BISA x CBA Presents:

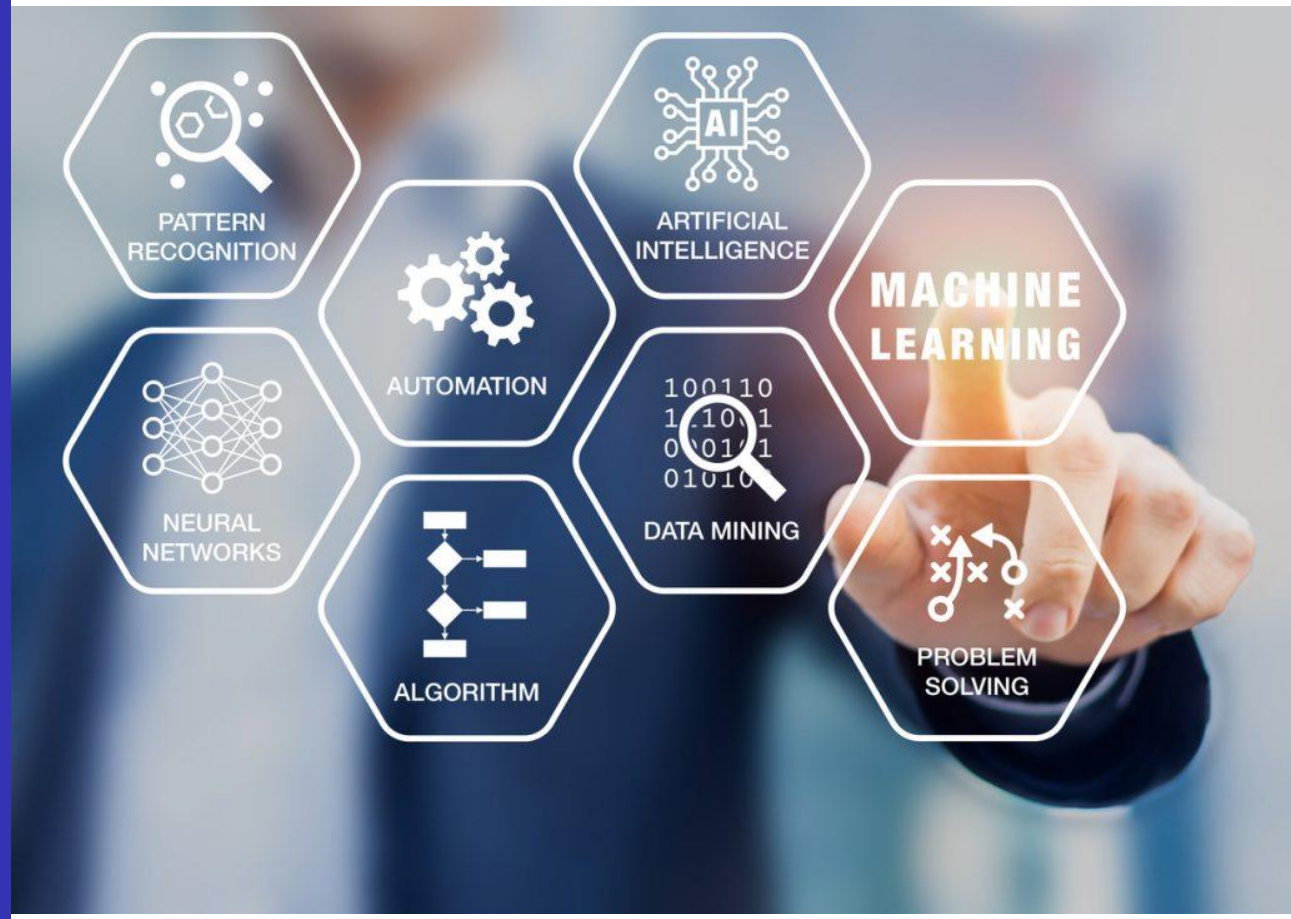
Introduction to Python for Data Science

Today's Breakdown

- Presentation
  - What is Data Science, Data Visualisation, Python
- Use Cases in CBA
- Q&A Session with CBA
- Networking / Food Break
- Popular Python Libraries
- Coding Activity 1: Basic fundamentals
- Coding Activity 2: Data visualisation on a real dataset
- Closing and extra resources

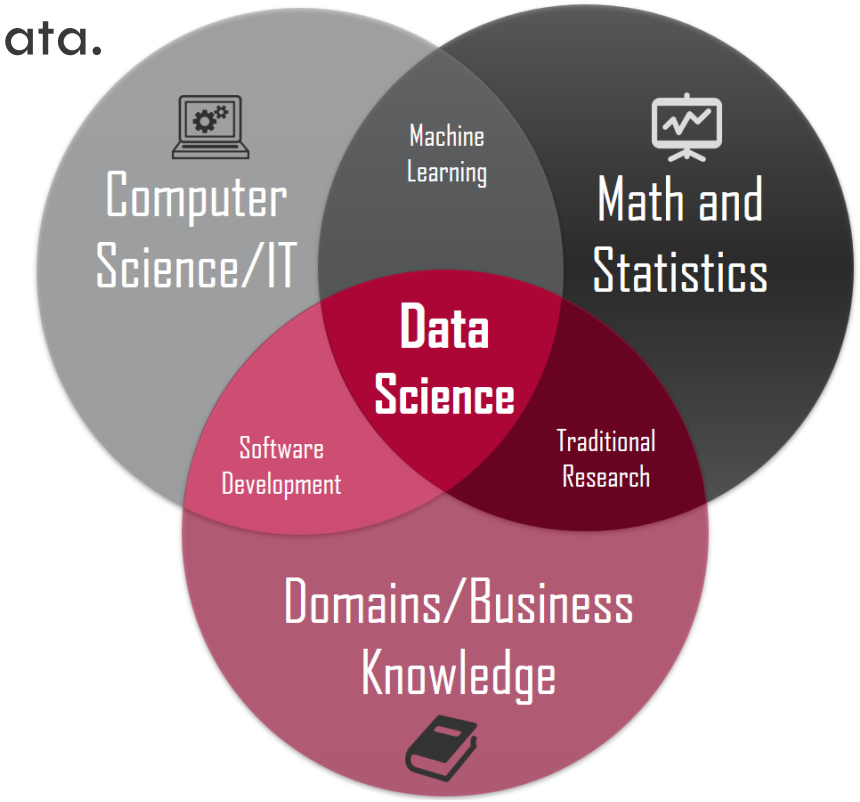
- **Also, if you haven't already:** facebook.com/ **BisaSydney**

What is Data Science?

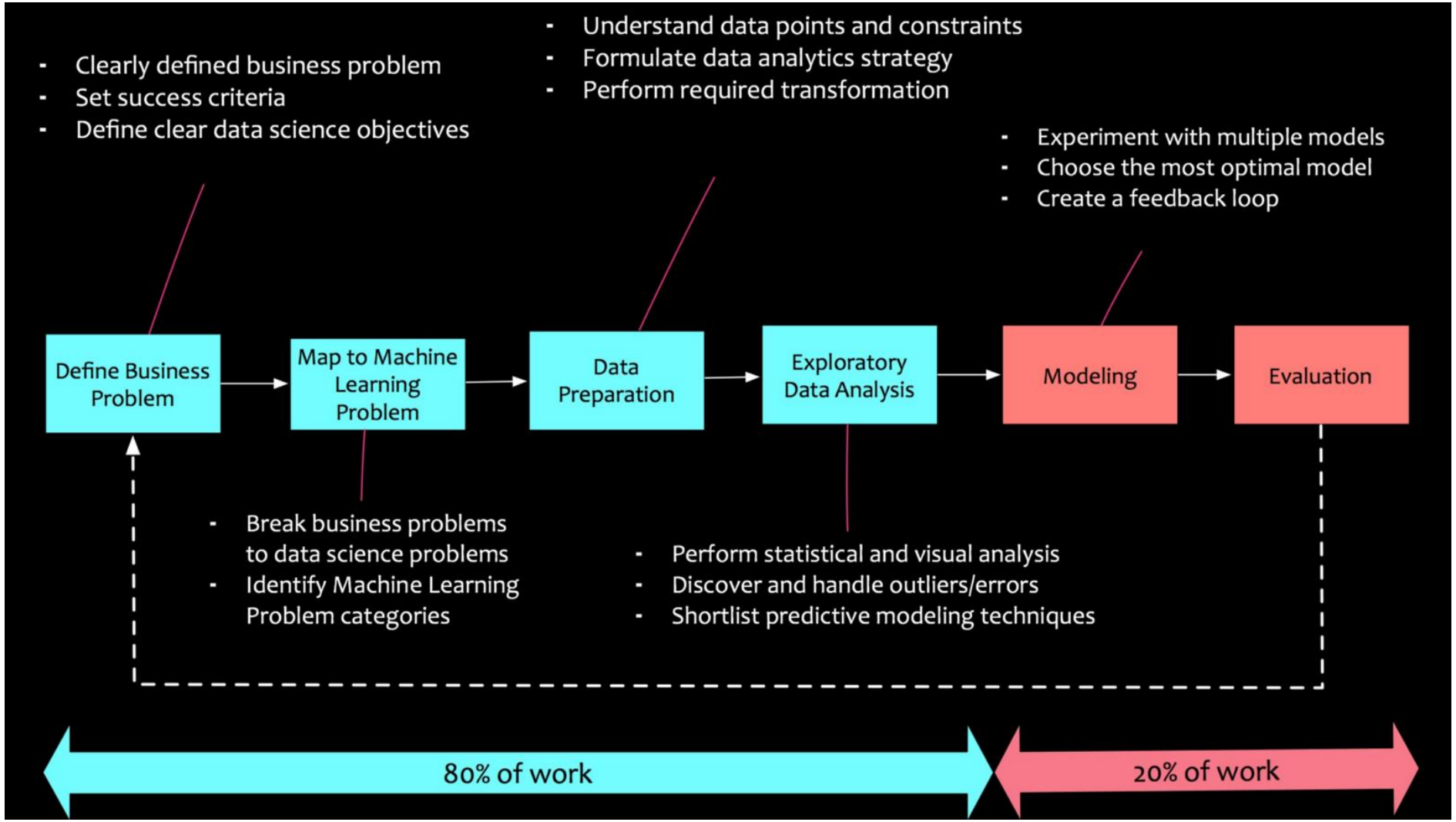


What is Data Science?

