MIT-WPU

॥ विश्वशान्तिर्ध्रुवं ध्रुवा ॥

Dr. Vishwanath Karad

MIT WORLD PEACE UNIVERSITY | PUNE

TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

UNIT - I
Introduction to Data Science

Data Science

T. Y. BTECH

SCHOOL OF COMPUTER ENGINEERING AND TECHNOLOGY

Prepared By Shilpa Sonawani

Data Science: Why all the Excitement?

Exciting new effective applications of data analytics

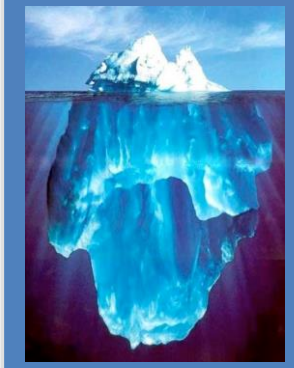
e.g., **Google Flu Trends (GFT):**
A web service for detecting outbreaks two weeks ahead of CDC data since 2008

CDC: Change Data Capture
A process of identifying and capturing changes in a database and then delivering those changes in real-time to a downstream process or system.

Ebola Virus: New models are estimating which cities are most at risk.
Prediction model is built on various data sources, types and analyses

“Big Data” Sources

It’s All Happening On-line



Every:
Click
Ad impression
Billing event
Fast Forward, pause,…
Server request
Transaction
Network message
Fault
…

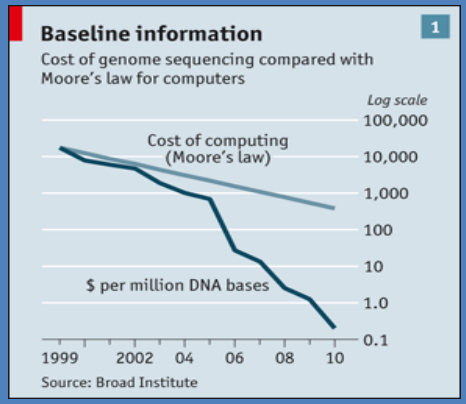
User Generated (Web & Mobile)




….
.


Internet of Things / M2M

Health/Scientific Computing

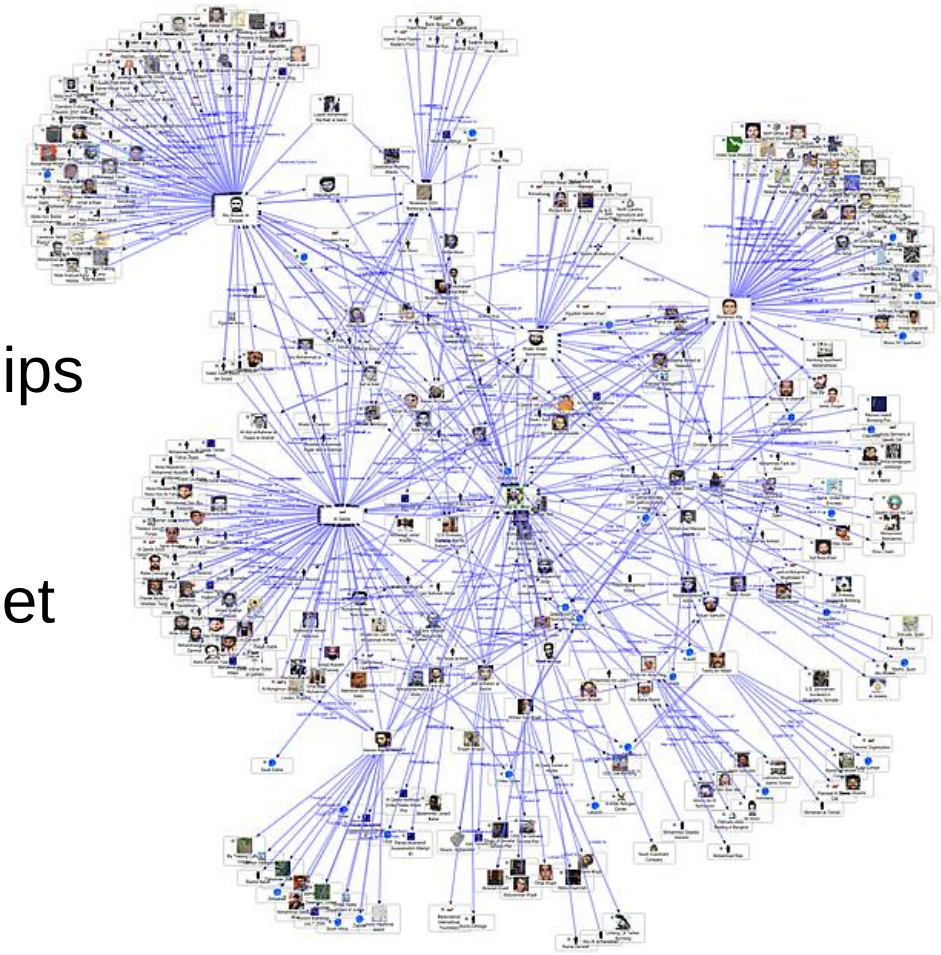


Graph Data

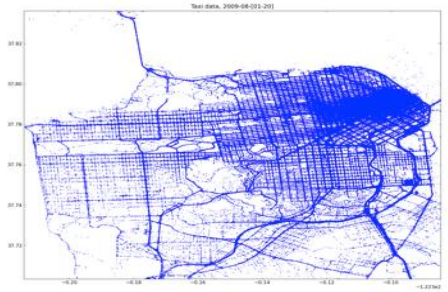
Lots of data generated has a graph structure:

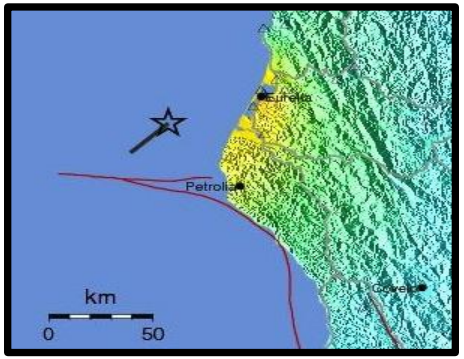
- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- …

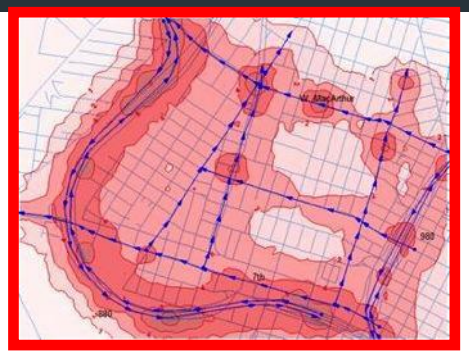
Some of these graphs can get quite large (e.g., Facebook* user graph)



What can you do with the data?







Crowdsourcing + Sensing + Physical Modeling + Data assimilation (Understanding)

From Alex Bayen, UCB

7V’s of Big Data

- Raw Data: Volume
- Change over time: Velocity
- Data types: Variety
- Data Quality: Veracity
- Information for Decision Making: Value
- Change in Data: Variability
- Presentation of Data: Visualization

DATA SCIENCE – WHAT IS IT?

Data Science – A Definition

Data Science is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.

