What are the Optimal Data Science and Machine Learning Competencies for Informatics Professionals?

George Hripcsak, MD, MS
Biomedical Informatics, Columbia University
Medical Informatics Services, NewYork-Presbyterian
Columbia offerings
by depth of data science

• PhD – data science track
• PhD
• NLM postdoc
• research masters – 3 years, research intensive
• applied masters – 1-1.5 years, more practical
• postdoc
• clinical informatics subspecialty fellowship
• certification in HIT
• medical student
Survey recent students about data science needs

1. Real world experience
   – Large databases like OHDSI network
Survey recent students about data science needs

2. Biomedical research methodology in the big-data era
   – Asking the **right research questions**
   – Setting up data science teams and promoting collaborative, **inter-disciplinary research**
   – **Understanding the data**, its fit with the research questions, and its fit with the methods to answer the question
   – **Evaluation methodology** (e.g., beyond intrinsic evaluation metrics, how to evaluate the validity of data-driven methods as part of the biomedicine and health ecosystems) and how to conduct informative error analysis
   – **Ethical implications** of data sciences (e.g., implications of data-driven analyses on care of individual patients and populations).
Survey recent students about data science needs

3. Dissemination in the biomedical big-data era
   – Understanding the *socio-technical barriers* to reproducibility in biomedical data science
   – Best practices for making *open-source tools* and making them useful to the research community in biomedicine and health, as well as in data science
   – Best practices for making data, metadata, and data proxies *available* and making them useful to the research community
   – Ensuring the *reproducibility* of research methods and findings
   – **Documenting**, presenting, and publishing for different audiences with different backgrounds.
Approach

• Strong focus on computational methods
  – Informatics research should be on par with CS, stats
    • At the extreme, inventing new areas of machine learning
      (e.g., deep exponential families, Ranganath)
  – Link to source fields: shared courses
    • Tenure usually requires letters from source field like CS

• Reproducibility

• All our graduates are informatics professionals
  – And can potentially influence health care
PhD – data science track

- **Quantitative reasoning**: linear algebra, probability theory, and functional data analysis.
- **Programming** for data-driven processing: python and R.
- Diving into large biomedical, behavioral, and clinical datasets: (i) An inventory of datasets and their properties, including quality, structure, and semantics; (ii) Data quality from cleaning and preprocessing to data integration techniques; and (iii) Techniques for exploratory data analysis and for information visualization.
- **Manipulating large datasets**: programming for large-scale data processing (including programming paradigms, and impact of different storage mechanisms).
- **Learning** from large datasets: (i) Machine learning and statistical inference methods in the context of biomedical research; and (ii) Incorporating and leveraging the existing biomedical knowledge resources.
PhD, NLM postdoc, research masters, applied masters

• Understand and apply ML systems in new ways
• Computational Methods in Biomedical Informatics
  – Survey of the computational methods underlying the field of medical informatics. Explores techniques in mathematics, logic, decision science, computer science, engineering, cognitive science, management science and epidemiology, and demonstrates the application to health care and biomedicine.
  – Emphasis on discussion of algorithms and metrics, interpretation of results, and assumptions and limitations of methods.
  – Requires significant programming, and students will apply machine learning to real problems: network analysis of public health datasets, clustering of patients according to their clinical characteristics, language modeling of PubMed abstracts, and classification of tumors.
• 2 computer science courses (AMA 1), 1 statistics course
• other methods, electives, ethics
Certification in HIT, CI fellows

• Strong focus on EHR and challenges in decision support

• CI fellows do certification, but more focus on clinical impact
  – May pick an ML project
Medical student

• Four-year curriculum, much in decision support with some exposure to ML benefits and challenges