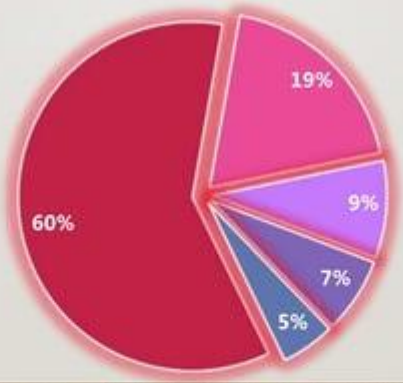
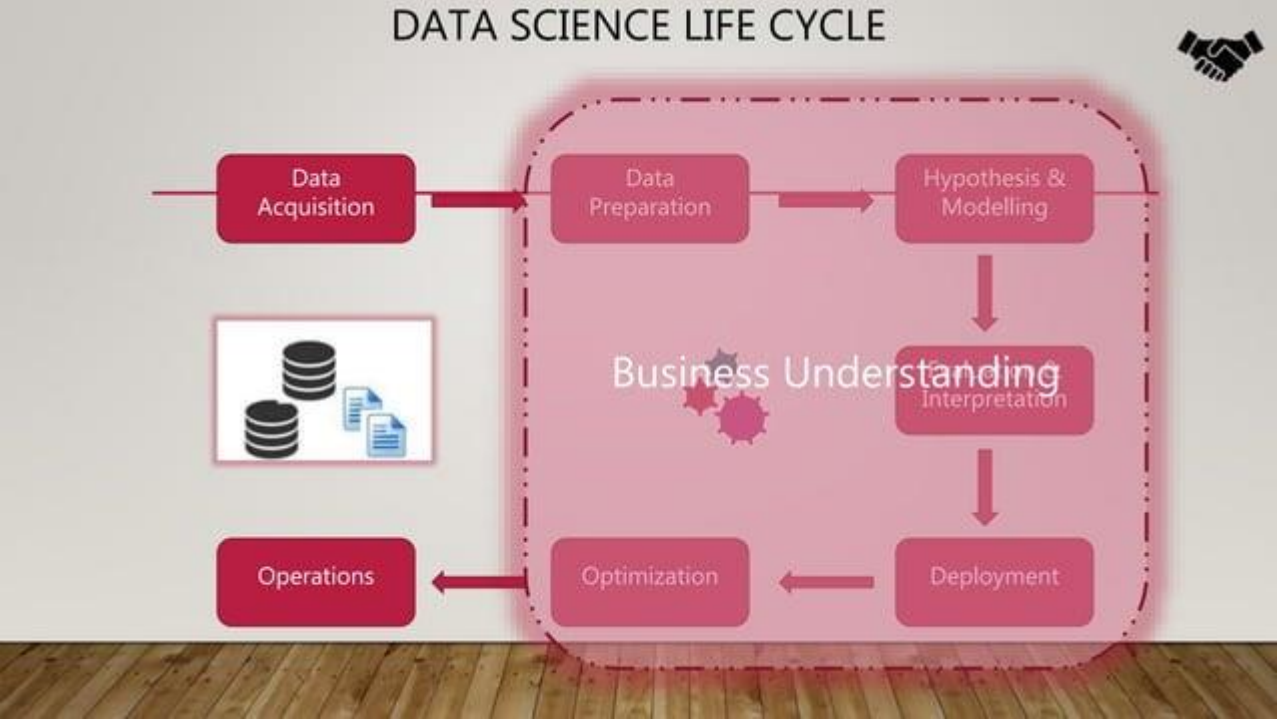
DataScience Lifecycle

