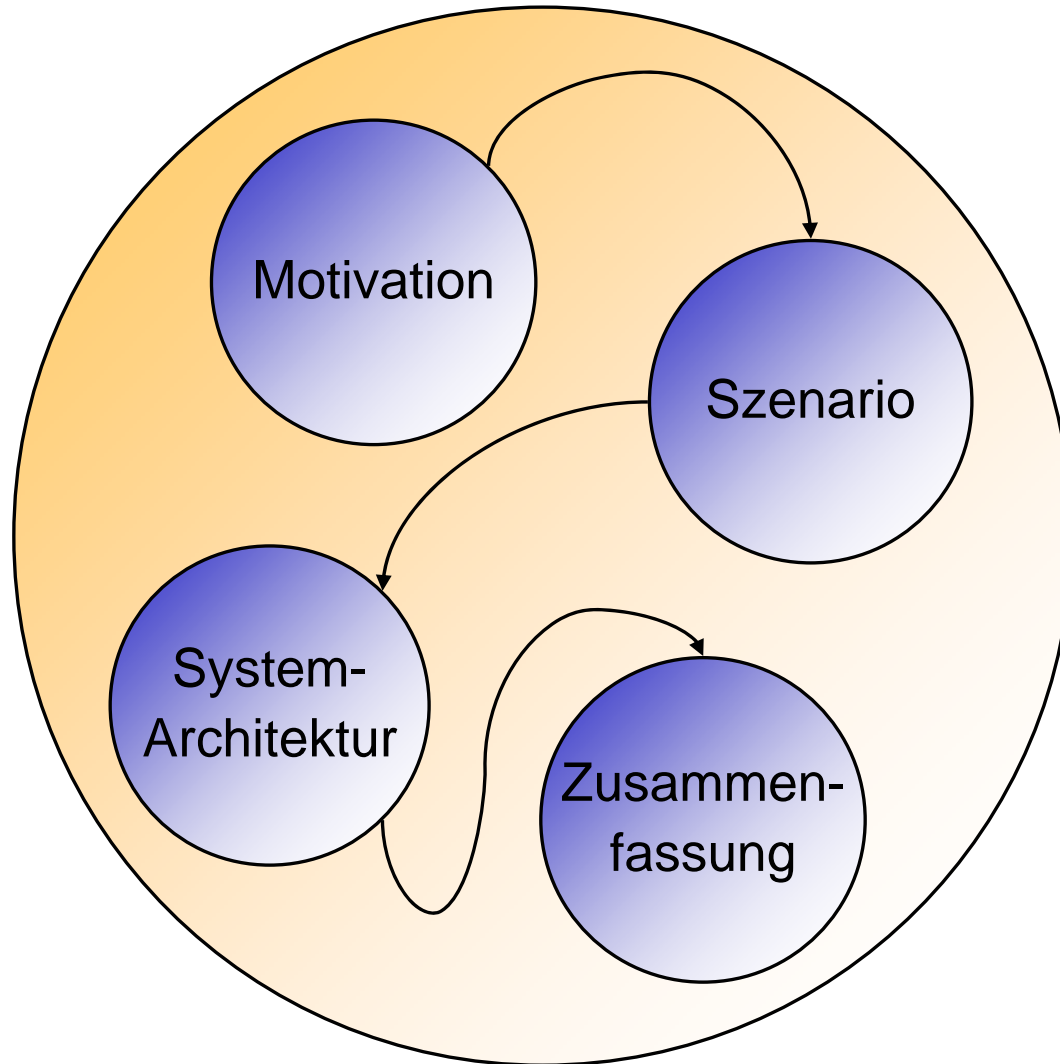


# Intelligente Suche in dynamischen Datenströmen

Dipl.-Inf. Thomas Hornung  
Prof. Dr. Georg Lausen



- Ein Großteil der Information im Internet wird dynamisch, z.B. bezüglich spezifischer Anfragen oder aktuellen Tagesgeschäften erzeugt und ist somit nicht von Suchmaschinen indexierbar.
- Das vorgestellte OntoGather-System ermöglicht eine vollautomatische, konsistente Verarbeitung von unstrukturierten Informationen aus web-basierten Applikationen.
- Die Hauptvorteile der Technologie gegenüber bestehenden Ansätzen liegen in der:
  - einfachen Handhabung
  - höheren Effizienz
  - deutlich geringeren Wartung

# Beispiel-Szenario: Semantische Integration von Web-Quellen

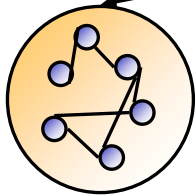
„digital camera“, price < 200 €, brand = „Canon“

The screenshot shows the eBay search results for 'digital camera'. The search filters are set to 'Canon' for the brand and '0' to '200' for the price range. The results list four Canon PowerShot models: A530, A430, A540, and 540. The prices are \$124.99, \$152.14, \$169.99, and \$178.10 respectively. The search results are displayed in a table format with columns for Manufacturer, Megapixels, Optical Zoom, Included Memory, Memory Type, Connectors, and Price.

