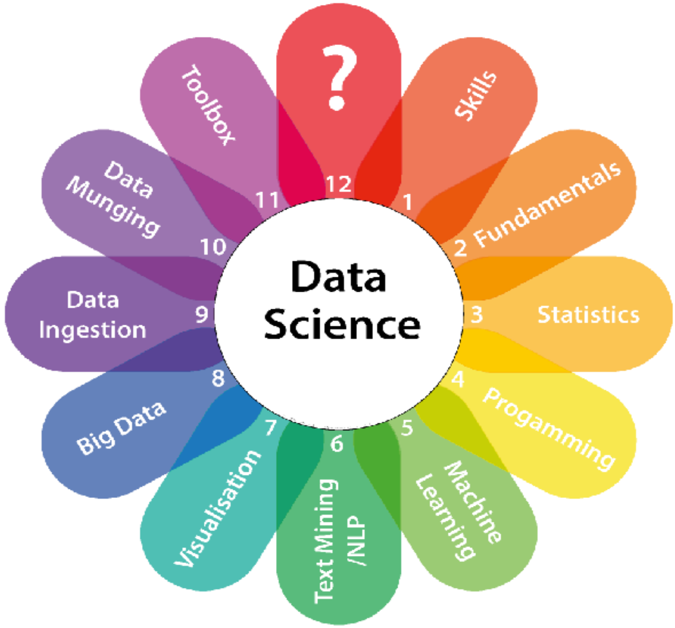
CS500-Data Science Tools and Technique
Introduction to Data Science

Bahar Ali
PhD Scholar,
National University Of Computer and Emerging Sciences,
Peshawar.

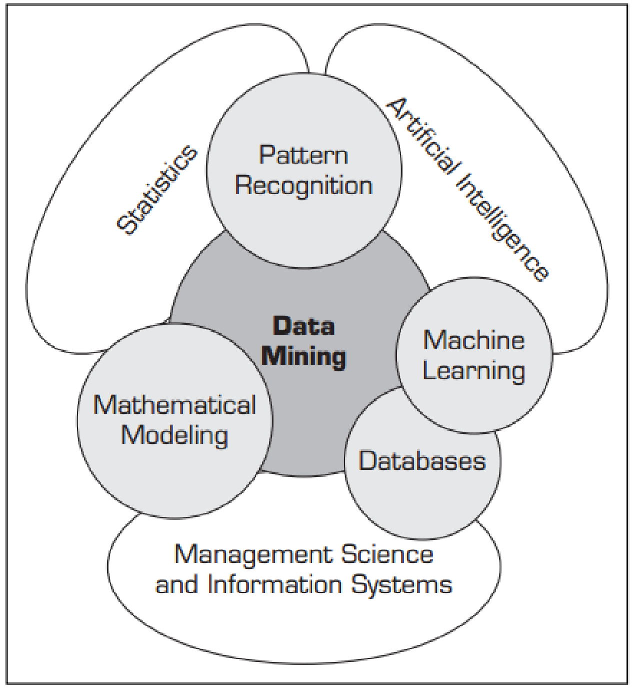
What is Data Science?

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data. It basically includes asking right questions, obtaining data, understanding data, building predictive models and Generating Visualizations



