

OpenAI und ChatGPT: Neuer Support in der IT!

Thomas Jäkel

brainymotion

Lead Trainer Cloud Infrastructure

Microsoft Certified Trainer since 1999

github.com/www42/openAI

brainymotion



Agenda Tag 2

1. Cloud Computing
2. Microsoft Azure OpenAI
3. Vector Search
4. Prompt Engineering Retrieval Augmented Generation RAG

$$12^{30} - 13^{15}$$

Tenant  Entra ID (AzureAD)

Microsoft Cloud

Azure

M365

SDN

VM
Container
SQL CosmosDB
App Service

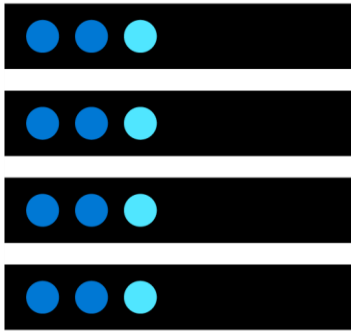
Exchange O
Share point
O365

Teams₄

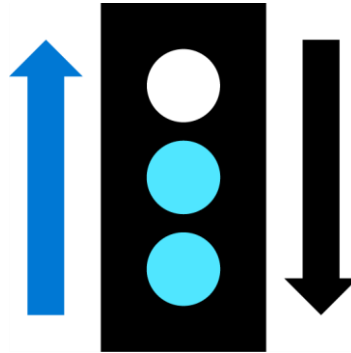
What is cloud computing?

AWS
Azure
GCP Google

Cloud Computing is the delivery of computing services over the internet, enabling faster innovation, flexible resources, and economies of scale.



Compute

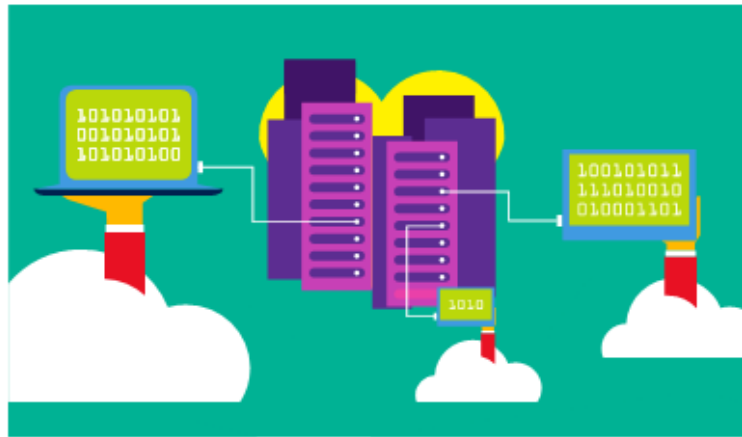


Networking



Storage

Hybrid cloud



Combines **Public** and **Private** clouds to allow applications to run in the most appropriate location.

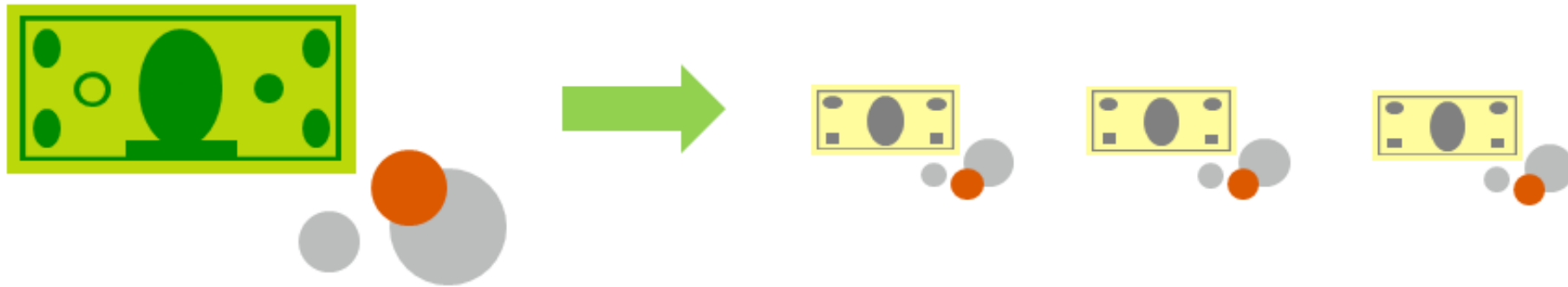
Compare CapEx vs. OpEx

Capital Expenditure (CapEx)

- The up-front spending of money on physical infrastructure.
- Costs from CapEx have a value that reduces over time.

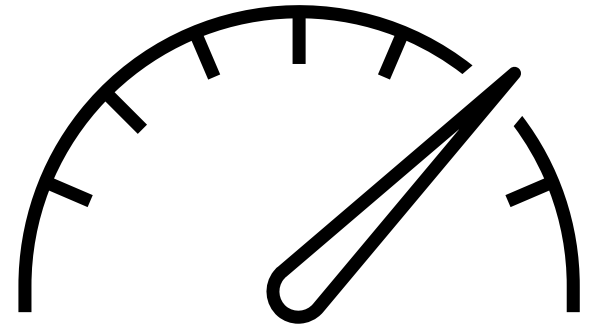
Operational Expenditure (OpEx)

- Spend on products and services as needed, pay-as-you-go
- Get billed immediately



Consumption-based model

- Cloud service providers operate on a consumption-based model, which means that end users only pay for the resources that they use. Whatever they use is what they pay for.
- Better cost prediction
- Prices for individual resources and services are provided
- Billing is based on actual usage



