Big Data Analytics

Dr. Sonali Agarwal
Associate Professor, Department of IT
Indian Institute of Information Technology Allahabad, India

Outline and Purpose of this course-I

To provide a simple introduction to:

- ✓ Introduction to Big Data,
- ✓ ML and Big Data,
- ✓ In memory and disk based computation of Machine Learning Algorithms
- ✓ Hands-on using Hadoop and Mahout, Flume, Kibana, Elastic Search

Outline and Purpose of this course-II

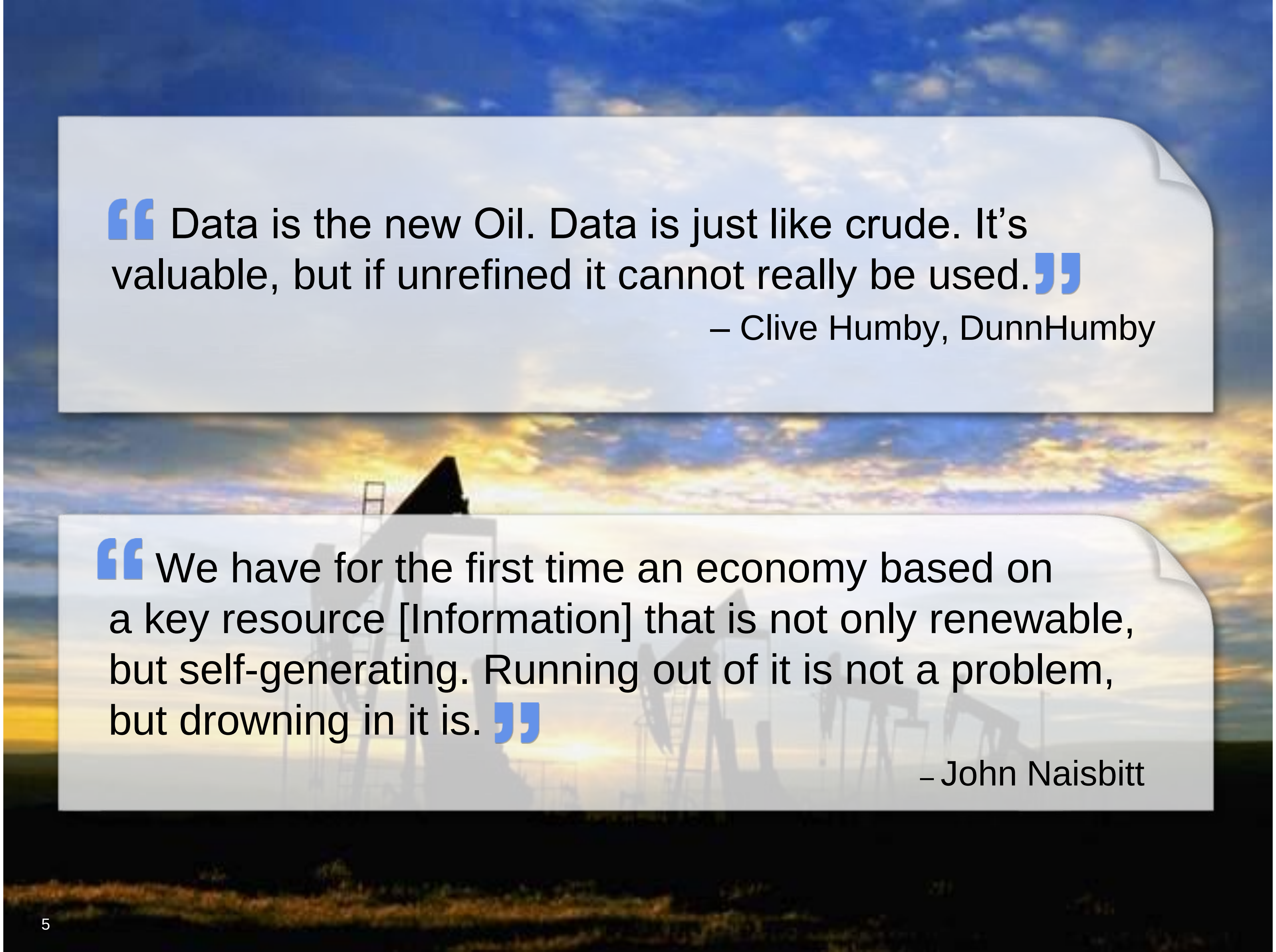
To provide a simple introduction to:

- ✓ Basic statistics, Data sources
- ✓ Pipelines, Extracting, transforming and selecting features,
- ✓ Classification and Regression
- ✓ Clustering, Collaborative filtering, Frequent Pattern Mining, Model selection and tuning, example and use cases
- ✓ Hands-on using Apache Spark

Course Project

Course Project will be of 3 types.

- **Dataset analysis:** select a dataset (for instance from your research) and apply at least two techniques seen in the course using Apache Spark, Dask or scikit-learn. You are not required to re-implement these techniques, but you need to discuss and interpret the results.
- **Technology evaluation:** perform a comparative study of at least two open-source technologies related to Big Data Analysis, for instance from the Hadoop project.
- **Algorithm implementation:** (Re-)implement at least two algorithms seen in the course or related to the themes seen in the course.



> “Data is the new Oil. Data is just like crude. It’s valuable, but if unrefined it cannot really be used.”

– Clive Humby, DunnHumby

> “We have for the first time an economy based on a key resource [Information] that is not only renewable, but self-generating. Running out of it is not a problem, but drowning in it is.”

– John Naisbitt

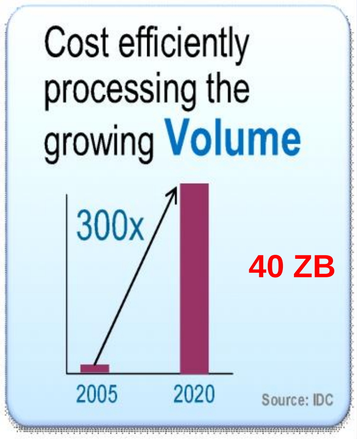
Big Data is the next Natural Resource

Big Data is the next Natural Resource

“We have for the first time an economy based on a key resource (Information) that is not only renewable, but self-generating.

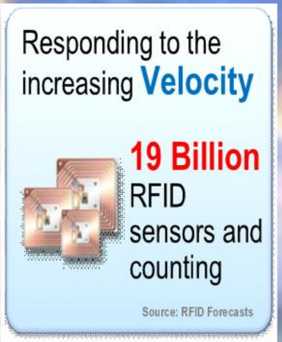
Running out of it is not a problem, but drowning in it is.”

— John Naisbitt



Cost efficiently processing the growing Volume

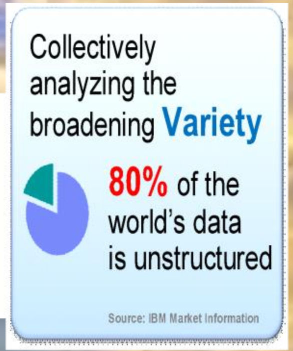
Source: IDC



Responding to the increasing Velocity

19 Billion RFID sensors and counting

Source: RFID Forecasts



Collectively analyzing the broadening Variety

80% of the world’s data is unstructured

Source: IBM Market Information

Establishing the Veracity of big data sources

1 in 3 business leaders don’t trust the information they use to make decisions

Source: IBM. BAO for the Intelligent Enterprise

Harvesting any resource requires Mining, Refining and Delivering

Big Data Vs Small Data

	Aspect	Big Data	Small Data
1	Size	Big volumes, often terabytes to petabytes	Relatively small and manageable
2	Focus	Broad, covering diverse topics and sources	Specific and targeted, focusing on relevant subsets
3	Context	Often lacks context, dealing with diverse sources	Contextually relevant, tied to specific domains or scenarios
4	Structure	It can be structured, semi-structured, or unstructured	Typically structured and organized
5	Accessibility	Requires significant resources and infrastructure	More accessible and readily available
6	Precision	Emphasizes identifying patterns and trends	Aims for precision and accuracy in analysis
7	Human-scale Interactions	Analyzes large-scale interactions, behaviors, or trends	Analyzes individual or small-scale interactions, often human-centric
8	Examples	Social media data, sensor readings, weblogs	Customer preferences, sales data, survey responses

