



MACHINE LEARNING & DATA MINING

GV. Vũ Đức Thịnh





Nội dung

Chương 1. Giới thiệu chung

Chương 2. Thu thập và tiền xử lý dữ liệu

Chương 3. Hồi quy

Chương 4. Phân cụm

Chương 5. Phân cụm phân cấp

Chương 6. Học dựa trên láng giềng (KNN)

Chương 7. Cây quyết định và rừng ngẫu nhiên

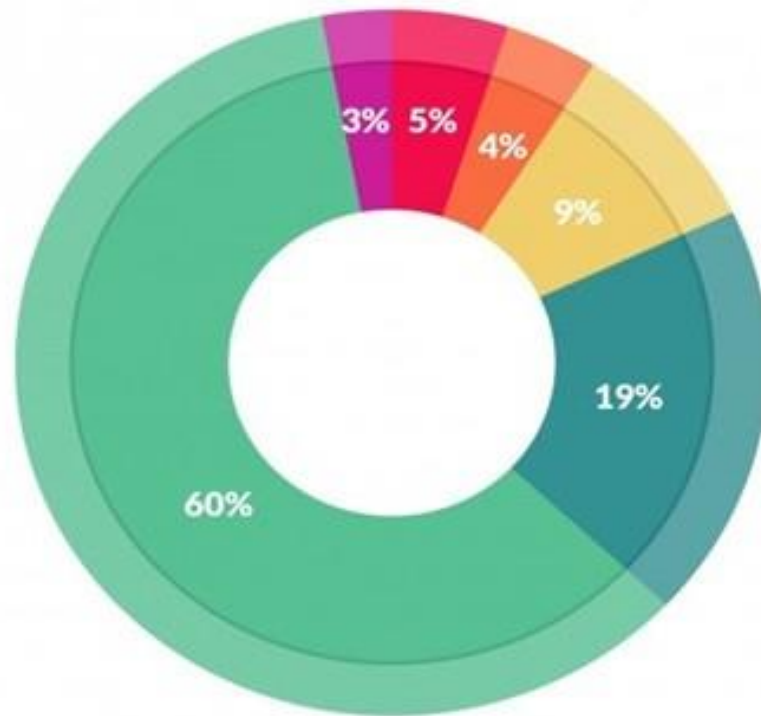
Chương 8. Máy vecto hỗ trợ (SVM)

Chương 9. Đánh giá hiệu quả của mô hình

Chương 10. Mạng nowrowrron nhân tạo

Chương 11. Mô hình xác xuất

Quỹ thời gian



CrowdFlower Inc., 2016

- Thời gian dành cho phân tích dữ liệu ra sao?
 - Thu thập dữ liệu: 19%
 - Thu xếp và làm sạch dữ liệu: 60%
 - Tạo tập dữ liệu huấn luyện: 3%
 - Khai phá: 9%
 - Cải thiện thuật toán: 4%
 - Khác: 5%

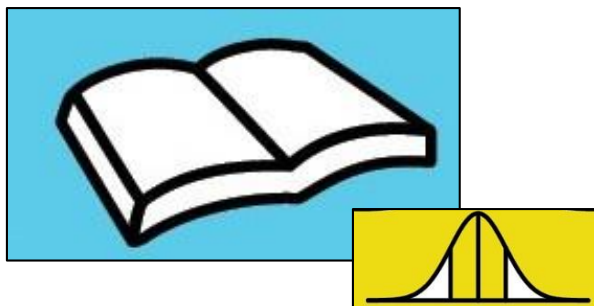
Why?

■ Tiền xử lý để làm gì

- Thuận tiện trong lưu trữ, truy vấn
- Các mô hình học máy thường làm việc với dữ liệu có cấu trúc: ma trận, vectơ, chuỗi,...
- Học máy thường làm việc hiệu quả nếu có **biểu diễn dữ liệu phù hợp**

