Just a three hour tour...

Machine Learning with SQL Server

Ginger Grant

May 25th 2018

Applied Agenda

- Python
- Machine Learning Process
- Data Science
- Algorithms
- Linear Regression
- Python and SQL Server

About Me

DESERT ISLE GROUP

Principal Consultant

@DesertIsleSql

www.desertislesql.com

Microsoft Most Valuable Professional

Microsoft Certified Trainer — MCT — 2018 - 2019

About You

Name

What you do (or where you work)

What you want to learn in this class

If you are on twitter 

Environment Setup

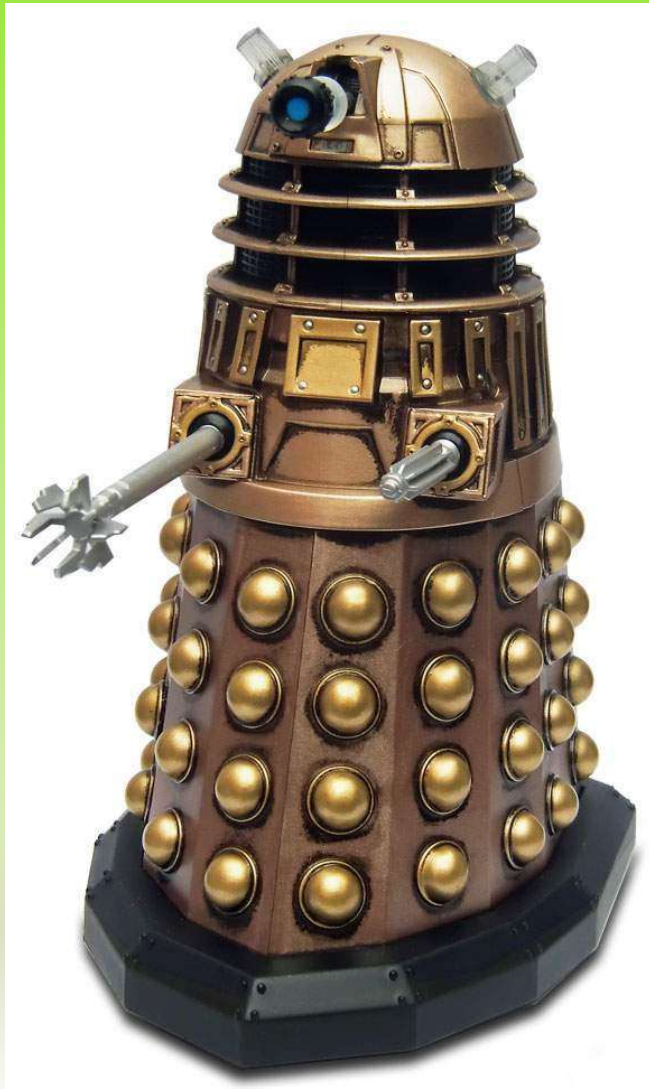
- SQL Server 2017 Developer Edition
  - Machine Learning Services
  - Python
  - R

- Azure Machine Learning Account

  studio.Azureml.net

What's The Difference?

- **Artificial Intelligence**
- **Machine Learning**
- **Data Science**



Machine Learning

- ***"Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions."*** **Techemergence Sept 2017**

Two Classes of Machine Learning

- **Supervised Learning has a defined set of inputs and outputs**

Two Classes of Machine Learning

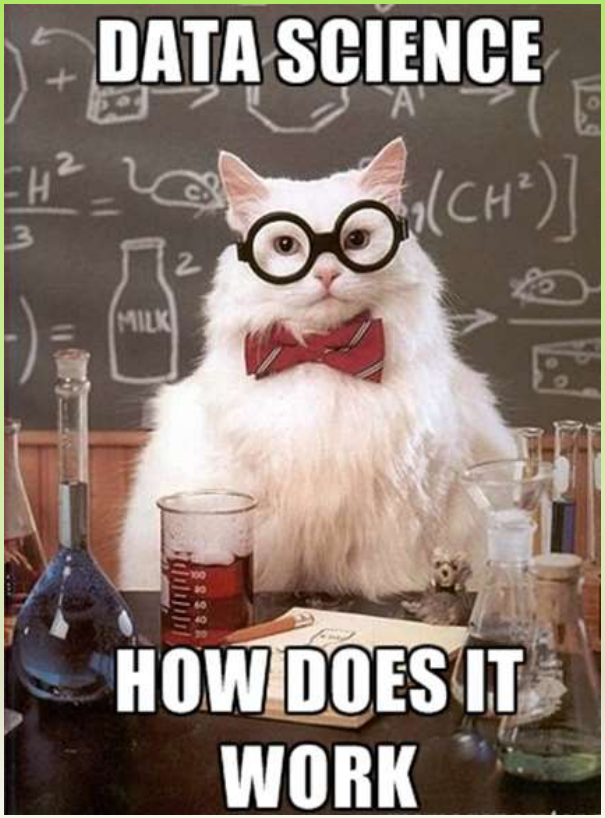
- **Supervised Learning has a defined set of inputs and outputs**
- **Unsupervised Learning is an open ended study of the inputs to find the outputs**

Algorithms

- Method used to teach the computer
- Using math to find the patterns in the data
- Number of different algorithms that can be used for any problem
- Implemented with libraries of pre-written code

Implement Machine Learning ?

- **R**
- **Python**
- **Dedicated Machine Learning Tools like Azure ML**

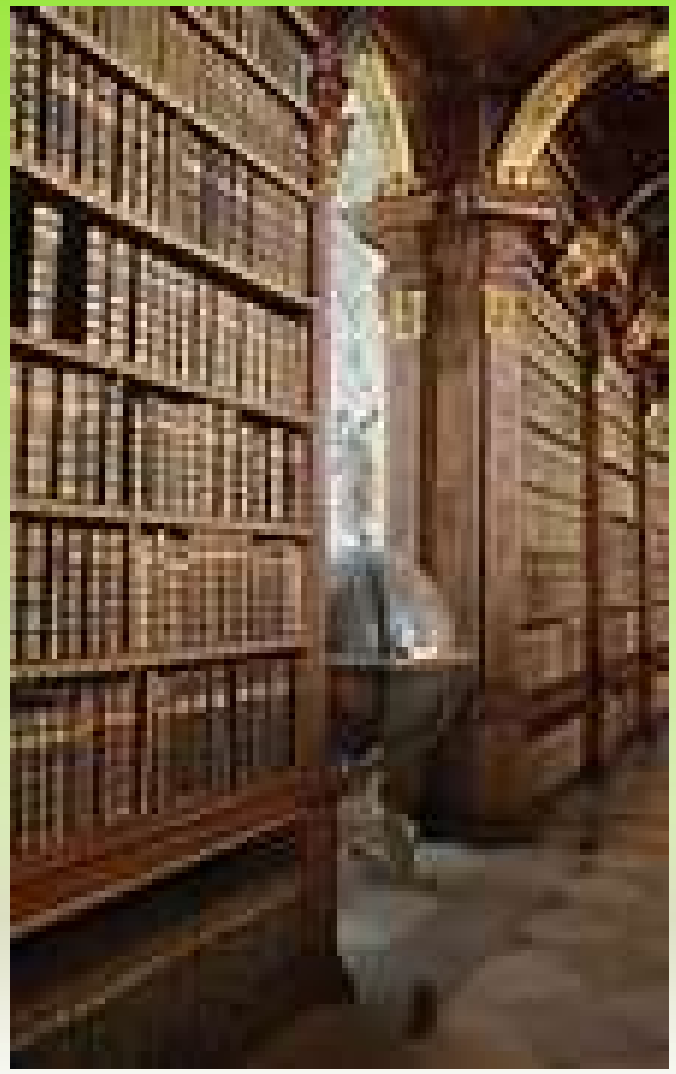


History of R

- Started as S Bell Labs in 1976
- 1996 Two Kiwis Ross Ihaka and Robert Gentleman created the Open Source version, called R
- Applies complex calculations of datasets and provides visualizations of the results
- In April 2015, Microsoft bought an R vendor and are incorporating R into their products

R Extensibility

- **Common Libraries for Graphics and Statistical Analysis**
- **Large User Community which generates functions to be used in R**
- **CRAN - Comprehensive R Archive Network**
- **Support included in a number of different applications**



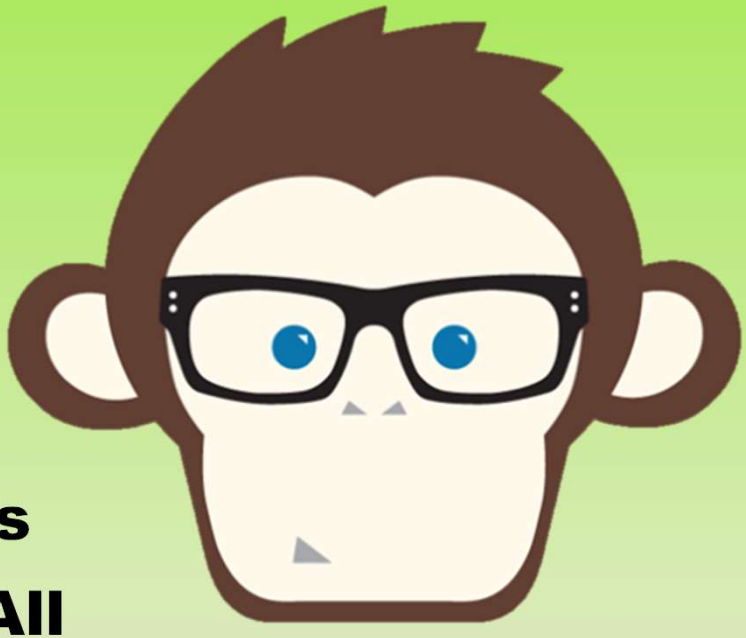
CRAN R

- R Project for Statistical Computing
- Open Source Version of R

  www.r-project.org
- Maintained by the R consortium
- Academic focus
- Many libraries written and Rewritten for R
- Includes R Console UI

Microsoft R Open

- Fully Compliant version of Open Source R
- Rewrote many of the commonly used statistical functions in C
- 38% Faster than Open Source R
- Uses Multi-threaded Math Libraries
- Nothing to do with SQL Server at All



UI Choices:
R Studio, Visual Studio, Jupyter

- R Studio is the UI written for R
- Visual Studio 2017 Community
- Interface appears very similar
- Intellisense
- Flexibility

Demo R UI

Login to Studio.AzureML.Net

Basic Language Constructs

- <- and =
- Everything is a Vector
- Null and NA
- For Loops Don’t Work Well
- Case Sensitive

Basic Language Constructs

- ***#*** **- comment**
- **Help**

help("lm")

?lm

Library Loading

```
if (!require("ggplot2")) {
  install.packages("ggplot2")
}
```

