BUCHANAN® & EDWARDS

Event: The Team Data Science Process

Presenter: Jon Tupitza, CTO Architect

BEOP.CTO.TP4
Owner: OCTO
Revision: 0001
Approved by: JAT
Effective: 08/30/2018

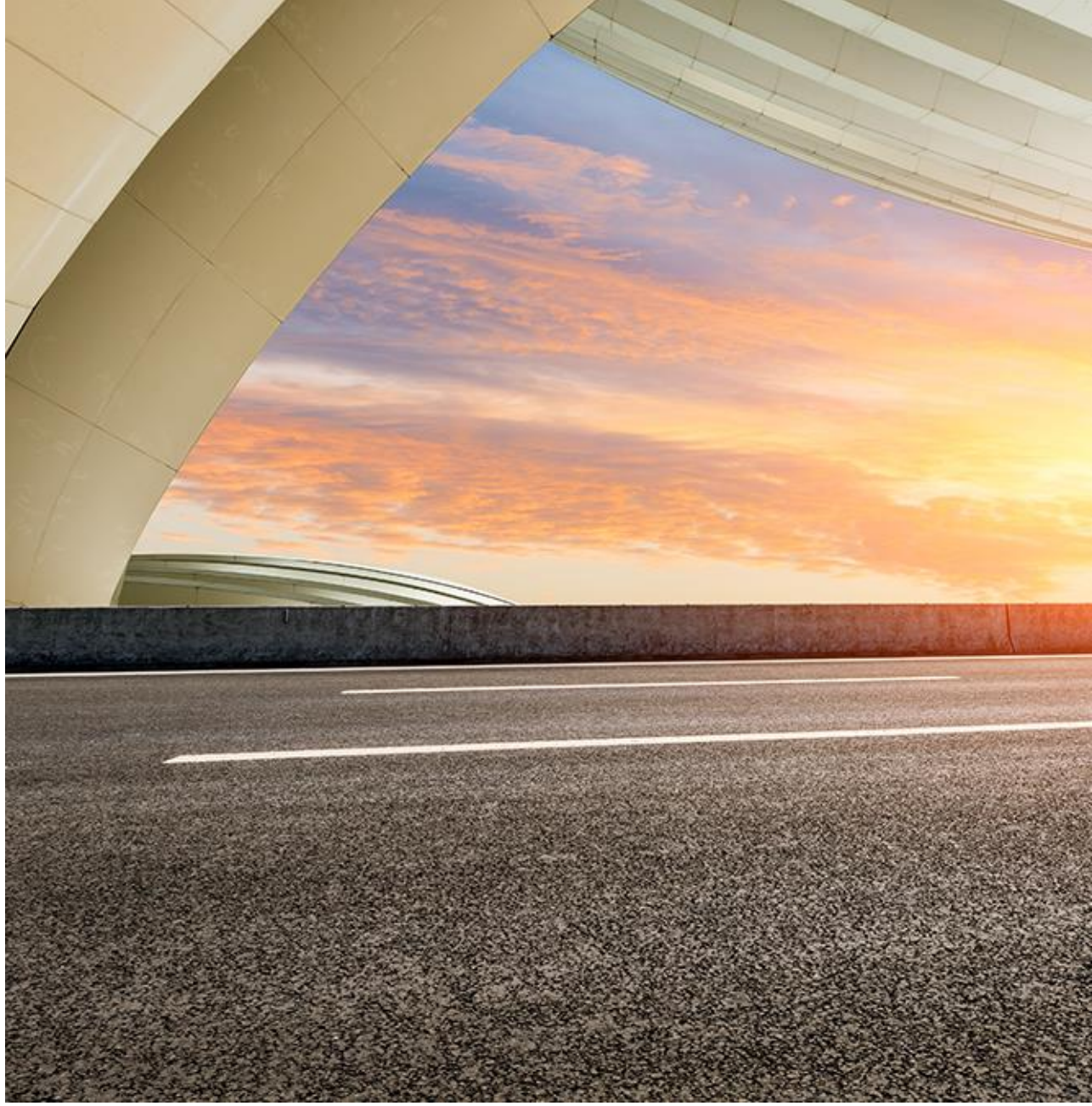
Implementing the Microsoft
Team Data Science Process

Developing Predictive Analytics using Python & Scikit-Learn

Jon Tupitza
Practice Director, Data Platform & Predictive Analytics

Take-Aways

- Understand the Microsoft Team Data Science Process
- Understand How to Implement Predictive Analytics using Python

