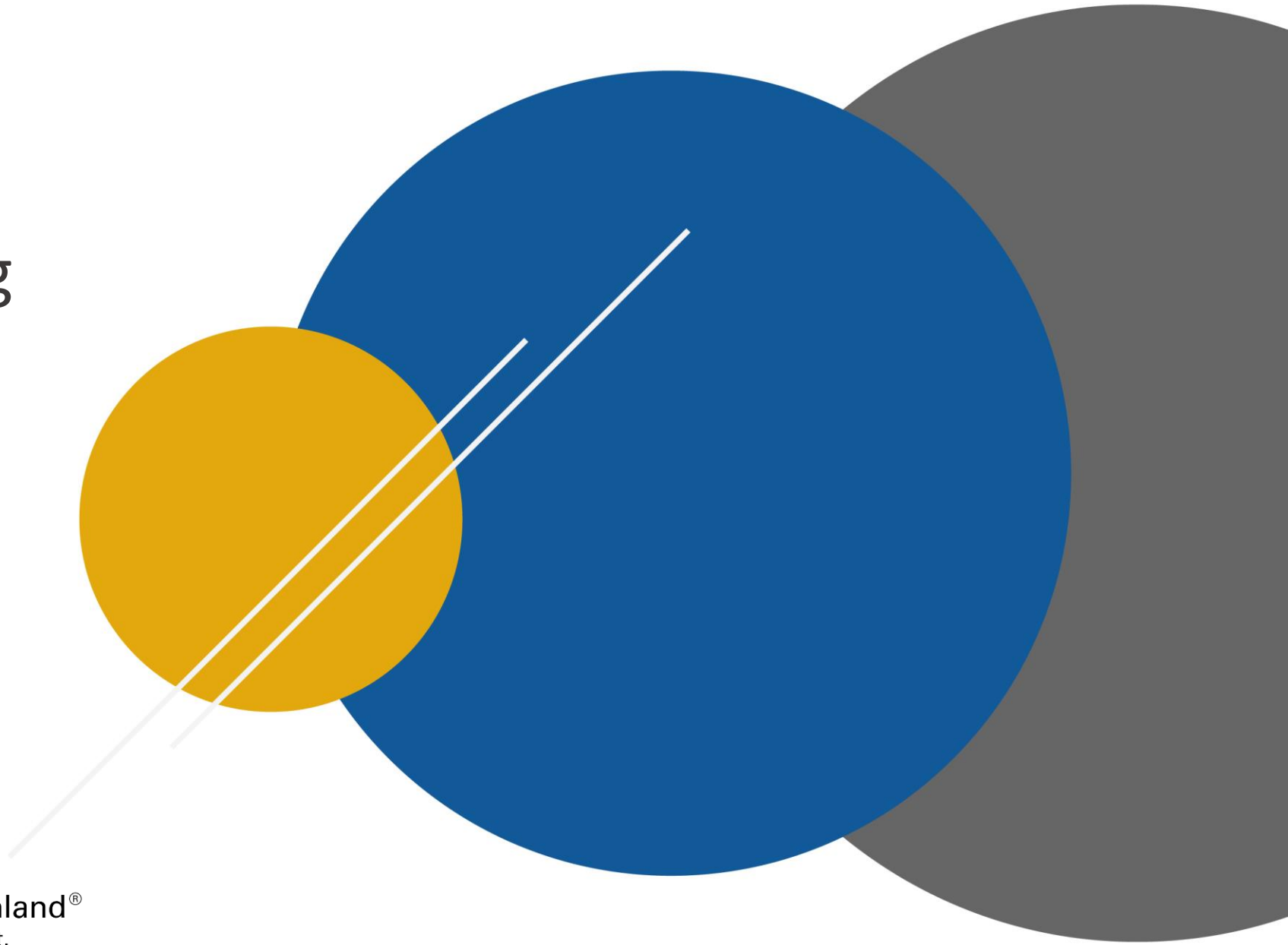


# Introduction to Big Data Area



# Agenda

---

- **Background & Motivation**
- Big Data Definition
- Computation Complexity
- Big Data, Data Science, and Data Analytics
- Big Data Taxonomy
- Big Data related Technology and Methodology
- Opportunity
- Challenge
- Structured vs Unstructured Data
- Artificial Intelligence & Machine Learning



