CIPRES Cyberinfrastructure for Phylogenetic Research

The CIPRES Science Gateway: Enabling High-Impact Science for Phylogenetics Researchers with Limited Resources

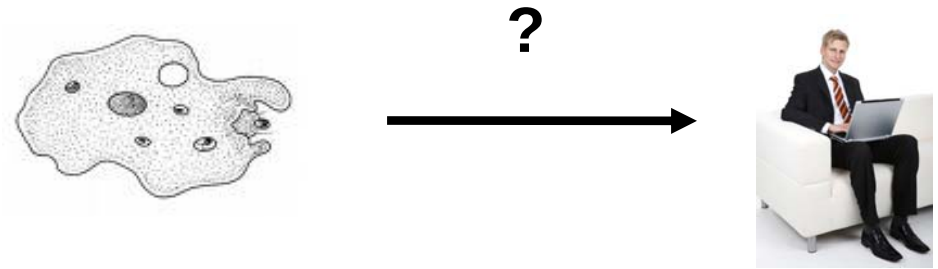
Mark A. Miller

San Diego Supercomputer Center

XSEDE Extreme Science and Engineering Discovery Environment

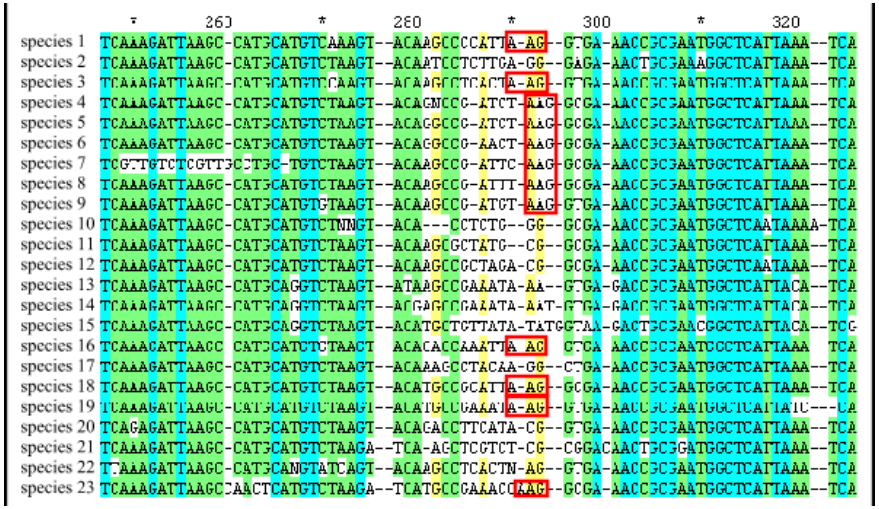
SDSC

Phylogenetics is the study of the diversification of life on the planet Earth, both past and present, and the relationships among living things through time

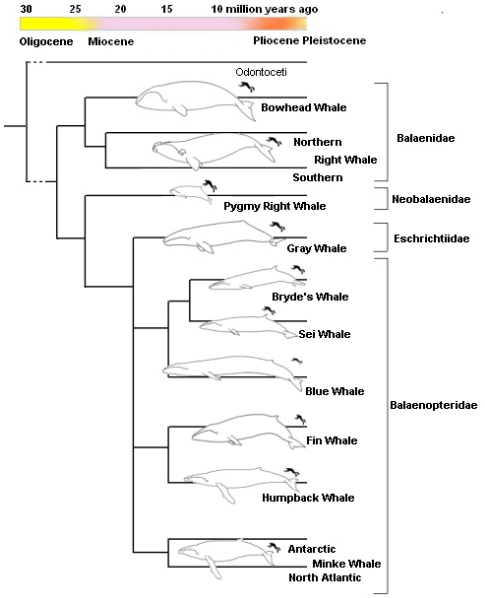


Evolutionary relationships can be inferred from DNA sequence comparisons:

