Developing and deploying AI solutions on the cloud using Team Data Science Process (TDSP) and Azure Machine Learning (AML)

Debraj GuhaThakurta
Microsoft, AI & Research
Cloud AI Platform
Algorithms and Data Science Group

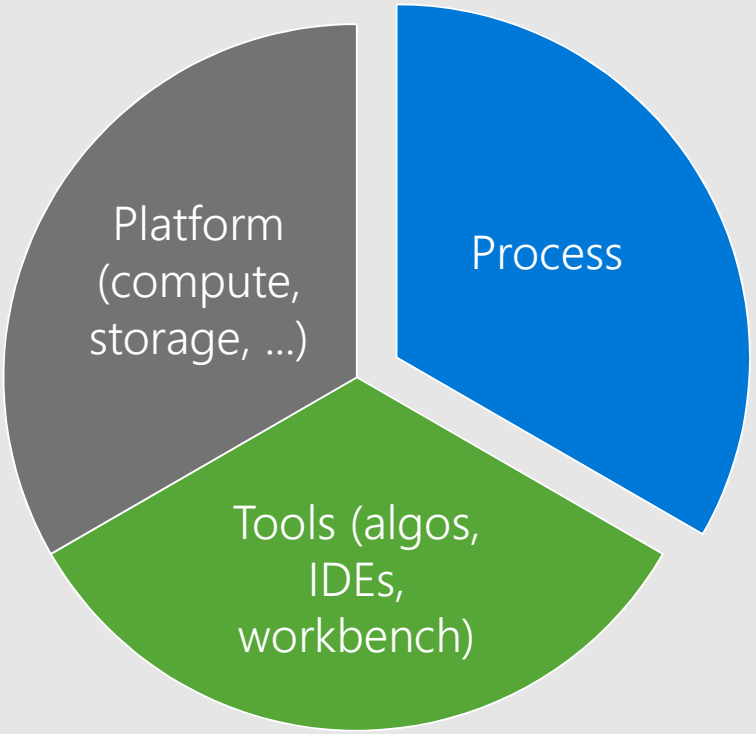
Global Artificial Intelligence Conference
January 18, 2018, Santa Clara, CA

Microsoft

Deck available @: https://aka.ms/tdsp-presentations

Agenda

- Team Data Science Process (TDSP)
  - Principle and objective
  - Key components
  - Adoption and use
- Azure Machine Learning (AML)
  - Key features
  - Developing and deploying AI solutions
- Using TDSP with AML (process + tools)
- Summary



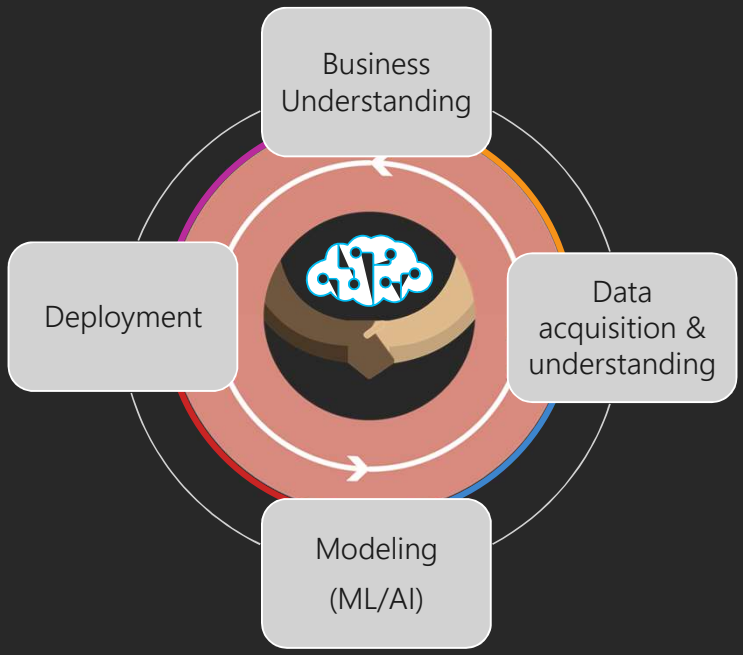
Process, tools and platforms for AI solutions

Team Data Science Process (TDSP)

Agile and iterative process to develop, deploy and manage AI applications in the cloud

https://aka.ms/tdsp

tdsp-feedback@microsoft.com



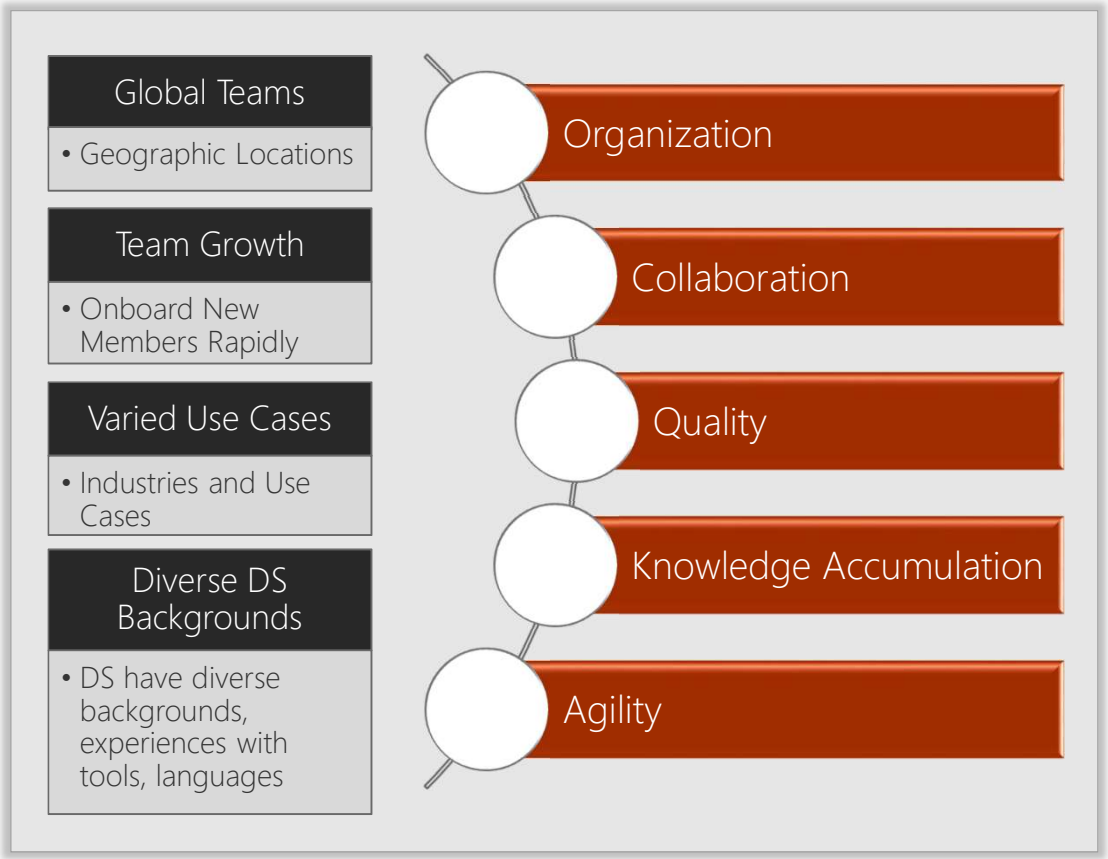
The opportunity and challenge of data science in enterprises

Opportunity: 17% had a well-developed Predictive/Prescriptive Analytics program in place, while 80% planned on implementing such a program within five years – Dataversity 2015 Survey

Challenge: Only 27% of the big data projects are regarded as successful – CapGenimi 2014

Tools & data platforms have matured -
Still a major gap in executing on the potential

One reason: Process challenge in Data Science



"Intelligent" application (ML/AI) development has unique complexity not always encountered in other Software Development scenarios

Why is a process useful for Data Science?

A process is a detailed **sequence of activities** necessary to perform specific business tasks

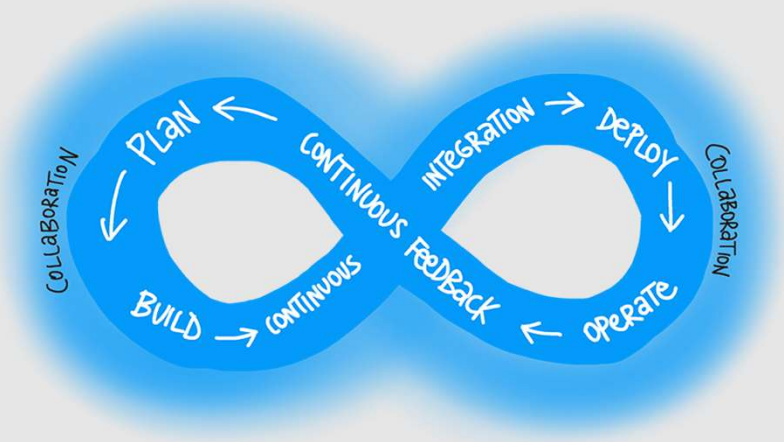
It is used to **standardize procedures** and **establish best practices**

Technology and tools are changing rapidly. A standardized process can provide continuity and stability of work-flow.
- Based on discussions with Luis Morinigo, Dir. IoT, NewSignature