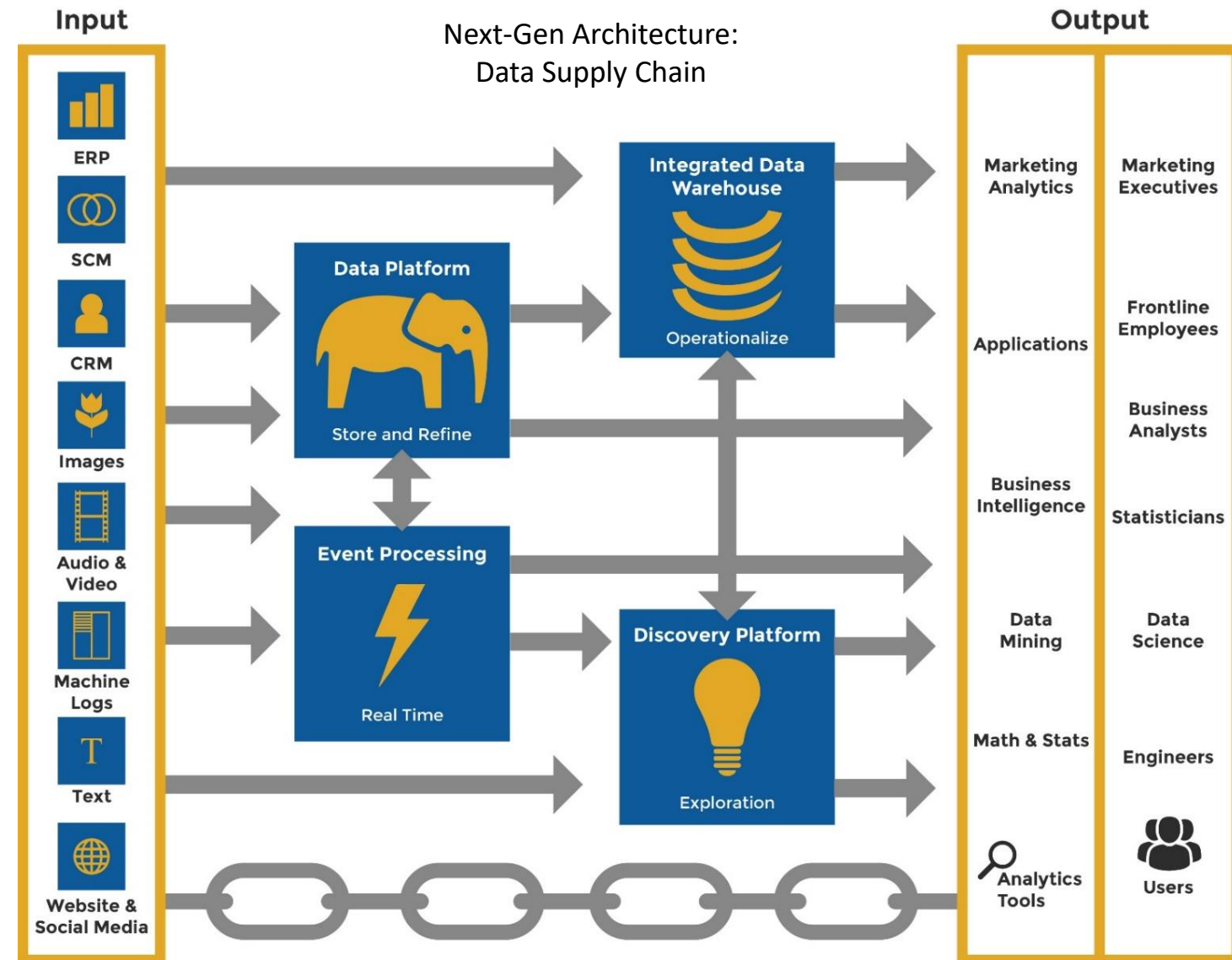
# Background

---

- Media social and internet technology factor
- Human data production (user generated content)
- Machine data production (internet of things)
- Cheap computing powers and Advance methodology



# Background



# Agenda

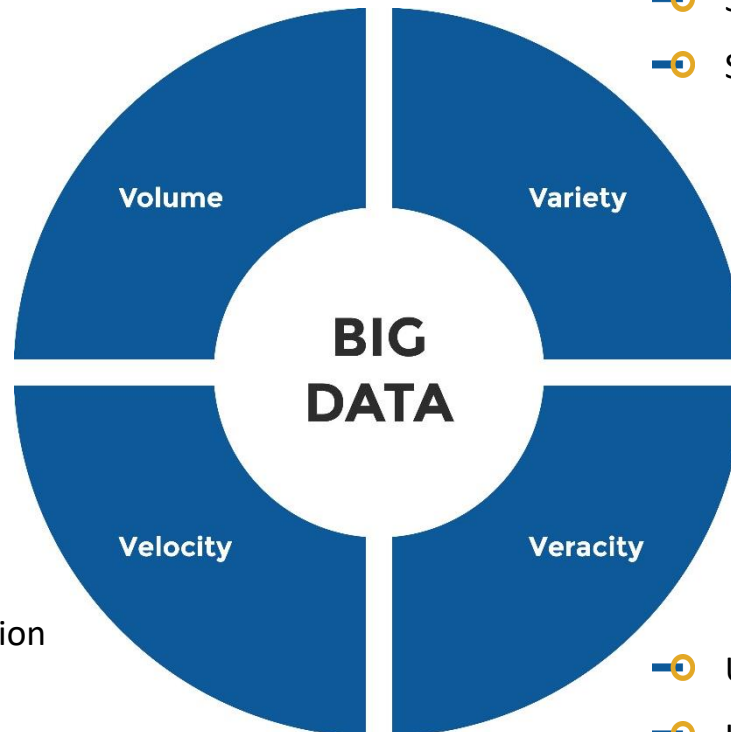
---

- Background & Motivation
- **Big Data Definition**
- Computation Complexity
- Big Data, Data Science, and Data Analytics
- Big Data Taxonomy
- Big Data related Technology and Methodology
- Opportunity
- Challenge
- Structured vs Unstructured Data
- Artificial Intelligence & Machine Learning



# Big Data Definition

- Click Stream
- Active/Passive Sensor
- Log
- Event
- Printed *Corpus*
- Speech
- Social Media
- Traditional



- Unstructured
- Semi-structured
- Structured

- Speed of Generation
- Rate of Analysis

- Untrusted
- Uncleansed

## The “V” Characteristics

### Volume

- The volume of persistent usable data in analytics system at any point in time.

### Variety

- The form and content of data structured (RDBMS), semi structured (Social Media) or unstructured (text/documents).

### Velocity

- How quickly the analytics system process the data to create insights.

### Veracity

- The degree to which data is accurate, precise and trusted.



# Big Data Definition



- A term → describe extremely large amounts of structured and unstructured data
- The activity → capture/storage/processing/sharing/reporting of data → beyond ability of legacy software tools and hardware infrastructure
- Related to many “science” branch → data analytics, data science, machine learning, artificial intelligence, IoT, and many more
- The application → on many field → efficient, cost effective, faster & accurate decision making.

