Information Systems Group

Hasso Plattner Institut | Universität Potsdam

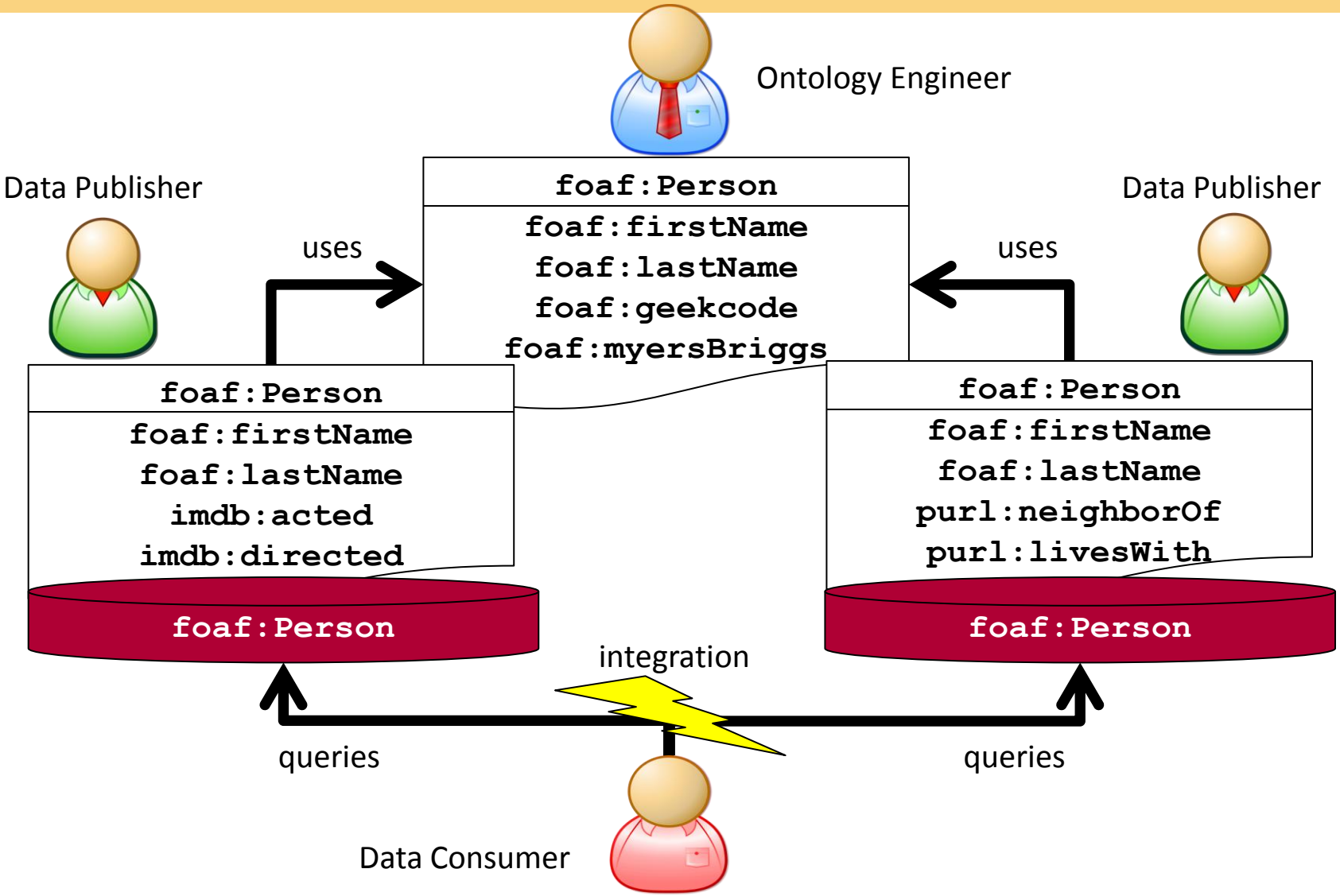
RDF Ontology (Re-)Engineering through large-scale Data Mining

