Text Analytics & Business Application

Introduction

Qinglai He

Department of Operations and Information Management

Wisconsin School of Business

WISCONSIN SCHOOL OF BUSINESS

Outline of Today's Class

- Course introduction
- Recap data analysis procedure
  - Data collection
  - Data cleaning

Our Course


data
techniques
Text
Mining
information
language
human
Abstraction
leverage
patterns
algorithms
learning
linguistics
problem
machine
common
method
another
called
visual
within
loss
novel
contexts
leverages
knowledge
developed
features
Data
user
order
field
new
dimensions

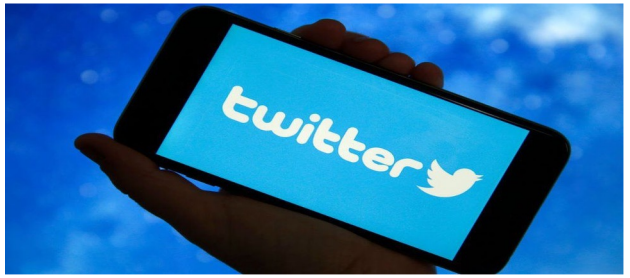
Text Mining Concepts

- Text mining ≈ Text analytics
- Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights.

Real-world Problems

- What is customers' overall sentiment about this restaurant?
- What is the most popular food here?
- What topics do customers talk about on Google review?

Real-world Problems



- How do people view a company? Is it positive or negative?
- Do we expect the stock price of a company to go up in the near future?
- What do people complain about a company?

NLP and Data Science Contractor

ITech Consulting Partners

Anywhere

Qualifications

- An academic background in a quantitative sciences, and or strong understanding of medical/clinical practice and areas of epidemiology, health economics and outcomes research
- at least 3 years of strong experience in Python programming and other software development/coding best practices
- Deep understanding with at least 2 years of experience in text mining, NLP pipelines, sentiment analysis, expertise in open-source text mining packages, algorithms
- Experience working in cloud-based environment, AWS services, knowledge graphs, semantic data models

Data Scientist

Whitespace

Remote • Remote

$95,000 - $125,000 a year - Full-time

You must create an Indeed account before continuing to apply

Apply on company site

Requirements:

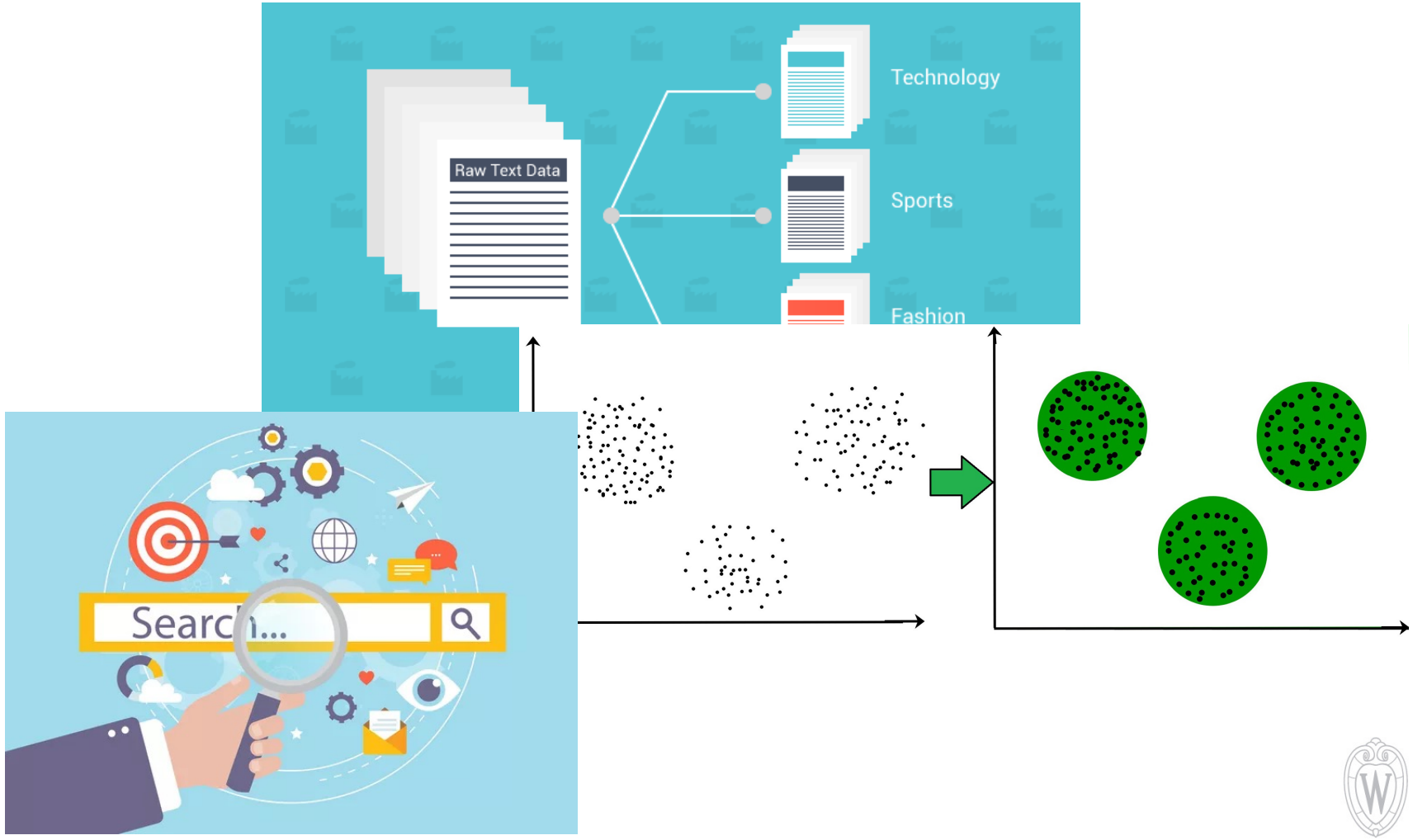
- Bachelors or Masters in a quantitative field (such as Engineering, Statistics, Math, Economics, or Computer Science with Modeling/Data Science), preferably with work experience of over 3 years.
- Ability to program in any high level language such as Python is required along with familiarity with common statistical analysis packages.
- Proven problem solving and debugging skills.
- Familiar with database technologies and tools (SQL/R/SAS/JMP etc.), data warehousing, transformation and processing. Work experience with real data for customer insights, business and market analysis will be advantageous.
- Experience with text analytics, data mining and social media analytics.