- Data Science is a field that comprises of everything that related to data cleansing, preparation, and analysis.
- Data Science is the combination of statistics, mathematics, programming, problem-solving, capturing data in ingenious ways…
- In simple terms, it is the umbrella of techniques used when trying to extract insights and information from data.

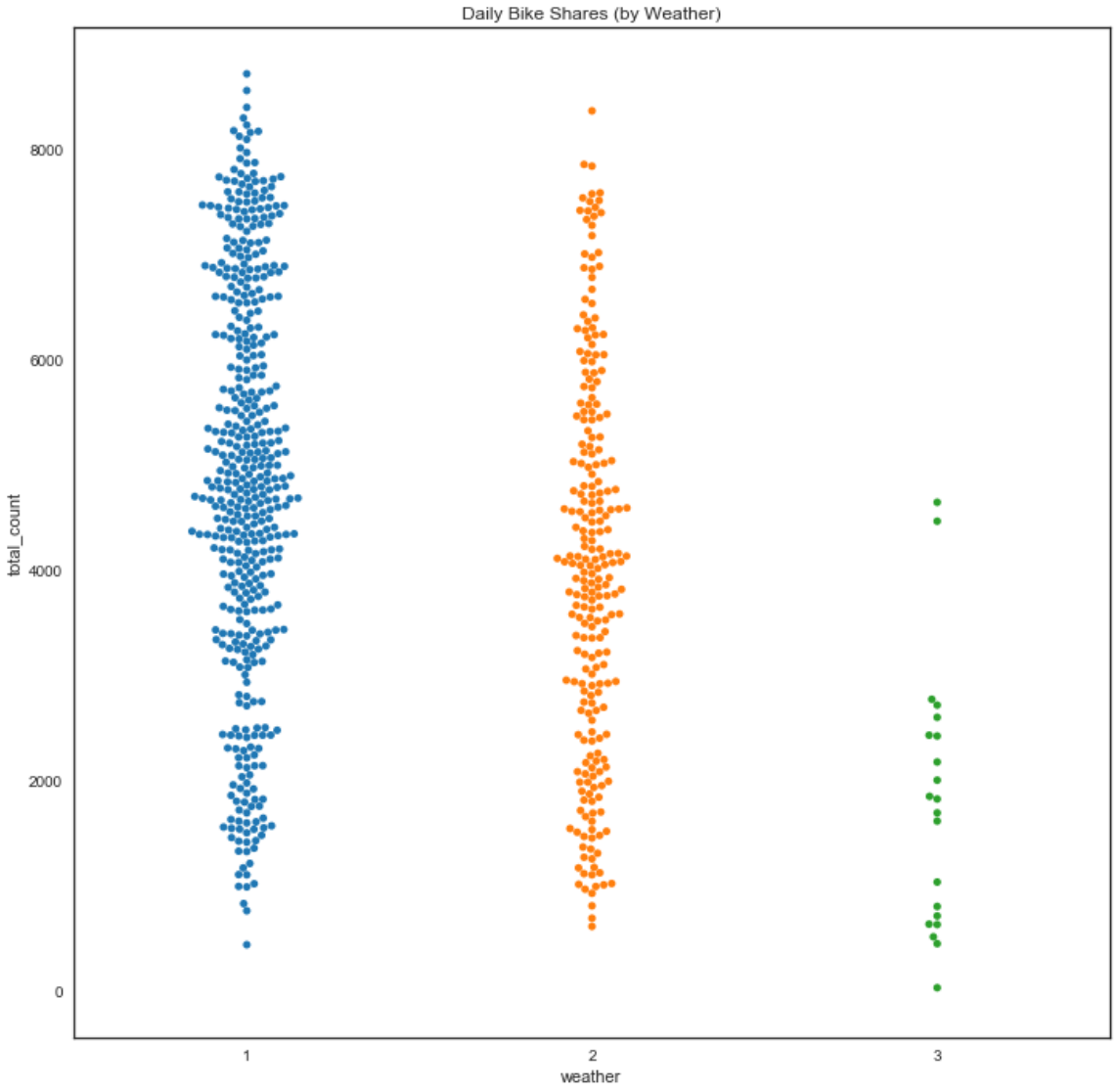


Data Science Process



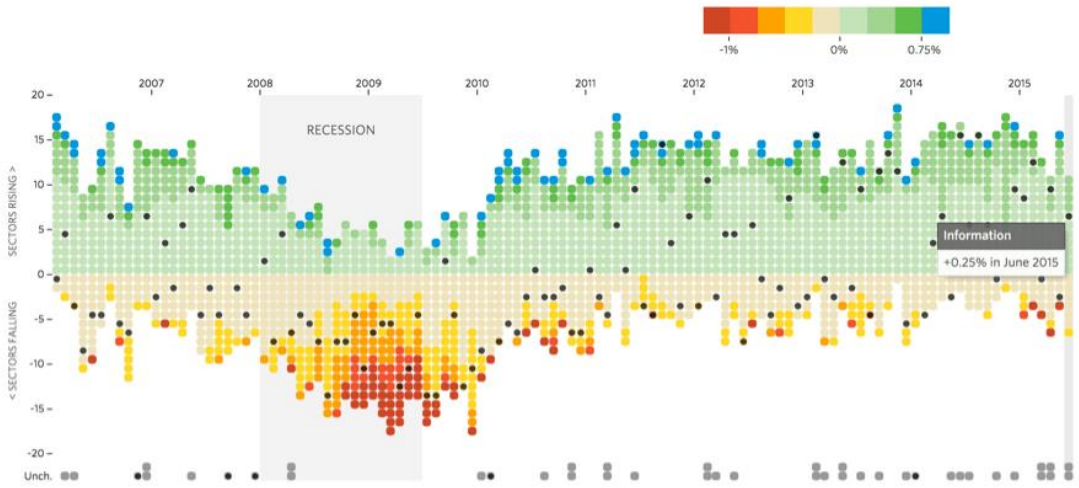
Source: https://www.datasciencecentral.com/profiles/blogs/data-science-simplified-principles-and-process

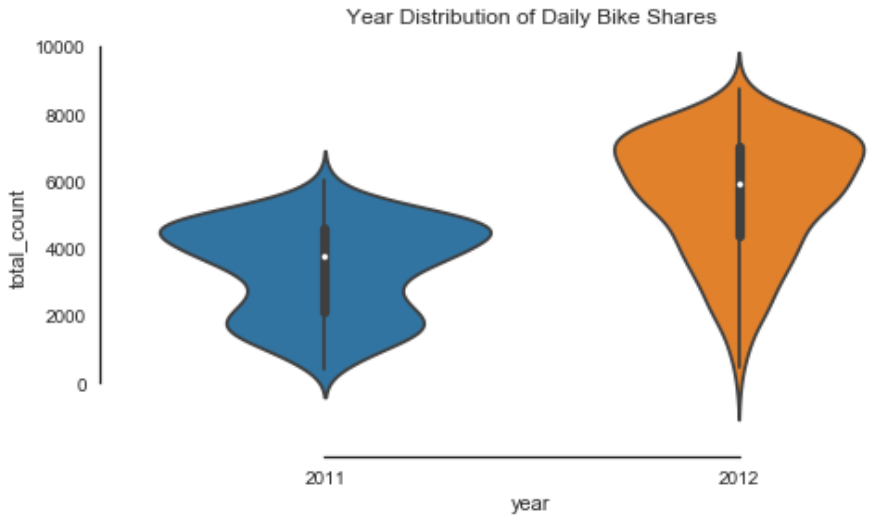
Data Visualisation



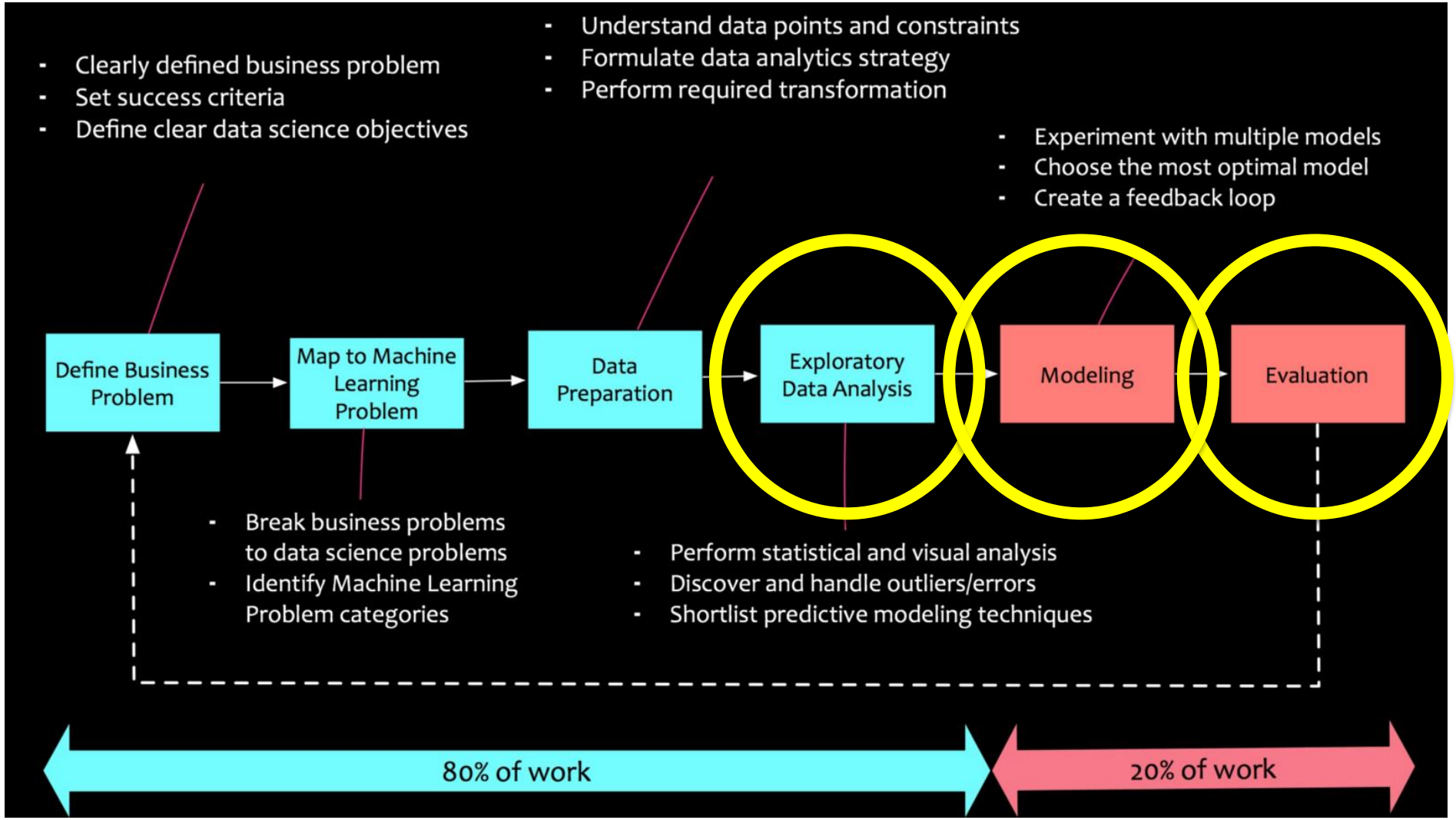
Why Data Visualisation?

