Syllabus and Introduction

Petchara Pattarakijwanich

Introduction to Data Science, 19 August 2022

Outline of This Week

- Syllabus of This Course

- What are Data Science and Machine Learning?
- Simple Examples of DS&ML

- Types of Problems in DS&ML
  (Regression vs Classification, Supervised vs Unsupervised)
- Example of Problems in Real-Life

- Methodology (Training $\rightarrow$ Validation $\rightarrow$ Application)
- Under/Over-fitting, Bias-Variance Tradeoff

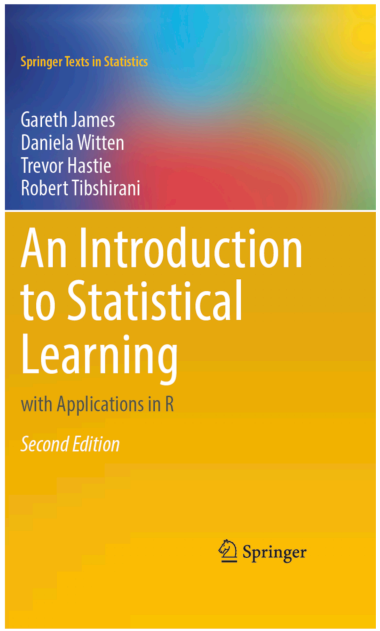
Syllabus of This Course

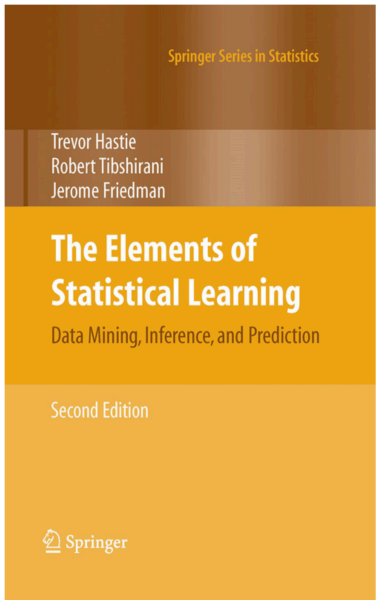
Topics:

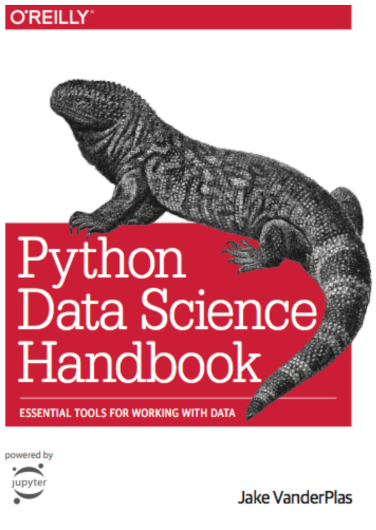
- Basic Programming in Python
- Regression and Classification
- Resampling Methods
- Subset Selection
- Shrinkage Methods (Ridge & Lasso)
- Dimensionality Reduction Method
- Decision Tree, Bagging, Boosting, Random Forest
- Support Vector Machine
- Neural Network
- Unsupervised Learning

Syllabus of This Course

- An Introduction to Statistical Learning (Gareth James et al.)
  https://www.statlearning.com
- The Elements of Statistical Learning (Trevor Hastie et al.)
  https://hastie.su.domains/ElemStatLearn/
- Python Data Science Handbook (Jake VanderPlas)
  https://jakevdp.github.io/PythonDataScienceHandbook/







Syllabus of This Course

Grading:

- Homework and Exercise 50%
- Final Project 50%

Syllabus of This Course

Calendar:

- Weeks with classes:
  - 19/26 Aug
  - 2/9/16/23/30 Sep
  - None in Oct
  - 4/11/18/25 Nov
  - 2 Dec
- Weeks with no class: 14/21/28 Oct
- Midterm exam Week: 3-7 Oct
- Final exam Week: 6-16 Dec

What is Data Science?

What is a Data Scientist?

Data scientists use technology to glean insights from large amounts of data they collect. It's a field that requires statistics, quantitative reasoning and computer programming skills. On top of all that, you need to be a good communicator so you can report your research findings and explain how they address a larger question you're trying to answer.

"A data scientist really is a scientist at heart," says Scott Beliveau, chief of the enterprise advanced analytics branch within the U.S. Patent and Trademark Office's Office of the Chief Technology Officer. "But rather than using chemicals or other things, a data scientist uses data – numbers, zeros, sometimes it's textual information – to try and solve and answer problems."

