

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**
- **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

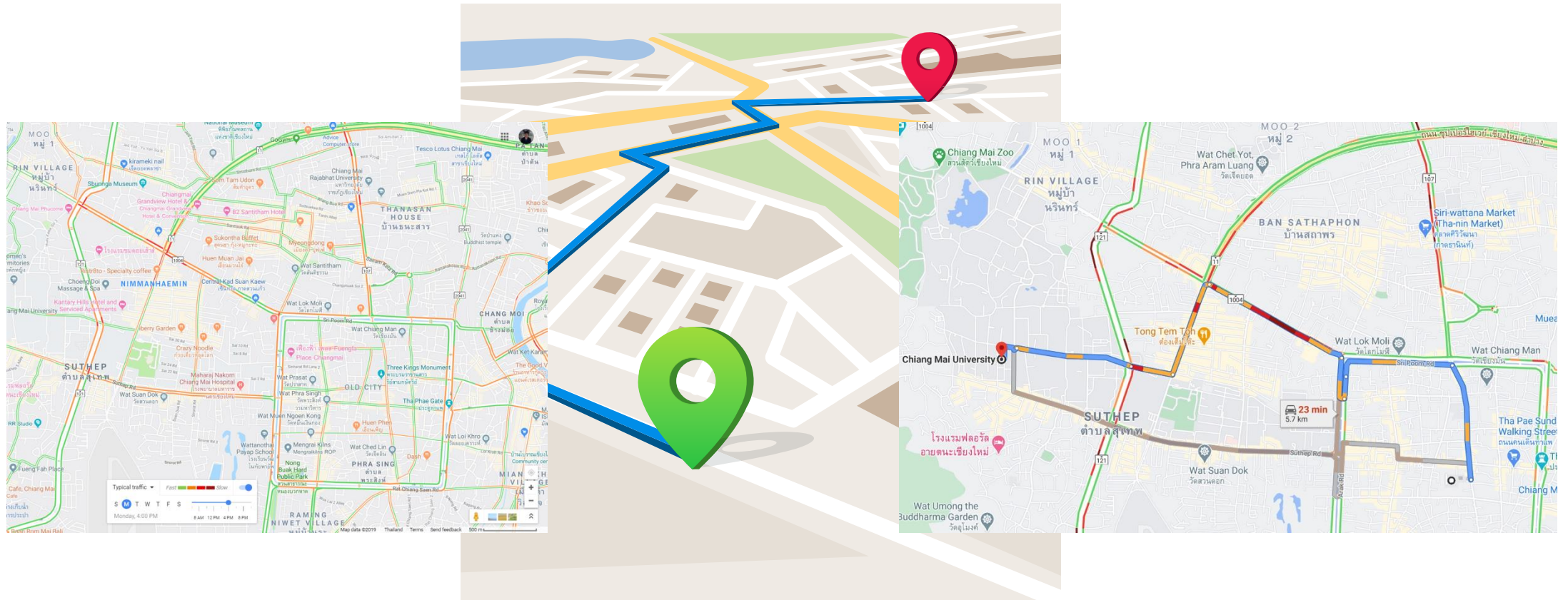
Data Science

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

Example Applications

Overview of Data Science

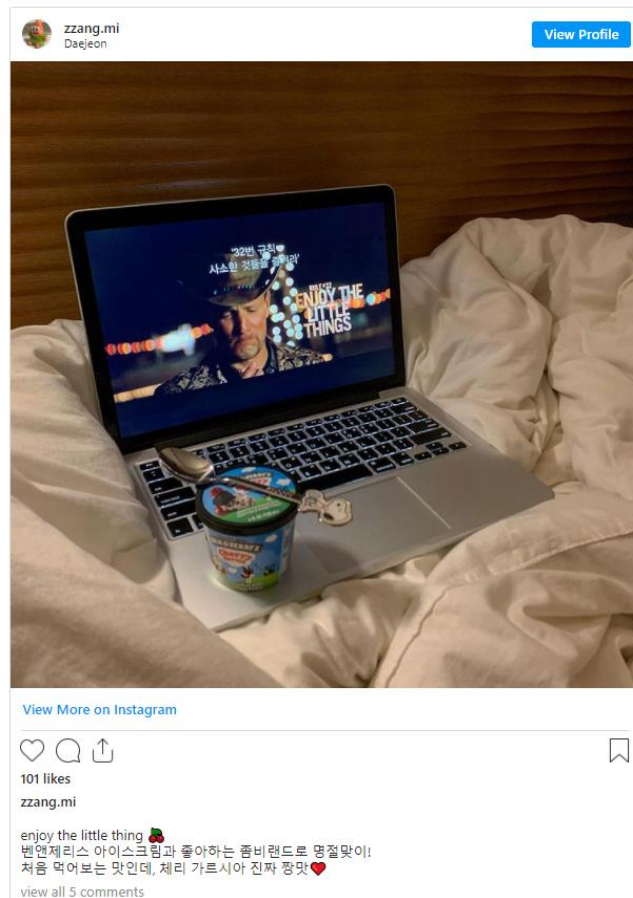
We use traffic data to discover the best time and route to go back to home.



