A Practical Introduction to Data Science

Part 8

Machine Learning Operations


Gergely Zsombor Haász


haasz.zsombi@gmail.com

Course Agenda

I.	Introduction to Data Science
II.	Business and Data Understanding
III.	Introduction to Supervised Learning
IV.	Advanced Supervised Learning
V.	Unsupervised Learning
VI.	Time Series Analysis
VII.	Deep Learning
VIII.	Machine Learning Operations

Review of the ML Lifecycle

Ways of working

- Business Collaboration
- Experimentation
- Iterative Process

Key Considerations

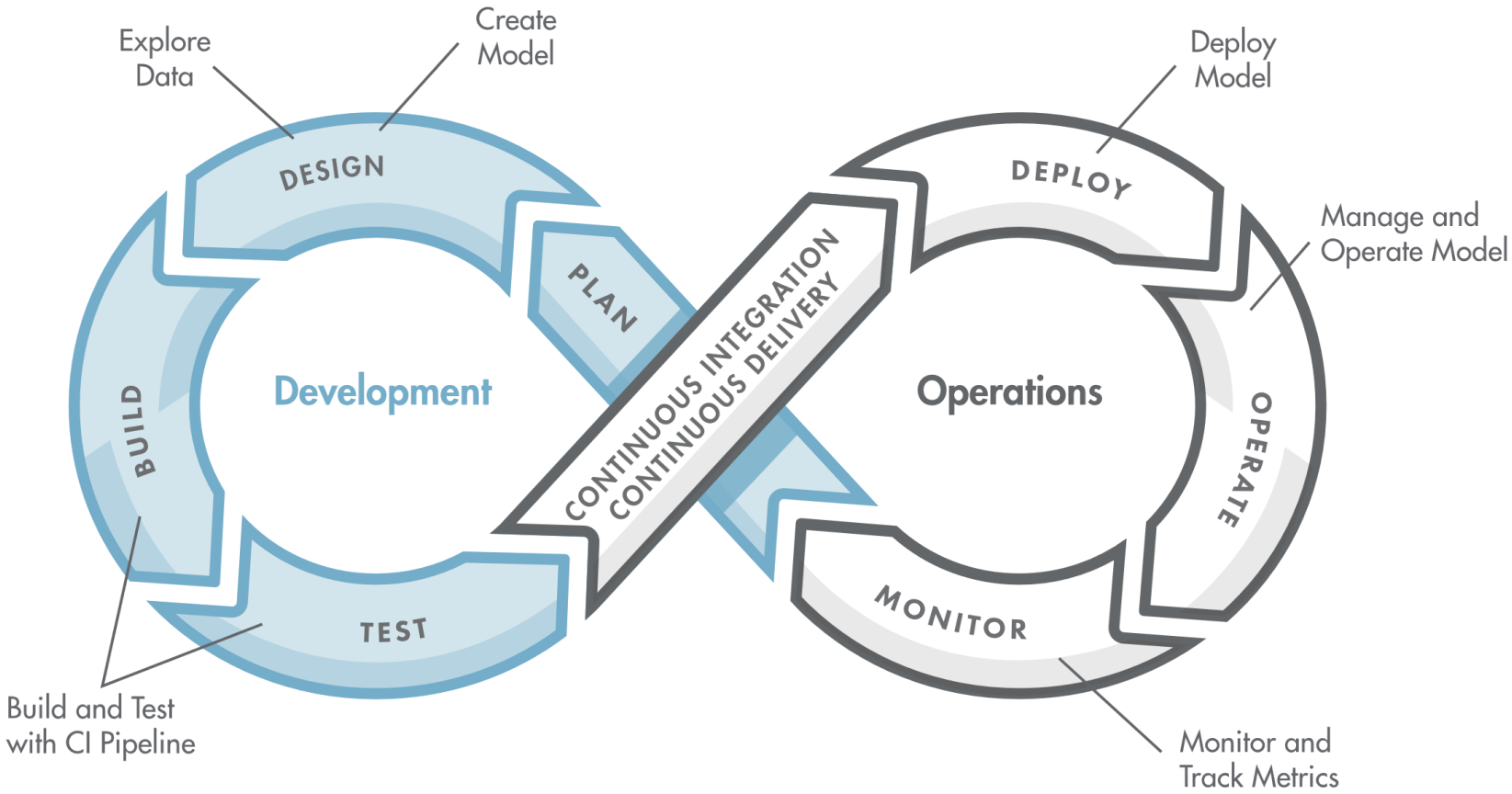
- Data Quality
- Generalization
- Data Leakage
- Business Impact

