



Business Understanding für Data Science

Stefanie Scholz

Data Science Day – 26. April 2022

Kontakt



Prof. Dr. Stefanie Scholz


Professorin für Sozialwirtschaft

stefanie.scholz@srh.de

- 🎯 KI-basierte Analysen unstrukturierter Text-Daten
- 🎯 Data driven Marketing
- 🎯 Advanced Analytics



Agenda

- 1) CRISP-DM als Grundlage für Business & Data Understanding
 - 2) Use Case
 - 3) Soziale Netzwerke
 - 4) Projektablauf einer UGC-Analyse
 - 5) Unser Beispiel: Subreddit „r/technology“
- 
- A long-exposure photograph of a highway at night, showing bright, curved light trails from vehicles. The trails are primarily white and yellow, curving from the bottom left towards the top right. There are also some orange and red trails, possibly from brake lights or slower-moving vehicles. The background is dark, with some distant lights visible on the horizon.

CRISP-DM

01

CRISP-DM → Cross Industry Standard Process for Data Mining



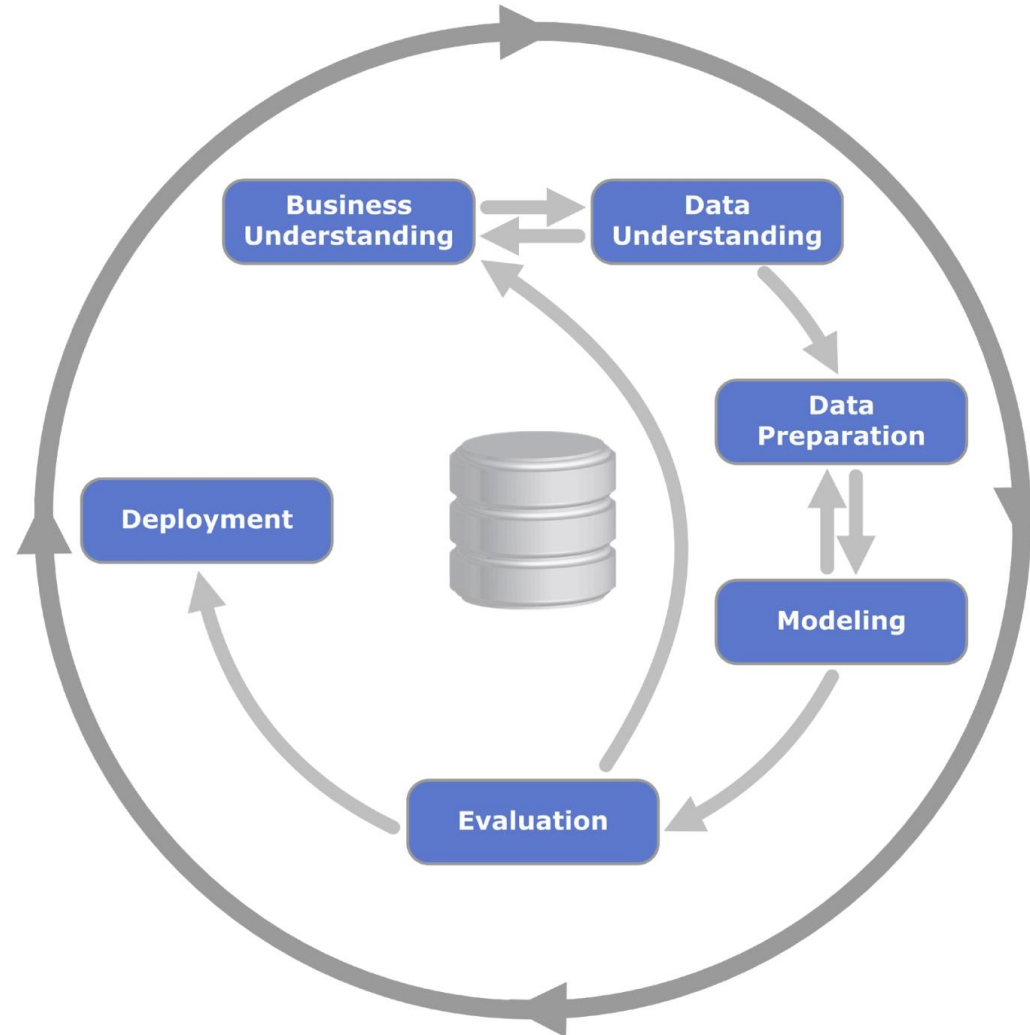
Moritz
Bunse



Christian
Koch



Shirin
Elsinghorst



Stefanie
Scholz



Christian
Winkler



Oliver
Zeigermann

https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

CRISP-DM → Cross Industry Standard Process for Data Mining

- 1) **Business Understanding: Anwendungsfall diskutieren und Geschäftsziel verstehen**
 - Verständnis der Ziele und Anforderungen aus der Business Perspektive & Transfer in die „Data Science-Welt“
 - **Business-Anforderungen** mit **möglichen Daten** korrelieren
 - **Realistische** Anforderungen stellen
 - **Interdisziplinäre** Zusammenarbeit besonders wichtig



