COURSE OVERVIEW

Dr. Le Hoang Son

GENERAL INFORMATION

- Subject: DATA MINING
- Credit: 2 (90 hours)
- Objectives:
  - Understand the principle of data mining
  - Understand some preliminary tasks in data mining
  - Do research in the final project (require making real applications)
- Requirements:
  - Mid-term exam: Paper test (30% grade)

  =========
  - Project: Write an application (30% grade)
  - Standalone Presentation (40% grade)

  =========
  - Research topic: paper & report for individual (70% grade)

TIMETABLE

- Week 1: Introduction & Data Mining Overview (10/2)
- Week 2: Data Processing (17/2)
- Week 3: Database Technology (24/2)
- Week 4: Data warehouse and OLAP (3/3)
- Week 5: Regression I (10/3)
- Week 6: Regression II (17/3)
- Week 7: Classification I (24/3)
- Week 8: Classification II (31/3)
- Week 9: Mid term paper test examination (7/4/2023)
- Week 10: Prediction I (14/4)
- Week 11: Prediction II (21/4)
- Week 12: Clustering I (28/4)
- Week 13: Clustering II (5/5)
- Week 14: Association Rules Mining (12/5)
- Week 15: Applications & Trends & Visualization Data Mining (19/5)
- Week 17: Final preparation with 6 projects (2/6/2023)

SCHEDULE

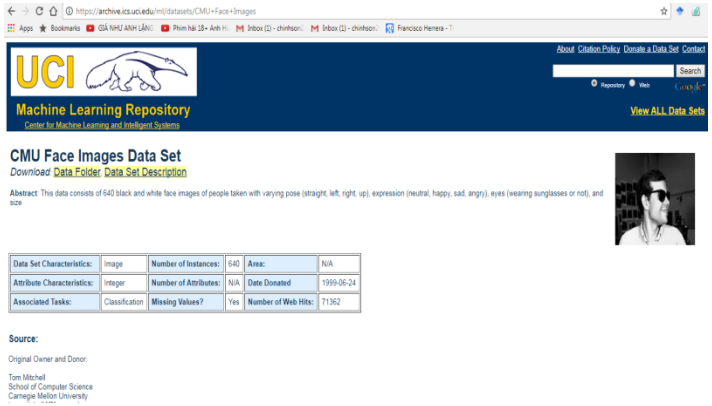
56	MAT3534	Khai phá dữ liệu	3	MAT3534 1	30/15/0	25	4	6-10	PM	Tiếng Việt	Lê Hoàng Sơn
57	MAT3534	Khai phá dữ liệu	3	MAT3534 2	30/15/0	25	4	6-10	PM	Tiếng Việt	Lê Hoàng Sơn
57	MAT3534	Khai phá dữ liệu	3	MAT3534 3	30/15/0	20	4	6-10	PM	Tiếng Việt	Lê Hoàng Sơn
57	MAT3534	Khai phá dữ liệu	3	MAT3534 4	30/15/0	20	4	6-10	PM	Tiếng Việt	Lê Hoàng Sơn

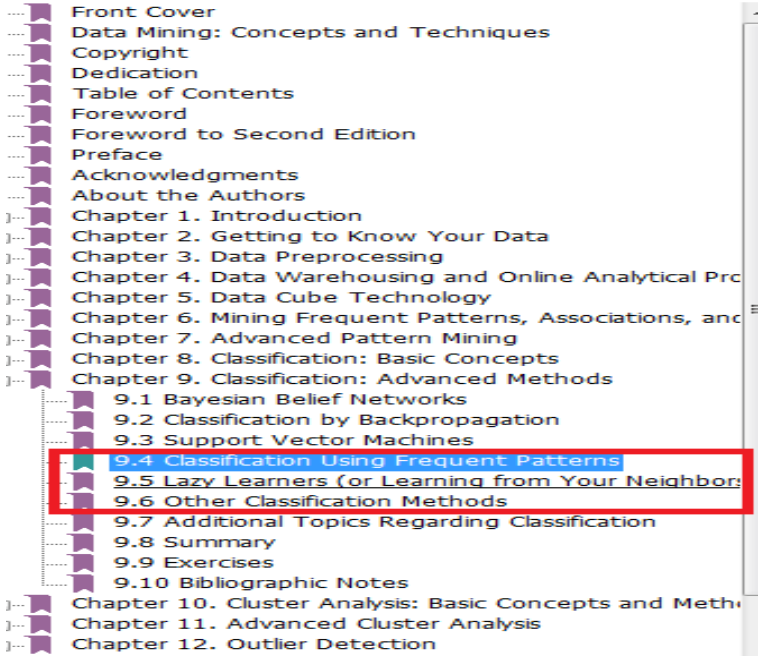
FINAL PROJECT INFORMATION

- Project 1: Face recognition
- Project 2: Video detection
- Project 3: Spam FB post
- Project 4: Cancer diagnosis
- Project 5: Customer behavior detection
- Project 6: Blog Feedback Prediction