Data Science can borrow processes from DevOps

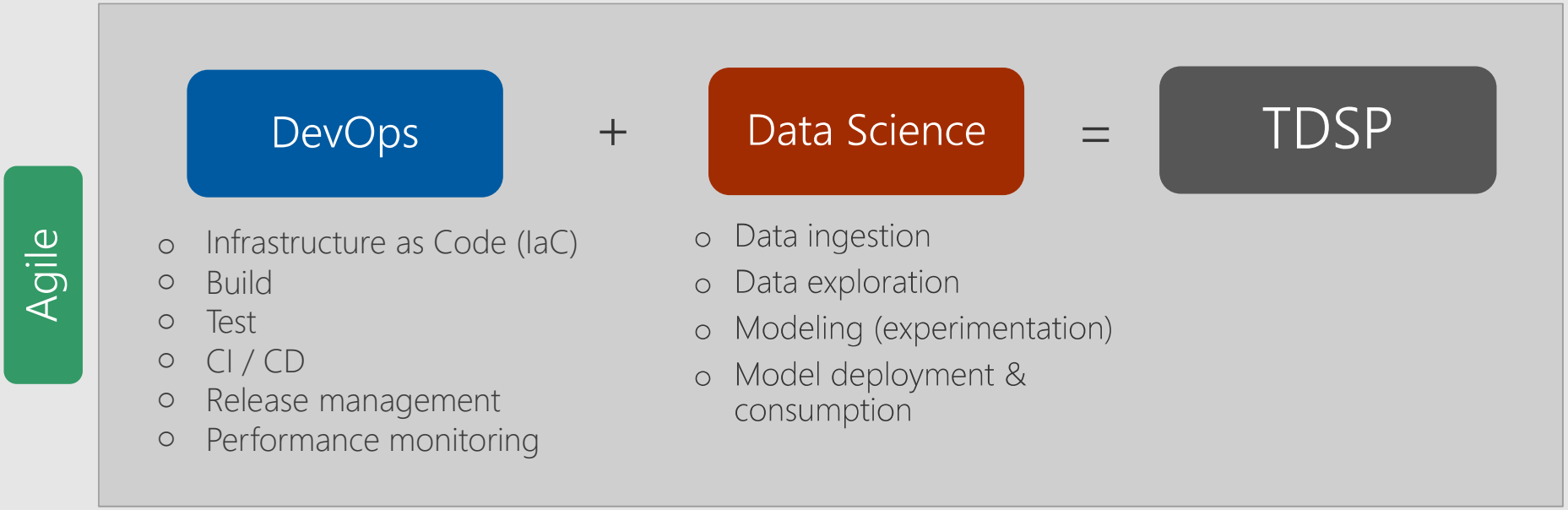
- Integrated Software Development & Operations (DevOps) has had much more time to mature, standardize, build in efficiency and develop best practices

- Data Science has unique complexity – but can learn standardized processes from DevOps



TDSP objective

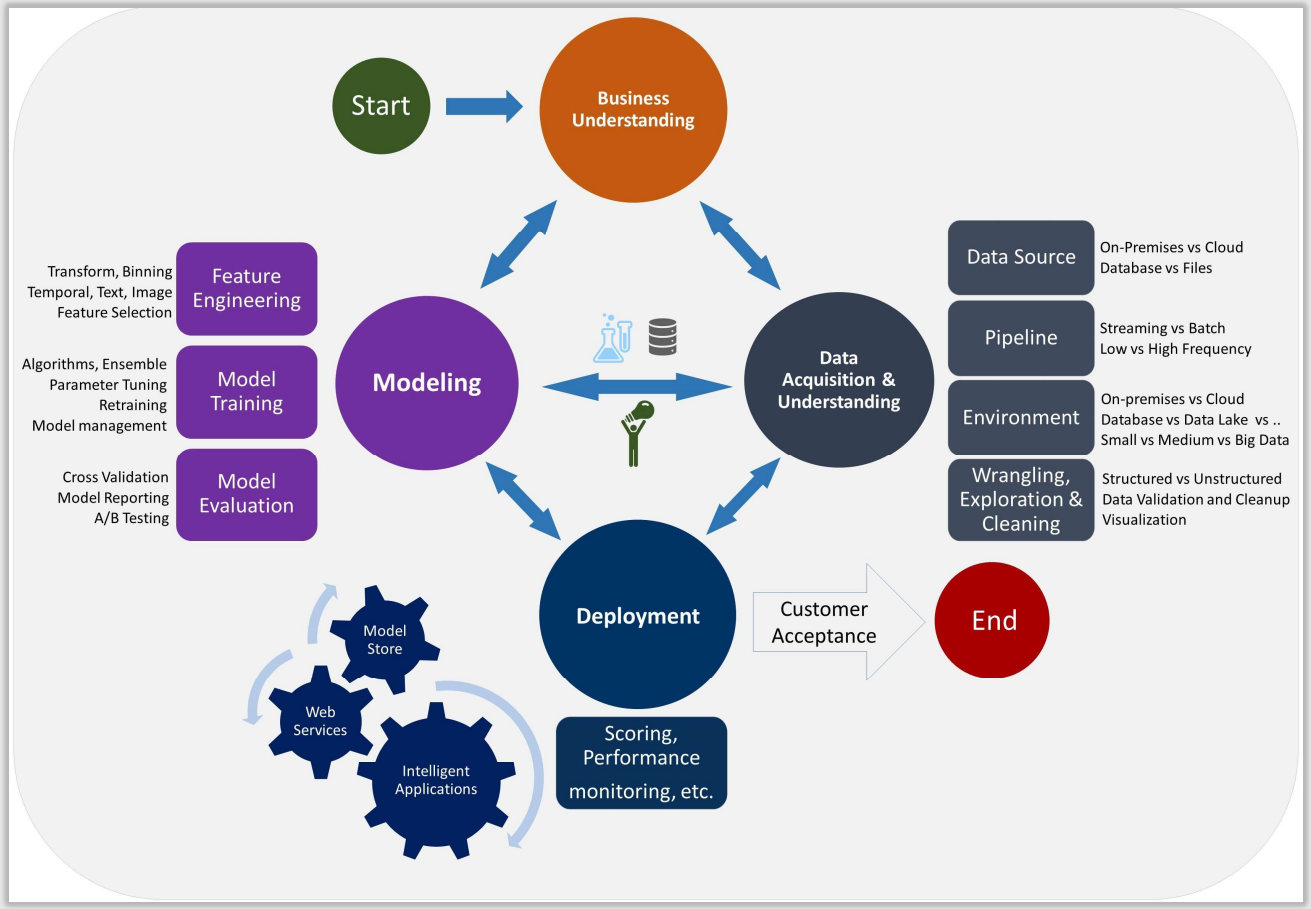
Integrate DevOps with data science workflows to improve collaboration, quality, robustness and efficiency in data science projects



TDSP features for data science teams

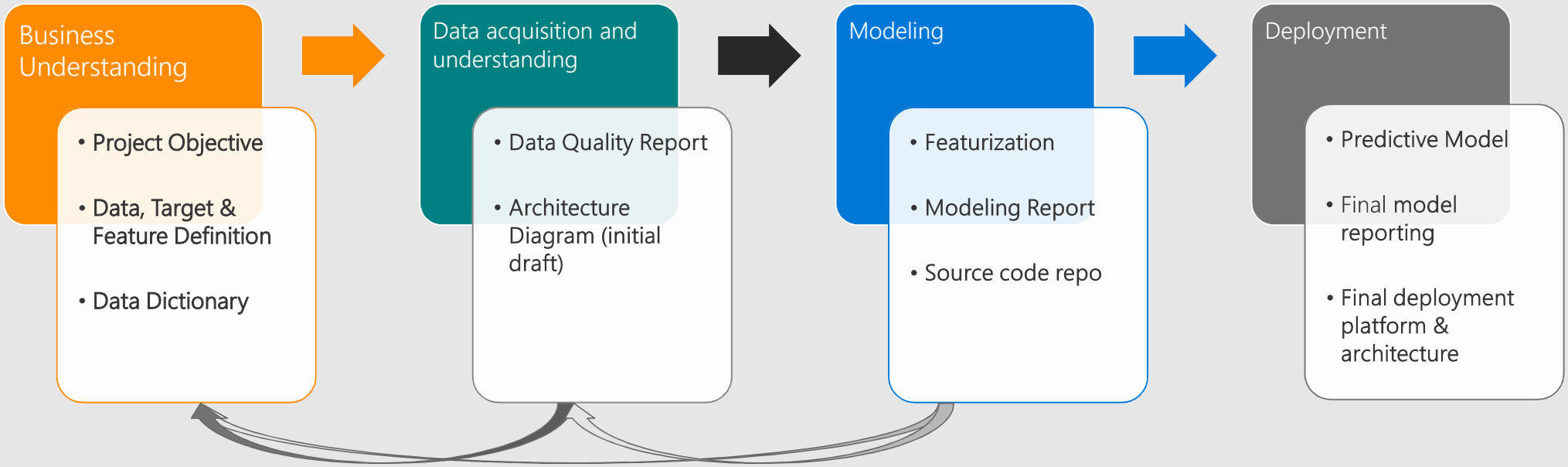
- Standardized Data Science Lifecycle
- Infrastructure & Toolkits
- Project Structure, Templates & Roles
- Project Execution (incl. DevOps components)
- Re-usable Data Science Utilities

