Azure Fundamentals

Cloud computing

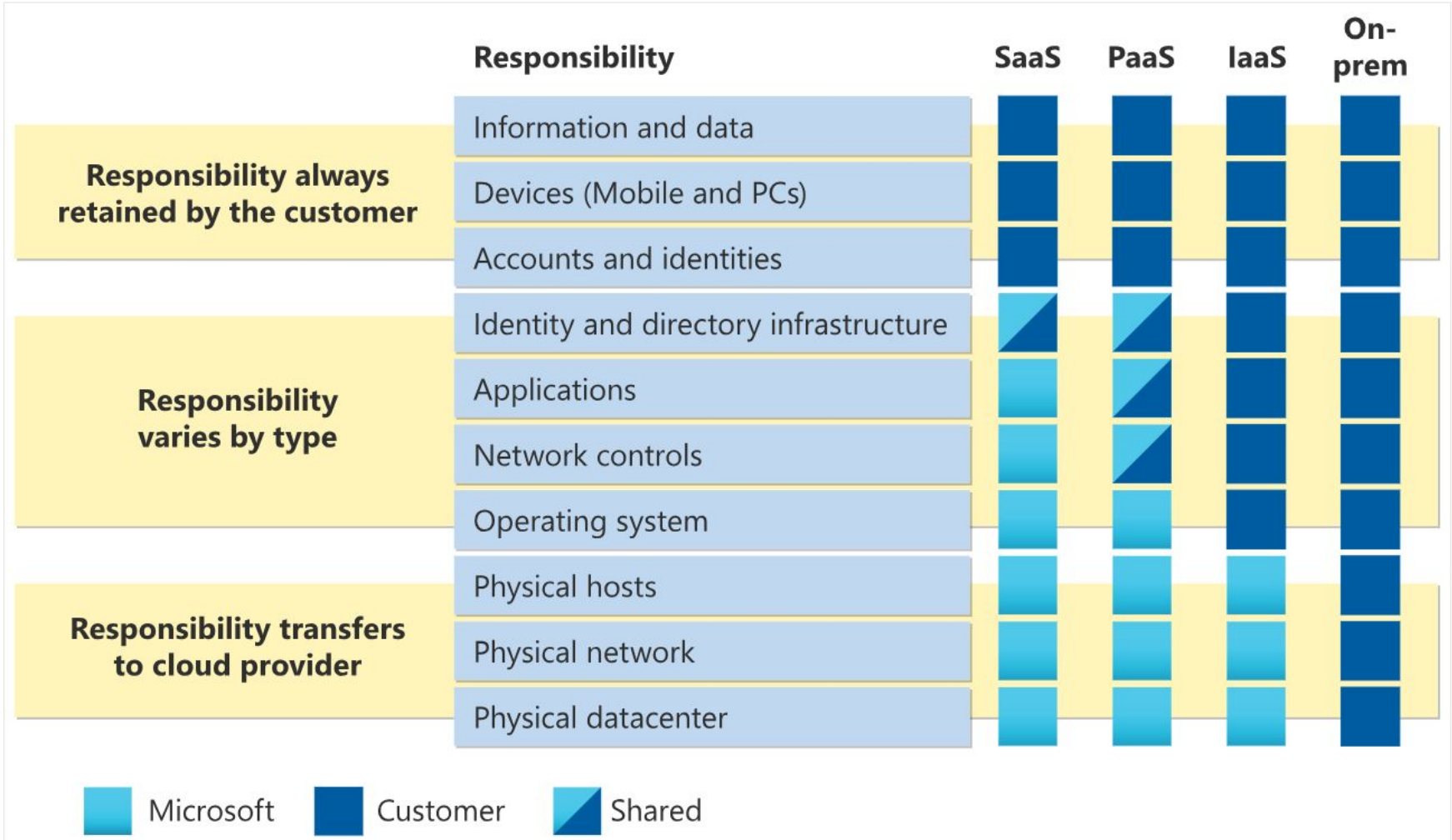
- Delivering computer services over the internet
- You only pay for the resources you use
- Doesn’t have to be constrained by physical infrastructure
- You don't have to wait until a new data center is built to increase your IT infrastructure
- Most important factors
  - Computing power
  - Storage

Shared Responsability model

- Traditional way:
  - The company is responsible for maintaining the physical space, ensuring security and maintaining the servers if anything happens

  - IT: responsible for
    - maintaining all infrastructure and SW needed to keep the Data center running
    - Keep all software patched

With the shared responsibility model these responsibilities are shared between the cloud provider and the consumer on a few different ways depending on the type of service.


Responsibility
SaaS
PaaS
IaaS
On-prem
Responsibility always retained by the customer
Information and data
Devices (Mobile and PCs)
Accounts and identities
Responsibility varies by type
Identity and directory infrastructure
Applications
Network controls
Operating system
Responsibility transfers to cloud provider
Physical hosts
Physical network
Physical datacenter
Microsoft
Customer
Shared

CLOUD MODELS

- Private Cloud
  - Natural evolution from a corporate datacenter
  - Cloud used by a single entity.
  - Privacy: Data is not collocated with other organizations data
  - Much grater control for the company’s IT
  - Costly
  - Missing some benefits the public cloud has
  - Can be hosted on-site or remotely by a 3rd party

CLOUD MODELS

- Public Cloud
  - Built, controlled, and maintained by a third-party cloud provider
  - General public availability: Anyone can access and buy resource
  - No capital expenditure to scale up
  - Applications are quickly provisioned and deprovisioned.
  - Pay only for what they use
  - Organizations don't have complete control over resources and security
  - Organization responsible for HW maintenance and updates.

Cloud Platform

- Ever-Expanding set of services to help you build solutions
- Web Services
- Fully virtualized computers to turn your custom software solutions
- Cloud-based services
  - Remote Storage
  - Database Hosting
  - centralized account management
- AI
- IoT

CLOUD MODELS

- Hybrid Cloud
  - Organizations determine where to run their applications:Users can choose which services to keep in public and which to deploy to their private cloud infrastructure.
  - Organizations in charge of security, compliance or legal requirements.
  - Provides the most flexibility
  -

CLOUD MODELS

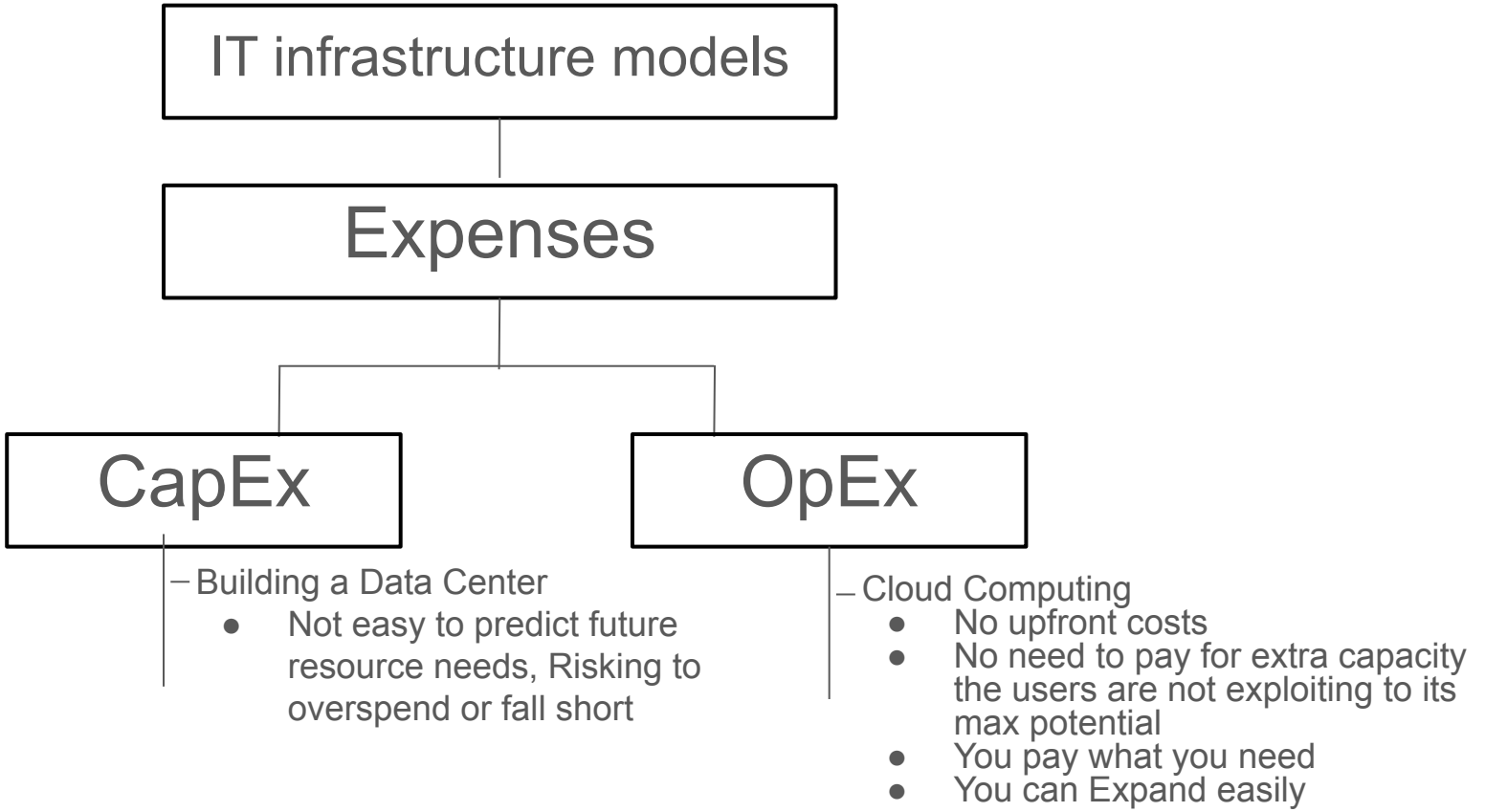
- Multi-Cloud
  - Deal with two or more public cloud providers
    - Maybe migrating from one to another
    - Or using features from one and the other

CLOUD MODELS

- Azure Arc
  - Set of tecnologies that help you manage your cloud environment
  - No matter if
    - Public Cloud solely on Azure
    - Private cloud in your datacenter
    - Hybrid..or
    - Even Multi-Cloud environment running on multiple cloud providers at once.

CLOUD MODELS

- Azure VMware Solution
  - For those who are already established with VMware in a private cloud environment but want to migrate to a public or hybrid cloud
  - Azure VMware solution lets you run your VMware workloads with seamless integration



The cloud shifts IT spend from a capital expense to an operational expense.

SLA

Uptime

- 100% Uptime is difficult and expensive to achieve
  - Allow no time for taking the service down for required maintenance/updates
  - Requires douplicating all
- More common: 99%, 99.9%,99.95% Uptime. Even 99.99

Downtime

- What is defined as down time?
- Having backup components kickoff immediately if something fails generating zero interruptions to the customer
-

Each Azure service has its own SLA

Scalability

Vertical Scaling

- Need more computing to develop an App?
  - Vertically Scale up to add RAM or CPU
- The opposite as well (Scale down)

Horizontal Scaling

- If you experience a steep jump on demand, you can scale out
  - I.e. Adding additional Virtual Machines
- The opposite as well (Scale in)

- Reliability
  - Ability of a system to recover
  - Cloud = Decentralized design
  - Resources deployed in regions around the world
    - In case of a catastrophic event, automatically switch to another region.

- Predictability

lets you move forward with confidence

  - Performance
    - Focuses on predicting the resources needed to deliver a positive experience for your customers
    - Concepts
      - Autoscaling
      - Load Balancing
      - High Availability
  - Cost

    Focused on predicting or forecasting the cost of the cloud spend. Monitor resources to ensure you're using them in the most efficient way.Use Data Analytics to find patterns and trends that help better plan resource deployments

    - Tools
      - Total cost of ownership
      - Pricing Calculator

Manageability

- Management of the Cloud
  Managing your cloud resources
  - Automatically scale resource deployment based on need
  - Deploy resources based on preconfigured template, no need for manual config.
  - Monitor the health or resources and automatically replace failing resources.
  - Receive automatic alerts based on configured metrics so you're aware of performance in real time.

- Management in the cloud
  How you manage your cloud environment and resources
  - Through a Web portal.
  - Using a command line Interface.
  - Using APIs.
  - Using powerShell.

IaaS

Most flexible category of cloud services
(Essentially renting the HW in a cloud datacenter)

- Cloud provider responsible
  - Maintaining the HW
  - Network Internet connectivity
  - Physical security

- User responsible of everything else
  - Operating System Installation,configuration and maint
  - Network config
  - Database and Storage config
  - etc

- Scenarios
  - Lift-and-shift migration: Replicating your on-premises datacenter to migrate it to the cloud.
  - Testing and development: When in need to create, replicate and shut down development environments and test environments rapidly.

PaaS

Middle Ground
Between IaaS and SaaS

- Cloud provider responsible of:
  - Maintaining the HW
  - Network Internet connectivity
  - Physical security
  - Operating Systems
  - Middleware
  - Development tools
  - Business intelligent services

- User optionaly responsible of:
  - Network config
  - Directory Infrastructure
  - Aplications

- Scenarios
  - Development Framework. Similar to the way you create an Excel macro. Developers create applications using built-in software components reducing the amount of coding needed.
  - Analytics (Business intelligence): Allow organizations to analyze and mine their data for business decisions.

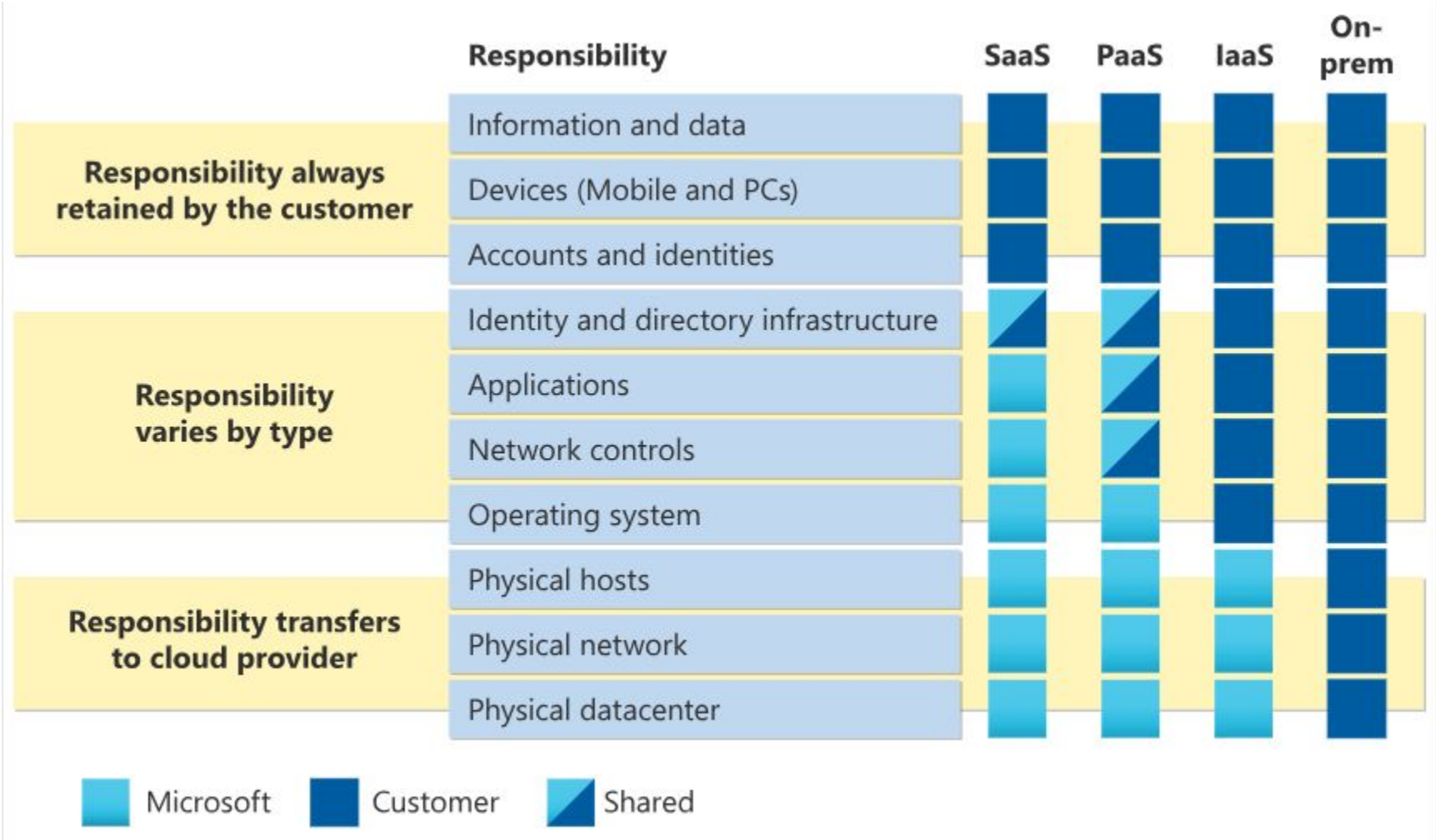
SaaS

Least Flexible
Easiest to get up and running
Requires the least amount of technical knowledge

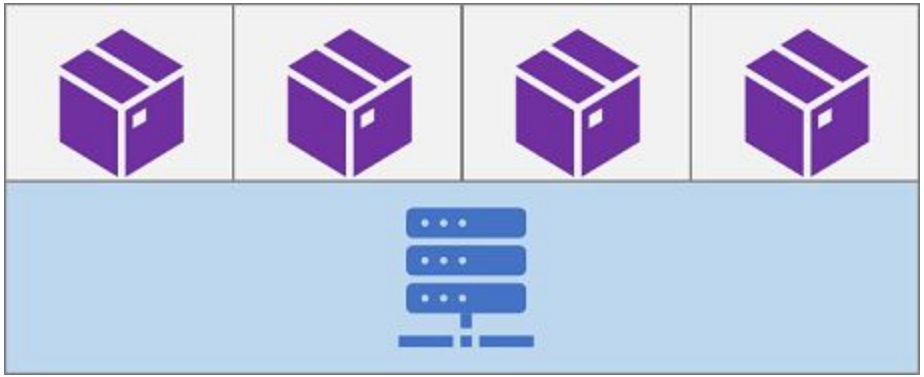
- The platform is responsible of:
  - Everything else

- User responsible of:
  - Data put into the system
  - Devices allowed to connect to the system
  - Users that have acces

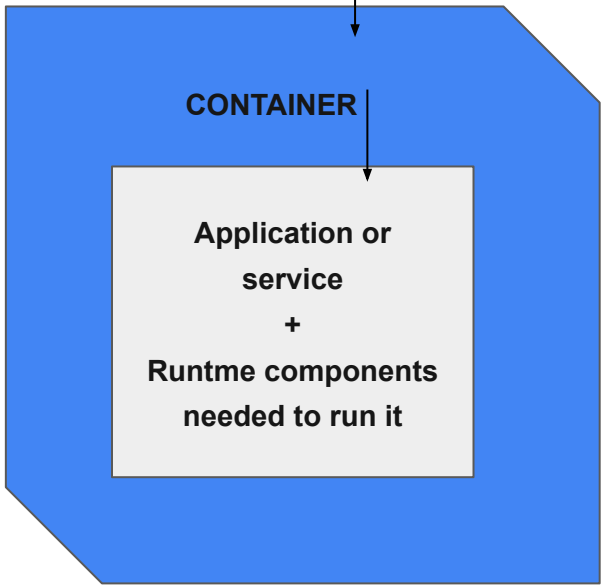
- Scenarios
  - Email and messaging.
  - Business productivity applications.
  - Finance and expense tracking.


Responsibility
SaaS
PaaS
IaaS
On-prem
Responsibility always retained by the customer
Information and data
Devices (Mobile and PCs)
Accounts and identities
Responsibility varies by type
Identity and directory infrastructure
Applications
Network controls
Operating system
Responsibility transfers to cloud provider
Physical hosts
Physical network
Physical datacenter
Microsoft
Customer
Shared

Deploy in Containers



Underlying operating System and Hardware



Examples of Container Host

- ★ Docker Server
- ★ Azure Container Instance (ACI)
- ★ Azure Kubernetes Service (AKS)

Benefits

- ★ Portables across hosts
- ★ Single container host can support multiple isolated containers
  - ○ Easier to consolidate multiple applications with different configuration requirements

Secure Azure AI Services

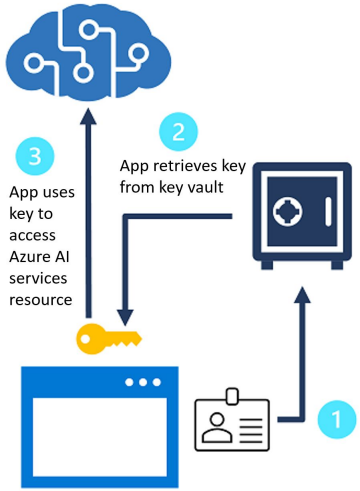
Authentication

Regenerate Keys

If you're using both keys in production:

1. Change all apps to use only 1
2. Regenerate the other
3. Switch everything the newly generated key
4. Regenerate the other
5. Switch back

Protect Keys with Azure Key Vault



Token-based authentification

- ★ Some AI services support (or require) token-based authentication .
- ★ Usual valid period: 10 mins.
- ★ Subsequent requests must present the token to validate the caller has been authenticated.