DATA SCIENTIST

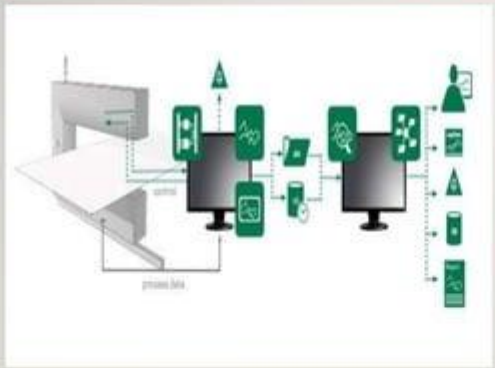
Effort



DATA SCIENCE LIFE CYCLE



DATA ACQUISITION

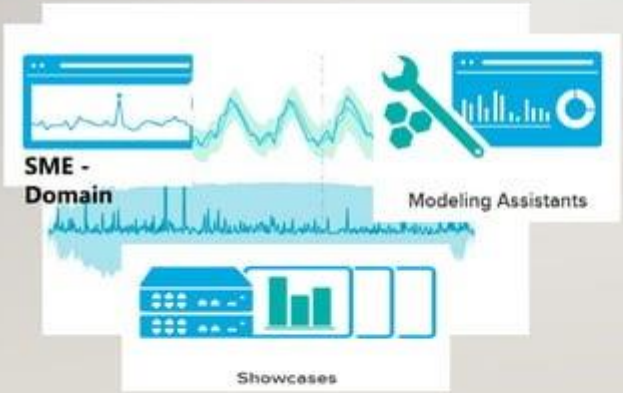


- **Static**
  - Feedback system
  - CSV Data sets / text files
- **Live**
  - Logs data, memory dumps
  - Sensors, controllers etc.
- **Virtual**
  - Data Virtualization
  - Caching , Storing

DATA SAMPLE DATASET INVENTORY

Dataset	Data Available and Accessible	Data Available, but not Accessible	Data to Collect	Data to Obtain from Third Party Sources
Products shipped	●			
Product Financials		●		
Product Call Center Data		●		
Live Product Feedback Surveys			●	
Product Sentiment from Social Media				●

PROJECT - PREDICTING FAILURE – PROACTIVE MAINTENANCE



- Baseline normal operational patterns by modelling the unstructured Log data.
- Use Domain Experts to identify patterns before failures.
- Use statistical measurements & Machine Learning to determine threshold.
- Identify patterns of activity to anticipate and react to circumstances that might otherwise disrupt operations

DATA PREPARATION

- Need for Data Preparation
  - Bad data or poor quality data can alter accuracy & lead to incorrect Insights
  - Gartner- Poor quality data costs an avg. organization $13.5M / year.
  - Dataset might contain discrepancies in the names or codes.
  - Dataset might contain outliers or errors.
  - Dataset lacks your attributes of interest for analysis.
  - All in all the dataset is not qualitative but is just quantitative.

- Steps Involved

DATA PREPARATION

- Includes steps to explore, preprocess, and condition data
- Create robust environment – analytics sandbox
- Data preparation tends to be the most labor-intensive step in the analytics lifecycle
  - Often at least 50 – 60% of the data science project's time
- The data preparation phase is generally the most iterative and the one that data scientists tend to underestimate most often ☺

- Database :

Airlines :
NYC FLIGHTS 13

- Understand the data

e.g. tailnum :
A **tail number** refers to an identification **number** painted on an aircraft, frequently on the **tail**

- Understand the Business

Goal : Predicting flight delays

Modelling

month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	origin
9	22	1514	1511	3	2000	1733	NA	UA	745	N515UA	LGA
9	22	1525	1530	-5	1921	1650	NA	EV	4580	N13958	EWR
9	22	1618	1559	19	2228	1750	NA	EV	3846	N18556	EWR
9	24	623	625	-2	NA	800	NA	MQ	3525	N735MQ	LGA
9	24	800	800	0	NA	933	NA	UA	544	N827UA	LGA
9	24	829	840	-11	1139	1020	NA	MQ	3531	N806MQ	LGA
9	25	1552	1559	-7	2048	1906	NA	UA	375	N408UA	EWR
9	25	2052	2045	7	332	53	NA	DL	347	N3731T	JFK
9	26	1331	1329	2	1923	1813	NA	UA	15	N67052	EWR
9	27	1332	1329	3	1629	1509	NA	AA	331	N565AA	LGA
9	27	2253	1945	188	NA	2146	NA	EV	5306	N605QX	LGA
9	28	555	600	-5	953	753	NA	EV	5068	N133EV	EWR
9	28	847	839	8	1130	959	NA	EV	4510	N14542	EWR
9	28	1010	1020	-10	1344	1222	NA	EV	4412	N12175	EWR
9	28	1714	1725	-11	1801	1510	NA	AA	300	N488AA	EWR

PREDICTING FLIGHT DELAYS - NYC FLIGHTS 13

- Exploratory Data Analysis of the flight data for inbound and outbound flights for year 2013 in NYC.
- Find patterns, benchmark, model and find predictors.
- To predict the flight delays for NYC Inbound/ Outbound flights.

