

Introduction to Data Science



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**
- **Data Science Process**
 1. **Business Problem**
 2. **Data Acquisition**
 3. **Data Preparation**
 4. **Exploratory Data Analysis**
 5. **Data Modeling**
 6. **Visualization & Communication**
 7. **Deploy & Maintenance**

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.

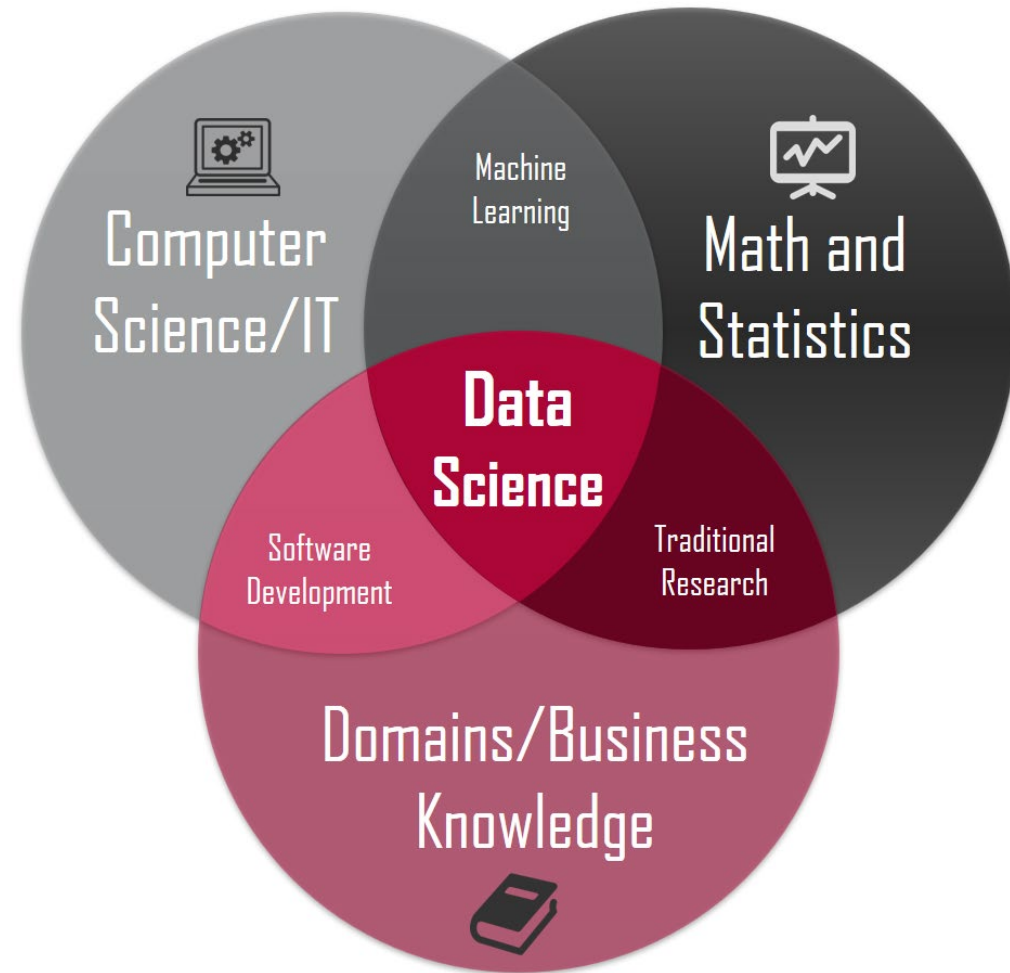
What is the Data Science?

