Vorlesung & Seminar Informationsintegration

Einführung und Themenvorstellung

Fabian Panse

panse@informatik.uni-hamburg.de

7. April 2017

Informationsintegration

Begrifflichkeit:

Der Begriff Informationsintegration beschreibt die logische Zusammenführung gleichartiger Informationen unterschiedlicher Herkunft.

Ziele:

- Jederzeit verfügbarer, effizienter Zugriff auf mehrere, heterogene Datenquellen.
- Vereinigung von Daten aus verschiedenen Systemen zu einen einheitlichen Datenbestand.
  - **virtuell** (z.B. fördertiertes Datenbanksystem)
    Integration zum Zeitpunkt der Anfrage.
  - **materiell** (z.B. Data Warehouse)
    Einmalige Integration, danach Anfragen auf physischen Datenbestand.

Probleme

Heterogenität:

- Uneinheitliche Modellierung von Informationen
  (z.B. verschiedene Datenformate, Schemata, Datenmodelle)

Autonomie:

- Eingeschränkter Zugriff auf Daten
- Fehlende Kontrolle (z.B. kein Einfluss auf Schemaänderungen)

Qualität:

- Fehlerhafte und unvollständige Daten

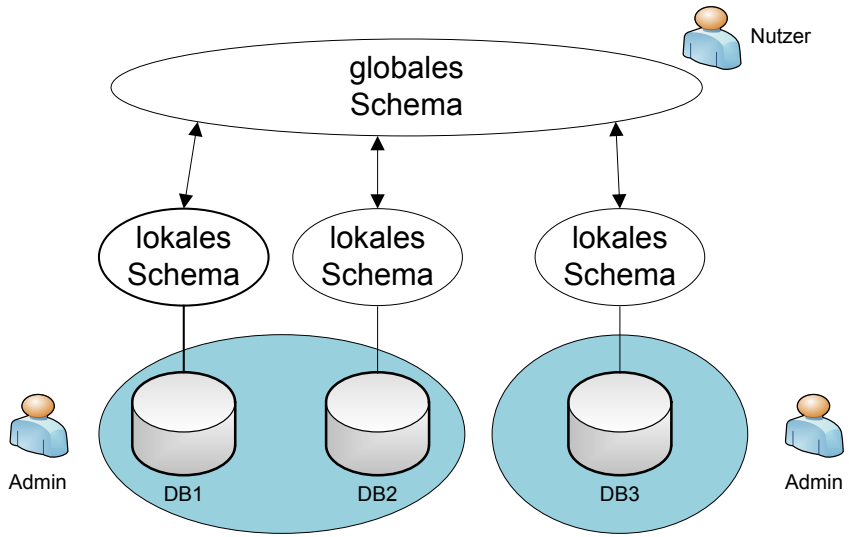
Komplexität:

- Automatische Integration erschwert
- Oft statisch integrierte Systeme

Integrationsablauf:

Schema Integration:

- Aufbau eines integrierten globalen Schemas
- Identifizierung von Korrespondenzen zwischen Elementen der lokalen Schemata und dem globalen Schema
- Generierung von Regeln zur Abbildung der Daten von den lokalen Schemata auf das globale Schema (z.B. durch SQL-Anfragen)

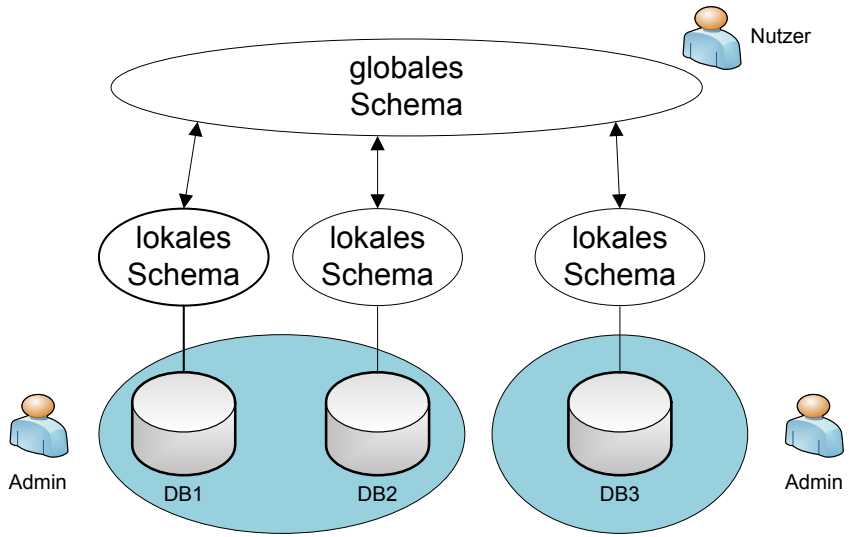


Integrationsablauf:

