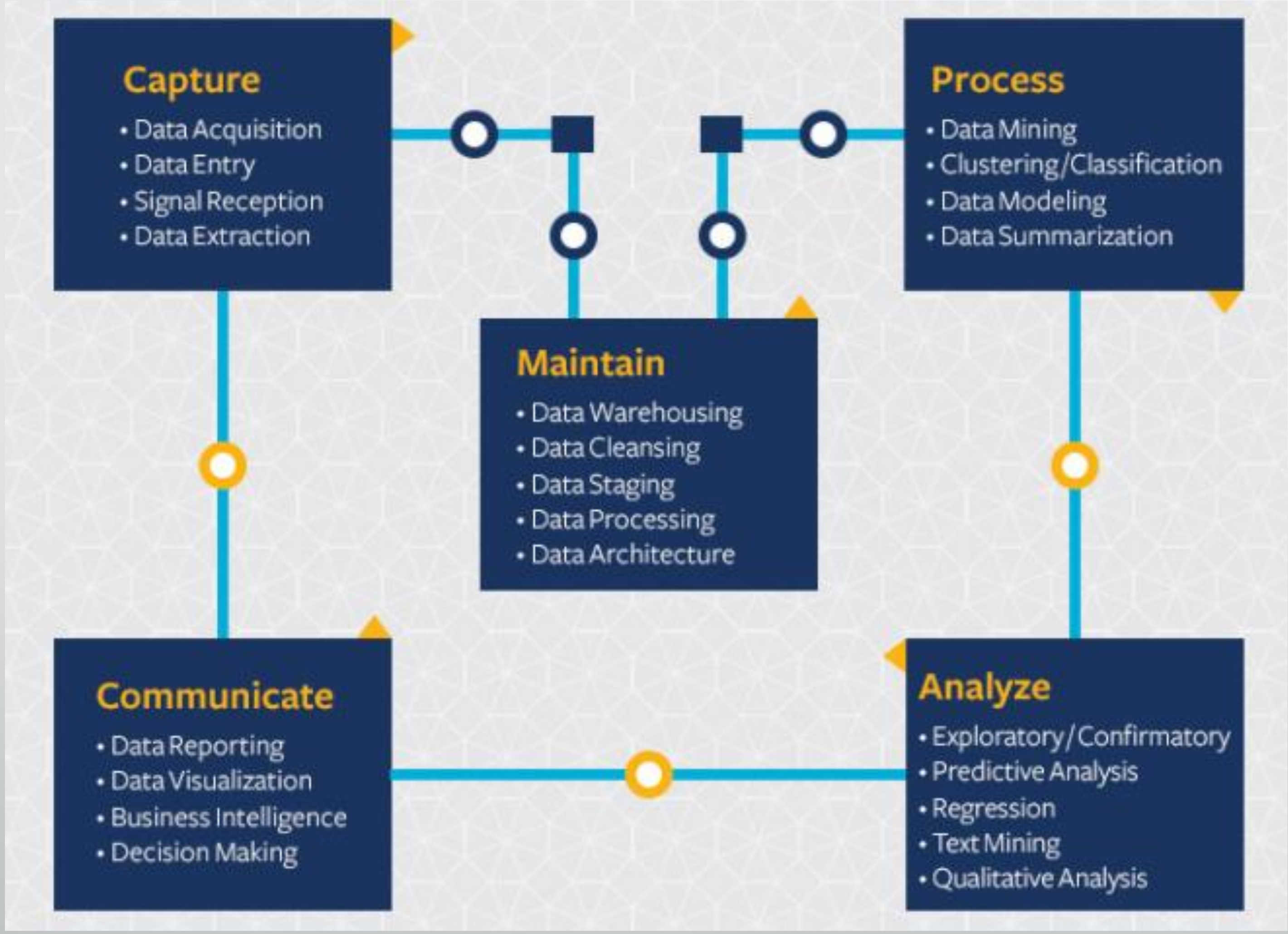
An Overview of Data Science

What is Data Science?

- Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.
- Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems to perform tasks that ordinarily require human intelligence.

