Course Introduction
Programming for Data Science

fit@hcmus

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

We do data science activities in daily life

Should I register for this course?

- Collect data
  - Syllabus
  - Reviews of students who have learned the course
  - …
- Run some analysis algorithm in the head
- Make decision

Is data correct and enough?

Should I buy this product on Tiki?

- Collect data: reviews
- Run some analysis algorithm in the head
- Make decision

blank line

Is data correct and enough?

My hypothesis:

In daily life, we observe data, and often rush to draw conclusions without thinking much about the quantity and quality of data

In this chaotic world, we often don't know whether our observed data is correct and enough
$\rightarrow$ The reasonable conclusion in most cases is probably no conclusion

What data science is: let’s look at its activities

- Ask a meaningful question
- Collect data
- Explore data: look at data to understand more about data, to identify problems in data, ...
- Preprocess data
- Analyze data (simple analysis: do computation and visualization to answer questions about things in our observed data; complex analysis: model data using machine learning to answer questions about things out of our observed data)

  $\longrightarrow$ the answer
- Communicate results / make decision

This data science process is not linear step-by-step

- It can start from a question, but it can also start from data and we will identify questions after exploring data
- From one step, we may need to go back to a step before to adjust things, and we may need to go back and forth many times

---

To do this process well, data scientists need to:

- Stay calm, objective, honest
- Master software computer tools for data science

What data science is: more abstract definitions

“A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician” – Josh Blumenstock

“Data science = computer science + statistics + domain knowledge” – Berkeley data 8 course

Data science is hot nowadays

Many Vietnamese companies need data scientists: tiki, lazada, shopee, zalo, grab, fpt, ...

2 main reasons make data science hot:

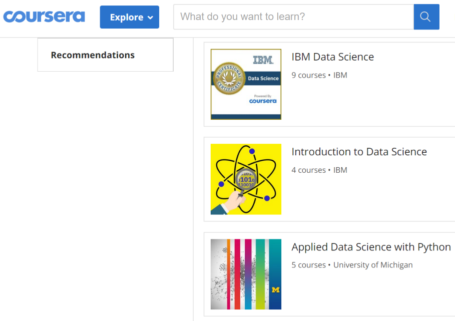
- Data gets bigger and bigger (beware: incorrect data and correct data mix together, we often don’t know which is which)
- Computers get faster and faster, allowing process big data in endurable time

Data science is hot nowadays: let’s make a tour of its applications in different fields

Recommender system

Question: what products will the user X like?

Data: history data of the user X (have liked/bought what products) and other users



IBM Data Science
9 courses • IBM

Introduction to Data Science
4 courses • IBM

Applied Data Science with Python
5 courses • University of Michigan

Web design for online shop

Question: how should the website be designed to attract customers?

Data: interaction data of customers with different design versions of the website

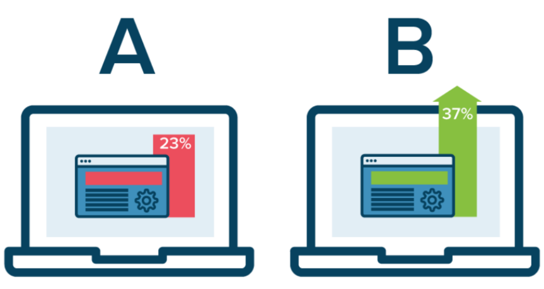


Image source: https://www.optimizely.com/optimization-glossary/ab-testing/

Sentiment analysis

Question: Does user community respond positively or negatively to the new product X?

Data: users’ reviews in social networks

Epidemic control

Question: what is the current state of the epidemic X in different locations?

Data: data about the epidemic X in different locations (which can be collected via newspapers, social networks, ...)

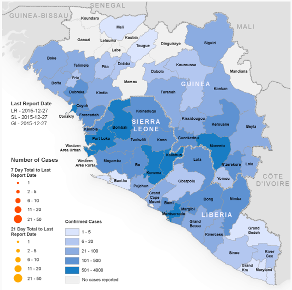


Image source: http://apps.who.int/ebola/sites/default/files/thumbnails/image/sitrep_casecount_31.png

Precision agriculture

Question: where in the farm field do we need supply water/fertilizer/...?

Data: data about the field collected via drone, ...

Image source: https://www.dronitech.com/drones-are-revolutionizing-the-future-of-agriculture-farming/

Education improvement

Question: what aspects of the course X do we need to improve?

Data: feedback data of students about the course

Criminal investigation

Question: what crimes were committed by the same offender?

Data: data about different crimes

Course contents

In this course, we will learn about tools for doing data science process:

- Linux commands
- Git & Github
- Conda
- Jupyter notebook
- Markdown
- Python
- Matplotlib
- Numpy
- Pandas

Data science process:

- Ask a meaningful question
- Collect data (this course: mostly use data which is already collected and published by others)
- Explore data
- Preprocess data
- Analyze data (this course: mostly do simple analysis, i.e. do computation and visualization to answer questions about things *in* our observed data)
  → the answer
- Communicate results / make decision

After successfully completing this course, you will be able to use learned tools to do learned data science process with new data

Q: What is the relationship between this course and the “Introduction to Data Science” course?