Ben Fry's Model: **Visualizing Data Process**

1. Acquire
2. Parse (Analyze and put in proper format)
3. Filter
4. Mine (Discovering patterns or knowledge from large datasets)
5. Represent
6. Refine
7. Interact

Jeff Hammerbacher's Model

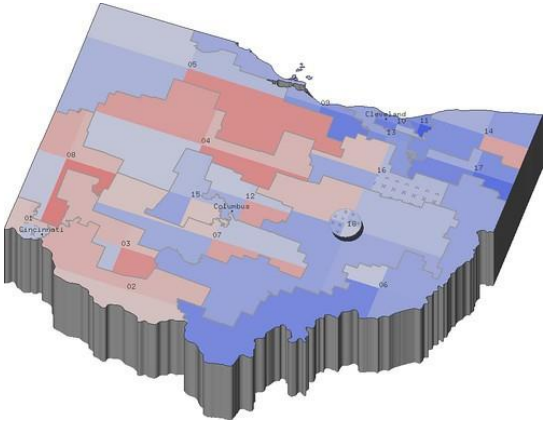
1. Identify problem

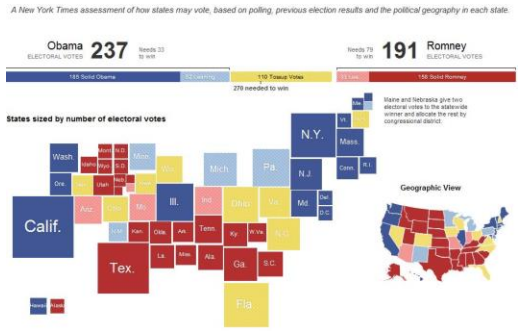
2. Instrument data sources

3. Collect data

4. Prepare data (integrate, transform, clean, filter, aggregate)

5. Build model





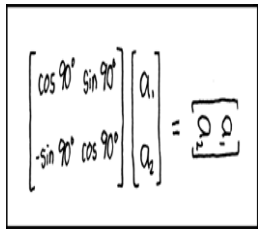
6. Evaluate model

7. Communicate results

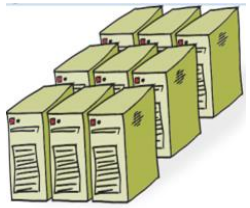
Data Scientist's Practice

Digging Around in Data

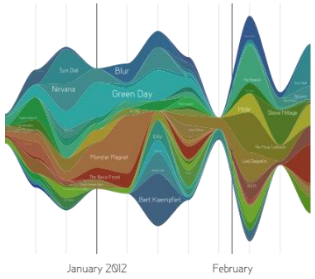
Clean, prep



Hypothesize Model

