Data Science

General Assembly Lecture 1

Instructor: Hamed Hasheminia

- Instructional Team:
    - Hamed Hasheminia
    - Karla Leibowitz
    - Joshua Cano

A little about me

- Projects with substantial data science components that I have been involved with:
  - Insurance Corporation of British Columbia
  - WestJet
  - Transport of Canada
  - Logico Carbon Solutions
  - Prince Rupert Authorities

ICBC

WESTJET

Transport Canada Transports Canada

Canada

LOGICO CARBON SOLUTIONS

PRINCE RUPERT PORT AUTHORITY
LINKING A WORLD OF OPPORTUNITY

SAN FRANCISCO STATE UNIVERSITY

UBC

A little bit about yourself

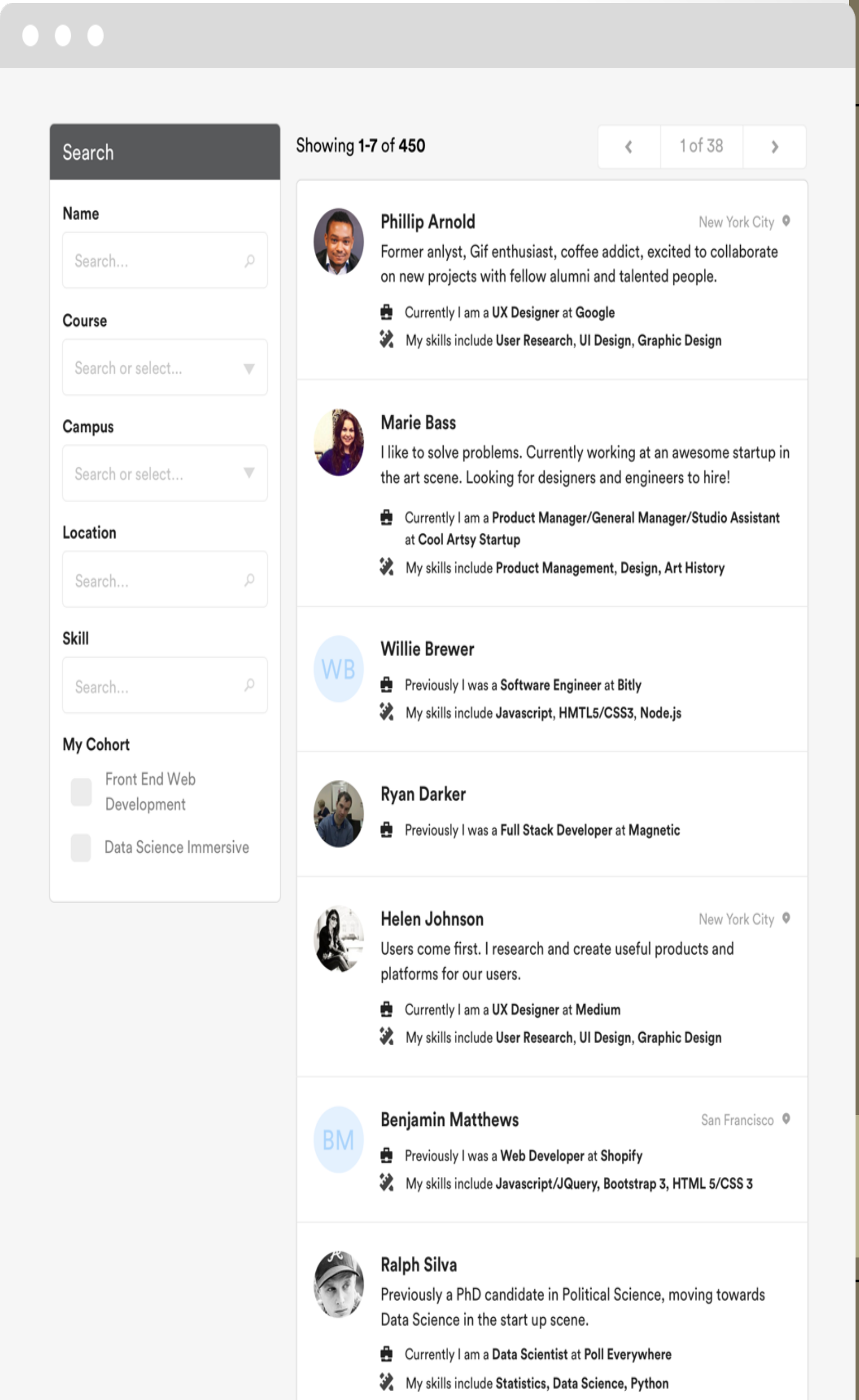
- Your Name
- Summary of your background
- What are you hoping/expecting to get out of this class
- Some interesting factoid about yourself

GA Directory

The GA Directory is a place for students, alumni and instructors to connect.

