# Application Form

# UMS Collaborative Program Support Fund

# 1. Proposed Program Title: Comprehensive Data Science

**2. Amount Requested:** $ 177,177

**3. Primary Applicant**

**Name:** Joseph Szakas

**Position Title:** Vice President of Academic Affairs/Provost

**Campus Address:** UMA, 46 University Drive, Augusta, ME 04330

**Contact Number:**(207) 621-3288**Email:** [szakas@maine.edu](mailto:szakas@maine.edu) \_

**Signature/Date:**

**4. Co-Applicant(s) Information**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Name** | **Email** | **Academic or Admin Office(s)** | **Role in Proposed Work or Program** | **Initials** |
| **1.** | Chris Bennett | [chris.bennett@maine.edu](mailto:chris.bennett@maine.edu) | Associate Professor of Computer Science | Expertise in data mining, computer science education, literacy and online delivery |  |
| **2.** | James Suleiman | [suleiman@maine.edu](mailto:suleiman@maine.edu) | Associate Professor of Management Information Systems | Expertise in data analytics, multidisciplinary learning, community engagement, curriculum development and course review. |  |
| **3.** | Bruce Segee | [segee@maine.edu](mailto:segee@maine.edu) | Director ACG/Prof ECE | Procure, install, maintain storage system and train users. |  |
| **4.** | Laura Wilson | [laura.wilson@maine.edu](mailto:laura.wilson@maine.edu) | 4-H Science Professional | Expertise in design and delivery of STEM (K-12) and experiential learning activities |  |

**5. Required Signatures:**

**UMA:**  Date:

Joseph Szakas, Provost

Date:

Diana Kokoska, Coordinator of CIS

**USM:**  Date:

Jeannine Uzzi, Provost

**UMF:**  Date:

Eric Brown, Provost

# Comprehensive Data Science

# Program Innovation Fund

# Proposal Application

# 2017-2018

## Description:

This proposal will provide support to scaffold a comprehensive Data Science offering within the UMS. There are five major components to this effort in collaboration with four institutions and existing System resources listed below:

**Data Science Plan

Text reading "UM STEM Ambassador (8th-12th grade) Data Science Modular development" in on the left of image. 

Arrow labled "Student Experience" moving left points towards text that reads "USM Data Science Masters Degree." 

Above Arrow READS "USM Data Science BS (Online).

Below arrow reads "UMF Data Science Modules (Service course, certificates and minors for non-majors)

Below that in a seperate box reads "Advanced Computing Group (ACG), Large Scale Data Storage to support UMS Faculty and Students"

**

1. Working with UM 4-H STEM Ambassadors, create data science modules for middle and high school students (UM) ($28,164)
2. Develop courses and modules to support data science efforts for non-majors. This work will result in the development of certificates and minors (UMF) ($27,500)
3. Continue course development to submit an Intent to Plan for a BS in Data Science, which would be offered online (UMA). UMA’s allocated funds would also coordinate meeting of PIs from around the system and joint development of curriculum ($15,000)
4. Develop a scalable data storage component that, once established, will be accessible system wide, and be made available to faculty and students. (ACG $106,513)
5. Collaborate with an existing PIF award last year to USM to build a Master’s degree in Data Science

Total request: $177,177

## Enrollment Demand/Enrollment Projections

A study by *CareerCast.com* states, “…data scientist jobs have the best growth potential over the next seven years, as they are one of the toughest jobs to fill.” The study further claims, “A quick search for data scientist jobs in the United States on *LinkedIn* reveals over 13,700 open positions. Additionally, this job trends tool by *Indeed*, which showcases the demand for data scientists, reveals that both data scientist job listings and job seeker interest are showing no signs of slowing down.” (<https://www.infoworld.com/article/3190008/big-data/3-reasons-why-data-scientist-remains-the-top-job-in-america.html>)

The *Bureau of Labor Statistics* claims 71% of all new jobs in STEM are in computing, and 8% of STEM graduates are in Computer Science. (https://www.bls.gov/)

We see the enrollment opportunity is strong and in line with other computer science/ computer technology/ computer information systems/ cyber security programs. It is worth noting the significant math requirements may be a negative factor (i.e., heavily based upon statistics), but the support needs from other programs, industry demand, and research potential, should offset this concern.

