Learning to attach semantic metadata to Web Services

Andreas Heß, Nicholas Kushmerick

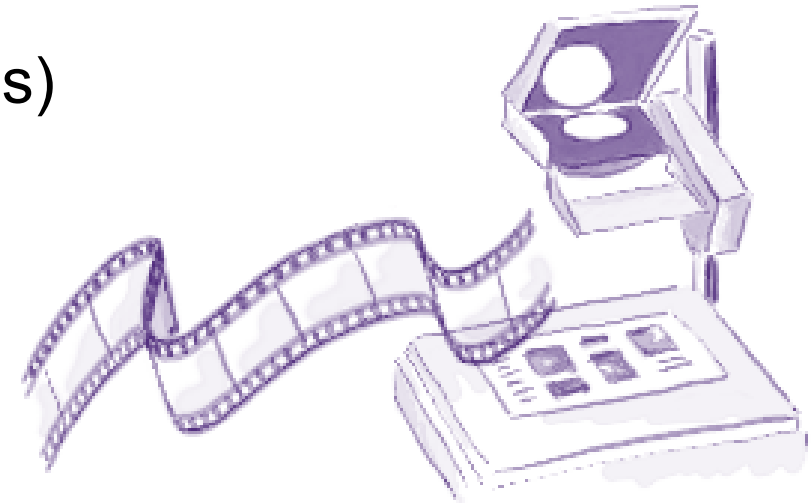
University College Dublin, Ireland

{andreas.hess, nick}@ucd.ie

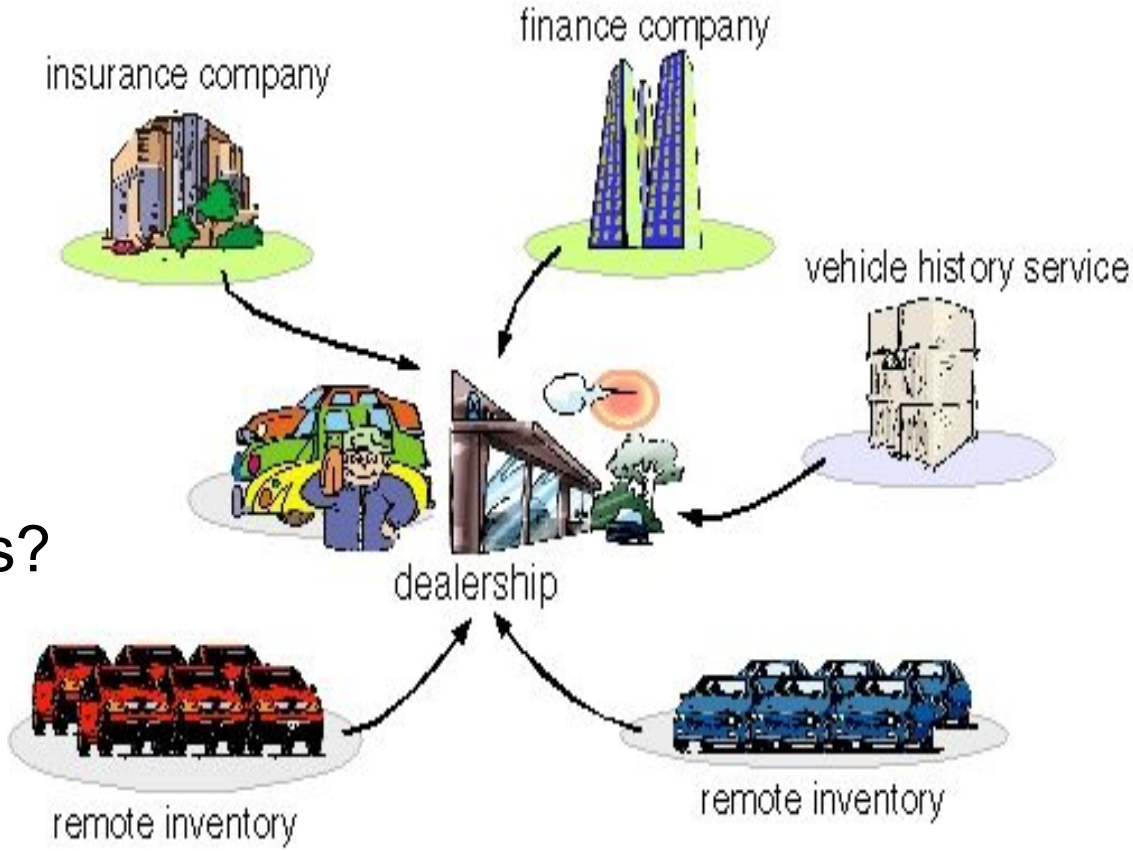
US Office of Naval Research

Science Foundation Ireland

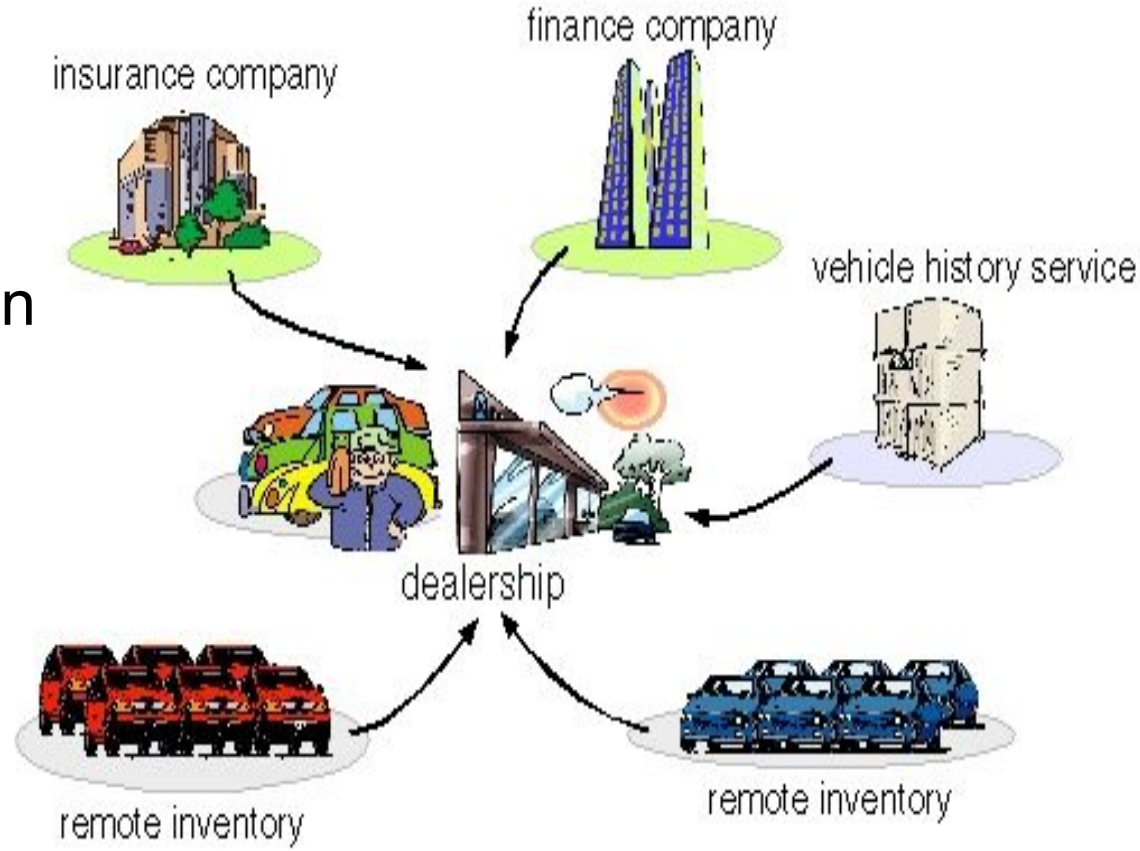
1. Introduction
2. Our Machine Learning Approach
3. Classification Tasks
   a) Domain & Datatype (Web Forms)
   b) Category (Web Services)
4. Conclusion & Discussion



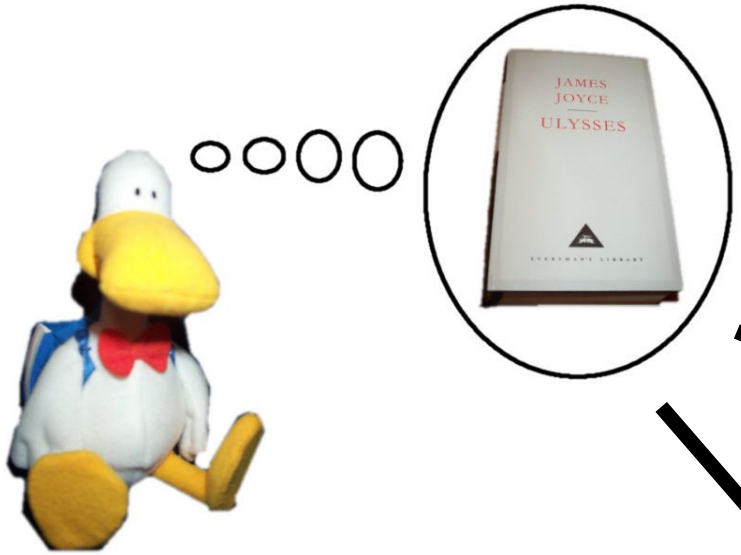
- Web Services
  - Web-accessible computational entities
  - Loosely coupled
  - Just RPC? Web forms?
  - Data Integration?