R Objects

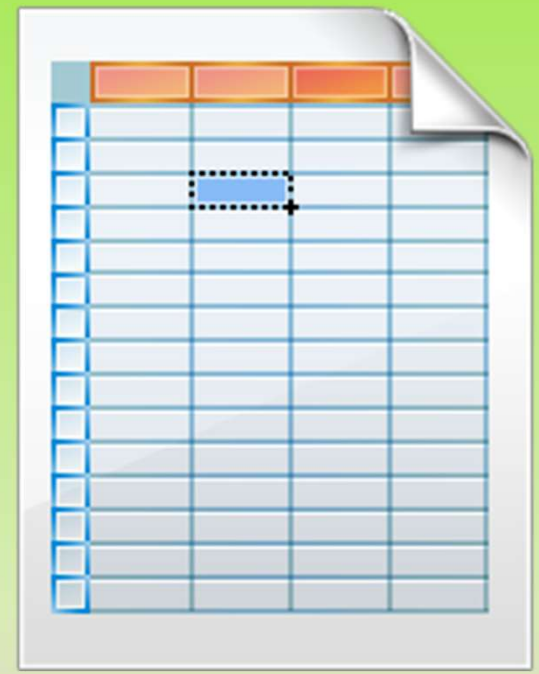
- Everything in R is an Object
- All Data is a Set
- Everything is Stored in Memory

R Objects and Storage Type

- Vectors
  - [23] – Vector with one item
- Matrices
  - Rows and Columns with one Data Type
- Data Frames
  - Rows and columns of different Data Types

List

Collection of other Objects



Vectors

- **Building Blocks for data objects in R**
- ***c* (combine) function to create a Vector**

```
v <- c(2, 3, 1.5, 3.1, 49)
```

- ***seq* function generates numeric sequences**

```
s <- seq(from = 0, to = 100, by = .1)
```

- ***rep* function replicates values**

```
r <- rep(c(1,4), times = 4)
```

- ***:* (Colon) creates a number sequence incremented by 1 or -1**

```
colon <- 1:10
```

- **length(var) returns length of vector**

```
length(colon)
```

Matrix

- Loading data done through binding
- Combine matrices by row or column
- rbind (row bind)
- cbind (column bind)

Data Frame

- **Most common way data is stored**
- **Very similar to a table in SQL**
- ***rownames* – extract row labels**
- ***colnames* – extract column labels**
- ***read.table, read.csv, readxl, RODBC***
  - **Different ways to create data frames**

List

- Combine multiple objects types into one object
  - vectors, matrices, data frames, list, functions
- Typically used by functions to output the model output
  - e.g. the output from the lm function

Packages – R Add ons

- **List packages already installed**
  library()
- **Display loaded Packages**
  Search()
- **Install package**
  install.package("dplyr2")
- **Load package to be used in R**
  library(dplyr2)

Univariate Analysis In R

- Plot
- Stem
- Hist
- Summary

Loading Data

- **Loaded into your Working Directory**

```
getwd()
```

- **Load data from a library**

```
data(package = "ggplot2")$results
```

- **Load Data from a File**

```
bankData <- read.table("c:/data/bank-
full.csv",header = T, sep = ";",
stringsAsFactors = F)
```

Demo R UI

Login to Studio.AzureML.Net

ExeRcise

Access Shared Folder

Login to studio.AzureML.net

Create a new R Notebook

Open up Rexercise.txt into a notebook

Paste and run the text in JupyterNotebooks

Why Learn Python?

"***For the fifth year in a row, Python retains it's #1 dominance***"

Code Eval Feb 2, 2016

"***Python edges out C and Java to become the most popular programing language***" **zdnet July 21, 2017**

"***Python overtakes R, becomes the leader in Data Science, Machine Learning platforms***" **KDNuggets Aug 2017**

Introduction to Python

Released in 1994 by Guido Van Rossum

Backwards incompatibility between versions 2 and 3 staring in 2008

Python 2 Code

```
5/2  == 2
```

Python 3 code

```
5/2 == 2.5
```

The Language was not named after a snake

Introduction to Python

Released in 1994 by Guido Van Rossum

Backwards incompatibility between versions 2 and 3 staring in 2008

Python 2 Code

```
5/2  == 2
```

Python 3 code

```
5/2 == 2.5
```

The Language was not named after a snake

It was named after Monty Python's Flying Circus as Guido is a fan

Python Development Environment

PyCharms

Spyder

Atom

Sublime Text

Visual Studio

Rodeo

Jupyter Notebook

Demo

Python GUIs

Python vs R

- History of R
- CRAN R Libraries
- Microsoft Investment

R Strengths

- **Loads data in Memory for Rapid Data Analysis**
- **Modular, weakly typed design is very forgiving**
- **Large number of pre-built code libraries**

R Weaknesses

- **Limited by available memory**
- **Developed by Statisticians, not developers**
- **Single threaded processing**

Data Science with Python

- **Newer Use of the Language**
- **Provides the ability to call libraries with algorithms**
- **Create visualizations for analysis**

Data Analysis Trends

- Kaggle more people using Python than R
- Advanced Library Development
- Continuum Analytics Library Re-write

Demo

Python Machine Learning Solution – Predictive Maintenance

Basic Language Constructs

- **Data Types**
  - `Integers, floating points, strings`
- **Tuples, Lists and Dict**
- **None**
- **Standard Loops – For and While**
- **Case Sensitive**
- **Functions, Classes and Objects**

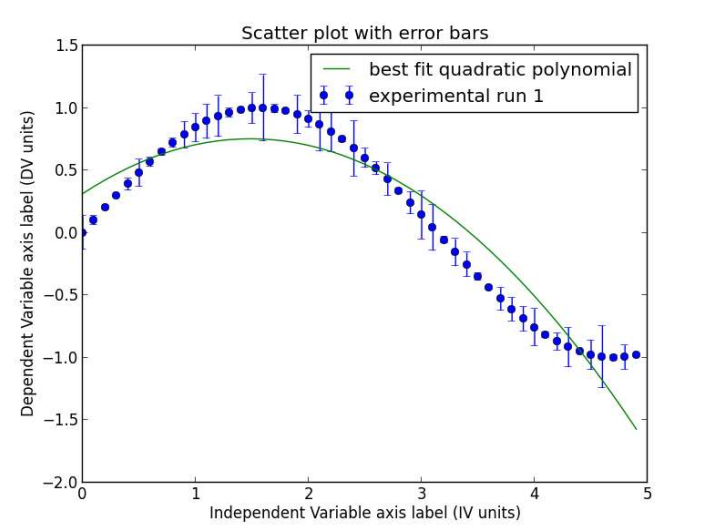
Demo

Sample Python Code

Python Packages to Know

- **Matplotlib**
- **NumPy**
- **Pandas**
- **SciPy**
- **Scikit-learn**
- **Anaconda**
- **Jupyter Notebook**

Matplotlib for Pictures of data



- **Primary method for visualizing data**
- **Standard reference is plt**
- **2D libraries**

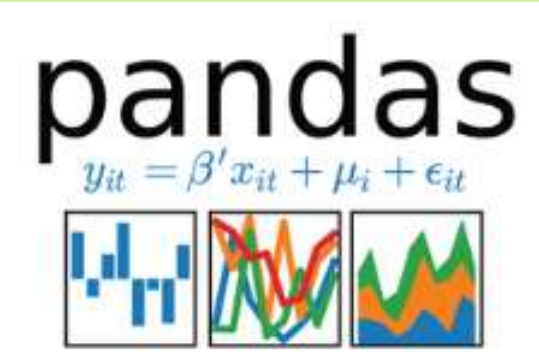
matplotlib

NumPy

- Developed for Linear Algebra computations
- Array structure where All values must be of one datatype
- Allow calculations across entire arrays
- Included as a dependency for a number of different libraries

Pandas

- **Built on the Numpy package**
- **Modeled after the R DataFrame**
- **Table-like structure of data**
- **Pandas.pydata.org**



SciPy

- The core package for scientific routines in Python
- Designed for use with Numpy Arrays
- Scientific Ecosystem includes the following
  - Numpy
  - Matplotlib
  - Pandas
  - Sympy
  - Ipython
  - Cython

SciKit-learn

- **Machine Learning Library**
- **Includes Supervised and unsupervised learning algorithms**
- **Dependent on a number of other Python Packages:**
  - **Numpy**
  - **SciPy**
  - **MatPlotlib**
  - **Sympy**
  - **Pandas**

Anaconda

- “Batteries included”
- Loads all of the other libraries
- “Fremium”
- Conda
- Miniconda
- Included in the release of SQL Server

ANACONDA®

Jupyter Notebook

- Interactive environment for code
- Provides documentation
- Number of different Languages Supported
- Included with Anaconda
- Free version with studio.azureml.net

Demo

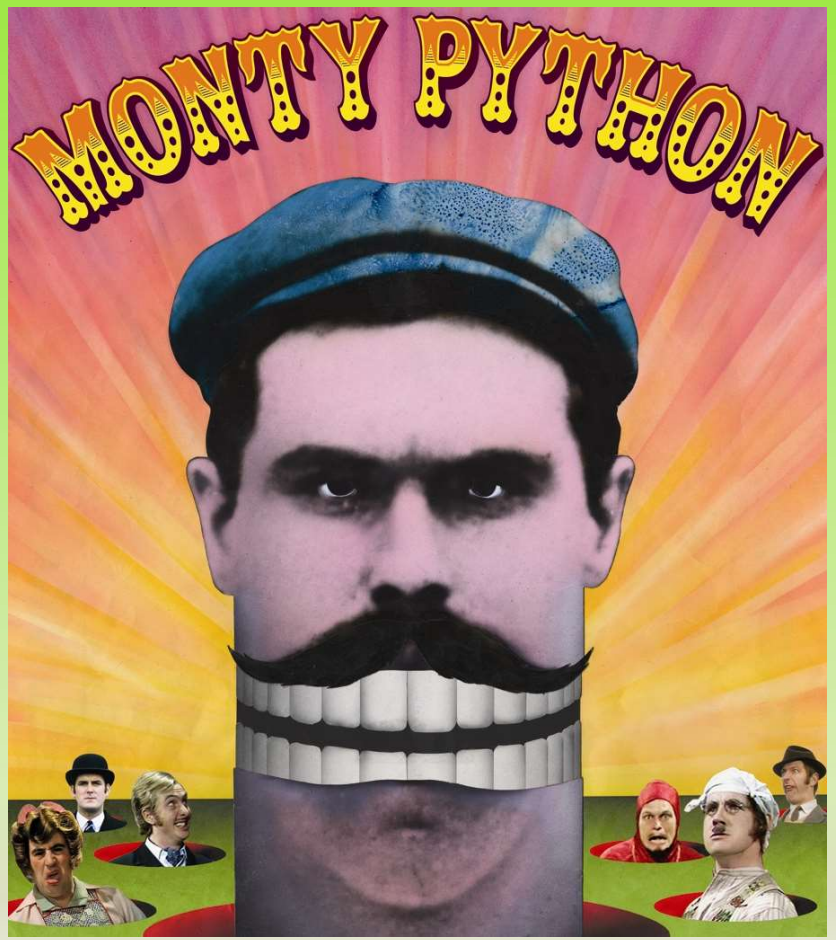
Anaconda

Exercise Python Code

Create a new Juypter Notebook

Copy the values from HelloWord.py to it

Monty exercise



exercise Review

Machine Learning And Data Science

- Machine Learning is the key tool that data scientists use to answer questions with data
- Uses algorithms to teach the computer
- Discover patterns in data
- Save the patterns in a model
- Use the model to analyze more data to determine a result

Data Science

"Data Scientists spend over 60% of their time cleaning and organizing data."