MEDIAN SALARY
$98,230

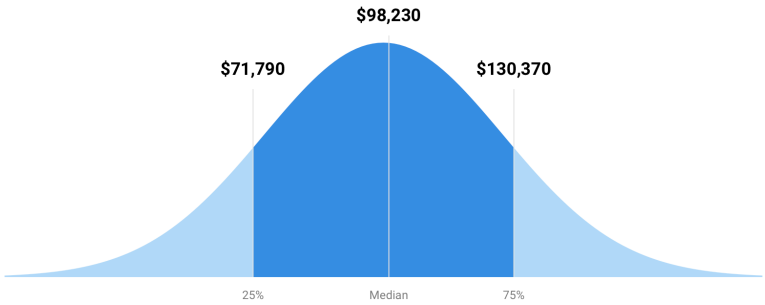
UNEMPLOYMENT RATE
2%

NUMBER OF JOBS
19,800

While data science is still a new career field, employers are increasingly recognizing the value of professionals with this expertise. Today, you'll find data scientists working at a range of organizations, including tech startups, government agencies, large companies and research institutions.

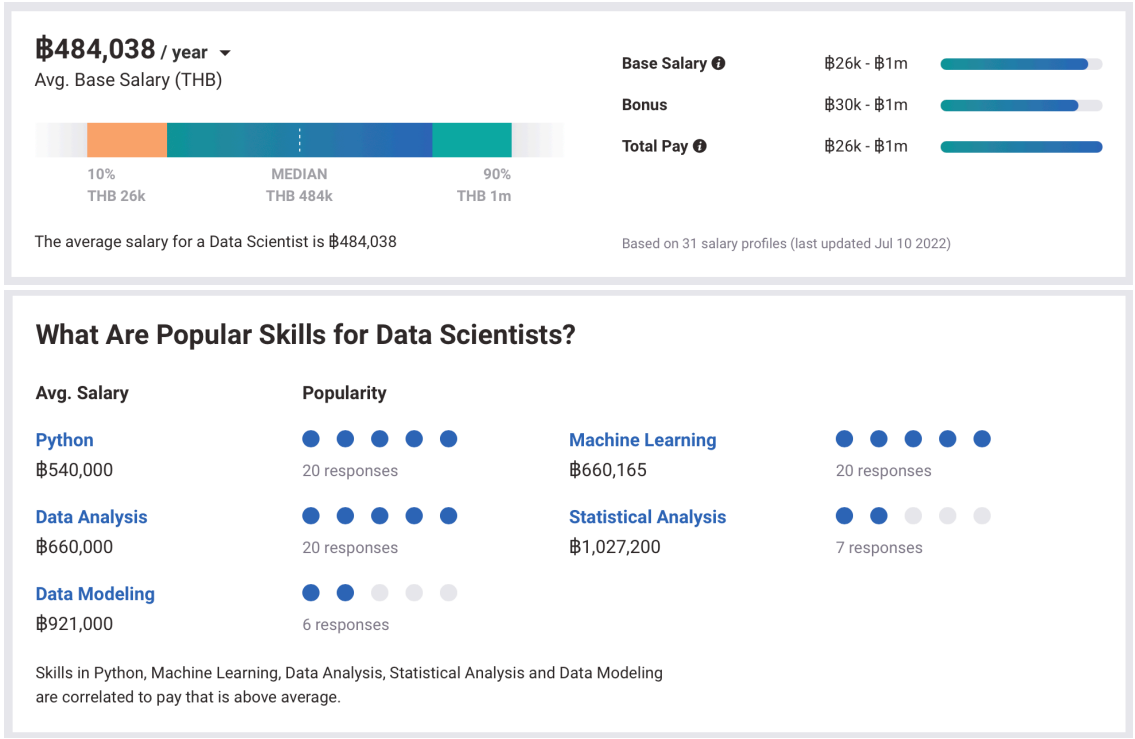
How Much Does a Data Scientist Make?

Data Scientists made a median salary of $98,230 in 2020. The best-paid 25 percent made $130,370 that year, while the lowest-paid 25 percent made $71,790.



