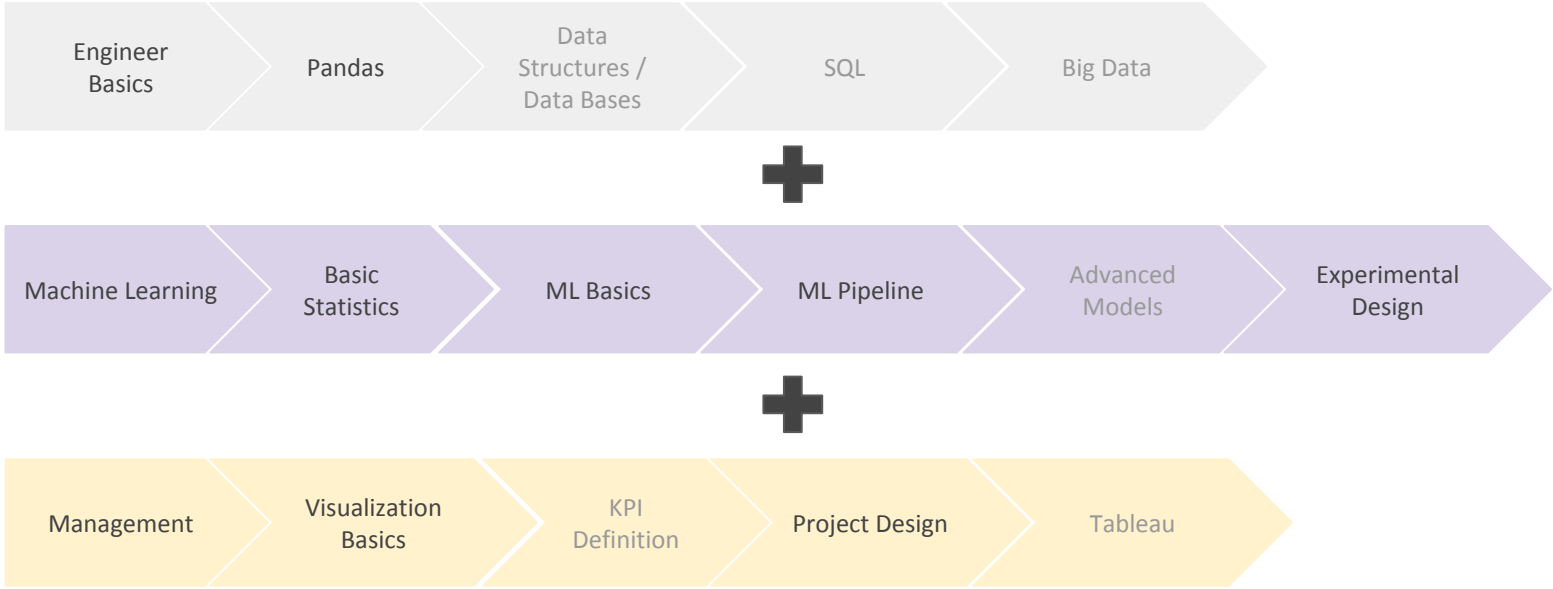
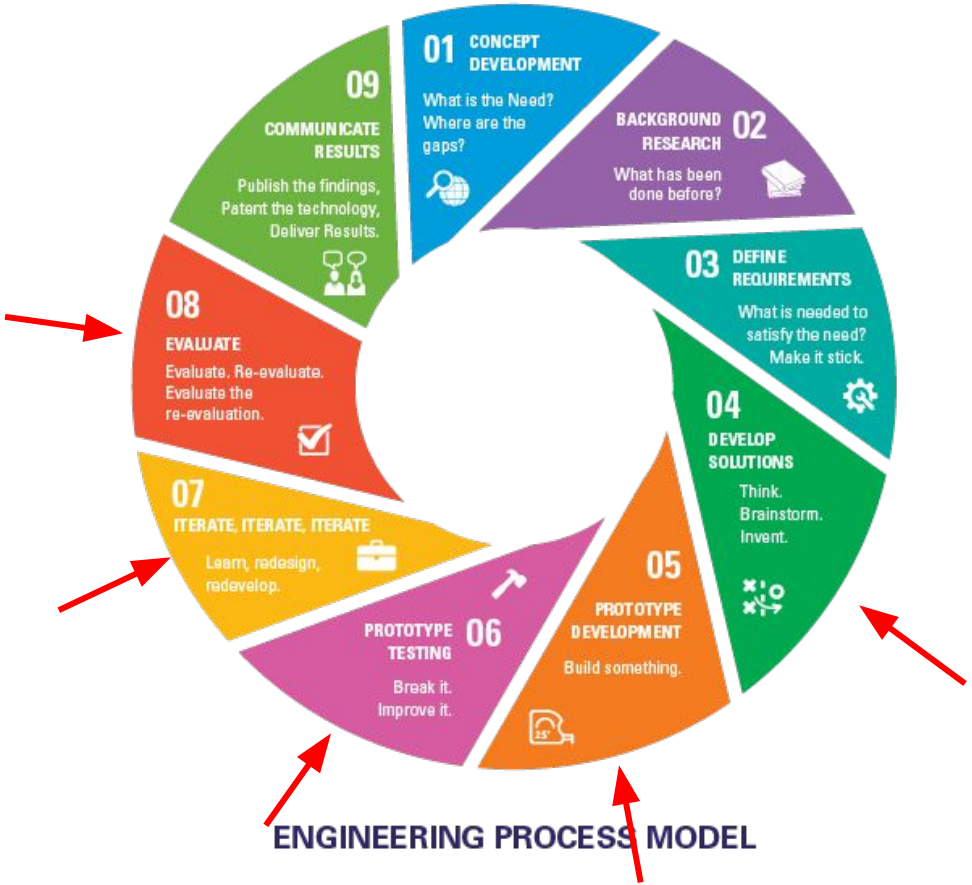
Course Overview