PROJECT 1: Face recognition

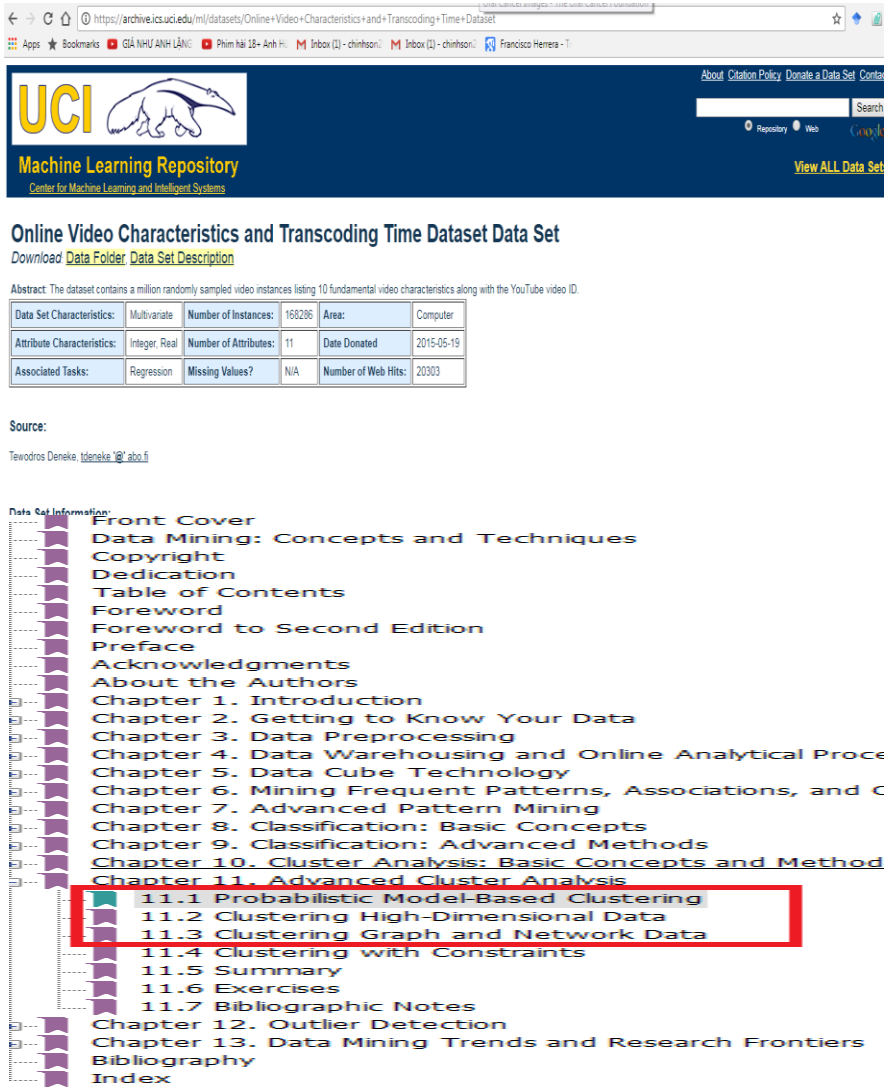
- Objective: Design a Web-based software that recognizes human face online
- Testing Datasets: **CMU Face Images Data Set** (https://archive.ics.uci.edu/ml/datasets/CMU+Face+Images)
- Requirement:
  - Implement 3 classification methods (Section 9: 9.4 - 9.6).
  - Show the Accuracy, F1 score of the methods.
  - Testing with real face datasets
- Outputs:
  - A software uploaded to https://sourceforge.net/
  - A report that presents the design, the algorithm and verification of the software followed by the standard coding document guidelines