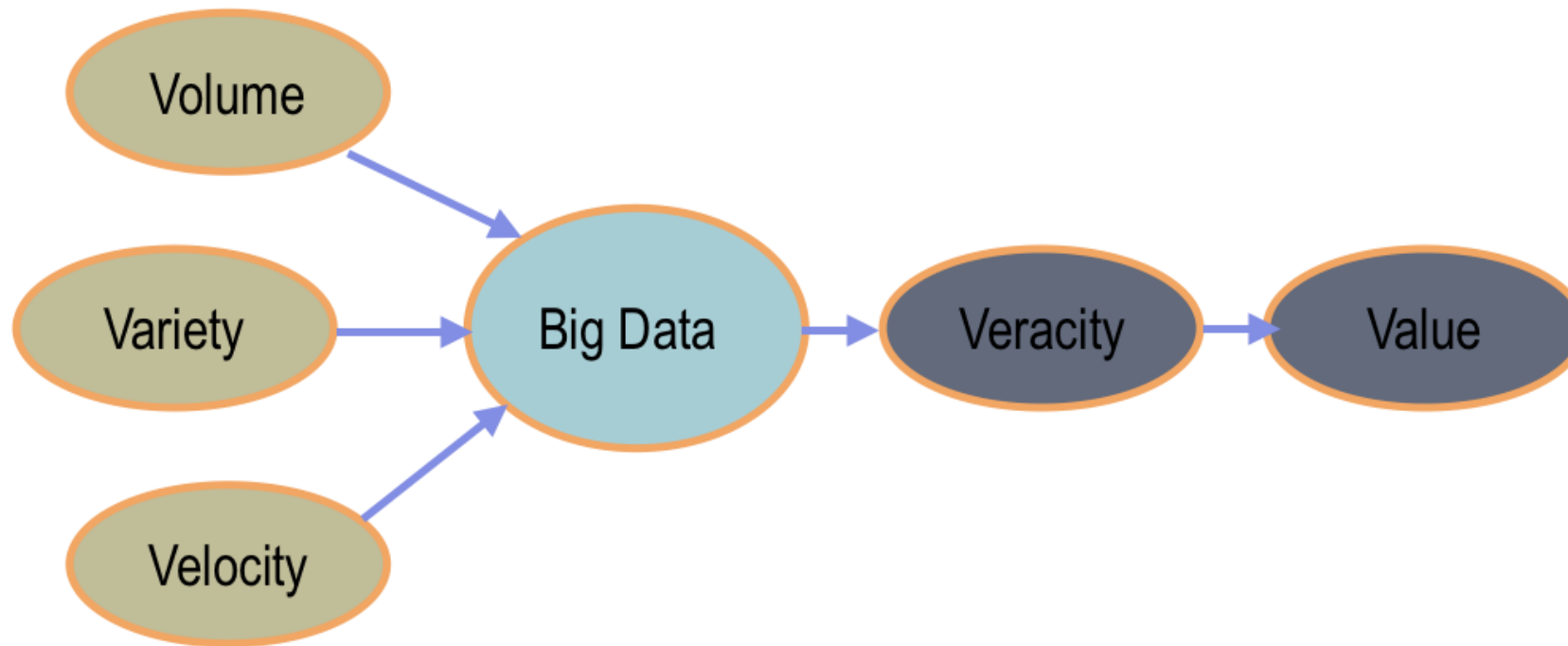
<b>Gigabyte</b>	$10^9 = 1.000.000.000$
<b>Terabyte</b>	$10^{12} = 1.000.000.000.000$
<b>Petabyte</b>	$10^{15} = 1.000.000.000.000.000$
<b>Exabyte</b>	$10^{18} = 1.000.000.000.000.000.000$
<b>Zetabyte</b>	$10^{21} = 1.000.000.000.000.000.000.000$

## Big Data Journey

1990	2010	Hadoop
Store 1400 MB	Store 1 TB	100 drives working at the same time can read 1 TB data in 2 minutes
Transfer Speed 4.5 MB/s	Transfer Speed 100 MB/s	
Read Drive ~ 5 Minutes	Read Drive ~ 3 Hours	



# The 5V's



- *Volume, Variety, and Velocity* are the "essential" characteristics of Big Data
- *Veracity, and Value* are the "quality" of Big Data





# Agenda

---

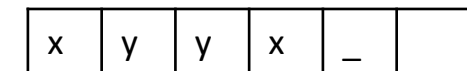
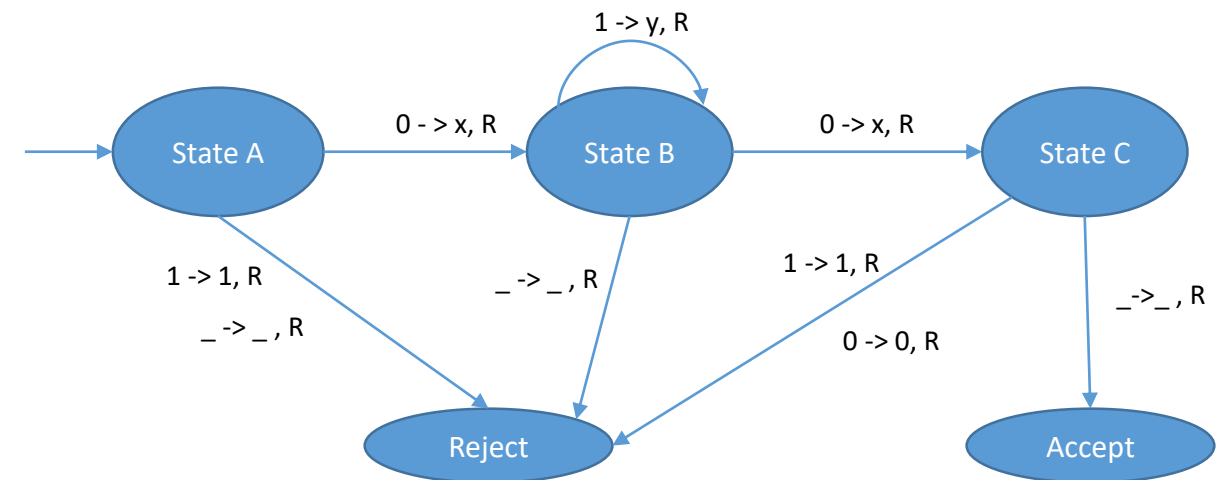
- Background & Motivation
- Big Data Definition
- **Computation Complexity**
- Big Data, Data Science, and Data Analytics
- Big Data Taxonomy
- Big Data related Technology and Methodology
- Opportunity
- Challenge
- Structured vs Unstructured Data
- Artificial Intelligence & Machine Learning



