A Vision for Implementing the Presidential Initiative in Data Science at the University of Oregon

A report providing guidance and recommendations for the implementation of the Presidential Initiative in Data Science

by

the University of Oregon Data Science Visioning Committee (Appendix 1)

TABLE OF CONTENTS

Executive Summary

Introduction. The sophisticated analysis of large and varied data sets is transforming society and scholarship within universities, including the University of Oregon. Faculty and students have initiated and grown diverse **data science** activities across campus, primarily in isolation of one another. Our University is now prepared to develop an integrated data science effort to support and enhance these activities by building upon our long history of being a liberal arts research university with a strong culture of interdisciplinary activities. Here we provide a roadmap for supporting the growth of data science through strategic efforts and investments that will significantly advance the University of Oregon's research and educational mission. The Provost convened a Data Science Visioning Committee in Fall 2017 that comprised constituents from across UO. Our charge was to generate guidance to address four major categories of consideration for the creation of a data science program: tools and technical resources, space requirements, educational framework, and initial intellectual foci. The recommendation was strongly informed by existing efforts at the University of Oregon, other programs at top universities, as well as the needs for novel training experiences for our students. The following document encompasses the major findings and recommendations of the Visioning Committee.

Core Principles for the Data Science Initiative

- Be well connected across the university and include both research and education.
- Build upon the long history of interdisciplinary research at UO.
- Capitalize on the identity and strength of UO being a liberal arts research university.
- Leapfrog historical structural weaknesses in applied math and computation
- Help build bridges to new academic and societal partners.

These principles led the committee to propose a data science ecosystem comprising numerous interacting and interconnected components both within and outside our university. This 'hub-and-spoke' model can promote connectivity along existing areas of scholarship and education at UO, and also support and guide the organic growth of the initiative over the next decade.

A proposed data science ecosystem that comprises three major categories

- A **methodological hub** that contains much of the research and educational expertise for implementation of data science such as data collection, database creation and curation, as well as statistical and computational analysis. In addition, several aspects of design, communication, ethics and law could also be components of the core.
- **Domain spokes** that are areas where the research and educational activities focused primarily on a particular field or problem. Faculty in these spokes may gather large amounts of data and use advanced analytic tools that are common to those in the hub, but their primary research and educational focus will always be first on their domain area of expertise.
- A **ring of connectivity** that facilitates UO to have greater impacts on the world by increasing connectivity with other academic institutions, government agencies, and industry partners.

The committee sees moving next to a phase of implementation beginning in AY2018/2019 under faculty and administrative leadership that encompasses the entire campus. Implementation will need to be sustained and grow through numerous years to support in order to expand research foci that will attract existing faculty, make concerted efforts to hire faculty who will be able to support key components of the hub and spokes, develop novel and integrative data science curricula for undergraduates and graduate students, compete successfully for significant 'center-type' support from federal agencies and private foundations, expand research collaborations in data science across sister institutions in Oregon and along the West Coast and develop attractive opportunities for philanthropic support.

I. Introduction

What is data science and why is important for the future of the University of Oregon?

The use of increasingly sophisticated quantitative and computational tools to analyze expanding data sets is transforming scholarly activity within universities. Our society is awash with data used in ways never before possible. Once disparate disciplines are now converging, providing the opportunity to create novel synergistic research and training programs in data science (see **Fig. 1** for a conceptual overview of Data Science). This emerging discipline comprises the development of advanced computational and quantitative tools and training programs for gathering and analyzing large amounts of data and associated metadata, and well as interpreting and communicating the findings and their implications for society. The University of Oregon has already transformed the way that it supports Data Science research and education in several ways. Transform IT has modernized UO's approach to providing information services. The Research Advanced Computing Services (RACS), with the Talapas supercomputer at its core, is a key new research facility. The Knight Campus for Accelerating Scientific Impact (KCASI) will attract computational and quantitative scientists. Most importantly, data science activities are occurring in numerous schools and departments across campus through the hard work of faculty and students. Our goal here is to provide a roadmap for supporting and capitalizing on these investments in data science through strategic efforts and investments that will significantly advance the University of Oregon's mission to produce world class research and provide excellent education.

Charge and goals of the visioning committee

The charge of this committee is to generate a proposal to me (Provost) that addresses the following four major categories of consideration for the creation of a data science program: tools and technical resources, space requirements, educational framework, and initial intellectual foci. I anticipate that your recommendation will be strongly informed by other programs at top universities, as well as the needs of potential employers of our students trained by our new program. I expect your committee to complete your work and present me with your initial recommendation by the end of the Winter Term (2018), and a final proposal by the end of Spring Term (2018). This recommendation should comprise both preferred and alternative paths and should address the following (and other) big questions.

- How will a data science program leverage existing UO strengths to give us a competitive advantage in this already crowded field? What attributes will make the program a desirable home for recruiting top researchers and an appealing target for philanthropy?
- The data science program is likely to be a nucleating hub with spokes of support for efforts growing organically across the campus already. What is the optimal degree of centrality vs. dispersion of such a structure to provide this support?
- How can this program serve as a nucleating factor for incentivizing new coordinated hires across departments and collaborative appointments with KCASI?
- What organizational structures and space resources would best provide an intellectual and physical home for interdisciplinary data science research and training activities?
- What are the key computational tools and technical resources needed for a successful data science program? What role might cloud computing play in the future?
- How will the program intersect with other schools in Oregon (particularly OHSU and OSU), the west coast and the rest of the country?
- What will the program's relationship be to entities outside the academy, such as government and industry partners like Intel and Microsoft?
- How can the University work to coordinate curriculum of advanced courses in data sciences across departments and other units?
- Should the program create degree granting programs at the graduate and undergraduate levels, and if so how would these capitalize on key strengths across our campus?
- What is the optimal timeline and strategy for standing up such a program?

Activities that led to the proposal

In late 2017 the University launched the Presidential Initiative in Data Science with the goal of supporting and enhancing already existing efforts in data science research and education, and also provide a nucleus to bring top notch students and researchers to campus working in this exciting new field. The Data Science Visioning Committee (Appendix 1) was charged with exploring the internal and external landscapes of data science activities, and to provide guidance on achieving this goal. We propose that the creation of an interdisciplinary program in data science would best serve to support and grow existing areas of interdisciplinary research and education across campus in data science.

