ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

IT4142E

Introduction to Data Science

Course presentation

Lecturer:

Muriel VISANI: murielv@soict.hust.edu.vn

Acknowledgements:

Khoat Than

Department of Information Systems
School of Information and Communication Technology - HUST

About the course

- Course name
  - Introduction to Data Science
- Volume
  - 15 x 3 x 45 minutes in total
  - 1 session of 3 x 45 mn each week, on Fridays afternoons
    - From 12h30 to 14h55, with a 10-minute break in the middle
- Lecturer
  - Muriel VISANI, Associate Professor
    - Contact: murielv@soict.hust.edu.vn
      - **Please use that email only, and Teams**
    - Department of Information Systems, SOICT, HUST
    - La Rochelle University, France

General information

- Prof. Visani (me)
  - Sent from La Rochelle University to Bach Khoa for at least 3 years
    - Since last year

La Rochelle University

- French public university, created in 1993
  - Provides 60 degrees in 4 faculties
  - Roughly 9,000 students, among which 10% foreigners
  - 466 lecturers / Ass. Prof. / Prof
  - Many student / researcher exchanges with Asia
    - In particular South-East Asia: Vietnam, Malaysia, Cambodia
    - In Hanoi, and in Computer Science / ICT:
      - 2 joint Master degrees with Hanoi universities
        - Vietnam National University (Trường Đại học Quốc gia Hà Nội)
        - University of Science and Technology of Hanoi (Trường Đại học Khoa học và Công nghệ Hà Nội)
      - Soon, hopefully, more joint degrees with Bach Khoa!

Contents of this course – in short

- You will learn to “take data”:
  - Understand
  - Process
  - Extract knowledge
  - 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

What will this course teach you?

- This course will introduce the basic ideas and techniques of Data Science: data scraping / pre-processing / cleaning and integration, Exploratory Data Analysis, Data visualization, Machine Learning, Big Data
- In this course, we will also present applications to Image and Video Analysis (Computer Vision)
- By doing a capstone project, students will gain practical experience in building a system based on DS.
- In addition, students will practice necessary skills for future work such as teamwork skills, research skills, writing reports and presentations.

Goals of this course

Goal	Description of the goal
M1	**Understand and be able to design and manage the systems which are based on Data Science (DS)**
M1.1	Identify and understand the components of the systems based on DS
M1.2	Identify, compare, and categorize the data types and systems in practice
M1.3	Be able to design systems based on DS in their future organizations
M2	**Identify and manage the opportunities from DS to boost the existing organizations, or develop new organizations**
M2.1	Understand and promote the use of DS to support their organizations
M2.2	Identify the (possible) impacts of Data Science on their organizations
M3	**Identify the newest trends in DS that are able to support development in organizations**
M3.1	Actively update and identify the most recent advances in Data Science
M3.2	Identify the opportunities from Data Science to develop their organizations

Contents of the course

- Chapter 1: Overview
- Chapter 2: Data scraping
- Chapter 3: Data cleaning, pre-processing and integration
- Chapter 4: Introduction to Exploratory Data Analysis
- Chapter 5: Introduction to Data visualization
- Chapter 6: Introduction to Machine Learning
  - Performance evaluation
- Chapter 7: Introduction to Big Data Analysis
- Chapter 8: Applications to Image and Video Analysis

Some technologies we will use / present

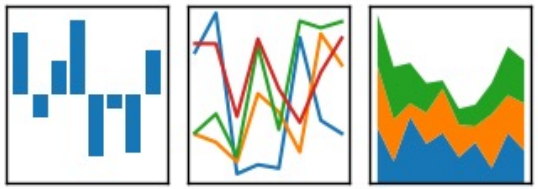
python™

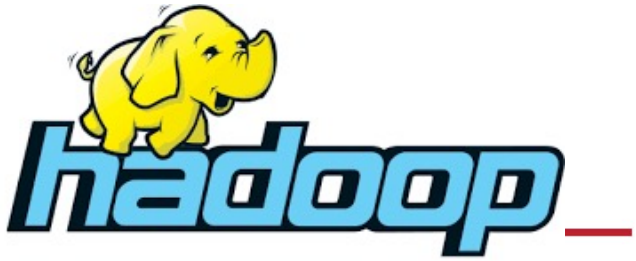
ANACONDA®
Powered by Continuum Analytics

TensorFlow

pandas

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





Evaluation

- **Final examination**: 60%
  - Written test or multiple-choice questions
- **Continuous assessment**: 40%
  - Activeness in class / homework / quizzes: 10%
  - Capstone project: 30% (determined later)
    - Students work in groups
    - Each group consists of 3-5 students
    - Study a problem, propose solutions, implement, and evaluate their effectiveness / efficiency
      - More details about the problem(s) will be given later on

Some information about the lectures

- In the teaching material, I often use *e.g.* and *i.e.*
  - Usually, common in English, but often HUST students don't know
  - *E.g.* = for example; *I.e.* = thus, therefore
- I create two versions of my slides: VStud and Vprof
  - For better **interactivity** during the lectures
  - Example of Vstud - VProf:

VStud

- **Question:**
  - What is the right solution to this problem?

VProf

- **Question:**
  - What is the right solution to this problem?
  - Answer: 42

- I'll give you the VStud beforehand, and play the VProf during the lectures
  - Please, fill in the blanks in the Vstud yourself during the lectures

Some information about the lectures

- For online sessions
  - Only the evaluation session(s) will be recorded
  - For other sessions, I will not record the lectures because of multiple reasons
    - Image reproduction right and personal privacy rights
    - Intellectual property rights
    - For re-usability of the teaching material
- But, to compensate for possible connection problems, I made the teaching material much more exhaustive than normal
  - I give you almost 2 times more slides than for offline lectures!!!
  - So that, if you cannot attend the lecture for connection problem, you still have the info you need
- If you cannot attend one lecture:
  - Please send me an email beforehand, or during the lecture:
    - murielv@soict.hust.edu.vn
  - Don't hesitate to ask questions on Teams to me / my TA / other students

Quizz

- In this course, we'll sometimes use Kahoot for interactive quizzes
  - Each student must create a Kahoot account
  - **Kahoot results might be used for student's evaluation**
  - You must follow strictly the instructions given in the slides for the Kahoot quizzes

Some educational recommendations

- Attend classes
  - If you can't, then send me an email in advance
- **Turn off your cell phone**
- Read the reference books
- **NO PLAGIARISM**
- Do not hesitate to
  - Ask questions **at any point during the lecture**
    - Even if it's only because of difficulties with English
  - Give your opinions / feedback
  - Discuss with me / my TA / other students on Teams

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.

Questions