Integrating Data Science into T32 Training Programs at IUPUI

Analysis and Recommendations for the Future

June 30, 2019

The effort and materials referenced in this report are work product from a grant from the U.S. National Library of Medicine of the National Institutes of Health under Award Number T15LM012502. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Executive Summary

Data science is critically important to the biomedical research enterprise. Many research efforts currently and in the future will employ advanced computational techniques to analyze extremely large datasets in order to discover insights relevant to human health. Therefore the next generation of biomedical scientists requires knowledge of and proficiency in data science.

With support from the U.S. National Library of Medicine, a team of faculty from Indiana University-Purdue University Indianapolis (IUPUI) facilitated curricula enhancement for National Institutes of Health (NIH) T32 research training programs with respect to data science. In collaboration with the existing NIH T32 Program Directors at IUPUI and the IU School of Medicine, the interdisciplinary team of faculty drawn from multiple schools and departments examined the existing landscape of data science offerings on campus in parallel with an assessment of the competencies that future biomedical and clinician scientists will require to be comfortable using data science methods to advance their research.

The IUPUI campus possesses a rich tapestry of data science education programs across multiple schools and departments. Furthermore, the campus is home to more than a dozen world-class T32 programs funded by the NIH to train biomedical and clinician scientists. However, existing training programs do not currently emphasize data science or provide specific curriculum designed to ensure T32 graduates possess basic competencies in data science. To position the campus for the future, robust T32 programs need to connect with the rapidly growing data science programs.

This report summarizes the rationale for the importance of connection and the competencies that future biomedical and clinical scientists will require to be successful. The report further describes the curriculum mapping efforts to link competencies with available degree programs, courses and workshops on campus. The report further recommends next steps for campus leadership, including but not limited to T32 Program Directors, the Office of the Vice Chancellor for Research, the Executive Associate Dean for Research Affairs at the IU School of Medicine, and the President and CEO of the Regenstrief Institute. Together we can strengthen the IUPUI campus and help ensure its T32 graduates are successful in their research careers.

Introduction

Background

Accessible, findable, well-organized, interpretable, secure, precise, and efficiently operated data resources are critical to modern biomedical research. Significant advancements in computation and information technologies make it easy to generate enormous volumes of data during the course of biomedical research, whether imaging the human brain or assessing the genetics of a single cell. These data must be wrangled, stored, retrieved, and analyzed efficiently by biomedical researchers. While computer scientists as well as informaticians specialize in the building of software to support research processes and analytical methods, ***all biomedical researchers need to be comfortable using advanced technical systems*** to capture, store, manage, use and share biomedical data that continues to increase in size, veracity, complexity, and variety. Therefore, biomedical researchers require knowledge of and proficiency in data science, defined as "the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data."

In its 2016-2020 Strategic Plan, the National Institute of Health (NIH) called out data science as essential to increasing the impact and efficiency of 'fundamental biomedical research.' The NIH subsequently published a Strategic Plan for Data Science in which it articulated specific priorities for the nation's biomedical research infrastructure:

1. Support a Highly Efficient and Effective Biomedical Research Data Infrastructure
2. Promote Modernization of the Data-Resources Ecosystem
3. Support the Development and Dissemination of Advanced Data Management, Analytics, and Visualization Tools
4. **Enhance Workforce Development for Biomedical Data Science**
5. Enact Appropriate Policies to Promote Stewardship and Sustainability

Specifically, in Objective 4-2, NIH asserts that

- "…it is essential that the next generation of researchers be equipped with the skills needed to take advantage of the growing promise of data science for advancing human health;" and
- "…NIH-funded training and fellowship programs [should] emphasize teaching of quantitative and computational skills and integrate training in data-science approaches throughout their curricula and during mentored research."

Programs funded by the NIH, such as the Big Data to Knowledge (BD2K) and NLM Institutional Training Grants for Research Training in Biomedical Informatics and Data Science (T15), provide programs and funding for training individuals who might specialize in data science. These programs further offer the opportunity to collaborate with Institutes and university-based training programs (T32, K-awardees) to offer data science training to the broader biomedical research workforce.

