Elements Of Data Science - F2024

Introduction

9/9/2024

Who am I?

Andi Cupallari

- PhD in Economics with a focus on AI and Deep Learning. Research Interests: AI, NLP (LLMs), causal inference, forecasting, advanced analytics

Director, Forecasting Capabilities @ Merck

MERCK

Past Experiences

Associate Director, Advanced Analytics @ Kite Pharmaceuticals

Kite
A GILEAD Company

Data Scientist, AI and NLP Expert

NOVARTIS

Who is this course for?

Who is this course for?

People new to (at least) one of:

Who is this course for?

People new to (at least) one of:

- Python

Who is this course for?

People new to (at least) one of:

- Python
- Data Science Python libraries

Who is this course for?

People new to (at least) one of:

- Python
- Data Science Python libraries
- Visualization

Who is this course for?

People new to (at least) one of:

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

Who is this course for?

People new to (at least) one of:

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

What will we be covering?

What will we be covering?

- Python DS tools
- Exploratory Data Analysis and Visualization
- Data Manipulation including cleaning and transformation
- Hypothesis Testing
- Predictive modeling using ML

What will we be covering? (cont)

What will we be covering? (cont)

- Clustering
- Dimensionality Reduction
- Natural Language Processing and Topic Modeling
- Dealing with Time Series data
- Recommendation Engines
- Interacting with Databases

Logistics

Columbia University email: ac5562@columbia.edu

Personal email: acupallari@gmail.com

TAs: See the course website

Office Hours:

- Andi: Mondays, 6pm, location TBA
- TAs: TBA, location TBA

Course Materials

Course Materials

- Course Website via Courseworks: https://courseworks2.columbia.edu/courses/185631

Course Materials

- Course Website via Courseworks:
  https://courseworks2.columbia.edu/courses/185631

- Slides and weekly quizzes:
  More instructions to come

Course Materials

- Course Website via Courseworks:
  https://courseworks2.columbia.edu/courses/185631

- Slides and weekly quizzes:
  More instructions to come

- Homeworks:
  More instructions to come

Slides

Slides

- written using Jupyter Notebook
    - in `notebooks` folder
    - open .ipynb files in jupyter

Slides

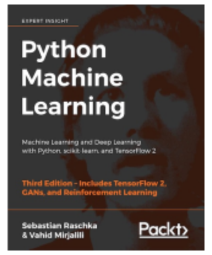
- written using Jupyter Notebook
  - in `notebooks` folder
  - open .ipynb files in jupyter

- also saved as pdf
  - in `slides_pdf` folder
  - open .pdf in a pdf viewer (chrome, acrobat, evince, etc.)

Textbooks

- (PDSH) **Python Data Science Handbook** by Jake VanderPlas
  - Free online
  - Columbia Library
  - 2nd Edition coming soon

- (PML) **Python Machine Learning (3rd Edition)** by Raschka and Mirjalili
  - Columbia Library
  - Associated Github repo
  - New Edition: Machine Learning with PyTorch and Scikit-Learn



Other Useful Texts

- **Data Science from Scratch, 2nd Ed.** by Joel Grus
- **Python for Data Analytics** by Wes McKinney (2nd Edition coming soon)
- **Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python** by Bruce, et al.
- **Effective Pandas** by Matt Harrison
- **SQL for Data Scientists** by Renée M. P. Teate

Quizzes, Homeworks and Exams

Quizzes, Homeworks and Exams

- **Weekly Quiz**, submit online (TBA)
    - 10% of grade, equally weighted
    - **no late submissions accepted**
    - **if you know there will be an issue, let me know in advance**

Quizzes, Homeworks and Exams

- **Weekly Quiz**, submit online (TBA)
    - 10% of grade, equally weighted
    - **no late submissions accepted**
    - **if you know there will be an issue, let me know in advance**

- **4 Homework Assignments**, submit online
    - 40% of grade, equally weighted
    - 2 free late days total over the semester to be used when you choose
    - 25% off for each late day

Quizzes, Homeworks and Exams

- **Weekly Quiz**, submit online (TBA)
  - 10% of grade, equally weighted
  - **no late submissions accepted**
  - **if you know there will be an issue, let me know in advance**

- **4 Homework Assignments**, submit online
  - 40% of grade, equally weighted
  - 2 free late days total over the semester to be used when you choose
  - 25% off for each late day

- **Midterm exam** 25% of grade

Quizzes, Homeworks and Exams

- **Weekly Quiz**, submit online (TBA)
    - 10% of grade, equally weighted
    - **no late submissions accepted**
    - **if you know there will be an issue, let me know in advance**

- **4 Homework Assignments**, submit online
    - 40% of grade, equally weighted
    - 2 free late days total over the semester to be used when you choose
    - 25% off for each late day

- **Midterm exam** 25% of grade

- **Final Exam** 25% of grade

In person Course

- In-class
- Use Ed Discussion for questions
- Zoom office hours (TBD)

Expectations

- Attend/view the weekly lecture
- Ask/answer questions via Ed
- 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

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

Questions re Logistics?

