Open Science Grid

Backpacking with Code: Software Portability for DHTC

Thursday morning, 9:00 am

Christina Koch (ckoch5@wisc.edu)
Research Computing Facilitator
University of Wisconsin - Madison

Open Science Grid

Goals for this session

- Understand the basics of...
  - how software works
  - where software is installed
  - how software is accessed and run
- ...and the implications for DHTC
- Describe what it means to make software "portable"
- Learn about and use two software portability techniques:
  - Run compiled code
  - Build installation and use it in jobs

Motivation

running a piece of software is like cooking a meal in a kitchen

The problem

Running software on your own computer = cooking in your own kitchen

Open Science Grid

The problem

In your own kitchen:

- You have all the pots and pans you need
- You know where everything is
- You have access to all the cupboards

On your own computer:

- The software is installed, you know where it is, and you can access it.

The problem

Running on a shared computer = cooking in someone else's kitchen.

Open Science Grid

The problem

In someone else's kitchen:

- You are guaranteed some things…
- …but others may be missing
- You don't know where everything is
- Some of the cupboards are locked

On a shared computer:

- Your software may be missing, un-findable, or inaccessible.

The solution

- Think like a backpacker
- Take your software with you
  - Install anywhere
  - Run anywhere
- This is called making software *portable*

Open Science Grid

Software

- How do we make software portable?
- First we have to understand:
  - What software is and how it works
  - Where software lives
  - How we run it

Open Science Grid

How software works

- A software program can be thought of as a list of instructions or tasks that can be run on an computer
- A launched program that is running on your computer is managed by your computer's operating system (OS)
- The program may make requests (access this network via wireless, save to disk, use another processor) that are mediated by the OS
- A single program may also depend on other programs besides the OS

How software works*

*Not to scale

Program
(software, code, executable, binary)

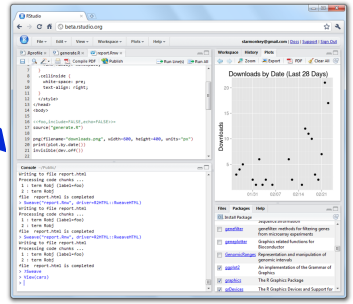
runs own tasks

makes requests

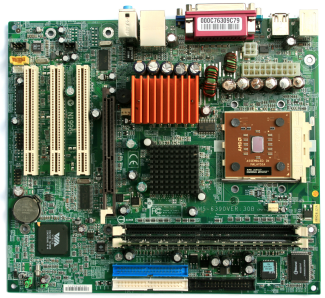
Operating System

translates program's request

launches to



monitors running programs



depends on

Running Program
(process, instance)

Hardware
(processors, memory, disk)

