Karlsruher Institut für Technologie

Clustering von Internetdiensten für aktive Ontologien

Masterarbeit

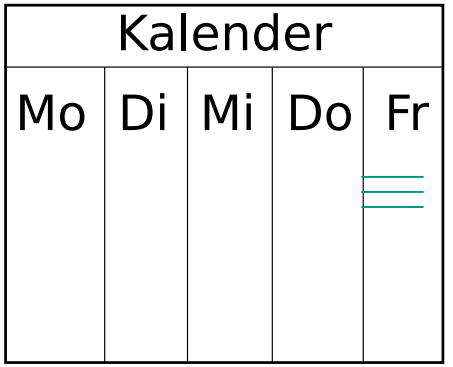
Philipp Lingel

Betreut von Martin Blersch und Mathias Landhäußer

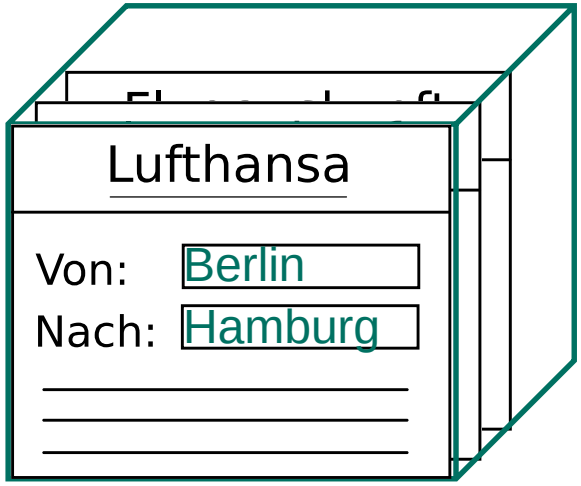
IPD Tichy, Fakultät für Informatik

Motivation


Aktive Ontologie
Termin
Titel
Wann
„Trage den Termin Präsentation am 25.09 ein“

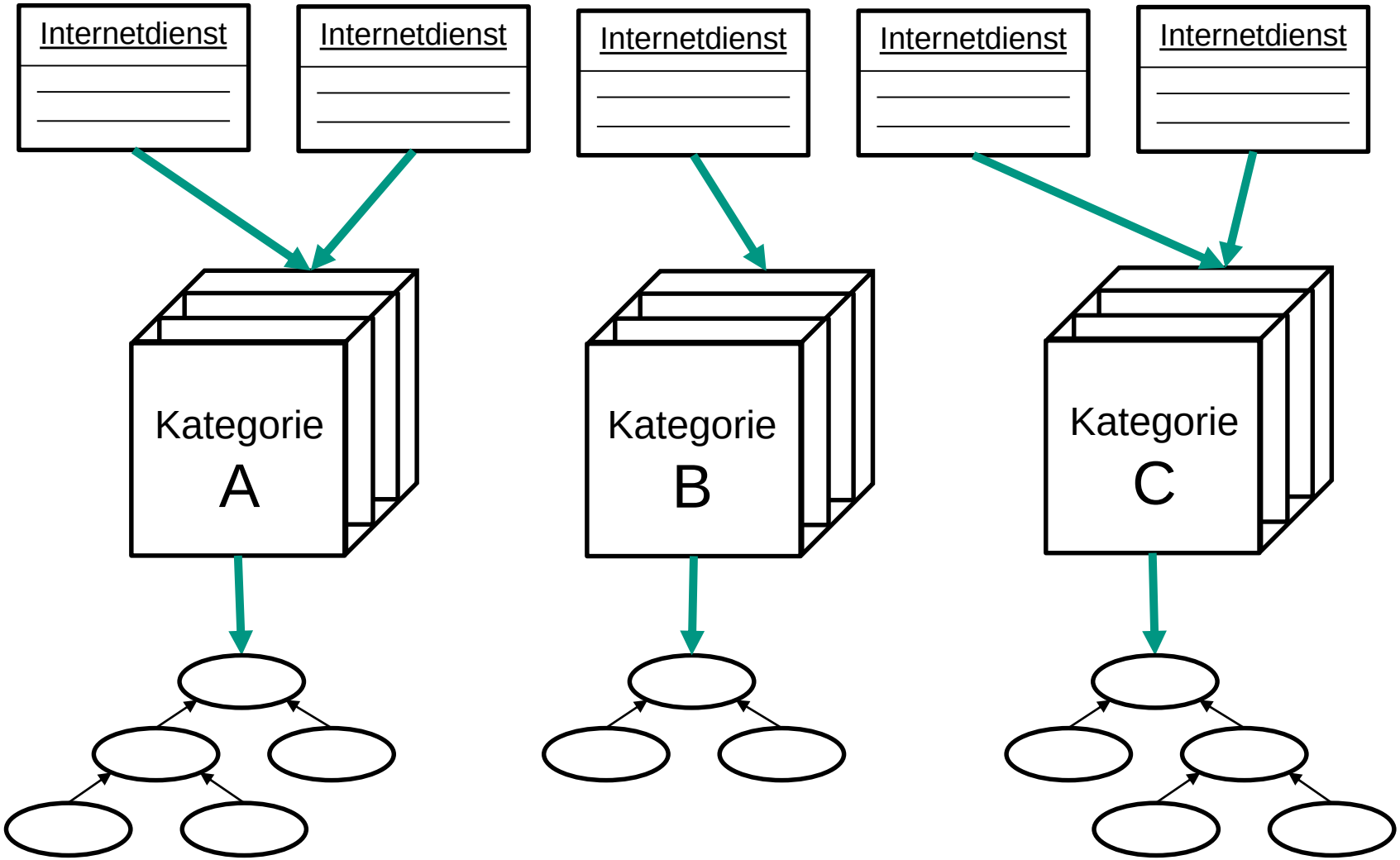

Kalender
Mo
Di
Mi
Do
Fr

Motivation

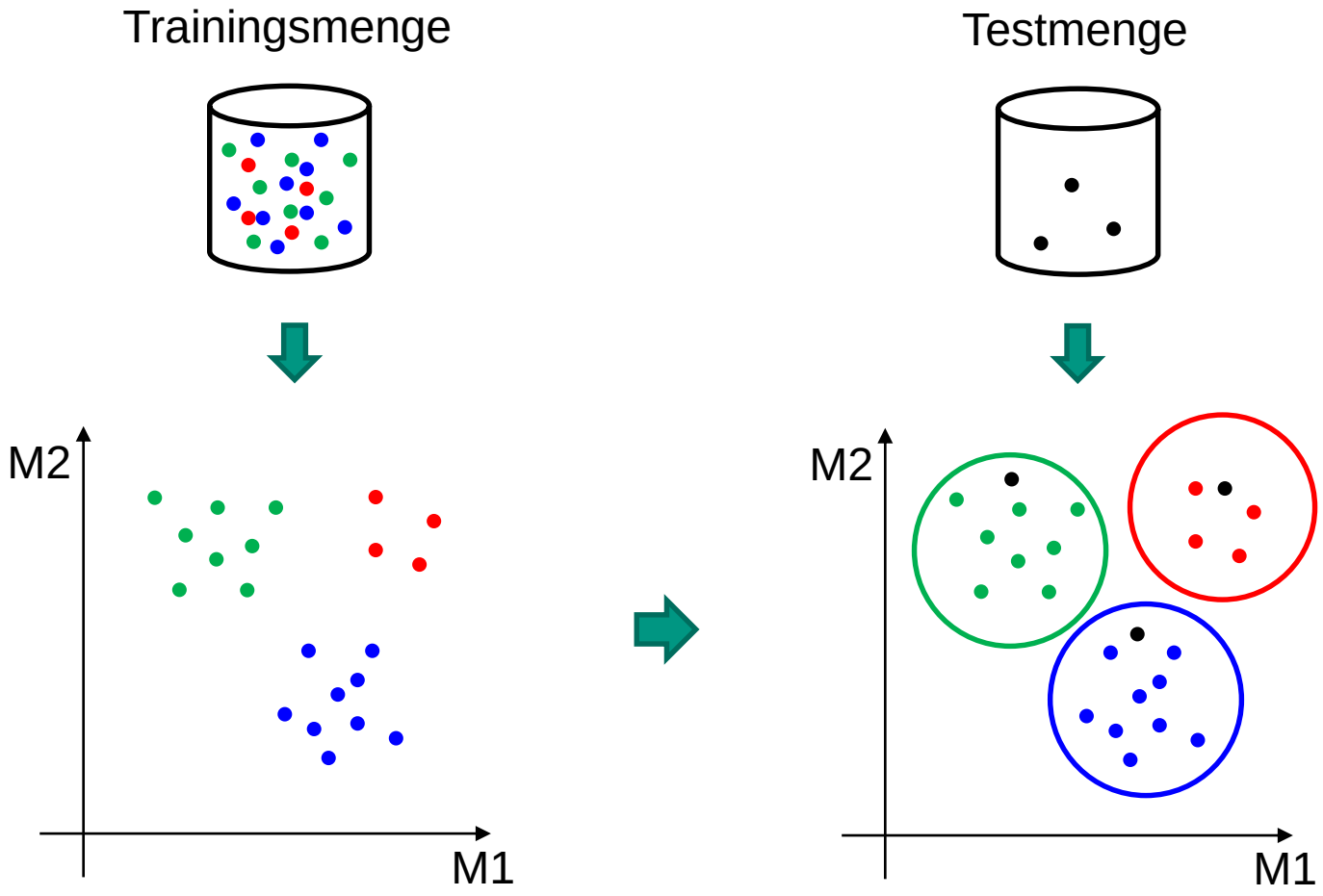


Flugauskunft

Ansatz

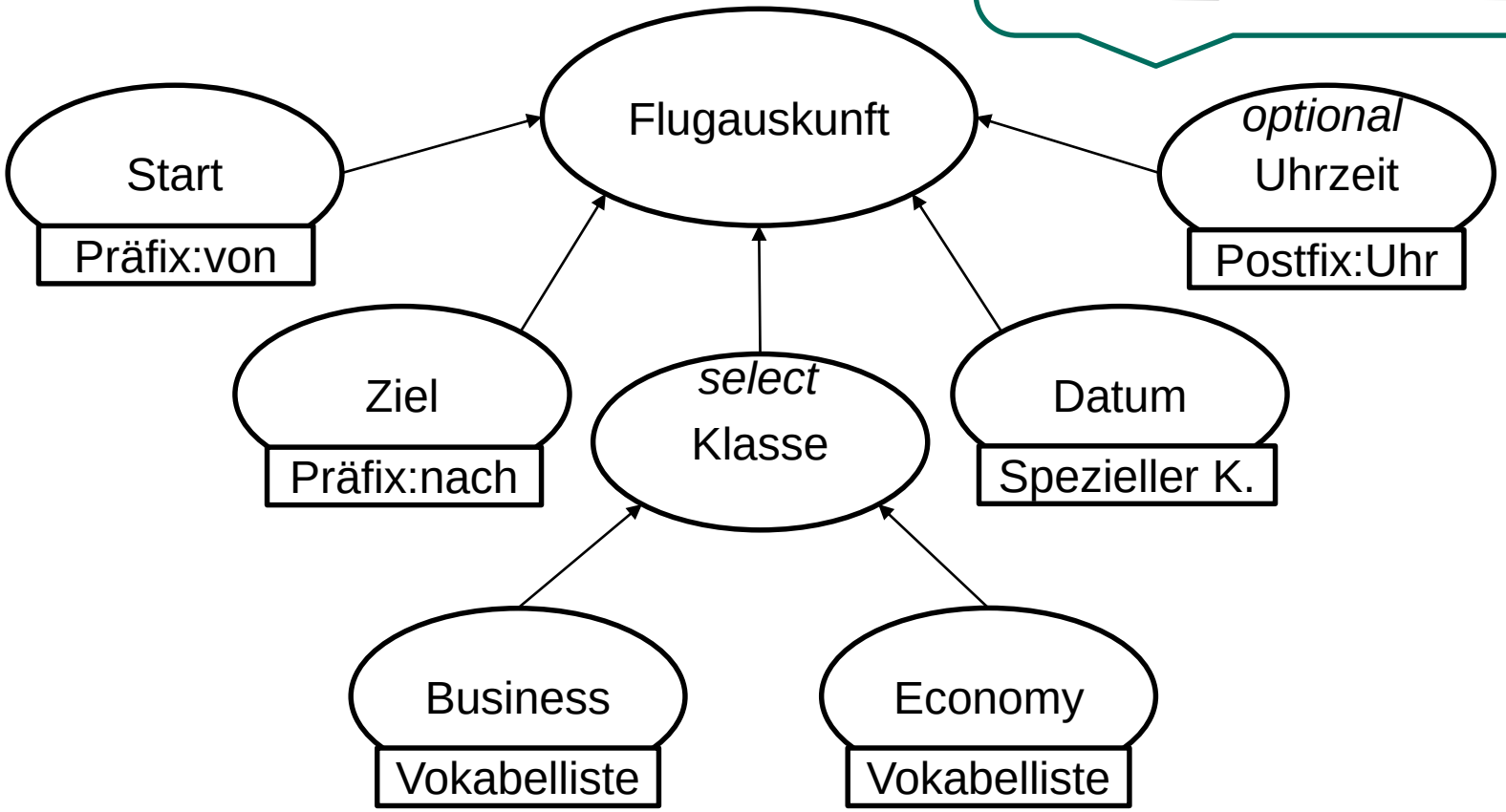

Internetdienst
Internetdienst
Internetdienst
Internetdienst
Internetdienst