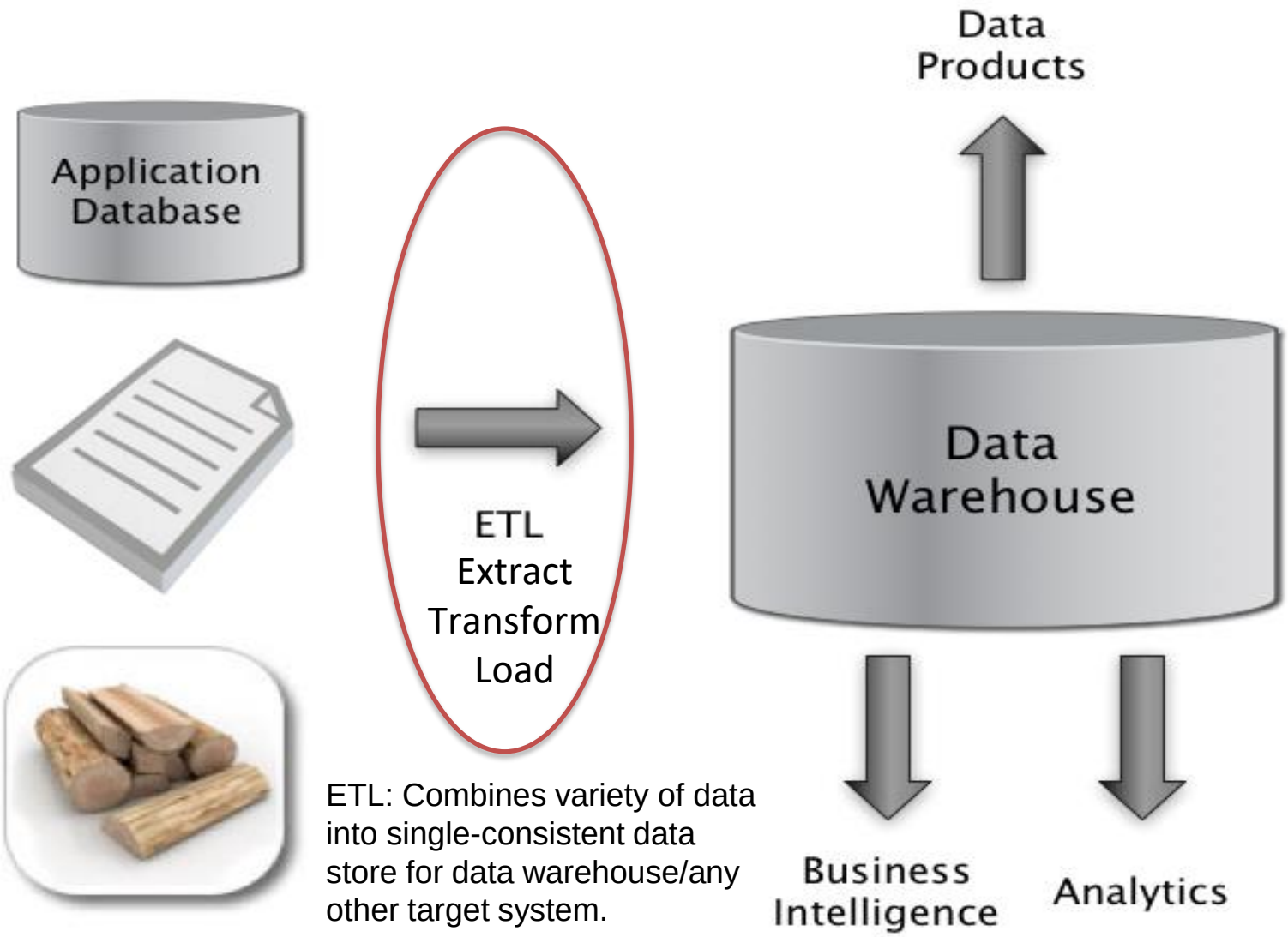


Large Scale Exploitation



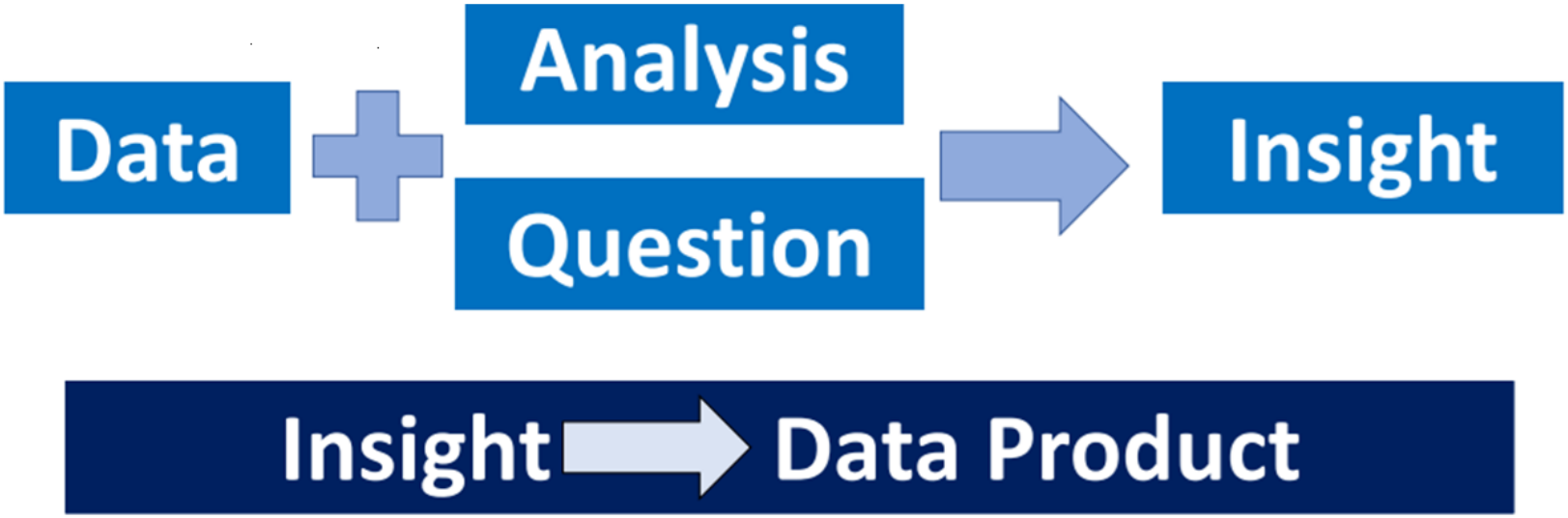
Evaluate Interpret

The Big Picture

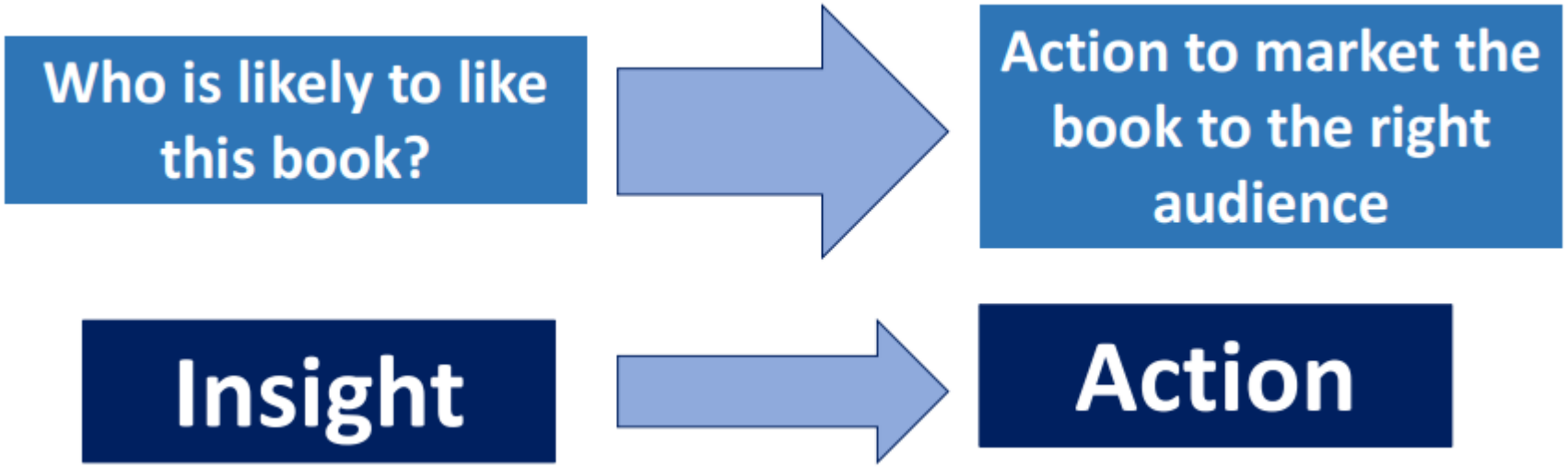


ETL: Combines variety of data into single-consistent data store for data warehouse/any other target system.

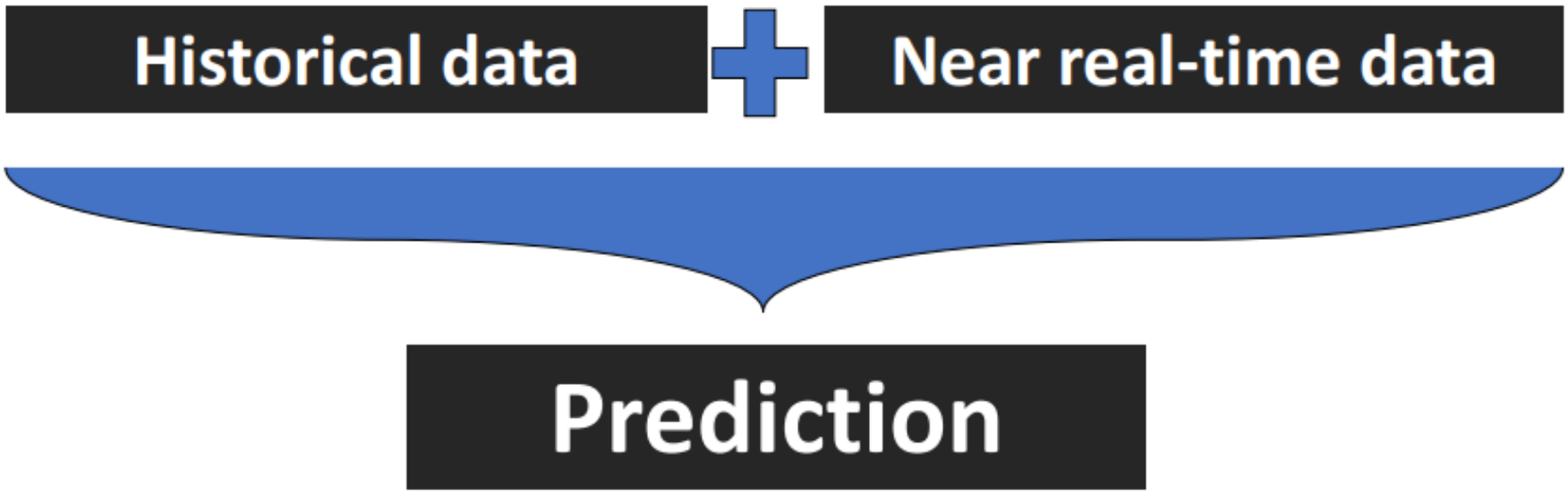
Data Science: Getting Value out of Data



Data Science: Getting Value out of Data

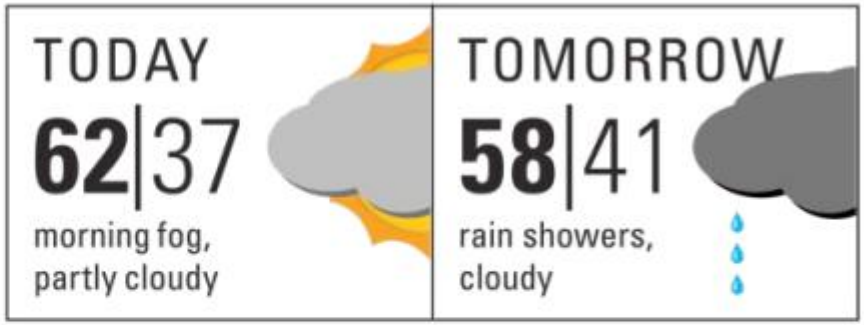


Data Science: Getting Value out of Data



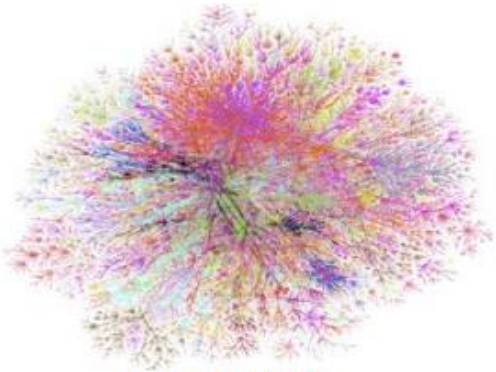
Data Science: Getting Value out of Data

Prediction



Action

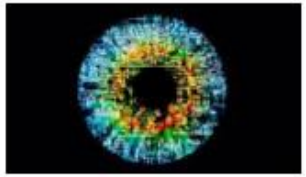
Why the Increased Interest in Data Science?



BIG DATA

COMPUTING AT SCALE

New era of data science!



