

1 Einführung und Modelle des Data Analysis Lifecycle

Modul: Angewandte Programmierung

Dennis Glösenkamp ▪ Köln ▪ 19. März 2020

© FOM Hochschule für Oekonomie & Management gemeinnützige Gesellschaft mbH (FOM), Leimkugelstraße 6, 45141 Essen

Dieses Werk ist urheberrechtlich geschützt und nur für den persönlichen Gebrauch im Rahmen der Veranstaltungen der FOM bestimmt.

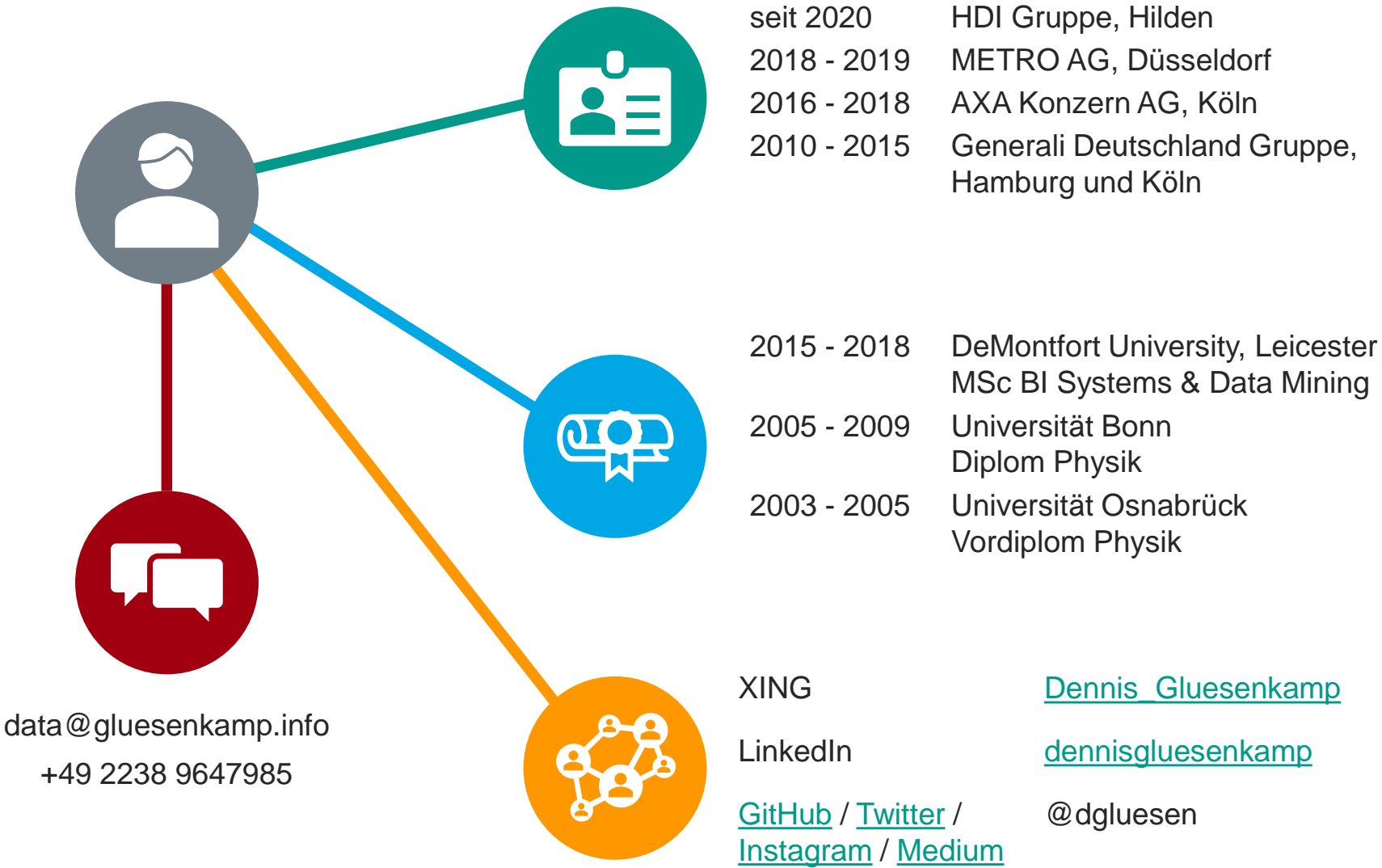
Die durch die Urheberschaft begründeten Rechte (u. a. Vervielfältigung, Verbreitung, Übersetzung, Nachdruck) bleiben dem Urheber vorbehalten.