Anfragebearbeitung:

Koordinierung und Ausführung der globalen Anfrage

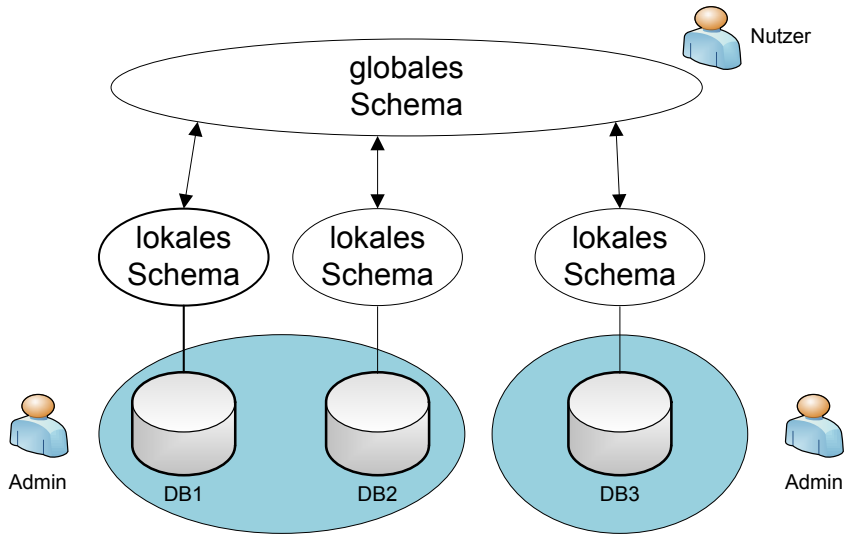
- Anfrageplanung (Local-As-View, Global-As-View)
- Anfrageübersetzung vom globalen auf die lokalen Schemata
- Anfrageoptimierung (lokal, global)



Integrationsablauf:

Daten Integration:

- Identifizierung von Datenobjekten, welche die gleiche Realwelt Entität darstellen (Duplikaterkennung)
- Zusammenführung mehrerer Duplikate zu einem einzelnen Datenobjekt (Datenfusion)



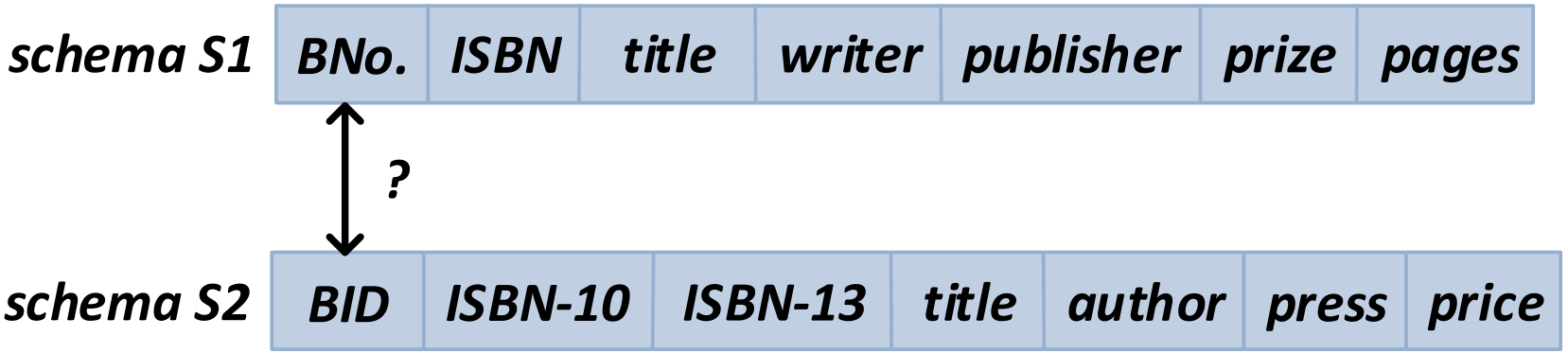
Finden von Korrespondenzen durch Schema Matching

