**Project Description**

**1. Vision**

The continued rapid evolution of computation and its impact on the world creates a new *challenge* for designing the learning experiences about computation at the university-level. Beyond the computer science discipline itself, deeply informed skills and knowledge about computation are needed in at least all STEM-H disciplines and new requirements are emerging in historically non-traditional fields (e.g., “digital humanities”). Furthermore, the growing awareness of “computational thinking” as a 21st century competency requires learning about computation to be positioned in a university’s general education curriculum. Weaving together the curriculum, pedagogy, and tools that engages learners with starkly different disposition and expectations about their learning of computation is a critical challenge. We address this challenge by crafting *authentic experiences* that engage *multi-disciplinary cohorts* of learners with computation both *in and across contexts* and whose efficacy is evaluated by *rigorous assessment*.

The continued rapid evolution of computation and its impact on the world also creates a new *opportunity* for designing the learning experiences about computation at the university-level. The simultaneous importance and availability of “Big Data” opens exciting possibilities for learning. Techniques for handling Big Data have become crucial to science, business, and policy making at all levels of government. Frequently in the news, Big Data provides a novel theme around which engaged learning can occur. The characteristics of Big Data match the requirements of the desired learning experiences. Big Data can underpin authentic experiences because the data derives from real phenomenon (e.g., geophysical events or social media), is from authoritative sources (e.g., US Geological Survey or Reddit), and is of genuine scale and complexity (not a “toy” version). Further, Big Data germane to a broad array of disciplines is available making it relevant to the learner’s concerns. However, technical difficulties have kept Big Data topics out of the introductory learning experiences. This omission deprives most students of the chance to gain an appreciation for the possibilities and risks associated with Big Data, and removes a novel and interesting class of problems from the set of learning opportunities. We address this deficiency by creating *carefully scaffolded technology* for manipulating Big Data streams available on the web which can be *easily deployed* by instructors to *quickly and seamlessly* incorporate this compelling dimension into new learning experiences.

In addition to its application in core CS1 and CS2 courses, revisions of the general education requirements at Virginia Tech have created a remarkable opportunity to develop and evaluate elements of curriculum, pedagogy, and technology for an introductory course open to students in all majors that: (1) serves as a model for other universities grappling with the challenge of providing computation for all STEM-H students or computational thinking for all students, (2) provides an onramp for developing minor courses of study in computer science, and (3) contributes to evolving discussions of how students are best introduced to the computer science discipline.

This proposal builds on the success of the RealTimeWeb project, our framework for rapidly building real-time web-data centered assignments in introductory courses. The RealTimeWeb tool chain employs novel, scaffolded technology to work with challenging, but motivating, dynamic and/or large-scale data. The framework has been successfully deployed in courses at Virginia Tech and the University of Delaware. Providing technical scaffolding to empower students to work with Big Data will leverage and enhance techniques that we have successfully applied to real time data.

**Intellectual Merit:**

By introducing authentic, massively-sized datasets of relevance into the early undergraduate curriculum, we can create a more engaging experience for students where they develop a foundational knowledge of Big Data concepts. This project will provide an excellent opportunity to improve the theoretical understanding of the problems and best practices of teaching Big Data, even as it offers a practical method for individual instructors. This project will also provide insights about the challenge of designing engaging and relevant contexts for projects that positively impact CS learning outcomes, while simultaneously enabling students to work with one of the most important technologies of the modern era.

**Broader Impact:**

This project will improve recruitment, retention, and engagement of students within the Computer Science disciplines. Prior research indicates that women are more likely to study and excel in Computer Science when content is contextualized and proven useful for solving real-world problems. Non-major students can be given realistic assignments that can more directly relate to their intended line of work, further increasing their motivation to succeed during their time in a CS course. Appropriately redesigning programming projects to involve interesting, contextualized Big Datasets is likely to improve the relevance and attractiveness of Computer Science to a wider community.

**2. Approach**

**2.1 Pedagogy**

Our approach is founded on *authentic problem-based learning* enabled by Big Data and a novel *multi-disciplinary cohort* model for courses that are not strictly CS-only.

Authentic problem-based learning comes from Situated Learning Theory, originally proposed by Lave and Wenger, which argues that learning normally occurs as a function of the activity, context, and culture in which it is situated [lave-situated]. Therefore, tasks in the learning environment should parallel real-world tasks, in order to maximize its *authenticity*. Authenticity has several different, interrelated meanings: (1) the problems are of realistic sizes, (2), the activity builds on students’ individual interests and meaningfully relate to the real-world to promote cognitive engagement with the task at hand, and (3) the concepts and techniques are applicable to tasks the students will encounter subsequently in their field of study. Contextualization is a key factor in these settings where learning is driven by the problem being solved, rather than the tools available -- therefore, the problem being solved should lead directly to the concepts being taught and the tools employed. Our project builds heavily on this educational theory.

High levels of authenticity are made possible through a Big Data framework that is scaffolded – an artificially simplified designed to support novices as they develop expertise. Scaffolding is designed to remove certain complexities until a student is prepared to handle them, or in order to reduce the cognitive load of a learner. The scaffolding we propose is less necessary for high-level courses, but it is critical for the introductory level. As students progress through the curriculum, the scaffolding is Faded – parts are incrementally removed, in order to give students a full understanding of the underlying systems.

