Data Science Graduate Education: Content, Projects, and Computer Languages

Data Analytics Across Disciplines
Niagara Falls, NY
5-18-2016

Randy Paffenroth
Associate Professor of Mathematical Sciences, and
Associate Professor of Computer Science
Data Science Program
Worcester Polytechnic Institute

http://www.wpi.edu/academics/datascience.html
A talk on education with a pop quiz!

- How many people know what Data Science is?
- How many people here would call themselves "Data Scientists"?
- How many people have taught others Data Science in a classroom settings?
- How about mentoring Data Scientists in a more informal setting?
Outline

• Data Science and the WPI Data Science program
• The learning goals of the WPI Data Science program
• How "Python" can be used to support those learning goals

Have a conversation! Please ask questions!
Meta Goal: Provide many links to additional resources
Drew Conway's Venn Diagram

Based upon Drew Conway's Data Science Venn Diagram

M.S. in Data Science Program

GRADUATE QUALIFYING PROJECT OR MS THESIS
(3 TO 9 CREDITS)

CONCENTRATION AND ELECTIVES
(9 TO 15 CREDITS)

MATHEMATICAL ANALYTICS (3 CREDITS)
DATA ACCESS & MANAGEMENT (3 CREDITS)
DATA ANALYTICS & MINING (3 CREDITS)
BUSINESS INTELLIGENCE & CASE STUDIES (3 CREDITS)

INTEGRATIVE DATA SCIENCE (3 CREDITS)
2014 Data Science Cohort

NATIONALITY
- CAMBODIA
- INDIA
- CHINA
- PAKISTAN
- TAIWAN
- IRAN
- U.S.A.
- BRAZIL
- NEPAL
- AFGHANISTAN
- INDONESIA

Backgrounds:
- MATHEMATICS
- COMPUTER SCIENCE
- BUSINESS SCHOOL
- INFORMATION TECHNOLOGY
- SENIOR RESEARCH ANALYST
- SENIOR BUSINESS ANALYST
- PATIENT FINANCIAL SERVICES
- DATA BASE ANALYST-ARCHITECT
- DECISION SCIENTIST
- MINISTRY OF FINANCE
- LAHEY HEALTH
- TECHNICAL PROGRAM MANAGEMENT
- U.S. DEPARTMENT OF STATE

GENDER
- 66.70% Male
- 33.3% Female
Skills Acquired by Our Students

**Fundamental/Technical:**
- SQL / Data Modeling / Cleaning
- Data Integration / Warehousing
- Statistical Learning / Machine Learning
- Distributed Computing
- Big Data Management
- Classif./Regression/Decision Trees
- Business Intelligence
- Distributed Mining Algorithms

**Professional Skills:**
- Business Use Cases / Entrepreneurship
- Interdisciplinary Teams / Leadership

**Tools:**
- Oracle / MySQL / DB2 / SQLServer
- R / SAS / SciKit
- Weka / RapidMiner / MatLab
- IBM Cognos / SPSS Modeler
- Hadoop / Mahout / Cassandra
- Python / Java / Cloud Computing
- Storm / Sparc / InfoSphere Streams
- Spotfire / Tableaux

**Professional Skills:**
- Story Telling / Visualization
- Presentations / Reports
Big Educational Opportunity

Taking students who come from varied background and provide them with a common baseline in Data Science.
Introduction to Data Science: Holy, moly!

This course provides an overview of Data Science, covering a broad selection of key challenges in and methodologies for working with big data. Topics to be covered include data collection, integration, management, modeling, analysis, visualization, prediction and informed decision making, as well as data security and data privacy. This introductory course is integrative across the core disciplines of Data Science, including databases, data warehousing, statistics, data mining, data visualization, high performance computing, cloud computing, and business intelligence.
Introduction to Data Science: Holy, moly!

Professional skills, such as communication, presentation, and storytelling with data, will be fostered. Students will acquire a working knowledge of data science through **hands-on projects and case studies** in a variety of business, engineering, social sciences, or life sciences domains. Issues of ethics, leadership, and teamwork are highlighted.
Python
Okay, human.

Huh?

Before you hit 'interpret', listen up.

You know when you're falling asleep, and you imagine yourself walking or something.

And suddenly you realize it's a dream, and start flying?

Yeah!

Well, that's what Python feels like. All the time!

Awesome!
Why Python?

- Great programming model
  - For those who don't like memory management...Garbage collection
  - For those who don't like templated containers...Everything is a first class type.
  - For those who really don't like templates...Duck typing

- Access to other languages
  - Can call C, C++, JAVA, etc.

