Applied Analytical Data Science
Teil 1: Überblick

Dr. Jörg-Uwe Kietz,
Vorlesung an der Univ. Zürich,
Mittwoch, 14:00-15:45 Uhr Vorlesung,
16:00-17:30 Uhr Übung

http://www.kietz.ch/AADS/

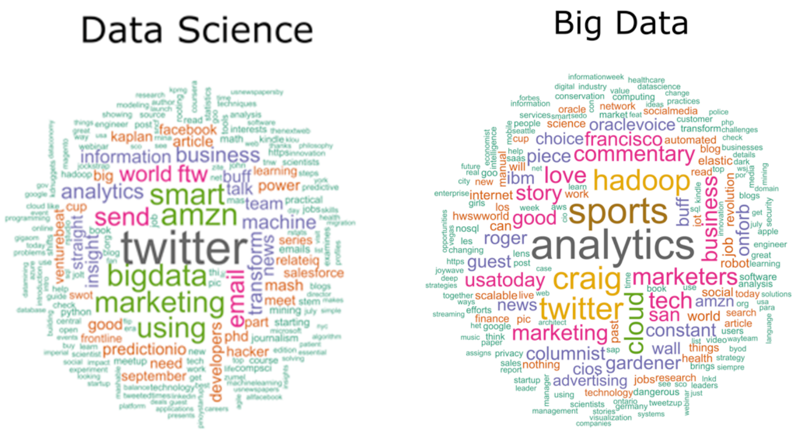
1

Heutiges Programm

- Analytical Data Science, Big Data or Data-/Text-Mining ⇔ How to use Machine Learning and Statistics Methods
- Data Mining, Wissensentdeckung in Datenbanken Was ist das?
- Wozu kann man Data Mining gebrauchen?
- Was braucht man zum Data Mining?
- Überblick über die Vorlesung

2

New Buzzwords: Data Science & Big Data



Twitter Post World Clouds for Data Science & Big Data, Source:
http://www.datasciencecentral.com/profiles/blogs/word-clouds-of-big-data-data-science-and-other-buzz-words

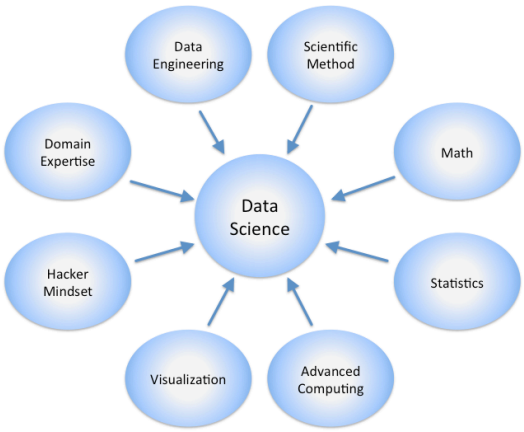
3

Data Scientist Job Posting Cloud

Source: http://blogs.gartner.com/svetlana-sicular/data-scientist-mystified/

4

Data Science a mash-up of disciplines



5

Heutiges Programm

- Analytical Data Science, Big Data or Data-/Text-Mining ⇔ How to use Machine Learning and Statistics Methods
- Data Mining, Wissensentdeckung in Datenbanken Was ist das?
- Wozu kann man Data Mining gebrauchen?
- Was braucht man zum Data Mining?
- Überblick über die Vorlesung

6

Wissensentdeckung in Datenbanken

Definitionen aus Fayad, Piatetsky-Shapiro & Smyth, 1996:

Knowledge Discovery in Databases (**KDD**) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable pattern in data.

7

Data Mining

Data Mining (Mustergewinnung, Statistik: Modellierung) ist ein Teil des KDD-Prozesses, der aus der Anwendung von Data Mining Algorithmen, die unter gewissen Ressourcenbeschränkungen ein Muster E_F aus einer gegebenen Faktenmenge F erzeugen.

- KDD umfaßt Data Mining als Teilprozeß, auch wenn Data Mining oft als Synonym für KDD verwendet wird.