Types of Big Data

Structured Data

Examples Of Structured Data

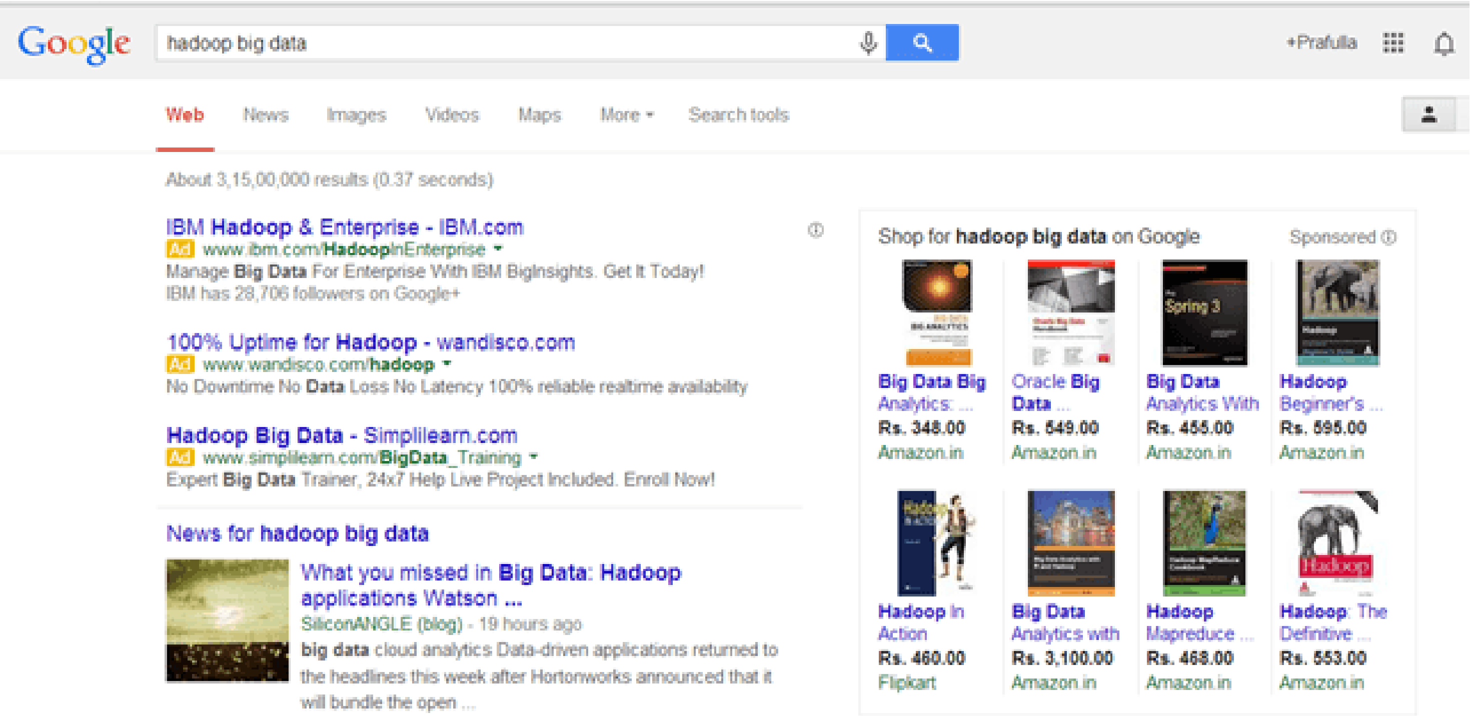
An 'Employee' table in a database is an example of Structured Data

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Types of Big Data

Un-structured Data

- The output returned by 'Google Search'



Google hadoop big data

+Prafulla

Web News Images Videos Maps More ▾ Search tools

About 3,15,00,000 results (0.37 seconds)

IBM Hadoop & Enterprise - IBM.com
Ad www.ibm.com/HadoopInEnterprise ▾
Manage Big Data For Enterprise With IBM BigInsights. Get It Today!
IBM has 28,706 followers on Google+

100% Uptime for Hadoop - wandisco.com
Ad www.wandisco.com/hadoop ▾
No Downtime No Data Loss No Latency 100% reliable realtime availability

Hadoop Big Data - Simplilearn.com
Ad www.simplilearn.com/BigData_Training ▾
Expert Big Data Trainer, 24x7 Help Live Project Included. Enroll Now!

News for hadoop big data

What you missed in Big Data: Hadoop applications Watson ...
SiliconANGLE (blog) - 19 hours ago
big data cloud analytics Data-driven applications returned to the headlines this week after Hortonworks announced that it will bundle the open ...

Shop for hadoop big data on Google Sponsored

Big Data Big Analytics: ...
Rs. 348.00
Amazon.in

Oracle Big Data ...
Rs. 549.00
Amazon.in

Big Data Analytics With
Rs. 455.00
Amazon.in

Hadoop Beginner's ...
Rs. 595.00
Amazon.in

Hadoop In Action
Rs. 460.00
Flipkart

Big Data Analytics with
Rs. 3,100.00
Amazon.in

Hadoop Mapreduce ...
Rs. 468.00
Amazon.in

Hadoop: The Definitive ...
Rs. 553.00
Amazon.in

Types of Big Data

Semi-structured Data

- Personal data stored in an XML file-

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

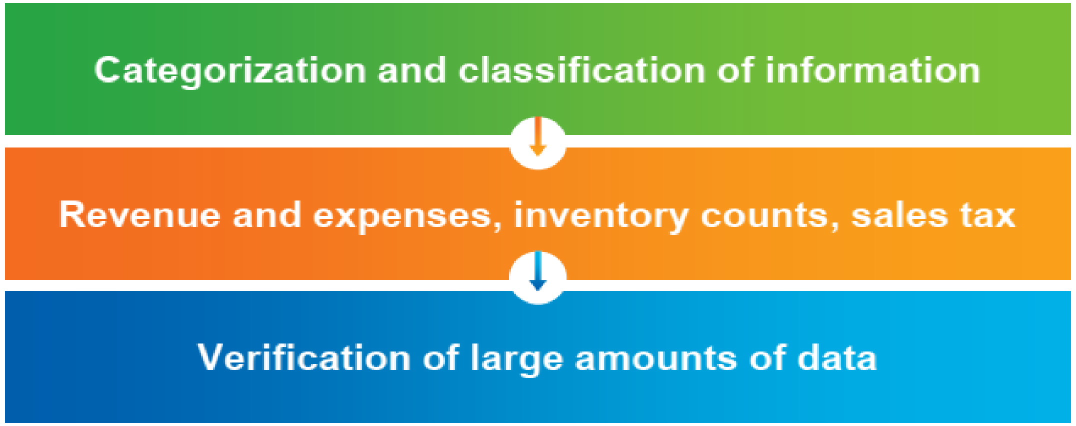
Types of Data Analytics

Introduction

Types of Data Analytics--------> Descriptive Analytics

- Descriptive analytics is the process of using current and historical data to identify trends and relationships.
- It is sometimes called the simplest form of data analysis because it describes trends and relationships but doesn’t dig deeper.
- Descriptive analytics helps business to understand, the number of time customers has visited the bank, types of transaction(s) carried out, how are they satisfied with the banks products and services.
- Tools: Microsoft Excel or data visualization tools

What is happening?



Categorization and classification of information

Revenue and expenses, inventory counts, sales tax

Verification of large amounts of data

[2] WOLNIAK, Radosław. "THE CONCEPT OF DESCRIPTIVE ANALYTICS." Scientific Papers of Silesian University of Technology. Organization & Management/Zeszyty Naukowe Politechniki Slaskiej. Seria Organizacji i Zarzadzanie 172 (2023).

