Data Science Bootcamp

Hyunsuk (hyssh@microsoft.com)
Microsoft Technology Center in Seattle

Microsoft Technology Centers

Bootcamp

Goal

The bootcamp will walk you through the concept of data science and enable you to start the data science project quickly

Agenda

- Data Science
  - Five questions
  - Ask the right question
  - How to make a prediction
- Data Science Process
  - Business understanding
  - Data acquisition and understanding
  - Modeling
  - Deployment
  - Customer Acceptance
- Data Science Tool
  - Access and process data
  - Train and deploy model
  - Monitor models and data

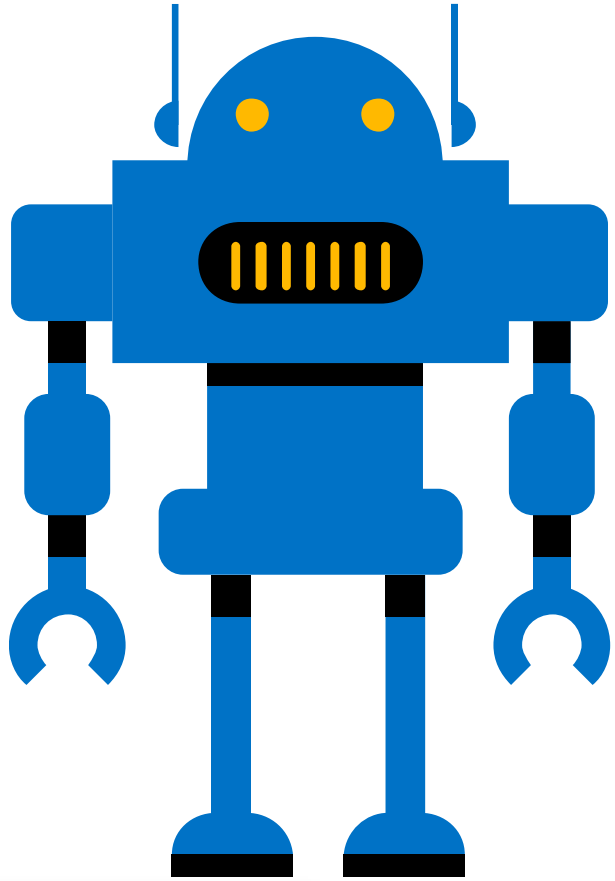
What is Data Science

The Formal one:

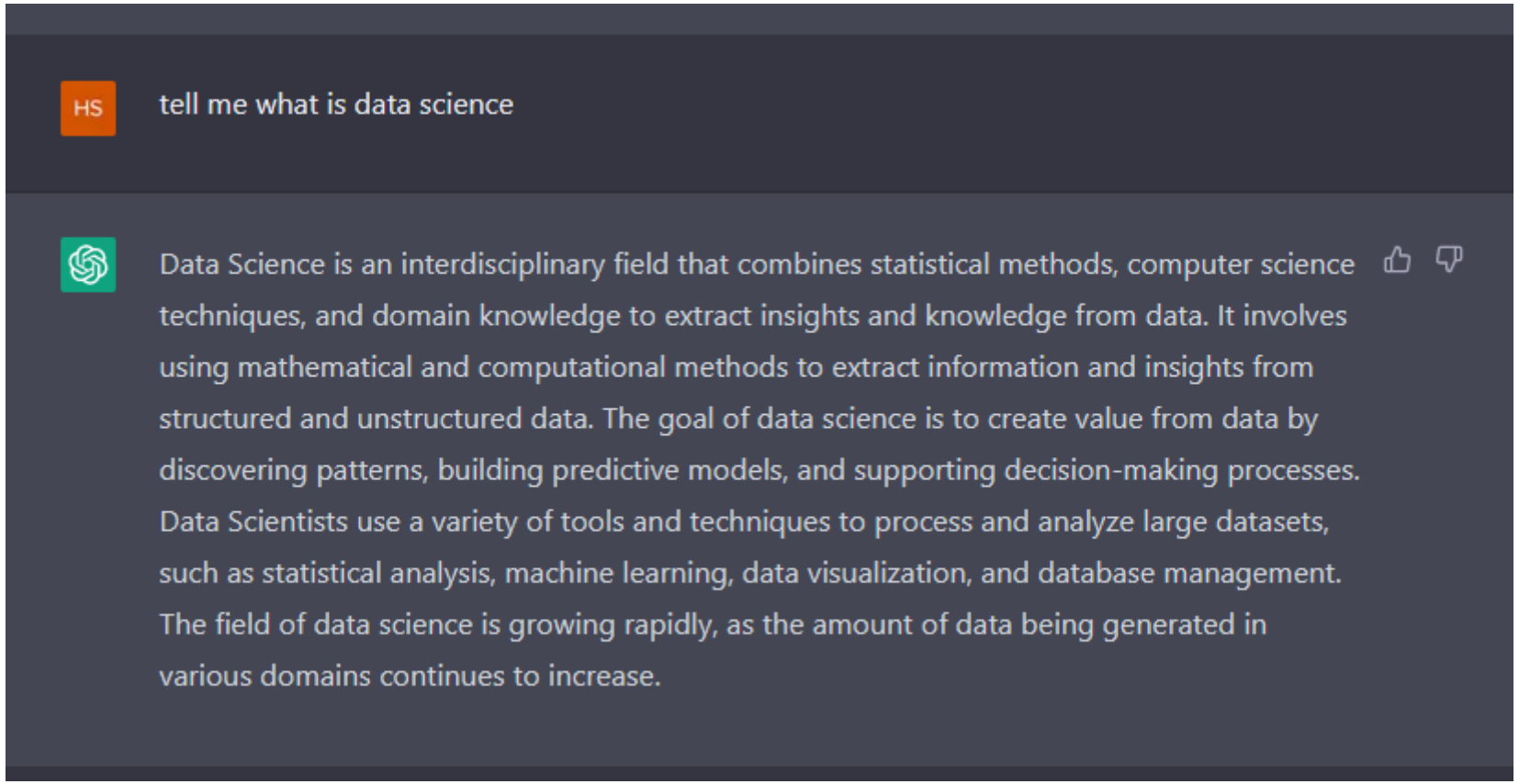
"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

A Practical Example:

Look at data. Do the thing. Better? No? Look at the data. Do something different. Better? Yes? *Do that again*. (Repeat)

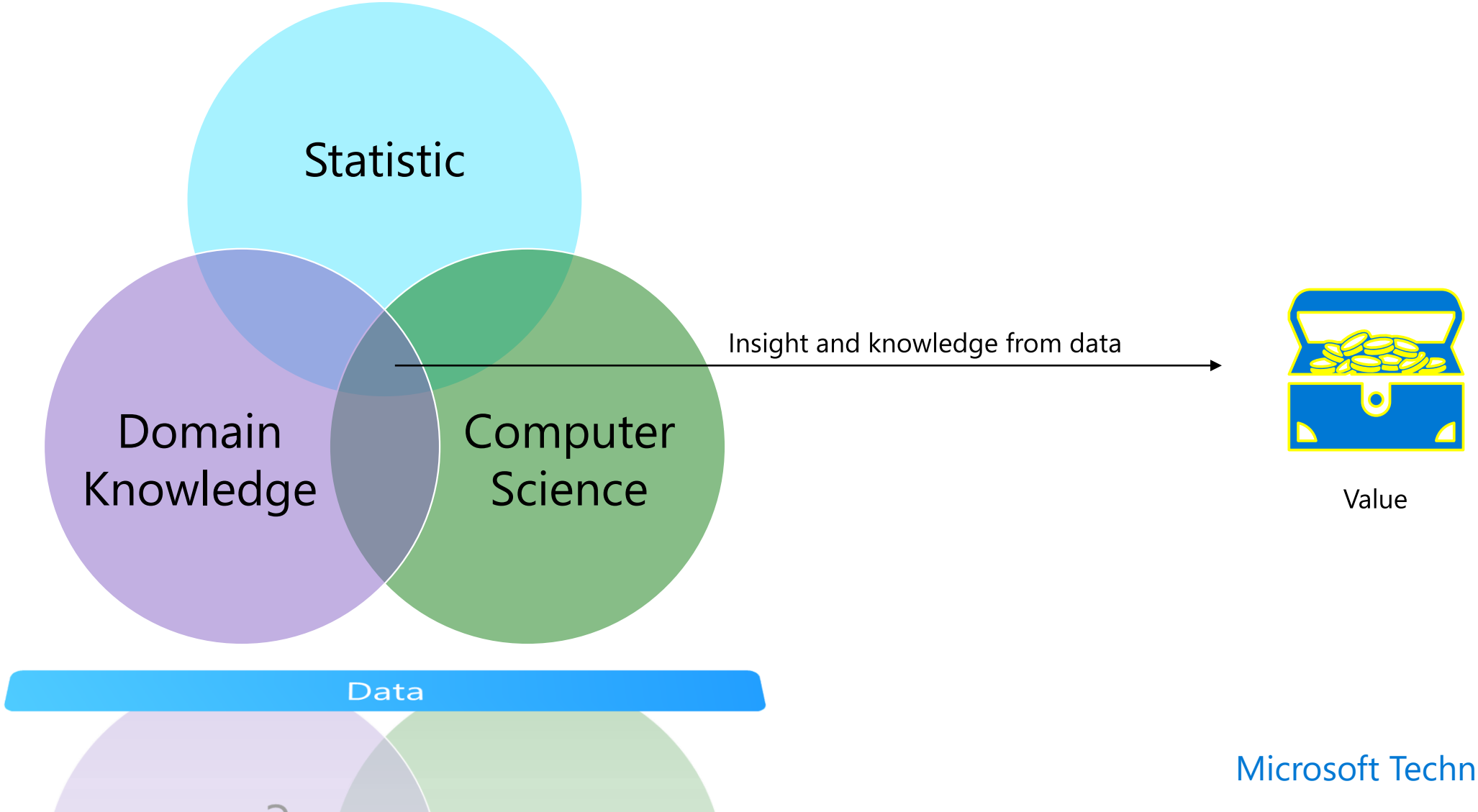


What is Data Science



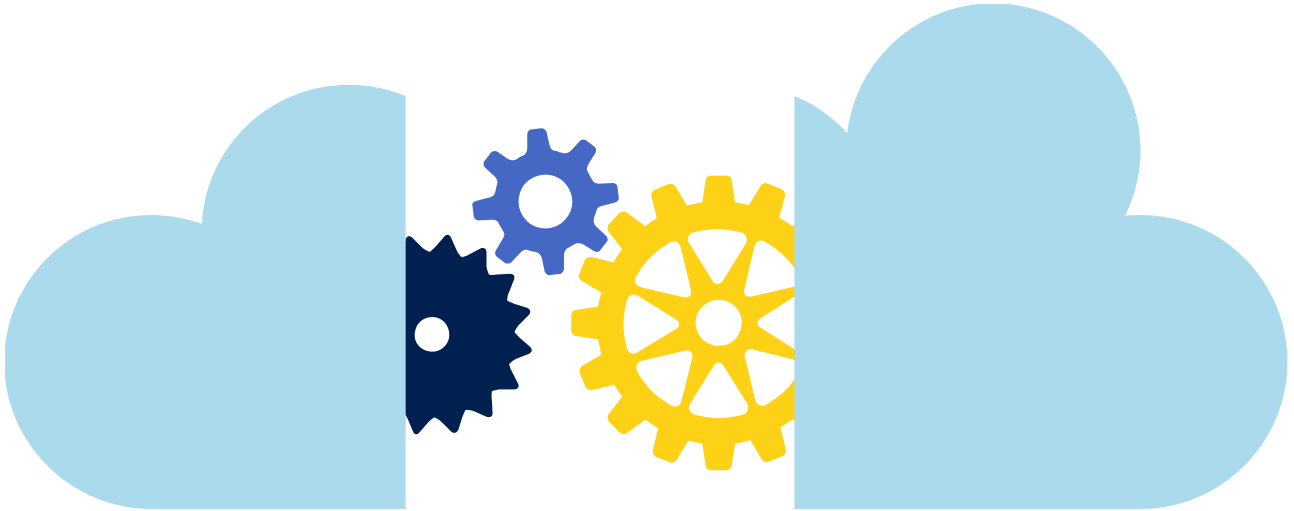
Why Data Science

Data science creates value from data by discovering patterns, building predictive models, and supporting decision-making processes.



Agenda

- **Data Science**
- **Data Science process – TDSP (Team Data Science Process)**
- **Data Science Tools**

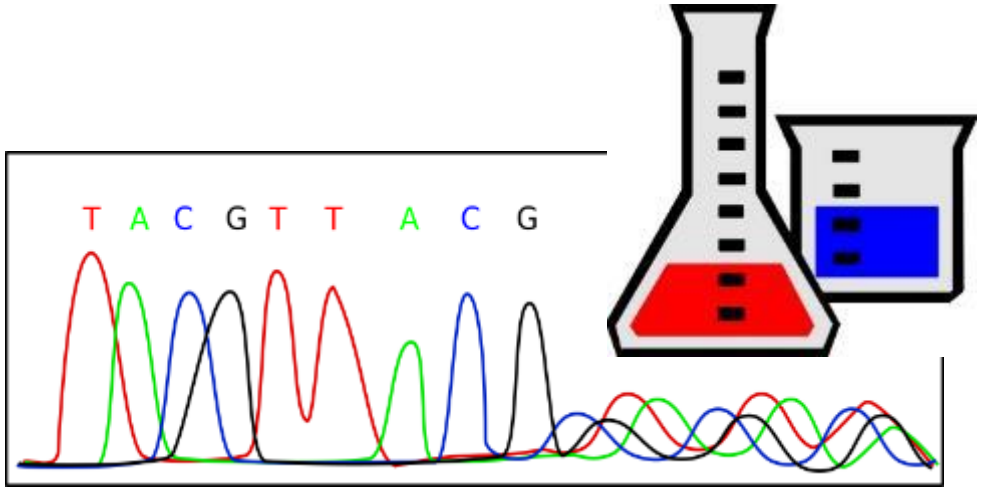


Microsoft Technology Centers

Data Science

Machine Learning

1. **Five questions**
2. **Is your data ready?**
3. **Ask the right question**
4. **Predict an answer**
5. **Copy other people's work**



Microsoft Technology Centers

1. Five questions

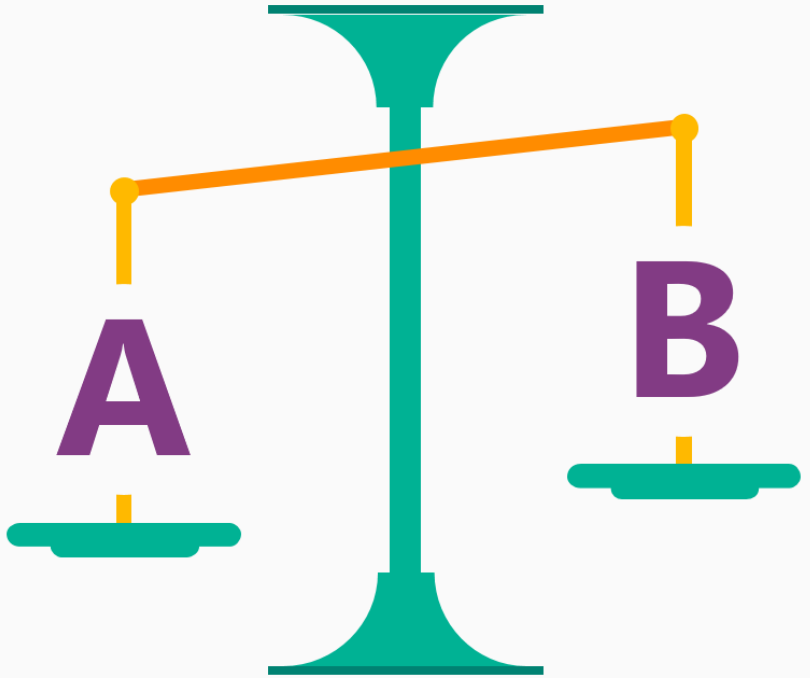
Machine Learning can answer

- Is this A or B?
- Is this weird?
- How much? How many?
- How is this organized?
- What should I do?

1. Five questions

Is this A or B?

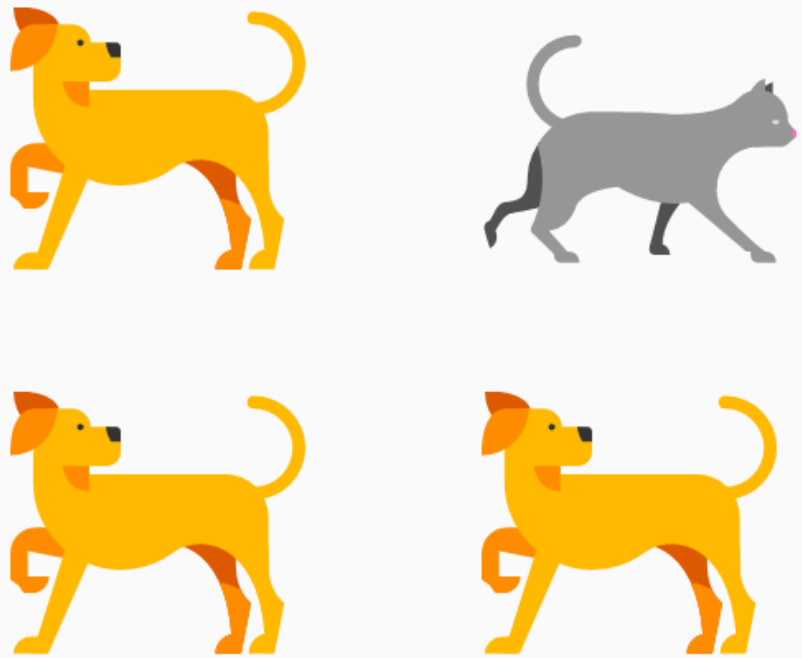
Classification algorithms



1. Five questions

Is this weird?

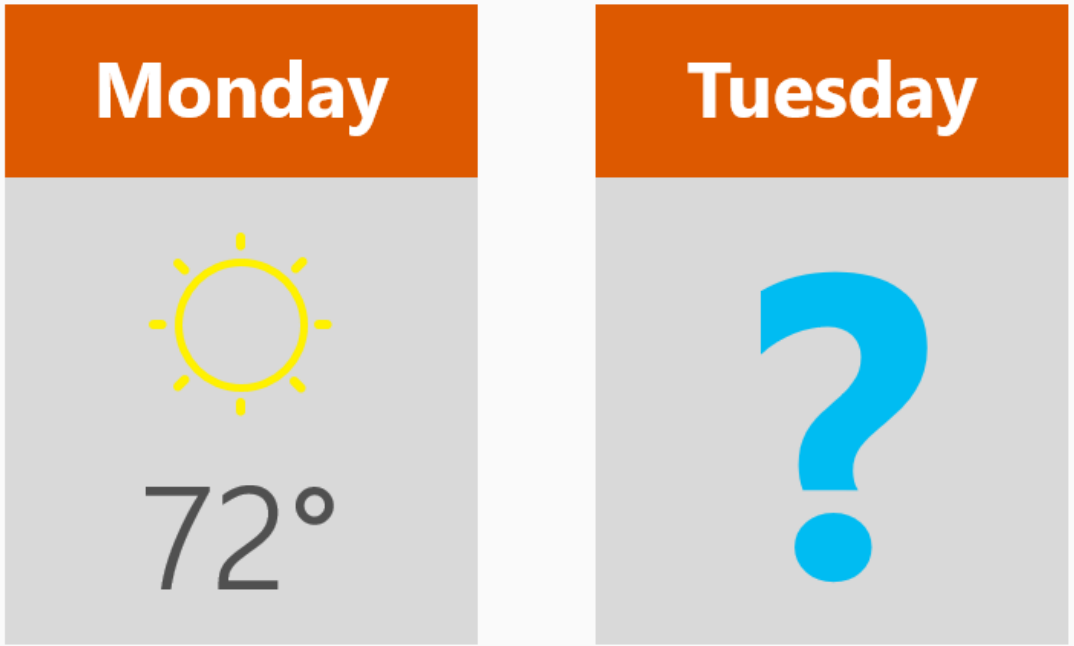
Anomaly detection algorithms



1. Five questions

How much? How many?