- Ref. DataSet

month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	origin
9	22	1514	1511	3	2000	1733	NA	UA	745	N515UA	LGA
9	22	1525	1530	-5	1921	1650	NA	EV	4580	N13958	EWR
9	22	1618	1559	19	2228	1750	NA	EV	3846	N18556	EWR
9	24	623	625	-2	NA	800	NA	MQ	3525	N735MQ	LGA
9	24	800	800	0	NA	933	NA	UA	544	N827UA	LGA
9	24	829	840	-11	1139	1020	NA	MQ	3531	N806MQ	LGA
9	25	1552	1559	-7	2048	1906	NA	UA	375	N408UA	EWR
9	25	2052	2045	7	332	53	NA	DL	347	N3731T	JFK
9	26	1331	1329	2	1923	1813	NA	UA	15	N67052	EWR
9	27	1332	1329	3	1629	1504	NA	AA	331	N565AA	LGA
9	27	2253	1945	188	NA	2146	NA	EV	5306	N605QX	LGA
9	28	555	600	-5	953	753	NA	EV	5068	N133EV	EWR
9	28	847	839	8	1130	959	NA	EV	4510	N14542	EWR
9	28	1010	1020	-10	1344	1222	NA	EV	4412	N12175	EWR
9	28	1714	1725	-11	1801	1510	NA	AA	500	N4RRAA	EWR

PREDICTING FLIGHT DELAYS - NYC FLIGHTS 13

- Exploratory Data Analysis of the flight data for inbound and outbound flights for year 2013 in NYC.
- Find patterns, benchmark, model and find predictors.
- To predict the flight delays for NYC Inbound/ Outbound flights.

- Ref. DataSet

	B	C	D	E	F	G	H	I	J	K	
	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tail
013	1	1	NA	1630	NA	NA	1815	NA	EV	4308	N18
013	1	1	NA	1935	NA	NA	2240	NA	AA	791	N3E
013	1	1	NA	1500	NA	NA	1825	NA	AA	1925	N3E
013	1	1	NA	600	NA	NA	901	NA	36	125	N61
013	1	2	NA	1540	NA	NA	1747	NA	EV	4352	N10
013	1	2	NA	1620	NA	NA	1746	NA	EV	4406	N11
013	1	2	NA	1355	NA	NA	1459	NA	EV	4434	N10
013	1	2	NA	1420	NA	NA	1644	NA	EV	4935	N75
013	1	2	NA	1321	NA	NA	1536	NA	EV	3849	N13
013	1	2	NA	1545	NA	NA	1910	NA	AA	133	NA
013	1	2	NA	1330	NA	NA	1640	NA	AA	753	N3F
013	1	2	NA	1601	NA	NA	1735	NA	UA	623	NA
013	1	3	NA	645	NA	NA	757	NA	EV	4241	N14

month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailnum	origin
9	22	1514	1511	3	2000	1713	NA	UA	745	N515UA	LGA
9	22	1525	1530	-5	1521	1654	NA	EV	4580	N13958	EWR
9	22	1618	1559	19	2228	1756	NA	EV	3846	N18556	EWR
9	24	623	625	-2	NA	808	NA	MQ	3525	N735MQ	LGA
9	24	800	800	0	NA	933	NA	UA	544	N827UA	LGA
9	24	829	840	-11	1139	1020	NA	MQ	3531	N806MQ	LGA
9	25	1552	1559	-7	2048	1904	NA	UA	375	N408UA	EWR
9	25	2052	2045	7	332	5	NA	DL	347	N3731T	JFK
9	26	1331	1329	2	1523	1811	NA	UA	15	N67052	EWR
9	27	1332	1329	3	1629	1625	NA	AA	331	N565AA	LGA
9	27	2251	1945	188	NA	2146	NA	EV	5306	N605QX	LGA
9	28	555	600	-5	932	753	NA	EV	5068	N13189	EWR
9	28	847	839	8	1130	959	NA	EV	4510	N14542	EWR
9	28	1010	1020	-10	1344	1222	NA	EV	4412	N12175	EWR
9	28	1714	1725	-11	1901	1510	NA	AA	301	N4YRAA	EWR

	B	C	D	E	F	G	H	I	J	K
	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier	flight	tailn
13	1	1	NA	1630	NA	NA	1815	NA	EV	4308	N18
13	1	1	NA	1935	NA	NA	2240	NA	AA	791	N3EI
13	1	1	NA	1500	NA	NA	1825	NA	AA	1925	N3F
13	1	1	NA	600	NA	NA	901	NA	B6	125	N63
13	1	2	NA	1540	NA	NA	1747	NA	EV	4352	N10
13	1	2	NA	1620	NA	NA	1840	NA	EV	4406	N13
13	1	2	NA	1355	NA	NA	1549	NA	EV	4434	N10
13	1	2	NA	1420	NA	NA	1644	NA	EV	4936	N75
13	1	2	NA	1321	NA	NA	1536	NA	EV	3849	N13
13	1	2	NA	1545	NA	NA	1910	NA	AA	133	NA
13	1	2	NA	1330	NA	NA	1640	NA	AA	751	N3H
13	1	2	NA	1601	NA	NA	1755	NA	UA	623	NA
13	1	3	NA	645	NA	NA	757	NA	EV	4241	N2A