Data Mining

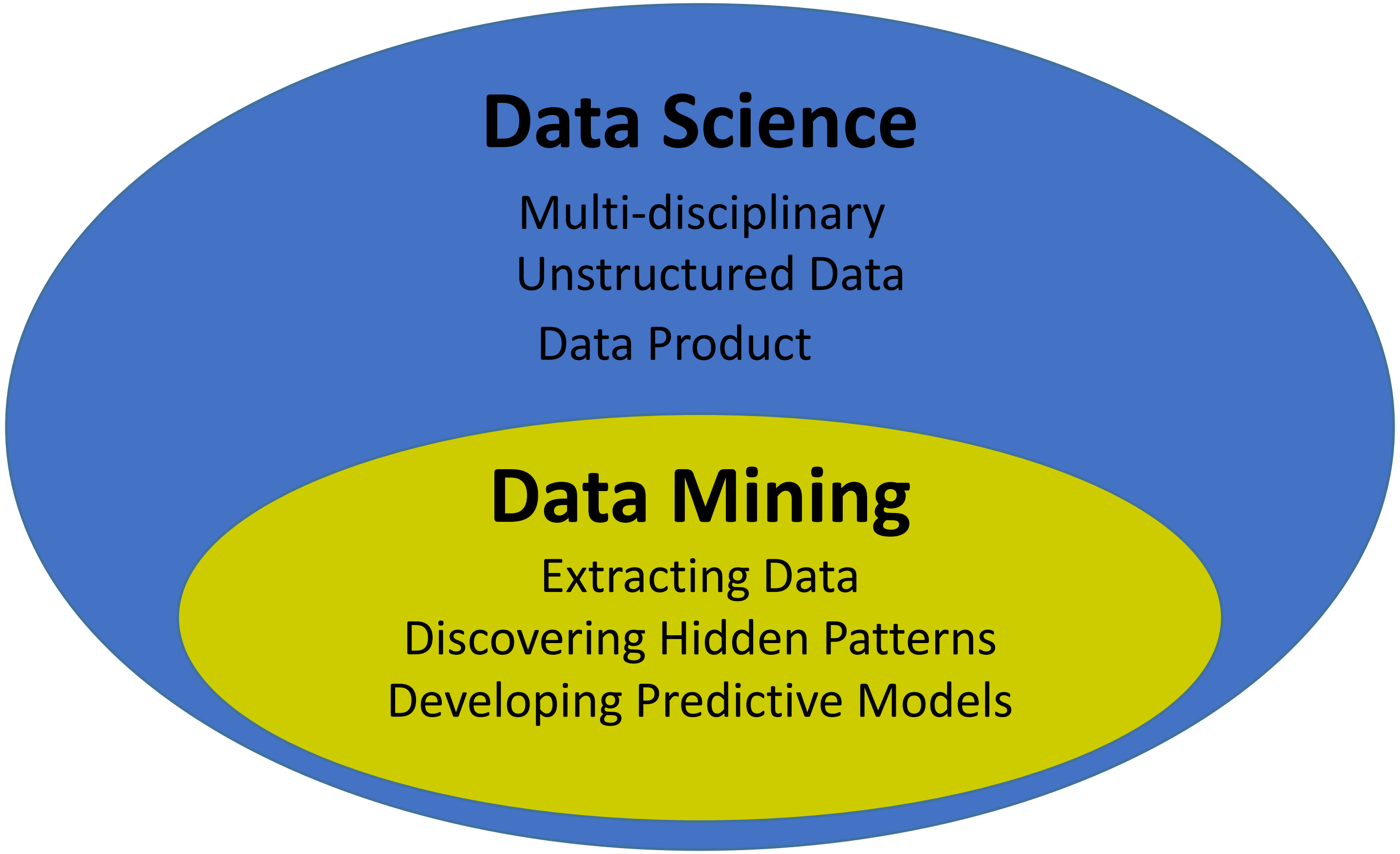
Data Mining is about finding the trends in a data set and using these trends to identify future patterns. It is an important step in the Knowledge Discovery process. It basically includes extraction of data, understanding of data, data preparation, and pattern analysis.



Data Mining vs Data Science

Basis for comparison	Data Mining	Data Science
What is it?	A technique	An area
Focus	Business process	Scientific study
Goal	Make data more usable	Building Data-centric products for an organization
Output	Patterns	Varied
Purpose	Finding trends previously not known	Social analysis, building predictive models, unearthing unknown facts, and more
Vocational Perspective	Someone with a knowledge of navigating across data and statistical understanding can conduct data mining	A person needs to understand Machine Learning, Programming, info-graphic techniques and have the domain knowledge to become a data scientist
Extent	Data mining can be a subset of Data Science as Mining activities are part of the Data Science pipeline	Multidisciplinary – Data Science consists of Data Visualizations, Computational Social Sciences, Statistics, Data Mining, Natural Language Processing, etc.
Deals with (the type of data)	Mostly structured	All forms of data – structured, semi-structured and unstructured
Other less popular names	Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction	Data-driven Science

Data Mining vs Data Science



Data Science

Multi-disciplinary

Unstructured Data

Data Product

Data Mining

Extracting Data

Discovering Hidden Patterns

Developing Predictive Models

Data Science Life cycle & Process

There are countless interpretations to data science lifecycle,
Here is a simple breakdown

1. Business Understanding/Asking Right Questions
2. Data Preparation
3. Understanding Data
4. Building Predictive Models
5. Generating Visualizations

Business Understanding

Success of any project depends on the quality of questions asked for the dataset.

