ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

iMeMex.org
Personal DataSpace Management System

iTrails: Pay-as-you-go Information Integration in Dataspaces

Marcos Vaz Salles   Jens Dittrich   Shant Karakashian
Olivier Girard   Lukas Blunschi

ETH Zurich

VLDB 2007

Outline

- Motivation
- iTrails
- Experiments
- Conclusions and Future Work

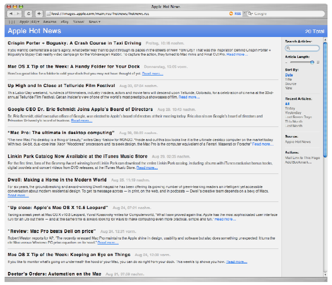
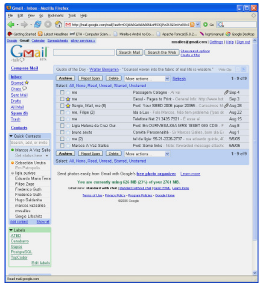
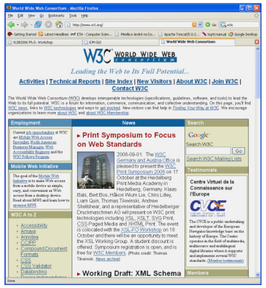
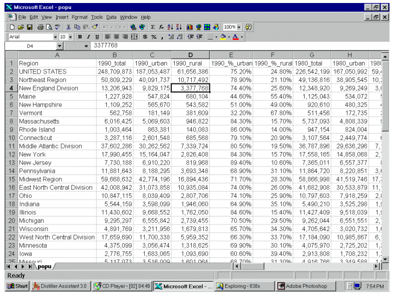
Problem: Querying Several Sources

Query

What is the impact of global warming in Zurich?

? ? ? ?

Systems

Data Sources


Laptop


Email Server

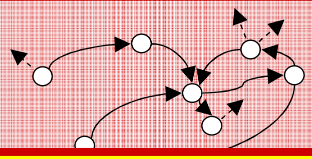

Web Server


DB Server

Solution 1: Use a Search Engine

Job!

Query global warming zurich

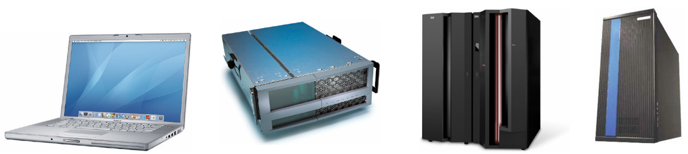


Graph IR
Search Engine

Drawback: Query semantics are not precise!

text, links
text, links
text, links
text, links

Data Sources



Laptop
Email Server
Web Server
DB Server

Jobs in Climate Change : Earthworks: PhD Student Mountain/Alpine ...
PhD Student Mountain/Alpine Soils and **Global Warming**, **Zurich**. A PhD position is open for an enthusiastic person interested in the response of high elevation ...
www.earthworks-jobs.com/climate/art7031.html - 6k - Cached - Similar pages

Impact of **global** dimming and brightening on **global warming**
Impact of **global** dimming and brightening on **global warming**. Martin Wild. Institute for Atmospheric and Climate Science, ETH **Zurich**, **Zurich**, Switzerland ...
www.agu.org/pubs/crossref/2007/2006GL028031.shtml - 7k - Cached - Similar pages

swissinfo - swissinfo talks to Swiss scientist Konrad Steffen ...
Iceman keeps his cool despite **global warming** ... set up the Swiss Camp in Greenland for the Federal Institute of Technology in **Zurich** in 1990 (swissinfo) ...
www.swissinfo.org/eng/feature/detail/Iceman_keeps_his_cool_despite_**global_warming**.html?siteSect=108&s... - 41k - Cached - Similar pages

SSRN-Uncertainty and **Global Warming**: An Option-Pricing Approach to ...

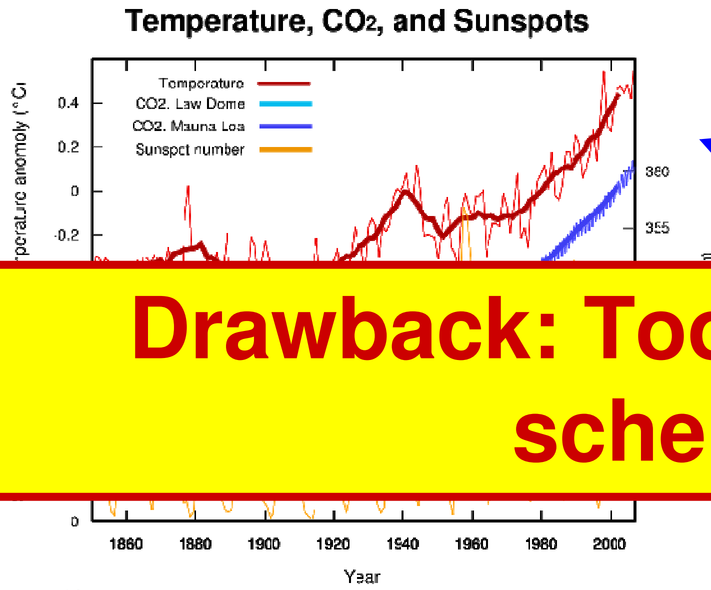
These scientists been warning about **global warming**, and its acceleration, for many years. For decades, the research institute at **Zurich** University has ...
www.rferl.org/featuresarticle/2007/02/13b23c06-e87e-41f4-9860-ae8a5b54d0bc.html - 41k - Cached - Similar pages

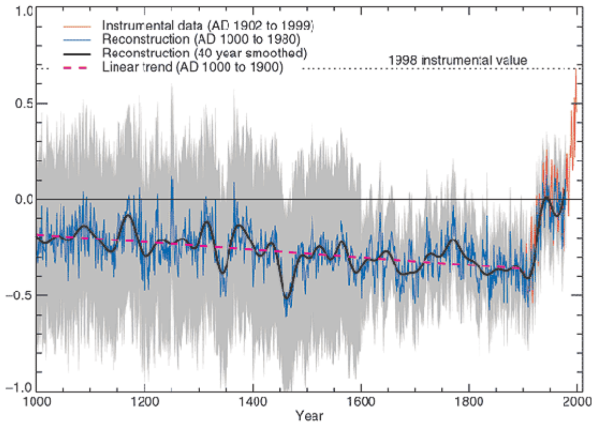
Decades of devastation ahead as **global warming** melts the Alps ...
Decades of devastation ahead as **global warming** melts the Alps ... Research by Davies - to be outlined this week at the **Zurich** conference - has discovered ...
observer.guardian.co.uk/international/story/0,6903,1001674,00.html - 48k - Cached - Similar pages

ETH - DUWIS - Atmosphäre und Klima - [Translate this page]
Umwelt, Umweltnaturwissenschaften, Studium, ETH **Zürich**, Environment, Environmental Sciences, Graduate Study Courses, ETH ZurichUmweltnaturwissenschaften, ...
www.env.ethz.ch/research/3 - 23k - Cached - Similar pages

peopleandplanet.net > climate change > newsfile > ski resorts ...
Ski resorts heading downhill owing to **global warming** ... for Economic Geography at the University of **Zurich**, and Dr Bruno Abegg, a travel journalist. ...
www.peopleandplanet.net/doc.php?id=2083 - 40k - Cached - Similar pages

Solution 2: Use an Information Integration System





`//Temperatures/*[city = "zurich"]`

Query

Information

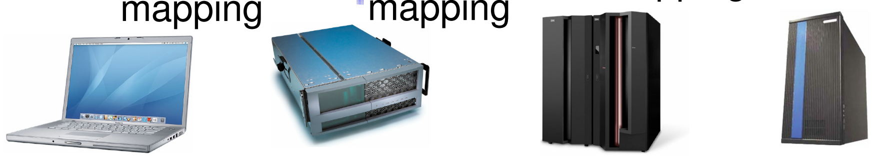
Drawback: Too much effort to provide schema mappings!