Data Science lifecycle



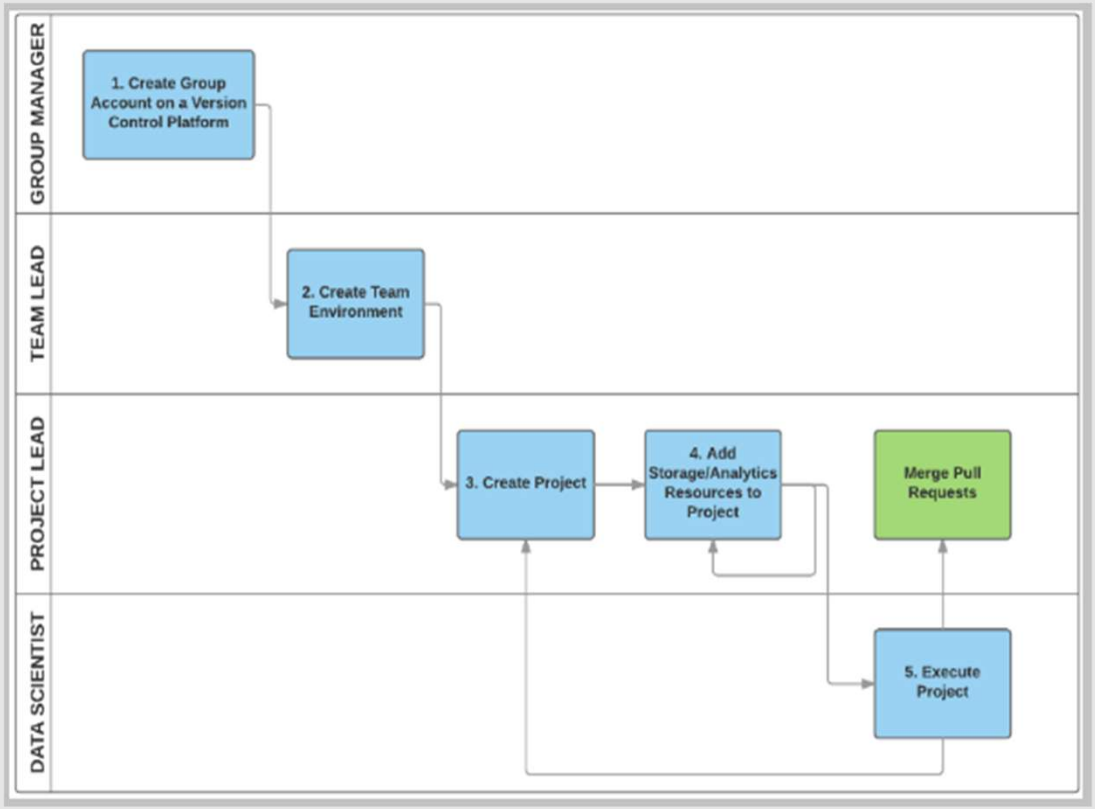
- Primary stages:
  - Business Understanding
  - Data Acquisition and Understanding
  - Modeling
  - Deployment

TDSP lifecycle stages can be integrated with specific deliverables & checkpoints



Project roles & tasks

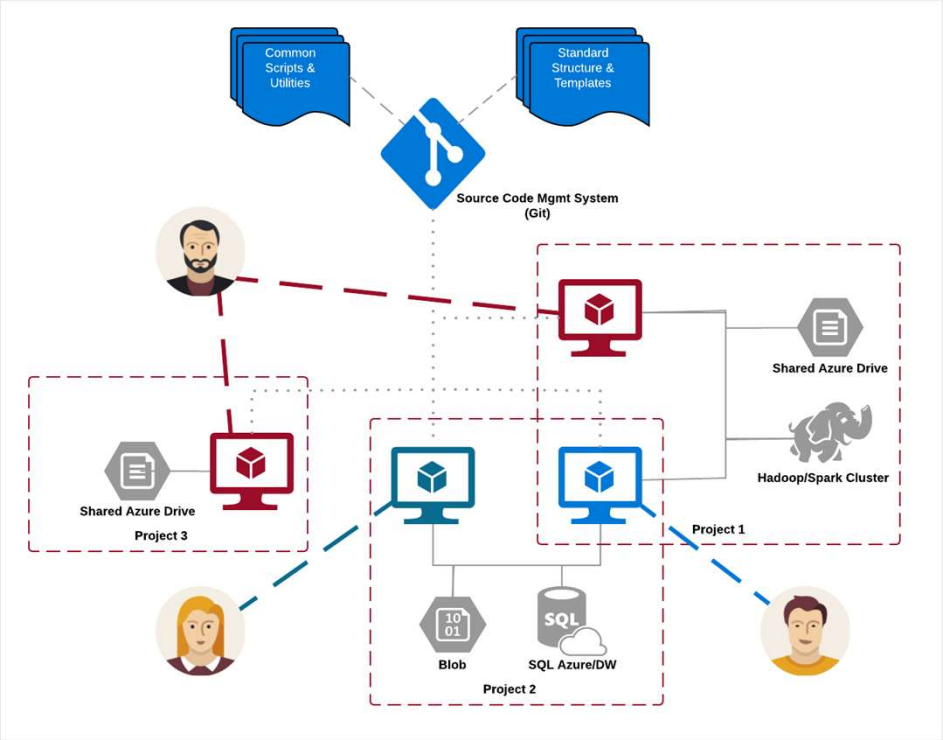
Personas: account managers, PMs, DSs, SWEs, architects



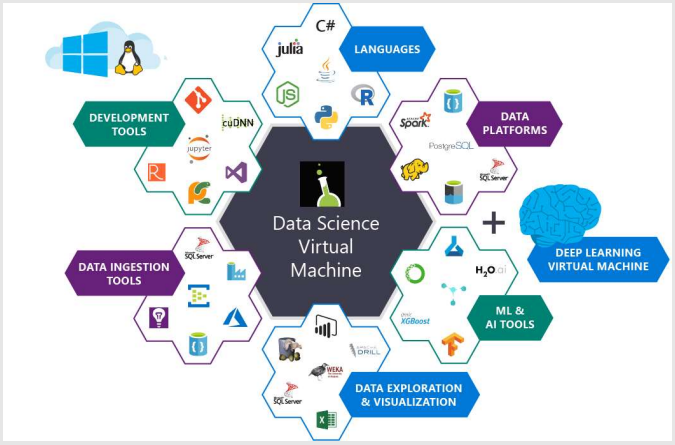
NOTE: Specific roles depend on organizations

- Governance and Project Management
  - Team lead
    - Git template repo & sever management, access control
  - Project lead
    - Business understanding, create project, work-items
- AI Developers
  - Data scientist
    - Modeling, exploratory analysis
  - SWE, data engineers, or architect
    - Data ingestion, deployment, solution architecture, integration into business applications

Shared and distributed infrastructure & toolkits



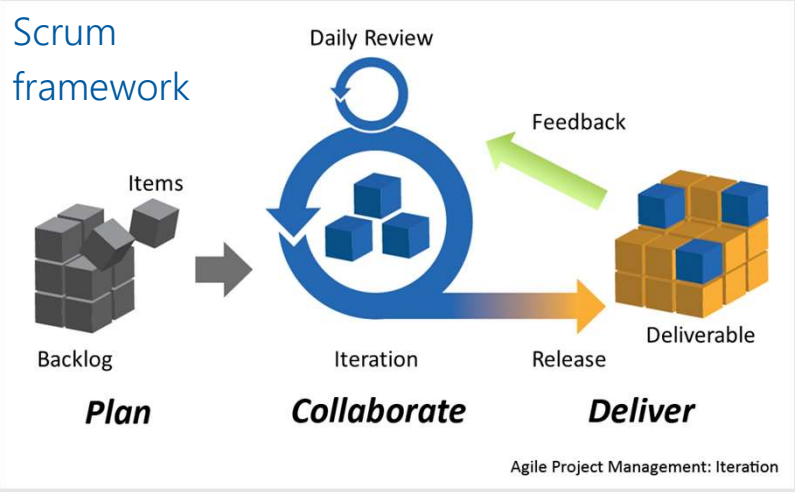
- Project artifacts & code stored in central git (version control) repositories.
- Data typically stored in cloud stores, such as blob or database
- Virtual machines (VMs), or clusters are disposable compute, added to projects as needed
- Many-to-many relationship between data scientists, compute and projects



Agile work planning and execution template

- Use Agile work planning & execution template (data science specific)
  - DS Projects: e.g. "Fraud Detection for Customer ABC"
  - DS Stages: correspond to the stages in TDSP lifecycle.
  - DS Stories: correspond to the life-cycle sub-stages.
  - DS Tasks: Tasks are assignable code or document work items to complete a specific data science story.

Scrum framework

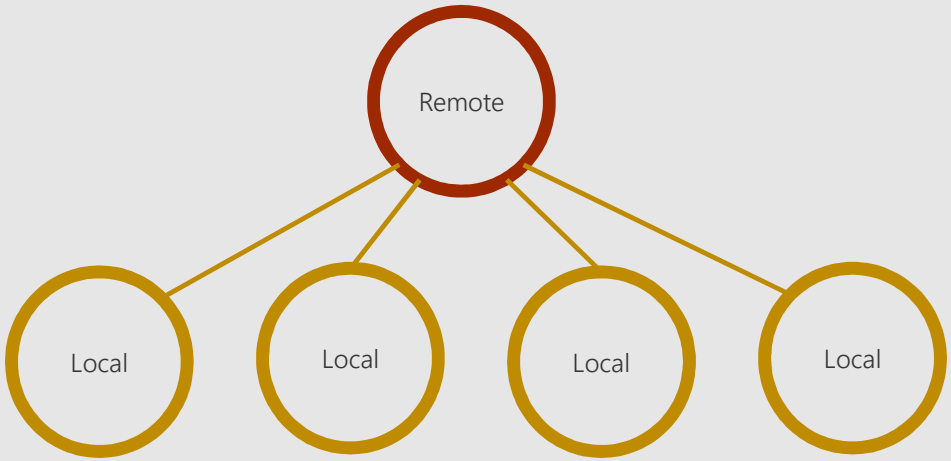


https://commons.wikimedia.org

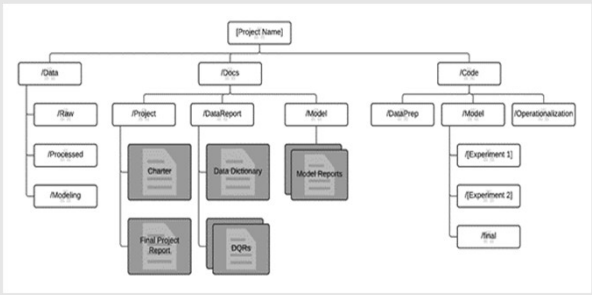
Work Item Type	Title	State
Data Science...	Fraud_Detection_CompanyABC	New
Business Und...	Define Objectives	Active
Data Science...	identify business variables	Active
Data Science...	Define success metrics	Active
Data Science...	define accuracy	New
Data Aquisition	Ingest Data	New
Data Aquisition	Explore First Derm data	New
Data Science...	Apply IDEAR to check data quality	New
Data Science...	Find missing data	New
Data Science...	Find numerical data distribution	New
Modeling	Feature Engineering	New
Data Science...	Compute various features	New
Data Science...	Compute categorical features	New
Data Science...	Compute text features	New
Deployment	Operationalize the model	New
Data Science...	Deploy model as a web services	New
Data Science...	set up web service with azure cloud services	New