Overview of Data Science

Data Science

is a multi-disciplinary field combining

- Computer Science
- Mathematics and Statistics
- Business Knowledge



Example Applications

Overview of Data Science

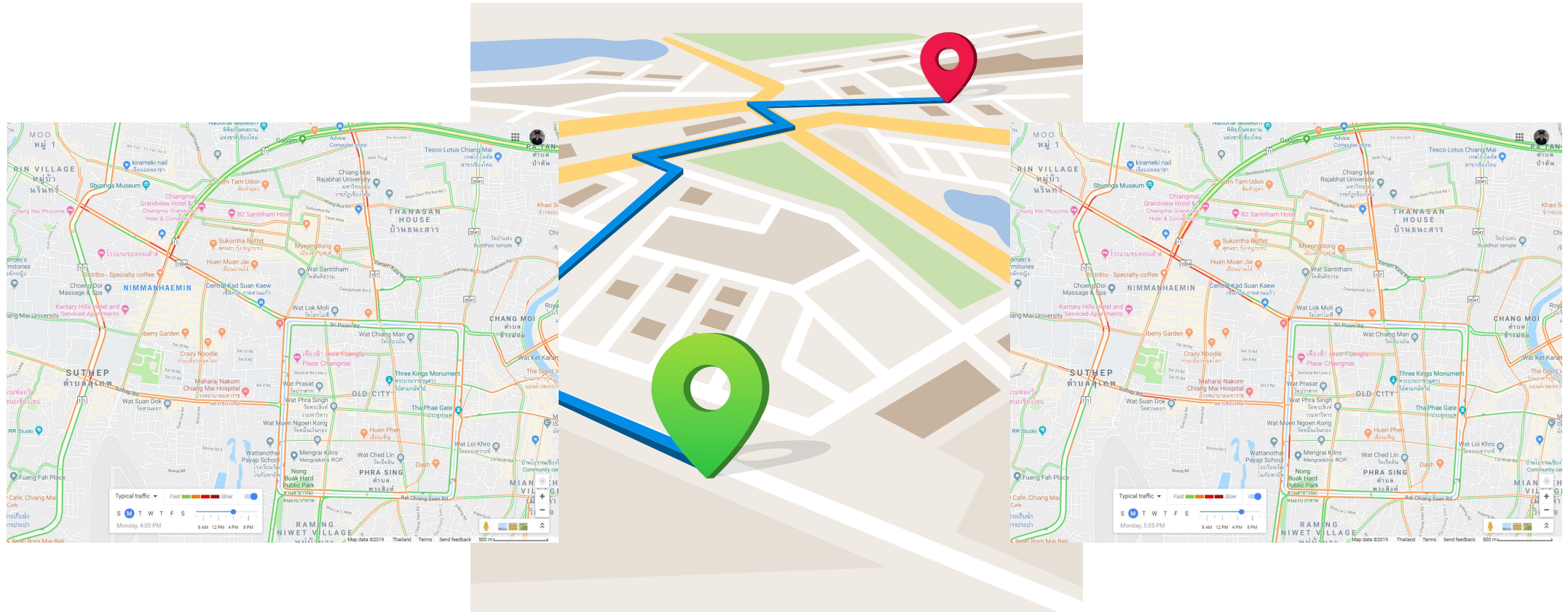
Genomic data provides a deep understanding of genetic issues in reaction to particular drugs and diseases



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

Amazon (online shop) suggests other goods that most people usually bought with what the customer buy.

The screenshot shows the Amazon product page for the book "Data Science (MIT Press Essential Knowledge series) Kindle Edition" by John D. Kelleher and Brendan Tierney. The page features a book cover, a price of \$9.87 for the Kindle edition, and a 4.5-star rating from 15 reviews. Below the main product information, there is a section titled "Customers who bought this item also bought" which displays seven related books from the MIT Press Essential Knowledge series, including "Information and Society", "Machine Learning: The New AI", "Metadata", "The Internet of Things", "Cloud Computing", "Computational Thinking", and "Computing: A Concise History".

