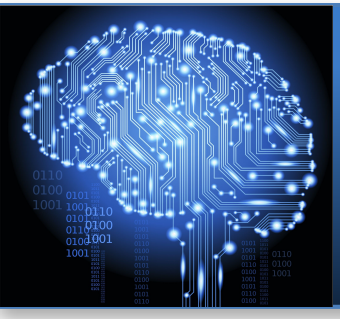


FIT1043 Lecture 2
Introduction to Data Science

Mahsa Salehi*

Faculty of Information Technology, Monash University

Semester 2, 2022

*with material from © Wray Buntine

Discussion: Data Science Jobs

Data Science Job Market in Australia

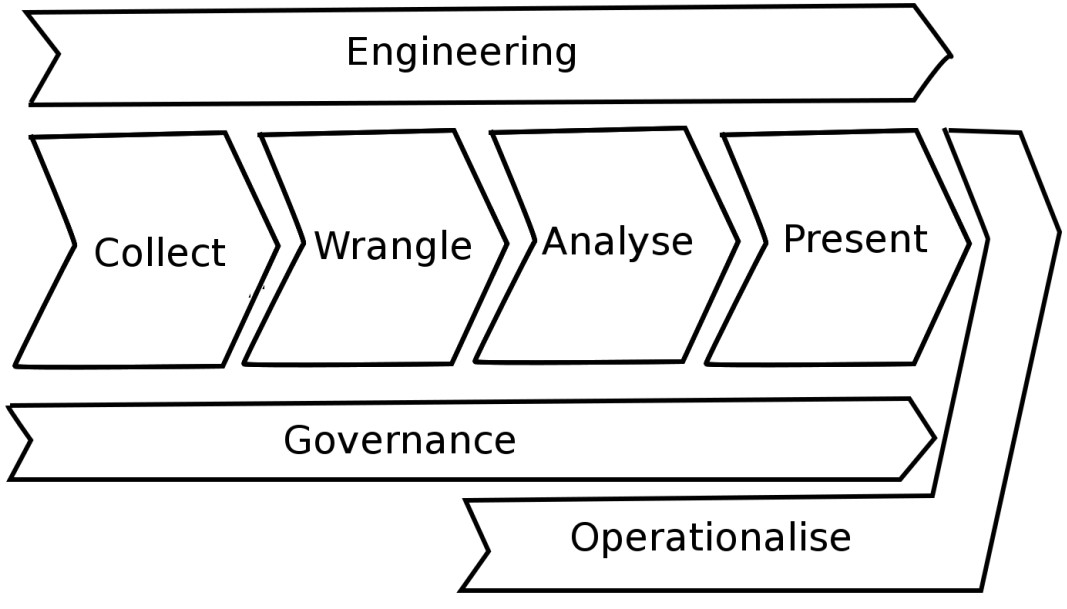
- smaller (per capita) market compared to USA & UK, where giant industry players are making better use of Data Science

Job Advertisements:

- communication skills and domain expertise are rated highly
  - different jobs require different toolset skills
- Week 3 pre-class activity
  - see Adzuna's CV upload page for an interesting application!

Our Standard Value Chain

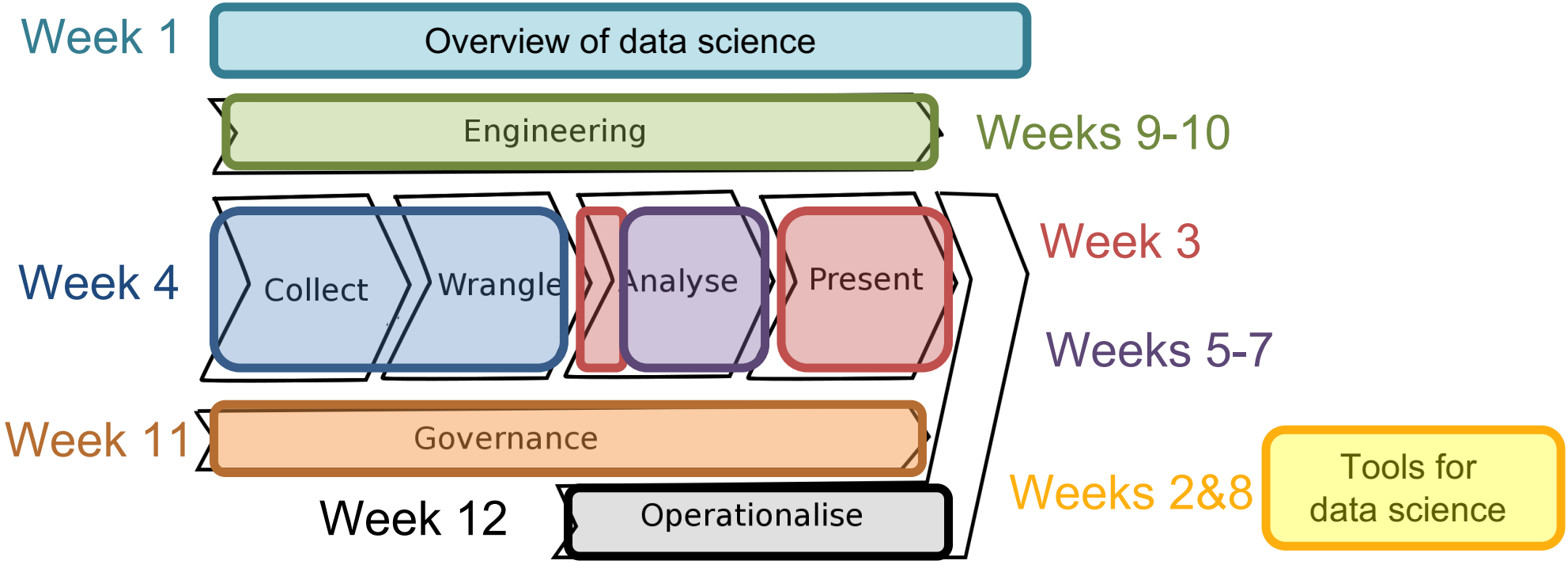
Collection: getting the data
Engineering: storage and computational resources
Governance: overall management of data
Wrangling: data preprocessing, cleaning
Analysis: discovery (learning, visualisation, *etc.*)
Presentation: arguing that results are significant and useful
Operationalisation: putting the results to work



We will refer to this throughout the semester!