The list of Data Science degrees is a growing academic area around the country. Over 400 data science business analytics at the bachelors, masters and doctoral levels exist <http://datascience.community/colleges>. Note: there were none listed in Maine.

The *Chronicle of Higher Education* (March 4th article) states, “…[T]hey heard it from the Obama administration, and from consultants like McKinsey & Company, which in 2011 projected that the United States could face a shortage of as many as 190,000 people with those skills by 2018. They heard it from business leaders, who described having to retrain new hires to make them versatile data scientists…” <https://www.chronicle.com/article/Colleges-Rush-to-Ride/242674>

## Curriculum Overview

The following components of curriculum development are described below:

1. Create data science modules for middle and high school students

This effort will be led by UM and the 4-H STEM Ambassadors (Co-PI Wilson) in consultation with the other PIs and institutions in the proposal. Data Science, like other emerging disciplines is hindered by a lack of knowledge about what is data science, what activities and duties are involved in data science, and what are the available educational opportunities in the state of Maine. It has been shown that students begin to make decisions on academic fields of interest as early as the 8th grade. Introducing the opportunities to these growth areas are key for a vibrant Maine economy. The 4-H STEM ambassadors have been providing STEM education activities and learning opportunities to K-12 students throughout the State of Maine for many years, and expanding into this new arena of data science, without losing its support of other STEM disciplines, is essential. Revising and expanding their inventory of ten educational kits starting with their Maine Lakes Kit which will develop new learning activities that include statistical analyses and data visualization of Maine Lake/water quality.

A break out of Staffing, Materials and Supplies, and Travel costs totaling $28,164, appears in Appendix A.

1. Continue to develop courses, minors and certificates to support other disciplines within the UMS

This effort will be led by UMF and their faculty Co-PI Bennett, Koban and Hardy. Data science is becoming a key core competency for many disciplines outside the IT arena. The ability for students, and faculty to create, manipulate, and analyze large quantities of data and information are skills needed today in almost every (if not all) academic disciplines. Given the wide range of students, this will serve traditional undergraduates looking to augment work in their majors as well as non-traditional students who may be looking to add new skills while they work. All required coursework would be online, reaching a student pool beyond the traditional student possible.

Fortunately, the area of Data Science is quite integrated with existing computer science and information science degrees within the UMS already, so the creation of certificates and minors to non-IT disciplines is not starting from scratch and approximately already 70% of the way there.

This component of the proposal will be achieved through four major components:

* Needs analysis of technical and data science-related skills for every program of study on campus, from the perspective of three different groups: students, academic programs, and employers.
* Inventory of existing applicable courses across the system and identification of candidates for modification and ability to move online.
* Development of new courses to address existing needs. Many of these courses will be what we call Technology Modules, discussed in more depth below.
* Modification and modularization of existing courses to make them available online and to make them available in 1 – 4 credit options.

This part of the proposal will incorporate courses from across disciplines and across campuses to take advantage of existing assets and capacity, minimizing the need for wasteful replication and, hopefully, achieving a whole greater than the sum of its parts. The minors and certificate programs will pull mathematics, technology, and writing courses from existing offerings while technology modules will address the needs identified by students, programs, and employers.

The curriculum will consist of a combination of new and existing courses, with the core courses in statistics, computer science, and professional writing to be offered in face-to-face settings in fall and spring semesters and online in winter and summer terms. The 1- and 2-credit technology modules will be offered entirely online in the winter and summer terms initially, potentially offering them in fall and spring based on demand in the future. Several new modules will be developed in response to demand from the needs analysis.

The specific anticipated requirements for both minors and certificates are listed below:

### Draft Minors

Data Analytics

|  |  |
| --- | --- |
| **Course** | **Credits** |
| MAT 120 - Introduction to Statistics1 | 4 |
| COS 140 - Introduction to Computer Science1 | 4 |
| MAT 228 - Intermediate Statistics2 | 4 |
| ENG 200 - Professional Writing1 | 4 |
| TEC Module Electives3 | 4 |

Technology

|  |  |
| --- | --- |
| **Course** | **Credits** |
| MAT 120 - Introduction to Statistics1 | 4 |
| COS 140 - Introduction to Computer Science1 | 4 |
| ENG 200 - Professional Writing1 | 4 |
| TEC Module Electives3 | 6 |

1 – Currently exists as traditional face-to-face, must be ported to online as well.

