Discovery Challenge - Initiative in Environmental Data Science

Summary ESF has a long and successful history in the environmental science disciplines but has yet to embrace powerful data science tools and concepts. This initiative will create the *Center for Environmental Data Science* to leverage existing ESF strengths, to address institutional weaknesses, and to be responsive to emergent data science themes being pushed by NSF and NIH. The center will also foster synergies with Empire Innovation Program-supported new hires here at ESF and SUNY Upstate. The Center will coordinate new course offerings, seed grants, graduate student support, campus computing (along-side ITS), innovative faculty training, and new partnerships with other academic institutions, government, and industry.

Background In the last five years, business and academic institutions have heralded a new era of data science. A May 2018 Bloomberg article titled "This is America's Hottest Job" (Michael Sasso) noted data science job postings on Indeed.com had increased 75% since 2015 with numerous companies unable to fill their openings. Academic institutions both nearby (University of Rochester, SUNY Binghamton) and distant (Duke, Stanford) have recently launched data science programs.

Data science encompasses data analysis in fields as diverse as health care, finance, energy systems, genomics, and marketing as well as the environment. Underpinning all data science is expertise within computer science and applied mathematics. Thus, with limited reach into these other disciplines, ESF is at a disadvantage when competing in the broad realm of data science.

> "Data Science [is] the multi-disciplinary field that combines data analysis with data processing methods and domain expertise, transforming data into understandable and actionable knowledge relevant for informed decision making…"—Gibert et al. (2018, Envir. Model. and Software, 106: 4-12)

However, ESF has distinct strengths within the sub-field of *environmental* data science which relies more heavily on domain knowledge and analog data acquisition (Figure 1) compared to the broad field of data science whose data likely originate in the digital realm (e.g., financial transactions, Facebook likes). A push for environmental data science would leverage ESF's exceptional domain knowledge, research properties beyond the Syracuse campus, and environmentally-geared faculty and student body to build a competitive data science program. This initiative would build off of existing ESF programs in aquatic health, geospatial analysis, hydrologic modeling, wildlife tracking, and others as well as connect to new data science-focused faculty lines in environmental health at both SUNY ESF and Upstate.

Specific challenges to conducting environmental data science include issues with data interoperability, treatment of data errors, data fusion, data storage and reproducibility, and selection of proper data streams (Gibert et al., 2018). Overcoming these challenges would move the environmental data science field forward as well as develop new capacity at ESF (particularly in the areas of data acquisition, analytics, and data processing, Figure 1).

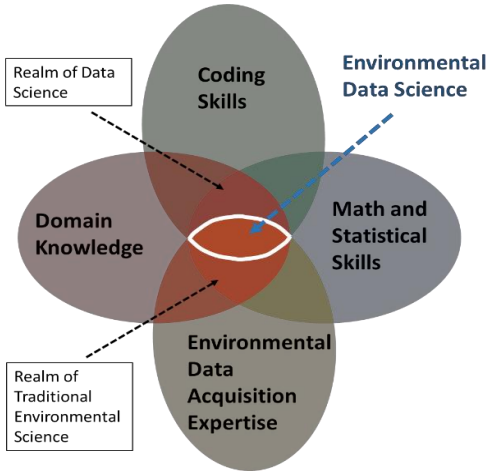

Figure 1. Environmental data science sits at the intersection of four disciplines. Competency in all four areas is critical for conducting impactful research at an institutional scale as well as for training individual students. While ESF has strong expertise in the more traditional areas of fundamental domain knowledge and math and statistics, this initiative would enhance ESF's capacities in data acquisition and coding/computer science.

Figure adapted from Conway 2013.

Goals The central goal of this initiative is to create a new Center for Environmental Data Science that broadly enhances environmental data science capacity at ESF in order to 1) be a leader in the field of environmental data science, 2) differentiate and enhance the academic programs ESF offers, 3) ensure our students remain highly sought after on the job market, 4) strengthen the research capabilities of ESF faculty 5) be a sought after partner to government, industry, and academic institutions, and 6) maximize the ability of ESF to solve real-world problems.

