TELKOM
DIGITAL TALENT
INCUBATOR 2020

Telkom University
Telkom Indonesia

DATA SCIENTIST

Eva Nurhazizah

Organized by:

CAE

TELKOM
DIGITAL TALENT
INCUBATOR 2020

Modul 1: Introduction to Big Data and Data Science

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Module Overview

Topics

- Big Data Phenomenon
- Big Data Characteristics
- Data Science Definition and Application
- Data Science Process

Activities

- Group Discussion

TELKOM DIGITAL TALENT INCUBATOR 2020

Module Objectives

- Understand Big Data and its characteristics
- Understands what its data science and its application
- Understand data science process

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Knowledge Check

- Let's do some pretest quiz:

Please open kahoot.it

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Our World Today

Presiden Jokowi dan Big Data: Kenapa Data lebih Berharga dari Minyak?

September 02, 2019

Dalam pidato kenegaraan Presiden Jokowi pada 16 Agustus 2019, dinyatakan bahwa "data adalah jenis kekayaan baru bangsa kita, kini data lebih berharga dari minyak". Pernyataan ini menunjukkan bahwa pemerintah RI dibawah pimpinan Presiden Jokowi telah menyadari betapa bernilainya potensi yang terkandung dalam suatu himpunan data. Dalam konteks nasional, data dapat dieksploitasi guna mewujudkan kemakmuran bangsa seperti [illegible] eksploitasi minyak bumi yang mendatangkan kemakmuran di negara-negara

KOMPAS.com PREMIUM JELAJAHI

ADVERTORIAL

Pentingnya "Big Data" untuk Pertumbuhan Bisnis

Senin, 28 Agustus 2017 | 14:53 WIB

Dalam dunia bisnis, big data (BD) banyak digunakan oleh para pemasar. Gunanya, untuk membantu mereka merumuskan strategi yang tepat untuk menembus pasar.

Dilansir dari *Smart-money.co*, Managing Partner Alpha JWC Ventures Will Ongkowidjaja mengatakan teknologi kecerdasan buatan (*artificial intelligent*/AI) dan analisis big data

Harvard Business Review Sign In

Artwork: Tamar Cohen, Andrew J Buboltz, 2011, silk screen on a page from a high school yearbook, 8.5" x 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

From the October 2012 Issue

Save Share $8.95

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt

Forbes

10,487 views | Mar 13, 2019, 02:12am EDT

Why Every Company Needs A Data Strategy For 2019

Bernard Marr Contributor

Enterprise Tech

Data matters to any company, regardless of size, sector or type. In fact, data is one of your biggest business assets, alongside your products, services, intellectual property, and people. But to get the most out of data, it's vital you have a data strategy in place.

Why Every Company Needs A Data Strategy For 2019

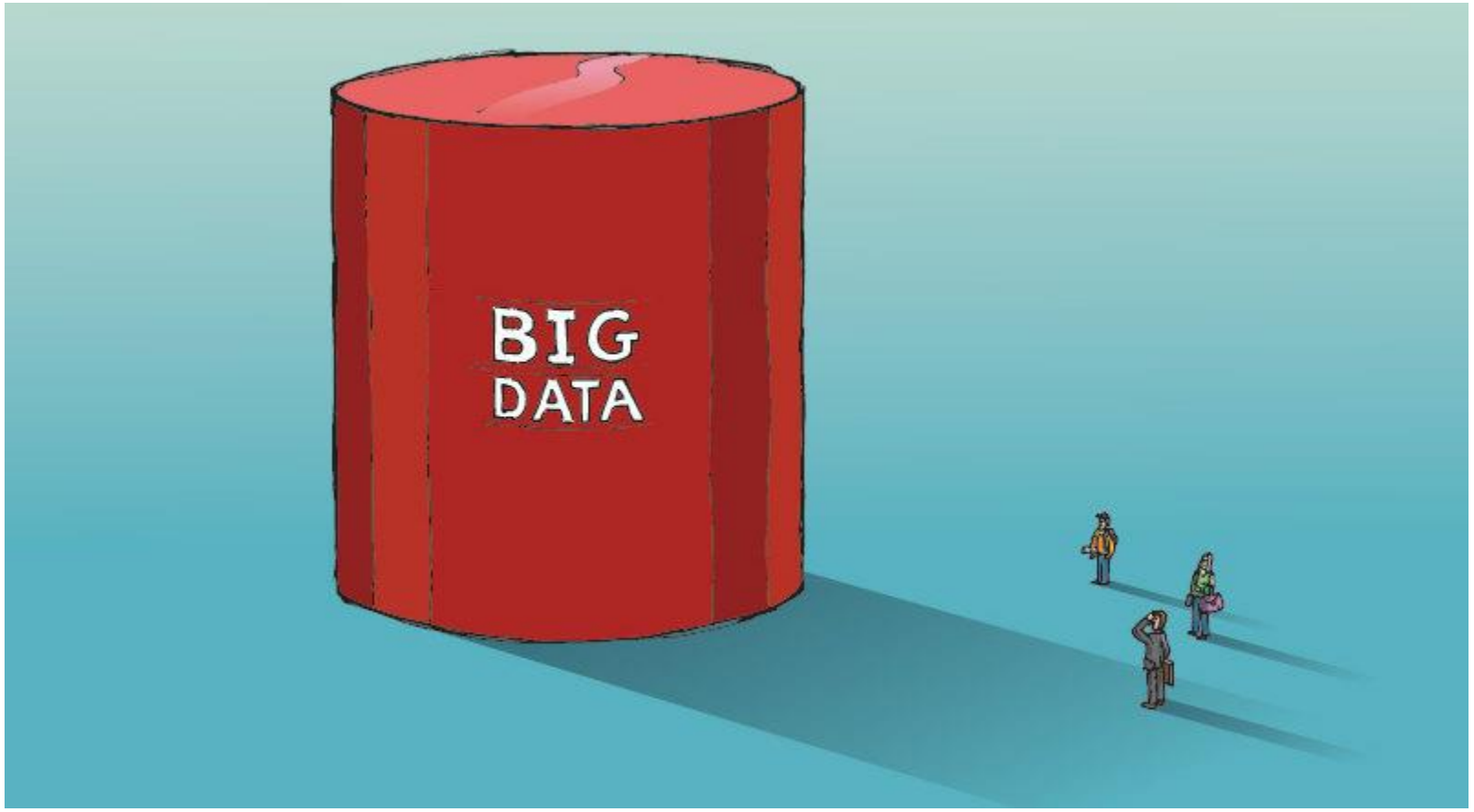
ADOBE STOCK

Organized by:

CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Big Data Phenomenon



Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

What is Big Data

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Big Data

- Big data is larger, more complex data sets, especially from new data sources.
- These data sets are so voluminous and make traditional data processing software can’t manage them.
- These massive volumes of data can be used to address business problems that you wouldn’t have been able to tackle before.