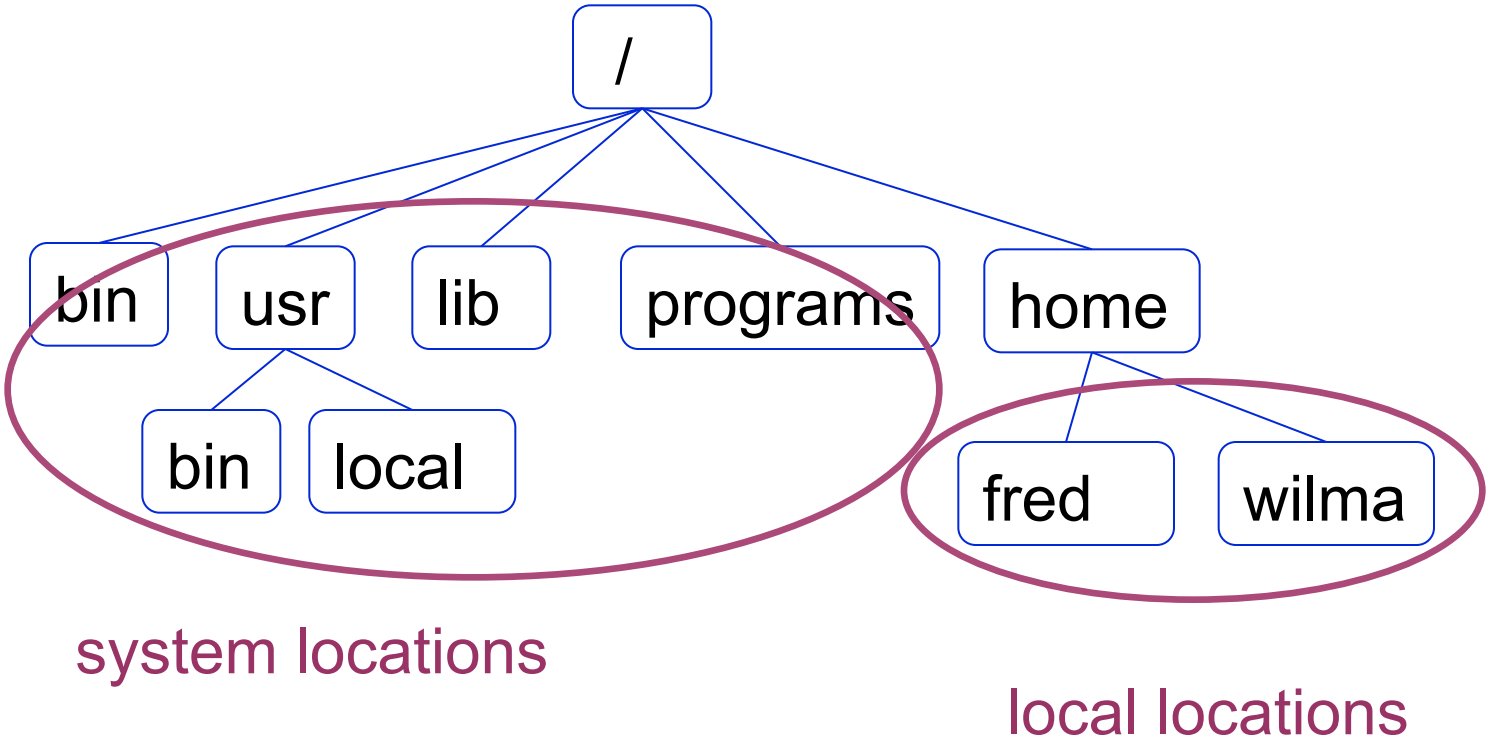
How software works

Implications for DHTC:

- Software must be able to run on target operating system (usually Linux)
- Request specific OS as job requirement
- Know what else your software depends on

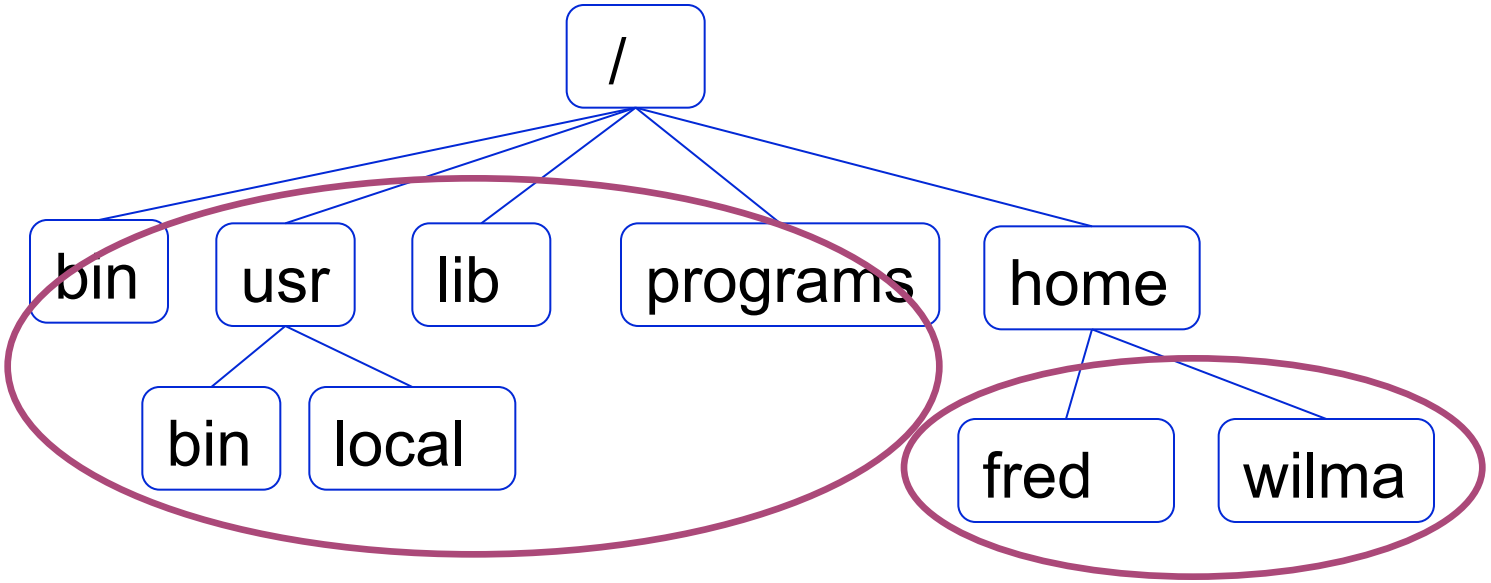
Location, location, location

- Where can software be installed?



Location, location, location

- Who can install the software?

