Arbeitsbereich NATS

Prof. Menzel

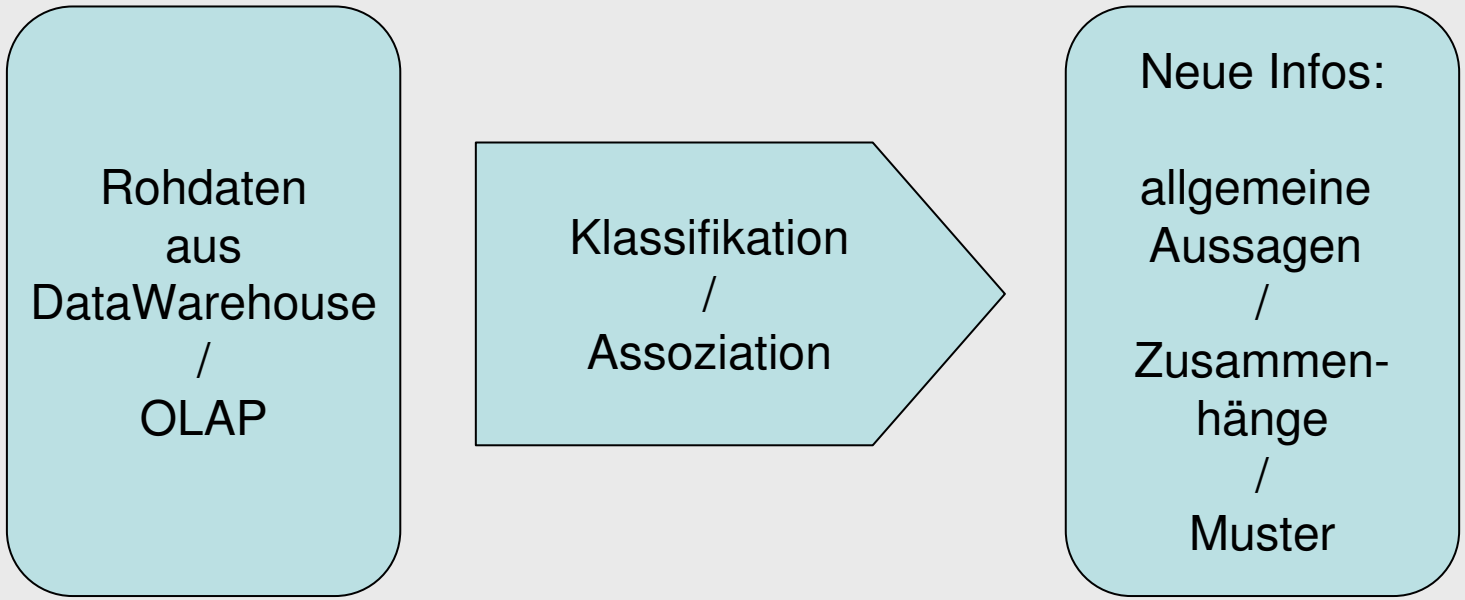
Seminar Data Mining

Datenschutzgerechtes Data Mining

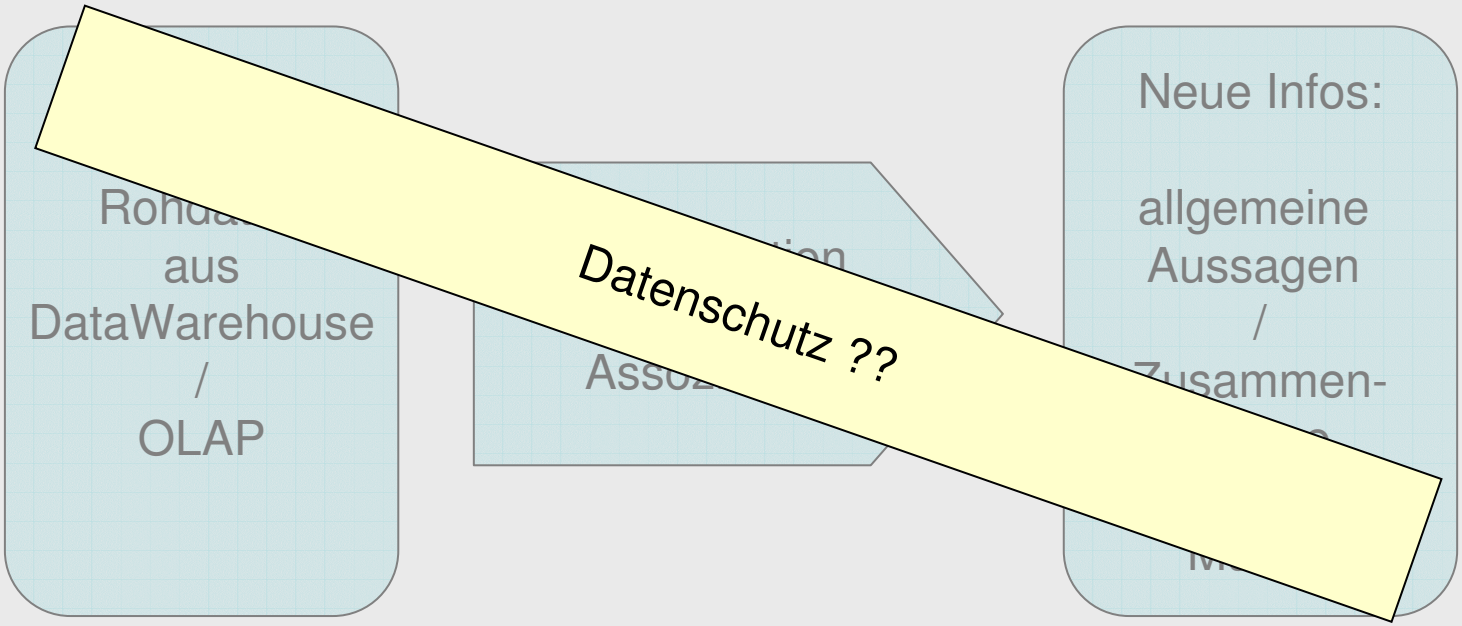
Seminarvortrag von Simon Boese

Student der Wirtschaftsinformatik

Wissensgewinnung



Wissensgewinnung



Agenda

- Ziele des Data Mining und Problemstellung
- Privatsphäre
- Was muss sich ändern?
- Vorgehensweisen
- Zusammenfassung

Was soll erreicht werden?