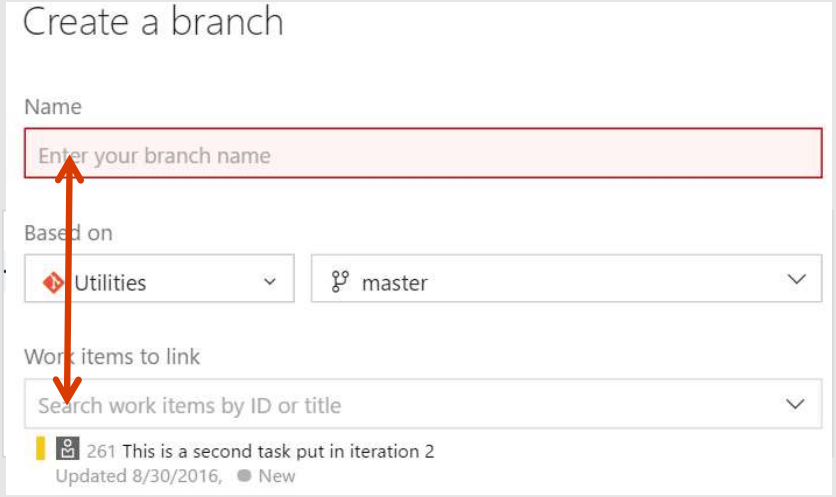
Collaborative development guidelines

- Version control and review
  - Git is a Version Control System
  - Each repo contains the full change history
  - Used in a distributed way with a single remote repo and several local repos (on local machine or a VM)

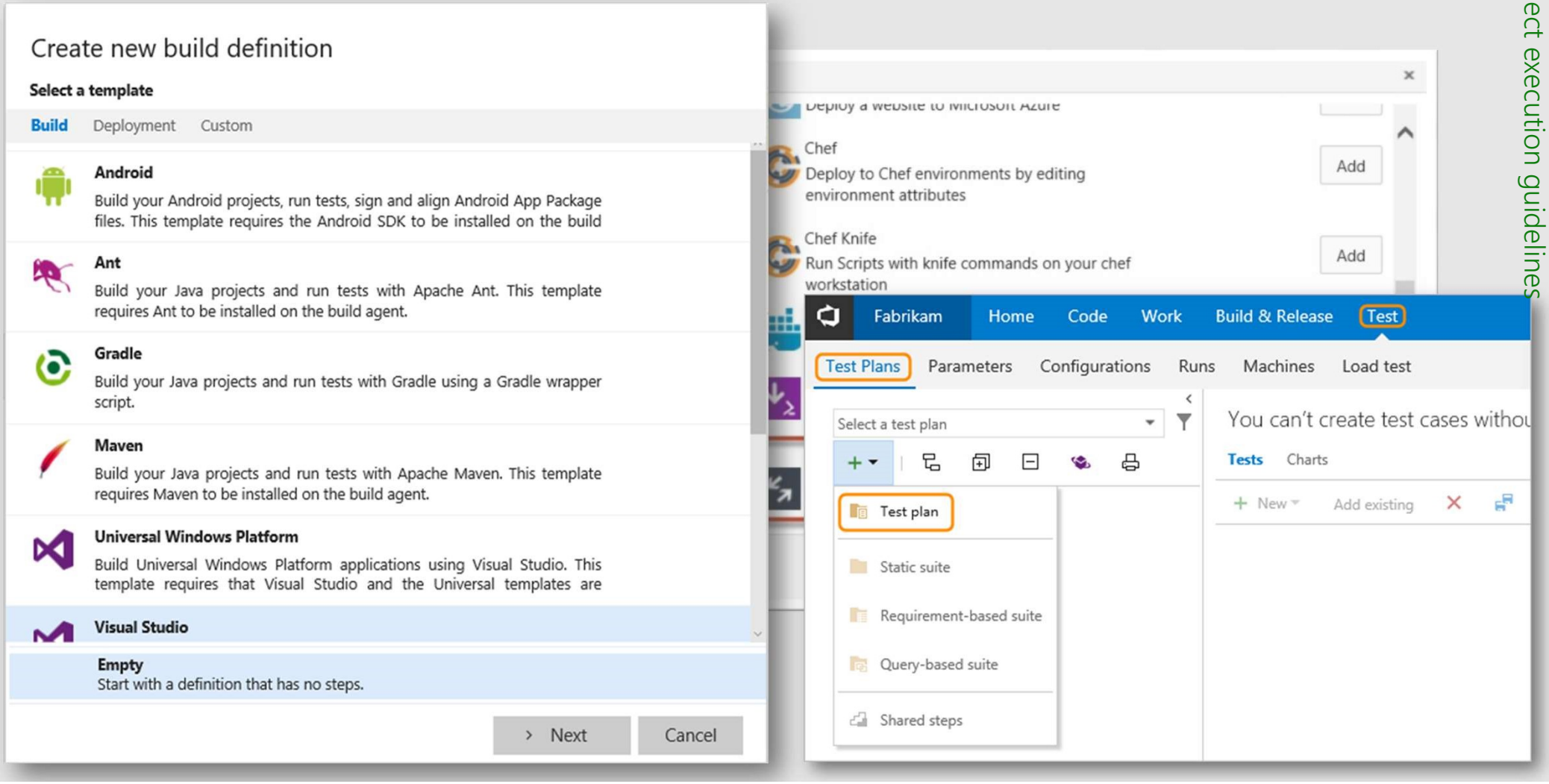


TDSP git Template





Integrating DevOps in projects

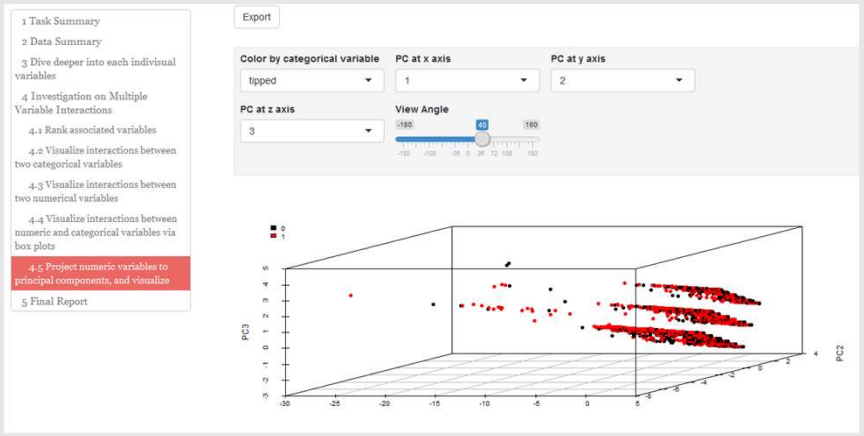


Re-usable data science utilities: Analytics

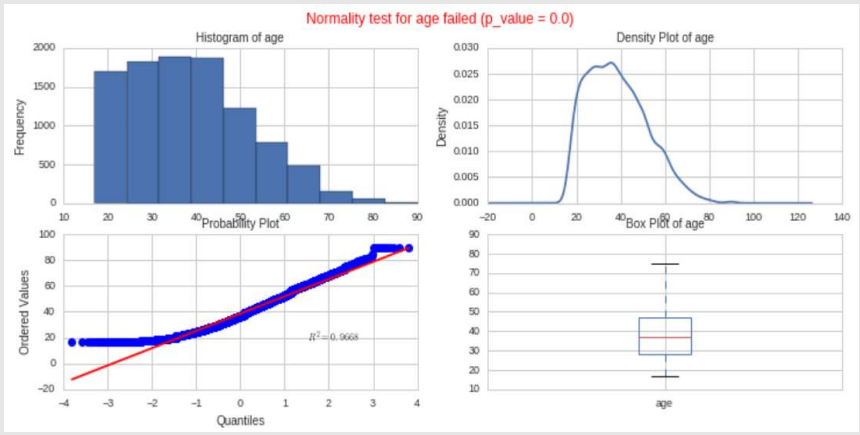
Interactive data exploration and reporting – IDEAR (Python, R, MRS)

- Data quality assessment
- Getting business insights from the data
- Association between variables
- Generating standardized data quality reports automatically