Data Science (MIT Press Essential Knowledge series) Kindle Edition
by John D. Kelleher (Author), Brendan Tierney (Author)
★★★★☆ 15 customer reviews

Kindle \$9.87 Audiobook \$0.00 Paperback \$12.96 Audio CD \$27.25

Read with Our Free App Free with your Audible trial 34 Used from \$10.24 74 New from \$4.99 3 Used from \$23.45 7 New from \$18.76

A concise introduction to the emerging field of data science, explaining its evolution, relation to machine learning, current uses, data infrastructure issues, and ethical challenges.

The goal of data science is to improve decision making through the analysis of data. Today data science determines the ads we see online, the books and movies that are recommended to us online, which emails are filtered into our spam folders, and even how much we pay for health insurance. This volume

Length: 282 pages Audible book: Available Similar books to Data Science (MIT Press Essential Knowledge series)

Follow the Authors

- John D. Kelleher + Follow
- Brendan Tierney + Follow

Customers who bought this item also bought

Book Title	Author	Rating	Price
Information and Society (MIT Press Essential Knowledge series)	Michael Buckland	★★★★☆ 2	Kindle Edition
Machine Learning: The New AI (MIT Press Essential Knowledge series)	Ethem Alpaydin	★★★★☆ 41	Kindle Edition
Metadata (MIT Press Essential Knowledge series)	Jeffrey Pomerantz	★★★★☆ 6	Kindle Edition
The Internet of Things (MIT Press Essential Knowledge series)	Samuel Greengard	★★★★☆ 28	Kindle Edition
Cloud Computing (MIT Press Essential Knowledge series)	Nayan B. Ruparelia	★★★★☆ 8	Kindle Edition
Computational Thinking (MIT Press Essential Knowledge series)	Peter J. Denning	★★★★☆ 23	Kindle Edition \$10.85
Computing: A Concise History (MIT Press Essential Knowledge series)	Paul E. Ceruzzi	★★★★☆ 23	Kindle Edition

Data Science Process

Overview of Data Science

1 Business Problem

Business Questions



Understand

Define Objectives



common data science problems:

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

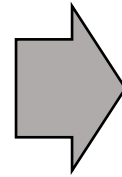
Data Science Process

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

?

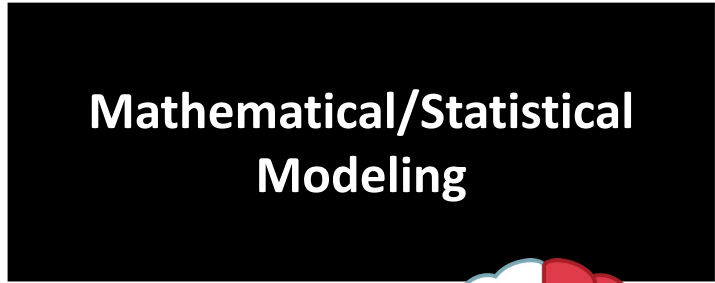
Data Science Process

Overview of Data Science

5 Data Modeling



Training Set



Association

- Itemset mining
- Summarizing Itemsets

Recognition

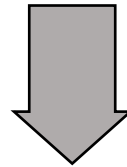
- Decision tree
- kNN
- SVM

Regression

- Linear regression
- Polynomial regression

Clustering

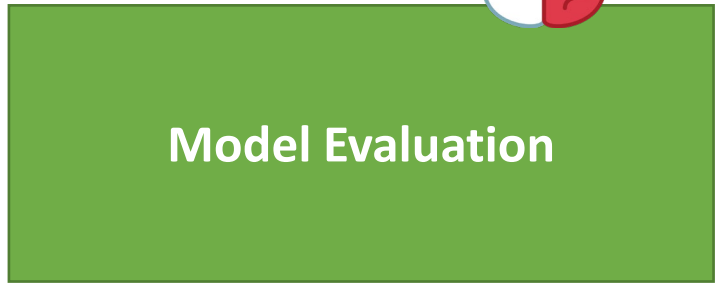
- K-mean
- Hierarchical clustering
- DBSCAN