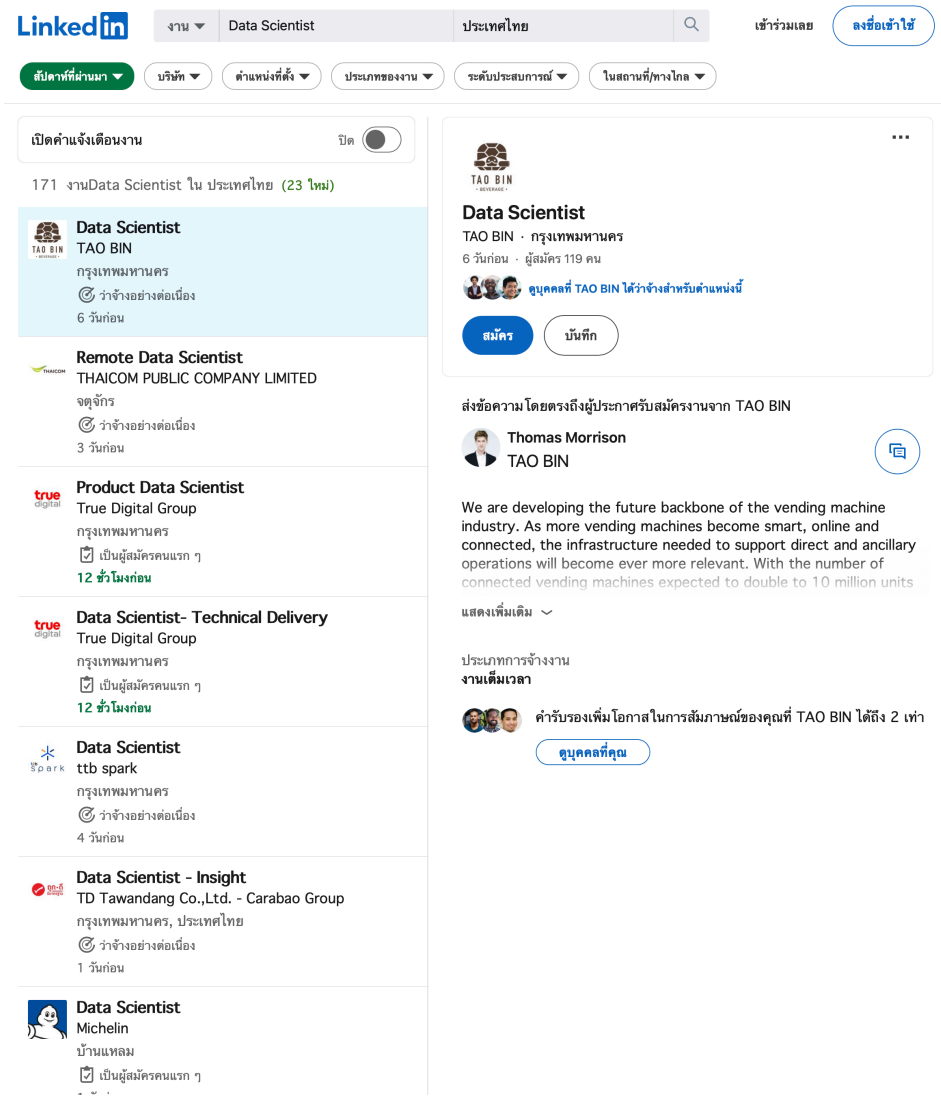
https://money.usnews.com/careers/best-jobs/data-scientist

What is Data Science?