missing schema mapping

missing schema mapping

schema mapping

schema mapping



Laptop

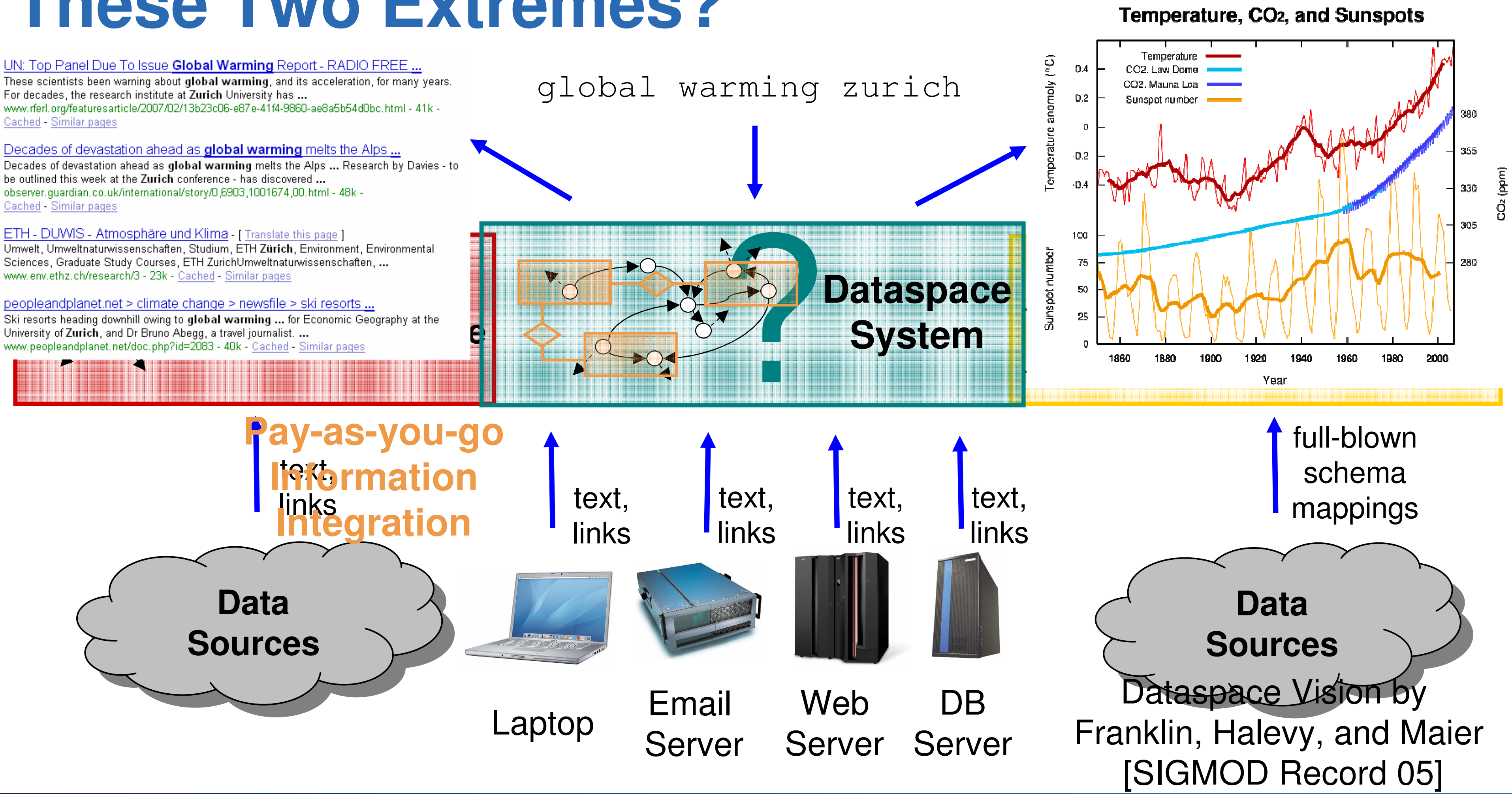
Email Server

Web Server

DB Server

Data Sources

Research Challenge: Is There an Integration Solution in-between These Two Extremes?



Outline

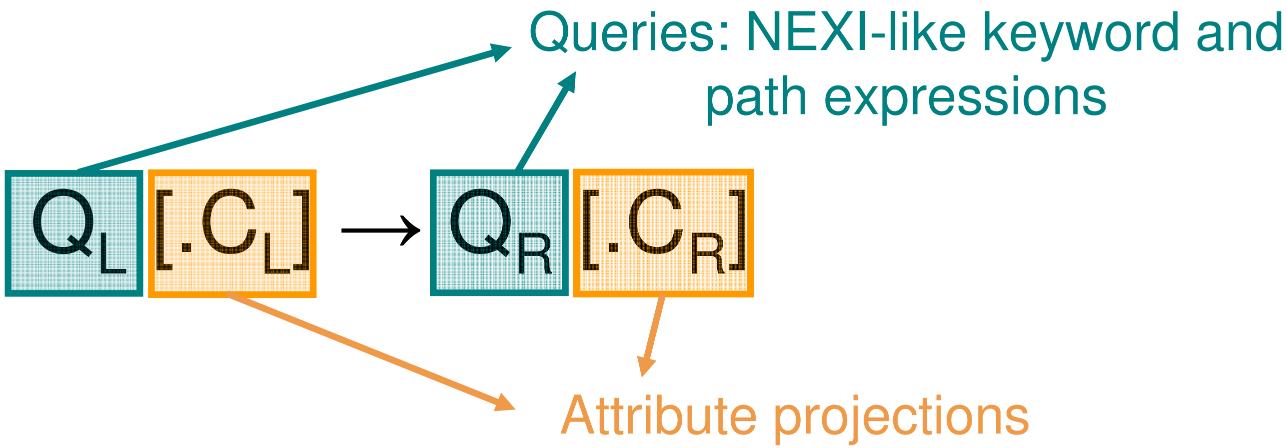
- Motivation
- iTrails
- Experiments
- Conclusions and Future Work

iTrails Core Idea: Add Integration Hints Incrementally

- **Step 1:** Provide a search service over all the data
  - Use a general graph data model (see VLDB 2006)
  - Works for unstructured documents, XML, and relations
- **Step 2:** Add integration semantics via hints (trails) on top of the graph
  - Works across data sources, not only between sources
- **Step 3:** If more semantics needed, go back to step 2
- **Impact:**
  - Smooth transition between search and data integration
  - Semantics added incrementally improve precision / recall

iTrails: Defining Trails

- **Basic Form of a Trail**



- **Intuition:**
  When I query for Q_L [.C_L], you should also query for Q_R [.C_R]

Trail Examples: Global Warming Zurich

DB Server

`global warming zurich`

Temperatures

date	city	region	celsius
24-Sep	Bern	BE	20
24-Sep	Uster	ZH	15
25-Sep	Zurich	ZH	14
26-Sep	Zurich	ZH	9

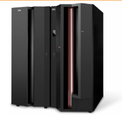
- **Trail for Implicit Meaning:** "When I query for `global warming`, you should also query for Temperature data above 10 degrees"

```
global warming →
   //Temperatures/*[celsius > 10]
```

- **Trail for an Entity:** "When I query for `zurich`, you should also query for references of zurich as a region"

