From the Frying Pan to the Fire: Bioinformatics Computing in Two Lectures

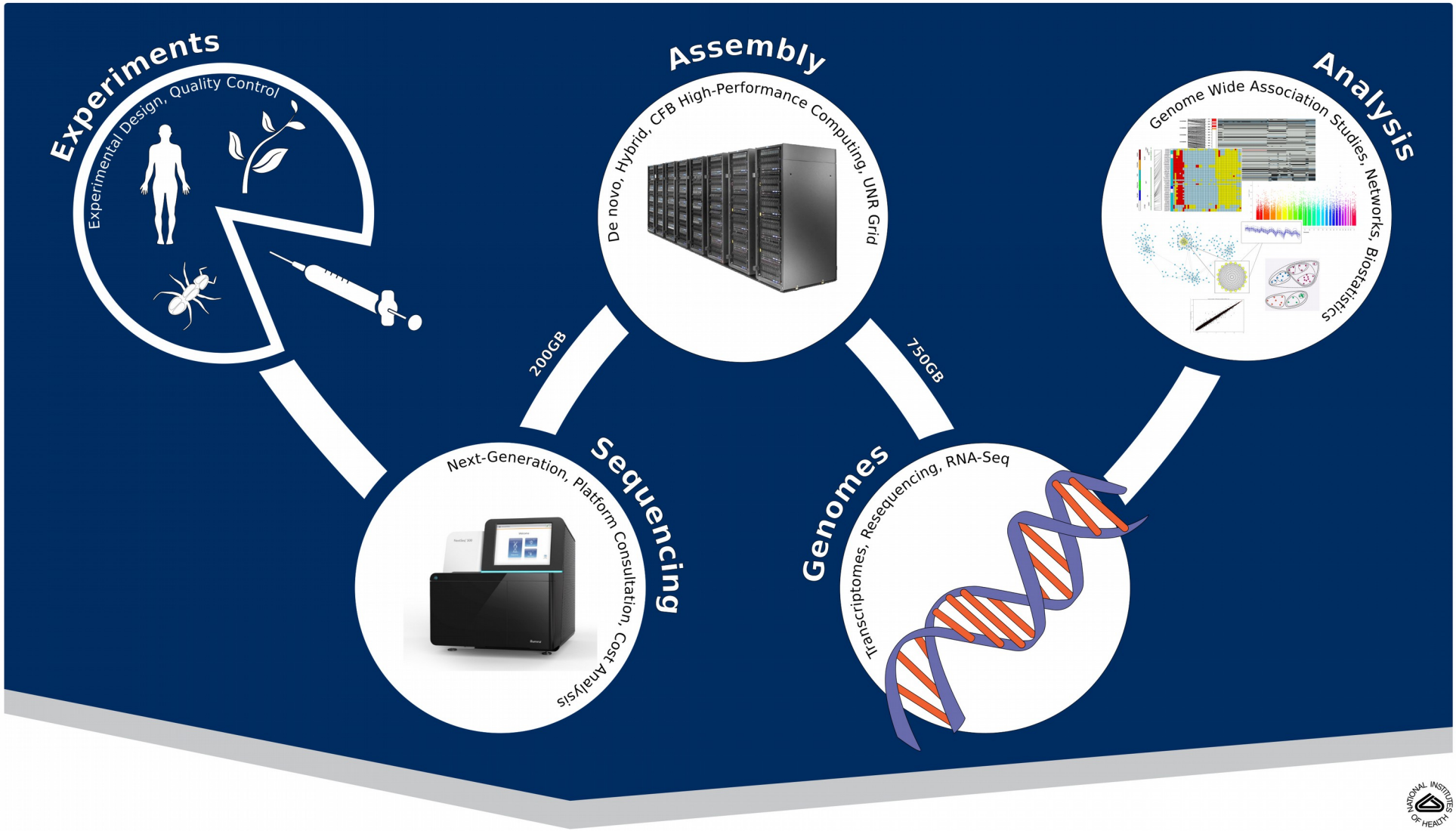
Sebastian Smith
Nevada Center for Bioinformatics
stsmith@unr.edu

BCH 709: Intro to Bioinformatics Fall 2015

Overview

- Introduction
- Linux in a nanosecond nutshell
  - Frying pan
- Hands-on genome assembly on UNR Grid
  - Fire
- Today = Computing
- Next = Biochem

WE BUILD GENOMES



Director: Karen Schlauch, Ph.D.
Richard L. Tillett, Ph.D.
Juli Petereit
Sebastian Smith