CRISP-DM → Cross Industry Standard Process for Data Mining



2) Data Understanding

→ Data Profiling = grundlegende Datenanalyse und Exploration

- Bestimmung von Häufigkeiten, Wertebereichen, Korrelationen und Verteilungen
- Analyse der Datenqualität (z.B. fehlende Werte, Ausreißer, Aktualität)

→ Welche Arten von Datenquellen kommen in Frage?

- Lexika (z.B. Wikipedia)
- Redaktioneller Content (kuratiert)
- User Generated Content

Vorab Data Engineering (Daten stehen bereits zur Analyse gut einlesbar zur Verfügung)

CRISP-DM → Cross Industry Standard Process for Data Mining



3) Data Preparation (Datenvorbereitung)

- Vorbereitung bzw. Aufbereitung der Input-Daten für das Data Mining
- Auswahl von Tabellen und Attributen, Festlegung von Filterbedingungen
- Transformation von Wertebereichen (z.B. Diskretisierung, Normalisierung)
- Datenbereinigung (z.B. Nullwerte, Ausreißer)

→ Analysemethoden

- (Deskriptive) Statistik
- Clustering, Selbstordnung
- Regression, Klassifikation und Anreicherung
- Zeitreihen, Trends und Vorhersagen

```
hidden_size = 256
learning_rate = 0.001

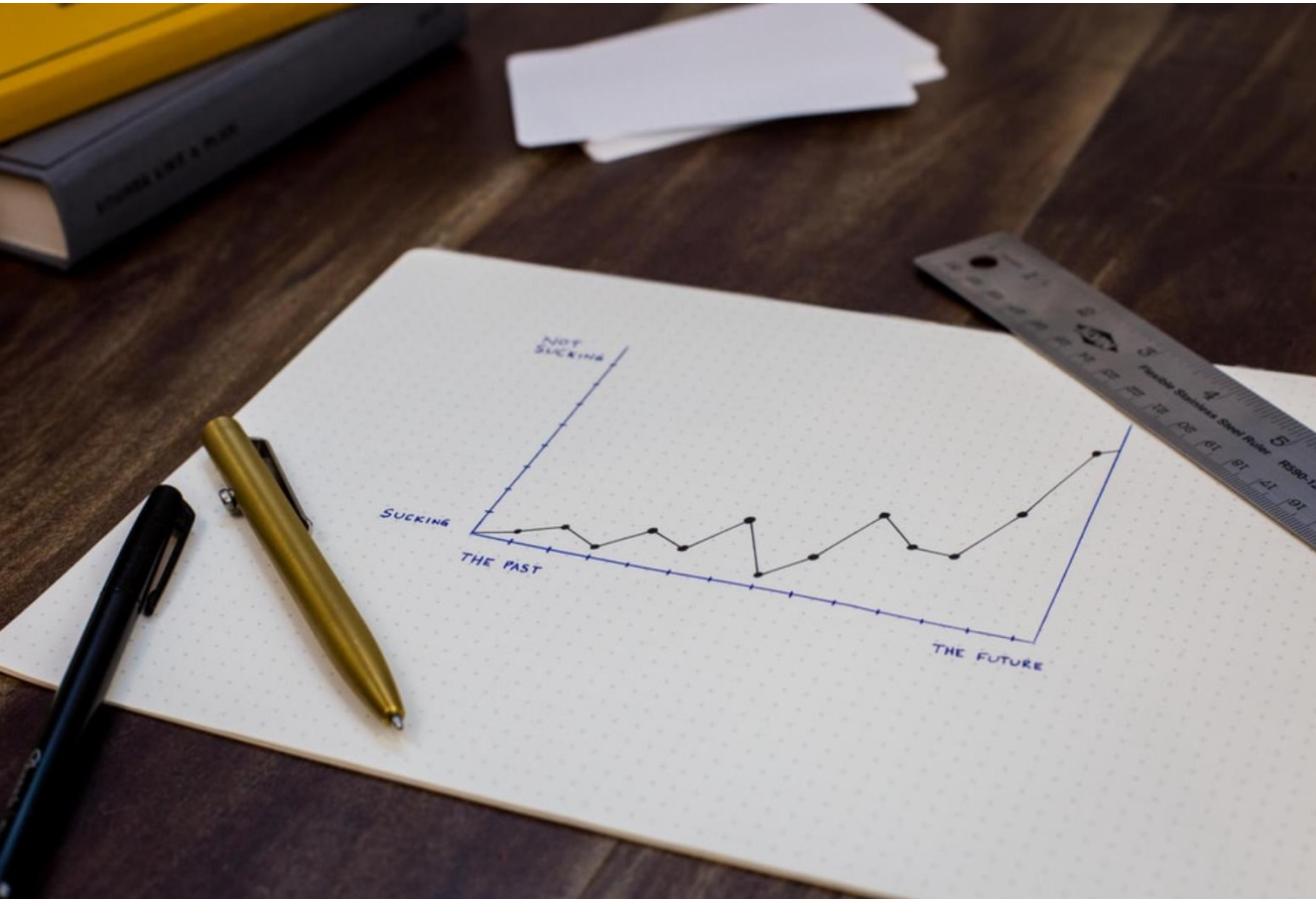
model = MyRNN(num_letters, hidden_size, num_langs)
criterion = nn.CrossEntropyLoss()
optimizer = torch.optim.Adam(model.parameters(), lr=learning_rate)

num_epochs = 2
print_interval = 3000

for epoch in range(num_epochs):
    random.shuffle(train_dataset)
    for i, (name, label) in enumerate(train_dataset):
        hidden_state = model.init_hidden()
        for char in name:
            output, hidden_state = model(char, hidden_state)
            loss = criterion(output, label)

        optimizer.zero_grad()
        loss.backward()
        nn.utils.clip_grad_norm_(model.parameters(), 1)
        optimizer.step()
```


