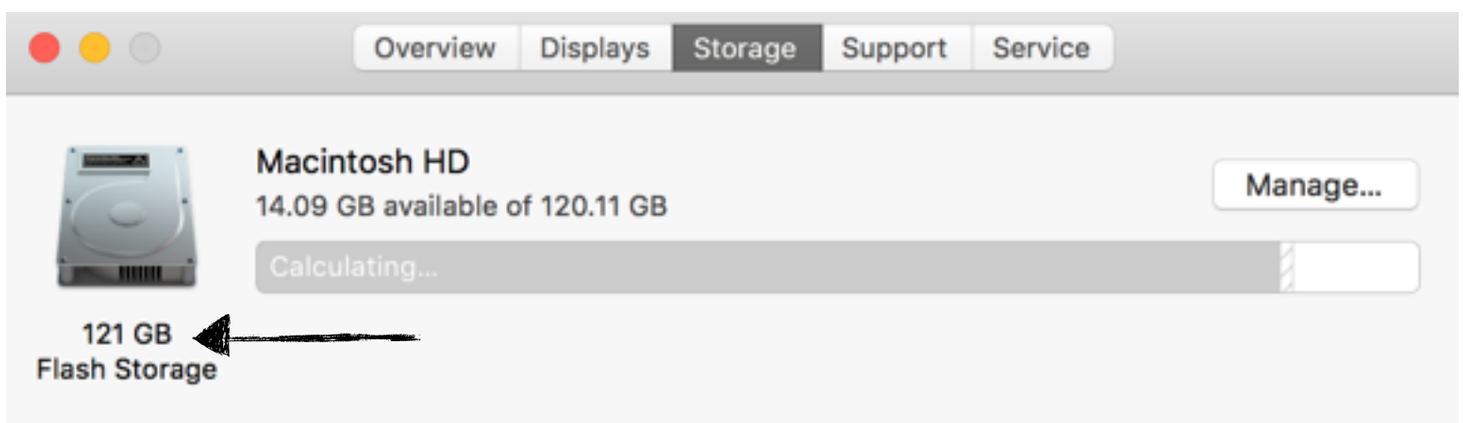
Big Data

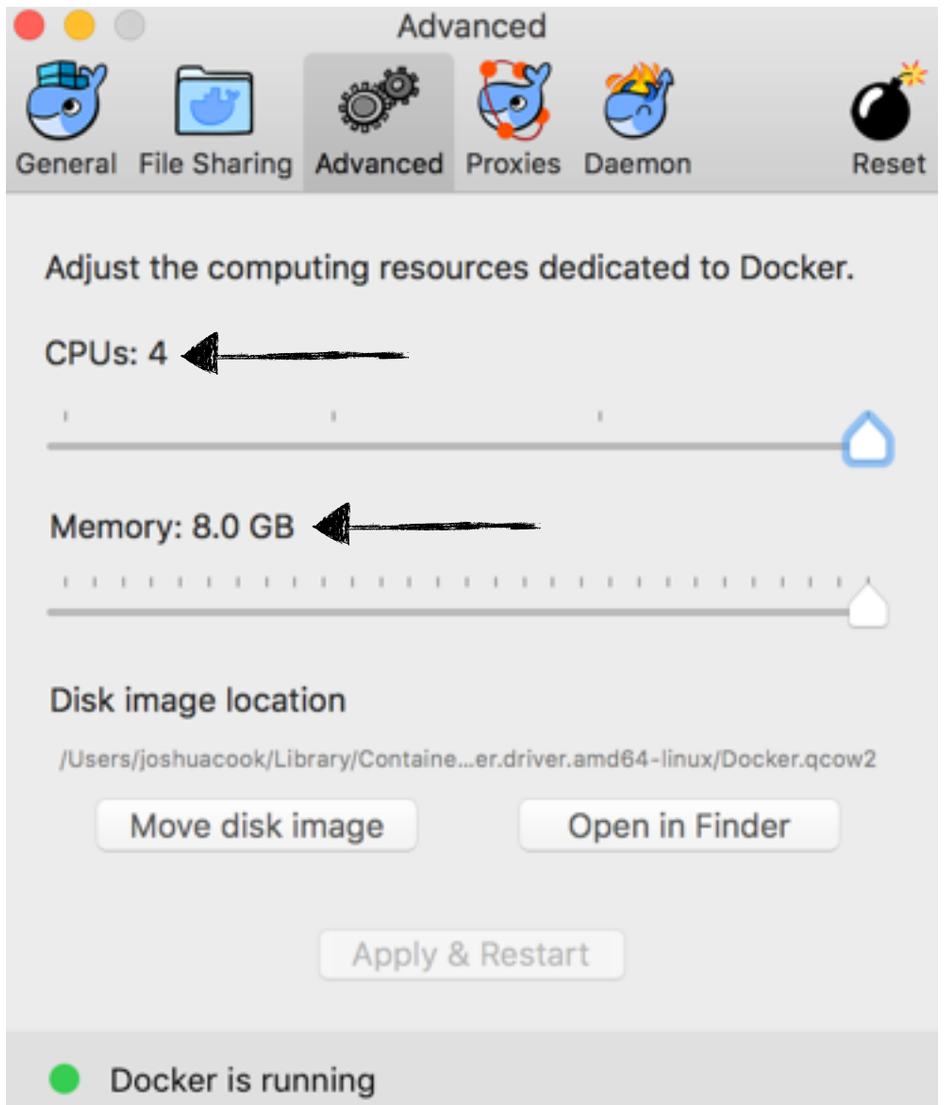
- A computational problem of scope, size, or complexity such that I have to think about the scope, size, or complexity.

- This is in terms of our available computing resources.

Computing Resources

processors • RAM • Hard Drive





Computing Resources

processors • RAM • Hard Drive

What do we use them for?

processors	RAM	Hard Drive
		we use the -v flag to connect dockerized Jupyter to our hard drive this is how we save files
	data we will do some immediate action	docker image cache is on the hard drive docker containers are also on the hard drive
loads a program	loads the resources of the program	data kept in csvs
Jupyter will use a processor: - to run python ("the kernel") - to run the notebook server		

Computing Resources

processors • RAM • Hard Drive

	vCPU	ECU	Memory (GiB)	Instance Storage (GB)	Linux/UNIX Usage
General Purpose - Current Generation					
t2.nano	1	Variable	0.5	EBS Only	$0.0059 per Hour
t2.micro	1	Variable	1	EBS Only	$0.012 per Hour
t2.small	1	Variable	2	EBS Only	$0.023 per Hour
t2.medium	2	Variable	4	EBS Only	$0.047 per Hour
t2.large	2	Variable	8	EBS Only	$0.094 per Hour
t2.xlarge	4	Variable	16	EBS Only	$0.188 per Hour
t2.2xlarge	8	Variable	32	EBS Only	$0.376 per Hour

https://aws.amazon.com/ec2/pricing/on-demand/

The “Modeling Problem” and The “Engineering Problem”

- The “Modeling Problem”
  - Develop the “best” model for a particular dataset and application
- The “Engineering Problem”
  - Build a system capable of
    - loading our particular dataset
    - Running our analysis tools
    - Training and using any number of models
  - In a Timely Fashion