New York University
Computer Science and Engineering

Big Data

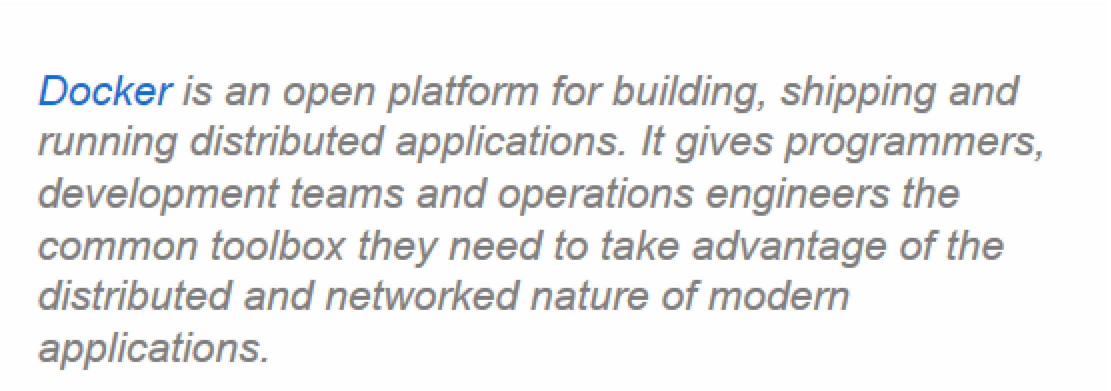
Course Tools

Juan Rodriguez
jcr365@nyu.edu

Course Tools

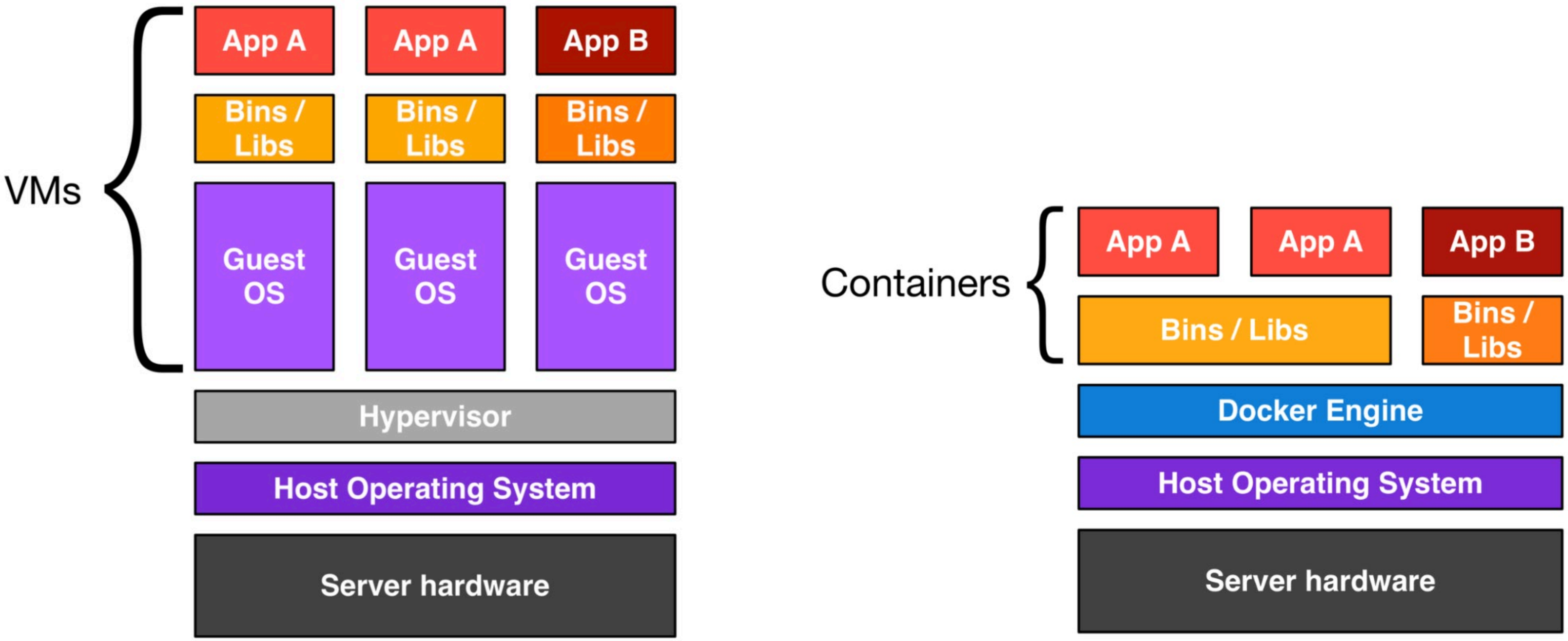
- **Virtualization**
  - Do **NOT** use the standalone latest version of Virtual Box (incompatible with Docker)
  - **Docker: https://www.docker.com/**
    - Windows 7, 10 Home: Docker Toolbox: https://www.docker.com/products/docker-toolbox
    - Windows 10 Pro: Docker for Windows: https://docs.docker.com/docker-for-windows/
    - Mac OS: Docker for Mac: https://www.docker.com/products/docker#/mac
    - Linux: containers (Linux kernel 3.6+): https://docs.docker.com/engine/installation/linux/
  - Windows required virtualization enabled in BIOS
  - Windows 10 Pro: required HyperV virtualization installed/enabled)