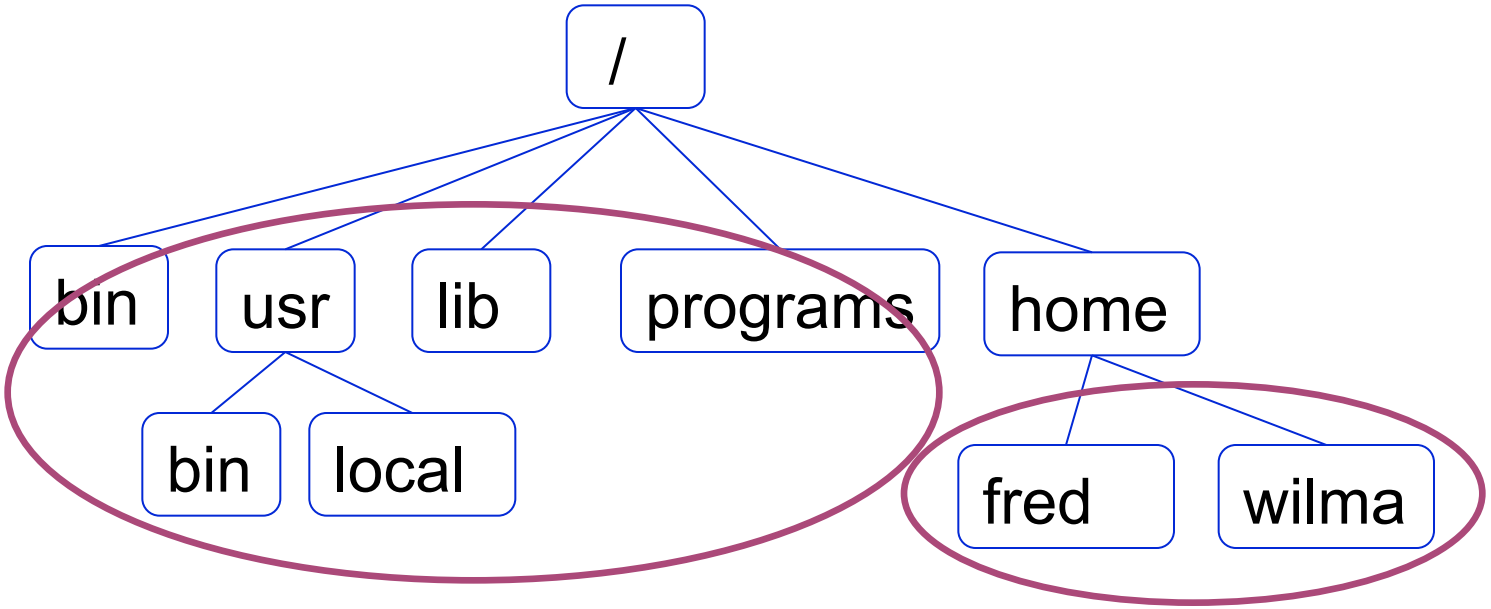


Usually requires administrative privileges

Owner of the directory

Location, location, location

- Who can access the software?



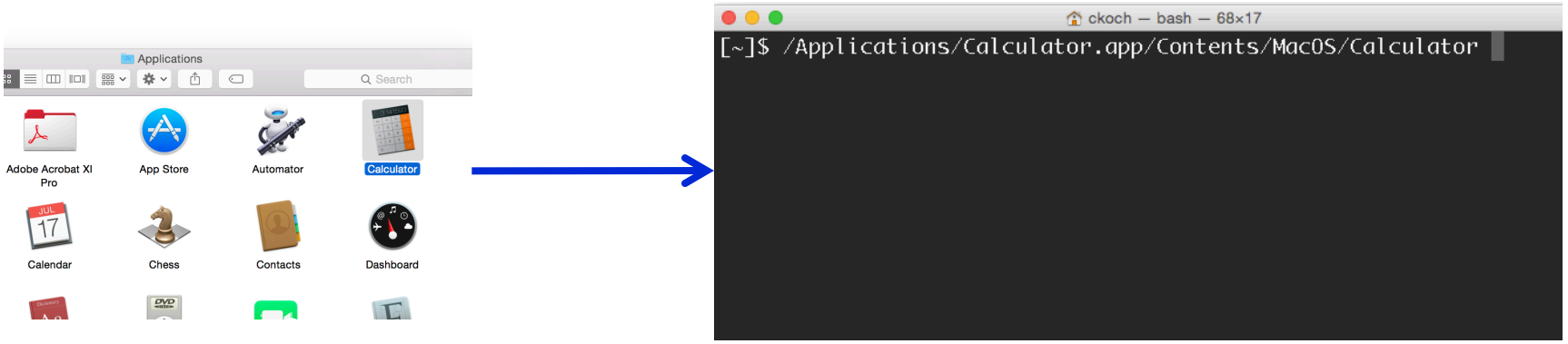
Location, location, location

Implications for DHTC:

- Software MUST be able to install to a local location
- Software must be installable without administrative privileges