CRISP-DM → Cross Industry Standard Process for Data Mining



4) Modeling (*eigentliches* Data Mining)

- Unterschiedliche Modellierungstechniken
- Nutzung von Data Mining zur Erstellung eines Modells (ggfs. erneut Data Preparation erforderlich)

→ Oliver Zeigermann: Machine Learning-Model (Sentiment-Detection & Sprachmodelle)

→ Christian Winkler: Trend-Detection-Modell & Jupyter Voilà

CRISP-DM → Cross Industry Standard Process for Data Mining

5) Evaluation (Bewertung)

- Modell ist erstellt und muss nun auf seine Qualität hin bewertet werden
- Berechnung von Qualitätsmetriken
- Überprüfung der Rahmenbedingungen und Einsatzfähigkeit
- Erfüllt das Modell den Geschäftszweck?
→ Relevant für Evaluation: Verständliche und ansprechende **Aufbereitung der Analyseergebnisse**

→ Shirin Elsinghorst: Data Storytelling & Visualisierung



CRISP-DM → Cross Industry Standard Process for Data Mining



6) Deployment (Anwendung, Operationalisierung)

- Anwendung des Modells im Entscheidungsprozess
- Integration von Scorings oder Regeln in operative Prozesse zur manuellen oder automatisierten Entscheidungsfindung
- Kontinuierliche Darstellung und Einbettung in operative Entscheidungsfindungsprozesse, z.B. mittels **Dashboards**

→ Christian Koch & Moritz Bunse:
Operationalisierung von Data Science
(MLOps)

Business Understanding: Use Case

02

Use Case "Automobilhersteller" → Branchenreports



Vom Auto zum Algorithmus

Autos zu bauen reicht nicht. Entscheidend ist, Bedürfnisse zu verstehen und zu bedienen.



Branchenreport Ausgabe 2020

AUTOMOBIL | GENIOS BranchenWissen Nr. 11 vom 10.11.2020



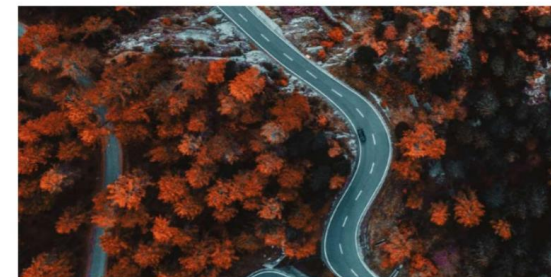
Branchenreport Automotive

Die deutsche Automobilindustrie und das Ende des „Weiter so“!