Kategorie
A
Kategorie
B
Kategorie
C

Grundlagen – Klassifikation per Clustering



Grundlagen – Aktive Ontologie

„Zeig mir für heute alle Business Class Flüge von Berlin nach Hamburg“

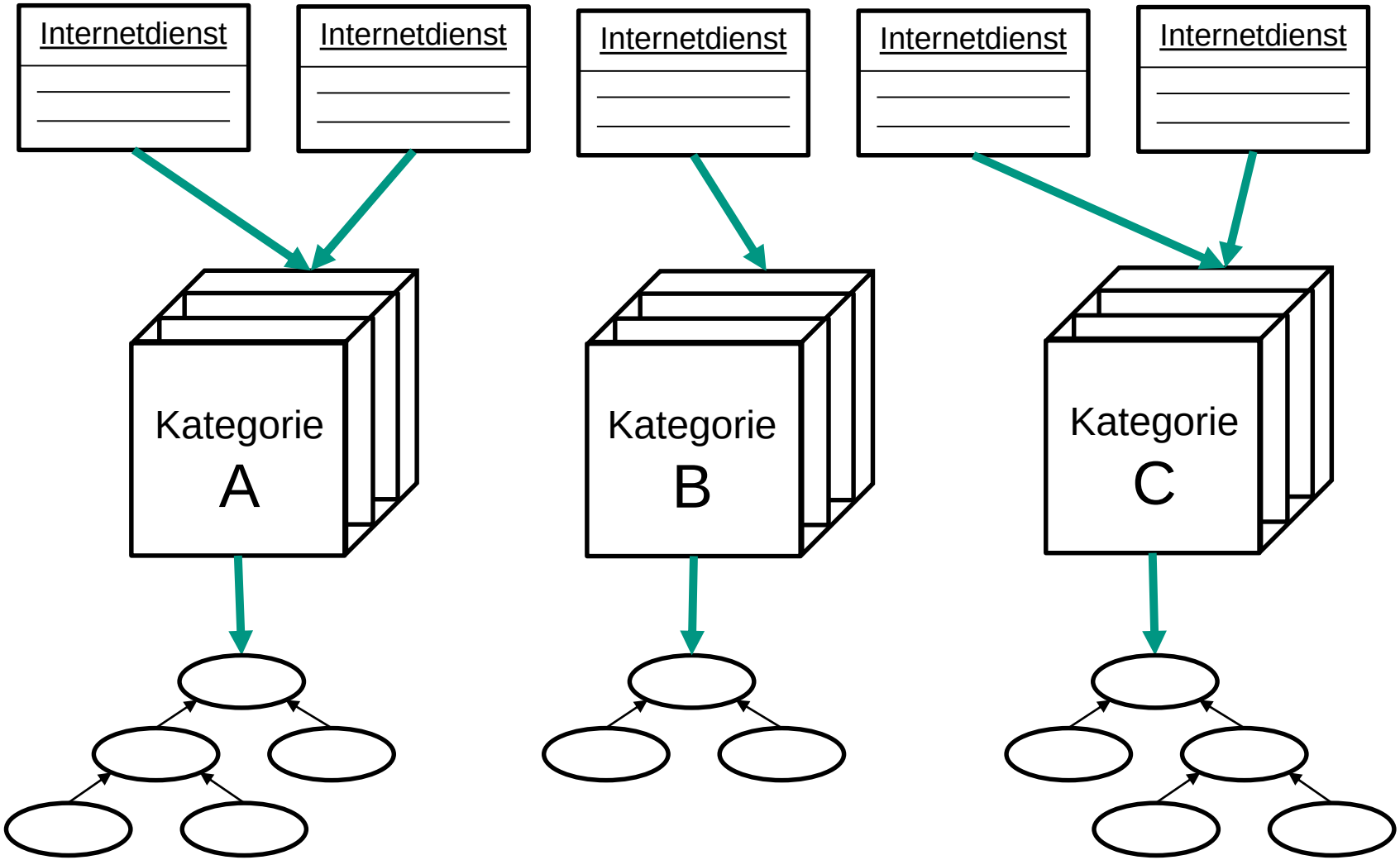


Verwandte Arbeiten

- Clustering
  - *„Web services discovery based on semantic similarity clustering“* [RD12]
  - *„Web service community discovery based on spectral clusteing“* [Zha09]

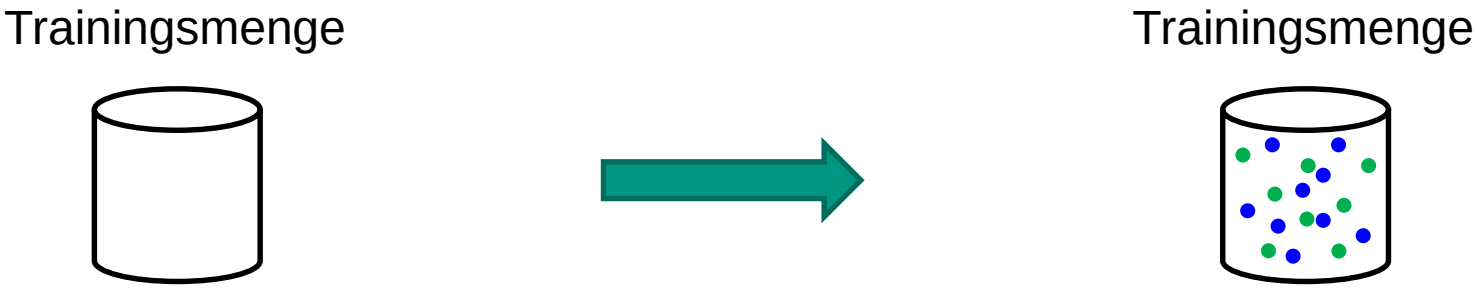
- Aktive Ontologien
  - *„OntoBuilder: Fully automatic extraction and consulidation of ontologies from web sources“* [GMJ04]
  - „Automatic ontologie matching using application semantics“ [Gal05]

