Modul 3

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

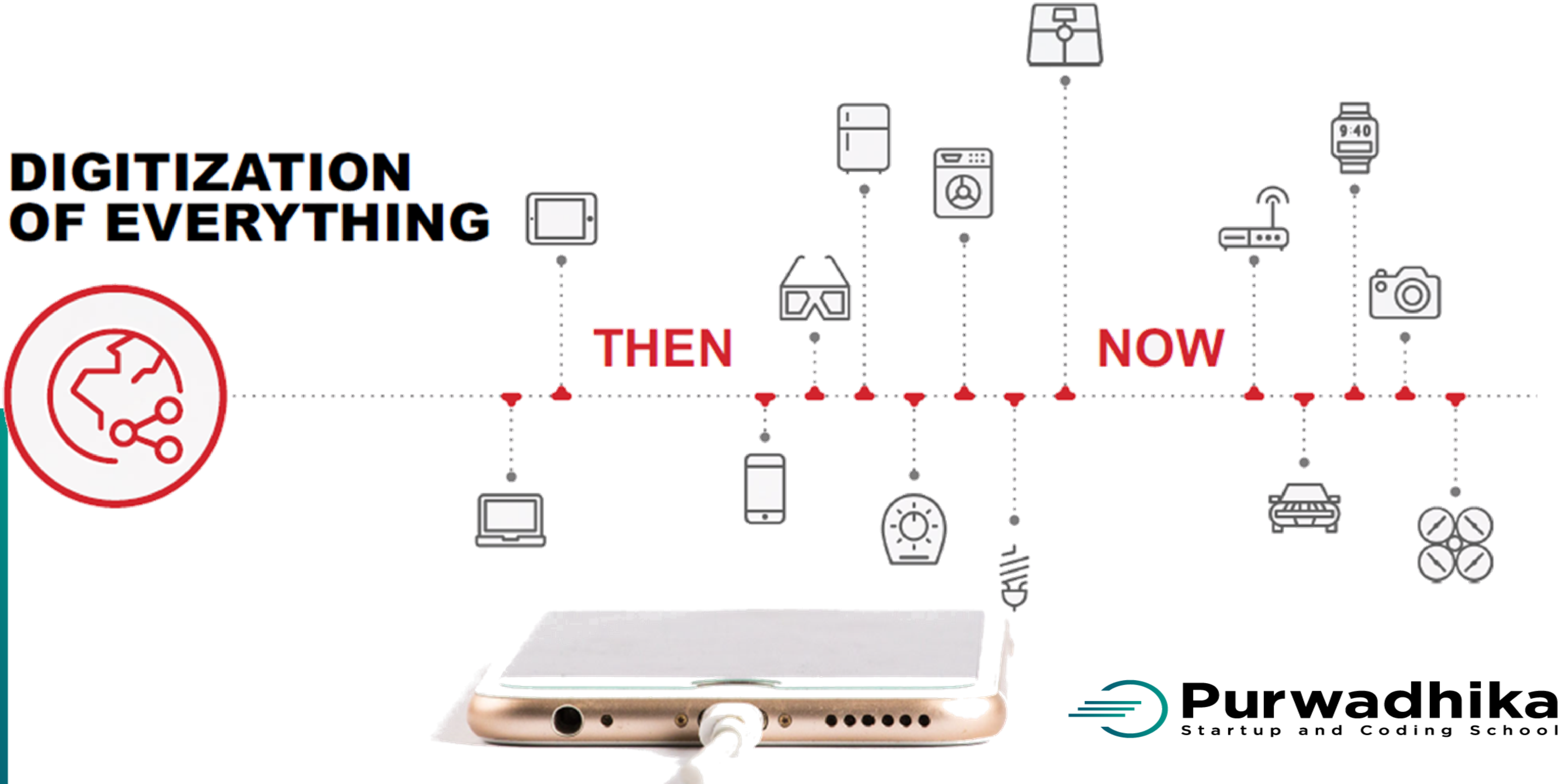
What is it?

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

Correlations to this term :

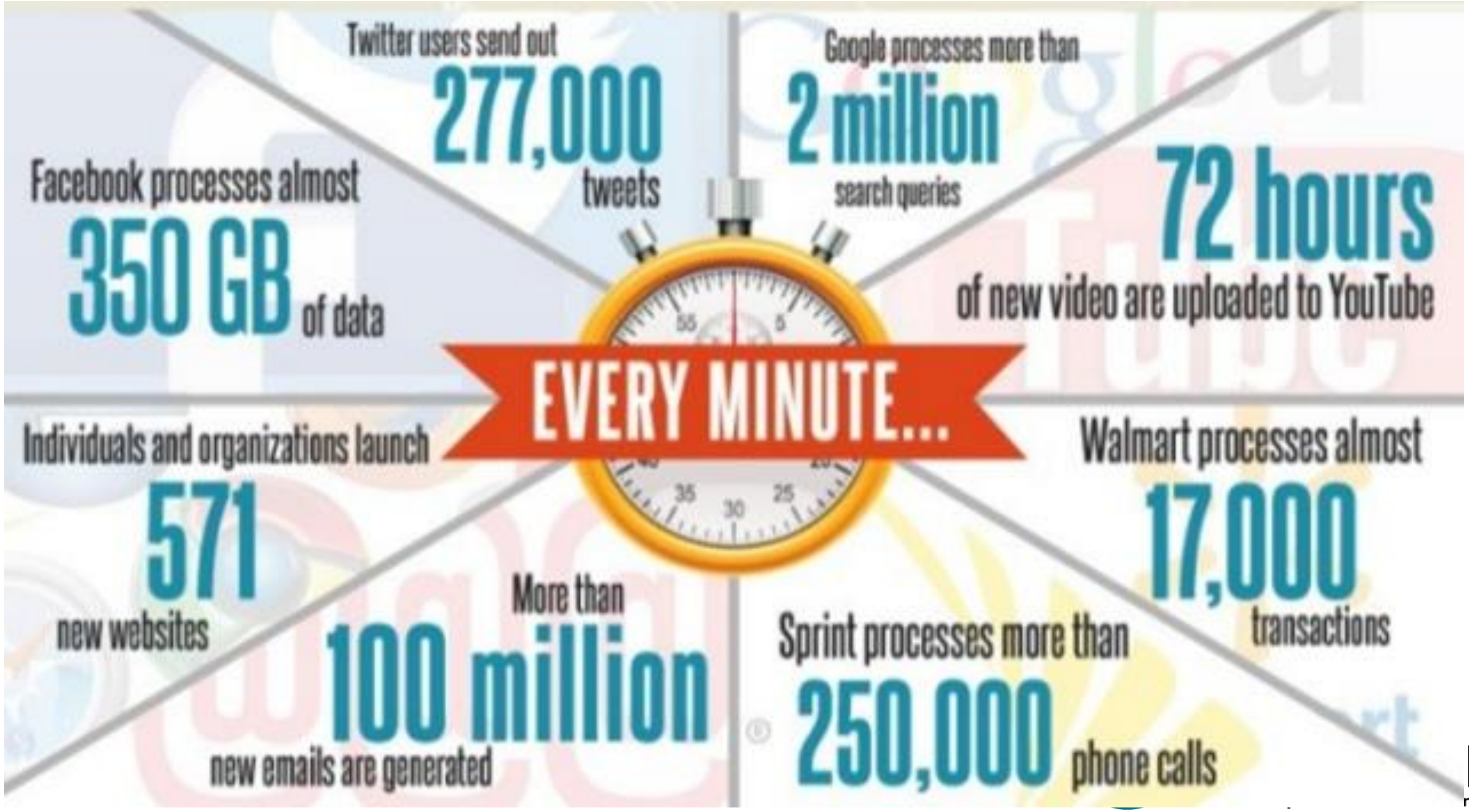
- Big Data
- Data Driven
- Machine Learning
- AI
- Distributed computing

