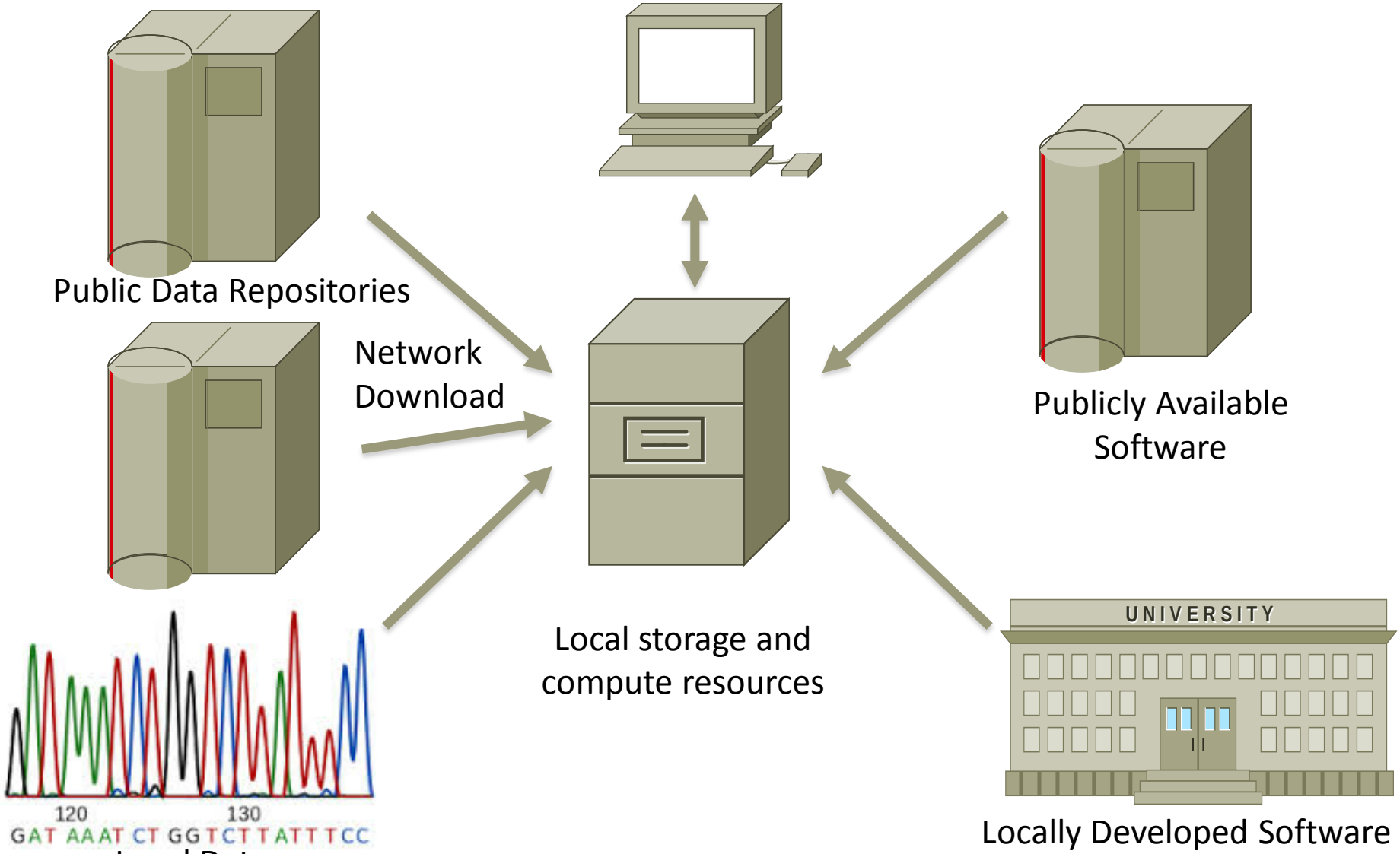
National Cancer Institute

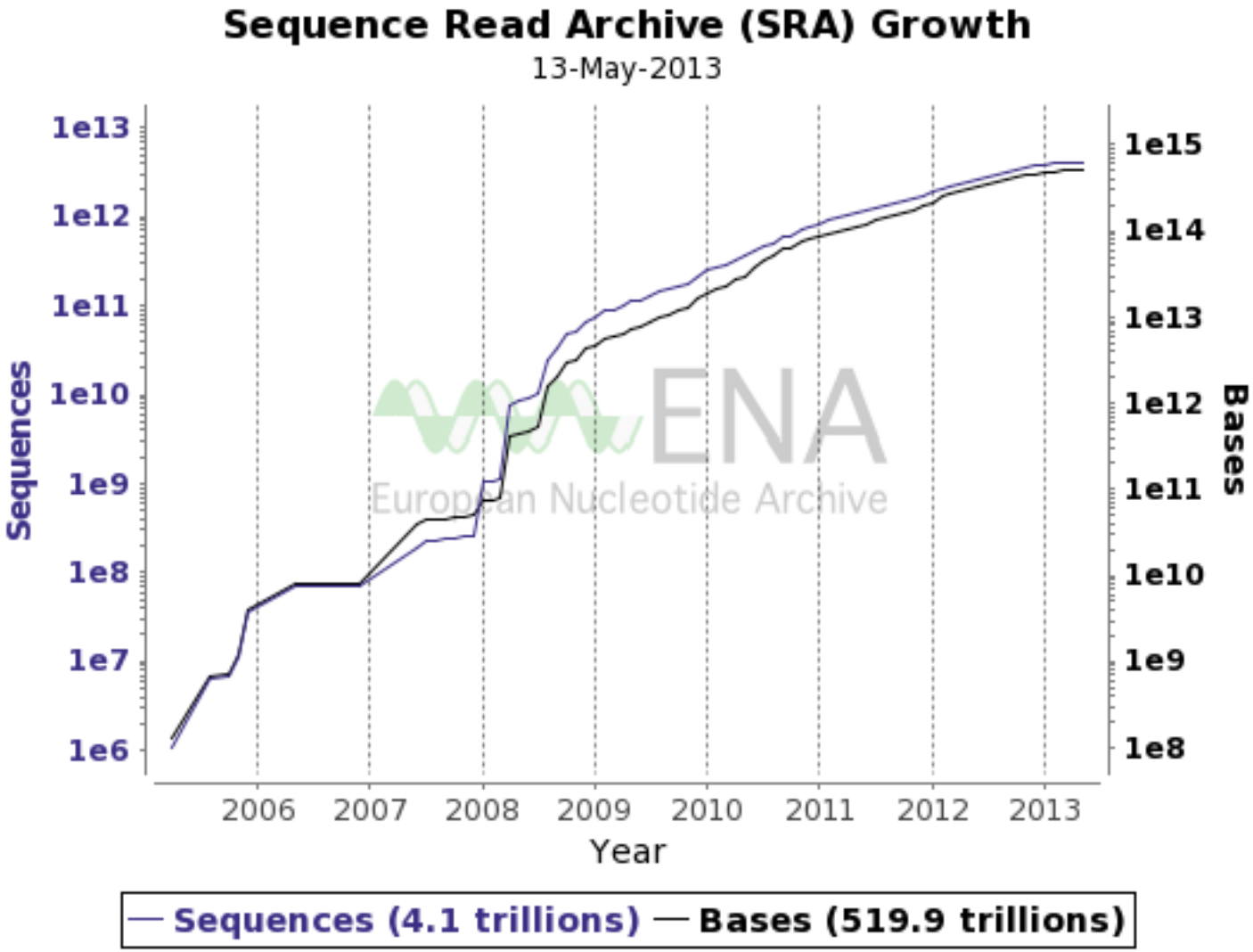
Cancer Genomics Cloud Pilots Concept Briefing to the NCAB/BSA

