

DATA SCIENCE 2

VORLESUNG - DATENSCHUTZ, ETHIK

PROF. DR. CHRISTIAN BOCKERMANN

HOCHSCHULE BOCHUM

SOMMERSEMESTER 2021

1 Daten und Analysen

2 Datenschutz

3 Learning Analytics

22.6. um 13 Uhr

- Vorstellung/ Besprechung der Gruppen Projekte
- Vortrag: Bot-Erkennung in Web-Daten (Projekt)

29.6. um 13 Uhr

- Status: Projekt, Fragerunde,...
- Vortrag: Big Data

6.7. um 13 Uhr

- Status: Projekt, Fragerunde,...
- Vortrag/Diskussion: Datenanalyse, Datenschutz und Ethik

13.7. um 13 Uhr

- Gastvortrag der Fa. CuraCon (Wirtschaftsprüfung)
*Das Geheimnis von der Datenanalyse - und warum sie den
Wirtschaftsprüfer (noch) nicht ersetzt!*

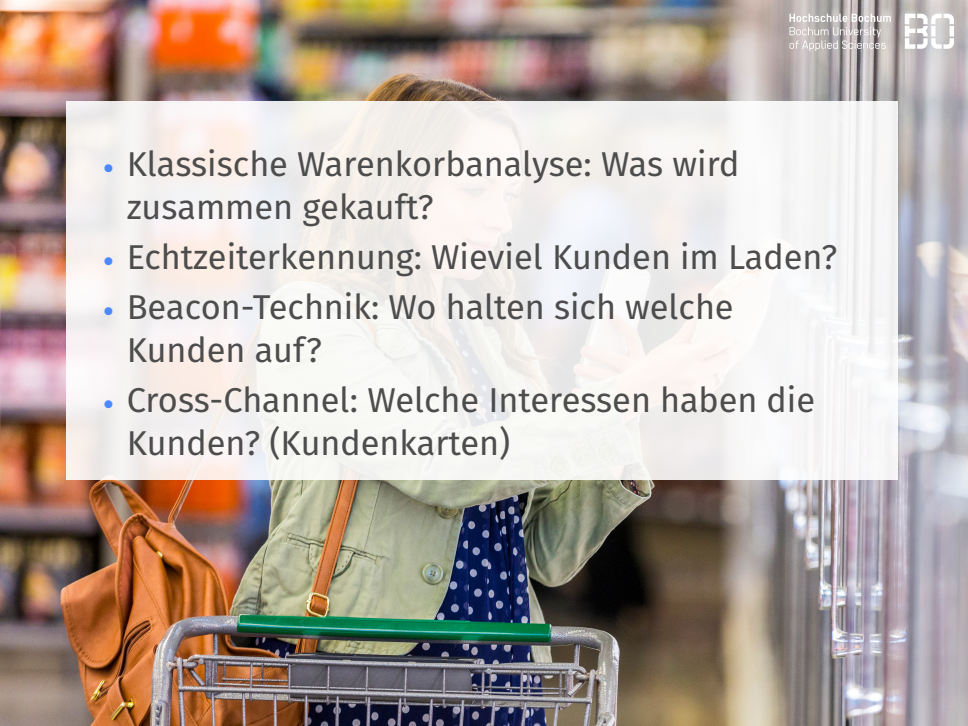
Vorwort

Diese Vorlesung erhebt **keinen Anspruch** auf

- Rechtliche Korrektheit und Verbindlichkeit
- Vollständigkeit und fachliche Korrektheit

A woman with long brown hair, wearing a light green jacket over a blue polka-dot top, is standing in a supermarket aisle. She is holding a white smartphone in her right hand and a brown paper bag in her left hand, looking at the phone. She has a brown shoulder bag and is pushing a metal shopping cart. The background shows shelves stocked with various products, including orange and red packages.

