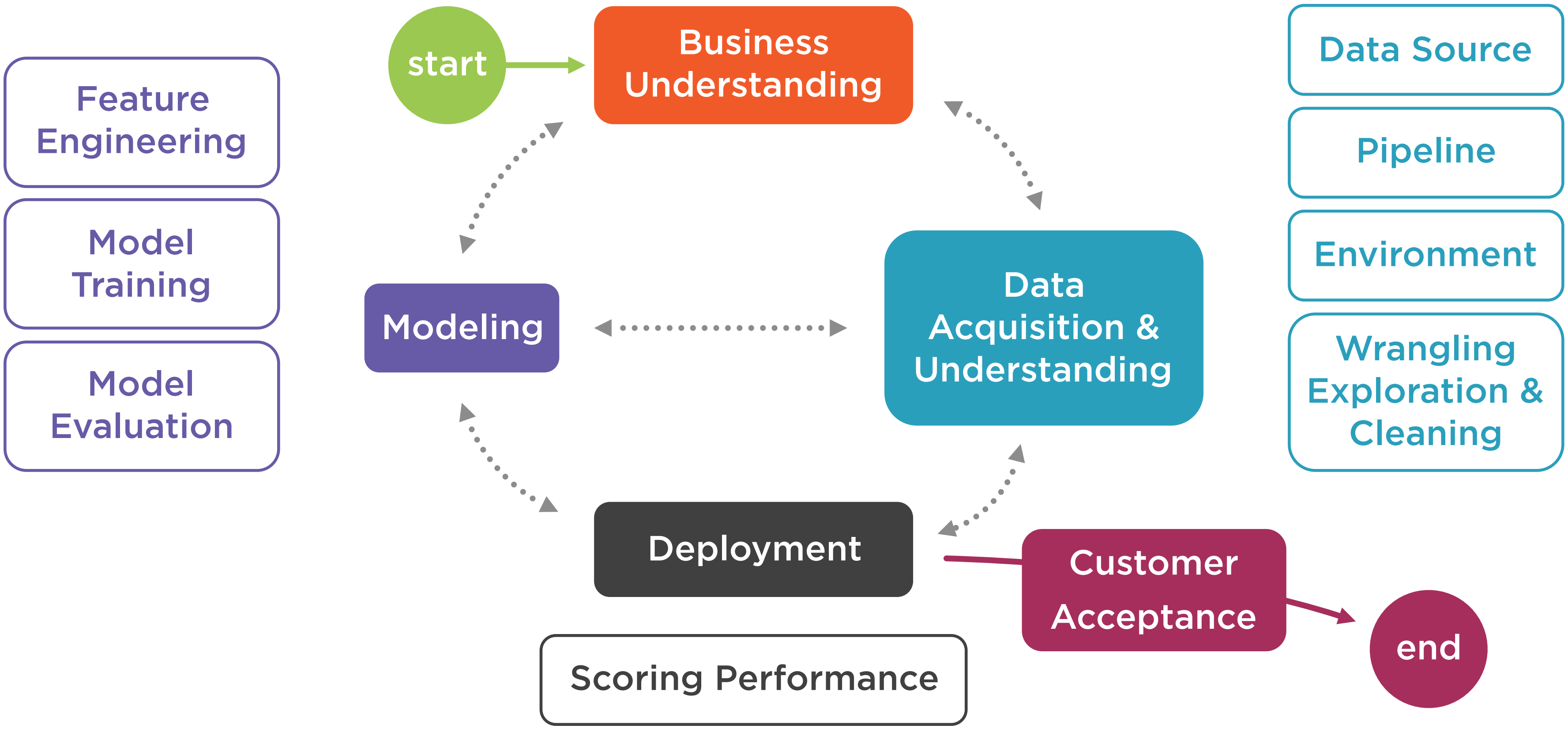
Understanding Data Ingestion Strategies

David Tucker
TECHNICAL ARCHITECT & CLOUD CONSULTANT
@_davidtucker_ www.davidtucker.net