Location and running software

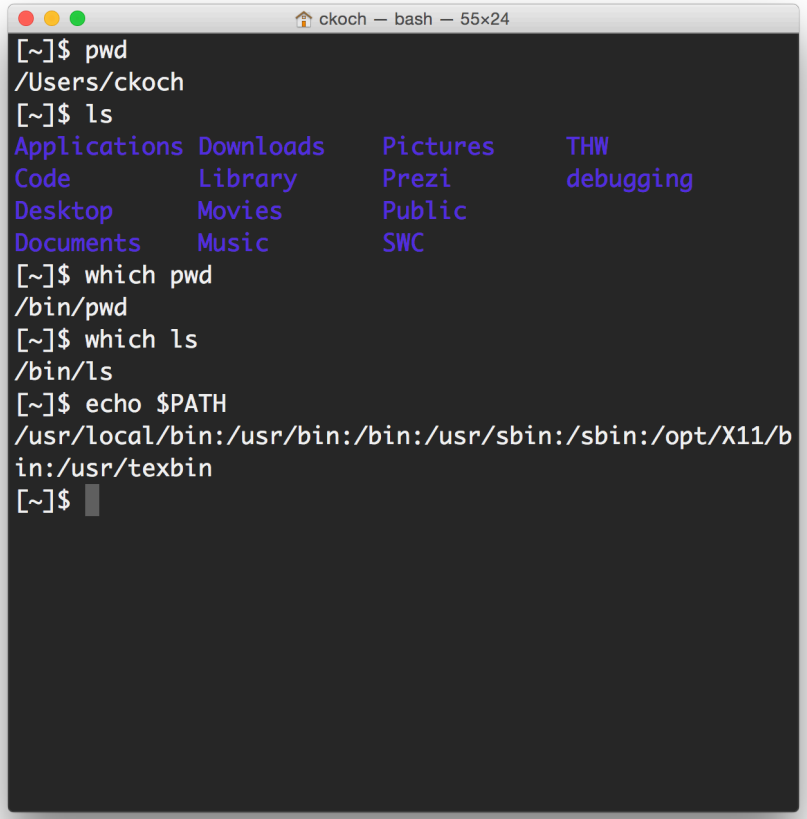
Instead of graphic interface… command line



- All DHTC jobs must use software that can be run from the command line.
- To run a program on the command line, your computer needs to know where the program is located in your computer's filesystem.

Common command line programs

- Common command line programs like `ls` and `pwd` are in a system location called `/bin`
- Your computer knows their location because `/bin` is included in your `PATH`

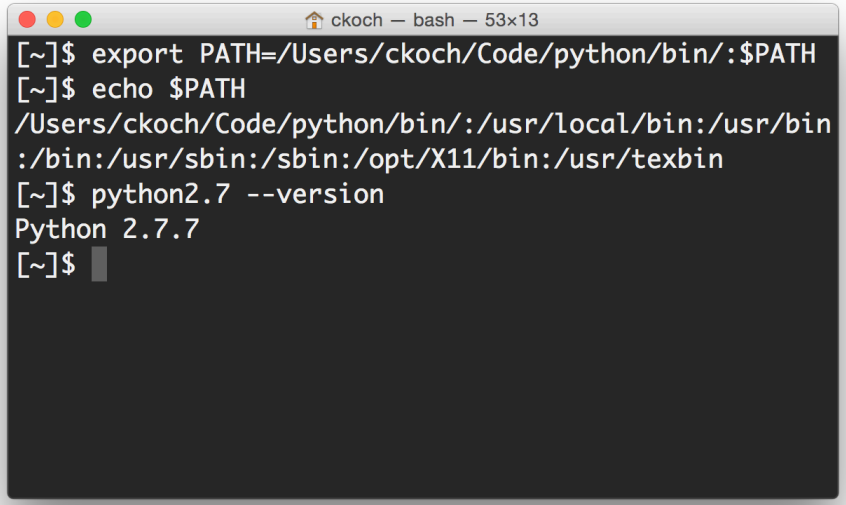


- The PATH is a list of locations to look for programs

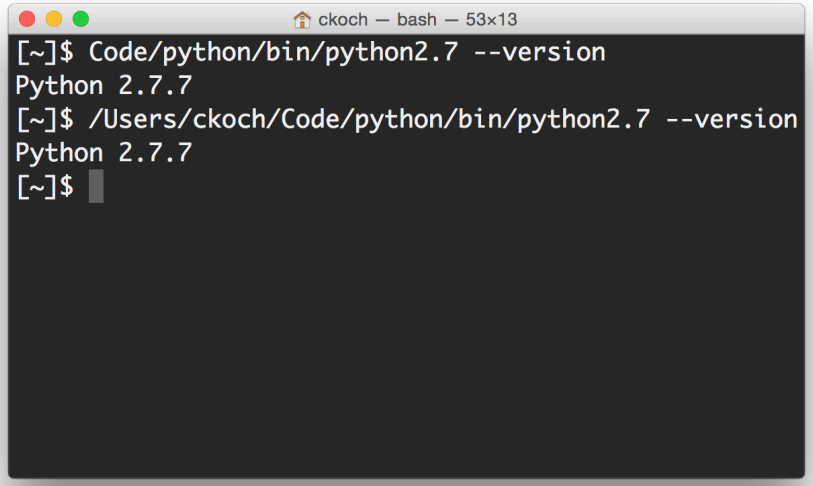
Other programs on command line

- Other programs may be installed in locations not listed in the PATH. You can access them by:

Adding their location to the PATH, then running



Using an relative or absolute path to the software



Command line

Implications for DHTC:

- Software must have ability to be run from the command line
- Multiple commands are okay, as long as they can be executed in order within a job
- There are different ways to “find” your software on the command line: relative path, absolute path, and PATH variable

Portability requirements

Based on the previous slides, we now know that in order to make software portable for DHTC, the software:

- Must work on target operating system (probably Linux)
- Must be accessible to your job (placed or installed in job's working directory)
- Must be able to run without administrative privileges
- Must be able to run from the command line, without any interactive input from you

Returning to our scenario:

In a DHTC situation, we are:

- Using someone else's computer
  - Software may not be installed
  - The wrong version may be installed
  - We can't find/run the installed software

Therefore:

- We need to bring along and install/run software ourselves

Open Science Grid

Portability methods

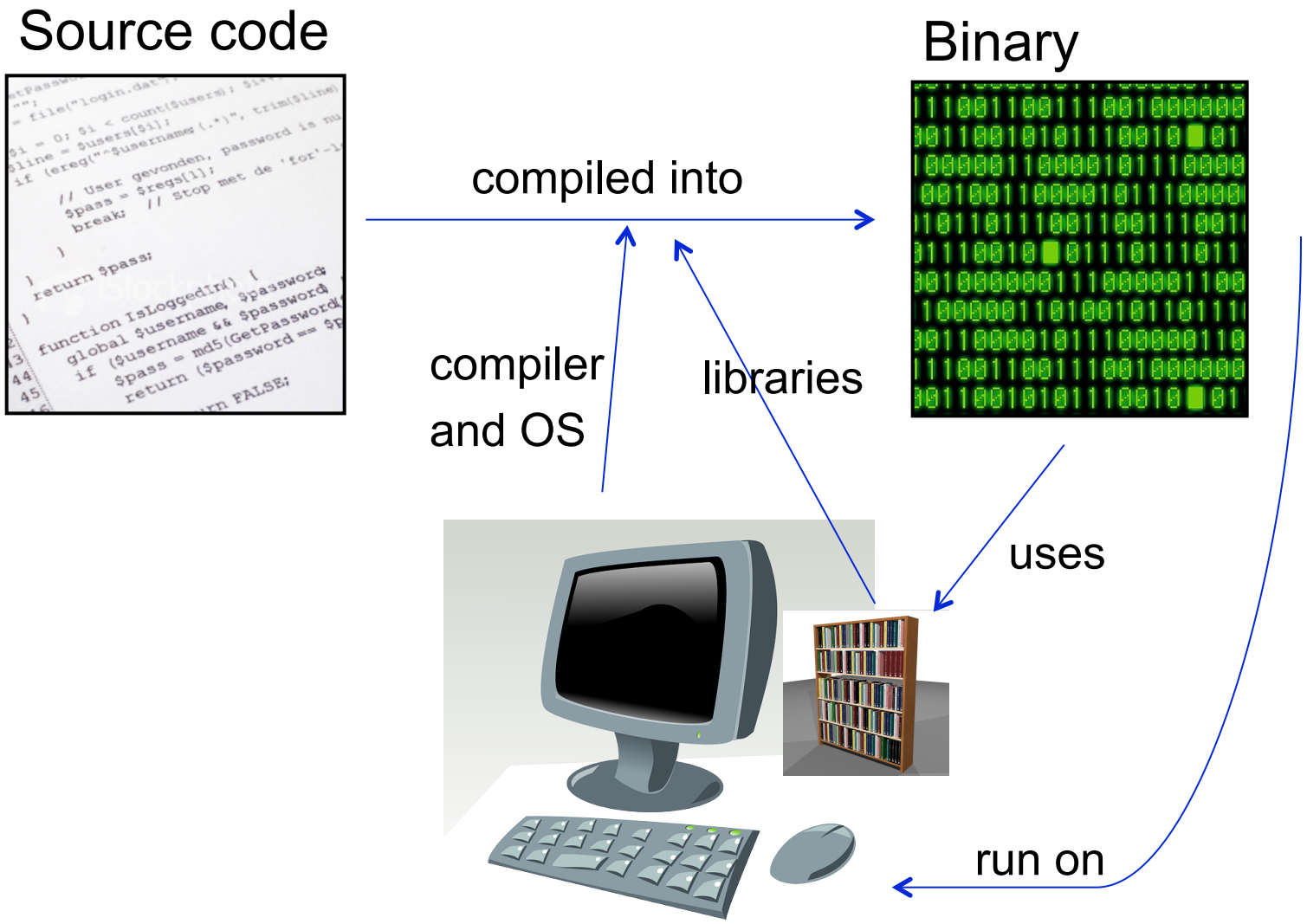
There are two primary methods to make code portable:

- Use a single compiled binary
  - Typically for code written in C, C++ and Fortran
- “Install” with every job
  - Can’t be compiled into a single binary
  - Interpreted languages (Matlab, Python, R)

Method 1

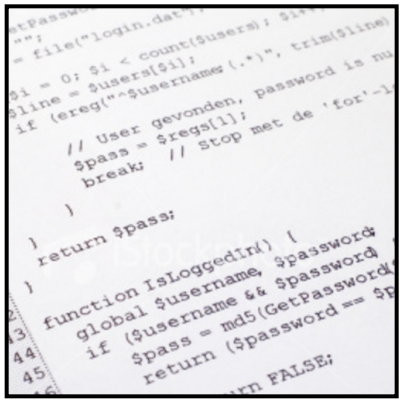
USE A COMPILED BINARY

What is compilation?