Clustering



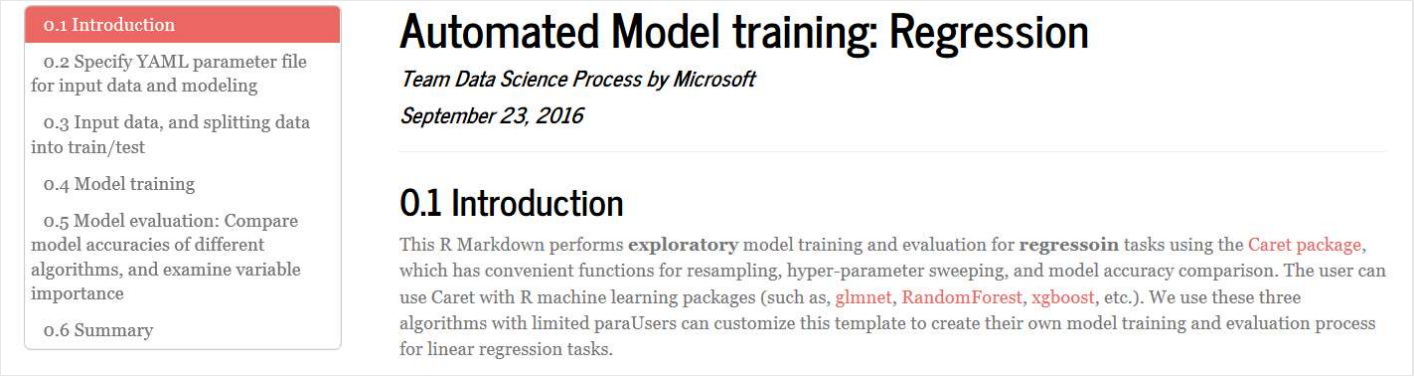
Distribution assessment



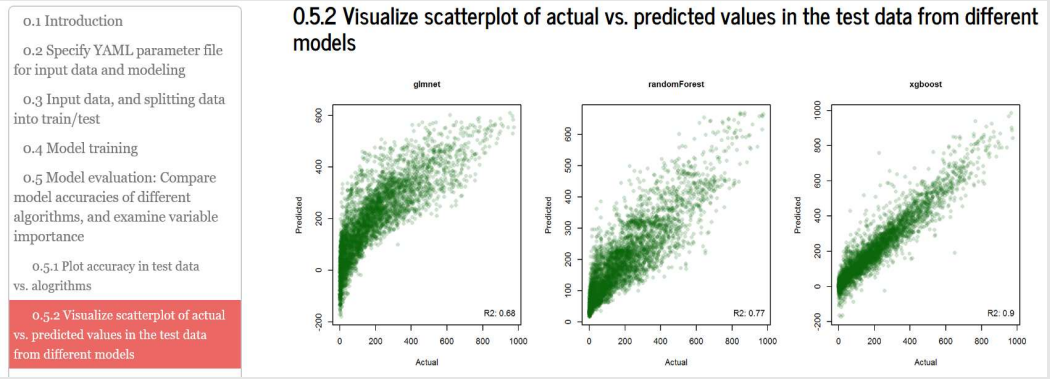
https://github.com/Azure/Azure-TDSP-Utilities

Re-usable data science utilities: Analytics - modeling

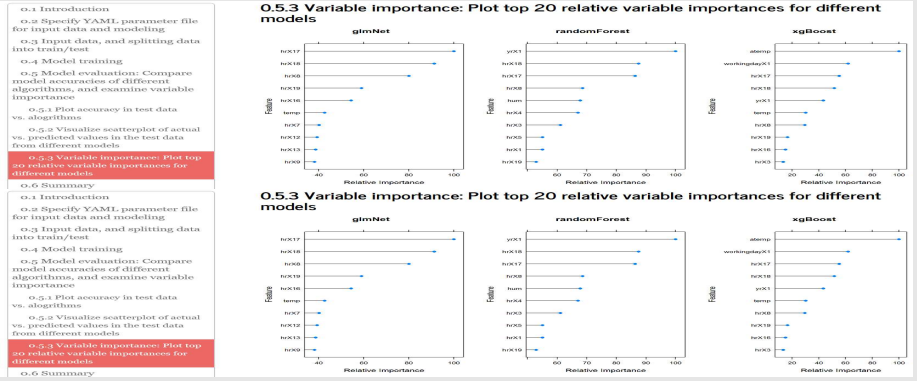
Automated modeling and reporting AMAR (R)



Predicted vs. Actual (multiple algorithms)

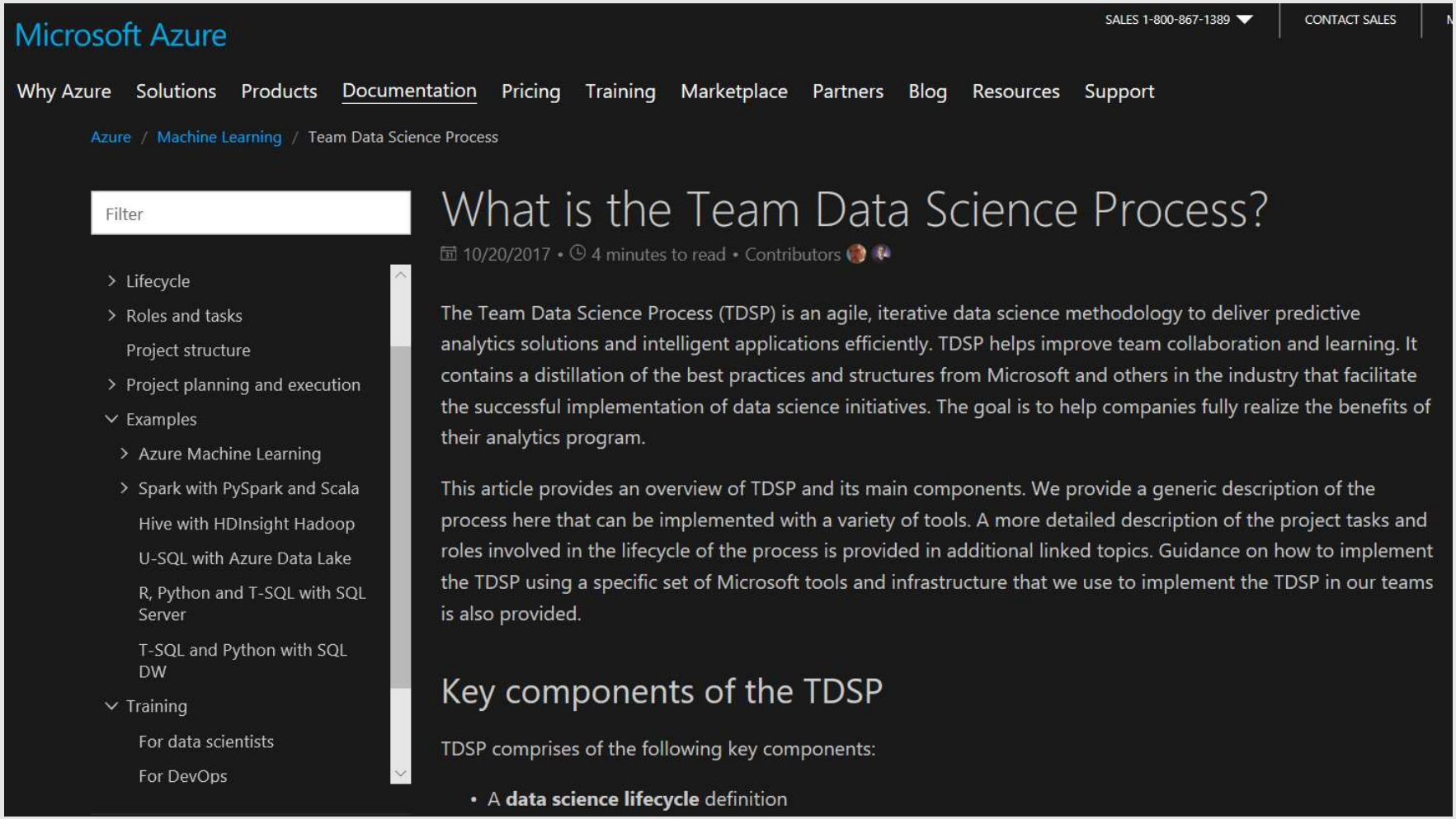


Feature Importance (multiple algorithms)

