

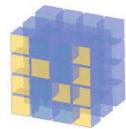
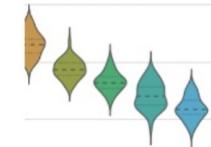


Pandas



matplotlib

Seaborn



NumPy



PYTORCH



SciPy



Cheminformatics With Open-Source Software

Pat Walters

February 24, 2022

Accessing the Interactive Materials For This Talk

1 <https://github.com/PatWalters>

A screenshot of a GitHub user profile for 'PatWalters'. The profile picture is a cartoon illustration of a man with glasses. The name 'Patrick Walters' and handle 'PatWalters' are displayed. Below the profile is an 'Edit profile' button and follower counts ('292 followers - 1 following'). A large circular callout '1' is centered over the profile picture. A white arrow points from this callout to a white circle containing the number '2' on the right side of the pinned repository card for 'EFMC'.

A screenshot of a GitHub repository page for 'EFMC'. At the top, it shows 'main' branch, '1 branch', and '0 tags'. Below is a list of files: LICENSE (Initial commit, 8 days ago), README.md (Update README.md, 8 days ago), example_compounds.sdf (initial commit, 8 days ago), tutorial_01_rdkit.ipynb (updated notebooks for colab, 8 days ago), tutorial_02_classification_model.ipynb (updated notebooks for colab, 8 days ago), and tutorial_03_regression_model.ipynb (notebook updates for colab, 8 days ago). A green 'Code' button is at the top right. A large circular callout '3' is centered over the 'README.md' file content area. A white arrow points from this callout to a white circle containing the number '3' on the right side of the first line of the README.md content.

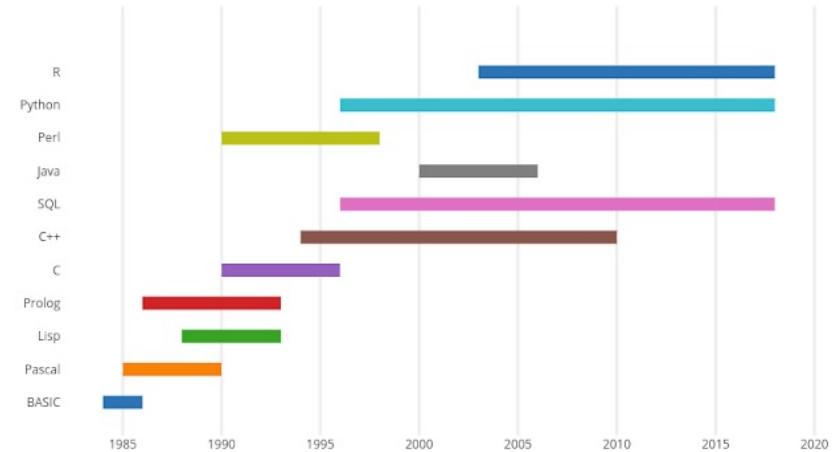
These 3 steps should enable you to launch Google Colab so that you can run the Jupyter notebooks for this talk

