From DevOps to DataOps:

Cloud Based Software Development and Deployment

(Data Science Projects Operationalisation)

Dr. Yuri Demchenko
University of Amsterdam

OpenInfra Summit 2022
7-9 June 2022, Berlin

Outline

- Motivation: Large scale Research Infrastructure and data driven research
- Data Science development process
- Data Science Process models and ML model formats
- Operationalising Data Science Analytics and ML: DataOps, MLOps and platforms
- Case studies:
  - AWS MLOps Framework with SageMaker
  - Azure MLOps: Data Science Workflow
- Teaching DataOps: Extensions to DevOps and Data Science curricula

Motivation and Goal

- Research Universities: Close relation between Research, Development, and Education/Training
- SLICES: Large scale project on building large scale Research Infrastructure (RI) for digital technologies experimentation
  - Experimenting and advancing technologies while catching up with the industry pace
  - Solving technical, organizational/staffing, and governance/legal issues
- Data driven research:
  - Data Science Analytics for experimental data processing
  - Data Management and Governance over the whole data lifecycle
- Experimental infrastructure management: ITIL, DevOps, SRE + DataOps

Motivation and Goal

- Research Universities: Close relation between Research, Development, and Education/Training
- SLICES: Large scale project on building large scale Research Infrastructure (RI) for digital technologies experimentation
  - Experimenting and advancing technologies while catching up with the industry pace
  - Solving technical, organizational/staffing, and governance/legal issues
- Data driven research:
  - Data Science Analytics for experimental data processing
  - Data Management and Governance over the whole data lifecycle
- Experimental infrastructure management: ITIL, DevOps, SRE + DataOps

Data Science and Big Data Environment (production/experiment)

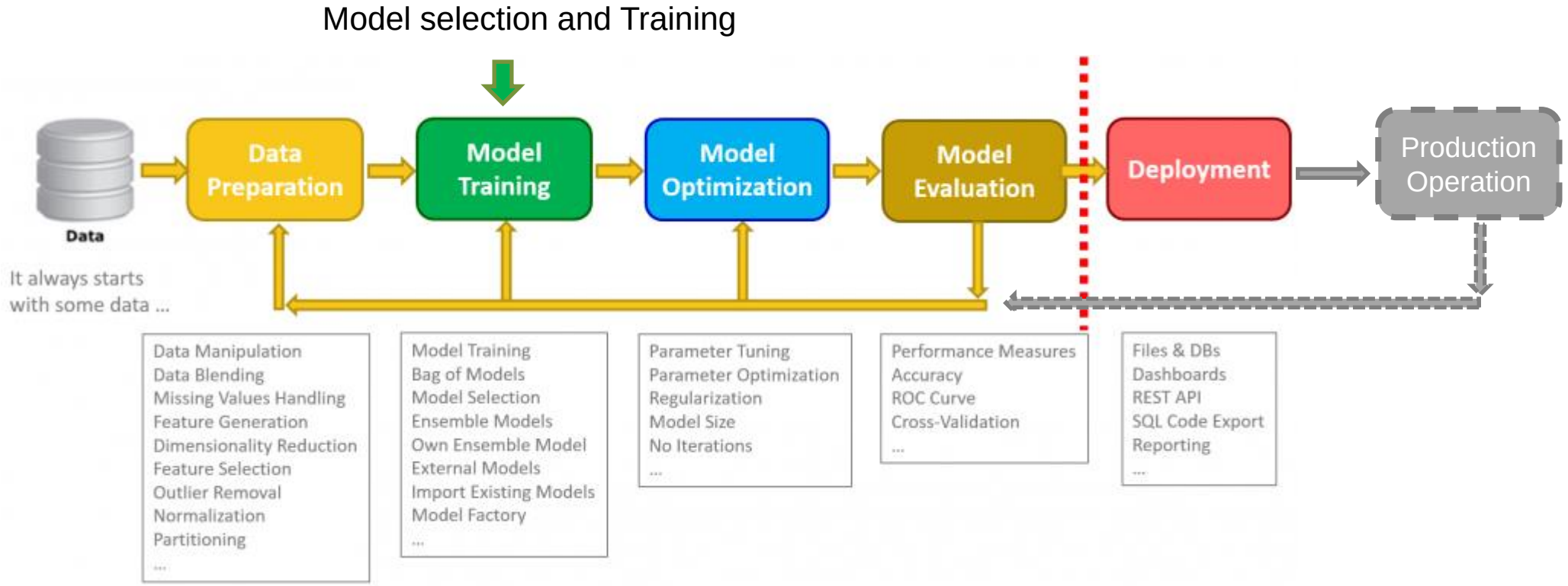
- Challenge: *Data Science is trapped in laptops*
  - There is a big step from laptop and Jupyter Notebook to production data and production application
- Productionalising Data Science Apps means working with Big Data
  - Means Big Data, Big Data Infrastructure and data driven applications
- Data Science team need to work with Software Engineers, Operators and Infrastructure Engineers
  - Data Science teams need to know their DevOps practices
  - DevOps Engineers need to know specifics of the Data Science projects

Data Science Development Process

Data Science process is dealing with the data pipelines that include stages:

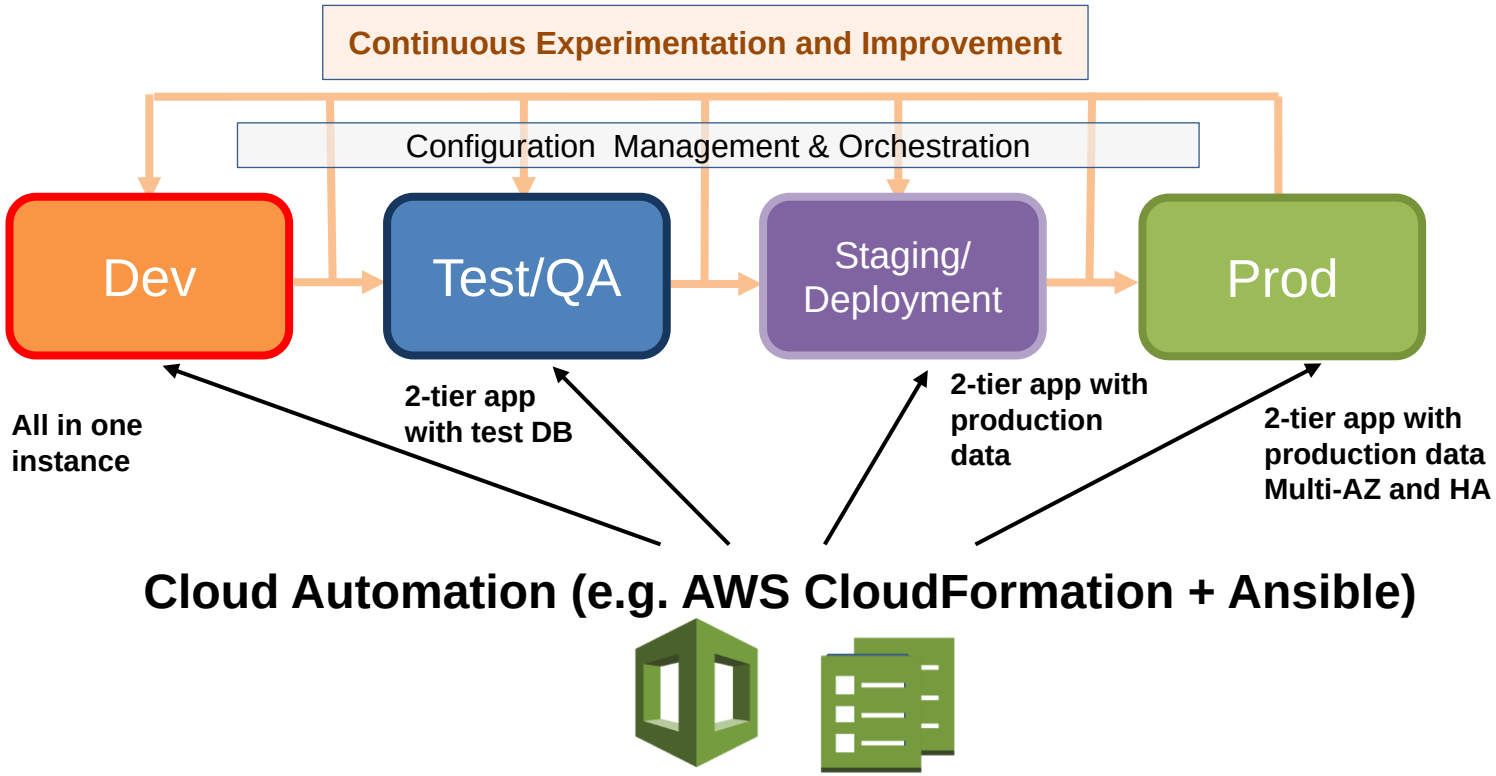
- **Collecting data** from multiple sources, also blending process or business data with external data such as environmental data or social media data that can be obtained via WebAPI or web scraping
- **Working with data** including data preparation, cleaning, filtering, and reformatting for modeling needs
- **Combing datasets** by joining on common attributes, consolidating attributes, build tabular data structure (such as used in popular analytics programming languages R, python, scala)
- **Feature engineering, algorithm selection, model training**
- **Testing** before production and **validating** the model during production, in particular, detecting drift in predictive models
- Implement changes and **deploy an updated model**
  - **Using standard DevOps CI/CD process**