What's this course about?

Common Text Mining Techniques

- Text classification
- Text clustering
- Information extraction
- Information retrieval

…



Text Mining Applications

- Risk management
- Customer service
- Fraud detection
- Social media analysis

Learning Objectives

- Retrieve, assemble, and clean text data for use in analytics applications.
- Use text representation approaches, including the bag of words and TF-IDF, to represent text data.
- Understanding basic concept of supervised and unsupervised data mining skills.
- Use LDA for topic modeling in the context of a business application.

Course Schedule

Try to use the lecture time effectively.

Week	Dates	Topic	In-class Exercises	Assignments
1	01/23	Course Introduction & Data Analytics Recap	IC1	
2	01/30	NLP Foundation and Pipeline	IC2	
3	02/06	Text Representation	IC3	*Project Proposal*
4	02/13	Text Classification	IC4	Project Milestone1 (2/14, 11:59 PM)
5	02/20	Text Clustering	IC5	Assignment 1 (2/21, 11:59 PM)
6	02/27	Business Application / Lab Experiment 1: Text Analytics in E-commerce	IC6	
7	03/05	Topic Modeling	IC7	Project Milestone 2 (3/6, 11:59 PM)
8	03/12	Business Application / Lab Experiment 2: Text analytics in Finance	IC8	
9	03/19	Information Extraction (IE)	IC9	Assignment 2 (3/20, 11:59 PM)
10	03/26	Spring Recess		
11	04/02	Business Application / Lab Experiment 3: Text Analytics in Healthcare	IC10	Project Milestone 3 (4/3, 11:59 PM)
12	04/09	Chatbots	IC11	
13	04/16	Recommedation System	IC12	Assignment 3 (4/17, 11:59 PM)
14	04/23	Text Summarization & Group Presentation 1		
15	04/30	Group Presentation 2 & Course Wrap-up		Project Milestone 4 (5/1, 11:59 PM)

Assessment & Point Distribution

Deliverables	Points
Exercises (24%)	120 points
Participation (6%)	30 points
Assignments (38%)	190 points
Projects (32%)	160 points
Total points possible:	500 points
Important Dates	
Classes Begin:	Jan 23th, 2024

Can we get extra credits? **YES**

+ Bonus Point

We may have pop quizzes during each class meeting.

An in-class exercise is due on 01/24 at 11:59 pm.

Week	Dates	Topic	In-class Exercises	Assignments
1	01/23	Course Introduction & Data Analytics Recap	IC1	
2	01/30	NLP Foundation and Pipeline	IC2	
3	02/06	Text Representation	IC3	*Project Proposal*
4	02/13	Text Classification	IC4	Project Milestone1 (2/14, 11:59 PM)
5	02/20	Text Clustering	IC5	Assignment 1 (2/21, 11:59 PM)
6	02/27	Business Application / Lab Experiment 1: Text Analytics in E-commerce	IC6	

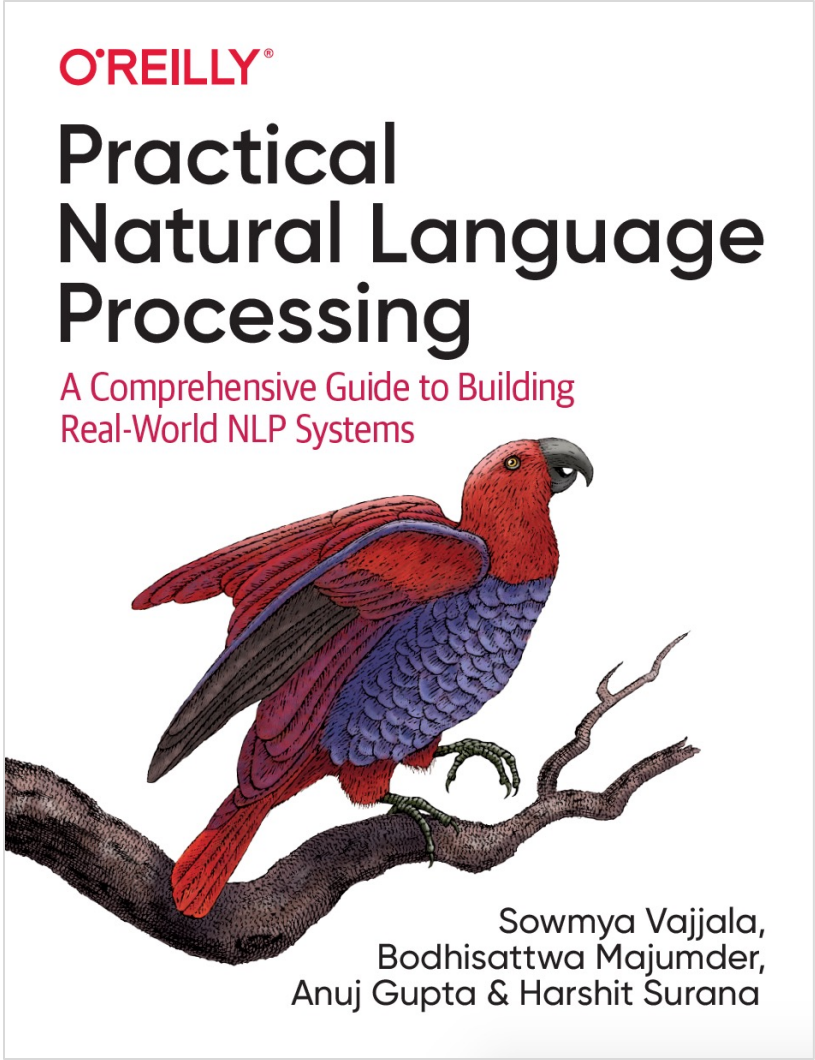
Project milestone (assignment) is due on 2/14 (2/21) at 11:59 pm.