About Me



**26 years in drug discovery
35 years writing software
A couple of years of blogging**

**Cheminformatics
Machine learning**



Journal of Molecular Graphics
Volume 11, Issue 2, June 1993, Pages 106-111



Papers

MOUSE: A teachable program for learning in conformational analysis

Daniel P. Dolata², W.Patrick Walters



Practical Cheminformatics

My Goals

Show you what's possible

Show you that's it's not very hard

Demystify all this AI stuff

This is aimed at people with very little programming or AI experience

What We're Going To Cover

A quick introduction to Jupyter notebooks

A classification model example

- Build a decision tree

A regression model

- Predict aqueous solubility

What We're Not Going To Cover

Neural networks

Learned molecular representations

Generative models

Imbalanced data

Hyperparameter tuning

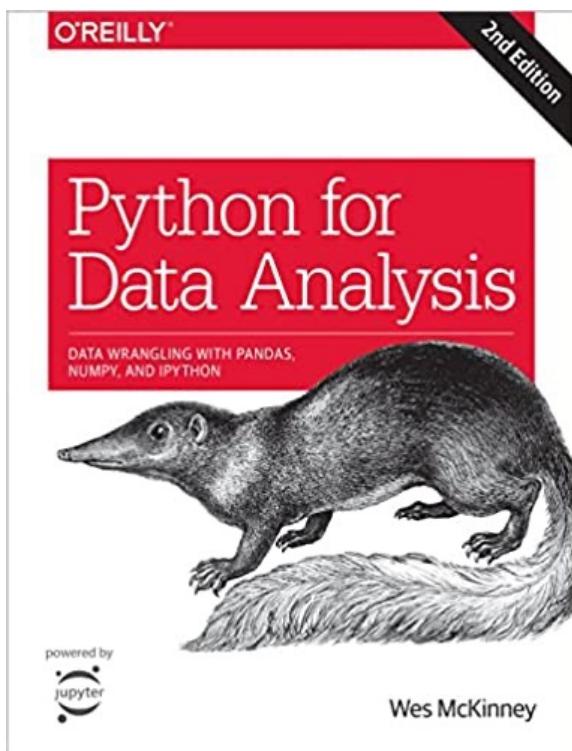
Applicability domain

Model confidence

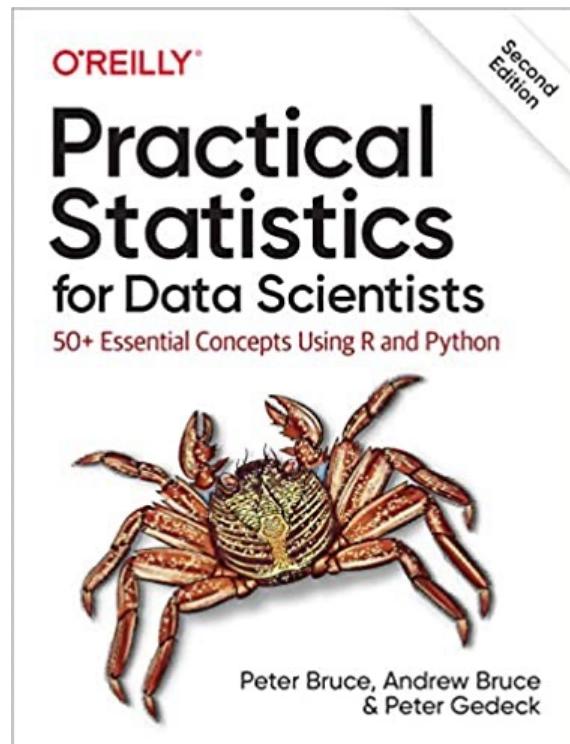
Missing data

A lot theory

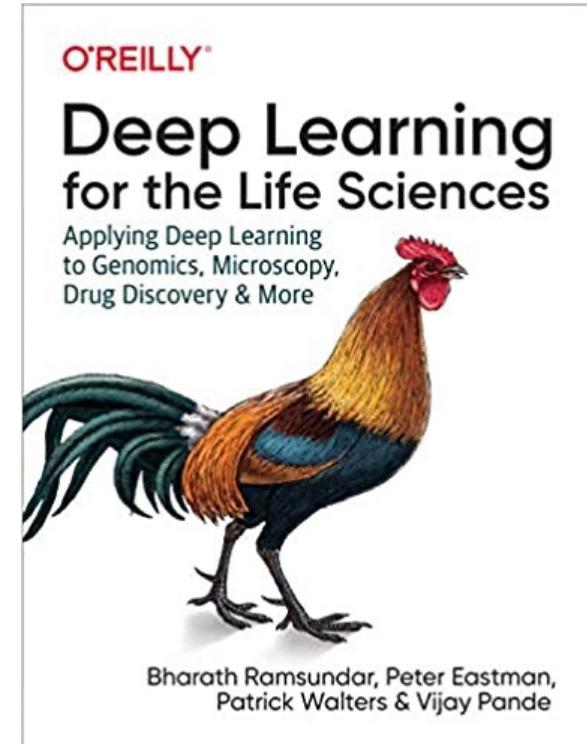
Recommended Reading



<https://www.amazon.com/Python-Data-Analysis-Wrangling-IPython/dp/1491957662/>



<https://www.amazon.com/Practical-Statistics-Data-Scientists-Essential/dp/149207294X/>



<https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837>

The RDKit Cookbook

The RDKit 2021.03.1 documentation » RDKit Cookbook

[previous](#) | [next](#) | [modules](#) | [index](#)

RDKit Cookbook

Introduction

What is this?

This document provides example recipes of how to carry out particular tasks using the RDKit functionality from Python. The contents have been contributed by the RDKit community, tested with the latest RDKit release, and then compiled into this document. The RDKit Cookbook is written in reStructuredText, which supports Sphinx doctests, allowing for easier validation and maintenance of the RDKit Cookbook code examples, where appropriate.

What gets included?

The examples included come from various online sources such as blogs, shared gists, and the RDKit mailing lists. Generally, only minimal editing is added to the example code/notes for formatting consistency and to incorporate the doctests. We have made a conscious effort to appropriately credit the original source and authors. One of the first priorities of this document is to compile useful **short** examples shared on the RDKit mailing lists, as these can be difficult to discover. It will take some time, but we hope to expand this document into 100s of examples. As the document grows, it may make sense to prioritize examples included in the RDKit Cookbook based on community demand.

Feedback and Contributing

If you have suggestions for how to improve the Cookbook and/or examples you would like included, please contribute directly in the source document (the .rst file). Alternatively, you can also send Cookbook revisions and addition requests to the mailing list: <rdkit-discuss@lists.sourceforge.net> (you will need to subscribe first).

Note

The Index ID# (e.g., RDKitCB_##) is simply a way to track Cookbook entries and image file names. New Cookbook additions are sequentially index numbered, regardless of where they are placed within the document. As such, for reference, the next Cookbook entry is RDKitCB_35.



Open-Source Cheminformatics
and Machine Learning

Table of Contents

RDKit Cookbook

- [Introduction](#)
 - [What is this?](#)
 - [What gets included?](#)
 - [Feedback and Contributing](#)
- [Drawing Molecules \(Jupyter\)](#)
 - [Include an Atom Index](#)
 - [Include a Calculation](#)
 - [Include Stereo Annotations](#)
 - [Black and White Molecules](#)
 - [Highlight a Substructure in a Molecule](#)
 - [Without Implicit Hydrogens](#)
 - [With Abbreviations](#)
- [Bonds and Bonding](#)
 - [Hybridization Type and](#)

<https://www.rdkit.org/docs/Cookbook.html>

Useful Blogs

[Practical Cheminformatics](#) - This is a blog where I post once a month or so. These posts typically contain code that demonstrates various aspects of cheminformatics; clustering, machine learning, data visualization, etc. I occasionally throw in posts containing opinions on things like AI and getting a job.

[Is Life Worth Living](#) - A great blog from Iwatobipen (aka pen), whose posts are chock full of great code examples. Pen always seems to be up on the latest methods and posts interesting examples on a variety of topics ranging from quantum chemistry to machine learning.

[The RDKit Blog](#) - Greg Landrum is the primary contributor to, and BDFL, of the RDKit. In addition to the latest and greatest features in the RDKit, Greg's posts also touch on a number of key issues in Cheminformatics, such as dealing with unbalanced datasets and the impact of fingerprint folding on similarity searching.

[Reverie Labs Engineering Blog](#) - The gang at Reverie Labs is doing some of the best work on applying machine learning in drug discovery. Their blog provides useful insights for those applying machine learning at scale.

[The OpenBench Blog](#) - A number of thoughtful posts on ADME modeling and the ways that we validate models.

[Cheminformania](#) - A set of very practical posts by Esben Jannik Bjerrum and friends that primarily focus on the applications of deep learning in drug discovery. These posts provide several useful code examples.

YouTube

https://www.youtube.com/results?search_query=data+professor

Latest from Data Professor

The screenshot shows the YouTube search results for the query "data professor". It features two video thumbnails. The top video is titled "DataPrep Python library for Easy Data Preparation and EDA" by Data Professor, with 1.1K views and uploaded 16 hours ago. The bottom video is titled "Data Slicing in Python with Mito" by Data Professor, with 1K views and uploaded 4 days ago. Both videos are marked as "New". A "More" button is visible at the bottom right of the search results.