Every Data Science Project/Research is based on Data Flow/Lifecycle (inspired by https://www.knime.com/blog/analytics-and-beyond)



- Each phase – data preparation, model training and evaluation, and model deployment – operates on **its own dataset**. All these data sets need to be isolated/split from each other. The pollution of data sets across the data science assembly line is one of the most frequent mistakes in model production.
- The data science journey/project always starts with some ***historical data*** *or* ***sample dataset*** lying around in a repository.
- ***Data blending*** involves additional data and connecting external sources

Compare: Services/Applications Development Lifecycle



- Easily creates test environment close to real
- Powered by cloud deployment automation tools
  - To enable configuration Management and Orchestration, Deployment automation
- Continuous development – test – integration
  - CloudFormation Template, Configuration Template, Bootstrap Template
- Can be used with Chef, Puppet and Ansible, deployment automation and management tools for clouds

DataOps and MLOps: DevOps for Data Analytics and ML

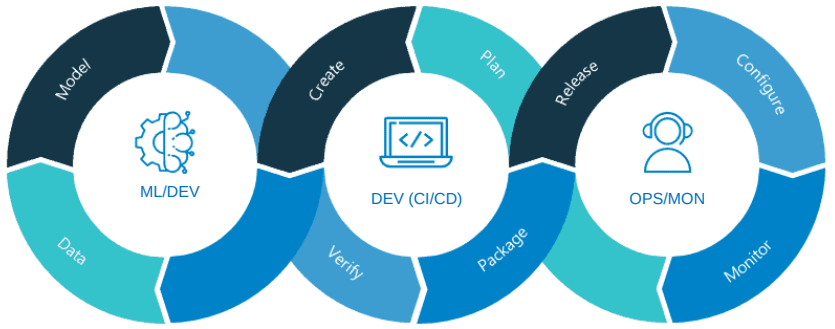
DataOps and MLOps are extension of **DevOps** to manage Data Analytics and Machine Learning data flow and process.

- DevOps is the practice of operations and development engineers participating together in the entire service lifecycle, from design through the development process to production support.
- Develop – Build – Deploy – Operate
- Cloud is an enabler for DevOps processes

DataOps, MLOps is about operationalizing ML and Data Analytics

- Different nature of processes at the stage of ML model development
- Benefit from the DevOps processes and culture
- Learn from DevOps experience

In fact, for the software product this is
ML + Dev (CI/CD) + Ops



DevOps Essentials (from Software Engineering)

- Better Software, Faster time to market
- Movement Comes from Open Source
- Synergy of Development and Operations
- Covers the *entire* Application Lifecycle

Data Science Process Models and Model Formats

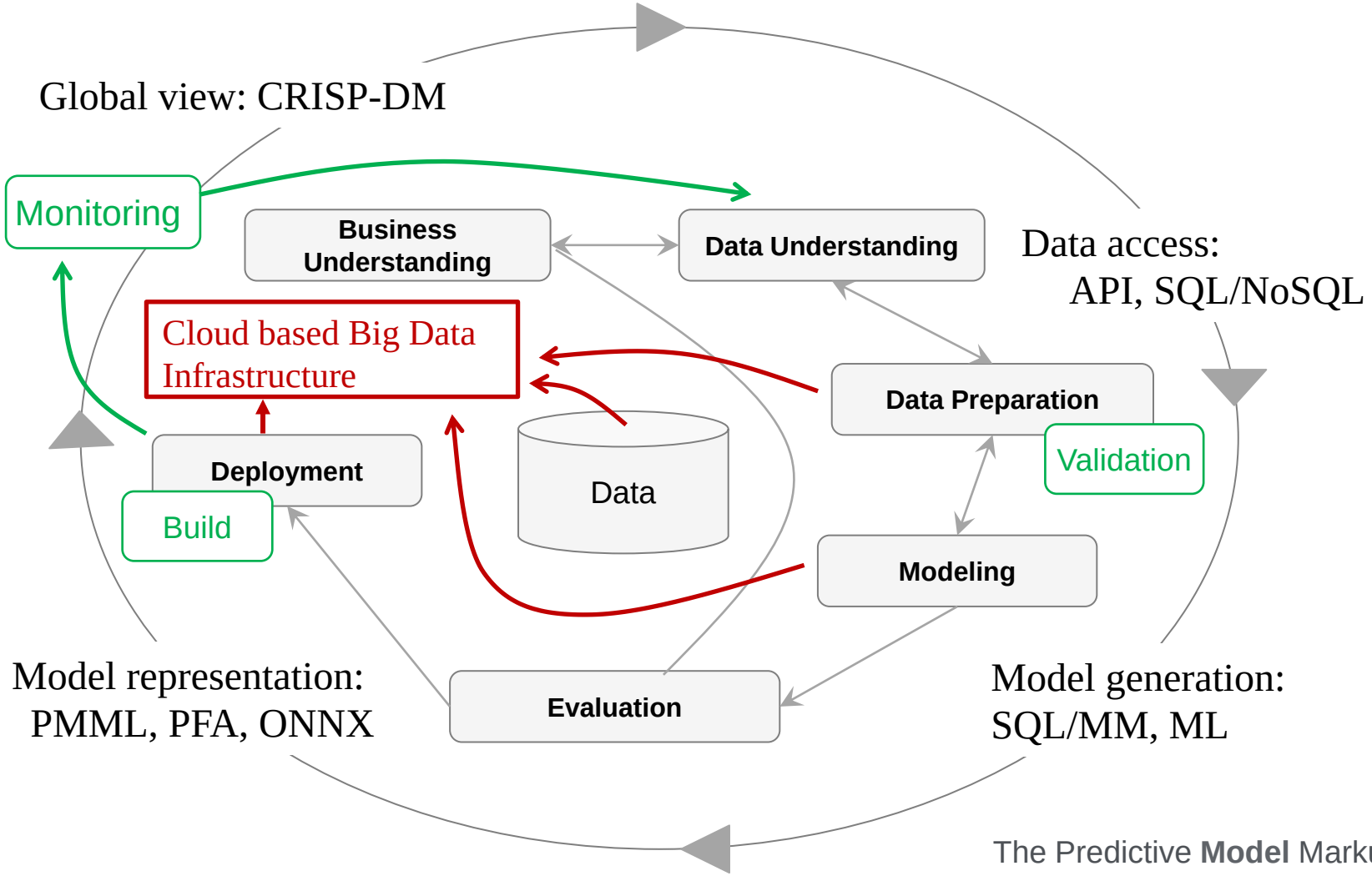
- Data Science process models
  - CRISP-DM, CRoss-Industry Standard Process for Data Mining
  - ASUM, Analytics Solutions Unified Method (IBM)
  - TDSP, Team Data Science Process (Microsoft)
  - KNIME Model Factory (KMF)

