Cornell Future Of Care

Karl Leswing

karl.leswing@gmail.com

FB/Insta/Twitter/Github: lilleswing

October 2018

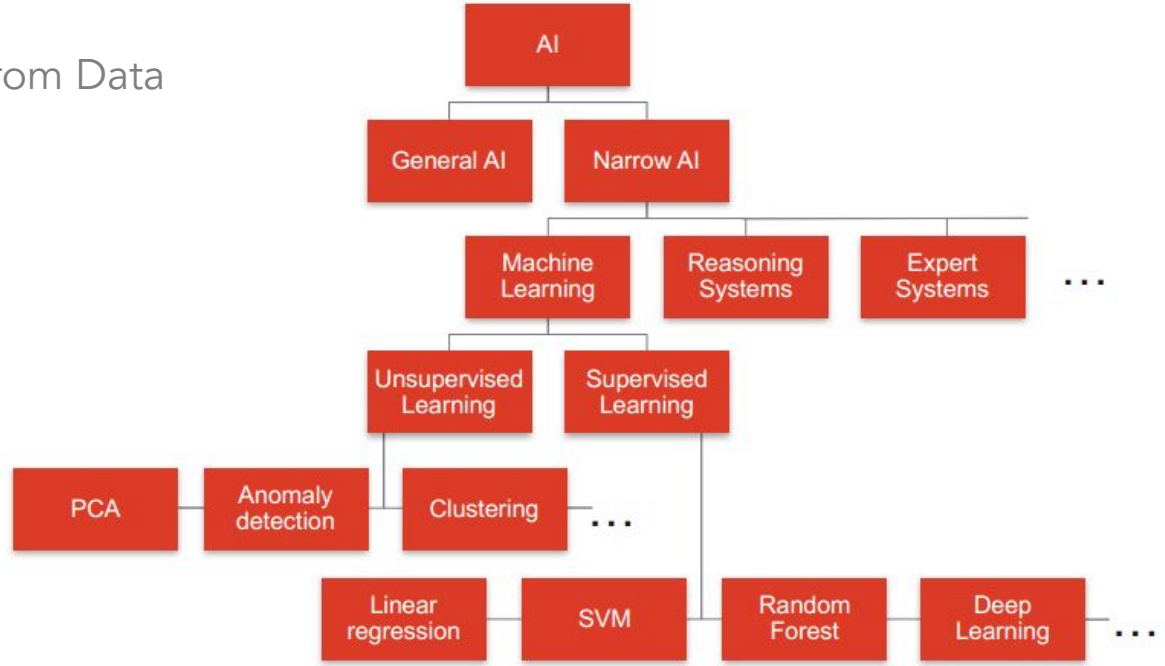
Grab Today's Code

https://github.com/lilleswing/future_of_care

Schrödinger Confidental

What Is Machine Learning

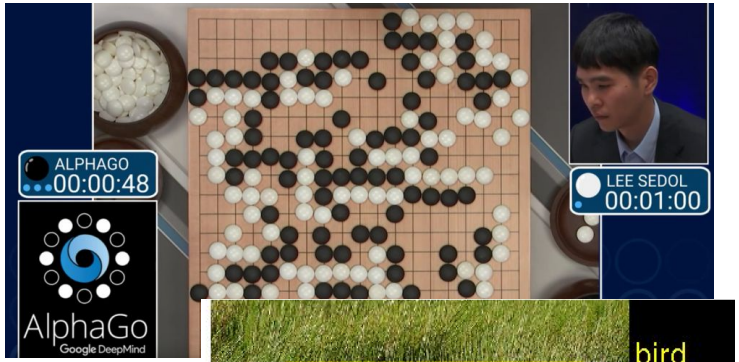
- Supervised Learning
  - Generate Functions From Data

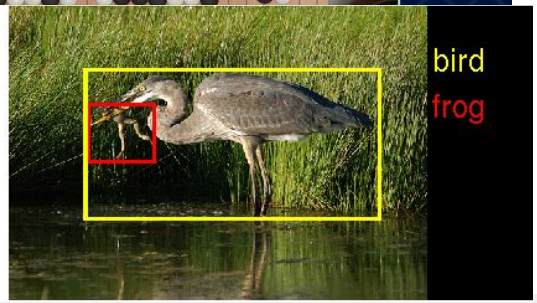


Schrödinger Confidental

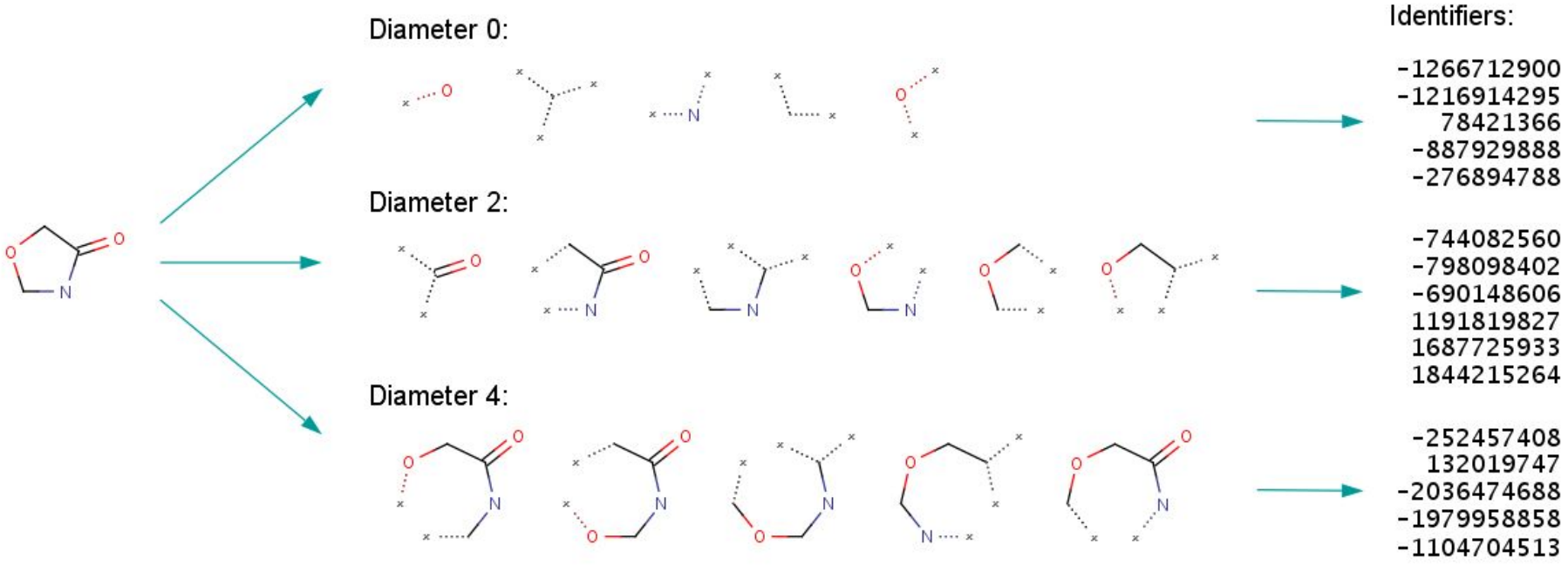
Deep Learning

- Deep learning methods are becoming very popular in image recognition, game playing, and question and answer systems.



