Defining the WSU Research Computing Portfolio

Research-associated computing at WSU is intricately connected to measurement science, understanding distributions of data, and using this data to predict and understand complex behavior across spatial and temporal scales. While the research portfolio covers diverse subject areas, there are many algorithms, methods, strategies, and software that cross-cut domain applications and therefore working outward from common computing core elements enables progress in seemingly unconnected fields.

Five major research themes are common in WSU's High Performance Computing (HPC)-dependent research:

1. Resource Management. As a land-grant institution, WSU's mission includes providing education and technological innovations directly relevant to daily life, supported by cooperation between the federal government and states through the county extension network. Resource management is essential to the land-grant mission in the 21st century and involves identifying ways to sustainably manage our land, water, air, and ecological resources in the face of global change without creating unintended consequences. Here, WSU leads cutting-edge research, from the efficient synthesis of high-value and commodity chemicals in ways that minimize waste, to "smart" farm and production management, and to system-scale analysis planning and management of our water resources. Integrating research and education with extension is now revolutionizing the scope of data-driven applications in ways previously unimaginable. Modeling and analyzing resource management can now include essential inputs from physics, chemistry, biology, ecology, psychology, engineering, and economics, where several scales of size and complexity are integrated. This research portfolio encompasses:

- *Resource Management at the Smallest Scale: Separating Atoms and Molecules:* Understanding the fundamental balance of forces that lets us separate complex mixtures of materials, from those that contaminate our environment to biofuels and high commodity chemicals.
- *A Telescopic viewing of Biology from Molecules to Organisms to Ecosystems:* Scaling from genes to ecosystems and explicitly coupling biotic and abiotic systems includes large-scale comparative genomics studies, environmental simulation, and incorporation of biogeochemical cycling. At the species scale, this means moving beyond characterizing communities, to understand the influence of species at different spatial and temporal scales in an ecosystem.
- *Organisms Adapting/Adapting organisms:* Altering a small number of genes in an organism can allow it to survive otherwise challenging environments—prominent examples include antibiotic, pesticide, and herbicide resistance. However, most traits are complex and may require simultaneous shifts in many genes. In the context of the changing environment, can organisms keep pace or will moving organisms or alleles facilitate the adaptive process?
- *Coupled Human-Natural Systems:* In the face of global change, it is critical to take an integrated (or system-scale) viewpoint of resource management. The Columbia River Basin is a perfect testbed to understand the food-energy-water nexus and the extent to which natural resource policies and best management plans will alleviate or exacerbate conflicts between each of these sectors. This area requires upscaling of social and biophysical processes occurring at finer spatial scales to assess how larger-scale decision making impacts the system as a whole.

2. Dynamic and Responsive Systems. Examples at WSU where dynamic and responsive systems are required include precision agriculture, smart and adaptable human environments,

and the power grid. Power grids exemplify many of the issues that are typical of these systems and show the role that research computing can play in making the systems more stable and allowing them to use fewer resources. The rapidly transforming electric power grid faces numerous and often unpredictable challenges due to uncertain and fluctuating supply and demand created by evolving structures in an uncertain world. Current problems can result from misaligned and unbalanced infrastructure, incorporation of alternative energy sources, and the increasing severity and frequency of extreme weather events. In power grid management, uncertainties in demand-supply imbalance have traditionally been managed by increasing and maintaining excess infrastructure capacity. However, recent changes in technology and regulations have focused on maintaining power delivery and reliability by incorporating demand response strategies that may include a detailed understanding of consumption patterns. Machine learning methods using research computing infrastructure is just one of several examples where specific patterns in the data, in this case consumption, can be used to provide predictive insight into future behavior to proactively optimize system infrastructure design.

