



Visual Studio **LIVE!** | San Diego
EXPERT SOLUTIONS FOR .NET DEVELOPERS

Predicting the Future Using Azure Machine Learning

Eric D. Boyd
Founder & CEO
responsiveX

Level: Intro-Intermediate

Code Again for
the First Time!

Visual Studio **25**
YEARS OF CODING INNOVATION

About Me



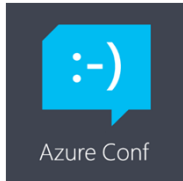
eric.boyd@responsiveX.com

responsiveX

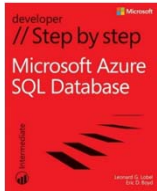


I...

speak at lots of conferences



write occasionally



try to be social on twitter

@EricDBoyd



Agenda

- Data Science and Machine Learning
- Azure Machine Learning
- Let's Model and Predict
- Questions



The Analytics Spectrum



The Value of Data

Companies that use data and business analytics to guide decision making are more productive and experience higher returns on equity than competitors that don't.

- Brad Brown, McKinsey Global Institute

By 2015, organizations integrating high-value, diverse new information types and sources into a coherent information management infrastructure will outperform their industry peers financially by more than 20%.

- Regina Casonato, Gartner



The Size of Data

35,000,000,000,000,000,000,000,000

According to IDC, the digital universe will grow to 35 zetabytes (35 trillion terabytes) globally by 2020.

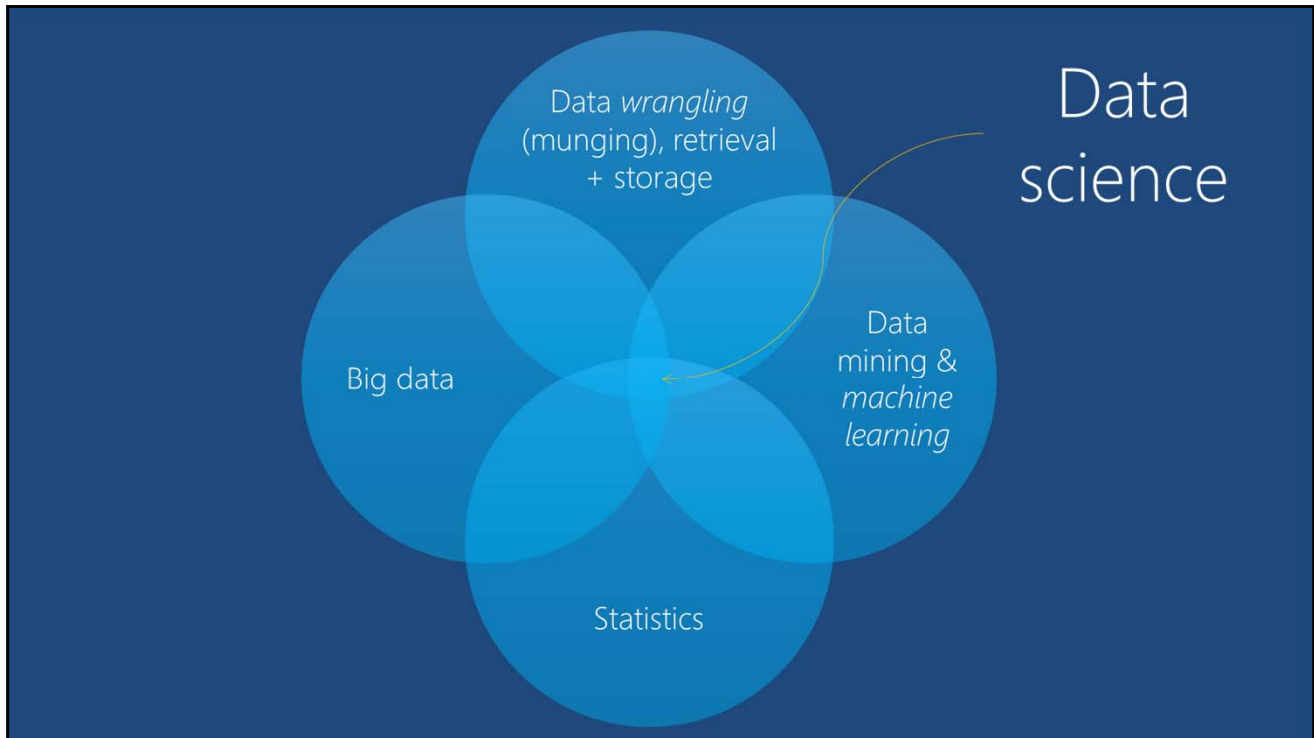


Why is Big Data Valuable?