https://www.docker.com/

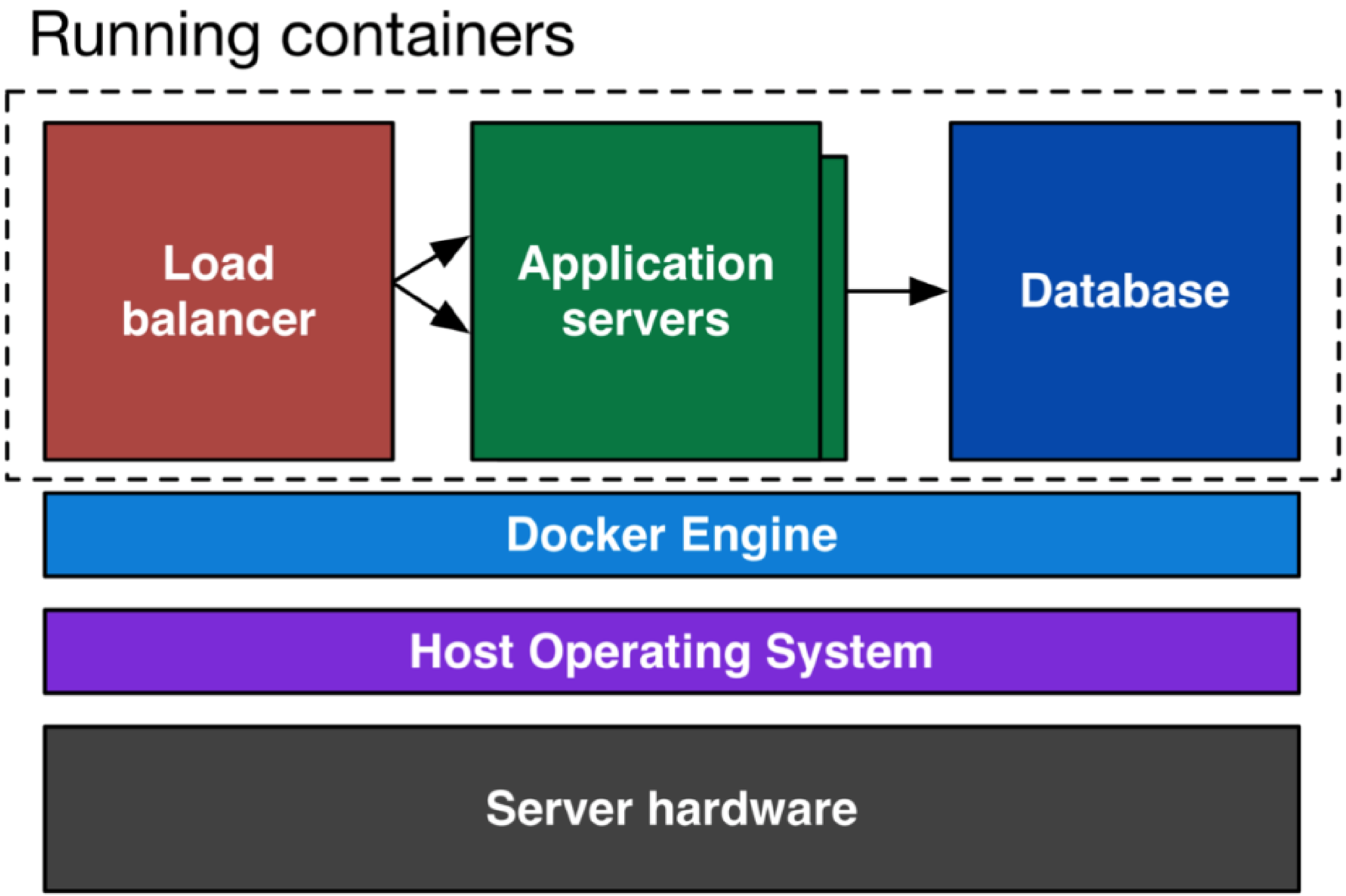

Docker is an open platform for building, shipping and running distributed applications. It gives programmers, development teams and operations engineers the common toolbox they need to take advantage of the distributed and networked nature of modern applications.

