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

Computational Resources at UCSB

Quick Aside….

Computational Resources at UCSB

Quick Aside….

Home Sweet Home

Computational Resources at UCSB

Important Points from my presentation

Ve're here to help you do your research.

Computational Resources at UCSB

Important Points from my presentation

Ve're here to help you do your research.

his is just an overview of our areas. There is more available a
CSB than what is presented today.

Computational Resources at UCSB

Important Points from my presentation

Ve're here to help you do your research.

his is just an overview of our areas. There is more available a
CSB than what is presented today.

now your data and be aware of UCs Human Subjects Guidelin

Computational Resources at UCSB

Important Points from my presentation

Ve're here to help you do your research.

his is just an overview of our areas. There is more available a
CSB than what is presented today.

now your data and be aware of UCs Human Subjects Guidelin

he Center for Scientific Computing has resources you can use
ree.

Computational Resources at UCSB

Important Points from my presentation

Ve're here to help you do your research.

his is just an overview of our areas. There is more available a
CSB than what is presented today.

now your data and be aware of UCs Human Subjects Guidelin

he Center for Scientific Computing has resources you can use
ree.

ome 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

Computational Resources at UCSB

Human Subjects Data Considerations

man Subject means a living individual about whom an
stigator (whether professional or student) conducting researc
ains (1) data or (2) identifiable private information through
rvention or interaction with the individual." – 45 CFR 46.102(f

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

se use their expertise!

Computational Resources at UCSB

Human Subjects Data Considerations

ter as an Example

k that it is safe to say that Twitter is public information (see Twitter policies etc. attached) and that
sions should not be needed when gathering data. (This is not true with other social media sites.)"
a Warren email to researcher, 5/2014

nds 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
ook, Twitter or other types of postings count as public or private behavior"
thical Decision-Making and Internet Research \ Recommendation from the AoIR Ethics Working
ittee
aoir.org/documents/ethics-guide/; see section "Data(Text)/Persons" for comment ""Will capturin
's Tweets cause them harm?"
nternational Journal of Internet Resources (http://ijire.net/)
nternet Research – Presentation to the SACHRP
www.hhs.gov/ohrp/sachrp/mtgings/2012%20Jul%20Mtg/pptinternetresearch.pdf; very interestin
tation; see slides 26 & 27 for discussion of privacy and suggestions)
General reference: IRB Review of the Use of Social Media in Research
www.quorumreview.com/wp-content/uploads/2012/12/IRB-Review-of-the-Use-of-Social-Media-i
ch_Gearhart_Quorum-Review_Monitor_2012_12_01.pdf"
sa Warren, email to researcher 5/2014

Human Subjects Data Considerations

then she recommends:

g provided these resources, we recommend approaching the use of data obtained from Twitter in
vative manner.

e recommend **anonymizing** the data (de-identifying without retaining a key code so that posts car
nked). Since Twitter states at the top of their Privacy Policy that "What you say on Twitter may be
l all around the world instantly" it appears reasonable to equate this to public behavior. **Thus, if
ing a person's Tweets does not cause them harm (see resource #3 above), and the proposed res
reater than minimal risk then the research may be eligible for exemption** under 45 CFR 46.101(b
ever, it is not possible to anonymize the data, then it's possible that a minimal risk study using Tw
rce data may be eligible for exemption under 45 CFR 46.101(b)(4) "Research involving the collectio
of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sou
blicly available or if the information is recorded by the investigator in such a manner that subjects
be identified, directly or through identifiers linked to the subjects." If category (b)(4) is utilized fo
t determination, to avoid harm to subjects, it's reasonable for the HSC to require de-identification
the conservative side." – Mellisa Warren, email to researcher 5/2014

Computational Resources at UCSB

Human Subjects Data Considerations

then she recommends:

g provided these resources, we recommend approaching the use of data obtained from Twitter in
vative manner.