Ansatz


Internetdienst
Internetdienst
Internetdienst
Internetdienst
Internetdienst
Kategorie
A
Kategorie
B
Kategorie
C

Internetdienste sammeln

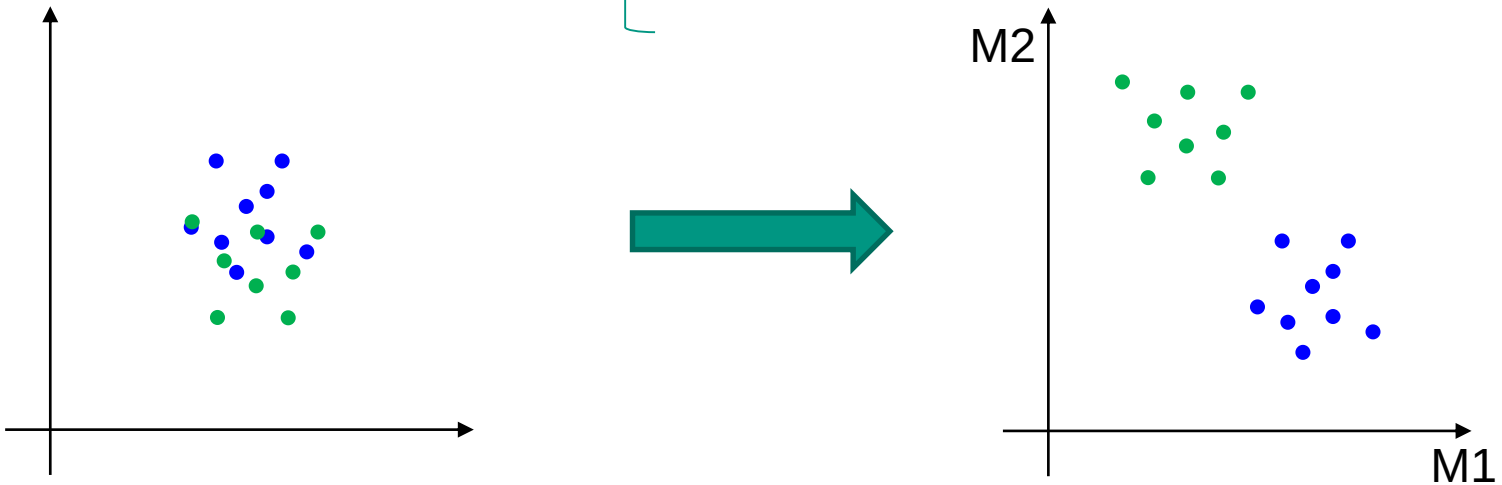
- Webcrawler
  - HTML-Formulare



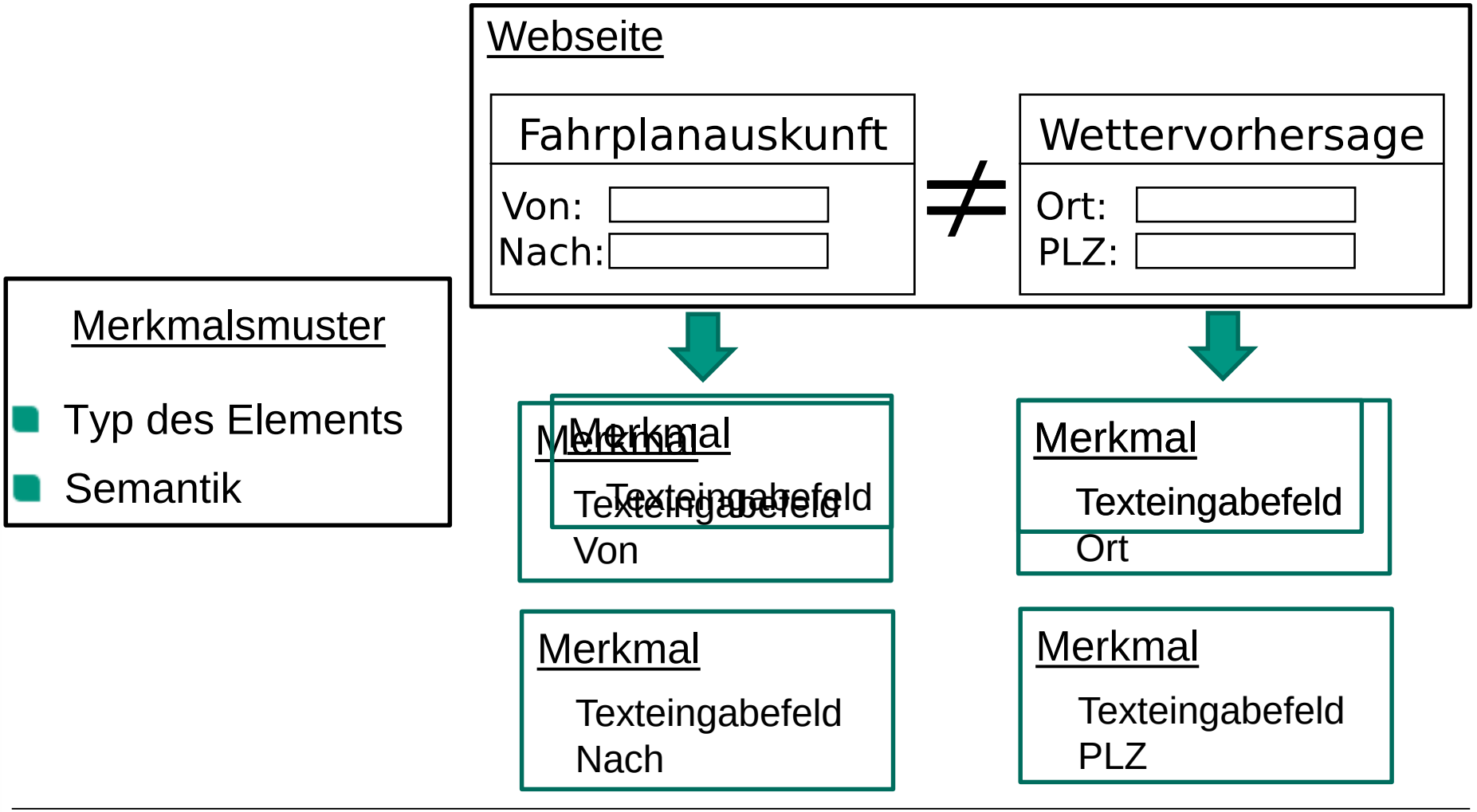
Merkmale erzeugen

- Webseitenbeschreibung
- Formularelemente
- Anzahl an Elementen

- Passworteingabe
- Einfache Auswahl
- Mehrfache Auswahl
- Einzeiliges Texteingabefeld
- Mehrzeiliges Texteingabefeld
- Spezielle Eingabefelder
- Link



Merkmalsmuster - Formularelemente



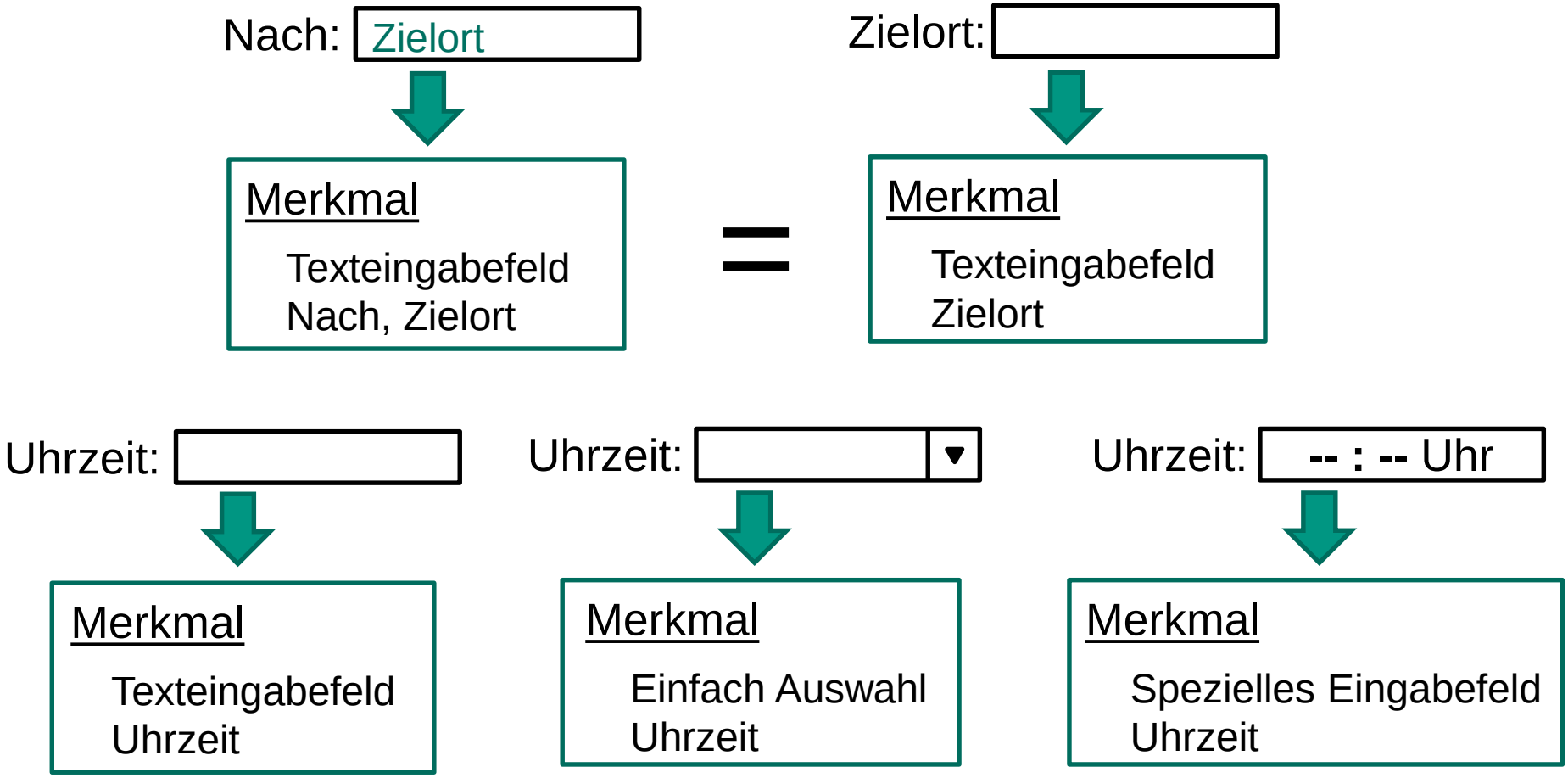
Semantik erkennen

Nach: Zielort Karlsruhe

```
<label>Nach:
 <input value=„Karlsruhe“, placeholder=„Zielort“, title=„Reiseziel“ name=„Endpunkt“>
</label>
```

