Elements Of Data Science - F2021

Introduction

9/13/2021

Who am I?

Bryan R Gibson, PhD

Russell Reynolds
ASSOCIATES

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

UNIVERSITY OF MICHIGAN

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)

- Clustering
- Dimensionality reduction
- Natural Language Processing and topic modeling
- Dealing with time series data
- Recommendation engines
- Interacting with databases

Logistics

Email: brg2130@columbia.edu

TAs: See the course website

Office Hours: See the course website

Course Materials

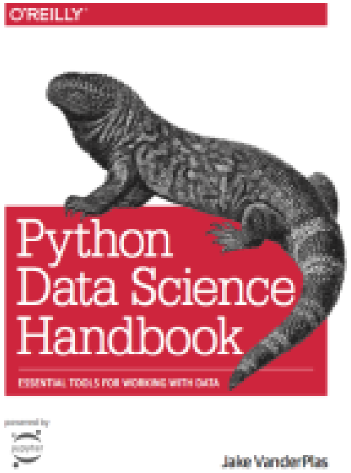
- Course Website via Courseworks:
  https://courseworks2.columbia.edu/courses/136835
- Slides and weekly quizzes via course git repo:
  https://github.com/bryanrgibson/eods-f21
- Homeworks via git:
  More instructions to come

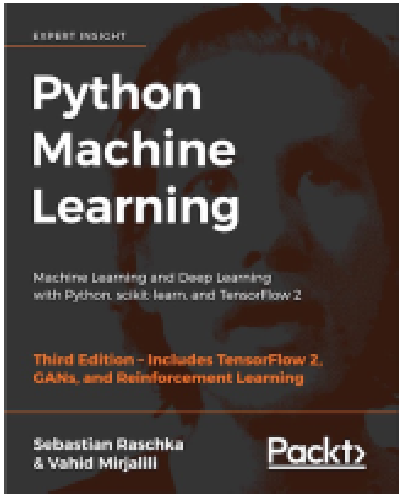
Slides

- 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
- (PML) **Python Machine Learning (3rd Edition)** by Raschka and Mirjalili
  - Via Amazon
  - Associated Github repo
  - 2nd Edition online via Columbia Library


O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
powered by jupyter
Jake VanderPlas


EXPERT INSIGHT
Python Machine Learning
Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2
Third Edition - Includes TensorFlow 2, GANs, and Reinforcement Learning
Sebastian Raschka & Vahid Mirjalili
Packt>

Other Useful Texts

- **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

- **Weekly Quiz**, submit online, graded on completion
  - 10% of grade, equally weighted
  - **no late days**
- **4 Homework Assignments**, submit online, equally weighted
  - 40% of grade, equally weighted
  - 2 free late days to be used when you choose
  - 25% off for each late day
- **Midterm exam** (online, week of 10/25) 25% of grade
- **Final Exam** (online, week of 12/13) 25% of grade

Online Course

- In-class and online (see course page for recordings)
- Use Ed Discussion for questions
- Zoom office hours

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

- 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

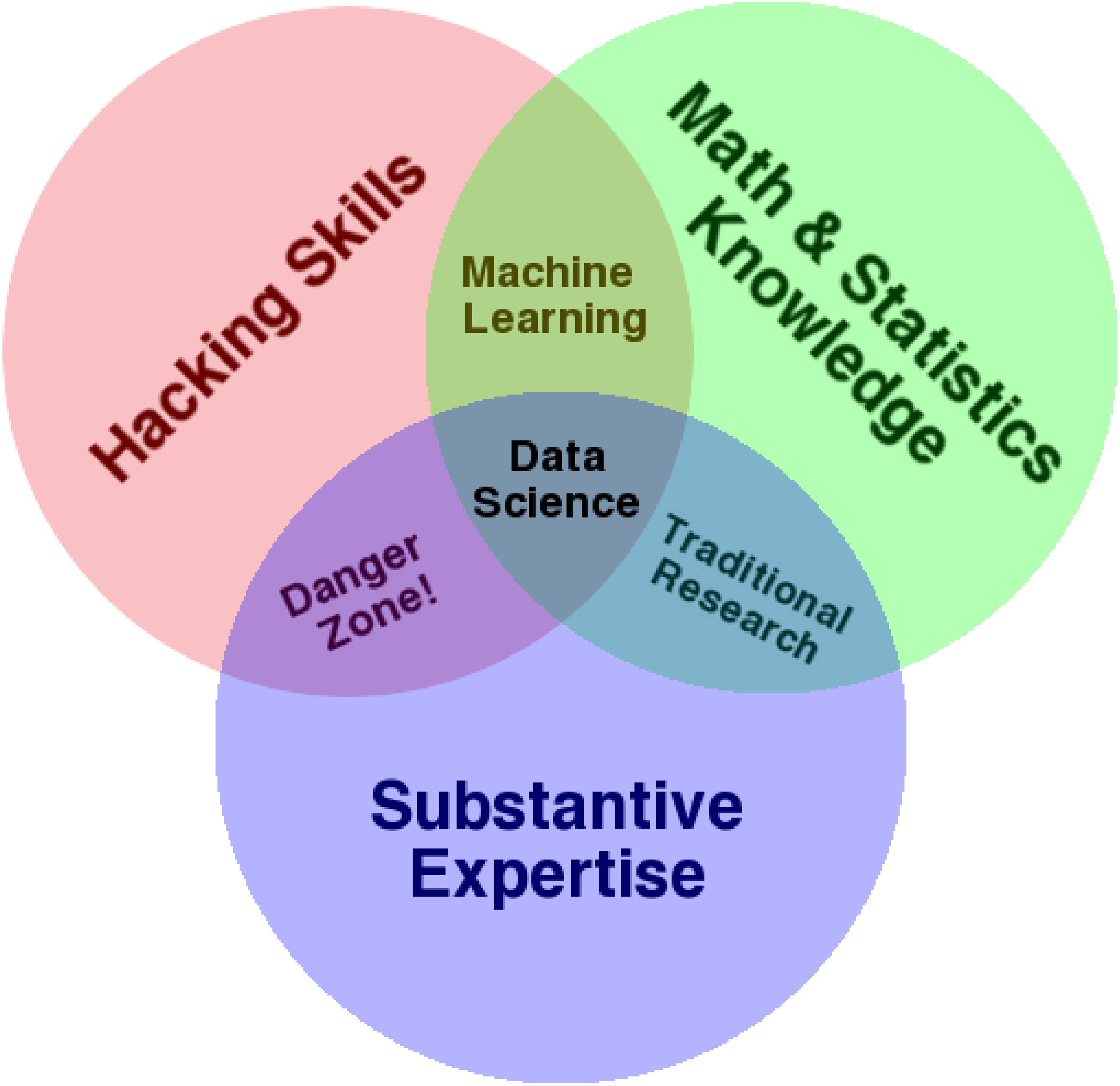
Questions re Logistics?

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?

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
- Demo: Example Flowershop

Questions?