Praktikum Semantic Web

TECHNISCHE UNIVERSITÄT DARMSTADT

Sommersemester 2012

Frederik Janssen, Heiko Paulheim

Fachgebiet Knowledge Engineering

Organisatorisches

TECHNISCHE UNIVERSITÄT DARMSTADT

- Zeitslot:
  - Donnerstags, 8:55-10:55
  - ca. jede zweite Woche
- Infos zur Veranstaltung:
  - http://www.ke.tu-darmstadt.de/lehre/ss12/praktikum_semantic_web
- Wichtig: bitte in TUCaN anmelden!

Programm heute

TECHNISCHE UNIVERSITÄT DARMSTADT

- Semantic Web in a Nutshell
- Was ist Ontology Matching?
- Gängige Matching-Verfahren

- Organisatorisches
- Software und Infrastruktur

Das „klassische“ Web

TECHNISCHE UNIVERSITÄT DARMSTADT

aus Sicht des Menschen:

```
<html>
  …
  <b>Dr. Mark Smith</b>
  <i>Physician</i>
  Main St. 14
  Smalltown
  Mon-Fri 9-11 am
  Wed 3-6 pm
  …
</html>
```

aus Sicht des Computers:

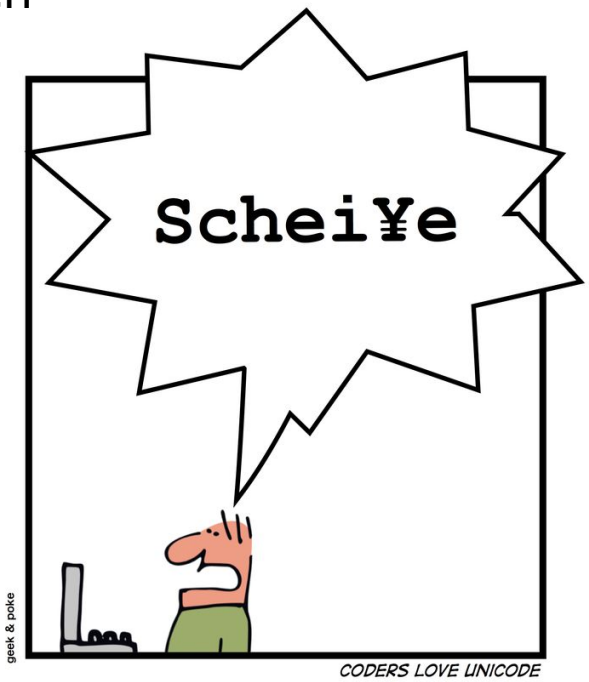
Dr. Mark Smith
Physician
Main St. 14
Smalltown
Mon-Fri 9-11 am
Wed 3-6 pm

Print in bold: „hmf298hmmhudsa“
Print in italics: „mj2i9ji0“
Print normal: „fdsah
02hfadsh0um2m0adsmf0ihm
asdfjköfdsa298ndsfmij32mio
lk2mjpoimjiofdpmsajiomjm“

Probleme des klassischen Webs

TECHNISCHE UNIVERSITÄT DARMSTADT

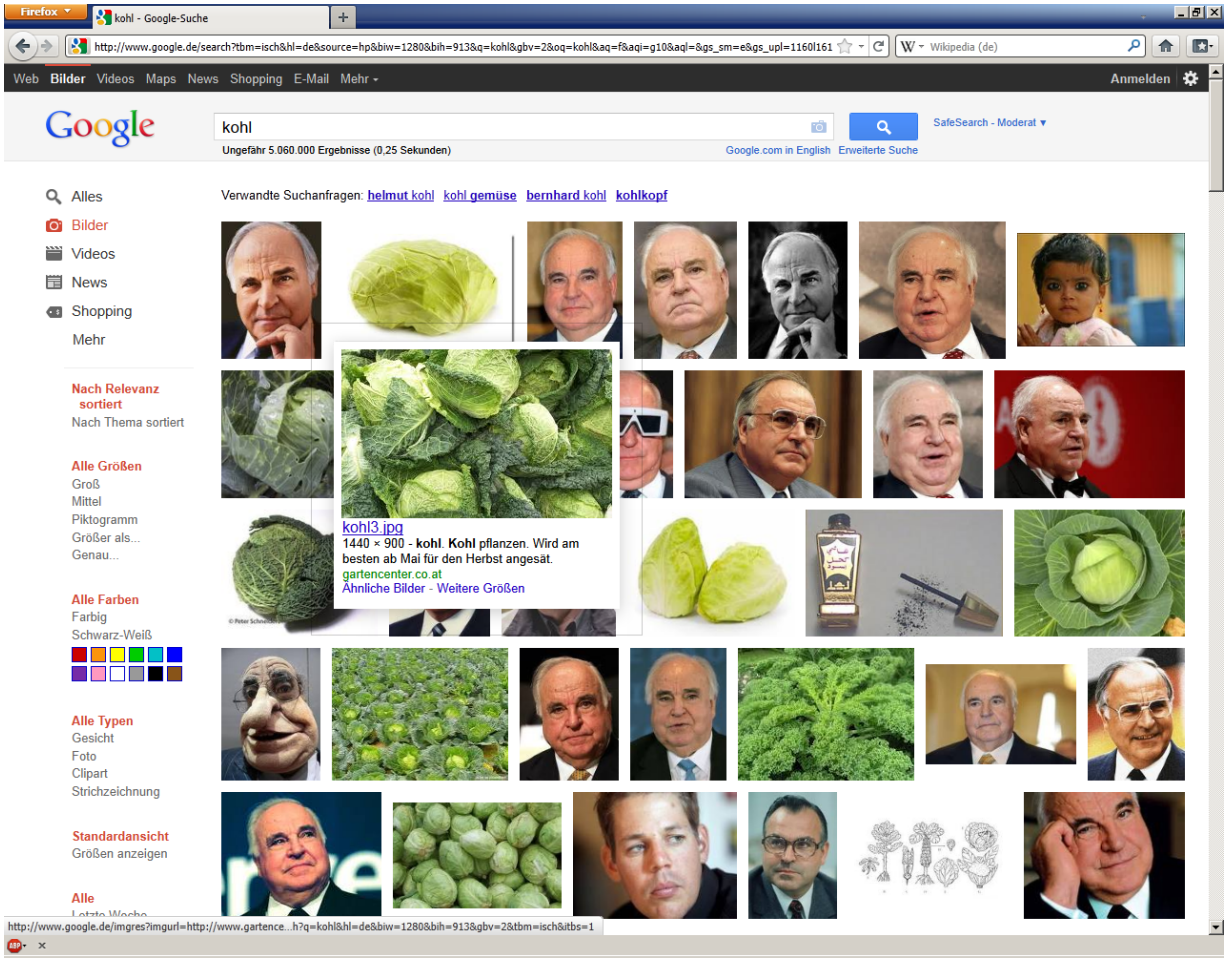
- Informationen finden
  - Stichwortbasierte Textsuche statt echter Fragen
  - verschiedene natürliche Sprachen
  - Homonyme/Polyseme
  - Synonyme
- Information verarbeiten
  - Formate (Encodings, Bilder, Videos, PDFs, …)
- Information verwerten
  - Verteilt auf verschiedene Seiten
  - Bsp.: Information zu Buchautor auf Verlagsseite, Adresse auf Uni-Seite