Our Standard Value Chain

Collection: getting the data
Engineering: storage and computational resources
Governance: overall management of data
Wrangling: data preprocessing, cleaning
Analysis: discovery (learning, visualisation, *etc.*)
Presentation: arguing that results are significant and useful
Operationalisation: putting the results to work



Unit Schedule

Week	Activities	Assignments
1	Overview of data science	Weekly quiz
2	Introduction to Python for data science	
3	Data visualisation and descriptive statistics	
4	Data sources and data wrangling	
5	Data analysis theory	Assignment 1
6	Regression analysis	
7	Classification and clustering	
8	Introduction to R for data science	
9	Characterising data and "big" data	Assignment 2
10	Big data processing	
11	Issues in data management	
12	Industry guest lecture	Assignment 3

Outline

- Introduction to Python for Data Science
  - Motivation to studying Python
  - Python data types
  - Essential libraries

- Overview of data science (cont)
  - Data science roles and skills
  - Impact of data science
  - Business models with data

Learning Outcomes (Week 2)

By the end of this week you should be able to:

- Comprehend essentials for coding in Python for data science
- Explain and interpret given **Python** codes
- Explain **different data science roles and skills** and comprehend the differences between them
- Explain **Impact** of data science
- Explain the **data business models** for organizations

Introduction to Python for Data Science

From Python Data Science Handbook by J. Vanderplas

The 2021 Top Programming Languages

Rank	Language	Type	Score
1	Python	Web, Enterprise, Embedded	100.0
2	Java	Web, Mobile, Enterprise	95.4
3	C	Mobile, Enterprise, Embedded	94.7
4	C++	Mobile, Enterprise, Embedded	92.4
5	JavaScript	Web	88.1
6	C#	Web, Mobile, Enterprise, Embedded	82.4
7	R	Enterprise	81.7
8	Go	Web, Enterprise	77.7
9	HTML	Web	75.4
10	Swift	Mobile, Enterprise	70.4

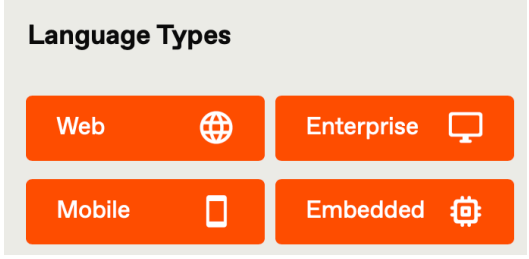
Language Types

- Web
- Enterprise
- Mobile
- Embedded

image src: IEEE

The 2021 Top Programming Languages

Rank	Language	Type	Score
1	Python		100.0
2	Java		95.4
3	C		94.7
4	C++		92.4
5	JavaScript		88.1
6	C#		82.4
7	R		81.7
8	Go		77.7
9	HTML		75.4
10	Swift		70.4



Language Rank	Types	Spectrum Ranking
1. Python		100.0
2. C++		99.7
3. Java		97.5
4. C		96.7
5. C#		89.4
6. PHP		84.9
7. R		82.9
8. JavaScript		82.6
9. Go		76.4
10. Assembly		74.1

2018

image src: IEEE

Data Science Preferred Tools

- **Python's Role in Data Science**

- Many tools out there for data science.

- Python has gained popularity over the last few years.
  - easy to learn
  - flexible and multi-purpose
  - great libraries
  - well designed computer language
  - good visualization for basic analysis

