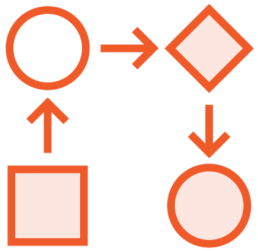
Designing Machine Learning Solutions on Microsoft Azure

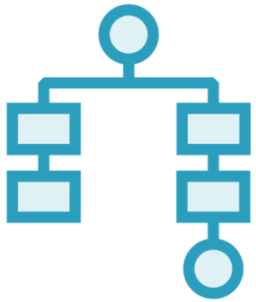
UNDERSTANDING THE AZURE MACHINE LEARNING WORKFLOW

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

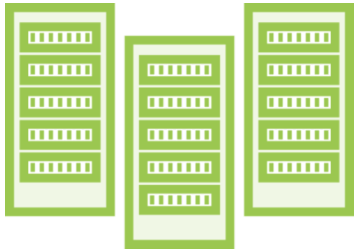
Azure Machine Learning Solution Components



Workflow



Data Pipeline



Infrastructure

Overview

- Review current challenges for implementing data science best practices
- Understand the benefits provided by the Team Data Science Process (TDSP)
- Review resources provided as a part of the TDSP
- Create an Azure Machine Learning Service workspace
- Provision compute infrastructure for Azure ML

Data Science Challenges

Data Science Disciplines

Computer Science

Data Management & Mining

Research & Domain Expertise

Mathematics & Statistics

Computer Science Best Practices

Version Control

Method to track team project assets over time

Continuous Delivery

Automated process for building, testing, and deploying a project

Scrum Methodology

Agile framework for how teams plan and execute work

Data Mining Processes

CRISP-DM
Cross Industry Standard Process for Data Mining

KDD
Knowledge Discovery in Databases

“87% of data science projects never make it into production.”

Venturebeat

Machine Learning Workflow Goal

To increase the likelihood of producing actionable insights that can be operationalized within your organization by leveraging cross-discipline best practices.

Introduction to Team Data Science Process

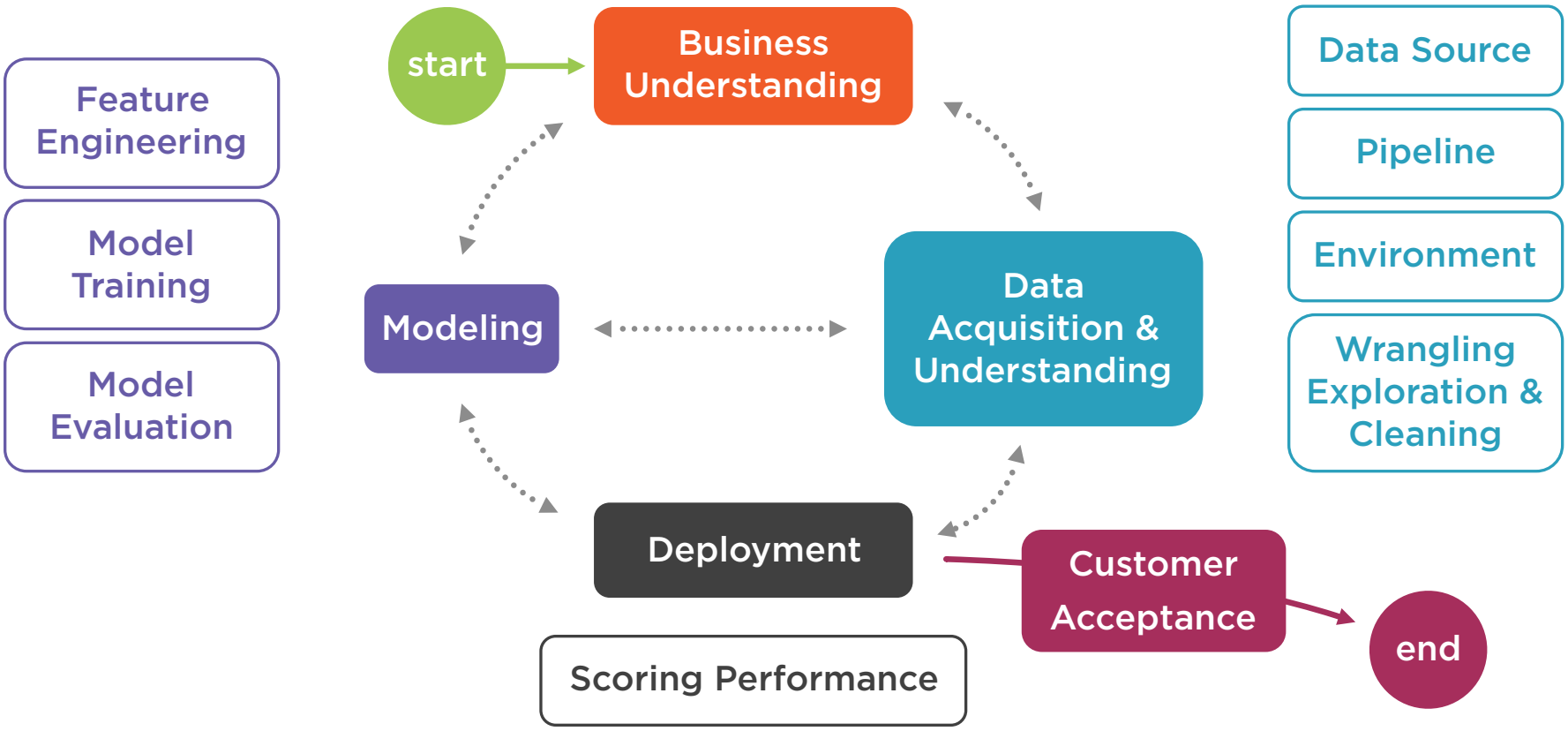
Team Data Science Process (TDSP)

An agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently.

TDSP Components

Data Science Lifecycle	Standardized Project Structure
Infrastructure and Resources	Tools & Utilities

Data Science Lifecycle



Project Structure

TDSP template available in Github

Enables the storage of all project assets for each phase in the repository

Includes templates for standard project deliverables

- Project charter
- Data reports
- Model reports
- Model performance metrics

“This infrastructure enables reproducible analysis. It also avoids duplication, which can lead to inconsistencies and unnecessary infrastructure costs. Tools are provided to provision the shared resources, track them, and allow each team member to connect to those resources securely.”

Microsoft, TDSP Documentation

Tools & Utilities

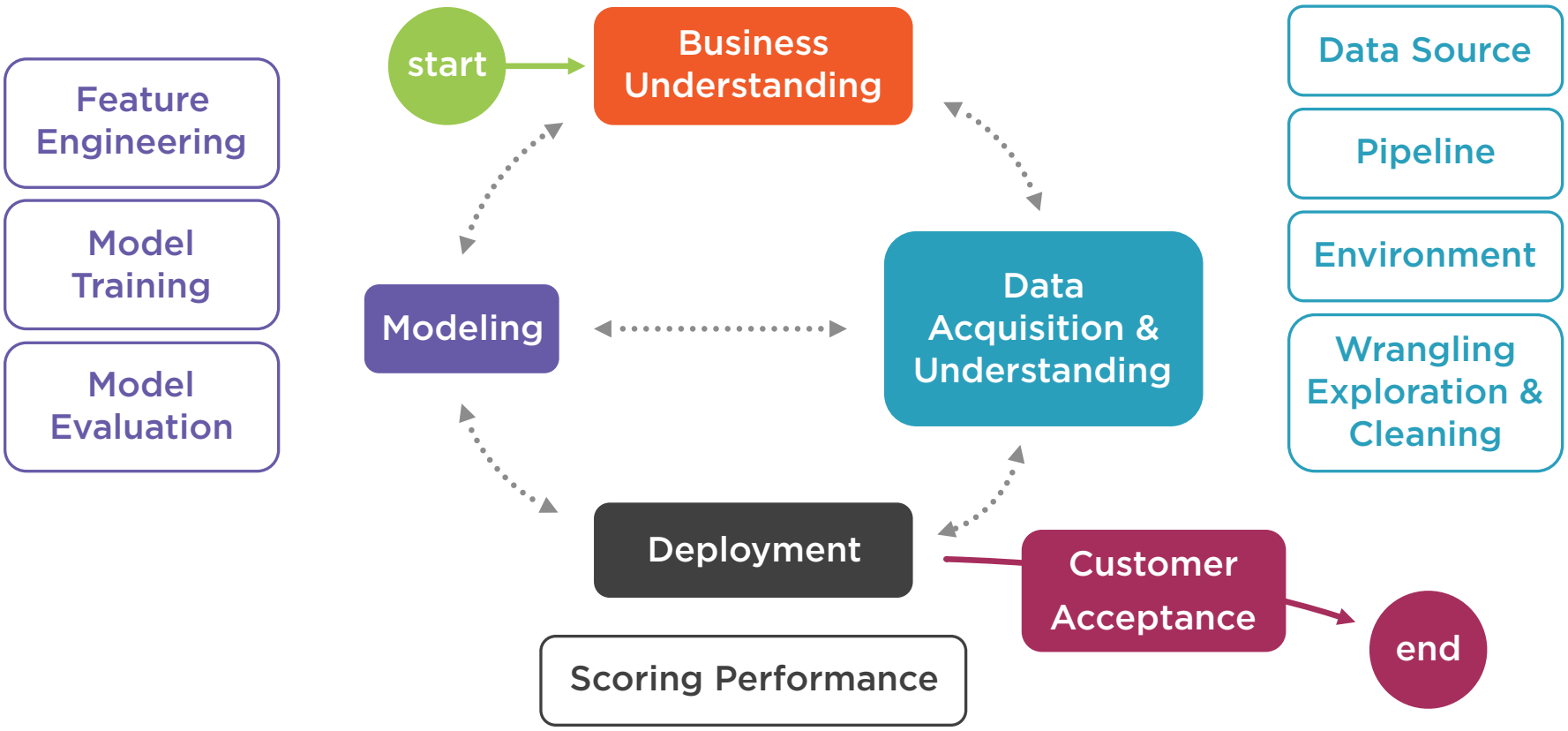
Interactive Data Exploratory Analysis and Reporting (IDEAR)