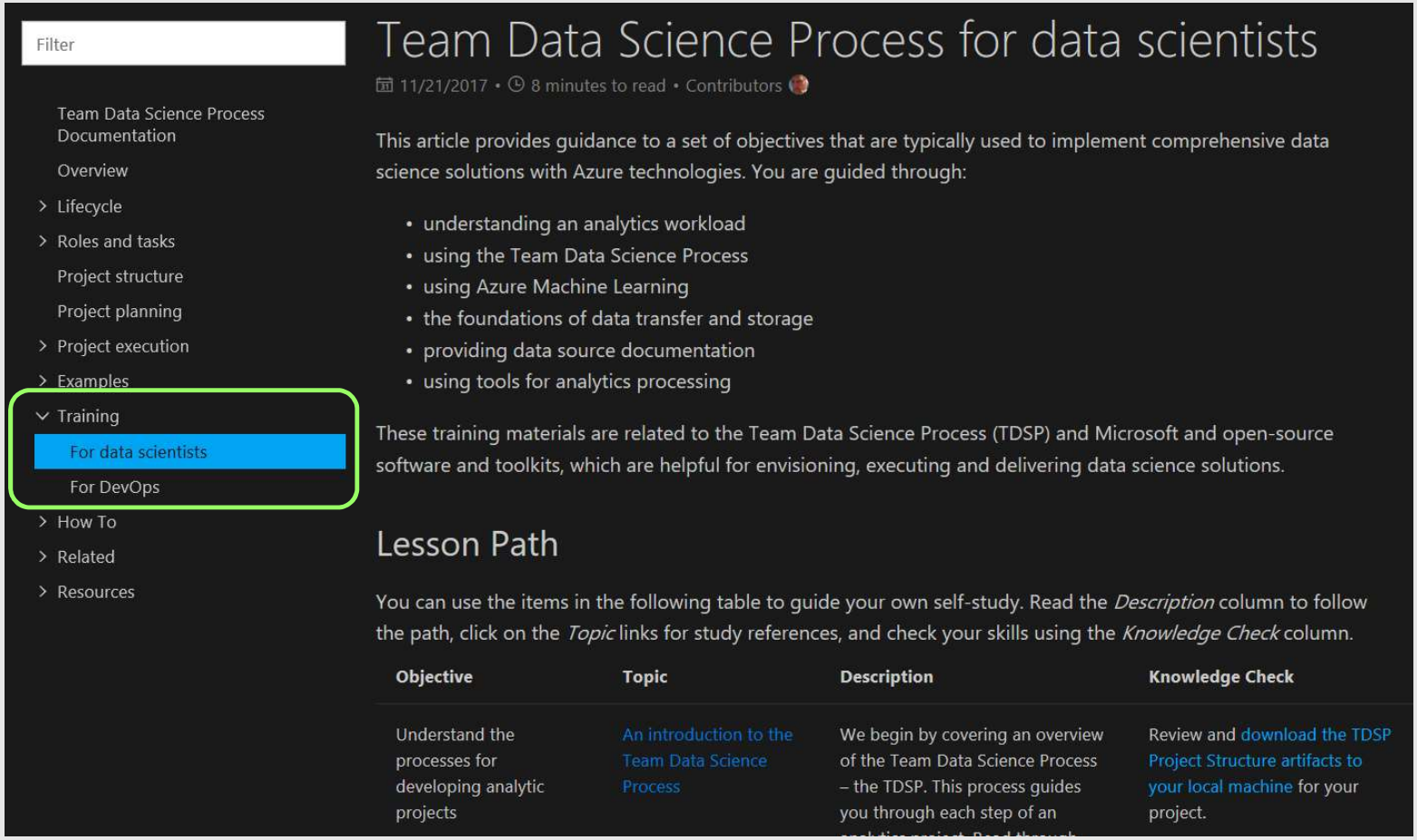


https://github.com/Azure/Azure-TDSP-Utilities

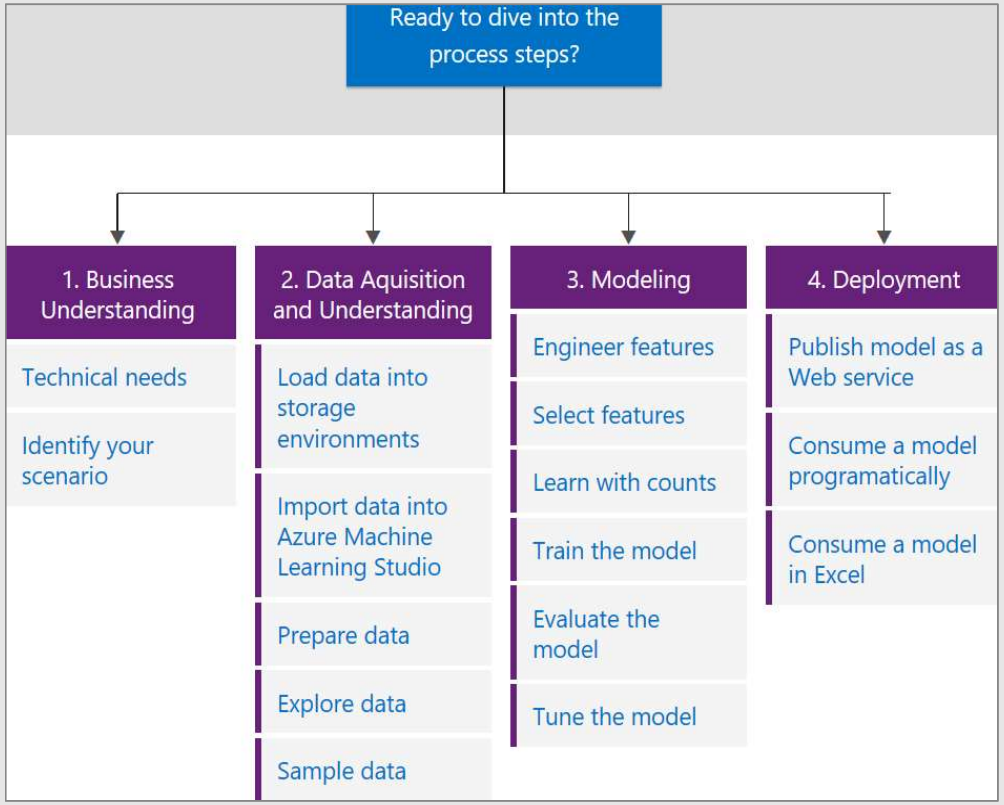
TDSP documentation: https://aka.ms/tdsp



TDSP trainings



TDSP in action: E2E worked-out samples



- Azure Machine Learning
- Azure HDInsight Spark
- Azure HDInsight Hadoop
- SQL-server with R and Python
- Azure SQL data warehouse
- Azure Data Lake

Adoption: How to stage (if needed)

- Data science teams may stage adoption as follows

Level 1

- One git repository per project
- Standard directory structure
- Standardized templates like charter, exit reports
- Planning and tracking of work items

Level 2

- Customize templates to fit team needs
- Create shared team utility repo (like IDEAR, AMAR)

Leve 3

- Develop process to graduate code from projects to the shared team utility repo
- Develop E2E worked-out templates
- Use mature work planning and tracking system (e.g. Agile)

Level 4

- Link git branch with work items
- Code review
- Manage and version model and data assets
- Develop automated testing framework
- Develop automated CI/CD

Adoption: Customers

- Microsoft internal
  - Microsoft consulting services (MCS)
  - AI & R Cloud Platform: Algorithm and data sciences team
  - Windows Devices DS team
  - …
  - …

Microsoft Consulting Services

- External partners
  - *New Signature*
  - *BlueGranite*

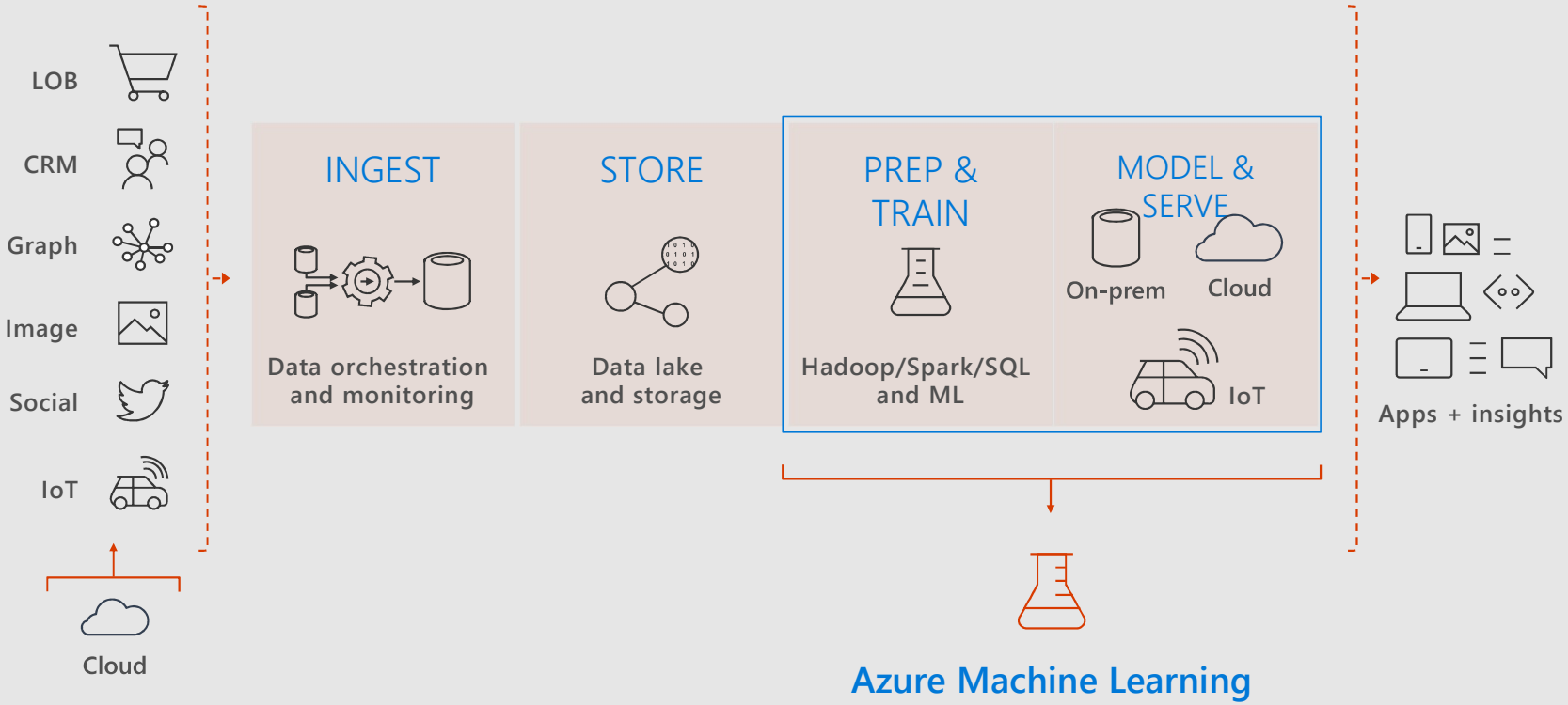
NEWSIGNATURE

BLUEGRANITE
BUSINESS INSIGHT. DELIVERED.

