Please Add Your Name and Email to the google sheet:

https://docs.google.com/spreadsheets/d/1KeARd5uDzs5ZBTWhtMEhYm2l4ZzQXMcBpcjRQTbLReI/edit?usp=sharing

(So I can send you the class materials)

QuantSprout

Python and Machine Learning

By Craig Sakuma

Introductions

Craig Sakuma

- Founder of QuantSprout
- Instructor at General Assembly
- MBA from Wharton
- B.Eng from Northwestern University

Fun Fact

Developed a novelty BBQ product that was featured in USA Today

QuantSprout

Class Introductions

- Name
- What’s your job?
- How do you plan to apply skills from the bootcamp?
- Fun Fact

Course Structure

- Lectures on topics
  - Interaction is good
  - Feel free to ask questions
  - If there's not enough time to cover questions, we'll put it in a parking lot for after class
- Hands on exercises
  - Pair programming
  - Mix up partners

Objectives for Class

- Get strong foundation of Python and Machine Learning
- Immediately use skills at work
- Remove barriers/frustration
- Develop skills to be self-sufficient after class
    - Learn and explore new topics
    - Troubleshoot problems

Course Outline

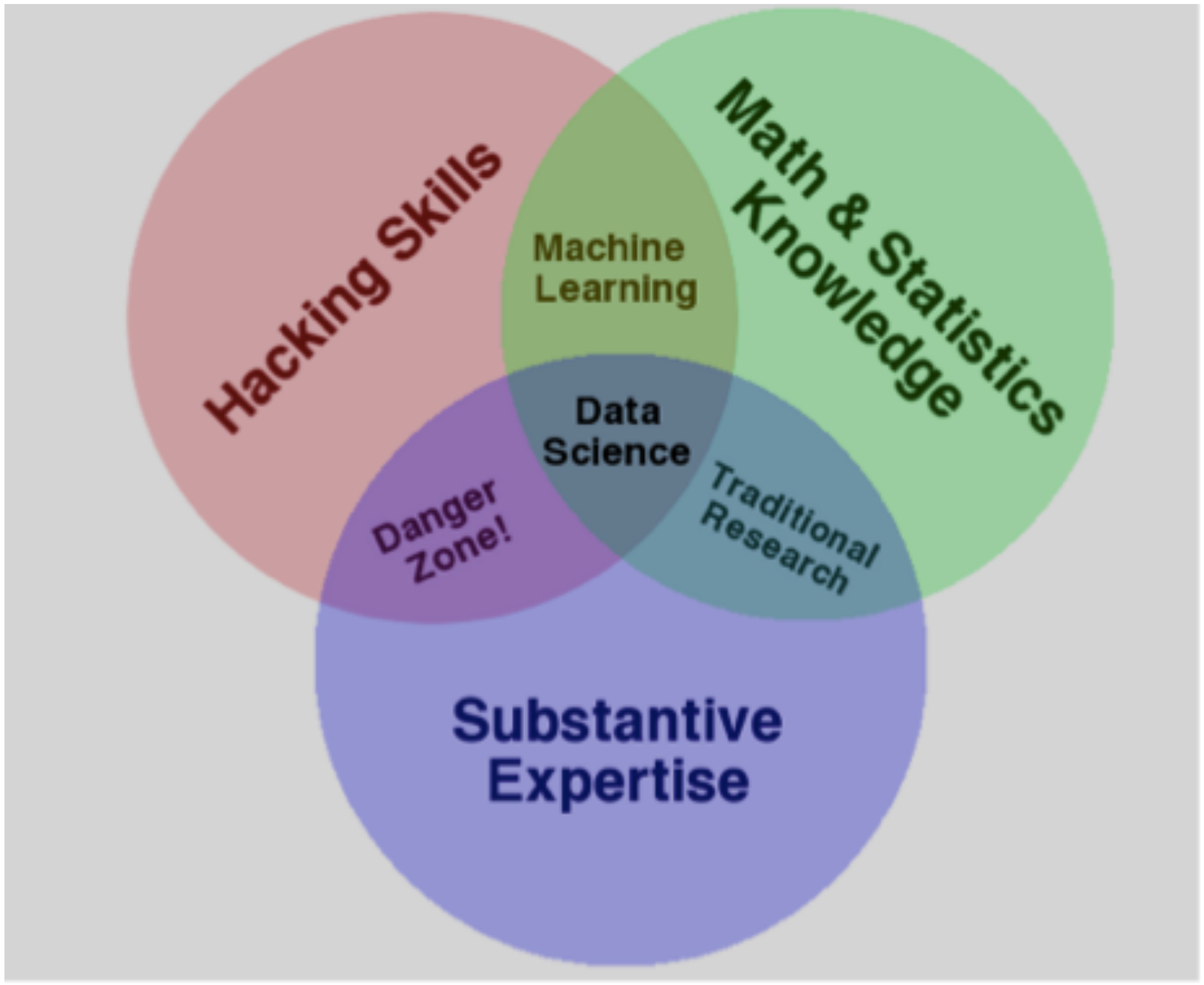
<u>Day 1</u>

- Overview of Data Science
- Intro to Pandas / Exploring Data
- Cleaning Data
- Classification Algorithms

<u>Day 2</u>

- Cross-validation
- Regression Algorithms
- Regularization

What is Data Science?



Data Science is OSEMN (Awesome)

<u>O</u>btain Data

<u>S</u>crub Data

<u>E</u>xplore

<u>M</u>odel Algorithms

i**<u>N</u>**terpret Results

80% (Obtain Data, Scrub Data, Explore)

20% (Model Algorithms, iNterpret Results)

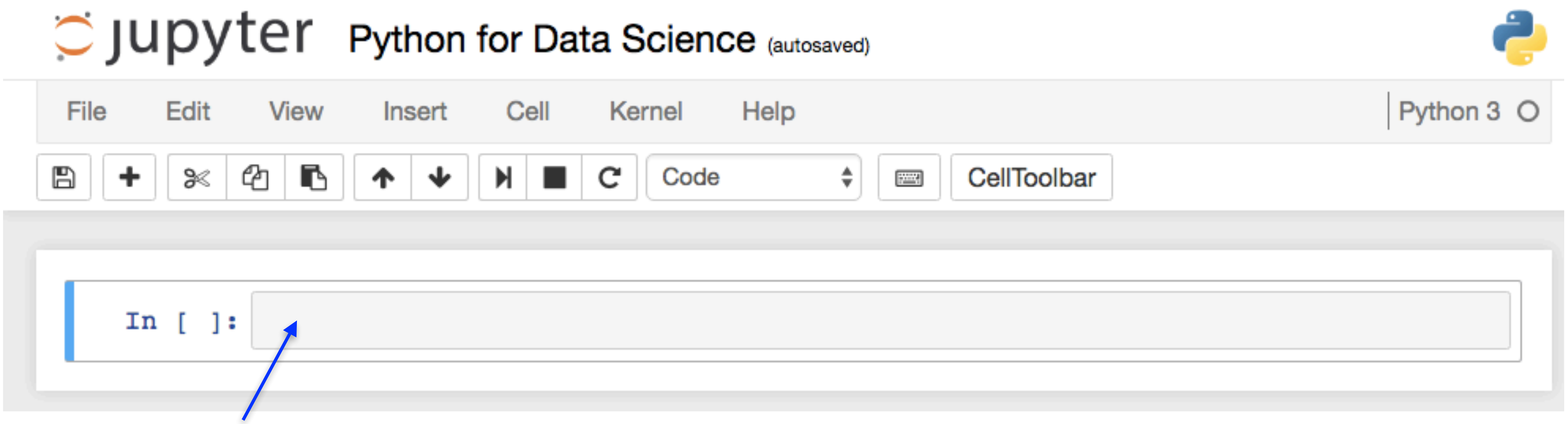
Majority of time is spent data munging

Why Python?

- Readability
- Flexibility
- Supports multiple programming paradigms
    - Procedural
    - Functional
    - Object oriented

