

TechEd

India 2014

Learn. Connect. Explore.



How to Become a Data Scientist with Azure Machine Learning

Saket Suman & Vikas Goyal
Microsoft Consulting Services

Data is the New Oil..



..And fortunately we have a plenty of it 😊

Data Scientists unearth the power of Data...



Data Scientist at work

..and we need many more of them 😞

Agenda (L200)

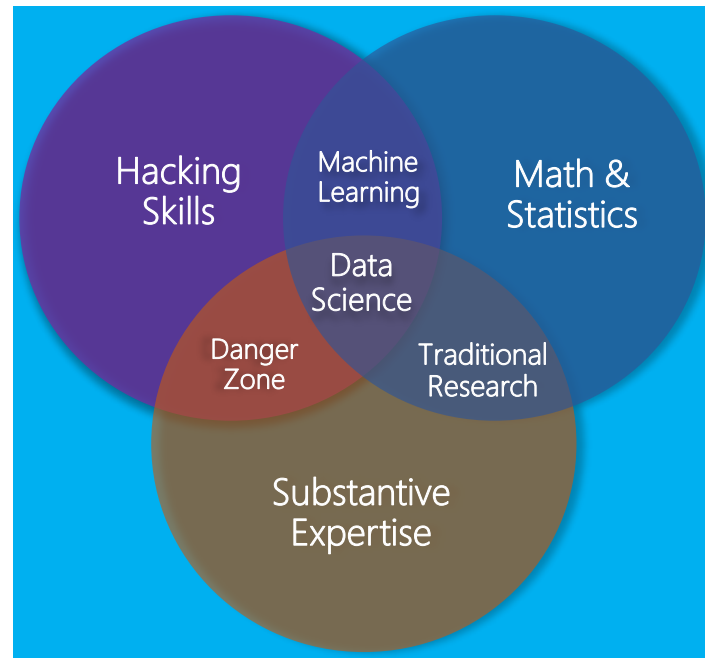
- Data Science & Data Scientists
- Machine Learning
- Azure Machine Learning (AML)
- Demos
- Q&A

Out of Scope

- R Language
- Model Details
- Model comparisons
- *Boring Talk & Demos*

What is Data Science

Machine Learning
Big Data R EDW
Data Mining KDD
Statistics Data Lake F#
IoT Data Hacking
Power BI



Predictive Analysis
Classification
Sentiment Analysis
A/B Testing Significance
Hypothesis Testing
Recommender
Data Exploration

Typical Backgrounds of Data Scientists

Mathematicians/
Statisticians

Programmers/
Hackers

Data
Practitioners

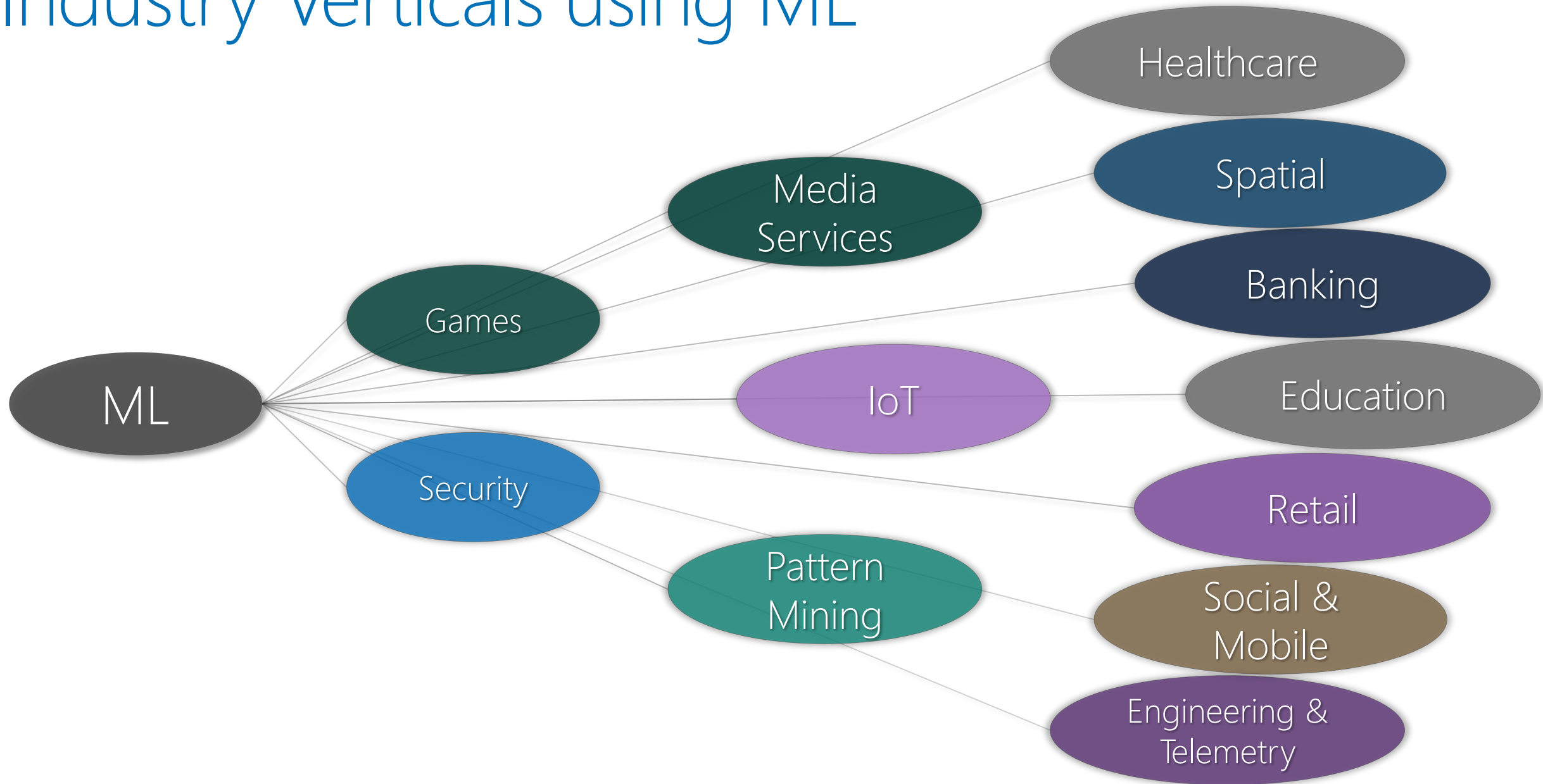
What is Machine Learning

"A computer program is said to *learn* from experience **E** with respect to some task **T** and some performance measure **P**, if its performance on **T**, as measured by **P**, improves with experience **E**."

Tom Mitchell (1998)

<https://www.coursera.org>

Industry Verticals using ML



Machine Learning in a box

Predictive Analytics (Supervised)