8

Daten und Muster über Daten

Daten: Eine Menge F von Fakten (Fälle "cases", Beispiele "examples"), z.B.:

- Tupel einer relationalen DB,
- Sätze in einer Datei

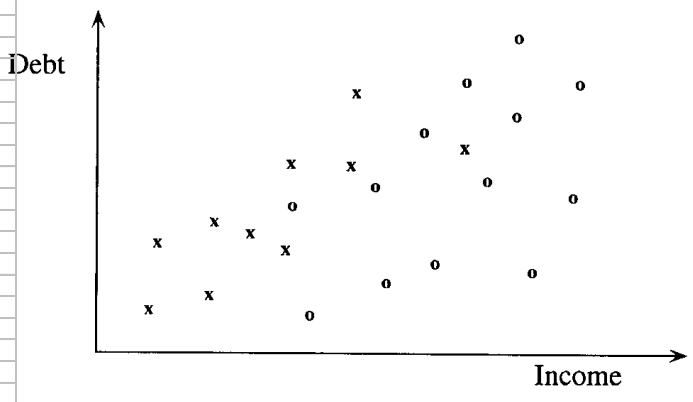
Muster („pattern", generiertes Wissen): Ein Ausdruck E einer Sprache L zur Beschreibung einer Teilmenge F_E von F. Wobei E einfacher (zu lesen, interpretieren, speichern, ...) als Aufzählung der Faktenmenge F_E ist, z.B.:

- Wertebeschränkung für DB-Felder
- Beziehung zwischen DB-Feldern
- Regeln

9

Beispiel: Daten und Muster

ID	Income	Debt	Defaulted
1	25	22	yes
2	30	58	yes
3	55	30	yes
4	58	70	yes
5	75	65	yes
6	92	55	yes
7	95	100	yes
8	95	78	no
9	105	20	no
10	125	100	yes
11	128	138	yes
12	135	88	no
13	142	35	no
14	160	118	no
15	165	48	no
16	180	108	yes
17	180	142	no
18	190	90	no
19	205	125	no
20	208	168	no
21	212	42	no
22	232	82	no
23	235	142	no



10

Beispiel: Daten und Muster

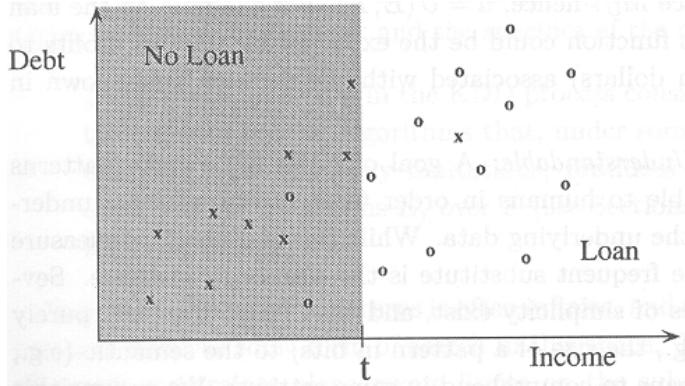


Figure 1.2
Using a single threshold on the income variable to try to classify the loan data set.

11

Beispiel: Daten und Muster

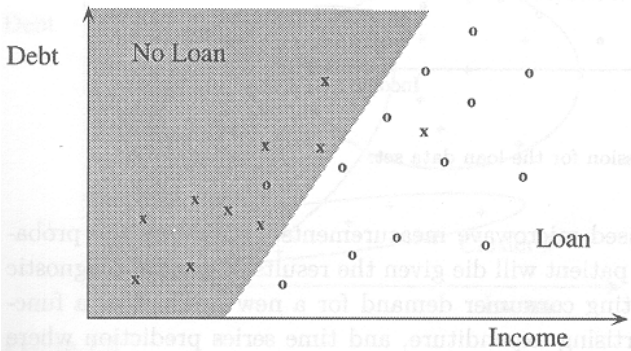


Figure 1.4
A simple linear classification boundary for the loan data set: shaded region denotes class "no loan."