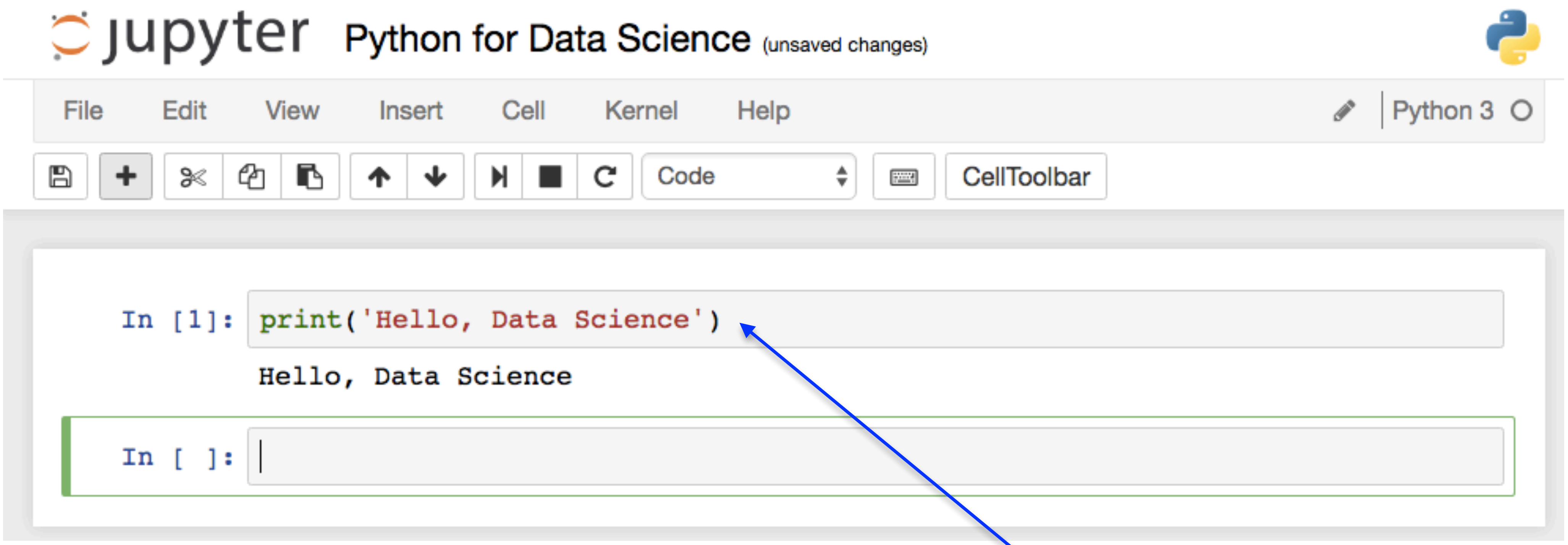
Libraries of Tools for Data Analysis

Jupyter Notebook Basics



Enter code here

Jupyter Notebook Basics

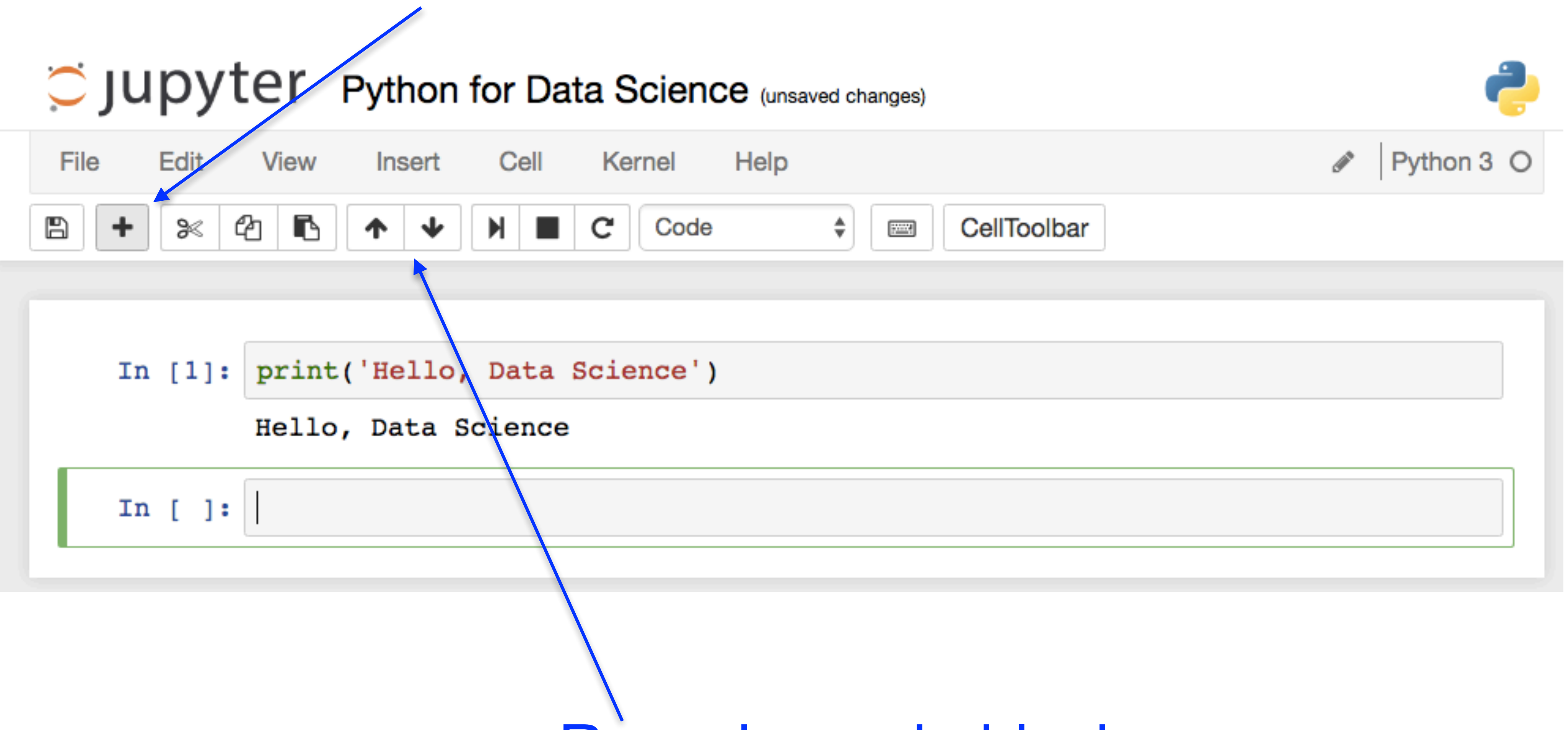


```
In [1]: print('Hello, Data Science')

Hello, Data Science

In [ ]:
```

Shift + Enter runs code and returns results

Jupyter Notebook Basics

Add more code blocks

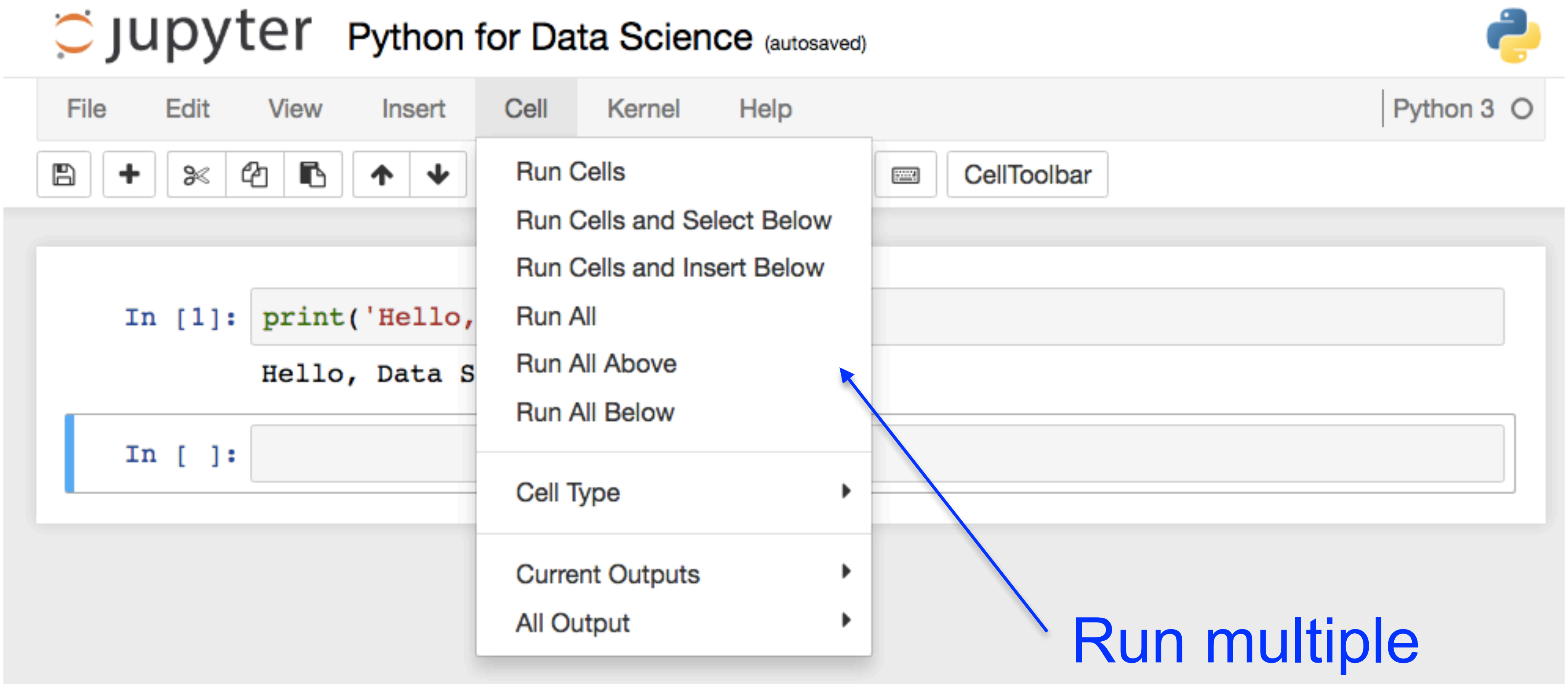


```
In [1]: print('Hello, Data Science')
        Hello, Data Science

In [ ]: 
```

Re-order code blocks

QuantSprout

Jupyter Notebook Basics

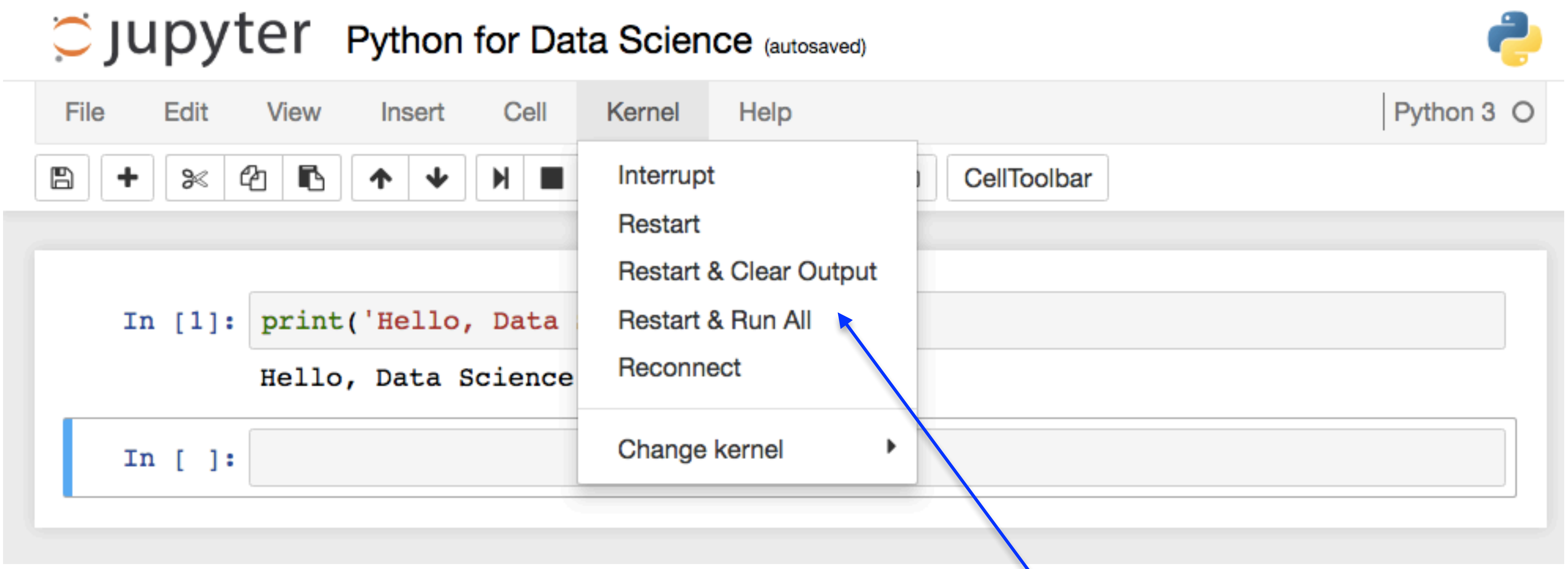


```
In [1]: print('Hello,
        Hello, Data S
In [ ]:
```

Run multiple code blocks

QuantSprout

Jupyter Notebook Basics

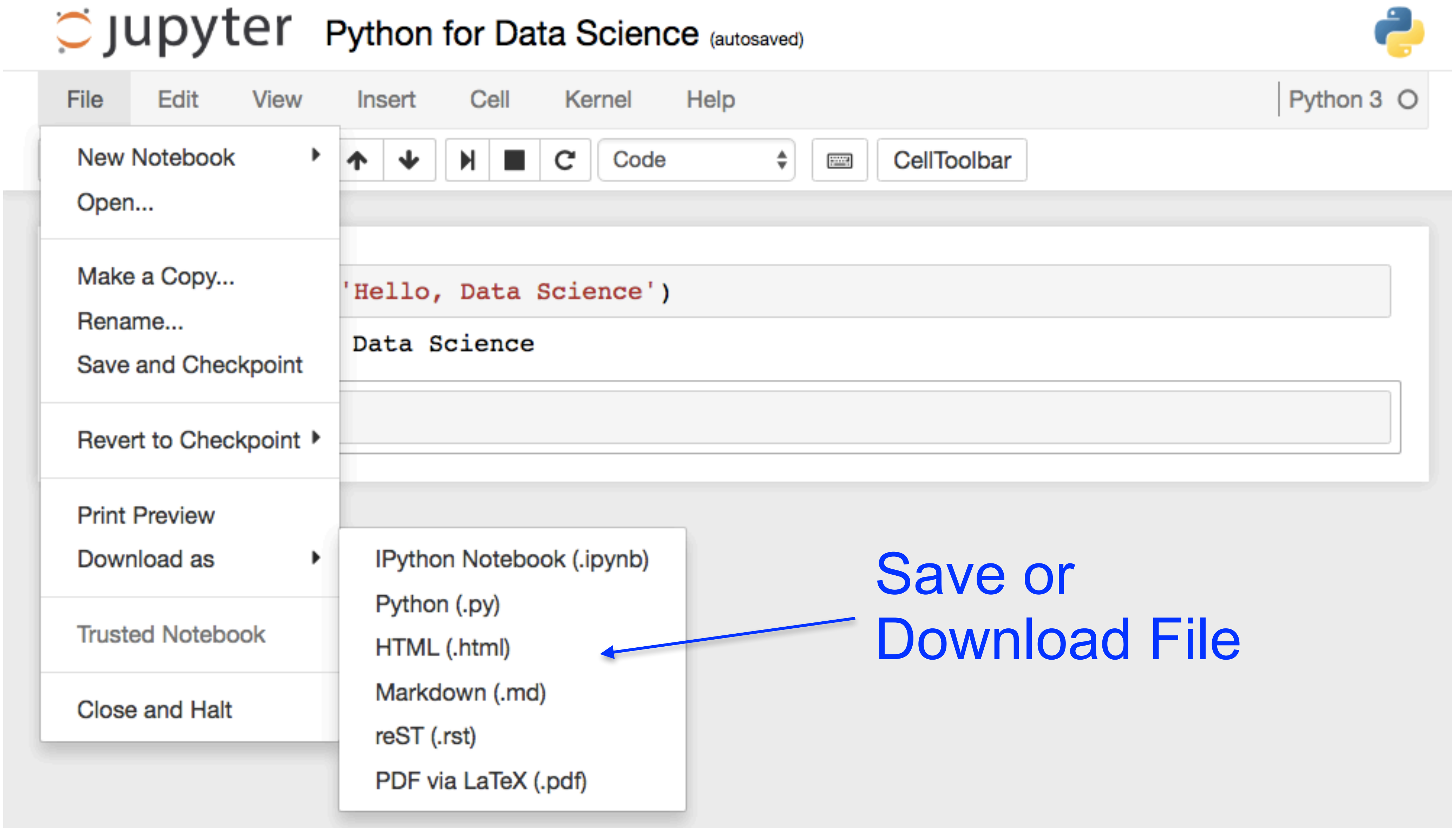


Python for Data Science (autosaved)

```
In [1]: print('Hello, Data
Hello, Data Science
In [ ]:
```