Experimental Design



Machine Learning Pipeline

01 Concept Development: **Recruit Restaurant Visitor Forecasting**

- What is the business question I want to answer?
  - Can I predict the amount of visitors at a given date?
  - Benefits: Plan personal and material required more efficiently
  - **Special interest on predicting most busy times and dates**
- What do I want to predict?
  - The number of visitors (continuous variable)
- Which data do I have available?
  - Weather data
  - Holidays
  - Historical number of visitors for the last 2 years
- Which type of problem do I have?
  - I have a regression problem
  - It requires a time series analysis

02 Background Research: **Recruit Restaurant Visitor Forecasting**

- ❏ Google: Time series analysis python
- ❏ Google: Visitor forecasting
- ❏ Google: How to evaluate time series analysis

03 Define Requirements: **Data Exploration**

- ❏ Is my data clean?
  - ❏ Remove outliers
  - ❏ Identify temporal gaps
  - ❏ Find nulls
- ❏ Do I have enough data?
  - ❏ Ensure that we have a period representative enough (covers different seasons)
  - ❏ Maybe we need external data (i.e. weather, special holidays of the city, events, etc)
- ❏ Is my training data representing well all the situations? Is it balanced?
  - ❏ Ensure that we have similar amount of data points for those more busy days than for those that are not
  - ❏ Overrepresent or remove data points
- ❏ How much training data do I have?
  - ❏ If big data:
    - ❏ Prototype will just require a small part of my dataset
    - ❏ I will need more storage, time for calculations and computational power
    - ❏ Take care that the big data is not just noise
  - ❏ If small data:
    - ❏ Use the minimum amount of features as possible to avoid overfitting (use a dimensions reduction approach)
    - ❏ Communicate to the business responsable on time the limitations on accuracy that it will have

04 Develop Solutions: **Data Mining & Feature Engineering**

- ❏ Remove outliers https://www.kdnuggets.com/2017/02/removing-outliers-standard-deviation-python.html
- ❏ Reduce dimensions if needed http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html#sklearn.decomposition.PCA
- ❏ Balance data http://contrib.scikit-learn.org/imbalanced-learn/stable/auto_examples/index.html
- ❏ Time series analysis / decomposition (remove seasonality and extract residuals) https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/
- ❏ Assign new external features to each event
- ❏ Extract new features out of input data
- ❏ Use unsupervised learning to find patterns and understand better your dataset

05 Prototype Development: **Design and Build Model Training Pipeline**

Classification Problem	Regression Problem	Recommendation System	Time Series Forecast
❑ Decision Tree ❑ Random Forest ❑ Logistic regression (binary) ❑ Xgboost ❑ Support Vector Machines ❑ Neural networks	❑ Linear Regression	❑ Collaborative Filtering ❑ Word embedding ❑ Boltzmann Machines ❑ Clustering	❑ ARIMA Method ❑ EWMA Method ❑ Recurrent Neural Network
Hyperparameters			
❑ Tree max depth ❑ Number of estimators ❑ Minimum samples leaf ❑ C (regularization strength) ❑ Number of iterations ❑ Weights		❑ Algorithm for distance calculation ❑ Number of means ❑ Maximum frequency (tf-idf) ❑ Number of grams ❑ Weights	❑ Smoothing parameter ❑ Weights
Metrics			
❑ Accuracy (binary) ❑ ROC curve (area under the curve) ❑ Precision ❑ Recall ❑ F1-score	❑ RMSE	❑ Similarity ❑ Cosine distance ❑ Euclidean distance ❑ Manhattan distance ❑ Homogeneity	❑ RMSE
A/B Test			

06/07/08 Prototype Testing, Iterate & Evaluate: **Optimize Model**

- ❏ Optimize Model
  - ❏ Grid Search for different hyperparameters
  - ❏ Evaluate performance on the selected metrics

- ❏ Choose Model that Performs Better
  - ❏ Repeat the model optimization for different models
  - ❏ Benchmark models with the same metrics

09 Communicate Results: **A visualization Challenge**

- ❏ **Be as much simple as possible**
- ❏ Do not talk about the model details, don't try to show off!
- ❏ Define business KPIs out of your selected metrics that answer the business question and are understandable
- ❏ Define an A/B Test to evaluate your model once it is deployed
- ❏ Be clear about the limitations and the strengths
- ❏ Use always simple plots to support your results

Now you will design your own experiment / project

01 Concept Development	Business Question	What do I want to predict?	Data Available	Type of Problem
02 Background Research	Which concepts should I google?			
03 Define Requirements	Number of samples	Number of useful features		
04 Develop Solutions (partially today)	New features?	External Data Sources?	Unsupervised learning useful?	
05 Prototype Development	Algorithm	Hyperparameters to optimize	Metrics	
06/07/08 Not today				
09 Communicate Results	Metric to communicate	KPI of impact	Risks when low performance	Type of visualizations