Group 14- Computational infrastructure necessary for future ML applications and data/metadata storage

Hervé Goeau
Nicky Nicholson
Gil Nelson
Nadya Williams

1. Collect user scenarios and establish what resources will be needed now and plan for approx 5 years from now. Based on the outcome propose an infrastructure that is scalable, reproducible, extensible and dynamic.
2. Who is the images aggregator? We may need to download some data from aggregators
    a. Get images to one place from aggregators and make sure it is replicated using Ceph or similar approach. The images must be available 24/7
    b. Plan to provide an ability to upload sets of images from users and to download results of computations. The extra sets and results require additional disk space capacity (may be a short term, some data may be long term if become public.
    c. Ability to share user-level data: (for example, annotations, training data sets) and provide a level of quality control or verification for such data that become “public”. Need to keep track of the metadata that provides information on how the data sets or results were obtained (software, method, algorithm, verison of all software involved, what control or testing was doen. etc)
3. Library of Virtual Machine images and containers with models for ML and other processing units:
    a. Easy to deploy
    b. Known software stack
    c. Easy to run on local (for testing on a small data set) or large (data centers, HPC) infrastructure

    The images must have a metadata associated with them which will provide information about how these images were built and what software stack they deploy.
4. Extensible and distributed storage
    a. Hard disk and SSD disks
    b. Replication of data (multisite), for example using Ceph
    c. Ease of adding new servers with additional disks.
5. Fast access to data and the ability to bring computations to data.
6. Access to heterogeneous GPU (low, high memory)
    a. Emphasize that GPU firepower key; different from a lot of current bioinformatic foci
    b. With local access to hard drives and SSDs
    c. Establish memory size and efficient moving of data from disk to memory
7. “Consolidating” multiple site resources into a hyper-cluster - take a look at the PRP (Pacific Research Platform) that built a hyper-cluster from the distributed hardware and

amassed knowledge how to use Kubernetes, Ceph, GPU. The website for the project https://ucsd-prp.gitlab.io/nautilus/namespaces/ (currently under development because of switching the website CMS). This link https://ucsd-prp.gitlab.io/nautilus/namespaces/ gives ideas what ML and AI projects people are running using this hypercluster. The properties of the system to support Phenology Machine Learning:

a. Includes nodes (servers) that provide access to the data and "compute" nodes that run jobs (a job is any part of the workflow, or a tool that runs some computations on data)
b. Fast access to data and the ability to bring computations to data
c. Access to fast networks
d. User-friendly Graphical interface (portal) for logging in. May be will need multiple portals for different workflows or workflow parts.
e. Documentation and testing of tools as soon as they become available. Deliver tools to the users in a way that is easy to use and adopt.
f. Accommodate changes in computational capacity (increase or decrease on demand based on usage)