PROSEMINAR: INFORMATIONSGEWINN DURCH EXPERIMENTE WS 09/10

DATA MINING ALS EXPERIMENT

VORTRAG: CHRISTOPH NÖLLENHEIDT

26.01.10

Ablauf

- Das CRISP-DM-Modell
- Zwei verschiedene Standpunkte über die Theoriebildung im Prozess
- Ein verändertes CRISP-DM-Modell
- Die Idee des „anything goes“ für das Data Mining
- Wie sieht das Experiment in RapidMiner aus?
- Literatur

CRISP-DM
(„Cross Industry Standard Process for Data Mining“)

- Inhaltliche Ausprägung vollständig aus Projekterfahrungen zum Data Mining abgeleitet

- Beteiligt an der Entwicklung seit 1996:

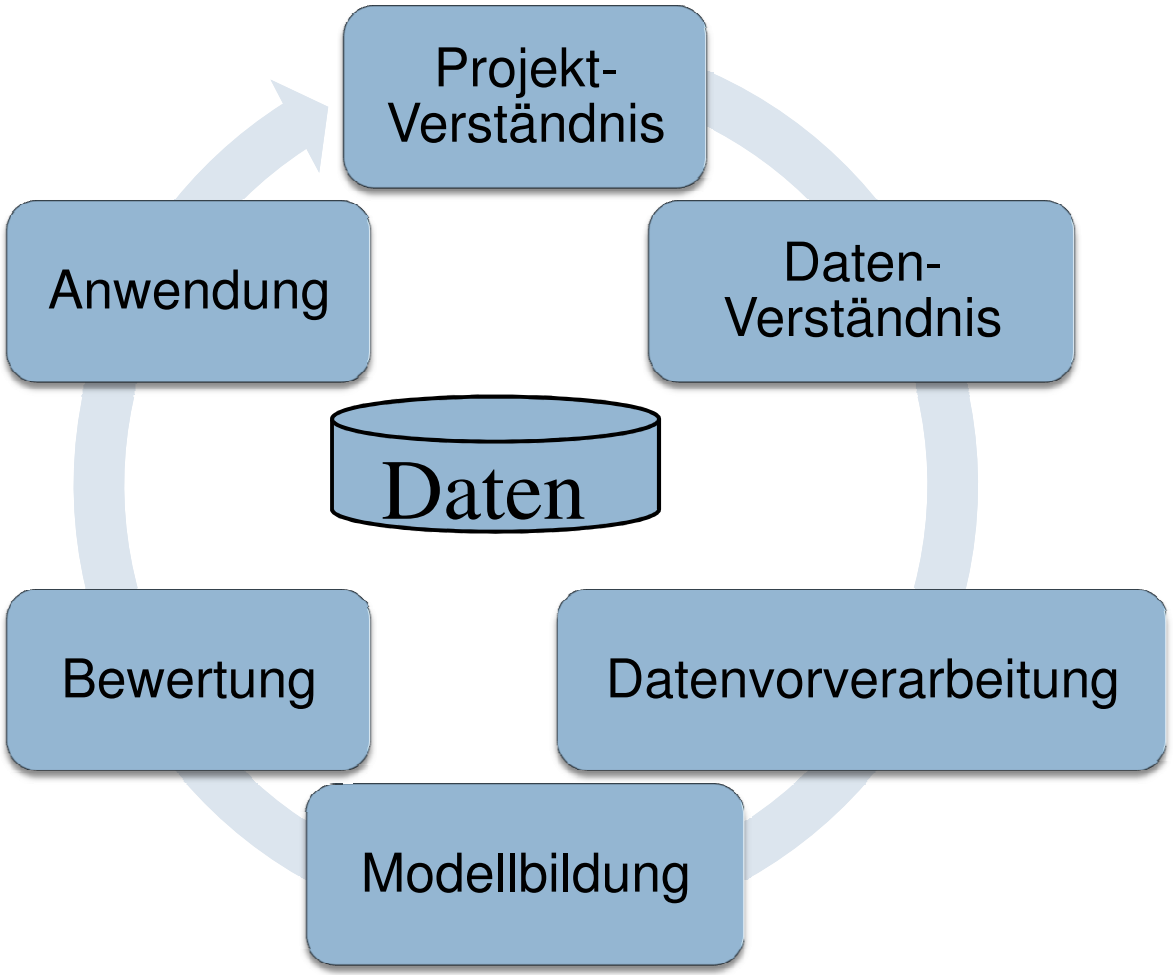
- DaimlerChrysler (damals Daimler-Benz)
- Integral Solutions Ltd. (seit 1999 Teil von SPSS)
- NCR („National Cash Register“)