Azure Machine Learning (AML)

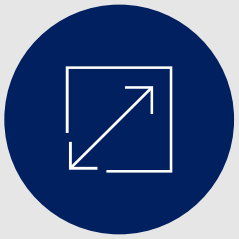
https://docs.microsoft.com/azure/machine-learning/preview

Azure Machine Learning



Bring AI everywhere

Benefit from the fastest AI developer cloud



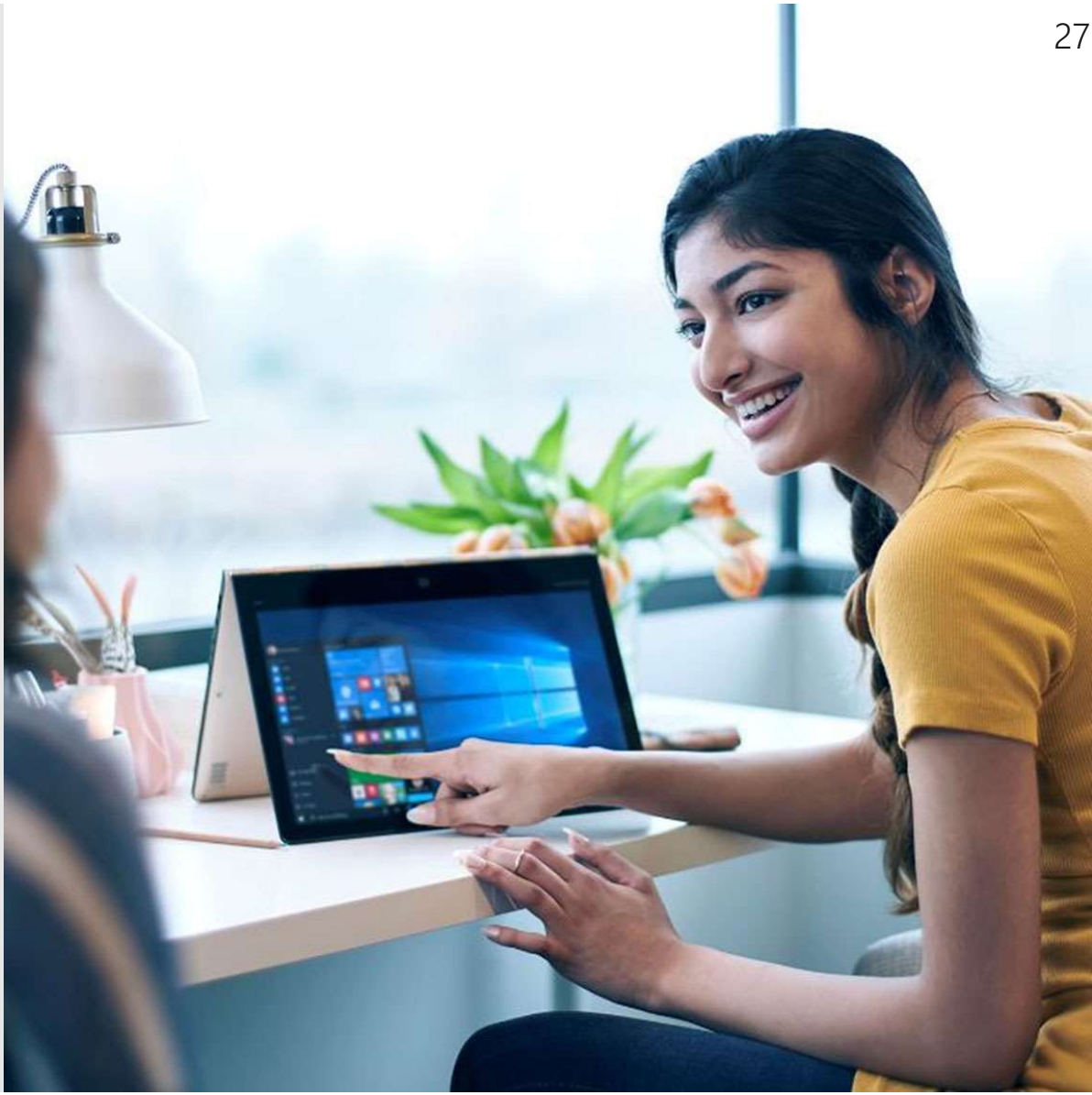
Build, deploy, and manage at scale

Boost productivity with agile development

Build with the tools and platforms you know

Build, deploy, and manage at scale

- Build and deploy everywhere – cloud, on-premises, edge, and in-data
- Deploy in minutes with data driven management and retraining of all your models
- Prototype locally then scale up and out with VMs, Spark clusters, and GPUs
- Real time, high through-put insights everywhere, including Excel integration



Manage models

Deployment and management of models as HTTP services

Container-based hosting of real time and batch processing

Management and monitoring through Azure (e.g., AppInsights)

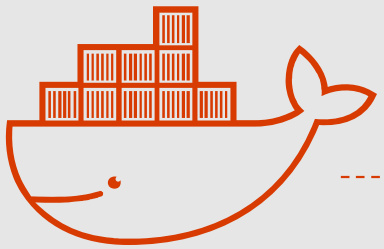
First class support for SparkML, Python, CNTK, TF, TLC, R, extensible to support others (Caffe, MXnet)

Service authoring in Python and .NET Core

```
# -n app name
# -f scoring script file name
# -m dependencies, in this case it is the pickled model file
# -r type of model, in this case it is the scikit-learn model
$ az ml service create realtime -n irisapp -f iris_score.py -m model.pkl -r scikit-py
```

```
$ az ml service run realtime -n irisapp -d '{"input": "[[1.0, 2.2, 2.3, 2.7]]"}'
2
```

Deploy everywhere



DOCKER

- Single node deployment (cloud/on-prem)
- Azure Container Service
- Azure IoT Edge
- Spark clusters

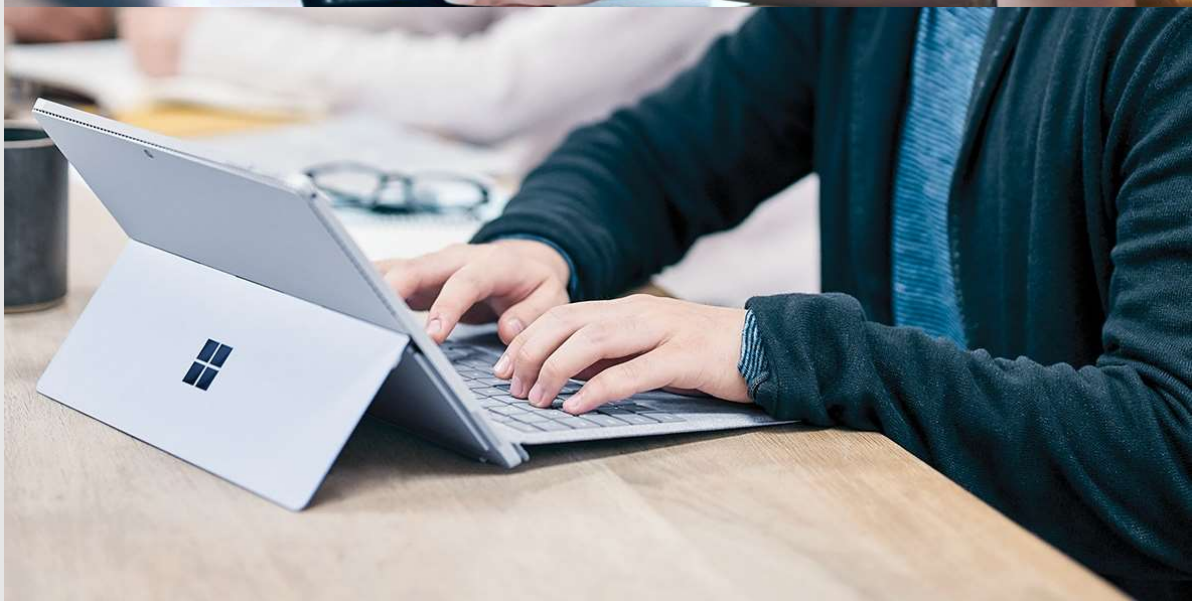
Boost productivity with agile development

- Spend more time modeling and less time prepping with built-in intelligent data wrangling
- Increase collaboration and sharing with notebooks and Git
- Avoid losing anything with version control and reproducibility
- Know the best performing models with metrics, lineage, run history, asset management, and more

Build with tools and platforms you know

- Choose between visual drag-and-drop or code-first authoring
- Use your favorite IDEs
- Call Azure Machine Learning services directly in VS Code*
- Build on any framework or library with the most popular languages
- Train quicker and easier with industry-leading Spark and GPUs