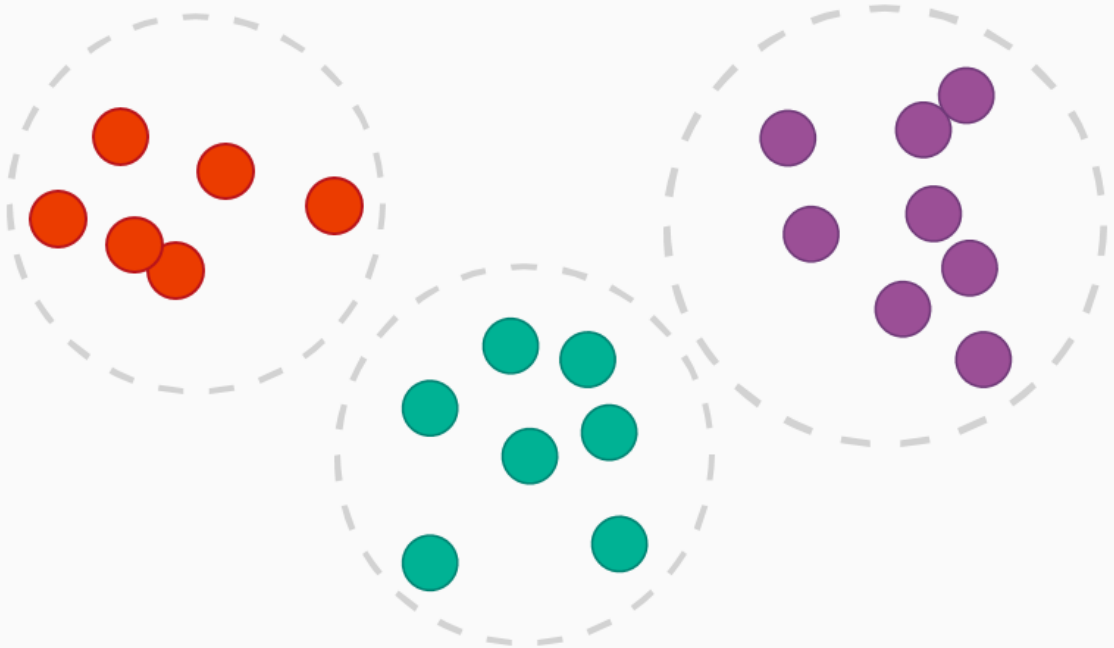
Regression algorithms



1. Five questions

How is this organized?

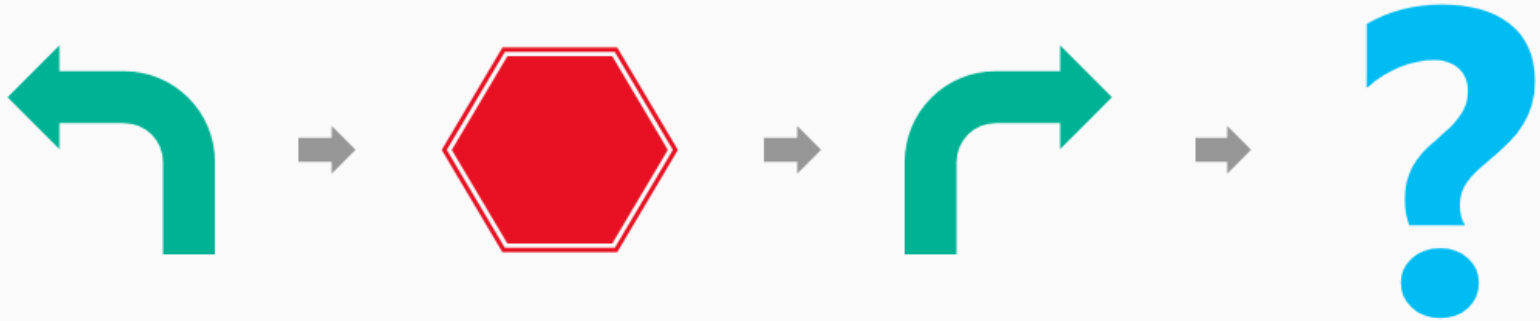
Clustering Algorithms



1. Five questions

What should I do now?

Reinforcement Learning Algorithms



2. Is your data ready?

- Relevant
- Connected
- Accurate
- Enough to work with

Irrelevant Data

Price of milk ($/gal)	Red Sox batting avg.	Blood alcohol content (%)
3.79	.304	.03
3.45	.320	.09
4.06	.259	.01
3.89	.298	.05
4.12	.332	.13
3.92	.270	.06
3.23	.294	.10

Relevant Data

Body mass (kg)	Margaritas	Blood alcohol content (%)
103	3	.03
67	5	.09
87	1	.01
52	2	.05
73	5	.13
79	3	.06
110	7	.10

2. Is your data ready?

- Relevant
- Connected
- Accurate
- Enough to work with

Disconnected Data

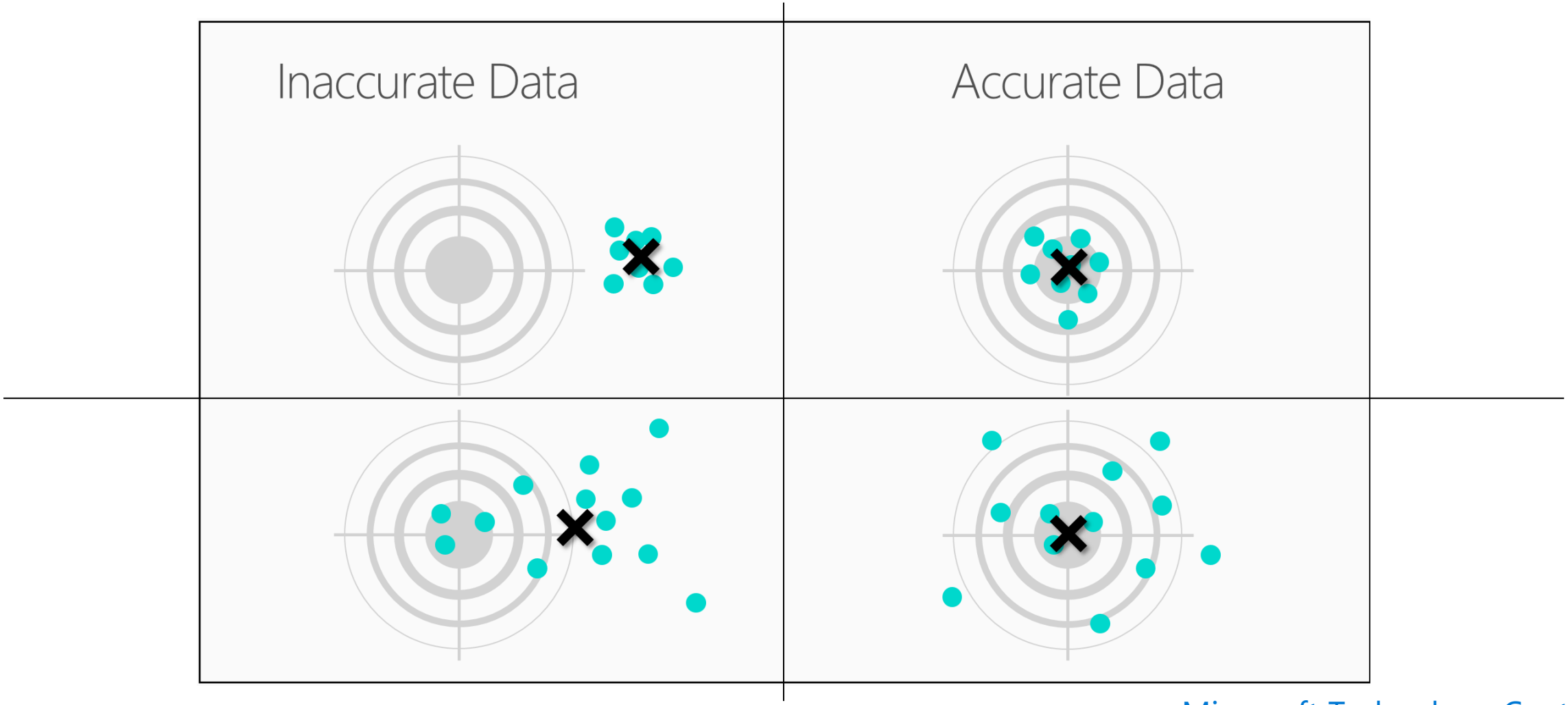
Grill temp. (Fahrenheit)	Weight of beef patty (lb)	Burger rating (out of 10)
	.33	8.2
	.24	5.6
550		7.8
725	.45	9.4
600		8.2
625		6.8
	.49	4.2

Connected Data

Grill temp. (Fahrenheit)	Weight of beef patty (lb)	Burger rating (out of 10)
575	.33	8.2
550	.24	5.6
550	.69	7.8
725	.45	9.4
600	.57	8.2
625	.36	6.8
550	.49	4.2

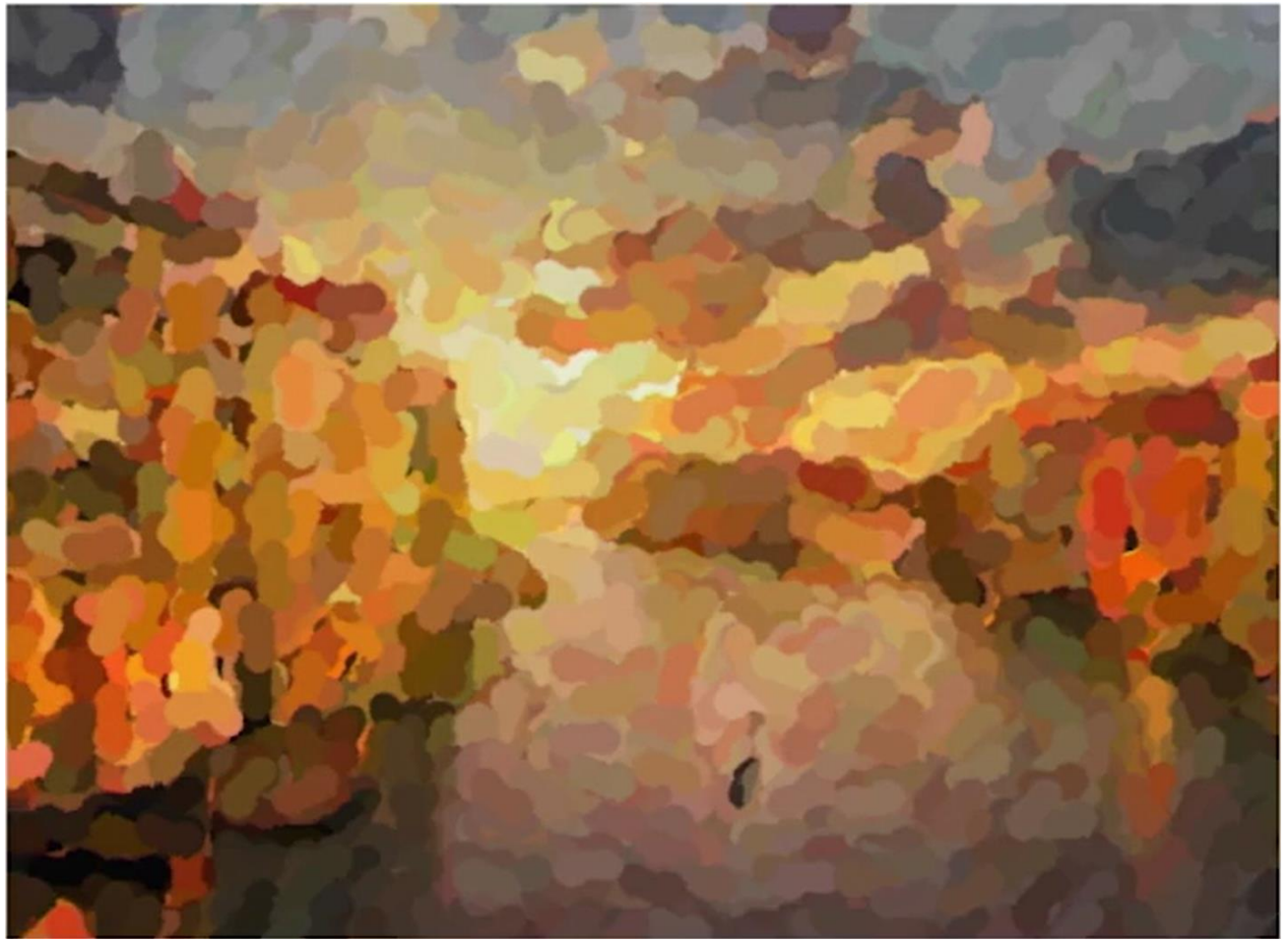
2. Is your data ready?

- Relevant
- Connected
- Accurate
- Enough to work with



2. Is your data ready?

- Relevant
- Connected
- Accurate
- Enough to work with



2. Is your data ready?

- Relevant
- Connected
- Accurate
- Enough to work with

2. Is your data ready?

- Relevant
- Connected
- Accurate
- Enough to work with

3. Ask the right question

Sharp question

3. Ask the right question

Sharp question

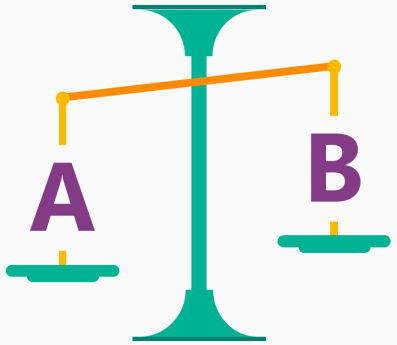
- Will this tire fail in the next 1,000 miles: Yes or no?
- Which brings in more customers: a $5 coupon or a 25% discount?
- If you have a car with pressure gauges, you might want to know: Is this pressure gauge reading normal?
- If you're monitoring the internet, you'd want to know: Is this message from the internet typical?
- What will my stock's sale price be next week?
- What will the temperature be next Tuesday?
- What will my fourth quarter sales be?
- Which printer models fail the same way?

3. Ask the right question

Reformulate your question

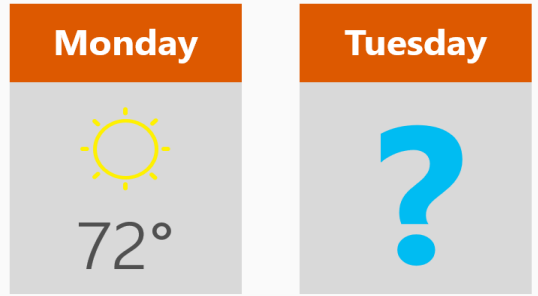
Is this A or B?

Classification algorithms



How much? How many?

Regression algorithms



What will my stock's sale price be next week?

What will the temperature be next Tuesday?

What will my fourth quarter sales be?

Which printer models fail the same way?

4. Predict an answer

How much 1.35 carat diamond cost these days?

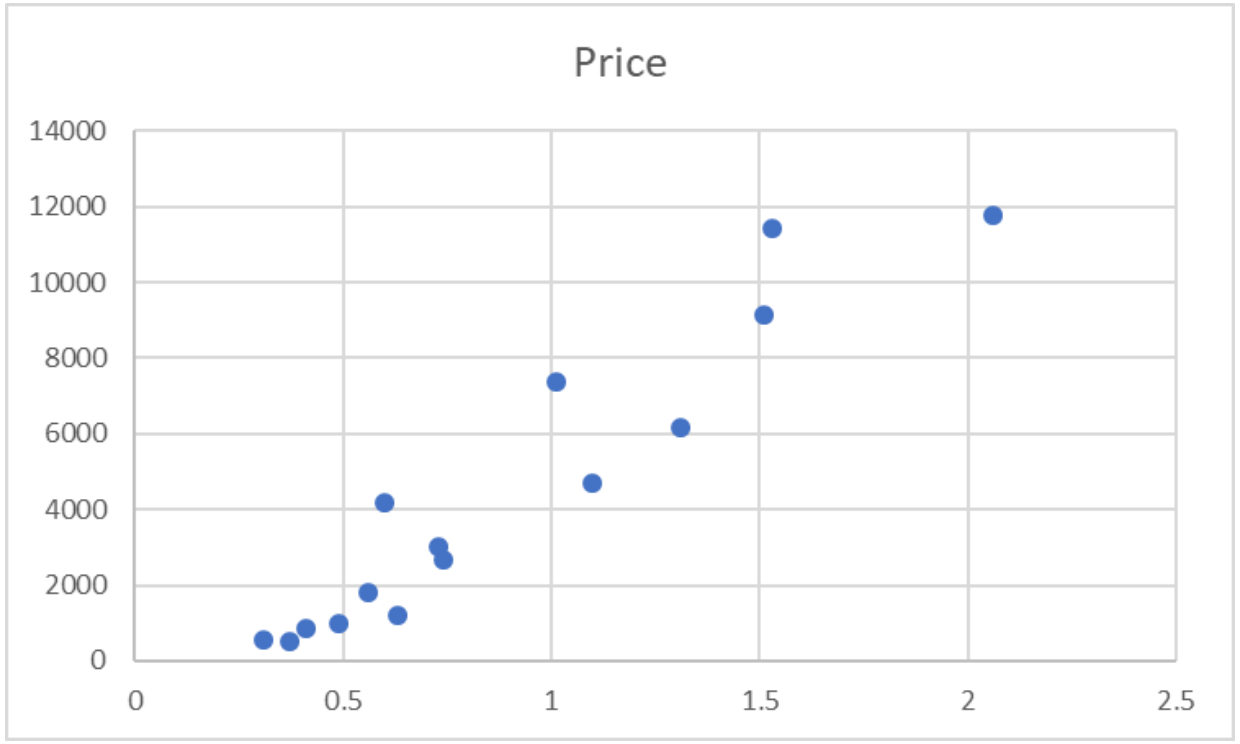
Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	9140
0.37	493
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	1190
0.6	4172
2.06	11764
1.1	4682
1.31	6171

4. Predict an answer

Understand data

- Plot the existing data

Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	9140
0.37	493
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	1190
0.6	4172
2.06	11764
1.1	4682
1.31	6171

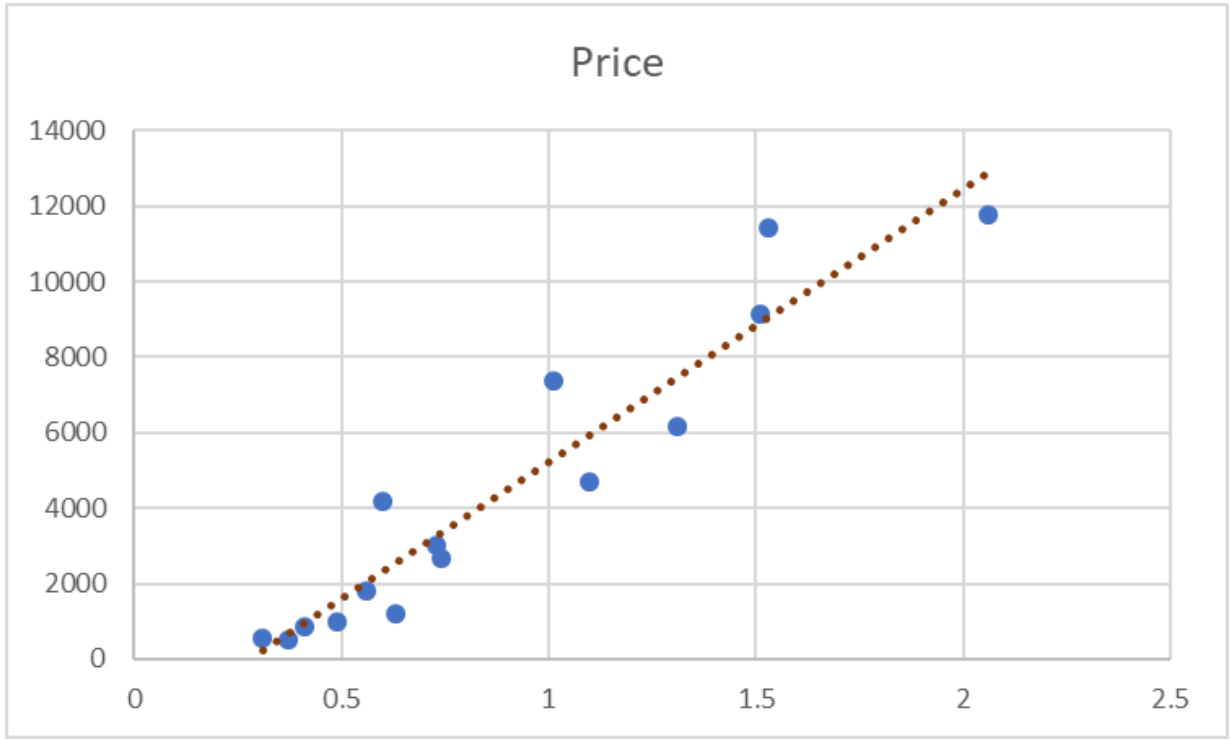


4. Predict an answer

Create a model

- Draw the model through the data points

Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	9140
0.37	493
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	1190
0.6	4172
2.06	11764
1.1	4682
1.31	6171