- Data Analytics Models Formats
  - Predictive Models Markup Language (PMML)
  - Portable Format for Analytics (PFA)
  - ONNX (Open Neural Network Exchange)
  - TensorFlow Model

What is CRISP-DM?

- Cross-Industry Standard Process for Data Mining
- Aim:
  - To develop an industry, tool and application neutral process for conducting Knowledge Discovery
  - Define tasks, outputs from these tasks, terminology and mining problem type characterization
- Founding Consortium Members: DaimlerChrysler, SPSS and NCR
  - CRISP-DM 2.0 (approx. 2008)
- CRISP-DM Special Interest Group ~ 200 members
  - Management Consultants
  - Data Warehousing and Data Mining Practitioners

CRISP DM Process: Processes and Data Lifecycle



Cross Industry Standard Process for Data Mining (CRISP-DM) model and stages

- Business understanding
- Data Understanding
- Data preparation
  - Data Validation
- Modelling
- Evaluation
- Deployment
  - Process monitoring

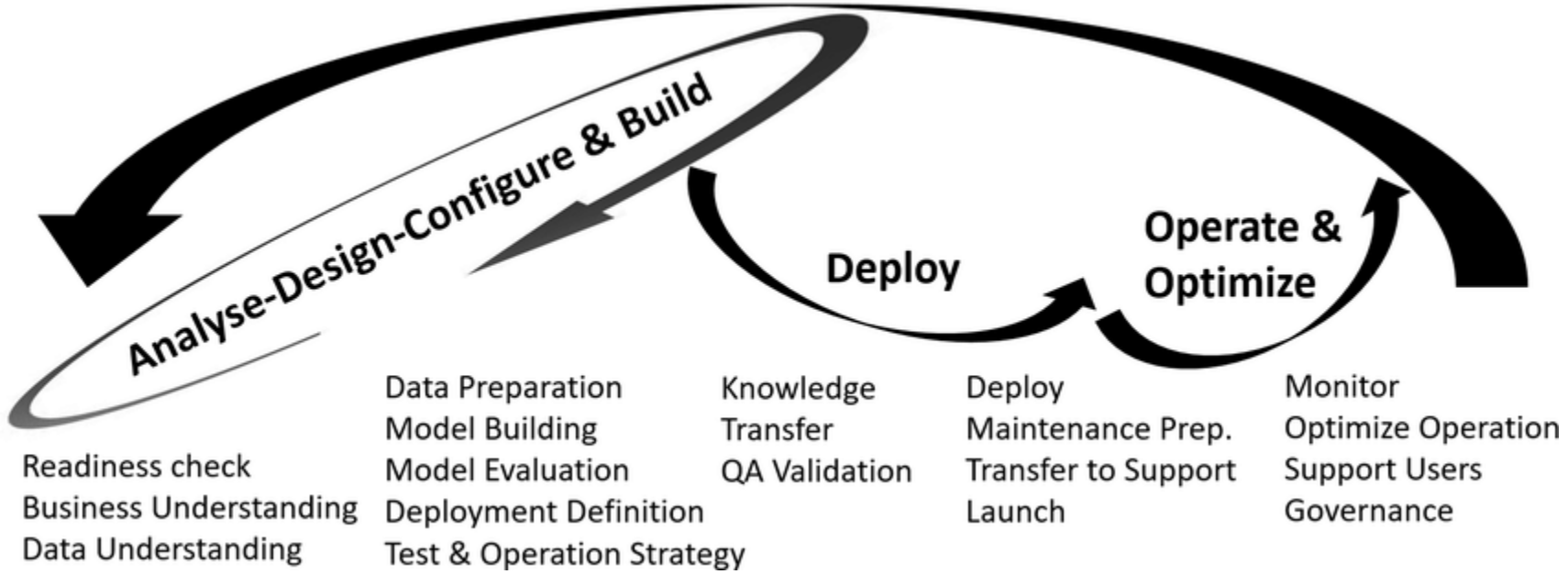
All stages are iterative with the goal to achieve effectiveness for business decision making

Phases, Tasks Generic and Special

The Predictive **Model** Markup Language (**PMML**)