- Weitere Attribute:
  - content
  - text
  - option

Merkmale vereinigen

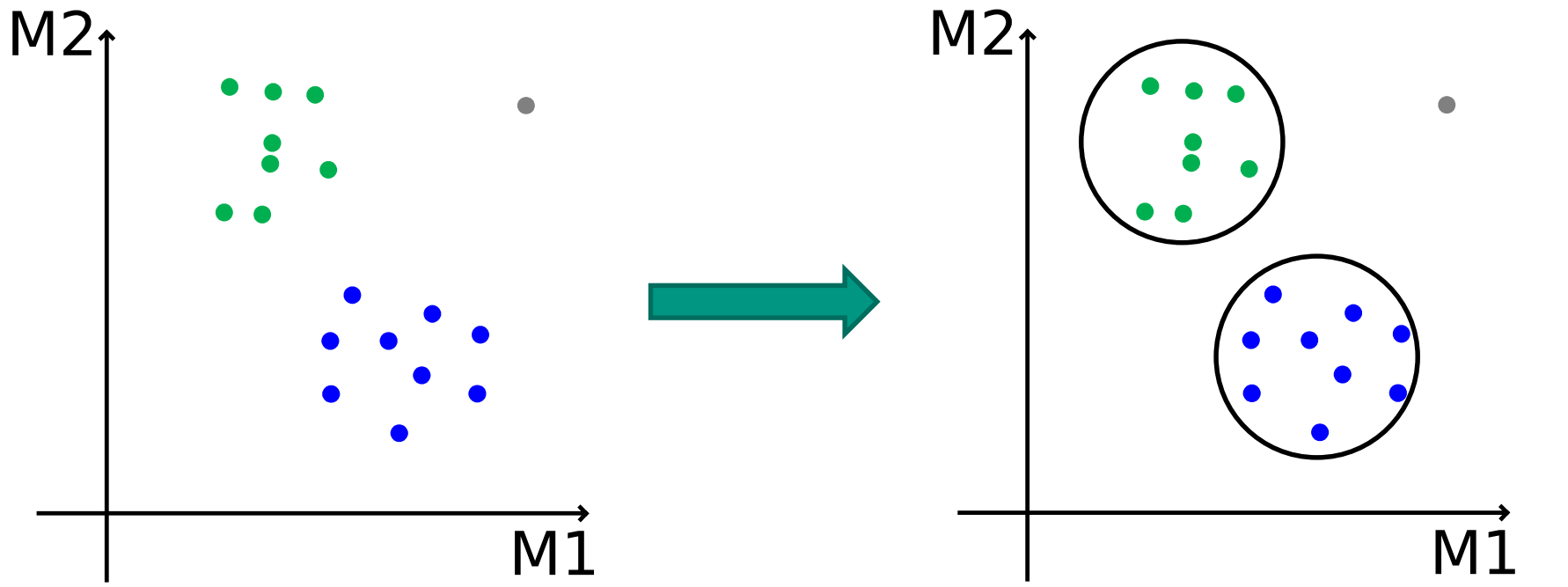


Merkmalen vereinigen

Elementtyp	Passworteingabe	Einfache Auswahl	Mehrfache Auswahl	Einzeiliges Texteingabefeld	Mehrzeiliges Texteingabefeld	Spezieller Eingabefelder	Link
Passworteingabe	x						
Einfache Auswahl		x		x			
Mehrfache Auswahl			x				
Einzeiliges Texteingabefeld				x			
Mehrzeiliges Texteingabefeld					x		
Spezielle Eingabefelder		x		x		x	
Link							x

Clustering

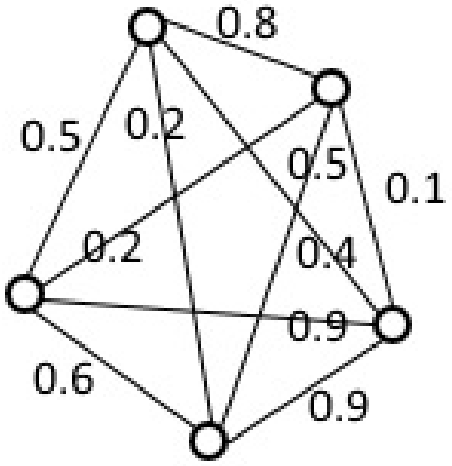
- Spectral Clustering
- Dichtebasierter Spatial Clustering (DBScan)



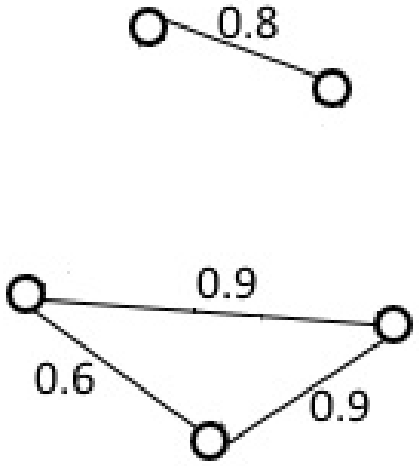
Internetdienste Merkmale Clustering Klassifikation

Clustering - Spectral Clustering

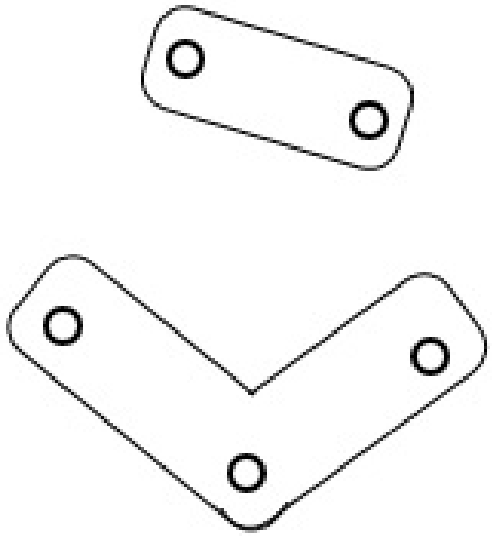
- Ähnlichkeitsfunktion
- Parameter k



Voll verbundener Graph



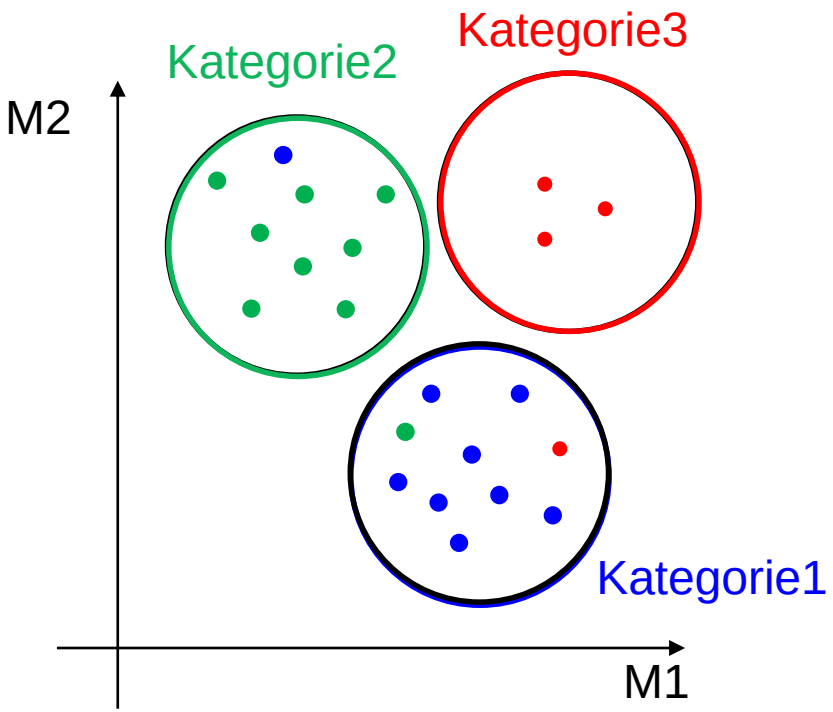
Kanten entfernen
(k = 0,55)