Concepts:

- Hypervisor
- Linux Containers
- System Images



Source: https://www.codementor.io/docker/tutorial/what-is-docker-tutorial-andrew-baker-oreilly



Running containers

Load balancer

Application servers

Database

Docker Engine

Host Operating System

Server hardware

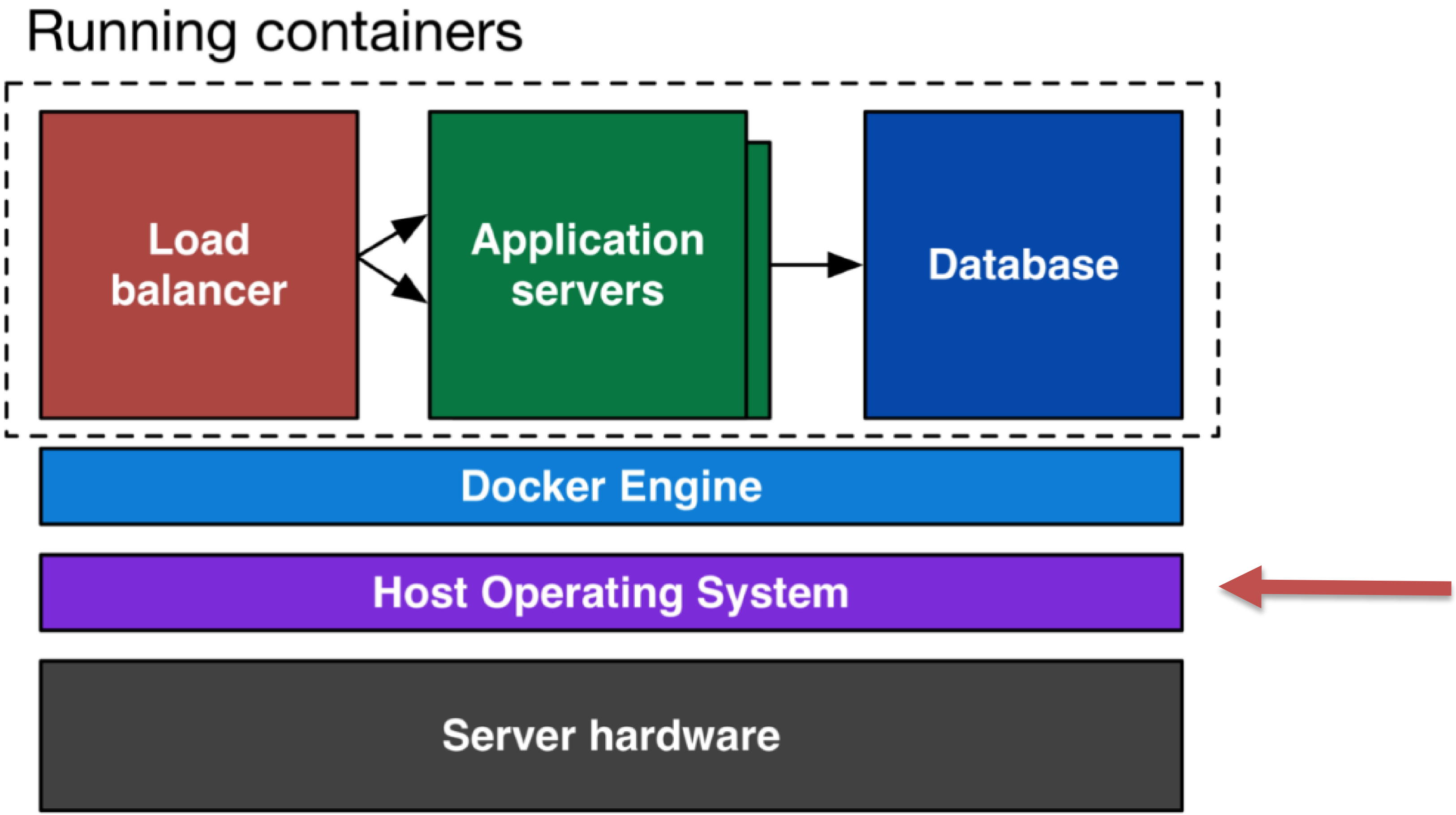
Components

- Docker daemon
- Docker client
- Hub
- Toolkit(s)