Restart Notebook to Clear Memory

Jupyter Notebook Basics

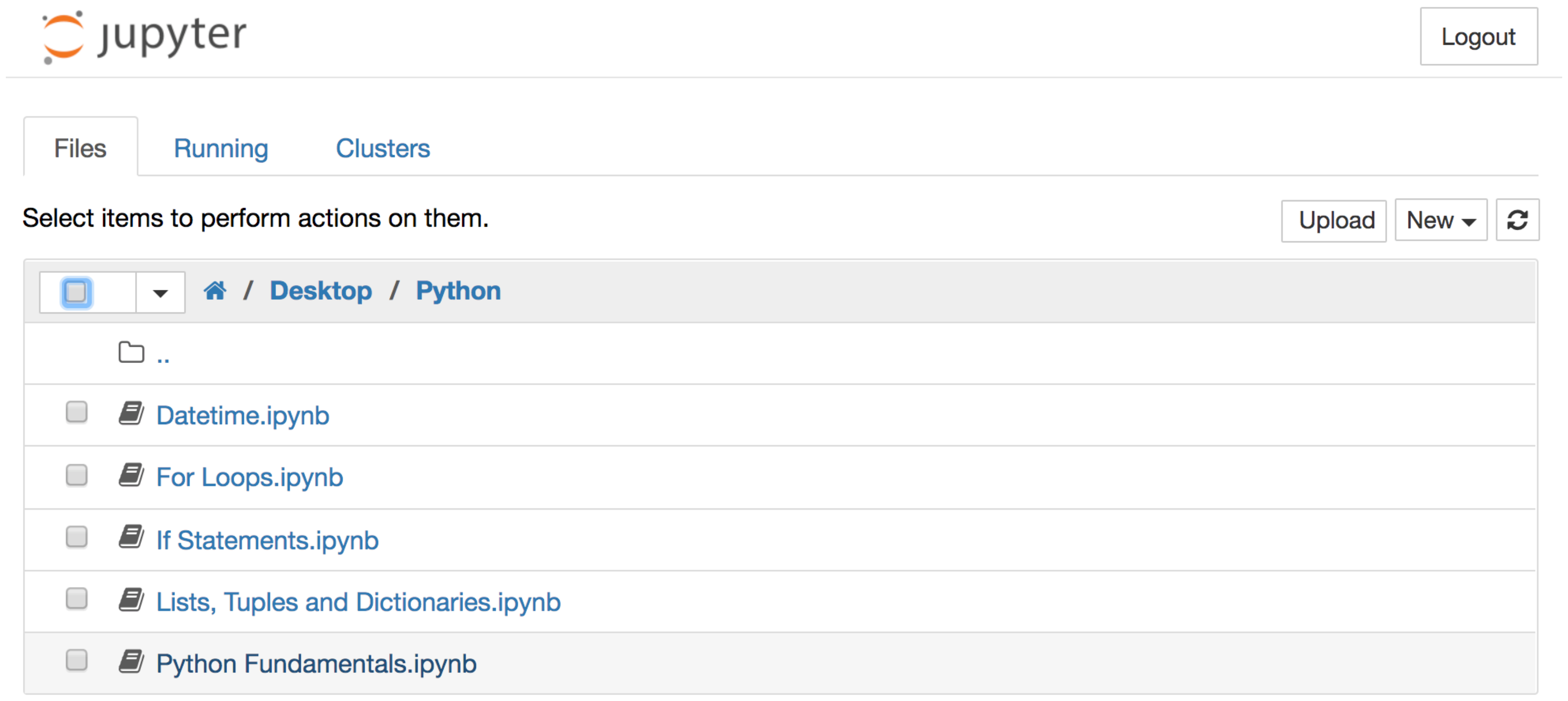


```
'Hello, Data Science')
Data Science
```

Save or
Download File

Set Up for Jupyter Notebooks

- Create a folder on your Desktop Called "python_ml"
- Copy the files from Slack into the folder
- Launch jupyter notebook



Select items to perform actions on them.

/ Desktop / Python

..

Datetime.ipynb

For Loops.ipynb

If Statements.ipynb

Lists, Tuples and Dictionaries.ipynb

Python Fundamentals.ipynb

Intro to Pandas

- Primary objects in Pandas are DataFrames
- DataFrames are like tables
  - Contain rows and columns of data
  - Columns have names
  - Rows have index values
- Pandas has easy functions for importing and exporting data
  - CSV files
  - Excel spreadsheets
  - SQL queries

Data Cleaning

- Missing Values
  - Identify Missing Values
  - Drop Values
  - Impute Values
    - Zero
    - Mean
- Categorical Data
  - Convert Text to Numbers
  - Encode Labels as Boolean Variables

What is Machine Learning?

“A field of study that gives computers the ability to learn without being explicitly programmed.” (1959)

- Arthur Samuel, AI Pioneer

Supervised vs. Unsupervised

Supervised

- Requires truth set of data for training algorithms
- Examples:
  - Forecasting sales
  - Classifying spam

Unsupervised

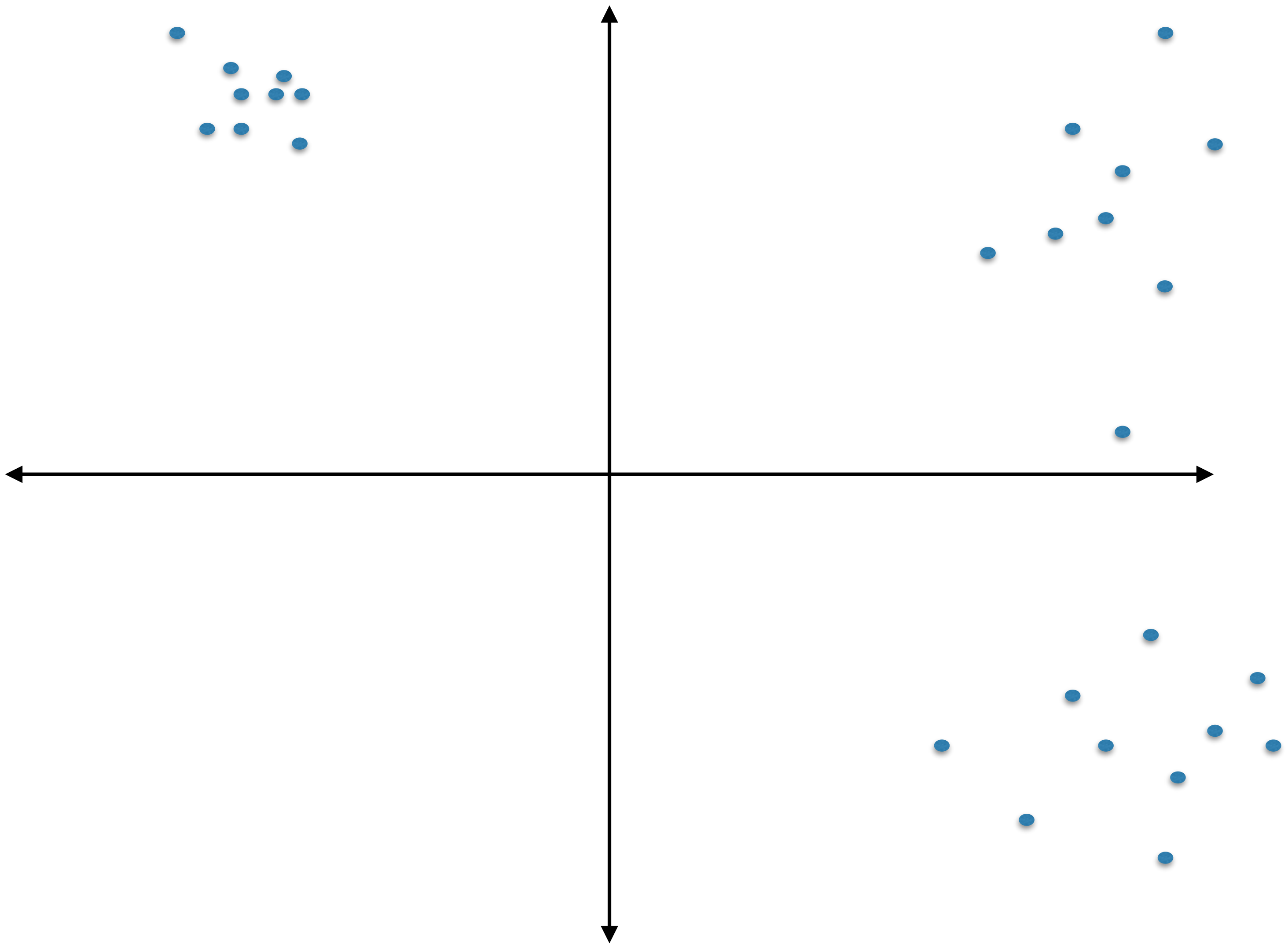
- Autonomous algorithm that requires no training
- Examples:
  - Cluster analysis
  - Anomaly detection