# Turing Machine

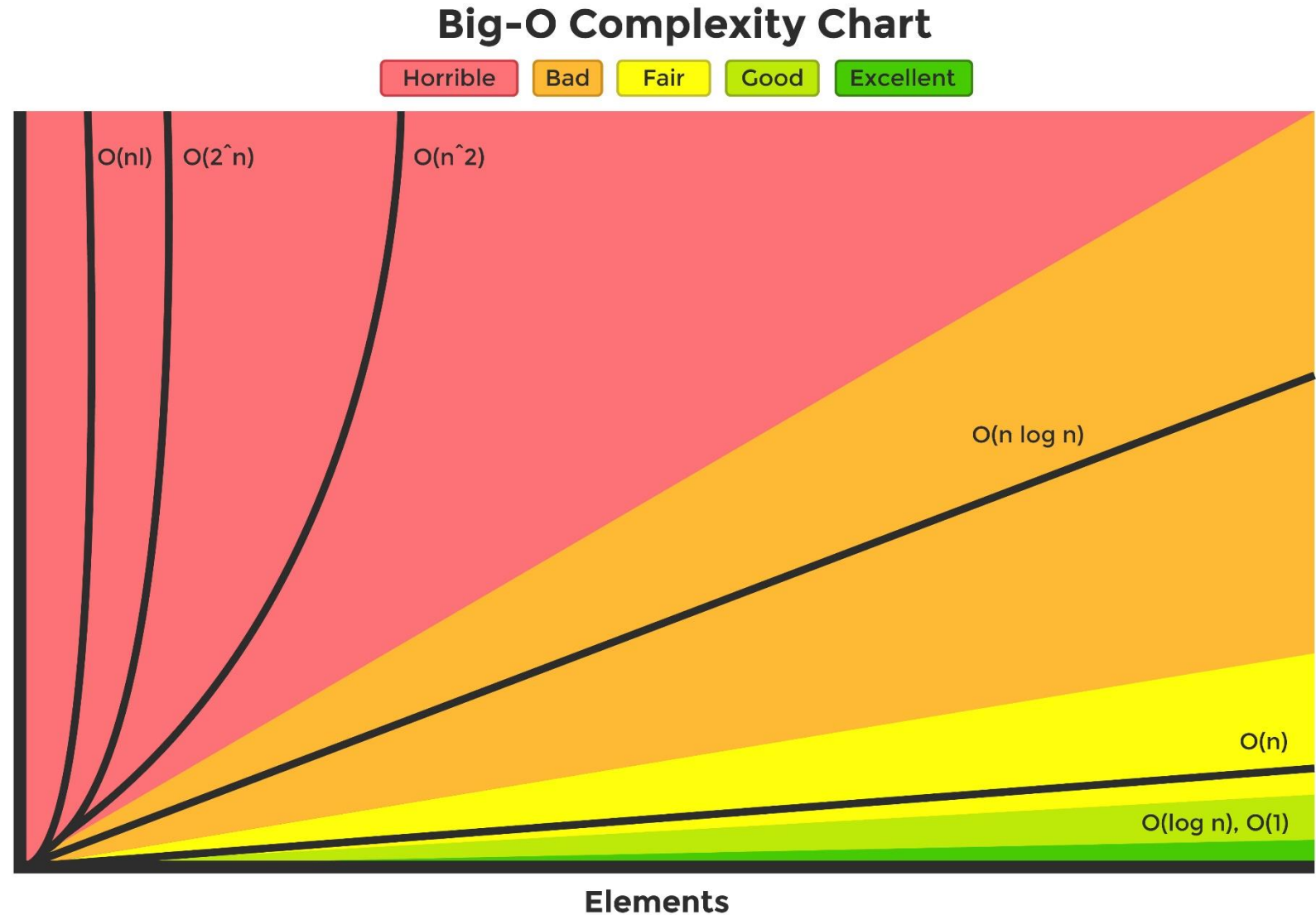
- Introduced by Alan Turing in 1936. A simple mathematical model of a computer. To solve computability questions.
- Turing Machine** provide a powerful computational model for solving problems in computer science and testing the limits of computation. It is an abstract computational model that performs computations by reading and writing to an infinite tape.
- A Turing machine consists of an infinite **tape** (as the memory), a **tape head** (a pointer to the currently inspected cell of memory), and a state transition table (to govern the behaviour of the machine). Each cell of the tape can have one of a predetermined finite set of symbols, one of which is the blank symbol.
- Example : Design a Turing Machine which recognize the language

$L = 01^*0$



write to tape

# Computation Cc



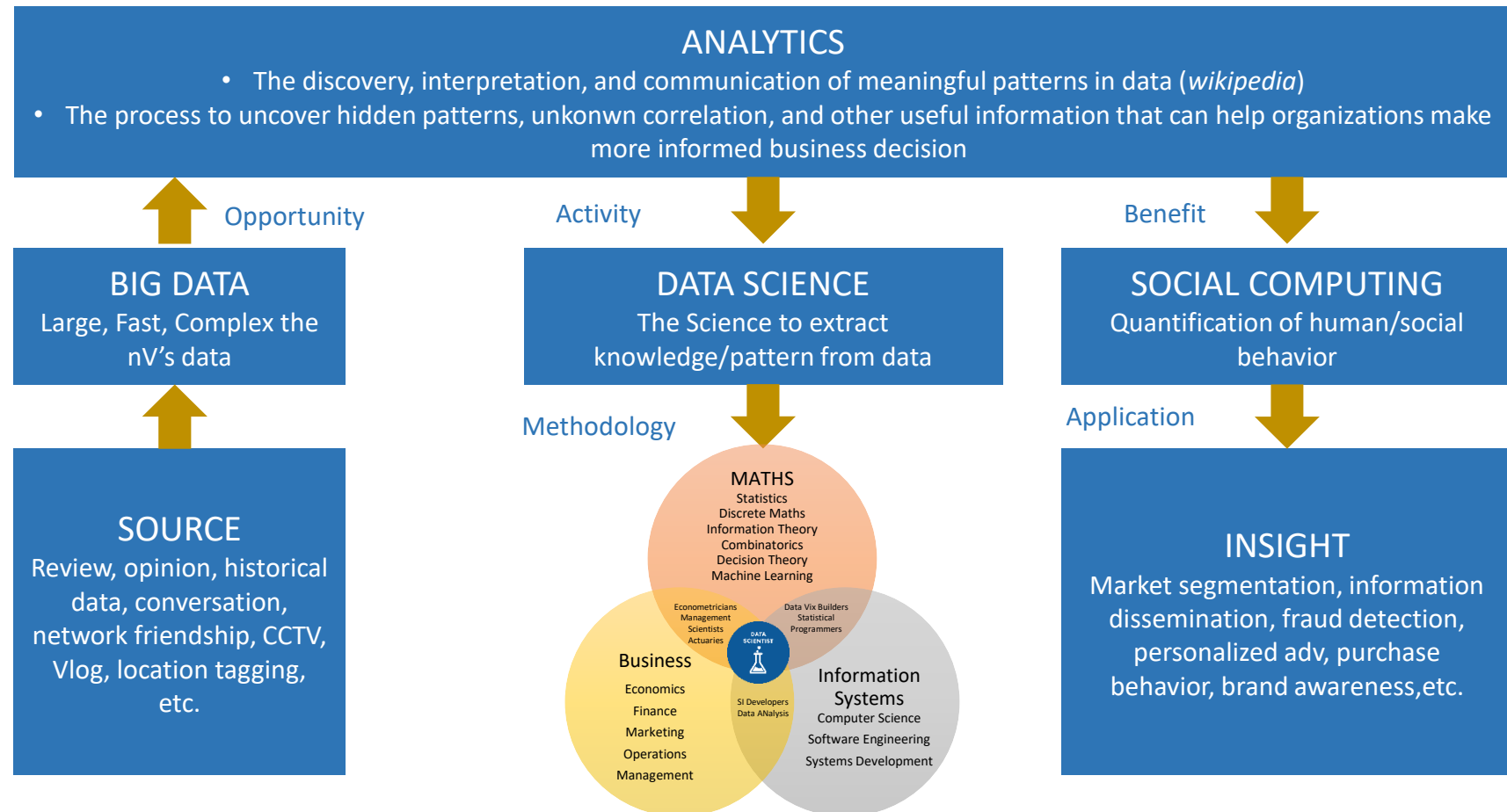
# Agenda

---

- Background & Motivation
- Big Data Definition
- Computation Complexity
- **Big Data, Data Science, and Data Analytics**
- Big Data Taxonomy
- Big Data related Technology and Methodology
- Opportunity
- Challenge
- Structured vs Unstructured Data
- Artificial Intelligence & Machine Learning



