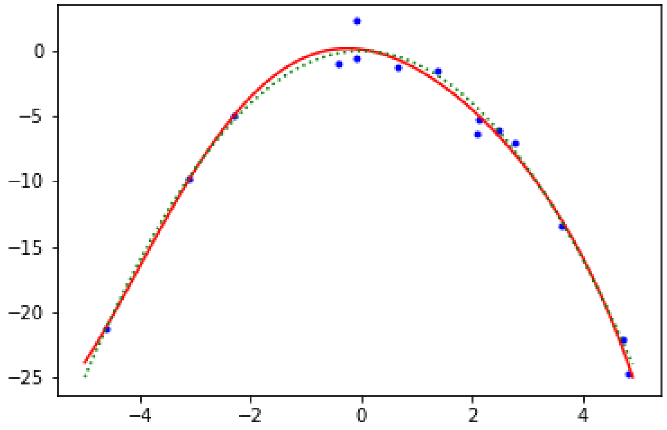
CSCI 303

Introduction to Data Science

1 - Welcome and Introduction



What is Data Science?

An NSF workshop on Data Science Education in 2016 concluded:

> "The integration of general scientific principles, computer science, information science, mathematics, statistics, and subject matter expertise creates an intersection that has as many definitions as there are academics attempting to define it."

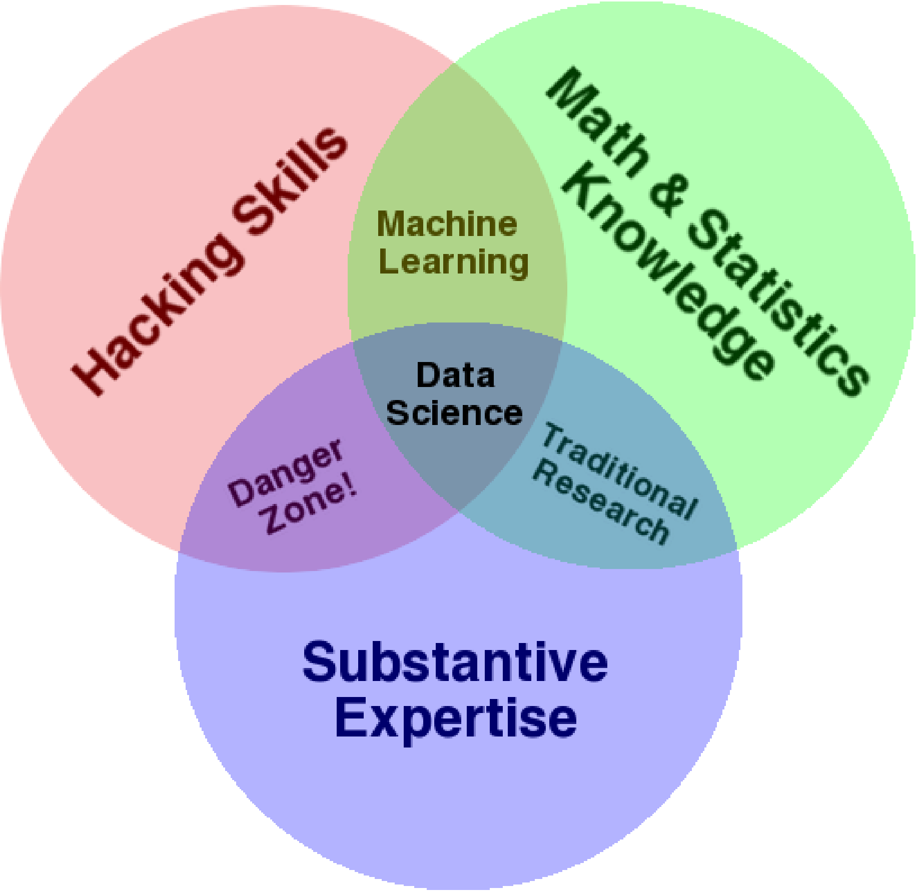
What is Data Science?

