



DATENINTEGRATION & DATENHERKUNFT

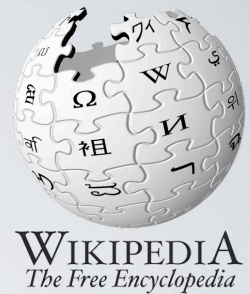
Übung Wintersemester 2010/11

Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

ZIELE DER ÜBUNG

- Projektbasierte Vertiefung der Vorlesungsinhalte
- Integration von Daten der Filmdomäne
- Teamarbeit
- Präsentation der Ergebnisse

FILMDATEN IM WEB



```
{{Infobox film
| name                = The Dark Knight
| image               = Dark_Knight.jpg
| image_size         = 
| caption             = Theatrical release poster
| director            = [[Christopher Nolan]]
| producer            = Christopher Nolan<br />[[Charles Roven]]<br />[[Emma Thomas]]
| screenplay           = Christopher Nolan<br />[[Jonathan Nolan]]
| story               = [[David S. Goyer]]<br />Christopher Nolan
| based on             = 
| [[Bill Finger]]     = 
| starring             = [[Christian Bale]]<br />[[Heath Ledger]]<br />[[Aaron Eckhart]]<br />[[Michael Caine]]<br />[[Maggie Gyllenhaal]]<br />[[Gary Oldman]]<br />[[Morgan Freeman]]<br />[[Monique Gabriela Curnen]]
<br />[[Aaron Eckhart]]<br />[[Michael Caine]]<br />[[Maggie Gyllenhaal]]<br />[[Gary Oldman]]<br />[[Morgan Freeman]]<br />[[Monique Gabriela Curnen]]
they are on the
| music               = 
| cinematography      = 
| editing              = 
| studio              = 
| distributor         = 
| released             = 
| country              = 
| language             = 
| budget              = 
| gross                = $1,001,921,825<ref name="revn">{{cite web|url=http://boxofficemojo.com/movies/?id=darkknight.htm|title=The Dark Knight (2008) |publisher=[[Box Office Mojo]] |accessdate=March 19, 2009}}</ref>
| preceded by         = ''[[Batman Begins]]''
}}
```

<http://en.wikipedia.org/wiki/Wikipedia:Film>

Daten ebenfalls erhältlich unter DBPedia 3.5 | <http://wiki.dbpedia.org/Downloads351>

FILMDATEN IM WEB



Daten erhältlich unter:
<http://www.imdb.com/interfaces>

The Dark Knight (2008)

8.9/10

Users: (470,927 votes) 3,345 reviews | Critics: 531 reviews

Batman, Gordon and Harvey Dent are forced to deal with the chaos unleashed by an anarchist mastermind known only as the Joker, as it drives each of them to their limits.

Director: [Christopher Nolan](#)

Writers: [Jonathan Nolan](#) (screenplay), [Christopher Nolan](#) (screenplay), and [3 more credits](#)

Release Date: 21 August 2008 (Germany)

Watch Trailer »

Full cast and crew | 103 photos | 33 videos »

Top 250 #11 | Won 2 Oscars. Another 79 wins & 61 nominations [See more](#) »

Photos

Cast

Cast overview, first billed only:

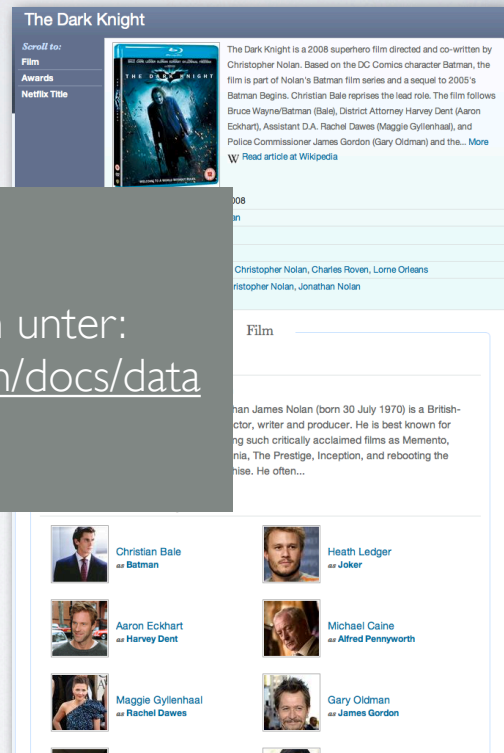
- [Christian Bale](#) ... [Bruce Wayne](#) / [Batman](#)
- [Heath Ledger](#) ... [Joker](#)
- [Aaron Eckhart](#) ... [Harvey Dent](#)
- [Michael Caine](#) ... [Alfred](#)
- [Maggie Gyllenhaal](#) ... [Rachel](#)
- [Gary Oldman](#) ... [Gordon](#)
- [Morgan Freeman](#) ... [Lucius Fox](#)
- [Monique Gabriela Curnen](#) ... [Det. Anna Ramirez](#)