We envision that courses would involve substantial projects chosen to reflect real-world concerns and involve issue of complexity and scale. Course material is delivered on-demand and in service of the students’ project work. To support the projects, part of our work is to construct a rich catalog of resources that can be manipulated with the software tools used in the course. The student’s project work focuses on “learning in context” because the student will choose projects that relate to his or her disciplinary area. The student engages in active learning because, as an individual, a student self-directs the selection, exploration, and completion of a project relevant to their major field of study. While minimum requirements will be set that all projects must meet, the student will have wide latitude to self-direct their project work.

Our approach also uses a novel *multi-disciplinary cohort* model for courses that are not strictly CS-only. The multidisciplinary cohort is a form of peer learning. For each project the students are organized into cohorts of 4-6 students so that each cohort has the greatest degree of diversity among the students’ actual or intended major fields of study. The activities of the cohort support “learning across contexts” because the peer interaction will provide each student a perspective of how computational thinking applies in the other students’ disciplines. As a member of the cohort a student is responsible for presentation, interaction, and support. These activities are:

* Presentation: describing to the other cohort members the significance of the project they have selected based on what the student has learned from their individual exploration.
* Interaction: asking questions and providing feedback about the projects of other cohort members. Through this interaction the student strives to gain some insight into the common computational techniques that are used across disciplines.
* Support: helping other members of the cohort with the mechanics of the tools and frameworks that are common across projects.

Part of the planned assessment will evaluate the efficacy of the multi-disciplinary cohort model and give great insight into intra-cohort dynamics.

**2.2 Technology**

Our approach builds on the success of the RealTimeWeb project to create a generalized, adaptable framework that can be instantiated for any Big Data source. RTW contains (a currently minimal) gallery of data streams drawn from publicly available web sources. For example, a current data source is large-scale weather data made available by the Weather Service. The essential service provided by RTW is to simplify the access of student code to the large and complex data steams available on the web. The data streams may be access in real-time or may be cached for later access.

Central to RTW are the client libraries of pre-built functions, one library for each data stream, which are used by the student’s code. The client library includes both the code and the documentation. These libraries allow the student to focus on the algorithms they must construct in order to process the data stream according to their project goals. While a prototype version of RTW has been constructed there is significant work needed in preparation for the proposed course.

To rapidly convert any third-party, massive, dataset into a practical educational resource, the proposed deliverable will utilize standardized Software Engineering methodologies and techniques. To that end, we will generate an extensible software framework. Libraries and frameworks are key building blocks for software development -- they provide predefined, extensible software components. Libraries and frameworks are similar, but subtly different in an important way. While both libraries and frameworks offer recyclable elements of functionality, frameworks also define an architecture that the developers can work around to build their software. A typical Object-Oriented software framework is composed of a collection of classes that define some central features. To create application-specific features, the developer can then add custom classes that extend abstract base classes. By developing our project as a framework, instructors that want to work with a new data source can simply extend our architecture.

They are free to cherry-pick the features and functionalities that we offer through our system.

**2.3 Curriculum**

In our experimental version of the course students will be introduced via a scaffolded approach to key programming concepts that allow them to complete significant Big Data projects. A combination of Blockly [ref] and Python will be used. Blockly is used to introduce the fundamentals of algorithms. With Blockly the syntactic detail of programming languages are set aside in favor of a visual representation of the algorithmic structure. Thus, students can gain confidence in their ability to construct algorithms before having to cope with the extra detail of the syntax of the programming language.

Blockly provides a collection of “blocks” which are shaped so that they can be assembled only in ways that “makes sense” syntactically. Blocks have slots into which other blocks can be inserted to form complex algorithms. Algorithms can be executed to see their effects. Simple, limited ability for input/output is provided. Blockly is extensible so that blocks with application specific semantics can be defined. An interesting aspect of Blockly is that the python code for a Blockly algorithm can be seen. This makes the transition from Blockly to Python a more progressive step for learners.

To provide learning materials for both Blockly and Python we have been in conversation with Runestone regarding their interactive on-line Python book [ref]. It is technically feasible to replace the interactive Python examples with equivalent examples using Blockly. Furthermore, with sufficient development support, it would be attractive to replace the existing interactive examples with ones which were specific to the data streams and Python structures being used in conjunction with RTW and to develop new Blockly blocks specific to RTW data streams as additional scaffolding.

**3. Proposed Work**

Support is needed to accomplish three main tasks:

1. Develop curriculum materials including plans for in-class demonstrations, presentation materials and associated learning resources.
   * Define project ideas for each newly identified data stream to give student users some potential targets for their work. This task involves interaction with disciplinary experts. Project ideas will be developed for CS 1 and CS 2 (Dr. Tilevich) and CS3 (Dr. Shaffer) and a computational thinking course (Dr. Kafura).
   * Expand initial contacts with Runestone to retarget the on-line interactive material to incorporate Blockly examples and RTW data streams.
2. Expand RealTimeWeb
   * Identify additional data streams for inclusion in the RTW gallery. These data streams must relate to as many disciplinary areas as possible.
   * Develop the code and documentation for the library elements for each newly identified data stream. This includes the development and testing of the framework code, writing and review of the documentation, and archiving of sample data streams.
   * Expand the RTW framework to include additional linkages to other services needed by student users. Extensions are needed for graphical, visualization, and statistical services.
3. Develop of assessment mechanisms
   * Quantitative assessment of key cognitive assessments against defined learning objectives
   * Qualitative assessment of inter-disciplinary cohorts using interviews and other observational techniques
   * Assessment of motivation using MUSIC model