Roadmap für die Automobilität der Zukunft

Grundstein einer Innovationspartnerschaft



Use Case "Automobilhersteller" → Ergebnisse aus Beratungsanalysen



Ausgangssituation

- Ergebnisse aus Branchenreports zeigen:

Innovationsdruck steigt für gesamte Branche

- Analyse durch Unternehmensberatung zeigt:

- massiver **externer Druck** (Porter's Five Forces)
 - bestehender Wettbewerb, neue Wettbewerber (Google, Apple), Kulturwandel bei Verbrauchern („Sustainability“: Car Sharing, E-Mobility)
 - **Interne Schwächen** (SWOT, insb. fehlendes Verständnis über „Customer's Voice“)
- Wunsch aus Management nach
- „**Schnellen** Erkenntnissen“ (keine Zeit für Wellen-/Panelbefragungen)
 - Identifikation von **Trendthemen**
 - Ableitung **konkreter** Implikationen für Positionierung
 - „Data driven Insights mittels **AI**“



Use Case "Automobilhersteller" → Business Understanding



→ **Commitment** bzgl. konkreter Fragestellungen der Stakeholder:

- Welche Themen werden von (potentiellen) Kunden bzw. Endverbrauchern bzgl. **innovativer Technologien** im Automobil-Sektor diskutiert?
- Können **Trends** im Bereich moderner Technologien identifiziert werden?
- Wie stehen Verbraucher bestimmten Themen gegenüber (**Assoziationen, Einstellungen**)?
- Differenzierungspotential über **Mobility als Service**?

→ Ziel: Besseres Verständnis über relevante Themen bei sog. „**Early Adopters**“, d.h. **Technologie-affinen** Verbrauchern für Positionierung und externe Kommunikation

→ Wichtig: **Erwartungsmanagement** bei internen Stakeholdern

→ Unrealistische Erwartungen frühzeitig „einfangen“

→ Iterativer, interdisziplinärer Analyseprozess



Datenquelle: Soziale Netzwerke

03

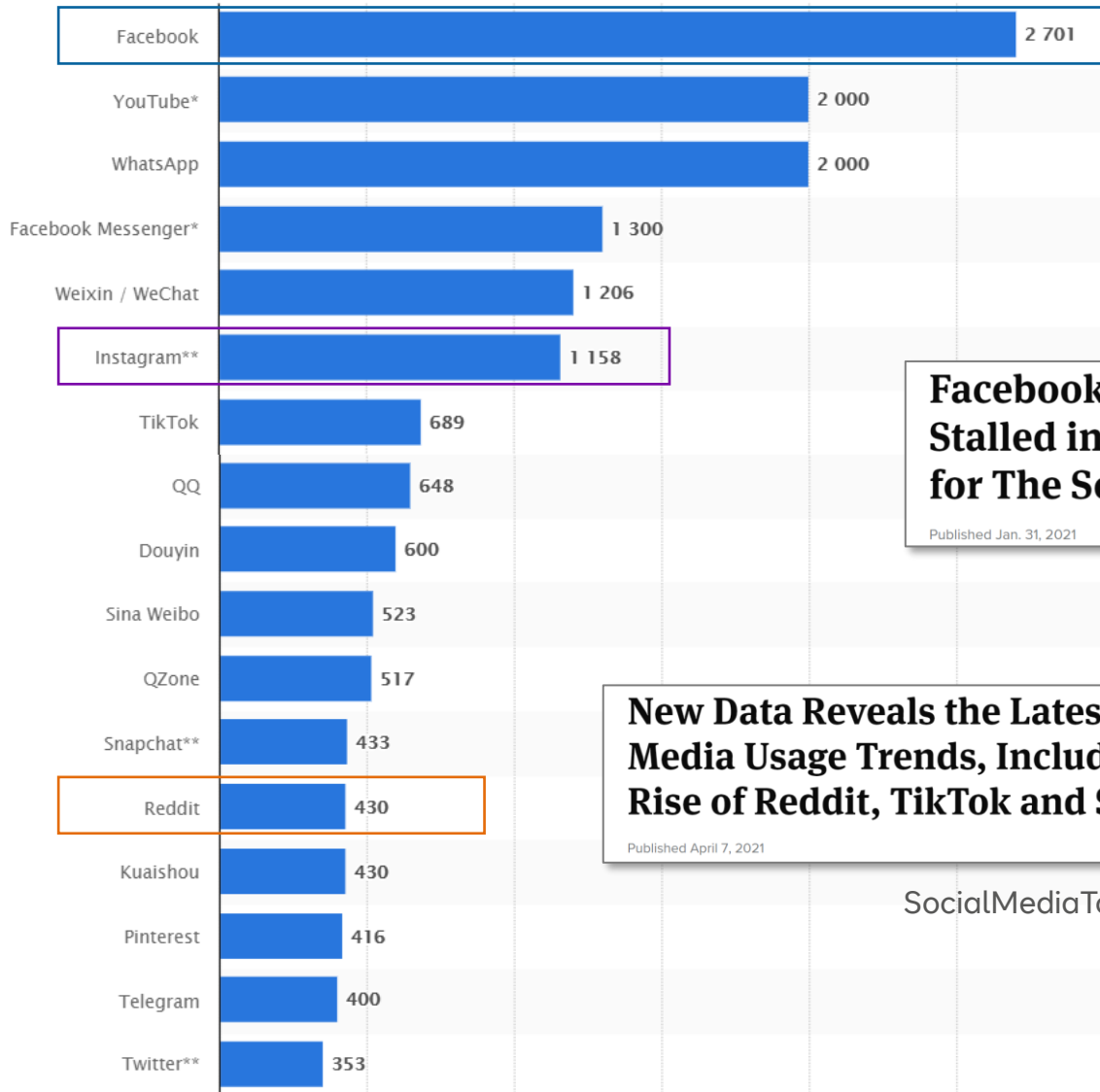
Datenquelle mit großem Potential: User Generated Content