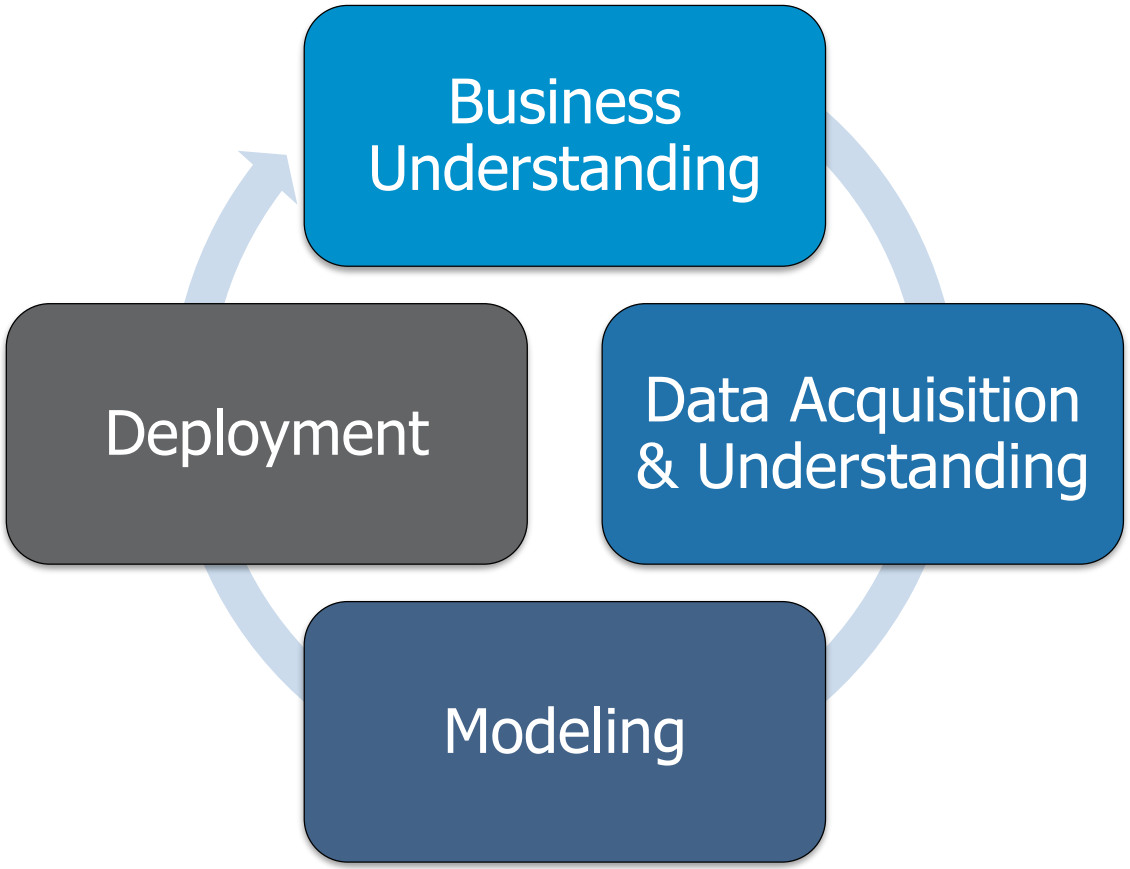


Agenda

- Introduction to the Microsoft Team Data Science Process
- Demonstrations:
  - Acquiring and Preparing Data
  - Exploring and Analyzing Data
  - Selecting Features
  - Reducing Dimensionality
  - Training, Testing & Evaluating Machine Learning Models
  - Using Pipelines
- Deployment Options