Working definition:

Data Science is an emerging discipline at the intersection of computer science and statistics which deals with the extraction of knowledge from data.

What is Data Science?

[Data Science Venn Diagram by Drew Conway (https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diagram/data_science.html)](https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diagram/data_science.html)



Why Data Science?

- Need:
  - Massive amounts of data becoming available
    - Marketing data (e.g. "clicks")
    - Scientific data
    - Socio-economic data
  - Business/science/government needs to stay current
- Timing:
  - Open source software largely supplanting proprietary
  - "Cloud" computing enabling very large data analytics
- The sexiest job of the 21st century? (Harvard Business Review article (https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century))

Some Data Science History

- Business intelligence/analytics (1960s+)
- "The Fourth Paradigm" (Jim Gray, 2007)
- Nate Silver
  - FiveThirtyEight (http://fivethirtyeight.com), 2008, 2012 election predictions
  - *The Signal and the Noise* (https://www.amazon.com/Signal-Noise-Many-Predictions-Fail-but/dp/0143125087)

This Course

- Python
  - basics
  - toolkits
- Database
- Machine learning
  - basics
  - supervised learning
  - unsupervised learning
  - intro to neural networks
- Ethics and miscellaneous topics

Python for Data Science

- Python and R roughly equally represented (for now)
- Python pros:
  - Sophisticated, modern PL
  - Powerful, consistent libraries
  - Easy interfacing (e.g. with other PL's)
  - Ecosystem
- Python cons:
  - Poor parallel execution support

Python Toolkits

- numpy - numerical computing
- pandas - data wrangling
- scikit-learn - machine learning
- matplotlib - visualization

Database

- SQL for querying (focus on SELECT)
- Access via Python

Machine Learning

- Fundamentals
- Programming
  - Some simple techniques from scratch (to help you understand what the tools are doing for you)
  - numpy, pandas, and scikit-learn toolkits

- Algorithms
  - Supervised, unsupervised learning
  - Feature selection/engineering
  - Dimensionality Reduction
  - More as time permits

Class Mechanics

- Jupyter notebooks
  - For lectures, projects
    - Online
      - https://jupyterhub.mines.edu/ (https://jupyterhub.mines.edu/)
      - Uses self-signed certificate for now - ignore warnings (proceed)
      - Login via usual Mines credentials
    - Local JNB
      - sudo pip3 install jupyter
      - sudo pip3 install matplotlib pandas scipy scikit-learn numpy
      - go into directory with JNB files and run: sudo jupyter notebook xx.ipynb
- Canvas
  - For syllabus, schedule, grading

Resources

Python Data Science Handbook, VanderPlas (http://shop.oreilly.com/product/0636920034919.do) *** This is our BOOK!!

Python for Data Analysis, 2nd ed., McKinney (http://shop.oreilly.com/product/0636920050896.do)

Introduction to Machine Learning with Python, Müller & Guido (http://shop.oreilly.com/product/0636920030515.do)

An Introduction to Statistical Learning, James et al. (http://www.springer.com/us/book/9781461471370)

Learning Python, Lutz (http://shop.oreilly.com/product/0636920028154.do)

Docs for Python, Python tools, Jupyter, etc.: See 'Help' in Notebook menu at top.

Further Reading

These books are *not* part of the course, but are general audience books related to data science.

The Signal and the Noise: Why So Many Predictions Fail--but Some Don't, Nate Silver (https://www.amazon.com/Signal-Noise-Many-Predictions-Fail-but/dp/0143125087)

The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research (various authors) (https://www.amazon.com/Fourth-Paradigm-Data-Intensive-Scientific-Discovery/dp/0982544200) - Note: $0.99 on Kindle!

Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Cathy O'Neil (https://www.amazon.com/Weapons-Math-Destruction-Increases-Inequality/dp/0553418815)