- Ziel: Etablierung eines Standards zur Vorgehensweise bei Data Mining Projekten

CRISP-DM-Modell

KDD

Data Mining



Die Phasen des CRISP-DM

1. Projekt-Verständnis

- Ziel: Eckpunkte eines Data Mining Projektes festlegen

Teilschritte:

- Situationsanalyse und Aufgabendefinition
- Formulierung der Data Mining Ziele
- Erstellung des Projektplans

Die Phasen des CRISP-DM

2. Verständnis der Daten erreichen

- Datenbeschaffung
- Erste Einblicke in die Daten
- Probleme mit der Qualität der Daten ausfindig machen
- Hinzunahme externer Daten?

Die Phasen des CRISP-DM

3. Datenvorverarbeitung

- schließt alle Aktivitäten ein, um den für die Modellbildung verwendeten Datensatz aus den Rohdaten zu erhalten
- für den Projekterfolg von großer Bedeutung
- enthält bereits Verfahren des maschinellen Lernens

Die Phasen des CRISP-DM

3. Datenvorverarbeitung

- Aufgaben sind z.B. Merkmalsauswahl, Transformation als auch Datenbereinigung
- Fehlerkorrektur: Fehlende Werte ersetzen
- Wie gehe ich mit Ausreißern um?
- Stichproben ziehen

Die Phasen des CRISP-DM

4. Modellbildung

- Anwendung der Data Mining Verfahren
- Ziel: hohe Aussage- bzw. Interpretationsfähigkeit des Analyseergebnisses
- Einige Verfahren haben spezielle Anforderungen an die Form der Daten, zur Vorverarbeitung zurückzukehren ist daher oftmals nötig

Die Phasen des CRISP-DM

5. Bewertung

- Bevor das gefundene Modell zur abschließenden Analyse verwendet wird, muss es bewertet werden
- Einzelne Schritte, die ausgeführt wurden, um das Modell zu bilden, werden überprüft
- Wurden die Data Mining Ziele erreicht?

Die Phasen des CRISP-DM

6. Anwendung

- Wissen für den Kunden darstellen

- Phase kann enthalten:

- Bericht

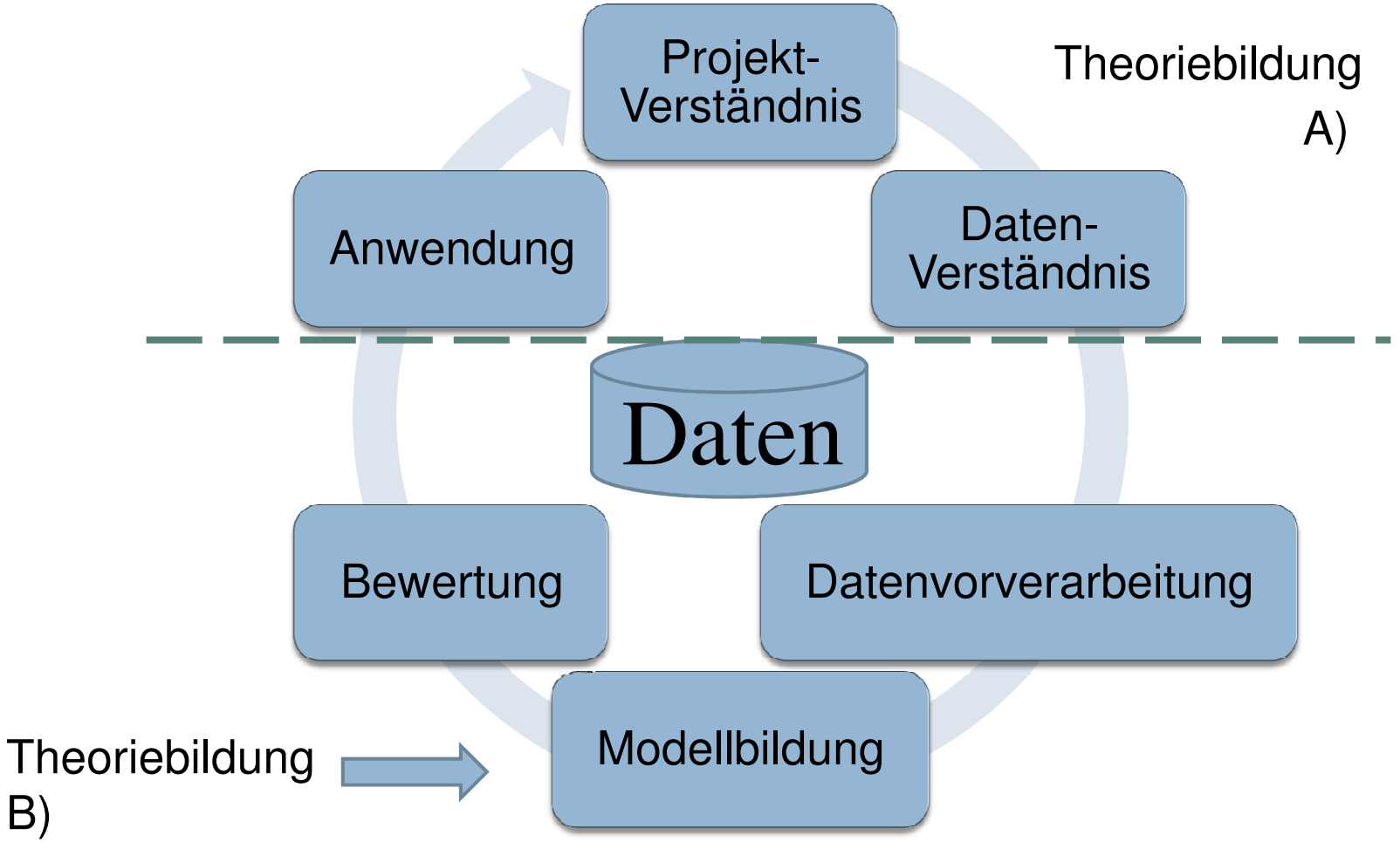
- wiederholbaren Data Mining Prozess implementieren

- Wichtig für den Kunden: Welche Aktionen müssen ausgeführt werden, um das Modell zu verwenden?

Experiment

- Experiment setzt zwingend Theorie voraus
- Messung macht ohne Theorie keinen Sinn (Kuhn)
- Bateson: „Theorien können sich im Laufe der Untersuchung ändern“
- 2 verschiedene Sichtweisen über die Bildung einer Theorie im Data Mining Prozess