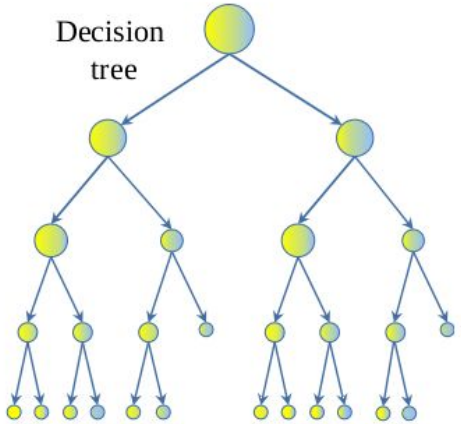


ECFP4


Diameter 0:
Diameter 2:
Diameter 4:
Identifiers:
-1266712900
-1216914295
78421366
-887929888
-276894788
-744082560
-798098402
-690148606
1191819827
1687725933
1844215264
-252457408
132019747
-2036474688
-1979958858
-1104704513

Decision Trees

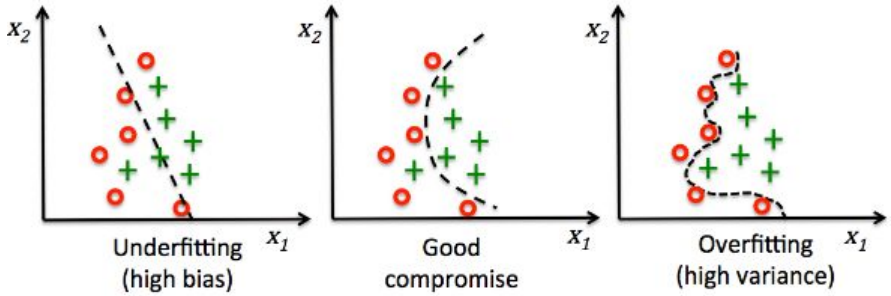
- Based on the data “split” on a cut off of a single feature
- Can use the most informative feature of all the samples
- Leaf nodes hold predicted values



Schrödinger Confidental

High Variance Low Bias

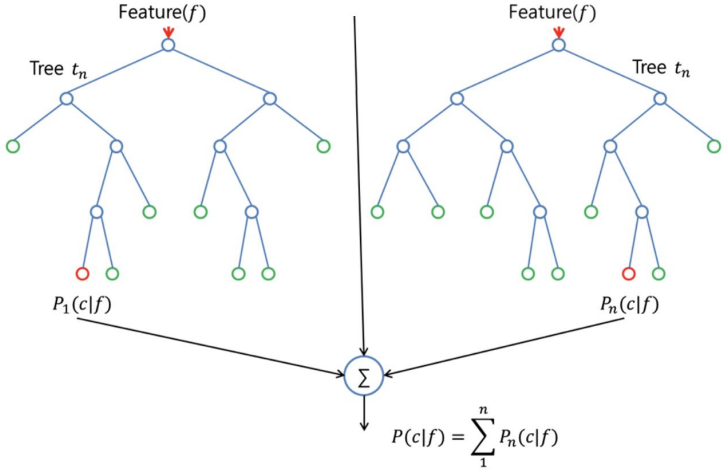
- Standard decision trees are prone to overfitting
- When there is “noise” in the response with respect to the input we need a more general model



Schrödinger Confidental

Bootstrap Aggregating (Bagging)

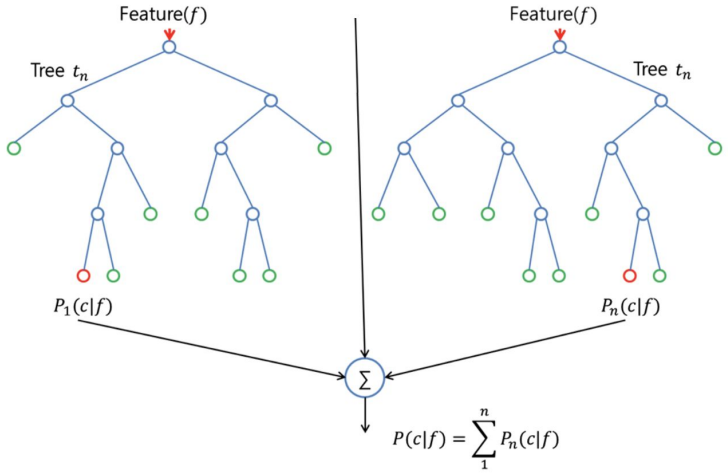
- Make many decision trees!
  - Each one on a subset of the training data, selected uniformly random WITH replacement
- Average results from all decision trees
- More robust to outliers



Schrödinger Confidental

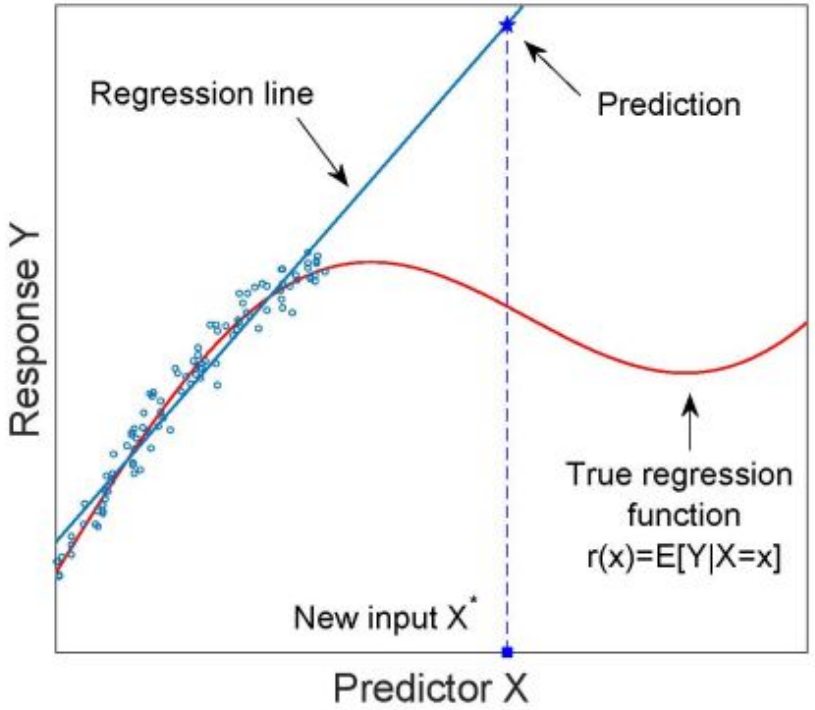
Feature Bagging (Random Forests)

- Many of the trees can be very similar
- E.X if one feature is very predictive all trees use this feature
  - No longer have good ensembling to lower variance
- Solution: At each split only select from a random subset of features



