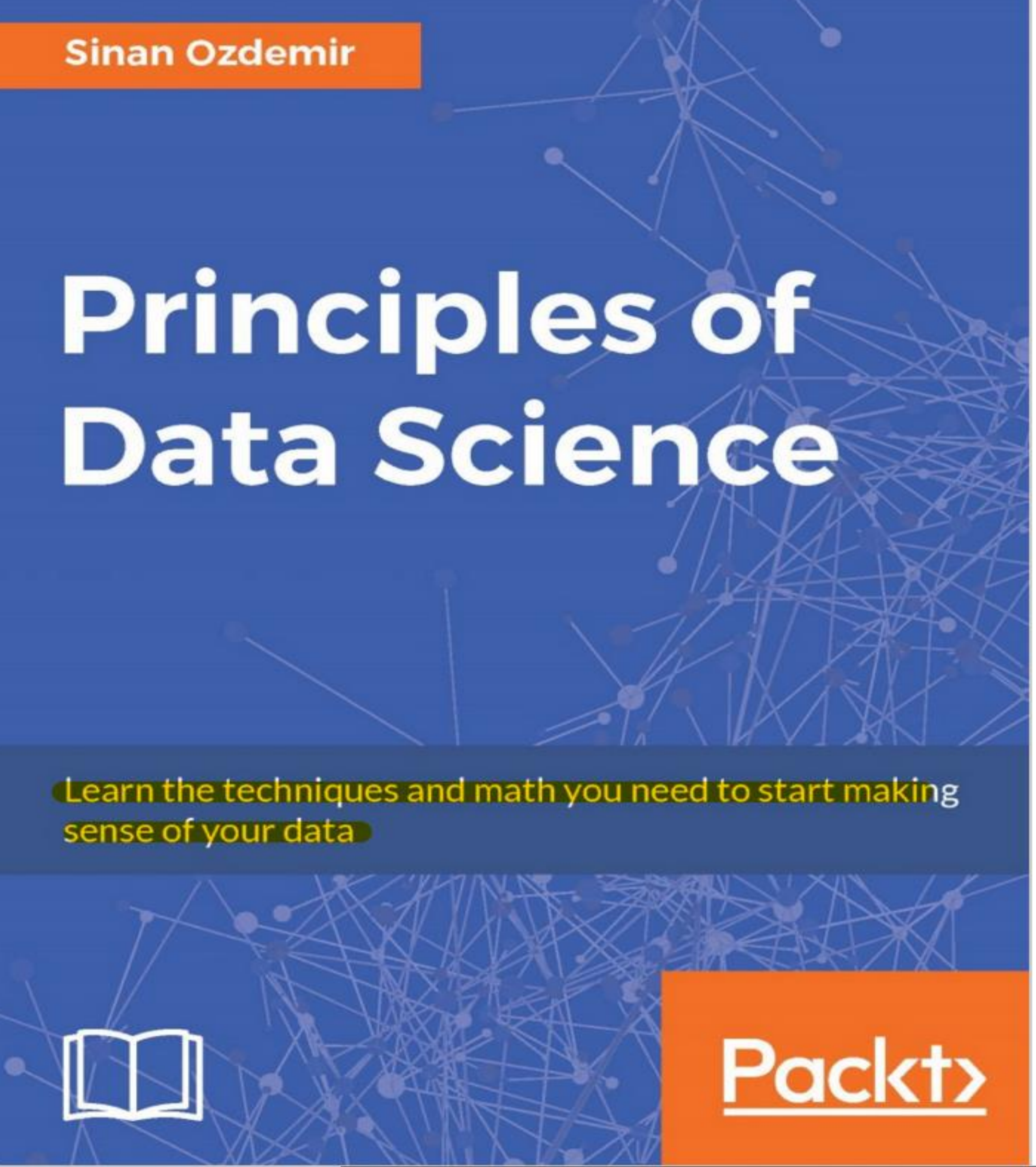
# PRINCIPLES OF DATA SCIENCE

2021 - 2022

Assoc. Prof. Dr. Salha Alzahrani

Department of Computer Science



Sinan Ozdemir

Principles of Data Science

Learn the techniques and math you need to start making sense of your data

Packt

# Principles of Data Science

Learn the techniques and math you need to start making sense of your data

**Sinan Ozdemir**

Packt

BIRMINGHAM - MUMBAI

TU

جامعة الطائف

TAIF UNIVERSITY

# CHAPTER 1: HOW TO SOUND LIKE A DATA SCIENTIST

- WHAT IS DATA SCIENCE?
- THE DATA SCIENCE VENN DIAGRAM
- PYTHON PRACTICES
- EXAMPLE OF BASIC PYTHON
- DOMAIN KNOWLEDGE
- SOME MORE TERMINOLOGY
- DATA SCIENCE CASE STUDIES

# HOW TO SOUND LIKE A DATA SCIENTIST

# Introduction

- No matter which industry you work in, IT, fashion, food, or finance, there is no doubt that data affects your life and work.
- At some point every day, you will either have or hear a conversation about data.
- News outlets are covering more and more stories about data leaks, cybercrimes, and how data can give us a glimpse into our lives.

- But why now? What makes this era such a hotbed for data-related industries?

# Introduction

- In the 19th century, the world was in the grip of the industrial age.
- By the 20th century, we were quite skilled at making huge machines; the goal now was to make them smaller and faster. The industrial age was over and was replaced by what we refer to as the information age.
- This left us with a problem in the 21st century, so much data; what we refer to as the data age.
  - About 1.8 trillion gigabytes of data in 2011 (take a moment to just think about how much that is).
  - Just one year later, in 2012, we created over 2.8 trillion gigabytes of data!
  - This number is only going to explode further to hit an estimated 40 trillion gigabytes of data creation in just one year by 2020.