Input

Vấn đề cần giải quyết
của
lĩnh vực



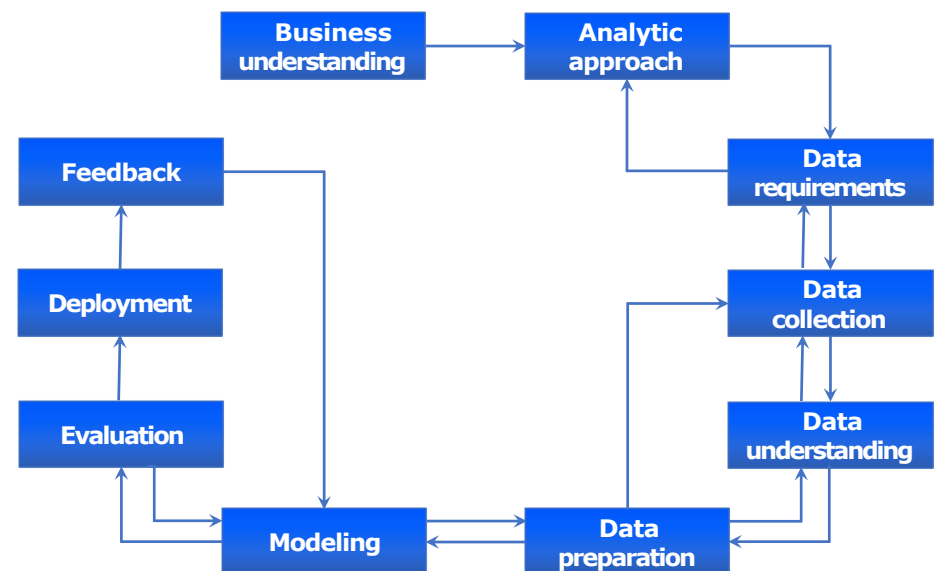
Output

Dữ liệu số - ma trận
vector

$$x^{(n)} = \begin{bmatrix} -0.0920 \\ 3.4931 \\ -1.8493 \\ \dots \\ \dots \\ -0.2010 \\ -1.3079 \end{bmatrix} \quad \mathcal{D} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \dots \\ x^{(n)} \end{bmatrix}$$

How?

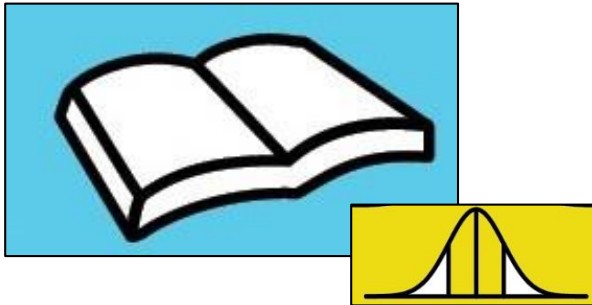
- Thu thập dữ liệu
 - Lấy mẫu (sampling)
 - Kỹ thuật: crawling, logging, scraping
- Xử lý dữ liệu
 - Lọc nhiễu, làm sạch, số hoá,...



Data collection

Input

Vấn đề cần giải quyết



Output

Mẫu dữ liệu

	A	B	C	D	E	F	G
1	Country	Region	Population	Under15	Over60	Fertil	LifeExp
2	Zimbabwe	Africa	13724	40.24	5.68	3.64	54
3	Zambia	Africa	14075	46.73	3.95	5.77	55
4	Yemen	Eastern M	23852	40.72	4.54	4.35	64
5	Viet Nam	Western P	90796	22.87	9.32	1.79	75
6	Venezuela (Bo	Americas	29955	28.84	9.17	2.44	75
7	Vanuatu	Western P	247				
8	Uzbekistan	Europe	28541				
9	Uruguay	Americas					

The image is a screenshot of a Wikipedia page. It shows the 'Welcome to Wikipedia' message at the top, followed by a 'From today's featured article' section. The featured article is about the London Transport Museum, mentioning its location in London and its collection of transport-related items. There are also links to 'In the news' and 'Did you know' sections.

Fundamentals :: Sampling

- **WHAT** - lấy tập mẫu nhỏ, phổ biến để đại diện cho lĩnh vực cần học.
- **WHY** - không thể học toàn bộ. Giới hạn về thời gian và khả năng tính toán
- **HOW** - thu thập các mẫu từ thực tế, hoặc các nguồn chứa dữ liệu web, database,...

"One or more small spoon(s) can be enough to assess whether the soup is good or not."



<https://www.coursera.org/learn/inferential-statistics-intro>

Fundamentals :: Sampling :: How

- **Variety** - tập mẫu thu được đủ đa dạng để phủ hết các ngữ cảnh của lĩnh vực.
- **Bias** - dữ liệu cần tổng quát, không bị sai lệch, thiên vị về 1 bộ phận nhỏ nào đó của lĩnh vực.

"One or more small spoon(s) can be enough to assess whether the soup is good or not."

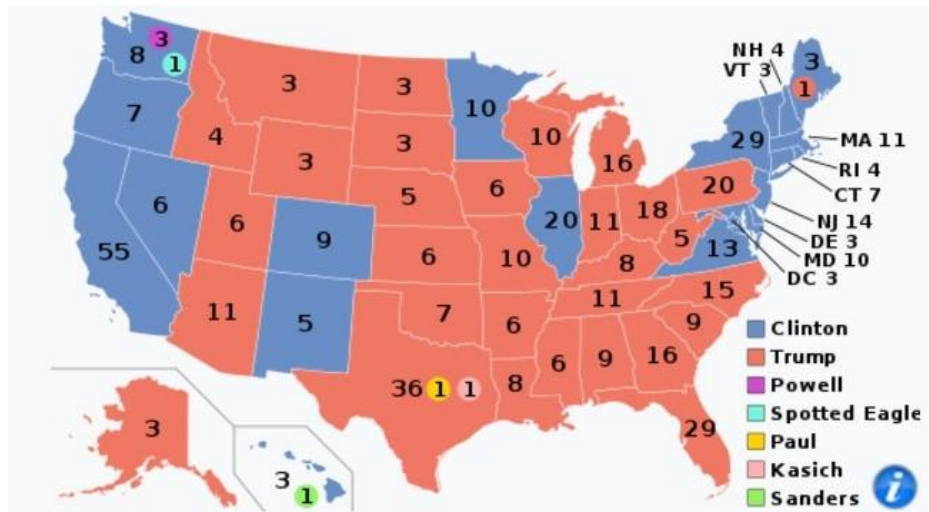
Remember to stir to avoid tasting biases.