- Regression
- Classification
- Time Series

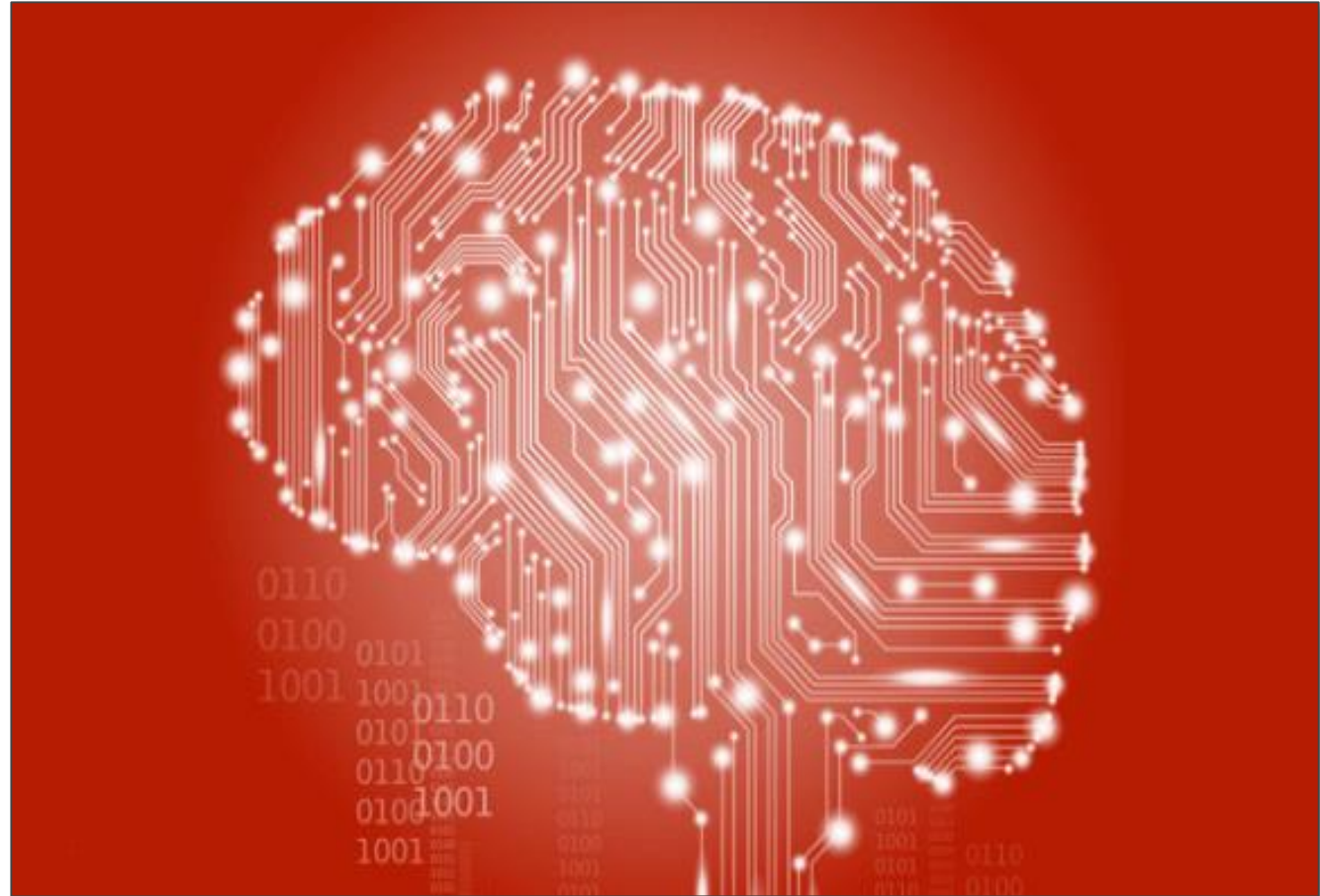
Unsupervised Learning

- Clustering
- Association
- Hidden Markov

*ML is the AI, Math & Science behind KDD

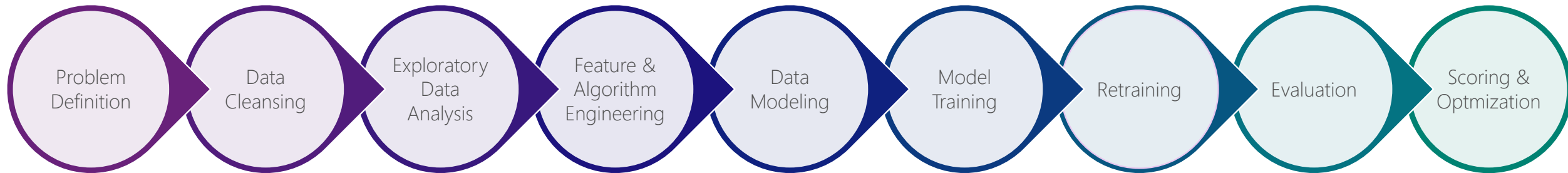
Other Methods

- Recommender Systems
- Reinforcement Learning



How do we enable Machine Learning

LEARNING = REPRESENTATION + EVALUATION + OPTIMIZATION



Yes! Machines do learn...with Statistics & Probability

A Learned Machine auto-programs itself

ML is NOT a 100% precise science

Applications of Machine Learning

Imagine what machine learning could do for your business.

Churn analysis

Market Basket Analysis

Image detection

Equipment Monitoring

Recommendation Engines

Forecasting with Trending

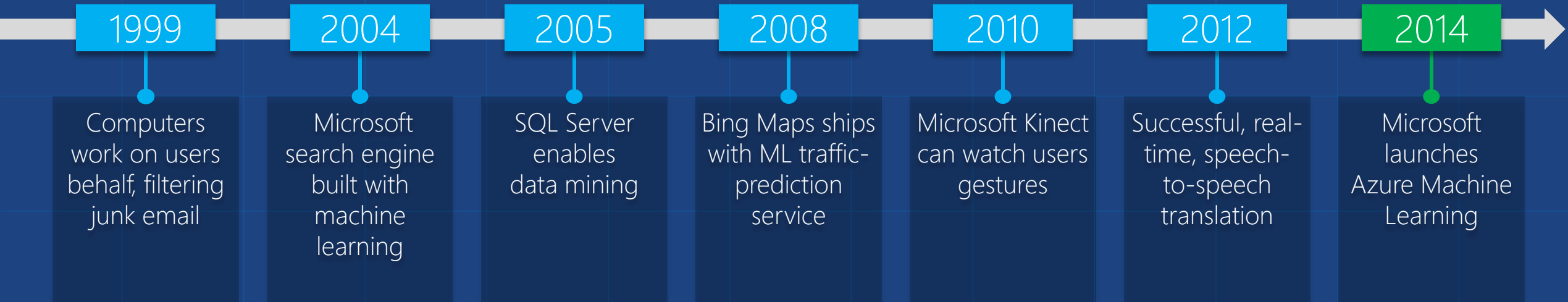
Spam filtering

Disease Outbreak Prediction

Anomaly detection

Microsoft & Machine Learning

15 years of realizing innovation



John Platt,
Distinguished scientist at
Microsoft Research

“Machine learning is pervasive throughout Microsoft products.”