- Find your classmates
- Reach out to alumni and instructors
- Hire talent based on skills and experience

directory.generalassemb.ly



Showing **1-7** of **450**

Phillip Arnold New York City

Former anlyst, Gif enthusiast, coffee addict, excited to collaborate on new projects with fellow alumni and talented people.

Currently I am a **UX Designer** at **Google**

My skills include **User Research, UI Design, Graphic Design**

Marie Bass

I like to solve problems. Currently working at an awesome startup in the art scene. Looking for designers and engineers to hire!

Currently I am a **Product Manager/General Manager/Studio Assistant** at **Cool Artsy Startup**

My skills include **Product Management, Design, Art History**

Willie Brewer

Previously I was a **Software Engineer** at **Bitly**

My skills include **Javascript, HMTL5/CSS3, Node.js**

Ryan Darker

Previously I was a **Full Stack Developer** at **Magnetic**

Helen Johnson New York City

Users come first. I research and create useful products and platforms for our users.

Currently I am a **UX Designer** at **Medium**

My skills include **User Research, UI Design, Graphic Design**

Benjamin Matthews San Francisco

Previously I was a **Web Developer** at **Shopify**

My skills include **Javascript/JQuery, Bootstrap 3, HTML 5/CSS 3**

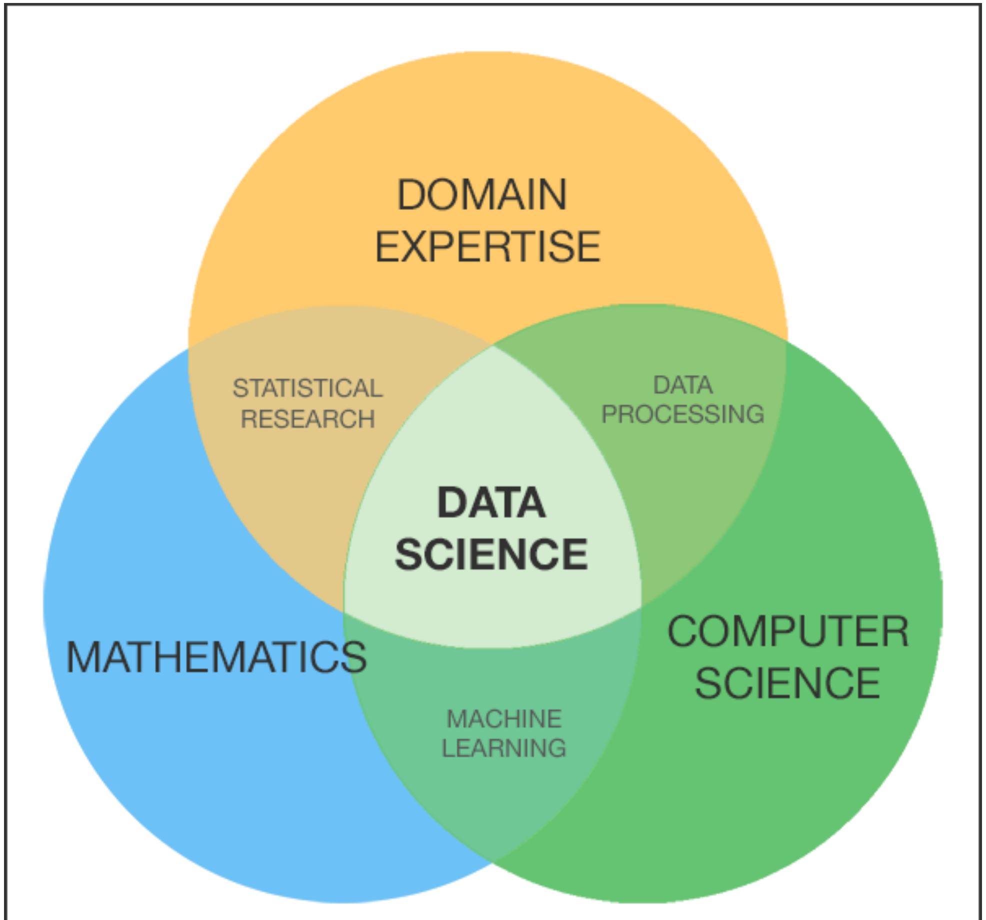
Ralph Silva

Previously a PhD candidate in Political Science, moving towards Data Science in the start up scene.

Currently I am a **Data Scientist** at **Poll Everywhere**

My skills include **Statistics, Data Science, Python**

What is Data Science?