<https://www.coursera.org/learn/inferential-statistics-intro>

Fundamentals :: Sampling :: How

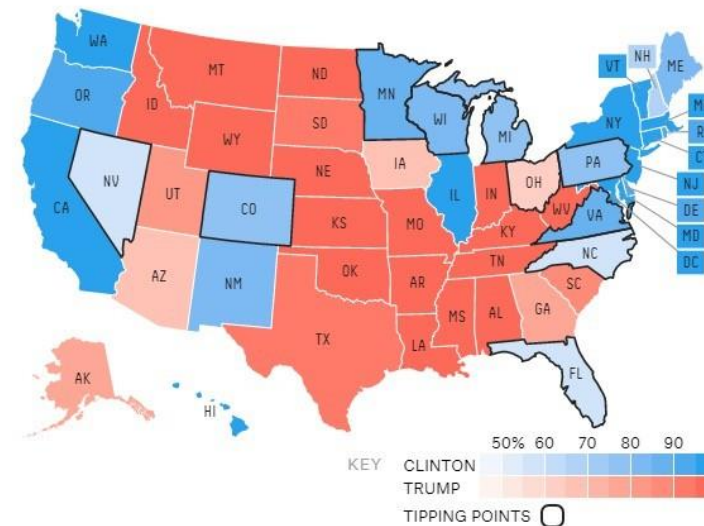
- **Variety** - các mẫu đủ đa dạng để phản ánh khách quan?



Actual results

<https://projects.fivethirtyeight.com/2016-election-forecast/>
<http://edition.cnn.com/election/results/president>
 Image credit: Wikipedia, FiveThirtyEight

Chance of winning



Electoral votes

Hillary Clinton	302 . 2
Donald Trump	235 . 0

Popular vote

Hillary Clinton	48 . 5%
Donald Trump	44 . 9%

ro

Techniques

- **Crowd-sourcing:** Survey - *thực hiện các khảo sát*
- **Logging:** lưu lại lịch sử tương tác của người dùng, truy cập sản phẩm,...
- **Scrapping:** tìm kiếm nguồn dữ liệu trên các website, tải về, bóc tách, lọc,...

Data preprocessing

Input

Mẫu dữ liệu thô
(text, ảnh, audio, ...)

	A	B	C	D	E	F	G
1	Country	Region					
2	Zimbabwe	Africa					
3	Zambia	Africa					
4	Yemen	Eastern M					
5	Viet Nam	Western P					
6	Venezuela (Bo	Americas					
7	Vanuatu	Wester					
8	Uzbekistan	Europe					
9	Uruguay	America					



Output

Dữ liệu số theo từng
ML/AI model(s)

$$x^{(n)} = \begin{bmatrix} -0.0920 \\ 3.4931 \\ -1.8493 \\ \dots \\ \dots \\ -0.2010 \\ -1.3079 \end{bmatrix} \quad \mathcal{D} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \dots \\ x^{(n)} \end{bmatrix}$$

Fundamentals :: Data “rawness”

Completeness (đầy đủ)

Từng mẫu thu thập nên đầy đủ thông tin các trường thuộc tính cần thiết

Integrity (trung thực)

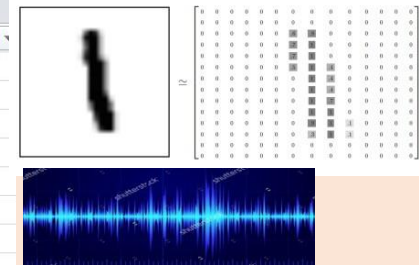
- Nguồn thu thập chính thống, đảm bảo mẫu thu được chứa giá trị chính xác trên thực tế.
- Jan. 1 as *everyone's* birthday? – *intentional (systematic) noises*

Homogeneity (đồng nhất)

- Rating “1, 2, 3” & “A, B, C”; or Age = “42” & Birthday = “03/07/2010” (*inconsistency*)
- Heterogenous data sources / schemas

Structures (cấu trúc)

C	D	E	F
Population	Under15	Over60	Fertil
13724	40.24	5.68	3.64
14075	46.73	3.95	5.77
23852	40.72	4.54	4.35
90796	22.87	9.32	1.79
29955	28.84	9.17	2.44
247	37.37	6.02	3.46
28541	28.9	6.38	2.38
3395	22.05	18.59	2.07