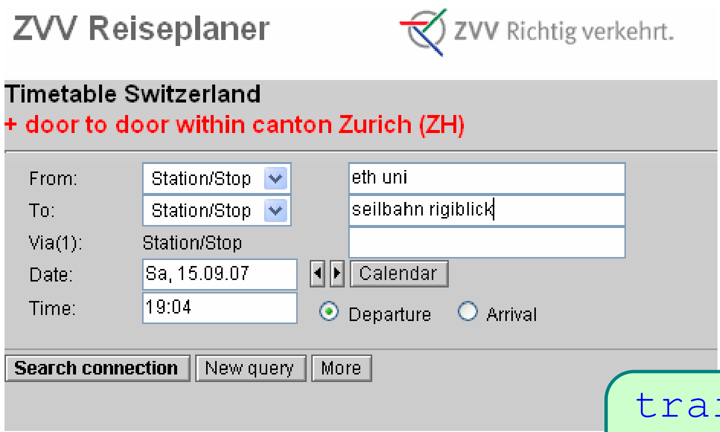
```
zurich → //*[region = "ZH"]
```

Trail Example: Deep Web Bookmarks



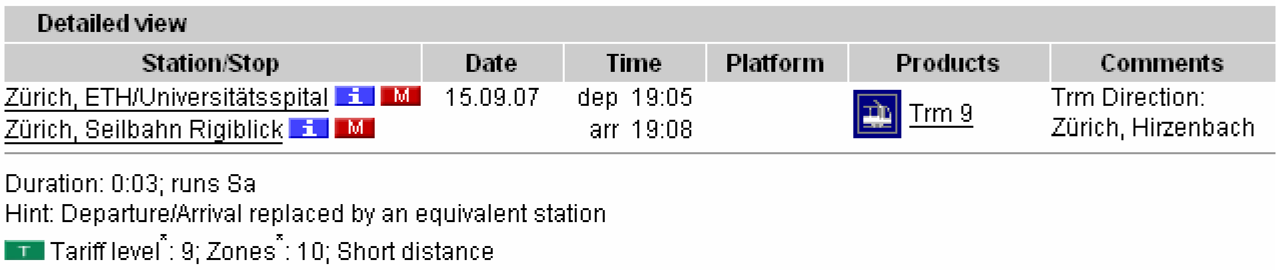
Web Server

`train home`



- **Trail for a Bookmark:** "When I query for `train home`, you should also query for the `TrainCompany`'s website with `origin` at `ETH Uni` and `destination` at `Seilbahn Rigiblick`"

```
train home →
   //trainCompany.com//*[origin="ETH Uni"
         and dest ="Seilbahn Rigiblick"]
```



Detailed view

Station/Stop	Date	Time	Platform	Products	Comments
Zürich, ETH/Universitätsspital	15.09.07	dep 19:05		Trm 9	Trm Direction:
Zürich, Seilbahn Rigiblick		arr 19:08			Zürich, Hirzenbach

Duration: 0:03; runs Sa
Hint: Departure/Arrival replaced by an equivalent station
T Tariff level*: 9; Zones*: 10; Short distance

Trail Examples: Thesauri, Dictionaries, Language-agnostic Search

Laptop Email Server

- **Trail for Thesauri:** "When I query for `car`, you should also query for `auto`"

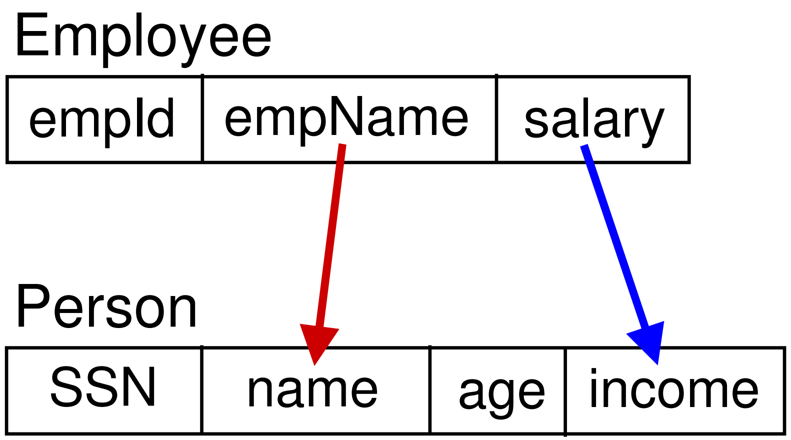
`car → auto`

- **Trails for Dictionary:** "When I query for `car`, you should also query for `carro` and vice-versa"

`car → carro`
`carro → car`

Trail Examples: Schema Equivalences

DB Server



- **Trail for schema match on names:** "When I query for `Employee.empName`, you should also query for `Person.name`"

```
//Employee//*.tuple.empName →
                //Person//*.tuple.name
```

- **Trail for schema match on salaries:** "When I query for `Employee.salary`, you should also query for `Person.income`"

```
//Employee//*.tuple.salary →
             //Person//*.tuple.income
```

Outline

- Motivation
- iTrails
- Experiments
- Conclusion and Future Work

- Core Idea
- Trail Examples
- How are Trails Created?
- Uncertainty and Trails
- Rewriting Queries with Trails
- Recursive Matches

How are Trails Created?

- Given by the user
  - Explicitly
  - Via Relevance Feedback
- (Semi-)Automatically
  - Information extraction techniques
  - Automatic schema matching
  - Ontologies and thesauri (e.g., wordnet)
  - User communities (e.g., trails on gene data, bookmarks)

Uncertainty and Trails

- **Probabilistic Trails:**
  - model uncertain trails
  - probabilities used to rank trails

$$Q_L\ [.C_L] \underset{p}{\longrightarrow} Q_R\ [.C_R],\ 0 \leq p \leq 1$$

  - Example: `car` $\underset{p\,=\,0.8}{\longrightarrow}$ `auto`

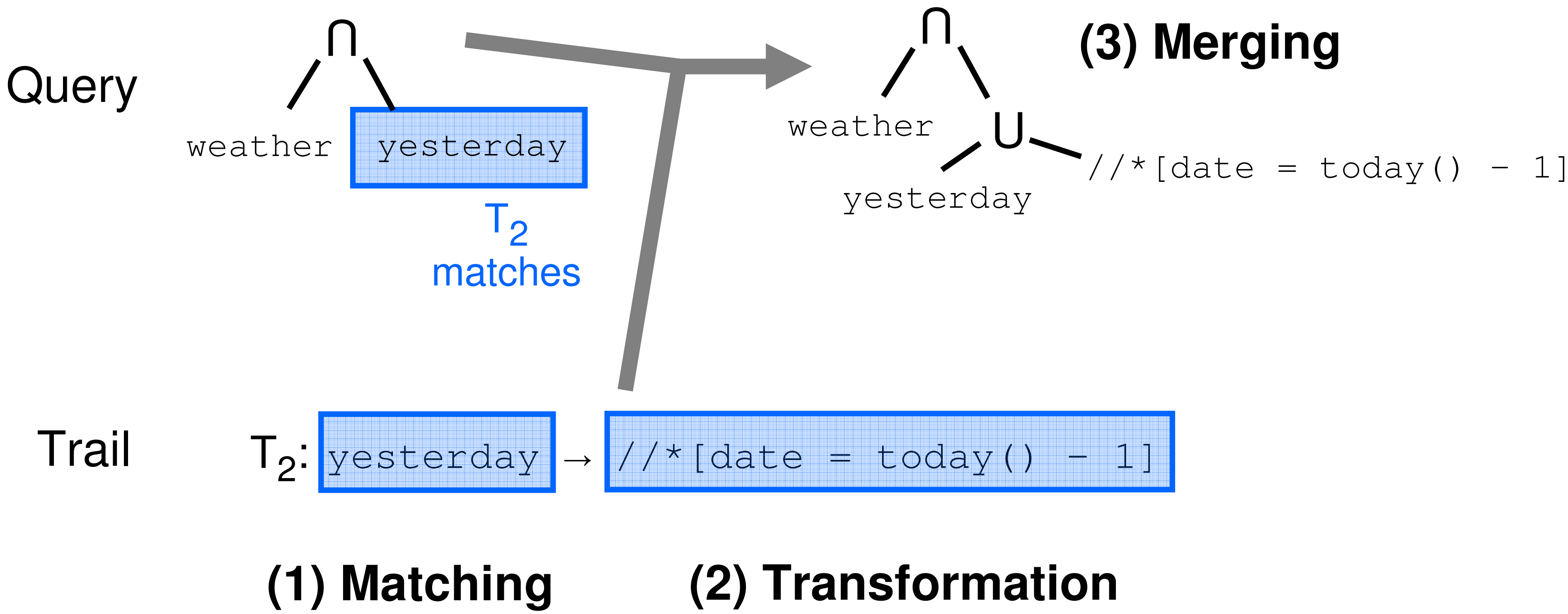
Certainty and Trails

- **Scored Trails:**
  - give higher value to certain trails
  - scoring factors used to boost scores of query results obtained by the trail