Clustering

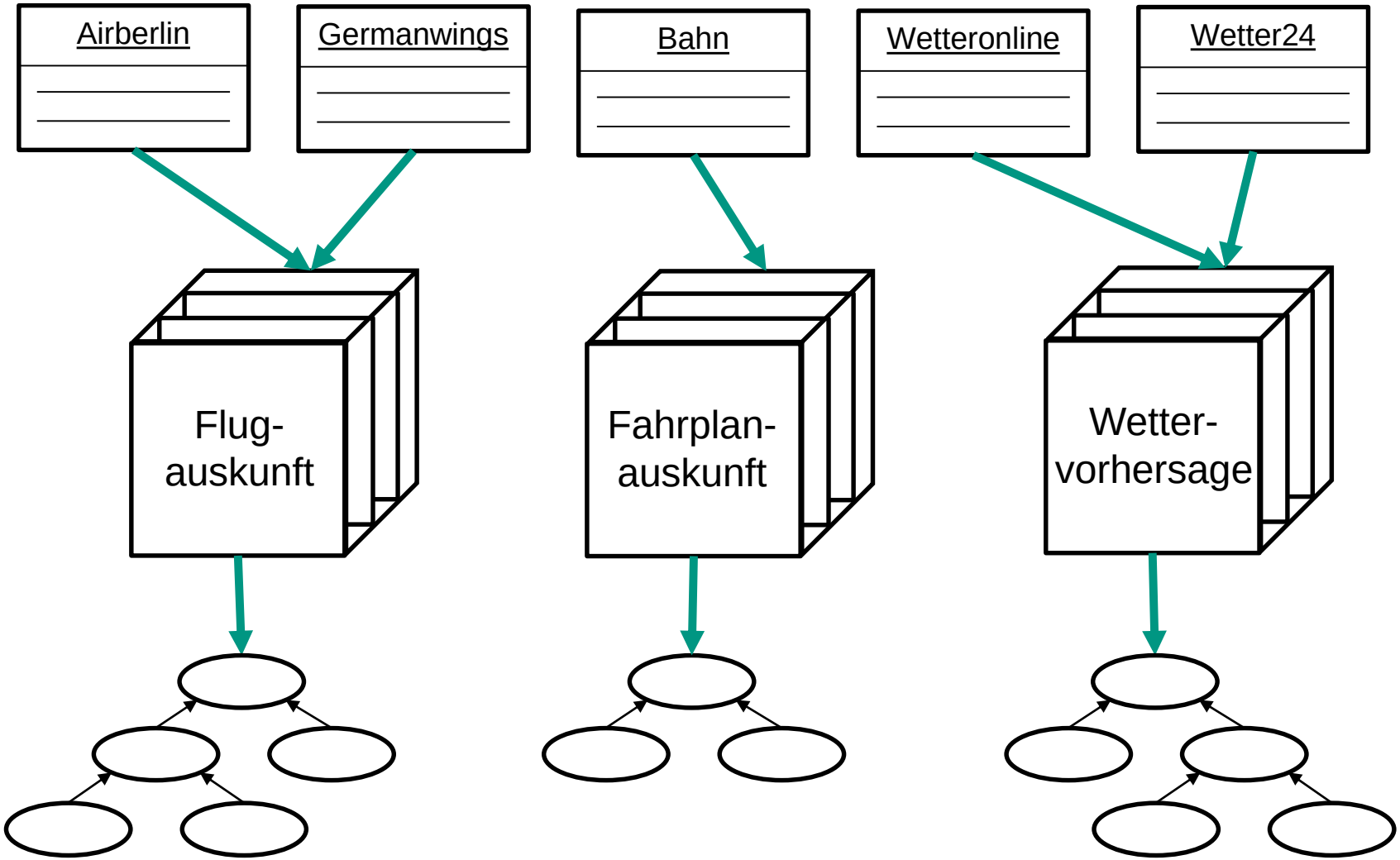
Clusterzuordnung

- Clusterzuordnung
  - Häufigkeit:



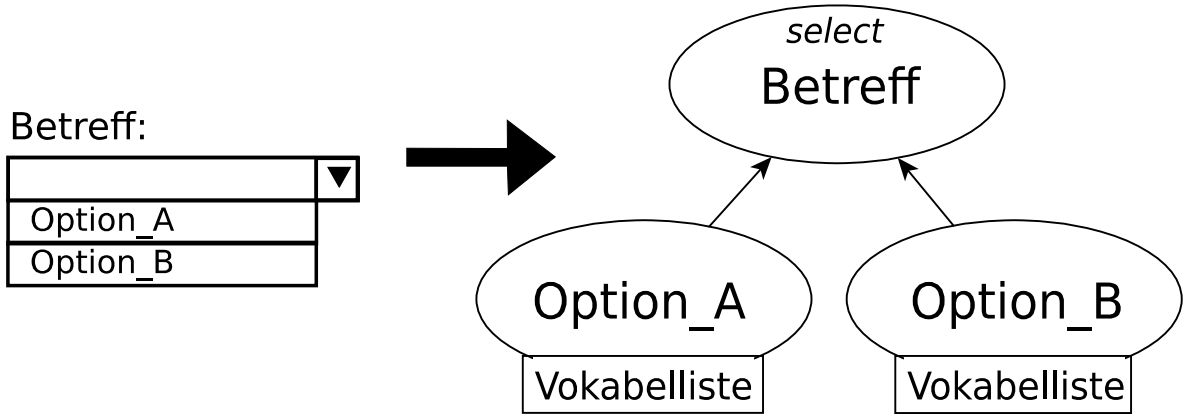
Internetdienstkategorie
Login
Registrierung
Fahrplanauskunft
Flugauskunft
Autovermietung
Unterkunftssuche
Newsletterabonnierung
Wettervorhersage

Ansatz

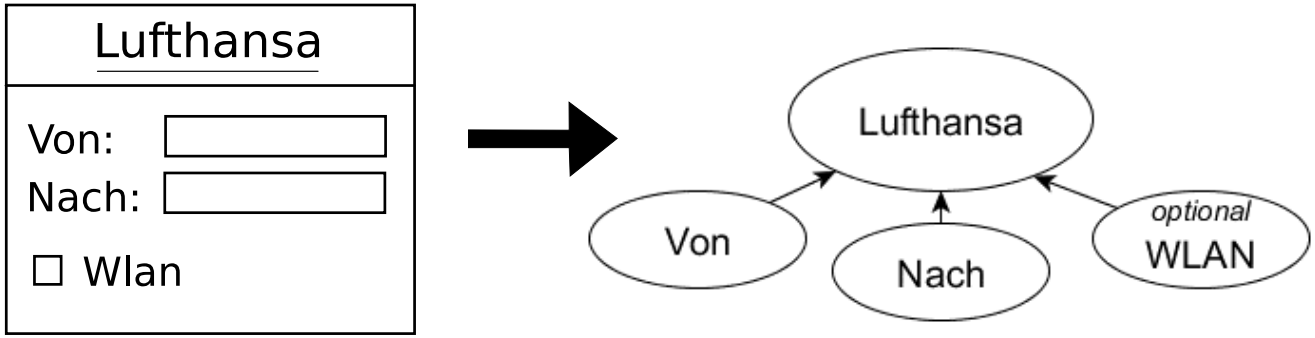

Airberlin
Germanwings
Bahn
Wetteronline
Wetter24
Flug-
auskunft
Fahrplan-
auskunft
Wetter-
vorhersage

