MBDH Initiative: Data Science in Small Colleges

Lior Shamir, Small College Data Science Collaboration
Kansas State University

Synopsis

While large R1 institutions receive the vast majority of the available federal funding, these institutions train less than 15% of the national IT workforce. The majority of the students are trained at community colleges (40%), Carnegie Master's universities (23%), liberal art colleges (7%), and others. That is, according to the existing models of higher education support, the vast majority of the students do not have access to the resources provided by the federal government.

Moreover, non-R1 institutions serve many of the least privileged student populations: the financially disadvantaged student who takes courses at their local community college, the working student who can only take evening classes, the student who has childcare responsibilities and takes classes at the institution nearby his/her house, or the first-generation college student and the underrepresented minority student who often do not see an exclusive elite school as their natural next step after graduating from high school. These students need the intervention of the federal government more than most other student populations, yet the institutions they attend receive disproportionally low federal support.

The initiative of Big Data and Data Science at non-R1 institutions aims at changing the existing reality to provide equal access to all Data Science students. With the Midwest Big Data Hub, we aim at identifying ways of sharing resources and services between small and large institutions, including the needs of non-R1 institutions in decision processes, and designing a new paradigm of federal and academic partnership. As a new discipline, we have the unique opportunity to design Data Science in an inclusive manner, embracing equity for all student populations, and creating an inviting and supportive environment for all genders and ethnicities.

Challenges

Education and Curriculum Development

Communication of importance of data science
- Getting past different language across domains; data science vs. data analytics
- Selling the importance of data science (what it can offer) to non-technical disciplines with good examples
- Why is data science different than math/CS/stats majors?

Interdisciplinary concerns in curriculum development
- Resolve course/resource conflicts across departments
- Faculty expertise spread across multiple departments
- Incorporating disciplinary perspectives within a set of 6-8 courses
- Who should teach the courses?
- Advising students across different departments
- Balance between programming and statistics skills

Student preparation
- Weak math preparation of incoming students is barrier to broader participation
- What are appropriate prerequisites?
- Developing courses at the appropriate level

Faculty development
- Training opportunities (e.g., HPC resources)
- Collaboration in lieu of each faculty member having enough time/expertise by themselves
- Release time to learn and develop teaching resources
- Finding adjunct faculty with domain/technical expertise who can also teach
- Quickly changing platforms

Sharing curriculum and course materials in a repository
- Identify common learning goals, core competencies, objectives
- Validated assessment
- Bachelors vs. Masters programs
- Common curriculum development
- Syllabi, course materials

Inclusion and Diversity

Overcoming Historic & Current Barriers to Enrollment
- Lack of Role Models for URG
  - Hiring faculty from URG (not hired and hard to find)
- Lack of enrollment in pipelines to Data Science
- Lack of awareness of Data Science as a Career, especially for diverse populations
- Overcoming Inherited Cultures in Departments Supporting Data Science
  - Department of first courses can set the culture for the entire program.
  - A bad experience with a department supporting only a portion of the degree could cause students to leave the program.

Externally Supporting Diverse Students
- Lack of available internships and fellowships in wide range of topics

Internally Supporting Diverse Students
- Inclusive Curriculum
  - Need for a wide variety of topics
  - Need to collaborate across departments and/or institutions
- Accessible Curriculum (no background or late exposure)
- Relevant and Approachable Data Sets
- Safe Spaces & Open Culture

Lowering Financial Barriers and other barriers to Success
- Free or low cost tools
- Free or low cost computers
- Flexibility in Schedules, in an already hard to set up interdisciplinary curricula
- Unaware of system, especially 1st generation

Industry Partnership

Internships:
- Being aware of the internships that are available.
- Making it feasible for students to find these opportunities.
- Developing internships that require a variety of skills, at differing levels.

Job opportunities:
- Identifying initial job opportunities for students completing programs with different skillsets.
- Design/Reverse-engineer aspects of our curriculum which reflect the actual needs of companies; to the extent possible, involving industry partners in the vision of curriculum.

Research opportunities
- Working with partners (industry, government, and academic) to identify relevant research problems, projects, and supply needed data sources.

Communication:
- Making industry aware that non-R1 institutions could be an untapped resource of students and faculty knowledge.
- We need to articulate clearly and persuasively about discipline-specific jargon in data analytics and data science and their applications to different domains.
- Communicate our ability to equip students for different levels of the data science work force (entry-level, research, policy)
- Establish formal means of communication with industry in a variety of venues such as symposia or workshops to showcase student portfolios and offer ways to engage the regional economy with the non-R1 institute resources.

Research

- Funding, including NSF funding for teaching and learning scholarship.
- Mid-sized awards
- Funding from smaller projects (smaller than a typical NSF award).
- Data access from industry
- Research opportunities for research topics that do not have research program at the school.
- Easy access to industry research, including data access and legal support
- Access to domain scientists.
- Partnership on grants between R1 and non-R1 institutions. Partnership with R1 researchers for grant proposals.

Resources

Hiring and training
Hiring: difficult to get and qualified people; in-demand field; pay is non-competitive
Retraining, skill development, and networking opportunities

Computing resources
- Access to HPC
- Access and cost of software and tools
- Keeping pace with industry changes and needs
- Computing resources in big data courses (availability?)
- Cost and availability

Real world projects
- Getting access to real projects or simulations production environment
- Opportunities for publication

Funding/resources (people and process)
- Lacking for non-R1 research
- Curriculum revision and development at pace with employer needs

Institutional Readiness

- Need a common language/framework that all students get in a common course.
- Moving past tradition.
- Support for data science from across campus -- strategies for explaining significance to colleagues in non-quantitative fields. Demonstrating applications to other disciplines, e.g., text mining, social networks.