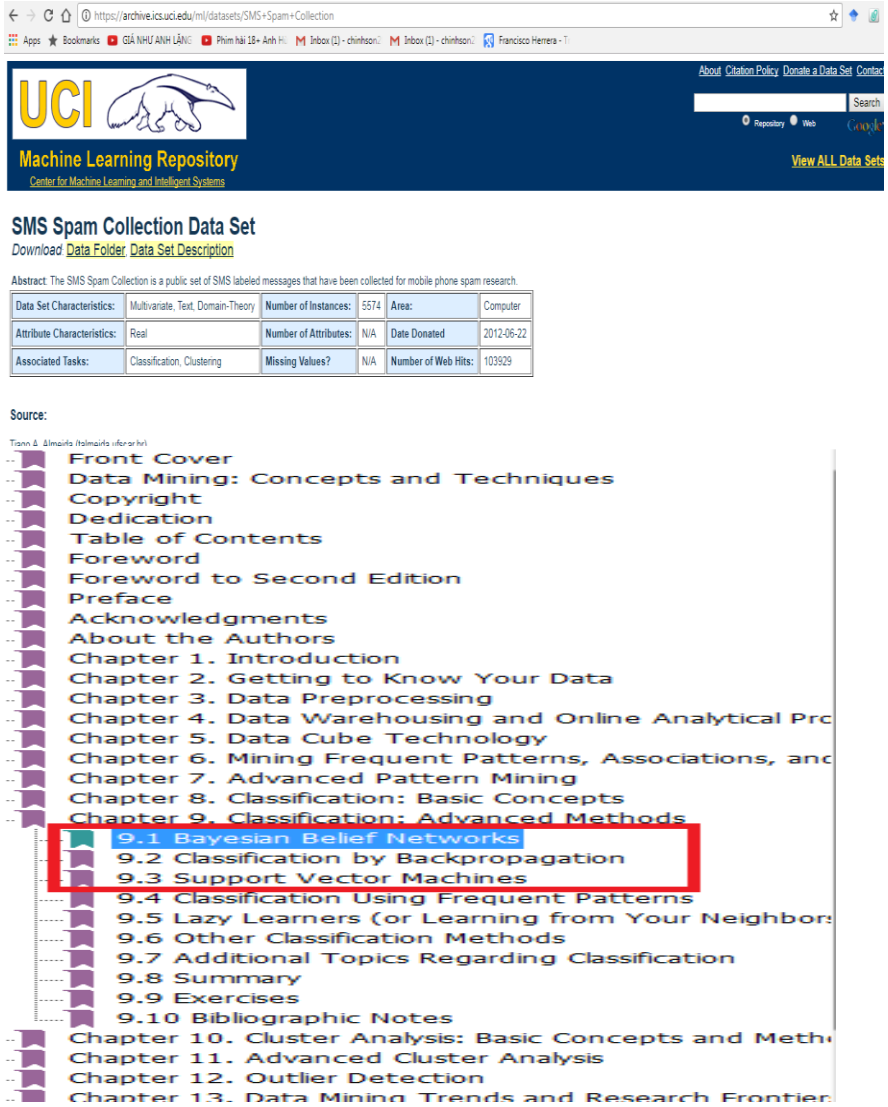
PROJECT 2: Video detection

- Objective: Design a Web-based software that detect all objects on a Youtube video
- Testing Datasets: **Online Video Characteristics and Transcoding Time Dataset Data Set** (https://archive.ics.uci.edu/ml/datasets/Online+Video+Characteristics+and+Transcoding+Time+Dataset)
- Requirement:
  - Implement 3 clustering methods (Section 11: 11.1 – 11.3).
  - Show the Running time and validity index of all methods.
  - Testing with Youtube video
- Outputs:
  - A software uploaded to https://sourceforge.net/
  - A report that presents the design, the algorithm and verification of the software followed by the standard coding document guidelines



