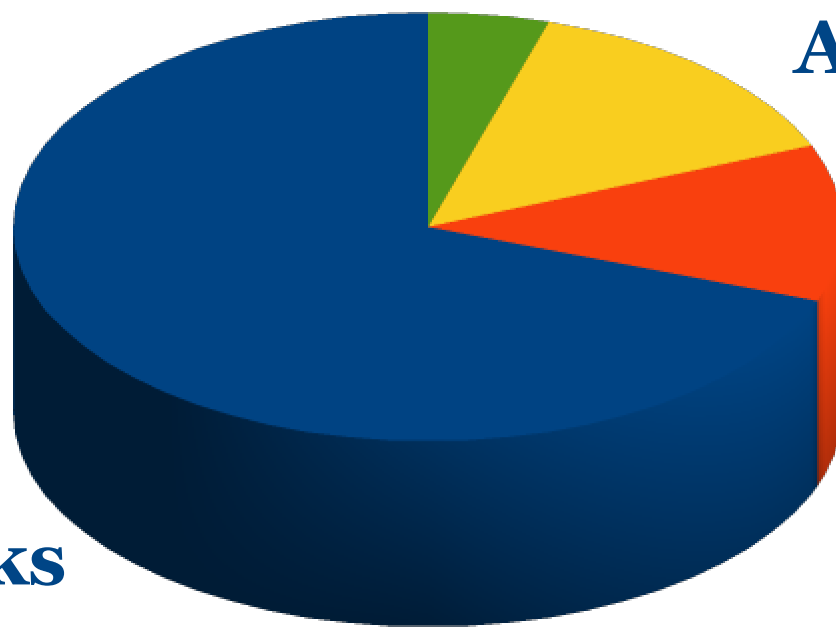
Transforming Data into Knowledge

using

Knowledge Engineering & Machine Learning

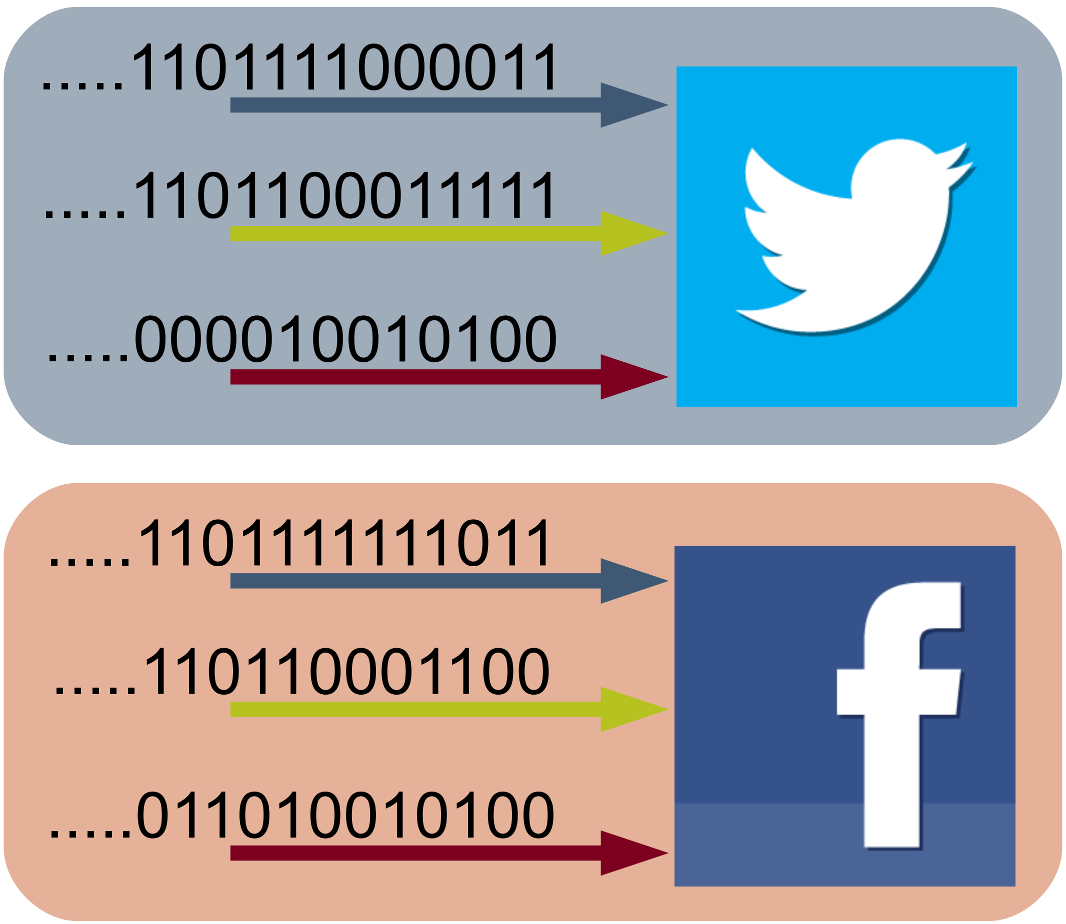
Presented by:
Atif Khan
infoTrellis

Scope


Tools
InfoTrellis ALLSight
DBpedia
Building Blocks
ontology
inference
machine learning

Motivation

Data Streams » Information?