Cloud Benefits

High availability

Scalability

Predictability

Governance

Elasticity

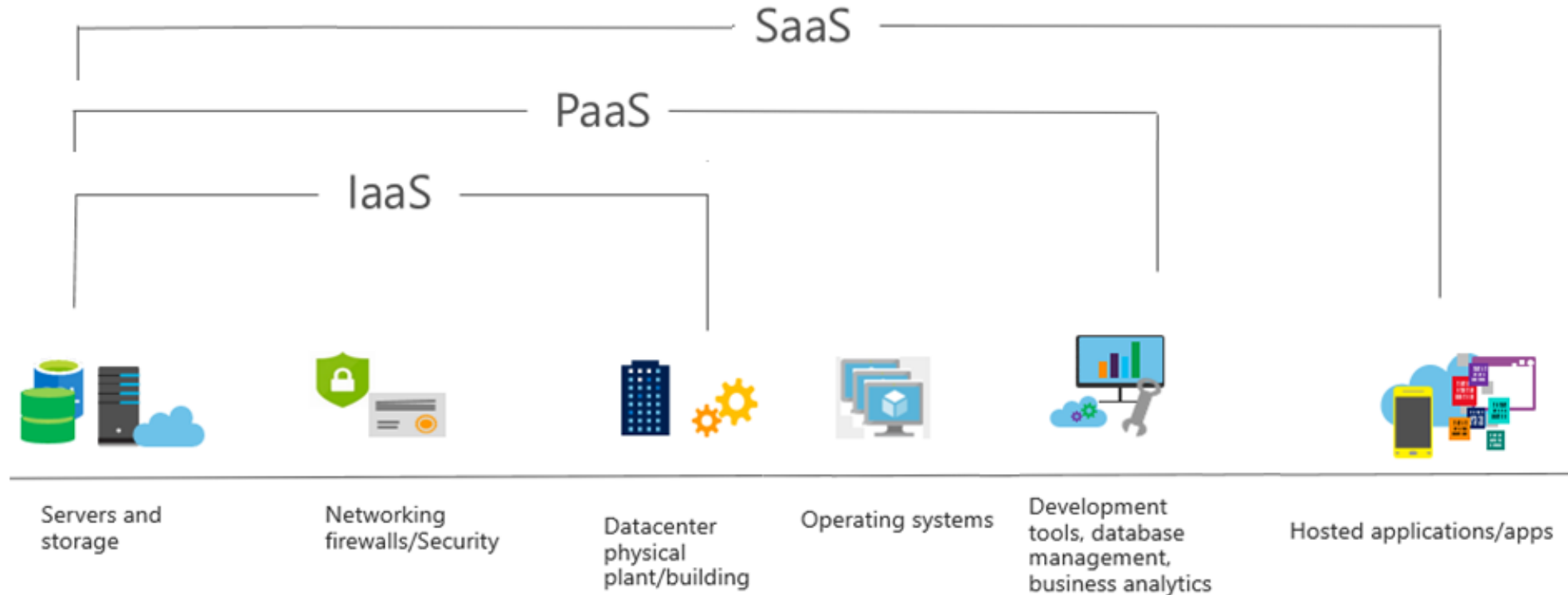
Reliability

Security

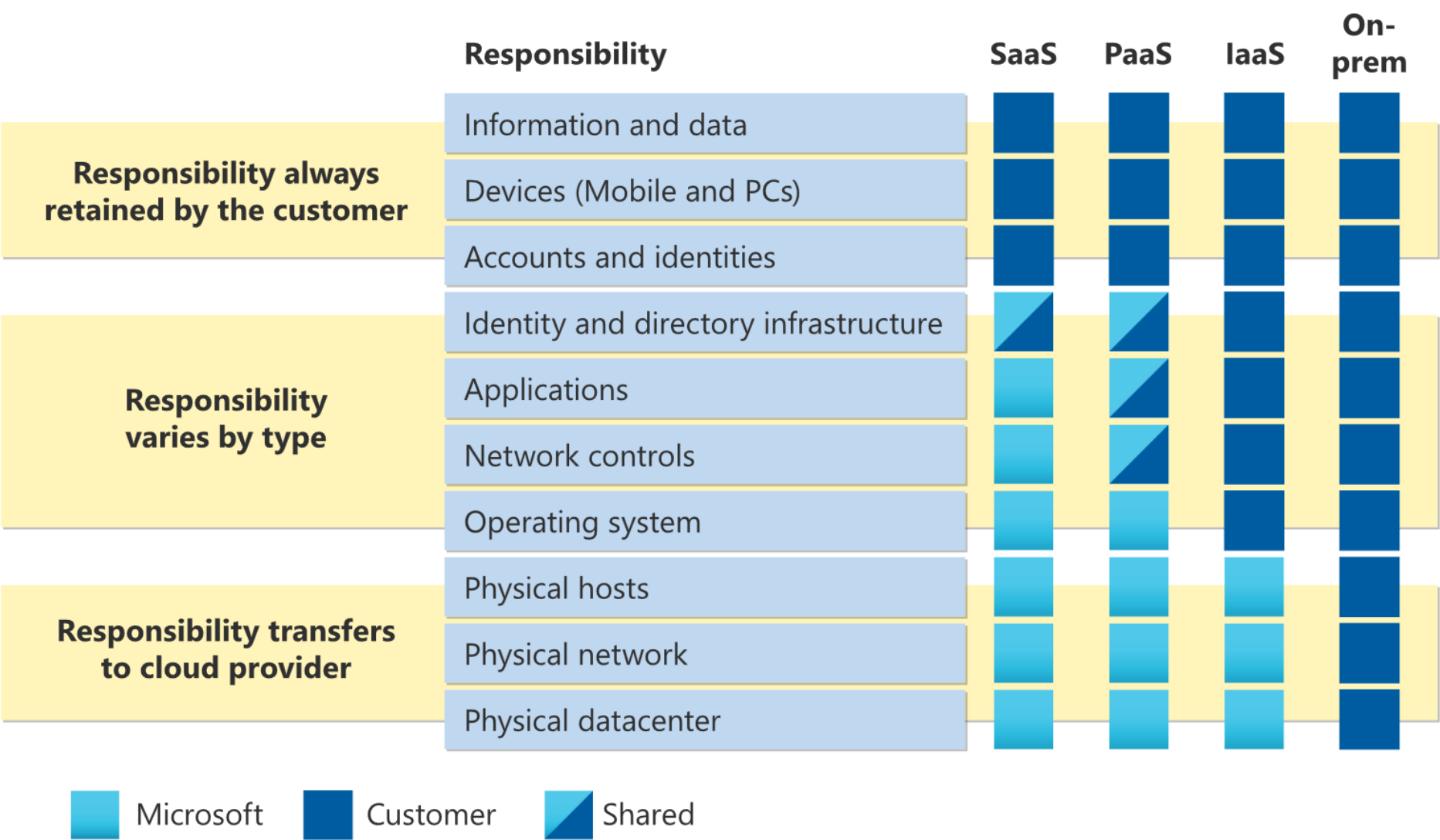
Manageability

Software as a Service (SaaS)

