Introduction to Supercomputing at USD

Doug Jennewein-Director of Research Computing

*Joseph Madison-Research Computing Assistant

Acknowledgements

- University of Oklahoma/Oklahoma State University
- USD – Information Technology Services
- USD – Department of Chemistry
- USD – A&S
- South Dakota's 2015-2020 NIH Institutional Development Award (IDeA) for Networks of Biomedical Research Excellence (INBRE): "South Dakota Biomedical Research Infrastructure Network." NIH-P20GM103443, $14M, USD is lead institution
- XSEDE

USD Research Computing Group

USD Supercomputing Team- Your Partners for Superpowers

- Doug Jennewein – Director of Research Computing
- Joseph Madison – Research Computing Assistant
- Wendi Sapp – Domain Expert, Chemistry
- Adam Erck – Domain Expert, Chemistry

USD Research Computing Group

Your poll will show here

1

Install the app from pollev.com/app

2

Make sure you are in Slide Show mode

Still not working? Get help at pollev.com/app/help

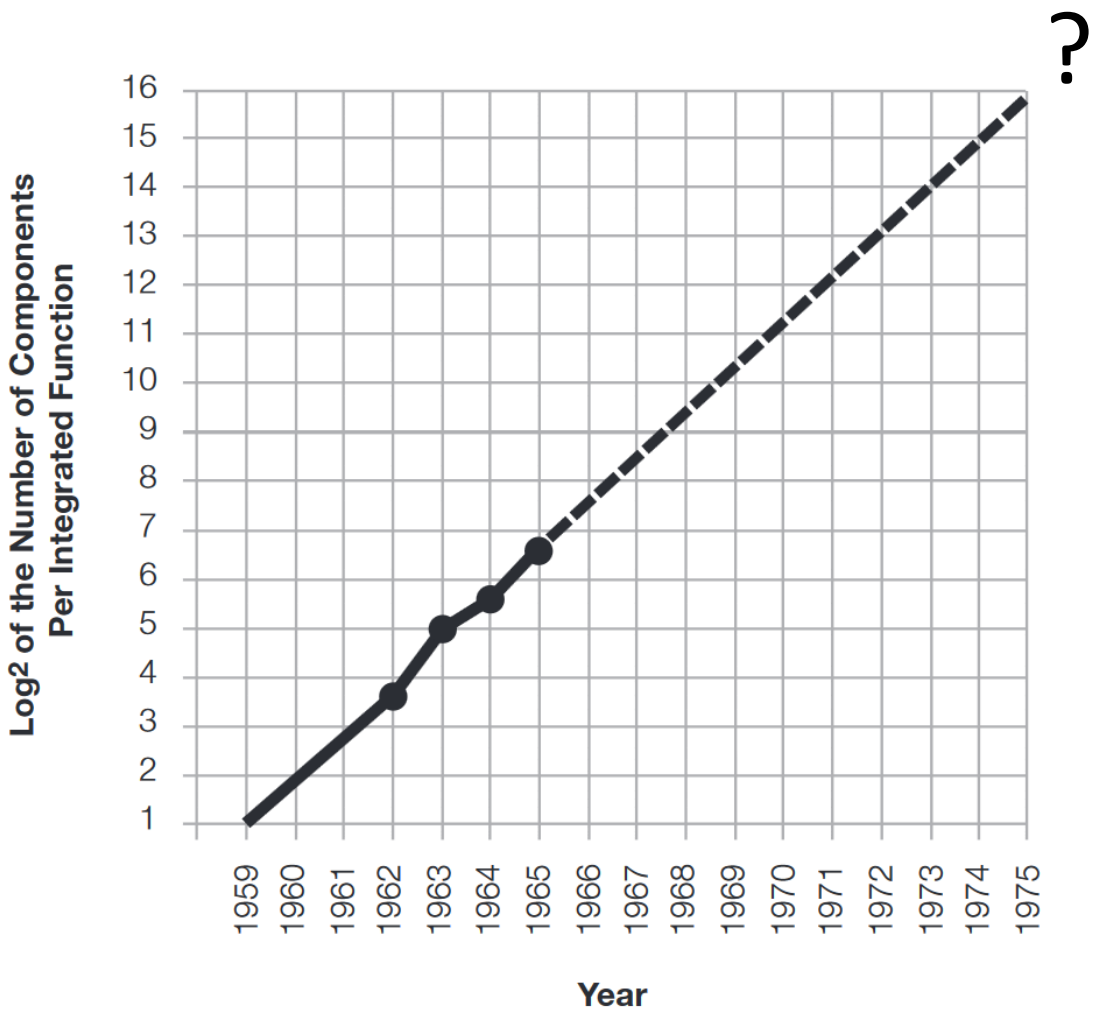
or

Open poll in your web browser

USD Research Computing Group

Moore's Law

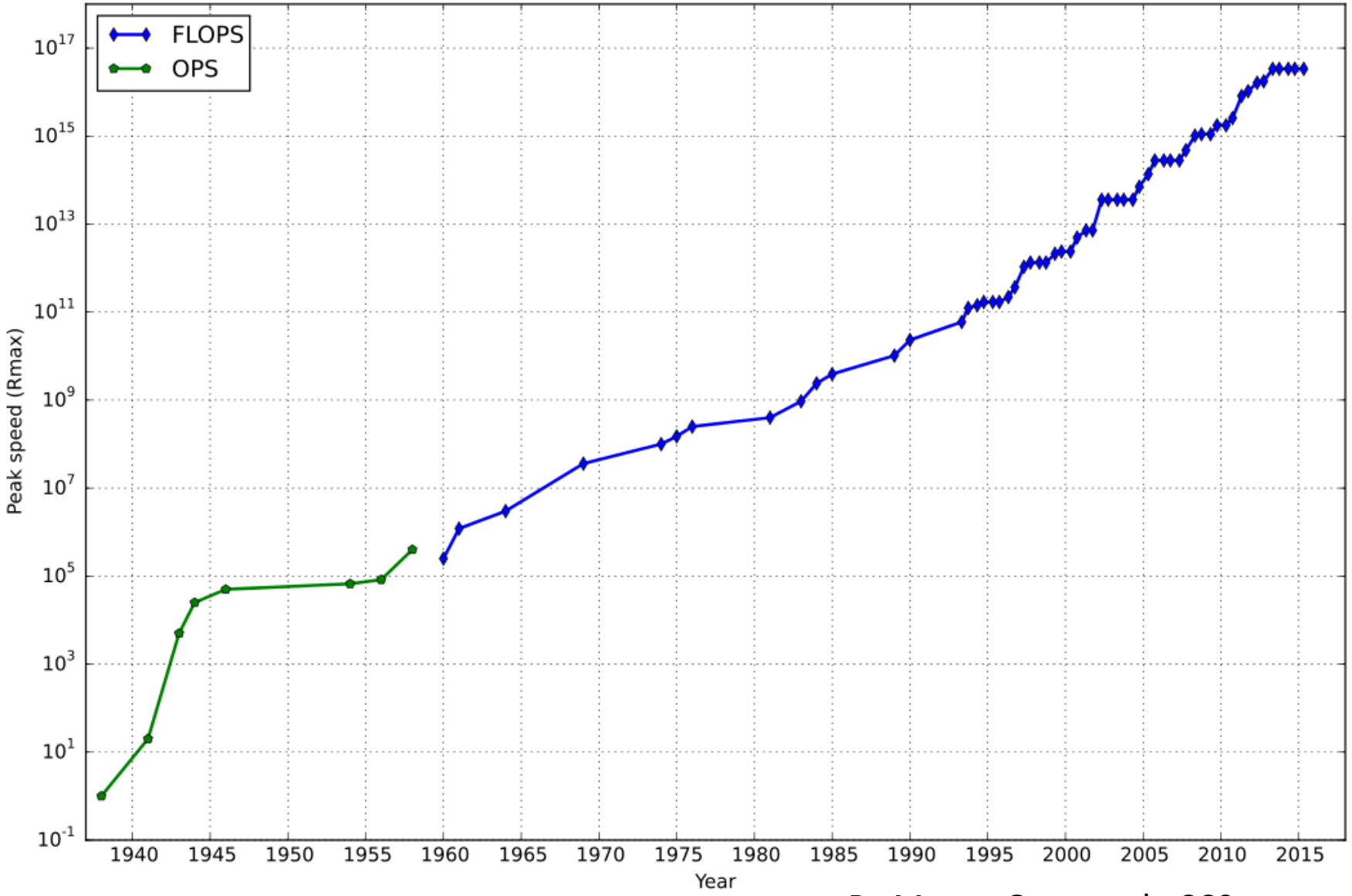
"Integrated circuits will lead to such wonders as home computers…"



USD Research Computing Group

Moore, G. (1965). Cramming More Components Onto Integrated Circuits, Electronics,(38) 8.

Top Supercomputer speed by year



By Morn - Own work, CC0

USD Research Computing Group

What is Supercomputing?

Supercomputing is the **biggest**, **fastest** **computing right this minute**. Likewise, a ***supercomputer*** is one of the biggest, fastest computers right this minute. So, the definition of supercomputing is **constantly changing**.

Rule of Thumb: A supercomputer is typically at least 100 times as powerful as a PC.