Forbes, March 23, 2016

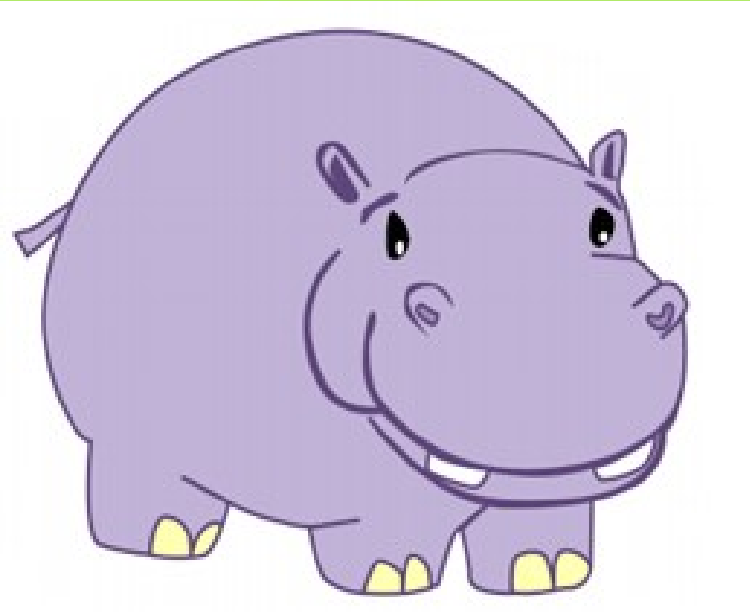
"Data Scientist:
The Sexiest Job of the 21st Century"

HBR, October 2012

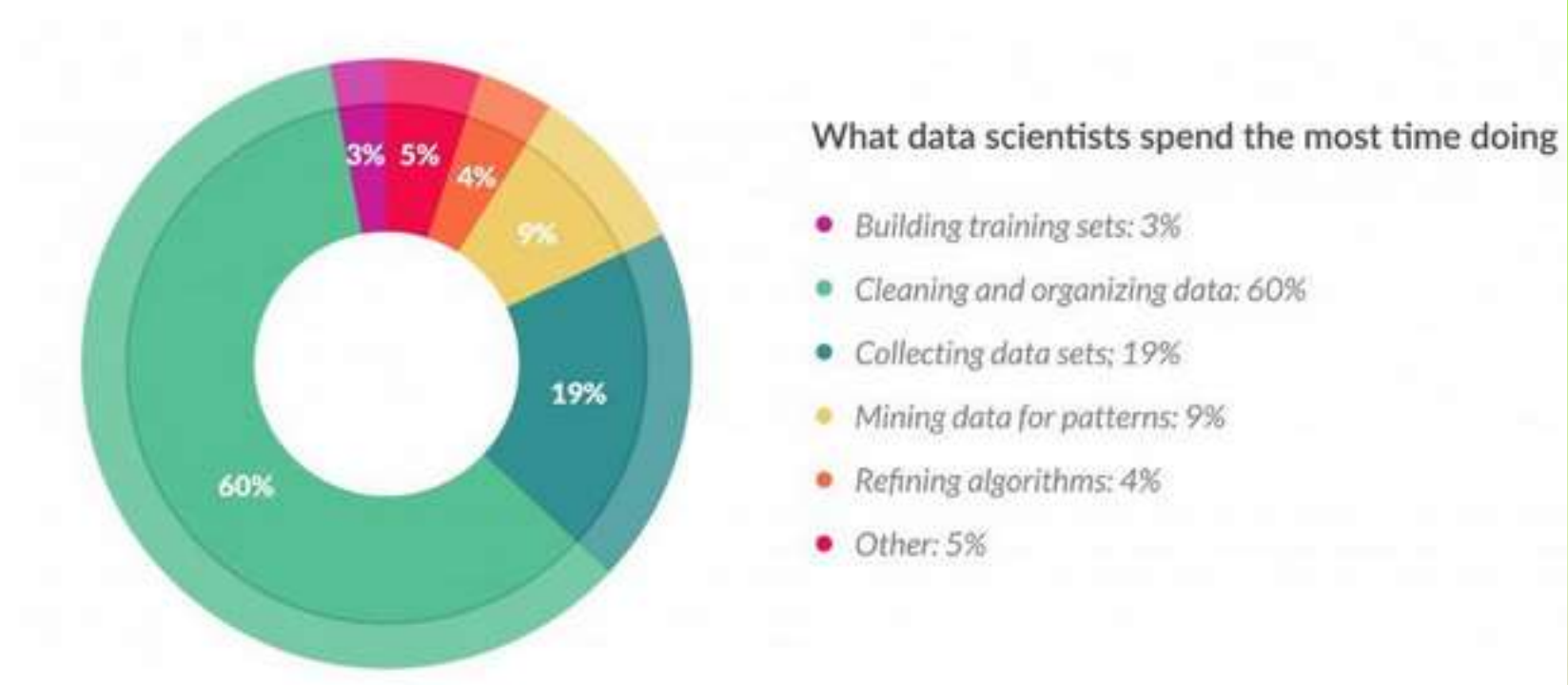
- **Professional with the training and curiosity to make discoveries in the world of big data**
- **Scientific methods, processes and systems to extract knowledge or insights from data**

Data Driven Decision Making

- **Data Analysis shown to be a superior method of supporting decision making**
- **Vast quantities of data now stored can be used to establish patterns**
- **H.I.P.P.O decision making statistically inferior**

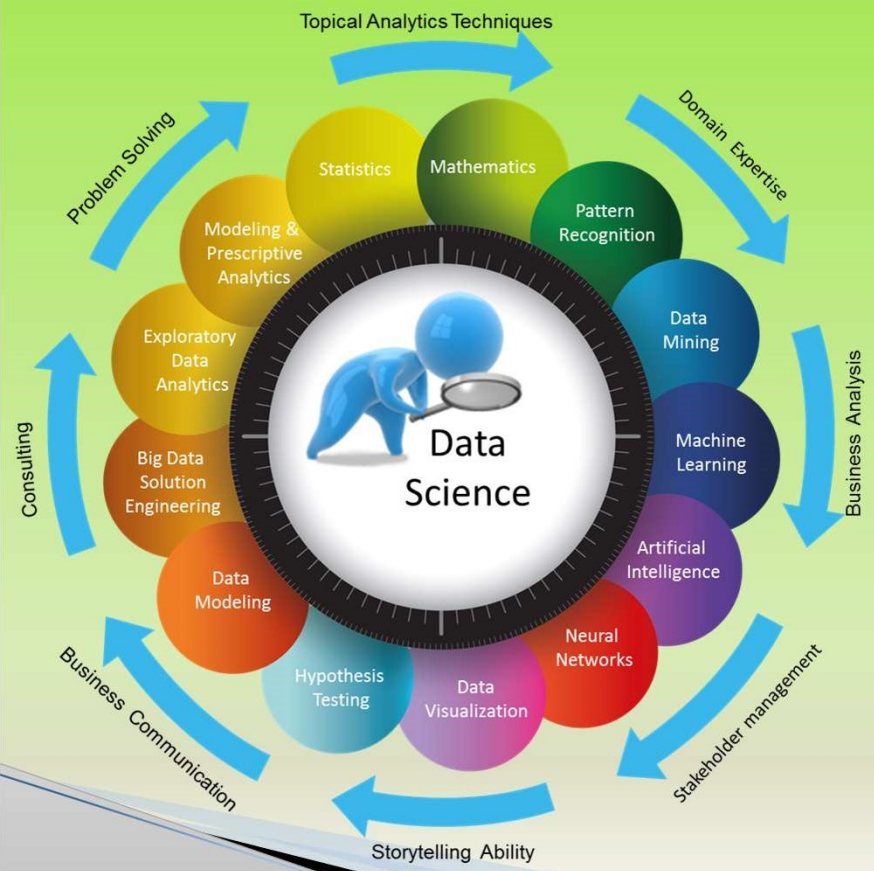


Data Scientist Spend lots of time on Data Prep



Forbes March 23, 2016 ***Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says***

Applied Data Science -CRISP

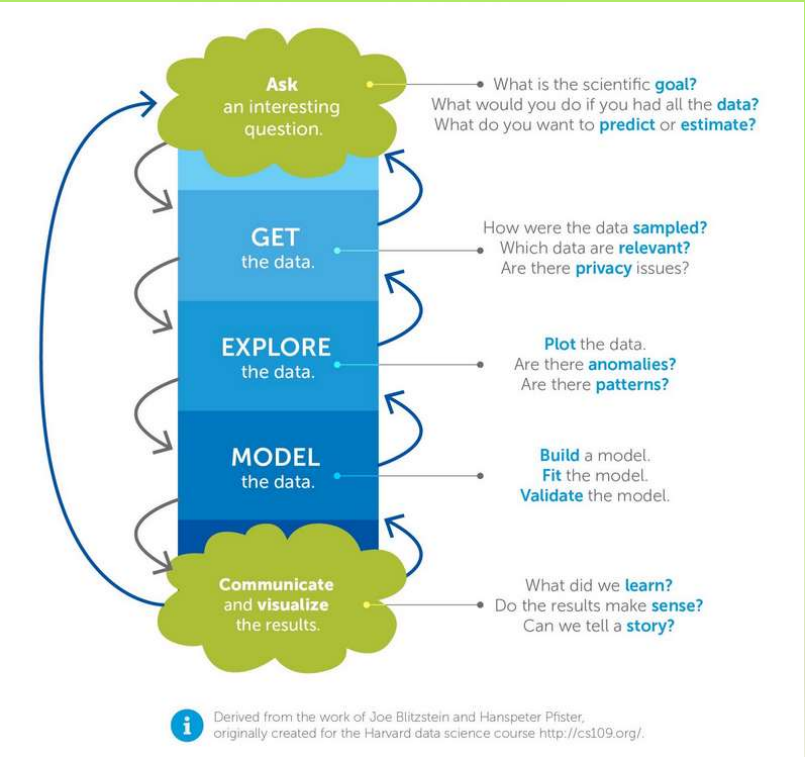


- **Get it out of the theoretical into the practical**
- **Learn the tools which have the processes created**
- **Term Familiarization**
- **Algorithm Use**
- **Discover Patterns in Data and Put them to use**

5 Questions Machine Learning Answers

- Is there a relationship here?
- Is there something wrong?
- Predict what will happen.
- How is this organized?
- What should I do next?

