.67	1	?
11	410	300	98	1	2	2.5	8.02	0	?

We want the generated model to predict the response values

SUPERVISED LEARNING

**This also illustrates the idea of
Supervised Learning**

- Teach the machine to do prediction with examples and the corresponding expected prediction

HANDS-ON: REGRESSION

HANDS-ON: REGRESSION

Dataset:

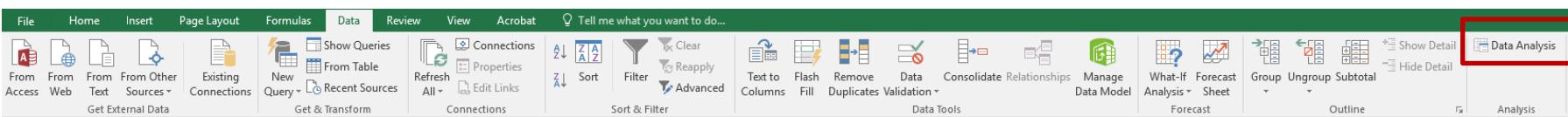
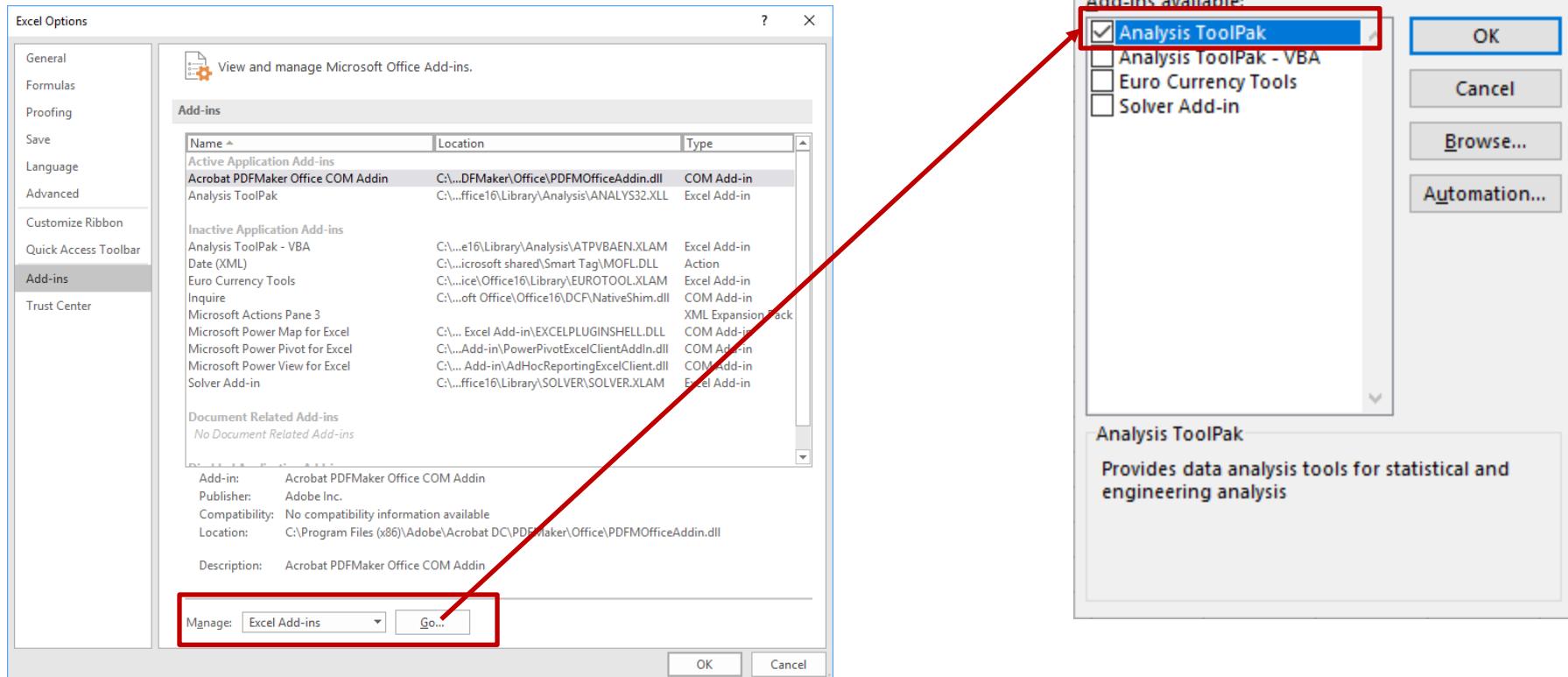
- <https://www.kaggle.com/mohansacharya/graduate-admissions>
- Prediction of Graduate Admissions from an Indian perspective
- 500 instances
- Manually divided into:
 - 400 instances training data (**Admission_Predict_Ver1.1.train.csv**)
 - 100 instances testing data (**Admission_Predict_Ver1.1.test.csv**)

Software:

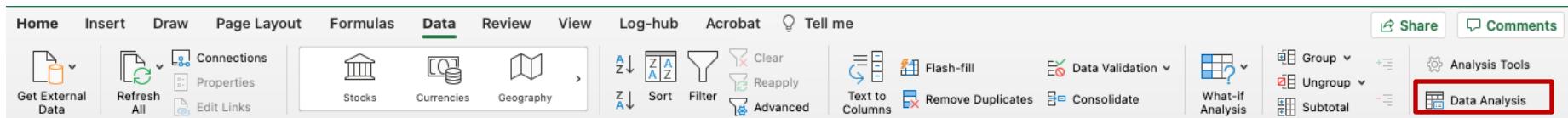
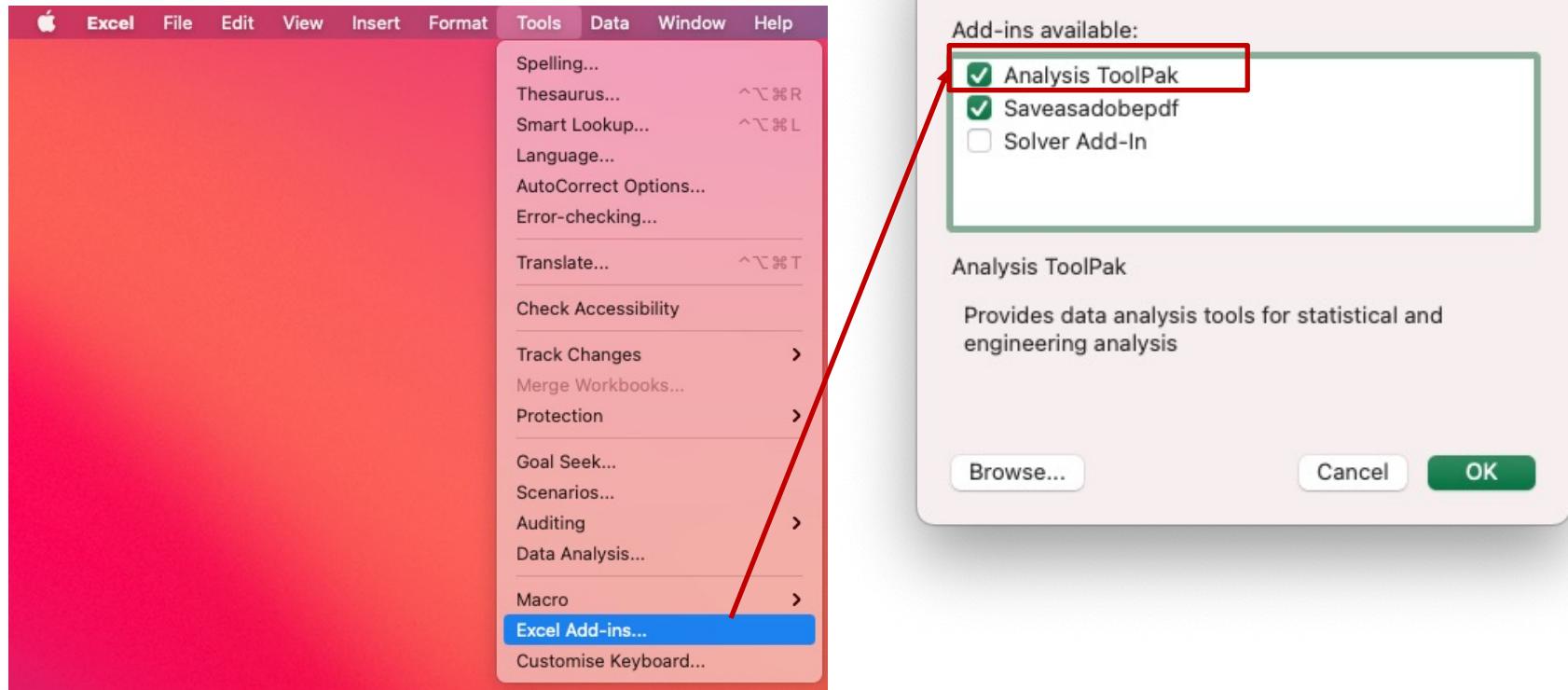
- Excel (with the **Analysis ToolPak**)



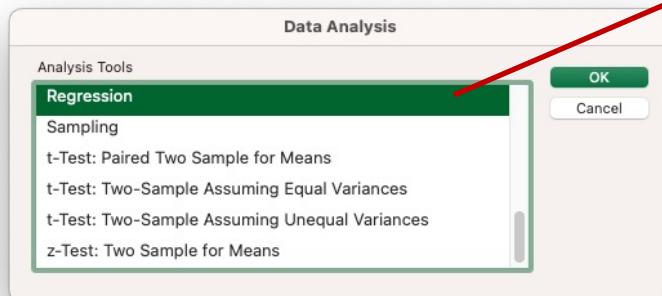
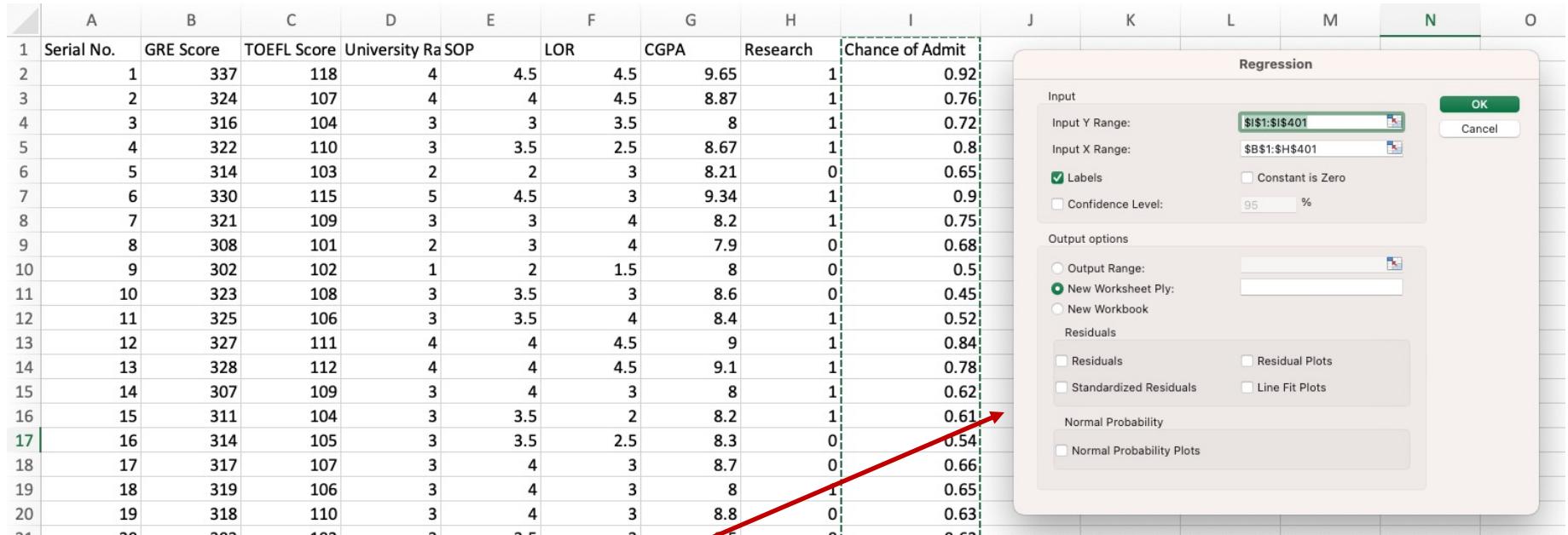
INSTALLING EXCEL ANALYSIS TOOLPAK (WIN)