THE RISE OF INTERNET



Purwadhika
Startup and Coding School

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

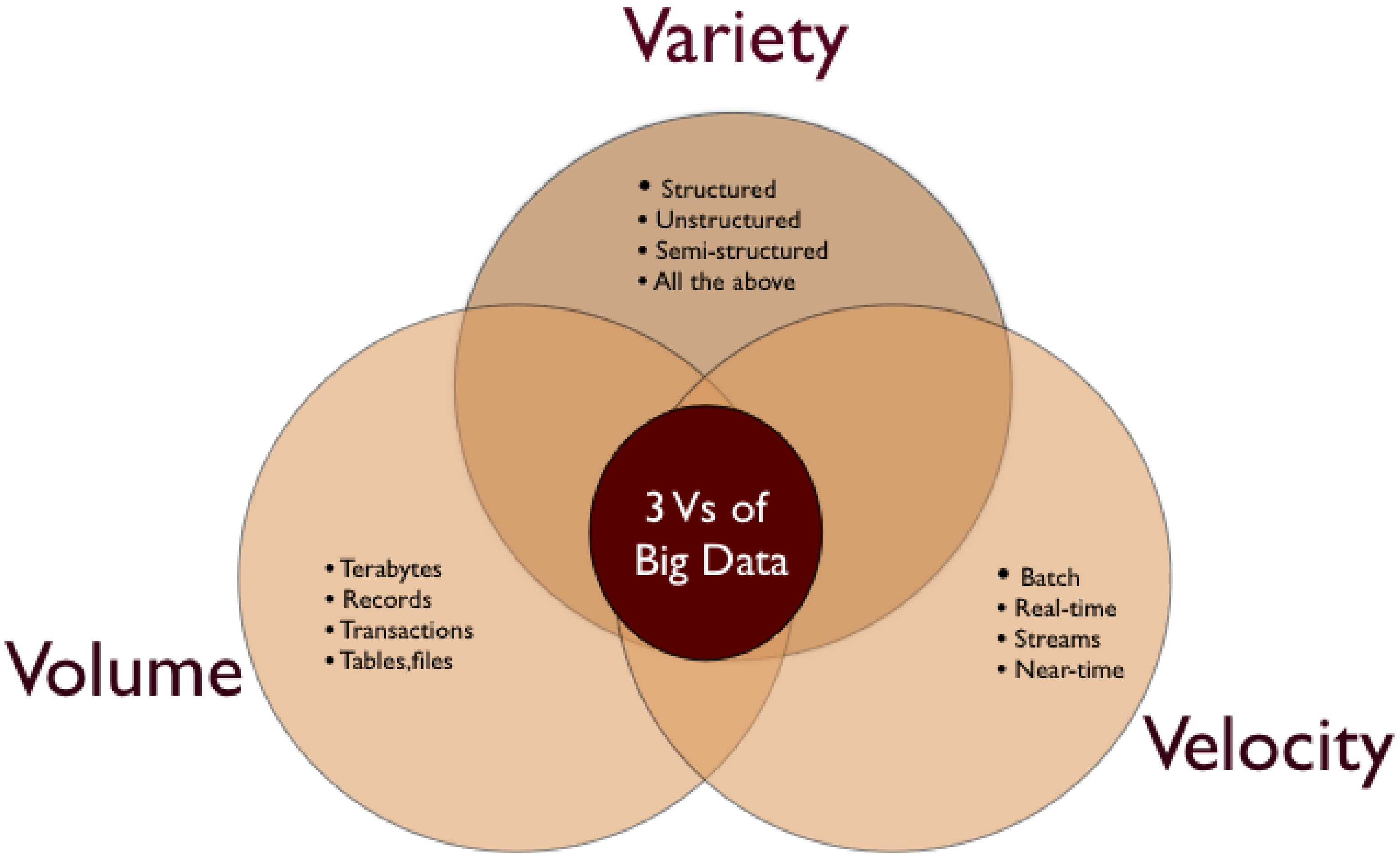
Individuals and organizations launch 571 new websites

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


Variety
• Structured
• Unstructured
• Semi-structured
• All the above
3 Vs of Big Data
Volume
• Terabytes
• Records
• Transactions
• Tables,files
Velocity
• Batch
• Real-time
• Streams
• Near-time

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

THE QUESTIONS

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

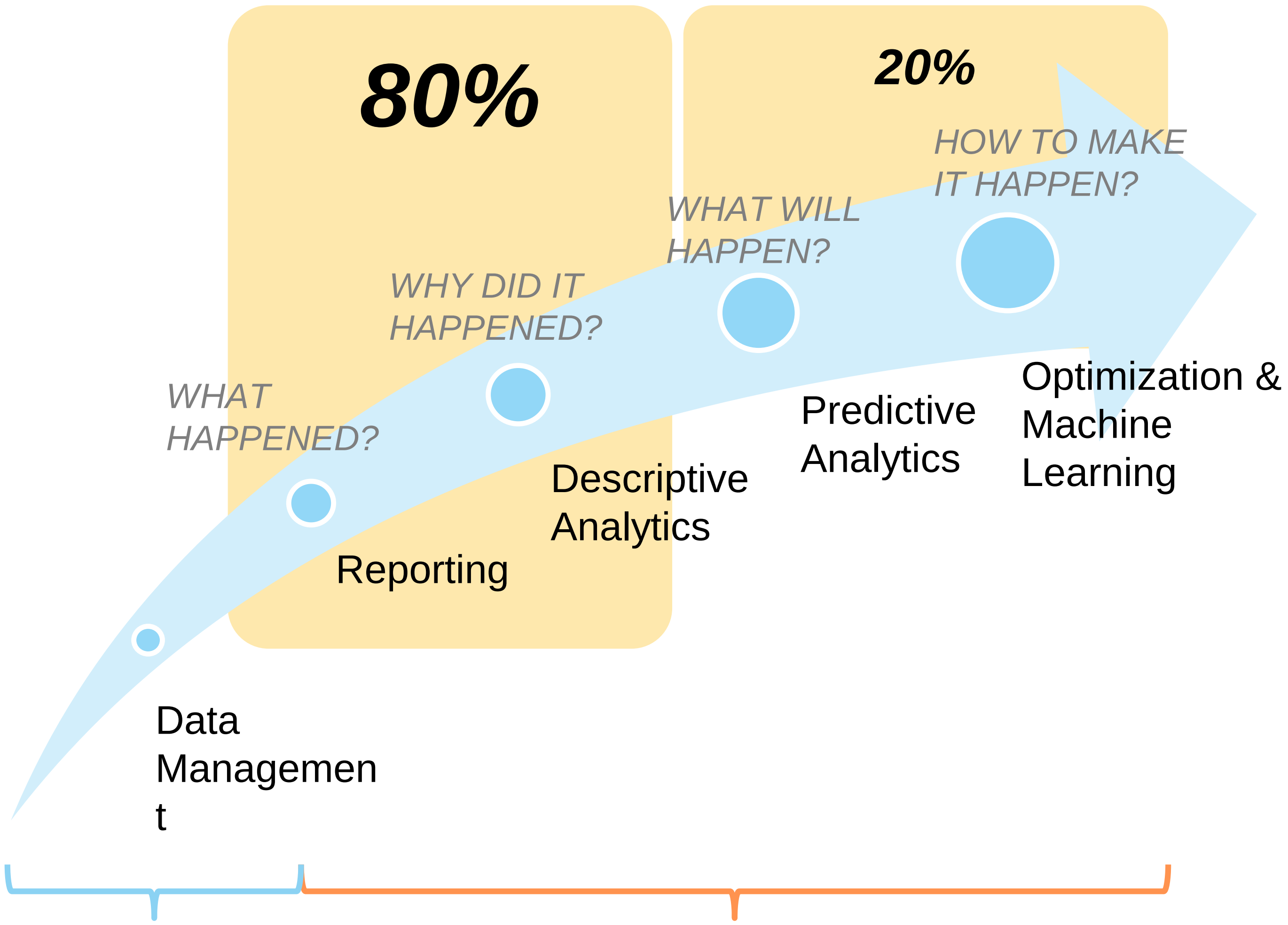
“Ada beberapa produk kami yang tidak laku, walau review sangat bagus”

“Kredit nasabah kami banyak yang macet”

“Stock barang selalu habis/terlalu banyak”

“Kami tidak tahu seberapa efisien sales person kami”

DATA SCIENCE CHALLENGES



80%

20%

WHAT HAPPENED?

Data Managemen t

Reporting

WHY DID IT HAPPENED?

Descriptive Analytics

WHAT WILL HAPPEN?

Predictive Analytics

HOW TO MAKE IT HAPPEN?