Das Werk oder Teile daraus dürfen nicht ohne schriftliche Genehmigung des Urhebers / der FOM reproduziert oder unter Verwendung elektronischer Systeme verarbeitet, vervielfältigt oder verbreitet werden. Dies schließt auch den Upload in soziale Medien oder andere digitale Plattformen ein.

Inhalt

1	Einführung und Organisation
	Persönliche Vorstellung
	Modulziele, Curriculum und Strukturierung
	Programmierungsumgebung, Tools und Datenquellen
	Prüfungsleistungen
2	Data Analysis Lifecycle und Prozessmodelle
	Motivation
	CRISP-DM
	Sonstige Prozessmodelle
	Übungsaufgabe

1 Einführung und Organisation



Modulziele

Die Studierenden können nach erfolgreichem Abschluss des Moduls

- den für Big-Data-Analysen typischen **Anwendungszyklus** beschreiben und in der Praxis **begleiten**;
- im Anwendungszyklus häufig eingesetzte Systeme, **Programmiersprachen** und **Programmierungsumgebungen** benennen;
- relevante **Programmiermodelle** beschreiben;
- in einer typischen Systemumgebung mithilfe ausgewählter Programmierwerkzeuge strukturierte, semistrukturierte und unstrukturierte **Daten**
 - für die Analyse **aufbereiten**,
 - in Analysesysteme **integrieren**,
 - **automatisch und manuell analysieren** und
 - **visualisieren** sowie
 - **Ergebnisse** für weitere Verarbeitungen **bereitstellen**;
- die eingesetzten **Methoden und Werkzeuge** im Rahmen von umfangreichen Analyse- und Consultingprojekten effektiv und programmgesteuert **anwenden**.

- Anwendungszyklus (Data Analysis Lifecycle)
- Typische Systemkomponenten
 - Datenbankmanagementsysteme
 - Apache Hadoop
- Programmiermodelle im Bereich Big Data
 - MapReduce
 - Funktional
 - SQL-basiert
 - statistisch und analytisch
 - Datenfluss-basiert
 - Bulk Synchronous Parallel
 - High-level Domain Specific Language (DSL)
- Gängige Programmiersprachen, Programmierumgebungen und Frameworks
 - Python
 - Java, Scala
 - R, PL/R
 - SQL
- Anwendung ausgewählter Programmiermodelle
 - Anbindung von Datenquellen
 - Extrahierung und Bereitstellung von Rohdaten
 - Aufbereitung und Integration von Daten
 - Datenanalyse und Bereitstellung von Ergebnisdaten
 - Datenvisualisierung und -bereitstellung
 - Operationalisierung
- Kommunikation von Ergebnissen

Was ist Kaggle?

- Online-Community für Data Scientists
- Plattform für
 - Data Science Competitions
 - freien Datensätzen
 - Beispielcodes
 - Diskussionen
- Mitglieder können Kenntnisse anwenden und vertiefen
- In diesem Moduls werden vor allem die dort angebotenen Datensätzen genutzt

kaggle

Was ist HackerRank?

- Training von Programmierkenntnisse in vielen verschiedenen Sprachen
- Übungen mit ansteigendem Schwierigkeitsgrad
- Browser-basierte Bearbeitung und Lösung der Aufgaben
- Auch Plattform für Jobangebote bzw. Testumgebung in Bewerbungsverfahren



Was ist git?

- Versionsverwaltung von Dateien
- Ähnliche Werkzeuge: BitKeeper, Bazaar
- konsistente Fortentwicklung von Programmcode



Was ist GitHub?

- Software mit Versionskontrolle und online verwalten
- Ähnliche Dienste: Bitbucket, GitLab
- Setzt auf Versionsverwaltungssoftware git auf
- Möglichkeit Projekte über eigenen Websites zu präsentieren
- Einfache, agile Tools inkludiert



Was ist R?

- Freie, auf statistische Anwendungen fokussierte Programmiersprache
- Eingaben über eine Kommandozeilenkonsole oder per Skript
- Integrierte Entwicklungsumgebung (IDE) ist z.B. RStudio



Was ist Anaconda?