INSTALLING EXCEL ANALYSIS TOOLPAK (MAC)



GENERATING REGRESSION MODEL (EXCEL)



Admission_Predict_Ver1.1.train.csv

GENERATING REGRESSION MODEL (EXCEL)

15									
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
17	Intercept	-1.259432478	0.124730747	-10.097209	1.841E-21	-1.5046574	-1.0142076	-1.5046574	-1.0142076
18	GRE Score	0.001737412	0.000597897	2.9058702	0.0038701	0.0005619	0.0029129	0.0005619	0.0029129
19	TOEFL Score	0.002919577	0.001089532	2.6796622	0.0076802	0.0007775	0.0050616	0.0007775	0.0050616
20	University Rating	0.005716658	0.004770425	1.1983539	0.2315032	-0.0036622	0.0150955	-0.0036622	0.0150955
21	SOP	-0.003305169	0.005561643	-0.5942792	0.5526682	-0.0142395	0.0076292	-0.0142395	0.0076292
22	LOR	0.022353127	0.005541485	4.0337793	6.599E-05	0.0114584	0.0332479	0.0114584	0.0332479
23	CGPA	0.118939454	0.012219435	9.7336294	3.382E-20	0.0949156	0.1429633	0.0949156	0.1429633
24	Research	0.024525106	0.007959756	3.0811379	0.0022076	0.008876	0.0401743	0.008876	0.0401743
25									

```
-1.259432478 + 0.001737412 * GRE + 0.002919577 * TOEFL + 0.005716658 * Uni_Rating - 0.003305169 * SOP + 0.022353127 * LOR + 0.118939454 * CGPA + 0.024525106 * Research
```

PREDICTING TESTING DATA (EXCEL)

Admission_Predict_Ver1.1.test.csv

$-1.259432478 + 0.001737412 * \text{GRE} + 0.002919577 * \text{TOEFL} + 0.005716658 * \text{Uni_Rating} - 0.003305169 * \text{SOP} + 0.022353127 * \text{LOR} + 0.118939454 * \text{CGPA} + 0.024525106 * \text{Research}$

	A	B	C	D	E	F	G	H	I	J
1	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit (Actual)	Chance of Admit (Predicted)
2	401	304	100		2	3.5	3	8.22	0	0.63
3	402	315	105		2	3	3	8.34	0	0.66
4	403	324	109		3	3.5	3	8.94	1	0.78
5	404	330	116		4	4	3.5	9.23	1	0.91
6	405	311	101		3	2	2.5	7.64	1	0.62
7	406	302	99		3	2.5	3	7.45	0	0.52

Predicted values

RECALL: TYPES OF DECISION PROBLEM (LEARNING PROBLEMS)

Regression (Supervised)

- Stock price prediction

Classification (Supervised)

- Weather forecast (sunny, rainy, cloudy, etc)

Clustering (Unsupervised)

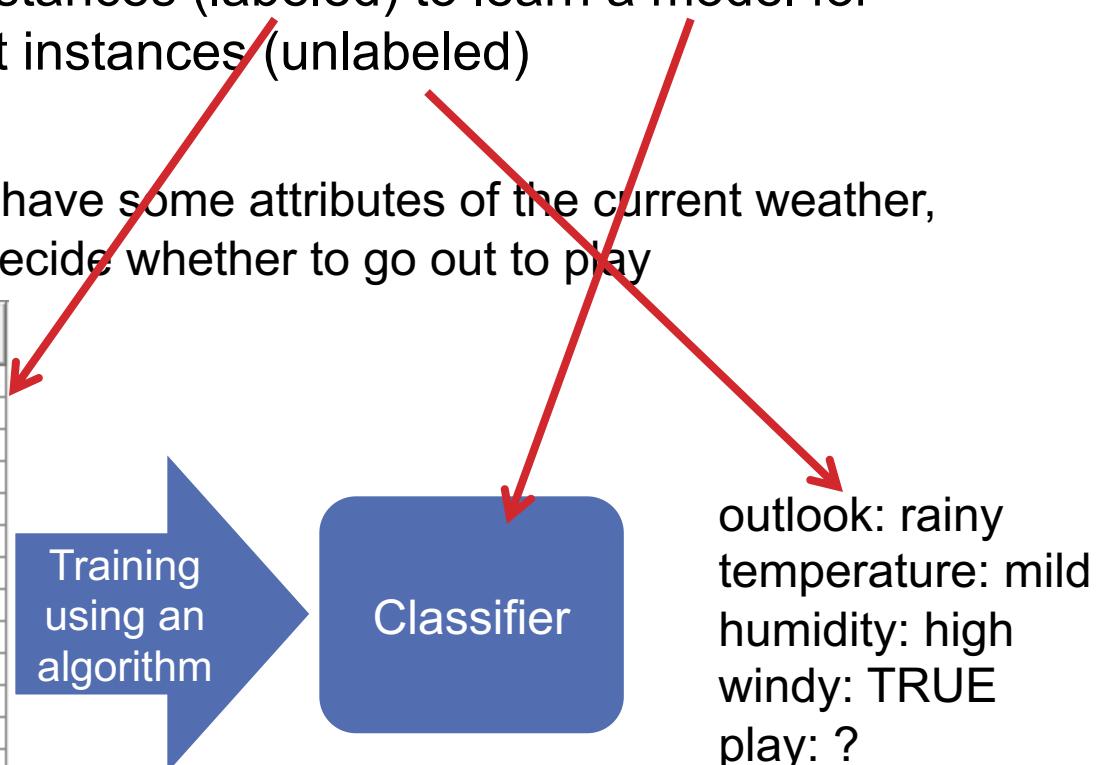
- Group Facebook users based on their interest