Microsoft's ML Investments

Azure Machine Learning Studio

SaaS on Azure

Microsoft Bing Predictions

Search Engine

Microsoft Social Listening

Sentiment Analysis

XBOX

Intelligent Gaming

Project Sage & Basket

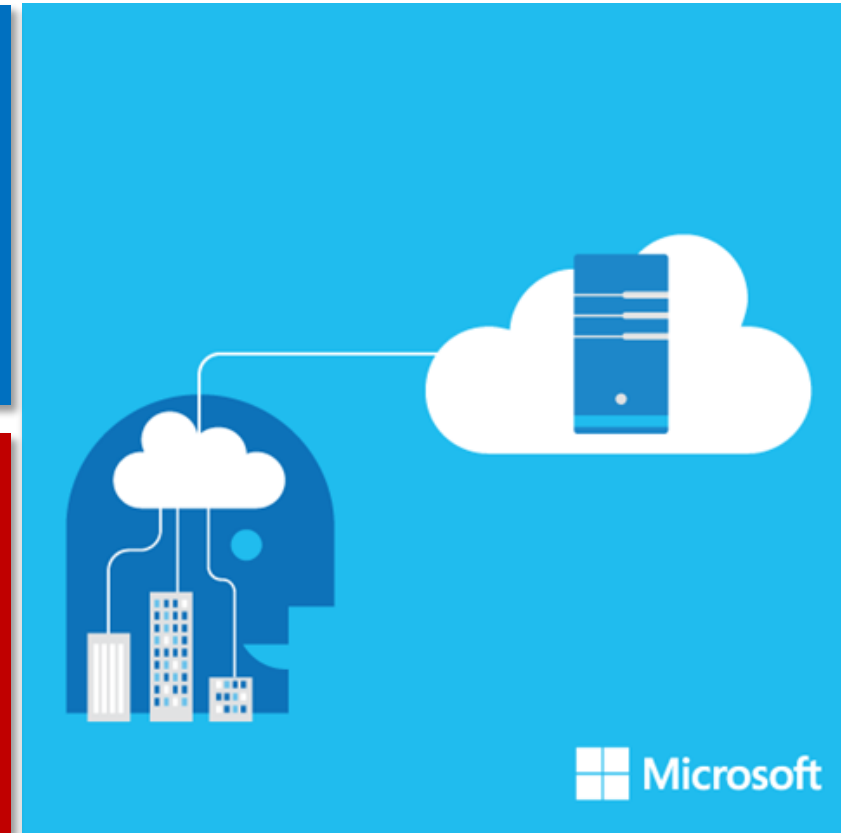
SaaS on Azure

Cortana

Phone Assistant

Power BI Forecasting

Prescriptive Analytics



What is Azure Machine Learning

Pay-per-use Predictive Analytics in Microsoft Azure

Data Scientist Toolkit

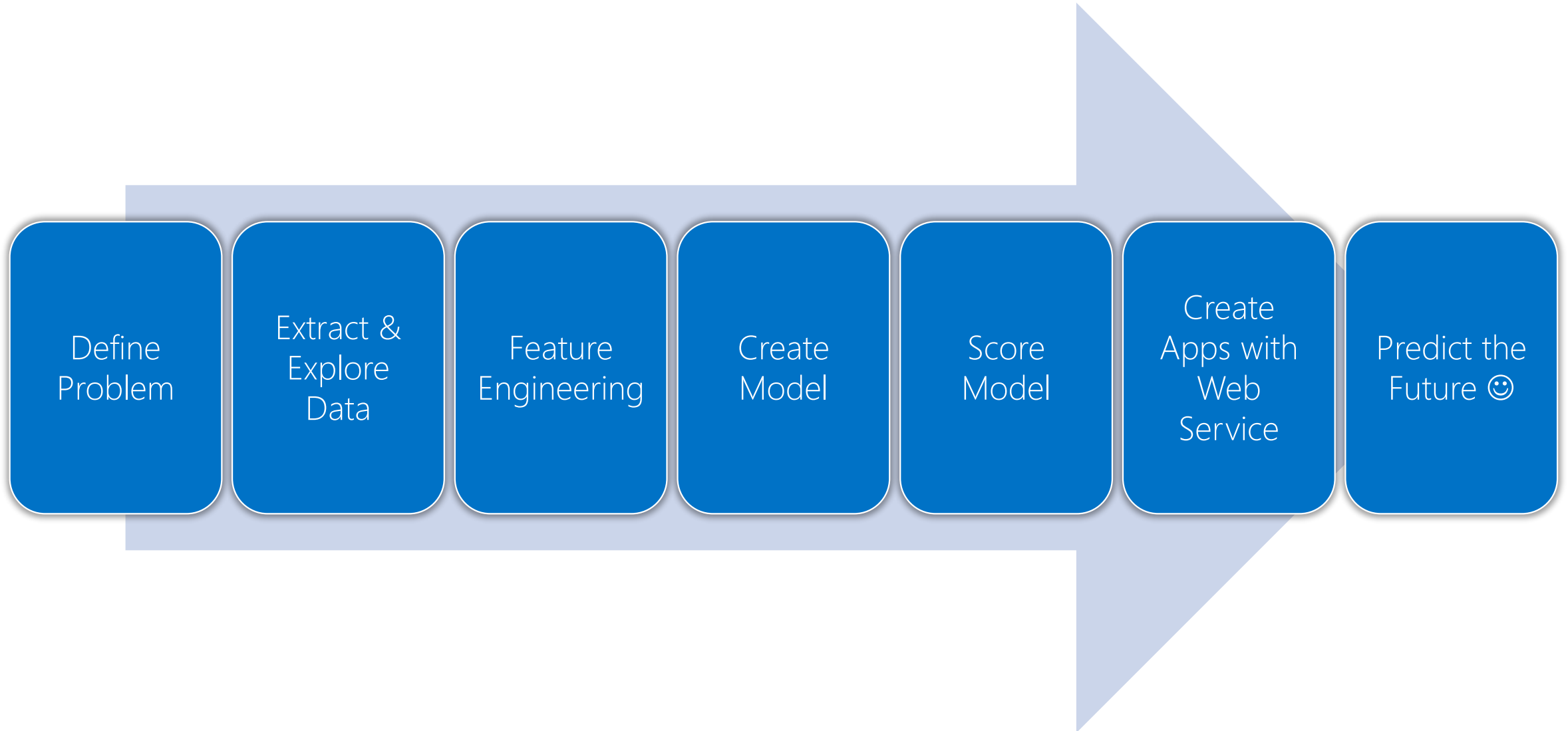
Deploy Predictive models to Production in minutes

Designed for Budding & Experienced Data Scientists

Industry Standard Algorithms & Support for R Language

Connectivity to HDI for Hadoop

How does Azure ML make Data Science easy



How it all works

The Environments

Azure Portal

ML Studio

ML API service

Visual Studio



The Team

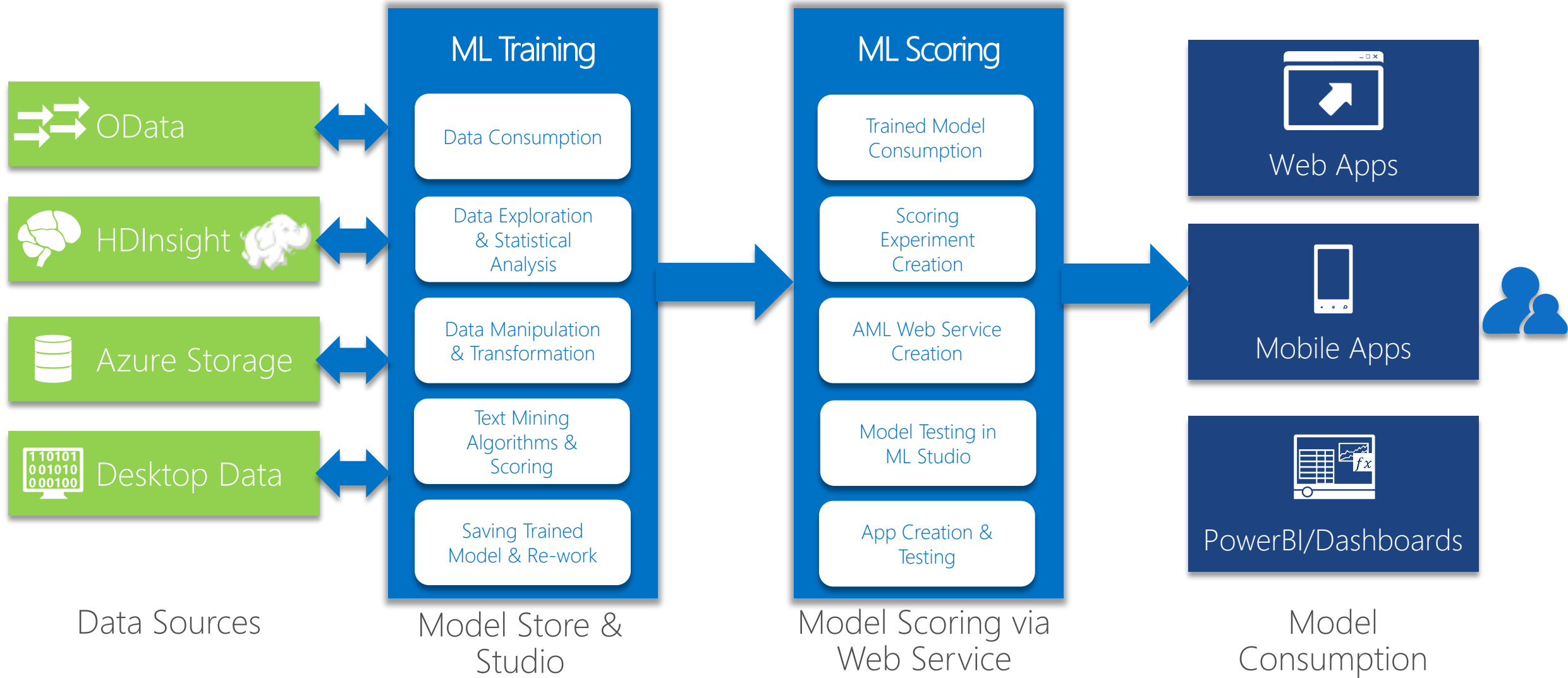
Azure Ops Team

Data Scientists

Developers

BI Users & Roles

Microsoft AML Architecture for Analytics



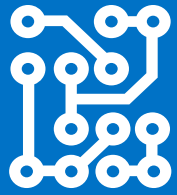
What is Predictive Analytics



- ✓ Supervised Learning
- ✓ History -> Prediction
- ✓ Examples -> Labels
- ✓ Probability
- ✓ Statistics
- ✓ Discrete/Continuous
- ✓ Predict N-class
- ✓ Algorithm Engineering
- ✓ Data Engineering