- To acquire the correct data, we should be able to understand the business.
- Asking questions about dataset will help in narrowing down to correct data acquisition.
- Every decision made in the company is supported by concrete data and that it is guaranteed to achieve results.
- In data science project, it is critical that you understand the problem you are trying to solve.

Business Understanding (continue)

- Identify the variables that need to be predicted.
- Data science is used to answer five types of questions:
  - How much or how many? (Regression)
  - Which category? (Classification)
  - Which group? (Clustering)
  - Is this unusual/ odd? (Anomaly detection)
  - Which option should be taken? (Recommendation)

Data Preparation

- One of the important tasks when analyzing data is to collect and prepare the data in a format appropriate for analysis of the samples.
- The most common steps for data preparation involve the following operations.
  - **Obtaining the data:**
    - The most traditional way of obtaining data is directly from files (CSV, TSV, TXT, JSON etc.).
    - Data can be obtained from database
    - Data might be obtained by scraping the web.
    - Another popular option to gather data is connecting to Web APIs.
    - Big Data Engineering

Data Preparation (continue)

- **Parsing the data:** The right parsing procedure depends on what format the data are in i.e. Plain text, CSV, XML, HTML, JSON etc.

- **Cleaning the data:** Survey responses and other data files are almost always incomplete. Sometimes, there are multiple codes for things such as, not asked, did not know, and declined to answer so there are always errors.
  - Remove or ignore incomplete records
  - Handle Nan/ Empty Records (Usually filled with mean values)
  - Removing noise
  - Format the data
  - Normalization (change values of numeric columns in a dataset to a common scale)

Data Preparation (continue)

- **Building data structures:** Once you read the data, it is necessary to store them in a data structure. If the data fit into the memory, building a data structure is usually the way to go. If not, usually a database is built, which is an out-of-memory data structure. Data structures which are usually used in data science are;
  - Array
  - Vector
  - Matrix
  - List
  - Tuple
  - Dictionary
  - Data Frame
  - Series

Understanding Data

Once your data is ready to be used, data scientists need to explore the data.

- Inspect the data and its properties.
- Identify type of data (numerical data, categorical data, ordinal and nominal data etc.)
- Calculate descriptive statistics to present data in a more meaningful way.
- Testing significant variables often is done with correlation.
- For features selection usually used the evaluation techniques like Pearson Correlation, Heat Map and Extra Tree Classifier

Building Predictive Models

Predictive Modeling is an essential part of Data Science. It is one of the final stages of data science to generate predictions.