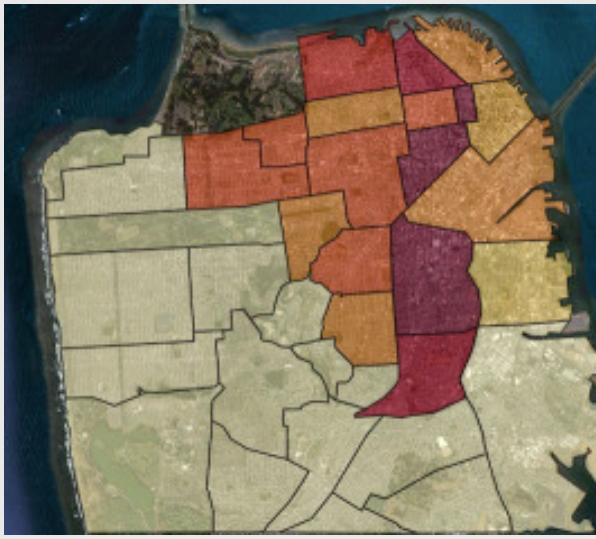
e recommend **anonymizing** the data (de-identifying without retaining a key code so that posts car
nked). Since Twitter states at the top of their Privacy Policy that "What you say on Twitter may be
l all around the world instantly" it appears reasonable to equate this to public behavior. **Thus, if
ing a person's Tweets does not cause them harm (see resource #3 above), and the proposed res
reater than minimal risk then the research may be eligible for exemption** under 45 CFR 46.101(b
ever, it is not possible to anonymize the data, then it's possible that a minimal risk study using Tw
rce data may be eligible for exemption under 45 CFR 46.101(b)(4) "Research involving the collectio
of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sou
blicly available or if the information is recorded by the investigator in such a manner that subjects
be identified, directly or through identifiers linked to the subjects." If category (b)(4) is utilized fo
t determination, to avoid harm to subjects, it's reasonable for the HSC to require de-identification
the conservative side." – Mellisa Warren, email to researcher 5/2014

Computational Resources at UCSB

Human Subjects Data Considerations

ow 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 o research are distributed fairly



cdotes

ber'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.

ornell/UCSF's Facebook Experiment

Human Subjects Data Considerations

w Best Practices…

grabbing data from a website that you do not have to log in t
ut requires you to adhere to its Terms and Conditions (as nea
ll do), print out/record those Terms and Conditions at the tim
f data collection.

grabbing data from a website that requires logins or
uthentication tokens, anonymize the data and/or contact
uman Subjects.

onsider your study from the perspective of your subjects. If th
vere identified, how might they react?

Computational Resources at UCSB

Secure Compute Research Environment

(Scary? :)

BER and ETS created a secure virtualized research environme

here you can do research on sensitive data.

tisfies Data Security Plan requirements for funding agencies.

a nutshell, it's the secure computer behind a locked door tha

ot connected to the internet.