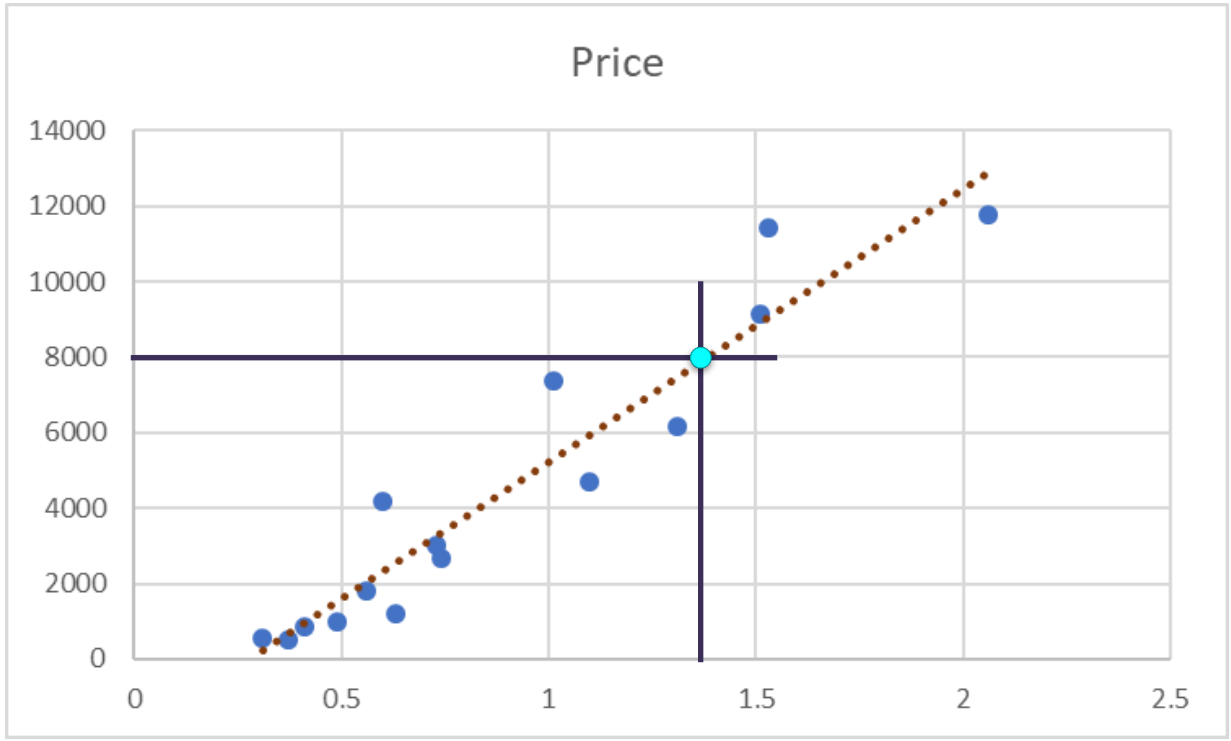


4. Predict an answer

Use the model to find the answer

- How much 1.35 carat diamond cost these days?

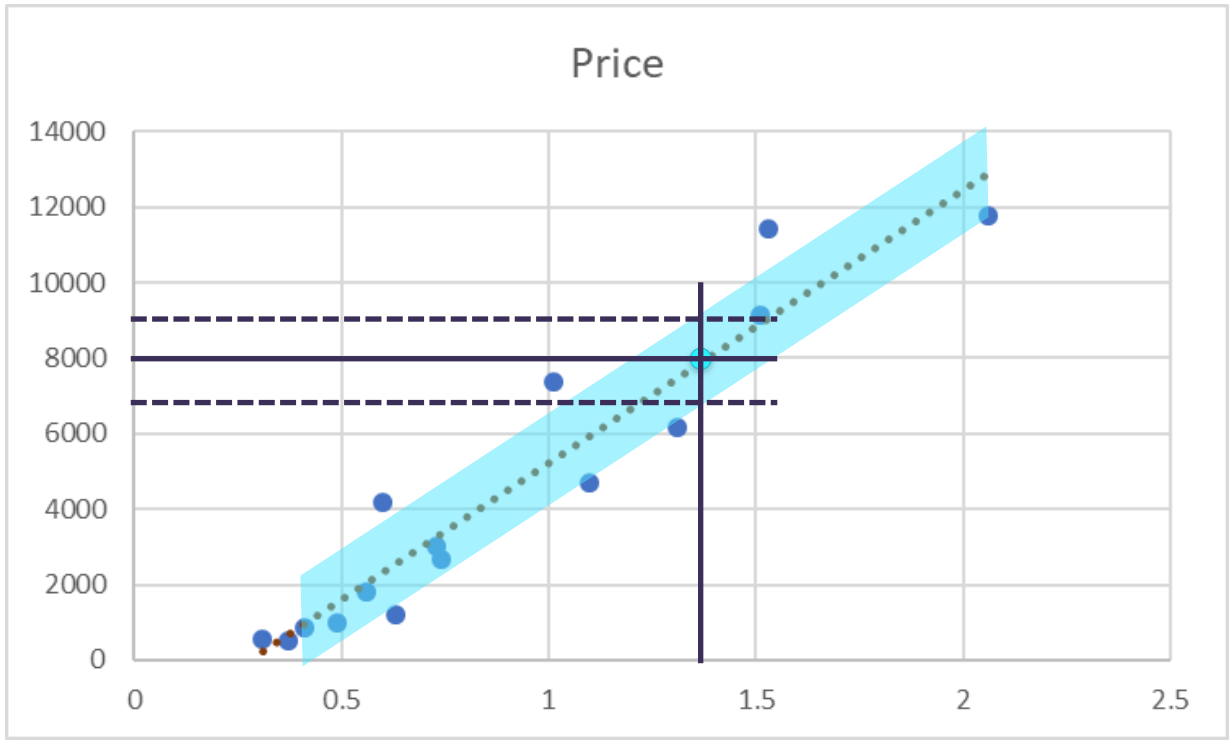
Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	9140
0.37	493
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	1190
0.6	4172
2.06	11764
1.1	4682
1.31	6171



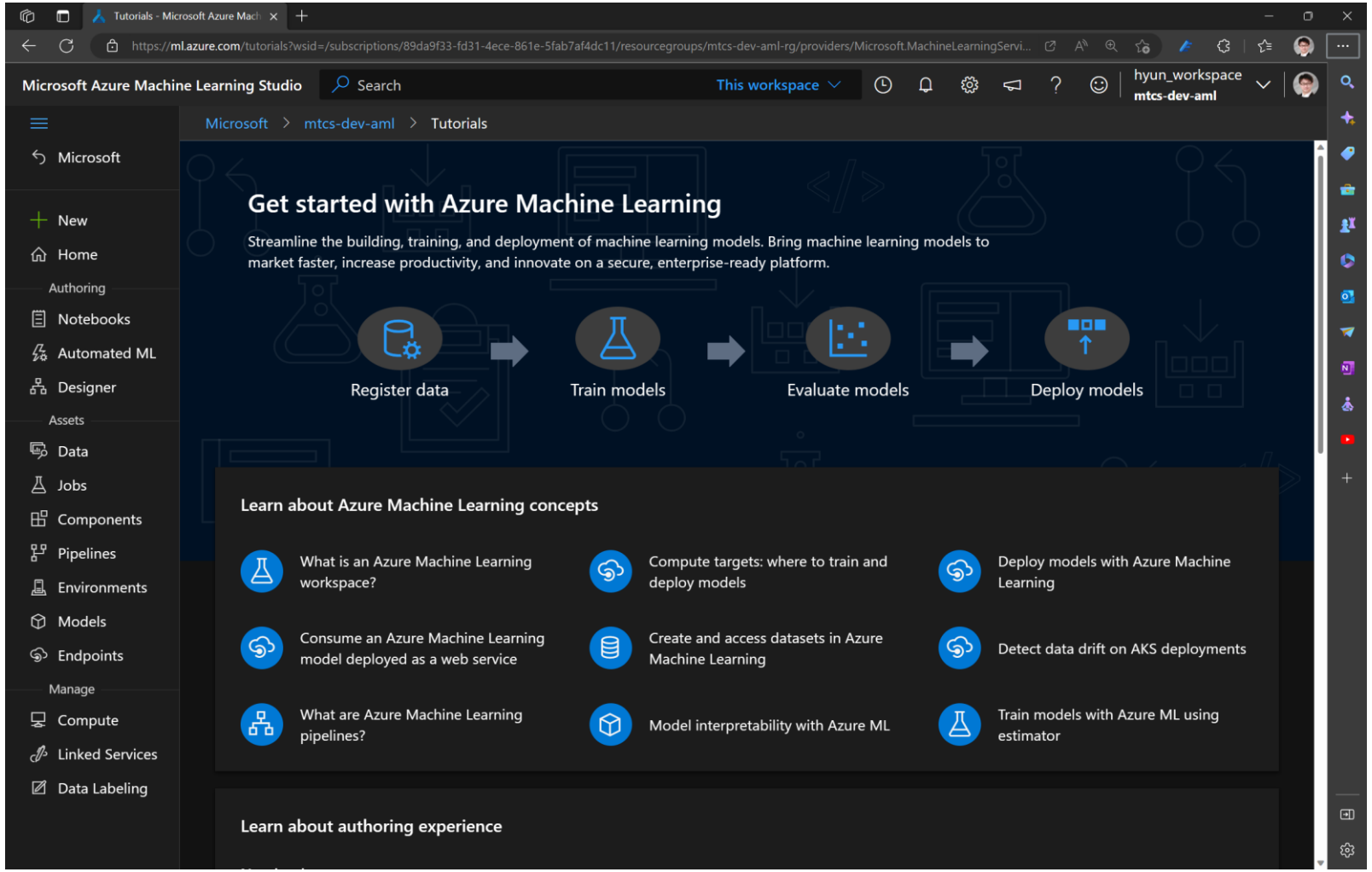
4. Predict an answer

Create a confidence interval

Carats	Price
1.01	7366
0.49	985
0.31	544
1.51	9140
0.37	493
0.73	3011
1.53	11413
0.56	1814
0.41	876
0.74	2690
0.63	1190
0.6	4172
2.06	11764
1.1	4682
1.31	6171



5. Copy other people's work

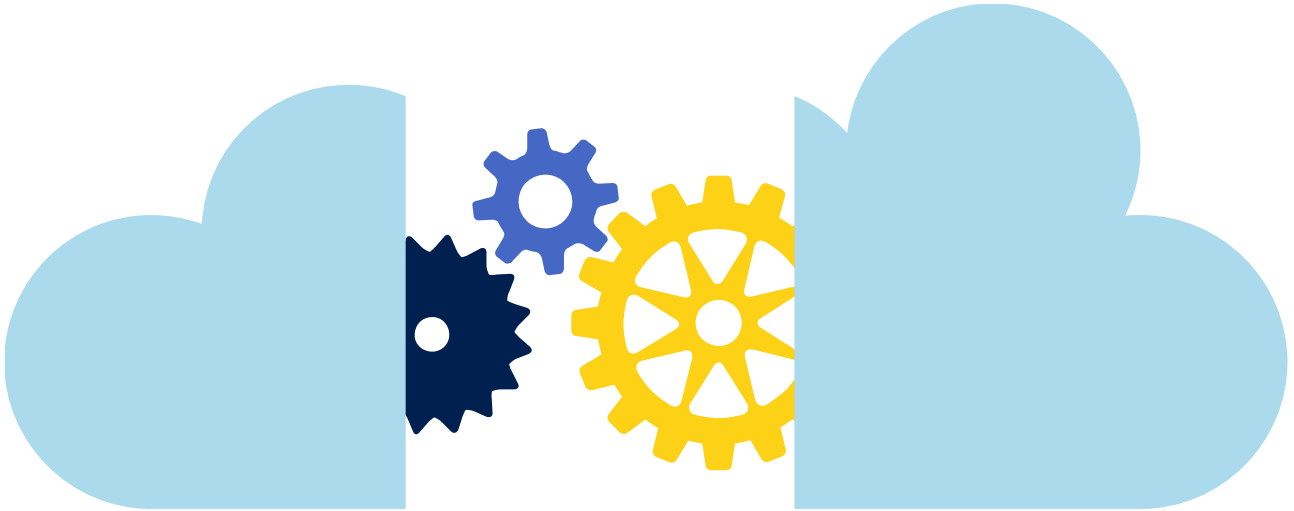


Coffee break

10 minutes

Agenda

- Data Science
- **Data Science process – TDSP (Team Data Science Process)**
- Data Science Tools

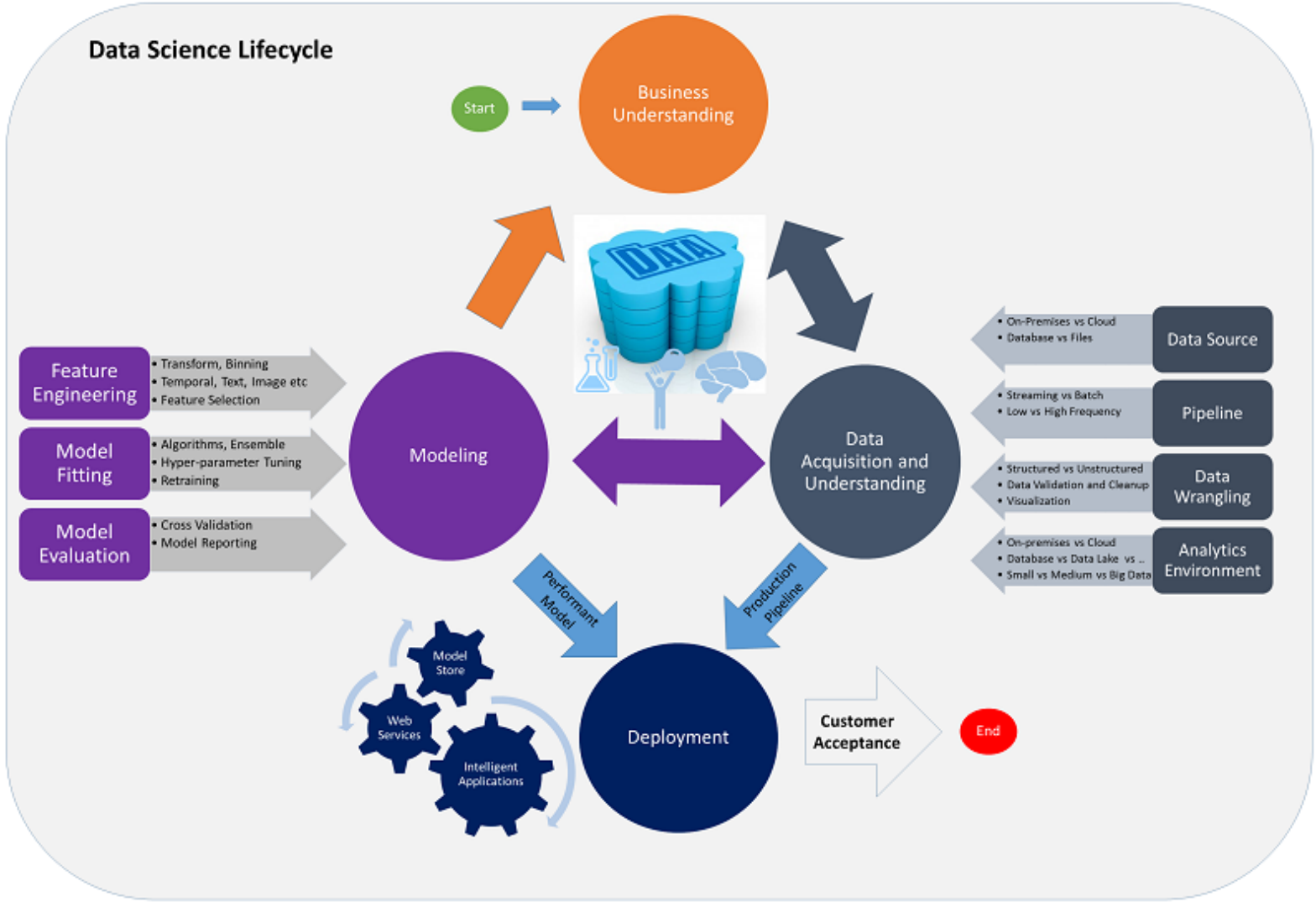


Microsoft Technology Centers

Data Science process – TDSP (Team Data Science Process)

Data Science Lifecycle

1. Business understanding
2. Data acquisition and understanding
3. Modeling
4. Deployment
5. Customer Acceptance



Microsoft Technology Centers

Cross-industry standard process for data mining - Wikipedia

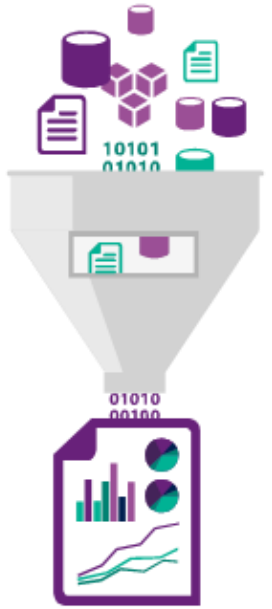
1. Business Understanding

Analyze business needs

- Define objectives
  - Identify key business variables
  - Project goal
  - Question
  - Success metrics

- Identify data sources
  - Data sources that contain known examples of answers
  - Data characteristics
  - Data quality
  - Tools and language
  - Security
    - Encryption, Audit, Access control

- Artifacts
  - Documents
  - Data sources
  - Data



1. Business Understanding

Business sample case

Wide World Importers is a company that imports and distributes products in multiple countries around the globe.

With several thousand employees, Information Technology is at the heart of our business operations, and has a significant cost.

Since we handle materials in multiple countries, we have a lot of private data, financial information, and other targets which have a high security profile. We are concerned with both external and internal attacks. In addition, many of our employees work in remote locations, some on ships and other challenging environments.

All of our IT systems have been modernized, and we're taking in a significant amount of semi-structured data from computing devices – most of it real-time. After talking with our IT leadership, we need a way to determine anomalies within the data streams we get, and have a way to observe the anomalies in a dashboard so that we can respond to outages, threats, and changes quickly.

1. Business Understanding

Design statements

Wide World Importers is a company that imports and distributes products in multiple countries around the globe.

With several thousand employees, Information Technology is at the heart of our business operations, and has a significant cost.

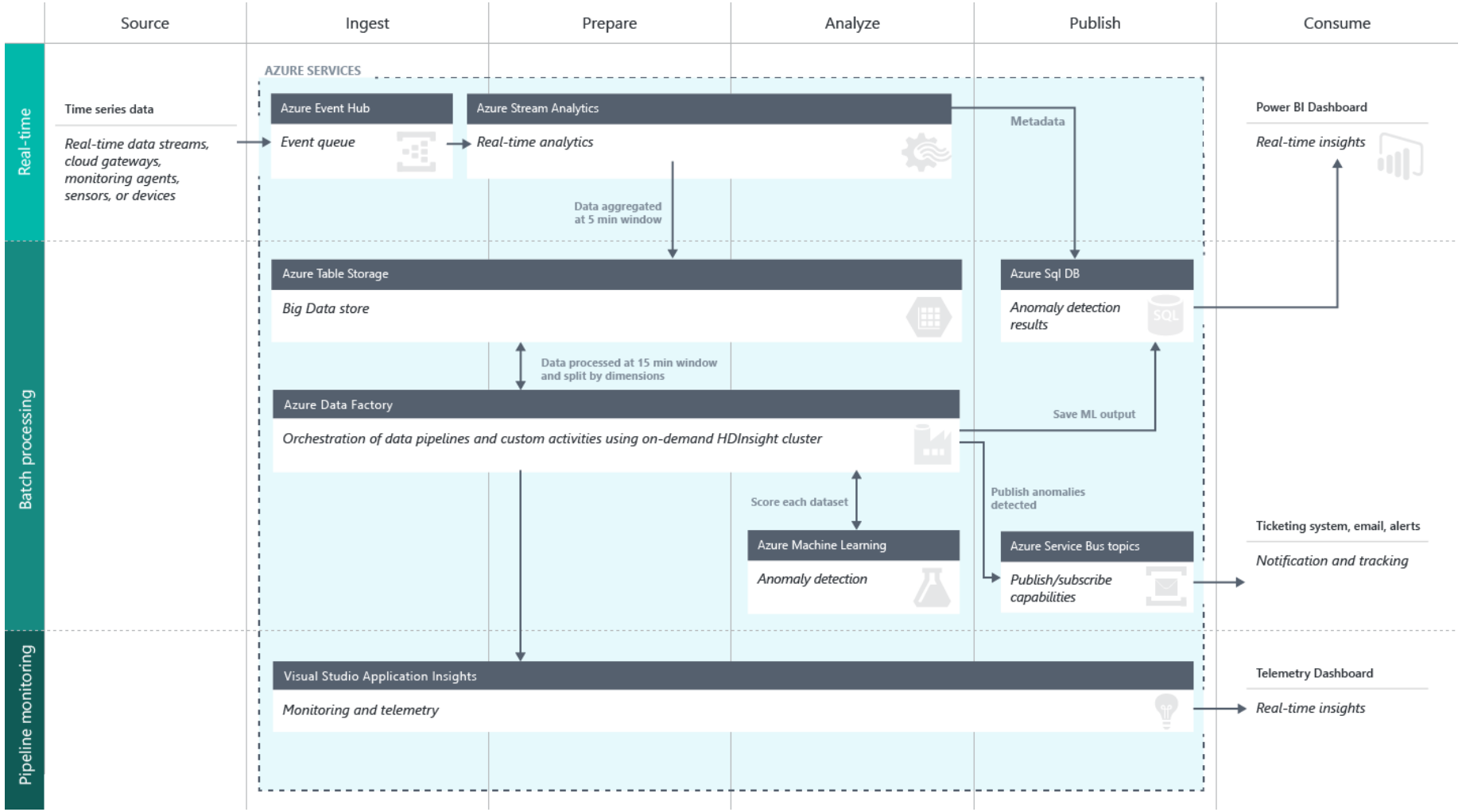
Since we handle materials in multiple countries, we have a lot of private data, financial information, and other targets which have a high security profile. We are concerned with both external and internal attacks. In addition, many of our employees work in remote locations, some on ships and other challenging environments.