Users connect to and use cloud-based apps over the internet: for example, Microsoft Office 365, email, and calendars

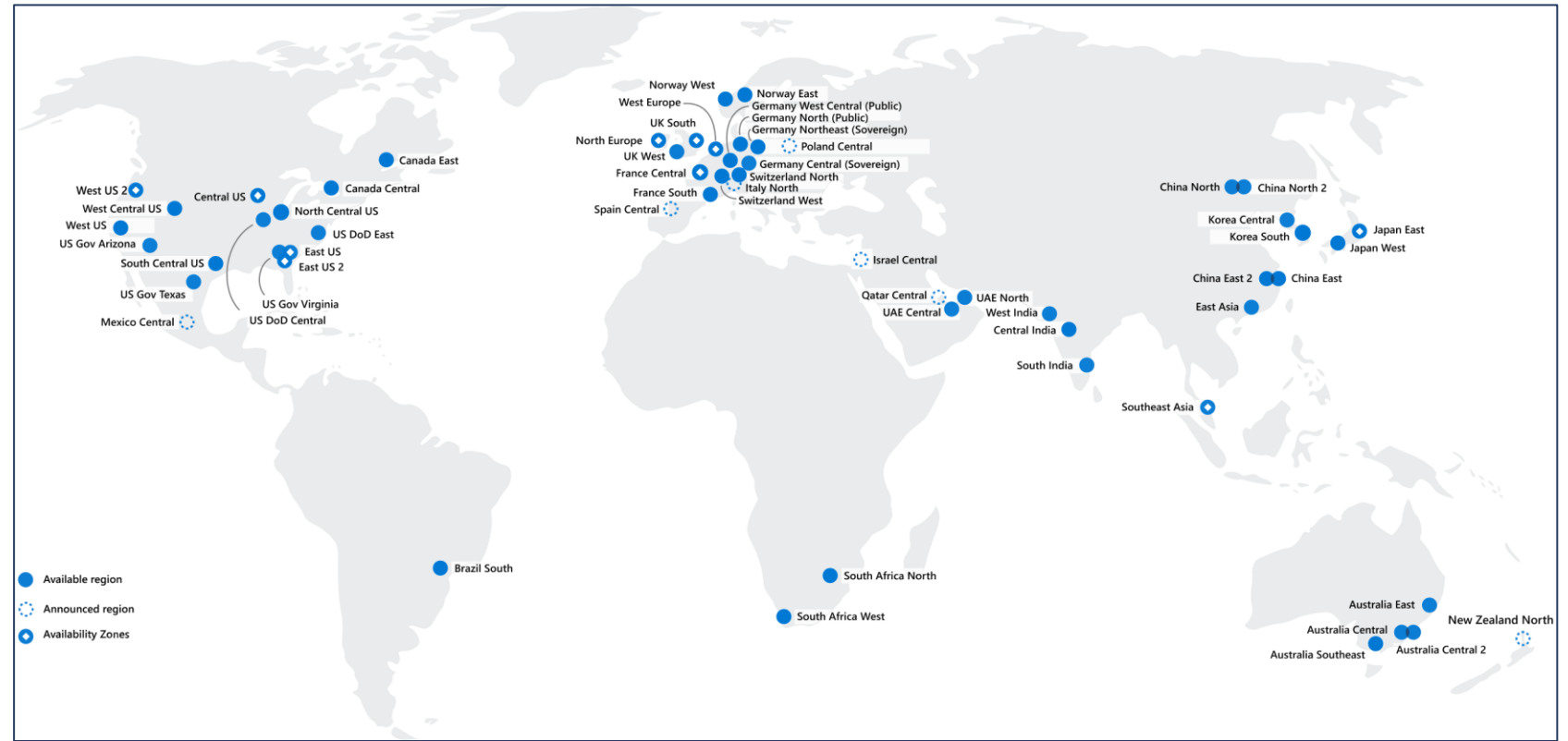


Shared responsibility model



Regions

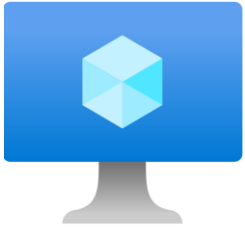
Azure offers more global regions than any other cloud provider with 60+ regions representing over 140 countries



- Regions are made up of one or more datacenters in close proximity.
- Provide flexibility and scale to reduce customer latency.
- Preserve data residency with a comprehensive compliance offering.

Azure Resources

Azure **resources** are components like storage, virtual machines, and networks that are available to build cloud solutions.



