Text Analytics & Business Application

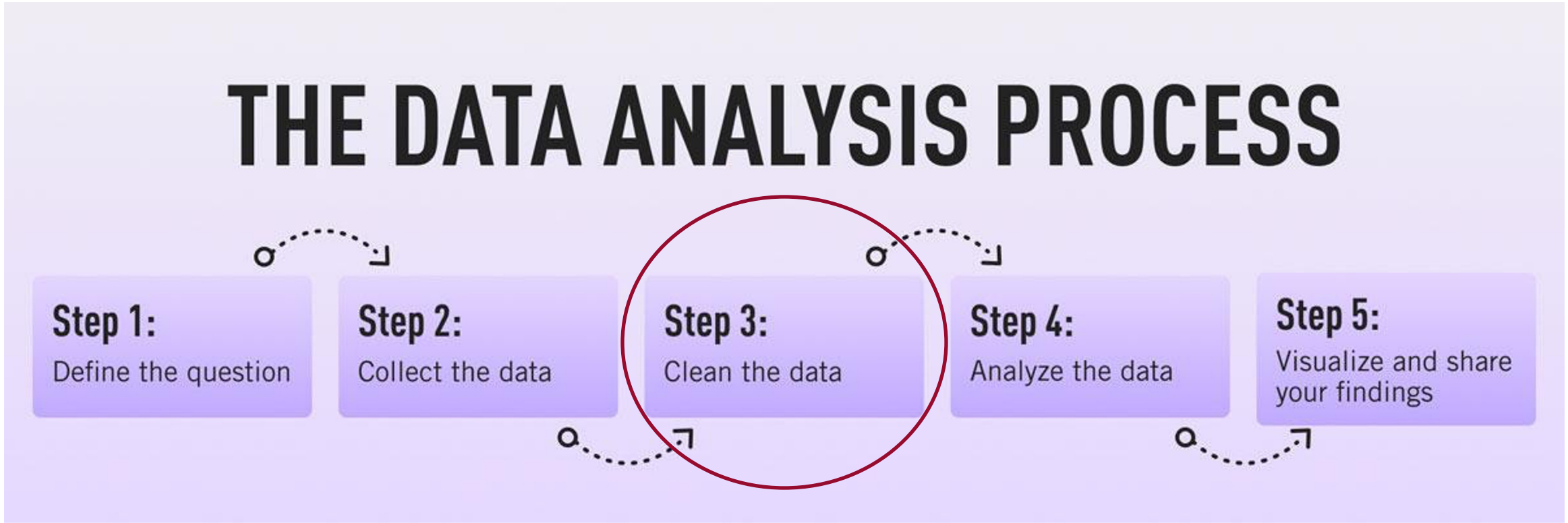
NLP Foundation & Pipeline

Qinglai He

Department of Operations and Information Management

Wisconsin School of Business

WISCONSIN SCHOOL OF BUSINESS


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

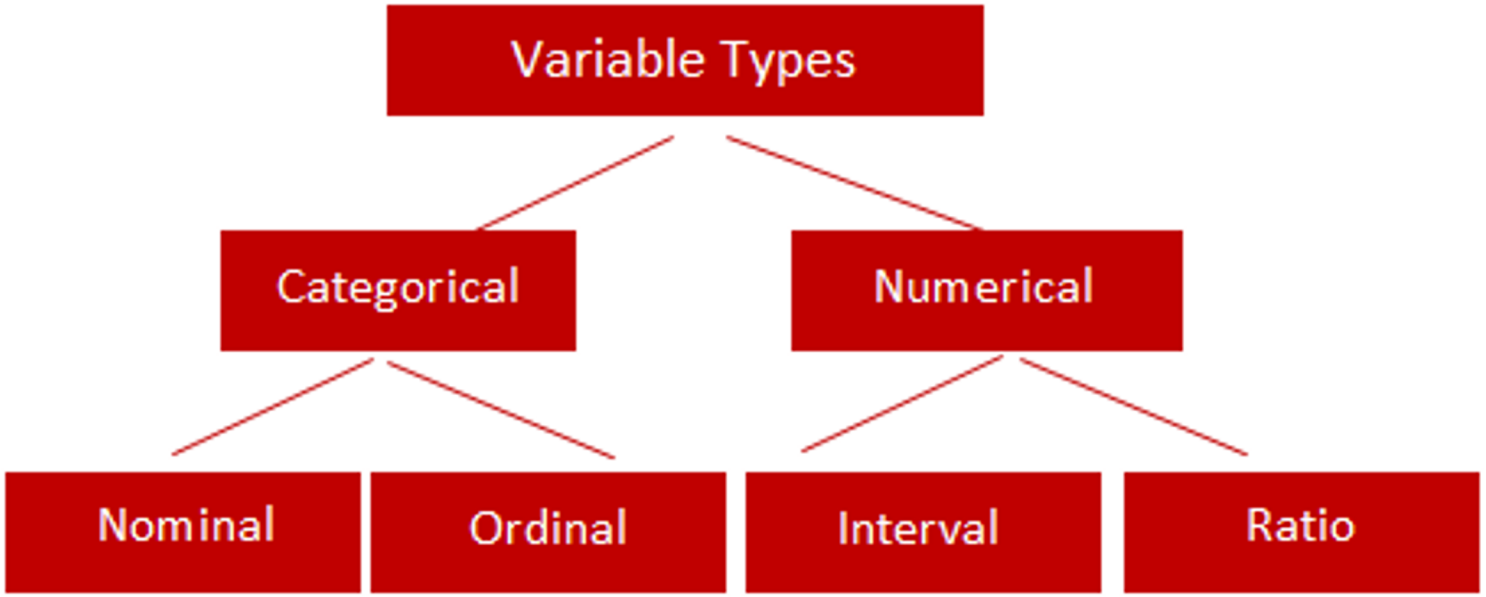
Outline of Today's Class

- Data Analytics Recap
  - Data Pre-processing
    - ~~Data cleaning~~
    - Data transformation
  - Four Types of Data Analytics
- Group Project & Milestone 1
- Intro to NLP
- Text Preprocessing

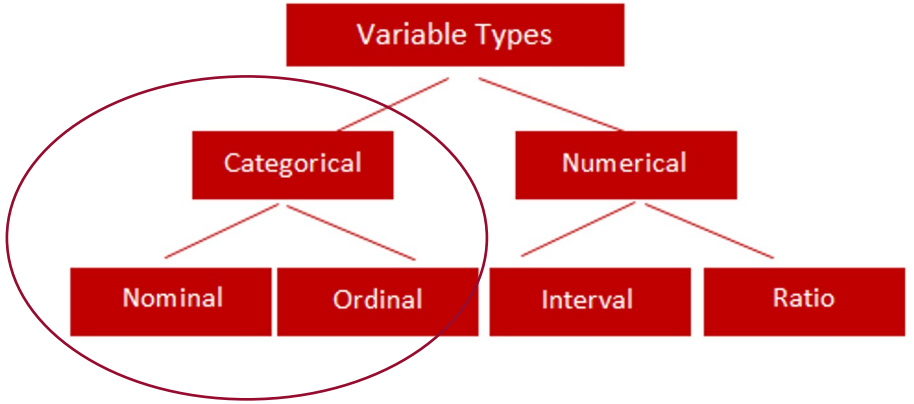
Variable Types

Variables

- A variable is a characteristic of interest that differs in kind or degree among various observations (records).