schema S1	BNo.	ISBN	title	writer	publisher	prize	pages

schema S2	BID	ISBN-10	ISBN-13	title	author	press	price

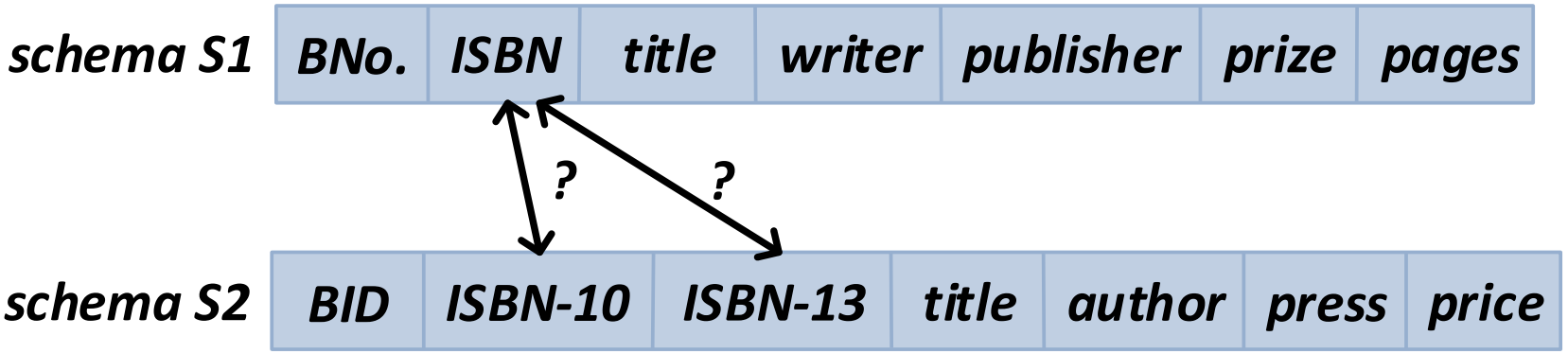
- Labelbasiert (z.B. linguistische Ähnlichkeit der Attributnamen)
- Instanzbasiert (z.B. Vergleich von Datenwerten)
- Strukturbasiert (z.B. Nachbarschaftsbeziehung, Integritätsbedingungen)

Finden von Korrespondenzen durch Schema Matching



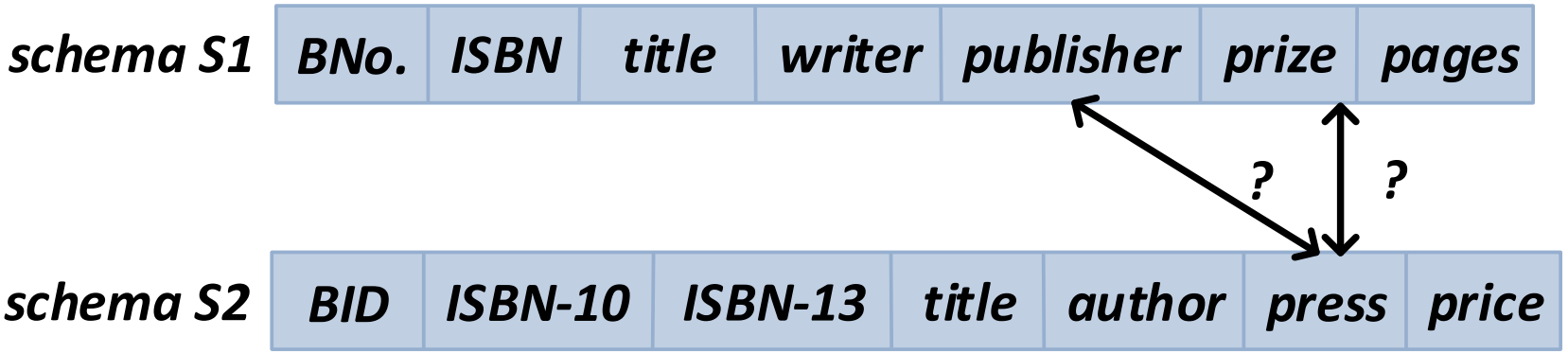
- Labelbasiert (z.B. linguistische Ähnlichkeit der Attributnamen)
- Instanzbasiert (z.B. Vergleich von Datenwerten)
- Strukturbasiert (z.B. Nachbarschaftsbeziehung, Integritätsbedingungen)