Motivation

Information » Knowledge

Charles Smith @profcharles
Just finished registering for **KDD 2012** in **Beijing** http://**www.kdd.org/kdd2012/**

Charles Smith shared a link
I am off to **knowledge discovery and data mining conference** in **Beijing**. Looking forward to **Michael Jordan**'s **keynote address**

Motivation

Information » Knowledge

Charles Smith

Charles Smith

Michael Jordan

KDD2012

Knowledge Discovery and Data Mining

URL
www..kdd2012/

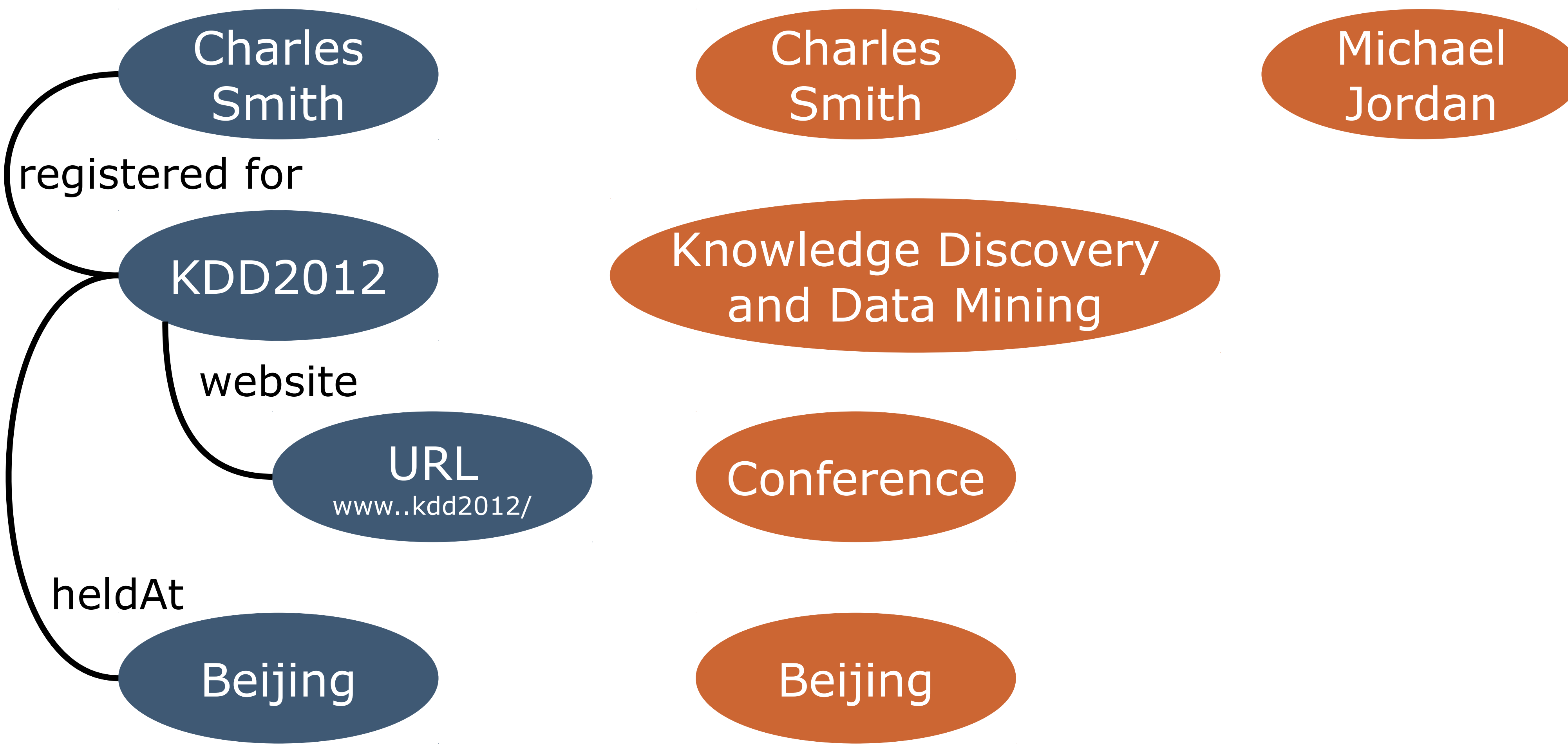
Conference

Beijing

Beijing

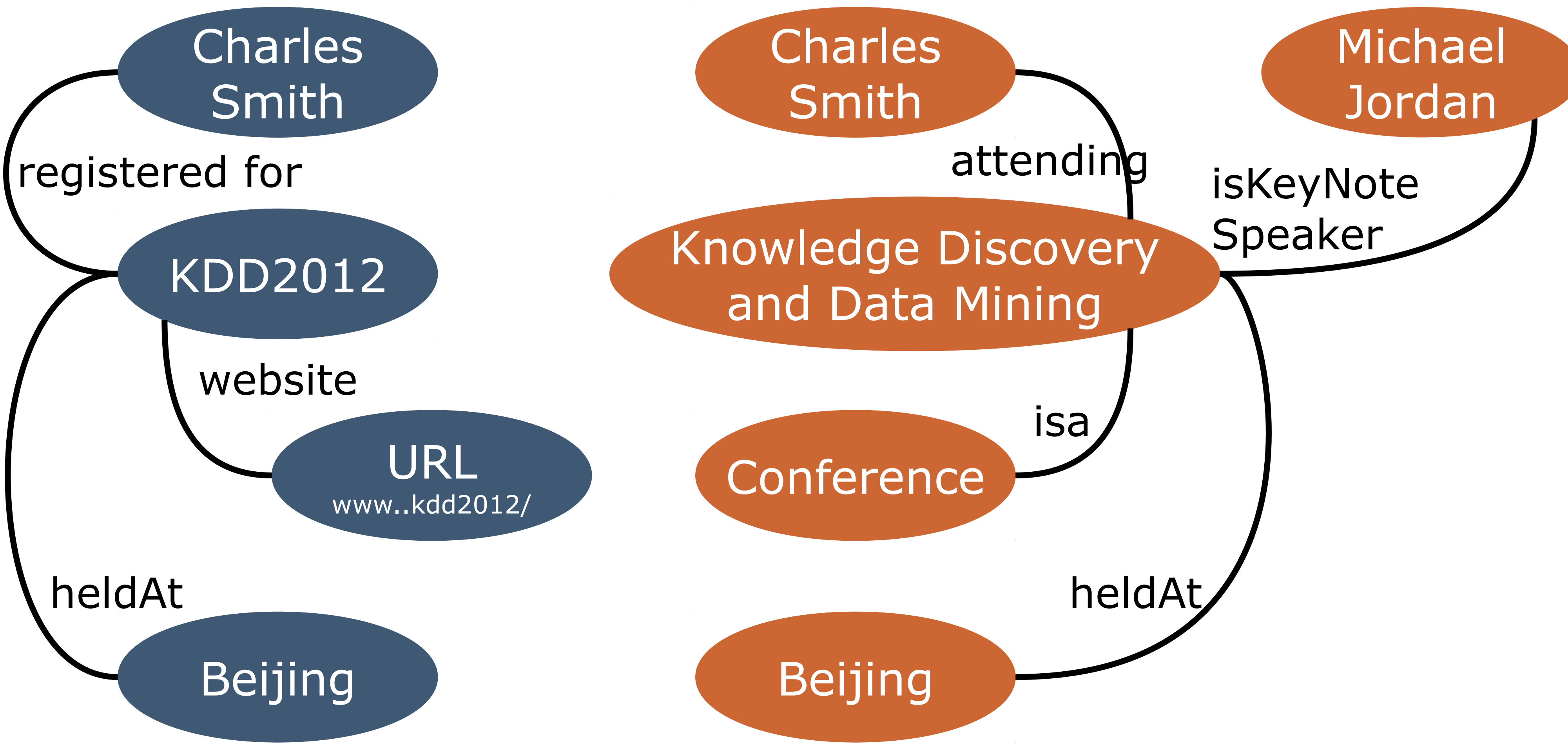
Motivation

Information » Knowledge



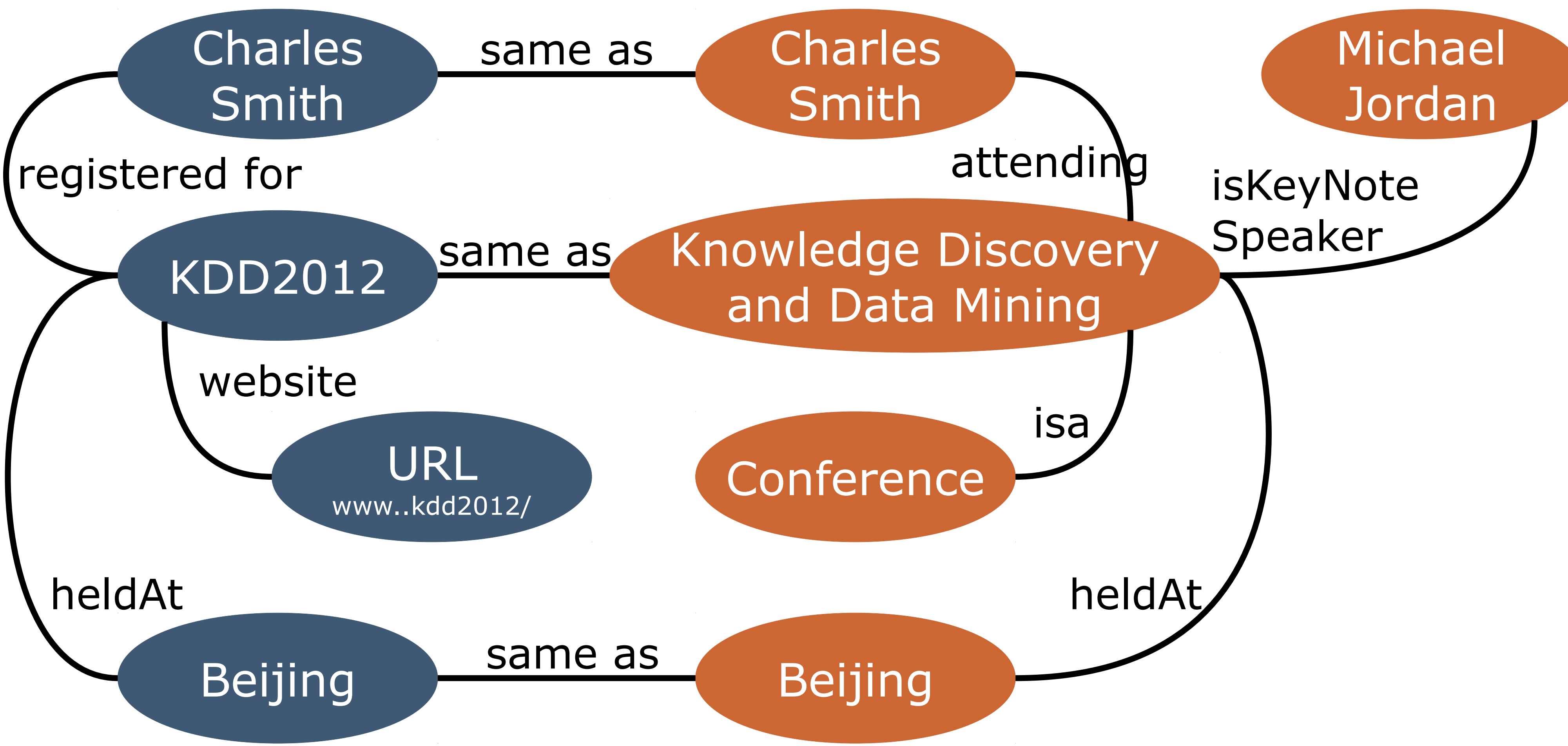
Motivation

Information » Knowledge