- Tasks / Problems
  - Automatic Discovery
  - Automatic Composition
  - Automatic Invocation



Scenario



?

?

Congo

- author
- title
- quantity

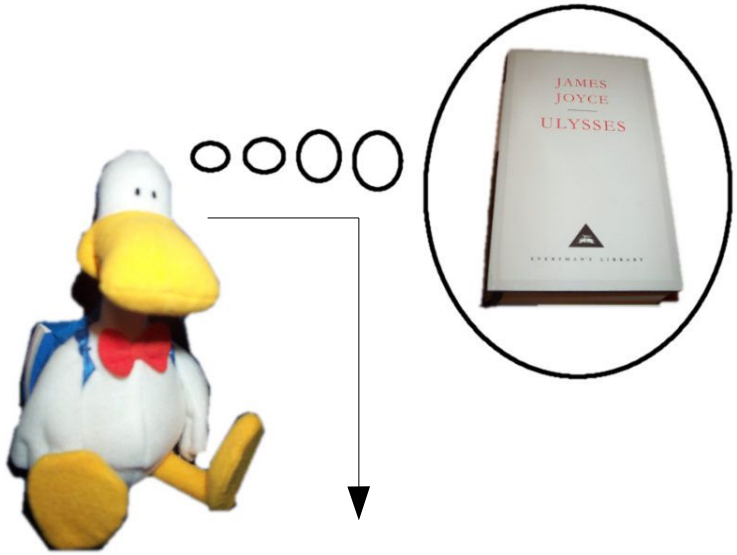
Winding Stair

- authName
- bookT
- ISBN

Teatime

- region
- qlty
- qty

Scenario



Global Ontology

- Item
  - Quantity
  - Price
  - Book
    - Author
    - Title
    - ISBN
  - Tea
    - Region
    - Quality

Congo

- author
- title
- quantity

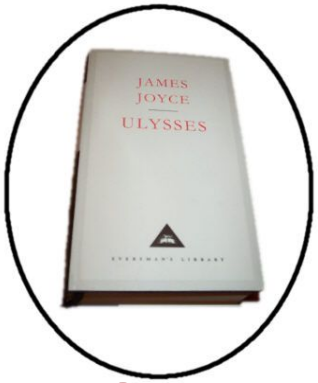
Winding Stair

- authName
- bookT
- ISBN

Teatime

- region
- qlty
- qty

Scenario



Semantic Metadata
e.g. DAML-S / OWL-S

Global Ontology

- Item
  - Quantity
  - Price
  - Book
    - Author
    - Title
    - ISBN
  - Tea
    - Region
    - Quality

(handcrafted)

Congo

- author
- title
- quantity

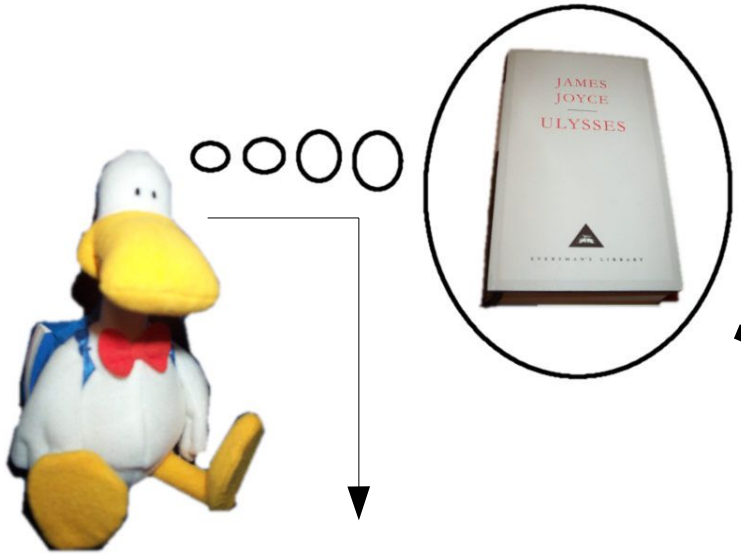
Winding Stair

- authName
- bookT
- ISBN

Teatime

- region
- qlty
- qty

Scenario



!

Congo

- author
- title
- quantity

Winding Stair

- authName
- bookT
- ISBN

Teatime

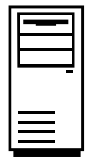
- region
- qlty
- qty

Global Ontology

- Item
  - Quantity
  - Price
  - Book
    - Author
    - Title
    - ISBN
  - Tea
    - Region
    - Quality

- Assumes:
  - semantic annotation
  - a shared ontology
- Semantic metadata needs to be handcrafted!!

- Our contribution: Use machine learning!