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

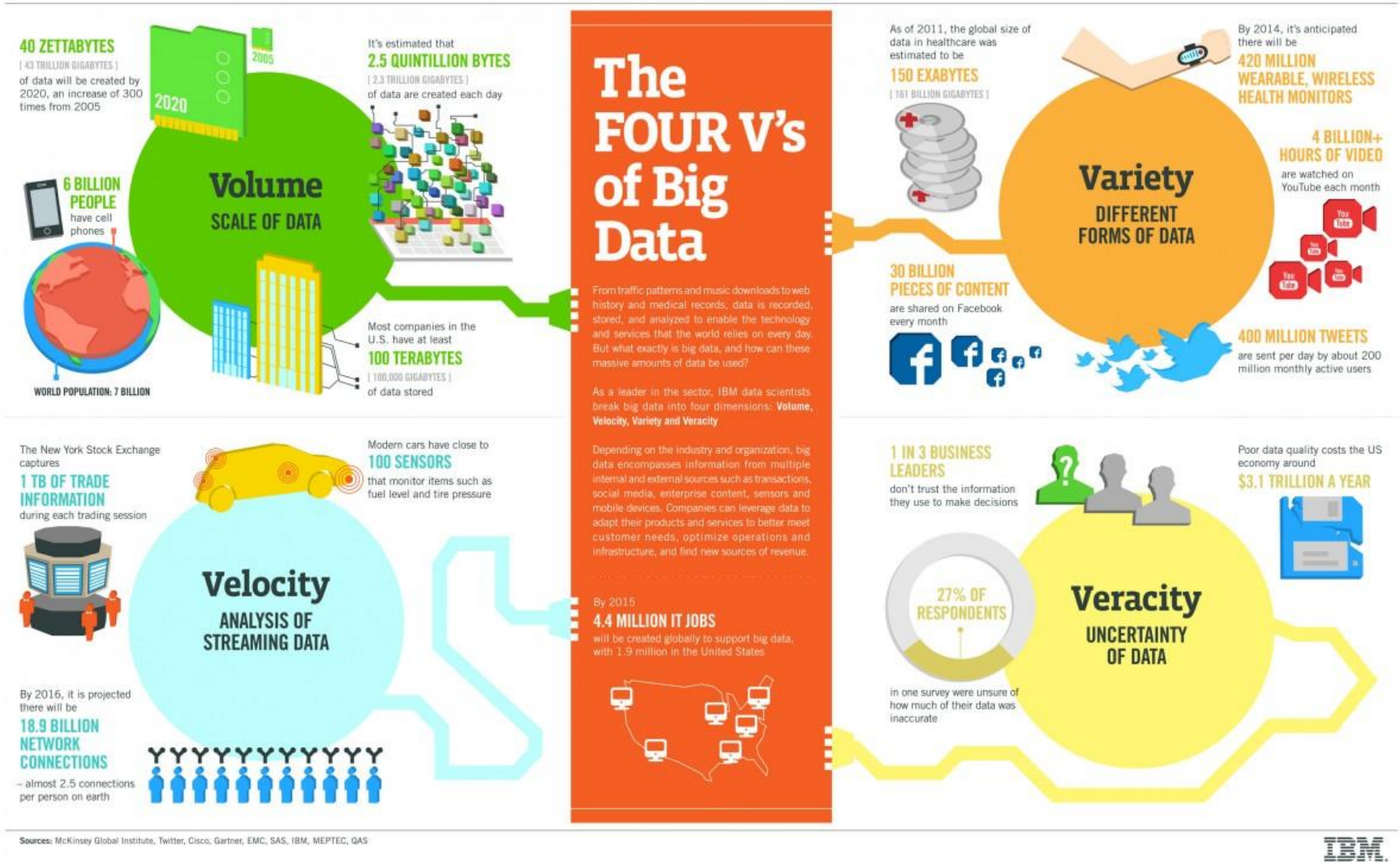
Source of Big Data

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Big Data Characteristic



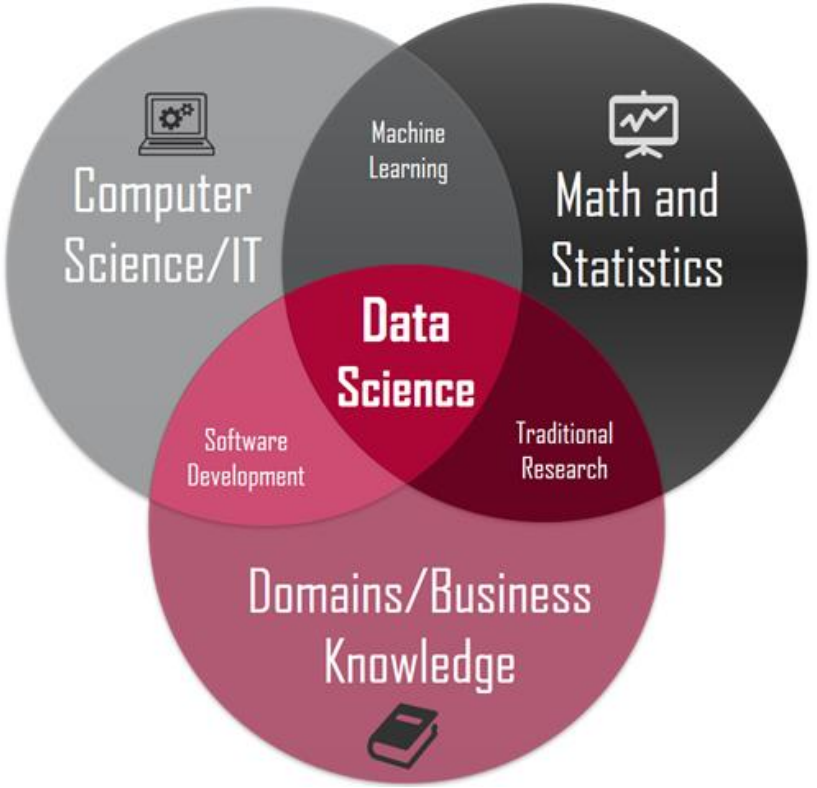
Organized by:

CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Data Science

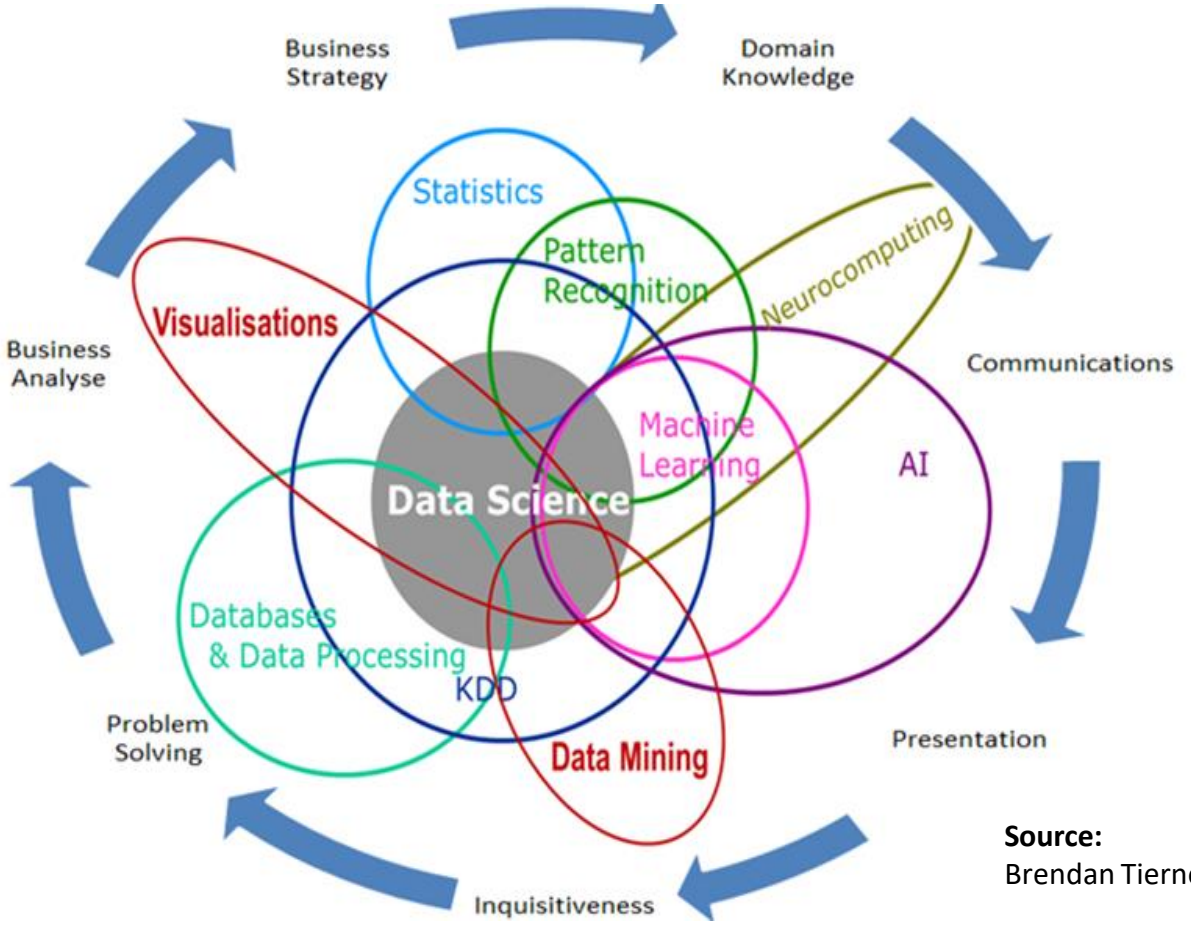
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data.
- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data.
- Data science principles apply to all data – big and small.



Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Data Science Multidisciplinary



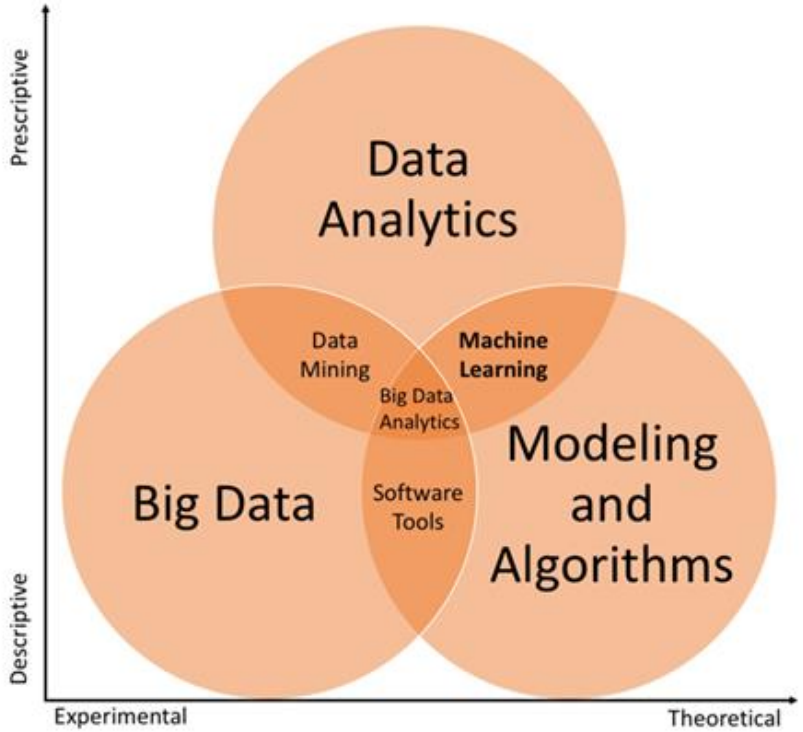
Source:
Brendan Tierney, 2012

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

The Fields of Data Science

- Data science takes all these considerations into account but also takes up other challenges.
- For example:
  - Capturing, cleaning, and transforming of unstructured social media and web data
  - Using big data technologies to store and process big, unstructured datasets



Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

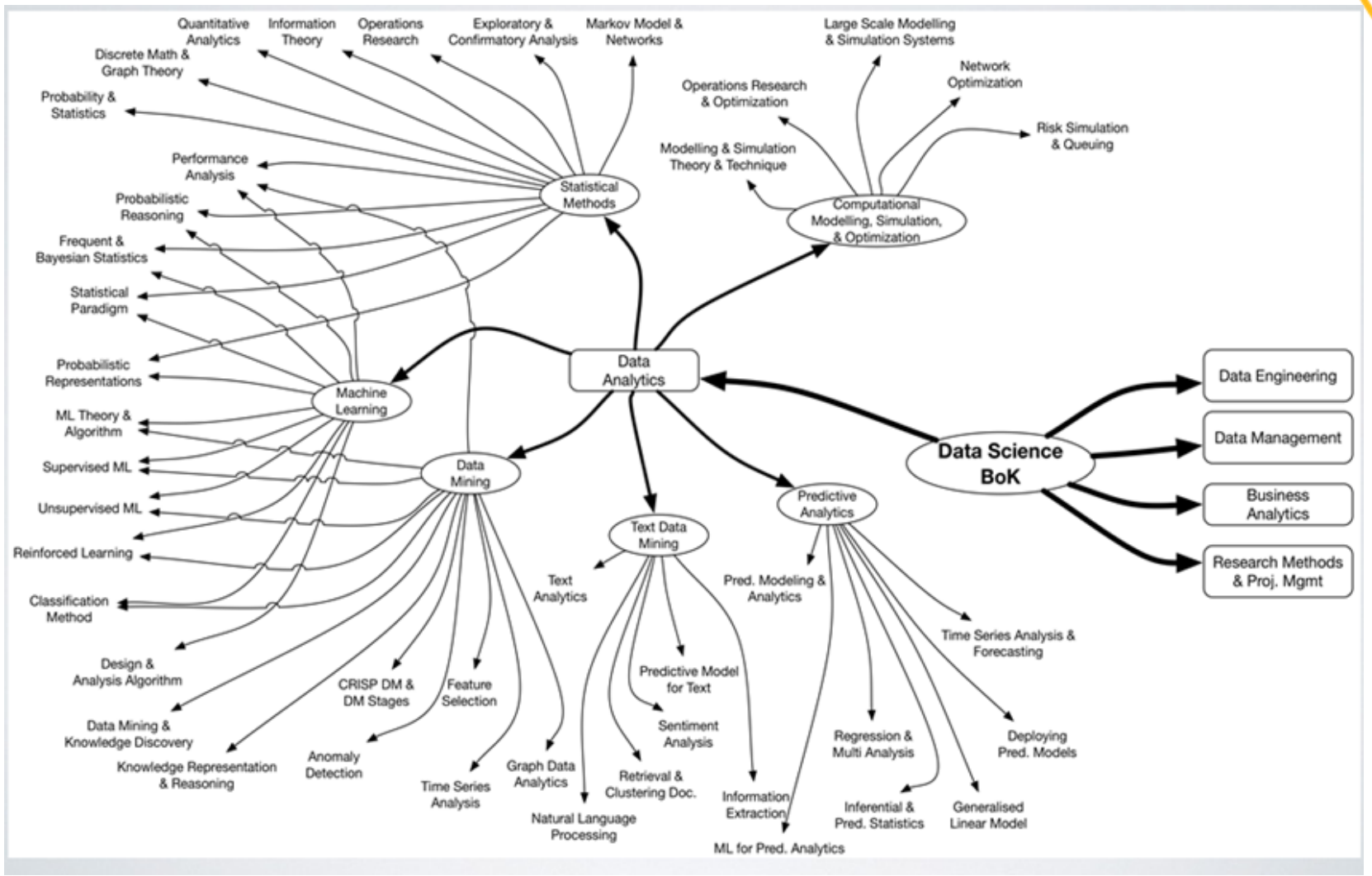
Data Science Body of Knowledge

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, Architeture & Organization, Computational Science, Graphic & Visualization, Information Management, Platform Based Dev., Software Engineering
3	Data Management	General Principle & Concepts in Data Mgmt 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 Mgmt, Scope Mgmt, Quality, Risk Mgmt)
5	Business Analytics	Business Analytics Foundation, Business Analytics Organisation and Enterprise Management	Business Analysis Planning & Monitoring, Requirement Analysis & Design Definition, Requirement Life Cycle Mgmt, Solution Evaluation & Improvements Recommendation
6	Domain Knowledge		

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Data Science Body of Knowledge

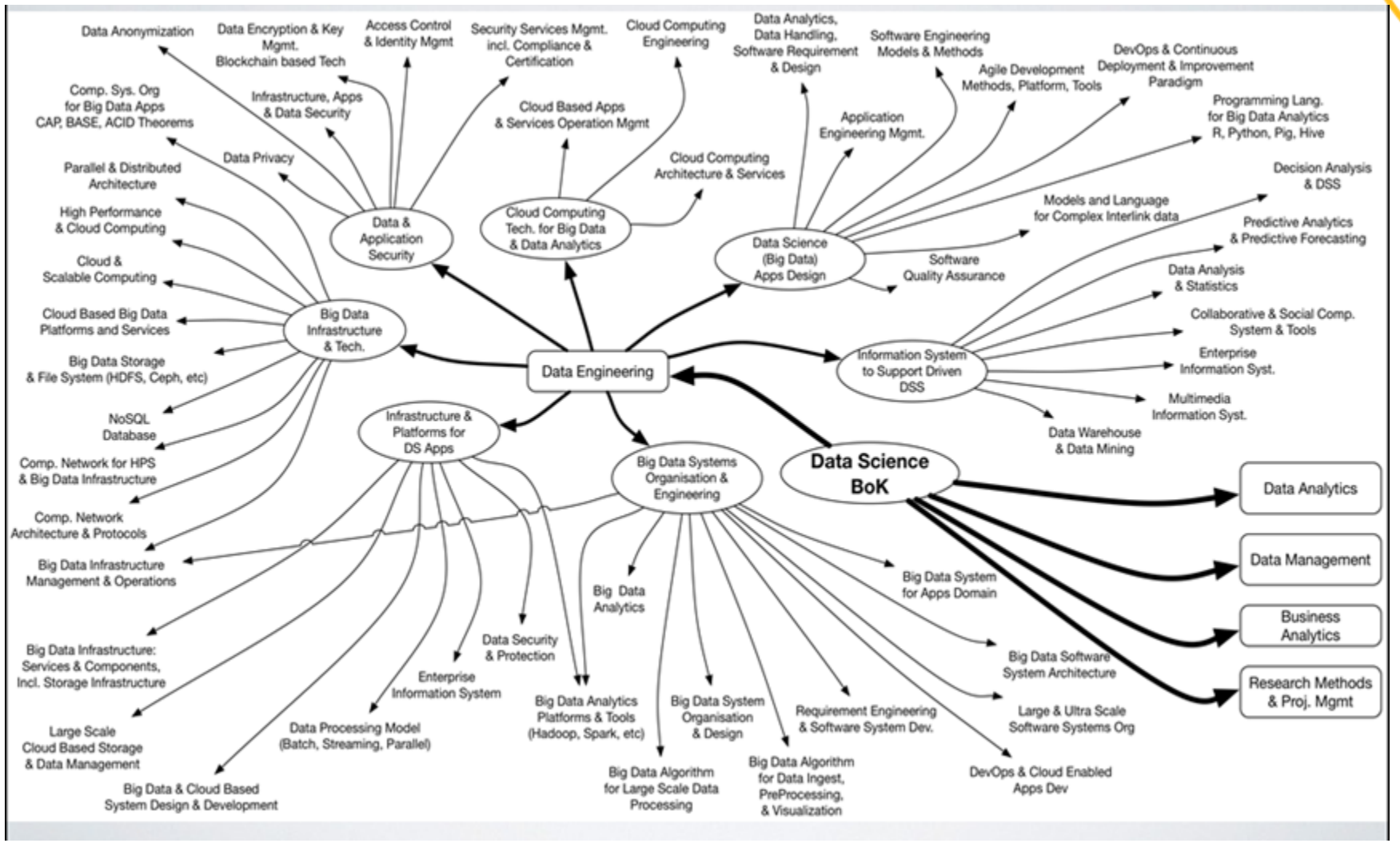


Organized by:

CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Data Science Body of Knowledge



Organized by:

CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Data Science Implementation

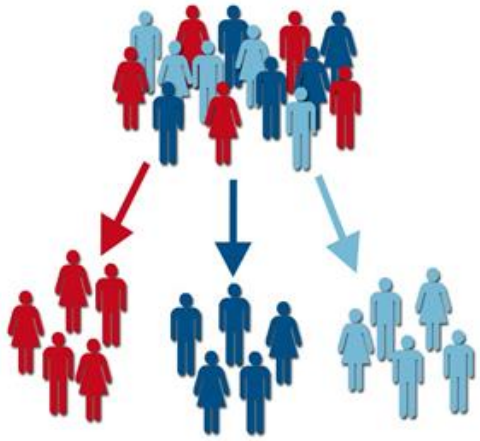
- Empowers management to make better decisions
- Helps identify trends to stay competitive
- Increase the efficiency in handling core tasks and issues
- Helps in selecting target audience
- Identifies and acts upon opportunities

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Data Science Implementation in Marketing

Customer Behavior Data Example: Website visitation intensity, user demographics, searching behavior, etc.