$$Q_L\ [.C_L] \xrightarrow[sf]{} Q_R\ [.C_R],\ sf > 1$$

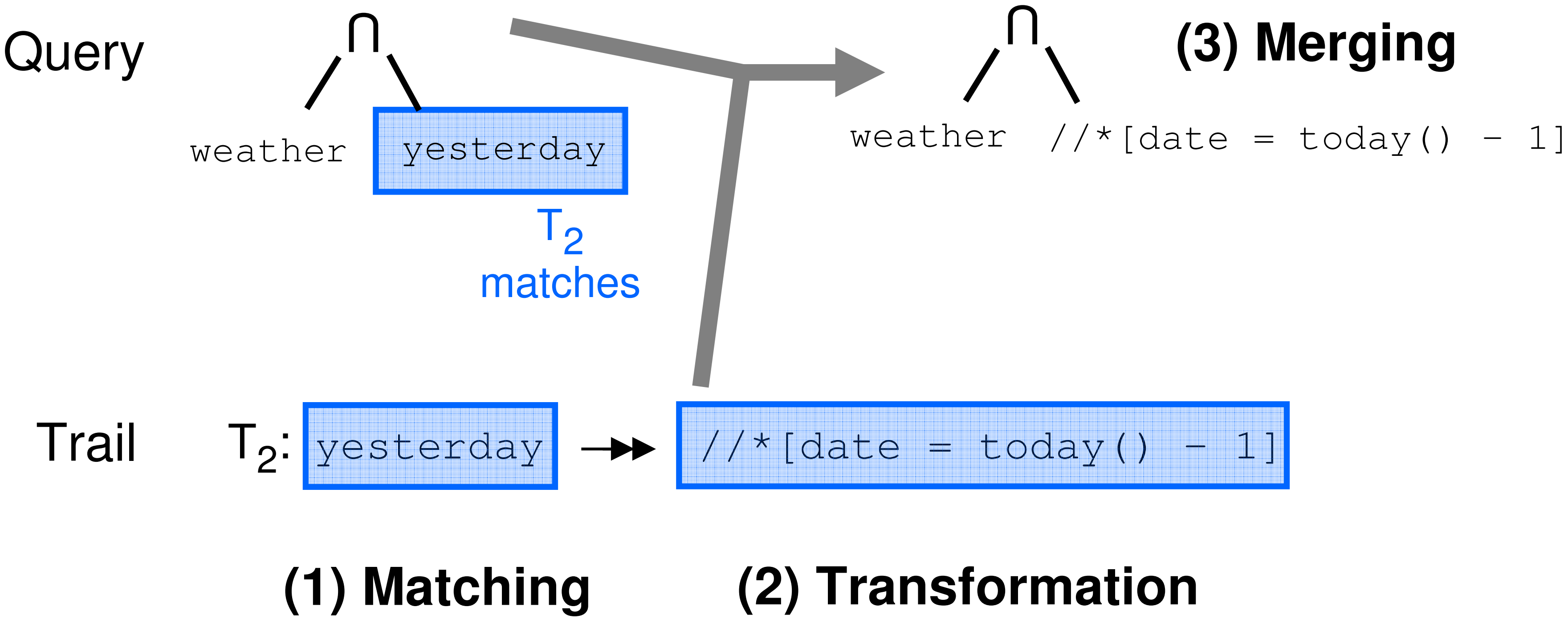
  - Examples:
    - T_1: `weather → //Temperatures/*`
      p = 0.9, sf = 2
    - T_2: `yesterday → //*[date = today() - 1]`
      p = 1, sf = 3

