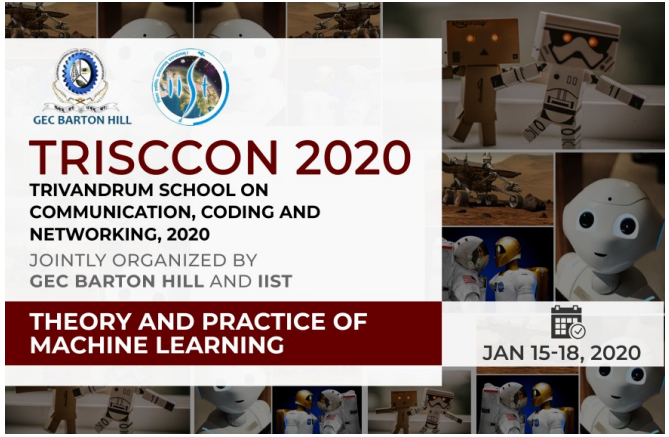



# Setting up a teaching environment for Machine Learning

Vineeth B. S.
Department of Avionics,
Indian Institute of Space Science and Technology

(Thanks to: Birenjith P. S.)

**Something to take away from these four days ...**

- TPML 2020
  - Courses in optimization, Python for machine learning, unsupervised, supervised, and reinforcement learning
  - Applying machine learning techniques to a variety of problems

- So what will you do now?
  - develop and teach machine learning courses
  - apply machine learning techniques to problems that you are working on
  - student projects in machine learning

## Some points to note ...

- In order to do any one of the above
  - Need to set up a mechanism for formulating, running, evaluating, and analyzing machine learning applications and experiments

- This talk touches on some aspects of setting up an environment to do machine learning teaching and research
  - Useful software components
  - Pointers to resources

- Talk is based on my experience. Apologies if it is biased and incomplete (if your favourite software package is not covered, please tell me!)
  - Talk does not cover the physical setup of a lab
  - Mainly targets an environment that can be easily deployed on to a student's laptop or desktop computers in a “programming lab”
  - Does not target “big data”, high performance computing, GPUs ...

## Plan for this talk ...

- Software tools for teaching ML
- Typical examples of how these software tools can be used
- How to package these software tools into a customized and easily deployable distribution?

## A pre-requisite for teaching ML: datasets

- https://toolbox.google.com/datasetsearch
- UCI machine learning datasets - https://archive.ics.uci.edu/ml/index.php
- Kaggle - https://www.kaggle.com/datasets
- Examples of Govt. sources
  - India - https://data.gov.in/
  - US - https://www.data.gov/
  - UK - https://www.ukdataservice.ac.uk/
- Financial data https://data.worldbank.org/, https://www.imf.org/en/Data
- MNIST handwritten digits dataset - http://yann.lecun.com/exdb/mnist/
- Classification of datasets
  - For each specific dataset type what is a good source of that dataset ?