Example Applications

Overview of Data Science

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



<https://twitter.com/benandjerrys>

Example Applications

Overview of Data Science

Large Language Models



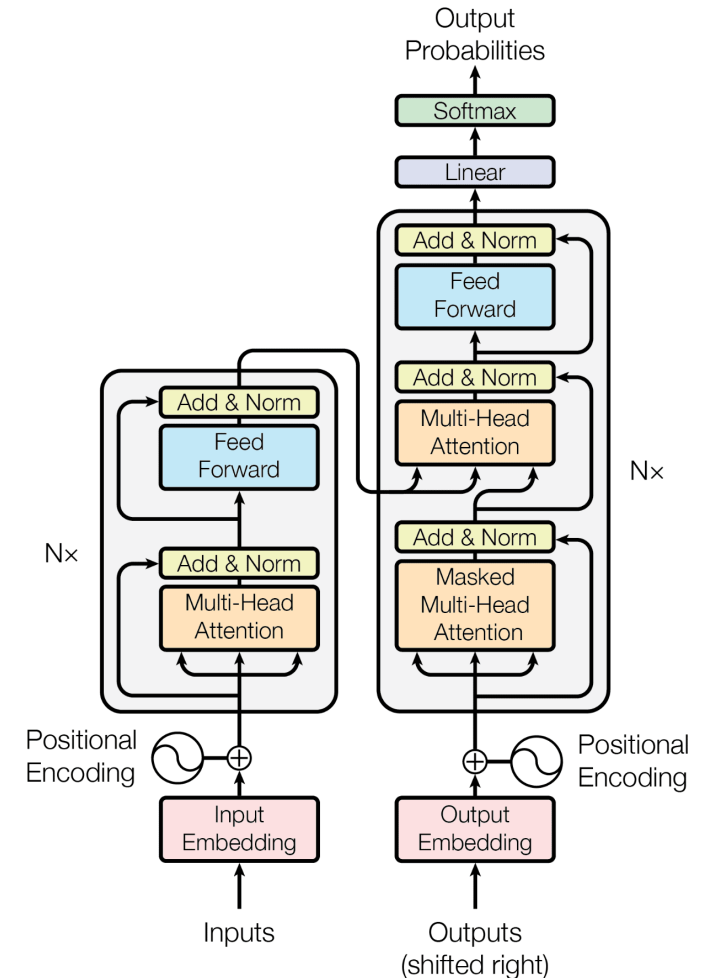
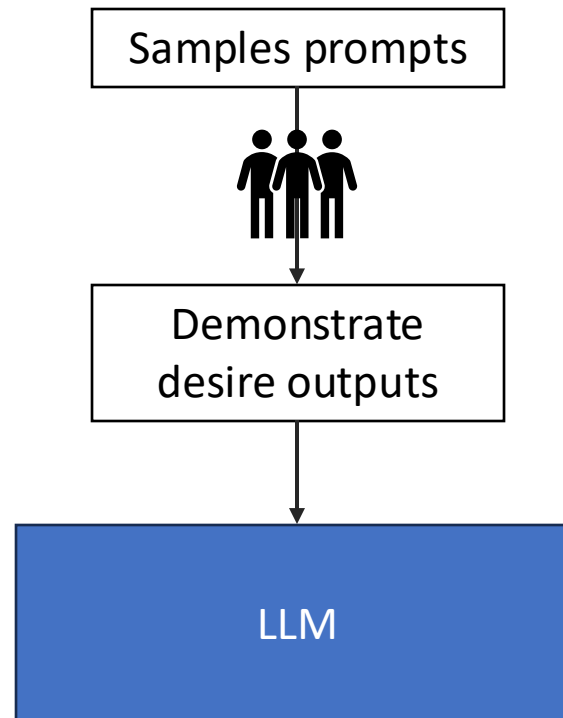
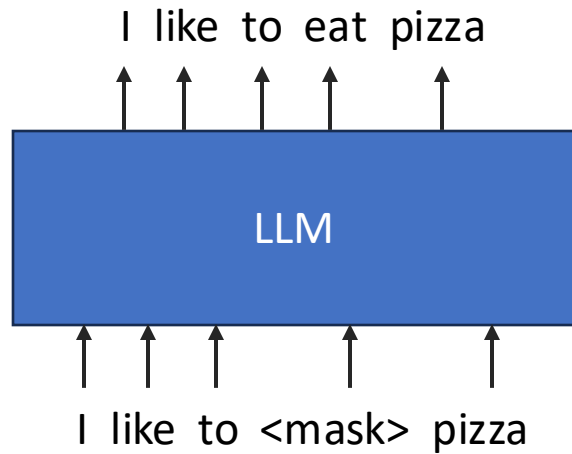
deepseek



Gemini



Claude



The encoder-decoder structure of the Transformer architecture

Vaswani, A., "Attention Is All You Need", arXiv e-prints, Art. no. arXiv:1706.03762, 2017. doi:10.48550/arXiv.1706.03762.