Many dynamic data-driven applications

Computer-Aided Drug Discovery

Smart Manufacturing

Personalized Precision Medicine

Smart Cities

Smart Grid and Energy Management

Disaster Resilience and Response

Data Science Applications

E-commerce

- Identifying Consumers
- Recommending Products
- Analyzing Reviews

Manufacturing

- Predicting Potential Problems
- Monitoring Systems
- Automating Manufacturing Units
- Maintenance Scheduling
- Anomaly Detection

Banking

- Fraud Detection
- Credit Risk Modeling
- Customer Lifetime Value

Healthcare

- Medical Image Analysis
- Drug Discovery
- Bioinformatics
- Virtual Assistants

Transport

- Self Driving Cars
- Enhanced Driving Experience
- Car Monitoring System
- Enhancing the safety of passengers

Finance

- Customer Segmentation
- Strategic Decision Making
- Algorithmic Trading
- Risk Analytics

Applications

- Climate change and weather
- Traffic control
- Agriculture
- Personalised healthcare
- Twitter data analysis
- Facebook information links
- Pollution and Weather

Contrast: Databases

	Databases	Data Science
Data Value	“Precious”	“Cheap”
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP*
Realizations	SQL	NoSQL: MongoDB, CouchDB, Hbase, Cassandra,…

ACID = Atomicity, Consistency, Isolation and Durability (properties of databases)

CAP = Consistency, Availability, Partition Tolerance

Databases

Databases:

Atomicity: When a database processes a transaction, it is either fully completed or not executed at all. If a single portion of the transaction fails, the whole transaction will fail.

Consistency: Is a property ensuring that only valid data following all rules and constraints is written in the database.

Isolation: Is a property that guarantees the individuality of each transaction, and prevents them from being affected from other transactions.

Durability: Durability is a property that enforces completed transactions, guaranteeing that once each one of them has been committed, it will remain in the system even in case of subsequent failures

Data science:

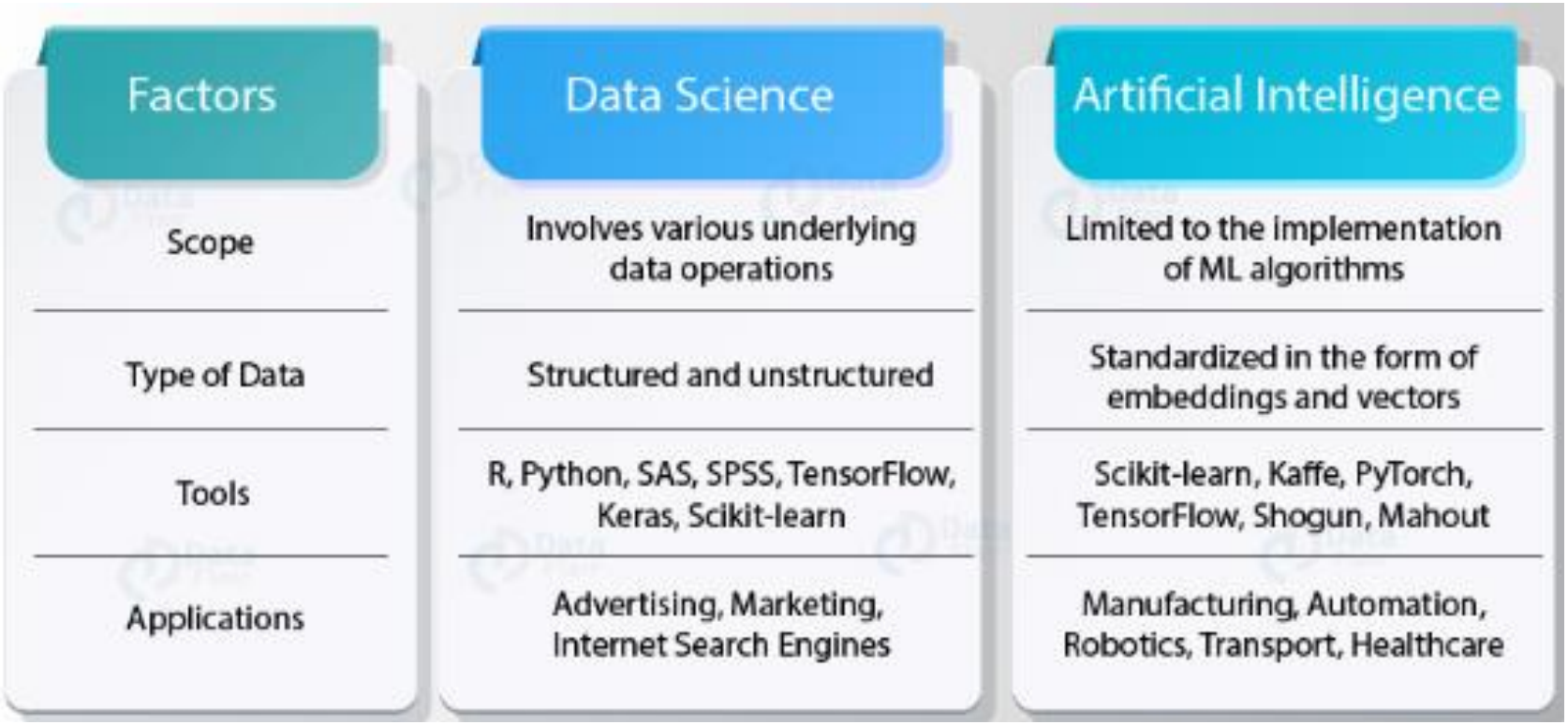
Consistency. All reads receive the most recent write or an error.
Availability. All reads contain data, but it might not be the most recent
Partition tolerance. The system continues to operate despite network failures

In case of network failure/slow network: Provides only Availability or Partition tolerance

Contrast: BI

Area	BI Analyst	Data Scientist
Focus	Reports, KPIs, trends	Patterns, correlations, models
Process	Static, comparative	Exploratory, experimentation, visual
Data sources	Pre-planned, added slowly	On the fly, as-needed
Transform	Up front, carefully planned	In-database, on-demand, enrichment
Data quality	Single version of truth	“Good enough”, probabilities
Data model	Schema on load	Schema on query
Analytics	Retrospective, Descriptive	Predictive, Prescriptive, Preventative