DataRobot, Inc. (2021). What Is Data Science Definition | Learn Why Data Science Is Important (datarobot.com)

Tools of Data Science



Capture

- Data Acquisition
- Data Entry
- Signal Reception
- Data Extraction

Process

- Data Mining
- Clustering/Classification
- Data Modeling
- Data Summarization

Maintain

- Data Warehousing
- Data Cleansing
- Data Staging
- Data Processing
- Data Architecture

Communicate

- Data Reporting
- Data Visualization
- Business Intelligence
- Decision Making

Analyze

- Exploratory/Confirmatory
- Predictive Analysis
- Regression
- Text Mining
- Qualitative Analysis

School of Information, University of California, Berkeley (2021). [What is Data Science? | The Data Science Career Path (berkeley.edu)]

What is Big? – Too big to fit in memory

- Bit -> 1 or 0
- Byte -> 8 bits
- Kilobyte -> 1024 bytes
- Megabyte -> 1024 Kilobytes
- Gigabyte -> 1024 Megabytes

- Terabytes -> 1024 Gigabytes
- Petabytes -> 1024 Terabytes
- Exabytes -> 1024 Terabytes
- Zettabytes -> 1024 Terabytes
- Yottabytes -> 1024 Exabytes

2.5 quintillion bytes of data are created daily or approximately 2.44 Exabytes (These are 2019 numbers) (Micro Focus. (2021))

Where does the data come from?

- 550 new social media users per minute
- 474,000 tweets per minute
- 4,333,560 videos viewed on YouTube per minute
- 300 hours of video uploaded to YouTube every minute
- 100,000,000 photos and videos uploaded to Instagram every day
- Every minute Facebook receives 293,000 status updates, 136,000 photos
- Facebook incurs 13 trillion likes daily
- 3.5 Billion Goggle searches every minute
- 100,000,000 text and in-app messages per minute

IoT - The Internet of Things

- 3.6 Billion in use worldwide in 2020
- Each phone has up to 14 sensors
- Temperature sensors
- Humidity Sensors
- Pressure sensors
- Proximity sensors
- Level sensors
- Accelerometers
- Gyroscopes
- Gas sensors
- Infrared sensors
- Optical sensors

Industries applying IoT solutions

- Agriculture
- Manufacturing
- Oil and Gas
- Utilities
- Shipping and receiving
- Smart Buildings
- Transportation
- Logistics

- Healthcare
- Retails
- Finance
- All levels of Government

Data Mining

- Data mining is the process of discovering patterns, correlations and anomalies in large datasets to predict outcomes.
- The term was coined in the 1990s, however, it is built on a long history of statistics, machine learning, and artificial intelligence.
- Data mining allows us to discover knowledge that is obscured by chaotic occurrences and repetitive noise.
- Once we discover what data is relevant to the pursuit of knowledge, we can predict relevant outcomes.
- Data mining is an interactive process with the system tweaked to improve the relevancy of the discoveries over time.

Data Warehouses

- For decades, across industries, data was local to each department.
- Each department handled their own data management systems.
- Most attempts to combined data across department failed miserably or incurred great cost-overruns.
- Attempts to centralize data did not fair much better.
- Eventually the concept of a data warehouse began to emerge.
- Local data emigrated to a central repository through a process called extraction, transformation, load, ETL/ELT
- The centralization of the data allowed for data mining and reporting.

Data Warehouses

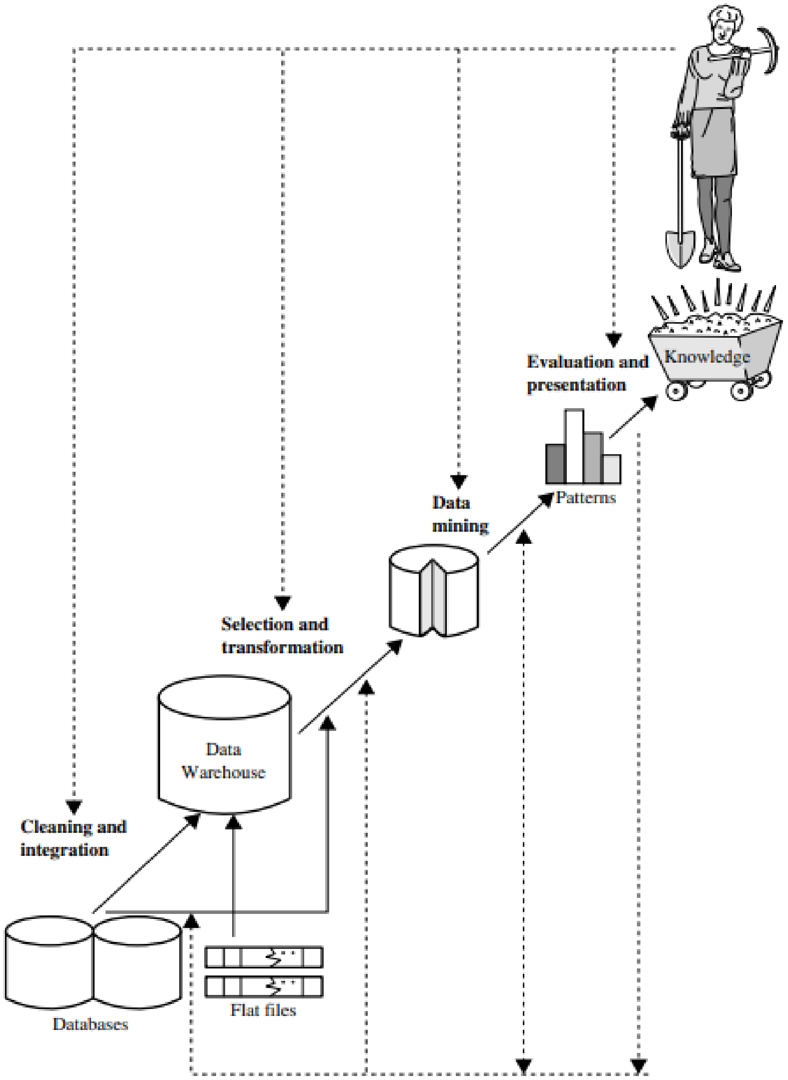
- Data warehouses allow for subject-oriented analysis, payroll, inventory, or functional areas.
- The data housed in the warehouse is integrated, the transformation process allows for disparate data types to house consistently.
- The data is non-volatile, stable and remains unchanging.
- Data warehouses are time-variant, which allows for analytics to explore data over time.