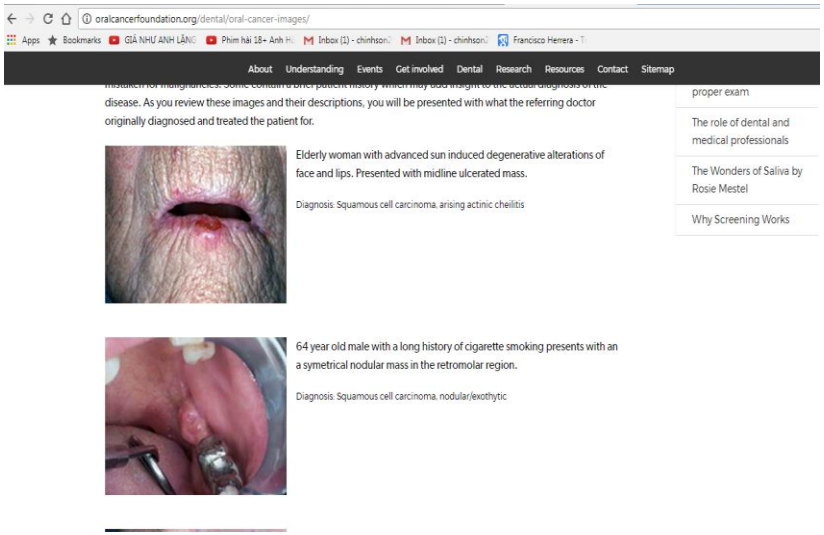
PROJECT 3: Spam FB post

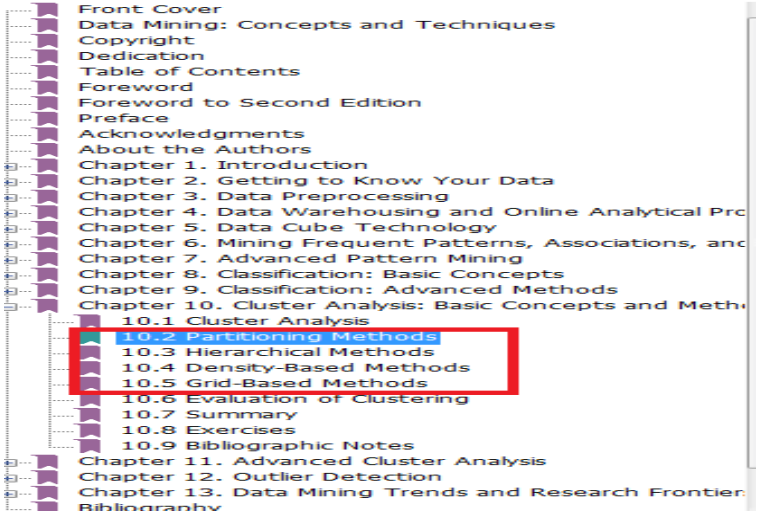
- Objective: Design a Web-based software that classify a FB post as spam or not?
- Testing Datasets: **SMS Spam Collection** (https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection)
- Requirement:
  - Implement 3 classification methods (Section 9: 9.1 - 9.3).
  - Show the Accuracy, F1 score of the methods.
  - Testing with real SMS datasets
- Outputs:
  - A software uploaded to https://sourceforge.net/
  - A report that presents the design, the algorithm and verification of the software followed by the standard coding document guidelines



PROJECT 4: Cancer diagnosis

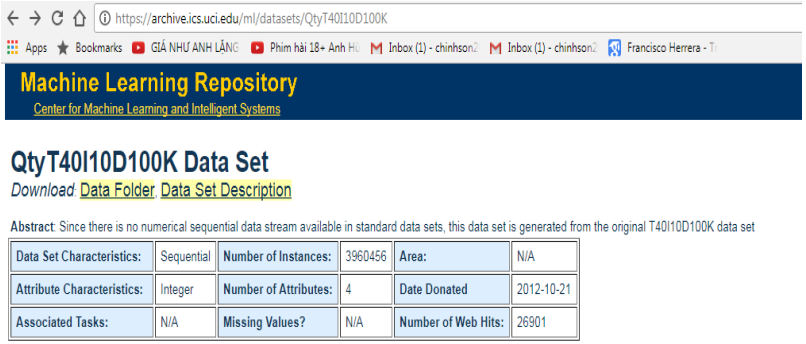
- Objective: Design a web-based software that analyze an oral image cancer or not?
- Testing Datasets: **Oral Cancer Images** (http://oralcancerfoundation.org/dental/oral-cancer-images/)
- Requirement:
  - Implement 4 clustering methods (Section 10: 10.2 – 10.5).
  - Show the Running time and validity index of all methods.
  - Testing with real datasets
- Outputs:
  - A software uploaded to https://sourceforge.net/
  - A report that presents the design, the algorithm and verification of the software followed by the standard coding document guidelines