Virtual Machines



Storage Accounts



Virtual Networks



App Services



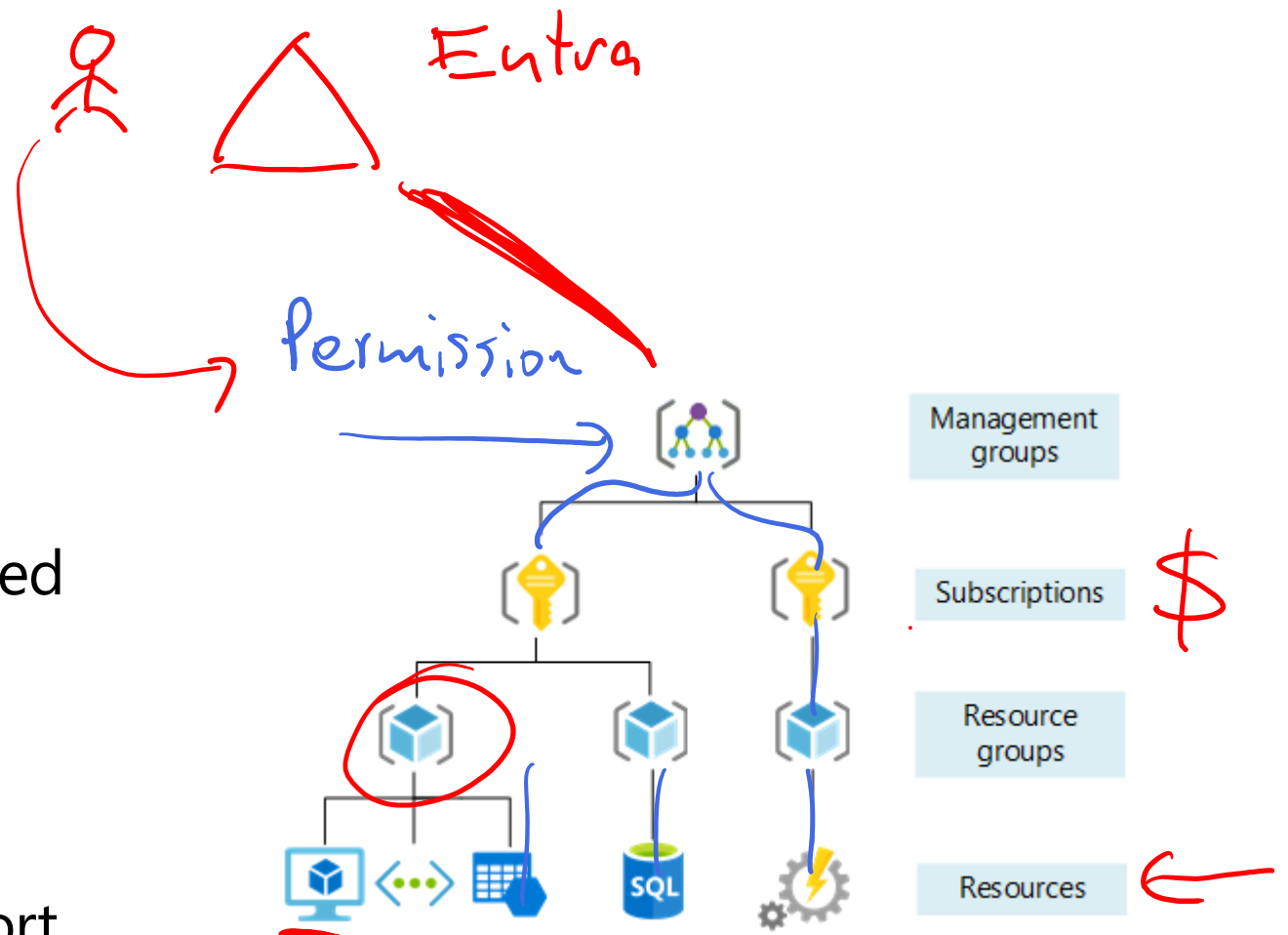
SQL Databases



Functions

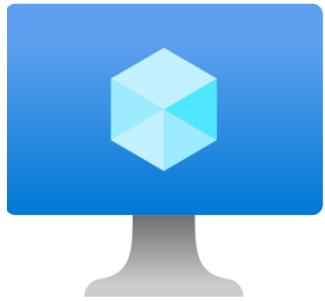
Management Groups

- Management groups can include multiple Azure subscriptions.
- Subscriptions inherit conditions applied to the management group.
- 10,000 management groups can be supported in a single directory.
- A management group tree can support up to six levels of depth.



Azure compute services

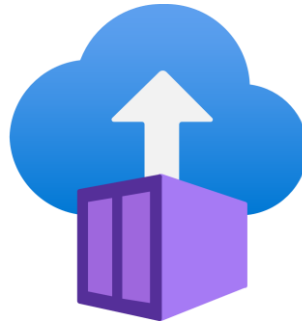
Azure **compute** is an on-demand computing service that provides computing resources such as disks, processors, memory, networking, and operating systems.



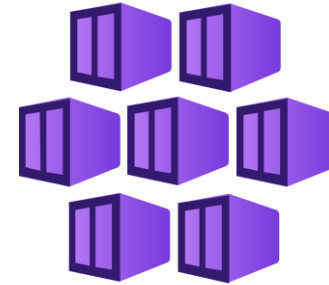
Virtual
Machines



App
Services



Container
Instances



Azure Kubernetes
Services (AKS)



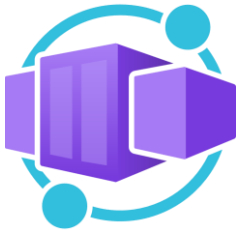
Azure Virtual
Desktop

Azure Container Services

Azure Containers are a light-weight, virtualized environment that does not require operating system management, and can respond to changes on demand.



Azure Container Instances: a PaaS offering that runs a container or pod of containers in Azure.



Azure Container Apps: a PaaS offering like container instances that can load balance and scale.

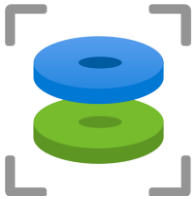


Azure Kubernetes Service: an orchestration service for containers with distributed architectures and large volumes of containers.

Azure storage services



Azure Blob: optimized for storing massive amounts of unstructured data, such as text or binary data.



Azure Disk: provides disks for virtual machines, applications, and other services to access and use.



Azure Queue: message storage service that provides storage and retrieval for large amounts of messages, each up to 64KB.