CLASSIFICATION

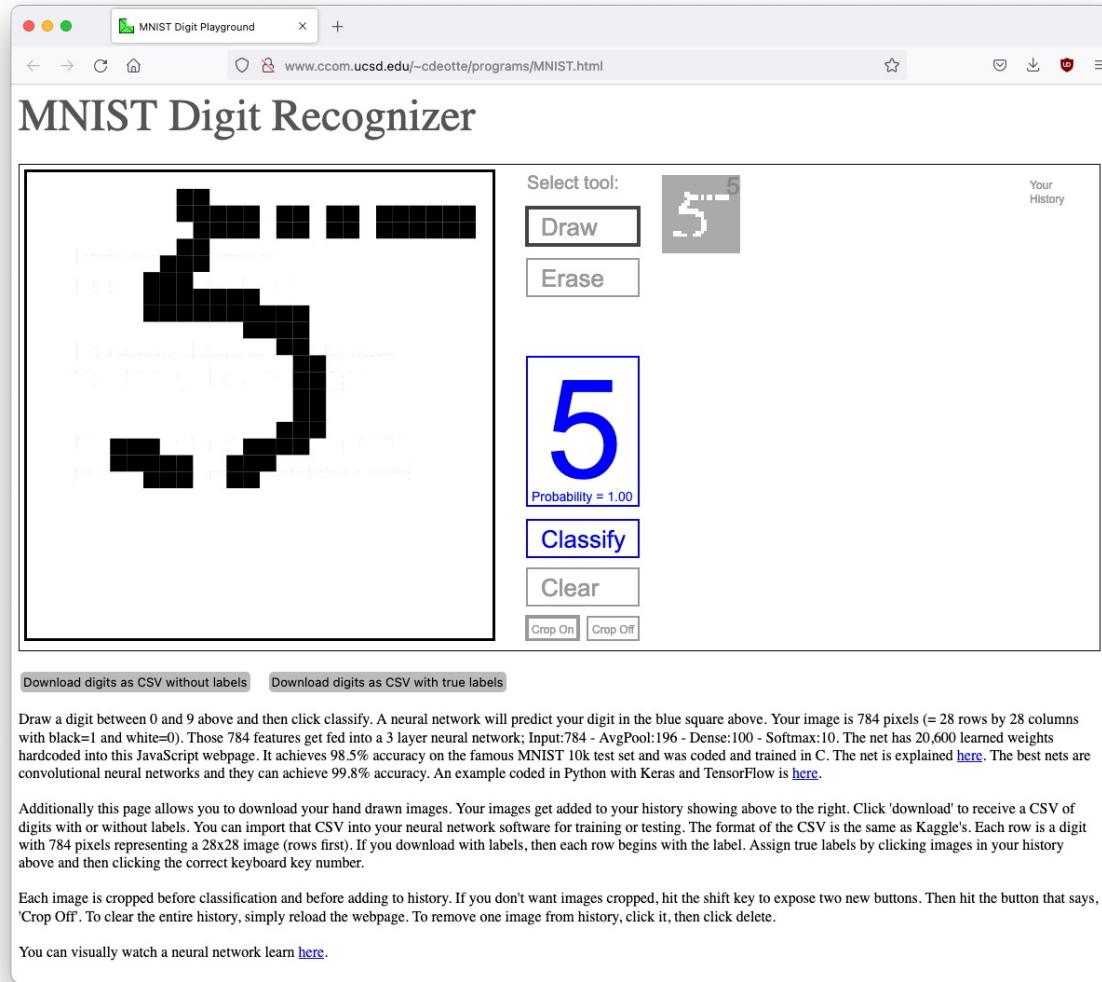
Classification (Supervised)

- Using existing data instances (labeled) to learn a model for predicting subsequent instances (unlabeled)
- Example:
 - Assume that you have some attributes of the current weather, and we need to decide whether to go out to play

No.	1: outlook Nominal	2: temperature Nominal	3: humidity Nominal	4: windy Nominal	5: play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