- **A picture is worth a thousand words** – especially when you're trying to find relationships and understand your data
- **Informative** – everyone is familiar with visuals like pie charts, bar graphs etc.
- A well designed visual is **efficient and engaging** in communicating the message concisely and without ambiguity





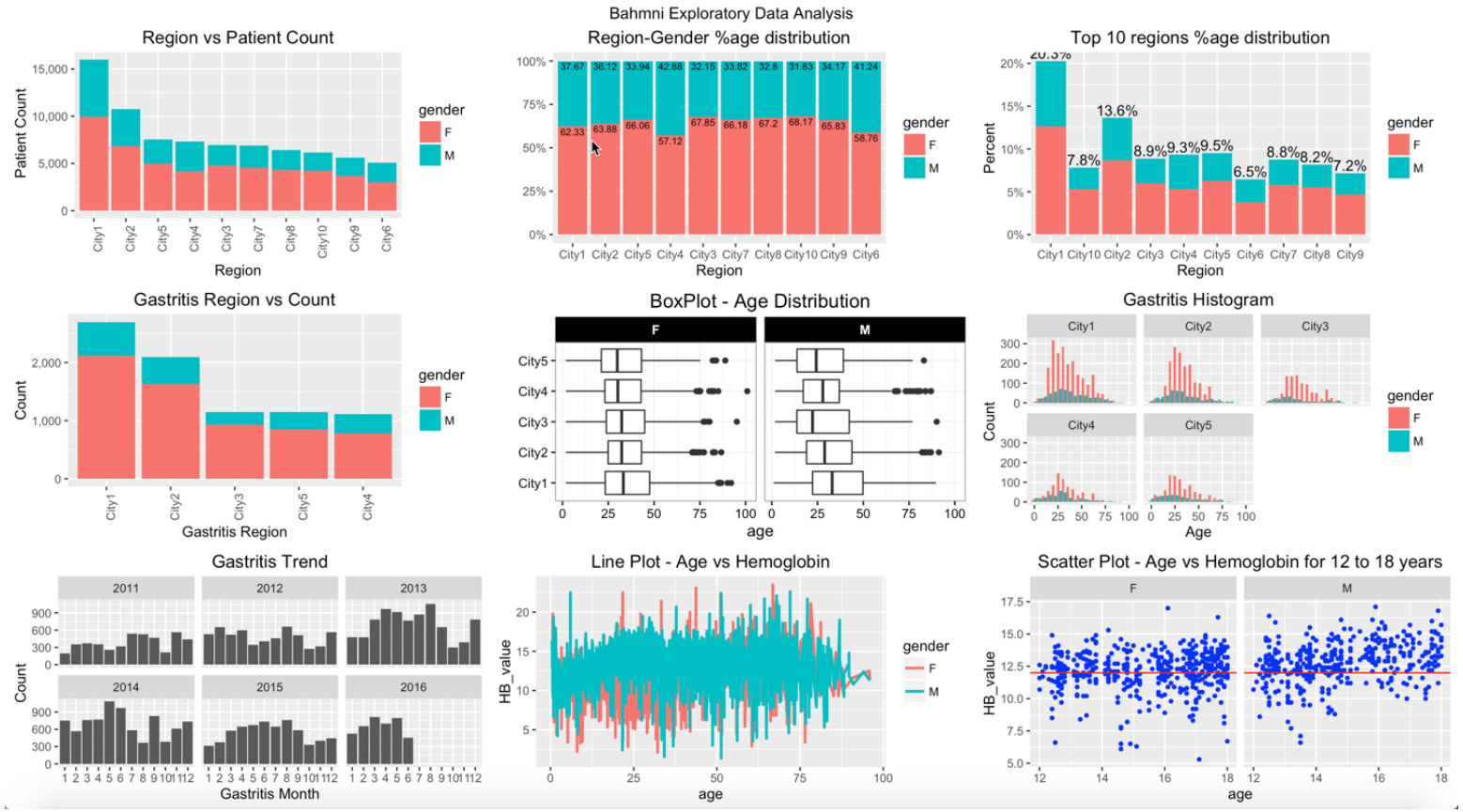
Where is Visualisation in the Data Science Process?



Source: https://www.datasciencecentral.com/profiles/blogs/data-science-simplified-principles-and-process

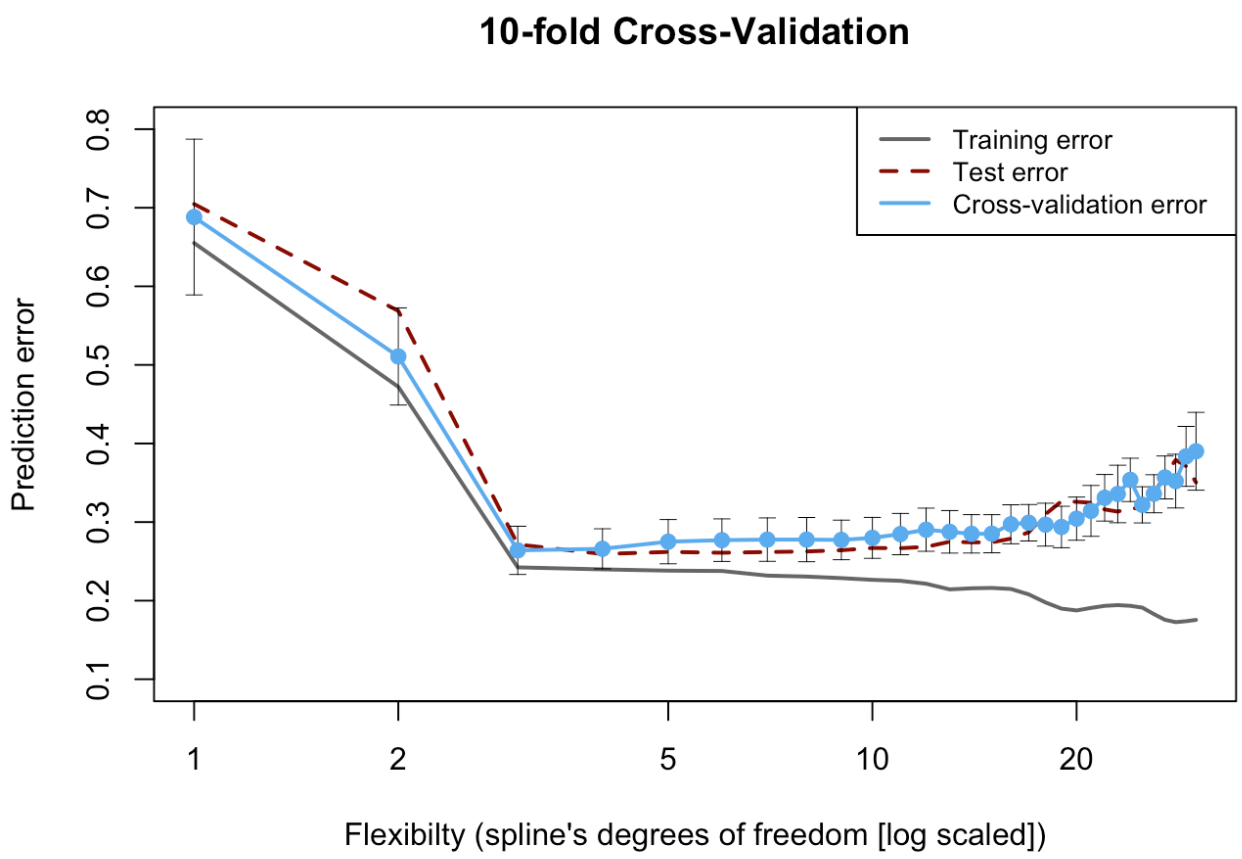
Visualisation for EDA

- You can use data visualisation on exploratory data analysis (getting to know your data)