Billion Triple Challenge 2011

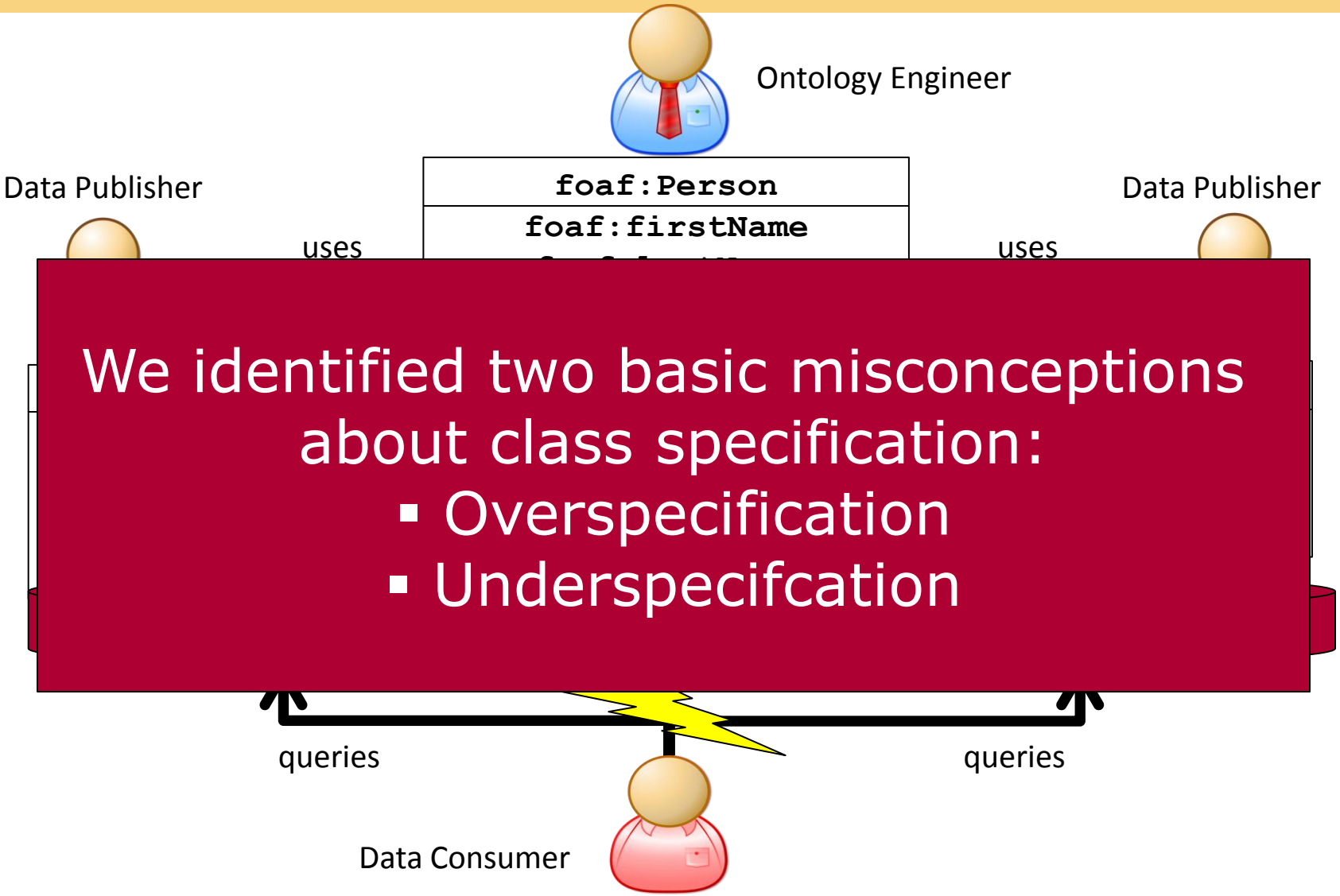
Johannes Lorey
Ziawasch Abedjan
Felix Naumann
Christoph Böhm