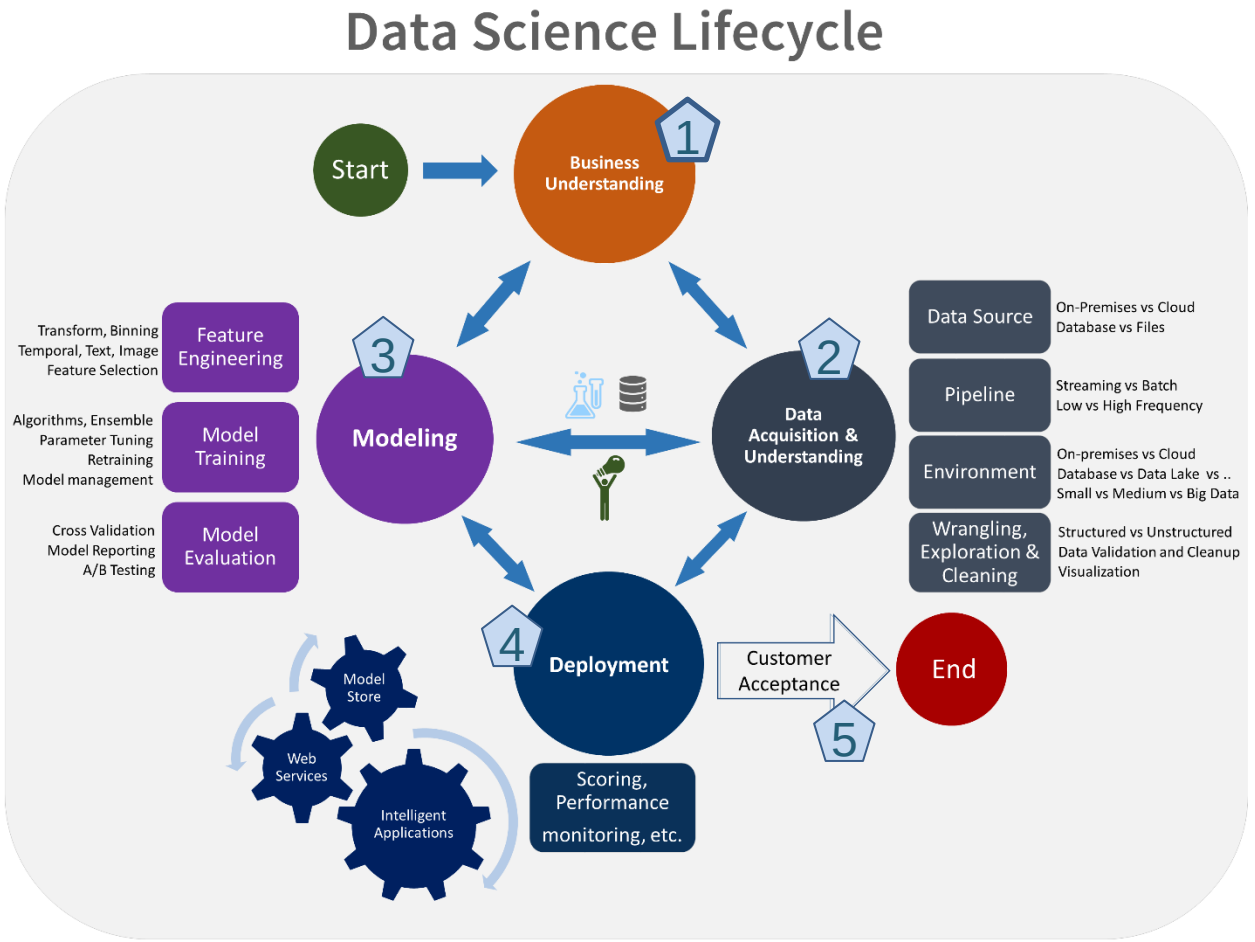
ASUM by IBM

- *Analyze*. Requirements specified and agreed; contract or services agreement is signed.
- *Design*. Define all components of the solution and their relationships and dependencies, identify necessary resources.
- *Configure and Build*. The solution is developed, all components are integrated and configured.
- *Deploy*. Create a plan to run and maintain the developed solution, including configuration management and migration plan if necessary.
- *Operate and Optimize*. The solution is operational is monitoring data are collected and maintained.



Team Data Science Process (Microsoft)

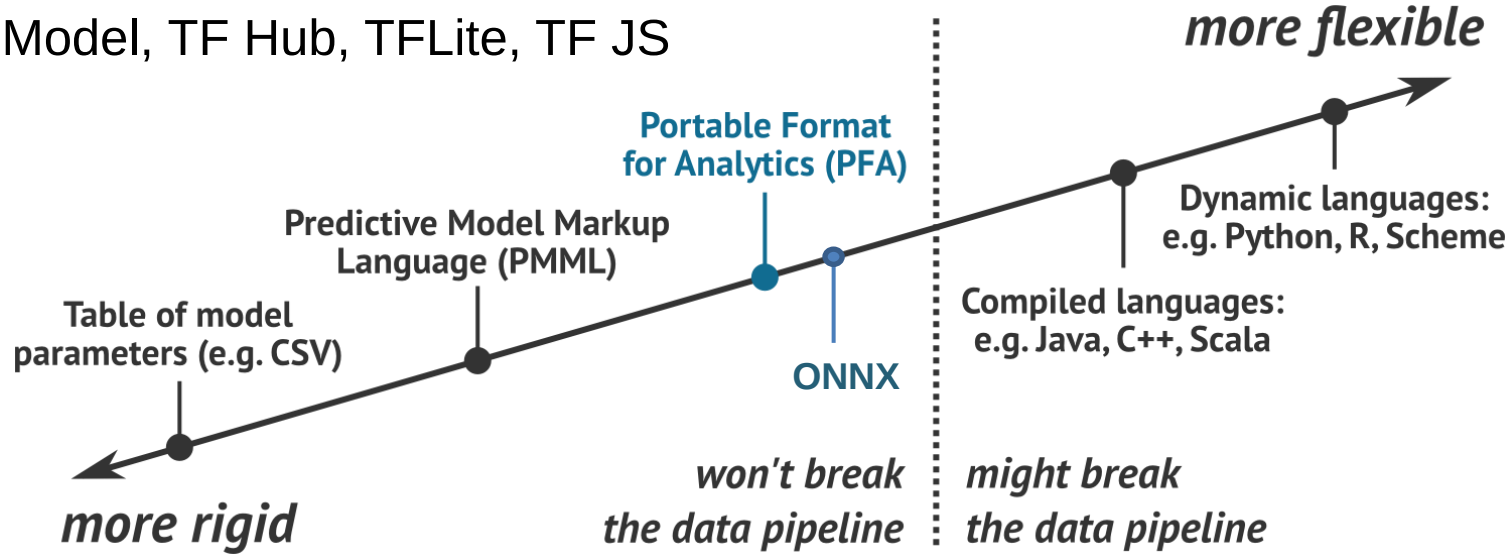
- TDSP is an agile and iterative process mod
  - Refactored into Azure DevOps in 2018
  - Currently supported by MLOps
- Includes components
  - A **data science lifecycle** definition
  - A **standardized project structure**
  - **Infrastructure and resources** recommended for data science projects
  - **Tools and utilities** recommended for project execution
- The lifecycle includes five sequential phase
  1. Business Understanding
  2. Data acquisition and understanding
  3. Modeling
  4. Deployment
  5. Customer acceptance



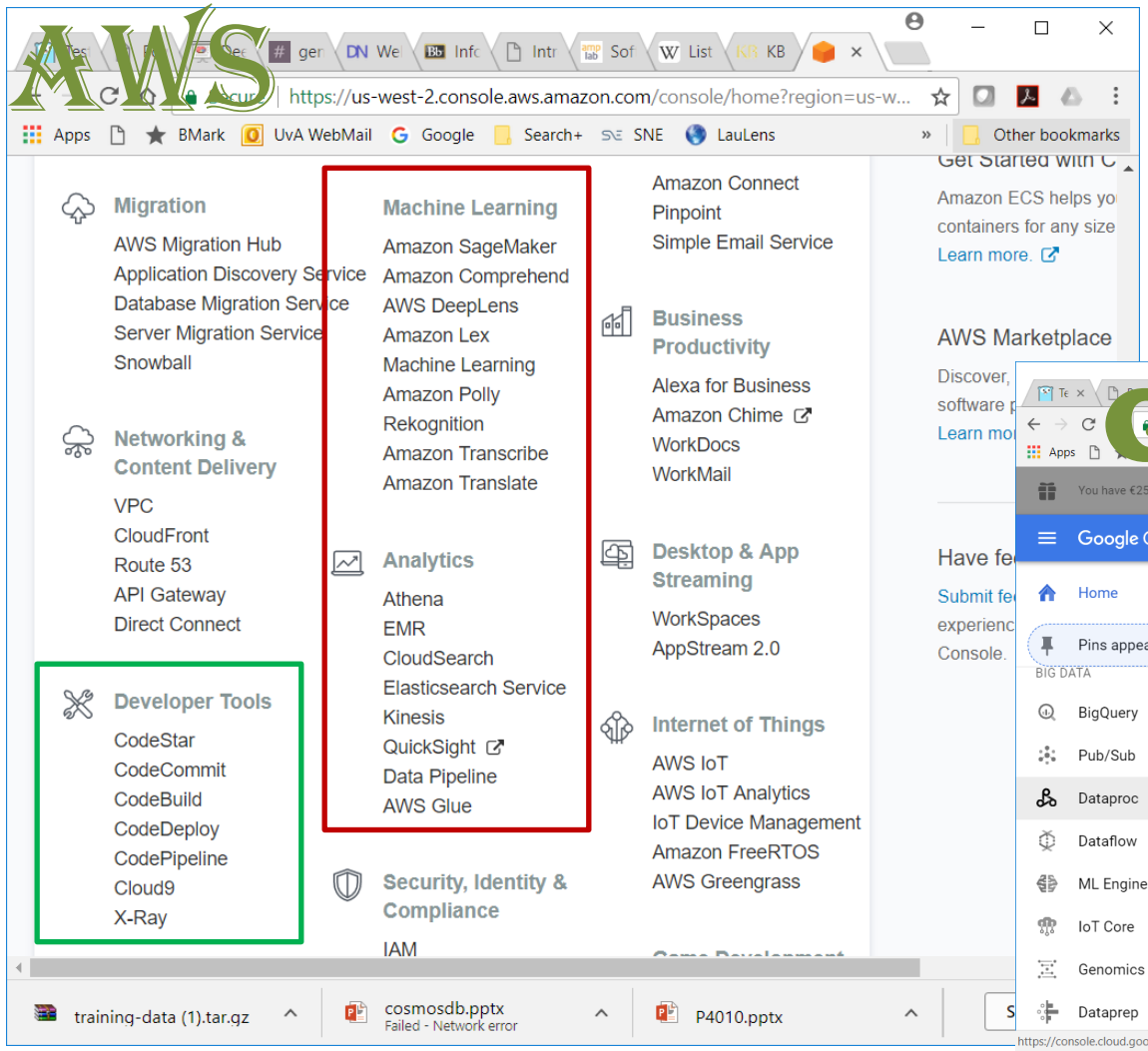
[ref] https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