Source: Cross-industry standard process for data mining - Wikipedia

Machine Learning Operations

Phases:

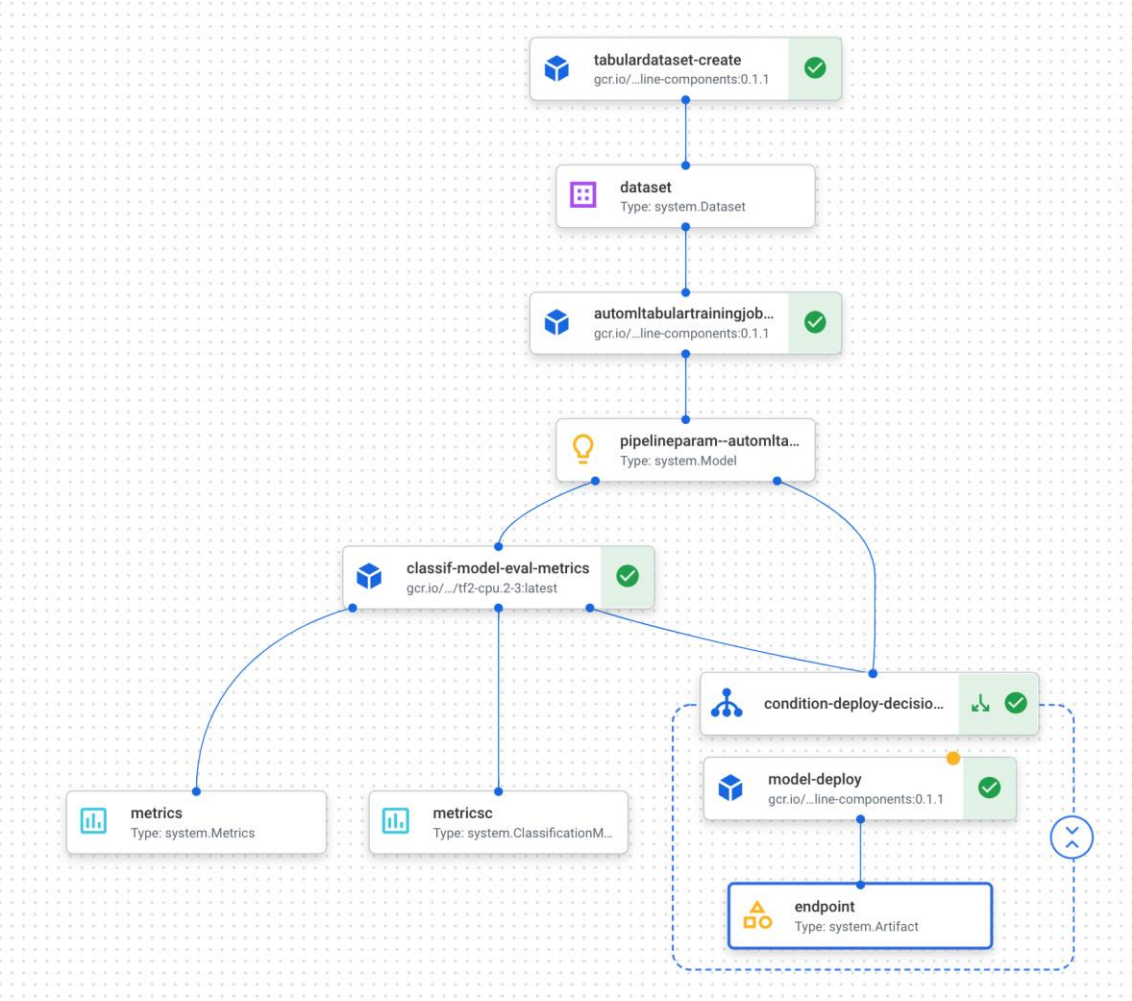
1. Pre-deployment
2. Deployment
3. Post-deployment