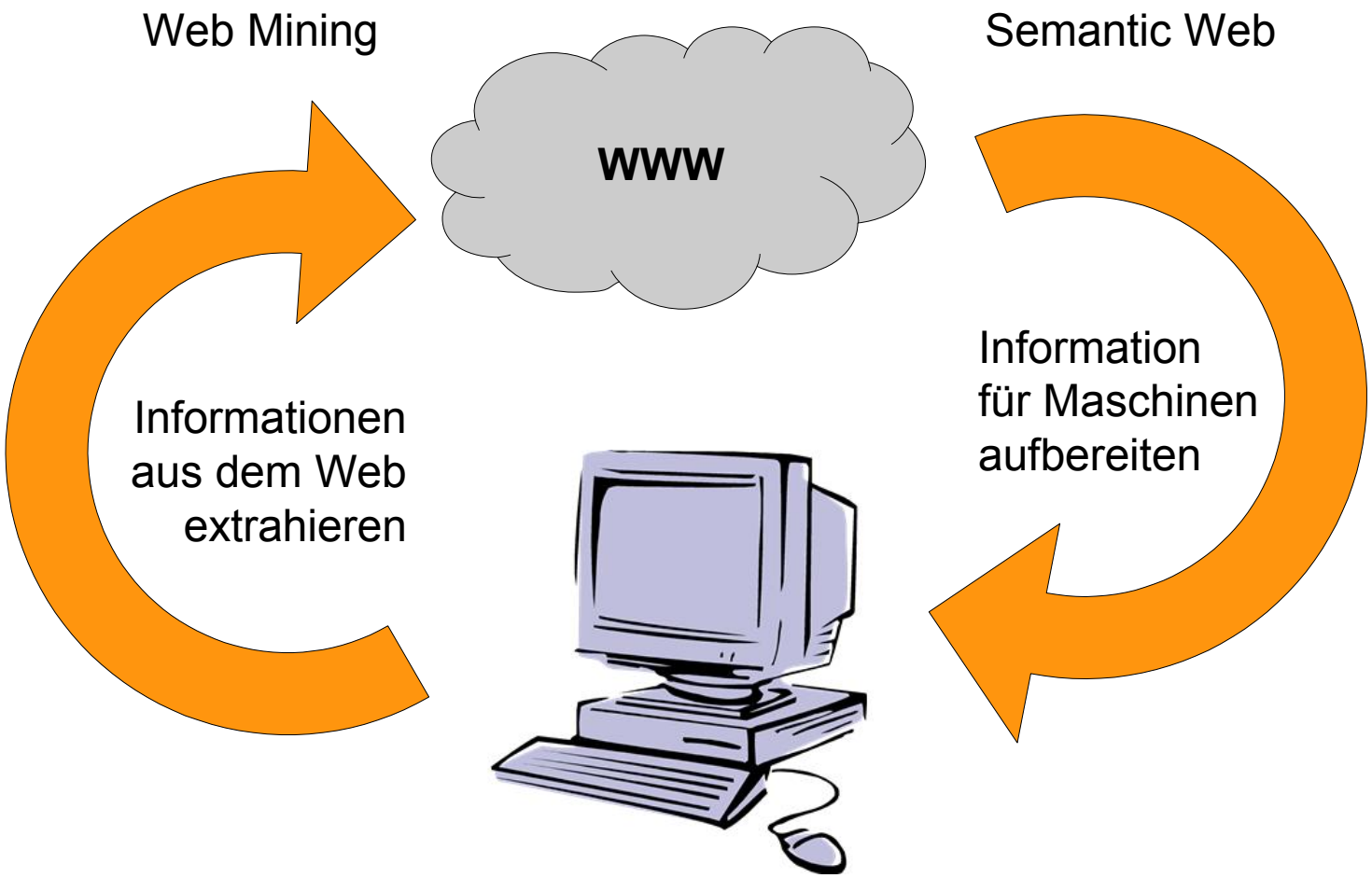
http://geekandpoke.typepad.com/geekandpoke/2011/08/coders-love-unicode.html

Probleme des klassischen Webs

TECHNISCHE UNIVERSITÄT DARMSTADT



Lösungsansätze



HTML vs. RDF

TECHNISCHE UNIVERSITÄT DARMSTADT

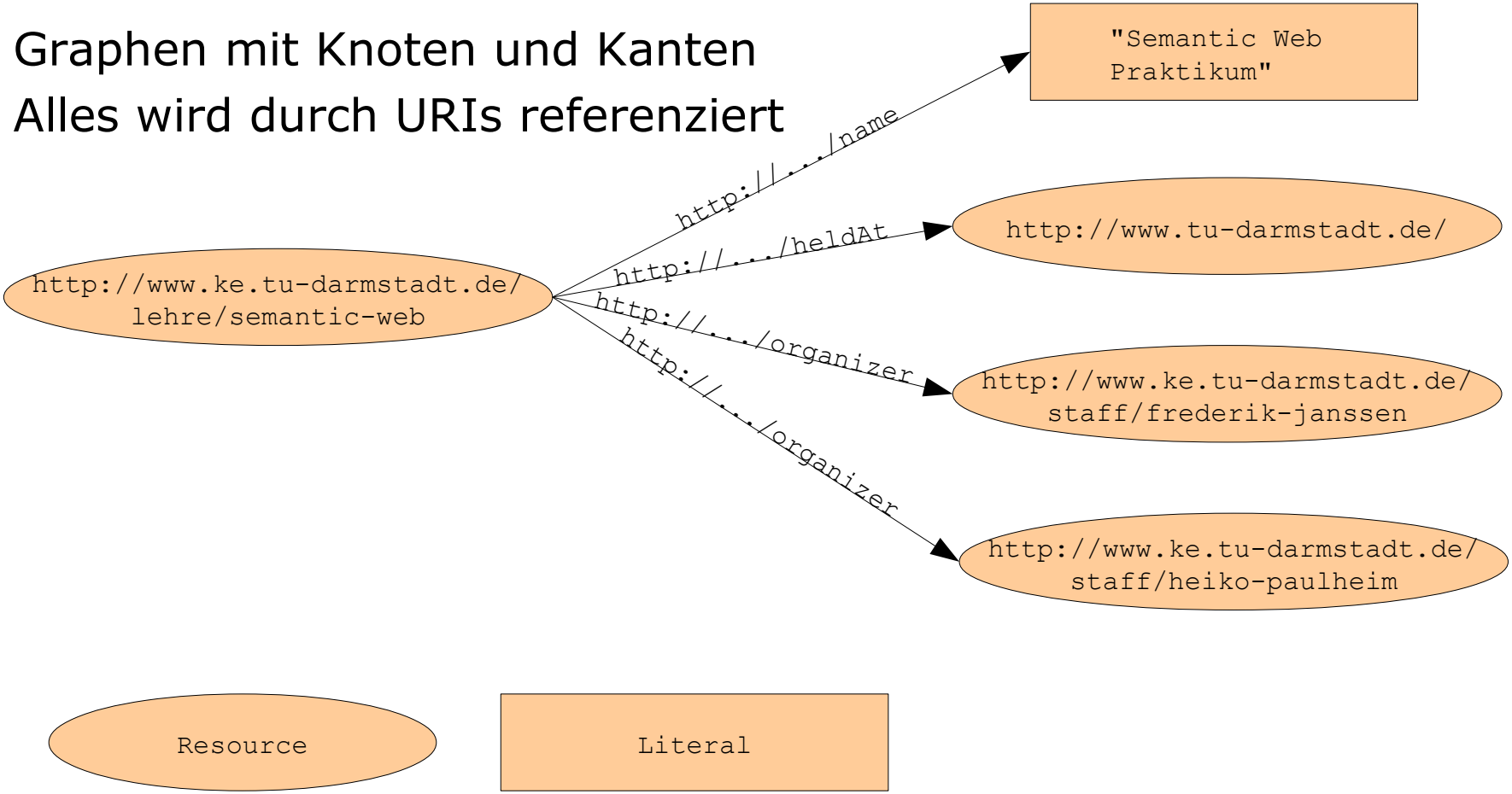
- Das "klassische" Web verwendet HTML
- Das semantische Web verwendet RDF
- Maschineninterpretierbares Format

```
<html>
  …
  <b>Dr. Mark Smith</b>
  <i>Physician</i>
  Main St. 14
  Smalltown
  Mon-Fri 9-11 am
  Wed 3-6 pm
  …
</html>
```

```
:p a :Physician .
:p :hasDegree "Dr." .
:p :hasName "Mark Smith" .
:p :hasAddress :a .
:a :street "Main Street" .
:a :number "14"^^xsd:int .
:a :city "Smalltown" .
:p :hasOpeningHours [
  a rdf:Bag ;
  [ :day :Monday;
    :from "9"^^xsd:int;
    :to "11"^^xsd:int;
  ]
  ...
```