Important parts of Prediction Datasets

Types of Data attributes



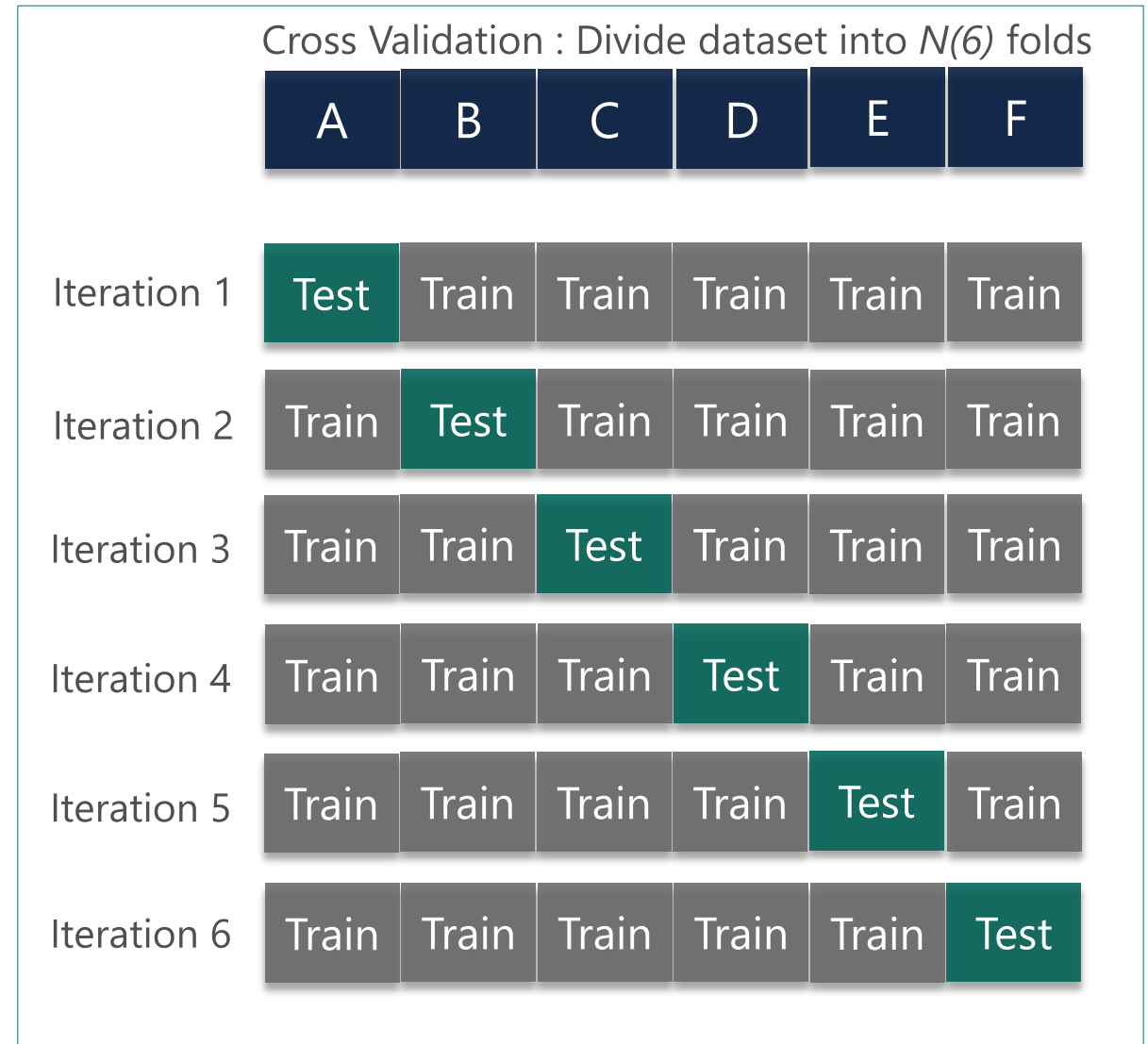
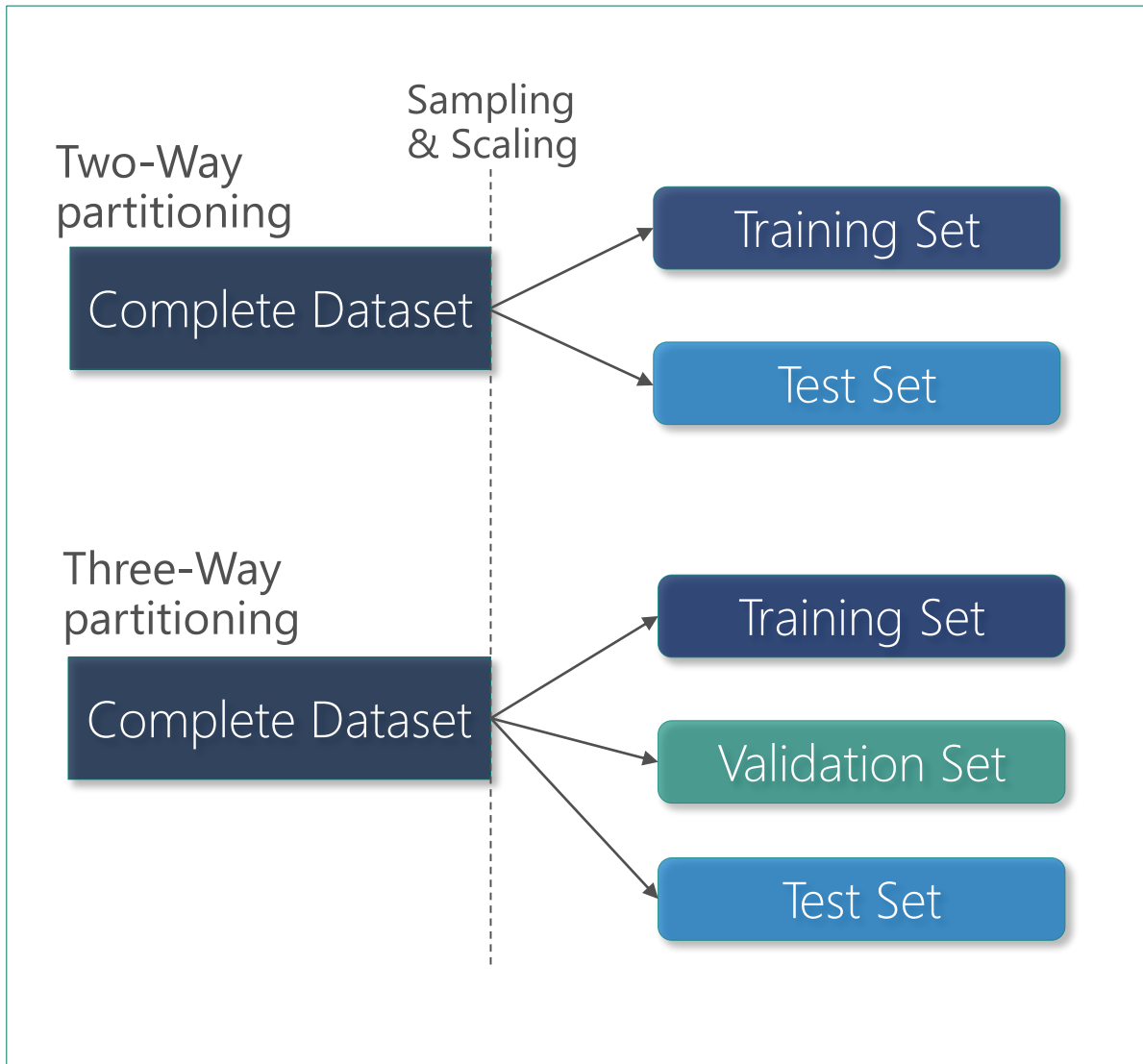
Features

