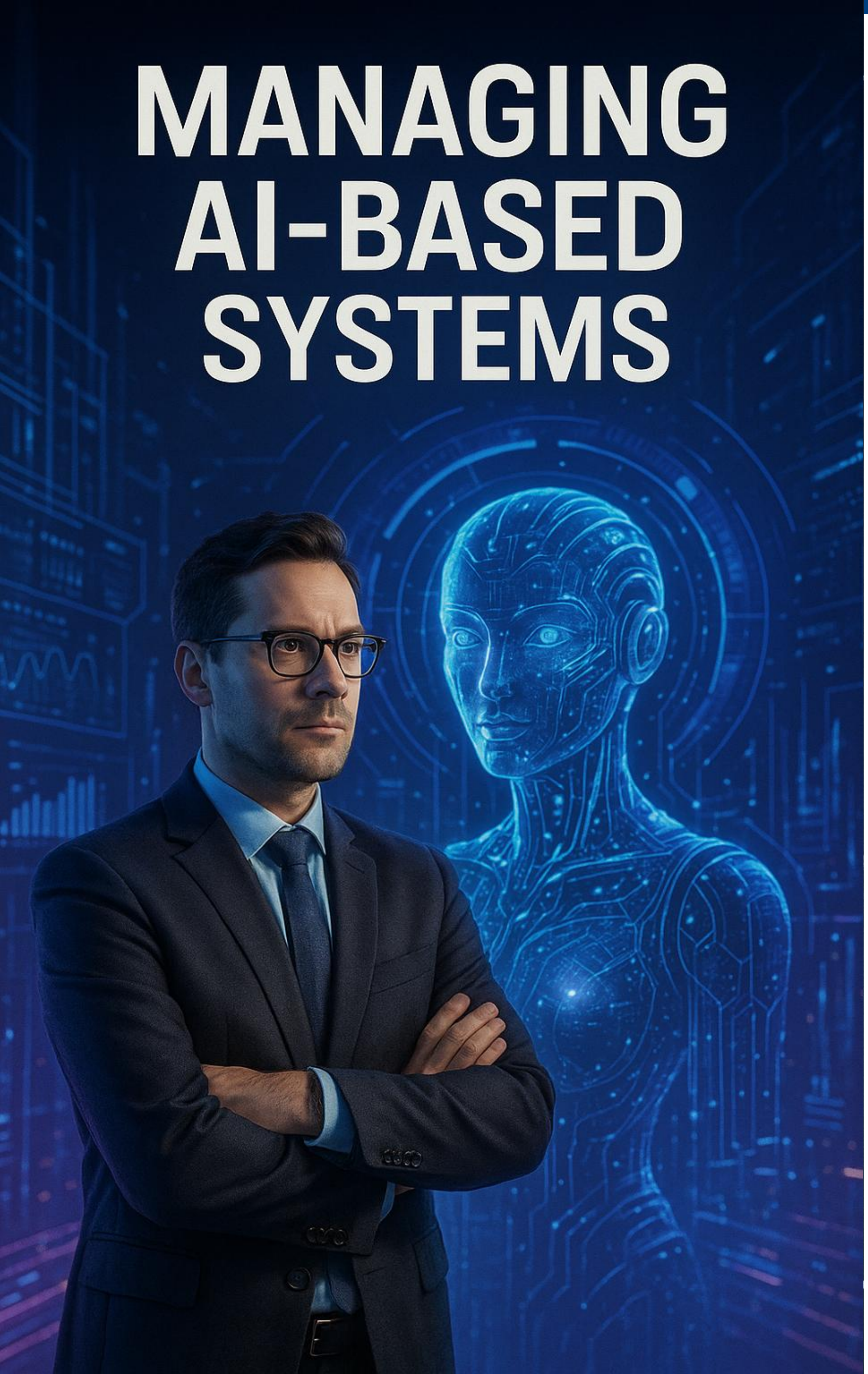


FRANKFURT UNIVERSITY OF APPLIED SCIENCES

dit lab
Research Lab for Digital Innovation & Transformation

AI FOR BUSINESS
BUSINESS FOR AI

Session 9: Architectures of AI applications

Managing AI-based Systems

Prof. Dr. Nils Urbach

Frankfurt University of Applied Sciences,
Research Lab for Digital Innovation & Transformation

FIM Forschungsinstitut für Informationsmanagement

Fraunhofer-Institut für Angewandte Informationstechnik FIT,
Institutsteil Wirtschaftsinformatik

www.ditlab.org
www.fim-rc.de
www.wirtschaftsinformatik.fraunhofer.de

Creative Commons Copyright

Course navigator

2
3
4
Basic understanding of AI and its business potential

AI Ideation

5 Application potentials of AI technologies

6 Identification, design and evaluation of AI use cases

AI Strategizing

7 Evaluation an organization's AI readiness

8 Management and governance of AI

AI Design & Development

9 **Architectures of AI applications**

10 Data Management and Model Transparency

11 Design of human-AI interactions

AI Operations at Scale

12 Monitoring and KPI-based control

13 Ethical, legal and social implications of AI

Implementation

Objectives of today's lecture

AI FOR BUSINESS
BUSINESS FOR AI

1. Understand the ML decision space
2. Learn how to select appropriate KPIs for AI initiatives
3. Understand the differences between data, training and deployment infrastructures

Agenda

Agenda

01 | **Knowing and understanding ML decision space**

02 | Data, training and deployment infrastructure

03 | Latest GenAI architectures

04 | From DevOps to MlOps

ML decision spaces

When introducing AI and machine learning in businesses, decision spaces refer to the **set of possible options and choices within the ML framework**. Companies must navigate these decision spaces to make informed choices at various stages of the ML process.

