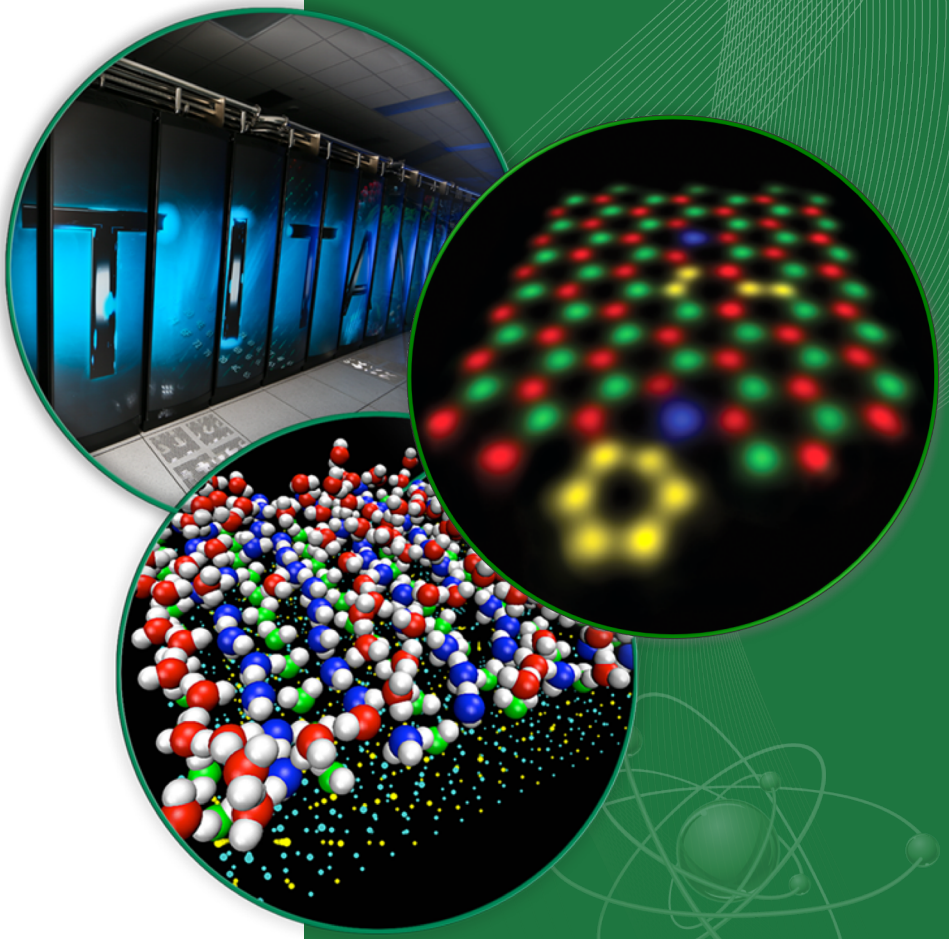
# Machine learning for materials science with python




**Rama Vasudevan**

CNMS User Meeting Workshop
Oak Ridge National Laboratory
August 15th 2018

# Pre-requisites

Details and data required for today's workshop are present here:

https://github.com/pycroscopy/CNMS_ML_in_MS_Workshop_2018

# Pre-requisites

To get the most out of this tutorial, you may want to work along with us on your own laptop. If you do, we highly recommend that you:

1. Download and install the following:
   - **Anaconda** 5.2 for python 3.6 - https://www.anaconda.com/download/
   - **Pycroscopy** - https://pycroscopy.github.io/pycroscopy/install.html
   - **HDFview** - https://support.hdfgroup.org/products/java/release/download.html
2. Familiarize yourself with basics concepts. Links to quick and helpful tutorials are available here - https://pycroscopy.github.io/pyUSID/external_guides.html . Topics include:
   - Basic **python** usage
   - **Jupyter notebooks** for running code
   - **Numpy** for numeric operations

   If you have any trouble or questions with installations etc., please go to:

   https://groups.google.com/forum/#!forum/pycroscopy

# Agenda

| Time | Details |
|---|---|
| 8:00-9:00 AM | Setup and installation |
| 9:00-9:30 AM | Introduction to machine learning in python |
| 9:30-10:15 AM | Linear unmixing methods for spectral analysis |
| 10:15-10:30AM | Coffee Break |
| 10:30-12:00 PM | Nonlinear unmixing: overview and applications |
| 12:00-1:00 PM | Lunch on your own |
| 1:00-2:00 PM | Traditional classification methods |
| 2:00-2:45 PM | Clustering methods applied to 4D STEM data |
| 2:45-3:00 PM | Break |
| 3:00-4:30 PM | Neural networks and deep learning for microscopy images |
| 4:30-5:00 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.