Optimization & Machine Learning

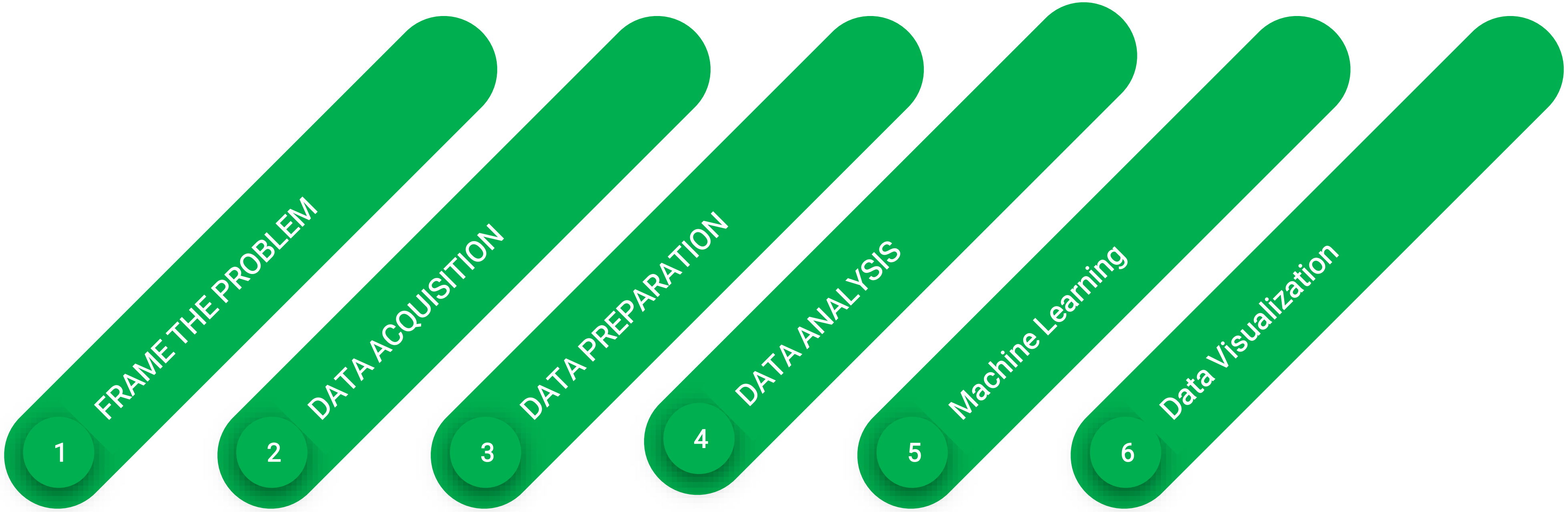
Data gathering, restructuring

Extracting information

Purwadhika
Startup and Coding School

DATA SCIENCE WORKFLOW

DATA SCIENCE WORKFLOW



1 FRAME THE PROBLEM

2 DATA ACQUISITION

3 DATA PREPARATION

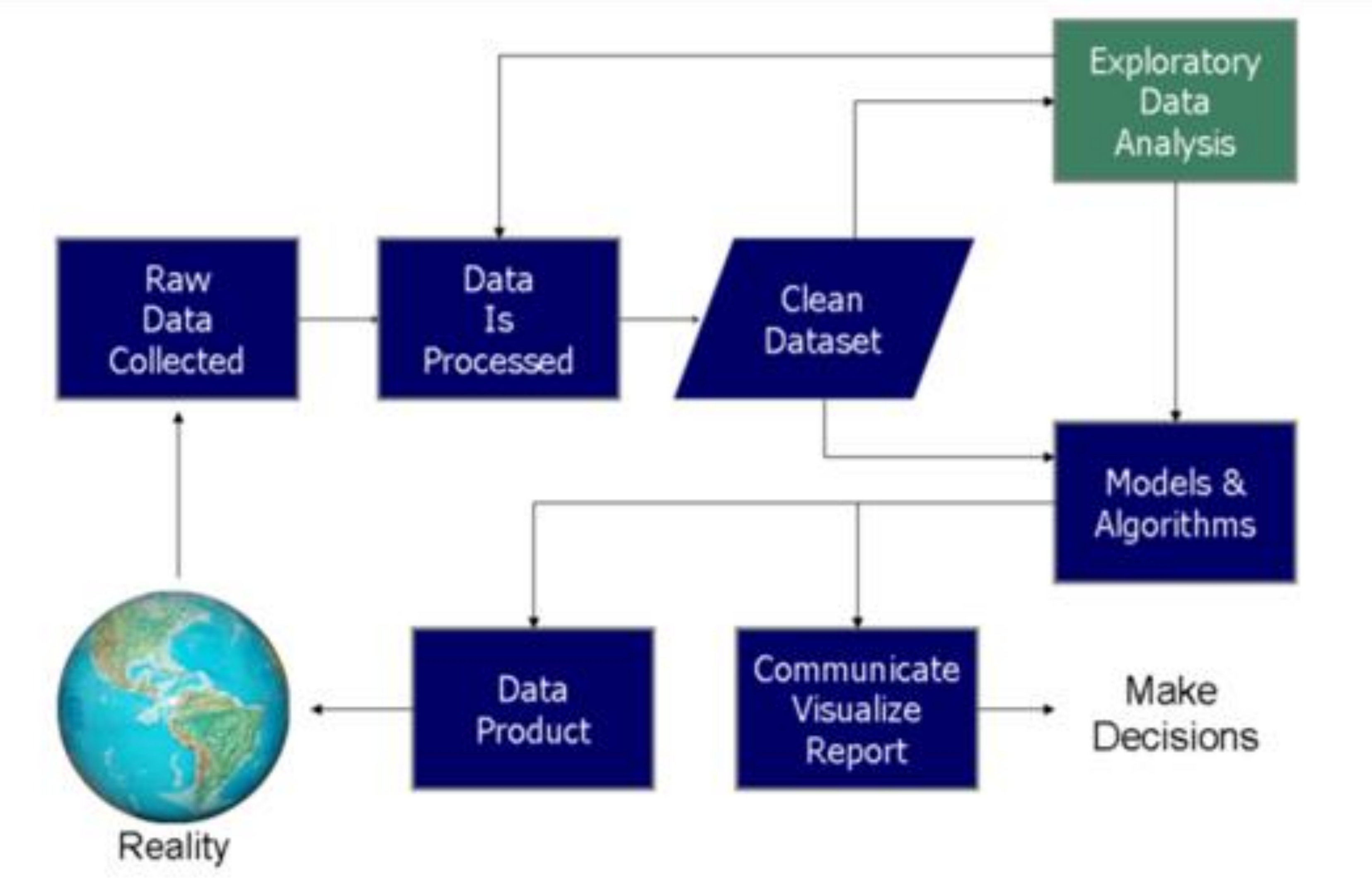
4 DATA ANALYSIS

5 Machine Learning

6 Data Visualization

Purwadhika
Startup and Coding School

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 ressources
- World trends/activity

DATA SCIENCE WORKFLOW

Data Location

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

DATA SCIENCE WORKFLOW

Data Format

- Hard copy
- Digital Documents
- Database
- Streams

DATA SCIENCE WORKFLOW

Data Types

- Numerical
- Text
- Image
- Audio
- Video

DATA SCIENCE WORKFLOW

Data Access

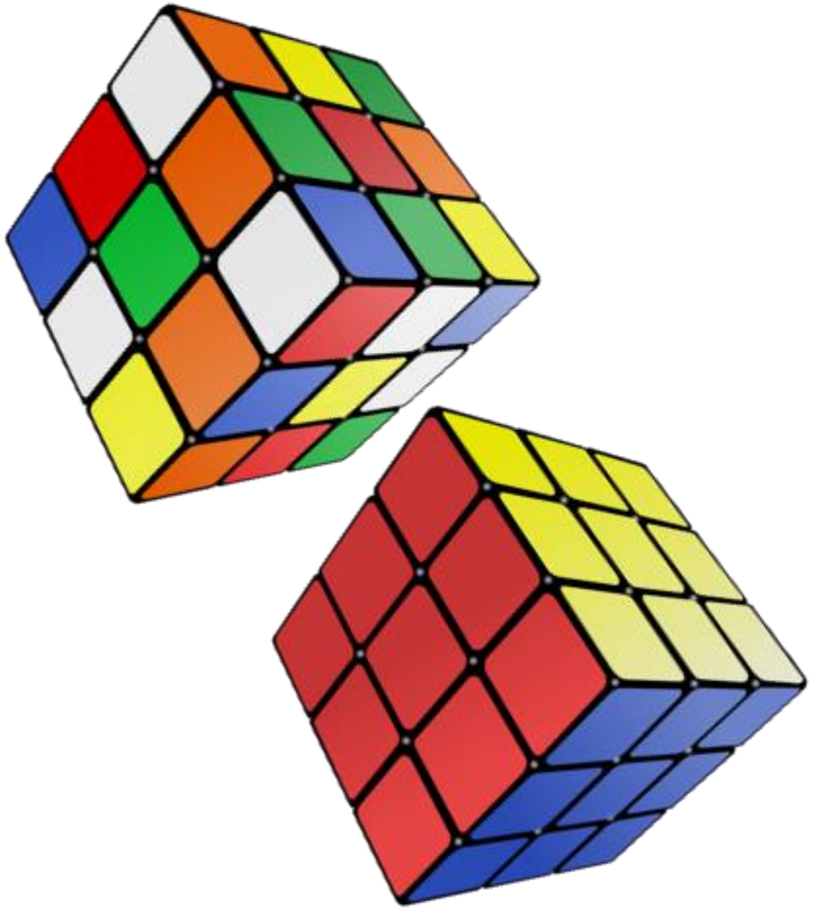
- Data warehousing
- REST API
- Web scraping

