Data Analytics, Sensor Data, Simulation and Cancer Research

Joel Saltz, Stony Brook University

Over the past few years, there has been major surge in work involving both multi-scale, mult-resolution imaging and molecular characterization as well as a related surge in cancer modeling efforts.

This white paper addresses two issues: 1) the need for supporting computational pipelines associated with the data gathering, integration and abstraction phases in physical, engineering, biomedical research efforts and 2) description of cancer research and treatment use cases with promise to leverage exascale computing to significantly improve understanding of cancer and to impact cancer treatment.

Exascale acquisition, analysis and integration of sensor data

Highly specific predictive modeling of realistic physical, engineering and biomedical scenarios is increasingly in reach. Well known examples include the need to predict detailed behavior of aircraft and space vehicles, . properties of manufactured objects in various use or stress scenarios, detailed dynamics of flame propagation and combustion in realistic scenarios and behavior of confined plasma in fusion reactors. Predictive modeling in biomedicine is increasingly prominent as will be described below. Much focus has been given to computational challenges, somewhat less to in situ analyses of data generated by simulations and comparatively little to requirements associated with processing, analysis, inference, integration and summarization of data obtained from sensors, imaging systems and cameras.

Exascale acquisition, analysis and integration of sensor data requires many of the same optimizations needed to optimize exascale computations – the need to maintain locality, drastically reduce data movement and communication to reduce power consumption and the need to tolerate hardware errors. There are some important differences as well. Raw data is generated by sensors, imaging systems or cameras hopefully with high bandwidth connectivity to the exascale computing platform. There are substantial opportunities for easy wins arising from recognizing when data can be discarded or summarized while in transit. This has the potential to avoid time and power consuming overheads needed to first move data to the exascale platform prior to carrying what are often inexpensive computations needed to discard or summarize data. High end experimental data analytics generally requires integration of information from multiple data sources – data integration can involve a variety of spatial joins carried out on combinations of spatially distributed raster type data and on descriptions of objects recognized and extracted by image analysis and feature extraction algorithms. Region templates are a middleware infrastructure designed by our group to optimize spatial locality, work and data distribution associated with computations carried out on multiple multi-resolution datasets.

In practice, simulation and data acquisition need to be coupled, for instance we see the promise of linking high-fidelity, coupled simulations with high bandwidth streaming sensor data to characterize and detect precursors of instabilities and disruptions in fusion tokamaks and to adaptively adjust magnetic fields to maintain fusion and avoid disruptions. In other cases, image acquisition needs to be steered by computation to control the choice of spatial region, resolution, wavelengths employed, intensity to be imaged as well as to control key components of image processing pipelines. An excellent example of this can be seen in DOE light sources e.g. National

Synchrotron Light Source II;, the ORNL Spallation Neutron Source; same principle applies to biomedical imaging devices and cancer radiation treatment systems. These scenarios fall into the application scenarios that have been called Dynamic Data-Driven Applications (DDDAS). A dynamic data-driven application system is "the integration of a simulation with dynamically assimilated data, multiscale modeling, computation, and a two way interaction between the model execution and the data acquisition methods". In the context of a recently awarded National Cancer Institute award and internal Stony Brook funding, a team from Stony Brook, Oak Ridge and Universidade de Brasília are developing an integrated system that includes the ADIOS system and Region Templates to target such applications.

Leveraging Exascale Computing in Cancer Research and Treatment

Studies of tumor initiation, development, heterogeneity, invasion, and metastasis all require detailed multi-resolution data that elucidates interplay between morphology and spatially mapped genetics and molecular data. In addition, there is a strong effort from the National Cancer Institute to develop applied hybrid multiscale imaging/"omic" methods able to predict outcome and response to treatment and to help steer treatment options.

Tumors are heterogeneous, highly complex spatio-temporal molecular systems (Hallmarks of Cancer below). In any given tumor there are many different categories of tumor cell, these tumor cells play varying roles in the development and spread of cancer. Non-cancer cells in an organ also play critical roles in the containment, spread or metastasis of cancer. While the precise number of cells involved in a tumor and surrounding region obviously varies from case to case, a tumor and embedded stroma may contain somewhere between 10^9 and 10^{12} cells arranged in multiple scale anatomic structures (e.g. glands, ducts, crypts, nephrons, blood vessels etc). In our view, biomedicine will move increasingly towards methods that detect, classify, characterize and spatially locate every cell and every multi-scale structure in a cancer specimen. The data management and computational challenges needed to deconstruct a tumor into its spatial/molecularly labeled elementary cellular components is massive; to date our group has leveraged Petascale architectures to tackle very limited special cases of ambitious problem and in our view, exascale resources will be required to support pipelines able to carry out reproducible multi-scale analysis. In addition, it seems likely that uncertainty quantification combined with a DDDAS approach to interactively steering sensor/image data acquisition will also be required.