- R
- MRS
- Python

Automated Modeling and Reporting in R (AMAR in R)

TDSP Data Science Lifecycle

Data Science Lifecycle



TDSP Team Roles

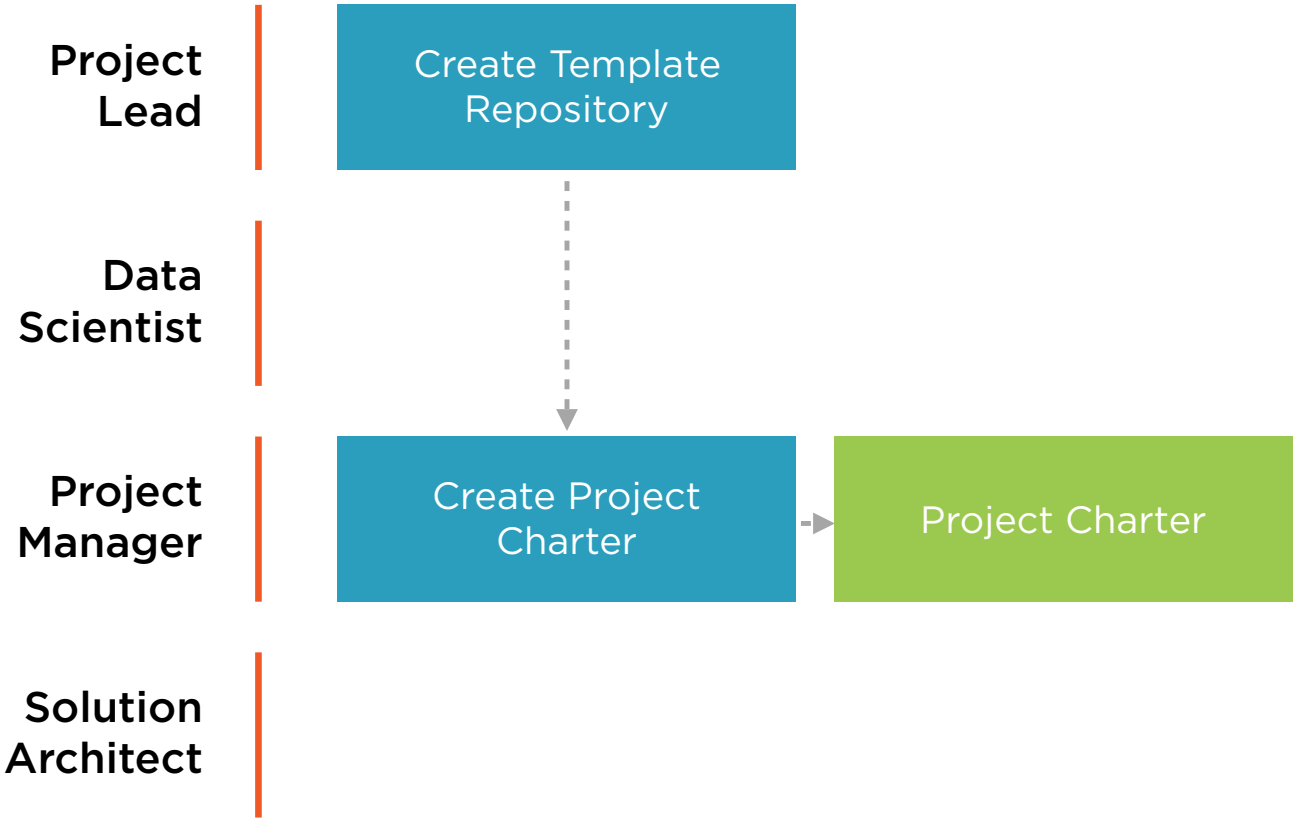
Project Lead

Data Scientist

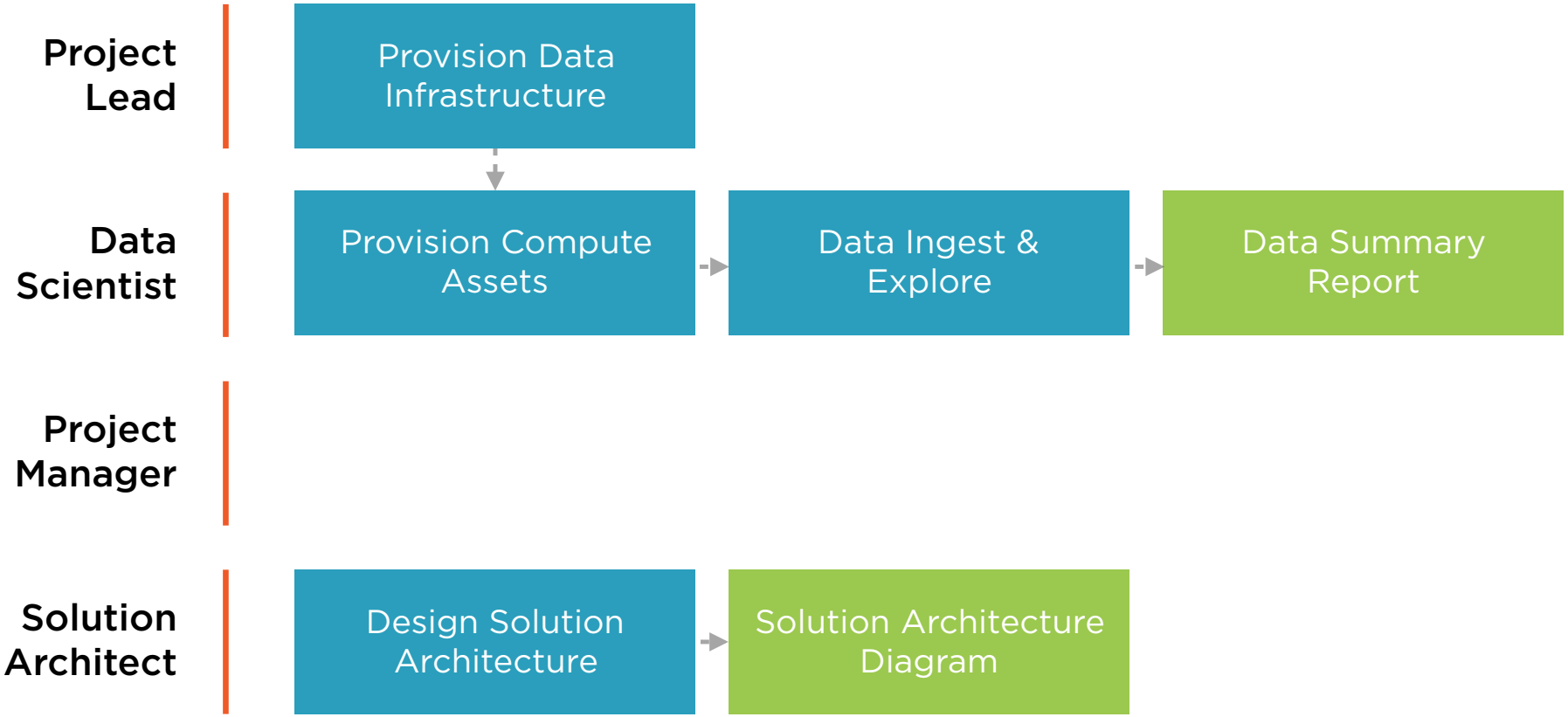
Project Manager

Solution Architect

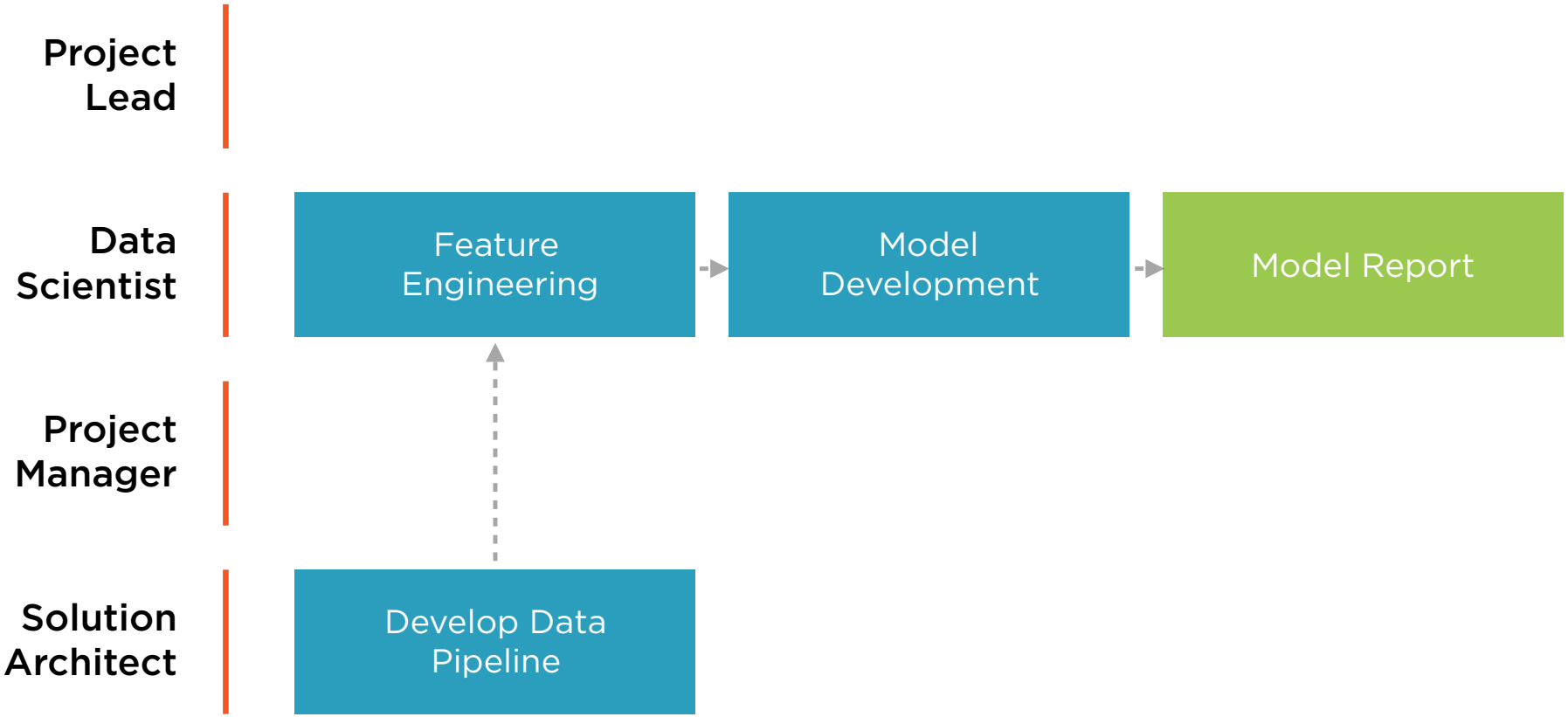
