OAK RIDGE National Laboratory

Machine Learning for Materials Science with python

Rama Vasudevan[1], Vinit Sharma[2]
[1]Center for Nanophase Materials Sciences,
[2]Joint Institute for Computational Sciences,
Oak Ridge National Laboratory

August 14th 2019

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

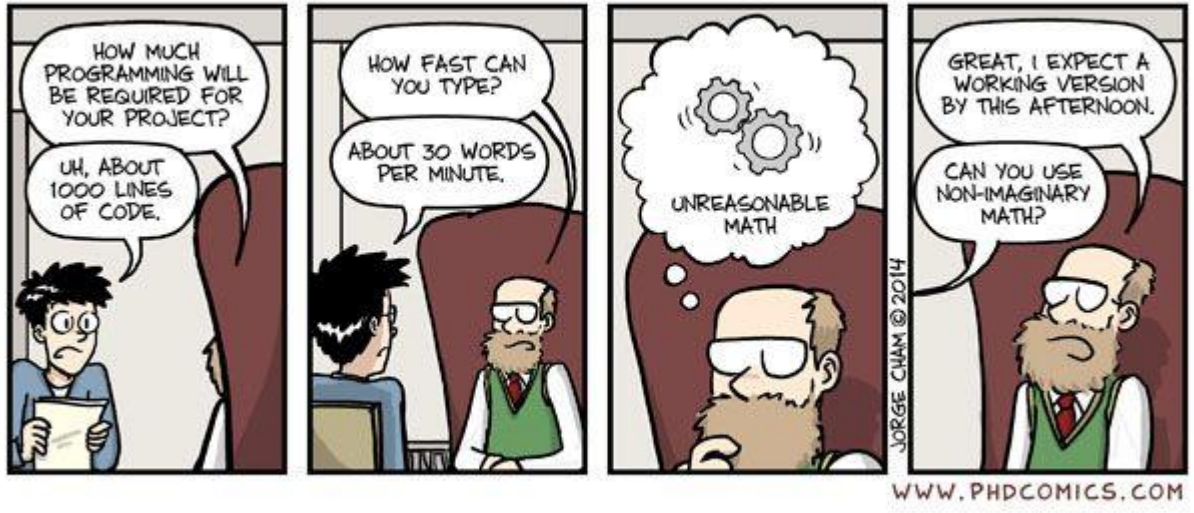
U.S. DEPARTMENT OF ENERGY

Pre-requisites

Details and data required for today's workshop are present at the shared Google drive that was sent out

Please let us know NOW if you do not have this

What you need for today



- An understanding of python and the python scientific software stack is desirable but not essential
- Laptop that has internet access
- Time (all of today!)

Agenda

Time	Details
8:00-9:00 AM	Setup, installation and introductions
9:00-9:30 AM	Introduction to machine learning in python
9:30-10:45 AM	Signal and Image Processing Methods
10:45-11:00AM	Coffee Break
2:00-3:00 PM	Guest speaker: Kamal Choudhary – Applications of ML to Materials Theory
4:00-4:30 PM	Deep Learning Methods: An overview
4:30 PM	Wrap up/Discussion/Summary

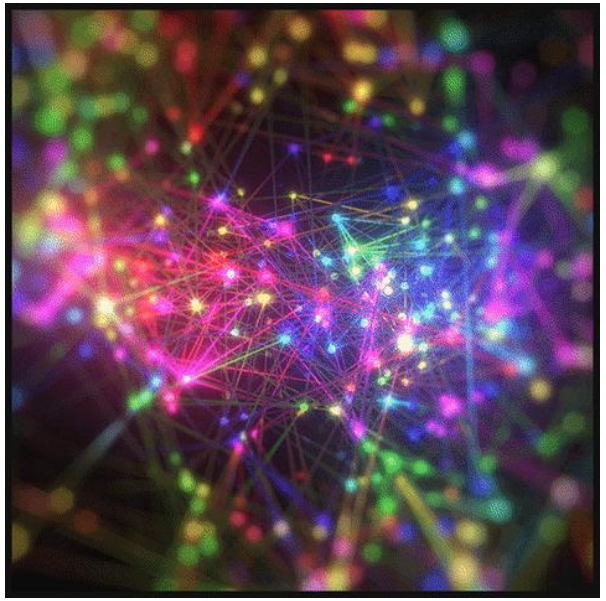
Why machine learning?

- You are here, so it must be important….
- More seriously, machine/statistical learning is about finding correlations and relationships in datasets
- With large data sizes and more data dimensions, it is increasingly more difficult to analyze data using traditional methods. Even visualizing trends in multidimensional datasets can be a formidable challenge
- Data science has developed tools to both analyze and visualize data, and do so in situations where the underlying relationships do not have easy analytical functional forms.

giphy.com

What *is* Machine learning?