2 – Currently being developed, must be ported to online as well.

3 – Some existing modules exist, many to be developed

### Draft Certificates

Data Analytics

|  |  |
| --- | --- |
| **Course** | **Credits** |
| MAT 120 - Introduction to Statistics1 | 4 |
| COS 140 - Introduction to Computer Science1 | 4 |
| MAT 228 - Intermediate Statistics2 | 4 |
| ENG 200 - Professional Writing1 | 4 |
| TEC Module Electives3 | 4 |
| TEC 300 Capstone Experience4 | 2 |

Technology

|  |  |
| --- | --- |
| **Course** | **Credits** |
| MAT 120 - Introduction to Statistics1 | 4 |
| COS 140 - Introduction to Computer Science1 | 4 |
| ENG 200 - Professional Writing1 | 4 |
| TEC Module Electives3 | 8 |
| TEC 300 Capstone Experience4 | 2 |

1 – Currently exists as traditional face-to-face, must be ported to online as well.

2 – Currently being developed, must be ported to online as well.

3 – Some existing modules exist, many to be developed

4 – To be developed

### Examples of Technology Modules

The technology modules are the key to the complementary nature of the minors and certificates, and their topics will be driven by input gathered during the needs analysis phase. We anticipate some of the modules to delve more specifically into broader topics such as (but in no way limited to) programming, data visualization, data mining, and advanced office applications). More specifically, courses may include:

* Statistical programming languages such as R and Python
* Data Visualization with Tableau, R, or Python
* Advanced MS Excel and MS Access
* Report Writing
* Data Mining with Python, Weka, and other platforms
* Web Technologies such as Content Management Systems and Web Programming
* Digital and Multimedia processing

True to their name, many of these courses will be modularized to offer a sequence of courses to keep offerings to 1 and 2 credits. This will also allow for threading different paths where multiple students may start with an introduction to Python but allow some to follow it down a statistics path, while others can pursue a data visualization path.

A detail of budget requests for this component is in Appendix B of this proposal.

1. Complete effort to develop remaining new courses for an Intent to Plan for an online BS in Data Science.

This continuing effort is led by UMA Co-PI Kokoska and Dube from UMA’s Computer Information Systems Department. UMA has been investing in curriculum development in the form of course development to both keep the CIS curriculum up to date, but also to prepare to support this new degree. As mentioned in the previous section the alignment with Data Science and Computer Information Systems is very tight, so approximately 70% of the courses are already created and running in support of the CIS degree.

A list of existing courses recently developed in support of both CIS and emerging Data Science includes the following:

* CIS 120 – Introduction to Data Structures
* CIS 150 – Introduction to Data Science
* CIS 218 – Introduction to SQL
* CIS 225 – Introduction to Health Informatics
* CIS 352 – Data Visualization
* CIS 449 – R Programming, Package Development, and Applications
* CIS 450 – Data Mining
* SOC 375 – Social Networks
* SOC 475 – Analysis of Social Networks

All of these courses are available online or at a distance.

And the final two proposed courses to be developed and funded through this grant proposal are:

* CIS 4XX - Statistical Quality Control
* CIS 4XX - Spatio-Temporal Information Science

Both of these courses will also be designed and made available online.

Although many of these courses have already run, there may have to be adjustments to align with the UMF and other UMS courses to simplify transfer. It should be noted that courses for majors and non-majors may have fundamental differences. Additionally, it is the hope that graduates either from this BS in Data Science or students with certificates or minors in Data Science would be well positioned to enter into the Masters of Data Science currently being developed at USM. To achieve that alignment, summer or fall curricular workshops are proposed to help support this curricular alignment effort.

A detail of budget requests is in Appendix C of this proposal.

1. Integrate the new large-scale data storage into the curriculum throughout the UMS

This effort is led by Co-PI Segee with The Advanced Computing Group (ACG)

To properly support the Data Science effort within all institutions of the UMS, there must be infrastructure created to support this data science effort both academically and for research for both faculty and students across the UMS. Large scale data sets needed by Data Science courses will need a place that is accessible as well as providing high performance. It makes no sense to create data storage silos among the institutions of the UMS. A centralized location that will support the storage, manipulation and analysis which can be accessed remotely is critical to this effort. Specific storage solutions and expertise are required when working with datasets as large as those needed by Data Science programs. Fortunately with the Advanced Computing Group, we have that expertise to assist in selecting the appropriate infrastructure that is best suited for Data Science. The proposed infrastructure has the advantage of being extremely scalable, so we can begin this at a modest size and know that we can grow as the data needs grow.