The Indiana University-Purdue University Indianapolis (IUPUI) campus supports a wide range of T32 grants that train approximately 85 next generation biomedical and clinical scholars each year. While these programs are diverse in their disciplines of medicine, nursing, engineering, psychology as well as biology, only one program formally includes competencies in data science.

Purpose of Document

This document provides a roadmap for the IUPUI campus to move towards ***integrating data science competencies into all biomedical research as well as clinical training programs***. It documents both existing programs and courses available to pre-doctoral and post-doctoral fellows in biomedical science. It further outlines potential pathways biomedical and clinical trainees might take to gain data science competencies during their career at IUPUI. This document is not meant to be prescriptive, but provide training program directors with options they can incorporate into their curricula to ensure biomedical researchers as well as clinical trainees leave IUPUI with the data science competencies they will need to be successful in their careers.

The document was developed by an interdisciplinary team of faculty from multiple schools and departments[1]. It was further reviewed by T32 program directors who provided feedback in which the team incorporated into the final version. The document is an information resource and should not be construed as a requirement for any program. While not all NIH training programs will require data science competencies to be integrated, we anticipate that most will require this component. And when not required, incorporating data science competencies into a training program will likely result in higher scores from reviewers since these competencies are increasingly recognized as important regardless of one's primary discipline.

Definition of Data Science

A critical component of our discussions with T32 program directors during the development of this document was the definition of data science. Therefore, we provide a summary of the discussions and a working definition of data science for IUPUI.

Researchers in the biomedical science fields have worked with data for decades, and the necessary skills for a successful research team have included a laundry list of components (from beginning to end): instrument design, including (physical and electronic) case report forms and documentation; data schema (either database or using an web-based data management software); data management (data cleaning for later analysis / visualization); data analysis; data reporting; and data curation over the course of the study and after a study has finished. Over the years, as more individuals have been trained in these sub-domains and the technologies utilized have become more familiar and widespread, a subset of individuals began to acquire skills in multiple domains, slowly becoming known as "data scientists".

While individuals have arguably been engaging in work that falls under the umbrella of data science for some time, deciding on the exact definition of what data science is has been elusive, and several governing bodies have developed their distinct, though fairly equivalent, definitions. The NIH, in their Strategic Plan for Data Science published in 2018,[2] define data science as "the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data". The National Academies of Science, Engineering, and Medicine[3] defines a data scientist as follows: "The term 'data scientist' typically describes a knowledge worker who is principally occupied with analyzing complex and

[1] Authors and team members listed at the end of the document.
[2] https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf
[3] http://nap.edu/25104

massive data resources. However, data science spans a broader array of activities that involve applying principals, for data collection, storage, integration, analysis, inference, communication, and ethics".

Synthesizing available definitions, IUPUI defines **data science** as "an interdisciplinary field of inquiry that applies mathematical, information science and computational methods across the 'Big Data pipeline,' including data collection, management, analysis and visualization; storytelling with and synthesis of data; dissemination and curation of data; and the ethics involved in all of these activities. Any individual who is engaged in one or more of these activities, in the context of Big Data, is engaging in data science. Most data scientists choose to specialize in a few of these activities and complete advanced training in the field. Furthermore, data scientists pursue continuous training in these areas due to the evolving nature of infrastructure, software development, and analysis methodologies, including statistical and machine learning methods.

