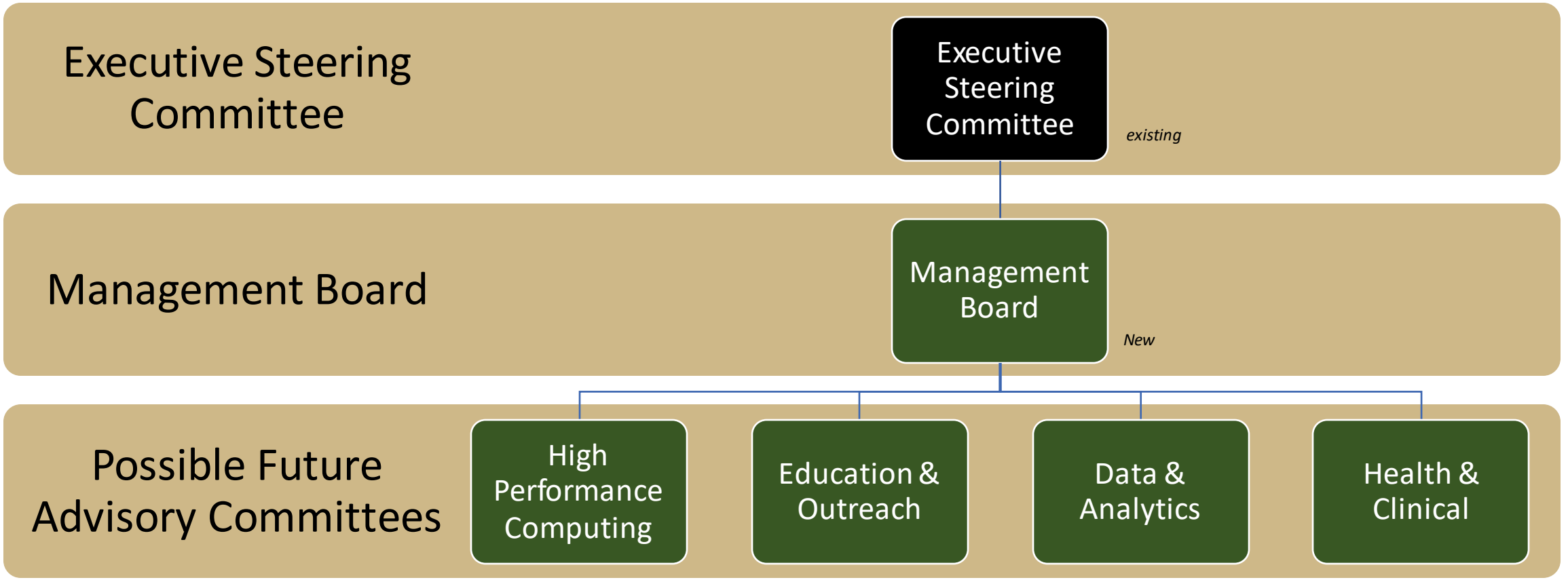
RCC Town Hall

Fall 2021

New RCC Governance Model

Bobby Sprinkle

New Governance Model - Research Computing



EXECUTIVE STEERING COMMITTEE MEMBERSHIP

- Laurel Fulkerson, Interim Vice President for Research
- Murray Gibson, Dean of the College of Engineering
- Sam Huckaba, Dean of the College of Arts and Sciences
- Rick Burnette, Interim CIO
- Bobby Sprinkle, Chief Technology Officer
- Jay Terry, Provost Representative
- Scott Stagg, Management Board Chair

MANAGEMENT BOARD CHARTER

- - Identify and approve infrastructure, staff, and resource allocation in RCC for support of research technology and consulting services.
- - Recommend 3rd party software licenses.
- - Recommend subsidies and the apportionment of costs for users.
- - Approve usage and scheduling policies for the cluster.
- - Provide requirements on usage reporting and user interfaces.
- - Incentivize researchers to use appropriate data storage and processing facilities.