Concepts

- Images
- Repositories
- Tags

- Commands
- Tags



Running containers

Load balancer

Application servers

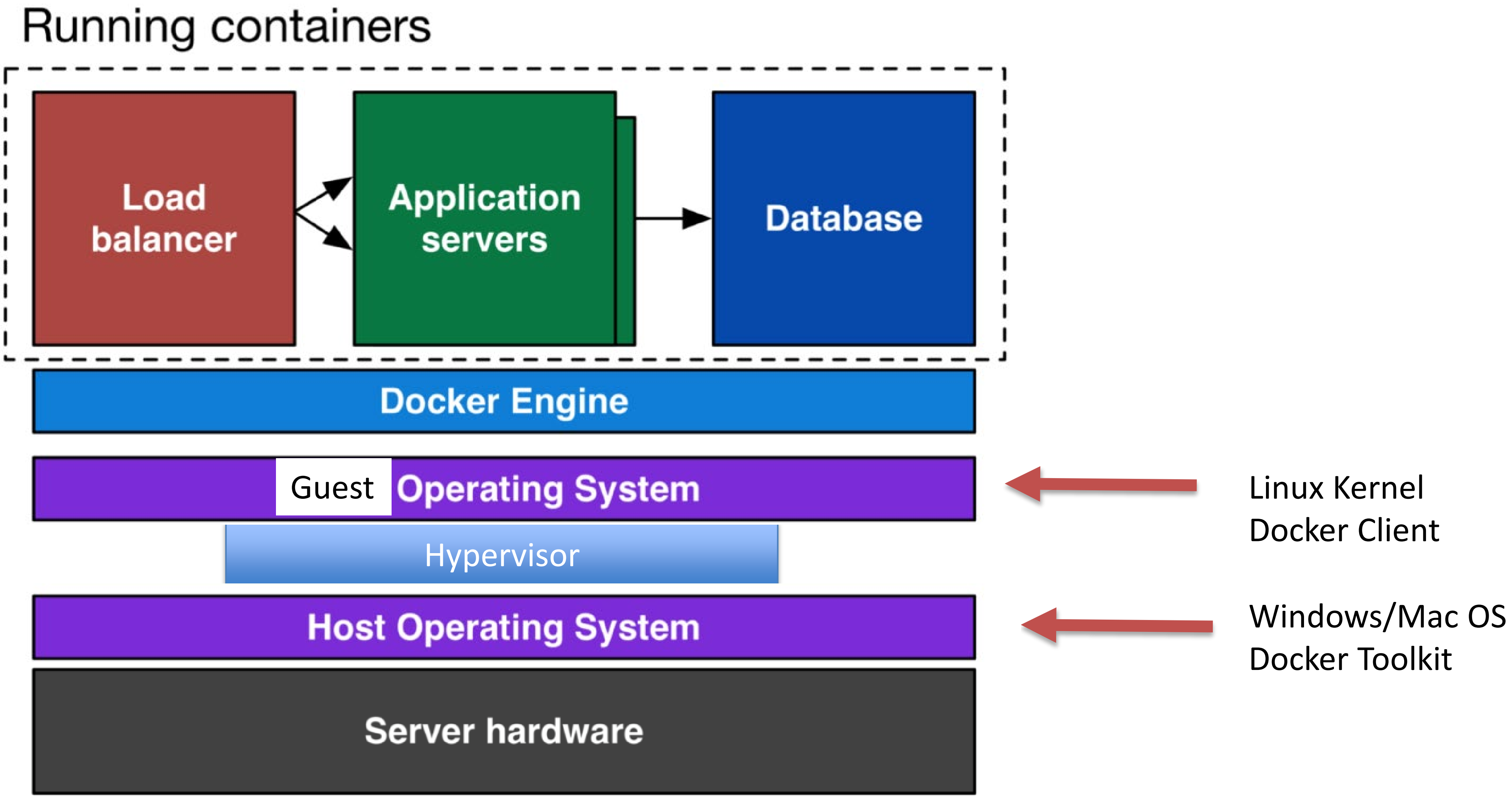
Database

Docker Engine

Host Operating System

Server hardware

Linux Kernel
Docker Client



Running containers

Load balancer

Application servers

Database

Docker Engine

Guest Operating System

Hypervisor

Host Operating System

Server hardware

Linux Kernel
Docker Client

Windows/Mac OS
Docker Toolkit

Docker Hub https://hub.docker.com/

Hello World: *docker pull hello-world*

Hadoop Quickstart: https://hub.docker.com/r/cloudera/quickstart/
docker pull cloudera/quickstart