Model selection:

- Consider different ML algorithms, architectures, and hyperparameters choosing the best model for a given task
- Assess trade-offs between model robustness, reusability, interpretability and performance, ensuring the selected model aligns with the specific business needs

Train & retrain process:

- Decision to retrain a once trained model due to changes in data distribution or business requirements, necessitating retraining to adapt the model to the updated situation.
- Trade-off between the cost-saving usage of a model trained for a similar operation and the better-performing training of a new model

Ashmore et al. (2021)

Goals of an AI model

AI FOR BUSINESS
BUSINESS FOR AI

Performant

- Considers quantitative performance metrics applied to the model when deployed in a system.
- E.g., receiver operator characteristic (ROC), mean squared error, classification accuracy

Robust

- Considers the model's ability to perform well in circumstances where the inputs encountered at runtime are different from those present in the training data.
- E.g., an image recognition model trained on supermarket fruits accurately identifies those at a farmer's market

Reusable

- Considers the ability of (components of) a model to be reused in systems for which they were not originally intended.
- E.g., facial recognition in an authentication system may have features that can be reused to identify operator fatigue

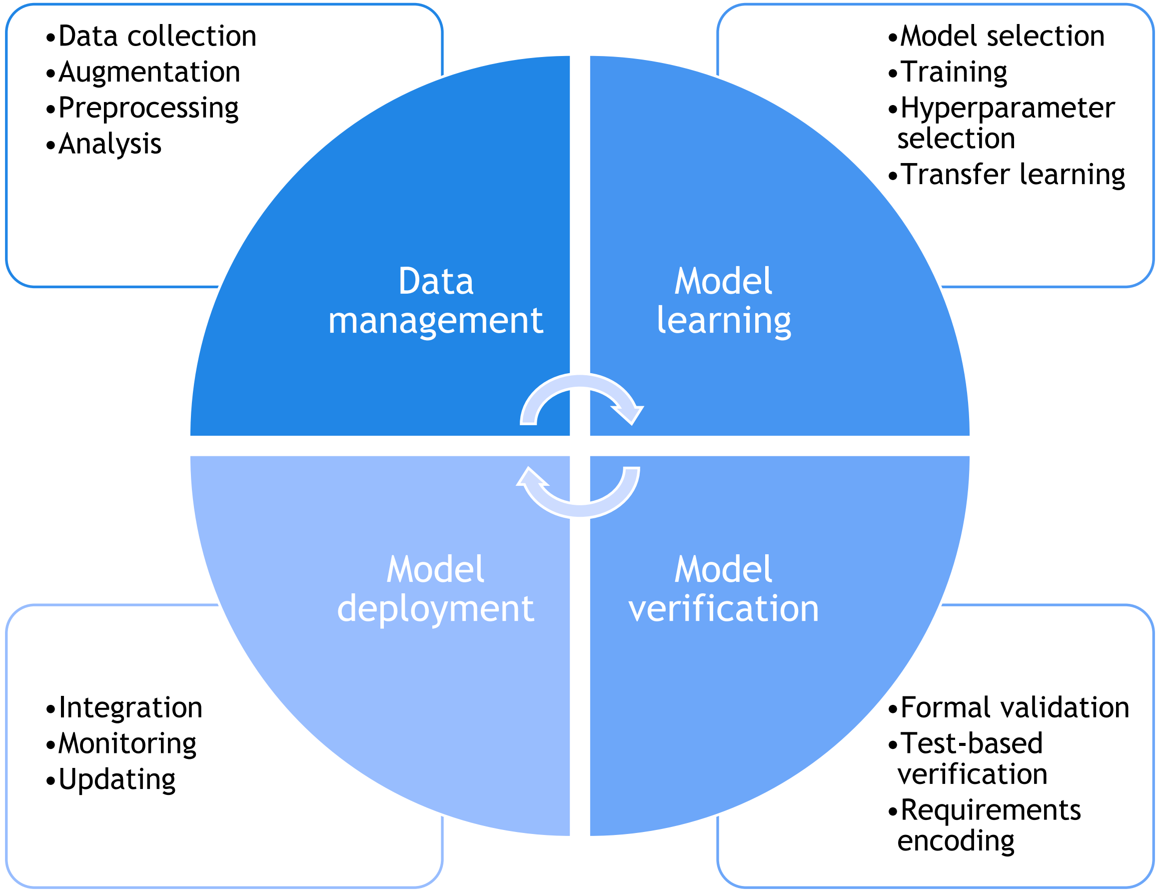
Interpretable

- Considers the extent to which the model can produce artefacts that support the analysis of its output, and thus of any decisions based on it.
- E.g., a decision tree may support the production of a narrative explaining the decision to hand over control to a human operator

Ashmore et al. (2021)

The decision space spans over all stages of the machine learning lifecycle

AI FOR BUSINESS
BUSINESS FOR AI



Ashmore et al. (2021)

Model learning activities

AI FOR BUSINESS
BUSINESS FOR AI

Model selection

Decision about the model type, variant, structure of the model to be produced in the model learning stage

Hyperparameter selection

Selection of the parameters associated with the training activity, controlling the effectiveness of training and performance of the resulting model.