Microsoft Entra ID authentication

using..

Service Principals

- Create a custom domain
  - Through:
    - Azure portal
    - Azure CLI, or
    - PowerShell.
- Assign a role to a service principal.

Managed identities

System-assigned

- Created and assigned to a specific resource.(i.e., Virtual Machine). When the resource is deleted so is the identity.

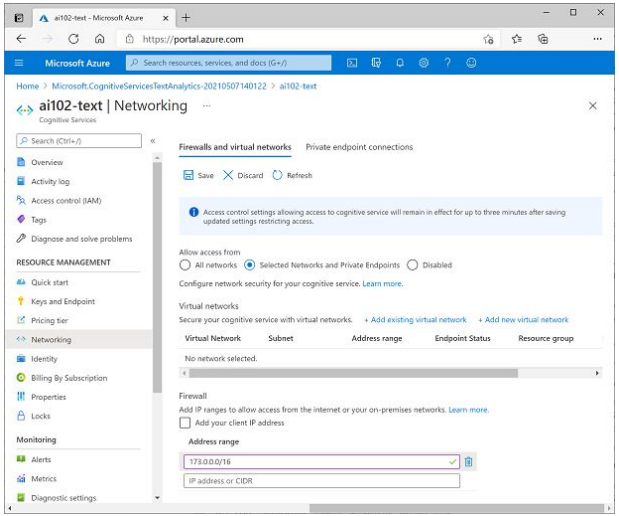
User-assigned

- Usable by multiple resources.

Network Security

Ensure unauthorized users can't reach the services that you're protecting. *Users can't compromise what they can't see.*

- **By default, Azure AI services are accessible from all networks.**
- **Some services resources can be conf to restrict access to specific networks.**
-



With network restrictions enabled, a client trying to connect from an IP address that isn't allowed will receive an *Access Denied* error.

Prepare to develop AI solutions on Azure

https://github.com/MicrosoftLearning/mslearn-ai-language

AI

Software that exhibits one or more human-like capabilities like:

Visual Perception

Use computer vision to accept, interpret, and process input from images, video streams and live cameras
i.e: Mobileye, security systems, face recognition, distracted driving detection

Text analysis and conversation

Use natural language processing (NLP) to read and generate realistic responses.
Extract semantic meaning from text

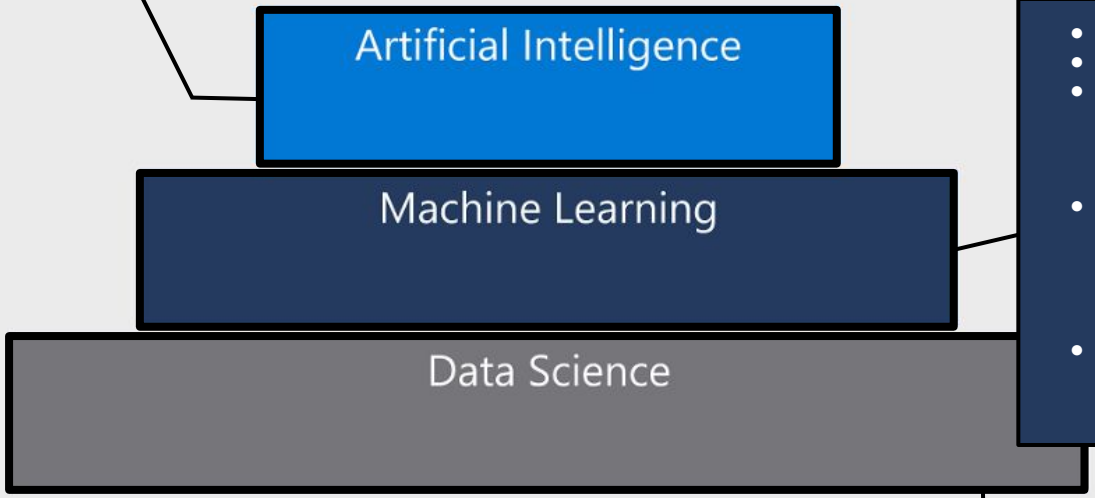
Speech

Recognize speech as input and synthesize spoken output. Speech capabilities + NLP = Conversational AI (human-compute interaction)
Users interact with AI agents (bots)

Decision making

Use past experience and learned correlations to assess situations and take appropriate actions
i.e.,Recognize anomalies in sensor readings and taking automated actions to prevent failure or system damage

- Usually (not always) Builds on (enpowers) machine learning to create software that emulates one or more characteristics of human intelligence
- I.e., In the endangered species example below: It may not be feasible to rely on human experts who can positively identify the animal in question, or to monitor a large area over a sufficient period of time to get an accurate count.
    - A predictive model can be trained to analyze image taken from motion-activated cameras, and predict whether a photograph contains a sighting of the animal and be used to identifying areas with dense animal populations that may be candidates for protected status



- Subset of Data Science
- Training and validation of predictive models
- Data scientist prep the data and uses it to train a model based on an algorithm that exploits the relationships between the features in the data to predict values for unknown labels

- A data scientist uses collected data to train a model that predicts annual growth or decline in population of endangered species based on factors such as the number of nesting sites observed, area of land designated as protected, human population in the area , daily volume of traffic,etc.
- This model can then be used as a tool to evaluate plans for housing, infrastructure and industrial development in the local area and assess their likely impact on the local wildlife.

- Discipline that focuses on the processing and analysis of data
- Applying statistical techniques to uncover and visualize relationships and patterns
- Define experimental models to help explore those patterns
- I.e., A data scientist may gather samples of data about the population of an endangered species in a geographical area and combine it with data about levels of industrialization and economic demographic in the same area

Model training

- Many AI models rely on predictive models that must be trained using simple data.
- The training process analyses the data and determines relationships between the features in the data and the label (the value that the model is being trained to predict).
- After the model has been trained you can submit new data that include know feature values and have the model predict the most likely label. Using this model to make predictions is referred to as inferencing.

Probability & Confidence scores

- No predictive model is infallible
- Predictions made by machine learning models are based on probability
- Predictions reflect statistical likelihood, not absolute truth
- In most cases predictions have an associated confidence score.

Responsible AI and ethics

- It's important for SW engineers to consider the impact of their SW on users, and society in general (ethical, etc)
- The decisions AI informs are based on probabilistic models, which are dependent on the quality of the data with which they were trained.
- The Human-Like nature of AI is a great benefit in making applications user-friendly but it also lead user to tend to trust in the application's ability to make correct decisions.
- Mayor Concern: The potential of harm (i.e., unfairness, mislead. Discrimination, etc) to individuals or groups through incorrect predictions or misuse of AI capabilities.
- AI-enabled solutions should apply due consideration to mitigate this risk.

Responsible AI

FAIRNESS

All systems should treat all people fairly

- Fairness of MLS is a highly active area of research
- Training data should be carefully reviewed to ensure it is potentially inclusive of all potentially affected subjects
- i.e.,Machine learning model that supports a loan approval application for a bank, should make predictions of whether or not the loan should be approved without any bias.

RELIABILITY AND SAFETY

AI systems should perform reliably and safely.

- AI-based software application dev must be subjected to rigorous testing to ensure they work as expected before release.
- SW engineers need to take into account the probabilistic nature of machine learning models, and apply appropriate thresholds when evaluating scores for predictions.
- I.e., AI-based SW for AV, MLM that diagnoses patient symptoms and recommends prescriptions.

PRIVACY AND SECURITY

AI systems should be secure and respect privacy.

- The MLM on which AI systems are based rely on large volumes of data that may contain personal details that must be kept private.
- Appropriate safeguards to protect data and customer content must be implemented.

INCLUSIVENESS

AI should bring benefits to all parts of society regardless of physical ability, gender, sexual orientation, ethnicity or other factors.

- One way to optimize it for inclusiveness is to ensure that the design, development , and testing includes input from as diverse group of people as possible.

TRANSPARENCY

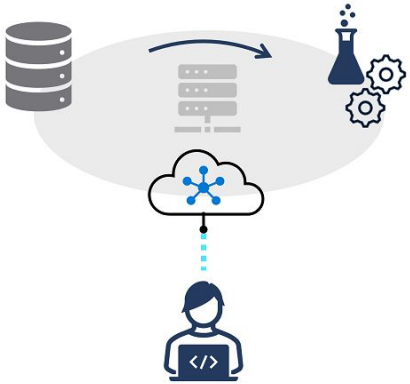
AI systems should be understandable

- Users made fully aware of the purpose of the system
- How it works
- Limitations to be expected
- I.e., make users aware of factors that may affect accuracy like: number of cases used to train the model, features that have th emost influence over its predictions.
- Confidence score
- In the case of Facial recognition: How the data is used and retained?. Who has access to it?

ACCOUNTABILITY

People should be accountable for AI systems

- Although they seem to operate autonomously, ultimately, it is the responsibility of the developers who trained and validated the models they use and defined the logic that bases decisions on model predictions to ensure that the overall system meets responsibility requirements.
- Work within a framework of governance and organizationsl principles to ensure the solution meets ethical and legal standards that are clearly defined.



Azure Machine Learning Service

Cloud based platform for running experiments at scale to train predictive models from data, and publish the trained models as services

Data scientists

- Ingest and prepare data
- Run experiments to explore data and train predictive models
- Deploy and manage trained models as web services

Automated Machine Learning

Enables non-expert to quickly create an effective machine learning model from data

Azure Machine Learning Designer

A graphical interface enabling no-code development of MLS

Data and compute management

Cloud-based data storage and compute resources that professional data Scientist can use to run data experiment code at scale

Pipelines

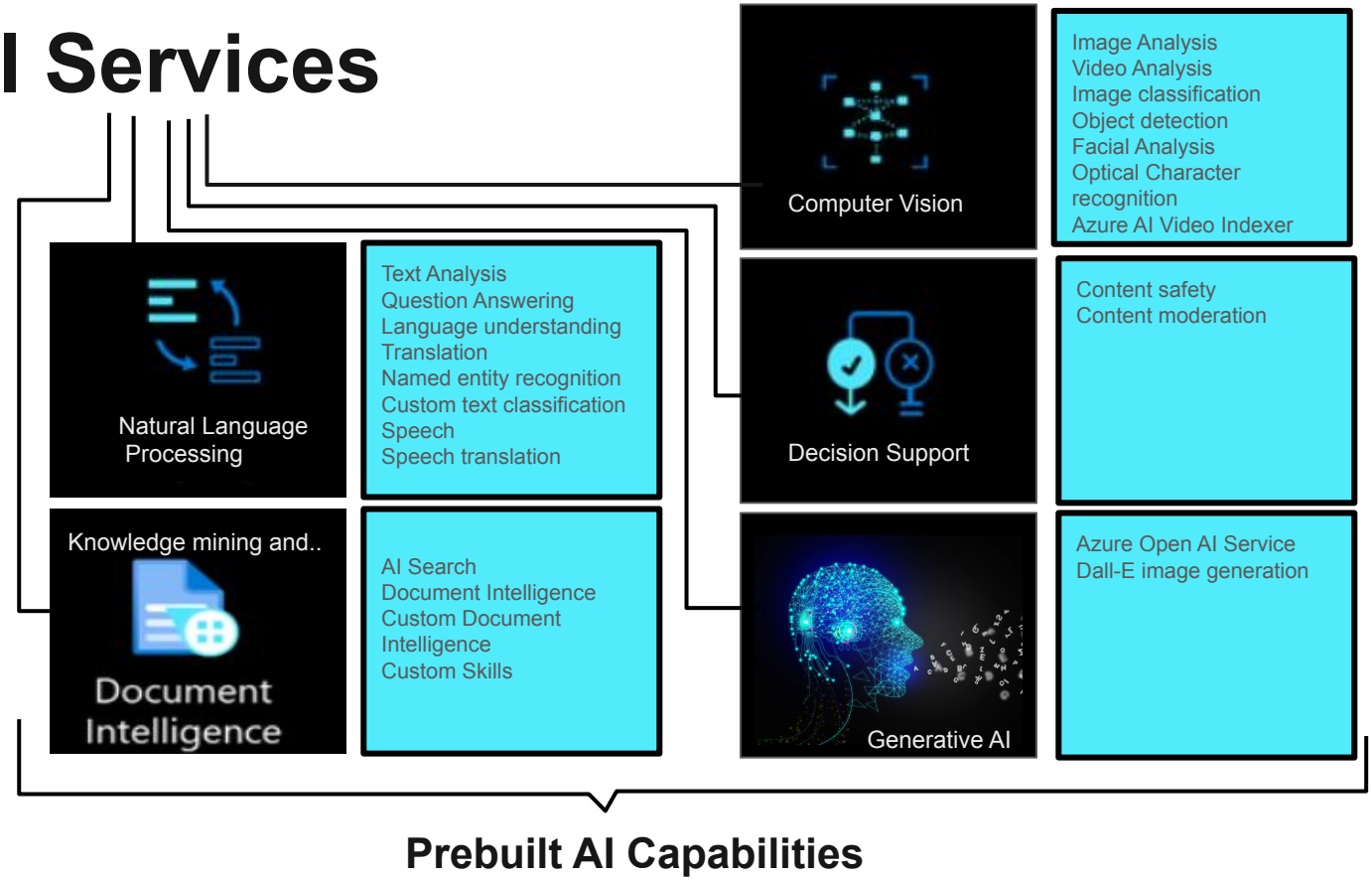
Data scientists, SW engineers, and IT operations professionals can define pipelines to orchestrate model training, deployment, and management tasks.

SW engineers

- Using automated machine learning or Azure Machine learning designer to train machine learning models and deploy them as services that can be integrated into AI-enabled applications
- Collaborating with Data scientist to deploy models based on common frameworks such as Scikit-Learn, PyTorch, and TensorFlow as web services, and consume them in applications.
- Using Azure ML SDKs or command-line Interface (CLI) scripts to orchestrate DevOps processes that manage versioning, deployment, and testing of MLMs as a part of an overall application delivery solution.

Azure AI Services

Rather than a single product , Azure is a set of individual services you can use as building blocks to compose sophisticated intelligent applications



Generative AI

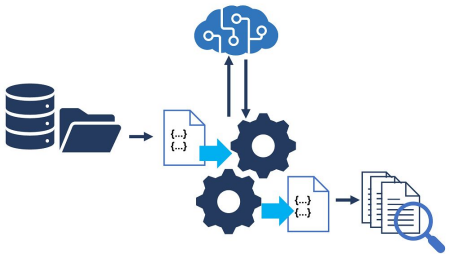
- ★ Relatively New
- ★ Quickly progressing
- ★ Focused in AI models that generate content
  - ○ Text
  - ○ Images
  - ○ Code
  - ○ More
- ★ Depend on Large Language Models LLMs
- ★ Queried with natural language prompts, generating impressively accurate responses when prompted correctly

Azure Open AI = OpenAI **In**

(both REST and language specific SDKs are available)

Azure Cognitive Search

- ➔ More useful search experience
- ➔ Insights generated can be analyzed and integrated into a data pipeline for a business intelligence solution



★ Applied AI service that enables you to ingest and index data from various sources and search the index to find, filter , and sort information from the source data

★ In addition to text-based indexing, Azure AI enables you to define an enrichment pipeline that uses AI skills to enhance the index with insights derived from the source data. (i.e., using computer vision and natural language processing to generate description of images, extract them from scanned docs and determine key phrases in large documents that encapsulate key points.

Check your knowledge

1. Which of the following best describes the predictions made by a machine learning model? *

- ( ) Absolutely correct values based on conditional logic.
- ( ) Randomly selected values with an equal chance of selection.
- (•) Probabilistic values based on correlations found in training data.

✓ That's correct. Machine learning models are trained using historic data, and rely on algorithms that find statistical relationships in the data. Predictions are generally based on probability; and while models are often extremely accurate, predictions are based on a confidence score that indicates a level of probability.

2. A data scientist has used Azure Machine Learning to train a machine learning model. How can you use the model in your application? *

- (•) Use Azure Machine Learning to publish the model as a web service.

✓ That's correct. You can use Azure Machine Learning to publish a trained model as a web service, and consume it from applications through its REST interface.

- ( ) Export the model as an Azure AI service.
- ( ) You must build your application using the Azure Machine Learning designer.

3. You want to index a collection of text documents, and search them from a mobile application. Which service should you use to create the index? *

- ( ) The Azure AI service
- (•) Azure AI Search

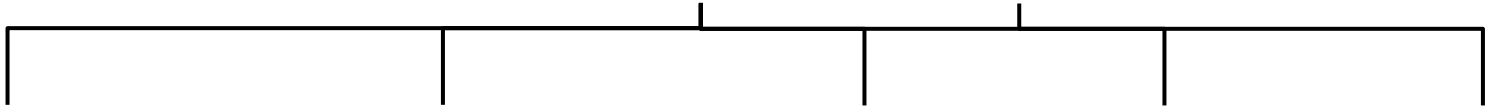
✓ That's correct. Use Azure AI Search to index documents for search.

- ( ) Azure OpenAI Service

Monitor

Monitor Azure AI Services

One of the main benefits of using cloud services is being able to pay only for the services you use



Plan Costs for AI services

- Azure pricing calc

View Costs

- Cost Analysis Tab

Create Alerts and Alert Rules

Based on events or metric thresholds

View Metrics

In the Azure portal

- Endpoint requests
- Data submitted
- Data returned
- Errors
- Etc

Add metrics to a dashboard

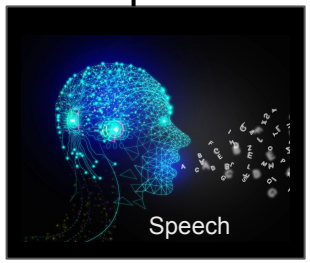
- Gain overall view of health and performance
- Add up to 100 dashboards

Create and consume

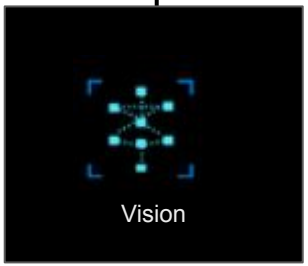
Rather than a single product , Azure is a set of individual services you can use as building blocks to compose sophisticated intelligent applications

Azure AI Services

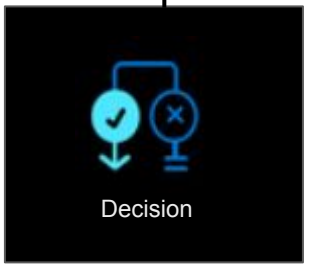
- ★ Azure AI Language
- ★ Azure AI Translator



- ★ Azure AI Speech



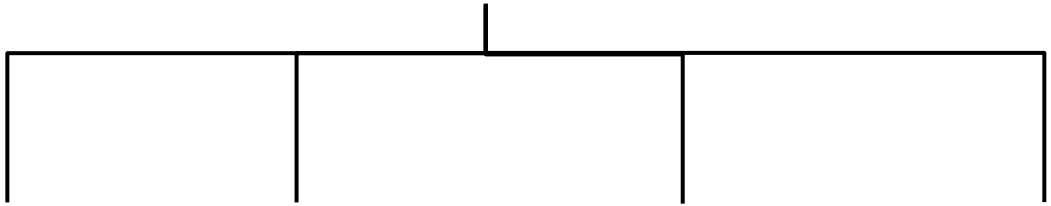
- ★ Azure AI Computer Vision
- ★ Azure AI Custom Vision
- ★ Azure AI Face



- ★ Azure AI Anomaly Detector
- ★ Azure AI Content Moderator
- ★ Azure AI Personalizer

Build your own AI solutions to provide out-of-the-box solutions for common AI scenarios

Azure AI Services include…



Azure AI Doc Intelligence

- ★ Optical Character Recognition (OCR)
- ★ Can extract semantic meaning from forms (invoices, receipts, etc)

Azure AI Immersive Reader

- ★ Reading solution
- ★ Supports people of all ages and abilities.

Azure Cognitive Search

- ★ Cloud-scale search
- ★ Extract insights from data and docs.

Azure OpenAI

- ★ Provide access to the capabilities of OpenAI GPT-4

To use any of the AI services…

- Create appropriate resources in an Azure subscription
- Define an endpoint where to consume the service
- Provide Access Keys for
  - Authenticated access
  - Manage Billing

Provisioning options

Multi Service resource

- Supports multiple different AI service
- I.e., A single resource that enables you to use AI Language + AI Vision
- Single credential to consume multiple services at a single endpoint and single billing.

Single Service resource

- Each AI service provisioned individually
- Enables you to use separate endpoints (i.e., different geographical regions)
- Manage access credentials and billing independently
- Generally offer a free tier
- Good choice to try out a service before using it in a production application.

Training and prediction resources

- Some services offer (or require) separate resources for model training and prediction.
- Ability to manage billing for training separately from model consumption
- Ability to use a dedicated service-specific resource to train model but generic to make the model available to application for inferencing.

Identify endpoints and keys

When you provision an Azure AI services service resource in you Azure subscription, you're defining an endpoint trough which it can be consumed by an application.