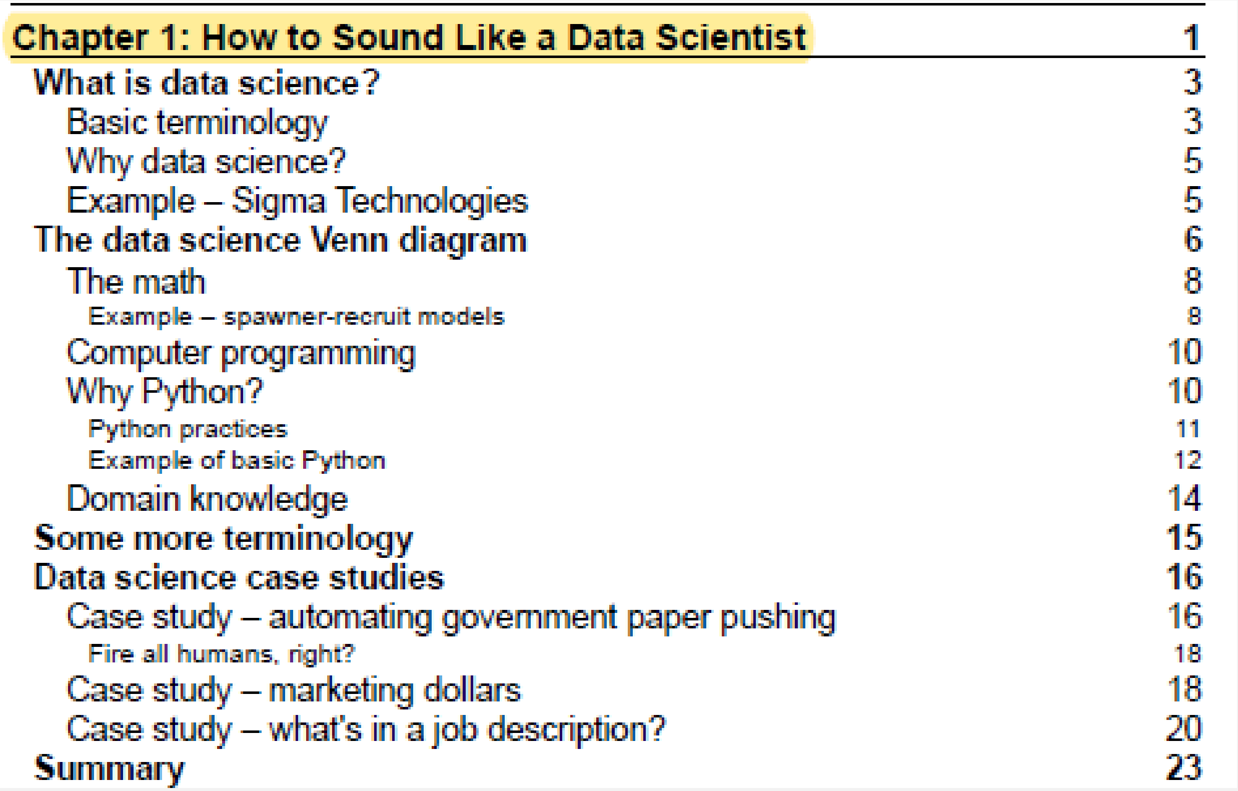
# Introduction

- People contribute to this every time they tweet, post on Facebook, save a new resume on Microsoft Word, or just send their mom a picture through smart phones.

- Not only we are creating data at an unprecedented rate, but we are also consuming it at an accelerated pace as well. In 2013, the average cell phone user used under 1 GB of data a month. Today, that number is estimated to be well over 2 GB a month.

# The book

- Chapter 1: will explore the terminology and vocabulary of the modern data scientist.
  - Basic terminology of data science
  - The three domains of data science
  - The basic Python syntax



**Chapter 1: How to Sound Like a Data Scientist** 1

**What is data science?** 3

Basic terminology 3

Why data science? 5

Example – Sigma Technologies 5

**The data science Venn diagram** 6

The math 8

Example – spawner-recruit models 8

Computer programming 10

Why Python? 10

Python practices 11

Example of basic Python 12

Domain knowledge 14

**Some more terminology** 15

**Data science case studies** 16

Case study – automating government paper pushing 16

Fire all humans, right? 18

Case study – marketing dollars 18

Case study – what's in a job description? 20

**Summary** 23

# WHAT IS DATA SCIENCE?

# What is data?

- Whenever we use the word "data", we refer to a collection of information in either an organized or unorganized format:

  - Organized data: This refers to data that is sorted into a row/column structure, where every row represents a single observation, and the columns represent the characteristics of that observation.
