TechEd

India 2014

Learn. Connect. Explore.

Microsoft

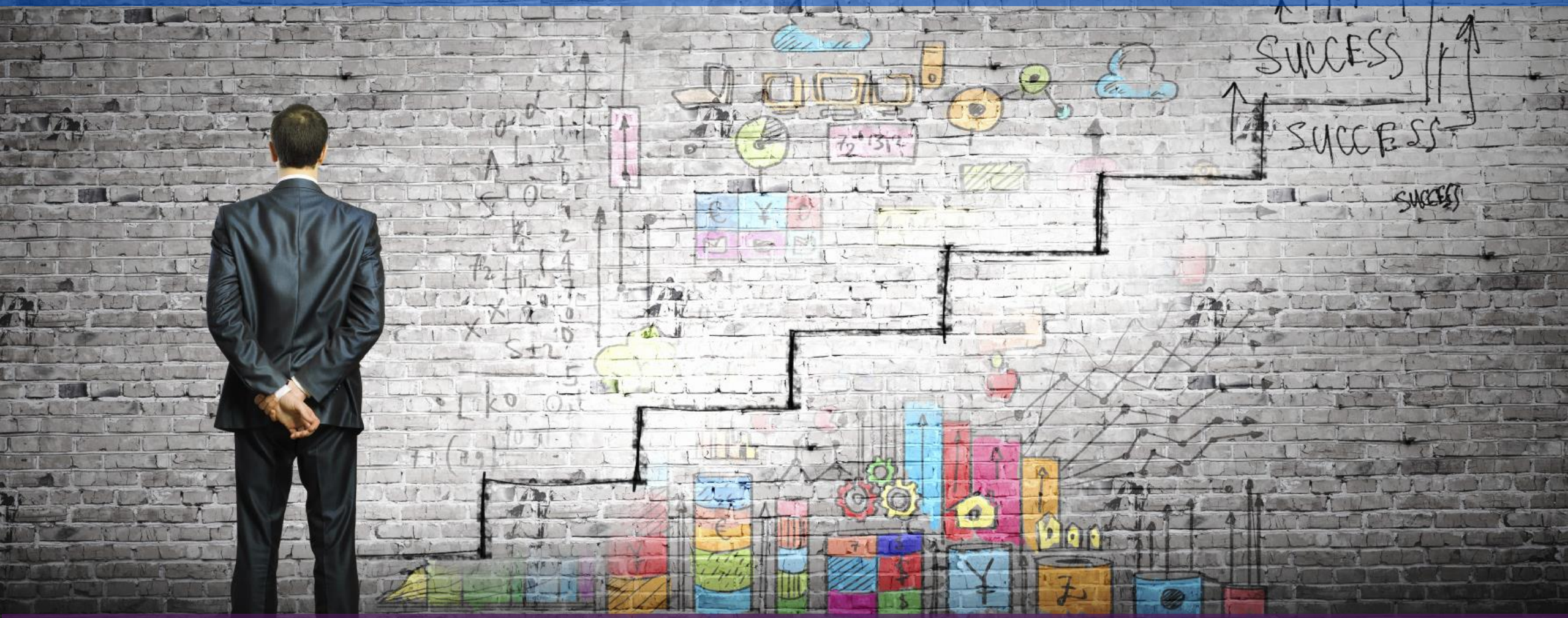
TechEd
India 2014

How to Become a Data Scientist with Azure Machine Learning

Saket Suman & Vikas Goyal
Microsoft Consulting Services

Microsoft

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

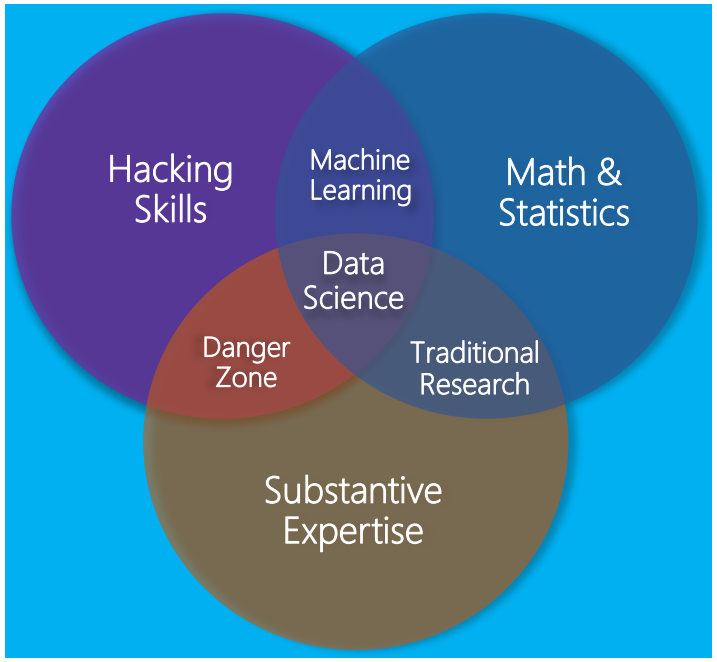
Big Data R EDW

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

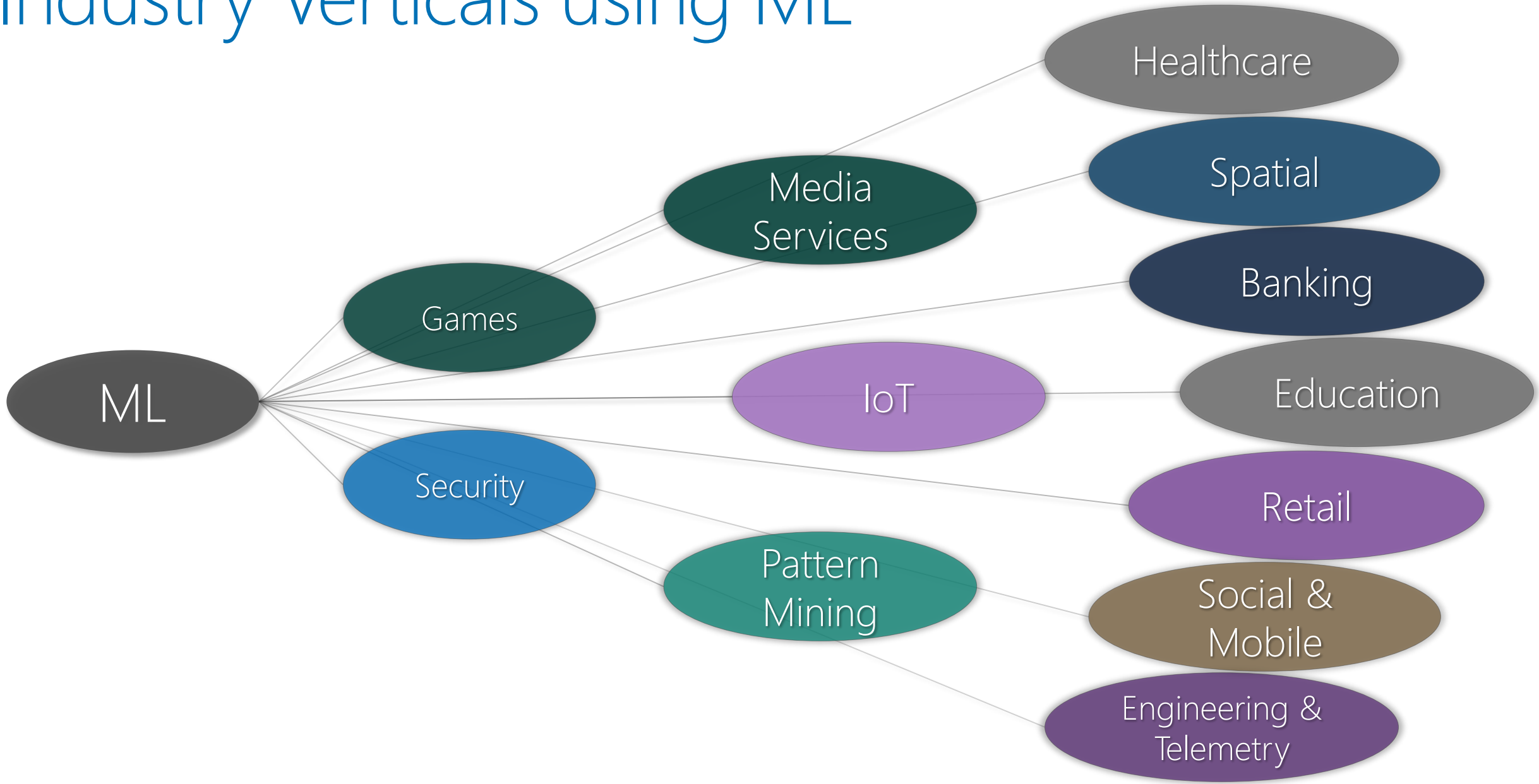
What is Machine Learning

"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

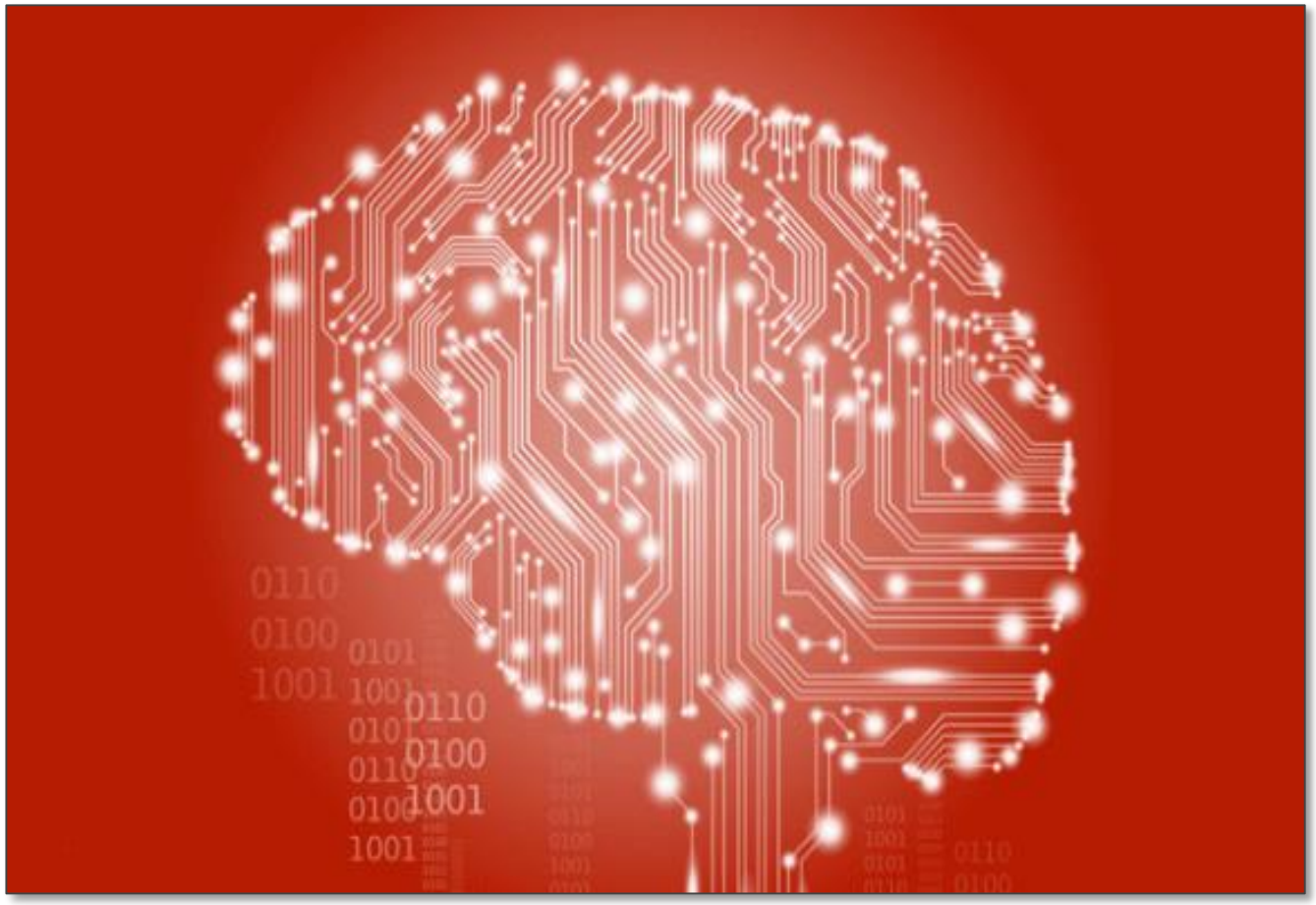
Unsupervised Learning

- → Clustering
- → Association
- → Hidden Markov