The Microsoft Team Data Science Process

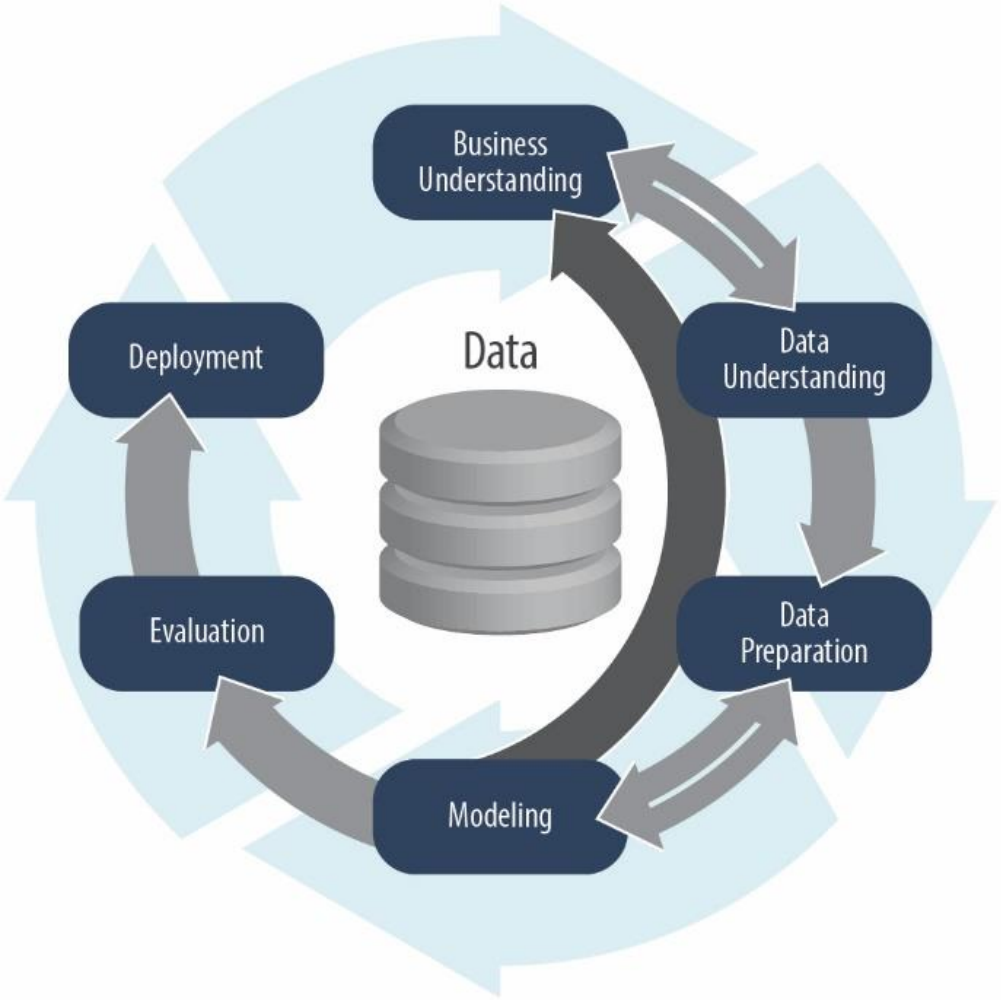
Largely Heuristic! Based on Conducting Experiments (i.e., Scientific Method)



Business Understanding	• Identify the Problem Domain • Identify the Solution Scenario
Acquire & Understand Data	• Load, Prepare & Explore Data • Identify Influential Features
Develop Machine Learning Models	• Select & Engineer Features • Train, Evaluate & Tune Models
Deployment	• Publish Models as Webservices • Consume Models Visually and Programmatically

CRISP-DM: Cross Industry Standard Process-Data Mining

This Seems Pretty Familiar: First Introduced in 1996!



Business Understanding	• Identify the Problem Domain • Identify the Solution Scenario
Data Understanding	• Load and Explore Data • Identify Influential Features
Data Preparation	• Remove Duplicates & Nulls • Impute Missing Values • Select & Engineer Features
Modeling	• Train Models Using a Variety of Algorithms • Tune Hyper-parameters
Evaluation	• Test Models' Performance & Predictive Power • Cross-Validate to Appraise Goodness-of-Fit • Select Most Effective Model for Deployment
Deployment	• Publish Models On-premises or in the Cloud • Consume Models Visually & Programmatically

Agenda

- Introduction to the Microsoft Team Data Science Process
- Demonstrations:
  - Acquiring and Preparing Data
  - Exploring and Analyzing Data
  - Selecting Features
  - Reducing Dimensionality
  - Training, Testing & Evaluating Machine Learning Models
  - Using Pipelines
- Deployment Options

Deployment: Operationalizing Machine Learning Models

- On-Premises:
  - Microsoft SQL Server 2016/2017 Machine Learning Services

- In the Cloud:
  - RESTful Web Service Endpoints
  - HDInsight with Hive
  - Apache Spark / Azure Databricks
  - Azure Container Registry
  - Azure Container Service with Kubernetes

On-Premises: SQL Server Machine Learning Services

- The First Commercial Database Server with Built-In Artificial Intelligence
- Enables Developers to Train, Evaluate and Deploy Machine Learning Models Inside of SQL Server Databases for Enterprise Production