Theoriebildung im Data Mining Prozess



CRISP-DM

- Im CRISP-DM gibt es keinen datengenerierenden Prozess

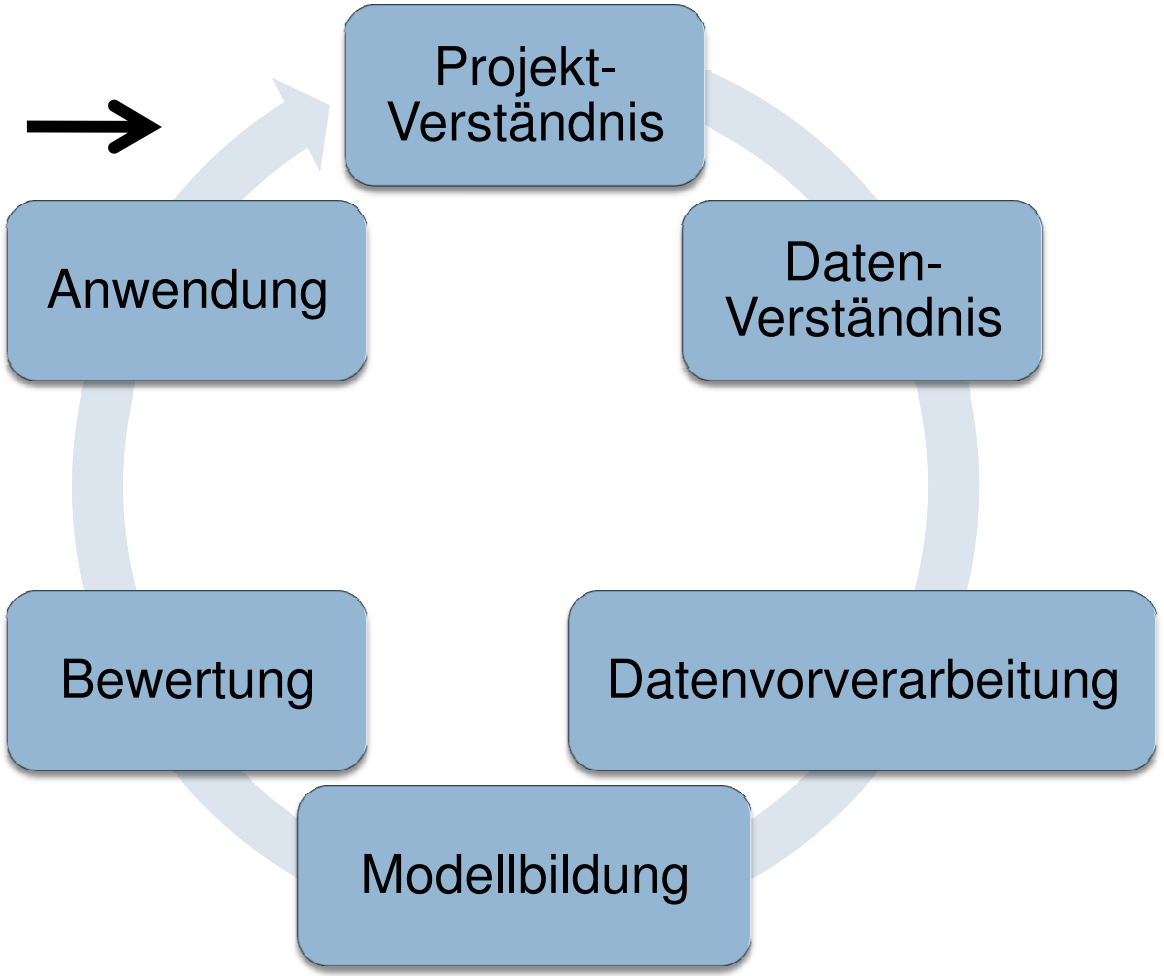
- Informatik: „Ihr gebt uns die Daten, ob wir damit die Realität abbilden, wissen wir nicht“

⇒ Erkenntnisse aufgrund der Daten

- Statistik: betont datengenerierenden Prozess

Was macht Wissenschaft aus?