In the post-exascale future, we also see an increasing effort to integrate cancer simulation methods into the above pipelines. In our view, meaningful cancer simulation requires precise high quality, highly specific data of the kind that will be generated by the above described cancer characterization pipelines. This will require development of multiscale analytical methods of the biology of cancer. This will include the microscopic (molecular/genetic), mesoscopic (cellular), and macroscopic (tumor/tissue) levels. We will also need to formulate techniques to relate and consistently link the models at the various scales. Indeed, models at higher scales should be derived as averages of models at lower scales. Microscopic (molecular) and mesoscopic (cellular) should be linked since genetics will drive the cellular behavior. Further mesoscopic (cellular) behaviors should drive the macroscopic tissue/tumor level behaviors.

Hallmarks of Cancer

In their seminal work, "Hallmarks of Cancer," Hanahan and Weinberg categorize the major changes in cells and tissue that characterize cancer; see Figure 1. The changes are: self-sufficiency in growth signals, insensitivity to anti-growth signals, tissue invasion and metastasis, limitless replicative potential, sustained generation of new blood vessels (angiogenesis), and evasion of the programmed cell death that normally

occurs in damaged or abnormal cells (apoptosis) . The authors emphasize that these “hallmarks” involve certain genetic mutations and a selection process (closely connected to the Darwinian model of evolution) leading to a malignant neoplasm. Most of the available biological evidence points to the fact that the onset of malignant cancer is preceded by a sequence of genetic mutations in various tissue sub-populations coupled with evolutionary selection.

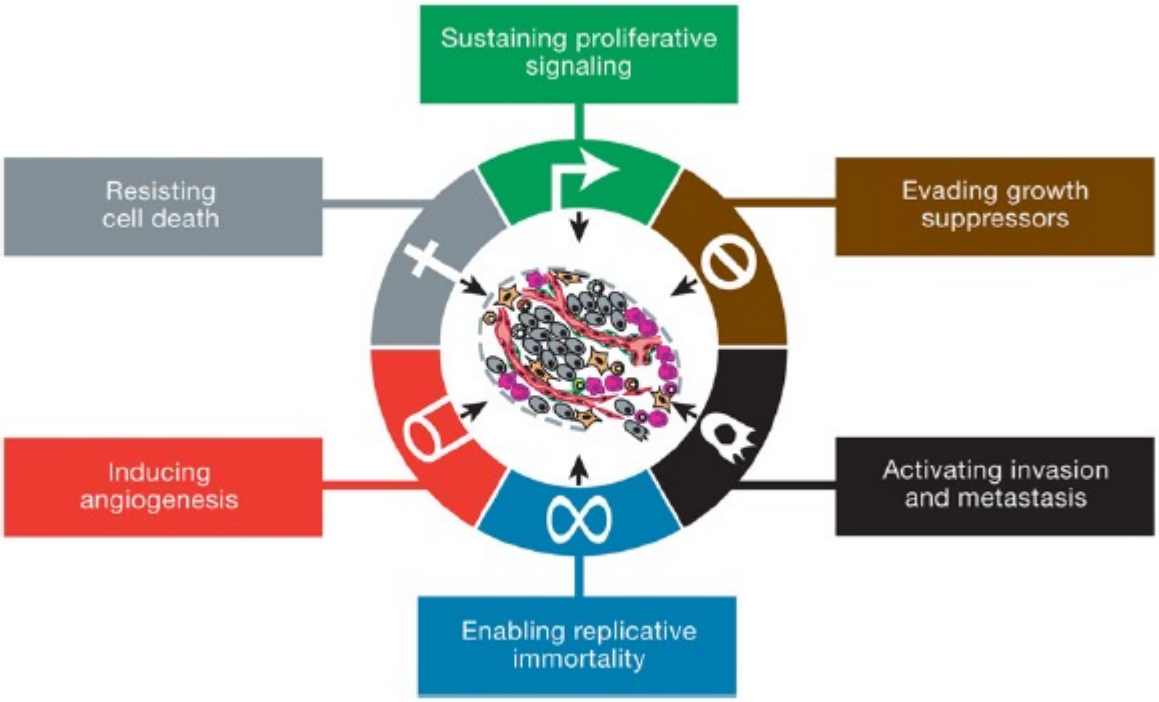

Figure 1: Six hallmarks of cancer taken from Hanahan-Weinberg, “Hallmarks of cancer: the next generation,” *Cell*, volume 144, 2011, pages 646-674.