Contrast: AI

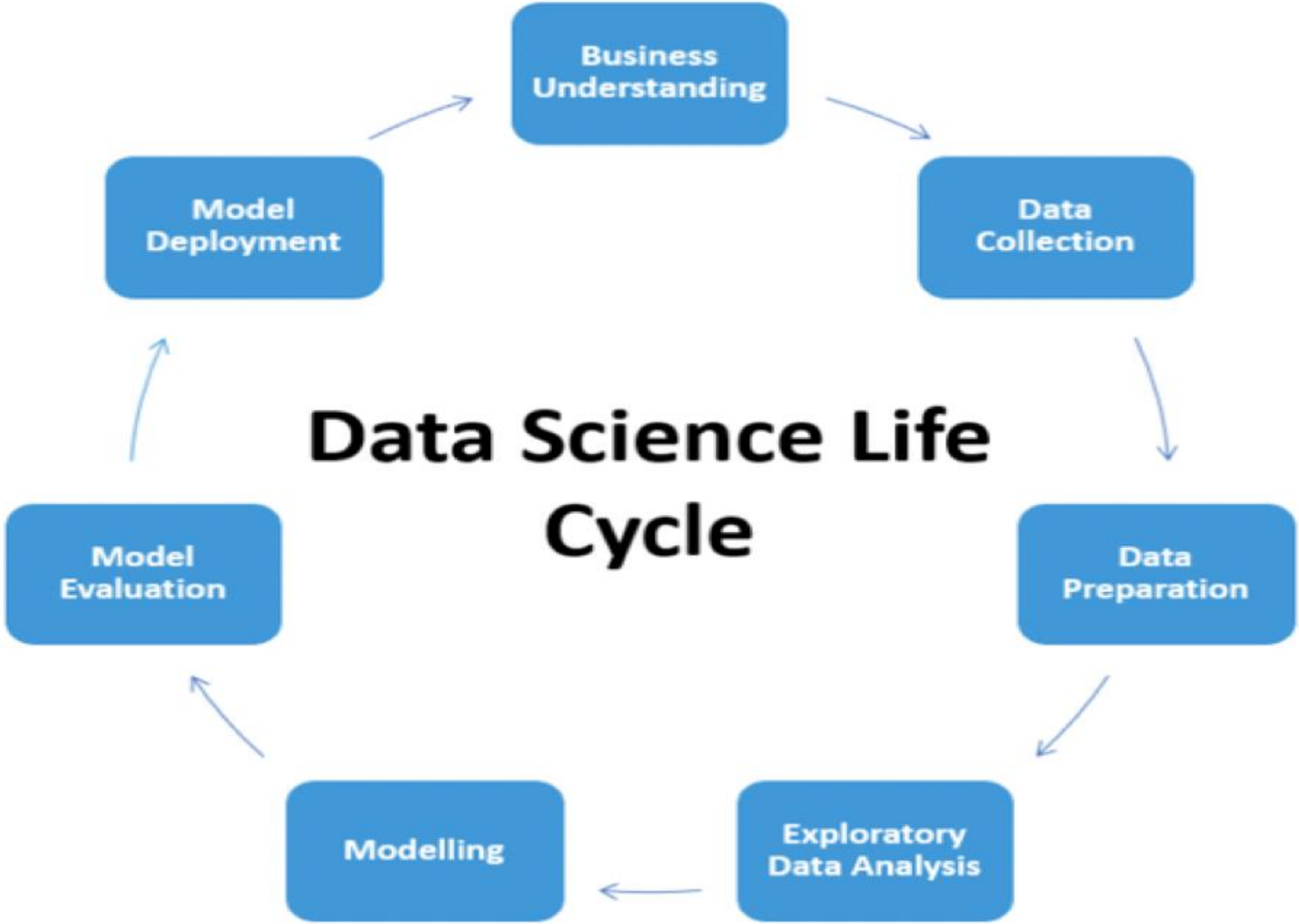


Factors	Data Science	Artificial Intelligence
Scope	Involves various underlying data operations	Limited to the implementation of ML algorithms
Type of Data	Structured and unstructured	Standardized in the form of embeddings and vectors
Tools	R, Python, SAS, SPSS, TensorFlow, Keras, Scikit-learn	Scikit-learn, Kaffe, PyTorch, TensorFlow, Shogun, Mahout
Applications	Advertising, Marketing, Internet Search Engines	Manufacturing, Automation, Robotics, Transport, Healthcare

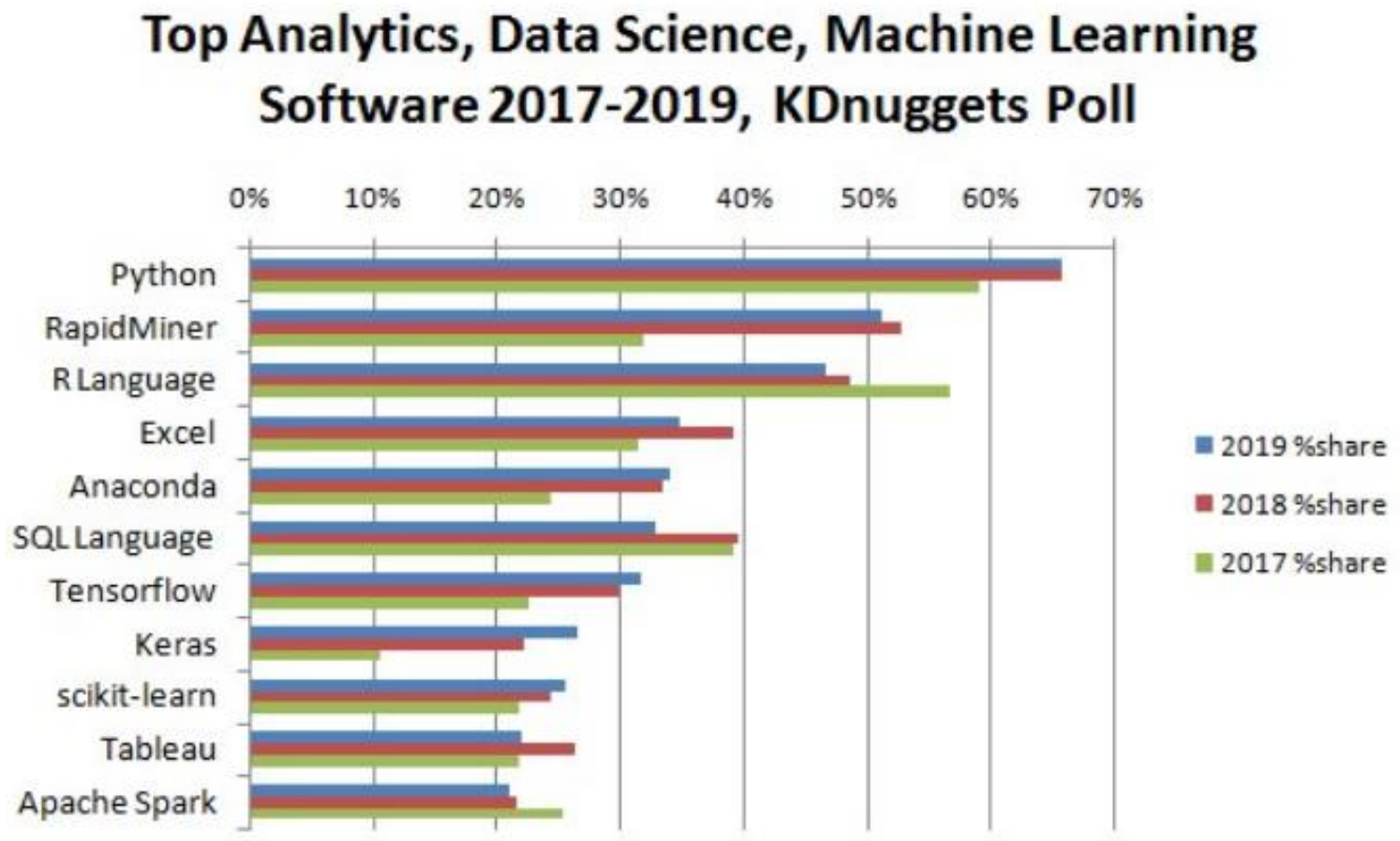
Modern Data Science Skills

- Programming in Python
- Statistics
- Machine Learning
- Scalable Big Data Analysis

DATA ANALYST SKILLS	DATA SCIENTIST SKILLS
Data Mining	Data Mining
Data Warehousing	Data Warehousing
Math, Statistics	Math, Statistics, Computer Science
Tableau and Data Visualization	Tableau and Data Visualization/Storytelling
SQL	Python, R, JAVA, Scala, SQL, Matlab, Pig
Business Intelligence	Economics
SAS	Big Data/Hadoop
Advanced Excel skills	Machine Learning