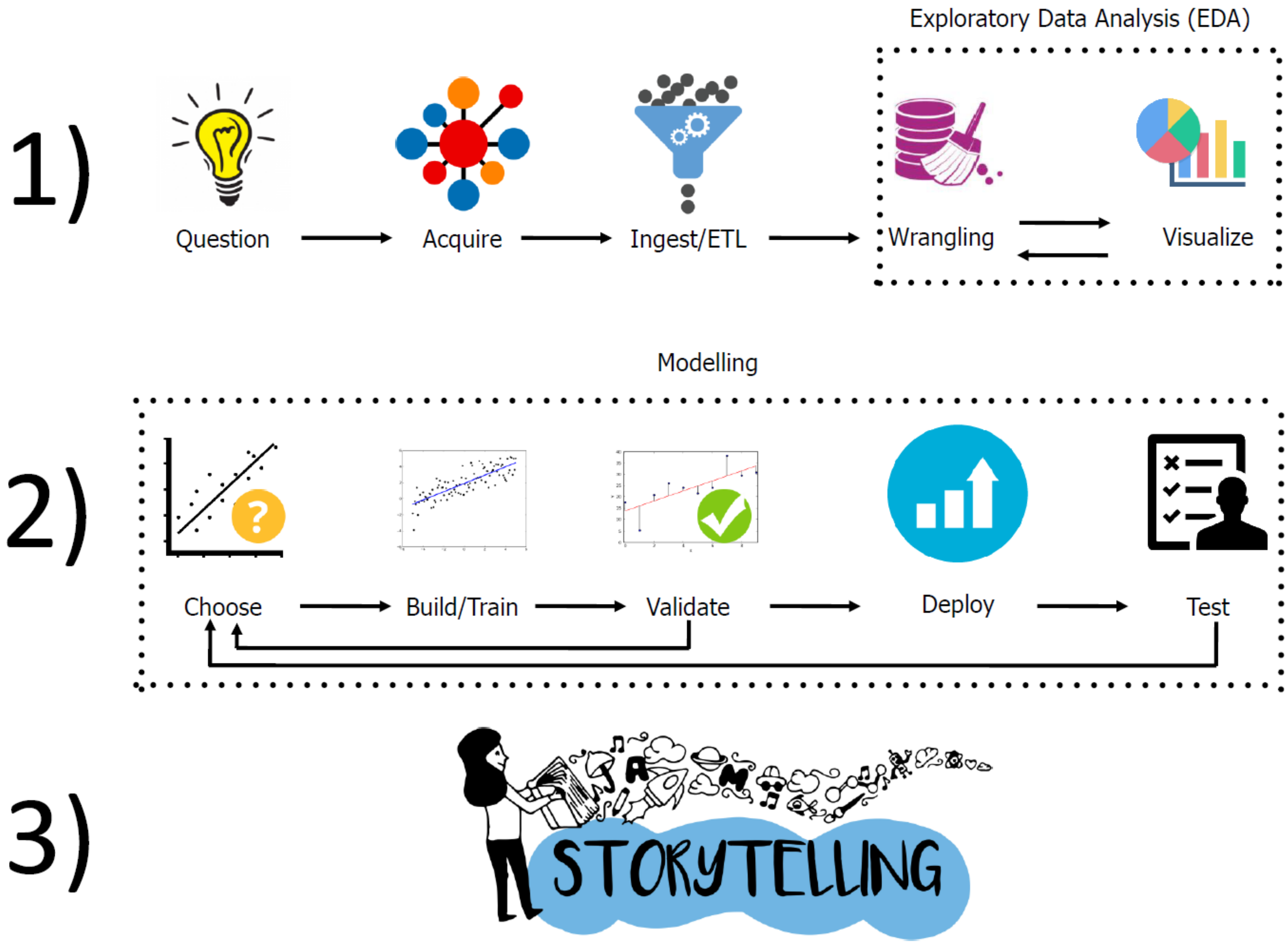
- Based on the business problem the right models should be selected.
- It is essential to identify the type of problem,
    - Is it a classification problem
    - Regression problem
    - Time series forecasting
    - Clustering problem.

- Once problem type is sorted out model could be implemented.
- After the modelling process, model performance measurement is required.
- Precision, recall, F1-score, ROC and AUC can be used for classification problem
- RMSE is the mostly widely used metric be used for regression problem

Generating Visualizations

The most crucial step of a data science project is to interpret the models and visualize the findings.

- The predictive power of a model lies in its ability to generalize unseen future data.
- To answer the business questions asked when the project is started.
- Visualize your findings in such a way that is useful to the organization, or else it would be pointless to your stakeholders.



1)

Exploratory Data Analysis (EDA)

Question → Acquire → Ingest/ETL → Wrangling ⇄ Visualize

2)