  - Unorganized data: This is the type of data that is in the free form, usually texts, images, videos, or raw audio/signals that must be parsed further to become organized.

# What is data science?

- Data science is the art and science of acquiring knowledge through data.
- What a small definition for such a big topic, and rightfully so! Data science covers so many things that it would take pages to list it all out.

- Data science is all about how we take data, use it to acquire knowledge, and then use that knowledge to do the following:
  - Make decisions
  - Predict the future
  - Understand the past/present
  - Create new industries/products

# Why data science?

- In this data age, it's clear that we have a surplus of data. But why should that necessitate an entire new set of vocabulary? What was wrong with our previous forms of analysis?
  - For one, the sheer volume of data makes it literally impossible for a human to parse it in a reasonable time.
  - Data is collected in various forms and from different sources, and often comes in very unorganized forms.
  - Data can be missing, incomplete, or just flat out wrong.
  - Often, we have data on very different scales and that makes it tough to compare.
  - One of the main goals of data science is to make explicit practices and procedures to discover and apply these relationships in the data.

# Example – Sigma Technologies

- Ben Runkle, CEO, Sigma Technologies, is trying to resolve a huge problem. The company is consistently losing long-time customers. He does not know why they are leaving, but he must do something fast. He is convinced that in order to reduce his churn, he must create new products and features, and consolidate existing technologies. To be safe, he calls in his chief data scientist, Dr. Jessie Hughan.

