NLP – PYTHON – ALTERYX

WENDY WONG

SENIOR ANALYTICS CONSULTANT AND DIGITAL ACCELERATOR AT PWC AUSTRALIA

AGENDA

Learning Objectives

- Define Python and compare it to other programming languages
- Discuss NLP and how it is used in different industries
- Define Alteryx, explain how it is used and why it is so popular today in the marketplace

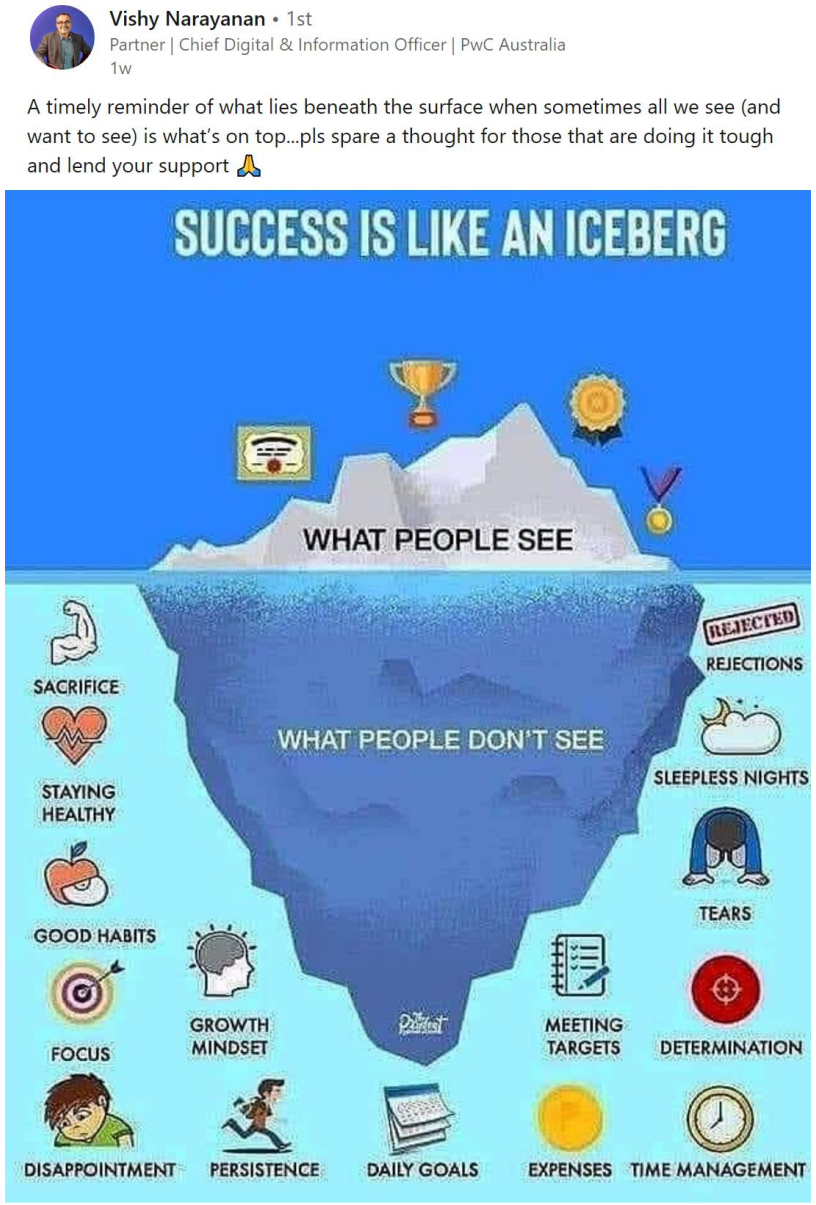
Disclaimer : The following slides do not reflect the opinions of my employer, but are my own personal opinions

ABOUT ME

ABOUT ME – FROM FINANCIAL PLANNING TO DATA SCIENCE

Tip of the Iceberg

LinkedIn post

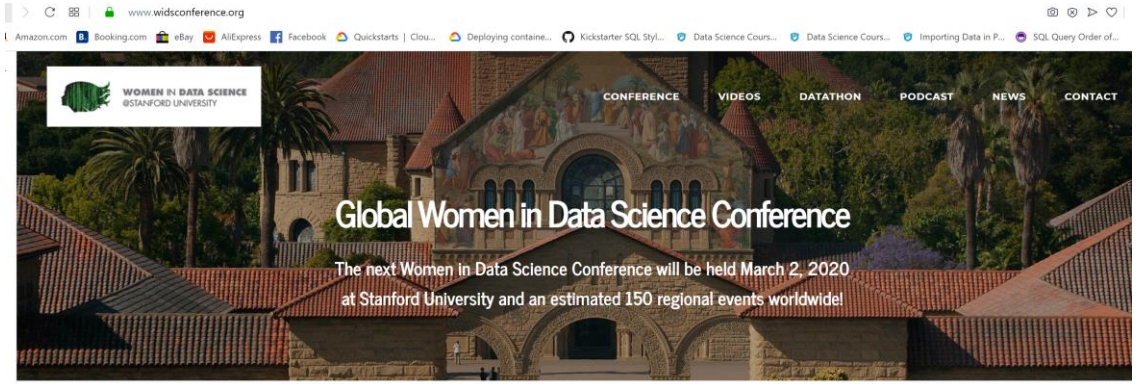


- You are making progress even when you do not feel you are.
- Grit
- GFC
- Reskill /Retrain
- Be uncomfortable

ABOUT ME – NEW LIFE IN DATA SCIENCE

iapa

GENERAL ASSEMBLY



+ableau SOFTWARE

UTS
UNIVERSITY OF TECHNOLOGY, SYDNEY

Towards Data Science

GovHack

About Women in Data Science

The Women in Data Science (WiDS) initiative aims to inspire and educate data scientists worldwide, regardless of gender, and to support women in the field.

WiDS started as a conference at Stanford in November 2015. Now, WiDS includes a global conference, with approximately 150+ regional events worldwide; a datathon, encouraging participants to hone their skills using a social impact challenge; and a podcast, featuring leaders in the field talking about their work, their journeys, and lessons learned.

WOMEN IN DATA SCIENCE
@STANFORD UNIVERSITY

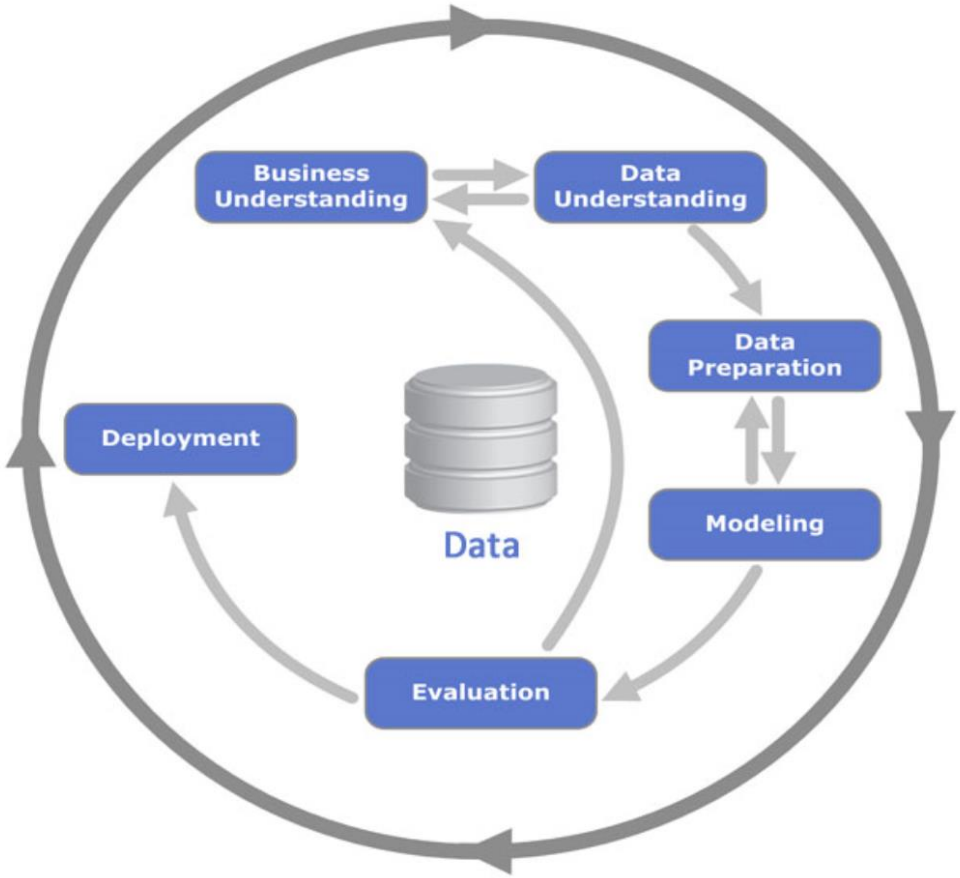
R-Ladies

DATA SCIENCE

WHAT IS CRISP DM?

Cross Industry Standard Process for Data Mining

CRISP-DM Process Diagram



- Data Science
- Data Analytics
- Consulting

Source: Kenneth Jensen

PYTHON

PYTHON

- Python is an open source language

What is Python?