✔ Introduction

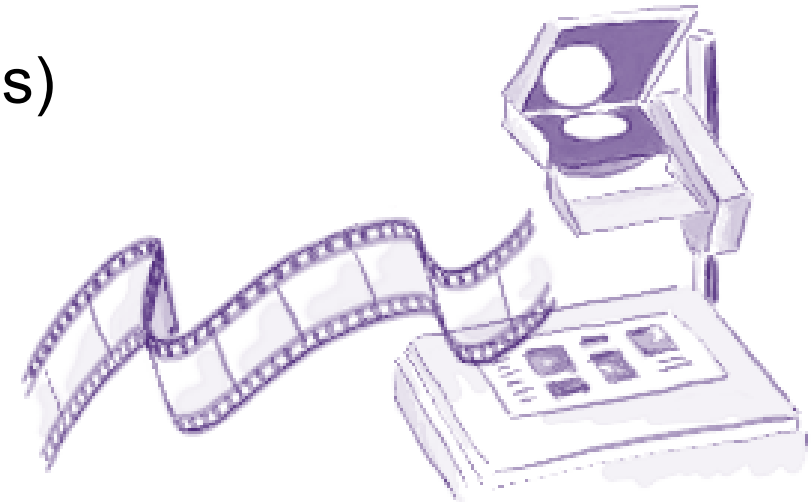
2. Our Machine Learning Approach

3. Classification Tasks

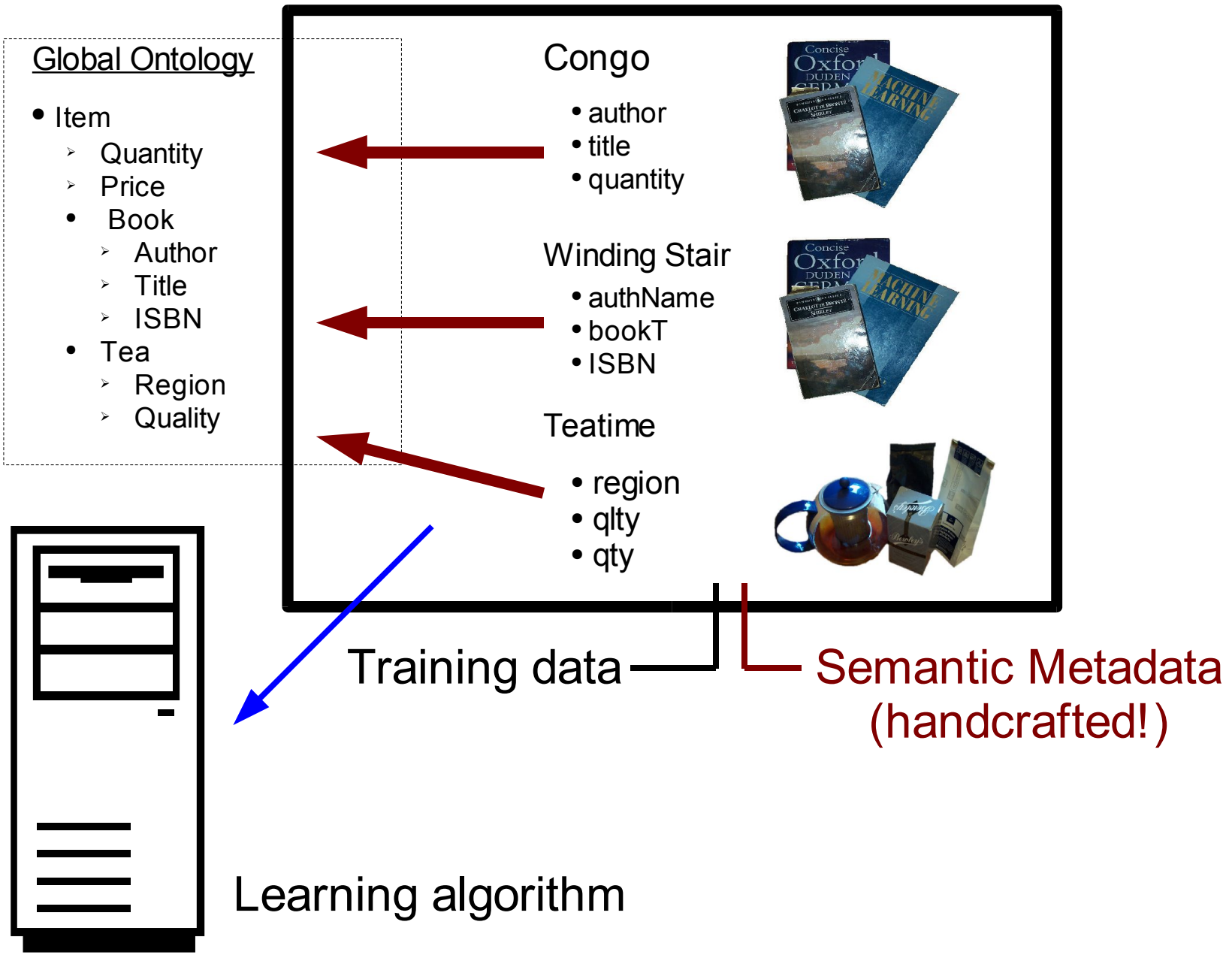
   a) Domain & Datatype (Web Forms)

   b) Category (Web Services)

4. Conclusion & Discussion



Machine Learning



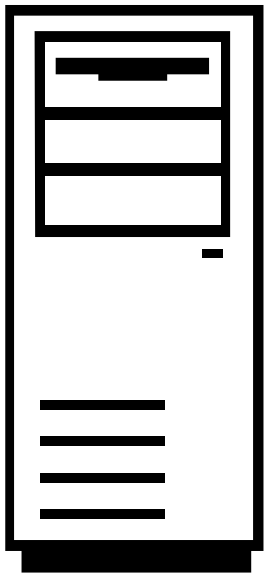
Machine Learning

Global Ontology

- Item
  - Quantity
  - Price
  - Book
    - Author
    - Title
    - ISBN
  - Tea
    - Region
    - Quality

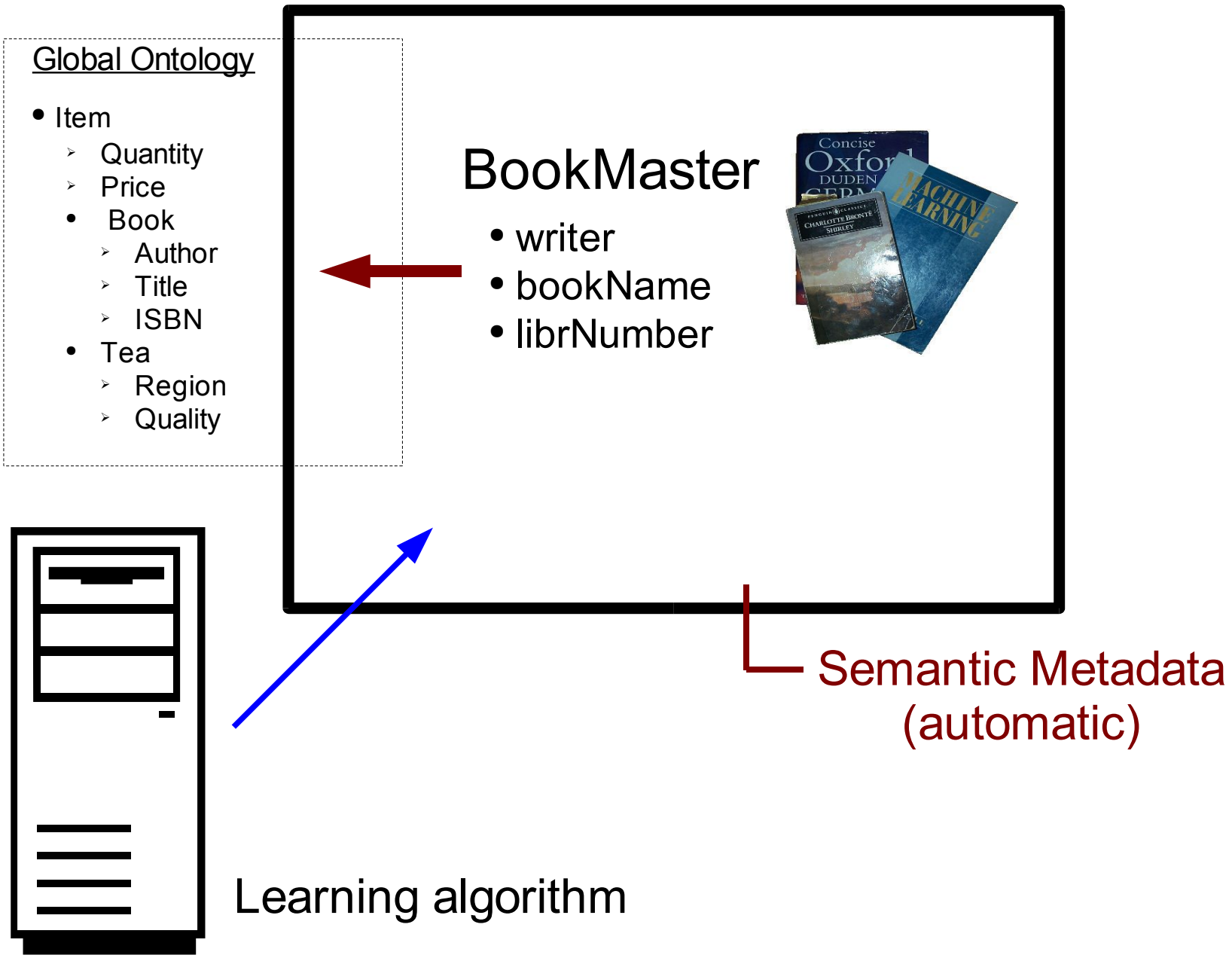
BookMaster

- writer
- bookName
- librNumber



Learning algorithm

Machine Learning



What does this function do?

```
public int nbgfuibhuf(int nvzfdubzuf, int cnuzdc) {

      int vfddf = 0;
      for (int ujz = 0; ujz < nvzfdubzuf; ujz++) {
         vfddf += cnuzdc;
      }
      return vfddf;
}
```