George Komatsoulis, Ph.D,
Director (interim)
Center for Biomedical Informatics and Information Technology

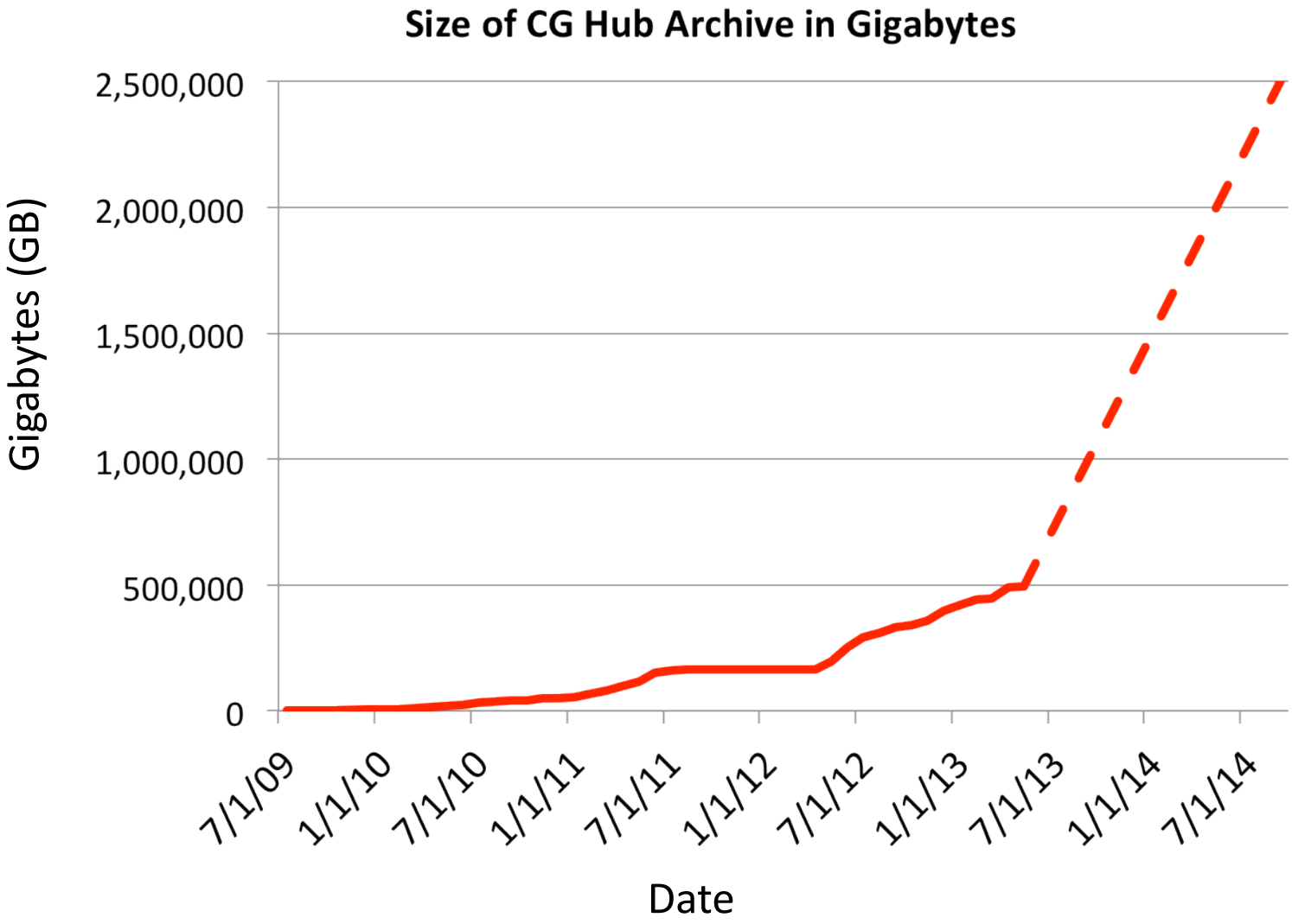
Standard Model of Computational Analysis