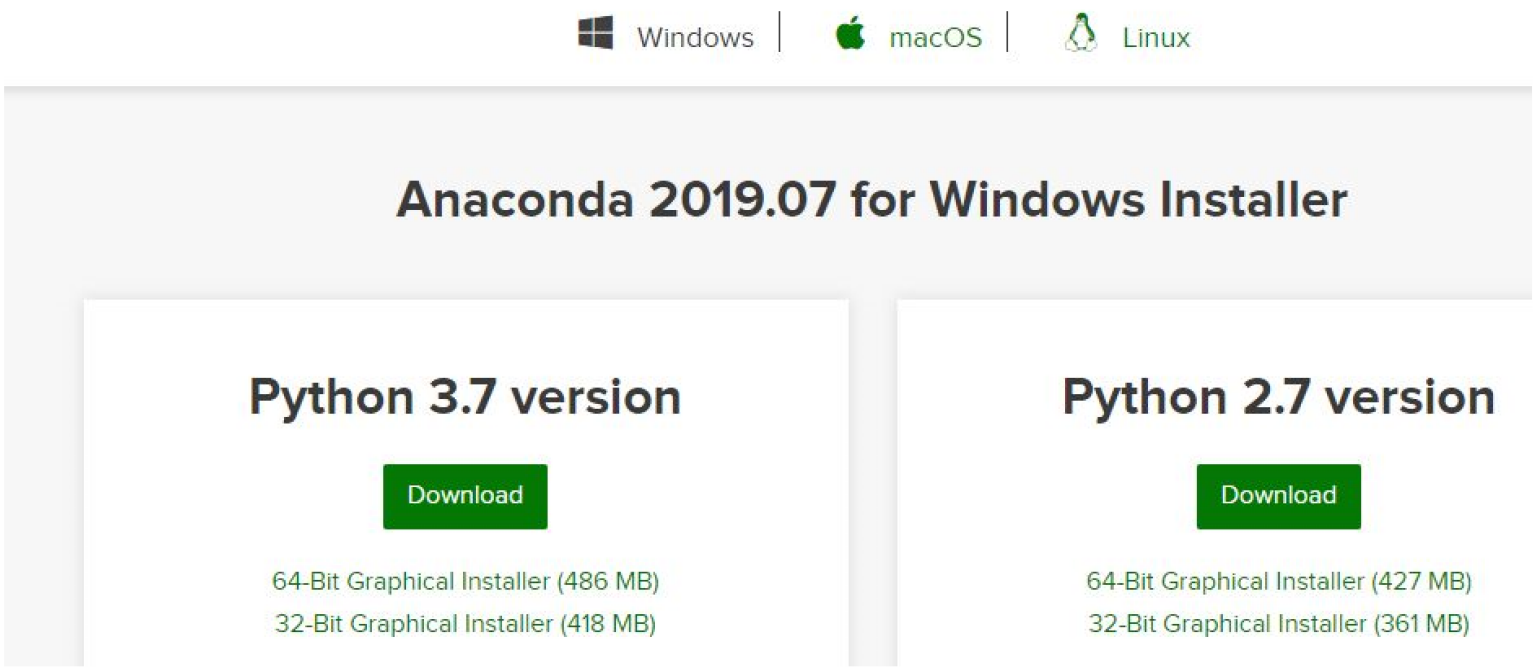
- Python is an open-source
- Python is free
- Python was created by Guido Van Rossum in 1991
- Object-oriented language can execute code that includes data attributes

Where is Python used today?	Machine Learning Algorithm
• **Recommendation Engines** e.g. Amazon, Netflix, Online shopping	Collaborative filtering
• **Virtual Assistants** e.g.Google Home, Alexa	Deep Learning (Recurrent Neural Networks) and Natural Language Processing
• **Commercial machine learning applications** e.g. customer segmentation by News Corp's readers in the cloud using AWS	K-means clustering
• **Rapid prototyping** e.g.Run an experiment which group of customers will be shown a new website by Atlassian	A/B Testing
• **Analyzing relationships** e.g. Analysing the Sydney trains network to improve Opal card services	Network Analysis, Graph databases
• **Prediction problem** e.g. Predict which class(segment) of customers Westpac will apply for a credit card	Multi-class Logistic regression
• **Data Analysis** e.g. Cleaning large datasets at PwC and reproducing the code. Reduce data entry error	Libraries Numpy (data manipulation with dataframes) and Scikit-Learn for mathematical calculations
• **Text Analytics** e.g. Allianz Insurance wants to analyze survey feedback	Natural Language Processing

PRE-WORK DOWNLOAD PYTHON 3.7

Download Python 3.7 version

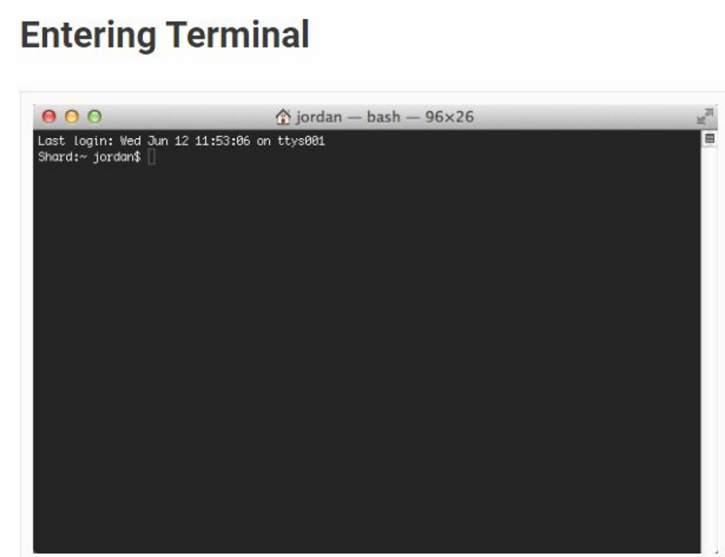
https://www.anaconda.com/distribution/

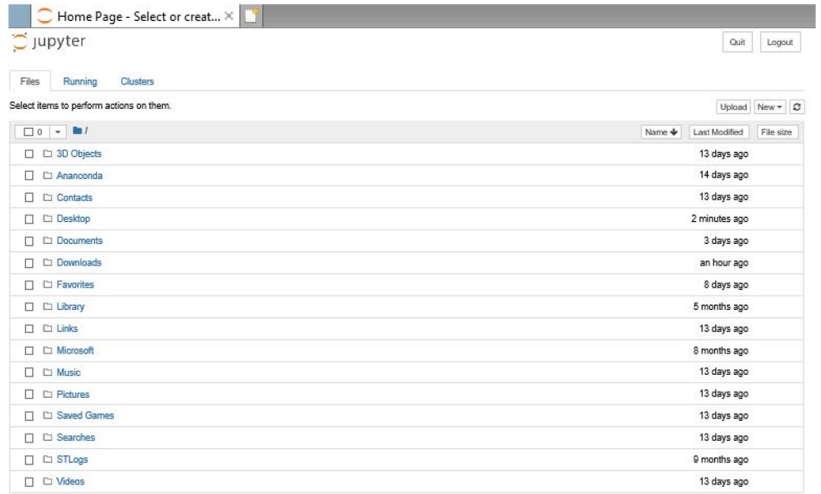


Launch Python 3.7 from Mac

1. Create a **new folder** e.g. Python Training from your Desktop.
2. In **Terminal**, type **Jupyter Notebook.**

Launch Jupyter Notebook

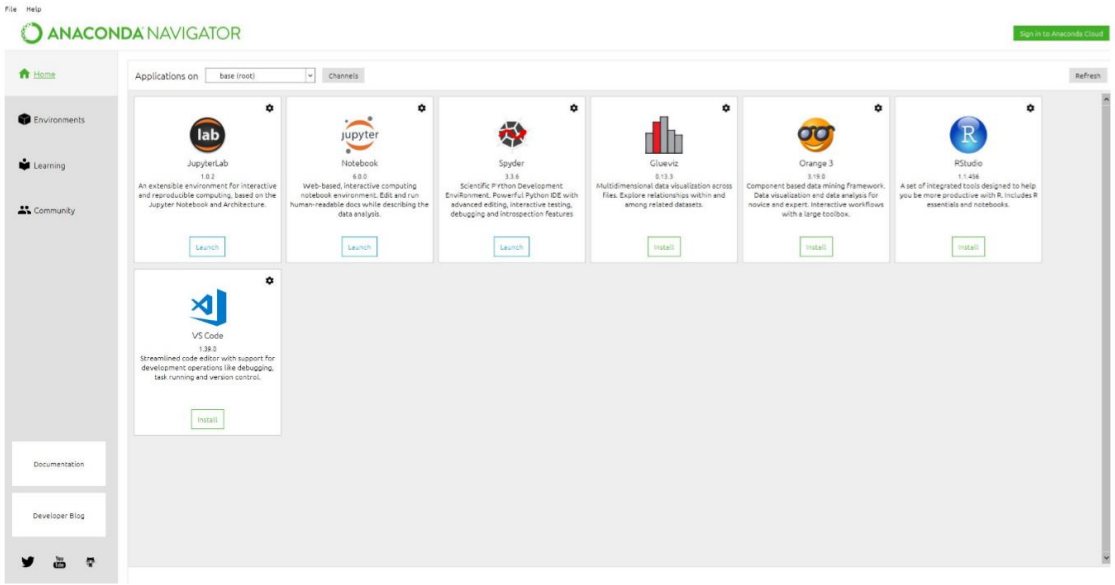




Launch Python 3.7 from Windows

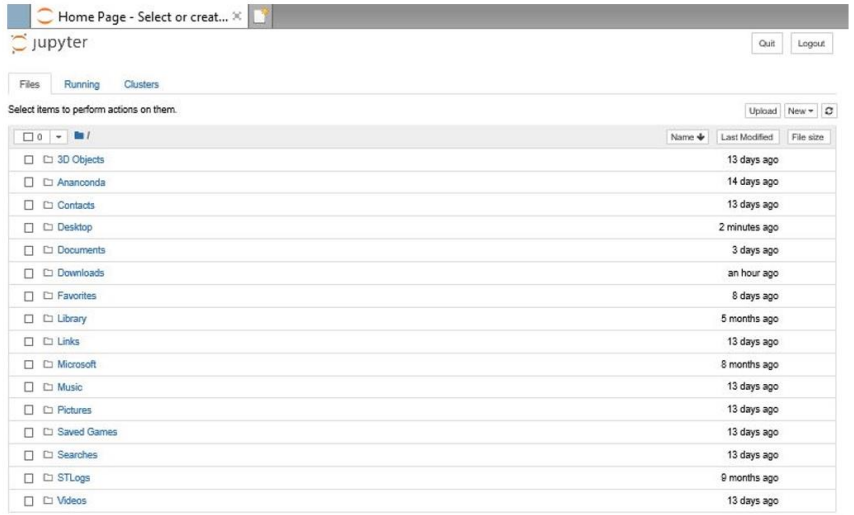
1. Create a **new folder** e.g. Python Training from your Desktop.
2. From your **Start Menu**, type **Anaconda Navigator (Anaconda)**.
3. Click **Ok** on the pop up message.
4. Click **Launch** from **Jupyter Notebook.**

Launch Jupyter Notebook



MAC Users:

In Terminal, type: **jupyter notebook** to launch your notebook



Top Analytics, Data Science, Machine Learning Software

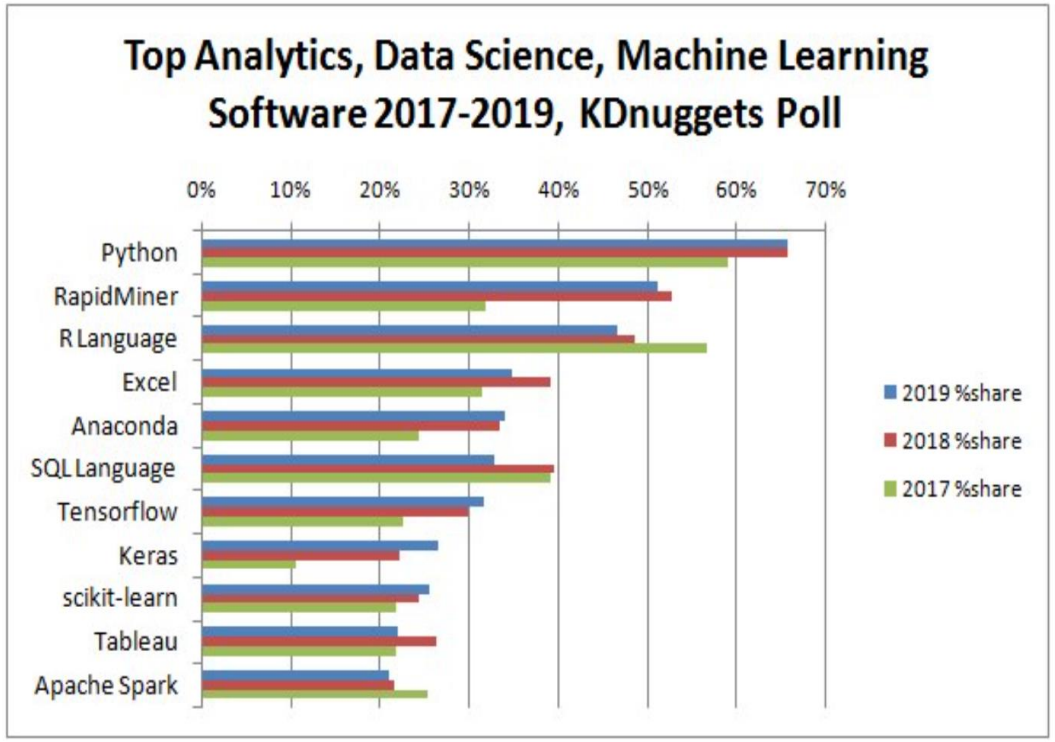


Fig 1: KDnuggets Analytics/Data Science 2019 Software Poll: top tools in 2019, and their share in the 2017, 2018 polls

PYTHON

- Reproducibility
- Turn your code into a slide deck for your boss
- Open Source
- Notebooks – code + output
- Data Visualisation
- Produce a report as a data analyst
- Deep Learning
- Machine Learning

LAB – PYTHON DEMO

CREATE A DATABASE ENGINE IN PYTHON

Slide Type Slide

Create database engine in Python

In [1]:

Slide Type Slide

```
1 # Install the package pyscopg2
```

In [2]:

Slide Type Slide

```
1 pip install psycopg2
```

```
Requirement already satisfied: psycopg2 in c:\users\wendy\anaconda3\anaconda3_1\lib\site-packages (2.8.4)
Note: you may need to restart the kernel to use updated packages.
```

In [3]:

Slide Type Slide

```
1 import psycopg2
```

PYTHON REFRESHER

3. Python Basics

Python as a Calculator

Division

```
In [1]: print(5/8)
```

```
0.625
```

Multiplication

```
In [2]: 10*4
```

```
Out[2]: 40
```

¶

```
In [3]: print(100*1.1**7)
```

```
194.87171000000012
```

Assign a variable

```
In [4]: weight = 55
```

```
In [5]: height = 1.79
```

```
In [6]: ### Calculate BMI = weight/Height ^2
```

```
In [7]: bmi = weight/height **2
```

```
In [8]: bmi
```

```
Out[8]: 17.165506694547613
```

Check the data type

What is the data type of bmi?

```
In [9]: type(bmi)
```

```
Out[9]: float
```

Slide Type Slide

Python - Connecting to the database

Slide Type Slide

```
### example: AWS PostgreSQL RDS database

* users - contains publicly-viewable data about each user in the MDSI Slack isntance.
* channels - contains data about each public channel in the MDSI Slack instance.
* messages - contains all messages posted in public channels in the MDSI Slack instance
```

In [4]:

Slide Type Slide

```
1 # Import packages
2 from sqlalchemy import create_engine
3 import pandas as pd
```

In [5]:

Slide Type Slide

```
1 # Create an engine that connects to the AWS PostgreSQL RDS database to the pysycop2 driver: engine
2 engine = create_engine("postgresql+psycopg2://dsp2019:oZkK6vgRbvDK@/mdsislack?host=mdsislack.clnutj7nhgyn.us-east-2.rds.
```

In [6]:

Slide Type Slide

```
1 # Print table names
2 print(engine.table_names())
```

```
['messages', 'user_analysis', 'users', 'channels']
```

CONNECTING TO A DATABASE - PYTHON

Slide Type Slide

Write a SELECT statement to query the 'users' table

In [8]:
```
# Perform query: rs
rs = con.execute('select * from users')
```

In [9]:
```
# Save results of the query to a pandas dataframe: df
df = pd.DataFrame(rs.fetchall())
```

In [10]:
```
# Close connection
con.close()
```

In [11]:
```
# Set the Dataframe column names
df.columns = rs.keys()
```

In [12]:
```
# Print the dataframe to the shell
print(df)
```

Slide Type Slide

Joining 'messages' and 'channels' tables with INNER Join

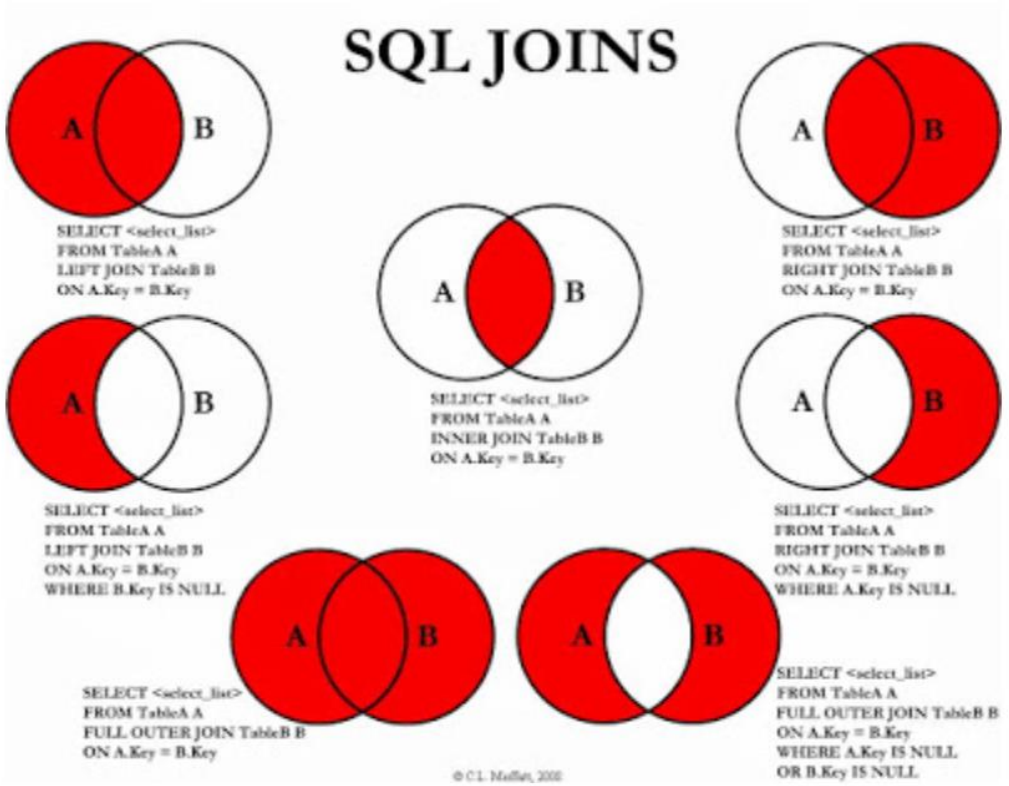
In [37]:
```
# Import packages
from sqlalchemy import create_engine
import pandas as pd
```

In [38]:
```
# Create an engine that connects to the AWS PostgreSQL RDS database to the pysycop2 driver: engine
engine = create_engine("postgresql+psycopg2://dsp2019:oZkK6vgRbvDK@/mdsislack?host=mdsislack.clnutj7nhgyn.us-east-2.rds.
```

In [39]:
```
# Open engine connection
con = engine.connect()
```

SELECT AND JOIN YOUR DATA

CHECK WHICH JOIN IS APPROPRIATE FOR YOU



- Do you want to preserve user names on the left?
- How will you treat N/A's?
- An inner join may reduce your data points
- What problem are you trying to answer?

https://javarevisited.blogspot.com/2012/11/how-to-join-three-tables-in-sql-query-mysql-sqlserver.html#axzz665Ch4Lx2

Write to CSV and read into R

In [*]:

Slide Type: Slide

```
# write dataframe to csv

df4 = df3.to_csv('df3.csv')
```

Slide Type: Slide

Install Feather to pass data between Python and R

In [*]:

Slide Type: Slide

```
pip install feather-format
```

WRITE PYTHON FILE TO R

Call Python from R Markdown

- https://rstudio.github.io/reticulate/

````
```{python}
import pandas
flights = pandas.read_csv("flights.csv")
flights = flights[flights['dest'] == "ORD"]
flights = flights[['carrier', 'dep_delay', 'arr_delay']]
flights = flights.dropna()
```

```{r, fig.width=7, fig.height=3}
library(ggplot2)
ggplot(py$flights, aes(carrier, arr_delay)) + geom_point() + geom_jitter()
```
````

In [101]:

Slide Type Slide

```
# Start with generating a word cloud on the variable 'user_name':
text = df3.user_name[0]

# Create and generate a word cloud image:
wordcloud = WordCloud().generate(text)

# Display the generated image:
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

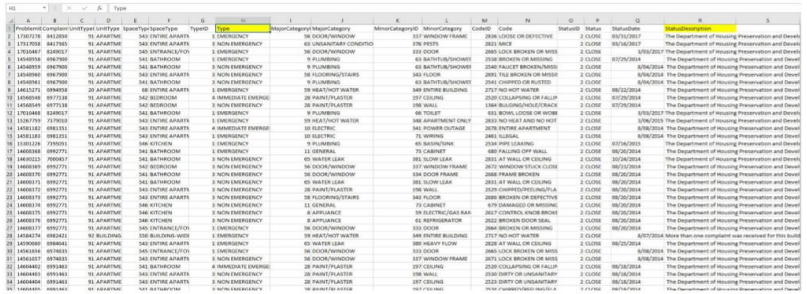
PYTHON WORD CLOUDS – TEXT DATA

LAB - PYTHON DEMO 2

DATA UNDERSTANDING

City of New York Complaints

Csv file



Dataframe

```
In [2]:
## Load the data into a Pandas Dataframe from csv file
import pandas as pd
df = pd.read_csv('Complaint_Problems.csv')
#### Inspect the first 5 values of the dataset
df.head()
```

Out[2]:

	ProblemID	ComplaintID	UnitTypeID	UnitType	SpaceTypeID	SpaceType	TypeID	Type	MajorCategoryID	MajorCategory	MinorCategoryI
0	17307278	8412850	91	APARTMENT	543	ENTIRE APARTMENT	1	EMERGENCY	56	DOOR/WINDOW	33
1	17317058	8417365	91	APARTMENT	543	ENTIRE APARTMENT	3	NON EMERGENCY	63	UNSANITARY CONDITION	37
2	17016467	8249017	91	APARTMENT	545	ENTRANCE/FOYER	1	EMERGENCY	56	DOOR/WINDOW	33
3	14548958	6967900	91	APARTMENT	541	BATHROOM	1	EMERGENCY	9	PLUMBING	6
4	14548959	6967900	91	APARTMENT	541	BATHROOM	3	NON EMERGENCY	9	PLUMBING	6

Variable

Slide Type Slide

Data Dictionary

There were 18 variables within the data set with 1045040 rows

* ProblemID
* ComplaintID
* UnitTypeID
* UnitType
* SpaceTypeID
* SpaceType
* TypeID
* Type
* MajorCategoryID
* MajorCategory
* MinorCategoryID
* MinorCategory
* CodeID
* Code
* StatusID
* Status
* StatusDate
* StatusDescription

Tasks:

1. Insights from the dataset
* Exploratory Data Analysis

2. Build a machine learning model that performs MultiClass classification to predict the outcome of complaint type
* Type: 1=EMERGENCY 2=HARZARDOUS 3=IMMEDIATE EMERGENCY 4=NON EMERGENCY
* StatusDescription
* UnitType
* SpaceType
* MajorCategory
* MinorCategory
* Code
* Status
* StatusDate

DATA PREPARATION

City of New York Complaints - Transform categorical variables

In [3]:

```
## Exploratory Data Analysis

#### Create a new dataframe for exploratory data analysis

df2 = df[['MajorCategory','Code','Status','StatusDate', 'Type', 'MinorCategory']]

df2.head()
```

Out[3]:

	MajorCategory	Code	Status	StatusDate	Type	MinorCategory
0	DOOR/WINDOW	LOOSE OR DEFECTIVE	CLOSE	3/31/17	EMERGENCY	WINDOW FRAME
1	UNSANITARY CONDITION	MICE	CLOSE	3/16/17	NON EMERGENCY	PESTS
2	DOOR/WINDOW	LOCK BROKEN OR MISSING	CLOSE	3/3/17	EMERGENCY	DOOR
3	PLUMBING	BROKEN OR MISSING	CLOSE	7/29/14	EMERGENCY	BATHTUB/SHOWER
4	PLUMBING	FAUCET BROKEN/MISSING/LEAKING	CLOSE	8/4/14	NON EMERGENCY	BATHTUB/SHOWER

In [6]:

```
y_enc = l_encoder.transform(y)
np.unique(y_enc)
```

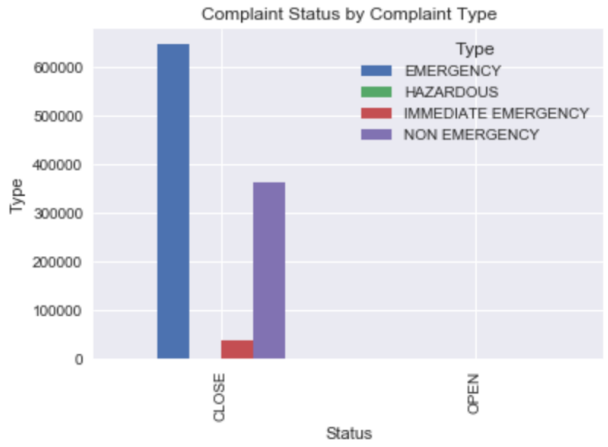
Out[6]:

```
array([0, 1, 2, 3])
```

In [44]:

```
# Barplot of Status grouped by Type
pd.crosstab(df2.Status, df2.Type).plot(kind='bar')
plt.title('Complaint Status by Complaint Type')
plt.xlabel('Status')
plt.ylabel('Type')

```

Out[44]: Text(0,0.5,'Type')

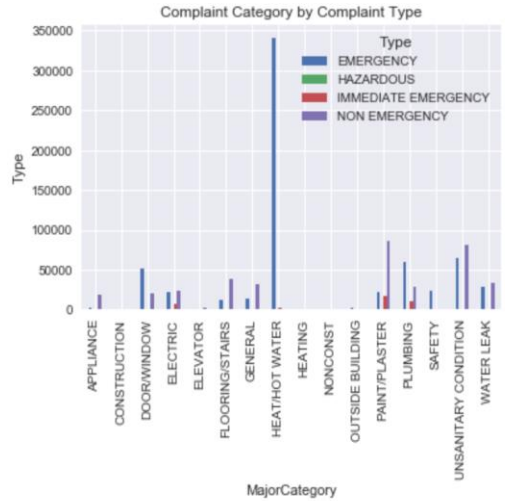


Imbalanced classes – resampling methods cross validation

In [38]:

```
# Barplot of MajorCategory grouped by Type
pd.crosstab(df2.MajorCategory, df2.Type).plot(kind='bar')
plt.title('Complaint Category by Complaint Type')
plt.xlabel('MajorCategory')
plt.ylabel('Type')
```

Out[38]: Text(0,0.5,'Type')



EXPLANATORY INSIGHTS

PREDICTIVE MODELLING

City of New York Complaints - Model Evaluation

Slide Type: Slide

Model Selection

We are now ready to experiment with different machine learning models, evaluate their accuracy and find the source of any potential issues.

We will benchmark the following four models:

1. Logistic Regression
2. (Multinomial) Naive Bayes
3. Linear Support Vector Machine
4. Random Forest

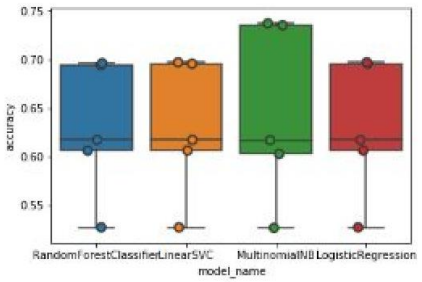
Multi-Class Model Selection

In [19]:

Slide Type: Slide

```
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import LinearSVC
from sklearn.model_selection import cross_val_score
models = [
    RandomForestClassifier(n_estimators=200, max_depth=3, random_state=0),
    LinearSVC(),
    MultinomialNB(),
    LogisticRegression(random_state=0),
]
CV = 5
cv_df = pd.DataFrame(index=range(CV * len(models)))
entries = []
for model in models:
  model_name = model.__class__.__name__
  accuracies = cross_val_score(model, features, labels, scoring='accuracy', cv=CV)
  for fold_idx, accuracy in enumerate(accuracies):
    entries.append((model_name, fold_idx, accuracy))
cv_df = pd.DataFrame(entries, columns=['model_name', 'fold_idx', 'accuracy'])
import seaborn as sns
sns.boxplot(x='model_name', y='accuracy', data=cv_df)
sns.stripplot(x='model_name', y='accuracy', data=cv_df,
              size=8, jitter=True, edgecolor="gray", linewidth=2)

plt.show()
```

- Logistic Regression
- Random Forest
- Multi-nomial Naive Bayes
- Linear Support Vector Machine



Visualise the ModelLinear Support vector

- Multinomial Naive Bayes and
- Linear Support Vector Machine

Performed better than the other classifiers

EVALUATE THE MODEL

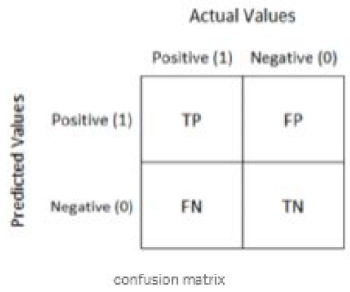
City of New York Complaints - Model Evaluation

```
In [22]:
model = MultinomialNB()
X_train, X_test, y_train, y_test, indices_train, indices_test = train_test_split(features, labels, df.index, test_size=
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
from sklearn.metrics import confusion_matrix
conf_mat = confusion_matrix(y_test, y_pred)
fig, ax = plt.subplots(figsize=(10,10))
sns.heatmap(conf_mat, annot=True, fmt='d',
            xticklabels=TypeID_df.Type.values, yticklabels=TypeID_df.Type.values)
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.show()
```

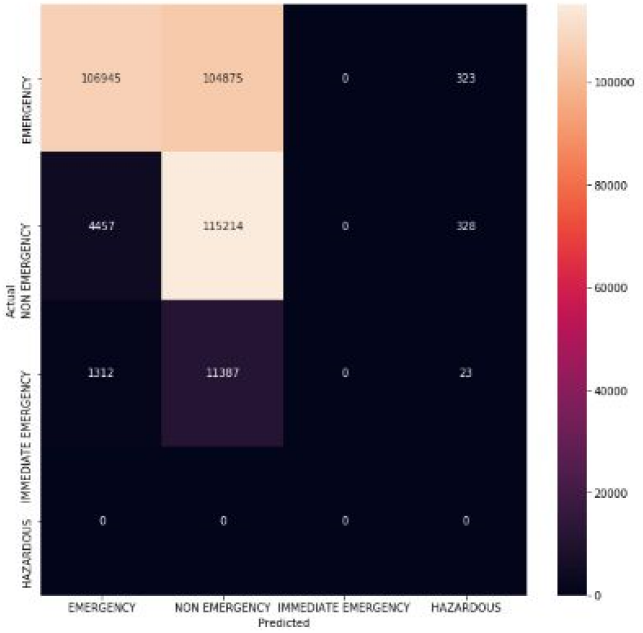
```
In [25]:
#### Misclassified complaint status are complaints that touch on more than one subjects (for example, complaints involv
#### Again, we use the chi-squared test to find the terms that are the most correlated with each of the categories:

model.fit(features, labels)
N = 2
for Type, TypeID in sorted(Type_to_id.items()):
  indices = np.argsort(model.coef_[TypeID])
  feature_names = np.array(tfidf.get_feature_names())[indices]
  unigrams = [v for v in reversed(feature_names) if len(v.split(' ')) == 1][:N]
  bigrams = [v for v in reversed(feature_names) if len(v.split(' ')) == 2][:N]
  print("# '{}':".format(Type))
  print("  . Top unigrams:\n       . {}".format('\n       . '.join(unigrams)))
  print("  . Top bigrams:\n       . {}".format('\n       . '.join(bigrams)))
```