DATA SCIENCE WORKFLOW

Data Privacy

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

DATA SCIENCE WORKFLOW



Structured Data

Vs

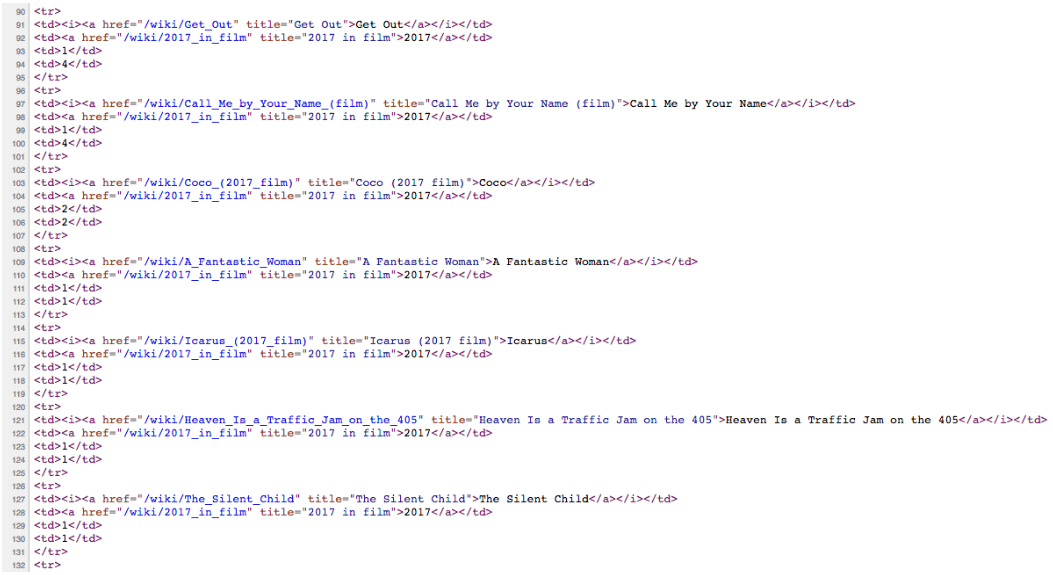
Unstructured Data

Purwadhika
Startup and Coding School

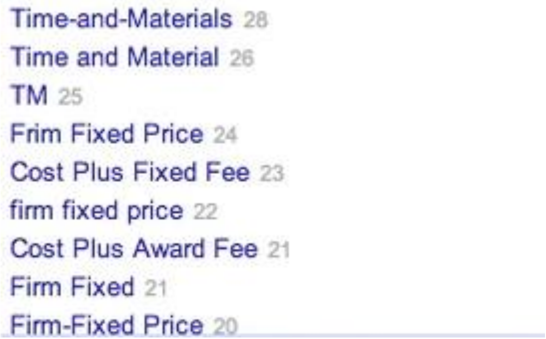
DATA SCIENCE WORKFLOW

Data preparation

- Data cleansing
  - Format normalization
  - Typing inconsistency
- Handling NULL values
- Handling outliers
- Feature Selection/Engineering



```html
<tr>
<td><i><a href="/wiki/Get_Out" title="Get Out">Get Out</a></i></td>
<td><a href="/wiki/2017_in_film" title="2017 in film">2017</a></td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td><i><a href="/wiki/Call_Me_by_Your_Name_(film)" title="Call Me by Your Name (film)">Call Me by Your Name</a></i></td>
<td><a href="/wiki/2017_in_film" title="2017 in film">2017</a></td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td><i><a href="/wiki/Coco_(2017_film)" title="Coco (2017 film)">Coco</a></i></td>
<td><a href="/wiki/2017_in_film" title="2017 in film">2017</a></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><i><a href="/wiki/A_Fantastic_Woman" title="A Fantastic Woman">A Fantastic Woman</a></i></td>
<td><a href="/wiki/2017_in_film" title="2017 in film">2017</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><i><a href="/wiki/Icarus_(2017_film)" title="Icarus (2017 film)">Icarus</a></i></td>
<td><a href="/wiki/2017_in_film" title="2017 in film">2017</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><i><a href="/wiki/Heaven_Is_a_Traffic_Jam_on_the_405" title="Heaven Is a Traffic Jam on the 405">Heaven Is a Traffic Jam on the 405</a></i></td>
<td><a href="/wiki/2017_in_film" title="2017 in film">2017</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><i><a href="/wiki/The_Silent_Child" title="The Silent Child">The Silent Child</a></i></td>
<td><a href="/wiki/2017_in_film" title="2017 in film">2017</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
```



Time-and-Materials	28
Time and Material	26
TM	25
Frim Fixed Price	24
Cost Plus Fixed Fee	23
firm fixed price	22
Cost Plus Award Fee	21
Firm Fixed	21
Firm-Fixed Price	20

Purwadhika
Startup and Coding School

DATA SCIENCE WORKFLOW : DATA ANALYSIS AND VISUALIZATION

DATA ANALYSIS

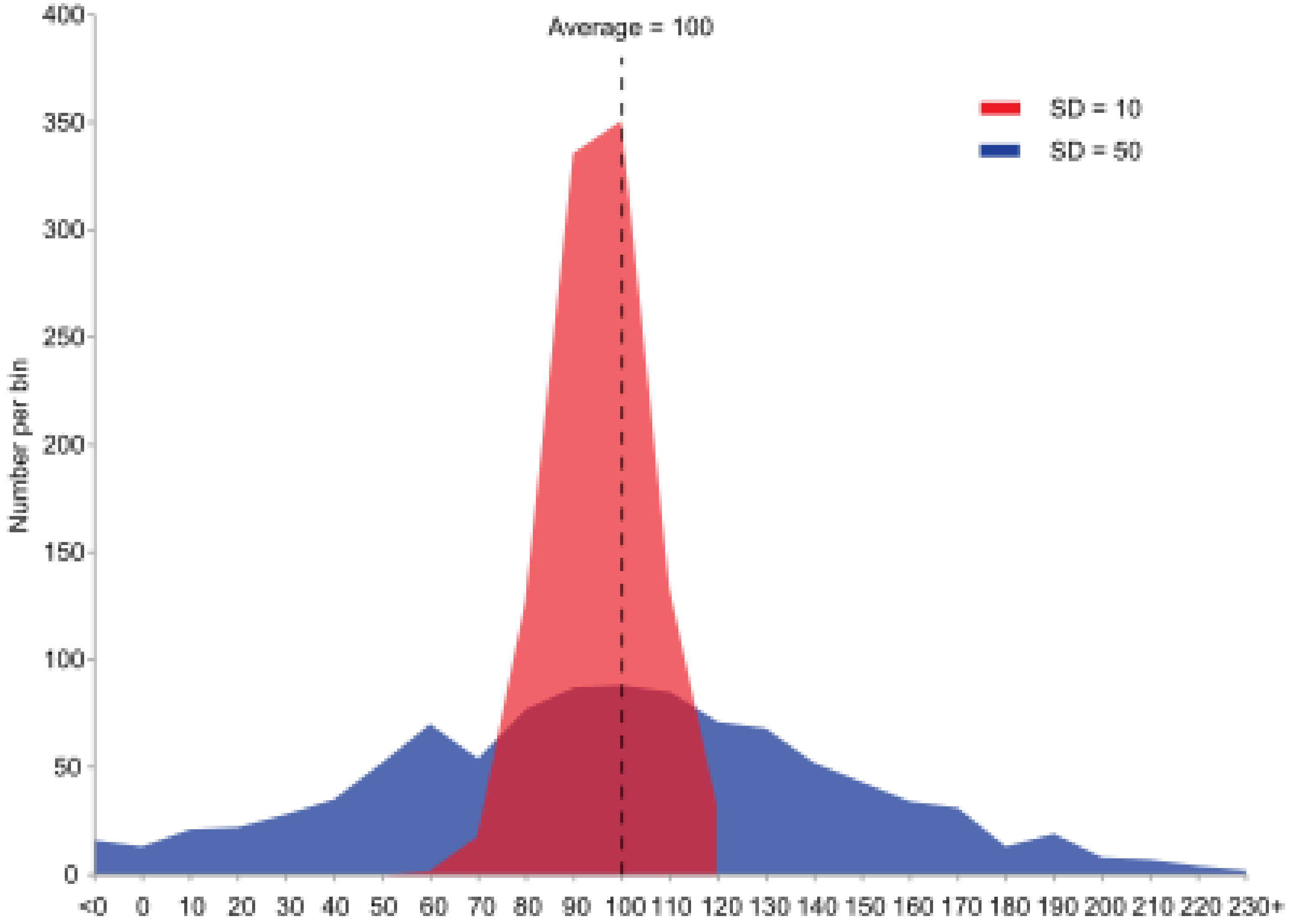
- Always aim to answer the problem definition
- Identify
  - Variations
  - Correlations
  - Trends
  - Outliers

Data Analysis Term

- **Variation** is a measure of how far a set of random numbers are spread out from their mean value. If one is familiar with Standard Deviation, variance could be calculated by squaring the Standard Deviation.
- **Correlation** is a degree of relationship between two or more numerical variable. In statistical term, it could be define as how pair of variable are linearly related.
- **Trend** is a general pattern of how condition, output, process, or general tendency of series data point move in a certain direction over course of time.
- **Outlier** is a data point that differ significantly from the other data observation

Variation

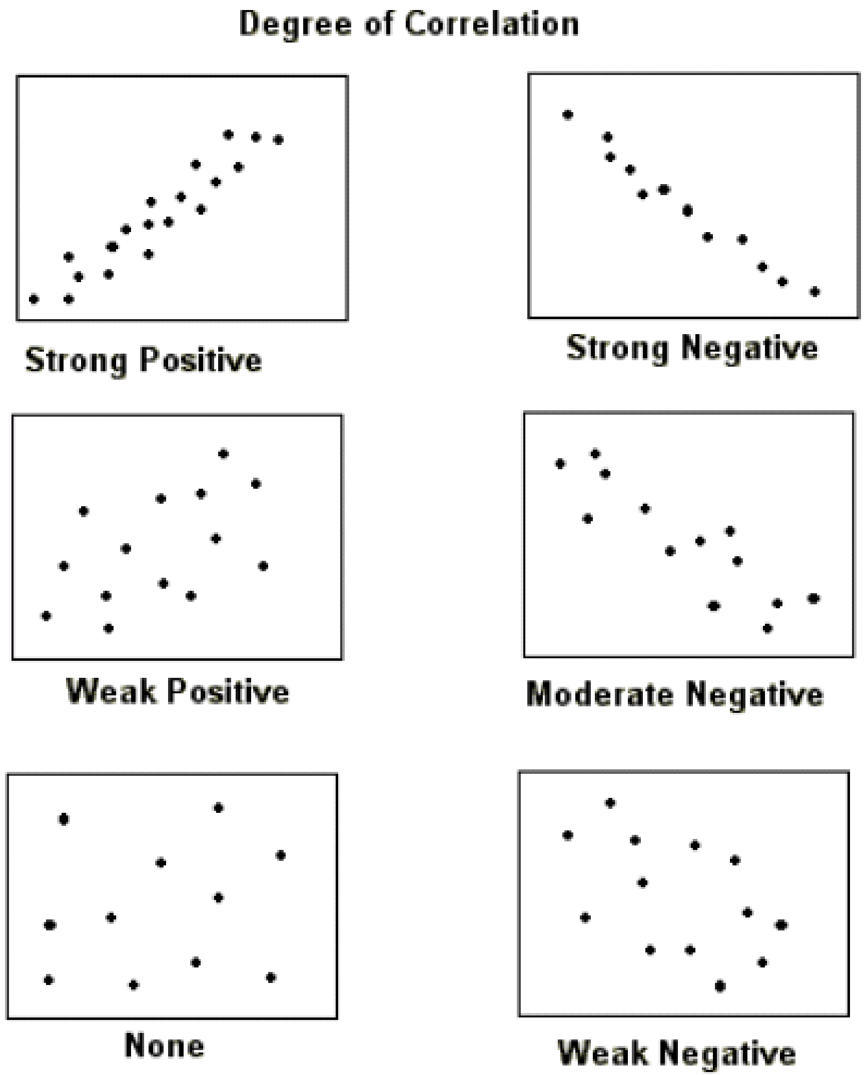
Example of samples from two populations with the same mean but different variances. The red population has mean 100 and variance 100 (SD=10) while the blue population has mean 100 and variance 2500 (SD=50).



Purwadhika
Startup and Coding School

Source: https://en.wikipedia.org/wiki/Variance

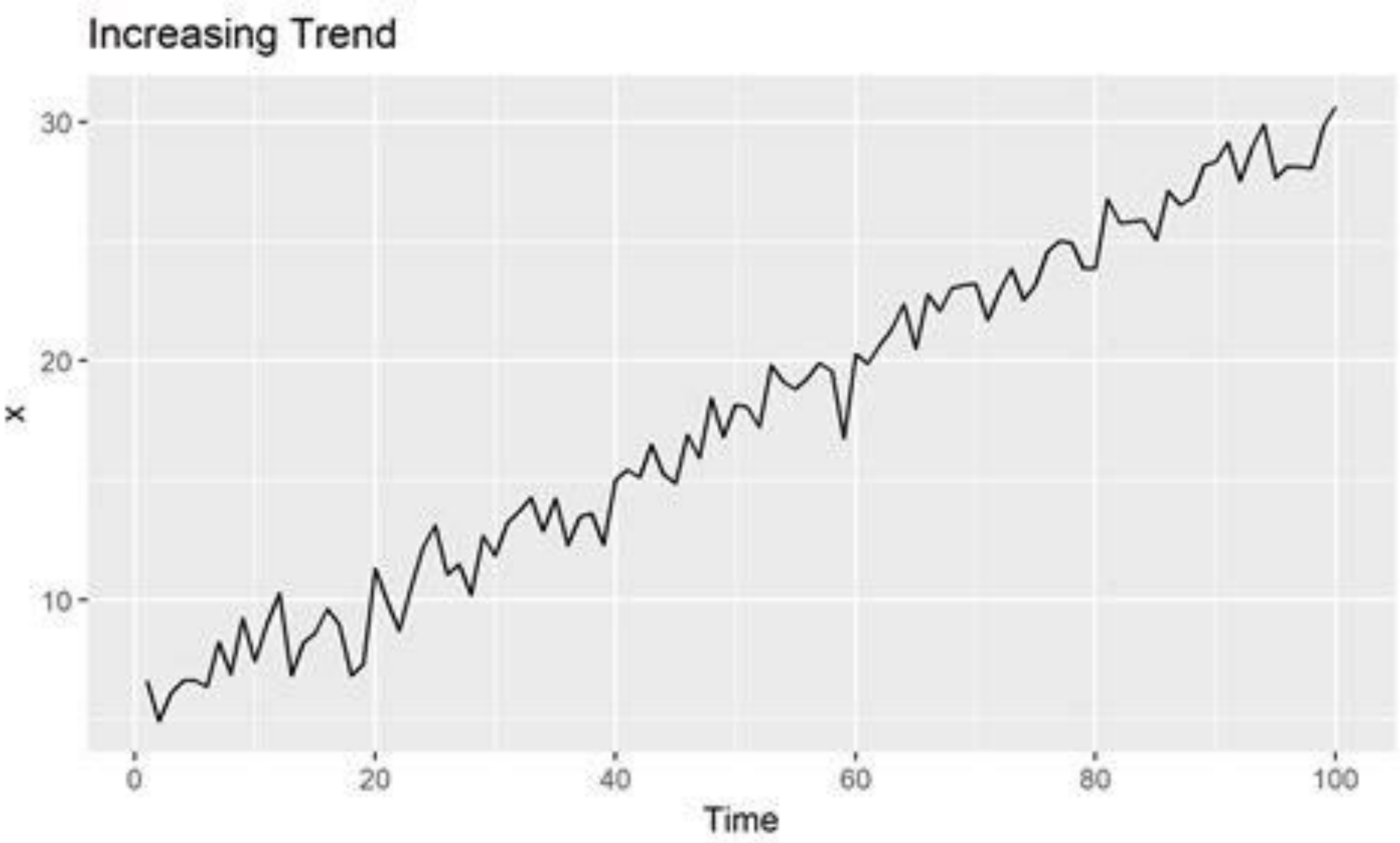
Correlation



Source: financetrain.com

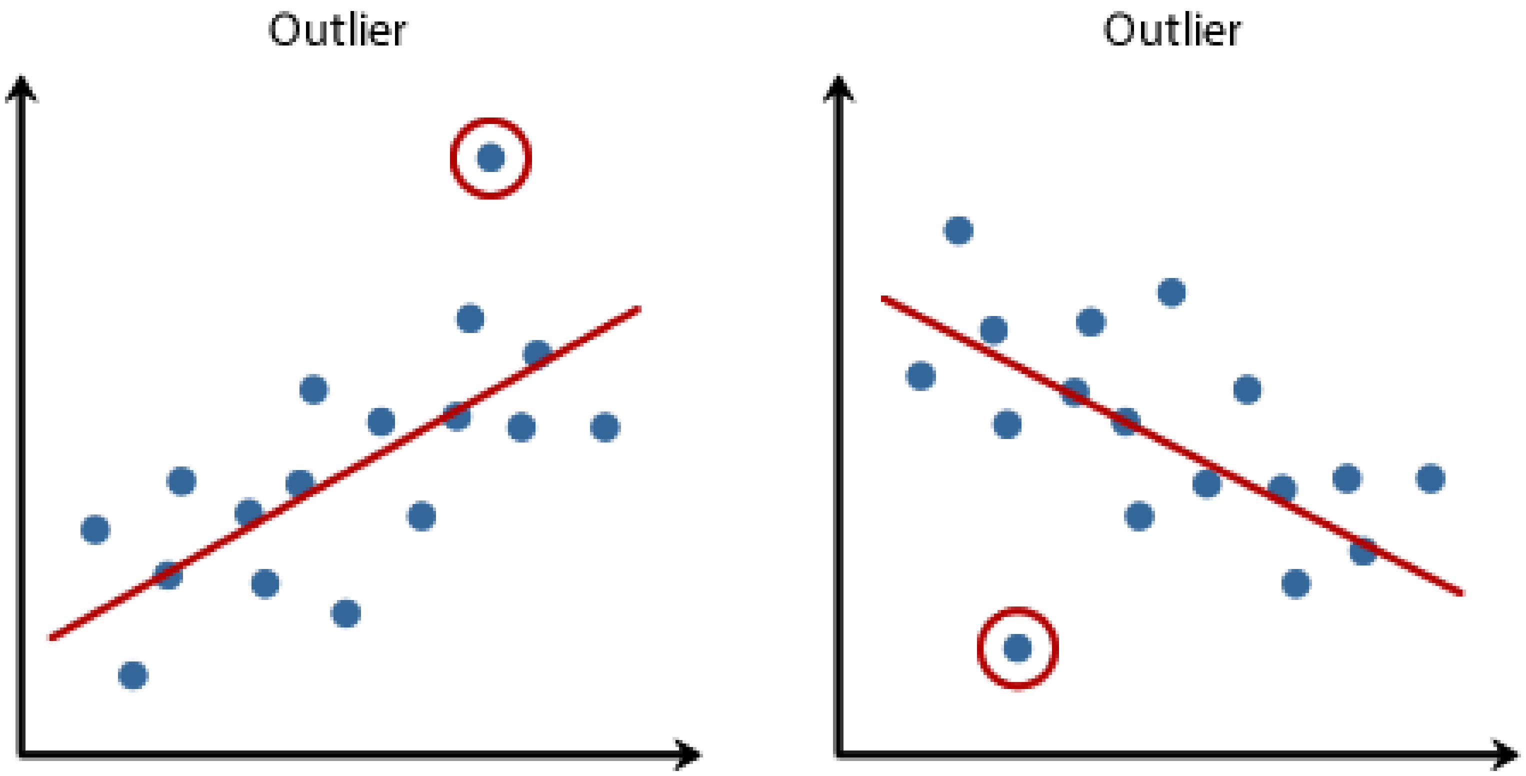
Purwadhika
Startup and Coding School

Trend



Source: datascience.com

Outlier



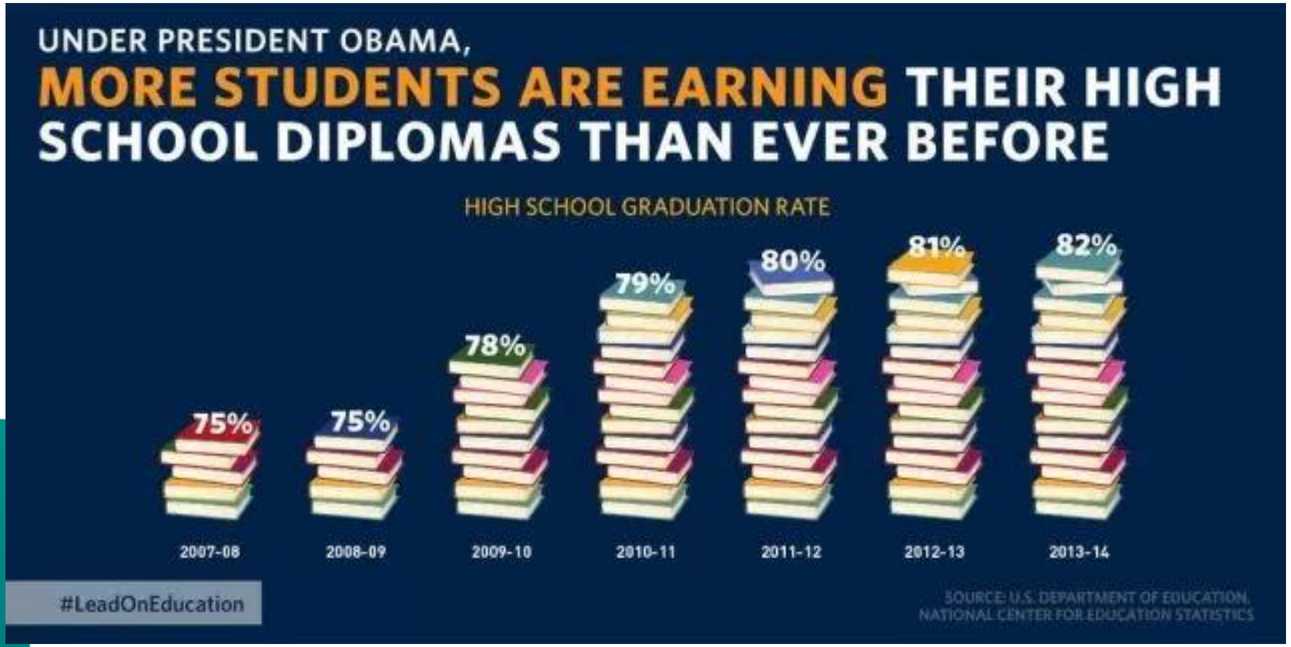
Copyright 2014. Laerd Statistics.

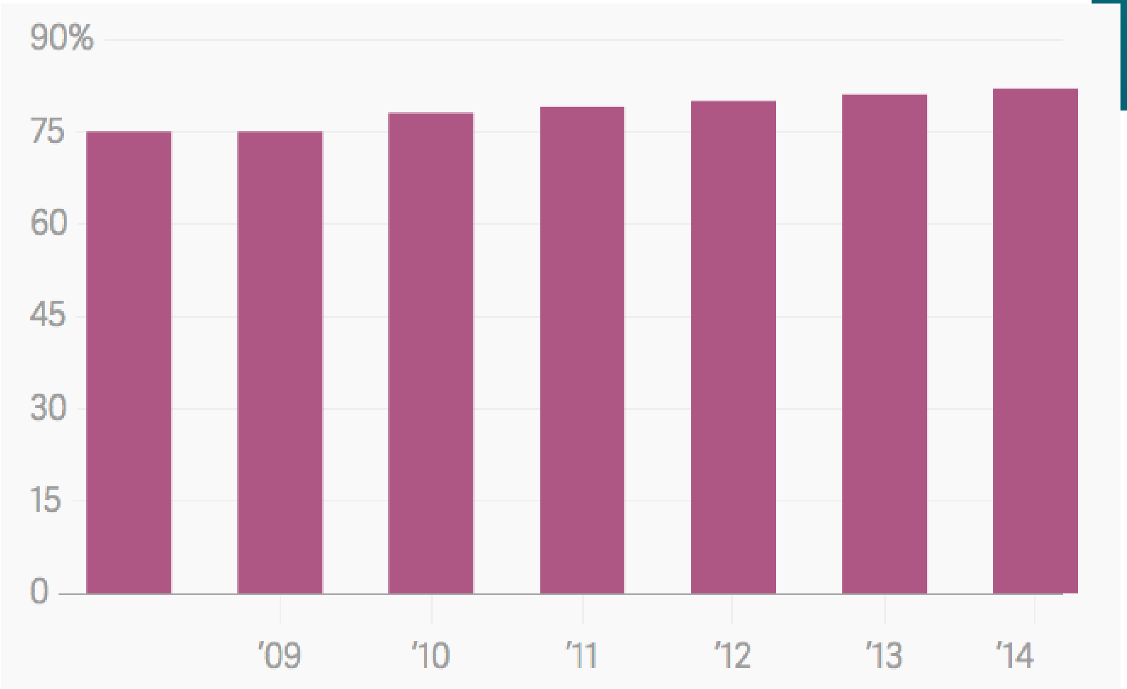
DATA SCIENCE WORKFLOW : DATA ANALYSIS AND VISUALIZATION

DATA Visualization

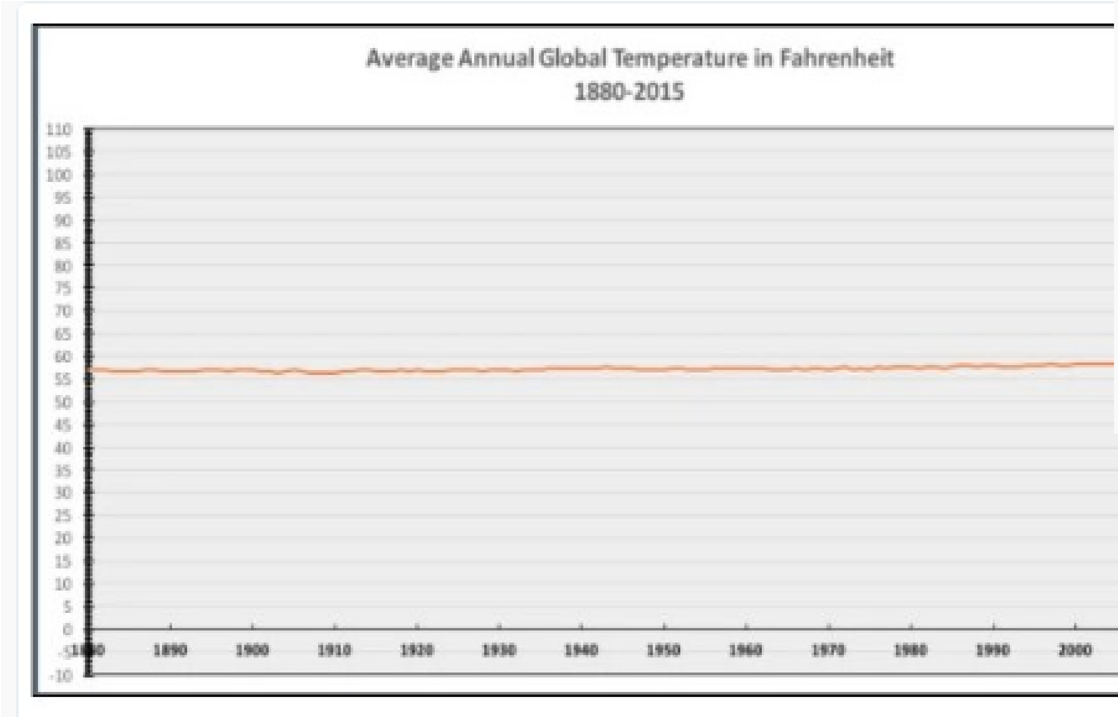
- Know the audience
- Visualization is all about perception

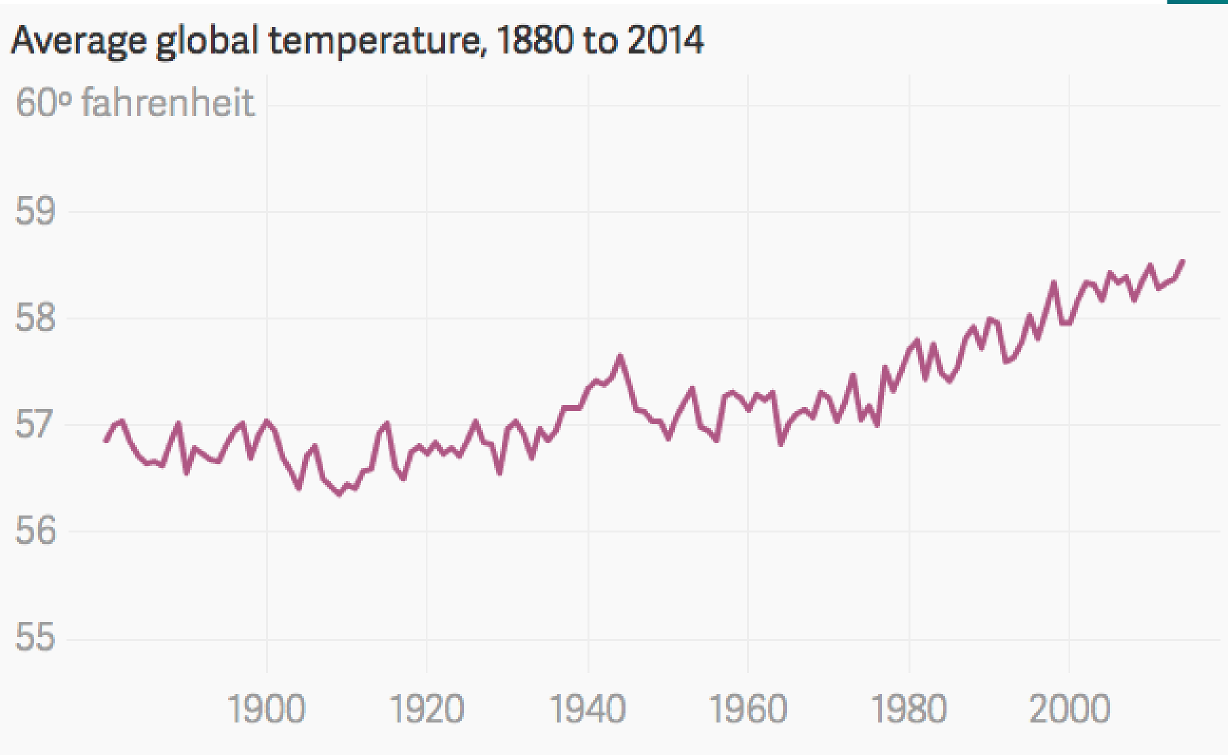
DATA SCIENCE WORKFLOW : DATA ANALYSIS AND VISUALIZATION





DATA SCIENCE WORKFLOW : DATA ANALYSIS AND VISUALIZATION





DATA SCIENCE ROLES

Data Scientist

Activities

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

Tools

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

Skills and Talents

- Statistical theories and methodologies
- Database systems
- Programming skills

DATA SCIENCE ROLES

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

DATA SCIENCE ROLES

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

DATA SCIENCE ROLES

Domain Expert

Activities

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

Skills and Talents

- Business and Domain knowledge
- Communication

Tools

- (depends on the field)

DATA SCIENCE ROLES

Other roles

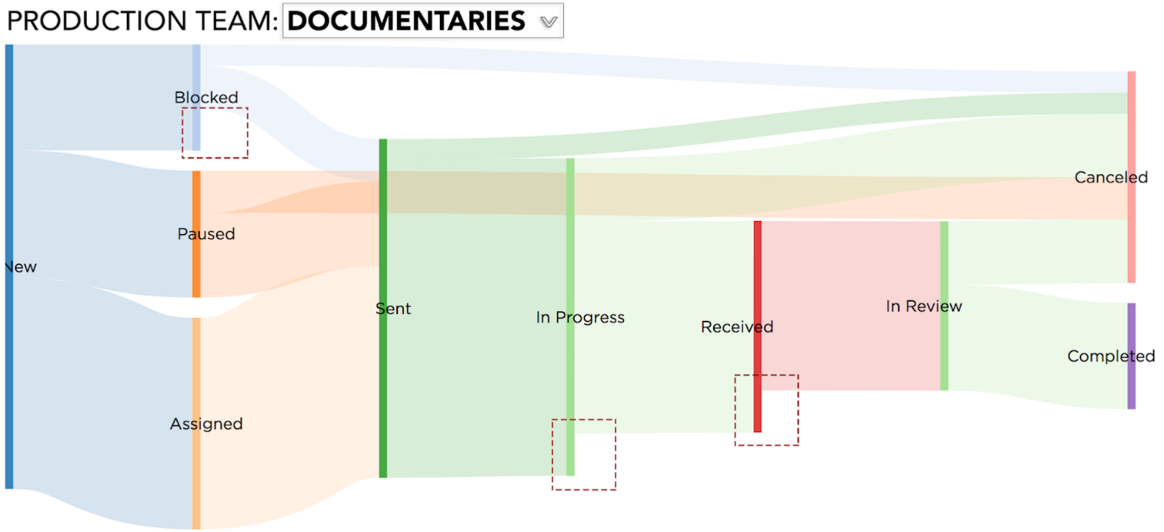
- Database Admin : Query/Prepare data to be processed/analyse
- Data Architect : Design information archtect
- Statistician :
- Developer

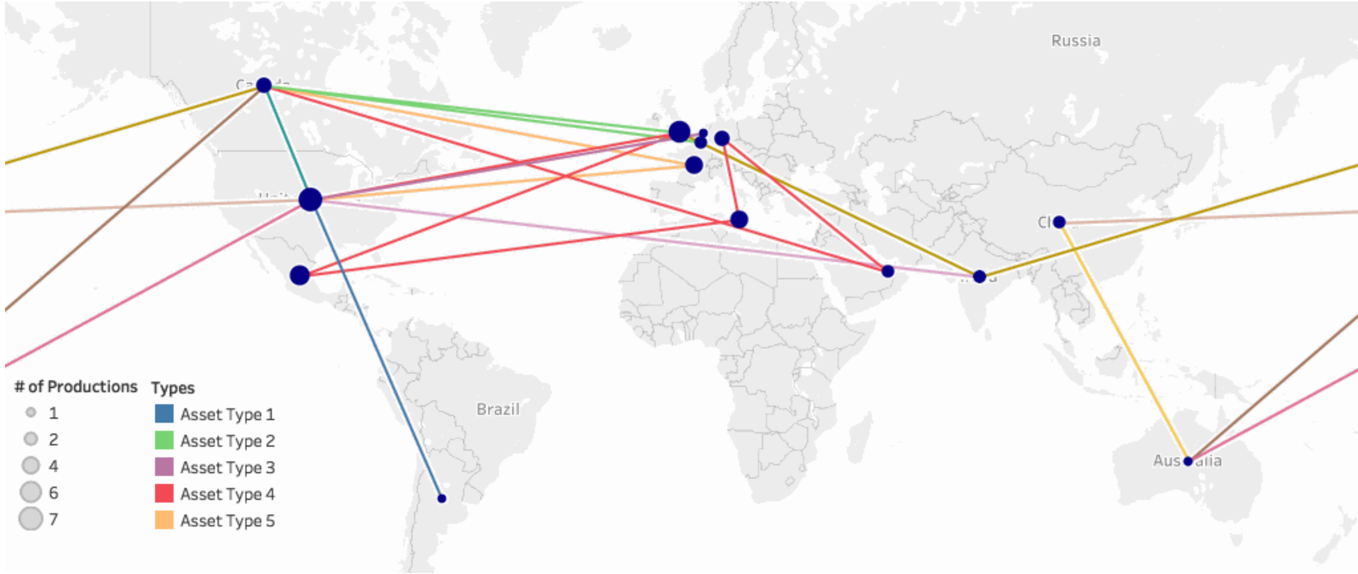
DATA SCIENCE ROLES

Keep it lean, grow as you go

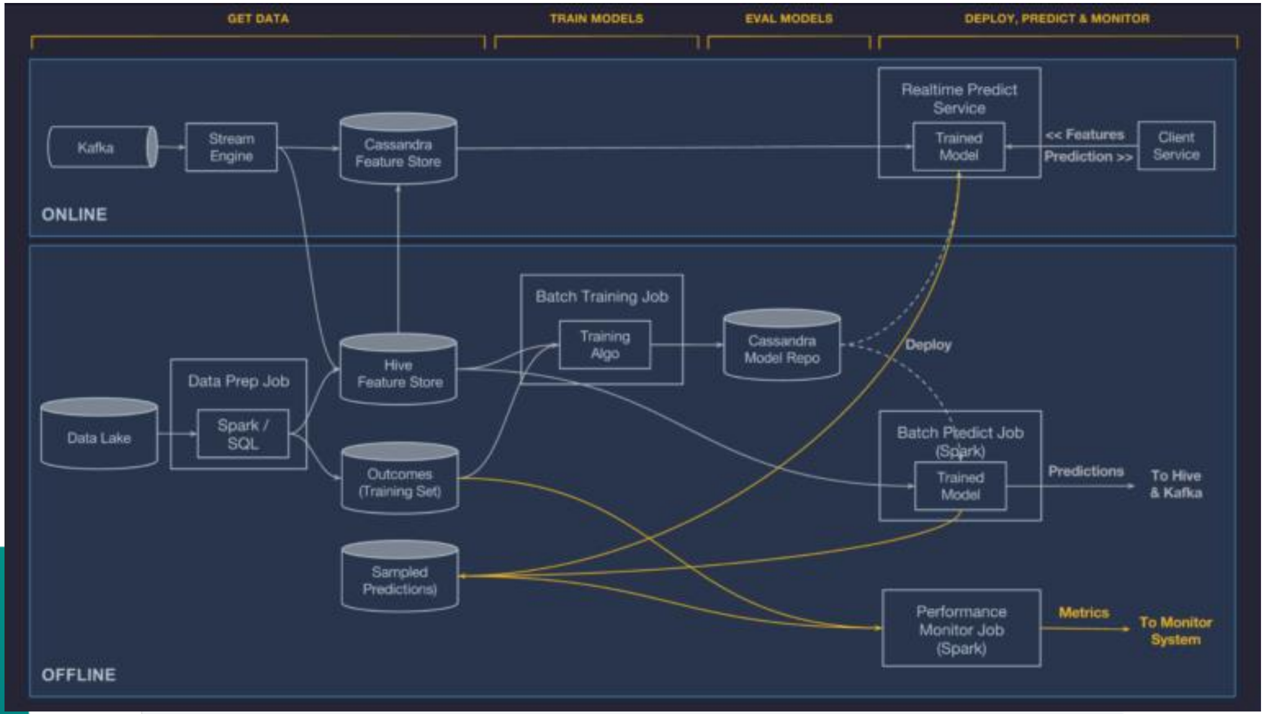
USECASE : NETFLIX

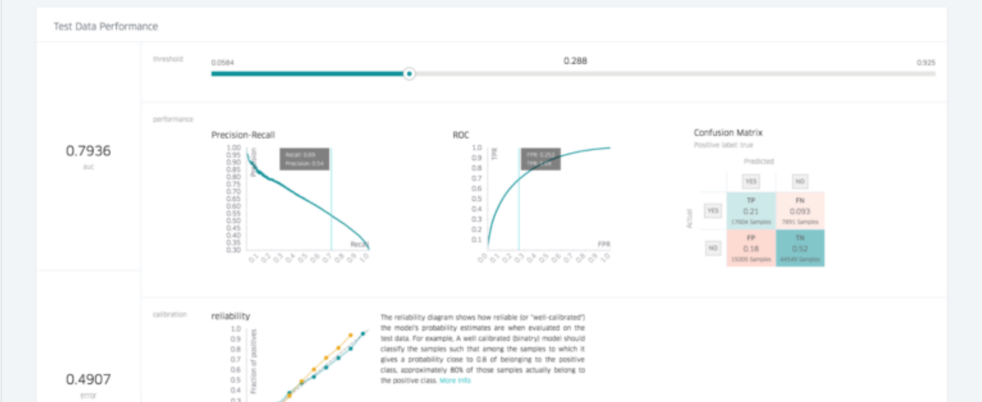
- Pre-production cost estimation
  - Location
  - Crews
  - Schedule
- Shooting schedules
- Post-Production assets progression
- Prioritazion of location


PRODUCTION TEAM: **DOCUMENTARIES**



Study Case : **UberEATS** estimated time of delivery model





The reliability diagram shows how reliable (or "well-calibrated") the model's probability estimates are when evaluated on the test data. For example, A well calibrated (binatry) model should classify the samples such that among the samples to which it gives a probability close to 0.8 of belonging to the positive class, approximately 80% of those samples actually belong to the positive class. More Info

Michelangelo: Uber's Machine Learning Platform

Machine Learning Workflow:

- Manage data
- Train models
- Evaluate models
- Deploy models
- Make predictions
- Monitor predictions

ASSIGNMENT

Challenge :

Unilever akan mengeluarkan varian shampo baru. Direksi meminta bantuan kepada tim Data Science untuk memberikan rekomendasi spesifikasi varian tersebut.

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

Purwadhika
Startup and Coding School

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