Analysis Process



- **Question**
- **Get the data**
- **Find patterns in the data Exploration**
- **Start creating hypotheses**
- **Distribution of Results – SQL Server**

Steps to Perform ML Analysis

- Acquire Relevant Data
- Exploratory Data Analysis
- Feature Engineering
- Split the data into Training and Test
- Apply an Algorithm (or two, or more)
- Analyze the results
- Determine the best model
- Deploy the Model

Data Acquisition

- Level of Granularity
- Missing Data
- Erroneous Data
- Timeframes
- Representative Sample

Exploratory Data Analysis

- **Univariate Analysis**
- **Summary information**
- **All sorts of graphics**

Demo

Exploratory Data Analysis with Power BI

Feature Engineering

- Reducing Data Complexity through Binning
- Numerical Encoding
- Transformation data to a common scale
- Handling Missing data
- Determine how you are going to process outliers
- Balancing Data

Algorithm Selection

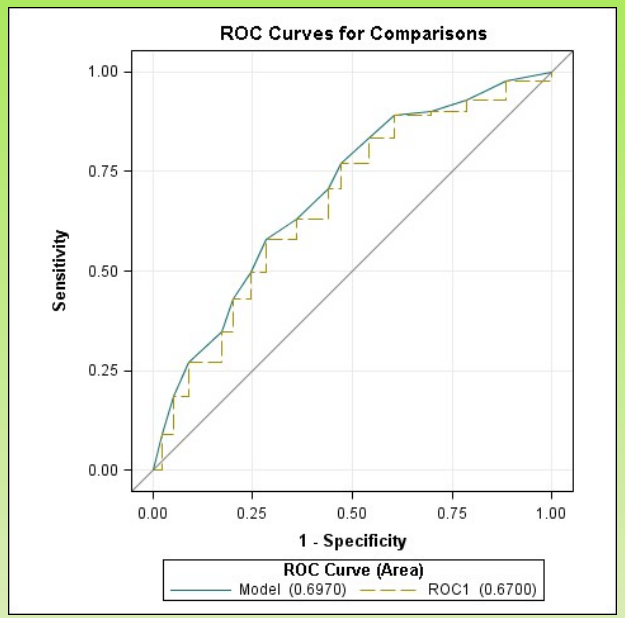
- Determine type of problem you are going to solve
- Look at algorithms which are used
- Evaluate a few

Data Splitting

- Not going to process all of the data at once
- Train the data
- Test against the control group
- Statistically significant Random sampling

Analyzing the Results

- Review visual models
- Compare to actual values
- Qualitative Review of Results
  - ROC Curves
  - Confusion Metrics



Evaluate the Model

- Probability of likely solution
- May not be able to determine anything
- Document the process through graphs

Deploy the model

- Remove test and training
- Compile solution
- Determine where it will be deployed

Demo

Azure Machine Learning

Classification Algorithms

Is there a relationship here?

- **Can my data show if X is an option?**
- **Facebook's "People You May Know"**
- **Are you a Gryffindor or a Slytherin?**

Anomaly Detection

Is there something wrong?

- Did this person cheat on a test?
- Is this a valid credit card transaction?
- Does this value mean the valve is going to blow?

Forecasting Algorithms

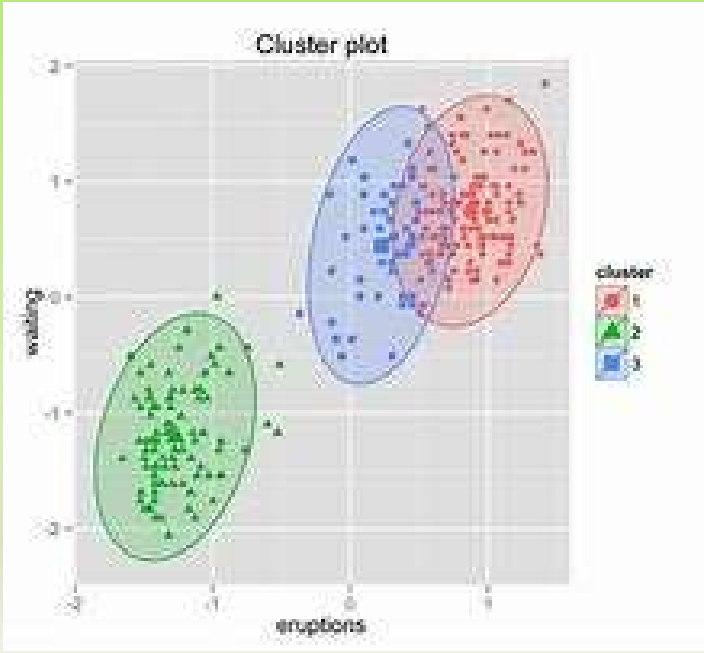
Predict what will happen

- What will sales be next quarter?
- Who is going to win the Super Bowl?
- If the weather predicts a blizzard, what will stores sell?

Clustering

How is this organized?

- Which things are bought together?
- Which customers are likely to buy more?
- What advertising needs to contain to motivate people to buy?
- How are these things related?



Reinforcement Learning

What should I do next?

- Algorithms act like Skinner Boxes
- Scheduling decision making
- Maintenance processing
- Solving a Rubix Cube
- Self driving Cars

Targeted Questions for Data Science

- Need to have the data available to provide an answer
  - Will this pump fail?
    - Need to have detailed information on performance
  - Did this person cheat on a test?
    - Need to know about questions answered correctly and incorrectly based on time and location compared to others

Gambling: Applied Data Science

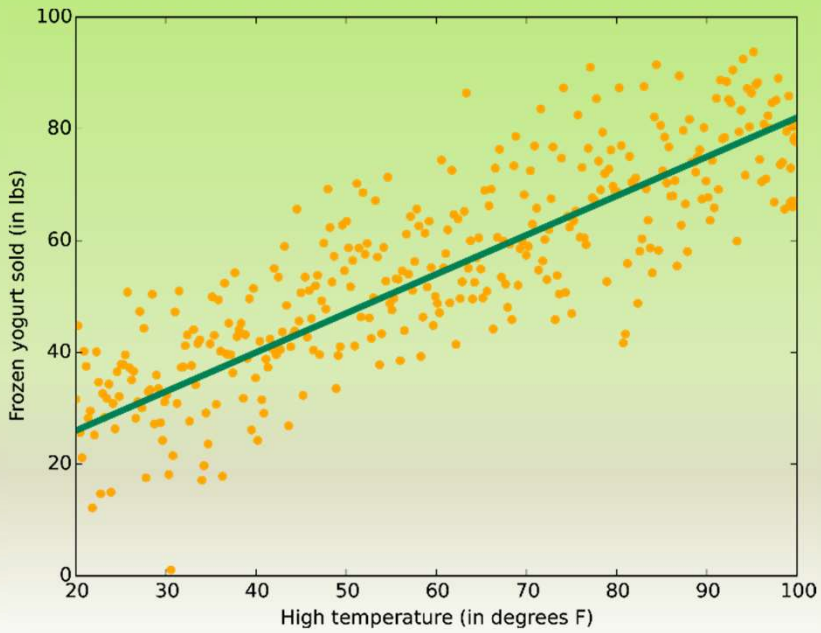
- Probability
- Correlation
- Significance

Terms

- **Linear Regression**
- **Forecasting**
- **Correlations**
- **Decision Forests**
- **Discretization**
- **Naïve Bayes**

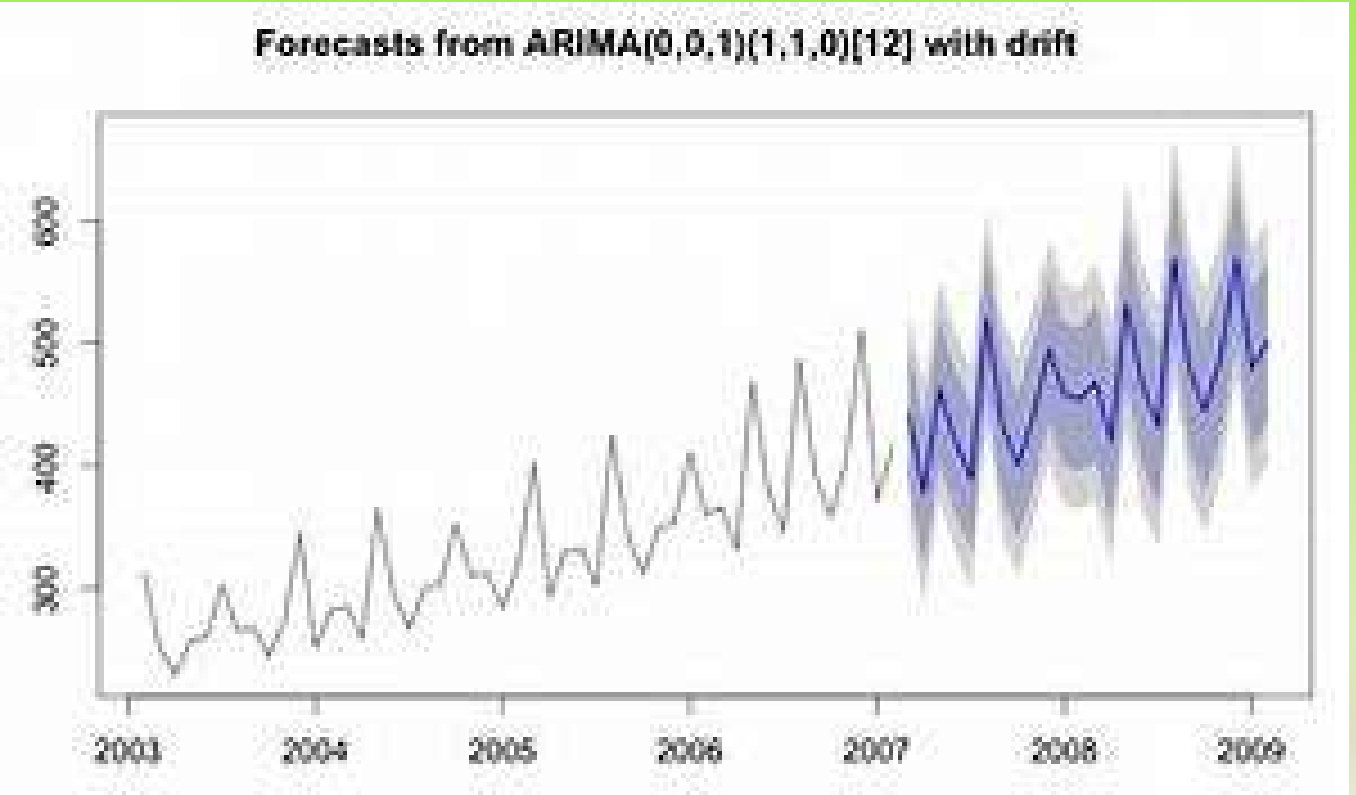
Linear Regression

- **Determine the relationship between two or more objects**
- **Look for patterns to see if things move in parallel**