A brief Overview of the ACG, a description of the storage solution, called Ceph, and a discussion about the configuration, properties and access throughout the UMS is described below:

#### About the ACG

The Advanced Computing Group serves research and education across the University of Maine System, providing High Performance Computing (HPC), Cloud Computing and a variety of storage solutions. The HPC services is capable of supporting a multitude of simultaneous users running jobs that utilize hundreds of processing cores each. The nodes are connected together by an ultra-high speed and low latency network connection. Although a month contains approximately 720 hours, the ACG routinely delivers over 1 million processing hours per month in HPC. Cloud Computing services allows users to create, use, and destroy virtual computers with a minimal amount of time and effort. These can be used for special purpose processing, classroom use, or as day-to-day machines. Cloud Machines can be accessed from virtually anywhere in the world using only a web browser. The ACG operates and maintains over 1 PetaByte (1 million gigabytes) of usable disk storage. All of this storage is stored redundantly, so that no hardware failure will result in data loss, and all data is backed up nightly to Portland, to guard against data loss in the event of a catastrophic event in the Orono datacenter. The storage system consists of multiple systems from multiple vendors networked together and is relatively time intensive to maintain. Recently, (several years) the ACG has been experimenting with a new type of storage system called Ceph, discussed below, and has also been working with SUSE who is developing enterprise tools in this space. .

#### About Ceph

Ceph is an Open Source system for creating a single unified storage system that is infinitely expandable, can be presented to the end user in a wide variety of ways, supports multiple replications of data within a filesystem, provides snapshotting and ease of management. Furthermore, by being Open Source, the motivation is high for development efforts beyond a single corporation. The basic Ceph layout is shown in Figure 1. It consists of multiple storage systems pooled together as a Storage Cluster. The Storage Cluster distributes I/O requests across the servers available. The servers typically contain redundant arrays of disks to guard against a disk failure and the storage cluster maintains multiple replications of data to guard against a server failure. Because the actual Disk I/O is spread across many servers, performance gets faster as more servers (and disk capacity) are added. That is, the system gets faster and more robust the larger it gets.

Figure 1: Basic Ceph layout

At the bottom of the graph are columns labled Server 1, Server 2, Server 3, Server 4, & Server N. The bottom block in each column reads "CPU Memory Disk Networks." Text to the left labels this row as "Commondity Hardware" 

The next block up in each row reads "OS." Text to the left labels this row "Linux OS"

The next block up in each row reads "CEPH". Text to the left labels this row "Ceph S/W."

On top of all columns is one block that spans all columns reading "CEPH Storage Cluster." Text to the left labels this block as "Distributed Storage.

On top of that block is three arrows pointing up labeld "Object Stroage," "Block Storage," and "File Storage." Text to the left lables this row of arrows as "Unified Storage Solution."

Arrows point up towards a box that reads "Clients."

Figure 1: Basic Ceph layout

<https://www.safaribooksonline.com/library/view/ceph-cookbook/9781784393502/ch01s05.html>

Figure 2: Ceph vs. traditional storage

On the left is a columns titled "Traditional Storage", and on the right titled "Ceph". Under each title are colums of 4 blocks. 

The blocks in the first row reads "Propoetary hardware" on the left columns, and "Commodity Hardware" on the right column.  A doule sided arrow pointing to the left and right is between them.

The blocks in the second row reads "Proprietary software" on the left columns, and "Open Source software" on the right column.  A doule sided arrow pointing to the left and right is between them.

The blocks in the third row reads "Life cycle enforced by vendor" on the left columns, and "Hardware" on the right column.  A doule sided arrow pointing to the left and right is between them.

The blocks in the fourth row reads "Hard scale limit" on the left columns, and "Exabyte scale" on the right column.  A doule sided arrow pointing to the left and right is between them.

The blocks in the third row shows 4 dollar signs on the left columns, and 2 dollar signs on the right column.  A doule sided arrow pointing to the left and right is between them.