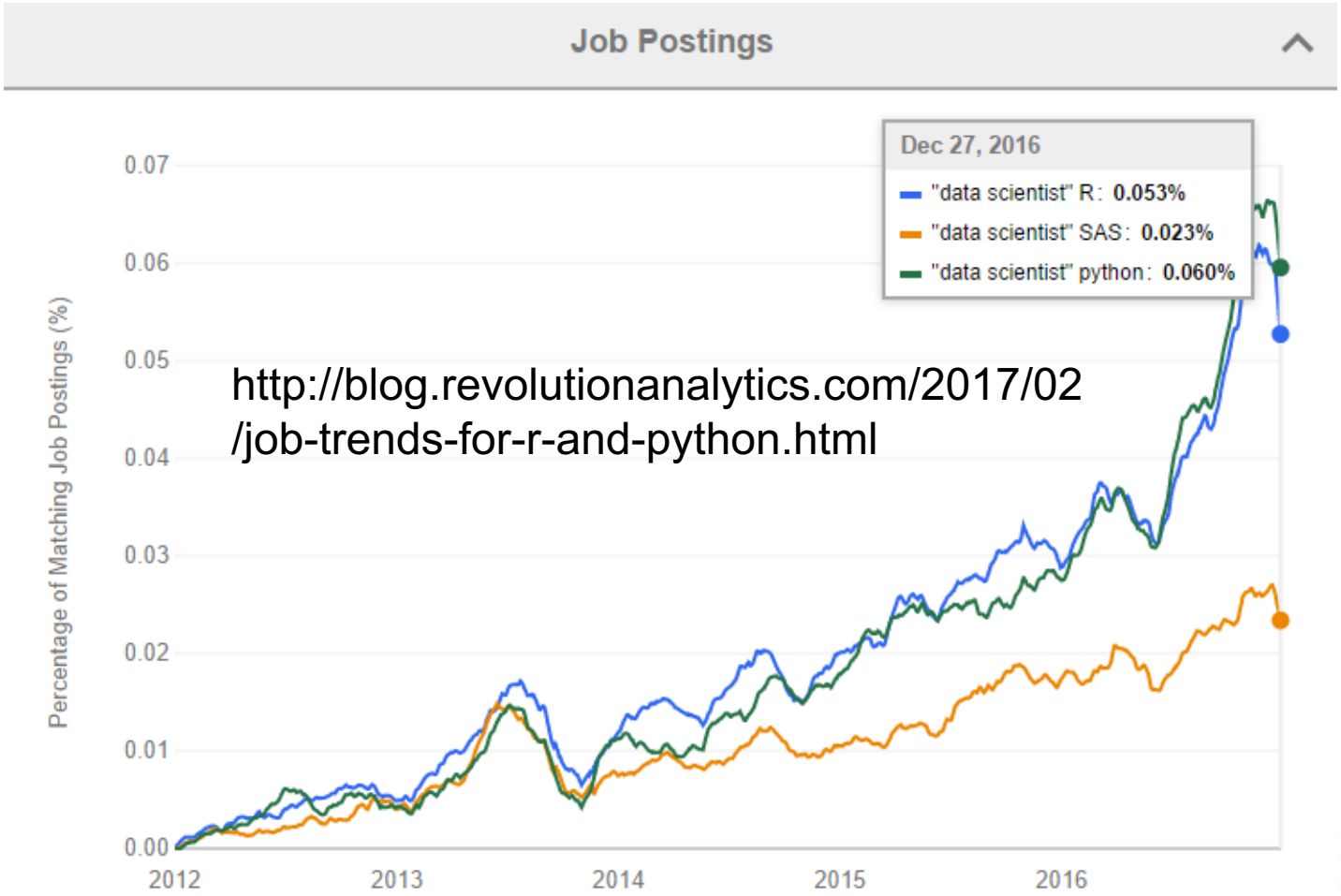
# What *is* Machine learning?