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Instructor Philosophy

- Adjustable pace
- Interactive Class
- Learn how to learn
- Learn by teaching
- Slow and constant progress

Content Philosophy

- Learn by doing
  - In-Class examples
  - Assignments
  - Course Project
  - Practice, Practice, Practice
- Extensive use of visuals
- Balance of depth with breadth

How to succeed

- You cannot drive by just sitting on the passenger seat
- Ask questions
- Answer questions
- Teach to your classmates
  - You only learn something when you can teach it to somebody else.
- Practice, Practice, Practice

Typical Class

- Review of previous materials
- Theoretical lecture – from PowerPoint slides
- Lab / Code Walk-through
- In-class exercises
- Homework assigned

GA GRADUATION REQUIREMENTS

1. Homework (80% of Homework and Labs)
2. Attendance (Miss no more than 2 classes)
3. Final Project

    a) Come up with 3 research ideas – due 3rd session

    b) Create an outline for your project. What you want to test, what are your goals – due 7th session

    c) Clean your data and do exploratory analysis (Due 12th session)

    d) Detailed iPython technical notebook with your models (18th session)

    e) Present your work (20 minutes presentations) (19th and 20th sessions)

Topics to be covered in the class

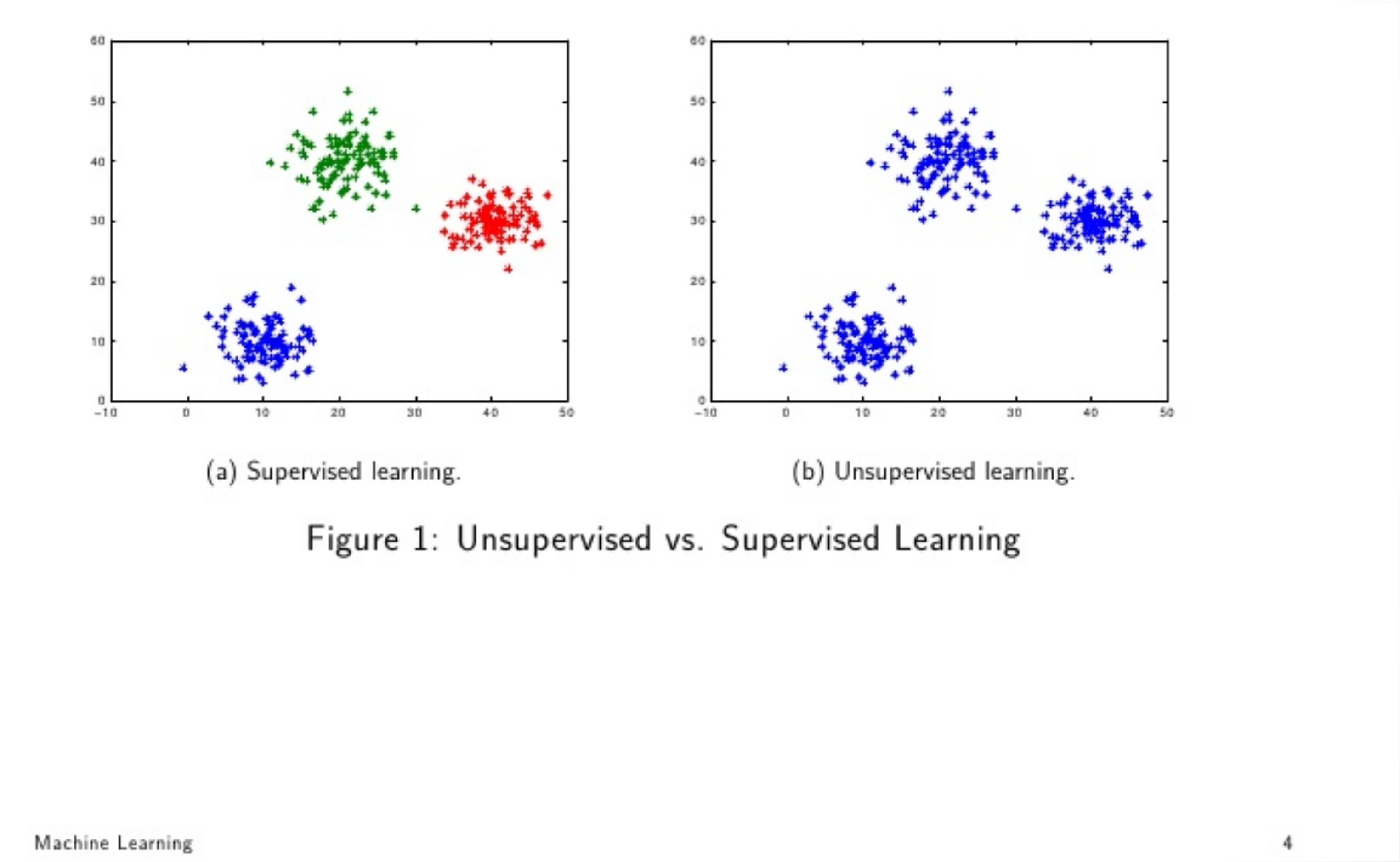
- Introduction/Basic Stat (2 sessions)
- Linear Models (2 sessions)
- Model Selection and Regularization (1 session)
- Missing Data and Imputation (1 session)
- K-Nearest Neighbors (1 session)
- Logistic Regression (2 sessions)
- In class Team Project (1 session)
- Decision Tree CART (1 session)
- Bagging, Boosting, Random Forest (1 session)
- Natural Language Processing (1 session)
- Principal Component Analysis (1 session)
- Time series Models (1 session)
- Naïve Bayes (1 session)
- Flex Session (to be decided)
- Ensemble models and summary (1 session)
- Project presentations (2 sessions)