Data Science Lifecycle



Feature Engineering

Model Training

Model Evaluation

start

Business Understanding

Modeling

Data Acquisition & Understanding

Deployment

Scoring Performance

Customer Acceptance

end

Data Source

Pipeline

Environment

Wrangling Exploration & Cleaning

Azure Machine Learning Solution Components

Workflow

Data Pipeline

Infrastructure

Overview

Review the data exploration utilities included with the TDSP

Create a data report utilizing the IDEAR notebook

Introduce machine learning pipelines for Azure ML

Create and use a machine learning pipeline

Review integrations for pipelines in other Azure services

Data Exploration & Reporting

Data Utilities with the TDSP

IDEAR

Interactive Data Exploration, Analysis and Reporting

AMR

Automated Modeling and Reporting in R

Tools & Utilities

Interactive Data Exploratory Analysis and Reporting (IDEAR)

- R
- MRS
- Python

Automated Modeling and Reporting in R (AMR in R)

IDEAR Utility

Aids in the data understanding phase by enabling you to visualize and analyze a data set and its correlations. This tool should enable you to fine tune your hypothesis.

IDEAR Utility with Python

Delivered through a Jupyter notebook

Two versions are provided:

- Specific version for Azure Notebooks
- General version that can be used anywhere

Requires specific modules to be installed prior to utilizing the notebook

IDEAR Capabilities

Preview of dataset

Statistics of numerical columns

Overview of categorical columns

Visualizations for variable correlations

Utilizing the IDEAR Data Tool

Demo

Uploading exercise files for the module

Loading sample data into Azure Blob Storage for the workspace

Downloading the IDEAR Jupyter notebook

Running the IDEAR notebook to gain insights on the sample data set

Machine Learning Pipelines

> “An **Azure Machine Learning pipeline** is an independently executable workflow of a complete machine learning task. Subtasks are encapsulated as a series of steps within the pipeline.”

Azure ML Documentation

Azure ML Pipelines

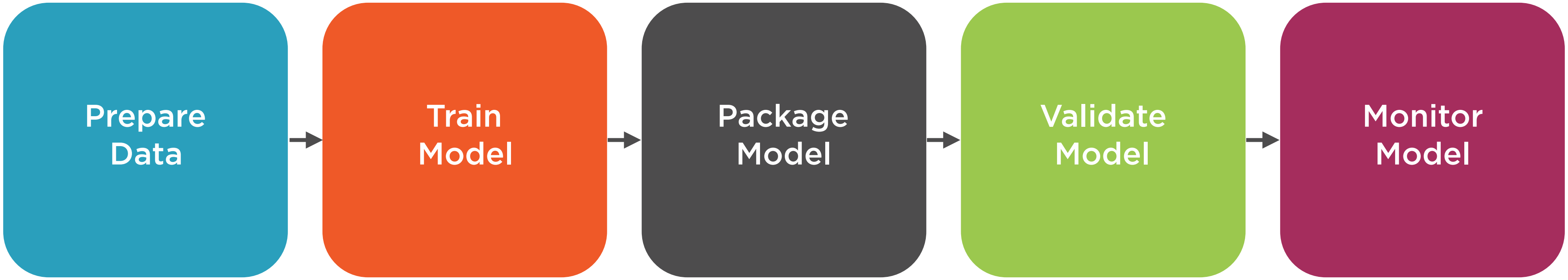
Connects to a single Azure ML Experiment

Automates common steps in machine learning iteration

Provides versioning and tracking of step inputs and outputs

Enables overall modularity to steps within the workflow

Example Pipeline Steps



Azure ML Pipeline Steps

Each step can utilize a separate infrastructure configuration

AzureML Python SDK provides several step items:

- EstimatorStep
- PythonScriptStep
- DataTransferStep
- ModuleStep

Dependencies are evaluated dynamically

Triggering Pipelines

Manually Triggered

Scheduled
(Time)

Scheduled
(File Change)