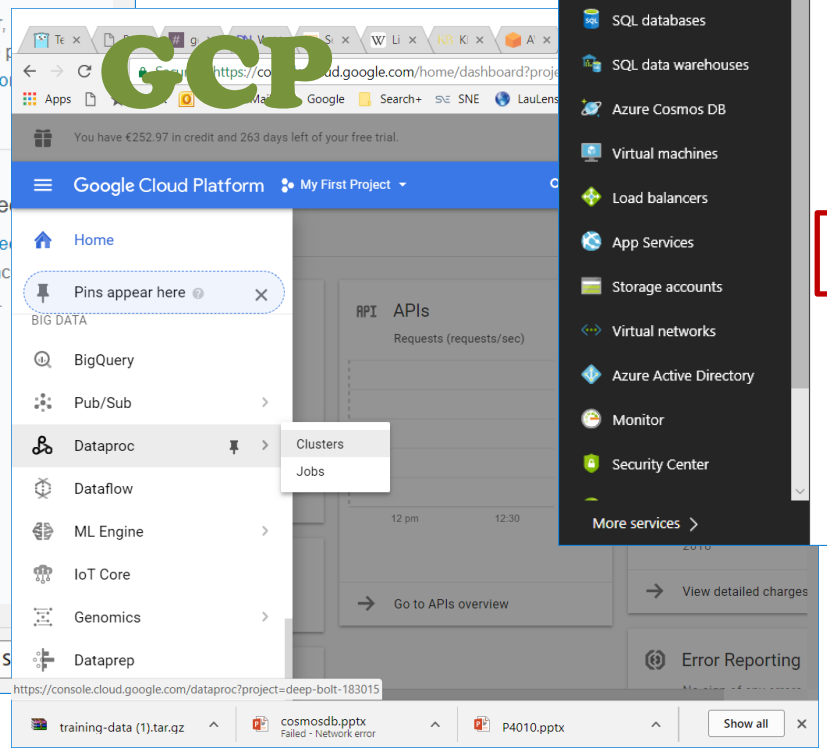
Data Analytics Models Formats – Step from ML Dev to Ops via CI/CD

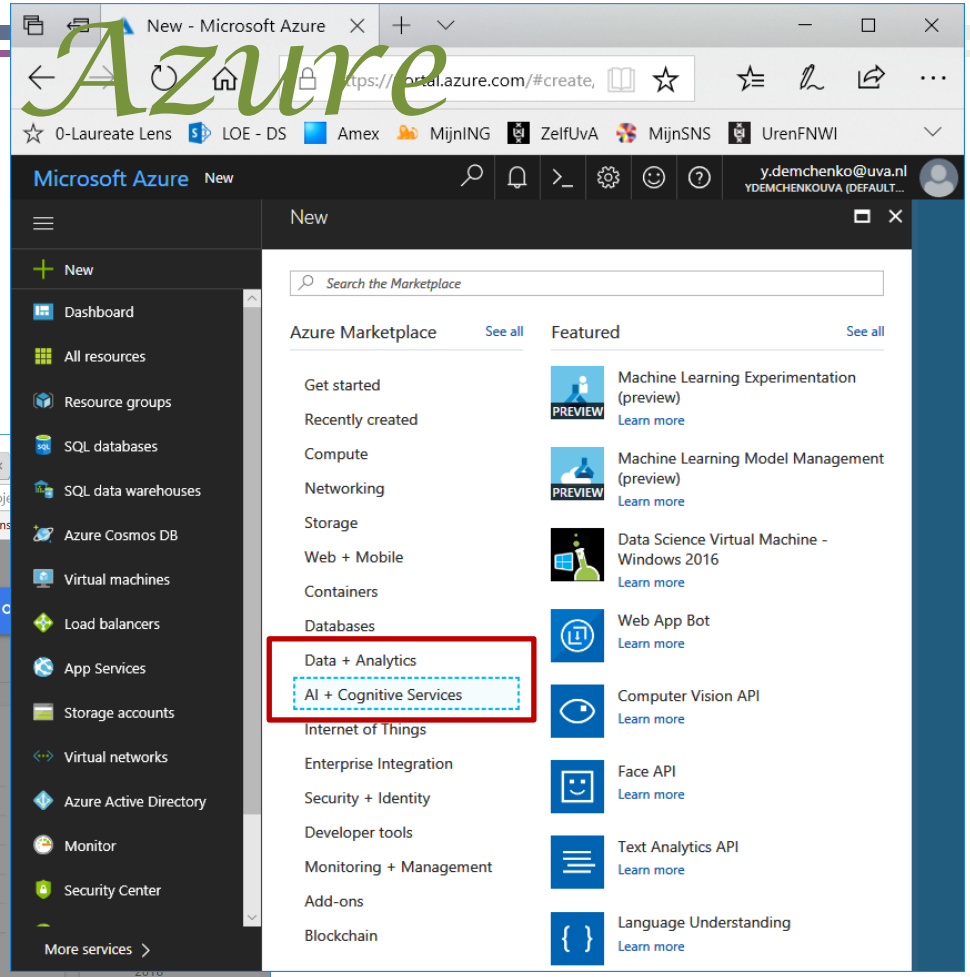
- **Predictive Models Markup Language (PMML)** that have benefits of transferring a developed model to production, access to coefficients
- **Portable Format for Analytics (PFA)**, an emerging standard for statistical models and data transformation engines to ease portability across systems with algorithmic flexibility by defining composable models, pre-processing, and post-processing functions that can be built into complex workflows
- **ONNX** (**Open Neural Network Exchange)** - an open format built to represent machine learning models. ONNX defines
  - Common set of operators - the building blocks of machine learning and deep learning models, and
  - Common file format to enable AI developers to use models with a variety of frameworks, tools, runtimes, and compilers.
- **TensorFlow Model** format: SavedModel, TF Hub, TFLite, TF JS