Zeppelin: https://hub.docker.com/r/dylanmei/zeppelin/
docker pull dylanmei/zeppelin

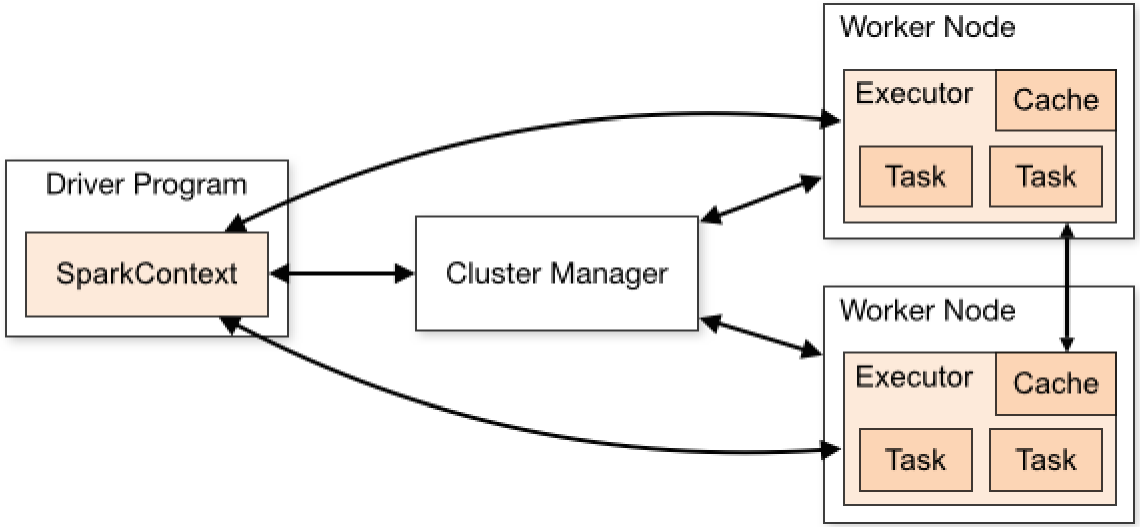
MongoDB: https://hub.docker.com/_/mongo/
docker pull mongo

Spark: (I recommend running in standalone mode; see next slide)
mhttps://hub.docker.com/r/gettyimages/spark/
docker pull gettyimages/spark

Apache Spark http://spark.apache.org/

- Install locally for this course. Deploy on cluster or the cloud for the final project
- Requirements: Java 7+

- Windows: Hadoop/Spark requires a special windows stub (windows-utils) to operate correctly)
  Download from http://www.barik.net/archive/2015/01/19/172716/
  Put in a folder and add the folder to your PATH environment variable



IDEs used by the professor

Java:

- IntelliJ IDEA: https://www.jetbrains.com/idea/
  Community Version is fine; their commercial offering is free with academic registration
- Eclipse: https://www.eclipse.org/downloads/packages/eclipse-ide-java-developers/neon2

Scala

- IntelliJ IDEA: https://www.jetbrains.com/idea/
  Community Version is fine; their commercial offering is free with academic registration
- Scala IDE: http://scala-ide.org/