12

Beispiel: Daten und Muster

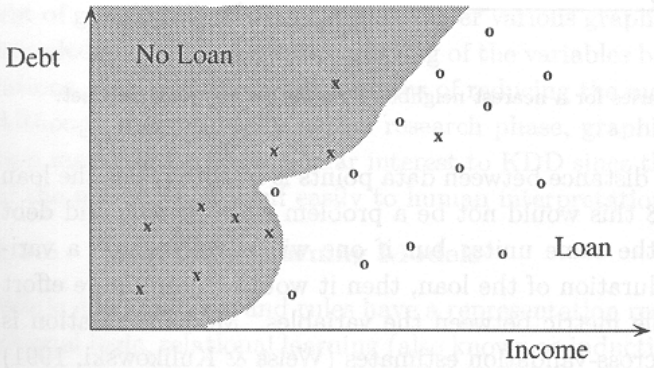


Figure 1.7
An example of classification boundaries learned by a non-linear classifier (such as a neural network) for the loan data set.

13

Beispiel: Daten und Muster

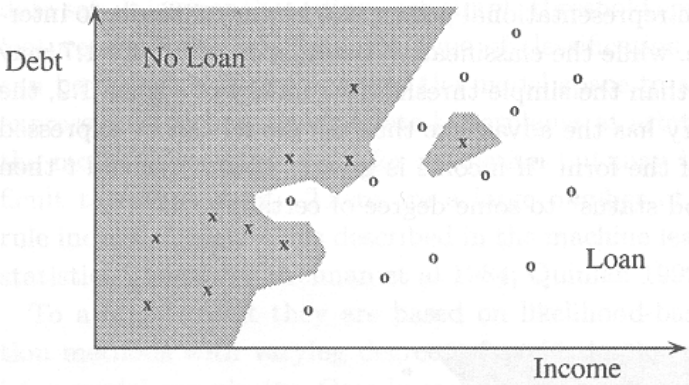


Figure 1.8
Classification boundaries for a nearest neighbor classifier for the loan data set.

14

Beispiel: Daten und Muster

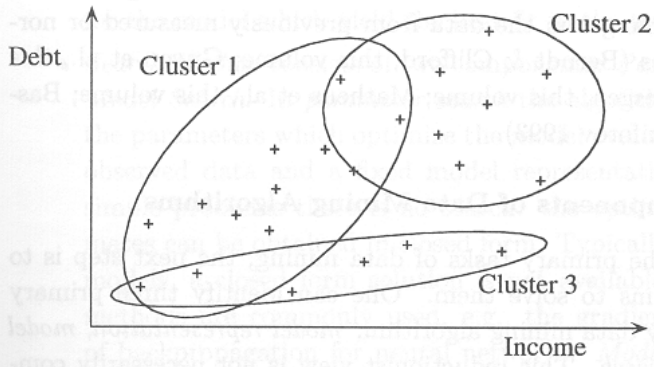


Figure 1.6
A simple clustering of the loan data set into 3 clusters. Note that original labels are replaced by +'s.

15

Beispiel: Daten und Muster

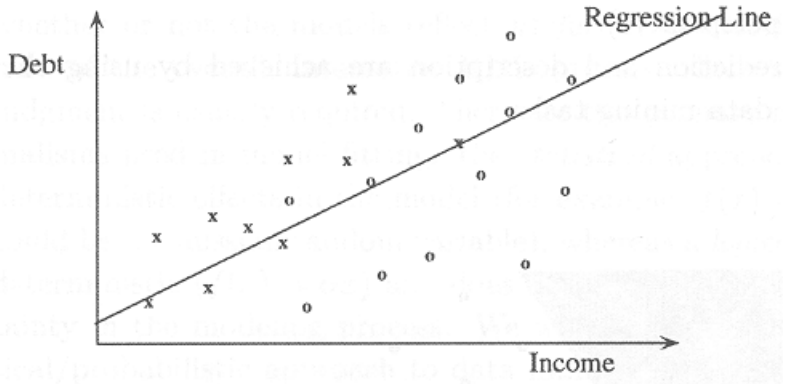


Figure 1.5
A simple linear regression for the loan data set.

16

Gültigkeit und Verständlichkeit von Mustern

Gültigkeit (validity): das gefundene Muster sollte mit gewisser Sicherheit für neue Daten zutreffend sein.

Verständlichkeit (ultimately understandable):gefundene Muster müssen für Menschen verständlich sein, d.h. wie die Muster am besten beschreiben/präsentieren?

