Course Overview

2110446 Data Science and Data Engineering

Assoc. Prof. Peerapon Vateekul, Ph.D.
Department of Computer Engineering,
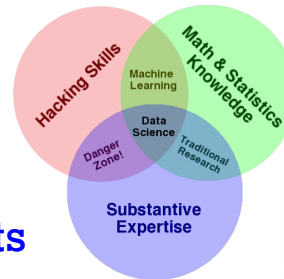Faculty of Engineering, Chulalongkorn University
Peerapon.v@chula.ac.th

CHULA ENGINEERING
Foundation toward Innovation
COMPUTER

* Part of this slide is modified from a slide of Prof.Natawut

What is this class about? (recap)

Dr. Virote
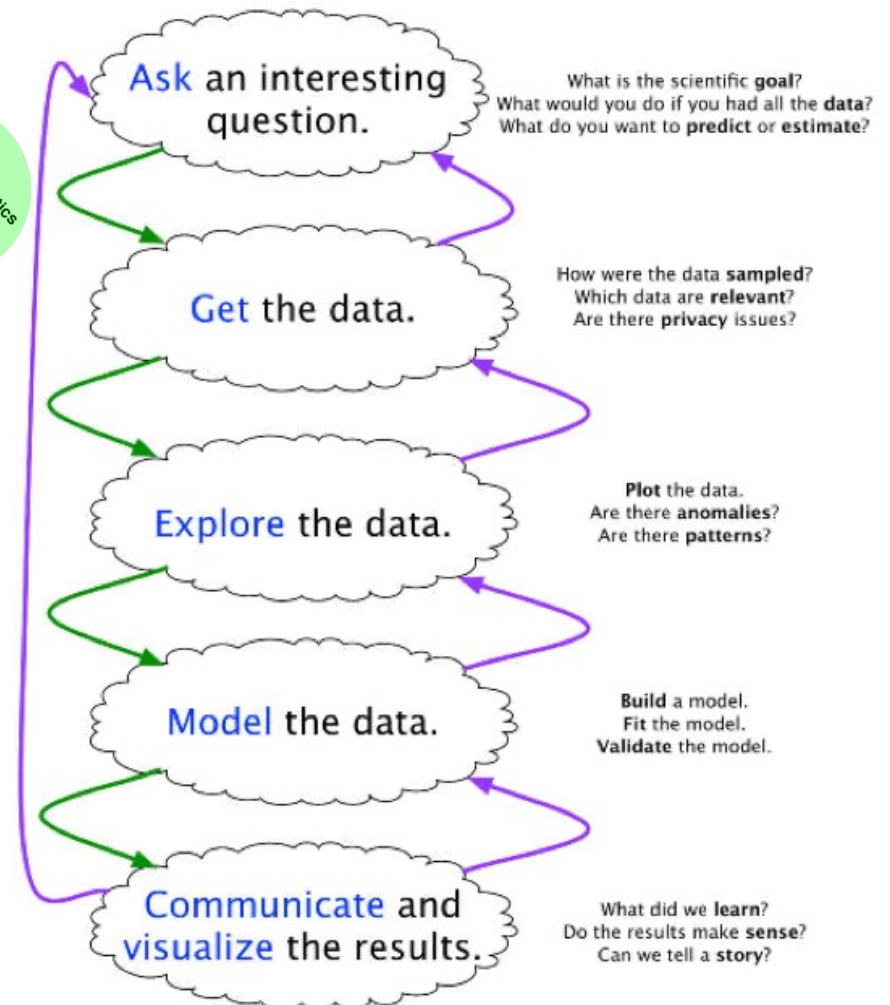
1. Transform data into valuable insights