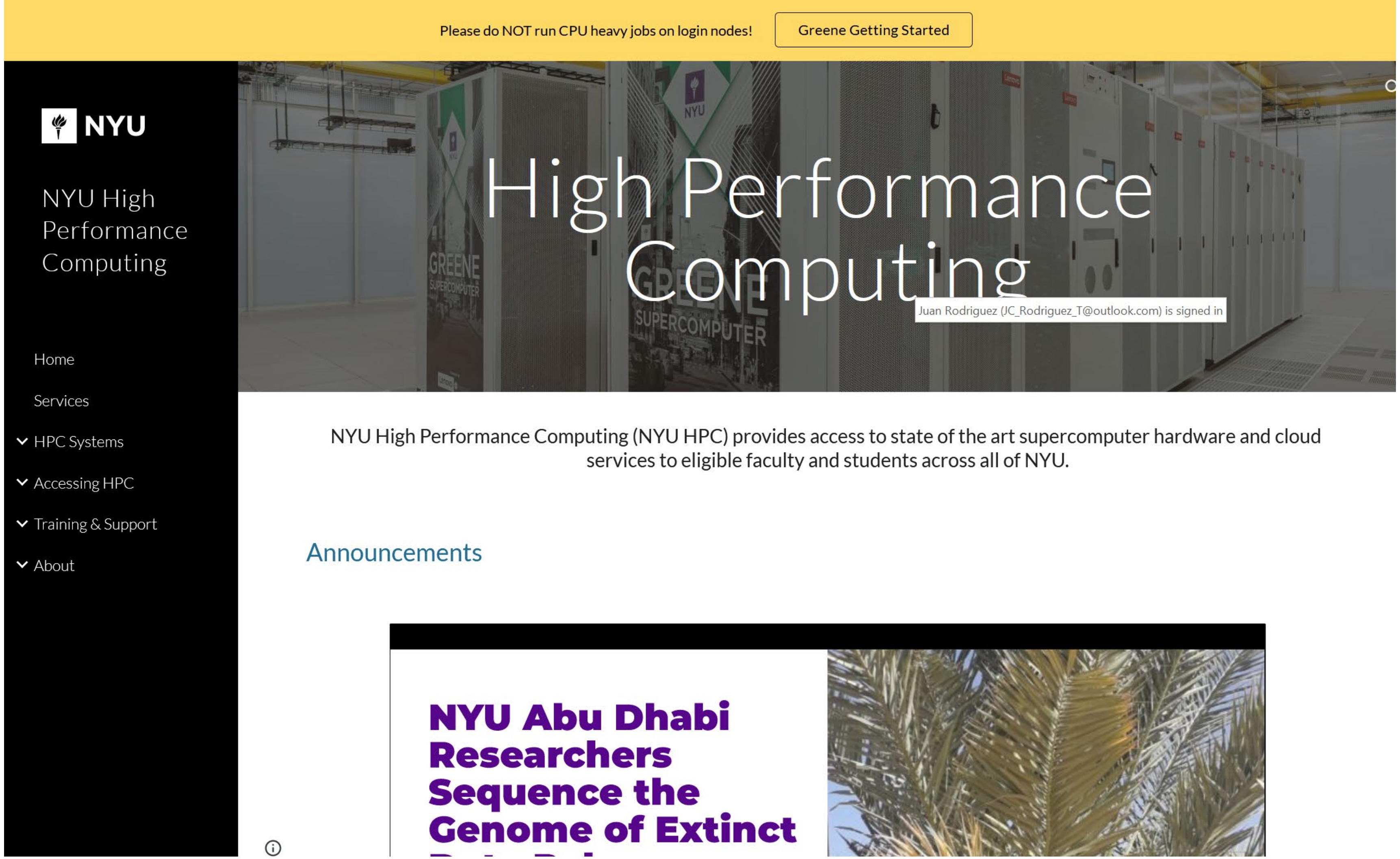
- ## Python
  IntelliJ Pycharm: https://www.jetbrains.com/pycharm/

Please note there are other options, though I do not/will not show in this course

Other tools you will see used during the course:

- Eclipse Plugin for Hadoop: https://github.com/winghc/hadoop2x-eclipse-plugin
- R: https://www.r-project.org/
- R-Studio: https://www.rstudio.com/
- H2O: http://www.h2o.ai/h2o/
- Knime: https://www.knime.org/knime-analytics-platform
- Visual Studio: https://www.visualstudio.com/vs/visual-studio-express/
- MySQL Workbench: https://www.mysql.com/products/workbench/

NYU High Performance Computing



Please do NOT run CPU heavy jobs on login nodes!

