Microsoft Azure

Intro to Azure Machine Learning and Security Considerations for the ML lifecycle

Who we are

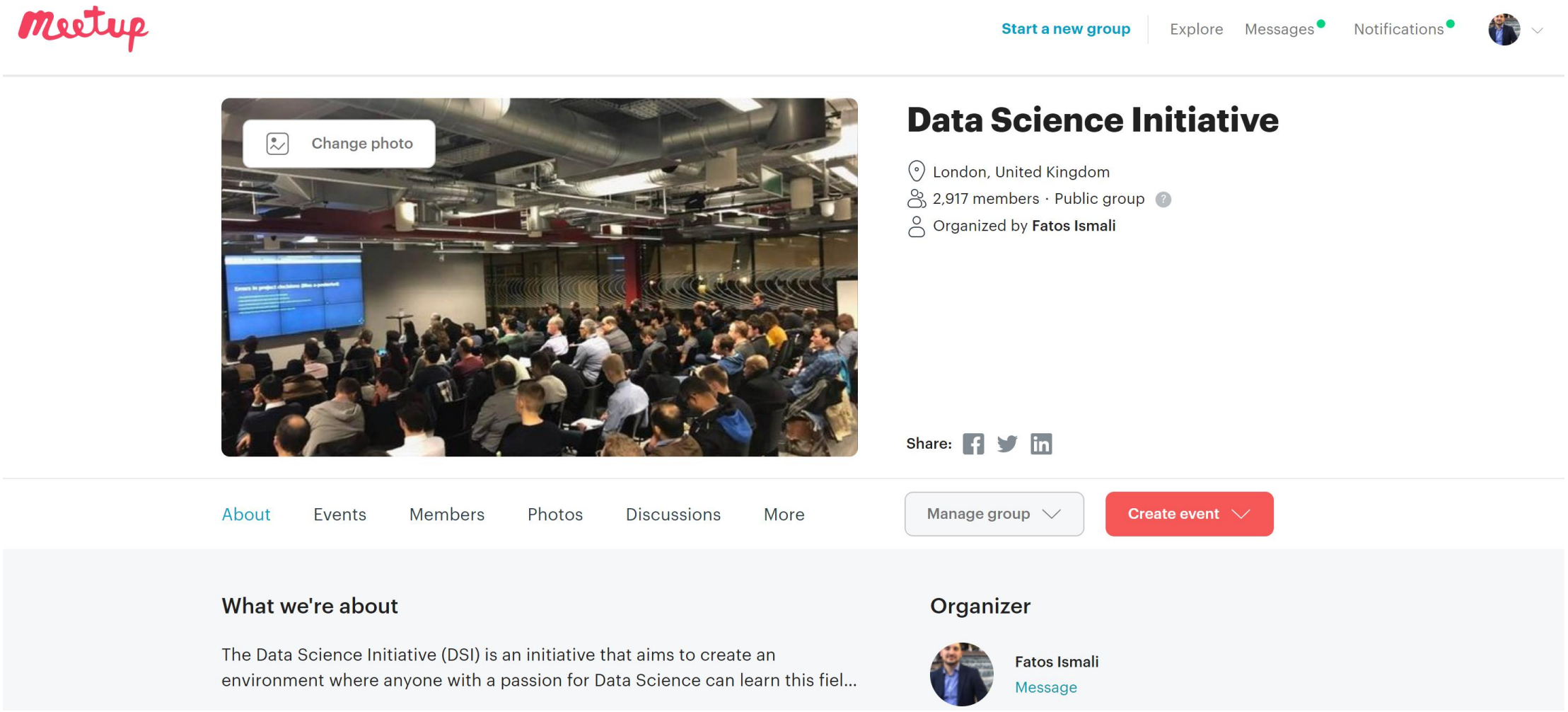
Fatos Ismali

Joe Plumb

Data & Analytics team @ Microsoft

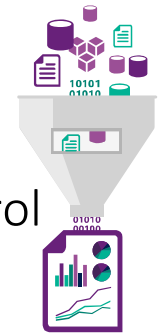
Data Science Initiative

https://www.meetup.com/data-science-initiative/



Data Science Described

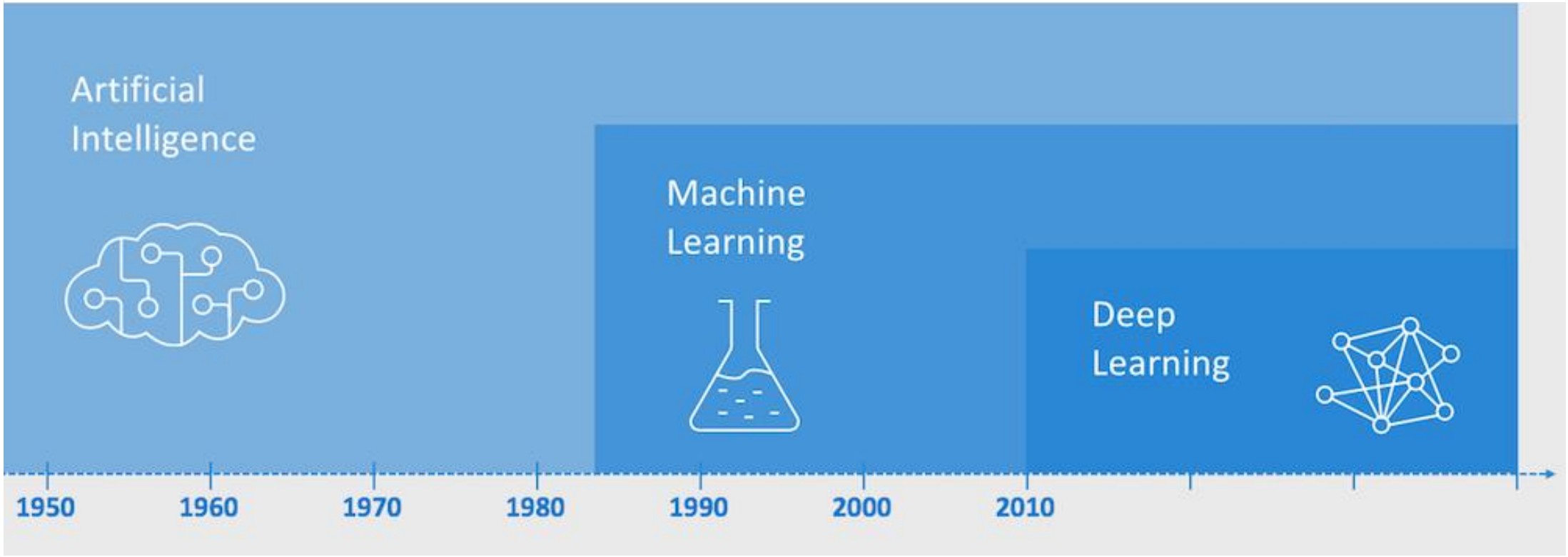
- **Data Analysis**
  - Create, Read, Update Delete
  - Programming and Control
  - Reporting
  - Business Intelligence
  - Statistics and Data Mining
  - Story Telling



- **Scientific Process**
  - Define the Question
  - (Create Hypothesis)
  - Create a Repeatable Test
  - Publish Results