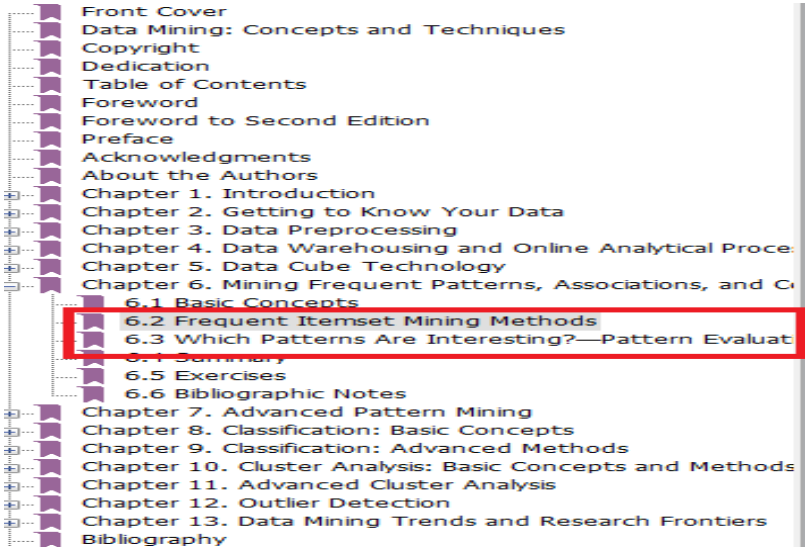




PROJECT 5: Customer behavior detection

- Objective: Design a web-based software that finds the customer behavior through buying products
- Testing Datasets: **QtyT40I10D100K Data Set** (https://archive.ics.uci.edu/ml/datasets/QtyT40I10D100K)
- Requirement:
  - Implement 2 frequent itemset methods (Section 6.2).
  - Display rules by decision trees
  - Evaluate rules
  - Testing with real datasets
- Outputs:
  - A software uploaded to https://sourceforge.net/
  - A report that presents the design, the algorithm and verification of the software followed by the standard coding document guidelines





PROJECT 6: Blog Feedback Prediction

- Objective: Design a web-based software that predict the number of comments of a Facebook posts in the upcoming 24 hours
- Testing Datasets: **BlogFeedback Data Set** (https://archive.ics.uci.edu/ml/datasets/BlogFeedback/)
- Requirement:
  - Implement regression methods (Slide).
  - Show the Running time and MSE of all methods.
  - Testing with real datasets
- Outputs:
  - A software uploaded to https://sourceforge.net/
  - A report that presents the design, the algorithm and verification of the software followed by the standard coding document guidelines

BlogFeedback Data Set

Download: Data Folder, Data Set Description

Abstract: Instances in this dataset contain features extracted from blog posts. The task associated with the data is to predict how many comments the post will receive.

Data Set Characteristics:	Multivariate	Number of Instances:	60021	Area:	Social
Attribute Characteristics:	Integer, Real	Number of Attributes:	281	Date Donated	2014-05-29
Associated Tasks:	Regression	Missing Values?	N/A	Number of Web Hits:	42332

Source:

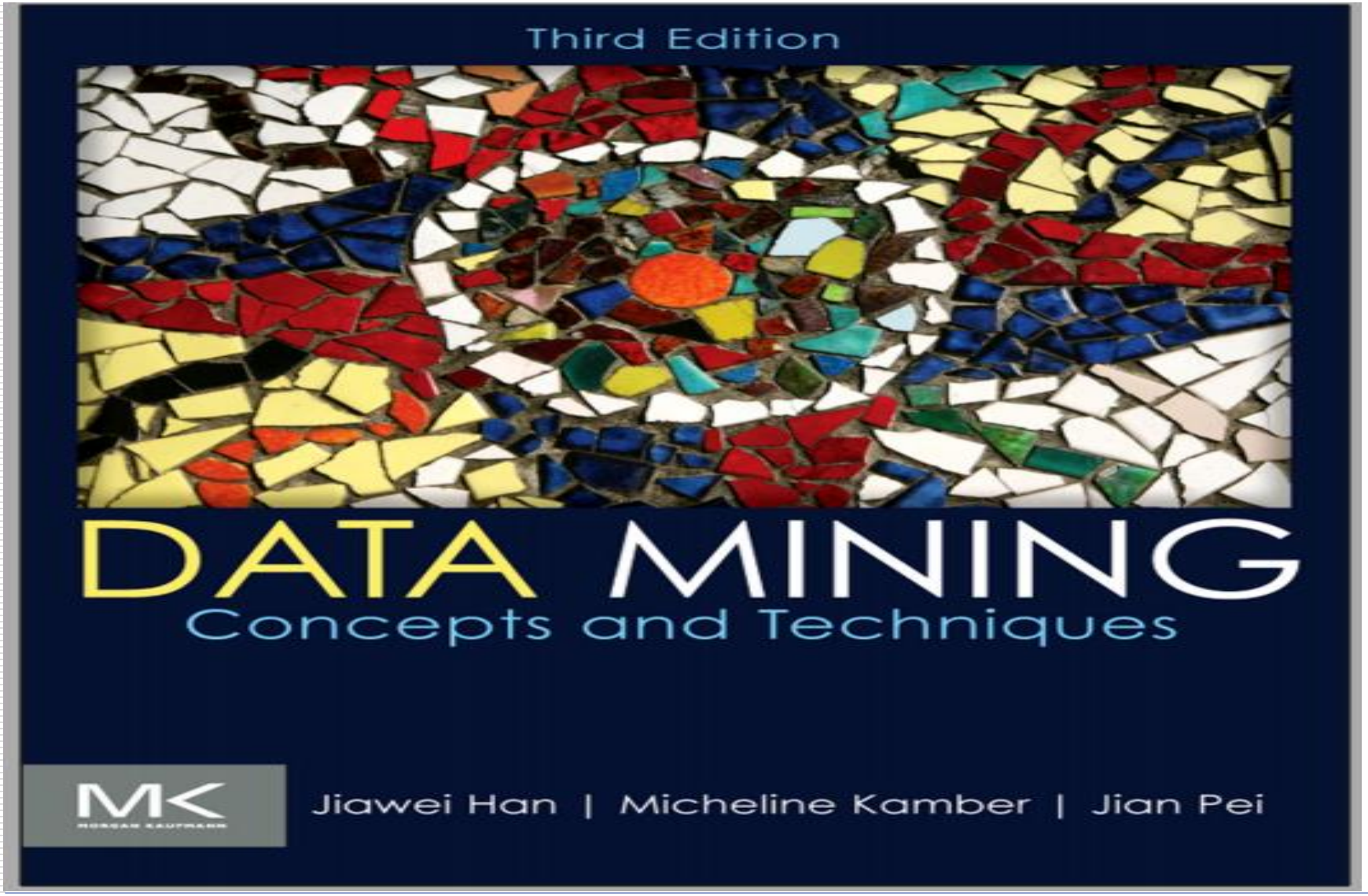
Krisztian Buza
Budapest University of Technology and Economics
buza '@' cs.bme.hu
http://www.cs.bme.hu/~buza