DataPrep Python library for Easy Data Preparation and EDA
1.1K views • 16 hours ago
Data Professor
In this video, I provide a quick overview on how you can use the DataPrep Python library to easily and quickly perform data ...
New

Data Slicing in Python with Mito
1K views • 4 days ago
Data Professor
In this video, Jake (co-Founder of Mito) will give us a step-by-step guide on using Mito to slice and dice data in Python. Particularly ...
New

+8 MORE

AI and “The Rise of the Machines”

↻ Andrew Chen Retweeted

 **Mat Velloso** @matvelloso · Nov 22

Difference between machine learning and **AI**:

If it is written in Python, it's probably machine learning

If it is written in **PowerPoint**, it's probably **AI**

166 6.6K 19K  

Show this thread

What is Machine Learning?

Machine learning is all about labeling
things using examples.

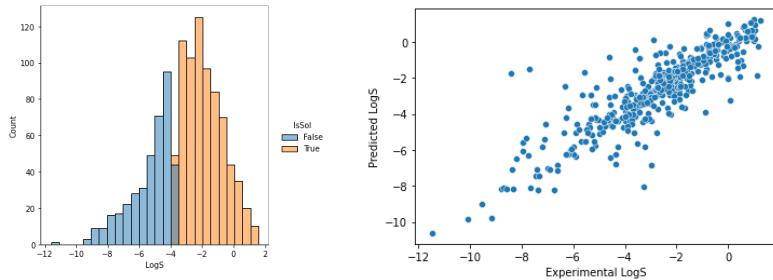
Cassie Kozyrkov, Google



Supervised vs Unsupervised Machine Learning

$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_n \rightarrow Y$

Supervised ML



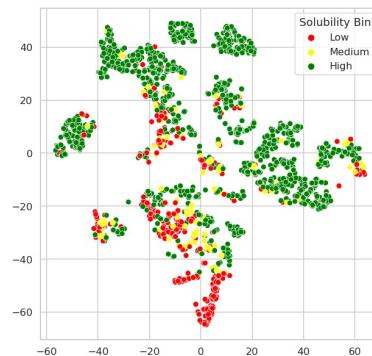