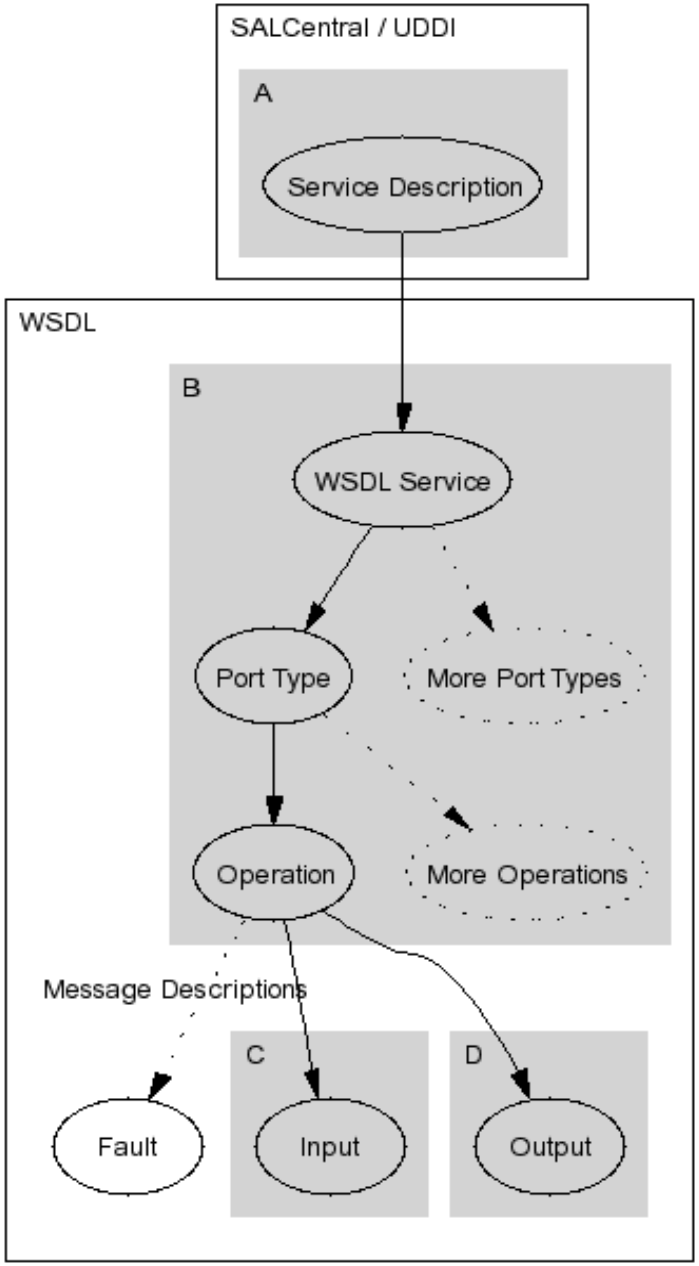
What does this function do?

```
public int multiply(int factor1, int factor2) {

      int product = 0;
      for (int n = 0; n < factor1; n++) {
         product += factor2;
      }
      return product;
}
```

What does this function do?

```
/**
 * This function multiplies two numbers in a very
 * inefficent way. It serves only as an example.
 */
public int multiply(int factor1, int factor2) {

     int product = 0;
     for (int n = 0; n < factor1; n++) {
        product += factor2;
     }
     return product;
}
```

A Text Classification Problem



Web Service classification

==

Text classification

Definitions

Category

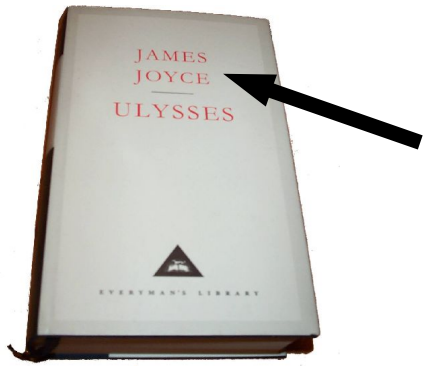
- Broad description of service as a whole
- e.g. e-commerce, weather, finance
- Profile hierarchy

Domain

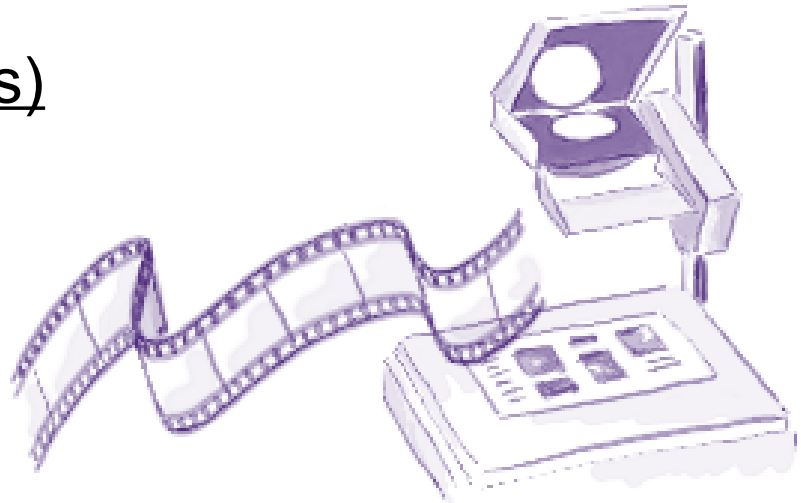
- Purpose of single operation
- e.g. query price, purchase book
- Atomic process

Datatype

- Meaning of single parameter
- e.g. author name, credit card number
- Property



- ✔ Introduction
- ✔ Our Machine Learning Approach
- 3. Classification Tasks
  - a) Domain & Datatype (Web Forms)
  - b) Category (Web Services)
- 4. Conclusion & Discussion



Assumption

Dependencies between Category⇔Domain⇔Datatype

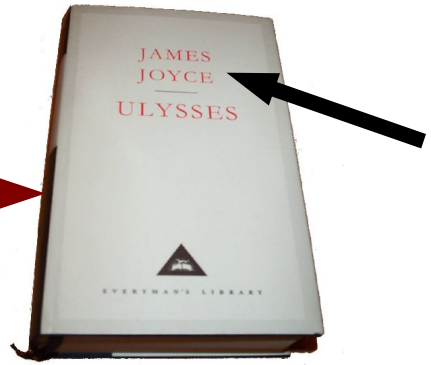
Category

Books

Domain

Query book price

Datatype



Book title

Assumption

Dependencies between Category⇔Domain⇔Datatype

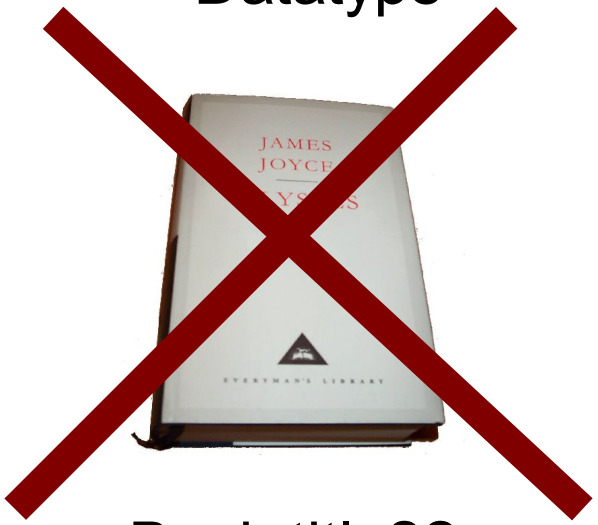
Category

Tea

Domain

Order tea

Datatype



Book title??