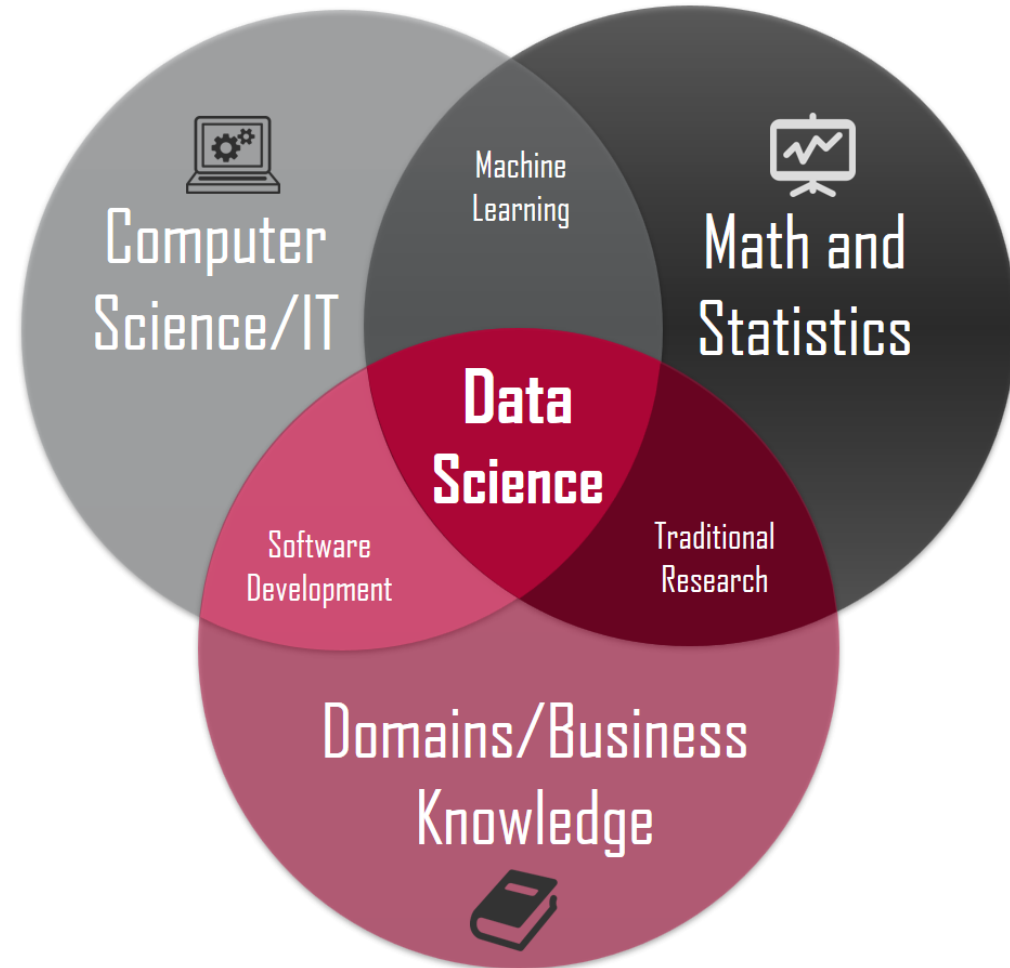
Related Study Fields

Overview of Data Science

Data Science

is a multi-disciplinary field combining

- Computer Science
- Mathematics and Statistics
- Business Knowledge



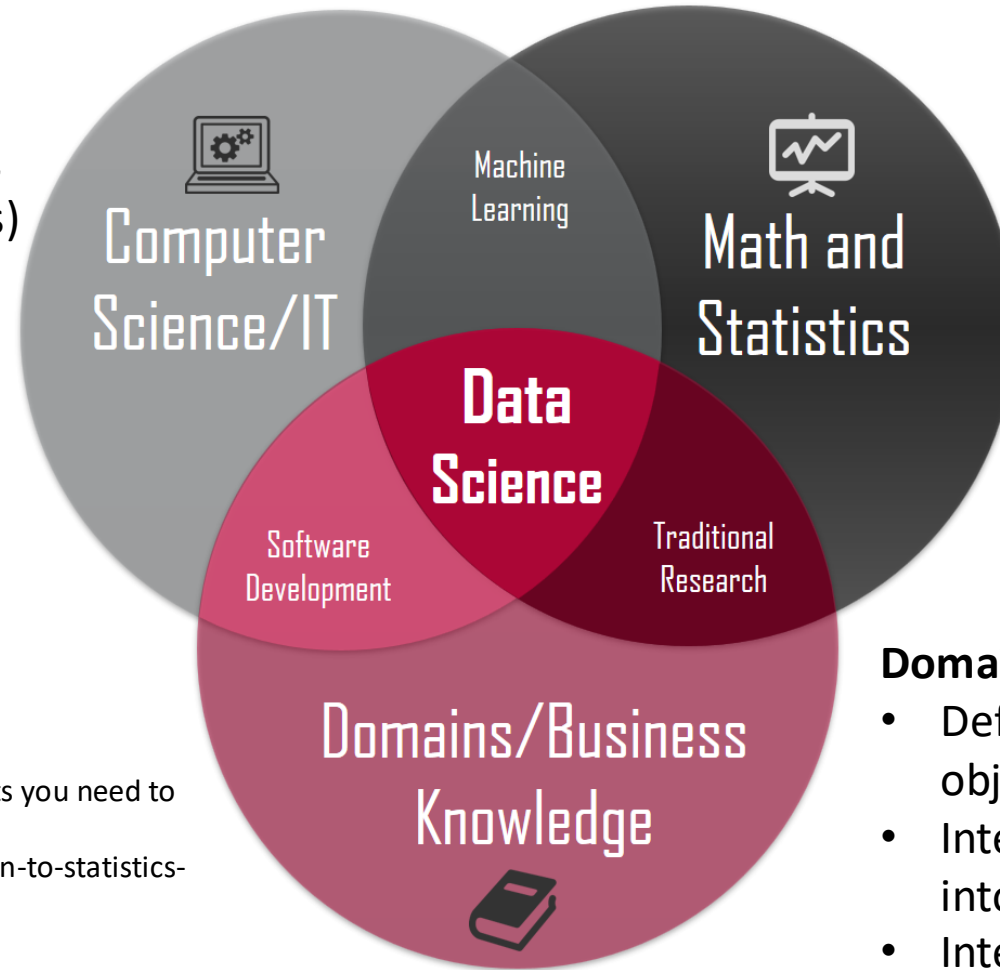
Michael Barber (2018). Data science concepts you need to know! Part 1. Retrieved from <https://towardsdatascience.com/introduction-to-statistics-e9d72d818745>

