Modul 1

Introduction to Data Science

Data Science Program

Purwadhika
Startup and Coding School

Outline

- Data Science Challenges
- Data Science Workflow
- Data Science Roles
- Group Assignment

What is Data Science?

And, why it is important skill nowadays?

What is Data Science?

What is it?

- Is it a Role or Position?
- Is it a Process?
- Is it a problem / challenge ?

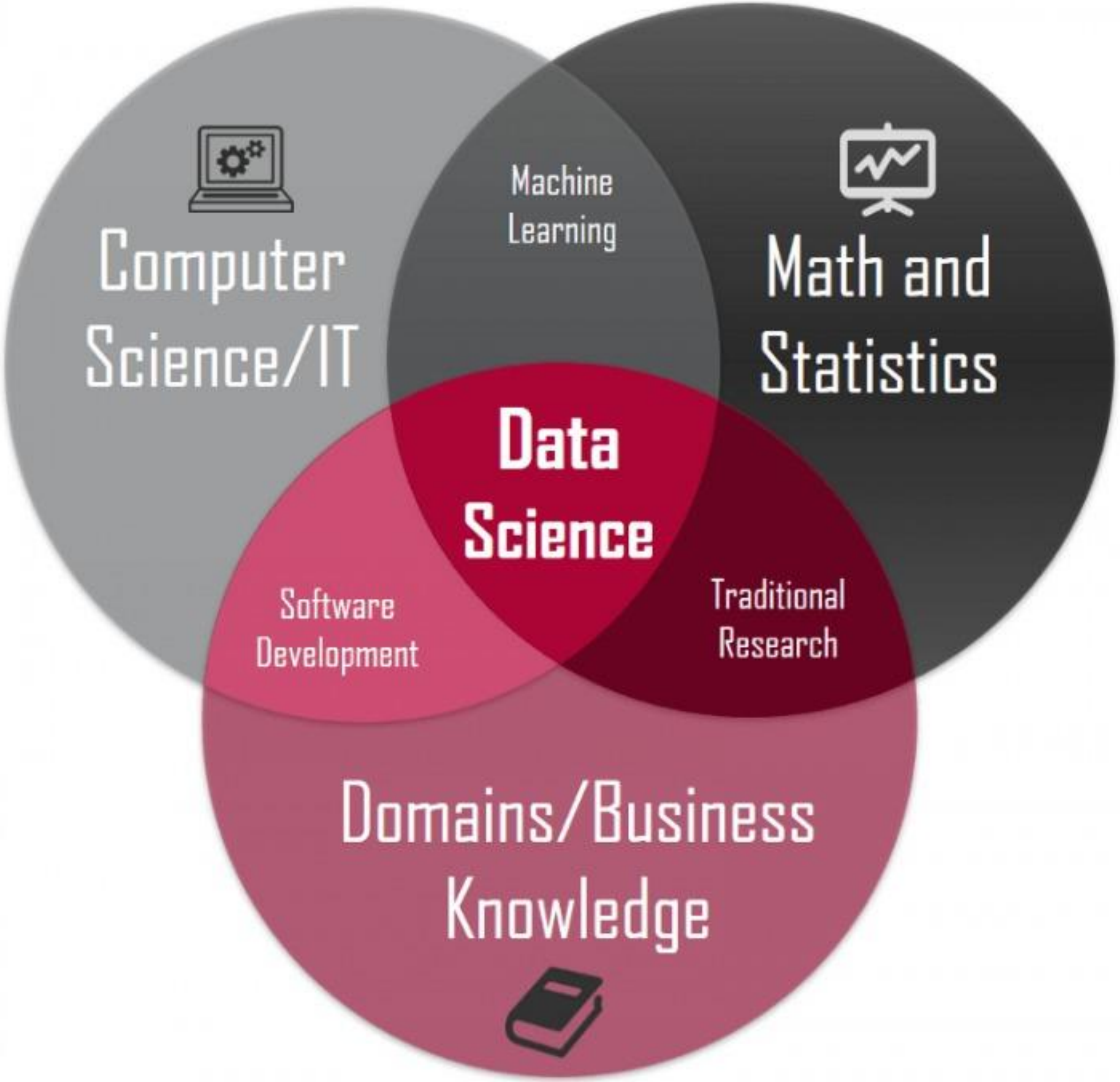
Correlations to this term :

- Big Data
- Data Driven
- Machine Learning
- AI
- Distributed computing

Data Science Definition

- **Data science** is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many **structural** and **unstructured data**.
- Data science is a "concept to unify statistics, data analysis, **machine learning** and their related methods" in order to **"understand and analyze actual phenomena" with data.**
- Data science is an interdisciplinary field focused on extracting knowledge from data sets, which are typically large (**big data**).

Data Science Skill



Machine Learning Definition

- **Arthur Samuel** (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.
- **Tom Mitchell** (1998) in his book “Machine Learning” provides a definition:

  A computer program is said to **learn** from experience ***E*** with respect to some tasks ***T*** and some performance measure ***P***, if its performance at tasks in ***T***, as measured by ***P***, improves with experience ***E***.

Machine Learning Analogy

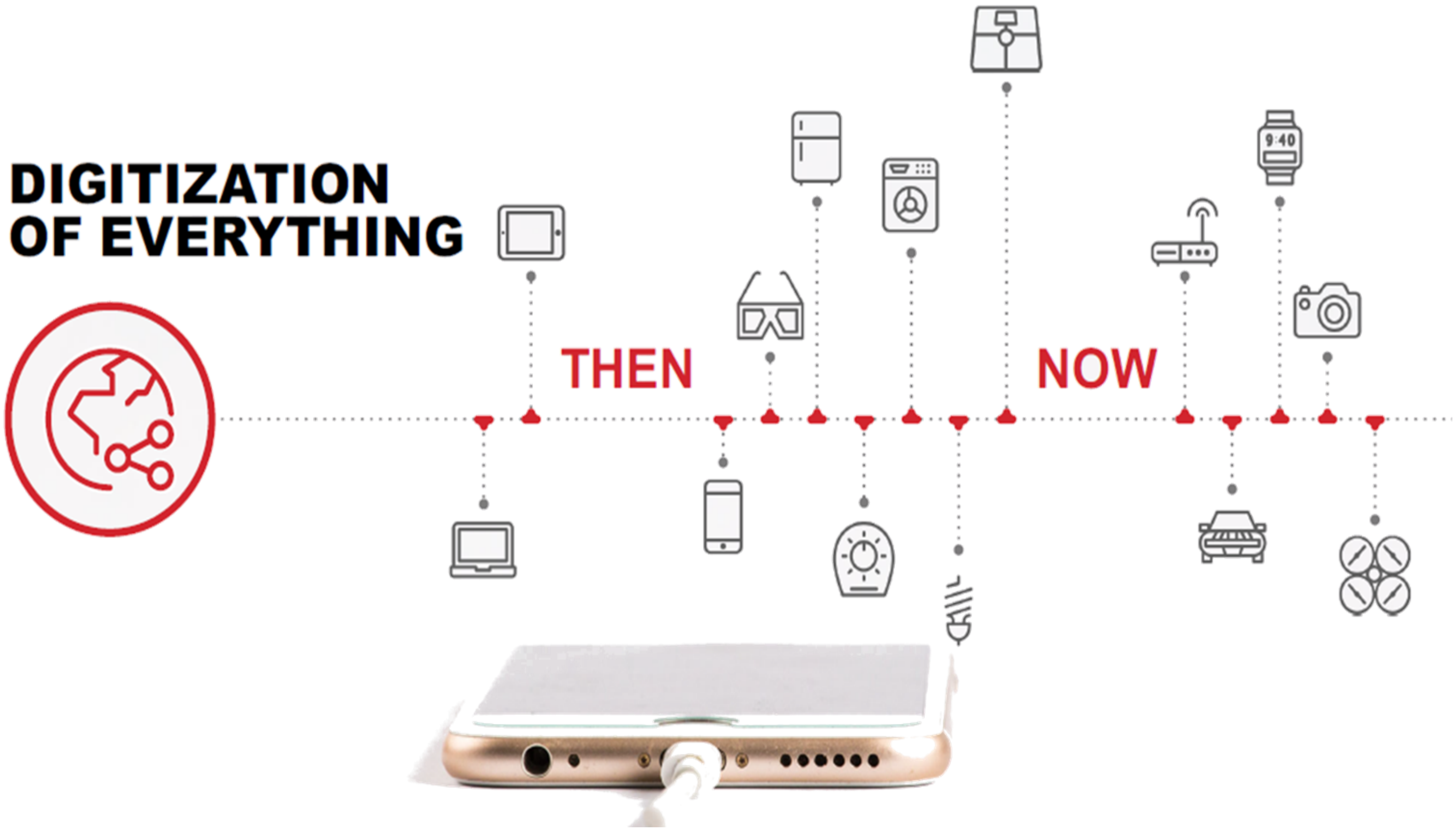
Membuat model Machine Learning (ML) bisa dianalogikan dengan mengajar seorang anak. Misalnya, kita ingin memberikan kemampuan untuk mengenali pohon atau bukan *(**classification problem**).*

