



JONES COLLEGE OF BUSINESS

Operationalizing Python for Machine Learning Data Mining and Predictive Analytics



Operationalizing Python for Machine Learning

Charlie H. Apigian
Interim Director - Data Science Institute
Middle Tennessee State University
August 20, 2018

I AM *true*
BLUE


Overview of Presentation

- Focus for today
 - Help debunk misperceptions of how and who can learn machine learning
 - Understand the generalities of ML and how it can be applied
 - Establish a process for operationalizing machine learning
 - Work through a problem showing one process
 - Understand where you fit into the Data Science Institute world and how you can get started.

mtsu.edu/dsi

using data for good

Outline of Presentation

- About  and the Data Science Institute at MTSU
- Setting the scope for Data Science and Machine Learning
 - And where you fit in
- Outlining the process for operationalizing a ML problem
- Create and work through a process that includes:
 - Supervised learning
 - Classification and Regression
- Understand what this means for you and your career
- YOUR next steps

What is the Data Science Institute?



- The Data Science Institute strives to be the leader in knowledge and research in the area of data science and big data concepts.
- Data Science Institute strategic focus areas:
 - Interdisciplinary Faculty Collaboration
 - Big Data Research Projects
 - Industry and Government Partnerships
 - Community Involvement

What a Data Scientist?

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand what a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting languages e.g. Python
- Statistical computing packages e.g. R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Hig
- Custom reducers
- Experience with xaaS like AWS

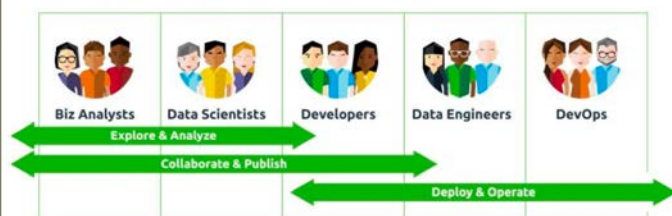
DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Resilient
- Proactive culture
- Outgoing, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Strong writing skills
- Translates data driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization

DATA SCIENCE AS A TEAM SPORT



ANACONDA

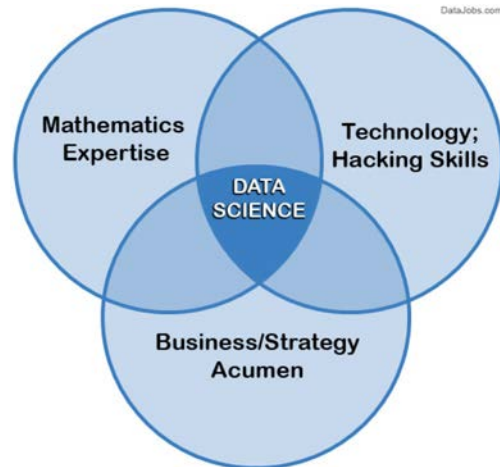
© 2017 Anaconda Inc. All Rights Reserved.

4

Start to think of Data Science as a team sport.

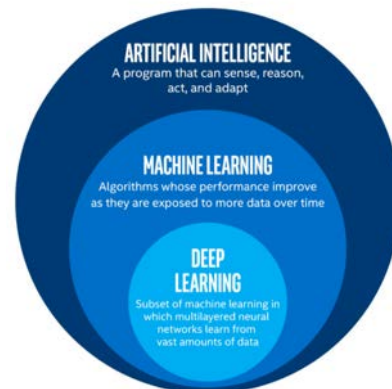
Which skills should you learn first?

Depends on your objective and your interest.



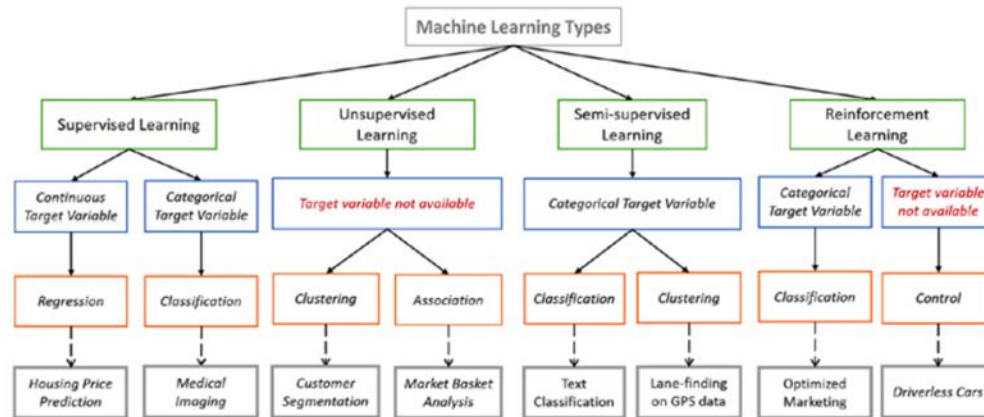
What is machine Learning?