Course policy & Other Info

#1 Recommended Textbook

- Practical Natural Language Processing.

https://www.amazon.com/Practical-Natural-Language-Processing-Pragmatic/dp/1492054054



#2 Other Resources

Various online resources can help you with many coding and concept related questions.

w3schools.com

coursera

kaggle

stackoverflow

Quora

#3 Participation

- Attendance is expected for all classes.
- Each student is allowed to miss a maximum of two classes throughout the semester. It is preferred that you inform your instructor in advance if you cannot attend the class.
- Every student has a full participation score (30 points) at the beginning of the semester. 10 points will be deducted from the participation score if students are absent from a class beyond two times without sending a notice to the instructor.

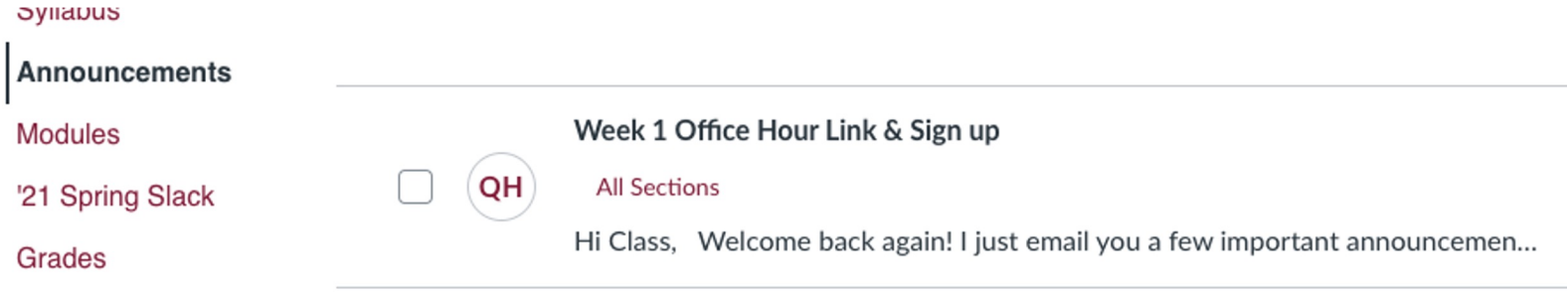
#4 Submission & Late submission

- Only one submission per exercise/assignment/group report is allowed.
- Each student has two opportunities for late submission. Submission that pass the deadline over 1 day will not be accepted.

#5 Grade Scale & Grade Appeal

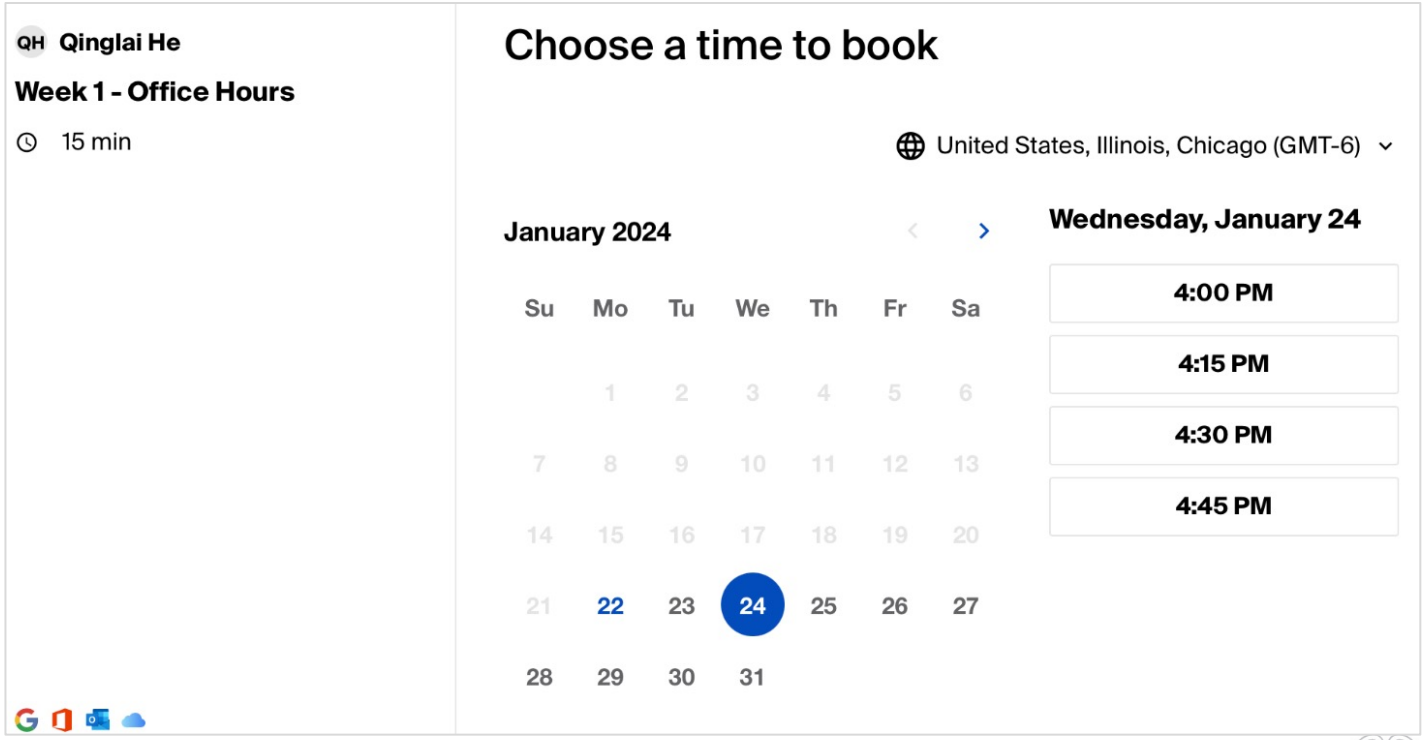
Grade	A	AB	B	BC	C	D	F
%	93-100	89-92.99	83-88.99	79-82.99	70-78.99	60-69.99	<60

- Grade appeal must be within 7 days of the posting of the grade.