1. Align sequences to determine evolutionary equivalence:



2. Infer evolutionary relationships based on some set of assumptions:



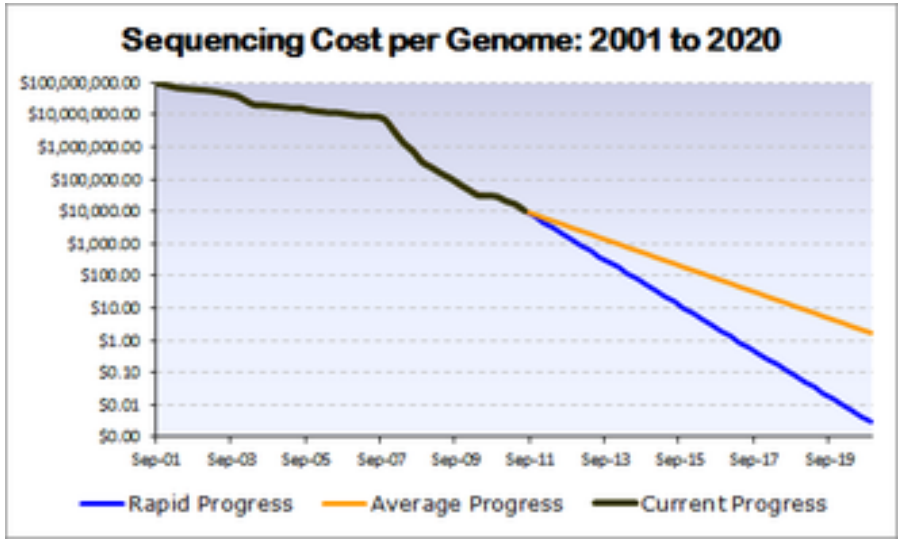
Inferring Evolutionary relationships from DNA sequence comparisons is powerful:

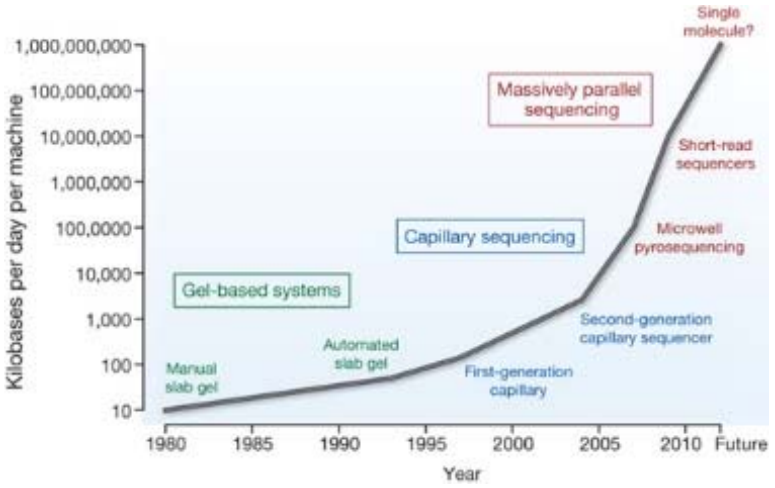
DNA sequences are determined by fully automated procedures.

Sequence data can be gathered from many species at scales from gene to whole genome.

The high speed and low cost of NexGen Sequencing means new levels of sensitivity and resolution can be obtained.

The speed of sequencing is still increasing, while the cost of sequencing is decreasing.





Data availability is no longer the limiting resource in inferring Evolutionary relationships…..

Inferring Evolutionary relationships from DNA sequence comparisons is powerful, BUT:

Current analyses often involve 1000's of species and 1000's of characters, creating very large matrices.

Sequence alignment and Tree inference are NP hard, so even with heuristics, computational power often limits the analyses (already).

The length of tree search analysis scales exponentially with number of taxa and with number of characters with codes in current use.

There are at least 10^7 species, each with 3000 - 30,000 genes, so the need for computational power and new approaches will continue to grow.

Inferring Evolutionary relationships from DNA sequence comparisons is powerful, BUT:

Current analyses often involve 1000's of species and 1000's of characters, creating very large matrices.

Sequence alignment and Tree inference are NP hard, so even with heuristics, computational power often limits the analyses (already).

The length of tree search analysis scales exponentially with number of taxa and with number of characters with codes in current use.