- Python arbeitet sehr stark mit verschiedenen Paketen
- Pakete müssen installiert und eingebunden werden
- Paketkombinationen und verschiedene -versionen können Konflikte hervorrufen
- Anaconda als Distribution für Python adressiert dieses Problem



- Für das Bestehen und die Benotung werden folgende **Prüfungsleistungen** erbracht:
 - Postersession, 25% der Gesamtnote
 - Klausur, 75% der Gesamtnote
 - Leistungen müssen jeweils mindestens ausreichend sein um das Modul zu bestehen
- **Postersession:**
 - Präsentation/Pitch im Rahmen einer Lehrveranstaltung (vsl. 04.06.2020)
 - Zeit maximal fünf Minuten pro Person
 - Möglichkeit zu Fragen und zum Austausch mit anderen Studierenden
 - Prämierung „Best Poster Award“ aus Studierendensicht
 - **Generelle Aufgabenstellung:** *Wählen Sie einen Datensatz aus, formulieren Sie eine zentrale Fragestellung, leiten Sie Erkenntnisse hierzu aus den Daten mit Hilfe der in der Vorlesung vorgestellten Methoden ab und präsentieren Sie Ihre Ergebnisse mit einem wissenschaftlichen Poster. Mögliche Arbeitsschwerpunkte: Datenbereinigung und -aufarbeitung, Visualisierung, Data-Story-Telling und/oder Modellbildung*
 - Thema/Datensatz wird selbst gewählt (vsl. 26.03.2020)
 - Vortrag und Poster nach Wahl in deutsch oder englisch
- **Klausur:**
 - Dauer von 90 Minuten, Termin: 20.06.2020
 - Thematischer Umfang der in der Vorlesung gezeigten und geübten Inhalte

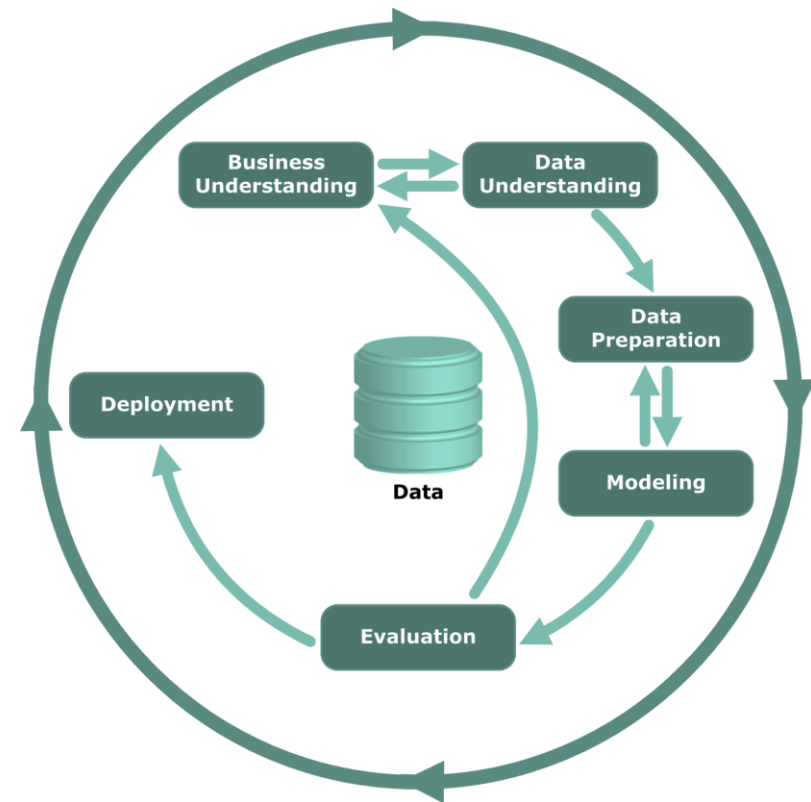
2 Data Analysis Lifecycle und Prozessmodelle

Verbindlichkeit durch Prozessmodell

- Prozessbeschreibung von Schritten, die bei der Durchführung von datengetriebenen Aktivitäten erforderlich sind, erzeugt
 - **Vollständigkeit**, da Schritte systematisch abgearbeitet werden
 - **Iterationsfähigkeit**, da in bestimmten Zyklen Fortentwicklung und Ergebnisbereitstellung erfolgen
 - **Anschaulichkeit**, da eine logische und sinnvolle Struktur abgearbeitet wird
 - **Transparenz**, da Rollen und Zuständigkeiten geklärt sind
 - **Sicherheit**, da Schutzmechanismen integriert werden können
- Vermeidung von Fehlern oder nicht notwendiger Ineffizienz durch mangelnde Organisation
- Steigender Komplexitätsgrad bei Daten-Projekten erfordert höheres Maß an Struktur um Sackgassen oder Chaos zu verhindern