Motivation

Information » Knowledge

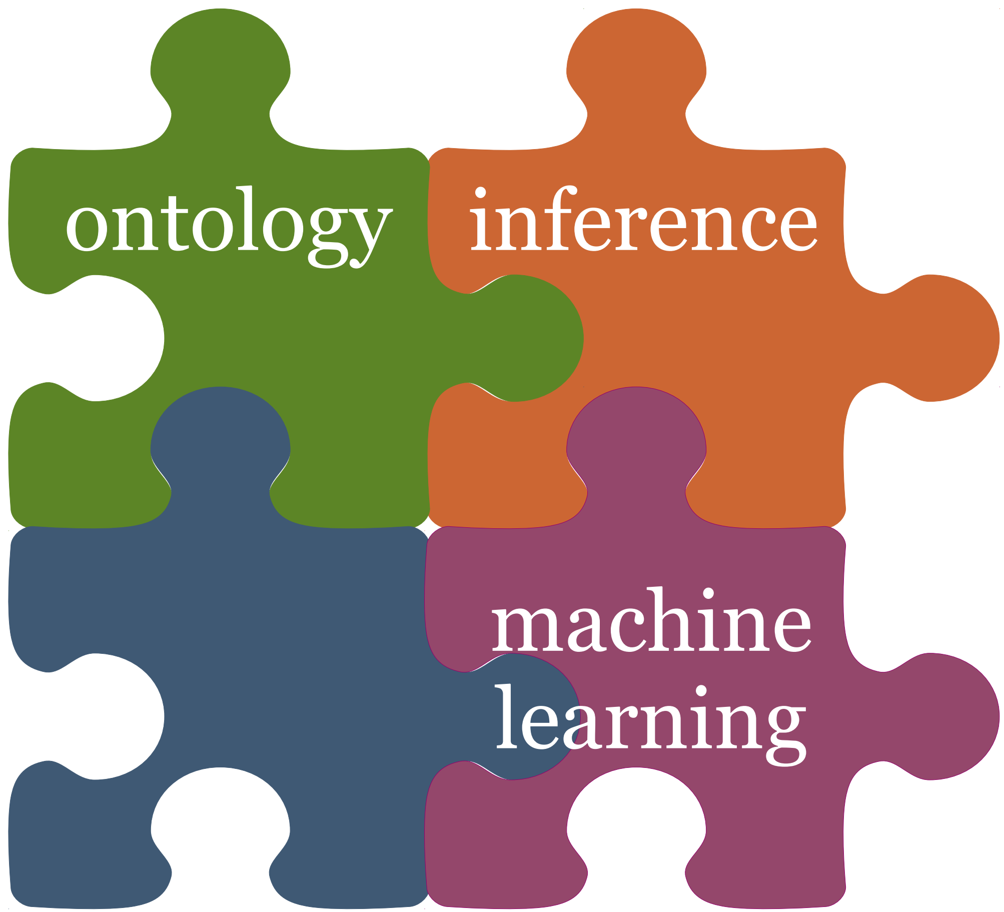


Motivation

Information » Knowledge

- Charles Smith is an academic
- KDD is a conference about
  data mining and *knowledge discovery*
- Michael Jordan is an influential
  academic in data mining community
- Charles Smith and Michael Jordan
  will both be in Beijing during KDD 2012

Building Blocks



Hybrid MDSS - Holmes**

Holmes

- a hybrid medical decision support system (MDSS)

Based on

- ontological knowledge representation
- logic-based inference
- machine learning to deal with noise

Atif Khan, John Doucette, Robin Cohen. *"Validation of an Ontological Medical Decision Support System for Patient Treatment Using a Repository of Patient Data: Insights into the Value of Machine Learning".* ACM Transactions on Intelligent Systems and Technology (TIST) 2012.

Hybrid MDSS

Line of Inquiry

- *which patients can be prescribed what sleep medications?*

Considerations

- patient-centric & evidence-based
- automated
- easy to explain and validate
- noise tolerant

Hybrid MDSS

Datasets

CDC–Behavioral Risk Factor Surveillance System (BRFSS) – 2010

Hybrid MDSS

Datasets

<u>multi-dimensional</u>
a. demographic information
b. medical information
c. behavioural Information

Hybrid MDSS

Datasets

<u>Characteristics</u>

a. 450K+ individuals
b. 400+ attributes/record
c. high "***missingness***"
d. numeric coding