Model



Test Set



Recognition/prediction Accuracy

Data Science Process

Overview of Data Science

6 Visualization & Communication

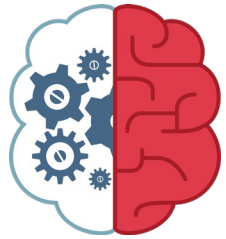
- Create powerful reports and dashboards
- Communicate business finding to convince the stakeholders



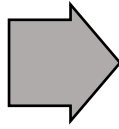
Data Science Process

Overview of Data Science

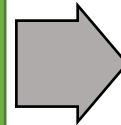
7 Deploy & Maintenance



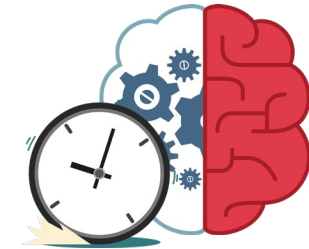
Model



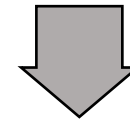
Test in pre-production environment



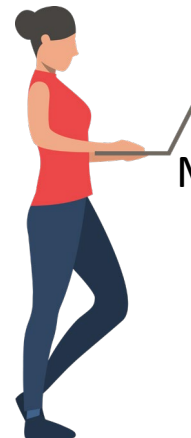
Deploy in production environment



Running Model



Real-time Analytics

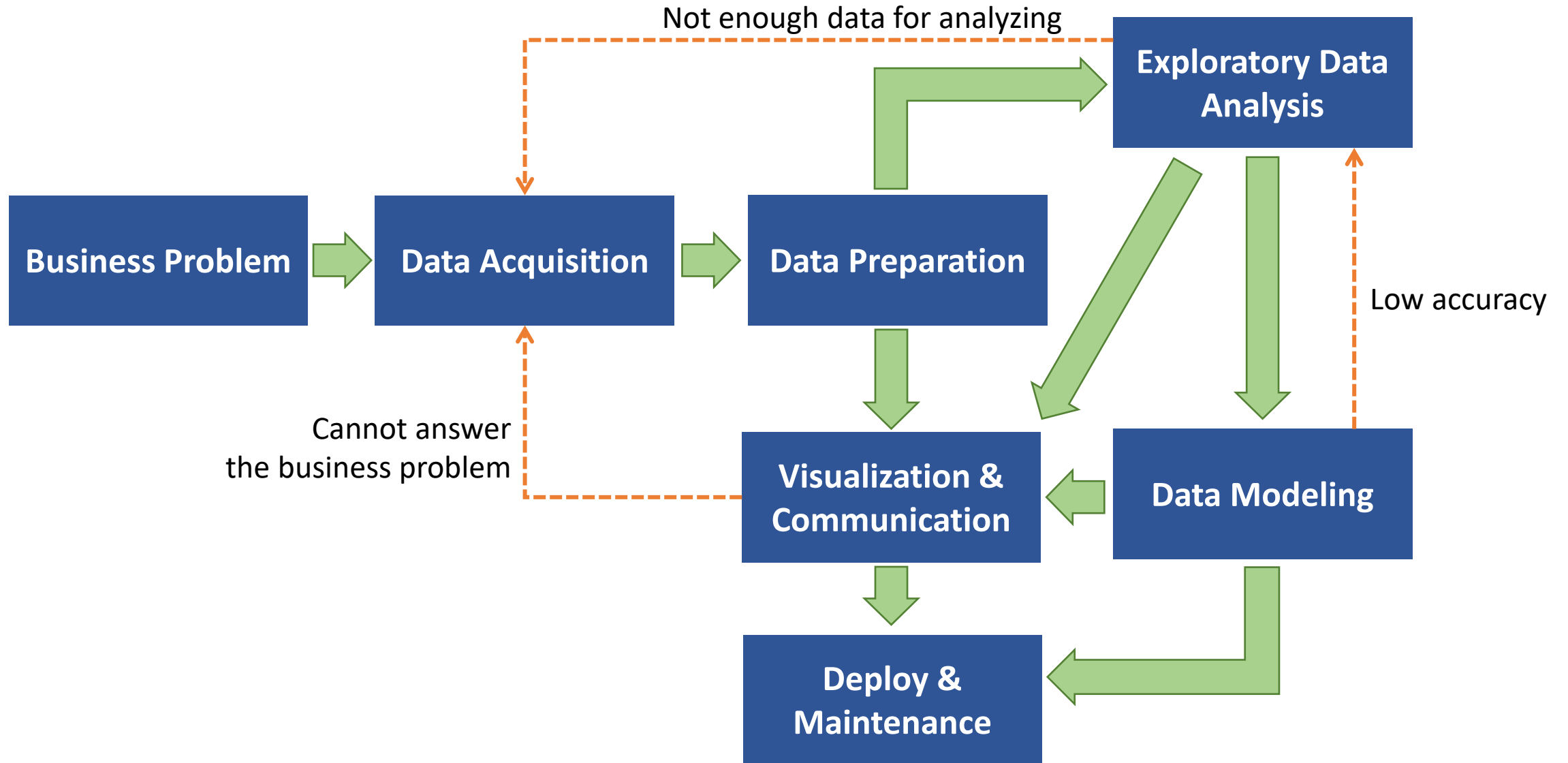


Monitoring



Data Science Process

Overview of Data Science

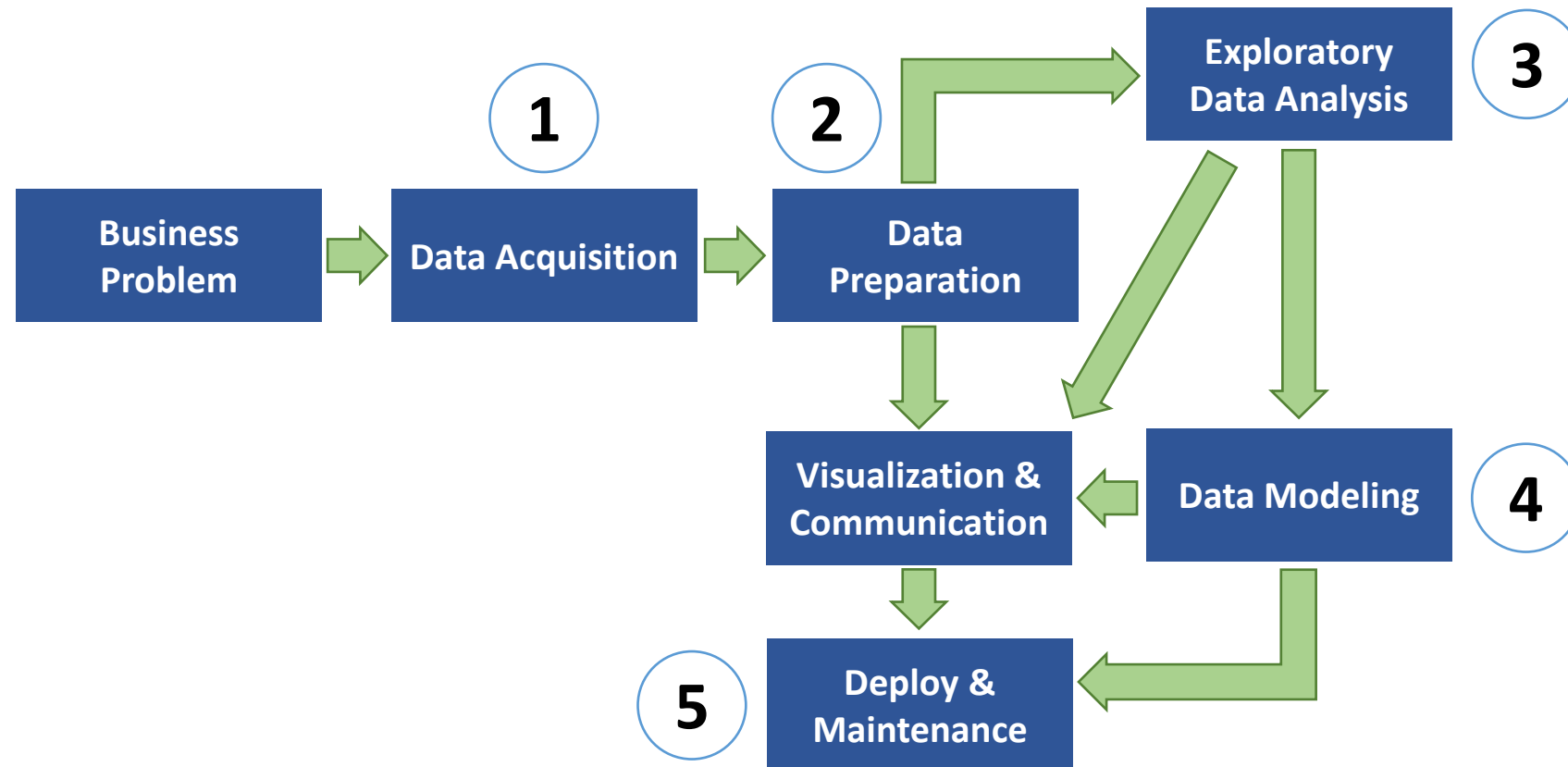


Overview of Data Science

A Simple Toy Problem

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