Machine Learning Categories

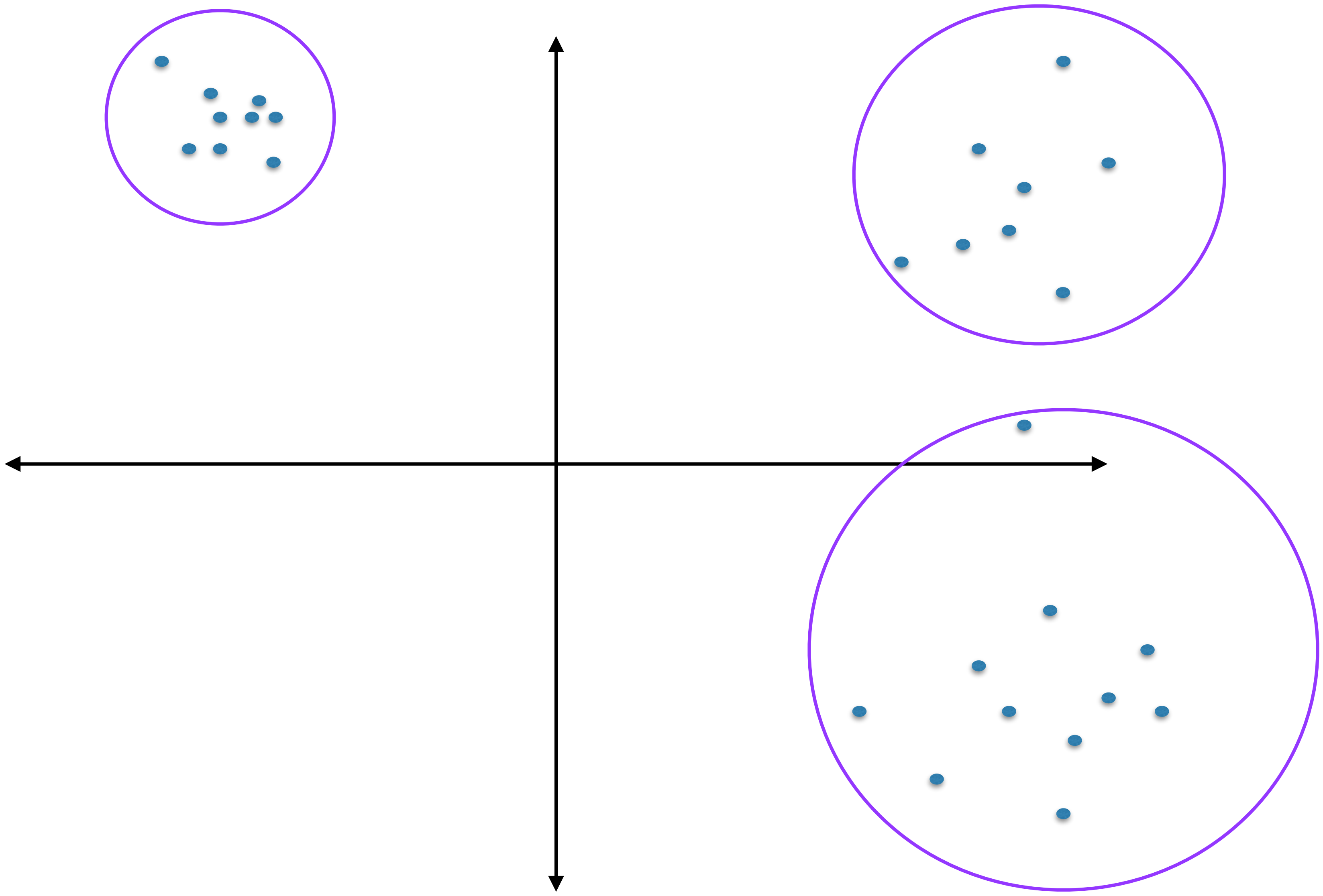
	Continuous	Categorical
Supervised	Regression	Classification
Unsupervised	Dimension Reduction	Clustering

