|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Computational Bioimaging | |
| **Vertical (area)** | | Scientific Research: Biological Science | |
| **Author/Company/Email** | | David Skinner**1**, [deskinner@lbl.gov](mailto:deskinner@lbl.gov)  Joaquin Correa**1**, [JoaquinCorrea@lbl.gov](mailto:JoaquinCorrea@lbl.gov)  Daniela Ushizima**2**, [dushizima@lbl.gov](mailto:dushizima@lbl.gov)  Joerg Meyer**2**, [joergmeyer@lbl.gov](mailto:joergmeyer@lbl.gov)  **1**National Energy Scientific Computing Center (NERSC), Lawrence Berkeley National Laboratory, USA  **2**Computational Research Division, Lawrence Berkeley National Laboratory, USA | |
| **Actors/Stakeholders and their roles and responsibilities** | | Capability providers: Bioimaging instrument operators, microscope developers, imaging facilities, applied mathematicians, and data stewards.  User Community: DOE, industry and academic researchers seeking to collaboratively build models from imaging data. | |
| **Goals** | | Data delivered from bioimaging is increasingly automated, higher resolution, and multi-modal. This has created a data analysis bottleneck that, if resolved, can advance the biosciences discovery through Big Data techniques. Our goal is to solve that bottleneck with extreme scale computing.  Meeting that goal will require more than computing. It will require building communities around data resources and providing advanced algorithms for massive image analysis. High-performance computational solutions can be harnessed by community-focused science gateways to guide the application of massive data analysis toward massive imaging data sets. Workflow components include data acquisition, storage, enhancement, minimizing noise, segmentation of regions of interest, crowd-based selection and extraction of features, and object classification, and organization, and search. | |
| **Use Case Description** | | Web-based one-stop-shop for high performance, high throughput image processing for producers and consumers of models built on bio-imaging data. | |
| **Current**  **Solutions** | **Compute(System)** | | Hopper.nersc.gov (150K cores) |
| **Storage** | | Database and image collections |
| **Networking** | | 10Gb, could use 100Gb and advanced networking (SDN) |
| **Software** | | ImageJ, OMERO, VolRover, advanced segmentation and feature detection methods from applied math researchers |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Distributed experimental sources of bioimages (instruments). Scheduled high volume flows from automated high-resolution optical and electron microscopes. |
| **Volume (size)** | | Growing very fast. Scalable key-value and object store databases needed. In-database processing and analytics. 50TB here now, but currently over a petabyte overall. A single scan on emerging machines is 32TB |
| **Velocity**  **(e.g. real time)** | | High-throughput computing (HTC), responsive analysis |
| **Variety**  **(multiple datasets, mashup)** | | Multi-modal imaging essentially must mash-up disparate channels of data with attention to registration and dataset formats. |
| **Variability (rate of change)** | | Biological samples are highly variable and their analysis workflows must cope with wide variation. |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues, semantics)** | | Data is messy overall as is training classifiers. |
| **Visualization** | | Heavy use of 3D structural models. |
| **Data Quality (syntax)** | |  |
| **Data Types** | | Imaging file formats |
| **Data Analytics** | | Machine learning (SVM and RF) for classification and recommendation services. |
| **Big Data Specific Challenges (Gaps)** | | HTC at scale for simulation science. Flexible data methods at scale for messy data. Machine learning and knowledge systems that drive pixel based data toward biological objects and models. | |
| **Big Data Specific Challenges in Mobility** | |  | |
| **Security & Privacy**  **Requirements** | |  | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | There is potential in generalizing concepts of search in the context of bioimaging. | |
| **More Information (URLs)** | |  | |
| **Note:** <additional comments> | | | |