```
# 'EMERGENCY':
  . Top unigrams:
       . complaint
       . violations
  . Top bigrams:
       . complaint closed
       . department housing
# 'HAZARDOUS':
  . Top unigrams:
       . violations
       . www
  . Top bigrams:
       . issued information
       . www nyc
# 'IMMEDIATE EMERGENCY':
  . Top unigrams:
       . violations
       . inspected
  . Top bigrams:
       . violations issued
       . conditions violations
# 'NON EMERGENCY':
  . Top unigrams:
       . violations
       . inspected
  . Top bigrams:
       . conditions violations
       . development inspected
```

Fitting the model



Confusion Matrix - produces **model evaluation metrics**



MODEL EVALUATION METRICS

The **confusion matrix**, which is a breakdown of predictions into a table showing correct predictions and the types of incorrect predictions made. Ideally, you will only see numbers in the diagonal, which means that all your predictions were correct!

- Precision is a measure of a classifier's exactness. The higher the precision, the more accurate the classifier.

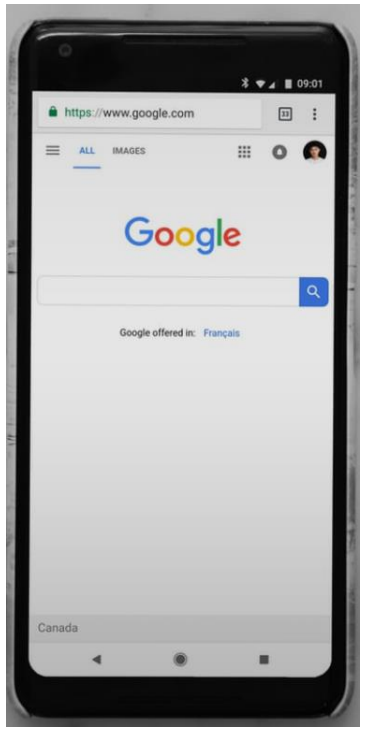
- Recall is a measure of a classifier's completeness. The higher the recall, the more cases the classifier covers.

- The F1 Score or F-score is a weighted average of precision and recall.

- Area under the ROC curve.

Source: Udacity

NLP

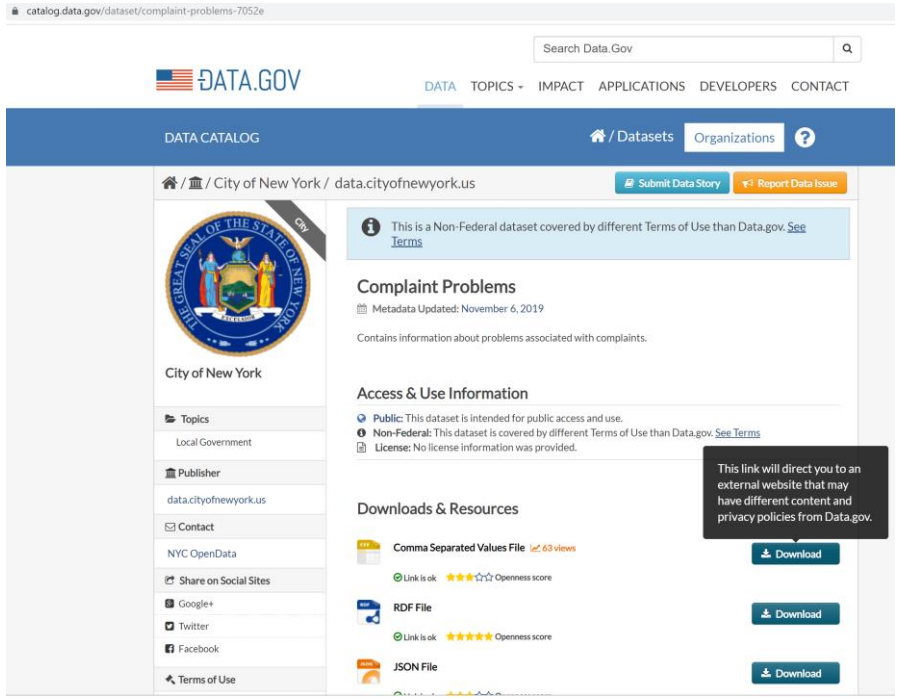


NLP – UNSTRUCTURED DATA
NATURAL LANGUAGE PROCESSING

Data Dictionary

There were 18 variables within the data set with 1045040 rows

- ProblemID
- ComplaintID
- UnitTypeID
- UnitType
- SpaceTypeID
- SpaceType
- TypeID
- Type
- MajorCategoryID
- MajorCategory
- MinorCategoryID
- MinorCategory
- CodeID
- Code
- StatusID
- Status
- StatusDate
- StatusDescription



Data Source

- City of New York complaint problems dataset is publicly available from the website data.gov via the link:

https://catalog.data.gov/dataset/complaint-problems-7052e

- Data published by: data.cityofnewyork.us
- Data last updated: 6 September 2019

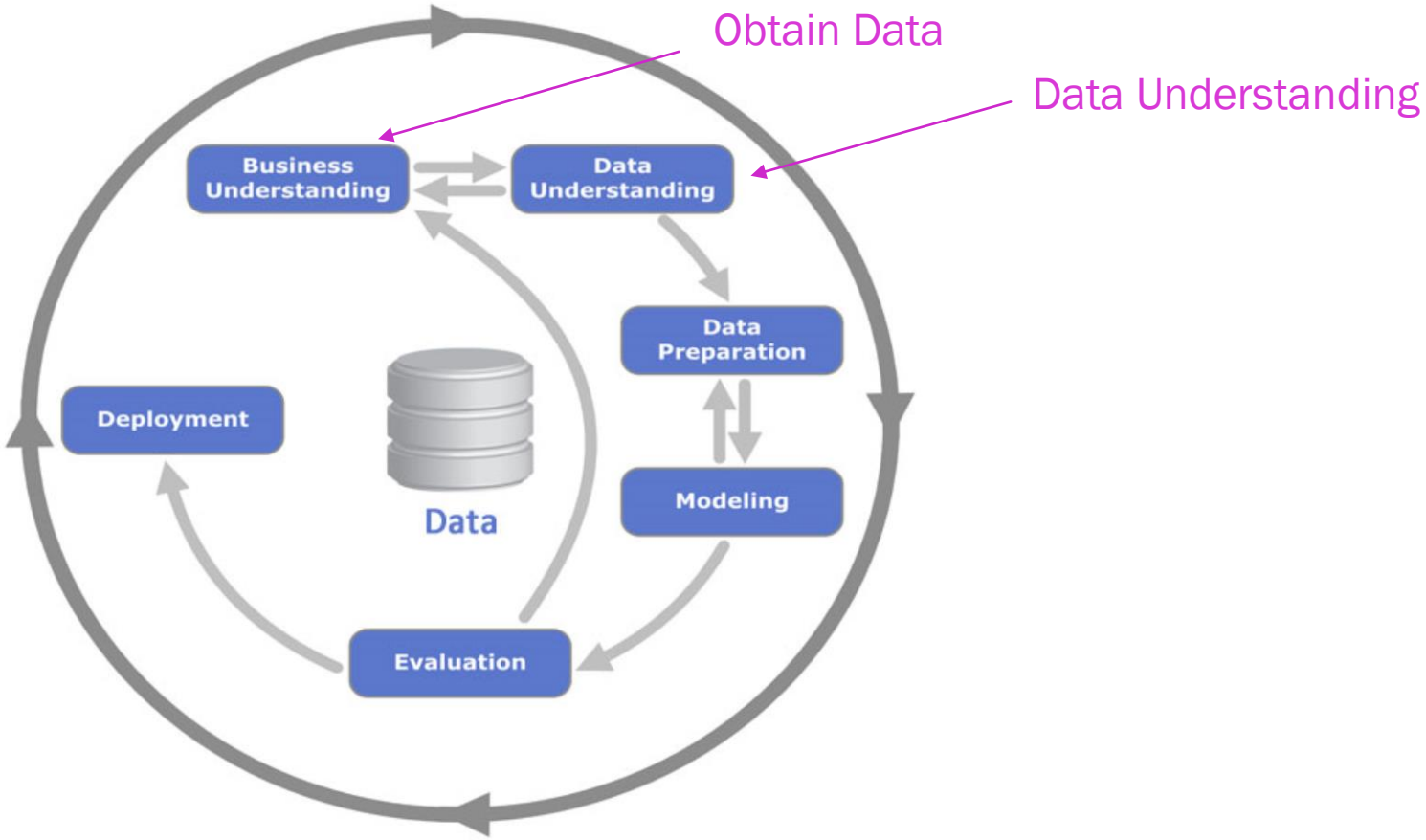
OBTAIN THE DATA

- Given a new complaint comes in, We want to assign it to one of 4 categories. The classifier makes the assumption that each new complaint is assigned to one and only one category.

WHAT IS CRISP DM?

Cross Industry Standard Process for Data Mining

CRISP-DM Process Diagram



Source: Kenneth Jensen

LAB – NLP DEMO

SPACY

In [77]:

Slide Type: Slide