Datenanalyse im Handel

- 
- A woman with brown hair, wearing a light green jacket over a blue polka-dot dress, is standing in a supermarket aisle. She is looking down at a smartphone held in her hands. In front of her is a silver shopping cart with a green handle. The background is blurred, showing shelves of various products in a brightly lit store.
- Klassische Warenkorbanalyse: Was wird zusammen gekauft?
 - Echtzeiterkennung: Wieviel Kunden im Laden?
 - Beacon-Technik: Wo halten sich welche Kunden auf?
 - Cross-Channel: Welche Interessen haben die Kunden? (Kundenkarten)

Warenkorb-Analyse

- Finden häufiger Mengen/Muster
- Vorlesung 6 in Data Science 2

Warenkorb-Analyse

- Finden häufiger Mengen/Muster
- Vorlesung 6 in Data Science 2

Kundenprofile

- Supermarkt kennt Produkte eines Kunden
- Über Kundenkarte zusätzlich Historie

Warenkorb-Analyse

- Finden häufiger Mengen/Muster
- Vorlesung 6 in Data Science 2

Kundenprofile

- Supermarkt kennt Produkte eines Kunden
- Über Kundenkarte zusätzlich Historie
- Kunden werden anhand ihrer Produkte gruppiert (Clustering, Vorlesung 5 in Data Science 2)
- Zielgerichtete Werbung, Promotions,...

Was weiss der Supermarkt?

- Einzugsbereich, Kundenaufkommen (Zeit)
- Einzelne Bons, Produktverkauf nach Tageszeit

Was weiss der Supermarkt?

- Einzugsbereich, Kundenaufkommen (Zeit)
- Einzelne Bons, Produktverkauf nach Tageszeit

# last 12 M	\$ last 12 M	Class	Schoki
4	623,38	Low	X
6	1.492,23	High	X
7	914,98	Mid	
8	1.378,43	High	
3	416,18	Low	X



Was weiss der Supermarkt?

- Einzugsbereich, Kundenaufkommen (Zeit)
- Einzelne Bons, Produktverkauf nach Tageszeit

# last 12 M	\$ last 12 M	Class	Schoki
4	623,38	Low	X
6	1.492,23	High	X
7	914,98	Mid	
8	1.378,43	High	
3	416,18	Low	X

Katzen	Vegan
X	
	X
X	X
	X
X	



Was weiss der Supermarkt?

- Einzugsbereich, Kundenaufkommen (Zeit)
- Einzelne Bons, Produktverkauf nach Tageszeit

# last 12 M	\$ last 12 M	Class	Schoki
4	623,38	Low	X
6	1.492,23	High	X
7	914,98	Mid	
8	1.378,43	High	
3	416,18	Low	X

Katzen	Vegan
X	
	X
X	X
	X
X	

Sport	Mode
X	X
	X
X	X
	X



Was weiss der Supermarkt?

- Einzugsbereich, Kundenaufkommen (Zeit)
- Einzelne Bons, Produktverkauf nach Tageszeit

# last 12 M	\$ last 12 M	Class	Schoki
4	623,38	Low	X
6	1.492,23	High	X
7	914,98	Mid	
8	1.378,43	High	
3	416,18	Low	X

Katzen	Vegan
X	
	X
X	X
	X
X	

Sport	Mode
X	X
	X
X	X
	X



3,522,031 views | Feb 16, 2012, 11:02am

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did





EXPLANATION OF BENEFITS

THIS IS NOT A BILL

Date: 06/20/13
Policy:

Provider of Service
NIC HOSPITAL INC
5/20/13- 5/20/13

Billed Charges	Not C...	eductible	Copy	Total P... Cost
20.76	101.76	20.76	449.00	
4.76	20.76	20.76	449.00	
20.76	20.76	20.76	449.00	
20.76	20.76	20.76	449.00	

Datenanalyse in der Medizin

Datenanalyse in der Medizin

- Erkennen von Krankheiten
(vgl. "Husten" – Vorlesung 5, Data Science 1)
- Individualisierung von Therapien/Medikamenten



Datenanalyse in der Medizin

- Erkennen von Krankheiten
(vgl. "Husten" – Vorlesung 5, Data Science 1)
- Individualisierung von Therapien/Medikamenten



Datenanalyse

HOM, 04.02.2014

Typ-2-Diabetes: Das kosten die Folgeerkrankungen



Datenanalyse in der Medizin

- Erkennen von Krankheiten
(vgl. "Husten" – Vorlesung 5, Data Science 1)
- Individualisierung von Therapien/Medikamenten