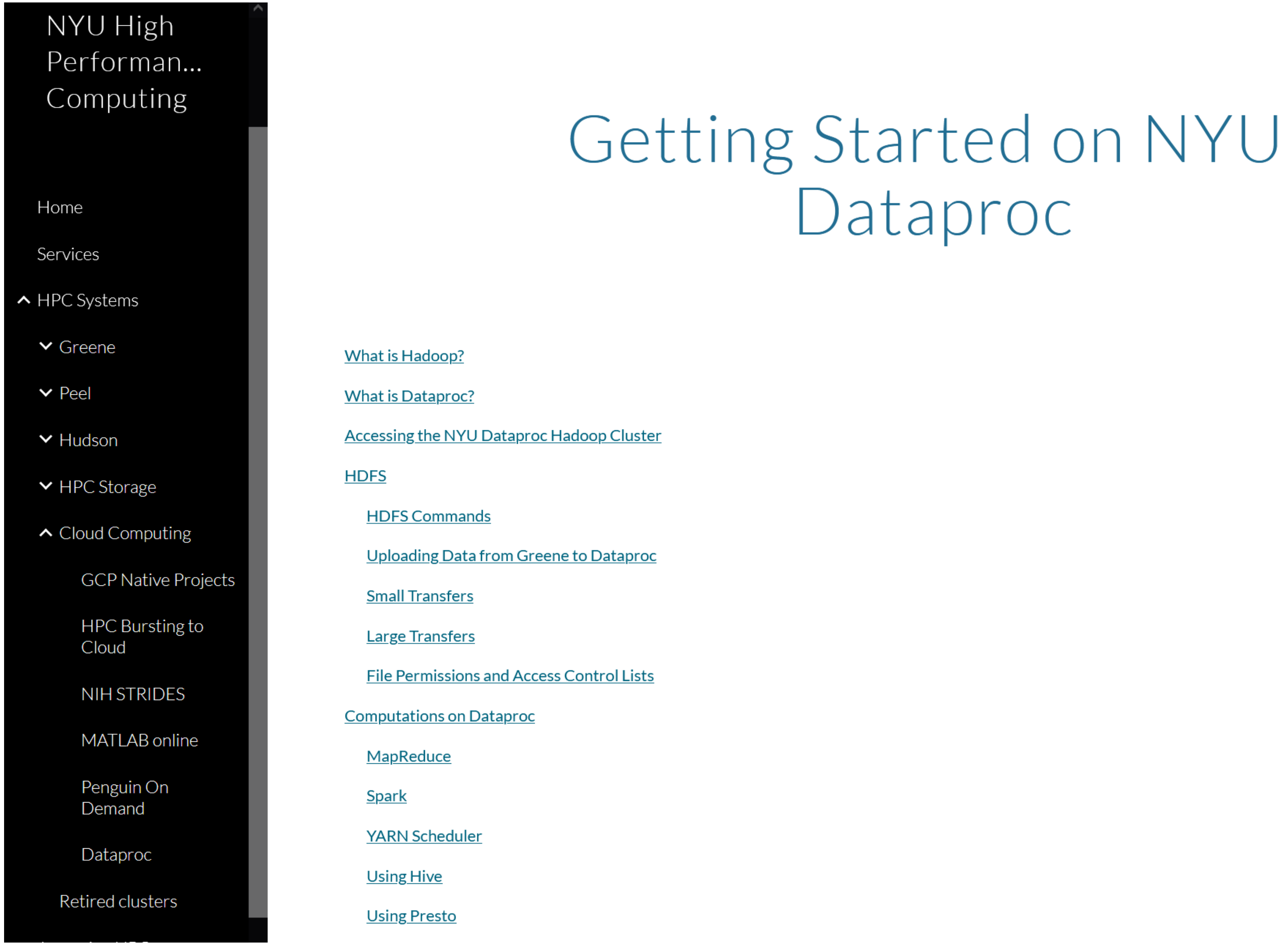
Greene Getting Started
NYU
NYU High Performance Computing
Home
Services
HPC Systems
Accessing HPC
Training & Support
About
High Performance Computing
Juan Rodriguez (JC_Rodriguez_T@outlook.com) is signed in

NYU High Performance Computing (NYU HPC) provides access to state of the art supercomputer hardware and cloud services to eligible faculty and students across all of NYU.