Schrödinger Confidental

Interpolation



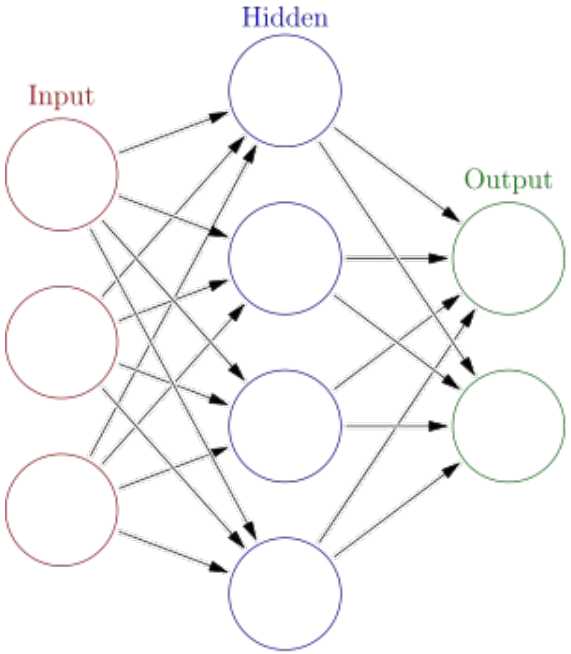
Schrödinger Confidental

Deep Learning

- Lots of excitement to try to use these methods in other contexts
- Should deep learning be used in materials?
- Where does it provide the greatest benefit?

Artificial Neural Network Overview

- Collection of units called neurons (Circles Here)
- Each neuron computes a function over its inputs (real numbers)
- Each neuron and can be connected to multiple outputs
- Trained using back propagation

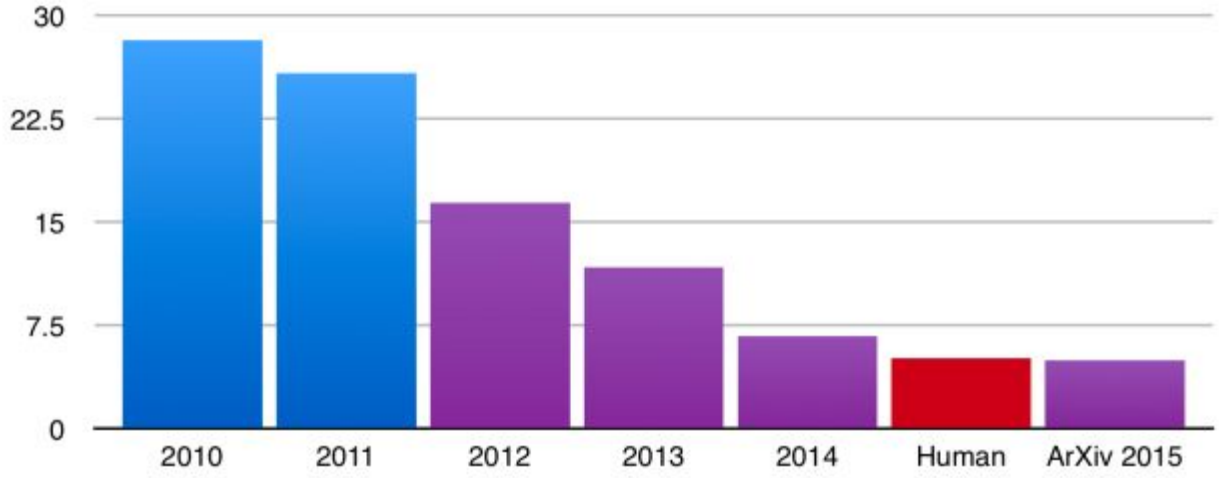


Universal Function Approximation Theorem

- Artificial Neural Networks can represent ANY function
- This does not pan out in practice
  - Limited data and compute power
- Requires us to create data and compute efficient models.

Deep Neural Network Image Classification

ImageNet Large Scale Visual Recognition Challenge Model Accuracy



As of 2015, a 27 layer DNN was more accurate than a human (Stanford student) at sorting 100,000 images into 1,000 different pre-specified categories

Deep Neural Network Image Classification

- The ImageNet classification challenge is *very* difficult:


Ruler


King crab

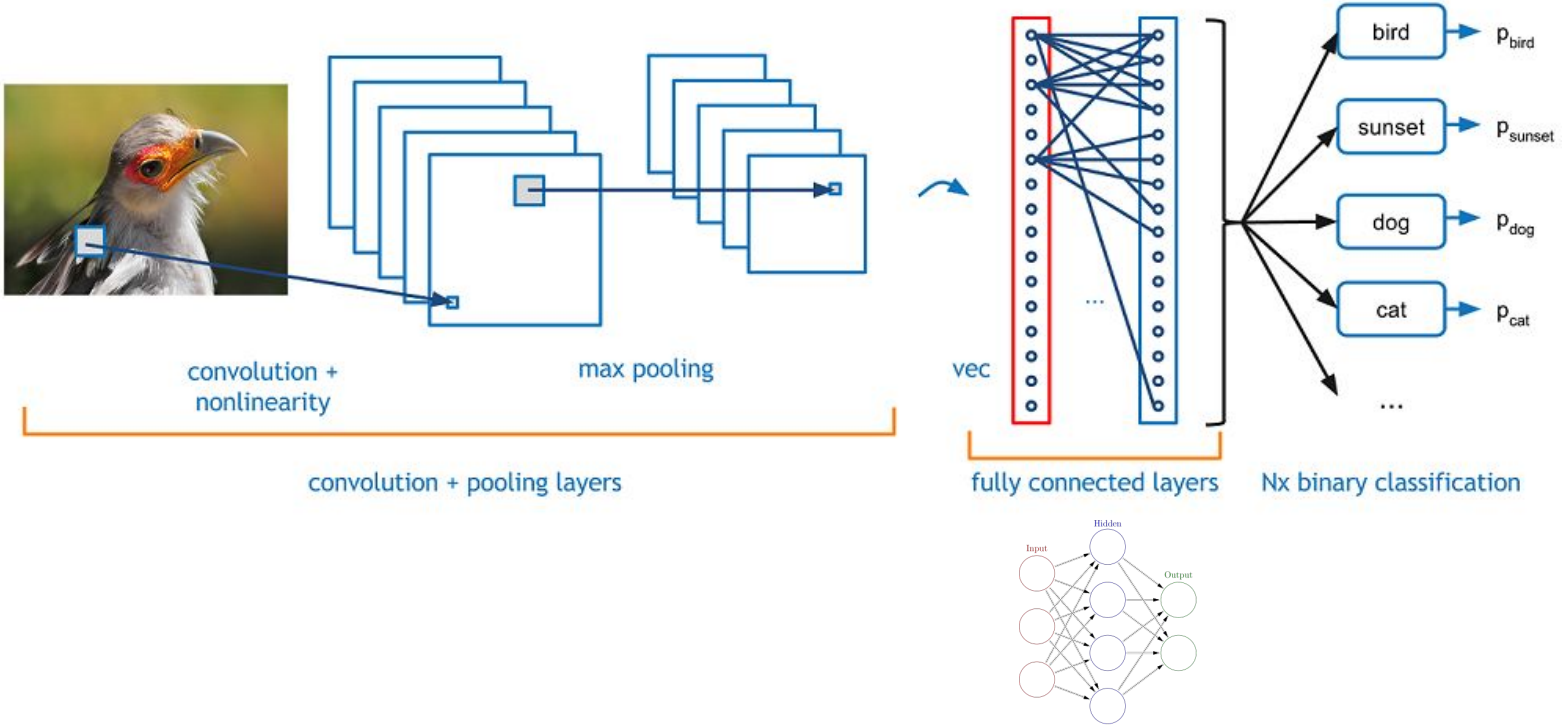

Sidewinder

Salt shaker


Reel


Hatchet

Convolutional Neural Networks



Convolution Layer

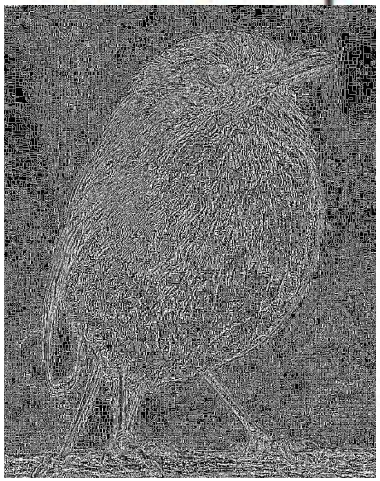
- Slide a learnable mask across the image.

Input image

Convolution Kernel

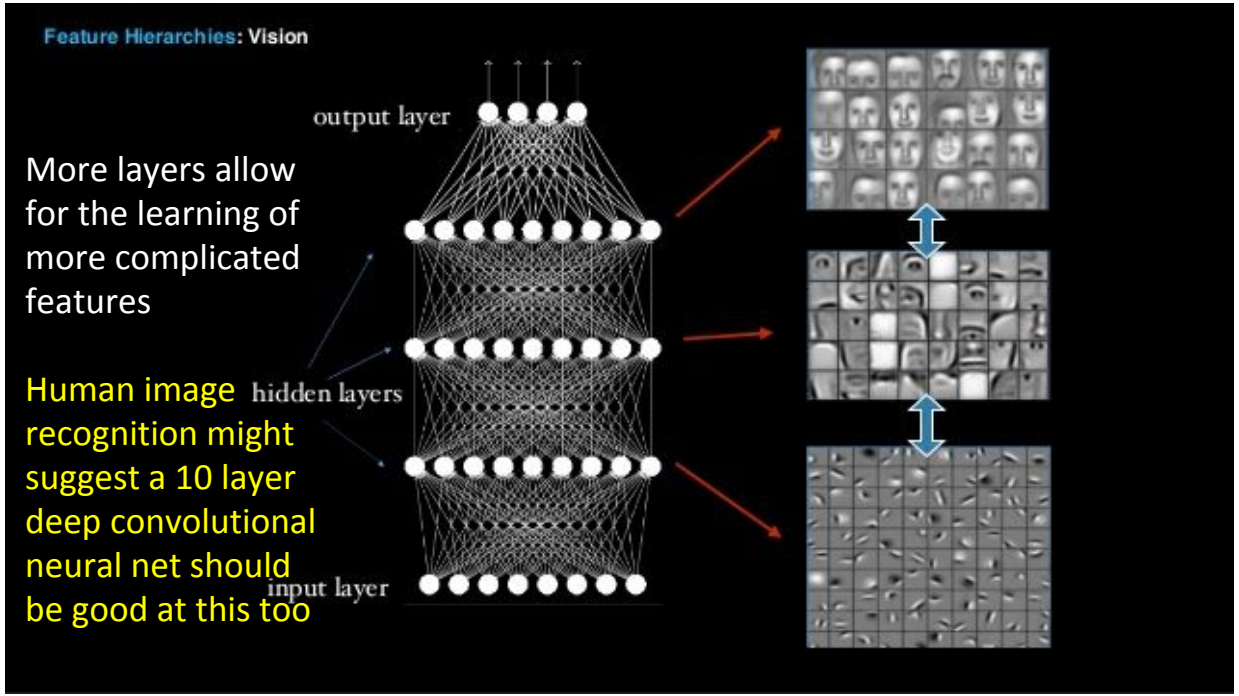
$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Feature map



Deep Neural Network Image Classification

- A unique aspect of Deep Learning is the ability learn new features as the network is trained:



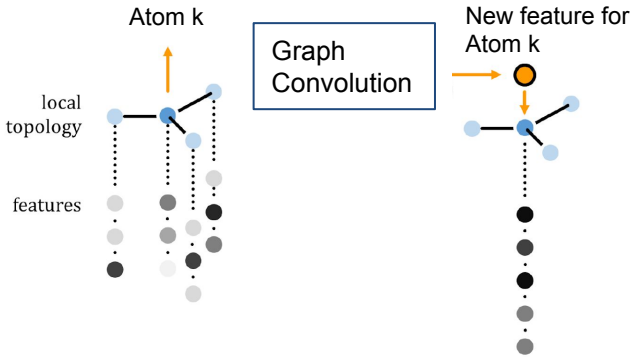
AutoQSAR w/ DeepChem Feature Generation

2D Graphic description of molecules

- Each node represents an atom
- Each edge represents a bond
- Atom features include atoms-type, valences, formal charges, and hybridization

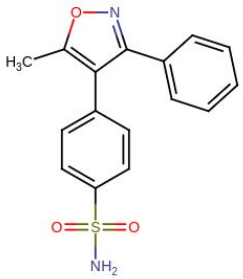
Graph Convolution

- **Automatically learn new local features that suit the endpoint**
- These new features are then converted to molecular feature which is feed to dense neural network for model building

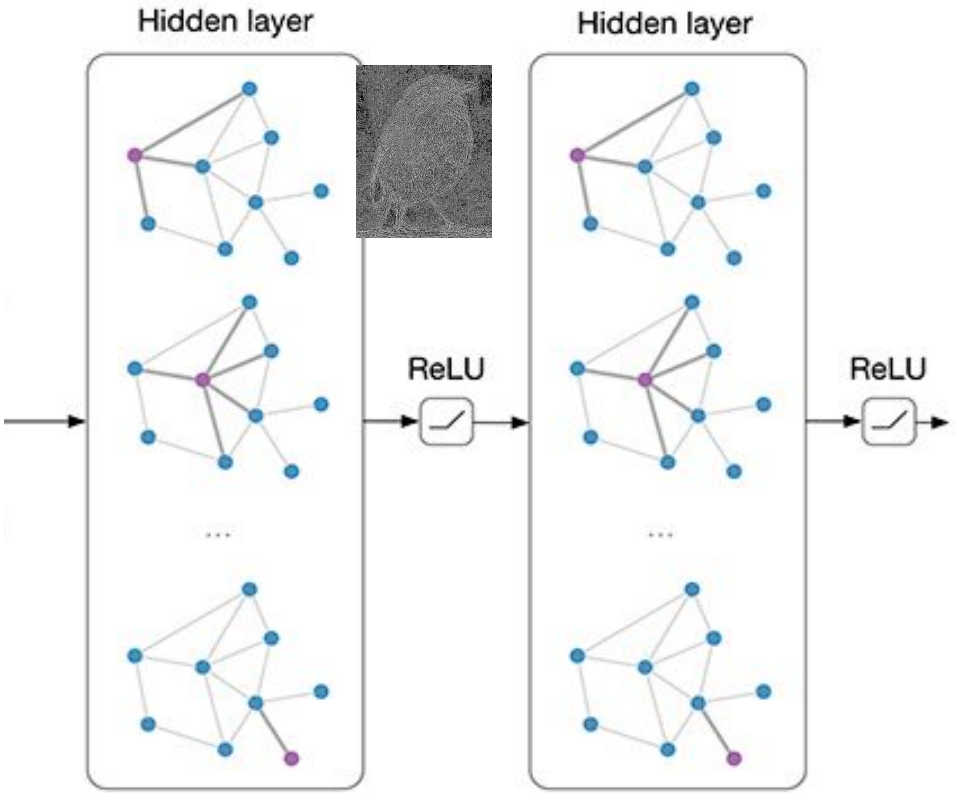


Graph Convolutions

Input



Valdecoxib (pain relief)