AWS, Azure, Google Big Data and ML Stacks





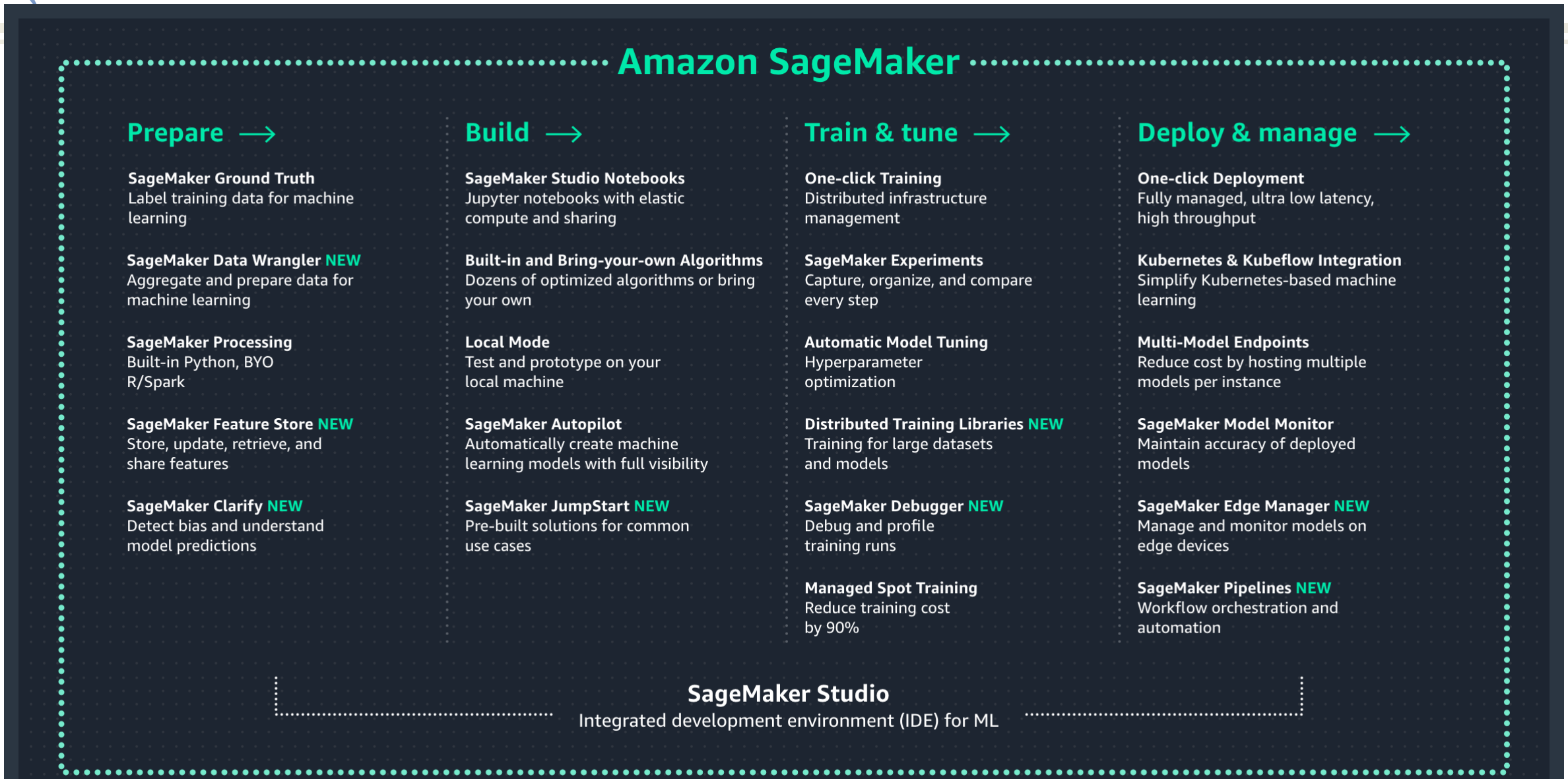


AWS MLOps Framework Components

- AWS CodeCommit
- AWS CodePipeline
- AWS CodeBuild
- AWS CodeDeploy

- AWS Sagemaker – Integrated IDE for ML
  - **Prepare - Build – Train&Tune – Deploy&Manage**
- AWS SageMaker Tools
  - **SageMaker Neo** enables machine learning models to train once and run anywhere in the cloud and at the edge
    - Supports Gluon, Keras, MXNet, PyTorch, TensorFlow, TensorFlow-Lite, and ONNX models
  - **SageMaker Model Monitor** is a capability of Amazon SageMaker that continuously monitors machine learning (ML) models in production, detects deviations such as data drift that can degrade model performance over time, and alerts you to take remedial actions.

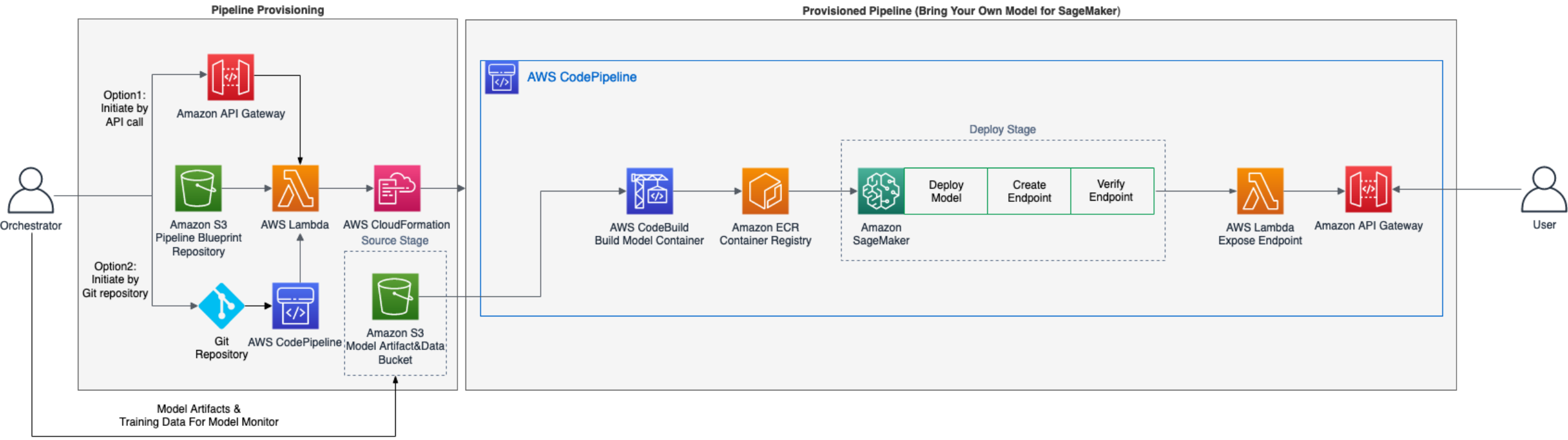
SageMaker Studio – Jupyter Notebook based