Techniques

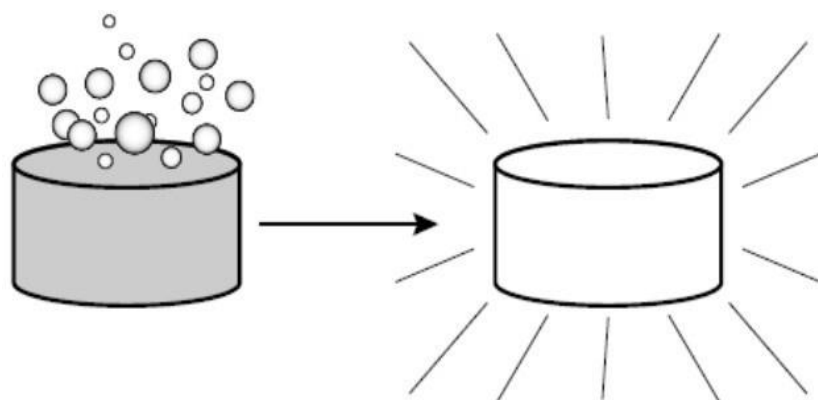
Cleaning

Integrating

Transforming

Techniques :: Cleaning

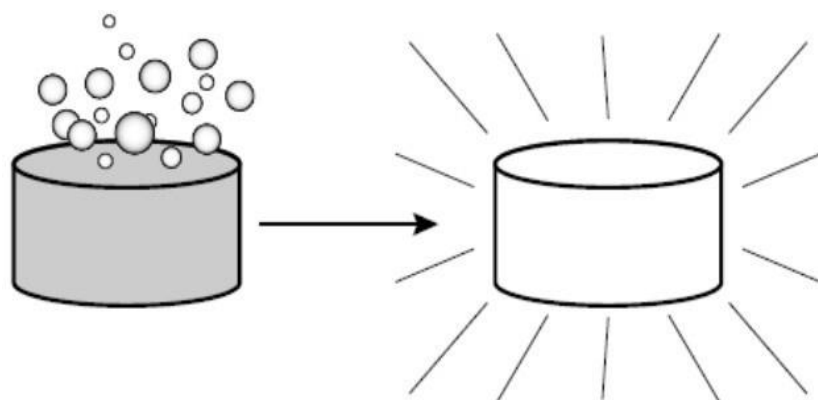
■ Tính đầy đủ + trung thực



- Mẫu dữ liệu cần được thu thập từ các **nguồn đáng tin cậy**. Phản ánh vấn đề cần giải quyết.
- Loại bỏ **nhiều** (ngoại lai): bỏ vài mẫu dữ liệu mà có khác biệt lớn với các mẫu khác.
- Một mẫu dữ liệu có thể **bị trống** (thiếu, chưa đầy đủ), cần có chiến lược phù hợp:
 - Bỏ qua, không đưa vào phân tích?
 - Bổ sung các trường còn thiếu cho mẫu?

Techniques :: Cleaning

■ Điền giá trị thiếu

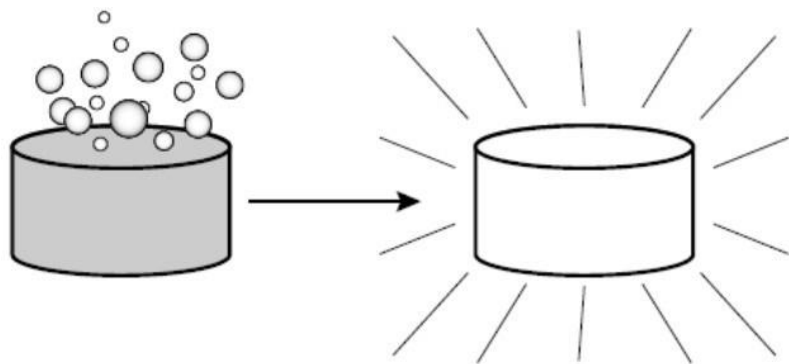


- Điền lại giá trị bằng tay
- Gán cho giá trị nhãn đặc biệt hay ngoài khoảng biểu diễn
- Gán giá trị trung bình cho nó.
- Gán giá trị trung bình của các mẫu khác thuộc cùng lớp đó.
- Tìm giá trị có xác suất lớn nhất điền vào chỗ bị mất (hồi quy, suy diễn Bayes,...)

A1	A2	A3	A4	A5	A6	A7	A8	y
?	3.683	?	-0.634	1	0.409	7	30	5
?	?	60	1.573	0	0.639	7	30	5
?	3.096	67	0.249	0	0.089	?	80	3
2.887	3.870	68	-1.347	?	1.276	?	60	5
2.731	3.945	79	1.967	1	2.487	?	100	4

Techniques :: Cleaning (cont.)

- Tính đồng nhất



Các mẫu dữ liệu cần có tính đồng nhất về cách biểu diễn, ký hiệu.

Ví dụ không đồng nhất:

Rating “1, 2, 3” & “A, B,

C”;

Age = 42 & Birthday = 03/08/2020

Techniques :: Integrating w/ some Transforming

Un-structured

	A	B	C	D	E	F	G
1	Country	Region	Population	Under15	Over60	Fertil	LifeExp
2	Zimbabwe	Africa	13724	40.24	5.68	3.64	54
3	Zambia	Africa	14075	46.73	3.95	5.77	55
4	Yemen	Eastern M	23852	40.72	4.54	4.35	64
5	Viet Nam	Western P	90796	22.87	9.32	1.79	75
6	Venezuela (Bo	Americas	29955	28.84	9.17	2.44	75
7	Vanuatu	Western P	247	37.37	6.02	3.46	72
8	Uzbekistan	Europe	28541	28.9	6.38	2.38	68
9	Uruguay	Americas	3395	22.05	18.59	2.07	77

```
{
  "code": "1473a6fd39d1d8fa48654aac9d8cc2754232",
  "title": "[Updating] Câu chuyện xuyên mưa về :",
  "url": "http://techtalk.vn/updating-cau-chuyen",
  "labels": "techtalk/Cong nghe",
  "content": "Vào chiều tối ngày 09/12/2016 vừa",
  "image_url": "",
  "date": "2016-12-10T03:51:10Z"
}
```

texts in websites, emails, articles, tweets



2D/3D images, videos + meta



spectrograms, DNAs, ...