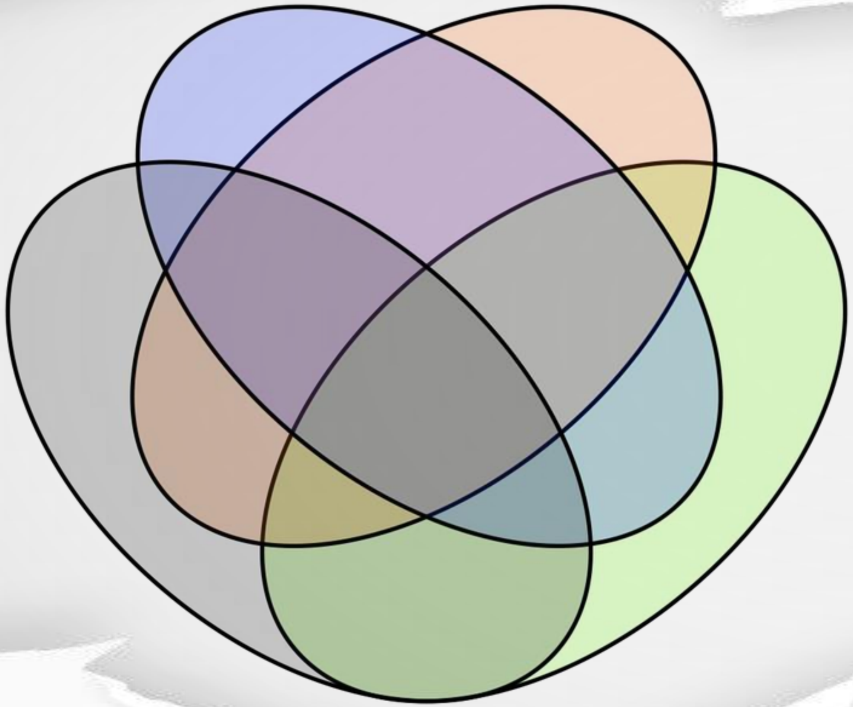
- However, she is not convinced that new products and features alone will save the company. Instead, she turns to the transcripts of recent customer service tickets. She shows Runkle the most recent transcripts and finds something surprising:
  - ".... Not sure how to export this; are you?"
  - "Where is the button that makes a new list?"
  - "Wait, do you even know where the slider is?"
  - "If I can't figure this out today, it's a real problem..."

# Example – Sigma Technologies

- It is clear that customers were having problems with the existing UI/UX, and upset due to a lack of features. Runkle and Hughan organized a mass UI/UX overhaul and their sales have never been better. Of course, the science used in the last example was minimal, but it makes a point.

- We tend to call people like Runkle, a driver. Today, CEO wants to make all decisions quickly and iterate over solutions until something works. Dr. Hughan is much more analytical. She wants to solve the problem just as much as Runkle, but she turns to user-generated data instead of her gut feeling for answers.