CRISP-DM Life Cycle

- **C**Ross-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining [1]
- 1996 im Rahmen von EU-Förderprojekt entwickelt
- **Offenes, freies Prozessmodell** zur Durchführung von Data Mining Vorhaben
- Methodik ist **flexibel und adaptierbar**
- Prozess kann **unabhängig von Branche, Toolset und Anwendung** verwendet werden
- Erweiterung [ASUM-DM](#) 2018 von IBM veröffentlicht



CRISP-DM Prozessmodelldiagramm

Image by Kenneth Jensen, distributed under a [CC BY-SA 3.0](#).

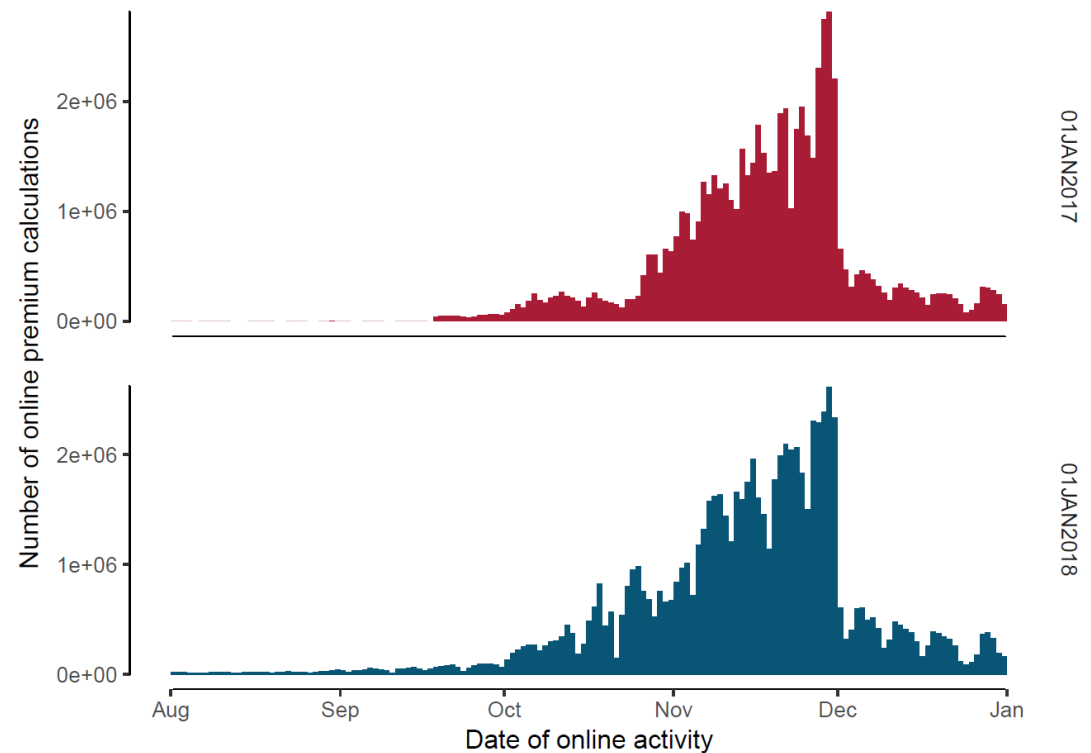
Phasen von CRISP-DM (1/3)

▪ Business Understanding

- Formulierung von konkreten Fragestellungen und Zielen
- Abgleich von Aufgaben und Erwartungen
- Vereinbarung eines Vorgehens/einer Planung
- Identifikation von wichtigen Einflussfaktoren
- Verständnis des Geschäftsmodells
- Definition von Erfolgskriterien

▪ Data Understanding

- Betrachtung des Datenbestands
- Auswertung der Datenverfügbarkeit, -reliabilität, -qualität
- Abstimmung zum Datenschutz



Anzahl von Preisanfragen für Kfz-Versicherungen über Aggregator- bzw. Vergleichswebsites bei einem deutschen Versicherer [2]