- **Artificial Intelligence**
  - **Machine Learning**
  - **Deep Learning**

Progression

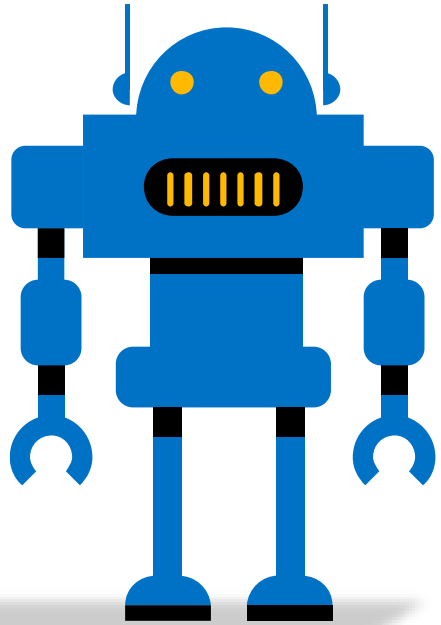


Machine Learning

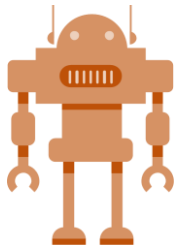
The Formal Definition:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

Tom M. Mitchell, Professor at Carnegie Mellon University, USA



Machine Learning Simplified



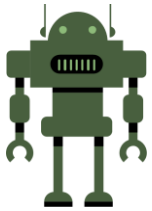
Features					Label
Image	OpenTop	ClosedBottom	HoldsLiquid	Hand-Sized	IsCup
	TRUE	TRUE	TRUE	TRUE	TRUE
	TRUE	TRUE	TRUE	TRUE	TRUE
	TRUE	TRUE	TRUE	TRUE	TRUE
	FALSE	FALSE	FALSE	FALSE	FALSE

Model - *If*:

- Open top = *TRUE*
- Closed bottom = *TRUE*
- Holds liquid = *TRUE*
- Hand-sized = *TRUE*
- …

