Einführung in Data Science

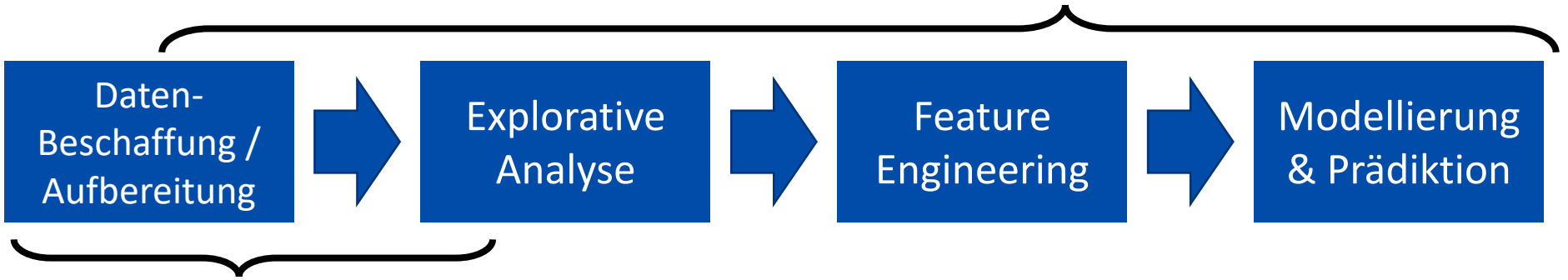
Unser Plan für heute:

1. Wiederholung
2. Fallbeispiel
3. Literatur und Lernunterlagen
4. Data Science Bibliotheken

Wdh | Data Science Überblick

Erkenntnisse
Vorhersagen
Daten-Produkte

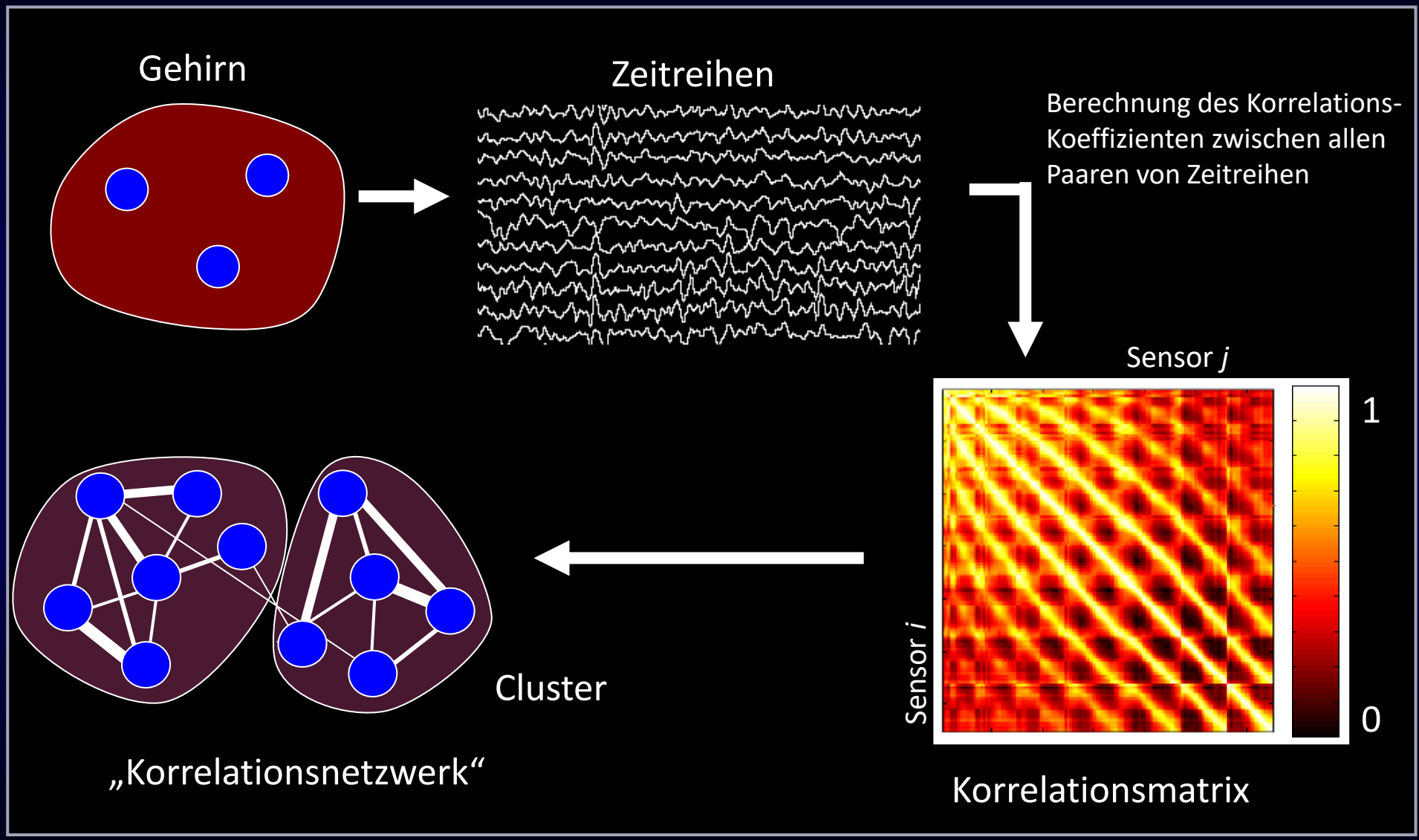
Data Scientist
Daten analysieren, visualisieren, interpretieren
Zusammenhänge erkennen