Rewriting Queries with Trails

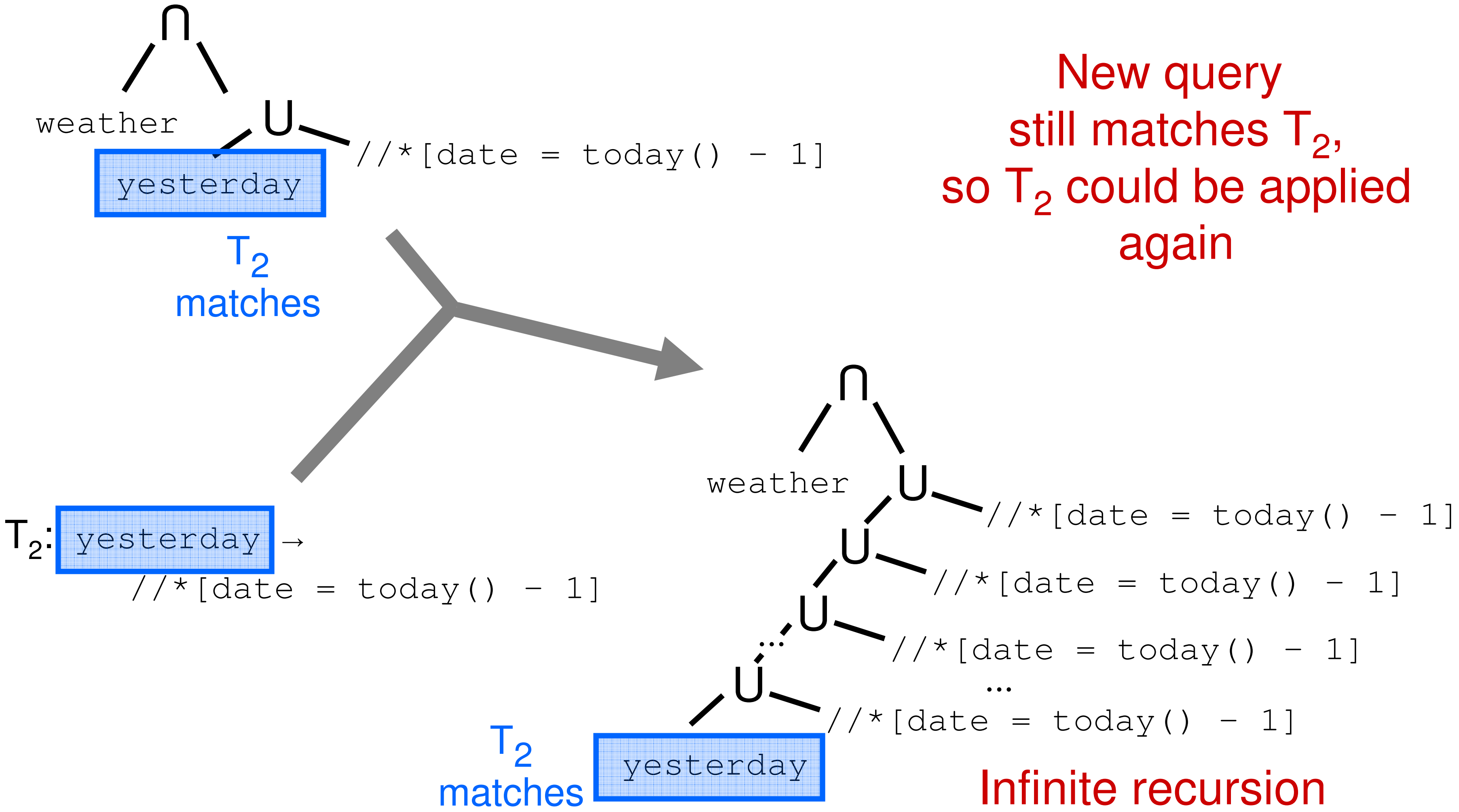


Replacing Trails

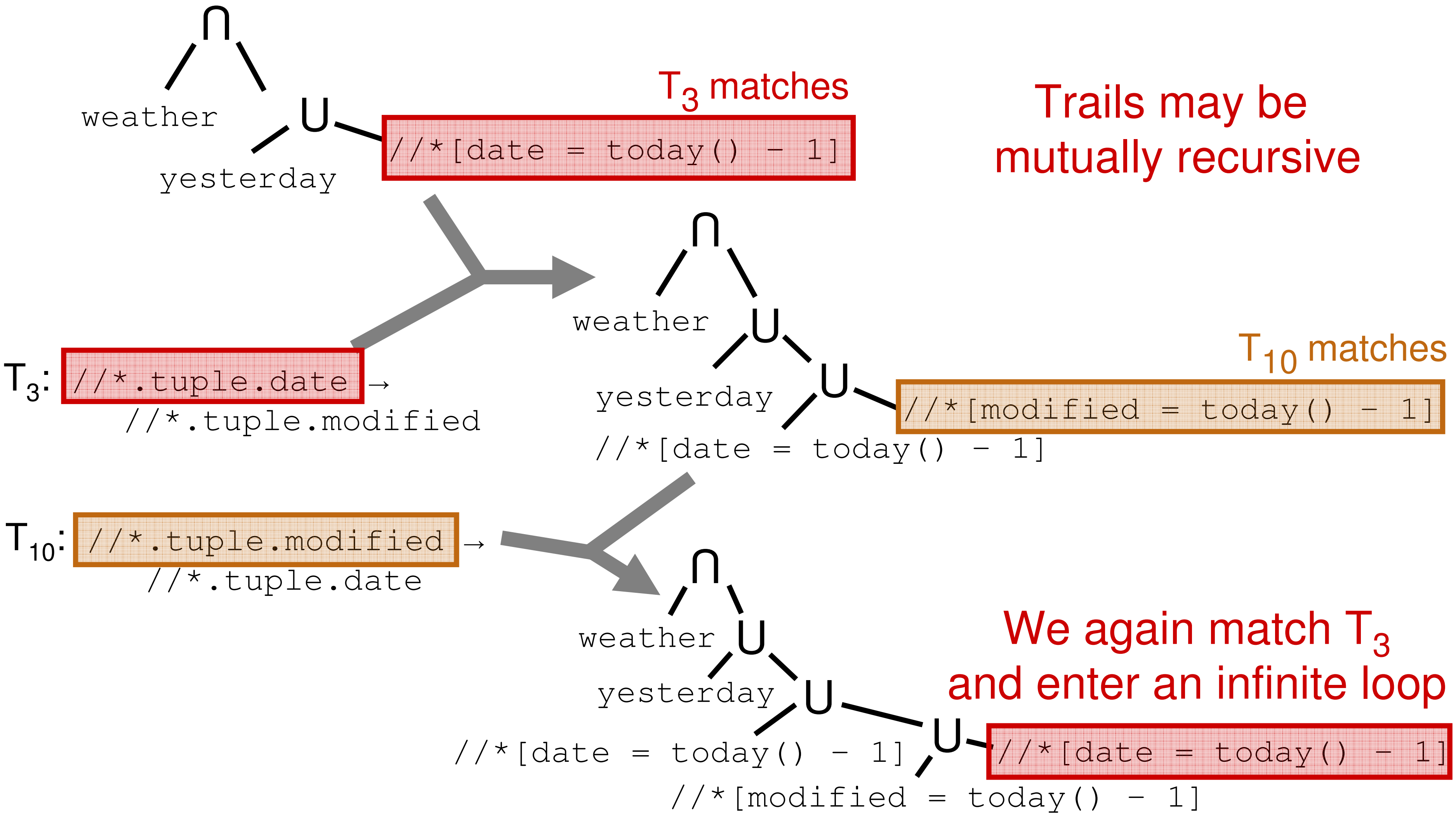
- Trails that use replace instead of union semantics



Problem: Recursive Matches (1/2)



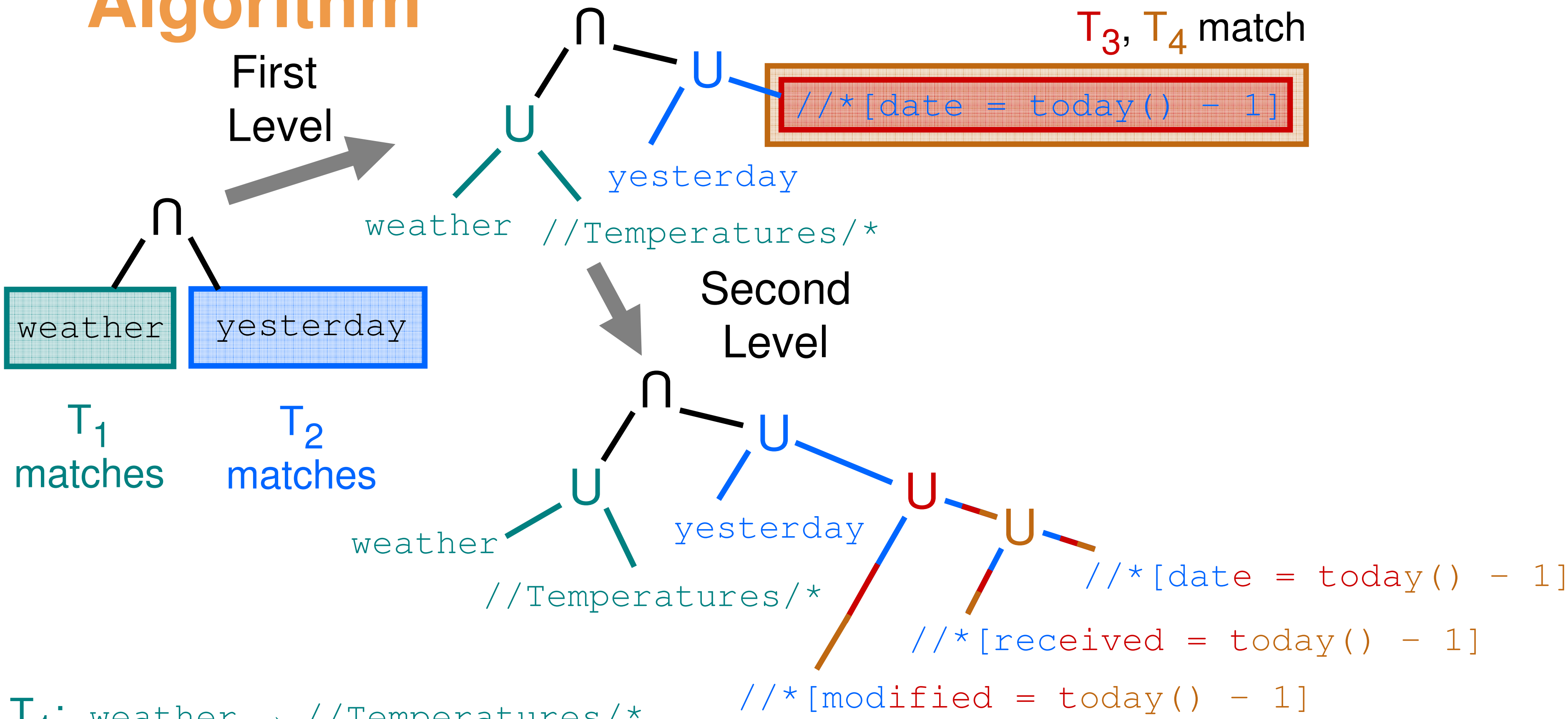
Problem: Recursive Matches (2/2)



Trails may be mutually recursive

We again match T_3 and enter an infinite loop

Solution: Multiple Match Coloring Algorithm



T_1: weather → //Temperatures/*

T_2: yesterday → //*[date = today() - 1]

T_3: //*.tuple.date → //*.tuple.modified

T_4: //*.tuple.date → //*.tuple.received

Multiple Match Coloring Algorithm Analysis

- **Problem:** MMCA is exponential in number of levels
- **Solution: Trail Pruning**
  - Prune by number of levels
  - Prune by top-K trails matched in each level
  - Prune by both top-K trails and number of levels

Outline

- Motivation
- iTrails
- Experiments
- Conclusion and Future Work

iTrails Evaluation in iMeMex

- **iMeMex Dataspace System:** Open-source prototype available at http://www.imemex.org

- **Main Questions in Evaluation**
  - Quality: Top-K Precision and Recall
  - Performance: Use of Materialization
  - Scalability: Query-rewrite Time vs. Number of Trails

iTrails Evaluation in iMeMex

- **Scenario 1:** Few High-quality Trails
  - Closer to information integration use cases
  - Obtained real datasets and indexed them
  - 18 hand-crafted trails
  - 14 hand-crafted queries