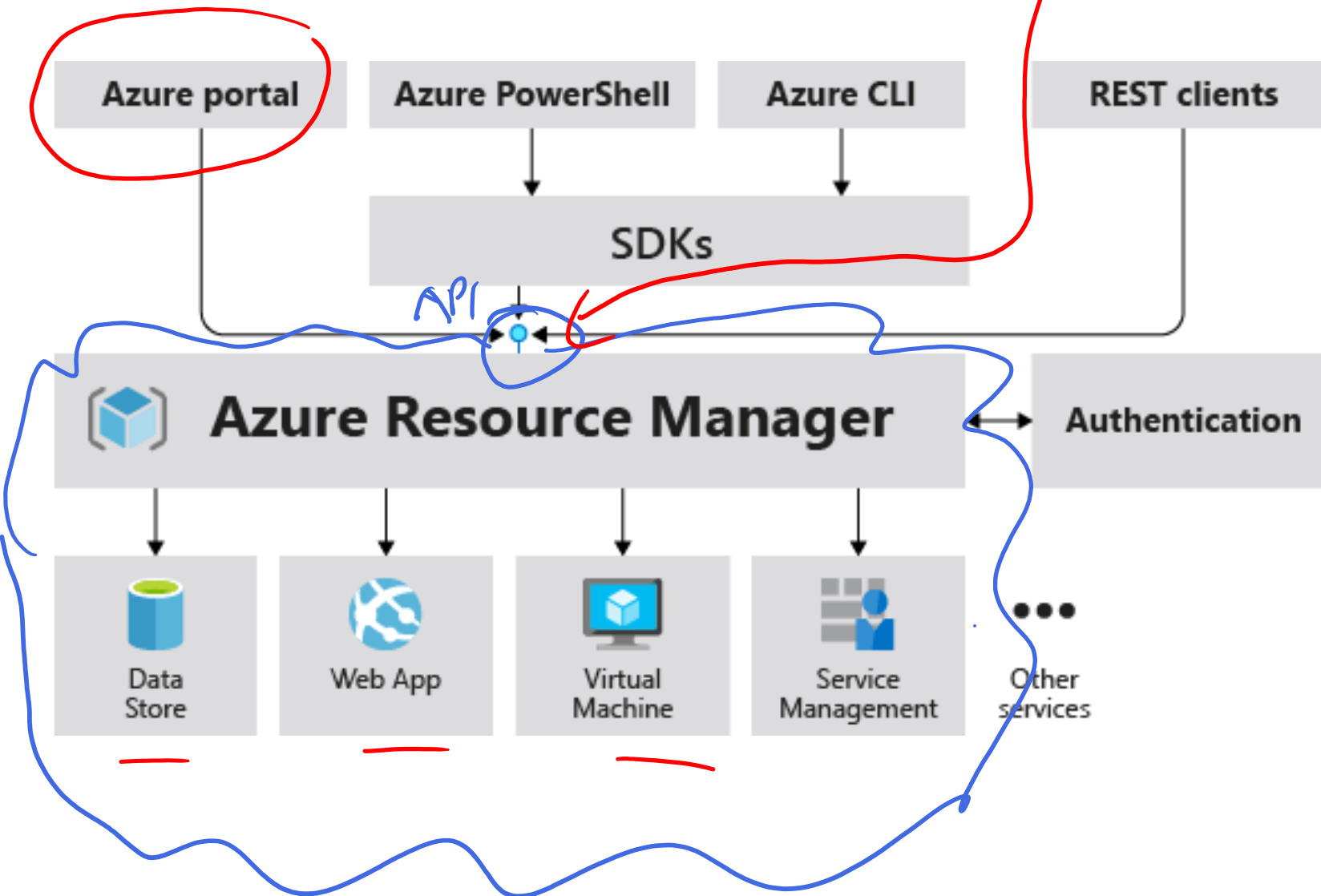


Azure Files: sets up a highly available network file share that can be accessed by using the Server Message Block protocol.



Azure Tables: provides a key/attribute option for structured non-relational data storage with a schema-less design.

Azure Resource Manager



ARM Templates

- declarative
- idempotent

The **Azure Resource Manager (ARM)** provides a management layer that enables you to create, update, and delete resources in your Azure subscription.

Semper idem!

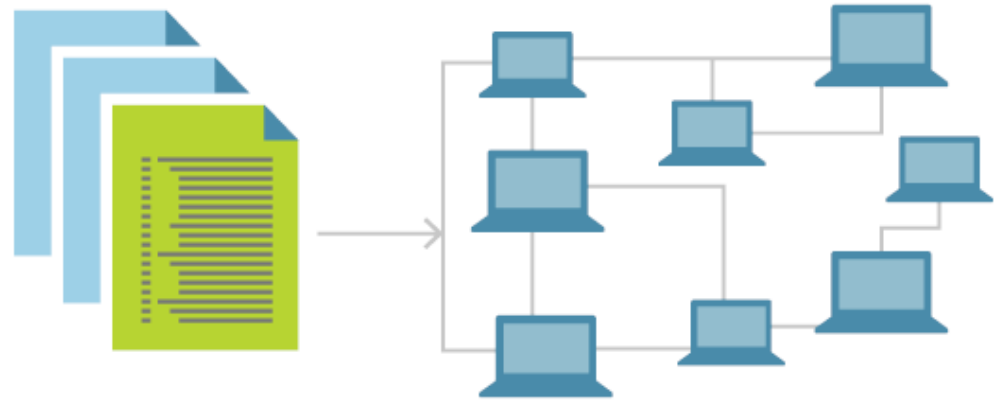
Infrastructure as code

- Ensure consistency in deployment across your cloud ecosystem.
- Manage configuration at scale.
- Rapidly provision additional environments based on a standard configuration and build.