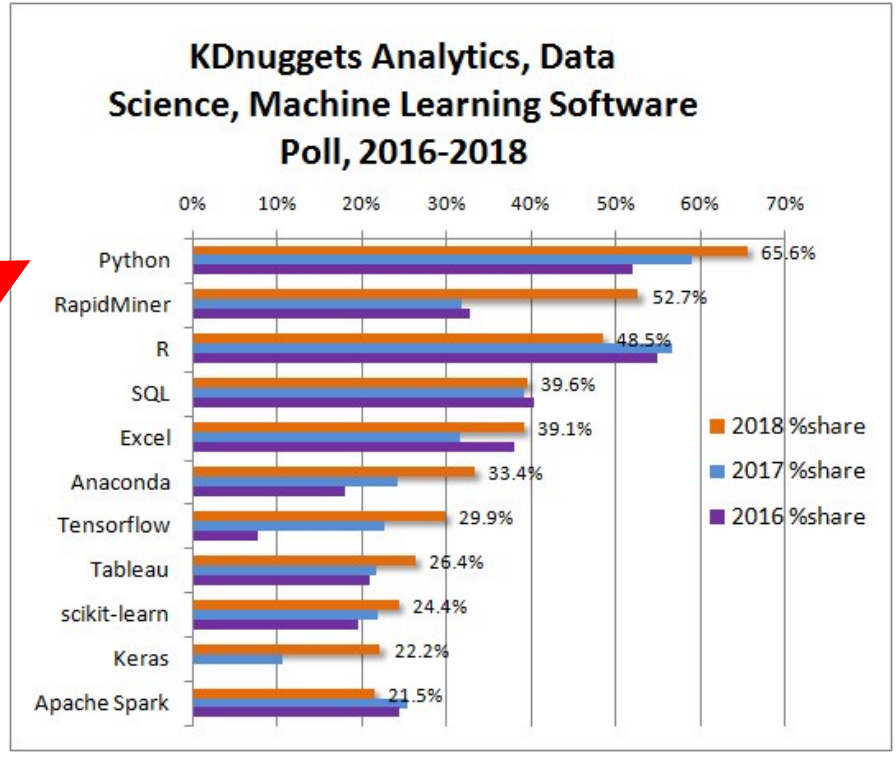


image src: kdnuggets.com

Setting Up Python Environment

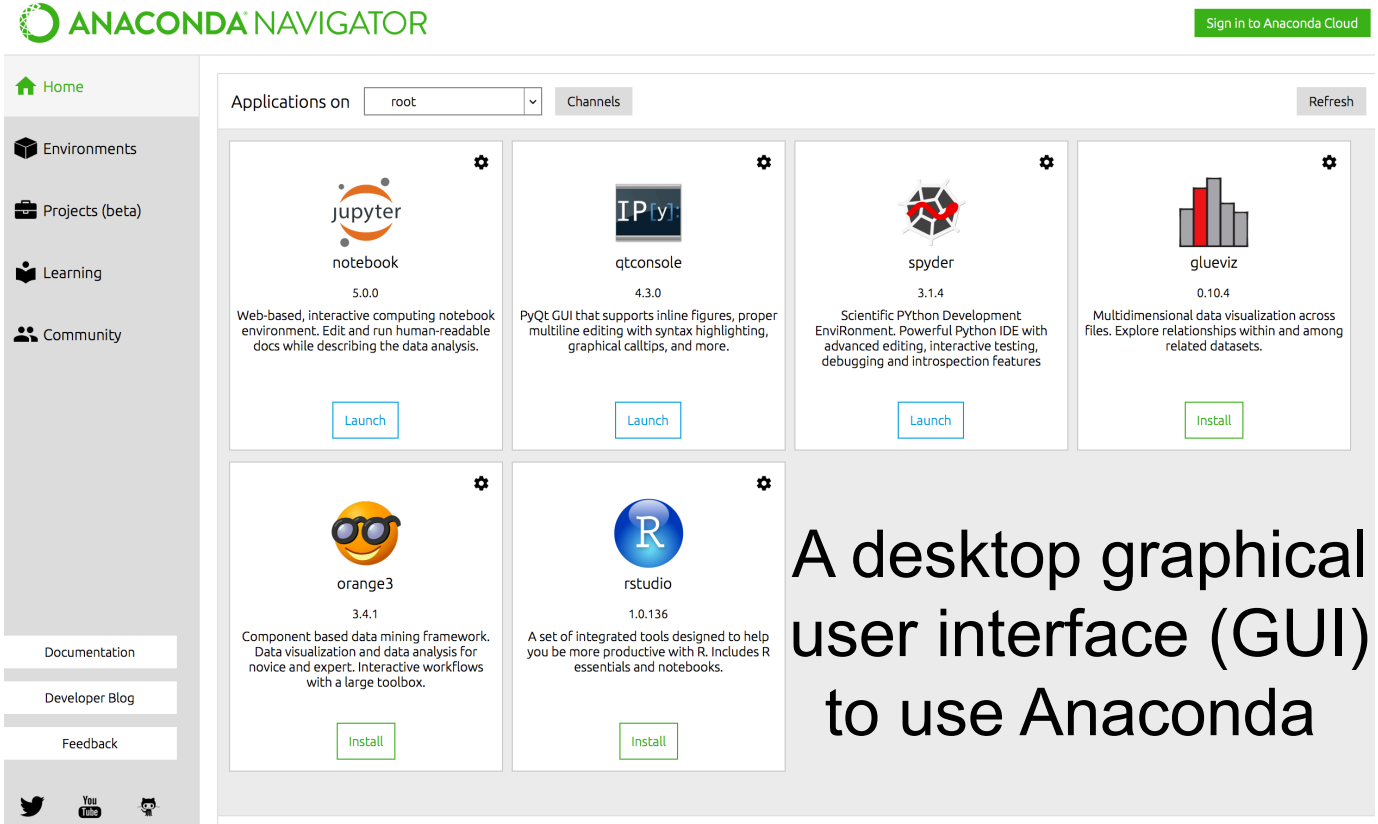
- Python 2.x vs 3.x

- IPython vs Jupyter Project
  - IPython (Interactive Python) is a useful interactive interface to Python, and provides a number of useful syntactic additions to the language
  - Jupyter provides a browser-based notebook useful for development, collaboration and publication of results.

IP[y]: IPython
Interactive Computing

Anaconda Project

- Anaconda is a package manager, an environment manager, a Python/R data science distribution, and a collection of over 1,500+ open source packages. Anaconda is free and easy to install.



A desktop graphical user interface (GUI) to use Anaconda

FLUX Question

What is .ipynb?