Target



Risks

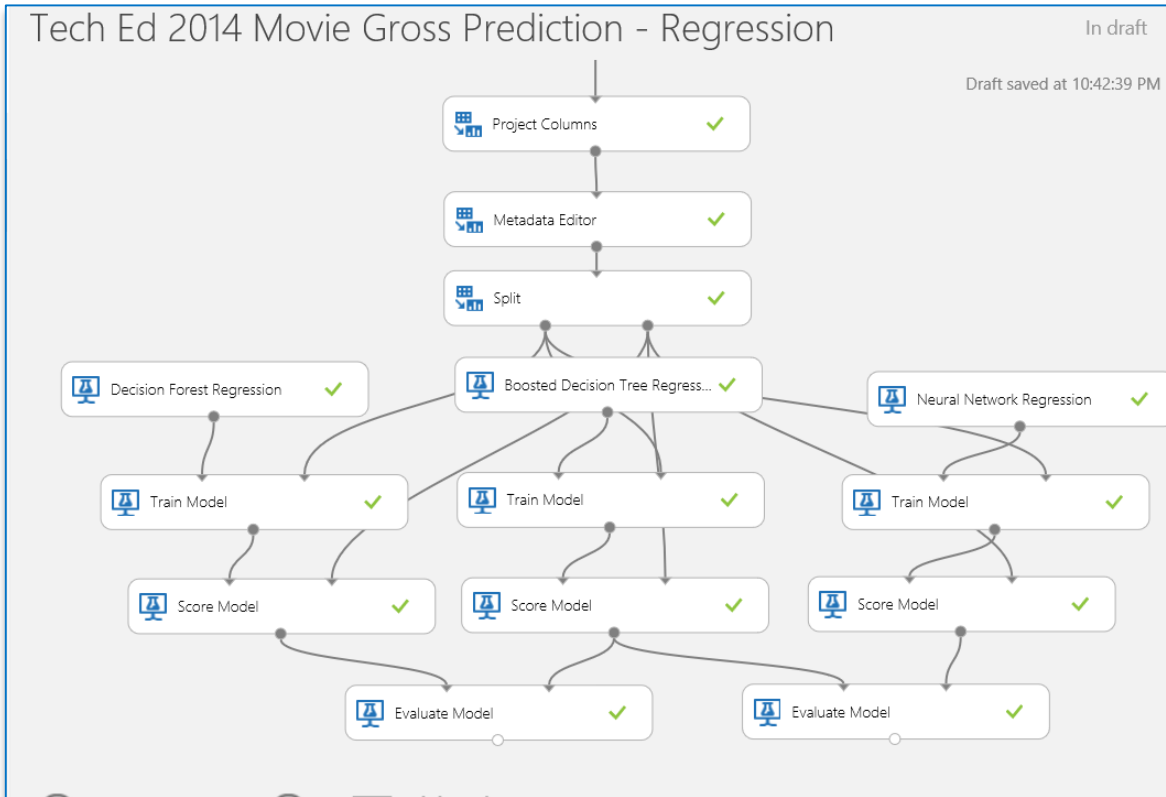
Model Training Methods



Demo

Classification & Regression

Movie Gross Prediction

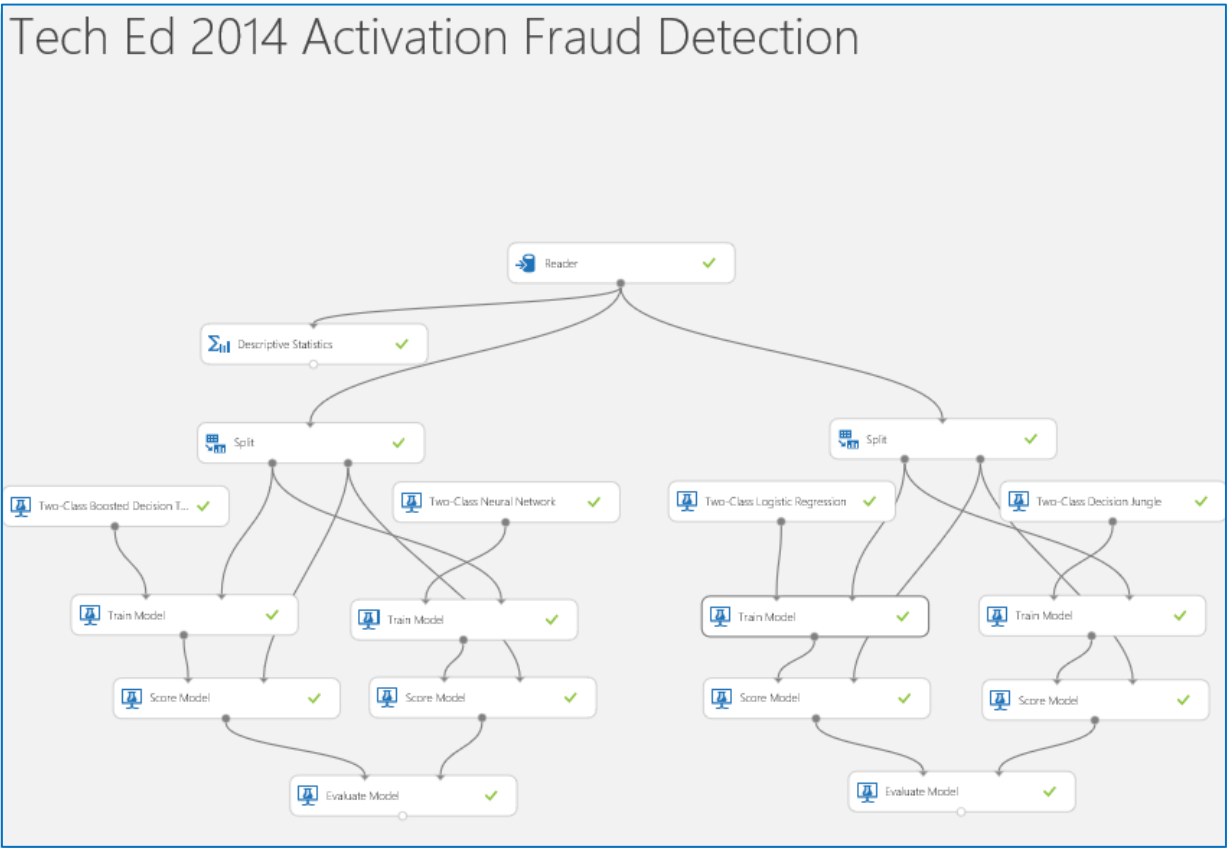


Tech Ed 2014 Movie Gross Prediction - Regressi... > Evaluate Model > Evaluation results

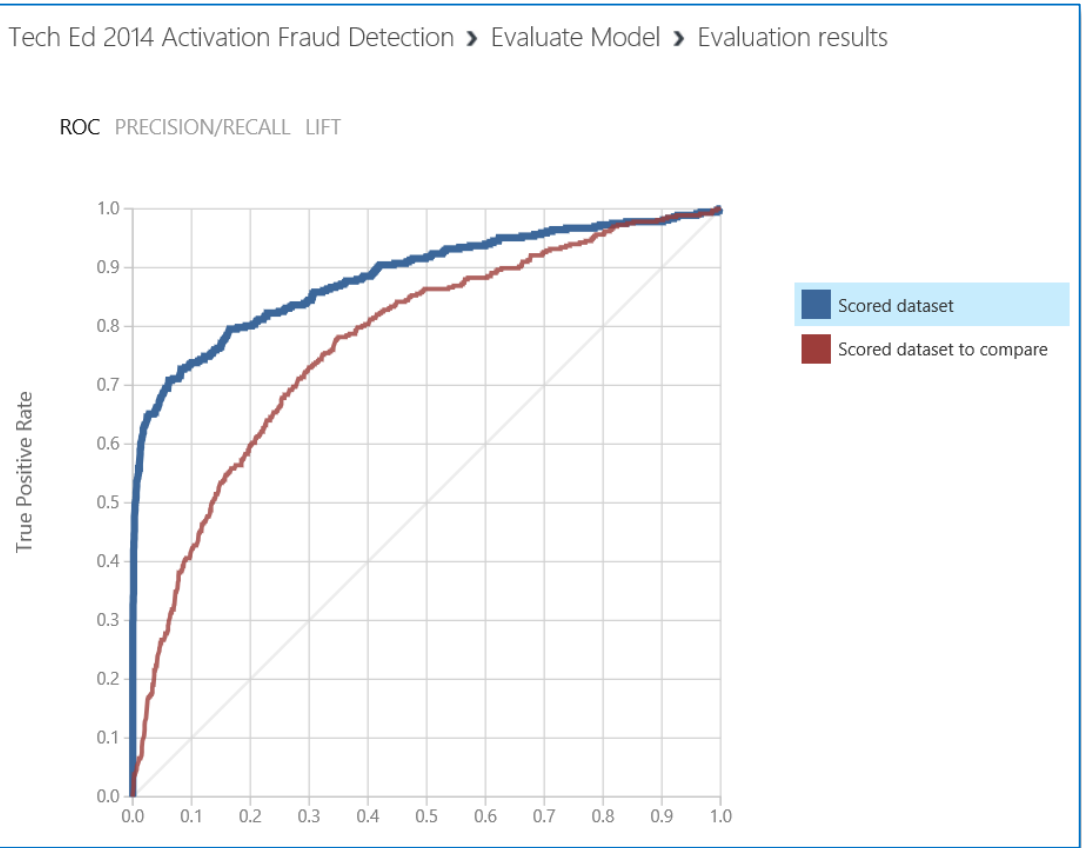
rows	columns	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
2	5	255392.992424	327050.877277	1.209493	1.555731	-0.555731
		615000	643386.68666	1.853425	3.831363	-2.831363

Activation Fraud detection

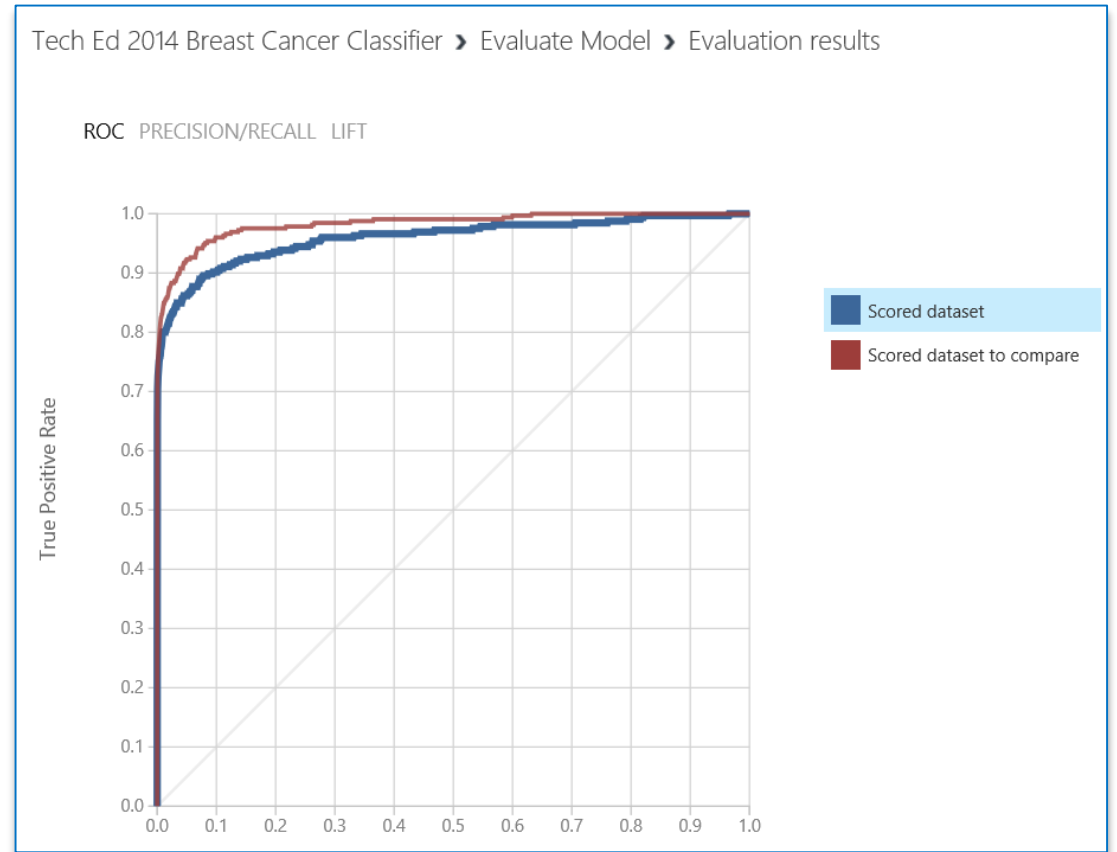
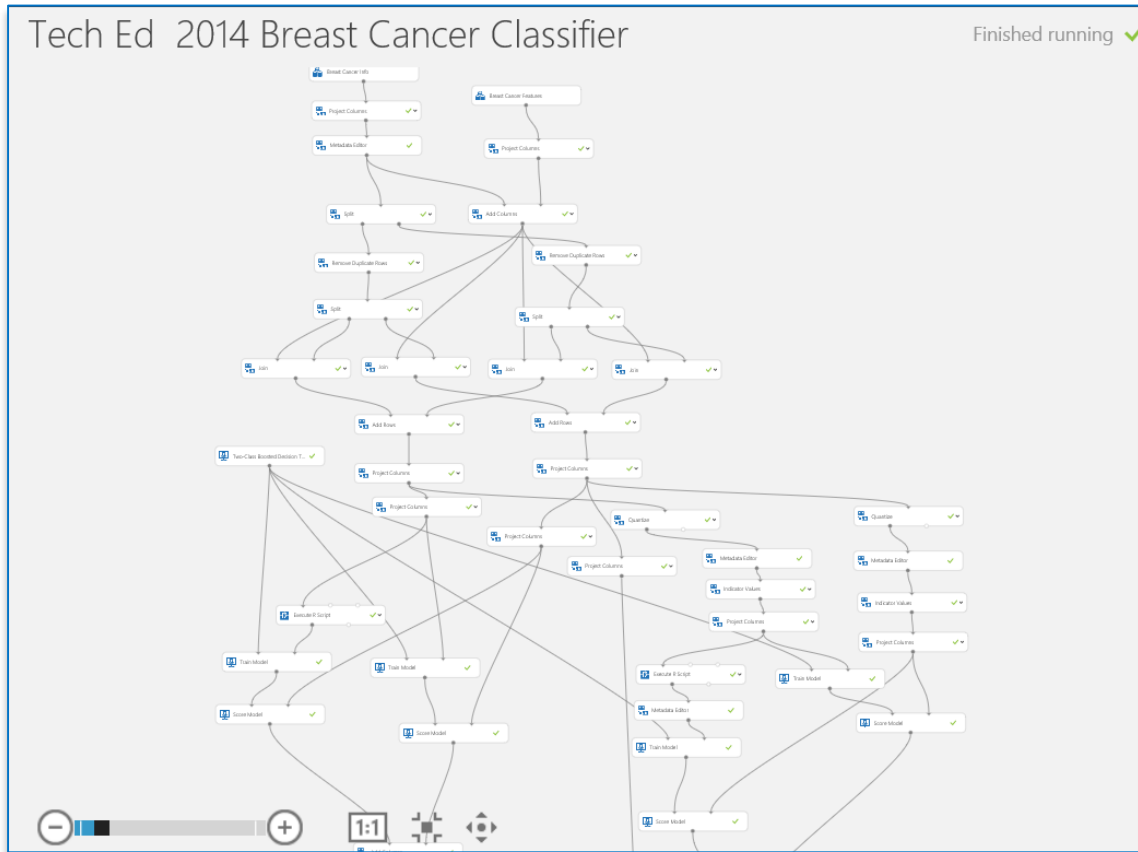
Tech Ed 2014 Activation Fraud Detection