Bicep → ARM

Microsoft

Terraform

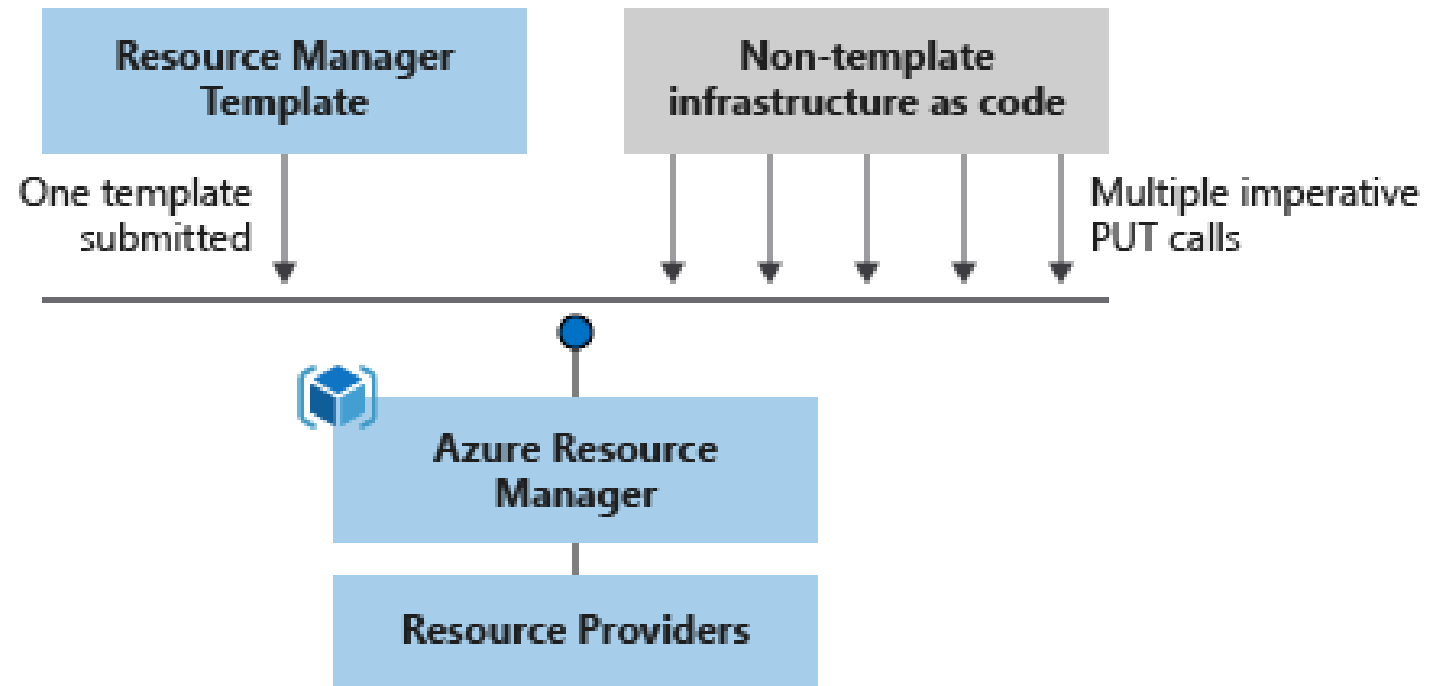


Helm

Azure Resource Manager (ARM) templates

Azure Resource Manager (ARM) templates are JavaScript Object Notation (JSON) files that can be used to create and deploy Azure infrastructure without having to write programming commands.

- Declarative syntax
- Repeatable results
- Orchestration
- Modular files
- Built-in validation
- Exportable code