Let’s get Started!

Agenda

- Supervised – Unsupervised learning
- Classification vs Regression
- Flexibility vs Interpretability
- Time series vs cross-sectional analysis
- Parsing data / data dictionary
- Intro to Numpy and Pandas

Supervised vs Unsupervised Learning



(a) Supervised learning.

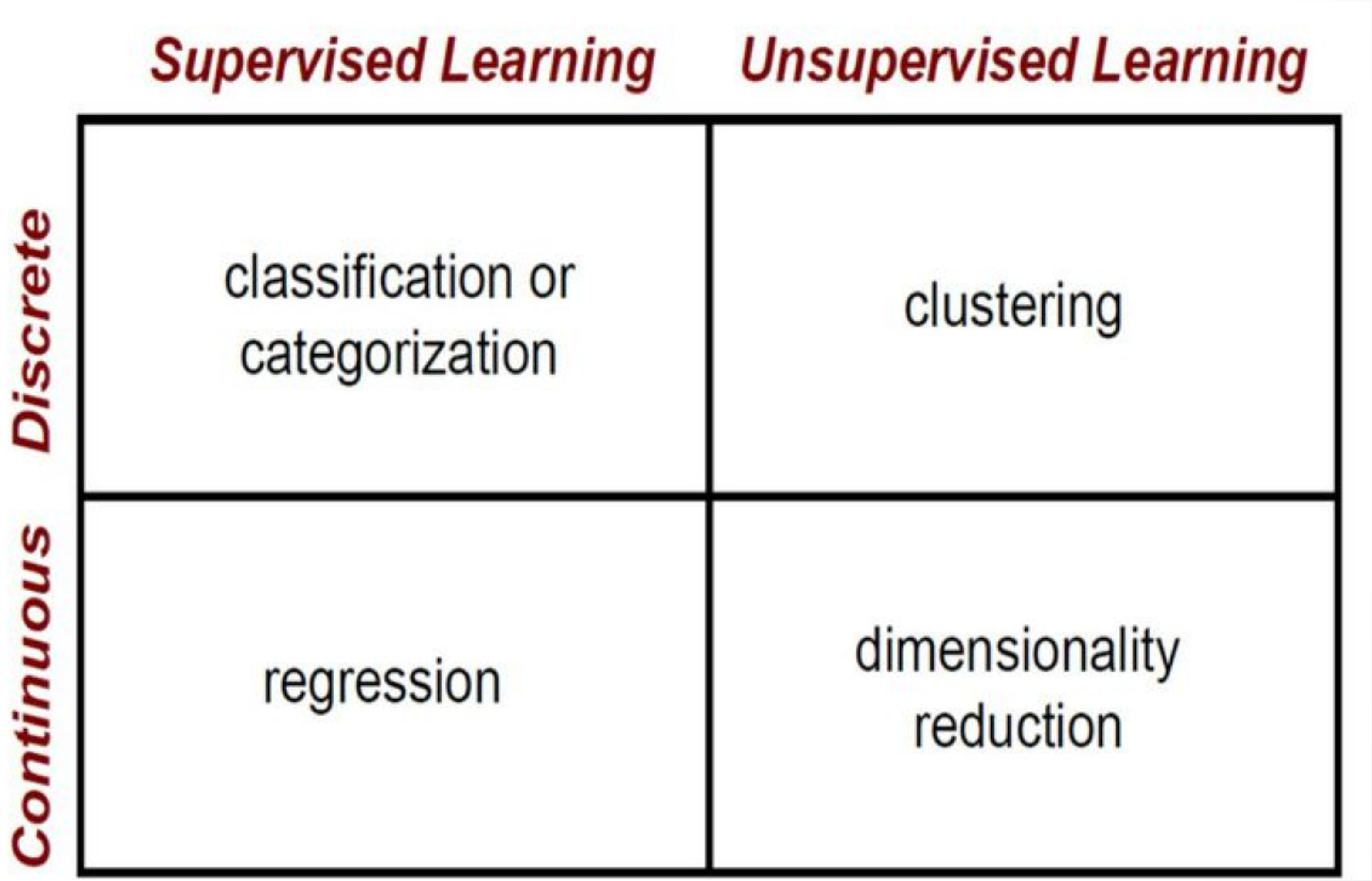
(b) Unsupervised learning.