Jargon: Supercomputing is also known as

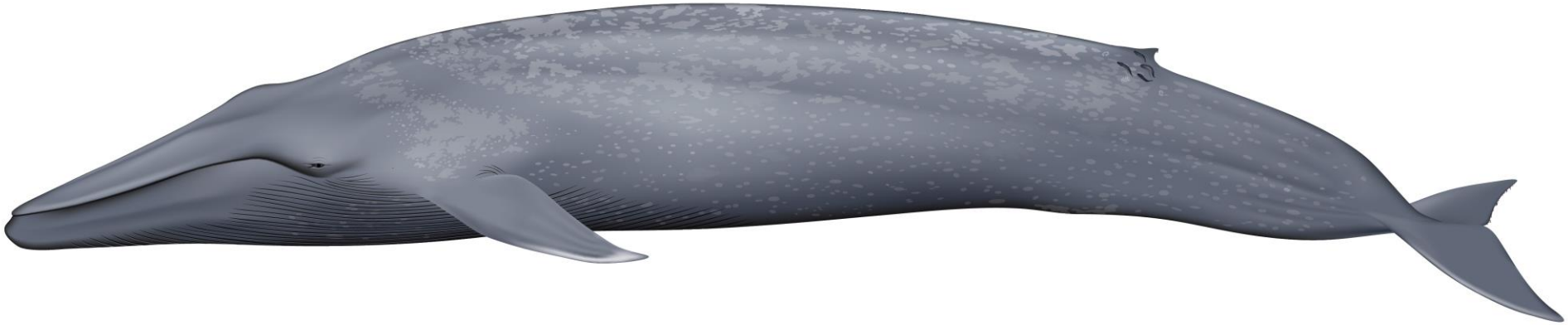
High Performance Computing **(HPC)** or

High End Computing **(HEC**) or

Cyberinfrastructure **(CI).**

USD Research Computing Group

What is Supercomputing About?



USD Research Computing Group

What is Supercomputing About?

- **Size**
  - Problems interesting to researchers need more RAM or disk than you could get on your PC

- **Speed**
  - Don’t want to wait a month for a job to run on a PC, do it on HPC in a few hours

USD Research Computing Group

Why Bother with HPC at All?

- Effective use of HPC takes quite a bit of effort, both learning how and developing software.
- That seems like a lot of trouble to go to just to get your code to run faster.
- It's nice to have a code that used to take a day, now run in an hour. But if you can afford to wait a day, what's the point of HPC?
- Why go to all that trouble just to get your code to run faster?

Why HPC is Worth the Bother

- What HPC gives you that you won't get elsewhere is the ability to do bigger, better, more exciting science. If your code can run faster, that means that you can tackle much bigger problems in the same amount of time that you used to need for smaller problems.
- HPC is important not only for its own sake, but also because what happens in HPC today will be on your desktop in about 10 to 15 years and on your cell phone in 25 years: it puts you ahead of the curve.

USD Research Computing Group

What is a Supercomputer?

- Similar idea to PC, but lots of PCs connected
- Each individual “computer” in cluster is referred to as a node
- Nodes have to “talk”
  - Interconnect, software needed

USD Research Computing Group

What a Supercomputer Looks like

USD Cluster

Nodes

Interconnect

IBM - Sequoia

USD Research Computing Group

Hardware Basics

- Central Processing Unit
- Primary Storage
- Secondary Storage
- Input Devices
- Output Devices

USD Research Computing Group

Central Processing Unit

- Components
  - Control Unit – figures out what to do next
  - Arithmetic/Logic Unit – performs calculations
  - Registers – data being used right now

USD Research Computing Group

Primary Storage

- Main memory
  - RAM
  - Data being used
- Cache
  - Small memory
  - About to be used or have been used recently

USD Research Computing Group

Secondary Storage

- Data that can be used in the future
- Non-volatile storage
- Examples
  - Blu-Ray
  - CD
  - Hard-drive
  - SSD

USD Research Computing Group

Input/Output

- Input
  - Keyboard, mouse
- Output
  - Speakers, printer

USD Research Computing Group

Software

- Command-line interface
  - Cygwin/PuTTY (Windows)
  - Linux terminal
  - Mac terminal
- Domain specific software
- Programming Languages

USD Research Computing Group

Parallelism

- Parallelism in software and hardware
- Hardware
  - Cluster, distributed, massively parallel, multi-core
  - NUMA, UMA
- Software
  - ILP, TLP, DLP
- Architecture considerations

The HPC Learning Curve

- Programming languages – Python, R, Java, C++, Fortran
- Should have a basic understanding of how a computing cluster works
- Learning domain specific software
- Where do I do this?

Contact the USD Research Computing Team!

Access Online Resources!

Talk to your peers!

Come to our Upcoming Seminar Series!