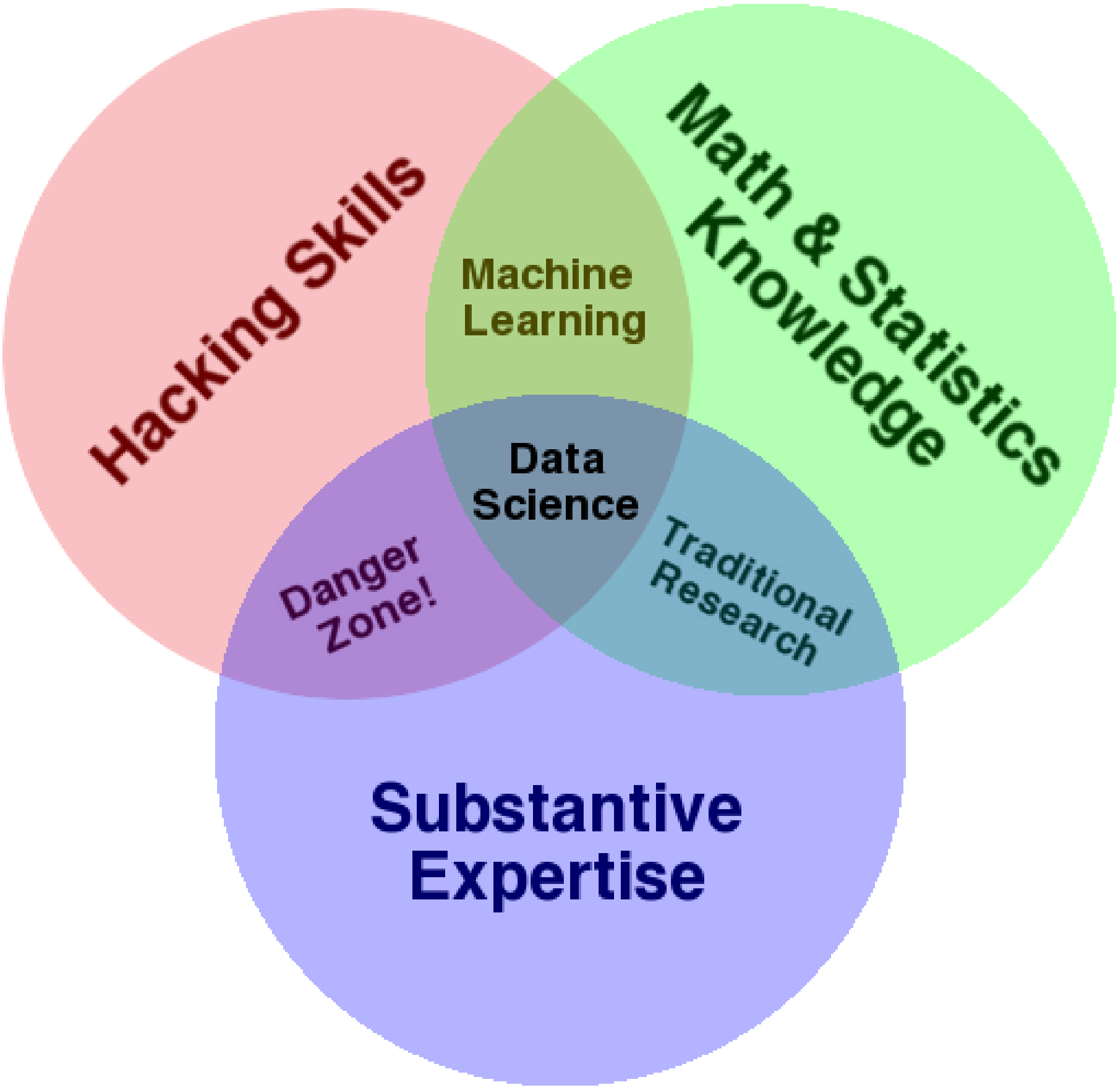
What is Data Science?

What is Data Science?

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 $\neq$ Magic

Data Science $\neq$ Magic

- "Can we find something in this data?" **Yes**

Data Science $\neq$ Magic

- "Can we find something in this data?" **Yes**
- "Will it solve our business problem?" **Maybe**

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

Data Science Workflow

- Business Need →

Data Science Workflow

- Business Need →
- DS Question →

Data Science Workflow

- Business Need →
- DS Question →
- **E**xtract-**T**ransform-**L**oad (ETL)→

Data Science Workflow

- Business Need →
- DS Question →
- **E**xtract-**T**ransform-**L**oad (ETL)→
- Experimentation →

Data Science Workflow

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

Data Science Workflow

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

Important Before You Start!

Important Before You Start!

1. What's the question?

Important Before You Start!

1. What's the question?

2. What does success look like?

Important Before You Start!

1. What's the question?

2. What does success look like?

3. How are we going to measure it?

Important Before You Start!

1. What's the question?

2. What does success look like?

3. How are we going to measure it?

Can't always get answers to these, but good to ask.

Example DS Projects

- Machine Bias in Criminal Sentencing, Propublica
- Analysis of OkCupid Data
- David Bowie Job Mentions
- NYC Crash Mapper
- NeurIPS 2019 Acceptance Stats
- NeurIPS 2021 Stats
- Demo: Example Flowershop

Questions?