# Big Data, Data Science, & Data Analytics



# Agenda

---

- Background & Motivation
- Big Data Definition
- Computation Complexity
- Big Data, Data Science, and Data Analytics
- **Big Data Taxonomy**
- Big Data related Technology and Methodology
- Opportunity
- Challenge
- Structured vs Unstructured Data
- Artificial Intelligence & Machine Learning



# Big Data Taxonomy :

## Data Science Body of Knowledge

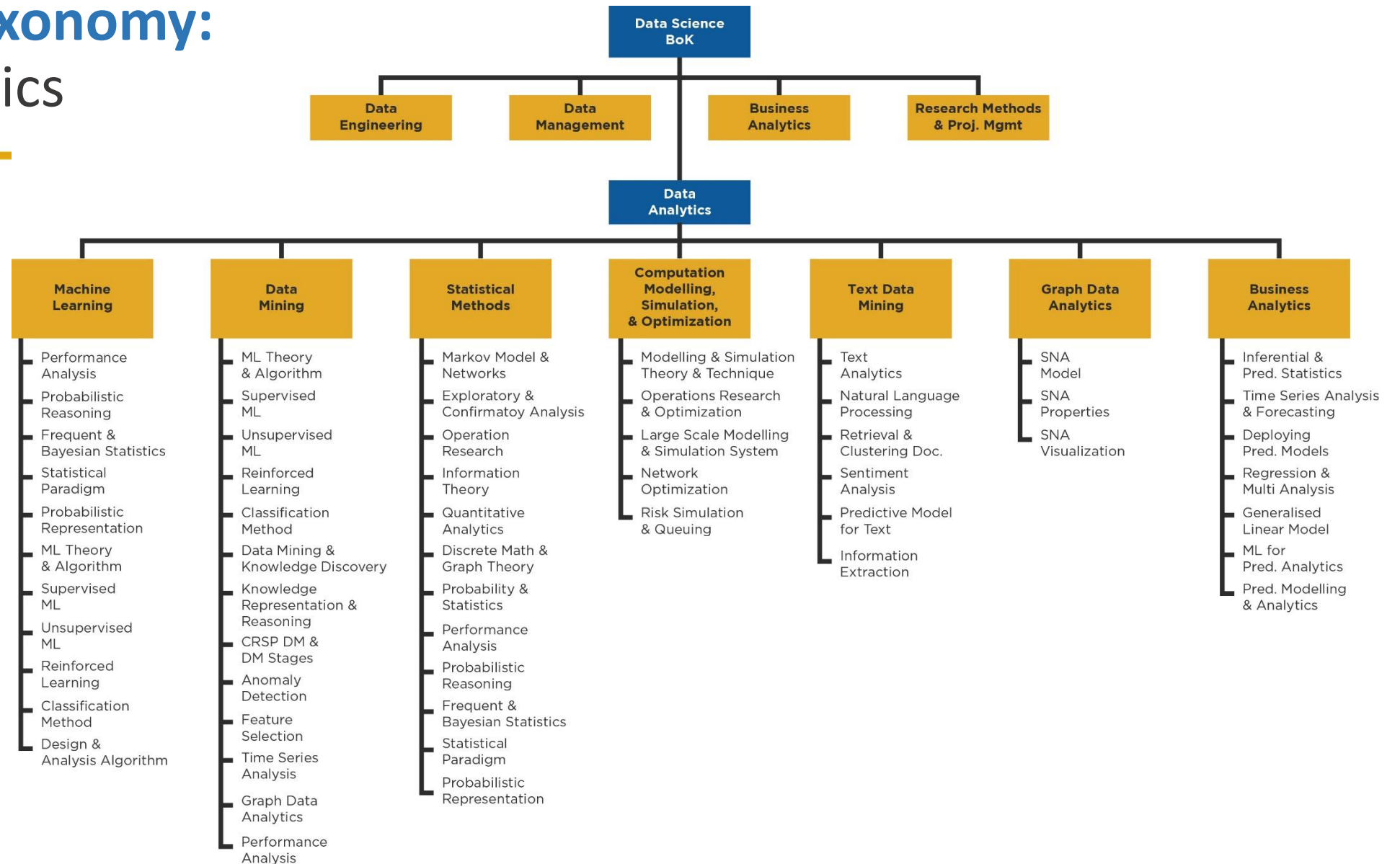
No	Name	Knowledge Area	Scientific Subject
1	Data Analytics	Statistical Methods, Machine Learning, Data Mining, Predictive Analytics, Computational Modeling/ Simulation/Optimization	Computing Methodologies, Mathematics of Computing
2	Data Engineering	Big Data Infrastructure & Technologies, Infrastructure & Platform for DS Apps, Cloud Computing Tech, Data & Apps Security, Big Data System Organization & Engineering, DS / Big Data Apps Design, IS to Support DSS	Algorithm & Complexity, Architecture & Organization, Computational Science, Graphic & Visualization, Information Management, Platform Based Dev, Software Engineering
3	Data Management	General Principle & Concepts in Data Management and Organization, Data Management Systems, Data Enterprise Infrastructure, Data Governance, Big Data Storage, Digital Library & Archives, Data Curation, Data Preservation	Data (Governance, Architecture, Model & Design, Storage & Operations, Security, Integration & Interoperability, Warehousing & BI, Quality), Metadata, Reference & Master Data
4	Research Methods and Project Management	Research Methods, Project Management	Project (Integration Management, Scope Management, Quality, Risk Management)
5	Business Analytics	Business Analytics Foundation, Business Analytics Organization and Enterprise Management	Business Analysis Planning & Monitoring, Requirement Analysis & Design Definition, Requirement Life Cycle Management, Solution Evaluation & Improvements Recommendation
6	Domain Knowledge		