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)

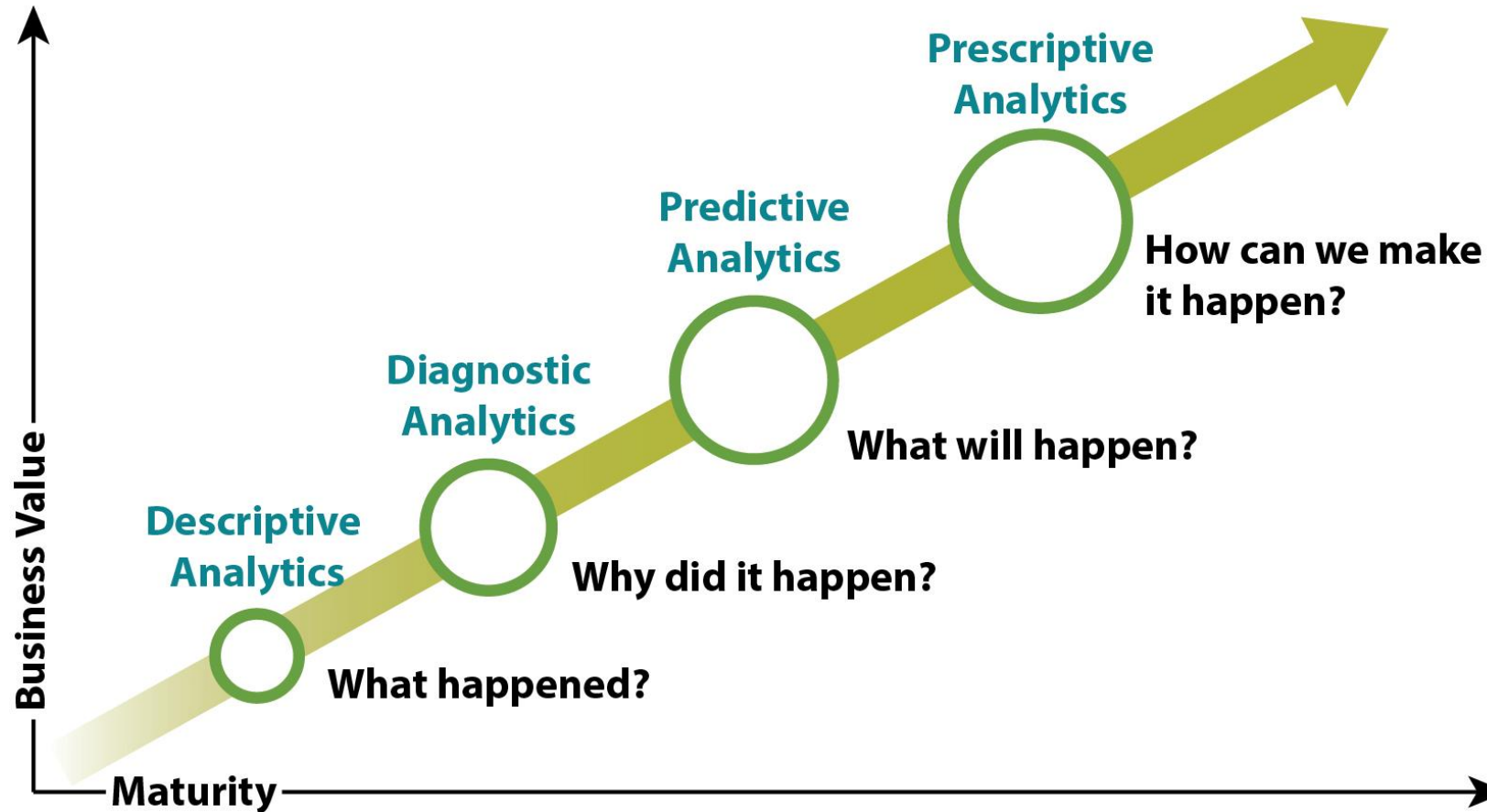
Domain Expertise

- 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?**”

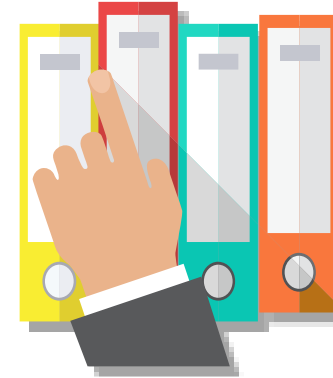
Levels of Data Analytics Maturity

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?**”