Manufacturer	Megapixels	Optical Zoom	Included Memory	Memory Type	Connectors	Price
Canon	5	4.0x	16MB	Secure Digital	USB, Video-out	\$124.99
Canon	3.9	4.0x	16MB	Secure Digital	USB, Video-out	\$152.14
Canon	5.9	4.0x	16MB	Secure Digital	USB, Video-out	\$169.99
Canon	3.9	3.0x	16MB	CompactFlash	USB, Video-out	\$178.10

# Beispiel-Szenario: Semantische Integration von Web-Quellen

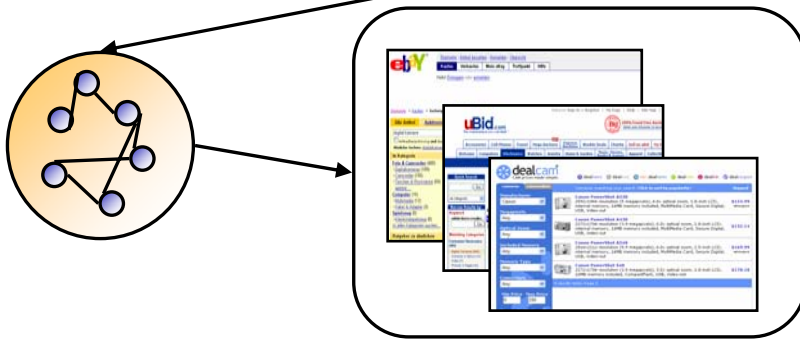
„digital camera“, price < 200 €, brand = „Canon“



Ressourcen-Auswahl mit  
Hilfe von Domänen-Ontologie

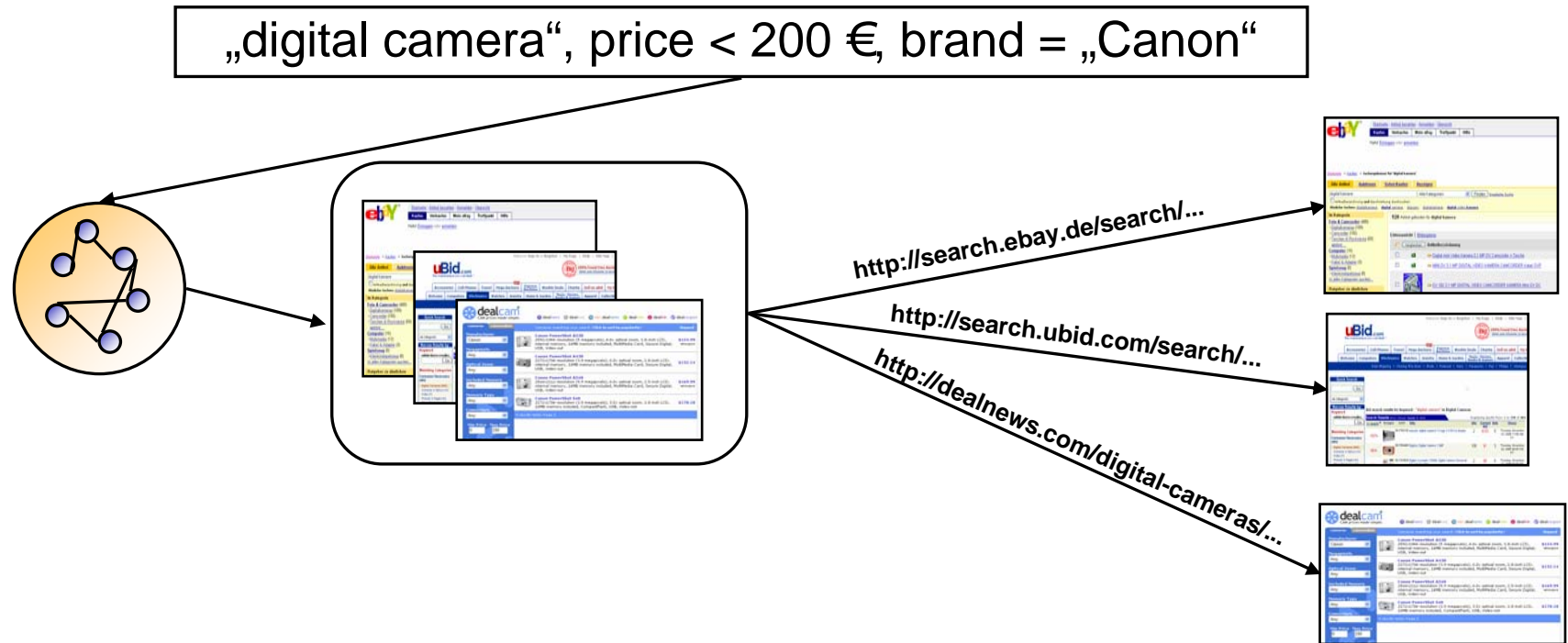
# Beispiel-Szenario: Semantische Integration von Web-Quellen