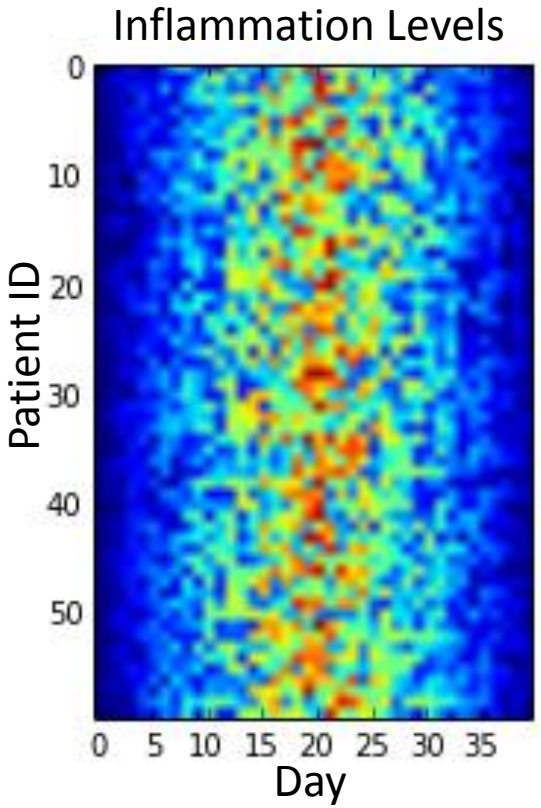
USD Research Computing Group

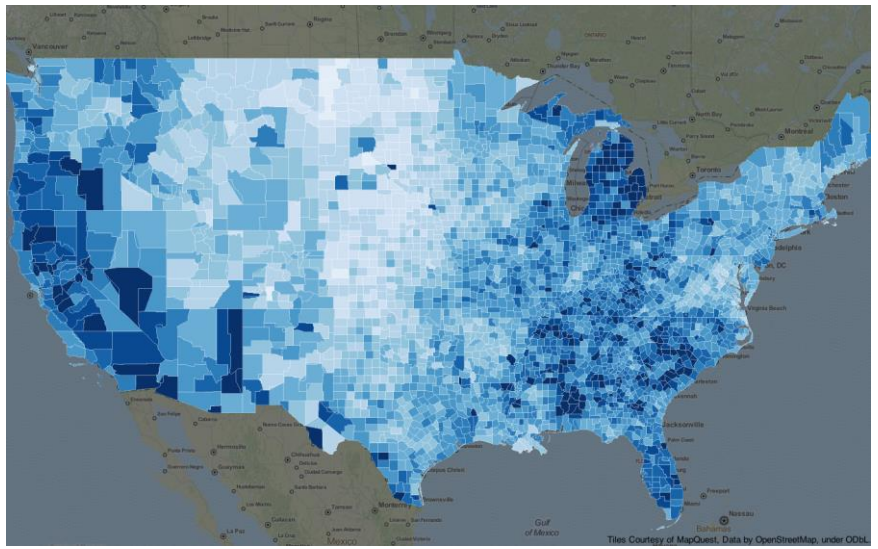
Are there reasons to learn a computer language other than for using the HPC?

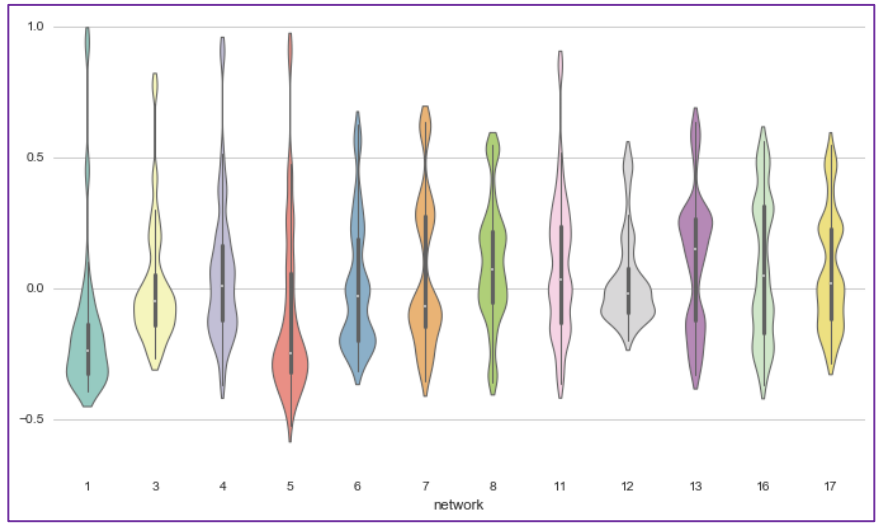
Yes!!!