RDF in a Nutshell

- Graphen mit Knoten und Kanten
- Alles wird durch URIs referenziert



RDF in a Nutshell

TECHNISCHE UNIVERSITÄT DARMSTADT

- Serialisierung N3: Subjekt – Prädikat – Objekt

```
@prefix ke: <http://www.ke.tu-darmstadt.de/>
ke:semweb2012 ke:organizer ke:staff/frederik-janssen .
ke:semweb2012 ke:organizer ke:staff/heiko-paulheim .
```

- Serialisierung RDF/XML:

```
<rdf:Description rdf:about="http://www.ke.tu-darmstadt.de/lehre/semweb2012">
  <ke:organizer rdf:resource="http://www.ke.tu-darmstadt.de/staff/frederik-janssen"/>
  <ke:organizer rdf:resource="http://www.ke.tu-darmstadt.de/staff/heiko-paulheim"/>
</rdf:Description>
```

Ontologien: bringen Semantik ins Spiel

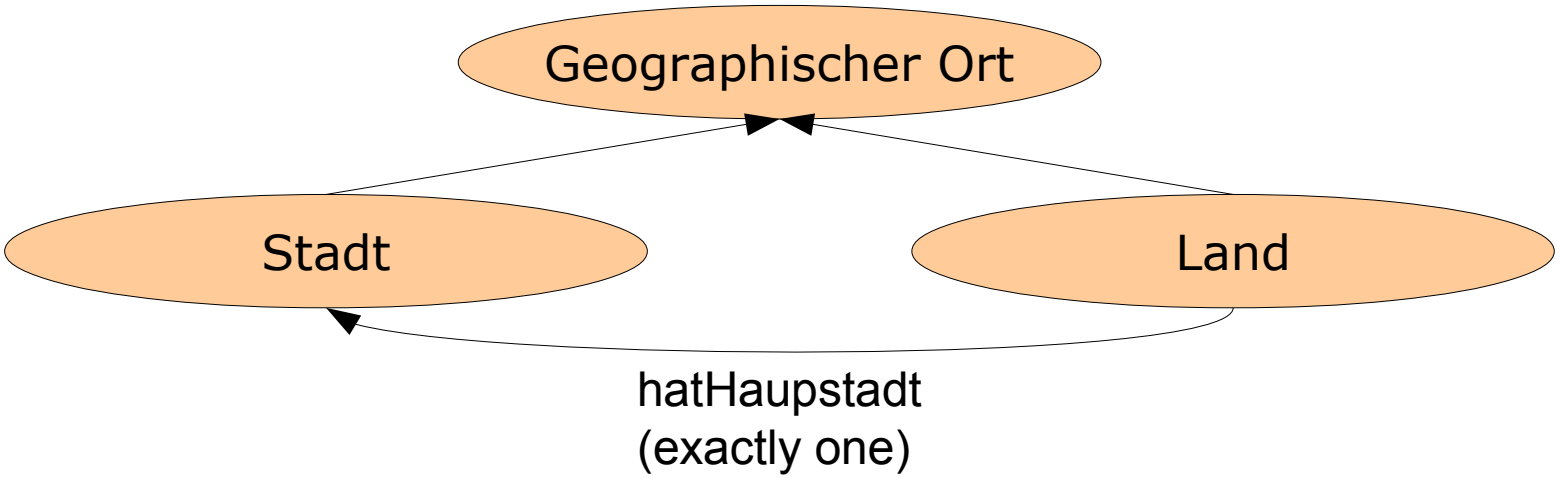
TECHNISCHE UNIVERSITÄT DARMSTADT

- RDF codiert Daten
  - z.B. "Madrid ist die Hauptstadt von Spanien"
- Ontologien codieren Hintergrundwissen
  - Hauptstädte sind Städte
  - Jedes Land hat genau eine Hauptstadt
- Damit kann man Schlussfolgern
  - z.B. "Madrid ist eine Stadt"
  - z.B. "Spanien ist ein Land"
  - z.B. "Paris ist nicht die Hauptstadt von Spanien"

Ontologien in a Nutshell

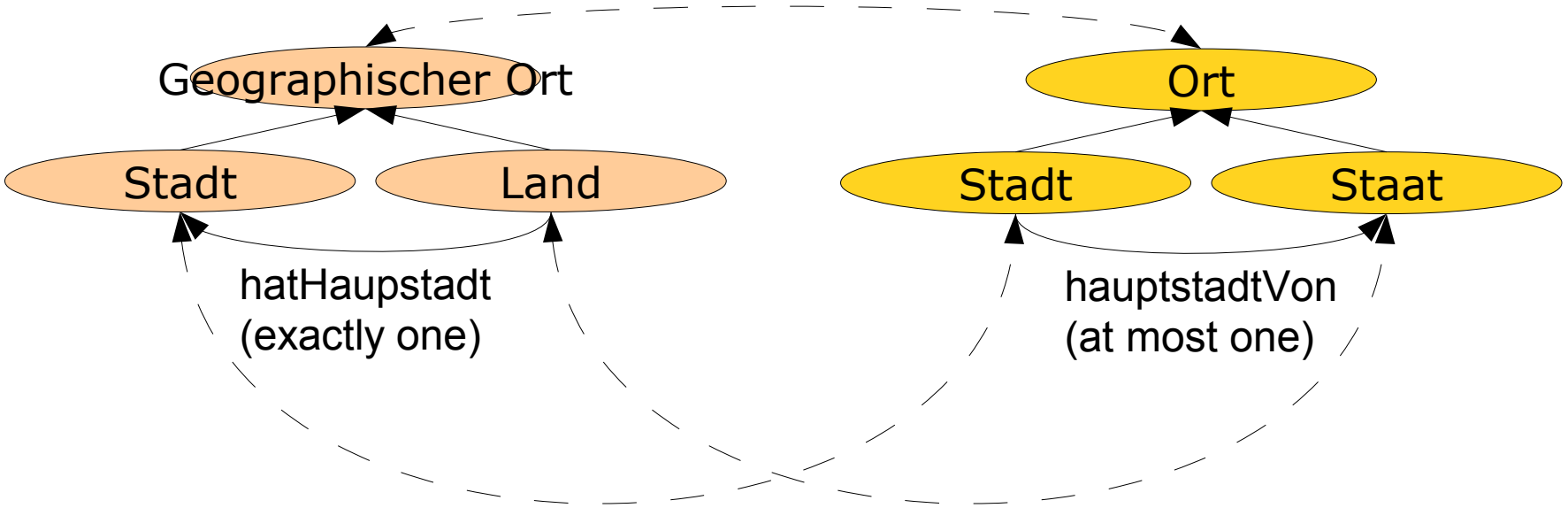
TECHNISCHE UNIVERSITÄT DARMSTADT

- Ontologien definieren
  - Klassen und Subklassen
  - Relationen
  - Einschränkungen



