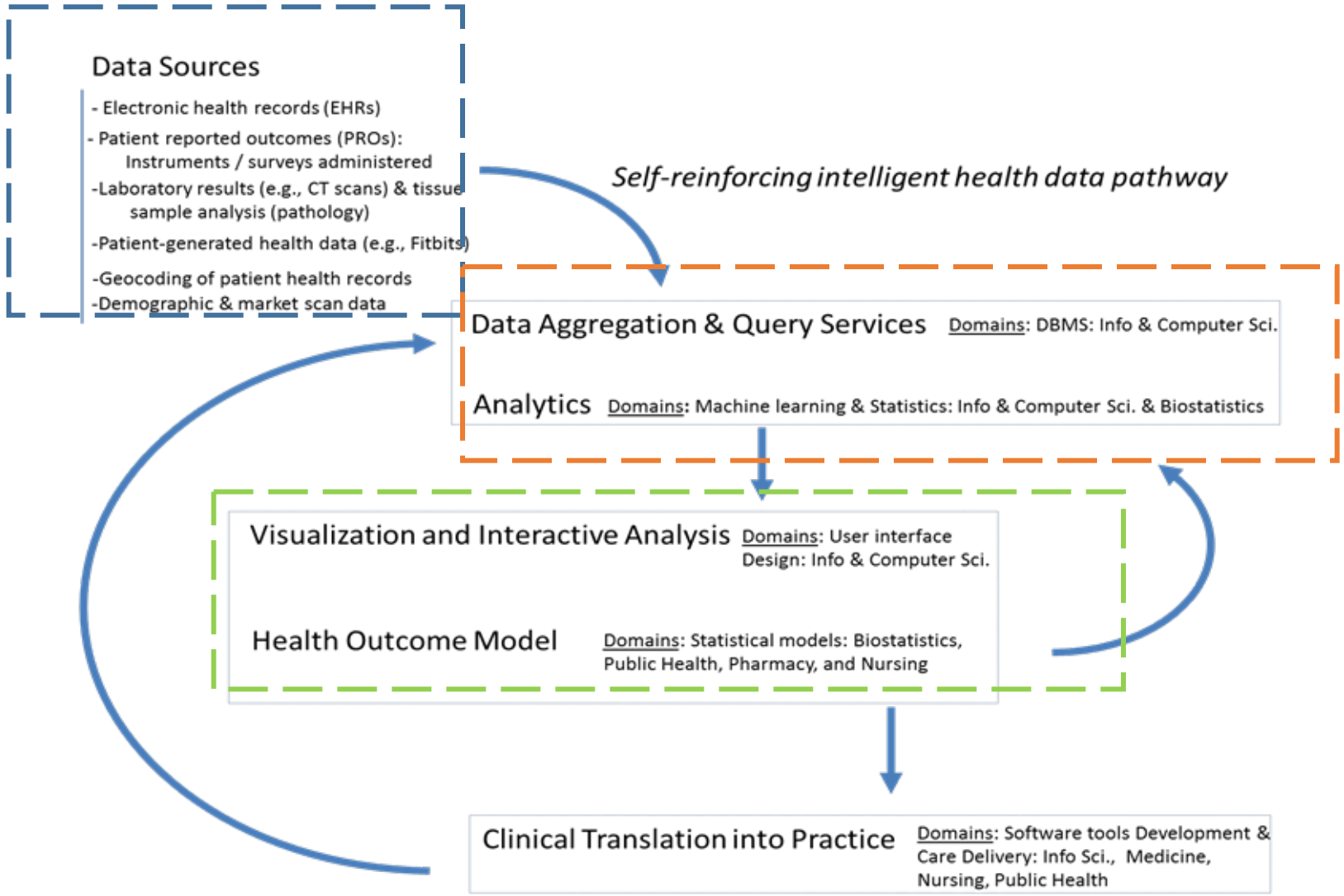
Interdisciplinary Program for Advanced Training in Health Data Analytics at UNC

Javed Mostafa
McColl Distinguished Term Professor
Information Science & Biomedical Research Imaging Center
The University of North Carolina at Chapel Hill

Outline

- Training program background
- NLM Supplement Key Activities
- Future plans

The Three Foundations of the Program & Program Pathway



Program Home & Team

- The home of the NLM T15 Training Program at UNC Chapel Hill is the Carolina Health Informatics Program (CHIP)

- CHIP is a highly interdisciplinary biomedical and health informatics program which is supported by jointly appointed faculty in 7 academic units: School of Medicine, School of Information & Library Science, Gillings School of Global Public Health, Eshelman School of Pharmacy, School of Dentistry, School of Nursing, and the Computer Science Department.