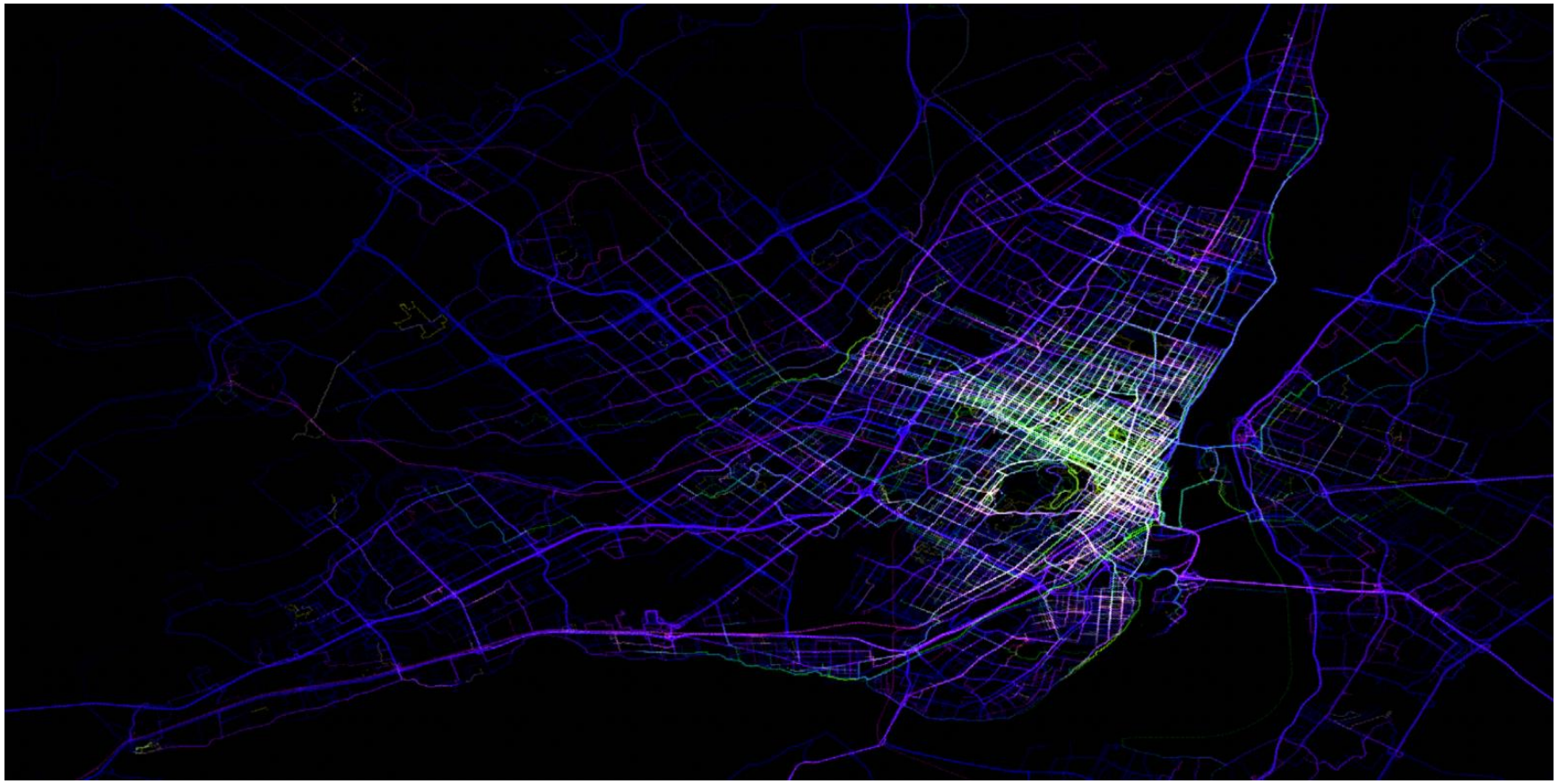
Visualisation for Modelling

- You can use data visualisation to get a visual intuition on how well your algorithm is doing e.g. plotting the predictions against real responses



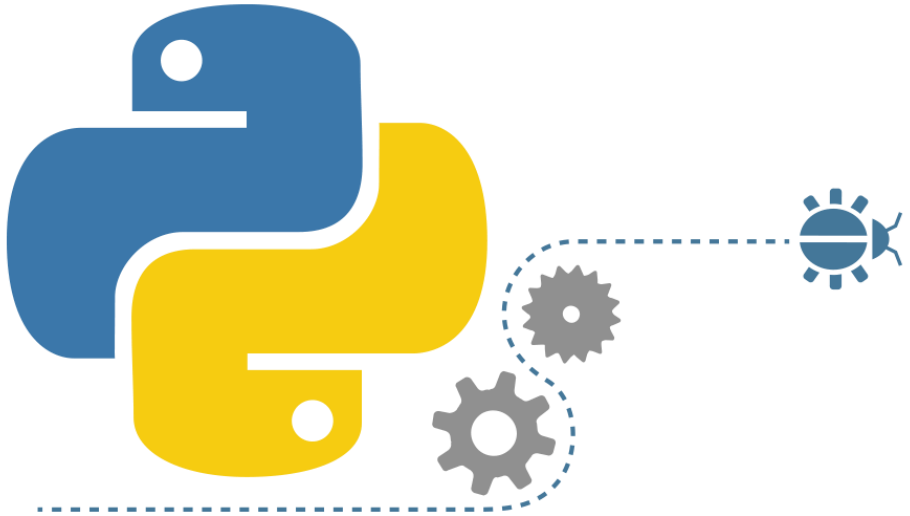
Visualisation for Communication

– You can use data visualisation for presentation/story-telling



Source: http://quorumetrix.blogspot.com/2017/12/geospatial-density-time-series-with.html

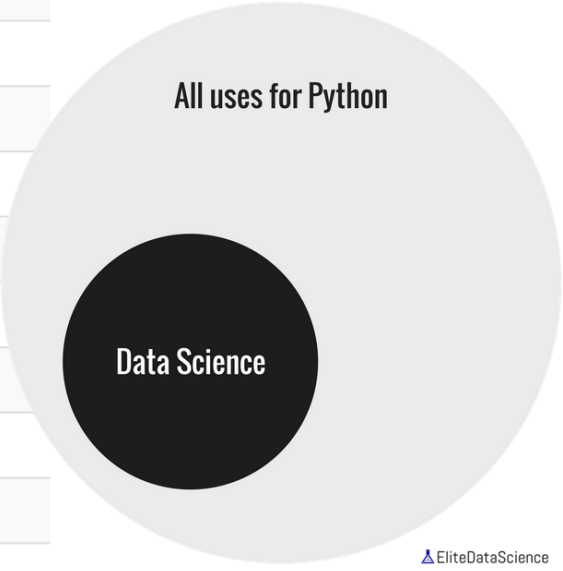
What is Python?



Python

– Python is a general purpose programming language

Aug 2018	Aug 2017	Change	Programming Language	Ratings	Change
1	1		Java	16.881%	+3.92%
2	2		C	14.966%	+8.49%
3	3		C++	7.471%	+1.92%
4	5	^	Python	6.992%	+3.30%
5	6	^	Visual Basic .NET	4.762%	+2.19%
6	4	v	C#	3.541%	-0.65%
7	7		PHP	2.925%	+0.63%
8	8		JavaScript	2.411%	+0.31%
9	-	^^	SQL	2.316%	+2.32%
10	14	^^	Assembly language	1.409%	-0.40%



Python – Simplicity

- Simplicity is Python's greatest strengths
- Python:

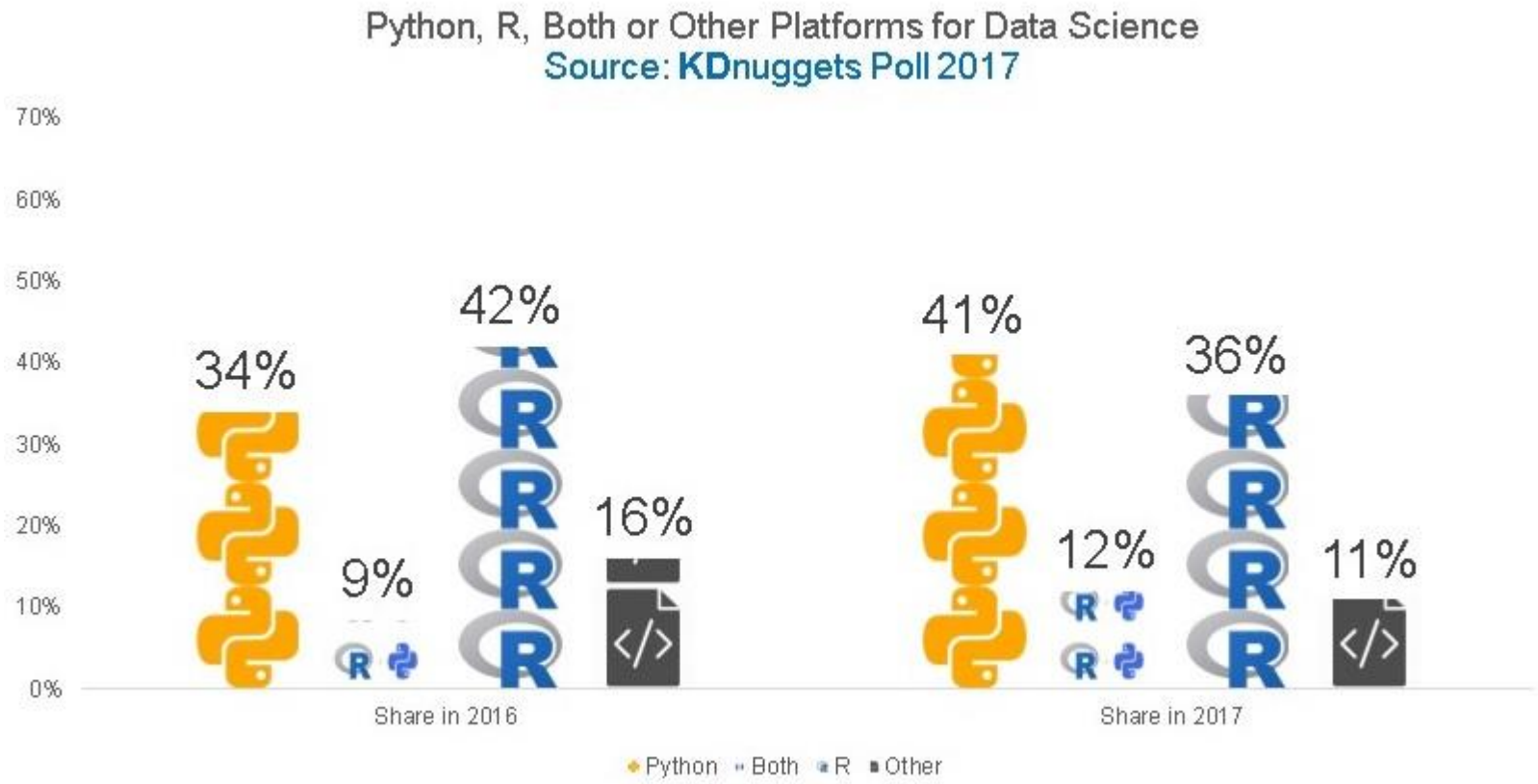
```python
print( "hello, world!" )
```

- Java:

```java
public class Main {
    public static void main(String[] args) {
        System.out.println("hello, world!");
    }
}
```

Python – for Data Analysis

- Popular along with the programming language R
- Easy to learn compared languages such as Java, C, C++ etc.
- Backed by a massive data science community
  - Googling any question about Python will definitely yield an answer you are looking for