*More IDEs coming soon



Use what you want

Use your favorite IDE

Leverage all types of platforms and tools/libraries

USE ANY FRAMEWORK OR LIBRARY

TensorFlow Microsoft CNTK Apache Spark ML

USE ANY TOOL

VS CODE Visual Studio jupyter PyCharm

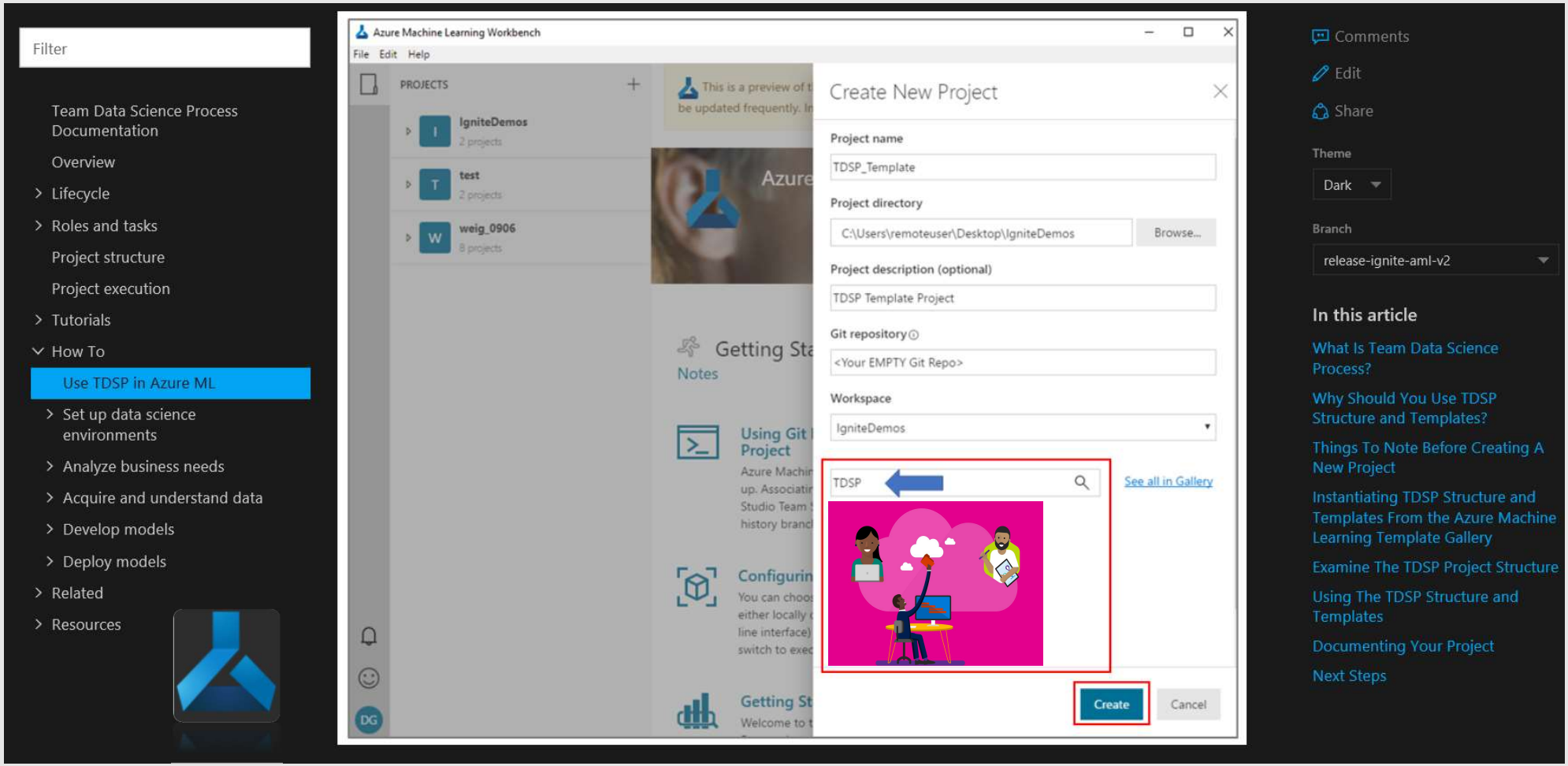
USE THE MOST POPULAR INNOVATIONS

git Apache Spark

docker python R

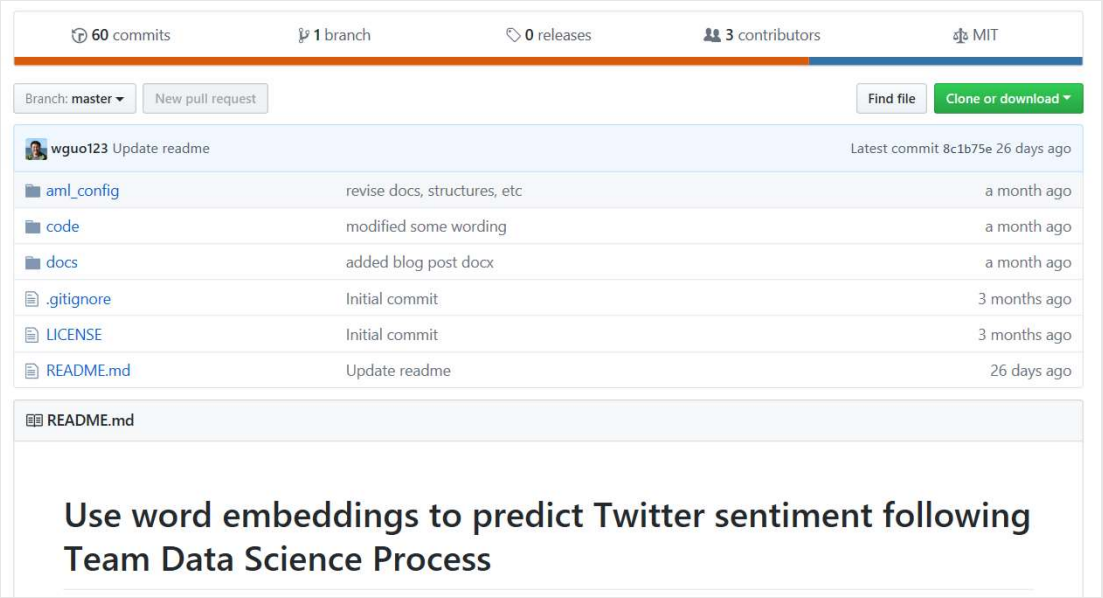
Using TDSP with AML

Using TDSP with Azure Machine Learning



TDSP – AML worked-out samples in AI

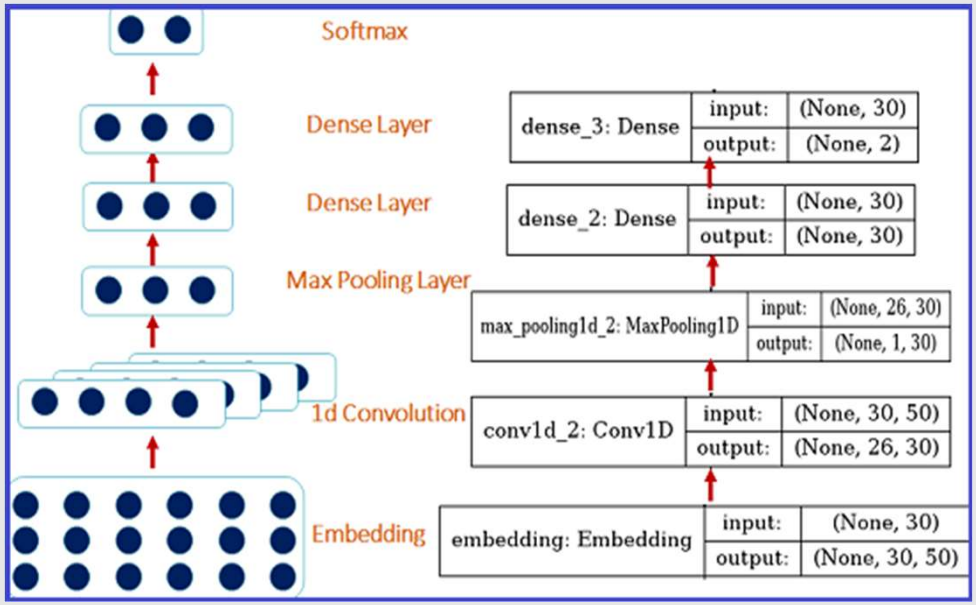
- Unstructured data (NLP) modeling with deep-learning
  - Sentiment classification using supervised word-embeddings
  - Biomedical (PubMed) Named entity recognition using LSTM
- Structured data
  - Income classification
- Public Github repositories for each sample

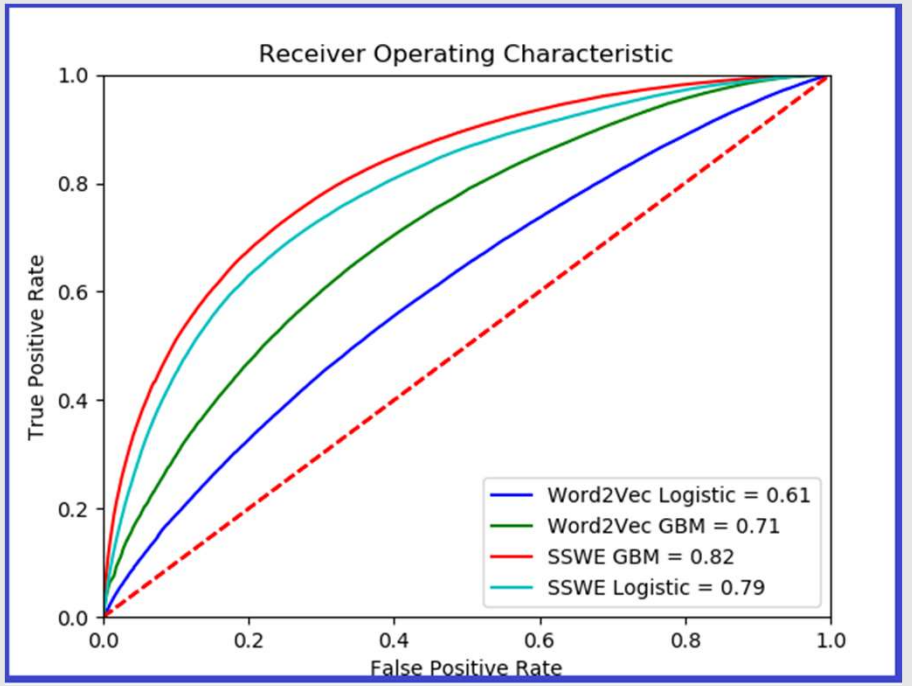


https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/walkthroughs

Sample: Sentiment-specific word embeddings (SSWE) improves classification of sentiments

CNN: Keras and CNTK





https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/predict-twitter-sentiment

http://www.aclweb.org/anthology/P14-1146

Summary

- A well-defined data science process is critical to bridge the gap between opportunities and challenges, and put AI models in production to impact businesses
- A combination of appropriate process (e.g. TDSP), tools (e.g. AML) and data-platforms (e.g. DSVM) is important for efficient development and deployment of AI solutions
  - TDSP – process
  - AML – tool
- A process such as TDSP can be used with a variety of tools and data platforms
- Resources are available to use TDSP and AML together, along with other data platforms on Azure to efficiently build and deploy AI solutions on the cloud

Thank you!

https://aka.ms/tdsp
tdsp-feedback@microsoft.com

Deck available @: https://aka.ms/tdsp-presentations