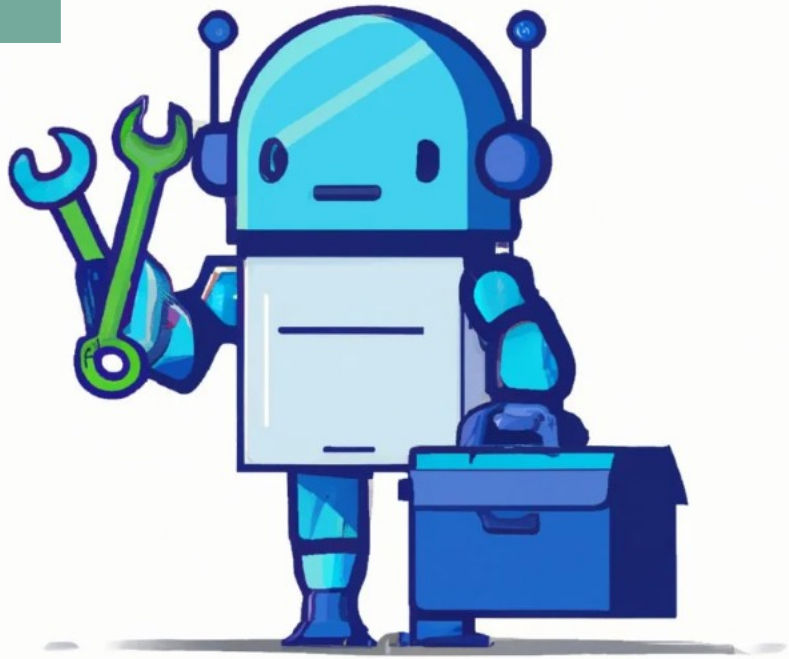
#6 Office Hour

- In-person and virtual office hours are offered
- M W 4-5 PM
- Please sign up for office hours via weekly Doodle link posted on Canvas.



#7 AI Policy

- You can get help from ChatGPT or other LLMs for coding and debugging, but you should acknowledge the usage of AI.
- You cannot use AI for exercises or assignments beyond coding questions.



#8 Course Feedback

- You are highly encouraged to provide course feedback to me so that we can make our course better for everyone!

UNIVERSITY of WISCONSIN–MADISON REPORT AN INCIDENT

Office of Student Conduct and Community Standards
Student Affairs

#9 Academic Integrity

- All in-class exercises and assignments are independent work.

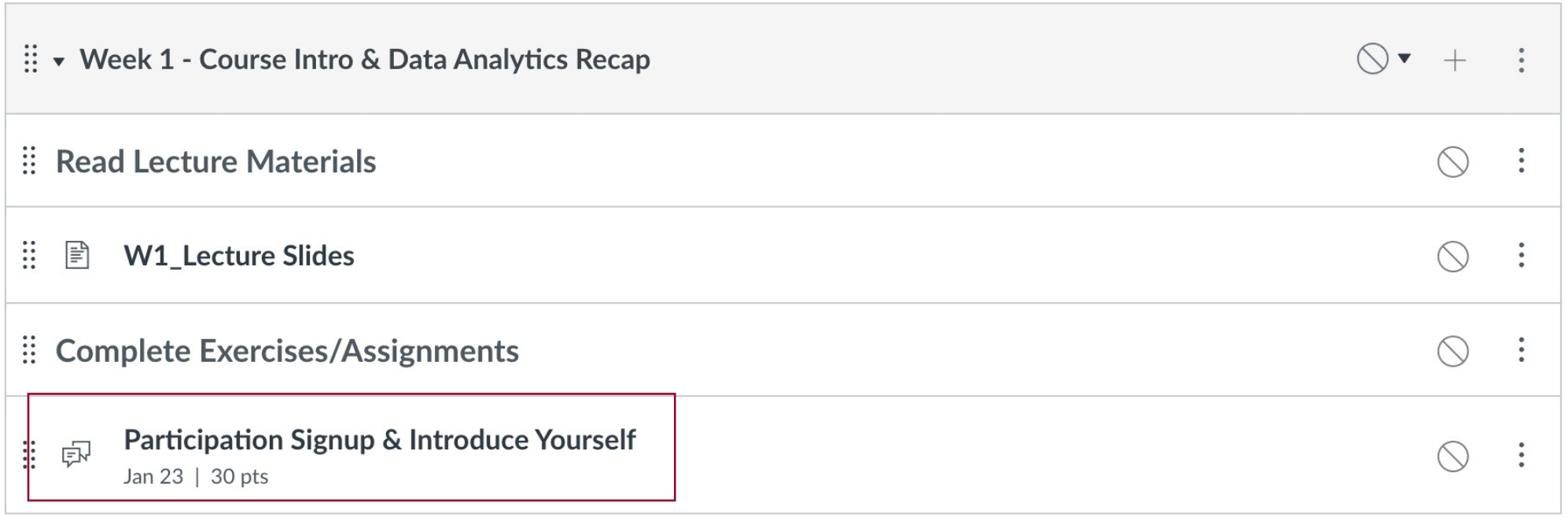
https://conduct.students.wisc.edu/academic-misconduct/

How to Succeed in This Course?

- Attend classes
- Understand the content and ask questions when you are confused about content
- Try to digest example code
- Complete and submit exercises/assignments on time
- Good team player

Any Questions?

What To Do Next?