MANAGEMENT BOARD CHARTER MEMBERSHIP

- Health and Life Sciences – 2 Members
- Physical Sciences – 2 Members
- Engineering – 2 Members
- Computational and Data Science – 2 Members
- Social Sciences and Humanities – 2 Member
- Director of RCC (ex officio)
- CTO for FSU ITS (ex officio)
- Library Representative (ex officio)

Management Board Current Membership

Faculty

- Scott Stagg – Chair
- Eric Chassignet
- Harrison Prosper
- Tarez Graban
- Sungmoon Jung
- Michelle Arbeitman
- Adrian Barbu
- Zhe He
- William Oates
- Yaacov Petscher

Ex Officio

- Bobby Sprinkle (CTO)
- Jean Phillips (Library Representative)
- Paul Van Der Mark (RCC)

Questions?

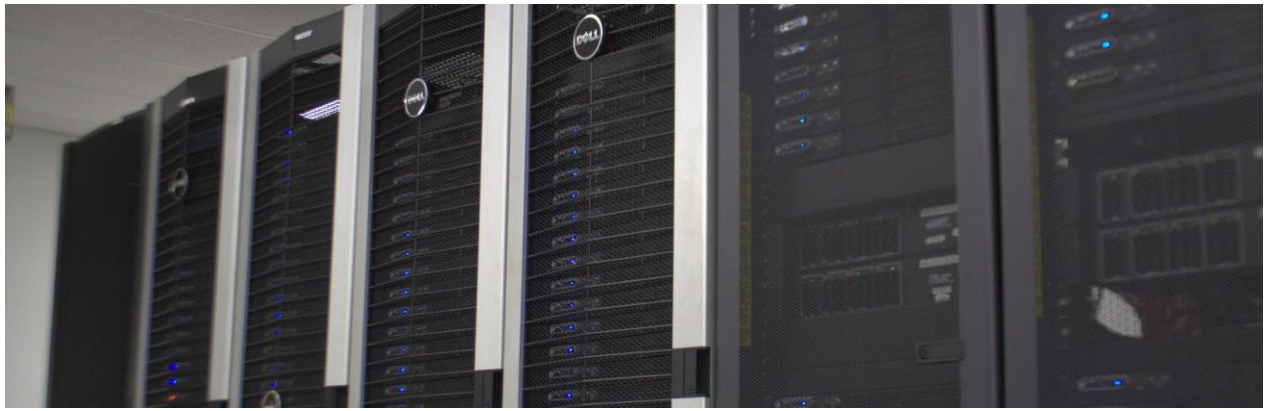
RCC History and Access

Paul Van der Mark

RCC History

- 1984: SCRI, CDC Cyber 205, ETA10, Cray
- 2007: Shared HPC facility (in academic dept)
  - Annual Buy-in, minimum of $25K, variable match
  - Started with 3 departments
  - Services limited to HPC and storage
- 2012: RCC (in university wide unit)
  - Auxilary model, fixed matching (50% on HPC)
  - New services consulting, archival, server colocation (Sliger)
- 2022: Change subsidy allocation, broaden user base.