Hybrid MDSS

Datasets

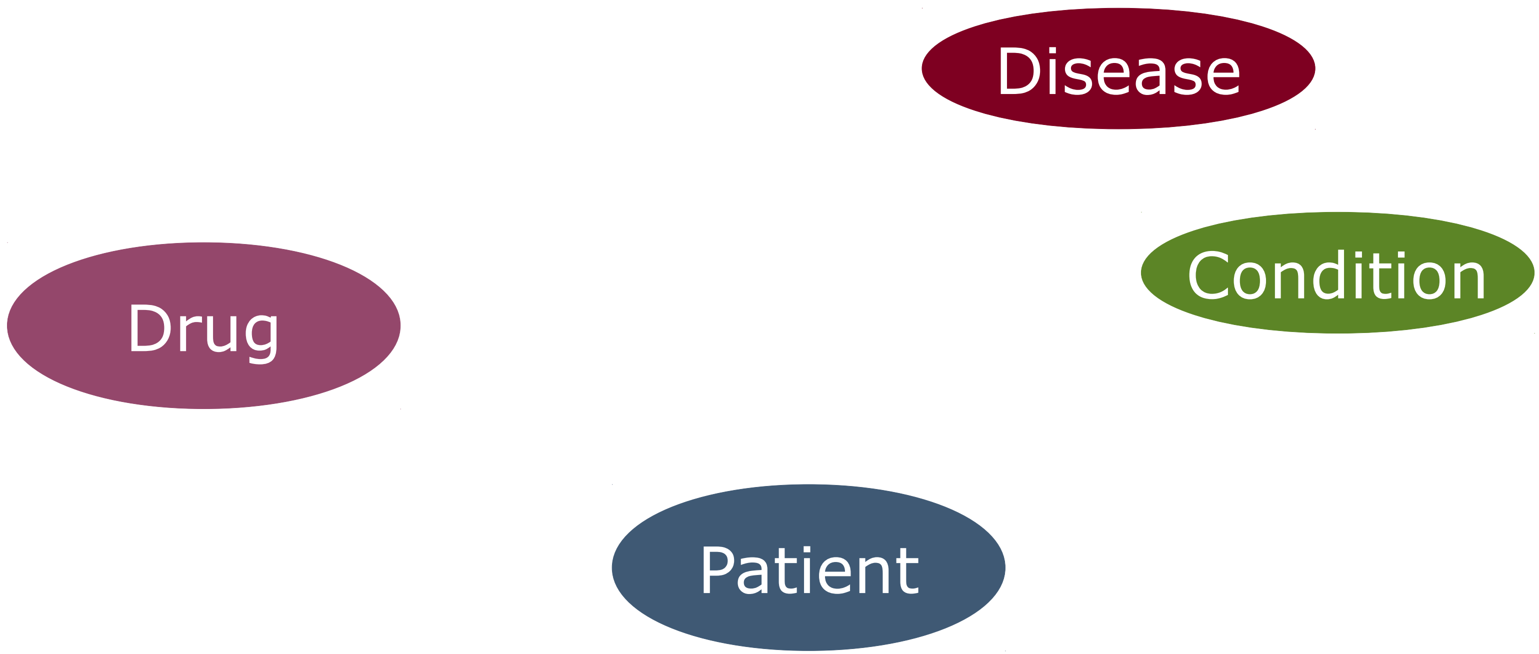
sleeping pill
prescription protocol
(*HTML*)