http://tinyurl.com/hpi-btc2011

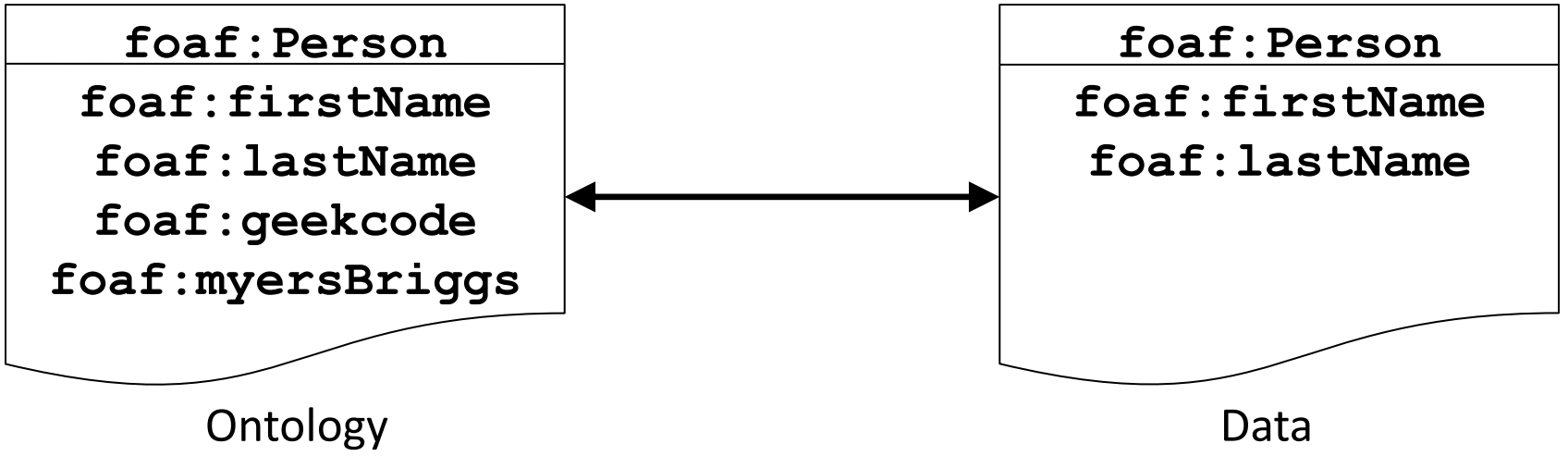
Motivation


Ontology Engineer
Data Publisher
Data Publisher
uses
uses
foaf:Person
foaf:firstName
foaf:lastName
foaf:geekcode
foaf:myersBriggs
foaf:Person
foaf:firstName
foaf:lastName
imdb:acted
imdb:directed
foaf:Person
foaf:Person
foaf:firstName
foaf:lastName
purl:neighborOf
purl:livesWith
foaf:Person
integration
queries
queries
Data Consumer

Motivation

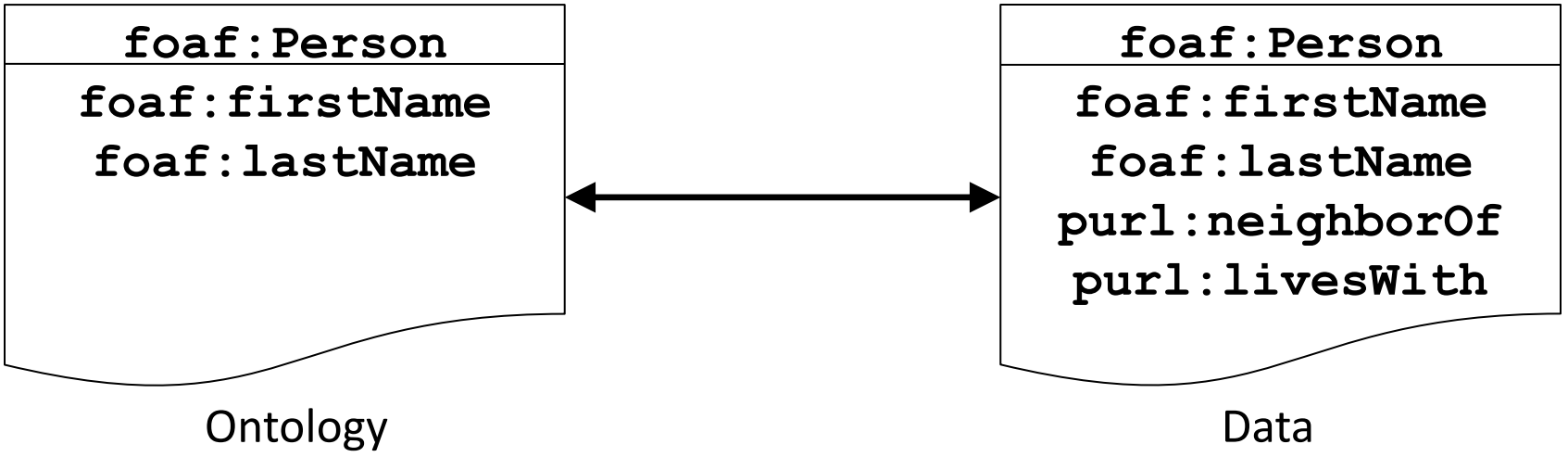

Ontology Engineer
Data Publisher
foaf:Person
foaf:firstName
Data Publisher
uses
uses
We identified two basic misconceptions about class specification:
▪ Overspecification
▪ Underspecifcation
queries
queries
Data Consumer