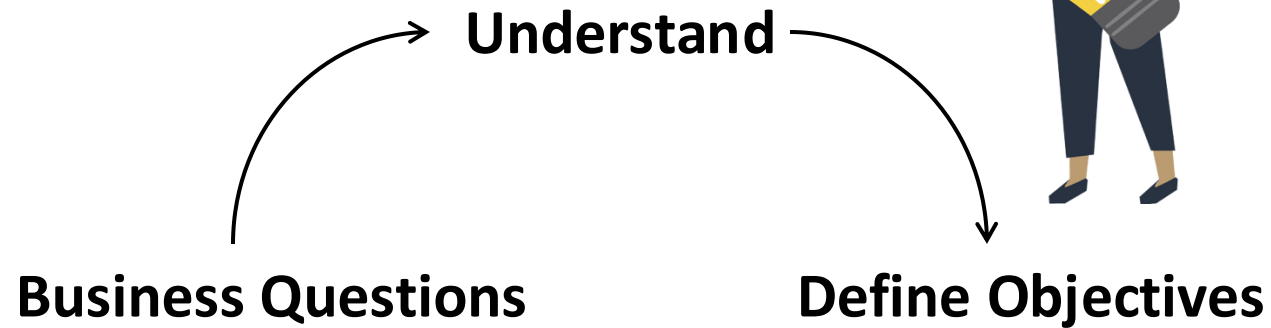
Prescriptive Analytics

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

Data Science Process

Overview of Data Science

1 Business Problem



common data science problems:

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

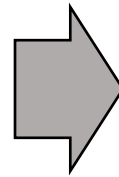
Data Science Process

Overview of Data Science

2 Data Acquisition

Data Sources

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



Data Science Process

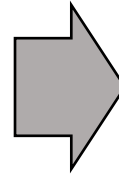
Overview of Data Science

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



Mathematical/Statistical Modeling

Association

- Itemset mining
- Summarizing Itemsets

Recognition

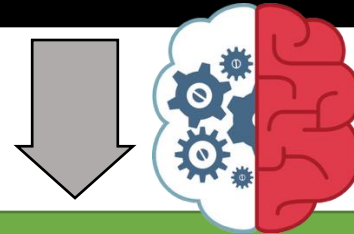
- Decision tree
- kNN
- SVM

Regression

- Linear regression
- Polynomial regression

Clustering

- K-mean
- Hierarchical clustering
- DBSCAN



Model

Model Evaluation

Recognition/prediction Accuracy

Data Science Process

Overview of Data Science

6 Visualization & Communication

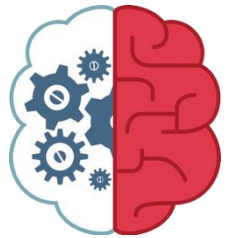
- Create powerful reports and dashboards
- Communicate business finding to convince the stakeholders



