

# Elements Of Data Science - S2021

## Introduction

1/11/2021

# Who am I?

Bryan R Gibson, PhD



# Who is this course for?

# Who is this course for?

People new to one of:

# Who is this course for?

People new to one of:

- Python

# Who is this course for?

## People new to one of:

- Python
- Data Science Python libraries

# Who is this course for?

## People new to one of:

- Python
- Data Science Python libraries
- Visualization

# Who is this course for?

## People new to one of:

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



# 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?

# 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)

# 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](mailto:brg2130@columbia.edu)

**TAs:** See the course website

**Office Hours:** See the course website

# Course Materials

# Course Materials

- Course Website via Courseworks:

<https://courseworks2.columbia.edu/courses/124237>



# Course Materials

- Course Website via Courseworks:

<https://courseworks2.columbia.edu/courses/124237>

- Slides via git:

<https://github.com/bryanrgibson/eods-s21>

# Course Materials

- Course Website via Courseworks:

<https://courseworks2.columbia.edu/courses/124237>

- Slides via git:

<https://github.com/bryanrgibson/eods-s21>

- Weekly quizzes via course git repo

# Course Materials

- Course Website via Courseworks:  
<https://courseworks2.columbia.edu/courses/124237>
- Slides via git:  
<https://github.com/bryanrgibson/eods-s21>
- Weekly quizzes via course git repo
- Assignments via git:  
More instructions to come

# Slides

# Slides

- written using Jupyter Notebook + RISE + reveal.js
  - open .ipynb in jupyter

# 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](#)



# 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

# Quizzes, Homeworks and Exams

- **Weekly Quiz**, submit online, graded on completion
  - 10% of grade, equally weighted
  - no late days

# 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

# 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**, 25% of grade

# 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**, 25% of grade
- **Final Exam**, 25% of grade

# Online Course

- Synchronously online / Asynchronous online (see course page for recordings)
- Use Piazza for questions
- Online office hours

# Expectations

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



# Plagiarism and Code copying

# Plagiarism 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 (Stack overflow, tutorials, etc. ) is fine

# Questions re Logistics?

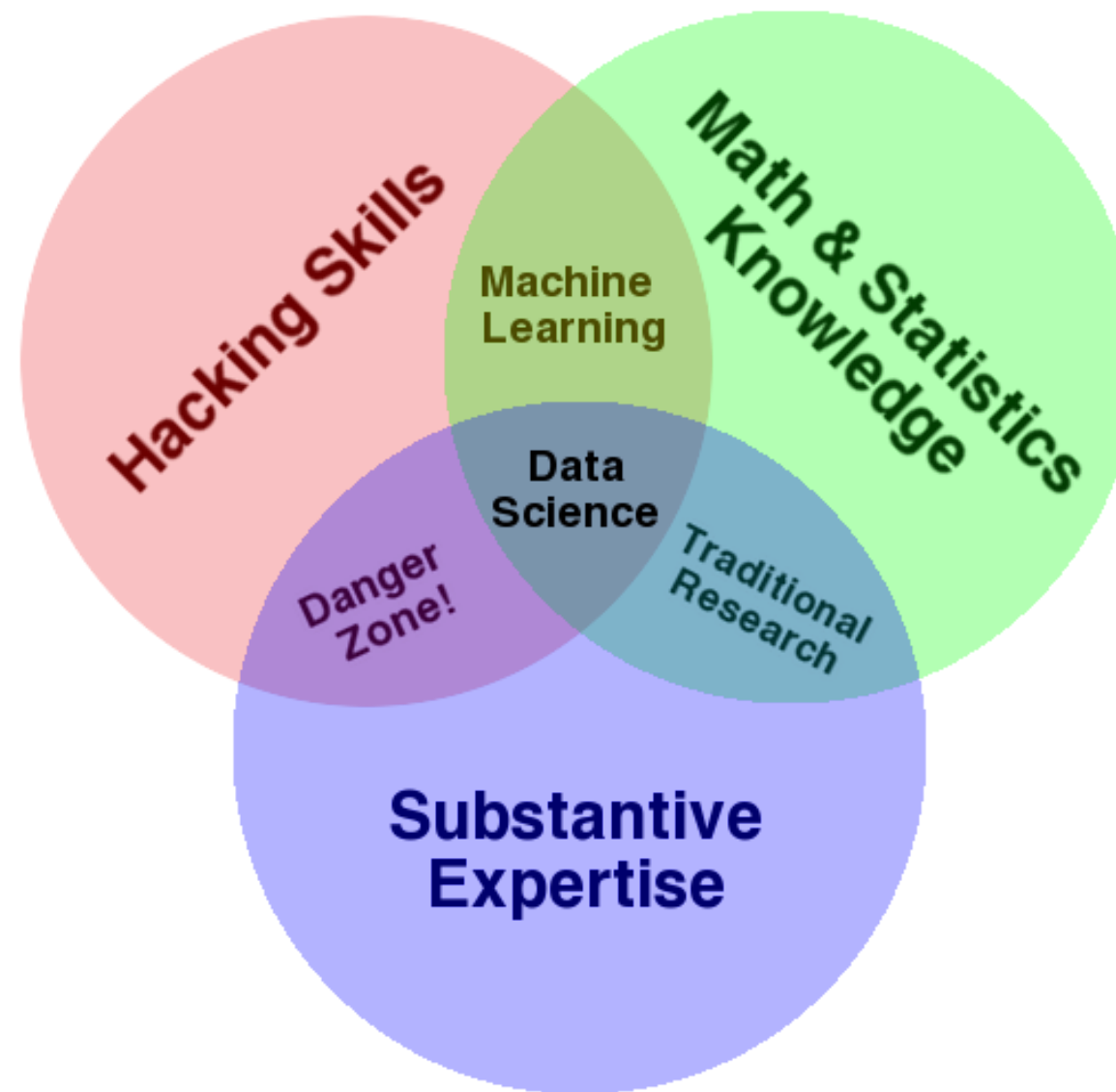
# What is Data Science?

# 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](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 $\neq$ 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 →
- **Extract-Transform-Load (ETL)**→

# Data Science Workflow

- Business Need →
- DS Question →
- **Extract-Transform-Load (ETL)**→
- Experimentation →

# Data Science Workflow

- Business Need →
- DS Question →
- **Extract-Transform-Load (ETL)**→
- Experimentation →
- API/Tool Creation →



# Data Science Workflow

- Business Need →
- DS Question →
- **Extract-Transform-Load (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](#)
- Demo: Example Flowershop

# Questions?

Please complete the poll at: [https://PollEv.com/free\\_text\\_polls/Ba4gBMLXJBATh86WFMYFz/respond](https://PollEv.com/free_text_polls/Ba4gBMLXJBATh86WFMYFz/respond)