*ML is the AI, Math & Science behind KDD

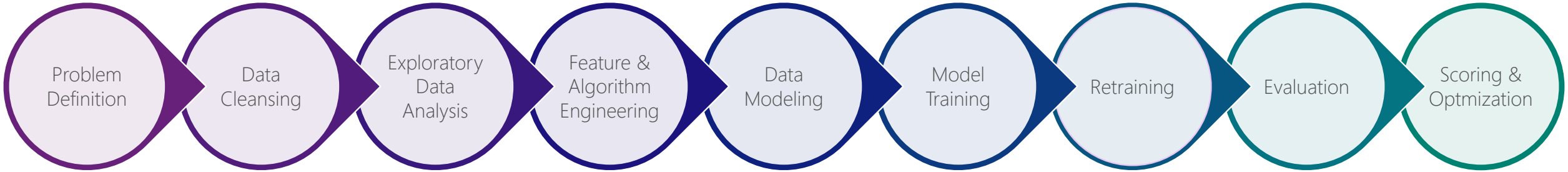
Other Methods

- → 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

1999 — Computers work on users behalf, filtering junk email

2004 — Microsoft search engine built with machine learning

2005 — SQL Server enables data mining

2008 — Bing Maps ships with ML traffic-prediction service

2010 — Microsoft Kinect can watch users gestures

2012 — Successful, real-time, speech-to-speech translation

2014 — Microsoft launches Azure Machine Learning

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

Designed for Budding & Experienced Data Scientists

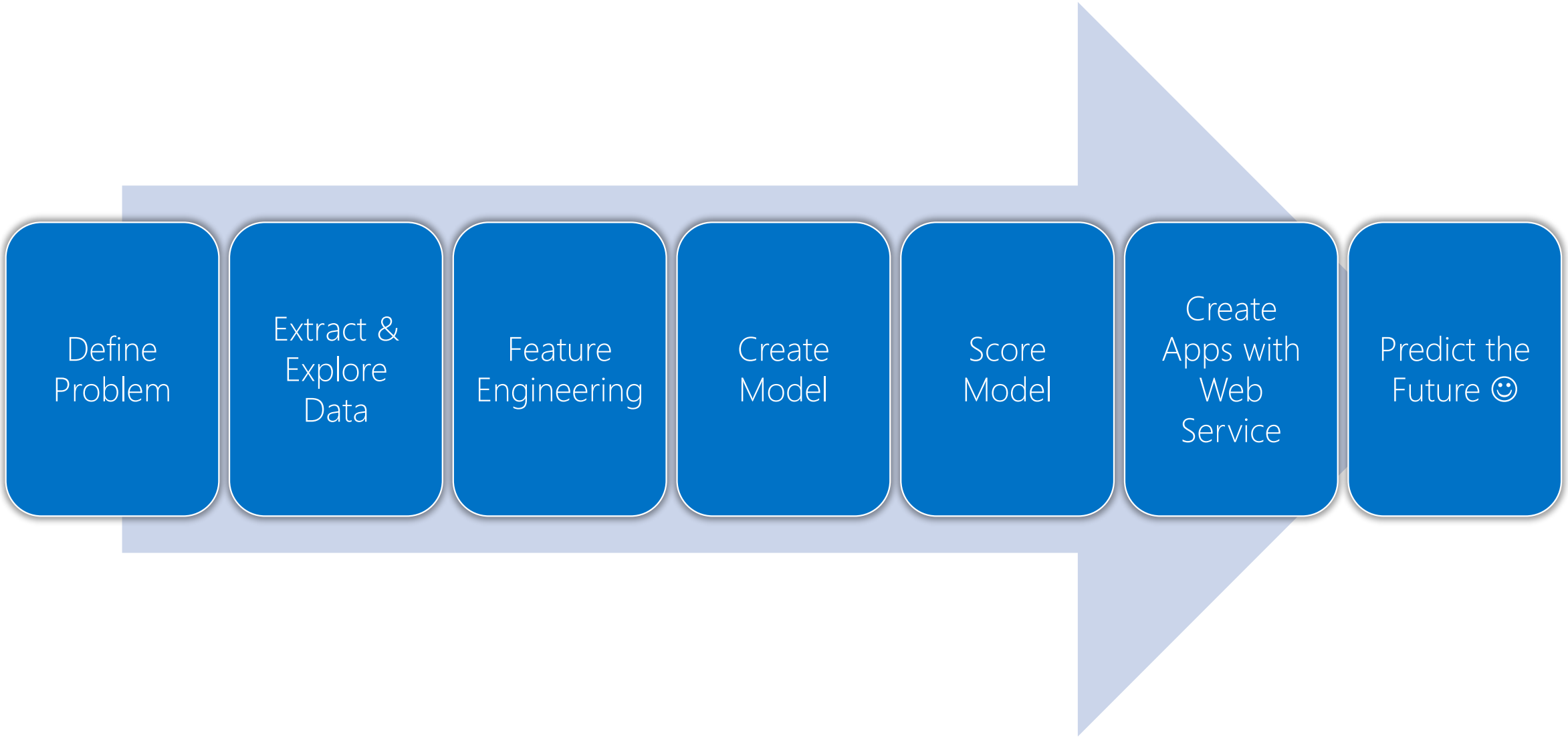