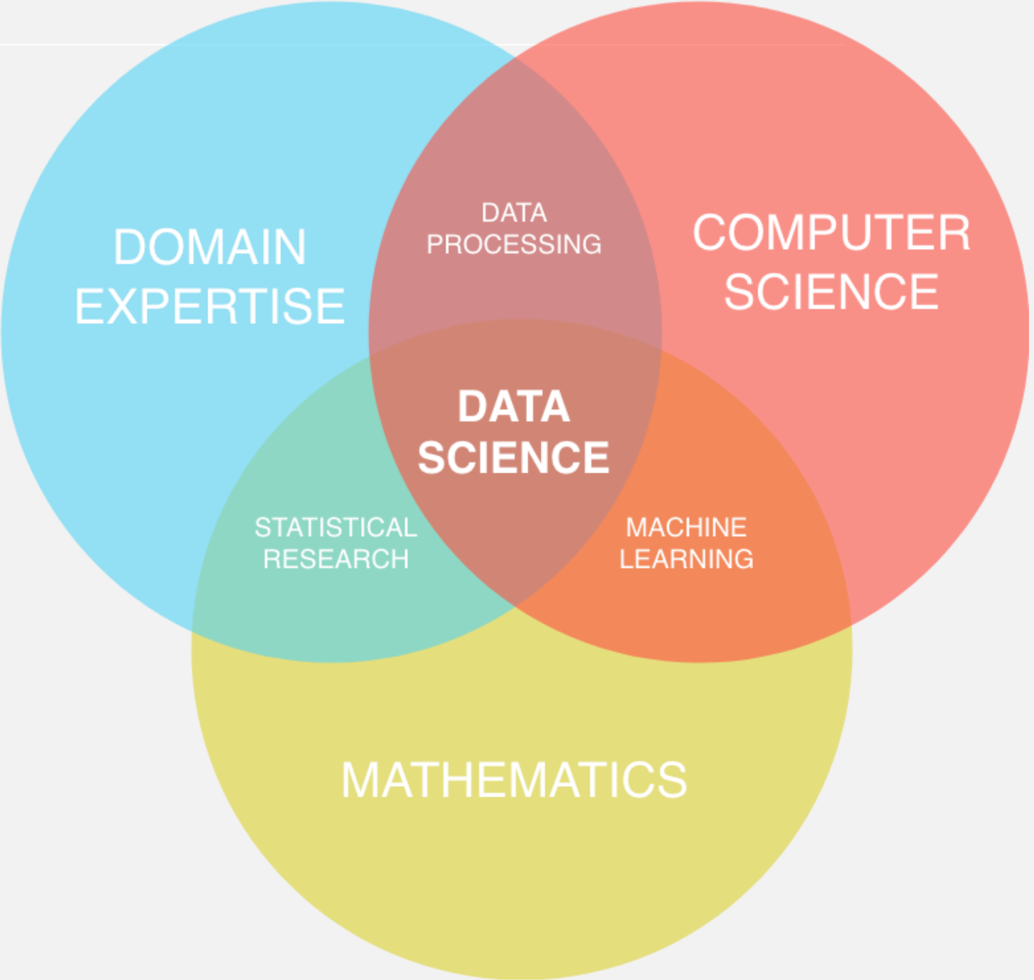
- Data science is about applying the skills of the analytical mind and using them as a driver would. Both of these mentalities have their place in today's enterprises; however, it is Hughan's way of thinking that dominates the ideas of data science—using data generated by the company as her source of information rather than just picking up a solution and going with it.



# THE DATA SCIENCE VENN DIAGRAM

# The data science Venn diagram

- Understanding data science begins with three basic areas:
  - Math/statistics: This is the use of equations and formulas to perform analysis
  - Computer programming: This is the ability to use code to create outcomes on the computer
  - Domain knowledge: This refers to understanding the problem domain (medicine, finance, social science, and so on)
- The following Venn diagram provides a visual representation of how the three areas of data science intersect:



# The Math

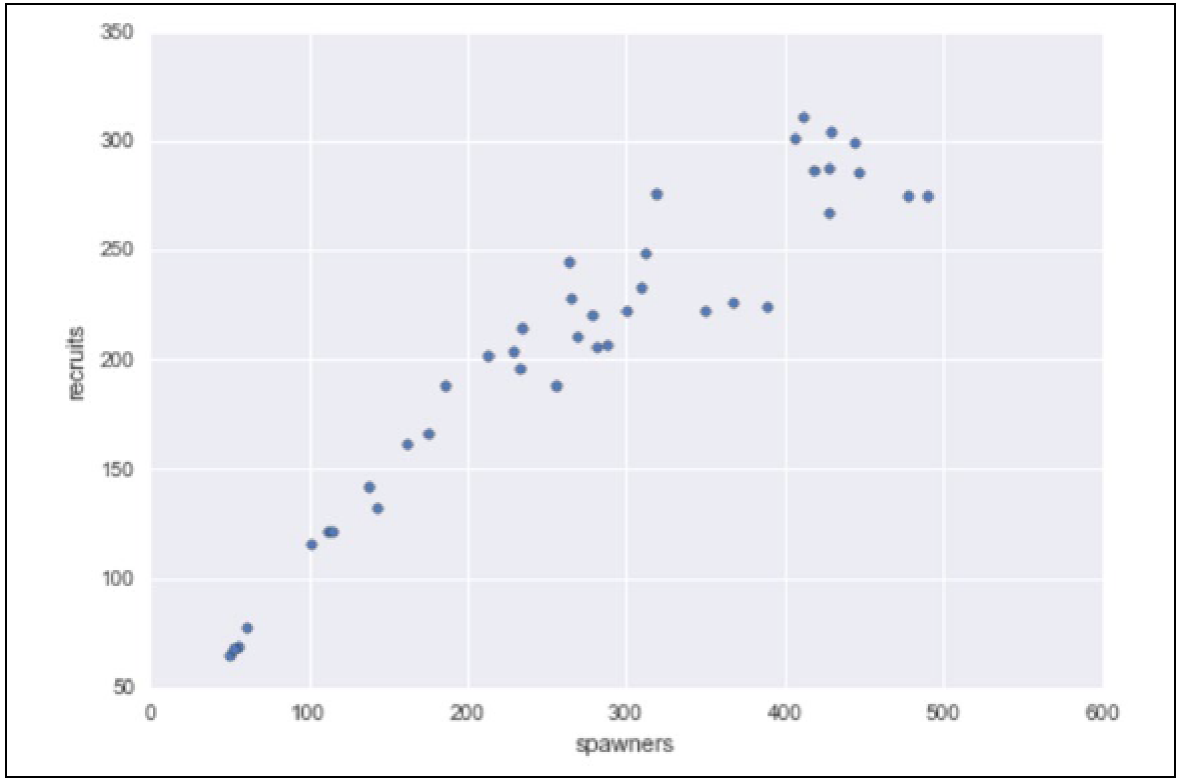
# The math

- The course will guide you through the math needed for data science, specifically statistics and probability. We will use these subdomains of mathematics to create what are called models.
- A data model refers to an organized and formal relationship between elements of data, usually meant to simulate a real-world phenomenon.
- Essentially, we will use math in order to formalize relationships between variables.
- Between the three areas of data science, math is what allows us to move from domain to domain.
- Understanding the theory allows us to apply a model that we built for the fashion industry to a financial model.

# Example – spawner-recruit models

- In biology, we use, a model known as the spawner-recruit model to judge the biological health of a species. It is a basic relationship between the number of healthy parental units of a species and the number of new units in the group of animals.

- In a public dataset of the number of salmon spawners and recruits, the graph was formed to visualize the relationship between the two. We can see that there definitely is some sort of positive relationship (as one goes up, so does the other).



The spawner-recruit model visualized

# Example – spawner-recruit models

- But how can we formalize this relationship? For example, if we knew the number of spawners in a population, could we predict the number of recruits that group would obtain, and vice versa?
- Essentially, models allow us to plug in one variable to get the other. Consider the following example:

$$Recruits = 0.5 * Spawners + 60$$

- In this example, let's say we knew that a group of salmons had 1.15 (in thousands) of spawners. Then, we would have the following:

$$Recruits = 0.5 * 1.15 + 60$$

$$Recruits = 60.575\,(in\,thousands)$$

- This result can be very beneficial to estimate how the health of a population is changing.

If we can create these models, we can visually observe how the relationship between the two variables can change.

# python™ Programming

# Computer programming

- Computer languages are how we communicate with the machine and tell it to do our bidding.
- A computer speaks many languages and, like a book, can be written in many languages; similarly, data science can also be done in many languages.
- Python, Julia, and R are some of the many languages available to us. This book will focus exclusively on using Python.