Finden von Korrespondenzen durch Schema Matching



- Labelbasiert (z.B. linguistische Ähnlichkeit der Attributnamen)
- Instanzbasiert (z.B. Vergleich von Datenwerten)
- Strukturbasiert (z.B. Nachbarschaftsbeziehung, Integritätsbedingungen)

Finden von Korrespondenzen durch Schema Matching

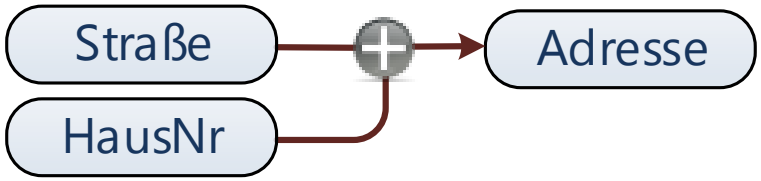


- Labelbasiert (z.B. linguistische Ähnlichkeit der Attributnamen)
- Instanzbasiert (z.B. Vergleich von Datenwerten)
- Strukturbasiert (z.B. Nachbarschaftsbeziehung, Integritätsbedingungen)

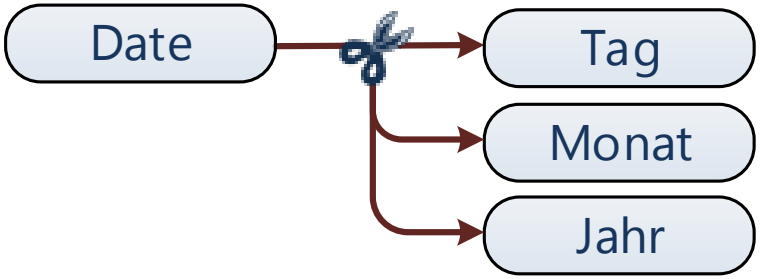
Arten von Schema Korrespondenzen

1:1 Korrespondenz

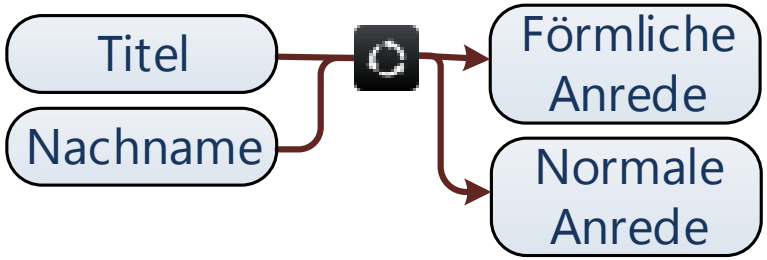
n:1 Korrespondenz



1:n Korrespondenz



n:m Korrespondenz



Konstruktion von Schema Mappings

Fragestellung: Welche Korrespondenzen müssen wie aufgelöst werden?

- Erkennen einer Menge an gefundenen Korrespondenzen die ein logisches Mapping bilden
- Erkennen der richtigen Funktionen um mehrerer Attribute bei 1:n, n:1 oder n:m Korrespondenzen zu kombinieren
  (speziell bei numerischen Attributen schwierig)

Ergebnis: Mapping zwischen lokalem und globalem Schema

- **Local-As-View:** Lokale Relation als (SQL-)Sicht auf globales Schema
- **Global-As-View:** Globale Relation als (SQL-)Sicht auf **alle** (relevanten) lokalen Schemata

Duplikaterkennung

ID	*first name*	*last name*	*size*	*gender*	*clusterID*
1	Jane Doe	*null*	5.48	ff	1
2	Meg Lee	*null*	5.61	f	4
3	Bill Hull	*null*	5.84	m	2
4	Jule	Smith	147	0	3
5	Meg	Li	171	0	4
6	William	Hall	178	1	2
7	Paul	Ryan	1.82	*null*	5
8	Bil	Hall	178	1	2