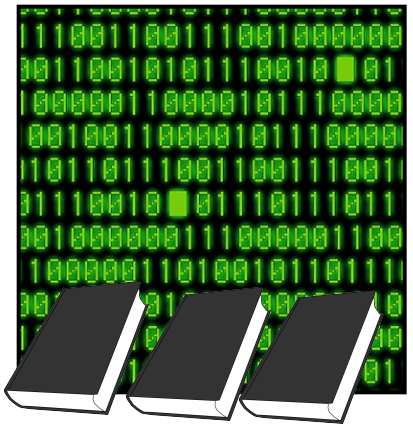
Static compilation

Source code



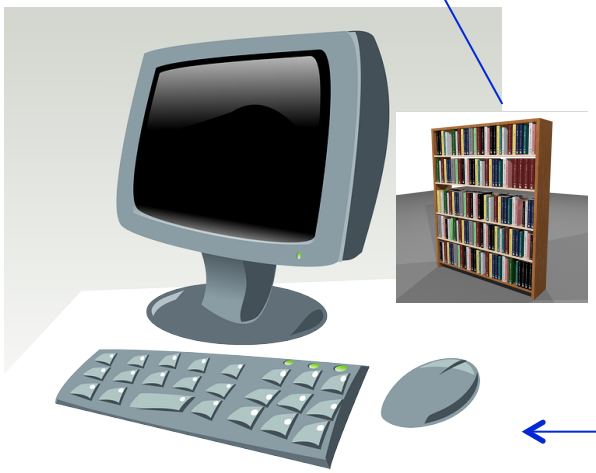
statically compiled into

Static binary