Assumption

Dependencies between Category⇔Domain⇔Datatype

Category

Tea

Domain

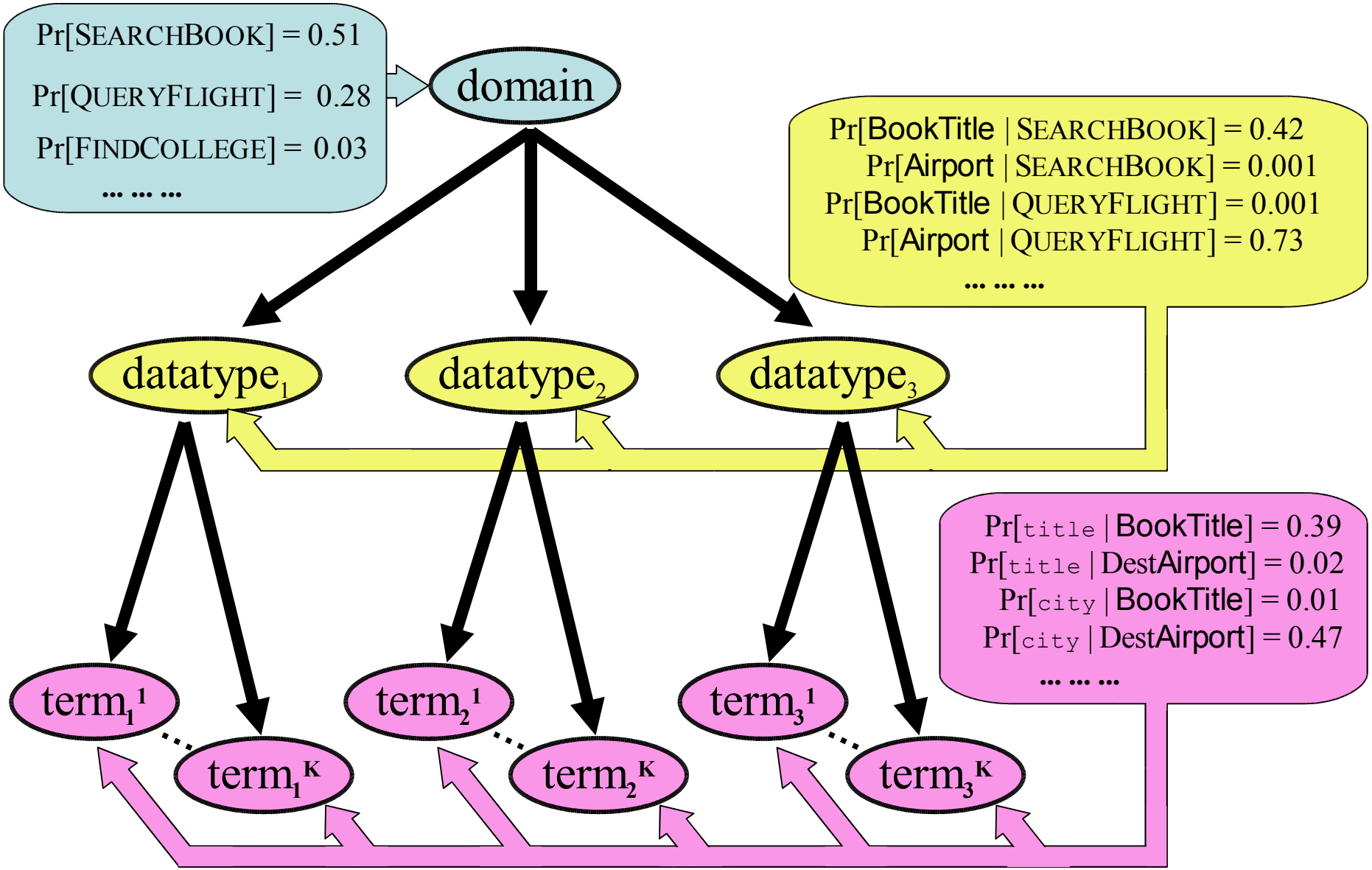
Order tea

Datatype

Credit card number

- Bayesian Networks
  - Nodes are random variables
  - Edges indicate conditional probabilites
  - Flow of evidence