FILMDATEN IM WEB



- Integrierte Daten aus
 - Wikipedia
 - ChefMoz
 - NNDB
 - MusicBrainz...
- User Content

Datenzugriff möglich unter:
<http://www.freebase.com/docs/data>



FILMDATEN IM WEB

- MoviLens
 - Filmdaten für Forschungszwecke
 - Recommender-Systeme
- Speichert Filme und Bewertungen



Daten verfügbar unter:
<http://www.grouplens.org/node/73>

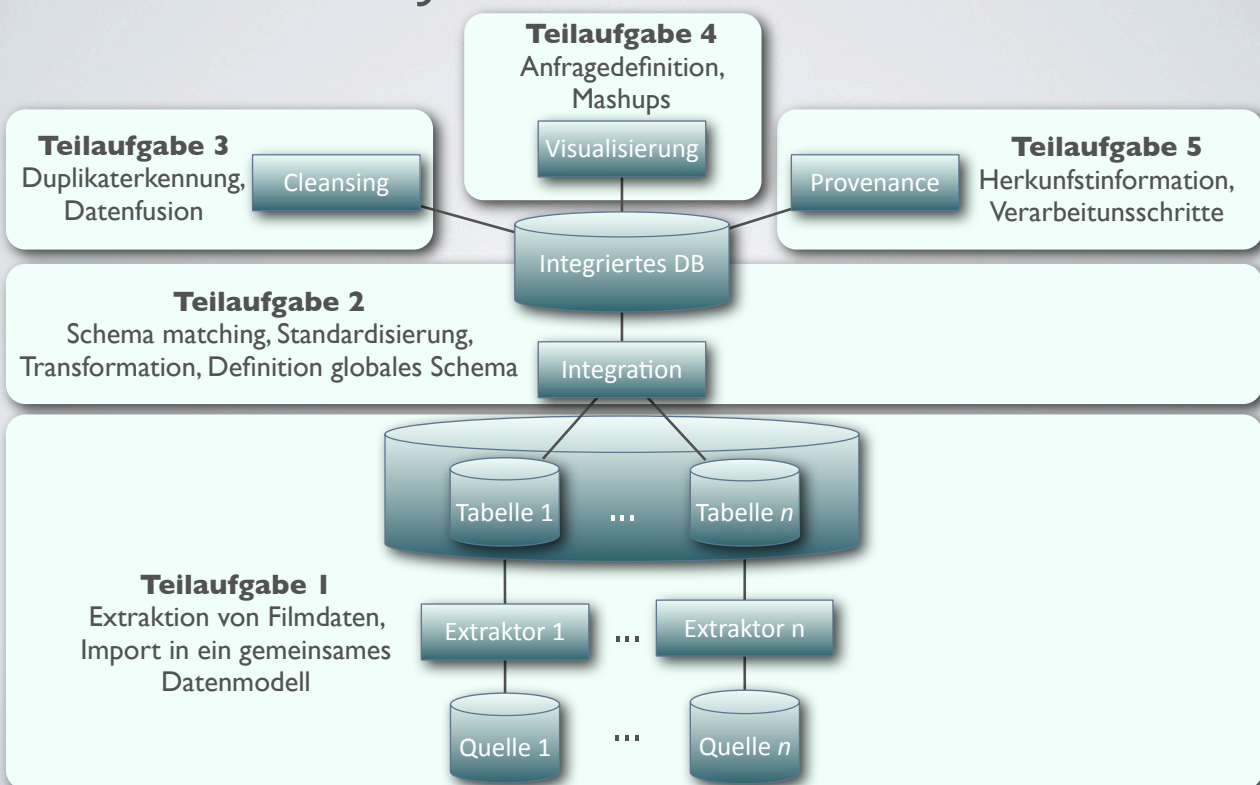
INTEGRATION VON FILMDATEN

Ihre Aufgaben

- 1.Extraktion von Daten aus mehreren Quellen
- 2.Integration der extrahierten Daten
- 3.Datenreinigung
- 4.Visualisierung interessanter Aspekte
- 5.Daten mit Herkunftsinformation anreichern

Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

PROJEKTÜBERSICHT



ZEITPLAN

Termin 1	Heute	Überblick + Einführung Extraktion	
Termin 2	10.11.2010	Einführung Integration	Vorstellungen Extraktion
Termin 3	24.11.2010		Vorstellungen Integration
Termin 4	8.12.2010	Einführung Datenreinigung	
Termin 5	22.12.2010	Einführung Visualisierung	Vorstellungen Datenreinigung
Termin 6	19.01.2011	Einführung Datenherkunft	Vorstellungen Visualisierung
Termin 7	26.01.2011		Vorstellungen Gesamtprojekt

Notenrelevant!

Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

AUFGABEN

- 2er Teams
- Jedes Team integriert unterschiedliche Datenquellen (mind. eine Quelle unterscheidet sich jeweils).
- Jedes Team muss alle Aufgaben bearbeiten.
- Jedes Team muss zu jeder Aufgabe einen Lösungsansatz präsentieren.
- Die finale Präsentation muss eine Demo einer funktionierenden Anfragebearbeitung beinhalten.

Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

BEWERTUNGSKRITERIEN

- Es geht zu gleichen Teilen in die Bewertung der Übung ein:
 - **Implementierung:**
 - Qualität, Allgemeingültigkeit und Umfang des Lösungsansatzes
 - Es sollte mind. eine spezifischen Fragestellung an die integrierten Daten korrekt beantwortet und visualisiert werden können.
 - **Präsentationen**
 - Qualität und Umfang der Inhalte
 - Klarheit der Darstellung
- Für **Masterstudenten**: die Gewichtung VL / Übung beträgt 75% / 25%

Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

EINFÜHRUNG EXTRAKTION

- Wählen Sie wie folgt Datenquellen aus
 - IMDB (Pflicht)
 - Eine Quelle, Wahl zwischen DBPedia, Freebase, MovieLens
 - Eine weitere Quelle Ihrer Wahl die ihre bisherigen Daten “in der Breite” komplementiert
 - z.B. Filmkritiken zu Filmen hinzufügt, Schauspielerdaten anreichert, DVD Preise angibt, ...
 - Die Quelle muss über Informationen von mindestens 500 Filmen verfügen.

Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

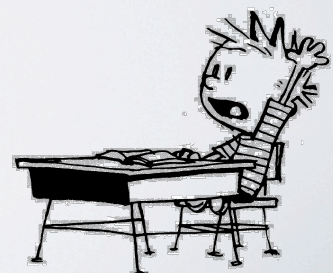
EINFÜHRUNG EXTRAKTION

- Für jede Datenquelle
 - Erstellen Sie ein relationales Schema (mind. eine Tabelle pro Quelle)
 - Ein Schema sollte viele interessante Attribute beinhalten, d.h., behalten Sie soweit möglich alle Attribute des Originalschemas.
 - Erstellen Sie einen Extraktor der die Daten in ihr Schema lädt.
 - Es empfiehlt sich, die Daten in eine relationale DB, z.B. PostgreSQL oder IBM DB2 zu laden.
 - Laden Sie die Daten in die von Ihnen gewählte DB.

Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen

ABGABE

- Erstellen Sie eine Präsentation Ihrer bisherigen Lösung.
- Schicken Sie mir Ihre Präsentation (PDF, Keynote, oder PPT(X)) mit Angabe Ihrer Namen und Matrikelnummern bis zum 9.11., 18 Uhr
- Am 10.11. stellt ihr Team seine Ergebnisse vor. Zusätzlich zu den Folien können Sie anhand Ihres Prototypen Ihre Lösung veranschaulichen.



Melanie Herschel | Lehrstuhl für Datenbanksysteme | Universität Tübingen