QuantSprout

Clustering

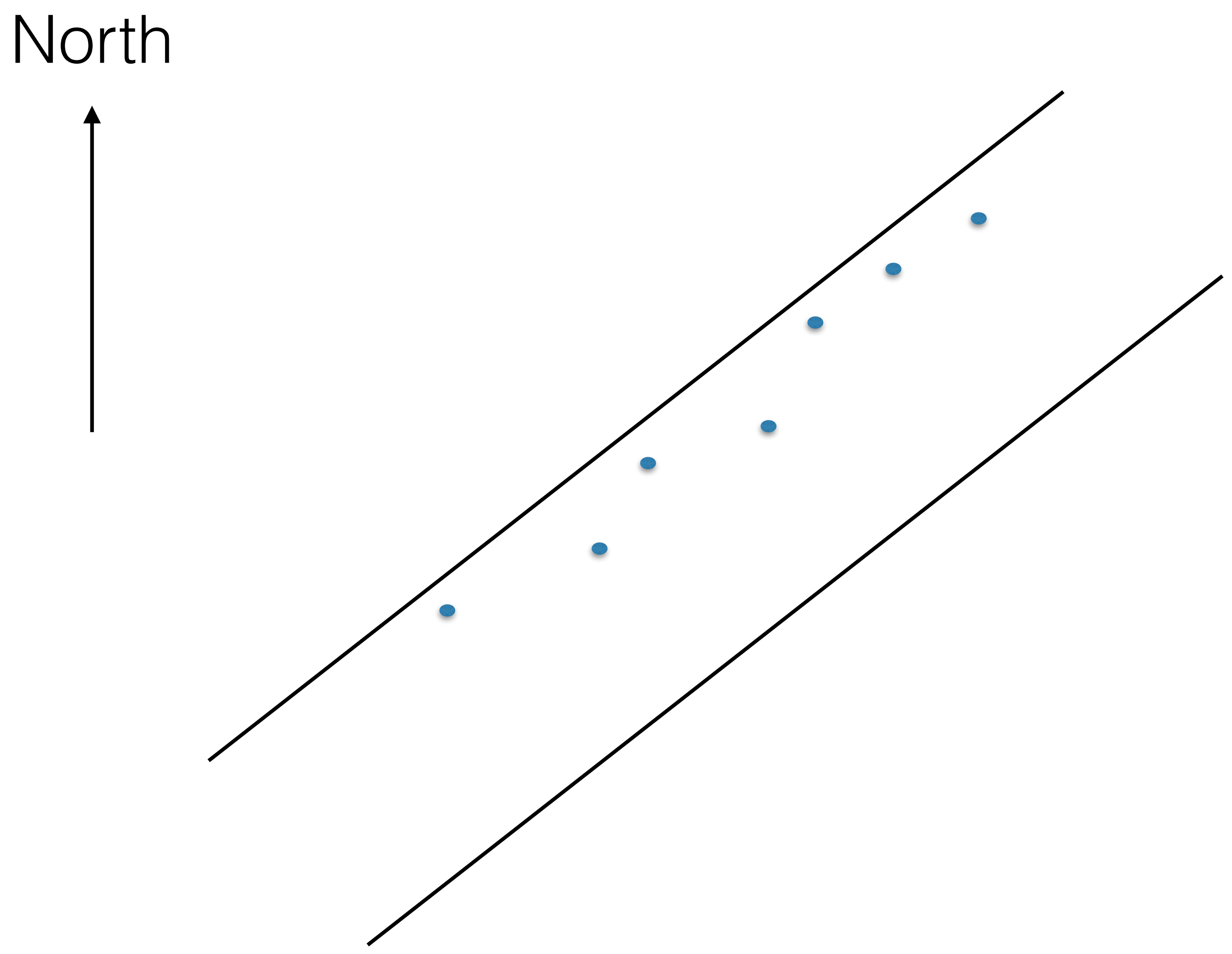


Clustering

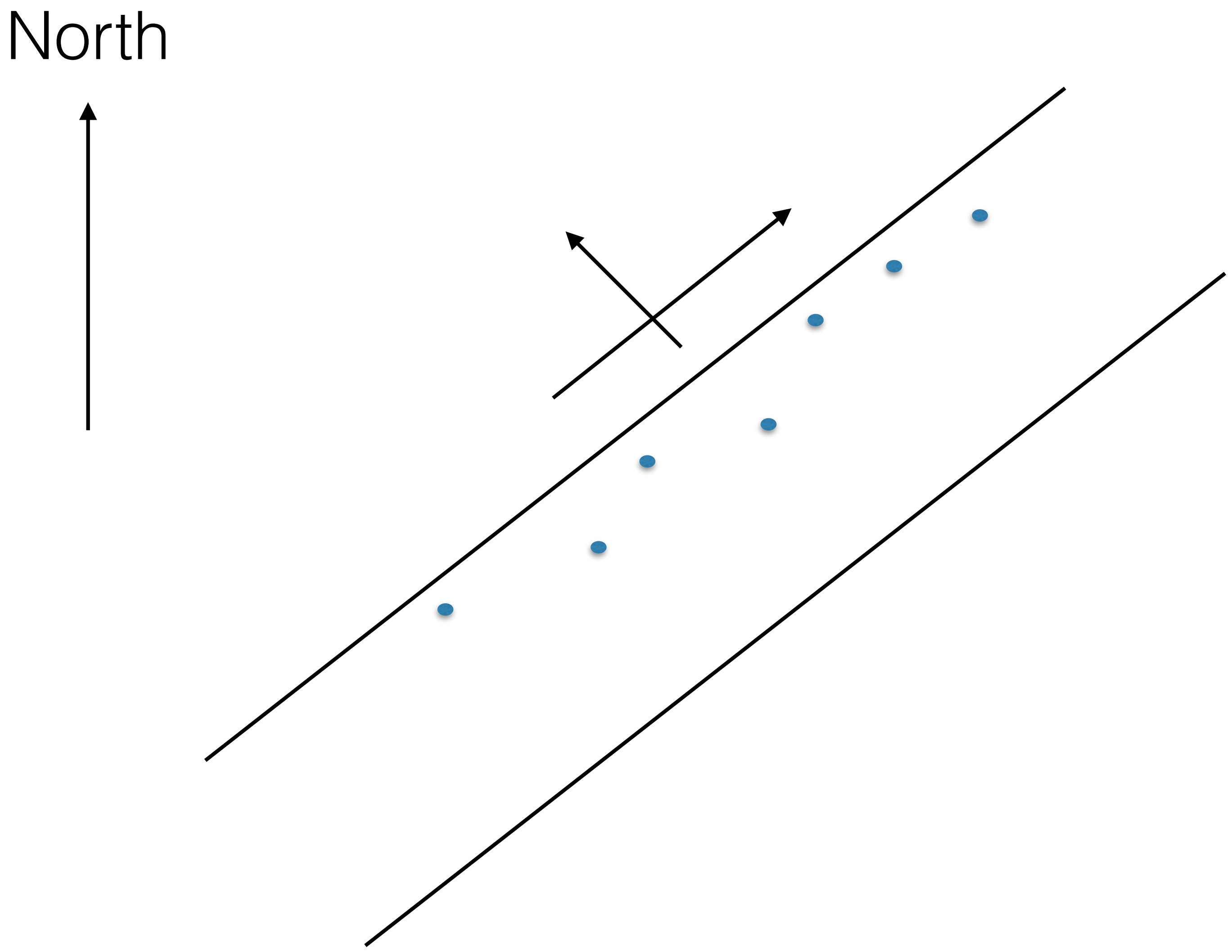


QuantSprout

Dimensionality Reduction

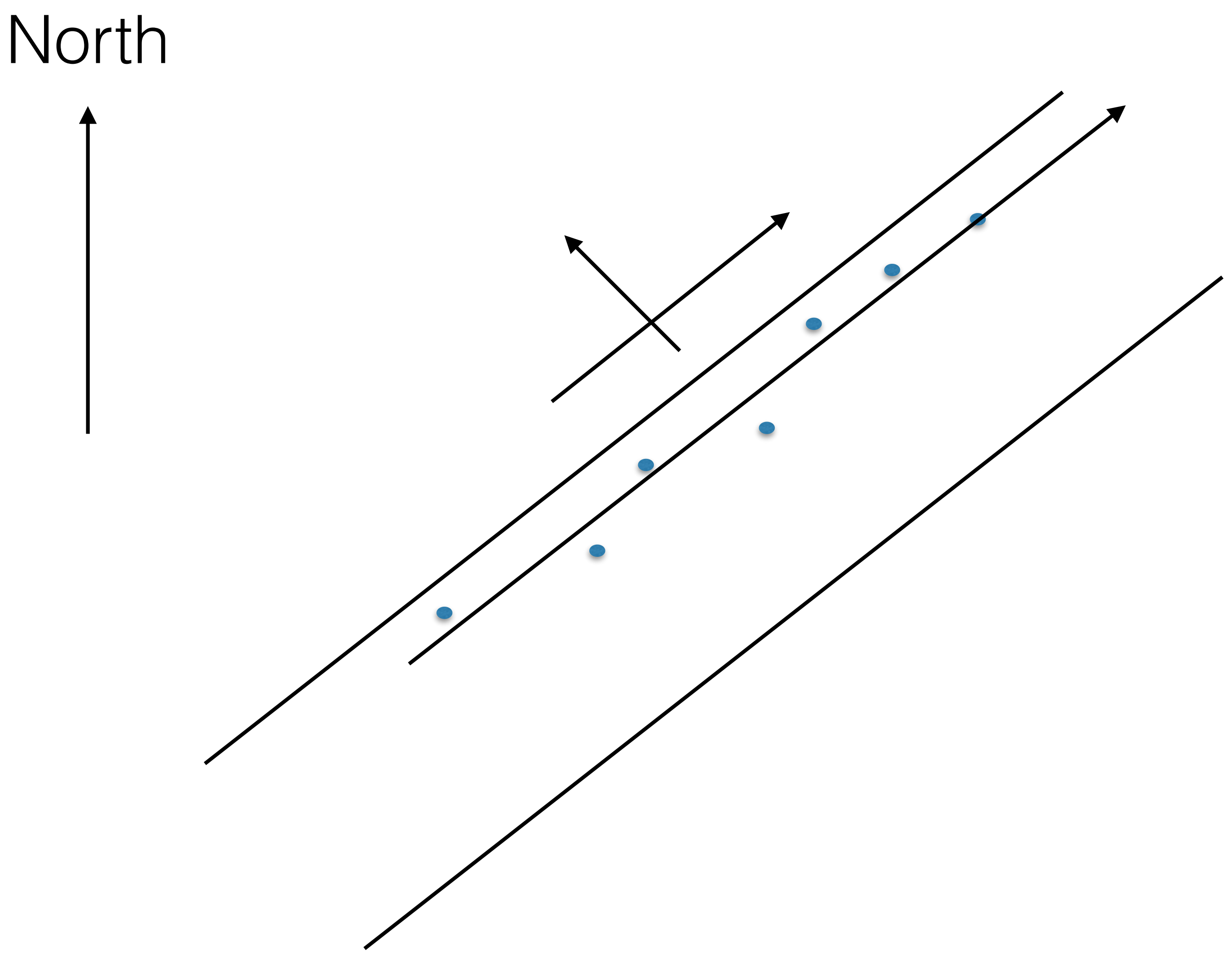


Dimensionality Reduction



QuantSprout

Dimensionality Reduction



Supervised Training

- Training Set
  - Data used to create coefficients for model
  - For example, data set used to create a linear regression
- Test Set
  - Data used to measure performance of trained model

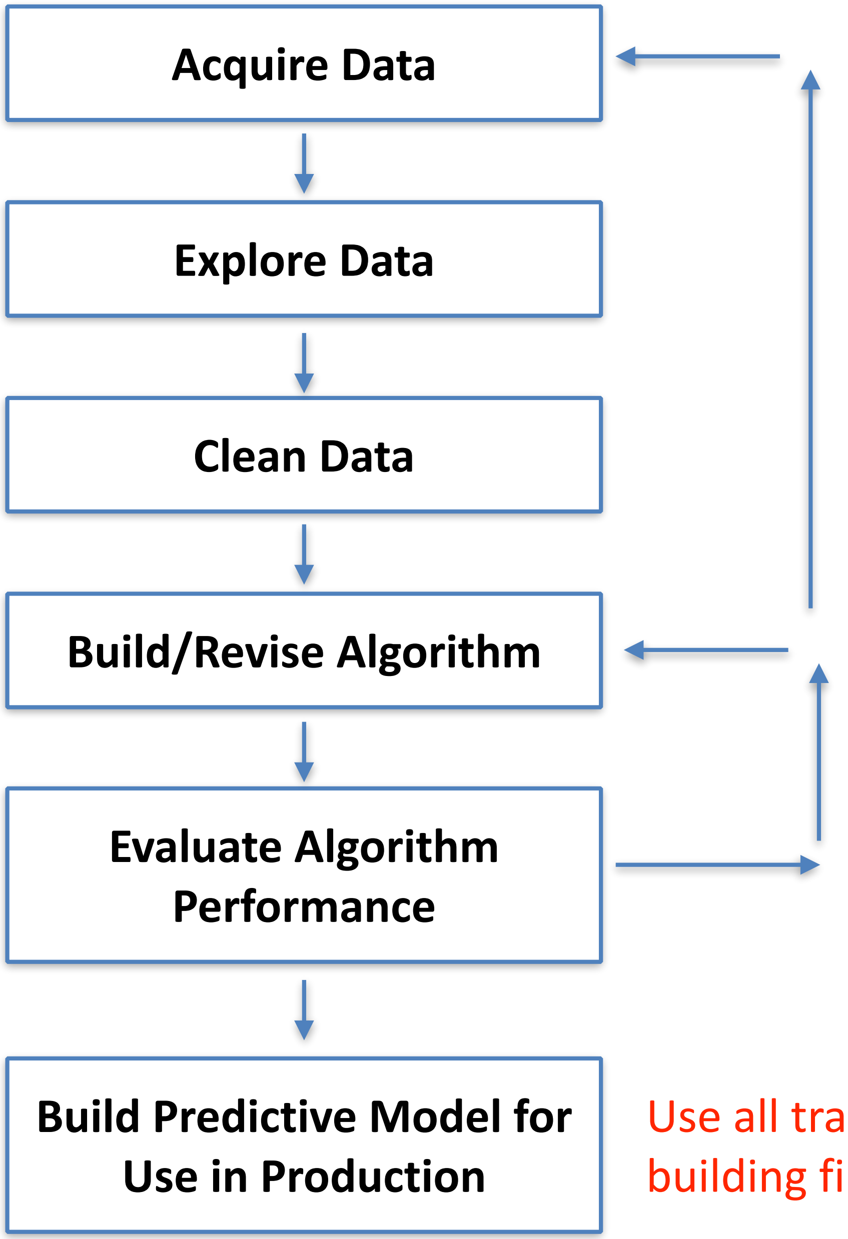
Training Set

Age	*Sex*	*Pclass*	*Survived?*
25	*Male*	*3*	*FALSE*
17	*Female*	*1*	*TRUE*
40	*Male*	*2*	*FALSE*
9	*Female*	*2*	*TRUE*

Classes (target)

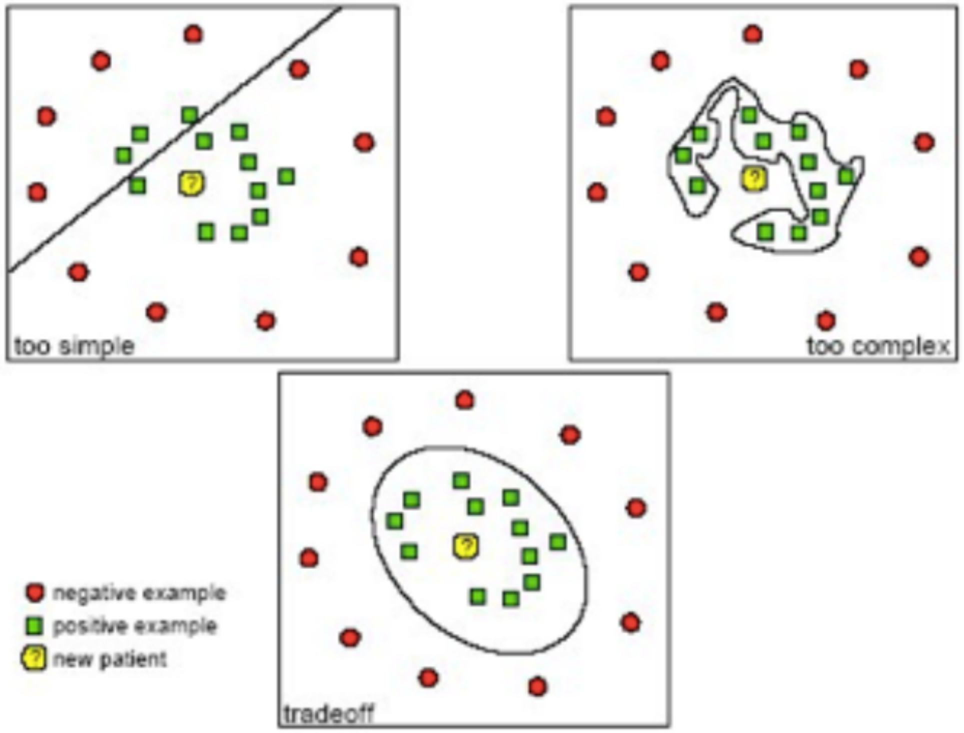
Independent Variables (features)

Data Science Process


Acquire Data
Explore Data
Clean Data
Build/Revise Algorithm
Evaluate Algorithm Performance
Build Predictive Model for Use in Production

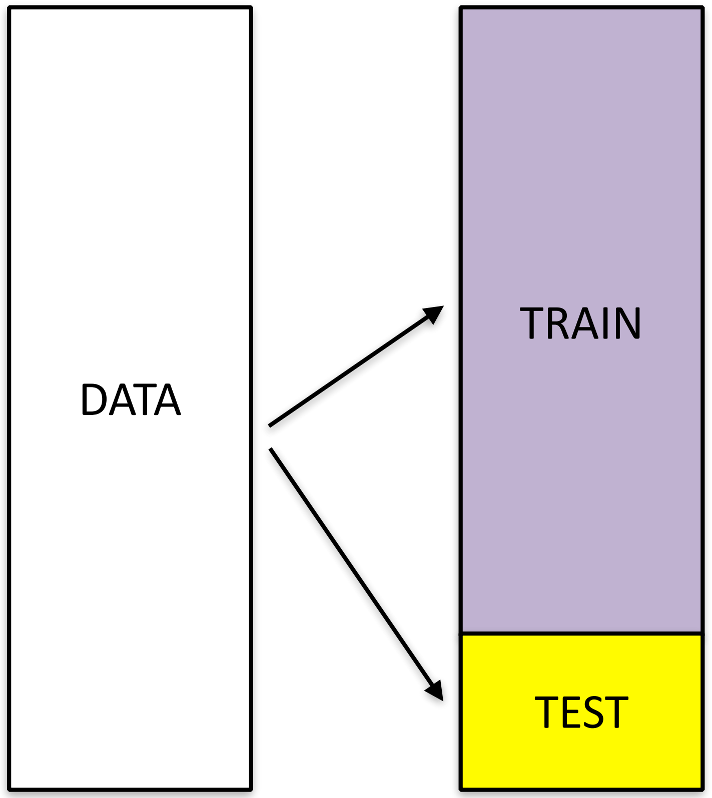
Use all training data when building final model

Underfitting and Overfitting



Cross-validation

Split Data Set



- Separate the data into two groups
- Use part of data to train the model
- Use trained model to predict out come for test data
- Compare predictions with actual results

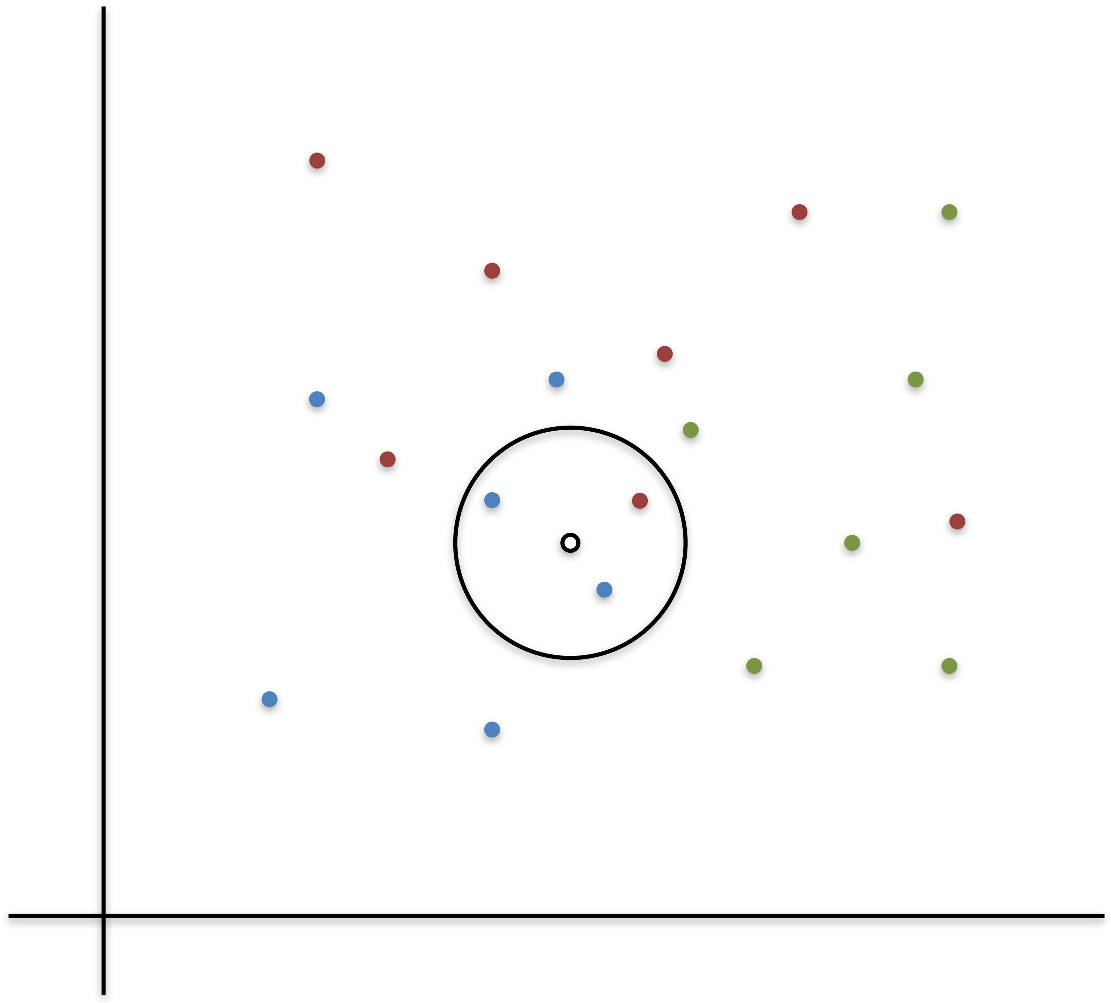
Measure model performance on accuracy of predictions