There are at least 10^7 species, each with 3000 - 30,000 genes, so the need for computational power and new approaches will continue to grow.

Inferring Evolutionary relationships from DNA sequence comparisons is powerful, BUT:

Current analyses often involve 1000's of species and 1000's of characters, creating very large matrices.

Sequence alignment and Tree inference are NP hard, so even with heuristics, computational power often limits the analyses (already).

The length of tree search analysis scales exponentially with number of taxa and with number of characters with codes in current use.

There are at least 10^7 species, each with 3000 - 30,000 genes, so the need for computational power and new approaches will continue to grow.

Inferring Evolutionary relationships from DNA sequence comparisons is powerful, BUT:

Current analyses often involve 1000's of species and 1000's of characters, creating very large matrices.

Sequence alignment and Tree inference are NP hard, so even with heuristics, computational power often limits the analyses (already).

The length of tree search analysis scales exponentially with number of taxa and with number of characters with codes in current use.

There are at least 10^7 species, each with 3000 - 30,000 genes, so the need for computational power and new approaches will continue to grow.

Inferring Evolutionary relationships from DNA sequence comparisons is powerful, BUT:

Current analyses often involve 1000's of species and 1000's of characters, creating very large matrices.

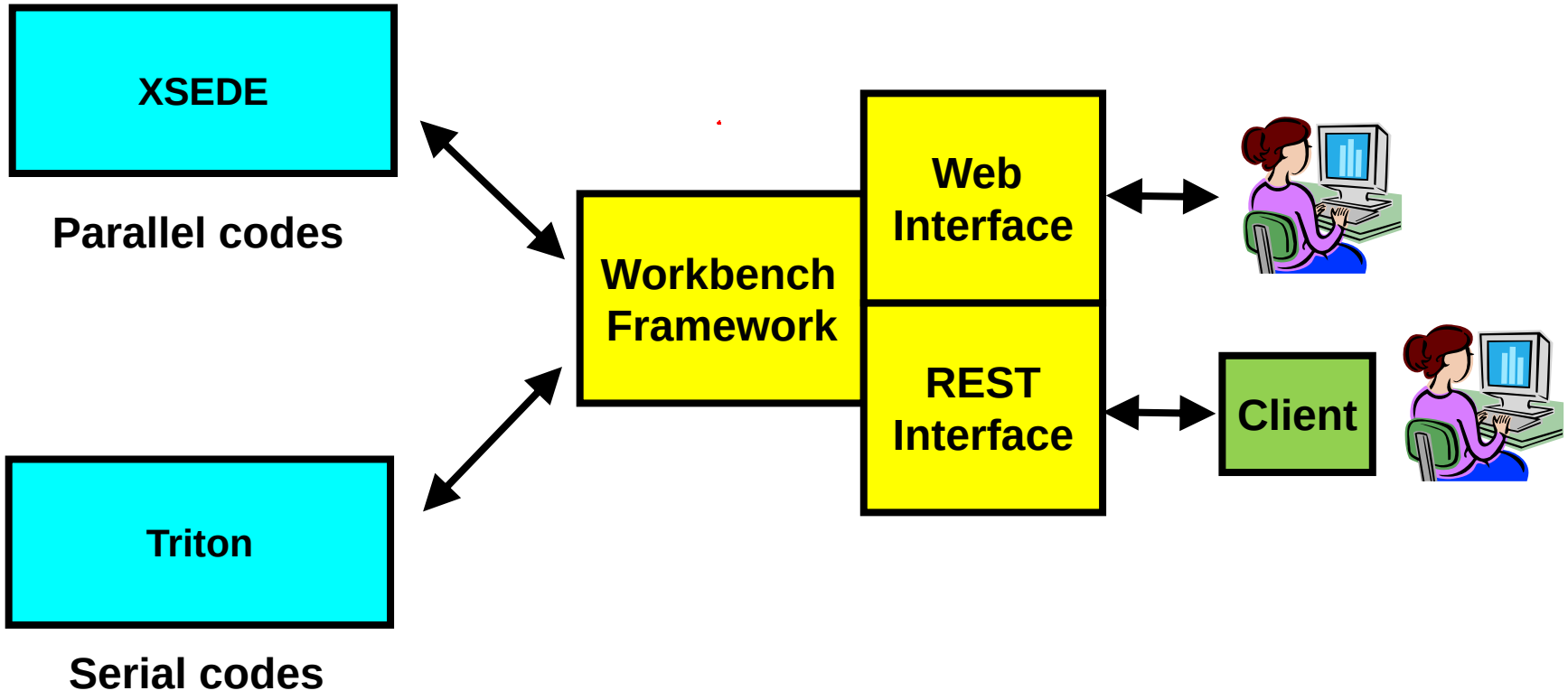
Sequence alignment and Tree inference are NP hard, so even with heuristics, computational power often limits the analyses (already).