Types of Forecasting

- Seasonal
- Clustering
- Average



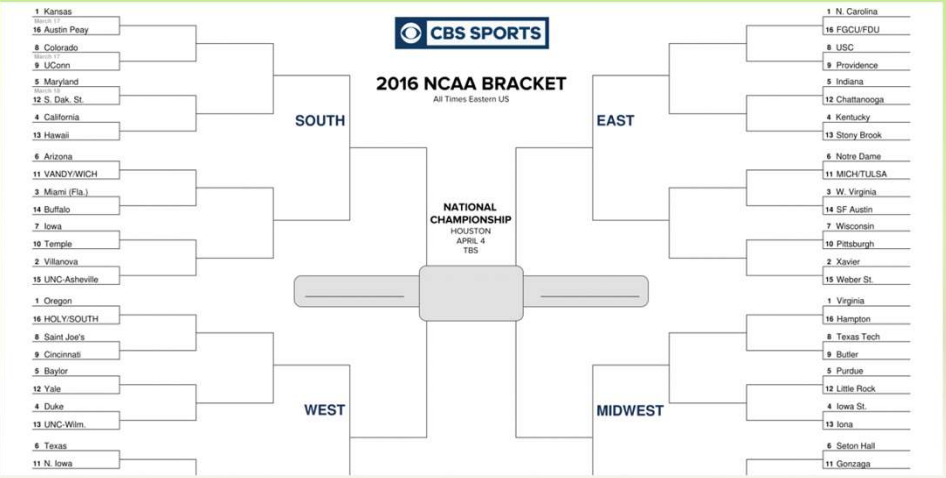
Correlation

- One of These Things is Not Like the Other
- Determine how things relate
- Used in a lot of different kinds of analysis

Decision Forests

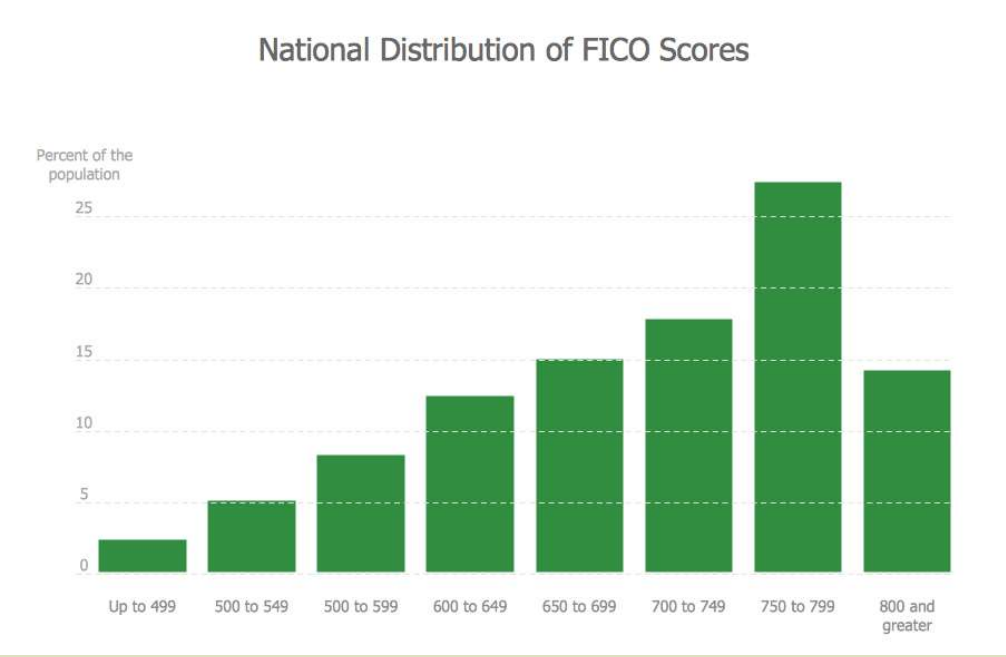
Series of decisions (or trees) that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Used to display an algorithm.

- **NCAA Brackets**



Discretization

- **Binning**
- **Grouping the values into equal groups**
- **Commonly displayed as a Histogram**



Naïve Bayes

- **Classify independent Variables and independently determine if they exist**
- **Features are not always independent, which is why it is Naïve**
- **Determine Fruit Types with it**

Which One is Faster?

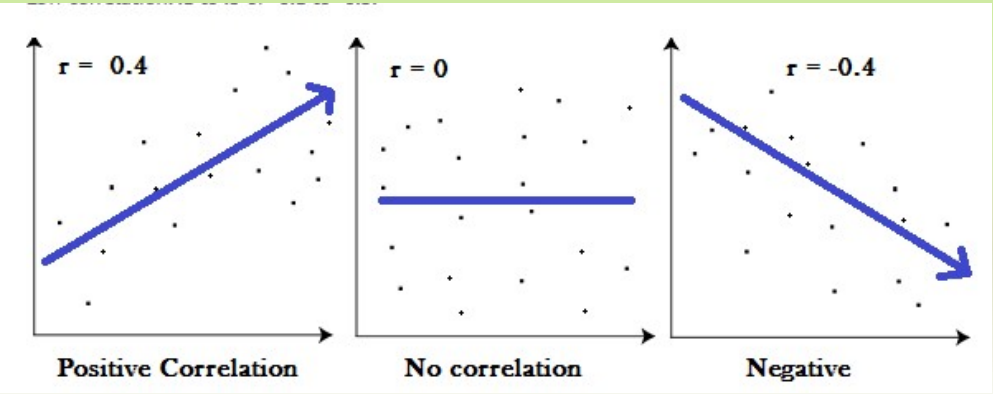
Common R Analysis:
Linear Regression and Correlation

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

- Value of r May Mean The Math was a waste
- Sometimes not looking for an exact number but looking for trends
- Important to know the answer for lots of data

Determining Linear Regression

- **Commonly determined by Scatterplots**
- **Look for a trend of points indicating relationships**
- **Hot weather and electric**

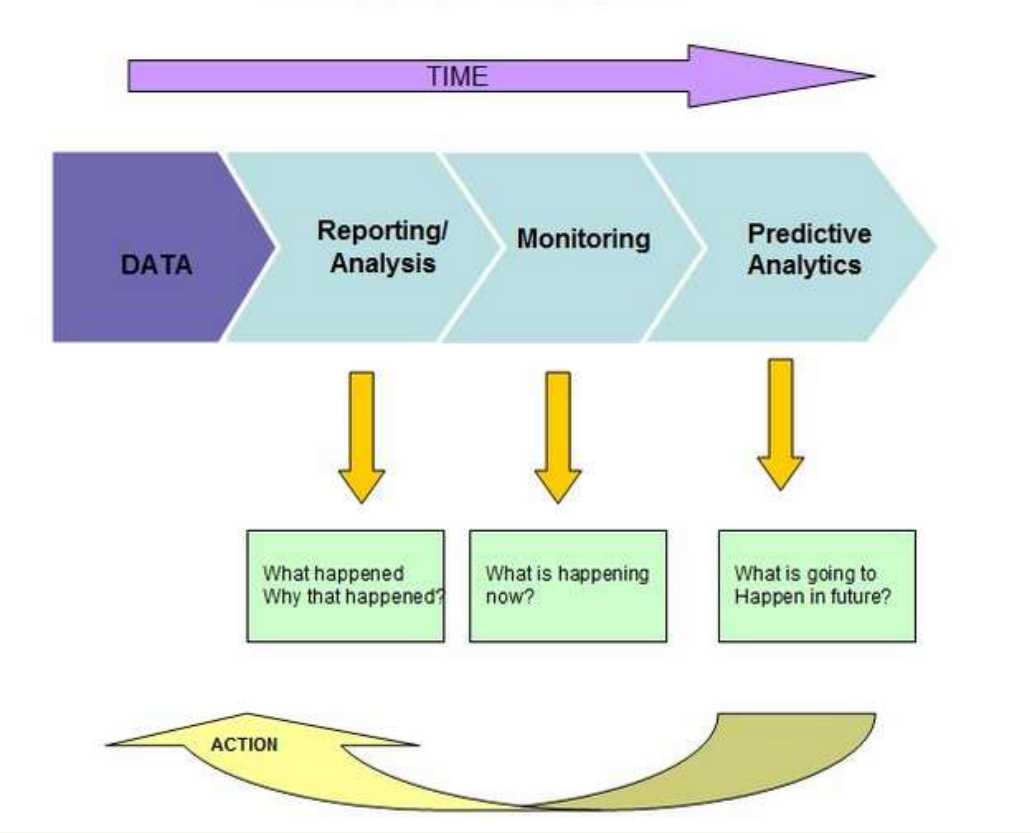


Demo

Linear Regression

Predictive Modeling

- Probability
- Supervised Segmentation
- Correlation
- Significance



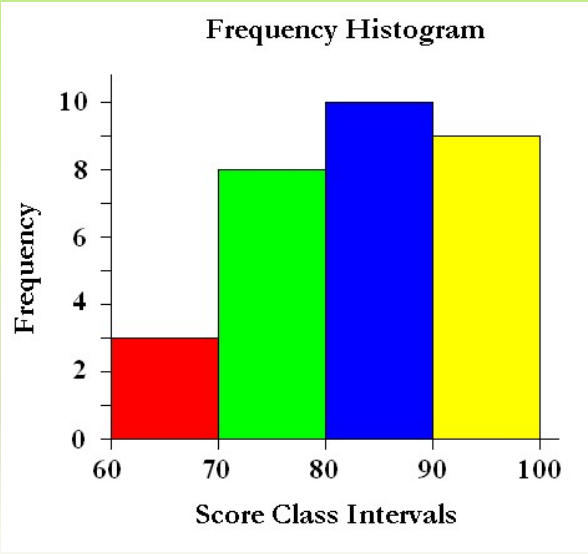
Source: predictiveanalyticstoday.com

Predictive Modeling

- Probability
- Correlation
- Significance

Supervised Segmentation

- Segmenting into subgroups with like attributes
- Attributes rarely split up the group perfectly
- Not all attributes are binary
- Binning numbers



Inductive Reasoning

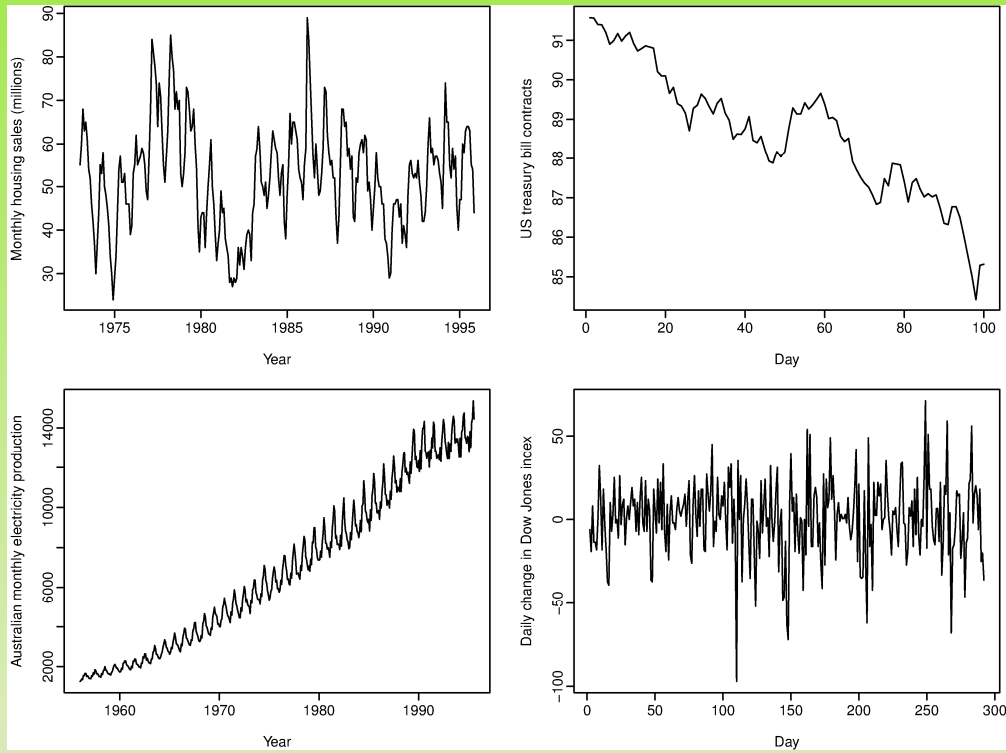
- Generalizing from specific cases to find general rules
- Models are general rules
- Procedure which creates the model is the induction algorithm or learner.
- Need models for classification and regression
- Target variable is the label