OBSERVATIONS - NYC FLIGHTS 13

REF. DATASET

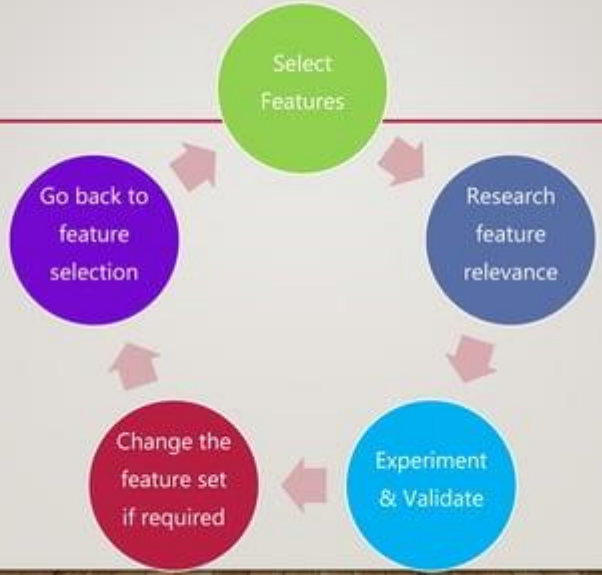
COMMON TOOLS - FOR DATA PREPARATION

- **Alpine Miner** provides a graphical user interface for creating analytic workflows
- **OpenRefine** (formerly Google Refine) is a free, open source tool for working with messy data
- Similar to OpenRefine, **Data Wrangler** is an interactive tool for data cleansing an transformation
- Alteryx and Informatica also can be tried.

HYPOTHESIS & MODELLING

- There are three main tasks addressed in this stage:
- **Feature engineering**: Create data features from the raw data to facilitate model training.
- **Model training**: Find the model that answers the question most accurately by comparing their success metrics.
- Determine if your model is **suitable for production.**

FEATURE SELECTION & ENGINEERING

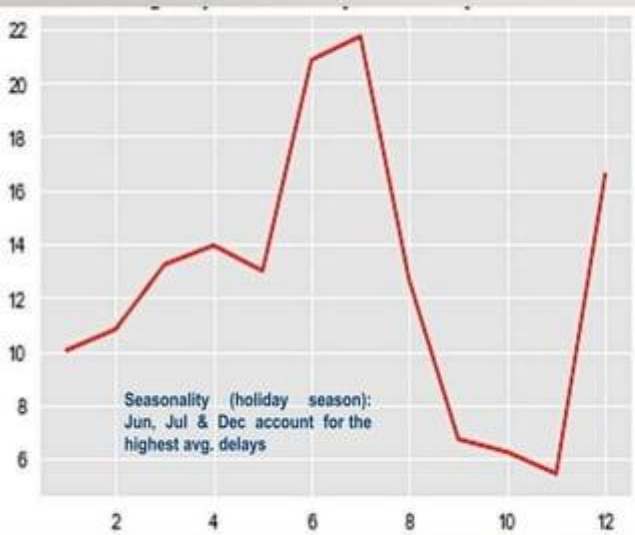


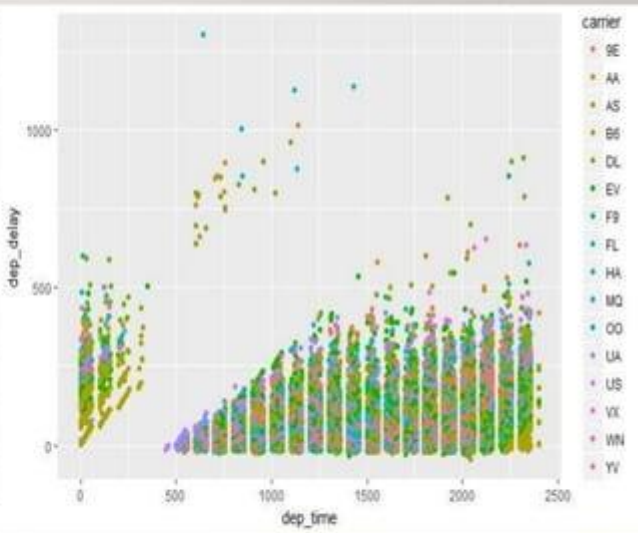
FEATURE ENGINEERING

Date	# footfalls in Dubai Mall
01/07/2017	124532
02/07/2017	65434
03/07/2017	12333
04/07/2017	60009
05/07/2017	46567
06/07/2017	98001
07/07/2017	146543
08/07/2017	112345
09/07/2017	76543