Figure 2: Ceph vs. traditional storage

<https://www.slideshare.net/gnyers/suse-storageagentleintroduction>

*Ceph at the ACG*

The ACG has been using Ceph storage for several years, during which time the maturity has increased dramatically. It has gone from "interesting" to "production ready" and the Ceph cluster currently supports our Cloud as well as our SeaFile data storage solution. We are working on merging our other storage solutions into the Ceph cluster and are currently working closely with SUSE, one of the biggest developers of Ceph, in extending the Ceph cluster from Orono to Portland. This work, if successful, will create a single, live, parallel file system that spans from Orono to Portland. (Currently the live data is in Orono, with Portland used for offsite backup.)

#### Needs to support Data Science across UMS

The proposed system-wide Data Science program will require growing the existing ACG Ceph cluster by an additional 400 TB of usable storage in order to support the large data sets necessary for true Big Data education and research. This will require four storage nodes at a cost of $25,000 each. Each storage node will contain 30 10TB disks, or 300 TB of storage. The four storage nodes contain 1.2 PB of storage, but because three copies of all data are maintained at all times to guard against data loss and improve throughput, the total usable storage is 400 TB. Because all of the necessary Ceph cluster equipment is set up, storage can be grown in 100 TB increments at a cost of $25K. With each additional storage node, not only does the amount of storage increase, but performance increases as well.

#### Access

The proposed disk storage array will be connected through existing fiber optic networking to every campus in the University of Maine Network.  Because of the nature of fiber optics, the data rate is not limited by the medium, but only by the end electronics. Thus, the networking speeds can continually increase without needing to replace hundreds of miles of cable.  Currently, access at the campus level is 40 Gb/s or higher, with individual researchers, educators, and students connected typically at 1 Gb/s. Naturally, those who use ACG resources for computation (i.e., HPC and cloud) will access the storage at the rate in the data center (40 Gb/s). The connectivity between the ACG and the commodity internet is very high. Thus a user at home or on campus would have the same experience with the data in Orono as at the local campus. The Office of Information Security routinely tests ACG resources for vulnerabilities to ensure the safety and security of user data. Various measures can be employed to further protect the data as warranted, including encryption of directories and two-factor authentication among others.

#### Training

The ACG has an NSF-funded Cyberinfrastructure Engineer position that is filled by Larry Whitsel who has a PhD in Computer Science.  The role of the Cyberinfrastructure Engineer is to assist faculty and students with using advanced computing resources. After the proposed storage system is commissioned and deployed, the ACG will provide no-cost training sessions on the effective use of the storage system.  The training sessions will be a combination of in-person and at-a-distance. These training sessions will be recorded and made available for future reference, both for new users and as a refresher for existing users. The ACG assumes all responsibility for operation and maintenance of the proposed infrastructure.  The ACG will also provide individual support, where appropriate, to assist with difficulties as they arise.

A breakout of the ACG budget is in Appendix D.

1. Align the undergraduate curriculum with the graduate curriculum that was a supported PIF awarded last year to develop a Master’s degree in Data Science in Fall 2019.

Awarded from a VCAA PIF in 2017, PI Professor James Suleiman have been working an Intent to Plan for a MS in Data Science at USM. This proposal aligns well and complements Professor Suleiman’s current effort. If funded, this proposal will expand the K-12 and undergraduate interest in Data Science, which will have a potential direct effect on enrollment for this new graduate program. The infrastructure alone would provide great benefits for the Master's degree education and research efforts (see attached letter of support from Professor Suleiman).

A brief description of the MS proposal is inserted below for your reference.

Data science is an interdisciplinary field focused on extracting value from data. It is “The art of turning data into action.” There is growing recognition nationally and in Maine of the need for data-driven decision making in all sectors and industries. Maine lags behind the nation in its capacity to fill this growing need. Yet, the University of Maine System (UMS) currently offers no undergraduate or graduate programming in Data Science. The Master of Science in Data Science Planning Initiative proposed in this document will develop the architecture and infrastructure necessary to close this critical workforce gap. The primary faculty and staff engaged in this initiative have collaborated successfully in USM’s Health Informatics Research Cluster (HIRC) for more than two years. They have now formed a planning team to design a Data Science program as the next logical step for USM and the UMS.