A (my view):

- The “Introduction to Data Science” course focuses more on data science process than tools (try to cover all steps in data science process, understand tools at a basic level)
- This course focuses more on tools than data science process (not try to cover all steps in data science process; instead, using time to go deep into important tools)

How many of you learned/are-learning the “Introduction to Data Science” course?

Let’s find out using data … (demo)

Course assessment

- **Individual homeworks** throughout the course: 50% of final score
- **Group final project**: 50% of final score
    - 2 students / group
    - Use Git & Github to control versions and collaborate
    - Find a public dataset about things your group is interested in, use tools to: explore data, identify meaningful questions which can be answered with this data, preprocess and analyze data to answer each question
    - Start near the end of the semester, do in ~3 weeks, and then present results to Teacher

Remember: the main goal here is to learn, truly learn

You can discuss ideas with others as well as consult Internet sources, but your writing and code must be your own, based on your own understanding

If you violate this rule, you will get 0 score for the course

References

- The “Stopping to Sharpen Your Tools” talk of Brandon Rhodes
- The “The Missing Semester” course of MIT
- The “Computational and Inferential Thinking” book of Berkeley
- Tool documents:
  - Git
  - Conda
  - Jupyter notebook
  - Python
  - Numpy
  - Matplotlib
  - Pandas

How to use tools in general: reflect and look forward

- In the “Stopping to Sharpen Your Tools” talk, Brandon talks about stopping to sharpen tools before continuing to use them to do the work
- Brandon gives example about a farmer using a blade to cut grass in the field
  - During this process, the blade will become blunt gradually
  - Instead of continuing to cut grass with this blunt blade, the work will be done faster if the farmer stops a little bit to sharpen his blade before continuing
- Using computer tools of us is similar: our work can be done faster if we stop to ”sharpen” our tools (change tools’ settings, learn more about how to use tools, change to other tools, ...) before continuing to use

Q: When is the right time to stop to sharpen our tools?

A:

Now – the beginning of the semester – is a good time to stop a little bit and reflect about how we have used our computers as well as how we should change for the better

Do you stay focused when using your computer (e.g. for online learning)?

Do you often use your computer for meaningless things?

How many times in a day do you check facebook, mail, …?
Do you aware of what you are doing?

Me: To avoid above problems, I try to have a clear plan before using my computer, and shutdown it when things are done

Q: When is the right time to stop to sharpen our tools?

A:

Now – the beginning of the semester – is a good time to stop a little bit and reflect about how we have used our computers as well as how we should change for the better

Assume that we have a deadline tonight and our computer suddenly become slow
Should we stop to reinstall our OS?

So, we need to consider our specific situation to decide whether we should stop to sharpen our tools

Q: What is the most important tool we need to keep sharp?

A:

Ourself, because we is “the tool” which use other tools

We need to watch ourself to know when we should stop to rest to recover our “sharpness” before continuing

Me: During a day, there are 4 periods I stop to rest, to calm my mind

- Before morning session
- After morning session and before afternoon session
- After afternoon session and before evening session
- After evening session

What tools do you have in your toolbox, and do you master these tools?

- OS commands?
- Word, Excel, PowerPoint?
- GoogleDocs, GoogleSheets, GoogleSlides?
- Latex?
- Markdown?
- C, C++?
- Python?
- Notepad++ / Atom / Sublime / VSCode / Vim / Emacs ...?

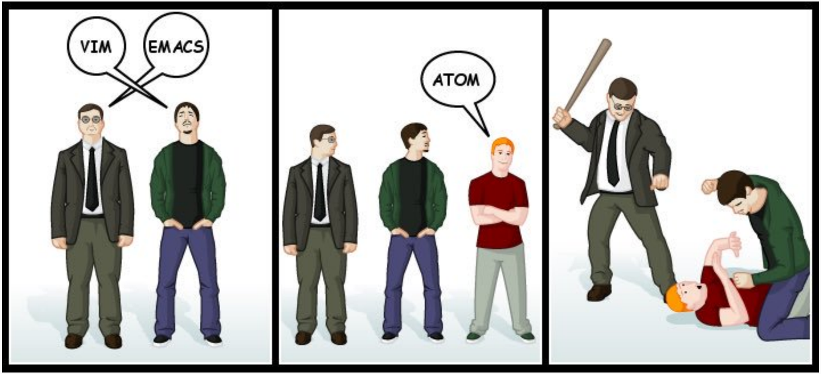


Image source: https://www.reddit.com/r/ProgrammerHumor/comments/6xs6zl/vim_vs_emacs_vs_atom/

What tools do you have in your toolbox, and do you master these tools?

- OS commands?
- Word, Excel, PowerPoint?
- GoogleDocs, GoogleSheets, GoogleSlides?
- Latex?
- Markdown?
- C, C++?
- Python?
- Notepad++ / Atom / Sublime / VSCode / Vim / Emacs ...?
- GoogleDrive / Dropbox ...?
- Git & Github?
- ...

Next lecture

Linux commands