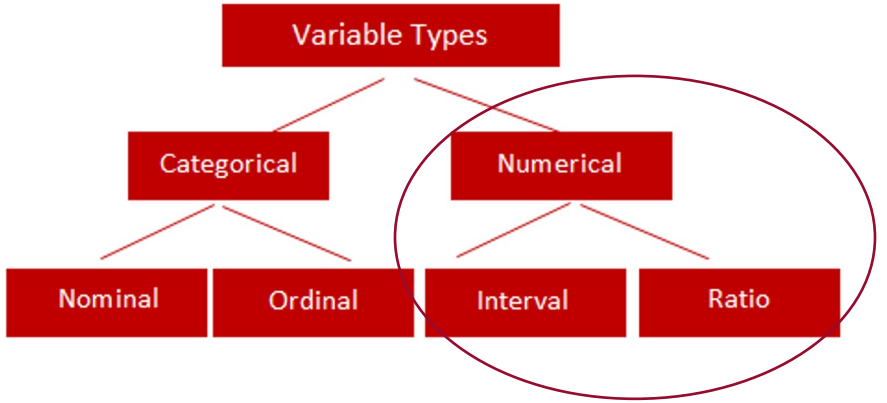


Categorical Variable



- Also called **Qualitative Variable**
- Analysis techniques depend on the type of data.
- Nominal
  - Least sophisticated
  - Values differ by label or name
  - Example: marital status (single/married)
  - Unique case: dummy/binary variables (e.g., TRUE/FALSE)
- Ordinal
  - Reflect labels or name, but can be ranked
  - Cannot interpret the difference between the ranked values
  - Example: reviews from 1 star (poor) to 5 stars (outstanding)

Numerical Variables



- Also called **Quantitative Variables**
- Interval
  - E.g., rank, differences are meaningful
  - Zero value is arbitrary and does not reflect absence of characteristic
  - Ratios are NOT meaningful
  - Example: temperature (74°F)
    - Why? (74°F - 72°F) = 2 °F (the result is meaningful), but (72°F)/(74°F) = ? (the result is not meaningful)
- Ratio
  - Most sophisticated
  - A true zero point, reflects absence of characteristic
  - Ratios are meaningful
  - Example: profits ($142,342,453)
    - Why? ($12 - $10) = $2 (the result is meaningful), and $12/$10 = 1.2 (the result is meaningful)

Q1. Price is a categorical variable.

A. True
B. False

Answer: B

Q2. Zipcode is a numerical variable.

A. True
B. False

Answer: B

Q3. COVID test result is a categorical variable.

A. True
B. False

Answer: A

Q4. The owner of a ski resort gathers data on tweens. Which of the following variable is nominal?

Tween	Music Streaming	Food Quality	Closing Time	Own Money Spent ($)
1	Apple Music	4	5:00 pm	20
2	Pandora	2	5:00 pm	10
⋮	⋮	⋮	⋮	⋮
20	Spotify	2	4:30 pm	10

A. Music Streaming

B. Food quality

C. Closing time

D. Own money spent

Answer: A

Q5. The owner of a ski resort gathers data on tweens. Which of the following variable is ordinal?

Tween	Music Streaming	Food Quality	Closing Time	Own Money Spent ($)
1	Apple Music	4	5:00 pm	20
2	Pandora	2	5:00 pm	10
⋮	⋮	⋮	⋮	⋮
20	Spotify	2	4:30 pm	10

A. Music Streaming

B. Food quality

C. Closing time

D. Own money spent

Answer: B

Q6. The owner of a ski resort gathers data on tweens. Which of the following variable is ratio?

Tween	Music Streaming	Food Quality	Closing Time	Own Money Spent ($)
1	Apple Music	4	5:00 pm	20
2	Pandora	2	5:00 pm	10
⋮	⋮	⋮	⋮	⋮
20	Spotify	2	4:30 pm	10

A. Music Streaming

B. Food quality

C. Closing time

D. Own money spent

Answer: D

Q7. The owner of a ski resort gathers data on tweens. Which of the following variable is interval?

Tween	Music Streaming	Food Quality	Closing Time	Own Money Spent ($)
1	Apple Music	4	5:00 pm	20
2	Pandora	2	5:00 pm	10
⋮	⋮	⋮	⋮	⋮
20	Spotify	2	4:30 pm	10

A. Music Streaming

B. Food quality

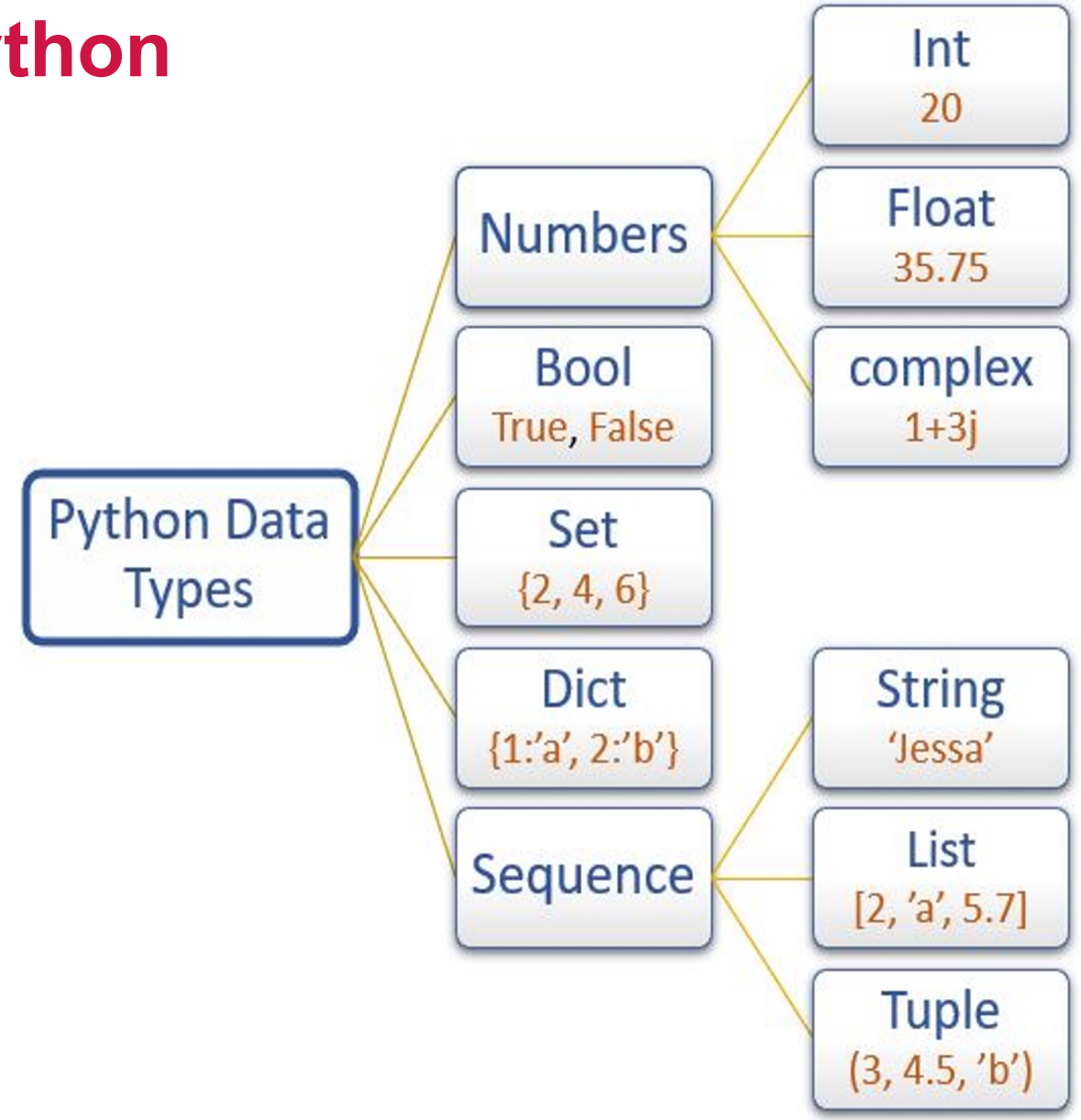
C. Closing time

D. Own money spent

Answer: C

Data/Variable Type in Python

- INT (numerical)
- Float (numerical)
- String (categorical)



What is Data Transformation?

- It is also called as data wrangling or data munging.
- The data conversion process from one variable type (format or structure) to another.
- It is an important step in bringing out the information in the data set, which can then be used for further data analysis.

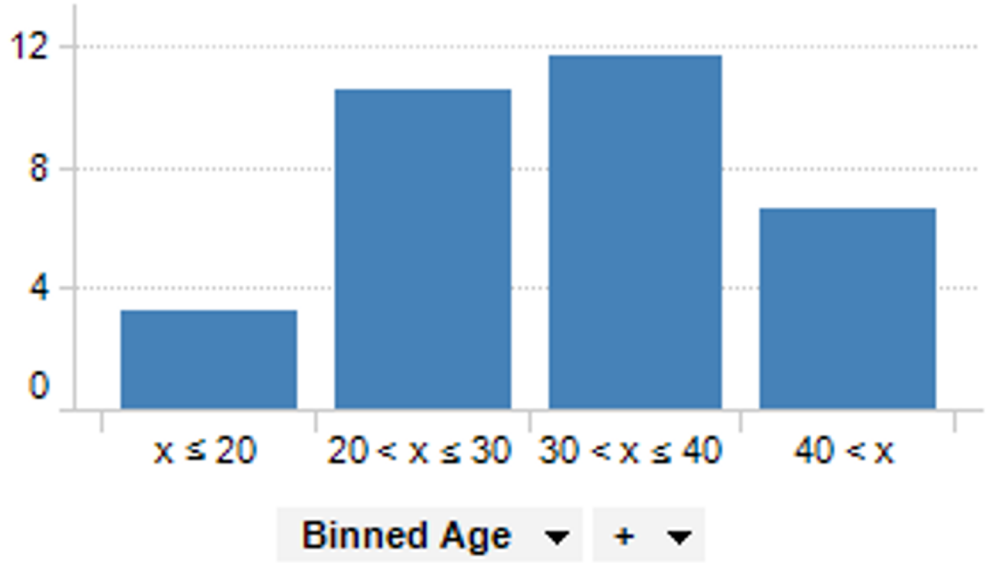
Data Cleaning vs. Data Transformation

- Data cleaning does not change the data/variable type.
- Data transformation changes the data/variable type.

