FH | JOANNEUM
University of Applied Sciences

Work smarter, not harder: Workflow-Management in der Datenanalyse

Praxisnaher Einstieg in die Grundlagen des Workflow-Management

Raphaele Raab, Hannah Wimmer
Data Science and Artificial Intelligence
Institut für Wirtschaftsinformatik und Data Science, FH Joanneum
Eckertstraße 30i, 8020 Graz
{raphaele.raab2, hannah.wimmer}@fh-joanneum.at

Workflow-Management in der Datenanalyse: kurzer Recap

Wo waren wir gerade? Ein kurzer Recap…

- Was ist bisher aus diesem Kurs ‚hängengeblieben'?
- Datenpipelines, Workflows, Automatisierung, …

Workflow-Management in der Datenanalyse: kurzer Recap

Alles noch einmal in einem Satz:

„Ein Workflow ist eine Abfolge von Schritten, die voneinander abhängen."

Was bedeutet das für uns?

- Die Reihenfolge ist nicht nur relevant – sie ist kritisch!
- Einzelne Komponenten unserer Pipeline bauen aufeinander auf
- Eventuelle Fehler akkumulieren sich

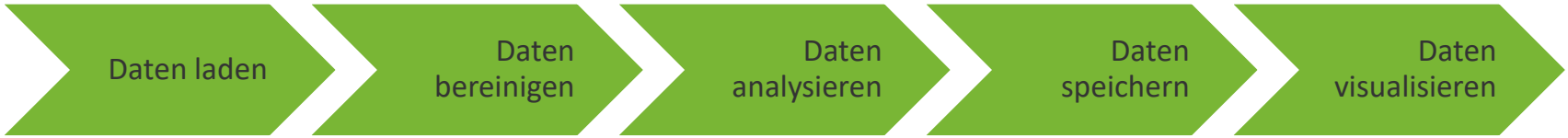
Workflow-Management in der Datenanalyse: kurzer Recap

Im Alltag verliert man oft den Überblick.

- Es wird (teils unter Zeitdruck) neuer Code in einer Codebasis hinzugefügt
- Es entstehen lange, unübersichtliche Skripten
- Einzelne Funktionalitäten gehen im Kontext unter
- Es ist am Ende unklar, wo was passiert, und was der konkrete Output sein soll

Was wir aber *eigentlich* gerne hätten…

- Klar erkennbare **Bauteile** in unserem Code
- Klar erkennbare **Linie,** die sich durch einzelne Prozesse zieht



Workflow-Management in der Datenanalyse: kurzer Recap

Das ist allerdings nicht alles. Fragen, die irgendwann automatisch aufkommen:

- Wer startet das Ganze?
- Was passiert, wenn irgendwo ein Fehler auftritt?
- Woher weiß ich, was schon fertig durchgelaufen ist?
- Muss ich das Skript jetzt jeden Montag erneut starten…?

→ Plötzlich reicht uns ein Skript **nicht mehr**

→ Wir stehen hier nun am Übergang zur **Workflow-Orchestrierung**

Themen des verbleibenden Kurses: praktische Beispiele

Wir schauen uns nun noch genau an:

- Wie man Workflows strukturiert und überwacht
- Wie man bereits *ohne* spezielle Tools Workflows erstellen kann
- Wie man *mit* Tools (hier Dagster) zusätzlich automatisieren und planen kann