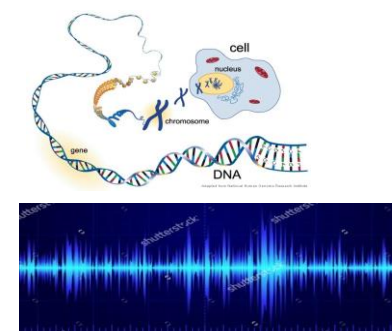


image credits: wikipedia, shutterstock, CNN

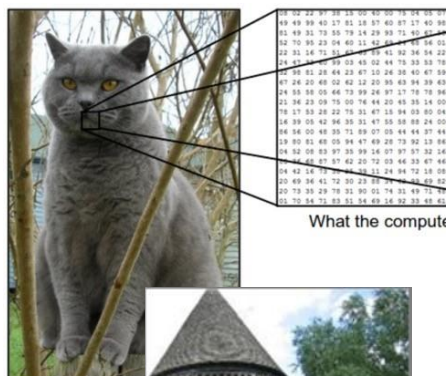
Techniques :: Transforming

Semantics?

Trích xuất các **đặc trưng ngữ nghĩa**, chuẩn hóa

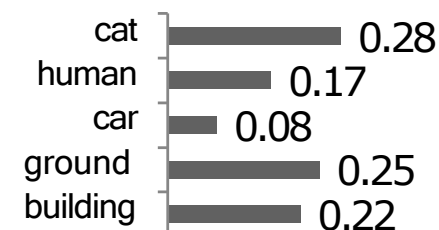
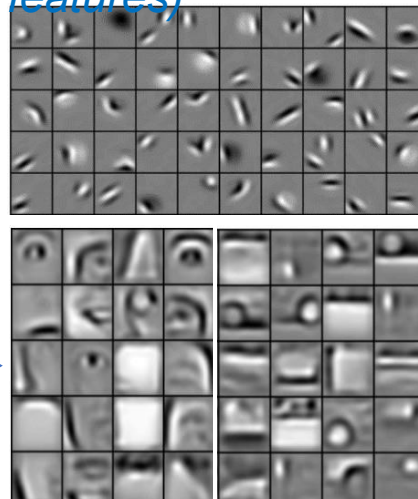
Semantics example: visual data

Low-level semantics
(raw pixels)



What the computer

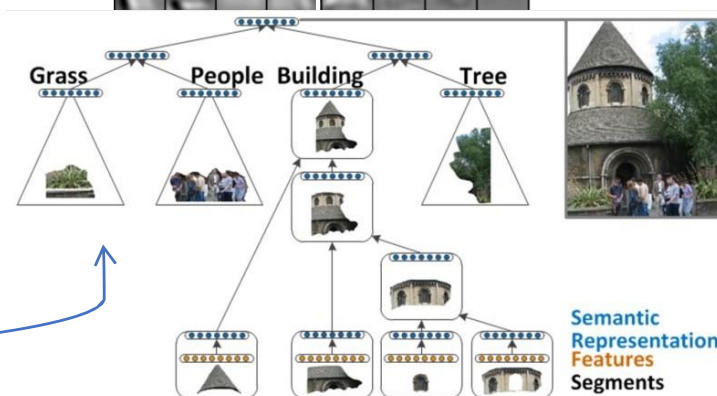
Mid-/High-level semantics
(e.g. human-interpretable features)



cat → **not on** → car
people ← **behind** ← building
car → **is** → red

Mức ngữ nghĩa tối thiểu để có thể hiểu:

- Phân loại văn bản
- Phân tích cảm xúc
- AI Chatbot (nhiều mức ngữ nghĩa khác nhau)



C	D	E	F
Population	Under15	Over60	Fertil
13724	40.24	5.68	3.64
14075	46.73	3.95	5.77
23852	40.72	4.54	4.35
90796	22.87	9.32	1.79
29955	28.84	9.17	2.44
247	37.37	6.02	3.46
28541	28.9	6.38	2.38
3395	22.05	18.59	2.07

Image credits: CS231n, Stanford University; Lee et al, 2009; Socher et al, 2011

Summary

(Take-home messages)

30

- Dữ liệu trong một lĩnh vực trước khi vào hệ thống học máy phải được thu thập và biểu diễn thành dạng cấu trúc với một số đặc tính: đầy đủ, ít nhiễu, nhất quán, có cấu trúc xác định.
- Dữ liệu thu thập cho quá trình học là tập nhỏ, tuy vậy cần phản ánh đầy đủ các mặt vấn đề cần giải quyết.
- Dữ liệu thô sau khi thu thập và tiền xử lý phải giữ được sự đầy đủ các đặc trưng ngữ nghĩa - các đặc trưng ảnh hưởng đến khả năng giải quyết vấn đề.
- Khoa học dữ liệu là một lĩnh vực rộng, ngoài việc sử dụng công cụ áp dụng, nắm vững được các kiến thức cơ bản là điều quan trọng.