Untuk membentuk kemampuan ini, anak ini diajak ke sebuah taman yang berisi beberapa jenis pohon. Di dalam taman, juga terdapat beberapa benda dan makhluk hidup yang bukan tergolong pohon. Anak ini diberi tahu mana yang termasuk pohon dan mana yang bukan pohon.

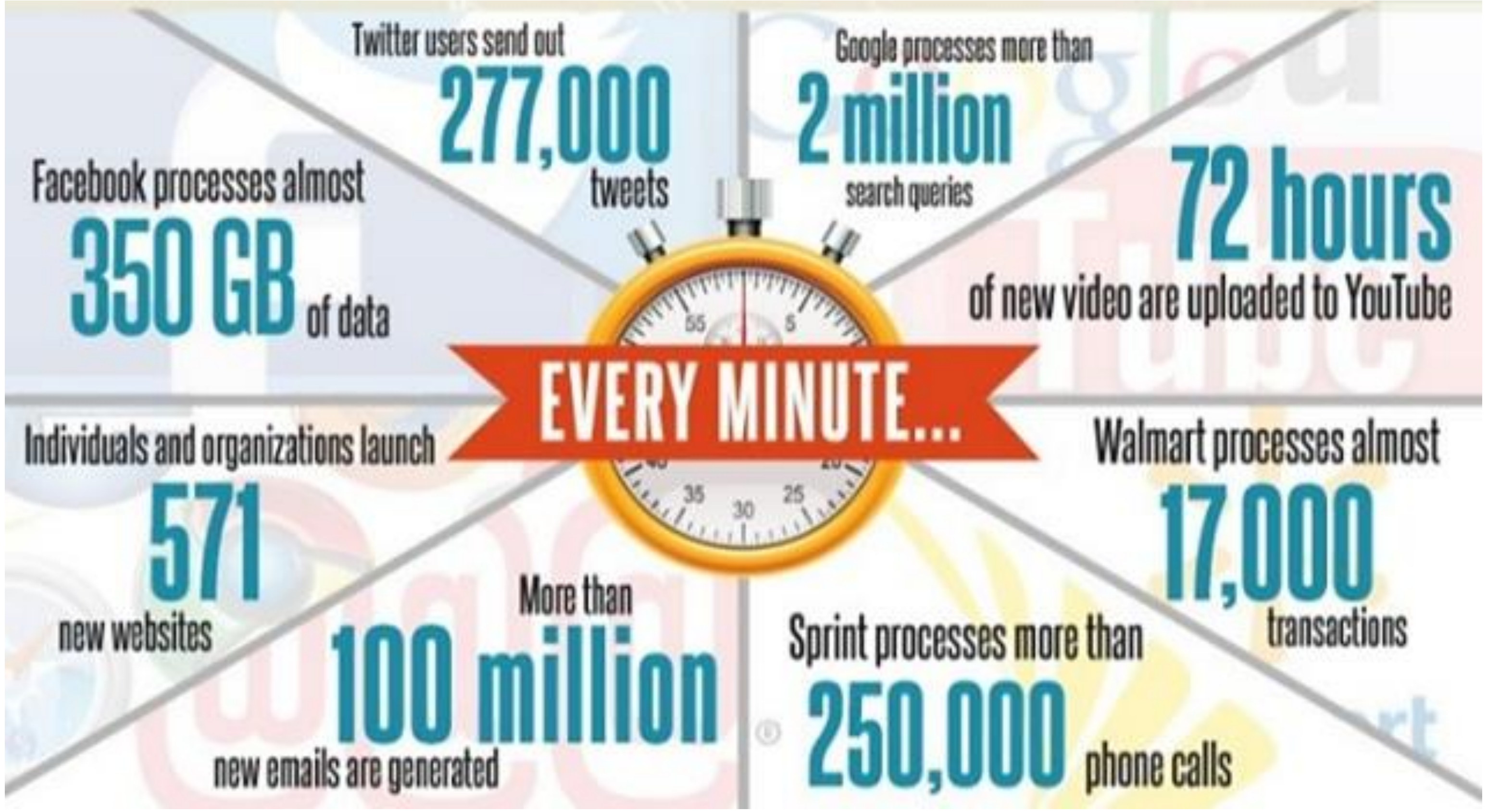
Untuk menguji kemampuan anak ini, kita ajak anak ini ke taman lain yang juga berisi pohon dan bukan pohon. Lalu kita uji seberapa tepat anak ini mampu mengenali pohon dan bukan pohon.

Tentu manusia dan mesin berbeda. Manusia memiliki multi-kemampuan. Sedangkan mesin memiliki kemampuan yang terbatas dan tergantung seberapa bagus data yang digunakan untuk belajar *(**data training**).*