- There are lots of definitions, but the main essence of machine learning is for the system to learn relationships from data, enabling generalizations to new situations. This is as opposed to being specifically programmed.
- Machine learning methods can be used for the following tasks:
  - Classification (can I group my observations in some manner?)
  - Regression (can I fit the data, even if the function cannot be expressed analytically?)
  - Dimensionality Reduction (can we express the data in a shorter form?)



giphy.com

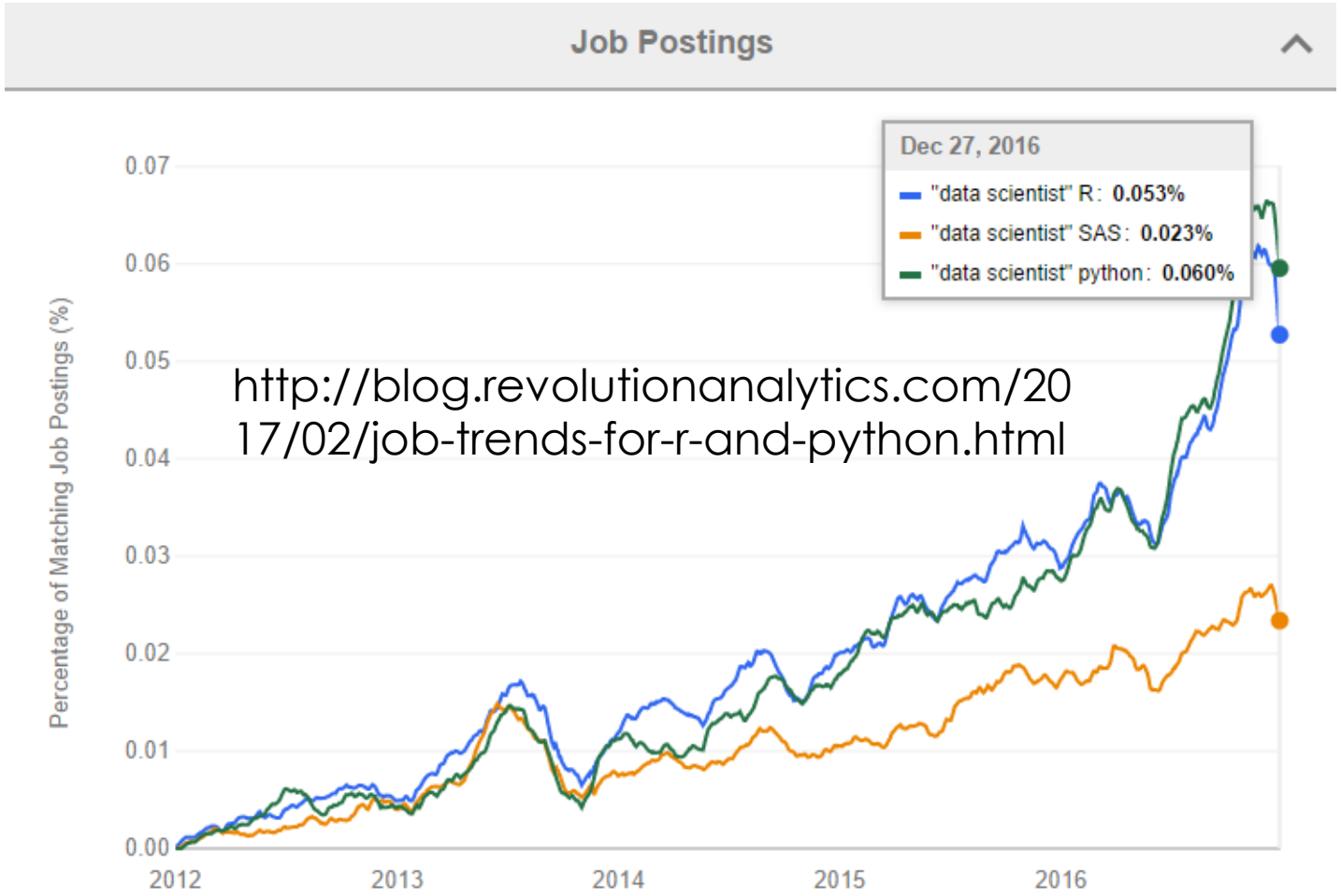
giphy.com

But why is it popular?