The length of tree search analysis scales exponentially with number of taxa and with number of characters with codes in current use.

There are at least 10^7 species, each with 3000 - 30,000 genes, so the need for computational power and new approaches will continue to grow.

Participating in modern phylogenetics research requires access to High performance computing resources…..

CIPRES provides access to scalable, sustainable resources available through NSF funded programs.



The CIPRES Science Gateway was designed to allow users to analyze large sequence data sets using community codes on significant computational resources.

The CSG provides

- **Login-protected personal user space for storing results indefinitely.**
- **Access to most/all native command line options for several codes.**
- **Support for adding new tools and upgrading to new versions as needed.**
- **Access up to 50,000 core hours of compute time per year at no cost.**

The CIPRES Science Gateway was designed to allow users to analyze large sequence data sets using community codes on significant computational resources.

The CSG provides

- **Login-protected personal user space for storing results indefinitely.**
- **Access to most/all native command line options for several codes.**
- **Support for adding new tools and upgrading to new versions as needed.**
- **Access up to 50,000 core hours of compute time per year at no cost.**

The CIPRES Science Gateway was designed to allow users to analyze large sequence data sets using community codes on significant computational resources.

The CSG provides

- **Login-protected personal user space for storing results indefinitely.**
- **Access to most/all native command line options for several codes.**
- **Support for adding new tools and upgrading to new versions as needed.**
- **Access up to 50,000 core hours of compute time per year at no cost.**

The CIPRES Science Gateway was designed to allow users to analyze large sequence data sets using community codes on significant computational resources.