Source: MLOps - MATLAB & Simulink

Pre-Deployment

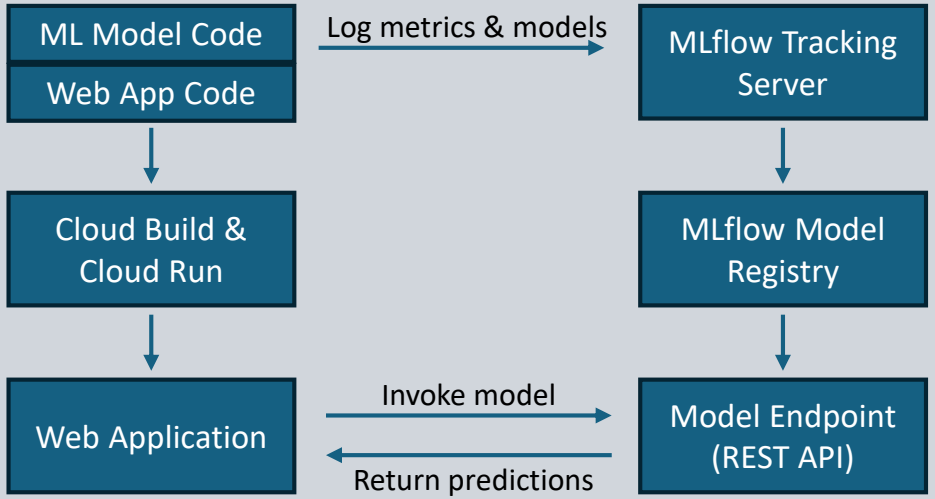
1. Data Engineering Pipeline
   - Reliable Data Ingestion
2. Model Development
   A. Experimentation
      - Notebooks
      - Experimentation Tracking (e.g. MLflow)
      - Version Control (Git)
   B. Machine Learning Pipeline
      - Directed Acyclic Graph (DAG)
      - Clean and Modular Code
      - Containerization
      - Automation
      - Reproducibility
      - Scalability

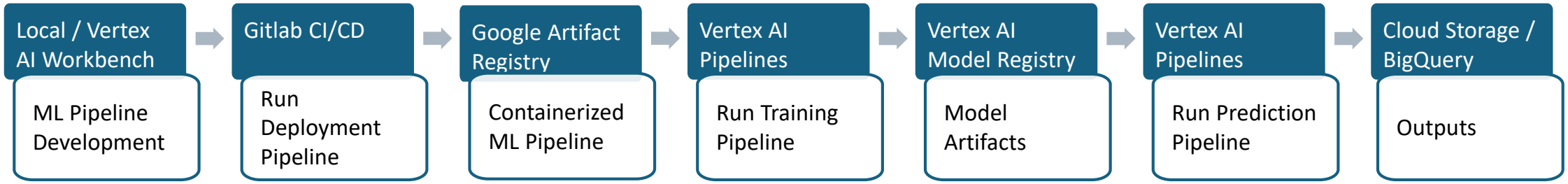


Source: Intro to Vertex Pipelines

Deployment

1. Rollout Strategy
2. Machine Learning Pipeline
3. Model Serving
   - Model Registry
   - Endpoints
4. Continuous Deployment
   - Environments: dev, test, prod
   - Stages: build, test, deploy