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

1. Transforming Numerical Data - Binning

- Numerical → categorical
  - Grouping numerical values into a small number of bins.
- Bins are consecutive and non-overlapping
- Reduce **noise** (e.g., outliers) in the data
  - If we believe that all observations in the same bin tend to behave the same way.



2. Transforming Numerical Data - New Variables

- Numerical → Numerical
- **Scale problem**. Sometimes, data on variables such as income, firm size, and house prices are highly skewed. Extremely high (or low) values of skewed variables significantly inflate the average for the entire data set. It is difficult to detect meaningful relationships with skewed variables
- **Data standardization or normalization**. We can apply feature scaling techniques (e.g., natural logarithm and square root) to reduce data skewness.

Supplementary Content 1- 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

Supplementary Content 1- 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

2. Transforming Numerical Data - New Variables (Cont.)

- Another common data transformation involves calendar dates.
- Statistical software usually stores date values as numbers.
  - Excel stores date values by using a reference value of 1 for January 1, 1900.
- Transforming “Date” values into “Weekday” helps enrich the data set by creating relevant variables to support subsequent analyses.

Date_value	Date	Weekday
25	1/25/1900	3
78	3/18/1900	7
156	6/4/1900	1
986	9/12/1902	5
431	3/6/1901	3

3. Transforming Categorical Data – Dummy Variables

- Categorical → numerical
- Categorical variables must first be converted into numerical variables
  - Many analysis techniques are limited in their abilities to handle categorical data
  - A dummy variable (value 0/1) is commonly used

Income	Age	Marital Status
$45,000	23	Single
$48,000	25	Single
$54,000	24	Single
$57,000	29	Single
$65,000	38	Married
$69,000	36	Single
$78,000	40	Married
$83,000	59	Divorced
$98,000	56	Divorced
$104,000	64	Married
$107,000	53	Married

→

Income	Age	Married	Divorced
$45,000	23	0	0
$48,000	25	0	0
$54,000	24	0	0
$57,000	29	0	0
$65,000	38	1	0
$69,000	36	0	0
$78,000	40	1	0
$83,000	59	0	1
$98,000	56	0	1
$104,000	64	1	0
$107,000	53	1	0

- Oftentimes, a categorical variable is defined by more than two categories.
  - Given k categories of a variable, the general rule is to use (k - 1) dummy variables in the analysis, using the last category as reference.

3. Transforming Categorical Data – Ordinal Data

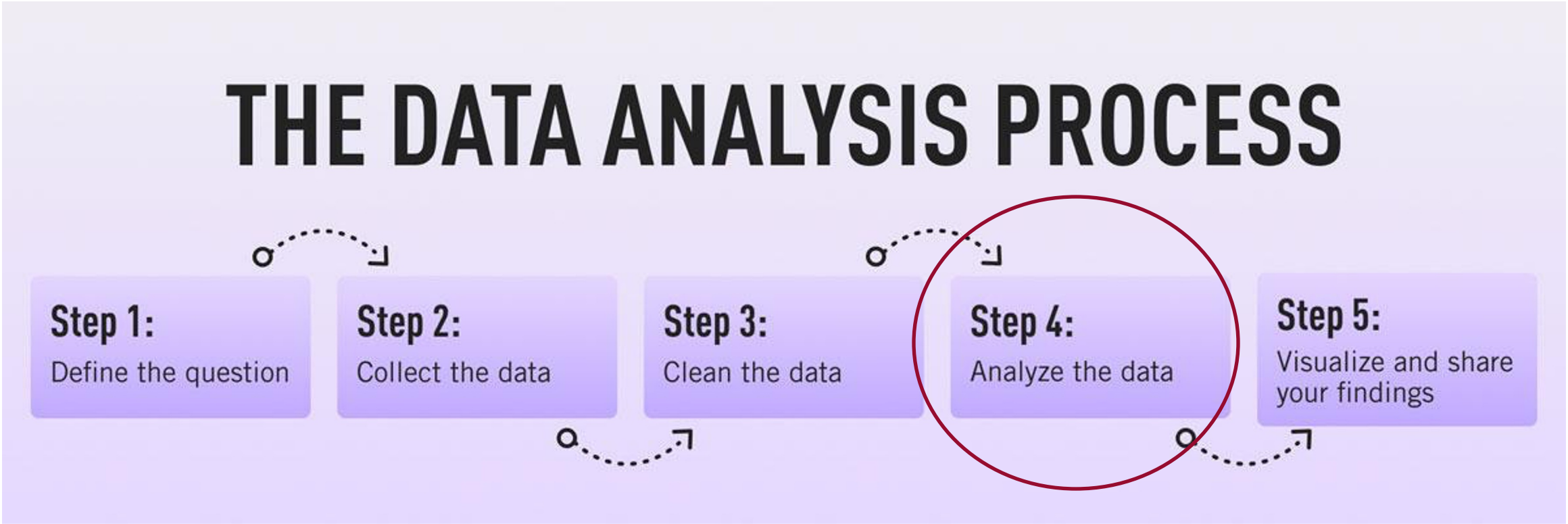
- Categorical → numerical
- Most appropriate if the data are ordinal and have natural, ordered categories.
- Allows the categorical variable to be treated as a numerical variable in certain analytical models.
- For an effective transformation, however, we assume equal increments between the category scores
    - May not be appropriate in certain situations.

Customer Satisfaction	Scores
Very Dissatisfied	1
Somewhat Dissatisfied	2
Neutral	3
Somewhat Satisfied	4
Very Satisfied	5

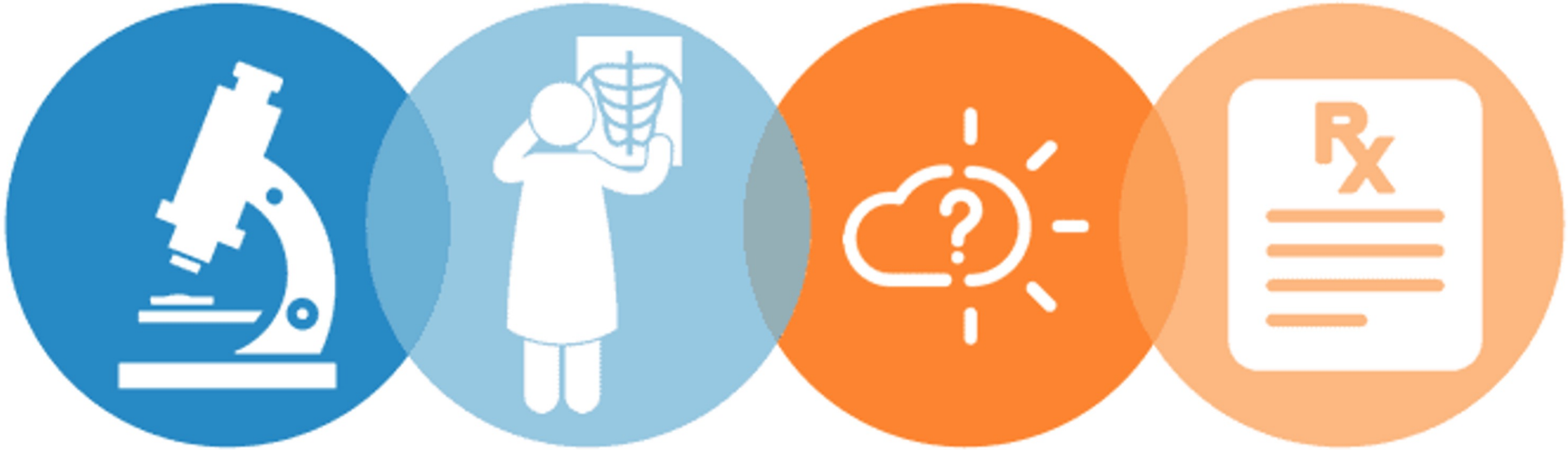
4. Transforming Categorical Data

- Categorical → Categorical
- It normally happens to category reduction - where we collapse some categories to create fewer non-overlapping categories.
- Why category reduction? Too many categories a variable has may lead to poor model performance.
- Determining the appropriate number of categories often depends on the data, context, and disciplinary norms, but there are a few general guidelines:
  - (1) Categories with very few observations may be combined to create the “Other” category.
  - (2) Categories with a similar impact may be combined.