Business Understanding



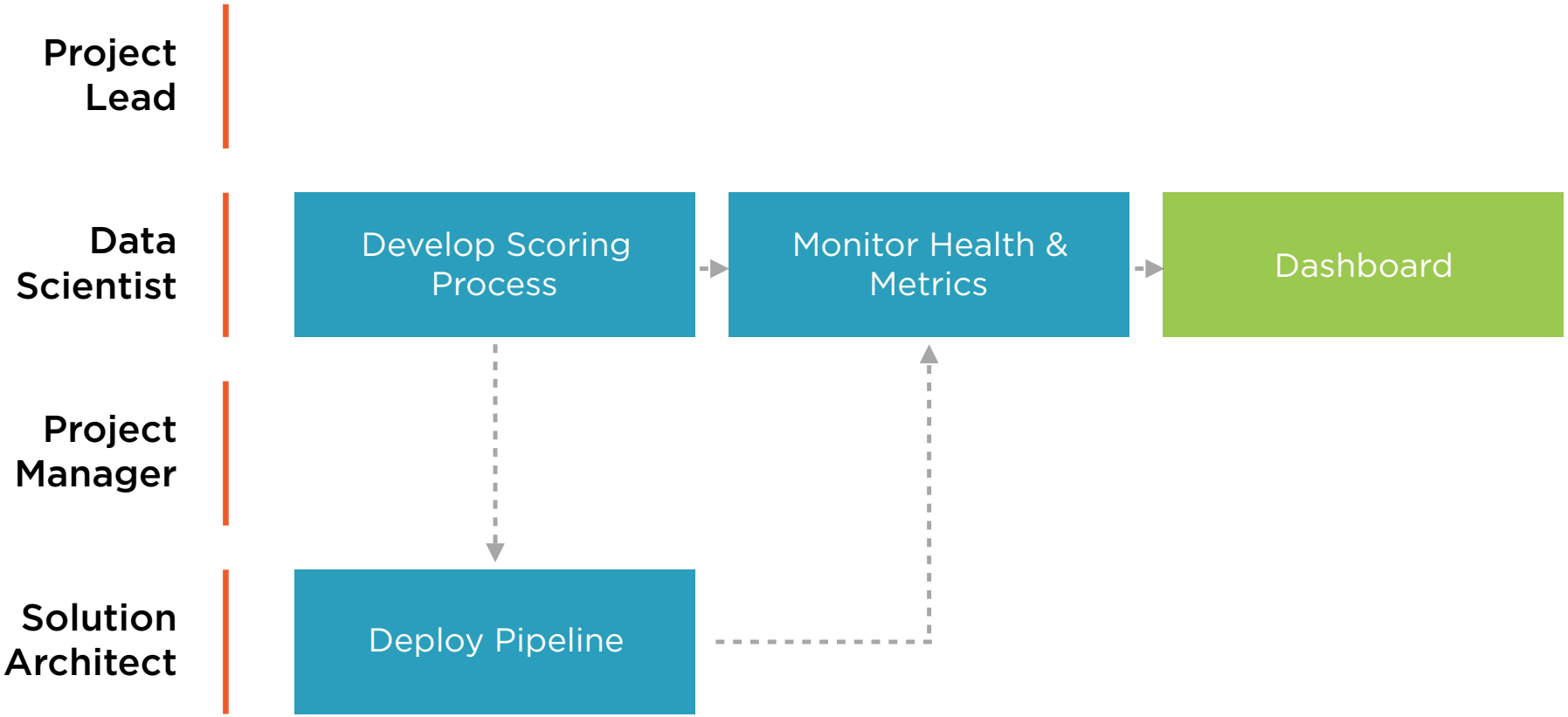
Data Ingest & Understanding



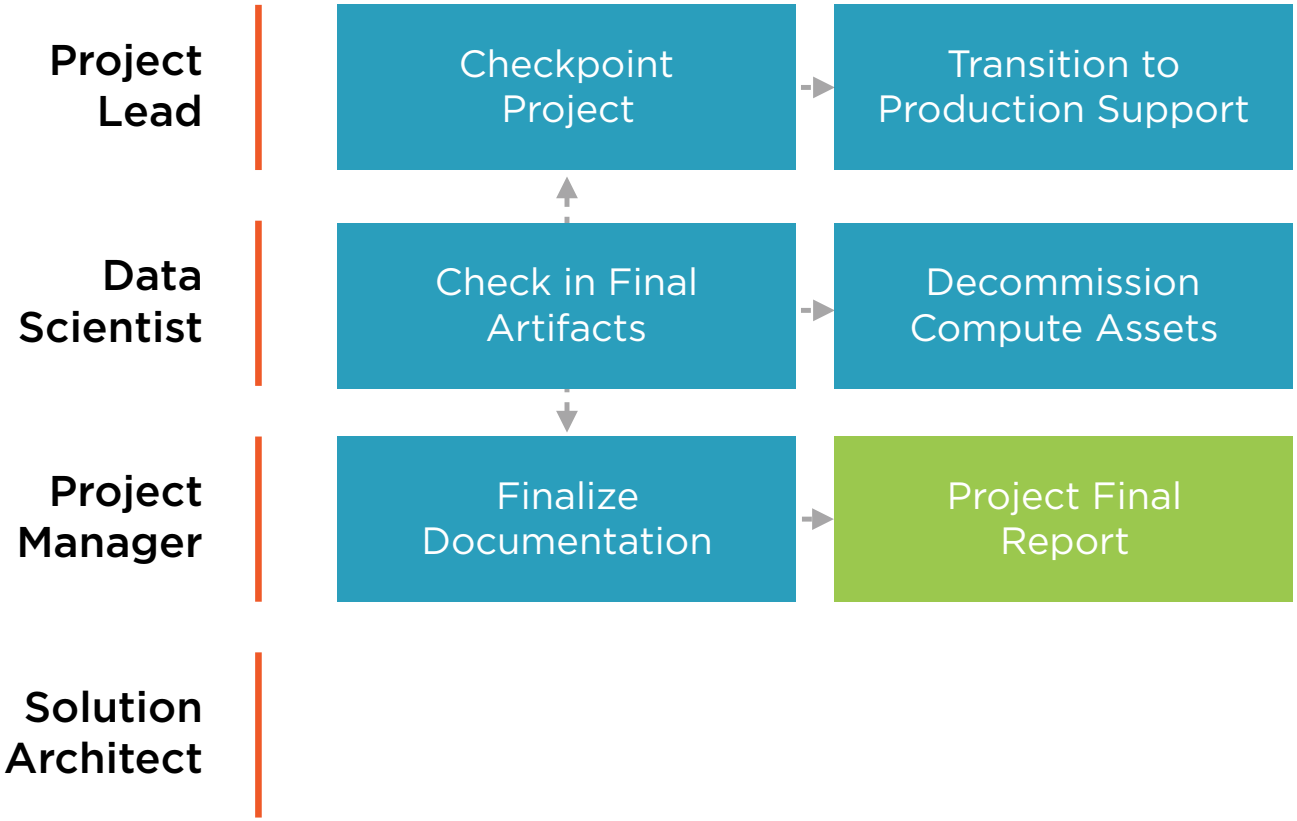
Modeling


Project Lead
Data Scientist
Project Manager
Solution Architect
Feature Engineering
Model Development
Model Report
Develop Data Pipeline

Deployment


Project Lead
Data Scientist
Project Manager
Solution Architect
Develop Scoring Process
Monitor Health & Metrics
Dashboard
Deploy Pipeline

Acceptance


Project Lead
Checkpoint Project
Transition to Production Support
Data Scientist
Check in Final Artifacts
Decommission Compute Assets
Project Manager
Finalize Documentation
Project Final Report
Solution Architect

Reviewing TDSP Resources

Demo

Review the TDSP template in Github

Review the TDSP Tools and Utilities in Github

Creating a TDSP Project

Demo

Create an Azure ML workspace

Clone the TDSP project template

Configure VS Code to work with Azure ML

Provision compute resources in Azure ML from the Python SDK

Decommission compute resources in Azure ML from the Python SDK

Summary

Overview

- Reviewed current challenges for implementing data science best practices
- Understood the benefits provided by the Team Data Science Process (TDSP)
- Reviewed resources provided as a part of the TDSP
- Created an Azure Machine Learning Service workspace
- Provisioned compute infrastructure for Azure ML