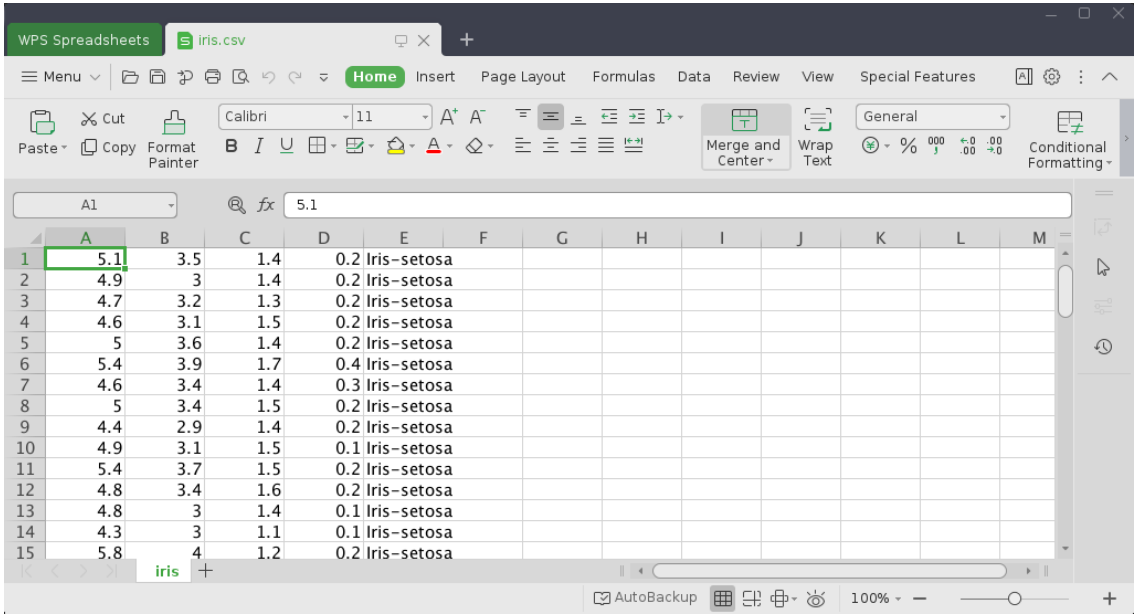
- Lists of datasets
  - https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research
  - https://elitedatascience.com/datasets
  - https://www.kdnuggets.com/2016/05/top-10-datasets-github.html
  - https://github.com/DataHackIL/DataSets
  - https://github.com/awesomedata/awesome-public-datasets

# Having a look at data - data exploration tools

- Let us look at 'iris' dataset and diabetes dataset

- Linux command line utilities

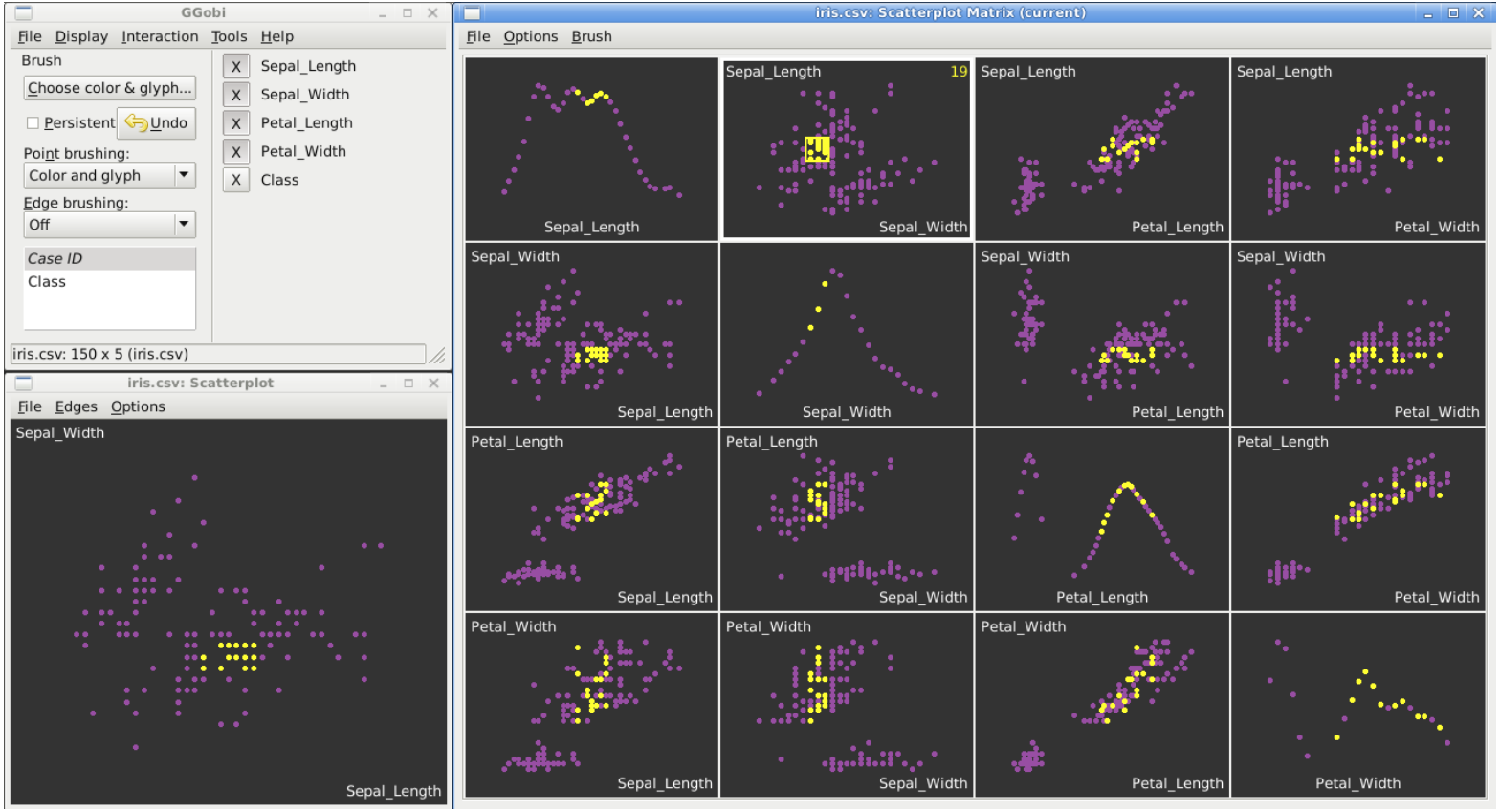
- Spreadsheet software
  - Libreoffice - https://www.libreoffice.org/
  - WPS office - https://www.wps.com/
  - Microsoft Excel




| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| 2 | 4.9 | 3 | 1.4 | 0.2 | Iris-setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| 4 | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| 5 | 5 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| 6 | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa |
| 7 | 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa |
| 8 | 5 | 3.4 | 1.5 | 0.2 | Iris-setosa |
| 9 | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa |
| 10 | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |
| 11 | 5.4 | 3.7 | 1.5 | 0.2 | Iris-setosa |
| 12 | 4.8 | 3.4 | 1.6 | 0.2 | Iris-setosa |
| 13 | 4.8 | 3 | 1.4 | 0.1 | Iris-setosa |
| 14 | 4.3 | 3 | 1.1 | 0.1 | Iris-setosa |
| 15 | 5.8 | 4 | 1.2 | 0.2 | Iris-setosa |

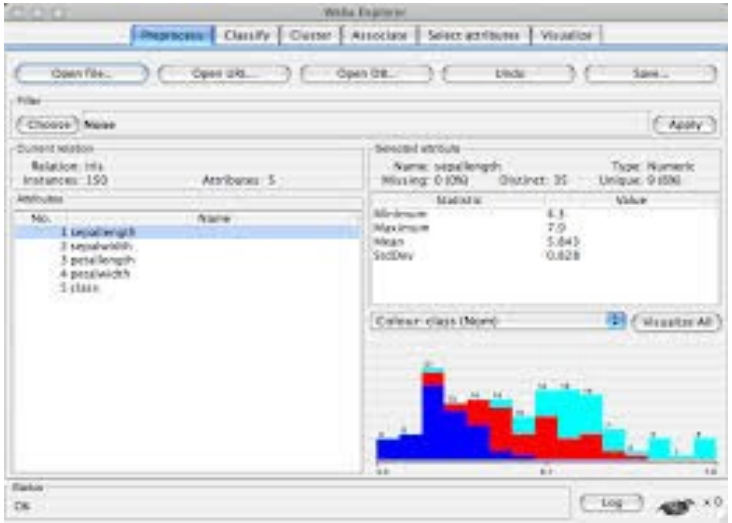
# Having a look at data - data exploration tools