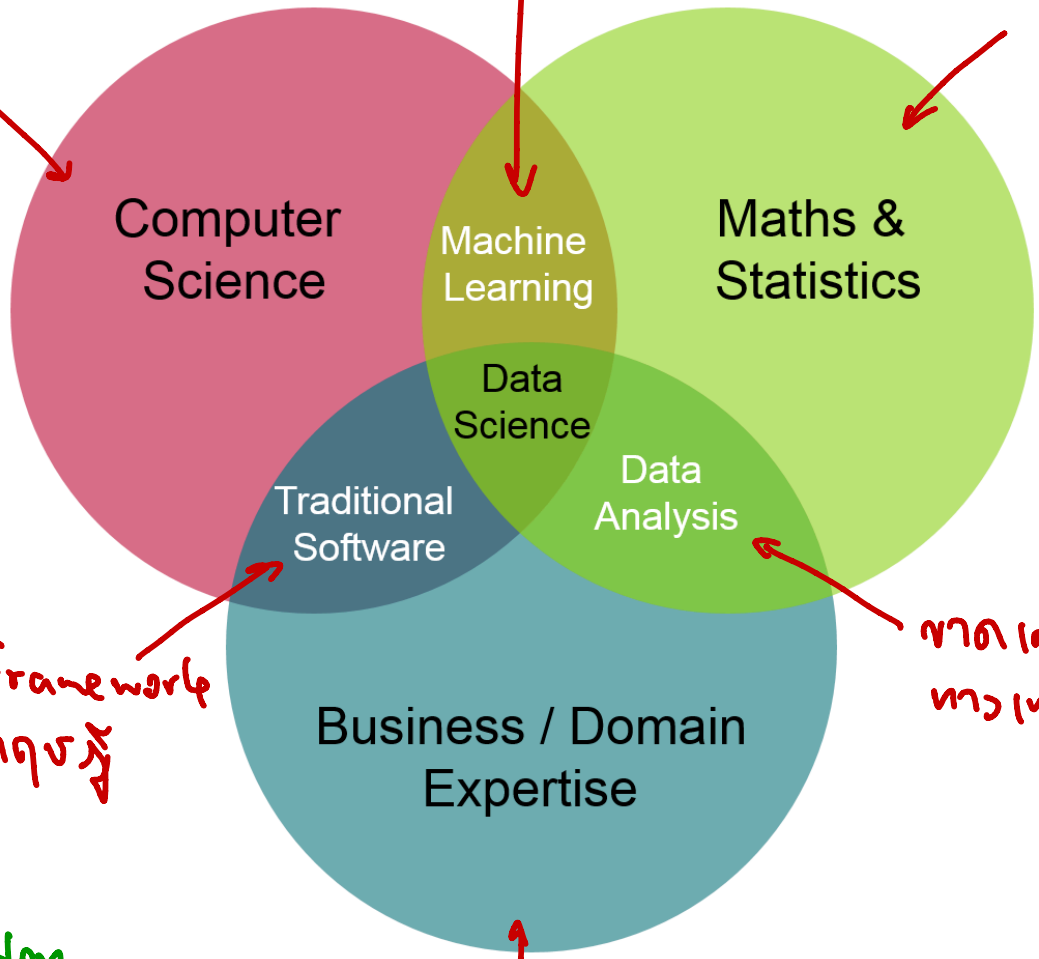
https://www.payscale.com/research/TH/Job=Data_Scientist/Salary