Applications require the following information:

endpoint URL

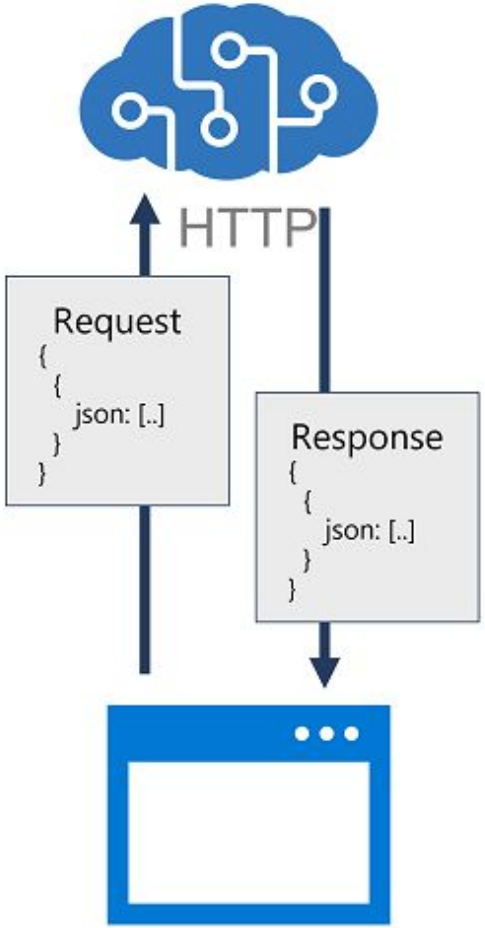
- **HTTP address at which the REST interface for the service can be accessed**
- **Most AI services Software Development Kits (SDKs) use URI as the endpoint to initiate connection.**

Subscription Key

- **Access is restricted based on a subscription key**
- **Client application must provide a valid key to consume the service.**
- **2 keys are created**
- **Applications can use wither key.**

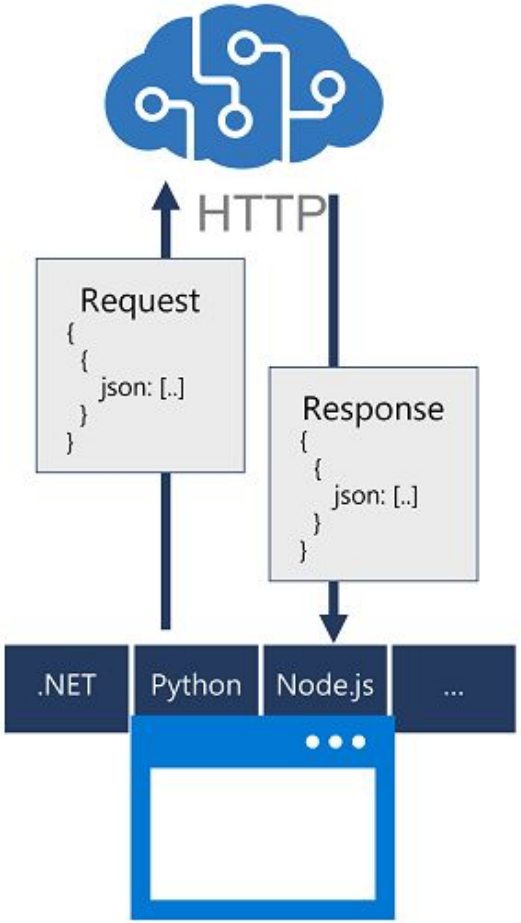
Resource location

- **When provisioned , resources are generally asigned ot a location.**
- **That will determine the Azure data center in which the resource is defined.**
- **While most SDK only use URI endpoint, some require the location.**



Use a REST API

- **Client applications use them to consume services.**
- **Service functions can be called by submitting data in JASON format over an HTTP request (POST, PUT, or GET).**
- **Results returned to the client as an HTTP response, often in JASON containing the output data from the function.**
- **Any programming language or tool capable of Submiting and receiving JASON over HTTP can be used to consume AI services.**
- **Programming languages**
  - **Microsoft C#**
  - **Python**
  - **JS**
- **Utilities**
  - **Postman**
  - **cURL**



Use an SDK

Software Development Kits

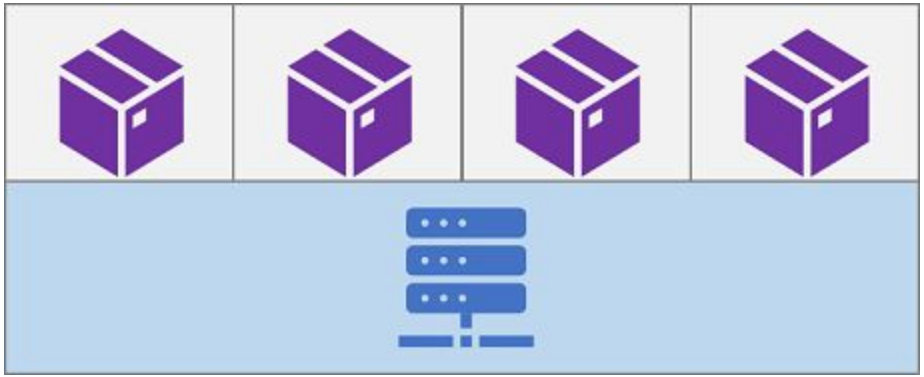
- Easier to build more complex solutions by using native libraries for the programming language in which you're developing the application

For most services there's an SDK for languages such as:

- **Microsoft C# (.NET Core)**
- **Python**
- **JavaScript (Node.js)**
- **Go**
- **Java**

- Each SDK includes packages that you can install in order to use service-specific libraries in your code
- Online documentation to help you determine the appropriate classes, methods and parameters used to work with the service

Deploy in Containers

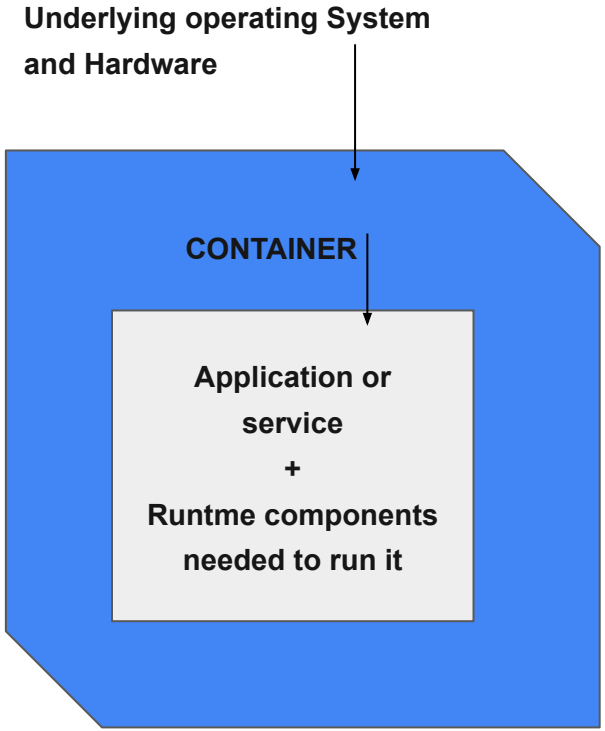


Examples of Container Host

- ★ Docker Server
- ★ Azure Container Instance (ACI)
- ★ Azure Kubernetes Service (AKS(

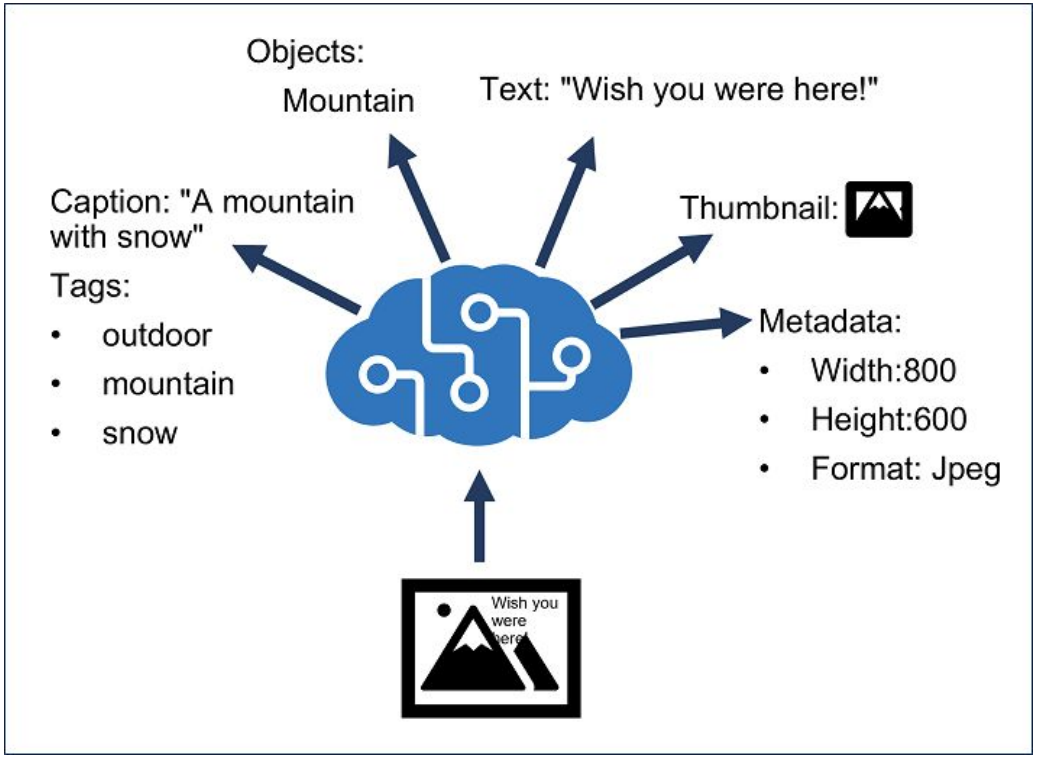
Benefits

- ★ Portables across hosts
- ★ Single container host can support multiple isolated containers
  - ○ Easier to consolidate multiple applications with different configuration requirements



Create a Computer Vision Solution

Azure AI Vision



Benefits

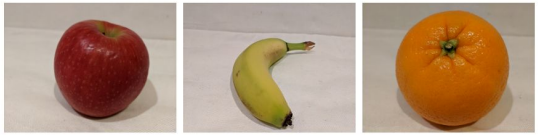
★ **Designed to Extract info from images**

- **Description and tag generation**
- **Object detection**
- **Image metadata, color, and type analysis**
- **Category identification**
- **Bakground removal**
- **Moderation Rating**
- **Optical Character recognition**
- **Smart Thumbnail generation**

★ **You can provision it as a single-service resource, or in a multi-service Azure AI Services resource**

Custom Model types

Image Classification

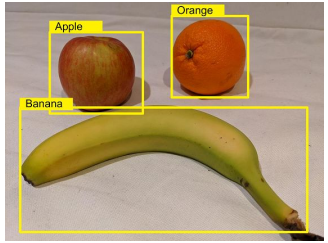


Apple Banana Orange

Model Trained to

- **Model trained to predict a label**
- **Based on the contents of the entire image**
- **Label relates to the main subject of the image**
- **Can be trained for multi-class classification**
  - **Each image can belong to only one class**
- **Or multi-label**
  - **Image can be associated with multiple labels.**

Object Detection



- ★ **Model trained to detect the presence and location of one or more classes of object in an image.**
- ★ **Uses**
  - **AI enabled checkout system in a grocery store able ot identify the type and location of items being purchased**
- ★ **Components**
  - **Class Label: i.e. banana, apple, orange**
  - **Location: Coordinates of a bounding box**

Product Recognition

- **Similar to Object detection**
- **Improved accuracy for**
  - **Product Labels**
  - **Brand Names**
- **Components**
  - **Class Labels**
  - **Location**

Understand Custom Model types

Components of a custom Vision project

Dataset + COCO file (labels)

- **Collection of images**
- **Coco defining the label information**
- **Dataset stored in an Azure blob storage container**

Type of Model

★ A

Budget (Time)

- **Similar to Object detection**
- **Improved accuracy for**
  - **Product Lables**
  - **Brand Names**
- **Components**
  - **Class Labels**
  - **Location**

STEPS

Json File with specific format that defines: images, annotations, and categories

- Create blob storage container
- Upload just training images

- Create data set
  - Define type:
    - Image classification
    - Object detection
    - Product recognition
- Connect it to blob storage container

You can use Vision Sudio or REST API

- Label your data in Azure Machine Data Labeling Project
- This creates the COCO file

- Connect the COCO file for the labeled dataset to your data set

- Train your custom model on the dataset and labels created.

- Verify performance
- Iterate if performance isn't meeting expectations

- Once you're happy with the performance, the model can be used in Vision Studio or in your own application.

Label and Train

Label

Be sure to accurately assign labels and completely label all instances of each class

ML Data Labeling

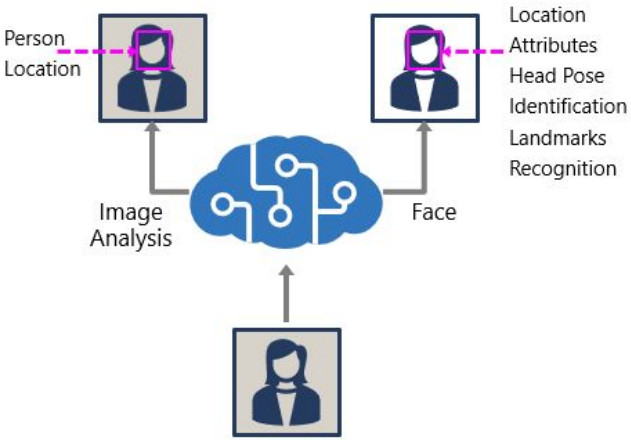
Create an Azure ML Data Labeling project to label your data and import it back to Vision Studio in the form of a COCO file.

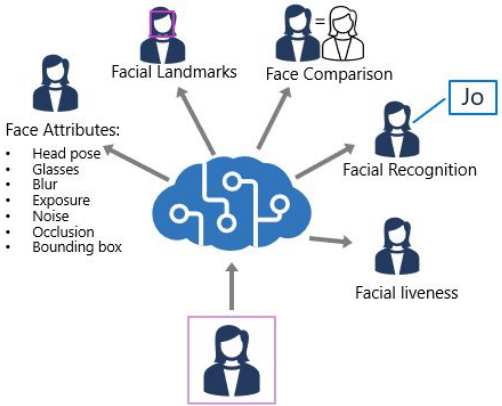
Detect Faces

Detect faces or people in images

The Azure AI Vision service

- **Detect People in an image**
- **Return bounding box for its location**





Face Service

-

Reading text

Two features read text from docs and images

Both are accessible through REST API or a client library.

Image Analysis

Optical Character Recognition (OCR)

- Use it to…
  - Read general, unstructured docs with smaller amount of text.
  - Read Images that contain text.
- In addition to extracting text….
  - Object detection
  - Describing an image
  - Categorizing an image
  - Generating smart-cropped thumbnails
- Results are returned immediately (synchronous) from a single API call
- Examples:
  - Street Signs
  - Hand-written notes
  - Store signs

Document intelligence

- Use it to…
  - Read small to large volumes of text from images and PDF docs
- Uses context and structure of the doc to improve accuracy
- Returns an asynchronous operation ID
  - Results are retrieved in a subsequent API call
- Examples:
  - Receipts
  - Articles
  - Invoices

1. Which API would be best for this scenario? You need to read a large number of files with high accuracy. The text is short sections of handwritten text, some in English and some of it is in multiple languages. *

- ○ A custom Language API
- ○ Document Intelligence API
- ◉ Image Analysis API

✔ Correct: The Image Analysis service OCR feature is best suited for short sections of handwritten text.

2. What levels of division are the OCR results returned? *

- ○ Only total content and pages of text.
- ◉ Blocks, words and lines of text.

✔ Correct: Results contain blocks, words and lines, as well as bounding boxes for each word and line.

- ○ Total content, image tags, pages, words and lines of text.

3. You've scanned a letter into PDF format and need to extract the text it contains. What should you do? *

- ○ Use the Azure AI Custom Vision service
- ○ Use the Image Analysis API of the Azure AI Vision service.
- ◉ Use the Document Intelligence API.

✔ Correct: The Document Intelligence API can be used to process PDF formatted files.

Analyze Video

Knowledge check

200 XP

3 minutes

Check your knowledge

1. You want Azure Video Indexer to analyze a video. What must you do first? *

- ( ) Use the Azure AI Vision service to extract key frames from the video.
- (•) Upload the video to Azure Video Indexer and index it.

  ✔ That's correct. You need to index a video before analyzing it.
- ( ) Store the video file in an Azure blob store container.

2. You want Azure Video Indexer to recognize brands in videos recorded from conference calls. What should you do? *

- (•) Edit the Brands model to show brands suggested by Bing, and add any new brands you want to detect.

  ✔ That's correct. You can both detect known brands, and well as include new brands you want to detect by providing information about it.
- ( ) Edit the conference call videos to include a caption of each brand seen on their first appearance.
- ( ) Embed the Azure Video Indexer widgets in a custom web site that has all the brand images stored for reference.

Facial Recognition

- Detecting the presence of individual in the image
- Requires Limited Access approval

Optical Character Recognition

- Reading Text in Video

Speech Transcription

Text Transcript of Spoken dialog in video

Azure VIDEO Indexer

Scene Segmentation

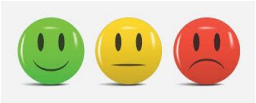
Breakdown of video into its constituent scenes

Identification of key topics discussed in video

Content Moderation

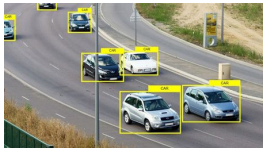
- Detection of adult or violent Themes in the video

Sentiment



How positive or negative segments within the video are

Labels



- Label tgs that identify key objects or themes throughout the video

Develop natural language processing solutions with Azure AI Services

Azure VIDEO Indexer

Predefined Models

- Recognize well-known celebrities
- Do OCR
- Transcribe spoken phrases into text

Or..

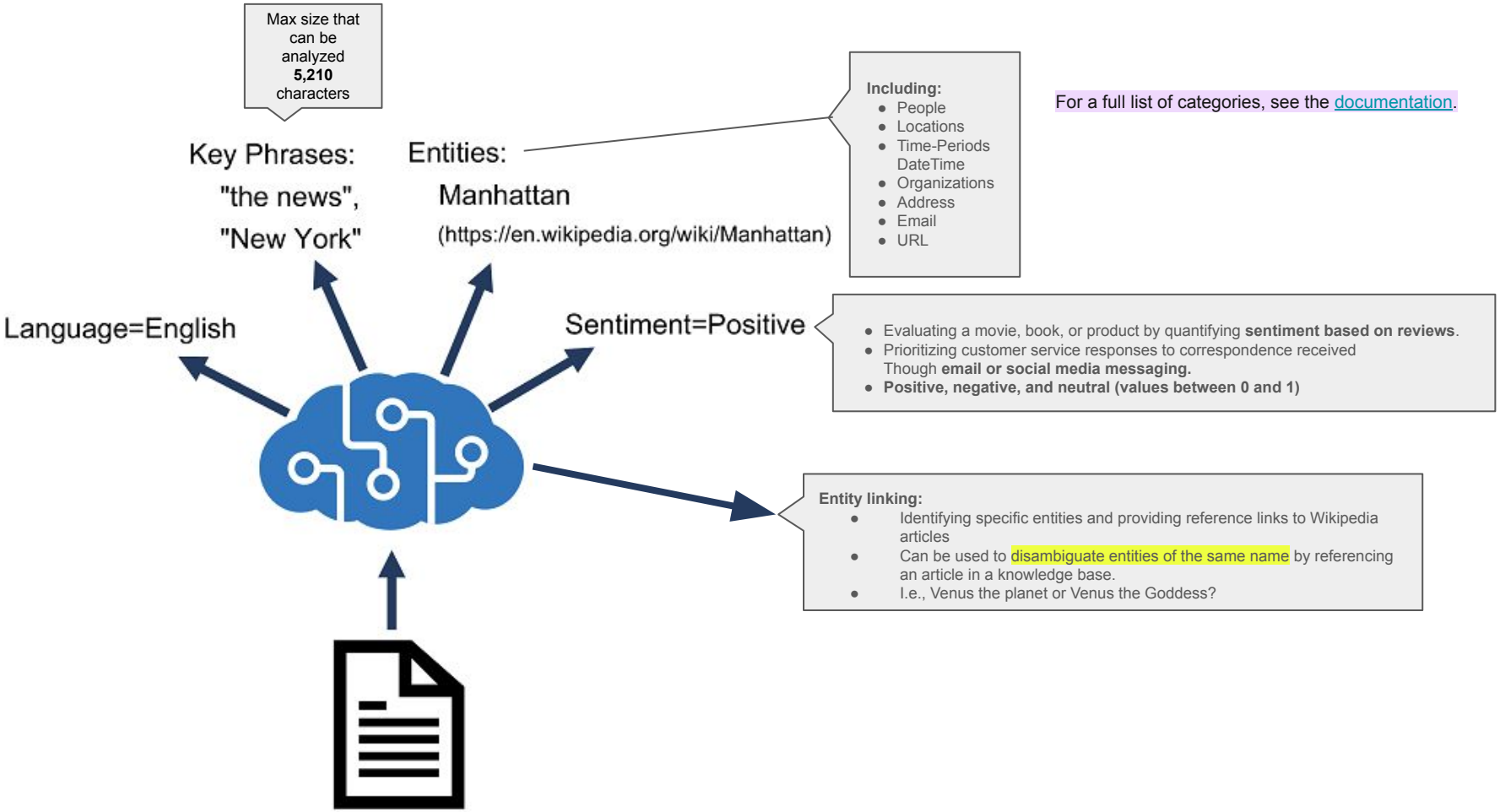
Create Custom Models

- People: Add images of the faces of people you want to recognize in videos.
- Language: Specific terminology used by your organization. Detect it and transcribe it.
- Brands: Train a model to recognize specific names and brands. I.e., identify a product or project or company that are relevant to your business.

