Introduction to Data Science

(Khoa học dữ liệu)

Khoat Than

Hanoi University of Science and Technology

khoattq@soict.hust.edu.vn

IT4142E, SOICT, HUST, 2019

About the course

- Period: 15 weeks
  - Lectures: 15 weeks
- Location & time:
  - D9-102,
  - Wednesday, 7:35 – 10:05
- Lecture directory
  http://is.hust.edu.vn/~khoattq/lectures/DS-intro-2019-9
- Join with us and discuss somethings:
  https://www.facebook.com/groups/230365807881078/
- Contact:
  - Email: khoattq@soict.hust.edu.vn
  - Address: room 1002, Building B1, HUST

This course

- You will learn to take data:
  - Understand
  - Process
  - Extract value
  - Visualize
  - Communicate
  - Make prediction

"The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data."
- Hal Varian, Google's Chief Economist

Contents of the course

- Introduction to Data Science
- Data crawling and processing
- Data cleaning and integration
- Exploratory data analysis
- Machine Learning
- Big data analysis
- Visualization
- Text analysis
- Image and video analysis
- Graph analysis
- Recommender system

Teachers/Instructors

Khoat Than

Viet-Trung Tran

Kiem-Hieu Nguyen

Oanh Nguyen

Mai-Anh Bui

Some technologies we will use

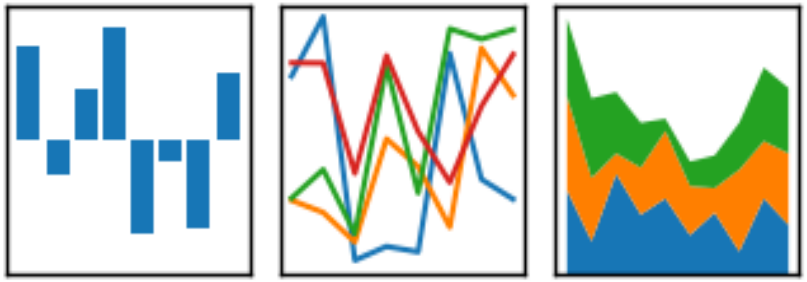
python™

ANACONDA
Powered by Continuum Analytics®

TensorFlow

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Evaluation (đánh giá)

- Attendance and activeness in the lectures
- Midterm test: **Capstone Project**
  - Students work in groups
  - Each group consists of 3-5 students
  - Each group chooses a data analysis problem
  - Propose solution, implement, and evaluate its effectiveness
- Final exam
  - Paper-based multiple-choice
- Overall: Midterm test (40%) + Final exam (60%)

Capstone Project: topic proposal

- Each group freely choose a problem/topic to be solved, datasets to be used, algorithms for analyzing data
- Each proposal should be precisely described
  - The problem: short description, input, output, data type, future application, …
  - The algorithms, tools to be used
  - Data sets to be used
- **Proposal registration: before 15/10/2019**
  - Via Google Form (TBA later)
  - Proposal: project name + description
  - List of members: student name + ID + Email

Capstone Project: requirements

- The result of the project will be presented in the ending period of this subject
  Every member is required to contribute to his/her project and presentation
- Project report:
  - **Source code**: save your code into one zip file
  - **Readme.txt**: describes clearly how to setup, compile, and run your program
  - **Written report:**
    - Introduce the problem to be solved, the data sets were used
    - Details about the methods for analyzing data
    - Results of different evaluations, new conclusions/findings, …
    - The main components of your code
    - The difficulties in this project, and your proposed solution

Capstone Project: evaluation

- The evaluation of each project will be based on
  - The difficulty of the problem of interest
  - The appropriateness & quality of the chosen method/solution
  - The rigor of the empirical evaluation and assessment on the chosen method for analyzing data
  - The quality of the presentation
  - The quality of the written report
- Each project will have 15' for slide presentation & demo
- **If you use some existing libraries/packages/codes, you have to clearly declare your usage in the written report and slide presentation**

References

- Reference books:
  - Grus, Joel. *Data science from scratch: first principles with python*." O'Reilly Media, Inc.", 2015.
  - Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition)*. Springer series in statistics, 2009.
  - Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.