Then

- CUP = TRUE

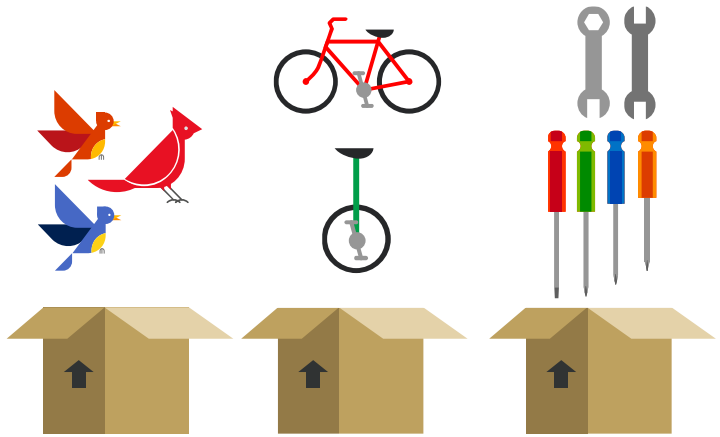


Score :

- CUP = TRUE
- CERTAINTY = 90%

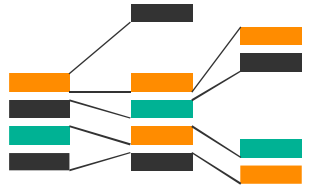
Machine Learning Uses and Algorithm Families

Which category
(Classification)



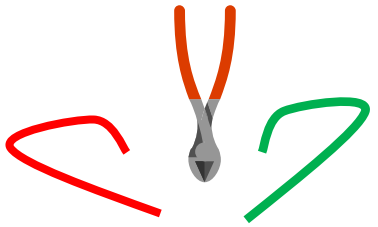
How much/many
(Regression)

Which group
(Clustering, Recommender)



Is it odd
(Anomaly)

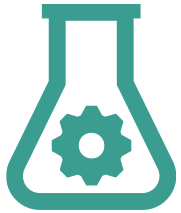
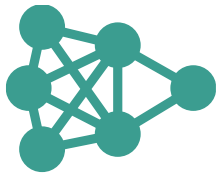
Which action
(Reinforcement Learning)



Languages and Libraries

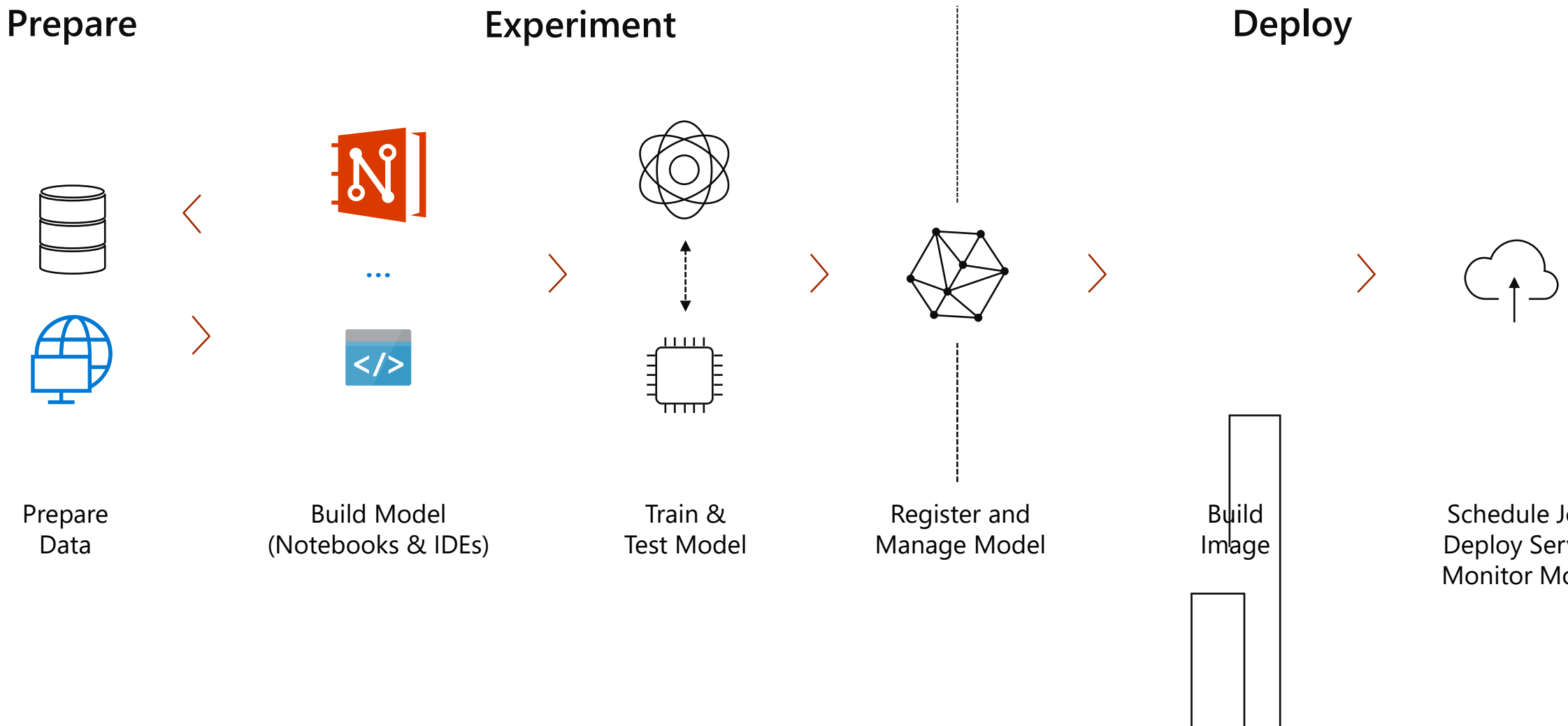
Language	Libraries
R	Standard Libraries and Packages, E1071, rpart, randomForest, caret, kernlab, glmnet, ROCR, gbm, party, arules, tree, klaR, Rweka, ipred, lars, earth, CORElearn, mboost
Python	TensorFlow, Scikit-Learn, Numpy, Keras, PyTorch, LightGBM, Eli5, SciPy, Theano, Pandas
Platform-Specific (Spark, Hadoop, etc.)	MLLib, PySpark, SparkR, … (Spark), Etc.
SQL (data selection and preparation)	N/A
Java (primarily data preparation)	ADAMS, Deeplearning4j, ELKI, JavaML, JSAT, Mahout, MALLET, Massive Online Analysis, RapidMiner, Weka.