Portfolio Ideas

- Black listed people detection for Casinos
-

Analyze Text


Max size that can be analyzed **5,210** characters
Key Phrases:
"the news",
"New York"
Entities:
Manhattan
(https://en.wikipedia.org/wiki/Manhattan)
Including:
People
Locations
Time-Periods
DateTime
Organizations
Address
Email
URL
For a full list of categories, see the documentation.
Language=English
Sentiment=Positive
Evaluating a movie, book, or product by quantifying **sentiment based on reviews**.
Prioritizing customer service responses to correspondence received Though **email or social media messaging.**
Positive, negative, and neutral (values between 0 and 1)
Entity linking:
Identifying specific entities and providing reference links to Wikipedia articles
Can be used to disambiguate entities of the same name by referencing an article in a knowledge base.
I.e., Venus the planet or Venus the Goddess?

Knowledge check

200 XP

3 minutes

1. How should you create an application that monitors the comments on your company's web site and flags any negative posts? *

- Use the Azure AI Language service to extract key phrases.
- Use the Azure AI Language service to perform sentiment analysis of the comments.

  ✔ Correct. Sentiment analysis helps you determine if text is negative or positive.
- Use the Azure AI Language service to extract named entities from the comments.

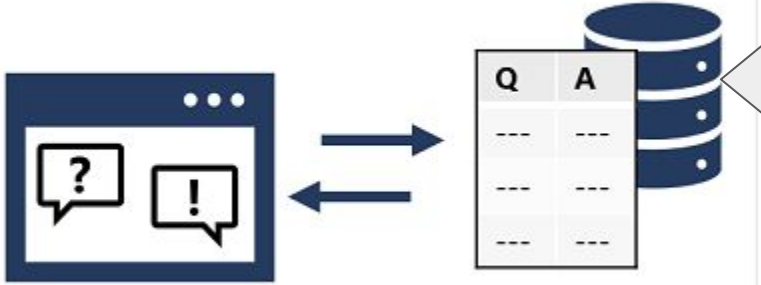
2. You are analyzing text that contains the word "Paris". How might you determine if this word refers to the French city or the character in Homer's "The Iliad"? *

- Use the Azure AI Language service to extract key phrases.
- Use the Azure AI Language service to detect the language of the text.
- Use the Azure AI Language service to extract linked entities.

  ✔ Correct. Linked entities enable you to disambiguate common entities of the same name.

Question Answering

Formerly: QnA Service (Stills exists as a standalone service)



Knowledge Base

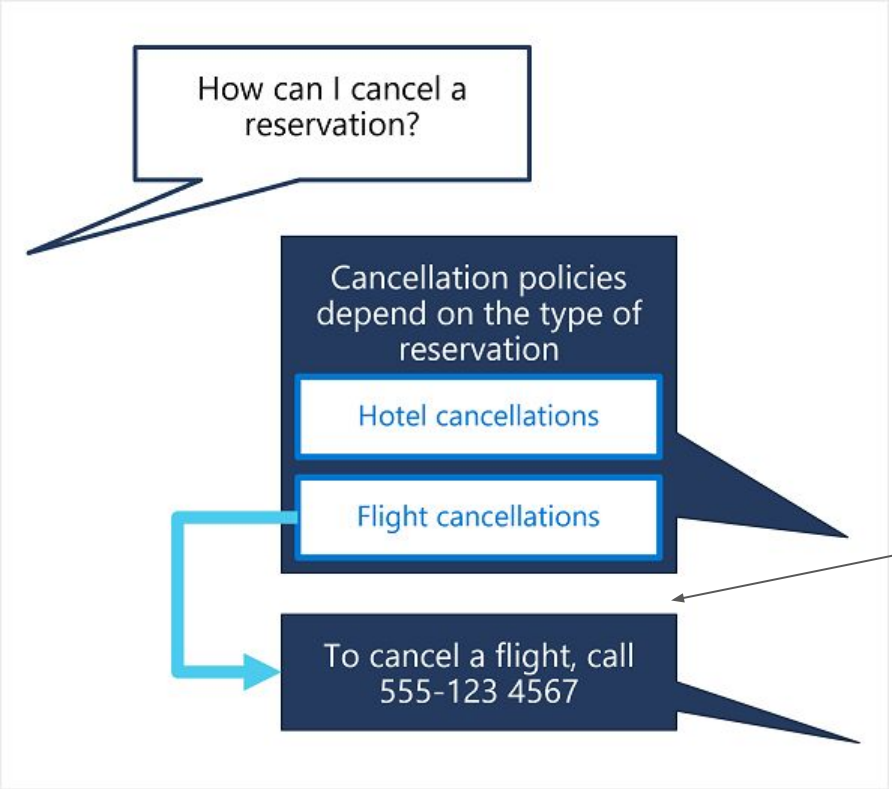
- **Can be created from existing sources like:**
  - **FAQ websites**
  - **Structured text Files**
    - **Brochures**
    - **User guides**
  - **Built-in chit chat question and answer pairs.**

Is a form of language model, which raises the question of when to use the ***conventional language understanding capabilities*** **of Azure AI language**

Question Answering vs Azure AI Language understanding

	Question answering	Language understanding
Usage pattern	User submits a question, expecting an answer	User submits an utterance, expecting an appropriate response or action
Query processing	Service uses natural language understanding to match the question to an answer in the knowledge base	Service uses natural language understanding to interpret the utterance, match it to an intent, and identify entities
Response	Response is a static answer to a known question	Response indicates the most likely intent and referenced entities
Client logic	Client application typically presents the answer to the user	Client application is responsible for performing appropriate action based on the detected intent

Multi-Turn conversations



Budget (Time)

..You can

- **Define answers to be taken by the AI from existing web page or document**

Or..

- **You can explicitly define follow-up prompts**
- **Example:**
  - **If cancellation policies depend on the type of reservation like shown here**

Check your knowledge

1. You want to create a knowledge base from an existing FAQ document. What should you do? *

- Create an empty knowledge base and manually enter the FAQ questions and answers.
- Create a new knowledge base, importing the existing FAQ document.

✔ Correct. You can create a knowledge base from an existing document or web page.

- Create a new knowledge base, selecting only the Professional chit-chat source.

2. How can you add a multi-turn context for a question in an existing knowledge base? *

- Add synonyms to the knowledge base.
- Add alternative phrasing to the question.

✕ Incorrect. To add a multi-turn context to a question, define a follow-up prompt.

- Add a follow-up prompt to the question.

✔ Correct. To add a multi-turn context to a question, define a follow-up prompt.

3. How can you enable users to use your knowledge base through email? *

- Add Friendly Chit-chat to the knowledge base.
- Enable Active Learning for the knowledge base and include the user's email address as the userId parameter in responses.
- Create a bot based on your knowledge base and configure an email channel.

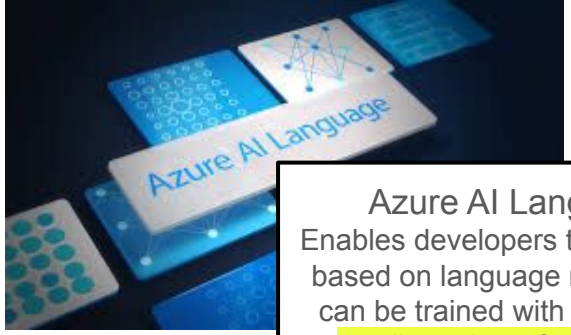
✔ Correct. You can create a bot for your published knowledge base and configure a channel for email communication.

Natural Language Understanding

Natural Language Processing (NLP)

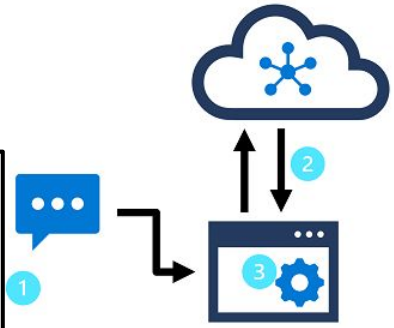
Natural Language Understanding (NLU)

Determining Semantic meaning from natural language



Azure AI Language

Enables developers to build apps based on language models that can be trained with a relatively small number of samples to discern a user's intended meaning



Pre-Configured Features

No modeling labeling or training needed.

Available for documents and conversations	Summarization
People, places, or companies	Named entity recognition
Email, address,IP, names, SSN	Personally identifiable Information (PII) detection
Pulls main concepts out of provided text	Key phrase extraction
How positive or negative a string or doc is	Sentiment Analysis
Identifies the language for each string or doc	Language detection

Full list of capabilities

Learned Features

Requires you to train, label and deploy a model to make it available to be used in your app

- Conversational language understanding (CLU)
- Custom named entity recognition
- Custom text classification
- Question answering

Full list of capabilities.

Utterances

Phrases a user might enter when interacting with an app that uses your LM

Get time
- What time is it?
- What is the time?
- Tell me the time

Use patterns to differentiate similar utterances

"Turn the {DeviceName} on"

Intents

- Vary the length
- Vary the location of the noun
- Use correct and incorrect grammar

GetTime
GetWeather
TurnOnDevice

None

Identifies utterances the user might submit, but for which there is no specific action required

Entities

Add specific context to intents

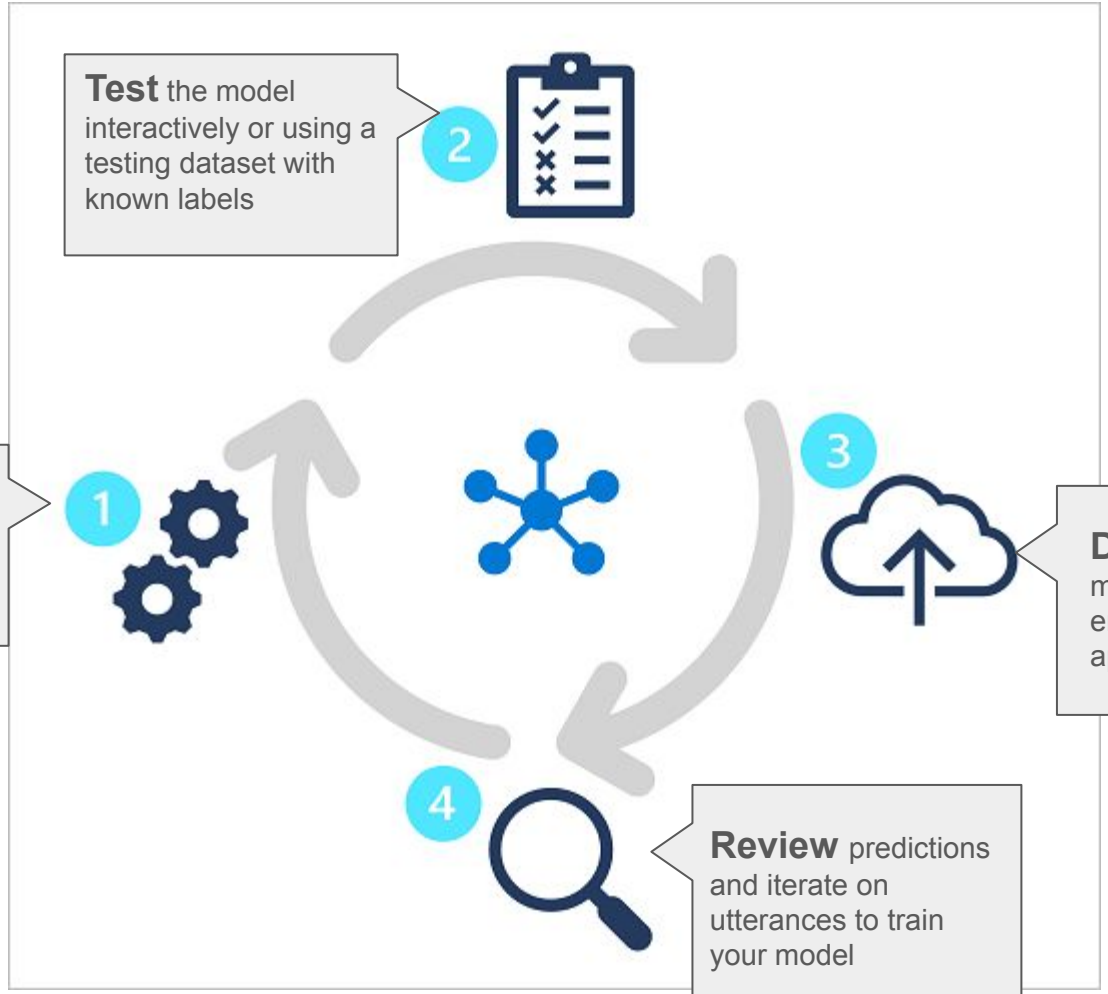
I.e.,
What is the time in London

Learned

List
(DayOfWeek)

Prebuilt

supported prebuilt entities.


Test the model interactively or using a testing dataset with known labels
2
Train a model to learn intents and entities from sample utterances
1
3
Deploy a trained model to public endpoint so client apps can use it
4
Review predictions and iterate on utterances to train your model

1. Your app must interpret a command such as "turn on the light" or "switch the light on". What do these phrases represent in a language model? *

- ○ Intents.
- ◉ Utterances.
  - ✓ Correct. Utterances are example phrases that indicate a specific intent.
- ○ Entities.

2. Your app must interpret a command to book a flight to a specified city, such as "Book a flight to Paris." How should you model the city element of the command? *

- ○ As an intent.
- ○ As an utterance.
- ◉ As an entity.
  - ✓ Correct. The city is an entity to which the intent (booking a flight) should be applied.

3. Your language model needs to detect an email when present in an utterance. What is the simplest way to extract that email? *

- ○ Use Regular Expression entities.
- ◉ Use prebuilt entity components.
  - ✓ Correct. When a language model needs to detect a common entity, use prebuilt components to have the Azure AI Language service automatically detect the entity.
- ○ Use Learned entity components.

Next unit: Summary

Continue >

NLP

- One of the most common AI problems
- Software must interpret text or speech in the natural form use
- Part of NLP is the ability to classify text
  - Including:
    - Sentiment
    - Language

Custom categories defined by user

- Single label classification
  - Only one class (label)
- Multiple label Classification
  - Can assign multiple class to each file

I.e., Classifying a video game as “Adventure” or as “Strategy”

I.e., Classifying a video game as “Adventure and Strategy”

Single vs Multiple

- Labeling
  - (Multiple more complex for quality control)
- Considerations to improving your model
  - **Recall:** Of all the actual labels, how many were identified?
    - True Positives/All Labels
  - **Precision:** How many of the predicted labels are correct?
    - True Positives/All Positives
  - **F1 Score:** Function of recall and precision
- API payload

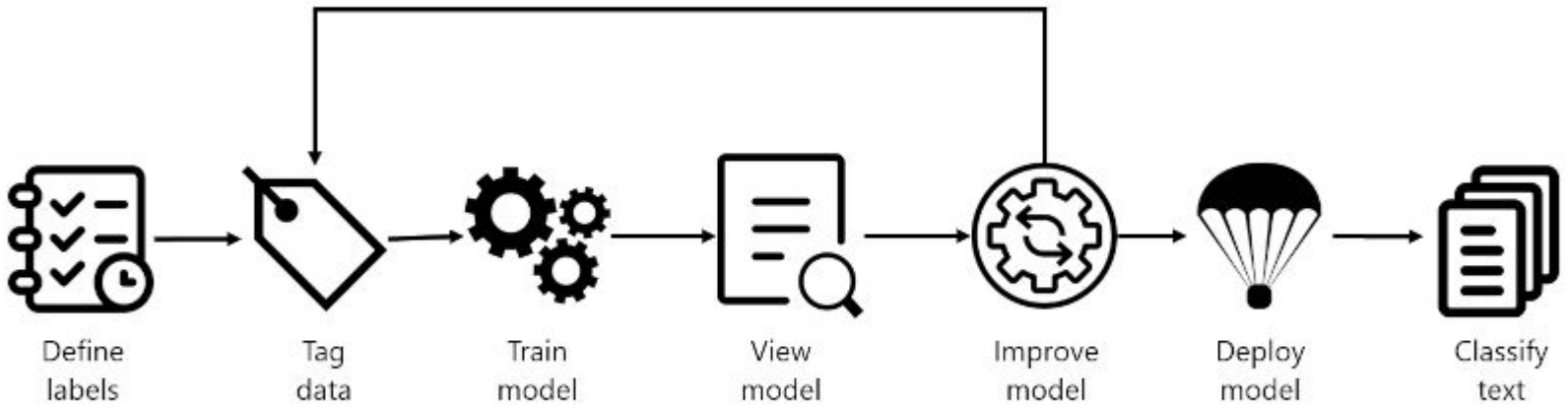
Intended to provide a single score to maximize for a balance for each component

How these metrics are calculated?

Custom Text Classification

How to Build text Classification projects

Custom



Define labels	Tag data	Train model	View model	Improve model	Deploy model	Classify text
● Understand the data ● Identify possible labels I.e., in the video game example: “Action”,”Adventure”, “Strategy”...	- Use the labels - The model will learn to classify future files. - Avoid ambiguity having different labels - Provide good examples	Train your model with the labeled data	- View the results - Score between 0 and 1 - Which genre didn’t perform well?	Improve by adding data to the classifications that failed	Deploy to make it available through API	Use it

How to split datasets for training

Training

- Used to actually train the model
- Data and labels fed into the ML algo to teach your model what data should be classified with which label
- Training dataset will be the larger (80% of labeled data)

Testing

- Labeled data used to verify your model after it's trained.
- Azure compares the model's output to how you labeled your data to determine how well the model performed

Azure AI Language allows to create multiple models and multiple deployment. (limit of 10 names)

FYI

Benefits:

- **Test** two models side by side
- **Compare** how the split of dataset impact performance
- Deploy **multiple versions** of your model

The API for the Azure AI Language service operates asynchronously for most calls.

FYI

We submit request and check back with another call to get status

Check your knowledge

1. You want to train a model to classify book summaries by their genre, and some of your favorite books are both mystery and thriller. Which type of project should you build? *

- ( ) A single label classification project
- (•) A multiple label classification project

  ✓ That answer's correct. Use a multiple label classification project to label books as multiple genres.
- ( ) A varied label classification project

2. You just got notification your training job is complete. What is your next step? *

- ( ) Label more data
- ( ) Deploy your model
- (•) View your model details

  ✓ That answer's correct. First view your model details to see how it scored, the classification distribution, and where it needs improvement.

3. You want to submit a classification task via the API. How do you get the results of the classification? *

- ( ) The result is in the response of the classification request.
- (•) Call an endpoint with your deployment name to get the most recent classification.

  ✗ That answer's incorrect. Get the value from the `operation-location` header in the request response, and use that to retrieve the results of the classification request.
- ( ) Call the URL provided in the header of the request response.

  ✓ That answer's correct. Get the value from the `operation-location` header in the request response, and use that to retrieve the results of the classification request.

Custom named entity recognition(NER)

= person, place, thing, event, skill, or value

AKA Entity Extraction

A full list of recognized entity categories is available in the NER docs.

NER

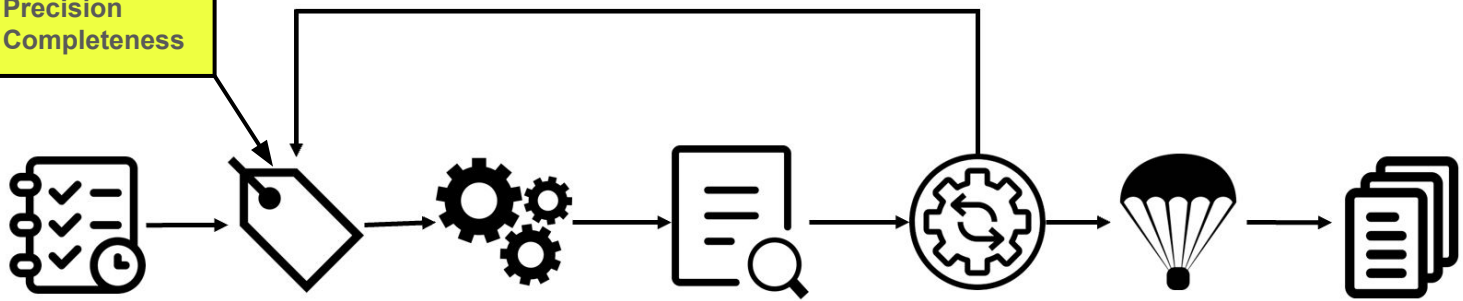
Built-in

Person
Location
Organization
URL

Custom

- The entities aren't part of the built-in
- Only want to extract specific entities
- i.e., Specific legal or bank data, Knowledge mining to enhance Catalog search, looking for specific text for audit policies.

Focus on…
Consistency
Precision
Completeness



Which part of the bank statement you want to extract?

Bank name, customer name, address, loan or account terms

Language Studio

LIMITS
- Training: >10<100K files
- Deployments: 10 names per project.
-APIs
  -Authoring: 10 POST and 100 GET max
  -Analyze: 20 GET or POST max
- 1 Storage per project
- 500 projects per resource
- 50 trained models per proj
- up to 200 Entities
- 500 char per entity

Considerations for data selection

- Diversity → As many different sources as possible
- Distribution → Appropriate distribution of document types
- Accuracy → Use as close to real world data as possible
- Entities → Avoid ambiguous entities. I.e., two names next to each other on a bank statement. If ambiguity is needed, then add more examples to better train the model.

