**Kurz**

Všechny informace ke kurzu najdeme zde: <https://microsoftlearning.github.io/mslearn-azure-ml/>

Github – Materiály k DataScience kurzu, které jsou ještě lepší než na microsoftu

<https://github.com/microsoftlearning>

<https://github.com/MicrosoftLearning/mslearn-azure-ml>

Learn on demand. Přihlašovací údaje mám v pracovním emailu. Můžeme se přihlásit ke svým Labům ve kterých jsme pracovali na kurzu: [**https://gopas.learnondemand.net/**](https://gopas.learnondemand.net/)

Pokud bychom chtěli získat certifikaci k Datascience v Azure, tak certifikace je platná na rok a pak se dá obnovit bez kontroly, která byla například u AWS.

U přípravy dat je třeba data nejdříve normalizovat. Pokud používáme například skalární data.

Stratifikace dat – Znamená například nediskriminovat někoho v datech. Například málo žen v kurzu Gopasu. Pro správný a komplexní sběr dat je nutné sbíraná data stratifikovat, tedy sbírat je do skupin podle námi zvoleného faktoru.

Klacifikace – Kategorizace, volební preference, supervised machine learning. Tzn už máme nějaká ověřená data.

Shluková analýza(Clustering) – společné znaky chování. Nevíme ještě jak lidi kategorizovat. Unsepervised. Snaží se hledat vzorce v datech vertikálně nebo horizontálně.

Regrese – Kolik objednat na sklad produktů. Je schopné odpovědět kolik objednat produktu napriklad na jednotlive dny. Pocet produktu mohou ovlivnovat jednotlive dny, pocasi, atp.

Time-series forecasting – Predictive maintance – Predikce toho kdy například vyměnit nástroj. Kdy se opotřebí, kde dojde k poruše atp. Supervised.

[**https://azure.microsoft.com/en-us/products/data-factory**](https://azure.microsoft.com/en-us/products/data-factory)

Nástroj k připojení k datům. ELT a ETL nástroj v Azure. Existuje zde mnoho konektorů, takže se lze připojit opravdu ke všemu.

**ETL (Extract, Transform, Load)** je proces, kdy se data nejprve extrahují z různých zdrojů, poté transformují do požadovaného formátu a nakonec se načítají do cílové databáze nebo datového skladu.

**ELT (Extract, Load, Transform)** je proces, kdy se data nejprve extrahují, následně se načítají do cílové databáze a teprve poté se transformují přímo v cílovém systému.

**Příprava dat a kontrola dat** – Zkontrolovat například pokud data generuje stroj, tak jestli nemají velké hodnoty. Například stroj generuje data od 0 do 100, ale najednou v datech je 1000 nebo nám někde chybějí data. Je třeba identifikovat tyto problémy a opravit je. Případně json může mít vždy jinou strukturu, takže je třeba dle toho následně upravit kod.

**https://www.datasciencecentral.com/**

**Azure Machine learning Workspace**

Místo kde se pracuje s datascience. **Container Azure block storage** něco jako S3 na AWS. Platí se od načatého GB a cca 6 eurocentů za měsíc. Jsou hodně levné.

**ADLS – Azure Data Lake Storage** je nadstavba storage. Každá data data jsou uložena duplicitně. Využití vyšší prostupnosti dat. ADLS je o něco dražší než na block storage.

Tenant – Je něco jako User v AWS, který se následně může přihlásit do Azure. Za Tenanta se neplatí.

Na Marketplacu existuje riziko, že mi to zvýší několikanásobně finance. Stáhnout si z market placu pouze aplikace nebo věci, které mi třeba někdo doporučil. Na Marketplace opravdu pozor.

Rozdíl mezi **Containery** a **File Share** je HTTPS a ABFFS.

Hot cold cool – můžeme je přesouvat z různých stagí, abychom ušetřili finance, ale může to být trochu komplikované, protože microsoftu trvá než data obnoví. Lektor nedoporučuje to používat.