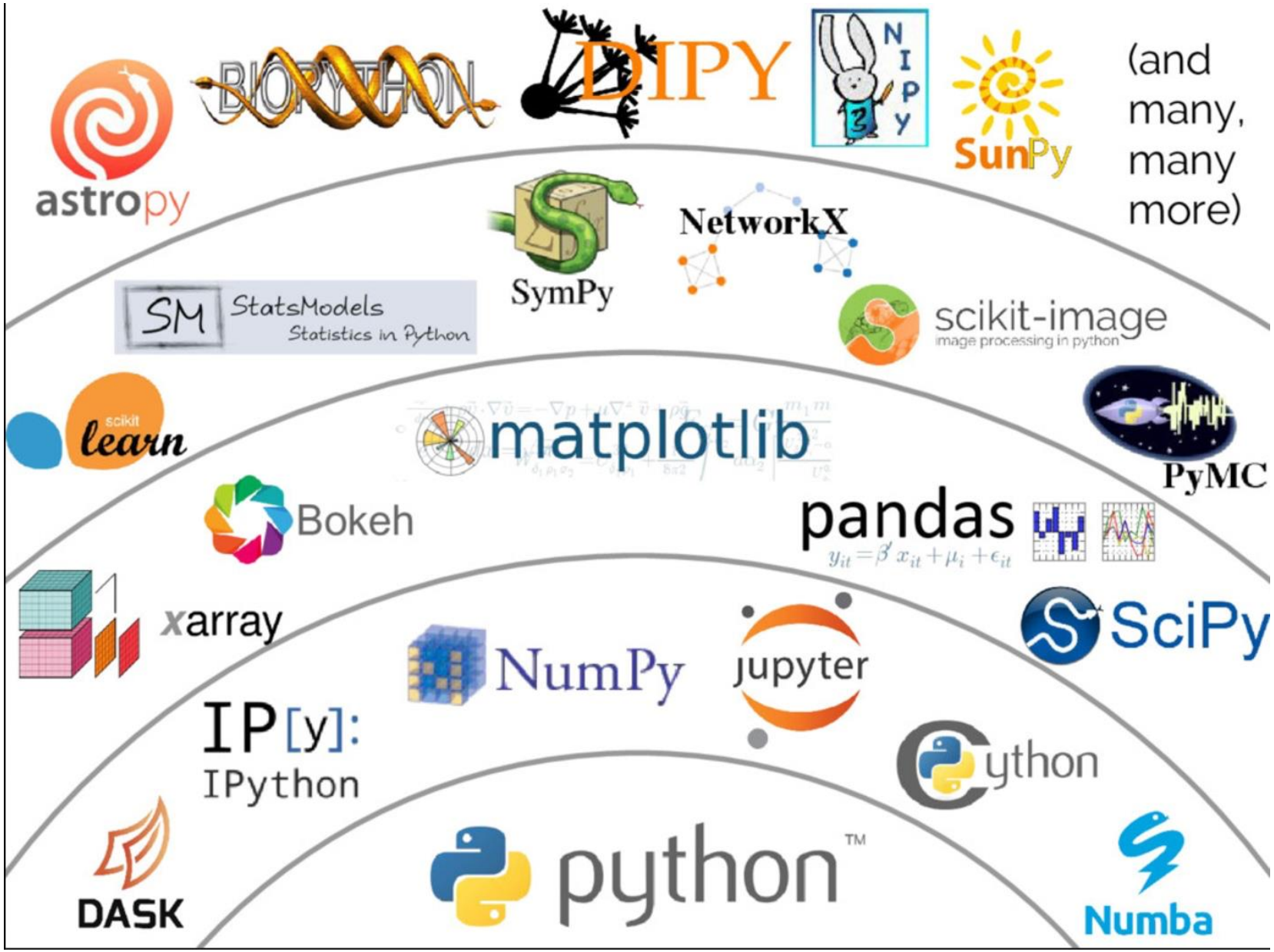
- Many reasons. Traditional ML has been around for decades. But, explosion in availability of labeled data has allowed these algorithms to ‘come to life’. Think Netflix, Amazon, Flight pricings, text to speech, etc.
- Often we have a lot of data, but we cannot make sense of it with traditional models. So,…
- It is much easier now, as the packages to apply ML have become popularized and adopted into mainstream.
- That’s where python comes into play!

giphy.com

So why python?



So why python?



From Jake van der Plas talk

But what's it used for in materials science?

- If you have repetitive tasks, it can be used to automate them
- If you need to fit functions but don't know the function, you can use an ML method
- If you need to identify and track objects in images or movies
- If you need to understand spectral datasets
- Predicting properties from structure or processing

But what's it used for in materials science?

- E.g. use of Neural networks for fitting potential energy landscapes (see https://aip.scitation.org/doi/abs/10.1063/1.5003074)
- Gaussian Processes for interpolation and prediction of experimental data
- Support vector machines for phase transitions (e.g., https://www.nature.com/articles/nphys3644)

Machine Learning

Supervised

“Give me some examples”

Unsupervised

“I don’t need no examples”

In both cases: Machine learning models learn from the data at hand

Supervised Methods

Training Phase

“Learning the wheels”

Testing Phase

“Let loose on the road”

Supervised Methods

In scikit-learn

The training phase is called “Fit”

The testing phase is called “Predict”

Where to go for help

- Online courses: Andrew Ng, Udacity, others
- Textbooks:
  - “Deep learning” (Goodfellow, Bengio, Courville) [https://www.deeplearningbook.org],
  - “The elements of statistical learning” (Hastie, Tibshirani, Friedman) [https://web.stanford.edu/~hastie/Papers/ESLII.pdf]
  - “An Introduction to Statistical Learning,” [http://www-bcf.usc.edu/~gareth/ISL/]
- Scikit-learn documentation and examples
  - http://scikit-learn.org/stable/
- Pycroscopy
  - https://pycroscopy.github.io/pycroscopy/about.html