QuantSprout

Classification Algorithms

- Logistic Regression
- Naive-Bayes
- Decision Trees
- Support Vector Machines
- Neural Networks
- K-Nearest Neighbors
- Random Forest

K Nearest Neighbors



Suppose you want to predict the white dot

1. Pick a value for k

2. Find colors of the k nearest neighbors

3. Assign the most common color

KNN Considerations

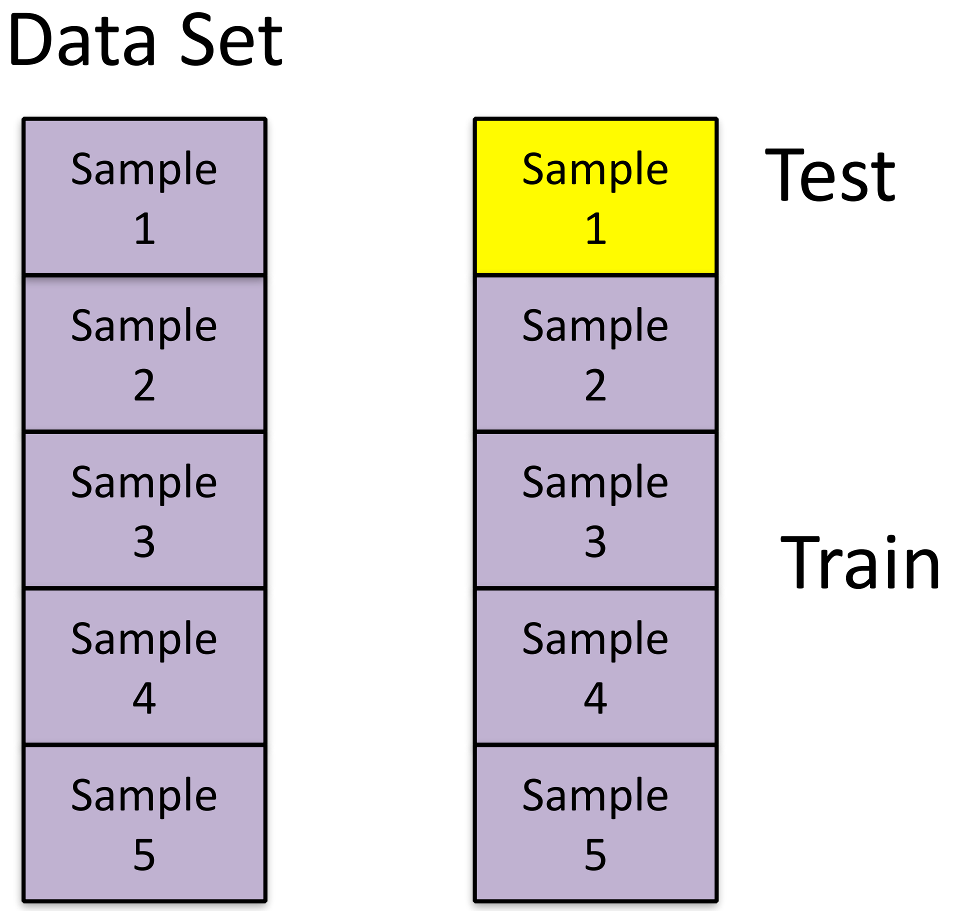
- Scaling of Data has large impact on algorithm (normalization is frequently used)
- Adjust the parameters of the algorithm
    - Number of neighbors
    - Treat data points uniformly or weighted by distance
- Can become computationally expensive at large scale

Evaluating Algorithms

- Model performance will vary based on how you split the data into training and testing groups
- Taking an average of the model performance across different splits improves the measurement

KFold Cross-validation

K-Folds Cross-validation

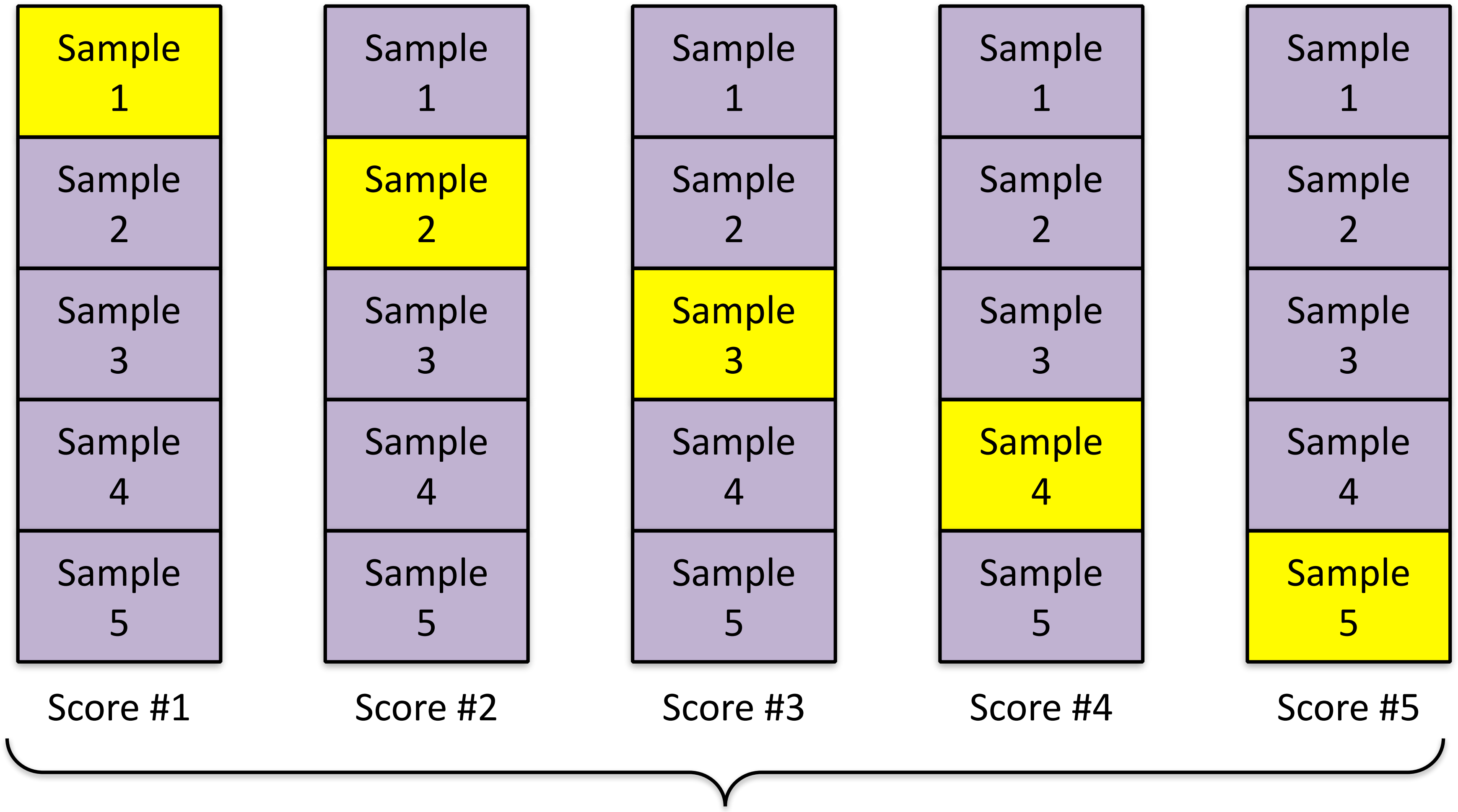


- Separate the data into K groups (e.g., 5)
- Each group maintains the same samples for all tests
- Train and test K times
- Sample group for test changes for each test

Every Data Point Is Tested Once

QuantSprout

K-Folds Cross-validation

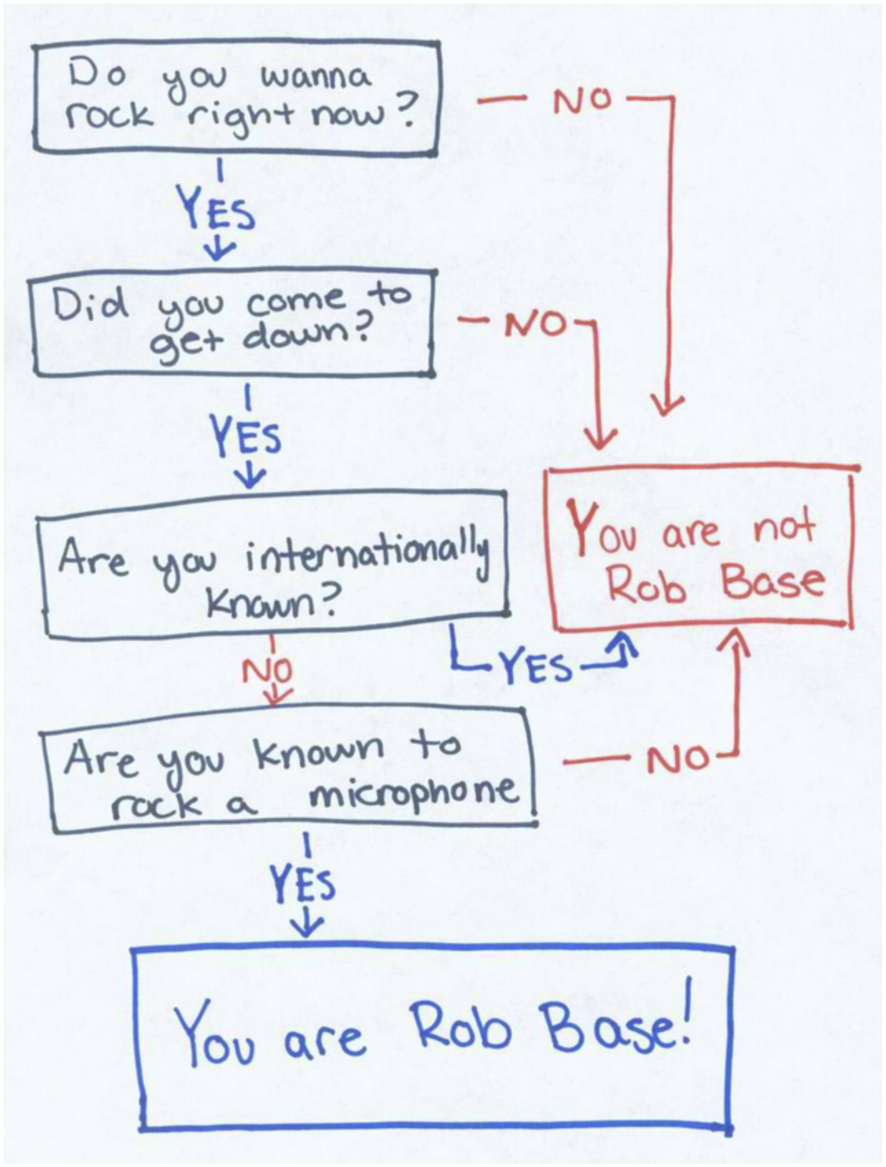


Average of Scores

Data Normalization

- Technique for adjusting the scale of data
- Calculate mean and standard deviation of all samples for each variable
- Subtract the mean and divide by the standard deviation

