Computational Resources at UCSB

Creating Connections to Enable Groundbreaking Research
(A lofty goal, to be sure)

UC SANTA BARBARA LIBRARY

ENTERPRISE TECHNOLOGY SERVICES | University of California
Santa Barbara

CALIFORNIA NANOSYSTEMS INSTITUTE

CENTER FOR SCIENTIFIC COMPUTING
UC SANTA BARBARA
Computational Resources at UCSB

Quick Aside....
Computational Resources at UCSB

Quick Aside....

Home Sweet Home
Important Points from my presentation

We’re here to help you do your research.
Computational Resources at UCSB

Important Points from my presentation

We’re here to help you do your research.

This is just an overview of our areas. There is more available at UCSB than what is presented today.
Computational Resources at UCSB

Important Points from my presentation

We’re here to help you do your research.

This is just an overview of our areas. There is more available at UCSB than what is presented today.

Know your data and be aware of UCs Human Subjects Guidelines.
Computational Resources at UCSB

Important Points from my presentation

We’re here to help you do your research.

This is just an overview of our areas. There is more available at UCSB than what is presented today.

Know your data and be aware of UCs Human Subjects Guidelines.

The Center for Scientific Computing has resources you can use for free.
Computational Resources at UCSB

Important Points from my presentation

We’re here to help you do your research.

This is just an overview of our areas. There is more available at UCSB than what is presented today.

Know your data and be aware of UCs Human Subjects Guidelines.

The Center for Scientific Computing has resources you can use for free.

Some Gaps and Future Projects
Computational Resources at UCSB

Who are We?

Center for Scientific Computing (CNSI, MRL, ETS)

Fuzzy Rogers – fuz@mrl.ucsb.edu
  High Performance Computing Sys Admin – MRL, MC-CAM, DowMI

Paul Weakliem – weakliem@cnsi.ucsb.edu
  Co-Director Center for Scientific Computing

Burak Himmetoglu - bhimmetoglu@ucsb.edu
  UCSB SuperComputing Consultant and XSEDE Campus Champion

B Library

Stephanie Tulley - stulley@ucsb.edu
  Acting Director of the Interdisciplinary Research Collaboratory
Human Subjects Data Considerations

A Human Subject means a living individual about whom an investigator (whether professional or student) conducting research obtains (1) data or (2) identifiable private information through intervention or interaction with the individual.” – 45 CFR 46.102(f)

UCSB’s Office of Research Human Subjects Contacts
Melissa Warren - warren@research.ucsb.edu
Dorin Donohoe - donohoe@research.ucsb.edu

Use their expertise!
Human Subjects Data Considerations

Twitter as an Example

I think that it is safe to say that Twitter is public information (see Twitter policies etc. attached) and that permissions should not be needed when gathering data. (This is not true with other social media sites.)” — Melissa Warren email to researcher, 5/2014

Sounds good, however...
Human Subjects Data Considerations

Using social media in your research
www.apa.org/gradpsych/2011/11/social-media.aspx; see comment “it's ethically unclear whether Facebook, Twitter or other types of postings count as public or private behavior”

Ethical Decision-Making and Internet Research/Recommendation from the AoIR Ethics Working Committee
aoir.org/documents/ethics-guide/; see section “Data(Text)/Persons” for comment “‘Will capturing someone else’s Tweets cause them harm?’”

International Journal of Internet Resources (http://ijire.net/)

Internet Research – Presentation to the SACHRP
www.hhs.gov/ohrp/sachrp/mtgings/2012%20Jul%20Mtg/pptinternetresearch.pdf; very interesting presentation; see slides 26 & 27 for discussion of privacy and suggestions)

General reference: IRB Review of the Use of Social Media in Research

Melissa Warren, email to researcher 5/2014
then she recommends:

When provided these resources, we recommend approaching the use of data obtained from Twitter in a conservative manner.

We recommend **anonymizing** the data (de-identifying without retaining a key code so that posts cannot be re-linked). Since Twitter states at the top of their Privacy Policy that “What you say on Twitter may be viewed all around the world instantly” it appears reasonable to equate this to public behavior. **Thus, if capturing a person’s Tweets does not cause them harm (see resource #3 above), and the proposed research is greater than minimal risk then the research may be eligible for exemption** under 45 CFR 46.101(b)(2). However, if it is not possible to anonymize the data, then it’s possible that a minimal risk study using Twitter data may be eligible for exemption under 45 CFR 46.101(b)(4) “Research involving the collection of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.” If category (b)(4) is utilized for exempt determination, to avoid harm to subjects, it’s reasonable for the HSC to require de-identification to the conservative side.” – Mellisa Warren, email to researcher 5/2014
then she recommends:

Having provided these resources, we recommend approaching the use of data obtained from Twitter in a conservative manner.