Hybrid MDSS

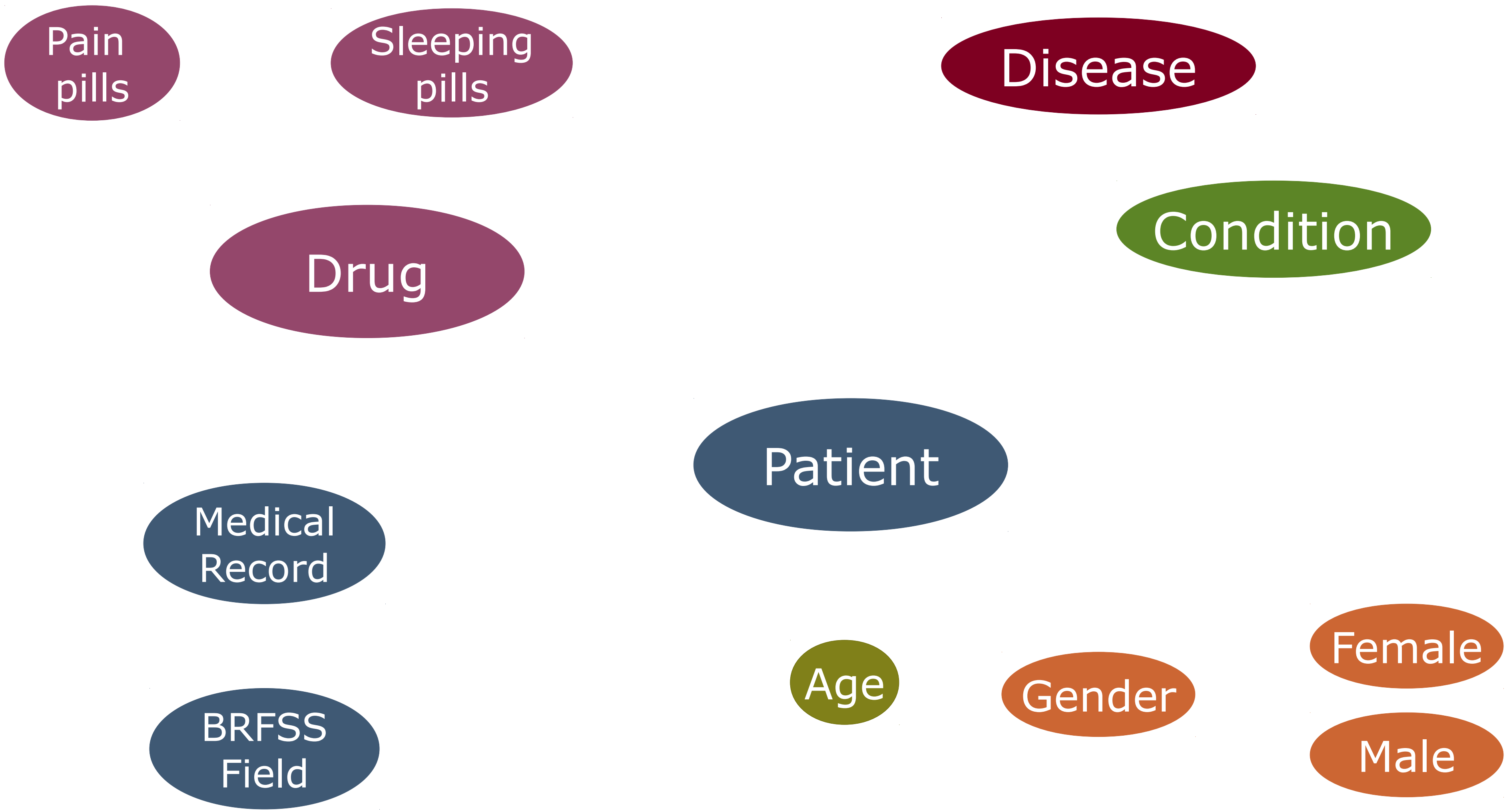
Datasets