All of our IT systems have been modernized, and we're taking in a significant amount of semi-structured data from computing devices – most of it real-time. After talking with our IT leadership, we need a way to determine anomalies within the data streams we get, and have a way to observe the anomalies in a dashboard so that we can respond to outages, threats, and changes quickly.

1. Business Understanding

Data Science process – TDSP

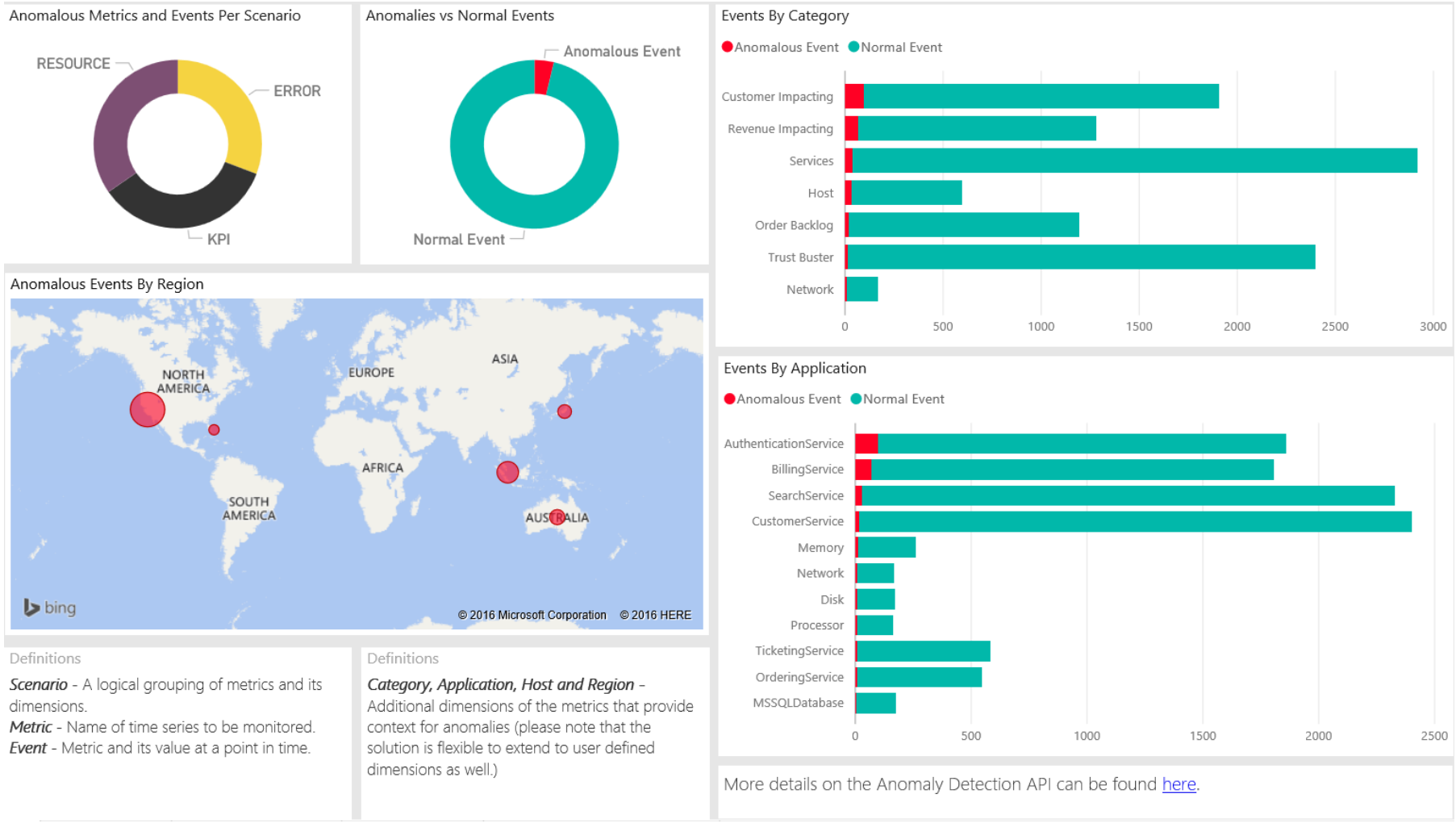
Solution Diagram



1. Business Understanding

Data Science process – TDSP

Business Insights



2. Data Acquisition and Understanding

Acquire and understand data

- Ingest the data
  - Move the data from source to target location
- Explore the data
  - Data cleaning
    - Incomplete
    - Noisy
    - Inconsistent
  - Review the quality of data
  - Have better understand the patterns
  - Iterative works
- Set up a data pipeline
  - Score new data
  - Refresh the data regularly
  - Pipeline may be batch-based or a streaming or hybrid
- Artifacts
  - Data quality report
  - Solution architecture
  - Checkpoint decision

Exploratory Data Analysis

Demo

3. Modeling

Train models

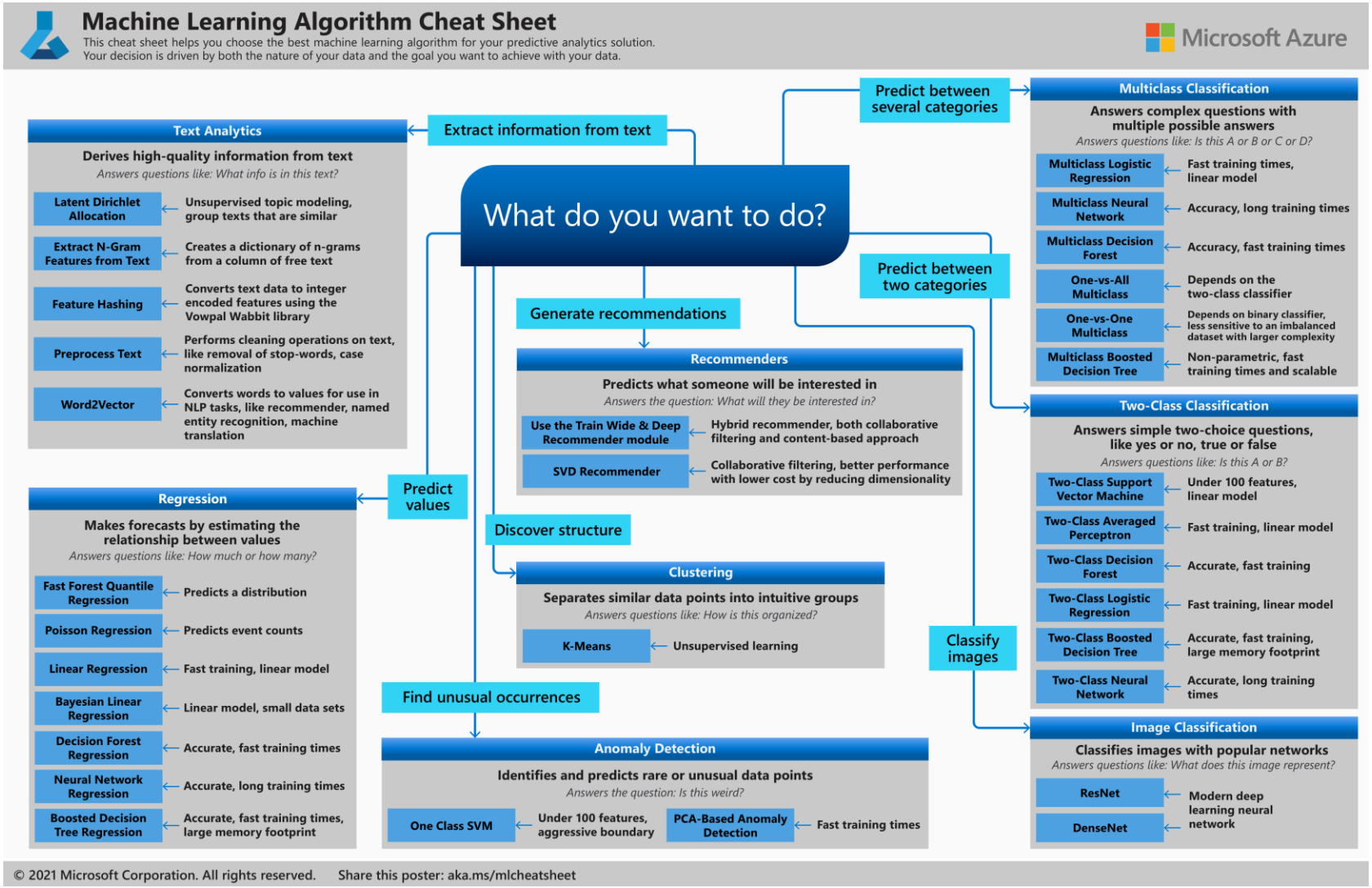
- Goals
  - Optimal data features for the machine learning model
  - Build informative machine learning model that predicts the target most accurately
  - Experiment for machine learning model that is suitable for production

- How to do it
  - Feature engineering
    - Aggregate and transform the raw variables to create the features
    - Requires a creative combination of domain expertise and insights
  - Model training
    - How to choose algorithms
    - Split, Build, Evaluate, Determine the best solution

- Artifacts
  - Feature sets
  - Modeling report
  - Checkpoint decision

3. Modeling

How to choose algorithms

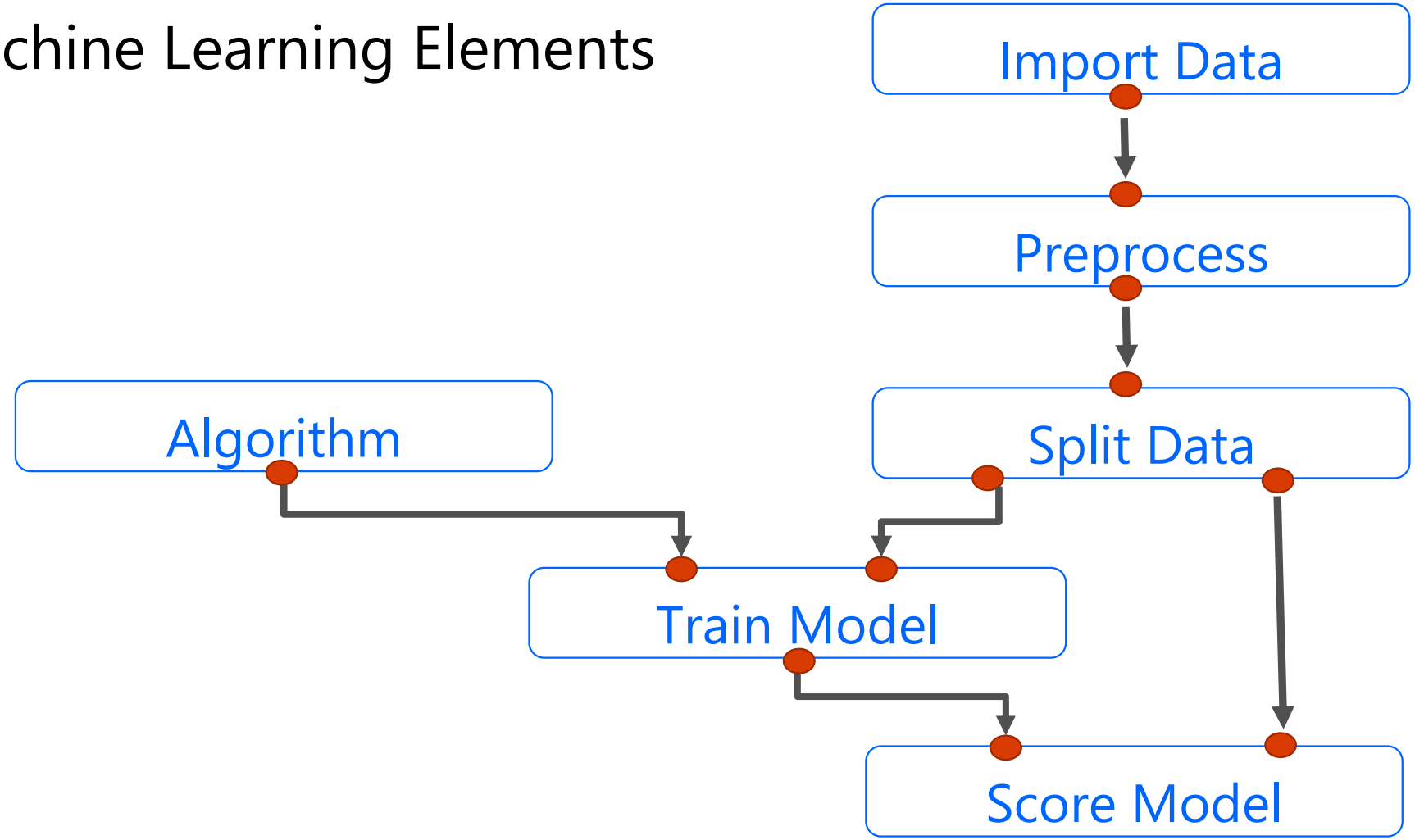


Machine Learning Algorithm Cheat Sheet - designer - Azure Machine Learning | Microsoft Learn

3. Modeling

Model Training

- Basic Machine Learning Elements



4. Deployment

Operationalize models

- Operationalize the model
  - Apply trained model to application for real-time or batch basis predictions
  - Expose the model as API interface to consume it easy

- Artifacts
  - Status dashboard of system health and key metrics
  - Final modeling report with deployment details
  - Final solution architecture document

5. Customer Acceptance

Finalize the project deliverables

- System validation
  - Confirm the deployed model and pipeline are meeting customer needs
  - Monitor systems
- Project hand-off
- Artifacts
  - Project final report

Data Science process – TDSP

Conclusion

Team Data Science Process Lifecycle

- Increase the chance of a successful completion of a complex data science project
- Continue to move a data science project forward toward a clear engagement end point
- Modeled as a sequence of iterated steps

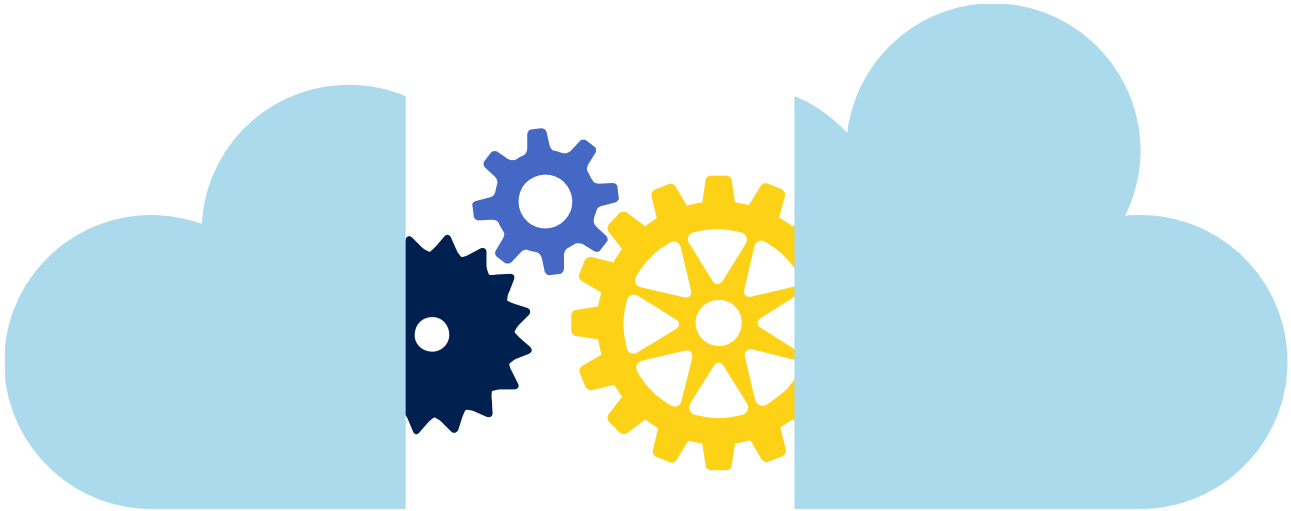
Image from Cross-industry standard process for data mining - Wikipedia