Figure 1: Unsupervised vs. Supervised Learning

Machine Learning

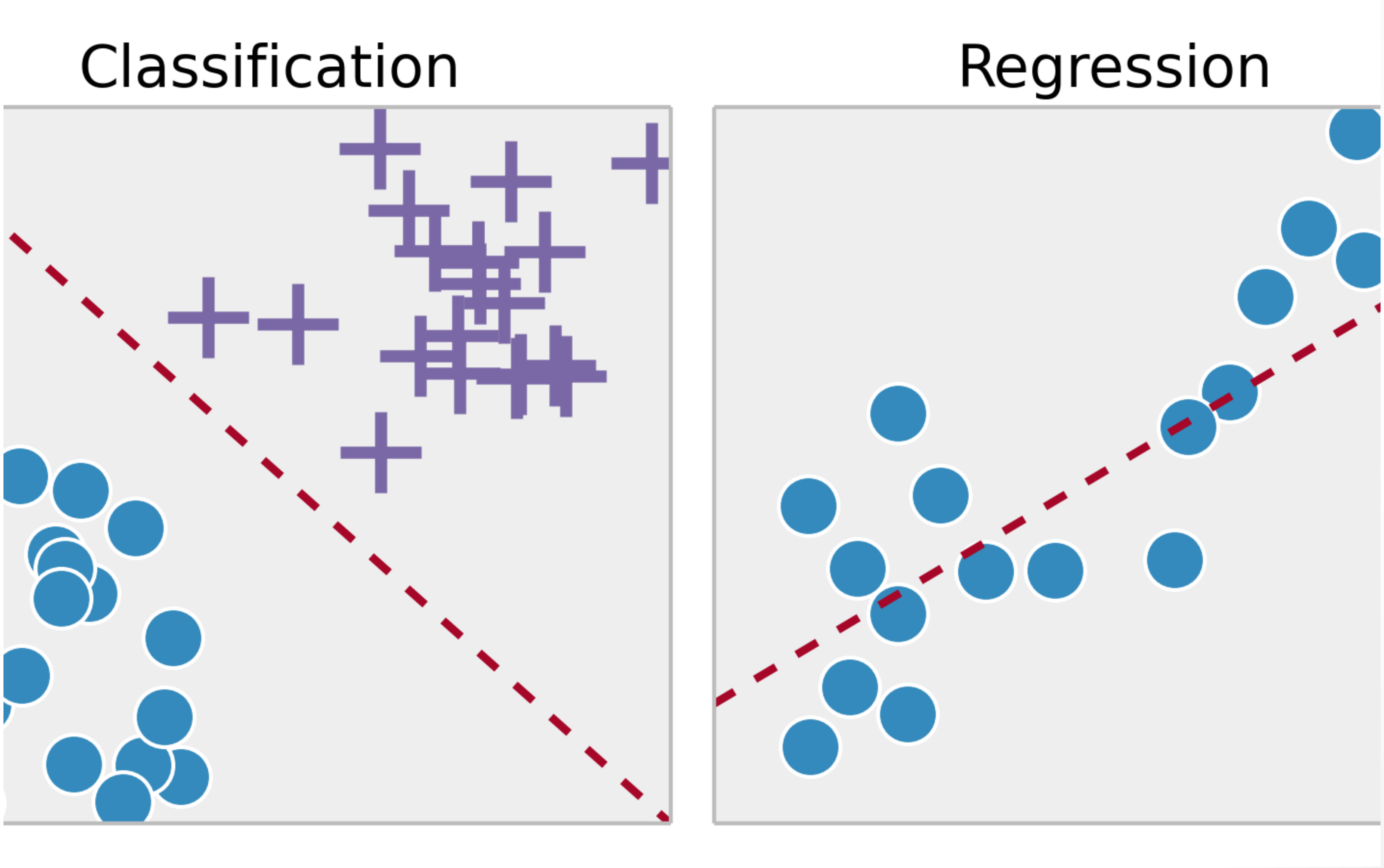
4

Supervised vs Unsupervised Learning

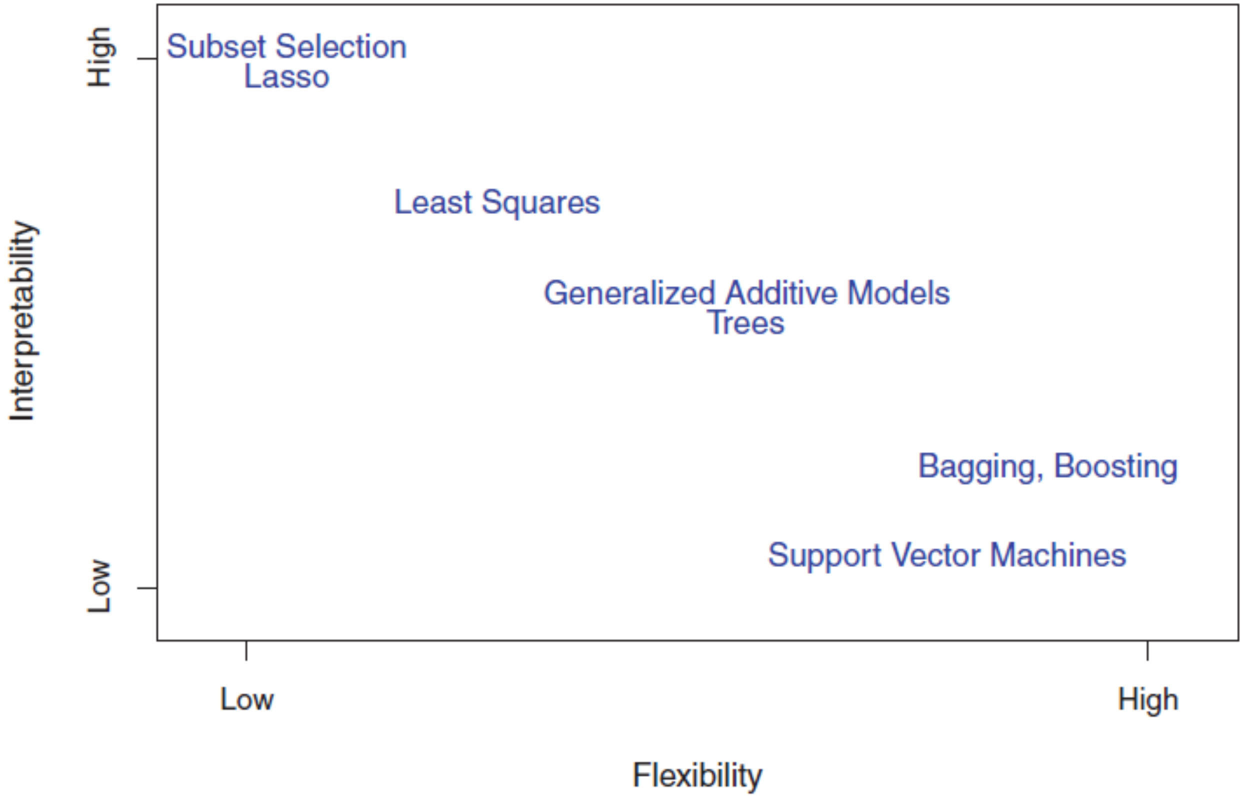


	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

Supervised Learning – Classification vs Regression



Flexibility vs Interpretability



Why data types matter?

- Different data types have different limitations and strengths.
- Certain types of analyses aren't possible with certain data types.
- There are 3 types of data which we may use for analysis:

1. Time series data: Time-Series data are data on one variable collected at a single point in time, e.g.
2. Cross-sectional data: Cross-sectional data are data on one or more variables collected at a single point in time, e.g.
3. Panel data, a combination of 1. & 2.

Cross-Section or Time Series?

- You would like to find the the relationship between company size and the return to investing in its shares. You survey different company sizes and their corresponding returns. What type of data did you collect?

- You would like to discover how the value of Apple’s stock price has varied when it announced the value of its dividend payment. What type of data will you collect to answer this question?

PARSE: UNDERSTANDING YOUR DATA

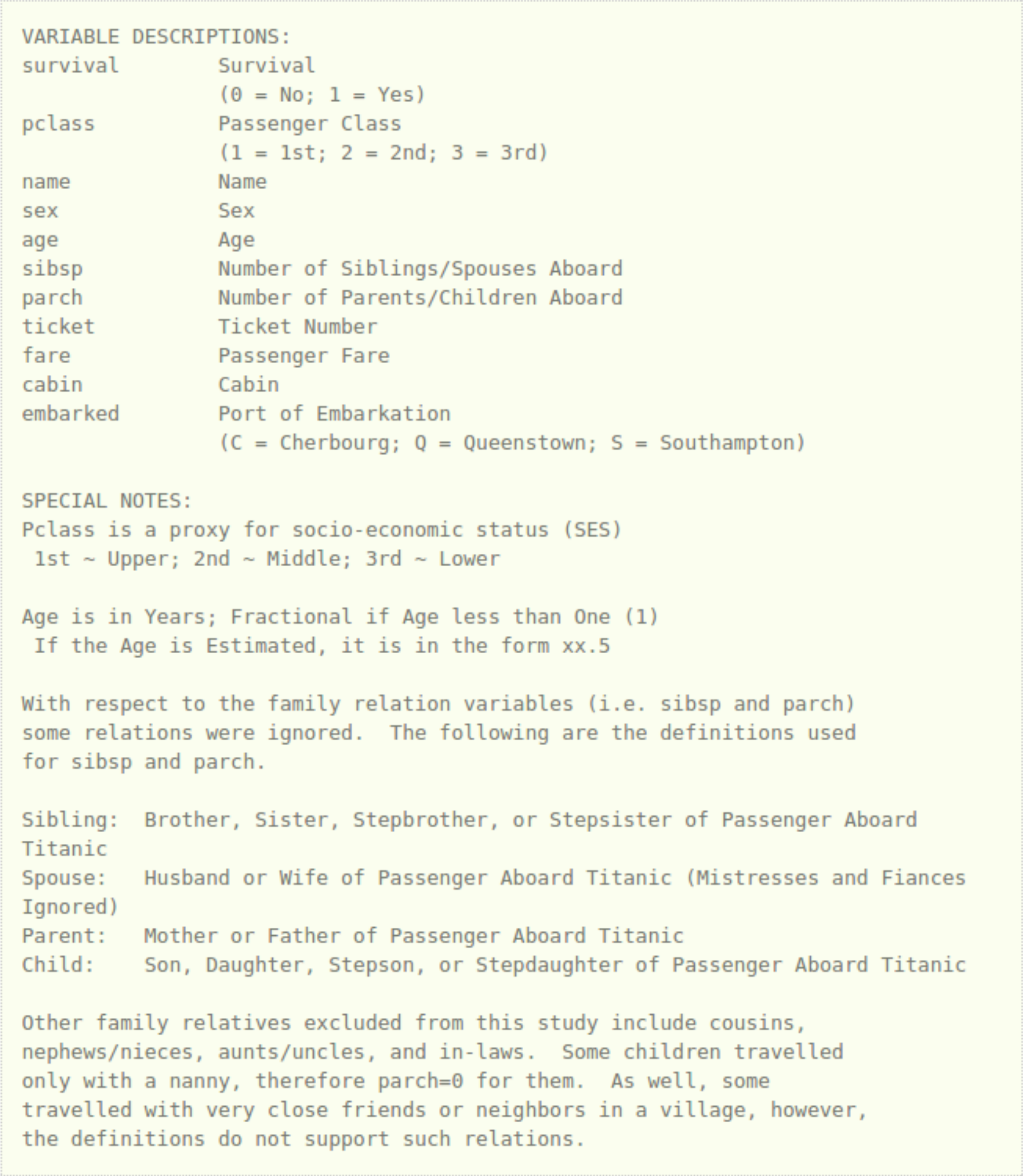
- You need to understand what you’re working with.

- To better understand your data
  - Create or review the data dictionary
  - Perform exploratory surface analysis
  - Describe data structure and information being collected
  - Explore variables and data types

INTRO TO DATA DICTIONARIES AND DOCUMENTATION

- Data dictionaries help judge the quality of the data.
- They also help understand how it’s coded.
  - Does gender = 1 mean female or male?
  - Is the currency dollars or euros?
- Data dictionaries help identify any requirements, assumptions, and constraints of the data.
- They make it easier to share data.

DATA DICTIONARY EXAMPLE: KAGGLE TITANIC DATA



```
VARIABLE DESCRIPTIONS:
survival        Survival
                (0 = No; 1 = Yes)
pclass          Passenger Class
                (1 = 1st; 2 = 2nd; 3 = 3rd)
name            Name
sex             Sex
age             Age
sibsp           Number of Siblings/Spouses Aboard
parch           Number of Parents/Children Aboard
ticket          Ticket Number
fare            Passenger Fare
cabin           Cabin
embarked        Port of Embarkation
                (C = Cherbourg; Q = Queenstown; S = Southampton)

SPECIAL NOTES:
Pclass is a proxy for socio-economic status (SES)
 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)
 If the Age is Estimated, it is in the form xx.5

With respect to the family relation variables (i.e. sibsp and parch)
some relations were ignored.  The following are the definitions used
for sibsp and parch.

Sibling:  Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard
Titanic
Spouse:   Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances
Ignored)
Parent:   Mother or Father of Passenger Aboard Titanic
Child:    Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

Other family relatives excluded from this study include cousins,
nephews/nieces, aunts/uncles, and in-laws.  Some children travelled
only with a nanny, therefore parch=0 for them.  As well, some
travelled with very close friends or neighbors in a village, however,
the definitions do not support such relations.
```

NUMPY AND PANDAS INTRO

NUMPY AND PANDAS INTRO

- What are Numpy and Pandas? Python packages

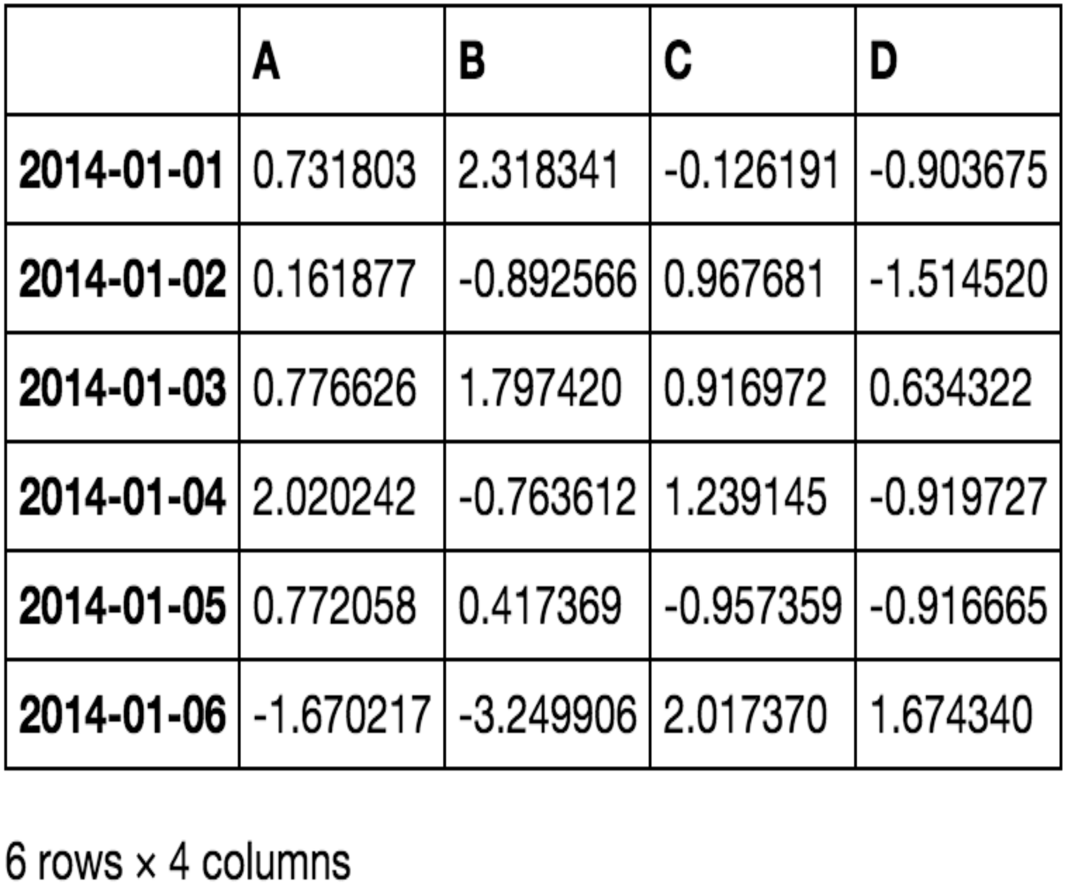
- Pandas is built on Numpy.

- Numpy uses arrays to do basic math and slice and index data.

- Pandas uses a data structure called a Dataframe.

- Dataframes are similar to Excel tables; they contain rows and columns.

NUMPY AND PANDAS INTRO



	A	B	C	D
2014-01-01	0.731803	2.318341	-0.126191	-0.903675
2014-01-02	0.161877	-0.892566	0.967681	-1.514520
2014-01-03	0.776626	1.797420	0.916972	0.634322
2014-01-04	2.020242	-0.763612	1.239145	-0.919727
2014-01-05	0.772058	0.417369	-0.957359	-0.916665
2014-01-06	-1.670217	-3.249906	2.017370	1.674340

6 rows × 4 columns

NUMPY AND PANDAS INTRO

- With these packages, you can select pieces of data, do basic operations, calculate summary statistics.

- Follow along and code along as we learn about Numpy and Pandas.