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



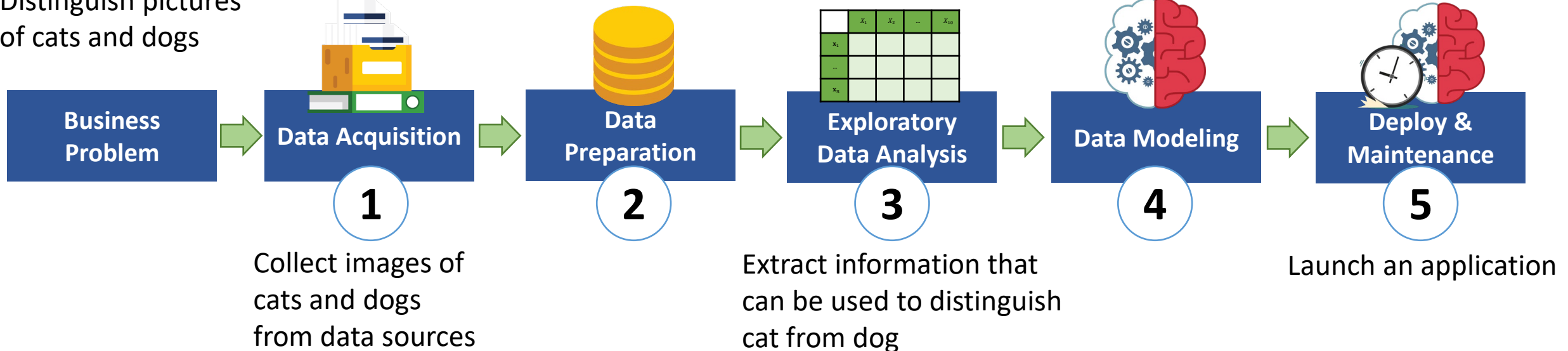
Overview of Data Science

A Simple Toy Problem

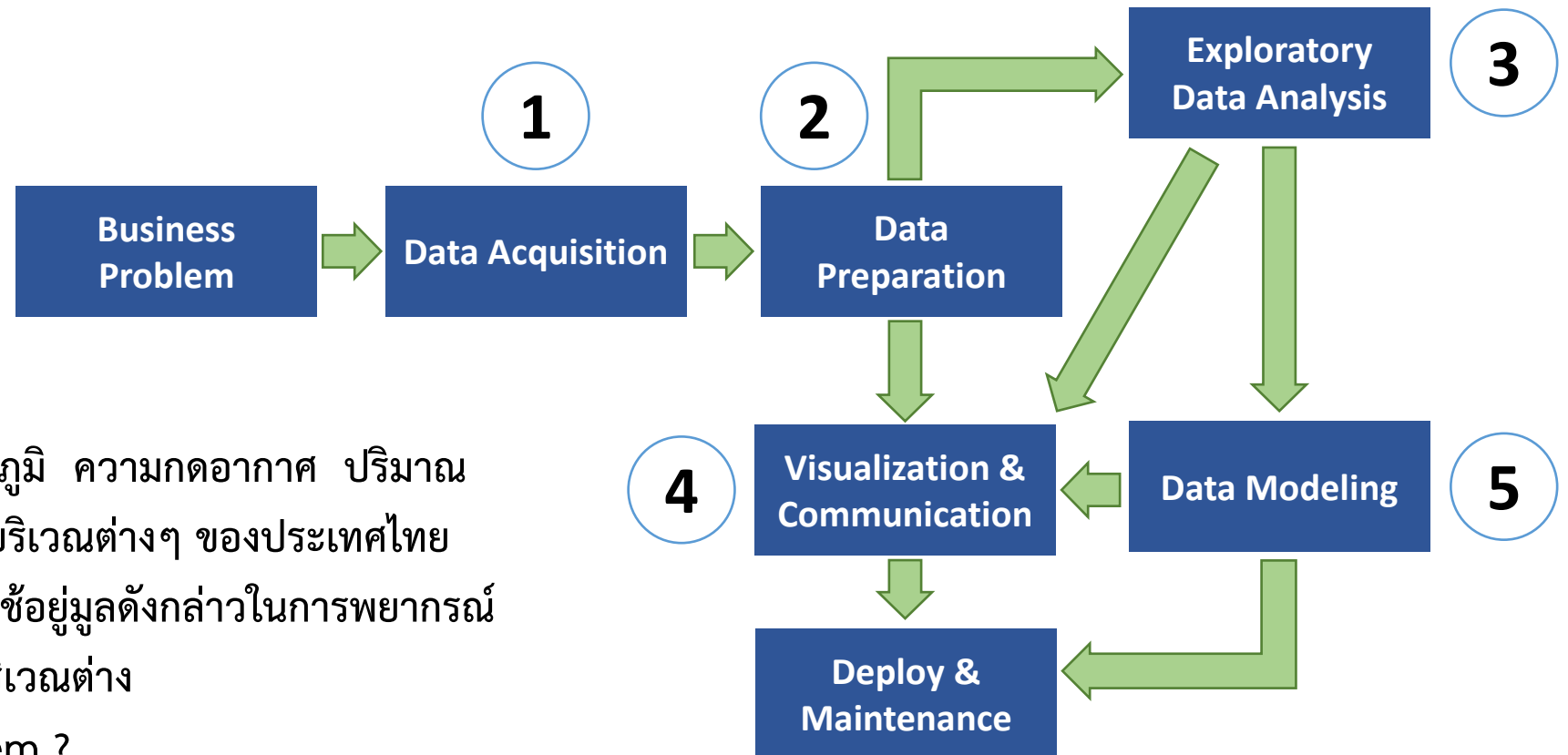
We want to teach a computer to distinguish pictures of cats and dogs.

- 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.