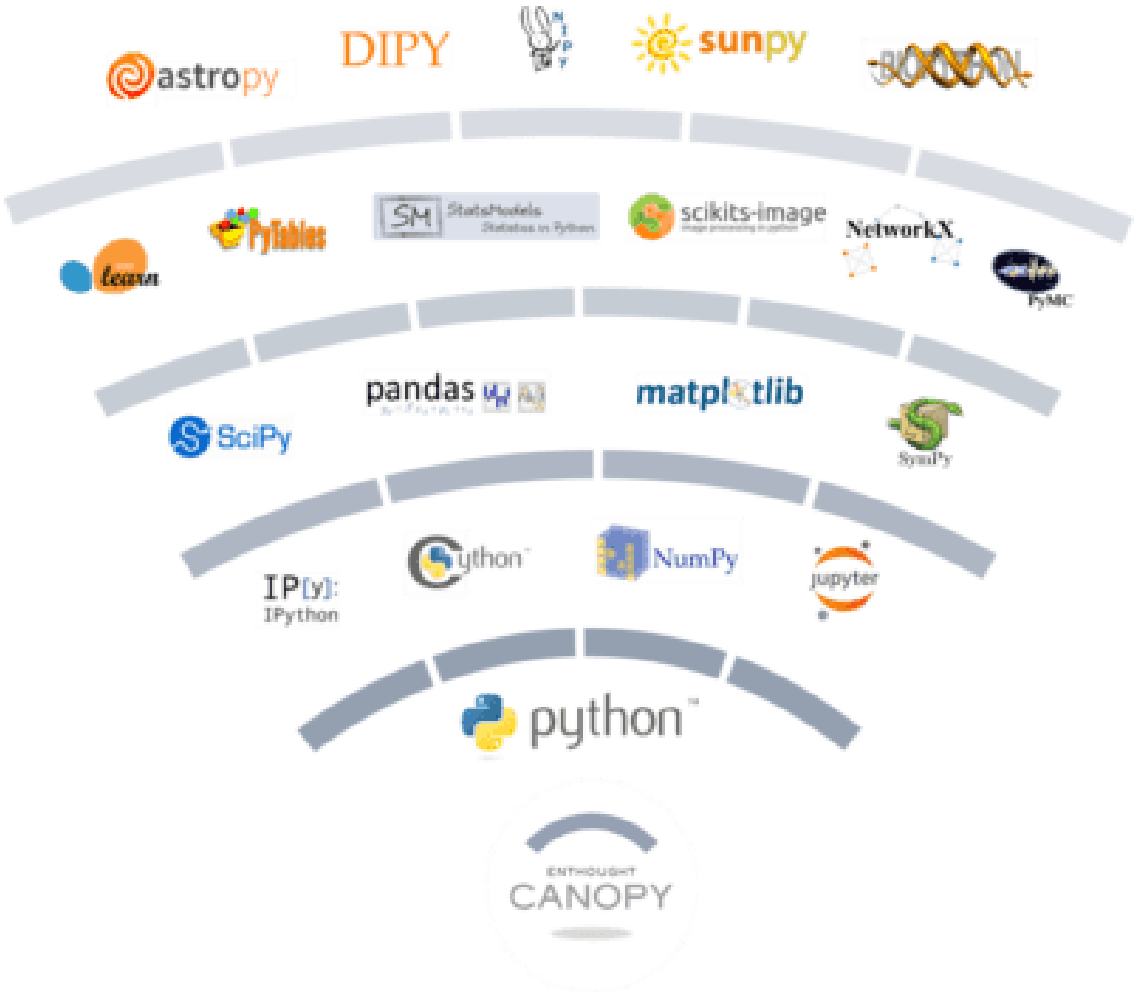
Use Cases in CBA's Environment

Q&A Session with CBA

Ask your burning questions on data science/analytics and programming languages e.g. Python, R!

Food Break / Networking

Python Libraries for Data Science



Python Libraries

– **Python library** is a collection of functions and methods that allows you to perform lots of actions without writing your own code. Some popular ones for data science:

SciPy

NumPy

matplotlib

IP[y]: IPython
Interactive Computing

pandas

scikit learn

Popular Python Libraries for Data Analysis

- **NumPy**
  - Core library for scientific computing in Python.
  - Provides a high-performance large multi-dimensional arrays, matrices and tools for working with these arrays.

- **Pandas**
  - Builds on top of NumPy
  - Uses “dataframes” to store data – similar to Excel
  - Used for data manipulation and analysis.
  - In particular, it offers data structures and operations for manipulating numerical tables and time series.

Pandas

- **Pandas** dataframe vs **Excel** (same dataset)
  - We will be using this dataset later today!

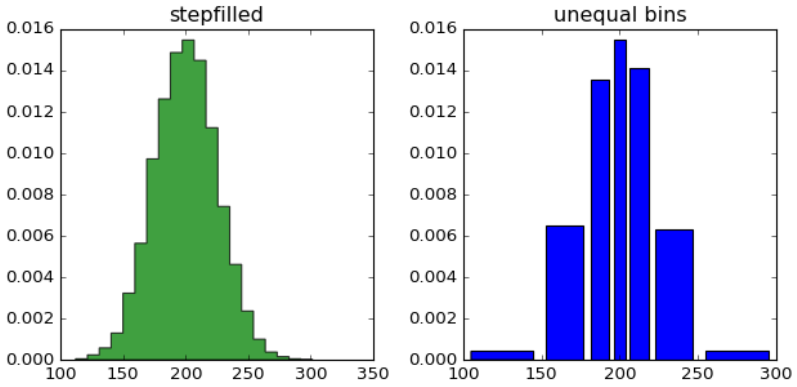
```
data.head(20)
```

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp
0	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625
1	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739
2	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405
3	4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122
4	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270
5	6	2011-01-06	1	0	1	0	4	1	1	0.204348	0.233209
6	7	2011-01-07	1	0	1	0	5	1	2	0.196522	0.208839
7	8	2011-01-08	1	0	1	0	6	0	2	0.165000	0.162254
8	9	2011-01-09	1	0	1	0	0	0	1	0.138333	0.116175
9	10	2011-01-10	1	0	1	0	1	1	1	0.150833	0.150888
10	11	2011-01-11	1	0	1	0	2	1	2	0.169091	0.191464
11	12	2011-01-12	1	0	1	0	3	1	1	0.172727	0.160473
12	13	2011-01-13	1	0	1	0	4	1	1	0.165000	0.150883
13	14	2011-01-14	1	0	1	0	5	1	1	0.160870	0.188413
14	15	2011-01-15	1	0	1	0	6	0	2	0.233333	0.248112
15	16	2011-01-16	1	0	1	0	0	0	1	0.231667	0.234217
16	17	2011-01-17	1	0	1	1	1	0	2	0.175833	0.176771
17	18	2011-01-18	1	0	1	0	2	1	2	0.216667	0.232333
18	19	2011-01-19	1	0	1	0	3	1	2	0.292174	0.298422
19	20	2011-01-20	1	0	1	0	4	1	2	0.261667	0.255050

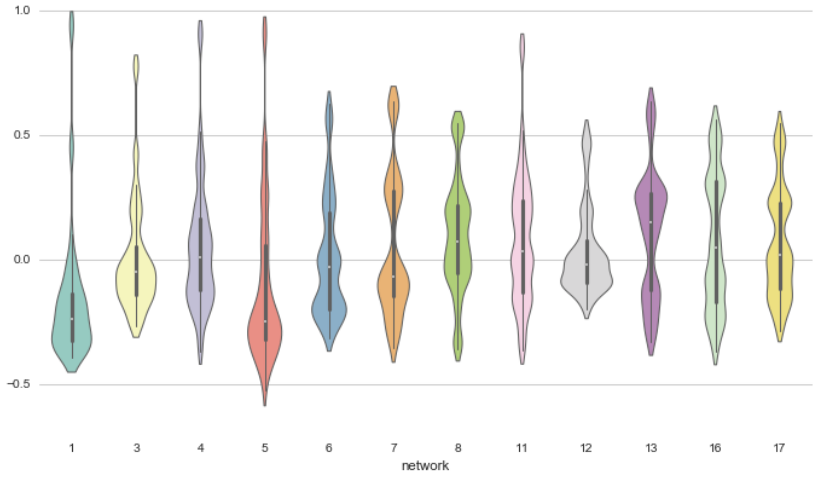
	A	B	C	D	E	F	G	H	I	J	K
1	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp
2	1	1/01/2011	1	0	1	0	6	0	2	0.344167	0.363625
3	2	2/01/2011	1	0	1	0	0	0	2	0.363478	0.353739
4	3	3/01/2011	1	0	1	0	1	1	1	0.196364	0.189405
5	4	4/01/2011	1	0	1	0	2	1	1	0.2	0.212122
6	5	5/01/2011	1	0	1	0	3	1	1	0.226957	0.22927
7	6	6/01/2011	1	0	1	0	4	1	1	0.204348	0.233209
8	7	7/01/2011	1	0	1	0	5	1	2	0.196522	0.208839
9	8	8/01/2011	1	0	1	0	6	0	2	0.165	0.162254
10	9	9/01/2011	1	0	1	0	0	0	1	0.138333	0.116175
11	10	10/01/2011	1	0	1	0	1	1	1	0.150833	0.150888
12	11	11/01/2011	1	0	1	0	2	1	2	0.169091	0.191464
13	12	12/01/2011	1	0	1	0	3	1	1	0.172727	0.160473
14	13	13/01/2011	1	0	1	0	4	1	1	0.165	0.150883
15	14	14/01/2011	1	0	1	0	5	1	1	0.16087	0.188413
16	15	15/01/2011	1	0	1	0	6	0	2	0.233333	0.248112
17	16	16/01/2011	1	0	1	0	0	0	1	0.231667	0.234217
18	17	17/01/2011	1	0	1	1	1	0	2	0.175833	0.176771
19	18	18/01/2011	1	0	1	0	2	1	2	0.216667	0.232333
20	19	19/01/2011	1	0	1	0	3	1	2	0.292174	0.298422