**SQL pools** - SQL pools jsou určeny pro analytiku a SQL dotazy nad strukturovanými daty

**Spark pools** - Spark pools pro zpracování velkých dat a strojové učení

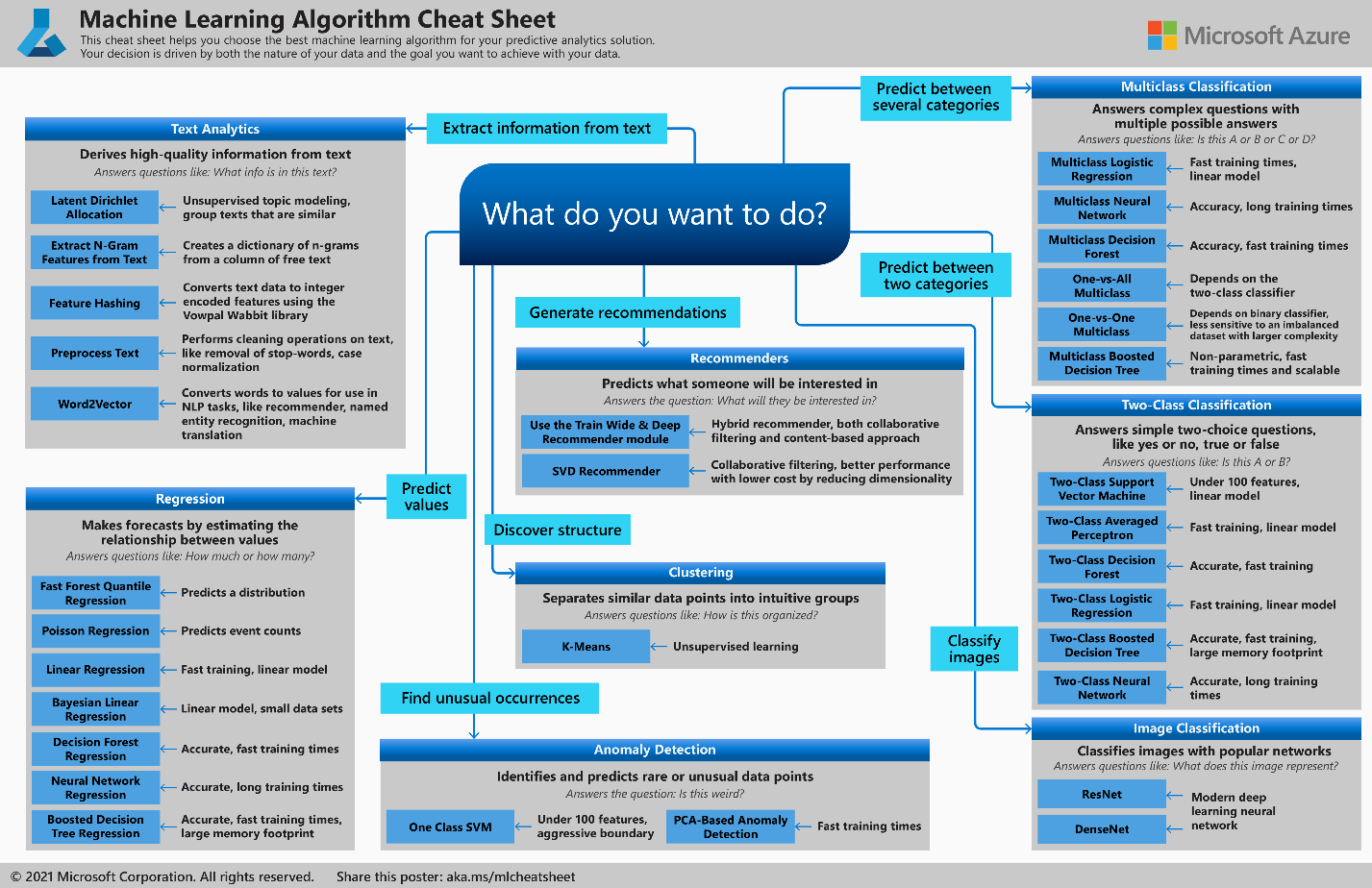
**Data explorer pools** - Data Explorer pools pro analýzu logů a telemetrických dat pomocí KQL

**Azure Synapse Analytics** – Využití k získání dat a uložení do blockstorage. Synapse notebooky ve kterých se pracuje s daty.

**Machine learning Studio** - Designer v Azure – Slouží k příkladu jak fungovat s DataScience modely v Azure. Většinou to bývá pomalejší než v Spark poolu nebo SQL poolu.

**Compute cluster** – Je Spark

<https://learn.microsoft.com/en-us/azure/machine-learning/algorithm-cheat-sheet?view=azureml-api-1>



**Scoring model** – Jdou do něho dva vstupy – testovací data a hotový model. Supervised data.

**Evaluate model** – kolik procent bylo správných výsledků, podíl správných a nesprávných. Porovnání skutečnosti s odhadem.

Machine learning mohu dělat v

**Azure AI Services –** Customize or consume prebuilt models

**MS Fabric –** One platform for all data engineering and data science projects at scale

**Azure Databricks –** Use notebooks for data engineering and data science at scale. Offers distributed compute (PySpark)

**Azure Machine Learning -**  Manage machine learning models from development to testing to production. Use Python in notebooks and scripts.

**Big Data Clustery –** Servery **Apache Hadoop**, **Spark clustery**, **NoSQL** clustery, **Cloudové Big Data** clustery pro příklad. Jsou to skupiny propojených výpočetních uzlů (nebo strojů), které jsou navrženy pro efektivní zpracování, analýzu a ukládání obrovských objemů dat, obvykle pomocí distribuovaných výpočetních frameworků.

**Materiály a kurzy k Databricks:** [**https://www.databricks.com/learn/training/home**](https://www.databricks.com/learn/training/home)

[**https://www.databricks.com/training/catalog/advanced-machine-learning-operations-3508**](https://www.databricks.com/training/catalog/advanced-machine-learning-operations-3508)

**Conversational AI** od Microsoftu se týká technologie, která umožňuje strojům rozumět, komunikovat a interagovat s uživateli prostřednictvím přirozeného jazyka, tedy způsobem, jakým komunikují lidé mezi sebou. Microsoft v tomto kontextu nabízí několik nástrojů a služeb pro vytváření chatbotů a dalších interaktivních aplikací, které využívají umělou inteligenci k rozpoznání, porozumění a generování textu nebo mluvené řeči.

**Microsoft Fabric –** Je to hrozně drahé