Datenanalyse

HDL 04.02.2014

Typ-2-Diabetes: Das kosten die Folgeerkrankungen



www.fotogrammi.de/Photo

Mit einem schmutzigen Trick kann Facebook rausfinden, wann Sie Ihre Periode haben

Sonntag, 23.02.2019, 10:15 – von FOCUS-Online-Redakteur Florian Boller

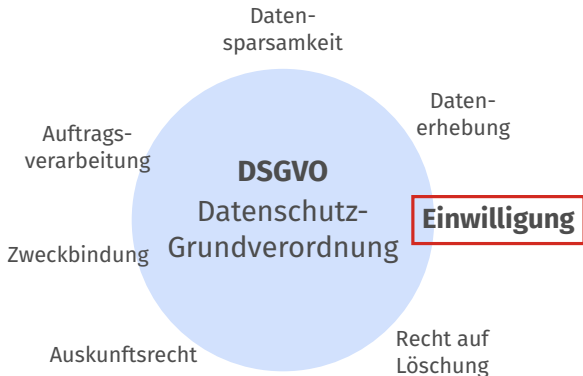


Doku: Das Dilemma mit den sozialen Medien (Netflix)



Datenschutz





Bevor Sie fortfahren

Google verwendet [Cookies](#) und Daten für Folgendes:

- Dienste anbieten und betreiben, z. B. Störungen prüfen und Maßnahmen gegen Spam, Betrug oder Missbrauch ergreifen
- Daten zu Zielgruppeninteraktionen und Websitestatistiken erheben, um zu verstehen, wie unsere Dienste verwendet werden

Wenn Sie zustimmen, verwenden wir Cookies und Daten auch für Folgendes:

- Qualität unserer Dienste verbessern und neue Dienste entwickeln
- Werbung ausliefern und die Effektivität von Werbung messen
- Personalisierte Inhalte anzeigen, abhängig von Ihren Einstellungen
- Personalisierte oder allgemeine Werbung bei Google und im Web anzeigen, abhängig von Ihren Einstellungen

Bevor Sie fortfahren

Google verwendet [Cookies](#) und Daten für Folgendes:

- Dienste anbieten und betreiben, z. B. Störungen prüfen und Maßnahmen gegen Spam, Betrug oder Missbrauch ergreifen
- Daten zu Zielgruppeninteraktionen und Websitestatistiken erheben, um zu verstehen, wie unsere Dienste verwendet werden

Wenn Sie zustimmen, verwenden wir Cookies und Daten auch für Folgendes:

- Qualität unserer Dienste verbessern und neue Dienste entwickeln
- Werbung ausliefern und die Effektivität von Werbung messen
- Personalisierte Inhalte anzeigen, abhängig von Ihren Einstellungen
- Personalisierte oder allgemeine Werbung bei Google und im Web anzeigen, abhängig von Ihren Einstellungen

Anonymisierung

Name	Vorname	PLZ	# last 12 M	\$ last 12 M	Class
Fox	Peter	13124	4	623,38	
Lennon	John	12921	6	1.492,23	
Jackson	Peter	13174	7	914,98	
Tyson	Mike	12511	8	1.378,43	
Brown	Alice	13124	3	416,18	

Anonymisierung

Name	Vorname	PLZ	# last 12 M	\$ last 12 M	Class
	Peter	13124	4	623,38	
	John	12921	6	1.492,23	
	Peter	13174	7	914,98	
	Mike	12511	8	1.378,43	
	Alice	13124	3	416,18	

Anonymisierung

Name	Vorname	PLZ	# last 12 M	\$ last 12 M	Class
	Peter	13124	4	623,38	
	John	12921	6	1.492,23	
	Peter	13174	7	914,98	
	Mike	12511	8	1.378,43	
	Alice	13124	3	416,18	

Anonymisierung

Name	Vorname	PLZ	# last 12 M	\$ last 12 M	Class
	Peter	13XXX	4	623,38	
	John	12XXX	6	1.492,23	
	Peter	13XXX	7	914,98	
	Mike	12XXX	8	1.378,43	
	Alice	13XXX	3	416,18	

Pseudonymisierung

ID	PLZ	# last 12 M	\$ last 12 M	Class
001	13XXX	4	623,38	
002	12XXX	6	1.492,23	
003	13XXX	7	914,98	
004	12XXX	8	1.378,43	
005	13XXX	3	416,18	

Personenbezogene Daten
durch Pseudonym ersetzen

China: Überwachungsstaat oder Zukunftslabor?