drug-to-drug interactions *(HTML)*

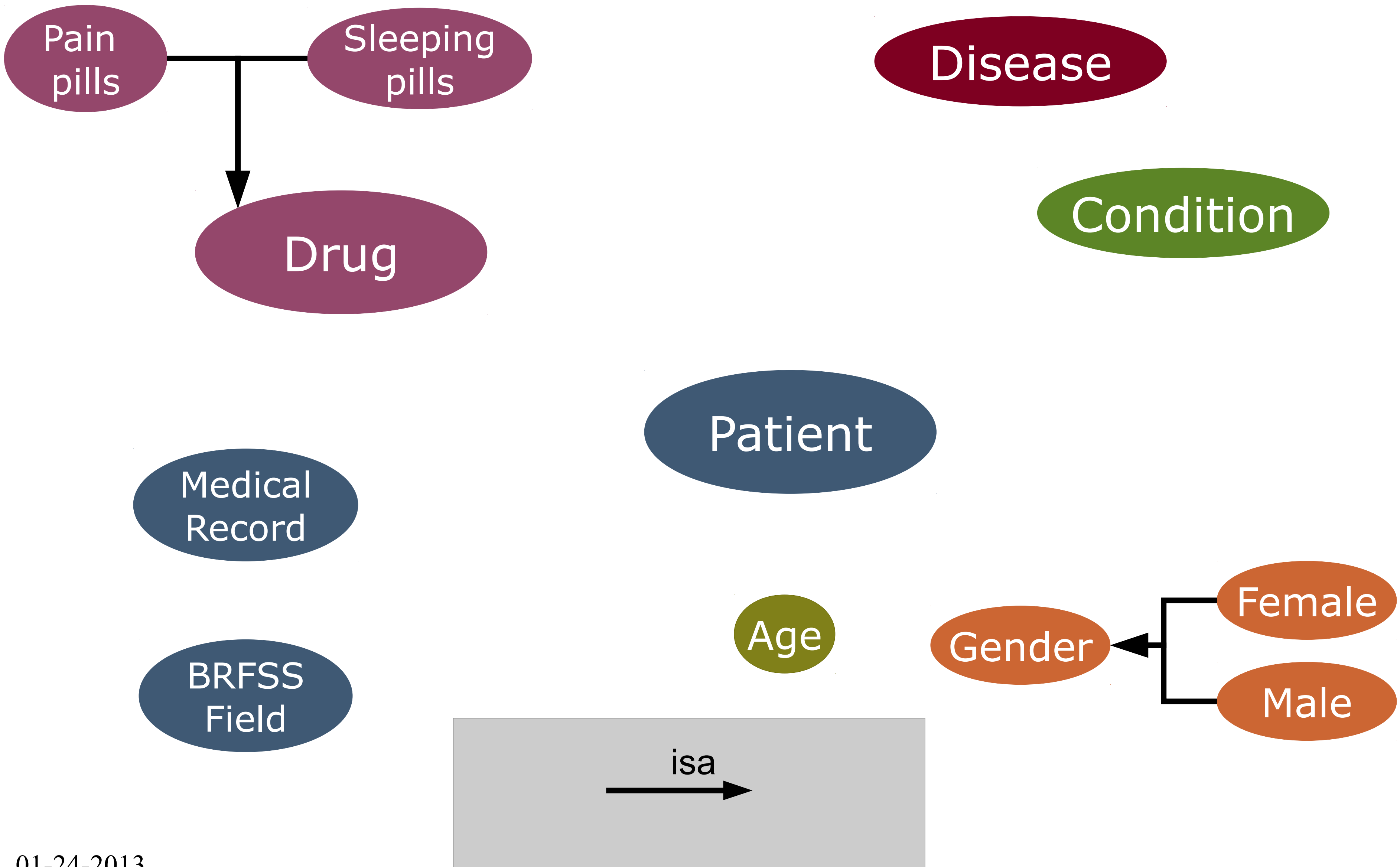