- 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?)

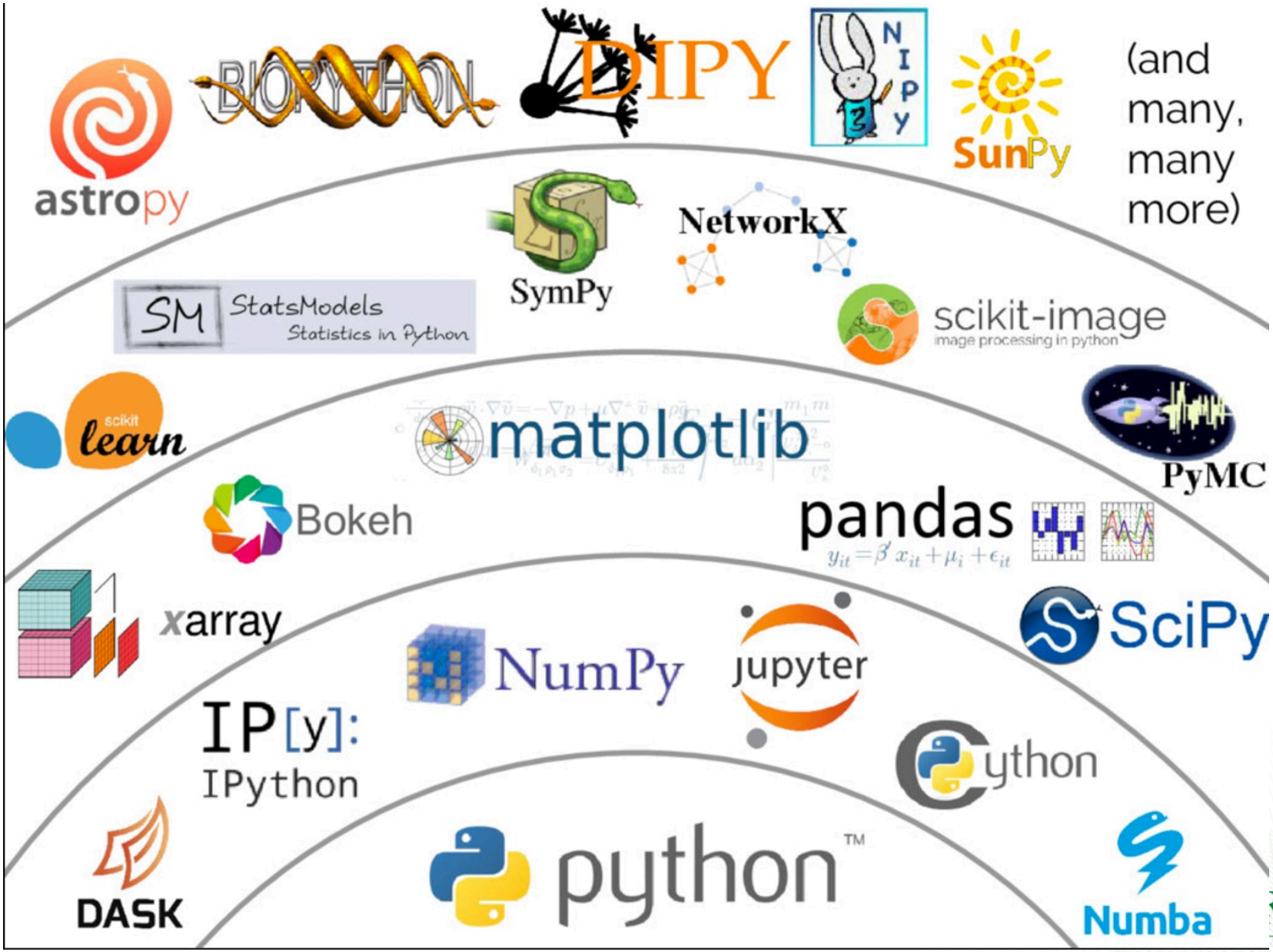
# But why is it popular?

- 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!

# So why python?



# So why python?



# 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