II. History of data science activities at UO

Incubation of programs at other universities

The development of data science programs is a new phenomenon, occurring mostly within the past 5 years. However, new data science programs often build upon deeper histories in areas such as statistics, applied math or computer science. At other research universities data science programs have often been incubated within existing schools or colleges that have historically had strong technical focus or applications. For example, schools of agriculture have often produced statistics departments, and many of these have subsequently been extended into programs in data science. Similarly, schools of medicine have developed advanced programs with a strong emphasis on biomedicine, and colleges of engineering have developed activities utilizing data science in the physical sciences and computer science. During the development of many of these data science programs at other institutions, the programs have often taken on a more targeted mission that is aligned with the unit in which it was incubated. Even at universities where the ultimate goal of the data science efforts was to create a wide-ranging structure that bridges traditional boundaries, the incubation of data science within an existing school or college with a strong history has led to the programs having a significant flavor that depends on the incubating unit. As a consequence, data science programs are diverse despite encompassing some common goals and approaches

The challenges and opportunities of a lack of similar incubating structures at UO

The University of Oregon was founded in 1876 with a focus on 'foundational' disciplines and as a consequence has strong programs in such traditional areas of the social and natural sciences, humanities, and mathematics. In contrast, our university does not have schools of engineering or medicine, and biomedical and engineering research has occurred in existing units on campus. Most notably the University has a relatively small department of Computer and Information Sciences, and while UO has a strong department in pure mathematics, we have no department of statistics. The lack of such incubating structures has created several defining features of UO and has important consequences (both positive and negative) for the full development of a data science initiative.

The University of Oregon has a long history of strength in the social sciences and humanities, and has nurtured the success of the Clark Honors College as an area of excellence in undergraduate education. This has led to the University having a long and proud tradition of being a liberal arts research university. In addition, the lack of some schools and colleges, such as engineering and medicine, which often support advanced interdisciplinary research and, has meant that faculty at UO have relied on creating new structures within which to support their work. Dozens of highly interdisciplinary and successful research and educational institutes have been developed, such as the Prevention Science Institute, Institute of Molecular Biology, Material Science Institute, and Environmental Studies Program. These institutes draw faculty from many different departments and schools, and these have been engines of innovation that have brought acclaim to UO for decades.

The long history of very interdisciplinary research at a liberal research university, paired with the absence of traditional silos that may retard the full development of a data science initiative, provide an excellent opportunity for the University of Oregon to set itself apart and launch a truly integrative data science program. Similar to the development of the Knight Campus for Accelerating Scientific Impact (KCASI), the can-do and nimble approach of the UO faculty has led to a willingness and desire to create truly integrative solutions. The counterpoint to such a future unfettered by the past is that it is also unsupported by existing investments in key programs and infrastructure. We have no department of statistics, so in many ways we will need to build this expertise for the data science initiative from scratch as compared to expanding it as would occur at other universities. Similarly, growth in areas of advanced computer science and analytics will need to occur, in many cases from the ground up. Lastly, although investments have been made at UO in supporting resources such as research and educational computing, we do not have as wide and strong of a base as exists at other research universities. We will need to continue to make investments. Despite these challenges, we do have a strong base on which to build, the disparate efforts that have been occurring across the domain aspects of data science.

Diverse data science efforts across campus in several domains

Because of the lack of existing domain infrastructure for incubating efforts in data science, faculty at the University of Oregon have been building out expertise and research collaborations primarily within separate domains across campus (see **Appendix 3**). Some of these efforts have been in areas which might be considered to be more traditionally fertile ground, such as hires in Computer and Information Sciences (CIS) in artificial intelligence and machine learning, statisticians in the Lundquist College of Business (LCB) focused on business analytics, and physicists working on large collaborative projects at CERN and LIGO. Some efforts have crossed traditional boundaries within schools and colleges, such as data science for social equity hires within the College of Arts and Sciences (CAS) that bridges geography, sociology and economics, or joint hires in math-biology-computer science that capitalize on the growing use of quantitative and computational approaches in the life sciences.

Another theme that emerged was the use of large amounts of distributed big environmental data from remote sensing that brought researchers together across physical and life sciences, as well as areas of social science and economics. Researchers in the School of Journalism and Communication (SOJC) and the College of Education (COE) have launched data science efforts from social interaction networks and communication (SOJC) to the causes of variation in health outcomes informed by family and community data (COE). As an indication of the truly broad possibilities for these efforts, the obesity prevention health promotion cluster bridges COE and CAS, recently adding a world-renowned geneticist with expertise in using machine learning to study human genetic data. Finally, significant efforts are underway in digital humanities and data ethics, as well as data science communication.

The lack of silos at UO has created fertile ground for interesting and exciting collaborative research in data science. The faculty and students who have been investing their time are the human capital upon which the initiative can be built. A challenge has been the lack of a defining overall vision for the ecosystem of data science at UO that can embrace and enhance these different efforts in domains by providing the resources and intellectual home to help them flourish. An additional challenge has been the historical lack of homes for more methodological or technical components of the ecosystem, such as robust statistics and artificial intelligence, as well as methodological expertise in communication, ethics and law as it pertains to data science. Because of the hard work of existing faculty and their students we now have a framework by which can strengthen these domain 'spokes' and increase their connectivity to other academic and societal institutions, and significantly add to the methodological capacity at the University of Oregon as well. We lay out the vision of the committee for fully building this ecosystem below.

III. The vision for building the data science initiative

Below we outline guidance for implementing the data science initiative at the University of Oregon. We start with goals and general principles and then continue to a proposed vision for the data science ecosystem at UO. We discuss how the data science initiative could expand research and education, increase connections to other university and societal partners, and increase the diversity of our UO community at all levels – faculty, staff and students. We highlight likely challenges and present possible ways to overcome these hurdles, including through strategic investments.

Overall principles

The committee agreed upon several guiding principles for the implementation of the initiative. These principles were informed by advancements in the field of data science in general, development of data science programs at other academic institutions, and needs for data science in society. The principles also build upon UO's history of research and educational excellence and existing efforts in data science, while also looking to areas of likely growth and impact. The data science initiative should…

- *Be broad and well connected across the university*. Because data science impacts nearly all areas of the university and society, the initiative should encompass the diversity of different units on campus that are willing to make substantive contributions. The initiative should have focal areas in each school and college on campus, but it should not be sequestered in any one existing school or college. A successful data science initiative will form a network of collaboration.
- *Include both research and education*. Adding novel areas of scholarship and discovery in data science will be a crucial component of a successful initiative, but so will training the next generation of data scientists. While some aspects of education in data science already exist at UO, more can be done or core methodologies and domain applications of data science.
- *Build upon interdisciplinary research at UO*. A long history of highly successful interdisciplinary research exists across campus supported in a variety of ways across campus. Schools and colleges, departments and institutes, and centers and programs have all brought together scholars from diverse backgrounds. Interdisciplinary research has been an engine for increasing collaborative research, support, and contributions that have significantly impacted the world. Incorporating this interdisciplinary ethos into the data science initiative will serve it well.
- *Capitalize on the identity and strength of UO being a liberal arts research university.* The social sciences and humanities have a long history of scholarship and educational excellence at UO. A modern data science initiative will also be built upon this strong liberal arts foundation. Doing so will allow data science to become an additional component of a liberal arts education, the historical goal of which is to provide students with a broad education to become successful in the world. It would also encourage the full inclusion of liberal arts knowledge and scholarship into the data science revolution in ways that would be quintessentially University of Oregon.
- *Leapfrog historical structural weaknesses.* A negative consequence of the lack of schools of medicine, engineering and agriculture at UO has meant the university has concentrated less effort on building in areas such as applied math, statistics and computer science. While these areas exist at UO, they are smaller in comparison to our peer institutions. One reason has been the appropriate focus on strength in 'pure' areas in the departments of mathematics and computer and information sciences. As a consequence, many of quantitative application areas have grown within domain foci. We can, however, fill these gaps by recruiting faculty with expertise in the front edges of these always-changing fields of application research.