Overspecification



- Some properties are specified in the ontology, but rarely used in instance data
- Reasons include:
  - Data publishers are unaware of specified attributes
  - Attributes are ambiguous/unfitting/omissible

Underspecification

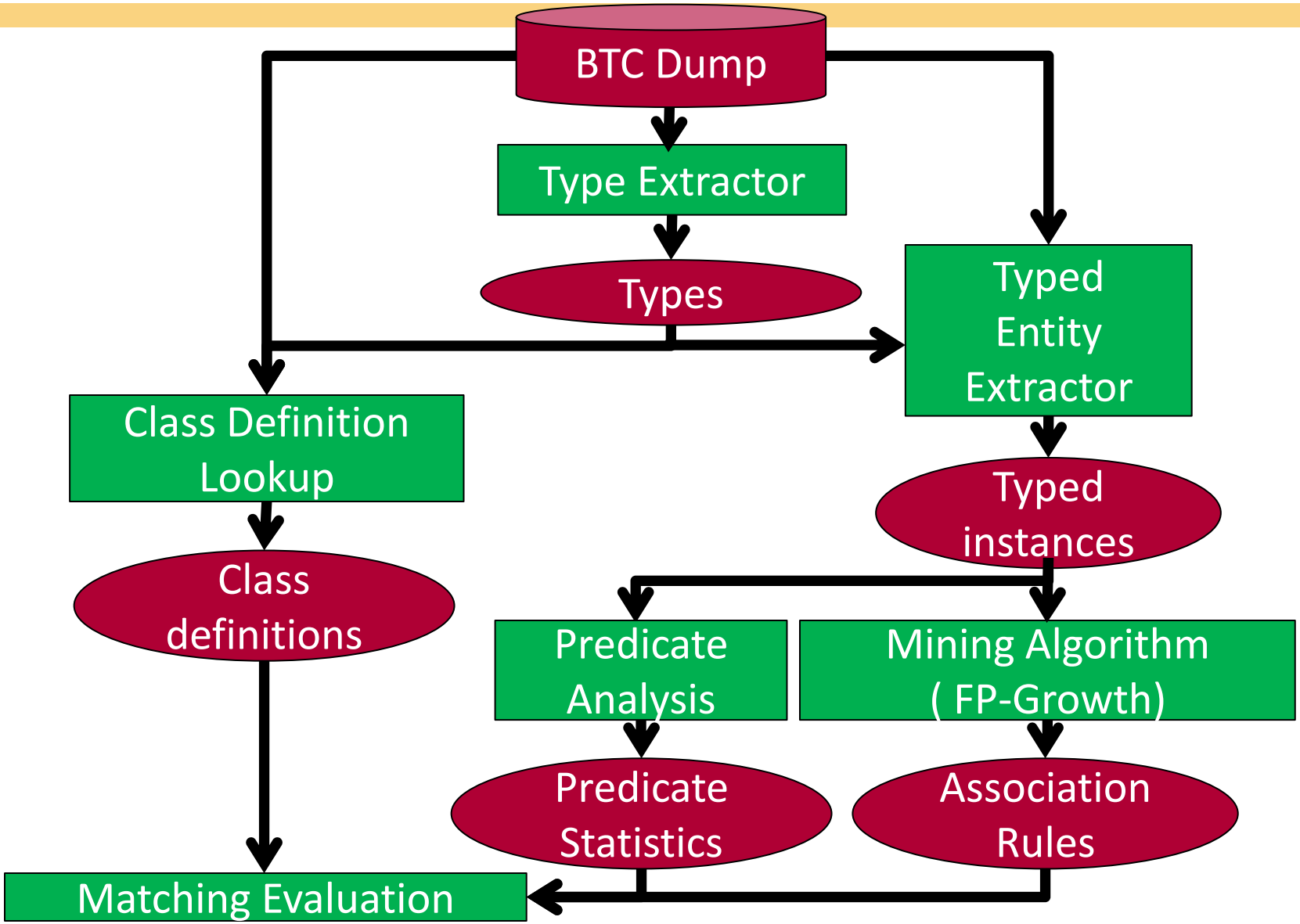


- Some properties occur frequently in instance data, but are not specified in the ontology
- Reasons include:
  - Data publishers introduce new attributes (although existing ones might be suitable)
  - Ontology definition lacks properties