What is Data Science?



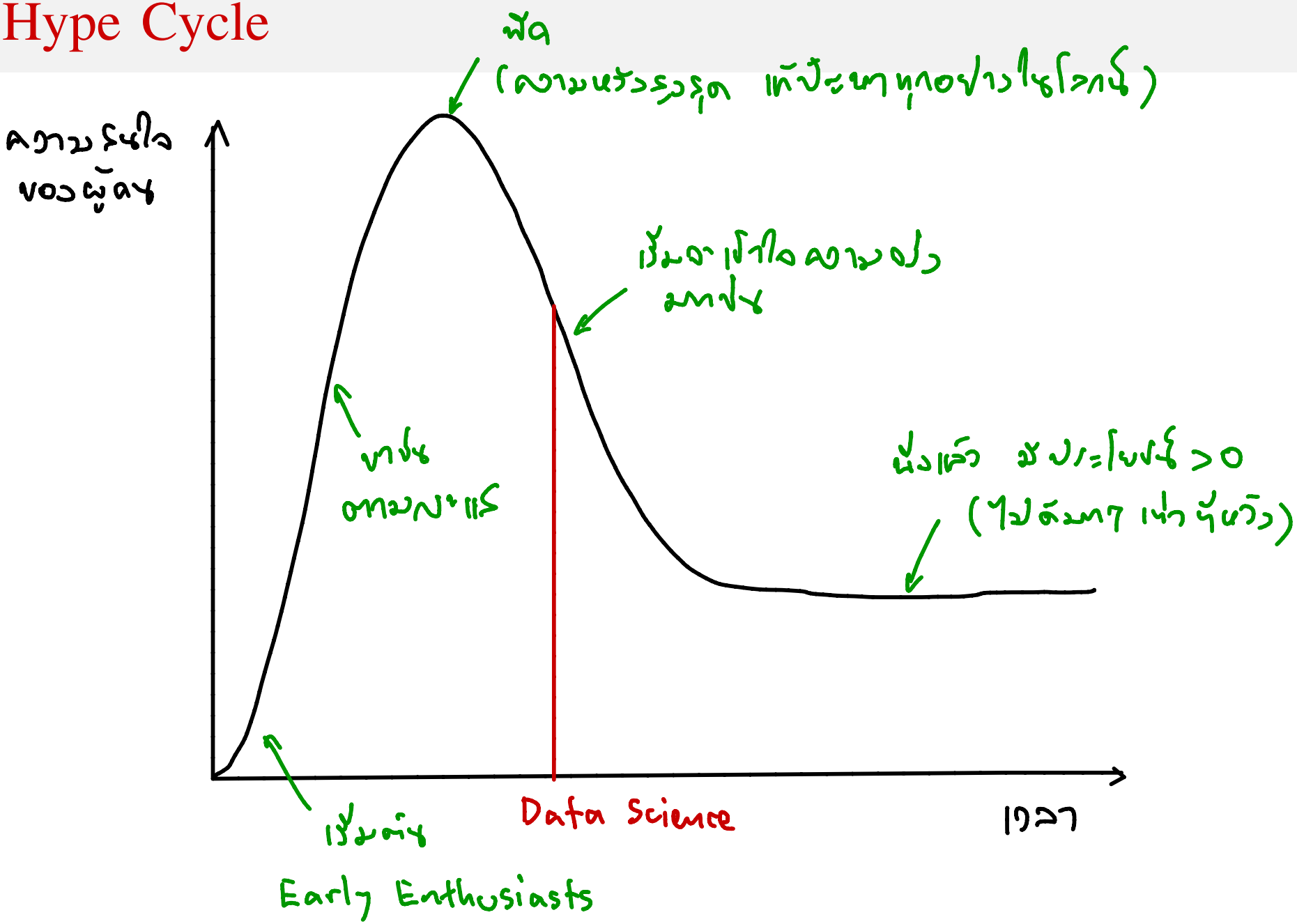
https://th.linkedin.com/

What is Data Science?



https://thedatascientist.com/data-science-without-programming/data_science_venn_diagram/

Hype Cycle



What is Data Science?

- Extract insights from data $\rightarrow$ Actionable Plan
- Use scientific methods
- Techniques from mathematics and statistics
- Algorithms and data management from computer science

Simple Example of Data Science

ฝึกนักศึกษาที่จบมา 2 ไฟล์

① เด็กปี 1 คณะวิทย์ 5 รุ่นที่ผ่านมา (≈ 1500 คน)
- GPA ม.ปลาย
- คะแนน ONET วิชา 1, 2, 3, ..
- อาชีพพ่อแม่
- รายได้พ่อแม่
- โรงเรียนที่จบมา (เล็ก/ใหญ่, รัฐบาล/เอกชน/อื่นๆ)
- เกรดตอนจบปี 1
- ผ่านปี 1 มั้ย?

ใช้หาคุณหา Pattern

② เด็กปี 1 คณะวิทย์ ปีนี้ (≈ 300 คน)
- รู้รีดำ
- ไม่รู้ที่เป็นรีแดง

ใช้ Pattern ที่หา เดาคำตอบ

ใช้คำตอบ => ไปใช้ประโยชน์

Simple Example of Data Science

พี่กบสัมภาษณ์น้องมา 2 ไฟล์ หา f ที่ทำให้ $\vec{y} = f(\vec{x})$

① เด็กปี 1 ตอนจบ 5 รุ่นที่ผ่านมา (≈ 1500 คน)

- GPA ม.ปลาย (x_1)
- คะแนน ONET วิชา 1, 2, 3, .. (x_2, x_3, x_4)
- อาชีพพ่อแม่ (x_5)
- รายได้พ่อแม่ (x_6)
- โรงเรียนที่จบมา (เล็ก/ใหญ่, รัฐบาล/เอกชน/อินเตอร์)