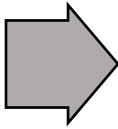
Data Science Process

Overview of Data Science

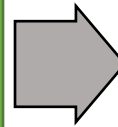
7 Deploy & Maintenance



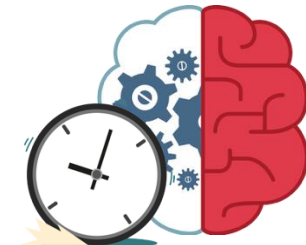
Model



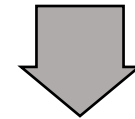
Test in pre-production
environment



Deploy in production
environment



Running Model



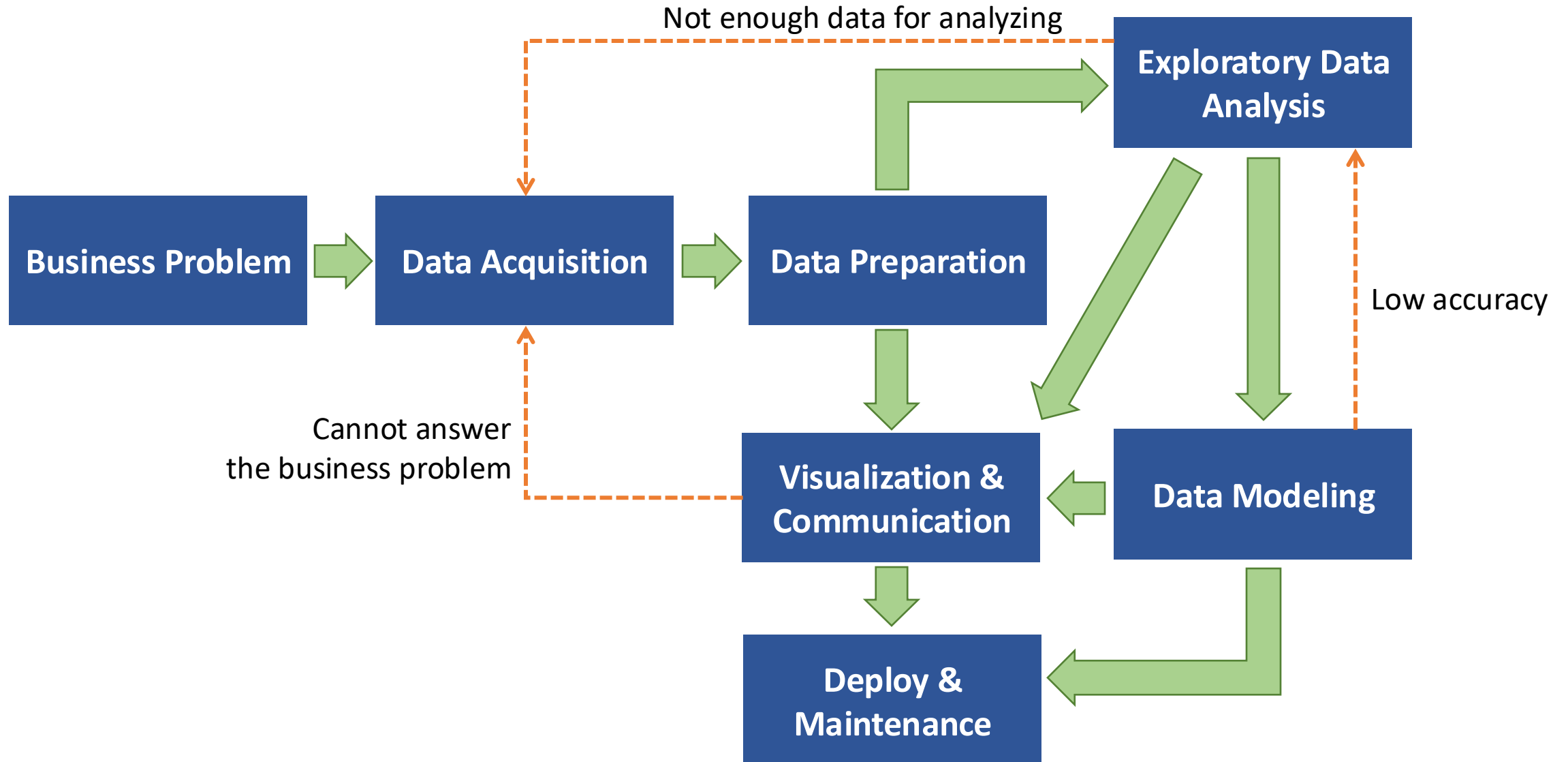
Monitoring



Real-time Analytics

Data Science Process

Overview of Data Science

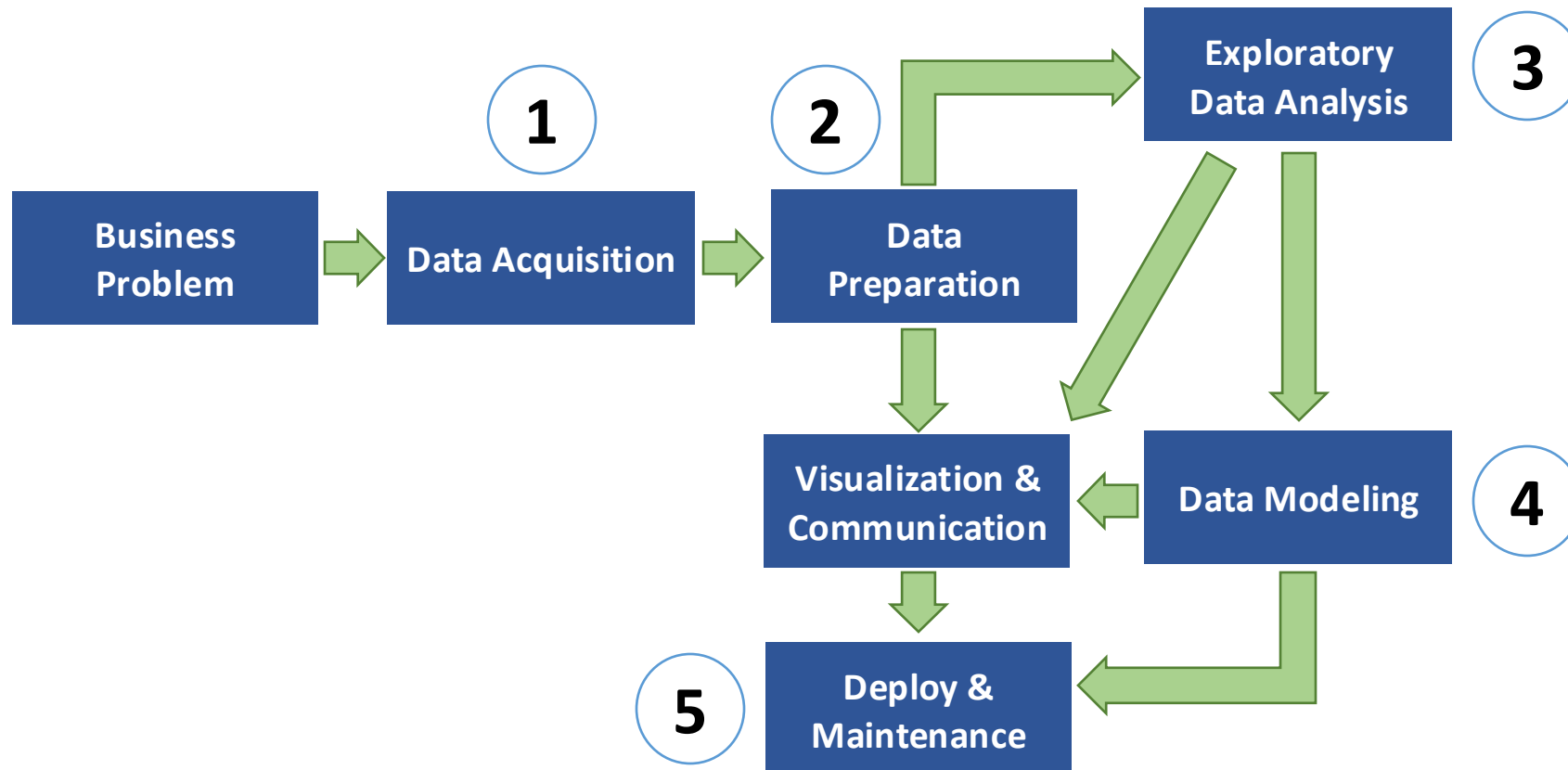


Overview of Data Science

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?