- *Advance new research and educational opportunities*. Several areas of scholarship and research in data science are underrepresented at UO, and a successful initiative will increase these areas. Similarly, there are segments of the undergraduate and graduate student population that are not being fully served by present course offerings. Data science is particularly well suited to new educational approaches, and a successful data science initiative could provide a framework for such educational innovation. The initiative will have core methodological components that lend themselves to common pedagogical approaches and will have connections to numerous different domain areas which provides a diversity of applications of these common approaches.
- *Help build bridges to new academic and societal partners*. A successful data science initiative should act as a portal for increased impact of UO on society, and a conduit for new research and educational opportunities to flow back into the university. The initiative should focus on building collaborative efforts with researchers at other academic institutions, particularly those within the state (e.g. OSU and OHSU) as well as along the West Coast. In addition, the initiative should increase the impact of UO scholarship and education on society by forming key relationships with key economic and governmental partners, as well as other segments of society.

A vision for the overall data science ecosystem at UO

The principles lead to proposed structures to support the data science ecosystem at UO. We use the term ecosystem intentionally, because the vision of a successful data science initiative comprises numerous interacting and interconnected components both within and outside our university. The committee has developed a vision that we have come to describe as a 'hub-and-spoke' model for the overall ecosystem (**Figure 2**), which attempts to capture these connections. The specific designations for each of the components are not meant to be restrictive or exclusive, but simply meant to provide real context for the overall conceptualization of the data science initiative. The goal of the hub-and-spoke ecosystem model is to support and promote connectivity along existing areas of scholarship and education, but also to support and guide the organic growth initiative over.

- The overall ecosystem. Because of the overall breadth of data science across the university, and its connection with partners in the wider world, we propose that the ecosystem vision comprise three major categories. Moving from the inside out are the following categories (**Figure 2**).

- Methodological hub. This is the area which contains much of the research and educational expertise for implementation of data science. Many of the technical aspects of data science, such as data collection, database creation and curation, as well as statistical and computational analysis, would be considered components of the hub. Because of the need to effectively present and communicate data, as well as make ethical and legal decisions regarding the appropriate use of these data, several aspects of design, communication, ethics and law could also be components of the core. A key aspect of the contributors affiliated with the hub is that they can pivot to work with many different spokes through research or education. For example, faculty who focus on machine learning or data science ethics can collaborate with many different domain data scientists and teach core competencies to students drawn from the different domains. Some examples of areas of inquiry within the hub are the following:
    - Applied statistics
    - Data management and cyber security
    - Machine learning and artificial intelligence
    - Data science interpretation, visualization and communication
    - Data ethics, law, and policy

- Domain spokes. These are areas where the research and educational activities are focused primarily on a particular field or problem (e.g. biomedical data science or business analytics). Faculty in these spokes may gather large amounts of data and use advanced analytic tools that are common to those in the hub, but their primary research and educational focus will always be first on their domain area of expertise. They can therefore collaborate with individuals in the hub and have connections and impact outside of UO but will primarily do so through their fields. Importantly, similar to how research institutes at UO cross traditional departmental and school boundaries, so too can the spokes, which may bring together researchers even from different colleges. Below are four examples of spokes that describe related research activities that are occurring at UO and which are natural avenues for focus and expansion.

    - **Business analytics** – collecting, curating and analyzing large amounts of market and supply-line focused data for competitive advantage in businesses.
    - **Environmental big data analytics** – harnessing immense amounts distributed remote sensing data, dense time series data, and novel computational models to make more rapid and appropriate decisions about environmental threats.
    - **Data science of social interactions and societal impacts** – using information on social interactions and factors that affect child, family and social well-being, and harnessing that information to make preventive interventions and change social policy.
    - **Biomedical data science**. The analysis of large amounts of diverse genetic, cellular, physiological and organismal data to improve human health.

- Ring of connectivity. While the hub and spokes are largely (but not exclusively) located within the university, the goal is to have greater impacts on the world by increasing connectivity with other academic institutions, government agencies, and industry partners. The overall structure of the data science ecosystem allows for the increased research and educational connectivity both through the hub as well as around the ring, allowing feedback to accelerate novel research and education within our university informed strongly by the wider world.

    - **Academic partners** in the state, along the West Coast and the Pacific Rim.
    - **Government agencies** at the local, state, regional national levels.
    - **Industry partners** across the spectrum of rapidly growing sectors.
    - **Societal impacts** on families, schools and communities.

Vision for the initiative to be an engine for expanded research

In order for the data science initiative to fulfill the goal of supporting advanced interdisciplinary research, it will require building the components of other successful research departments and institutes at UO. These include the development of intellectual support networks such as shared physical spaces and computational resources, as well as activities such as seminar series and journal clubs. In addition, administrative support for developing, submitting and managing grant proposals will be required. For the initiative to be maximally successful it will also develop mechanisms for incubating novel research ideas and directions.

The most important component of a successful research institution or department involves the involvement of faculty, students, and staff. In order for the initiative to succeed as an initiative, it should develop various degrees of relationship with the program and commensurate expectations.

For tenure track faculty (TTF), these affiliations could be the following:

- Affiliation of existing or new faculty who have their complete tenure home in another unit.
- Partial tenure in the data science initiative with a partner unit somewhere else in the university.
- Full tenure home in the data science initiative.

For non-tenure track faculty (NTTF), these affiliations could be the following:

- Affiliation for NTTF with appointments in other units in a manner that is similar to the affiliations of TTF (see above).
- Appointments of research teams (e.g. research assistants or postdoctoral scholars) working with a senior TTF or NTTF data science researcher.
- As the appointment home for more senior research scholars such as Research Associates of Research Professors.

An important goal of the first year of the implementation of the initiative will be to determine the roles and responsibilities of each level of affiliation for TTF and NTTF, and to develop these into clear policies that align with University of Oregon policies.

Growing educational opportunity through implementation of the initiative

The data science initiative has the opportunity to increase both the quality of education at the University of Oregon, and number of students being served, both at the undergraduate and graduate levels. For example, one of the largest undergraduate classes at UC Berkeley is their introduction to data science class with over 1000 students taking the course annually. In addition, several data science programs at other institutions have developed robust online opportunities and 'flipped' classrooms, where the overall content is delivered electronically with much more detailed, hands-on exercises where students interact directly with faculty. The initiative should facilitate novel educational directions with partnerships with existing colleges and departments through the development of new degree programs and by innovating novel approaches to teaching. A successful data science initiative will develop and implement educational programs in each of these areas. The following were important points of discussion by the committee in making these recommendations:

- Training for some students may be focused solely on the methodological aspects of data science, but in many other cases the data science training will be to advance expertise in a specific domain or discipline. The training opportunities provided by the data science initiative should cover both and therefore cover a range of pedagogical approaches.
- The committee felt that several educational models should be explored deeply during the initial years of implementation of the data science initiative. Because data science covers so many academic areas, some combination of focused course delivery within the initiative, and collaborative course offerings with partner departments, is likely to be optimal.
- Lastly, the educational solutions may differ between the undergraduate and graduate levels. From the standpoint of data science training for domain purposes (e.g. biomedical data science), the overall training of a student may involve deep understanding within a domain as well as core data science methodologies. Whereas a graduate student may come join the initiative to add data science expertise on to an undergraduate degree in a domain area, undergraduates at UO may benefit from concurrent education in a domain area with a concentration in data science.