} Independent variables, Features, Inputs

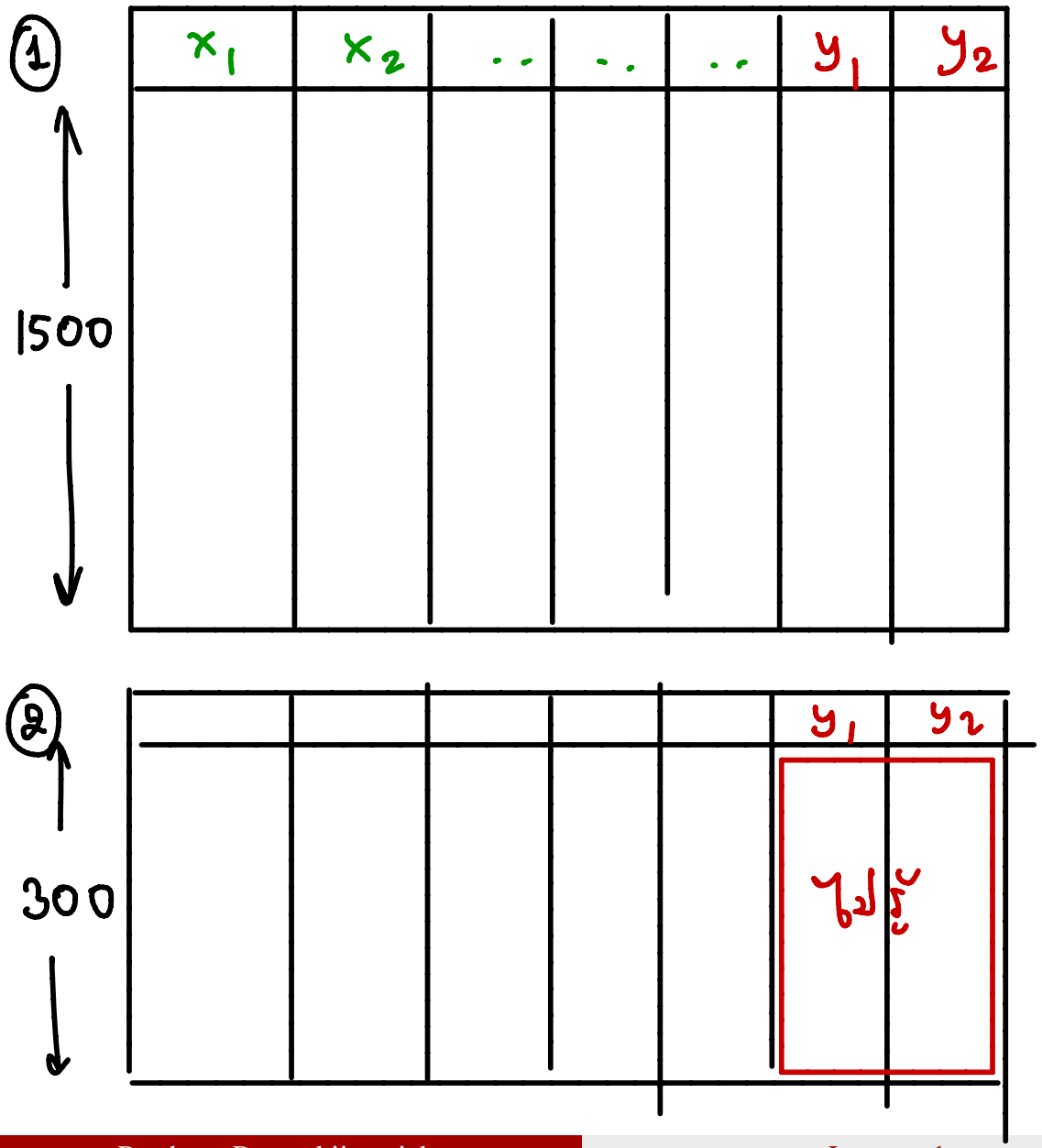
- เกรดตอนจบปี 1 (y_1)
- ผ่านปี 1 มั้ย? (y_2)

} Dependent variables, Targets, output

② เด็กปี 1 ตอนจบ ปีนี้ (≈ 300 คน)

- รู้ $\vec{x}$
- ไม่รู้ $\vec{y}$ แล้ว ↙ $\vec{y}_{predicted} = f(\vec{x})$

Simple Example of Data Science

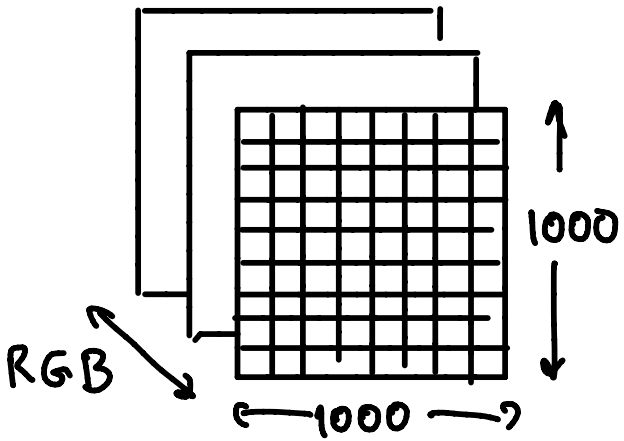


หา f ที่ทำให้ $y = f(x)$

ได้ f ที่ทำให้ $y = f(x)$

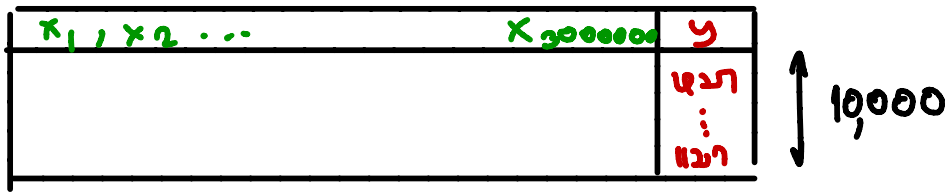
Simple Example of Data Science

โจทย์ : Image classification ภาพนี้เป็น หมา รึแมว ?