„digital camera“, price < 200 €, brand = „Canon“



Ressourcen-Auswahl mit  
Hilfe von Domänen-Ontologie

# Beispiel-Szenario: Semantische Integration von Web-Quellen



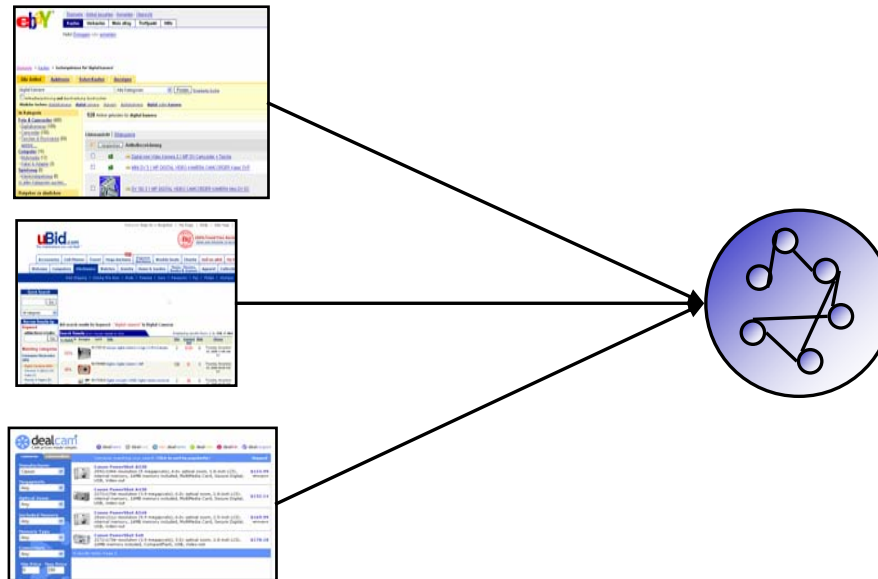
Ressourcen-Auswahl mit  
Hilfe von Domänen-Ontologie