Introduction

Descriptive Analytics

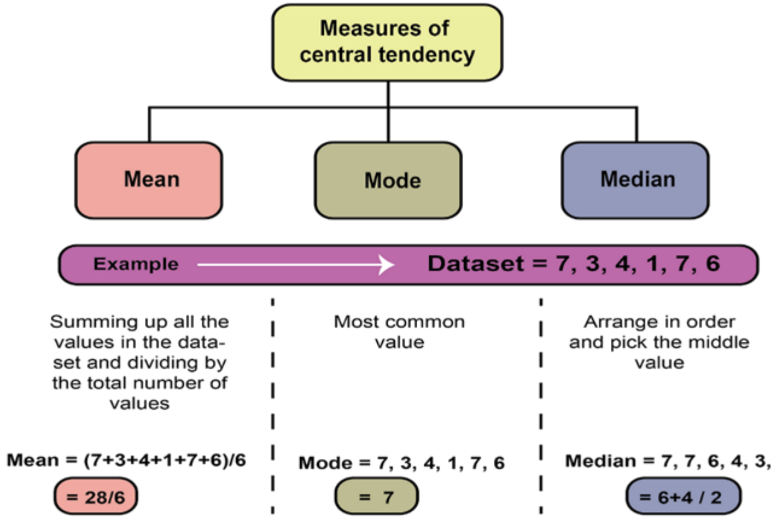
(1) Measure of central tendency

(2) Interquartile range

(3) Skewness

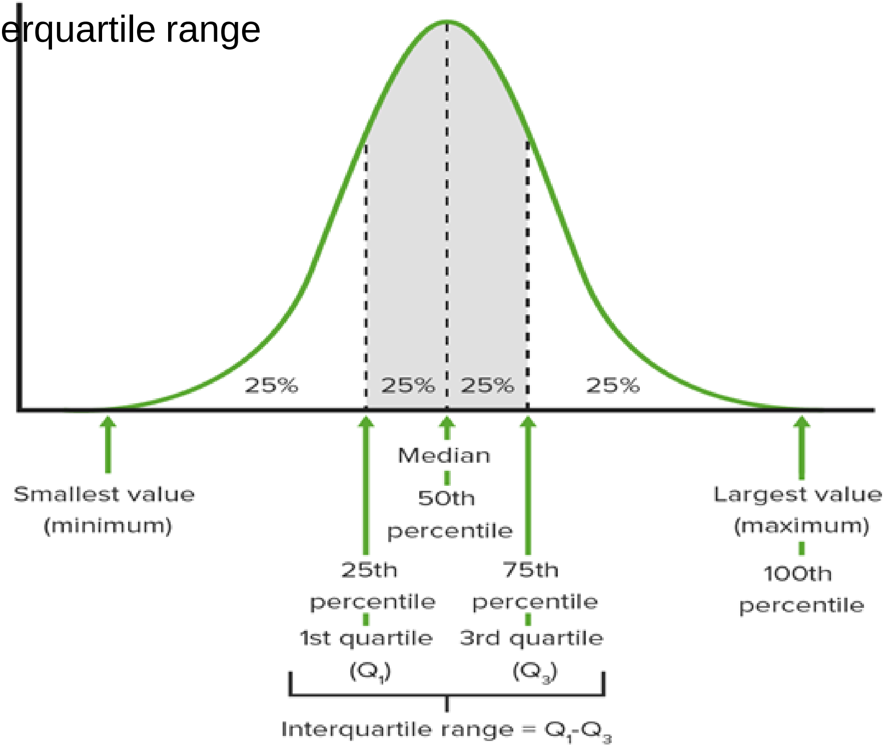
(4) Kurtosis

Measures of central tendency help you find the middle, or the average, of a data set.
The 3 most common measures of central tendency are the mode, median, and mean.
Mode: the most frequent value.
Median: the middle number in an ordered data set.
Mean: the sum of all values divided by the total number of values.

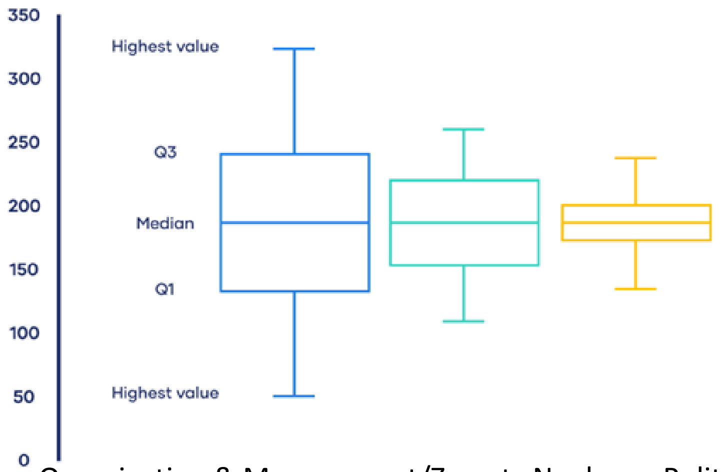


Measures of central tendency
Mean
Mode
Median
Example
Dataset = 7, 3, 4, 1, 7, 6
Summing up all the values in the data-set and dividing by the total number of values
Most common value
Arrange in order and pick the middle value
Mean = (7+3+4+1+7+6)/6
= 28/6
Mode = 7, 3, 4, 1, 7, 6
= 7
Median = 7, 7, 6, 4, 3, 1
= 6+4 / 2

interquartile range



The interquartile range gives you the spread of the middle of your distribution.



[2] WOLNIAK, Radosław. "THE CONCEPT OF DESCRIPTIVE ANALYTICS." Scientific Papers of Silesian University of Technology. Organization & Management/Zeszyty Naukowe Politechniki Slaskiej. Seria Organizacji i Zarzadzanie 172 (2023).

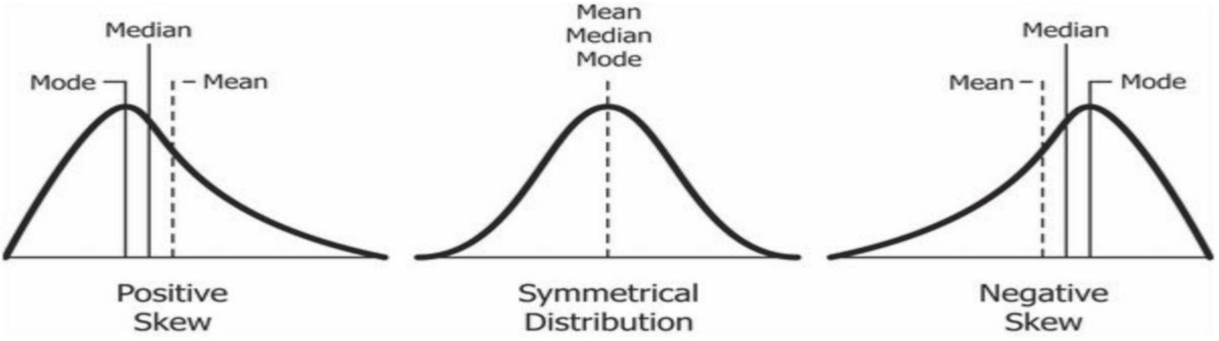
Introduction

Descriptive Analytics techniques

Skewness

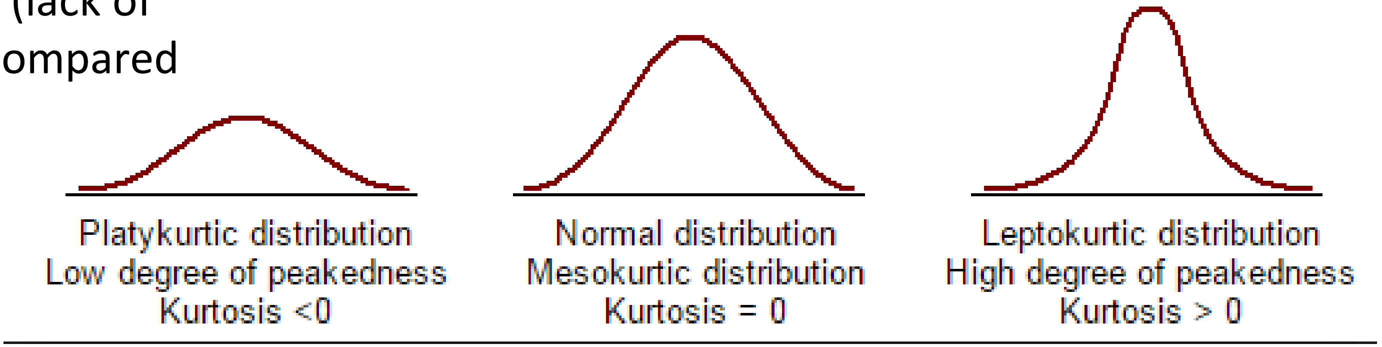
$$Skewness = \frac{3\,(Mean - Median)}{Std\ \ Deviation}$$

- Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean
- A perfectly symmetrical data set will have a skewness of 0. Example: The normal distribution has a skewness of 0
- The skewness value can be positive, zero, negative, or undefined



Kurtosis

- Kurtosis describes whether the data is light-tailed (lack of outliers) or heavy-tailed (outliers present) when compared to a Normal distribution.
- There are three kinds of Kurtosis:

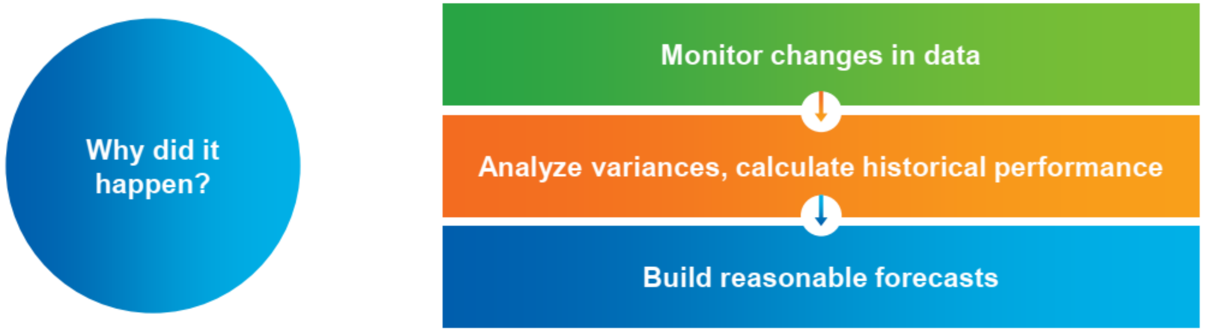


[2] WOLNIAK, Radosław. "THE CONCEPT OF DESCRIPTIVE ANALYTICS." Scientific Papers of Silesian University of Technology. Organization & Management/Zeszyty Naukowe Politechniki Slaskiej. Seria Organizacji i Zarzadzanie 172 (2023).