There are several key points that should be addressed in order to fully understand Data Science today. Among them are the relative importance of data cleaning skills, the wide breadth of the field of data science, and the increasing importance of cloud infrastructures. First, while data organization, data analysis, machine learning, and dissemination are all important parts of the Big Data pipeline, one should not understate the importance of data cleaning and pre-processing. The majority of data encountered in the wild is of widely varying formats, naming schemes, sizes, and levels of security. This leads to data cleaning and pre-processing making up an estimated 80% of data science workflows on a day-to-day basis. Data science professionals will find sufficient experience with data wrangling of great benefit in order to contribute in this domain. However, due to the very broad nature of Data Science as a field, data science professionals will likely choose to focus on a smaller set of domains, for instance two to three of the domains mentioned earlier in this document, and a team-based approach will be utilized for building data science products. Finally, as cloud infrastructure becomes more convenient, both in terms of physical limitations such as space and cost considerations, as well as cheaper to acquire, a transition from on-site computing resources to cloud-based resources is naturally occurring. It will be important for future data science professionals to understand and be comfortable with utilizing a cloud infrastructure to complete their work.

Data Science Competencies

Training in the biomedical and health sciences has increasingly moved towards competency-based education. Therefore, our team sought to establish a set of core competencies for data science before we identified or created training opportunities in data science for IUPUI.

Based on discussions with T32 program directors, we recognized that biomedical and clinical trainees require unique pathways. Therefore, we established unique but complementary sets of competencies for each group. The pathway for biomedical researchers emphasizes two levels. The clinical research pathway has three levels that loosely align with the divide between undergraduate medical education, fellowship, and post-doctoral training. For both groups we advocate training programs to try to ensure all trainees achieve mastery of the Level 1 competencies. The other level(s) would be goals for trainees who wish to further pursue mastery of advanced data science skills.

Biomedical Research Track Competencies in Data Science

The following competencies were designed for biomedical researchers. These competencies are designed to produce biomedical researchers who can leverage data science methods in the conduct of

their primary discipline work and collaborate with data scientists in alignment with 'team science.' Mastery of all competencies is not equivalent with training in data science as a primary discipline. Individuals seeking to become data scientists would likely require advanced study and training.

Level 1 Competencies in Data Science for Biomedical Research

1. Define data science and explain how the field facilitates the conduct of biomedical research.
2. Recognize statistical foundations such as Hypothesis testing, including Type I error rate, adjustment for multiple comparisons, and Power
3. Ability to explain importance of and central role of data management and curation of data, including access and data wrangling/cleaning
4. Describe the principles of data description and visualization, including commonly used aggregate measures such as the mean, median, and visualizations for summarizing different data types
5. List and describe common sources of data used by researchers to generate hypotheses and assess pre-existing hypotheses
6. Identify commonly used data standards in biomedical research
7. Describe the importance of using data standards when collecting research trial data or using preexisting biomedical data sources
8. Explain and document the role of data modeling and assessment of data that has already been wrangled and transformed into a suitable format
9. Provide justification and background regarding importance of workflow documentation (with respect to scientific data and information) and reproducibility of analytical work
10. Explain the critical role that a data scientist or informatician brings to the research team (e.g., team science)
11. Demonstrate how communication and teamwork facilitate completion of an analysis of a large and possibly unorganized data set
12. Summarize unique needs determined by domain-specific considerations, i.e. difference between clinical trials and big data in genome wide analyses / neuroimaging data
13. Recognize the importance of utilizing ethical problem solving in all data science activities, including example of unintended consequences of unchecked models (e.g., Weapons of Math Destruction example)

Level 2 Competencies in Data Science for Biomedical Research

ALL Basic Competencies, plus the following:

14. Successfully wrangle data, including filtering / sub-setting operations, creation of new variables (feature engineering, phenotyping)

15. Visualize data in an appropriate fashion based on the type of data being utilized (quantitative, qualitative, continuous, discrete, nominal, ordinal, etc.)
16. Connect to a biomedical/health database or web service to extract data for cleaning/wrangling/visualization
17. Demonstrate proficiency in using reproducibility workflows, for instance one of the following: jupyter notebook, RMarkdown, or another similar tool
18. Employ statistical tests or modeling techniques using a scripting language such as R or Python
19. Run a task in parallel using a HPC or cloud-based environment solution
20. Explain the role of a system like Github or Dryad (what do these systems do? What is their role in science?)
21. Write code for or analyze code for GitHub or Dryad

Clinical Data Science Competencies

These are preliminary competencies for clinicians as well as clinician scientists but designed to align with the IU School of Medicine overall plan for clinical training. The competencies have also been created with psychologists, pharmacists, nurses, and other clinicians in mind.