- **Machine learning** is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.



Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

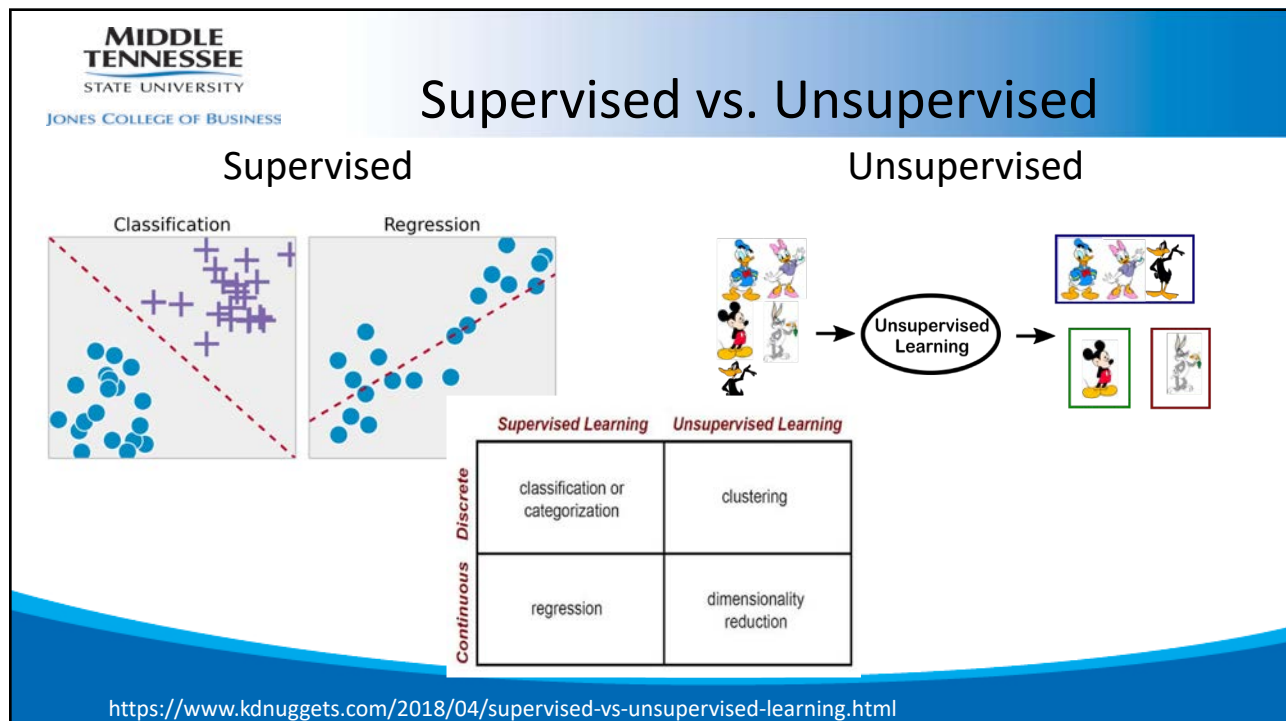
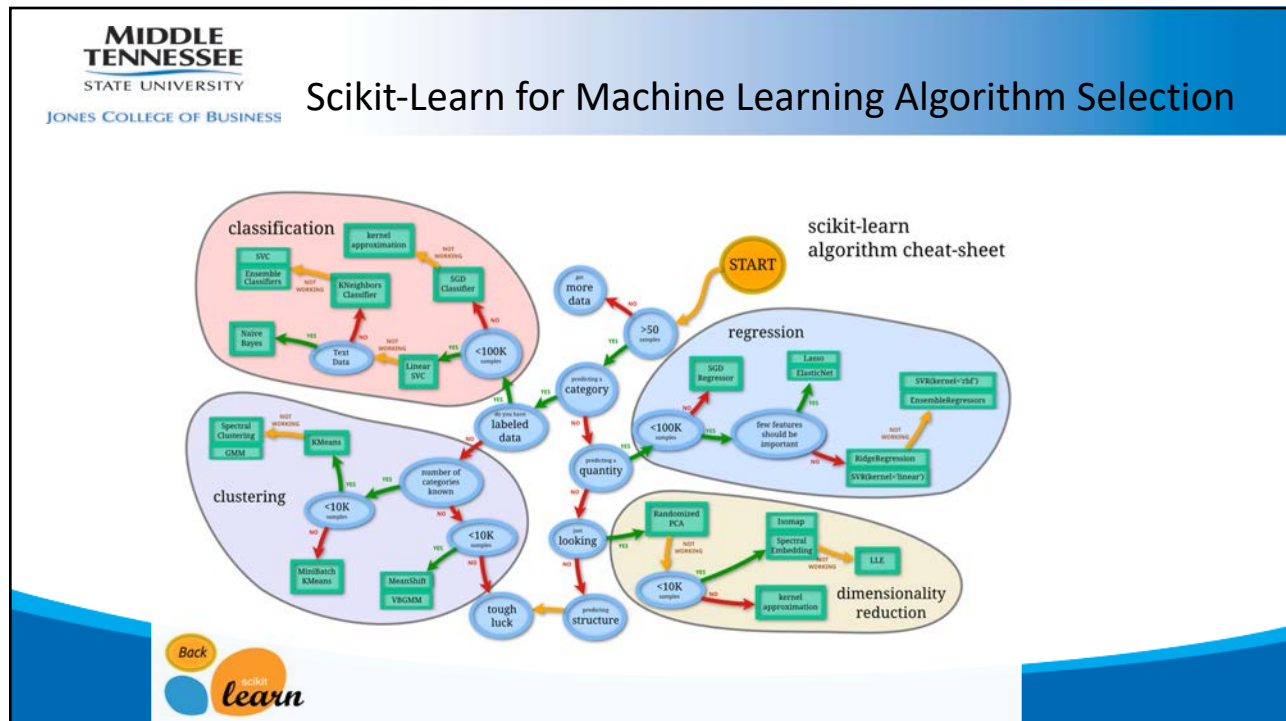
Different Types of Machine Learning