Post-deployment

1. Model-based decisions (actions)
2. Monitoring
   - ML Pipeline
   - Business
   - Model Performance Backtesting
   - Concept Drift and Data Drift
   - Business Impact
3. Retraining
   - Regular or performance driven

Post-deployment

Steps

1. Model-based decisions (actions)
2. Monitoring
   - **ML Pipeline**
   - Business
   - Model Performance Backtesting
   - Concept Drift and Data Drift
   - Business Impact
3. Retraining
   - Regular or performance driven

- Failures
- Runtime
- Input data
  - Availability and Format
  - Quality (missing values, duplicates)
  - Volume
- Output Volume and Quality

Post-deployment

Steps

1. Model-based decisions (actions)
2. Monitoring
   - ML Pipeline
   - **Business**
   - Model Performance Backtesting
   - Concept Drift and Data Drift
   - Business Impact
3. Retraining
   - Regular or performance driven

- Portfolio Size
- Business Targets
- Profit and Loss

Post-deployment

Steps

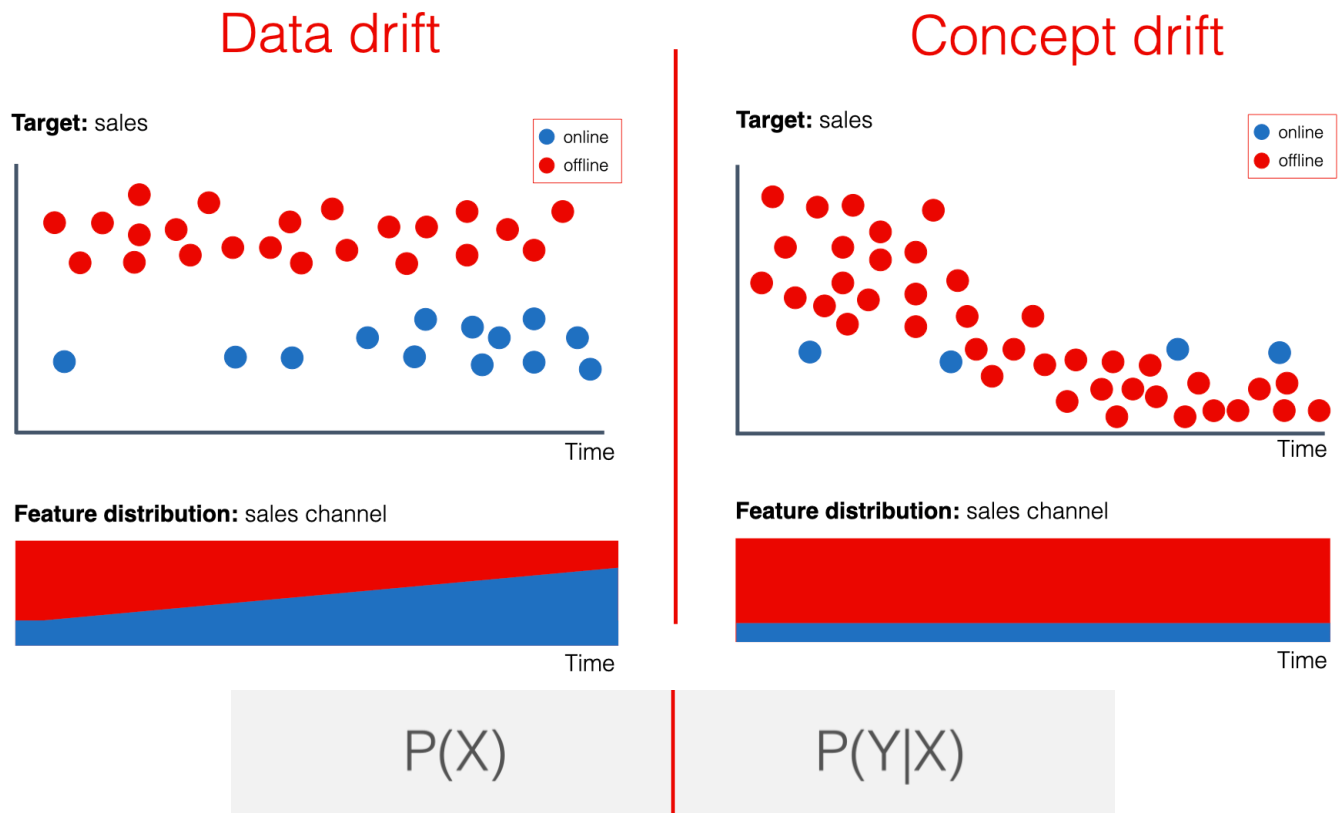
1. Model-based decisions (actions)
2. Monitoring
   - ML Pipeline
   - Business
   - **Model Performance Backtesting**
   - Concept Drift and Data Drift
   - Business Impact
3. Retraining
   - Regular or performance driven