A. An illegal file extension.
B. Interactive Python NoteBook.
C. Intelligent Python Nota Bene.
D. Typo, it should be ‘pinyin’

Python Basic Types

- Integers
- Floating-Point Numbers
- Boolean
  - True/False
- Strings

Integers (int)

- Python interprets a sequence of decimal (power of 10) digits without any prefix (0b, 0o or 0x) to be a decimal number:
  - 0b is interpreted as a binary sequence of numbers

```
>>> print(0b10)
2
```

  - 0o is interpreted as a octal sequence of numbers (rarely used)

```
>>> print(0o10)
8
```

  - 0x is interpreted as a hexadecimal sequence of numbers

```
>>> print(0x10)
16
```

Floating Point (float)

- The values are specified with a decimal point.

```
>>> 4.2
4.2
>>> type(4.2)
float
>>> 4.
4.0
```

- For scientific notation style, the character `e` followed by a positive or negative integer may be used.

```
>>> .4e7
4000000.0
>>> type(.4e7)
float
>>> 4.2e-4
0.00042
```

Boolean (bool)

- Note that this type is only available in Python 3 and it is not in Python 2.
- Boolean type (in any language) has one of two values, True or False

```
>>> type(True)
bool
>>> type(False)
bool
>>> print(True | False)
True
```

Strings (str)

- Strings are delimited using either the single or double quotes.
- Only the characters between the opening delimiter and matching closing delimiter are part of the string.

```
>>> print("I am a string.")
I am a string.

>>> type("I am a string.")
str
```

Strings (str)

- Handling strings can be a bit more complicated than we initially think.
- For example, if we want to include quotes.
  - You aren't simple

```
>>> print('you aren't
simple')
SyntaxError: invalid
character in identifier

>>> print("you aren't
simple")
you aren't simple
```

Strings (str)

- The earlier example is just for the basics of putting the sequence of characters between the delimiters as a string.

- There are many other considerations to cater for special characters in strings handling.

- Use \ (back-slash) as the escape character.

```
>>> print('you aren\'t simple')
you aren't simple
```

- There are a few reserved special escape characters:

- `\t` Tab
- `\n` New line
- `\uxxxx` 16-bit unicode character

Dynamic Typed Language

For those who learned programming with static typed languages, you will need to declare the variables, e.g., in C.

```
int x;
```

In Python, there is no declaration and it is only known at run-time.

```
>>> x = 10
>>> print(type(x))

>>> x = 'Hello, world'
>>> print(type(x))
```

Built-in Functions

- There are more than 65 built-in functions in the current Python version. These functions cover
  - Maths
  - Type Conversions
  - Iterators
  - Composite Data Types
  - Classes, Attributes, and Inheritance
  - Input/Output
  - Variables, References, and Scope
  - Others

- You can refer to them here

Operators and Strings Manipulation

- Arithmetic operators
  +, -, *, /, % etc.
- Comparison operators
  >, <, <=, >=, !=, ==
- String operators
  +, *, in

```
>>> s = 'foobar'
>>> s[0]
'f'
>>> s[3]
'b'
>>> len(s)
6
>>> s[len(s)-1]
'r'
>>> s[-1]
'r'
```

Strings(useful for Data Science)

- String subset

```
>>> s = 'foobar'
>>> s[2:5]
'oba'
>>> s[0:4]
'foob'
>>> s[2:]
'obar'
>>> s[:4] + s[4:]
'foobar'
>>> s[:4] + s[4:] == s
True
```

- Striding

```
>>> s = 'foobar'
>>> s[0:6:2]
'foa'
>>> s[1:6:2]
'obr'
```

More Python Data Types

Lists and tuples are useful Python data types.

- A Python list is a collection of objects (not necessary the same).
- Lists are defined by square brackets that encloses a comma-separated sequence of objects `([])`

```
>>> a = ['foo', 'bar',
'baz', 'qux']
>>> print(a)
['foo', 'bar', 'baz', 'qux']
```

- Lists are ordered.
- Lists can contain any arbitrary objects.
- List elements can be accessed by index.
- Lists can be nested to arbitrary depth.
- Lists are mutable.
- Lists are dynamic.

More Python Data Types

Tuple

- Tuples are identical to lists in all aspects except that the content are immutable (fixed).
- Tuples are defined by round brackets (parentheses) that encloses a comma-separated sequence of objects `()`.

Dictionary

- Dictionary is similar to a list in that it is a collection of objects.
- Only difference is that list is ordered and indexed by their position whereas dictionary is indexed by the key.
  - Think of it as a key-value pair.
  - This maps nicely to Data Science when there is access to NoSQL databases that stores items in key-value pairs.

Dictionary

```
d = dict([
    (<key>, <value>),
    (<key>, <value),
      .
      .
      .
    (<key>, <value>)
])
```

```
>>> person = {}
>>> person['fname'] = 'Ian'
>>> person['lname'] = 'Tan'
>>> person['age'] = 19
>>> person['pets'] = {'dog':
'Barney', 'cat': 'Dino'}
>>> person
{'fname': Ian', 'lname':
'Tan', 'age': 19, 'pets':
{'dog': 'Barney', 'cat':
'Dino'}}
```

Controls

Conditions

```
if <expr>:
    <statement>
elif <expr>:
    <statement(s)>
elif <expr>:
    <statement(s)>
else:
    <statement(s)>
```

Iterations

```
while <expr>:
    <statement(s)>
```

Python `for` loops link

Note: Python uses indentation!

Essential Python and Data Science

Specific libraries that are considered as the “starter pack” for Data Science:

- Numpy: Scientific computing, support for multi-dimensional arrays
- Pandas: Data structures as well as operations for manipulating numerical tables.
- Matplotlib: library for visualization
- Scikit-learn: Python machine learning library that provides the tools for data mining and data analysis

For some, you may also want to look at

- NLTK: Natural Language ToolKit to work with human language data

Loading Libraries

The general syntax to include a library:

```
>>> import numpy as np
>>> import pandas as pd
>>> from matplotlib import pyplot as plt
>>> import matplotlib.pyplot as plt
```

Let' Start!