Ontology Matching

- Unterschiedliche Anwendungen, Services, Datensets
  - nutzen unterschiedliche Ontologien
  - sind daher nur beschränkt kompatibel
- Lösung: Ontologien abbilden



Ontology Matching

TECHNISCHE UNIVERSITÄT DARMSTADT

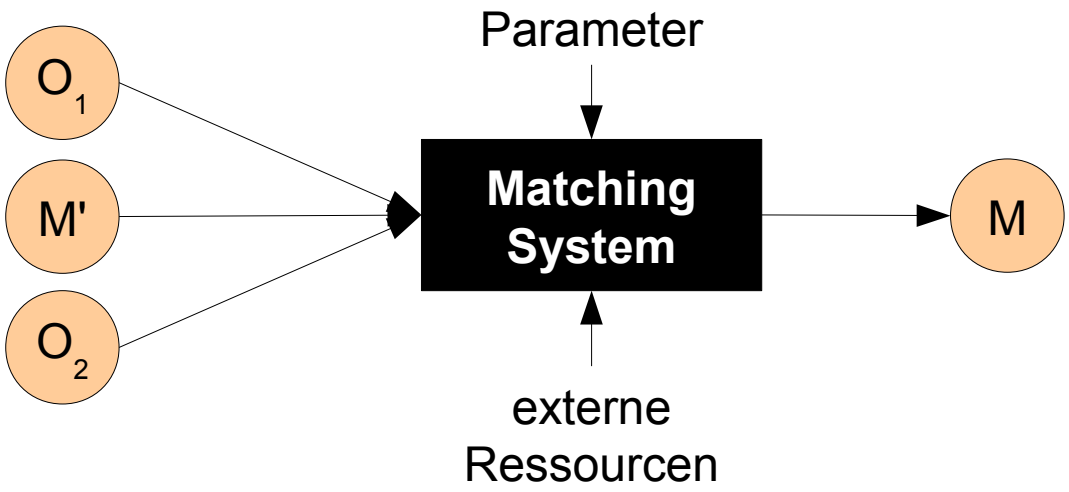
- Ziel: Abbildungen automatisch finden
  - Zwischen Klassen und Klassen
  - Zwischen Relationen und Relationen
    - unterschieden nach Objekt- und DataProperties

- ...und das möglichst gut
  - d.h. vor allem: möglichst exakt
  - NFRs: Skalierbarkeit, Performance

Ontology Matching

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ontology Matching:
  - automatisches Finden von Mappings
  - Gegeben: zwei Ontologien



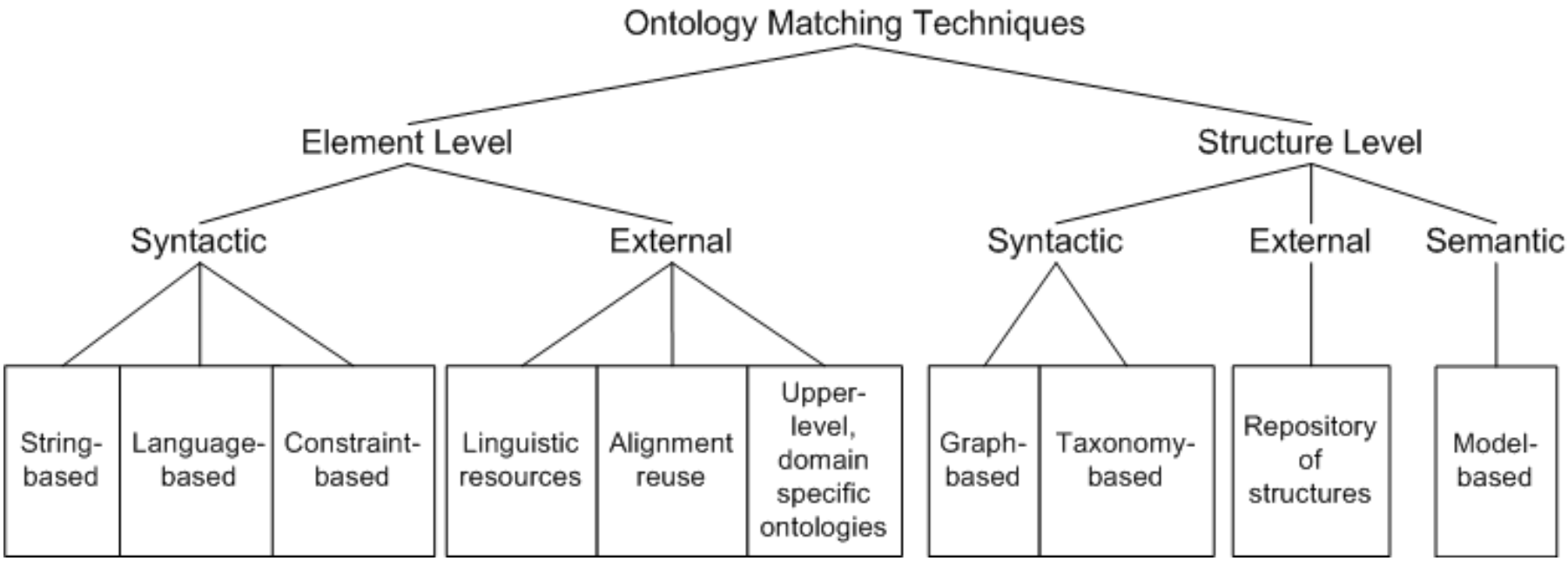
Euzenat & Shvaiko: Ontology Matching (2007)

Ontology Matching

TECHNISCHE UNIVERSITÄT DARMSTADT

- Hands on!

Ontology Matching – Verfahren



Euzenat & Shvaiko: Ontology Matching (2007)

Simple elementbasierte Techniken

TECHNISCHE UNIVERSITÄT DARMSTADT

- Element-basierte Verfahren vergleichen in der Regel Bezeichner
  - z.B. URI-Fragmente
    - `ex1:name`
    - `ex2:hasName`
  - z.B. Labels
    - `ex1:name rdfs:label "A person's name"@en .`
    - `ex2:hasName rdfs:label "The name of a person"@en .`
  - z.B. comments
    - `ex1:name rdfs:comment "Usually the family name"@en .`
    - `ex2:name rdfs:comment "Usual order: family name, given name"@en .`

String-basierte Verfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

- Direkte Stringgleichheit
  - z.B. ex1:Person, ex2:Person
- Suche von gemeinsamen Präfixen
  - z.B. ex1:phone, ex2:phoneNumber
- Suche von gemeinsamen Postfixen
  - z.B. ex1:name, ex2:hasName

- Verfeinerung:
  - Längenrelation von gemeinsamem Präfix/Suffix und String als Konfidenzmaß
  - z.B. ex1:phone, ex2:phoneNumber => c = 5/11
  - z.B. ex1:name, ex2:hasName => c = 3/7

String-basierte Verfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

- Edit-Distanz
  - Minimale Anzahl von Editier-Schritten, um von String 1 zu String 2 zu kommen, geteilt durch Länge des längeren Strings
    - ein Zeichen einfügen
    - ein Zeichen löschen
    - ein Zeichen ändern
  - z.B. ex1:hasName, ex2:firstName
    - hasName → fhasName → fhastName → fiastName → firstName
    - Edit-Distanz = 4/9

String-basierte Verfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

- N-Gramm-Analyse
  - wie viele Buchstabengruppen der Länge n stimmen überein geteilt durch Anzahl n-Gramme in längerem String (= Länge – n + 1)
  - häufig verwendet: n=3
  - z.B. ex1:hasName, ex2:firstName
    - übereinstimmende 3-Gramme: Nam, ame
    - gesamte 3-Gramme: fir, irs, rst, stN, tNa, Nam, ame