Data Scientist Toolkit

Industry Standard Algorithms & Support for R Language

Deploy Predictive models to Production in minutes

Connectivity to HDI for Hadoop

How does Azure ML make Data Science easy



How it all works

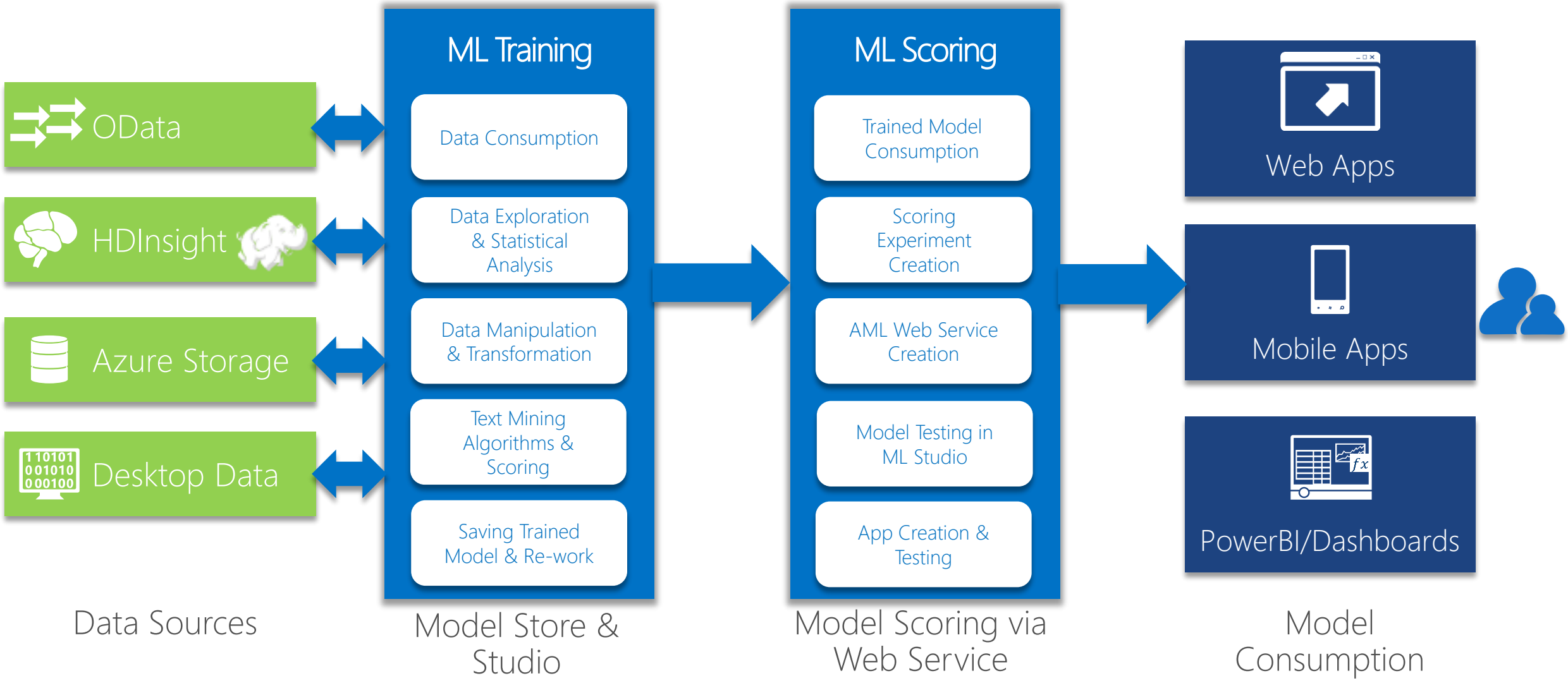
The Environments

- 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



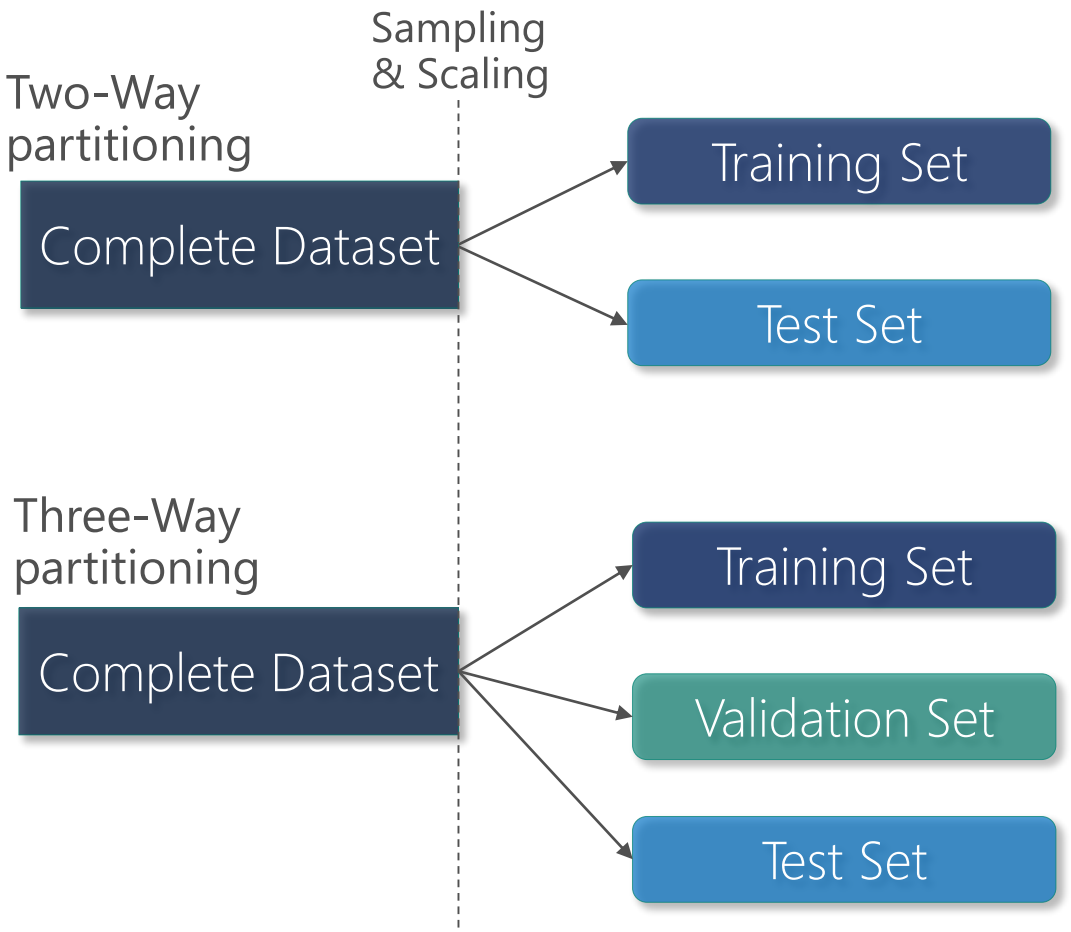
What is Predictive Analytics

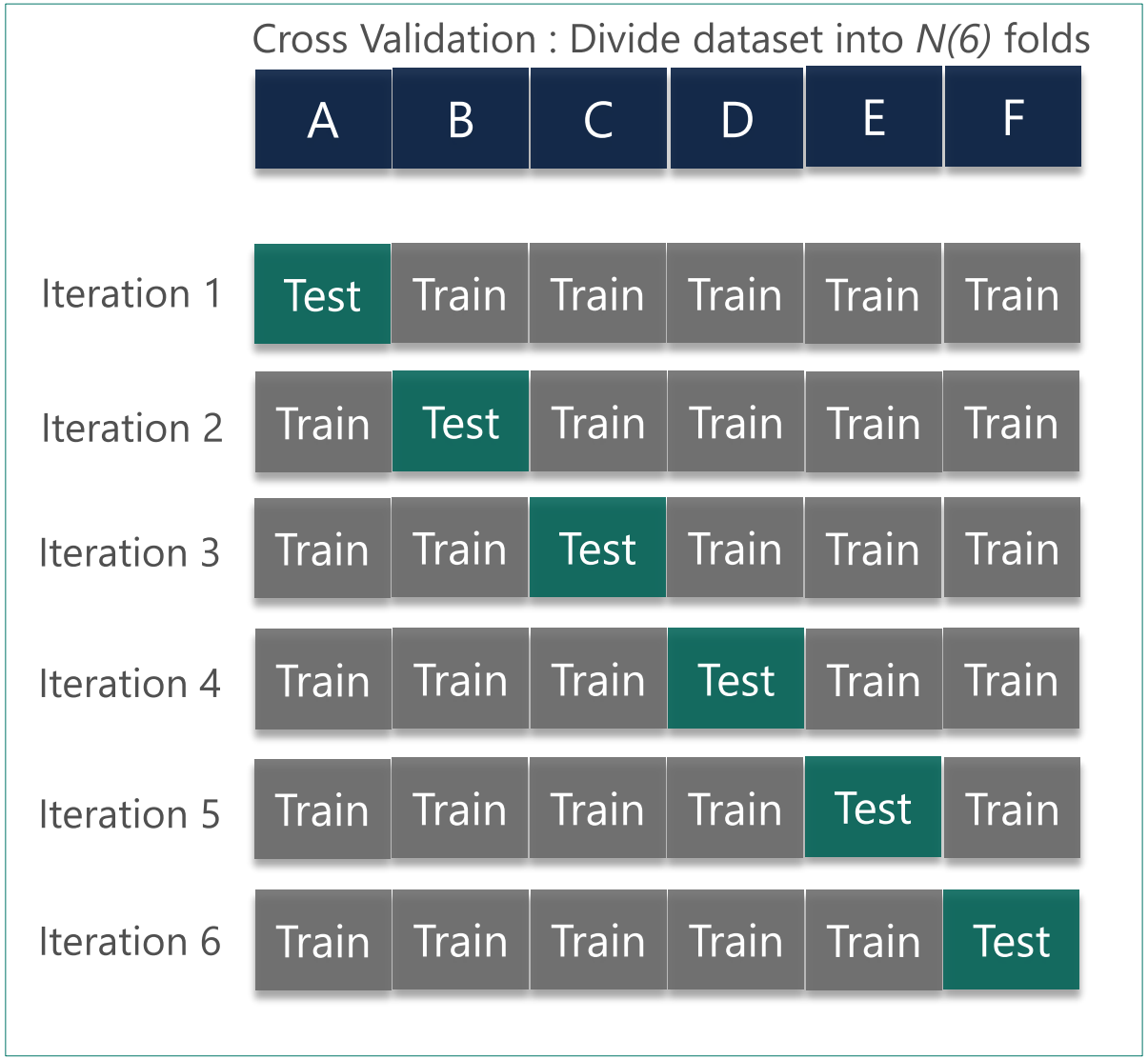
- ✓ Supervised Learning
- ✓ History -> Prediction
- ✓ Examples -> Labels
- ✓ Probability
- ✓ Statistics
- ✓ Discrete/Continuous
- ✓ Predict N-class
- ✓ Algorithm Engineering
- ✓ Data Engineering