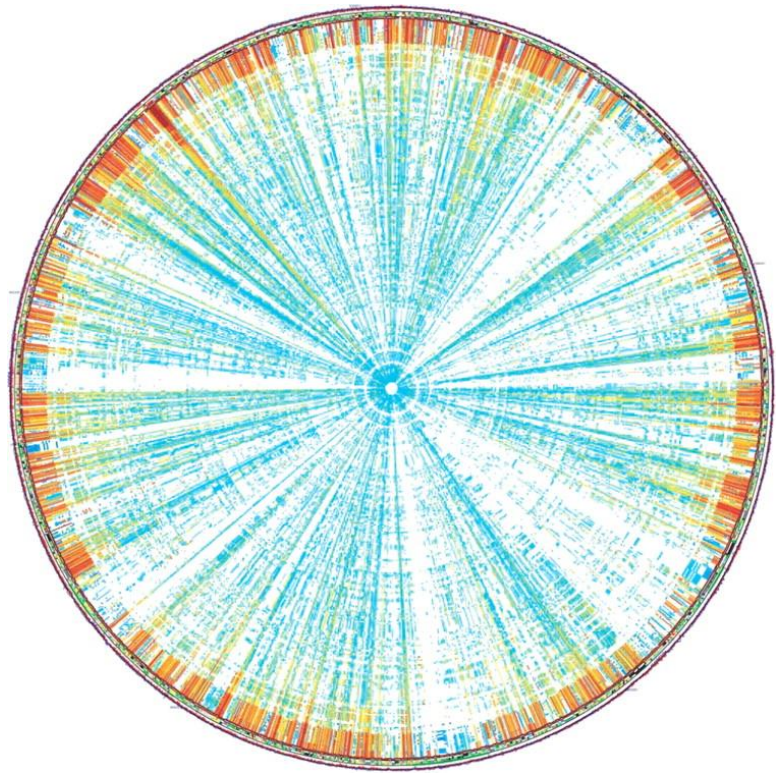
USD Research Computing Group

Data visualization using basic Python code



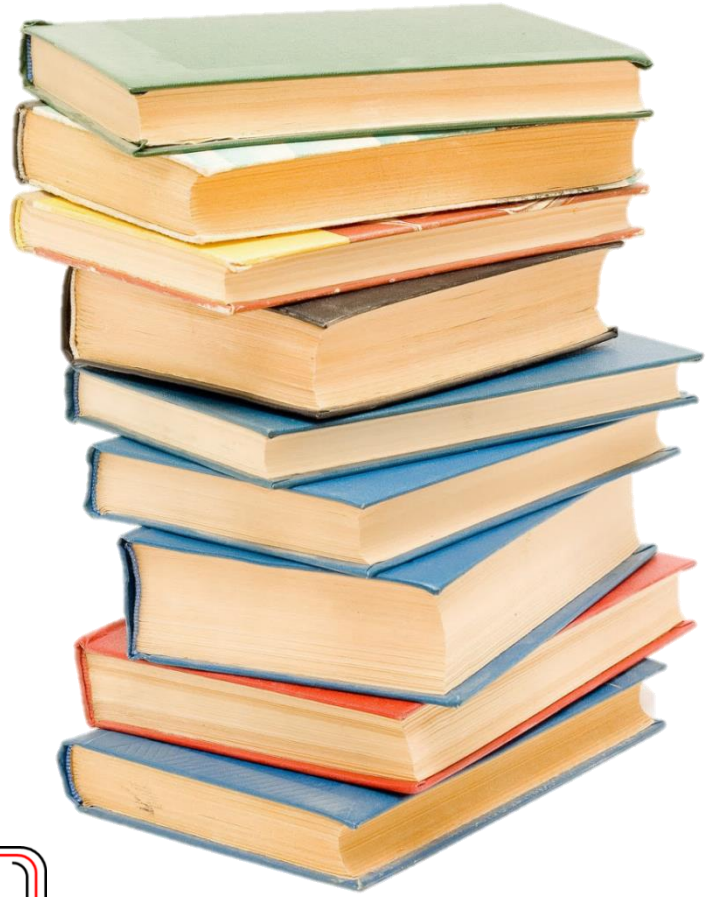






USD Research Computing Group

How learning some basic programming can automate your tedious tasks



Problem:

I am writing a report about gender roles and word associations from the early 1800s. I have chosen to look at the six Jane Austen novels as my source of text.

One Solution:

Open each text file and use the ‘Find’ feature to search for words that I choose. I could use a spreadsheet to track words that are adjacent to the selected words.

A Different Way:

Use R to perform text mining in order to get word repetitions, word associations, and multiple word visualizations.

USD Research Computing Group

How learning some basic programming can automate your tedious tasks

Problem:

I have a folder full subfolders with multiple data files in each subfolder. I need to feed these files into a program but they are the wrong format. They are also named inconsistently.

One Solution:

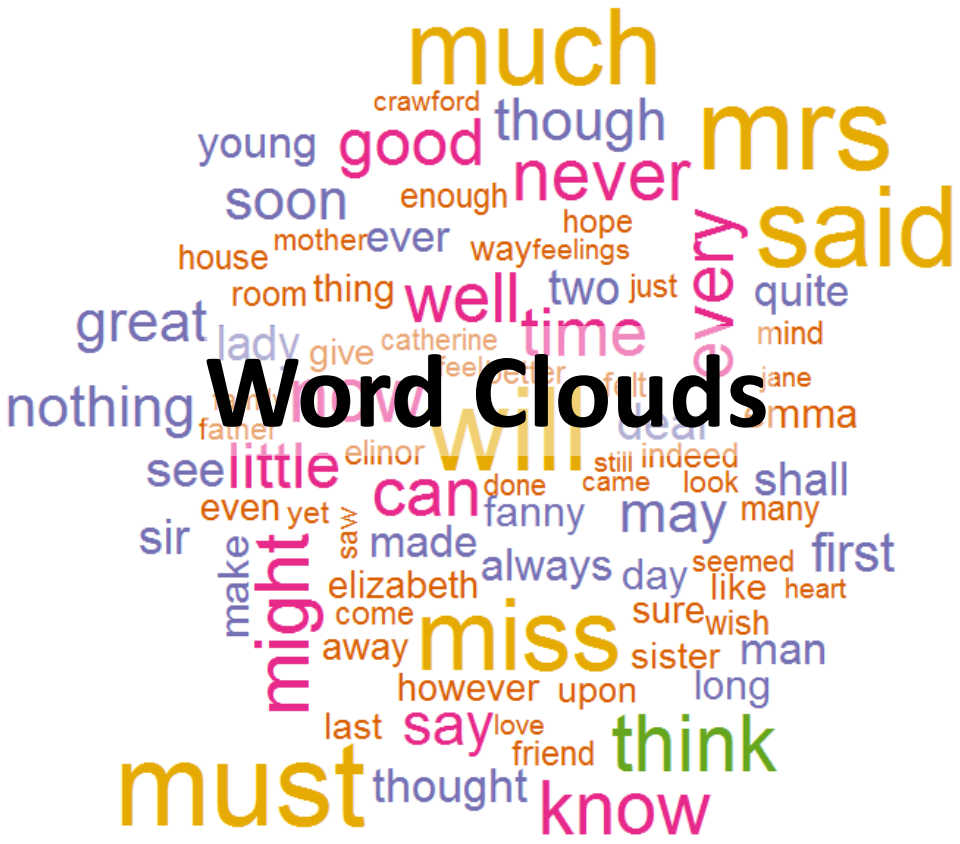
Open each folder and file individually in order to resave in the appropriate format. Rename each file one-by-one. This could take hours depending on the number of files in the folder.

A Different Way:

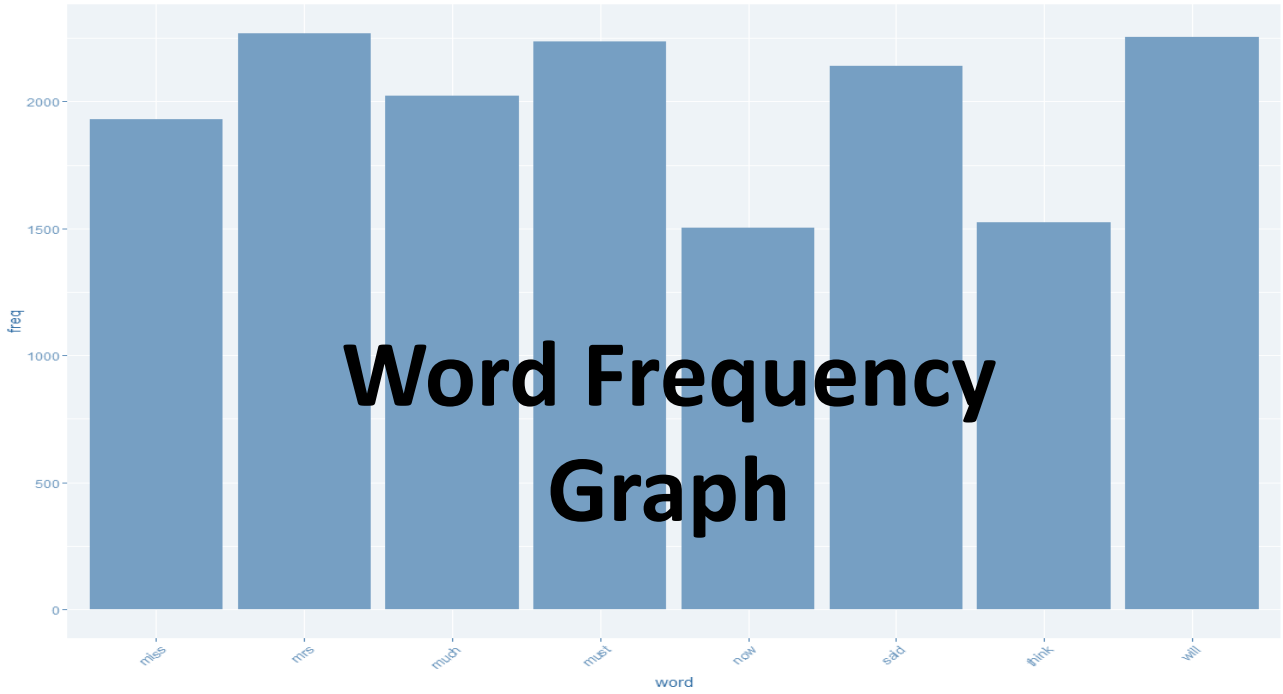
Write a few lines of code that will convert the files and rename them. Writing the code takes 5-10 minutes and the code executes in only a few minutes.

