Elements Of Data Science - S2022

Introduction

1/18/2022

Who am I?

Haiyuan Wang, Ph.D.

- B.S. Engineering, Tongji University, Shanghai, China
- M.S., Ph.D. Operations Research and Statistics, Rensselaer Polytechnic Institute, Troy, NY
- Multiple years of modeling and research experience in financial companies including Morgan Stanley and BlackRock.
- Adjunct faculty in Applied Analytics, DSI, and Statistics since 2016.

Acknowledgement

Jake VanderPlas

- Fantastic textbook on Data Science

Aurelien Geron

- Fantastic textbook on Applied Machine Learning

Bryan Gibson, Ph.D.

- Developed this course
- Have used and enhanced this set of materials for multiple years

Who is this course for?

People new to one of:

- Python
- Data Science Python libraries
- Visualization
- Hypothesis Testing
- Machine Learning

What will we be covering?

- Python DS tools
- Data exploration and visualization
- Exploratory data analysis and hypothesis testing
- Data manipulation, cleaning and transformation
- Predictive modeling using ML

What will we be covering? (cont)

- Regression
- Decision trees and ensemble methods
- Support vector machines
- Clustering
- Dimensionality reduction
- Natural Language Processing and topic modeling
- Dealing with time series data
- Recommendation engines
- Interacting with databases

Logistics

Email: hw2592@columbia.edu
TA: Yufan Cao, yc3906@columbia.edu
Office Hours: TBD on canvas

Course Materials

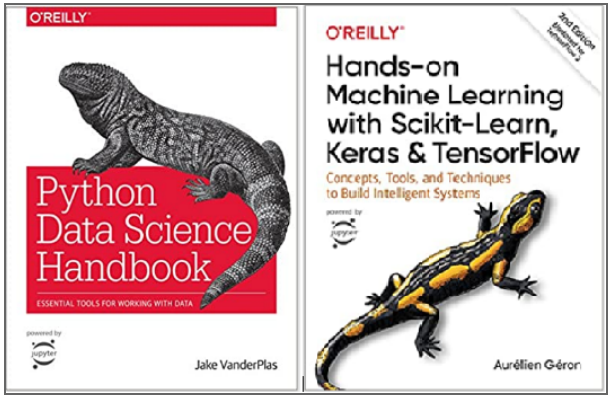
- Course Website via Courseworks:
  https://courseworks2.columbia.edu/courses/136835

Slides

- written using Jupyter Notebook + RISE + reveal.js
    - open .ipynb in jupyter
- also saved as pdf (slides_pdf folder)
    - open in a pdf viewer (acrobat, evince, etc.)

Textbooks

- (PDSH) **Python Data Science Handbook** by Jake VanderPlas
    - **Free online**
- (HOML) **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd Edition)** by Aurelien Geron
    - **Via Amazon**
    - **Associated Github repo**



Other Useful Texts

- **Python Machine Learning (3rd Edition)** by Raschka and Mirjalili
- **Data Science from Scratch, 2nd Ed.** by Joel Grus
- **Python for Data Analytics** by Wes McKinney
- **Practical Statistics for Data Scientists** by Bruce and Bruce

Additional Resources

- See the course website...

Quizzes, Homeworks and Exams

- **Potentially Weekly Quiz**, submit online, graded on completion
  - 20% of grade, equally weighted
  - **no late days**
- **4 Homework Assignments**, submit online, equally weighted
  - 30% of grade, equally weighted
  - **no late days**
- **Two projects** (end of March, and Beginning of May) 30% of grade
- **Final Exam** (end of Semester) 20% of grade

Course

- In-class and online (see course page for zoom recordings)
- Use Canvas Discussion for questions or email
- Zoom office hours

Expectations

- Attend/view the weekly lecture
- Ask/answer questions via Canvas or email
- Attend Office Hours for additional help
- Complete all quizzes and homeworks on time
- Hopefully learn enough to get through a junior DS job interview

Plagarism and Code copying

- Homeworks may be checked for plagiarism
- Copied code will result in 0 points for all involved
- Copying from my slides or online sources, not recommended but common practice

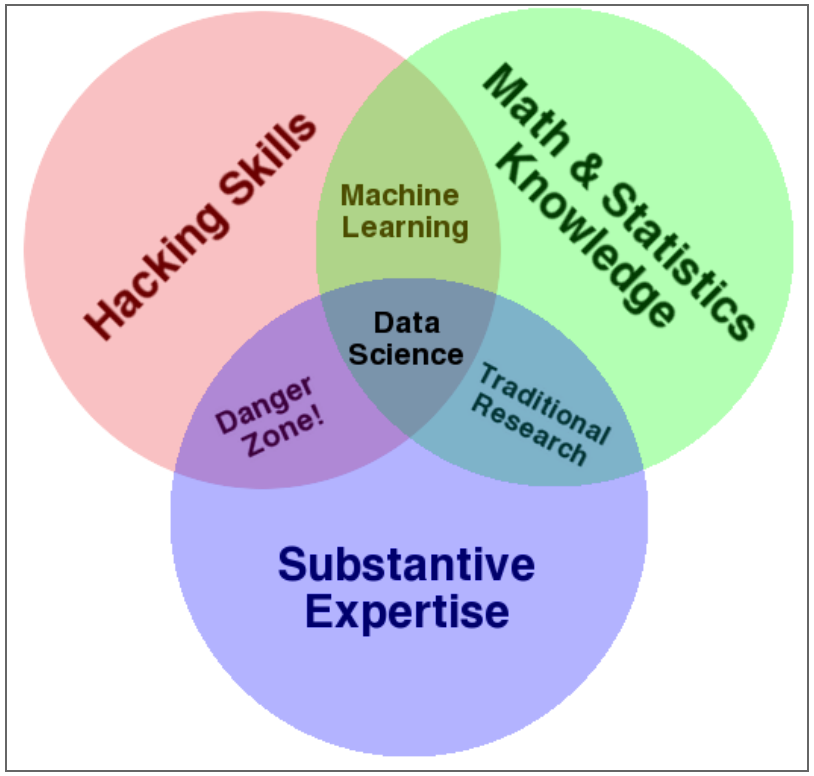
Questions re Logistics?

What is Data Science?

Data science, also known as data-driven science, is **an interdisciplinary field** about scientific methods, processes, and systems **to extract knowledge or insights from data in various forms**, either structured or unstructured, similar to data mining.

https://en.wikipedia.org/wiki/Data_science

What is Data Science?



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data Science ≠ Magic

- "Can we find something in this data?" **Yes**
- "Will it solve our business problem?" **Maybe**
- "Will it be easy?" **Probably not**

Data Science Workflow

- Business Need →
- DS Question →
- **E**xtract-**T**ransform-**L**oad (ETL)→
- Experimentation →
- API/Tool Creation →
- Reporting

Important Before You Start!

1. What's the question?

1. What does success look like?

1. How are we going to measure it?

Example DS Projects

- **Machine Bias in Criminal Sentencing, Propublica**
- **Analysis of OkCupid Data**
- **David Bowie Job Mentions**
- **NYC Crash Mapper**
- **NeurIPS 2019 Acceptance Stats**
- Demo: Example Flowershop

Questions?