Coffee break

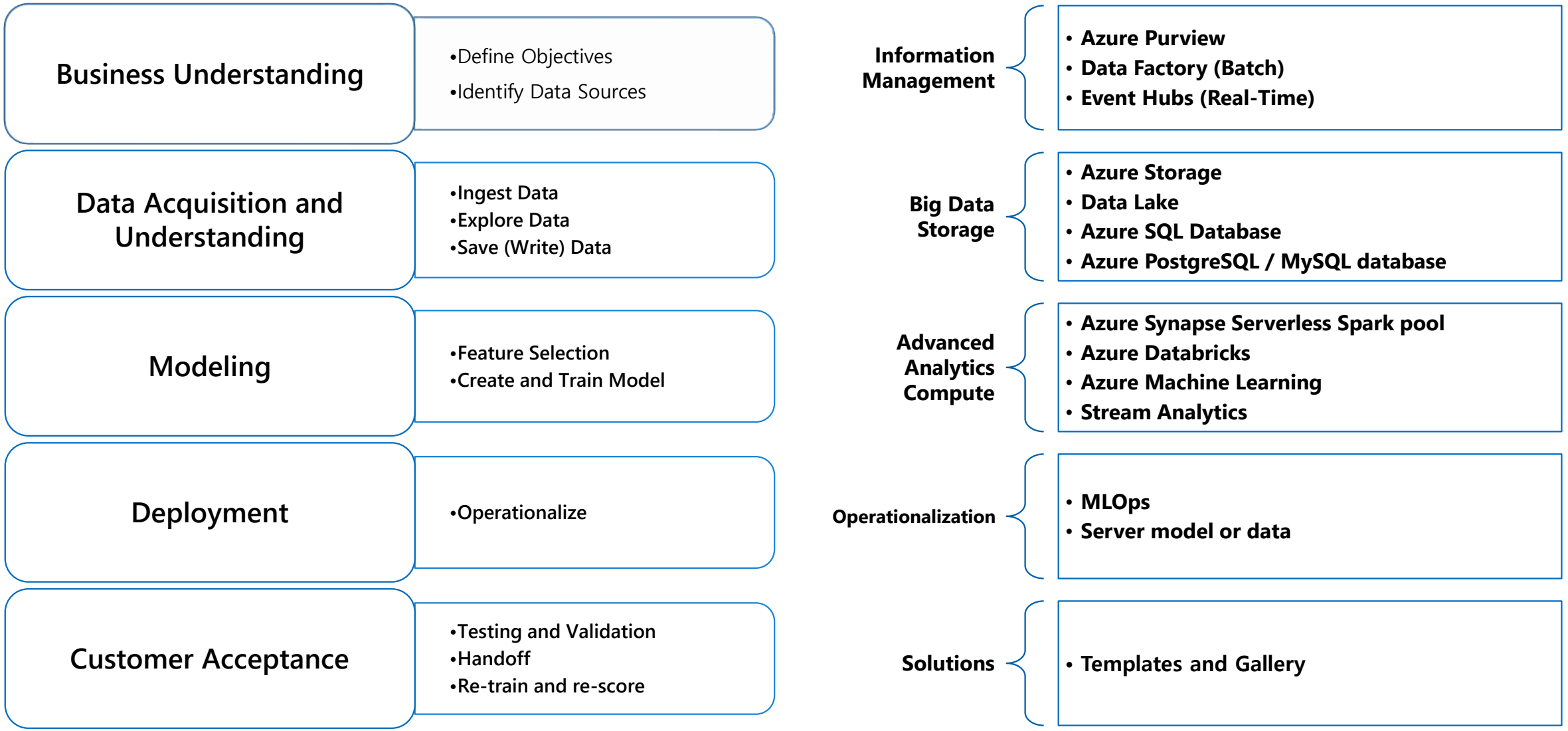
10 minutes

Agenda

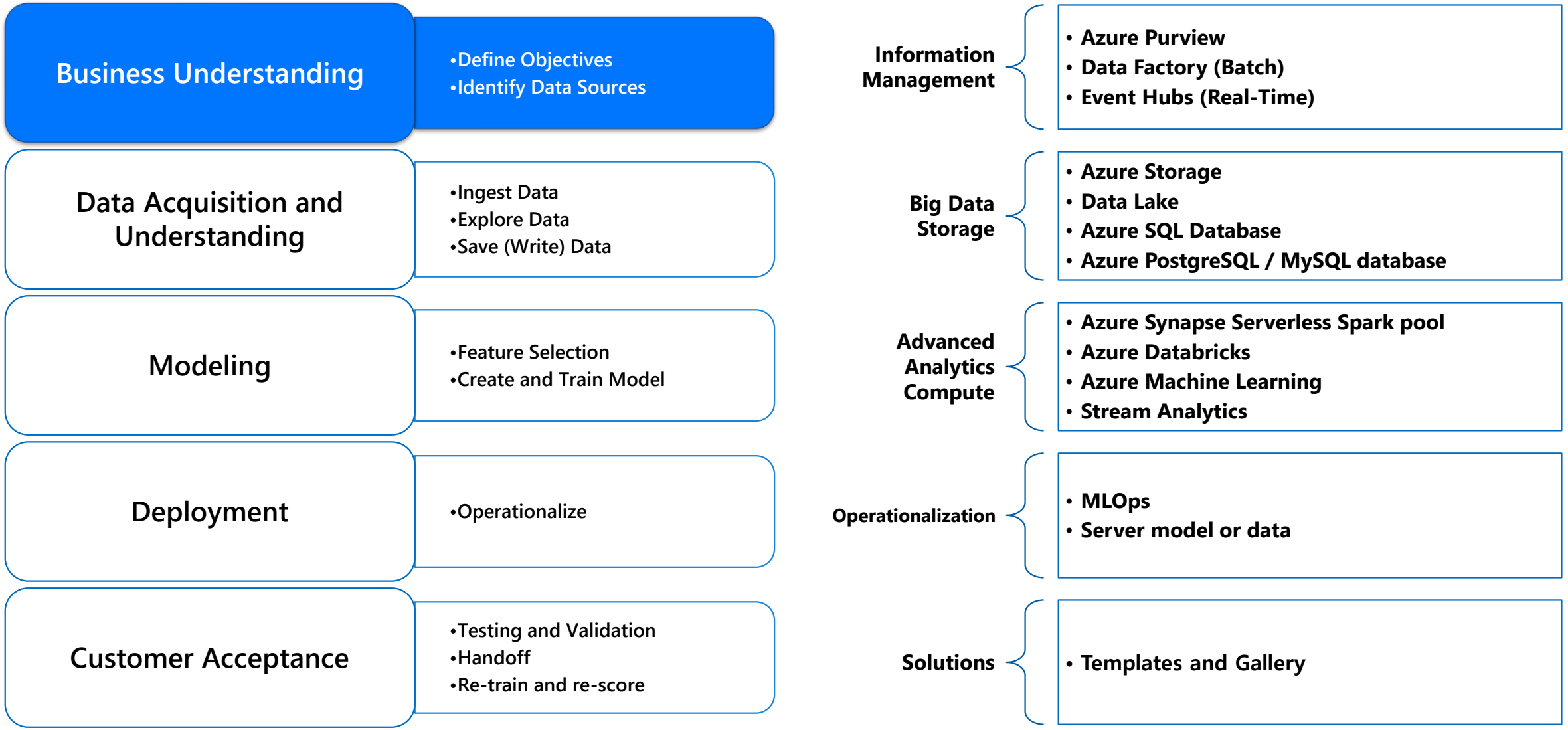
- Data Science
- Data Science process – TDSP (Team Data Science Process)
- **Data Science Tools**



Data Science Tools



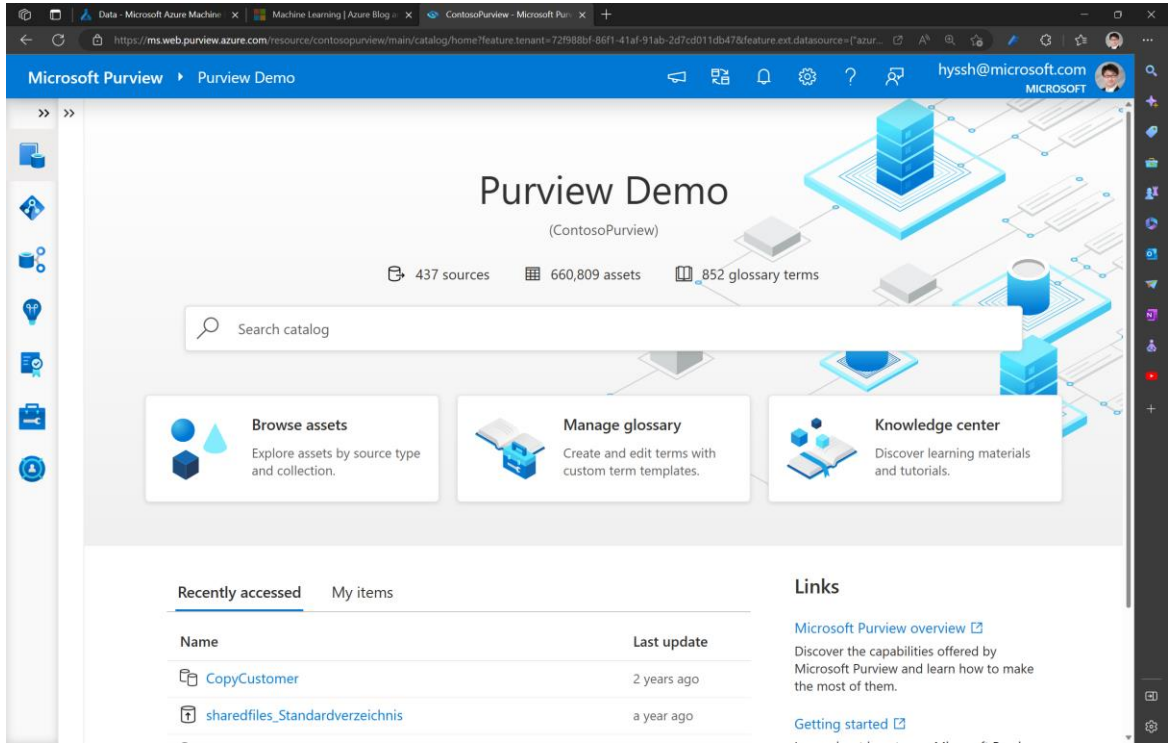
Data Science Tools



1. Business Understanding

Discover data

- Find data sources using Azure Purview



- Spend less time looking for data, and more time getting value from it
- Register enterprise data sources, discover data assets and unlock their potential, and capture tribal knowledge to make data understandable
- Bridge the gap between IT and the business, allowing everyone to contribute their insights, tags, and descriptions
- Intuitive search and filtering to understand the data sources and their purpose
- Let your data live where you want; connect using tools you choose
- Integrate into existing tools and processes with open REST APIs

Data Science Tools

Business Understanding	•Define Objectives •Identify Data Sources
Data Acquisition and Understanding	•Ingest Data •Explore Data •Save (Write) Data
Modeling	•Feature Selection •Create and Train Model
Deployment	•Operationalize
Customer Acceptance	•Testing and Validation •Handoff •Re-train and re-score

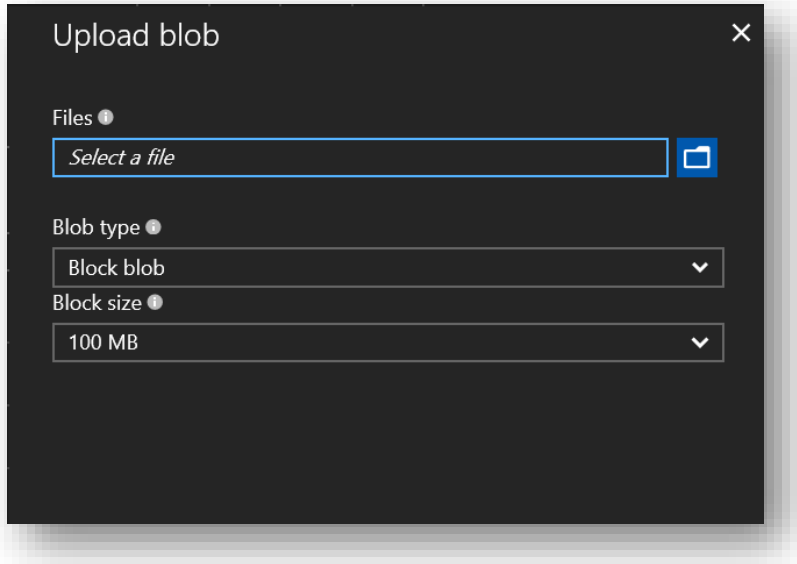
Information Management	• **Azure Purview** • **Data Factory (Batch)** • **Event Hubs (Real-Time)**
Big Data Storage	• **Azure Storage** • **Data Lake** • **Azure SQL Database** • **Azure PostgreSQL / MySQL database**
Advanced Analytics Compute	• **Azure Synapse Serverless Spark pool** • **Azure Databricks** • **Azure Machine Learning** • **Stream Analytics**
Operationalization	• **MLOps** • **Server model or data**
Solutions	• **Templates and Gallery**

2. Data Acquisition and Understanding

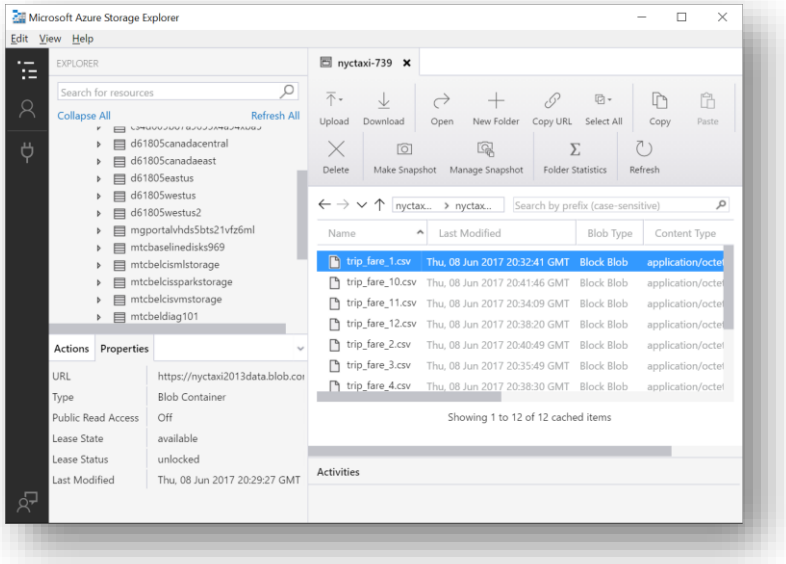
Data Science Tools

Acquire and understand data

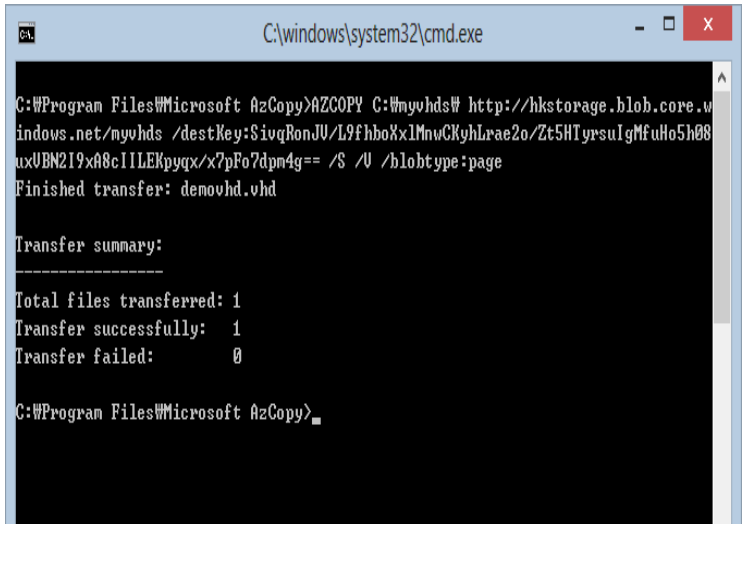
- Azure portal
- PowerShell
- Azure Data Factory
- Azure Event Hubs
- Azure storage SDKs (.NET, Node.js, python, C++, etc.)
- AzCopy (blob, file, and table only)
- Import/Export service



[Azure Portal]



[Azure Storage Explorer]



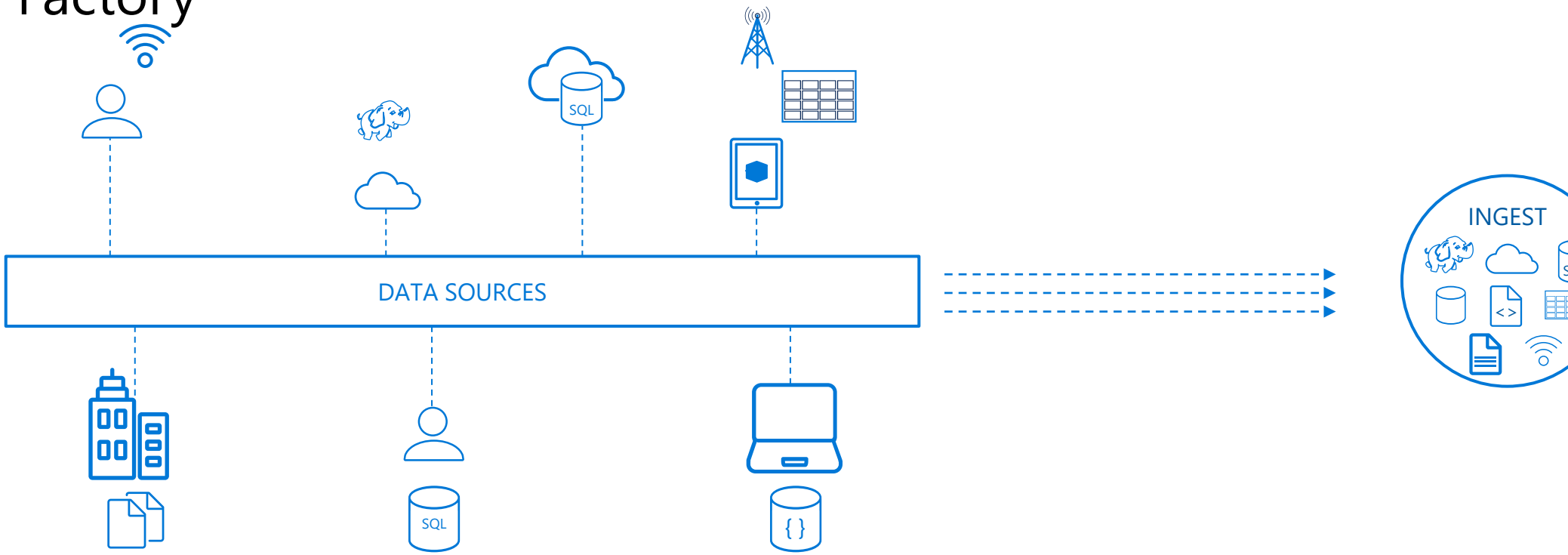
[AZcopy]

Microsoft Technology Centers

2. Data Acquisition and Understanding

Ingest Data

- Azure Data Factory



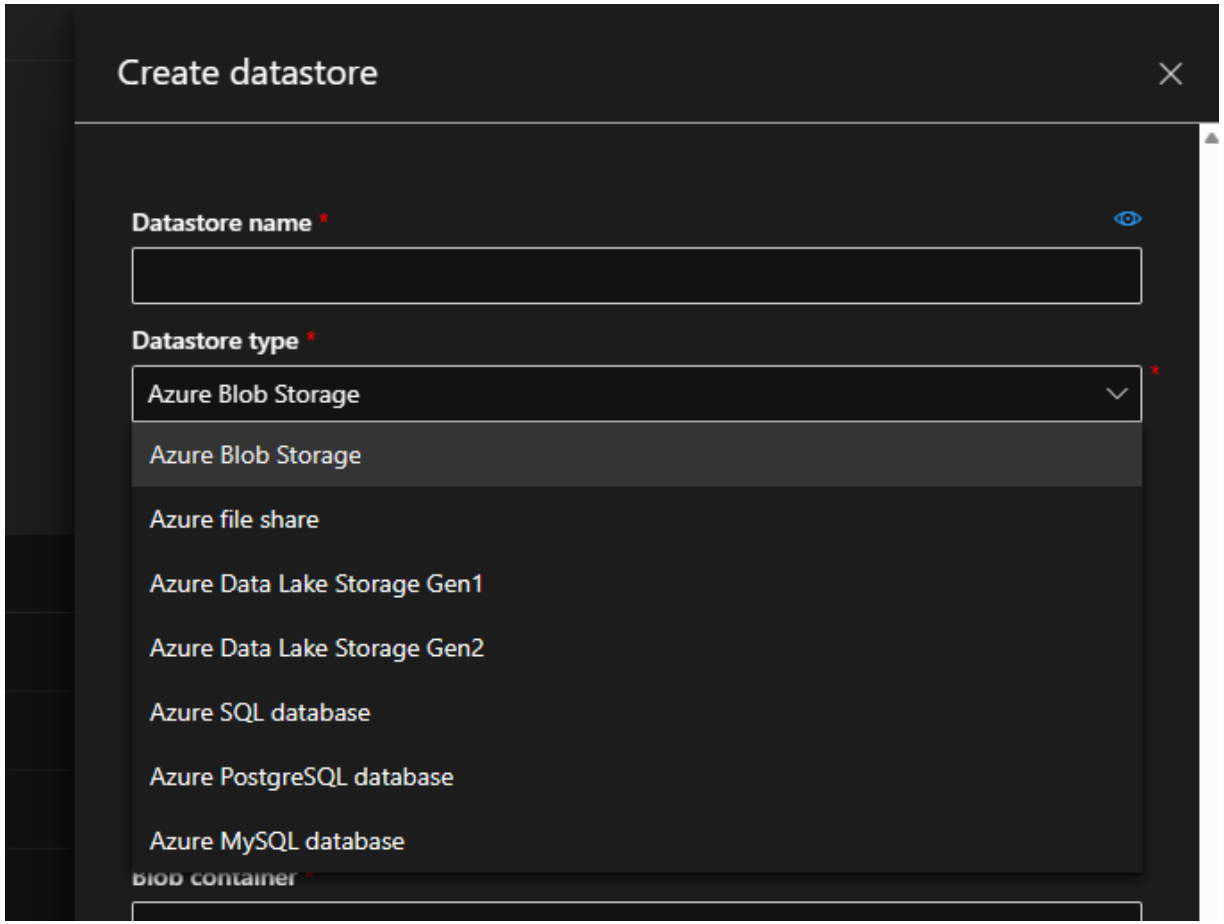
- Create, schedule, orchestrate, and manage data pipelines
- Visualize data lineage
- Connect to on-premises and cloud data sources
- Monitor data pipeline health

- Automate cloud resource management
- Move relational data for Hadoop processing
- Transform with Hive, Pig, or custom code