These competencies would be geared toward clinical postdocs, or physician scientists doing career award training. Some competencies might be relevant to medical students or residents.

Level 1 Competencies in Data Science for Clinical Research

1. Define data science and explain how the field supports care delivery
2. Recognize statistical foundations such as Hypothesis testing, including Type I error rate, adjustment for multiple comparisons, and Power
3. Explain the critical role that a data scientist or informatician brings to a clinical research team (e.g., team science)
4. Recognize the importance of utilizing ethical problem solving in all data science activities, including example of unintended consequences of unchecked models
5. List and describe common sources of data used by clinicians to deliver care and researchers to conduct clinical trials
6. Identify commonly used data standards in health care
7. Describe the importance of using data standards when collecting clinical or research trial data
8. Explain the concept of clinical decision support (CDS) and its role in supporting high quality care delivery

9. Identify commonly used tools by researchers to analyze data and clinicians to access the results from computational models.
10. Describe the process of clinical phenotyping and explain its importance to both clinical practice as well as clinical research.

Intermediate Competencies in Data Science for Clinical Research

11. Distinguish and classify data based on its type (e.g., numeric, text) and scale (e.g., velocity, variety)
12. Employ basic statistical tests such as Hypothesis testing, including Type I error rate, adjustment for multiple comparisons, and Power using SAS or R
13. Explain the importance of and central role of data management and curation of data, including access and data wrangling/cleaning
14. Describe the principles of data description and visualization, including commonly used aggregate measures such as the mean, median, and visualizations for summarizing different data types
15. Explain and document the role of data modeling and assessment of data that has already been wrangled and transformed into a suitable format
16. Provide justification and background regarding importance of workflow documentation (with respect to scientific data and information) and reproducibility of analytical work
17. Explain the challenges to implementing CDS systems in health care organizations
18. Demonstrate how communication and teamwork facilitate completion of an analysis of a large and possibly unorganized data set
19. Summarize unique needs determined by domain-specific considerations, i.e. difference between clinical trials and big data in genome wide analyses / neuroimaging data
20. List and explain the FAIR (Findable, Accessible, Interoperable, Reusable) principles.
21. Explain the concept of data stewardship and identify responsibilities for data stewardship by members of the research team.

Level 3 Competencies in Data Science for Clinical Research

22. Wrangle data, including filtering / sub-setting operations, creation of new variables (feature engineering)
23. Perform clinical phenotyping using common health data standards in combination with other variables in a dataset
24. Create a predictive model using common regression techniques
25. Evaluate the performance of a predictive model

26. Visualize data in an appropriate fashion based on the type of data being utilized (quantitative, qualitative, continuous, discrete, nominal, ordinal, etc.)
27. Connect to a clinical/health database or web service to extract data for cleaning/wrangling/visualization
28. Demonstrate proficiency in using reproducibility workflows, for instance one of the following: jupyter notebook, RMarkdown, or another similar tool
29. Employ statistical tests or modeling techniques using a scripting language such as R or Python
30. Run a task in parallel using a HPC or cloud-based environment solution
31. Explain the role of a system like Github or Dryad for archiving and/or distributing reusable code or data sets in accordance with the FAIR principles.

Existing Data Science Curriculum at IUPUI

Before proposing new training workshops or courses for IUPUI, our team conducted a comprehensive review of available programs (e.g., MS, PhD), courses, and workshops offered by schools or the university. We searched the Indiana University as well as the IUPUI course catalog and websites of relevant units (e.g., School of Informatics and Computing) to identify potential programs and courses. We further searched the main IU and IUPUI websites to identify workshops or online trainings available from UITS (University IT Services) or similar units. This list was reviewed by our team of which was composed of faculty from various schools and units, and it was further reviewed by T32 program directors. While we did our best to identify all possible programs, courses, or workshops, our list might be incomplete due to the fact that no central clearinghouse for data science exists at IUPUI.

