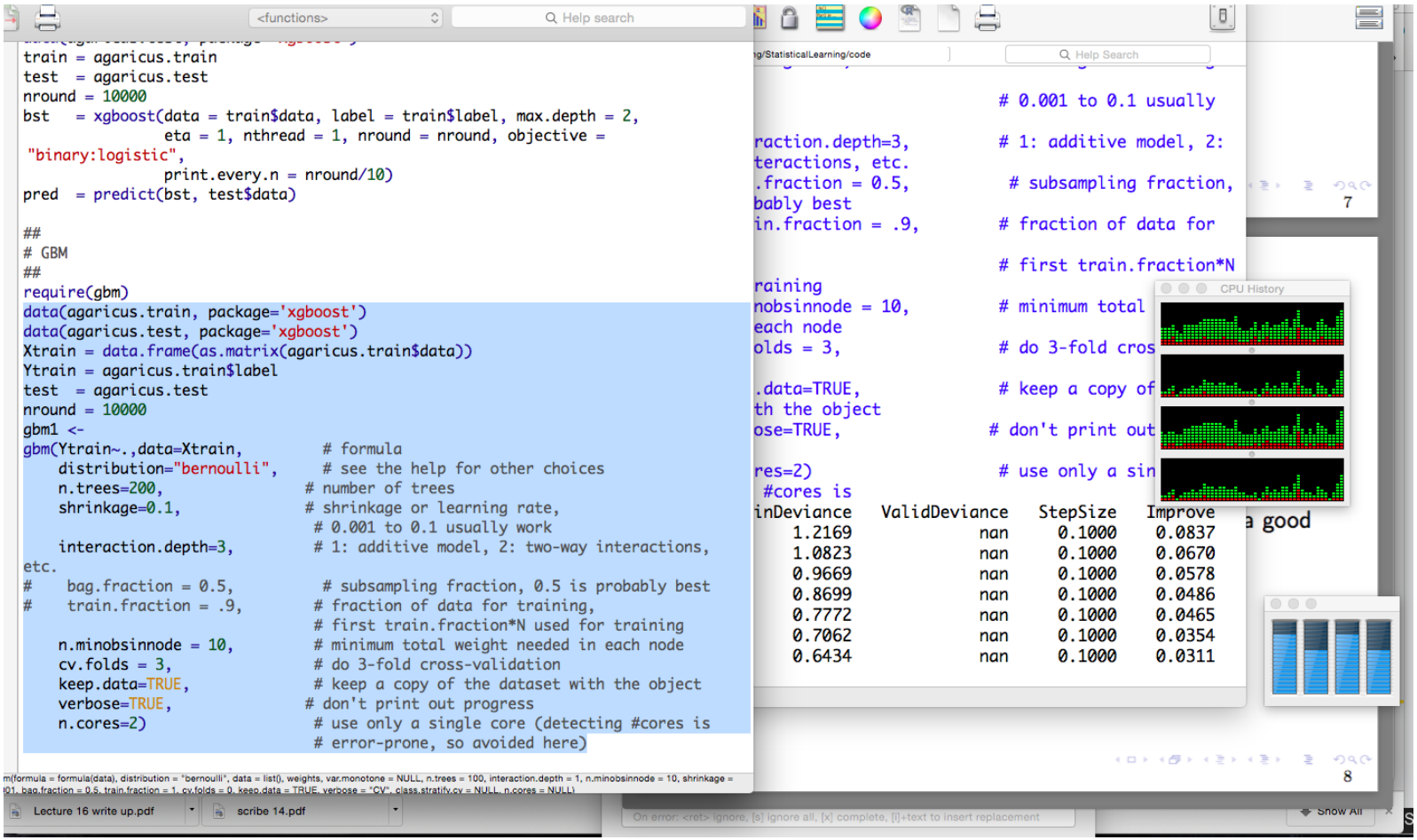
Quick Discussion on Parallelism

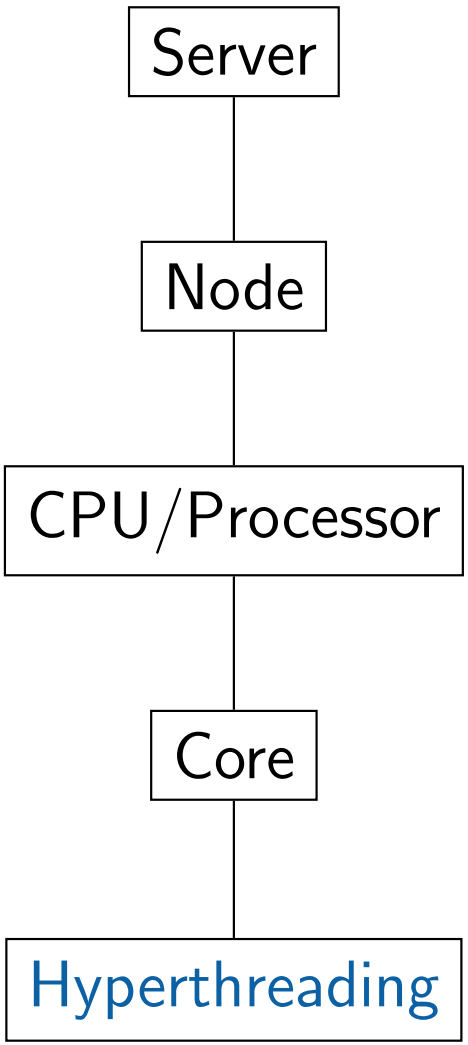
-Applied Analytics-

Lecturer: Darren Homrighausen, PhD

Introduction



Distributed computing hierarchy



Example: A server might have

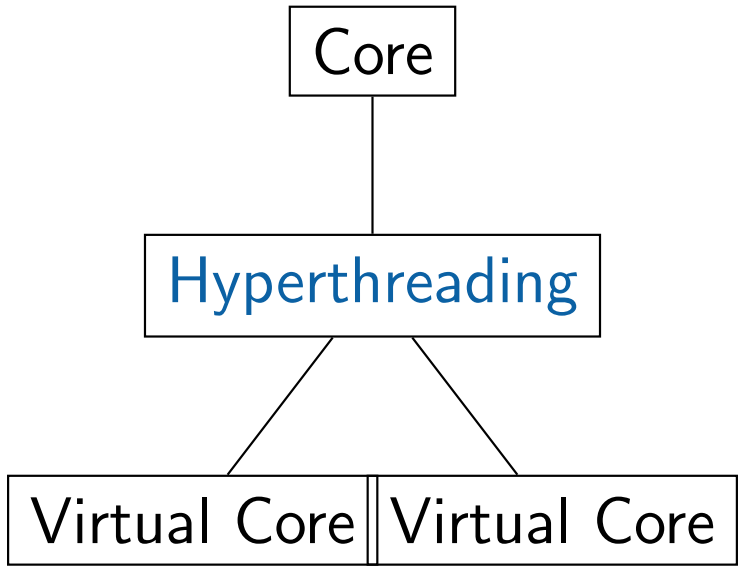
- 64 nodes
- 2 processors per node
- 16 cores per processor
- hyper threading

The goal is to somehow allocate a job so that these resources are used efficiently

Jobs are composed of threads, which are specific computations

Hyperthreading

Developed by Intel, Hypertheading allows for each core to pretend to be two cores



This works by trading off computation and read-time for each core

Embarrassingly parallel

Suppose we want to make a large number of computations, but it can be broken up intro independent chunks

Examples:

- K-fold CV: Compute the performance estimate for each fold and then combine them together
- Bootstrap draws: We can recompute the statistic for each bootstrap draw
- Matrix multiplication: We can break matrix into chunks of rows and do the multiplication for each chunk
- Split-apply-combine: We can assign each 'group' in group_by to a different core
- You are simulating at a large number of parameter settings.
- Grid search optimization....

K-fold CV

We split our data up in K non-overlapping chunks $(\mathcal{D}_k, k = 1, \ldots, K)$

Now, we can allocate to each core the following:

1. Fit the model on $\mathcal{D}_1, \ldots, \mathcal{D}_{k-1}, \mathcal{D}_{k+1}, \ldots, \mathcal{D}_K$
2. Get the performance estimate on $\mathcal{D}_k$, $\hat{E}_k$

After all the cores are done, we can recombine:

$$\hat{E} = \frac{1}{K}(\hat{E}_1 + \cdots, \hat{E}_K)$$

Some considerations

- Make sure the problem is large enough. Using parallelism on small problems can increase the processing time

  (This generally means the processing time is longer than you can accept)
- Pay attention to load balancing
- Make sure the processes are independent

  (Example: for K-fold, if you train using one processor and get the performance estimate on another)
- It's best practice to run in 'batch mode' without having an IDE/interpreter open

  (For this class, don't worry about this point. This is more for your future careers. Batch mode would be, e.g. R CMD BATCH script.R &)