Popular Python Libraries for Data Visualisation

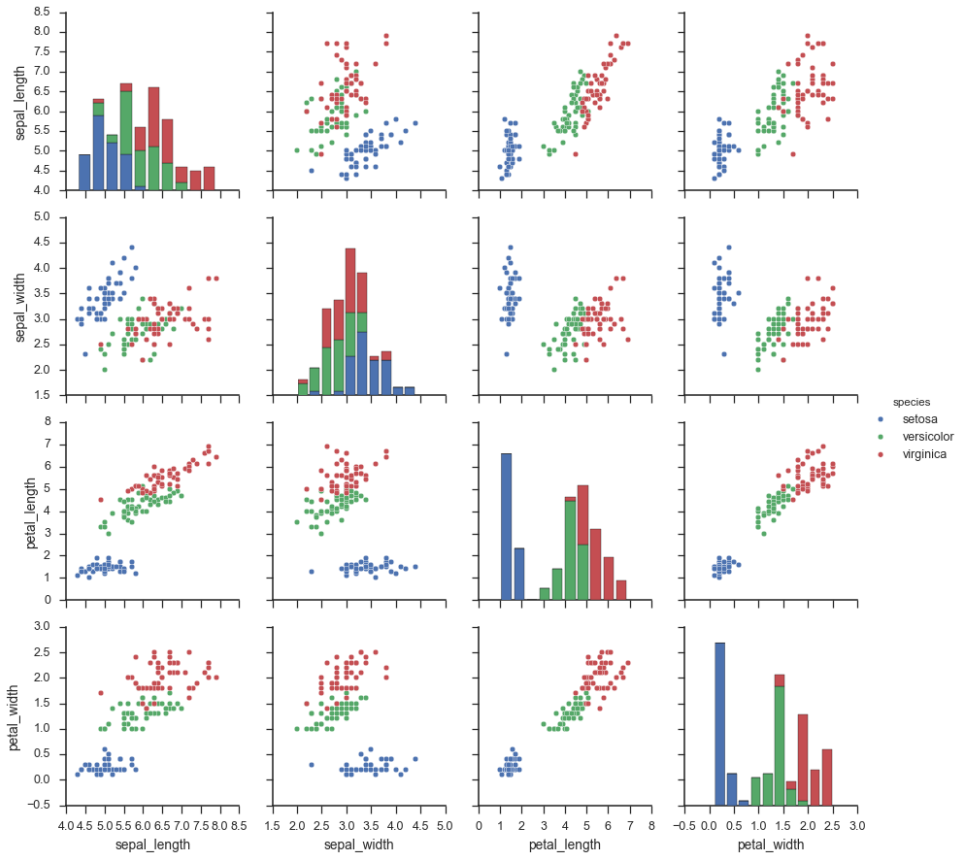
- **Matplotlib**
  - Most widely used library for plotting in Python
  - But looks outdated



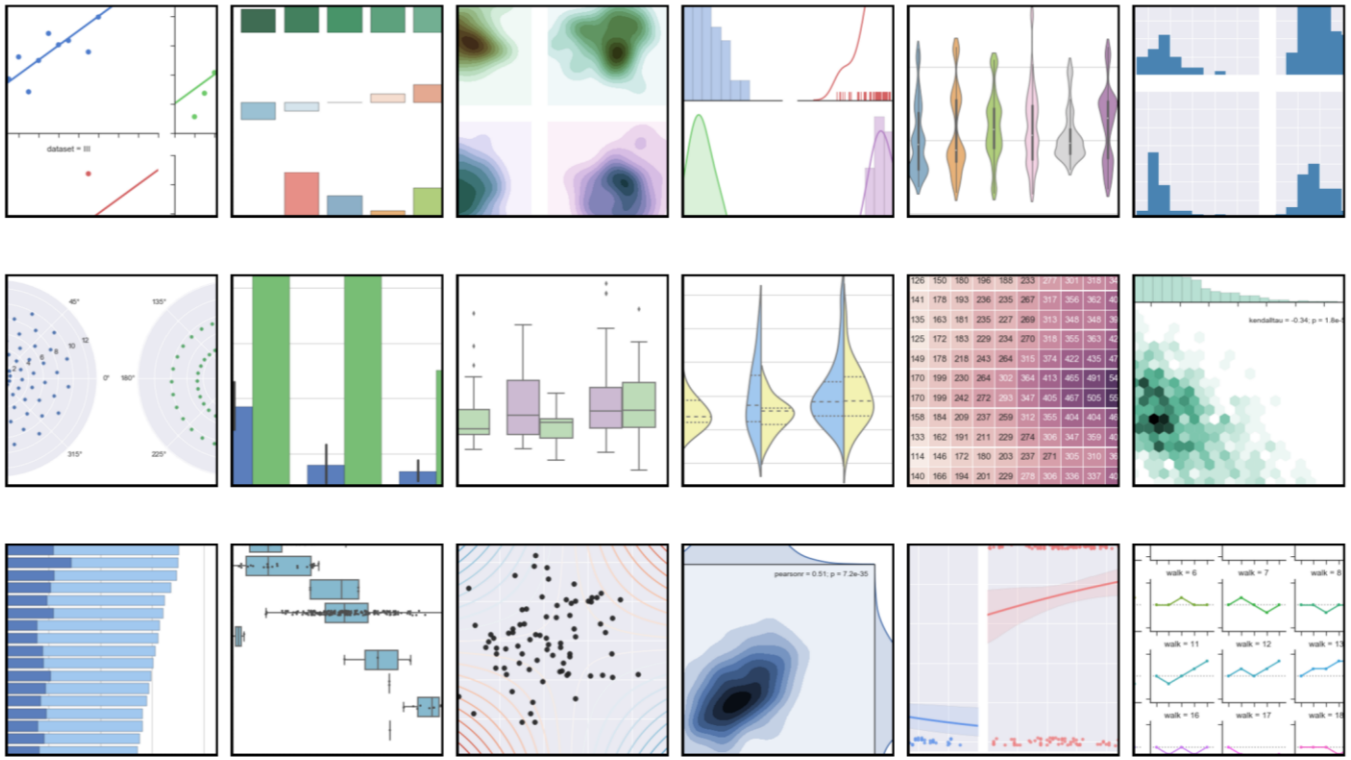
- **Seaborn**
  - Built on top of matplotlib, but much more beautiful and aesthetically pleasing
  - Therefore you will need to know matplotlib commands
  - Easier customisations with styles and colour palettes
  - More suitable for visualising regression, classification etc.



Seaborn

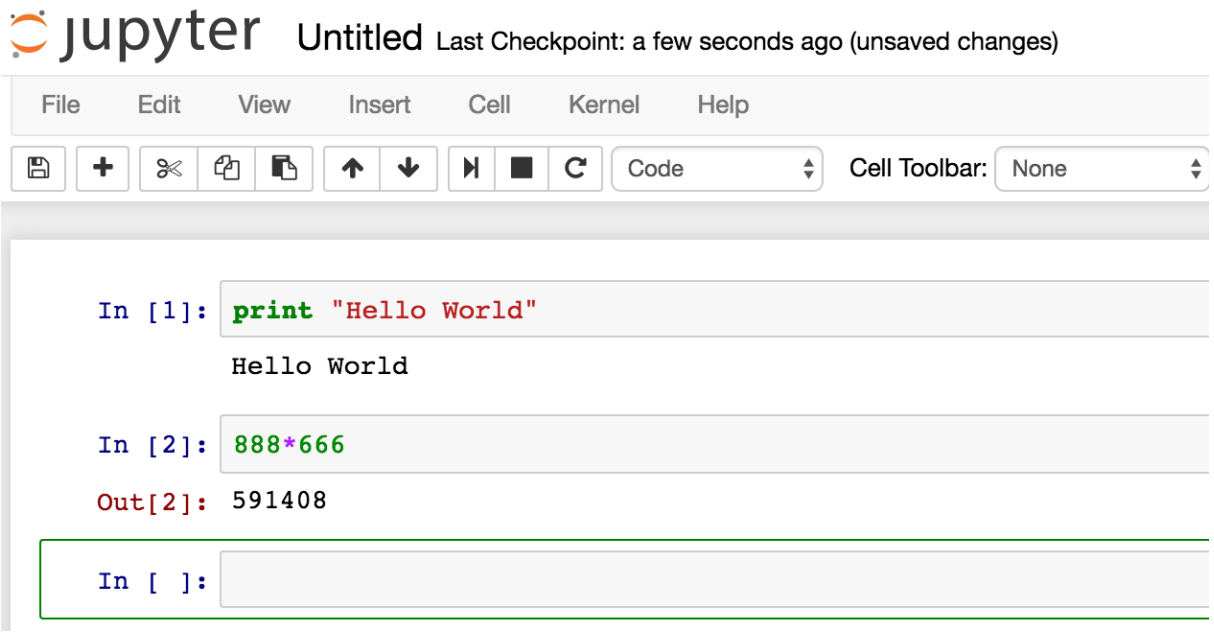


Example gallery



Coding Activity 1 – Python Fundamentals

Lets use Jupyter Notebook



Jupyter Notebook

- Jupyter Notebook is a popular machine learning environment (with Python).
- It is browser based interface and is simple to use.

jupyter activity_2_solution Last Checkpoint: a few seconds ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

Getting to know your dataset

Below, it is clear that Pandas dataframe resembles Excel worksheet, except you cannot edit them (you would need to do this via coding).

We call `.head()` to show the first 5 rows. You can specify the number of rows to display by inputting a parameter e.g. `.head(12)`.