Goal: The goal of this multidisciplinary initiative is to develop a competency-based curriculum for a graduate program in Data Science that will launch in Fall 2019. The Master of Science in Data Science (MSDS) will be housed at the University of Southern Maine (USM) and designed to serve as a statewide resource for the UMS.

The budget request to support this is found within the UMA component where Professor Suleiman will be invited to participate in all curriculum and Ceph Training events.

## Faculty/Personnel

### UMA:

Diana Kokoska is the Coordinator of the Computer Information Systems program at the University of Maine at Augusta. She is a pioneer in online course development and has successfully spearheaded efforts to guide the UMA CIS program into expanding areas of technology. Her academic areas of expertise are web design & development, graphic design, advanced productivity applications, and online databases with PHP and MySQL. She co-authored the current PHP Programming with MySQL textbook published by Cengage Learning. Diana is an active member of the Mainely Tech Women group whose mission is “to promote educational, networking, career, and collegial opportunities for girls and young women in Maine who are sparked by technology.”

Dr. Matthew Dube is an Assistant Professor of Computer Information Systems at the University of Maine at Augusta. He holds a Ph.D. and an M.S. in Spatial Information Science and Engineering from The University of Maine and a B.A. in Mathematics and Statistics from The University of Maine. He currently teaches courses in data science areas ranging from Data Mining to Database Design and is an active researcher in Qualitative Spatial Reasoning, Mathematical and Computational Redistricting Procedures and Analysis, and Equine Statistics, serving also as a reviewer for numerous geospatial information journals and conferences. His research work crosses data science disciplines such as geoinformatics, data mining, and knowledge communication and discovery. He also teaches courses for the University of Maine at Augusta Honors Program centering on Interdisciplinarity, a key staple of this rapid growth and dynamic content field.

### UMF:

Chris Bennett is an Associate Professor of Computer Science at the University of Maine at Farmington. His research interests include data mining and computer science education, and he has worked on many projects to improve computer science literacy at every level of education, including online delivery of computing curricula.

Lori Koban is a Professor of Mathematics at the University of Maine at Farmington. Her research is in combinatorics, especially matroids and their connections with graph theory. She has developed interests in both actuarial education and mathematics education, including online delivery of mathematical curricula.

Sarah Hardy is a Professor of Mathematics at the University of Maine at Farmington. Her expertise is in statistics and applied mathematics. Her research areas include data science and data mining, advocating for an integration of mathematical and technical skills across the liberal arts.

### USM:

James Suleiman, PhD, Associate Professor of Management Information Systems, will participate in the curriculum development and course review for the MSDS, as well as the community-engaged conversations with industry leaders. Dr. Suleiman has connections to the southern Maine business community and leads the Business School’s graduate certificate program and MBA concentration in business analytics. These MBA courses are likely to serve the MSDS program, and Dr. Suleiman has demonstrated a strong commitment to multidisciplinary learning in data science by cross-listing these courses with other graduate programs including public health and regularly has students from UMaine graduate programs in his classes. As a result, the data analytics courses in the MBA program have served as a testing ground for this new venture and Dr. Suleiman’s experience with them will provide critical information to the planning process.

### UM:

Laura Wilson is a 4-H Science Professional with the University of Maine Cooperative Extension. She has extensive experience in the design and delivery of STEM outreach to both adult and youth audiences. Wilson co-developed the 4-H STEM Ambassador program, and coordinates the program for the UMaine and UMA. She also oversees the design, development, and testing of experiential learning activities (4-H Science Toolkits).

### ACG:

Dr. Bruce E. Segee is Director of Advanced Computing for the University of Maine System as well as the Henry R. and Grace V. Butler Professor of Electrical and Computer Engineering at the University of Maine. His research is focused in high performance computing, cloud computing, visualization, and big data. He interacts routinely with K-12, undergraduate, and graduate students, and researchers and educators from all disciplines. He has recently received a $2M NSF award to teach K-12 students the basics of computer programming using the popular game Minecraft. Other recent funding has been targeted to improving Maine's cyberinfrastructure for research, education, and economic development.

## Program Assessment

The proposal will be measured by the distinct academic components that are created:

* Number of Minors and Certificates for non-majors created
* Number of STEM Science Kit activities adapted for Data Science exploration
* An online BS in Data Science
* A working large-scale data storage that is available to all institutions within the UMS
* Alignment of K-12 and undergraduate efforts with the MS in Data Science.