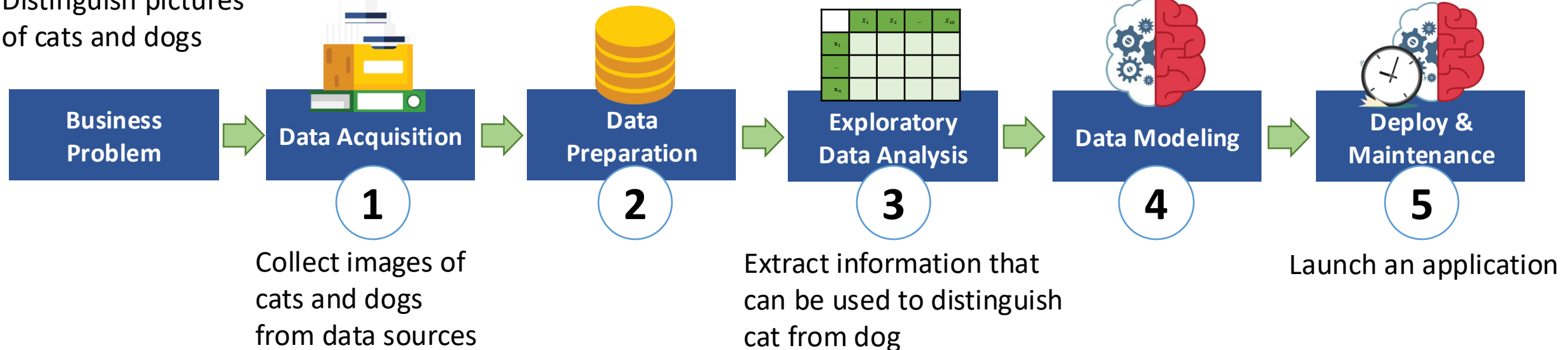
Overview of Data Science

A Simple Toy Problem

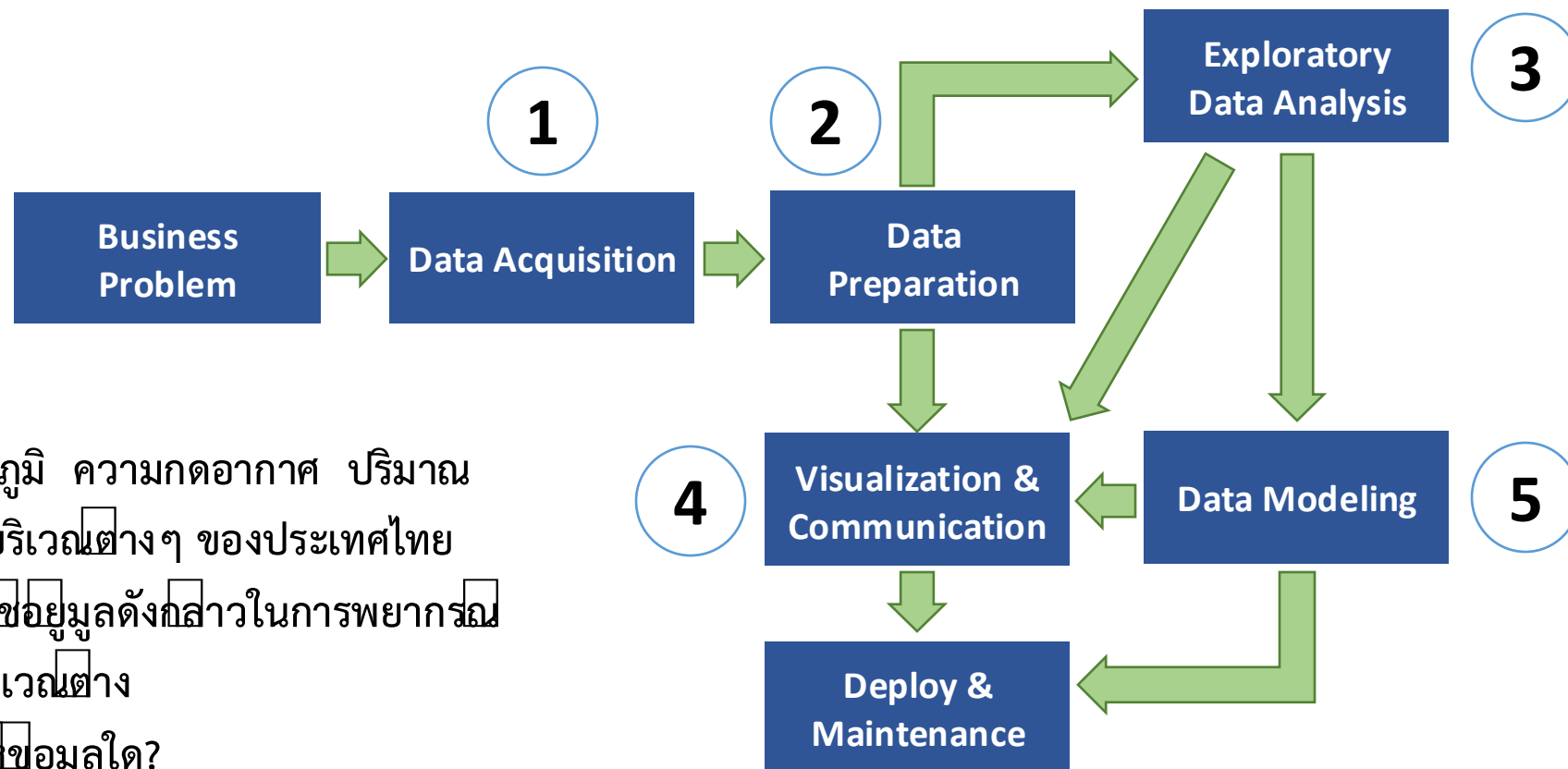
We want to teach a computer to distinguish pictures of cats and dogs.