towardsdatascience.com

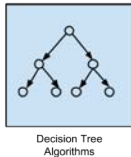
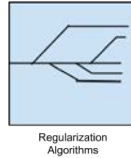
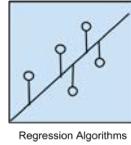
Machine Learning Algorithms Mind-Map



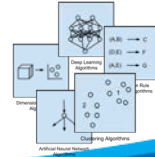
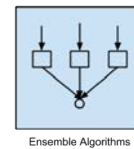
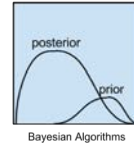


Types of Algorithms

- Regression Algorithms
 - Ordinary Least Squares Regression (OLSR)
 - Linear Regression
 - Logistic Regression
 - Stepwise Regression
- Regularization Algorithms
 - Ridge Regression
 - Least Absolute Shrinkage and Selection Operator (LASSO)
 - Elastic Net
- Decision Tree Algorithms
 - Classification and Regression Tree (CART)
 - Iterative Dichotomiser 3 (ID3)
 - Chi-squared Automatic Interaction Detection (CHAID)



- Bayesian Algorithms
 - Naive Bayes
 - Gaussian Naive Bayes
 - Multinomial Naive Bayes
- Ensemble Algorithms
 - Boosting
 - Bootstrapped Aggregation (Bagging)
 - AdaBoost
 - Gradient Boosting Machines (GBM)
 - Random Forest
- Clustering Algorithms
- Association Rule Learning Algorithms
- Dimensionality Reduction Algorithms
- Deep Learning Algorithms
- Artificial Neural Network Algorithms
- ...and many others



Process for Machine Learning

1. Frame the Problem?
 - What is the problem and objectives?
2. Setup the workspace
 - Setup the editor
 - Import the libraries – pip install if needed
 - Folder Management
3. Get the data
4. Explore the data
 - Visualize, describe, group, etc.
5. Cleanse and merge the data
6. Transform the data
 - Change label columns to numerical/dummy variables
7. Split the data
 - 7. Train Test Split
 - 8. Standardize X_train and X_test
8. Fine-tune the model
 - 7. K-folds
 - 8. For loop alpha scores
 - 9. Grid Search
9. Evaluate the final model
 - Accuracy scores
 - Confusion Matrix

Process for Machine Learning

1. Identify the business problem and objectives
2. Import the data and libraries
3. Cleanse and merge the data
4. Transform the data
5. Create target and independent variables
6. Create test and train data
7. Identify appropriate test for the data
 - Supervised – Classification or regression
 - Unsupervised – Clustering or recommender systems
8. Train the algorithm
9. Fit with the training data
10. Predict with the testing data
11. Measure its accuracy

Platforms for Data Science?

