Data Wrangling

DATA 542

Fatemeh Hendijani Fard

Winter 2023/2024 – Term 2

The big questions

- Who are we?
  - Professor, TA, students?
- Why are we here?
  - Why learn about data wrangling?
- What is it all about?
  - Course goals?
  - Logistics?

Who are we?

Who are we? Professor!

Professor:	Fatemeh Hendijani Fard (Assistant Prof. CS)
Office Location	FIP 305
Office Phone	250-807-9607
Email	Fatemeh.fard@ubc.ca
Credit Hours	3.0
Presentation format	Lecture 3 hrs/wk, Tuesdays, Thursdays
Prerequisite:	NA

- Course website:
  - Canvas– Check frequently

More about my research

- Build intelligent tools to help software developers
- Techniques:
  - Deep neural networks and natural language processing
  - Software engineering and mining software repositories (Stack Overflow, GitHub, etc.)
- Applications:
  - Code analysis (automatic comment generation, method name prediction)
  - User feedback analysis (Twitter, Google Play, etc.)
- I am always hiring good students for Masters or PhD
  - A fully funded PhD scholarship covering both tuition fees and living expenses
  - Strong Coding and Math background
  - Preferred background: CS/SE/EE

More about me!

- Instructor
- Researcher
  - Research area: Software Analytics, AI for Code and developers
  - Experienced with: machine learning, Big data analytics and technologies, deep neural networks
  - Software developer, collaborator with my industry partners, writer, learner, reader

- Reviewer
- Mentor
- And I have my personal life!

More about me!

- I cannot answer all individual emails (more than 300 students from other courses). Contact TAs first.
- I am not working 24/7 on the course development.
  - But what you learn and how you learn it is super important for me.
- I sometimes forget to upload what I promised on time.
  - I am a human and I make mistakes too.
- You are all my special course assistants to have a better experience in this course.
  - Your feedback matters.

More about me!

- I have been in your position:
  - knowing what skills are required when you apply for a job
  - knowing what are the frustrations to work with data (and so much fun of course)
  - Knowing that data is not ready to analyze

Who are we? TAs

Experienced graduate students

- Amanat Ullah (amanat7@mail.ubc.ca)

Who are we? Students
Why are you here?

- Introduce yourself
  - Name
  - Background
  - Motivation to be here
  - Goals

Why are we here?

Why learn about data wrangling?

My Goals

- **My Goals for this course:**
  - Have you be successful in the course by learning and understanding the materials.
  - Summarize and document the information in a simple, concise, and effective way for learning.
  - Be available for questions during class time, office hours, and at other times as needed.

  - **Learn and understand data wrangling in practice!**

Why are we here?

- I’m here because I see practical usage of data wrangling in my work!
- This is especially important when you work with textual data and big datasets!

Why this course?

- The overall goal of DATA 542 is for you to:

Apply fundamental pre-processing techniques to data and prepare it for data analytics

- This course will cover essential skills required for data wrangling for real world problems using programming techniques with Python.

