UNIVERSIDADE CATÓLICA PORTUGUESA

BRAGA

Machine Learning

Session 26 - T

Automated Machine Learning

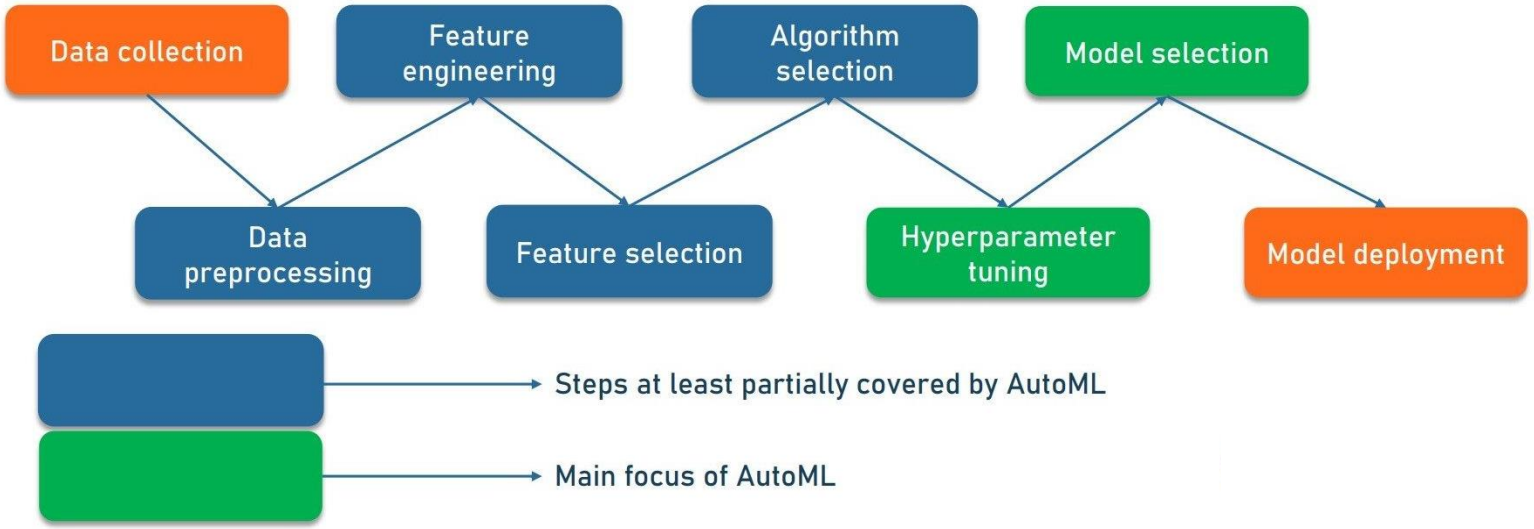
Degree in Applied Data Science

2024/2025

What is Automated Machine Learning (AutoML)?