↳ Wissensgenerierung

- Individuelle Informationen bilden die Basis
  - Kundenverhaltensdaten etc.
- Klassifikation, Regression, Assoziation, Clustering
- Optimierung der
  - Vertriebsstruktur
    - Produktplatzierung
    - gezielte Werbung
  - SCM
    - Produktionsverfahren
    - Lieferantenbeziehungen

…was noch?

- Überwachung
- Sicherheit
- Prävention

Wer darf
was
und
wozu
?

EU-Richtlinie

(95/46/EG)

Richtlinie über „die Verarbeitung personenbezogener Daten und den Schutz der Privatsphäre in der elektronischen Kommunikation“

- Jede Person hat das Recht, sich keiner (teil-) automatisierten Entscheidung zu unterwerfen
  - Entscheidungen aufgrund Data Mining Resultate
- Ausnahmen für Strafverfolgung und Geheimdienste

Data Mining Reporting Act[1]

US-Senat 2003

Definition des Data Mining

Eine Anfrage, Suche oder Analyse auf einer oder mehrerer Datenbanken, wobei

- mind. eine DB von nicht-staatlichen Stellen erzeugt/erhoben oder kontrolliert wurde/wird ODER die ursprüngliche Erhebung durch staatliche Stelle keinen geheimdienstlichen oder strafverfolgungsrelevanten Hintergründe hatte.

Data Mining Reporting Act$_2$

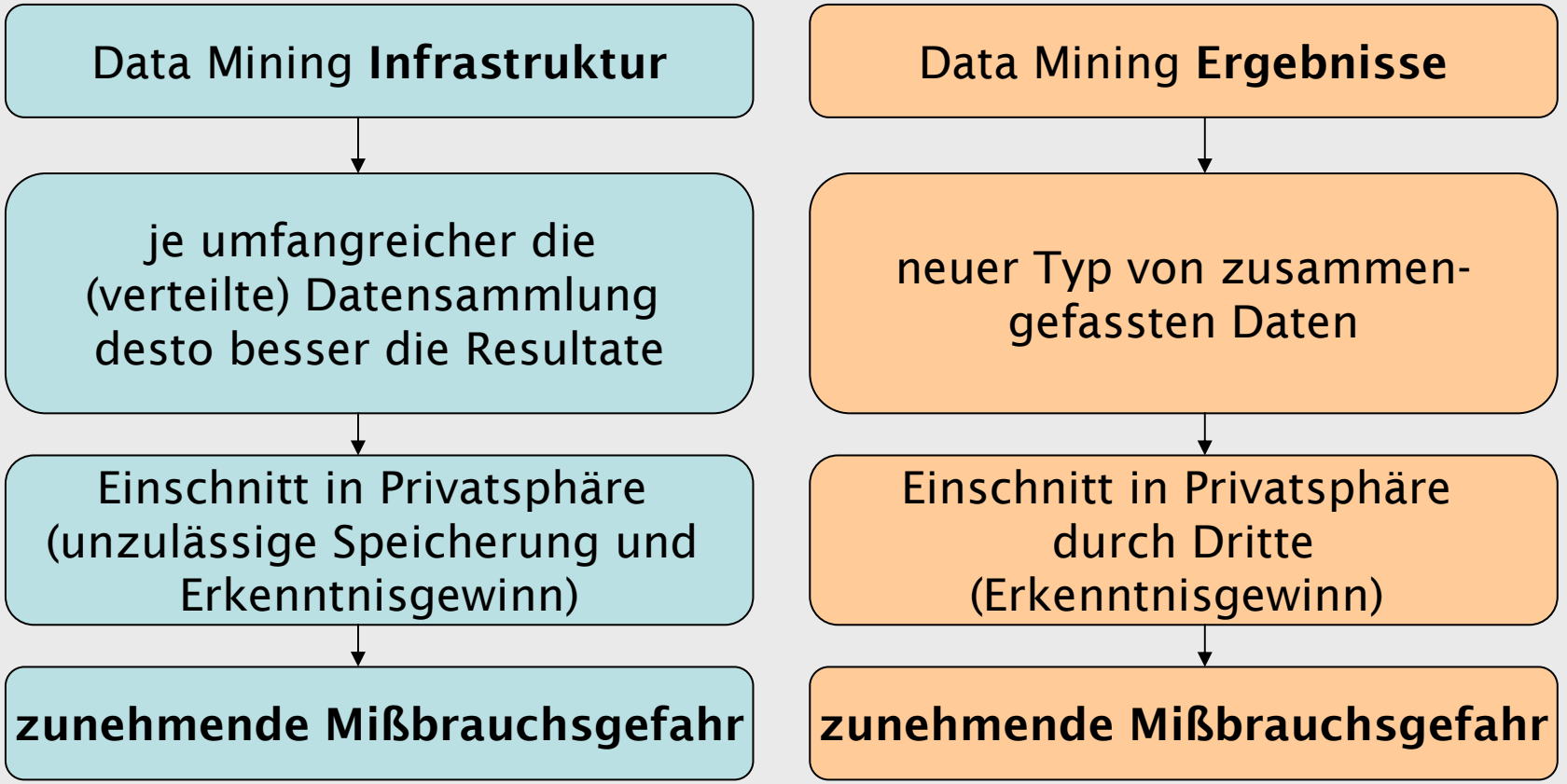
US-Senat 2003

Definition des Data Mining

Eine Anfrage, Suche oder Analyse auf einer oder mehrerer Datenbanken, wobei

- bei der Suche keine personenbezogenen Daten verwendet werden dürfen, um nach Informationen über diese bestimmte Person zu suchen,
- eine staatl. Behörde auch nach Mustern, die auf terroristische oder andere kriminelle Aktivitäten hinweisen, suchen darf.