Introduction

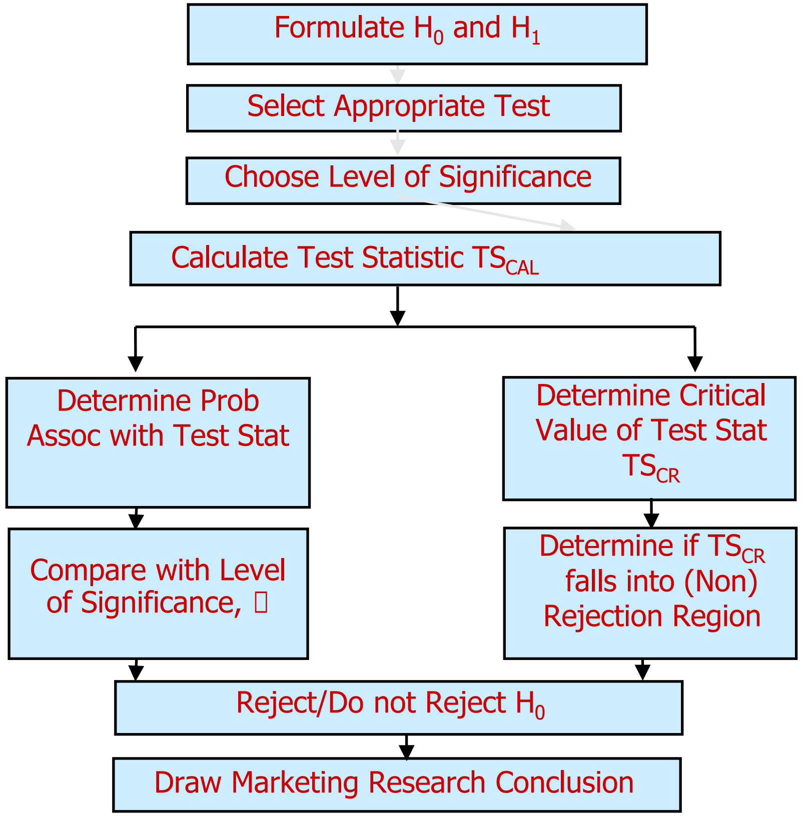
Types of Data Analytics--------> Diagnostic Analytics

- Diagnostic analytics helps address the question of why something happened by analyzing data.
- This techniques are **(1) Hypothesis Testing (2) Root Cause Analysis and (3) Anomaly Detection**, aiming to identify the cause-and-effect relationships behind the observed trends.



Why did it happen?
Monitor changes in data
Analyze variances, calculate historical performance
Build reasonable forecasts

Hypothesis Testing



Formulate H_0 and H_1
Select Appropriate Test
Choose Level of Significance
Calculate Test Statistic TS_{CAL}
Determine Prob Assoc with Test Stat
Determine Critical Value of Test Stat TS_{CR}
Compare with Level of Significance,
Determine if TS_{CR} falls into (Non) Rejection Region
Reject/Do not Reject H_0
Draw Marketing Research Conclusion

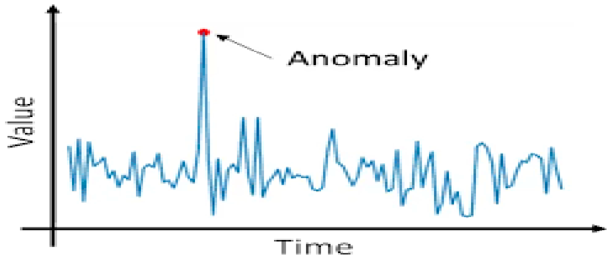
[3] https://online.hbs.edu/blog/post/diagnostic-analytics

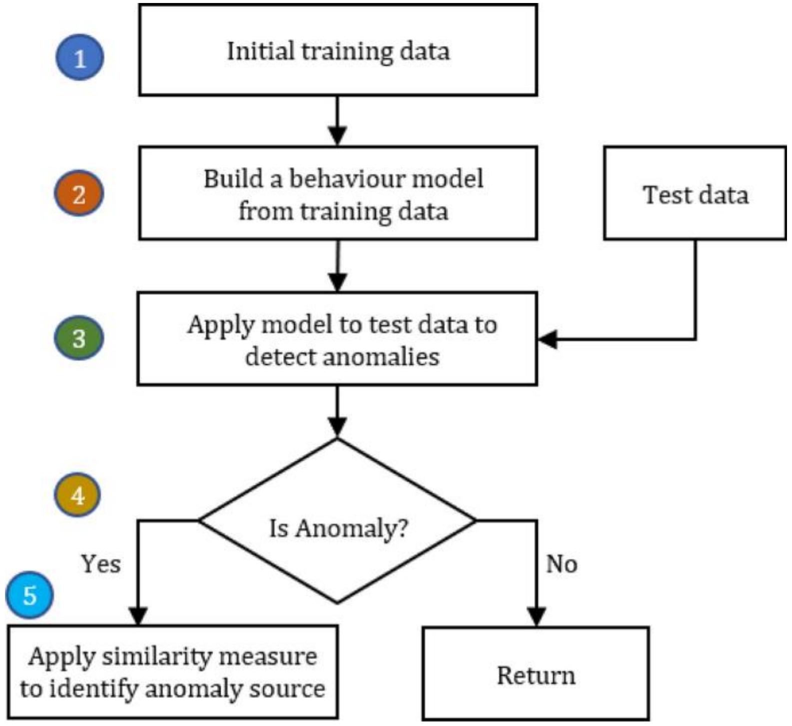
[4] WOLNIAK, Radosław, and Wes GREBSKI. "THE CONCEPT OF DIAGNOSTIC ANALYTICS."

Introduction

Diagnostic Analytics Techniques

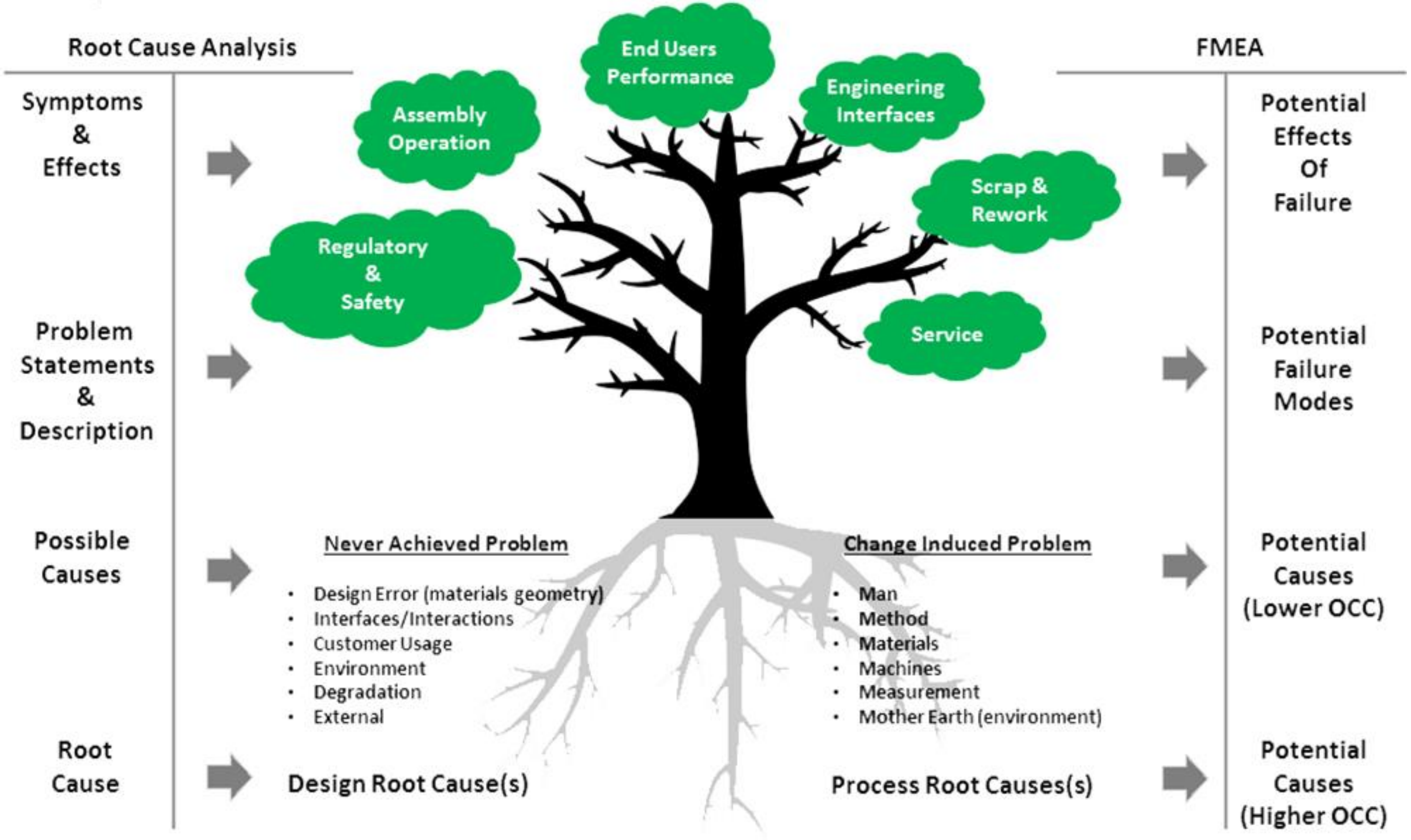
Anomaly Detection





1
Initial training data
2
Build a behaviour model from training data
Test data
3
Apply model to test data to detect anomalies
4
Is Anomaly?
Yes
No
5
Apply similarity measure to identify anomaly source
Return