Problemstellung

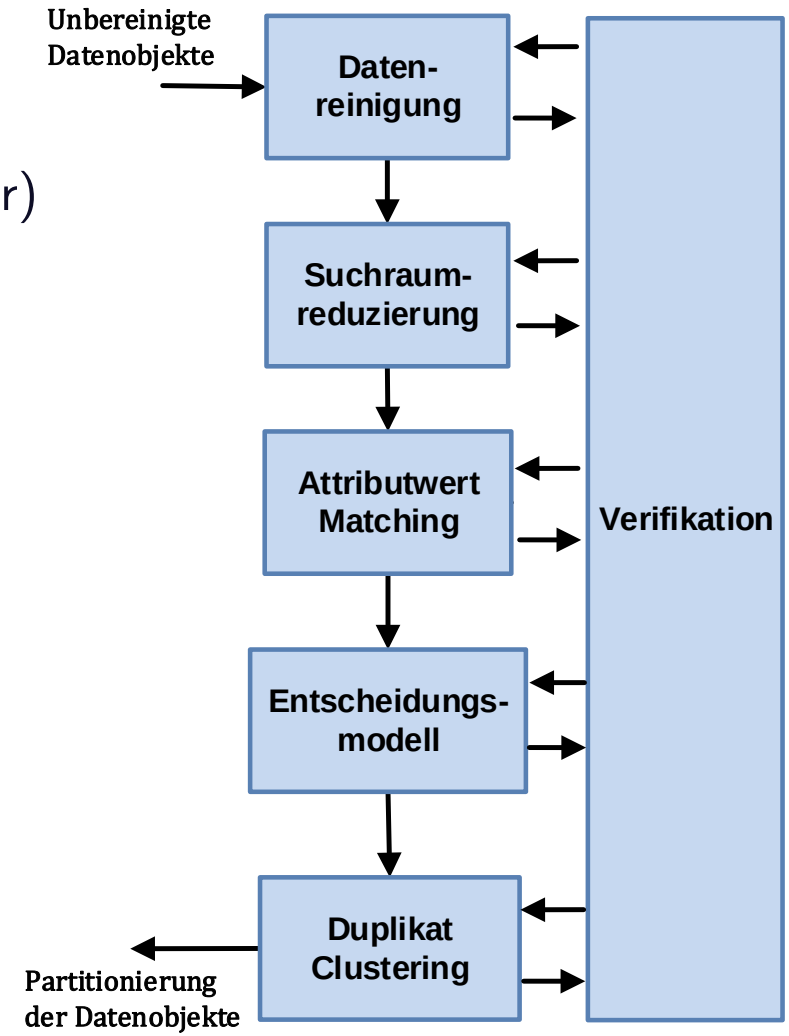
- Falsche, fehlende oder veraltete Werte
- Unterschiedliche Formatierungen, Codierungen oder Maßeinheiten

Ergebnis: Duplikatclustering

- Partitionierung aller Datenobjekte (ein Cluster pro Realwelt Entität)

Duplikaterkennung - Prozessablauf

- Datenreinigung
  (Standardisierung, Entfernung einfacher Fehler)
- Suchraumreduzierung
  (Effizienzerhöhung)
- Attributwertmatching
  (Messen von Attributwertähnlichkeiten)
- Entscheidungsmodell
  (ähnlichkeitsbasierte Entscheidungsfindung)
- Duplikat Clustering
  (Clustern der paarweisen Entscheidungen)
- Verifikation/Evaluation
  (Abschätzung der Ergebnisqualität)



Organisation: Vortrag

Vortrag:

- Dauer: ca. 25 Minuten (einzeln) und ca. 45 Minuten (zu zweit)
- Anschließende Diskussion
- Bewertung:
  - Inhaltlich
    - Inhaltliche Richtigkeit
    - Tiefgang
    - Nutzung von Beispielen
    - Kompetenz in der Fragenbeantwortung
  - Präsentation
    - Vortragsstil
    - Foliengestaltung
    - Verständlichkeit
- Schriftliches Feedback durch alle Zuhörer

Organisation: Seminararbeit

Seminararbeit:

- Umfang (netto): 10-12 Seiten (einzeln) und 15-20 Seiten (zu zweit)
- Bewertung:
  - Inhaltlich
    - Inhaltliche Richtigkeit
    - Tiefgang
  - Präsentation
    - Aufbau der Arbeit
    - Illustrationen
    - Nutzung von Beispielen
  - Arbeitstechnik
    - Selbständige wissenschaftliche Arbeitsweise
    - Selbständige Quellensuche
    - Korrekte Zitierung

Richtlinien: Prüfung & Prüfungsvoraussetzungen

Prüfung:

- mündlich (Dauer ca. 30 Minuten)
- Inhalt von Vorlesung und Saalübungen
- Seminarnote dient als Tie-breaker

Prüfungsvoraussetzung gilt als nicht erfüllt wenn:

- Präsentation oder Seminararbeit nicht den Anforderungen genügt
- Präsentation am vereinbarten Termin nicht erfolgt
- Seminararbeit nicht angefertigt wird
- Die Anwesenheitspflicht beim Seminar missachtet wurde