China: Überwachungsstaat oder Zukunftslabor?

31.05.2021 · [Reportage & Dokumentation](#) · Das Erste



China baut ein riesiges digitales Überwachungssystem auf. Beim Staat laufen gigantische Datenmengen zusammen - und die Bürger machen bereitwillig mit. Denn die Angebote sind praktisch - und wer sich an die Regeln hält, wird belohnt.

Video verfügbar:

bis 31.05.2022 -
23:59 Uhr

Learning Analytics

Learning Analytics – Wie gut ist (e-) Learning?

Learning Analytics – Wie gut ist (e-) Learning?

- Messen und Analyse von (e-)Learning
- Gezielte Förderung von Studierenden
- Planung/Gestaltung von Studienangeboten (vgl. Umfrage zu Master in DataScience)
- Privacy/Datenschutz hat höchste Priorität

Beispiel: Log Daten des Data-Science Servers

- Zugriffsprotokoll auf Web-Seite
- Zweck ist die technische Überwachung des Servers

Web-Server Access Logs

Typische Felder sind:

IP Adresse	Quelle IP Adresse der Verbindung
Datum	Datum+Uhrzeit des Zugriffs
Method+URI	Zugriffsmethode und URL
Status-Code	Status der Antwort (Ok, Fehler,..)
Referer	Vorangegangene URL
User-Agent	Bezeichnung des Browsers

Beispiel: datascience.hs-bochum.de

- Log Daten von 5.10.2020 bis 21.6.2021
- 1.405.942 Anfragen in ca. 260 Tagen
- von 1200 verschiedene IP-Adressen
- 24.419 verschiedene URLs
- 326 verschiedene Browser-Typen

Fragestellung

Führt die Nutzung des Jupyter-Servers zu besseren Noten in den Hausarbeiten/Klausuren des Kurses Data Science?

Fragestellung

Führt die Nutzung des Jupyter-Servers zu besseren Noten in den Hausarbeiten/Klausuren des Kurses Data Science?

- Nutzungsverhalten pro Teilnehmer:in
- Vorhersage: Nutzungsverhalten \leftrightarrow Note?

Fragestellung

Führt die Nutzung des Jupyter-Servers zu besseren Noten in den Hausarbeiten/Klausuren des Kurses Data Science?

- Nutzungsverhalten pro Teilnehmer:in
- Vorhersage: Nutzungsverhalten <-> Note?
- **Aktivitätsprofile**: Nachteile oder nicht?

Fragestellung

Führt die Nutzung des Jupyter-Servers zu besseren Noten in den Hausarbeiten/Klausuren des Kurses Data Science?

- Nutzungsverhalten pro Teilnehmer:in
- Vorhersage: Nutzungsverhalten <-> Note?
- **Aktivitätsprofile**: Nachteile oder nicht?
- **Endgeräte**: Wer nutzt iPhone/Android/Windows/Linux/Mac?

Wieviel würden Sie preisgeben?

Learning Analytics in Data Science:

- Teilnahme an der Vorlesung/Übung
- Studiengang, Berufsziel
- Interessen, Hobbies
- Dauer der Nutzung des Servers
- Dauer der Arbeit an bestimmten Notebooks/Übungsblättern
- Bearbeitung/Nicht-Bearbeitung von einzelnen Ü-Aufgaben
- Exakte Zeitpunkte der Nutzung des Servers
- Endgerät, Browser,...
- Ungefährer Ort (Stadt) bei der Nutzung

Welche Plattformen/Apps nutzen Sie?

- TikTok, WhatsApp, Signal
- FaceBook, Twitter, Instagram
- ...