2. Data Acquisition and Understanding

Link data to Azure Machine Learning Studio

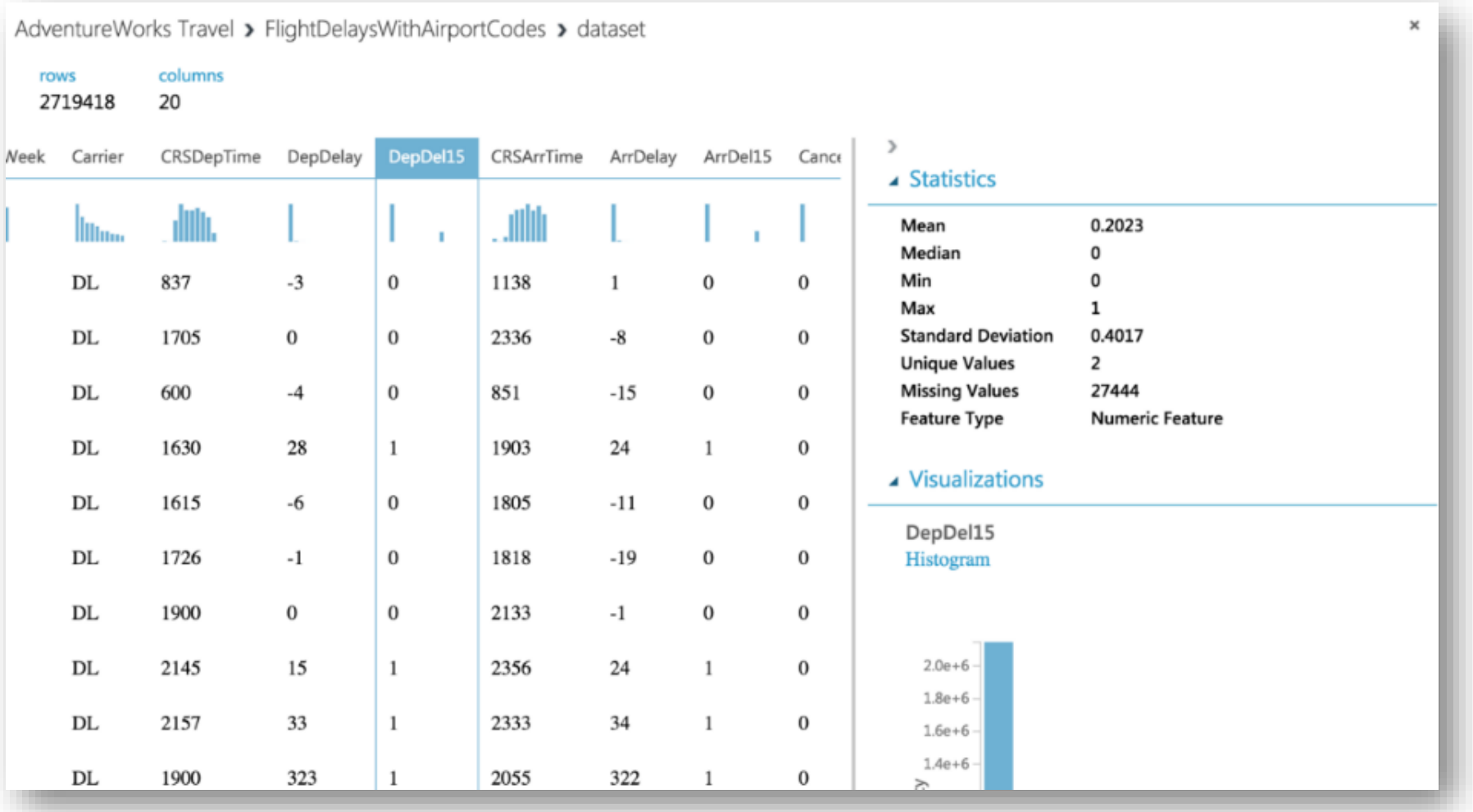
- Datastore



2. Data Acquisition and Understanding

Tools to understand data

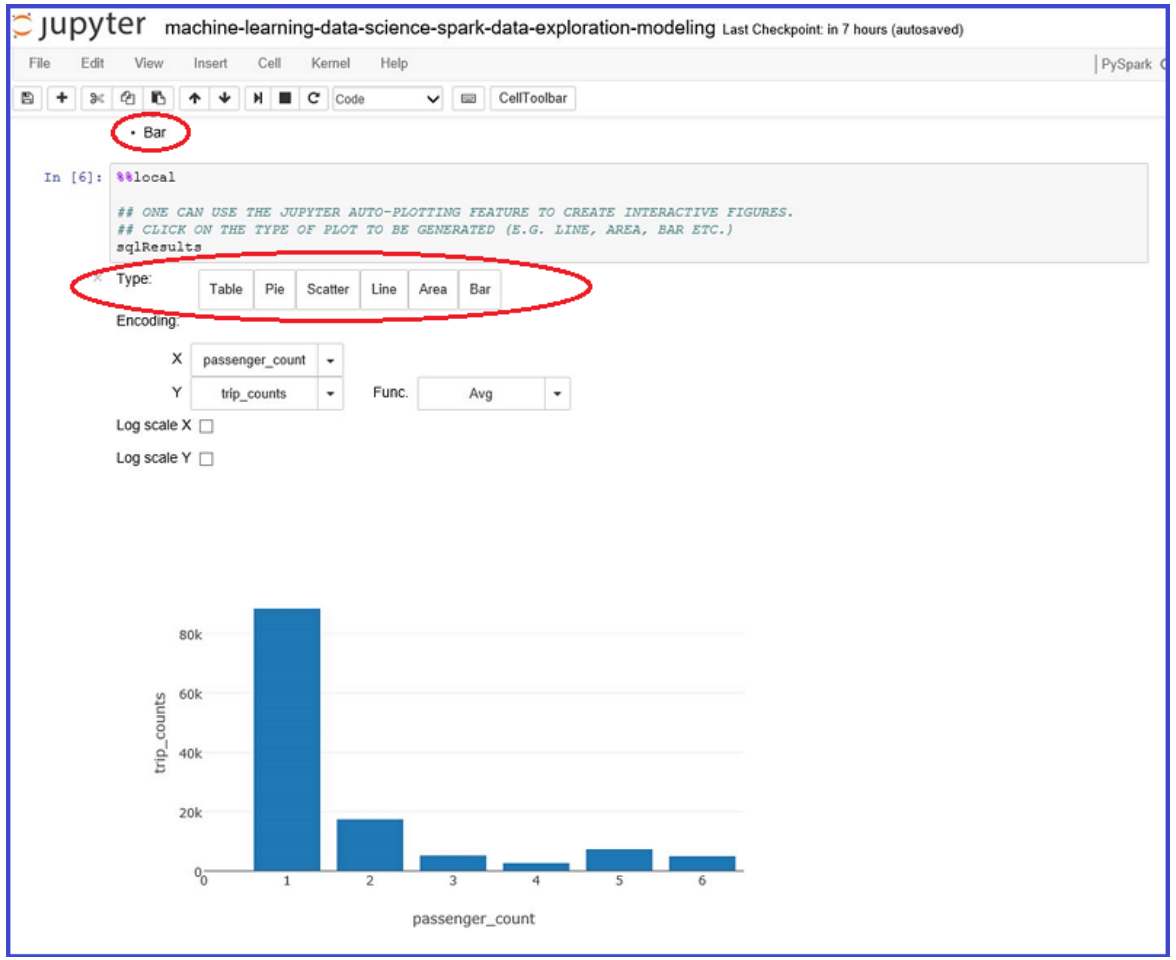
- Data Asset in Azure Machine Learning Studio

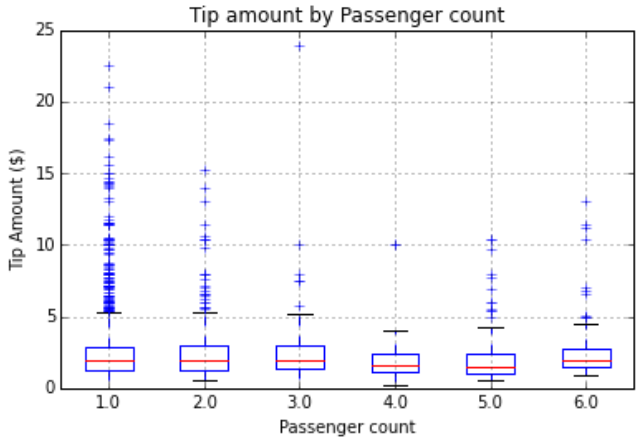


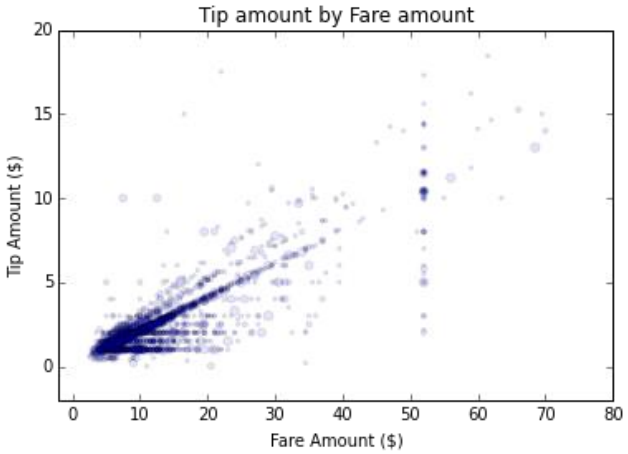
2. Data Acquisition and Understanding

Tools to understand data

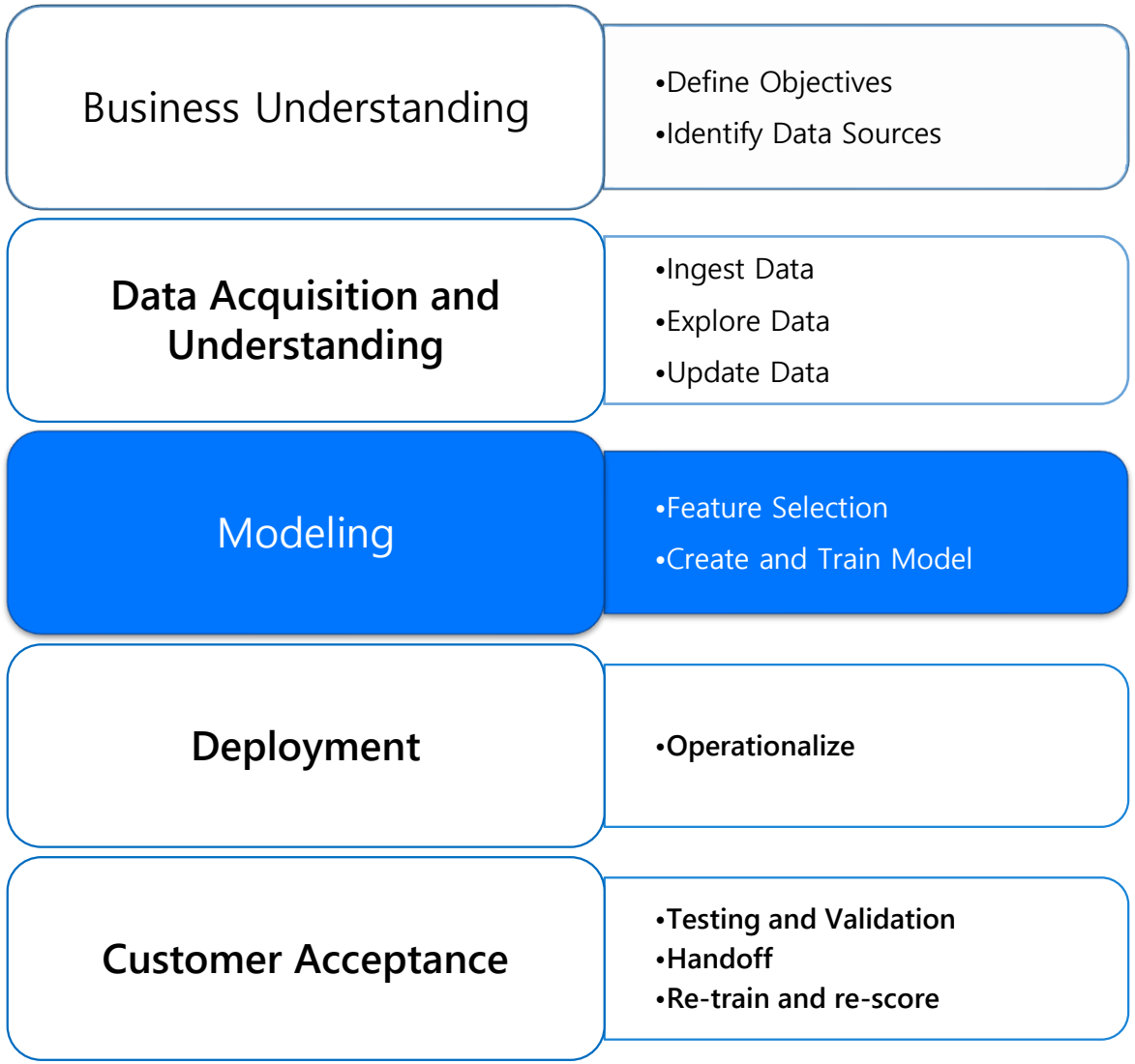
- Use Seaborn, Matplotlib in Python

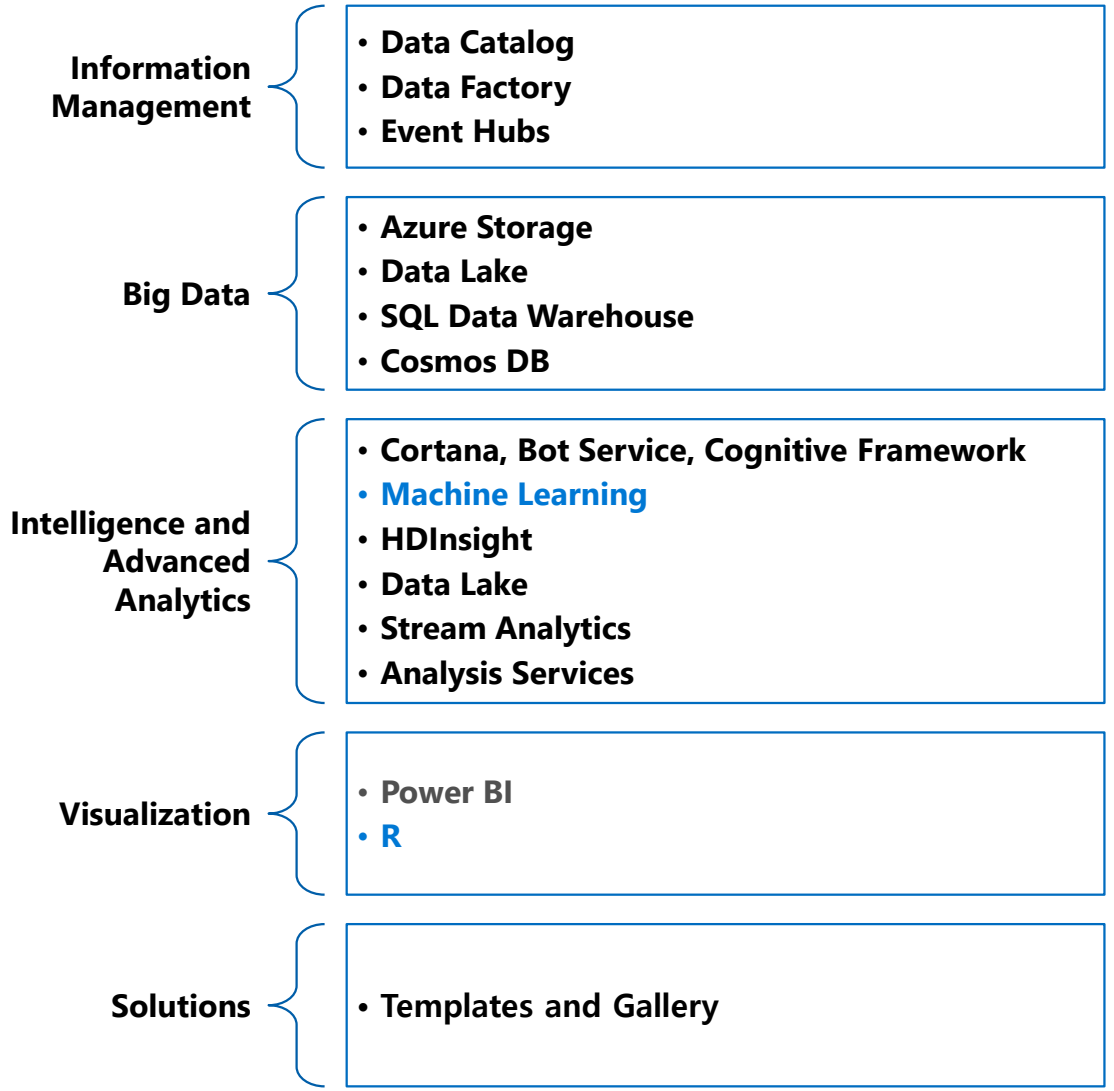






Data Science Tools

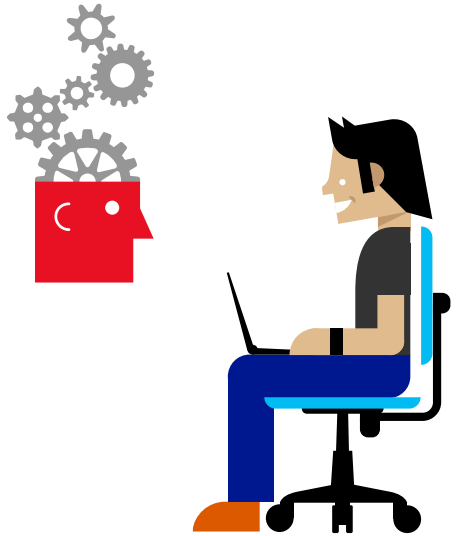




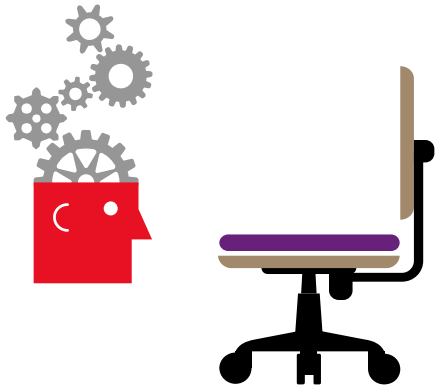
3. Modeling

Machine Learning Algorithms

- Split into two main categories
  - Supervised learning
    - Predicting the future
    - Learn from known past examples to predict future
    - Labels provided

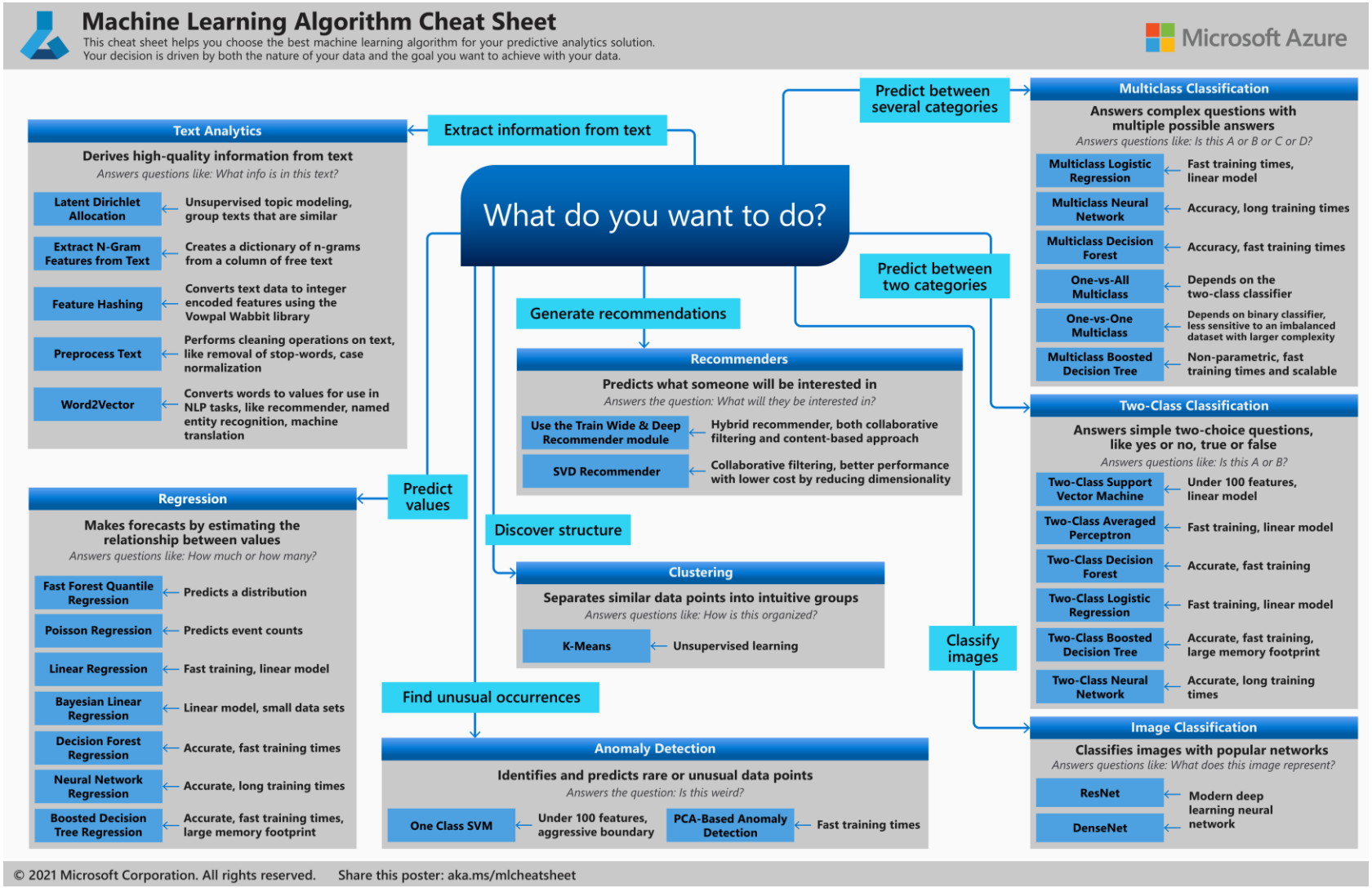


  - Unsupervised learning
    - Making sense of data
    - Understanding the past
    - Learning the structure of data
    - Labels no provided