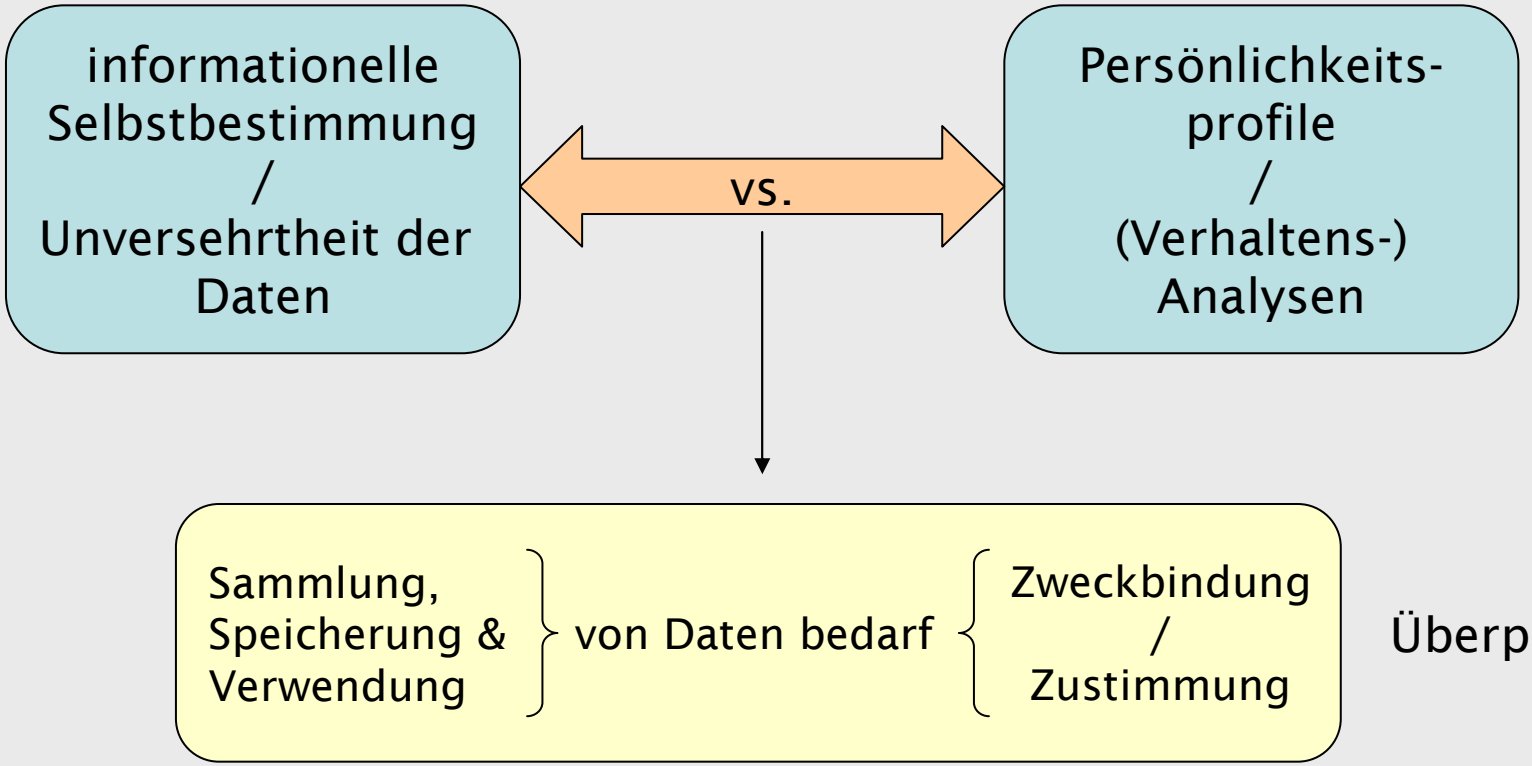
Problemstellung



Agenda

- Ziele des Data Mining und Problemstellung
- Privatsphäre
- Was muss sich ändern?
- Vorgehensweisen
- Zusammenfassung

Privatsphäre$_1$



Privatsphäre$_2$

Wann ist die Privatsphäre verletzt?

- Solange keine konkrete Person identifiziert werden kann, stellen Data Mining Ergebnisse keinen Eingriff dar.
- Jede Wissenserweiterung (über die Basisdaten hinausgehend) ist bereits eine Störung der Privatsphäre.

Privatsphäre$_3$

Wissenserweiterung ist jedoch oberstes Ziel!

- Der Erkenntnisgewinn muss bewertet werden und den Möglichkeiten, aus den Daten Rückschlüsse auf Individuen zu ziehen, gegenüber gestellt werden:
  - Messung der Schwere des Eingriffs
  - Definition von Akzeptanzwerten

Erhebungsprozess

Die Privatsphäre muss während des gesamten Prozesses geschützt bleiben:

- Datenerhebung (Anonymisierung und Pseudonymisierung)
- Teilnehmende Parteien (Datensicherheit)
- Veröffentlichung/Nutzung

Agenda

- Ziele des Data Mining und Problemstellung
- Privatsphäre
- Was muss sich ändern?
- Vorgehensweisen
- Zusammenfassung

Anforderungen$_1$

Datenschutzgerechtes Data Mining muss zusichern, dass die offen gelegten Daten…

- nicht zu einem Individuum zurückverfolgt werden können
- keine Verletzung der Privatsphäre darstellen

Anforderungen$_2$

Datenschutzgerechtes Data Mining muss zusichern, dass die offen gelegten Daten…

- das Risiko einer Identifikation minimieren
- auch nicht in Kombination mit anderen verfügbaren Daten zu einer Verletzung der Privatsphäre führen