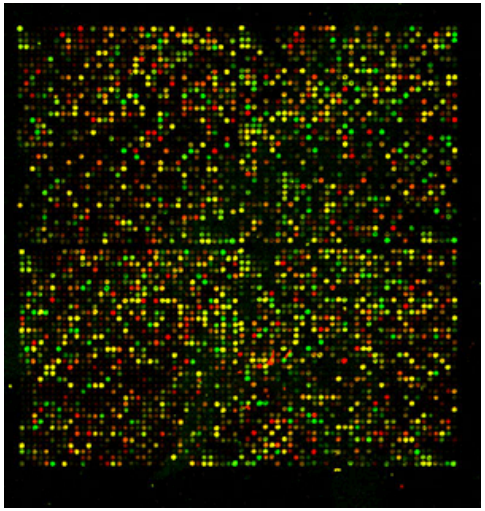
Growth of Sequence Data

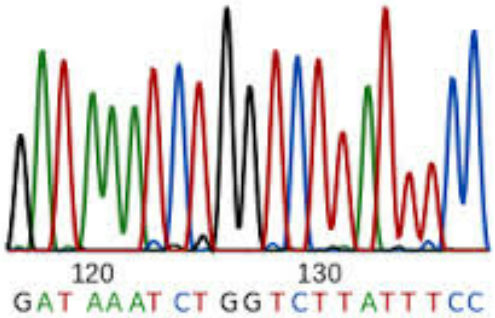


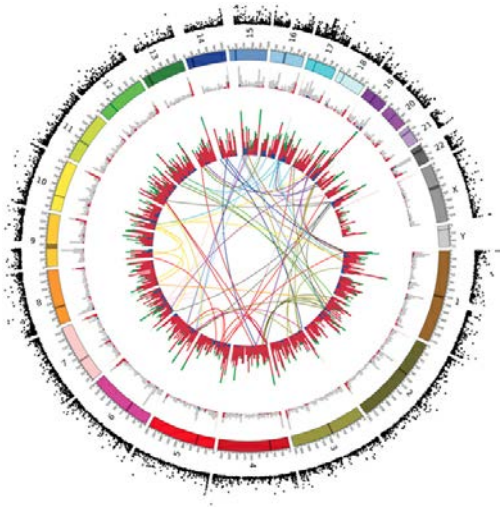
Growth of Sequence Data (CG Hub)

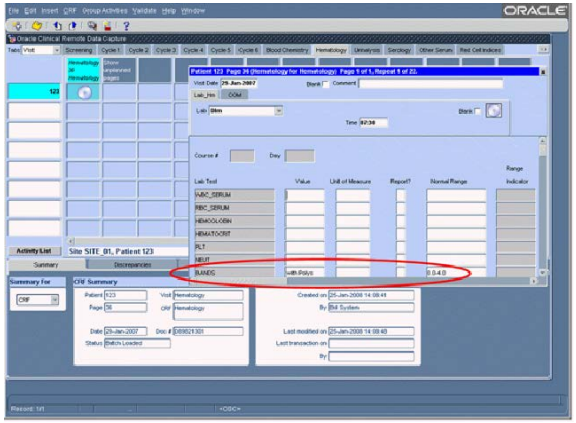


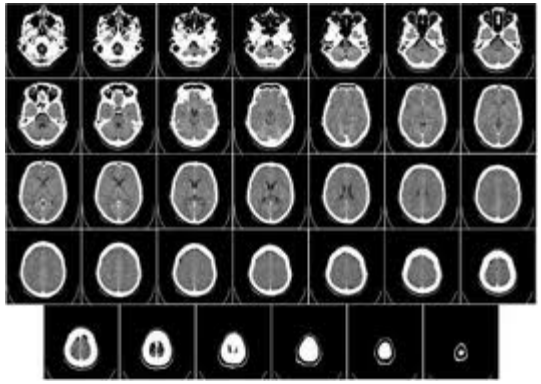
Multiple orthogonal data types











Limitations of the standard model for large data sets

- Assuming the 2.5 PB TCGA data set
- Storage and Data Protection cost approximately $2,000,000 per year.
- Downloading TCGA data at 10 Gb/sec = 23 days
- Size + high dimensionality = high computational requirements that grow quickly

IT limitations in the real world

Dr. Harold Varmus Requests

ncip.nci.nih.gov/nci-cloud-initiative

National Cancer Institute | U.S. National Institutes of Health | www.cancer.gov

NCIP National Cancer Informatics Program

Follow us on: Search

Home | Program Announcements | National Cancer Informatics Program Launch Meeting | Contact Us

You are here: Home → Dr. Harold Varmus Requests Input On NCI Cloud Initiative

Dr. Harold Varmus Requests Input On NCI Cloud Initiative

by janene — last modified 2013-05-20 13:47

To the Cancer Research Community:

The advent of a variety of new research technologies has dramatically increased the rate at which biological data is generated, a rate of change that has strained conventional mechanisms for distributing and analyzing this information. The NCI (and the biomedical community broadly) is investigating next generation computational capabilities to support the biomedical research community. One possibility is the creation of public "cancer knowledge clouds", namely data repositories with co-located computational resources, allowing investigators to bring their analytic tools to the data rather than move the data to their analytic tools. Such clouds have the potential to increase the speed of discovery and democratize access to cancer genomics data, which is too often the province of organizations that can support the high cost of maintaining these enormous data sets. While there is an emerging consensus that such computational resources are going to be essential components of the resource environment in the near future, there is not yet a consensus on the implementation of such an environment. NCI believes that the correct approach is to conduct a series of pilots that will lead to the development of a cancer knowledge cloud or clouds.

In order to ensure that our initial efforts in this area are directed at the practical problems faced by biomedical researchers, we are writing to request your advice about the scientific questions that should guide the process of developing infrastructure to analyze these high volume data sets. Specifically, we are requesting that you provide us with information on:

1. Situations where limitations in information technology capabilities are either preventing you from carrying out certain types of high value research, or where these limitations, while not preventing your research, are slowing the pace of discovery significantly.
2. Your experiences using high performance computing environments for biological research, including custom environments and commercial clouds (private or public).
3. Metrics that could be utilized to determine the level of success of pilot "cancer knowledge clouds".

Your responses (along with those from other scientists and relevant oversight committees) will help to guide the NCI in its efforts to ensure that the research community has the tools it needs. If possible, we would like to receive your comments by May 3, 2013; they can be sent electronically to ncicloud@mail.nih.gov.

Thank you for your prompt attention to this matter.