Undergraduate Education

The committee agreed that an important goal of the data science initiative would be to bring new communities of undergraduate students to UO who might presently be drawn elsewhere. We felt that this could be achieved through developing a combination of new core course offerings in data science methodologies and increased course offerings of data science applied to a range of domain areas. If successful, these novel educational opportunities will increase the size of the undergraduate body at UO, as it has other institutions. The committee discussed two models of undergraduate education, but several more should examined during the implementation of the initiative.

'***Data Science + Domains****' model for undergraduate education.* In this model the data science initiative would develop a wide-ranging set of courses that covers the ecosystem of data science and offer degree programs in these areas. These course offerings and degrees would include the core methodologies as well as degrees for which data science is focused on a domain area. This model would resemble course offerings within traditional departments, making the data science initiative at the undergraduate level akin to a new school or college (albeit small).

Benefits of such a model revolve around its familiarity. The inside out model makes it easier to create a unitary identity for the initiative that is separate from other aspects of UO, and models for administration and tuition return can simply be adopted from other parts of campus. A significant disadvantage of the inside out model is that without very conscience bridge building to other parts of campus, it may create barriers to participation from the wide diversity of individuals working on data science in domain areas across campus already.

'***Domains + Data Science****' for undergraduate education.* In this model the data science initiative would develop course offerings and degree programs affiliated with some aspects of core methodology as in the previous model. However, more educational development of data science applied to domain areas would occur through collaborative efforts with existing schools and colleges. For example, the Department of Biology currently has several areas of emphasis for its undergraduate degree (e.g. molecular biology, neuroscience, etc.). Instead of developing its own degrees in data science applied to the life sciences, the data science initiative could work with the Biology Department to develop a new area of emphasis such as Biological Data Science emphasis for the Biology degree.

This model has several benefits. The initiative would be able to more quickly increase the breadth of educational offerings at UO by working within the established academic and administrative structures as compared to needing to obtain approval at university and state levels. Another benefit accrues because departments can have new course offerings developed by initiative associated faculty as compared to hiring faculty with this expertise themselves. This model provides a mechanism by which the data science initiative remains closely aligned with – and an engine for – integrative data science activities in domains across campus. Undergraduate education can therefore serve to advance innovation by incentivizing faculty from different disciplines interact. If done well, this model could serve as an 'intellectual coffee kiosk' to bring together faculty from different disciplines.

Several challenges exist. Because of the inherent differences in pedagogical approaches across disciplines, the nature and quantity of workload would need to be clearly defined *a priori*. An additional challenge will be to determine the appropriate amount of return of tuition dollars for the initiative itself. Simply coordinating across multiple campus units in different colleges could be a challenge, which would require strong senior leadership and support. If successful, the initiative will help increase the overall undergraduate student body at the university, and therefore cover a proportion of the costs of the faculty lines within the initiative. However, the administrative costs would largely exist within existing departments. The overall business model should be developed to accommodate these realities.

Graduate Education

The committee felt that graduate education at UO could be significantly enhanced through the implementation of the data science initiative. At the **Ph.D. level**, faculty associated with the initiative will almost certainly want to have graduate students participating in their research and scholarship, and as with other disciplines will therefore develop specific educational tracts to train the next generation of scholars. Similar to the undergraduate level, these graduate tracts may comprise a combination of data science methodology specific programs within the initiative (e.g. machine learning for data science), as well as graduate degrees within domain areas (e.g. business analytics). Similar to the arguments laid out above for undergraduate degrees, the domain focused graduate degrees could follow either an inside out or outside in model. Many of the same benefits and deficits arise. The committee felt that evaluating these (and other) models of graduate education should be a focus of the first two years of implementation of the data science initiative.

The committee felt that there were significant potential benefits for expanding offerings in data science at the **Masters level**. Hiring of data scientists has accelerated rapidly in nearly all areas of industry and society, and training in core competencies of data science that build upon a deeper domain expertise is currently desirable. Many students who finish an undergraduate degree in a specific domain would likely be attracted to a focused Masters degree in data science to launch their careers. These Masters degrees could be offered in collaboration with explicit internship programs, an example of which is the Bioinformatics and Genomics Masters Program (BGMP) that is part of the Knight Campus Internship Program (KCIP). The BGMP provides 1.5 years of training, including 9 months of focused data science training in genomics and bioinformatics at UO followed by a 9 month internship with academic or industrial partners. A similar model could be applied in many domain areas.

IV. Key steps and components for launch

Faculty and Institutional Leadership

Successfully launching an initiative that covers a width swath of the university will be challenging. Increasing the complexity is that existing faculty will initially be added through a process of affiliation before hiring new faculty. Like all units, these faculty should provide self-governance within the context of the institutional structures that allow decisions to be optimized at the level of the data science initiative as well as the university. In addition, having broad participation, input and advising from across campus will increase the likelihood of success.

The committee felt that the question of which individuals and units will be involved with the data science initiative should be determined primarily by a demonstrable commitment of resources by those individuals and units such as time, faculty lines, and philanthropic resources to the growth of the initiative. As a consequence, it will be important to develop a robust leadership structure that guides the develop of the initiative and facilitates these commitments. We suggest a structure (**Figure 3**) that comprises **five** primary components: Data Science Faculty, Faculty Leadership, Director, Steering Committee and the Provost. We describe the goals of each component, with example composition, but refrain from suggestions of the precise representation or individuals.

- Data Science Faculty. These comprise the complete faculty affiliated with the Data Science Initiative. We propose two categories of faculty, those who are core and those who are affiliated. Core tenure track faculty (TTF) may have full or partial tenure homes in the new data science initiative, whereas affiliated faculty will have associations with the initiative similar to the relationships that have TTF with tenure homes in departments have affiliations with research centers or institutes. As the initiative grows, non-tenure track research and instructional faculty (NTTF) are likely to be added as well, and they will become core faculty.