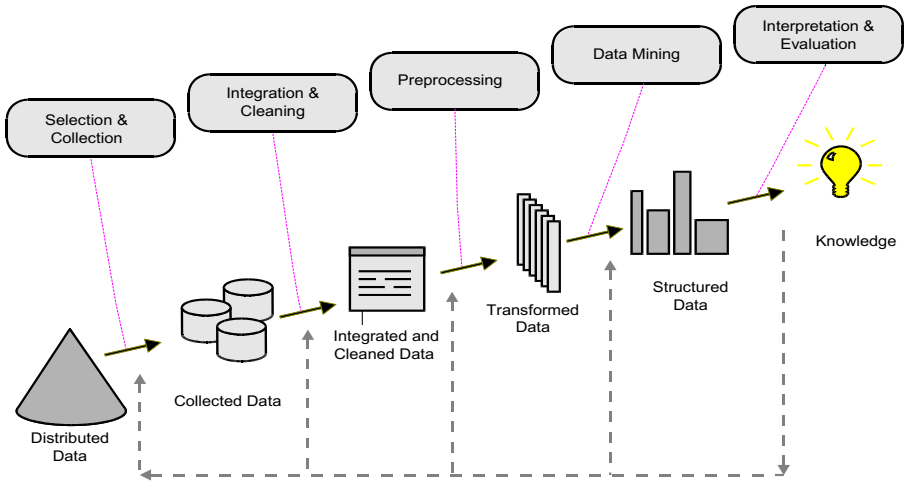
17

KDD-Prozess

KDD-Prozess ist der Prozess der nötig ist, um Data Mining Methoden (Algorithmen) zur Extraktion (Identifikation) von Wissen auf Daten F anzuwenden. Neben dem Data Mining umfasst er weitere Schritte, wie Dateninspektion und -aufbereitung, und Ergebnisevaluation und -interpretation.

18

Der KDD-Prozess



19

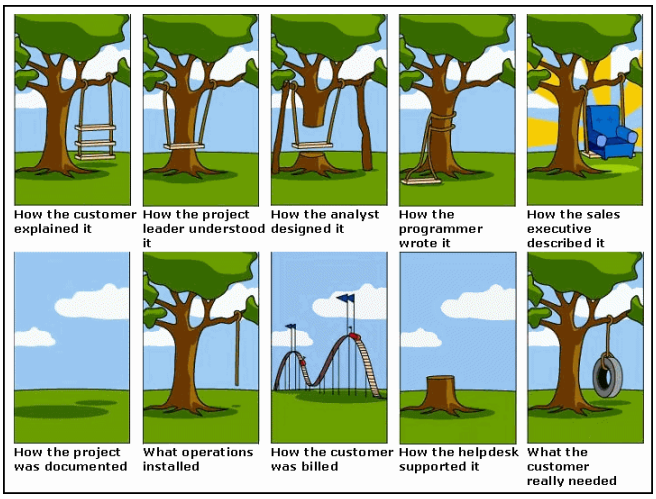
Prozessschrittdauer und -wichtigkeit

	Time	Importance
• Business Understanding	20	80
– a) Exploring the problem	10	15
– b) Exploring the solution	9	14
– c) Implementation Specification	1	51
• Data Preparation & Mining	80	20
– a) Data Preparation	60	15
– b) Data Exploration	15	3
– c) Modeling (data mining)	5	2

(Pyle, 2000)