## Proposed Budget

Summary of the budget was listed in the description section. Each institution’s budgets are listed in the following appendices.

# Appendix A

# UM 4-H STEM Ambassador program and 4-H Toolkits

**1.  Staffing**: one half-time 4-H Community Education Assistant.  20 hours/week $14/hour for one year. This position will facilitate development and testing of the new data science related activities with students from multiple campuses, as well as provide coordination assistance for the 4-H STEM Ambassador programs at UMA and USM.  Their role will be to support the program, allowing Laura Wilson and Sarah Sparks (STEM Ambassador coordinator, USM and UMF) to develop and test the data science activities.

$14/hour, 20 hours a week, for an annual cost of $14,560, fringe of $7,804, total cost - **$22,364**

**2.  Travel:** Travel is requested for the new staff member to train and mentor 4-H STEM Ambassadors.   UMA STEM Ambassadors are located around the state, and training/support/follow-up needs to occur on site   **$800**

**3.  Materials and supplies:**  We propose to revise at least one existing 4-H Science toolkit (curricula 4-H STEM Ambassadors use when they facilitate activities with youth) and adapt activities to embed data science components.  Our first candidate is the Maine Lakes kit.  This kit includes environmental science data sources referenced; we will develop learning activities that include statistical analyses and data visualization of Maine lake water quality data.  We estimate for the 10 existing kits to be revised a budget of **$5000** in materials and supplies.

# Appendix B

# UMF Minor and Certificate Development

The budget is broken down by the two major components of the project. The first is needs analysis and inventorying of existing courses for creation, adaptation, and modularization and the second is the creation of new courses as well as the porting and modularization of other courses. The first component will be completed by January 2019, but output from the work would inform course creation and adaptation starting Fall 2019.

There will be a coordinator who will oversee all steps of the work at UMF and act as liaison with other campuses and employers. The work of the coordinator is expected to last the full academic year of 2018-19.

For course creation and porting, we have budgeted $750 per credit hour. We anticipate the need to port four courses while creating at least eight new modules. Some of these will be 1 credit while others will be 2 credits so we are budgeting for a total of 12 credit hours of these modules.

This work requires breaking up an existing online course worth greater than 2 credits into smaller courses of 1 or 2 credits. The number of these modularizations is difficult to determine until after the needs analysis and existing resource cataloging, but we budget for 12 credits of this work.

Coordinator $3500

Course Porting

COS140 @ 750/credit $3000

MAT120 @ 750/credit $3000

MAT228 @ 750/credit $3000

ENG200 @ 750/credit $3000

New Technology Modules

1 credit module x 6 @ 750/credit $4500

2 credit module x 3 @ 750/credit $4500

Modularization of Existing Courses

12 credits @ 250/credit $3000

-------------------------------------------------------------------------

Total $27,500

# Appendix C

# University of Maine at Augusta

New Technology course

3 credit CIS 4XX Statistical Quality Control module $4000

3 credit CIS 4XX Spatio-Temporal Information Science $4000

Software license/ Materials to support curriculum effort $3000

Summer/Fall curriculum PIs all day workshop and Ceph Training

Stipends $500 x 8 PIs (3 UMA, 3 UMF, 1 UM, 1 USM) $4000

Total $15,000

# Appendix D

# ACG budget to purchase 4 Ceph Nodes

   Storage node:            $21,863.15

   SUSE support:            $ 3,465.00

​   Cables:                  $    100.00

   .25 of a Rack and PDU's: $    800.00

   Infiniband Switch port    $    400.00

   ======================== ==========

   Total per node:          $26,628.15

4 nodes requested $26,628.15 x 4 = $106,512.60

# Data Science, Bachelor of Science (DRAFT)

## Bachelor Degree Requirements:

* Minimum 121 Credit Hours
* Writing Intensive Course
* Minimum Cumulative G.P.A.: 2.00
* 30 Credit Hours of Residency Courses
* 9 Credits of Upper-Level Major Residency Courses
* Minimum G.P.A. in the Major: 2.00