- Stability and Drop in Performance
- Model Calibration
- Underperforming Segments
- Bias and Fairness

Post-deployment

Steps

1. Model-based decisions (actions)
2. Monitoring
   - ML Pipeline
   - Business
   - Model Performance Backtesting
   - **Concept Drift and Data Drift**
   - Business Impact
3. Retraining
   - Regular or performance driven

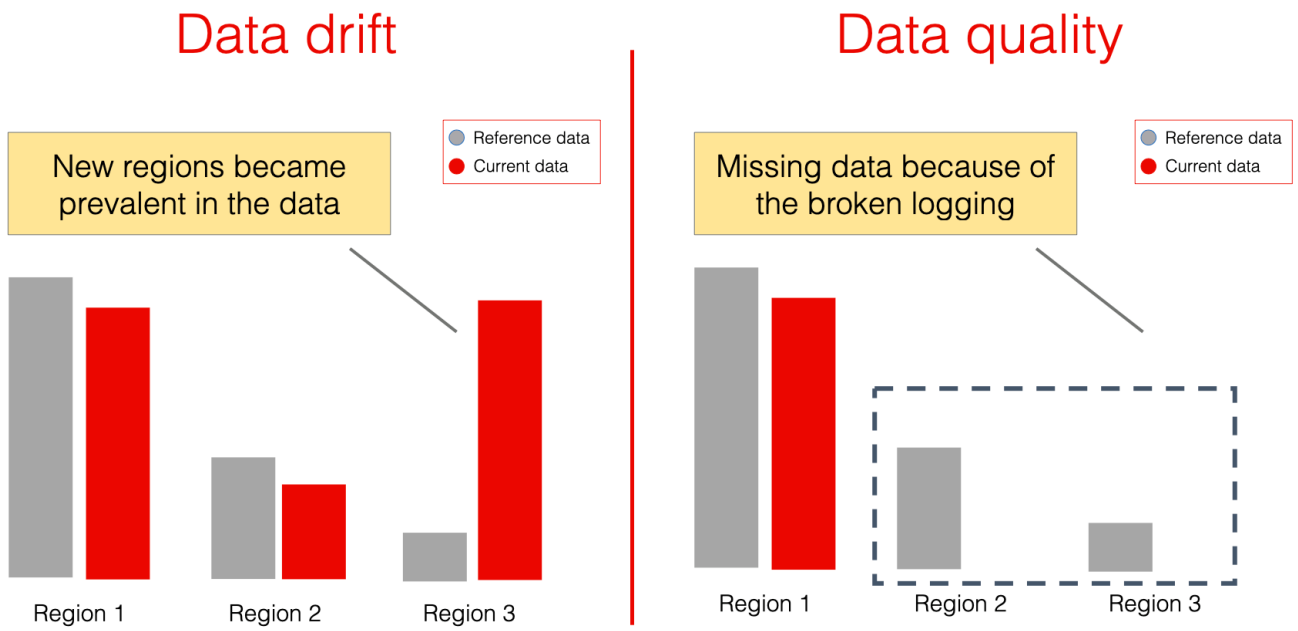


Source: What is data drift in ML, and how to detect and handle it

Post-deployment

Steps

1. Model-based decisions (actions)
2. Monitoring
   - ML Pipeline
   - Business
   - Model Performance Backtesting
   - **Concept Drift and Data Drift**
   - Business Impact
3. Retraining
   - Regular or performance driven