We recommend *anonymizing* the data (de-identifying without retaining a key code so that posts cannot be linked). Since Twitter states at the top of their Privacy Policy that “What you say on Twitter may be read all around the world instantly” it appears reasonable to equate this to public behavior. *Thus, if* a person’s Tweets *does not cause them harm* (see resource #3 above), and the proposed risk is greater than minimal risk then the research may be eligible for exemption under 45 CFR 46.101(b)(2). However, it is not possible to anonymize the data, then it’s possible that a minimal risk study using Twitter source data may be eligible for exemption under 45 CFR 46.101(b)(4) “Research involving the collection of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.” If category (b)(4) is utilized for exempt determination, to avoid harm to subjects, it’s reasonable for the HSC to require de-identification on the conservative side.” – Mellisa Warren, email to researcher 5/2014
Computational Resources at UCSB

Human Subjects Data Considerations

Follow the protocol of the Belmont Report’s 3 Ethical Principles:

- Respect for persons
- Beneficence – obligation to protect persons from harm...
- Justice – obligation to ensure that the benefits and burdens of research are distributed fairly

Anecdotes

- Uber’s Rides of Glory Blog Post
  The most one-night stands originated in Chinatown, the Mission, Downtown, Bernal Heights, Russian Hill, the Marina, and the Castro-Upper Market area.

- Cornell/UCSF’s Facebook Experiment
Human Subjects Data Considerations

A Few Best Practices...

- If grabbing data from a website that you do not have to log in to but requires you to adhere to its Terms and Conditions (as nearly all do), print out/record those Terms and Conditions at the time of data collection.

- If grabbing data from a website that requires logins or authentication tokens, anonymize the data and/or contact Human Subjects.

- Consider your study from the perspective of your subjects. If they were identified, how might they react?
Secure Compute Research Environment

(Scary? :)

BER and ETS created a secure virtualized research environment where you can do research on sensitive data.

Satisfies Data Security Plan requirements for funding agencies. In a nutshell, it’s the secure computer behind a locked door that is not connected to the internet.

http://www.ets.ucsb.edu/services/secure-compute-research-environment
Computational Resources at UCSB

The Center for Scientific Computing

An Academic Senate recognized center for campus-wide High Performance Computing

Compute Clusters – 2 of which are available to all campus

Beowulf Clusters are groups of computers of similar OS’s connected by specialized high speed networking and high performance file systems.
Computational Resources at UCSB

Our Clusters

QSR (quite old)
32 nodes, 4 core 2.2GHz, 8 GB RAM/node
Myrinet Interconnect, 4 TB storage, CentOS 6.6

Knot
119 standard nodes, 1400 cores 2.6GHz (12-20 cores/node)
48-64 GB RAM/node, CentOS 6.2
6 GPU nodes, 2 NVIDIA M2050s per node, 1 8-Phi node
4 FAT nodes, 32 cores/node 2.6GHz, 768G-1TB RAM/node
all nodes use Infiniband interconnect and 60 TB HP storage
How can they be used?

Many use cases for both QSR and Knot

Serial Jobs (short or long running)

MPICH / Multi-threaded Jobs

Parallel MPI jobs
  Often require advanced programming and/or parallel capable binaries

Note that compute nodes typically cannot see the internet – the data to process has to be on the local high performance storage.
Computational Resources at UCSB

How can they be used?

Knot has specialized compute nodes

Fat Nodes – 1TB RAM
  Perfect for really big memory jobs (Agent Based Modeling)
  Large MatLab arrays

GPU Nodes – GPUs have hundreds of small instruction cores
  Many codes beginning to take advantage of GPUs
  Requires specialized programming (CUDA)

Phi Nodes – Intel’s answer to GPUs
  Currently in its infancy
Computational Resources at UCSB

How do you use them?

Account Requests and Documentation available at:

http://csc.cnsi.ucsb.edu

Clusters are Linux based, using queuing systems (Torque/Maui) jobs submitted to queues to run unattended.

[fuz@knot ~]$ more Rsnowtest.job
#!/bin/bash
#PBS -l nodes=4:ppn=4

cd $PBS_O_WORKDIR
cat $PBS_NODEFILE > nodes

mpirun -np 1 /sw/bin/R --no-save < Rsnow.R

[fuz@knot ~]$
Computational Resources at UCSB

[fuz@knot ~]$ more nodes
node3
node3
node3
node3
node3
node3
node3
node50
node50
node50
node50
node50
node56
node56
node56
node56
node56
node63
node63
node63
node63
node63
[fuz@knot ~]$
[fuz@knot ~]$ more Rsnow.R
library(Rmpi)
library(snow)

# Initialize SNOW using MPI communication. The first line will get the
# number of MPI processes the scheduler assigned to us. Everything else
# is standard SNOW

np <- mpi.universe.size()
cluster <- makeMPLICluster(np)

# Print the hostname for each cluster member
sayhello <- function()
{
  info <- Sys.info()[c("nodename", "machine")]
  paste("Hello from", info[1], "with CPU type", info[2])
}

names <- clusterCall(cluster, sayhello)
print(unlist(names))