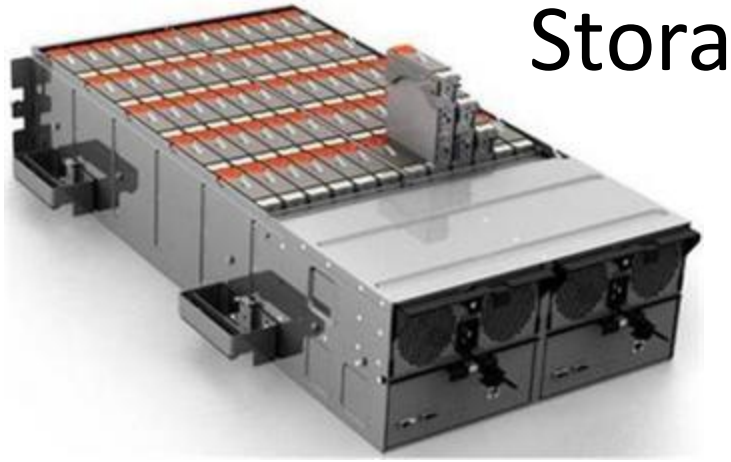
RCC Services



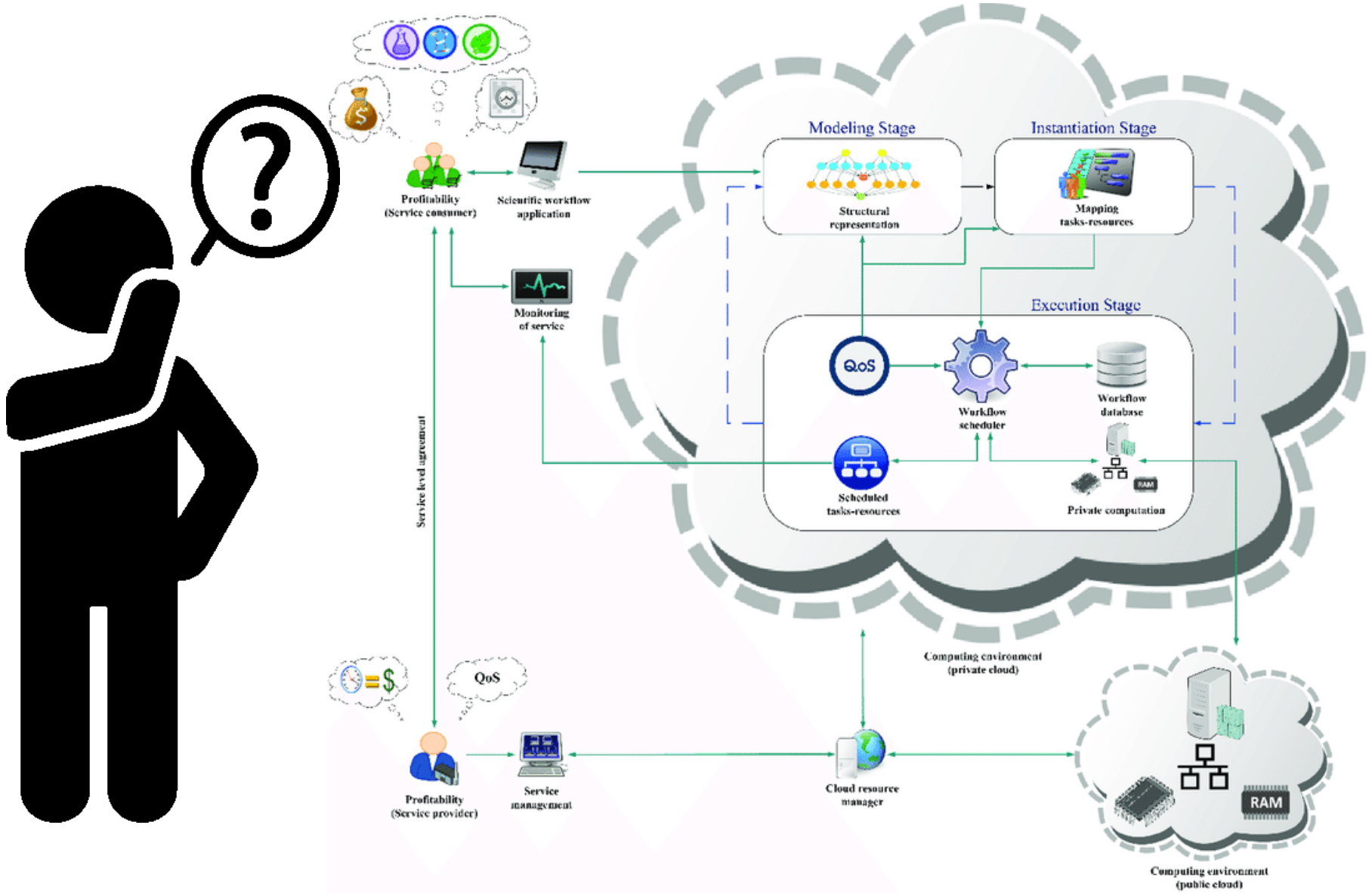
Compute

GPU

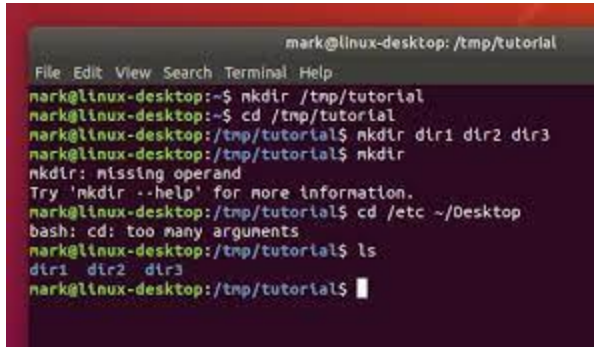
Storage



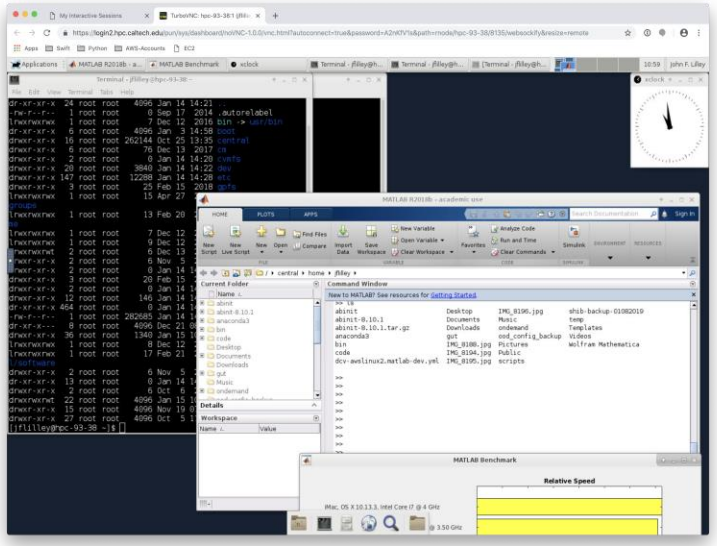
RCC Consulting



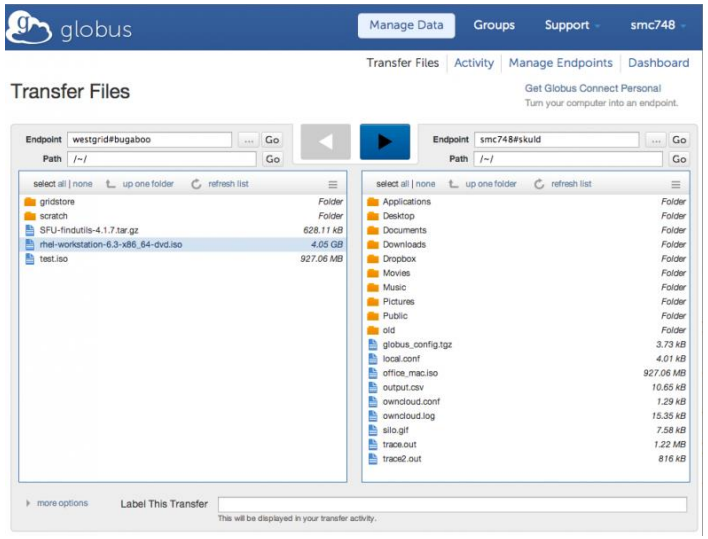
Access RCC Services



Command Line



Virtual Desktop



Fast, secure, and reliable data transfer (globus)

RCC Funding Model

- Redistribute subsidy and new rate calculation
  - More emphasis on storage: lower rates with 50%
  - NCU prices lower and 7 years instead of 5.
  - disclaimer: Budget Office and Sponsored Research
- Allocate part of subsidy to free services
  - Currently 30TB sccratch & free older genacc cores
  - New:
    - 350 new cores to general access partitions
    - 400TB storage pool
      - Annual application for part of that pool
      - Lightweight application

Flexible Compute Servers

- Base node configuration
  - Fast CPUs (Xeon or AMD server-type)
  - 128GB Ram
  - 1 or 10GbE network
  - RCC software stack (linux)
- Upgrades: Infiniband, GPUs, Memory, SSD
- Infrastructure is free, server paid by PI
- Charge pass-through rates or **purchase in spearmart**
- Work with vendor on “better” rates