      → 3-Gramm-Distanz = 2/7

Sprachbasierte Techniken

TECHNISCHE UNIVERSITÄT DARMSTADT

- Werden insbesondere zur Vorverarbeitung genutzt, um die Treffgenauigkeit zu verbessern
- Linguistische Verfahren
- Eliminierung von Stop-Words
  - ex1:locatedIn → ex1:located
- Lemmatisierung/Stemming:
  - ex1:located, ex2:location
  - beides wird zu locat-
- Zerlegen in Token (Tokenization)
  - ex1:graduated_from_university → {graduated,from,university}
    ex2:isGraduateFromUniversity → {is,Graduate,from,University}
  - Token werden einzeln weiterverarbeitet

Sprachbasierte Verfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

- können Treffgenauigkeit verbessern
  - ex1:located, ex2:location
  - Stemming: ex1:locat-, ex2:locat-
  - Edit-Distance: 0 → hohe Ähnlichkeit
- Gegenbeispiel:
  - ex1:locationOf, ex2:locatedIn (Inverse Properties!)
  - Erster Schritt: Eliminierung von Stop-Words: ex1:location, ex2:located
  - Zweiter Schritt: Stemming: ex1:locat-, ex2:locat-
  - Edit-Distance: 0 → hohe Ähnlichkeit
- dagegen Edit-Distance ohne Vorverarbeitung: 0.5 → nicht so hoch

Constraint-basierte Verfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

- Zusätzlicher Abgleich der Eigenschaften von gemappten Entitäten
  - z.B. verwendete Multiplizitäten von Restriktionen
  - z.B. Datentypen bei DatatypeProperties
- Im letzten Beispiel:
  - ex1:locationOf
    - definiert als InverseFunctionalProperty
    - Kardinalität ist nicht eingeschränkt
  - ex2:locatedIn
    - definiert als FunctionalProperty
    - exactCardinality=1
- Damit können wir die Ähnlichkeit von ex1:locationOf und ex2:locatedIn reduzieren

Matching mit linguistischen Ressourcen

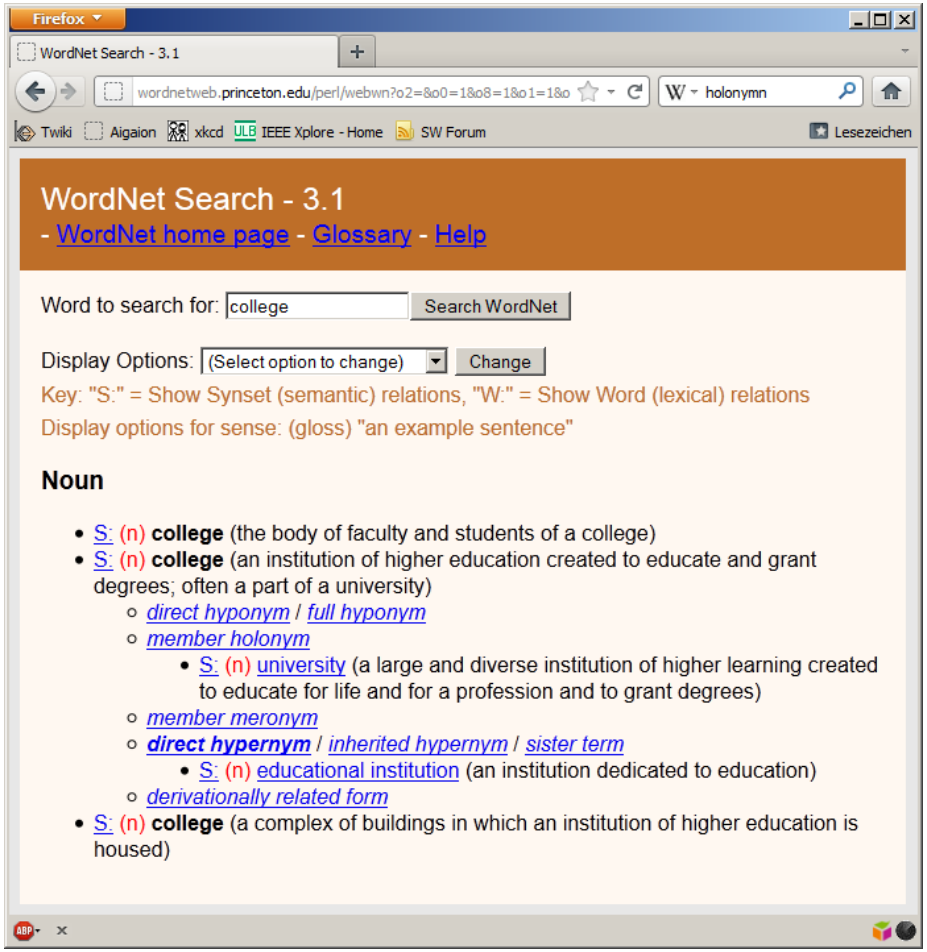
TECHNISCHE UNIVERSITÄT DARMSTADT

- Externes Wissen
- Beispiel: Synonyme
  - ex1:Verfasser, ex2:Autor
  - Edit-Distance: 8/9 (sehr hoch!)
  - Synonymwörterbuch kann hier eine Lösung sein
- Beispiel: Ontologien in unterschiedlichen Sprachen
  - ex1:Stadt, ex2:city
  - Edit-Distance: 1
  - Lexikalische Ressource kann hier eine Lösung sein

Matching mit linguistischen Ressourcen

TECHNISCHE UNIVERSITÄT DARMSTADT

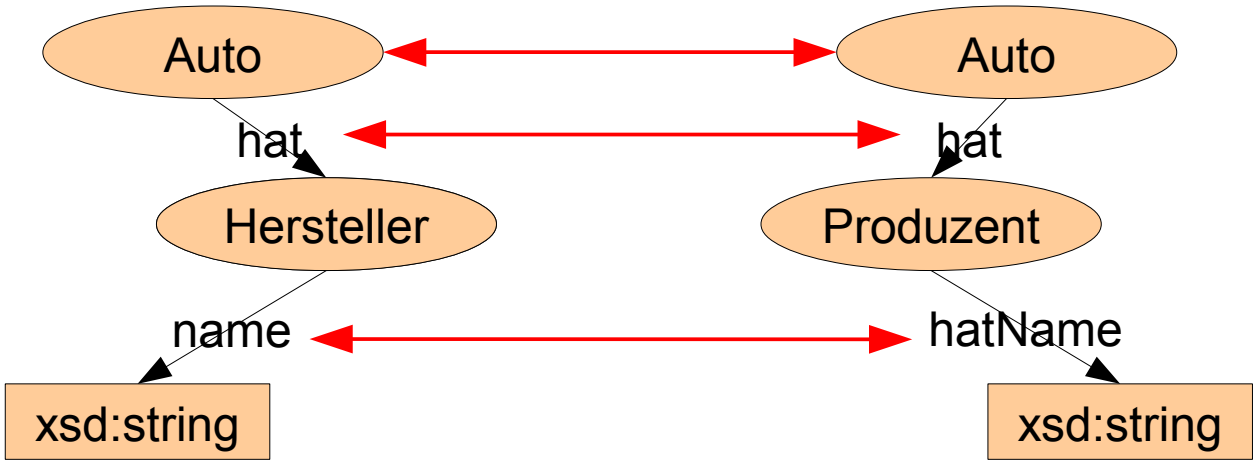
- WordNet
  - strukturiert für englisch
  - Synonyme, Hyponyme, Hyperonyme
  - Holo- und Meronyme