The CSG provides

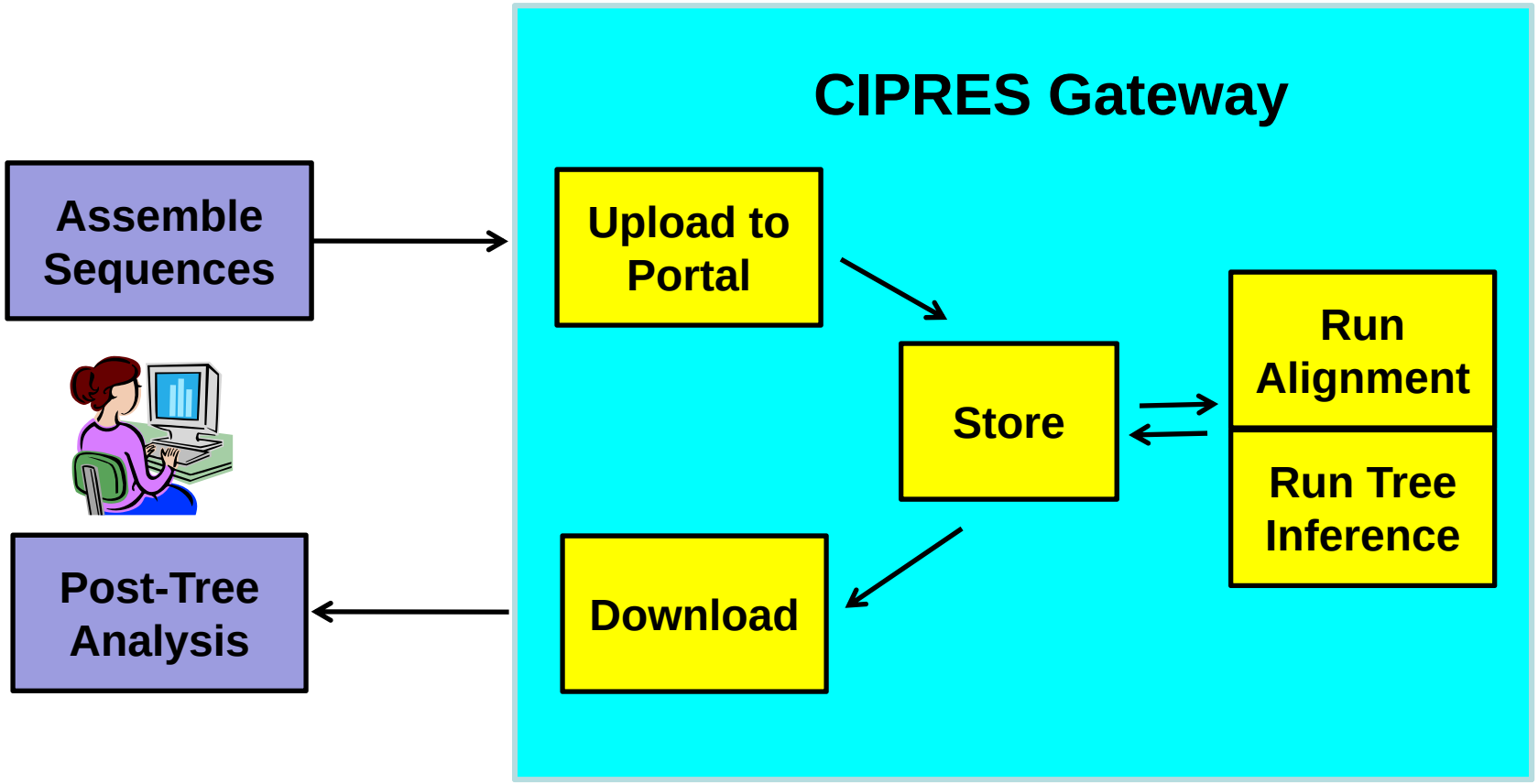
- **Login-protected personal user space for storing results indefinitely.**
- **Access to most/all native command line options for several codes.**
- **Support for adding new tools and upgrading to new versions as needed.**
- **Access up to 50,000 core hours of compute time per year at no cost.**

The CIPRES Science Gateway was designed to allow users to analyze large sequence data sets using community codes on significant computational resources.