Workflows ohne ‚Schnickschnack‘

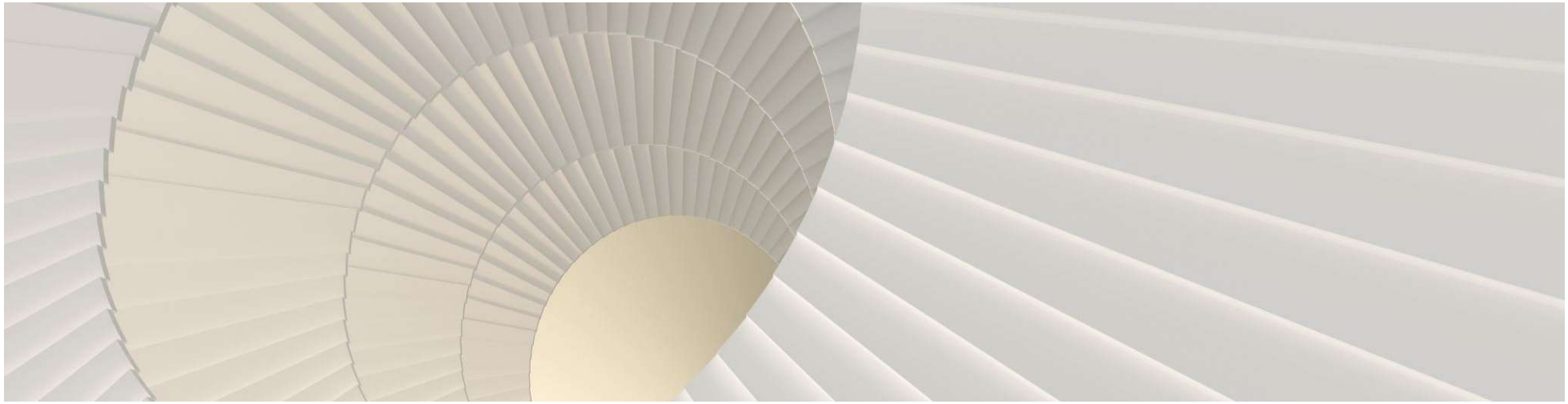
→ Mit einem *Workflow-Mindset* coden

Workflows ohne ‚Schnickschnack'

Ein Workflow braucht nicht zwangsläufig irgendwelche Tools.

- Ein sauber strukturiertes Skript / Jupyter Notebook kann bereits ein **Workflow** sein
- Erst bei **großen Datenmengen**, **regelmäßigen** oder **fehleranfälligen Abläufen** lohnt es sich, über Orchestrierung nachzudenken

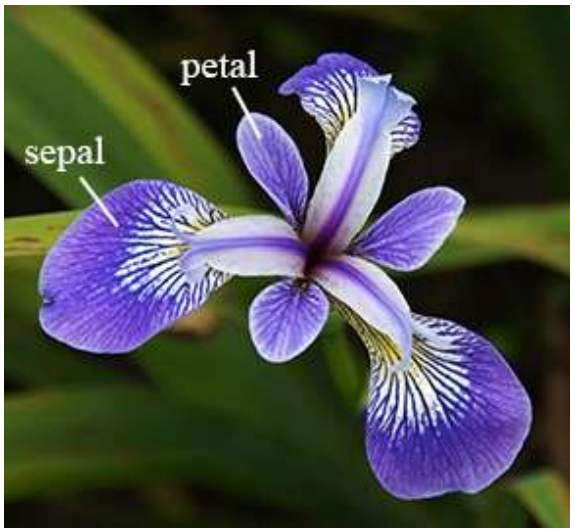
Ein funktionierender Workflow ist also zunächst nur eine **sinnvolle, reproduzierbare Abfolge von Schritten.**



Workflows ohne ‚Schnickschnack'
Beispiel 1

Coding Beispiel: ein einfacher ML-Workflow mit dem Iris Datensatz

- Iris („Schwertlilien") Datensatz:
  - Drei Arten von Schwertlilien: Iris setosa, Iris virginica, Iris versicolor
  - Jeweils 50 Proben
  - Je vier Merkmale: Sepal length & width, Petal length & width



Taken from [1]

→ Ein Standard-Datensatz für Machine-Learning-Beispiele

→ Wir wollen mit diesem Datensatz einen einfachen Workflow in einem **Jupyter Notebook** erstellen:

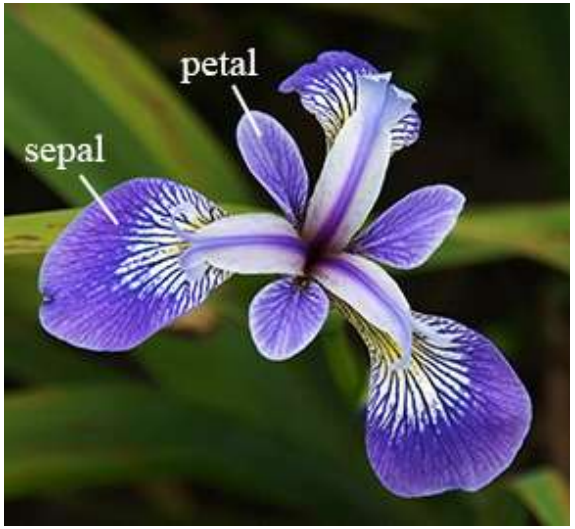
[1] Python: Analysis of Iris Dataset Using Pandas and Matplotlib

Workflows ohne ‚Schnickschnack'
Beispiel 1

Coding Beispiel: ein einfacher ML-Workflow mit dem Iris Datensatz

Worauf wir achten müssen, damit der Workflow **sauber** ist:

- Klare Reihenfolge der Schritte
- Train-/Test-Trennung
- Nachvollziehbare Parameter
- Reproduzierbare Ausführung
- Messbare Evaluation
- Klar dokumentiertes Ergebnis



Taken from [1]

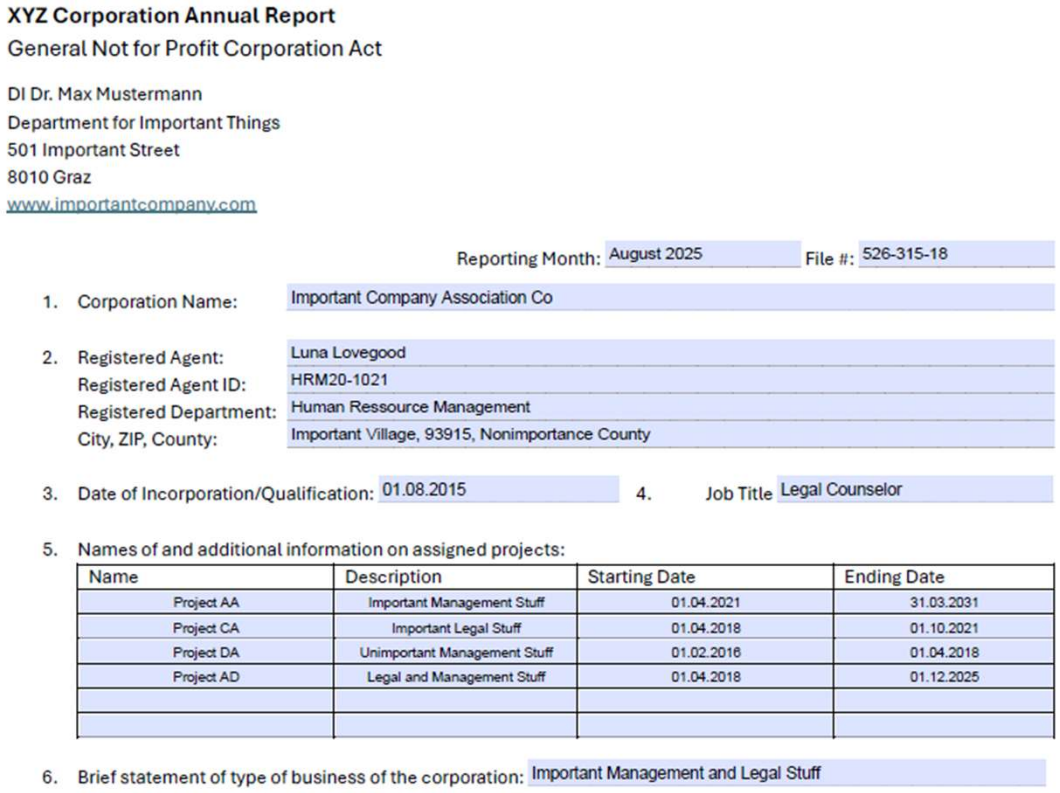
[1] Python: Analysis of Iris Dataset Using Pandas and Matplotlib

Workflows ohne ‚Schnickschnack'
Beispiel 2

Coding Beispiel: Datenextraktion aus einem PDF-Formular

- PDF-Formular mit vorgefertigten Feldern
- Mischung aus Metadaten und Tabellen
- Teils fehlende Einträge
- Teils Auswahlfelder (Yes/No)