Classification

Regression

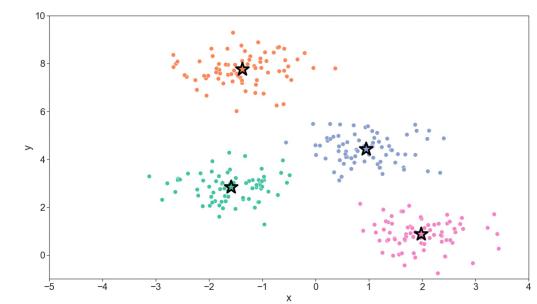
Predict

$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_n$

Unsupervised ML



Dimensionality Reduction

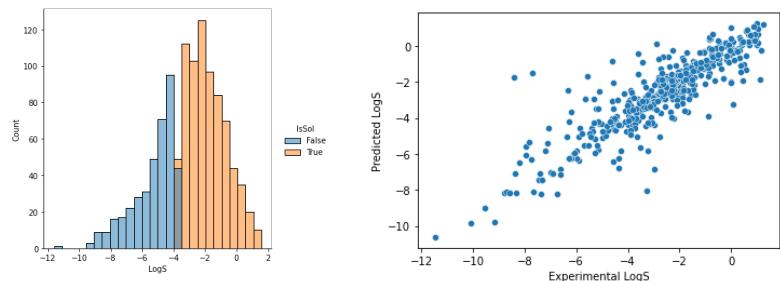


Clustering

Group and Understand

Supervised vs Unsupervised Machine Learning

Supervised ML

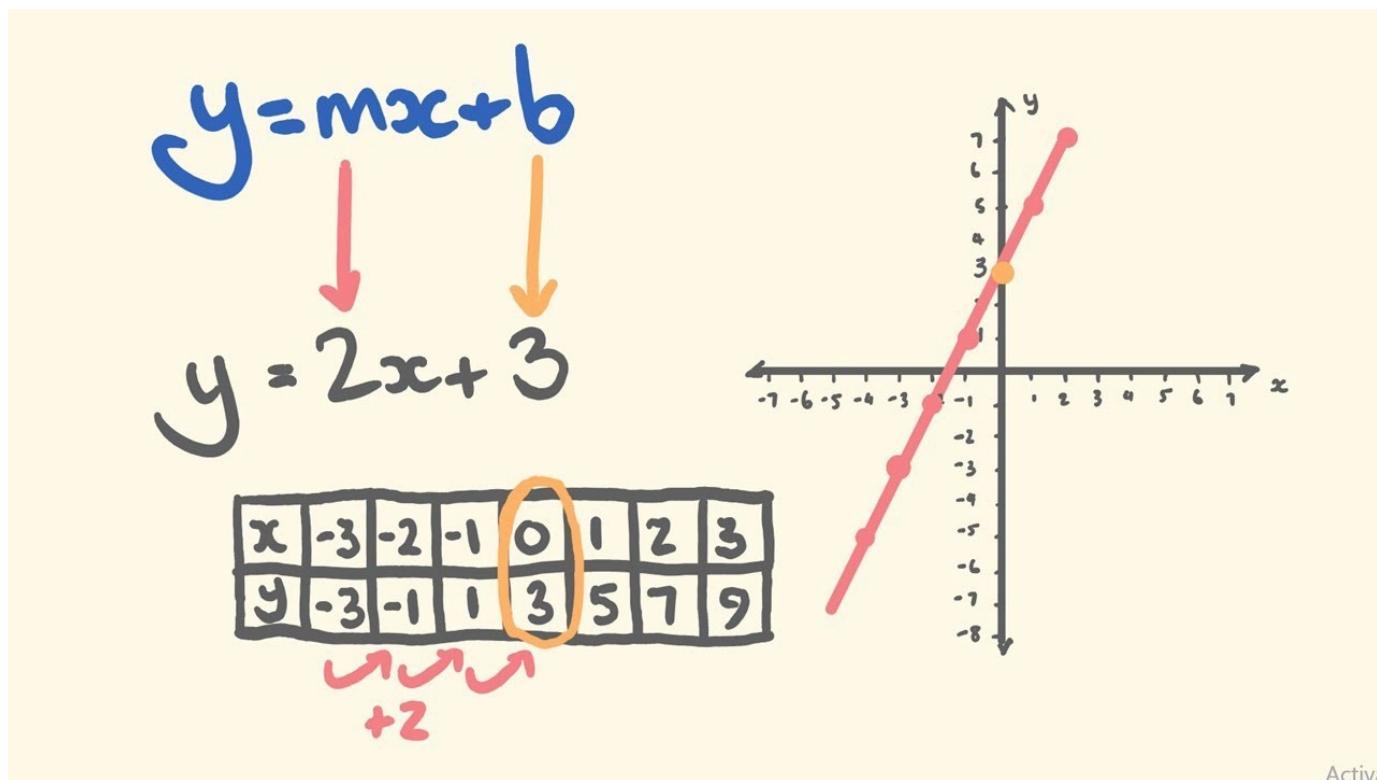


Classification

Regression

Predict

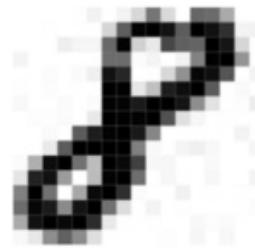
What Are the Observables (X)? What is Being Predicted (Y)?



Activat

Much of Computation is Based on Vector Representations

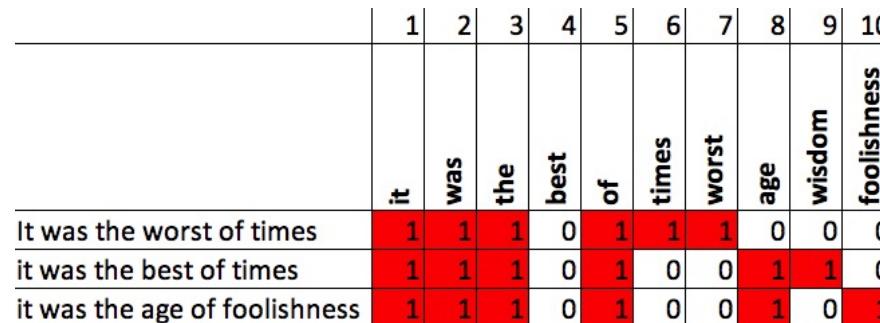
images



1

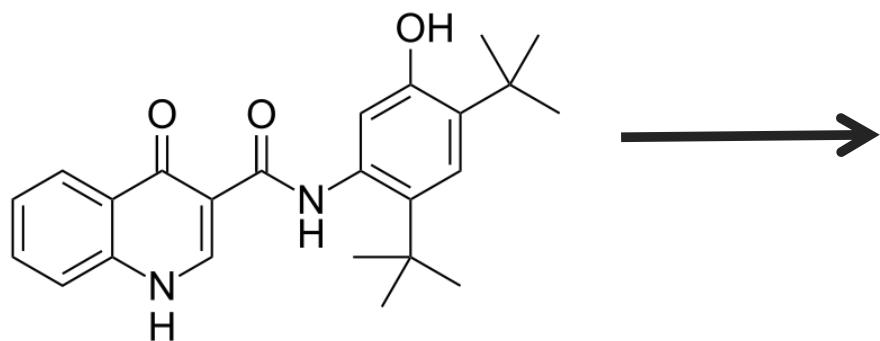
Text

Word	Index
It	1
was	2
the	3
best	4
of	5
times	6
worst	7
age	8
wisdom	9
foolishness	10



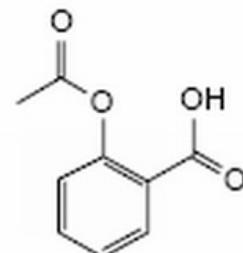
Can We Create a Similar Vector Representation of a Molecule?

Encode molecular features in computer
readable format

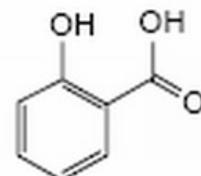


1	0	0	1	1	1	0	0	1	0	0	1	1	1	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Chemical Fingerprints Represent Molecular Features