3. Modeling

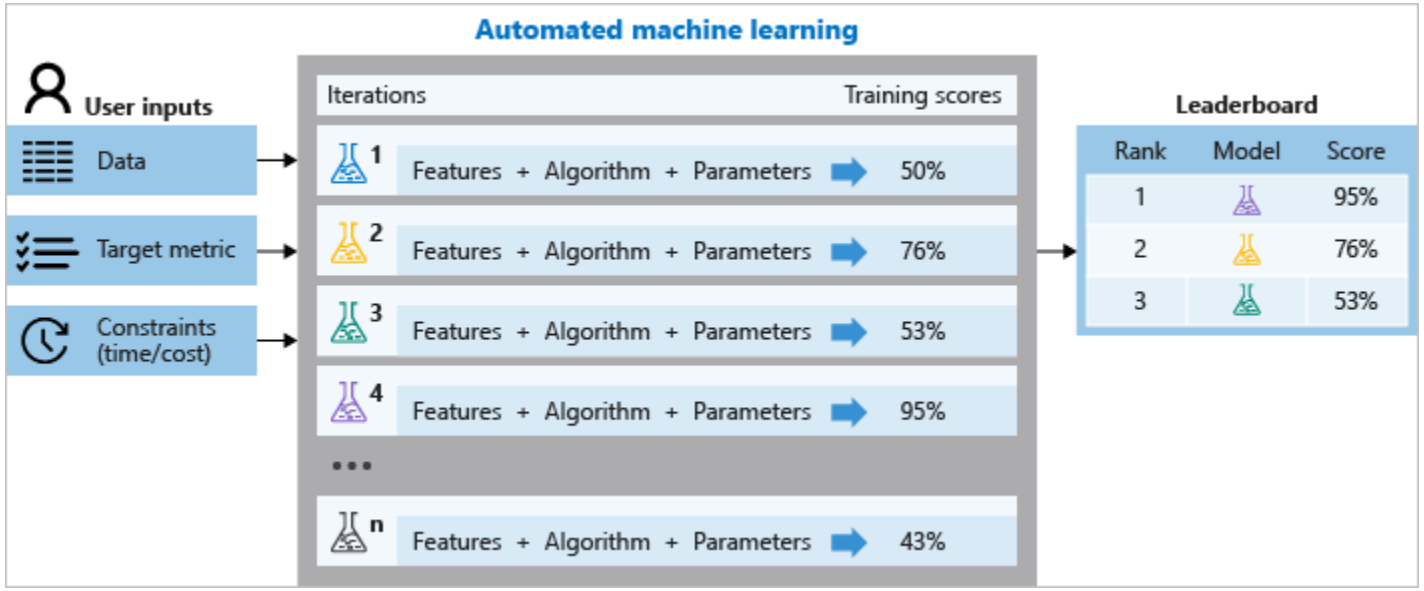
Machine Learning Algorithms



3. Modeling

Model training

- AutoML

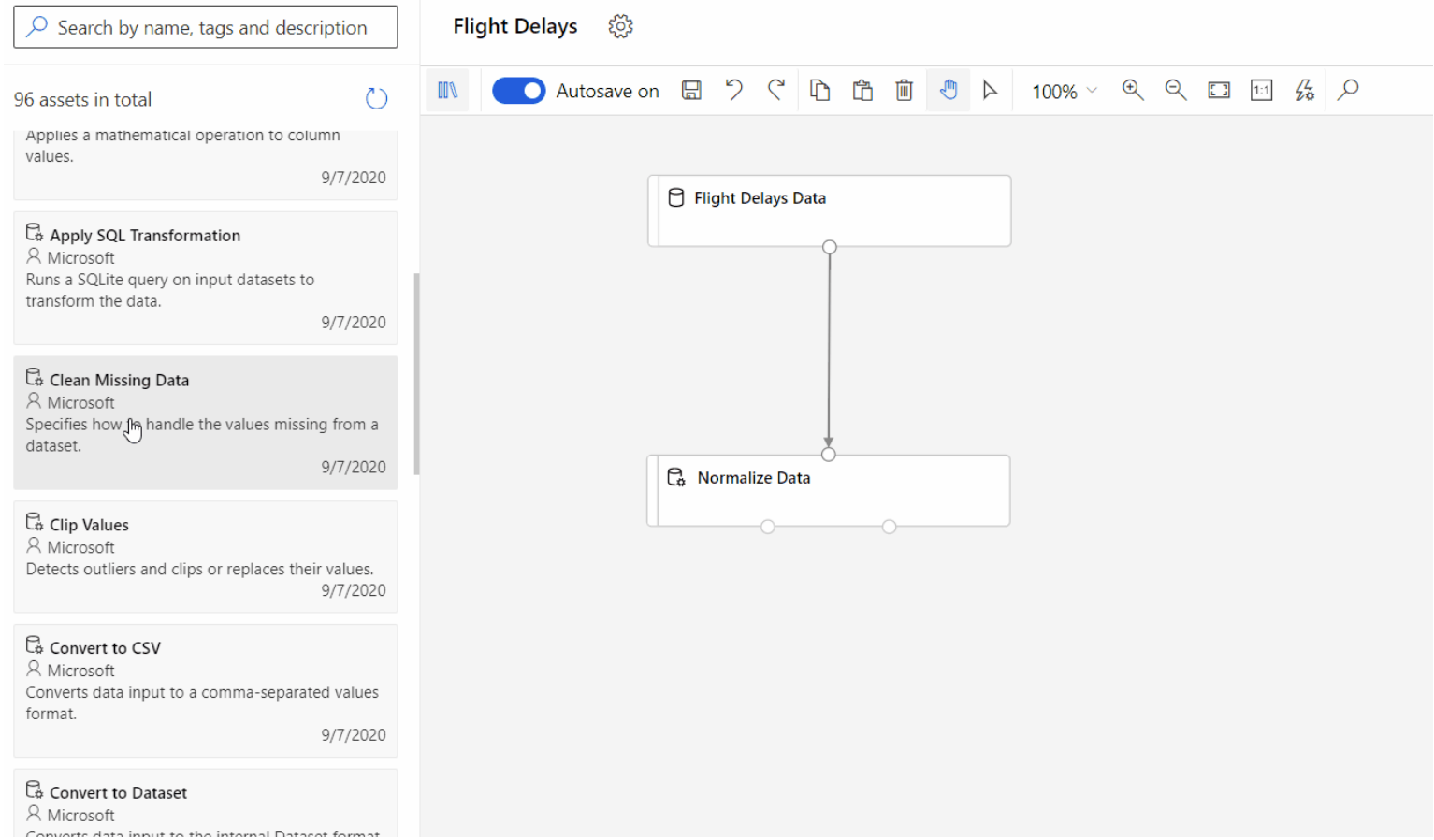


Automated machine learning, also referred to as automated ML or AutoML, is the process of automating the time-consuming, iterative tasks of machine learning model development. It allows data scientists, analysts, and developers to build ML models with high scale, efficiency, and productivity all while sustaining model quality. Automated ML in Azure Machine Learning is based on a breakthrough from our Microsoft Research division.

3. Modeling

Model training

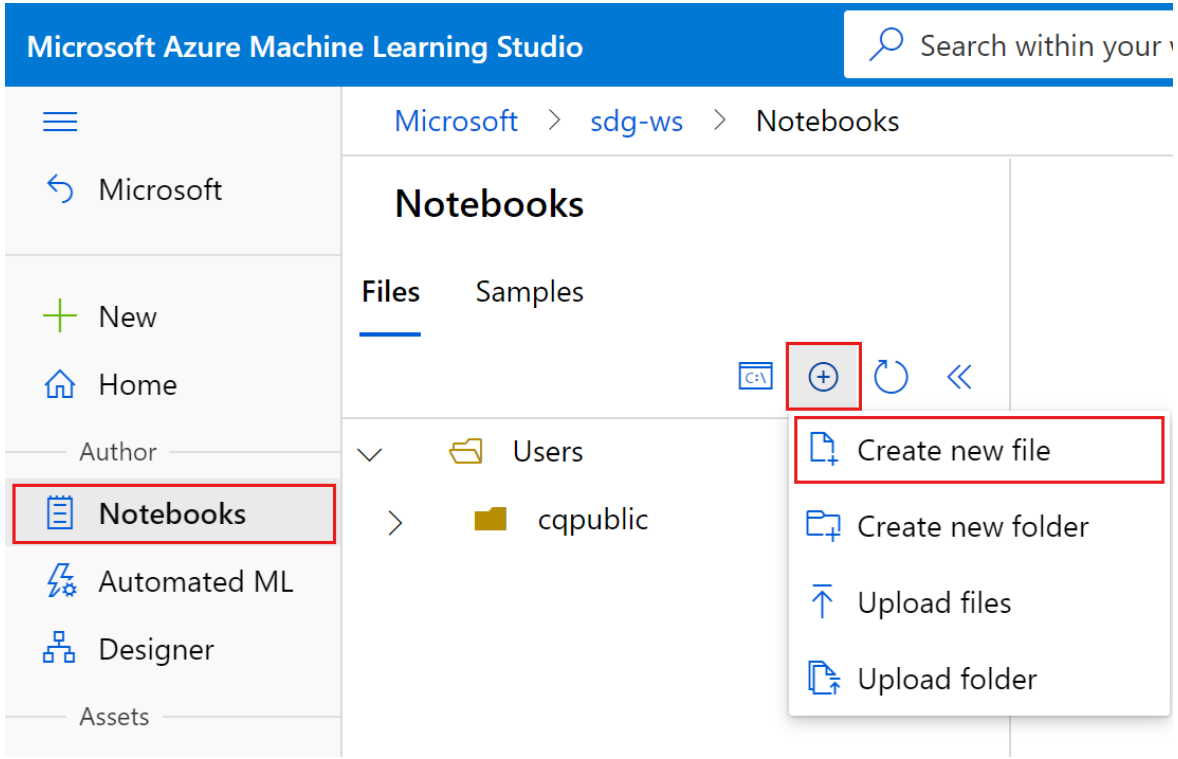
- Designer

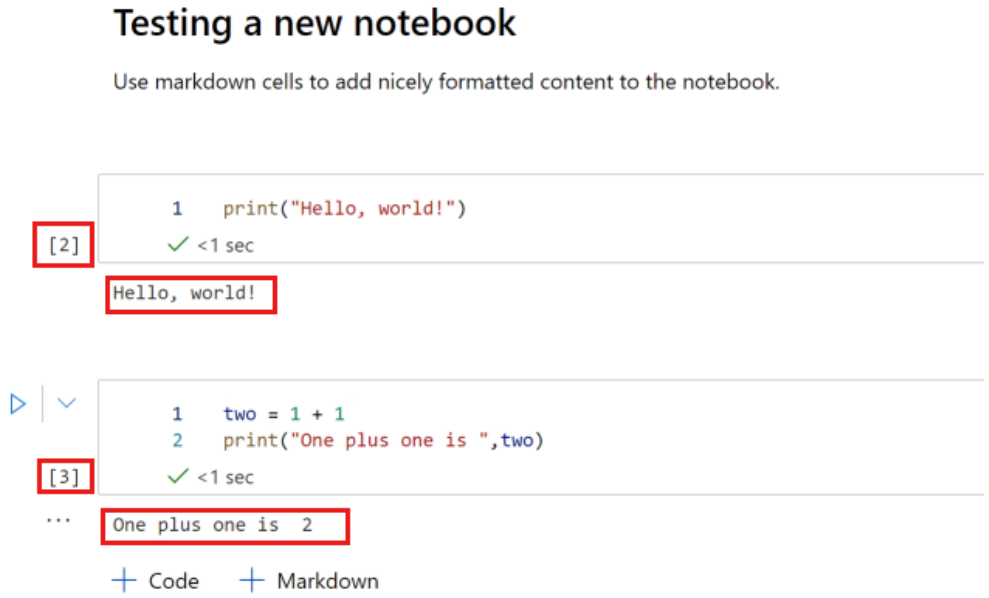


3. Modeling

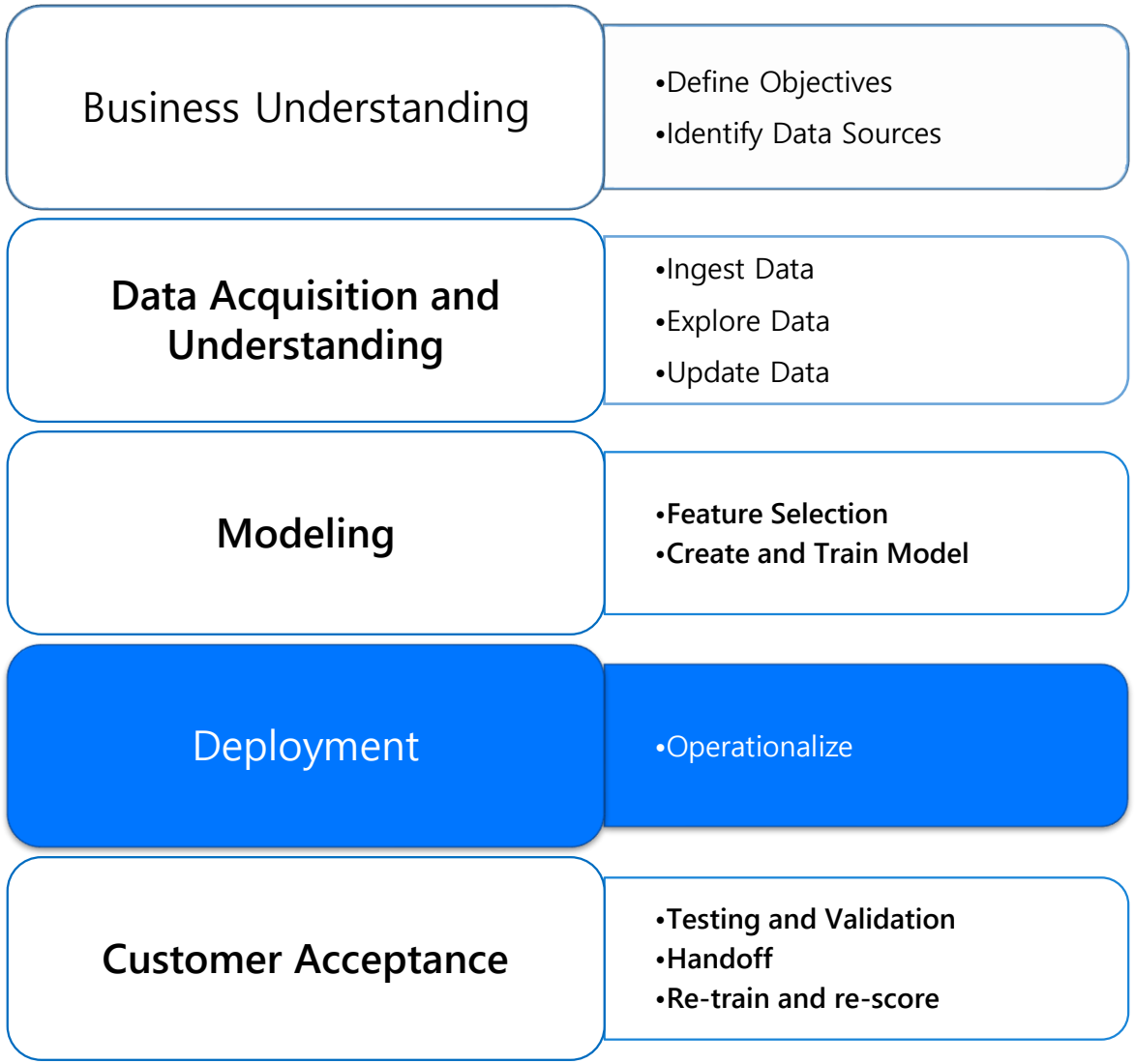
Model training

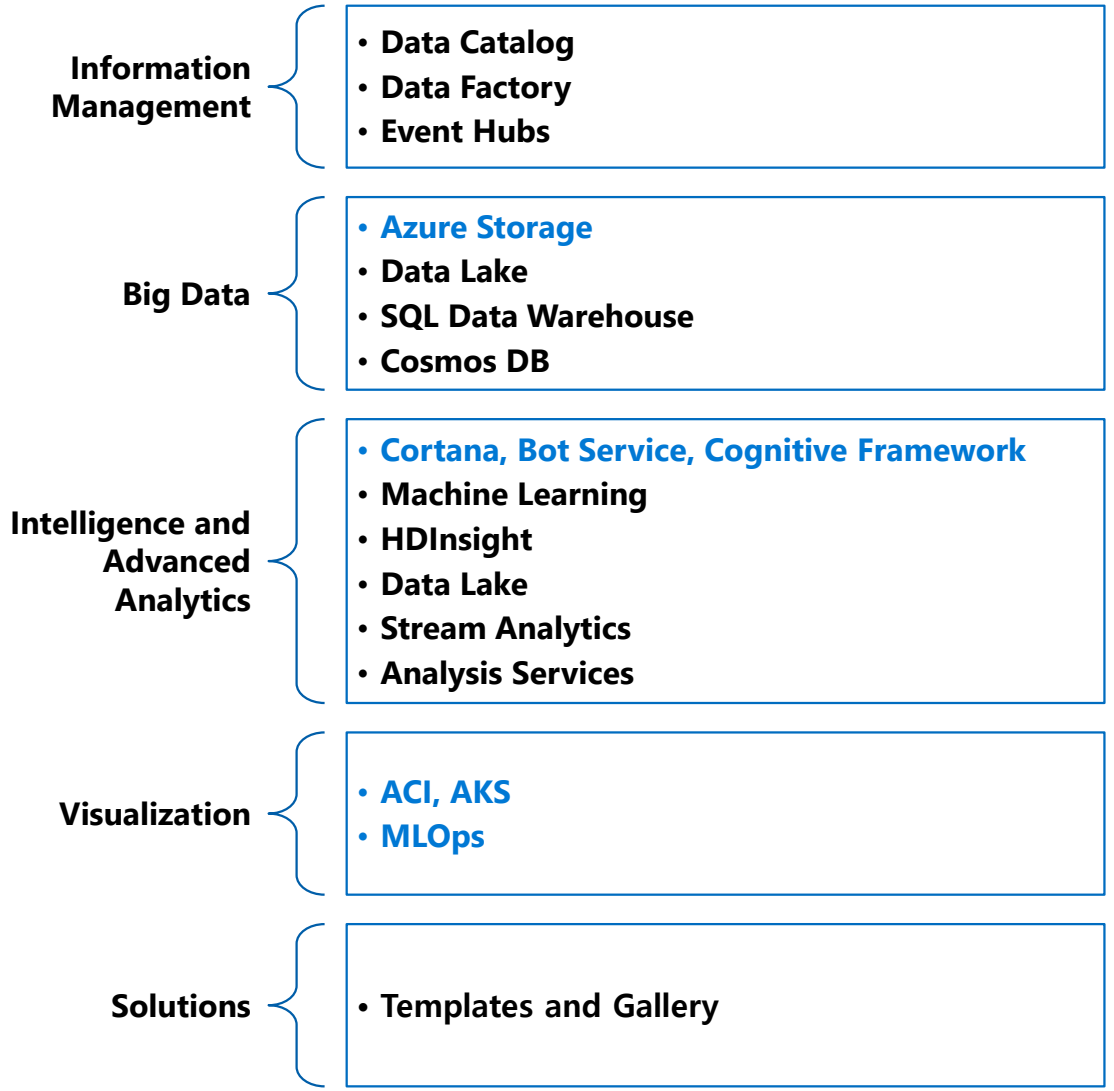
- Notebook (Code First)