Source: What is data drift in ML, and how to detect and handle it

Post-deployment

Steps

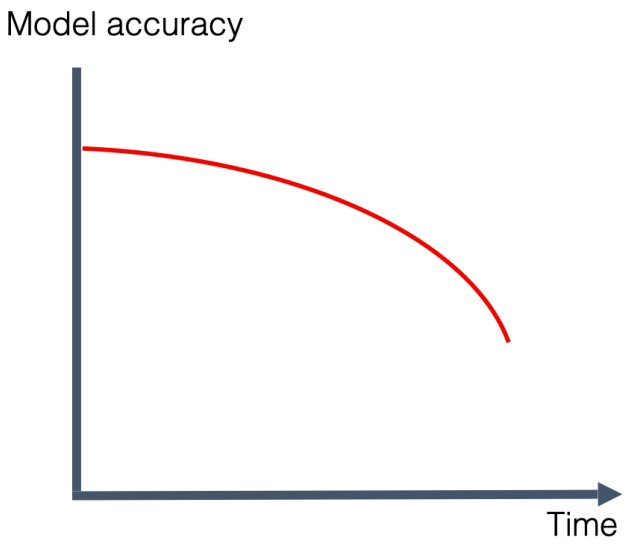
1. Model-based decisions (actions)
2. Monitoring
   - ML Pipeline
   - Business
   - Model Performance Backtesting
   - Concept Drift and Data Drift
   - **Business Impact**
3. Retraining
   - Regular or performance driven

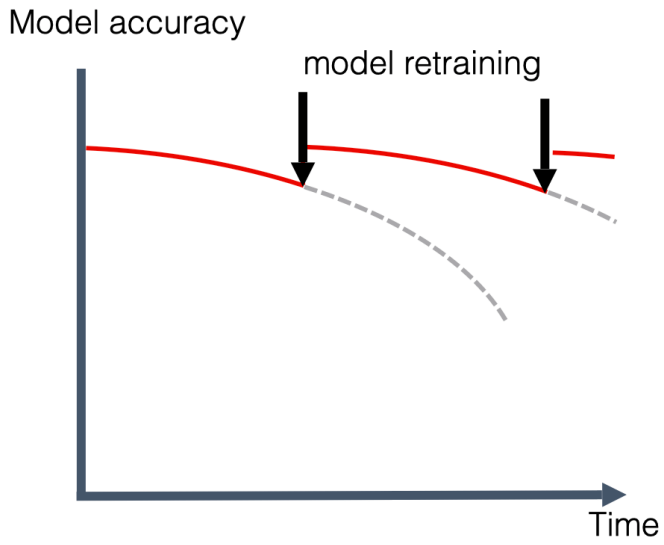
- A good model is not enough
  - Business decision, timing, target group
  - A/B testing
- Feedback Loops:
  - Predictions/actions may influence results and future samples
  - Backtesting and future modelling may be affected
    - Healed customers
    - Self-fulfilling prophecies
    - Altered distributions (e.g. excluded segments)
  - Action tracking
- External Factors
  - Environment, competitors etc.