Harold E. Varmus, M.D.
Director, NCI

George A. Komatsoulis, Ph.D.
Director (interim)
Center for Biomedical Informatics and Information Technology

Louis M. Staudt, M.D., Ph.D.
Acting Co-Director

http://ncip.nci.nih.gov/nci-cloud-initiative

Major Themes of 68 Responses to Letter

- Data access
- Computing capacity
- Data interoperability
- Training
- Usability
- Governance

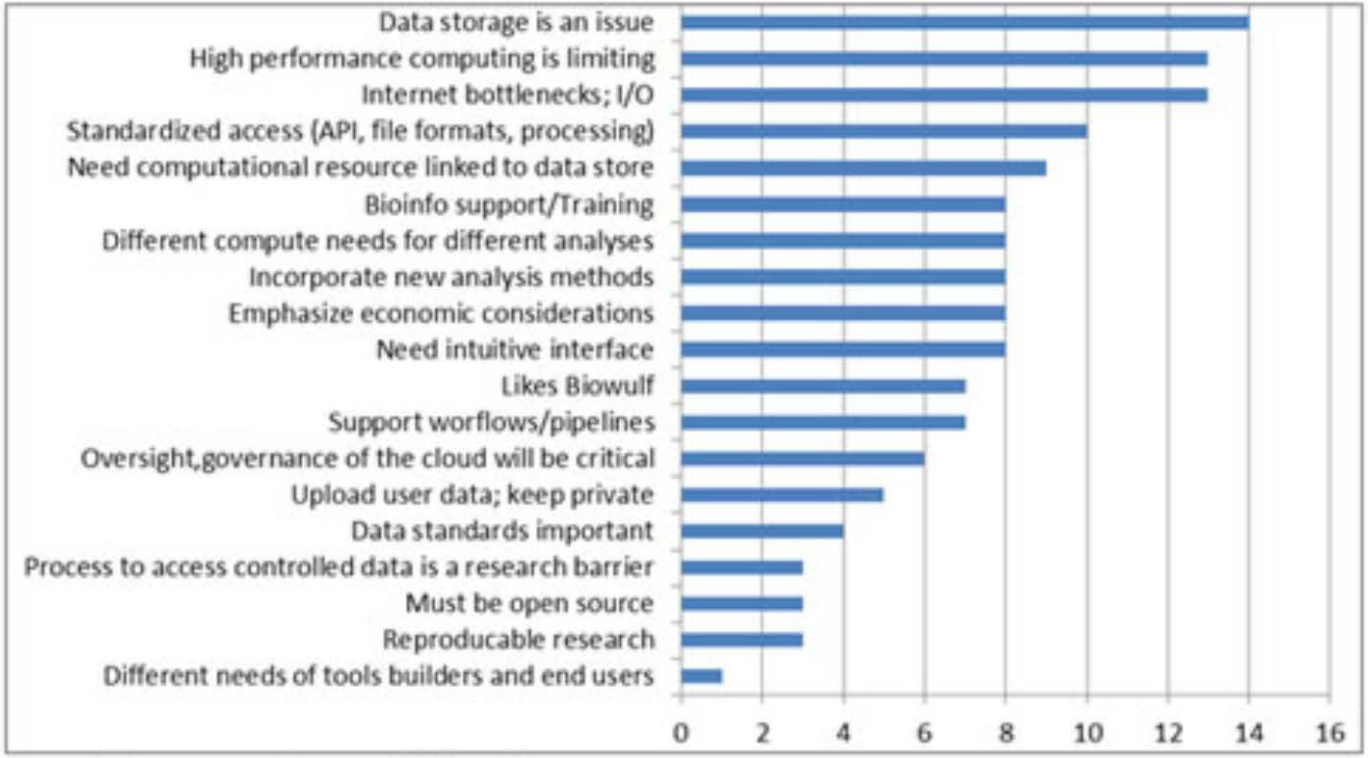
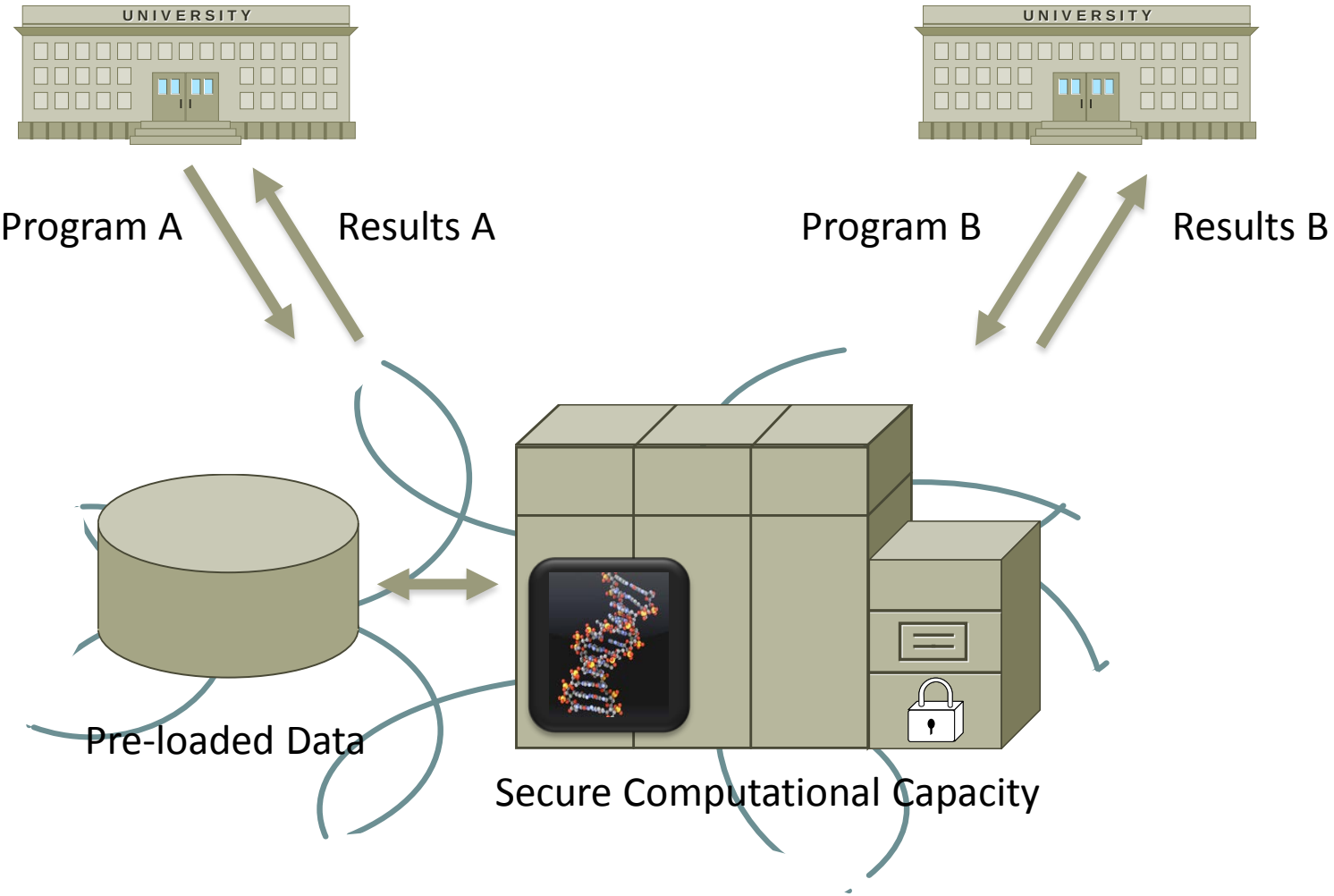
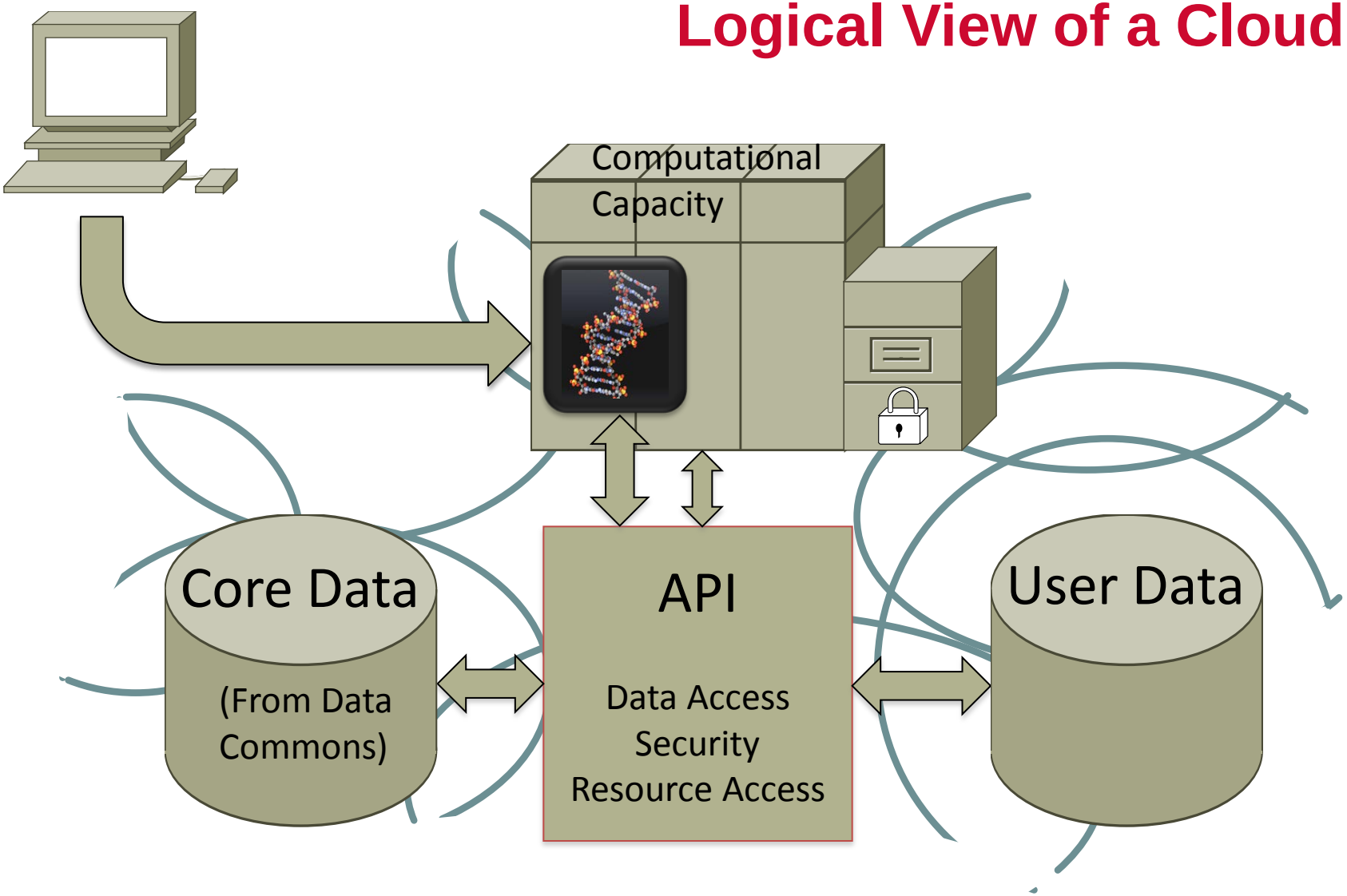