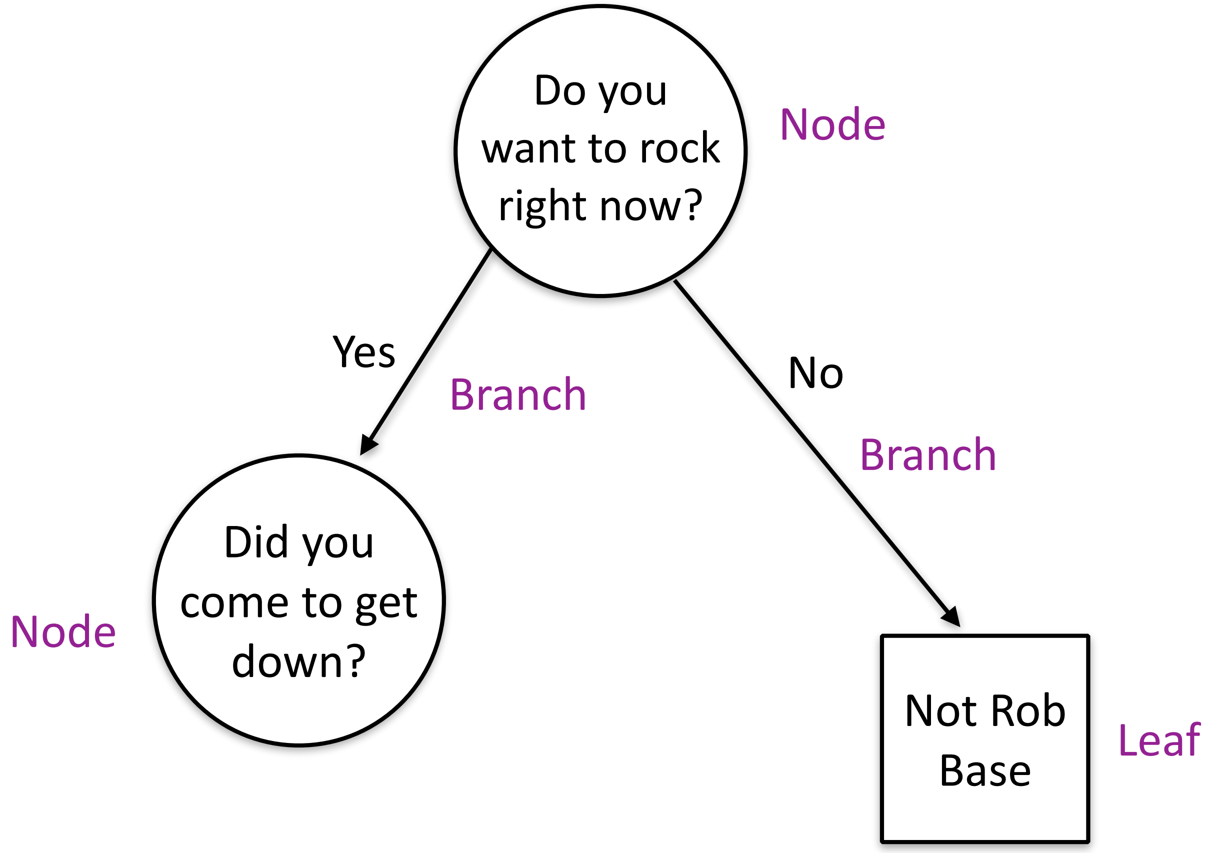
Is This a Decision Tree?


Do you wanna rock right now?
NO
YES
Did you come to get down?
NO
YES
Are you internationally known?
YES
NO
You are not Rob Base
Are you known to rock a microphone
NO
YES
You are Rob Base!

QuantSprout

Decision Trees

- Trees consist of:
  - Nodes (questions)
  - Branches (answers)
  - Leaves (end points)
- Acyclic - flows in one direction
- No split is necessary when all records are from same class



Decision Trees

- Algorithm selects optimal node that creates the largest increase in purity
- Decision trees are susceptible to overfitting
- Techniques for preventing overfitting
    - Minimum number of records for leaf
    - Maximum depth for branches
- One technique is to purposely overfit, but manually prune branches

Random Forest

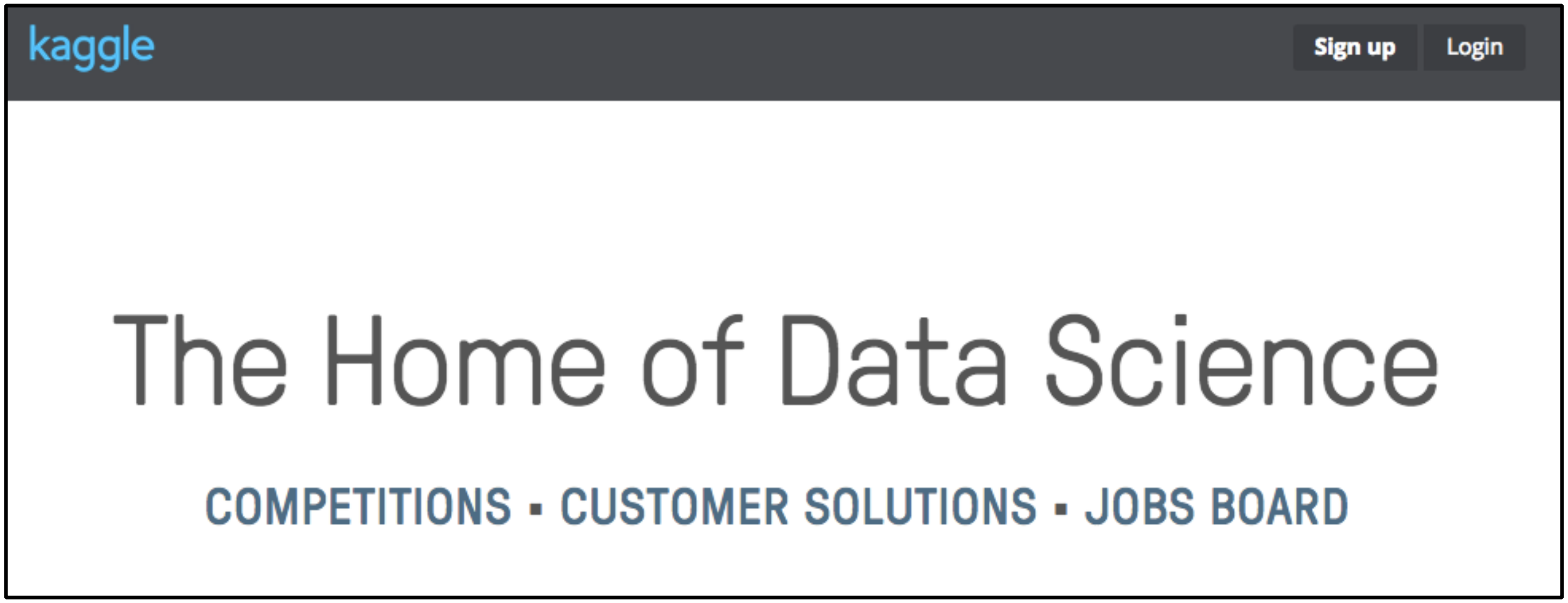
- Ensemble algorithm (mix of many models)
- Collection of decision trees
- Evaluates predictions from many models and selects the most common classification
- Features are randomly selected for each decision tree
- Bagging (Bootstrap Aggregating) is also applied to each tree
  - Sample of the data set is used for training
  - Samples are drawn with replacement

Why Random Forest Is Popular?

- Add all your data and algorithm will prioritize
- Not susceptible to overfitting
- Doesn’t require normalization
- Solid performance in wide range of applications

Kaggle

- Data prediction competitions



The Home of Data Science

COMPETITIONS ▪ CUSTOMER SOLUTIONS ▪ JOBS BOARD

Kaggle

- Create an account
- Find “Titanic: Machine Learning from Disaster”
- Submission Instructions:

  “You should submit a csv file with exactly 418 entries plus a header row. This must have exactly 2 columns: PassengerId (which can be sorted in any order), and Survived which contains your binary predictions: 1 for survived, 0 for did not.”

Make Predictions

- Train your machine learning algorithm with training data
- Read the test.csv file
- Clean test data
- Predict outcomes for test data
- Convert predictions into a DataFrame
- Save DataFrame as a CSV file (remember to exclude the index)
- Submit predictions to Kaggle site

Types of Regression Models

- Linear
  - Ordinary Least Squares
  - Generalized Linear Models
- Regularized Models
  - Ridge
  - Lasso
  - Elastic Net

Types of Regression Models

- Non-linear
  - Logistic
  - Polynomial
- Trees
  - Random Forest
  - Gradient Boosted Trees
- Autoregression

Regression Process

Data Exploration

Model Selection

Feature Selection

Model Evaluation

- **Visualization of data**
- **Data cleaning**
- **Correlation between features and target**
- **Correlation between different features**
- **Outlier detection**

Regression Process

Data Exploration

Model Selection

Feature Selection

Model Evaluation

- **Linear vs. Nonlinear**
- **Regularization**
- **Evaluate model pre-requisites**

QuantSprout

Regression Process

Data Exploration

Model Selection

Feature Selection

Model Evaluation

- **Under vs. Overfitting**
- **Feature Engineering**
- **Data Transformations**
- **Forward and Backward Stepwise Selection**
- **Multicollinearity**

Regression Process

Data Exploration

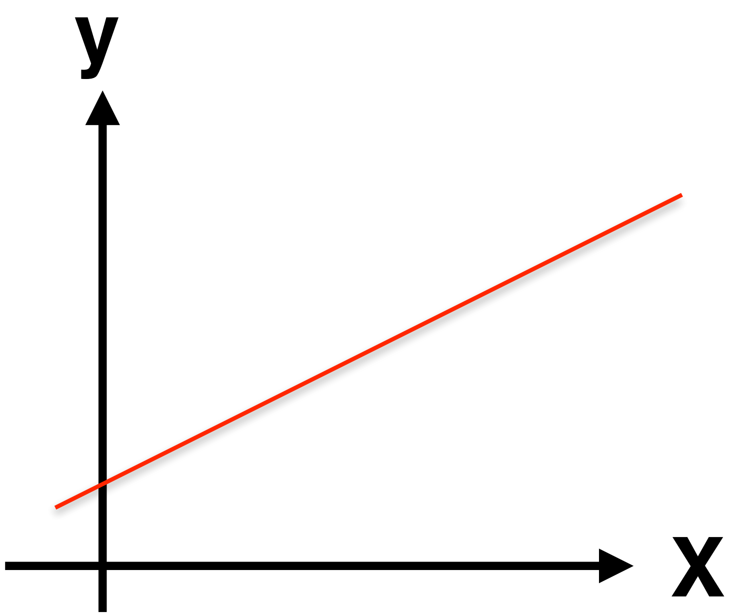
Model Selection

Feature Selection

Model Evaluation

- **Root Mean Square Error**
- **R-squared**
- **Residual Analysis**
- **Statistical Significance of Coefficients**
- **Crossvalidation**