Agenda

- Ziele des Data Mining und Problemstellung
- Privatsphäre
- Was muss sich ändern?
- Vorgehensweisen
- Zusammenfassung

k-Anonymität$_1$

- Grundgedanke: Bildung einer Gruppe von k Datensätzen
  - Maximal: Identifikation einer Gruppe

- Quasi-Identifikatoren (QI): alle Informationen, die in Kombination mit regulär zugänglichen Daten die Identifikation eines Einzelnen ermöglichen

k-Anonymität$_2$

Definition:

Kein Datensatz darf in seinen QI einzigartig sein. Es müssen immer mind. k Datensätze die selben QI aufweisen

k-Anonymität$_3$

- Um die QI bereinigten Daten, machen die Identifikation eines Individuums unmöglich
- Unsicherheit bei Informationen über eine konkrete Person
- Rückschlüsse nur auf eine Gruppe

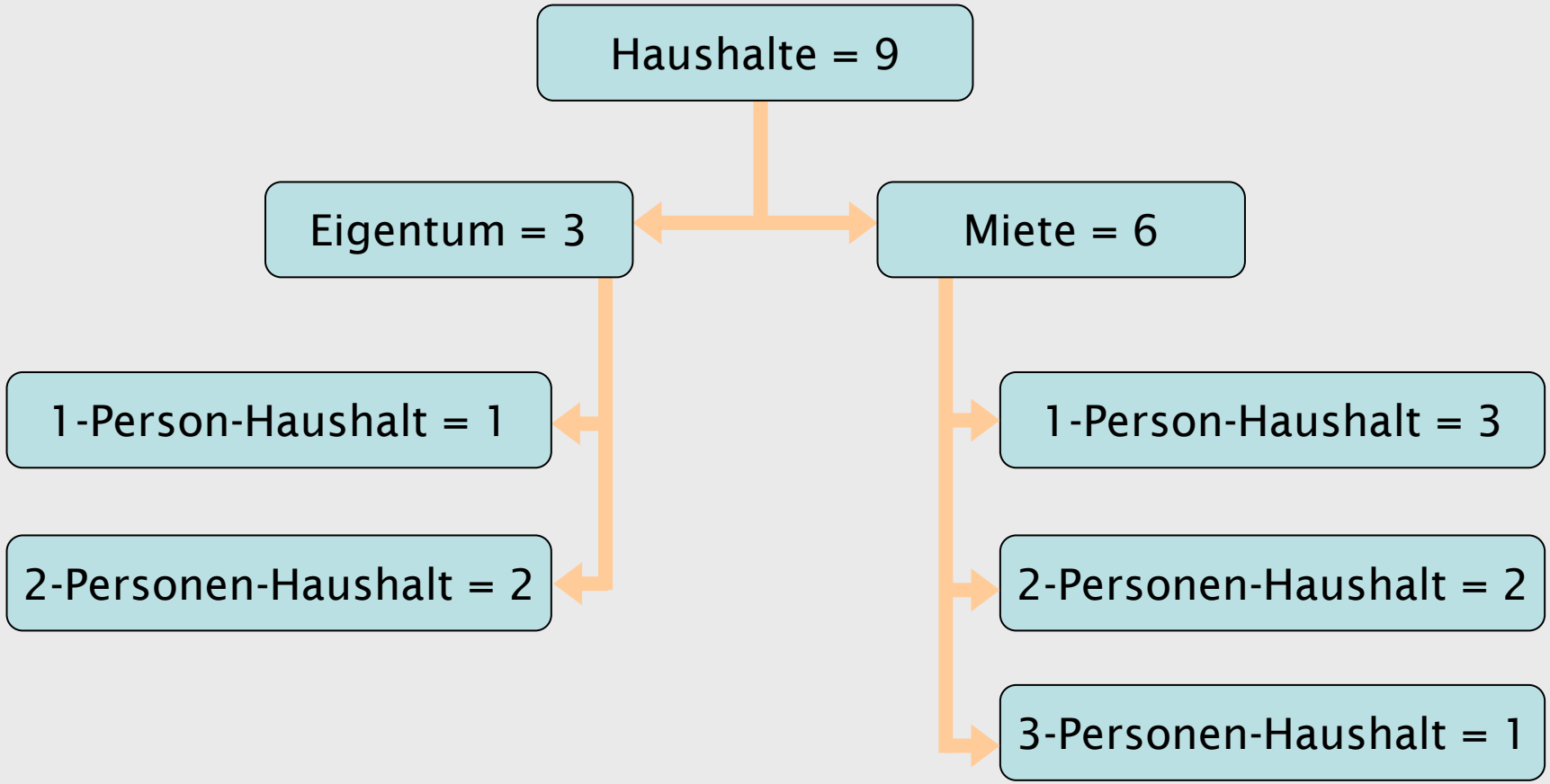
Verschleierung

- Anhand einer bestimmten Verteilung die Originaldaten verzerren.
- Statistisch ungenau machen.
- Verwendbarkeit der Ergebnisse?
  - Verzerrung mittelt sich möglicherweise heraus.
  - Rückrechnung schwierig, falls überhaupt die Methode der Verschleierung bekannt ist.