Data pre-processing statistics

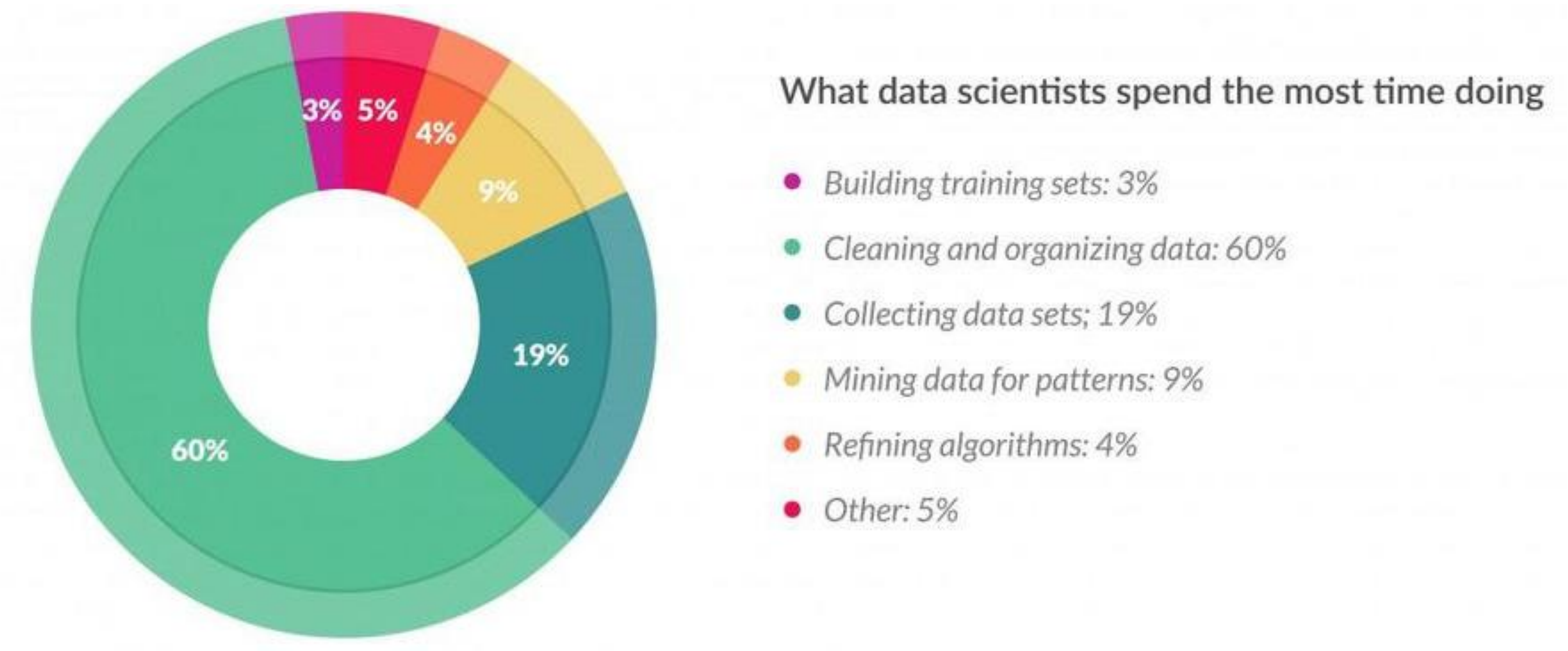
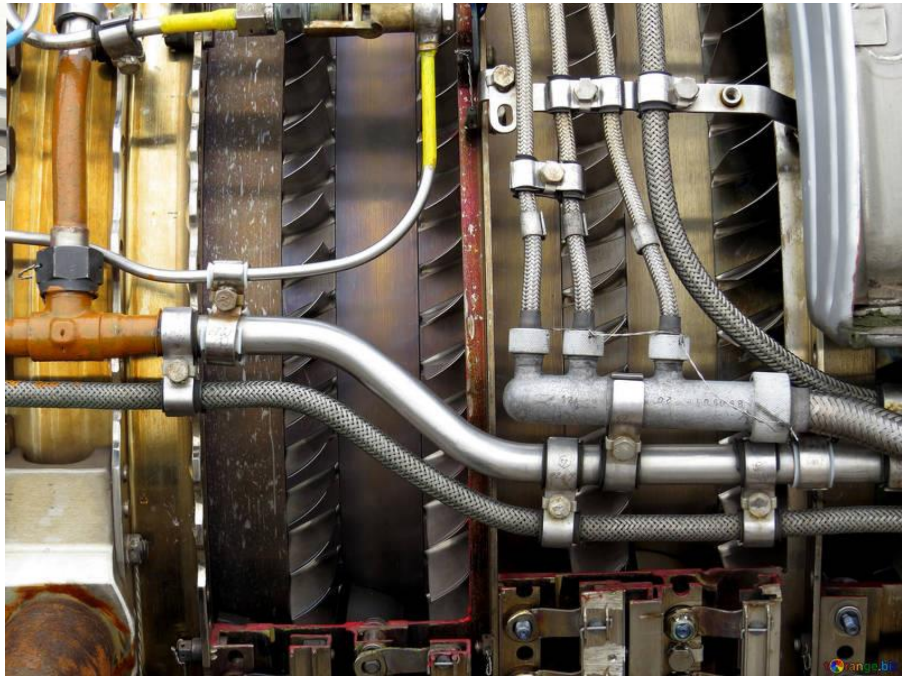


Image source: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#3c4da0836f63

Reality vs expectation

So Fabulous
Such Beauty
What I Think I Look Like when I Blow Dry my hair
V-ART-579
What In The World is this odd Creature?!
What I Actually Look Like



https://www.deviantart.com/v-art-579/art/Expectations-vs-Reality-Blowdrying-My-Hair-658122944

Not convinced yet?

Why Study Data Wrangling?

This is a required course!

What is it all about?

Course goals/overview

Logistics

Course goals

- Upon successfully completing this course, you will be able to:

Perform your data pre-processing in a programming environment

Manage different types of data

Import, scrape, and export data

Index, subset, reshape and transform data

Filter, sort, and group

Clean, convert, and parse your data in different formats, including times and dates

Join separate sources of data

Visualize data

Perform basic data analytics and statistics

Course Topics

- Load data
- Clean data
- Process data
- Wrangle Data
- Exploratory Data Analysis
- Data Analysis
- Export reports/data analyses and visualizations

Required

- Energetic students
- Jupyter notebook
- Python
- Numpy
- Pandas
- Seaborn and matplotlib
- Scikit learn

Resources

- Python Data Science Handbook by Jake VanderPlas:
- https://github.com/jakevdp/PythonDataScienceHandbook

- **Python for Data Analysis**
- Data Wrangling with Pandas, NumPy, and IPython
- By William McKinney

Other Resources

- Google ☺
- Your classmates

How to get most out of the course?

- Attend lectures and labs
- Interact
- Practice
- Solve assignments yourself and ask your peers
- ***Excel in Data Science:***
  - Online sources
  - Kaggle: 5 day challenges, tutorials, competitions
  - Stackoverflow
  - Practice
  - Learn about techniques and technologies
  - Excel in programming
  - Don’t rely on towardsdatascience articles!

Lab assignments

- 4 labs
- Starting TODAY
- 80% of total score

In-class activities

- In class questions and participation
- Discussions and group activities in class: **Discuss ….**
- Questions: I will ask groups to announce the discussion/answer for class

Grading

- Labs: 80%
- Quiz: 20%

Expectations

- Participation
- Interaction
- Prepare for class
- Review course lecture

Let’s Start Our Business!