# Big Data Taxonomy:

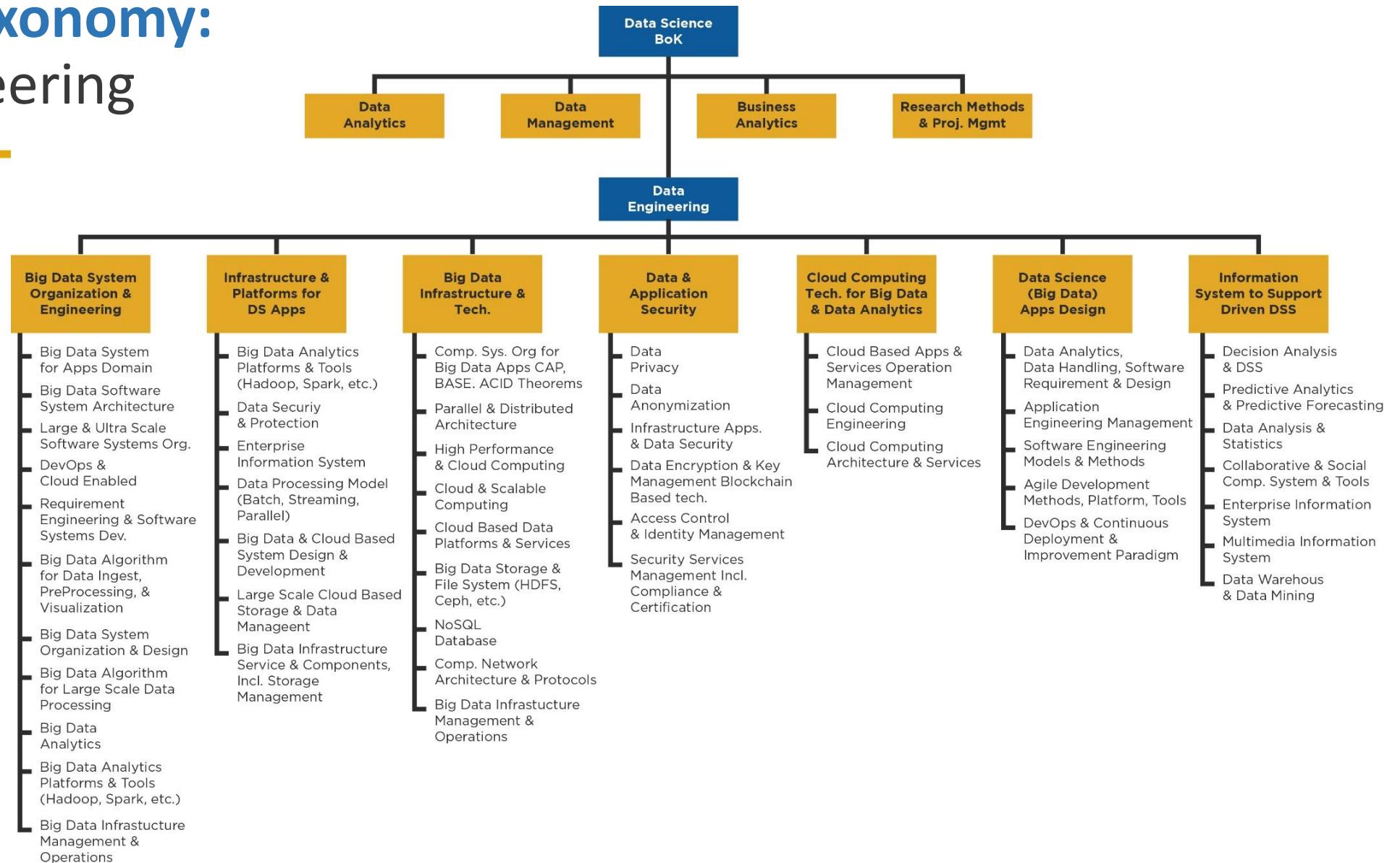
## Data Analytics





# Big Data Taxonomy:

## Data Engineering



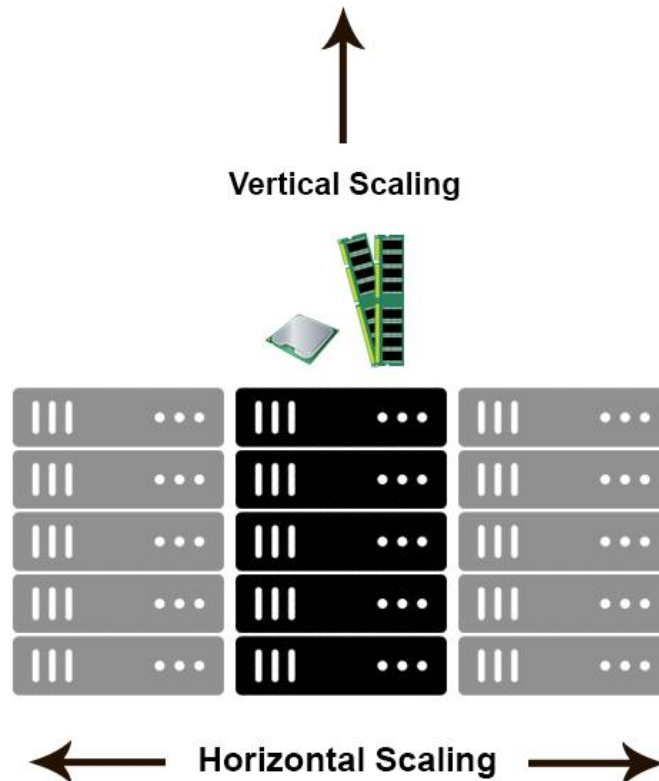
# Agenda

---

- Background & Motivation
- Big Data Definition
- Computation Complexity
- Big Data, Data Science, and Data Analytics
- Big Data Taxonomy
- **Big Data related Technology and Methodology**
- Opportunity
- Challenge
- Structured vs Unstructured Data
- Artificial Intelligence & Machine Learning



# Horizontal vs Vertical Scaling



- Horizontal Scaling (scale-out) means scaling by adding more machines into an existing pool of resources.
- Vertical Scaling (scale-up) means scaling by adding more power (CPU, RAM, and others) to a single machine.







*Load-Balancer* is responsible to distribute user requests (load) among the various back-end systems/machines/nodes in the cluster (scale-out system).