Wie kommt man an valide Daten über bestimmte Zielgruppen **ohne Surveys**?

- **Vorhandene Informationen** nutzen
- **UGC**: enorme, unstrukturierte Datenmenge – automatisierte Auswertung mittels NLP nötig
- Bis vor kurzem Content nicht **systematisch handelbar** (nur „händisch“)

Soziale Netzwerke – aktuelle Entwicklungen



Wie sehr die Facebook-Dominanz quer durch alle Altersgruppen seit 2015 gesunken ist

28. April 2021, Autor: Michael Kroker

Seit 2015 ist Facebook als meistgenutztes soziales Netzwerk von gut zwei Drittel auf erstmals weniger als Hälfte aller Social-User gesunken.

Wirtschaftswoche (2021)

Facebook's Daily Active Usage Has Stalled in the US - A Sign of Concern for The Social Network?

Published Jan. 31, 2021

SocialMediaToday (2021)

New Data Reveals the Latest Social Media Usage Trends, Including the Rise of Reddit, TikTok and Snapchat

Published April 7, 2021

SocialMediaToday (2021)

HOME > TECH

Facebook reported a decline of 2 million daily active users in the US and Canada

BusinessInsider (2020)

- ➔ Facebook im Downturn, verliert User
- ➔ Instagram wächst seit 2013 (zunehmend langsamer)
- ➔ **Reddit** unter den Top 10-Websites (7) in den USA (Instagram 18) vs. Top 20 (18) global (Instagram 22) gemessen an MAUs und DAUs

