DATA SCIENCE & MACHINE LEARNING COURSE
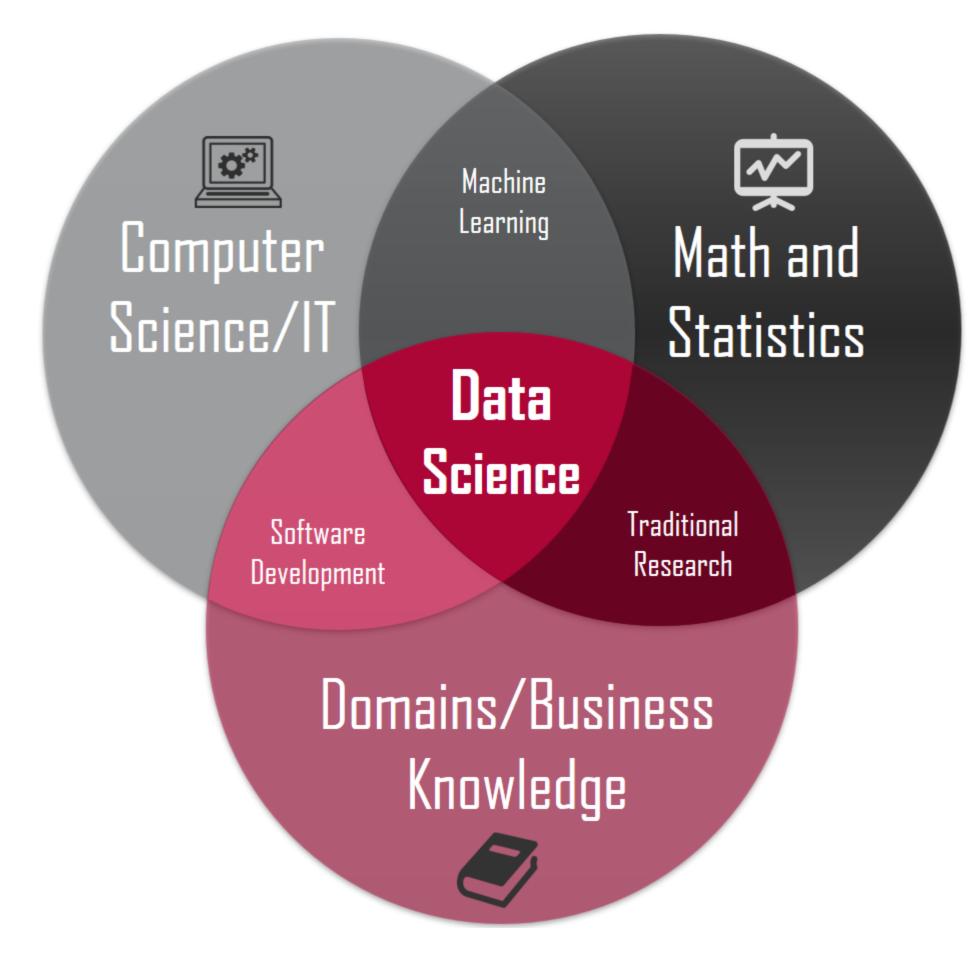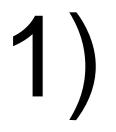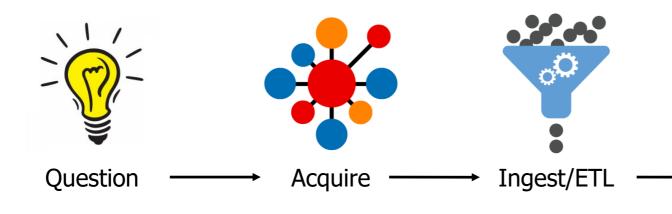
https://www.facebook.com/diceanalytics/
https://pk.linkedin.com/company/diceanalytics
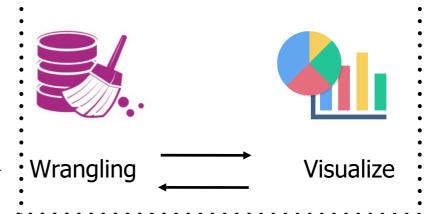
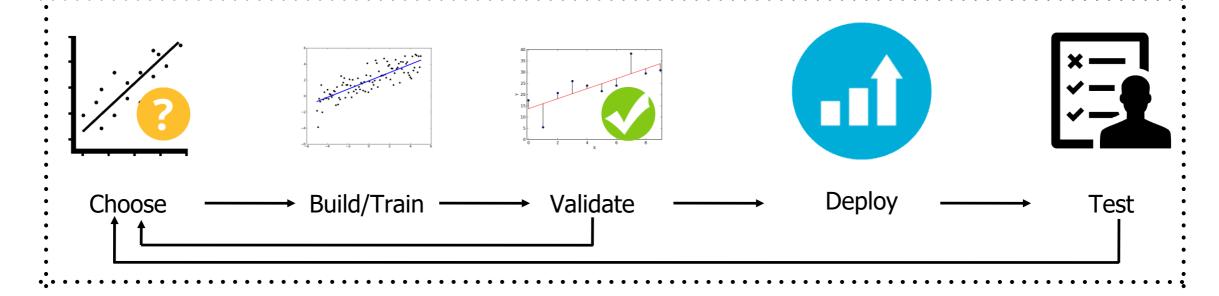**1)**