Nevada Center for Bioinformatics

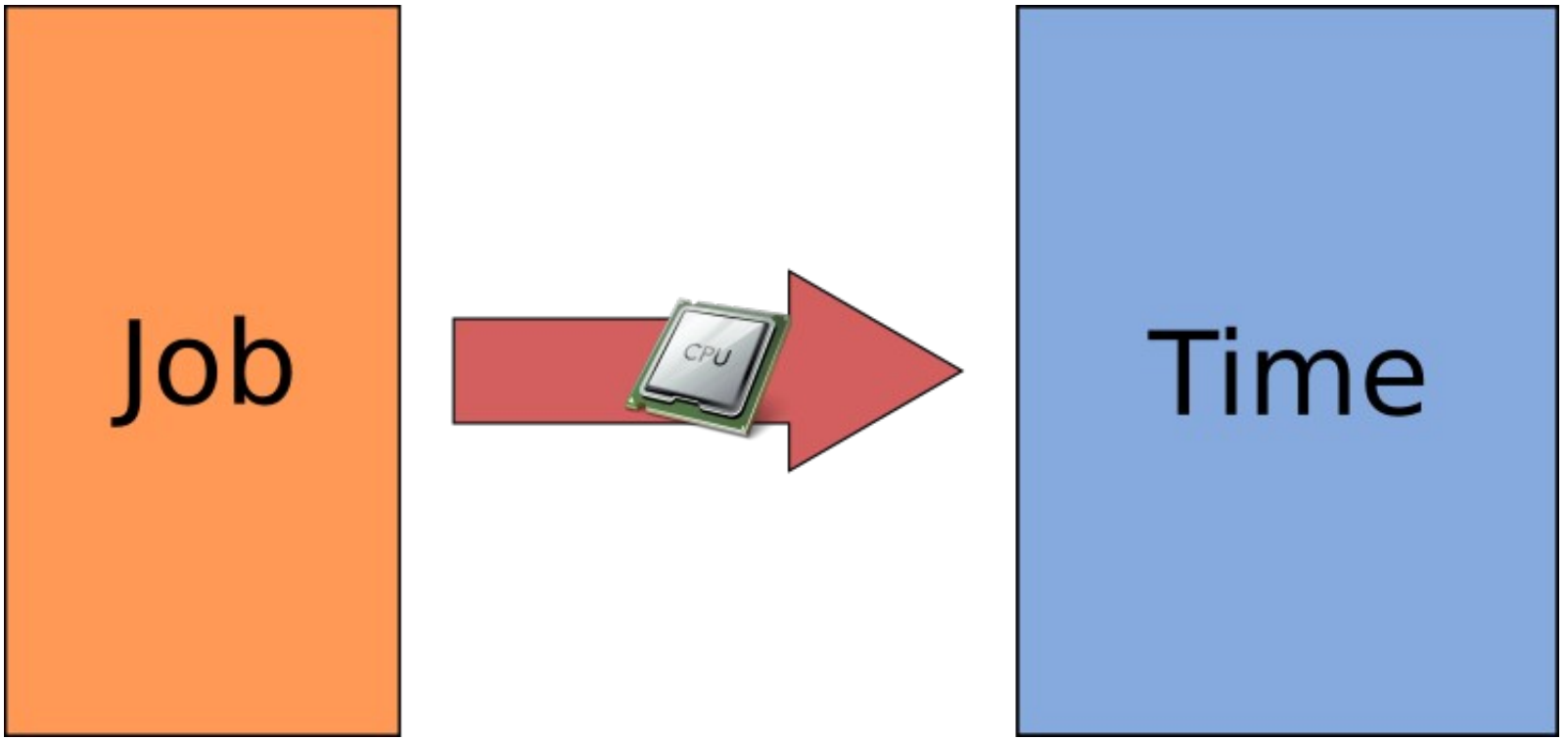
www.unr.edu/bioinformatics

This poster was made possible by a grant from the NIH NIGMS P20GM103440

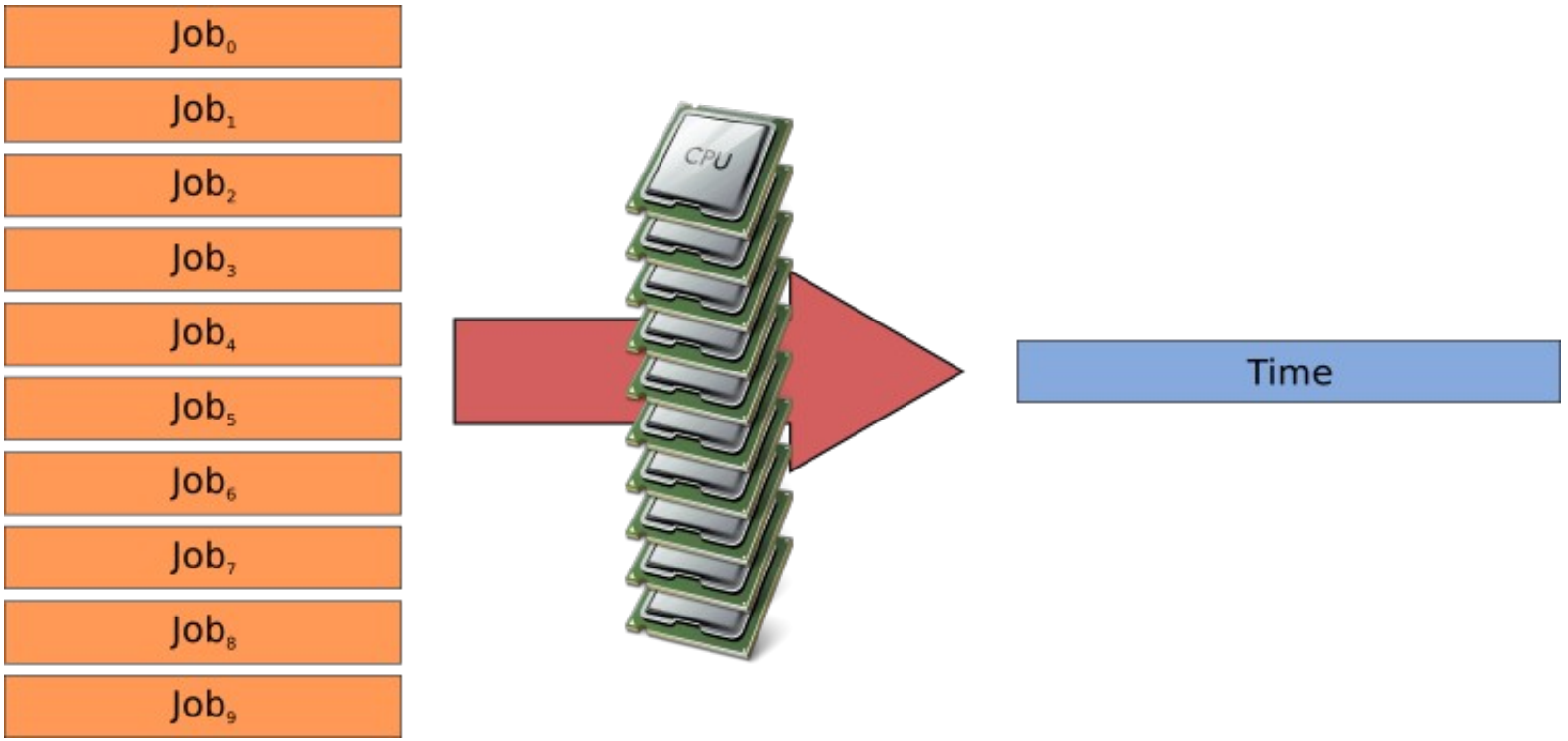
CFB Computing

- Example: Assembly of a new genome exceeds 20,000 hrs of computation
  - >2.25 years for high-quality results
  - Our goal is to complete this computation in ***7*** days
- How?
  - High-performance software and systems
    - Big computers
    - Advanced algorithms
    - Parallelism
    - Automation

HPC: The Gist



HPC: The Gist



This Class

- 10,000ft view of HPC at CFB
- You will use UNR research computing to:
  - Assemble a genome
  - Perform an analysis
- Goals:
  - Provide a HPC “starter kit”
    - Use as a template for your own work
  - Introduce you to UNR HPC resources

Into the Frying Pan

UNR Grid

- Primary research computer
- Cluster/Grid
  - Collection of computers that work together as one
  - 27 computers
    - 16 CPU, 256GB RAM
    - 432 total CPUs, 4TB RAM
  - 40+TB disk
- Why use it?
  - Annihilate work
  - ~$200k computer you don't have to buy
- Uses ***Linux***
  - Centos 6.7

Linux

- Monolithic kernel
  - Manages applications, computer hardware and their interaction
- Operating system
  - Highly modular
  - Distributions
    - Manage collections of software, configuration, security
    - Make it easy for end-users
    - Ubuntu, Debian, RedHat, CentOS, Fedora, Mint, Arch, Suse

Why Linux?

- It's free
  - Open source
  - Powerful, free development ecosystem
- Versatile
  - Highly configurable
    - Real time systems, desktop computing, supercomputing, specialty
- Scalable
  - Refrigerators to supercomputers
- Performant
- Stable
- Pervasive
  - 28% mainframe, 29% embedded, 36% web, 53% mobile, 97% top500 supercomputing
  - 48% of 2009 businesses use it

CFB Software

- Non-interactive
  - Automatic, daemon
- Shell
  - Command Line Interface
    - Simple human interface
      - Read-Eval-Print Loop (REPL)
    - Have to memorize commands, complex to learn
    - Fast and powerful
    - Powerful programming languages
      - Automation
      - Control self and other computers
    - Low-bandwidth remote-control