- Possible platforms
 - R
 - SPSS
 - SASS
 - Minitab
 - Excel and its plug-ins
 - Python

Why Python?

- Python libraries continue to grow, which makes it a viable option for data science.
- Its growth in data science has exploded over the past few years, mainly due to libraries like pandas and scikit-learn.

<https://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html>

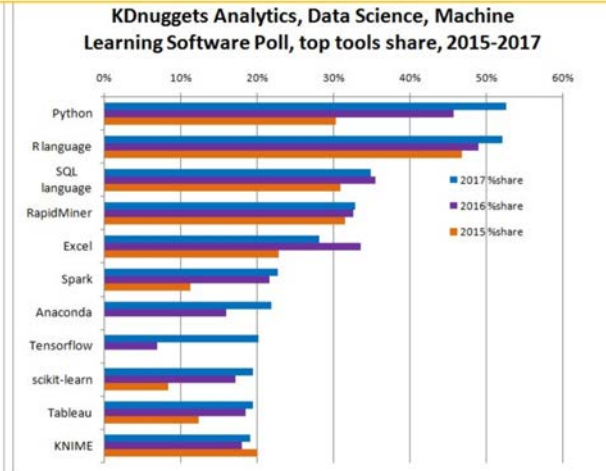


Fig 1: KDnuggets Analytics/Data Science 2017 Software Poll: top tools in 2017, and their usage in the 2015-6 polls

What will we use?