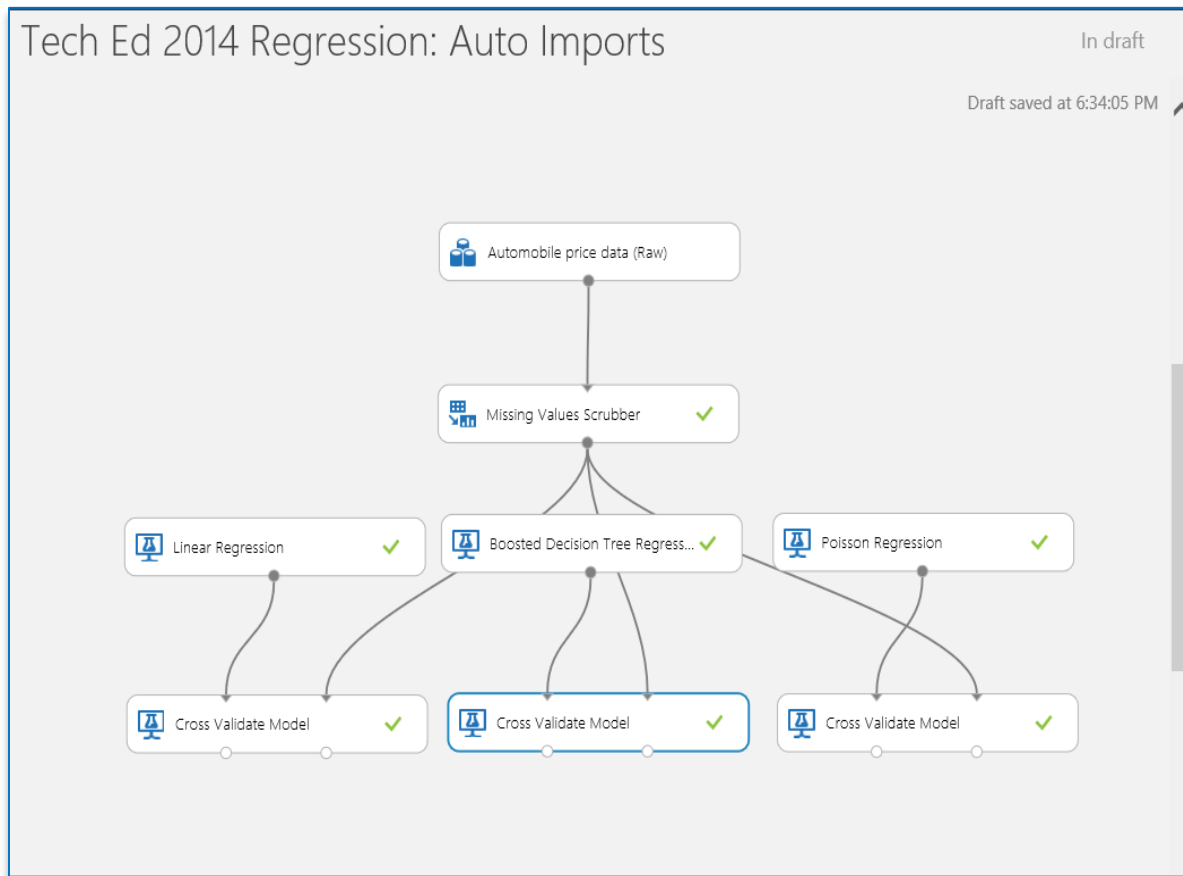
Tech Ed 2014 Activation Fraud Detection > Evaluate Model > Evaluation results



Breast Cancer Prediction



Automobile Price Prediction



Tech Ed 2014 Regression: Auto Imports > Cross Validate Model > Evaluation results by fold

rows: 12, columns: 8

Fold Number	Number of examples in fold	Model	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error
0	20	Boosted Trees Regression (FastRank)	3592.302637	4869.292302	0.890838	1.160533
1	21	Boosted Trees Regression (FastRank)	2823.884928	4657.849637	0.681399	0.740715
2	20	Boosted Trees Regression (FastRank)	6791.005005	11205.280223	0.869277	1.222005
3	21	Boosted Trees	5447.980748	9133.432145	0.909056	1.215119

Statistics

Mean	7596.2005
Median	8004.1508
Min	2310.0669
Max	11205.2802
Standard Deviation	2671.5876
Unique Values	12
Missing Values	0
Feature Type	Numeric

Visualizations

Root Mean Squared Error Histogram

compare to:

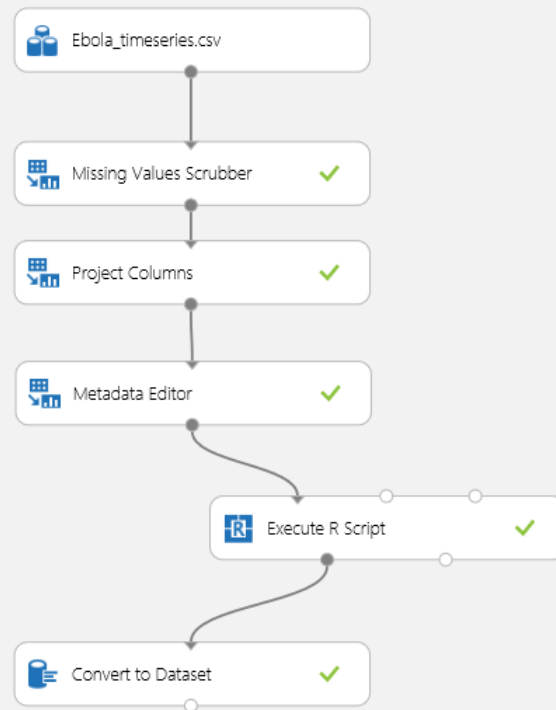
Demo

Time Series Prediction

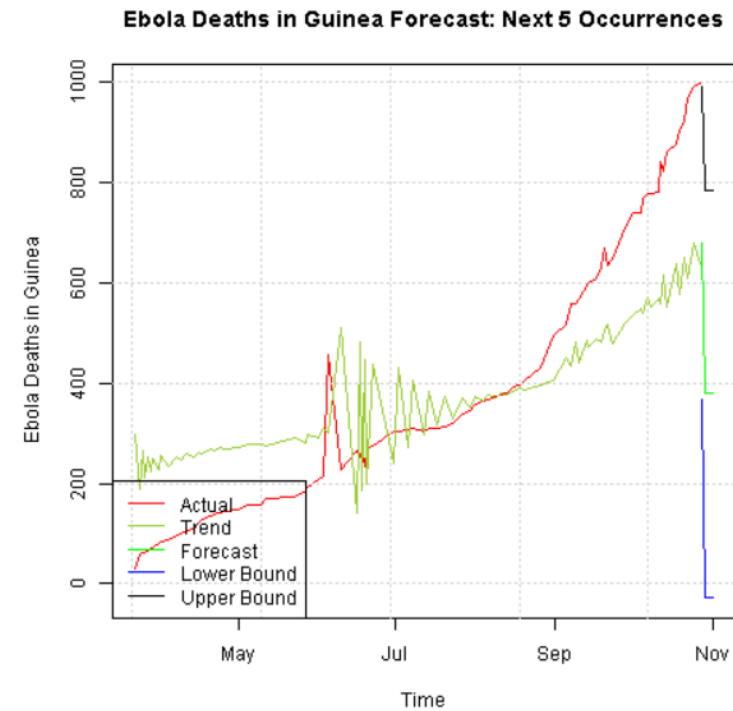
Dataset Source: <https://github.com/cmriivers/ebola>

Ebola Prediction

Ebola TimeSeries Prediction for Guinea using R ARI



Ebola TimeSeries Prediction for Guinea using R... ▶ Execute R Script ▶ R Device



USD to INR conversion rate ARIMA/STL

Tech Ed 2014 USD to INR Rate Prediction Finished running ✓

```
graph TD; A[USDINR_Data.zip] --> B[Execute R Script]; B --> C[Convert to CSV]; C --> D[Writer];
```

Properties

Writer

Please specify data destination
AzureBlobStorage

Please specify authentication type
Account

Azure account name
teched2k14ml

Azure account key
.....

Path to blob beginning with container
data/USDINRConv.csv

Azure blob storage write mode
Error

File format for blob file
CSV

Write blob header row

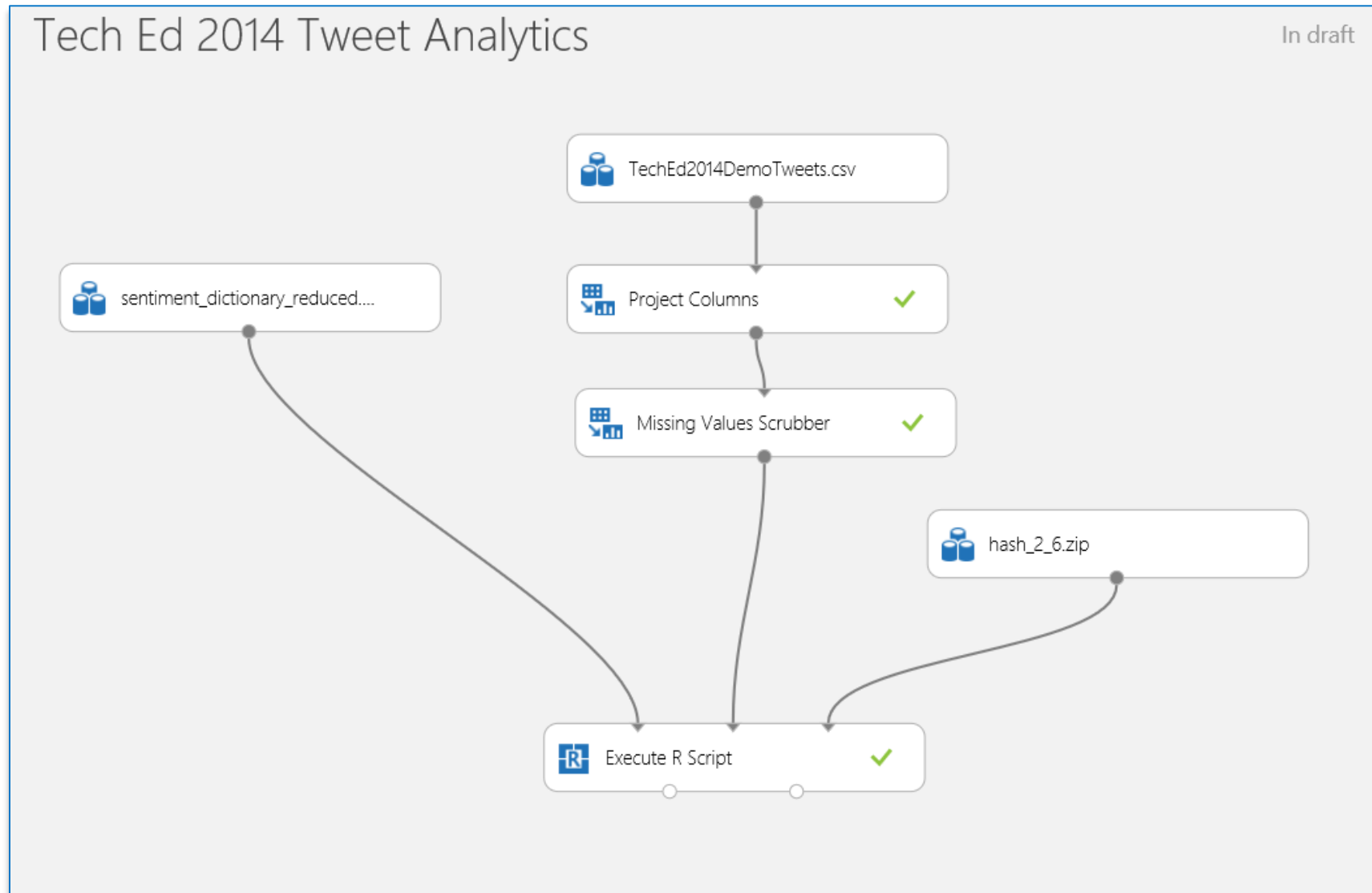
Writer

Write a dataset to Windows Azure BLOB storage, an SQL Azure table, HDFS
[\(more...\)](#)