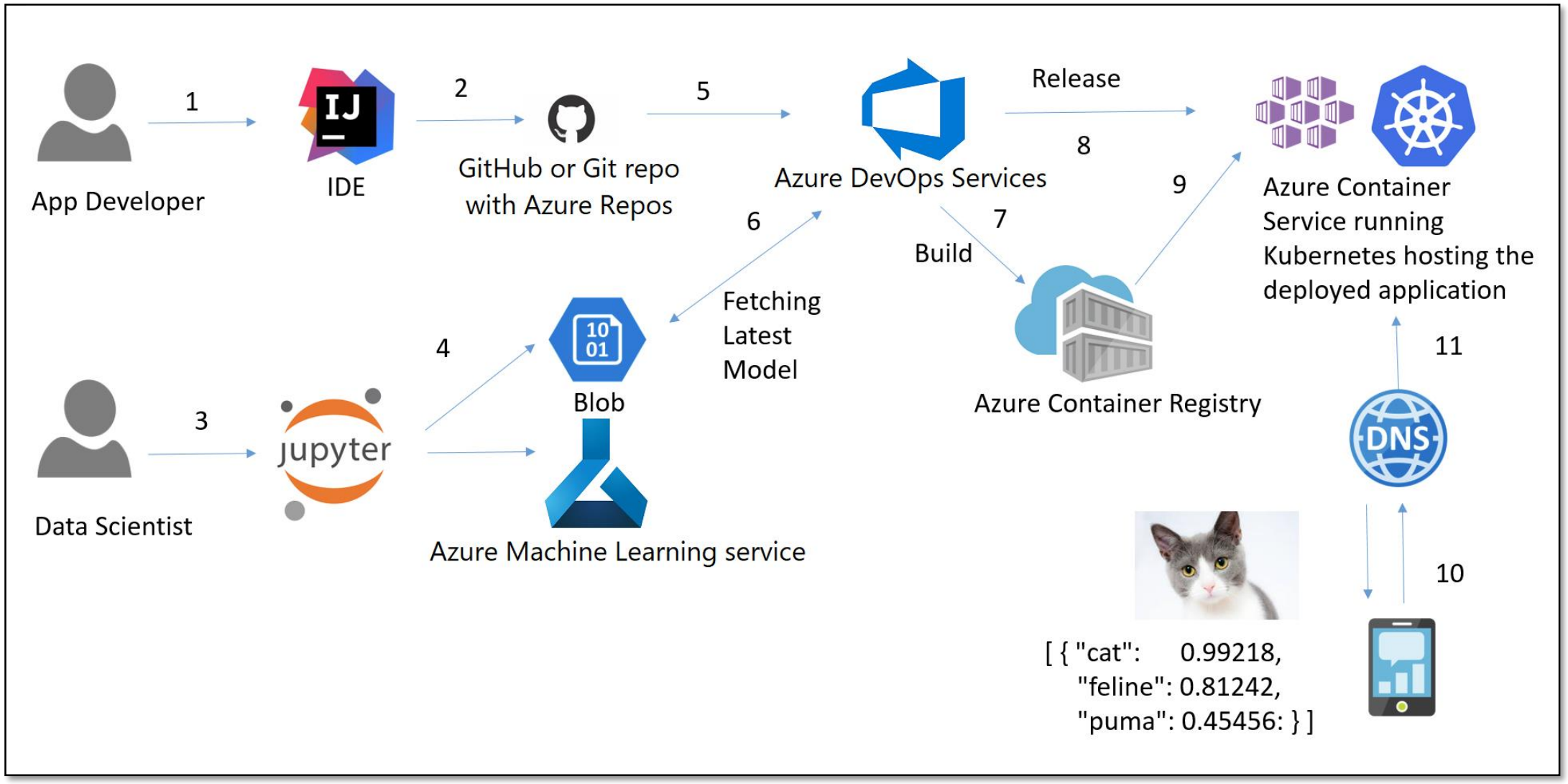
Overcomes Some Major Limitations Inherent to Statistical Software

- System Memory has been limited to the capacity of client workstations
- Data Movement has been saturating networks between remote storage and the development environment
- Performance and Scale have been limited by a lack of multi-threading and parallel processing capabilities

Provides a Convenient Way to Operationalize Machine Learning

- Access ML Algorithms using familiar T-SQL stored procedures
- Manage Machine Learning Models in SQL Server database tables
- Store Predictive Outcomes in SQL Server database tables
- Leverage database mechanisms like security, governance and monitoring

In the Cloud: Continuous Integration & Deployment



Resources

- Python Documentation
- Scikit-Learn Documentation
- Microsoft Docs:
  - Team Data Science Process
  - Tutorials for SQL Server Machine Learning Services
- Microsoft Machine Learning Server Blog:
  - Basics of R and Python Execution in SQL Server

Questions