2. Transform data into data products

3. Transform data into interesting stories

Aj. Natawut

1. Measurement (decision)

2. Insights (knowledge)

3. Data Products (Innovation, Intelligent)

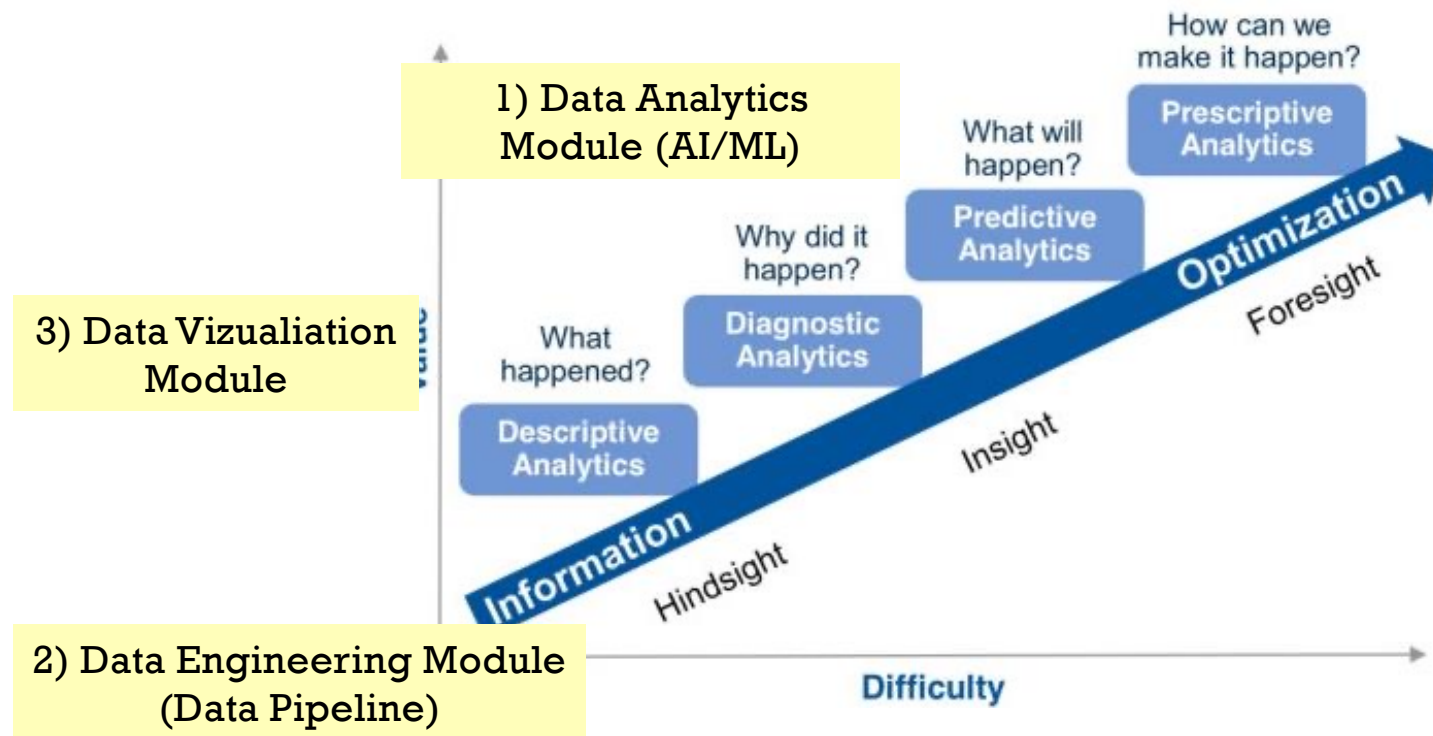The Data Science Process

Ask an interesting question.
What is the scientific goal?
What would you do if you had all the data?
What do you want to predict or estimate?

Get the data.
How were the data sampled?
Which data are relevant?
Are there privacy issues?

Explore the data.
Plot the data.
Are there anomalies?
Are there patterns?

Model the data.
Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.
What did we learn?
Do the results make sense?
Can we tell a story?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

What is this class about? 3 Modules

1) Data Analytics Module (AI/ML)

3) Data Vizualiation Module

2) Data Engineering Module (Data Pipeline)

How can we make it happen?

Prescriptive Analytics

What will happen?

Predictive Analytics

Why did it happen?

Diagnostic Analytics

What happened?