- Data Science needs DATA
  - Reading data
  - Writing data

- We can read data from different sources
  - Flat files
  - **CSV files**
  - Excel files
  - Image files
  - Relational databases
  - NoSQL databases
  - Web

Reading from CSV

- Python has a built in CSV reader but for Data Science purposes, we will use the pandas library.

- Assuming your file name is filename.csv

```
>>> import pandas as pd
>>> data = pd.read_csv("filename.csv")
>>> data.head()

>>> X = data[["Age"]]
>>> print(X)
```

Usual 1st Step upon Obtaining Data

- A description or a summary of it.

- Sometimes, referred to as ***five number summary*** if the data is numeric.
  - Minimum, maximum, median, 1st quartile, 3rd quartile

- Work with pandas DataFrames.

```
>>> df = pd.DataFrame(data)
>>> print(df)

>>> df.describe()
```

Working with DataFrames (Basic)

- Select a column by using its column name:

```
>>> df['Name']
0 Braund, Mr. Owen Harris
1 Cumings, Mrs. John Bradley (Florence Briggs
Thayer)
```

- Select multiple columns using a list of column names:

```
>>> df[['Name', 'Survived']]
Name Survived
0 Braund, Mr. Owen Harris 0
1 Cumings, Mrs. John Bradley (Florence Briggs
Thayer) 1
```

- Select a value using the column name and row index:

```
>>> df['Name'][3]
'Futrelle, Mrs. Jacques Heath (Lily May Peel)'
```

Working with DataFrames (Basic)

- Select a particular row from the table:

```
>>> df.loc[2]
PassengerId                              3
Survived                                 1
Pclass                                   3
Name           Heikkinen, Miss. Laina
Sex                                 female
Age                                     26
SibSp                                    0
Parch                                    0
Ticket                  STON/O2. 3101282
Fare                                 7.925
Cabin                                  NaN
Embarked                                 S
Name: 2, dtype: object
```

Working with DataFrames (Basic)

- Select all rows with a particular value in one of the columns:

```
>>> df.loc[df['Age'] <= 6]
```

Pair Discussion

Save the Data

- Assuming you just want to analyse a part of the data and you want to save a resulting data frame to a CSV file.

```
>>> df2 = df.loc[df['Age'] >= 12]
>>> df2.to_csv ('output.csv', index =
None, header=True)
```

- We have now read, describe, basic data exploration and save the data.

Working With Data

- There are some basic data pre-processing that are usually done or at least taken into consideration.
  - Removing duplicates
  - **Categorical data**
  - Dealing with dates
  - Missing data
  - **Subsetting data**
  - Concatenating
  - Transforming
  - **Aggregating**
- More will be explored in week 4

Categorical Data

- A categorical data is one that has a specific value from a limited set of values. The options are fixed.

- A ticket class is generally categorical, i.e. 1st class, 2nd class & 3rd class.

```
>>> df.loc[df['Class'] == 1]
```

- We can create our own categories, e.g.

```
>>> import pandas as pd
>>> tix_class = pd.Series(['1st','2nd','3rd'],
dtype='category')
```

Subsetting Data

- We actually already have done this a few slides before ☺
- Extract only those that survived

```
>>> df.loc[df['Survived'] == 1]
```

- What does the code below return?

```
>>> df.loc[(df['Sex'] == 'female') &
(df['Survived'] == 1)]
```

Slicing Data

- Slice rows by row index.

```
>>> df[:5]
>>> df[3:10]
```

- If we only want certain columns, e.g. Age, Name, Sex, Survived

```
>>> df.loc[:,
('Age','Name','Sex','Survived')]
```

Aggregating

- Like our 5 number statistic, we can also obtain aggregated values for columns. The total fare can be easily obtained by

```
>>> df['Fare'].sum()
4385.095600000001
```

- Or we can get the average age of the passenger by

```
>>> df['Age'].mean()
28.141507936507935
```

- Check the answers against the `df.describe()` earlier

Aggregating

- Like in SQL, we often want to know the aggregated values for certain values from another column. Similarly, we can use the groupby function:

```
>>> df.groupby('Sex')['Age'].mean()
Sex
female    24.468085
male      30.326962
Name: Age, dtype: float64
```

Aggregating

- What does the following mean?

```
>>>df.loc[df['Survived']==1].groupby('Sex')['Age'].mean()
Sex
female    26.265625
male      23.314444
Name: Age, dtype: float64
```

- Compare it with the previous statement, what can you tell from it?

FLUX Question

What is a dataframe?

A. An array.

B. A list.

C. A theory about data.

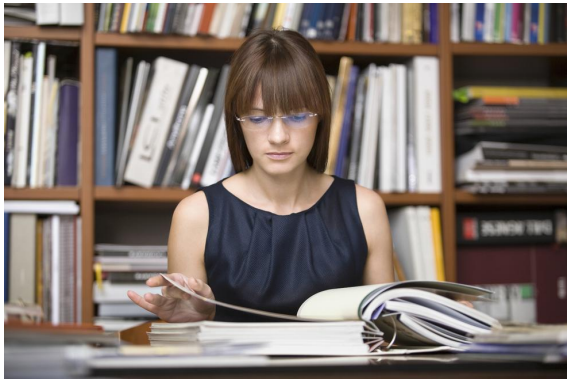
D. A structure that stores tabular data

Next few weeks

- We will be using Python for the next few weeks
  - MatPlotLib
  - Scikit-Learn

Suggested Reading

- You can easily look for Python resources online, to be specific, Python for Data Science.
  - An excellent online course will be from DataCamp



Tutorial/Lab week 2

- Introductory Python for data science
- Make sure participate in the within class activities
- Use forum if you wish to swap tutorials
- Email the role account for any other questions:
  - Clayton: fit1043.clayton-x@monash.edu

Roles of a Data Scientist

better understanding the different kinds of data scientists:

- reviewing:
  - from *Analyzing the Analyzers* from Harris, Murphy and Vaisman
- interviews
  - from *Data Analytics Handbook*

Roles of a Data Scientist 1:

Reviewing *Analyzing the Analyzers*

What is the Difference Between ...