```
# Visualise the input text with displacy

displacy.render(nytimes, style = "ent",jupyter = True)
```

"But this our purpose now is twelve month old DATE ,""And bootless 'tis to tell you we will go:""Therefore we meet not now. Then let me hear""Of you, my gentle cousin Westmoreland PERSON ,""What yesternight our council did decree""In forwarding this dear expedience.""My liege, this haste was hot in question,""And many limits of the charge set down""But yesternight: when all athwart there came""A post from Wales GPE loaden with heavy news,""Whose worst was, that the noble Mortimer PERSON ,""Leading the men of Herefordshire GPE to fight""Against the irregular and wild Glendower,""Was by the rude hands of that Welshman ORG taken,"" A thousand CARDINAL of his people butchered,""Upon whose dead corpse there was such misuse,""Such beastly shameless transformation,""By those Welshwomen done as may not be""Without much shame retold or spoken of.""It seems then that the tidings of this broil""Brake off our business for the Holy Land ORG .""This match'd with other did, my gracious lord,""For more uneven and unwelcome news""Came from the north and thus it did import:"" On Holy-rood day WORK_OF_ART , the gallant Hotspur there," "Young Harry Percy and WORK_OF_ART brave Archibald PERSON ,""That ever-valiant and approved Scot PERSON ,""At Holmedon GPE met,""Where they did spend a sad and bloody hour TIME ,""As by discharge of their artillery,""And shape of likelihood, the news was told,""For he that brought them, in the very heat""And pride of their contention did take horse,""Uncertain of the issue any way.""Here is a dear, a true industrious friend,""Sir Walter Blunt PERSON , new lighted from his horse.""Stain'd with the variation of each soil"" Betwixt NORP that Holmedon GPE and this seat of ours,""And he hath brought us smooth and welcome news."" The Earl of Douglas WORK_OF_ART is discomfited:"" Ten thousand CARDINAL bold Scots ORG , two CARDINAL and twenty CARDINAL knights,""Balk'd in their own blood did Sir Walter PERSON see""On Holmedon GPE 's plains. Of prisoners, Hotspur took"" Mordake the Earl of Fife WORK_OF_ART , and eldest son""To beaten Douglas PERSON , and the Earl of Athol WORK_OF_ART ,""Of Murray, Angus PERSON , and Menteith:""And is not this an honourable spoil?""A gallant prize? ha, cousin, is it not?""In faith,""It is a conquest for a prince to boast of."" Yea PERSON , there thou makest me sad and makest me sin""In envy that my Lord Northumberland PERSON ""Should be the father to so blest a son,""A son who is the theme of honour's tongue,"" Amongst WORK_OF_ART a grove, the very straightest plant,""Who is sweet Fortune ORG 's minion and her pride:"" Whilst I, by looking on the praise of him WORK_OF_ART ,""See riot and dishonour

SPACY

In [34]:

Slide Type Slide

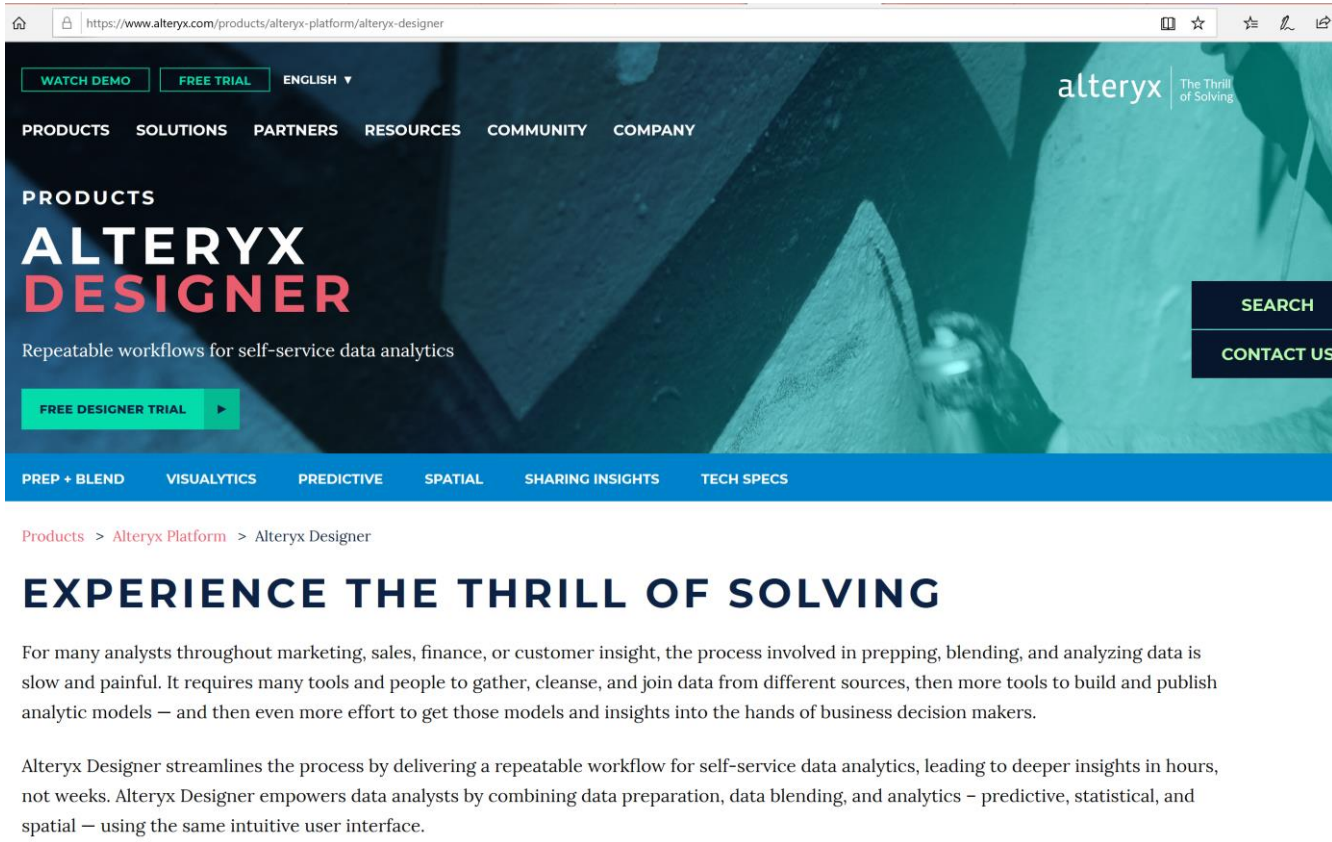
```
# Find named entities, phrases and concepts