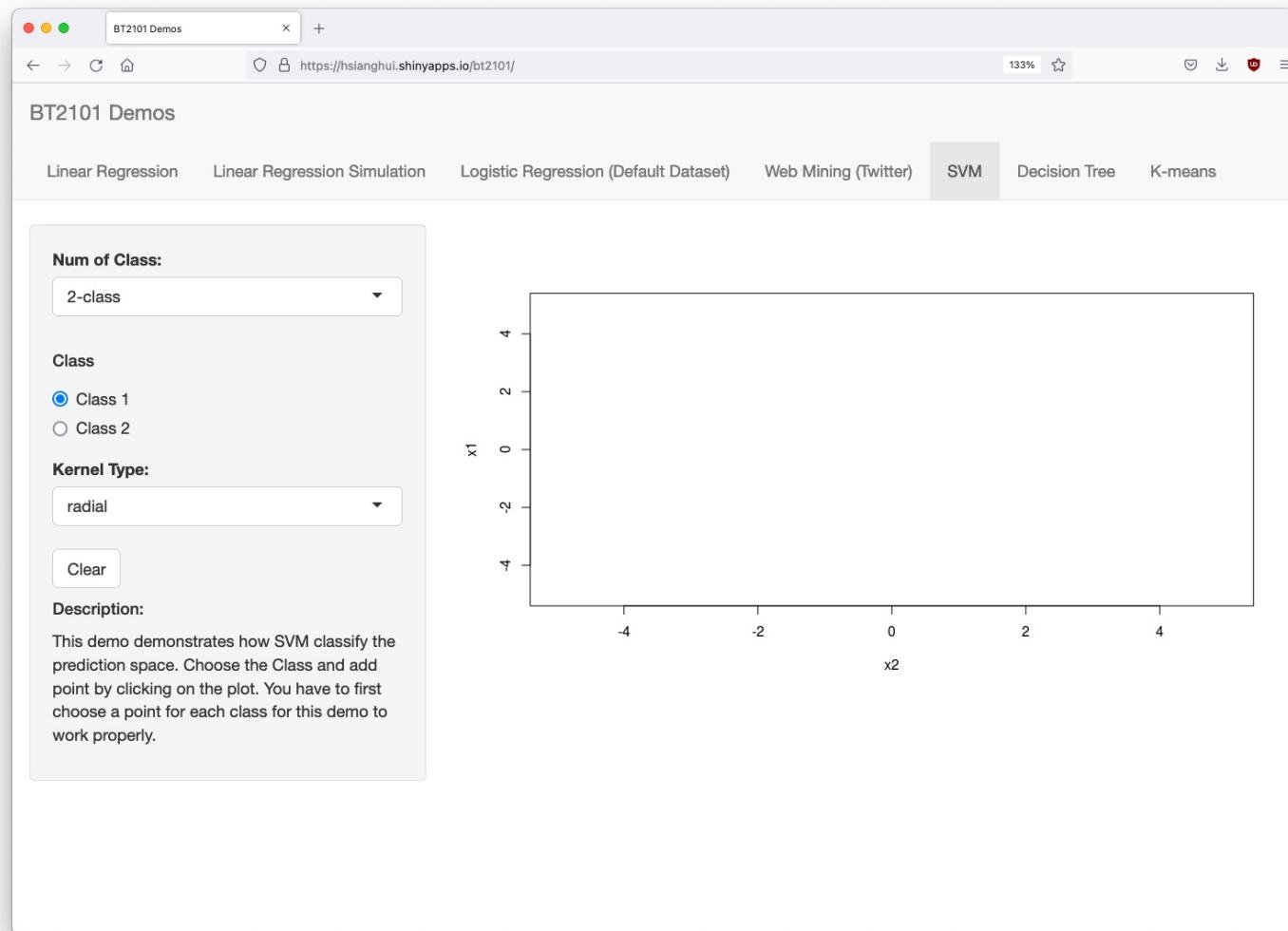


CLASSIFICATION DEMO



<http://www.ccom.ucsd.edu/~cdeotte/programs/MNIST.html>

CLASSIFICATION DEMO



<https://hsianghui.shinyapps.io/bt2101/>

UNSUPERVISED LEARNING



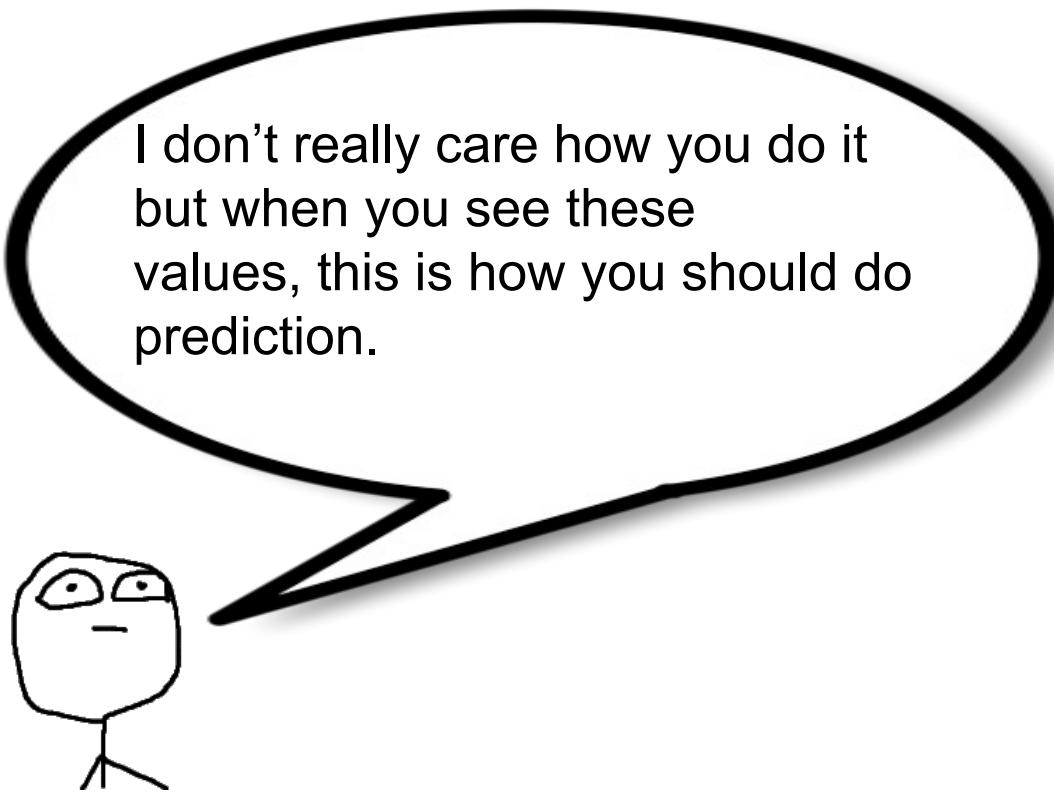
Introduction
to Web
Mining

Supervised
Learning

Unsupervised
Learning

Web
Scraping
Demo

MACHINE LEARNING



MACHINE LEARNING

Ok I've learnt something
about the data but I get
better at this if you give
me more examples!