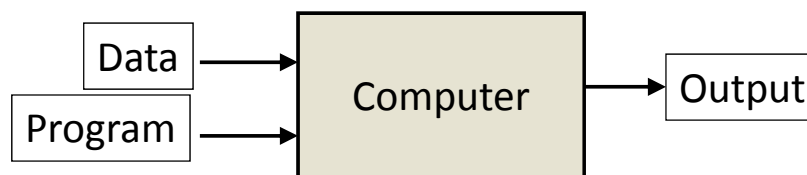
Statistically, big data is important because of the confidence interval (margin of error)

The Confidence Interval has an inverse relationship with the sample size

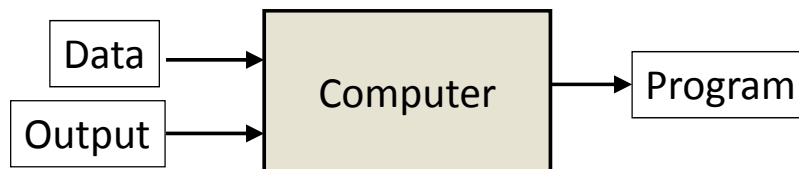




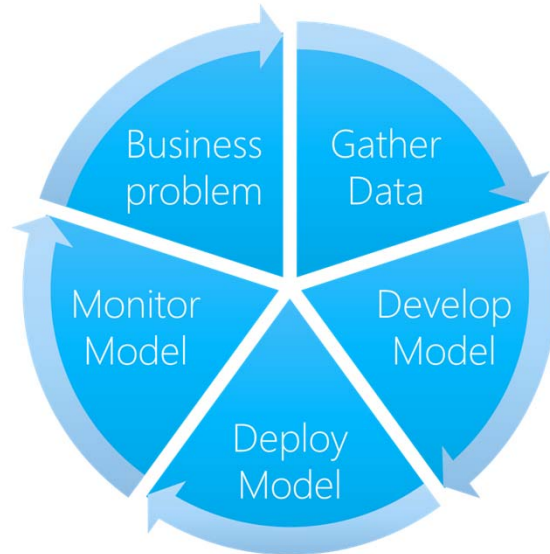
Traditional Programming



Machine Learning



How to do data science?

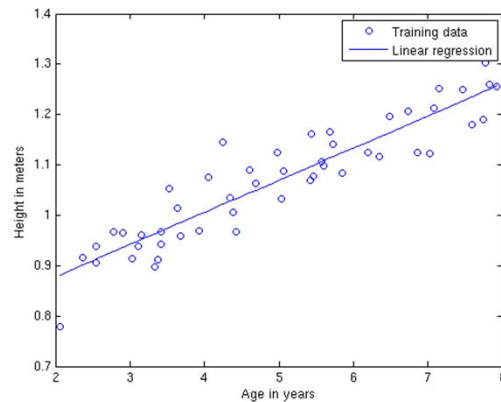


Types of Algorithms

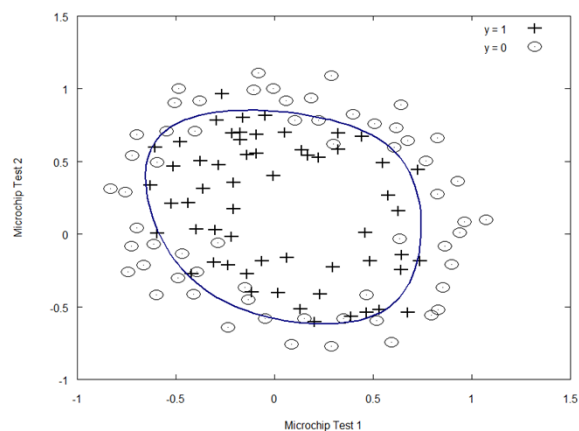
Classification	Categorize data
Clustering	Discover natural groupings and patterns
Regression	Predict numerical values
Recommendations	Suggestions from associations
Ensembles	Multiple models and algorithms together



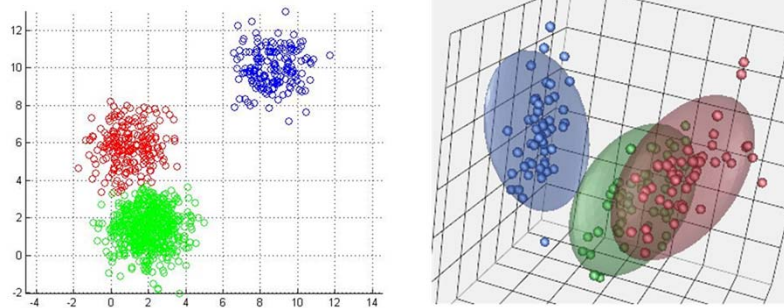
Regression



Classification



Clustering



Ensemble Models



Ensemble Strategies

- Bagging
 - different subsets of the data to train each model
- Boosting
 - Additional models focus on misclassified data



Do we train?