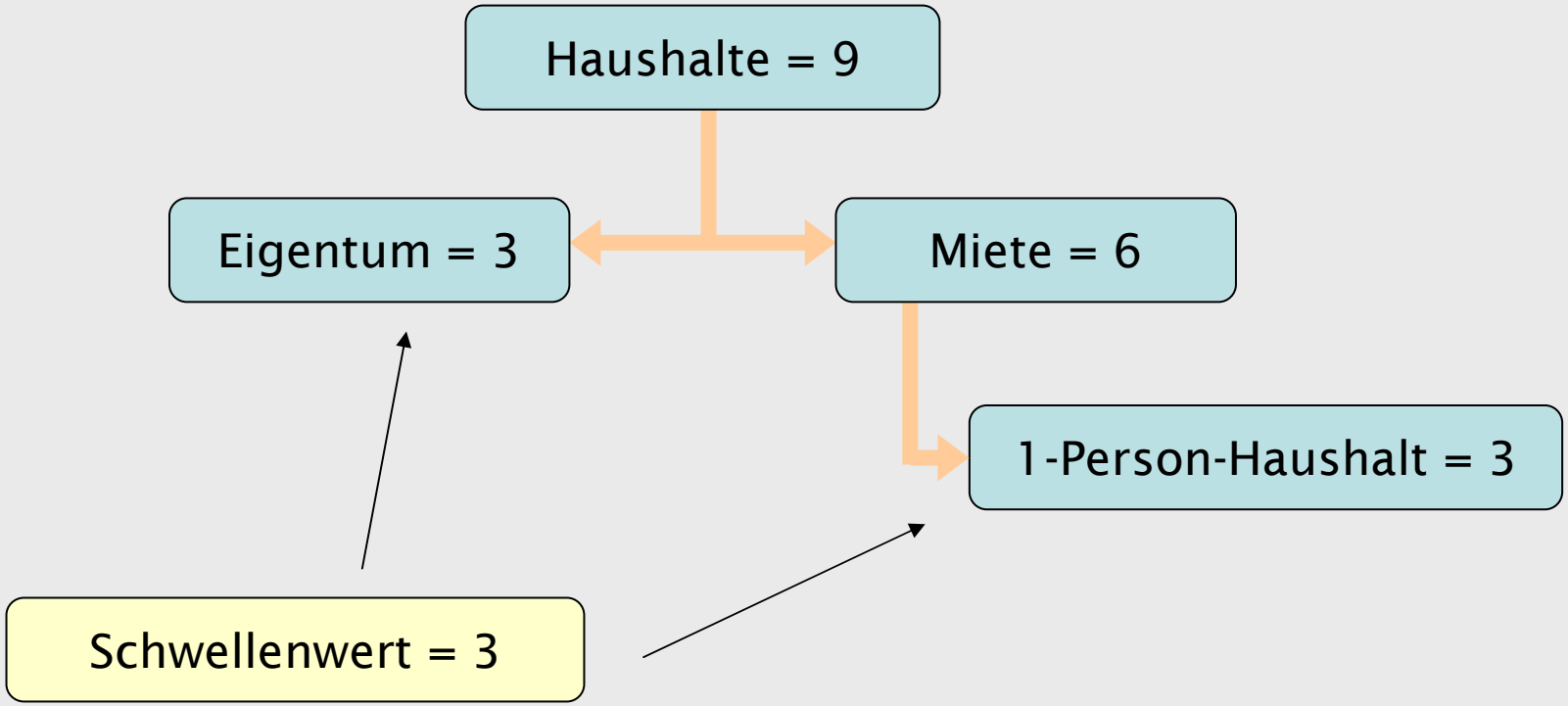
Schwellenwerte und Generalisierung$_1$

- Schwellenwerte:
  - Kriterien erst ab gewisser Anzahl an Treffern veröffentlichen
- Generalisierung:
  - weniger Details, Zusammenfassung
- „beschneiden“ des Entscheidungsbaumes
- können auch bei der k-Anonymität verwendet werden

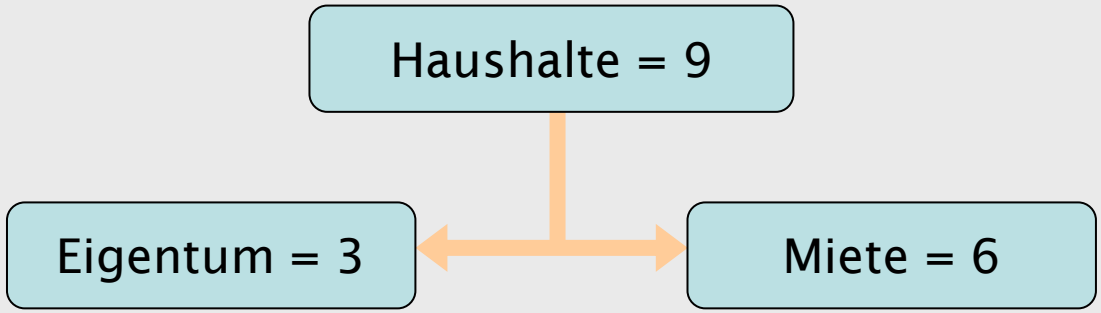
Schwellenwerte und Generalisierung$_2$ – Bsp.



Beispiel: Schwellenwerte



Beispiel: Generalisierung



weniger Detail

Sicherheit

- Bisheriger Fokus: Data Mining Ergebnisse
  - Verletzung der Privatsphäre einzelner Individuen bei Veröffentlichung
- Neuer Fokus: Data Mining Prozesse
  - „collaboration meets competition“
    - Datensicherheit
    - Kryptographie

Verteilung der Daten$_1$

- Wann muss überhaupt datenschutzgerechtes Data Mining eingesetzt werden?
  - Wenn Daten zentralisiert bzw. verteilt vorliegen, aber nur von einer Stelle kontrolliert, besteht keine zusätzliche Gefahr einer Verletzung der Privatsphäre.
    - ↳ anderer Ansatz

Verteilung der Daten$_2$

- Wann muss überhaupt datenschutzgerechtes Data Mining eingesetzt werden?
  - Liegen Daten verteilt vor und werden von unterschiedlichen Stellen kontrolliert, könnten durch Kooperation dieser Stellen sensible Daten offen gelegt werden.