CHALLENGES OF SUPERVISED LEARNING



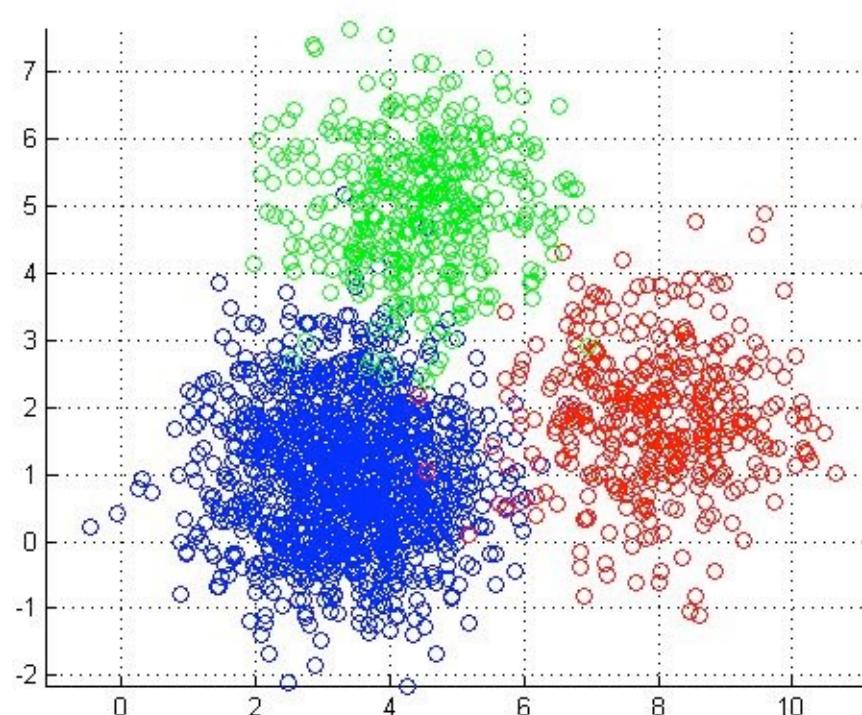
UNSUPERVISED LEARNING

Unsupervised

- Assume we only have unlabeled data and we want to either:
 - label the data instances
 - or group up data instances into subsets sharing common characteristics
- One commonly application of unsupervised approach is clustering
- What do you think is the accuracy of unsupervised approaches compared to supervised?
- Why are unsupervised approaches important?

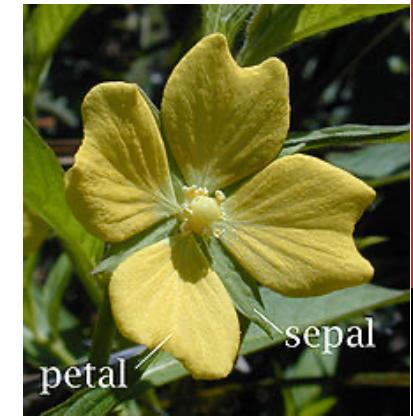
CLUSTERING EXAMPLE

Points are clustered together because they are closer to each other



UNSUPERVISED

Clustering Example (Iris flower)



Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
6.2	2.2	4.5	1.5	versicolor
5.7	2.6	3.5	1	versicolor
5.5	2.4	3.8	1.1	versicolor
5.5	2.4	3.7	1	versicolor
7.2	3.6	6.1	2.5	virginica
6.5	3.2	5.1	2	virginica
6.4	2.7	5.3	1.9	virginica
6.8	3	5.5	2.1	virginica