Initiative Components (Approximate fund allocation indicated by number of $)

1. Development of New Curriculum for Undergraduate, Graduate, and Professional Programs ($)

To differentiate ESF students from those at other institutions and to prepare them for in-demand jobs, environmental data science would become a core curricula element at all levels. We would develop an environmental data science undergraduate minor centered around an introductory level programming course and an intermediate level general data science class, including exposure to cutting edge analytic approaches (i.e., machine learning). At the graduate level, an interdisciplinary core course would be developed. Professional programs could draw from aspects of these new undergrad and graduate courses. Graduate students could also participate in an environmental data science "boot camp" at the highly-instrumented, long-term biomonitoring station in the Archer Creek watershed at the ESF campus in Newcomb. The "boot camp" would focus on hands-on deployment of sensors, maintenance of cyberinfrastructure, and data analytics.

2. Investment in High-Quality Graduate Students ($$$)

Central to this initiative would be training a cohort of graduate students in the elements of environmental data science. These graduate students could drive collaboration among faculty, test out new approaches to graduate education, bring existing skills of their own (namely in computing, data acquisition, and analytics), and help advance faculty research. Put simply, any new research initiative at ESF is highly dependent on having highly motivated and capable graduate students behind it. To attract high-quality students with the appropriate background and an interest in a PhD, funds would be used to enhance the standard TA stipends (e.g., a supplement of $8k/yr for 6 to 10 students).

3. Research Seed Grants ($$$) To initiate new environmental data science research and to encourage new collaboration with faculty inside and outside ESF, small research grants would be

offered with the intent of fostering a pathway to future extramural funding by supporting travel, materials and supplies, and summer salary. Competitive seed grants would require grant recipients to demonstrate new collaborative relationships as well as support of the broader initiative (e.g., by contributing materials to curriculum development, attending training sessions, etc.) and submission of proposals to extramural funding competitions.

4. Coordination of Improvements to Computing and Sensing Resources at ESF ($$)
High performance computing at ESF remains ad hoc and non-centralized, both in terms of hardware but also in terms of personnel. An initial review of computer resources to identify immediate areas for investment will be conducted. While investments in computing hardware may be needed, this review would also consider support staff who provide training and direct assistance in technical aspects of research, as well as the development of a formal policy for adapting and utilizing cloud computing. A central aspect of this initiative also addresses data collection. There are numerous, rapidly evolving means of collecting data: microbial community sequencing, UAV's, embedded network sensor systems, RF ID's, flux towers, and hand-held devices. Some funds would also be used to purchase such data collection hardware.

5. Coordination with Other Regional Data Science Programs ($)
Nearby institutions have recently made large investments in data science (U of R, Binghamton) or have advanced capabilities in computing (Cornell). In many cases, such institutions welcome collaboration and partnerships. Additionally, the initiative would support the emergent relationship being established between ESF and SUNY Upstate in environmental health.

6. Faculty Workshops and Training ($$)
Training workshops would be used to jump start faculty involvement and to rapidly provide faculty with new skill sets. Additionally, workshops could be used to gather structured guidance from outside experts in data science to develop a road map for advancing environmental data science at ESF. Entities such as Data Carpentry (https://datacarpentry.org/) are well suited to provide this type of professional training.

Strategy for Continued Growth In the first two years, the primary goal would be to receive funding for an NSF Research Experience for Undergraduates (REU) and an NSF Research Training (NRT) grant. These awards would provide a focal point for on-campus collaboration in environmental data science as well as provide an avenue to develop novel methods to train students and to provide indirect support for research. The NRT would be a particular emphasis since an NSF priority funding area for coming years is "Harnessing the Data Revolution," one of NSF's 10 Big Ideas for the next decade. NIH has also identified data science training as a major goal in their latest Strategic Plan for Data Science (Goal 4). We will also develop new relationships with state agencies (NYS DEC, NYS Canal Corp, NYSERDA, NYS DOH, etc.) to train staff and develop new approaches to environmental monitoring within the state. As capacity was built, growth would also come from sponsored research, awarded from competitive grants or from industry and governmental partners.