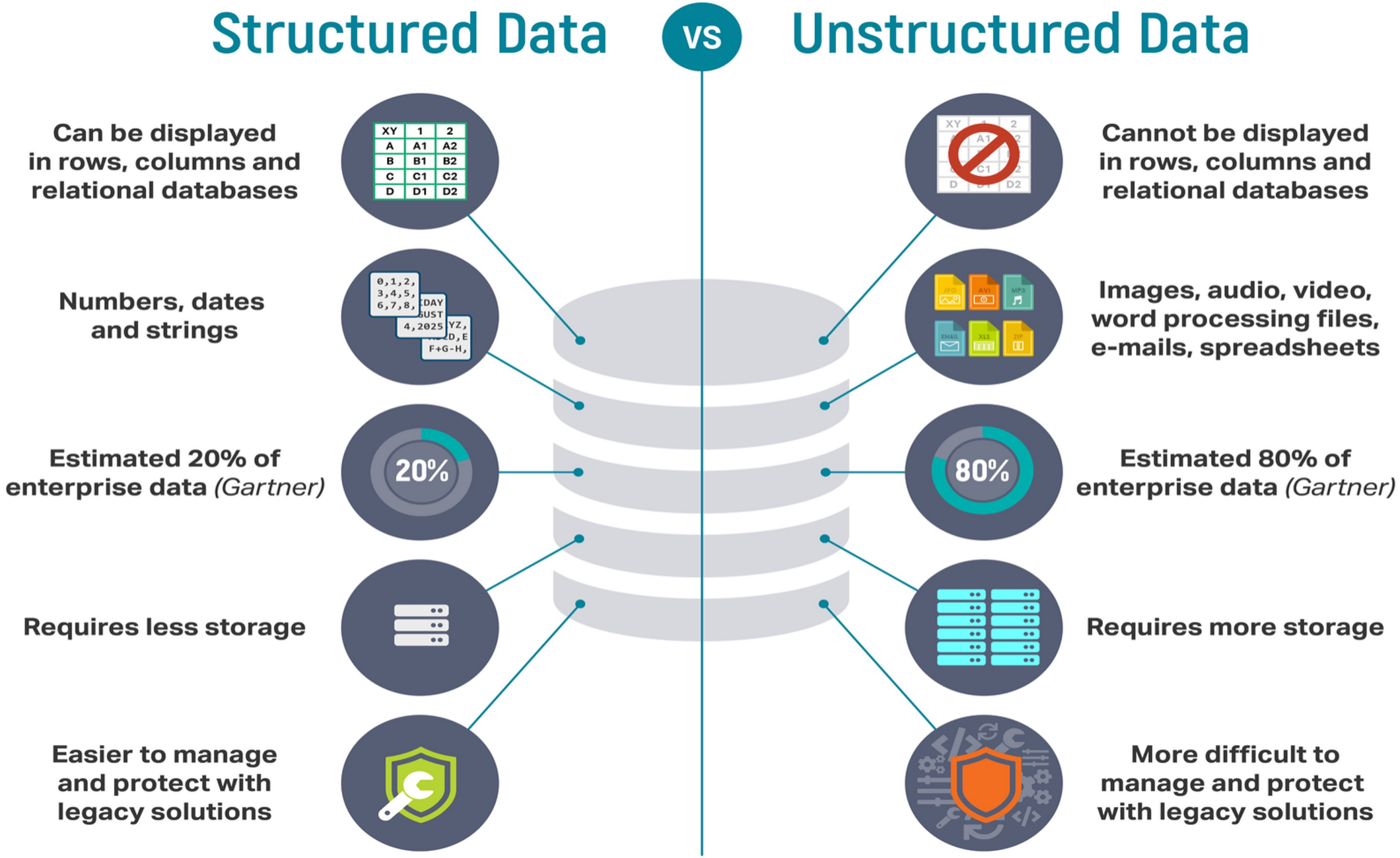
Let’s take 15 minutes break…

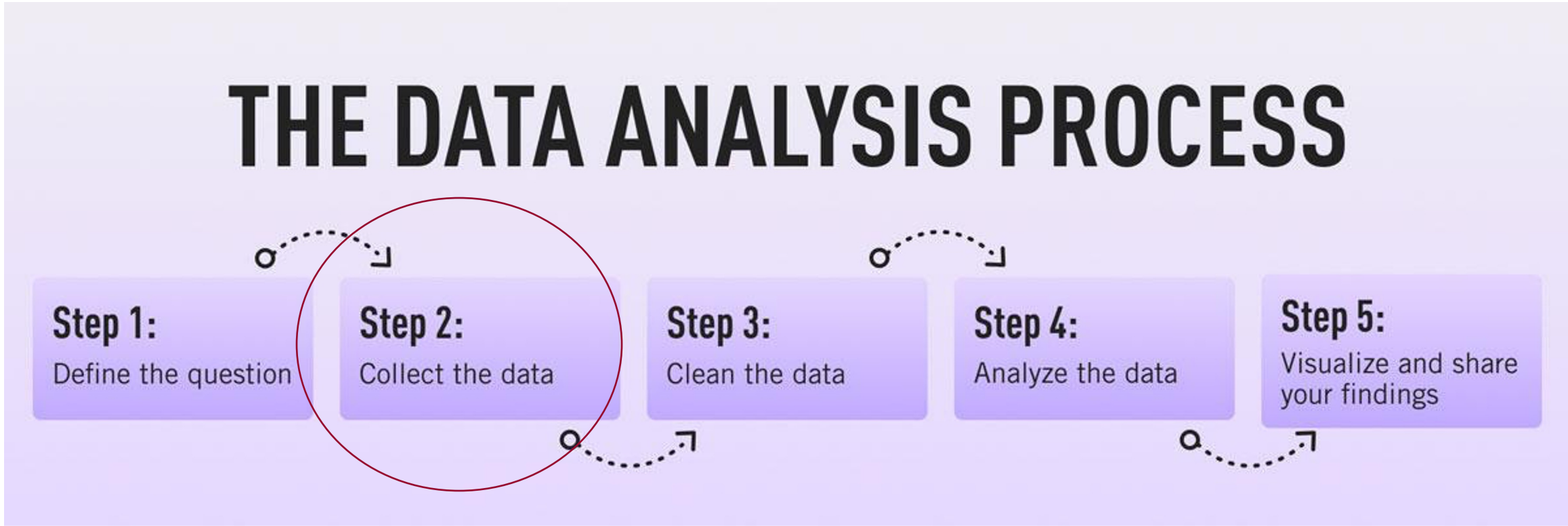
Data Analytics Recap

- Data type
- Data analysis process
  - Data collection – where to find data?
  - Data quality, pre-processing & transformation