Remote Control

- You're on Windows... how the heck do we control the UNR super computer?
- Secure Shell (ssh)
  - Secure, remote CLI
  - Windows = Putty
  - Linux = ssh
- Secure CoPy (scp)
  - Securely transfer files
  - Windows = WinSCP
  - Linux = scp

Connect to the Grid

- Execute Putty.exe
- Type “login.research.unr.edu” into the “Host Name” box
- Type “22” into the “Port” box
- Click “Open” button
- Input NetID as user name
- Input NetID password as password

The Grid Shell

- Command interpreter
- Bash
  - Bourne-again shell
- Important features
  - Scripting language
    - Can read commands from a file
  - Tab completion
    - Minimize spelling errors

Commands

- Software application
  - Input from the shell
  - Output to the shell
- Structure
  - *command [OPTIONS] <ARGUMENTS>*
  - Type command name, options, arguments and press enter to execute

$ rsync -avzP stsmith@bioinformatics.unr.edu:~/class/ stsmith@login.research.unr.edu:~/

Files

- Everything is a file
- Filesystem is a directed graph
  - Root = /
  - Home = /home
  - Your home = /home/<NetID>
- Folders = directories
- Current Working Directory (CWD)
  - Where you are currently located
  - Your home directory when you log in
- Path
  - Specifies a location in the filesystem
  - Absolute
    - Points to the same location regardless of CWD
    - Always starts at root
  - Relative
    - Points to a different location dependent on CWD
- “/” = path delimiter

- Shortcuts
  - ~/ = home directory
  - ./ = current directory
  - ../ = parent directory

Manipulating Files

- pwd
  - List the current working directory
- ls [path]
  - List directory contents
- cd <path>
  - Change directory
- mkdir <path>
  - Make a directory
- mv <src path> <dest path>
  - Move/rename a file or directory
- cp <src path> <dest path>
  - Copy a file
  - Copy a directory with the -r option
- rm <path>
  - Remove a file
  - Remove a directory with the -r option
- file <file path>
  - Determine file type
- less <file path>
  - Read contents of text files

Editing Text Files

- nano <file path>
- ctrl + x to exit
  - Will prompt to save on exit

Processes

- Three types
  - Interactive, automatic (at, batch, cron), daemon
- ps
  - Report processes
  - Report all processes on system with -ef options
- top
  - Display tasks in table format
  - Updates periodically
- Kill [-s signal] <process id>
  - Sends a signal to processes
  - Kill a process with -s 9 option

Cluster Computing

- All users share a common resource
- Resource starvation/software failure if we all run at the same time
- Take turns using the resource
- Batch-queuing system
  - Job queue
  - Job scheduler
  - Program that manages background program execution
- UNR Grid uses Sun/Oracle Grid Engine (SGE)

Cluster Management

- qsub <script>
  - Submit a job to SGE queue
- qstat
  - Show status of SGE jobs
- qdel <job id>
  - Delete SGE job from queue

Into the Fire

We Know Enough To Be Dangerous

- Run a genome assembly on The Grid
- We've packaged a small, simplified assembly task
  - Explore the package on your own time to learn how it works
  - ~5 hour runtime on 16 CPUs
    - 80 CPU hours

1

- Log in to the grid head node
  - SSH to login.research.unr.edu
  - Username = NetID
  - Password = NetID password

2

```
$ cd ~/scratch
```

3

- Get class files from Github
  - https://github.com/UNR-CFB/bch-709-intro-bioinformatics-2015f

  $ git clone https://github.com/UNR-CFB/bch-709-intro-bioinformatics-2015f.git

- Makes a bch-709-intro-bioinfomratics-2015f directory

4

$ cd bch-709-intro-bioinformatics-2015f

5

- Execute the setup script

  $ ./setup.sh

- This links a directory containing large Illumina sequencer reads that already exists on the grid
  - Faster than copying the files to everyone

6

$ cd day-1/sge

7

- Submit the assemble_geno.sge job to the queue

  $ qsub assemble_geno.sge

8

- Watch the status of your job change

  $ watch qstat

9

- Let it bake
- 4.5..5 hours

What Did We Just Do?

$ less assemble_geno.sge

- We told the SGE queue to find a computer that has 16 free CPUs and run the assemble_geno.sge script on it
- assemble_geno.sge runs an assembler called Spades on the Illumina sequencer reads to produce a genome
- We will use the genome to answer questions tomorrow

Questions?

Getting Help

- Nevada Center for Bioinformatics
  - We help with all stages of research
  - Training, systems administration
  - Sebastian Smith (systems and software)
    - stsmith@unr.edu
  - Richard Tillett (biochem)
    - rltillett@unr.edu
- UNR Grid
  - John Anderson
    - jra@unr.edu
  - http://www.unr.edu/it/research-resources/the-grid