for entity in doc.ents:
    print(entity.text, entity.label_)
```

```
ACT ISCENE I. London ORG
KING HENRY PERSON
the EARL of WESTMORELAND ORG
WALTER BLUNT PERSON
hoofsOf PRODUCT
one CARDINAL
one CARDINAL
shockAnd ORG
March DATE
one CARDINAL
kindred PERSON
Christ ORG
crossWe GPE
Forthwith PERSON
English LANGUAGE
fourteen hundred years ago DATE
the bitter cross ORG
```

ALTERYX

ALTERYX DESIGNER

- Excel on steroids
- Self service analytics – you do not need to code
- What can it do?
  - Workflow Optimization – compress file size, speed up processing
  - Data Preparation and Data Blending (joins)
  - Writing fast and accurate expressions
  - Using Macros in workflows
  - Containers
  - Data Parsing – Text to Columns
  - Outputs files to csv, an Alteryx workflow and even a Tableau hyper file
  - Predictive analytics

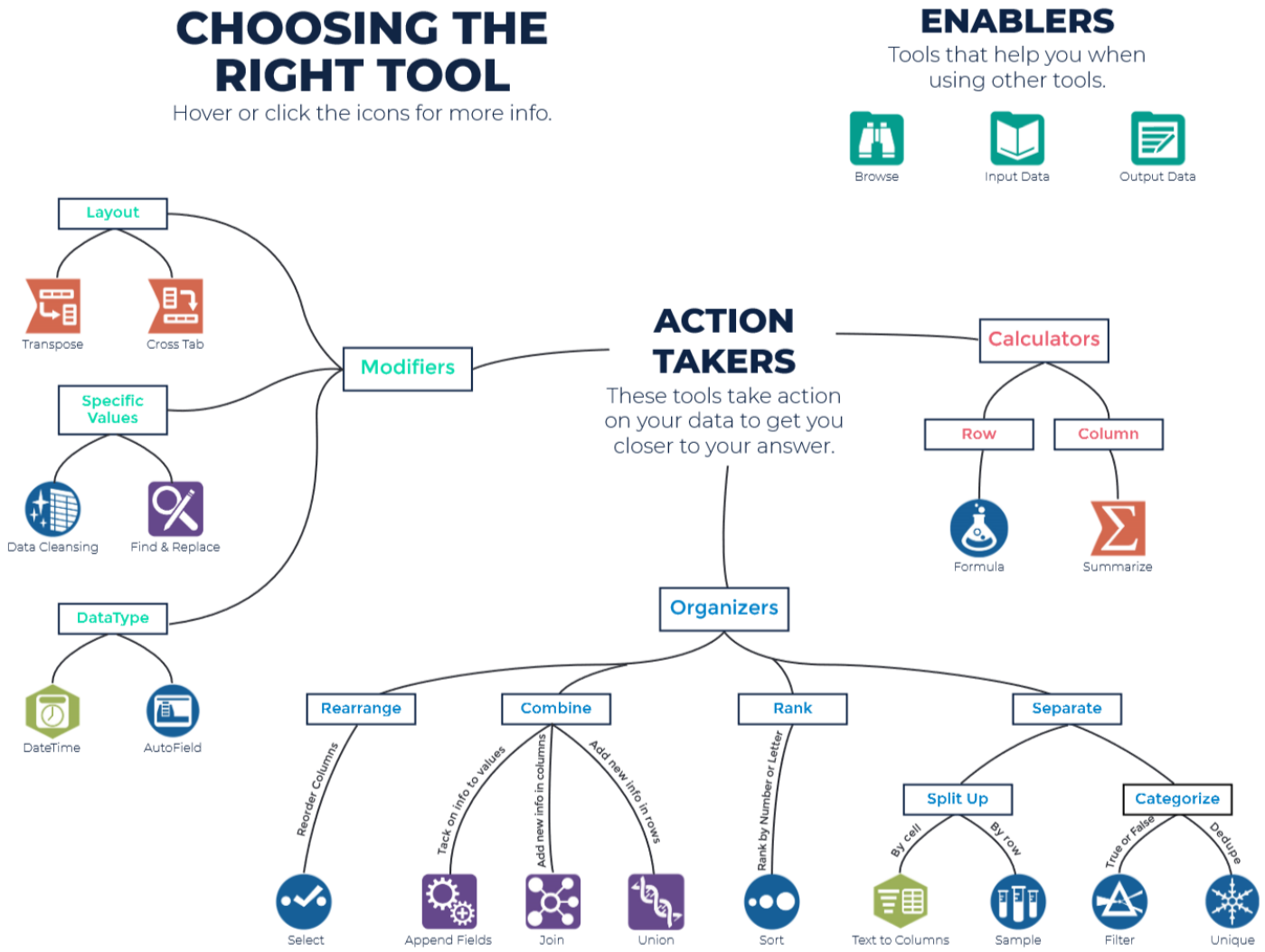


Machine Learning Platforms

Fig. 1: Gartner 2019 Magic Quadrant for Data Science and Machine Learning Platforms (as of Nov 2018)

- Open source languages **Python** and **R** do not appear in Gartner's 2018 magic quadrant
- Alteryx is a 'challenger' in machine learning and data science

Choosing the right Alteryx Tool



https://community.alteryx.com/t5/Alteryx-Community-Resources/Designer-Cheat-Sheet/ta-p/371793?attachment-id=36217

CHOOSING THE RIGHT ALTERYX TOOL

ACTIONS YOU MAY WANT TO TAKE...	TOOLS THAT CAN DO THAT IN DESIGNER
Change Datatype	Select, Formula, Autofield
Row to Column	Transpose
Column to row	Cross Tab
Split one cell into multiple cells	Text to Column, Formula
Combine lists by adding rows	Union
Combine lists by adding columns	Join, Find & Replace, Append Fields
Group Information	Union, Find & Replace, Summarize
Rank Data	Sort, Summarize
Get rid of columns	Select
Get rid of empty values	Formula, Filter
Get rid of rows	Formula, Filter, Sample
Get rid of puncutation or whitespace	Data Cleansing
Perform a calculation	Summarize, Formula
Work with dates	DateTime, Formula
Find a value (unique, min, max)	Summarize, Unique, Find & Replace
Identify records with a unique ID	Record ID
Replace a value	Find & Replace, Formula
Input Data	Input Data
Rename Fields	Select
Reorder Fields	Select
View Results	Browse
Ouput Results	Output Data

FUNCTIONS

When using functions in Designer, keep in mind that datatype is very important. The table on the right shows the function category and an X indicates that functions in that category are compatible with that column's corresponding datatype. This is not an exhaustive list. Rather, use this table to match your data's type and find a category that is compatible with that datatype to ensure the function will work. Note that you may need to change your data's datatype if you wish to use it with a particular function.

	String	Numeric	DateTime	Boolean	Spatial
Conditional	X	X	X	X	X
Conversion	X	X			
DateTime	X		X		
File	X				
Finance		X			
Math		X			
Math: Bitwise		X			
Min/Max		X			
Operators	X	X	X	X	X
Spatial		X			X
Specialized	X	X	X	X	X
String	X				
Test	X	X	X	X	X

TERMINOLOGY

Blend - merging data from different sources into one dataset, such as data from different spreadsheets, databases, or other sources into one complete dataset.

Concatenate - joining one or more text strings together.

Datatype - an attribute of data which lets the computer know how to interpret that value. There are 5 main datatypes in Designer (string, numeric, DateTime, Boolean, Spatial). Datatypes can be changed for particular values.

Delimiter - a sequence of one or more characters that creates a boundary between values. Common delimiters include commas, pipes, and quotes.

Filter - filtering separates your data into two streams: True containing the data met your criteria, and False containing the data that did not meet your criteria.

Flag - flagging data is a technique used to categorize data. This is usually accomplished with a conditional statement which checks values against a set of criteria and creates a corresponding flag in another column.

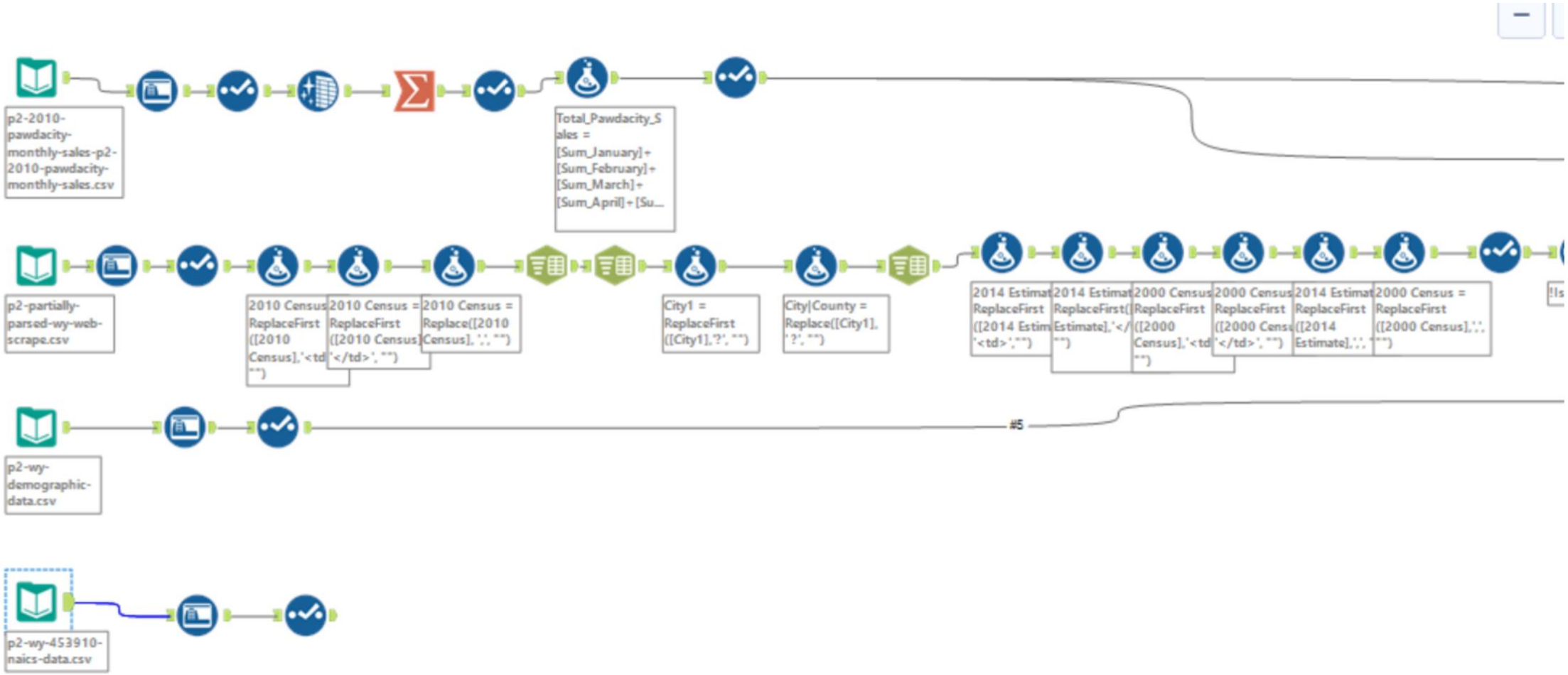
Parse - parsing separates values based on delimiters. Examples include: separating keywords from phrases, separating numbers from letters, or area codes from phone numbers.