Konstruktionsplan

- Formularelement ⟶ Aktive Ontologie

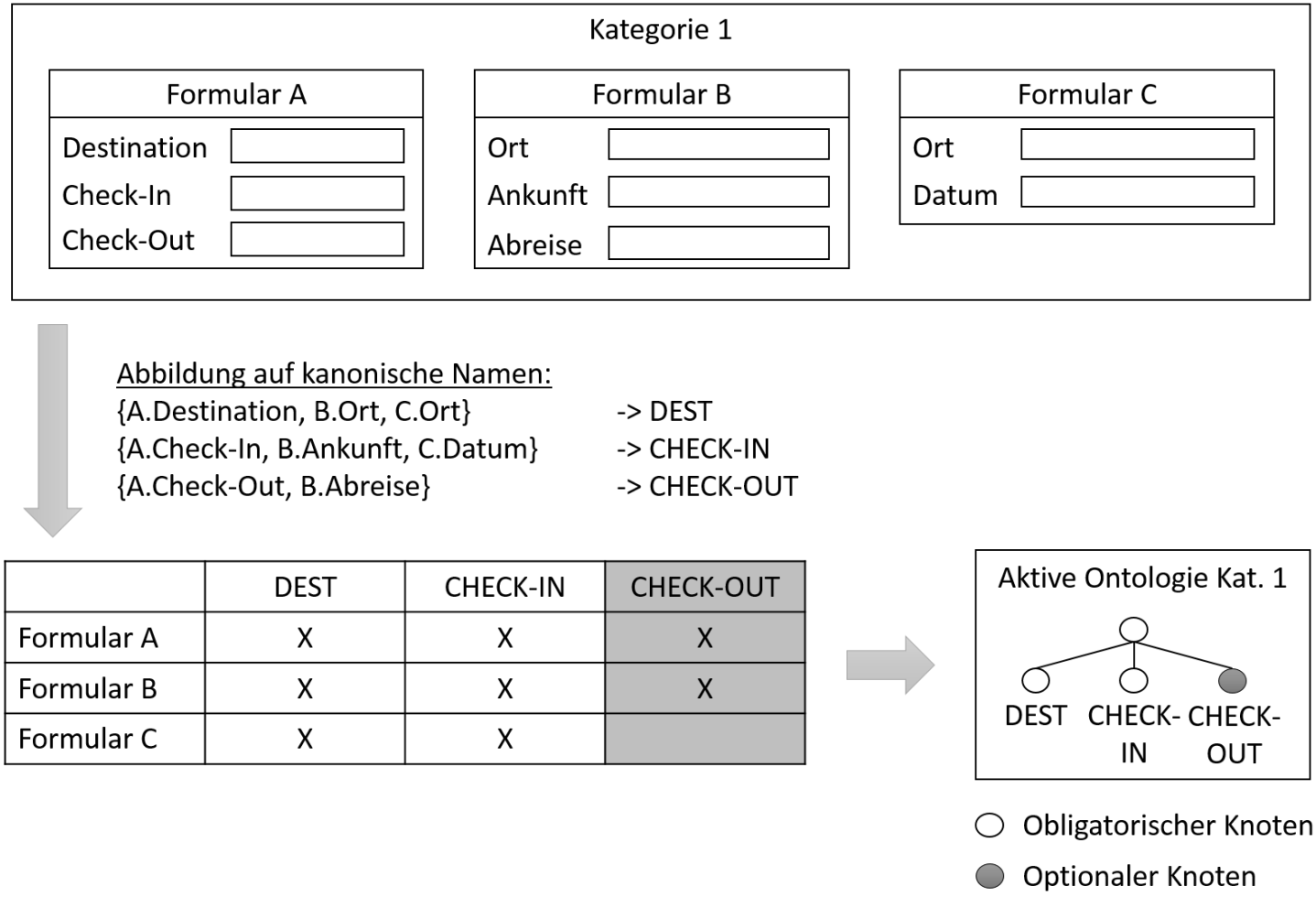


- Internetdienst ⟶ Aktive Ontologie



Konstruktionsplan

- Kategorie ⟶ Aktive Ontologie

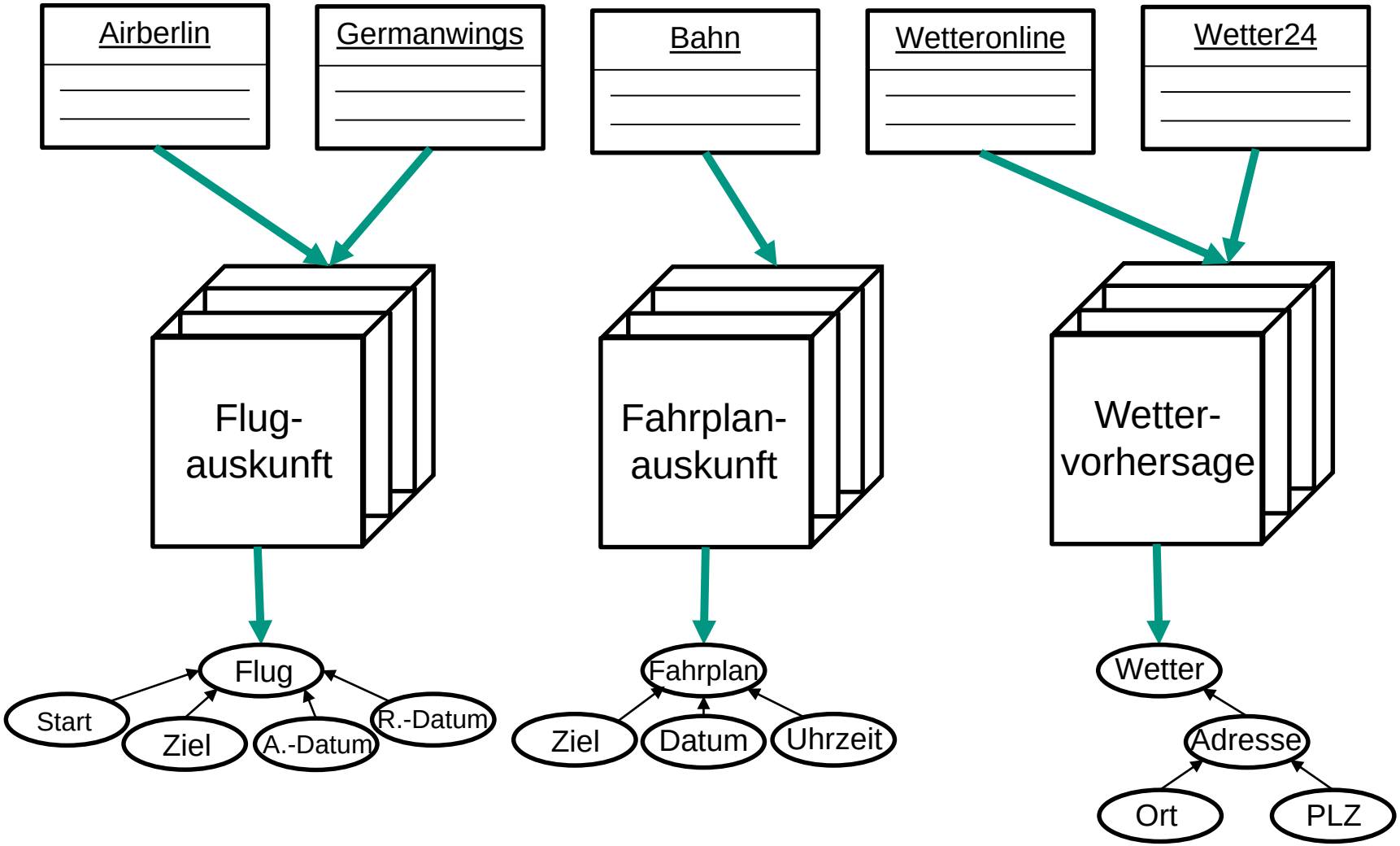


	DEST	CHECK-IN	CHECK-OUT
Formular A	X	X	X
Formular B	X	X	X
Formular C	X	X	

Flugauskunft

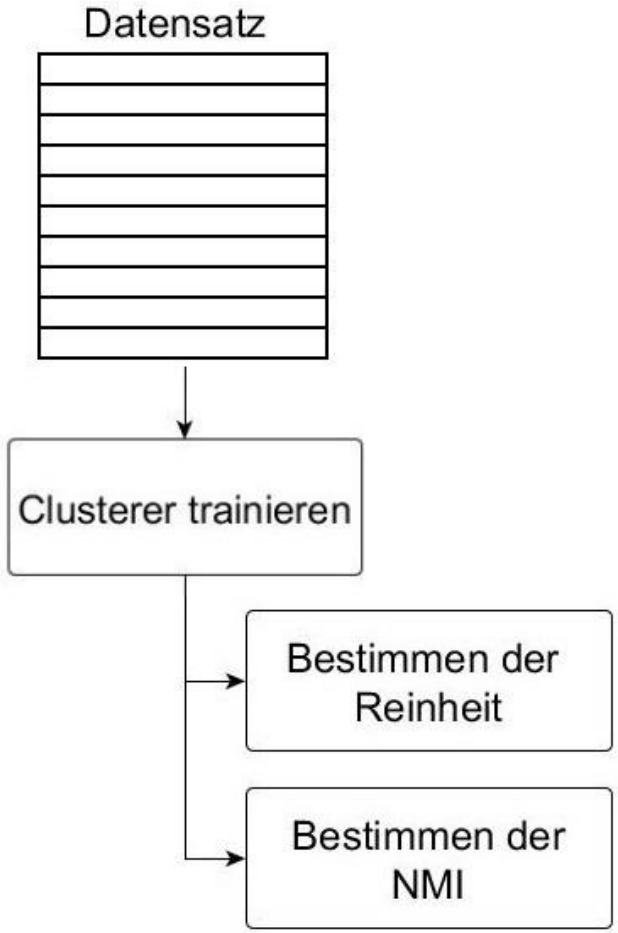
Merkmal:	Dienstanbieter:	Travelvision	Condor	Britishairways	Germanwings	Airberlin	Wow-Air	Ryanair	Fly
Merkmalsmuster	**Semantik**								
Einzeiliges Texteingabefeld	Reisestart	X	X	X	X	X	X	X	X
Einzeiliges Texteingabefeld	Reiseziel	X	X	X	X	X	X	X	X
Spezielles Eingabefeld	Abreisedatum	X	X	X	X	X	X	X	X
Spezielles Eingabefeld	Rückreisedatum	X	X	X	X	X	X	X	X
Spezielles Eingabefeld	Abreisezeit	X							
Spezielles Eingabefeld	Rückreisezeit	X							
Einfache Auswahl	Nur Hinflug		X	X	X	X	X	X	X
…	…								

Ansatz

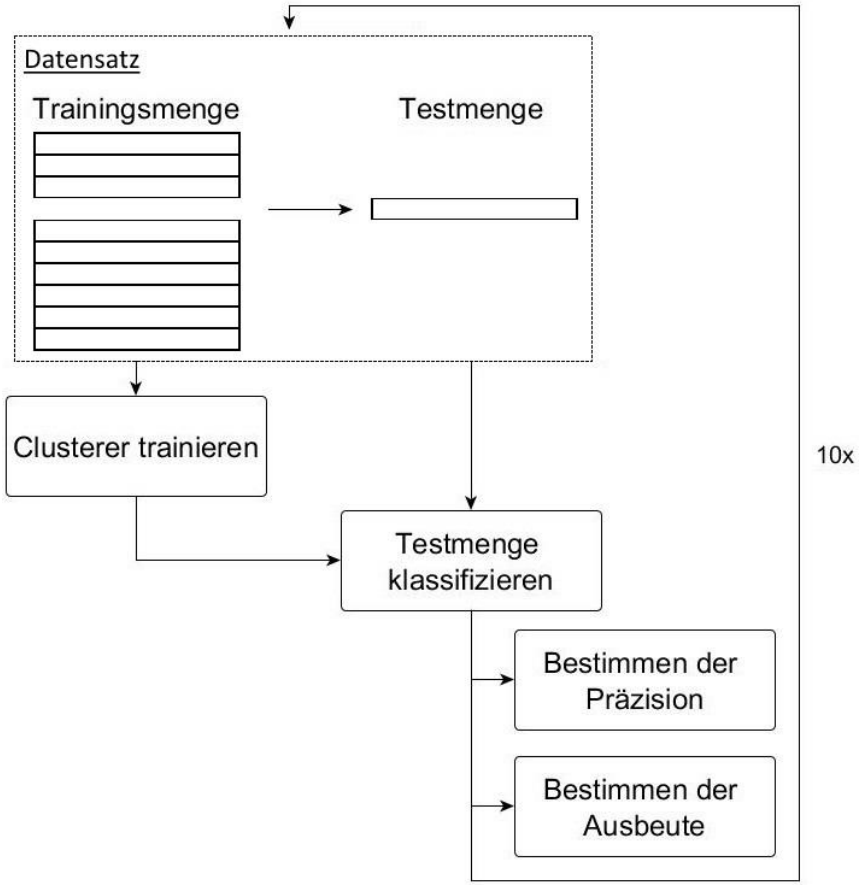

Airberlin
Germanwings
Bahn
Wetteronline
Wetter24
Flug-
auskunft
Fahrplan-
auskunft
Wetter-
vorhersage
Flug
Start
Ziel
A.-Datum
R.-Datum
Fahrplan
Ziel
Datum
Uhrzeit
Wetter
Adresse
Ort
PLZ