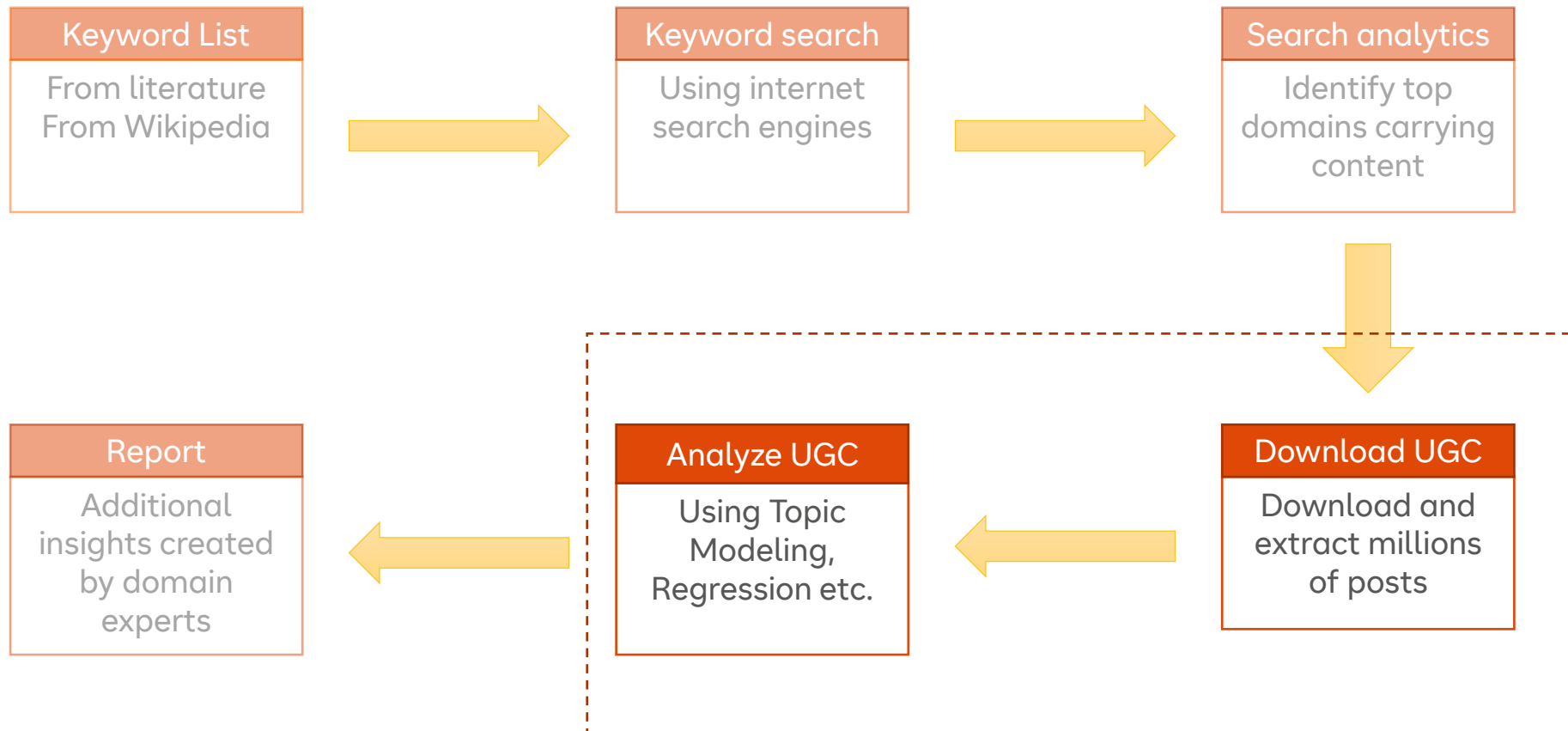
Alexa (2021), Hypestat (2021), pewresearch.org (2019)

Leading social networks worldwide (Oct. 2020), ranked by number of active users (in millions), Statista (2020)