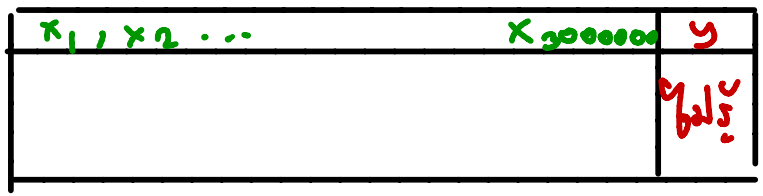


⇒ มี $1000 \times 1000 \times 3$ ตัวเลข

⇒ $x_1, x_2, \cdots x_{3000000}$

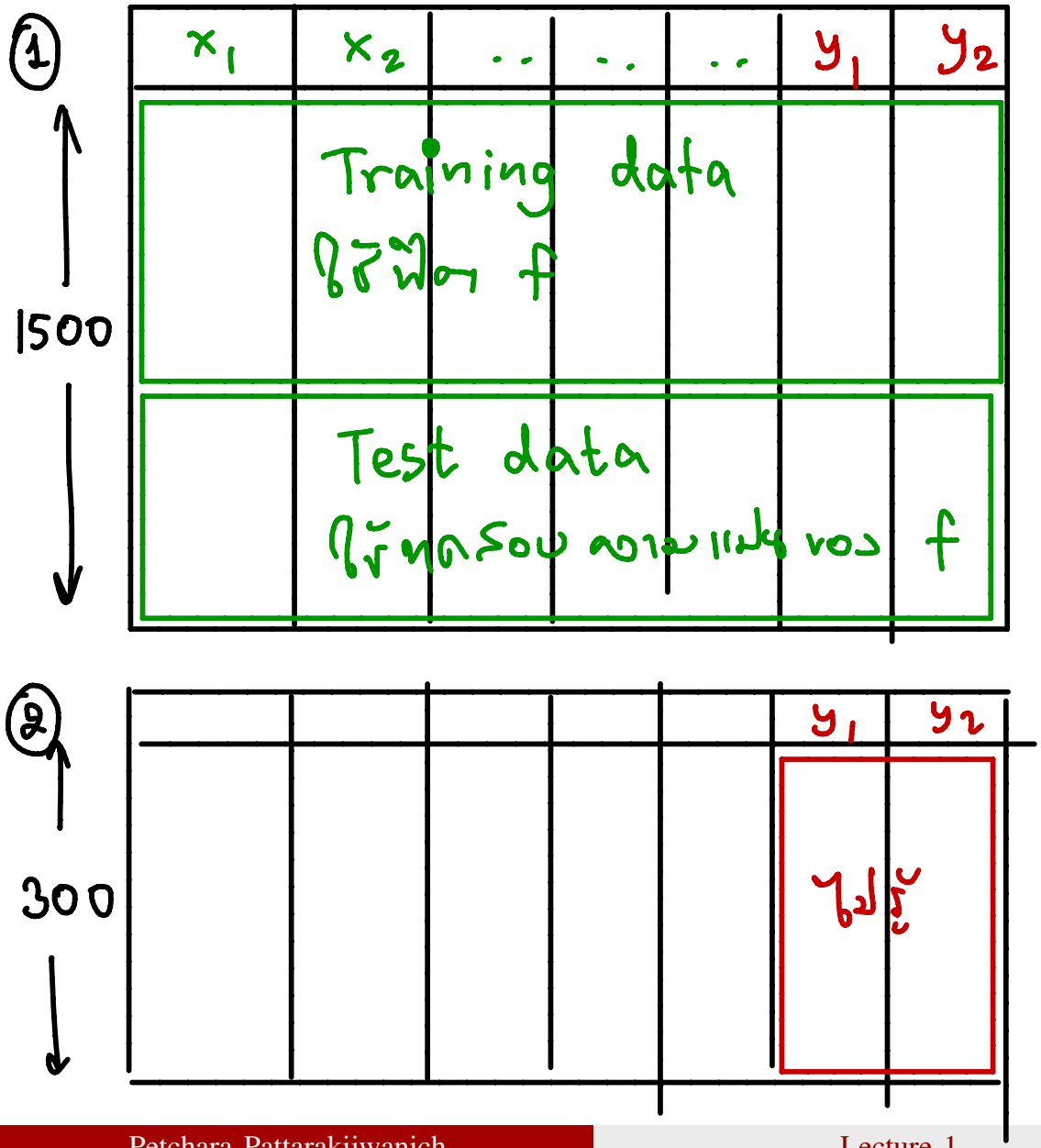
① 

พัฒนา f ตามข้อมูล ①

② 

ใช้ f ที่ทำตาม ②

Simple Example of Data Science

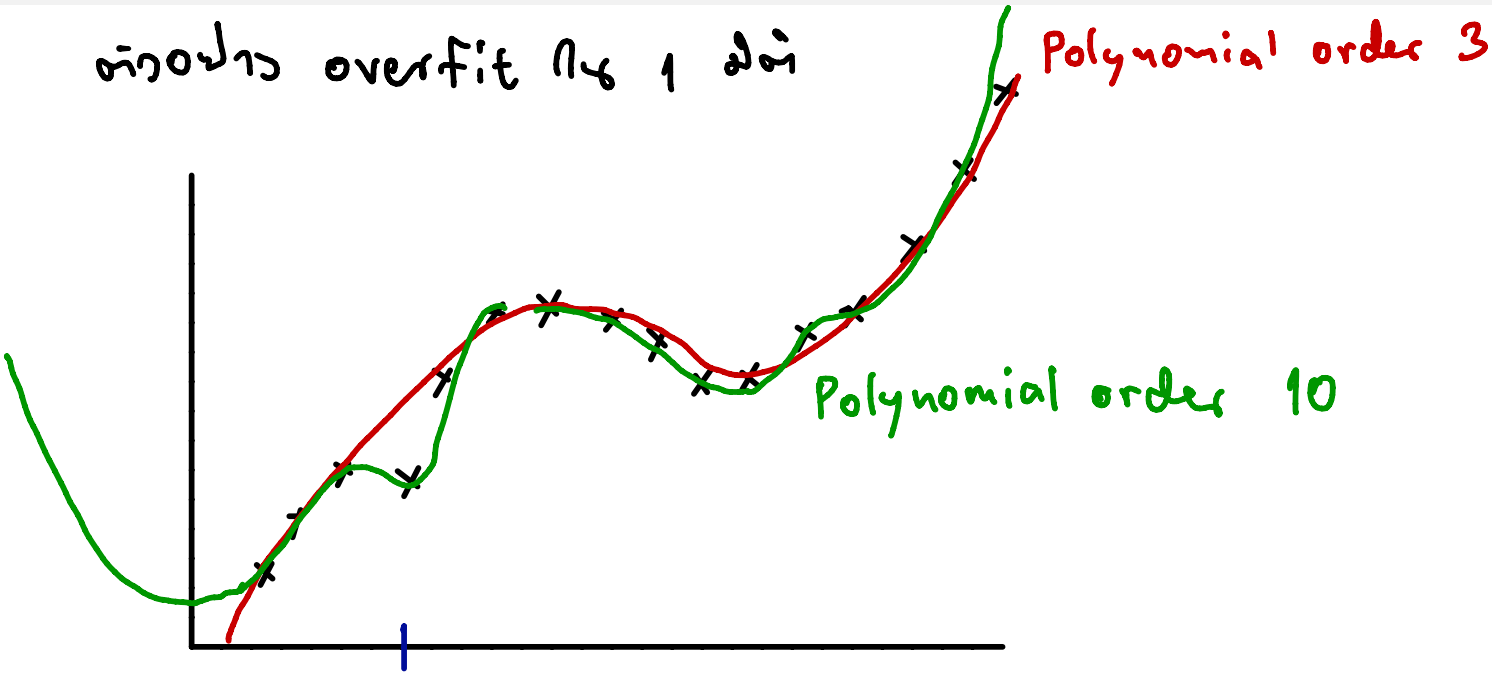


หา f ที่ทำให้ $y = f(x)$

รู้ได้ไงว่าไม่ overfit ?

ใช้ f ทำนาย $y = f(x)$

Simple Example of Data Science



Simple Example of Data Science

Input $x_1, x_2, \ldots, x_i$ $\Rightarrow$ Output y_1, y_2

ประเภทตัวแปร

- Quantitative (Real Number) [GPA , คะแนน ONET , ...]
- Qualitative / Categorical (เป็นกลุ่มๆ) [รัฐต่างๆ , เพศ , ...]
- Ordered Categorical [Letter Grade (A, B, C ...) , ดีมาก/ดี/พอใช้]

Output

- Quantitative $\rightarrow$ Regression
- Qualitative $\rightarrow$ Classification

} Supervised Learning

Real-life Example of Data Science

- Identification of cells, tissues, bacteria, etc. from medical images
- Identification of dog breed from images
- Reading handwritten text from image (image to text)
- Personal identification from image (facial recognition)
- Automatic caption (speech to text)
- Predict property prices from location and other information
- Predict sales of businesses
- Predict whether a debt will be defaulted

Real-life Example of Data Science

- Predict what diseases a person might have from medical record
- Predict a person's health from DNA (personalized medicine)
- Predict performance of employees from their CVs
- Predict health/activity from smart watch data
- Identify writers from writing styles
- Predict of credit card fraud
- Predict consumer buying behavior
- Predict election outcome from social media trend