Parallele Anfrage  
an Ressourcen

# Beispiel-Szenario: Semantische Integration von Web-Quellen

„digital camera“, price < 200 €, brand = „Canon“



Parallele Anfrage  
an Ressourcen

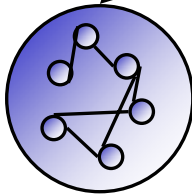


Integration der Resultate  
In Domänen-Ontologie

# Beispiel-Szenario: Semantische Integration von Web-Quellen

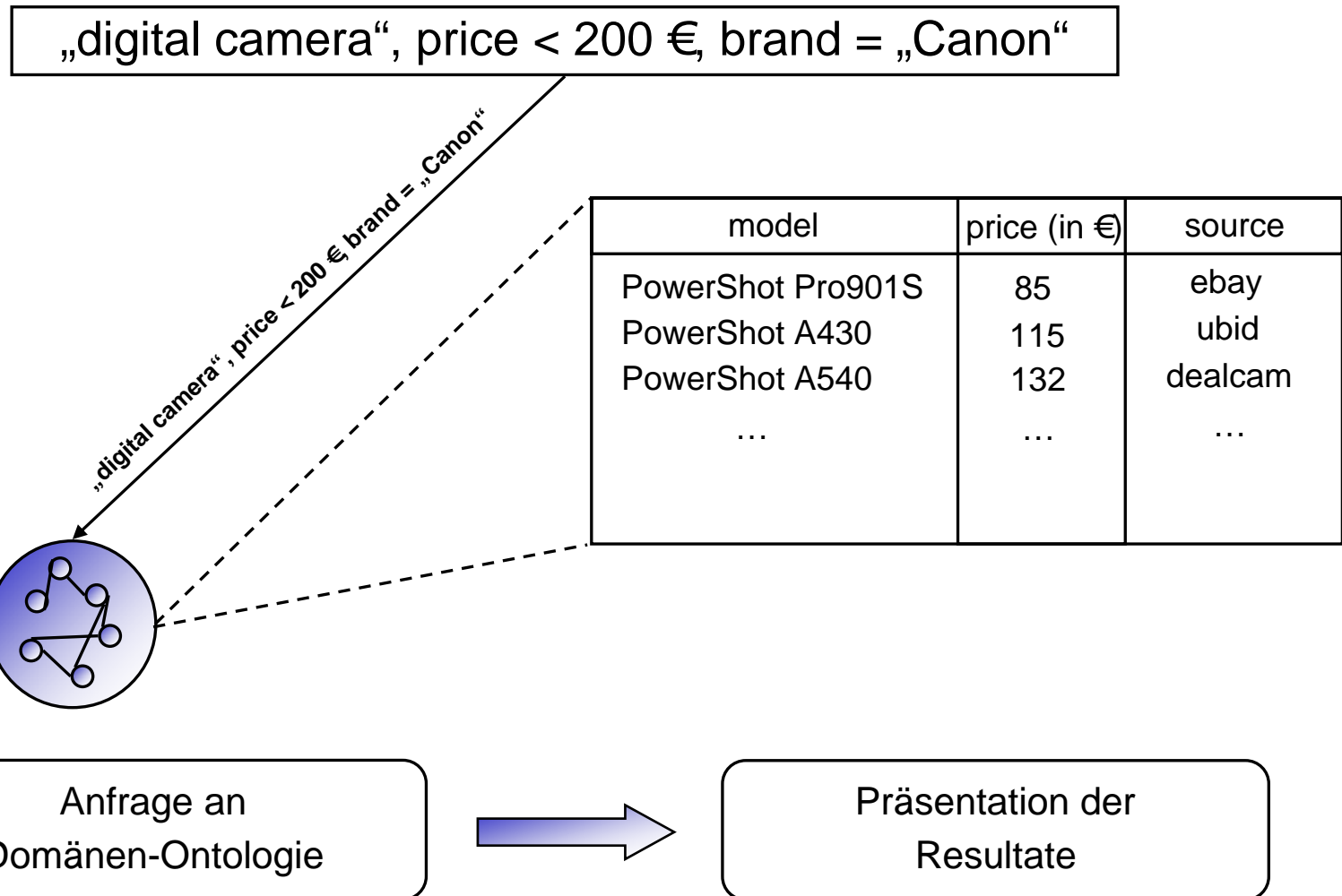
„digital camera“, price < 200 €, brand = „Canon“

„digital camera“, price < 200 €, brand = „Canon“

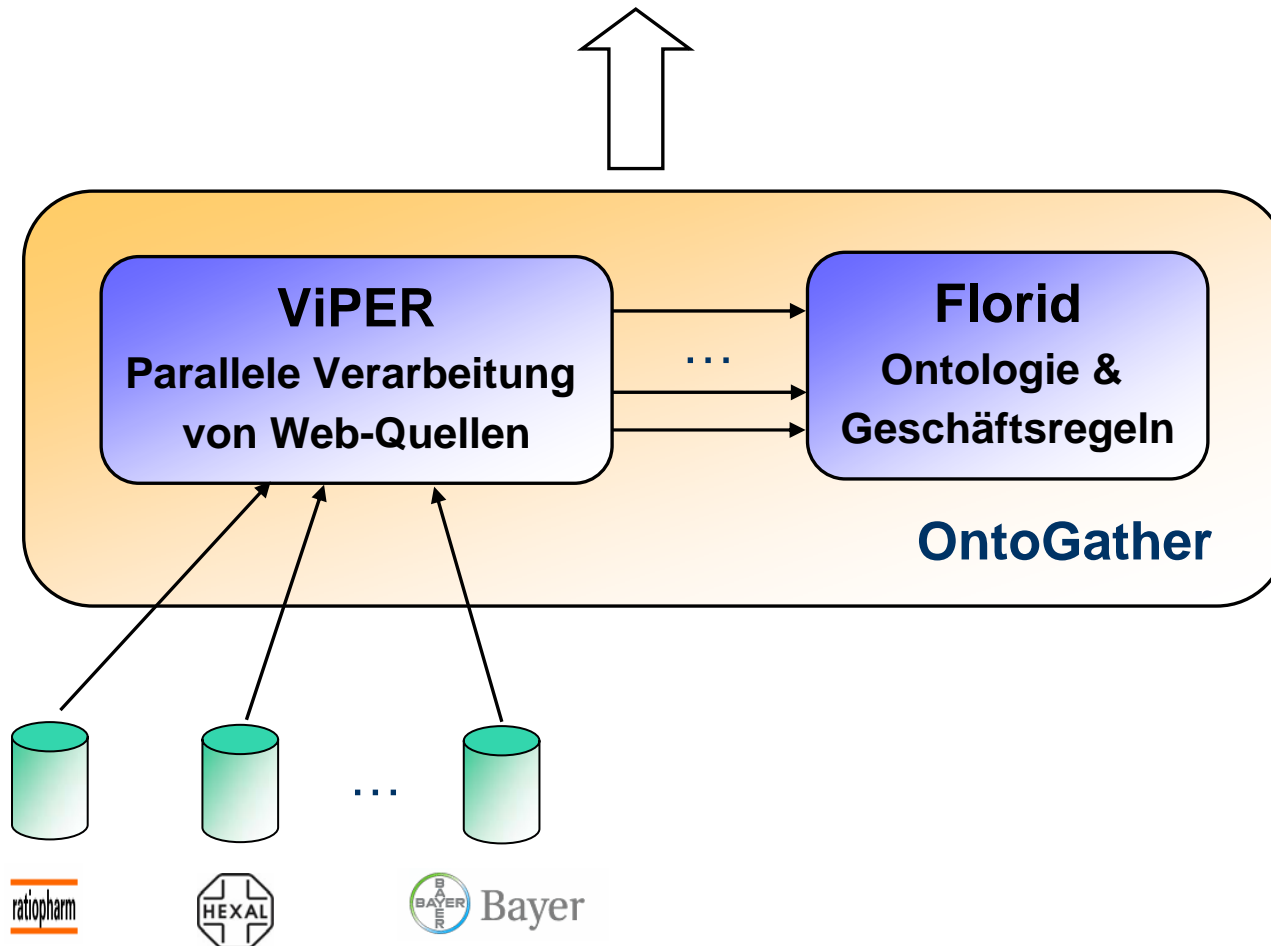


Anfrage an  
Domänen-Ontologie

# Beispiel-Szenario: Semantische Integration von Web-Quellen




„Wir sollten den Preis für unser Produkt erniedrigen“



## ViPER [ **Visual Perception-based Extraction of Records** ]

- Zielsetzung:
  - Einfache Integration heterogener unstrukturierter Web-Informationsquellen und deren Infrastrukturen.
- Features
  - operiert anhand einer Web Seite
  - benötigt mindestens zwei ähnlich strukturierte, direkt aufeinanderfolgende Datensätze
  - benutzt visuelle Informationen für die Segmentierung und das Gewichten der Datensätze
  - Ordnet die extrahierten Daten tabellarisch an, so dass eine einfache Notation der Daten möglich ist.

 einfach, schnell, effizient, wartungsarm

## FLORID [ **F-Logic Reasoning In Databases** ]

- Zielsetzung:
  - Verwaltung von Ontologien in F-Logik (OWL-Unterstützung in Vorbereitung)
  - Effiziente Beantwortung von F-Logik-Anfragen
- Features:
  - Implementiert in C++
  - Direkter Einfluss auf Integration von Daten aus Web-Seiten möglich
  - Geschäftsregeln, Domänen-Ontologie und Anfragen können in derselben Sprache modelliert werden

 einfache, konsistente Wartung

## OntoGather

### Vorteile:

- einfache Handhabung
  - höhere Effizienz
  - deutlich geringere Wartung
- im Vergleich zur herkömmlichen Vorgehensweise

### Stand der Entwicklung:

- Prototyp in Java entwickelt

### Nächste Schritte:

- Implementierung der Streaming-Verarbeitung in Florid-Komponente
- Berücksichtigung von Text Mining-Techniken zur Verbesserung der Datenextraktion