Important parts of Prediction Datasets

Types of Data attributes

Model Training Methods



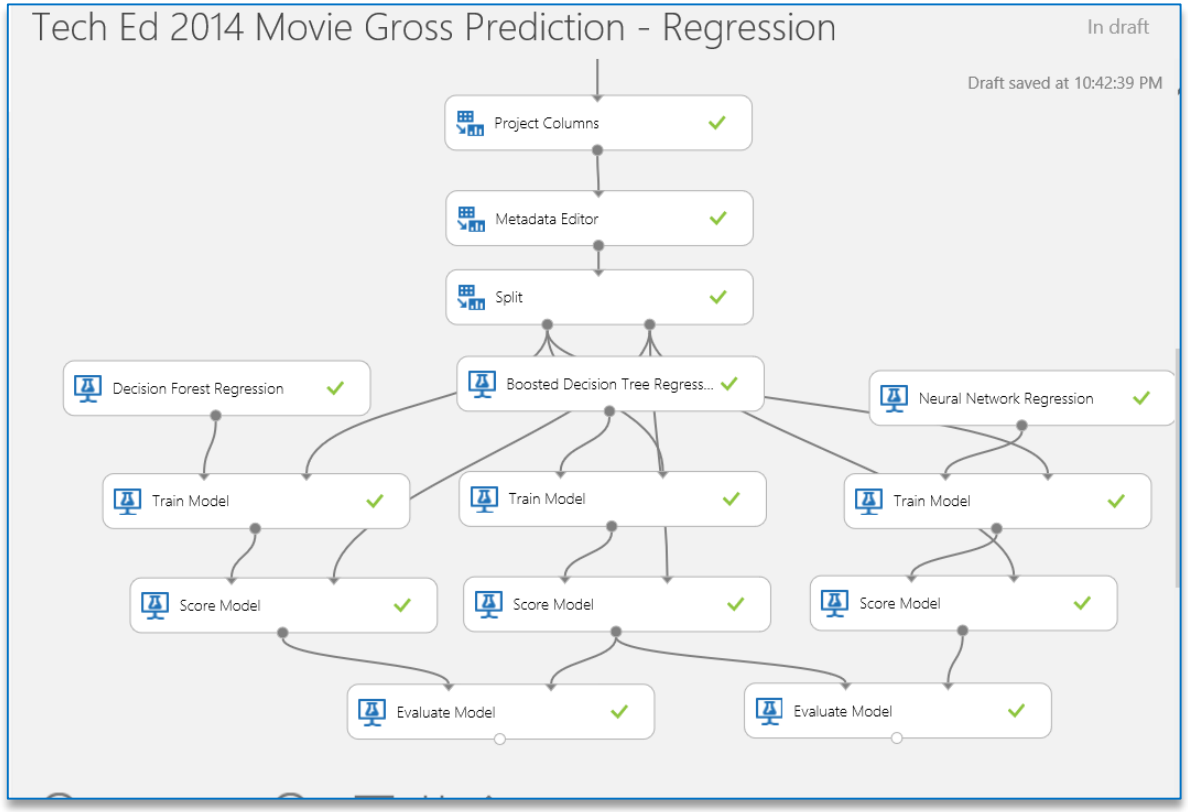


Source: From a book by Thomas W. Miller

Demo

Classification & Regression

Movie Gross Prediction



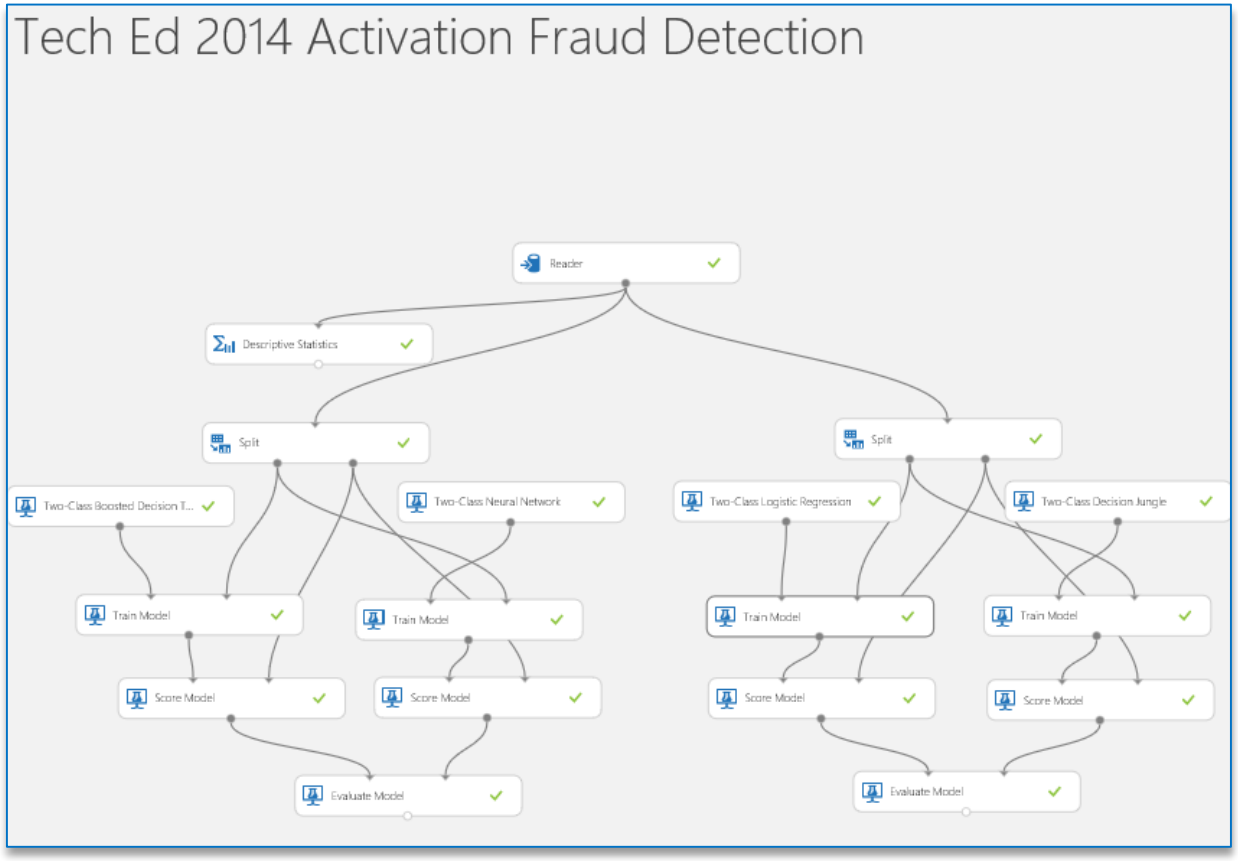
Tech Ed 2014 Movie Gross Prediction - Regressi... › Evaluate Model › Evaluation results