Next, we mapped the available curriculum at IUPUI onto the consensus-based competencies developed for this report. A faculty member in the IU School of Medicine created the first version of the curricular map with support from the Center for Teaching and Learning. Next, a faculty member whose primary research is at the intersection of data science and population health completed the curricular map by examining program materials as well as course syllabi. This document was then reviewed by our team for accuracy and completeness and was further circulated to T32 program directors for review and input.

The document, entitled “IUPUI/IUSM Data Science Program Map for T32 Grant Programs,” is available on Box for access by T32 Program Directors as well as other interested parties.

https://iu.box.com/s/ni8rfxsik45jjh6tvtft3148hcrby3e0

The document contains three pathways:

1. Track 1 – PhD Pathway; this pathway is designed for pre-doctoral trainees who are seeking a PhD in a biomedical discipline. The pathway contains suggested courses, minors, and other forms of training that could augment a student’s primary concentration. Individual courses could further be added to a trainee’s program of study.

2. Track 2 – Postdoc Pathway; this pathway is designed for postdoctoral fellows who are focused on advancing their research career. The pathway contains suggested certificate programs and Master's degree programs along with workshops that could augment the primary research training program. Individual courses could further be added to a trainee's program of study.
3. Track 3 – Clinical Pathway; this pathway is designed for clinical fellows or postdoctoral physician scientist programs. The pathway features certificate programs and workshops that might augment the primary research training. Individual courses could further be added to a trainee's program of study.

Courses, programs, and workshops are listed in the rows of the spreadsheet. Each row represents an individual course or workshop. The competencies are represented as columns. When an individual cell contains an 'X' this indicates that the corresponding course or workshop in the row covers the corresponding competency in that column. In example, PBHL E647 (Introduction to Population Health Analytics) covers a total of three (3) competencies; two (2) in Level 1 and one (1) in Level 2. The course syllabus specifies eight learning objectives, so the competencies are mapped by individual learning objective. Many courses do not specify learning objectives in the syllabus available on the university site; therefore, our team could not map competencies to individual learning objectives. Instead, competencies were mapped to the overall course number based on our review of the syllabus. Columns shaded grey indicate that our team could not map a given competencies to an available course or workshop on campus. Rows shaded light grey indicate that a given course could not be mapped to any of the competencies in Level 1 or Level 2.

We envision that a T32 or other biomedical science training program faculty could use the curriculum map to identify programs (e.g., Graduate Certificate in Biomedical Analytics), individual courses, or university-based workshops to recommend (or require) of their trainees. Some programs in data science might naturally complement a trainee's primary discipline. For example, the SOIC doctoral minor in Bioinformatics might complement a trainee's plan of study in a medical genomics PhD program. In other cases, trainees might need to take 2-3 individual courses to obtain a specific set (or Level) of competencies in data science. Training program faculty will make these determinations.

NOTE: Our team's efforts concluded on June 30, 2019, at the conclusion of targeted funding from the NLM to conduct the work described in this document. While the document could be updated in the future, there is no expectation that the team members will update the document on a regular basis. The document is not meant to provide an up-to-date assessment of curricular offerings but a one-time environmental scan to inform the recommendations in this document.

Recommendations for Data Science in IUPUI T32 Programs

After open discussions with T32 program directors, our team recommends the following plan for integrating data science into T32 programs based at IUPUI:

1. The campus should offer a workshop that will cover the Level 1 competencies for biomedical and clinical researchers on campus.
    a. Existing workshops, courses, and programs do not completely cover all of the competencies identified by our interdisciplinary team. Therefore, the campus would benefit from a concerted effort to develop a comprehensive, introductory workshop available to T32 trainees.

b. The workshop would ideally be offered online and available year-round. Alternatively, an in person workshop could be offered during the summer.
    i. The workshop was designed to be offered either online or in a flipped classroom approach where students view lectures online then come to class to perform more hands-on tasks such as data wrangling, analysis, visualization of data using university computing resources.
    ii. The workshop could be offered as two half-day events where students work on exercises after viewing didactic content online.
c. Two distinct workshops would be offered – one geared towards biomedical researchers and one geared towards clinical researchers. Some of the modules might overlap, but there are distinctions in the Level 1 competencies for these two pathways.