Training

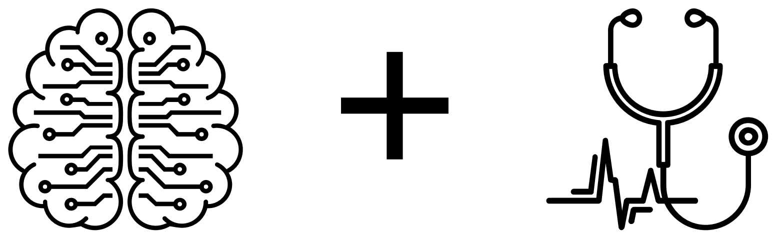
Optimization of the ML model performance with respect to an objective function that captures the specific requirements and goals for the model.

Transfer learning

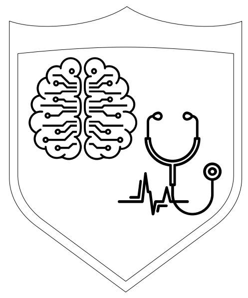
Reuse of a once trained model or use it as a starting point to retrain it as a second model, significantly reducing the training time and costs.

Ashmore et al. (2021)

Assurance methods for the model learning stage



The use of ML in safety-critical applications requires a high level of security.



Assurance of ML is about finding evidence that ML is sufficiently safe for its intended use.

Associated Activities Method	Model Selection	Training	Hyperparam. Selection	Transfer Learning
Performance measures	✓	✓		
Ensemble learning	✓	✓		✓
Batch normalization		✓	✓	
Model zoos	✓	✓		✓

The different methods provide evidence across the model learning stage so that the model is safe for its application. Assurance methods also exist for the data management stage, the model verification stage and the model deployment stage.

Ashmore et al., 2021

Assurance methods for the model learning stage

1

Performance measures

- Used to compare and evaluate the performance of models
- Examples: accuracy, recall, specificity

2

Ensemble learning

- Combining multiple models to create model with superior performance
- lower bias and less variance errors

3

Batch normalization

- Penalizes parameter sensitivity to increase robustness

4

Model zoos

- Making use of models that were used before in a related area

Ashmore et al., 2021

Train & retrain AI models

1. Train

Initial training of a machine learning model.

During this phase, the model is exposed to a set of training data to **learn patterns and relationships** within the data.

The raw model **adjusts its parameters** based on the given data to make predictions or classifications on new, unseen data in the future.

2. Retrain

Retraining an already trained model to adapt it to new data or changed requirements.

Model retraining **does not change the number of parameters and variables used** in the model.

It **adapts the model weights to the current data** so that the model gives better outputs.

Dilmegani (2023)

Retraining

Why?
Model drift – decreasing accuracy of predictions

Concept Drift

Happens when the data that a model was trained on are no longer representative. This can be due to changing user behaviors, shifts in the economy, environmental changes, new goals, etc.

Data Drift

Data drift refers to the divergence between the data used in production and the data initially employed for testing and validating a model prior to its deployment.

What data?

Offline learning/batch

If a concept drift has occurred and the old dataset does not reflect the new environment, it is better to replace the entire dataset, even if this is costly.

Online learning

If there's a constant stream of new training data, online learning continuously retrains the model by setting a time window that includes only new data.

When?

Periodic retraining

This is useful when underlying data changes within measurable time intervals. Frequent retraining can be computationally costly so determining the correct time interval is important.

Trigger-based retraining

This method involves determining performance thresholds. Models can be retrained automatically when the model's performance drops below this threshold.

Dilmegani (2023)

Agenda

Deployment models for IT Infrastructures

AI FOR BUSINESS
BUSINESS FOR AI

Edge Computing

- Deploys applications, data, and processing at the logical extremes of a network
- Brings resources and services **closer to the data-generating sources**
- Aim: reduce latency and improve scalability by processing data locally instead of sending it to centralized data centers

On-Premise Hosting

- Traditional approach: organization owns, manages, and maintains IT infrastructures within its **physical location** (data centers, local servers)
- Offers full control over data and applications (suitable for sensitive data and regulatory compliance)

Cloud Computing

- Enables ubiquitous, on-demand network access
- Shared pool of configurable computing resources.
- Rapid provisioning and release with minimal management effort
- Access to networks, servers, storage, applications, and services

Escamilla-Ambrosio et al. (2018);

Edge computing providers

AI FOR BUSINESS
BUSINESS FOR AI

Edge computing solutions, depending on the requirements

Best allrounder: Amazon Web Services

Intelligence at the edge: Microsoft Azure

IoT at the edge: ClearBlade

Analytics, management, scaling, and optimization at the edge: Dell Technologies

Edge data centers: EdgeConneX

Edge deployment of containerized applications: Section

enterprisenetworkingplanet.com (2023)