## Program Major Requirements (91 credit hours):

### Program Core (72 credit hours):

* CIS 101 Introduction to Computer Science (3) ★
* CIS 110 Programming Fundamentals (3)
* CIS 120 Introduction to Data Structures (3)
* CIS 150 Introduction to Data Science (3)
* Complete any 200-level programming course (3):

*CIS 212 Introduction to Visual Basic Programming*

*CIS 214 Introduction to Java Programming*

*CIS 215 Introduction to C++ Programming*

*CIS 216 Programming in C# and .NET*

* CIS 218 SQL (3)
* CIS 225 Intro to Health Informatics (3)
* CIS 255 Database Design (3)
* CIS 280, 380 or 480 Internship (3) would move out of the core and become a requirement of the concentration
* Complete any advanced-level programming course (3):

*CIS 312 Advanced Visual Basic*

*CIS 314 Advanced Java Programming*

*CIS353 HCI*

* CIS 350 Database Management (3)
* CIS 351 Database Management Systems: Oracle (3)
* CIS 352 Data Visualization (3)
* CIS 354 Algorithms and Data Structures (3)
* CIS 360 Geographical Information Systems (3)
* CIS 449 R Programming, Package Development and Applications (3)
* CIS 450 Data Mining (3)
* CIS 460 Computers & Culture (3)
* CIS 4XX Statistical Quality Control (3)
* CIS 4XX Spatio-Temporal Information Science (3)
* INT 208 Introduction to Interdisciplinary Studies (3)
* SOC 375 Social Networks (3)
* SOC 475 Analyzing Social Media (3)
* SSC 220 Basic Research Methods (3)

### Mathematics (19 credit hours):

* MAT 112 College Algebra (3) ★
* MAT 115 Elementary Statistics I (3)
* MAT 125 Calculus I (4)
* MAT 261 Applied Linear Algebra (3)
* MAT 280 Discrete Mathematical Structures (3)
* MAT 315 Applied Statistics and Data Analysis (3)

## Other Program Requirements (28 credit hours):

* Complete any 100-level Communications course (3) ★
* ENG 101 College Writing (3) ★
* ENG 317W Professional Writing (3) ★
* Complete one of the following Fine Arts electives (3) ★:

*ART 1xx any 100-level Art course*

*DRA xxx any Drama course*

*ENG 351 Creative Writing I*

*ENG 452 Creative Writing II*

*MUH 1xx any 100-level Music History course*

*MUS 1xx any 100-level Music course*

* Complete two of the following Humanities electives (6) ★:

*AME xxx any American Studies course*

*ARH 105 History of Art & Architecture I*

*ARH 106 History of Art & Architecture II*

*DRA xxx any Drama course*

*ENG xxx any English course (except ENG 101 or 317w)*

*FRE xxx any French course*

*HGH xxx any Holocaust, Human Rights & Genocide course*

*HTY xxx any History course*

*HUM xxx any Humanities course*

*MUH xxx any Music History course*

*PHI xxx any Philosophy course (except PHI 135 or 335)*

*SPA xxx any Spanish course*

*WGS xxx any Women and Gender Studies course*

* Complete any 100-level Laboratory Science course **BIO 110 recommended** (4) ★
* Complete two of the following Social Science electives (6) ★:

*ANT 1xx any 100-level anthropology course*

*ECO 1xx any 100-level economics course*

*ECO 201 Macroeconomics*

*ECO 202 Microeconomics*

*JUS 1xx any 100-level justice studies course*

*POS 1xx any 100-level political science course*

*PSY 1xx any 100-level psychology course*

*SOC 1xx any 100-level sociology course* ***SOC 101 recommended***

*SSC 1xx any 100-level social science course*

## GENERAL ELECTIVES (2 CREDIT HOURS):

Complete any 100-level or higher general elective **BIO320 recommended** (2)

**General Education:**

It is the intention of the University of Maine at Augusta that every degree graduate will be prepared to function in our society as an effective and informed citizen. To this end, the faculty has designed a set of minimum expectations that students are expected to satisfy. The aspirations are defined by core skills, competencies, and abilities as well as knowledge based learning experiences that are the grounds for the General Education Requirements.

Courses notated by a ★ symbol represent a select minimum of courses within this major that satisfy the UMA general education requirements.

Students are encouraged to contact their faculty advisor and the Advising Center for academic advising and support services throughout their stay at UMA.