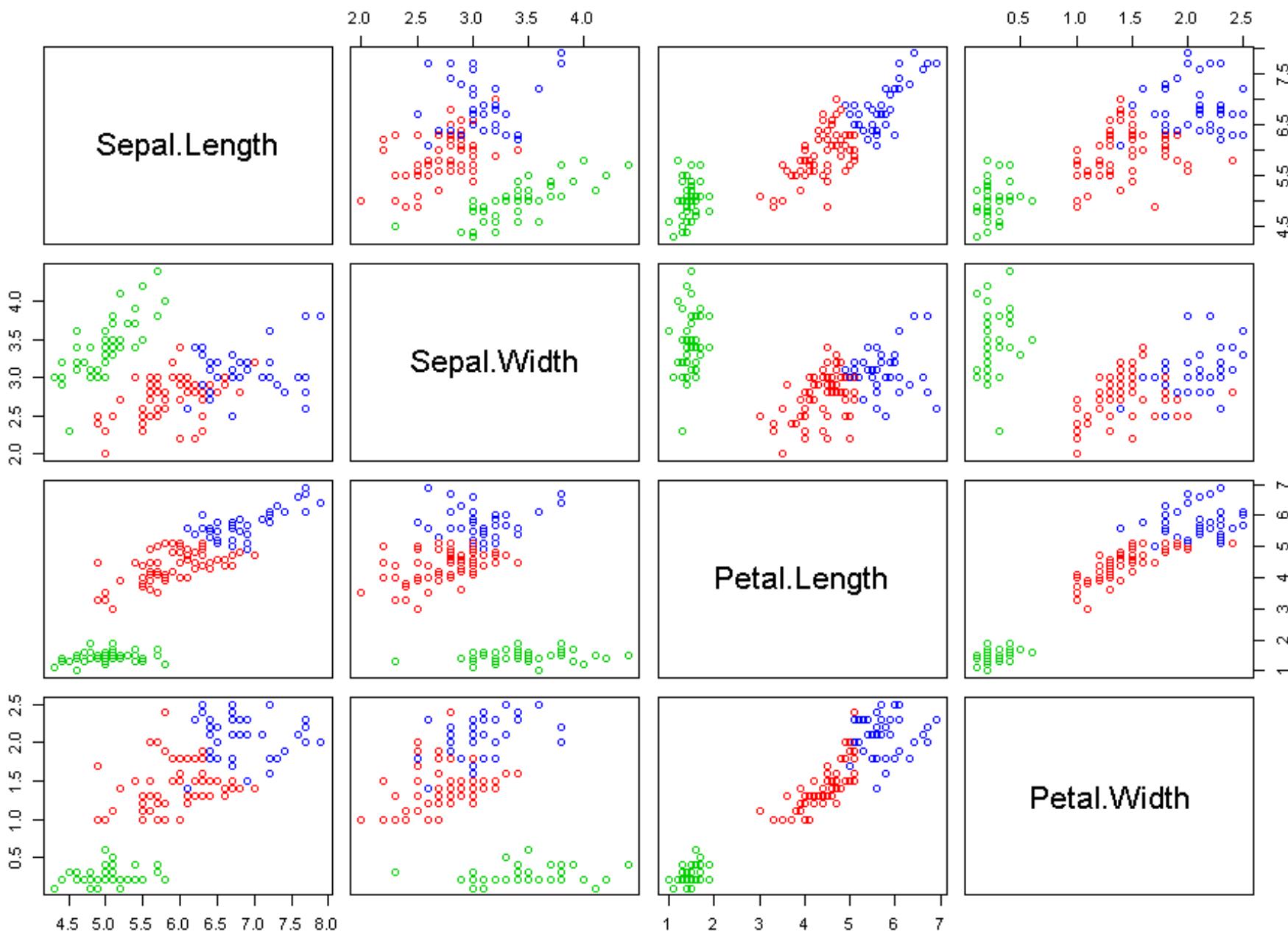
Iris Setosa



Iris Versicolor



Iris Virginica



OTHER TYPES OF DECISION PROBLEMS

Regression (Supervised)

- Stock price prediction

Classification (Supervised)

- Weather forecast (sunny, rainy, cloudy, etc)

Clustering (Unsupervised)

- Group Facebook users based on their interest

OTHER TYPES OF DECISION PROBLEMS

Data mining techniques:

- Regression
- Classification
- Clustering
- Association Rules
- Neural Networks
- Deep Learning
- etc



Already discussed briefly earlier

ASSOCIATION RULES

Association Rules

- Quite different from Regression and Classification
- Still using data but there is no clear-cut prediction that we can derive
- Also different from Clustering in that we are clear in our objectives
- We are interested in mining **Association Rules** which look like:
 - $X \rightarrow Y$ (X implies Y)
 - E.g. $\{\text{diaper, milk}\} \rightarrow \{\text{beer}\}$

SUPERMARKET PROBLEM

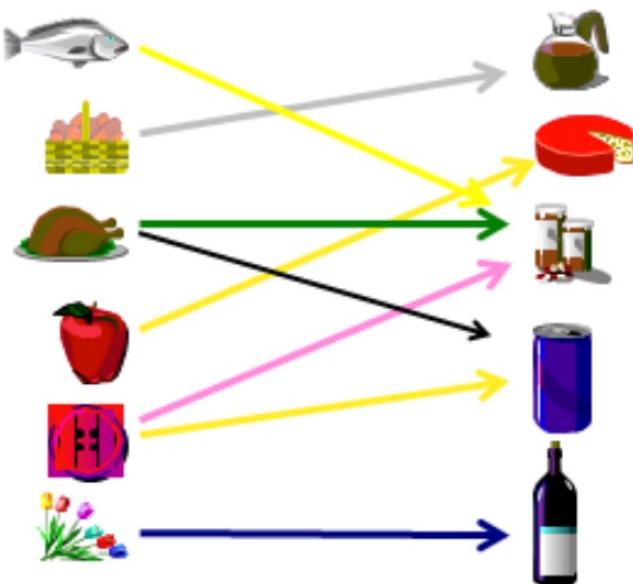
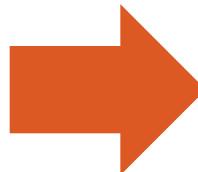


One of the key problems that a supermarket wants to tackle is how to place the products

Place similar category products together for better searching?

MARKET BASKET ANALYSIS

Market Basket Analysis (MBA) aims to find **association** between groups of items based on the **transactions**



98% of people who purchased items A and B also purchased C

SUPERMARKET PROBLEM



Probably better to place products that users are likely to buy together in the same location

- Likely to generate more sales that way

WEB SCRAPING DEMO



Introduction
to Web
Mining

Supervised
Learning

Unsupervised
Learning

Web
Scraping
Demo

WEB SCRAPING DEMO: AMAZON.COM

The screenshot shows the Amazon.com Today's Deals page. At the top, there is a navigation bar with the Amazon logo, delivery options (Deliver to Singapore), search bar, account information (Hello, Sign In / Account & Lists), and a shopping cart icon (0 items). Below the navigation bar, there is a secondary menu with links like All, Today's Deals, Customer Service, Registry, Gift Cards, and Sell.

The main content area is titled "Today's Deals" and features a grid of product thumbnails. Each thumbnail includes a deal summary (e.g., "Up to 36% off Top deal") and a "Sort by: Featured" dropdown. The products shown include:

- Magical Flames Cosmic Fire Color Packets (Up to 36% off, Top deal)
- Amazfit Smartwatches and Bands (Up to 30% off, Top deal)
- Thermacell E-Series Mosquito Repellers (26% off, Top deal)
- Outdoor Power & Lawn Equipment (Up to 30% off, On deal)
- Wag JERKY (Jerky product)
- A small image of a colorful toy or container.

On the left side of the main content area, there is a sidebar with filters and categories:

- All deals
- Available
- Upcoming
- Watchlist
- Delivery
- prime
- Departments
- Select All
 - Amazon Devices
 - Arts, Crafts & Sewing
 - Automotive & Motorcycle
 - Baby
 - Baby Clothing & Accessories
 - Beauty
 - Books
 - Boys' Fashion
 - Camera & Photo
 - Cell Phones & Accessories
 - Computers & Accessories

CHALLENGE

Are you able to get the review text of a product?