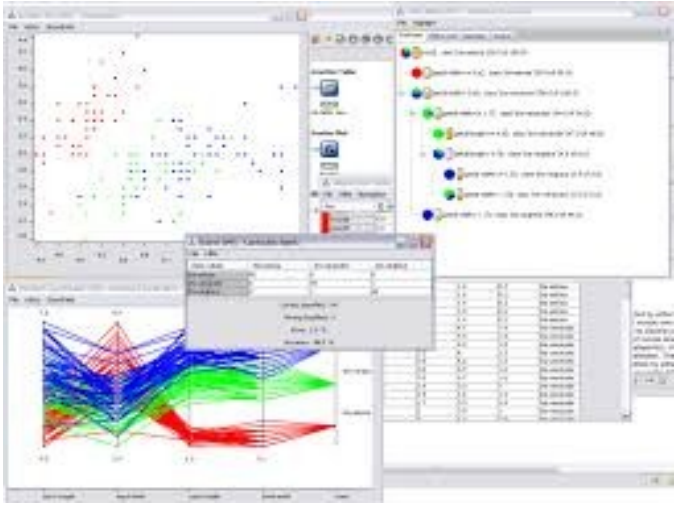
- GGobi (http://ggobi.org/)

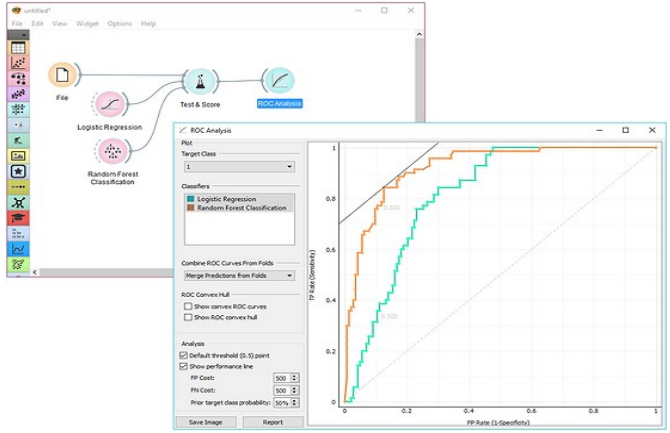


# Teaching ML without (much) programming background

- WEKA
- RapidMiner
- KNIME
- Orange(3)





# A quick introduction to WEKA

- WEKA - Waikato Environment for Knowledge Analysis

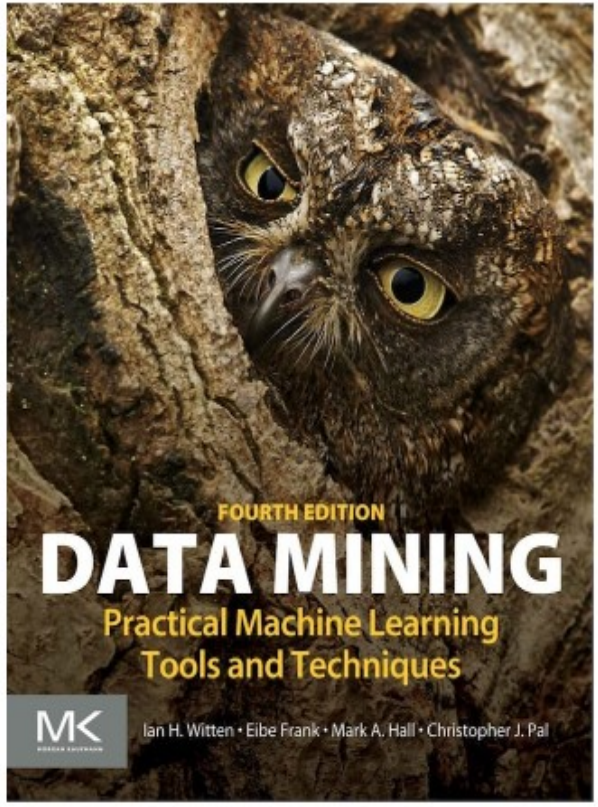
- Inside ...
  - 100+ in-built algorithms for classification
  - 70+ in-built algorithms for pre-processing of data
  - 25 feature selection algorithms
  - “Well-known” clustering, classification algorithms