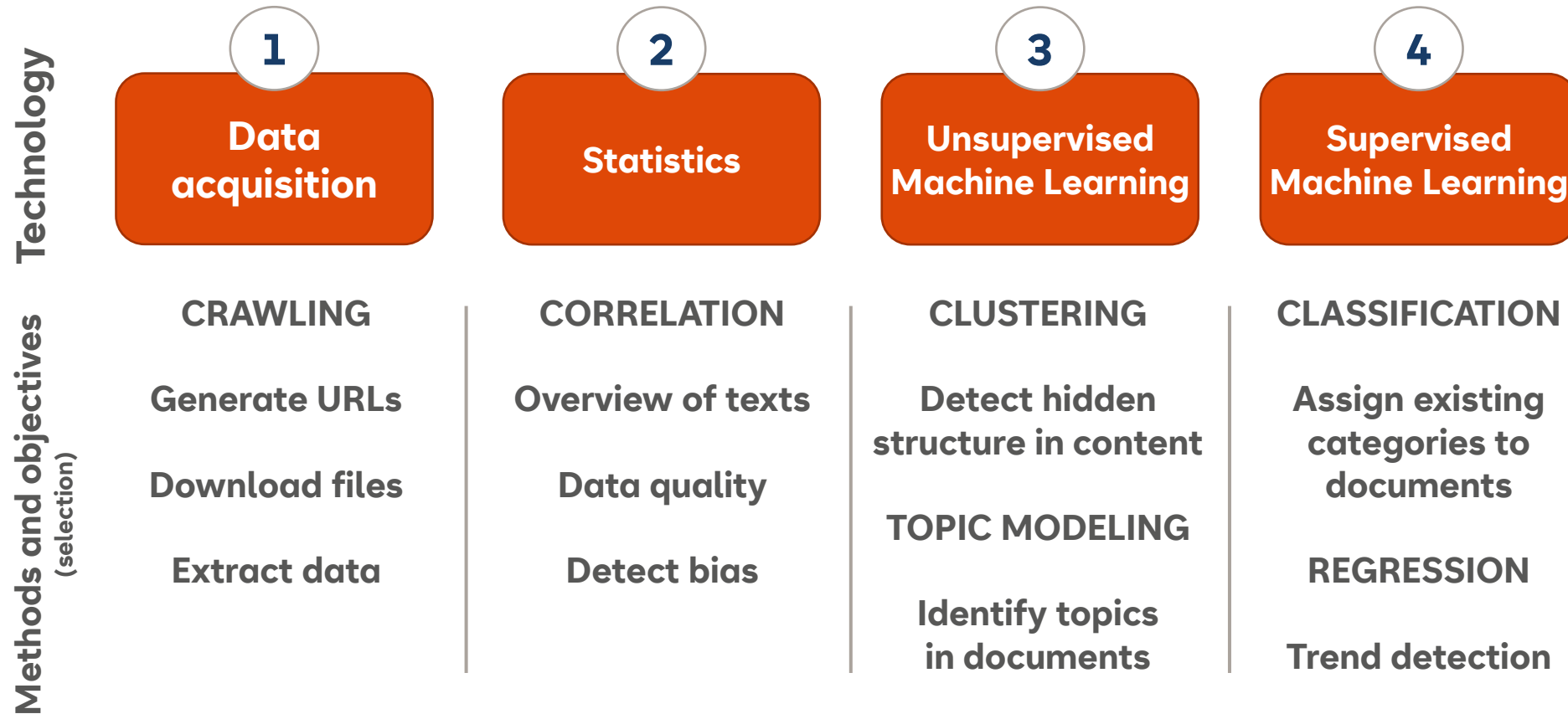
Projektablauf einer UGC-Analyse

04

Projektphasen



Technologien – Teil 1



Technologien – Teil 2 (hier nicht verwendet)

Technology	5	6	7	8
	Word Embeddings	Contextualized Embeddings	Transfer Learning	Question Answering
Methods and objectives (selection)	WORD SEMANTICS	CONTEXT	TRANSFER	UNDERSTANDING
	<p>Word similarity</p> <p>word2vec</p> <p>fastText</p> <p>GloVe</p>	<p>Words not as isolated entities</p> <p>Build language model</p> <p>Model meaning of words</p>	<p>Training with large corpus</p> <p>Billions of parameters</p> <p>Transfer to specific data</p>	<p>Use transfer learning</p> <p>Additional training with SQuAD corpus</p> <p>Answer question = prediction</p>

Unser Beispiel: Subreddit „r/technology“

05

Beispiel "r/technology"

Inhaltlich

- Inhaltlich relevanter Content
- Aktualität
- Ergiebige Auswertungsdimensionen

Technisch


- API vorhanden
- Download erlaubt

Über diese Community

Subreddit dedicated to the news and discussions about the creation and use of technology and its surrounding issues.

11.8m
Mitglieder

62.6k
Online

 Am 25. Jan. 2008 erstellt

 r/technology Themen

Technology

Nach Flair filtern

TechnologySupport

Business

Society

Software

Privacy

Social Media

Nanotech/Materials

Transportation

Space

Politics

Networking/Telecom

Hardware

Beispiel: Hierarchische Struktur bei Reddit



↑ 26.6k ↓ | Twitter to accept Elon Musk's \$45 billion bid to buy company Business

iv131012 · vor 3 Std.
tell me how I can do this, because I may want to try this.
↑ 906 ↓ Antworten Teilen ...

TheTjallan · vor 3 Std.
First of all, become a billionaire
↑ 1.2k ↓ Antworten Teilen ...