A *quote from Quora from Jason Widjaja*:

> **Data analysts** *are primarily people who develop insights with data. ....*
>
> **Data scientists** *are primarily people who develop data models and products, that in turn produce insights. ...*
>
> **Data engineers** *are primarily people who manage data infrastructure, automate data processing and deploy models at scale. ...*

(Note the use of the word “primarily”!)

see also *Job Comparison* – Data Scientist vs Data Engineer vs Statistician

FLUX Question

Consider the definition given for data science, is the boundary between data science, data engineering and data analysis fixed?

A. TRUE
B. FALSE

Skills of Data Scientists

Analyzing the Analyzers, Harris, Murphy and Vaisman, 2013

Business: product development, business

Machine learning/Big data: machine learning, big and distributed data

Mathematics/Operations research: optimisation, mathematics, graphical models, algorithms

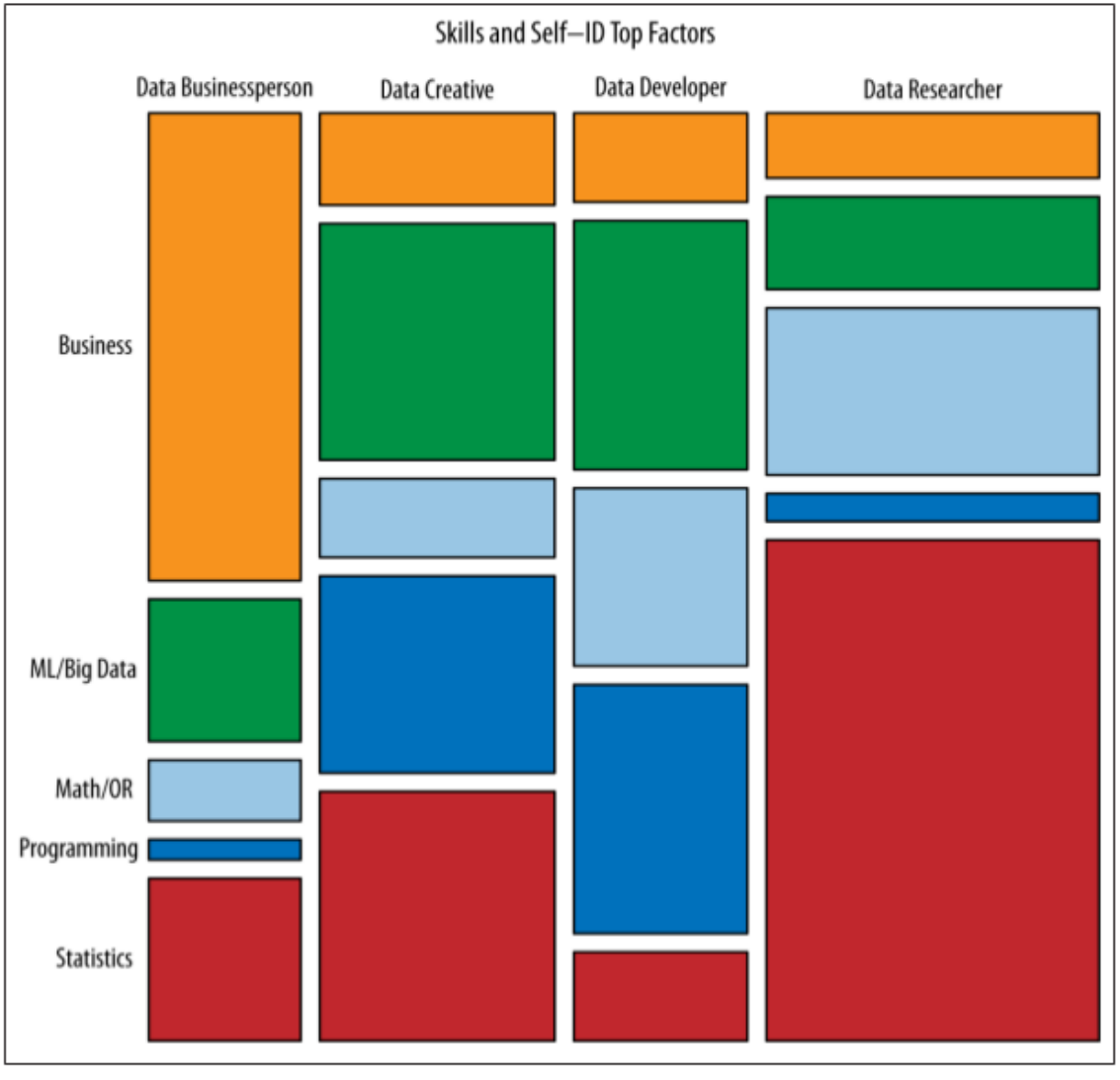
Programming: systems administration, back end programming, front end programming

Statistics: visualisation, temporal statistics, spatial statistics, science, data manipulation

NB. typical data scientist doesn't have to know all of these!