Exploratory Data Analysis (EDA)

Question → Acquire → Ingest/ETL → Wrangling ⇄ Visualize

2)

Modelling

Choose → Build/Train → Validate → Deploy → Test

3)

STORYTELLING

DICE ANALYTICS

Data Science Deconstructed

Ask a Lot of Questions
- Translate an ambiguous request into a concrete, well-defined problem
- Identify business priorities & strategy decisions that will influence your work

Identify All Available Datasets
- Web, internal/ external databases, etc.

Extract Data Into Usable Format
- .csv, .json, .xml, etc.

Identify Business Insights
- Return back to the business problem

Visualize Your Findings
- Keep it simple & priority-driven

Tell a Clear & Actionable Story
- Effectively communicate to non-technical audiences

Examine Data at a High-Level
- Understand every column; identify errors, missing values & corrupt records

Clean the data
- Throw away, replace, and/or filter corrupt /error prone / missing values

Create a Predictive Model
- Use feature vectors from step #4

Evaluate & Refine Model
- Perhaps return to step #2, 3, or 4

Play Around With the Data
- Split, segment, & plot the data in different ways

Identify Patterns & Extract Features
- Use statistics to identify & test significant variables

THE DATA SCIENCE PROCESS

- 01 Frame the Problem
- 02 Collect Raw Data
- 03 Process the Data
- 04 Explore the Data
- 05 Perform In-Depth Analysis
- 06 Communicate Results

DICE ANALYTICS

SKILLS REQUIRED

01 **FRAME THE PROBLEM**
- **Domain Knowledge** (needs)
- **Product Intuition** (metrics)
- **Business Strategy** (priorities)
- **Teamwork** (people & resources)

02 **COLLECT RAW DATA**
- **Database Management**
  - Systems: mySQL, postgreSQL, Oracle, MongoDB
- **Querying Structured Databases**
  - SQL
- **Retrieving Unstructured Info**
  - Informational Retrieval / Text Mining
- **Distributed Storage**
  - Hadoop HDFS, Spark, Flink

03 **PROCESS THE DATA**
- **Scripting Language**
  - Python or R
- **Data Wrangling & Cleaning**
  - Python "Pandas" library
- **Distributed Processing**
  - Hadoop MapReduce / Spark

04 **EXPLORE THE DATA**
- **Scientific Computing**
  - Python: numpy, matplotlib, scipy, pandas
- **Inferential Statistics**
  - hypothesis testing
  - correlation vs. causation
- **Experimental Design**
  - A/B tests, controlled trials

05 **PERFORM IN-DEPTH ANALYSIS**
- **Machine Learning**
  - Supervised / Unsupervised algorithms
  - Contextual pros/cons)
- **ML Tools Library**
  - Python: scikit-learn
- **Advanced Math**
  - Linear Algebra & Multivariate Calculus

06 **COMMUNICATE RESULTS**
- **Business Acumen**
  - Non-technical terminology
- **Data Visualization Tool(s)**
  - Tableau, D3.js, Google visualize, matplotlib, ggplot, seaborn
- **Data Storytelling**
  - presenting & speaking
  - reporting & writing

DICE ANALYTICS