On-Premise as a Service: A hybrid alternative

AI FOR BUSINESS
BUSINESS FOR AI

Blend of traditional on-premise and off-premise public cloud computing

Key Features

- Data kept on-site like traditional on-premise data center
- Eliminates substantial upfront costs for IT infrastructure
- Pay-as-you-go model for storage capacity

Distinguishing Factors

- Not a traditional data center setup managed solely by in-house IT
- Third-party service provider owns and manages on-site hardware and equipment

Cost Structure

- Replaces upfront capital expenses with operating expenses
- Payment for actual storage capacity consumed

Service Provision

- More than a lease – comprehensive service offering
- Includes storage, computing, networking, expertise, and technical support

Flexibility and Scalability

- Allows adjustment of service capacity based on business progression
- Retains the benefits of on-site data control with added flexibility

zadara

First OPaaS provider

DELL Technologies

On-Premise Infrastructure

Lenovo

TruScale
On-Premise Infrastructure

vallous.com (2022)

Cloud computing providers

Software-as-a-Service

Provision of applications that run on a cloud infrastructure and are accessible from client devices

SAP S/4 HANA

ERP

HubSpot

CRM Software

Platform-as-a-Service

Provision of computing power, storage space, networks and other basic computer resources

App Engine

Application Dev System

Azure

Runtime Environment

Infrastructure-as-a-Service

Provision of programming languages, libraries, services and development tools

Amazon EC2

Computing Capacity

Amazon S3

Storage

Mell and Grance (2011)

Data infrastructure

Data infrastructure refers to the systems, tools, and resources required for data capture, storage, processing, and management. Data is crucial for the success of AI and machine learning applications as these models need to be trained on large volumes of high-quality data to achieve effective results.

Edge Computing

- Data captured and processed at or near the source
- Reduces data traffic to centralized locations
- Enables faster data processing and response
- Suitable for real-time applications

On-Premise Hosting

- Physical location owned by the organization
- Data stored and processed locally
- Provides full control and ownership of data

Cloud Computing

- Data resources located in one or more data centers worldwide
- Data filed and processed remotely in the cloud
- Accessible from anywhere with internet connectivity

Escamilla-Ambrosio et al. (2018);

Training infrastructure

The training infrastructure includes the resources that are needed for the development and training of AI and machine learning models. Training a model requires intensive computing power and storage resources as it involves processing large amounts of data and performing complex calculations.

Edge Computing

- Not practical due to limited computational and storage capabilities of edge devices

On-Premise Hosting

- Suitable for model training with powerful hardware and infrastructure
- Requires investment in GPUs, servers, or specialized hardware

Cloud Computing

- Offers dedicated resources for AI model training
- Powerful GPUs and TPUs available for high-performance computing
- Provides high scalability, allowing resources to be adjusted as needed

Escamilla-Ambrosio et al. (2018);

Deployment infrastructure

Deployment infrastructure refers to the environment where a trained model is put into production to make real-time predictions or decisions.

Edge Computing

- Trained model runs directly on the edge device/infrastructure
- No continuous connection to external infrastructure needed
- Enables fast and low-latency inference for real-time applications

On-Premise Hosting

- Preferred for local deployment of trained models
- Ensures data sovereignty, compliance, and security
- Satisfies regulatory requirements and data privacy concerns

Cloud Computing

- Trained model runs on cloud servers
- Cloud performs inference and returns prediction results
- Prediction requests are sent to the cloud for processing

Escamilla-Ambrosio et al. (2018);

Agenda

01 | Knowing and understanding ML decision space

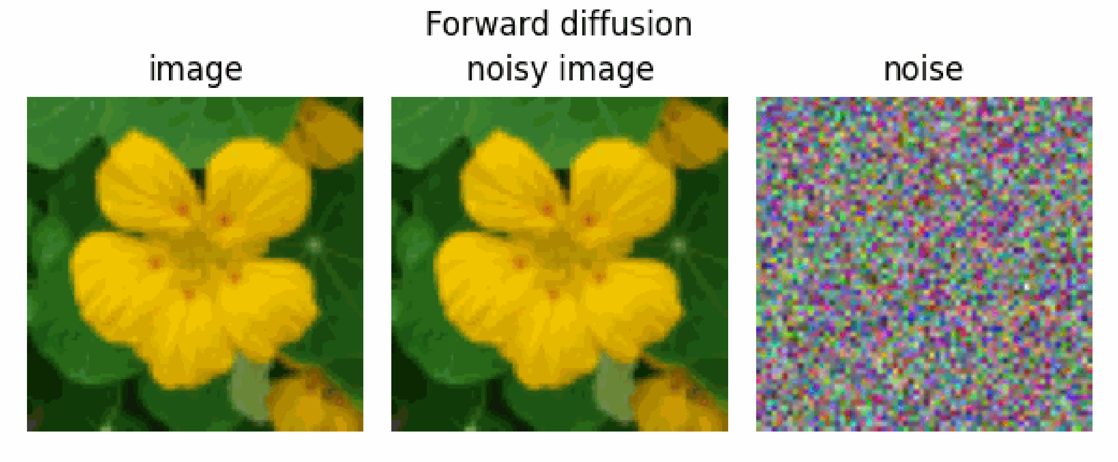
02 | Data, training and deployment infrastructure

03 | Latest GenAI architectures

04 | From DevOps to MlOps

Diffusion models help generate visual content, e.g., images

A **diffusion model** creates new data by gradually **adding and then removing** noise from a data sample, leading to high-quality and realistic outputs. Diffusion process gradually adds noise to an input (image) until only noise is left. Model learns to reverse the process, i.e., denoising starting from random noise to obtain image.