Hướng dẫn cài đặt môi trường chạy demo các thuật toán ML&DM

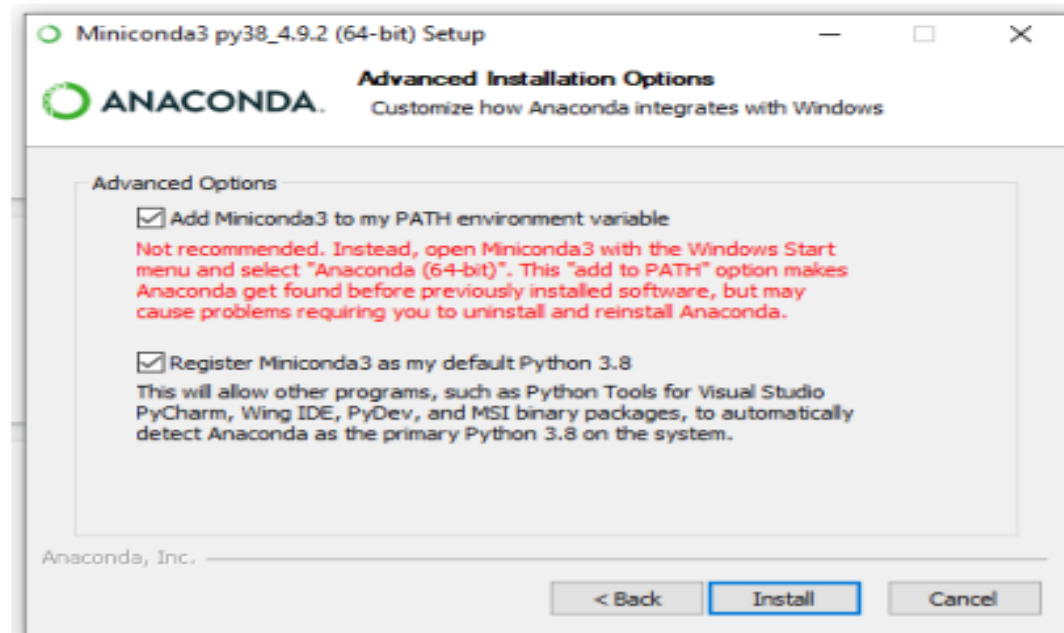
1. Windows 10

1.1. Các yêu cầu về phần cứng và phần mềm

- Máy chạy Windows 10, 64 bits
- Có quyền admin

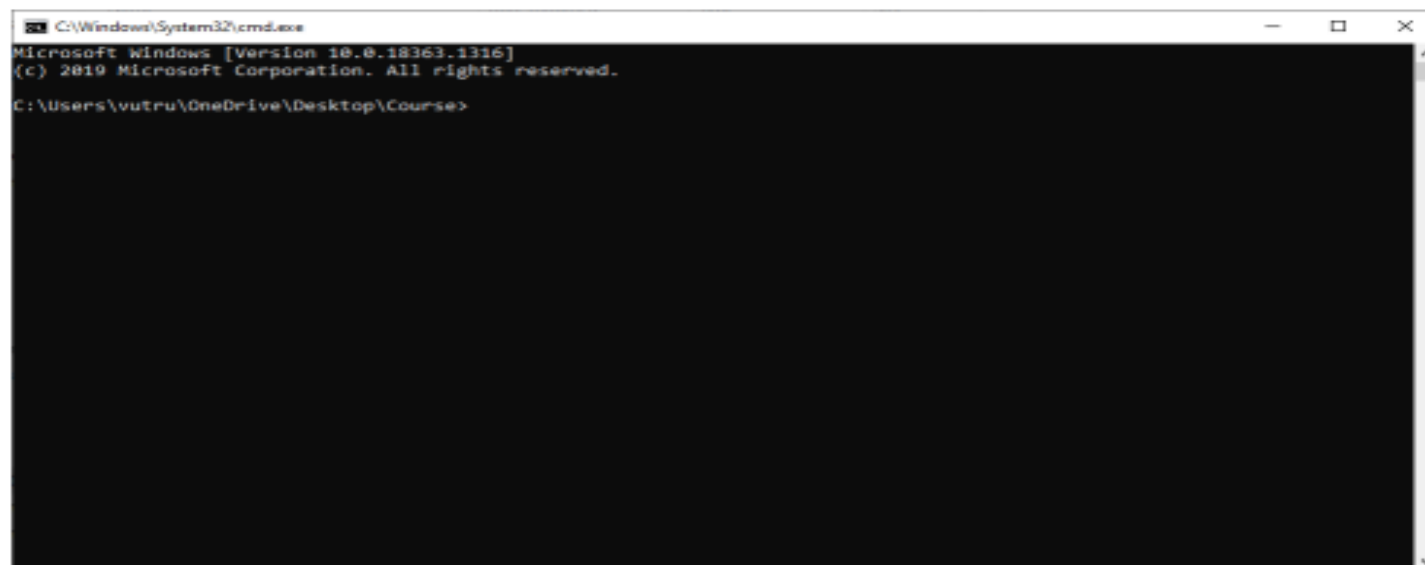
1.2. Cài đặt Miniconda

- Tải Miniconda tại link sau: https://repo.anaconda.com/miniconda/Miniconda3-latest-Windows-x86_64.exe
- Sau khi tải về chạy file .exe và chọn theo hướng dẫn. Lưu ý cần tích vào ô đầu tiên như hình dưới để thêm conda vào biến môi trường.