→ Wir wollen mit diesem PDF-Formular nun einen einfachen Workflow in einem **Jupyter Notebook** erstellen:



XYZ Corporation Annual Report
General Not for Profit Corporation Act

DI Dr. Max Mustermann
Department for Important Things
501 Important Street
8010 Graz
www.importantcompany.com

Reporting Month: August 2025 File #: 526-315-18

1. Corporation Name: Important Company Association Co

2. Registered Agent: Luna Lovegood
Registered Agent ID: HRM20-1021
Registered Department: Human Ressource Management
City, ZIP, County: Important Village, 93915, Nonimportance County

3. Date of Incorporation/Qualification: 01.08.2015 4. Job Title Legal Counselor

5. Names of and additional information on assigned projects:

Name	Description	Starting Date	Ending Date
Project AA	Important Management Stuff	01.04.2021	31.03.2031
Project CA	Important Legal Stuff	01.04.2018	01.10.2021
Project DA	Unimportant Management Stuff	01.02.2016	01.04.2018
Project AD	Legal and Management Stuff	01.04.2018	01.12.2025

6. Brief statement of type of business of the corporation: Important Management and Legal Stuff

Workflows ohne ‚Schnickschnack'

Coding Beispiel: Datenextraktion aus einem PDF-Formular

Worauf wir achten müssen, damit der Workflow **sauber** ist:

- Klare Reihenfolge der Schritte
- Getrennte Funktionalitäten für Laden, Bereinigen, Aufbereiten der Daten
- Validierung der Ergebnisse
- Klar dokumentiertes Ergebnis

XYZ Corporation Annual Report
General Not for Profit Corporation Act

DI Dr. Max Mustermann
Department for Important Things
501 Important Street
8010 Graz
www.importantcompany.com

Reporting Month: August 2025 File #: 526-315-18

1. Corporation Name: Important Company Association Co
2. Registered Agent: Luna Lovegood
   Registered Agent ID: HRM20-1021
   Registered Department: Human Ressource Management
   City, ZIP, County: Important Village, 93915, Nonimportance County
3. Date of Incorporation/Qualification: 01.08.2015 4. Job Title Legal Counselor
5. Names of and additional information on assigned projects:

Name	Description	Starting Date	Ending Date
Project AA	Important Management Stuff	01.04.2021	31.03.2031
Project CA	Important Legal Stuff	01.04.2018	01.10.2021
Project DA	Unimportant Management Stuff	01.02.2016	01.04.2018
Project AD	Legal and Management Stuff	01.04.2018	01.12.2025

6. Brief statement of type of business of the corporation: Important Management and Legal Stuff

Workflows mit Dagster

→ Mit Tools automatisieren und planen

Workflows mit Workflow-Management-Tools

Bei größeren Workflows machen Workflow-Management-Tools durchaus Sinn.

- Tools wie Dagster geben einen guten Überblick über das Geschehen quer durch die ‚Pipeline'
- Fehler lassen sich genau zurückverfolgen
- Abläufe lassen sich planen und automatisiert ausführen

Ein Python-Skript kann ohne großen Aufwand in einen **Dagster-Workflow** umgebaut werden.

Workflows mit Workflow-Management-Tools
Beispiel 3

Coding Beispiel:
Szenensegmentation, Transkription und Zusammenfassung von Videos

- Videodaten sind normalerweise groß und divers – oft bleibt keine Zeit, sich alles anzusehen

Was wir also gerne machen würden:

- Video in zusammenhängende Szenen segmentieren
- Diese Szenen transkribieren
- Die Transkription zusammenfassen
- Das alles strukturiert darstellen

→ Das ist jetzt eine längere Pipeline

→ Wir sehen uns alle Teilbereiche zunächst in einem **Jupyter Notebook** an, dann wechseln wir zu **Dagster**

WORK SMARTER, NOT HARDER: WORKFLOW-MANAGEMENT IN DER DATENANALYSE

Diskussion und Wrap-Up