Graphenbasierte Verfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ontologien bilden Graphen
- Idee: Ähnlichkeiten in den Graphen propagieren
- z.B.
  - #gemappte Nachbarknoten / #Nachbarknoten
  - #gemappte Kanten / #Kanten



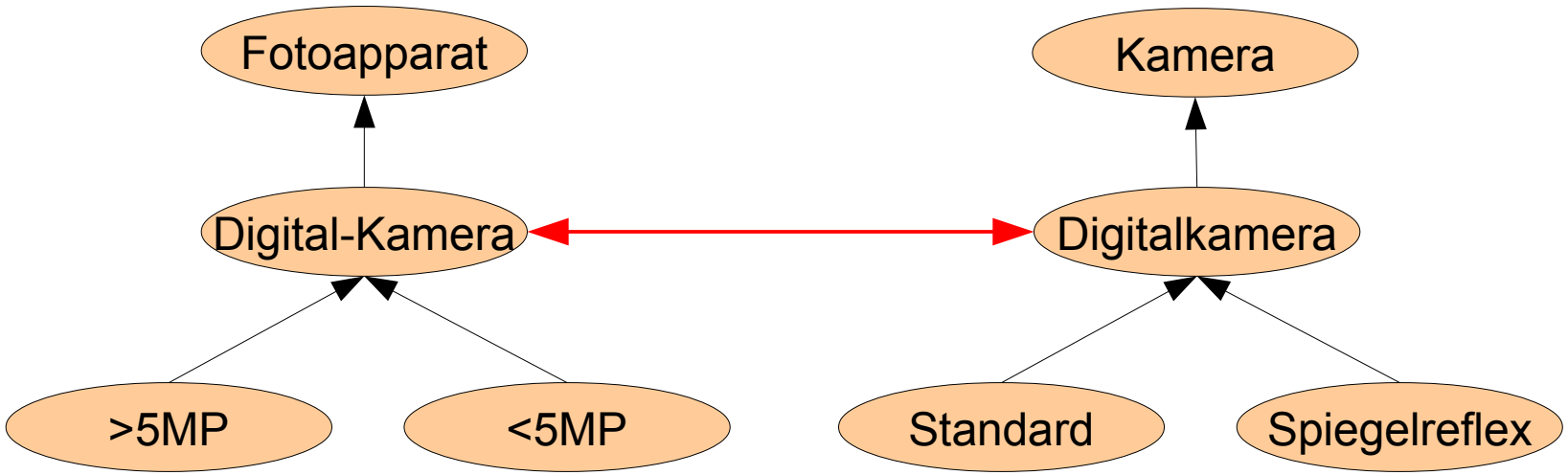
Graphenbasierte Verfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

- Sind auch geeignet, um Mappings für Properties zu bestimmen oder zu verfeinern:
  - wenn je zwei Klassen durch ein Property verbunden sind, dann kann dieses mit hoher Wahrscheinlichkeit gemappt werden

Taxonomiebasierte Verfahren

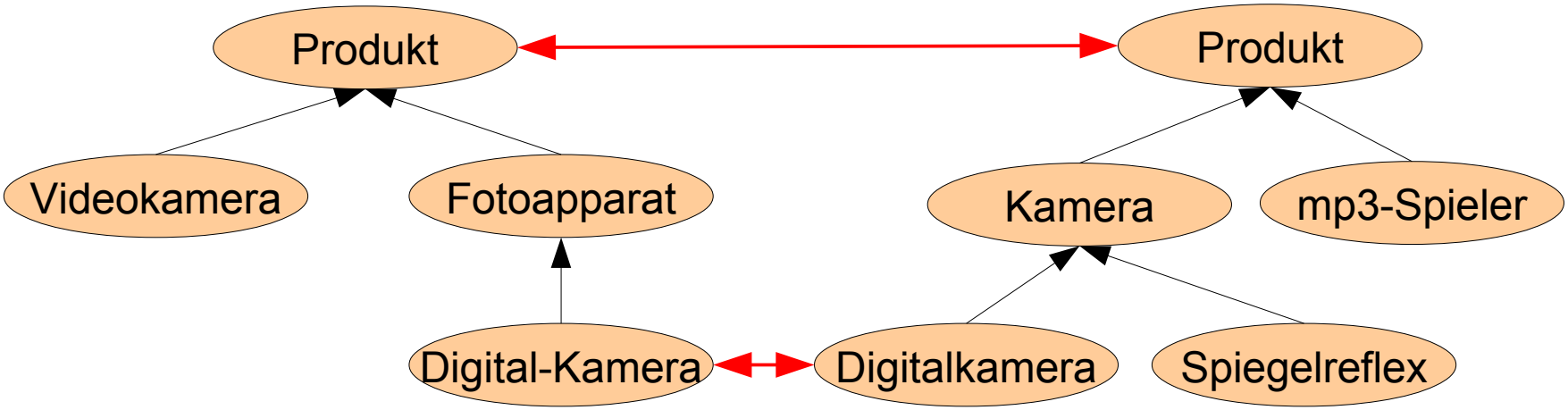
- Klassenhierarchien sind essentiell in Ontologien
- Super/Subkonzeptregeln: Kinder und Eltern von gemappten Elementen sind Kandidaten



Taxonomiebasierte Verfahren

TECHNISCHE UNIVERSITÄT DARMSTADT

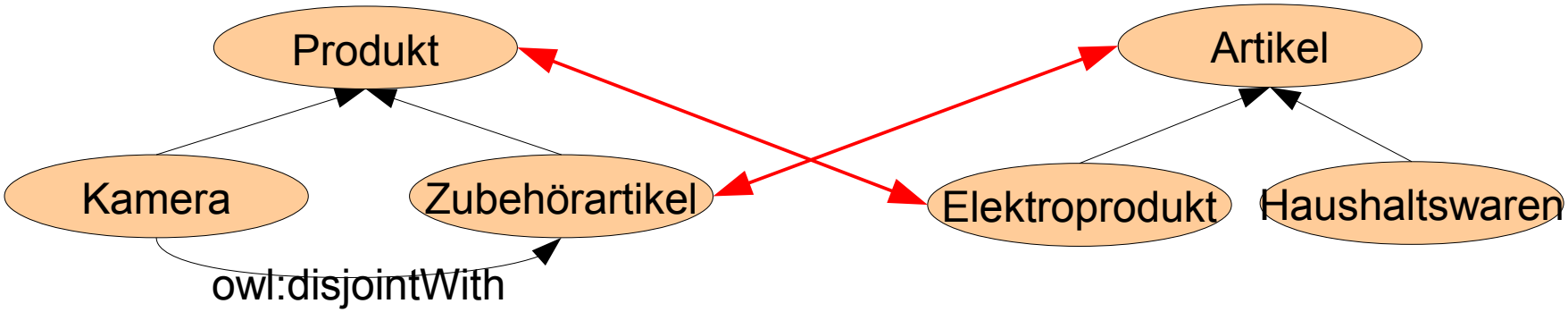
- Klassenhierarchien sind essentiell in Ontologien
- Bounded Path Matching: Auf Pfaden zwischen gemappten Elementen liegen wahrscheinlich weitere Kandidaten