Modelling

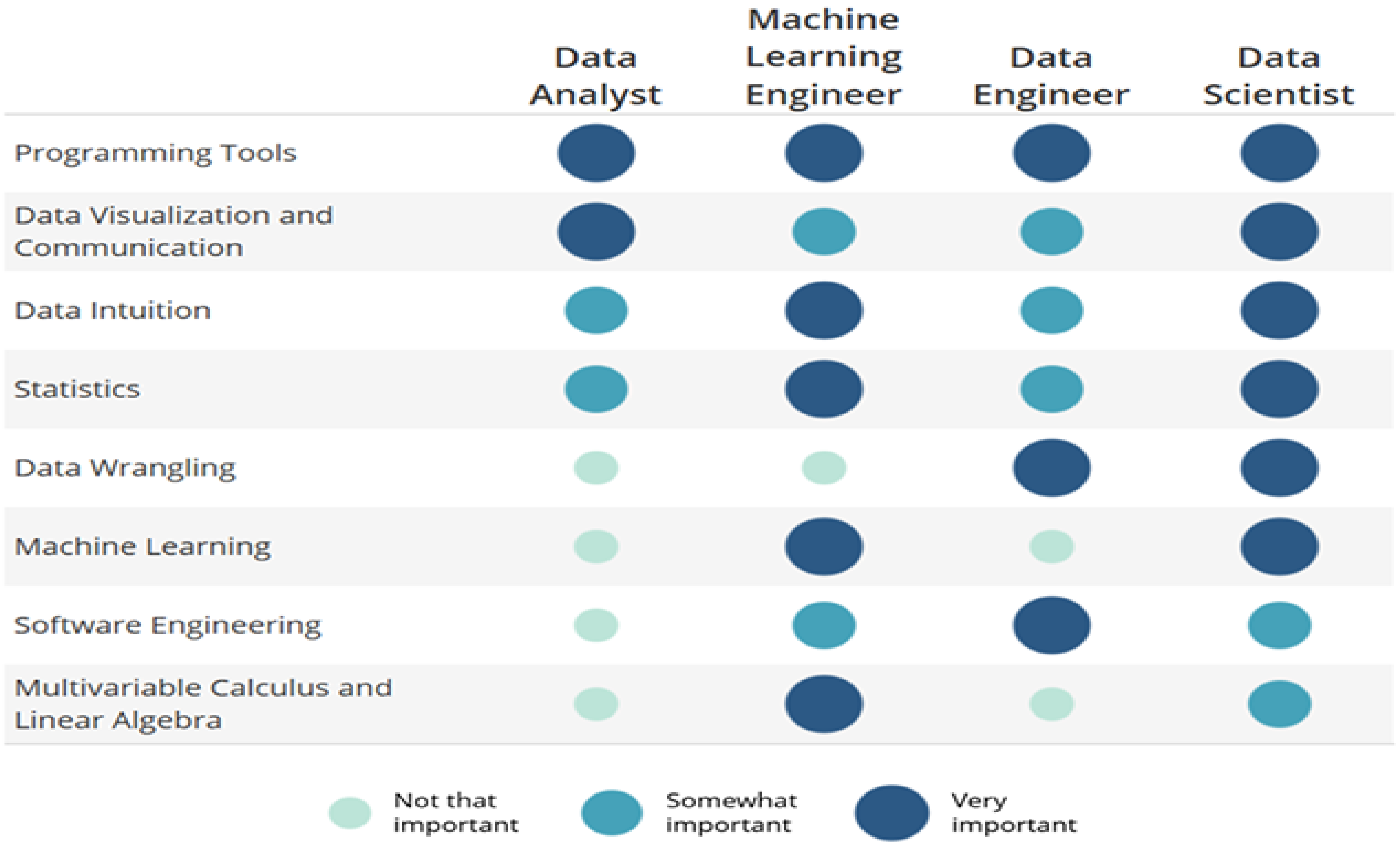
Choose → Build/Train → Validate → Deploy → Test

3)