- Faculty Leadership. At maturity, the data science initiative will comprise numerous core TTF and NTTF who will provide the leadership in ways similar to what occurs in departments and institutes across UO. During the startup phase of the initiative, however, all of the faculty will be affiliates, and a faculty leadership committee will be needed to help the Director implement the growth of the program. These leaders should be representatives of the different domain and methodology areas of data science drawn from both spokes and hub. They should be chosen to also focus on creating new avenues for interdisciplinary research and education. It is likely that these will be senior TTF with strong institutional perspectives.
- Director. The Director of the data science program will act as the primary academic leader of the unit. As the data science initiative begins, the Director will have responsibilities that would be similar to the Head of a department or the Director of an institute. As the data science initiative grows, the roles and responsibilities of the Director will as well to encompass academic and leadership roles that are similar to an associate dean or dean. The Director will be responsible to the faculty, be guided by the Steering Committee, and report to the Provost.
- Steering Committee. Because of the goal of creating a data science initiative that is broad enough to bridge traditional college and school boundaries within the university, we propose the creation of a Steering Committee that represents academic leadership drawn from across campus, likely a mix of individuals at the level Dean, Vice President, Head of unit, and senior faculty in other units. Key leadership from the Faculty Senate would also be desirable. The goal of this Steering Committee is to support the Director and advise the Provost on institutional perspectives for growth of the data science initiative. In addition, the Steering Committee will be positioned to be able to provide guidance on realignment of resources such as lines and space that will help the data science initiative advance the mission of the university. In addition, because of the goal of having the data science initiative increase connectivity with academic and societal partners, it may be beneficial to have outside members drawn from those sectors.
- Provost. The Provost will receive advice and input from all other levels and will make final decisions on major changes in direction of the initiative.

Develop processes for initial growth of the initiative

The data science initiative will need to grow organically over the next decade (see timeline below). To do so will require a well-planned series of phases, as well as a morphing of roles and responsibilities as the initiative grows. This incubation phase will likely be the most challenging and cover the first 2 to 3 years of the implementation of the data science initiative and will require startup support from UO. Below we provide guidance for avenues of growth.

- Growth of Faculty. As mentioned previously, the data science initiative will ultimately comprise a set of standing core faculty (both TTF and NTTF). However, the initiative will initially need to grow through the work of existing UO faculty who will be affiliates of the data science initiative. This initial data science faculty will therefore act in a manner that is most similar to members of a research institute and will therefore set the overall intellectual direction of the data science initiative. Concurrently, the faculty leadership and Director will need to develop policies and best practices for the growth of the faculty through new hires or possible movement of FTE from other units on campus. The faculty leadership will also serve as the coordinators for hires within domain areas that are affiliated with the data science initiative and will ultimately provide guidance and leadership to develop the policies for tenure and promotion.

- Administrative Support Staff. The fully-formed data science initiative will have affiliated support and technical staff that will be supported through mechanisms of indirect grant returns and tuition (general fund) allocations as the fully-formed faculty are writing and receiving awards and teaching new students. During the initial growth phase, however, it is unlikely that there will be support or need for several FTE worth of staff. Because different expertise will be needed (e.g. research vs. educational support), it is likely that sharing FTE with an existing unit on campus would be most efficacious. For example, the data science initiative could be incubated within an existing school or college using a shared services model.

- Physical space. Similar to the need for administrative support staff, the data science initiative will eventually require space allocation for new TTF and NTTF for offices and teaching spaces. Initially, the affiliates and core faculty hired as split FTE may be housed within existing units on campus. However, for the data science initiative to successful launch, it will still require a physical presence to serve as an intellectual nexus. This space could be on loan within an existing school or college for the first three years of the program, with the initiative leadership helping implement a vision for creating new space. This will most likely require significant state and philanthropic investments. Acquiring these funds should be a focus of the leadership of the data science initiative over the initial incubation phase.

- Growth of computational resources. Many of the new faculty and students recruited to UO for the data science initiative will require the use of advanced computing resources, as well as connectivity with collaborators and resources around the world. Investments in high performance computing, research computing, and IT infrastructure over the past three years have right-sized these resources for the present and immediate future needs of our faculty and students (see Appendix 4). However, to support the growth of the data science initiative, as well as the Knight Campus, to maturity will require additional investments in computational and IT resources, most likely around year 5 (see timeline).

Establish unique UO identity

A concerted effort will be required over the first two years of implementation of the initiative to create and promote an identity that is deeply seated within the University of Oregon. Doing so will require the development of physical spaces that will serve as an intellectual home for faculty and students, and a touchstone for academic and industry partners. This physical space will serve as the bricks and mortar 'storefront' of the initiative and will be home to a small number of interactive spaces, flexible teaching areas, and offices for faculty and staff. Commensurately, the initiative will need to greatly expand its web presence for both internal and external communication. Finally, a concrete plan for communication and advertising of the program should be developed and implemented that targets potential new faculty and students, as well as academic and societal partners. These should be fully in place by the end of the second year and sustained thereafter.

V. Challenges and investments needed to fulfill the vision

Challenges

Incubating the initiative with a clear path

- Lack of existing structures will be both liberating and limiting
- Will need to create and incubate initial structure
- Will simultaneously recruit existing faculty and hire new faculty

Appropriate balance between focus and inclusion

- We need to harness all efforts across campus to compete with other programs
- However, we don't want to become too broad or diffuse
- Leadership will need to provide constant attention
- Cost of inclusion is effort - with feet, dollars and time invested

Building educational programs

- Vision for graduate and undergraduate programs
- Work with faculty leadership and administration
- HECC and other approvals
- Recruit top, diverse students

Building bridges with academic and industry partners

- We will be OK by ourselves, but much more successful as part of a network
- Connections with industry (government/society) will need to be built
- Significant investments during startup will need to be placed on building those relationships

Investments

Attracting and retaining top faculty

- Top notch faculty are expensive
- Will need to have an appropriate compensation structure
- Chairs and salary augmentation
- Strong communications and advertising plan

Attracting and retaining top students

- We do not yet have a brand or image
- We will need to create a 'brand'
- Fellowships and other student support
- Strong communications and advertising plan

Infrastructure Investments

- Computational resources (hardware, software and people)
- Research and educational computing infrastructure
- Physical space for
  - Intellectual engagement
  - New faculty
  - Student home
  - Presence for our partnerships

Advertising and communication

- We are new to this game, and we don't have 'halo' from other units (e.g. stats)
- Need to establish our brand to be able to stand out
- Communicate out to faculty and students
- Communicate out to academic, industry and societal partners

VI. Timeline

Year 1 (AY2018/2019) – Implementation

- Establish leadership team and support staff
- Build institute-like structures to support research proposals and intellectual growth
- Attract existing UO faculty as affiliates
- Coordinate searches with faculty leadership
- Target of opportunity hires
- Begin building degree granting programs
- Establish programs to attract top students and faculty

Year 2 (AY2019/2020) – Launch

- Launch first degree programs
- Begin hiring new core faculty associated with initiative
- Additional world class target of opportunity hires
- Continue building regional and national presence
- Obtain major 'center-like' funding from federal or private agency
- Launch philanthropic efforts that are appropriately scaled and sized to need

Year 3 to 5 – Maturation

- Develop full research vision of program/institute
- Graduate and undergraduate educational programs running
- New online opportunities explored and developed
- Certificate programs explored and developed
- Evaluation and recalibration of research and educational approaches
- Continue to expand major 'center-like' research efforts

Year 5 to 10 – Excellence

- National prominence of initiative in quintessentially UO way
- Significant increase quality and quantity of student populations
- Full partnerships with other R1 schools
- Deep relationships outside of the academy

FIGURES

Figure 1. Conceptual overview of the field of data science in general.