20

Die Wichtigkeit des Business Understandings



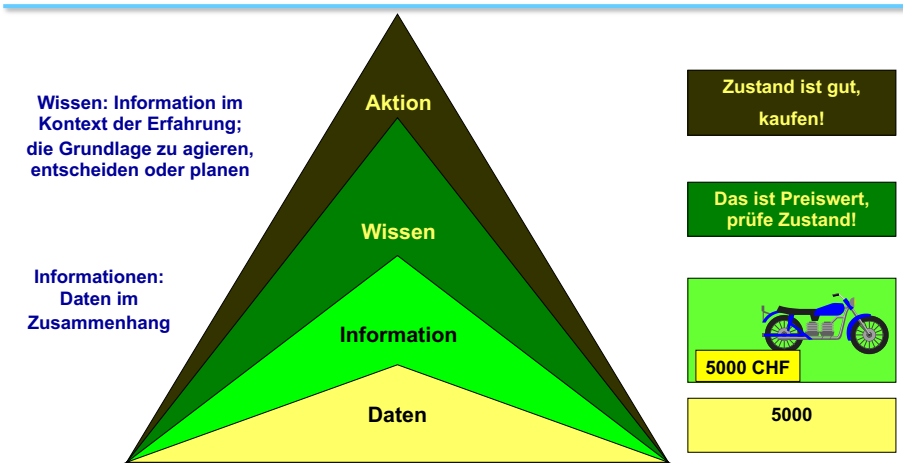
21

Daten, Informationen und Wissen

- Daten: Die Repräsentation von Fakten über Objekte der Welt.
  - "Alles was man erfassen (messen) und speichern kann."
- Informationen: Betrifft die Kommunikation von Daten, ein Datum kann zur Information werden, wenn es für den Empfänger informativ (neu, relevant, ...) ist.
  - "Die richtigen Daten zum richtigen Zeitpunkt."
- Wissen: Daten in einem Kontext zur Nutzung der Daten.
  - "Zu wissen, was mit den Daten zu tun ist."

22

Wissen ist die Grundlage der Aktion



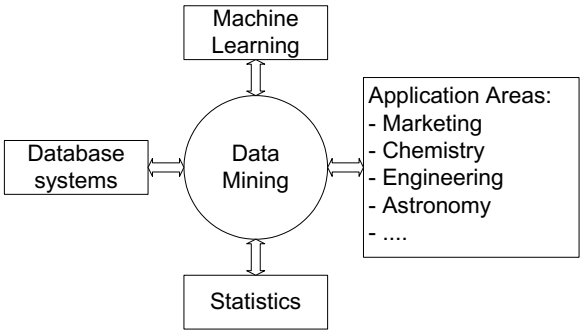
23

Kann Wissen durch Data Mining (KDD) in Daten entdeckt werden?

- Computer an sich erzeugen nur Daten aus Daten, d.h. Data Mining erzeugt „Meta-Daten“ über die Daten.
- **Aber**, diese durch Data Mining gewonnen „Meta-Daten“ können für den **Benutzer** (Datenanalysten) informativ sein, d.h. wichtige **Informationen** darstellen.
- Der **KDD-Prozess** soll (und kann) den notwendigen **Kontext** (Geschäftsverständnis, Problem & Lösungsanalyse) bereitstellen, in dem die gewonnenen „Meta-Daten“ zu wertvollem **Wissen** werden können.

24

Data Mining als Wissenschaft



Data Mining adaptiert Methoden des Maschinellen Lernens und der Statistik mit Hilfe von Datenbank-Techniken um sie auf grosse Datenmengen eines Anwendungsgebiets anwenden zu können, und dort neues Wissen zu produzieren.

25

Heutiges Programm

- Analytical Data Science, Big Data or Data-/Text-Mining ⇔ How to use Machine Learning and Statistics Methods
- Data Mining, Wissensentdeckung in Datenbanken Was ist das?
- Wozu kann man Data Mining gebrauchen?
- Was braucht man zum Data Mining?
- Überblick über die Vorlesung