Data Lakes

- The transformation process can be a disadvantage to the data integrity.
- The data lake is loaded with the raw data.
- Data is transformed when extracting content from the lake to the required structure needed by the analytics, maintaining the raw data integrity.

KDD - Knowledge Discovery from Data

- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation



Han, J. Kamber, M. Pei, J. (2012). *Data Mining Concepts and Techniques*. Morgan Kaufmann Publishers, Waltham, MA

Business Intelligence

- Business intelligence provides the tools to analyze historical, current and predictive views of operations at the strategic and tactical levels.
- Business intelligence allow for organizations to perform effective market analysis.
- Business intelligence allows organizations to preform customer feedback analysis on their products.
- Business intelligence allows organizations to discover their strengths and weaknesses compared to their competition.
- Business intelligence allows organizations to manage customer expectations
- Business intelligence allows organizations make smart business decisions.

Machine Learning

- In 1959, Arthur Samuel defined machine learning as the field of study that gives computers the ability to learn without being explicitly programmed.
- In 1997, Tom Mitchell applied an engineering twist to the definition, stating a computer program is said to learn from experience E with respect to some task T, and some performance measure P, if its performance on T, as measured by P, improves with experience E.
- Machine learning is designed to be a stand-alone system that does not require any modification once application is in place.

Machine Learning

- System specifics rules are defined.
- A dataset is defined that contains to desired functionality.
- The dataset is divided into training and testing elements, 70/30.
- The system is trained by analyzing the training set.
- The system is verified by analyzing the results of the testing set.
- If the testing results are unacceptable, the system rules are modified.
- The training and testing phases are again launched
- The results of the testing set are again analyzed.
- If the results are acceptable, the system is set to launch.

Forms of Machine Learning

- Supervised learning – The data contains labels
- Unsupervised learning – The data contains no labels
- Reinforcement leaning – The system is "rewarded" for making the correct classification.
- Semi-Supervised learning – Some of the data contain labels
- Self-Supervised learning – Unlabeled data is used to solve an alternative or prefix task.
- Multi-Instance Learning – The data is not labeled but arraigned in collections or bags or containers that are labeled.

Forms of Machine Learning

- Inductive learning – Drawing general conclusions from specific observations.
- Deductive learning – Using general rules to determine specific outcomes.
- Transductive learning – Specific observations are used to determine specific outcomes.
- Multi-Task learning – A supervised technique where the model is trained on multiple tasks in a manner that improves the performance across all tasks when compared to any single task.

Forms of Machine Learning

- Active learning – Where a model is allowed to query outside sources to resolve ambiguities during the learning process.
- Online learning – The observations provided over time and the probability distribution of the observations are expected to change over time. Reacting in real-time to changes in streaming data.
- Transfer learning – Where the model is trained for a specific task and then all or parts of the model transition to related tasks.
- Ensemble learning- Where two or more models are trained with the same data and the predicted results of the models are combined.

Examples of Automated Decision Making

- 2020 Mars Perseverance Rover – Picture perfect most difficult Mars landing
- Successful flights of the Ingenuity Mars Helicopter
- Autonomous vehicles
- Recommender systems
- Autocompletion assistances on email
- Spam detection

Deep Learning

- Deep learning is a subset of machine learning that incorporates artificial neural networks
- Neural networks are a series of algorithms that endeavor to analyze relationships through processes that mimic the human brain.

PlayGround.Tensorflow.org

- Lets go play!