- Demonstrations
  - Data exploration
  - Classification of a dataset

# Resources for setting up ML practice lab sessions using WEKA

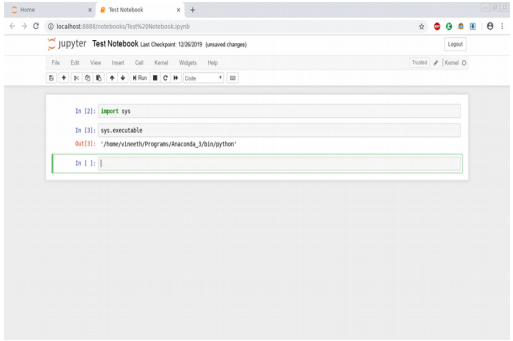
- A possible set of experiments ...
  - Data exploration & preprocessing on multiple datasets
    - Missing data, Describing data
  - Supervised learning algorithms
    - Classification
  - Unsupervised learning algorithms
    - Clustering

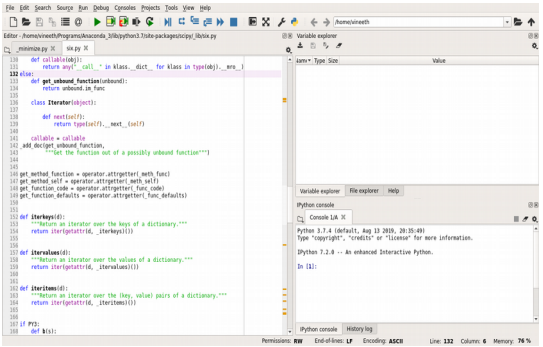
- Free resources for starting up ...
  - https://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/
  - https://www.cs.waikato.ac.nz/ml/weka/mooc/moredataminingwithweka/
  - https://www.cs.waikato.ac.nz/ml/weka/mooc/advanceddataminingwithweka/
  - https://www.cs.waikato.ac.nz/~ml/weka/book.html

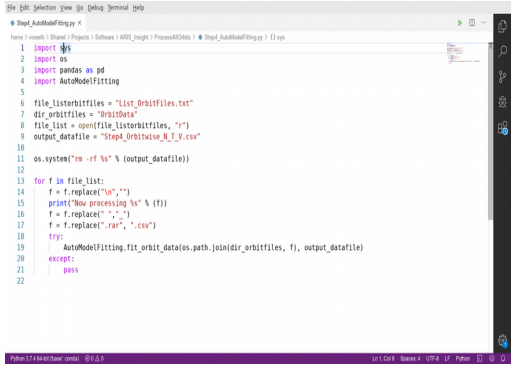


# Teaching ML with the use of a programming language

- Python is widely used as a language for building and studying ML algorithms
- Several options to work with Python
  - Jupyter notebook
  - Spyder
  - Visual code
- A number of useful libraries
  - Numpy
  - Scipy
  - Pandas
  - Scikit-learn
  - Matplotlib






- Resources for teaching python
  - Google's Python class - https://developers.google.com/edu/python
  - https://docs.python.org/3/tutorial/
  - https://www.learnpython.org/
  - https://www.coursera.org/learn/python

## Teaching ML with programming background (Python) - online

- Kaggle notebook
- Google CoLab

- AWS SageMaker
- Google Cloud Datalab
- Domino Data Lab
- DataBrick Notebooks
- Azure Notebooks
- Datalore
- CoCalc
- Binder

## Setting up Python

- Python - usually available on your computer
  - Maybe used for scripts used in administration of your linux distribution
  - Not really a good idea to modify this for your machine learning needs