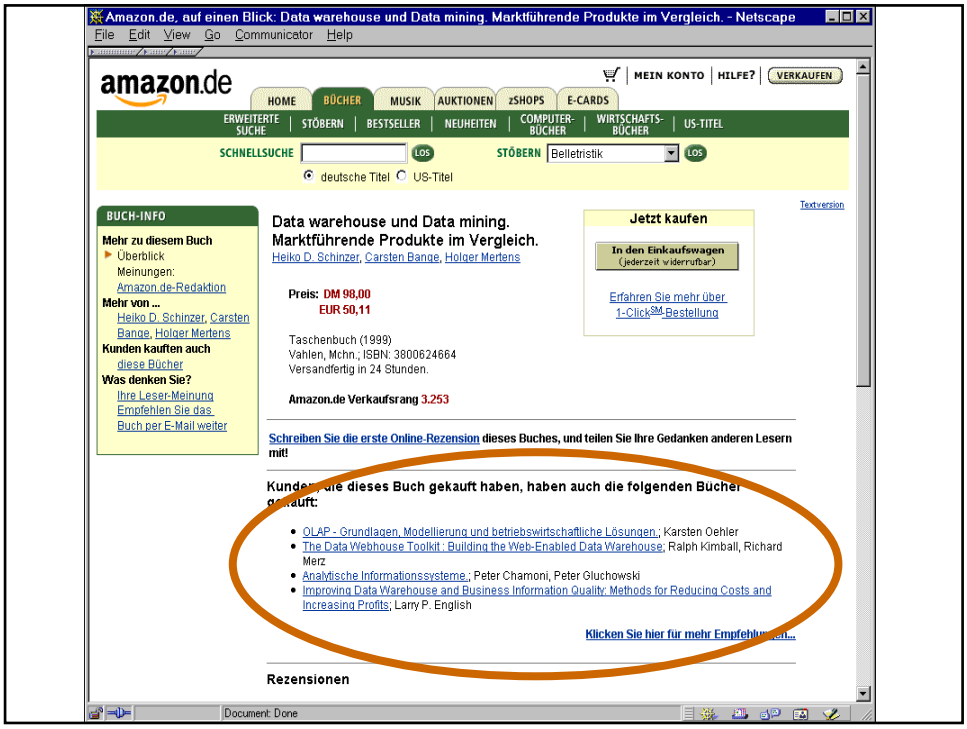
26

Wozu kann man Data Mining gebrauchen?

- Einzelhandel: Kaufverhalten der Kunden (Warenkorbanalyse)
- Marketing: Zielgruppenidentifikation, Marketing-Aktionen
- Banken: Kredit Risiko Bewertung, Missbrauchsentdeckung
- Telekommunikation: Missbrauchsentdeckung
- Produktion: Prozess Analyse
- Versicherungen: Individuelle Risiko Prämien
- e-Business: Individuelle Produkt (oder Banner) Vorschläge (yahoo, amazon)
- Wissenschaft: Auswertung von Messdaten, z.B. Satellitenfotos, Astronomie, ...

27

28



29

Heutiges Programm

- Analytical Data Science, Big Data or Data-/Text-Mining ⇔ How to use Machine Learning and Statistics Methods
- Data Mining, Wissensentdeckung in Datenbanken Was ist das?
- Wozu kann man Data Mining gebrauchen?
- Was braucht man zum Data Mining?
- Überblick über die Vorlesung

30

Was braucht man zum Data Mining?

- Einen Anwendungsproblem dessen Lösung den Aufwand des DM rechtfertigt.
- Daten zu analysieren, die für das Anwendungsproblem relevant sind.
  - z.B.: (Nicht-) Kundendaten, Produktdaten, Verkaufsdaten
  - Data Warehouse
- Eine Idee, wie das Anwendungsproblem mit DM zu lösen ist.
- Data Mining Tools für die verschiedenen Data Mining Tasks
  - z.B.: Assoziationsregeln, Klassifikation
- Zeit von Experten (Data Mining Experten, Anwendungsexperten).

31

Fazit

- Knowledge Discovery in Databases (KDD) ist ein Prozess.
- Data Mining ist der KDD-Prozessschritt der Analyse von Datenbanken mit Methoden der Statistik und des Maschinellen Lernens.
- Computer manipulieren Daten, aber das Ergebnis kann für den Benutzer eine Information darstellen oder gar zu Wissen werden.
- Data Mining kann nahezu überall eingesetzt werden, wo es
  - ein Anwendungsproblem gibt, dessen Lösung den Aufwand des Data Mining rechtfertigt.
  - für das Anwendungsproblem relevante Daten vorhanden sind.

32

Heutiges Programm

- Analytical Data Science, Big Data or Data-/Text-Mining ⇔ How to use Machine Learning and Statistics Methods
- Data Mining, Wissensentdeckung in Datenbanken Was ist das?
- Wozu kann man Data Mining gebrauchen?
- Was braucht man zum Data Mining?
- Überblick über die Vorlesung