RCC Pilot Program (3y)

- Goal: assist non-traditional RCC users and non-funded (development) researchers
- 1 FTE digital humanities support position
  - Collaborate with library
- Dedicated hardware
  - $1M over 3 years
- Paid for by RCC auxilary cash reserve

User Cases

Traditional Usage

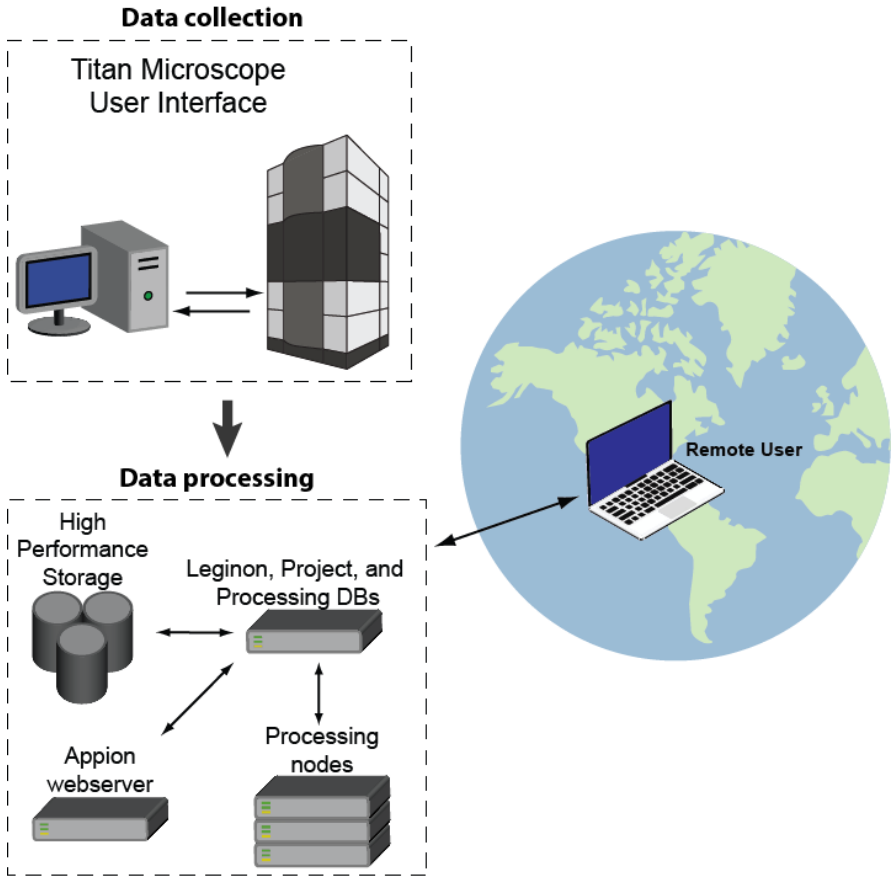
Scott Stagg

High performance computing

- COAPS
  - Simulate ocean currents
  - 360 cores current + 300 more to come
  - Some jobs run on hundreds of cores at once for 6 to 12 months at a time
- 720 Jobs submitted per month
- 103,451 CPU hours computed per month

Storage and processing

- Cryo-EM @ FSU
  - Thousands of images per day
  - Automated data collection, data assessment tools
  - Web-based monitoring of data collection
  - Dozens of users from around the Southeast
- Archival system is used for short term secure data housing and transfer
- Scalable
  - Supports the Southeastern Consortium for Microscopy of Macromolecular Machines (SECM[4])



RCC and Genomics at FSU

Michelle Arbeitman
Biomedical Sciences Department
College of Medicine
Florida State University
Michelle.arbeitman@med.fsu.edu

RCC usage for Genomics

- NovaSeq Illumina Sequencing at FSU is performed by the Translational Sciences Lab core in the College of Medicine. Data are provided to the user on the RCC.
- Depending on the workflow you will perform, you will likely continue to use the RCC for your analysis.
- Often users will map Illumina Sequencing Reads to a reference genome using the RCC. Common use applications that require read mapping are RNA-seq for gene expression studies, and ChIP-seq for protein binding in the genome.
- The RCC is also used for genome assembly and large-scale comparisons across fully sequenced genomes.
- Many additional applications—new genomics workflows are developed constantly.