Best Practices for Pipelines

Pipeline steps should be broken down into independent steps

Avoid tight coupling between steps

Once you transition to iterating on a hypothesis, transition to using a pipeline

Creating a Machine Learning Pipeline

Demo

Configuring dependencies for pipeline steps

Creating references to pipeline data inputs and outputs

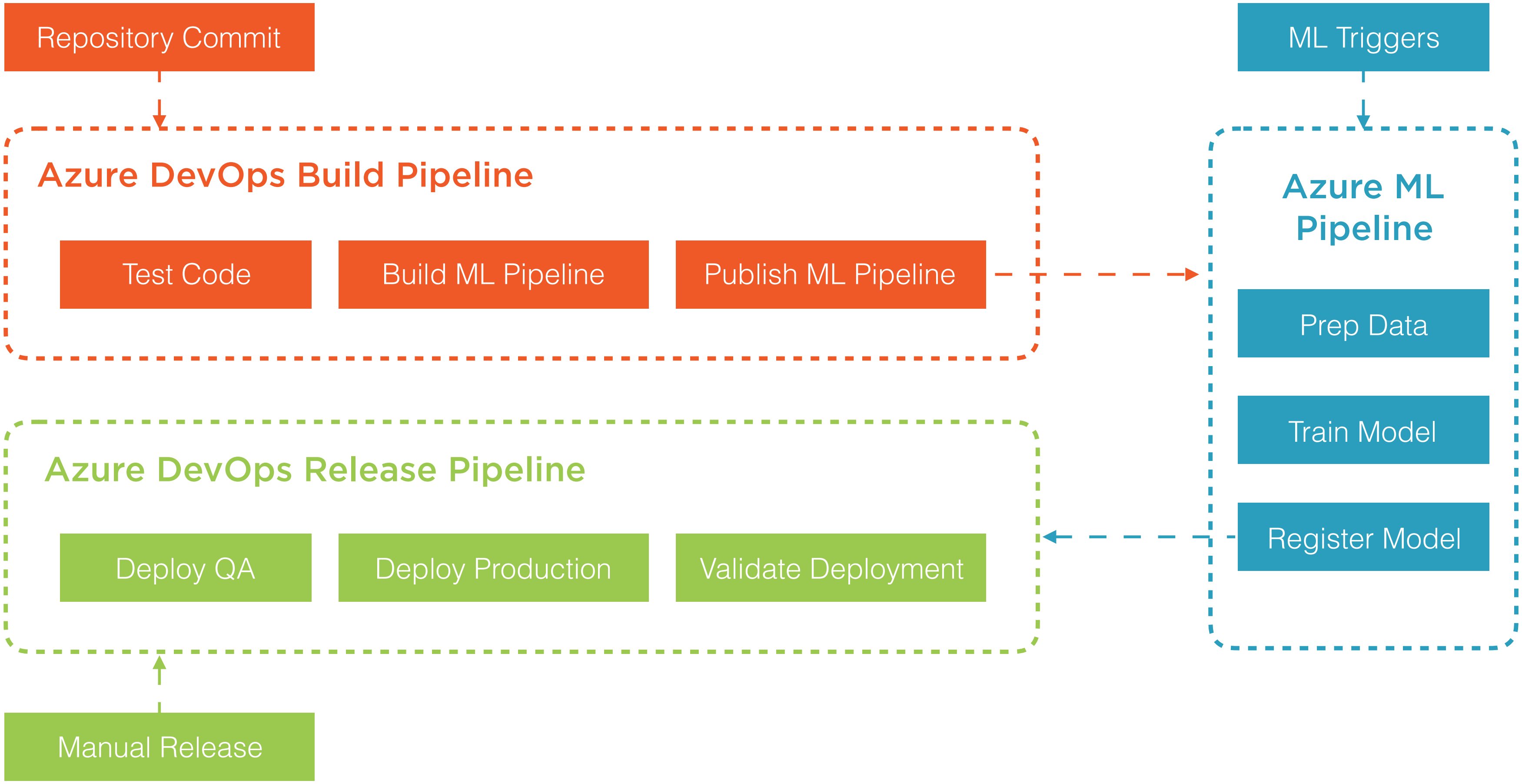
Creating pipeline steps for data prep, model training, and registering model

Publishing and executing the pipeline

Utilizing the included component to review pipeline execution progress

Azure Integration for Machine Learning

Complete Machine Learning Lifecycle



Azure Service Integration

Azure DevOps can support both the build and release pipelines

Build pipeline can trigger the Azure ML pipeline execution

Release pipeline can be triggered by saving a new model into the workspace

Summary

Summary

Reviewed the data exploration utilities included with the TDSP

Created a data report utilizing the IDEAR notebook

Introduced machine learning pipelines for Azure ML

Created and used a machine learning pipeline

Reviewed integrations for pipelines in other Azure services