Distinguish pictures of cats and dogs



Practice

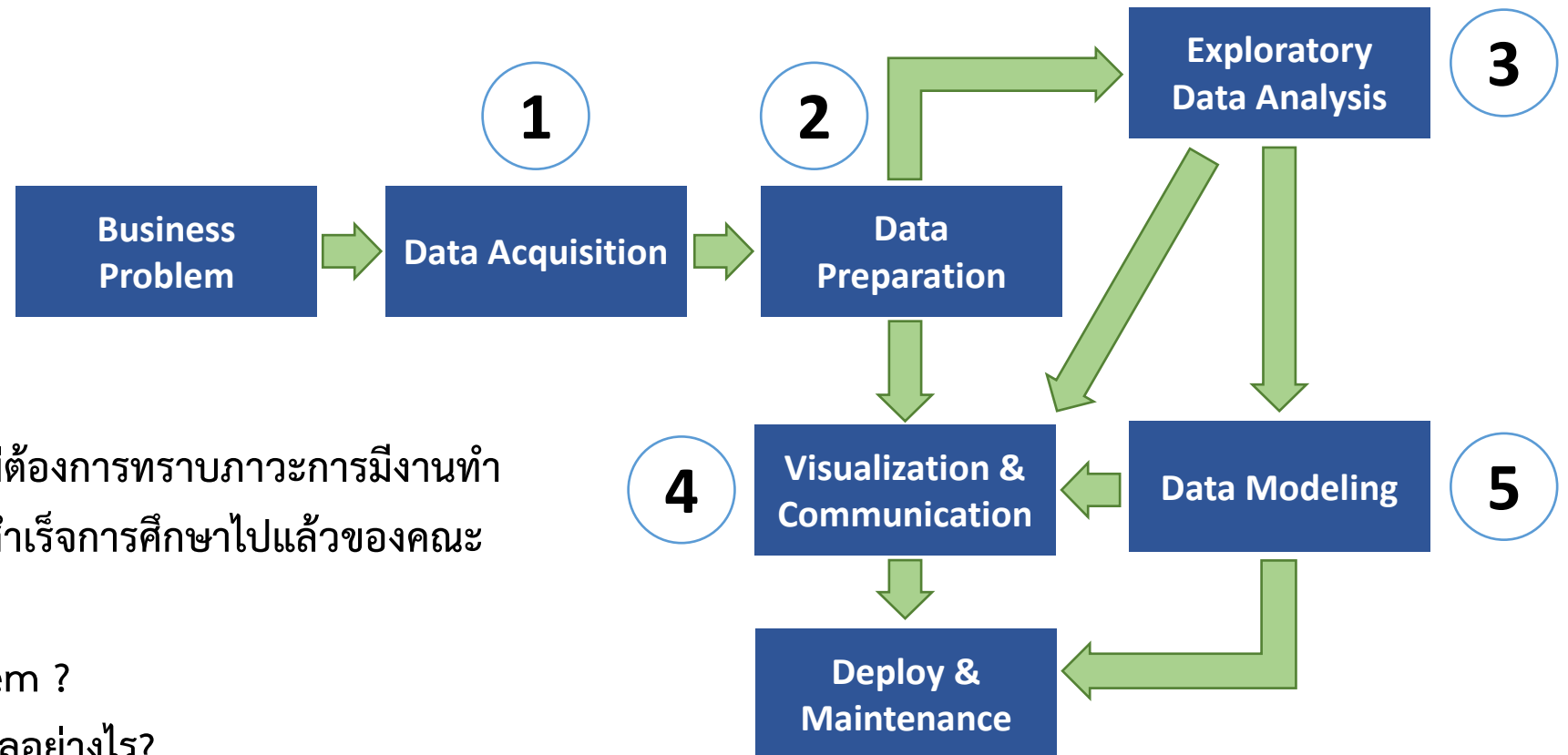


Problem

กรมอุตุนิยมวิทยามีข้อมูลอุณหภูมิ ความกดอากาศ ปริมาณ ความชื้น และกระแสลม ณ บริเวณต่างๆ ของประเทศไทย เจ้าหน้าที่กรมคนหนึ่งต้องการใช้อยู่มุดังกล่าวในการพยากรณ์ ความเป็นไปได้ที่ฝนจะตกในบริเวณต่าง

- อะไรคือ Business Problem ?
- มีขั้นตอนการวิเคราะห์ข้อมูลอย่างไร?

Practice



Problem

ผู้บริหารมหาวิทยาลัยเชียงใหม่ต้องการทราบภาวะการปฏิบัติงานทำและสายอาชีพของนักศึกษาที่สำเร็จการศึกษาไปแล้วของคณะต่างๆ ในมหาวิทยาลัย

- 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.