Data Types

Structured Data vs Unstructured Data

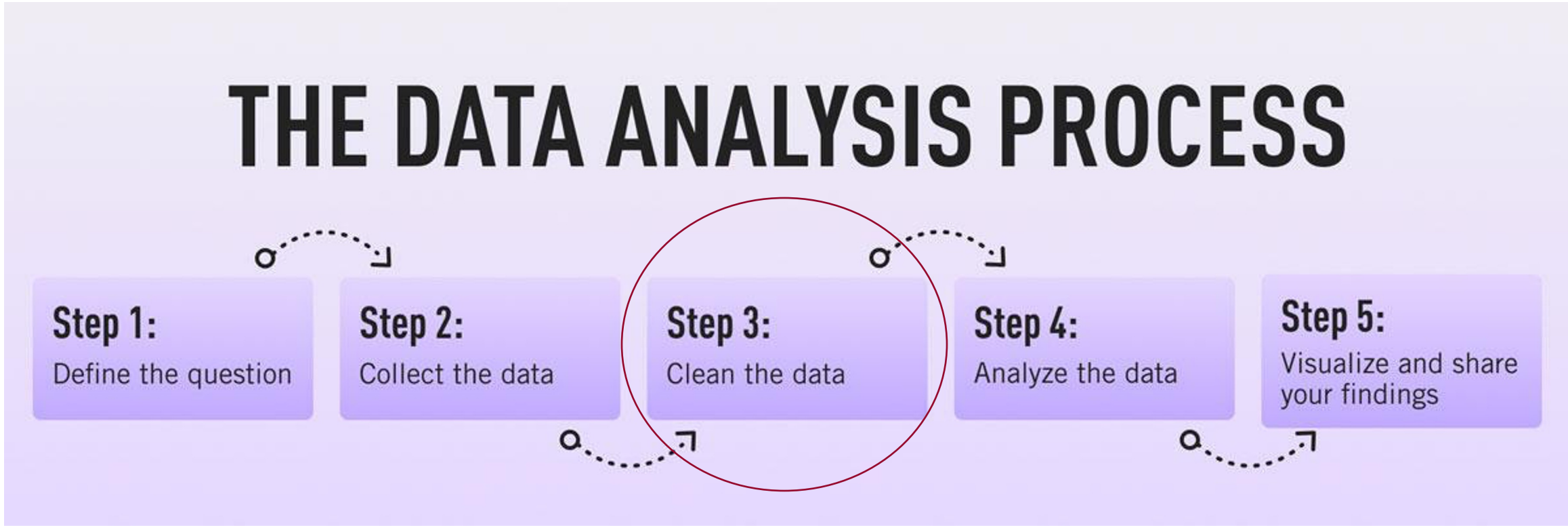



THE DATA ANALYSIS PROCESS
Step 1:
Define the question
Step 2:
Collect the data
Step 3:
Clean the data
Step 4:
Analyze the data
Step 5:
Visualize and share your findings

Data Collection – Where to Find Data?

- Public government data
  - New York City Open Data https://opendata.cityofnewyork.us/
  - Austin Open Data https://data.austintexas.gov/
  - U.S. Climate Data https://www.ncei.noaa.gov/weather-climate-links#ghcn
  - U.S. Census Bureau's Small Area Income and Poverty Estimates https://www.census.gov/programs-surveys/saipe.html
- Social media API
  - Twitter https://developer.twitter.com/en/docs/twitter-api
  - Reddit https://www.reddit.com/dev/api/
- Online code/data repository
  - Github (for example, COVID data ran by Johns Hopkins University https://github.com/CSSEGISandData/COVID-19)
  - Kaggle https://www.kaggle.com/
- Internal dataset
- Scaping data from the website? (be cautious about this. Need to follow platforms' term of use)
- Tableau list many free health, social impact, climate, government, education data.

  https://www.tableau.com/learn/articles/free-public-data-sets


THE DATA ANALYSIS PROCESS
Step 1:
Define the question
Step 2:
Collect the data
Step 3:
Clean the data
Step 4:
Analyze the data
Step 5:
Visualize and share your findings

Why Data Cleaning is Important?

Why Data Cleaning is Important?

- Garbage in, garbage out!
- Quality data beats fancy algorithm
- The dirty job you cannot ignore
- Most of the data analysis effort may be spent on data cleaning

Data Quality

- **Accurate** – correct, precise and up to date Inaccurate/incorrect value (e.g., typo)
- **Complete** – all possible data that is required is present Incomplete data
- **Conformant** – data is stored in an appropriate and standardized format Incorrect format
- **Consistent** – there are no conflicts in information within or between systems Inconsistent data (Can happen in the same column or cross columns)
- **Timely** – data is created, maintained and available quickly and as required Outdated data
- **Unique** – where appropriate, there are no duplicates or redundant data elements Duplicated and redundant data
- **Valid** – data is authentic, and proven to be valid, and derived from authentic and known sources Incorrect data

https://www.winshuttle.com/blog/good-data-quality-worth/

What's Wrong with This Data?

RESTAURANT	BORO	BUILDING	STREET	CUISINE DESCRIPTION	GRADE
1	Brooklyn	366	UTICA AVENUE	Caribbean	A
2	Manhattan	2245	1 AVENUE	Hawaiian	
3	Brkly	922A	4 AVENUE	Spanish	
4	Brooklyn	2	METROTECH CTR	Asian/Asian Fusion	A
5	Brooklyn	5324	8 AVENUE	Chinese	A
6	Brooklyn	5324	8 AVENUE	Chinese	A
7	Brooklyn	367	METROPOLITAN AVE	Mexican	B
8	Brooklyn	6918	3 AVENUE	Pizza	B
9	Brooklyn	733	KNICKERBOCKER AV	Indian	

Inconsistent data

What's Wrong with This Data?

RESTAURANT	BORO	BUILDING	STREET	CUISINE DESCRIPTION	GRADE
1	Brooklyn	366	UTICA AVENUE	Caribbean	A
2	Manhattan	2245	1 AVENUE	Hawaiian	
3	Brkly	922A	4 AVENUE	Spanish	
4	Brooklyn	2	METROTECH CTR	Asian/Asian Fusion	A
5	Brooklyn	5324	8 AVENUE	Chinese	A
6	Brooklyn	5324	8 AVENUE	Chinese	A
7	Brooklyn	367	METROPOLITAN AVE	Mexican	B
8	Brooklyn	6918	3 AVENUE	Pizza	B
9	Brooklyn	733	KNICKERBOCKER AV	Indian	

Duplicated data

What's Wrong with This Data?

RESTAURANT	BORO	BUILDING	STREET	CUISINE DESCRIPTION	GRADE
1	Brooklyn	366	UTICA AVENUE	Caribbean	A
2	Manhattan	2245	1 AVENUE	Hawaiian	
3	Brkly	922A	4 AVENUE	Spanish	
4	Brooklyn	2	METROTECH CTR	Asian/Asian Fusion	A
5	Brooklyn	5324	8 AVENUE	Chinese	A
6	Brooklyn	5324	8 AVENUE	Chinese	A
7	Brooklyn	367	METROPOLITAN AVE	Mexican	B
8	Brooklyn	6918	3 AVENUE	Pizza	B
9	Brooklyn	733	KNICKERBOCKER AV	Indian	

Missing data

What is Data Cleaning?

- Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

- If data is **incorrect**, outcomes and algorithms are **unreliable**, even though they may look correct.

- Common types of data issues:
  - Missing, Inconsistent, Irrelevant, Duplicated, Outliers

- No absolute way to prescribe the exact steps in the data cleaning process
  - The processes will vary from dataset to dataset.

https://www.tableau.com/learn/articles/what-is-data-cleaning

1- Missing Data

- A common quality problem found in data sets both large and small.
  - Missing data can cause significant reduction in usable observations and bias results
  - Example: 20 variables with 5% of the values randomly missing across observations and variables => only 35.85% of the observations are usable

- Understanding **why** the values are missing. The reasons could be..
  - For survey data, respondents decline to provide the information due to its sensitive nature
  - For survey data, some of the items do not apply to every respondent
  - Human errors, sloppy data collection, or equipment failures

Dealing with Missing Data - Omission

- Omission: complete-case analysis
  - Exclude observations with missing values
  - Appropriate when the amount of missing value is **small** or concentrated in a small number of observations
  - If the variable that has many missing values is deemed **unimportant** or can be represented using a proxy variable that does not have missing values, the variable may be excluded from the analysis.

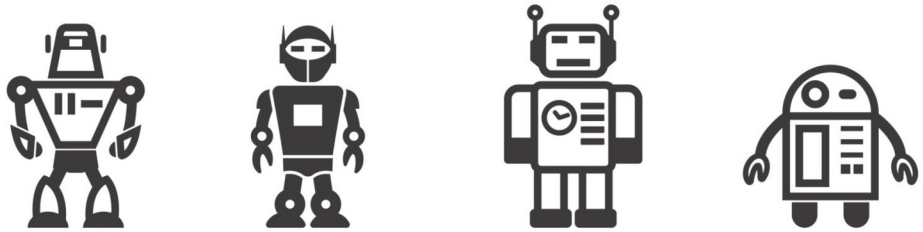
Dealing with Missing Data - Imputation

- Imputation: replace missing values with some reasonable values.
- For **numerical variables**, the mean/average is often used.
  - Easy to implement
  - Doesn't increase the variability in the data set
  - If a large number of values are missing, mean imputation will likely distort the relationships among variables, leading to biased results
- For **categorical variables**, the most frequent category is often used.
  - Flag the missing data by creating an "unknown" category
  - Useful for when data are missing for a reason
- Based on the data context, choose the right way to fix the missing data.

2 - Inconsistent Data

- Data inconsistency include **strange naming conventions**, **typos**, or **incorrect capitalization**. These inconsistencies can cause mislabeled categories or classes.
    - For example, you may find “N/A” and “Not Applicable” both appear, but they should be analyzed as the same category.

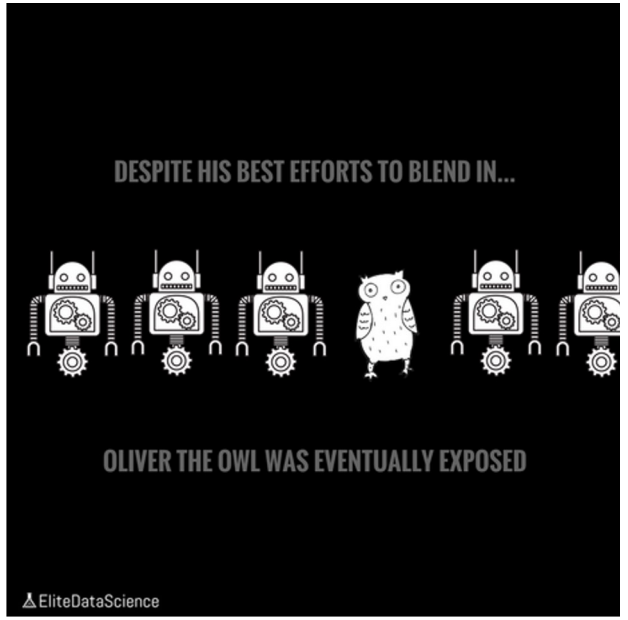
- Solutions: Modify the inconsistent values to enforce the data format and values are consistent.



3 - Duplicated Data

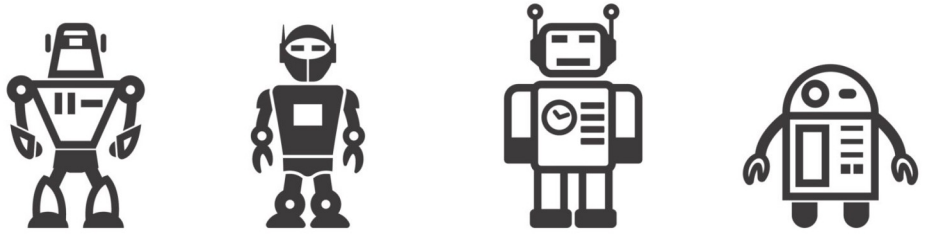
- When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data.

- Solutions: remove duplicate observations/rows



4 - Irrelevant Data

- Irrelevant data are when you notice observations that do not fit into the specific problem you are trying to analyze.
    - For example, if you want to analyze data regarding millennial customers, but your dataset includes older generations, you might remove those irrelevant observations.
- Solutions: Remove irrelevant observations/rows or variables/columns.