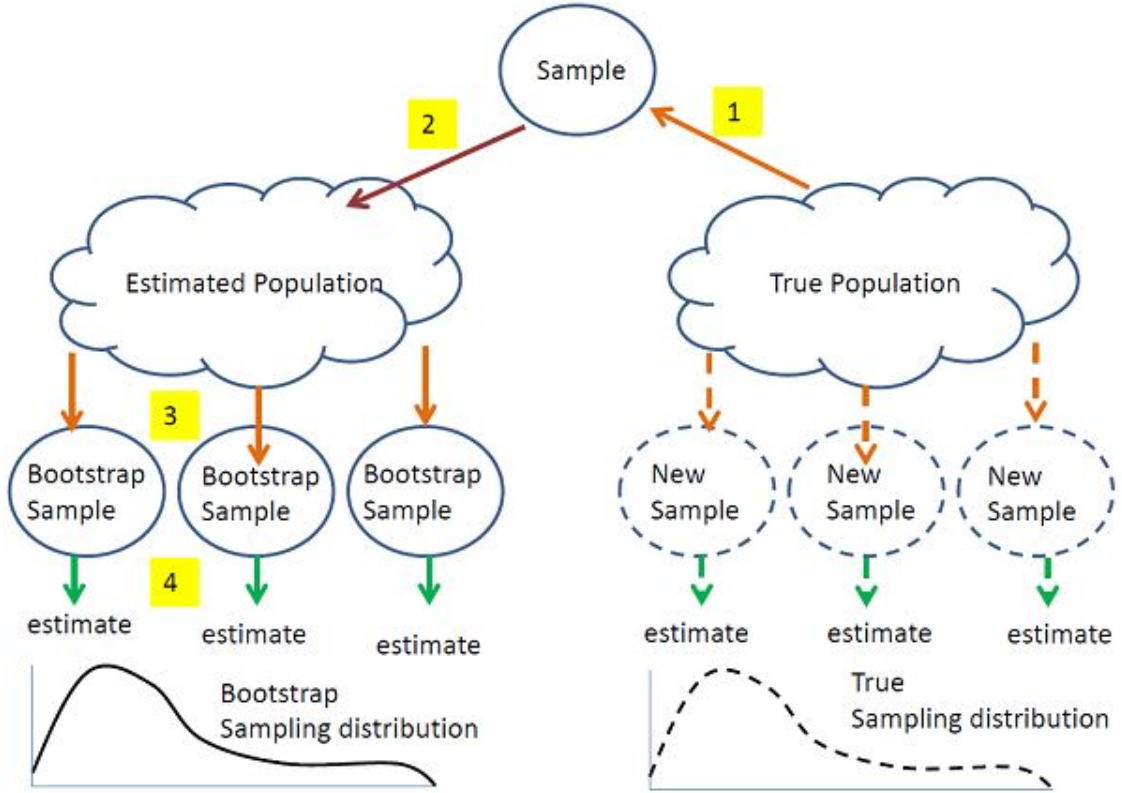
Can We Detect Bad/Interesting Labels In Chemical Data?

08/28/2018

Schrödinger Confidental

Bootstrapping Interesting Datapoint Identification

- Repeatedly train a model on a random 80% of the data
- Predict on remaining 20%
- Find samples whose predictions are farthest from labels



Schrödinger Confidental

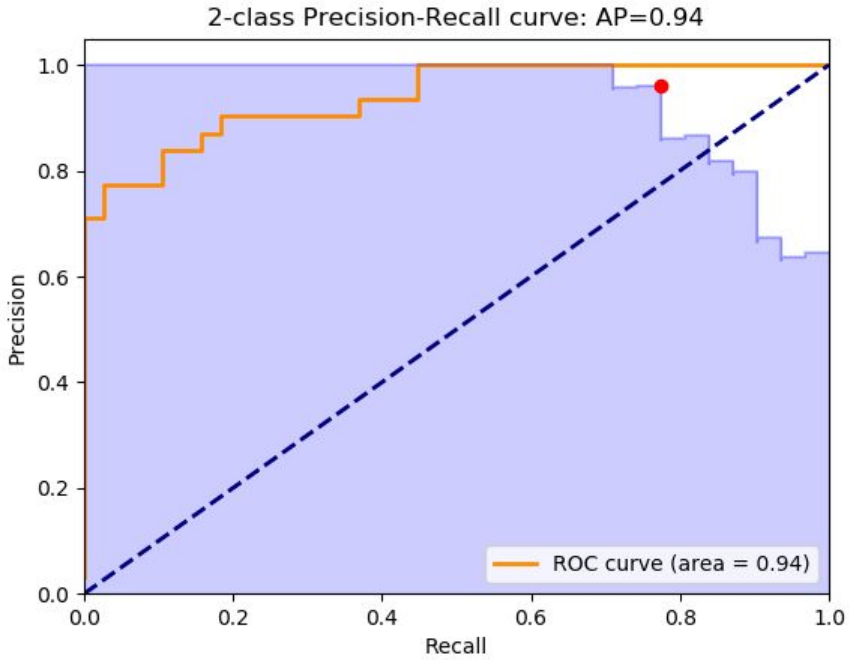
Models

- Classification Model **Descriptors.MolLogP(m) > 0**

- **We will randomly incorrectly label x% [0,10,20,50] of compounds and see if we can find the molecules we incorrectly labeled**