- 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

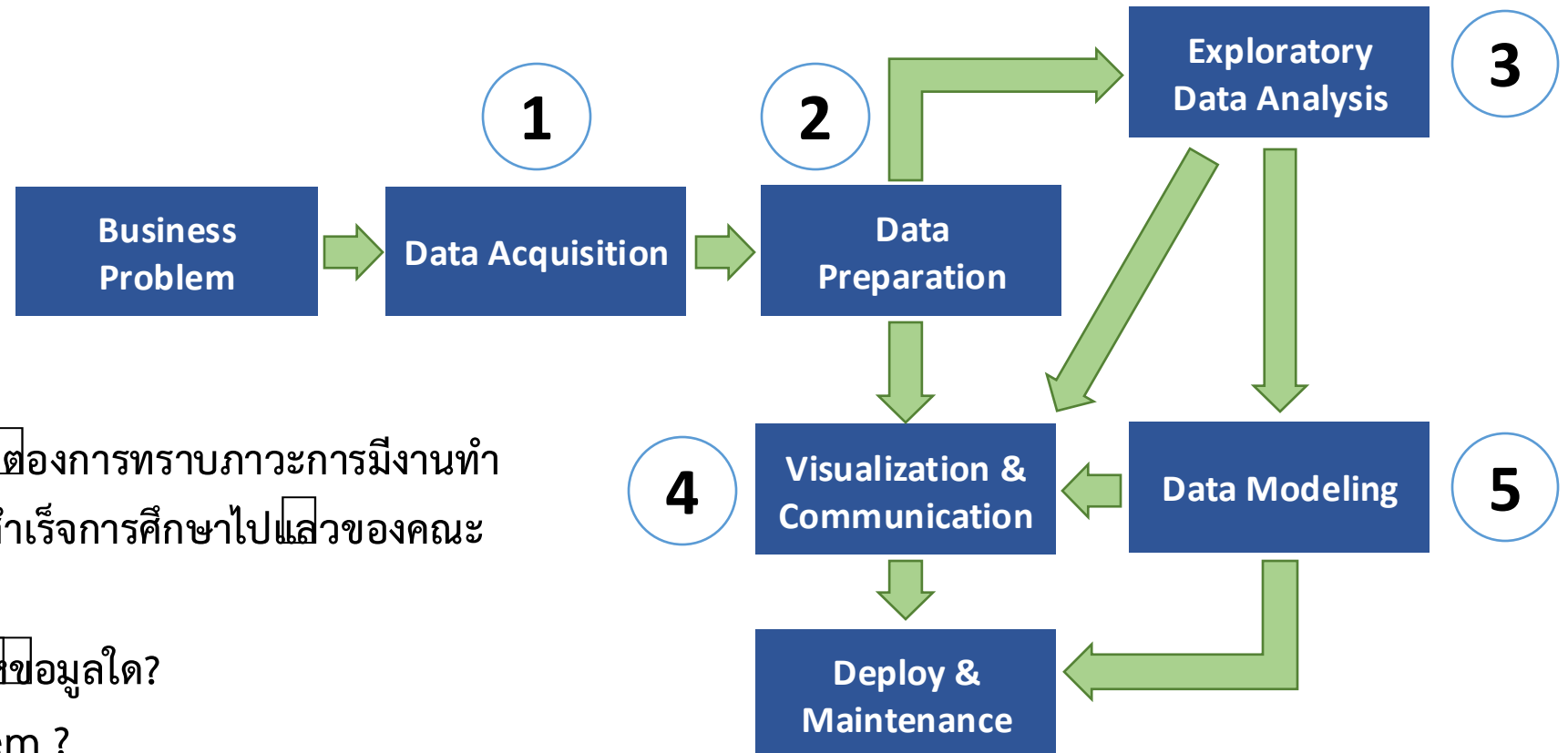


Problem

กรมอุตุนิยมวิทยามีข้อมูลอุณหภูมิ ความกดอากาศ ปริมาณ
ความชื้น และกระแสลม ณ บริเวณต่างๆ ของประเทศไทย
เจ้าหน้าที่กรมคนหนึ่งต้องการใช้อุปกรณ์ดังกล่าวในการพยากรณ์
ความเป็นไปได้ที่ฝนจะตกในบริเวณต่าง

- อยู่ในลำดับขั้นการวิเคราะห์ข้อมูลใด?
- อะไรคือ Business Problem ?
- มีขั้นตอนการวิเคราะห์ข้อมูลอย่างไร?

Practice



Problem

- ผู้บริหารมหาวิทยาลัยเชียงใหม่ต้องการทราบภาวะการปฏิบัติงานทำและสายอาชีพของนักศึกษาที่สำเร็จการศึกษาไปแล้วของคณะ □ ต่างๆ ในมหาวิทยาลัย
- □ อยู่ในลำดับขั้นการวิเคราะห์ข้อมูลใด?
- อะไรคือ Business Problem ?
- มีขั้นตอนการวิเคราะห์ข้อมูลอย่างไร?
- □ ยกตัวอย่างข้อมูลที่จะต้องเก็บรวบรวม (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.