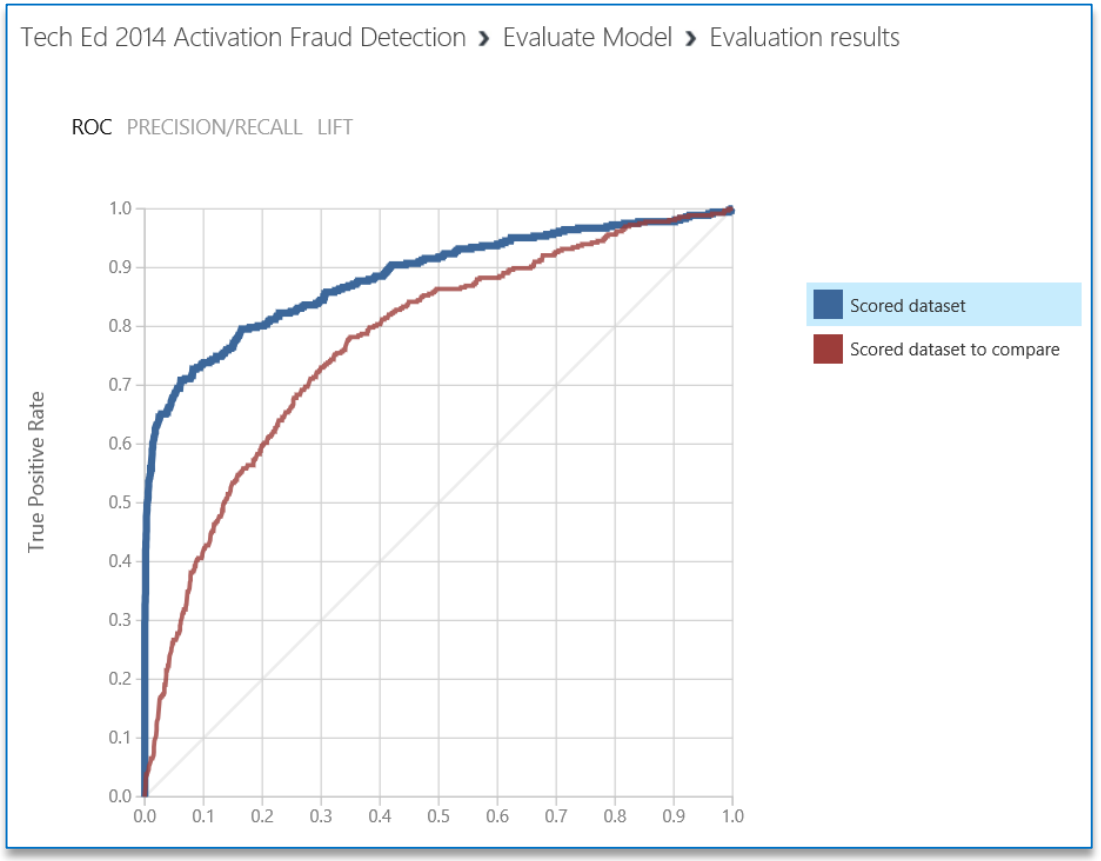
rows 2 columns 5

view as

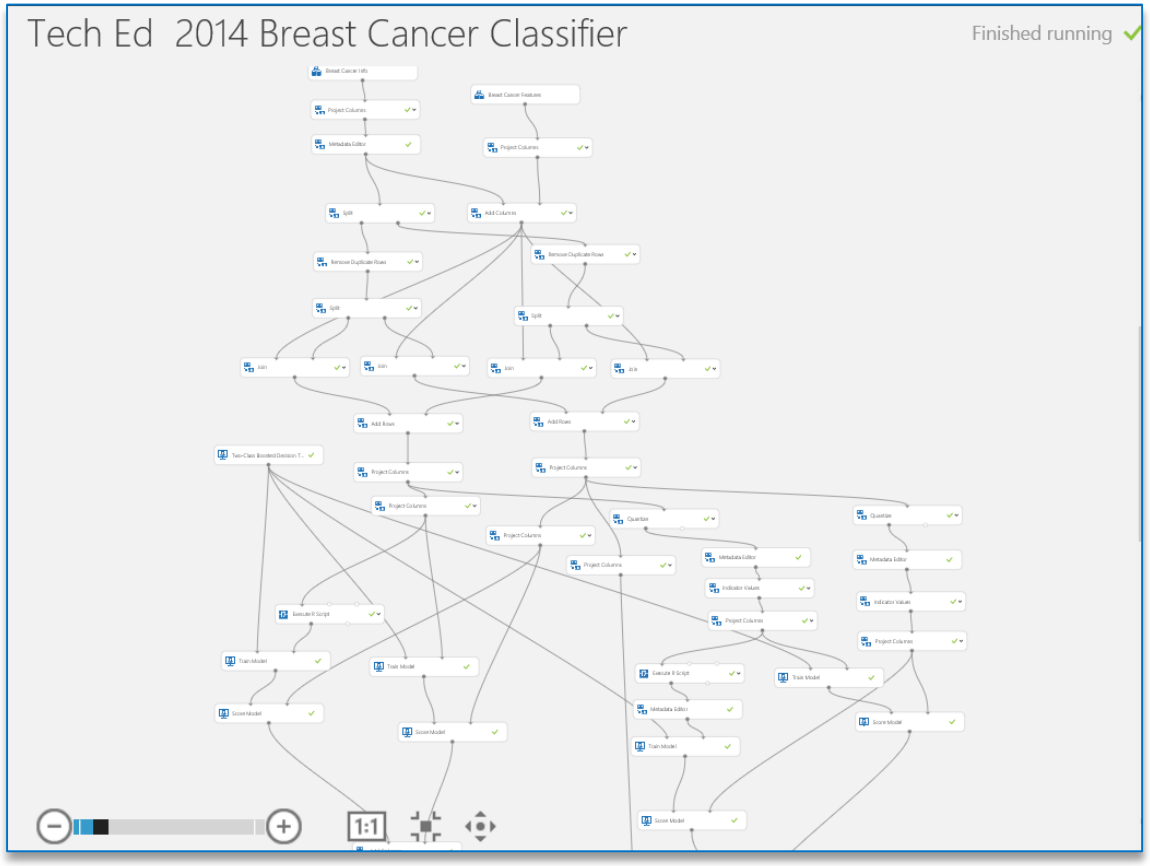
	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
	255392.992424	327050.877277	1.209493	1.555731	-0.555731
	615000	643386.68666	1.853425	3.831363	-2.831363