Core Team: Steve Shaw, Hyatt Green, Mary Collins, James Gibbs, Colin Beier, Lindi Quackenbush
Collaborating Staff and Faculty: Tim Morin, Chuck Kroll, Jim Sahm, Bahram Salehi, Brian Leydet

Initiative in Environmental Data Science (EDS): Supplemental Information

New technologies have led to a flood of data from satellites, medical record keeping, internet searches, social media posts, financial transactions, gene sequencing, embedded sensor networks, and a myriad of other sources. Most areas of the economy as well as most academic disciplines are undergoing at least some degree of transformation due to this new wealth of data. **To remain institutionally relevant, it is essential for ESF to be involved in the ongoing transformation of the economy and academia by the proliferation of data**. ESF does not have the institutional expertise to lead at the mathematics/statistics/computing science edge of data science, but there is potential for ESF to establish a niche in environmental data science. A series of recent interviews with leading data scientists (Gutierrez, 2014. Data Scientists at Work) emphasize that a core element of effective data science is understanding one's data; namely, domain knowledge (e.g. biology, ecology, hydrology, chemistry, etc.) needs to go hand-in-hand with the mathematics/statistics/computing science of data analysis. Thus, the central strategy of the Environmental Data Science (EDS) Initiative is to 1) identify partnerships with other institutions that have less domain knowledge but more mathematics/statistics/computing science expertise and 2) to develop a unique hands-on educational program to cross-train practitioners and academics in traditional domain-specific environmental science tasks as well as data science.

Project Start-Up Start-up would focus on three main efforts: 1) position ESF to receive NSF-funded training grants, 2) create training workshops and generate valuable data flows based on environmental monitoring efforts at our satellite campuses, and 3) develop partnerships with other regional institutions doing work in data science. The two NSF funded training grants we would focus on are the Research Experience for Undergrads (REU) and the NSF Graduate Training Program (NRT). An REU provides $300-400k in funds while an NRT is $3M in funds. An on-going priority area of the NRT program is the NSF theme "Harnessing the Data Revolution", a concept closely aligned with the EDS initiative. A letter of intent to submit an NRT grant this year has been submitted (PI-Shaw) with the expectation it may take more than one year before it is funded. A central criteria on which NRT proposals are evaluated is prior experience in implementing training elements. Requested startup funds would allow for piloting of: a seminar series, partnering with SUNY Poly on ESD workshops, and a data-science training "boot-camp" on ESF's northern Forest Properties, which will also serve as an *essential data source* for our growth in this area. Building on the data-rich long-term monitoring programs ongoing at Huntington Forest since the 1970s, start-up funds will support expansion of monitoring to other campuses (e.g., CLBS, TIBS and Ranger School) and link all ESF campuses via data flows that feed into our computer labs, design studios, and research and service projects. Our third emphasis would be on regional partnerships. Data science is a new "area of excellence" at SUNY Binghamton; SUNY Poly is starting an MS in data science; University of Rochester has funded a $50M data science program (the Goergen Institute); and SUNY Upstate has made key hires in data analytics. ESF is too small to go it alone in the realm of data science; institutionally we need to reach out and forge better connections. Funds would be used to support fellowships and seed grants on projects that promote these new partnerships and support relationship-building with foundations and donors interested in big-data frontiers.

Continued Growth At its core, the EDS Initiative is intended to develop capacity that ESF does not currently have. Data science will continue to be a growth area for decades with myriad applications in environmental science, engineering and policy. The EDS initiative will position ESF to compete for industry, philanthropic, state, and federal funds in a rapidly growing area.