Post-deployment

Steps

1. Model-based decisions (actions)
2. Monitoring
   - ML Pipeline
   - Business
   - Model Performance Backtesting
   - Concept Drift and Data Drift
   - Business Impact
3. **Retraining**
   - Regular or performance driven





Source: What is data drift in ML, and how to detect and handle it

MLOps Tools

Tracking	Model Registry	ML Pipeline	CI/CD	Other
MLflow	MLflow	Airflow	Gitlab CI/CD	Docker
W&B	Vertex AI	Kubeflow	Github Actions	Kubernetes
	SageMaker	Vertex AI	Jenkins	FastAPI

Read more about MLOps:

- ML Ops: Machine Learning Operations
- The Post Deployment Data Science Blog - by nannyML

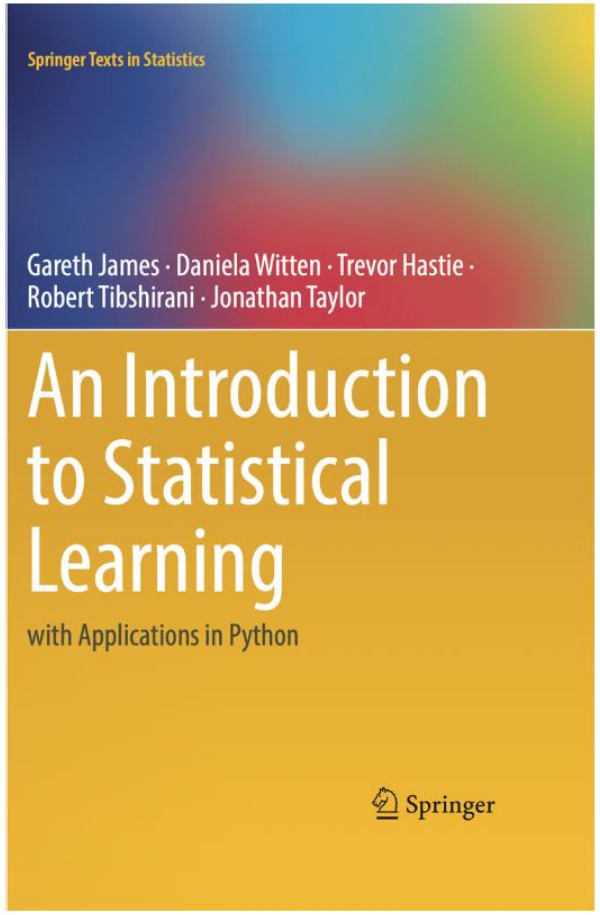
What we have covered

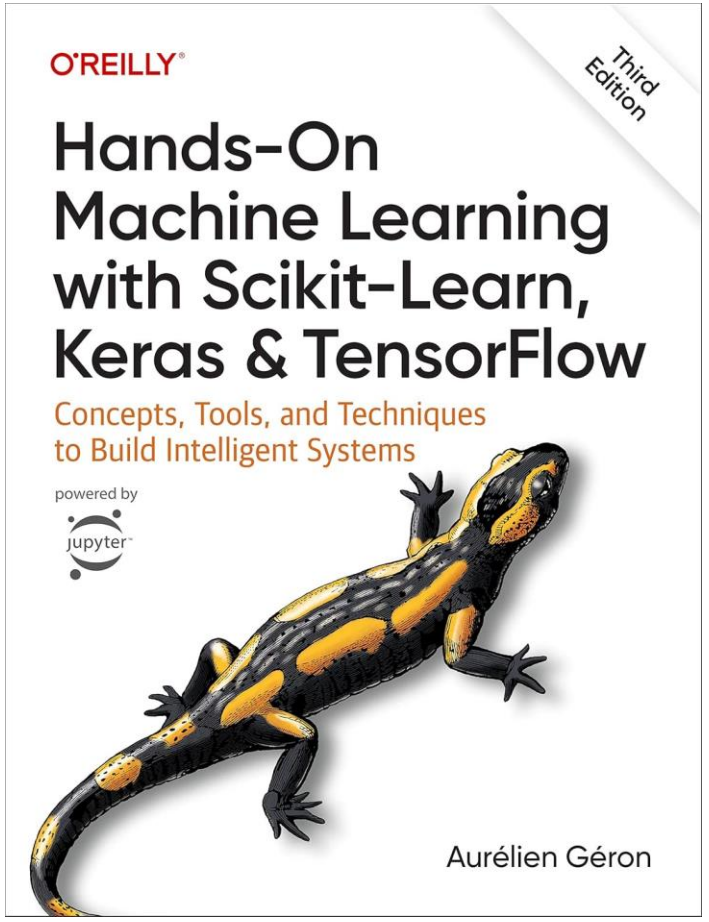
I.	Introduction to Data Science
II.	Business and Data Understanding
III.	Introduction to Supervised Learning
IV.	Advanced Supervised Learning
V.	Unsupervised Learning
VI.	Time Series Analysis
VII.	Deep Learning
VIII.	Machine Learning Operations