- Python package management
  - pip installs python packages from PyPI - python package index
  - repository of open source third party python packages
  - pip install <package_name> can be used to easily install packages

## Python virtual environment

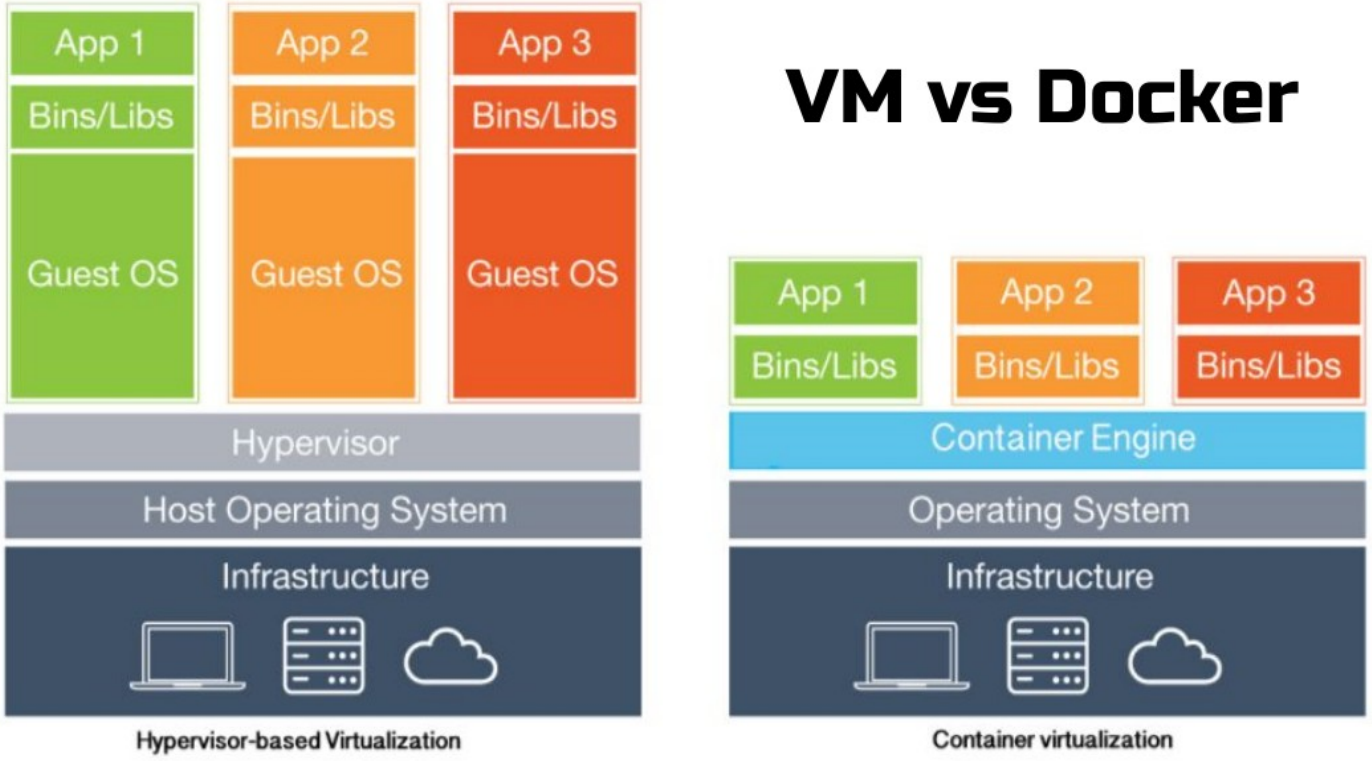
- Creation of isolated environments for Python
- Demonstration of how to set up an isolated environment (https://pyvideo.org/pycon-us-2011/pycon-2011--reverse-engineering-ian-bicking--39-s.html)

- virtualenvwrapper provides a set of commands which makes working with virtual environments much more pleasant.
- It also places all your virtual environments in one place.