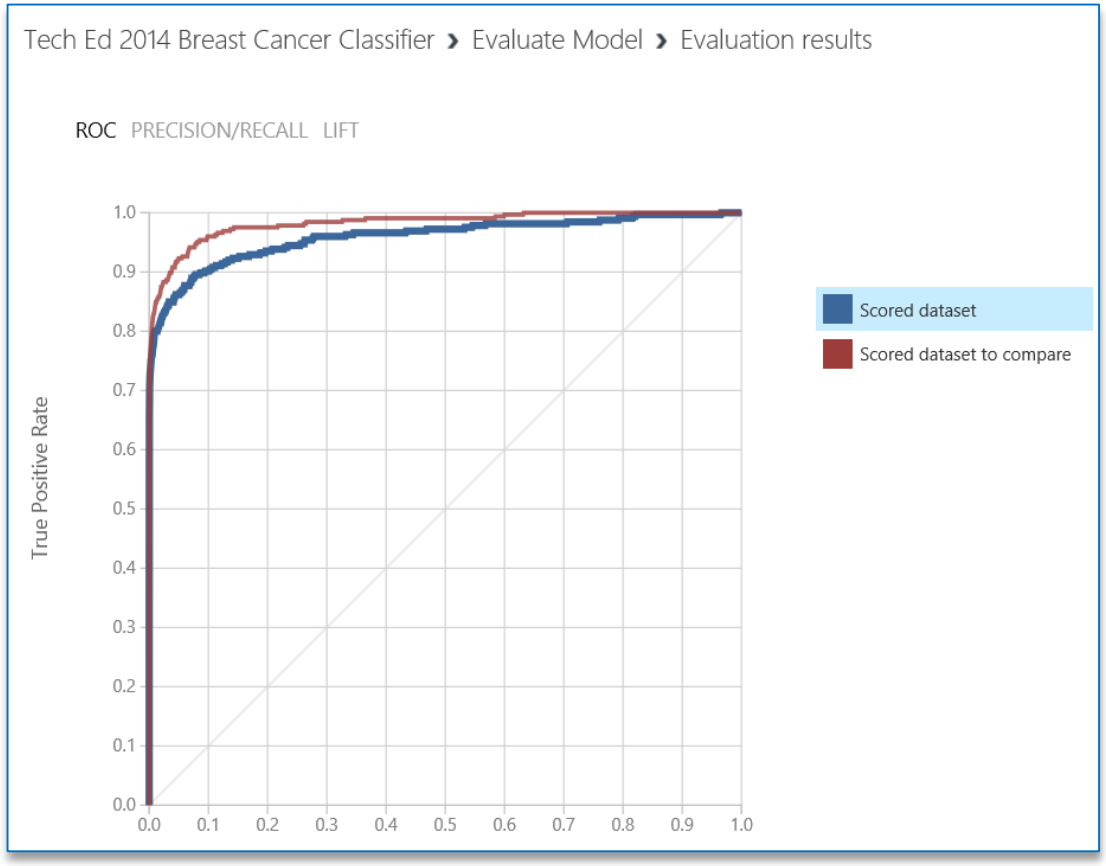
Activation Fraud detection



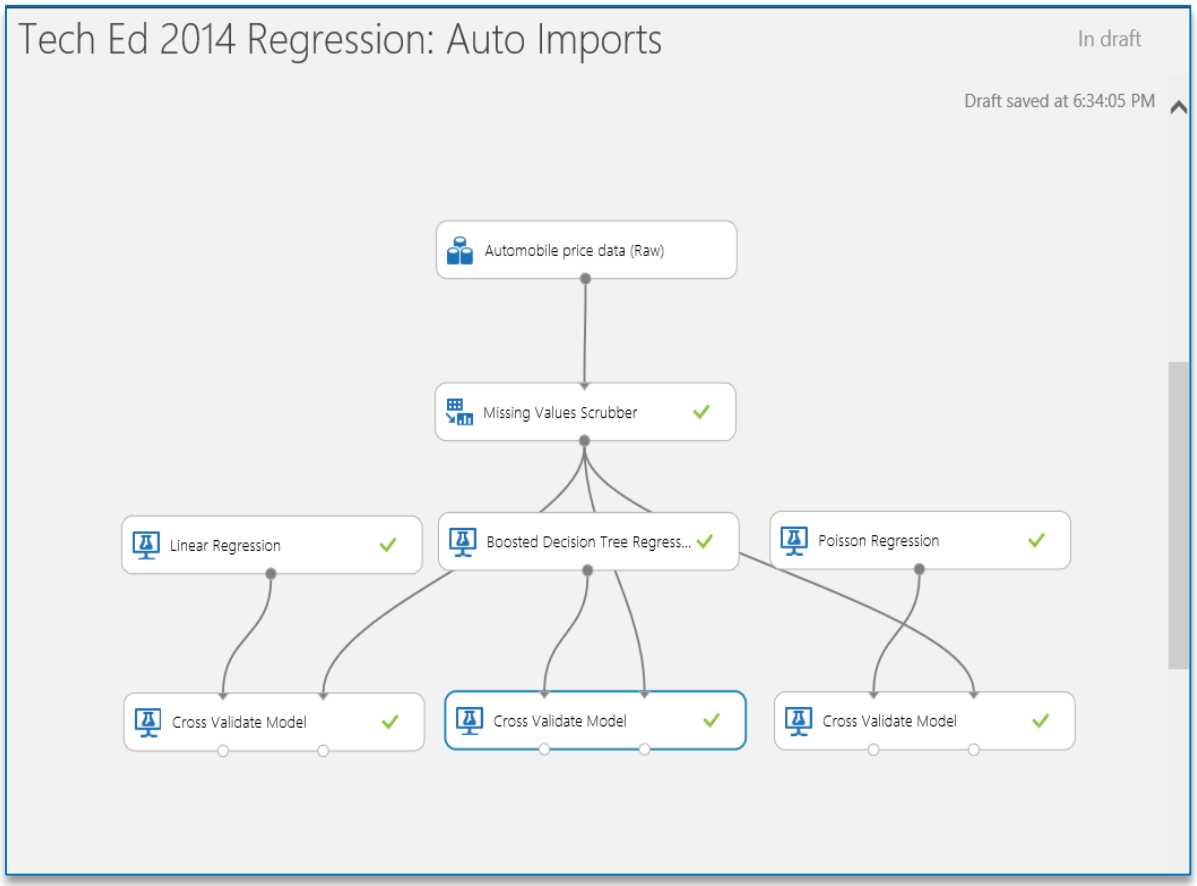


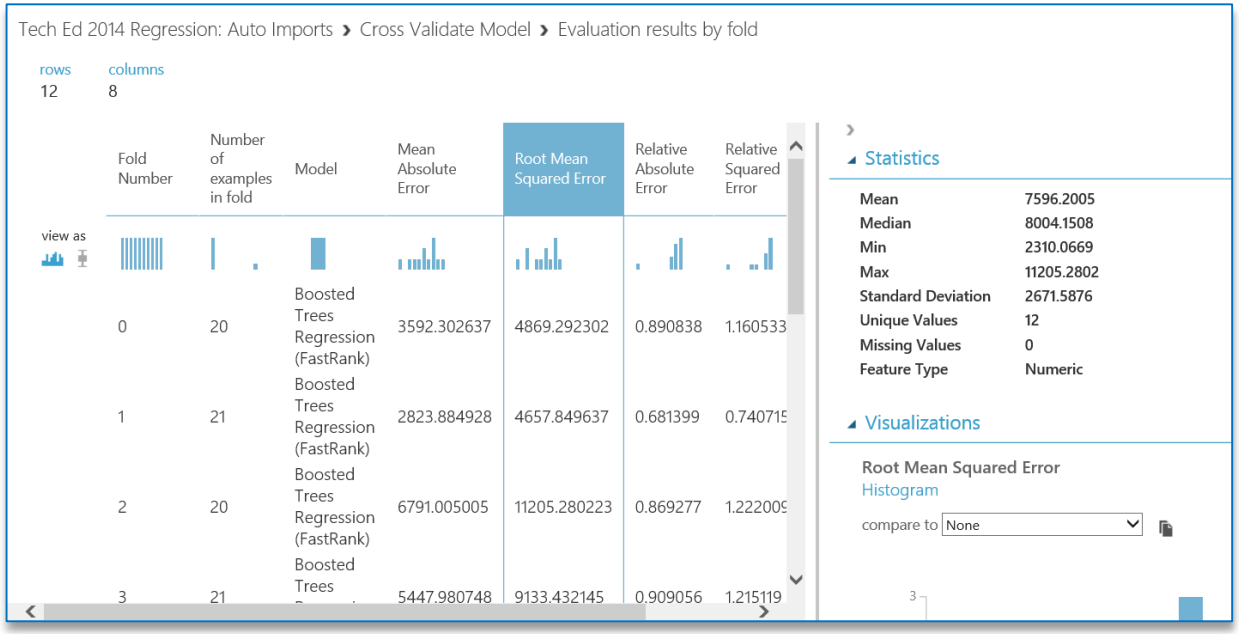
Breast Cancer Prediction





Automobile Price Prediction