Current Program Core Faculty

Name	Title and Affiliation*	Expertise
Javed Mostafa, PhD	CHIP, Director & Professor, Biomed. Research Imaging Center (Sch. of Med.) & Sch. of Information & Lib. Science	Information retrieval, data mining, and user interface design
Kevin Jeffay, PhD	Gillian Cell Distinguished Professor and Chair, Department of Computer Science	Multimedia systems, networking, and improving communication systems over broadband
Ashok Krishnamurthy, PhD	Deputy Director, RENCI; Director of Biomedical Informatics, NC TraCS Institute	High performance computing, image processing, data Mining
Arcot Rajasekar, PhD	Professor, School of Information and Library Science; Chief Scientist, DICE and RENCI	Large-scale data management, data grids, digital libraries, big data analytics platforms
Arlene Chung, MD, MHA, MMCi	Assistant Professor of Medicine and Pediatrics, School of Medicine;	User-computer interaction, workflows, and mobile health computing tools
Haibo Zhou, PhD	Professor, Gillings School of Global Public Health	Bioinformatics and novel techniques for analyzing large-scale genomics and environmental data
David Gotz, PhD	Associate Professor, School of Information & Library Science; Assistant Director of CHIP	Visual analytics and biomedical data mining
Shariar Nirjon, PhD	Assistant Professor, Department of Computer Science	Data analytics, embedded systems, and wireless networks
Debbie Travers, PhD, RN, FAEN	Associate Professor, School of Nursing	Natural Language Processing and health information system usability
Sam Cykert, PhD	Professor, School of Medicine	Outcome assessment and health disparity tracking using informatics

UNC NLM T15 Supplement Aims

- To develop a *repository* with Clinical data linked to Genomic and Proteomic Data and to Image data (Radiological and Pathological images)
- To develop and test a set of *big data analytic methods* on this repository
- To develop a *curriculum for a course on Big Data Analytics* with 16 modules along with Use Cases, Teaching Problems with known results and a code repository of well-functioning code in Python and R.
- To *test the course at each of our partner* institution with student feedback and subsequent publication and dissemination of our results

UNC NLM T15 Supplement Collaboration

- UNC is directly collaborating with two other NLM T15 Supplement Sites, namely the University of Buffalo and Yale University

- The broader aim is to develop the course collaboratively, each institution taking responsibility for a subset of modules, share resources (data sets and software) and expertise, and upon assessment of content disseminate course, associated content, and observations/findings among all the T15 sites.

- The course modules follow …

Biomedical & Health Data Analytics Course

Module Title	Site	Content
1. Precision Medicine	Yale	This module will discuss the process from raw sequencing reads to annotated somatic mutations in cancer and immunological diseases, and neoepitope prediction. Quality control, types of variant callers, workflow management, annotation of missense variants. In the cancer domain, we will cover neoepitope prediction, including HLA calling and MHC affinity prediction. In the immunological domain, we will cover V(D)J assignment and somatic hypermutation calling for next-generation B and T cell receptor repertoire sequencing data sets.
2. Sequential Data Learning	Yale	Sequential data is ubiquitous in biomedical informatics (e.g., DNA sequences, biomedical sensor data, clinical data along time axis). This module will explore similarities among data types, investigate common analytical strategies, both for unsupervised and supervised learning for sequential data.
3. FAIR data sharing	Yale	This module will discuss ontology-based annotations, linked data sharing, including RDF encoding, triple stores, SPARQL endpoints. It will also discuss related topics including controlled vocabularies, and minimal standards for Omics Data
4. Fundamentals of High Performance Computing	UB / UNC	This course will teach parallel programming, HADOOP, and database optimization and indexing. We will teach distributed computing. It will teach NOSQL databases such as MONGO DB and Berkley DB. It will teach complexity theory, and how problems scale computationally.
5. Natural Language Processing and data reliability	UB	We will teach the students natural language processing in the context of indexing clinical and image data. We will discuss data reliability including data cleaning, missing data, duplicate data, conflicting data and unreliable data.
6. Biomedical Ontology	UB	We will teach students the principles of ontology. This will give the students a logical basis for how biomedical knowledge is represented stored and retrieved reliably. We will teach OWL DL and storage of RDFS triples in both a triple store and in a graph database. Students will learn how to reason over data represented in OWL DL.
7. Data Mining and Machine Learning	UNC / UB	Based on a standard and widely accepted set of software and software platforms and multi-format reference data sets, the module will teach students supervised, unsupervised, mixed learning, and deep learning approaches for addressing practical health care questions
8. Image Data Analytics	UNC / UB /Yale	Students will learn about the standard modalities of health image generation, as well as image storage, curation, and manipulation using standard analytics tools
9. Population Health Analytics	UNC / UB	Aggregation, annotation, storage, and data warehousing of large population-centric data sets and their practical applications based on visual and interactive data manipulation in the context of health care will be taught
10. Research Ethics, Privacy and Security in Big Data Science	UB/UNC	Privacy and security policies associated with health data that are rooted in national and local regulations and laws in the US will be taught, along with the ethics associated with data collection, long-term preservation, manipulation, and sharing. Students will also learn about related IT advances that support privacy-preserving data mining and secure health care data management.
11. Clinical Decision Support in the Era of Big Data	UNC	Access to comprehensive and longitudinal data on patients has made the potential of developing highly accurate and robust decision support aids. The design, work-flow analysis, architecture, deployment, change management, usability, and finally knowledgebase maintenance that go with modern CDS systems will make up the primary topics in this module.