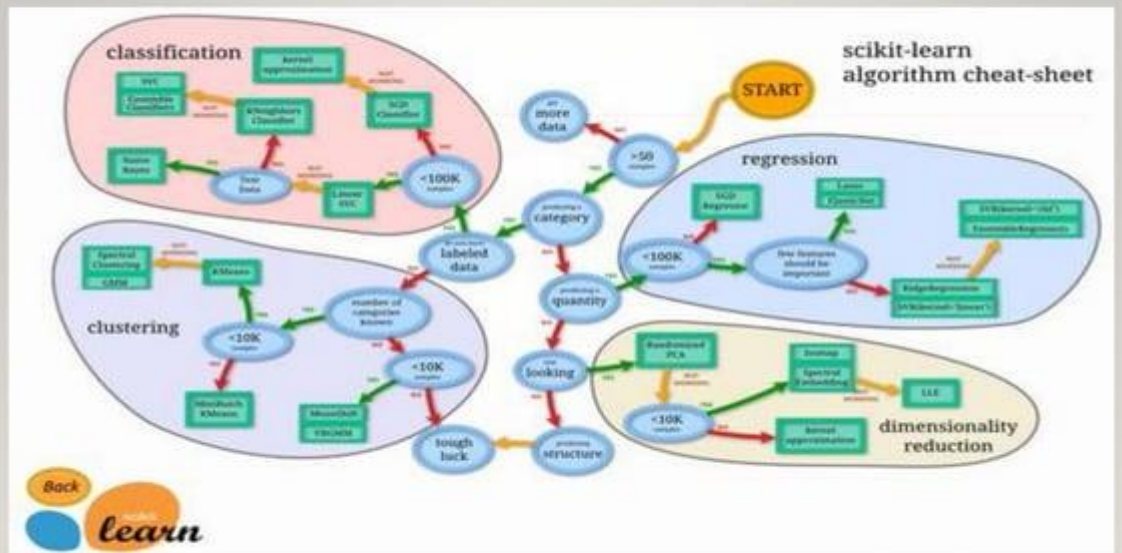
Date	# footfalls in Dubai Mall	IsHoliday?
01/07/2017	124532	Yes
02/07/2017	65434	No
03/07/2017	12333	No
04/07/2017	60009	No
05/07/2017	46567	No
06/07/2017	98001	No
07/07/2017	146543	yes
08/07/2017	112345	yes
09/07/2017	76543	No

FEATURE ENGINEERING





MODELLING


scikit-learn
algorithm cheat-sheet
START
classification
regression
clustering
dimensionality
reduction
labeled
data
category
quantity
structure
tough
luck
looking
more
data
<100K
<10K
<50
learn

CREATE YOUR MODEL & EVALUATE

- **Split the input data** randomly for modeling into a training data set and a test data set.
- **Build the models** by using the training data set.
- **Evaluate** the training and the test data set. Use a series of competing machine-learning algorithms along with the various associated tuning parameters (known as a *parameter sweep*) that are geared toward answering the question of interest with the current data.
- **Determine the "best" solution** to answer the question by comparing the success metrics between alternative methods.

CREATE YOUR MODEL & EVALUATE

- Supervised Learning
  - Naive Bayes
  - KNN
  - Support Vector Machines (SVM)
  - Linear Regression
- Unsupervised Learning
  - Principal Component Analysis.
  - K Means

- Classification Metrics
  - Accuracy Score
  - Classification Report
  - Confusion Matrix
- Regression Metrics
  - Mean Absolute Error.
  - Mean Squared Error
  - R2 Score
- Clustering Metrics
  - Adjusted Rand Index.
  - Homogeneity
  - V - measure

DEPLOYMENT

After you have a set of models that perform well, you can operationalize them for other applications through APIs or other interface to consume from various applications, such as:

- Online websites
- Spreadsheets
- Dashboards
- Line-of-business applications
- Back-end applications

Data Science Skill Tree

THANK YOU !