Route Cause Analysis



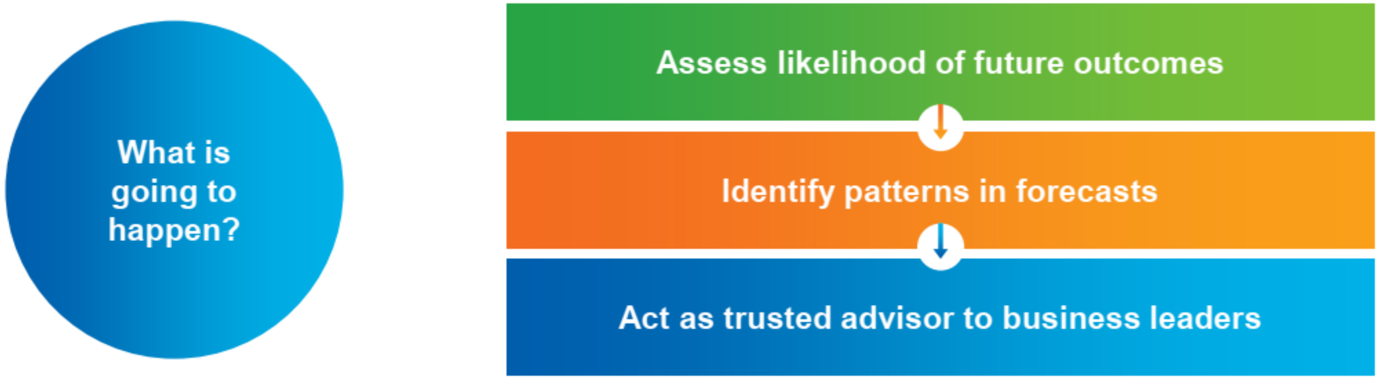
Source: http://theroutex.com/root-cause-analysis-rca/

[5] N assif, Ali Bou, Manar Abu Talib, Qassim Nasir, and Fatima Mohamad Dakalbab. "Machine learning for anomaly detection: A systematic review." Ieee Access 9 (2021): 78658-78700.

Introduction

Types of Data Analytics--------> Predictive Analytics

- Predictive analytics is the process of using data to forecast future outcomes. The process uses data analysis, machine learning, artificial intelligence, and statistical models to find patterns that might predict future behavior.
- Predictive analytics help banks and financial institutions to predict consumer behaviors and preferences.
- Understanding customer patterns allows businesses to gain a competitive advantage in forecasting, planning, and making decisions aligning with the best interests of their clients.



What is going to happen?

Assess likelihood of future outcomes

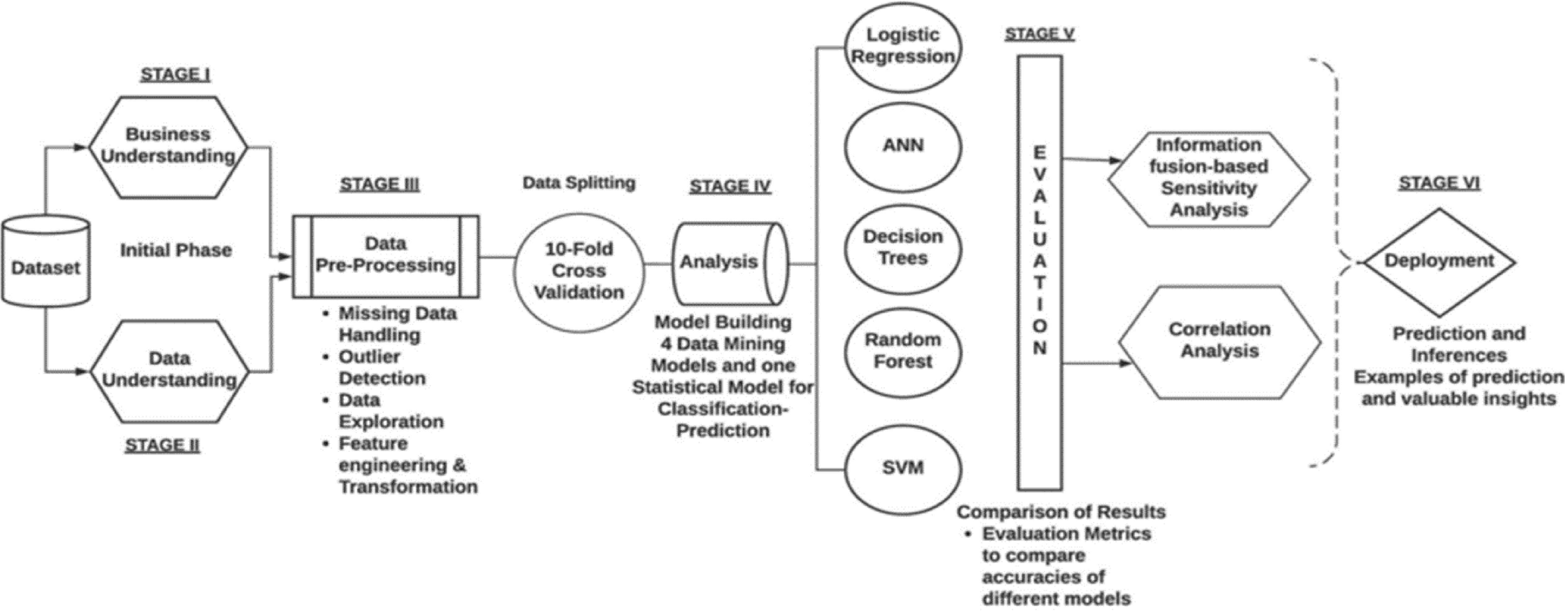
Identify patterns in forecasts

Act as trusted advisor to business leaders

[6] Kumar, Vaibhav, and M. L. Garg. "Predictive analytics: a review of trends and techniques." *International Journal of Computer Applications* 182, no. 1 (2018): 31-37.

Introduction

Predictive Analytics Process Flow

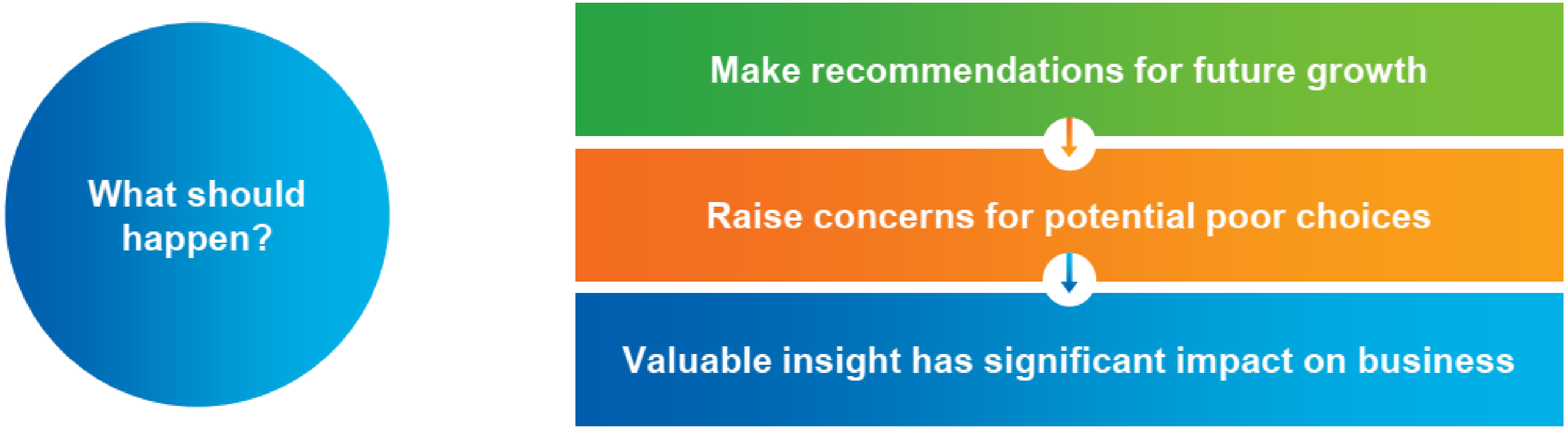


[7] Nasir, Murtaza, Nichalin Summerfield, Ali Dag, and Asil Oztekin. "A service analytic approach to studying patient no-shows." Service Business 14 (2020): 287-313.

