In [ ]:

```
%matplotlib inline
import matplotlib
import seaborn as sns
sns.set()
matplotlib.rcParams['figure.dpi'] = 144
```

Introduction to Machine Learning

Module description

Machine learning has come to the forefront of data analytics in the last 10-15 years and is a much-heard term in the news and other media. But what is it?
The proliferation of cheap sensor devices, the ubiquity of mobile phones (and the data they can collect), social media, and other data sources provides information from which machines can try to learn valuable patterns from structured data.

This module introduces the basics of machine learning (with a focus on "supervised learning"), including concepts such as regression, classification, model evaluation metrics, over/under fitting, variance versus bias, ensemble methods, model selection, and hyperparameter optimization. We also give an introduction to "unsupervised learning" including clustering and dimensionality reduction.

We cover the powerful and widely-used Python package `Scikit-Learn` and other common tools to provide a strong understanding of core concepts in machine learning, as well as the ability to efficiently train and benchmark accurate predictive models. Students gain hands-on experience building complex "Extract-Transform-Load" (ETL) pipelines to handle data in a variety of formats, developing models with tools such as feature unions and pipelines to reduce duplicate work, and practicing parallelization to speed up prototyping and development.

Learning outcomes

At the conclusion of this module, students should:

- Know the goal of machine learning, and how this differs from statistical analysis
- Understand the two general supervised machine learning tasks:
  - Regression, i.e. prediction of a continuous variable
  - Classification, i.e. prediction of a categorical variable
- Have familiarity with commonly-used regression and classification tools in Scikit-Learn including:
  - Linear Regression
  - Ridge Regression
  - Decision Trees
  - Nearest Neighbors
  - Logistic Regression
- Understand the Bias-Variance tradeoff in model building, and how this is handled using tools such as `GridSearchCV` in Scikit-Learn to help choose appropriate hyperparameters
- Be able to work with the Scikit-Learn API, specifically:
  - Know how to create custom estimators and transformers to manipulate data effectively and build specialized prediction methods
  - Be comfortable with data preprocessing techniques such as feature scaling and one-hot encoding, and the knowledge for when these preprocessing methods are suitable
  - Be able to effectively use built-in tools such as `Pipeline` and `FeatureUnion` to produce a "start-to-end" ETL workflow
- Know what is the purpose and uses of unsupervised machine learning and have familiarity with some basic Scikit-Learn tools used to perform that task

Prerequisites

We assume familiarity with the Python programming language, including data types such as `list`, `dict`, and `tuple`, list and dictionary comprehensions, writing functions, and basic object-oriented programming techniques in Python.

In addition, successful students will have understanding of the Pandas `DataFrame` data structure, and methods to manipulate data in a `DataFrame` such as selecting, filtering, creating new columns from existing columns, and understanding the various ways of indexing and slicing DataFrames. Finally, being able to create and manipulate NumPy arrays is also necessary. NumPy arrays and Pandas DataFrames are commonly-used data structures for interacting with the Scikit-Learn library.

Suggested notebook order

1. `ML_Introduction`
2. `ML_Regression`
3. `ML_Scikit_Learn_API`
4. `ML_Classification`
5. `ML_Overfitting`
6. `ML_Scikit_Learn_Workflow`
7. `ML_Transformers_and_Preprocessing.ipynb`
8. `ML_K_Nearest_Neighbors`
9. `ML_Unsupervised_Learning`

Additional resources

General Machine Learning

- Wikipedia's articles on
    - Machine Learning: https://en.wikipedia.org/wiki/Machine_learning (https://en.wikipedia.org/wiki/Machine_learning)
    - Supervised (Machine) Learning: https://en.wikipedia.org/wiki/Supervised_learning (https://en.wikipedia.org/wiki/Supervised_learning)
    - Unsupervised (Machine) Learning: https://en.wikipedia.org/wiki/Unsupervised_learning (https://en.wikipedia.org/wiki/Unsupervised_learning)

Scikit-Learn

- The online Scikit-Learn documentation is extensive: https://scikit-learn.org/stable/ (https://scikit-learn.org/stable/)
- Scikit-Learn's introductory tutorial: https://scikit-learn.org/stable/tutorial/basic/tutorial.html (https://scikit-learn.org/stable/tutorial/basic/tutorial.html)
- The Scikit-Learn algorithm cheat sheet: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html (https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html)

Pandas and NumPy

Pandas DataFrames and NumPy arrays are commonly-used data structures that interact well with the Scikit-Learn Machine Learning API.

Pandas

- A cheat sheet with a summary of commonly-used Pandas operations: http://pandas.pydata.org/Pandas_Cheat_Sheet.pdf (http://pandas.pydata.org/Pandas_Cheat_Sheet.pdf)
- A blog post discussing the connection between Pandas and NumPy: https://blog.thedataincubator.com/2018/02/numpy-and-pandas/ (https://blog.thedataincubator.com/2018/02/numpy-and-pandas/)

NumPy

- A quick-start tutorial on NumPy: https://docs.scipy.org/doc/numpy/user/quickstart.html (https://docs.scipy.org/doc/numpy/user/quickstart.html)
- A NumPy cheat sheet geared for those transitioning from MATLAB: https://docs.scipy.org/doc/numpy/user/numpy-for-matlab-users.html (https://docs.scipy.org/doc/numpy/user/numpy-for-matlab-users.html)
- 100 NumPy exercises: http://www.labri.fr/perso/nrougier/teaching/numpy.100/ (http://www.labri.fr/perso/nrougier/teaching/numpy.100/)

Data Sources

Where can one find data sets to try out machine learning algorithms? Here's a few online sources.

- University of California Irvine Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php (https://archive.ics.uci.edu/ml/index.php) (This is a well-known repository of common and some not-so-common data sets.)
- Google Dataset Search: https://toolbox.google.com/datasetsearch (https://toolbox.google.com/datasetsearch) (Provides links to various online data sets, try a keyword search.)

The Mathematics of Machine Learning

You can do machine learning without the need to understand the fine details of the mathematics that underlies it, but if you want to delve into this world, here is a good starting point:

- **The Elements of Statistical Learning (https://web.stanford.edu/~hastie/ElemStatLearn/)** by Trevor Hastie, Robert Tibshirani, and Jerome Friedman (A rather comprehensive treatment, available for free online.)
- **An Introduction to Statistical Learning (https://www-bcf.usc.edu/~gareth/ISL/)** by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (A gentler introduction, i.e. less mathematically intense, than the above book.)