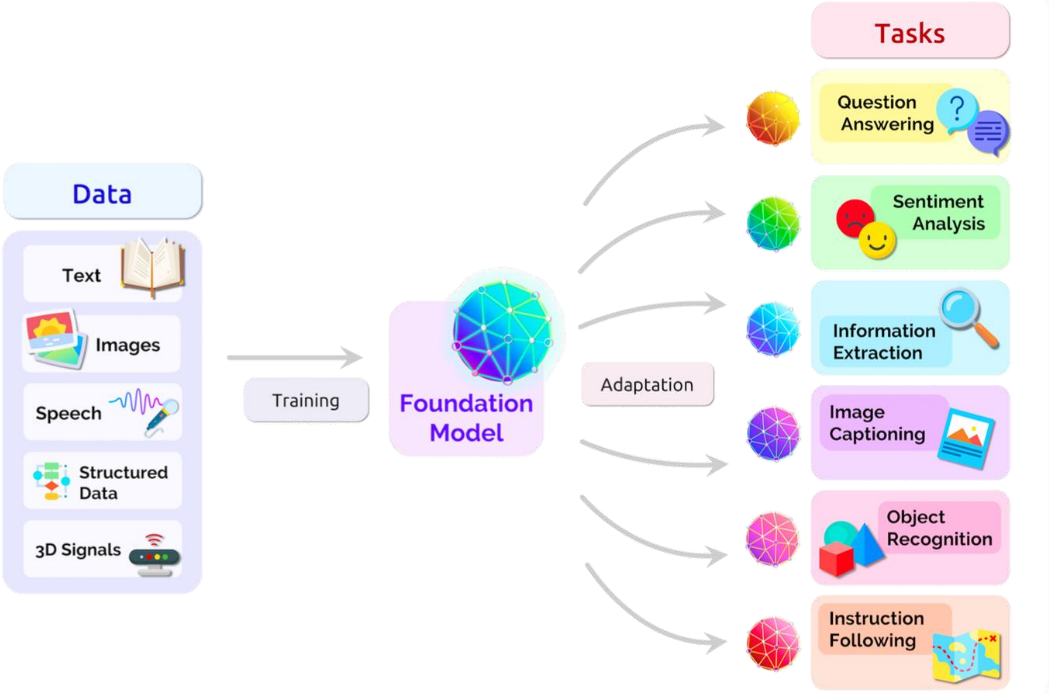


Application Areas: generate/optimize/augment synthetic data (e.g., time series analysis, generation of new molecular structures, etc.)

Béres (2022)

Foundation models are flexible and efficient in adapting to new tasks without retraining from scratch

A **foundation model** is any model that is **"trained on broad data** (generally using self-supervision at scale) that **can be adapted** (e.g., fine-tuned) **to a wide range of downstream tasks"**