- Better resources at the Career Development Office in non-R1 institutions to offer pointers to students looking for professional opportunities in data science.
- Pathways to establish connections between non-R1 universities and hubs to access resources for research, career development, education and training.
- Articulating the value of investing in non-R1 universities to industry and government sponsors and clarifying the value proposition offered by non-R1 universities.
- Addressing the challenges to retrain or augment the skills of existing IT professionals to incorporate data science tools and thereby meeting industry needs.
- Understanding that non-R1 universities can meet the industry needs on various levels (not at the scale of R1 universities).
- Creating layered networks of industry sponsors who are willing to invest in non-R1 institutions as an alternative to using crowdsourcing for their data science needs.

- Non-R1s do not have an effective office to communicate with industry partners.
- Non-R1 need often smaller grants that are not provided by the NSF.
- Non-R1 need grants for teaching and learning scholarship in data science.

Training for faculty and students at non-R1 institutions
Development of standardization of curriculum
- Basic data science for professionals in industry
- Outside of the typical disciplines
- Online videos/courses created for the expressed purpose of training faculty, students, and industry managers
- Workshops in strategic locations that allow the smaller institutions to attend more easily
- Formalized mentorships with faculty in data science at the larger institutions

OER (open educational resources)
- Seed grants to modify, improve existing curriculum
- Access to HPC and software
- Questionnaires with, for example, type of research, resources needed to better assist in aligning available resources
- Hiring Resources

Vision and Activities

Setup materials
- Information to share with Administration about need or possible benefits

Program Level Material
- Create a common language/framework
- Define Core Material versus Elective Material
- Example Curriculum

Course Level Material
- Lecture layouts
- Assignments
- Data sets
- A center for online quality data science courses
  - Expert courses in topics that cannot be covered at non-R1 institutes
  - Ran as a typical distance course
  - Collaborate teaching
  - Varying Levels
  - Some with minimum or no prerequisites

Create an online resource with:
- 5-10 best practices for diversity inclusion in data science programs
- 5 optimized, inclusive data science exercises
- Diversity-friendly student opportunities for fellowships or internships
- A discussion and resource area for contributions

Identify diversity experts in data science
- A group of experts who can be contacted to help programs become more inclusive or diverse.
- Analysis of different data science programs and the potential benefits or hindrances to diversity and inclusion.

Courses:
- Engaging collaborative intro to data science courses: i.e. data science in economics, biology, etc… especially non-STEM courses
- HUB sponsored Diversity and Inclusion webinars.

Companies having standing contracts maintained and held by the HUB
- Companies have contracts built around projects that they would like to have done
- Colleges/universities can go to this repository of projects and select projects that might fit for them
- Could be categorized by analytical or technical level (low, med, high)

Resource center: location where an employer can find graduates or active students based on skill sets taught in the various programs
- Best practices available like bringing industry representatives into the classroom to discuss what is expected/needed
- Maybe supported with some funding?
- Workshops to bring industry partners (or potential partners) together with academics (colleges and universities) to discuss how to work with each other, expectations, timelines, defining mutually beneficial projects, etc.
- Webinars from industry partners
- What is data science at GM?
- What is data science at Amazon?
- Students/faculty and active projects or expertise
- This could be a searchable database

Industry sponsorships/grants/scholarship
- Showcasing success stories
- Skill development workshops/webinars
- Internships

Data Access
- A hub of resources where institutions can find sources of data
- Access to research in best practices in data science education and teaching.
  - Techniques to attract and introduce students to data science
  - Information on curriculum development and funding
- Access to sample datasets in various domains that are immediately available for use.
- A hub where industry and government agencies can find research centers (including faculty experts and student research groups)
- Access to specialized computing resources, e.g. high-performance computing and big data resources.

Information on Data Science tools
- Commonly used software
- Open source tools
- Commercial software with educational programs

Best practices in obtaining data and ensuring privacy.
- Policies and procedures
- Sample documents such as: release forms, NDAs

Domain-expert access: A hub to find experts in diverse disciplines - including STEM and liberal arts. (Any field for which a research group would need domain-expertise.)
- Networking:
- A resource for connections between peers.
- A resource for listing venues where research can be presented.
- A resource for connecting with peers to establish conferences/workshops/symposia to disseminate research findings and attract stakeholders in data science research and educat

Collaboration
- A resource for collaboration between researchers in:
  - non-R1 institutions
  - R1 and non-R1 institutions.
- A resource for students to work on research projects in
  - non-R1 institutions
  - R1 and non-R1 institutions.
- A resource for larger grants from the NSF or other agencies, which can be shared among collaborators.

Data Access
- An archive of "real world" data sets that could be used in classes
- Organized to highlight diversity of data
- consistent data definitions, metadata, ethics, compliance, reproducibility, intellectual property

Networking
- Student networking opportunities
  - Industry matching for student projects and graduate program matching
- Faculty collaborations between industry and academia
- Meet-ups and informal gatherings

Funding
- Federal agencies and foundations
- Industry/academic partnerships
- Awards for Faculty and students
- Grant management
- Scholarships/fellowships