# Compute row sums in parallel using all processes,
# then a grand sum at the end on the master process
parallelSum <- function(m, n)
{
  A <- matrix(rnorm(m*n), nrow = m, ncol = n)
  row.sums <- parApply(cluster, A, 1, sum)
  print(sum(row.sums))
}

parallelSum(500, 500)

stopCluster(cluster)
mpi.exit()
Type 'q()' to quit R.

```r
> library(Rmpi)
> library(snow)
> # Initialize SNOW using MPI communication. The first line will get the
> # number of MPI processes the scheduler assigned to us. Everything else
> # is standard SNOW
> np <- mpi.universe.size()
> cluster <- makeMPICluster(np)
> 16 slaves are spawned successfully. 0 failed.
>
> # Print the hostname for each cluster member
> sayhello <- function()
+ {
+   info <- Sys.info()[c("nodename", "machine")]
+   paste("Hello from", info[1], "with CPU type", info[2])
+ }
>
> names <- clusterCall(cluster, sayhello)
> print(unlist(names))
[1] "Hello from node3 with CPU type x86_64"
[2] "Hello from node3 with CPU type x86_64"
[3] "Hello from node3 with CPU type x86_64"
[4] "Hello from node50 with CPU type x86_64"
[5] "Hello from node50 with CPU type x86_64"
[6] "Hello from node50 with CPU type x86_64"
[7] "Hello from node50 with CPU type x86_64"
[8] "Hello from node56 with CPU type x86_64"
[9] "Hello from node56 with CPU type x86_64"
[10] "Hello from node56 with CPU type x86_64"
[11] "Hello from node56 with CPU type x86_64"
[12] "Hello from node63 with CPU type x86_64"
[13] "Hello from node63 with CPU type x86_64"
[14] "Hello from node63 with CPU type x86_64"
[15] "Hello from node63 with CPU type x86_64"
[16] "Hello from node3 with CPU type x86_64"
>
> # Compute row sums in parallel using all processes,
> # then a grand sum at the end on the master process
> parallelSum <- function(m, n)
+ {A <- matrix(rnorm(m*n), nrow = m, ncol = n)
+  row.sums <- parApply(cluster, A, 1, sum)
+  print(sum(row.sums))
+ }
>
> parallelSum(500, 500)
[1] -1148.675
>
> stopCluster(cluster)
```
How does the cluster run jobs?

Queuing system prioritizes your job
Torque / Maui
When? depends on cluster usage and open cores
Fairshare priorities – the more you use the cluster the less priority you have (calculated over a week)
No restriction on runtimes
Jobs run for minutes, days, weeks, even months
Caveat – Fires, floods, outages, crashing nodes
Checkpointing is critical for long running jobs
Computational Resources at UCSB

What Software is Available?

Software is constantly evolving – Let us know if you need a package installed

Currently Installed on Knot (which might be of interest to you): R, Rstudio, Rsnow, Rmpi, and R packages as necessary

Agent Based Modeling – FLAME (mpi), NetLogo, Rsimecol

MatLab, Mathematica

Python, python for mpi, standard scripting languages

Use X2Go for software requiring GUIs
Condo Clusters

Professors, Research Groups, or Departments buy nodes

CSC pays for infrastructure & management

Infrastructure = disk & backup, networking, racks, etc.

Node types/OS are restricted to fit in with current condos

Fairshare = buy-in / condo size + 5%

3 Condos in use by about 14 different groups

Lattice (60 nodes), Guild (60 nodes), Braid (93 nodes)
Computational Resources at UCSB

Storage

- Knot - ~60TB High Speed, 250 TB standard and shared
  Both backed up weekly

Globus Endpoint ( [http://globusonline.org](http://globusonline.org) )

Knot’s storage is at ucsb#knot-storage
Both /home (60TB) and /csc/central (250TB) available
Perfect for large quantities of data
Restores broken connections

Box.com
Not at production speeds yet (davfs yields 1 MB/s)
Computational Resources at UCSB

How you can help us

If you use the CSC clusters please acknowledge us in your publications:

acknowledge support from the Center for Scientific Computing at the CNSI and MRL: an NSF MRSEC (DMR-1121053) and NSF CNS-0960316

In 5 years, approximately 270 publications acknowledge the use of Knot alone.
Computational Resources at UCSB

Future Projects and Gaps

• Revolution R Open (RRO)
• Greater Reproducibility of results
• Multi-thread capabilities with MKL
• Analytics Gateway (https://wratheomatics.shinyapps.io/tags)
• Text Mining through Web Interface
• Campus Cloud (http://cio.ucsb.edu/resources/UCSBCyberinfrastructurePlan.pdf)
• AWS for all campus researchers (in the works)

Gaps as I see them

• Visualization (none of us are experts)
• Social Sciences representative on informal research sys admins group
Computational Resources at UCSB

XSEDE and Burak Himmetoglu

Once you grow beyond the local resources you will want to utilize the National SuperComputing Centers.

Burak Himmetoglu can assist.

Thanks for listening and contact me with questions!

fuz@mrl.ucsb.edu