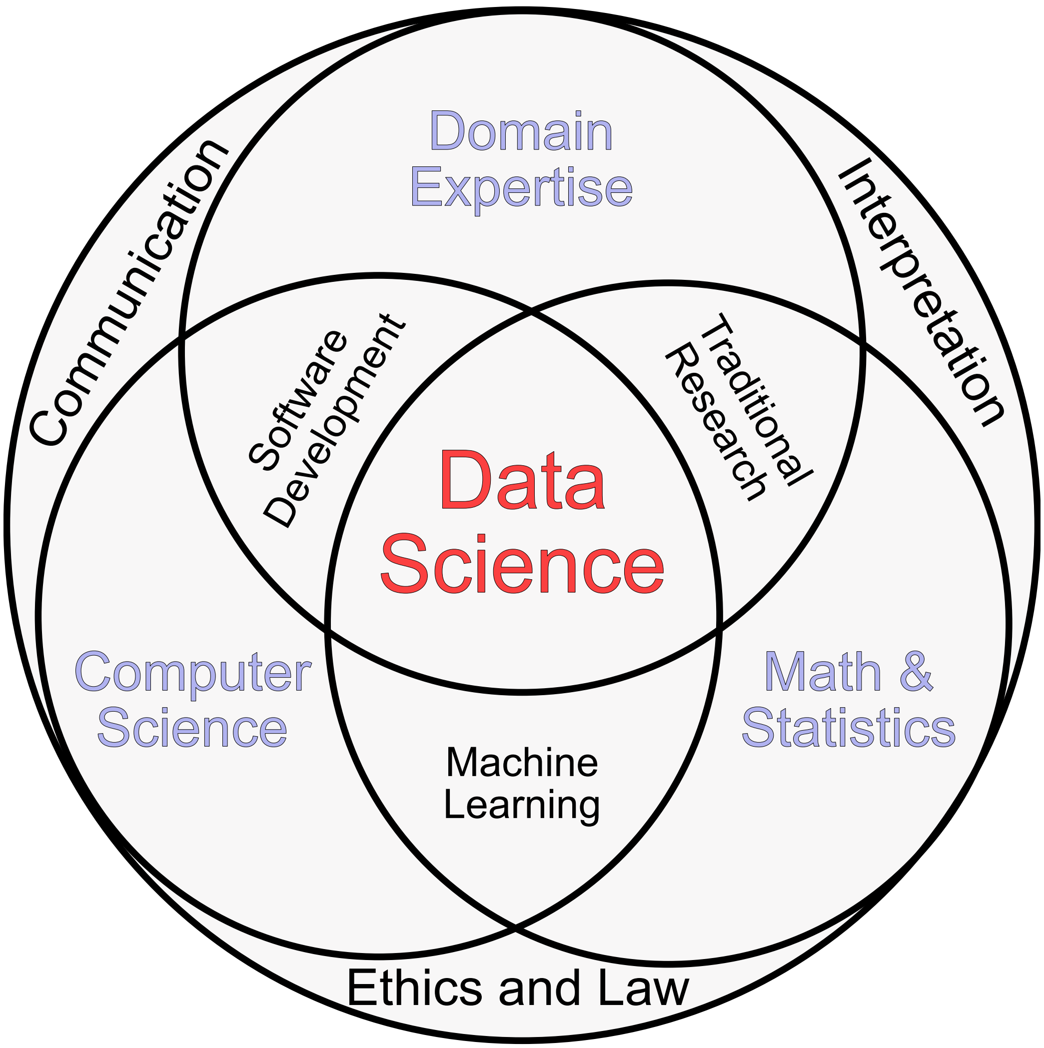


Figure 2. Conceptual diagram of hub and spoke model with example domain spokes.

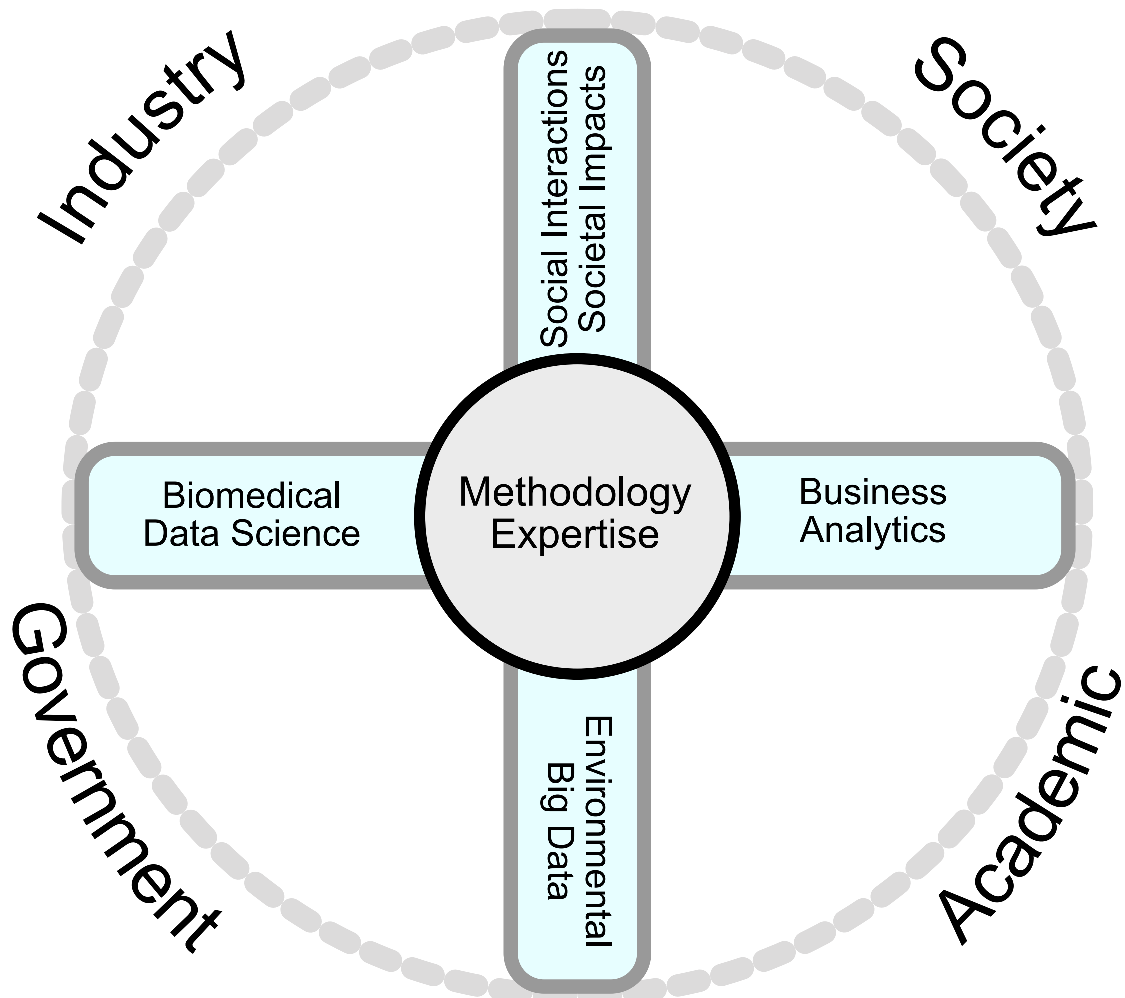
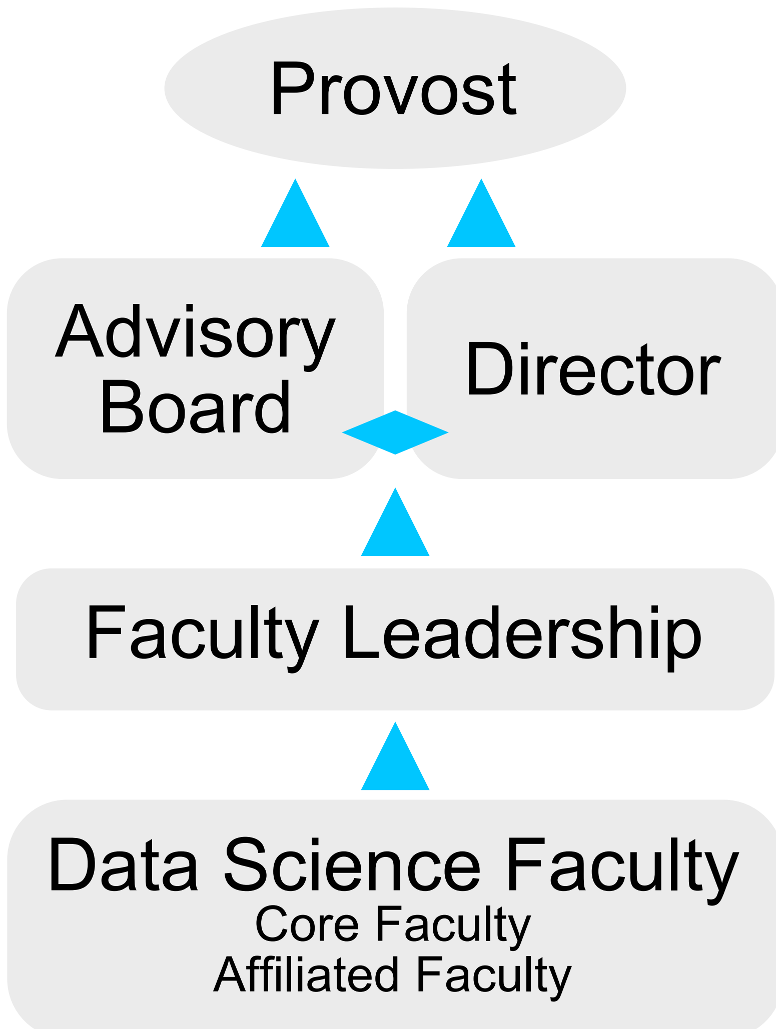