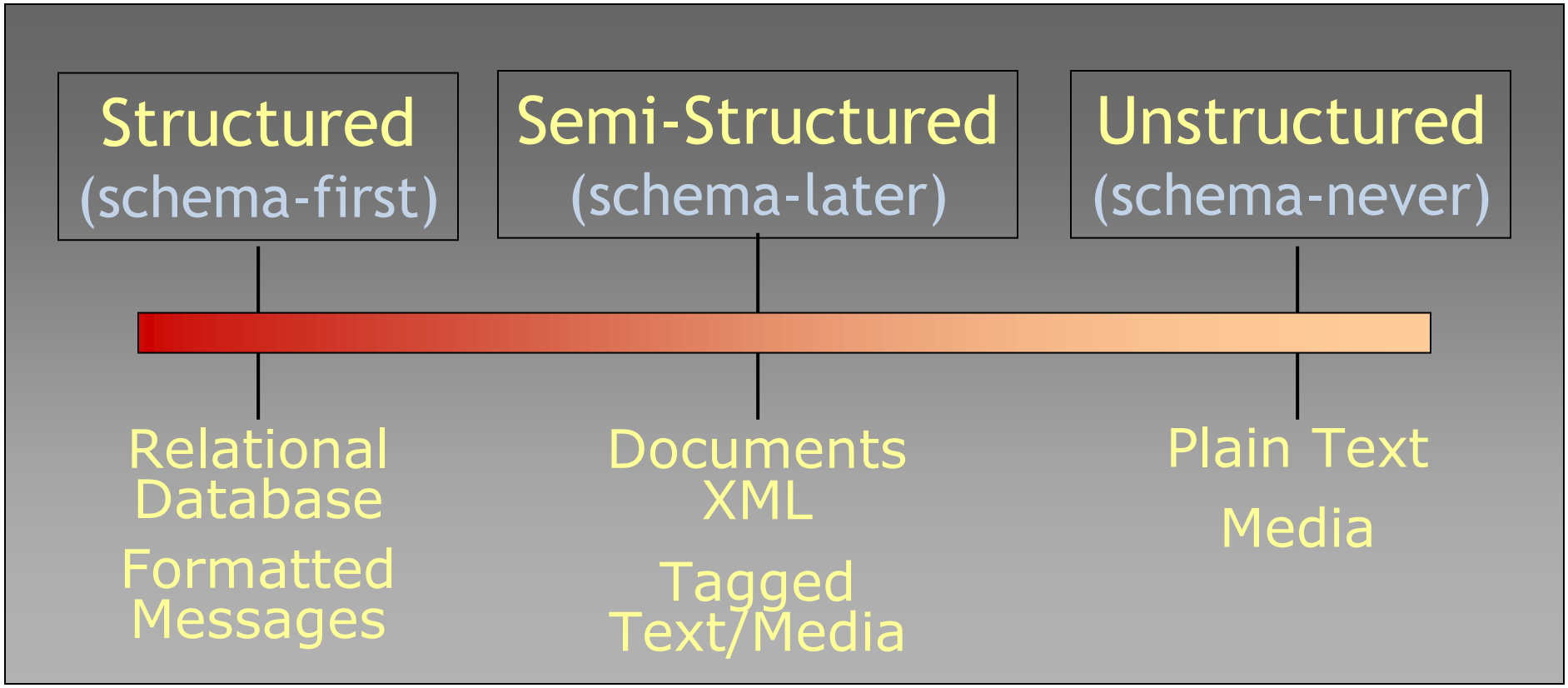

Business Understanding
Data Collection
Data Preparation
Exploratory Data Analysis
Modelling
Model Evaluation
Model Deployment
Data Science Life Cycle

Why Python for Data Science???



https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

The Structure Spectrum



Key Concept: Structured Data

A *data model* is a collection of concepts for describing data.

A *schema* is a description of a particular collection of data, using a given data model.

NumPy/Python

- NumPy is a Python library used for working with arrays.
- It also has functions for working in domain of linear algebra, fourier transform, and matrices.
- NumPy stands for Numerical Python.
- In Python we have lists that serve the purpose of arrays, but they are slow to process.
- NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.
- The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.
- Arrays are very frequently used in data science, where speed and resources are very important
- NumPy arrays are stored at one continuous place in memory unlike lists, so processes can access and manipulate them very efficiently

Pandas

- Pandas officially stands for ‘Python Data Analysis Library’, **THE** most important Python tool used by Data Scientists today.
- Pandas is an open source Python library that allows users to explore, manipulate and visualise data in an extremely efficient manner. It is literally Microsoft Excel in Python.
- It is easy to read and learn
- It is extremely fast and powerful
- It integrates well with other visualisation libraries
- Pandas can take in a huge variety of data, the most common ones are csv, excel, sql or even a webpage.

Applications

Amazing real-time Data Science Applications:

Recommendation- Most of the apps and websites like Amazon, YouTube, Flipkart, etc. give recommendation over as per the viewer's interest. Online music applications like Spotify give recommendations as per your taste in music. So these are good examples of data science recommendation applications.

Search Results- Machine Learning algorithms used to find the most relevant search for Google search engines. Such an algorithm used for the most visited sites on google chrome.

Intelligent Assistant- Google assistant, Siri are examples of intelligent assistants. The advanced machine learning algorithm converts voice input into text output. These smart assistants recognize the voice and provide the required information in both voice and text outputs.

Autonomous driving vehicles- Automobile companies like Waymo and Tesla looking for the next generation of autonomous vehicles. 3D images were taken by the cameras and the information provided to the algorithms for further processing.

Piracy Detection- YouTube is an example of piracy detection using machine learning algorithms. Due to the big database, copied contents cannot be detected manually. So it helps to detect and remove the copied content to reduce human efforts.

Image Recognition- Facebook is the application that uses image recognition by data science and machine learning for the friend suggestion. Even Google lens uses an image recognition algorithm to provide the related information to you.

Data Cleaning-Dirty Data

- The Statistics View:
  - There is a process that produces data
  - We want to model ideal samples of that process, but in practice we have non-ideal samples:
    - **Distortion** – some samples are corrupted by a process
    - **Selection Bias** - likelihood of a sample depends on its value
    - **Left and right censorship** - users come and go from our scrutiny
    - **Dependence** – samples are supposed to be independent but are not (e.g. social networks)

Dirty Data

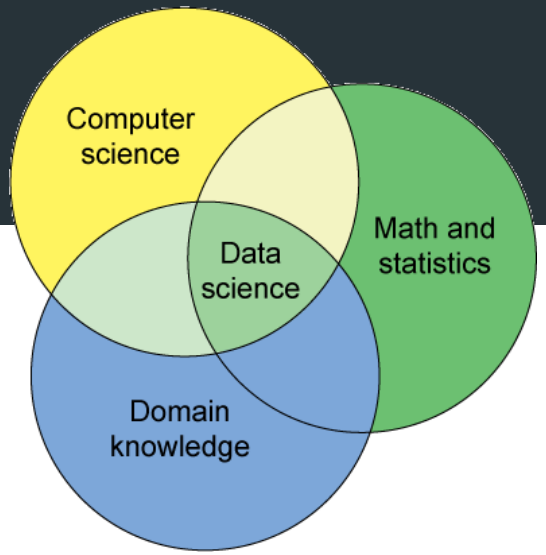
- The Database View:
  - Some of the values are missing, corrupted, wrong, duplicated
  - Results are absolute (relational model)

  - You get a better answer by improving the quality of the values in your dataset

Dirty Data

- The Domain Expert's View:
  - This Data Doesn't look right
  - This Answer Doesn't look right
  - What happened?