Verteilung der Daten$_3$

An k verschiedenen Orten P werden in einer Datensammlung D bestehend aus den Spaltenbezeichnungen I und den Entitäten E Informationen gespeichert:

$$D_k = (E_k, I_k)$$

- Horizontal/Homogen verteilte Daten
- Vertikal/Heterogen verteilte Daten

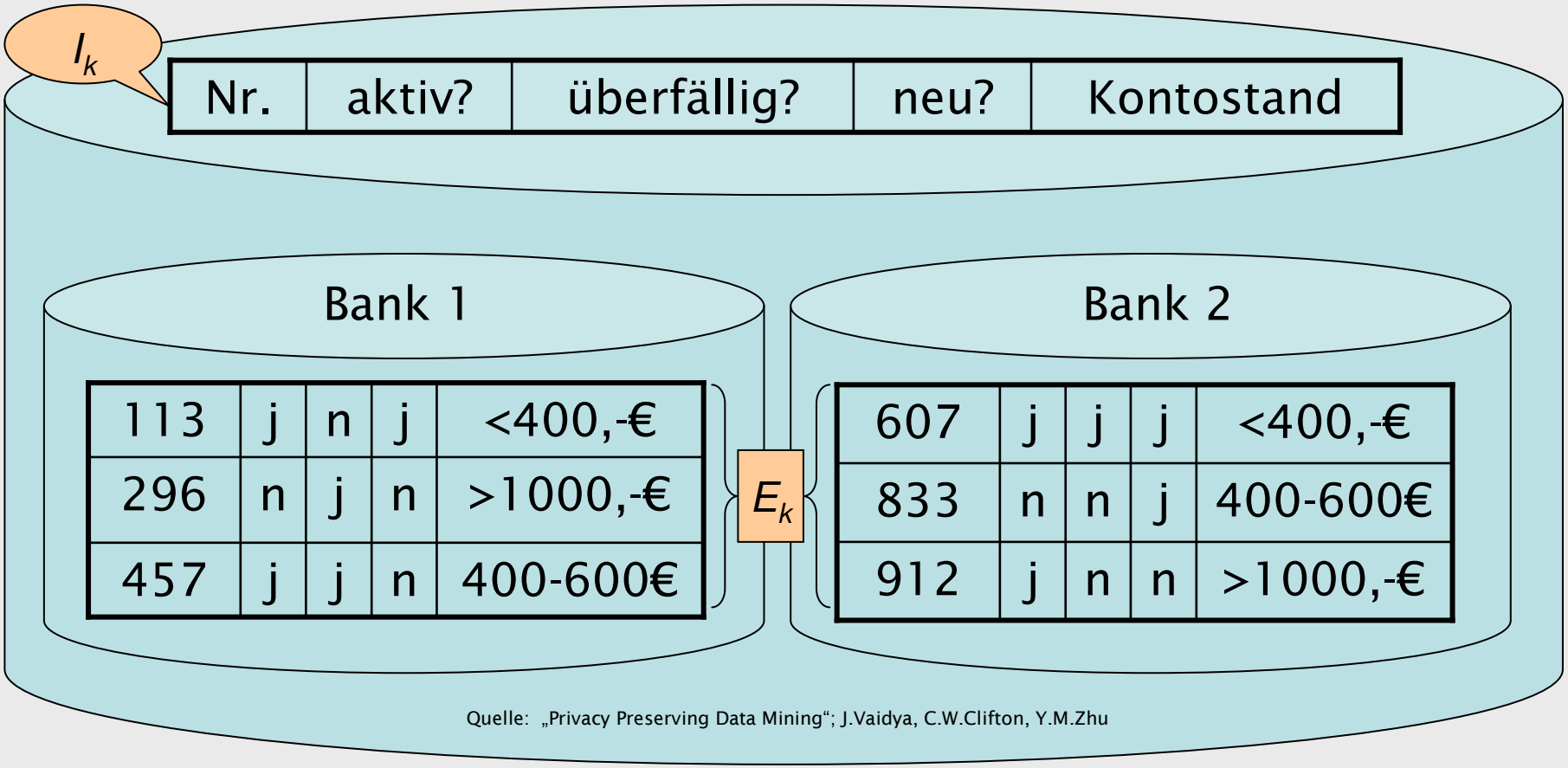
Horizontale Datenverteilung$_1$

- an verschiedenen Orten sind gleiche Daten über verschiedene Objekte vorhanden
- Datenmenge erhöhen, statistische Genauigkeit steigt