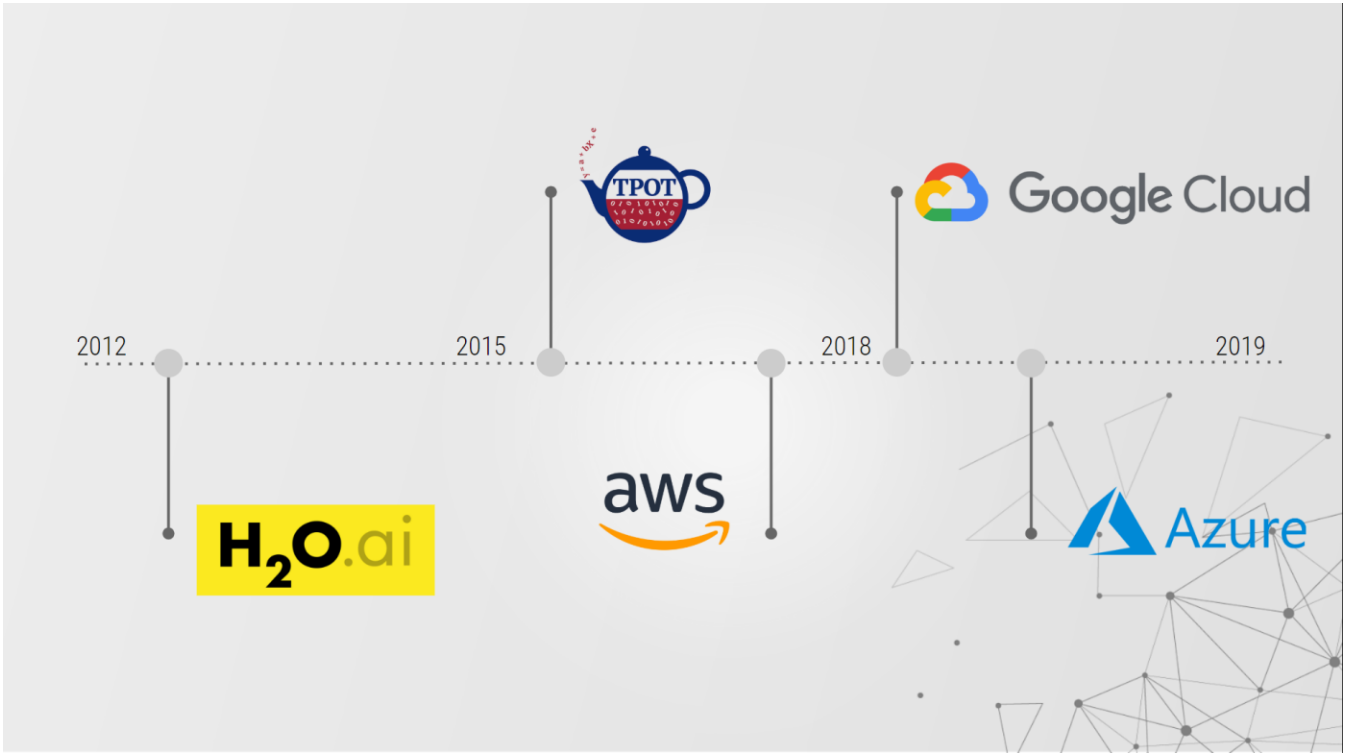
- AutoML is the process of automating the end-to-end process of applying machine learning to real-world problems;

- Simplify and speed up the development of machine learning models, making it accessible to non-experts.



AutoML – Tools and Frameworks

- Google AutoML
- H2O.ai
- Auto-sklearn
- TPOT
- Microsoft Azure AutoML
- …



Advanced Topics in AutoML

- **Neural Architecture Search** (NAS): Automating the design of neural network architectures.
- **Meta-Learning**: Learning how to learn; leveraging past experiences to improve future AutoML tasks.
- **Fairness and Ethics**: Addressing bias, transparency, and ethical considerations in automated systems.

Resources

- Automated Machine Learning. (2019). In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), The Springer Series on Challenges in Machine Learning. Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5