2. Trainees who complete the workshop and desire to continue training in data science methods and techniques should add a program or course to their training plan.
    a. Most training plans are individualized, especially at the postdoctoral level.
    b. The curriculum map provides several options for trainees based on their interests and primary discipline. Faculty should work with trainees to navigate the map and select the right fit for each trainee.
    c. PhD students may wish to pursue a doctoral minor in data science to better position themselves in the marketplace post-graduation.
    d. Postdoctoral fellows should consider a certificate in analytics or data science to enhance their marketability following completion of the fellowship.

Suggested Curriculum for a Workshop to Introduce Data Science

Our team recommends the following structure and curriculum for the two workshops. Although this team was not charged with offering the workshop, we provide our thoughts to assist those who may take up the charge to organize and implement the workshop on the IUPUI campus.

The sample curricula are available for download from Box. The suggested sequencing and divide between didactic and hands-on activities are suggestions only. Organizers are free to amend the curricula and further break topics apart into smaller modules. Please note that the design assumes that the workshops will be offered using a flipped classroom approach in which online didactic material is available to trainees ahead of in-person, hands-on sessions during which studies work through real-world biomedical or clinical scenarios by applying concepts from the online modules.

Workshop for Biomedical Researchers

https://iu.box.com/s/sbxug9cjzj6tb84soewn8ew4yw7qkwux

Workshop for Clinical Researchers

https://iu.box.com/s/oa6kn5t417eteklfnhr073ea8kjgpaqx

Our team was not commissioned to develop the actual curriculum for the workshop or organize it for implementation during the 2019-2020 academic year. Team members could be engaged in planning or teaching components of the workshop by a program or faculty who desires to organize the workshop.

We recommend that the Office of the Vice Chancellor for Research (OVCR) in partnership with the IU School of Medicine identify one or more faculty willing to organize the workshop and offer it starting in mid-2020. The effort would benefit the multiple T32 programs on campus. The effort could be funded through an internal grant mechanism or by pursuing a supplement to an existing NIH grant.

Report Authors and Contributors

Principal Author

Brian E. Dixon, MPA, PhD, FACMI, FHIMSS
Co-Director of the Indiana Training Program in Public Health Informatics (NLM T15 Grant based at IUPUI)
Associate Professor, Department of Epidemiology, IU Richard M. Fairbanks School of Public Health
Director of Public Health Informatics, Center for Biomedical Informatics, Regenstrief Institute

Co-Authors and Major Contributors

Suranga N. Kasthurirathne, PhD
Visiting Assistant Research Professor, Department of Epidemiology, IU Fairbanks School of Public Health
Research Scientist, Center for Biomedical Informatics, Regenstrief Institute

Julia C. Stumpff, MSLIS
Instructional Design Librarian, Ruth Lilly Medical Library
Indiana University School of Medicine

Spencer Lourens, PhD
Assistant Professor, Department of Biostatistics
IU Richard M. Fairbanks School of Public Health

Other Contributors to the Work

Kun Huang, PhD, FAIMBE
Professor, Department of Medicine
IUSM PHI Chair for Genomic Data Science
Director of Data Science and Informatics, Precision Health Medicine
Indiana University School of Medicine

Yunlong Liu, PhD
Professor, Medical & Molecular Genetics, Biostatistics, and BioHealth Informatics
Director, Center for Computational Biology and Bioinformatics
Director, Center for Medical Genomics
Indiana University School of Medicine

Sarath Janga, PhD
Associate Professor, Department of BioHealth Informatics
School of Informatics and Computing, Indiana University Purdue University Indianapolis