Check your knowledge

1. You've trained your model and you're seeing that it doesn't recognize your entities. What metric score is likely low to indicate that issue? *

- (●) Recall
  - ✔ That answer's correct. Recall indicates how well the model extracts entities, regardless of which entity that is.
- ( ) Precision
- ( ) F1 score

2. You just finished labeling your data. How and where is that file stored to train your model? *

- (●) JSON file, in my storage account container for the project
  - ✔ That answer's correct. The JSON file lives next to the dataset in your container for the model to use during training.
- ( ) XML file, in my local project folder
- ( ) YAML file, anywhere in my Azure account

3. You train your model with only one source of documents, even though real extraction tasks will come from several sources. What data quality metric do you need to increase? *

- ( ) Distribution
- ( ) Accuracy
- (●) Diversity
  - ✔ That answer's correct. Having the right data diversity will lead to better extraction performance.

Translate Text with Azure

90 Supported Languages

Custom Translations

You can create and train a model For businesses or industries that Have specific vocabularies of terms that require custom translations

One-to-many

Translation:

- en: "Hello"
- fr: "Bonjour"

Translation Options

- Word alignment
- Sentence length
- Profanity filtering
  - No Action/Deleted/Marked(***)
  - profanityMarker value of Tag (enclosed in XML tags)

Where to get the resource?

- Azure AI Translator resource

- Multi-Service Azure AI Services
→Text analytics API

Detect:

Language: ja

Transliteration:

"Kon'nichiwa"

Converting text from its native script to an alternative script

こんにちは

Knowledge check

200 XP

2 minutes

1. What function of Azure AI Translator should you use to convert the Chinese word "你好" to the English word "Hello"? *

- ○ Detect
- ◉ Translate

✔ Correct. Translation converts text from one language to another.

- ○ Transliterate

2. What function of Azure AI Translator should you use to convert the Russian word "спасибо" in Cyrillic characters to "spasibo" in Latin characters? *

- ○ Detect
- ○ Translate
- ◉ Transliterate

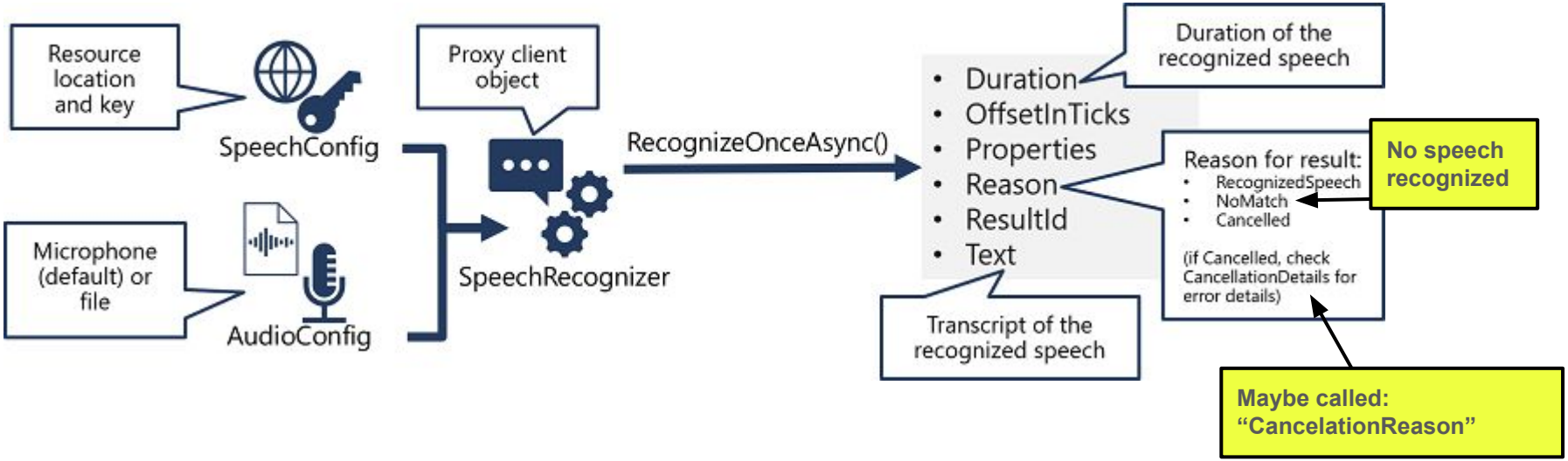
✔ Correct. Transliteration converts text from one script to another.

Next unit: Summary

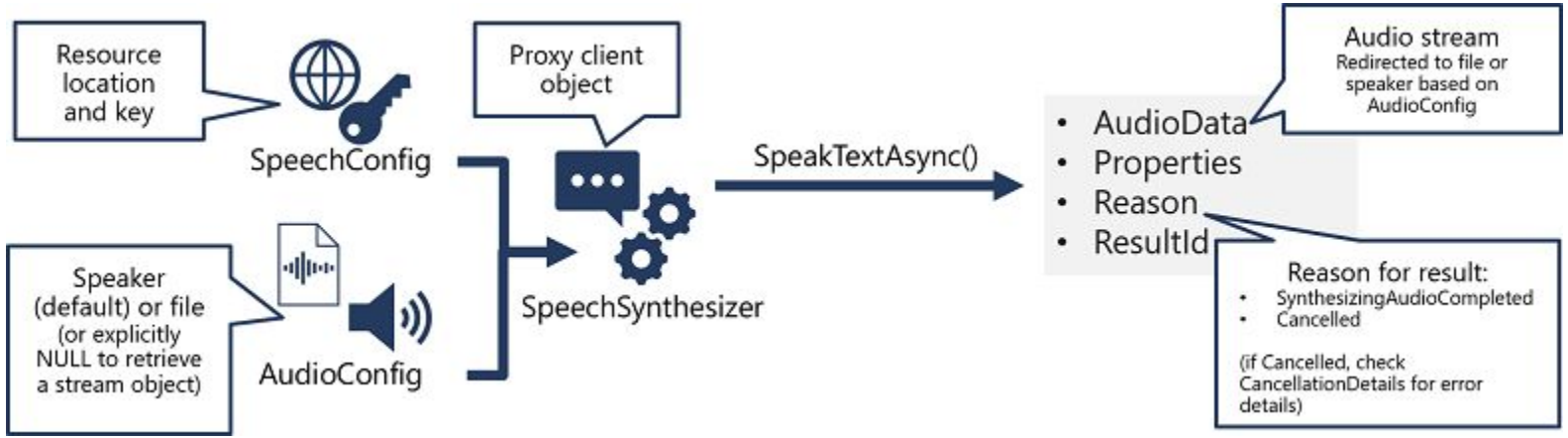
Continue >

Speech-enabled Apps

Azure AI Speech SDK



Text to speech API



You can use plain tex or…

Use Speech Synthesis Markup Language (SSML)

- **XML-based syntax**
- **for better control on how the spoken output sounds**

```
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
                     xmlns:mstts="https://www.w3.org/2001/mstts"xml:lang="en-US">
    <voice name="en-US-AriaNeural">
        <mstts:express-asstyle="cheerful">
          I say tomato
        </mstts:express-as>
    </voice>
    <voice name="en-US-GuyNeural">
        I say <phoneme alphabet="sapi" ph="t ao m ae t ow"> tomato </phoneme>.
        <break strength="weak"/>Lets call the whole thing off!
    </voice>
</speak>
```

Use SpeechConfig object to customize the audio returned by the Azure AI Speech service

- **Audio Format: file type / Sample-rate / Bit-depth**
- **Voices**
  - **Standard voices**
    - **Synthetic voices created from audio samples**
  - **Neural Voices**
    - **More natural sounding created using deep neural networks**

Full list of supported formats

More Info.

Knowledge check

200 XP

3 minutes

1. What information do you need from your Azure AI Speech service resource to consume it using the Azure AI Speech SDK? *

- [x] The location and one of the keys

  ✔ **Correct. The Azure AI Speech SDK requires the location and a key to connect to the Azure AI Speech service.**

- [ ] The primary and secondary keys
- [ ] The endpoint and one of the keys

2. Which object should you use to specify that the speech input to be transcribed to text is in an audio file? *

- [ ] SpeechConfig
- [x] AudioConfig

  ✔ **Correct. Use an AudioConfig to specify the input source for speech.**

- [ ] SpeechRecognizer

3. How can you change the voice used in speech synthesis? *

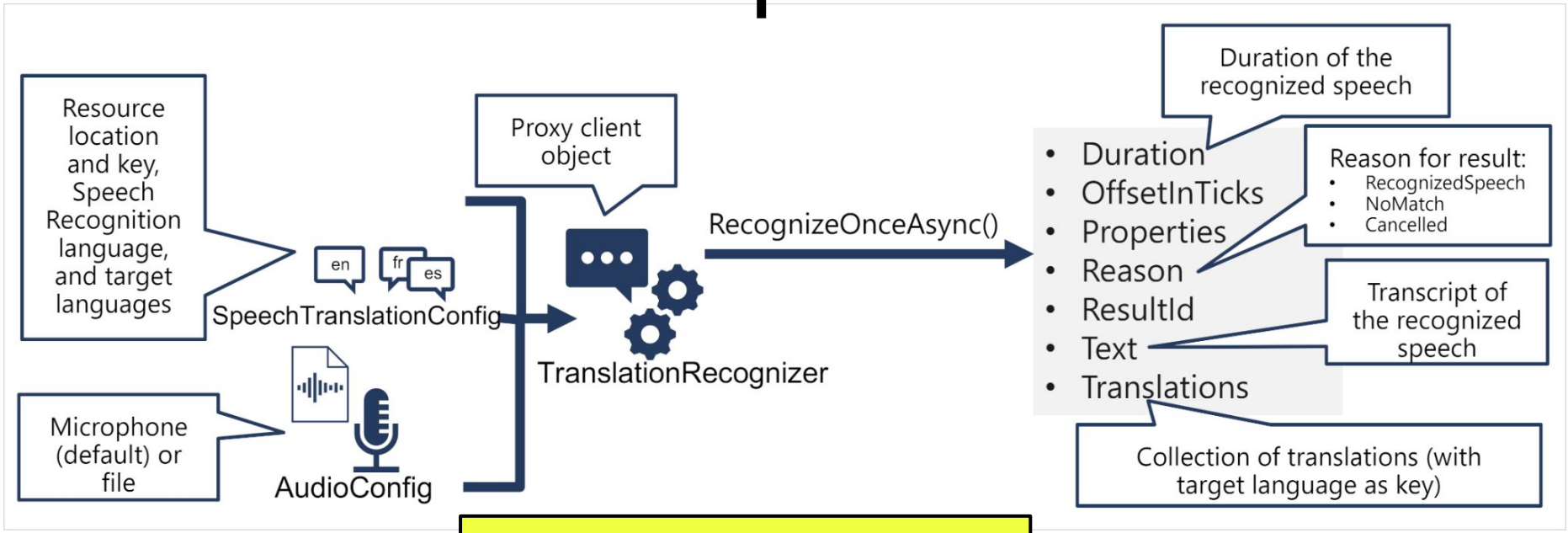
- [ ] Specify a SpeechSynthesisOutputFormat enumeration in the SpeechConfig object.
- [x] Set the SpeechSynthesisVoiceName property of the SpeechConfig object to the desired voice name.

  ✔ **Correct. To set a voice, set the SpeechSynthesisVoiceName property of the SpeechConfig to a voice name, such as "en-GB-George".**

- [ ] Specify a filename in the AudioConfig object.

Translate Speech

Translate Speech to text



To synthesize the translation as speech…

- **Capture the translation with an event-based synthesis**
- **Manual Synthesis**

Knowledge check

200 XP

3 minutes

1. Which SDK object should you use to specify the language(s) into which you want speech translated? *

- SpeechConfig
- SpeechTranslationConfig
  - ✔ Correct. Specify target languages in the SpeechTranslationConfig object.
- AudioConfig

2. Which SDK object should you use as a proxy for the Translation API of Azure AI Speech service? *

- TranslationRecognizer
  - ✔ Correct. Use a TranslationRecognizer to call the Translation API of the Azure AI Speech service.
- SpeechRecognizer
- SpeechSynthesizer

3. When translating speech, in which cases can you use the Synthesizing event to synthesize the translations and speech? *

- Only when translating to a single target language.
  - ✔ Correct. You can only use event-based synthesis when translating to a single target language.
- Only when translating to multiple target languages.
- When translating to one or more target languages.

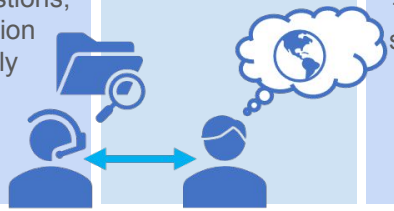
Azure AI Search Solution

Knowledge Mining

Justification

All companies rely on information to make decisions, answer questions, and function efficiently

And now a days they have **A LOT** of it



So much, that is not easy to **find and extract the information** from the massive set of documents, databases, and other sources.

Example: Margie's Travel agency. Sources: **Brochures**, **reviews** of hotels submitted by customers, **websites**, **articles**, etc.

To optimize for scalability and availability…

Replicas

- Instances of the search service
- More replicas = More Capacity to manage multiple query request at the same time while managing indexing.

Partitions

- Divisions of an Index into multiple storage locations
- Split I/O operations (i.e., Querying, rebuilding an index.

Search Units (SU) = R X P

Solution

- Provides a cloud-based solution for **Indexing** and **Querying** a wide range of data sources and **creating** comprehensive and high scale **SEARCH SOLUTIONS**
  - **Index** docs and data from different sources.
  - Use cognitive skills to **enrich index data**
  - **Store extracted insights** in a knowledge store for analysis and integration

TIERS

Free (F) - **Explore** and **test**
Basic (B) - max **15 indexes and 5GB of data**
Standard (S) - **Enterprise Scale** Solution
Storage Optimized (L) - **Large** indexes **less speed**

Choose wisely You can't change the pricing tier for your solution later. You'll have to create a new resource from scratch!

SEARCH
COMPONENTS

Facetable is typically used in a presentation of search results that includes a hit count by category.

Data Source

- Unstructured files in Azure blob storage containers
- Tables in Azure SQL DB
- Documents in Cosmos DB.
- Applications can push JSON data directly into an index.

Skillset

- AI-Enriched Search (Enrich the source data to be indexed)
- Apply AI skills as part of the indexing process

Language / Key phrases / sentiment / locations / description of images / custom skills that you develop to meet specific requirements

Indexer

- Engine that drives the overall indexing process
- Takes the skillset + Data extracted from original source and maps them to fields in the index
- Automatically run or can be scheduled to add more documents to the index.

Index

- Searchable result of the Indexer process
- Collection of JASON documents containing the values extracted during indexing

Attributes:
Key / searchable / filterable / sortable / facetable / retrievable

The Indexing process

Creates a document for each entity

Combining data from the data source with enriched fields extracted by AI

- document
  - Metadata_storage_name
  - Metadata_author
  - content

When the data contains images

You can configure the Indexer to extract the image data and place in a normalized collection

- document
  - metadata_storage_name
  - metadata_author
  - Content
  - Normalized_images
    - image 0
    - image 1

Each skill adds fields to the JSON

I.e., A skill that detects the language in which the document is written stores its output in a language field

- document
  - metadata_storage_name
  - metadata_author
  - Content
  - Normalized_images
    - image 0
    - image 1
  - language

Skills are applied hierarchically to a specific context

You could run an OCR for each image to extract any text they might contain

- document
  - metadata_storage_name
  - metadata_author
  - Content
  - Normalized_images
    - image 0
      - Text
    - image 1
      - Text
  - language

Output can be used as input

We could use a merge skill to combine The original text content with the text extracted from each image

- document
  - metadata_storage_name
  - metadata_author
  - Content
  - Normalized_images
    - image 0
      - Text
    - image 1
      - Text
  - Language
  - merged_content

Fields are mapped to index in one of two ways…

1. Fields extracted from the source data are all mapped to index field
   - Implicit mapping
     - Autom mapped to field with the same name
   - Explicit mapping
     - A map defines the better source/index match
2. Output fields from the skillset are mapped from their hierarchical location in the output to the target field in the index.

Search an index

An index could be queried based on a simple text matching, but most search solutions use **full text search** semantics to query an index.

Search solutions that parse text-based document to find query terms.
In Azure AI is based on the **Lucene query Syntax**

- **Simple**: Intuitive syntax that makes it easy to perform basic searches
- **Full**: Extended Syntax. Supports complex filtering, regular expressions and sophisticated queries.

Client applications submit queries to Azure AI search
Some common parameters include....

- search - Terms to be found
- queryType - The Lucene syntax to be evaluated (simple or full)
- searchFields - The index fields to be searched
- select - the fields to be included in the results
- searchMode - Criteria for including results based on multiple search items

For the search: Comfortable hotel
"Any" returns docs that contain **"comfortable", "hotel"**, or **both.**

"All" restricts results that contain both **"Comfortable"** and **"hotel"**

QUERY PROCESSING

Query parsing
Search expression is evaluated and reconstructed as a tree of subsequent subqueries.
term queries / phrase queries / prefix queries

Lexical analysis
The query terms are analyzed and refined based on linguistic rules
Text converted to lowercase / nonessential words removed / word converted to their root form(comfortable->comfort)/ composite words are split.

Document retrieval
The query terms are matched against the indexed terms. The set of matching documents is identified.

Scoring
A relevance score is assigned based on TF/IDF
(**T**erm **F**requency/ **I**nverse **D**ocument **F**requency)

FILTERING & SORTING (search query API)

EXAMPLE To find docs containing the text ***London*** that have an author field of ***Reviewer***

You can filter queries in 2 ways

- By including filter criteria in a simple search expression

  search=London+author='Reviewer'
  queryType=Simple

- You can use OData filter in a $filter parameter with A full Lucene expression

  search=London
  $filter=author eq 'Reviewer'
  queryType=Full

⚠ OData $filter expressions are case-sensitive

Filtering Criteria based on a field value in a result set

Filtering with facets

1- Specify facetable fields for which you want to retrieve the possible values in an initial query

Return all the possible values for the **author** field:
search=*
facet=author

2.- Results will include a collection of discrete facet values that you can display in the UI for the user to select

3- In a subsequent query, you can use the selected facet value to filter results

search=*
$filter=author eq 'selected-facet-value-here'

Sorting results

- By default, results are sorted based on the relevancy score assigned by the query process. (Highest scoring matches first)
- You can override this sort order by including an OData **orderby** parameter that specifies one or more **sortable** fields and a sort order (asc or desc)

EXAMPLE

To sort the results so that the most recently modified docs are listed first:

search=*
$orderby=last_mofified desc

Enhance the index

Basic index is alright but with Azure AI search you can enhance an index to provide a better user experience

Search-as-you-type

By adding a ***suggester*** you can enable 2 forms of search-as-you-type experience

- **Suggestions** -List of suggested results in the search box as the user types.
- **Autocomplete** - complete partially typed search terms based on values in index fields

Custom Scoring & result boosting

Customize the default TF/IDF scoring algorithm by creating a scoring profile that applies a weighting value to specific field
You can boost results based on field values (i.e. date of modification or size)

Synonyms

Synonym maps to help users find the information they need

United Kingdom
UK
Great Britain
GB

https://github.com/MicrosoftLearning/mslearn-knowledge-mining

Knowledge check

200 XP

3 minutes

1. You want to find information in Microsoft Word documents that are stored in an Azure Storage blob container. What should you do to ensure Azure AI Search can access the files? *

- ( ) Add a JSON file that defines an Azure AI Search index to the blob container
- ( ) Enable anonymous access for the blob container
- (•) In an Azure AI Services resource, and add a data source that references the container where the files are stored

✔ Correct. To search files in a blob container, you should create a data source

2. You're creating an index that includes a field named modified_date. You want to ensure that the modified_date field can be included in search results. Which attribute must you apply to the modified_date field in the index definition? *

- (•) `searchable`

✕ Incorrect. Making a field searchable means that it can be queried for search terms. It doesn't mean the field can be included in the results.

- ( ) `filterable`
- ( ) `retrievable`

✔ Correct. To enable a field to be included in the results, you must make it retrievable.

3. You created a data source and an index. What must you create to map the data values in the data source to the fields in the index? *

- A synonym map
- An indexer

✔ Correct. Use an indexer to map data to index fields.

- A suggester

4. You want to create a search solution that uses a built-in AI skill to determine the language in which each indexed document is written, and enrich the index with a field indicating the language. Which kind of Azure AI Search object must you create? *

- Synonym map
- Skillset

✔ Correct. A skillset enables you to define an enrichment pipeline composed of AI skills.

- Scoring Profile

5. You want your search solution to show results in descending order of the file_size field value. What is the simplest way to accomplish this goal? *

- (●) Create a scoring profile that boosts results based on the file_size field

  ✗ Incorrect. A scoring profile calculates a relevancy score based on factors like term-frequency. You can boost scores based on a field, such as file_size; but other factors are also considered in the overall score.

- ( ) Make the file_size field facetable, and include a facet parameter that specifies the file_size field in queries.

- ( ) Make the file_size field sortable, and include an orderby parameter that specifies the file_size field in queries.