Bayesian Networks



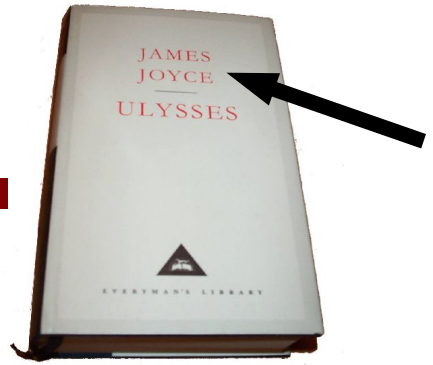
Flow of evidence

Dependencies between Category⇔Domain⇔Datatype

Category

Domain

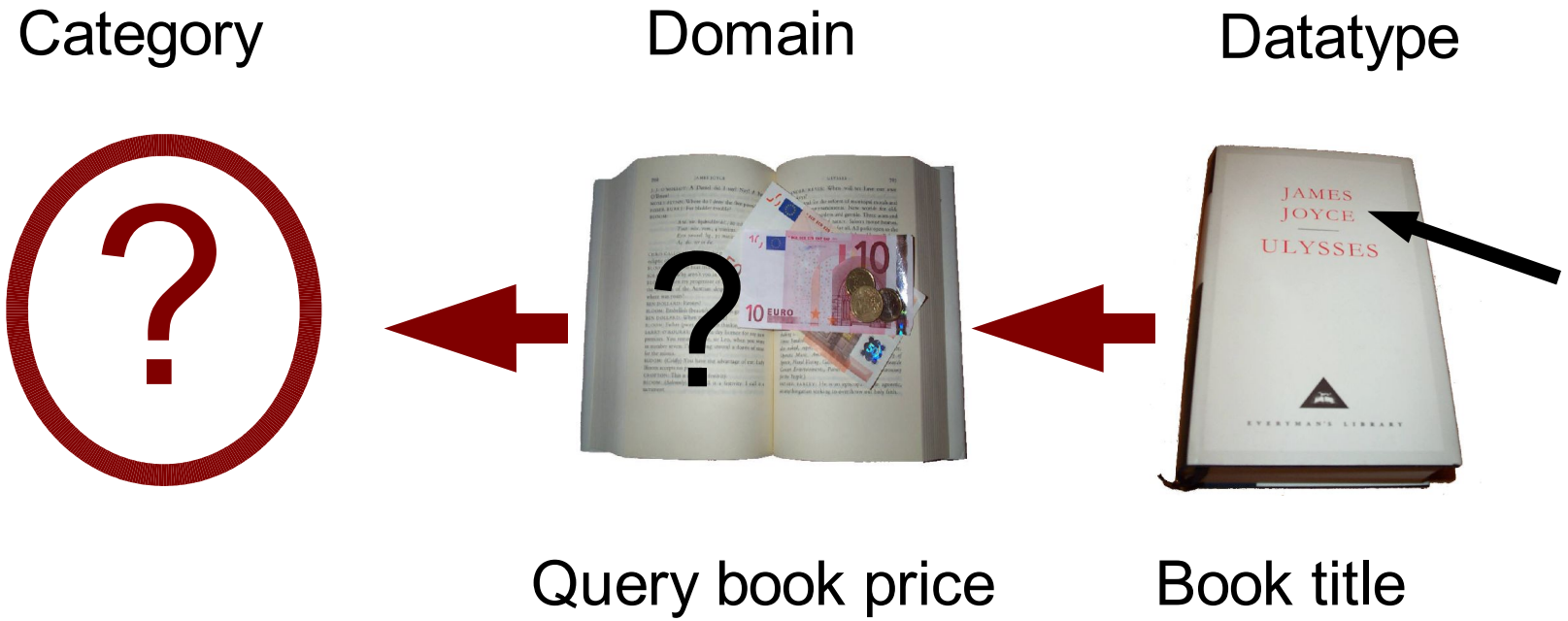
Datatype



Book title

Flow of evidence

Dependencies between Category⇔Domain⇔Datatype



Flow of evidence

Dependencies between Category⇔Domain⇔Datatype

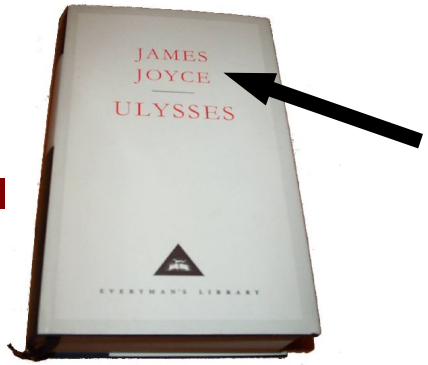
Category

Books

Domain

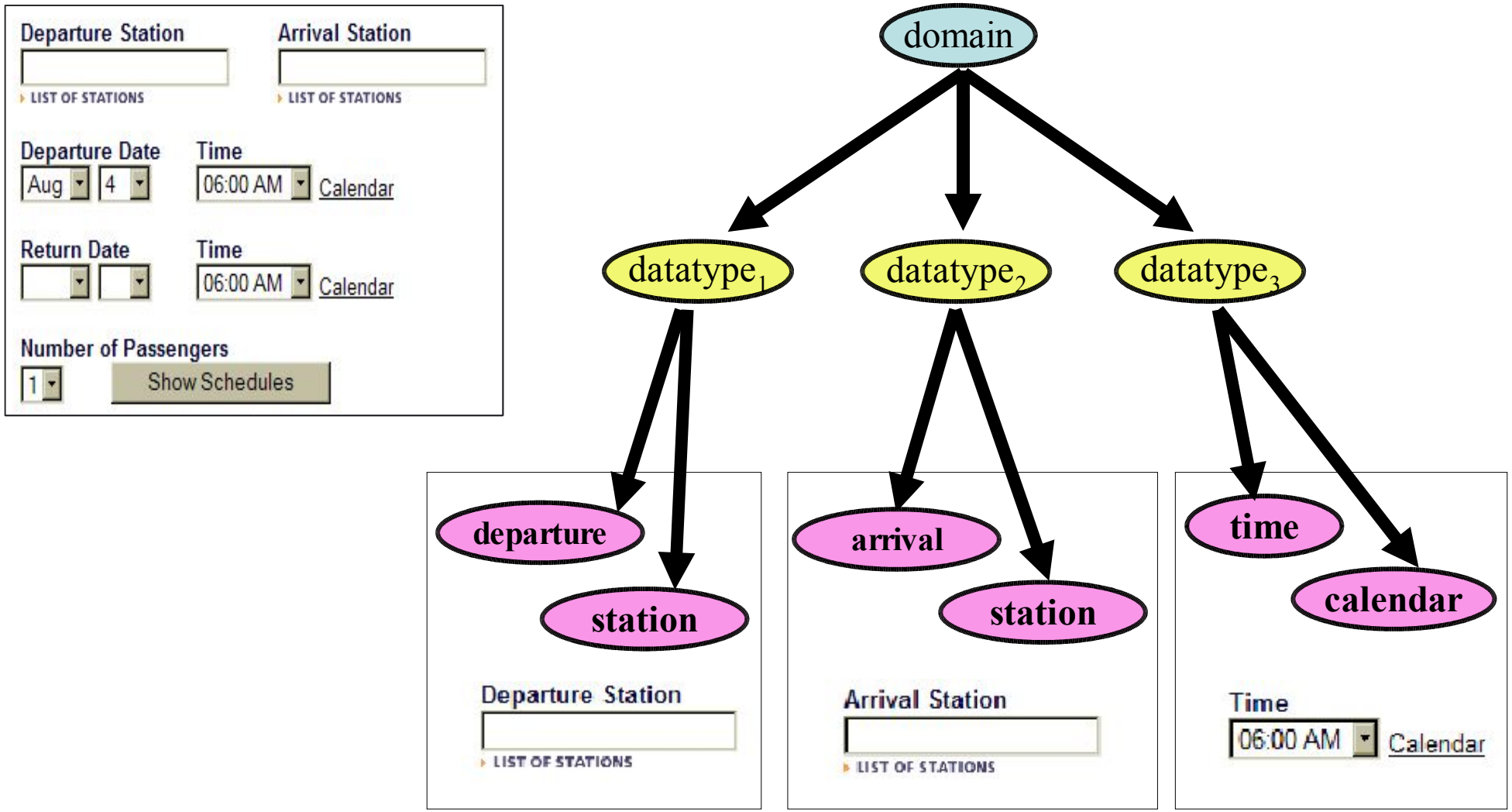
Query book price

Datatype



Book title

Web Forms


Departure Station
LIST OF STATIONS
Arrival Station
LIST OF STATIONS
Departure Date
Aug
4
Time
06:00 AM
Calendar
Return Date
Time
06:00 AM
Calendar
Number of Passengers
1
Show Schedules
domain
datatype1
datatype2
datatype3
departure
station
Departure Station
LIST OF STATIONS
arrival
station
Arrival Station
LIST OF STATIONS
time
calendar
Time
06:00 AM
Calendar

Evaluation

129 real Web forms

656 fields

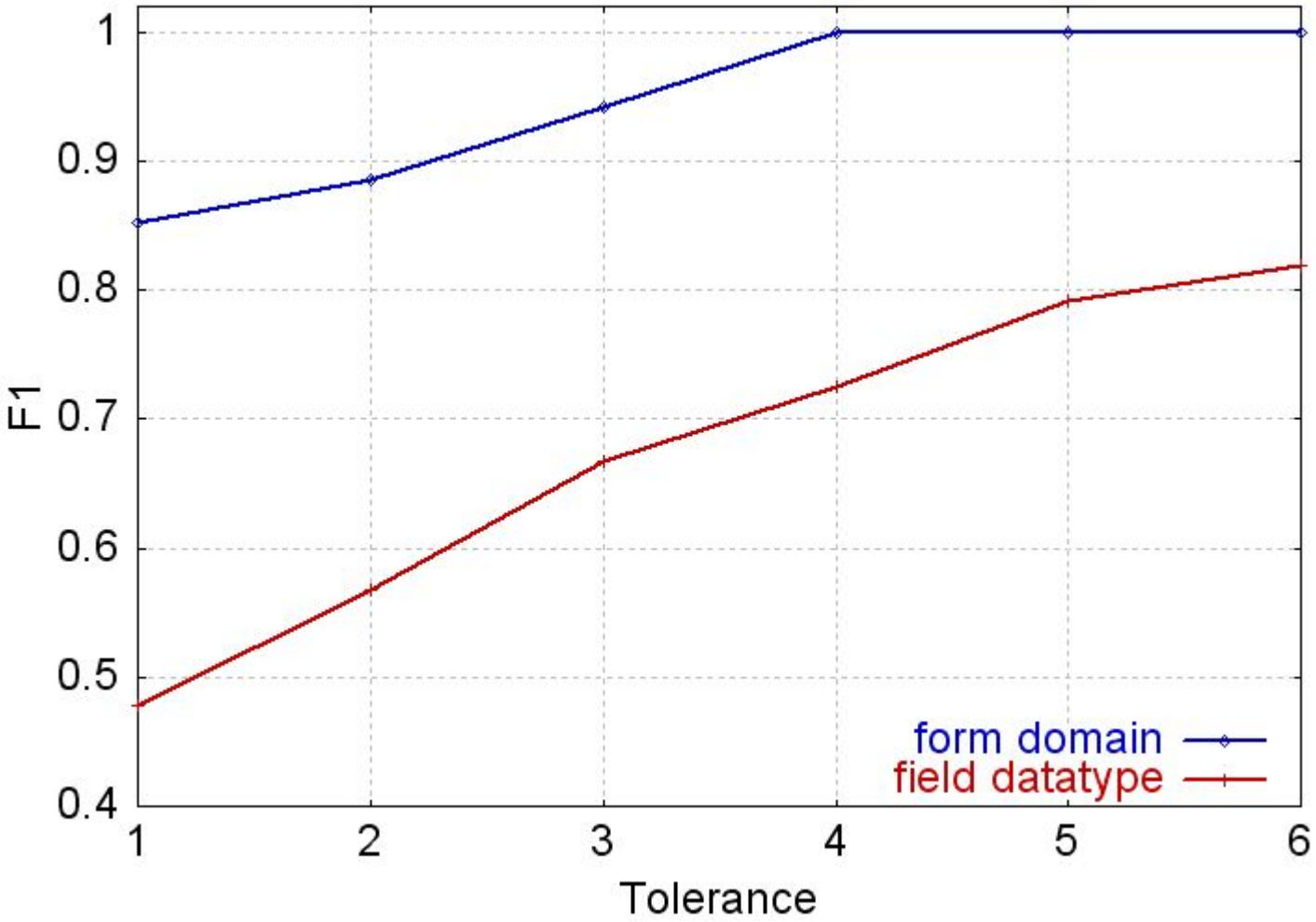
6 domains (number of services in domain)

Search Book (44)	Find College (2)	Search College Book (17
Query Flight (34)	Find Job (28)	Find Stock Quote (9)

72 datataypes (illustrative sample)

Book Details	Book Edition	Book Format	Book Search Type
Book Title	Number of Children	City	Class
College Subject	Company Name	Country	Currency
...	...	...	...
Ticker Symbol	Time	Time Return	Time Period
Travel Type	US State	ZIP	