Nokia

Github

OpenAI & Microsoft



*Ensure that artificial
general intelligence (AGI)
benefits humanity*



*Empower every person and
organization on the planet
to achieve more*

GPT-3.5 and GPT-4

Text

ChatGPT

Conversation

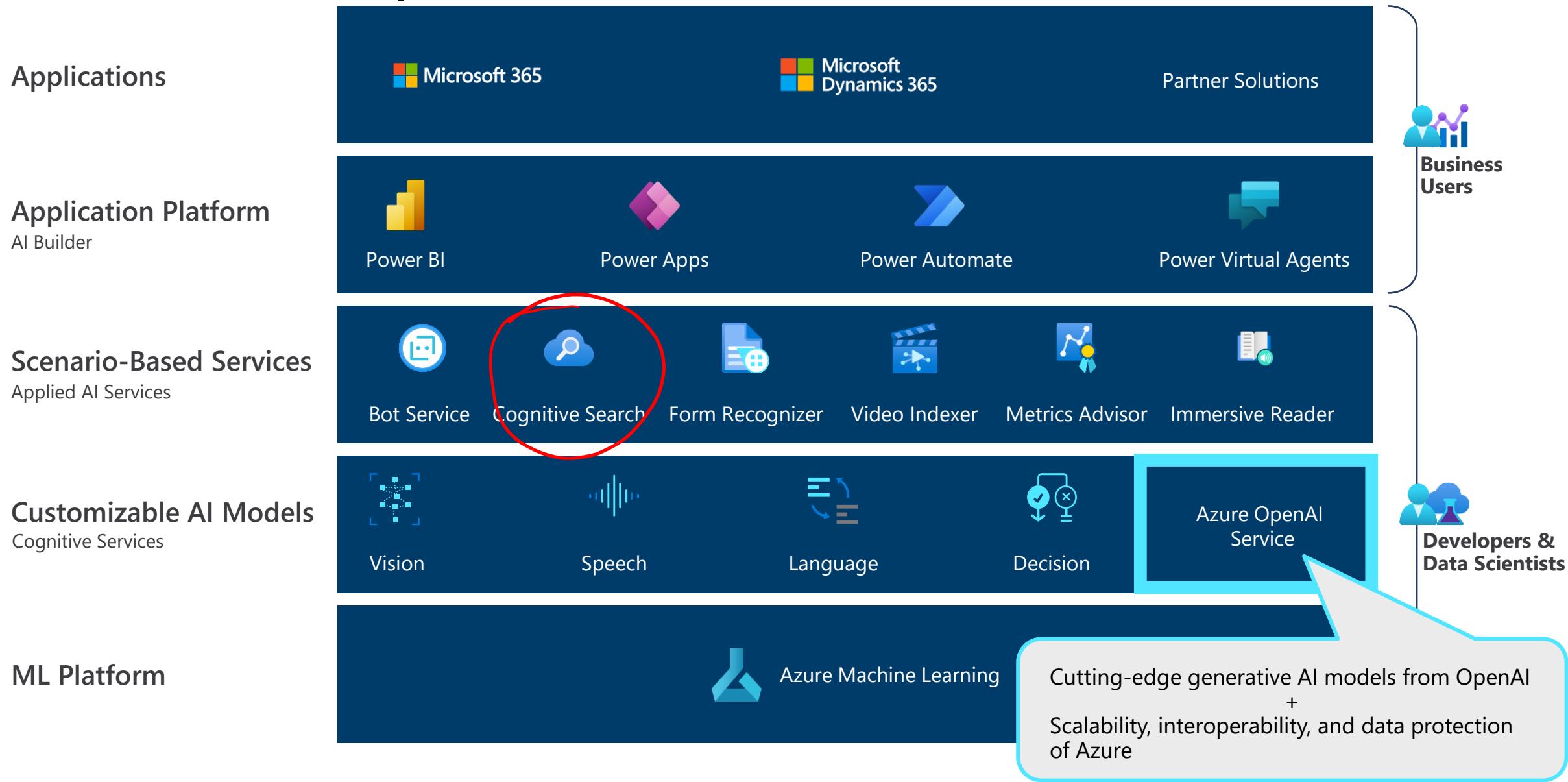
Codex

Code

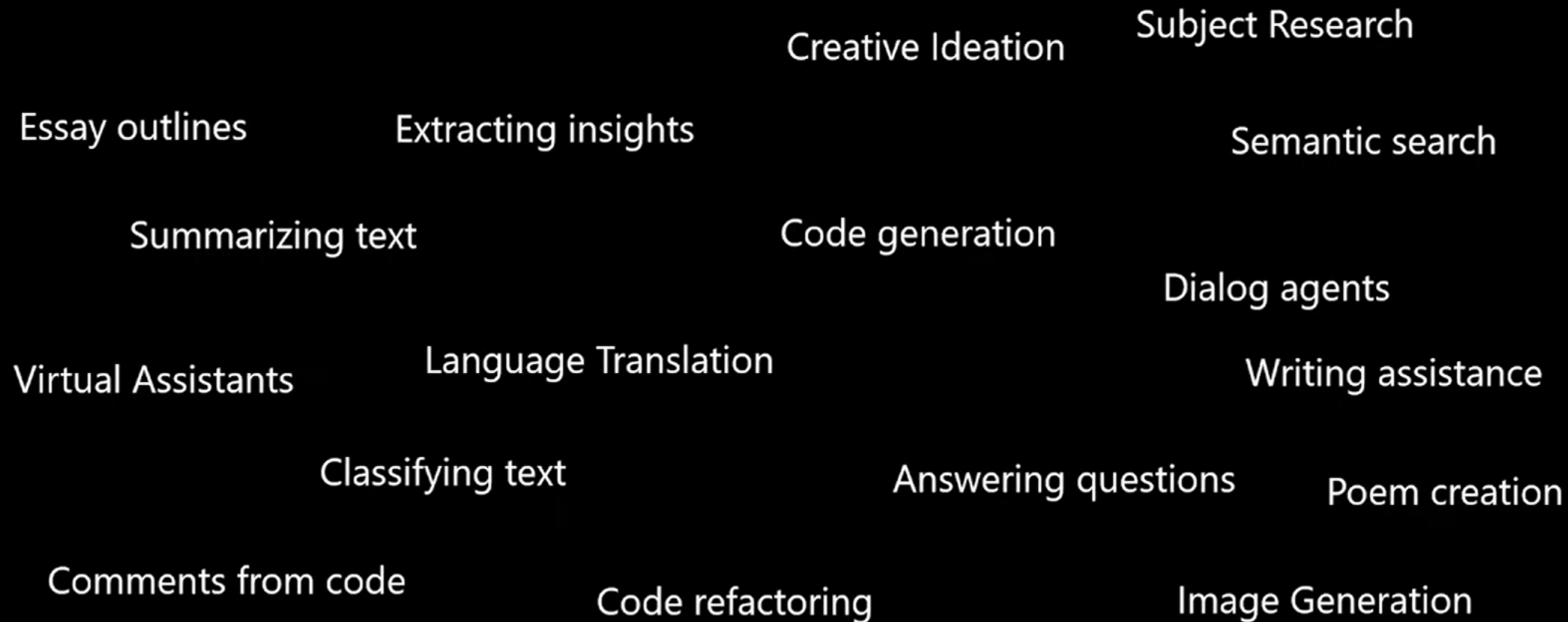
DALL·E 2

Images

What is Azure OpenAI Service?




| Azure OpenAI Service Capabilities



Azure OpenAI model capabilities

- The service include multiple models, optimized for specific tasks
- Models generate responses based on natural language *prompts*

	Language Generation	Code Generation	Image Generation
Prompt:	Write a haiku about marmalade	Write a Python function to add two numbers	Paint a pink fox in a field in the style of Monet
Output:	<i>Orange sunrise, sweet Spread on toast with morning tea A marmalade treat</i>	<pre>def add_two_numbers(a, b): return a + b</pre>	

Using prompts to get completions from models

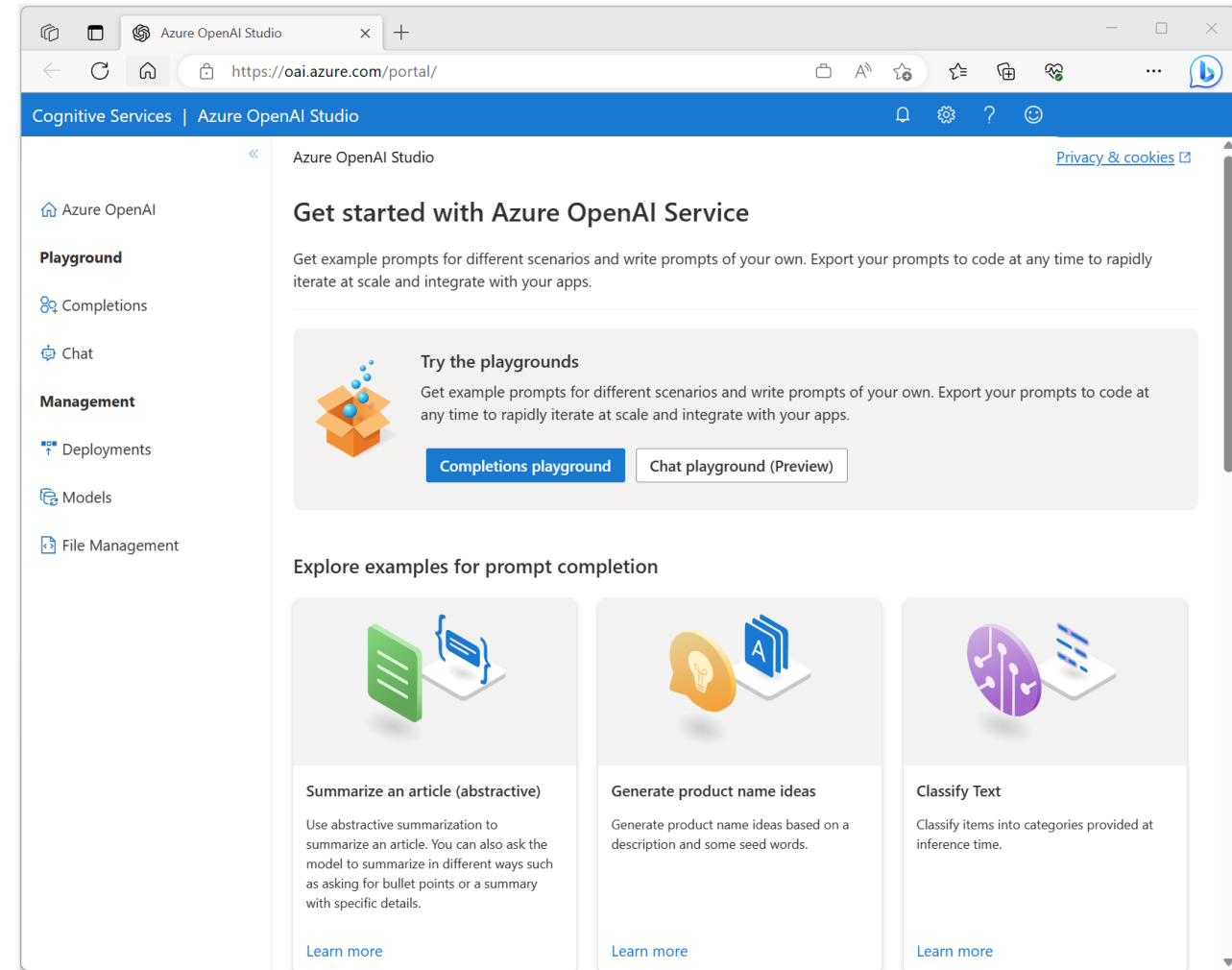
Task	Prompt	Completion
Classifying content	Tweet: I enjoyed the training course. Sentiment:	Positive
Generating new content	Write a poem about databases	Databases, oh databases, You keep our information safe, From the small to the large, You store our data in a place.
Transformation/Translation	English: Hello French:	Bonjour
Summarization	Scotland is [<i>long description of Scotland...</i>] Summarize the previous text	Scotland is [<i>summarized description...</i>]
Continuation	One way to grow tomatoes is to	start with seeds...
Question answering	How many moons does Earth have?	Earth has one moon.
Chat	<i>Setup, followed by messages...</i>	<i>A sequence of relevant responses</i>