Example Code for Data Transformation


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

Four Types of Data Analytics



Descriptive
Explains what happened.

Diagnostic
Explains why it happened.

Predictive
Forecasts what might happen.

Prescriptive
Recommends an action based on the forecast.
simulation, or optimization.

1. Descriptive Analytics

- What has happened?
  - Gather, Organize, Tabulate, Visualize, Summarize
  - Focus on the past
  - Easy to visualize
- Examples
  - A firm's marketing expenses and sales
  - Financial reports
  - Crime rates across regions and time
- Exploratory data analysis (EDA) is an important approach to do descriptive analytics

Exploratory Data Analysis (EDA)

- An important descriptive data analytics
- An important method of data diagnosis/inspection
- Often draws out insights using visualizations
  - e.g., graphs and plots.

https://careerfoundry.com/en/blog/data-analytics/exploratory-data-analysis/

Why EDA?

- Spot missing and incorrect data
- Determine error margins
- Calculate the most important statistics values
- Understand the underlying structure of your data
- Identify the most appropriate statistical tools
- …

https://careerfoundry.com/en/blog/data-analytics/exploratory-data-analysis/

Common Ways to do EDA

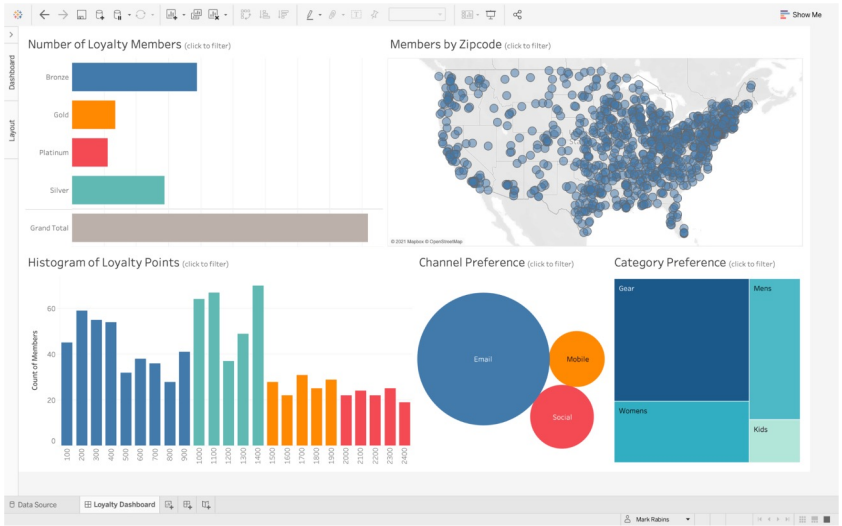
Values Diagnosis

- Summarizes the features and characteristics (e.g., summary statistics) of a dataset.
- Check the unique values of each variable. Any strange values? Any missing values? Inconsistent values?
- Check the minimum/maximum/mean/median value and standard deviation of data
- Check the correlation of different variables
- With the aggregate data, we can look at the value distribution of variable
- …

	A	B	C	D	E	F	G	H
1	Country	Canada						
2								
3	**Row Labels**	**Sum of Amount**						
4	Banana	33775						
5	Apple	24867						
6	Orange	19929						
7	Broccoli	12407						
8	Mango	3767						
9	**Grand Total**	**94745**						
10								
11								
12								
13								
14								
15								

Tools

- Excel, Tableau, R, Python…
- Visualization
- …



https://help.tableau.com/current/pro/desktop/en-us/what_chart_example.htm

https://r4ds.had.co.nz/exploratory-data-analysis.html

EDA Example - Netflix

- The longest movie/TV-show?
- The oldest movie/TV-show?
- The most common rating on Netflix?
- The least common rating on Netflix?
- Average duration of movie produced in 80s?

2. Diagnostic Analytics

- Why did something happen?
  - Also focus on the past.
  - Look for cause and effect to illustrate why something happened. The objective is to compare past occurrences to determine causes.
- Explaining why things are happening.
- Rely on causal inference methods (*not the focus of this course*), and help troubleshoot issues.
- Examples
  - Work from home boost the work productivity by 20% through the reduced commute time.
  - Machines lead to decreased demand in employee in low-end position because routine work has been substituted by the machines.
  - Machines lead to increased demand in employee in manager role because more managerial roles are needed to manage new production procedure and non-routine tasks.

3. Predictive Analytics

- What will happen in the future?
  - Use historical data to make predictions
  - Analytical models help identify correlation
  - Build models that help an organization understand what might happen in the future
  - Commonly use statistical methods and data mining
- Examples
  - Identifying customers who are most likely to respond to specific marketing campaigns
  - Transactions that are likely to be fraudulent
  - Incidence of crime at certain regions and times

4. Prescriptive Analytics

- What should we do next? (*not the focus of this course*)
  - Commonly use statistical, data mining, simulation, optimization
- Based on current data analytics, predefined future plans and goals
- Focus on providing advice and suggesting course of action
- Examples
  - Scheduling employees' works hours
  - Select a mix of products to manufacture
  - Choose an investment portfolio

Q8. Please answer what analytics the following statement is about
"The unemployment rate peaked above levels seen in the Great Recession"

A. Descriptive analytics
B. Diagnostic analytics
C. Predictive analytics
D. Prescriptive analytics

Answer: A

Q9. Please answer what analytics the following statement is about
"Production and transportation occupation had the second-largest increases. The pandemic and effort to contain it had a substantial impact on these occupations."

A. Descriptive analytics
B. Diagnostic analytics
C. Predictive analytics
D. Prescriptive analytics

Answer: B

Q10. Please answer what analytics the following statement is about
"Companies should offer flexible work-from-home option to attract talents in 2022 ."