  ✓ Correct. Making a field sortable enables you to apply an orderby parameter to sort results by that field.

6. You created a search solution. Users want to be able to enter a partial search expression and have the user interface automatically complete the input. What should you add to the index? *

- (●) A suggester

  ✓ Correct. A suggester makes it possible to implement autocomplete and suggestions.

- ( ) A synonym map.

- ( ) A scoring profile.

Custom Skills

For Azure AI Search

I.e., You could train a model on the synopsis on the back cover of books to automatically identify a books genre.

Store search data

- Container needs to be accessible
- Choose Container instead of private container
- Need a way to assign classifications for each document:
  - ○Language Studio
  - ○Manually
  - ○At the JSON file before creating the Language project
- Types of classification projects
  - ○ Single label
  - ○ Multi label

Create Language Studio project

- Create your Azure AI Language project using the Azure portal.
  - ○ Creating it from the Language studio is less flexible.
- Select custom text classification

Train model

- Need data to train it
- The model needs to see examples on how to map data to a class and have some examples to test.
- By default the model with use 80% of the data for training and 20% for testing.
- You can choose specific doc you might want to be used for blind test by label them for testing.

Create search index

Coming from custom skill set in AI Search

5 things the function app needs to know

1. Text to be classified
2. Endpoint for the model
3. Primary Key for the custom text class proj
4. Project name
5. Deployment name

Create Function App

- You can choose language
- Needs to be able to
- pass JSON to the custom classification endpoint
- to process the JSON response from the model.
- Return a structured JSON message back to a custom skillset in AI Search.

Found in Language Studio

Update cognitive search

3 changes in the Azure portal to enrich your search index

1. **Add a field to your index** to store the custom text classification enrichment
2. Add **custom skillset to call your function app** with the text to classify
3. Map the **response from the skillset into the index**

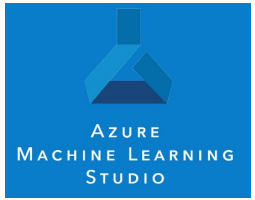
Create Machine Learning workspace

Train model

Edit scoring code

Create endpoint

Update cognitive search



- Has to be a Webs service endpoint
- Has to be an Azure Kubernetes Service (AKS)
  - AMLStudio can create and manage it for you
- Container instances not supported
-

https://github.com/MicrosoftLearning/mslearn-knowledge-mining

```
PS C:\Users\Student\RRAI\mslearn-knowledge-mining\Labfiles\02-search-skill> ./setup
Creating storage...
Uploading files...
Finished[#################################################################]  100.0000%
Creating search service...
(If this gets stuck at '- Running ..' for more than a couple minutes, press CTRL+C then select N)
-------------------------------------
Storage account: ai102str146303259
{
  "connectionString":
"DefaultEndpointsProtocol=https;EndpointSuffix=core.windows.net;AccountName=ai102str146303259;AccountKey=5a/GwP/LZh2RyXj98zclZdwrQtWOKJChOJsRtdK6disuYglLGjppejiZfWq03WMzqXZanpnwmdvp+AStGjyrfw==;BlobEndpoint=https://ai102str146303259.blob.core.windows.net/;FileEndpoint=https://ai102str146303259.file.core.windows.net/;QueueEndpoint=https://ai102str146303259.queue.core.windows.net/;TableEndpoint=https://ai102str146303259.table.core.windows.net/"
}
----
Search Service: ai102srch
Url: https://ai102srch146303259.search.windows.net
Admin Keys:
{
  "primaryKey": "sfDX2hvTr345jjhq7eLQSBkkGcEAkk9q5Kt6ccv5XoAzSeC20gIc",
  "secondaryKey": "5AvCO9awq81RfXmuV3FMLnK1KgZhBL34556zdoLK1AzSeCuo1B6"
}
 Query Keys:
[
  {
    "key": "rQvOvIfwM23haPg3fDzbW1223qVQSVZjTRZrlTIItwJAzSeD5m4Th",
    "name": null
  }
]
```

```
PS C:\Users\Student\RRAI\mslearn-knowledge-mining\Labfiles\02-search-skill\create-search> ./create-search
-----
Creating the data source...
-----
Creating the skillset...
-----
Creating the index...

Waiting for 0 seconds, press CTRL+C to quit ...
-----
Creating the indexer...
PS C:\Users\Student\RRAI\mslearn-knowledge-mi
```

Knowledge check

200 XP

Module assessment • 5 minutes

ⓘ Great job! You passed the module assessment. ×

1. You want to include a sentiment score for each document in an index. What should you do? *

- ○ Create a custom skill that uses an Azure Machine Learning model to predict the sentiment for a document
- ○ Create a custom skill that calls the Azure AI Language service and predicts the sentiment of each document.
- ◉ Add the built-in Sentiment skill to the skillset used by the indexer.

✔ **Correct. The built-in sentiment skill can be used to accomplish the goal in this scenario.**

2. You implemented a custom skill as an Azure function. You want to include the custom skill in your Azure AI Search indexing process. What should you do? *

- ◉ Add a WebApiSkill to a skillset, referencing the Azure function's URI

✔ **Correct. To integrate an Azure function custom skill into an indexing process, you must define a skillset containing a WebApiSkill with the URI for the function.**

- ○ Create a JSON document with the input schema for your function, and save it in the folder where the documents to be indexed are stored.
- ○ Submit each document to the function, and store the output in a separate data source. Then use the Merge skill to add the results to the index.

3. When you create an Azure AI Language project, if you let the model automatically split your training data, what percentage of the documents will it use to train the model, by default? *

- 20%
- 50%
- 80%

✔ Correct. If you let the model automatically split your training data, it uses 80% of the documents to train the model, by default.

4. When you create an Azure Machine Learning custom skill, what type of endpoint does the URI have to use? *

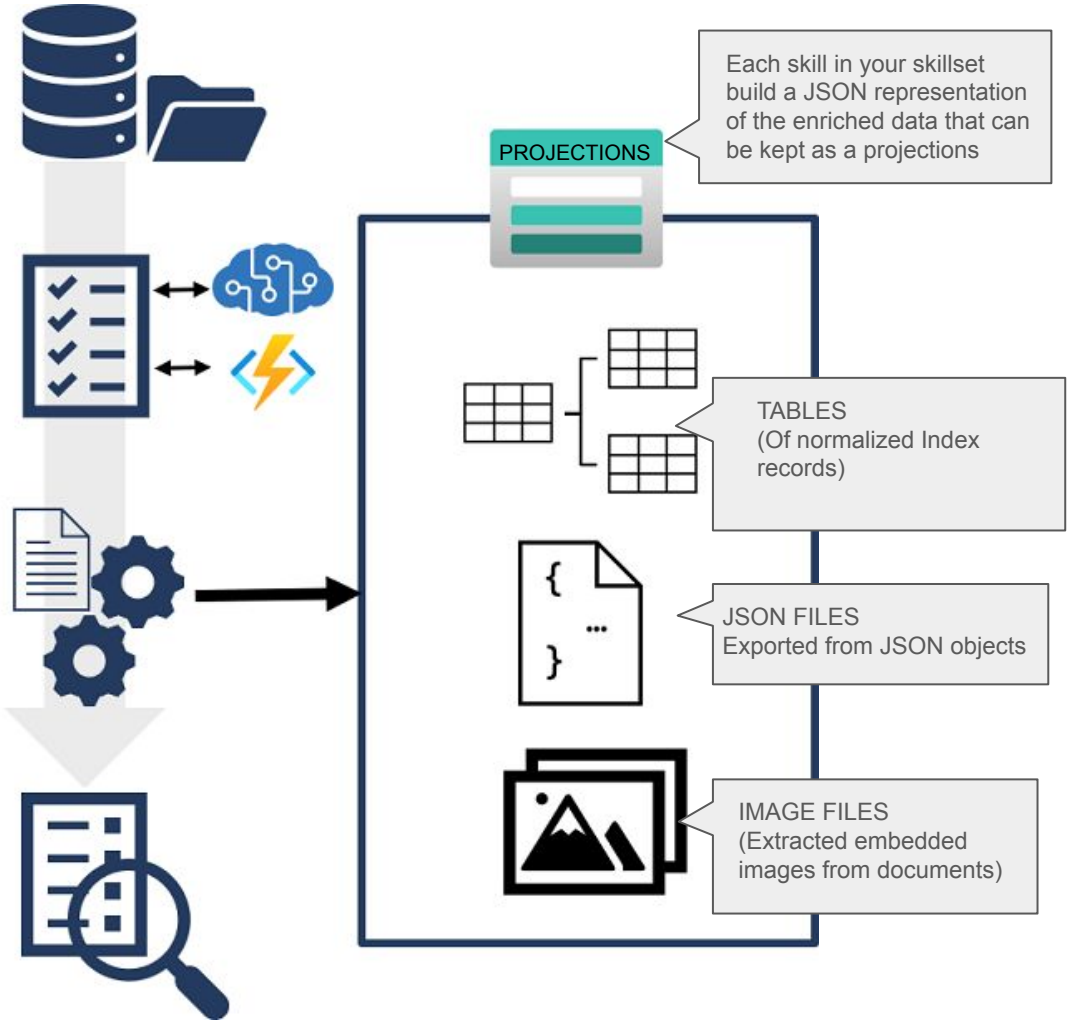
- The URI has to use an HTTPS endpoint

✔ Correct. The URI has to use an HTTPS endpoint.

- The URI has to use an HTTP endpoint
- The URI has to use an FTP endpoint

Knowledge Store

with Azure AI Search



Knowledge Store

Defined in the skillset that encapsulates your enrichment pipeline

GitHub

https://github.com/MicrosoftLearning/mslearn-knowledge-mining

Knowledge check

200 XP

Module assessment • 3 minutes

ⓘ Great job! You passed the module assessment. ×

1. You want to create a skillset that includes a knowledge store definition. Which type of skill should you use to map the enriched fields extracted by your skillset to the desired structure for the knowledge store data? *

- ○ Merge
- ◉ Shaper

✔ Correct. A shaper skill enables you to define a custom document structure for your enriched fields.

- ○ Split

2. You want to create a knowledge store that contains JSON representations of the indexed documents. What kind of projection should you define? *

- ◉ Object

✔ Correct. Object projections are JSON representations of an indexed document.

- ○ File
- ○ Table

3. You want to create a knowledge store that contains a relational schema for your enriched data. What kind of projection should you define? *

- ( ) Object
- ( ) File
- (•) Table

✔ Correct. Table projections define a relational schema of tables for your enriched data.

4. You want to create a knowledge store that contains the images extracted from your indexed documents. What kind of projection should you define? *

- ( ) Object
- (•) File

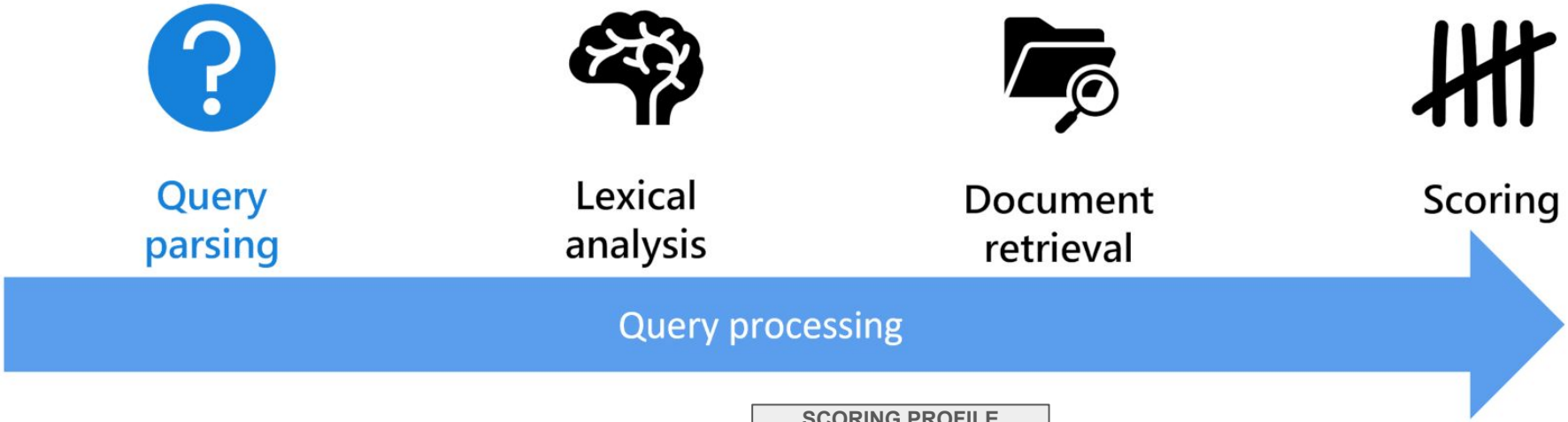
✔ Correct. File projections create a .jpg file for each image extracted from a document.

- ( ) Table

Advanced Search Features

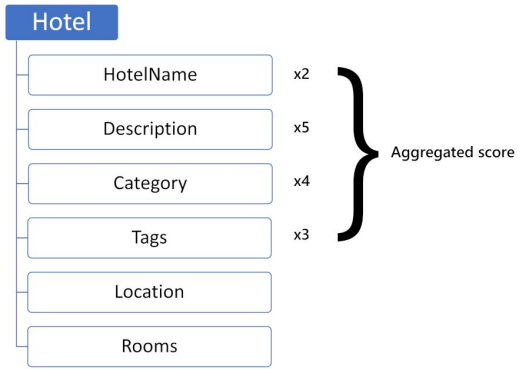
In Azure AI Search

I.e., Change ranking on documents, boost terms, and allow searching in multiple languages



SCORING PROFILE
(You can add up to **100** to a search index)

Can also include **functions**. I.e.,distance, freshness. **You can also define the boosting duration** applied to newer docs before they score the same as older docs.



The search engine scores the documents returned from the first three phases. The score is a function of the

- # of times identified search terms appear in a document
- Doc size
- Rarity of the term

By default search results are ordered by search score

Analyzers and tokenized terms

TOKENS

- Azure AI Search is configured to analyze text and identify tokens that will be helpful in your index
  - The right tokens ensure that user can find the docs quickly
  - In most cases , default config produces an optimal index
    - However, when you have unusual or unique fields you might want to configure exactly how text is analyzed.

To control how the content of a field is split into tokens for inclusion in the index we use a **CUSTOM ANALYZER**

What do Analyzers do?

- Break text into words using whitespace and punctuation as delimiters
- Remove Stopwords (i.e., “the”, “it:, etc.)
- Reduce words to their root form (i.e., ran -> run)

If no Analyzer is specified for a field, the default Lucene analyzer us used.
Alternatively there are pre-built analyzers into AI Search:

- Language analyzers
  - If you need advanced capabilities for specific languages (Analyzers for 50 lang avail)
    - Lemmatization — Changing/Changed/Change = Change
    - Decompounding — Wissenschaftskolleg (science college), the atoms are Wissenschaft (science) and Kolleg (college)
    - Entity Recognition
- Specialized analyzers (language-agnostic)
  - Zip codes
  - Product IDs
  - Separators like comma(,) using PatternAnalyzer

Sometimes you need an analyzer with an unusual behavior for a field

- Character Filters - Enables operations on the text before is divided into tokens — html_strip, mapping, pattern_replace
- Tokenizers - Divides the text into tokens to be stored in the index — A postal address, URL or email address, words based on the grammar of a specific language
- Token filters - Add or remove stuff to/from the Tokens — Arabic_normalization, apostrophe, removing english possessives and dots from acronyms, remove tokens that doesn’t include certain words, filter by length, white space removal,

TO CREATE SPECIFY IT WITH JSON WHEN DEFINING THE INDEX

Enhance an index to include multiple languages

Add new fields to the index

Add Translation skillsets

Map the translate output to the index in the indexer

```
{
  "name": "description_jp",
  "type": "Edm.String",
  "facetable": false,
  "filterable": false,
  "key": false,
  "retrievable": true,
  "searchable": true,
  "sortable": false,
  "analyzer": "ja.microsoft",
  "indexAnalyzer": null,
  "searchAnalyzer": null,
  "synonymMaps": [],
  "fields": []
},
```

```
skills": [
  {
    "@odata.type":
"#Microsoft.Skills.Text.TranslationSkill"
    "name": "#1",
    "description": null,
    "context": "/document/description",
    "defaultFromLanguageCode": "en",
    "defaultToLanguageCode": "ja",
    "suggestedFrom": "en",
    "inputs": [
      {
        "name": "text",
        "source": "/document/description"
      }
    ],
    "outputs": [
      {
        "name": "translatedText",
        "targetName": "description_jp"
      }
    ]
  }
```

```
"outputFieldMappings" : [
  {
    "sourceFieldName" : "/document/description/description_jp" ,
    "targetFieldName" : "description_jp"
  },
  {
    "sourceFieldName" : "/document/description/description_uk" ,
    "targetFieldName" : "description_uk"
  }
]
```

Results by distance from a given reference point

geo-spatial functions

Make sure your index includes location for result
With data type Edm.**GeographyPoint** and store **latitude** and **longitude**

- geo.distance - Straight line distance from a point to the location
  - search=(Description:luxury OR Category:luxury)$filter=geo.distance(location, geography'POINT(-122.131577 47.678581)') le 5&$select=HotelId, HotelName, Category, Tags, Description&$count=true
  - search=(Description:luxury OR Category:luxury)&**orderby**=geo.distance(Location, geography'POINT(2.294481 48.858370)') asc&$select=HotelId, HotelName, Category, Tags, Description&$count=true

When you sue geo.distance in a filter eq and ne are not supported.
Use lt, le, gt, or ge

Because it returns the distance in Km, you can use it in an orderby clause

- geo.intersects - location is inside a geofence (TRUE/FALSE)
  - search=(Description:luxury OR Category:luxury) AND geo.intersects(Location, geography'POLYGON((**2.32 48.91, 2.27 48.91, 2.27 48.60, 2.32 48.60, 2.32 48.91**))')&$select=HotelId, HotelName, Category, Tags, Description&$count=true

geo.intersects returns a boolean value so can't be used in an orderby clause

In Polygons, you must specify the points in counterclockwise order and the polygon must be closed, so the 1st and last points must be the same

Compares a location with a polygon you specify with 3 or more points

Knowledge check

200 XP

Module assessment • 3 minutes

ⓘ Great job! You passed the module assessment. ×

1. What character do you add after a search term boost the term? *

- ○ +.
- ◉ ^.

  ✔ Correct. ^ used in combination with a numerical value boosts a term.
- ○ !.

2. Which of the following options is a function you can use in a scoring profile? *

- ◉ Tag.

  ✔ Correct. You can alter scores based on common tag values.
- ○ Volume.
- ○ Staleness.

3. What Azure product can you use to enrich an index with different language translations? *

- ○ Azure AI Search.
- ○ Azure Speech Service.
- ◉ Azure AI Services.

  ✔ Correct. Azure AI Services provides translation services.

Azure Data Factory

Search data outside Azure AI search

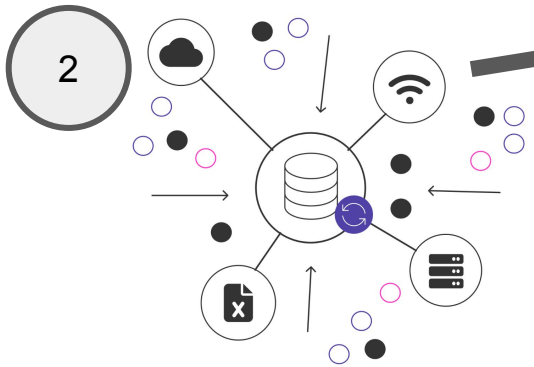
Two main ways to get data into a search index…..

1

Azure Data Factory

- Connections to nearly 100 datas stores (AKA Sinks)
- Connectors like HTTP and REST

2



AI Search PUSH API

INDEX

AI Search

Index data from external data sources using Azure Data Factory (ADF)

- Search API
  - Default port: 443
  - Must include api-version
  - Header must include api-key attribute
  - Use an HTTP POST request
  - If request successful API will return a 200 status code (Full list codes)
  - For better performance batch your uploads to a max of 1,000 dcs or 16MB in total size
  - Apply a code to find the best batch size
  - If your index starts to throttle due to overload (code 503 or 207) code a backoff strategy (pausing for some time before retrying your request)
  - Use threading to improve performance (implementing backoff strategy in threads ,for example the number of cores your processor has)

Feature	Operations
Index	Create, delete, update, and configure.
Document	Get, add, update, and delete.
Indexer	Configure data sources and scheduling on limited data sources.
Skillset	Get, create, delete, list, and update.
Synonym map	Get, create, delete, list, and update.

Knowledge check

200 XP

Module assessment • 3 minutes

ⓘ Answer 100% of questions correctly in order to pass. Retake ✕

1. What is the limitation of using the Azure Search linked service as a sink in a copy data task? *

- (●) You can only upload one document at a time.
  - ✕ Incorrect. You can upload multiple documents if they're defined in the source data.
- ( ) The JSON can't contain complex data types like arrays.
  - ✓ Correct. At the moment, the linked service only supports a limited number of field types.
- ( ) You have to define the index in the Azure portal first.

2. Which feature of the REST API would you use to upload documents into a search index? *

- (●) Index.
  - ✓ Correct. You use the index REST API focused on documents.
- ( ) Indexer.
- ( ) Skillset.