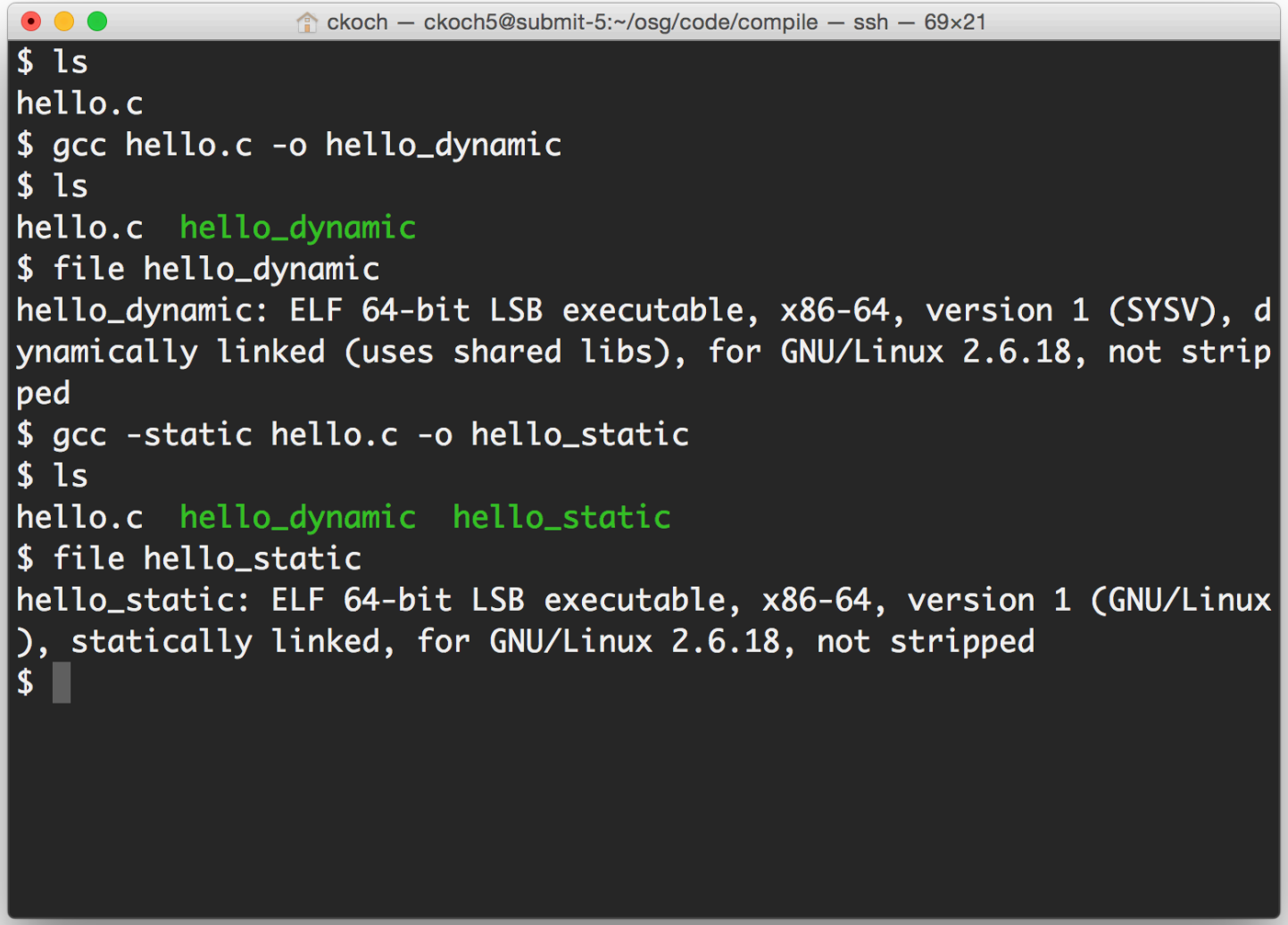
compiler and OS

libraries

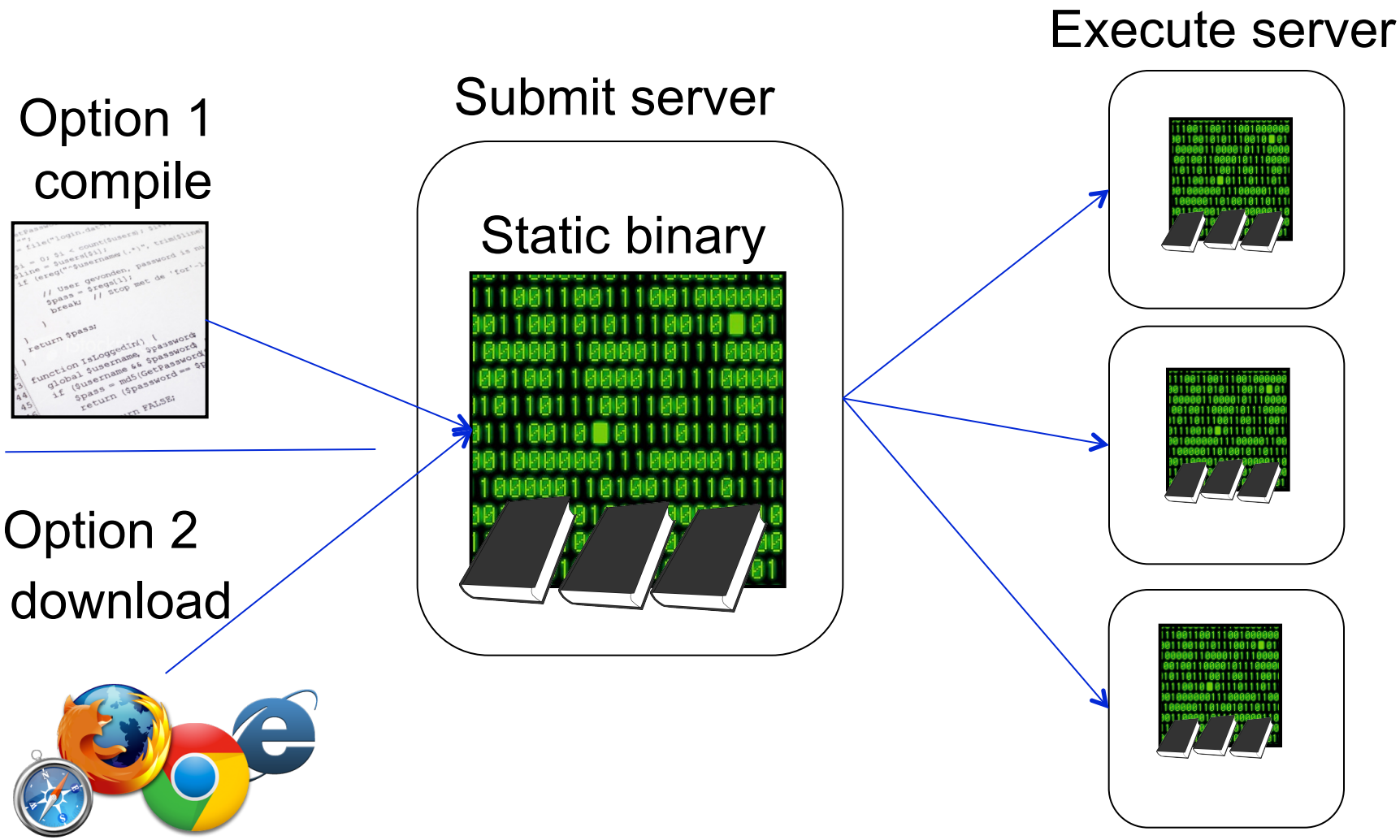


run anywhere

Compilation (command line)