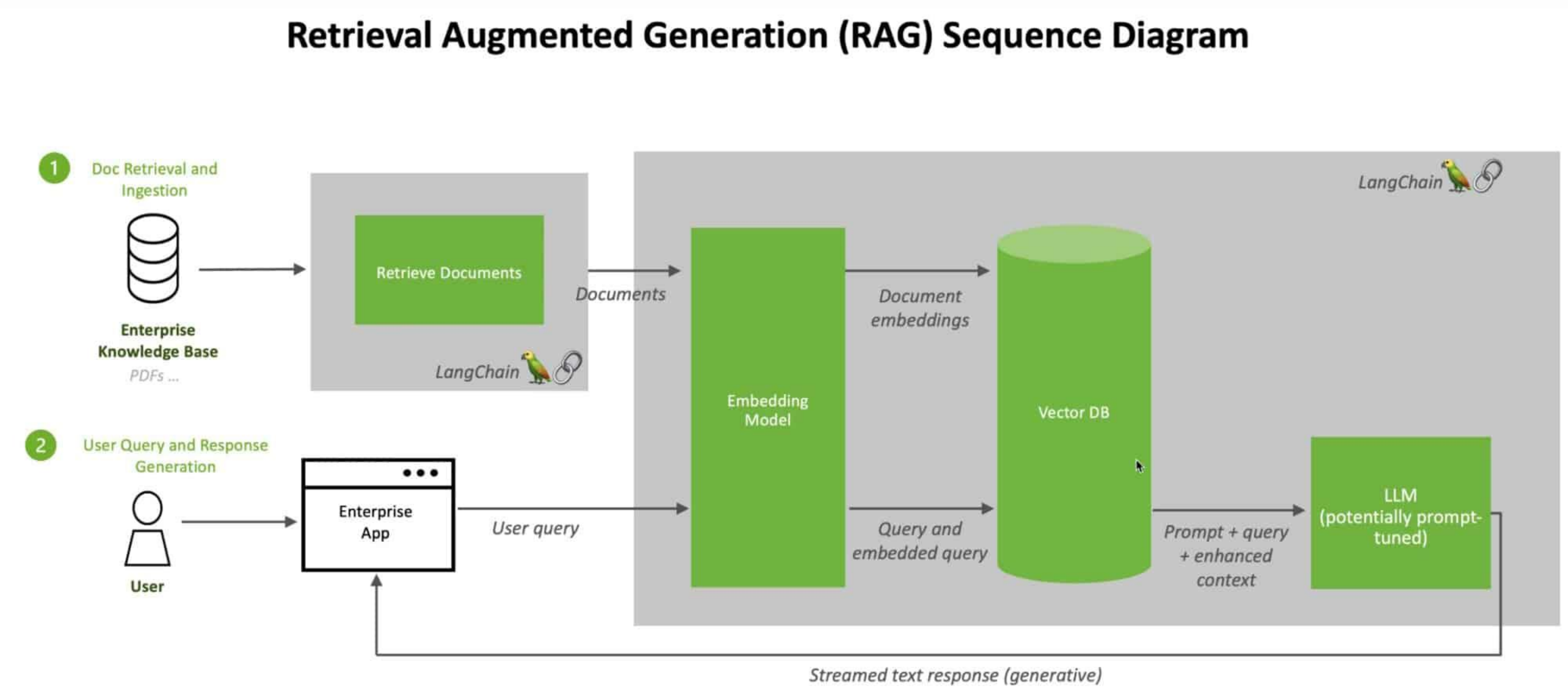


Application Areas: natural language processing, speech recognition, predictive text completion, knowledge extraction

Liang et al. (2021)

RAG can help reduce hallucinations and enable verifications of sources when using LLMs

Generative AI applications using Retrieval-Augmented Generation (RAG) integrate information into the generative process by fetching relevant information from a database in response to a query, producing more accurate, detailed, and contextually relevant responses.



Lewis (2020), Dataiku (2023), nvidia (2024)