33

Programm der Vorlesung

	Vorlesung	Übung
19.02.20	Überblick	DME: Data Mining Environment
26.02.20	KDD-Prozess + Data Understanding	Business + Data Understanding
04.03.20	Preprocessing I	DME: Preprocessing
11.03.20	Preprocessing II	Data Preprocessing
18.03.20	Classification + Regression	DME: Predictive Modeling
25.03.20	Regression + Assoziation	DM - Classification
01.04.20	Clustering + Deviation	DME: Descriptive Modeling
08.04.20	Multi-Relationale Daten	DM - Regression
15.04.20	Osterferien	
22.04.20	Raum + Zeit	DME: Data Transformation
29.04.20	Text Mining	Model Application
06.05.20	Text Mining II	DME: Text Mining
13.05.20	Data Mining & Text Mining Projekt	
20.05.20	Recommender Systeme + IDA	Uni geschlossen (Auffahrt)
27.05.20	DM Anwendungen	Business Präsentation
03.06.20		
10.06.20	Mündliche Prüfungen	

34

Infopaket zur Vorlesung

- Wahlvorlesung Informatik mit 6 Punkten, benotete Prüfung, An und Abmeldedung wie üblich.
- Vorlesungsunterlagen: Folien in PDF, Übungsaufgaben und Literaturliste auf meinem Web-Server
  - http://www.kietz.ch/AADS/
- Bitte Übungsgruppen bilden und Email (juk@ifi.uzh.ch) mit Name, Email und Matrikel-Nr der Gruppenmitglieder schicken.
- Übung mit
  - RapidMiner https://rapidminer.com
  - R http://www.r-project.org

35

Literatur zur Vorlesung

- Jiawei Han, Micheline Kamber, Jian Pei: Data Mining: Concepts and Techniques, Morgan Kaufmann, 2011
- Maimon, Oded & Rokach, Lior. (2010). Data Mining and Knowledge Discovery Handbook, 2nd ed.
- Pyle, D.: Data Preparation for Data Mining, Morgan Kaufmann, 1999.
- Chapman, P.; Clinton, J.; Khabaza, T; Reinartz, T.; Wirth, R.: The CRISP-DM Process Model.

36

Literatur zur Vorlesung

Weiter Gesamtüberblicke:

- Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, Lukasz A. Kurgan: Data Mining: A Knowledge Discovery Approach, Springer Verlag, 2007.
- Hippner; Küsters; Meyer; Wilde (Eds.): Handbuch Data Mining im Marketing,Vieweg, 2001.
- Weiss, S.; Indurkkhya, N.: Predictive Data Mining - a practical guide, Morgan KaufmannPublishers Inc, 1998.

Einführung:

- **Fayyad, U.; Piatetetsky-Shapiro, G.; Smyth, P.: From Data Ming to Knowledge Discovery: An Overview, In: (Fayyad, etal 1996).**
- Brachman, R.; Anand, T. The Process of Knowledge Discovery in Databases: A Human-Centered Approach, In: (Fayyad, etal 1996).
- Fayyad, U.; Piatetetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R.: Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, 1996.

37

Literatur zur Vorlesung

Data Mining Methods:

- **Witten, I.; Frank, E.: Practical Machine Learning Tools and Techniques with Java implementations, Morgan Kaufmann, 2000.**
  Java-sourcen at http://www.cs.waikato.ac.nz/ml/weka
- K. Morik; S. Wrobel; T. Joachims: "Maschinelles Lernen und Data Mining" In: G. Görz, J. Schneeberger und C.-R. Rollinger (Hrsg.), »Handbuch KI«, Oldenbourg Verlag, 2000.
- Wrobel, S.:An algorithm for multi-relational discovery of subgroups, In: Komorowski, J.; Zytkow, J.: Principles of Data Mining and Knowledge Discovery: First European Symposium (PKDD'97), Springer Verlag, 1997.
- Dzeroski, S.: Inductive Logic Programming and Knwoledge Discovery in Databeses, In: (Fayyad, etal 1996).

38

Literatur zur Vorlesung

Preprocessing:

- **Pyle, D.: Data Preparation for Data Mining, Morgan Kaufmann, 1999.**
- Liu, H.; Motoda, H.: Feature Selection for Knowledge Discovery and Databases, Kluwer, 1998.

Anwendungen:

- Berry, M.; Linoff, G.: Mastering Data Mining: The Art and Science of Customer Relationship Management, John Wiley, 2000.

Data Mining Projekt Handbuch:

- Chapman, P.; Clinton, J.; Khabaza, T; Reinartz, T.; Wirth, R.: The CRISP-DM Process Model, http://www.crisp-dm.org/