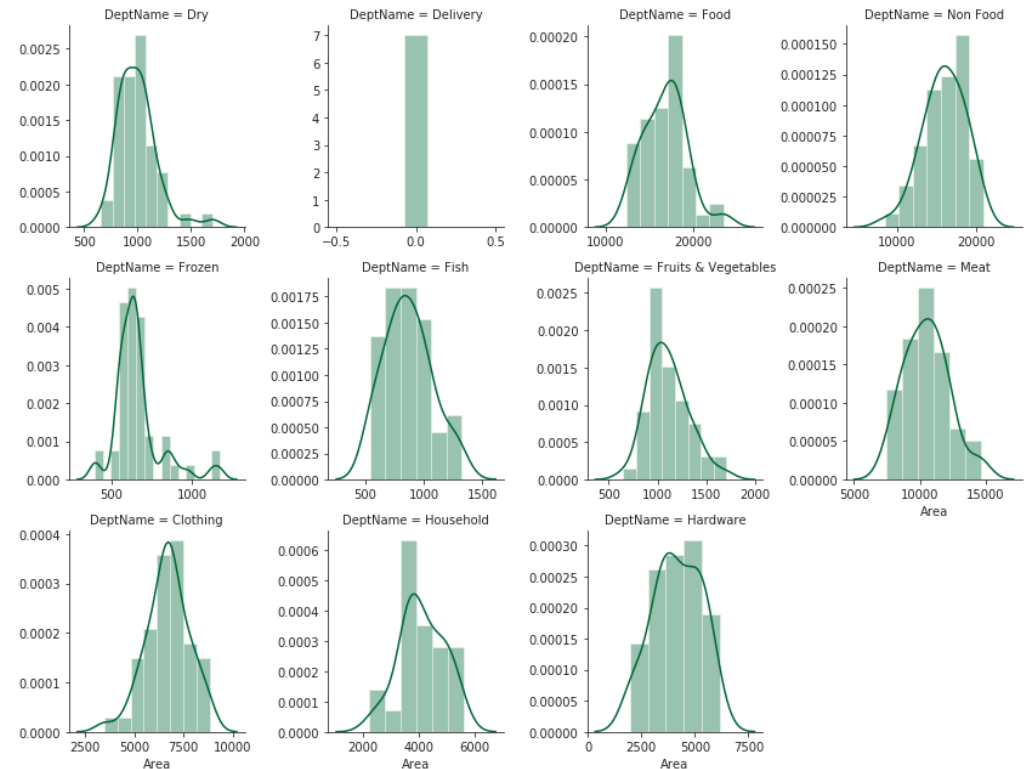
Phasen von CRISP-DM (2/3)

■ Data Preparation

- Datenbereinigung und Transformationen
- Datenverknüpfung und -aggregation
- Feature Engineering
- Feature Selection

■ Modeling

- Definition der Annahmen und Rahmenbedingungen der Modellierung
- Auswahl von geeigneten Algorithmen
- Test Design
- Training des Modells



Verteilung von Verkaufsflächen von verschiedenen Märkten eines fiktiven Handelskonzerns, getrennt nach Fachabteilungen [3]

Phasen von CRISP-DM (3/3)