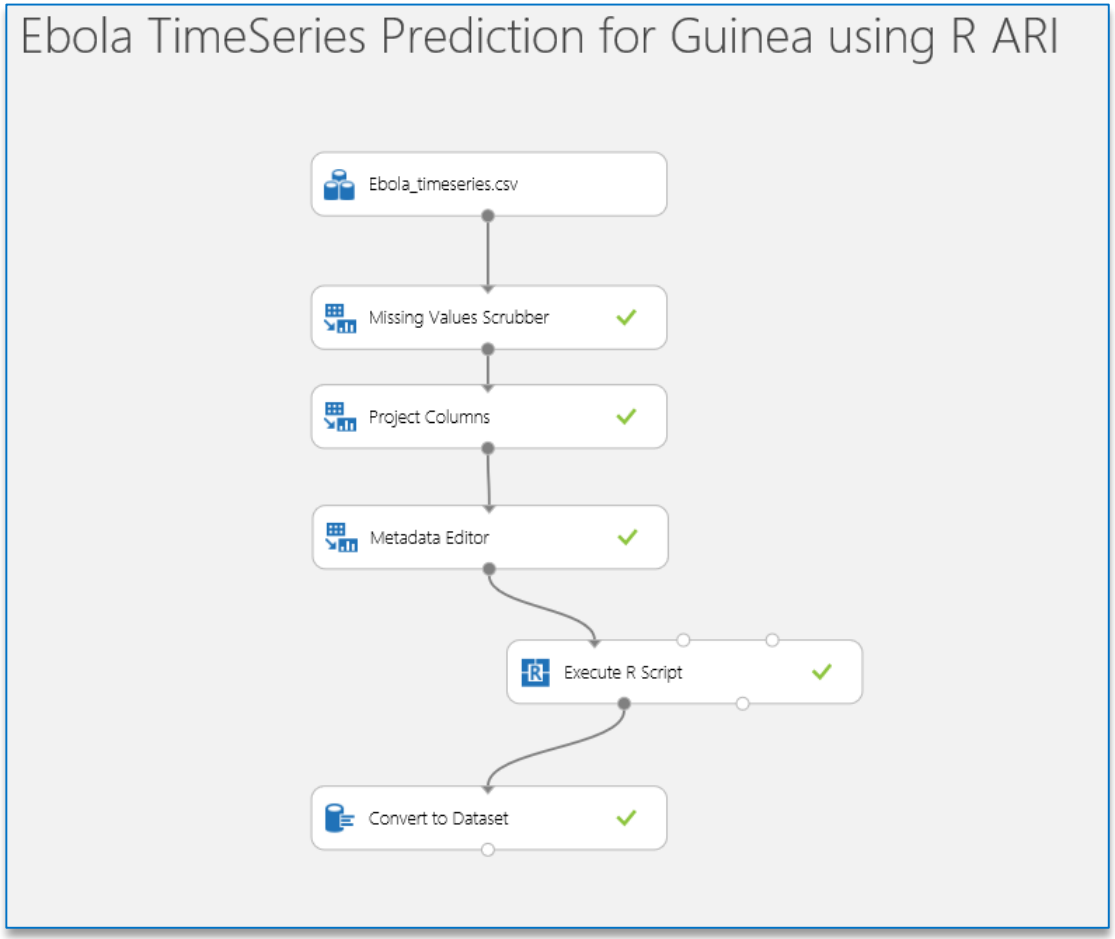


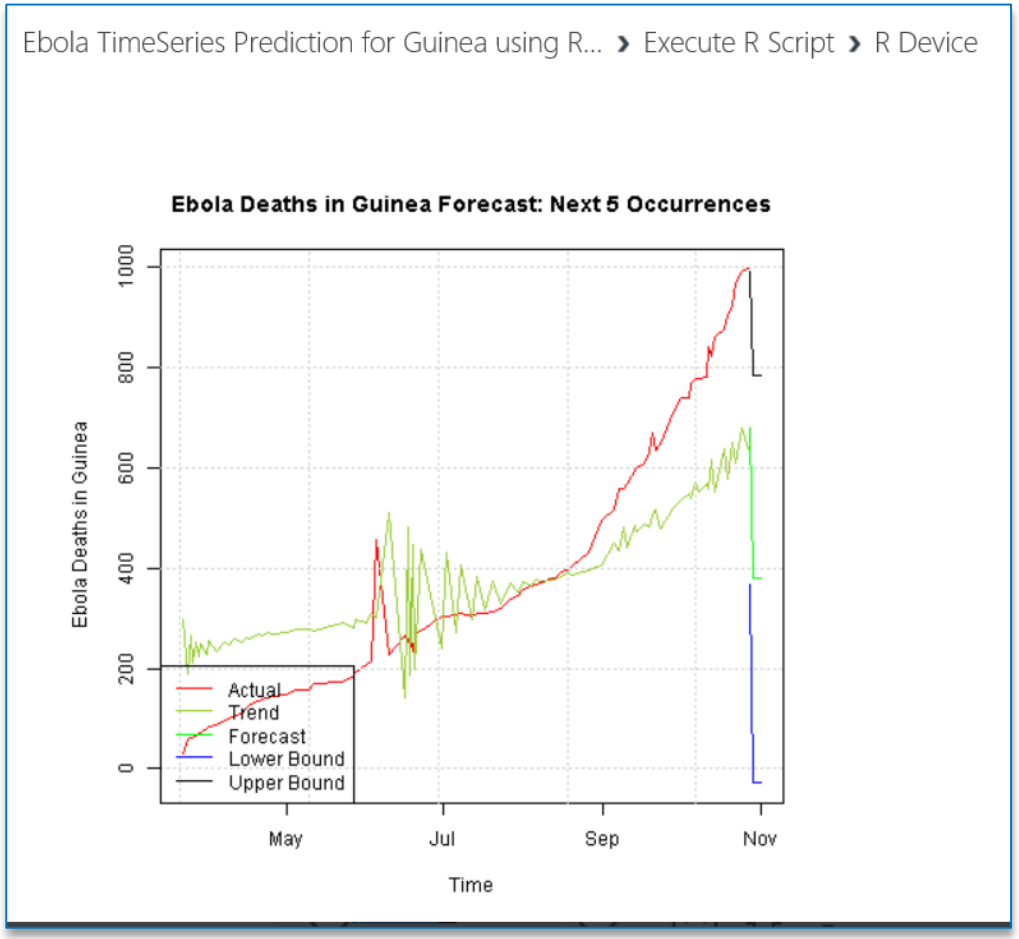
Demo

Time Series Prediction

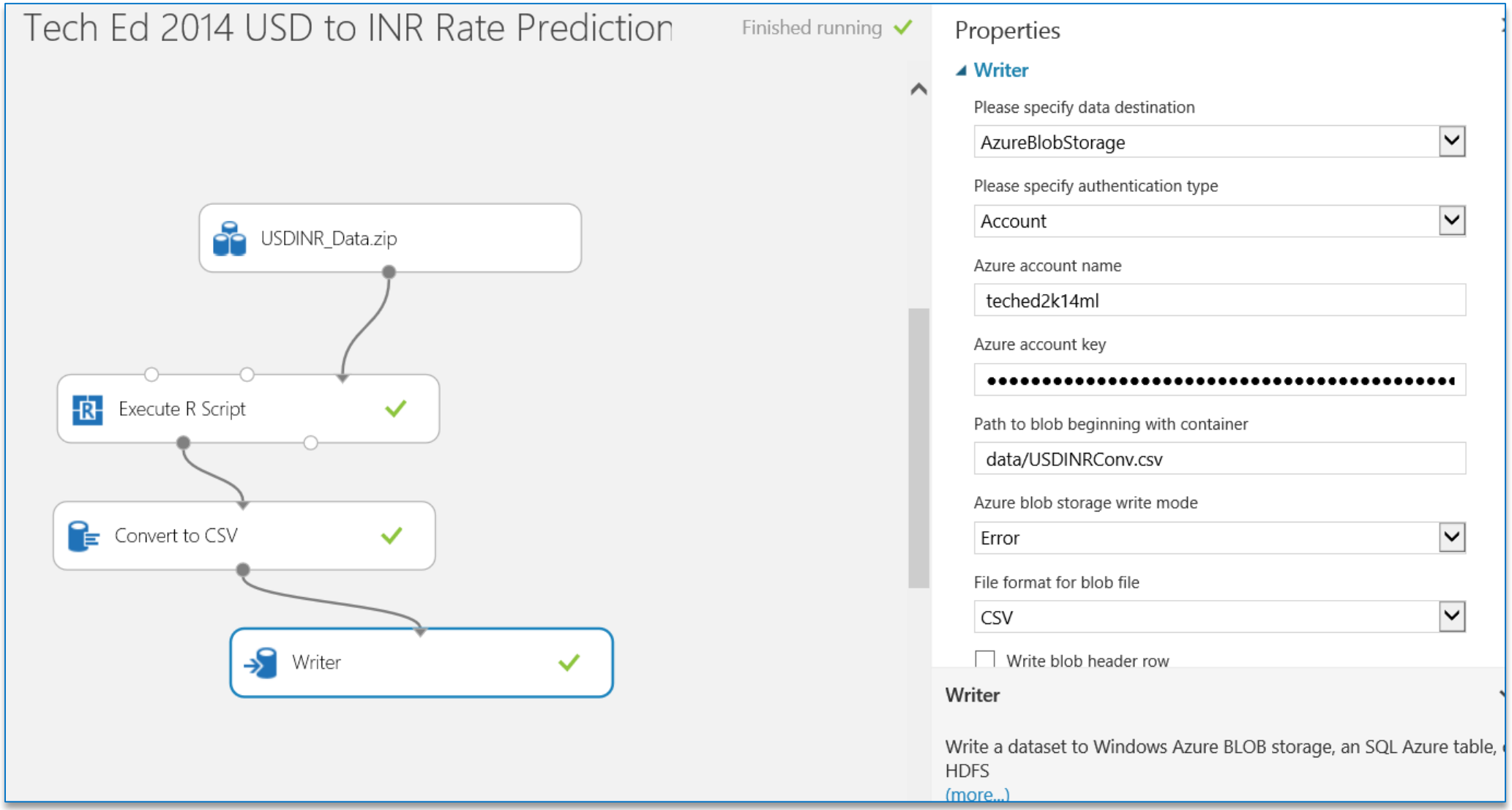
Dataset Source: https://github.com/cmrivers/ebola