*Load Balancer* distributes load by maintaining state of each machine, how many requests are being served by each machine, which machine is idle, which machine is over-loaded with queued requests etc.

Load balancing algorithm considers such things before redirecting the request to an appropriate server machine. It also takes into account the network overhead and might choose the server in the nearest data center provided it is available to service the requests



# Distributed Computing

- 
**A computer cluster** is a set of loosely / tightly connected computers that work together, in many respects, they can be viewed as a single system. Each computer/node set to perform the same task, controlled and scheduled by software.
- 
 Computer clusters emerged as a result of convergence of a number of computing trends including the availability of low-cost microprocessors, high-speed networks, and software for high performance distributed computing.
- 
**Distributed computing** is a computer science field that studies distributed computer systems, where computers are located on different network, in which each communicate and coordinate their actions by passing messages to one another.
  - Cluster/Distributed Objectives : **Scalability** and **High Availability**
  - Cluster/Distributed Abilities : **Load Balancing** and **Fault Tolerance**
- 
 High Availability : If a node in a cluster fails, the services running on this node can be taken over by other service nodes
- 
 Load Balancing : distribute tasks to computing and network resources in a clustered environment.
- 
 Fault Tolerance : provide mechanism to recover the system, when failure occurs



# Big Data Related Technology / Methodology

---

Data Science

Artificial Intelligence

Open Data

Internet of Things

Data Analytics

Machine Learning

Social Media Analytics

Prescriptive / Predictive Analytics

Stream Processing

Parallel Processing / Hadoop

Data Mining

Text / Video Analytics

Graph Analytics

NoSQL



# Agenda

---

- Background & Motivation
- Big Data Definition
- Computation Complexity
- Big Data, Data Science, and Data Analytics
- Big Data Taxonomy
- Big Data related Technology and Methodology
- **Opportunity**
- Challenge
- Structured vs Unstructured Data
- Artificial Intelligence & Machine Learning



# Big Data

## Analytics Role

---

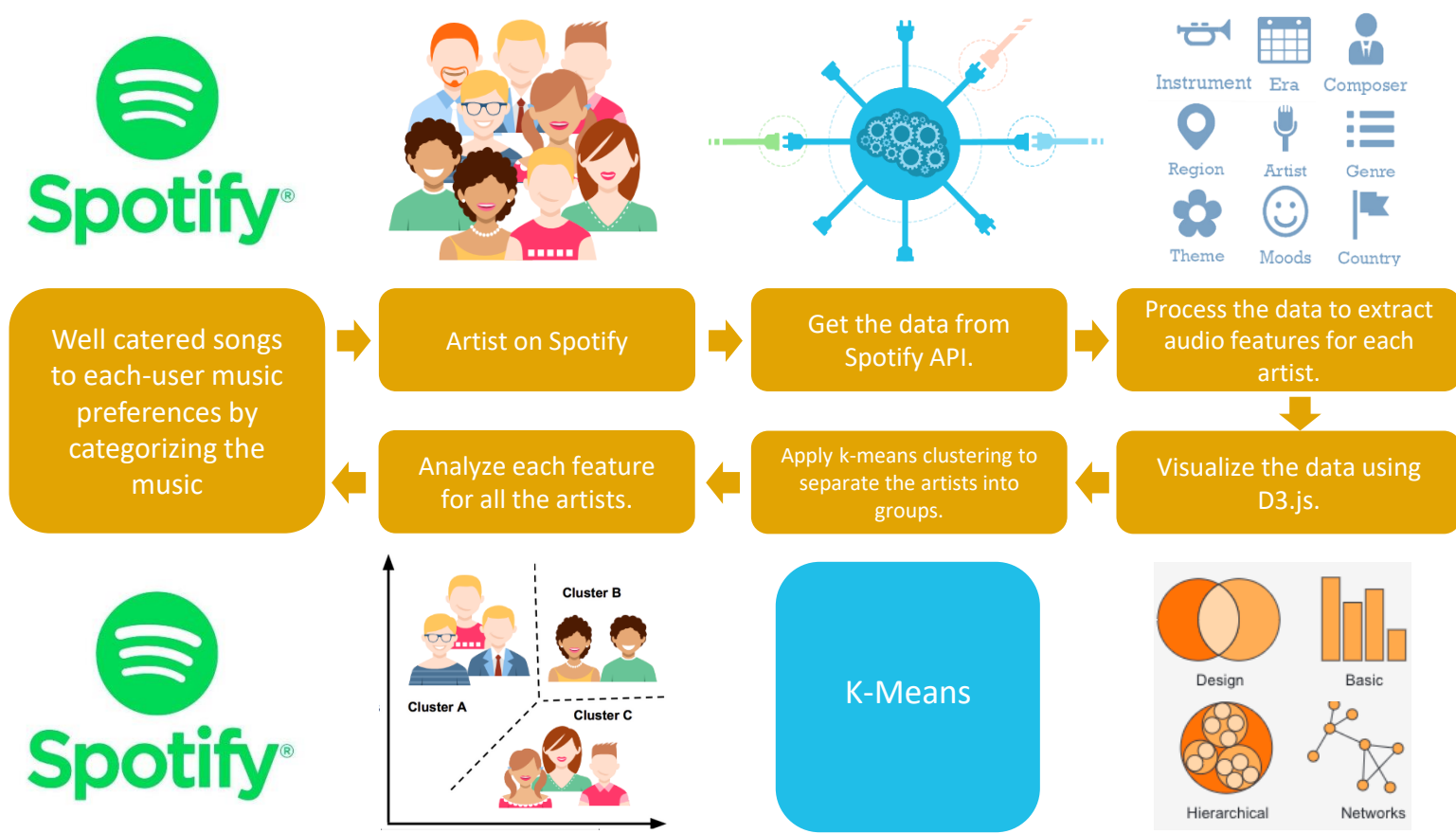
- By **describing** the phenomenon, by **predicting** the value, by **estimating** the future outcome, by **optimizing** the resources and the decision, by **simulating** all possible scenarios





# Practical Example

## Data Analytics Challenges





# Opportunity

---

- Finding new (hidden) pattern from large scale data
- Provide granular / micro (or macro) data
- Provide relatively fast and cheap process
- Modelling behavior lead to prediction analytics
- Complementary to legacy methods based on sampled data
- Fulfill the need of real time analytics
- Event simulation (ANN)
- Problem Optimization



# Agenda

---

- Background & Motivation
- Big Data Definition
- Computation Complexity
- Big Data, Data Science, and Data Analytics
- Big Data Taxonomy
- Big Data related Technology and Methodology
- Opportunity
- **Challenge**
- Structured vs Unstructured Data
- Artificial Intelligence & Machine Learning