Figure 1: Summary of community input

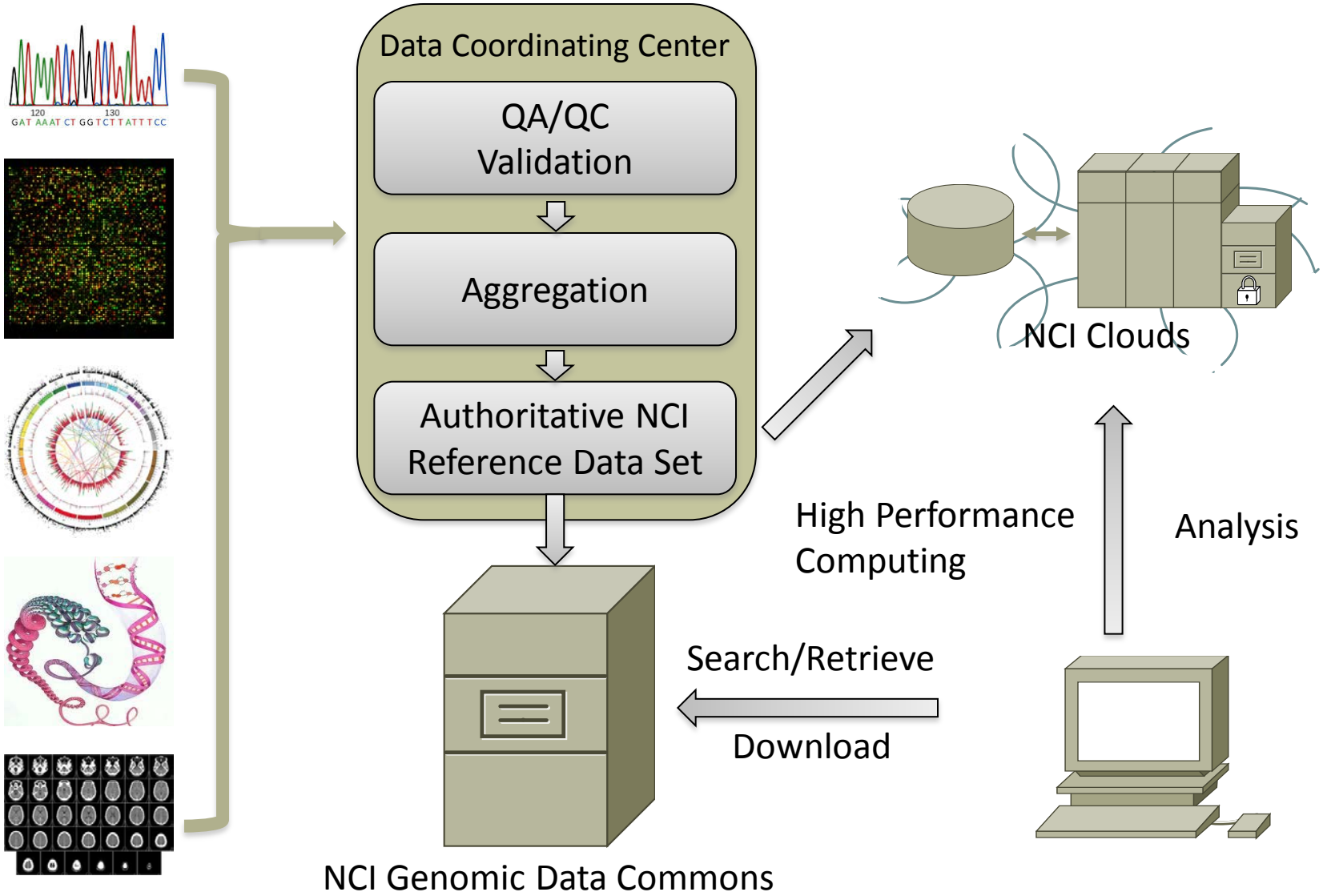
Biomedical Cloud Alternative



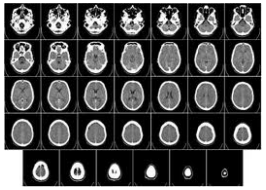
Logical View of a Cloud

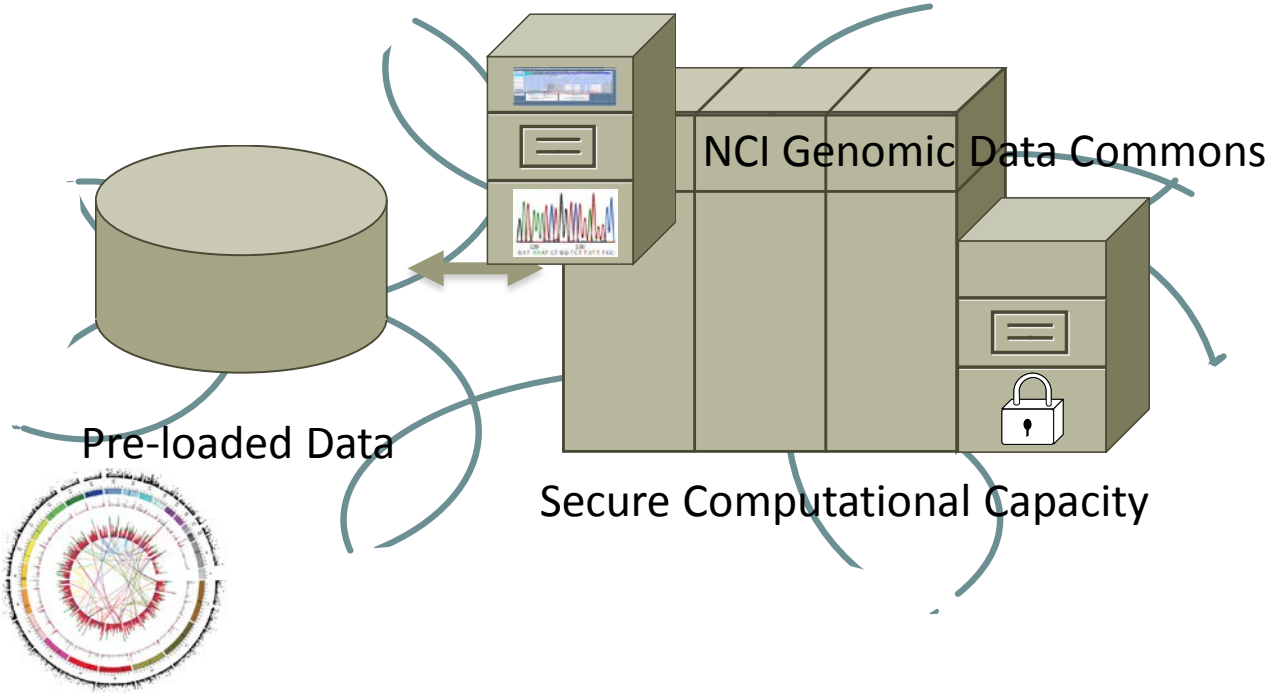


A full cancer data pipeline

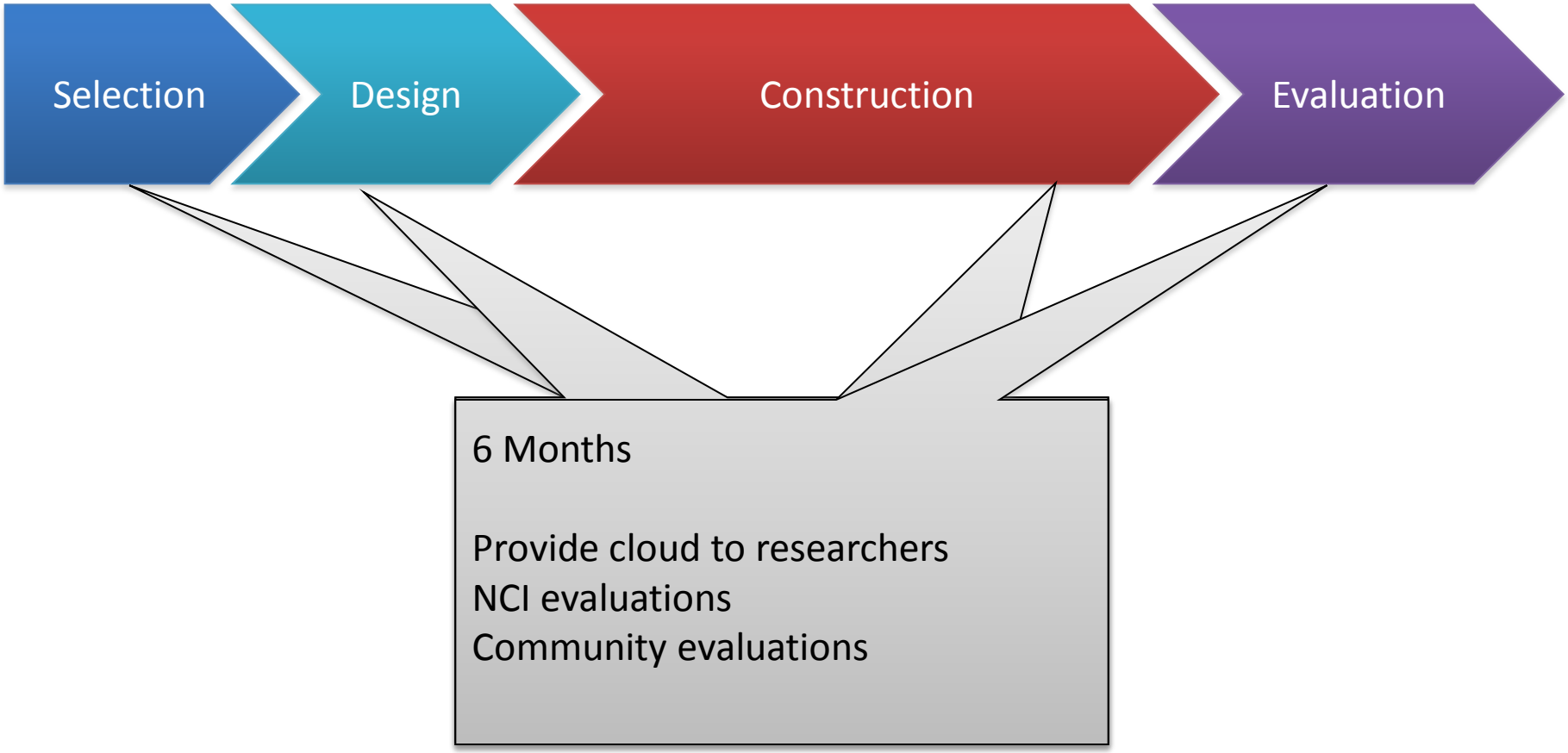


Cloud Pilot Consortium





Project Schedule and Deliverables



Considerations – Design and Intellectual Property

- All possible designs will be considered, whether they use commodity cloud technology, specialized implementations using dedicated, commodity hardware, or specialized hardware.

- Designs must be released under a non-viral, open source license that allows any entity to implement a version of the cloud for either commercial or non-commercial use.

- Commercial products can be used so long as they are available under standard commercial terms and do not inhibit the IP rights needed by the NCI

Considerations: Extensibility and Interoperability