USD Research Computing Group

Jane Austen text analysis results



Word Clouds



Word Frequency Graph

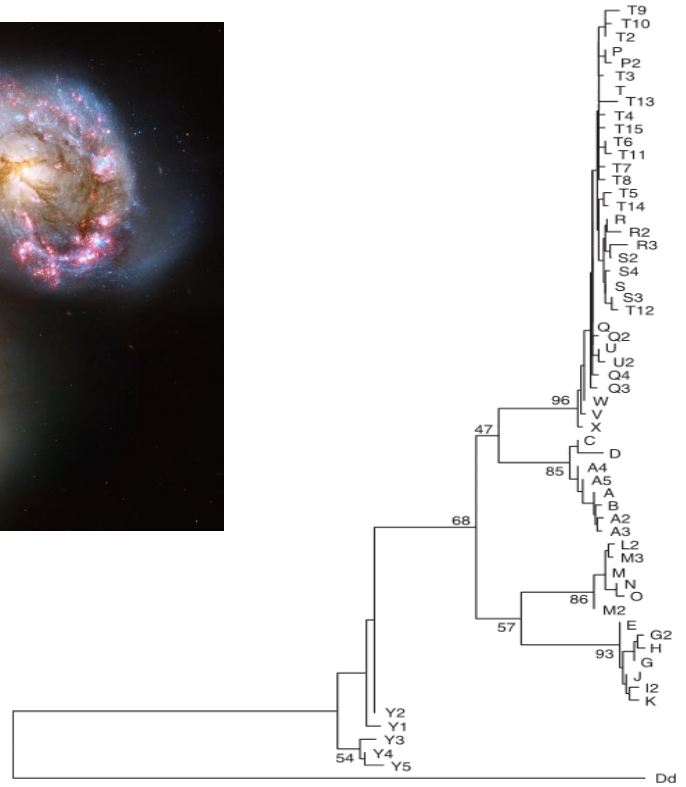
```
$mrs
 situation        lik    mistake
      0.99       0.98       0.98
      call    collect comfortable
      0.97       0.97       0.97
   improve    parties
      0.97       0.97
```

Word Associations

USD Research Computing Group

What Things can I use HPC for?

- Data Mining
  - Gene sequencing
  - Signal processing
- Visualization
  - Lots of data
- Simulation
  - Galaxy formation
  - Evolution
  - Weather



Vianna, J. A., et al. (2006). Phylogeography, phylogeny and hybridization in trichechid sirenians: implications for manatee conservation. Molecular Ecology, 15(2), 433-447.

USD Research Computing Group

Your poll will show here

1

Install the app from
pollev.com/app

2

Make sure you are in
Slide Show mode

Still not working? Get help at pollev.com/app/help
or
Open poll in your web browser

USD Research Computing Group

Applications of HPC in Chemistry

- Prediction of energies, molecular structures, and vibrational frequencies of molecular systems
- Study molecules and reactions under a wide range of conditions, including both stable species and transition structures.
- Electronic structure investigations of many body systems
  - DFT, Kohn-Sham, Hartree-Fock, etc.

Applications of HPC in Chemistry-Cont.

- Supercomputing in materials sciences
  - Data mining
  - The Materials Project – Berkeley Lab

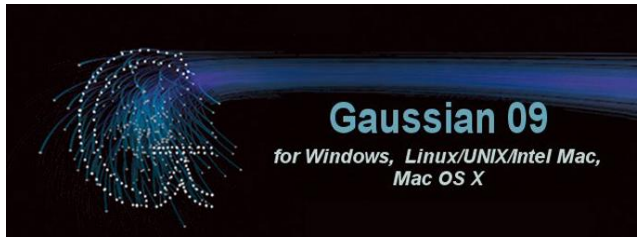
USD Research Computing Group

Jain, A., Persson, K. A., & Ceder, G. (2016). Research Update: The materials genome initiative: Data sharing and the impact of collaborative ab initio databases. APL Materials, 4(5), 053102.

A. Jain*, S.P. Ong*, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson (*=equal contributions). The Materials Project: A materials genome approach to accelerating materials innovation. APL Materials, 2013, 1(1), 011002.

Applications of HPC in Chemistry-Cont.

- Gaussian 09
- NWChem
- VASP
- Many others that can be optimized for use on HPC



USD Research Computing Group

Valiev, M., et al. (2010). NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations. Computer Physics Communications, 181(9), 1477-1489.

Applications of HPC in Physics

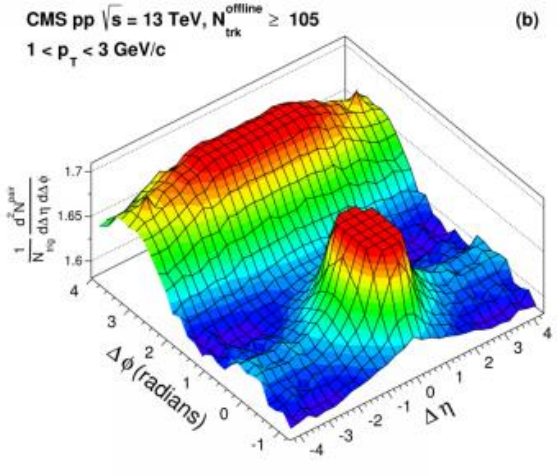
- Astronomical/cosmological data
- Detection Experiments
  - LUX (Large Underground Xenon) experiment looking for dark matter
  - IceCube – Neutrino Detection
- Simulations, models

LUX Experiment – Lead, SD

USD Research Computing Group

Applications of HPC in Physics-Cont.

- Root
  - Big data processing, analysis, visualization
- Geant4
  - Toolkit using monte carlo methods



Agostinelli, S., et al. (2003). GEANT4—a simulation toolkit. Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 506(3), 250-303.

Antcheva, I., et al. (2009). ROOT—A C++ framework for petabyte data storage, statistical analysis and visualization. Computer Physics Communications, 180(12), 2499-2512.