- Supervised learning
 - Trained
 - Model created from known datasets
 - Examples include classification and regression
- Unsupervised learning
 - Untrained
 - Model created by self-analyzing data and inferring patterns and structure
 - Examples include clustering



Supervised Learning Dataset

Inputs
a.k.a. Features or Vectors

Outputs
a.k.a. Labels or Supervisory Signal

Age	Occupation	Education	Income
30	Sales	Bachelors	\$75,000
40	Physician	Doctorate	\$190,000



Azure Machine Learning Studio



Grouping of machine learning assets

Build models with experiments

Experiments published as web services

IPython Jupyter Notebook

Uploaded data for the experiments

Experiments that have been trained and converted to models

Where you add users and auth tokens



Creating and Training the Model

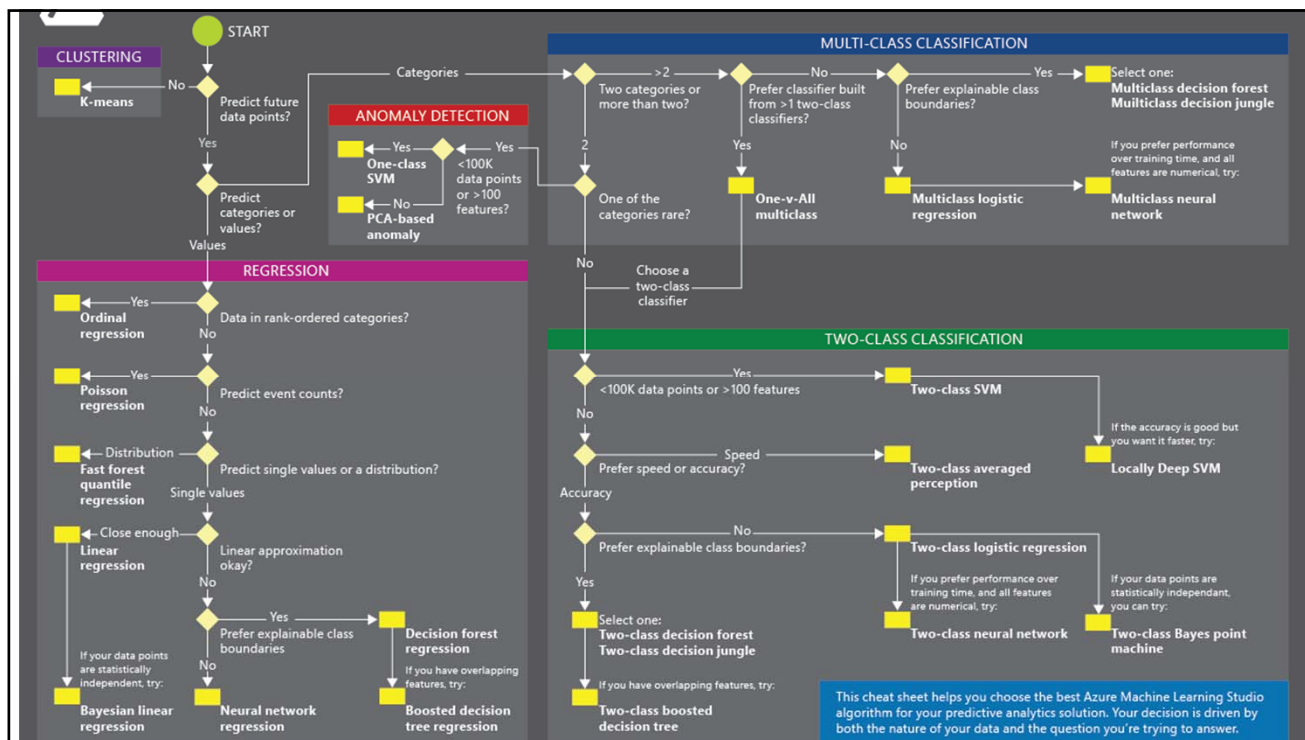
- Clean Data
- Select Columns
- Split Data for training
- Choose an Algorithm



Model Evaluation

- **Mean Absolute Error (MAE):** The average of absolute errors (an *error* is the difference between the predicted value and the actual value).
- **Root Mean Squared Error (RMSE):** The square root of the average of squared errors of predictions made on the test dataset.
- **Relative Absolute Error:** The average of absolute errors relative to the absolute difference between actual values and the average of all actual values.
- **Relative Squared Error:** The average of squared errors relative to the squared difference between the actual values and the average of all actual values.
- **Coefficient of Determination:** Also known as the **R squared value**, this is a statistical metric indicating how well a model fits the data.





Resources

- Azure Machine Learning eBook
 - <https://mva.microsoft.com/ebooks>
- <https://azure.microsoft.com/en-us/documentation/services/machine-learning/>
- <https://azure.microsoft.com/en-us/documentation/articles/machine-learning-create-experiment/>

Q&A



eric.boyd@responsiveX.com

responsiveX