3. Which response code will require you to implement a backoff strategy? *

- ( ) 200 and 201.
- ( ) 404 and 501.
- (●) 207 and 503.
  - ✓ Correct. 503 is the response means the system is under heavy load and your request can't be processed at this time. 207 means that some documents succeeded, but at least one of them failed.

Maintain an Azure AI Search solution

Performance, Cost, Reliability

Security approaches

Data in transit

HTTP TLS 1.3 encryption over port 446

You can use your own encryption keys by using the Azure Key Vault. Double encryption will be enabled on all objects you use your custom keys on

Azure AI Search encrypts the data it stores at REST with service managed keys.

SKILL SETS

DATA SOURCES

INDEXER DEFINITIONS

INDEXES

SYNONYM MAPS

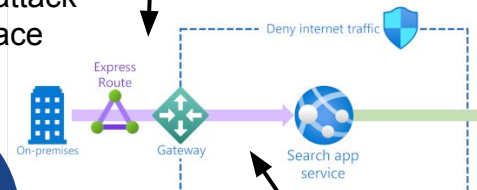
Inbound search requests

Outbound Requests

Secure data at the document-level

Reducing the attack surface

If your service is going to be used by on-premise resources only



If accessed externally from the internet or apps.

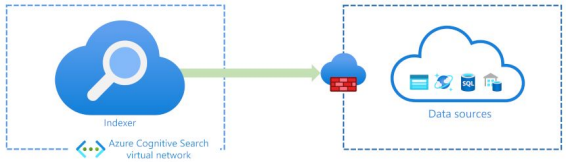
Also endpoint can be switched to private using a link. Very secure but more expensive

Authenticate requests to your search solution…..

- Key-based authentication
  - Admin Keys
    - Write and Query permissions (2 per service)
  - Query Keys
    - Read permissions. Used by users or apps (max 50)

Role-based access control (BETA)

- Built-in
  - Owner, Contributor, Reader, Search service Contributor, Search Index Data Contributor, Search index Data Reader



The outbound connection supports:

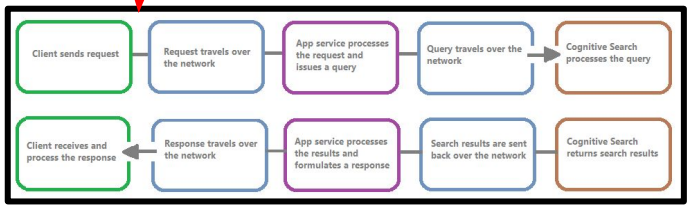
- Key-based authentication
- Database logins
- Microsoft Entra logins
- System or user-assigned managed Identity
- Firewall that only allows your service
  - If enriching your indexes with AI, need to allow all IP addresses in the AzureCognitiveSearch service tag

- Restrict documents someone can search. (i.e., restrict searching contractual docs to people in your legal department)
- Add a security field to the docs that contains the user or group ID that can access it
- Add the searh.in filter to all search queries so it returns results only if ID allowed.

Optimize the performance of an Azure AI Search Solution

1. Measure your current search performance
2. Check if your search service is throttled
3. Check performance of individual queries
4. Optimize your index size and schema
5. Improve the performance of your queries
6. Use best service tier for your search needs

- Enable Diagnostic logging : Monitoring/Diagnostic settings/+Add diagnostic setting
- It is important to capture this diagnostic info at the search service level. Your en-users or apps might be getting performance issues in any stage of the process



Throttling: limiting the bandwidth for some resources.

- Searches & indexes can be throttled
- It will appear in Log Analytics as an **503** HTTP response for the searches, and a **207** HTTP for the indexes.
- Monitoring/Logs
- Run query

```
AzureDiagnostics
| where TimeGenerated > ago(7d)
| summarize count() by resultSignature_d
| render barchart
```

Use a client tool like Postman.
Azure will always return an 'elapsed-time' value for how long it took to the service to complete the query
To calculate the time it took to send and then receive the response: Total round trip (Time) - elapsed-time

The smaller and more optimized your indexes, the fast Azure AI Search can respond to queries.

- Review that all the documents in your index are still relevant.
- Can you reduce the complexity of the schema?
- Need to be searchable, facetable, filterable?
- Do you need all the same skillsets?
- Support for filters, facets and sorting = x4 storage needs.

Your Search works? You can tune your queries to drastically improve performance

1. Only specify the fields you need to search using the SearchFields parameter.More fields require more processing
2. Return only the fields you need to render on your results.
3. Avoid partial search terms like prefix terms or regular expressions (more computationally expensive)
4. Avoid using high skip values
5. Limit using facetable and filterable fields to low cardinality data (low unique values)
6. Use search functions instead of individual values in filters
   Use: search.in(userid,'123,143,563,121',',') instead of $filter=userid eq 123 or userid eq143, or.....

Ways to scale out

- Add partitions
- Add replicas
- Using a higher tier

You need to know the approximate total size of storage you're going to need.
Largest now is 12 partitions, 24 TB

Search Units (SU) = Replicas * Partitions

Manage Costs of an Azure AI Search solution

ESTIMATE

Your search solutions baseline costs

- Azure AI Search pricing calculator
  i.e.: An S2 tier search sol, using 4 SU, extracting 80K images, and using 200K semantic queries:

  S2 tier 4SU = $981.12*4 = $3,924.48
  Cracking Images = $1*80=$80
  Semantic Search = $500 (up to 250K searches plan) (wouldn't the $1 per 1,000 be better?)

  $4,504.48 per month
  + Data ingestion + Processing

UNDERSTAND

The billing model

- Hourly cost: $3924.48÷744 = 5.27 per hour approx.
- The other premium features are billed per use
  - Indexer usage (per 1000 API calls)
  - Image extraction (per 1000 records)
  - Built-in skills (# transactions. 20 per indexer per day for free)
  - Custom Entity Lookup skill (per 1000 text records)
  - Semantic Search (# of queries)
  - Private Endpoints (Per endpoint bandwidth)
- $0 for # of search queries, responses or docs ingested

TIPS

To reduce the costs of your search sol.

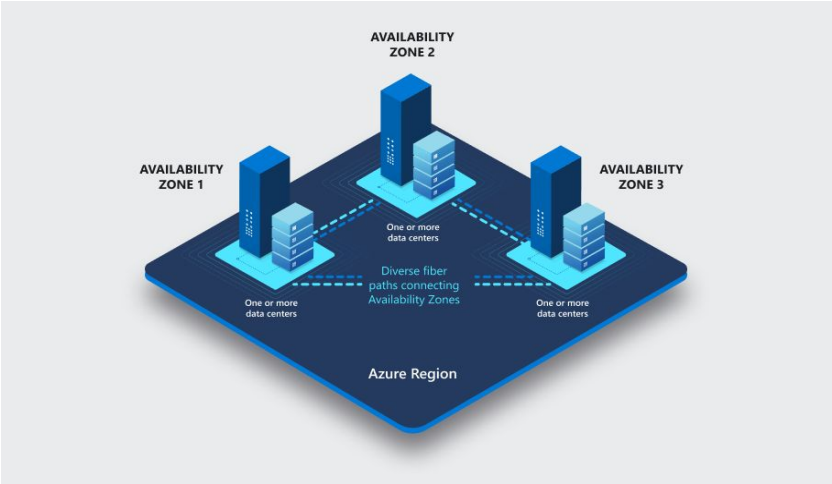
1. **MINIMIZE BANDWIDTH COSTS** by using as few regions as possible. Ideally all resources in the same region
2. **SCALE UP FOR INDEXING AND BACK FOR YOUR REGULAR QUERYING** if your indexing of new data happens in predictable patterns.
3. **KEEP YOUR SEARCH REQUESTS INSIDE THE AZURE DATACENTER BOUNDARY** by using an Azure Web App Front-end as your search app.
4. **ENABLE ENRICHMENT CACHING**, if you're using AI enrichment on blob storage.

MANAGE

Search service costs using budgets and alerts

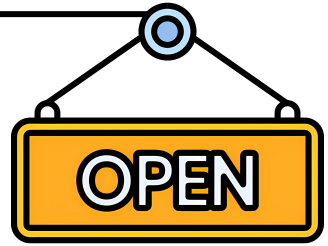
- Monitor how much you're spending, and take action if the costs have increased over your budget : Tutorial Here
- Home/Cost Management
- Create budgets and alerts

Improve reliability of an Azure AI Search solution



Make your solution highly available

- ★ Increase the number of replicas
  - ○ Availability guarantees based on # of replicas
    - ■ 2 replicas = 99% availability for queries
    - ■ 3+ = 99.9% availability for queries & indexing
- ★ Add more redundancy with Availability Zones
  - ○ Requires (at least) a standard tier
  - ○ When you add replicas, you can choose to host them in different Availability Zones
    - ■ Benefit: They're physically located in different data centers



Distribute your search solution globally

- ★ Most cost-efficient way to architect = Single resource group and region

However…. If availability and performance are important..

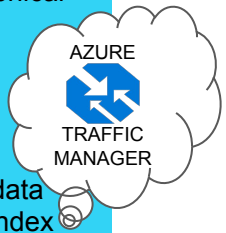
- ➔ Host multiple versions of your SS in different geographical regions

Pros:

- ◆ Protection against failure in a region
- ◆ Improved response time if your users are global

Cons:

- ◆ You'll have to make sure all indexers have same data
- ◆ Route requests to the fastest responding search index



At present (Dec 2024)
Azure doesn't offer a formal backup and restore mechanism for Azure AI Search.
You'll have to built your own tools to **backup index definitions as series of JSON files** to re-create your search indexes using those files in case of a data loss.

Monitor an Azure AI Search solution

Monitor

Default in Overview/Monitoring: Search latency + Queries per second + % Throttled queries

Overview/Usage: What resources your search solution is using

You can go further with some more configuration. Azure monitor can be used to monitor all Azure resources

Once you have started using Log Analytics:

- AzureActivity - Shows you tasks that have been executed (i.e., scaling the search service)
- AzureDiagnostics - All the query and indexing operations
- AzureMetrics - Data used for metrics that measure the health and performance of your search service

Metrics

Using charts is a powerful way to view how your search service is performing

Monitoring/Metrics

- DocumentsProcessedCount
- SearchLatency
- Search QueriesPerSecond
- SkillExecutionCount
- ThrottledSearchQueriesPercentage

EXAMPLE

Plot search latency against % Throttled queries to see its effect

Kusto

Monitoring/Logs

Log Analytics allows you to write any Kusto query against captured log data

I.e.: List of recent operations and how many times they happened:

```
AzureDiagnostics
| summarize count() by OperationName
```

Alerts

- Search Latency - What latency triggers the laert (in seconds)
- Throttled search percentage
- Delete Search Service - Be notified if your search service is deleted
- Stop Search Service - Be notified if your SS is stopped (i.e., scale up/down or needs restart)

Debug Search issues using Azure portal

When you first create your search service, **you need to make some assumptions about the data you're indexing**. You make choices about the index and how to index that data. **Until you run your created indexer** you can't be certain that you made all the right choices

Azure Debug Session Tool

- Lets you step through the enrichment pipeline of a document while being enriched.
- You can make changes and fixes to each individual skill and rerun the indexer in real-time
- Once fixed, you can republish the indexer and rerun it

Debug a skillset with Debug Sessions

Create a Debug Session

Overview / Search Management / Debug Sessions / +Add
Storage connection string = General-purpose account for caching the debug session
Indexer template = Skillset that drives the indexer you want to debug

Explore and edit a skill

1. In the graph, select a skill
2. Executions tab / OUTPUTS / select "/" to open Expression evaluator
3. Skill Settings tab / Edit the JSON of the skill. Save
4. To test: Select RUN and COMMIT if OK

Validate field mappings

1. Select **Skill Graph** and check that **Dependency graph** is selected
2. Field Mappings
3. Make changes to where data should be mapped
4. Save
5. Select Output Field Mappings
6. Fix Field Mappings
7. Save
8. **Commit changes** if OK

Ask Learn

Knowledge check

200 XP

Module assessment • 3 minutes

ⓘ Great job! You passed the module assessment.

1. An organization wants to improve the reliability of a search service. It's important that both read and write operations are 99.9% available. Which of these architectures would ensure this reliability? *

- ○ Create an Azure AI Search service with a Storage Optimized service tier and at least two replicas.
- ◉ Create an Azure AI Search service with any Standard service tier and at least three replicas.

✔ Correct. The main factor is that the search service has three replicas.

- ○ Create an Azure AI Search service with a High-density service tier and one replica.

2. After an Azure AI Search service has been created, which three metrics can be viewed in graphs without any other configuration? *

- [x] Search latency, queries per second, and the percentage of throttled queries.

✔ Correct. These three metrics are graphed on the overview pane.

- [ ] Count of documents processed, count of skills executed, and the search latency.
- [ ] Number of errors per indexer, number of warnings per indexer, and the total number of documents indexed.

3. Which of the following option is the best way to manage your search service costs? *

- [ ] Enable enrichment caching if you're using AI enrichment on blob storage.
- [ ] Keep your search requests and responses inside the Azure datacenter boundary.
- [x] Monitor and set budget alerts for all your search resources.

✔ Correct. This option is the most effective way to manage your costs.

Next unit: Summary

‹ Previous Next ›

Search reranking with Semantic Ranking

Semantic Ranking

What is Semantic Ranking?

Capability of **improving ranking search** results by **using language understanding** to more accurately **match the context** of the original query

BM25

Default **ranking function** that ranks search result based on the **frequency** that the search term appears within a document. No relevance is placed to semantics

This works great for some searches like vehicle parts codes, but not so great for searches like "What's the capital of France?" where semantics are important.
Semantic Ranking = BM25 + Understanding Language Models

Semantic Captions and Answers

Semantic Captions: Extract summary sentences from the document (Verbatim) and highlights the most relevant text in the summary sentences

Semantic Answers: (optional feature) Provides answers to questions. If the search quarry appears to be a questions and the results contains texts that appears to be a relevant answer then the semantic answer is returned.

How semantic ranking woks?

1. Takes the top 50 results from BM25
2. Split results into multiple fields
3. Convert fields into text strings and trimmed to 256 unique tokens (~1word)
4. Pass tokens trough machine reading comprehension to find phrases and sentences that best match the quarry
5. Return a semantic **Caption** or **Answer.**
6. Semantic captions are ranked by relevance

Advantages

- Can rank results to more closely match the semantics of the original query.
  - More likely the most useful doc will appear on top
- Can find strings within the results to
  - Render as a caption on the search results page
  - Provide an answer to a question

Limitations

- It is applied to BM25 results so it wont provide any additional documents that weren't returned by it.
- Considers **ONLY** the **top 50** results from the BM25

Pricing

- Up to 1000 semantic ranking queries/month -> Free of charge
- > 1,000 queries a month = Standard Pricing
  https://azure.microsoft.com/en-us/pricing/details/search/

Setting up Semantic Ranking

- Semantic Ranking not available in every region. Check here

To choose the semantic ranking plan…

1. Select your search service
2. Navigation pane / Settings / Semantic ranker
3. Select the appropriate service plan

To configure semantic ranking….

1. Select your search service
2. Navigation bar / Search management / Indexes
3. Select your Index
4. Semantic configurations / Add semantic configuration
5. Name
6. Title = Select the field that describes the document
7. Content fields / Field name / Select a content field
8. Repeat 6-7 for additional content fields
9. Keyword fields / Field name / Select field with keyphrases
10. Repeat # 9 for additional keyword fields
11. Save
12. On index page, Save

Knowledge check

200 XP

Module assessment • 3 minutes

Great job! You passed the module assessment.

1. How many results are returned by semantic ranking? *

- (•) Up to 50.
  - ✓ Correct. Semantic ranking returns 50 results, or as many results as the BM25 ranking function, whichever is lower.
- ( ) As many results as the BM25 ranking function returns.
- ( ) Up to 25.

2. Which services is a prerequisite for semantic ranking? *

- (•) Azure AI Search service with a billable tier.
  - ✓ Correct. Azure AI Search service with a billable tier is required for semantic ranking.
- ( ) Azure AI services with a billable tier.
- ( ) Azure AI Language service.

3. What are semantic captions? *

- (•) Verbatim summary sentences from the document.
  - ✓ Semantic captions extract summary sentences from the document verbatim and highlight the most relevant text in the summary sentences.
- ( ) A summary of the content from the highest ranked document.
- ( ) A summary of the content from all documents.

Next unit: Summary

Previous Next

Vector Search and Retrieval

Vector Search

- Type of data representation that is used by machine learning models.
- Semantic meaning of a piece of text
- Can be visualized as an array of numbers
  - The numerical distance between two embeddings represents their semantic similarity
- Embedding models: Similarity search Text search, and Code Search embeddings
- Embedding space: Core of vector queries comprising all the vectors fields from the same embedding model. (Abstract space. Not very comprehensible by humans). i.e.,[2,6,4,5], [-2,-1,0,1]

What it it?

Capability to index, store and retrieve vector embedding from a search index.

Can be used to **match criteria** across **different** types of source data by providing a **mathematical representation of the content** generated by MLM.

When to use it?

- Use OpenAI to encode text, and use queries encoded as vectors to retrieve documents
- Similarity search across encoded [illegible]es, text, video and audio, or a mixture (multi-modal).
- Hybrid searches from vector and searchable text fields as vector searches are implemented at field level.
- Apply filters to text and numeric fields and include this in your query to reduce the data your vector search needs to process
- Create a vector database to provide an external knowledge base or use as a long term memory.

Limitations

- Need to provide the embeddings using Azure OpenAI or a similar open source solution, as Azure AI Search doesn;t provide these for your content.
- Customer Managed Keys (CMK) are not supported
- There are storage limitations applicable so you should check what your service quota provides

If your documents are large , consider **chunking**

PREPARE YOUR SEARCH

- Encode your query by sending it to an embedded model
- The response is then passed to a search engine to complete a search over the vector fields

IN ORDER FOR THE QUERY TO WORK, YOU NEED TO DO THE FOLLOWING TASKS

1

CONVERT QUERY INPUT INTO A VECTOR

- Check if your search has vector by running an empty search, the result includes a vector field with a number array, or….
- Look for a field named vectorSearch (type Collection(Edm.single)

2

CHECK YOUR INDEX HAS VECTOR FIELDS

- You can only quarry a vector field with a query vector
- En users provide a query string which your app converts into a vector by using the embedding library you used to create the source document embeddings.

Knowledge check

200 XP

Module assessment • 3 minutes

ⓘ Great job! You passed the module assessment. ×

1. When would you use a vector search? *

- ○ To create a search to match text input.
- ◉ When you need to find matches across different types of data from a search index.
  - ✔ Correct. A vector query can be used to match criteria across text, video, image, and audio data sources.
- ○ To upload and index a document library.

2. What do you need to run a successful vector query? *

- ◉ Your search service URL and an admin key
  - ✔ Correct. These are inserted into the header information of your query.
- ○ Your Storage account name and location.
- ○ Your Azure subscription ID.

3. What type of vector search would you use to capture semantic similarity? *

- ○ A Text search.
- ◉ A Similarity search.
  - ✔ Correct. Use Similarity search embeddings to capture the semantic similarity between pieces of text.
- ○ A Code search

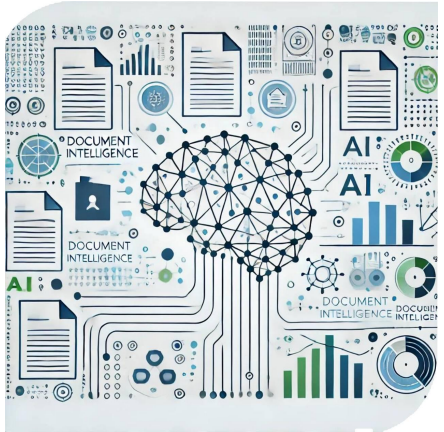
Next unit: Summary

< Previous | Next >

Document Intelligence Solution

Form Recognizer

Azure AI Document Intelligence



What it it?

Azure service that you can use to analyze forms (Either hand-written or digital) completed by people and extract the data they contain.

This avoids manual input costs and errors.

Similar to AI vision OCR

- Use **AI Vision OCR** for
  - Extract **simple words and text** from a picture without contextual information.
  - You have your own analysis code.
- Use **AI document intelligence** for **more sophisticated analysis** of documents. I.e., identify key/value pairs, tables and context-specific fields

Use a model or build one

Pre built:

- General Document Analysis
  - Read
  - General document
  - Layout
- Document type-specific
  - Invoice
  - Receipt
  - W2 US tax declaration
  - ID Document
  - Business card
  - Health insurance card

If you want to extract more specific information, you can create and train a custom model

You can also associate multiple custom models, trained on different types of models into a single model known as **Composed Model.**

Responsible use of AI

- **Fairness**
  Should treat people fairly regardless of race, belief, gender, sexuality, etc.
- **Reliability and safety**
  Reliable answers with quantifiable confidence levels
- **Privacy and security**
  All AI systems should secure and protect sensitive data and operate within applicable data protection laws
- **Inclusiveness**
  Available for all no regardless of their abilities
- **Transparency**
  All AI systems should operate understandably and openly
- **Accountability**
  All AI systems should be run by people who are accountable for the actions of those systems

How to use it

Visual tool:
Azure AI Document Intelligence Studio

To integrate into your own apps:
APIs:

- C#/.NET
- Java
- Python
- JavaScript

To use other languages call Azure AI Document Intelligence by using its RESTful web service

Prebuilt Models Available

General document analysis model

Use it to analyze docs with unpredictable structures

- Read
  - Extract words and lines from printed & handwritten docs
  - Also detects the language
  - You can use the pages parameter to fix a page range for the analysis
- General document

  Can extract values from structured, semi-structured, and unstructured documents
  - Read + extract key-value pairs, **entities**, selection marks, and tables

    Only model that supports entity extraction. A text might return both a key-value and an entity
- Layout
  - Extract text, tables, and structure information from forms
    Each table cell extracted with content, bounding box, Header yes/no, and row/col
  - Can recognize selection marks . (Checkboxes and radio buttons)

FEATURES OF PRE BUILT MODELS

- Text Extraction
  - Handwritten or printed text
- Key-value pairs
  - I.e., Weight (Key), 31Kg (value)
- Entities
  - People,locations, dates.
- Selection marks
  - Radio buttons ,Checkboxes
- Tables
- Fields
  - i.e., CustomerName and Invoicetotal fields in the invoice model.