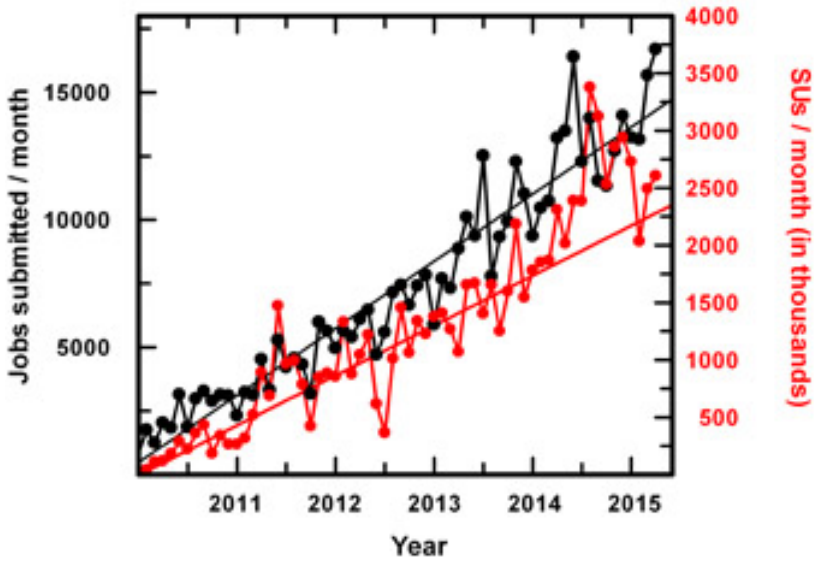
The CSG provides

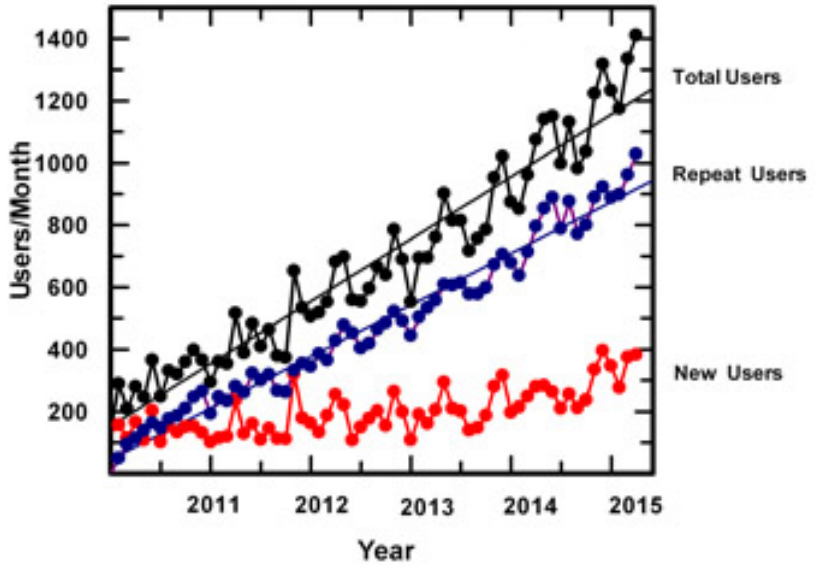
- **Login-protected personal user space for storing results indefinitely.**
- **Access to most/all native command line options for several codes.**
- **Support for adding new tools and upgrading to new versions as needed.**
- **Access up to 50,000 core hours of compute time per year at no cost.**

Workflow for the CIPRES Gateway:



CIPRES Science Gateway Usage Statistics 12/1/2009 - 3/31/2015





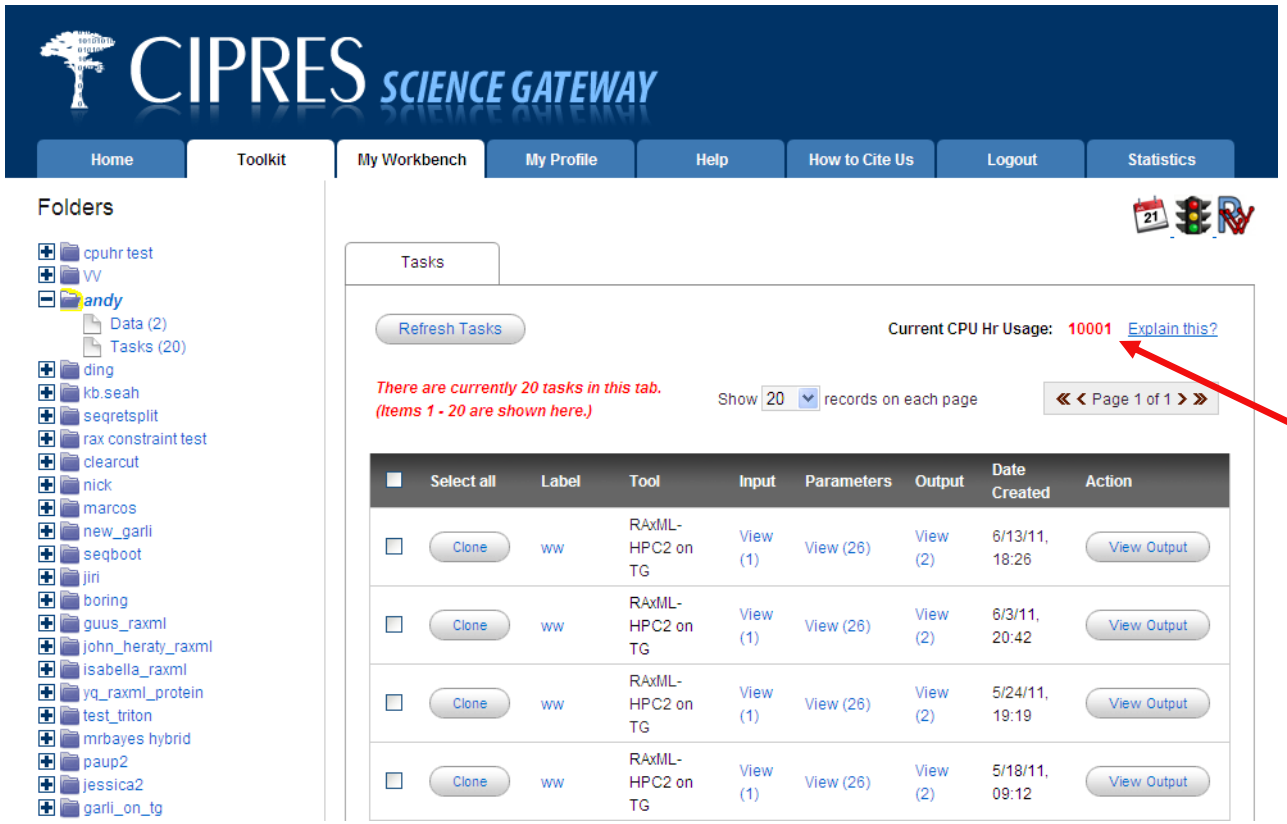
Jobs - Black

Core hours - Red

Innovations in response to our growing user base :

- ability to halt submissions from a given user account
- ability to monitor usage by each account automatically
- help users track their resource consumption / forecast resource cost of jobs
- ability to charge to a user's personal XSEDE allocation
- automatic sunset of inactive accounts and delete data
- store each data item only once
- *give users the ability to download bulk data/ delete bulk data.*

Help users track their resource consumption:



Notify users of their usage level

Innovations driven by evolving requirements:

- implement queue policies that allow 2 week long runs
- make job tracking robust against loss of communication between server and remote resource
- ability to return output files that are > 4 GB in size
- ability to run from inside another application

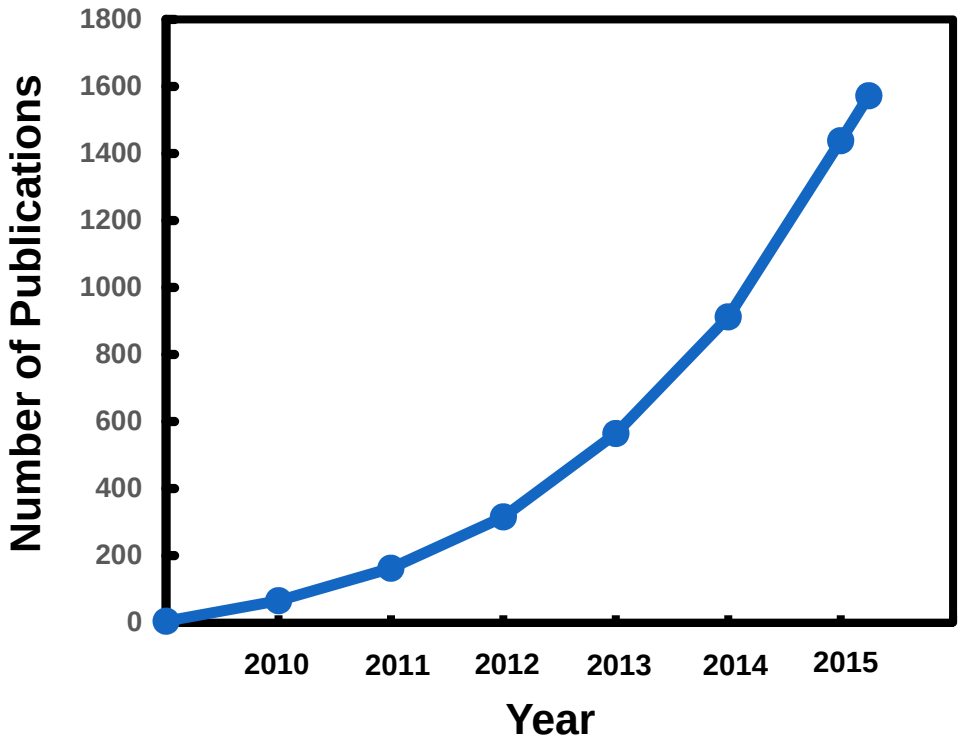
Broad Impact:

- **In Q1 2015, 29% of all XSEDE users who ran jobs ran them from CIPRES.**
- **50% of users said they had no access to local resources, nor funds to purchase access on cloud computing resources**
- **Used for curriculum delivery by at least 76 instructors.**
- **Routine submissions from Harvard, Berkeley, Stanford, as well as many non-PhD granting institutions……**
- **55% of users are in the US or have a collaborator in the US**

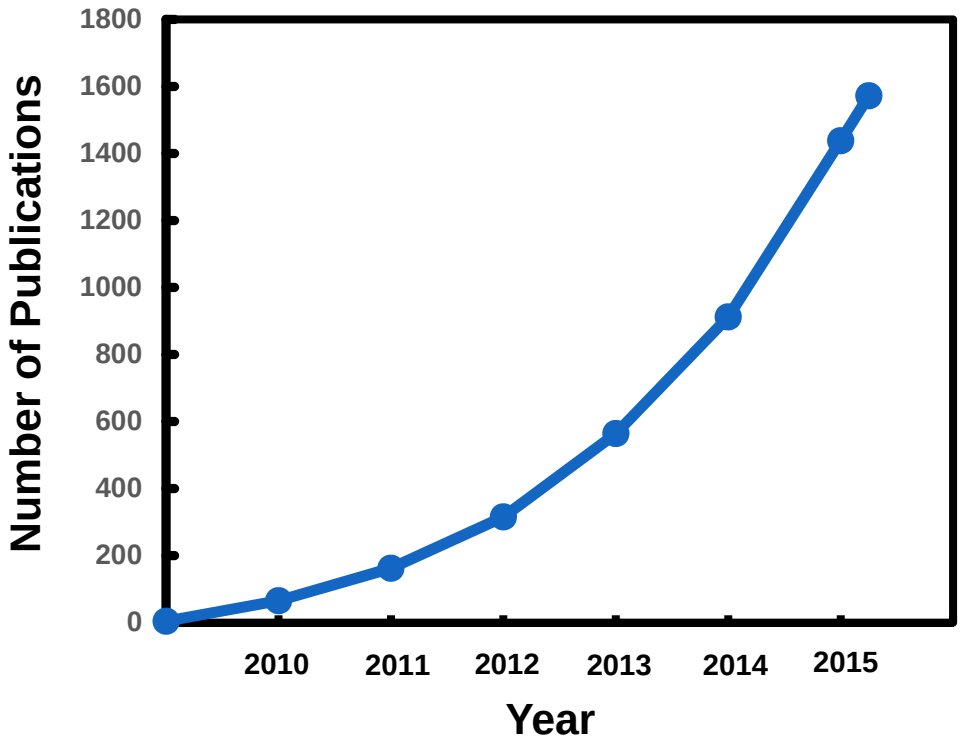
Broad Impact:

"It is hard for me to imagine how I could work at a reasonable pace without this resource, especially when things like MS or grant submission deadlines loom…."

Publications enabled by CIPRES:



Publications enabled by CIPRES:



About 1 publication for every 50,000 core hours of compute time…..

Broad Impact:

“It is an easy-to-use cluster to run BEAST analyses in a short time. This allows students to run analyses that actually converge in a single class.**”**

“I found it is important to be able to let the student explore the analysis 'all the way', i.e. not just show the principle but actually let them run an entire Markov chain and let them evaluate the results. For that I found that having access to the Cipres Science Gateway to be crucial.”

CIPRES Science Gateway Usage Statistics 12/1/2009 - 3/31/2015

- 472,832 TeraGrid/XSEDE jobs submitted by 12,667 unique users.
- Average of 297 new XSEDE users registered in each of the last 12 months.
- 81.3 million core hours of TeraGrid/XSEDE time distributed to scientists.
- Used for curriculum delivery by at least 76 instructors.
- Supported at least 1570 publications.