In [31]:
```
data.head(5)
```

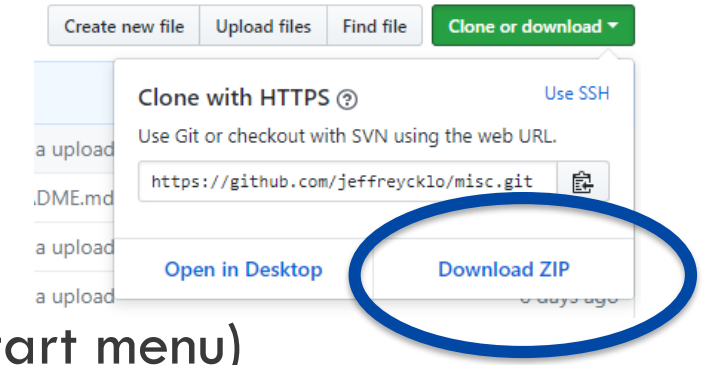
Out[31]:

	date	season	year	month	holiday	weekday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	total_count
0	2011-01-01	1	2011	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
1	2011-01-02	1	2011	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	2011-01-03	1	2011	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	2011-01-04	1	2011	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	2011-01-05	1	2011	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600

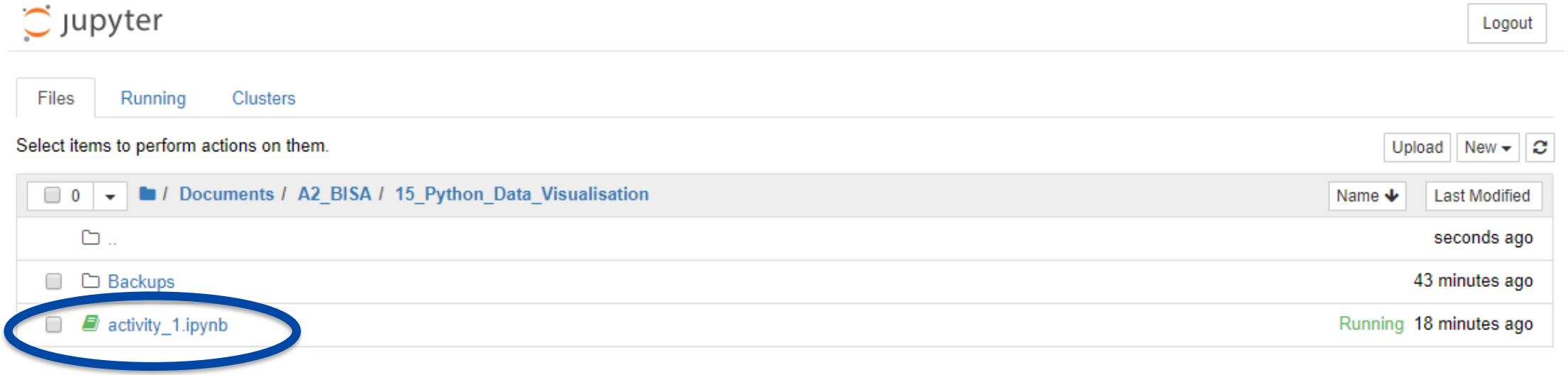
Similarly, `.tail()` will show the last 5 rows.

In [9]:
```
data.tail(5)
```

Coding Activity 1

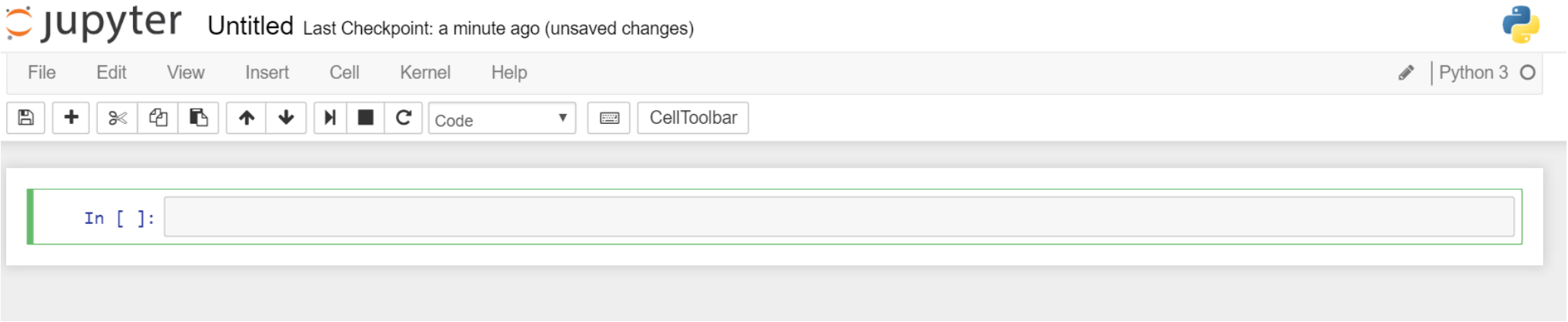


1. Open this link: **www.tinyurl.com/BISApython**
2. Download the ZIP folder from the above link
3. Then unzip the folder
4. Open Jupyter Notebook on your PC (search "Jupyter" in start menu)
   - Once Jupyter opens up, you will see a screen that looks like the one below.
   - In the main body you will see all files in the directory from which you started Jupyter.
5. On Jupyter, navigate to your downloaded folder and open ***activity_1.ipynb***



Jupyter Notebook

- **Lets do a quick demo of the features of Jupyter Notebook!**
- The basic elements of Jupyter Notebook are the cells, as in the next figure.
- Each cell holds is interpreted as code by default. You can type (or copy and paste) as much code as you like in a single cell (use enter for a new line).
- When you want **to run the code, press Shift + Enter**.
- **Exercise** – run the cell as a calculator (e.g. 1 + 1) and run the code as above.



Short break – 5 mins

Feel free to do the exercises in the meantime, or just have a short break ☺

Coding Activity 2 – Data Visualisation

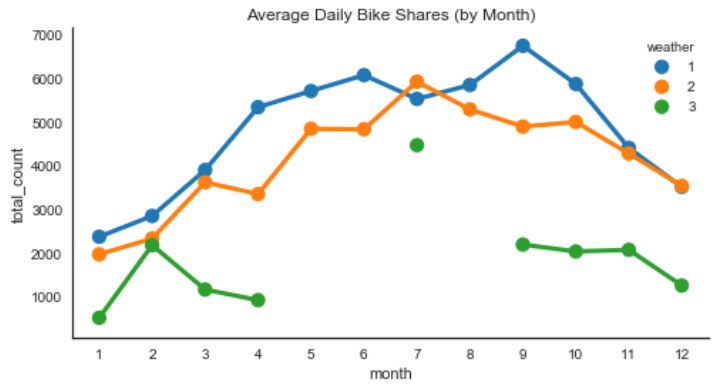
On a real dataset

jupyter activity_2_solution Last Checkpoint: 11 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

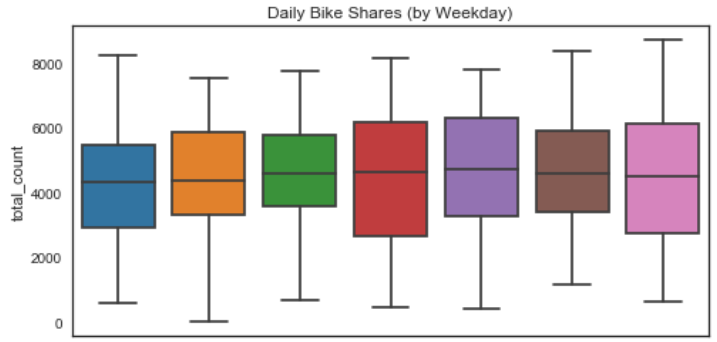
Run Markdown

In [20]:
```
fig = plt.figure(figsize=(8,4))
sns.pointplot(data=data, x='month', y='total_count', hue='weather', ci=None)
sns.despine()
plt.title("Average Daily Bike Shares (by Month)")
plt.show()
```



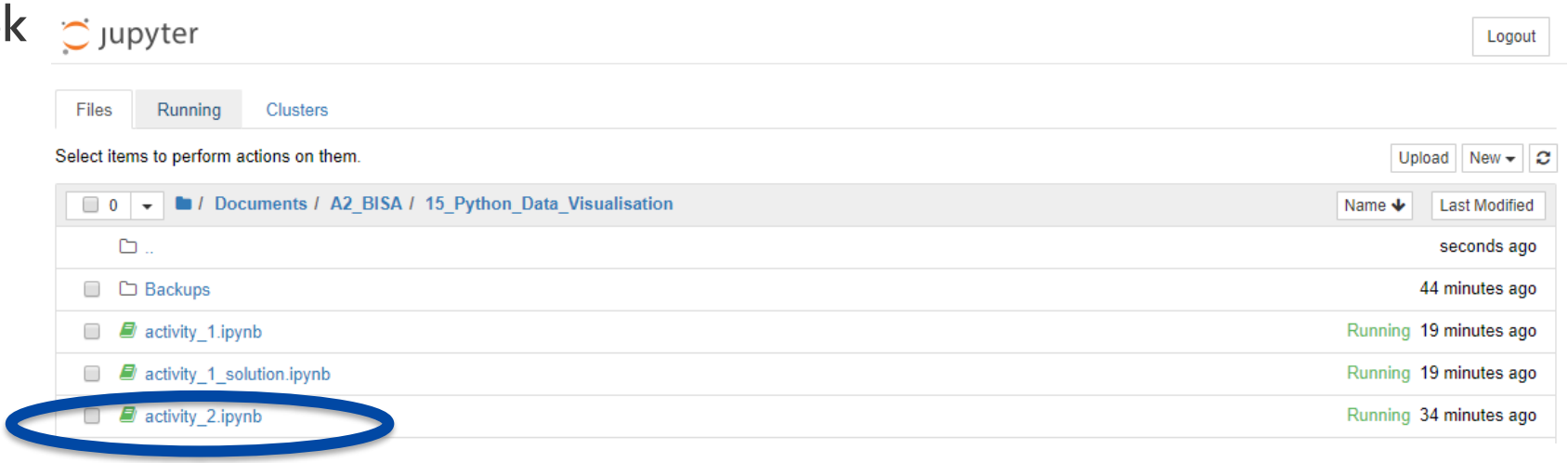
Distribution of Daily Bike Shares (by Weekday)

In [21]:
```
fig = plt.figure(figsize=(8,4))
sns.boxplot(data=data, x='weekday', y='total_count')
plt.title("Daily Bike Shares (by Weekday)")
plt.show()
```



Coding Activity 2

- From the previous downloaded folder, open ***activity_2.ipynb*** from your Jupyter Notebook



- We will be using a small dataset – Bike Sharing Dataset from Kaggle

```
In [3]: data.head()
```

Out[3]:

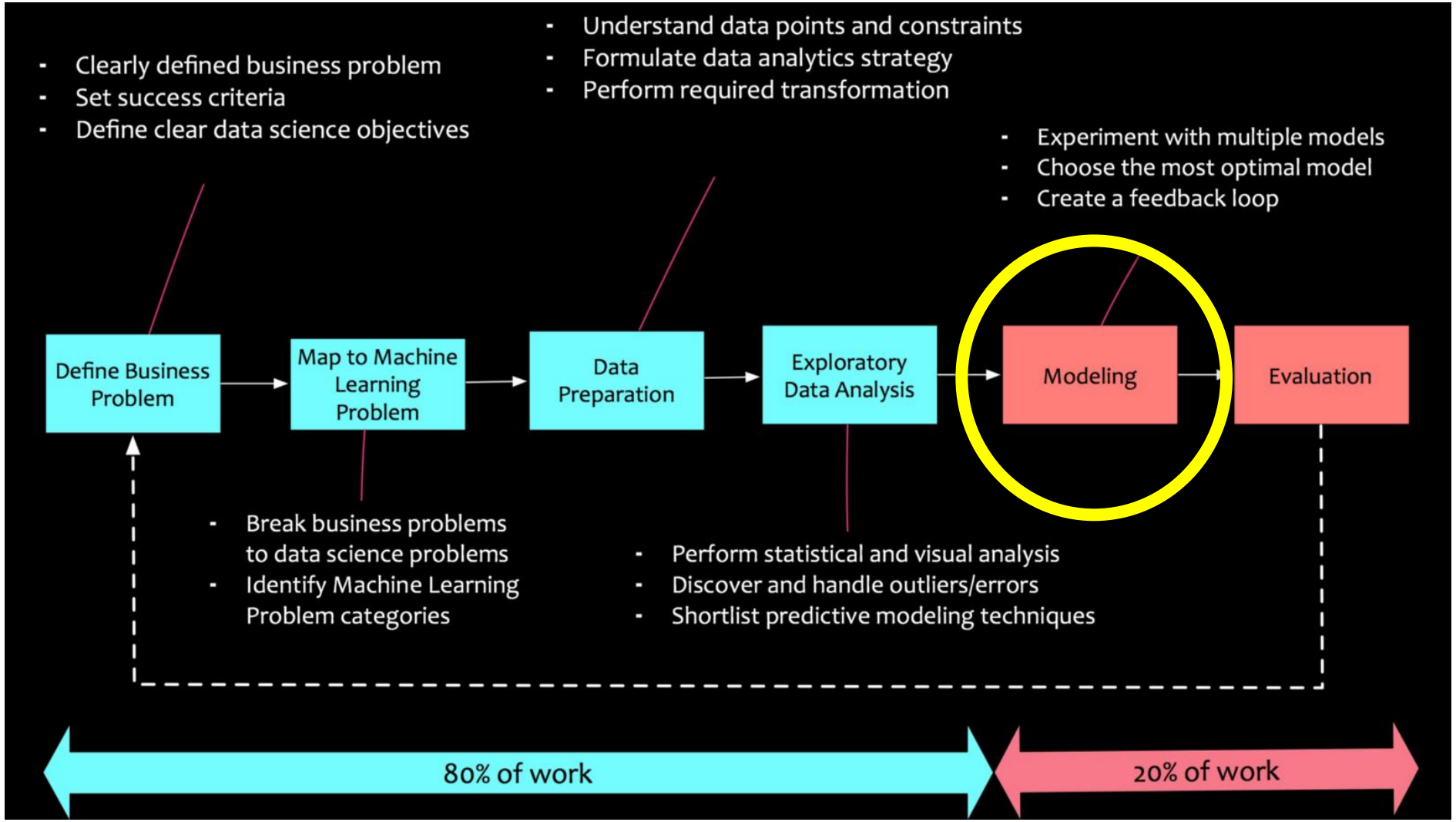
	date	season	year	month	holiday	weekday	weather	temperature	humidity	windspeed	casual	registered	tota
0	1/01/2011	Spring	2011	1	0	Sun	Cloudy	14.11	80.58	10.75	331	654	985
1	2/01/2011	Spring	2011	1	0	Mon	Cloudy	14.90	69.61	16.65	131	670	801
2	3/01/2011	Spring	2011	1	0	Tues	Clear	8.05	43.73	16.64	120	1229	1349
3	4/01/2011	Spring	2011	1	0	Wed	Clear	8.20	59.04	10.74	108	1454	1562
4	5/01/2011	Spring	2011	1	0	Thurs	Clear	9.31	43.70	12.52	82	1518	1600

Key Takeaways from the Activities

- It is important to learn the fundamentals
- Practice with small datasets
- Using Seaborn is not as hard as you think (at least to get started)
  - To get a completely different plot with Seaborn, oftentimes it is possible to just replace one word e.g. replace “sns.barplot” with “sns.boxplot” to get a completely different plot
- Much of the data visualisation we learnt today is mostly applicable in the exploratory data analysis (EDA). You need to know how the data is like before you start moulding, transforming the features, and before you can feed the data into predictive algorithms.

Extra Resources

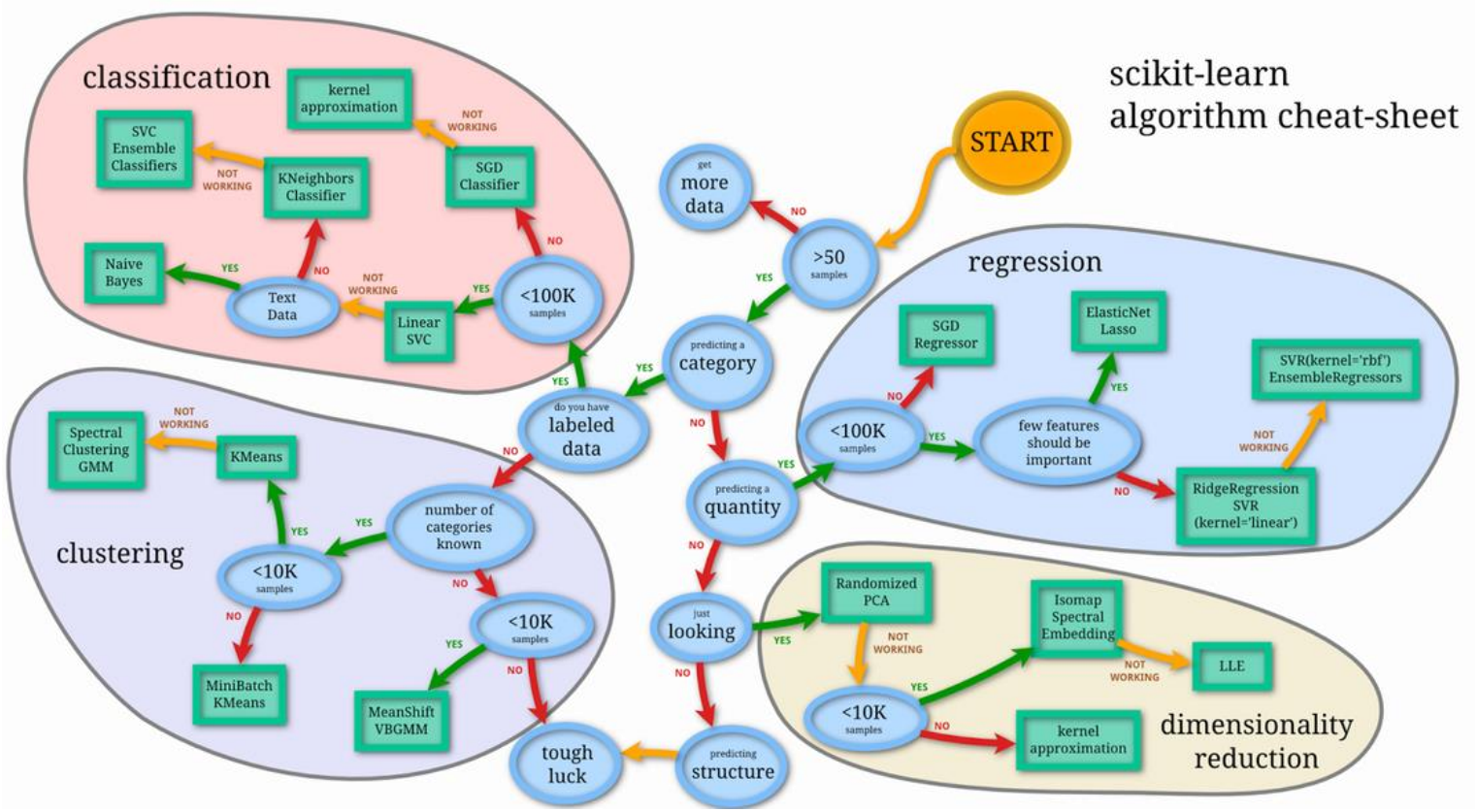
Where does Machine Learning Models sit?



Source: https://www.datasciencecentral.com/profiles/blogs/data-science-simplified-principles-and-process

Popular Python Library for Machine Learning

- **Scikit-Learn**
  - Machine learning library with various classification, **regression** and clustering algorithms including support vector machines, random forests, gradient boosting etc.

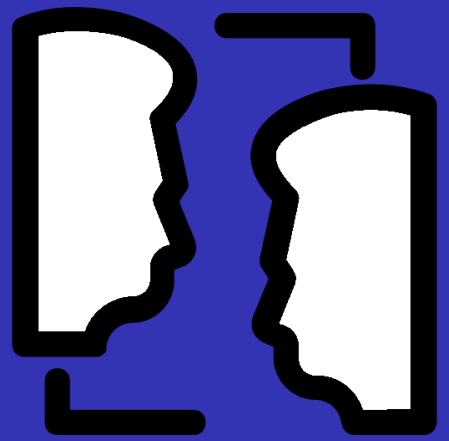


Extra Resources for Learning Python and Data Science

Here are some:

- Dataquest.io
- Seaborn Gallery
- Data Science Guides by EliteDataScience
- Python Cheat Sheet for Data Science
- https://automatetheboringstuff.com/
- Data Science Learning Plan

Thanks for coming!
Feel free to chat to our CBA / BISA reps!



If you have any questions for BISA or CBA, feel free to stick around!

facebook.com/ **BisaSydney**

BISA

References

- https://blog.modeanalytics.com/python-data-visualization-libraries/
- https://i2.wp.com/analyticsweek.com/wp-content/uploads/2017/11/DataScienceWorkflow.jpg
- https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport.pdf
- https://data-visualization.cioreview.com/cxoinsight/what-is-data-visualization-and-why-is-it-important-nid-11806-cid-163.html
- https://i1.wp.com/generalassemb.ly/blog/wp-content/uploads/2015/08/Track-National-Unemployment-Job-Gains-and-Job-Losses-%E2%80%93-Wall-Street-Journal-.png?ssl=1
- https://au.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article
- https://cdn-images-1.medium.com/max/1200/1*mgXvzNcwfpnBawI6XTkVRg.png
- https://www.flaticon.com