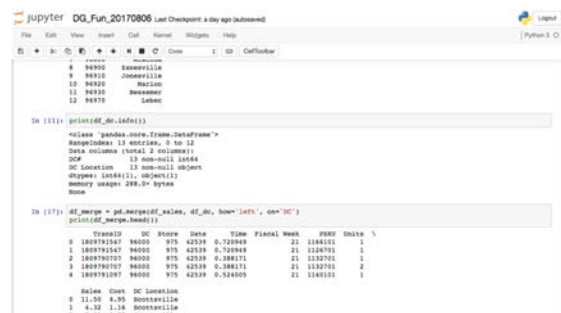
Anaconda

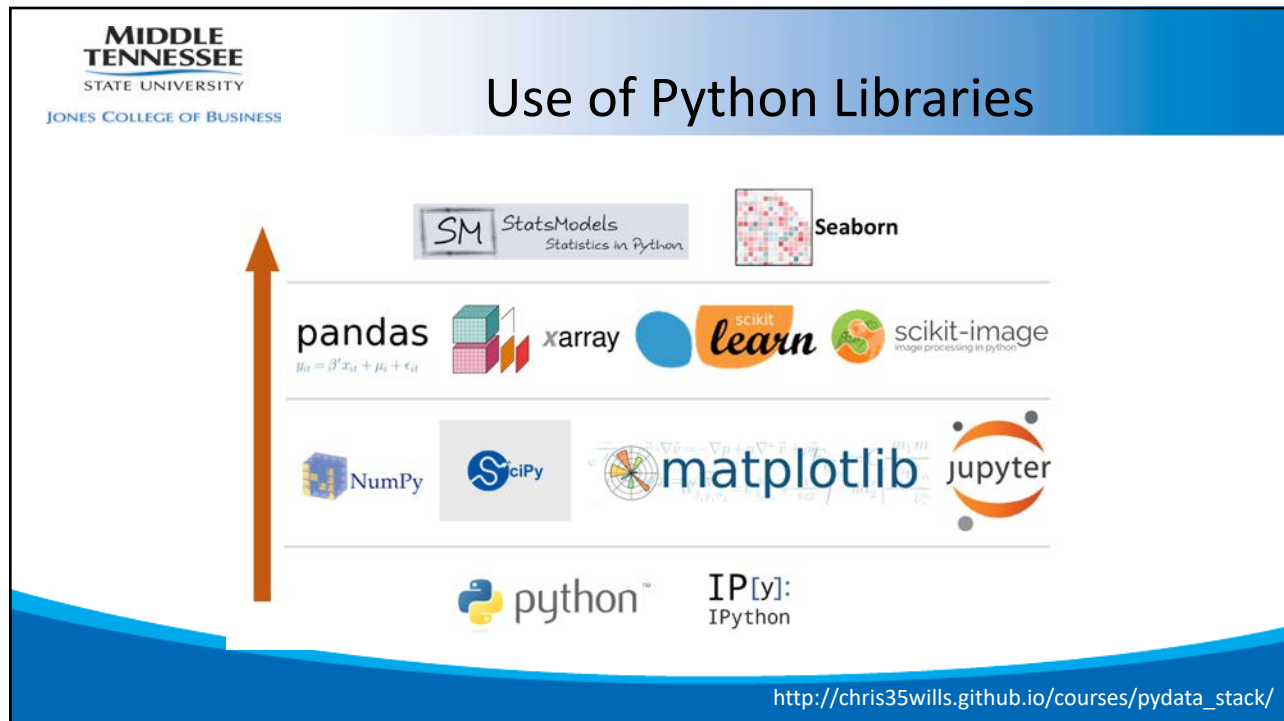
<https://anaconda.org/>



Jupyter Notebooks

<http://jupyter.org/>





MIDDLE TENNESSEE STATE UNIVERSITY
JONES COLLEGE OF BUSINESS

Python - basic building blocks

- The real power of Python for data analysis comes when we move up a row. Numpy, SciPy, Matplotlib and JuPyter notebooks are the fundamental building blocks for your data analysis scripts:
 - NumPy**: provides N-dimensional numerical arrays (or matrices in MATLAB speak), linear algebra, Fourier transforms.
 - SciPy**: builds closely on NumPy, providing more advanced numerical methods, integration, ordinary differential equation (ODE) solvers...if you've come across the book called 'Numerical Recipes', there's a good chance that you'll find those algorithms implemented in SciPy.
 - Matplotlib**: Python's main graphing/plotting library. Make all the pretty plots. The documentation on the Matplotlib website is good, especially the [gallery](#). **All the packages on this page with plotting capabilities rely on Matplotlib under the hood.**
 - Jupyter**: rather than using the interactive IPython command line, you might want to use Python in a 'notebook' style from inside your web browser, which keeps your commands and their outputs together in a single document that you can re-open later on. Particularly worth looking into if you do a lot of statistics.

http://chris35wills.github.io/courses/pydata_stack/

Analysing and manipulating your data

- [Pandas](#) - Number-one most important tool for data science. High-performance data structures and data analysis tools. Takes a whole load of complexity out of loading tabular data into Python for analysis, especially CSV files, Excel files, SQL databases... Labels your data nicely with column headings and indexes. Does lots of basic statistics and offers plotting facilities to quickly take a look at your data. Particularly good if you have to work with time series data. Work through the [course tutorial](#) on Pandas.
- [scikit-learn](#) - machine learning tools for Python. Increasingly popular, contains all the main algorithms used in this field such as K-means clustering. Check here before deciding that you need to write your own algorithm from scratch!
- **The top level - advanced statistics**
 - [Statsmodels](#) - provides implementations of all the major statistical algorithms. Preferentially works with Pandas DataFrames. Has the option of using R-like syntax, which you'll probably like if you're familiar with R.
 - [seaborn](#): a set of statistical plotting tools. The plots look very elegant. Well worth looking at if you do a lot of statistical work. Takes Pandas DataFrames as stan

http://chris35wills.github.io/courses/pydata_stack/

Python example

Let's get started with Python and
Jupyter notebooks



Predicting a Bad Loan

- Lending Club is a peer-to-peer loan company. Appleton provides personal and commercial loans to borrowers, which are then backed by investors in the loans.
- The peer-to-peer model allows for lower interest-rates to borrowers and for investors to earn money on the interest paid back to them.
- Over 80% of the loans provided by Appleton are personal. These loans are mostly made by borrowers in order to consolidate debt or pay off credit cards, but they may be provided for numerous reasons such as weddings, vacations, and for small businesses.

Appleton Lending

- **Business Problem**
 - Appleton would like to predict if a loan will be good or bad when the applicant applies for the loan.
- What data about the applicant would you want to analyze?
- What demographic data would you want to analyze?
- What type of testing would you want to do?