Mapping Styles to Skills

Analyzing the Analyzers, Harris, Murphy and Vaisman, 2013



X-axis:
different roles

Y-axis:
different skills

which might you be?

Roles of a Data Scientist 2:

Interviews from *Data Analytics Handbook*

From *Data Analytics Handbook*

The *Data Analytics Handbook* is a four volume set of long interviews from industry and academic professionals in the field.

Volume 1 deals with practitioners:

- What exactly do the sexy “data scientists” do?
- What other professions are there in big data?
- What tools do they use to accomplish their tasks?
- How can I enter the industry if I don’t have a Ph.D. in Statistics?

Lessons from the DA Handbook

1. Communication skills are underrated.
2. The biggest challenge for a data analyst is the **Collection** and **Wrangling** steps.
3. A data scientist is better at statistics than a software engineer and better at software engineering than a statistician.
4. The data industry is still nascent and the roles less well defined so you get to interact with many parts of the company from engineering to business intelligence to product managers.
5. Keep a **curiosity** about working with data, a quality as important as your technical abilities.

Career as Data Scientist

To become a specialist you need:

- solid machine learning and statistics
- related mathematics (1st+2nd year in many degrees)
- solid prototyping (R, Python, Java)
- perhaps Unix experience (Linux, Mac OSX)

Note: This unit provides an introduction and background only.

Data Science skills, Have vs Want

Kdnuggets, By **Gregory Piatetsky** (co-founder of KDD), 2020

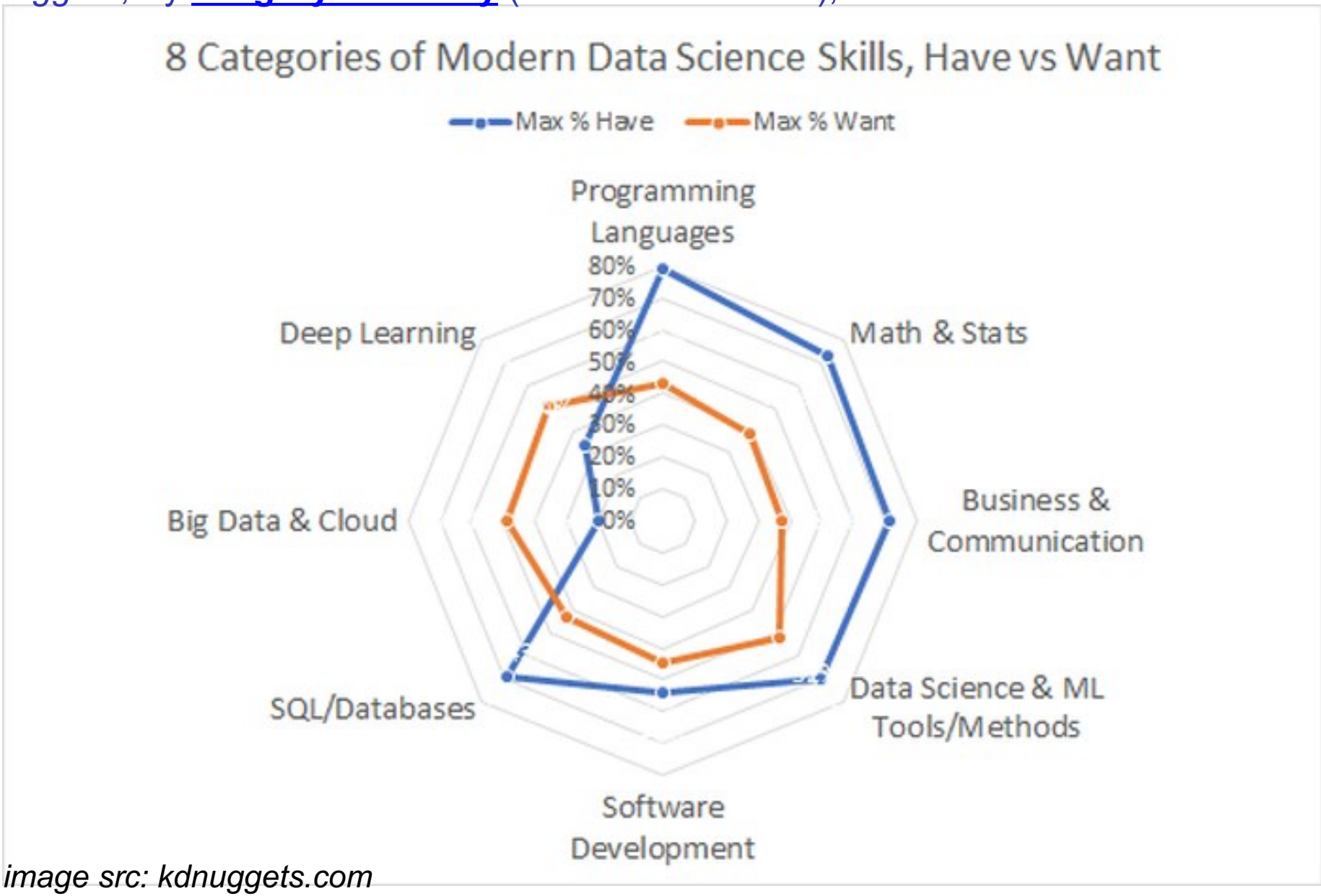


image src: kdnuggets.com

Impact of Data Science

some examples of how data science is impacting others:

- your life in the cloud
  - datafication of you
- social good
  - numerous examples and very rewarding
- futurology
  - healthcare and automobiles as examples

Impact of Data Science:
Your life in the cloud

datafication of you

Your Life on the Cloud

From *Year Zero: Our life timelines begin*

Our personal information is increasingly stored in the cloud:

- social life (Facebook datafies our friendship),
- career (LinkedIn datafies our professional accomplishments),
- location history (Google Map datafies our location),
- health and medical (Fitbit datafies our activities),
- music (Apple datafies our music taste), ...