Evaluierungsaufbau

1. Evaluation der Clusterbildung



2. 10-fache Kreuzvalidierung

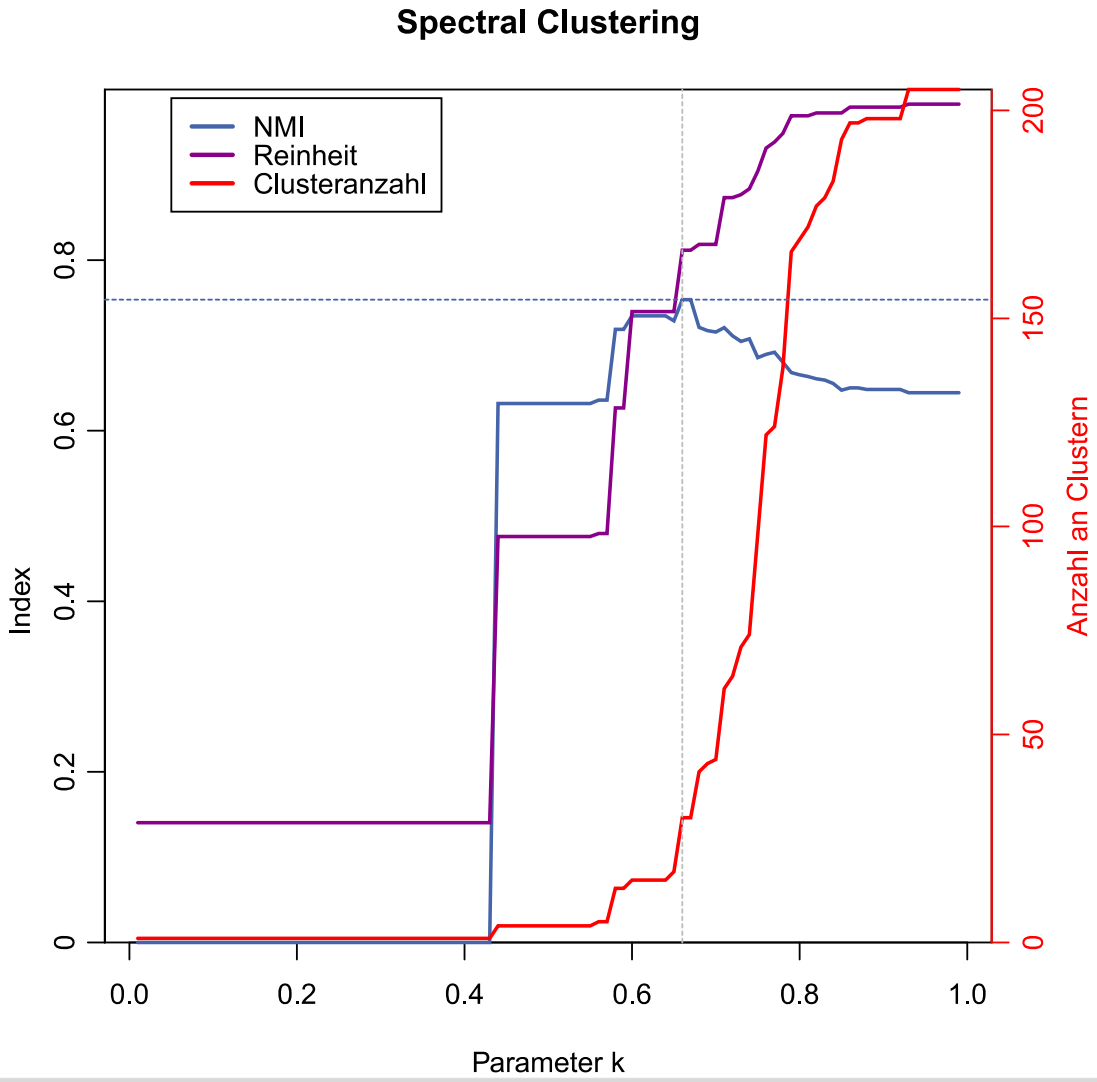


Datensatz

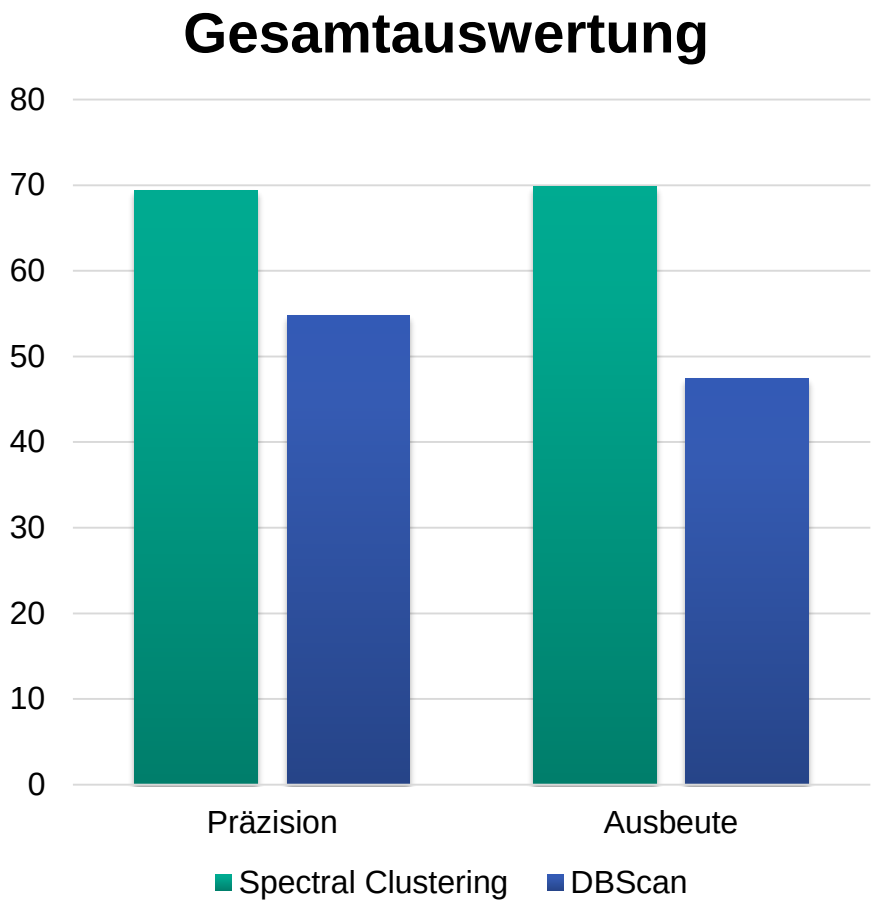
Internetdienstkategorien	Anzahl
Login	36
Registrierung	12
Fahrplanauskunft	41
Flugauskunft	23
Autovermietung	19
Unterkunftssuche	41
Kontaktformular	39
Newsletterabonnierung	24
Wettervorhersage	25
Gesamt	**260**

Internetdienstkategorien	Anzahl
Sprachauswahl	13
Mehrfache Auswahl	3
Veranstaltungsfilter	1
Parkplatzverfügbarkeit	1
Buchung verwalten	2
Filterung	4
Ticketrechner	3
Feedbackformular	1
Suche	4
Gesamt	**32**