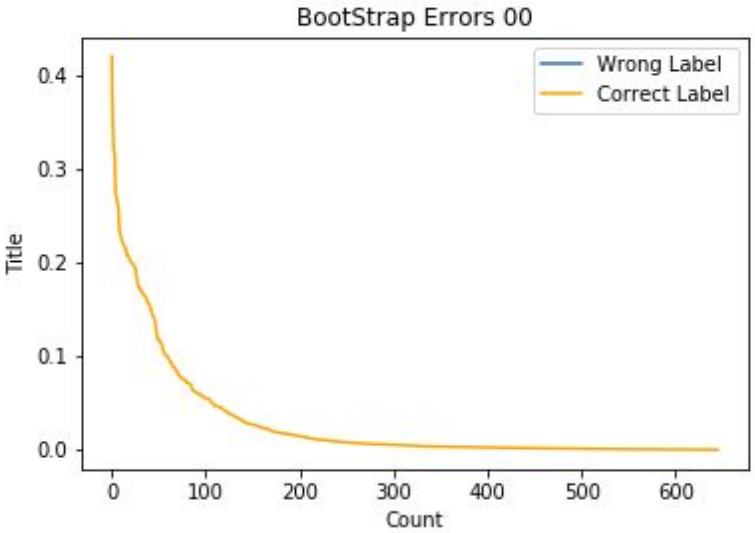
Schrödinger Confidental

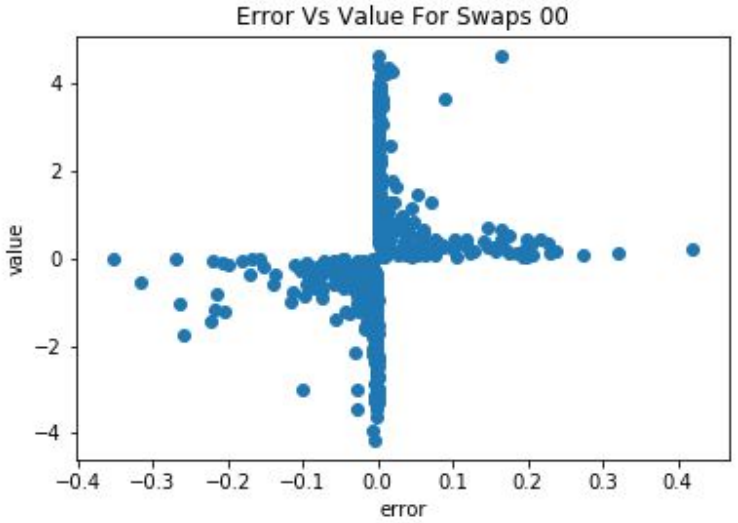
No Flips



Schrödinger Confidental

No Flips

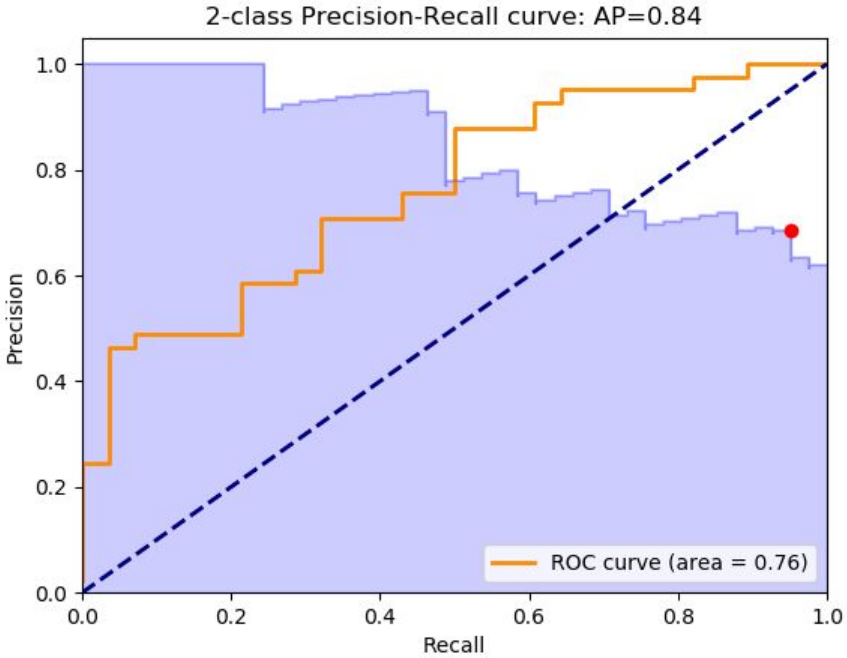


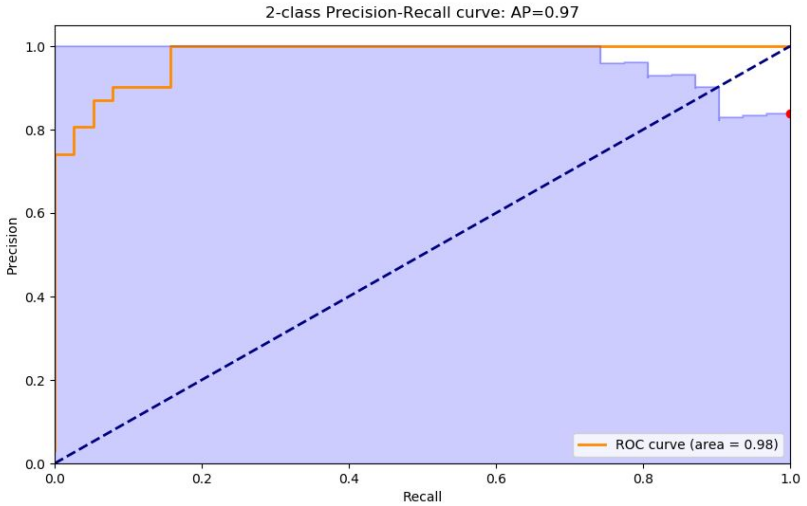


Schrödinger Confidental

10% Flips

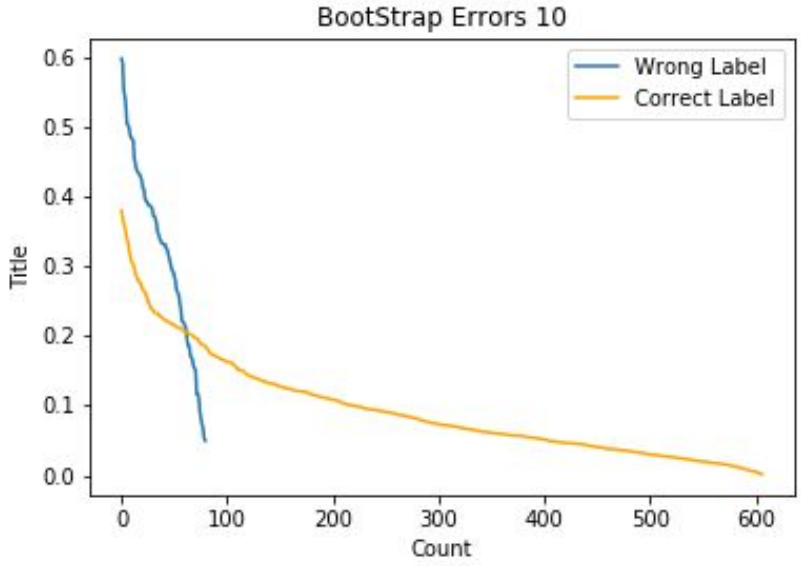
Errored Holdout

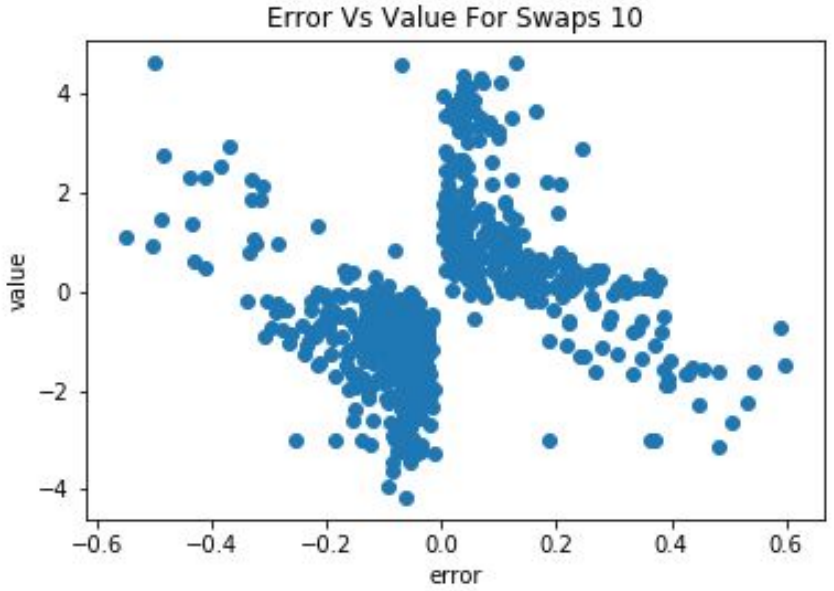




Schrödinger Confidental