Announcements

NYU Abu Dhabi Researchers Sequence the Genome of Extinct

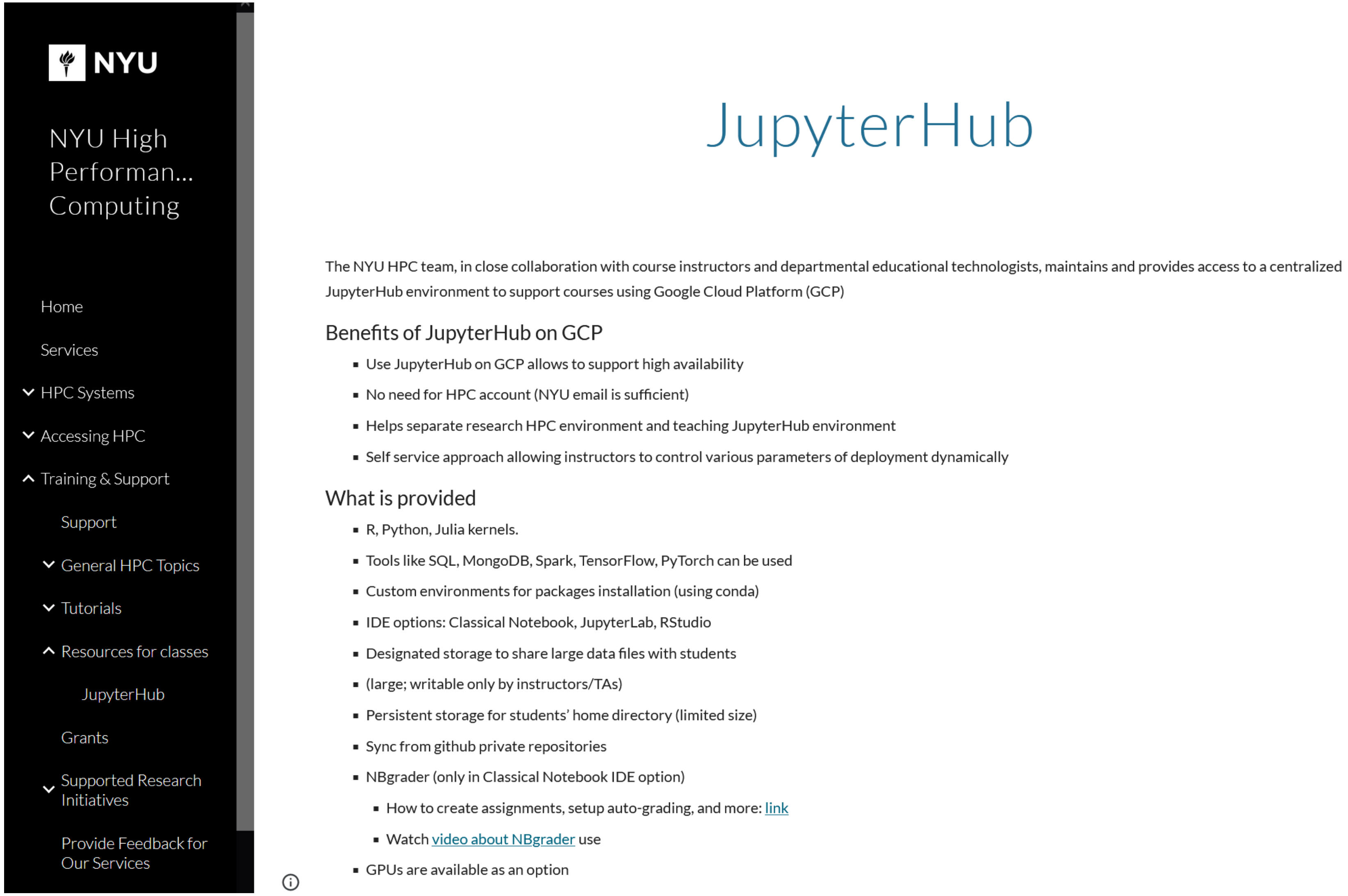
NYU High Performance Computing - Dataproc



Getting Started on NYU Dataproc

- What is Hadoop?
- What is Dataproc?
- Accessing the NYU Dataproc Hadoop Cluster
- HDFS
  - HDFS Commands
  - Uploading Data from Greene to Dataproc
  - Small Transfers
  - Large Transfers
  - File Permissions and Access Control Lists
- Computations on Dataproc
  - MapReduce
  - Spark
  - YARN Scheduler
  - Using Hive
  - Using Presto

NYU High Performance Computing - JupyterHub



JupyterHub

The NYU HPC team, in close collaboration with course instructors and departmental educational technologists, maintains and provides access to a centralized JupyterHub environment to support courses using Google Cloud Platform (GCP)

Benefits of JupyterHub on GCP

- Use JupyterHub on GCP allows to support high availability
- No need for HPC account (NYU email is sufficient)
- Helps separate research HPC environment and teaching JupyterHub environment
- Self service approach allowing instructors to control various parameters of deployment dynamically

What is provided

- R, Python, Julia kernels.
- Tools like SQL, MongoDB, Spark, TensorFlow, PyTorch can be used
- Custom environments for packages installation (using conda)
- IDE options: Classical Notebook, JupyterLab, RStudio
- Designated storage to share large data files with students
- (large; writable only by instructors/TAs)
- Persistent storage for students' home directory (limited size)
- Sync from github private repositories
- NBgrader (only in Classical Notebook IDE option)
  - How to create assignments, setup auto-grading, and more: link
  - Watch video about NBgrader use
- GPUs are available as an option

- Reach out to the TAs for assistance with any of these tools

- Assigment handed out during week 2 or 3 requires you have the programming environments ready ….