Project Roles in Machine Learning

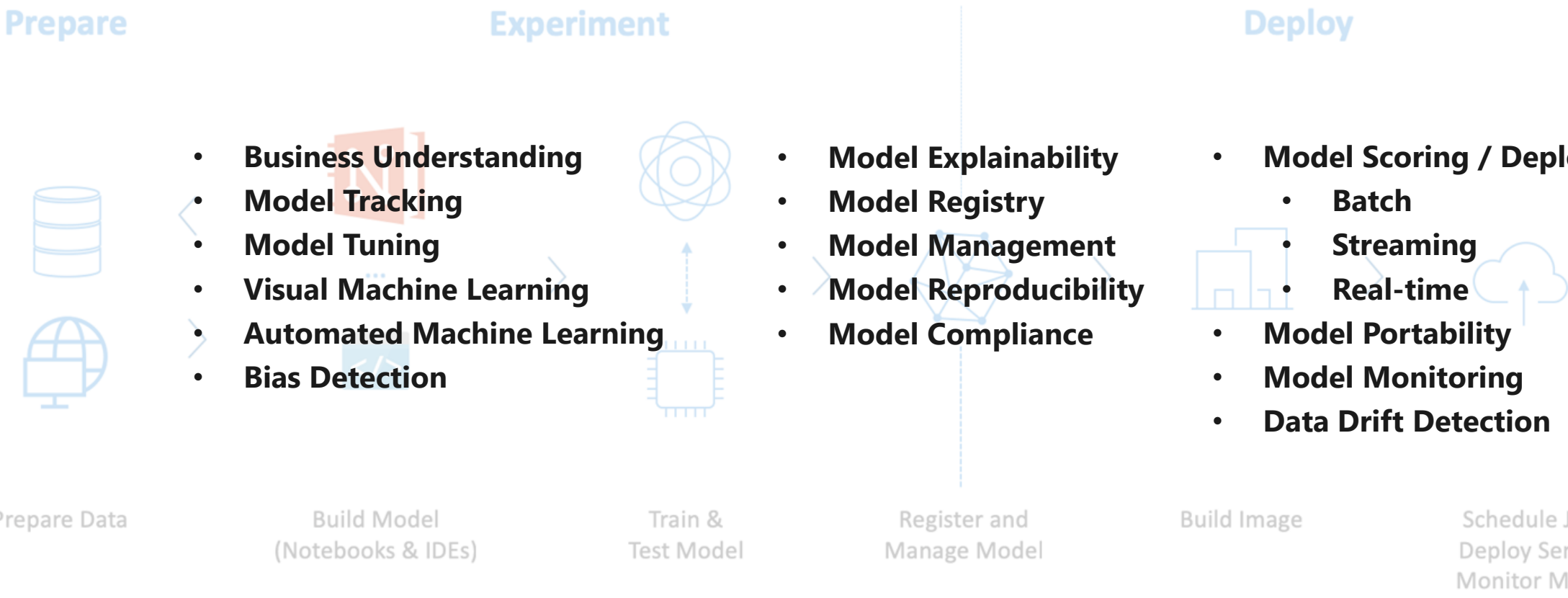




Data Scientist	ML Engineer	Business Analyst	IT	AI Architect
• Is this going to let me do new or better kinds of models and analysis?	• Will this let me put this into production and operate at scale, but still give me visibility?	• Is this something I can use myself without becoming a DS PHD?	• Is this going to be supported, secure, compliant, and integrated with our IT infrastructure	• Does my organization have the resources and infrastructure to successfully use AI/ML?