Amazon.com: Customer reviews X +

https://www.amazon.com/Apple-iPhone-XR-64GB-White/product-reviews/B08BGD4G36

Top reviews All reviewers All stars All formats Text, image, video

45,886 total ratings, 9,390 with reviews

From the United States

Alan marnik

★★★★★ As advertised

Reviewed in the United States on April 27, 2019

Size: 128GB | Color: Blue | Service Provider: Unlocked | Product grade: Renewed | **Verified Purchase**

I was a little sceptical buying this phone because some of the reviews said that the phone they received was locked or had slight damage. However, the one I received was perfect. It came fully unlocked and undamaged. It looks as if I had bought it straight from apple. So far an overall 5 stars and no complaints.

1,125 people found this helpful

Helpful Report abuse

Diamond Lovemore

★★★★★ Trustworthy

Reviewed in the United States on June 12, 2019

Size: 64GB | Color: Coral | Service Provider: Unlocked | Product grade: Renewed | **Verified Purchase**

This seller is amazing was a bit skeptical at first but this phone met my every expectation, very trustworthy in every way no it doesn't come with headphones. At least it comes with a charger & a 90 day warranty I also got a brand new spanking battery I've checked the serial number online it is worthy!!

942 people found this helpful

Helpful Report abuse

griffin cannon

★★★★★ FRAUDS- phones SOLD ARE STOLEN

Reviewed in the United States on August 20, 2019

Size: 64GB | Color: Black | Service Provider: AT&T | Product grade: Renewed | **Verified Purchase**

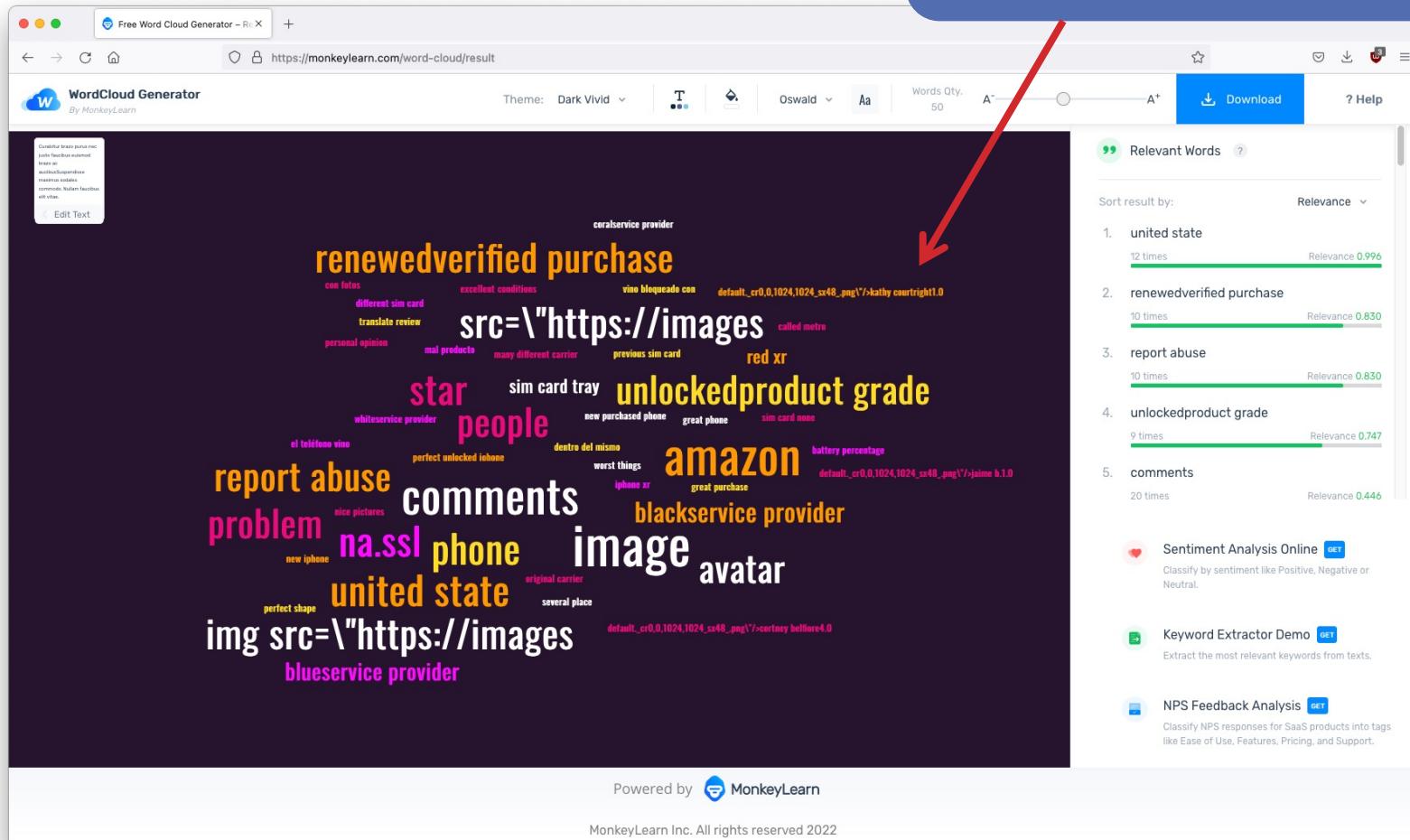
Just was told by my carrier, ATT, that this phone is blacklisted. The phone was blacklisted after 1 week of receiving the phone. THANKS AMAZON

698 people found this helpful

Helpful Report abuse

CHALLENGE

And use a word cloud generator to summarize the insights



<https://monkeylearn.com/word-cloud>

WHAT'S NEXT?

Cross-Industry Standard Process for Data Mining (CRISP-DM) & Predictive Analytics I