UNIVERSIDADE CATÓLICA PORTUGUESA

BRAGA

Machine Learning

Session 26 - T

Model Deployment and Monitoring

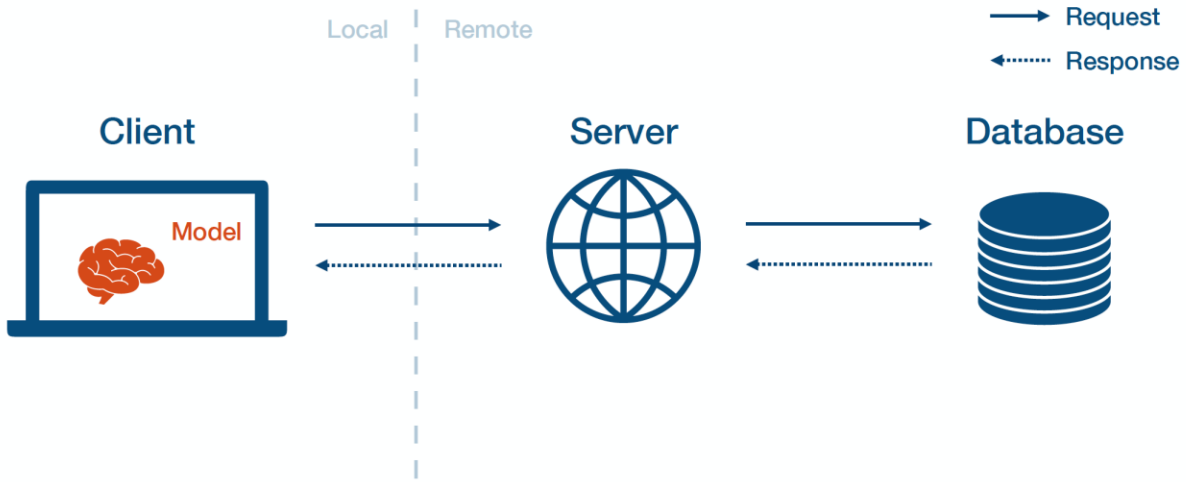
Degree in Applied Data Science

2024/2025

Model Deployment

- Model deployment is the process of making a machine learning model **available for use in production environments**;

- Models can be deployed in various environments, including on-premise **servers**, **cloud platforms**, and **edge devices**. Each scenario comes with its own challenges and considerations.



Model Optimization for Deployment

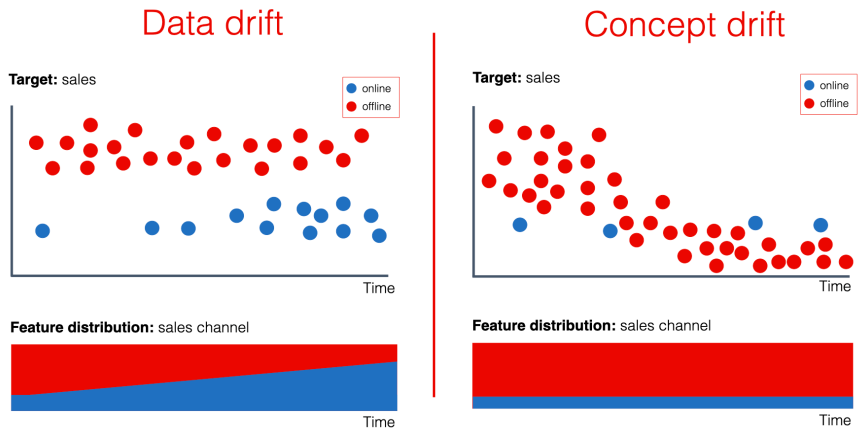
UNIVERSIDADE CATÓLICA PORTUGUESA
BRAGA

- **Model Compression Techniques**: To improve deployment efficiency, models can be compressed using techniques like **quantization**, **pruning**, and **knowledge distillation**, reducing their size and computational complexity.

- **Latency and Throughput**: Optimizing models for **low latency** and **high throughput** is essential for real-time applications. Techniques such as **hardware acceleration** and **architectural optimizations** can help achieve these goals.

Model Monitoring

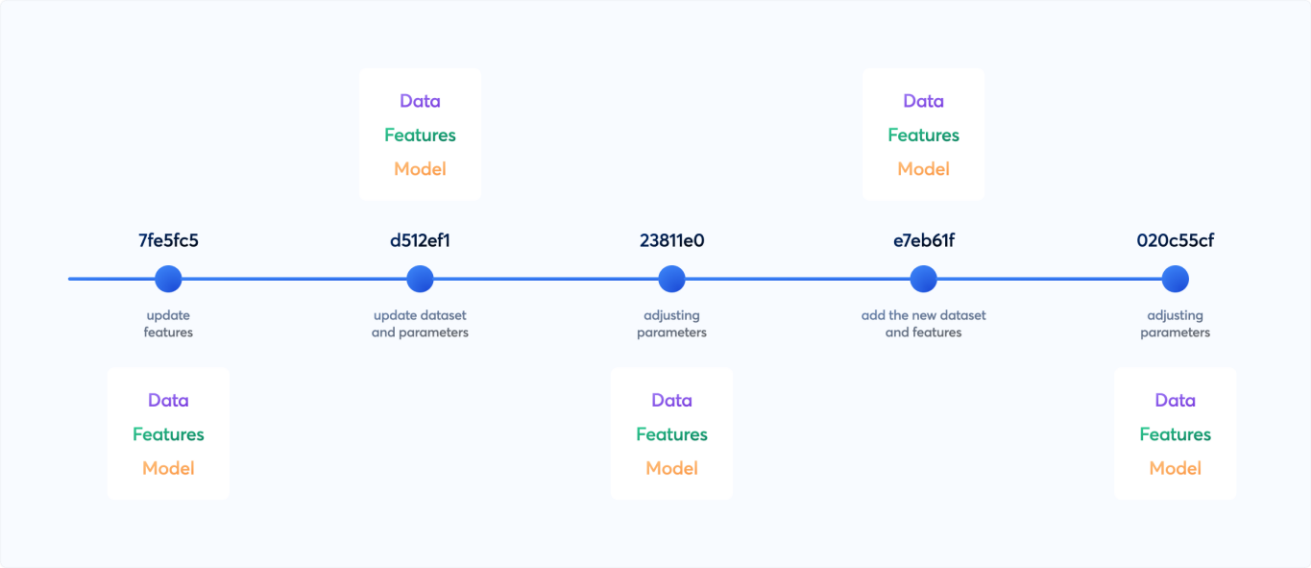
- Monitoring ensures that deployed models **perform as expected over time**;
- **Track metrics** like accuracy, latency, and throughput to detect performance issues;
- Monitor for **concept drift** (changes in the data distribution) and **data drift** (changes in data characteristics).