[Ref] https://aws.amazon.com/sagemaker/

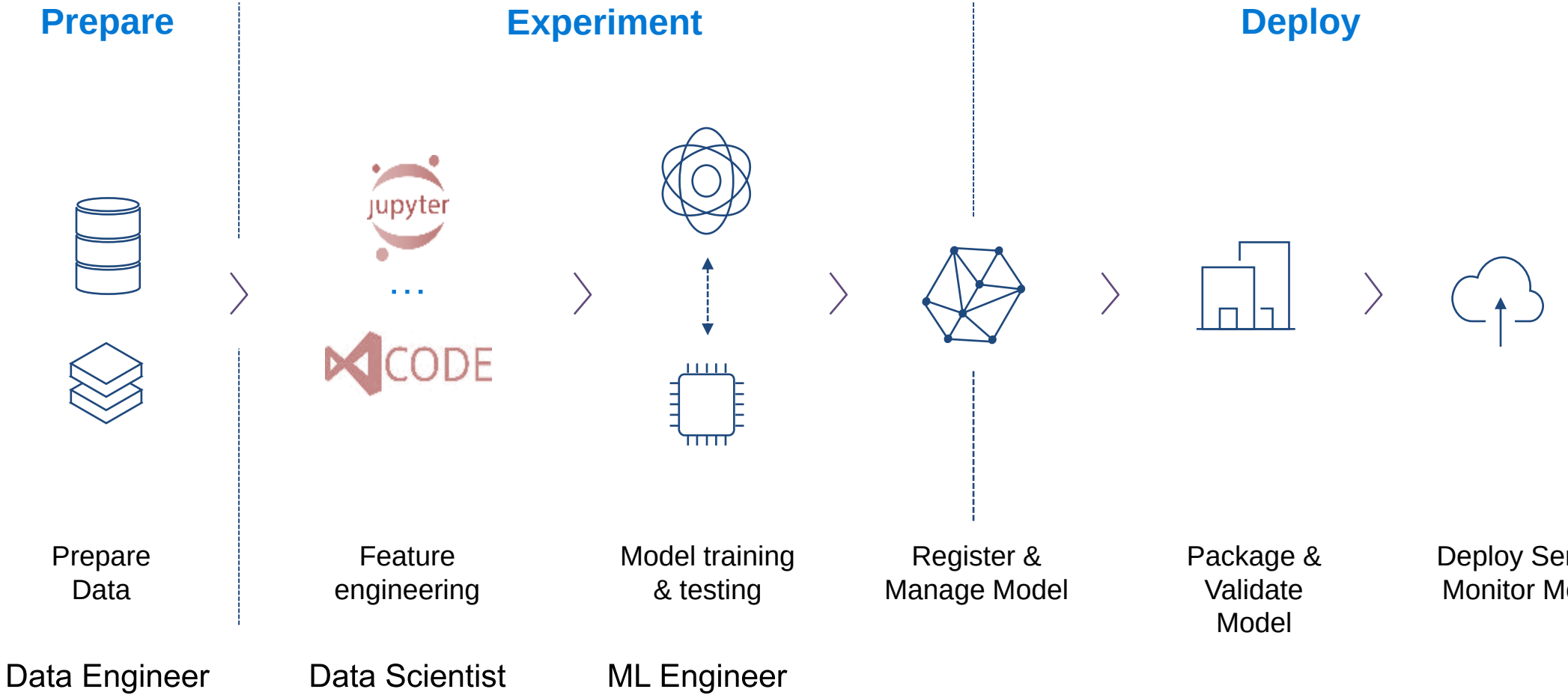
AWS MLOps Framework

https://aws.amazon.com/solutions/implementations/aws-mlops-framework/#



- MLOps pipeline provisioning CodePipeline and SageMaker for Model deployment
  - Using CloudFormation template - https://s3.amazonaws.com/solutions-reference/aws-mlops-framework/latest/aws-mlops-framework.template
  - Source code https://github.com/awslabs/aws-mlops-framework
  - Container based – using ECR (Elastic Container Registry)
- Configured for using AWS CDK - Code Development Kit (https://aws.amazon.com/cdk/), can also integrates with Terraform
- SageMaker Neo enables machine learning models to train once and run anywhere in the cloud and at the edge

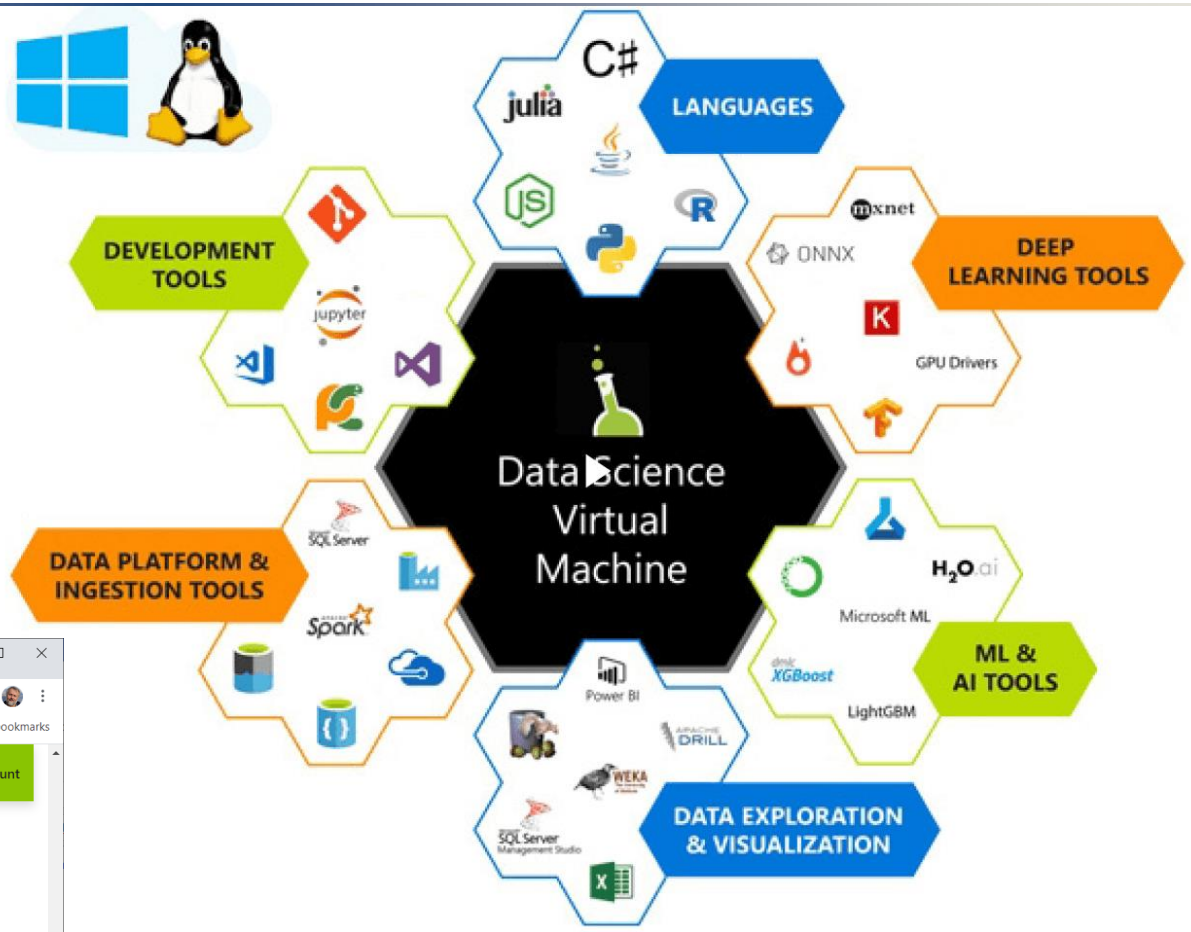
Azure MLOps: Data Science Workflow
Empowering Data Science Process with DevOps and MLOps