10% Flips

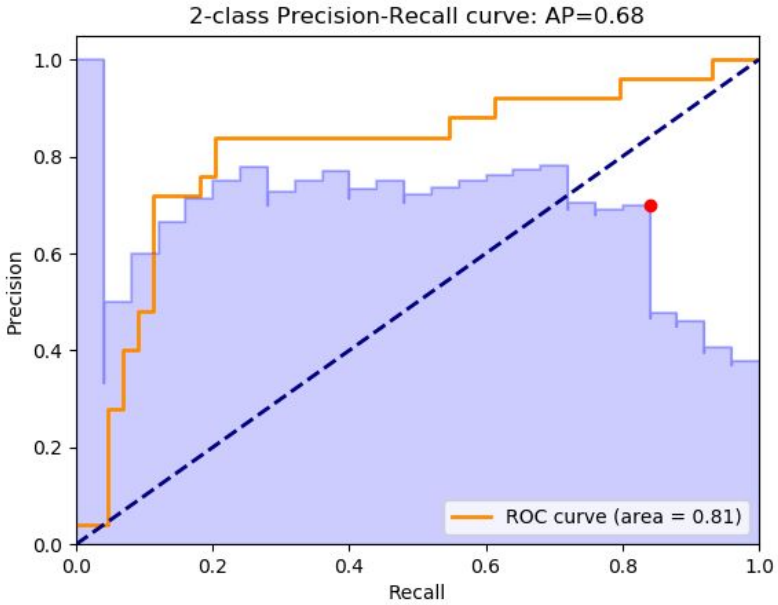


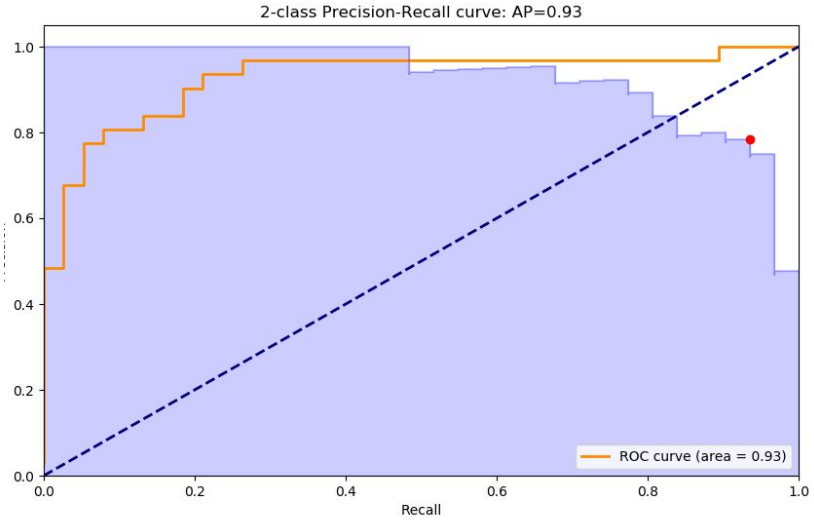


Schrödinger Confidental

20% Flips

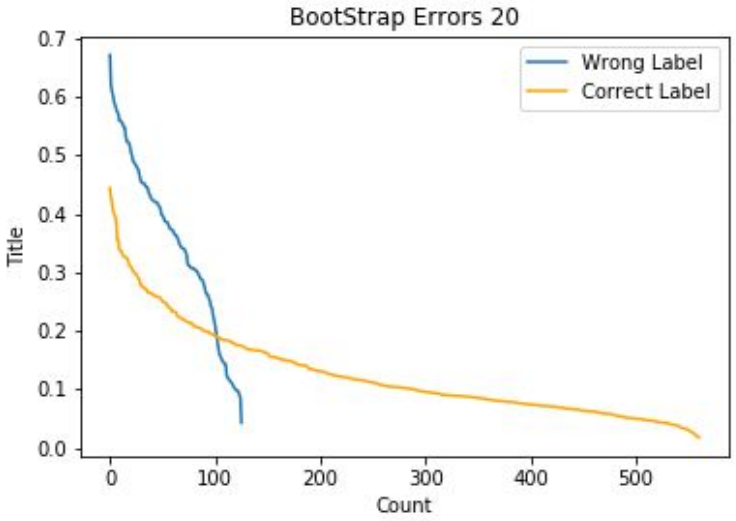
Errored Holdout

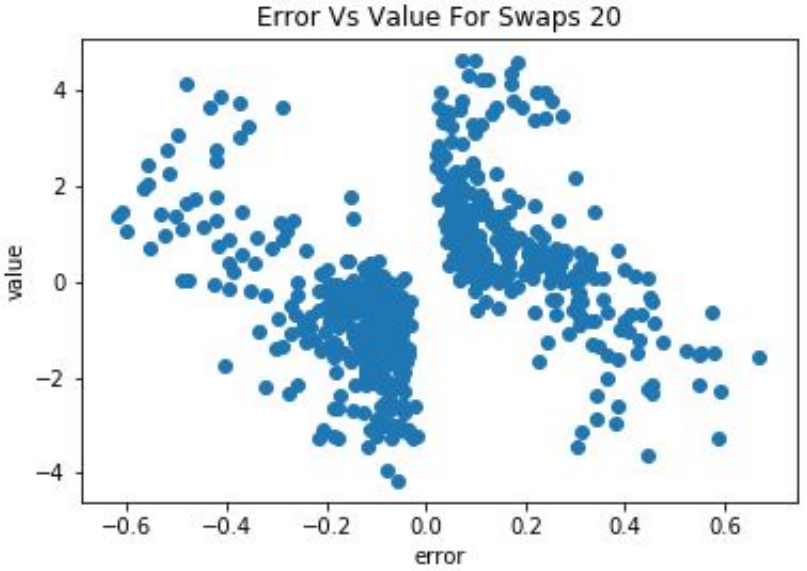




Schrödinger Confidental

20% Flips

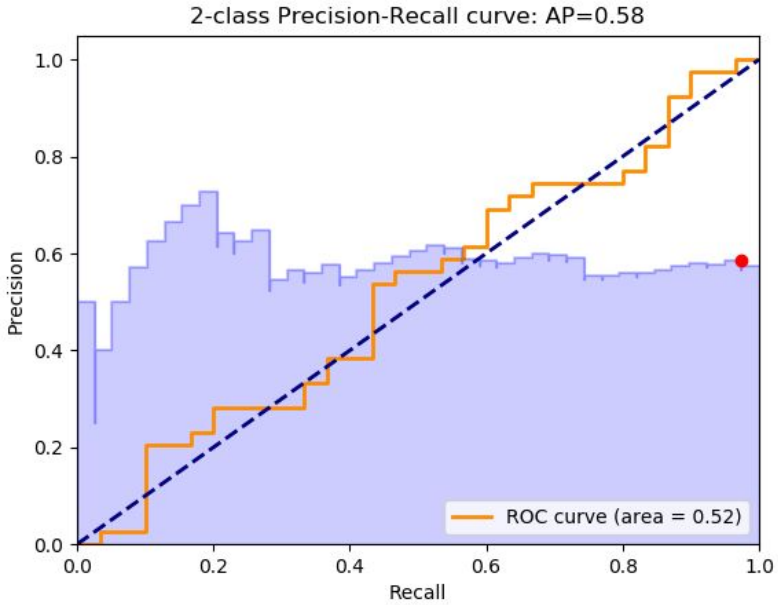


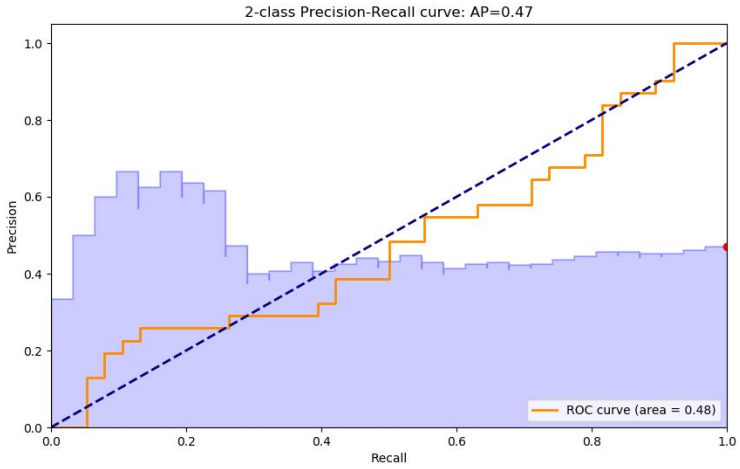


Schrödinger Confidental

50% Flips