Profiling Customer based on behavioral similarity

Promote product to targeted customer

Organized by:

CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Spotify

Spotify

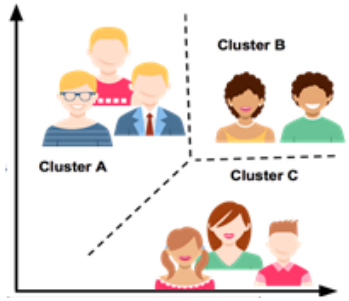
Artist on Spotify

Get the data from Spotify API.

Process the data to extract audio features for each artist.

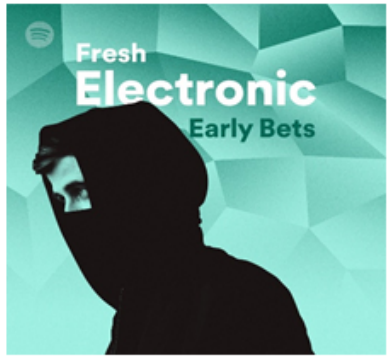
Apply k-means clustering to separate the artists into different groups.

K-Means



Analyze each feature for all the artists.

Well catered songs to each-user music preferences by categorizing the music



Organized by:

CAE

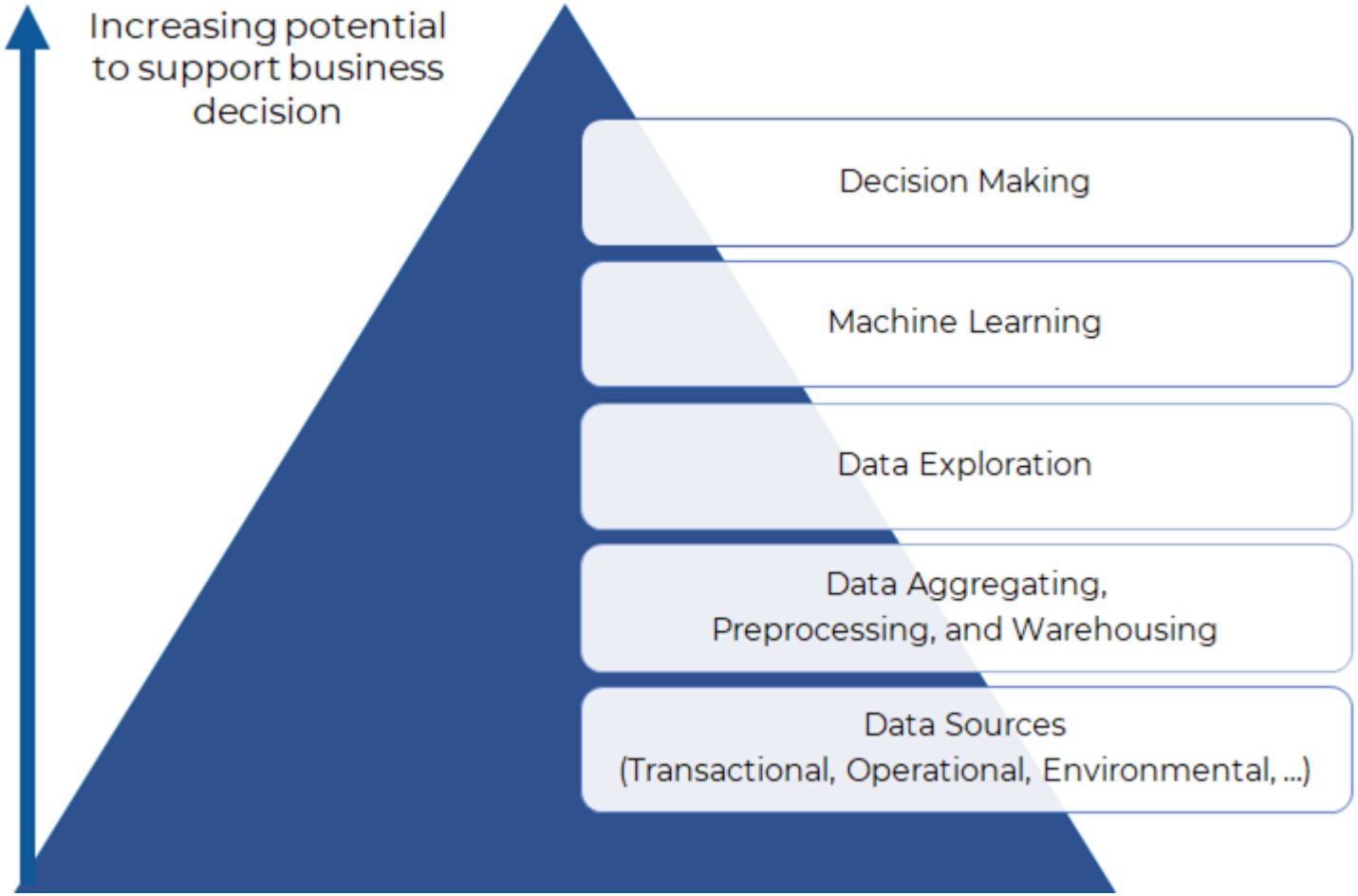
TELKOM DIGITAL TALENT INCUBATOR 2020

Zara

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Data Science Pyramid



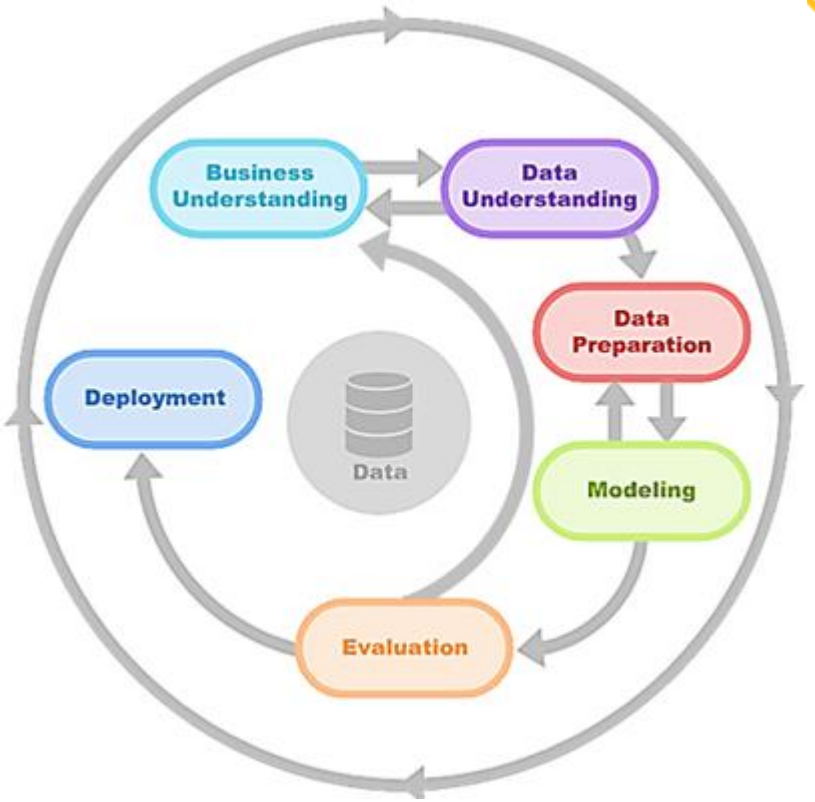
Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Data Science Process: CRISP-DM