Algorithms for Predictive Modeling

- Time Series
- Regression
- Association
- Clustering
- Decision Trees

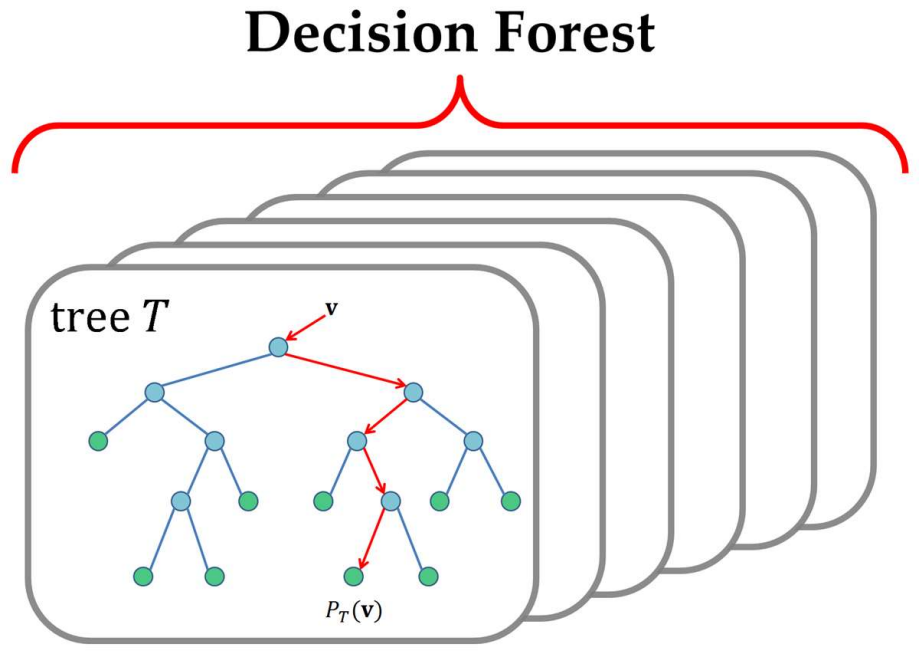
Time Series

- **Trend**
- **Seasonal**
- **Cyclic**



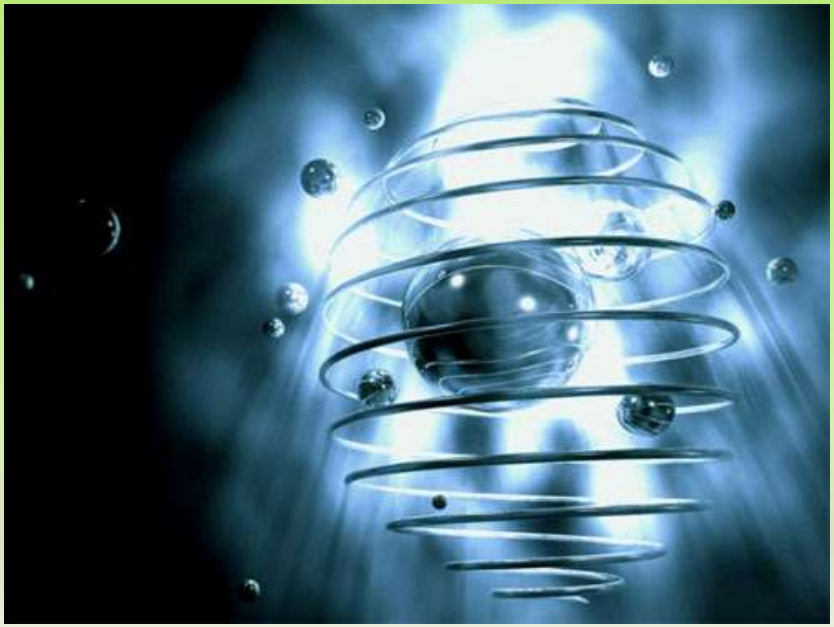
Visualizing Segmentations NCAA charts

- **Trees as sets of Rules**
- **Probability estimation trees**



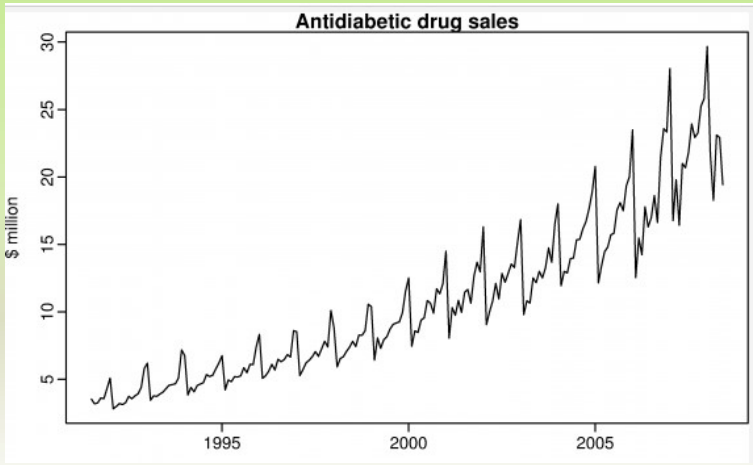
Forecasting: What Kinds of Things

- Well understood attributes and the weight of their contributions to the outcome
- Quantity of Available Data
- Whether the forecasts can affect the thing we are trying to forecast



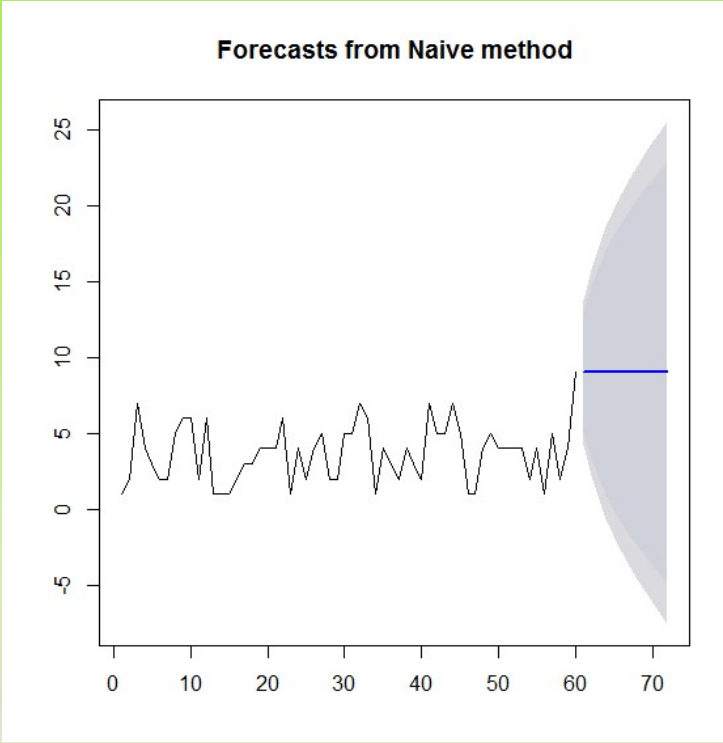
Forecasting

- **Quantitative – Past Patterns Provide Patterns to Future response**
- **Time Series – Observations Over Time**
- **Judgement – Predicting the Unpredictable**



Applied Forecasting

- **Average Method**
- **Naive Method**
- **Seasonal Naive Method**
- **Drift Method**



Evaluating Forecasts

- Mean Error
- Root Mean Squared Error
- Mean Absolute Error
- Mean Percentage Error
- Mean Absolute Percentage Error
- Mean Absolute Scaled Error
- Autocorrelation of errors at lag 1
- Theil's U

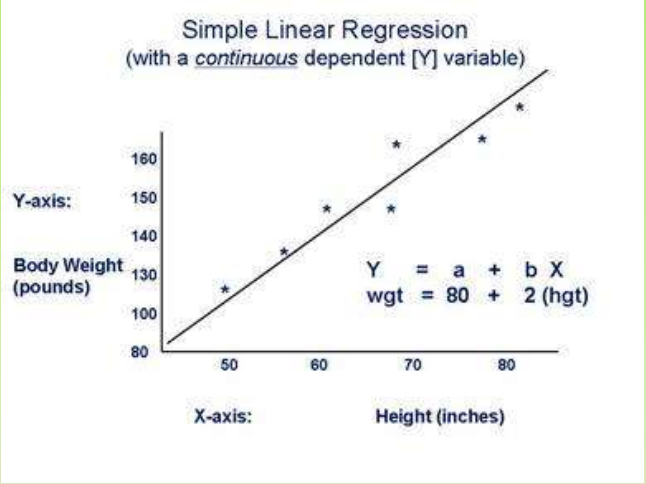
R Demo on Forecasting

Applying the Concepts

- While a lot of the analysis involves visualization, it is not required
- Implementation is based on numerical references
- Returns Decision Points

Linear Regression with No Pictures

- **Make a decision with data**
- **Process Based on results**
- **Return Numerical references**



Pearson Correlation Coefficient

- **Results are between -1 and 1**
- **Measures the dependence between two variables**
- **Significant dependence would be absolute value of .5 to 1**

$$\rho_{X,Y} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2] - \mathrm{E}[X]^2}\,\sqrt{\mathrm{E}[Y^2] - \mathrm{E}[Y]^2}}.$$