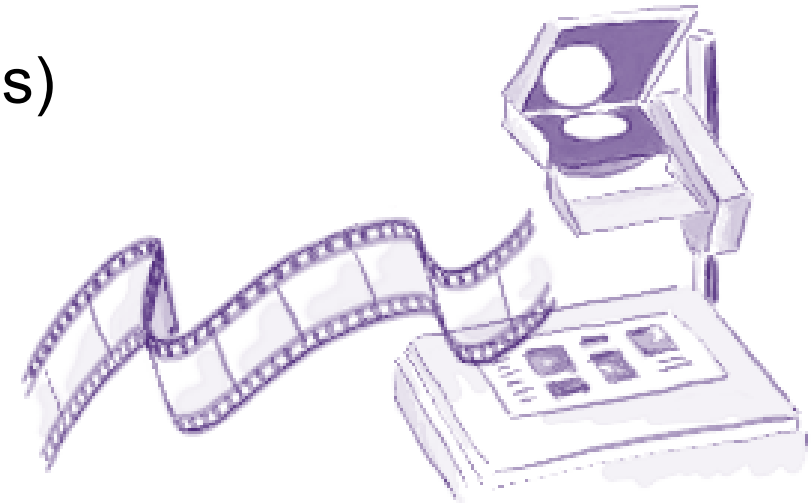
Evaluation


1
0.9
0.8
0.7
0.6
0.5
0.4
F1
1
2
3
4
5
6
Tolerance
form domain
field datatype

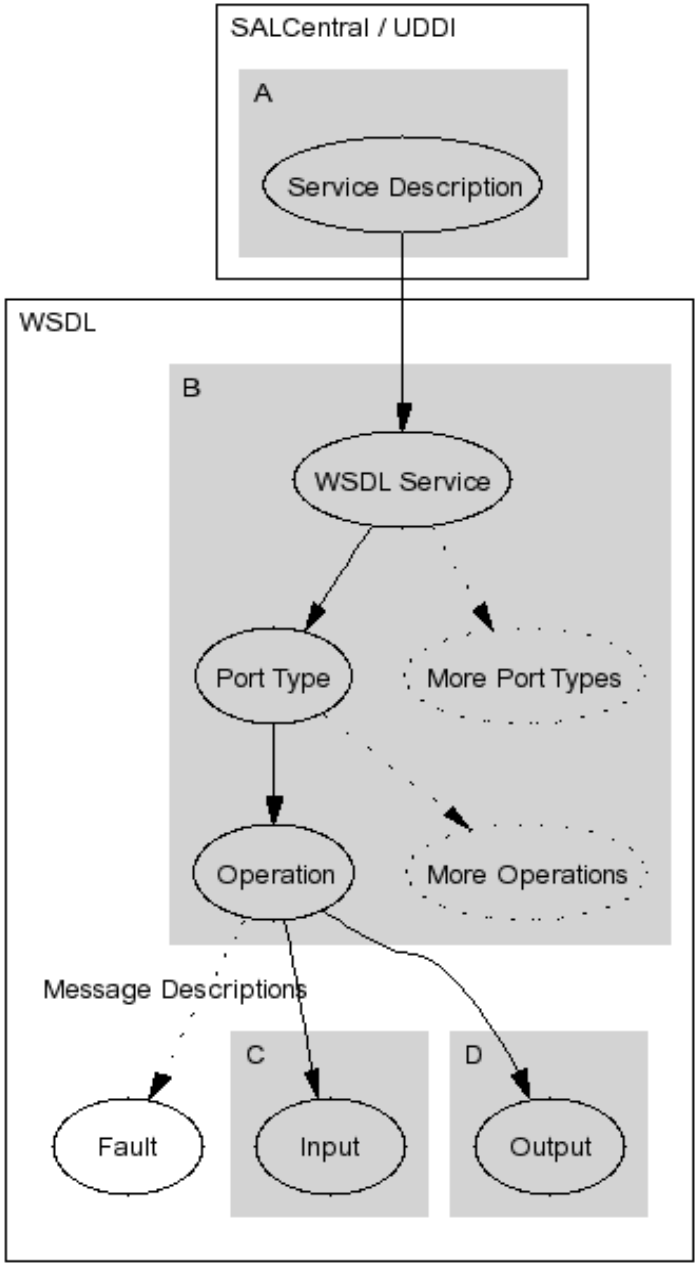
Web Forms?

- Why forms? Not talking about services?
  - Fully annotated Web Services not available at time of experiments
  - Web Services actually easier: HTML parsing causes noise
  - Next experiment will show real Web Services!

- ✔ Introduction
- ✔ Our Machine Learning Approach
- 3. Classification Tasks
  - ✔ Domain & Datatype (Web Forms)
  - b) Category (Web Services)
- 4. Conclusion & Discussion



A Text Classification Problem



- Text sources:

A) Service Description
(plain text, e.g. from UDDI)

B) WSDL:
service, portType, operation

C) WSDL: Input message

D) WSDL: Output message

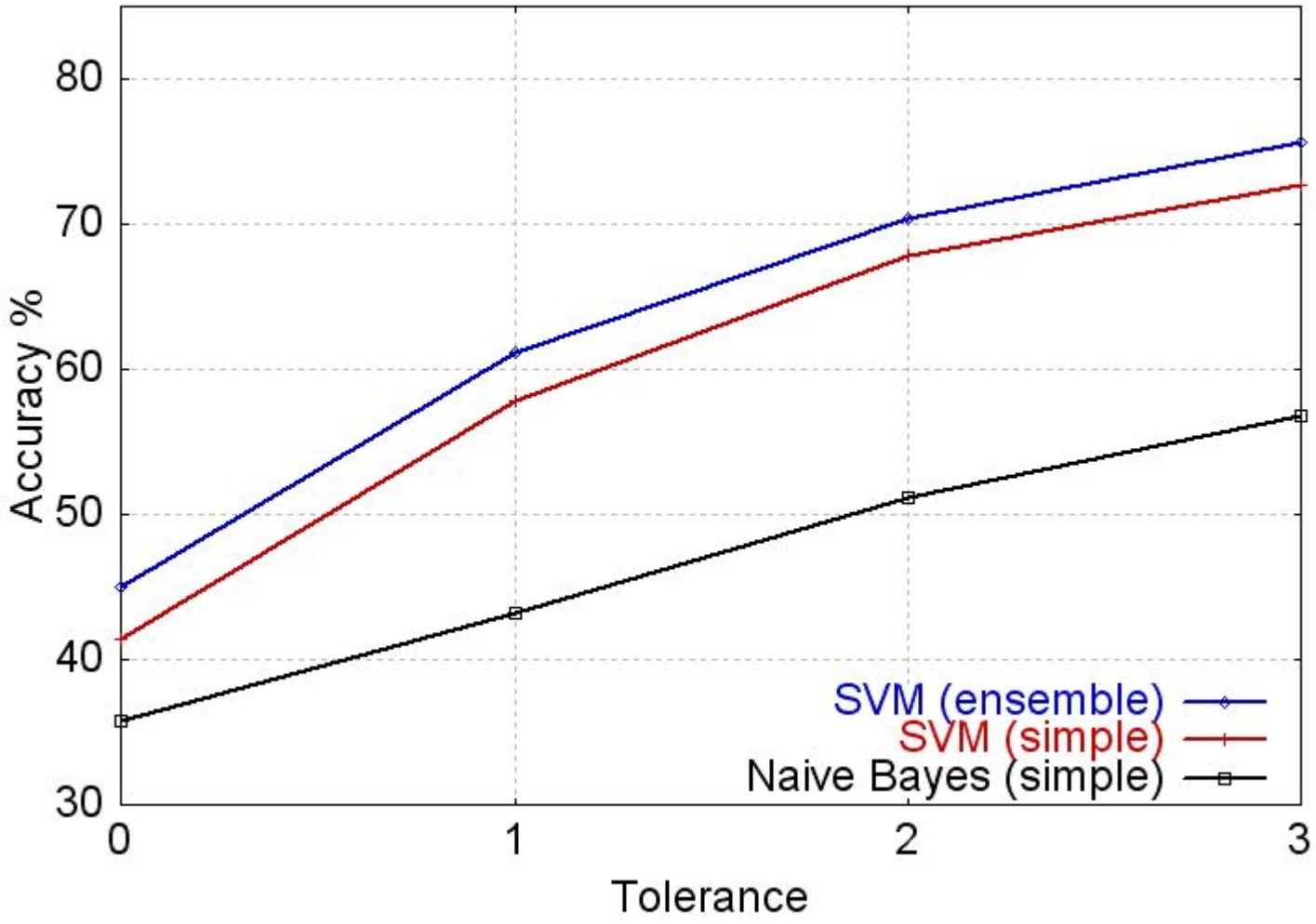
- Ensemble Learning
  - Each text source contains different words
    (e.g. operation “buyBook”, message part “author”)
  - Using seperate learners is more accurate