- why does this work?
  - a copy of the python binary is kept in your virtual environment directory
  - libraries and modules are searched in places relative to the location of the python binary - this is a feature of python and not of virtualenv - means that we can have other versions of libraries and modules placed with respect to this new location of the python binary.

## Deployment options

- Virtual machines
- Docker




Courtesy: Nan Li

## Deployment using virtual machines

- A very simplified view - how does a program run on your computer?
- Should be possible to emulate this in software - a virtual computer.

- Setting up a virtual machine ...
  - Virtualbox
  - VMware
- Setting up of a virtual machine for deployment in ML teaching ...

## Docker

- Machine learning labsheets as well as associated software can be given to students or participants in the form of containers.

- Multiple containers can co-exist without any conflict and have their own software, libraries, and configuration
- Containers are run on a operating system (like linux) and are not as resource intensive as virtual machines. Containers share the operating system kernel.

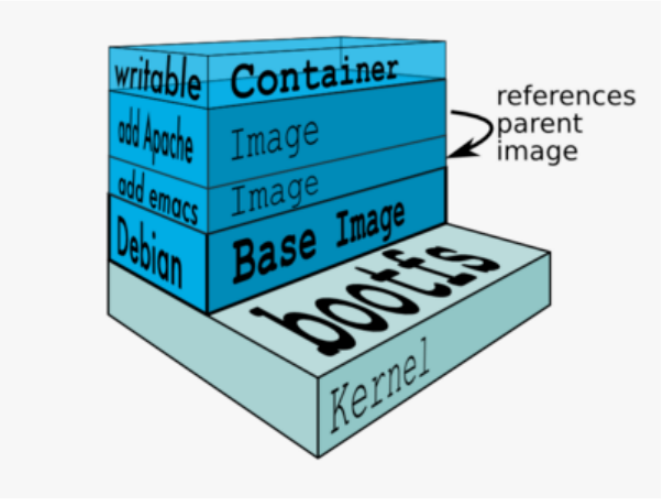
- Containers from images
- Images contain layers
  - Starting from an operating system, required libraries, application, and then configuration
  - Only layers that have changed needs to be communicated

## Docker images

- An image is a set of files that forms the root file system.
- Images are made up of layers which can be visualized as being stacked one on top of the other
- Layers can add, change, and remove files
- Sharing of layers for optimizing performance is also possible

- An image is like a read only file system
- A container provides a writable copy of that filesystem
  - Containers and images are like instances and classes

- Our need is to build a docker image which will contain all the necessary software for our machine learning lab.
  - Demonstration ...
- This process can be automated using a Dockerfile
  - A series of directives (FROM, RUN, COPY)




Courtesy: Tutorials on docker.org

## A possible sequence of experiments for an ML lab with Python

- Lab 1: Datasets and their exploration
  - Classification of datasets
  - Different kinds of attributes
  - Missing data
  - Outliers
  - Generating synthetic data
- Lab 2:
  - Python introduction
  - Jupyter notebooks
- Lab 3:
  - Use of basic python libraries
  - Numpy and Scipy libraries
  - Basic plotting in Python using Matplotlib
- Lab 4:
  - Introduction to Pandas & Scikit-Learn
  - Feature engineering
  - Dimensionality reduction
- Lab 5:
  - Unsupervised learning using scikit-learn
  - Clustering algorithms
- Lab 6:
  - Training, Validation, Test split
  - Supervised learning using scikit-learn
- Other labs from other topics can be added as required

## Resources for ML

- https://github.com/josephmisiti/awesome-machine-learning/blob/master/books.md
- https://github.com/josephmisiti/awesome-machine-learning/blob/master/blogs.md
- https://github.com/ujjwalkarn/Machine-Learning-Tutorials
- https://github.com/NajiElKotob/Awesome-ML