Agenda

MLOps

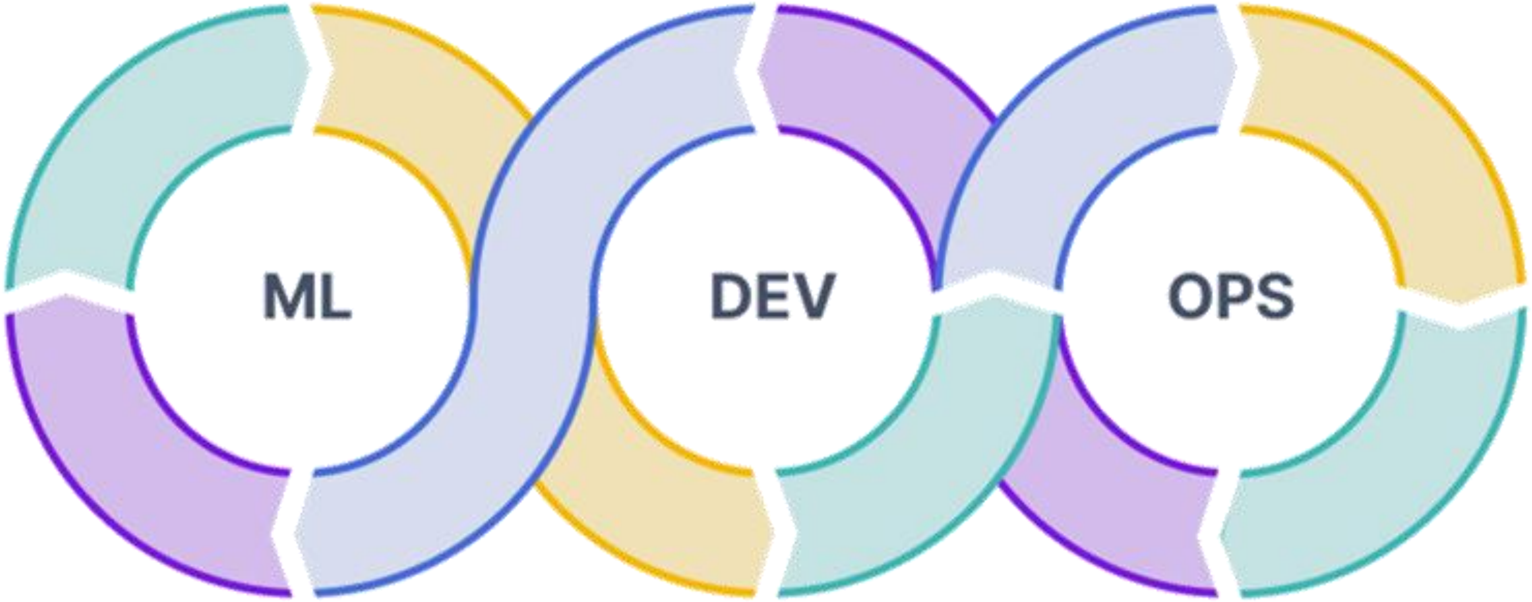


Image: https://www.v7labs.com/blog/mlops-machine-learning-ops-guide

AI-specific processes and organizational models: DevOps

AI FOR BUSINESS
BUSINESS FOR AI

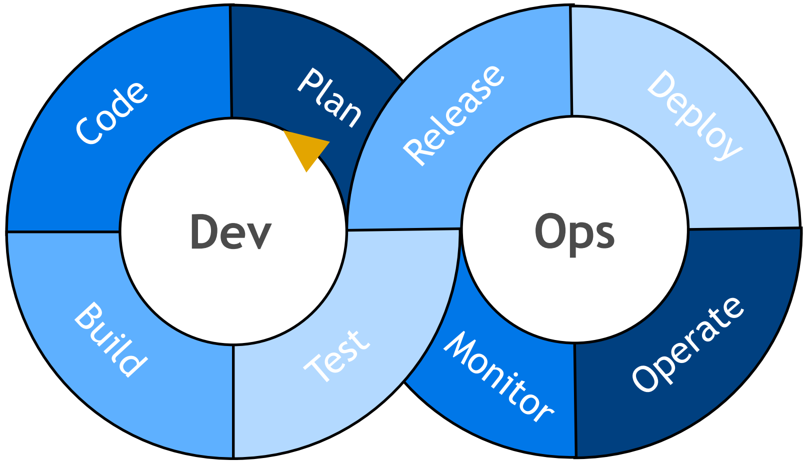
Defining DevOps

- Set of practices that integrates the traditionally distinct domains of **development and operations** through the automation of development, deployment, and infrastructure monitoring processes
- Paradigm shift within organizations, moving **away from siloed teams** that handle specific functions separately towards **cross-functional teams** that collaborate to achieve a **continuous flow of operational features and enhancements**

DevOps Lifecycle

Development

- **Plan:** Definition of requirements, initial execution planning
- **Code:** Coding according to agreed standards and best practices
- **Build:** Evaluation of the software artifacts
- **Test:** Ensuring the quality of the software artifacts



Operations

- **Release:** Releasing software after manual and automated tests
- **Deploy:** Focus on (re)deploying and the software continuously
- **Operate:** Maintaining and troubleshooting applications within a production environment
- **Monitor:** Ensuring stability

DevOps facilitates to build, test and deploy software and therefore reduces the time to market

Ebert et al. (2016), Subramanya et al. (2022)

AI-specific processes and organizational models: MLOps

AI FOR BUSINESS
BUSINESS FOR AI

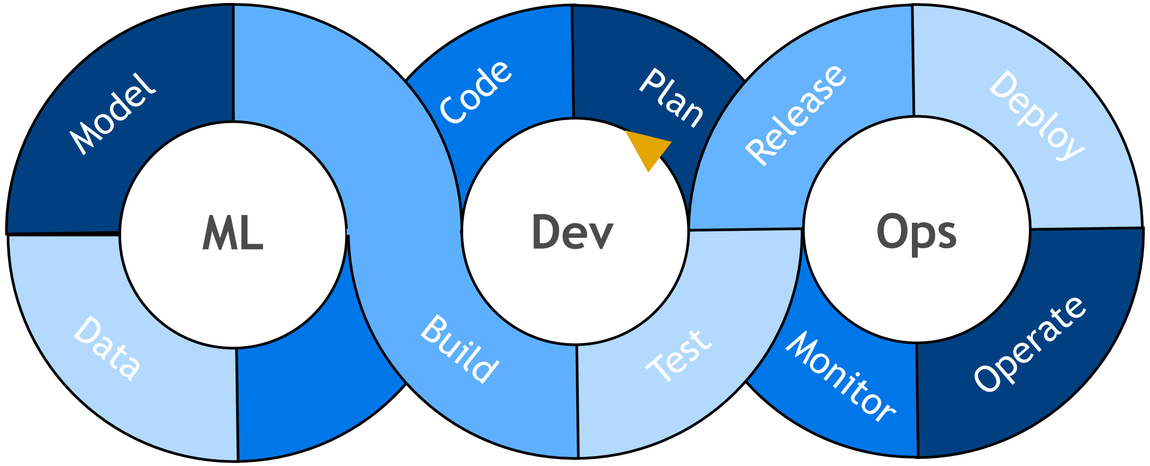
Defining MLOps

- Machine Learning Ops is a set of practices to ensure the consistent and efficient maintenance and deployment of machine learning code and their models
- Just like in DevOps, the goal is to achieve a shorter code-build-deploy loop but with a **focus on the fast development and deployment of ML models** with **high quality**, **reproducibility** and **end-to-end tracking**

MLOps Lifecycle

Machine Learning

- **Model:** Heart of ML application (e.g., Neural networks, …)
- **Data:** Data analysis and operations are crucial for MLOps since data is the base for ML



Differences compared to DevOps

- **Test:** Mostly related to tests regarding the convergence of models and specific model behavior
- **Deploy:** New data triggers retraining and redeployment of models
- **Monitor:** Assists in comprehending model performance and initiating retraining if required

MLOps specifically addresses the challenges of managing machine learning models and workflows to enable efficient and reliable implementation of AI technologies in production environments

Subramanya et al. (2022)