A complementary approach can be taken with agriculture, where a combination of measurement, modeling and prediction, can be used to inform best practices. One early example developed at WSU included the integration of weather information, area-wide trapping of insects, and life history information of insects and crop plants, to predict where and when to apply pesticides so as to lower insect populations through timed and targeted application; in turn this helped maximize crops, and minimize application costs and environmental impact. Newer technologies, including GIS and aerial monitoring of crop growth, are being used to precisely target fertilizer application and make similar management decisions on a very fine scale. Such precision farming will be sensitive to a changing landscape of cultivars, markets, and weather conditions. For farming strategies to be tested and implemented, it may be important for data sources to be able to transmit directly to the HPC or for interfaces to be accessible via alternative routes, such as web-based applications or cloud-based data collection and analysis.

3. Complex Systems and Emergent Phenomena. In many cases, the appropriate description of a system depends on the scale at which the system is studied. Take water for example, at everyday scales, a glass of water is described by a few properties such as its temperature and volume (thermodynamics). Probed a little more deeply, one notice flow in the glass (hydrodynamics). This picture is very different, however, from the one that would be adequate at a microscopic level to describe hydration of a protein, where the position of individual water molecules must be described using molecular dynamics. At an even smaller level, these water molecules must be described using quantum mechanics.

This behavior is typical of *complex systems*: different techniques are required at different scales to understand the system due to *emergent phenomena* – interesting behavior at large scales “emerge” from properties of their substructures in ways that are not obvious when those substructures are considered individually. Understanding complex systems is thus intrinsically a *multiscale* problem where computing plays a central role in both modeling these systems at a given scale, and connecting phenomena at different scales.

High-performance computing (HPC) has played a transformative role in understanding complex systems. For example, massively parallel processing enables microscopic theories to be solved in ways that extend to larger scales, allowing researchers to predict and explore emergent properties by directly calculating with microscopic theories. HPC thus plays an incredibly important role in discovering new phenomena with practical applications, or in resolving long-standing mysteries. At WSU, a wide range of such research is carried out across many fields, some of which are described below. The following list provides several categories of complex

systems where we project continued growth, competitiveness, and funding opportunities that are suited to the skills and areas of interest to WSU research programs:

- *Complexity and emergence in biological systems:* Correlating relationships in biological systems are notoriously challenging to disentangle and there remain significant opportunities for using data science techniques beyond for post-processing to formulate hypotheses and optimize experimental studies. Such efforts transcend scales of relationships to reveal new inter-dependencies and guide resource management efforts.
- *Transcending phases of matter:* The accurate predication of phase transformations is essential to a vast range of technological, industrial, and fundamental science questions (from astrophysics to nuclear energy). Phase diagrams, and prediction of the properties of different phases of matter, often rely upon a complex suite of physics models that may include relativity, electron correlation, and can be of increasing dimensionality in composition phase space.
- *Bridging length and timescales (accounting for both static and dynamic properties):* Simulation methodologies are generally limited to specific length and timescales and it remains a grand challenge to create an integrated multiscale computing software paradigm. Data science methods in conjunction with method development is essential toward this task.
- *Synthesis of heterogeneous information/data to predict emergence in complex systems*: The concept of emergence fundamentally derives from properties of a complex system that are absent in its constituent components. Emergence may only occur at specific conditions or be associated with a wide range of physicochemical characteristics. Thus, predicting emergence is increased if a wide range of data is employed that is measured or simulated across vastly different length and timescales. Fusing heterogeneous data in a manner that facilitates uncertainty quantification is a significant research challenge.

4. Data-intensive Science. Improvements in data collection technologies, computing, statistics, and the use of machine learning across multiple scientific domains, are creating copious amounts of data and are driving a shift in how research is accomplished. The shift is resulting in a new fundamental paradigm of scientific discovery known as data science. Collecting, storing, analyzing, and managing huge data sets creates major challenges for researchers in disciplines that span the physical, biological, and social sciences and many areas of the humanities. Big data and data science are increasingly necessary for both basic and applied science especially for forming, testing, and understanding hypothesis about complex systems.