An instance-data driven approach

- Perform **predicate frequency** analysis
- Perform **rule mining** to identify usage patterns
  - Positive rules: predicates co-occur frequently (**`foaf:gender` ⇒ `foaf:weblog`**)
  - Negative rules: predicates occur exclusive (**`foaf:weblog` ⇒ ¬ `foaf:homepage`**)
- **Match against class definition** to offer re-engineering suggestions such as
  - Adding new properties to class definition
  - Removing properties from class definition
  - Creating subclasses and pushing properties down

Workflow using BTC data



The data

- around **475GB** of uncompressed data
- **213,382** RDF types
- **441,461,669** typed instances
- type definitions for around **90%** of all instances disovered within BTC dataset
- we ran our approach on the most frequent types

RDF Type Usage in the BTC dataset

Type	#Instances
foaf:Person	362,590,928
foaf:OnlineAccount	2,938,416
foaf:Document	1,252,681
rdf:Statement	887,363
foaf:Image	876,863
mo:MusicArtist	310,529
foaf:Agent	204,435

- 3,893 „new properties“ found for foaf:Person)

foaf:nick	361,937,602	Status:testing
foaf:member_name	19,083,000	Undefined property
foaf:tagLine	19,062,451	Undefined property
foaf:Image	18,033,515	Misusage
foaf:name	3,597,768	Most common “name” property

RDF Type Usage in the BTC dataset

Type	#Instances
foaf:Person	362,590,928
foaf:OnlineAccount	2,938,416
foaf:Document	1,252,681
rd	
fo	
m	
fo	

The BTC dataset provides a large-scale heterogenous snapshot of the Web of Data and reveals **common** (mis-)use patterns.

foa		ting
foa		erty
foaf:tagLine	19,062,451	Undefined property
foaf:Image	18,033,515	Misusage
foaf:name	3,597,768	Most common “name” property

(Re-)Engineering overspecified class definitions

Resource	#Instances
foaf:Person	362,590,928
• foaf:myersBriggs	9
• foaf:geekcode	7
• foaf:plan	1

- Data publishers do not provide suitable values for these properties
- Data consumers will have no luck in trying to query for these properties

➢ Remove the properties from this class definition (or possibly move them to subclasses)

(Re-)Engineering underspecified class definitions

Resource	#Instances
foaf:OnlineAccount	2,938,416
• foaf:accountName	2,859,090
• http://rdfs.org/sioc/ns#account_of	2,411,233

- Data publishers add more properties (and information) to a predefined class
- Data consumers considering the original class will retrieve unexpected information
- However, blindly adding this property to the class definition might result in fragmentation

➢ Consider mined rules for re-engineering suggestions

(Re-)Engineering underspecified class definitions

Resource	#Instances
foaf:OnlineAccount	2,938,416
• foaf:accountName	2,859,090
• http://rdfs.org/sioc/ns#account_of	2,411,233

Mining results:

`foaf:accountName ⇒ http://rdfs.org/sioc/ns#account_of`

- Originally specified property co-occurs frequently with new unspecified property
- Add new property to ontology

(Re-)Engineering underspecified class definitions

Resource	#Instances
mo:MusicArtist	310,529
• musicbrainz:isInstrumentalArtistOf	4,015
• musicbrainz:isEngineerOf	1,594
• musicbrainz:isMixEngineerOf	1,223

Mining results:

musicbrainz:isInstrumentalArtistOf ⇒
¬ musicbrainz:isEngineerOf

musicbrainz:isInstrumentalArtistOf ⇒
¬ musicbrainz:isMixEngineerOf

- Pairwise negative rules for frequent attributes
- Split up class into disjunct subclasses

Summary

- We found certain divergence types between class definitions and usage patterns
- We propose predicate frequencies and association rules as means to suggest possible alterations

- Goals:
  - Encourage ontology reengeneering
  - Better (typed) data for consumers