- Initial clouds will focus on a set of “core datatypes”
- Cloud Designs must be capable of extending to additional datatypes without major refactoring of the existing system (extensibility)
- Cloud designs should be interoperable to the maximum degree practical

Considerations: Scalability and Sustainability

- Scalability: Clouds must be designed to support expansion of up to 100 fold in the areas of data size and usage

- Sustainability: Cloud deliverables include a sustainment plan that includes cost assessments for operating at current scale and at 10- and 100-fold increases in storage, compute and usage.

- Operational costs in production modes will vary based on the technical solution
  - Commodity clouds: Low capital/higher operating expenses
  - Dedicated hardware: High capital costs, lower operating costs

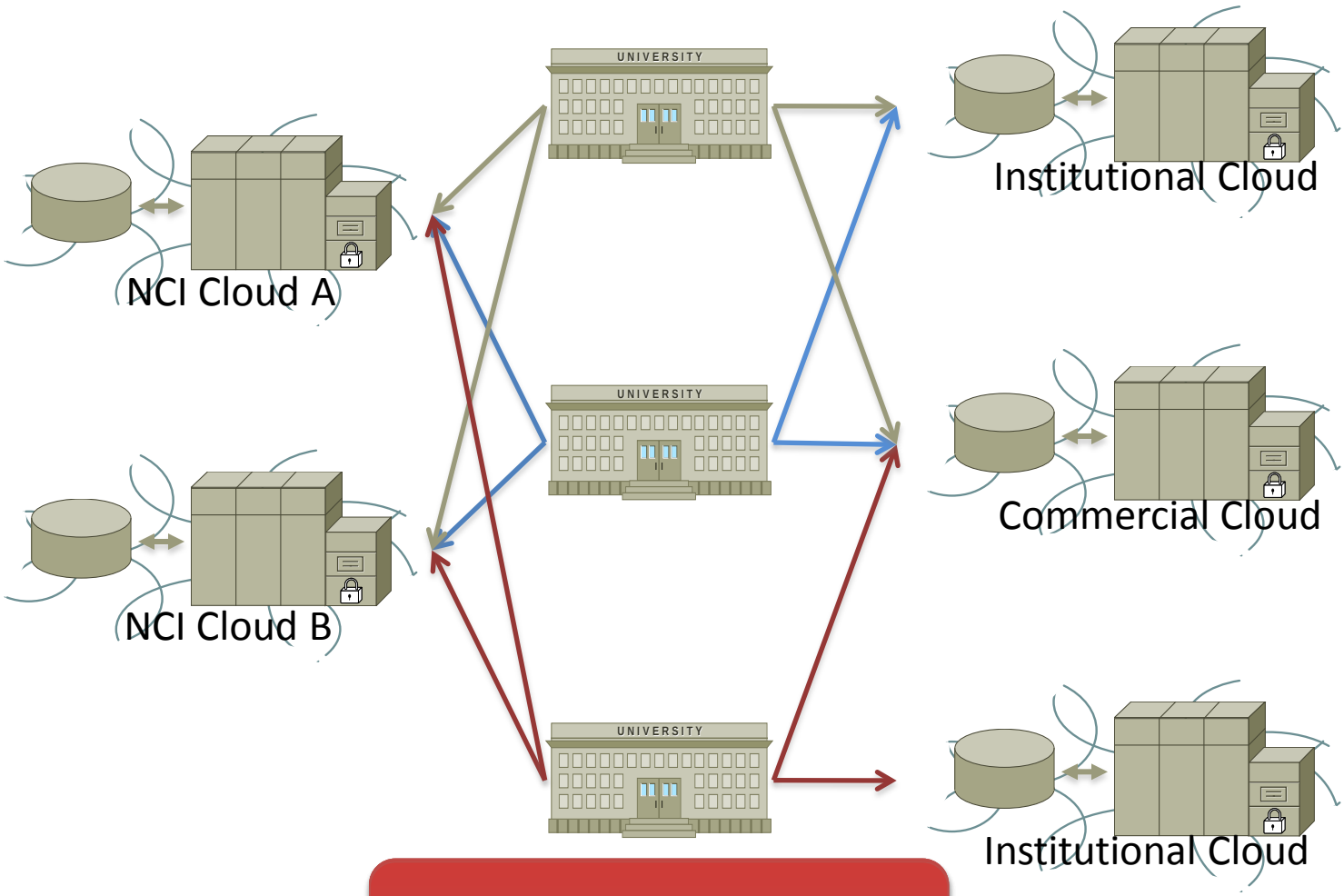
Considerations: Acquisitions and Evaluations