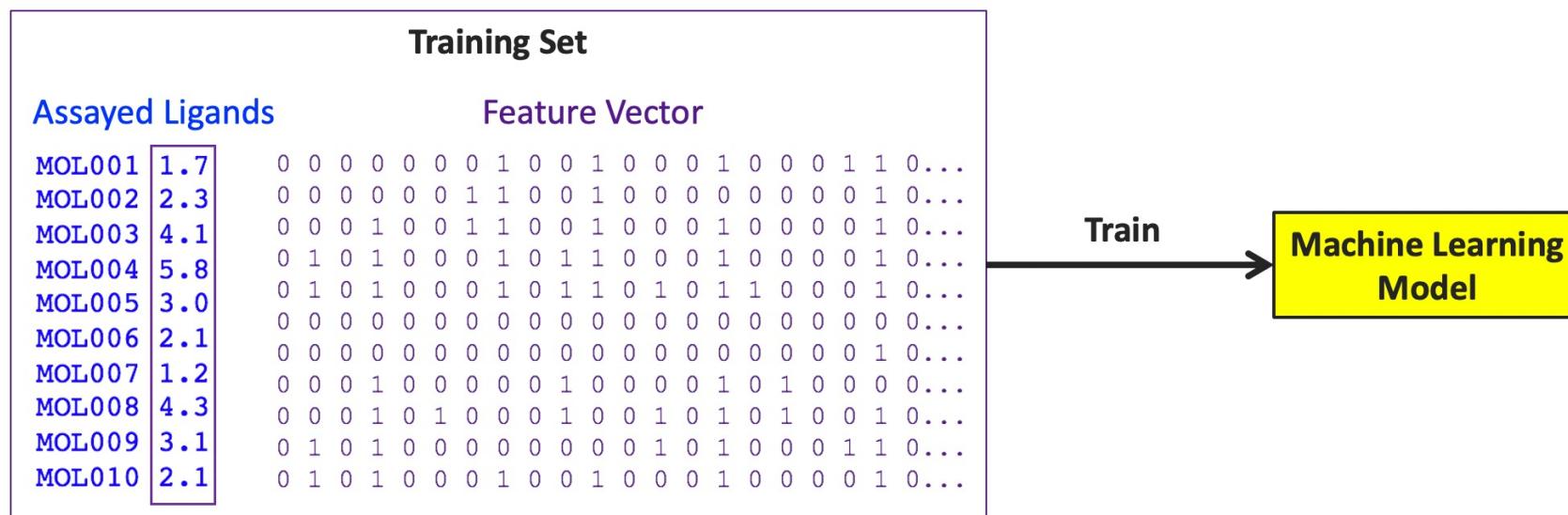
1



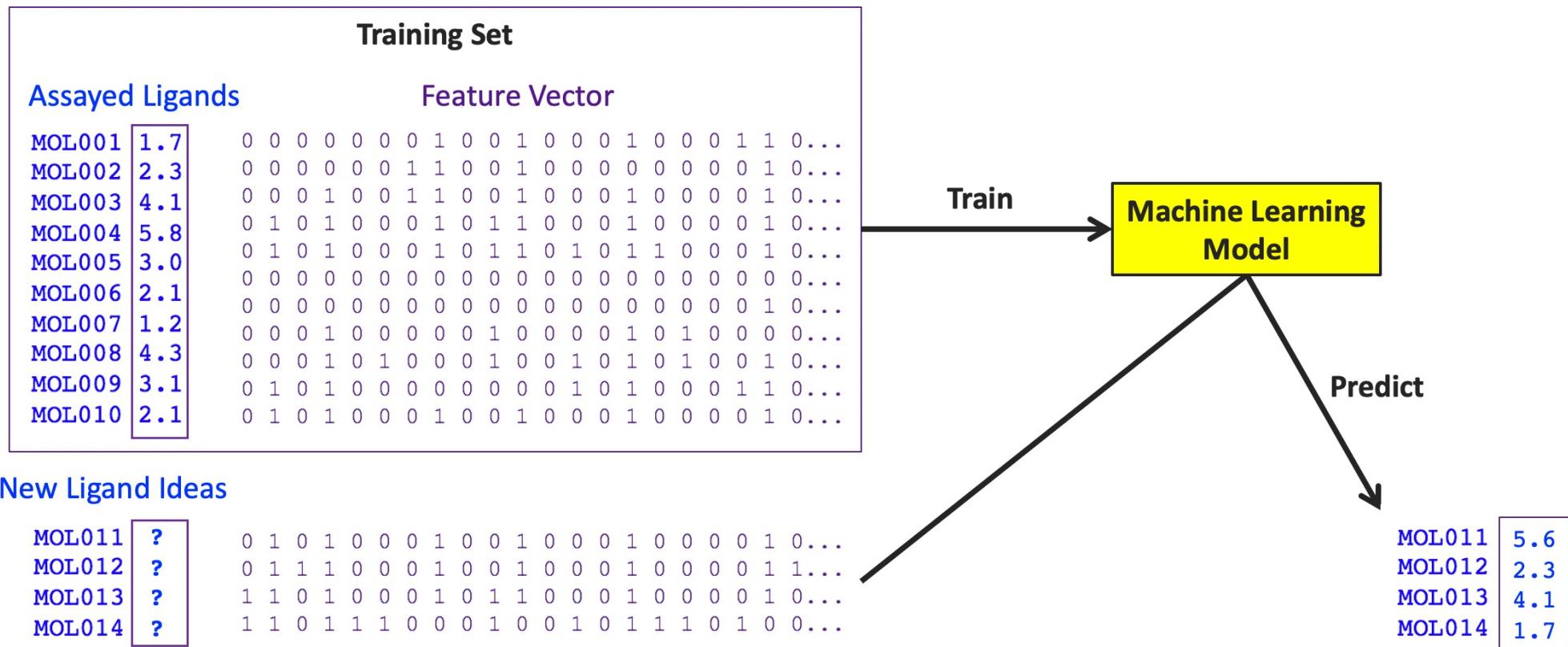
2

1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0

Training a Machine Learning Model



Making Predictions With a Machine Learning Model



Important Jupyter Notebook Keyboard Shortcuts

Shift-Enter	Execute the current cell
Esc-a	Insert cell above
Esc-b	Insert cell below
Esc-x	Delete cell(s)
Esc-m	Convert cell type to Markdown
Esc-y	Convert cell type to Code
!	Execute a shell command, e.g. “!ls”