Introduction

Types of Data Analytics--------> Prescriptive Analytics

- Prescriptive analytics is a statistical method that focuses on finding the ideal way forward or action necessary for a particular scenario, based on data.
- Prescriptive analytics uses both descriptive and predictive analytics but the focus here remains on actionable insights rather than data monitoring.
- In banking, prescriptive analytics can help optimize operational processes and decision-making. For instance, banks can use prescriptive models to determine the most profitable pricing strategies, allocate resources efficiently, or optimize loan portfolios.



What should happen?

Make recommendations for future growth

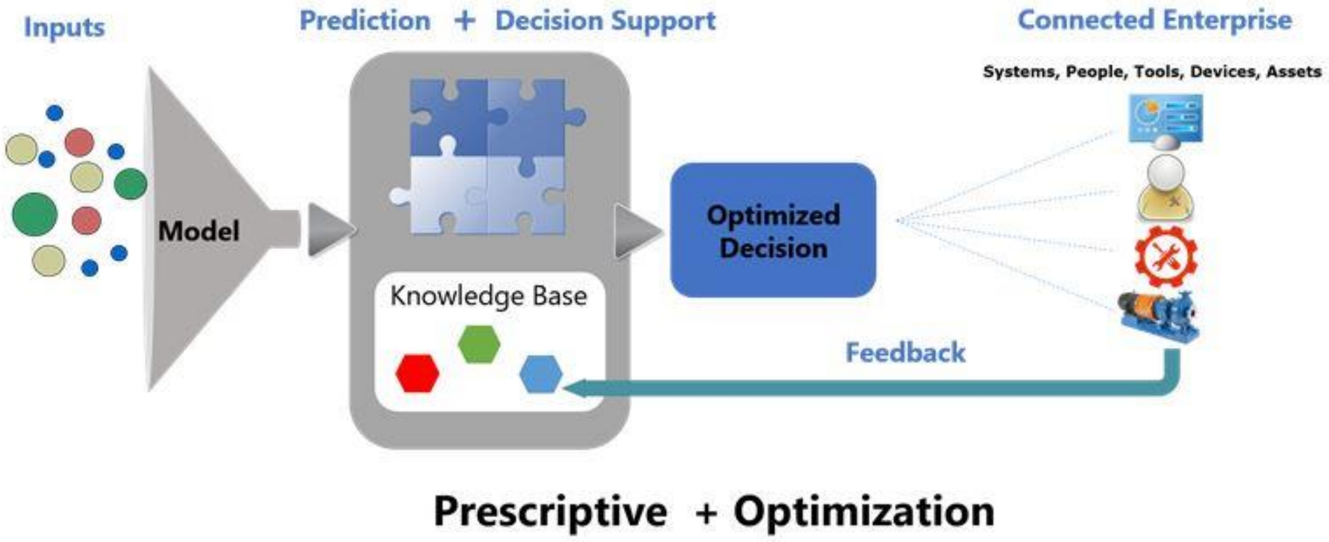
Raise concerns for potential poor choices

Valuable insight has significant impact on business

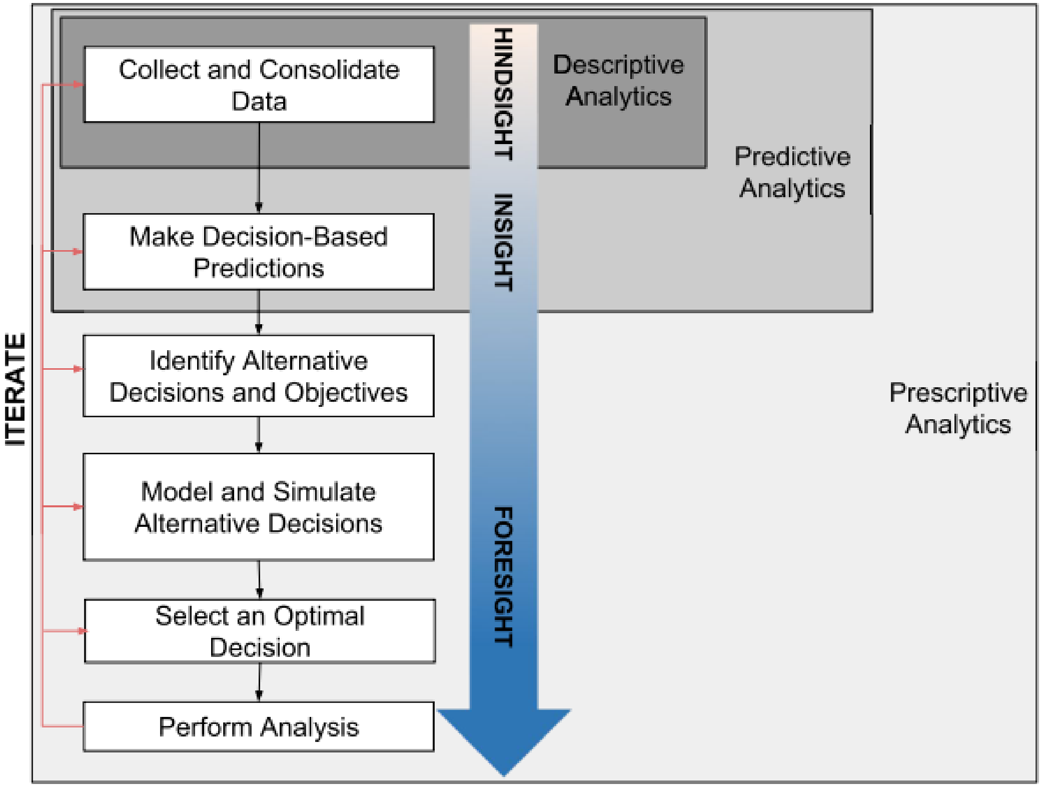
[8] Lepenioti, Katerina, Alexandros Bousdekis, Dimitris Apostolou, and Gregoris Mentzas. "Prescriptive analytics: Literature review and research challenges." International Journal of Information Management 50 (2020): 57-70.

Introduction

Prescriptive Analytics Techniques


Inputs
Prediction + Decision Support
Connected Enterprise
Systems, People, Tools, Devices, Assets
Model
Knowledge Base
Optimized Decision
Feedback
Prescriptive + Optimization

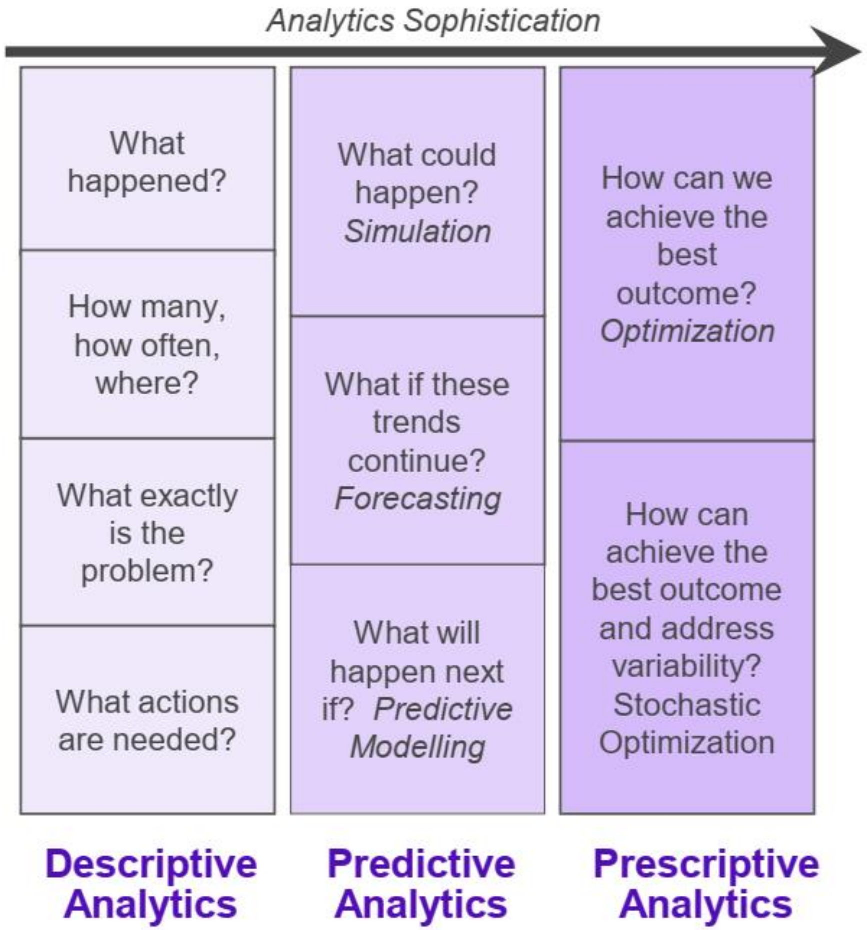
https://www.arcweb.com/industry-best-practices/how-prescriptive-analytics-defined