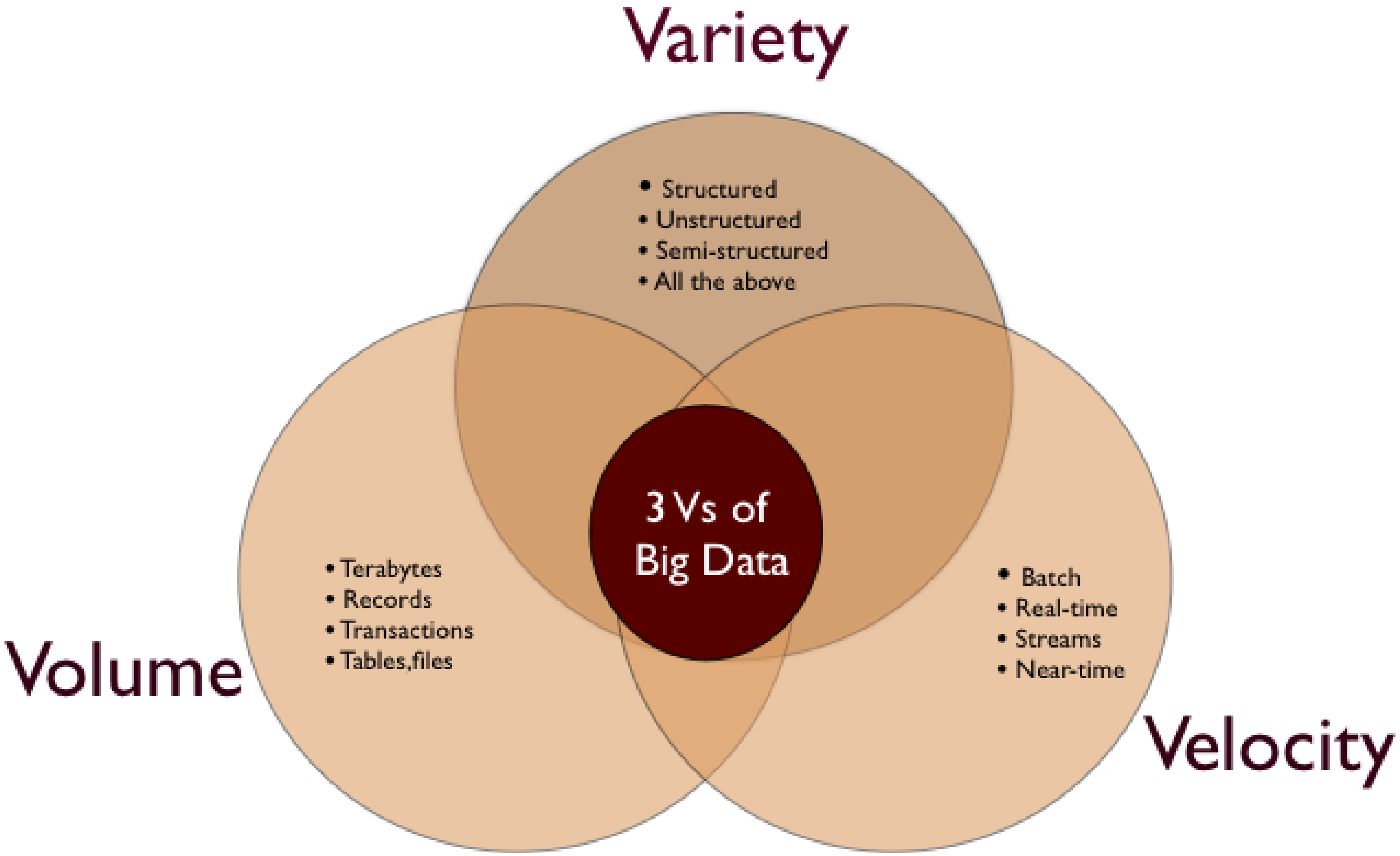
THE RISE OF INTERNET



THE RISE OF INTERNET


EVERY MINUTE...
Facebook processes almost
350 GB of data
Twitter users send out
277,000 tweets
Google processes more than
2 million search queries
72 hours of new video are uploaded to YouTube
Individuals and organizations launch
571 new websites
More than
100 million new emails are generated
Sprint processes more than
250,000 phone calls
Walmart processes almost
17,000 transactions

BIG DATA : 3V



adhika
Coding School

MULTI DISCIPLINARY

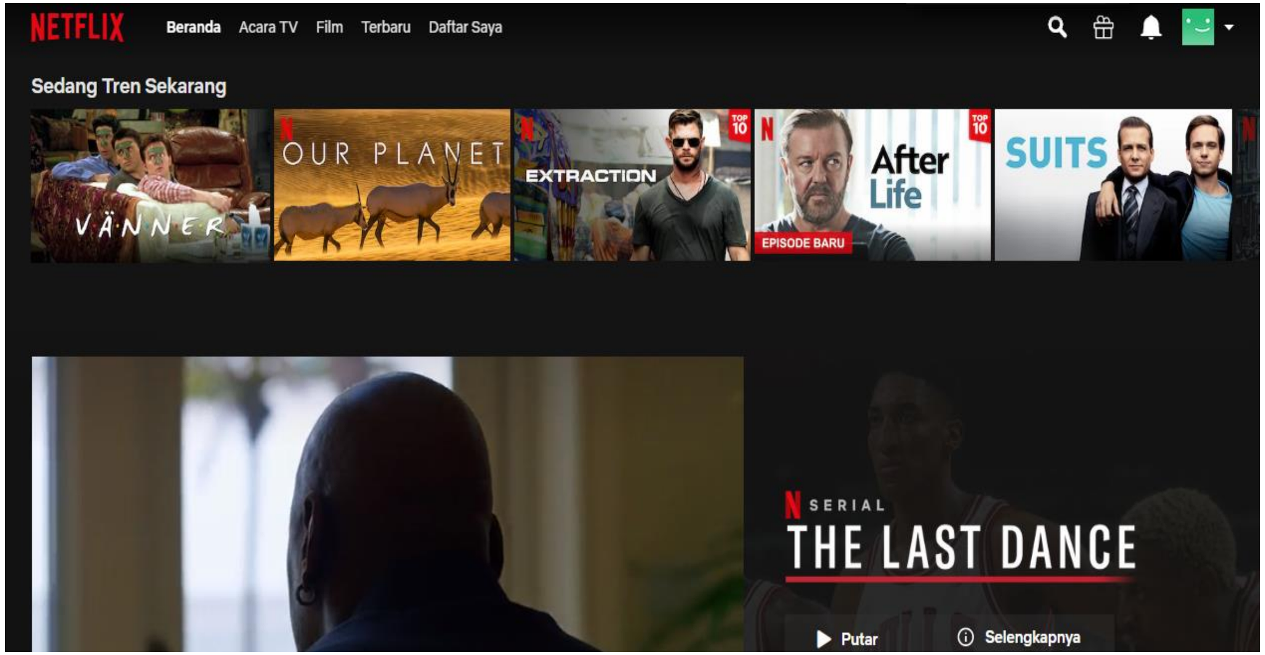
BANKING AND FINANCIAL SERVICES
PUBLIC SECTOR SERVICES
BANKING AND FINANCIAL SERVICES
SPORTS AND EDUCATION
OIL & GAS AND MANUFACTURING
HEALTHCARE AND MEDICAL SERVICES
TRANSPORTATION AND LOGISTICS
TELECOM AND ICT SERVICES
MEDIA AND ENTERTAINMENT
TRAVEL AND HOSPITALITY
CONSTRUCTION AND REAL ESTATE

We need to understand the PROBLEM

1. How the management think
2. How the customer think
3. How the market shifts

Purwadhika
Startup and Coding School

Use case: Netflix



Netflix mampu memahami perilaku konsumennya *(customer behavior)* dengan membaca pola konsumsi konsumen.

Sehingga, rekomendasi film atau series yang ditawarkan oleh Netflix sangat *customize* sesuai dengan preferensi setiap konsumen. Ini satu contoh penerapan *Machine Learning* jenis ***Recommender System***.

Another Use Case

- Credit Approval
- Customer Churn
- NLP
- House Price Prediction
- Fraud Detection
- Image Recognition
- Etc.