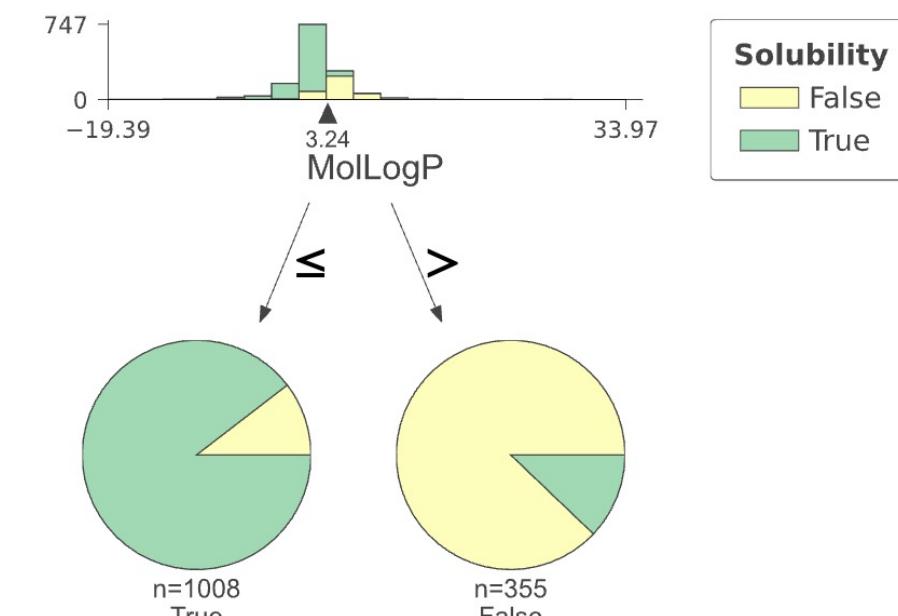
Hands-on Part 1



The Decision Tree – A Simple ML Classifier

Molecular Descriptors

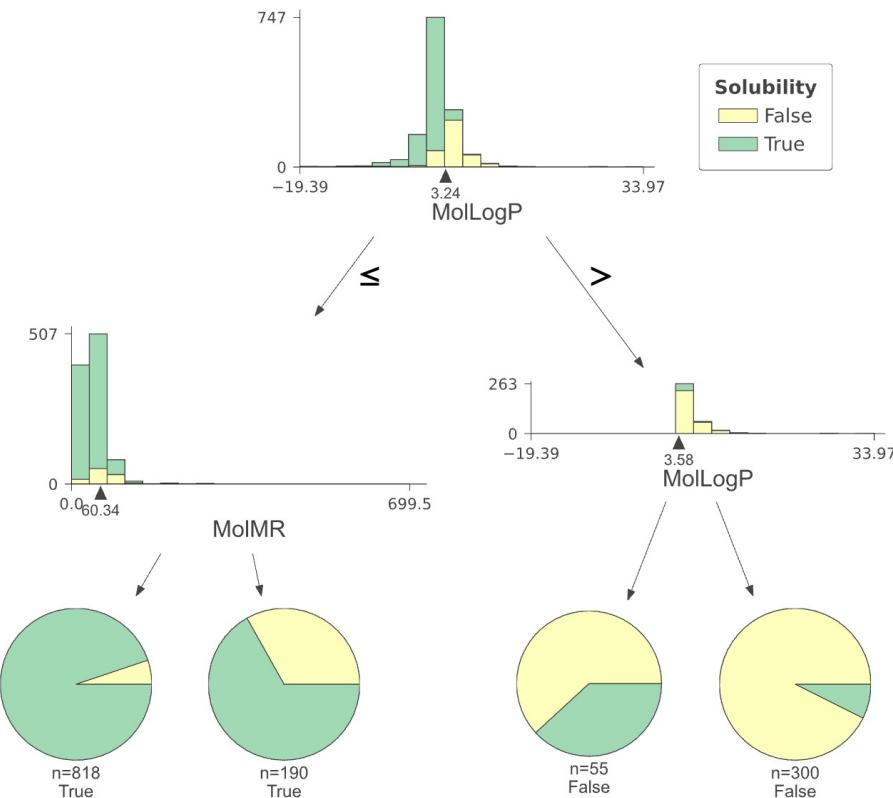
MolWt
MolLogP
MolMR
HeavyAtomCount
NumHAcceptors
NumHDonors
NumHeteroatoms
NumRotatableBonds
NumValenceElectrons
NumAromaticRings
NumSaturatedRings
NumAliphaticRings
RingCount
TPSA
LabuteASA



Increase the Decision Tree Depth to 2

Molecular Descriptors

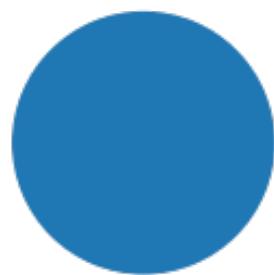
- MolWt
- MolLogP
- MolMR
- HeavyAtomCount
- NumHAcceptors
- NumHDonors
- NumHeteroatoms
- NumRotatableBonds
- NumValenceElectrons
- NumAromaticRings
- NumSaturatedRings
- NumAliphaticRings
- RingCount
- TPSA
- LabuteASA



Select Decision Tree Splits Using Gini Impurity

$$G = 1 - \sum_{k=1}^n p_k^2$$

Gini impurity



0.00



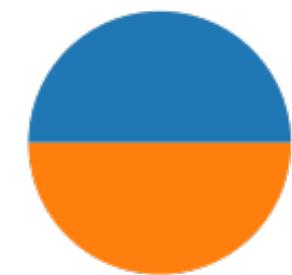
0.18



0.24



0.40



0.50

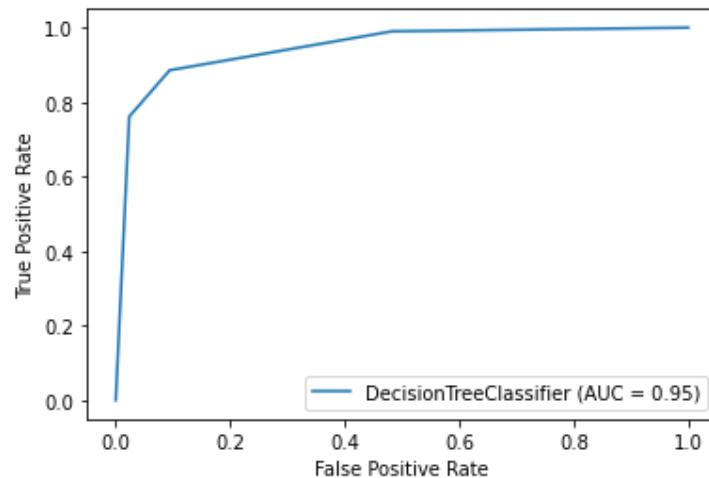
Evaluating Classification Models

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Confusion Matrix

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$



ROC AUC Curve

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Matthews Correlation Coefficient

$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)}$$

Cohen's Kappa

Building A Classification Model With SciKit Learn

```
train, test = train_test_split(df)
train_X = train[descriptors]
train_y = train.experiment
test_X = test[descriptors]
test_y = test.experiment
```

Construct training and test sets

```
my_model = RandomForestRegressor()
my_model.fit(train_X, train_y)
pred = my_model.predict(test_X)
```