Modellbasierte Verfahren

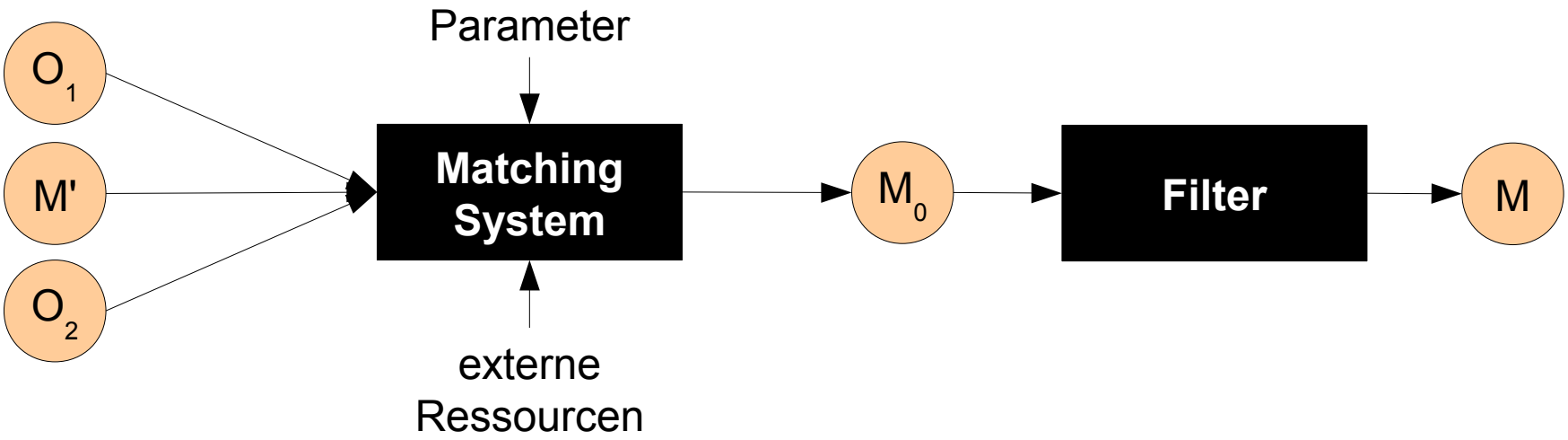
TECHNISCHE UNIVERSITÄT DARMSTADT

- Verwendung von einem Reasoner
- Verfahren:
  - beide Ontologien zu einer vereinigen
  - gefundene Mappings einfügen
  - mit Reasoner auf Widerspruchsfreiheit prüfen



Finalisierung der Mapping-Ergebnisse

- Ontology Matcher liefert
  - $<e_1,e_2,r,c>$
  - Übliches Verfahren: Rückgabe aller Mappings, für die c ≥ t gilt
  - Auch möglich: Filtern mit Plausibilitätsprüfung



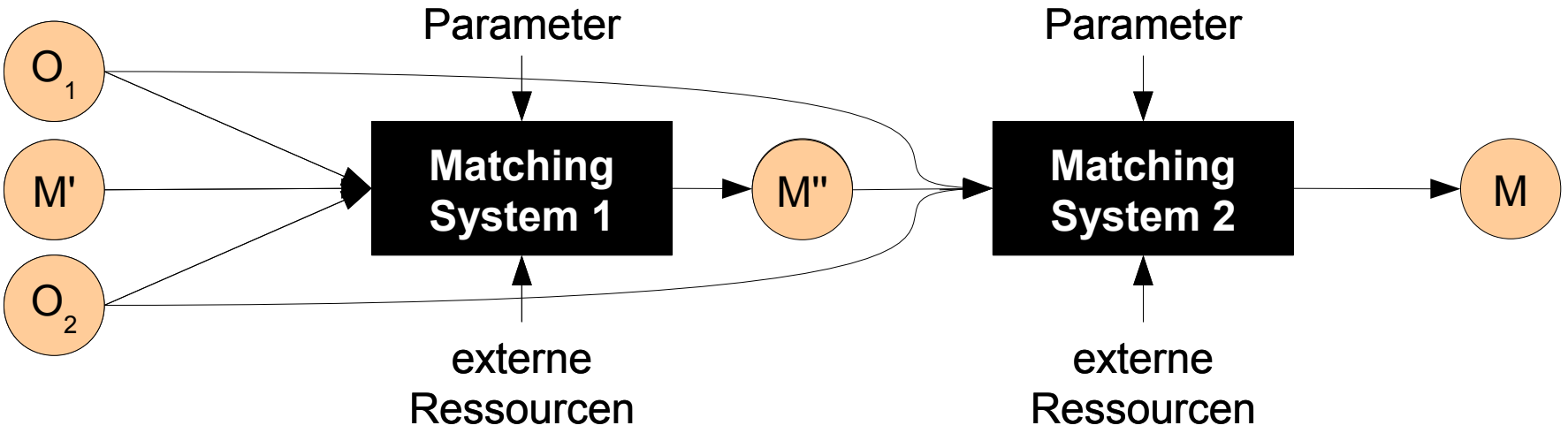
Kombination von Matchern

TECHNISCHE UNIVERSITÄT DARMSTADT

- Was wir bis jetzt gesehen haben
  - eine Reihe von Strategien
  - jede hat Stärken und Schwächen
  - in der Regel werden mehrere Verfahren kombiniert