Figure 3. Initial leadership model figure for the data science initiative at UO.



APPENDICES

Appendix 1. Committee Membership

Name		**Affiliation**
Cresko	William	AVPR and Professor of Biology, CAS
Del Guercio	Diane	Senior Associate Dean, Professor of Finace, LCB
Huter	Steve	Director of the Network Startup Resource Center (NSRC)
Ives	Colin	Art and Technology Program Director, Associate Professor Digital Arts, COD
Piger	Jeremy	Professor of Economics, CAS
Koopman	Colin	Associate Professor of Philosophy, CAS
Leve	Leslie	AVPR and Professor of Counseling Psychology and Human Services, COE
Lewis	Seth	Shirly Pape Chair in Emerging Media, SOJC
Lobben	Amy	Professor of Geography, CAS
Maggio	Nick	Director of Research Advanced Computing Services (RACS), OVPRI
Minton	Jessie	Chief Information Officer, IS
Moseley	Cass	AVPR, Research Professor and Director, Institute for a Sustainable Environment, COD
Nutter	Sarah	Edward Maletis Dean, LCB
Pangburn	Michael	Head, Department of Operations and Business Analytics, LCB
Sadofsky	Hal	Divisional Dean of the Natural Sciences, CAS
Scalise	Kathleen	Associate Professor of Educational Methodoogy, Policy and Leadership, COE
Shelton	Brad	Executive Vice Provost, Academic Operations , OPAA
Coonrod	Leslie	Knight Campus Internship Program (committee analyst)
Wagner	Stacey	Knight Campus Internship Program (committee analyst)

Appendix 2. Relevant program comparisons at other institutions

Benchmarking Existing Data Science Initiatives
Executive Summary

1. Methodology

Data science related programs at external Universities were included based on the following criteria:

- Top ranked Data Science MS programs (Forbes)
- Programs of interest due to location and connections (e.g. OHSU)
- Key words: Data Science
- Top ranked Statistics Departments (US News and World Report)

2. General Findings

a. Degrees

While PhDs and MS programs are common for traditional Statistics programs, Data Science programs are commonly at the MS level and more focused on career opportunities than research.

A popular model includes

- Core competencies
- Capstone project in domain to apply competencies

b. Structures

Three general structures for the organization of the faculty

- **Hub model** – TTF and associates are affiliated from multiple domains/departments to accomplish several goals including but not limited to research collaboration, code development, company/industry engagement and collaboration, and education. Centers typically employ core staff (a director, scientists without additional departmental affiliations etc).
- **Decentralized model** – Affiliated faculty from different departments are part of the unit to achieve common goal, typically educating graduate students.
- **Department** – traditional academic department model
  - **Mixed model** – hub and department – Some faculty are affiliates while other's home department is the hub. Their focus is typically not restricted to a single domain.

c. Domains

Domains and affiliated departments vary.

- Domain
  - Humanities
  - Social Science
  - Arts
  - Law
  - Life Sciences
  - Physical sciences
  - Business
  - Engineering
- Methodological
  - Statistics and mathematics
  - Computer science

Example of successful program
eScience Institute at University of Washington (Seattle)
http://escience.washington.edu/

Education

1. Undergraduate specialization
    a. Required topics
        i. Data science framework, Implemented by all departments
            1. Programming
            2. Machine Learning
            3. Societal implications
            4. Data Management
            5. Data visualization and Communication
            6. Probability and stats
        ii. Domain requirements
2. Graduate Programs
    a. Integrative Graduate Education and Research Traineeship (IGERT) Data Science PhD Program
        i. IGERT requirements maps to departments, some go beyond
        ii. Astonomy, Biology, ChemE, CS&E, Genome Sciences, Oceanography
    b. Advanced Data Science Option
        i. Overlay of department requirements
            1. Data management
            2. Machine Learning
            3. Stats
            4. Big Data Seminar
        ii. Departments: Applied Math, Astronomy, Biology, ChemE, CS&E, Genome Sciences, Math, Oceanography, Psychology, Stats
    c. Graduate Data Science Option
        i. Appears to be new
        ii. Courses in following areas
            1. Software Development
            2. Stats and Machine learning
            3. Data Management and visualization
            4. eScience Community Seminar
        iii. Departments: Astonomy (PhD), ChemE (MS & PhD), Phsycology (PhD)
3. Certificates
    a. Part of greater UW system, subset are data science related topics
    b. Format
        i. Online, self or group paced
        ii. Classroom, noncredit
    c. 4-8 months
    d. From project management to machine learning
    e. ~3-5k