- CRISP-DM or Cross Industry Standard Process for Data Mining is a framework that has been commonly used to perform data science processes.

- The CRISP-DM life cycle consists of six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

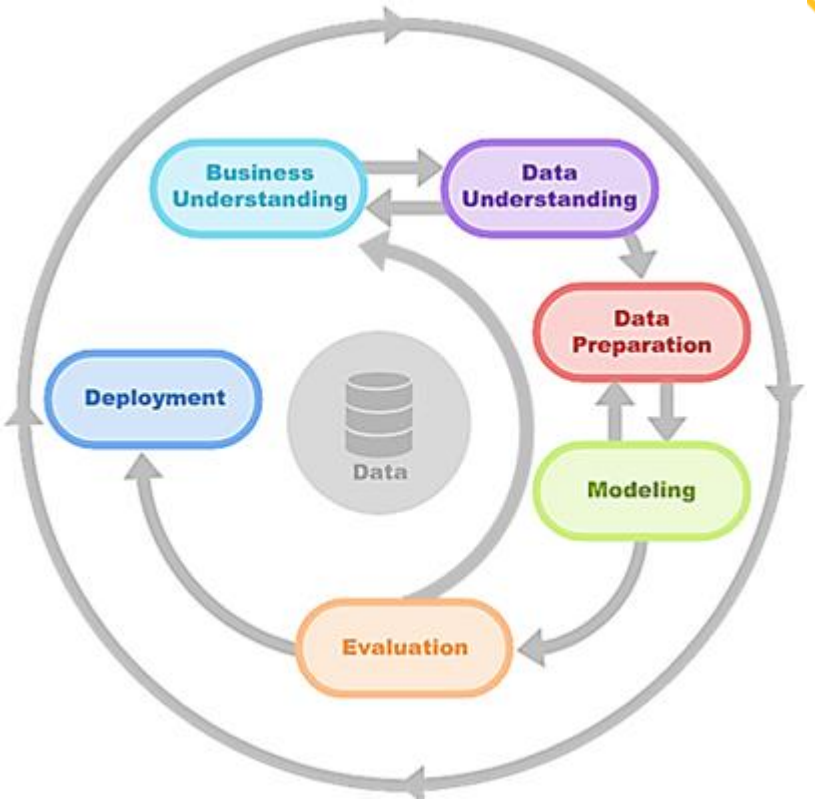


Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

CRISP-DM Steps

- Consists of six phases
- Data are at the center of all data science activities
- Arrows indicating the typical direction of the process and the most frequent dependencies between phases
- The process is semi-structured



Organized by:
CAE

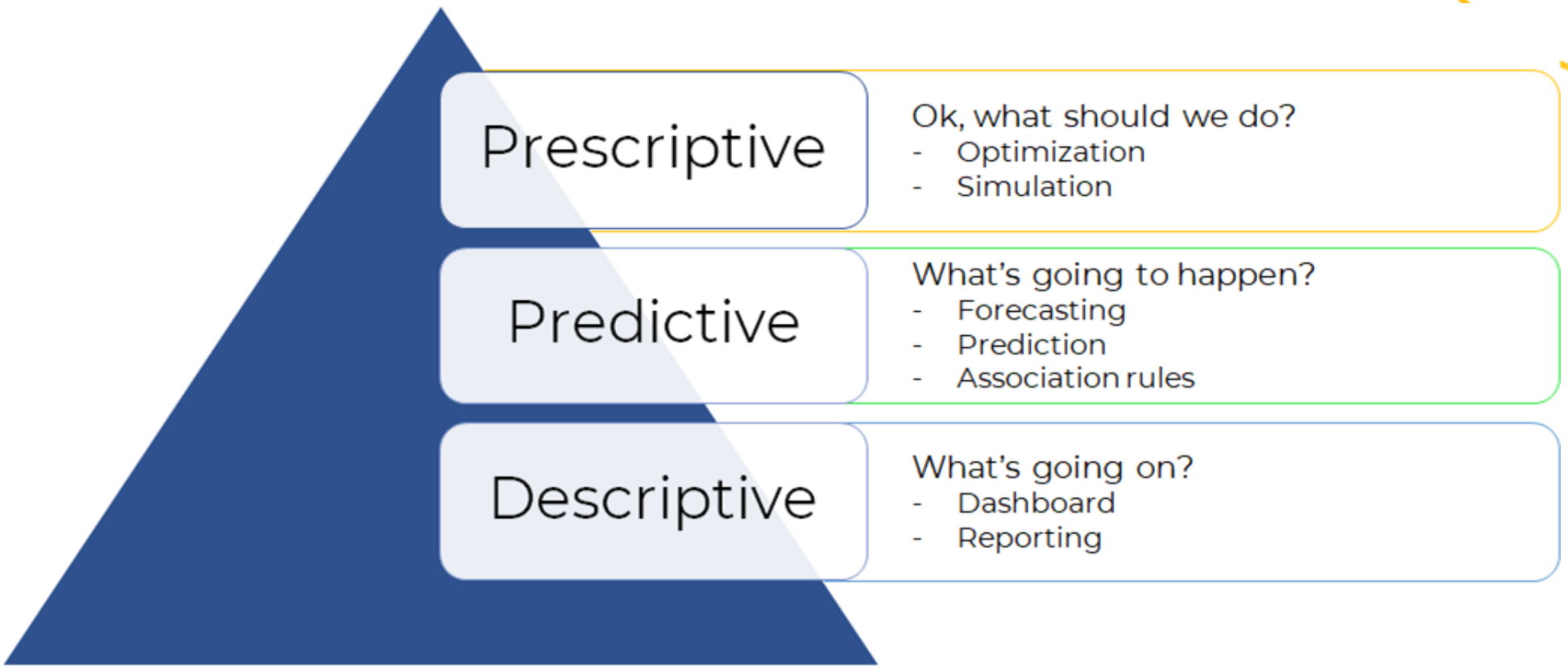
TELKOM DIGITAL TALENT INCUBATOR 2020

Business Understanding

- What outcomes your company expects from data mining?
- Ensure that everyone is on the same page before expending valuable resources
- The process include:
  - Determining Business Objectives
  - Assessing the Situation
  - Determining Data Mining Goals
  - Producing a Project Plan

TELKOM DIGITAL TALENT INCUBATOR 2020

Data Mining Goal



Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Data Understanding

- What data is available to be processed through data mining?
- Avoiding unexpected problems during the next phase (data preparation) which is typically the longest part of a project
- The process include:
  - Collecting Initial Data
  - Describing Data
  - Exploring Data
  - Verifying Data Quality

TELKOM DIGITAL TALENT INCUBATOR 2020

Data Type

Type of Data	Definition	Example
Numeric	True numeric values that allow arithmetic operations	Price, Age
Interval	Values that allow ordering and subtraction, but do not allow other arithmetic operations	Date, Time
Ordinal	Values that allow ordering but do not permit arithmetic	Size measured as small, medium, or large
Categorical	A finite set of values that cannot be ordered and allow no arithmetic	Country, Product type
Binary	A set of just two values	Gender
Textual	Free-form, usually short, text data	Name, Address

TELKOM DIGITAL TALENT INCUBATOR 2020

Verifying Data Quality

Issue	Definition
Missing data	include values that are blank or coded as a non-response (such as $null$, ?, or 999)
Data errors	are usually typographical errors made in entering the data
Measurement errors	include data that are entered correctly but are based on an incorrect measurement scheme
Coding inconsistencies	typically involve nonstandard units of measurement or value inconsistencies, such as the use of both M and male for gender.
Bad metadata	include mismatches between the apparent meaning of a field and the meaning stated in a field name or definition.

TELKOM DIGITAL TALENT INCUBATOR 2020

Data Preparation

- Data preparation is one of the most important and often time-consuming aspects of data mining.
- In fact, it is estimated that data preparation usually takes more than 70% of a project's time and effort.
- The process include:
  - Selecting Data
  - Cleaning Data
  - Constructing New Data
  - Integrating Data
  - Formatting Data

TELKOM DIGITAL TALENT INCUBATOR 2020

Cleaning Data

Issue	Decision
Missing data	Exclude rows or characteristics. Or, fill blanks with an estimated value
Data errors	Use logic to manually discover errors and replace. Or, exclude characteristics
Measurement errors	Decide upon a single coding scheme, then convert and replace values
Coding inconsistencies Bad metadata	Manually examine suspect fields and track down correct meaning

TELKOM DIGITAL TALENT INCUBATOR 2020

Modelling

- Modeling is usually conducted in multiple iterations. There are many ways to look at a given problem.
- Rare for an organization's data mining question to be answered satisfactorily with a single model and a single execution.
- The process include:
  - Selecting Modeling Techniques
  - Generating a Test Design
  - Building the Models
  - Assessing the Model

TELKOM DIGITAL TALENT INCUBATOR 2020

Evaluation

- The key to ensure that the organization can make use of the results obtained.
- Rare for an organization's data mining question to be answered satisfactorily with a single model and a single execution.
- Rare for an organization's data mining question to be answered satisfactorily with a single model and a single execution.
- The process include:
  - Evaluating the Results
  - Review Process
  - Determining the Next Steps

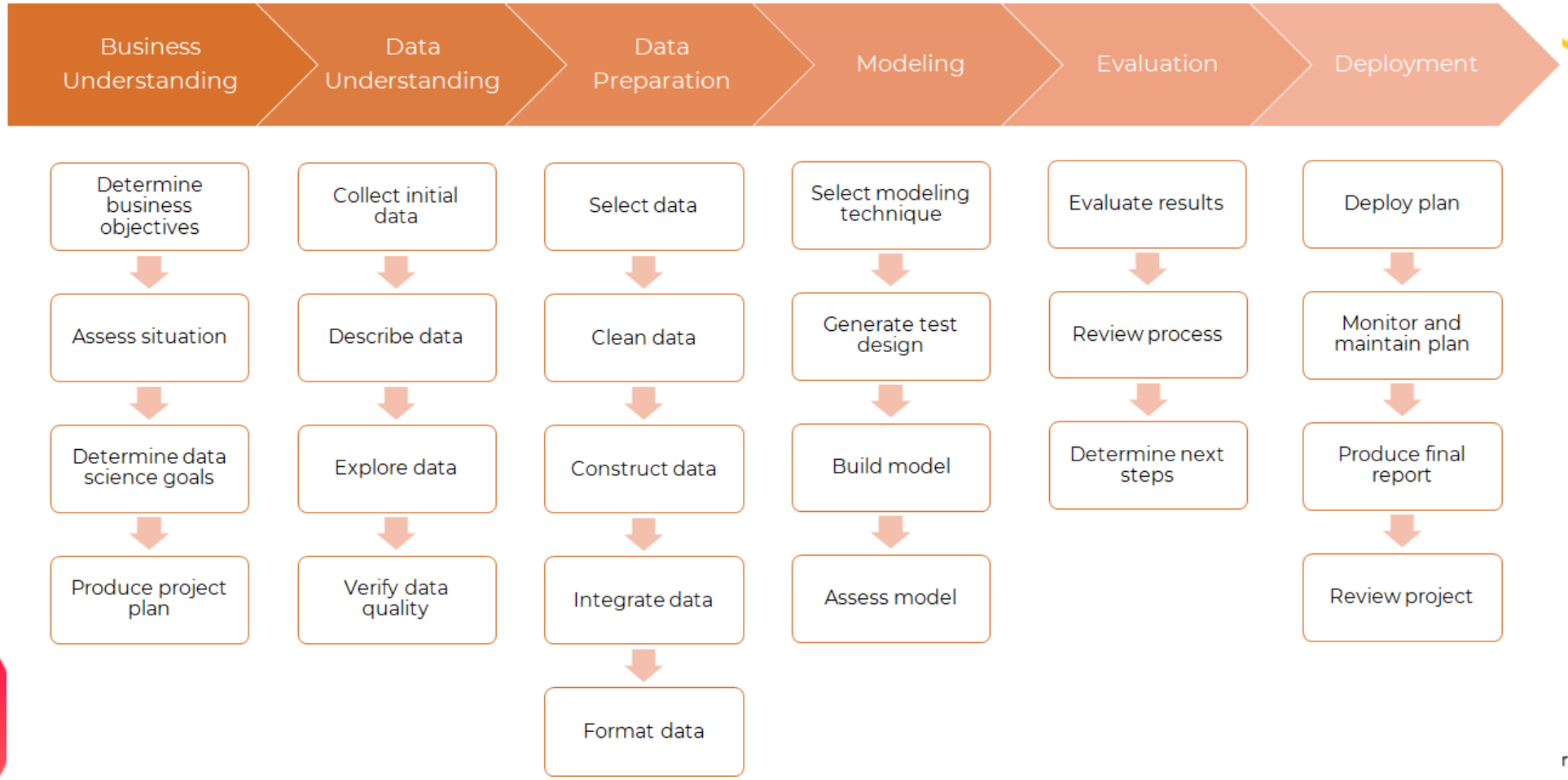
TELKOM DIGITAL TALENT INCUBATOR 2020

Deployment

- Use the insights gained from data mining to elicit change in the organization
- In general, the deployment phase of CRISP-DM includes two types of activities:
  - Planning and monitoring the deployment of results
  - Completing wrap-up tasks such as producing a final report and conducting a project review

TELKOM DIGITAL TALENT INCUBATOR 2020

Overall Process: CRISP-DM



rganized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

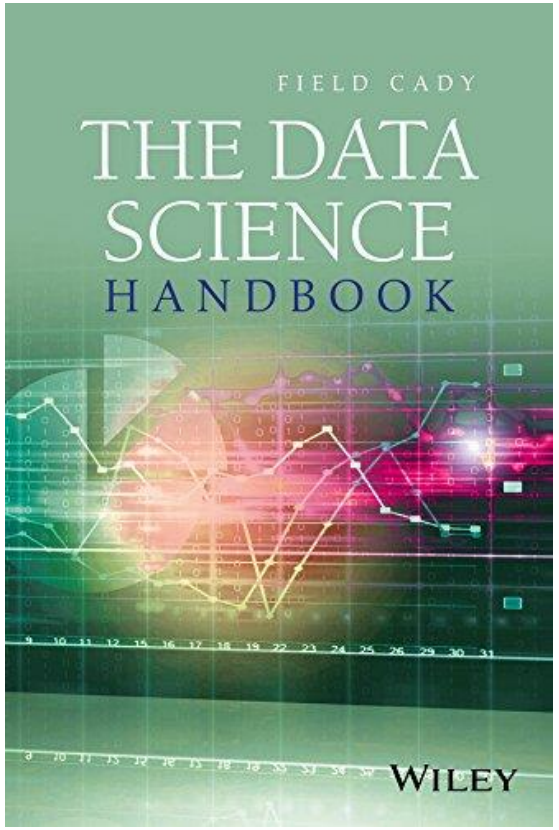
Module Summary

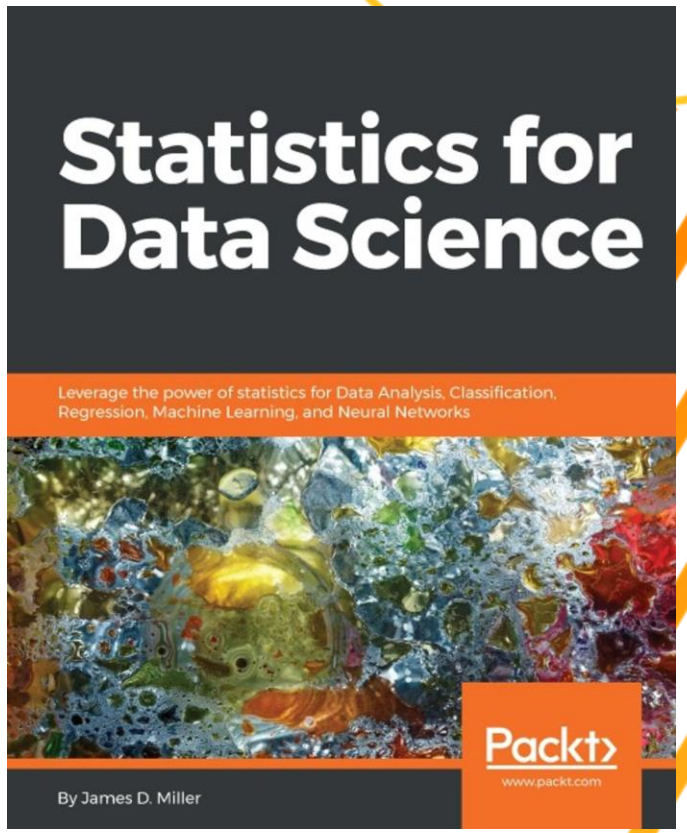
- Big data is larger, more complex data sets, especially from new data sources.
- Big data have 5 characteristics Volume, Variety, and Velocity Veracity, and Value.
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data.
- CRISP-DM or Cross Industry Standard Process for Data Mining is a framework that has been commonly used to perform data science processes.
- The CRISP-DM life cycle consists of six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

TELKOM DIGITAL TALENT INCUBATOR 2020

References/Additional Resources

- The Data Science Handbook - *1st Edition* by Field Cady.
- Statistics for Data Science ebook by James D. Miller



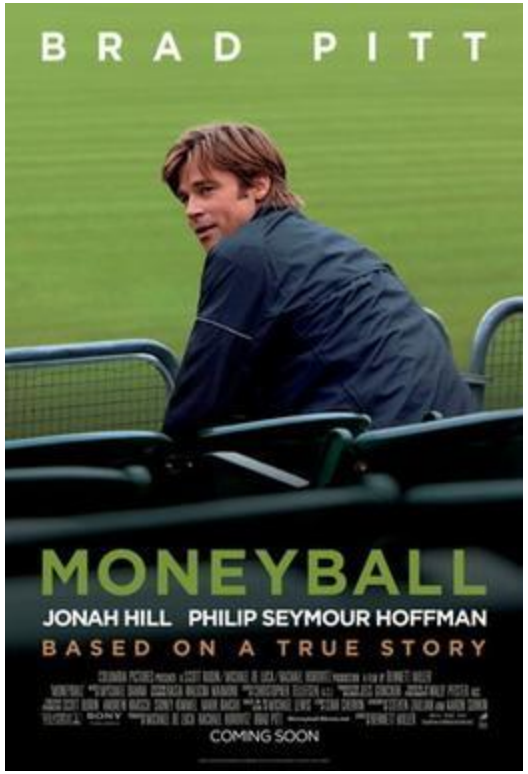


Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

Assignment

- Please Watch Moneyball(2011) Movie.
- Make Summary about the film, and explain how they use Data Science
  - Write it in Indonesian
  - Post it on Medium, Blog, or other online writing media.
  - Not allowed to take other people's writings.
  - Make it as informative as possible
  - Submit the link through: _____
  - **Deadline: 23rd September 2020, 23.59**



Organized by:
CAE