391 real Web Services

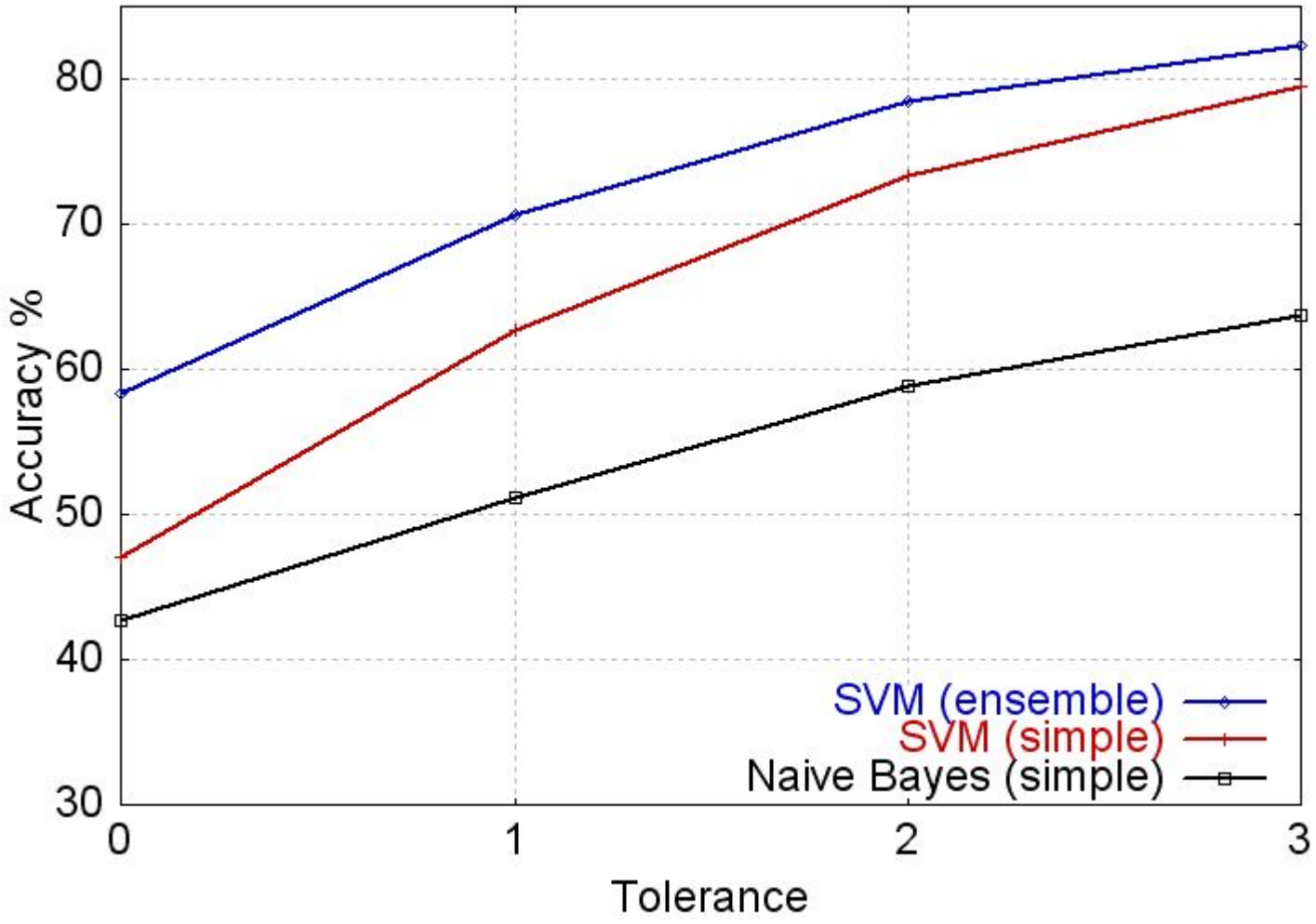
11 categories (number of services in category)

Business (22)
Country Information (62)
Games (9)
News (30)
Communication (44)
Developers (34)
Mathematics (10)
Web (39)
Converter (43)
Finder (44)
Money (54)
discarded (33)

WSDL only

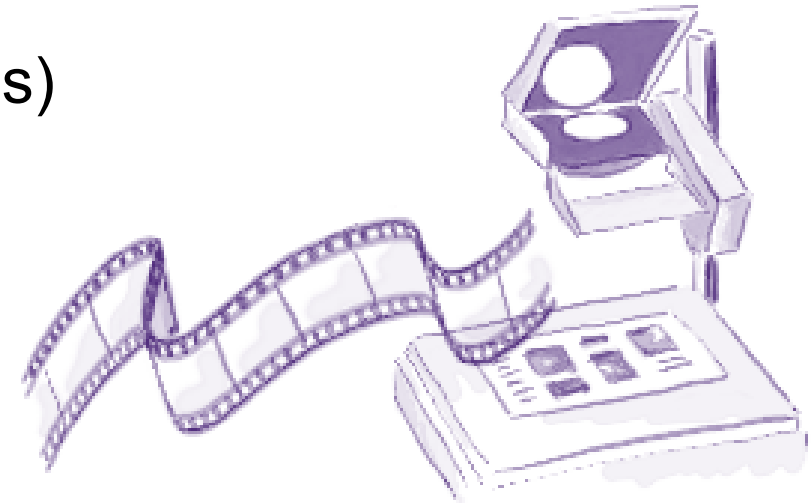

80
70
60
50
40
30
Accuracy %
0
1
2
3
Tolerance
SVM (ensemble)
SVM (simple)
Naive Bayes (simple)

WSDL plus plain text descriptions (easier)



- Clustering
  - Unsupervised – no training data needed
  - Algorithm discovers meaningful groups of services
- Evaluation
  - All clustering algorithms outperform baseline
  - See paper for details

- ✔ Introduction
- ✔ Our Machine Learning Approach
- ✔ Classification Tasks
  - ✔ Domain & Datatype (Web Forms)
  - ✔ Category (Web Services)
- 4. Conclusion & Discussion

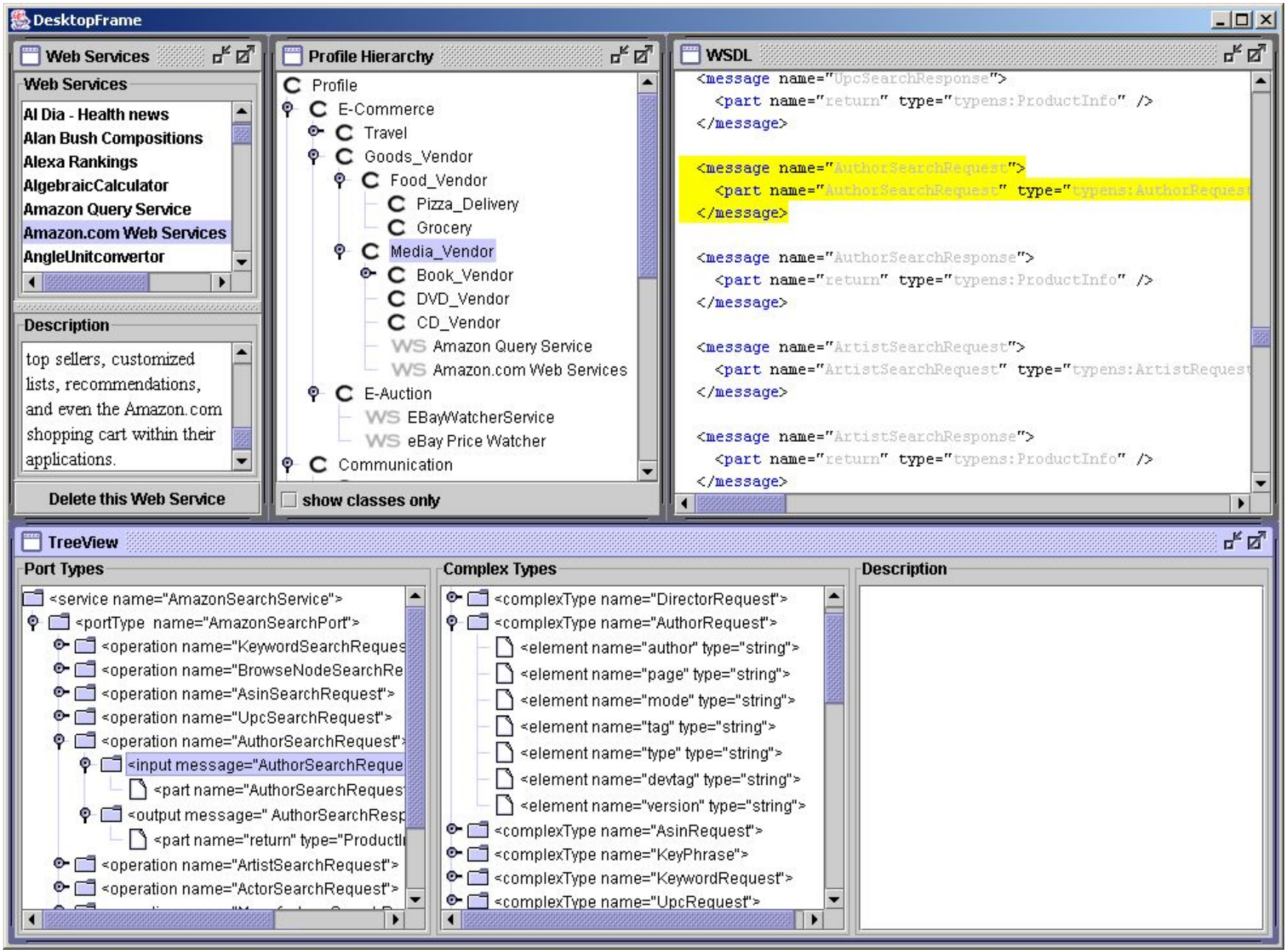


- Summary
  - Bayesian inference algorithm for classifying forms
  - Ensemble learning for classifying Web Services
  - Clustering of Web Services

Open Issues & Limitations

- Open issues
  - Larger Web Service corpus ⇨ training data
  - Predictions not good enough yet
- Limitations: We do not handle...
  - Process Model
  - Compostion & Workflow

Current & Future Work



Discussion & Questions

- We are gathering annotated Web Services
- Contact us to use them, or to contribute:
  - (Partially) annotated Web Services in large numbers?
  - Global / Shared Profile Hierarchy?
- Questions?