Datengenerierende Prozess



Paul Feyerabend

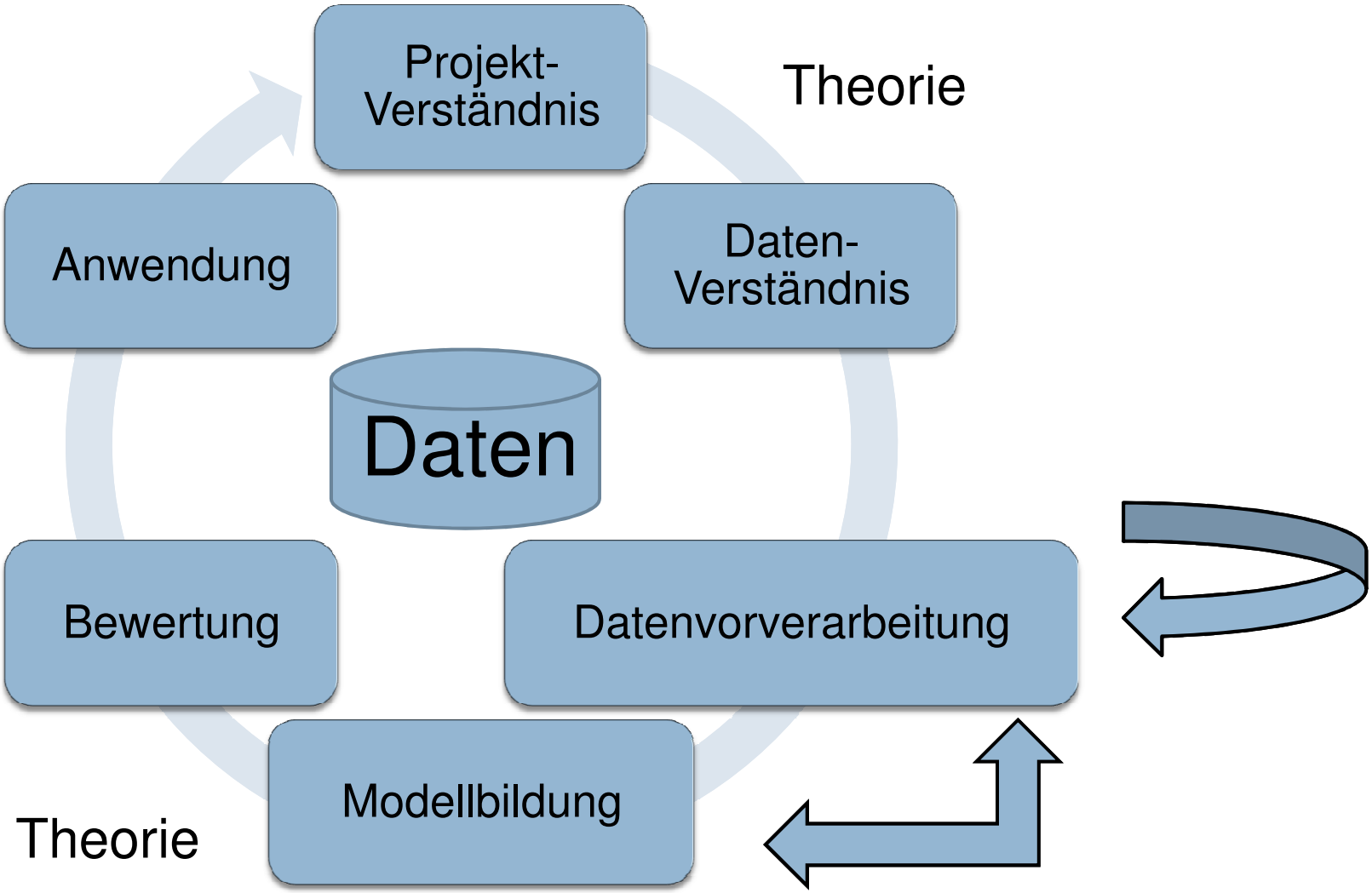
- „anything goes“
- Wir müssen nicht ab einem gewissen Punkt anfangen und an einem gewissen Punkt ankommen
- Wissenschaft sollte nicht „durch Zwänge eingeschränkt werden“

Idee des „anything goes“

Bedeutung im Data Mining:

- Kreativität
- Vielleicht erhalten wir durch neue Betrachtungen bessere Ergebnisse

Idee des „anything goes“



RapidMiner (vormals YALE)

- Open-Source Umgebung für Data Mining
- Rapid-I (http://www.rapid-i.com)
- Mittlerweile meist genutzte open-source Umgebung für Data Mining
- Anwendungen in Forschung und Praxis

RapidMiner

- Wichtigste Eigenschaft: Prozesse des Data Mining als Operator-Bäume modellieren
- Operator in RapidMiner bedeutet: einzelne Methode, z.B. Methode des maschinellen Lernens
- Experimente können auf der Grundlage einer großen Anzahl von beliebig ineinander verschachtelbaren Operatoren durchgeführt werden

Literatur

- Mierswa, I. et al., *YALE: Rapid Prototyping for Complex Data Mining Tasks*, In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), 2006.
- Gabriel, R. et al.: *Data Warehouse & Data Mining* (2009), W3L-Verlag, Herdecke, Witten.
- Für das CRISP-DM-Modell:

http://www.crisp-dm.org