Data Engineer
Daten sammeln, speichern, bereinigen,
zusammenführen, bereitstellen, ETL

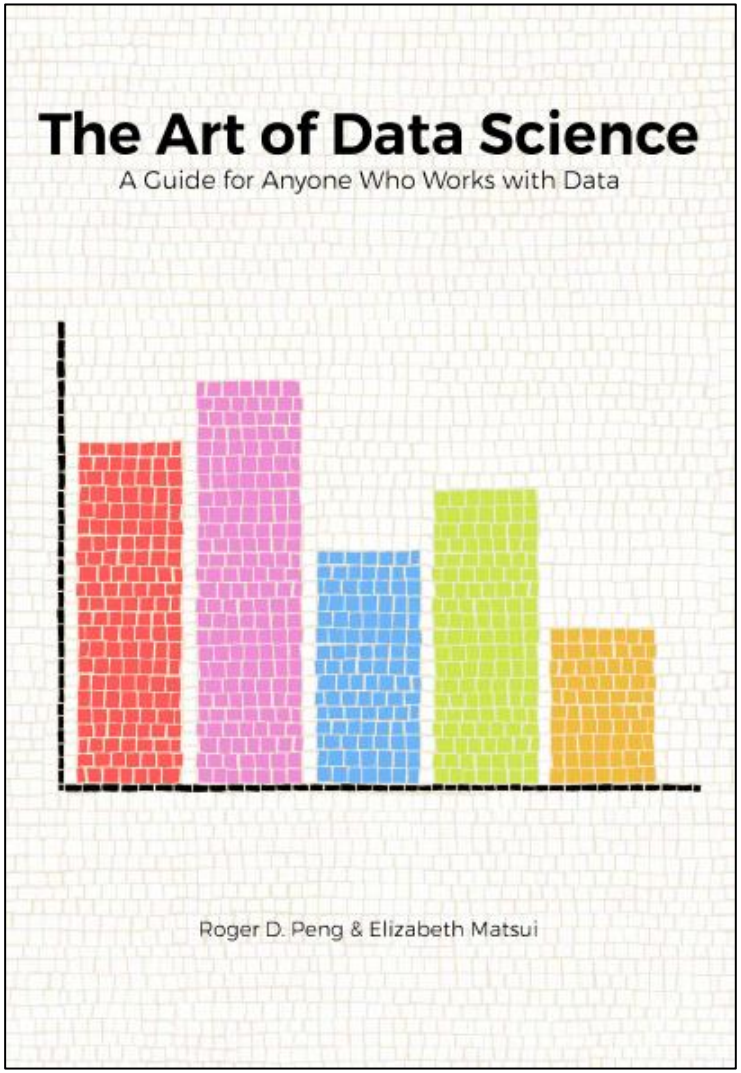
IT-Infrastruktur
Daten-Schnittstellen

Beispielpipeline für Vorhersage epileptischer Anfälle



3. Literatur und Lernunterlagen

Literatur und Lernunterlagen

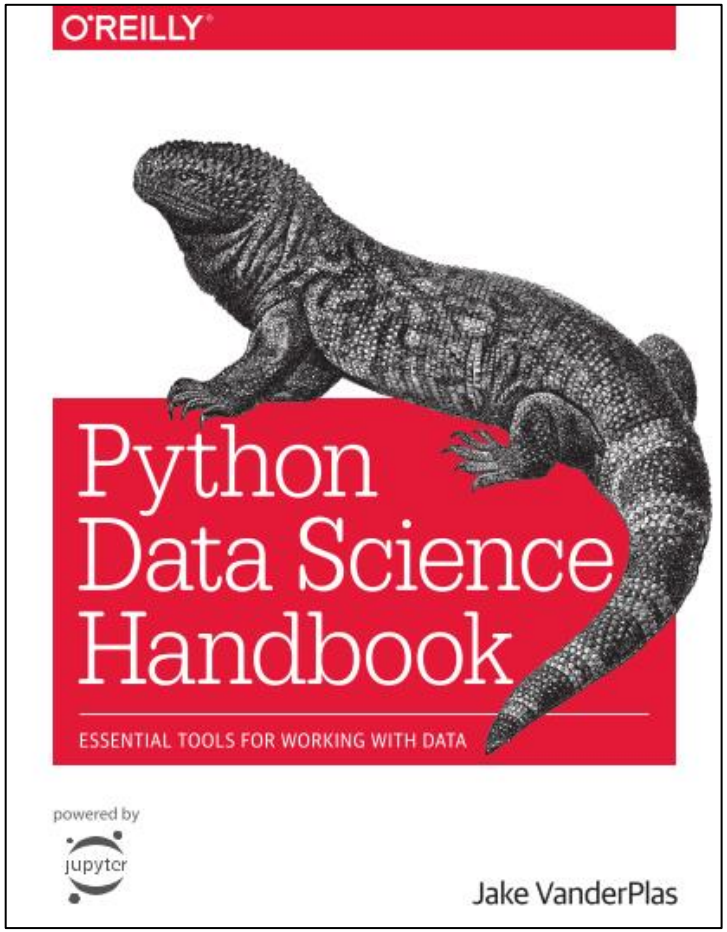


The Art of Data Science
von R. D. Peng, E. Matsui

- Darstellung des Arbeitszyklus im Bereich Data Science
- zur Orientierung über die Arbeitsweisen gut geeignet; aber: kein Arbeitsbuch
- Schwerpunkt Data Science statt Data Engineering
- Autoren sind Professoren an der Johns Hopkins Universität, Maryland (USA)

- Als HTML-Buch online verfügbar:
  https://bookdown.org/rdpeng/artofdatascience/

Literatur und Lernunterlagen



Python Data Science Handbook
von Jake VanderPlas

- für anwendungsorientierte Leser
- Übersicht über wichtigste Data Science Software-Bibliotheken für Python (548 Seiten)
- Autor ist Data Scientist, Universität Washington; Entwickler und Maintainer (scikit-learn, scipy)
- Online abrufbar: https://jakevdp.github.io/PythonDataScienceHandbook/ (vorläufige Fassung) finale Fassung im FH-Aachen Netz: https://bit.ly/2uXAxNF

Curriculum

1. Grundbegriffe / Überblick
2. Zentrale Softwarebibliotheken