# Why Python?

We will use Python for a variety of reasons:

- Python is an extremely simple language to read and write, even if you've never coded before, which will make future examples easy to ingest and read.
- It is one of the most common languages, both in production and in the academic setting (one of the fastest growing, as a matter of fact)
- The language's online community is vast and friendly. This means that a quick Google search should yield multiple results of people who have faced and solved similar (if not exactly the same) situations.
- Python has prebuilt data science modules that both the novice and the veteran data scientist can utilize

# Why Python?

- The last is probably the biggest reason we will focus on Python. These prebuilt modules are not only powerful, but also easy to pick up. By the end of the first few chapters, you will be very comfortable with these modules. Some of these modules are as follows:
  - pandas
  - sci-kit learn
  - seaborn
  - numpy/scipy
  - requests (to mine data from the Web)
  - BeautifulSoup (for the Web-HTML parsing)

# Python practices

In Python, we have variables that are placeholders for objects.

- int (an integer) : Examples: 3, 6, 99, -34, 34, 11111111
- float (a decimal) : Examples: 3.14159, 2.71, -0.34567
- boolean (either True or False) :
  - The statement, Sunday is a weekend, is True
  - The statement, Friday is a weekend, is False
  - The statement, pi is exactly the ratio of a circle's circumference to its diameter, is True (crazy, right?)
- string (text or words made up of characters)
  - "I love hamburgers" (by the way, who doesn't?)
  - "Matt is awesome"
  - A Tweet is a string
- list (a collection of objects) : Example: [1, 5.4, True, "apple"]

# Example of basic Python

- In Python, we use spaces/tabs to denote operations that belong to other lines of code.
- Note that the following list variable, my_list, can hold multiple types of objects. This one has an int, a float, boolean, and string inputs (in that order):
  - my_list = [1, 5.7, True, "apples"]
  - len(my_list) == 4    # 4 objects in the list
  - my_list[0] == 1    # the first object
  - my_list[1] == 5.7    # the second object

# Terminology

# Domain knowledge

- Domain knowledge is about the topic or area you are working on. For example, if you are a financial analyst working on stock market data, you have a lot of domain knowledge. If you are a journalist looking at worldwide adoption rates, you might benefit from consulting an expert in the field.

- Does that mean that if you're not a doctor, you can't work with medical data? Of course not! Great data scientists can apply their skills to any area, even if they aren’t fluent in it. Data scientists can adapt to the field and contribute meaningfully when their analysis is complete.

- A big part of domain knowledge is presentation. Depending on your audience, it can greatly matter how you present your findings. Your results are only as good as your vehicle of communication. You can predict the movement of the market with 99.99% accuracy, but if your program is impossible to execute, your results will go unused. Likewise, if your vehicle is inappropriate for the field, your results will go equally unused.

# Some more terminology

- Machine learning: refers to giving computers the ability to learn from data without explicit "rules" being given by a programmer. We have seen the concept of machine learning earlier in this chapter as the union of someone who has both coding and math skills. Here, we are attempting to formalize this definition. Machine learning combines the power of computers with intelligent learning algorithms in order to automate the discovery of relationships in data and create of powerful data models.

- Speaking of data models, we will concern ourselves with the following two basic types of data models:
  - Probabilistic model: refers to using probability to find a relationship between elements that includes a degree of randomness.
  - Statistical model: refers to taking advantage of statistical theorems to formalize relationships between data elements in a (usually) simple mathematical formula.

# Some more terminology

- Exploratory data analysis (EDA) refers to preparing data in order to standardize results and gain quick insights. EDA is concerned with data visualization and preparation. This is where we turn unorganized data into organized data and also clean up missing/incorrect data points. During EDA, we will create many types of plots and use these plots to identify key features and relationships to exploit in our data models.

- Data mining is the process of finding relationships between elements of data. Data mining is the part of data science where we try to find relationships between variables (think spawn-recruit model).

# Please Read The Case Studies

Page 16-22

s.zahrani@tu.edu.sa