THE QUESTIONS

“Kami mau pasang iklan, tapi tidak tahu channel mana yang paling efektif”

“Ada beberapa produk kami yang tidak laku, walau review sangat bagus”

“Kredit nasabah kami banyak yang macet”

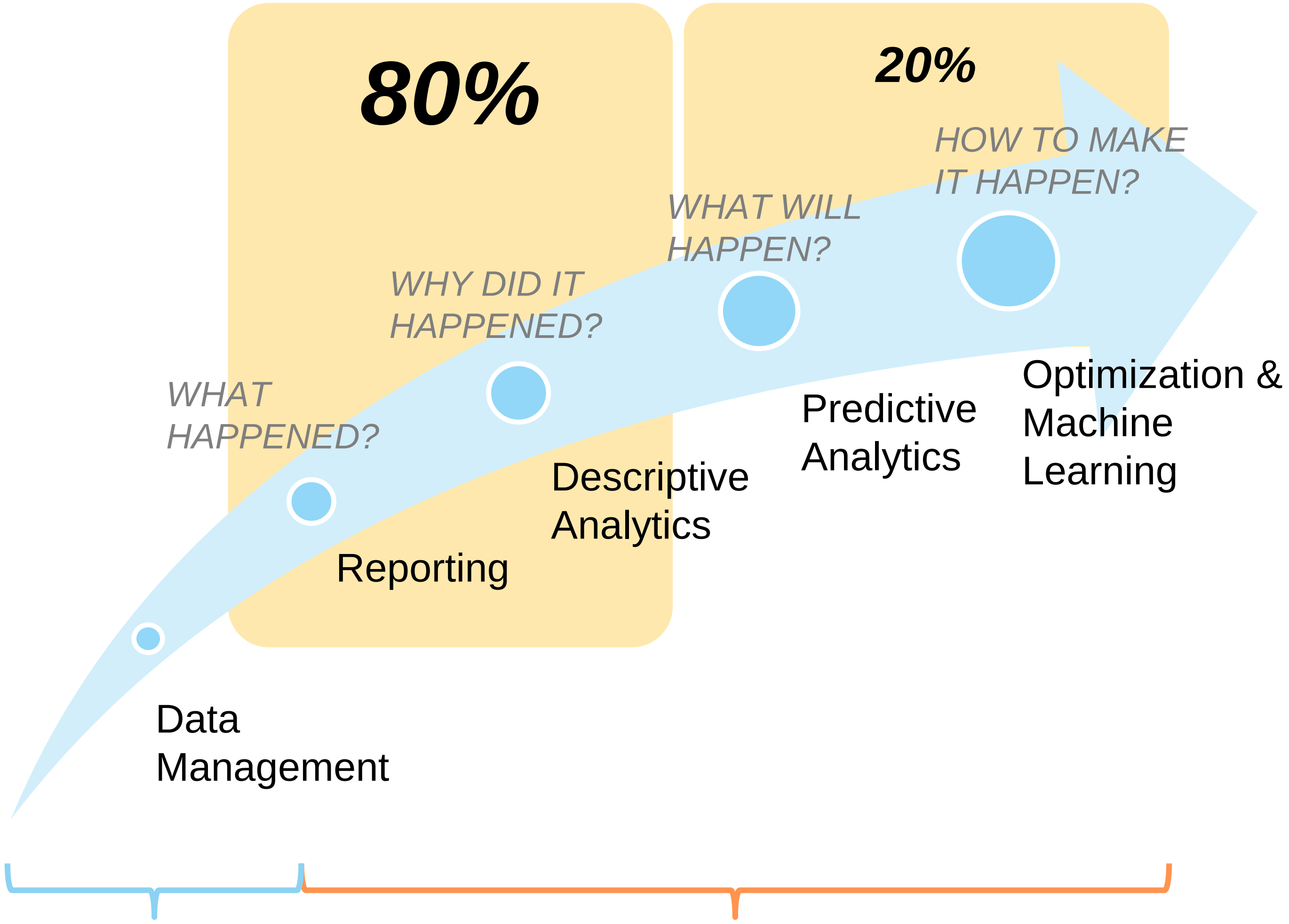
“Stock barang selalu habis/terlalu banyak”

“Kami tidak tahu seberapa efisien sales person kami”

Data Science Challenges

Purwadhika
Startup and Coding School

DATA SCIENCE CHALLENGES



80%

20%

WHAT HAPPENED?

WHY DID IT HAPPENED?

WHAT WILL HAPPEN?

HOW TO MAKE IT HAPPEN?

Data Management

Reporting

Descriptive Analytics

Predictive Analytics

Optimization & Machine Learning

Data gathering, restructuring

Extracting information

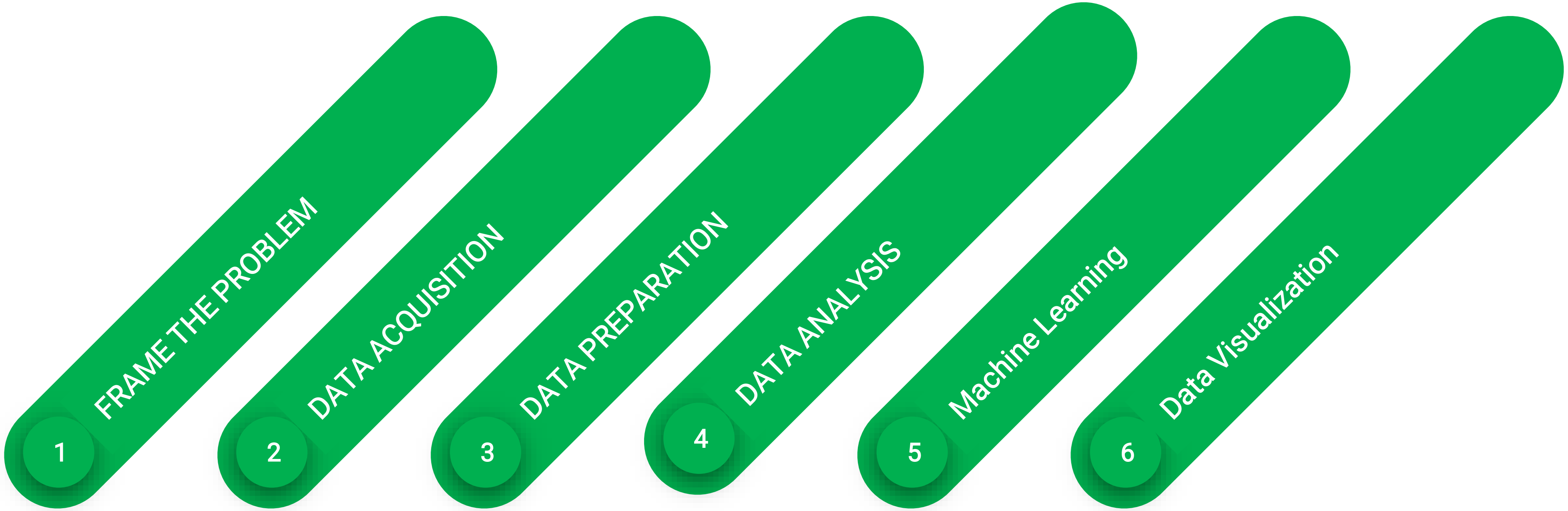
Purwadhika
Startup and Coding School

DATA SCIENCE WORKFLOW

Step to step solving the problems as Data Scientist

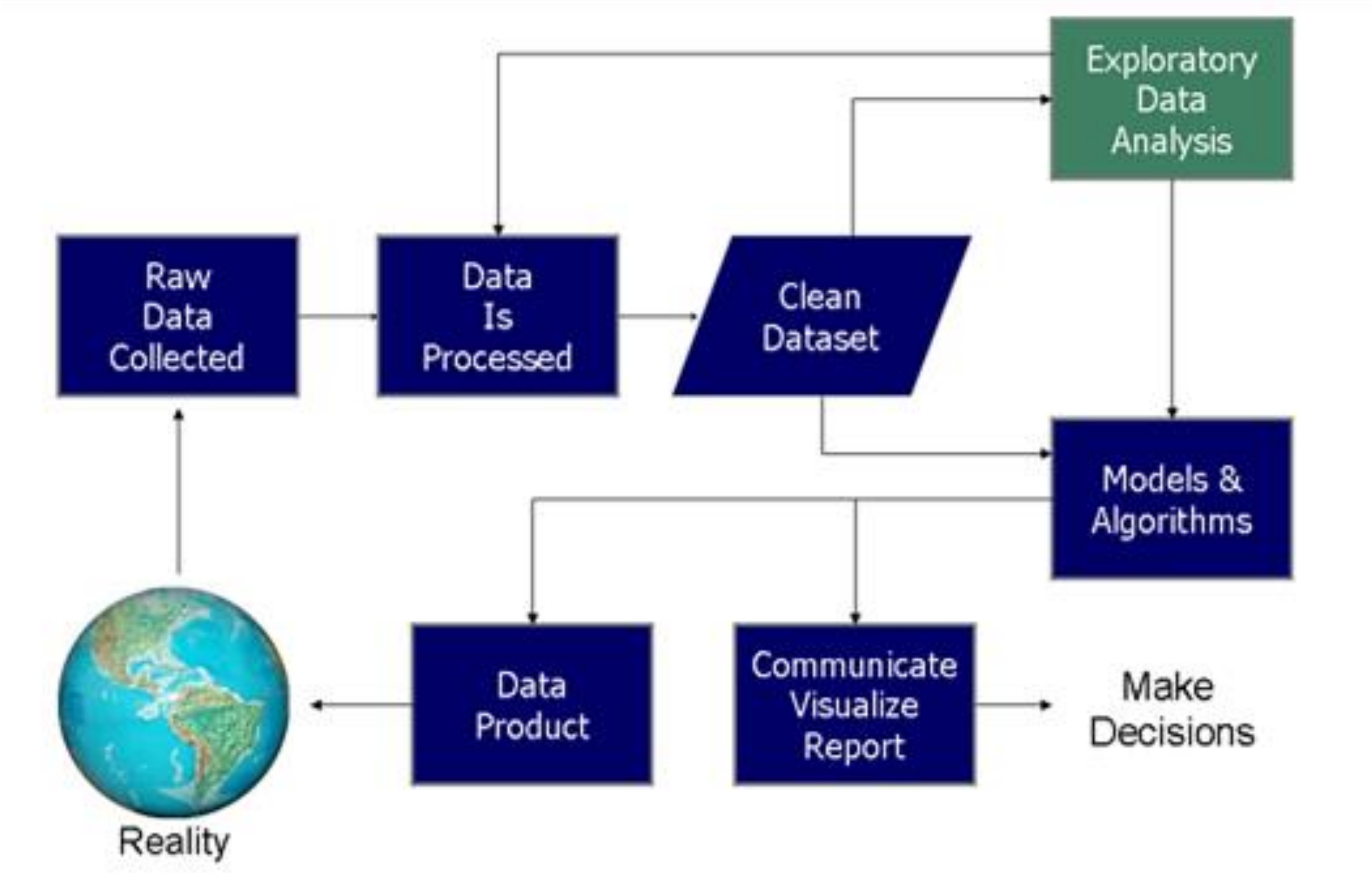
Purwadhika
Startup and Coding School

DATA SCIENCE WORKFLOW



1. FRAME THE PROBLEM
2. DATA ACQUISITION
3. DATA PREPARATION
4. DATA ANALYSIS
5. Machine Learning
6. Data Visualization

DATA SCIENCE WORKFLOW



DATA SCIENCE WORKFLOW

Ask Questions

- Who are the customers?
- Why are they buying our product?
- How do we predict if a customer is going to buy our product?
- What is different from segments who are performing well and those that are performing below expectations?
- How much money will we lose if we don’t actively sell the product to these groups?

DATA SCIENCE WORKFLOW

What needs to be considered:

- Data Sources
- Data Location
- Data Format
- Data Types
- Acquisition Methods
- Data Privacy

DATA SCIENCE WORKFLOW

Data Sources:

- Users Profile
- Users Activity/transaction
- Enterprise resources
- World trends/activity

Data Location:

- Inter Department
- Across Department
- External Data
- Public Data

DATA SCIENCE WORKFLOW

Data Format:

- Hard copy
- Digital documents
- Database
- Streams

Data Types:

- Numerical
- Text
- Image
- Audio
- Video

DATA SCIENCE WORKFLOW

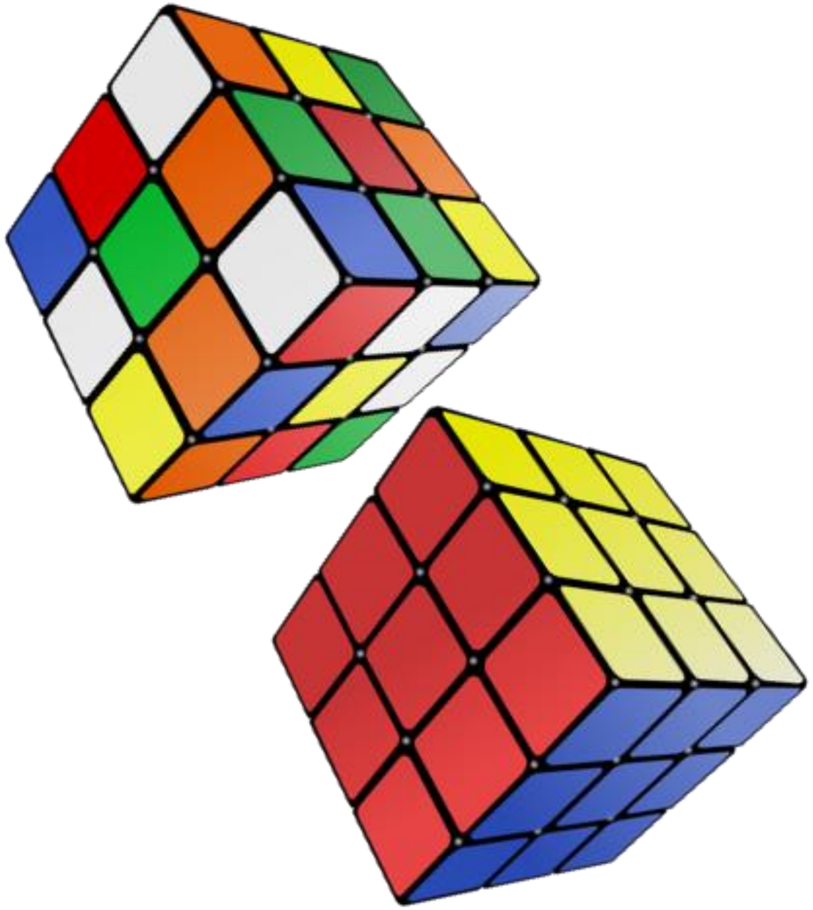
Data Access:

- Data Warehousing
- REST API
- Web Scraping

Data Privacy:

- User Consent: User needs to give consent for any usage purposes
- Data Privacy Law:
  - EU General Data Protection Regulator
  - RUU Perlindungan Data Pribadi

DATA SCIENCE WORKFLOW



Structured Data

Vs

Unstructured Data

Purwadhika
Startup and Coding School

DATA SCIENCE WORKFLOW

Structured Data

Structured data is most often categorized as quantitative data, and it's the type of data most of us are used to working with. Think of data that fits neatly within fixed fields and columns in relational databases and spreadsheets.

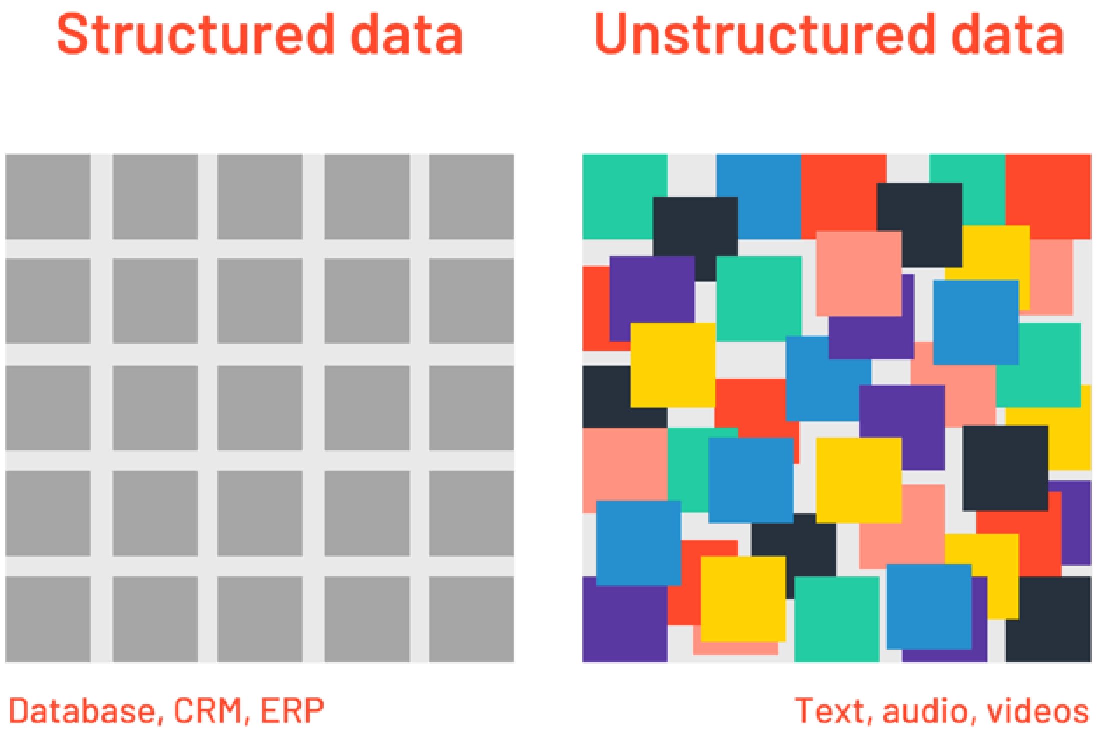
Examples of structured data include names, dates, addresses, credit card numbers, stock information, geo-location, and more.

Unstructured Data

Unstructured data is most often categorized as qualitative data, and it cannot be processed and analyzed using conventional tools and methods.

Examples of unstructured data include text, video, audio, mobile activity, social media activity, satellite imagery, surveillance imagery – the list goes on and on.

DATA SCIENCE WORKFLOW



More than **80 percent** of all data generated today is considered **unstructured**, and this number will continue to rise with the prominence of the internet of things.

Unstructured data is difficult to deconstruct. Instead, non-relational, or **NoSQL** databases, are best fit for managing **unstructured data**.

The programming language used for managing **structured data** is called structured query language, also known as **SQL**.

DATA SCIENCE WORKFLOW

DATA PREPARATION & ANALYSIS

Data Preparation

- Data cleansing
  - Format normalization
  - Typing inconsistency
- Handling NULL values
- Handling outliers
- Feature selection/ engineering

Data Analysis

- Always aim to answer the problem definition
- Identify:
  - Variations
  - Correlations
  - Trends
  - Outliers

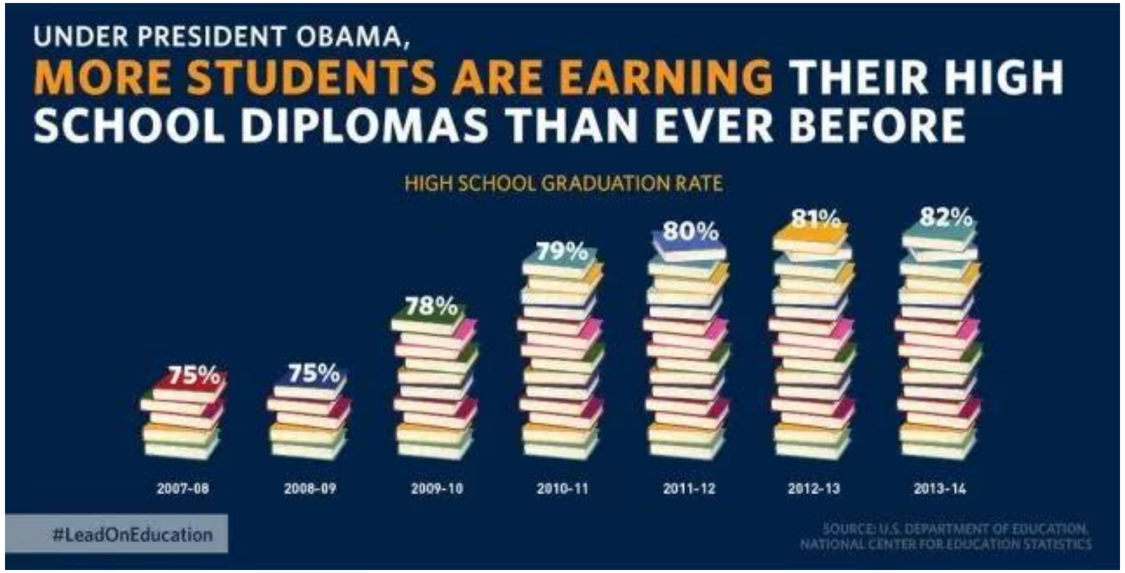
Purwadhika
Startup and Coding School

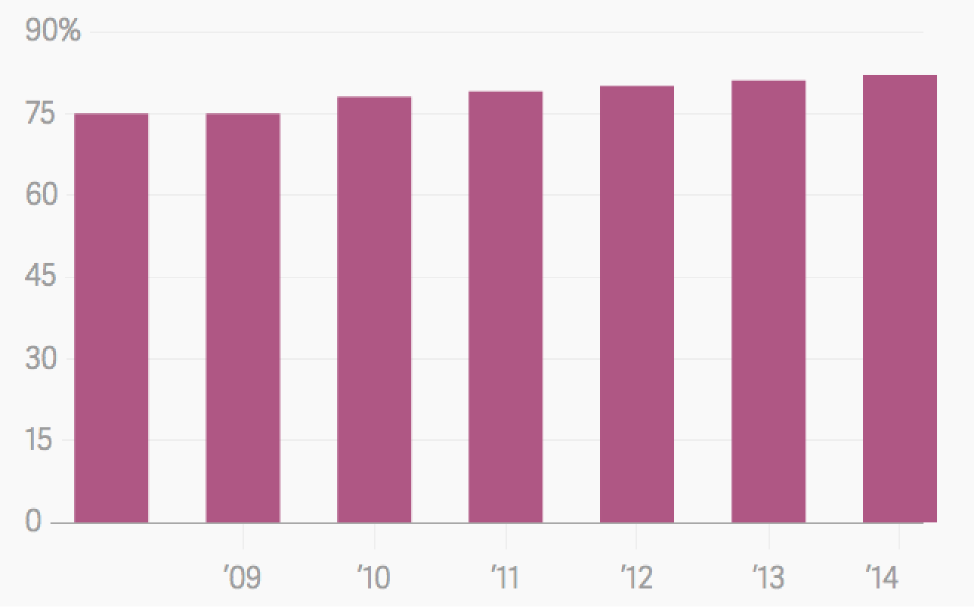
DATA SCIENCE WORKFLOW

DATA PREPARATION & ANALYSIS

Data Visualization

- Know the audience
- Visualization is all about perception

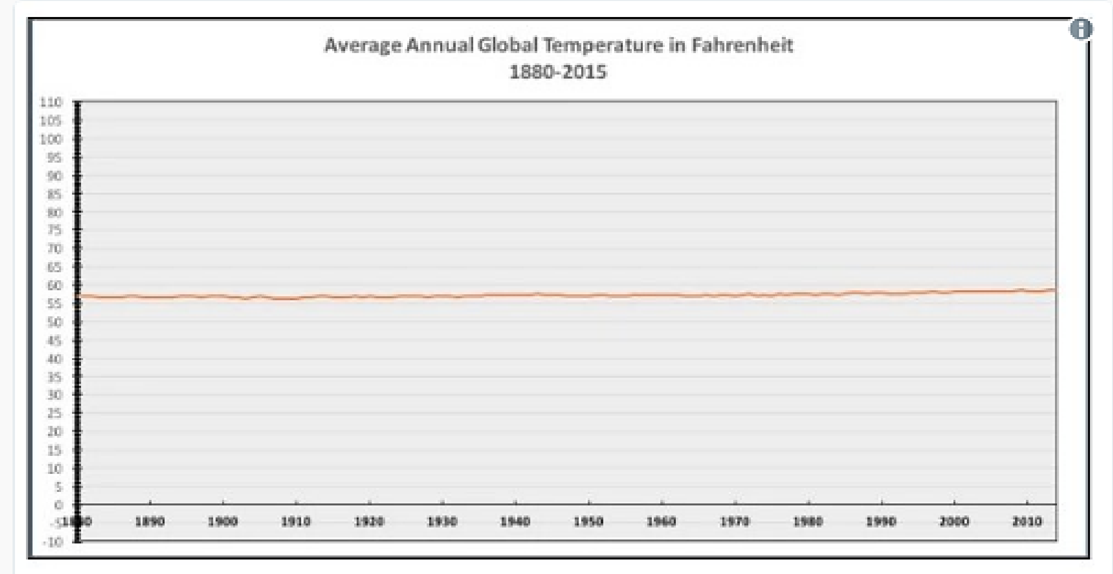


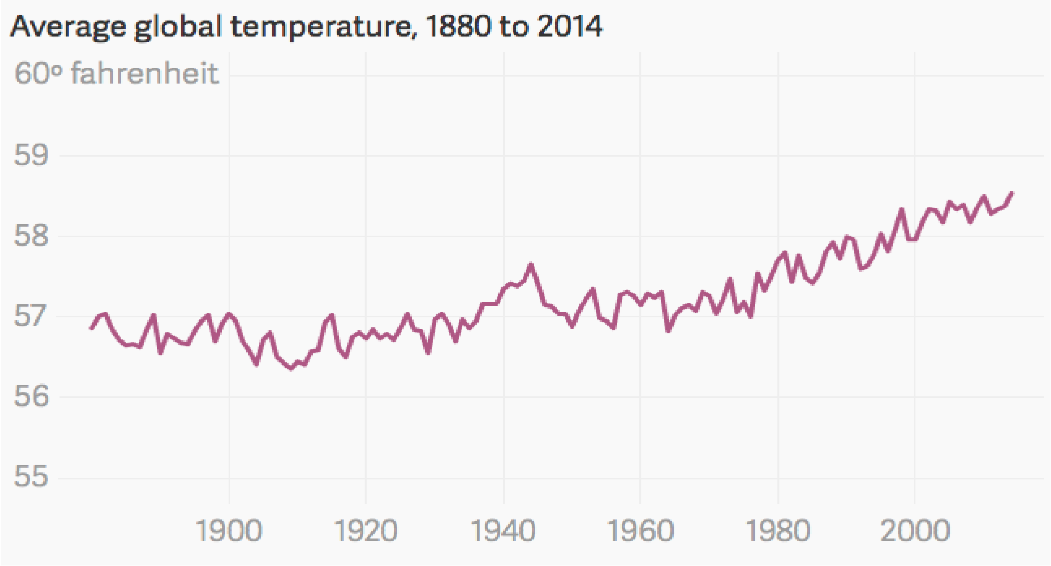


DATA SCIENCE WORKFLOW

DATA PREPARATION & ANALYSIS

Data Visualization





DATA SCIENCE ROLES

Data Scientist, Data Engineer, Business Analyst, Domain Expert, etc.

Data Scientist

Activities

- Data cleansing and Preparation
- Evaluating statistical models
- Build ML Model

Tools

- R
- Python
- Matlab
- Stata
- SQL
- Spark

Skills and Talents

- Domain Knowledge/Business Understanding
- Statistical theories and methodologies
- Database systems
- Programming skills

Data Engineer

Activities

- Data Integration
- Product Development (Dashboard, API)
- Scalability and Automation

Tools

- Database systems: SQL, NoSQL
- Python, Node
- Google Cloud Platform, Amazon AWS
- Distributed System

Skills and Talents

- Programming skills
- Database system and modelling
- IT Infrastructure and Cloud environment

Business Analyst

Activities

- Framing the problem
- Data Exploration
- Presenting Analysis insights

Skills and Talents

- Business and Domain knowledge
- Communication
- Database query language

Tools

- Dashboard
- Visualization tools :Tableau, QlikView
- Open Refine
- Powerpoint and Excel

Domain Expert

Activities

- Framing the problem
- Provides Consultation to the real world problems

Tools

- (depends on the field)

Skills and Talents

- Business and Domain knowledge
- Communication

Other roles

- Database Admin: Query/Prepare data to be processed/analyze
- Data Architect: Design information architect
- Statistician
- Developer

“Keep it **lean,**
Grow as you go”

Purwadhika
Startup and Coding School

ASSIGNMENT

Challenge :

Sebuah perusahaan *financial technology* (fintech) ingin mengetahui karakteristik nasabah yang berpotensi kreditnya macet *(Credit Risk).* Tujuannya agar tim marketing memahami karakteristik target calon nasabah yang kreditnya tidak berpotensi macet.

Direksi meminta bantuan kepada tim Data Science untuk memahami karakteristik nasabah yang macet dan lancar.

ASSIGNMENT

PROBLEM IDENTIFICATION

Define the problem, identify the questions:

- What is the problem ?
- Who is having the problem ?
- When is it happening ?
- Where is it happening ?
- What are the expected output?
- What have happened in the past?

ASSIGNMENT

Plan the data driven Process!

- **Data Acquisition :**
  What data do I need, and how to access them?

- **Data Preparation :**
  Define the ideal data format, and ways to prepare them

ASSIGNMENT

Plan the data driven Process!

- **Data Acquisition :**
  What data do I need, and how to access them?

- **Data Preparation :**
  Define the ideal data format, and ways to prepare them

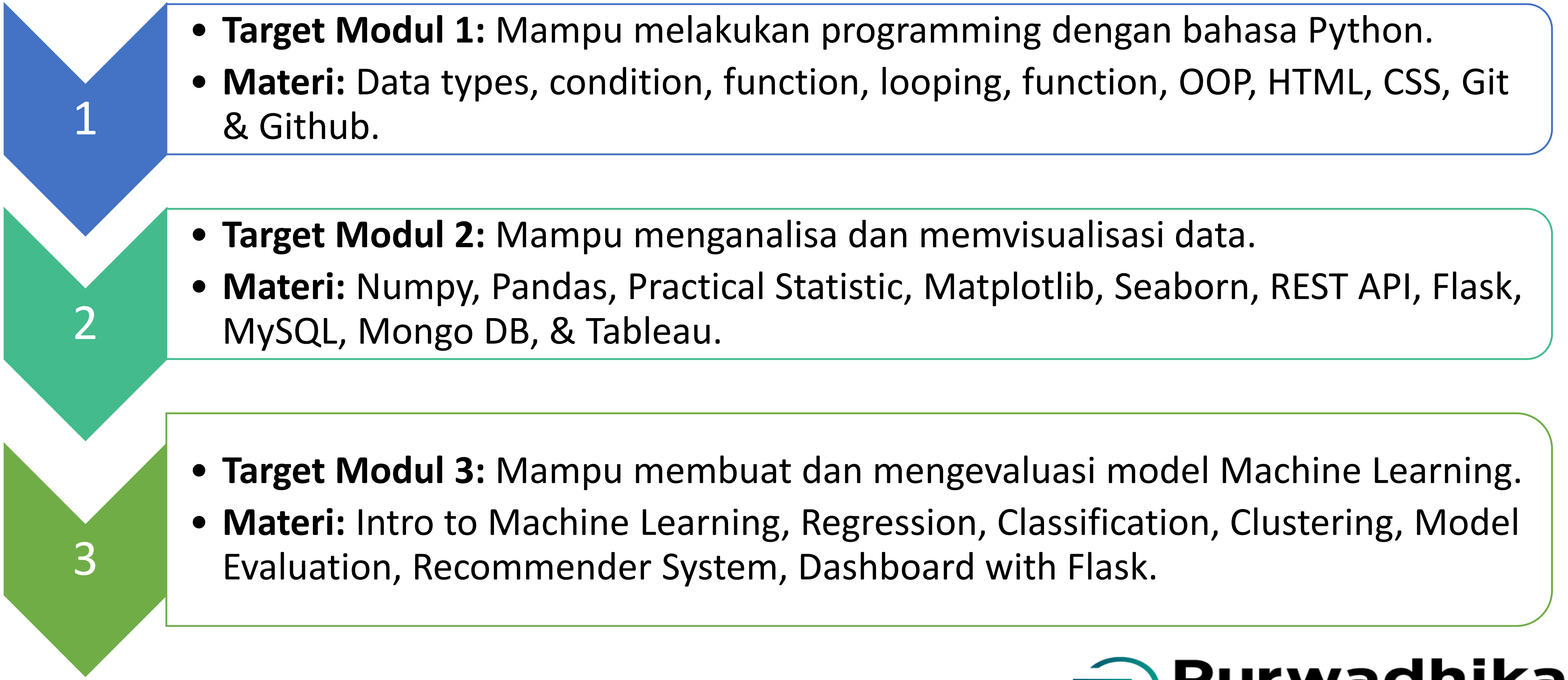
- **Data Analysis :**
  What insigths do you need, and how to analyse them?

- **Data Visualization:**
  How and to whom do you share your insights

DATA SCIENCE MODULE

Purwadhika
Startup and Coding School

Data Science Module



1

- **Target Modul 1:** Mampu melakukan programming dengan bahasa Python.
- **Materi:** Data types, condition, function, looping, function, OOP, HTML, CSS, Git & Github.

2

- **Target Modul 2:** Mampu menganalisa dan memvisualisasi data.
- **Materi:** Numpy, Pandas, Practical Statistic, Matplotlib, Seaborn, REST API, Flask, MySQL, Mongo DB, & Tableau.

3

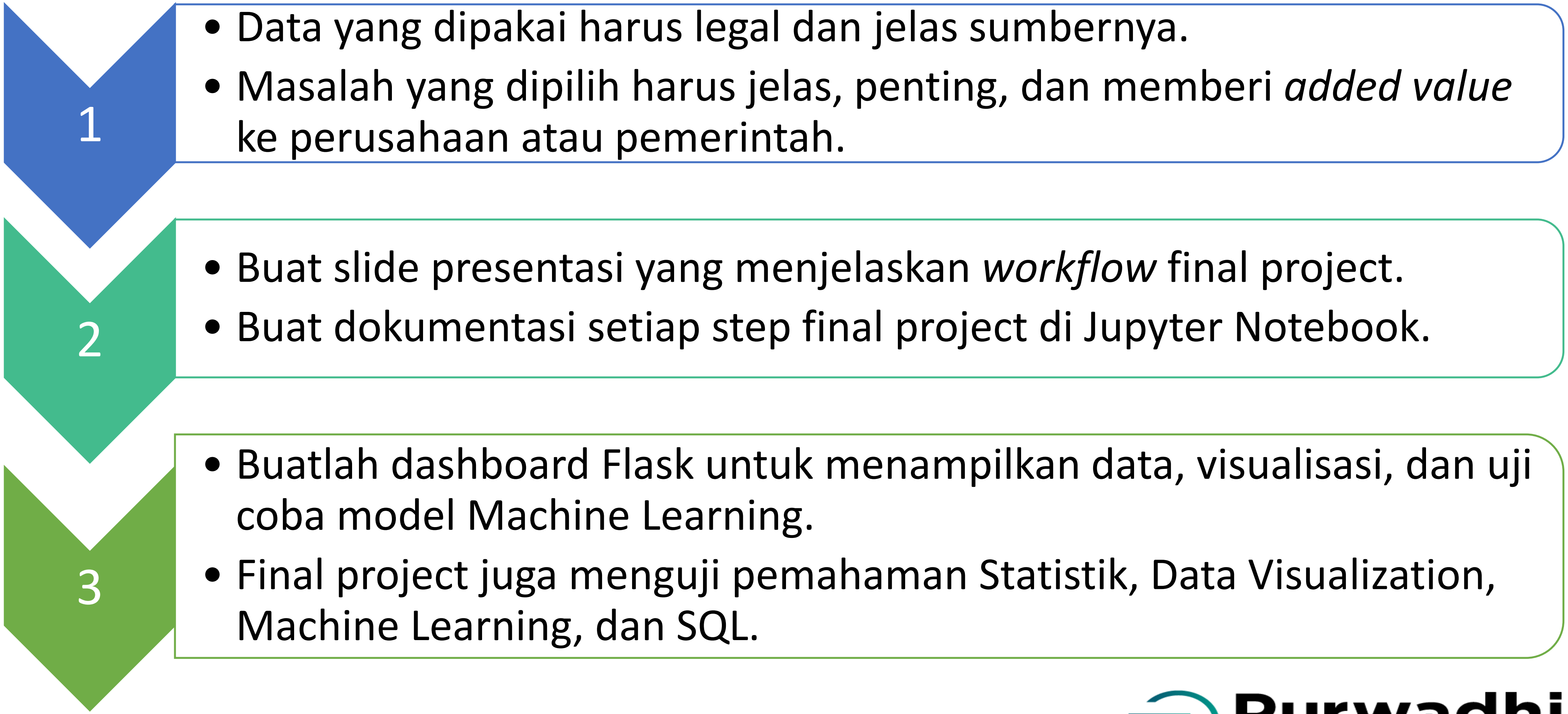
- **Target Modul 3:** Mampu membuat dan mengevaluasi model Machine Learning.
- **Materi:** Intro to Machine Learning, Regression, Classification, Clustering, Model Evaluation, Recommender System, Dashboard with Flask.

Purwadhika
Startup and Coding School

FINAL PROJECT GUIDANCE

Purwadhika
Startup and Coding School

Final Project Guidance



1

- Data yang dipakai harus legal dan jelas sumbernya.
- Masalah yang dipilih harus jelas, penting, dan memberi *added value* ke perusahaan atau pemerintah.

2

- Buat slide presentasi yang menjelaskan *workflow* final project.
- Buat dokumentasi setiap step final project di Jupyter Notebook.

3

- Buatlah dashboard Flask untuk menampilkan data, visualisasi, dan uji coba model Machine Learning.
- Final project juga menguji pemahaman Statistik, Data Visualization, Machine Learning, dan SQL.