Collect and Consolidate Data
Make Decision-Based Predictions
Identify Alternative Decisions and Objectives
Model and Simulate Alternative Decisions
Select an Optimal Decision
Perform Analysis
ITERATE
HINDSIGHT
INSIGHT
FORESIGHT
Descriptive Analytics
Predictive Analytics
Prescriptive Analytics

[9] Frazzetto, Davide, Thomas Dyhre Nielsen, Torben Bach Pedersen, and Laurynas Šikšnys. "Prescriptive analytics: a survey of emerging trends and technologies." The VLDB Journal 28 (2019): 575-595.

Introduction

Analytics is evolving in sophistication to now encompass Cognitive Capabilities



Analytics Sophistication

What happened?

How many, how often, where?

What exactly is the problem?

What actions are needed?

What could happen? *Simulation*

What if these trends continue? *Forecasting*

What will happen next if? *Predictive Modelling*

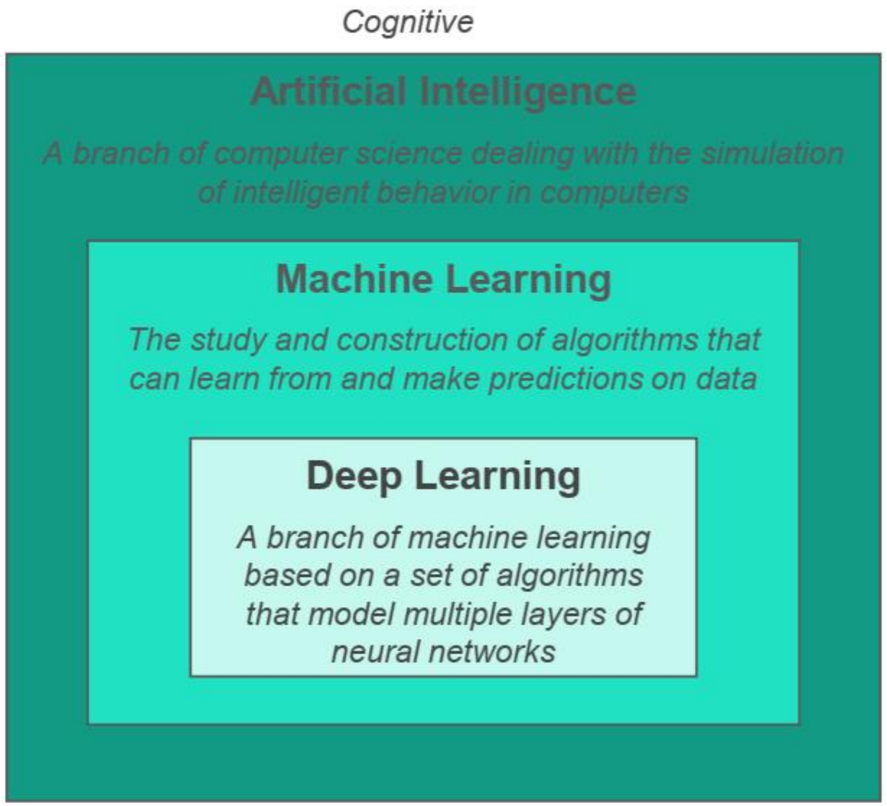
How can we achieve the best outcome? *Optimization*

How can achieve the best outcome and address variability? Stochastic Optimization

Descriptive Analytics

Predictive Analytics

Prescriptive Analytics



Cognitive

Artificial Intelligence

A branch of computer science dealing with the simulation of intelligent behavior in computers

Machine Learning

The study and construction of algorithms that can learn from and make predictions on data

Deep Learning

A branch of machine learning based on a set of algorithms that model multiple layers of neural networks

[10] Handfield, Robert, Seongkyoon Jeong, and Thomas Choi. "Emerging procurement technology: data analytics and cognitive analytics." International journal of physical distribution & logisti management 49, no. 10 (2019): 972-1002.

Introduction to Basics of Big Data

Where Is This "Big Data" Coming From ?

16+ TBs
of tweet data
every day

? TBs of
data every
day

25+ TBs of
log data
every day

40 billion RFID tags
today
(1.3B in 2005)

4.6
billion
camera
phones
world
wide

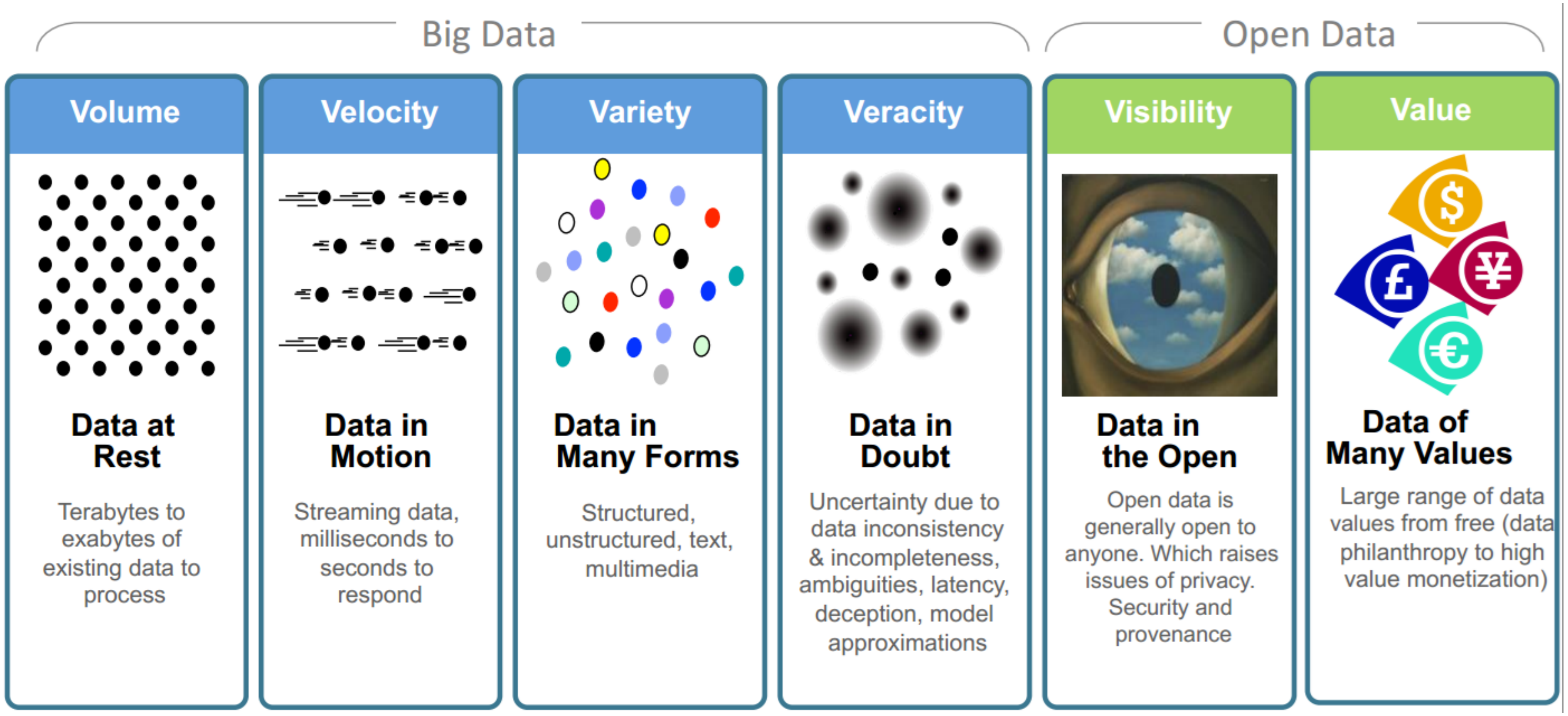
100s of
millions of
GPS
enabled
devices
sold
annually

6+ billion
people on
the Web
by end
2019

76 million smart
meters in 2009…
400M by 2019

Big Data: Vs??????????