1.4. Cài đặt môi trường python và thư viện

- Vào thư mục vừa giải nén. Sau đó nhấn tổ hợp phím **Ctrl + L**
- Xóa dòng chữ được bôi đậm đi thay bằng **cmd** và chọn **Enter**
- Sau bước này hiện ra một giao diện cửa sổ dòng lệnh như hình bên dưới (không cần giống hoàn toàn vì cấu hình và vị trí thư mục các máy là khác nhau)



```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.18363.1316]
(c) 2019 Microsoft Corporation. All rights reserved.
C:\Users\vutru\OneDrive\Desktop\Course>
```

- Nhập câu lệnh sau: **conda create -n course python=3.8**
- Sau đó chọn **Enter**, tiếp đó nhập **y** và chọn **Enter** để xác nhận cài đặt
- Sau khi cài đặt xong chạy câu lệnh sau để vào môi trường vừa được tạo ra: **conda activate course**
- Chạy câu lệnh sau để cài đặt thư viện: **pip install -r requirements.txt**
- Sau khi cài đặt xong thì tắt cửa sổ đi

1.5. Chạy chương trình trên Jupyter Lab

- Vào thư mục **Course** vừa giải nén phía trên. Sau đó nhấn tổ hợp phím **Ctrl + L**
- Xóa dòng chữ được bôi đậm đi thay bằng **cmd** và chọn **Enter**
- Chạy câu lệnh sau để vào môi trường cài đặt phía trên: **conda activate course**
- Chạy lệnh sau để vào giao diện jupyter lab: **jupyter lab**


```

C:\Users\vutru\OneDrive\Desktop\Course>conda create --name course python=3.8
Collecting package metadata (current_repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: C:\Users\vutru\miniconda3\envs\course

added / updated specs:
- python=3.8

The following packages will be downloaded:

package                                     build                                     122 KB
ca-certificates-2021.1.19                 haa95532_0
certifi-2020.12.5                         py38haa95532_0
openssl-1.1.1i                            h2bfff1b_0
pip-20.3.3                                py38haa95532_0
setuptools-51.3.3                         py38haa95532_4
vc-14.2                                    h21ff451_1
vs2015_runtime-14.27.29016                h5e58377_2
wheel-0.36.2                              pyhd3eb1b0_0
Total: 8.6 MB

The following NEW packages will be INSTALLED:

ca-certificates pkgs/main/win-64::ca-certificates-2021.1.19-haa95532_0
certifi          pkgs/main/win-64::certifi-2020.12.5-py38haa95532_0
openssl         pkgs/main/win-64::openssl-1.1.1i-h2bfff1b_0
pip             pkgs/main/win-64::pip-20.3.3-py38haa95532_0
python          pkgs/main/win-64::python-3.8.5-h5fd99cc_1
setuptools      pkgs/main/win-64::setuptools-51.3.3-py38haa95532_4
sqlite         pkgs/main/win-64::sqlite-3.33.0-h2a8f88b_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-0.36.2-pyhd3eb1b0_0
wincertstore   pkgs/main/win-64::wincertstore-0.2-py38_0
zlib           pkgs/main/win-64::zlib-1.2.11-h62dcd97_4

Proceed ([y]/n)? y