Model Versioning

- **Managing Versions**: Keep track of different versions of deployed models to facilitate rollback if necessary.

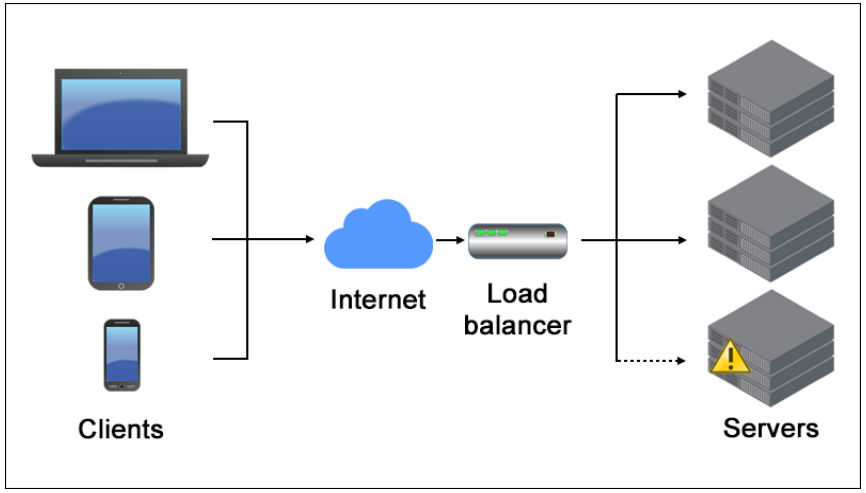
- **Rollback Strategies:** Plan for reverting to previous model versions in case of issues with new deployments.

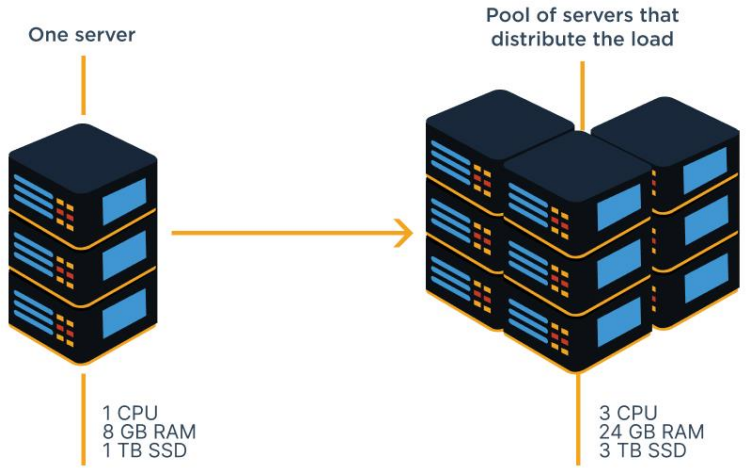


Scalability and Performance

- **Scaling Strategies**: Implement techniques like load balancing and horizontal scaling to handle increased demand;

- **Performance Optimization**: Optimize model inference speed and resource utilization for efficient deployment.





Resources

- Islam, J. (2022). Machine Learning Model Serving Patterns and Best Practices: A definitive guide to deploying, monitoring, and providing accessibility to ML models in production. Packt Publishing.