Skills Required

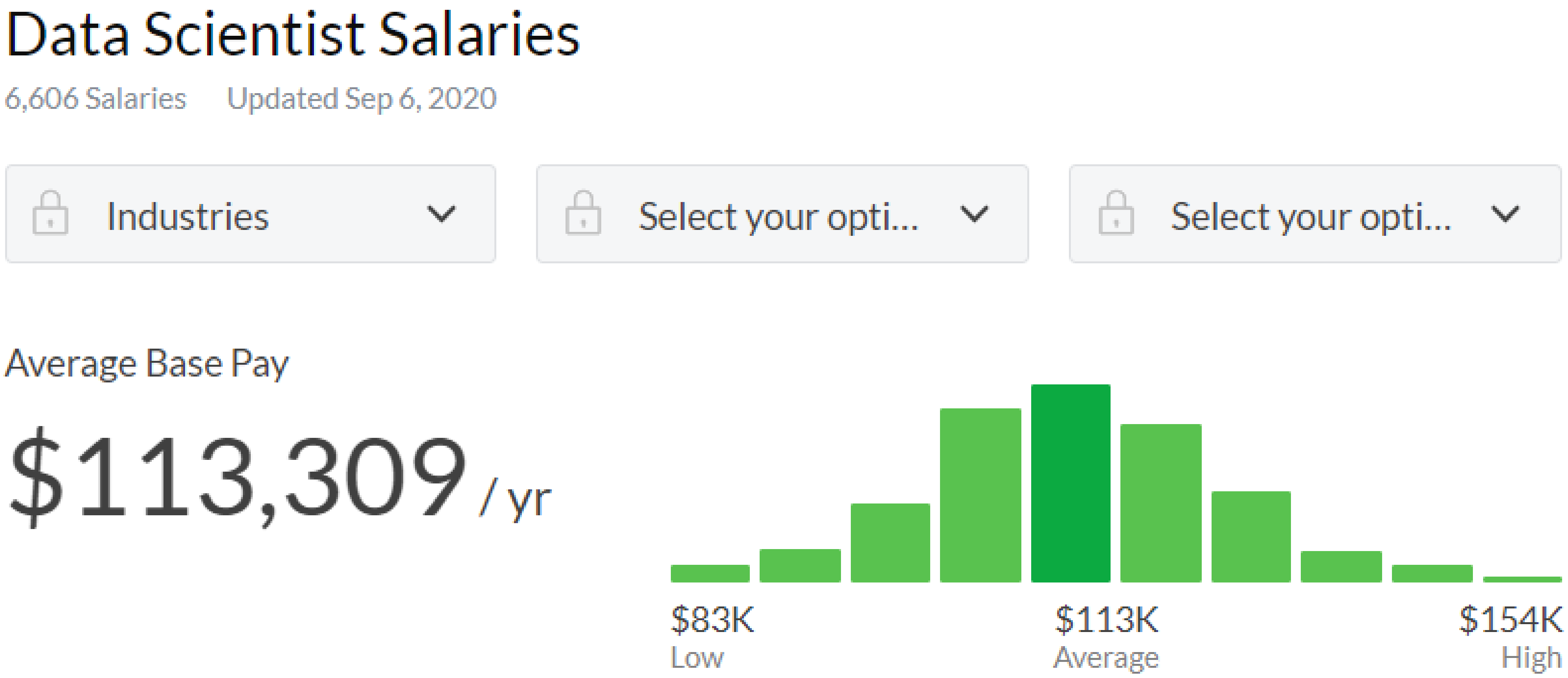
1. Business Understanding
   - Domain Knowledge (Needs)
   - Product Intuition (Metrics)
   - Business Strategy (priorities)
   - Teamwork (People and resources)
2. Data Preparation
   - Database Management (MySQL, PostgreSQL)
   - Querying Structured Database (SQL)
   - Retrieving unstructured data (Text Mining)
   - Distributed Storage (Hadoop HDFS, Spark)
   - Scripting Language (Python, R)
   - Data Wrangling (Python Pandas library)
   - Distribute programing paradigms (MapReduce/ Spark)
3. Understanding Data
   - Scientific computing (numpy, matplotlib, scipy, pandas)
   - Inferential statistics (Hypothesis, correlation)
   - Experimental Design (A/B Test)
4. Building Predictive Models
   - Machine Learning (Supervised/Unsupervised)
   - ML Tools Library (scikit-learn)
   - Advanced Math (Linear Algebra, Calculus)
5. Generating Visualization
   - Business Acumen (Non-technical terminology)
   - Data Visualization Tools (Tableau, matplotlib, seaborn)
   - Data Storytelling (Presenting and speaking, reporting and writing)

Data science roles and skills



	Data Analyst	Machine Learning Engineer	Data Engineer	Data Scientist
Programming Tools	●	●	●	●
Data Visualization and Communication	●	◑	◑	●
Data Intuition	◑	●	◑	●
Statistics	◑	●	◑	●
Data Wrangling	○	○	●	●
Machine Learning	○	●	○	●
Software Engineering	○	◑	●	◑
Multivariable Calculus and Linear Algebra	○	●	○	◑

○ Not that important ◑ Somewhat important ● Very important



Data Scientist Salaries

6,606 Salaries   Updated Sep 6, 2020

Industries ⌄   Select your opti... ⌄   Select your opti... ⌄

Average Base Pay

$113,309 / yr

https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm