Introduction to Data Science

Last Update: 28 Nov 2020

Chapter 1
Overview of Data Science

Papangkorn Inkeaw, PhD

Department of Computer Science, Faculty of Science
Chiang Mai University

Outline

Overview of Data Science

- **What is the Data Science?**
- **Example Applications**
- **Related Study Fields**
- **Levels of Data Analytics Maturity**
- **Data Science Process**

What is the Data Science?

Overview of Data Science

Data

- Facts and statistics collected together for reference or analysis.
- A set of values of subjects with respect to qualitative or quantitative variables.

Science

A systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe.

Use Science to Understand Data

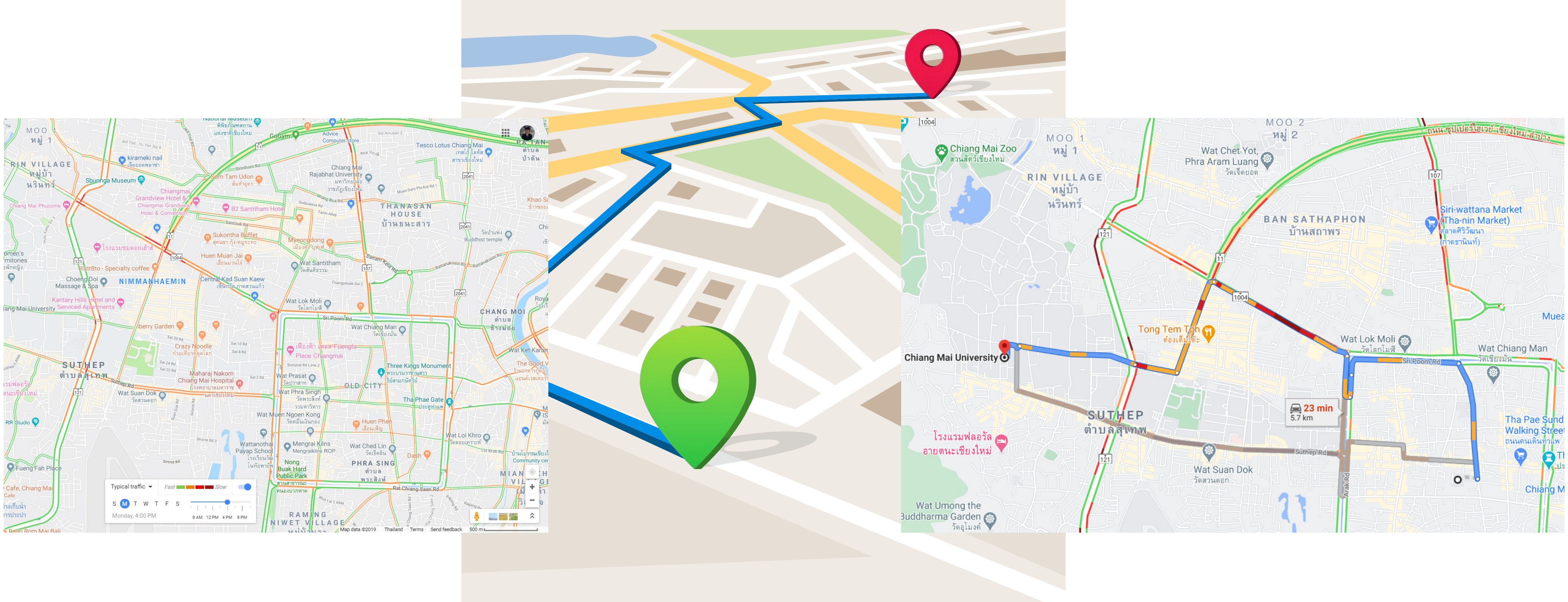
Data Science

Uses of scientific methods, processes, algorithms and systems to extract knowledge and insights from data.

Example Applications

Overview of Data Science

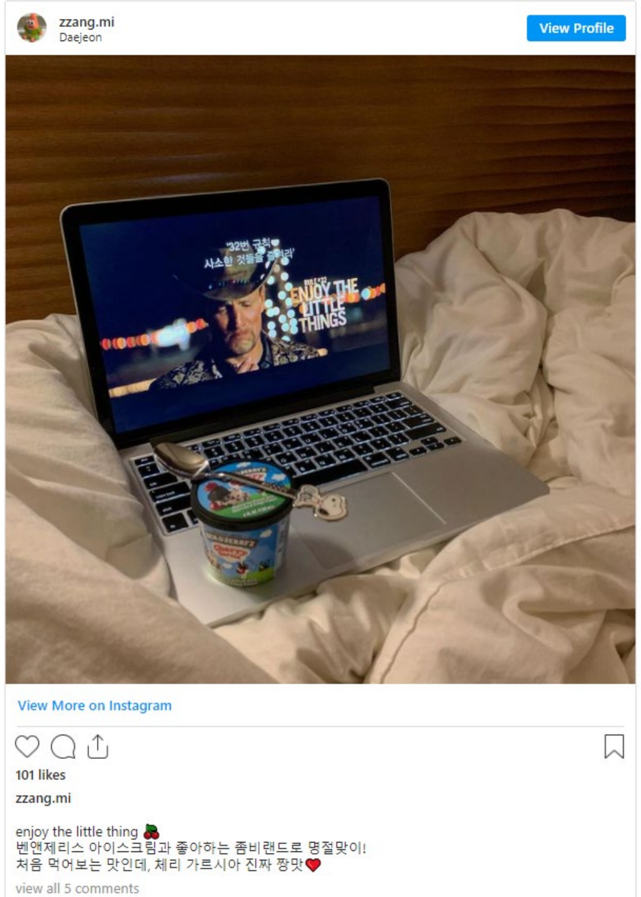
We use traffic data to discover the best time and route to go back to home.



Example Applications

Overview of Data Science

New ice cream flavor 'Netflix & Chill’d’ from social listening tool



zzang.mi
Daejeon

101 likes

zzang.mi

enjoy the little thing 🍒
벤앤제리스 아이스크림과 좋아하는 좀비랜드로 명절맞이!
처음 먹어보는 맛인데, 체리 가르시아 진짜 짱맛❤️

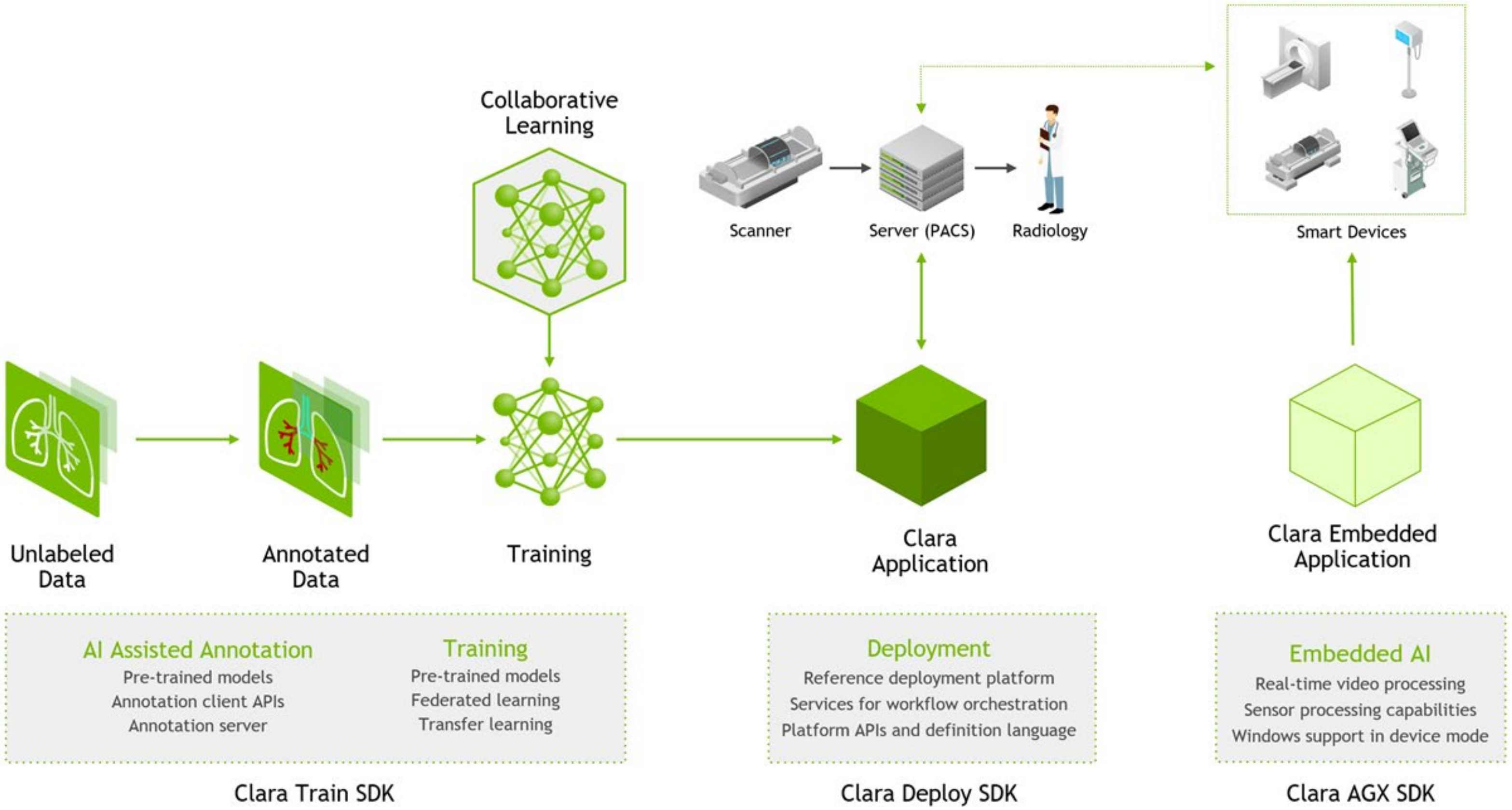
view all 5 comments

https://twitter.com/benandjerrys

Example Applications

Overview of Data Science

Develop computer-aided medical diagnosis using NVIDIA Clara



https://developer.nvidia.com/clara

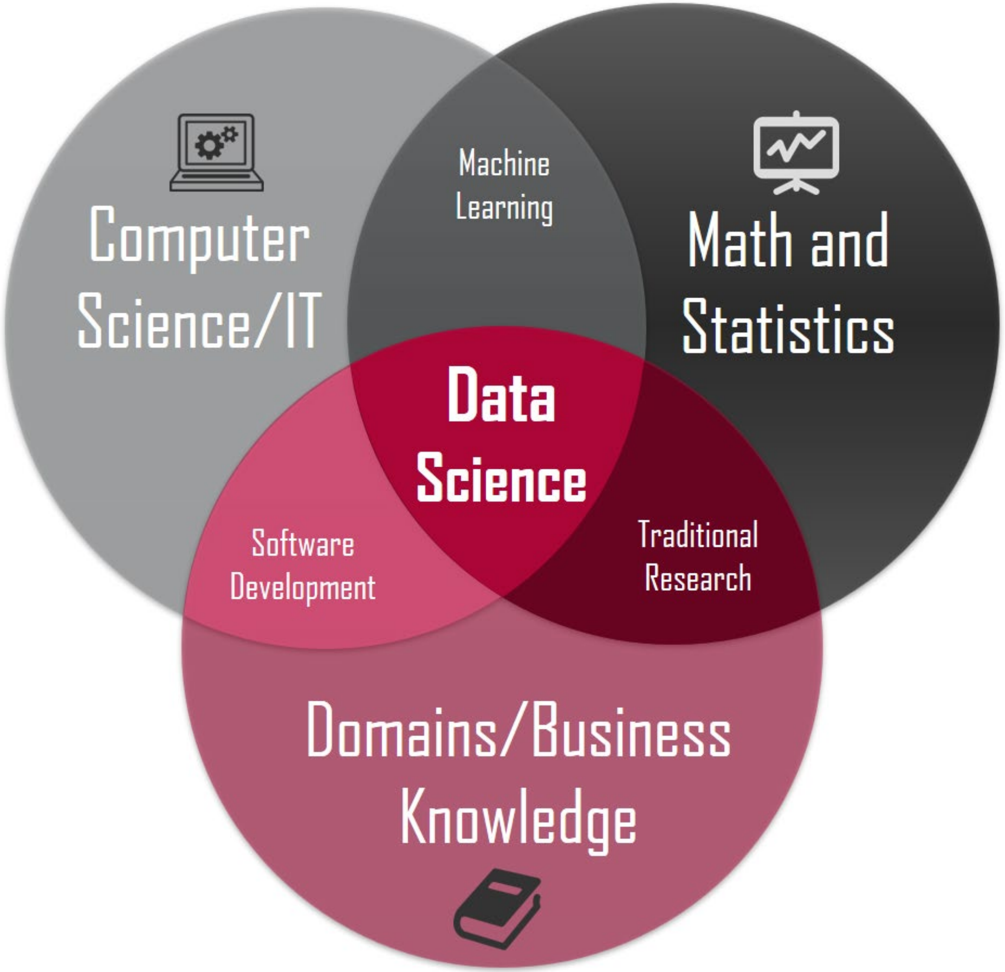
Related Study Fields

Overview of Data Science

Data Science

is a multi-disciplinary field combining

- Computer Science
- Mathematics and Statistics
- Business Knowledge



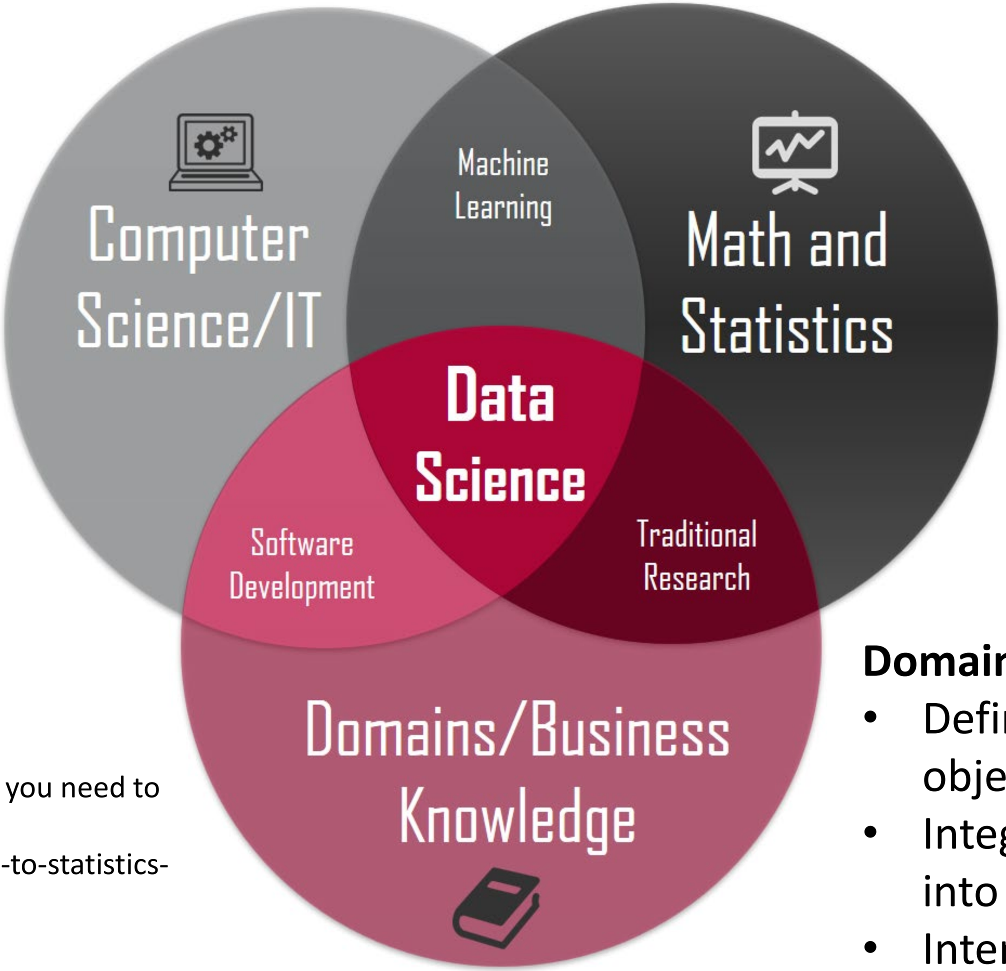
Michael Barber (2018). Data science concepts you need to know! Part 1. Retrieved from https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

Related Study Fields

Overview of Data Science

Computer Science

- Provide tools (Software, Programming languages)
- Data Engineering



Mathematics and Statistics

- Analytics tools (Math and state models, method)

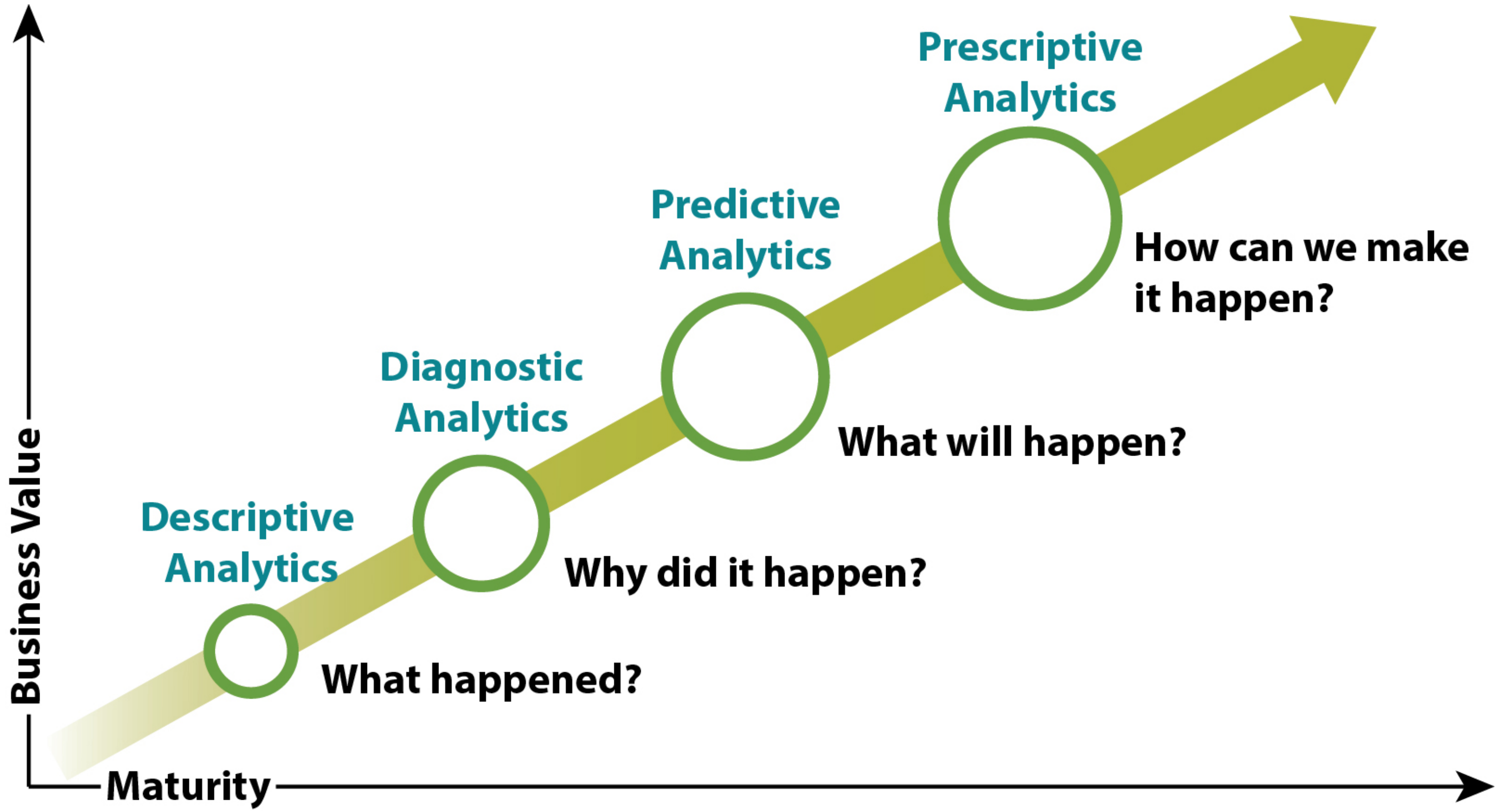
Domain Expertise

- Define business problem and objectives
- Integrate business knowledge into data analysis process
- Interpret results

Michael Barber (2018). Data science concepts you need to know! Part 1. Retrieved from https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

Levels of Data Analytics Maturity

Overview of Data Science



Levels of Data Analytics Maturity

Overview of Data Science

Descriptive Analytics

- Univariate Analysis
- Use descriptive statistics
- Ask the question “**What happened?**”

Diagnostic Analytics

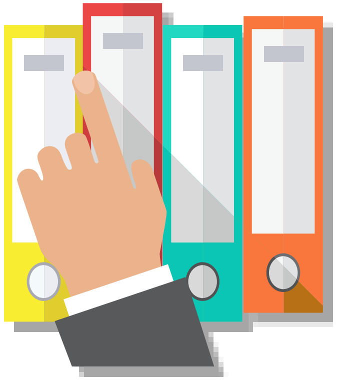
- Multivariate Analysis
  - Correlation Analysis
  - Association Analysis
- Ask the question “**Why did it happen?**”

Levels of Data Analytics Maturity

Overview of Data Science

Predictive Analytics

- Predict the futures or unknown values
- Construct a predictive model
- 2 approaches:
  - Statistical hypothesis test
  - Machine Learning
- Ask the question "**What will happen?**"



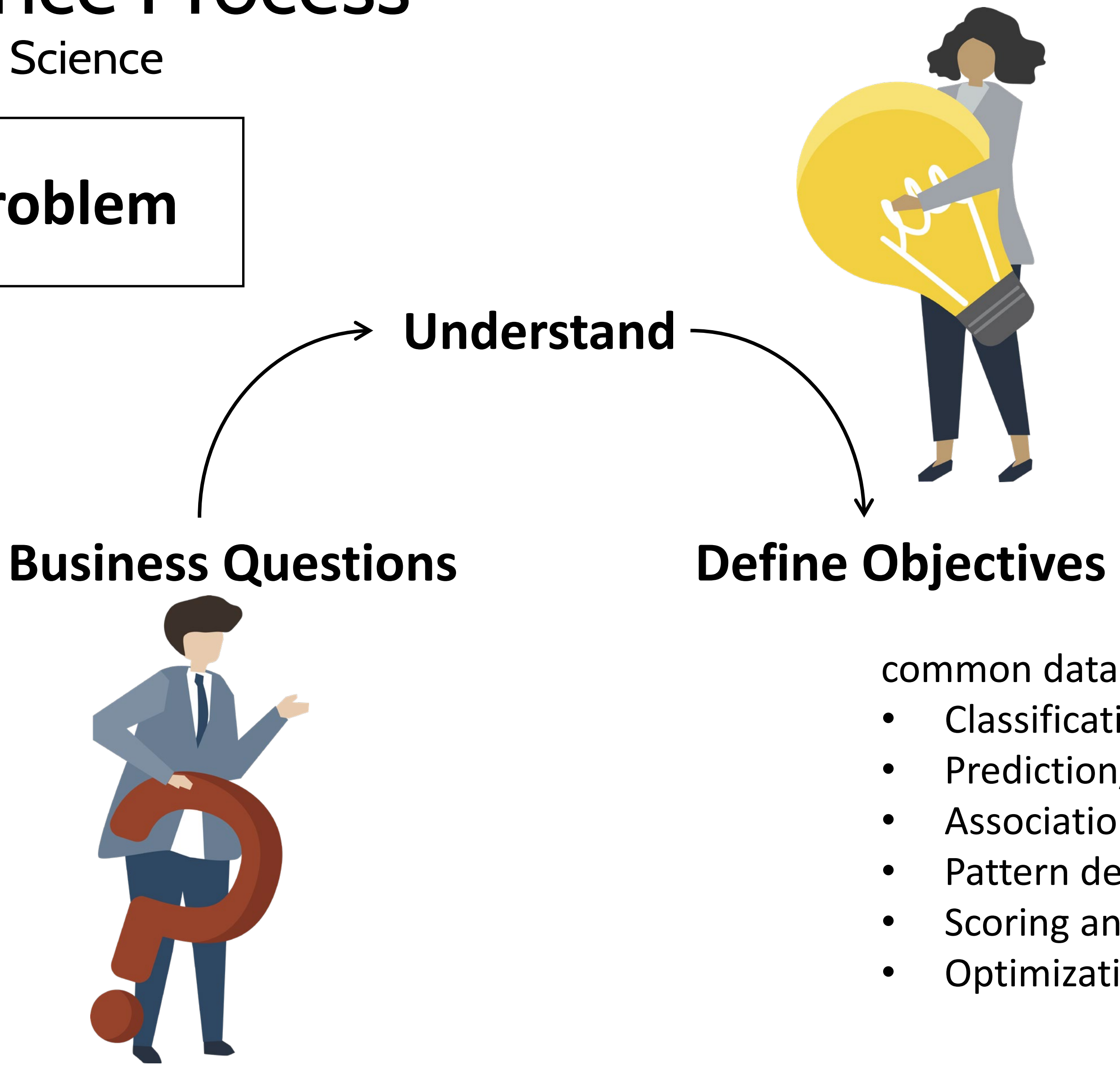
Prescriptive Analytics

- Decision support
- Optimization
- Simulation
- Ask the question
  "**How can we make it happen?**"

Data Science Process

Overview of Data Science

1 Business Problem



Understand

Business Questions

Define Objectives

common data science problems:

- Classification/Recognition
- Prediction/Regression
- Association
- Pattern detection/clustering
- Scoring and ranking
- Optimization

Data Science Process

Overview of Data Science

2 Data Acquisition

Data Sources

- Questionnaires
- Web servers
- Web services (API)
- Database
- Logs
- Online repositories

Data Science Process

Overview of Data Science

3 Data Preparation

Data Cleaning

- Inconsistent datatypes
- Missing data
- Duplicate data
- etc.

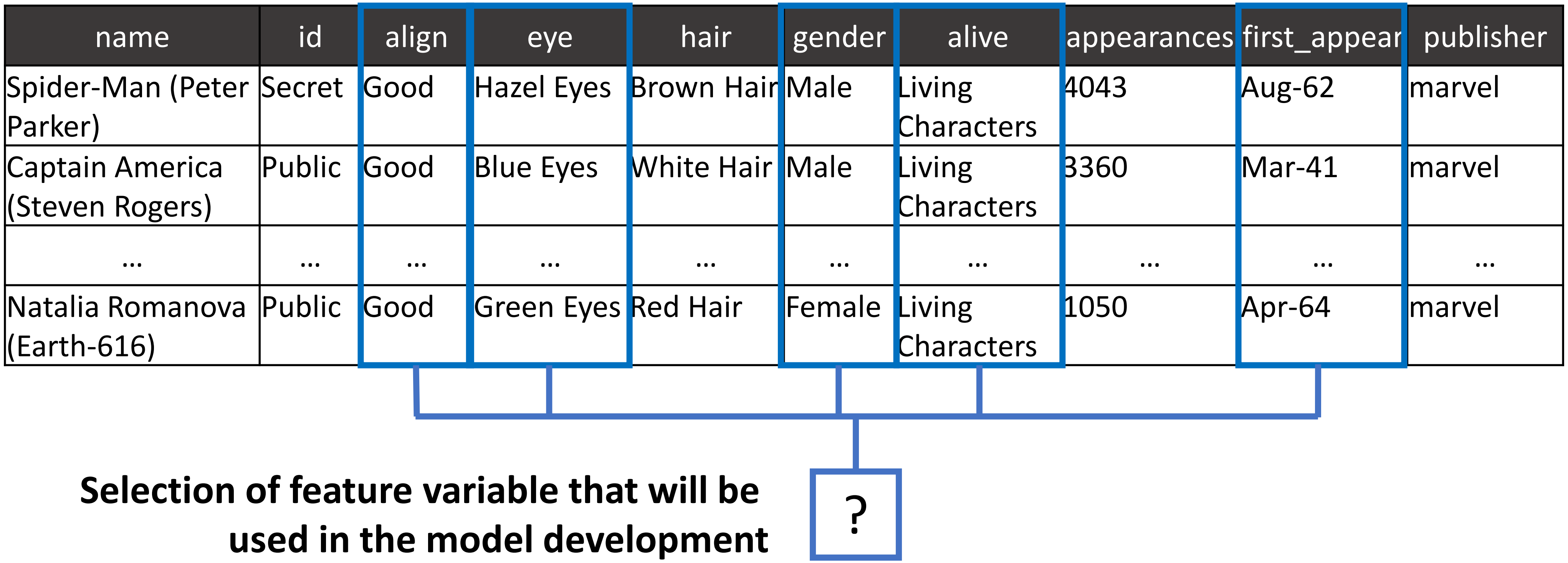
Data Transformation

- Converting data from one format/structure into another format or structure.
- Help to understand data structure
- Understand what we actually can do with the data

Data Science Process

Overview of Data Science

4 Exploratory Data Analysis



name	id	align	eye	hair	gender	alive	appearances	first_appear	publisher
Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
…	…	…	…	…	…	…	…	…	…
Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

Selection of feature variable that will be used in the model development

Data Science Process

Overview of Data Science

5 Data Modeling

Association

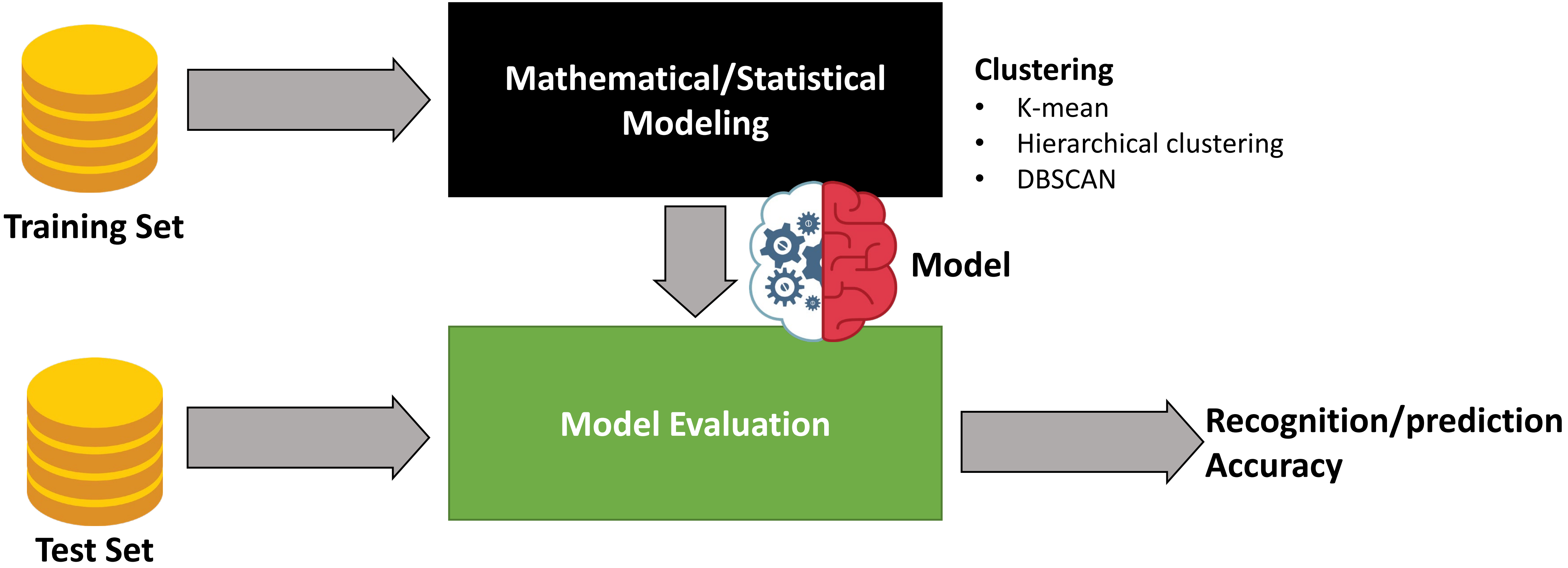
- Itemset mining
- Summarizing Itemsets

Recognition

- Decision tree
- kNN
- SVM

Regression

- Linear regression
- Polynomial regression



Training Set

Mathematical/Statistical Modeling

Clustering

- K-mean
- Hierarchical clustering
- DBSCAN

Model

Test Set

Model Evaluation

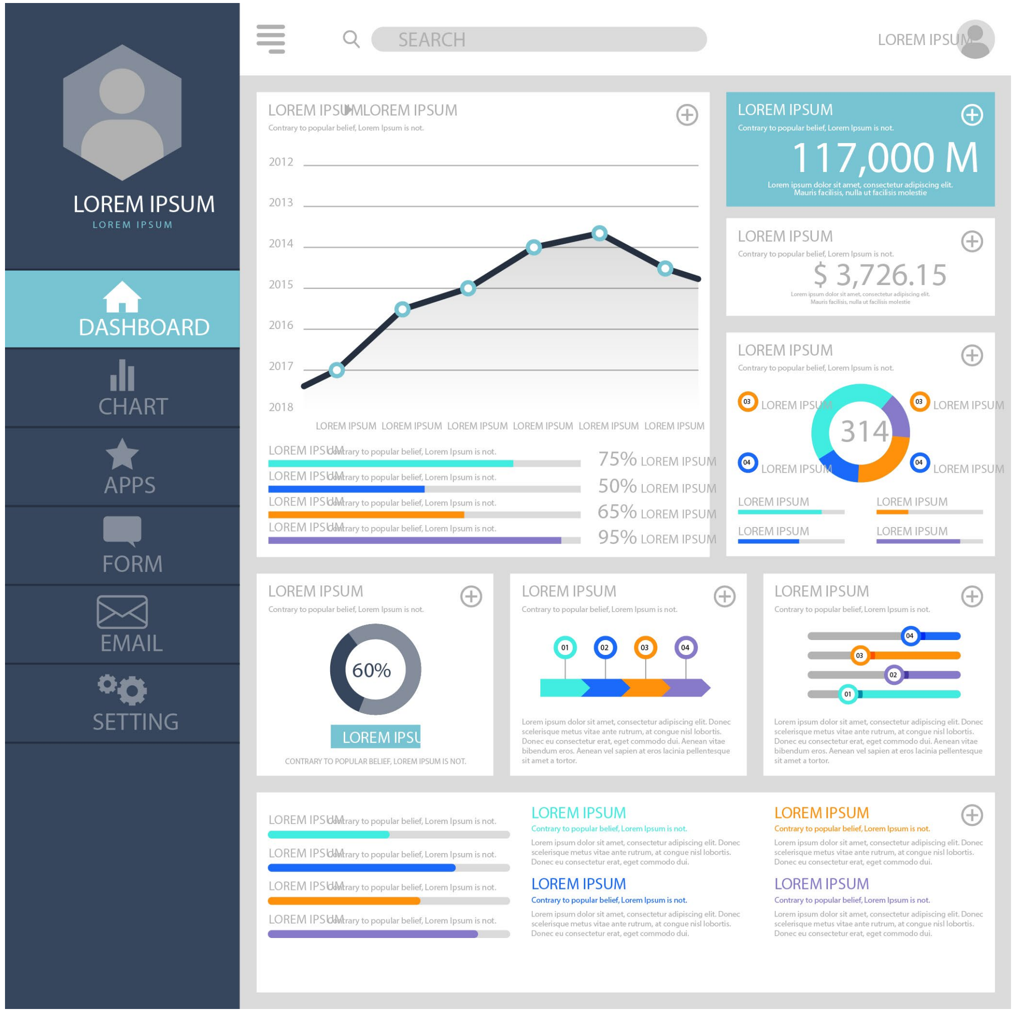
Recognition/prediction Accuracy

Data Science Process

Overview of Data Science

6 Visualization & Communication

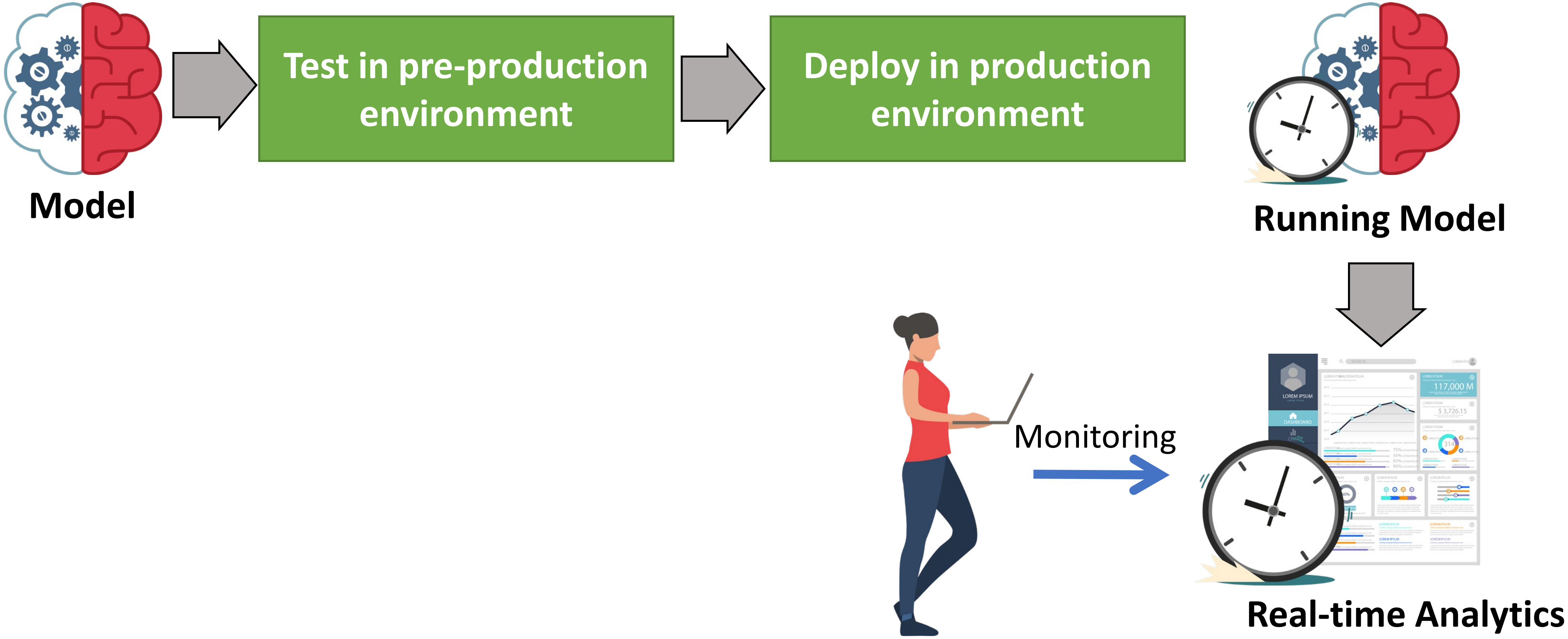
- Create powerful reports and dashboards
- Communicate business finding to convince the stakeholders



Data Science Process

Overview of Data Science

7 Deploy & Maintenance



Model

Test in pre-production environment

Deploy in production environment

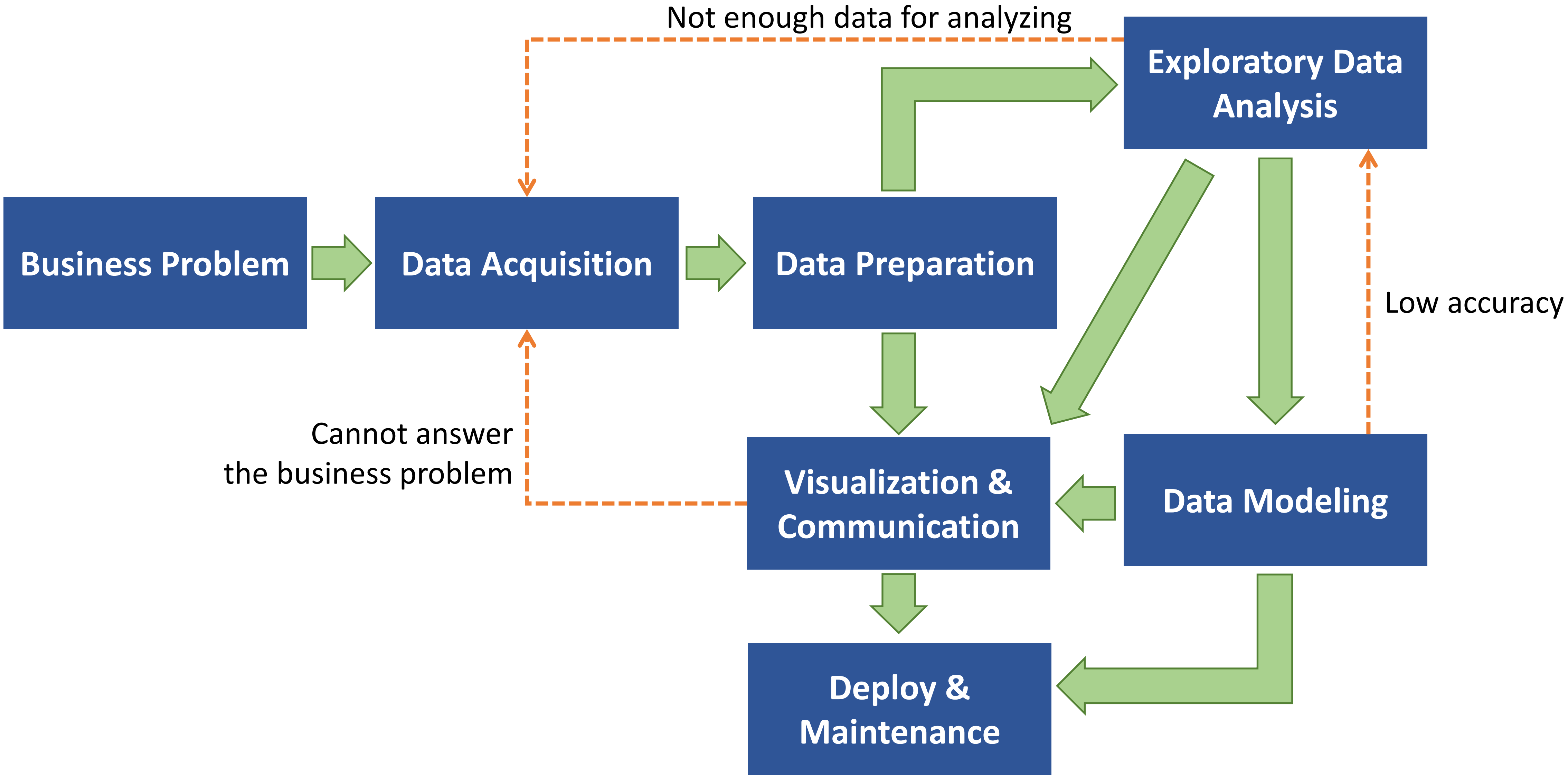
Running Model

Monitoring

Real-time Analytics

Data Science Process

Overview of Data Science



Business Problem

Data Acquisition

Data Preparation

Exploratory Data Analysis

Data Modeling

Visualization & Communication

Deploy & Maintenance

Not enough data for analyzing

Low accuracy

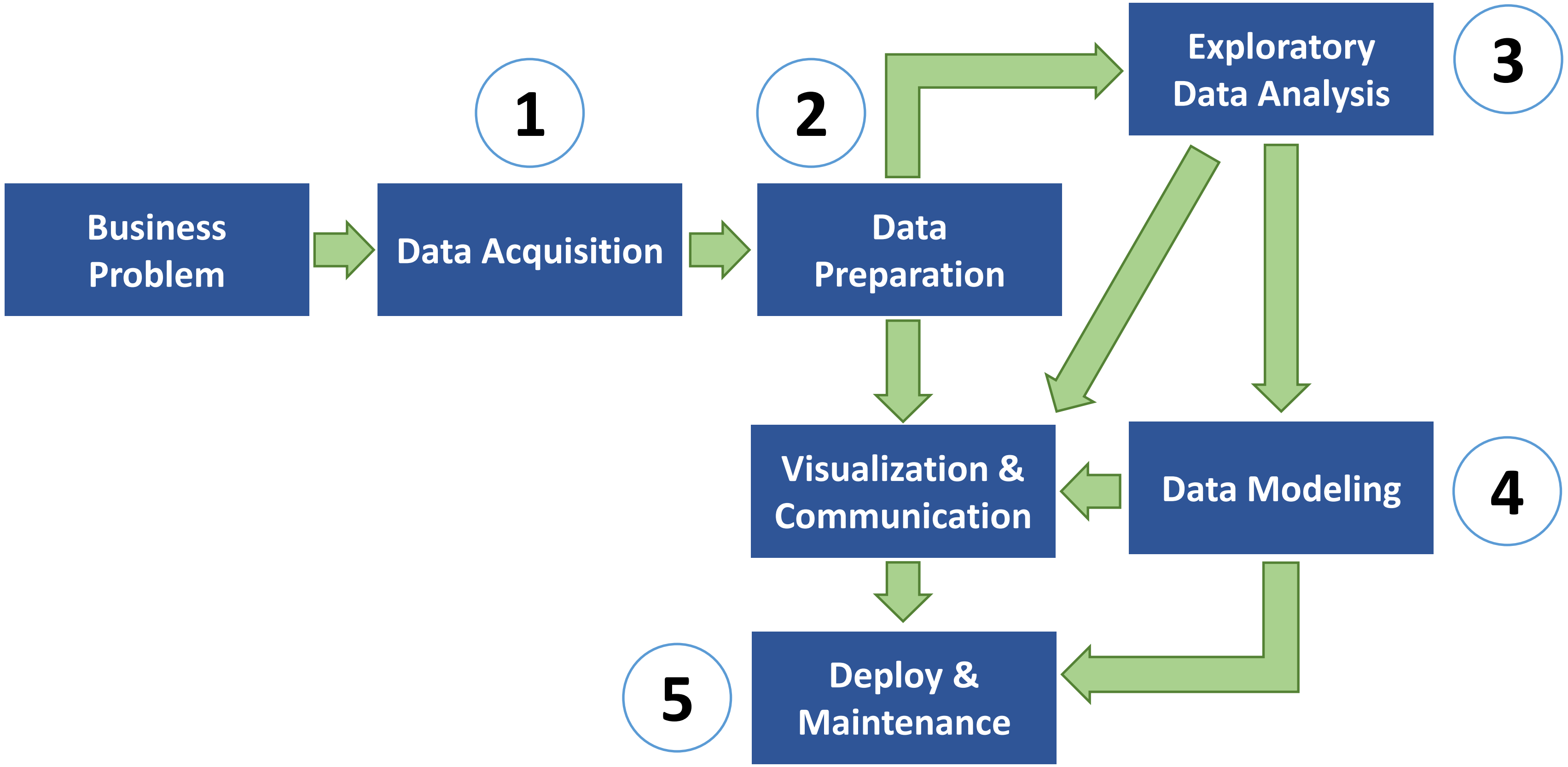
Cannot answer the business problem

Overview of Data Science

A Simple Toy Problem

We want to tech a computer to distinguish pictures of cats and dogs.

How we can perform a data science process to solve the problem?



Overview of Data Science

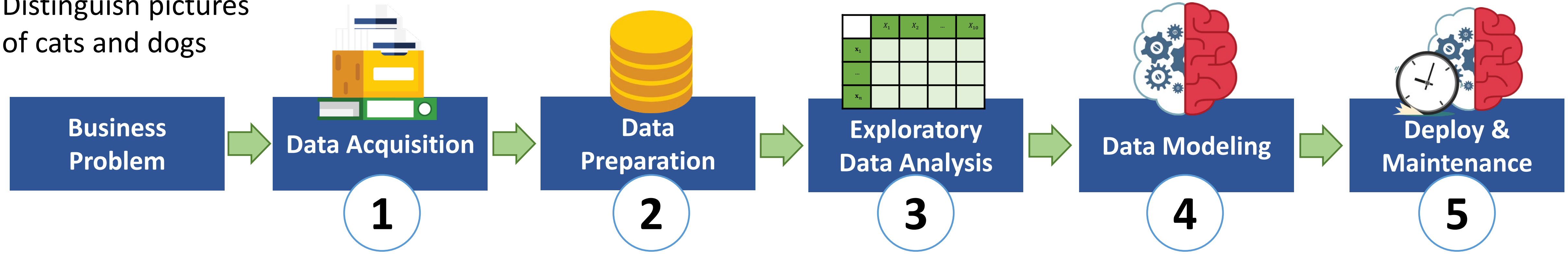
A Simple Toy Problem

We want to tech a computer to distinguish pictures of cats and dogs.

- Extract images that contain only cat or dog
- Scale images to have the same size (Options)
- Store cat's and dog's images in different directories.