- **Scenario 2:** Many Low-quality Trails
  - Closer to search use cases
  - Generated up to 10,000 trails

iTrails Evaluation in iMeMex: Scenario 1

- Configured iMeMex to act in three modes
  - **Baseline:** Graph / IR search engine
  - **iTrails:** Rewrite search queries with trails
  - **Perfect Query:** Semantics-aware query
- Data: shipped to central index

	Desktop	Wiki4V	Enron	DBLP	$\sum$
Net Data size	1,230	26,392	111	713	28,446

sizes in MB


Laptop


Web Server

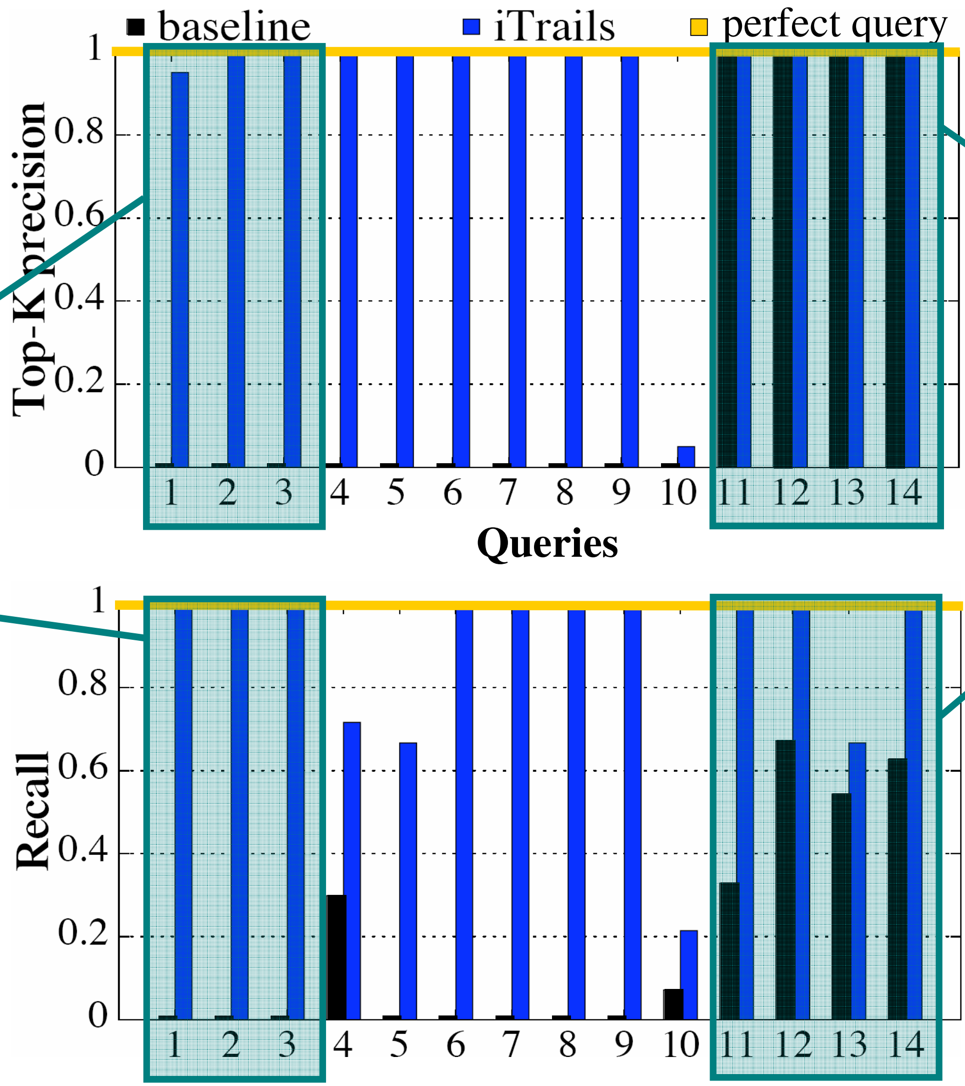

Email Server


DB Server

Quality: Top-K Precision and Recall

K = 20

Scenario 1:
few high-quality trails
(18 trails)



Search Engine misses relevant results

Perfect Query always has precision and recall equal to 1

Search Query is partially semantics-aware

Q3: `pdf yesterday`

Q13: `to = raimund.grube@enron.com`

Performance: Use of Materialization

Scenario 1:
few high-quality trails
(18 trails)

Trail merging adds overhead to query execution

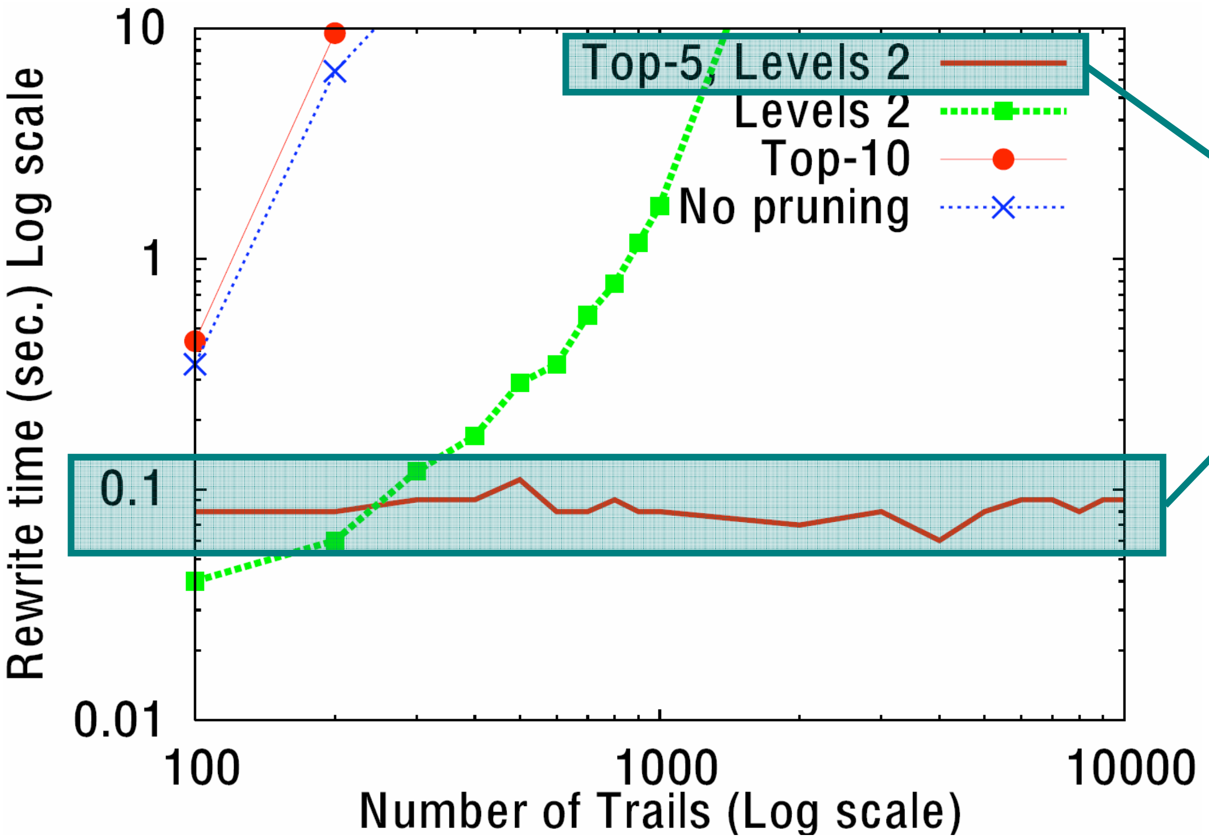
Trail Materialization provides interactive times for all queries

Q. No.	iTrails with Basic Indexes	iTrails with Trail Mat.
1	2.18	0.21
2	0.74	0.52
3	10.72	0.39
4	1.86	0.07
5	0.56	0.44
6	0.32	0.05
7	1.73	0.67
8	5.27	0.48
9	179.02	1.50
10	10.14	0.29
11	0.60	0.60
12	0.60	0.60
13	0.49	0.44
14	0.14	0.14

response times in sec.

