Welcome to the *Pittsburgh Supercomputing Center* Big Data Workshop

John Urbanic

Parallel Computing Scientist
Pittsburgh Supercomputing Center

Distinguished Service Professor
Carnegie Mellon University

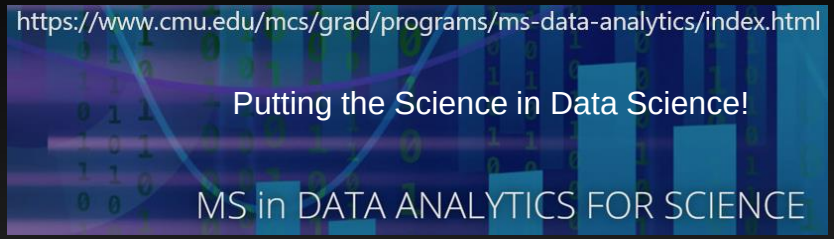
Who are we?

- If you find this a useful resource, you are welcome to apply for time to continue as a member of our research community.
- It is free.
- It is easy.

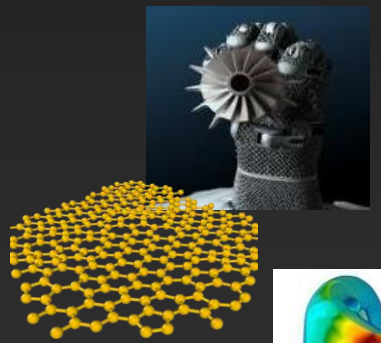
Who am I?

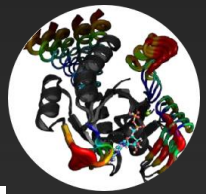
John Urbanic

Distinguished Service Professor
Carnegie Mellon University



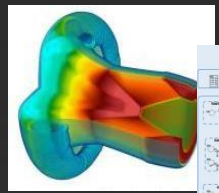
Undergrad Advanced Computational Physics
Graduate Large Scale Computing
Data Science Capstone Projects

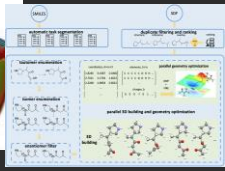




Parallel Computing Scientist
Pittsburgh Supercomputing Center

Code, code, code, on
Parallel platforms: MPI, OpenMP, OpenACC, ...
Big Data platforms: Spark, ...
Machine Learning: Spark, TensorFlow, PyTorch, ...





PSC HPC Monthly Workshop Schedule

- October 15-16 — *HPC Monthly Workshop*: Big Data & Machine Learning
- November 13 — *HPC Monthly Workshop*: GPU
- December 10-11 — *HPC Monthly Workshop*: MPI
- January 7-8 — *HPC Monthly Workshop*: Big Data & Machine Learning
- February 19 — *HPC Monthly Workshop*: OpenMP
- March 4-5 — *HPC Monthly Workshop*: MPI
- April 9 — *HPC Monthly Workshop*: GPU
- **May 13-14 — *HPC Monthly Workshop*: Big Data & Machine Learning**

- **More to come!**

HPC Monthly Workshop Philosophy

- Workshops as long as they *should* be.

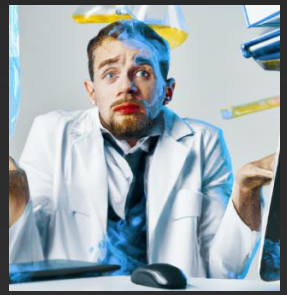
- You have real lives…
  in different time zones…
  that don't come to a halt.

- Learning is a social process
  - This is not a MOOC
  - This is the Wide Area Classroom
    so raise your expectations!

Our particular motivation

Is this you?

- Machine learning in the sciences went from "science fiction" 10 years ago, to "maybe there is something to this" 5 years ago, to "this works way better than anything else" today.
- Meanwhile, all the knowledge is still walled off in the CS community. Usually in semester long courses.
- This ~~panic~~ exciting situation has left a huge knowledge gap for practicing scientists at all levels.
- So here we are.



PITTSBURGH SUPERCOMPUTING CENTER

Agenda

Tuesday, May 13th

11:00	Welcome
11:25	A Brief History of Big Data
12:20	Intro to Spark
1:00	Lunch Break
2:00	More Spark and Exercises
3:00	Intro To Machine Learning
4:30	*Maybe Recommender early start*
5:00	Adjourn

Wednesday, May 14th

11:00	Machine Learning: A Recommender System
12:30	*Maybe Deep Learning early start*
1:00	Lunch break
2:00	Deep Learning
5:00	The Big Picture
5:30	Adjourn

We do this all the time, but…

- This is a very ambitious agenda.
- We are going to cover the guts of a semester course.
- We may get a little casual with the agenda.
- The reasons we can attempt this now:
    - Tools have reached the point (Spark and TF) where you can do some powerful things at a high level.
    - Worked last time. Feedback is very positive.

Biggest Potential For Disappointment

- We absolutely, definitely, without question, wish we had more hands-on exercise time.
- This is by design and demand. The topics we cover are all greatly requested and attempts to delete any of them provoke outrage in our surveys. This demand has compressed our hands-on sessions.
- One solution is for you to use the remainder of our short days to do further work.
- We also assume you will use your extended access to do exercises.
- Use your time wisely, and ask questions relentlessly.

Resources

Your local TAs

Questions from the audience

On-line talks:

https://www.psc.edu/resources/training/hpc-workshop-series/

The old XSEDE YouTube Channel

These are the last WAC versions of these events and have much of this content. Find them on the XSEDE Monthly Workshop Training Channel:

XSEDETraining

They will be incrementally appearing in the coming months on a new channel. Subscribe and give us feedback.

Copying code from PDFs is very error prone. Subtle things like substituting “-” for “-” are maddening. I have provided online copies of the codes in a directory that we shall shortly visit. I strongly suggest you copy from there if you are in a cut/paste mood.

Getting Connected

The first time you use your account sheet, you must go to apr.psc.edu to set a password. You may already have done so, if not, we will take a minute to do this shortly.

We will be working on bridges2.psc.edu. Use an ssh client (a Putty terminal, for example), to ssh to the machine.

At this point will be on a login node. It will have a name like “bridges2-login012”. This is a fine place to edit and compile codes. However we must be on compute nodes to do actual computing. We have designed Bridges to be the world’s most interactive supercomputer. We generally only require you to use the batch system when you want to. Otherwise, you get your own personal piece of the machine. For this workshop we will use

```
interact
```

to get a regular node of the type we will be using with Spark. You will then see name like “r251” on the command line to let you know you are on a regular node. Likewise, to get a GPU node, use

```
interact -gpu
```

This will be for our TensorFlow work tomorrow. You will then see a prompt like “v32”.

Some of you may follow along in real time as I explain things; some of you may wait until exercise time, and some of you may really not get into the exercises until after we wrap up tomorrow. It is all good.

Modules

We have hundreds of packages on Bridges. They each have many paths and variables that need to be set for their own proper environment, and they are often conflicting. We shield you from this with the wonderful modules command or containers. You can load the two packages we will be using as

Spark

```
module load spark
```

Tensorflow

```
singularity shell --nv  /ocean/containers/ngc/tensorflow/tensorflow_23.04-tf2-py3.sif
```

Editors

For editors, we have several zero-setup options:

- emacs
- vim
- nano: use this if you aren’t familiar with the others

It you are comfortable with using a remote editor like VS Code, feel free. But we can't afford the time to help you configure it.

For this workshop, you can get by just working from the command line. We will mostly be working in a Python shell.

Programming Language

Warning! Warning!

Several of the packages we are using are very prone to throw warnings about the JVM or some python dependency.

We've stamped most of them out, but don't panic if a warning pops up here or there.

In our other workshops we would not tolerate so much as a compiler warning, but this is the nature of these software stacks, so consider it good experience.

- We have to pick something
- Pick best domain language
- Python
- But not "Pythonic"
- I try to write generic pseudo-code
    - If you know Java or C or R, etc. you should be fine.

Our Setup For This Workshop

After you copy the files from the training directory, you will have:

```
/BigData
        /Clustering
        /MNIST
        /Recommender
        /Shakespeare
```

Datasets, and also cut and paste code samples are in here.

Preliminary Exercise

Let's get the boring stuff out of the way now.

- Log on to apr.psc.edu and set an initial password if you have not.
- Log on to Bridges-2.

```
ssh username@bridges2.psc.edu
```

- Run the setup script that will copy over the BigData directory we will all use..

```
~training/Setup
```

- Edit a file to make sure you can do so. Use emacs, vim or nano (if the first two don't sound familiar).
- Start an interactive session.

```
interact
```