ML Ops - Lifecycle



ML Ops – Lifecycle Complexity



- **Business Understanding**
- **Model Tracking**
- **Model Tuning**
- **Visual Machine Learning**
- **Automated Machine Learning**
- **Bias Detection**

- **Model Explainability**
- **Model Registry**
- **Model Management**
- **Model Reproducibility**
- **Model Compliance**

- **Model Scoring / Deployment**
  - **Batch**
  - **Streaming**
  - **Real-time**
- **Model Portability**
- **Model Monitoring**
- **Data Drift Detection**

Model complexity

Type I

Simple single-threaded models leveraging traditional supervised and unsupervised machine learning techniques on easily accessible datasets

Type II

Single-threaded models leveraging more advanced techniques in addition to traditional ML. For instance, deep learning and reinforcement learning, natural language processing, computer vision, sound processing.

Type II-A (Parallel training)

Traditional ML models and more complex models (e.g. Deep Learning) leveraging parallel training of multiple models

Type II-B (Distributed training – Spark based)

Leveraging distributed processing engines such as Spark for distributing training of ML models and data processing

Machine Learning on Azure

Domain specific pretrained models
To simplify solution development


Vision


Speech


Language


Search

Familiar Data Science tools
To simplify model development


Visual Studio Code


Azure Notebooks

Jupyter


Command line

Popular frameworks
To build advanced deep learning solutions


PyTorch


TensorFlow

Scikit-Learn


ONNX

Productive services
To empower data science and development teams


Azure Databricks


Azure Machine Learning


Machine Learning VMs

Powerful infrastructure
To accelerate deep learning


CPU


GPU


FPGA


From the Intelligent Cloud to the Intelligent Edge

Machine Learning on Azure

Domain specific pretrained models
To simplify solution development


Vision


Speech


Language


Search

Familiar Data Science tools
To simplify model development


Visual Studio Code


Azure Notebooks

Jupyter


Command line

Popular frameworks
To build advanced deep learning solutions