Big Data: 6V in Summary



Big Data

Open Data

Volume

Data at Rest

Terabytes to exabytes of existing data to process

Velocity

Data in Motion

Streaming data, milliseconds to seconds to respond

Variety

Data in Many Forms

Structured, unstructured, text, multimedia

Veracity

Data in Doubt

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Visibility

Data in the Open

Open data is generally open to anyone. Which raises issues of privacy. Security and provenance

Value

Data of Many Values

Large range of data values from free (data philanthropy to high value monetization)

Transforming Energy and Utilities through Big Data & Analytics. By Anders Quitzau@IBM

Other V’s

- **Variability**
  - Variability refers to data whose meaning is constantly changing. This is particularly the case when gathering data relies on language processing.
- **Viscosity**
  - This term is sometimes used to describe the latency or lag time in the data relative to the event being described. We found that this is just as easily understood as an element of Velocity.
- **Virality**
  - Defined by some users as the rate at which the data spreads; how often it is picked up and repeated by other users or events.
- **Volatility**
  - Big data volatility refers to how long is data valid and how long should it be stored. You need to determine at what point is data no longer relevant to the current analysis.
- More V’s in the future …

Can you spot some difference between Traditional Analytics and Big Data Analytics ?

The Big Data Approach to Analytics is Different

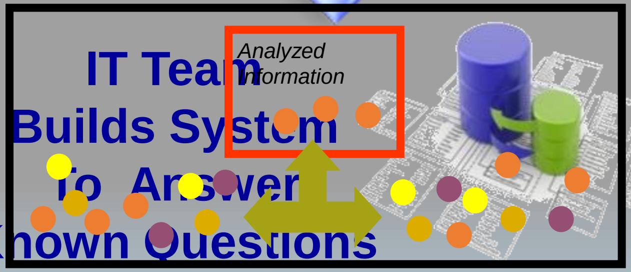
Traditional Analytics

Structured & Repeatable
Structure built to store data

Business Users Determine Questions

Analyzed Information
Available Information

Capacity constrained down sampling of available information

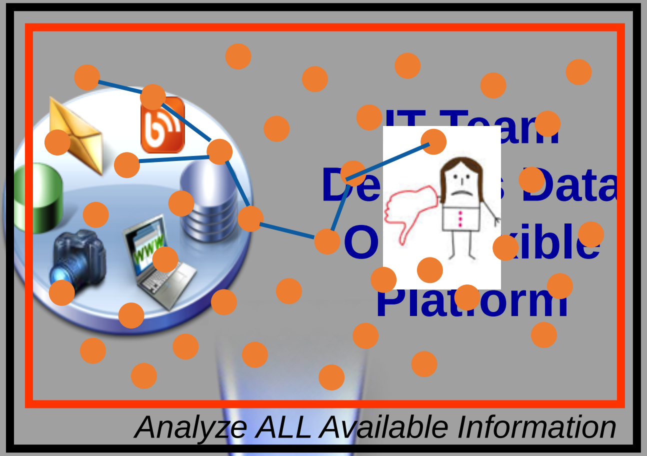


IT Team Builds System To Answer Known Questions

Carefully cleanse a small information before any analysis

Big Data Analytics

Iterative & Exploratory
Data is the structure



IT Team Delivers Data On Flexible Platform

Whole population analytics connects the dots

Business Users Explore and Ask Any Question

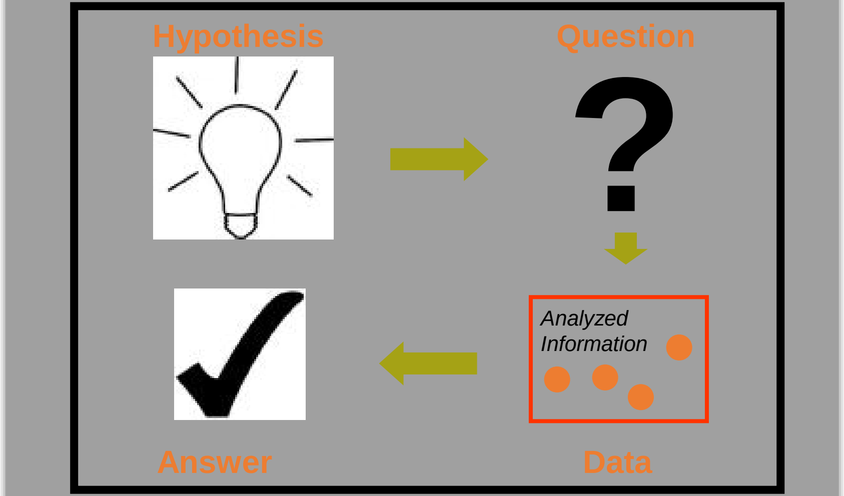
Analyzed Information

Analyze information as is & cleanse as needed & existing repeatable

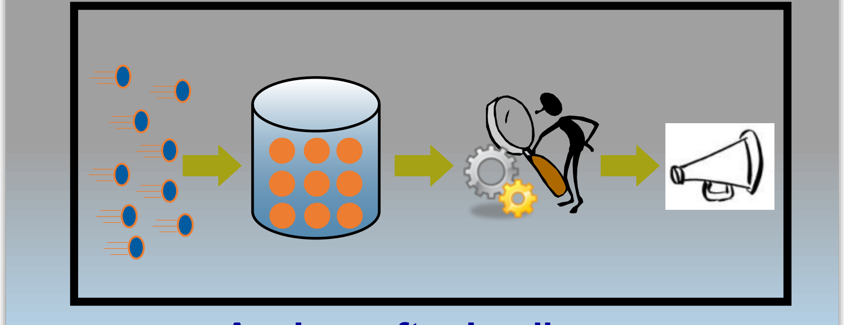
The Big Data Approach to Analytics is Different

Traditional Analytics

Structured & Repeatable
Structure built to store data



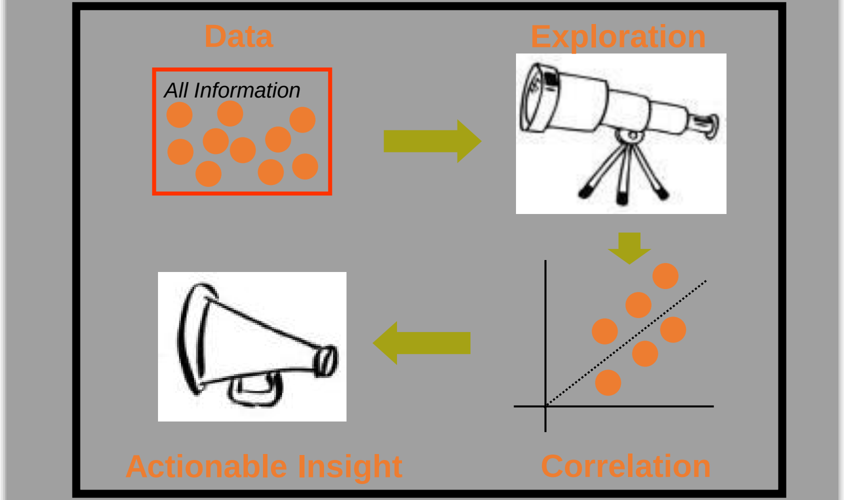
Start with hypothesis
Test against selected data



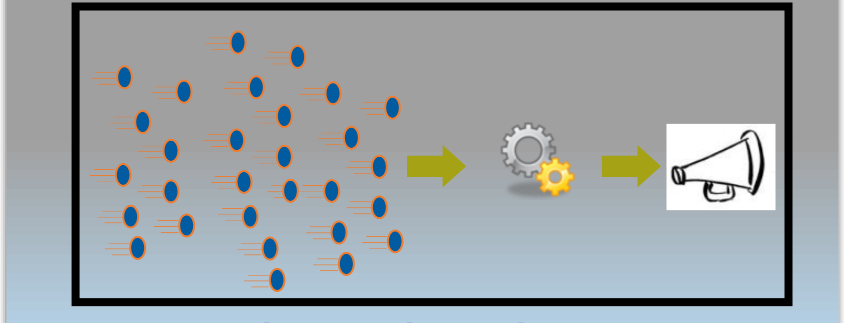
Analyze after landing…

Big Data Analytics

Iterative & Exploratory
Data is the structure



Data leads the way
Explore *all* data, identify correlations



Analyze in motion…

Big Data Trends for 2024

01

Data Gravity

VOLUME

VELOCITY

Democratizing Real-Time Data

02

03

Augmented Data Management

VARIETY

VERACITY

Identity Authenticity

04

05

Data Monetization

VALUE

VIRTUE

Adaptive Data Governance

06

07

AI-Driven Storytelling & Augmented Analytics

VISUALIZATION

VIRALITY

Data Marketplace

08

09

DevOps – DataOps – XOps

VISCOSITY