Descriptive Analytics

Optimization

Foresight

Insight

Information

Hindsight

Difficulty

#	Tue (1PM-4PM)	Contents	Instructor	Module
1	7-Jan-25	Introduction & data preparation	Aj.Peerapon	Data Science
2	14-Jan-25	Traditional ML (1)	Aj.Peerapon	Data Science
3	21-Jan-25	Traditional ML (2)	Aj.Peerapon	Data Science
4	28-Jan-25	DL (1)	Aj.Peerapon	Data Science
5	4-Feb-25	DL (2)	Aj.Peerapon	Data Science
6	11-Feb-25	Advanced topics & Guest speakers * Modern ML techniques (Transformer) * Sharing data scientist experience (in Thailand & Abroad) [2-3 guest speakers]	Aj.Peerapon	Data Science
7	18-Feb-25	Big data architecture + data storage	Aj.Natawut	Big Data Eng.
8	25-Feb-25	Web scraping	Aj.Natawut	Big Data Eng.
	5-Mar-25	Midterm Exam Week (3-7 Mar)		
9	11-Mar-25	Data ingestion	Aj.Natawut	Big Data Eng.
10	18-Mar-25	Big data processing (Spark)	Aj.Natawut	Big Data Eng.
11	25-Mar-25	MLOps: Orchestration (Airflow) and serving (FastAPI, Seldon Core)	Aj.Natawut	Big Data Eng.
12	1-Apr-25	Data visualization	Aj.Veera	Data Viz.
13	8-Apr-25	Python visualization, Streamlit, and Dash	Aj.Veera	Data Viz.
	15-Apr-25	Songkran Holiday		
14	22-Apr-25	Graph visualization	Aj.Veera	Data Viz.
	30-Apr-25	Final Exam Week (28 Apr -14 May) TBD: Tue 6 May 2025 at 1PM - 4PM		

Evaluation

- Module1 Assignment (data analytics)   15%

- Module2 Assignment (data engineering)   15%

- Module3 Assignment (data visualization)   10%

- Midterm Exam (Kaggle)   15%

- Project   10%

- Final Exam   30% (Lab Test)

- Attendance   5%
  - Students need to attend the class on-site at least 80% (at least 12 weeks) as a mandatory criterion to "pass" this course.

Remarks

- **This course assumes that:**

- It is <u>just</u> a fundamental course to perform a data science task properly. Then, you can choose your own specialty in AI domains. <mark>(Year 3)</mark>

- It aims to utilize all tools and algorithms **without** theoretical proofs.

- All students must have a strong background in Python programming. <mark>(CP only)</mark>

- **There are many courses related to AI/ML/DS?**

- AI

- Data Mining

- Machine Learning

- Neural Networks

- ASR

- NLP

- Computer Vision

- Etc.

ปีการศึกษา 2567/ภาคการศึกษาปลาย

ทวิภาค

2110446 DATA SCI/ENG

วิทยาศาสตร์ข้อมูลและวิศวกรรมข้อมูล

DATA SCIENCE AND DATA ENGINEERING

คณะวิศวกรรมศาสตร์ (ภาควิชาวิศวกรรมคอมพิวเตอร์)

3.0 **CREDIT HOURS =** LECT 3.0 CR

(LECT 3.0 HR)

เงื่อนไขรายวิชา : -

วันสอบกลางภาค : TDF (รอประกาศ) วันสอบปลายภาค : TDF (รอประกาศ)

ตอนเรียน	วิธีสอน	วัน-เวลาเรียน	อาคาร	ห้อง	ผู้สอน	หมายเหตุ	จำนวนนิสิต Regis/Max
21	LECT	TU 13:00-16:00	EN100	201A	NNP,VMS,PVK	APPROVED CP	100/100

CHULA **ΣNGINEERING** **COMPUTER**

+ All students need to fill this form by Wed 8 Jan 2025.
We will reply you on Thu 9 Jan 2025.

แบบสอบถามสำหรับการลงทะเบียน
วิชา 2110446 Data Science and Data Engineering (2024/2)

Submission deadline is Friday (12/01/2567)

* Indicates required question

Student ID *

Your answer

First name *

https://forms.gle/32jKaKuiSeVNK8zM6

CHULA ΣNGINEERING COMPUTER
Foundation toward Innovation

On Thu 9 Jan 2025, we are going to add you to the grader system (https://dsde.nattee.net/)

- 01_pandas_01_2024s2

Task: Pandas Basic

The file "scores.csv" contains the data with two columns

1. id -> id of student
2. score -> score of student

Problem: Let's Learn how to use pandas !

Input:

1. Input file name
2. Question number from 1 to 4

Output:

For Q1, please find the shape of the data

For Q2, please find the max score of the data

For Q3, please find the total student that have score equal or more than 80 points

Otherwise, just print "No Output"

You always need to bring your laptop to the class. Please install VS Code & conda. ([Link](#))

Conda + VScode Handbook

• • •

DSDE CP 2110446 2024/2

Tools

Python Virtual Environment Management Tool

1. Anaconda Navigator - UI
2. Conda - command line

IDE

1. VSCODE

Any questions? ☺