Demo

Python Machine Learning Solution

Python and SQL Server

Python and SQL Server

- **Performance improvement negligible**
- **Enhanced libraries available**
- **Compatible with existing code**

revoscalepy

Based on the **RevoScaleR** package for R

Supports Remote and Local compute

Contains optimized Specific Functions

RxLocalSeq

RxInSqlServer

RxFileData

RxXdfData

RxOdbcData

python

Machine Learning

Includes Functions Designed for Advanced Anal

rx_btrees_ex

rx_dforest_ex

Rx_dtree_ex

rx_lin_mod_ex

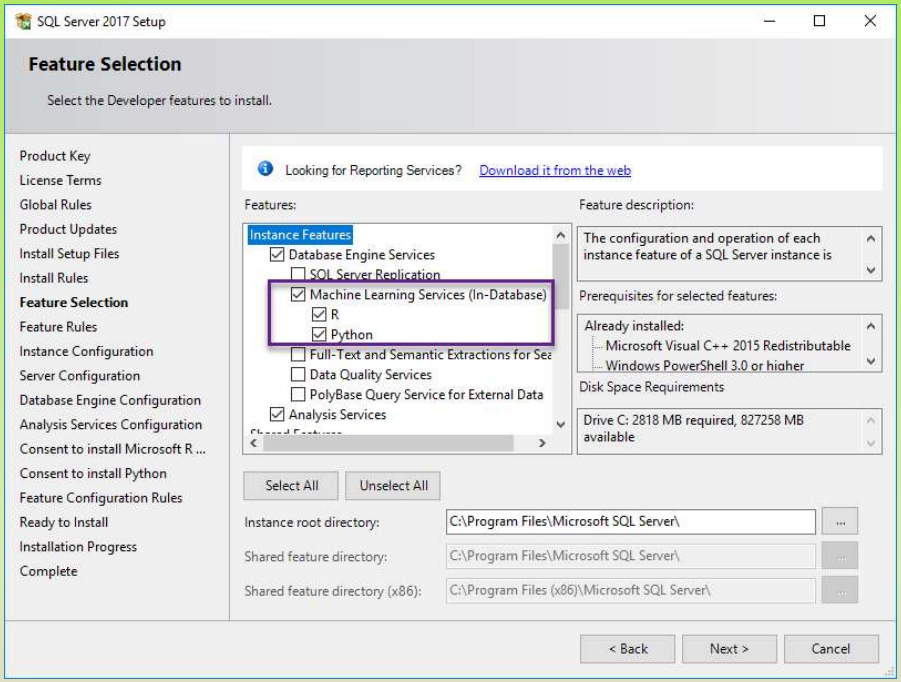
rx_logit_ex

rx_predict_ex

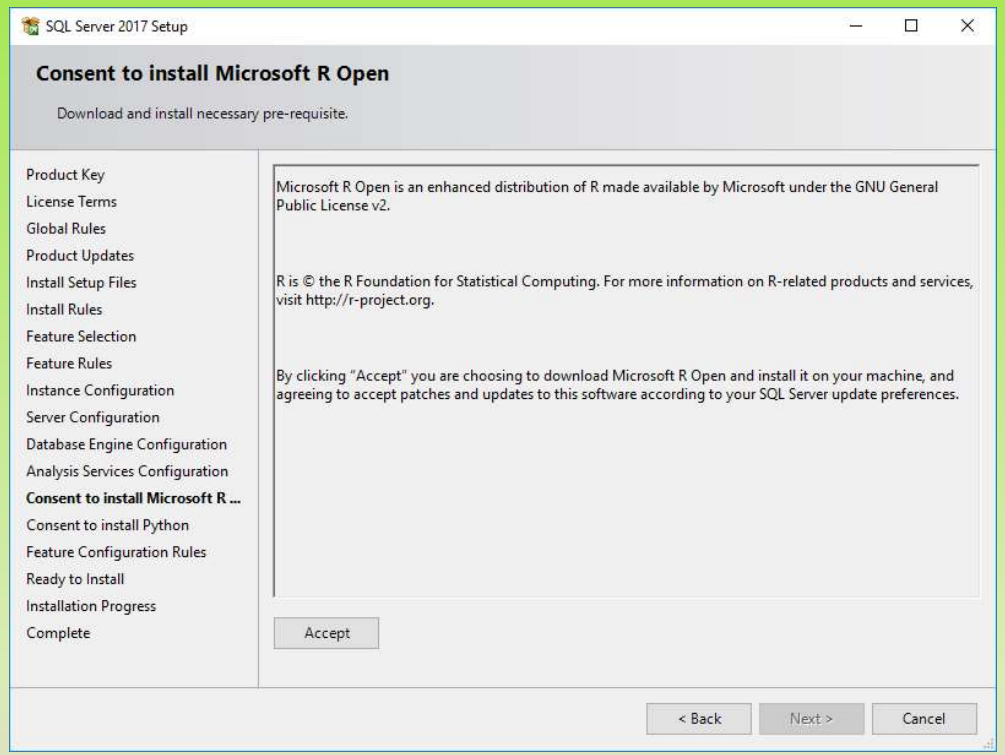
rx_summary

SQL Server Machine Learning Services

- **Replaced R Server**
- **CPython interpreter version 3.5**
- **Part of Anaconda packages are included**
- **Microsoft's RevoScalePy package**
- **Available in All Versions of SQL Server 2017**



SQL Server R Install

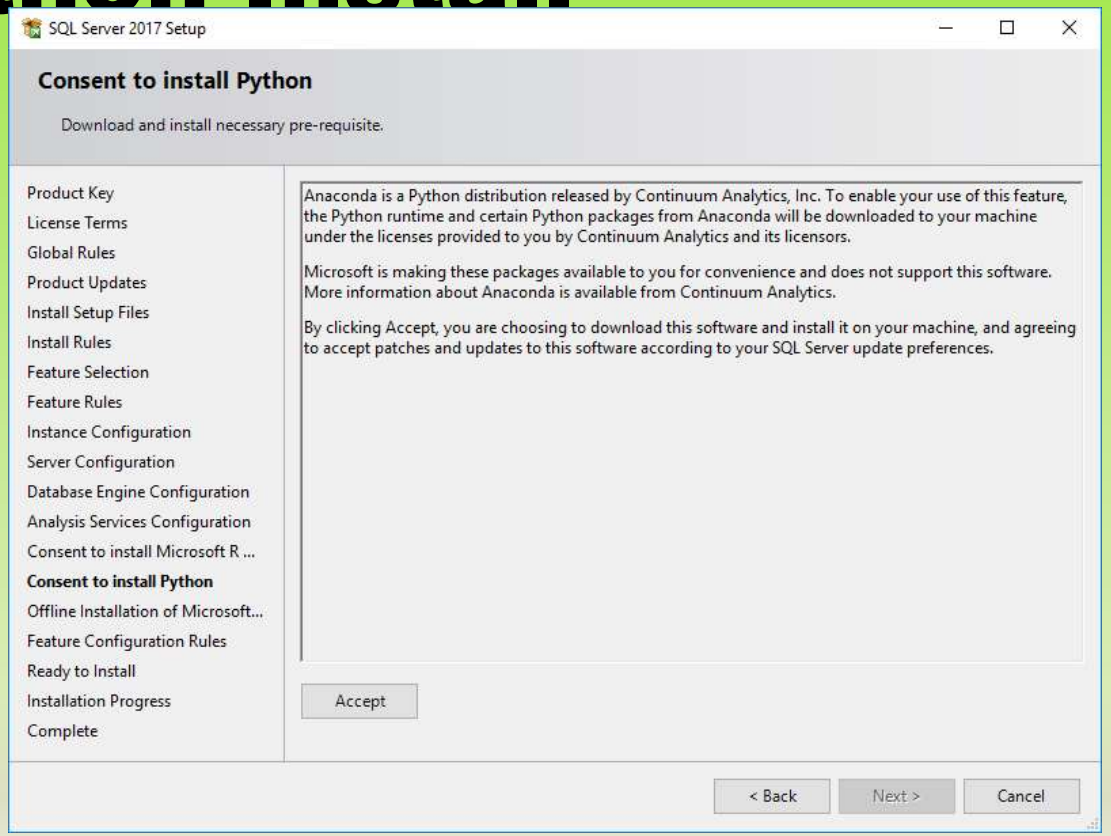


- **Goes to the internet to get R**
- **Installs Open Source R and Microsoft R Open**
- **Directory for all of the R Services**

```
<sql install dir>\Microsoft SQL Server\MSSQL14.<instance>\R_SERVICES
```

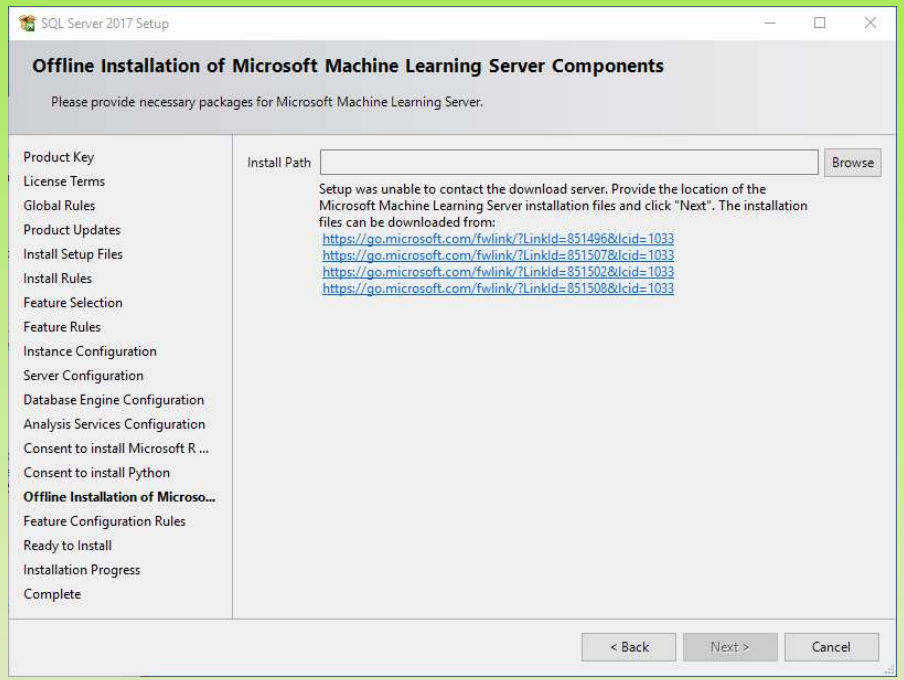
SQL Server Python Install

- **Internet is required**
- **Commonly used Data Science Library**
- **Other libraries may need to be added later**