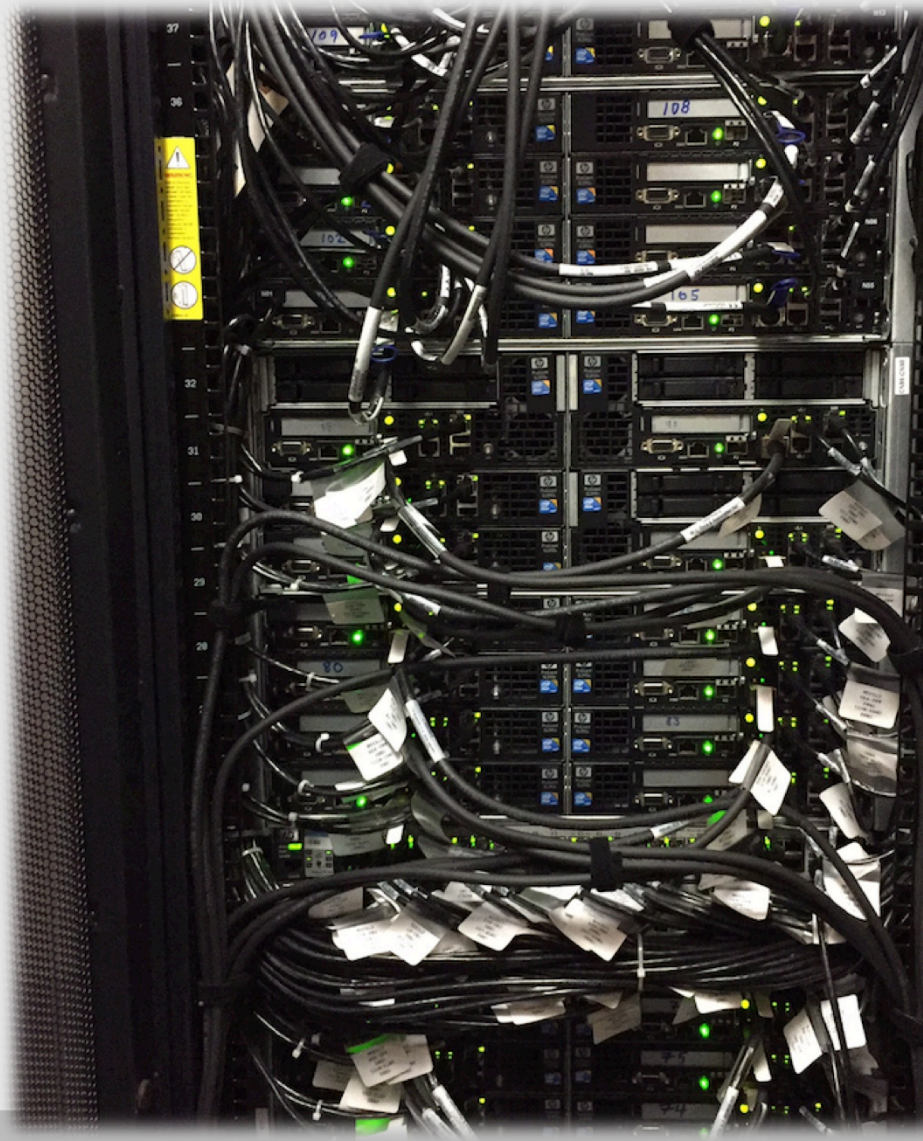
//www.ets.ucsb.edu/services/secure-compute-research-environm

Computational Resources at UCSB

The Center for Scientific Computing

n Academic Senate recognized
enter for campus-wide High
erformance Computing
Compute Clusters – 2 of which
re available to all campus

wulf Clusters are groups of
puters of similar OS's connected
pecialized high speed networking
high performance file systems.



Computational Resources at UCSB

Our Clusters

R (quite old)

- 32 nodes, 4 core 2.2GHz, 8 GB RAM/node
- Myrinet Interconnect, 4 TB storage, CentOS 6.6

ot

- 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

Computational Resources at UCSB

How can they be used?

ny 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

te that compute nodes typically cannot see the internet – the
a to process has to be on the local high performance storage.

Computational Resources at UCSB

How can they be used?

- ot 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?

count Requests and Documentation available at:
http://csc.cnsi.ucsb.edu

sters 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
node50
node50
node50
node50
node56
node56
node56
node56
node63
node63
node63
node63
[fuz@knot ~]$
```

Computational Resources at UCSB

```
[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 <- makeMPIcluster(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()
[fuz@knot ~]$
```

Computational Resources at UCSB

```
Type 'q()' to quit 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)
[1] 1
```

Computational Resources at UCSB

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?

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

rrently 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

e X2Go for software requiring GUIs

Computational Resources at UCSB

Condo Clusters

essors, Research Groups, or Departments buy nodes

pays for infrastructure & management

nfrastructure = 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

ot - ~60TB High Speed , 250 TB standard and shared

- Both backed up weekly

bus Endpoint (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

x.com

- Not at production speeds yet (davfs yields 1 MB/s)

Computational Resources at UCSB

How you can help us

ou use the CSC clusters please acknowledge us in your publications:

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

years, approximately 270 publications acknowledge the use of Knot alone.

National Science Foundation
WHERE DISCOVERIES BEGIN

Computational Resources at UCSB

Future Projects and Gaps

lution R Open (RRO)
eater Reproducibility of results
ulti-thread capabilities with MKL
Analytics Gateway (https://wrathematics.shinyapps.io/tags)
xt Mining through Web Interface
pus Cloud (http://cio.ucsb.edu/resources/UCSBCyberinfrastructurePla
for all campus researchers (in the works)

Gaps as I see them

Visualization (none of us are experts)
l Sciences representative on informal research sys admins group

Computational Resources at UCSB

XSEDE and Burak Himmetoglu

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

Burak Himmetoglu can assist.

Thanks for listening and contact me with questions!
fuz@mrl.ucsb.edu