Personnel

Organized into working groups

1. Career Paths and Alternative Metrics
   Non-traditional fellows and faculty lines
   - Institutional policy
   - Foster community
   - Mentorship and job satisfaction
   - How are they funded?
2. Education and Training
   - Learning free from departmental politics
   - To implement innovative, multidisciplinary training
   - Boot camps, summer schools, tutorials
3. Software Tools Environments and Support
   - Remove obstacles to creating reusable software
4. Reproducibility and Open Science
   - Develop tools and techniques
5. Data Science Studies
   - Science and technology subfield; study of data science
6. Working Spaces and Culture
   - Design spaces that promote collaboration
7. Neuroinformatics

Appendix 3. Brief descriptions of *some* existing efforts in data science at UO

Center for Cyber Security and Privacy

The Center for Cyber Security and Privacy (CCSP) was established as the university's primary hub for cyber security research and education, with research focus on DDoS, DDoS defense, internet routing and infrastructure security, online social networking security, internet of things security and privacy, and privacy in the digital age. CCSP members span departments including Computer and Information Science, Philosophy, Law, Business and Information Services.

College of Education

The College of Education (COE) currently offers a 5-course data science specialization for graduate students and is in the process of developing a MS degree in the Department of Educational Methodology, Policy and Leadership (EMPL), and an Educational Data Science Ph.D. emphasis area in EMPL. Research initiatives include submitting several proposals to the NSF, including "Integrative Strategies for Understanding Neural and Cognitive Systems" program solicitation, and a proposal for a seed grant between UO and OHSU. Additionally, the COE is moving to hire an assistant professor in Education Data Science within EMPL.

Department of Computer and Information Sciences

The Department of Computer and Information Sciences (CIS) offers several courses for undergraduate science majors. Research activities include topics such as artificial intelligence, data mining, information extraction, big data visualization, machine learning, and others.

Data and Media Research Collaboratory

The Data and Media Research Collaboratory (DMRC) is a cross-disciplinary faculty research group studying how data is developed and used, the values and ethics of data & media, how media shape data, and how data and media contribute to the dramatically changing social and historical formations in which we find ourselves today. The DMRC organizes small research seminars and workshops as well as major public events, including the Living Data: Inhabiting New Media conference. Members include faculty from Art History, History, Philosophy, Comparative Literature, Media Studies, English, and Romance Languages.

Department of Economics

The undergraduate major in economics contains substantial required and elective coursework in theoretical and applied econometrics, and offerings in big-data and data-science specific courses are expanding. The Master's program is undergoing a redesign to incorporate more data-science content, and new Ph.D. level seminars have been introduced covering topics such as machine learning, web scraping, remote sensing, and Bayesian econometrics. The department has hired three new faculty with expertise in data-science methods, including one as part of the "Big Data and Social Equity" cluster hire.

Department of Geography

The Department of Geography currently offers a Spatial Data Science and Technology major for undergraduates. This 48-hour undergraduate major includes 4 compulsory courses and 8 elective courses.

Compulsory Courses:

- GEOG 181: Our Digital Earth
- GEOG 281: The World & Big Data

- CIS 122: Introduction to Programming
- GEOG 481: GIScience I

Elective Courses:

- GEOG 482: GIScience II
- GEOG 485: Remote Sensing I
- GEOG 486: Remote Sensing II
- GEOG 490: Special Topics in GIScience*
- GEOG 491: Advanced GIS & Python
- GEOG 493: Advanced Cartography
- GEOG 494: Spatial Analysis
- GEOG 495: Geographic Data Analysis
- GEOG 496: Location-Aware Systems
- GEOG 498: Geospatial Project Design
- CIS 210: Computer Science I
- CIS 211: Computer Science II
- GEOG XX: 300 or 400-level course not listed above**
- CIS XX: 400-level course not listed above**

Department of Sociology

The Department of Sociology has recently added coursework in Social Network Analysis, Advanced Statistics, and will offer a new course next year in Computational Social Science. Research efforts in the department focus on issues related to data science and computational social science, and a faculty member has recently received one of the inaugural Interdisciplinary Awards in the Humanities and Social Sciences from UO for using Twitter and GIS data to estimate environmental hazards. A cluster hire in data science is being proposed for a sociologist actively using “big data” and advanced quantitative methods to study social or environmental impacts.

Lundquist College of Business

The Lundquist College’s Department of Operations and Business Analytics (OBA) spans the areas of operations management and applied statistics. OBA offers multiple 400/500 level courses including Information Analysis for Managerial Decisions, Analyzing Big Data, Predictive Analytics, Business Analytics, Business Database Management Systems, and Sports Analytics. Research in the OBA domain focuses on the development of models and techniques for improving firms’ operational efficiency. Such efforts often requiring integrating analytic methodologies from industrial engineering, statistics, and economics. Additionally, OBA has hired four tenure track faculty with data-analytics focused research programs.

Prevention Science Institute

The Prevention Science Institute currently has multiple funding lines through the NIH and the Office of Juvenile Justice and Delinquency to fund research projects synthesizing big data using meta-analytic techniques. Projects include examining the effectiveness of brief substance use interventions and using predictive modeling to advance precisions medicine for substance use prevention, and examining the effectiveness of juvenile drug treatment courts and synthesizing big data with predictive modeling to identify personalized treatment recommendations. Other applied research methodologies include expertise in using integrative data science approaches to harmonize and synthesize big data across clinical trials, and expertise in analyzing social media and other social interaction data.

Appendix 4. Key infrastructure investments at UO that will support data science

Improvements to campus information technology (IT) infrastructure

- Redesigned campus network to provide 100G capable core with enhanced redundancy for high speed data transfer and resilient structure.
- Connected campus core network at 100G to Internet2 providing high speed connectivity and peering for researchers for the first time off campus.
- Collaborating with Oregon's research universities and state government to build statewide 100G research and education network.

Improvements to research computing services

- The University built a brand-new supercomputer, Talapas, that has capabilities to perform large amounts computations simultaneously on large, diverse data sets.
- Established Research Advanced Computing Services (RACS) and hired 5 highly skilled staff (3 PhDs) to support computational research.
- Hired an Associate Director of Artificial Intelligence to head a new data science/machine learning rapid response team.
- Launched UO's first high-speed data transfer service allowing researchers to move, and share, large data sets with colleagues around the world.
- RACS, CIS, and Pres. Initiative in Data Science sponsored a series of hosted workshops on advanced computational techniques in data science.