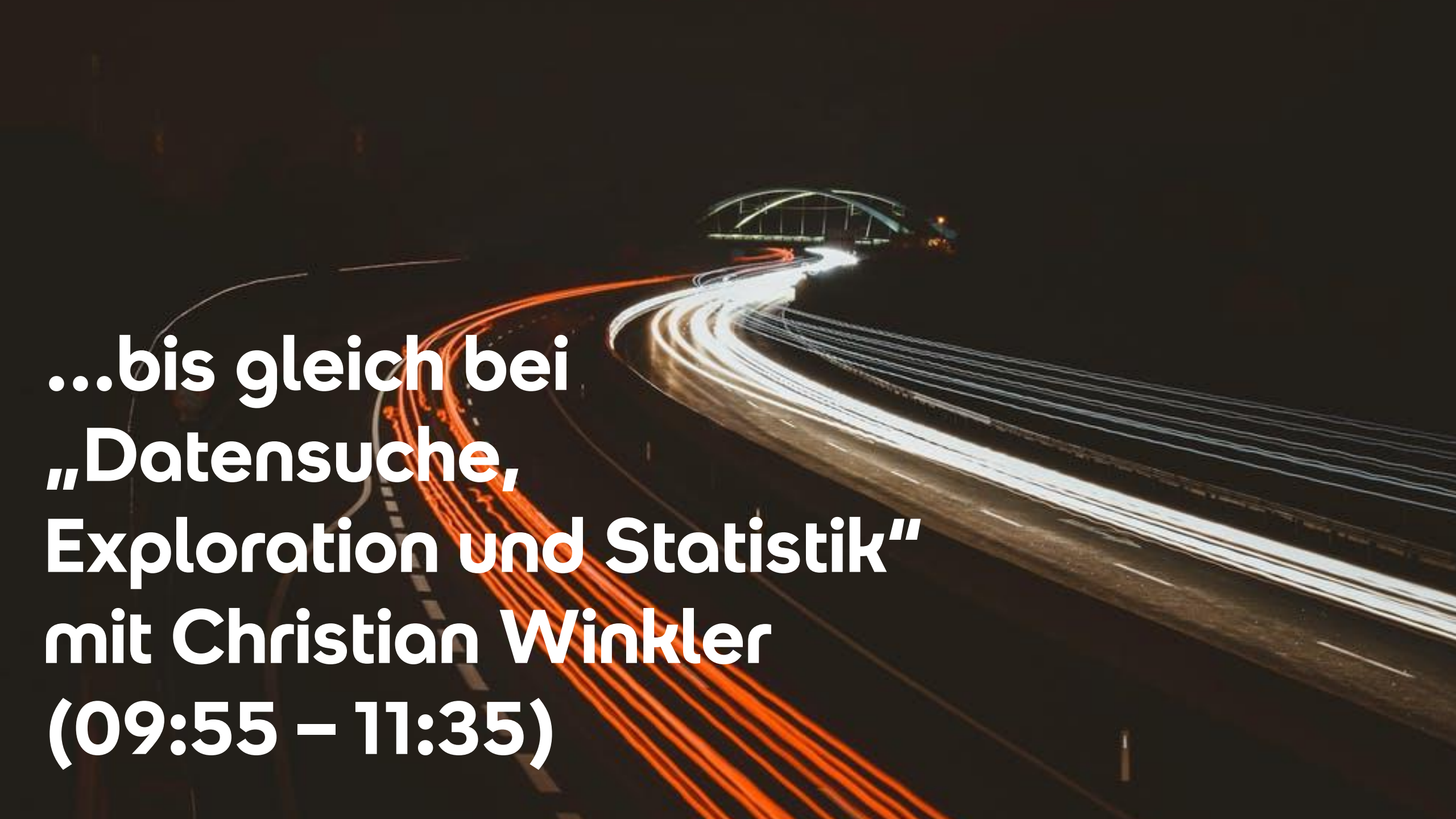
rcjlfk · vor 3 Std.
Seems easy. Step 2?
↑ 461 ↓ Antworten Teilen ...

emdave · vor 3 Std.
Don't not be a billionaire...
↑ 337 ↓ Antworten Teilen ...

arewehavinfunyet · vor 3 Std.
Okay i think that can easily be achieved in about 6 months according to some influencer if i click the link in their bio. Then what?
↑ 223 ↓ Antworten Teilen ...

drakefish · vor 2 Std.
Step 3: Get enough followers on twitter to manipulate the market using cryptic tweets
↑ 112 ↓ Antworten Teilen ...
[23 weitere Antworten](#)

MrDERPMcDERP · vor 2 Std.
Do shit posts all day
↑ 30 ↓ Antworten Teilen ...
[4 weitere Antworten](#)
[7 weitere Antworten](#)
[2 weitere Antworten](#)

A long-exposure photograph of a highway at night. The image shows multiple lanes of traffic with light trails from cars. On the left, there are several bright orange and red trails, likely from taillights. On the right, there are many white and blue trails, likely from headlights. In the background, a bridge with a curved arch is visible, illuminated with blue and white lights. The overall scene is dark, with the light trails providing the main source of illumination.

**...bis gleich bei
„Datensuche,
Exploration und Statistik“
mit Christian Winkler
(09:55 – 11:35)**