Train a classification model and test the model

Distinguish pictures of cats and dogs



Business Problem

1 Data Acquisition

2 Data Preparation

3 Exploratory Data Analysis

4 Data Modeling

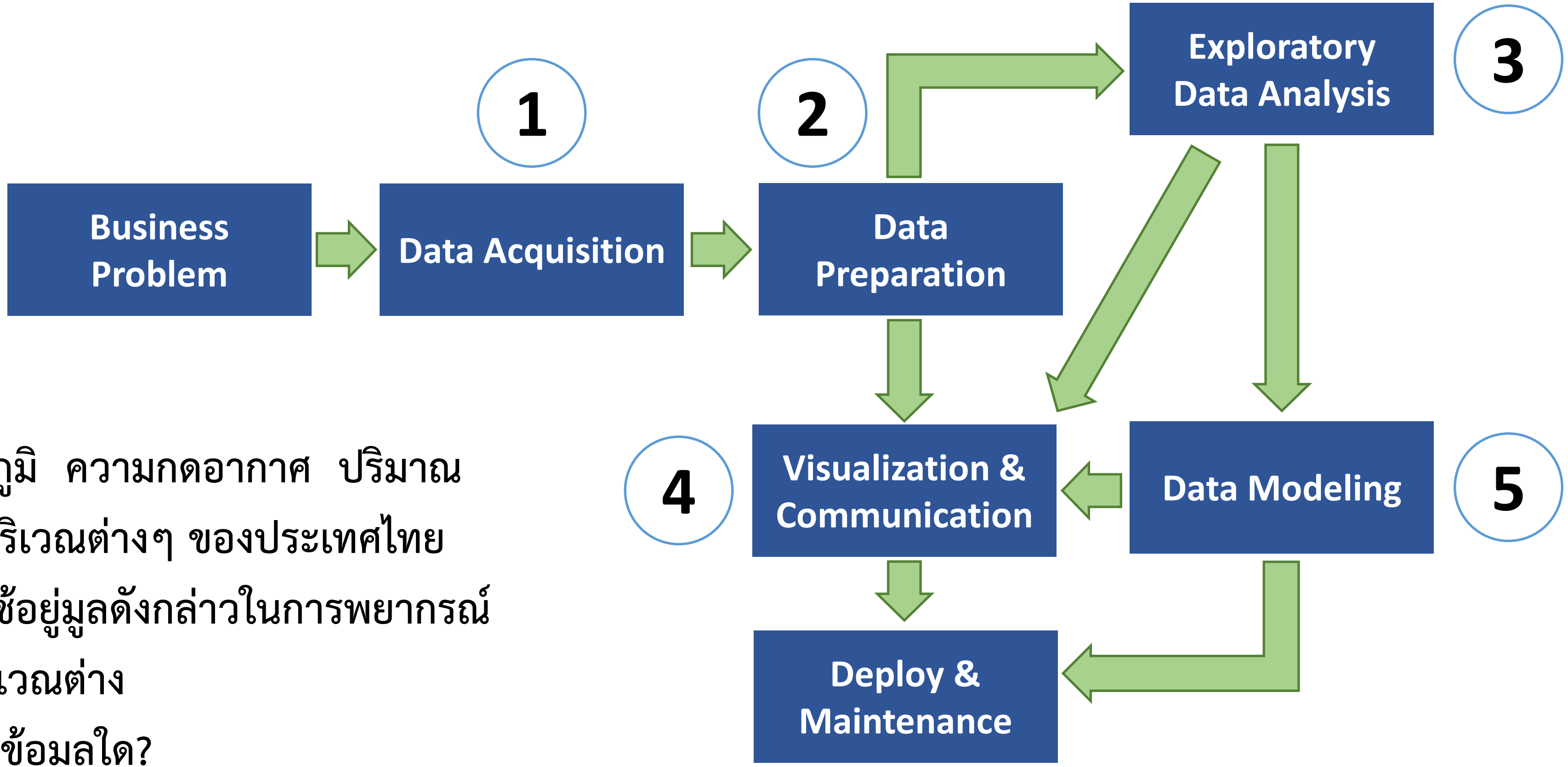
5 Deploy & Maintenance

Collect images of cats and dogs from data sources

Extract information that can be used to distinguish cat from dog

Launch an application

Practice

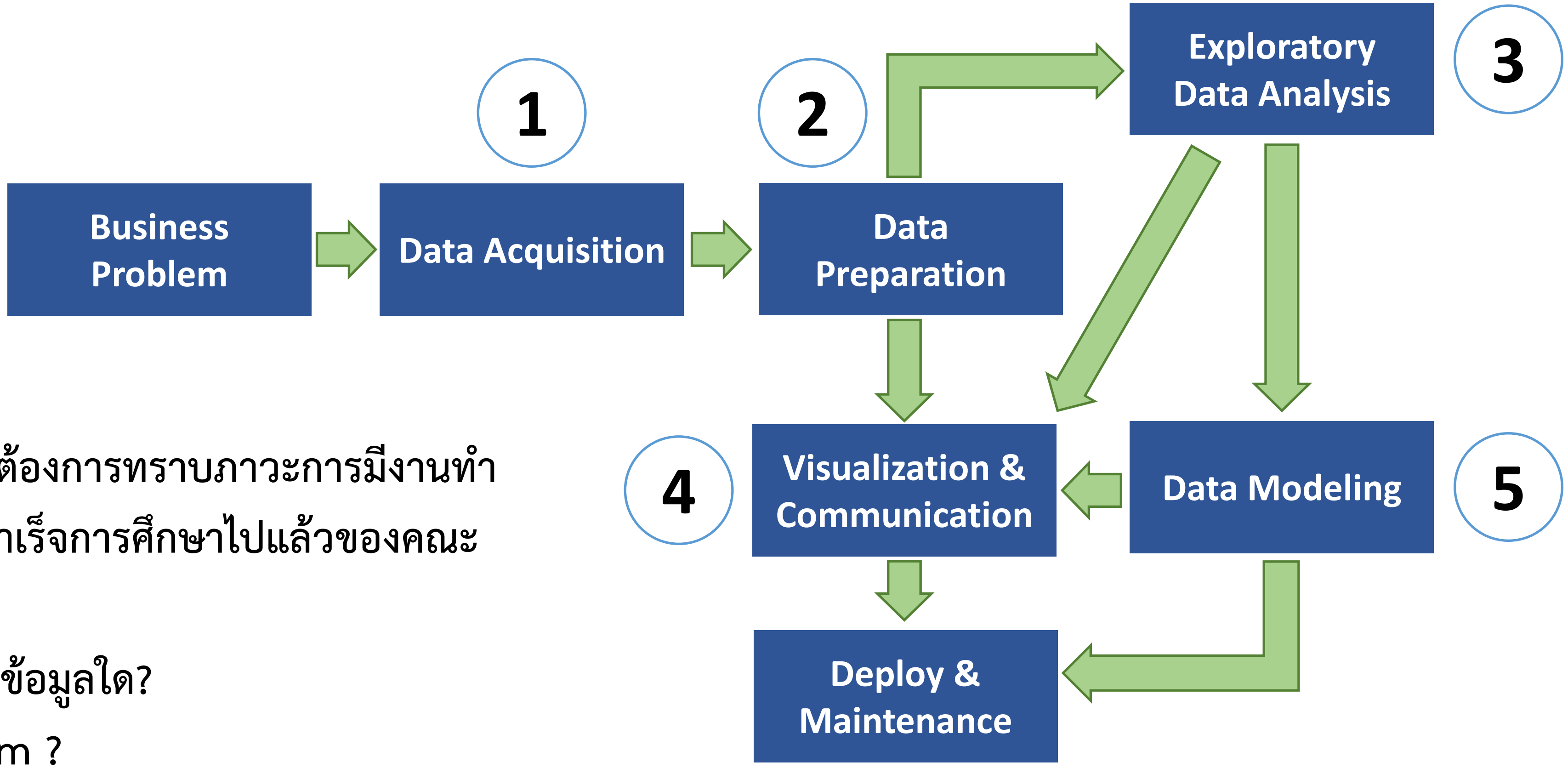


Problem

กรมอุตุนิยมวิทยามีข้อมูลอุณหภูมิ ความกดอากาศ ปริมาณความชื้น และกระแสลม ณ บริเวณต่างๆ ของประเทศไทย เจ้าหน้าที่กรมคนหนึ่งต้องการใช้อยู่มูลดังกล่าวในการพยากรณ์ความเป็นไปได้ที่ฝนจะตกในบริเวณต่าง

- อยู่ในลำดับขั้นการวิเคราะห์ข้อมูลใด?
- อะไรคือ Business Problem ?
- มีขั้นตอนการวิเคราะห์ข้อมูลอย่างไร?

Practice



Problem

ผู้บริหารมหาวิทยาลัยเชียงใหม่ต้องการทราบภาวะการมีงานทำและสายอาชีพของนักศึกษาที่สำเร็จการศึกษาไปแล้วของคณะต่างๆ ในมหาวิทยาลัย

- อยู่ในลำดับขั้นการวิเคราะห์ข้อมูลใด?
- อะไรคือ Business Problem ?
- มีขั้นตอนการวิเคราะห์ข้อมูลอย่างไร?
- ให้ยกตัวอย่างข้อมูลที่จะต้องเก็บรวบรวม (2 ตัวอย่าง)

Further Study

- **Video**: Data Science In 5 Minutes | Data Science For Beginners | What Is Data Science? – Simplilearn https://www.youtube.com/watch?v=X3paOmcrTjQ&t=201s
- **Online lesson**: วิทยาการข้อมูลเบื้องต้น (Introduction to Data Science) – CMU MOOC https://thaimooc.org/courses/course-v1:CMU-MOOC+cmu034+2019_T1/about
- **Books**:
  - Jeremy Watt, Reza Borhani and Aggelos K. Katsaggelos. **Machine Learning Refined: Foundations, Algorithm, and Application**. Cambridge University Press, 2016.