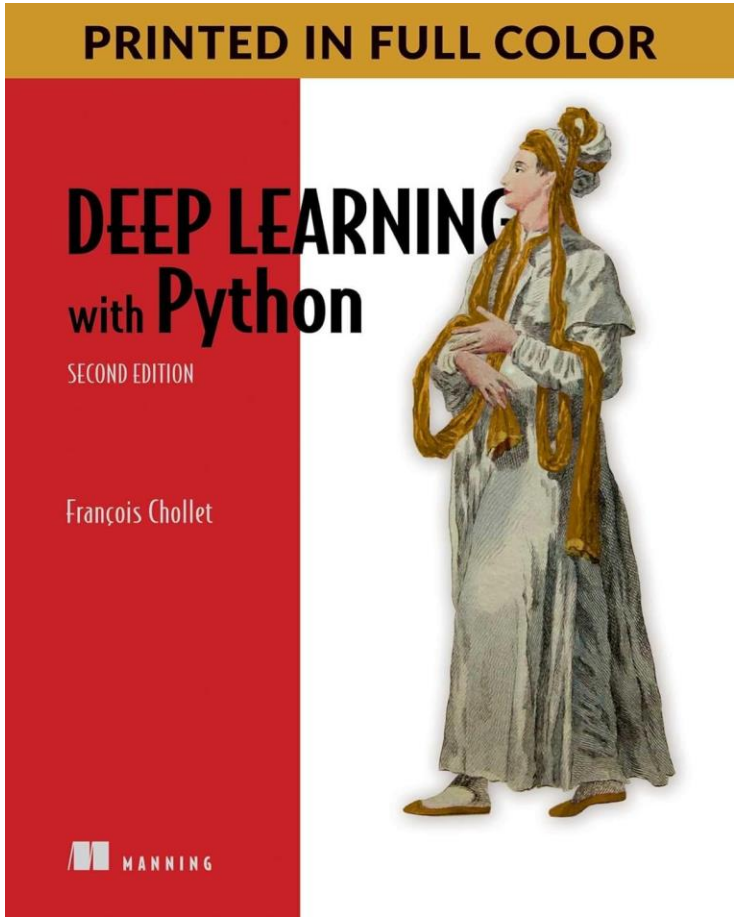
Free Online Resources

- Making Friends with Machine Learning: The Entire Course
- StatQuest with Josh Starmer
- User Guides and Tutorials – e.g. scikit-learn
- 3Blue1Brown – Neural networks
- Andrej Karpathy – Deep Dive into LLMs like ChatGPT
- Lecture series from top universities and other tutorials on YouTube
  - Stanford CS229 I Machine Learning I Building Large Language Models (LLMs)
  - MIT Deep Learning

Books


Springer Texts in Statistics
Gareth James · Daniela Witten · Trevor Hastie ·
Robert Tibshirani · Jonathan Taylor
An Introduction to Statistical Learning
with Applications in Python
Springer


O'REILLY
Third Edition
Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by
jupyter
Aurélien Géron


PRINTED IN FULL COLOR
DEEP LEARNING with Python
SECOND EDITION
François Chollet
MANNING

Thank you for your attention!

Your feedback would be much appreciated:

Any Questions?

Gergely Zsombor Haász

haasz.zsombi@gmail.com