Simple Linear Regression



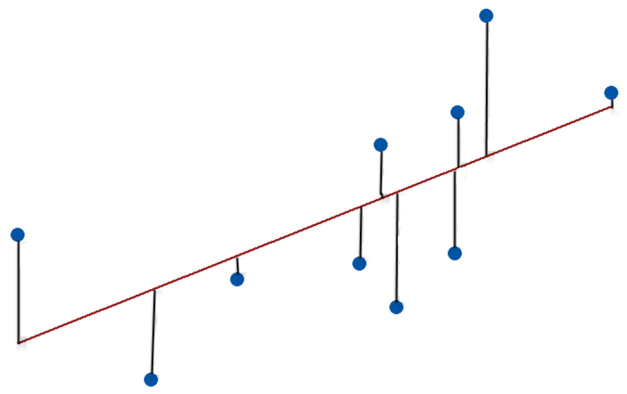
$$y = mx + b$$

- y - dependent variable (target)
- x - independent variable (feature)
- m - slope (change in y for each unit change in x)
- b - intercept with y axis

QuantSprout

Ordinary Least Squares (OLS)

- Algorithm to estimate coefficients of linear model (slope **m** and intercept **b**)
- Coefficients are adjusted to minimize an error term
- Error term is calculated by summing the square of the distance between the predictions and the actual data points

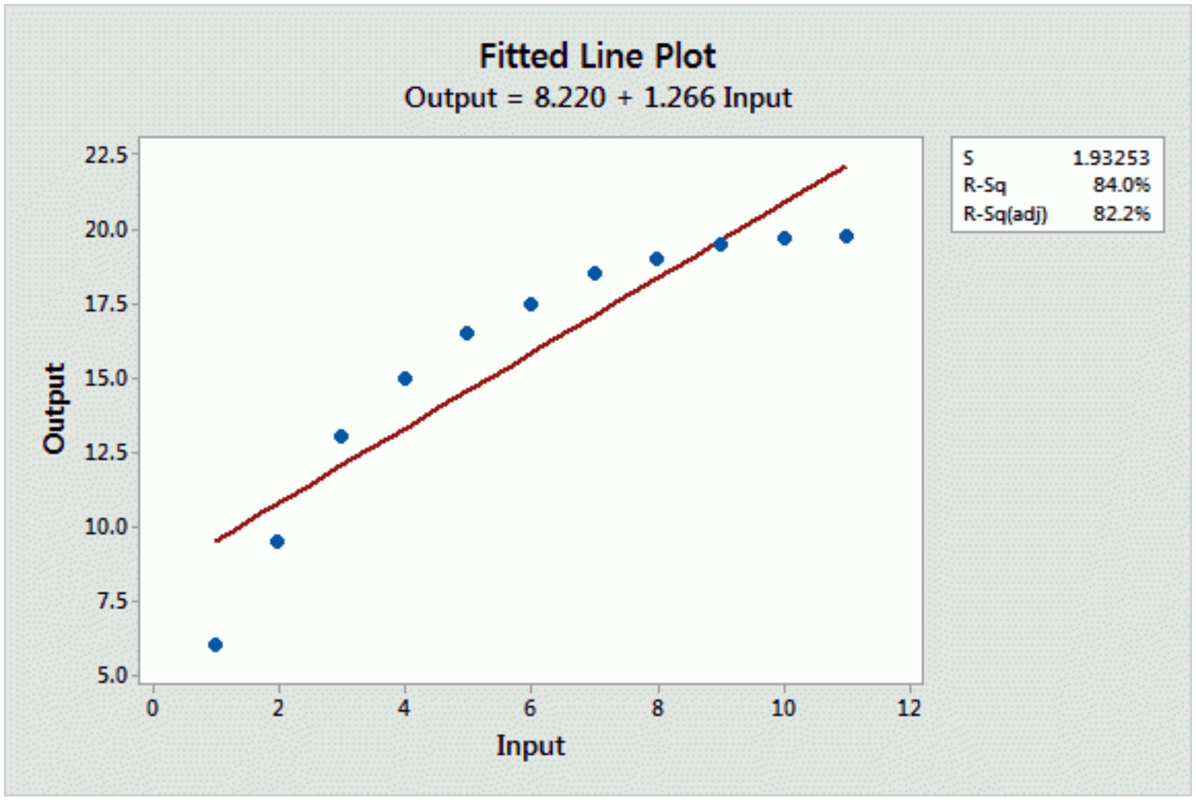


$$y = mx + b$$

QuantSprout

OLS Requirements

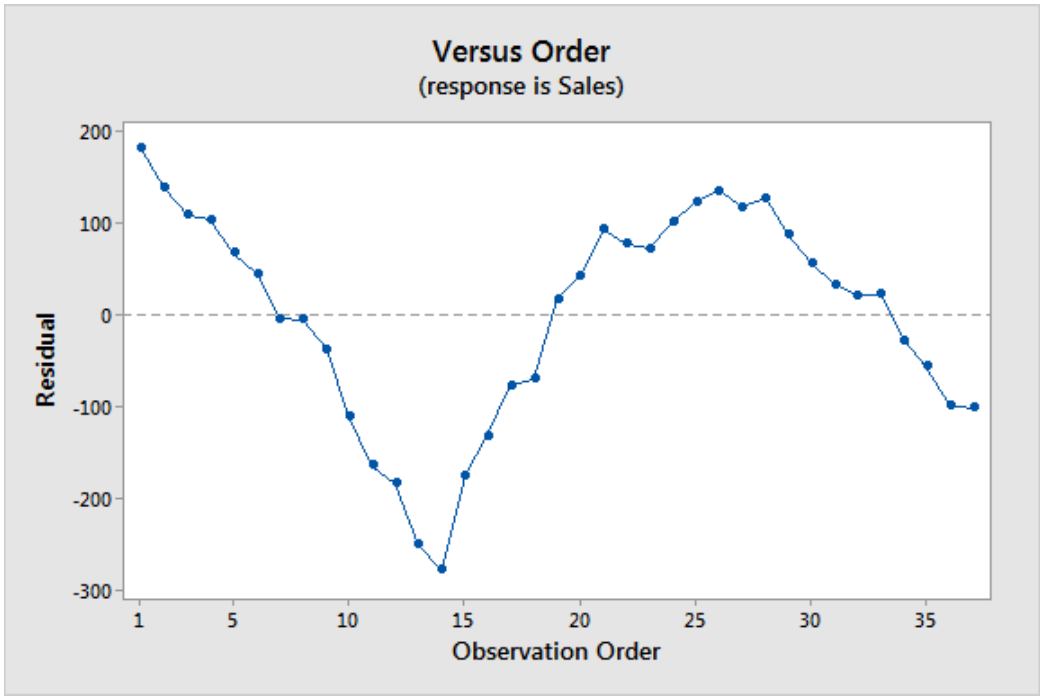
- Linear relationship between independent and dependent variables



QuantSprout

OLS Requirements

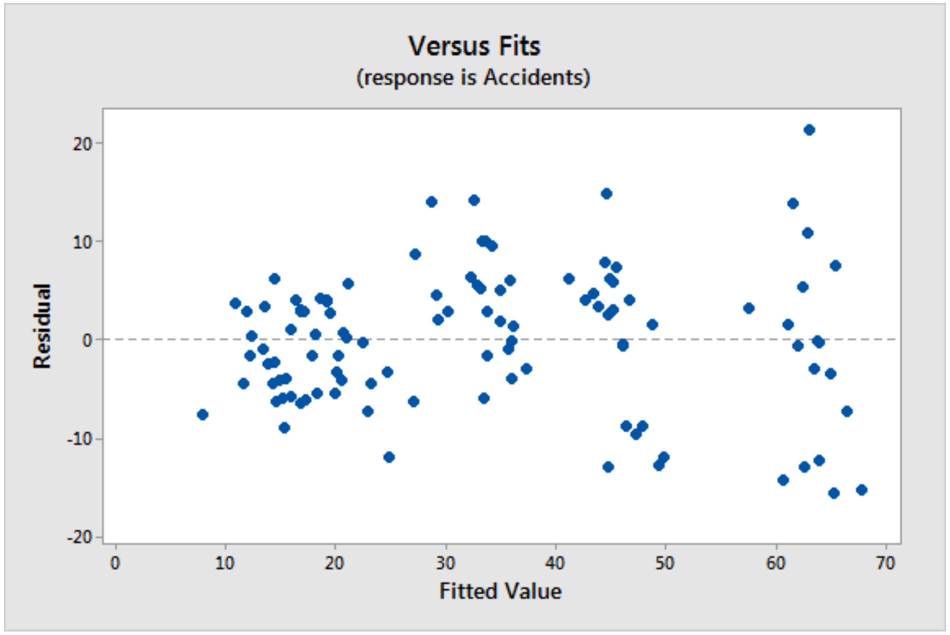
- Error terms has population mean of zero
- Independent variables are uncorrelated with error terms
- Error terms are uncorrelated with each other



QuantSprout

OLS Requirements

- No independent variable is a perfect linear function of other explanatory variables
- Error term has constant variance (no heteroscedasticity)



Root Mean Squared Error

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(P_i - O_i\right)^2}{n}}$$

- Standard deviation of residuals
- Metric for model accuracy
- Measurement in units of dependent variable

R-squared

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

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

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

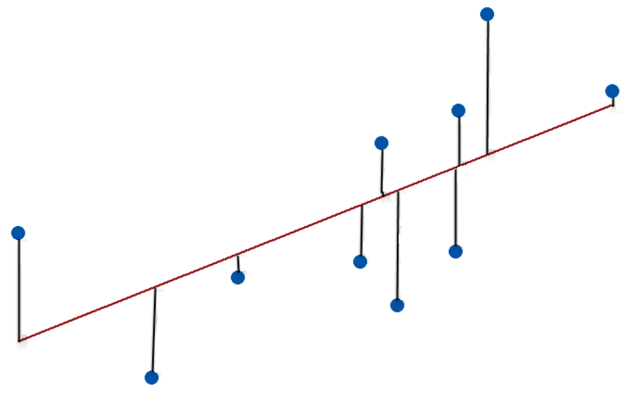
- Coefficient of determination
- Metric for model fit
- Proportion of variance in dependent variable that is predictable from independent variable
- Range: 0.0 ~ 1.0

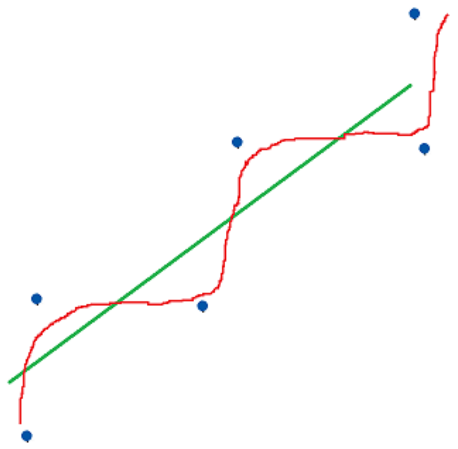
Outliers

- Not representative of population
- Skew calculation of coefficients
- Data visualization for detection

Overfitting

- Error term from unobserved random variable / noise
- Simplify models (reduce features) to avoid overfitting
- Overfit models can misinterpret noise as a signal





$$y = mx + b + e$$

QuantSprout

Multicollinearity

- Independent variables are correlated
- Coefficients can swing wildly and become sensitive to small changes in model
- Difficult to interpret and trust coefficients
- p-values can become untrustworthy and hard to validate statistical significance of coefficients

Regularized Models

- Models add penalties to complexity
- Naturally reduces overfitting
- Helpful for selecting features from large feature set
- Try Lasso model and Ridge Regression

Next Steps

- Practice, Practice, Practice
  - Find a fun project
  - Kaggle Competitions
  - Pair programming buddy
- Join a Python Meetup
  - San Francisco Python - recommend Project Night event
  - Bay Area Python Interest Group (BayPIGgies)

yelp.com/biz/quantsprout-san-francisco