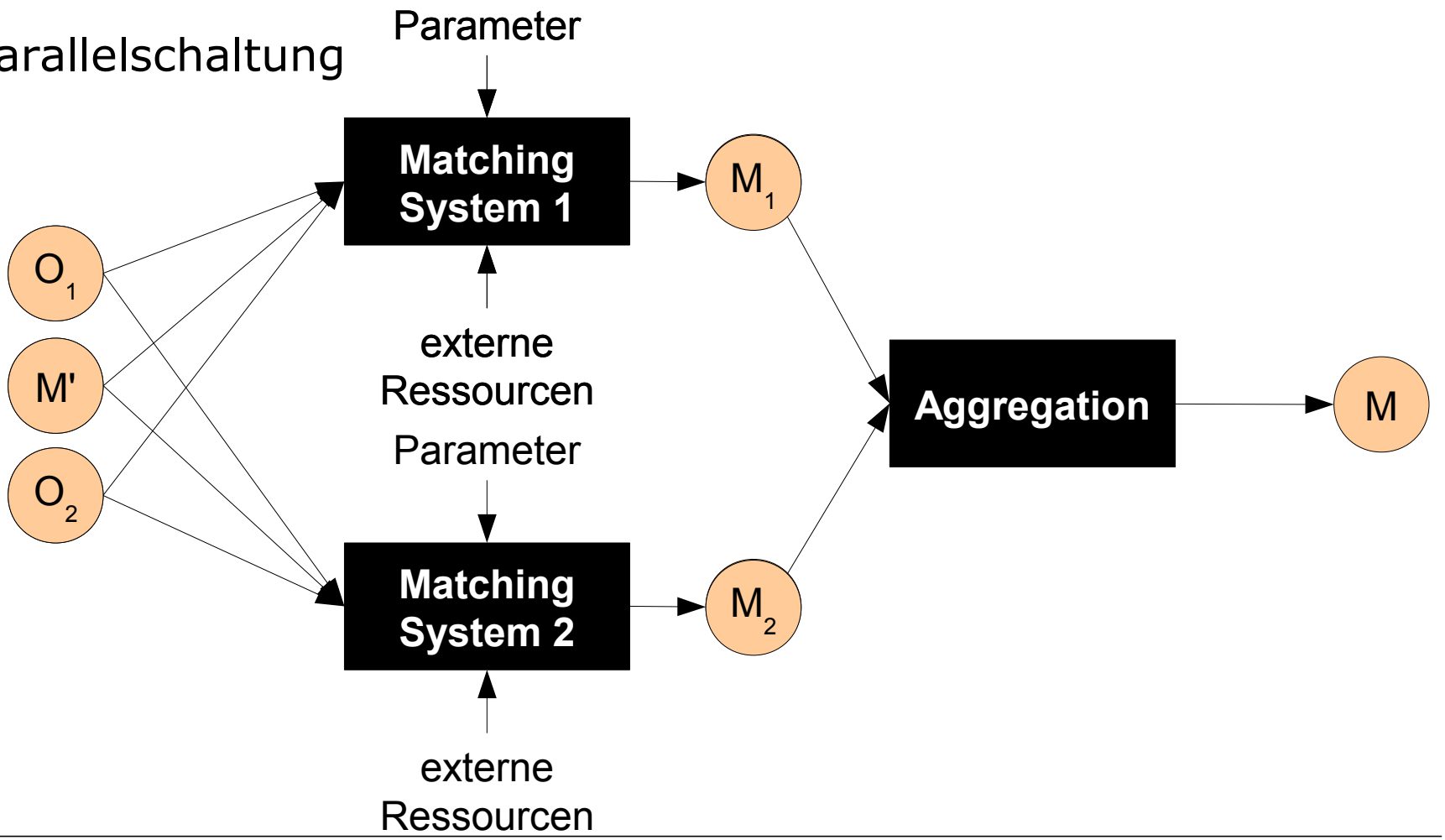
Kombination von Matchern

- Reihenschaltung
  - gut für Matcher, die Kandidaten benötigen:
    - Strukturbasierte Verfahren, z.B. Bounded Path
    - Modellbasierte Verfahren (Validierung von Kandidaten)



Kombination von Matchern

- Parallelschaltung



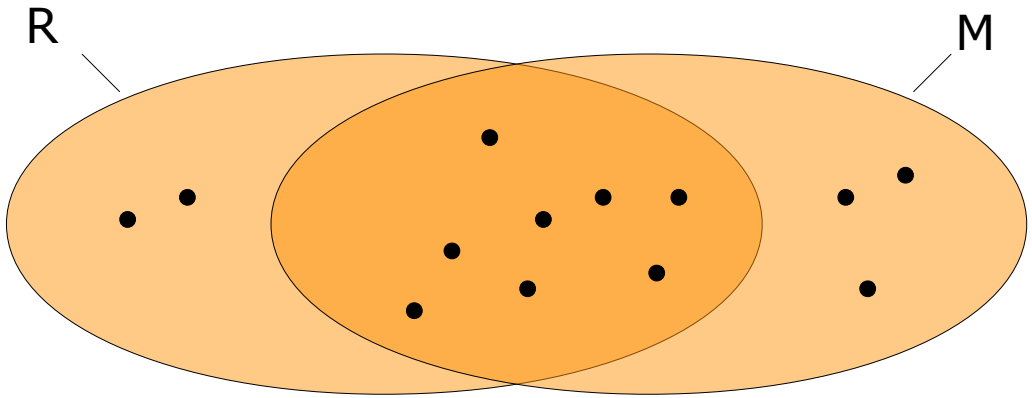
Kombination von Matchern

TECHNISCHE UNIVERSITÄT DARMSTADT

- Parallelschaltung
- ermöglicht
  - z.B. verschiedene elementbasierte Techniken
  - unterschiedliche Techniken auf unterschiedlichen Ontologieteilen
- Aggregation der Ergebnisse
  - auf Basis der Konfidenz
  - z.B. gewichtetes Mittel, Minimum, Maximum

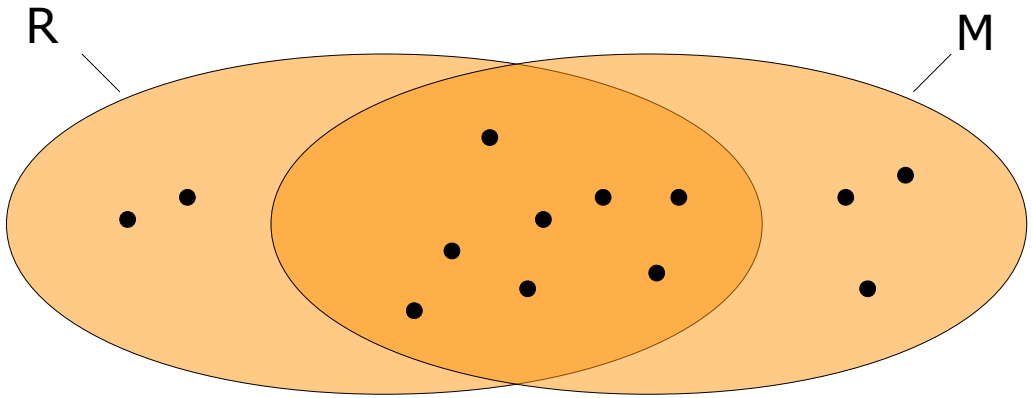
Auswertung von Mappings

- Gegeben:
  - das vom Matcher gefundene Mapping (M)
  - das tatsächliche Referenz-Mapping (R)
- Recall: der Anteil der gefundenen Elemente des Referenz-Mappings
  - rec = #(R ∩ M) / #R
  - im Beispiel: rec = 8/10 = 0.8



Auswertung von Mappings

- Gegeben:
  - das vom Matcher gefundene Mapping (M)
  - das tatsächliche Referenz-Mapping (R)
- Precision: der Anteil der korrekten Elemente unter den gefundenen
  - rec = #(R ∩ M) / #M
  - im Beispiel: rec = 8/11 = 0.73



Auswertung von Mappings

TECHNISCHE UNIVERSITÄT DARMSTADT

- Recall und Precision sind allein leicht zu optimieren
  - Recall: gib einfach alle möglichen Mappings zurück (Kreuzprodukt)
  - → rec = 1
- Precision
  - gib nur Elemente zurück, die fast sicher stimmen
  - → prec nahe 1 sehr wahrscheinlich

- Zwischen beidem existiert offenbar ein Trade-off
- Daher in der Praxis oft genutzt: F-Measure
  - harmonisches Mittel von Recall und Precision
  - 2*prec*rec / (prec +rec)

Organisatorisches

TECHNISCHE UNIVERSITÄT DARMSTADT

- Unser Ziel: gemeinsam einen guten Matcher bauen
  - aus Einzelverfahren, in Kleingruppen entwickelt
  - gemeinschaftliche Lösung
  - zur OAEI einreichen
  - Fokus: Benchmark+Conference

- Vorgehen:
  - Entwicklung von einzelnen Matchern
  - Kombination der Einzellösungen
  - Profitieren von den Lösungen anderer

Organisatorisches

TECHNISCHE UNIVERSITÄT DARMSTADT

- Treffen ca. alle zwei Wochen
  - Terminliste siehe Webseite
- Präsentation des Zwischenstandes
  - implementierte Verfahren
  - aktuelle Ergebnisse auf OAEI-Ontologien
    - Recall/Precision/F-measure
    - Verbesserungen und Verschlechterungen
    - jeweils für Benchmark und Conference
  - gut gelöste Fälle
  - problematische Fälle
  - geplante nächste Schritte

Organisatorisches

TECHNISCHE UNIVERSITÄT DARMSTADT

- Wiki und Forum zum Diskutieren
- SVN für gemeinsamen Code
  - jede Gruppe legt sich ein Unterverzeichnis an
  - Basis-Matcher ist gegeben

Basis-Matcher + Evaluierung mit SEALS

TECHNISCHE UNIVERSITÄT DARMSTADT

- Live-Demo

Erste Aufgabenstellung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Gegebenen Matcher weiterentwickeln
  - einige Basis-Verfahren implementieren
  - Fortschritte analysieren und dokumentieren

Praktikum Semantic Web

TECHNISCHE UNIVERSITÄT DARMSTADT

Sommersemester 2012

Frederik Janssen, Heiko Paulheim

Fachgebiet Knowledge Engineering