# Challenges

---

- Data privacy (ethical issues)
- Choosing the suitable algorithm based on data characteristics (type) => Machine Learning methods and evaluation maturity
- Finding the right balance between speed and accuracy of model formation
- Incorporate to legacy scientific questions and hypotheses
- Streaming data and online / real time analytics
- Data quality + costs + security < benefit
- Hard to find the right talent



# Agenda

---

- Background & Motivation
- Big Data Definition
- Computation Complexity
- Big Data, Data Science, and Data Analytics
- Big Data Taxonomy
- Big Data related Technology and Methodology
- Opportunity
- Challenge
- **Structured vs Unstructured Data**
- Artificial Intelligence & Machine Learning



# Structured vs Unstructured Data

## Structured Data

- High Degree of Organization, such as a relational database

Column	Value
Patient	John Brown
Date of Birth	12/07/1993
Date Admitted	02/03/2011

## Unstructured Data

- Information that is difficult to organize using traditional mechanisms

“The patient came in complaining of chest pain, shortness of breath, and lingering headaches.. Smokes 2 packs a day.. Family history of heart disease.. Has been experiencing similar symptoms for the past 12 hours...”



# Structured vs Unstructured Data (Characteristics)

---

## Structured Data

- Well defined content
- Easily Understood
- Stored in RDBMS
- Easy to enter, store and analyze
- Example: data in database table  
(customer data, sales data, sensor data)

## Unstructured Data

- Structure not obvious
- Process data to understand
- RDBMS not a good fit
- Difficult and costly to analyze
- Example: email, videos, audio, web pages,  
social media feeds, presentation



# Unstructured Data

## From all sorts of **player**

- Company/Brands
- Consumers
- Prospects
- Internet Users
- Employees

## From all sorts of **forms**

- Long and complex sentences in formal language
- Short sentences with spelling and/or grammar mistakes
- Non-coordinated concepts
- In any others languages

## From all sorts of **media**

- Web pages
- Blog posts
- Posts and answer on social networks
- E-Newsletters/Webzines
- Press Releases
- White papers/Product brochures/Online or offline scanned documents
- Survey verbatims
- Complaints collected by brand
- Answer to open-ended questions
- CRM text data
- Internal collaborative platform
- Online and offline content



# SQL vs NOSQL

## SQL

- Structured and organized data
- Structured Query Language (SQL)
- Data and its relationships stored in separated tables
- Data manipulation Language, Data Definition Language
- Tight Consistency
- BASE Transaction

## NoSQL

- No Declarative Query Language
- No predefined schema
- Key-Value pair storage, Column Store, Document store, Graph Databases
- Eventual consistency rather ACID property
- Unstructured and unpredictable data
- CAP Theorem
- Prioritize high performance, high availability and scalability



	NoSQL	SQL
<b>Model</b>	Non-Relational Stores data in JSON documents, key/value pairs, wide column stores, or graphs	Relational Stores data in table
<b>Data</b>	Offers flexibility as not every record needs to store the same properties	Great for solutions where every record has the same properties
	New properties can be added on the fly	Adding a new property may require altering schemas or backfiling data
	Relationships are often captured by denormalizing data and presenting it in a single record	Relationships are often captured in a using joins to resolve references across tables
	Good for semi-structured data	Good for structured data
<b>Schema</b>	Dynamic or flexible schemas Database in schema-agnostic and the schema is dictated by the application. This allow for agility and highly iterative development.	Strict schema Schema must be maintained and kept in sync between application and database
<b>Transactions</b>	ACID transaction support varies per solution	Support ACID transactions
<b>Consistency</b>	Consistency varies per solution, some solutions have tunable consistency	Strong consistency supported
<b>Scale</b>	Scales well horizontally	Scales well vertically



# Agenda

---

- Background & Motivation
- Big Data Definition
- Computation Complexity
- Big Data, Data Science, and Data Analytics
- Big Data Taxonomy
- Big Data related Technology and Methodology
- Opportunity
- Challenge
- Structured vs Unstructured Data
- **Artificial Intelligence & Machine Learning**



# Artificial Intelligence

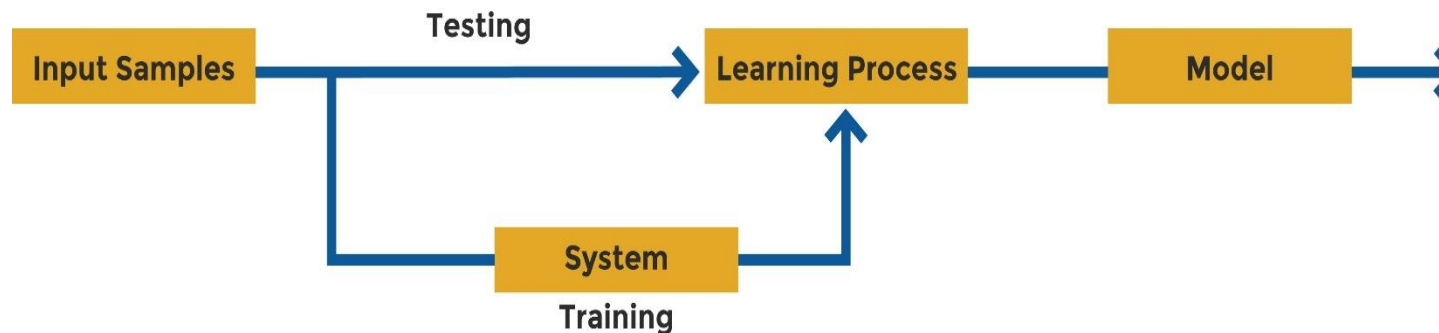
---

- It is intelligence exhibited by machines. In computer science, an ideal “Intelligent” machine is a flexible rational agent that perceives its environment and takes actions that maximize its chance of success at some goal.
- The science and engineering of making intelligent machines
- The study and design of intelligent agents
- Several domain of AI : Expert System, Natural Language Processing, Speech Recognition, Computer Vision, Robotics, Automatic Programming.



# Machine Learning

- A branch of artificial intelligence, concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data.
- As intelligence requires knowledge, it is necessary for the computers to acquire knowledge.
- Learning System Model



- Like human learning from past experiences.
- A computer does not have “experiences”. an application domain.
- A computer system learns from data, which represent some “past experiences” of an application domain.



# Learning Methodology