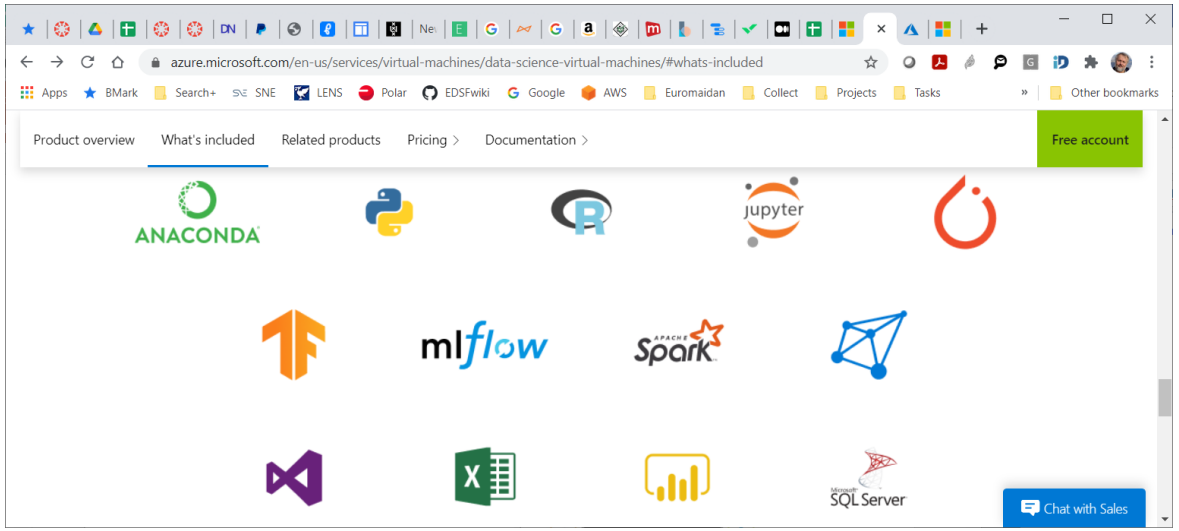


Azure – Data Science Virtual Machine (DSVM)

https://azure.microsoft.com/en-us/services/virtual-machines/data-science-virtual-machines/

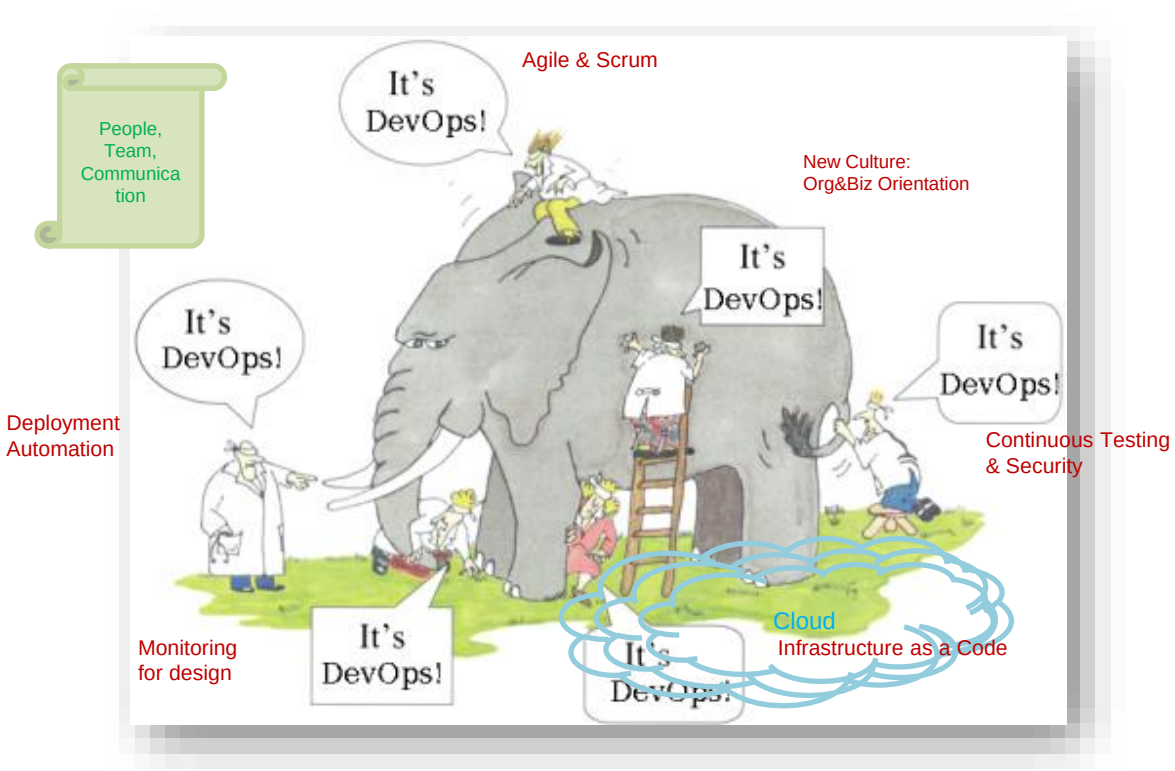
- Pre-Configured virtual machines in the cloud for Data Science and AI Development
  - Python, R, ONNX, + Microsoft AutoML
  - https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/tools-included
- Standard Configuration
  Standard_DS3_v2 – 4 vcups – 1 GiB
  - @ 328.12 USD/mo
- Expected to be used as a virtual desktop
- DSVM configuration
  - **mchmin :: BDIT****ndows

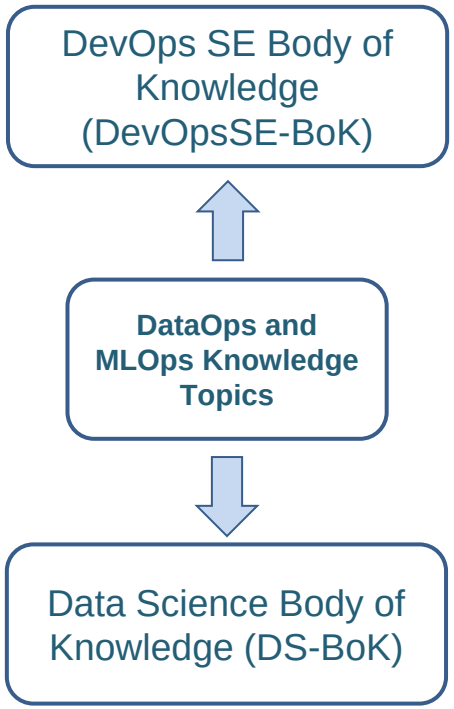




Teaching DataOps: Cross-domain Competences, Skills, Knowledge

- DevOps and Cloud Based Software, Software Engineering at the University of Amsterdam [online] https://studiegids.uva.nl/xmlpages/page/2020-2021-en/search-course/course/80049
- Big Data Infrastructure Technologies for Data Analytics at NTUU “Kyiv Polytechnic Institute” http://pma.fpm.kpi.ua/uk/students/courses
- SLICES Research Infrastructure design and training https://slices-ri.eu/





Summary and take away

- With the rapid development of Data Science Analytics and ML applications demand for agile technologies and processes increases – with focus on the full applications and data lifecycle
- DataOps and MLOps is based on applying DevOps principles to Data Science Analytics and ML projects
  - Both domains are well developed, and their fusion facilitates operationalization of DSA and AI/ML
- Data driven research and applications require wider processes extensions “left” and “right” to development and operational stage – Site Reliability Engineering (SRE) practices to be assessed
- Emerging MLOps and DataOps open new professional possibilities for SE and DevOps engineers
  - DataOps Knowledge topics extension for DevOpsSE-BoK and DS-BoK

- Exploring benefits of OpenStack ecosystem for data driven research and education

Discussion and Questions

- Research
- Development
- Skills Management, Teaching and Training

- Extended version
  http://www.uazone.org/demch/presentations/http://www.uazone.org/demch/presentations/openinfra2021-dataops-datascience-operationalising-v03-extended.pdf