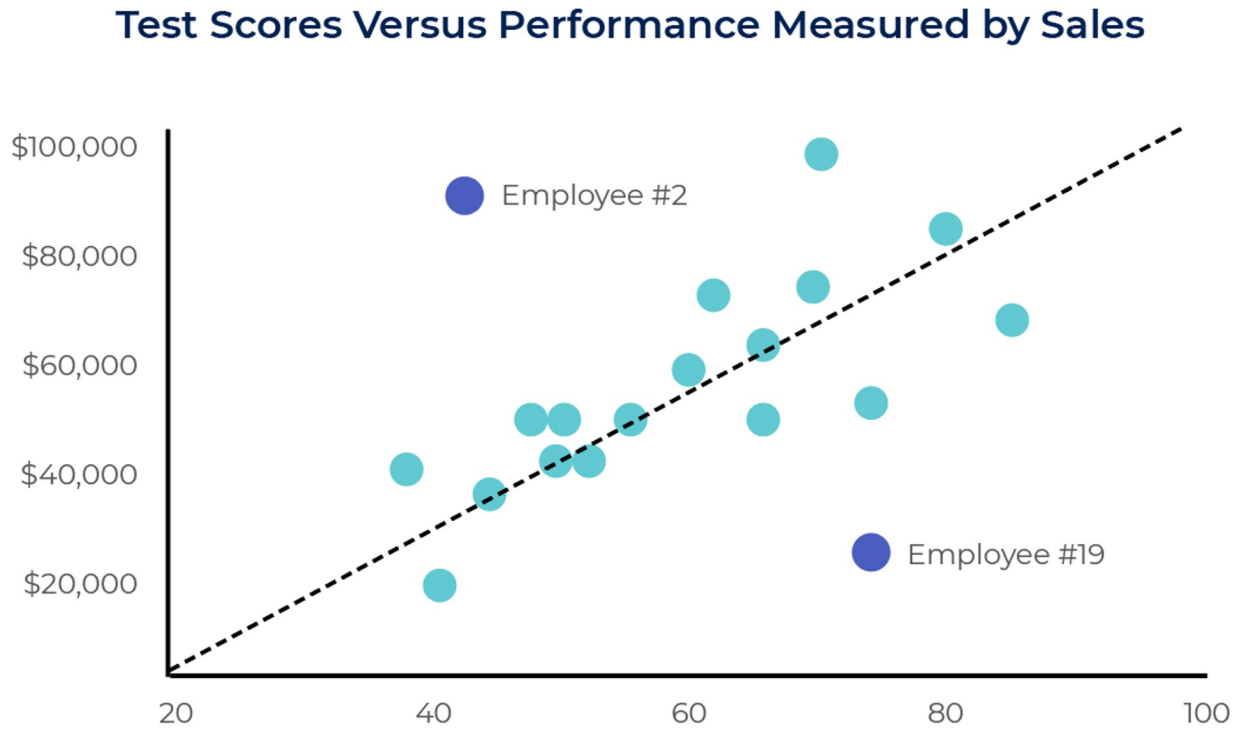


5 - Outliers

- Outliers are extremely large or small observations in data sets
- Influence summary statistics
    - Mean, standard deviation,...

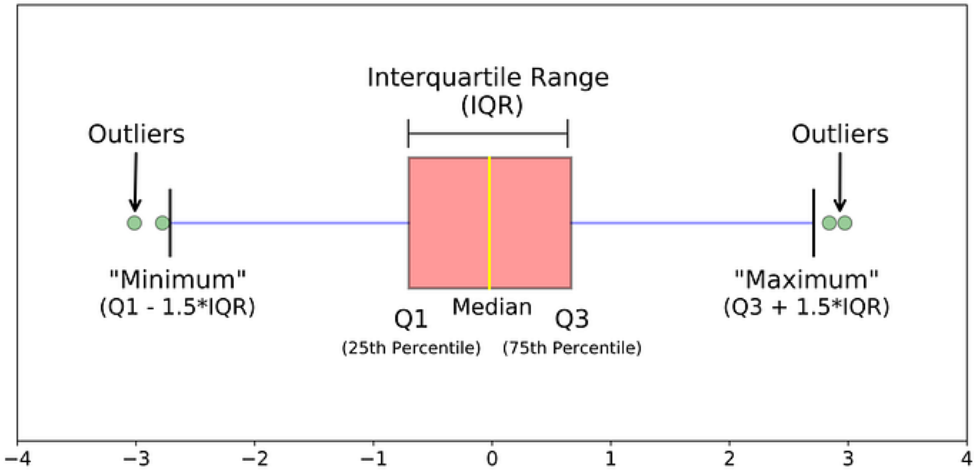
What causes outliers?

- Random variations
    - Solution: Keep it
- Incorrectly recorded observations
    - Solution: Correct or delete it



Identify Outliers

- Visual inspection
  - Boxplot
  - Histogram
  - Q-Q plot (data distribution is known)
- Interquartile Range (IQR)
  - $IQR = Q_3 - Q_1$
  - Left fence: $Q_1 - 1.5IQR$
  - Right fence: $Q_3 + 1.5IQR$
- Z-score method
  - $z = \frac{x-\mu}{\sigma}$
  - Threshold: -3, 3 (i.e. $\mu \pm 3\sigma$ in normal distribution)
- Modified Z-score method
  - $z = 0.6745\frac{x-\mu}{MAD}$ $(MAD: median\ absolute\ deviation)$
  - Minimize the influence of outliers on Z-score



Review_id	Review_cnt	Review_norm
1	10	0.009
2	20	0.018
3	1000	0.909
4	40	0.036
5	30	0.027

https://medium.com/analytics-vidhya/how-to-remove-outliers-for-machine-learning-24620c4657e8

https://towardsdatascience.com/outlier-detection-part1-821d714524c

Deal with Outliers

- Deleting observations
  - When outlier is caused by data entry error, data processing error or outlier observations are very small in numbers.
  - Delete the observation is not a good idea when we have small dataset.
- Transforming values
  - Scaling: normalization, standardization
  - Log transformation
  - Cube Root Normalization
  - BoxCox transformation
- Imputation
  - Mean, median, 0…
- Separately treating

6 – Feature Scaling

- Importance of feature scaling
  - Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled.
  - Outliers are extremely large or small observations in data sets Influence summary statistics.
  - Tree-based models are less affected by scaling.

- Algorithms require scaling
  - K-nearest neighbors with a Euclidean distance
  - Logistic regression, SVM, perceptron, neural network
  - K-means
  - Linear discriminant analysis, PCA, kernel principal component analysis

https://www.geeksforgeeks.org/normalization-vs-standardization/ https://dataakkadian.medium.com/standardization-vs-normalization-da7a3a308c64

Feature Scaling methods

	Standardization (z-score)	Normalization (min-max)
Formula	$\frac{x - mean}{std}$	$\frac{x - min}{max - min} (if\ x = x_{min}, x = 0; if\ x = x_{max}, x = 1)$
Range	No boundaries	[0,1]
Outliers	Much less affected by outliers	Affected by outliers
Sklearn	*StandardScaler*	*MinMaxScaler*
Use case	• Ensure zero mean and unit standard deviation • When the feature distribution is Gaussian	• When features are of different scales • When we don't know about the distribution

Set up Google Colab & Example Code for Data Cleaning