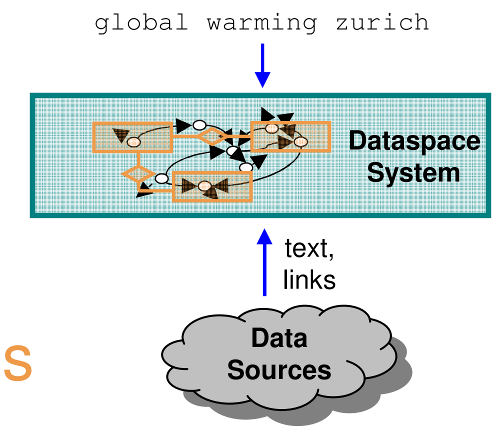
Scalability: Query-rewrite Time vs. Number of Trails

Scenario 2:
many low-quality
trails



Query-rewrite time can be controlled with pruning

Conclusion: Pay-as-you-go Information Integration



- **Step 1:** Provide a search service over all the data
- **Step 2:** Add integration semantics via trails
- **Step 3:** If more semantics needed, go back to step 2
- **Our Contributions**
  - iTrails: generic method to model semantic relationships (e.g. implicit meaning, bookmarks, dictionaries, thesauri, attribute matches, ...)
  - We propose a framework and algorithms for Pay-as-you-go Information Integration
  - Smooth transition between search and data integration

Future Work

- Trail Creation
  - Use collections (ontologies, thesauri, wikipedia)
  - Work on automatic mining of trails from the dataspace
- Other types of trails
  - Associations
  - Lineage

Questions?