▪ Evaluation

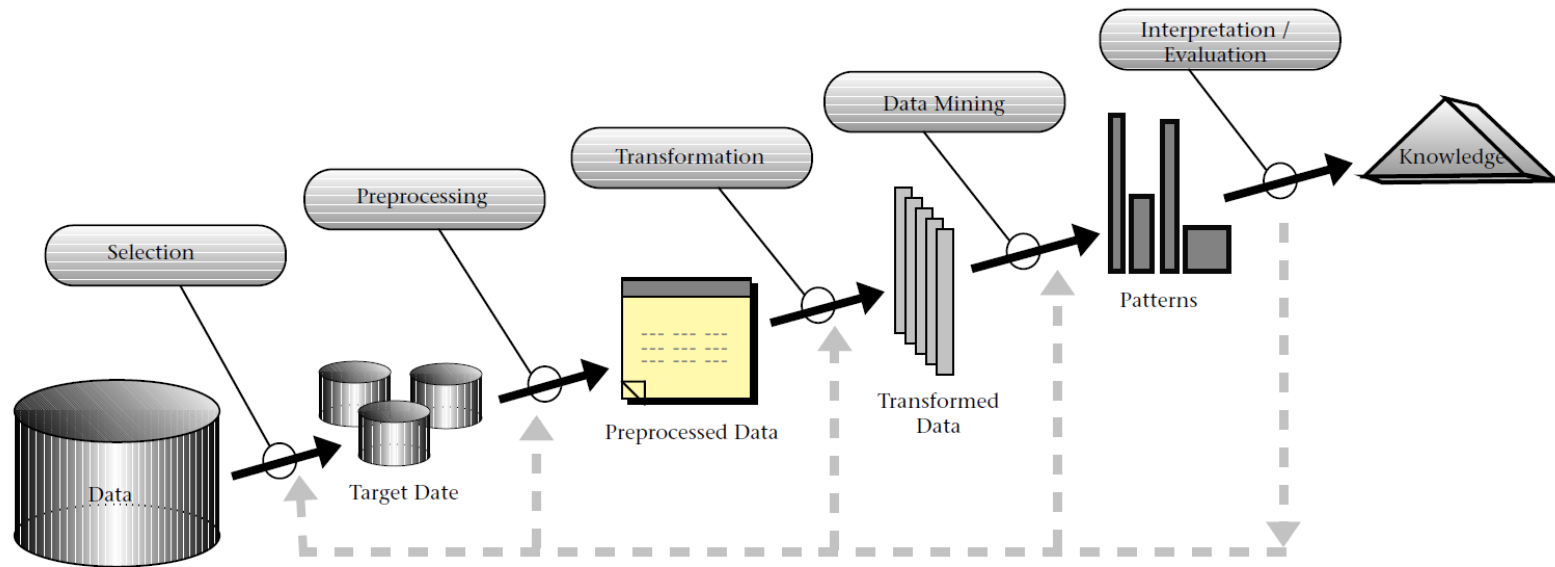
- Vergleich der verschiedenen Modelle anhand von Gütekriterien
- Betrachtung der Interpretierbarkeit des Modells
- Kritische Analyse des Modellierungsprozesses
- Abgleich mit (wirtschaftlichen) Erfolgskriterien
- Definition von Folgeaktivitäten

▪ Deployment

- Kommunikation der Ergebnisse
- Integration des Modells in die Systemlandschaft und Entscheidungsprozesse
- Wartung und Pflege des Modells
- Dokumentation der Erkenntnisse und Funktionsweise

KDD und SEMMA

- KDD ist Prozessmodell für **K**nowledge **D**iscovery in **D**atabases [4]
- 1996 von Fayyad, Piatetsky-Shapiro, Smyth publiziert



KDD Prozessschritte [4]

- **S**ample, **E**xplore, **M**odify, **M**odel, and **A**ssess ist ein von SAS vorgeschlagenes Prozessmodell [5]
- Prozess wird trotz Tool-Unabhängigkeit vornehmlich in enger Verknüpfung zu SAS-Lösungen genutzt

Übungsaufgabe

In Vorbereitung auf die nächste Vorlesung am 26. März 2020 bearbeiten Sie bitte folgende Aufgabenstellung:

- Bitte wählen Sie ein mögliches Data Science/Mining Vorhaben (Use Case) aus, dass Ihnen aus Ihrer beruflichen Praxis, den Medien oder aufgrund von persönlichen Interessen bekannt ist.
- Diskutieren Sie anhand dieses Beispiels mögliche Fragen, Arbeitsschritte und Ergebnisse in den sechs Schritten des CRISP-DM Lifecycles.
- Wählen Sie zusätzlich zum Vergleich ein anderes Lifecycle-Modell aus und diskutieren Sie die Gemeinsamkeiten und Unterschiede zu CRISP-DM im Rahmen Ihres Beispiels. (Hilfreich könnte hier der Artikel von Azevedo und Santos sein [6])

Anhang

- [1] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc*, 9, 13..
- [2] Gluesenkamp, D. (2018). Prediction of customer churn with premium online calculation data in insurance business. DeMontfort University, Leicester, United Kingdom.
- [3] Gluesenkamp, D. (2019). Wrangling and cleansing business data. Retrieved from <https://dgluesen.github.io/wrangling-sales-workload/>
- [4] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37-37.
- [5] SAS Institute. SAS® Enterprise Miner. Retrieved from https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-enterprise-miner-101369.pdf Publisher website: <https://www.sas.com/>
- [6] Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.