UNIVERSITY of WASHINGTON

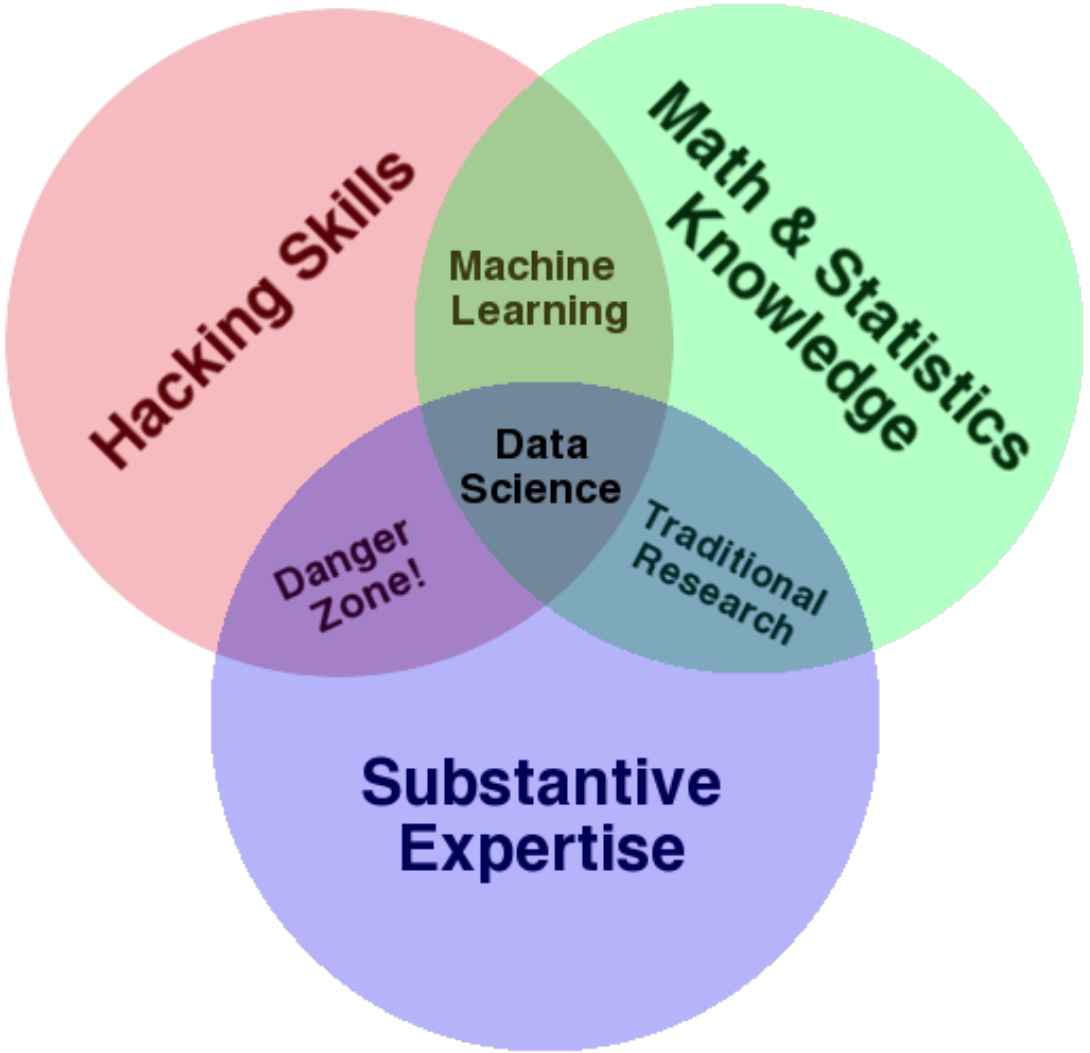
Introduction to Data Science

Bill Howe, PhD
Director of Research,
Scalable Data Analytics
University of Washington
eScience Institute

What is Data Science?

- Fortune
  - "Hot New Gig in Tech"
- Hal Varian, Google's Chief Economist, NYT, 2009:
  - "The next sexy job"
  - "The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill."
- Mike Driscoll, CEO of metamarkets:
  - "Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics."
  - "Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools & materials, coupled with a theoretical understanding of what's possible."

Drew Conway's Data Science Venn Diagram



What do data scientists do?

"They need to find nuggets of truth in data and then explain it to the business leaders"

-- Rchard Snee, EMC

Data scientists "tend to be "hard scientists", particularly physicists, rather than computer science majors. Physicists have a strong mathematical background, computing skills, and come from a discipline in which survival depends on getting the most from the data. They have to think about the big picture, the big problem."

-- DJ Patil, Chief Scientist at LinkedIn

Mike Driscoll's three sexy skills of data geeks

- Statistics
  - traditional analysis
- Data Munging
  - parsing, scraping, and formatting data
- Visualization
  - graphs, tools, etc.

"Data Science refers to an emerging area of work concerned with the collection, preparation, analysis, visualization, management and preservation of large collections of information."

An Introduction to Data Science

Jeffrey Stanton
Syracuse University School of Information Studies

"A data scientist is someone who can obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product."

-- Hilary Mason, chief scientist at bit.ly

"data wrangling"
"data jujitsu"
"data munging"

Three types of tasks:

1) Preparing to run a model

Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping, massaging

2) Running the model

3) Communicating the results

Data Science is about *Data Products*

- "Data-driven apps" (Mike Loukides)
  - Spellchecker
  - Machine Translator
- Interactive visualizations
  - Google flu application
  - Global Burden of Disease
- Online Databases
  - Enterprise data warehouse
  - Sloan Digital Sky Survey

Data science is about building data products, not just answering questions

Data products empower others to use the data.

May help communicate your results (e.g., Nate Silver's maps)

May empower others to do their own analysis (e.g., Global Burden of Disease)