A. Descriptive analytics
B. Diagnostic analytics
C. Predictive analytics
D. Prescriptive analytics

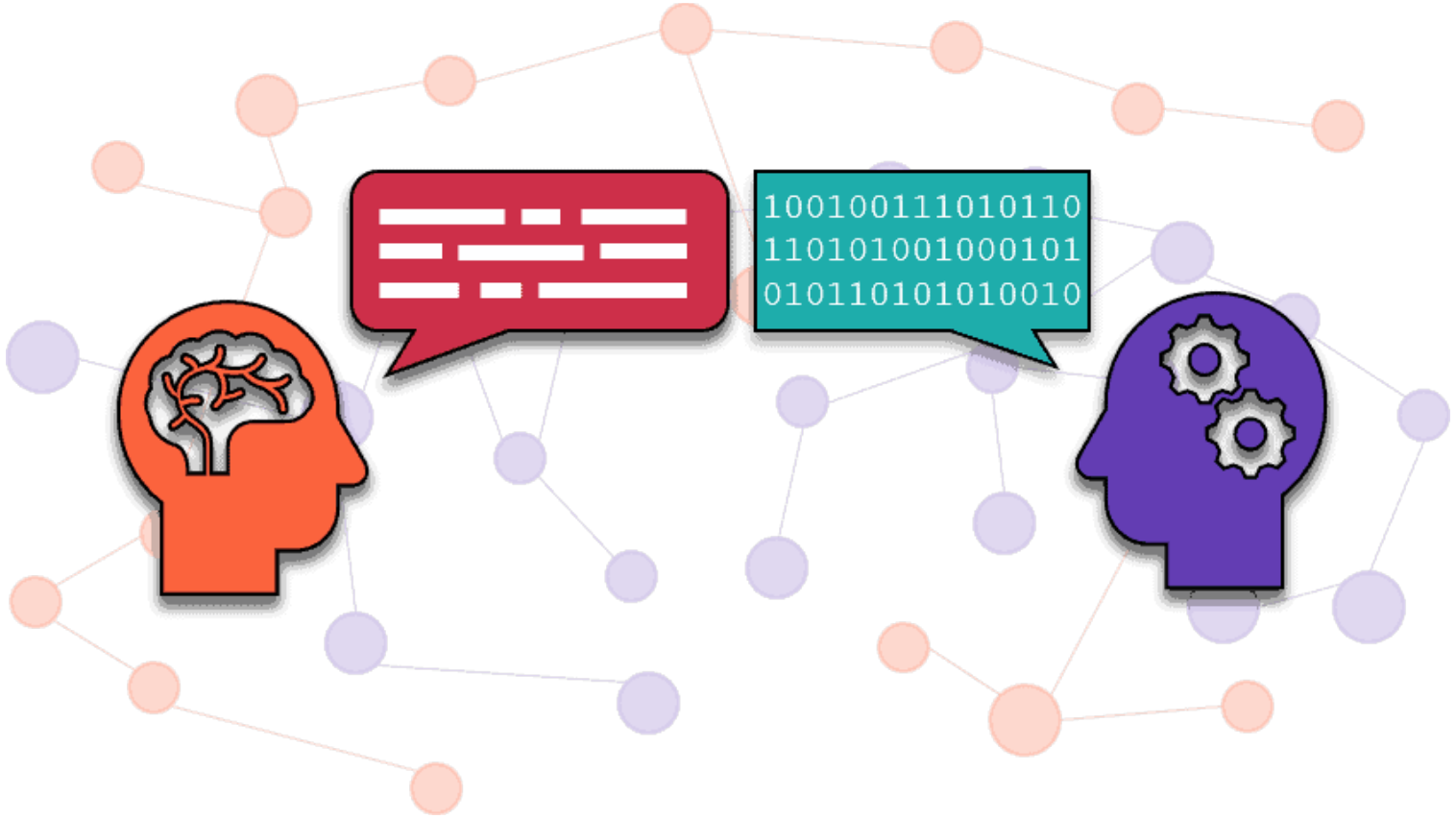
Answer: D

Q11. Please answer what analytics the following statement is about
"Experts forecast that the employment rate in 2022 will decrease."

A. Descriptive analytics
B. Diagnostic analytics
C. Predictive analytics
D. Prescriptive analytics

Answer: C

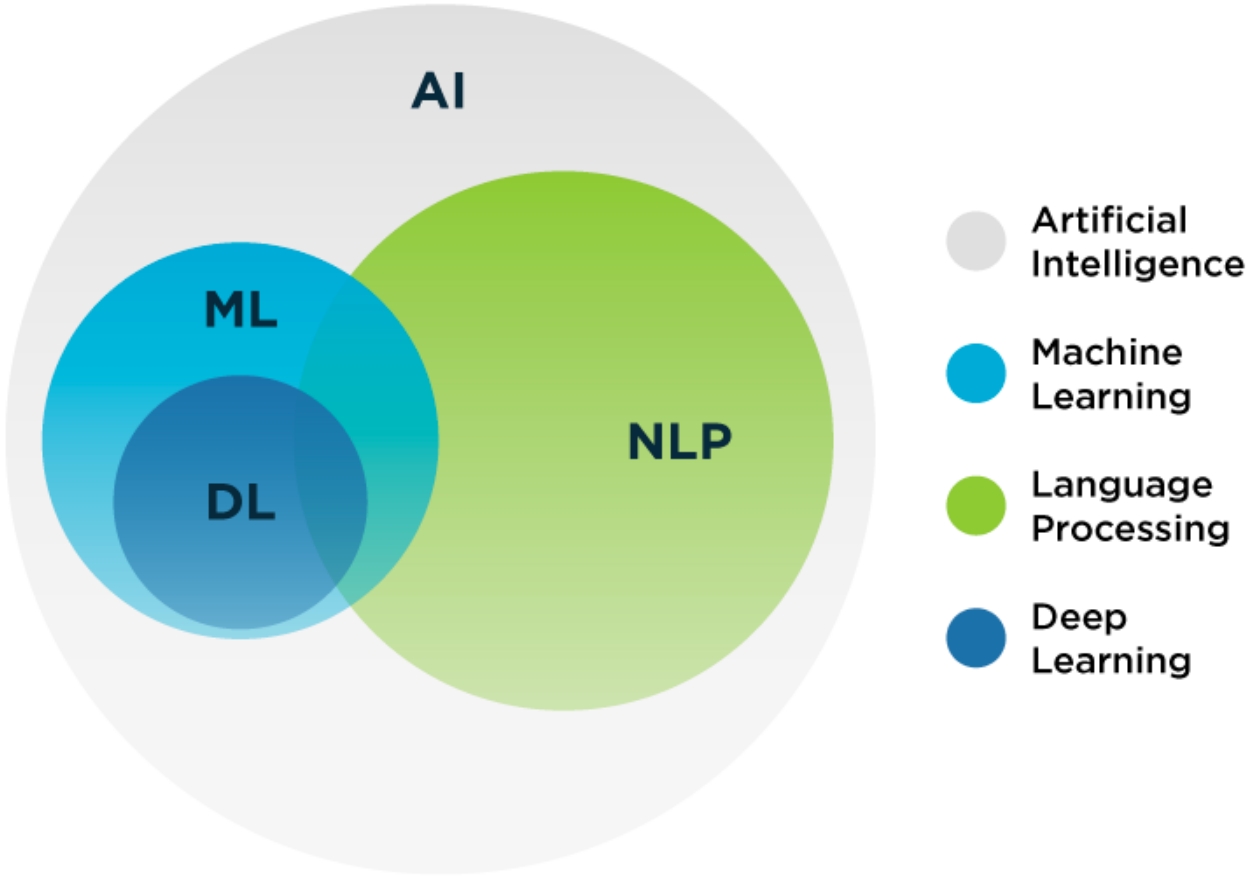
Group Project – Milestone 1



What is Natural Language Processing?

NLP Concepts

- **Natural Language Processing** (NLP) is a part of computer science and artificial intelligence which deals with human languages.



An Example of NLP (Natural Language Processing)

Arabic text كلب هو مطاردة صبي في الملعب.

How can a computer make **sense** out of this **string**?

Morphology	- What are the basic units of meaning (words)? - What is the meaning of each word?
Syntax	- How are words related with each other?
Semantics	- What is the "combined meaning" of words?
Pragmatics	- What is the "meta-meaning"? (speech act)
Discourse	- Handling a large chunk of text
Inference	- Making sense of everything

Reference: Dr. Hongning Wang, Text Mining Course. http://www.cs.virginia.edu/~hw5x/Course/TextMining-2019Spring/_site/

Why is NLP Challenging?

1. Ambiguity

- Natural language is designed to make human communication efficient. Therefore,
  - We omit a lot of “common sense” knowledge, which we assume the hearer/reader possesses
  - We keep a lot of ambiguities, which we assume the hearer/reader knows how to resolve
- This makes EVERY step in NLP hard
  - Ambiguity is a “killer”!
  - Common sense reasoning is pre-required

Reference: Dr. Hongning Wang, Text Mining Course. http://www.cs.virginia.edu/~hw5x/Course/TextMining-2019Spring/_site/