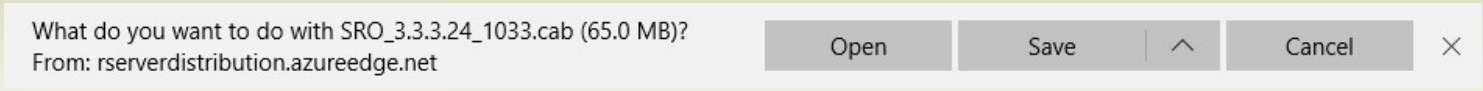


Machine Learning Install

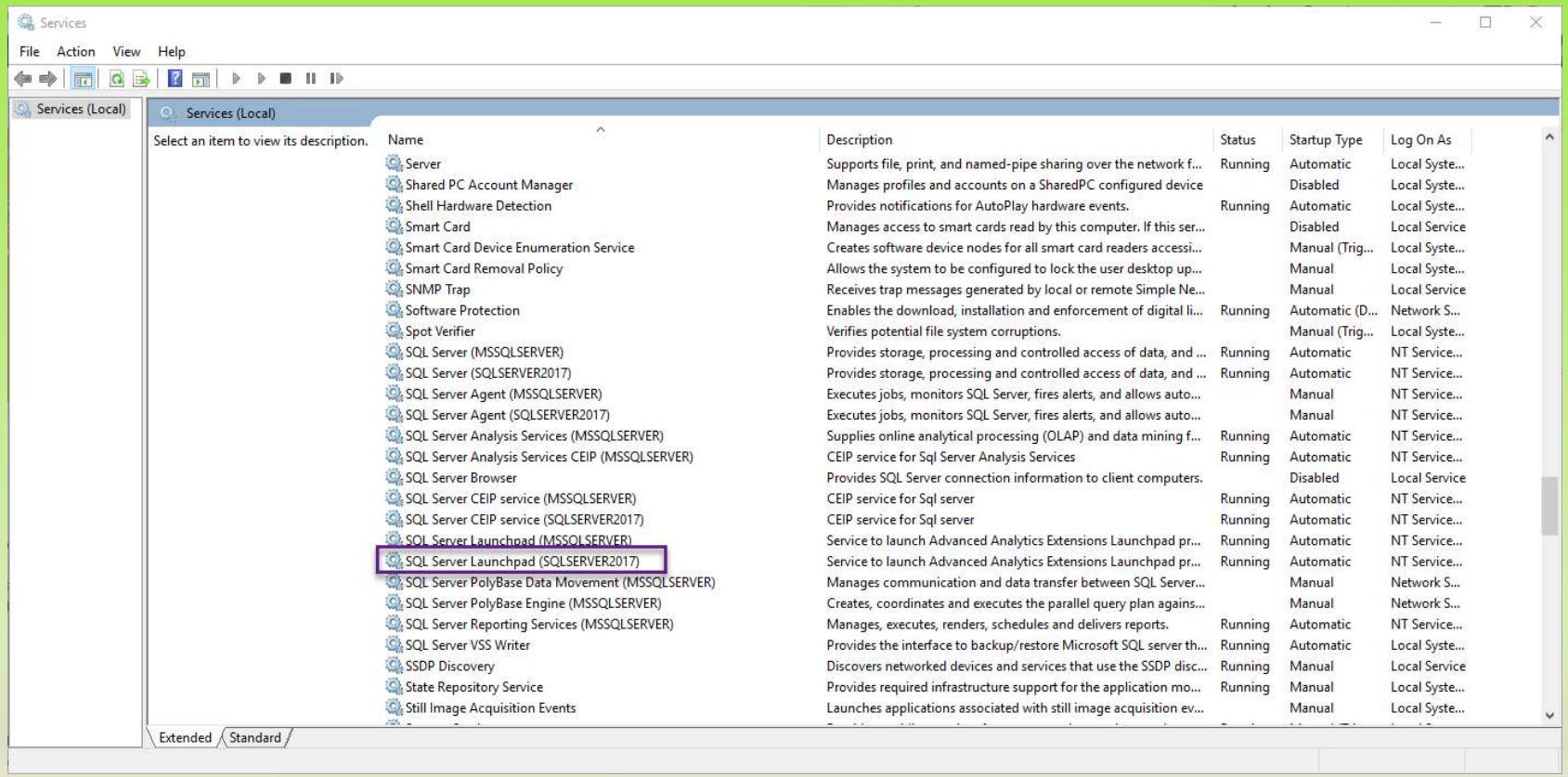
No Internet



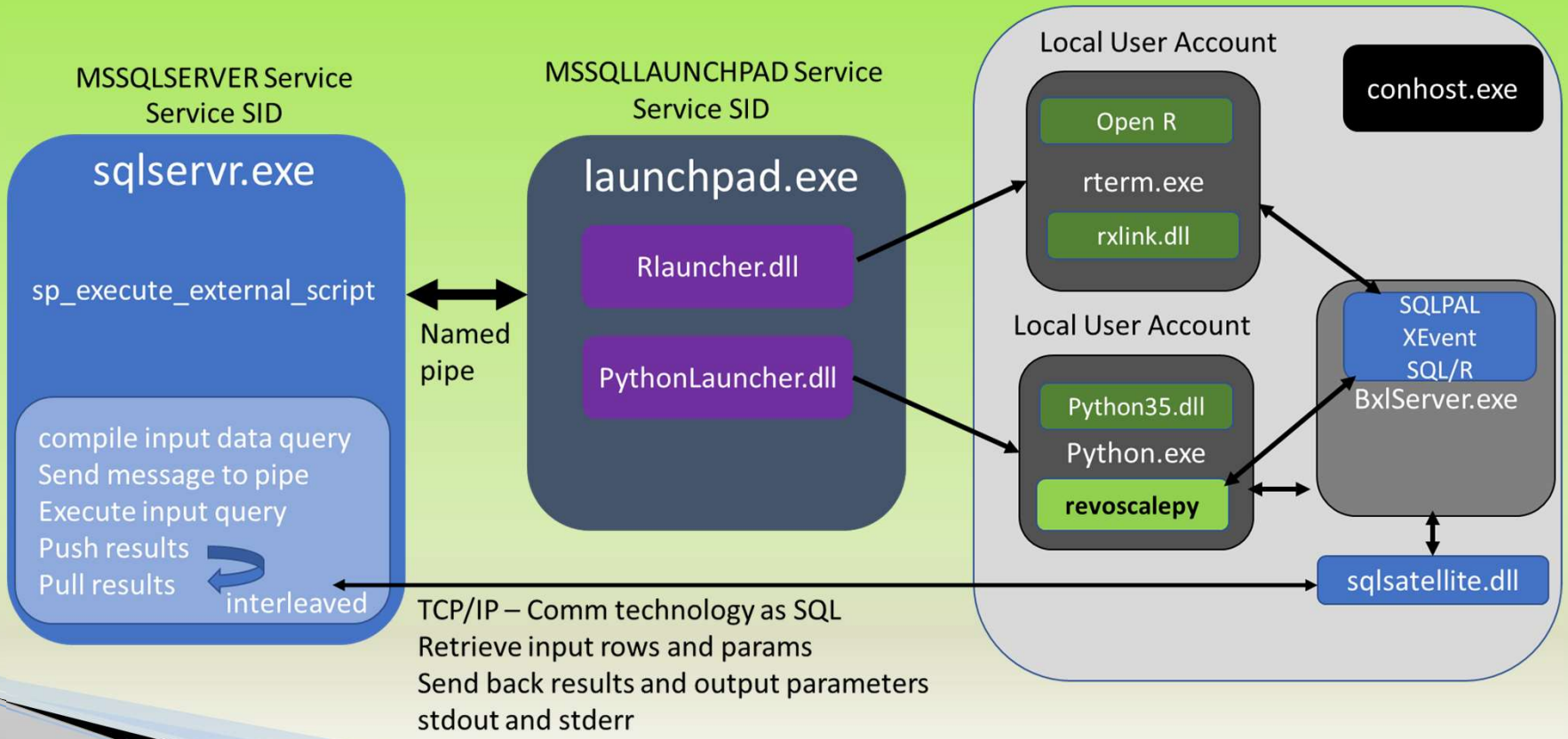
- **Detects there is no internet access**
- **Provides instructions for how to get the installs for Python and R**



Machine Learning Services



SQL Server ML Services Internals



SQL Server Install

```
sp_configure 'external scripts enabled', 1
reconfigure
GO

alter role db_rrerole add member UserName
```

Integrating Python in SQL Server Procs

```
DECLARE @ParamINT INT = 1234567
DECLARE @ParamCharN CHAR(6) = 'INPUT '
DECLARE @RowsPerRead INT = 5


EXEC sp_execute_external_script  @language =N'Python',
@script=N'
import sys
import os
if ParamINT == 1234567:
ParamINT = 1
else:
ParamINT += 1
ParamCharN="OUTPUT"
OutputDataSet = InputDataSet
global daysMap
```

- `daysMap = { "Monday" : 1, "Tuesday" : 2, "Wednesday" : 3, "Thursday" : 4, "Friday" : 5, "Saturday" : 6,  "Sunday" : 7 }`
- `OutputDataSet["DayOfWeek"]=pandas.Series( [daysMap[i] for i in OutputDataSet["DayOfWeek"]] )`
- `',`
- `@input_data_1 = N'SELECT top 10000 DayOfWeek, CRSDepTime  from AirlineDemoSmall',`
- `@params = N'@r_rowsPerRead INT, @ParamINT INT OUTPUT, @ParamCharN CHAR(6) OUTPUT',`
- `@r_rowsPerRead =@RowsPerRead,`
- `@paramINT =@ParamINT OUTPUT,`
- `@paramCharN = @ParamCharN OUTPUT`
- `with result sets (("DayOfWeek" int null, "ArrDelay" float null))`

Best Practices and Performance Considerations

- Balance memory needed by SQL Server and external pool
- Launchpad needs specific privileges
- Be sure SQLMLUserGroup has log on local rights
- Restart the SQL Server Service not stop/start (Launchpad is dependent)
- Remote ODBC execution requires SQLMLUserGroup login
- 20 unique users allowed to execute ML scripts concurrently by default
- SQL Server Query and Index design still apply

.

Managing, Monitoring, and Troubleshooting

- DMVs
  - dm_external_script_execution_stats – monitor ScaleR API executions
  - dm_external_script_requests – active script execution
- Perfmon counters: SQL Server:External Scripts

ML Services in other Programs

- XEvents
  - Events for SQL Server, Launchpad, and external processes
  - Configuration file needed for Launchpad and external processes

SQL Server External Resource Pools

internal, default, “user”, and now external

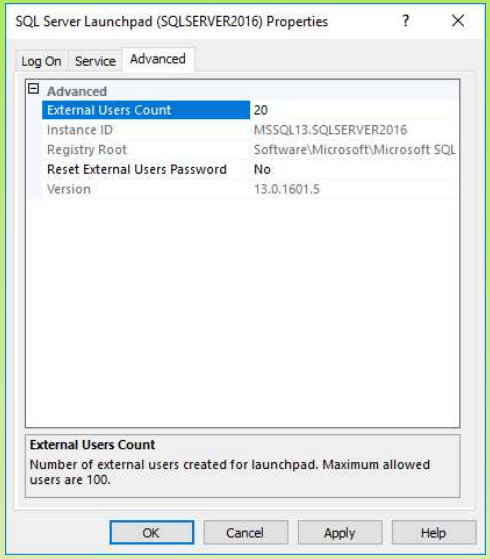
Controls resources for external processes through Launchpad.

Default external pool and user external pools. User classifier function supported

Review the settings for the following controls, as tweaking may be required

- MAX_CPU_PERCENT – Max CPU percentage for external processes
- MAX_PROCESSES – Max number of external processes
- MAX_MEMORY – Max committed memory % for external processes
- AFFINITY – Control NODEs or CPUs for external processes

SQL Server Config Settings



- Default Settings for are Low
- Create a Resource Pool
- Set the Memory use for External Resources

Demo - SQL Server ML Services Configuration

DevOps and Data Science

Using SQL Server to implement a solution

PASS SQLSATURDAY
SOUTH ISLAND NZ | 26 MAY 2018

Dev Ops Process

- Configure
- Code
- Build
- Test
- Package
- Release
- Monitor

Implementation Process

- **Database Lifecycle Management**
- **ETL Process Development**
- **Analytics Implementation**

Configuration of Analytics

- Integration of Teams who may not work together
- Data Sources used in Development and Production
- Library Management
- Results Distribution
- Planned Obsolescence

Tools to Assist the Process

- Source Control
- Extended Events for Reviewing Libraries
- Performance Monitoring

Summary

- Practical Data Science can be achieved by applying some concepts with the right datasets and some R
- Data can be analyzed quickly by reviewing visualizations of the results.
- There are a number of different ways of analyzing data and there are a lot of samples to show how

Questions?

www.desertislesql.com