| Azure OpenAI | GPT-3 Family of Models

Model	Request	Description, performance, cost	Use cases
Davinci	4,000 tokens	Most capable GPT-3 model. Can do any task the other models can do, often with <i>higher quality, longer output</i> and <i>better instruction-following</i> .	Complex intent, cause and effect, summarization for audience
Curie	2048 tokens	Very capable , but <i>faster</i> and <i>lower cost</i> than Davinci.	Language translation, complex classification, text sentiment, summarization
Babbage	2048 tokens	Capable of straightforward tasks, <i>very fast</i> , and <i>lower cost</i> .	Moderate classification, semantic search classification
Ada	2048 tokens	Capable of very simple tasks, usually the <i>fastest</i> model in the GPT-3 series, and <u>lowest cost</u> .	Parsing text, simple classification, address correction, keywords

Azure OpenAI Studio

- Web portal for working with Azure OpenAI models:
<https://oai.azure.com/>
- View and deploy base models
- Manage fine tuning and data files for custom models
- Test models in visual playgrounds:
 - **Completions** (GPT-3 and earlier models)
 - **Chat** (GPT-3.5-Turbo and later models)



Vector Search

"LOREM IPSUM
DOLOR SIT AMET,
CONSECTETUR..."

DATA

gpt-4

embedding
model

-0.001
0.006
-0.014
...
-248.6

vector

open source Pinecone
MS Azure Cognitive
search

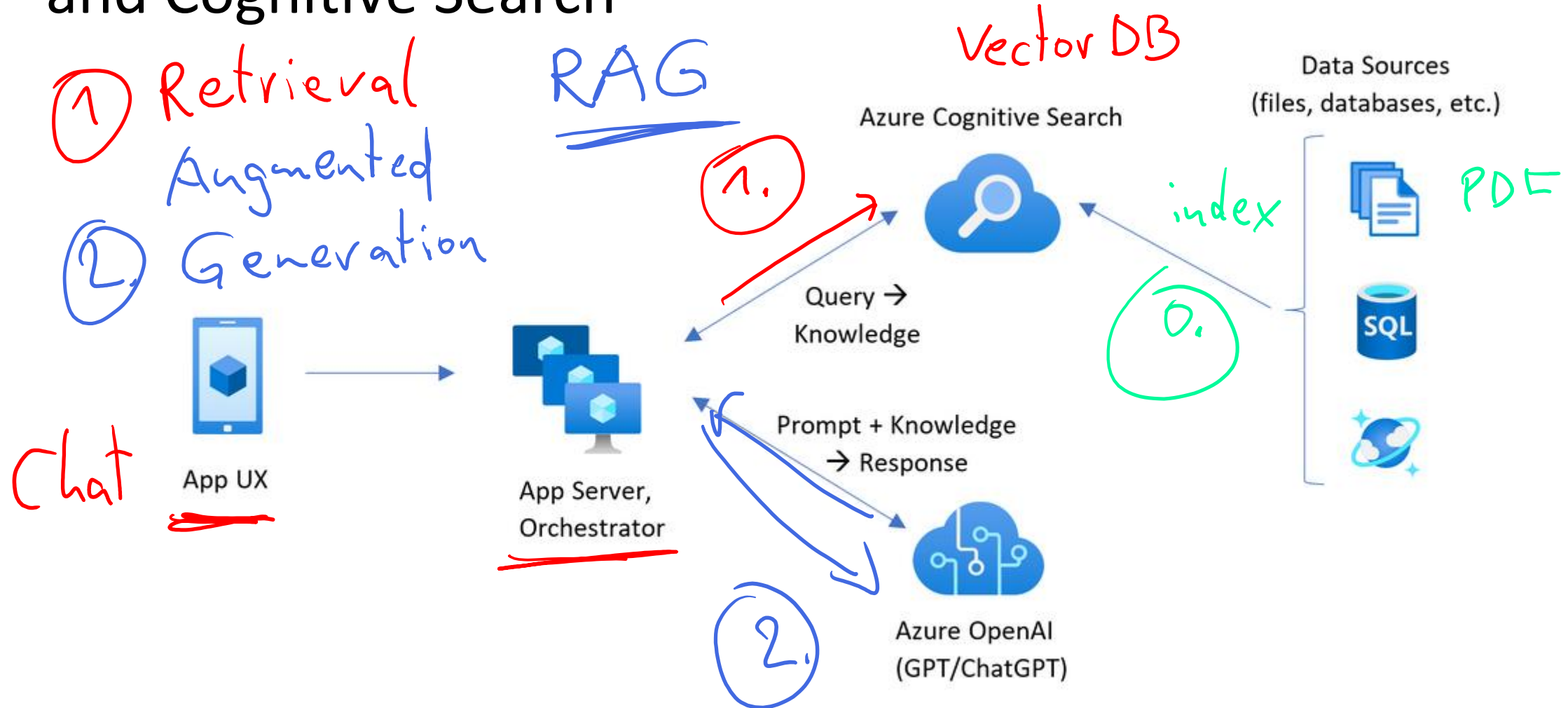
vector
DB





Prompt Engineering Retrieval Augmented Generation (RAG)

ChatGPT + Enterprise data with Azure OpenAI and Cognitive Search



ChatGPT

Eine Bereicherung für die täglichen Aufgaben