Dirty Data



- The Data Scientist's View:
  - Some Combination of all of the above

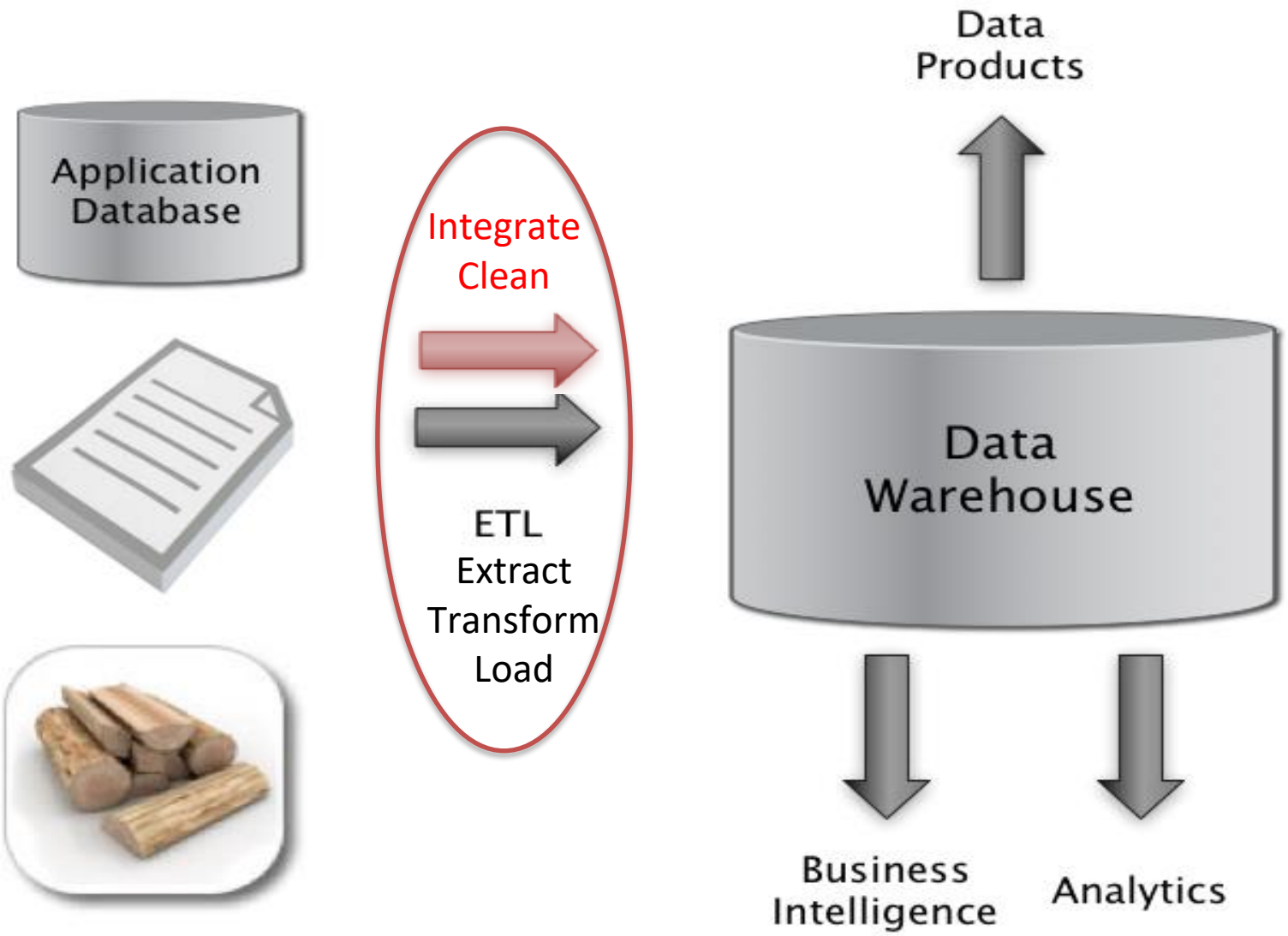
Solution:

- **Data Preprocessing**
- **Data Wrangling**

Data Quality Problems

- (Source) Data is dirty on its own.
- Transformations corrupt the data (complexity of software pipelines).
- Data sets are clean but integration (i.e., combining them) screws them up.
- “Rare” errors can become frequent after transformation or integration.
- Data sets are clean but suffer “bit rot” (the tendency for digital information to degrade or become unusable over time)
  - Old data loses its value/accuracy over time
  - Redundant obsolete trivial (ROT) data refers to the digital information that has little or no business value to the organization but is still stored
- Any combination of the above

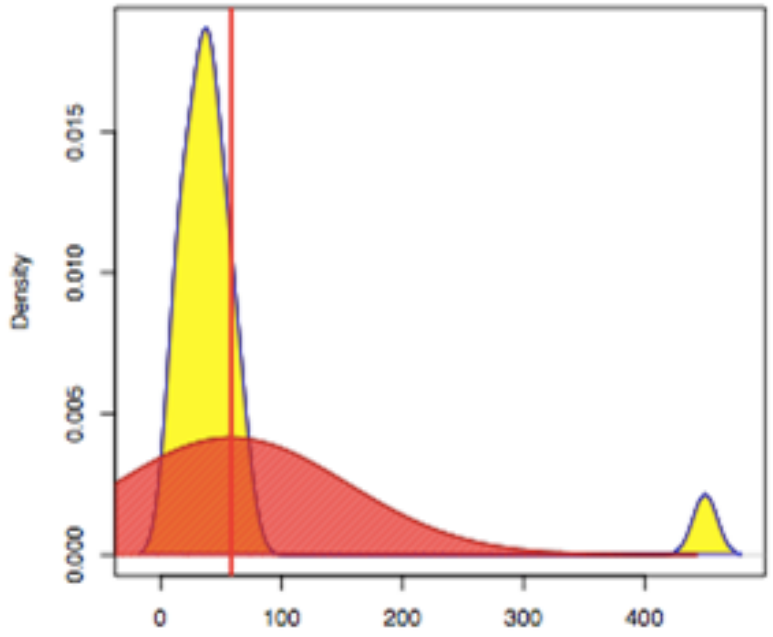
Big Picture: Where can Dirty Data Arise?



Numeric Outliers

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	68	450

ages of employees (US)



- median 37
- mean 58.52632
- variance 9252.041

Adapted from Joe Hellerstein's 2012 CS 194 Guest Lecture

Dirty Data Problems

1) Parsing text into fields (separator issues)
2) Naming conventions: NYC vs New York
3) Missing required fields (e.g. key field)
4) Different representations (2 vs Two)
5) Fields too long (get truncated)
6) Primary key violation (from un- to structured or during integration)
7) Redundant Records (exact match or other)
8) Formatting issues – especially dates
9) Licensing issues/Privacy/ keep you from using the data as you would like?

Conventional Definition of Data Quality

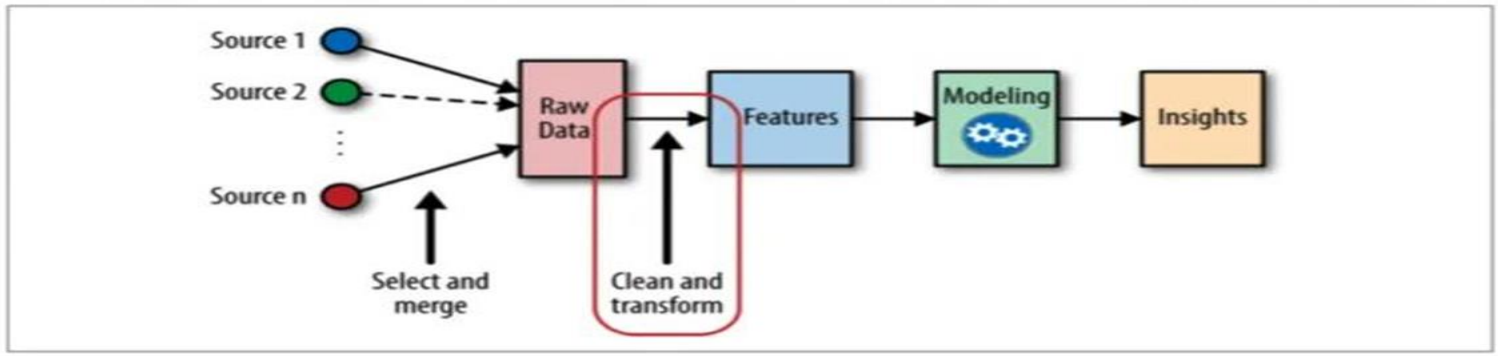
- Accuracy
  - The data was recorded correctly.
- Completeness
  - All relevant data was recorded.
- Uniqueness
  - Entities are recorded once.
- Timeliness
  - The data is kept up to date.
    - Special problems in federated data (multiple databases to function as one): time consistency.
- Consistency
  - The data agrees with itself.

Data Preparation

Data Preparation is an important part of Data Science. It includes two concepts such as **Data Cleaning** and **Feature Engineering**. These two are compulsory for achieving better accuracy and performance in the Machine Learning and Deep Learning projects.

Data Preprocessing is a part of data preparation.

Feature Engineering and Modeling



Data Preprocessing

Data Preprocessing is a technique that is used to convert the raw data into a clean data set.

Data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

The set of steps is known as Data Preprocessing. It includes –

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

Tasks of Data Preparation (including Data Preprocessing)

- **Data Cleaning :**This is the first step which is implemented in Data Preprocessing. In this step, the primary focus is on handling missing data, noisy data, detection, and removal of outliers, minimizing duplication and computed biases within the data.
- **Data Integration :**This process is used when data is gathered from various data sources and data are combined to form consistent data. This consistent data after performing data cleaning is used for analysis.
- **Data Transformation :**This step is used to convert the raw data into a specified format according to the need of the model. The options used for transformation of data are given below –
- **Normalization –** In this method, numerical data is converted into the specified range, i.e., between 0 and 1 so that scaling of data can be performed.
- **Aggregation – T**his method is used to combine the features into one. For example, combining two categories can be used to form a new group.
- **Generalization –** In this case, lower level attributes are converted to a higher standard (e.g age 20, 40 – may be taken as Young, Old, etc)
- **Data Reduction:** After the transformation and scaling of data duplication, i.e., redundancy within the data is removed and efficiently organize the data.