PyTorch


TensorFlow

Scikit-Learn


ONNX

Productive services
To empower data science and development teams


Azure Databricks


Azure Machine Learning


Machine Learning VMs

Powerful infrastructure
To accelerate deep learning

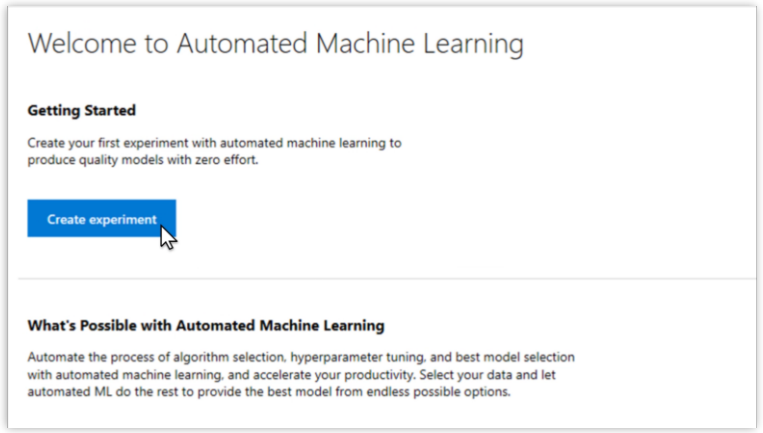

CPU


GPU

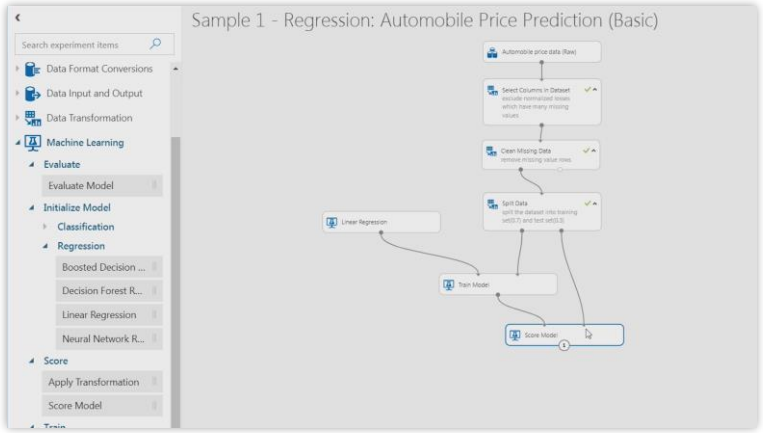

FPGA


From the Intelligent Cloud to the Intelligent Edge

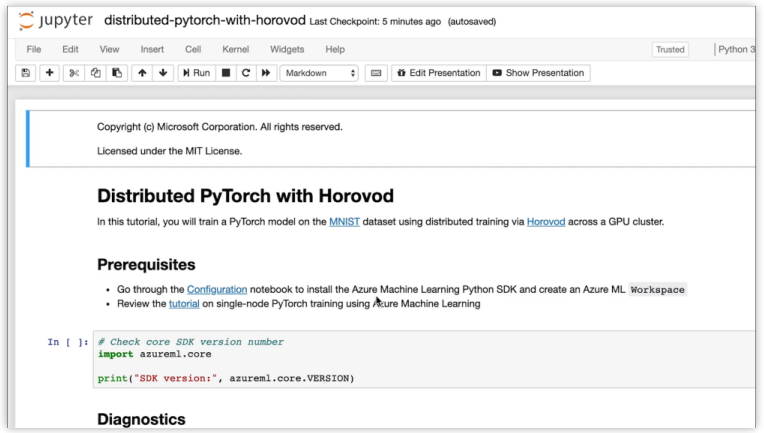
Machine learning for any skill level



Automated machine learning UI

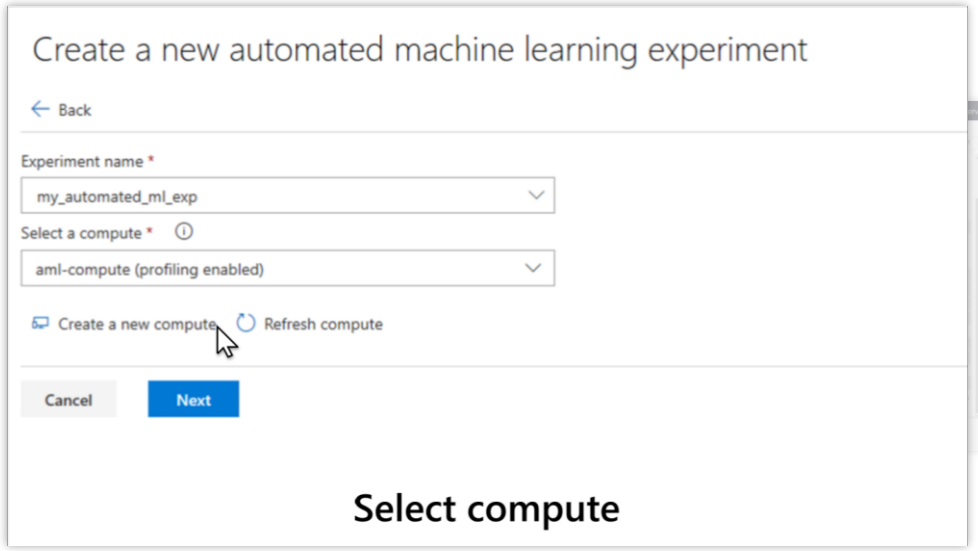


Visual interface



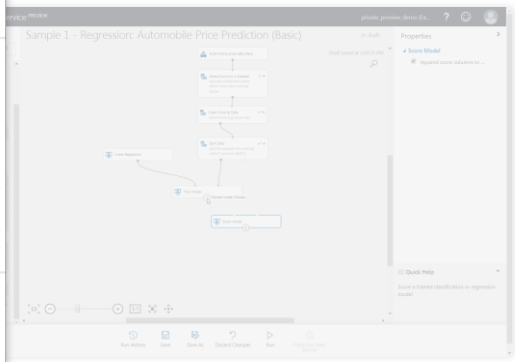
Machine learning notebooks

Machine learning for any skill level

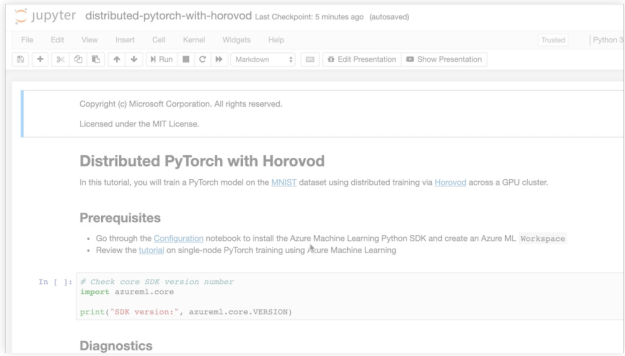


Select compute

Automated
machine learning UI



Visual interface

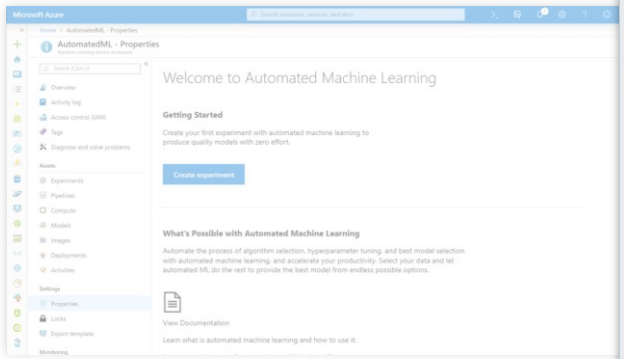


Machine learning notebooks

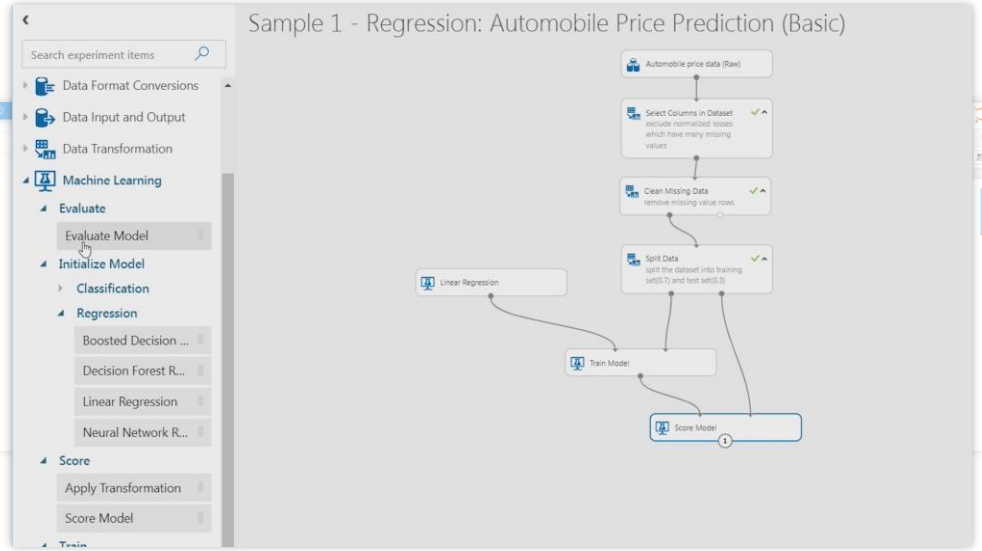
https://arxiv.org/abs/1705.05355

Machine learning for any skill level

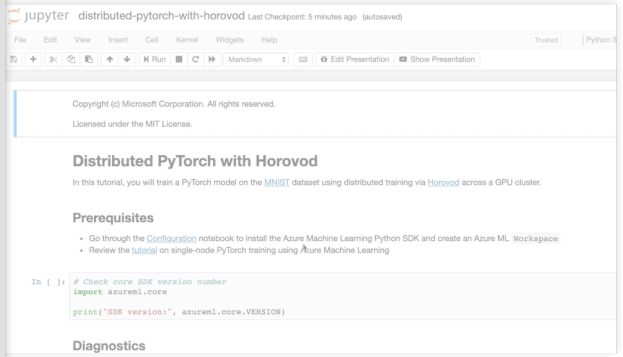
New capabilities in Azure Machine Learning service