$$E_G = \bigcup_i E_i = E_1 \cup \ldots \cup E_k$$

$$I_G = \bigcap_i I_i = I_1 \cap \ldots \cap I_k$$

Horizontale Datenverteilung$_2$



Quelle: „Privacy Preserving Data Mining“; J.Vaidya, C.W.Clifton, Y.M.Zhu

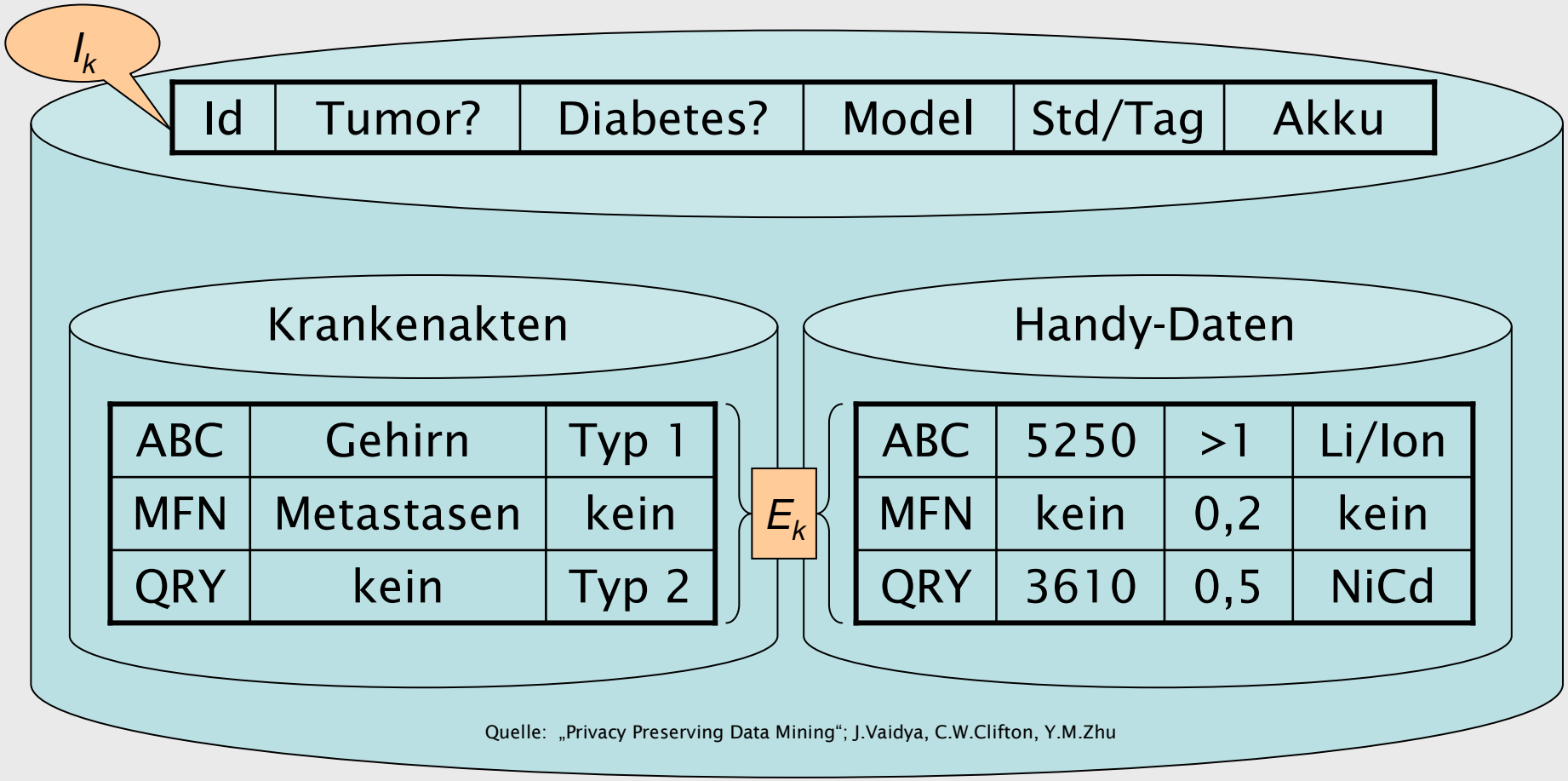
Vertikale Datenverteilung$_1$

- an verschiedenen Orten liegen unterschiedliche Daten über gleiche Objekte
- Datenumfang erhöhen, globale Zusammenhänge erkennen

$$E_G = \bigcap_i E_i = E_1 \cap \ldots \cap E_k$$

$$I_G = \bigcup_i I_i = I_1 \cup \ldots \cup I_k$$

Vertikale Datenverteilung$_2$



Quelle: „Privacy Preserving Data Mining“; J.Vaidya, C.W.Clifton, Y.M.Zhu

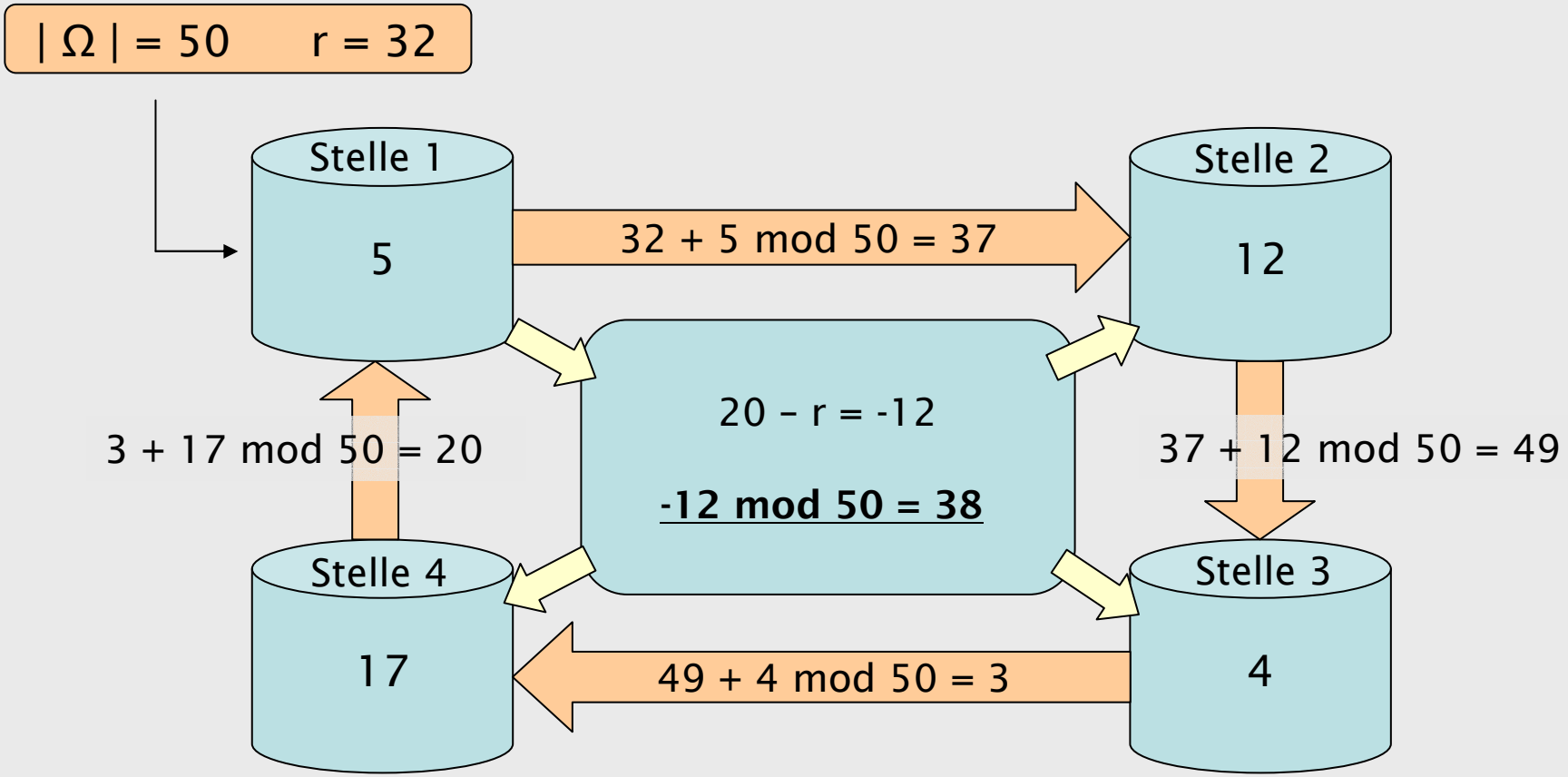
Secure Multi-party Computation (SMC)