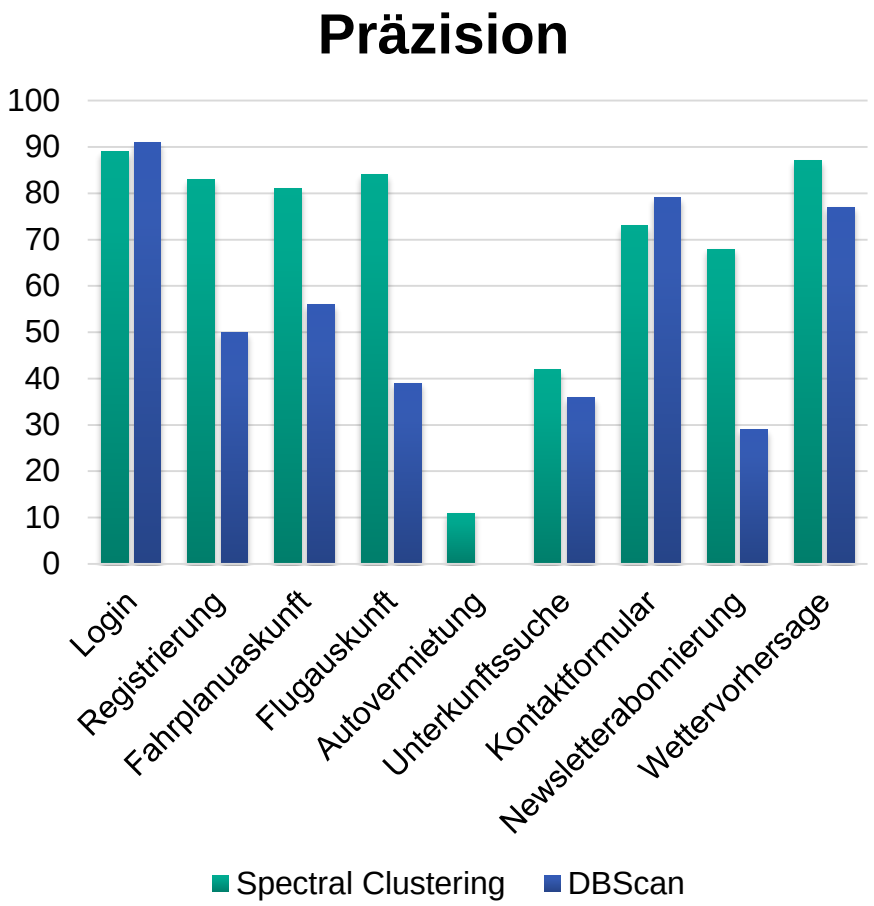
Evaluation des Clusteringalgorithmus

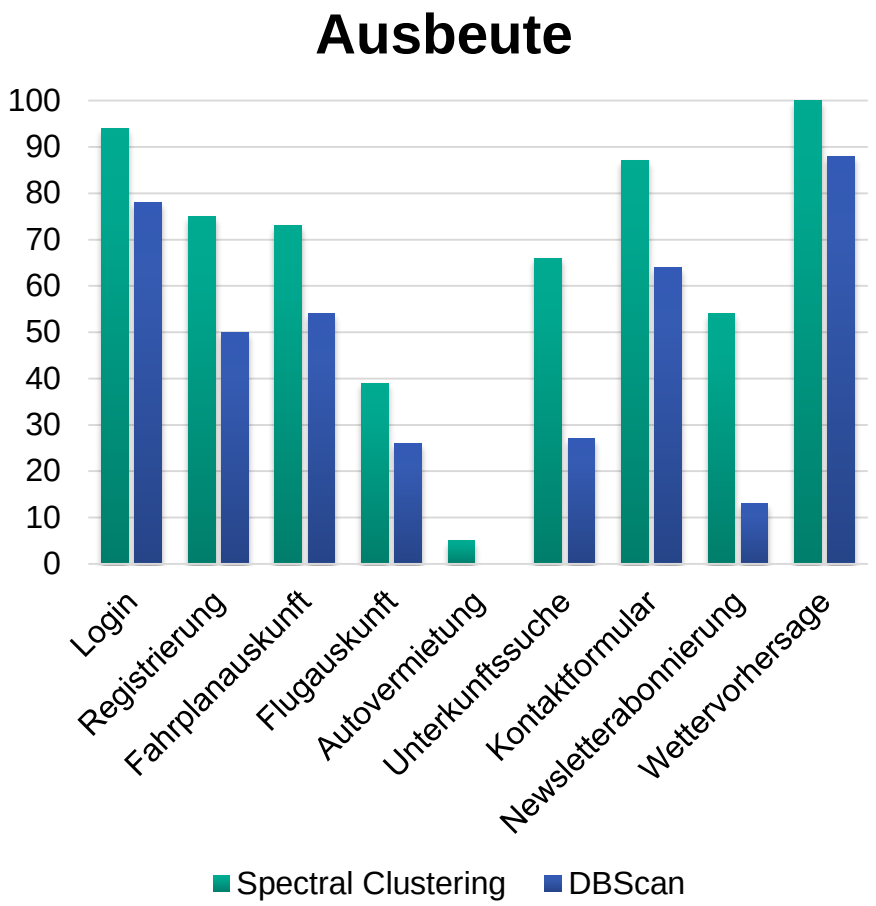


Evaluation des Klassifikators



Evaluation des Klassifikators





Fazit

- Klassifikation
  - DBScan -
  - Spectral Clustering +
- Konstruktionsplan
  - Vorschrift +
  - Manuell -
- Ausblick
  - Konstruktionsplan automatisieren
  - Semantik Web

Karlsruher Institut für Technologie

Danke
für ihre
Aufmerksamkeit

Literatur

[Gal05] Avigdor Gal u. a., “Automatic ontology matching using application semantics“, AI magazine 26.1, 2005.

[GMJ04] Avigdor Gal, Giovanni Modica und Hasan Jamil, „Ontobuilder: Fully automatic extraction and consolidation of ontologies from web sources“, Engineering, 2004, Proceedings. 20th International Conference on. IEEE, 2004.

[Guz08] Didier Guzzoni. „Active: a unified platform for building intelligent applications“, Diss. Ecole Polytechnique Federale De Lausanne, 28. Jan. 2008.

[RD12] P Ravinder Reddy und A Damodaram, “Web services discovery based on semantic similarity clustering“, Engineering (CONSEG), CSI Sixth International Conference on. IEEE, 2012.

[Zha09] Xizhe Zhang u. a., “Web service community discovery based on spectrum clustering“, Computational Intelligence and Security, 2009. CIS'09. International Conference on. Bd. 2. IEEE. 2009.

Evaluation des Klassifikators

Karlsruher Institut für Technologie

Spectral Clustering

Internetdienstkategorie	Anzahl	Präzision	Ausbeute
Login	36	89%	94%
Registrierung	12	83%	75%
Fahrplanauskunft	41	81%	73%
Flugauskunft	23	84%	39%
Autovermietung	19	11%	5%
Unterkunftssuche	41	42%	66%
Kontaktformular	39	73%	87%
Newsletterabonnierung	24	68%	54%
Wettervorhersage	25	87%	100%
Gesamtauswertung	**260**	**69,38%**	**69,86%**

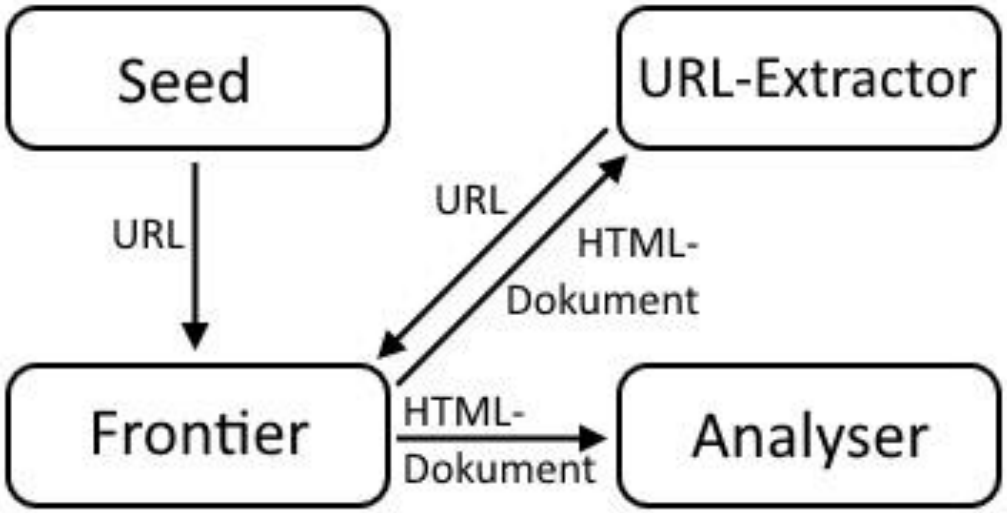
Evaluation des Klassifikators

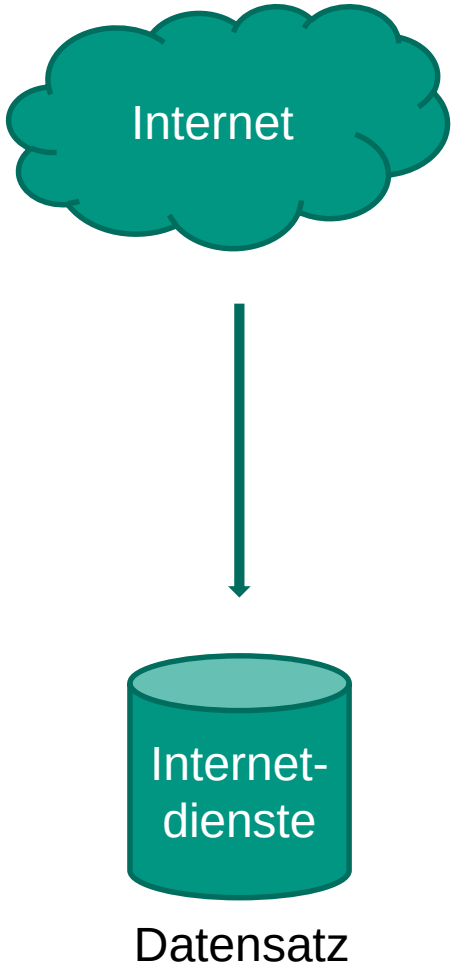
Spectral Clustering

Internetdienstkategorie	Anzahl	Präzision	Ausbeute
Sprachauswahl	13	58%	62%
Mehrfache Auswahl	3	0%	0%
Buchung verwalten	2	0%	0%
Filterung	4	25%	25%
Ticketrechner	3	0%	0%
Suche	4	0%	0%
Unbekannt	3	0%	0%
Gesamtauswertung	**32**	**26,69%**	**28,31%**

Internetdienste sammeln - Webcrawler

- Programmiersprache: Python
- HTTP- & HTTPS Protokolle





Internetdienste sammeln – Webcrawler

- Problem: zyklische Verlinkung
  - Lösung: Host- und URL-gesehen-Test
- Host- und URL-gesehen-Test:
  - Problem: Speicherkapazität
  - Annahme: Ein Internetdienst ist von der Startseite aus, nach höchstens einer Verlinkung erreichbar.