Ebola Prediction





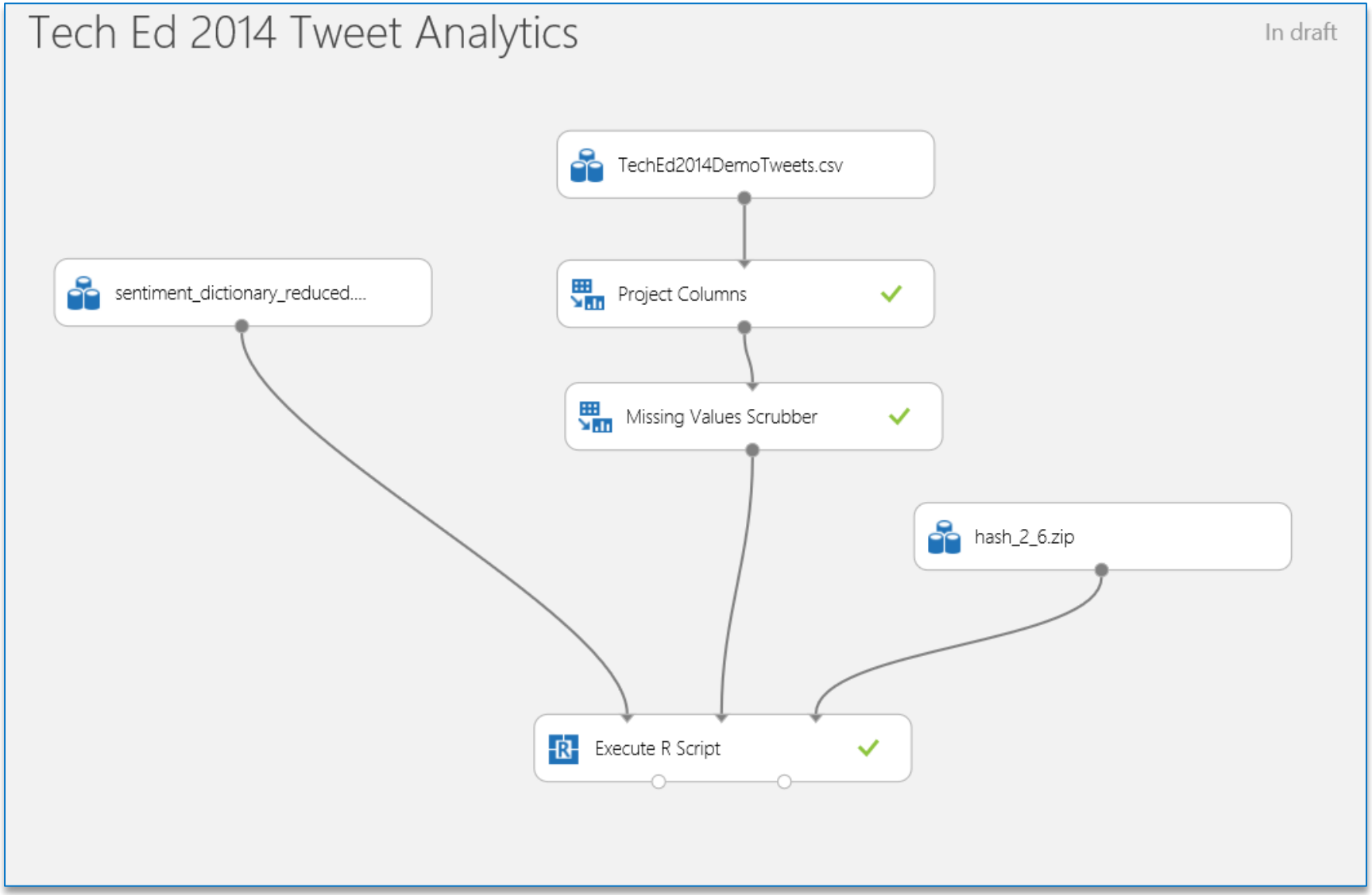
USD to INR conversion rate ARIMA/STL



Demo

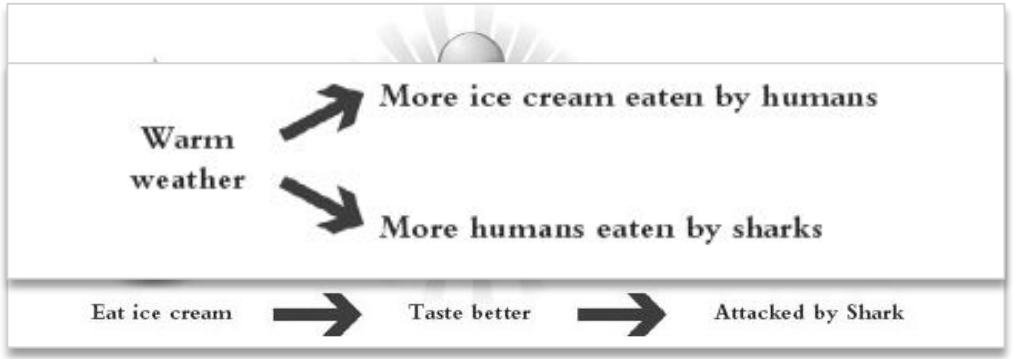
Text Analytics

Tweet Analysis

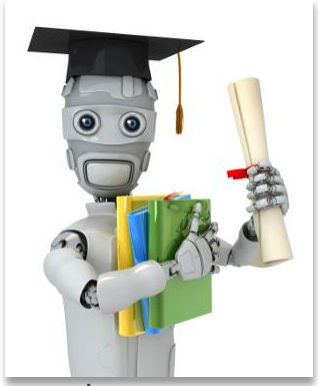


Tips for budding Data Scientists

- *Correlation* v/s *Causation*



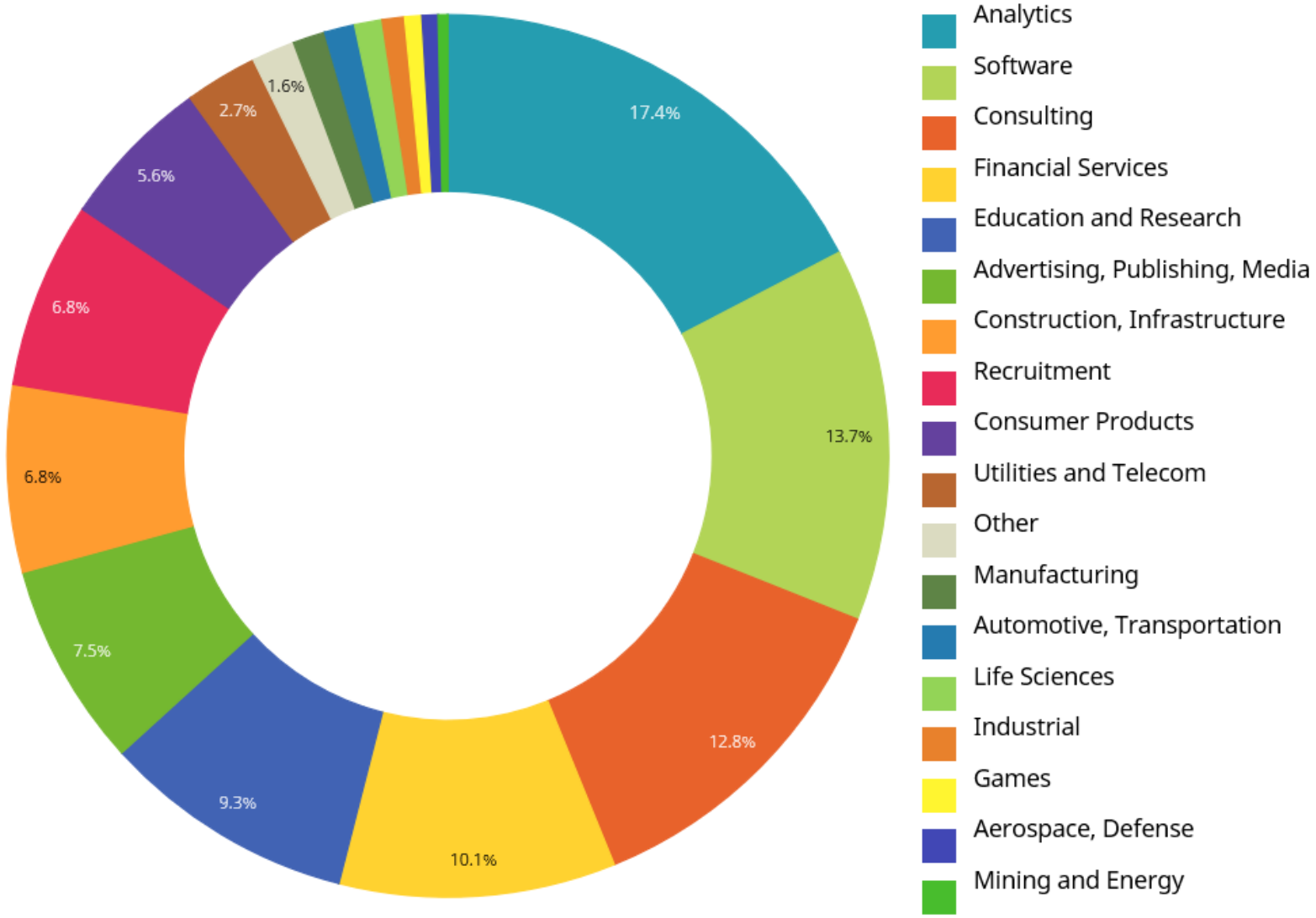
- You don't need a **Ph.D.** for this



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

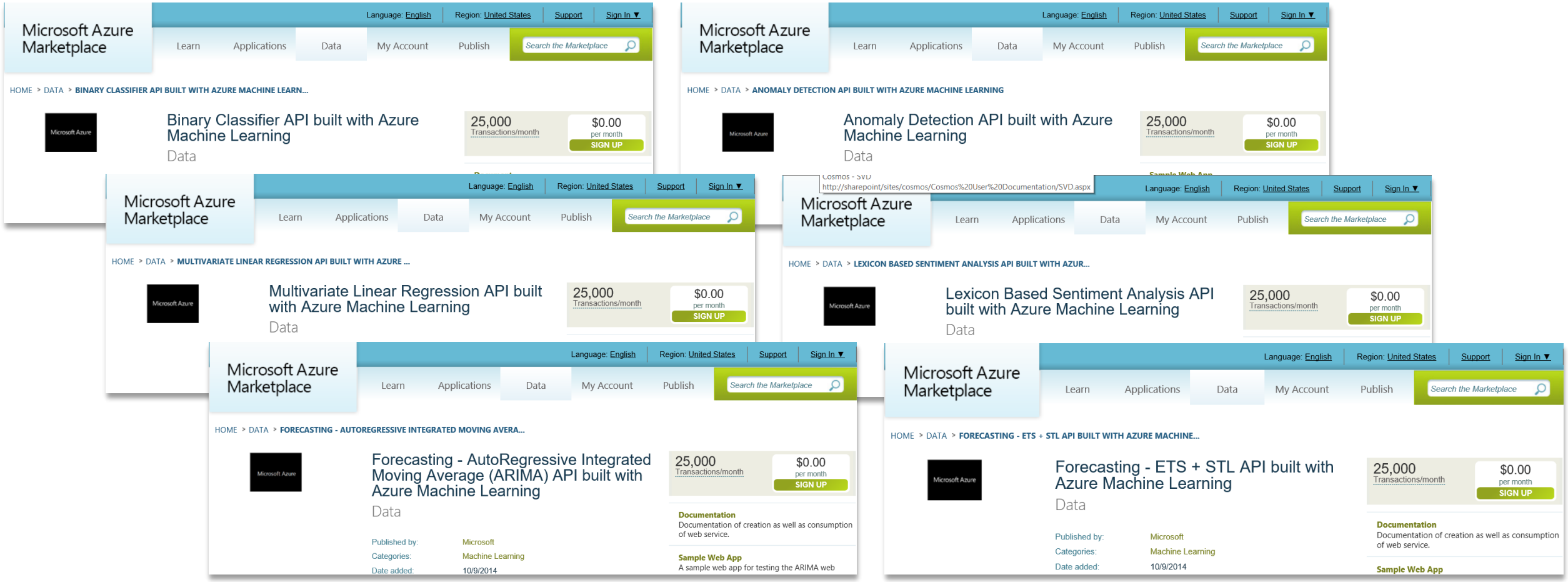
Ref:
Predictive Analytics: The power to predict who will click, buy, lie, or Die by Eric Siegel

Data Scientist Jobs ☺



https://data-scientists-count.silk.co/explore/donutchart/collection/data-scientist-sector-breakdownv2/numeric/data-scientists/order/desc/data-scientists

What Next for ML @ Microsoft



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

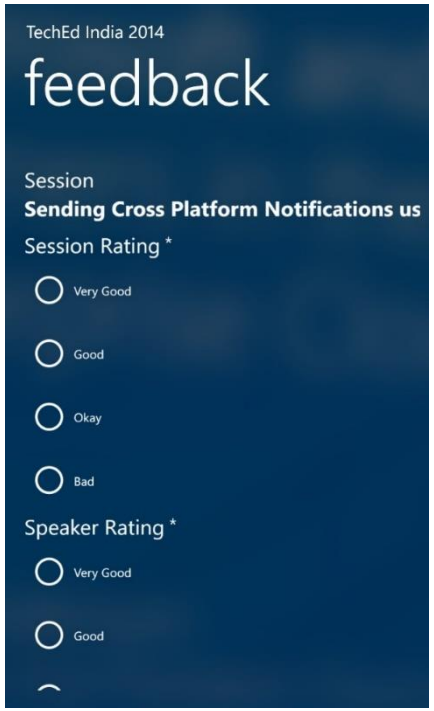
Fill out evaluation of this session and help shape future events.

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



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

OPTION 3: **Feedback stations** outside the hall

TechEd India 2014

Microsoft

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)
Searn (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

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