Data Science Tools

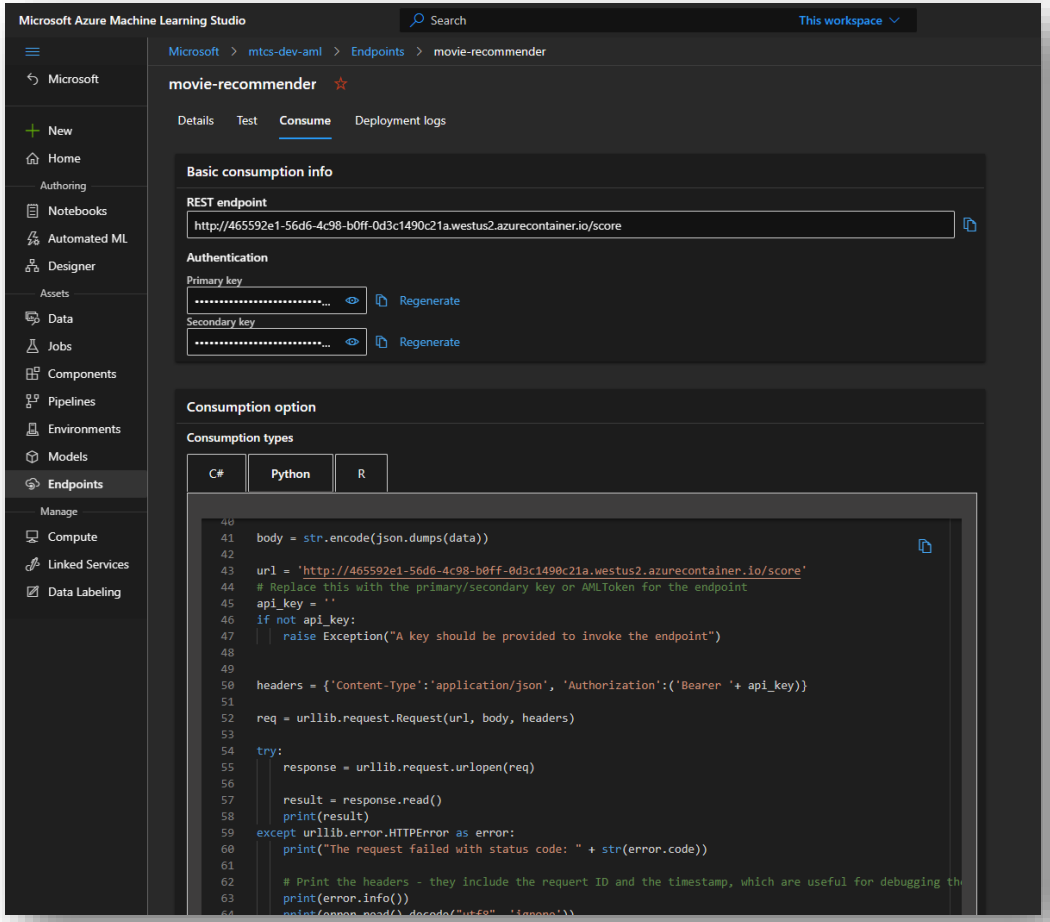


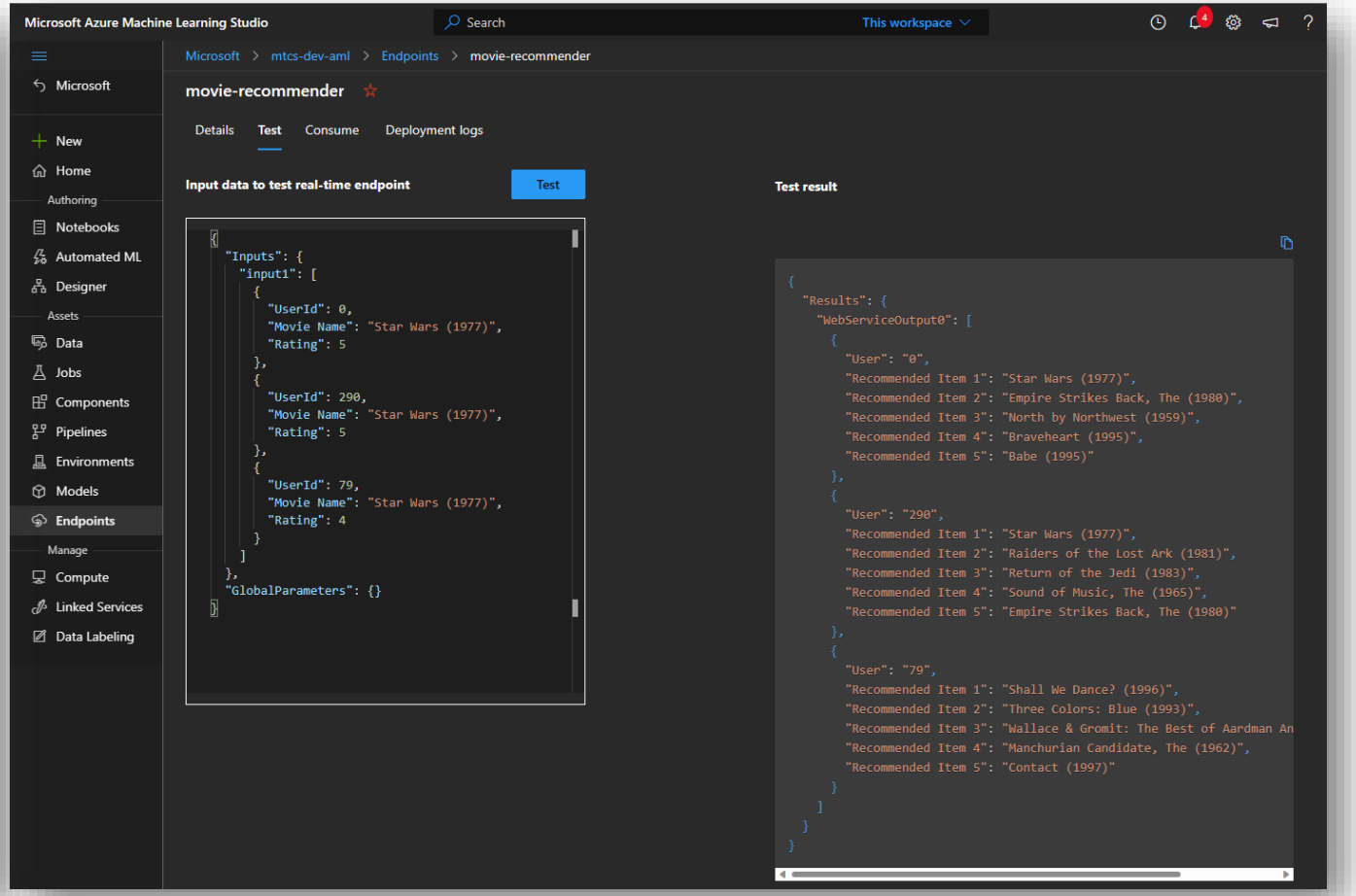


4. Deployment

Operationalize models

- Azure Machine Learning - Publish as a Web Service

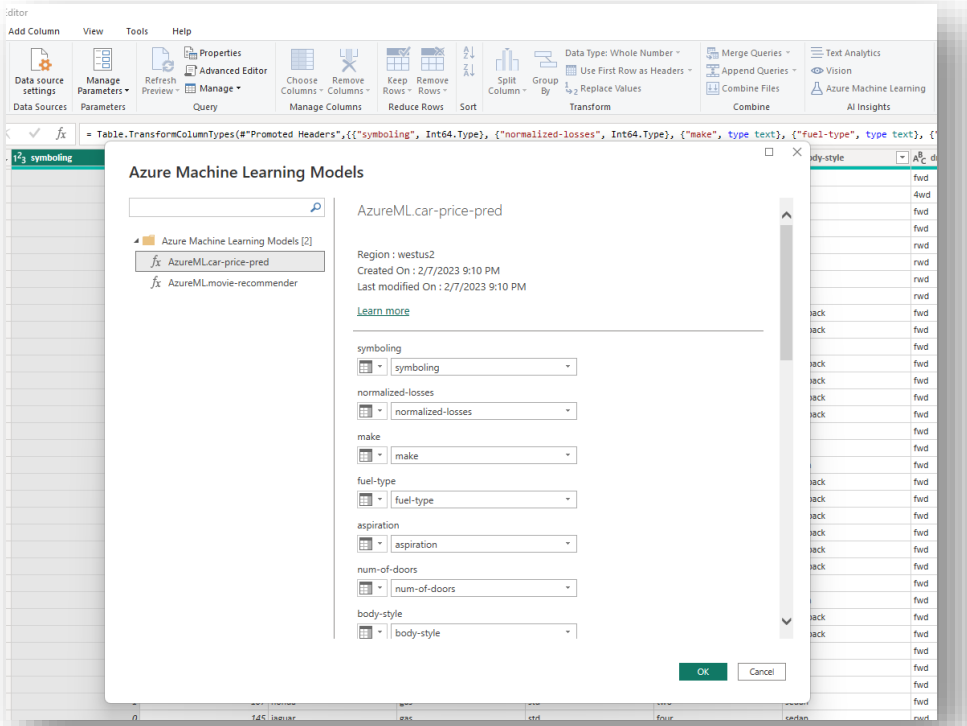


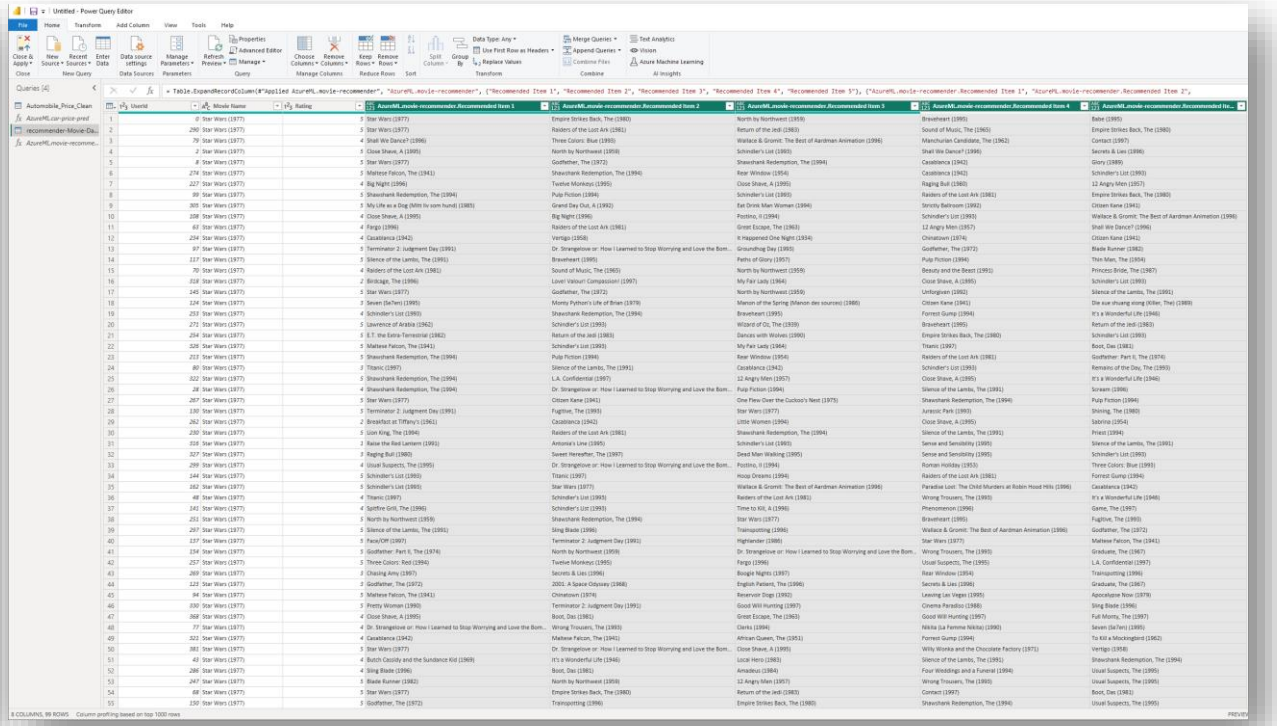


4. Deployment

Operationalize models

- Power BI Integration

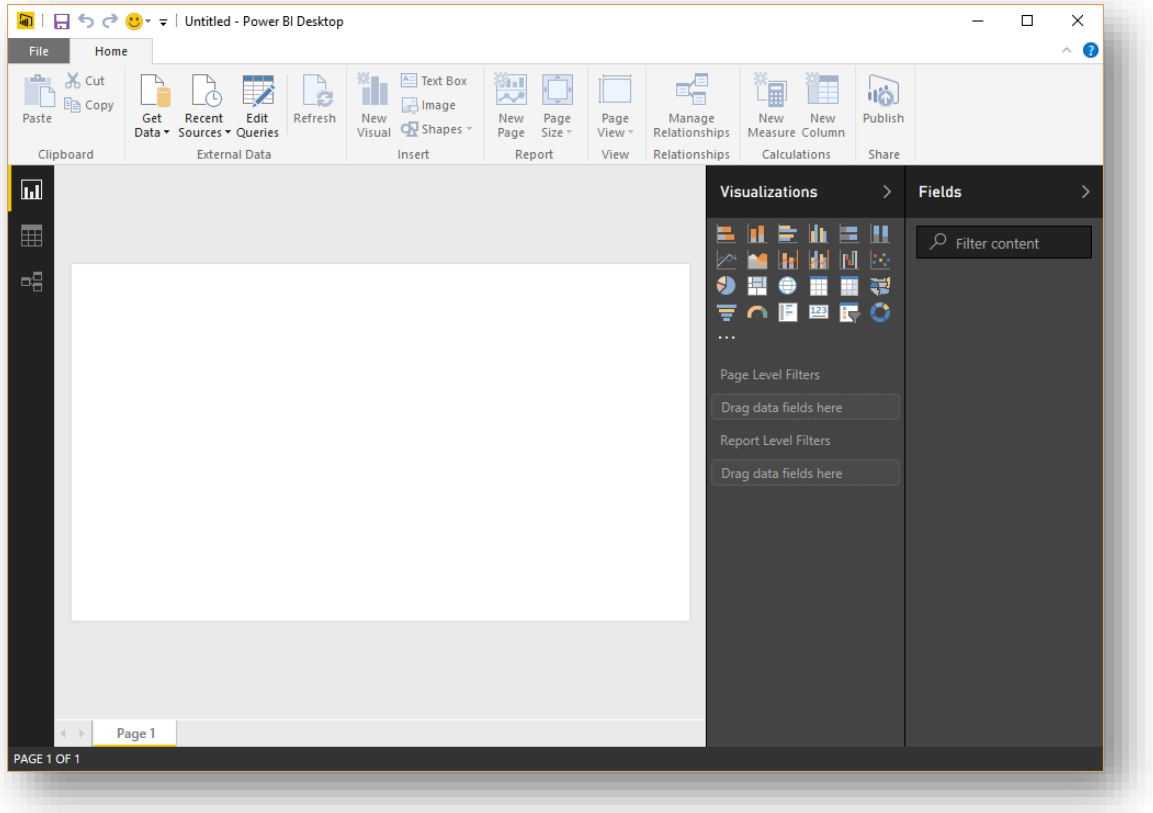




4. Deployment

Report

- Power BI
  - A Reporting System for Multiple Data Sources
  - Available in:
    - Web Portal
    - Power BI Desktop
    - Microsoft Excel
    - Mobile apps (iOS, Android, Windows)
  - Author
    - Connect to Data
    - Shape the Data
    - Model the Data
    - Report on the Data
  - Publish
    - Local
    - To Service

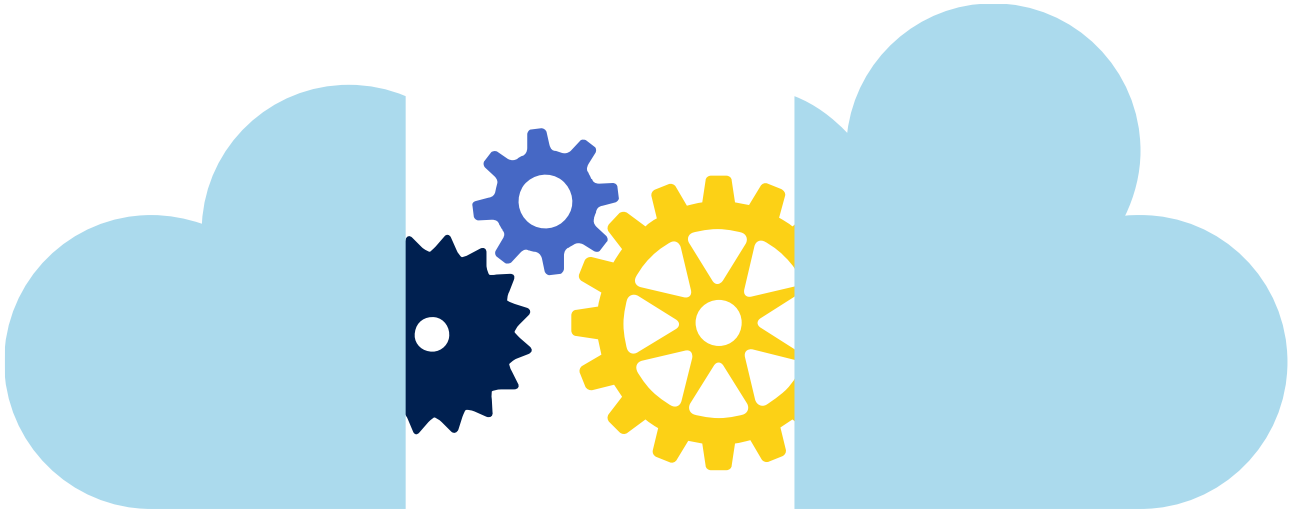


Data Science Tools

Business Understanding	•Define Objectives •Identify Data Sources
Data Acquisition and Understanding	•Ingest Data •Explore Data •Save (Write) Data
Modeling	•Feature Selection •Create and Train Model
Deployment	•Operationalize
Customer Acceptance	•Testing and Validation •Handoff •Re-train and re-score

Agenda

- Data Science
- Data Science process – TDSP (Team Data Science Process)
- Data Science Tools
- **And next**



Microsoft Technology Centers

How to start Data Science Project at work?

HS how to start data science project at work?

Starting a data science project at work can be a complex and challenging task, but it can also be very rewarding. Here are some steps to help you get started:

1. Define the problem: Start by identifying a specific problem or business challenge that you aim to solve with your data science project.
2. Gather data: Collect relevant data that can be used to address the problem. This may involve acquiring data from internal sources, purchasing data from external providers, or scraping data from websites.
3. Clean and prepare the data: Clean and pre-process the data to make it ready for analysis. This may involve dealing with missing values, converting data into a format that can be easily analyzed, and normalizing data to remove any biases.
4. Explore the data: Use data visualization and other exploratory data analysis techniques to gain a deeper understanding of the data and identify any patterns or relationships.
5. Develop a model: Based on the insights gained from the exploratory data analysis, develop a model to make predictions or support decision-making.
6. Validate and evaluate the model: Validate the model by comparing its predictions to actual outcomes, and evaluate its performance to determine if it meets the desired accuracy criteria.
7. Deploy the model: Deploy the model into a production environment, where it can be used to make real-time predictions or support decision-making.
8. Monitor and maintain the model: Continuously monitor the model to ensure it is functioning correctly, and make any necessary updates or improvements over time.

These are just general guidelines, and the specific steps involved in starting a data science project will vary depending on the problem being solved, the data being used, and the methods being employed. It's important to have a solid understanding of statistics, machine learning, and computer programming, and to work with a team of skilled data scientists, engineers, and domain experts.