Demo

Text Analytics

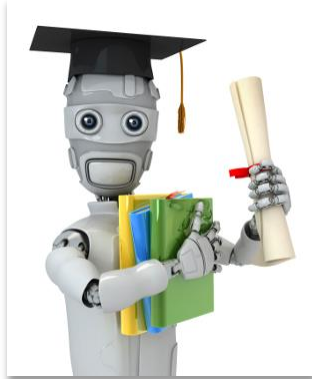
Tweet Analysis



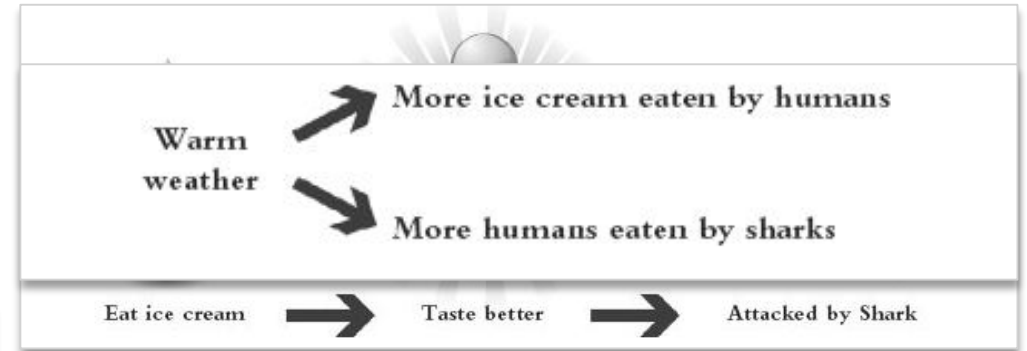
Tips for budding Data Scientists

- *Correlation v/s Causation*

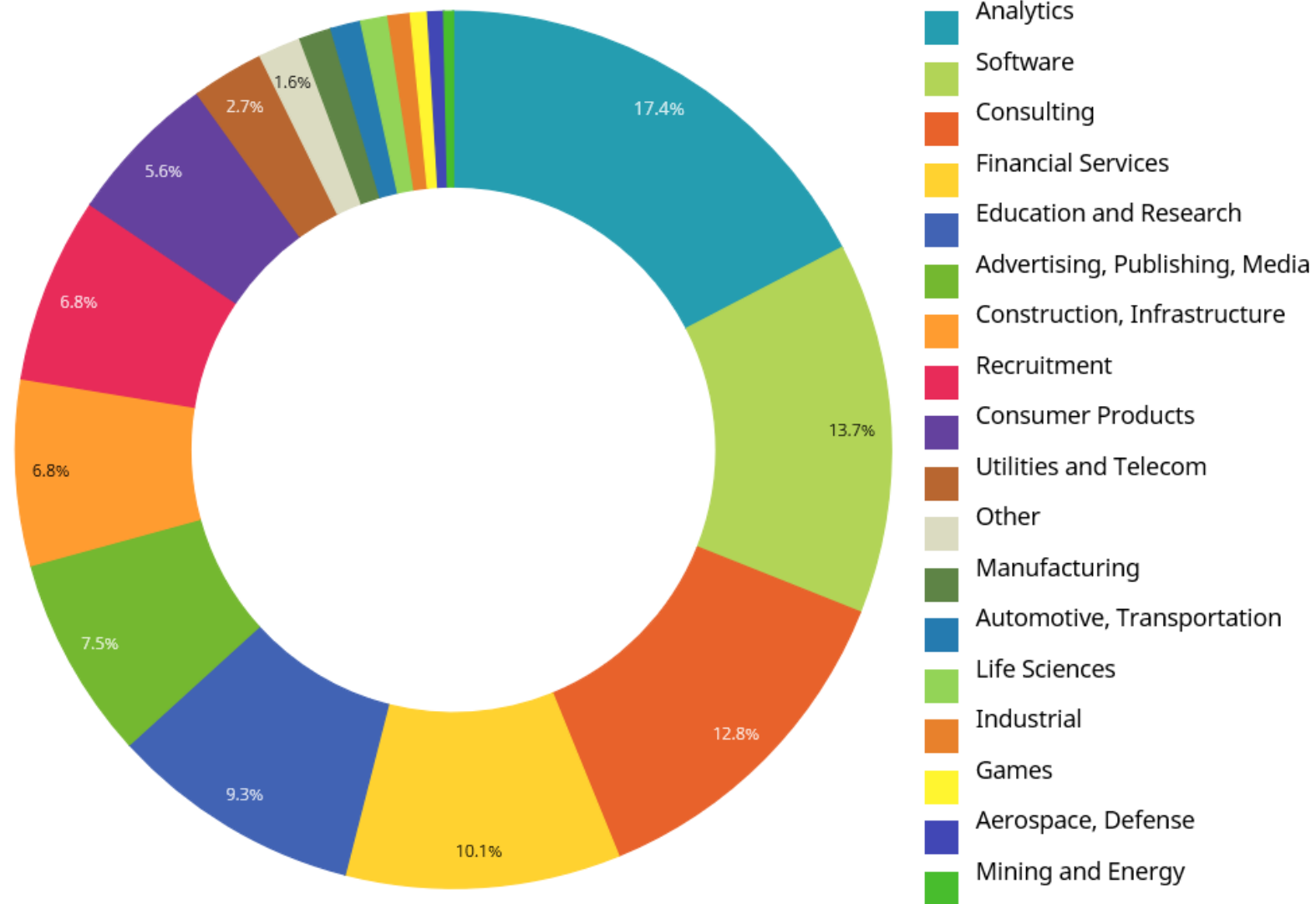
- You don't need a Ph.D. for this



- But you do need *Common Sense* and *Ability* to derive sense out of inferences made by the *Machine(s)!!!*



Data Scientist Jobs 😊



What Next for ML @ Microsoft

Microsoft Azure Marketplace

Language: English | Region: United States | Support | Sign In

Learn Applications Data My Account Publish Search the Marketplace

HOME > DATA > BINARY CLASSIFIER API BUILT WITH AZURE MACHINE LEARN...

Binary Classifier API built with Azure Machine Learning

Data

25,000 Transactions/month | \$0.00 per month | SIGN UP

Microsoft Azure Marketplace

Language: English | Region: United States | Support | Sign In

Learn Applications Data My Account Publish Search the Marketplace

HOME > DATA > ANOMALY DETECTION API BUILT WITH AZURE MACHINE LEARNING

Anomaly Detection API built with Azure Machine Learning

Data

25,000 Transactions/month | \$0.00 per month | SIGN UP

Microsoft Azure Marketplace

Language: English | Region: United States | Support | Sign In

Learn Applications Data My Account Publish Search the Marketplace

HOME > DATA > MULTIVARIATE LINEAR REGRESSION API BUILT WITH AZURE ...

Multivariate Linear Regression API built with Azure Machine Learning

Data

25,000 Transactions/month | \$0.00 per month | SIGN UP

Microsoft Azure Marketplace

Language: English | Region: United States | Support | Sign In

Learn Applications Data My Account Publish Search the Marketplace

HOME > DATA > LEXICON BASED SENTIMENT ANALYSIS API BUILT WITH AZUR...

Lexicon Based Sentiment Analysis API built with Azure Machine Learning

Data

25,000 Transactions/month | \$0.00 per month | SIGN UP

Microsoft Azure Marketplace

Language: English | Region: United States | Support | Sign In

Learn Applications Data My Account Publish Search the Marketplace

HOME > DATA > FORECASTING - AUTOREGRESSIVE INTEGRATED MOVING AVERA...

Forecasting - AutoRegressive Integrated Moving Average (ARIMA) API built with Azure Machine Learning

Data

25,000 Transactions/month | \$0.00 per month | SIGN UP

Documentation
Documentation of creation as well as consumption of web service.

Sample Web App
A sample web app for testing the ARIMA web

Published by: Microsoft
Categories: Machine Learning
Date added: 10/9/2014

Microsoft Azure Marketplace

Language: English | Region: United States | Support | Sign In

Learn Applications Data My Account Publish Search the Marketplace

HOME > DATA > FORECASTING - ETS + STL API BUILT WITH AZURE MACHINE...

Forecasting - ETS + STL API built with Azure Machine Learning

Data

25,000 Transactions/month | \$0.00 per month | SIGN UP

Documentation
Documentation of creation as well as consumption of web service.

Sample Web App

Published by: Microsoft
Categories: Machine Learning
Date added: 10/9/2014

Q&A

References

Related references for you to expand your knowledge on the subject

<http://datamarket.azure.com/dataset/amla/recommendations>

<http://datamarket.azure.com/dataset/amla/mba>

<http://azure.microsoft.com/en-us/services/machine-learning/>

http://en.wikipedia.org/wiki/Predictive_analytics

<http://blogs.technet.com/b/machinelearning/archive/2014/09/17/extensibility-and-r-support-in-the-azure-ml-platform.aspx>

<http://blogs.technet.com/b/saketbi/archive/2014/08/20/microsoft-azure-ml-amp-r-language-extensibility.aspx>

Predictive Analytics: The power to predict who will click, buy, lie, or Die

by Eric Siegel

TechNet
technet.microsoft.com/en-in

mva
Microsoft Virtual Academy
aka.ms/mva

Developer Network
msdn.microsoft.com/

Follow us online

Twitter: #AMLTechEd2014

Email:

vikas.goyal@microsoft.com
saket.suman@microsoft.com



Facebook
facebook.com/MicrosoftDeveloper.India



Twitter
twitter.com/msdevindia

Your Feedback is Important

Fill out evaluation of this session and help shape future events.

You'll also be entered into a daily prize drawing!

OPTION 1



Scan the QR code to evaluate this session on your mobile device.

OPTION 2

A screenshot of a mobile application interface for feedback. The title is "TechEd India 2014" and "feedback". The session is "Sending Cross Platform Notifications us". Under "Session Rating *", there are four radio button options: "Very Good", "Good", "Okay", and "Bad". Under "Speaker Rating *", there are two radio button options: "Very Good" and "Good". A back arrow is visible at the bottom.

You can fill out evaluation of this session directly through the App

OPTION 3: Feedback stations outside the hall



Microsoft

© 2014 Microsoft Corporation. All rights reserved. Microsoft, Windows, and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries. The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

Appendix

Machine Learning Examples

Credit Card Fraud Detection – Anomaly detection

Digit Recognition for handwriting recognition – Pattern Recognition

Facial & Image recognition engines – Pattern Recognition

Recommendation Engines for Products in e-commerce – MBA with Recommender

Stock Trading – ARIMA on Time Series

Medical Diagnosis of Diseases & Lifespan – Clustering & Association

Crash prediction of machines in the Cloud - Classifier

Vowpal Wabbit from MSR

The VW program supports:

- Multiple supervised (and semi-supervised) learning problems:

Classification (both binary and multi-class)

Regression

Active learning (partially labeled data) for both regression and classification

- Multiple learning algorithms (model-types / representations)

OLS regression

Matrix factorization (sparse matrix SVD)

Single layer Neural net (with user specified hidden layer node count)

Search (Search and Learn)

Latent Dirichlet Allocation (LDA)

Stagewise polynomial approximation

Recommend top-K out of N

One-against-all (OAA) and cost-sensitive OAA reduction for multi-class

Weighted all pairs

Contextual-bandit

- Multiple loss functions:

Squared error

Quantile

Hinge

Logistic

- Multiple optimization algorithms

Stochastic gradient descent (SGD)

BFGS

Conjugate gradient

- Regularization (L1 norm, L2 norm, & elastic net regularization)

Flexible input - input features may be

Binary

Numerical

Categorical (via flexible feature-naming and the hash trick)

Can deal with missing values/sparse-features

- Other features

On the fly generation of feature interactions (quadratic and cubic)

On the fly generation of N-grams with optional skips (useful for word/language data-sets)

Automatic test-set holdout and early termination on multiple passes bootstrapping

User settable online learning progress report + auditing of the model