Various Types of Ambiguity

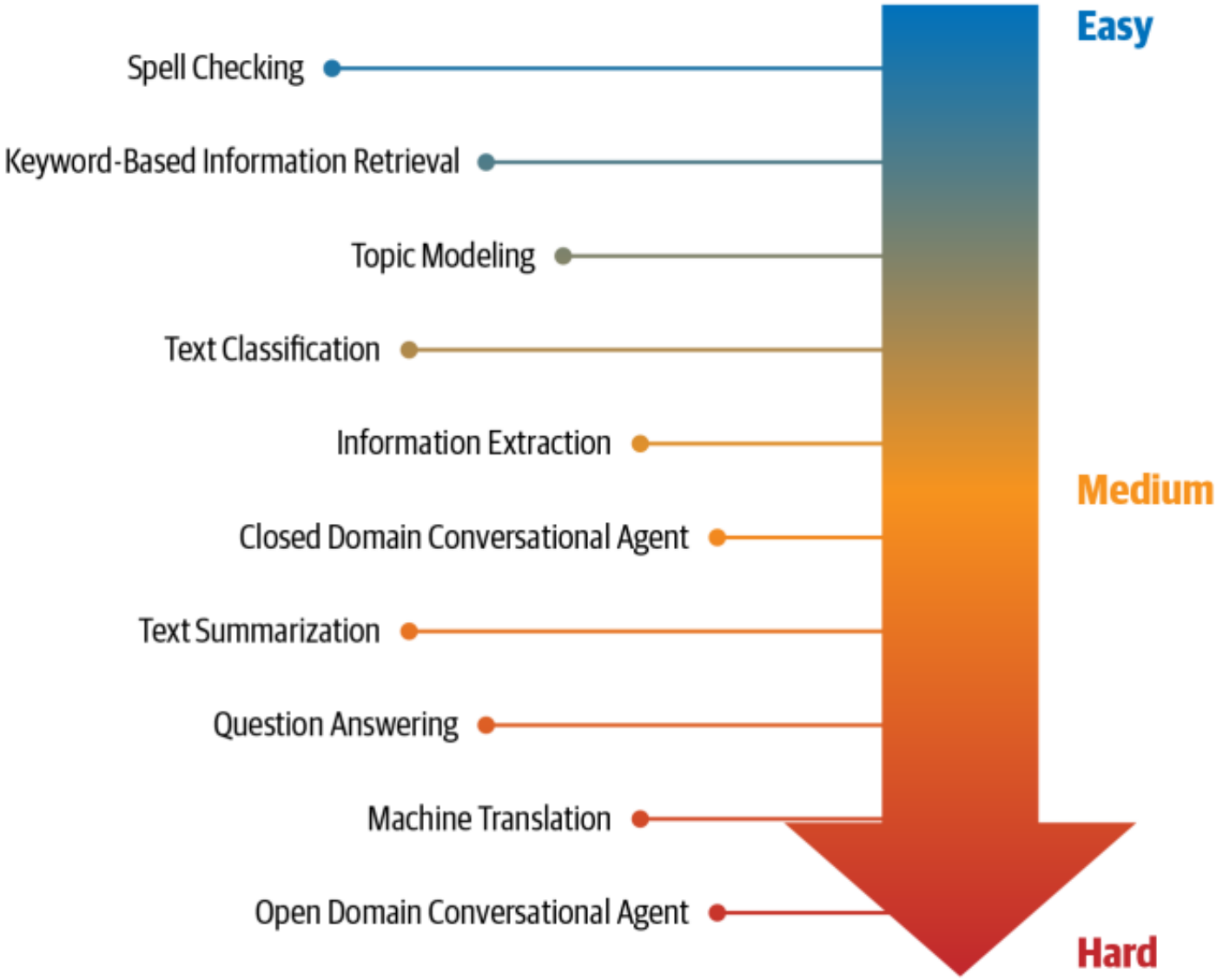
- Word-level ambiguity
  - “design” can be a noun or a verb (Ambiguous part of speech)
  - “root” has multiple meanings (Ambiguous sense)
- Syntactic ambiguity
  - “natural language processing” (Modification)
  - “A man saw a boy *with a telescope*.” (PP Attachment)
- Anaphora resolution
  - “John persuaded Bill to buy a TV for himself.” (himself = John or Bill?)
- Presupposition
  - “He has quit smoking.” implies that he smoked before.

Reference: Dr. Hongning Wang, Text Mining Course. http://www.cs.virginia.edu/~hw5x/Course/TextMining-2019Spring/_site/

2. Other Challenges

- Common knowledge
  - Common knowledge is a set of facts that most humans are aware of.
  - “Man bit dog” vs. “Dog bit man”
- Creativity
  - Language is not just rule-driven. It has creative part.
  - Various styles, dialects, genres, and variations are used in any language.
  - E.g., poems
- Diversity across languages
  - There is no direct mapping between the vocabularies of any two languages.
  - A solution that works for one language might not work at all for another language.

NLP Tasks


Spell Checking
Keyword-Based Information Retrieval
Topic Modeling
Text Classification
Information Extraction
Closed Domain Conversational Agent
Text Summarization
Question Answering
Machine Translation
Open Domain Conversational Agent
Easy
Medium
Hard

Approaches to NLP

- Heuristic-based NLP
  - Early attempts at building NLP systems were based on building rules for the task at hand.
  - Such systems normally require resources like dictionaries and thesauruses.
  - E.g., Lexicon-based sentiment analysis, Wordnet, regular expression
- Machine learning for NLP
  - Naïve baye, support vector machine, hidden Markov model, conditional random fields…
- Deep learning for NLP
  - Recurrent neural networks (RNN), long short-term memory (LSTM), convolutional neural networks (CNN), transformers…

NLP Pipeline

- The step-by-step processing of text is known as pipeline.
- It always involves going back and forth between individual steps.

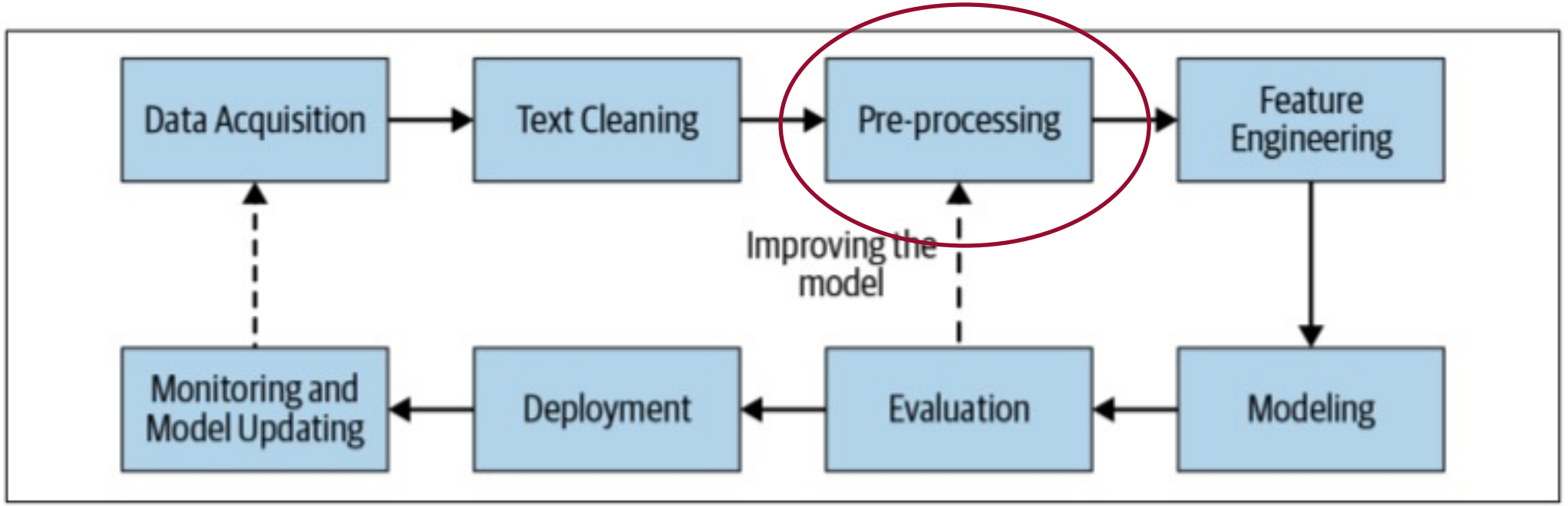


Figure 2-1. Generic NLP pipeline

Text Preprocessing

- A must-have step for text mining tasks.
- Transform text into a more digestible form so that machine learning algorithms can perform better.
- Note that not all text mining tasks need to go through every step in the pre-processing pipeline. It depends on the data and tasks.
- Also, the text preprocessing sequence might change in some real-world applications.

Is This a Positive or Negative Review? And Why?

Indie Coffee

4.7 ★★★★½ 735 reviews · $

Coffee shop

Indie Coffee is one of my favorite coffee shops/brunch spots in Madison. I first started coming here in college and I keep coming back because it's just so good. Great waffles, tasty breakfast sandwiches, and unique seasonal coffees as well!

Indie used to be my favorite coffee shop and I used to go almost every day without any issues. Recently I have been having issues every time. I get there maybe once a week now and the last two times it's taken about an hour to get my order, and both times one of the food items has been completely wrong. How it's happened where it's taken 4x longer than expected to get my food and getting the wrong item back to back is beyond me. The quality has dropped drastically in the last year. I have a hard time justifying my time and money to go onto campus for an hour just to get the wrong food. Really disappointed and would love for them to bring the quality back up.