This provides many, many advantages:

- *e.g.* personal agents, computerised support for health

But also some disadvantages:

Your Life on the Cloud (cont.)

But

- corporate leakage to government (security, tax, *etc.*)
- what if you don't have rights to access/delete data?
- security and privacy breaches
- what if we've changed our ways?
- the department of pre-crime

Impact of Data Science:
Social good

Data Science for Social Good

Data Science for Social Good movement training data scientists to support community and charity.

Many researchers in the Faculty of IT- at Monash work on data science for social good research projects, such as:

Improve road safety

Improve cardiac arrest survival rates

image src: emotiv.com, monash.edu

Impact of Data Science:
Futurology

some areas where significant impact is to be made in the future

Health Care Futurology

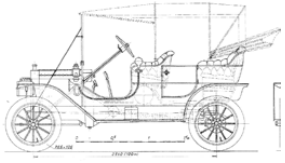
see “Big data – 2020 vision” talk by SAP manager John Schitka

- your stomach can be instrumented to assess contents, nutrients, *etc.*
- your bloodstream can be instrumented too assess insulin levels, *etc.*
- your “health” dashboard can be online and shared by your GP
- GP will know about your icecream/beer binge last night and you missing your morning run
- longitudinal studies feasible

longitudinal studies:= a method in which data is gathered for the same subjects repeatedly over a period of time

Car Computing Evolution Since Pre-1980s = Mechanical / Electrical → Simple Processors → Computers

Pre-1980s
Analog / Mechanical
Used switches / wiring to route feature controls to driver

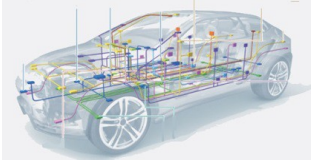


1980s (to Present)
CAN Bus (Integrated Network)
New regulatory standards drove need to monitor emissions in real time, hence central computer

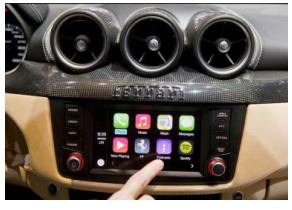
1990s (to Present)
OBD (On-Board Diagnostics) II
Monitor / report engine performance; Required in all USA cars post-1996

1990s-2010s
Feature-Built Computing + Early Connectivity
Automatic cruise control...
Infotainment...Telematics...GPS / Mapping...

Today = Complex Computing
Up to 100 Electronic Control Units / car...
Multiple bus networks per car (CAN / LIN / FlexRay / MOST)...
Drive by Wire...



Today = Smart / Connected Cars
Embedded / tethered connectivity...
Big Tech = New Tier 1 auto supplier (CarPlay / Android Auto)...



Tomorrow = Computers Go Mobile?...
Central hub / decentralized systems?
LIDAR...
Vehicle-to-Vehicle (V2V) / Vehicle-to-Infrastructure (V2I) / 5G...
Security software...

"The Box" (Brooks & Bone)

Source: KPCB Green Investing Team, Darren Liccardo (DJI); Reilly Brennan (Stanford); Tom Denton, "Automobile Electrical and Electronics Systems, 3rd Edition," Oxford, UK: Tom Denton, 2004; Samuel DaCosta, Popular Mechanics, Techmor, US EPA, Elec-Intro.com, Autoweb, General Motors, Garmin, Evaluation Engineering, Digi-Key Electronics, Renesas, Jason Aldag and Jhaan Elker / Washington Post, James Brooks / Richard Bone, Shareable

FLUX Question

Referring to the two slides about the car industry:

First, they underwent a digitization process, followed by a _____ process

FLUX Question

Using a word or short phrase, name a non-automotive industries that have had similar developments in recent decades.

How do you expect the datafication process to change <*these industries*>?
More effective in _____

Automobile Futurology

see “Big data – 2020 vision” talk by SAP manager John Schitka

Self driving cars:

- how does the city replace traffic fine revenue?
- can you drink and drive if the car is automatic?
- what happens to the taxi industry?
- what happens to the auto insurance industry?
- what happens to people still “self” driving, and their insurance?

Business Models with Data

what kinds of businesses do we have operating in the Data Science world?

Business Models

From Wikipedia:

> *A business model describes the rationale of how an organization creates, delivers, and captures value, in economic, social, cultural or other contexts.*

Examples of general classes:

- retailer versus wholesaler
- luxury consumer products
- software vendor
- service provider

What kinds of businesses do we have operating in the Data Science world?

Business Models with Data:
Data business models

what are some business models specific to data science?

Amazon.com

- an assembly line for the retail industry, with support for embedded online retailers
- huge stock of books, DVDs, CDs, *etc.* easily searchable
- extensive customer reviews

Amazon.com (cont.)

Information-based differentiation: satisfies customers by providing a differentiated service:

- superior information including reviews about products
- superior range

Information-based delivery network: they deliver information for others:

- retailers in the Amazon marketplace get customers directed to them
- retailers can advertise

Other Data Business Models

information brokering service: buys and sells data/information for others.

information provider: business is based on selling the data/information it collects.

"What a Big-Data Business Model Looks Like" by Ray Wang in the Harvard Business Review claims these are unique in the data world.

Suggested Reading

From *Data Analytics Handbook* read the interviews of

- Abraham Cabangbang (2 pp)
- Ben Bregman (2 pp)
- Leon Rudyak (3 pp)