Richtlinien: Folien und Seminararbeit

Vortragsfolien:

- Empfohlene Verwendung der Folien-Vorlagen (PowerPoint, OpenOffice Impress oder LaTeX)
- Empfohlene Folienanzahl: 15-40
- Bevorzugte Verwendung von Vektorgraphiken

Seminararbeit:

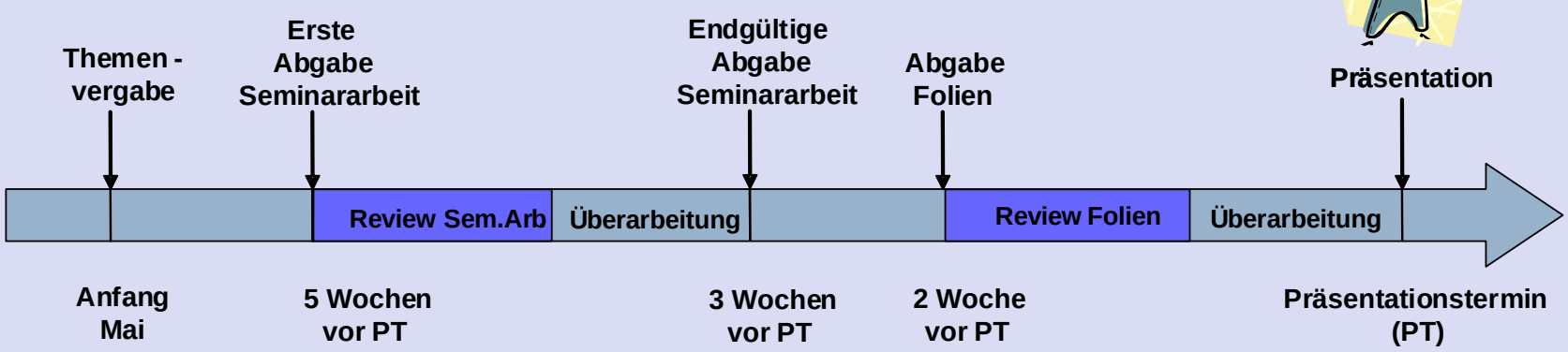
- Empfohlene Verwendung der Seminararbeits-Vorlage (MS Word, OpenOffice Writer oder LaTeX)

Vorlagen unter:

`https://vsis-www.informatik.uni-hamburg.de/vsis/teaching/templates`

Richtlinien: Folien und Seminararbeit

Idealer Zeitlicher Ablauf:



Abgaben an: panse@informatik.uni-hamburg.de

Termine

Vorschlag 1:

- **Vorlesung:** 28.08. - 01.09.2017 und 04.09. - 06.09.2017, täglich 4 Stunden (z.B. 09:30 - 13:30 Uhr)
- **Seminar:** 07.09. - 08.09.2017, vermutlich täglich 7 Stunden (z.B. 09:30 - 16:30 Uhr)

Vorschlag 2:

- **Vorlesung:** 04.09. - 08.09.2017 und 11.09. - 13.09.2017, täglich 4 Stunden (z.B. 09:30 - 13:30 Uhr)
- **Seminar:** 14.09. - 15.09.2017, vermutlich täglich 7 Stunden (z.B. 09:30 - 16:30 Uhr)

Ablauf: Seminar

- Begutachtung der vorgeschlagenen Themen:

  `vsis-www.informatik.uni-hamburg.de/oldServer/teaching/ss-17/ii/seminar/SeminarThemen.php`

- Themen (bzw. Kombinationen) die für mehr als eine Person geeignet sind werden dementsprechend markiert.
- Ab Montag dem 08.05.2017 um 20 Uhr, pro Teilnehmer/Team drei Themenwünsche an panse@informatik.uni-hamburg.de senden.
- Zuweisung der Themen (bei Konflikt zählt Zeitpunkt der E-Mail: *First-come, first-served*)
- Achtung: Aus Gründen der Fairness werden E-Mails die vor dem 08.05.2017 um 20 Uhr eintreffen bei Konflikten nicht priorisiert.
- Die Zuweisung der Themen wird über die Webseite bekannt gegeben.

weitere Informationen:

... gibt es auf der Homepage zum Seminar:

https://vsis-www.informatik.uni-hamburg.de/vsis/teaching/coursekvv/417