3. Univariate explorative Analyse (EDA) Visualisierung (Teil 1)
4. Visualisierung (Teil 2), Multivariate explorative Analyse (Teil 1)
5. Multivariate explorative Analyse (Teil 2)
6. Dimensionsreduktion (Teil 1): PCA
7. Dimensionsreduktion (Teil 2): MDS, Isomap
8. Clustering: K-Means, HCA
9. Clustervalidierung
10. Probeklausur
11. Feature Engineering, Datengetriebene Modellierung (Teil 1)
12. Datengetriebene Modellierung (Teil 2)

Überblick / Begriffe

Explorative Analyse (EDA)

Feature Engineering & Modellierung

4. Python – Data Science Bibliotheken

Python – Data Science Bibliotheken

Laden, Verarbeiten:

- **NumPy** (numerical Python)
  Bibliothek zur effizienten Handhabung mehrdimensionaler Arrays.
- **Pandas** (= „**P**ython **an**d **D**ata **A**nalysi**s**)
  arbeiten mit in Tabellen organisierten Daten

Visualisieren:

- **Matplotlib**
  Klassiker; Erstellung (nicht-interaktiver) Abbildungen.
- **Seaborn**
  ergänzt Matplotlib mit zusätzlichen Abbildungstypen und Stilen
- **Plotly**
  erstellt interaktive Abbildungen; Nutzung u.a. in Dashboards

Untersuchen, Modellieren

- **SciPy** – Werkzeuge für EDA, Feature Engineering u.v.m; enthält: numpy
- **Scikit Learn** – datengetriebene Modellierung / Machine Learning