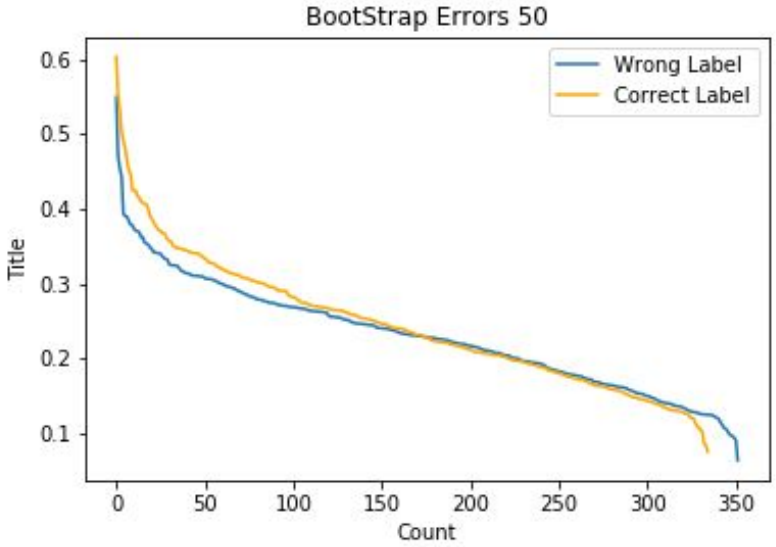
Errored Holdout

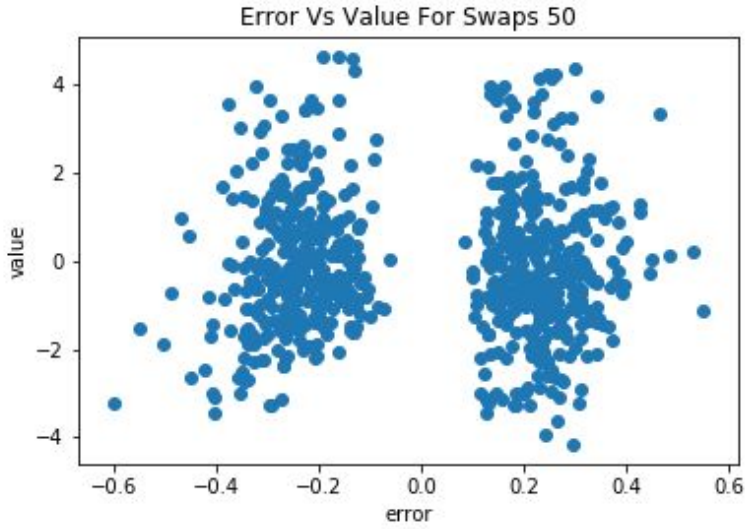




Schrödinger Confidental

50% flips

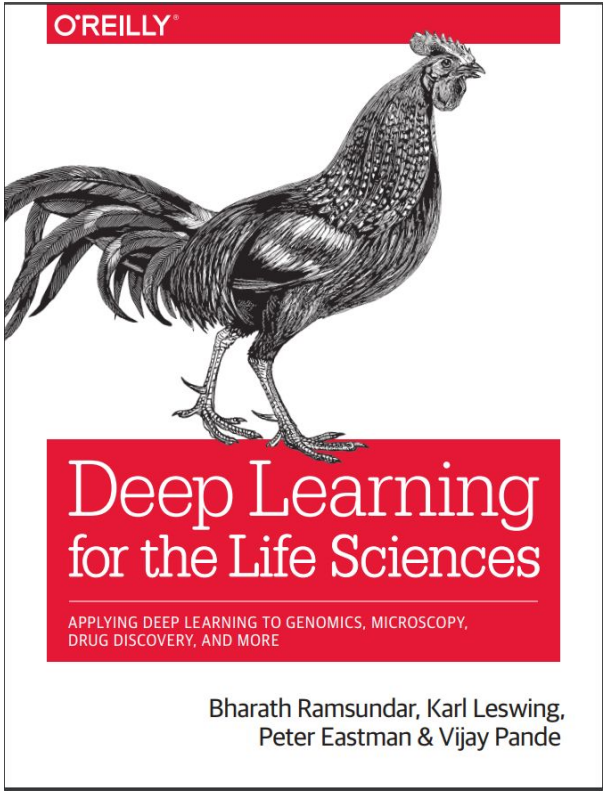




Schrödinger Confidental

Interested in Learning More?

- Book is in pre-release looking for feedback!
- https://www.facebook.com/groups/1362916627160962/
  - Facebook Group
- https://gitter.im/deepchem/Lobby



Schrödinger Confidental