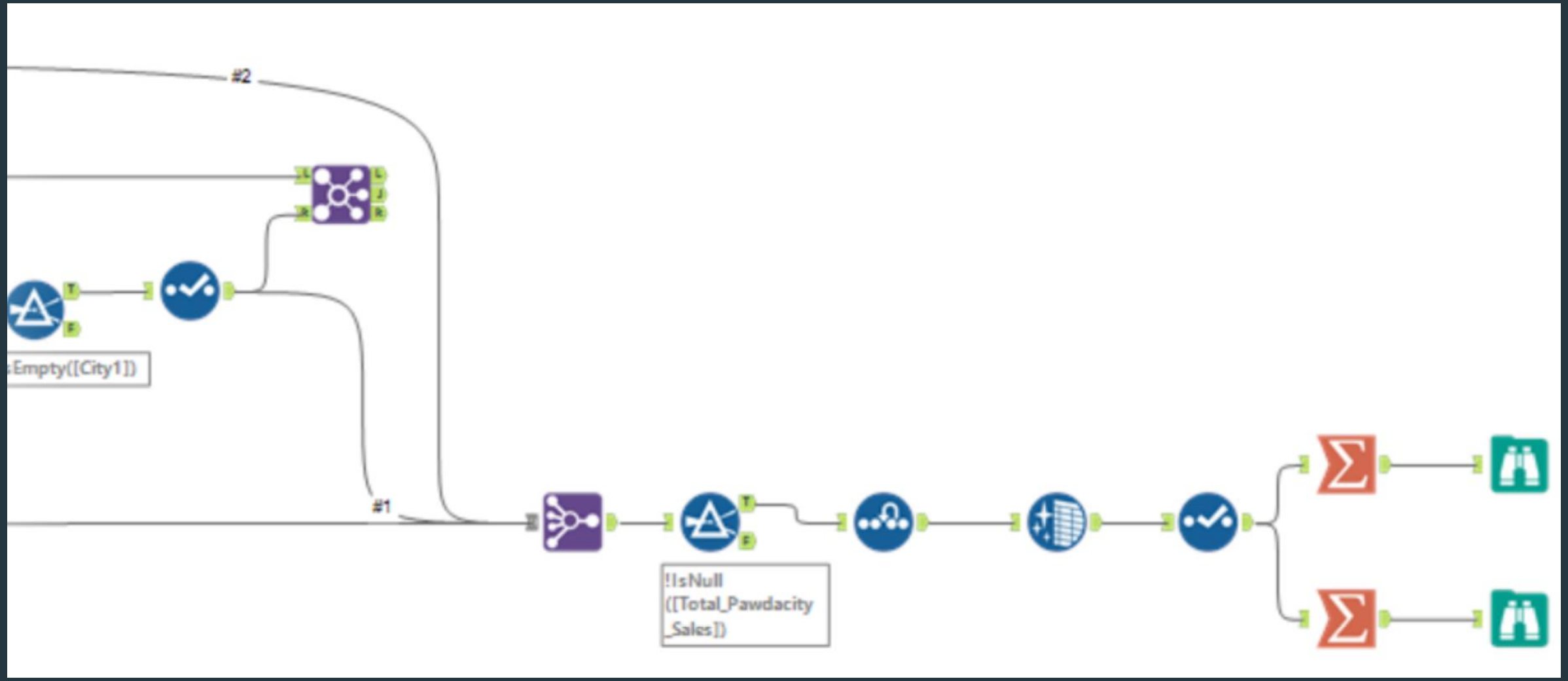
Sort - ranking items in ascending or descending order.

WEBSCRAPING

- **Python** and **Alteryx**
- We know you want to impress your boss and do web scraping
- Check firstly if your company allows you to scrape someone else's data
- Speak with your business risk team
- You may be breaching copyright laws under some company terms and conditions
- Some websites are locked down by organizations

LAB - ALTERYX LIVE DEMO

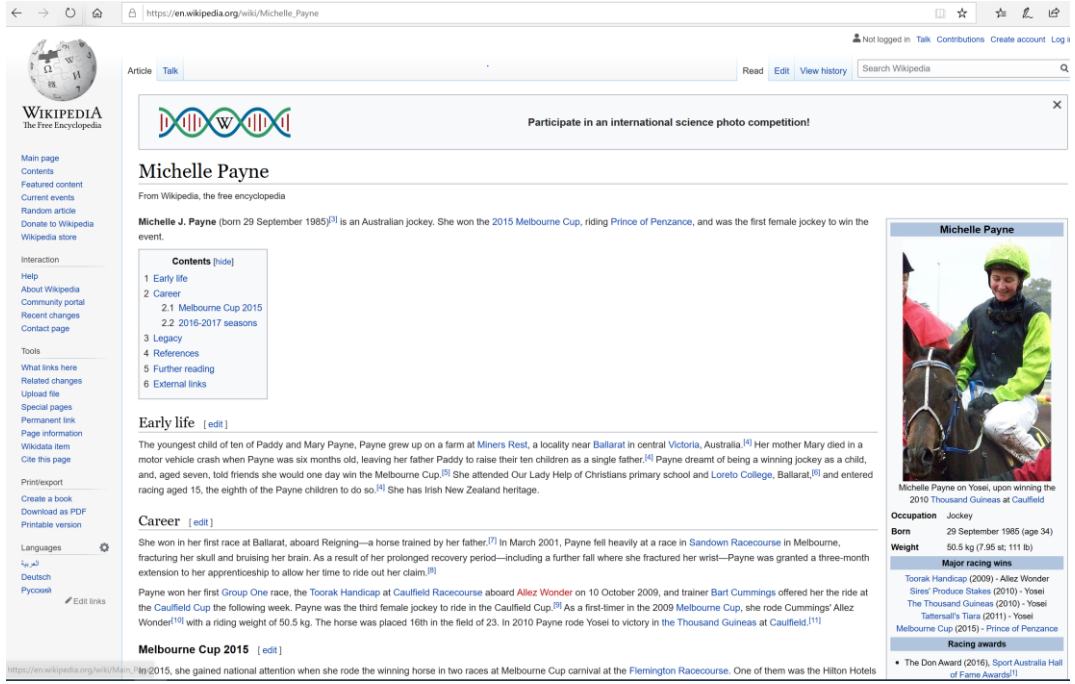

p2-2010-pawdacity-monthly-sales-p2-2010-pawdacity-monthly-sales.csv
Total_Pawdacity_Sales = [Sum_January]+[Sum_February]+[Sum_March]+[Sum_April]+[Su...
p2-partially-parsed-wy-web-scrape.csv
2010 Census = ReplaceFirst([2010 Census],'<td>','')
2010 Census = ReplaceFirst([2010 Census],'</td>','')
2010 Census = Replace([2010 Census],',','')
City1 = ReplaceFirst([City1],'?','')
City|County = Replace([City1],'?','')
2014 Estimat ReplaceFirst([2014 Estim '<td>','')
2014 Estimat ReplaceFirst([Estimate],'</ '')
2000 Census ReplaceFirst([2000 Census],'<td '')
2000 Census ReplaceFirst([2000 Cens '</td>','')
2014 Estimat ReplaceFirst([2014 Estimate],',','
2000 Census = ReplaceFirst([2000 Census],',','')
!Is
#5
p2-wy-demographic-data.csv
p2-wy-453910-naics-data.csv


#2
Empty([City1])
#1
!IsNull
([Total_Pawdacity
_Sales])

CAREER ADVICE

MY TWO CENTS – DATA SCIENCE CAREER TIPS

- Meetup.com e.g data science, data engineering, Rladies, Sydney Women in Machine Learning and Data Science, Women who code, PyData, Docker, Kubernetes, AWS
- Lunch & Learn
- Hackathons - Data Science is a team sport
- Data Science community – Get Involved
- Data Science blogs
- Learn Python, R and SQL
- Internship



Roles and responsibilities

- Design and development of solutions to greatly increase the efficiency of our internal operations
- Producing visualisations and automated reporting that can be presented or emailed to key stakeholders
- Designing automated notifications for internal and external stakeholders using real time analytics
- Operational analytics and fraud detection
- Helping to design any IT implementation required to enable the above to be executed

Skills & Experience

Skills & Experience

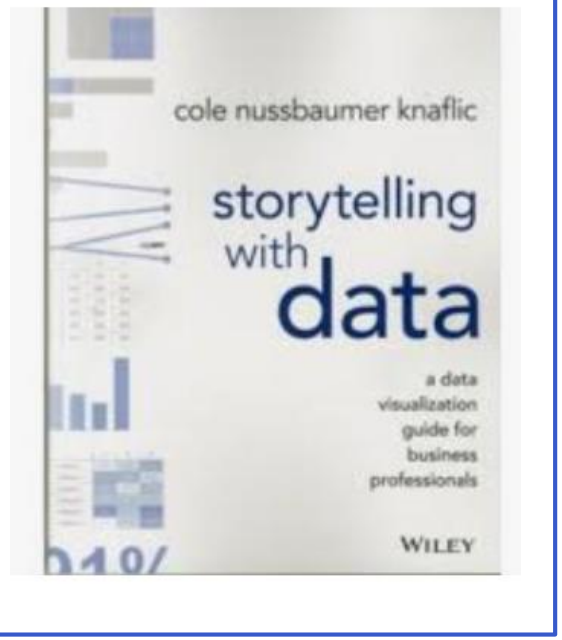
- Advanced programming skills, including in Python, REST APIs
- Advanced knowledge and experience in SQL
- Advanced knowledge and experience using AI / Machine Learning algorithms
- Advanced knowledge of analytics and data processing technologies and architectures.
- Strong analytical and problem solving skills
- Understanding of geospatial data related technologies.
- Experience scoping/running analytics projects
- Experience in analysing large volumes of unstructured data
- Experience in building analytics applications for deployment in real or near real time
- Clear communicator (verbal and written) in order to discuss complex information, analysis and findings to technical, and non-technical audiences
- Familiar with CSS, HTML, IMAP, Visual Basic, Node JS, XML, Tableau, Tosca or Selenium

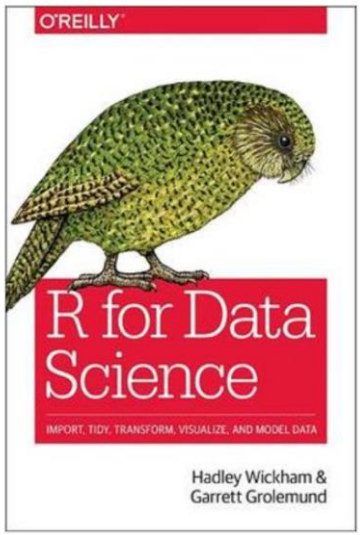
Data Science Interviews

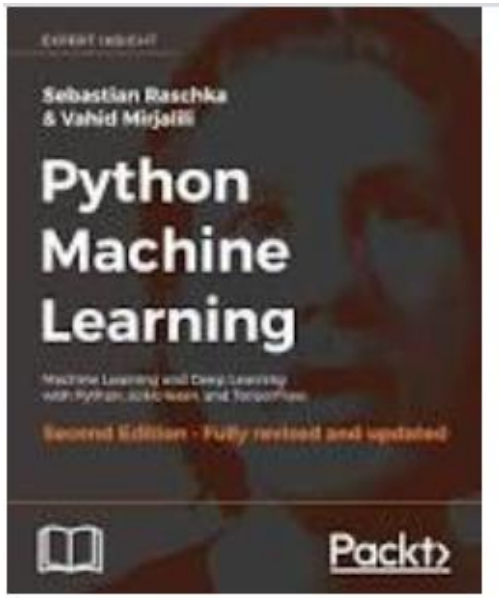
Job Boards– Buyer Beware

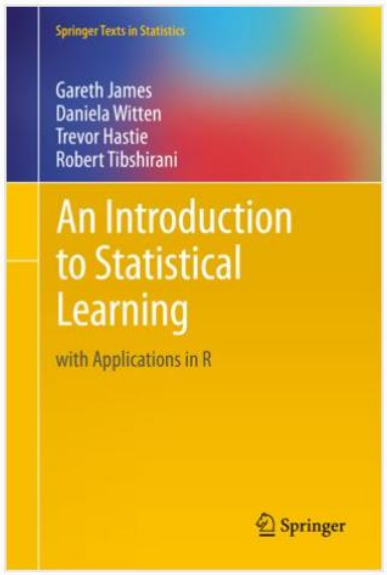
- Ask the **recruiter** what duties you will be performing in your data scientist role
- Ask the **hiring manager** what your duties you will be performing in your data scientist role
- Do not proceed to a further interview if the hiring manager says ‘yes you will be doing reporting and we may do data science in 18 months time’ - Run for your life!
- Keep on applying for data science roles, get a mentor, speak to the community, do you have peace?

RESOURCES









BOOKS I RECOMMEND

HANDY RESOURCES

- https://github.com/REMitchell/python-scraping
- Tutorials
- Blogs
- Meetups
- Data Sets – AWS, Data.gov

Turn your Jupyter Notebook into slides

- **File -> Download as -> Slides (slides.html)**

In Windows

- **A short tutorial for slide conversion in Mac:**
- https://medium.com/learning-machine-learning/present-your-data-science-projects-with-jupyter-slides-75f20735eb0f

In Mac

TURN JUPYTER NOTEBOOK INTO SLIDES

THANK YOU

https://www.linkedin.com/in/wendywong7/

@abc_wendsss

Q & A