https://www.anaconda.com/distribution/

https://github.com/vmmadathil/HackSMU

https://scikit-learn.org/stable/index.html

Let’s Do Data Science

Visakh Madathil

Agenda

- What is Data Science
- Data Science Pipeline and Lifecycle
- Demo and Practice

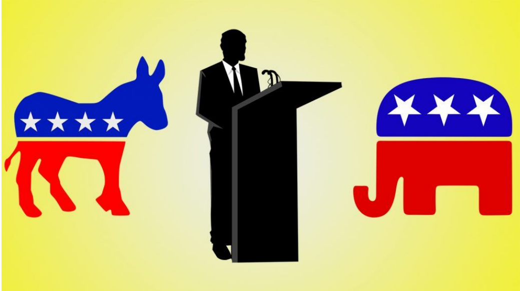
Who Am I?

It’s Visakh (like G-shock)

It’s not:

- V-sock
- V-sack
- Vy-sack

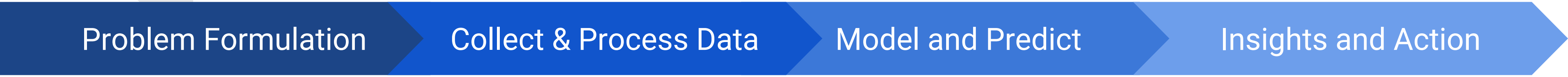
SMU



What is Data Science?

“A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician. … We'll say that a data scientist is someone who extracts insights from messy data.”

- Joel Grus



Problem Formulation

- Identify favorable outcomes
- Identify data sources
- Identify possible biases

Collect & Process Data

- Clean and Pre-Process Data
- Prepare Data for Modeling

Model and Predict

- Use Machine Learning and Statistical Modeling
- Explore causation and correlation

Insights and Action

- Translate results into actions
- Feed results into research/ business pipeline and processes

Modeling Techniques

Statistical (Machine) Learning

Machine learning techniques are very common tools for data analysis.

ML techniques are increasingly important: they are *the way* that many problems are being attacked.

Statistical (Machine) Learning

The basic premise of a [supervised] machine learning problem:

- Given some inputs, we want to predict the (most likely) correct output.
- We have many examples of correct input + output: training data.
- A model is trained with the known data, and then used to predict on new inputs.

Statistical (Machine) Learning

To get anywhere, we need to have a lot of correct input/output pairs.

We will use most of them to train the model. Hopefully it will find whatever relevant structure/patterns are in the data, and make good predictions of the output later.

Statistical (Machine) Learning

But how will we know if good predictions are being made?

Usually, we want to break up the known input/outputs into two sets: training data to train the model and testing data to test how good the predictions are.

```
from sklearn.model_selection import train_test_split
X = known_inputs
y = corresponding_outputs
X_train, X_test, y_train, y_test = train_test_split(X, y)
```

Scikit-learn

Scikit-learn

The scikit-learn module implements many machine learning algorithms and corresponding tools.

It's still important to understand how the models work: you won't know their strengths and weaknesses otherwise. But implementing them can be done by somebody else.

Scikit-learn

The models implemented in scikit-learn all have the same general API. First create the model with whatever parameters it needs:

```
model = SVC(kernel='linear',
C=0.05)
```

Then train it with the training data:

```
model.fit(X_train, y_train)
```

Scikit-learn

Once you have a model, you can check how it does on the testing data:

```
print(model.score(X_test, y_test))

 0.97
```

You can tend manipulate the parameters and model to achieve greater accuracy

Scikit-learn

And finally, make some predictions on new inputs:

```
model.predict(X_new)
```

Let’s Practice