Build a predictive model

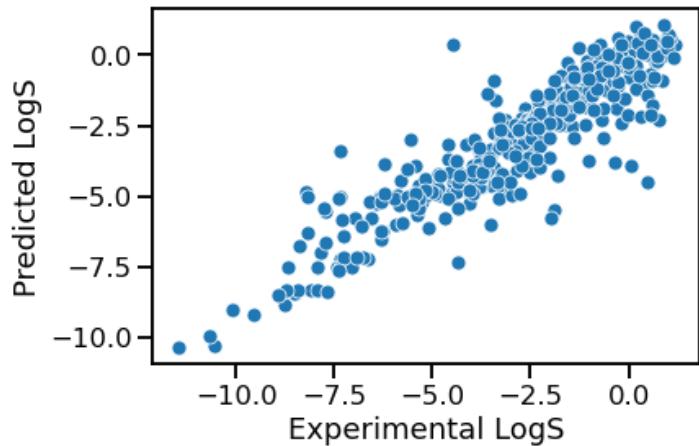
```
auc = roc_auc_score(test_y,pred)
mcc = matthews_corrcoef(test_y,pred)
kappa = cohen_kappa_score(test_y,pred)
```

Evaluate the predictive model

Hands-on Part 2



Evaluating Regression Models



$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Coefficient of Determination

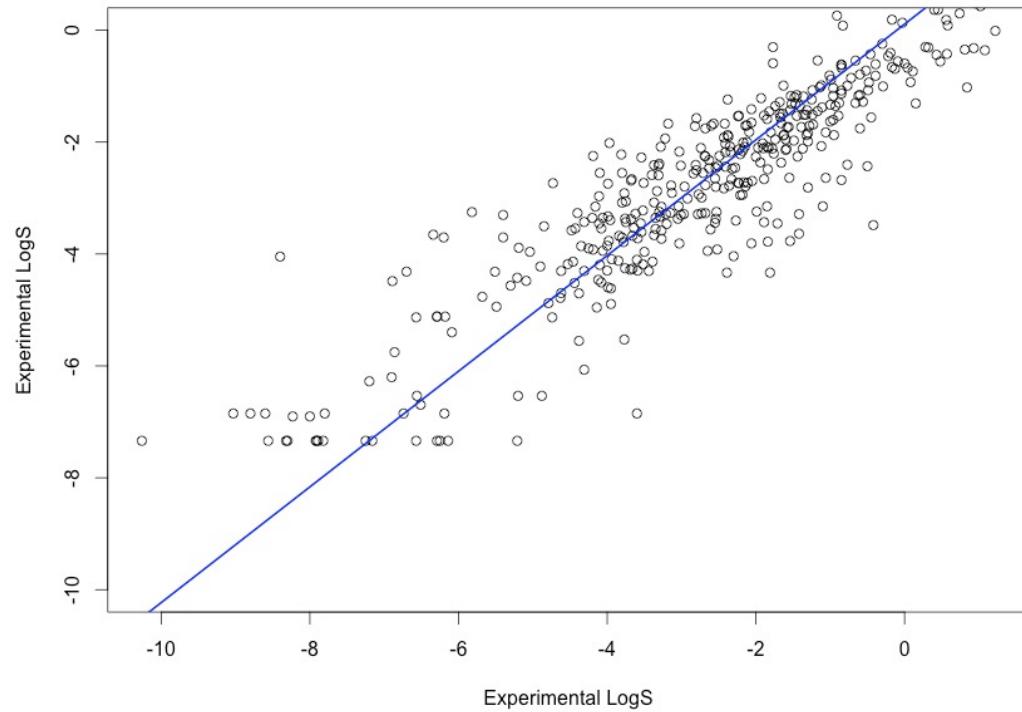
$$\text{RMSD} = \sqrt{\frac{\sum_{t=1}^T (\hat{y}_t - y_t)^2}{T}}.$$

Root Mean Squared Deviation (Error)