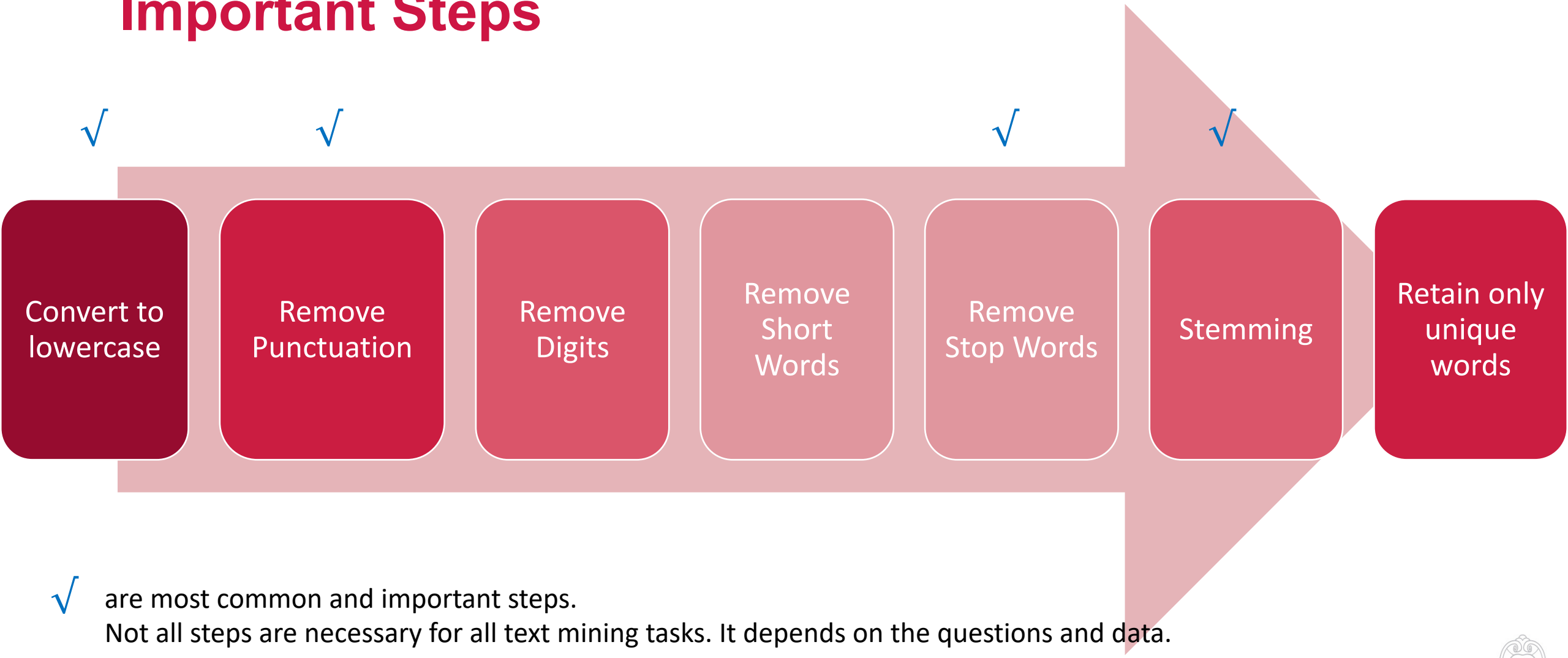
What are the key words? What information are unnecessary?

Pirates of The Caribbean is quite simply Hollywood's best pirate film in ages; a funny, rollicking swashbuckler that pays homage to the great films of the 1930's and 1940's featuring the likes of Errol Flynn, Charles Laughton, among others.

Text Preprocessing

- Preliminaries
  - Sentence segmentation (Split the text into sentences)
  - Word tokenization (split/tokenize a sentence into words)
- Frequent steps
  - Stop word removal, stemming and lemmatization, removing digits/punctuation, lowercasing, etc.
- Other steps
  - Normalization, language detection, code mixing, transliteration, etc.
- Advanced processing
  - POS (part of speech) tagging, parsing, coreference resolution, etc.

Let's Follow This Sequence to Discuss a Few Important Steps



√ are most common and important steps.
Not all steps are necessary for all text mining tasks. It depends on the questions and data.
Also, the text preprocessing sequence might change in some real-world applications.

Running Example

Note that this code is slightly different from the Colab Example code.

```
from nltk.corpus import stopwords
from stemming.porter2 import stem
import string
import re
# review from Amazon.com
text = """Pirates of The Caribbean is quite simply Hollywood's
best pirate film in ages; a funny, rollicking swashbuckler that
pays homage to the great films of the 1930's and 1940's
featuring the likes of Errol Flynn, Charles Laughton, among
others."""
```

1. Convert Text to Lowercase

- Python code:

```
text = text.lower()
print ("Lower Case:")
print (text)
```

- Output:

```
Lower Case:
pirates of the caribbean is quite simply hollywood's best
pirate film in ages; a funny, rollicking swashbuckler that
pays homage to the great films of the 1930's and 1940's
featuring the likes of errol flynn, charles laughton, among
others.
```

2. Remove Punctuation

- Python code:

```
regex = re.compile('[%s]' % re.escape(string.punctuation))
text = regex.sub('', text)
print ("No Punctuation:")
print (">>>> ", string.punctuation)
print (text)
```

- Output

```
No Punctuation:
>>>>  !"#$%&'()*+,-./:;<=>?@[\]^_`{|}~
pirates of the caribbean is quite simply hollywoods best pirate film in
ages a funny rollicking swashbuckler that pays homage to the great films
of the 1930s and 1940s featuring the likes of errol flynn charles
laughton among others
```

3. Remove Numbers

- Remove numbers if they are not relevant to your analyses. Usually, regular expressions are used to remove numbers.
- Python code:

```
regex = re.compile('[%s]' % re.escape(string.digits))
text = regex.sub('', text)
print ("No Digits:")
print (">>>> ", string.digits)
print (text)
```

- Output:

```
No Digits:
>>>>  0123456789
pirates of the caribbean is quite simply hollywoods best pirate film in ages
a funny rollicking swashbuckler that pays homage to the great films of the s
and s featuring the likes of errol flynn charles laughton among others
```

4. Remove Short Words

- Python code:

```
text = ' '.join([word for word in text.split() if (len(word)>=4)])
print ("Only Words >= 4 Chars Long:")
print (text)
```

- Output:

```
Only Words >= 4 Chars Long:
pirates caribbean quite simply hollywoods best pirate film ages
funny rollicking swashbuckler that pays homage great films
featuring likes errol flynn charles laughton among others
```

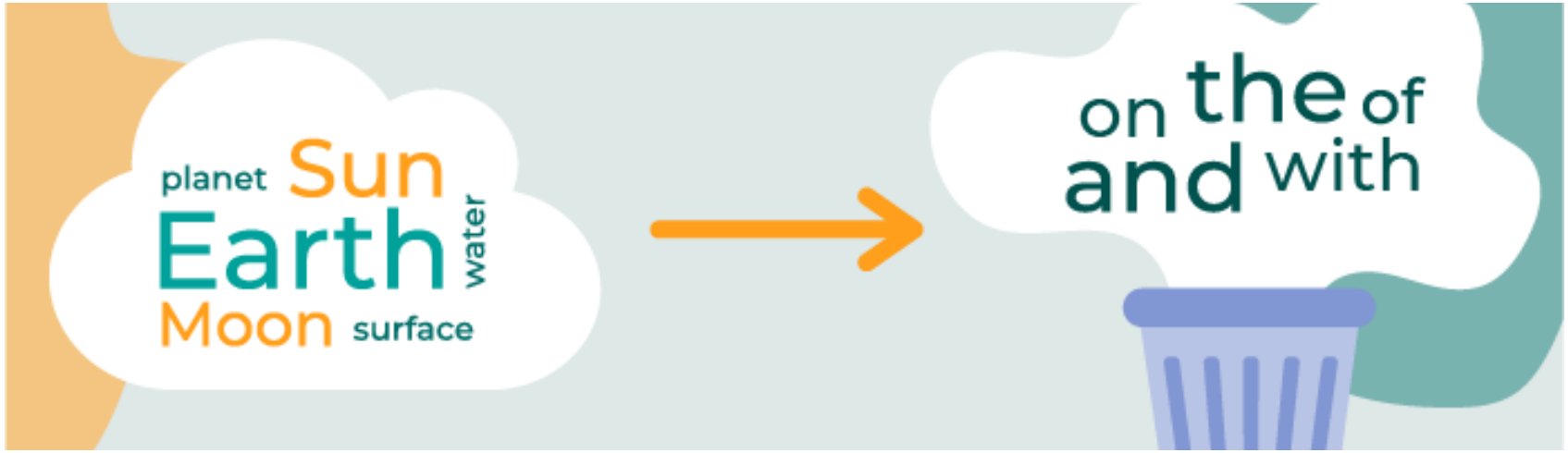
5. Remove Stop Words

- Python code:

```
sw = stopwords.words("english")
text = ' '.join([word for word in text.split() if word not in sw])
print ("No Stop Words:")
print (sw)
print (text)
```

Why do We Remove Stop Words?