- Acquisition via Broad Agency Announcement
  - Allows for maximum flexibility/creativity
  - Selection utilizes peer review
- NCI testing and evaluation

No expansion of clouds until reviewed by appropriate groups (NCAB *ad hoc* Informatics Working Group, SPL, NCAB/BSA, etc.)

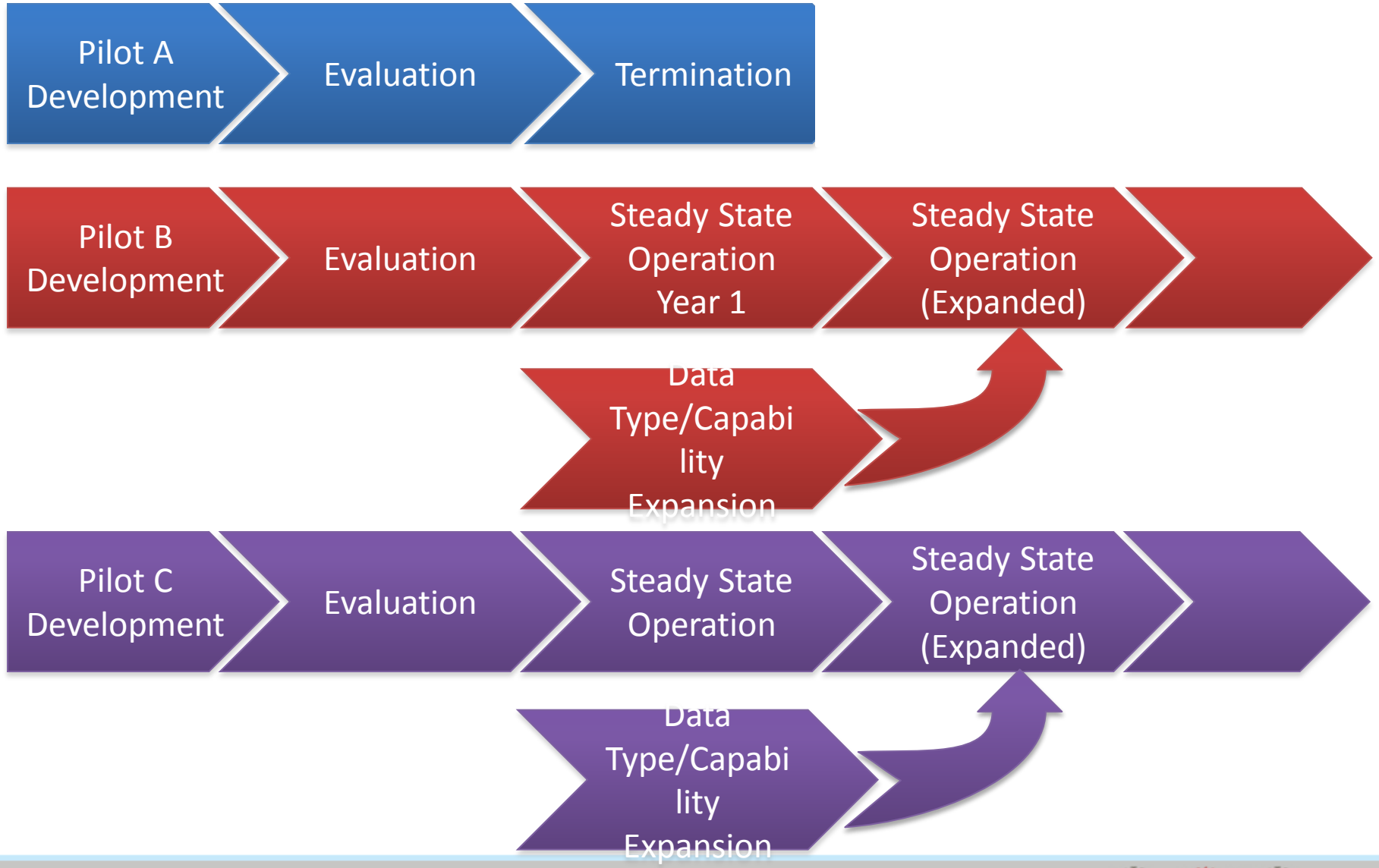
-
  - Which systems did community members like using?
  - Use challenge.gov/TopCoder type contests to incentivize experimentation with the clouds
- Sustainability
  - Evaluate cost and sustainment models for various clouds

Possible Models for Long-Term Support



HYBRID MODEL

Operation and Modernization Strategy



Cost Estimate

- Design and implementation
  - $3,000,000 - $5,000,000 per cloud pilot (FY 14)
- Evaluation period
  - $500,000 per cloud pilot for operations (FY15)
- Operational phases (if successful)
  - $3,000,000 - $5,000,000 per cloud (FY16 and beyond)
- Estimates are widely variable due to the large number of options (commodity clouds vs. dedicated hardware, etc.) inherent in the design of a new capability

Acknowledgments

- Anthony Kerlavage, Ph.D.
  - Chief, Informatics Program Branch, CBIIT
- Daoud Meerzaman, Ph.D.
  - Head, Computational Genomics Section, CBIIT
- Juli Klemm, Ph.D.
  - Head, Cancer Biology and Genomics Section, CBIIT
- Ishwar Chandramouliswaran, MS, MBA
  - Program Manager, CBIIT
- Tanja Davidsen, Ph.D.
  - Program Manager, CBIIT

- Barbara Wold, Ph.D.
  - Professor, Caltech and Acting Director, CCG
- Louis Staudt, MD,Ph.D.
  - Director, CCG
- Stephen Chanock, MD
  - Acting Co-Director, CCG

QUESTIONS