Meeting the challenges of data science is critical to US national competitiveness. This has been recognized by funding agencies, which are increasingly addressing data-intensive science within their strategic planning documents. Funding opportunities are available in many agencies to support investigator-led creation of cyberinfrastructure, tools for data collection and analysis, relevant workforce education and training, metadata standardizations, implementation of FAIR data principles (**F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets), etc. The trend is clear. To remain competitive, WSU must provide support for data science across its education, research, and extension missions in order to remain competitive for national-level funding and to attract, and appropriately train, ambitious students and researchers.

CIRC is uniquely placed to address challenges of data science, and as a strategic unit catalyzes success for individual programs in many research areas and their long-term ability to collaborate with peers at other institutions. CIRC has an objective of helping researchers access heterogeneous computing infrastructure that provides broad access to relevant computing technologies. It also supports WSU researchers developing and optimizing scientific workflows locally in order to scale their analyses to larger, competitively available national resources or to predict costs for more defined applications of cloud computing. The Kamiak cluster further

provides access to compute facilities that extend research between funded projects. Finally, as all scientific domains confront “big data”, CIRC plays an important role in recruiting new faculty researchers—access to local computer facilities and support is critical for data-intensive disciplines. New faculty members do not want to re-invent the computer infrastructure and resources they need. There are many opportunities for CIRC to help increase WSU’s competitiveness in computational literacy and involvement in nationally funded data science programs.

5. Software and Public Database Development. As a land-grant institution, research computing can play a pivotal role not only fundamental and applied science and creative output, but also in the creation of tools that can be used broadly for the public good. The creation of software, new algorithms and methodologies, and the curation of data for the use and learning of others, is an essential aspect of this mission. A few representative project are:

- Genetic databases: Consider that technological innovation over the last two decades has propelled agricultural science into an era of big data-driven discovery. Fueled by these high-tech advances and the critical need to provide food security in a rapidly changing global environment, scientists are now routinely generating and analyzing larger, and ever more complex genomic, genetic, and breeding (GGB) datasets. The value of these rich data sets significantly increases when they are organized, annotated, and integrated with other data, and shared FAIRly through online relational databases with access to easy to use analysis and visualization tools. These databases are built using the resource-efficient, open source, interoperable Tripal platform, that is being adopted and actively developed by many database communities around the world. Through connecting these platforms to HPC, massive amounts of data can be routinely collected, analyzed, curated, integrated and made available to scientists in formats that best meet their diverse needs, as well as to provide web-based analysis tools to facilitate basic, translational and applied research.
- Computational fluid dynamic (CFD) software for additive manufacturing: 3D printing based additive manufacturing (AM) has been implemented in many industries due to significant cost savings in conventional tooling, mass customization, shortened supply chains, and greatly reduced time to market. Despite the growing popularity of AM technologies, there are still issues facing current AM technologies such as expensive hardware cost, low production rate, and geometric and property variations. Physics-based predictive models can enable improved understanding of manufacturing processes, thus making it possible to address the issues. A process modeling tool based on the open-source Gerris software has been developed for an innovative inkjet-based high-speed sintering AM technology at WSU Vancouver. It is a predictive package to link the processing parameters to microstructure of the parts and eventually to mechanical performance.
- Development of novel high-performance computing (HPC) system software: I/O pipelines for large scientific data analysis can be long and complex because they comprise many “stages” of analytics across different layers of the I/O stack of HPC systems. Performance limitations at any I/O layer or stage can cause an I/O bottleneck resulting in longer than expected end-to-end I/O latency. Software-defined Storage Resource Enclaves (SIREN) was developed as a novel data management infrastructure. It can enforce the end-to-end policies that dictate an I/O pipeline’s performance. Results on the Kamiak HPC cluster and the Titan supercomputer demonstrate that SIREN provides performance isolation among workflows while maintaining high system scalability and resource utilization.