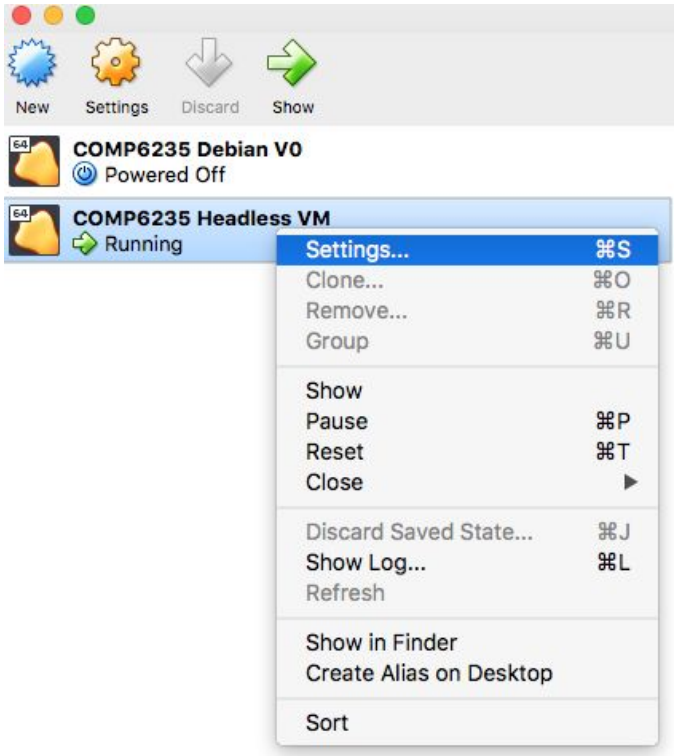
Comp 6235 Apache Spark

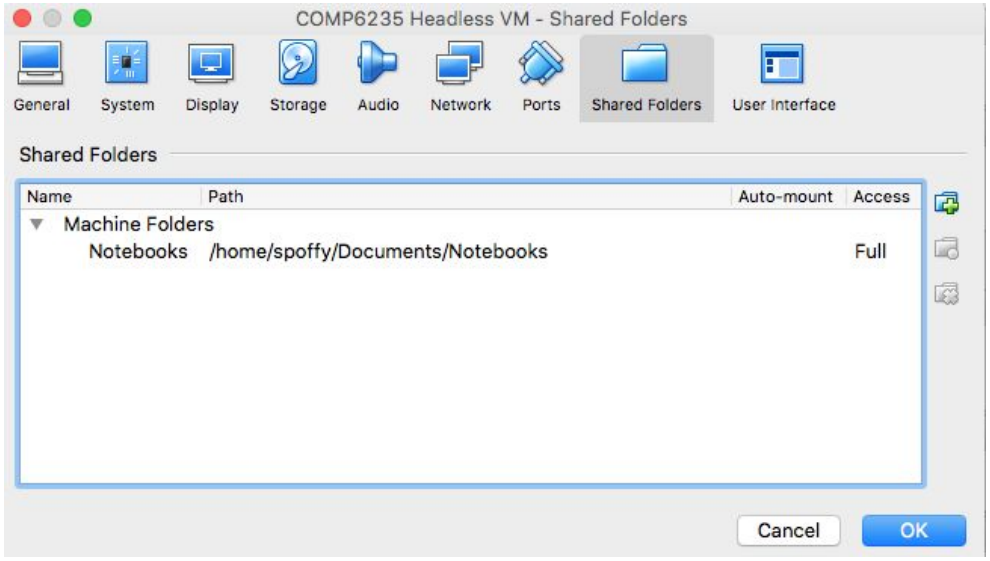
Set up VM

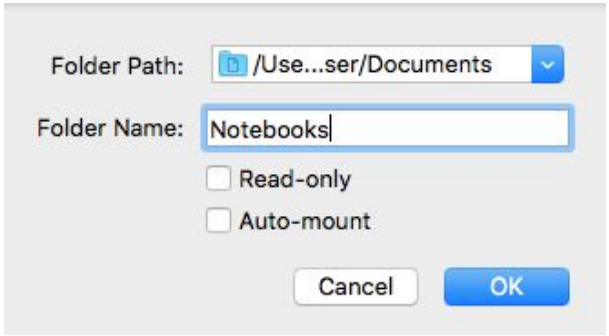
1. Copy .ova file from usb to your machine (recommended to put in you h drive)
2. Double click .ova file (import as new appliance)
3. Modify the setting (right-click on the VM then click Settings...) for Shared Folders (Folder Path set to your Documents folder , then at Folder Name field put Notebooks , see examples below)
4. Start the VM (*** note: do not start the vm if you have not done step 3.)
5. Type ./run-jupyter
6. Open a browser on your local machine, type in http://localhost:8888/ (password is comp6235)
7. Open sparkTutorial.ipynb in the browser
   or
   Download sparkTutorial.ipynb to your local machine from the wiki and upload the file in the Jupyter notebook

Overview of Apache Spark

- Open-source distributed cluster computing framework
- At the core are resilient distributed datasets (RDDs)
- Interoperable with Python, Scala, Java, SQL and R
- Originally created by a PhD student in the University of California


New
Settings
Discard
Show
COMP6235 Debian V0
Powered Off
COMP6235 Headless VM
Running
Settings... ⌘S
Clone... ⌘O
Remove... ⌘R
Group ⌘U
Show
Pause ⌘P
Reset ⌘T
Close
Discard Saved State... ⌘J
Show Log... ⌘L
Refresh
Show in Finder
Create Alias on Desktop
Sort


COMP6235 Headless VM - Shared Folders
General
System
Display
Storage
Audio
Network
Ports
Shared Folders
User Interface
Shared Folders
Name
Path
Auto-mount
Access
Machine Folders
Notebooks
/home/spoffy/Documents/Notebooks
Full
Cancel
OK


Folder Path:
/Use...ser/Documents
Folder Name:
Notebooks
Read-only
Auto-mount
Cancel
OK