FINAL RESEARCH TOPIC

- Topic 1: Nền tảng tính toán mềm (Soft Computing Foundation)
- Topic 2: Hệ xử lý tri thức với dữ liệu lớn (Knowledge-Based Systems)
- Topic 3: Học máy tích hợp trong đa phương tiện thông minh (Integrated Machine Learning for Multimedia Intelligence)
- Topic 4: Ứng dụng AI đa môi trường (Multi-modal & environmental AI)
- Works-to-do:

1) Read papers
2) Propose solutions and implement
3) Write a report

Reference

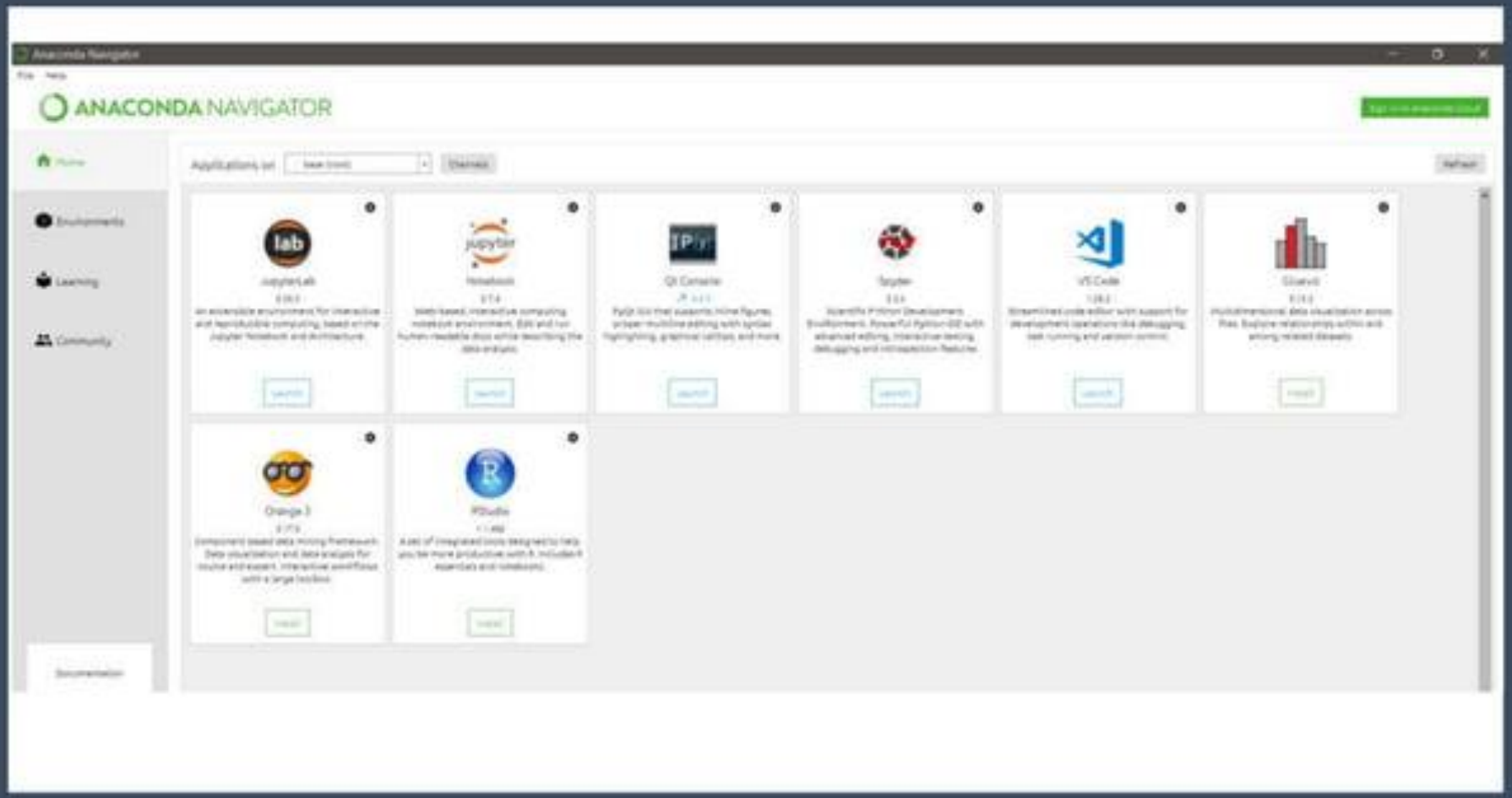

Third Edition
DATA MINING
Concepts and Techniques
Jiawei Han | Micheline Kamber | Jian Pei

LIST TOOLS FOR PRACTISING

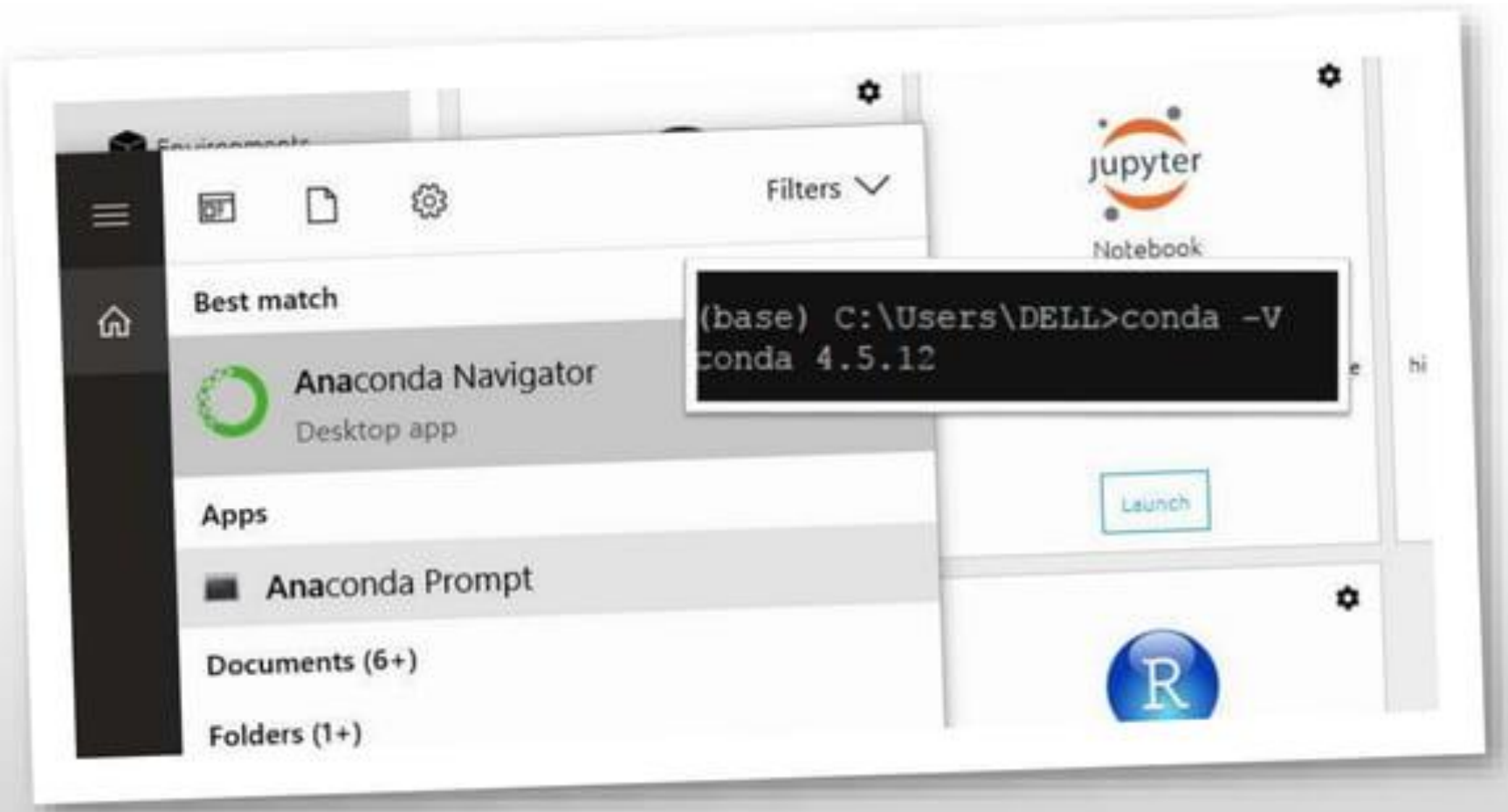
- Python https://www.python.org/downloads/
- Anacoda https://www.anaconda.com/
- SublimeText: https://www.sublimetext.com/

