



HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY  
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY



HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY  
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

# Introduction to Data Science (IT4142E)

# About the course

- Period: 15 weeks
  - 11 lectures
  - 1 week: progress report for capstone project
  - 3 weeks: final report for capstone project
- Location & time
- Join with us and discuss somethings:  
<https://www.facebook.com/groups/230365807881078/>
- Contact:
  - 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

- ❑ Lecture 1: Overview of Data Science
- ❑ Lecture 2: Data crawling and preprocessing
- ❑ Lecture 3: Data cleaning and integration
- ❑ Lecture 4: Exploratory data analysis
- ❑ Lecture 5: Data visualization
- ❑ Lecture 6: Multivariate data visualization
- ❑ Lecture 7: Machine learning
- ❑ Lecture 8: Big data analysis
- ❑ Lecture 9: Capstone Project guidance
- ❑ Lecture 10+11: Text, image, graph analysis
- ❑ Lecture 12: Evaluation of analysis results

# Instructors



**Khoat Than**



**Viet-Trung Tran**



**Kiem-Hieu  
Nguyen**



**Oanh Nguyen**



**Mai-Anh Bui**

## Some technologies we will use



## Evaluation (đánh giá)

- Attendance and activeness in the lectures
- Midterm test: **Capstone Project**
  - Students work in groups
  - Each group consists of 3-4 students
  - Each group chooses a data analysis problem
  - Propose solution, implement, and evaluate its effectiveness
- Final exam
  - Multiple-choice test
- 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:**
  - TBA
  - Proposal: project name + description
  - List of members: student name + ID + Email

---

## Capstone Project: **requirements**

- The result 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

- Lectures: <https://users.soict.hust.edu.vn/khoattq/lectures/IT4142E-DS/>
- 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*. Springer, 2017.
  - Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.



**25**  
**SOICT**

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG  
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

**Thank you  
for your  
attentions!**

 [soict.hust.edu.vn/](http://soict.hust.edu.vn/)  [fb.com/groups/soict](https://fb.com/groups/soict)