INPUT REQUIREMENTS

- **JPEG**, **PNG**, **BMP**, **TIFF**, or **PDF** format
- **Microsoft Office** files accepted by the Read model
- File must be smaller than 500 MB (standar) and 4MB (Free)
- Images between 50 x 50 px and 10,000 x 10,000 px
- PDF dimension less than 17 x 17 inches or A3
- PDF must NOT be protected by password
- Text-embedded PDF are preferable for better text recognition
  - Only first 200 pages and first 2 on free tier

Specific document type models

- Invoice (English and Spanish)
- Receipt (printed and handwritten)
- W-2 (Extract Data from US W-2 tax declaration form
- ID document (US driver's license and int passports)

  Only biological pages (not visas)
- Business Card
- Check Model
- Health insurance card model
- Marriage certificate
- Credit/Debit card model
- Mortgage documents
- Banks statement model
- Pay stub Model

More models being released regularly. Before training a custom model check if your use case can be analyzed accurately with an existing pre built models.

Custom Models

- To train a custom model…
  - Supply at least 5 examples (The more examples, the more confidence)
  - The more varied your documents are in structure and terminology, the greater the number of example document you will need to supply to train a reliable model.
  - You can supply a labeled dataset or let the model identify key-value pairs and table data based on what it finds.
  - Make sure your training forms include examples that spans the full range of possible input (i.e., written/printed entries).

- There are 2 types of custom model
  - Custom template models
    - Most adequate for forms with consistent visual templates
    - Support 9 different languages for handwritten text and more for printed
    - If you have a few variations of the templates train a model for each and then compose.
  - Custom neural models
    - Works with structured and unstructured documents
    - Works on english with the highest accuracy and marginal drop for Ger,Fr,It,Sp and Dutch.

Composed Models

If you're using the Standard pricing tier, you can add up to 100 custom models into a single composed model. With the free, only 5.

- Consists of multiple custom models
- Typical scenarios: When you don't know the submited document and want to classify and then analyze it, or when you have multiple variations of a form and train a model for each. Doc Int will choose the one that better fits.

Knowledge check

200 XP

Module assessment • 10 minutes

ⓘ Great job! You passed the module assessment. ×

Choose the best response for each of the questions below.

Check your knowledge

1. You want to create an Azure AI Document Intelligence model where the documents are in one of three formats: wills, probate declarations, and affidavits. Each has their own specific layout. What type of model should you use that will understand the format of the three document categories? *

- ○ A Read model.
- ○ A Layout model.
- ◉ A Composed model.

✓ Correct. A Composed model consists of multiple custom models. Each submitted form is categorized as one of the custom form types and analyzed using the corresponding custom model.

2. You have developed a custom model that analyzes health assessment forms returned by patients to a medical practice. You've observed too much inaccuracy in the values that the model extracts for each field. What should you do to address this problem? *

- ◉ Retrain the model with a larger number of example forms.

✓ Correct. The larger the number of example forms you use to train a model, the more accurate it will be and the higher the confidence levels will be.

- ○ Change from a custom model to the general document model.
- ○ Change from the free tier to the standard tier.

3. You want to call your Azure AI Document Intelligence solution from a mobile app by using an API. Which of the following programming languages is natively supported as an Azure AI Document Intelligence SDK? *

- [x] Python
  - ✔ Correct. Microsoft publishes a Python API you can use to call Azure AI Document Intelligence services.
- [ ] Go
- [ ] R

4. Which of the following is an Azure AI Document Intelligence prebuilt model? *

- [ ] Employment record
- [ ] Resume
- [x] Receipt
  - ✔ Correct. The receipt model can identify commonly used fields and their values in scanned or photographed receipt documents.

Next unit: Summary

< Previous Next >

You work for a company that conducts polls for private companies and political parties. Participants submit their responses as paper forms or as online PDFs. You currently spend a lot of time and money entering these responses into databases. You want to assess Azure AI Document Intelligence to find out if you can use it to streamline this process

Knowledge check

200 XP

Module assessment • 3 minutes

ⓘ Great job! You passed the module assessment. ×

Choose the best response for each of the questions below.

Check your knowledge

1. You have a large set of documents with varying structures that contain customer name and address information. You want to extract entities for each customer. Which prebuilt model should you use? *

- ○ Read model.
- ◉ General document model.
  - ✓ Correct. The general document model is the only one that supports entity extraction.
- ○ ID document model.

2. You are using the prebuilt layout model to analyze a document with many checkboxes. You want to find out whether each box is checked or empty. What object should you use in the returned JSON code? *

- ◉ Selection marks.
  - ✓ Correct. Selection marks record checkboxes and radio buttons and include whether they're selected or not.
- ○ Bounding boxes.
- ○ Confidence indicators.

3. You submit a Word document to the Azure AI Document Intelligence general document model for analysis but you receive an error. The file is A4 size, contains 1 MB of data, and is not password-protected. How should you resolve the error? *

- ○ Change from the free tier to the standard tier.
- ○ Submit the document one page at a time.
- ◉ Convert the document to PDF format.
  - ✓ Correct. Word documents are not supported by Azure AI Document Intelligence but PDF documents are supported. Azure AI Document Intelligence is designed to analyze scanned and photographed paper documents, not documents that are already in a digital format so you should consider using another technology to extract the data in Word documents.

Next unit: Summary

< Previous Next >

3. You submit a Word document to the Azure AI Document Intelligence general document model for analysis but you receive an error. The file is A4 size, contains 1 MB of data, and is not password-protected. How should you resolve the error? *

- ○ Change from the free tier to the standard tier.
- ○ Submit the document one page at a time.
- ◉ Convert the document to PDF format.
  - ✓ Correct. Word documents are not supported by Azure AI Document Intelligence but PDF documents are supported. Azure AI Document Intelligence is designed to analyze scanned and photographed paper documents, not documents that are already in a digital format so you should consider using another technology to extract the data in Word documents.

Input requirements

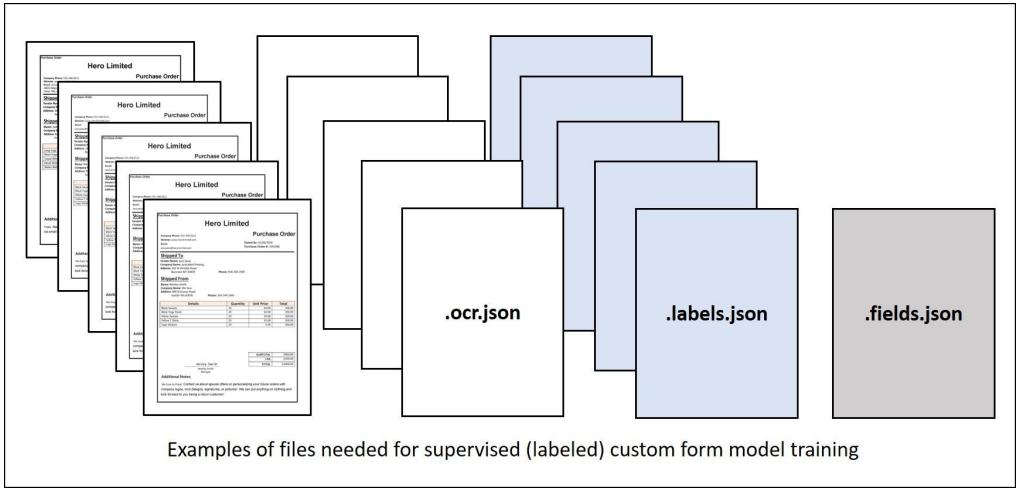
The prebuilt models are flexible but you can help them to return accurate and helpful results by submitting one clear photo or high-quality scan for each document.

You must also comply with these requirements when you submit a form for analysis:

- The file must be in JPEG, PNG, BMP, TIFF, or PDF format. Additionally, the Read model can accept Microsoft Office files.
- The file must be smaller than 500 MB for the standard tier, and 4 MB for the free tier.
- Images must have dimensions between 50 x 50 pixels and 10,000 x 10,000 pixels.
- PDF documents must have dimensions less than 17 x 17 inches or A3 paper size.
- PDF documents must not be protected with a password.

Document Intelligence Solution

Train Custom Models



Examples of files needed for supervised (labeled) custom form model training

To train a custom model

1. **Store sample forms in a blob container**
   - Along with JSON files (ocr.json, labels.json, fields.json)
   - You can generate OCR.json with the Azure Document Intelligence's Analyze Document function
2. **Generate Shared Access Security (SAS) URL for the container**
3. **Use Build model REST API function (or equivalent SDK)**
4. **Use Get model REST API function (or equivalent SDK) to get the trained model ID**

or…

Use the Azure Document Intelligence Studio to label and train.

Two types of models

- **Custom Template Models**
  - Labeled key-value pairs, selection marks, tables, regions and signatures
  - More than 100 languages supported
- **Custom Neural models**
  - Labeled Fields
  - Best for semi-structured or unstructured docs.

Knowledge check

200 XP

Module assessment • 3 minutes

ⓘ Great job! You passed the module assessment. ×

In this module, you have learned how to use the Azure Document Intelligence service to extract data from forms.

Consider the following review questions to check your understanding of the topics discussed in this module.

1. A person plans to use an Azure Document Intelligence prebuilt invoice model. To extract document data using the model and REST API language, what are two calls they need to make to the API? *

- ○ Train Model and Get Model Labels
- ◉ Analyze Invoice and Get Analyze Invoice Result

  ✓ Correct: The Analyze Invoice function starts the form analysis and returns a result ID, which they can pass in a subsequent call to the Get Analyze Invoice Result function to retrieve the results.
- ○ Create Azure Document Intelligence and Get Analyze Invoice Result

2. A person needs to build an application that submits expense claims and extracts the merchant, date, and total from scanned receipts. What's the best way to build the application? *

- ○ Use the Read API of the Computer Vision service.
- ◉ Use Azure Document Intelligence's prebuilt receipts model

  ✓ Correct: Use the Azure Document Intelligence's prebuilt receipts model. It can intelligently extract the required fields even if the scanned receipts have different names in them.
- ○ Use Azure Document Intelligence's Layout service

3. A person is building a custom model with Azure Document Intelligence services. What is required to train a model? *

- ◉ Along with the form to analyze, JSON files need to be provided.

  ✓ Correct: The labels needed in training are referenced in the ocr.json files, labels.json files, and single fields.json file.
- ○ Training must be done through language-specific SDKs.
- ○ Nothing else is required.

Compossed Models

- **A composed model consist in multiple custom models.**
-
  When you submit a form for analysis, Azure AI categorizes it and selectes the best custom model for the analysis

Max number of custom models per tier

TYPE	FREE	STANDARD
Custom	500	5000
Custom Neural	100	500
Composed	5	200

HOW?

1. Assemble your set of custom models into a composed model
   a. Either in Azure AI Document Intelligentce Studio or by using the method StartCreateComposedModelAsync() in your code.

2. Submit your form for analysis as you usually do
   a. Don;t forget to specify the model ID of the composed model

COMPATIBILITY

- Custom template models are composable with other custom template models across 3.0 and 2.1 API versions
- Custom Neural are composable with other Custom Neural
- Custom Neural can’t be composed with custom template

Great job! You passed the module assessment.

Choose the best response for each of the questions below.

Check your knowledge

1. You have a composed model that consists of three custom models. You're writing code that sends forms to the composed model and you need to check which of the custom models was used to analyze each form. Which property should you use from the returned JSON? *

- ○ modelId.
- ○ status.
- ◉ docType.

✓ Correct. The docType property includes the model ID of the custom model that was used to analyze the document.

2. You're trying to create a composed model but you're receiving an error. Which of the following should you check? *

- ◉ That the custom models were trained with labels.

✓ Correct. Only custom models that have been trained with labeled example forms can be added to a composed model.

- ○ That the custom models all have the same model ID.
- ○ That the custom models all have the same list of fields.

Next unit: Summary

< Previous | Next >

Develop GenAI Solutions

with Azure OpenAI Service

Azure OpenAI Service

- Azure AI Foundry : https://ai.azure.com/
  - Model Management
  - Deployment
  - Experimentation
  - Customization
  - Learning resources
- When Deploying Models
  - Select base models which Tokens Per Minute (TPM) are within The deployment's quota
  -
- Prompt engineering
  - Process of designing and optimizing prompts to better utilize AI models
  - Prompts must be: Relevant, Specific, Unambiguous, and Well structured.
  - Gen AI models have a ton of parameters and the logic it follows is unknown to users.
  - Some Prompt engineering methods
    - Providing clear instructions, Contextual Context, Cues or few-shot examples, and correctly ordering content in your prompt.

Azure OpenAI Service

- Adjusting Params
  - temperature and top_probability are the most likely to impact a model's response as they both control randomness in the model, but in different ways.

  - Higher values ….
    - produce more creative and random responses but less consistent and focused
    - Responses expected to be fictional or unique (vs consistent and concrete)
    - A high temperature allows more variation in sentence structure
    - A high top_p allows for more variation in words used (using a variety of Synonyms).
  - It's recommended to change either temperature or top_p at a time but not both.

Azure OpenAI Service

- Write Effective Prompts

Try asking for **exactly** what you want to see in the result, and you may be surprised at how well the model satisfies your request

    - Write clear instructions :Include specifics
        - Optional:
            - Include complex instructions
                - i.e.,
                    - Bulleted lists of details
                    - Length of response
                    - Desider formats to be included in output
    - Format of instructions
        - Recency bias can affect models: Information located towards the end of the prompt can have more influence on the output.
        - You may have better responses if you repeat instructions at the end of the prompt.

Azure OpenAI Service

- Write Effective Prompts
  - Primary, supporting and grounding content
    - Including content for the model to use
      - Primary content
        - i.e. , an article we want to summarize. Add it between --- blocks and then end with an instruction

          ---

          <full article>

          Summarize this article and identify three takeaways in a bulleted list
      - Supporting content
        - Content that might alter the response but it isn't the focus or subject of the prompt
        - Names, preferences, future date to include in response, etc.

          <full email here as primary content>

          <the next line s the supporting content>

          Topics I'm very interested in: AI,webinar dates, submission deadlines

          Extract the key points from the above email , and put them in a bulleted list

Azure OpenAI Service

- Write Effective Prompts
  - Primary, supporting and grounding content
    - Including content for the model to use
      - Primary content
      - Supporting content
      - Grounding content
        - Allows the model to provide reliable answers by providing content to draw answers from
        - Essais, articles, company FAQ document, information more recent than the data the model was trained on.
        - It is different from primary content because it is used to answer the prompt but it is not being operated on for summarization, translation etc. Example: providing as a grounding content an unpublished research paper on the history of AI so the model can answer questions using that grounding content.

          <unpublished paper on the history of AI here, as grounding content>

          Where and when did the field of AI start?

Azure OpenAI Service

- Write Effective Prompts
  - Primary, supporting and grounding content
    - Including content for the model to use
      - Primary content
      - Supporting content
      - Grounding content
  - Cues
    - Leading words for the model to build upon, and often help shape the response in the right direction
    - Often used with instructions but not always
    - Particularly useful if prompting the model for code generation

Azure OpenAI Service

- Provide text with prompt engineering
  - Request output composition: Specifying the structure of the output

    “Write a table in markdown with 6 animals in it, with their genus and species”
  - System message

    {“role”: “system”,”content” : “You are a casual, helpful assistant. You will talk like an American old western film character.” }
  - Conversation history Enables model to continue responding in a similar way (tone, formatting) and allow the user to reference previous content in subsequent queries
  - Few shot learning Using s user defined example conversation

    User: That was an awesome experience
    Assistant: positive
    User: Iwon;t do that again
    Assistant: negative
    …
  - Break down a complex task: Divide complex prompts into multiple queries
  - Chain of thought : What sport is easiest to learn but hardest to master? Give a step by step approach of your thoughts, ending in your answer

Azure OpenAI Service

- Construct code with Natural language
  - Write functions
  - Change coding language
  - Understand unknown code
  - Complete code and assist the development process
  - Write unit tests (test code)
  - Add comments and generate documentation
  - Fix bugs and improve performance of your code
  - Refactor inefficient code

Knowledge check

200 XP

Module assessment • 3 minutes

Great job! You passed the module assessment.

1. What Azure OpenAI base model can you deploy to access the capabilities of ChatGPT? *

- text-davinci-003
- gpt-35-turbo
  - Correct. Only the gpt-3.5-turbo and later models can be used to access the chat capabilities.
- text-embedding-ada-002 (Version 2)

2. Which parameter could you adjust to change the randomness or creativeness of the completions returned? *

- Temperature
  - Correct. The temperature parameter can be adjusted to change the randomness or creativeness of the completions returned.
- Frequency penalty
- Stop sequence

3. Which Azure AI Foundry playground is able to support conversation-in, message-out scenarios? *

- Images
- Chat
  - Correct. The Chat playground is able to support conversation-in, message-out scenarios.
- Bot

Next unit: Summary

Previous Next

Retrieval Augmented Generation (RAG)

with Azure OpenAI Service

Retrieval Augmented Generation
=
Connecting pretrained models to your own data sources

(OpenAI on your data searches and add relevant data chunks of it to the prompt as grounding data before sending it)

Notes

- OpenAI on your data encourages (but don't require) the model to respond using only your data
  - This setting can be unselected
    - This may result in the model choosing to use its pretrained knowledge over your data

Fine-tuning vs RAG

- Fine-tuning: Create a custom model by training an existing foundational model (i.e., gpt-35-turbo) with a database of additional training data
  - COSTLY AND TIME INTENSIVE PROCESS
  - Higher quality requests than prompt engineering alone
  - Examples larger than what can be fit in a prompt
  - Allow the user to provide fewer examples to get the same quality response
- RAG
  - No training needed
  - Connects to the model via stateless API
  - AI Search first finds the useful information to answer the prompt and adds it to the prompt as grounding data
  - Azure OpenAI then forms the response based on that info

Add your own data source

Can be done through:

- The Azure AI Studio
- Chat playground
- Specifying your data source in an API call

If uploading or using files already in a storage account, Azure Open AI on your data supports:

.md, .txt, .html, .pdf, and Microsoft Word or PowerPoint files

Knowledge check

200 XP

Module assessment • 2 minutes

ⓘ Great job! You passed the module assessment. ×

1. What does RAG with Azure OpenAI enable developers to do? *

- Create their own AI chat models
- Access Azure OpenAI without an approved subscription
- (●) Use supported AI chat models that can reference specific sources of data

✔ **Correct. Azure OpenAI on your data allows developers to use supported AI chat models that can reference specific sources of data to ground the response.**

2. What is the recommended way to add data when using Azure OpenAI on your data? *

- Using any data source option available for Azure OpenAI on your data.
- (●) Using Azure AI Studio to create the search resource and index.

✔ **Correct. Using Azure AI Studio allows the appropriate chunking to happen when inserting into the index, yielding better responses.**

- Connecting to files in a storage account without using Azure AI Studio.

3. What are some recommended prompt engineering techniques when using RAG with Azure OpenAI on your own data? *

- (●) Break down the task and use chain of thought prompting.

✔ **Correct. Breaking down the task and using chain of thought prompting can help the model respond more effectively within the token limit.**

- Include as much conversation history as possible in your call.
- Use a single long prompt to provide all necessary information.

Generate images

with Azure OpenAI Service

DALL·E

Neural network based model that can generate graphical data from natural language input

- **The images are original,not a result of a search**

DALL-E 3 models are only available in Azure OpenAI service resources in the East US, Australia East, and Sweden Central regions.

Knowledge check

200 XP

Module assessment • 3 minutes

Great job! You passed the module assessment.

1. You want to use a model in Azure OpenAI to generate images. Which model should you use? *

- [x] DALL-E
  - ✔ The DALL-E model is used to generate images based on natural language prompts.
- [ ] GPT-35-Turbo
- [ ] Text-Davinci

2. Which playground in Azure AI Studio should you use to utilize the DALL-E model? *

- [ ] Completions
- [ ] Chat
- [x] Images
  - ✔ The Images playground is used to explore image generation models.

3. In a REST request to generate images, what does the n parameter indicate? *

- [ ] The description of the desired image.
- [x] The number of images to be generated
  - ✔ The number of images to be generated is specified in the n parameter.
- [ ] The size of the image to be generated

Next unit: Summary

Previous Next

Still 2-do

- Read documentation
  - https://learn.microsoft.com/en-us/azure/ai-services/
- Implement decision-support solutions
  - Implement content moderation solutions (Deprecated)
  - Now: Azure AI Content Safety
  -
-