- Generelles Problem bei der Berechnung von Funktionen, wenn die Input-Daten verteilt – vertikal/horizontal – vorliegen
- Kommunikation via Protokoll
- Ziel: Außer dem Ergebnis erfährt keine Partei etwas Neues
- Erfolgt eine Berechnung sicher, dann war sie auch privat! (Goldreich et al. basierend auf Yaos Millionärs Protokoll)

SMC – Beispiel$_1$

- Berechnung einer Summe
- Jede (beteiligte) Partei hält einen Wert, der nicht bekannt gegeben werden soll
- Annahme: der Wertebereich Ω der zu berechnenden Summe ist bekannt
- gleichverteilte Zufallszahl $r \in \Omega$ als Verschleierung

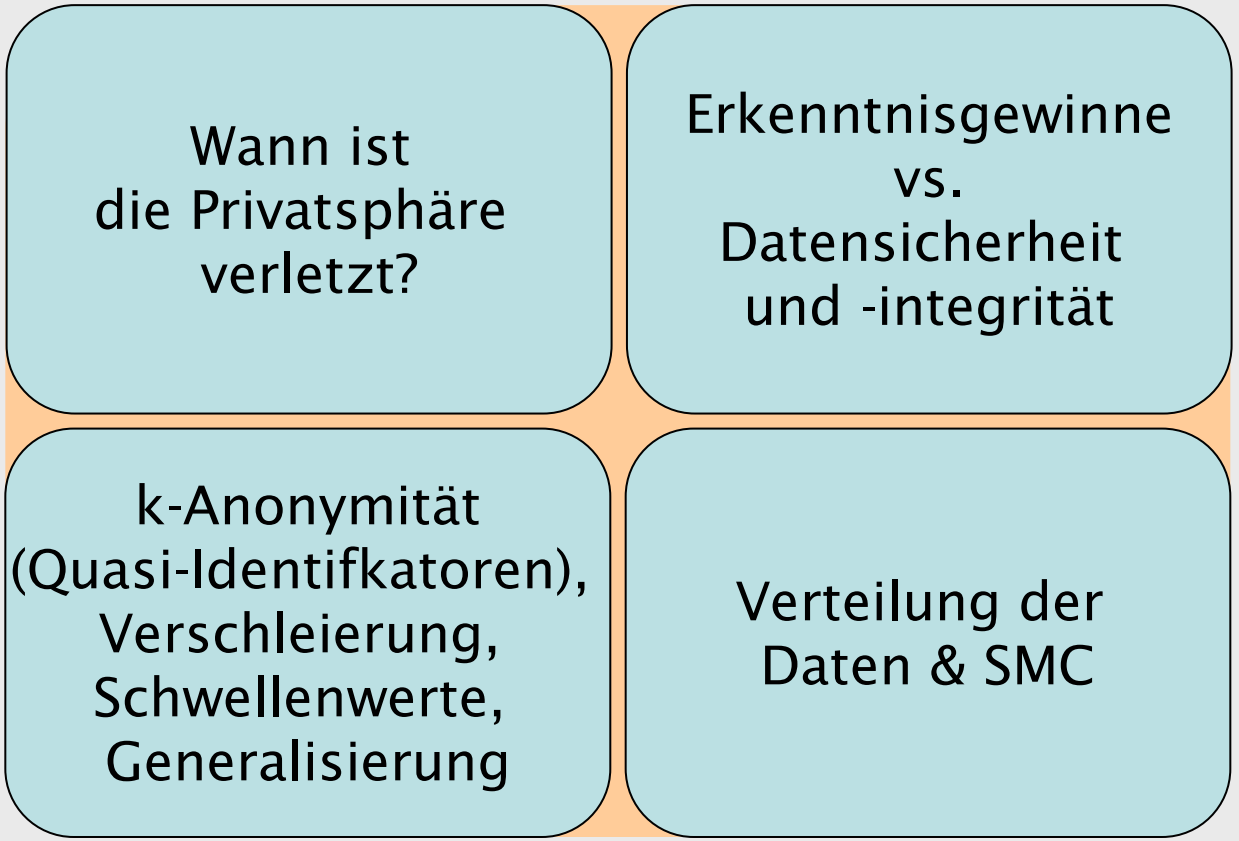
SMC – Beispiel$_2$



Quelle: „Privacy Preserving Data Mining"; J.Vaidya, C.W.Clifton, Y.M.Zhu

Agenda

- Ziele des Data Mining und Problemstellung
- Privatsphäre
- Was muss sich ändern?
- Vorgehensweisen
- Zusammenfassung


Wann ist die Privatsphäre verletzt?
Erkenntnisgewinne vs. Datensicherheit und -integrität
k-Anonymität (Quasi-Identifkatoren), Verschleierung, Schwellenwerte, Generalisierung
Verteilung der Daten & SMC

Literatur

- J. Vaidya, Ch.W. Clifton, Y.M. Zhu; **„Privacy Preserving Data Mining“**; Springer Science+Business Media; New York 2006

Ende

Fragen , Anregungen und/oder Kritik?

Vielen Dank für Ihre/Eure Aufmerksamkeit.