```

```

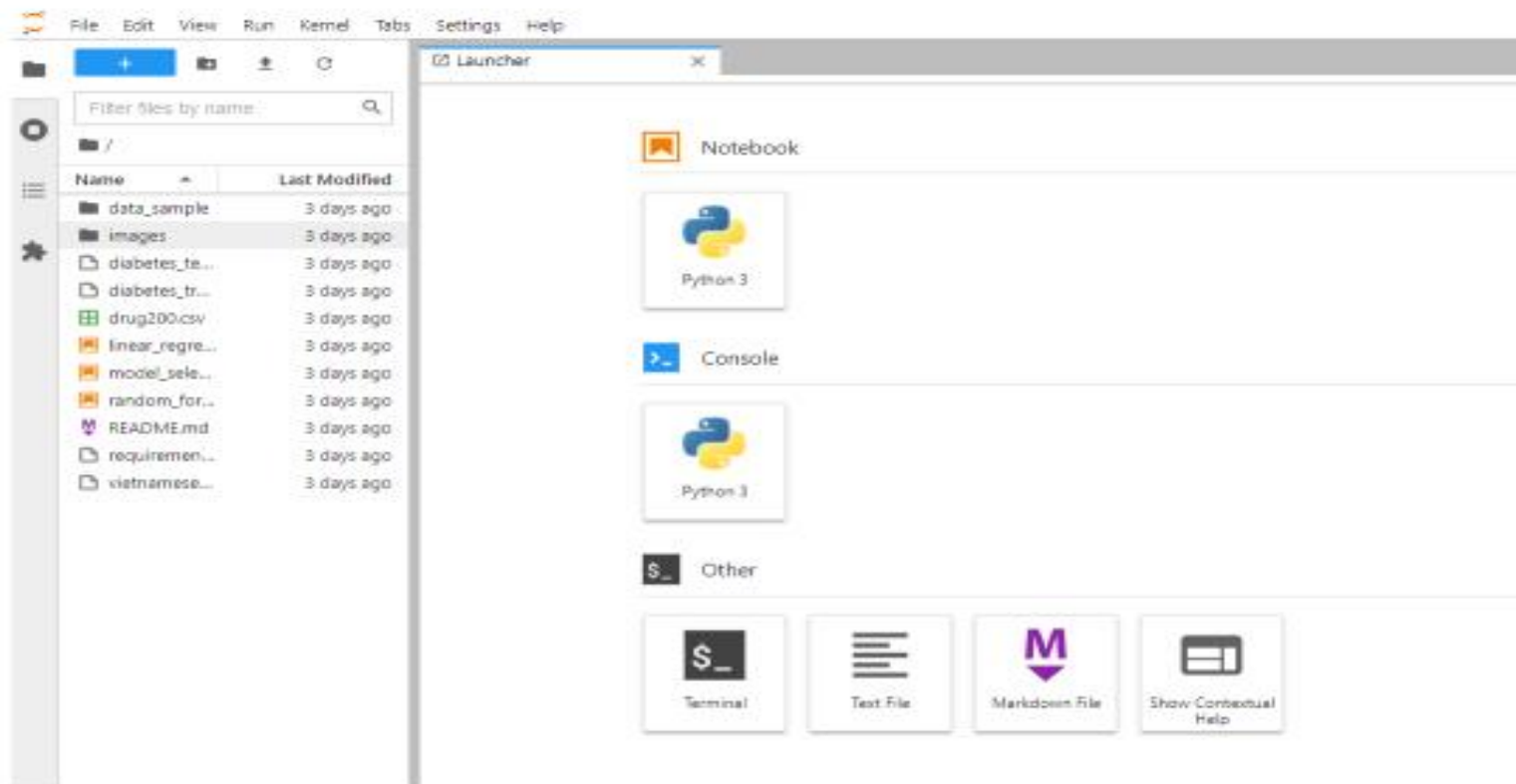
(course) C:\Users\vutru\OneDrive\Desktop\Course>jupyter lab
[I 2021-01-23 18:25:58.904 ServerApp] jupyterlab | extension was successfully linked.
[I 2021-01-23 18:25:58.982 ServerApp] Writing notebook server cookie secret to C:\Users\vutru\AppData\Roaming\jupyter\runtime\jupyter_cookie_secret
[W 2021-01-23 18:25:58.998 ServerApp] The 'min_open_files_limit' trait of a ServerApp instance expected an int, not the NoneType None.
[I 2021-01-23 18:25:59.040 LabApp] JupyterLab extension loaded from c:\Users\vutru\miniconda3\envs\course\lib\site-packages\jupyterlab
[I 2021-01-23 18:25:59.041 LabApp] JupyterLab application directory is c:\Users\vutru\miniconda3\envs\course\share\jupyter\lab
[I 2021-01-23 18:25:59.047 ServerApp] jupyterlab | extension was successfully loaded.
[I 2021-01-23 18:25:59.386 ServerApp] nbclassic | extension was successfully loaded.
[I 2021-01-23 18:25:59.387 ServerApp] Serving notebooks from local directory: C:\Users\vutru\OneDrive\Desktop\Course
[I 2021-01-23 18:25:59.388 ServerApp] Jupyter Server 1.2.2 is running at:
[I 2021-01-23 18:25:59.392 ServerApp] http://localhost:8888/lab?token=e1296069b1785526c6380247c40fd1c3dd5ba7cdf5dffffa
[I 2021-01-23 18:25:59.393 ServerApp] or http://127.0.0.1:8888/lab?token=e1296069b1785526c6380247c40fd1c3dd5ba7cdf5dffffa
[I 2021-01-23 18:25:59.393 ServerApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).

[C 2021-01-23 18:25:59.448 ServerApp]

To access the server, open this file in a browser:
File:///C:/Users/vutru/AppData/Roaming/jupyter/runtime/jpserver-10156-open.html
Or copy and paste one of these URLs:
http://localhost:8888/lab?token=e1296069b1785526c6380247c40fd1c3dd5ba7cdf5dffffa
or http://127.0.0.1:8888/lab?token=e1296069b1785526c6380247c40fd1c3dd5ba7cdf5dffffa
[I 2021-01-23 18:26:03.917 LabApp] Build is up to date

```

Hình 1: Kết quả sau khi chạy lệnh `jupyter lab`



Hình 2: Giao diện jupyter lab

2. Ubuntu 20.04 LTS

Việc cài đặt trên Ubuntu 20.04 LTS tương tự như với Windows 10, chỉ khác ở bước tải và cài đặt miniconda. Tải file cài đặt miniconda tại: https://repo.anaconda.com/miniconda/Miniconda3-py38_4.10.3-Linux-x86_64.sh

Cấp quyền thực thi cho file này bằng lệnh: **chmod +x Miniconda3-py38_4.10.3-Linux-x86_64.sh**

Thực hiện cài đặt bằng cách chạy file trên: **./Miniconda3-py38_4.10.3-Linux-x86_64.sh**

Sau đó mở file: **/home/{username}/.bashrc** và thêm thông tin về conda như sau vào cuối file:

PATH=/home/{username}/miniconda3/bin:\$PATH

Chuyển hướng vào thư mục **/home/{username}** và load lại file mới được cập nhật bằng lệnh: **source .bashrc**

Các bước còn lại thực hiện tương tự so với Windows 10.