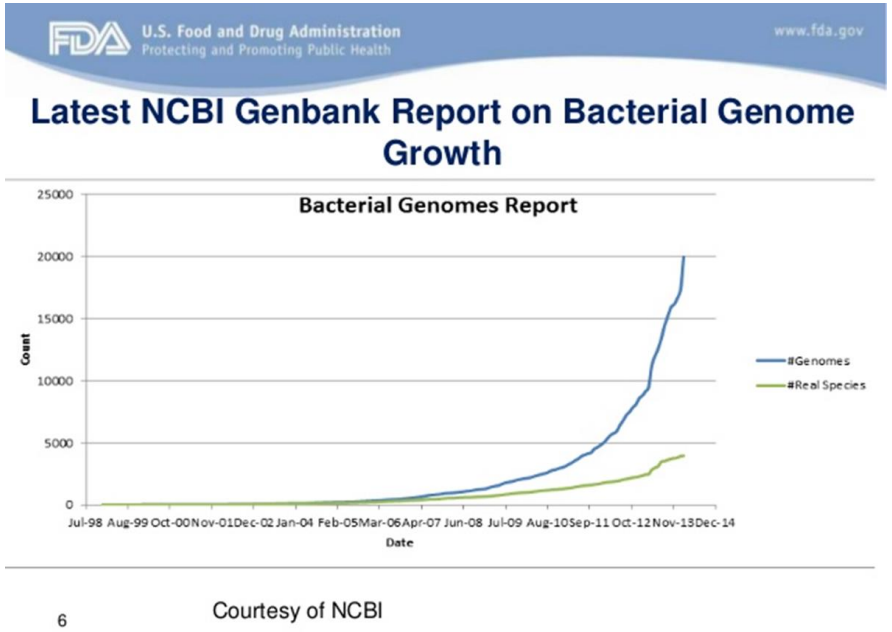
Applications of HPC in Biology

- Sequencing data
  - Evolutionary questions/modeling
  - Elucidation of molecular pathways
  - Phylogenomics (NP hard)
- Applications in population modeling, epidemiology
- Explorative Data Mining
  - gene clustering (FCM), gene ontologies

USD Research Computing Group

Applications of HPC in Biology-Cont.

- Can use R for parallel computing on HPC
  - SNOW(fall), others



Courtesy of NCBI

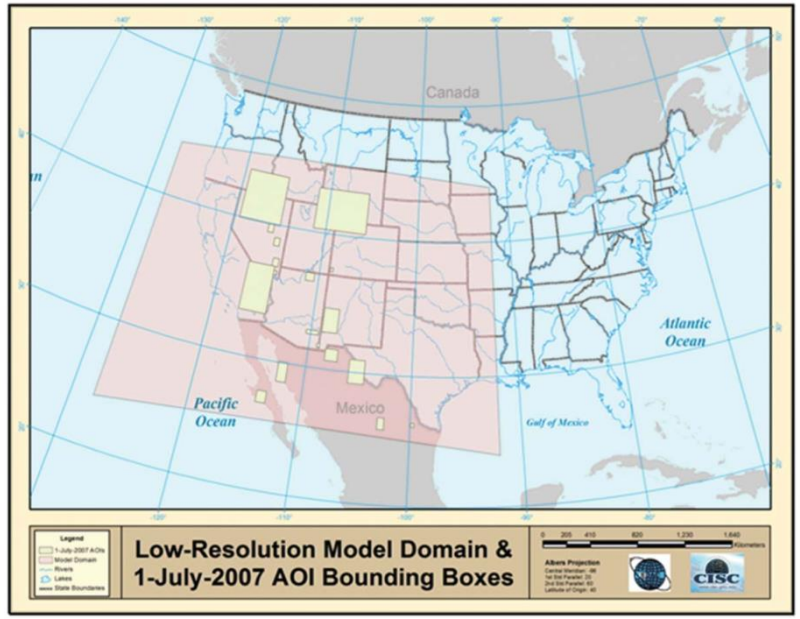
USD Research Computing Group

Baele, G., & Lemey, P. (2013). Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency. Bioinformatics, btt340.

Other Applications

- GIS (Geographic Information System) problems
- Research in humanities, social sciences

Sandstorm!
Everyone take cover!



USD Research Computing Group

Huang, Q., Yang, C., Benedict, K., Rezgui, A., Xie, J., Xia, J., & Chen, S. (2013). Using adaptively coupled models and high-performance computing for enabling the computability of dust storm forecasting. International Journal of Geographical Information Science, 27(4), 765-784.

In Case you Missed it…

- NSF Major instrumentation grant for new supercomputer
- Lawrence will come online May/June
- New opportunities
  - Visualization
  - More power and speed!
- XSEDE-Jetstream access
  - Free Matlab!

USD Research Computing Group

Resources for YOU!

- GitHub Group – USDRCG
- https://hpcc.okstate.edu/content/hpc-education-resources
- https://www.xsede.org/home

XSEDE

Extreme Science and Engineering Discovery Environment

USD Research Computing Group

Questions?

Your poll will show here

1

Install the app from
pollev.com/app

2

Make sure you are in
Slide Show mode

Still not working? Get help at pollev.com/app/help
or
Open poll in your web browser

USD Research Computing Group