Missing, Noisy and inconsistent Data

How we can deal with the missing data

How missing data can be handled. Three different steps can be executed which are given below –

- **Ignoring the missing record –** It is the simplest and efficient method for handling the missing data. But, this method should not be performed at the time when the number of missing values are immense or when the pattern of data is related to the unrecognized primary root of the cause of statement problem.
- **Filling the missing values manually –** This is one of the best-chosen methods. But there is one limitation that when there are large data set, and missing values are significant then, this approach is not efficient as it becomes a time-consuming task.
- **Filling using computed values –** The missing values can also be occupied by computing **mean, mode or median** of the observed given values. Another method could be the predictive values that are computed by using any Machine Learning or Deep Learning algorithm.

How we can deal with the noisy data

- **Data Binning :** In this approach sorting of data is performed concerning the values of the neighborhood. This method is also known as local smoothing.
- **Clustering :** In the approach, the outliers may be detected by grouping the similar data in the same group, i.e., in the same cluster.
- **Machine Learning :** A Machine Learning algorithm can be executed for smoothing of data. For example, **Regression Algorithm** can be used for smoothing of data using a specified linear function.
- **Removing manually:** The noisy data can be deleted manually by the human being, but it is a time-consuming process, so mostly this method is not given priority.

What Is Data Wrangling?

- Data preprocessing before you build an analytic model

- Data wrangling is used in step during EDA and modeling to adjust data sets interactively while analyzing data and building a model.
  - Steps: Process of removing errors and combining complex data sets to make them more accessible and easier to analyze.
  - It is used to convert the raw data into the format that is convenient for the consumption of data
  - **It** executed at the time of making an interactive model.

- Data wrangling
  - extract the data from different data sources
  - sort of data using certain algorithm is performed
  - decompose the data into a different structured format
  - finally store the data into another database.

  Data is converted to the proper feasible format before applying any model to it.

  By performing filtering, grouping and selecting appropriate data accuracy and performance of the model could be increased.

Why is Data Wrangling Important?

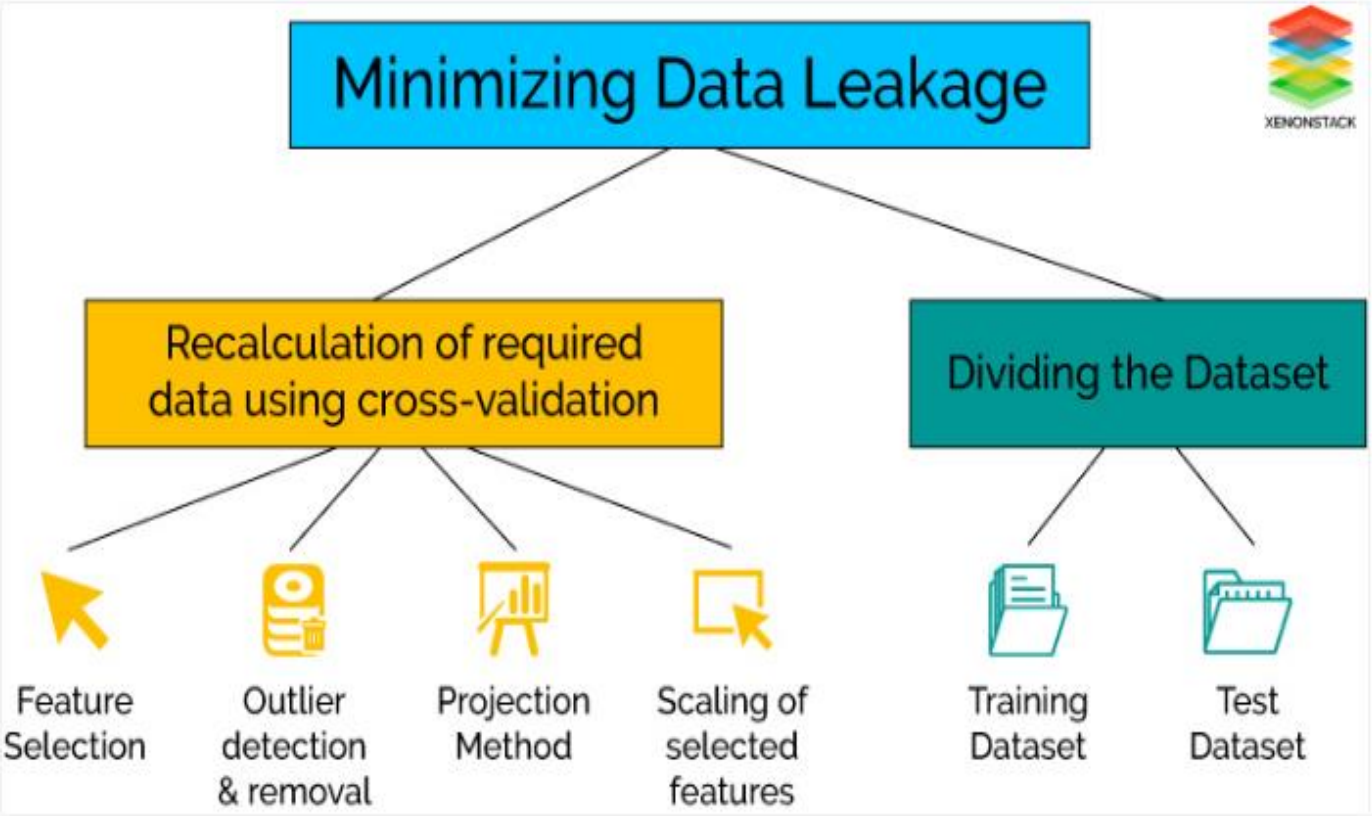
- Data Wrangling is used to handle the issue of **Data Leakage** while implementing Machine Learning and Deep Learning.
- Data Leakage is responsible for the cause of invalid Machine Learning/Deep Learning model due to the over optimization of the applied model.

Data Leakage

Data Leakage can be demonstrated in many ways that are given below –

- The Leakage of data from test dataset to training data set.
- Leakage of computed correct prediction to the training dataset.
- Leakage of future data into the past data.
- Usage of data outside the scope of the applied algorithm
- In general, the leakage of data is observed from two primary sources of Machine Learning/Deep Learning algorithms such as feature attributes (variables) and training data set.
- **Checking the presence of Data Leakage within the applied model**

Minimizing Data Leakage



How is Data Wrangling performed?

- If one considers the complete data set for normalization and standardization, then the cross-validation is performed for the estimation of the performance of the model leads to the beginning of data leakage.
- The effect of Data Leakage could be minimized by recalculating for the required Data Preparation during the cross-validation process that includes feature selection, outliers detection, and removal, projection methods, scaling of selected features and much more.
- Another solution is that dividing the complete dataset into training data set that is used to train the model and validation dataset which is used to evaluate the performance and accuracy of the applied model.

Tasks of Data Wrangling

- **Discovering:** Firstly, data should be understood thoroughly and examine which approach will best suit. For example: if have a weather data when we analyze the data it is observed that data is from one area and so primary focus is on determining patterns.
- **Structuring :**As the data is gathered from different sources, the data will be present in various shapes and sizes. Therefore, there is a need for structuring the data in proper format.
- **Cleaning :**Cleaning or removing of data should be performed that can degrade the performance of analysis.
- **Enrichment :**Extract new features or data from the given data set to optimize the performance of the applied model.
- **Validating:** This approach is used for improving the quality of data and consistency rules so that transformations that are applied to the data could be verified.
- **Publishing :**After completing the steps of Data Wrangling, the steps can be documented so that similar steps can be performed for the same kind of data to save time.