Capacity Building: Data Science in the University At Rensselaer Polytechnic Institute (RPI)

Professor James Hendler
Director, Rensselaer Institute for Data Exploration and Analytics
The Rensselaer IDEA

- **Institute for**
  - Ubiquitously linked to everything at Rensselaer; potentiating major strengths

- **Data**
  - The basis of the 4th Industrial Revolution

- **Exploration &**
  - Leading in advancing “data science” through basic research

- **Applications**
  - Linked to global challenges
Data Dexterity Requirement

March 22, 2018

Rensselaer Introduces First in the Nation “Data Dexterity” Requirement for All Undergraduate Students

Requirement prepares students in all disciplines to use datasets to define and solve complex real-world problems

By Mary L. Martialay
Objectives for Data Dexterity

- Identify different types of data, information and evidence within the relevant discipline, and be able to discuss issues of data curation, validation, and uncertainty.
- Identify appropriate problems to which data can be applied, and discuss limitations, assumptions and interpretations.
- Effectively communicate about problems/issues in this field in which data is a relevant tool, including writing about, presenting on, and visualizing data.
- Discuss the ethical issues surrounding data in this field, including, but not limited to, responsible conduct of research, privacy, provenance, privatization, monetization, social implications.
Drew Conway’s 2013 Venn Diagram

Data science, due to its interdisciplinary nature, requires an intersection of abilities: hacking skills, math and statistics knowledge, and substantive expertise in a field of science.

Hacking skills are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.

Math and statistics knowledge allows a data scientist to choose appropriate methods and tools in order to extract insight from data.

Substantive expertise in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.

Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

The "danger zone"

DATA SCIENCE SKILLSET

- Hacking Skills
- Substantive Expertise
- Math and Statistics Knowledge
- Machine Learning
- Traditional Research

Data science, due to its interdisciplinary nature, requires an intersection of abilities: hacking skills, math and statistics knowledge, and substantive expertise in a field of science.

- **Hacking skills** are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.
- **Math and statistics knowledge** allows a data scientist to choose appropriate methods and tools in order to extract insight from data.
- **Substantive expertise** in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.
- **Traditional research** lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.
- **Machine learning** stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.
- **Danger zone!** Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

Teaching Data Science

• Science and interdisciplinary from the start!
  – Not a question of: do we train scientists to be technical/data people, or do we train technical people to learn the science
  – It’s a skill/ course level approach that is needed
• Teach methodology and principles over technology
• Data science is a skill, and natural like using instruments, writing/using codes
• Team/ collaboration aspects are key
• Foundations and theory must be taught

Lessons learned from Peter Fox
https://www.slideshare.net/brandsteve/data-science-for-every-student-at-rpi
Foundations must be taught
Hands on experience is crucial

• 3 day Analytics "Boot camp" offered January 2018
  • 52 participants (mostly SoE and SoS)
  • 7 faculty, 10 pdoc/staff, 35 graduate students
  • No cost to participants
  • Exploring a slightly less intense version to offer “on the road” in FY19
Team Collaboration
Data INCITE Lab: Learning by doing

Innovative pipeline for creating next generation of agile data scientist and data users

- Novel low-barrier course “on ramps”
- Partners with healthcare, industry and institutions for research projects
- Leads to student internships/coops/careers

About 200 students have done INCITE projects to date.

K. Bennett, RPI
Data INCITE Creates Impactful Research

- Data INCITE undergraduate researchers partnered with Neural Stem Cell Institute to understand RNA-Seq data from “brain in a dish” model
- Became Rensselaer Tool for Identifying Corticogenesis (Brain Development) windows of susceptibility to disease
- Zika-virus induced “microcephaly” in infants has window of susceptibility in first 30 days of development
- Students developed R to CAMPFIRE workflow
Using Data Science (and AI) to teach other topics

Cognitive and Immersive Classroom – The Mandarin Project

Cognitive and Immersive Systems Laboratory at EMPAC
Conclusions

• Capacity Building in Data Science requires
  – Teaching the Skills
  – Understanding the Technologies
  – Developing the Science
  – Practice, Practice, Practice