Evaluating Correlation

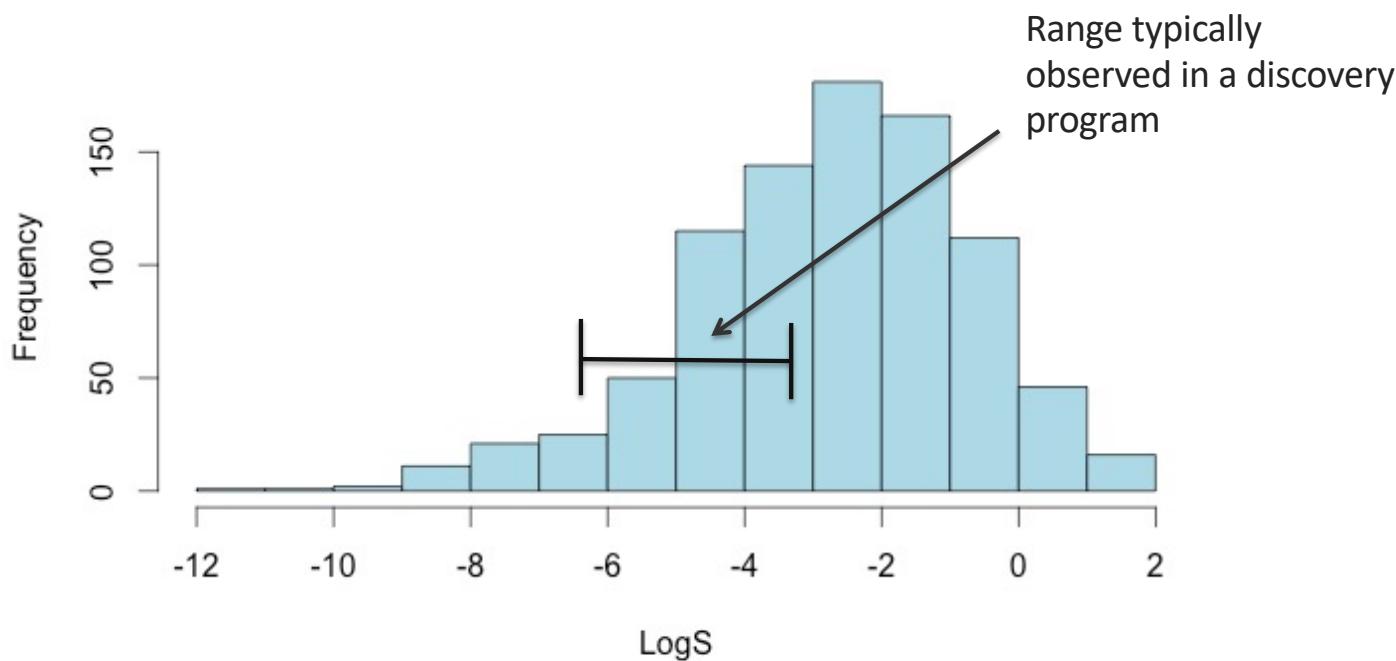
$R^2=0.8$

Median Absolute Error =0.54



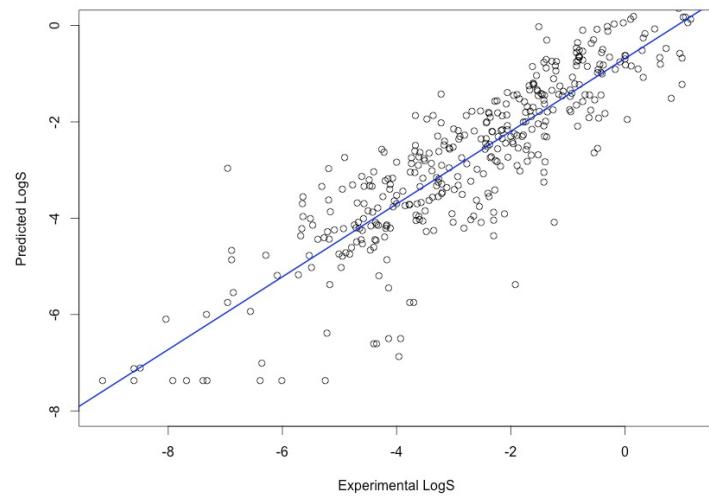
Let's Understand the Data

- This data spans a very wide range
- Much wider than typically observed in a drug discovery program

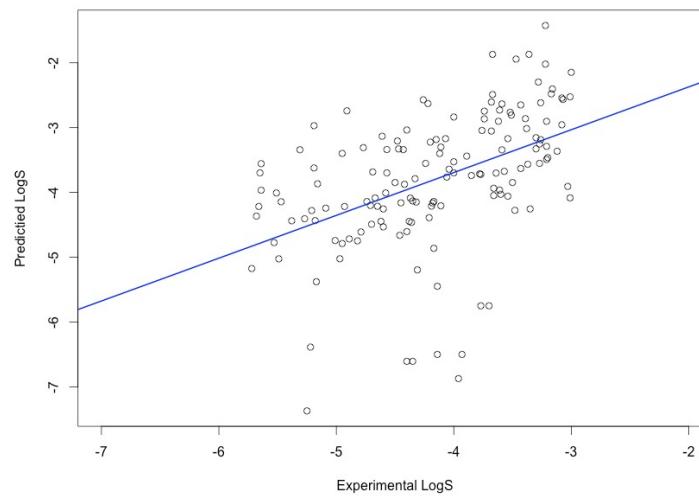


Correlation Changes With Dynamic Range

This is the same dataset. On the left we consider the entire set, which has an unrealistically large (~10 log) dynamic range. On the right we consider a more realistic subset with a 3 log dynamic range. Note the change in correlation.

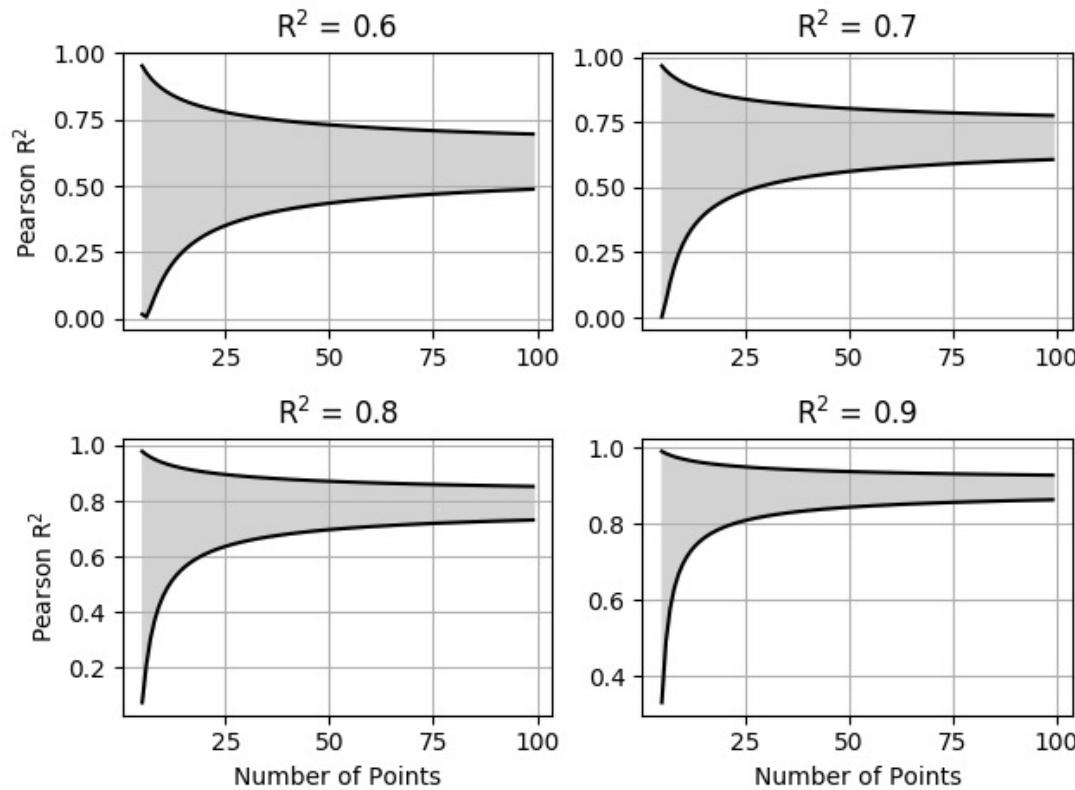


$$R^2 = 0.76$$
$$MAE = 0.55$$



$$R^2 = 0.22$$
$$MAE = 0.69$$

Remember That Correlation Have Confidence Limits



Questions?



<https://practicalcheminformatics.blogspot.com/>



@wpwalters



<https://www.linkedin.com/in/wpwalters/>