- Lots of libraries
  - Numpy, scipy, pandas, pycuda, mpi4py, etc.

- It is free!
Python is easy to install

- For more advanced students, they just install everything from packages.
- For students with less computing background you have: Enthought Canopy!
- Makes everything easy to install!
- Academic version
  - https://store.enthought.com/#canopy-academic
IPython/Jupyter Notebooks

- Assignment, IDE, and submission all in one package.
  - LaTeX, Markdown, code, and results all in one place.
Introduction to Data Science

• The most important learning tool?

Case studies using real world data!
Introduction to Data Science: Case Study One

- Getting data (which you kind of need for Data Science :-).  
- Introduction to IPython notebooks.  
- Twitter Streaming API - https://dev.twitter.com/streaming/overview
  - Python Twitter - https://code.google.com/p/python-twitter/
  - Russell, Matthew A. "Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More." O'Reilly Media, Inc., 2013.
Introduction to Data Science: Case Study Two

- Data wrangling
  - MovieLens 1M Data Set - http://grouplens.org/datasets/movielens/
  - NumPy - http://www.numpy.org/
  - Matplotlib - http://matplotlib.org/
  - Pandas - http://pandas.pydata.org/
Introduction to Data Science: Case Study Three

- Machine learning
- Movie Review Data v2.0 from - http://www.cs.cornell.edu/people/pabo/movie-review-data
  - Natural language processing problem
  - “Term Frequency times Inverse Document Frequency” - sklearn.feature_extraction.text.TfidfTransformer
- Vast array of algorithms
  - K-nn, SVM, LDA, QDA, etc.
Introduction to Data Science: Case Study Four

- Distributed processing
  - Hadoop
- Python saves the day again
  - mrjob - https://pythonhosted.org/mrjob/
Cautionary tale: Regression Analysis

- Python is almost too good!
- Almost everything they need is available if they look hard enough.
- Need to be careful that learning objectives and ground rules are clearly stated.
Research

- My students and I are mainly interested in Data Science (with a focus on semi-supervised and unsupervised methods) as applied to cyber defense.

http://www.internet2.edu/
There is extensive support in the community for research.

- Deep learning: Theano http://www.deeplearning.net/software/theano/
- Statistical modeling: Statsmodels http://statsmodels.sourceforge.net/
- HPC: mpi4py http://mpi4py.scipy.org/, pycuda and pyopencl http://mathema.tician.de/software
WPI GQP program

GQP 0 Project – IOMICS. BIOASSAY CONCEPT MAPPING: LIMITATION of CURRENT NLP TECHNOLOGIES. Students: Mohammed Ayub, Bir B. Kafle, Suman Lama, Hai Liu, Nikita Mutta

GQP 1 Project - IBM: AIRCRAFT HEALTH PREDICTIVE ANALYTICS Students: Kakar Tabassum, Junwei Guan, Dutta Suwodi, Sean Viseth

GQP 2 Project – MassInc: PUBLIC SCHOOL WEBSITE DESIGN Students: Liu Yang, Chen Ming, Zhang Chi, Wang Feiyi

GQP 3 Project - Mathworks: UNIVERSITY LOOKUP Students: Li Hongnan, Yang Sijing, Saber Jahanpour, Chitra

GQP 4 Project - Pfizer: SCIENTIFIC DATA CLOUD Students: Azharudin Priyotomo, Yanpu Li, Congyuan tang

GQP 5 Project - EnerNoc: CLUSTERING OF TIMESERIES DATA BASE BASED ON A GIVEN DISTANCE FUNCTION Students: Cory Hayward, Marcus Moyses, Sachin Sudarshana

GQP 6 Project - Seceon: ANOMALY DETECTION IN CYBERSECURITY Students: Xing Liu, Pan Qijie, Barman Snehasish

GQP 7 Project - GRC: DATA ON DEMAND Students: Feiyi Wang, Nikita Mutta

Project Coordinator: Fatemeh Emdad, Ph.D.
Acknowledgements

• Data Science Steering Committee at WPI:
  • Prof. Elke Angelika Rundensteiner (Director of Data Science)
  • Prof. Mohamed Eltabakh
  • Prof. Eleanor T. Loiacono
  • Prof. Joseph D. Petruccelli
  • Prof. Carolina Ruiz
  • Prof. Diane M. Strong
  • Prof. Andrew C. Trapp
  • Prof. Domokos Vermes
  • Prof. Jian Zou
Questions?
rcpaffenroth@wpi.edu
http://www.wpi.edu/~rcpaffenroth

http://www.wpi.edu/academics/datascience.html
Backup
Pretty picture