Today's lecture at a glance

- We have discussed model selection and train & retrain processes within ML decision spaces
- We have understood which KPIs are important in the context of AI initiatives and how suitable KPIs can be selected
- We have worked out the differences between data, training and deploment infrastructures

Questions, comments, observations

Scientific references

- Ashmore, R., Calinescu, R. & Paterson, C. 2021. Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. ACM Comput. Surv. 54, 5, Article 111 (June 2022), 39 pages. https://doi.org/10.1145/3453444
- BAIER, L., Jöhren, F., Seebacher, S. Challenges in the Deployment and Operation of Machine Learning in Practice. In: ECIS. 2019.
- Botchkarev, A. (2019): „A NEW TYPOLOGY DESIGN OF PERFORMANCE METRICS TO MEASURE ERRORS IN MACHINE LEARNING REGRESSION ALGORITHMS", Ryerson University.Toronto, Canada
- Breck, Eric; Cai, Shanqing; Nielsen, Eric; Salib, Michael; Sculley, D. (2017): The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction. In: Proceedings of IEEE Big Data.
- Escamilla-Ambrosio, P.J., Rodríguez-Mota, A., Aguirre-Anaya, E., Acosta-Bermejo, R., Salinas-Rosales, M. (2018). Distributing Computing in the Internet of Things: Cloud, Fog and Edge Computing Overview. In: Maldonado, Y.,
- Fortuna, C., Mušić, D., Cerar, G., Čampa, A., Kapsalis, P., Mohorčič, M. (2023). On-Premise Artificial Intelligence as a Service for Small and Medium Size Setups. In: Shinkuma, R., Xhafa, F., Nishio, T. (eds) Advances in Engineering and Information Science Toward Smart City and Beyond. Engineering Cyber-Physical Systems and Critical Infrastructures, vol 5. Springer, Cham. https://doi.org/10.1007/978-3-031-29301-6_3
- Haas, Christian (2019): The Price of Fairness - A Framework to Explore Trade-Offs in Algorithmic Fairness. In: Fortieth International Conference on Information Systems.
- Chai, S. Y. W., Phang, F. J. F., Yeo, L. S., Ngu, L. H., & How, B. S. (2022). Future era of techno-economic analysis: insights from review. Frontiers in Sustainability, 3, 924047.
- Even, A., Shankaranarayanan, G.: Utility-Driven Assessment of Quality. In: The DATA BASE for Advances in Information Systems 38 (2007) 2, S. 75-93.

Scientific references

AI FOR BUSINESS BUSINESS FOR AI

- Overhage, S., Birkmeier, D.Q. & Schlauderer, S. Qualitätsmerkmale, -metriken und -messverfahren für Geschäftsprozessmodelle. Wirtschaftsinf 54, 217–235 (2012). https://doi.org/10.1007/s11576-012-0335-1
- Pawar, C.S., Ganatra, A., Nayak, A., Ramoliya, D., Patel, R. (2021). Use of Machine Learning Services in Cloud. In: Pandian, A., Fernando, X., Islam, S.M.S. (eds) Computer Networks, Big Data and IoT. Lecture Notes on Data Engineering and Communications Technologies, vol 66. Springer, Singapore. https://doi.org/10.1007/978-981-16-0965-7_5
- Pipino, L., Lee, Y. W., Wang, R. Y.: Data Quality Assessment. In: Communications of the ACM 45 (2002) 4, S. 211-218.
- Rácz, Anita; Bajusz, Dávid; Héberger, Károly (2019): Multi-Level Comparison of Machine Learning Classifiers and Their Performance Metrics. In: Molecules 24 (15). DOI: 10.3390/molecules24152811.
- Silk, D., Mazzali, B., Gargalo, C. L., Pinelo, M., Udugama, I. A., & Mansouri, S. S. (2020). A decision-support framework for techno-economic-sustainability assessment of resource recovery alternatives. Journal of cleaner production, 266, 121854.
- Trujillo, L., Schütze, O., Riccardi, A., Vasile, M. (eds) NEO 2016. Studies in Computational Intelligence, vol 731. Springer, Cham. https://doi.org/10.1007/978-3-319-64063-1_4
- Ying, X. (2019, February). An overview of overfitting and its solutions. In Journal of physics: Conference series (Vol. 1168, p. 022022). IOP Publishing.

Non-scientific references

- https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861
- https://research.aimultiple.com/model-retraining/
- https://truera.com/ai-quality-management-key-to-driving-business-value/#:~:text=In%20short%2C%20AI%20Quality%20encompasses,robustness%2C%20reliability%20and%20data%20quality.
- https://www.qlik.com/us/kpi
- https://www.investopedia.com/terms/m/metrics.asp
- https://www.investopedia.com/terms/k/kpi.asp
- https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce
- https://www.techslang.com/definition/what-is-on-premises/
- www.klipfolio.com/resources/articles/what-is-a-key-performance-indicator
- https://www.youtube.com/watch?app=desktop&v=ctzDFIINSrI
- https://arxiv.org/abs/2005.11401
- https://crfm.stanford.edu/assets/report.pdf
- https://keras.io/examples/generative/ddim/

Pictures

- https://www.campaign-services.de/glossar/kpi/
- https://medium.com/@senapati.dipak97/grid-search-vs-random-search-d34c92946318