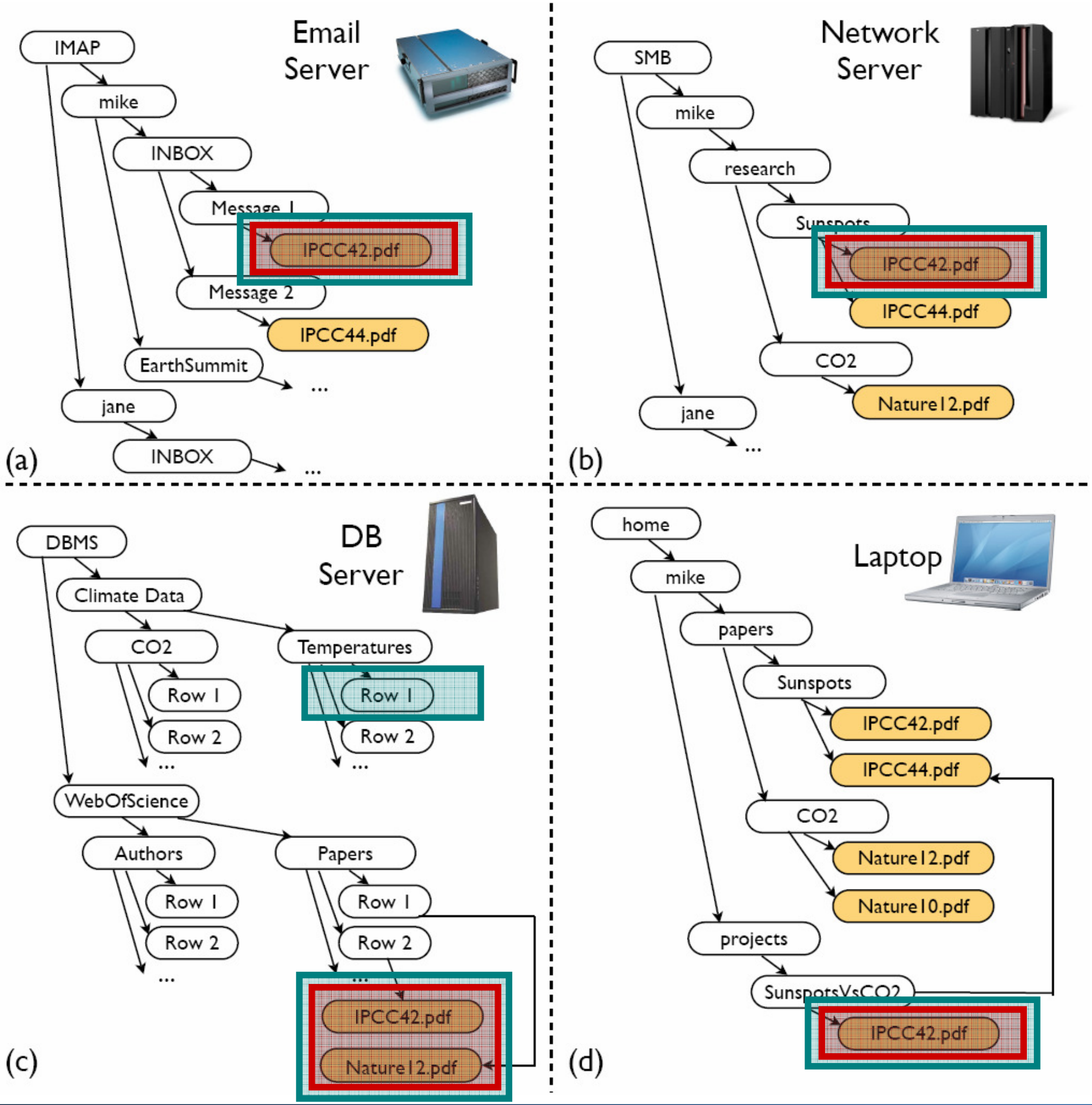
Thanks in advance for your feedback! ☺

marcos.vazsalles@inf.ethz.ch

http://www.imemex.org

Backup Slides

Problem: Global Warming in Zurich

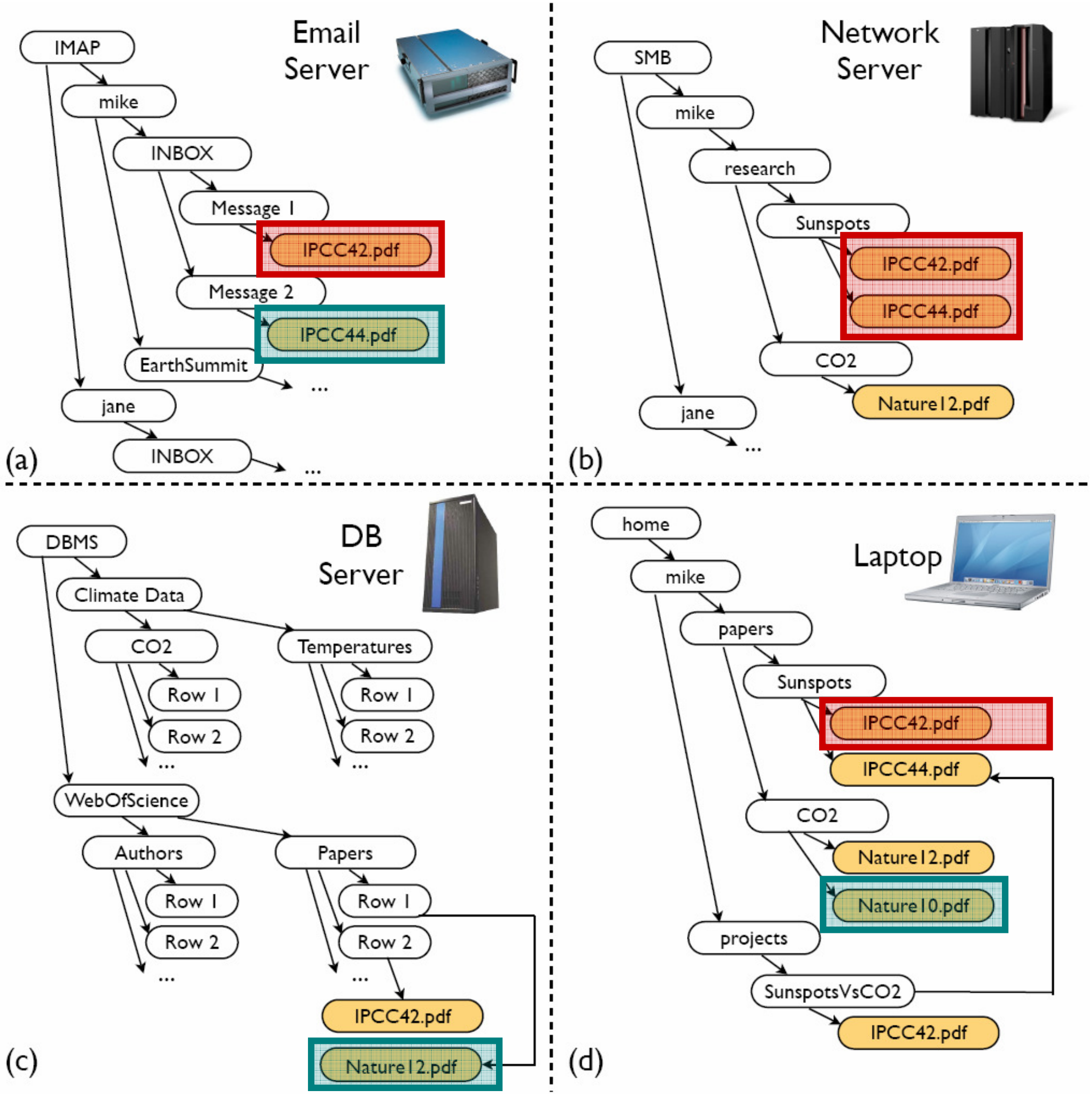


- Query: “What is the impact of global warming in Zurich?”
- Search for:

`global warming zurich`

- Meaning of keyword query
  - `global warming` should lead to query on Temperatures
  - `zurich` should lead to a query for a city

Problem: PDF Yesterday



- Query: “Retrieve all PDF documents added/modified yesterday”
- Search for:

  `pdf yesterday`

- Meaning of keywords `pdf` and `yesterday`
- Different sources, different schemas:
  - Laptop: modified
  - Email: received
  - DBMS: changed

Related Work: Search vs. Data Integration vs. Dataspaces

		Integration Solution		
		Search	**Dataspaces**	**Data Integration**
Features	**Integration Effort**	Low	Pay-as-you-go	High
	Query Semantics	Precision / Recall	Precision / Recall	Precise
	Need for Schema	Schema-never	Schema-later	Schema-first

Personal Dataspaces Literature

- Dittrich, Salles, Kossmann, Blunschi. **iMeMex: Escapes from the Personal Information Jungle (Demo Paper).** VLDB, September 2005.
- Dittrich, Salles. **iDM: A Unified and Versatile Data Model for Personal Dataspace Management.** VLDB, September 2006
- Dittrich. **iMeMex: A Platform for Personal Dataspace Management.** SIGIR PIM, August 2006.
- Blunschi, Dittrich, Girard, Karakashian, Salles. **A Dataspace Odyssey: The iMeMex Personal Dataspace Management System (Demo Paper).** CIDR, January 2007.
- Dittrich, Blunschi, Färber, Girard, Karakashian, Salles. **From Personal Desktops to Personal Dataspaces: A Report on Building the iMeMex Personal Dataspace Management System.** BTW 2007, March 2007
- Salles, Dittrich, Karakashian, Girard, Blunschi. **iTrails: Pay-as-you-go Information Integration in Dataspaces**. VLDB, September 2007

iDM: iMeMex Data Model

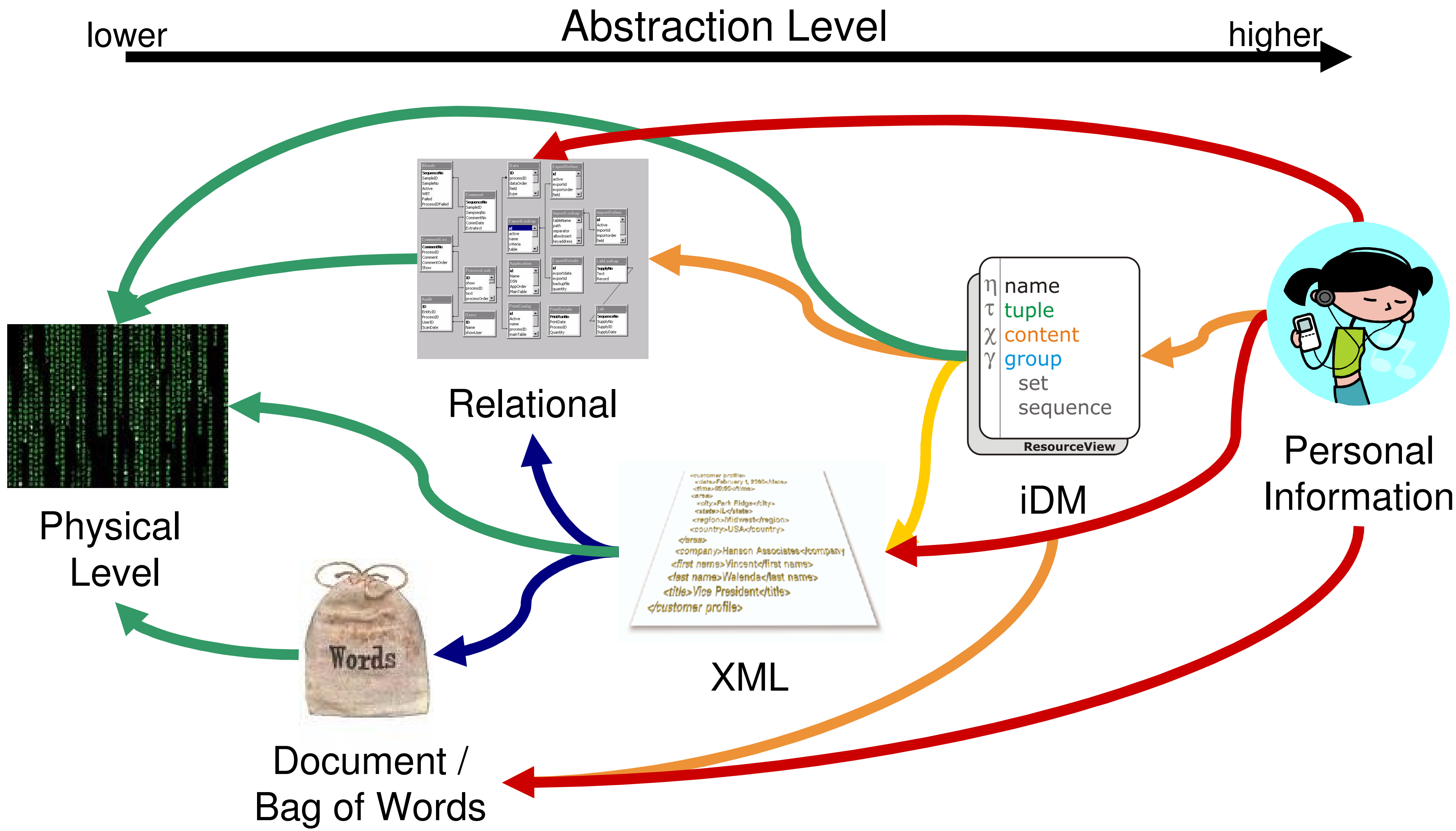
- **Our approach:** get the data model closer to personal information – not the other way around
- Supports:
  - Unstructured, semi-structured and structured data, e.g., files&folders, XML, relations
  - Clearly separation of logical and physical representation of data
  - Arbitrary directed graph structures, e.g., section references in LaTeX documents, links in filesystems, etc
  - Lazily computed data, e.g., ActiveXML (Abiteboul et. al.)
  - Infinite data, e.g., media and data streams

See VLDB 2006

Data Model Options

		Data Models			
		Bag of Words	Relational	XML	iDM
Support for Personal Data	Non-schematic data	✓	⊘	✓	✓
	Serialization independent	✓	✓	⊘	✓
	Support for Graph data	⊘	Specific schema	Extension: XLink/ XPointer	✓
	Support for Lazy Computation	⊘	View mechanism	Extension: ActiveXML	✓
	Support for Infinite data	Extension: Document streams	Extension: Relational streams	Extension: XML streams	✓

Data Models for Personal Information



Architectural Perspective of iMeMex

Complex operators
(query algebra)

Indexes&Replicas access
(warehousing)

Data source access
(mediation)