Installing Anaconda

DATA MINING



Choose Anaconda Navigator



Choose Powershell Prompt

TensorFlow

Installing TensorFlow

Guidelines and screen shots of each step

conda create --name tensorflow python=3.7

```
(base) F:\thongph\Giang day\datamining\test>conda create --name tensorflow python=3.7
Collecting package metadata (current_repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: E:\Anacoda\envs\tensorflow

  added / updated specs:
    - python=3.7


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    ca-certificates-2023.01.10 |       haa95532_0         121 KB
    certifi-2022.12.7          |   py37haa95532_0         149 KB
    openssl-1.1.1s             |       h2bbff1b_0         5.5 MB
    pip-22.3.1                 |   py37haa95532_0         2.7 MB
    python-3.7.16              |       h6244533_0        17.2 MB
    setuptools-65.6.3          |   py37haa95532_0         1.1 MB
    sqlite-3.40.1              |       h2bbff1b_0         889 KB
    wheel-0.37.1               |     pyhd3eb1b0_0          33 KB
    wincertstore-0.2           |   py37haa95532_2          15 KB
    ------------------------------------------------------------
                                           Total:        27.7 MB

The following NEW packages will be INSTALLED:

  ca-certificates    pkgs/main/win-64::ca-certificates-2023.01.10-haa95532_0
  certifi            pkgs/main/win-64::certifi-2022.12.7-py37haa95532_0
  openssl            pkgs/main/win-64::openssl-1.1.1s-h2bbff1b_0
  pip                pkgs/main/win-64::pip-22.3.1-py37haa95532_0
  python             pkgs/main/win-64::python-3.7.16-h6244533_0
  setuptools         pkgs/main/win-64::setuptools-65.6.3-py37haa95532_0
  sqlite             pkgs/main/win-64::sqlite-3.40.1-h2bbff1b_0
  vc                 pkgs/main/win-64::vc-14.2-h21ff451_1
  vs2015_runtime     pkgs/main/win-64::vs2015_runtime-14.27.29016-h5e58377_2
  wheel              pkgs/main/noarch::wheel-0.37.1-pyhd3eb1b0_0
  wincertstore       pkgs/main/win-64::wincertstore-0.2-py37haa95532_2


Proceed ([y]/n)?
```

conda activate tensorflow

```
(base) F:\thongph\Giang day\datamining\test>conda activate tensorflow

(tensorflow) F:\thongph\Giang day\datamining\test>_
```

pip3 install tensorflow

Create file hellow.py:

```
import tensorflow as tf
msg = tf.constant('Hello, TensorFlow!')
tf.print(msg)
```

Then run: python hellow.py:

```
(tensorflow) F:\thongph\Giang day\datamining\test>python hellow.py
2023-02-05 11:53:36.828851: I tensorflow/core/platform/cpu_feature_guard.cc:193] This TensorFlow binar
y is optimized with oneAPI Deep Neural Network Library (oneDNN) to use the following CPU instructions
in performance-critical operations:  AVX AVX2
To enable them in other operations, rebuild TensorFlow with the appropriate compiler flags.
Hello, TensorFlow!
```

Thank You !