Automated machine learning UI



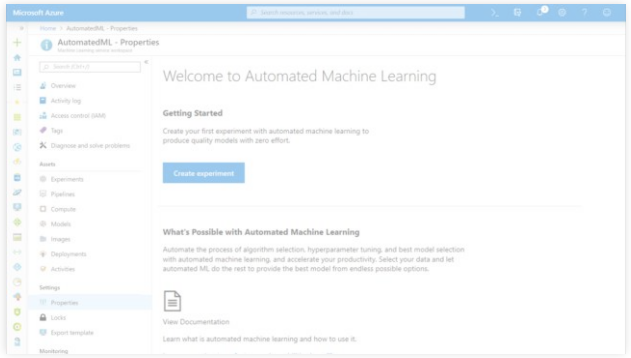
Visual interface



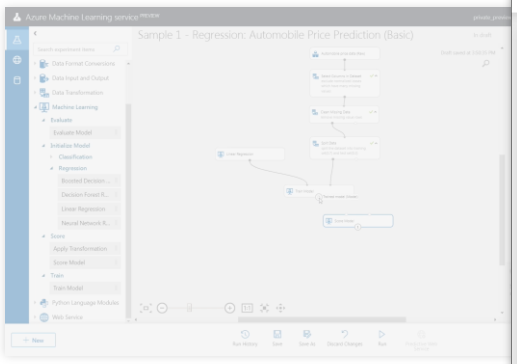
Machine learning notebooks

Machine learning for any skill level

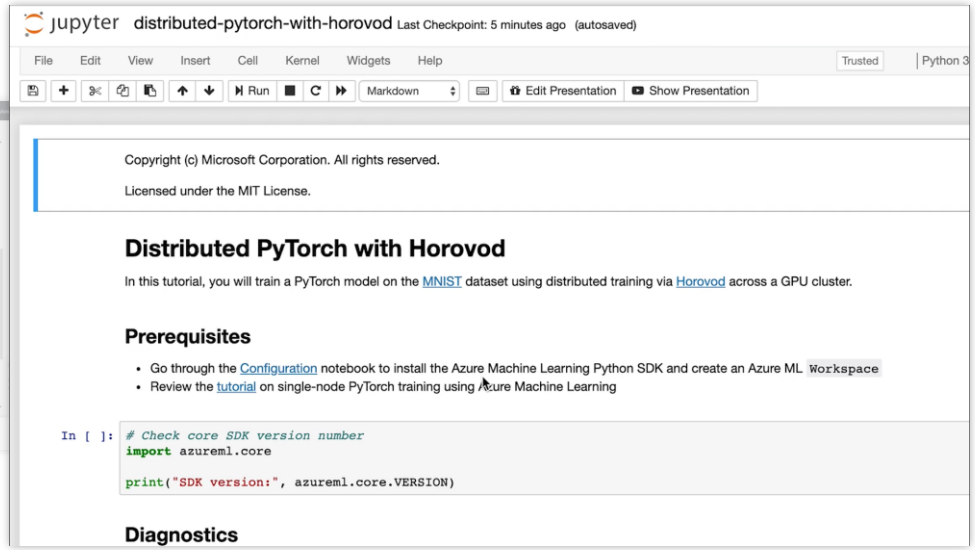
New capabilities in Azure Machine Learning service



Automated machine learning UI



Visual interface



Machine learning notebooks

Microsoft Azure

Security Considerations for the ML lifecycle

Security considerations for *any system*

USER ACCESS

SECRETS MANAGEMENT

DATA ENCRYPTION

NETWORK PERIMETER

Security Components – User Access

Authentication and Authorization

- **Users** authenticate with their user identity to access and provision services in Azure. Same identity for portal, PaaS, cli. Fine-grained RBAC across environments.
- **Service Principal** is a security identity granted access rights to applications and services to enable secure, automated execution of tasks.
- Create a Service Principal, then grant access permissions to services via management groups (further reading).
- Apply principle of least privilege i.e. only those privileges which are essential to perform its intended function.
  - **Azure Machine Learning**: Custom role allowing Microsoft.MachineLearningServices/workspaces/*/write (believe new AAD roles are coming soon)
  - **Azure Blob/Azure Data Lake Storage**: Storage Blob Data Reader (watch out for ACLs!)
  - **Azure Databricks**: Contributor permissions on the Databricks workspace. Generate authentication token manually (PAT token) or via CLI and store in Key Vault
  - **Azure DevOps**: New SP created when ARM Service Connection created.

Security components – Secrets Management Key Vault

- Secure store for secrets in your pipelines and workflows. Easy to provision, store, and get secrets.
- Create and store your Service Principal secret in Key Vault
- AML creates its own key vault on deployment to manage its secrets – you'll need to create your own for secrets for your project
- Use a separate Key Vault per application per environment. Reference secrets using the same variable name to ensure code portability between environments
- Configure key rotation with Azure Automation for keys that you manage. Storage accounts used by AML can be updated using the SDK.
- Setup and store variables and service connections in Azure DevOps/GitHub/Jenkins

Security Components – Data encryption
Safeguard data according to your needs

- **Encryption at rest** - AML stores snapshots, output, and logs in the Storage account that's tied to the workspace. All the data stored in Azure Blob storage is encrypted at rest with Microsoft-managed keys. You can bring your own keys too.
- If your workspace contains sensitive data, recommended to set the hbi_workspace flag while creating your workspace.
- **Encryption in transit –** Ensure external scoring endpoints are secured using TLS.

Security components – Networking

Perimeter security

- Azure Machine Learning relies on other Azure services for compute resources. These can be created in a virtual network.
- You can enable Azure Private Link for your AML workspace. Private Link allows you to restrict communications to your workspace from an Azure Virtual Network. **Currently in preview**, and is available in the US East, US West 2, US South Central regions.
- Private Link vs Service endpoints
  - "The main difference between the two is Private Link introduces a private IP for a given instance of the PaaS Service and the service is accessed via the private IP. A Service Endpoint uses the public IP address of the PaaS Service when accessing the service." [Source]
  - Private Link = VNET to PaaS **Instance.** Service endpoint = VNET to PaaS **Service**.
- Vnet integration
  - https://docs.microsoft.com/en-us/azure/machine-learning/how-to-enable-virtual-network

Microsoft