Static compilation workflow



Method 2

INSTALL WITH EVERY JOB

Install software with every job

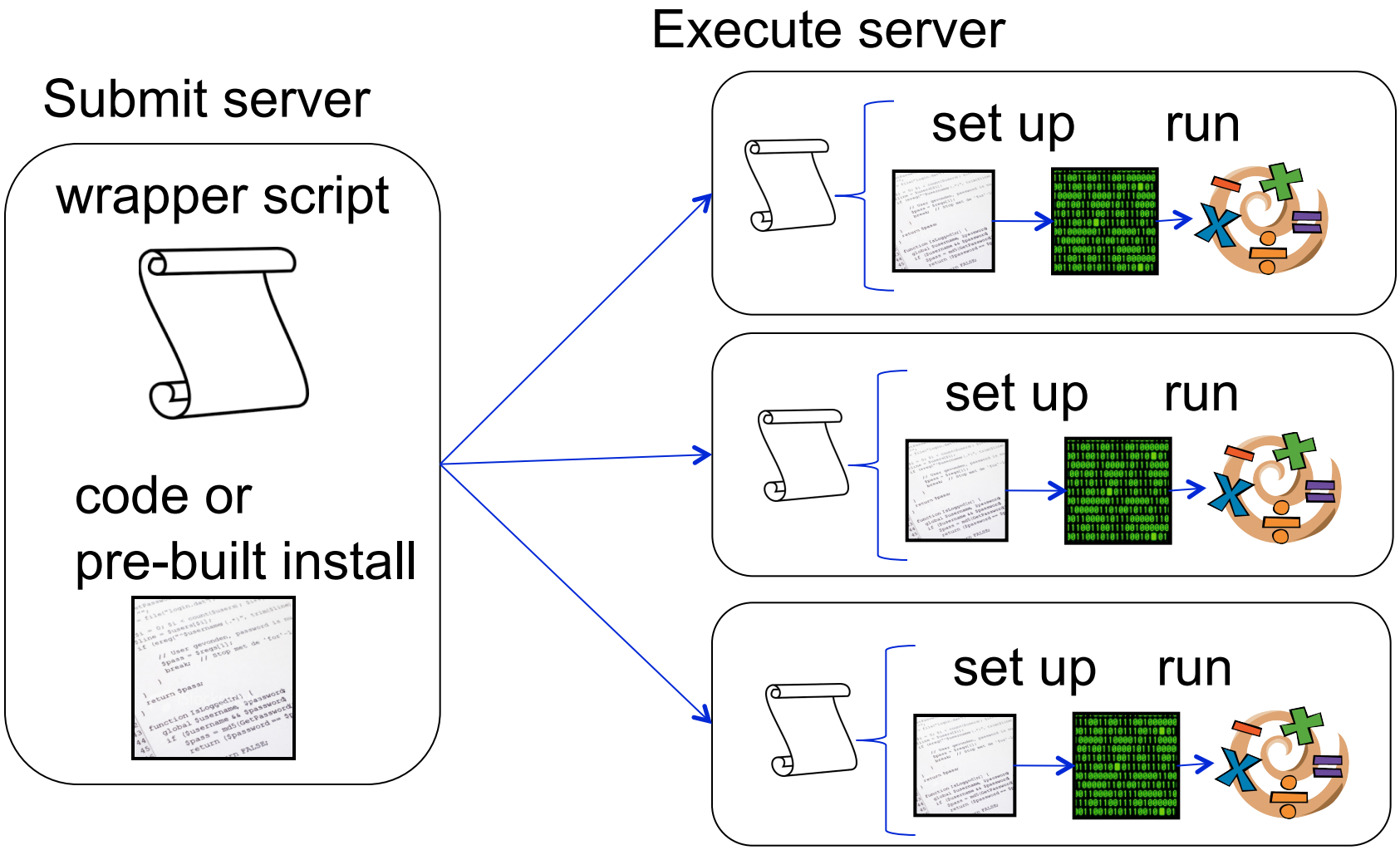
- Good for software that:
  - Can't be statically compiled
  - Uses interpreted languages (Matlab, Python, R)
  - Any software with instructions for local installation
- Method: write a wrapper script
  - Contains a list of commands to execute
  - Typically written in bash or perl (usually common across operating systems/versions)

Wrapper scripts

- Set up software in the working directory
  - Bring along pre-built software and unpack

    or
  - Bring along source and install a fresh copy
- Run software
- Besides software: manage data/files in the working directory
  - Move or rename output
  - Delete installation files before job completion

Wrapper script workflow



When to pre-build?

Pre-built installation

- Install once, use in multiple jobs
- Faster than installing from source code within the job
- Jobs must run on a computer similar to where the program was built

Install with every job

- Computers must have appropriate tools (compilers, libraries) for software to install
- Can run on multiple systems, if these requirements are met
- Longer set-up time

Preparing your code

- Where do you compile code? Pre-build code? Test your wrapper script?
- Guiding question: how computationally intensive is the task?
  - Computationally intensive (takes more than a few minutes, as a rule of thumb)
    - Run as interactive job, on a private computer/server, or with a queued job
  - Computationally light (runs in few minutes or less)
    - Run on submit server (or above options, if desired)

Open Science Grid

Exercises

- Software is a compiled binary
  - Exercise 1.1: statically compile code and run (C code)
  - Exercise 1.2: download and run pre-compiled binary (BLAST)
- Install software with each job
  - Exercise 1.3: create a pre-built installation, and write a wrapper script to unpack and run software (GROMACS)

Questions?

- Feel free to contact me:
  - ckoch5@wisc.edu
- Now: Hands-on Exercises
  - 9:30-10:30am
- Next:
  - 10:30-10:45am: Break
  - 10:45am-12:15pm: Other research software considerations: licenses and interpreted languages
  - 12:15-1:15pm: Lunch