RCC usage for Genomics

- The RCC also provides workshops on using different bioinformatics tools. The RCC recently hosted the group called “Data Carpentries” that did a workshop on implementing computational genomic approaches.
- The RCC has installed some standard algorithms/analysis software used in genomics.

- Feedback on how the RCC can better serve genomics users is appreciated. For example, packages for usage across users that takes into account that the typical user works in bursts on the RCC.

RCC and Communications Use Cases

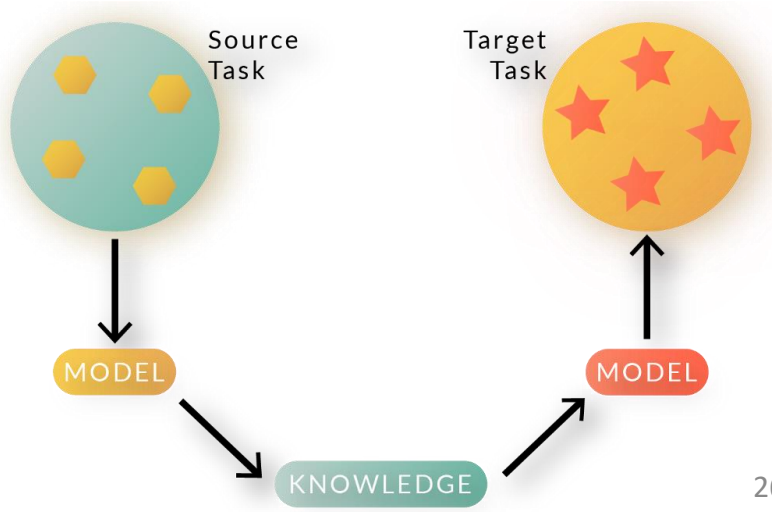
Zhe He
MSIT Program Chair
School of Information
Florida State University

RCC Use Case 1

- CORD-19 Open Research Data Challenge
  - CORD-19 is a resource of over 500,000 scholarly articles, including over 200,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease.
- Tasks:
  - What do we know about COVID-19 risk factors?
  - What do we know about vaccines and therapeutics?
  - What do we know about transmission, incubation, and environmental stability?

RCC Use Case 2

- Transfer learning with pre-trained language models for various applications
  - OpenAI's GPT-3: 175 billion parameters
  - Google's BERT
    - Devlin and his colleagues trained the BERT on English Wikipedia (2.5B words) and BooksCorpus (0.8B words) and achieved the best accuracies for some of the NLP tasks in 2018.
- Further pretrain BERT with domain-specific text corpora
- Downstream NLP Applications
  - Question & Answering
  - Text classification
  - Named entity recognition
  - Relation extraction



Aspirational Use Cases for the Digital Humanities (DH)

Sarah Stanley
University Libraries - Office of Digital Research and Scholarship
scstanley@fsu.edu

Text Analysis Support

Large Scale (or medium scale) Text Analysis

- Libraries provide limited support for text analysis, but faculty often need more
- HathiTrust Data
  - rsync for acquiring large (or medium-sized!) datasets
  - Ability to build corpora using specific queries of textual metadata
  - Knowledge of data formats for non-consumptive text mining
  - When necessary, the ability to conduct analysis on large (think millions of words) textual corpora

Humanities Data

Database Design / Ontologies / Cultural Heritage Data

- Creating humanities data often involves modeling complex relationships
  - Humanities training involves *theorizing* these relationships, but training rarely extends to modeling them as data
- Using or combining existing ontologies + creating new ones
- Putting faculty-created data in context with existing cultural heritage/library/humanities data
- Expertise in SPARQL/SQL for querying existing humanities data

Question and Answer Session