- Stop words are relatively useless for document analysis
  - Why removing stop words? Not all words are informative. We remove these words to reduce vocabulary size.
  - There is no universal set of words called stop words
  - Risk: break the original meaning and structure of text
    - E.g., this is not a good option → after removing stop words: option
      to be or not to be → after removing stop words: null



Some Common Stop Words

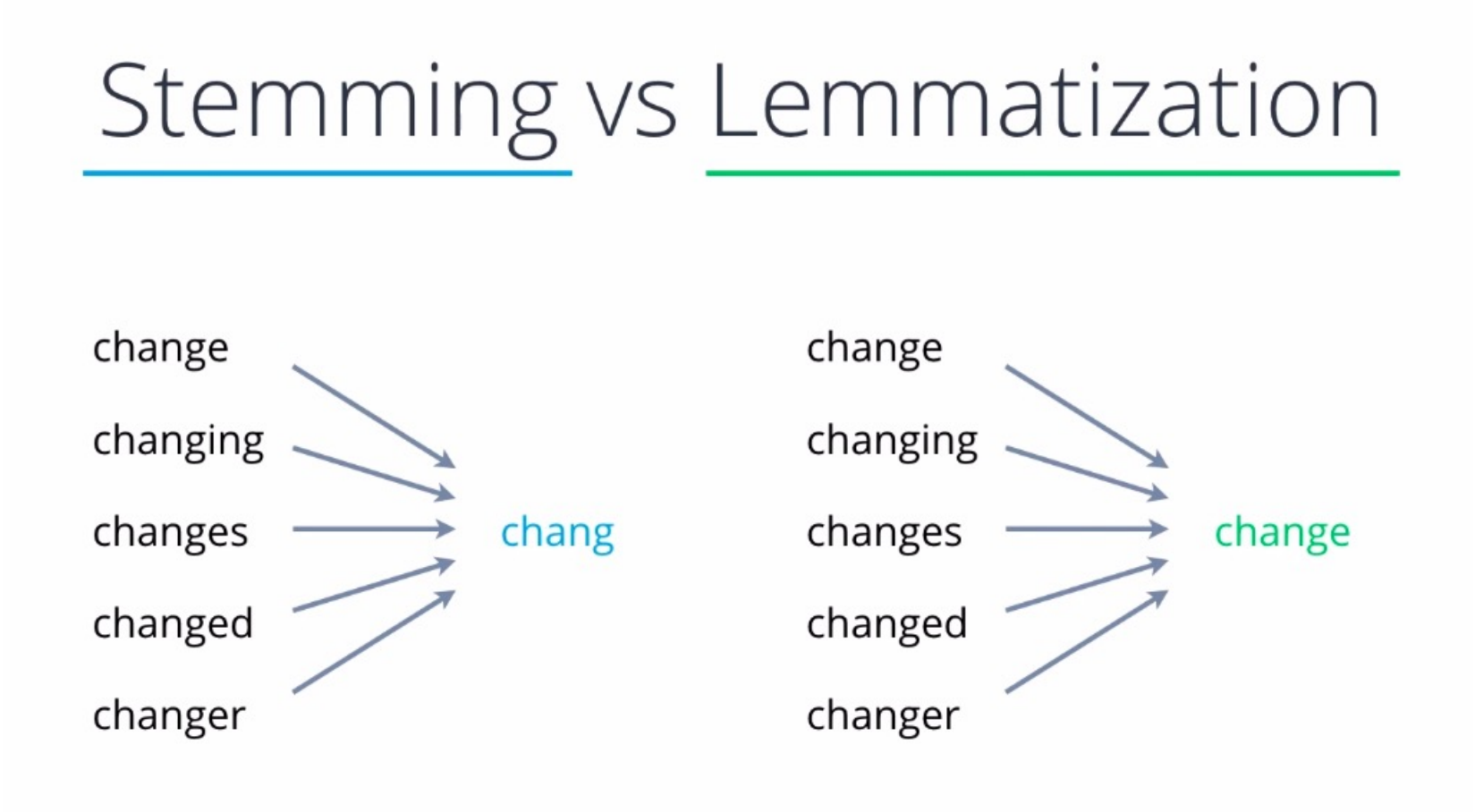
Nouns	Verbs	Adjectives	Prepositions	Others
1. time	1. be	1. good	1. to	1. the
2. person	2. have	2. new	2. of	2. and
3. year	3. do	3. first	3. in	3. a
4. way	4. say	4. last	4. for	4. that
5. day	5. get	5. long	5. on	5. I
6. thing	6. make	6. great	6. with	6. it
7. man	7. go	7. little	7. at	7. not
8. world	8. know	8. own	8. by	8. he
9. life	9. take	9. other	9. from	9. as
10. hand	10. see	10. old	10. up	10. you
11. part	11. come	11. right	11. about	11. this
12. child	12. think	12. big	12. into	12. but
13. eye	13. look	13. high	13. over	13. his
14. woman	14. want	14. different	14. after	14. they
15. place	15. give	15. small	15. beneath	15. her
16. work	16. use	16. large	16. under	16. she
17. week	17. find	17. next	17. above	17. or
18. case	18. tell	18. early		18. an
19. point	19. ask	19. young		19. will
20. government	20. work	20. important		20. my
21. company	21. seem	21. few		21. one
22. number	22. feel	22. public		22. all
23. group	23. try	23. bad		23. would
24. problem	24. leave	24. same		24. there
25. fact	25. call	25. able		25. their

The OEC: Facts about the language

6. Stemming & Lemmatization

- Stemming and Lemmatization are text normalization techniques within the field of NLP that are used to prepare text, words, and documents for further processing.
- **Stemming**: the process of producing morphological variants of a root/base word.
  - Stemming programs are commonly referred to as stemming algorithms or stemmers.
- **Lemmatization:** considers a language's full vocabulary to apply a morphological analysis to words.
  - The lemma of 'was' is 'be' and the lemma of 'mice' is 'mouse'.
  - Lemmatization is typically seen as much more informative than simple stemming

Example 1 - Stemming & Lemmatization



Example 2 - Stemming & Lemmatization

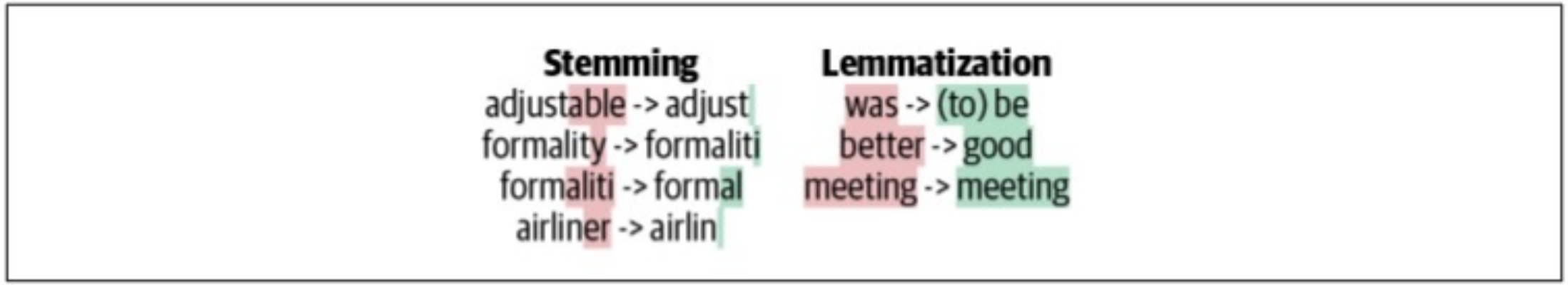


Figure 2-7. Difference between stemming and lemmatization [33]

Stemming Example

- Python code:

```
text = ' '.join([stem(word) for word in text.split()])
print ("Stem Words:")
print (text)
```

- Output:

```
Stem Words:
pirat caribbean quit simpli hollywood best pirat film age funni rollick
swashbuckl pay homag great film featur like errol flynn charl laughton
among other
```

7. Retain Unique Words

- Python code:

```
text = ' '.join(set(text.split()))
print ("Unique Words:")
print (text)
```

- Output:

```
Unique Words:
homag great errol rollick best featur charl laughton other
flynn among caribbean film swashbuckl pirat like pay quit
funni hollywood age simpli
```

Cautious

- Remember that not all of these preprocessing steps are always necessary, and not all of them are performed in the order in which they're discussed here.
- For example, if we were to remove digits and punctuation, what is removed first may not matter much. However, we typically lowercase the text before stemming.

Exercises using Google Colab