Ontology Model



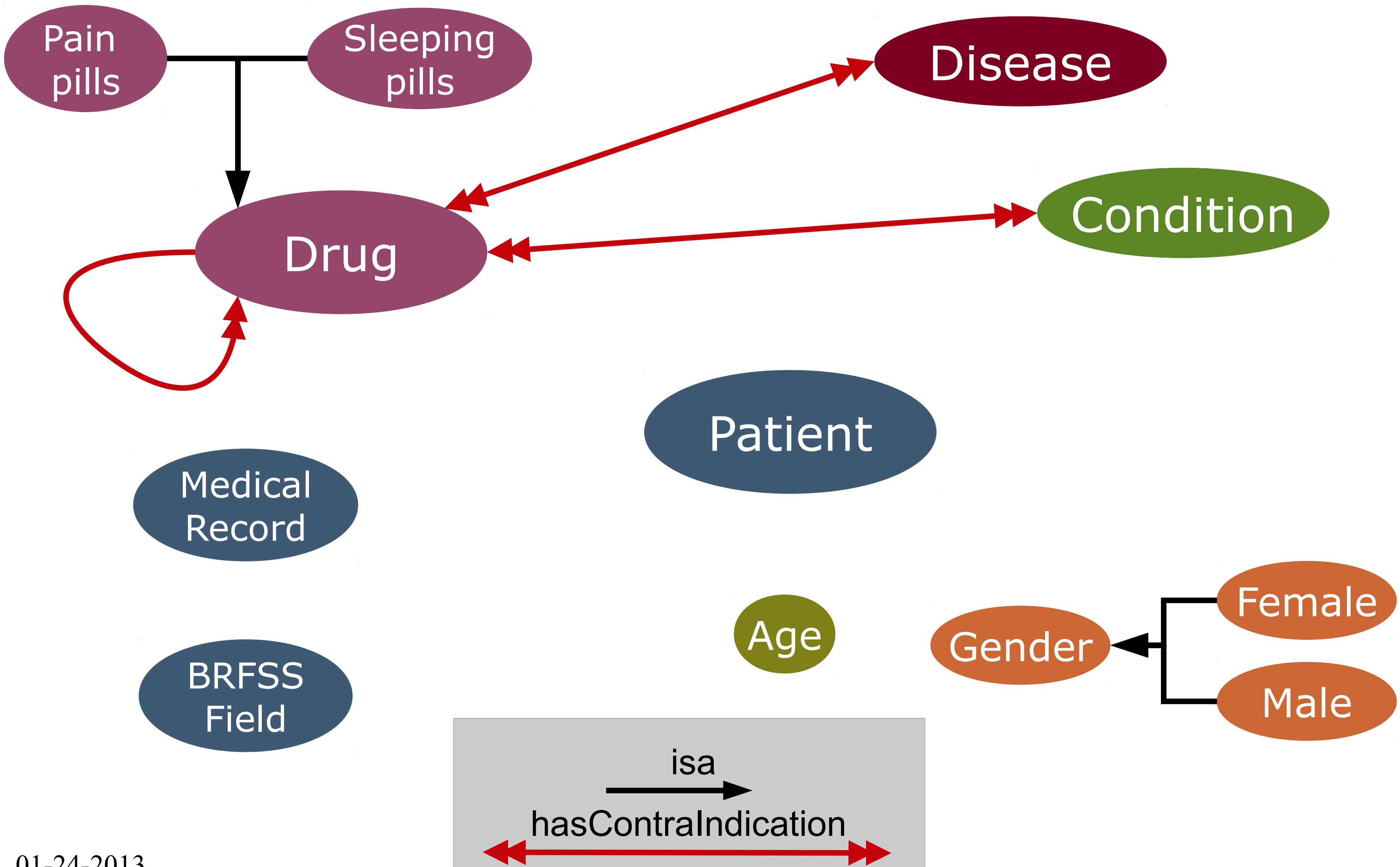
Ontology Model



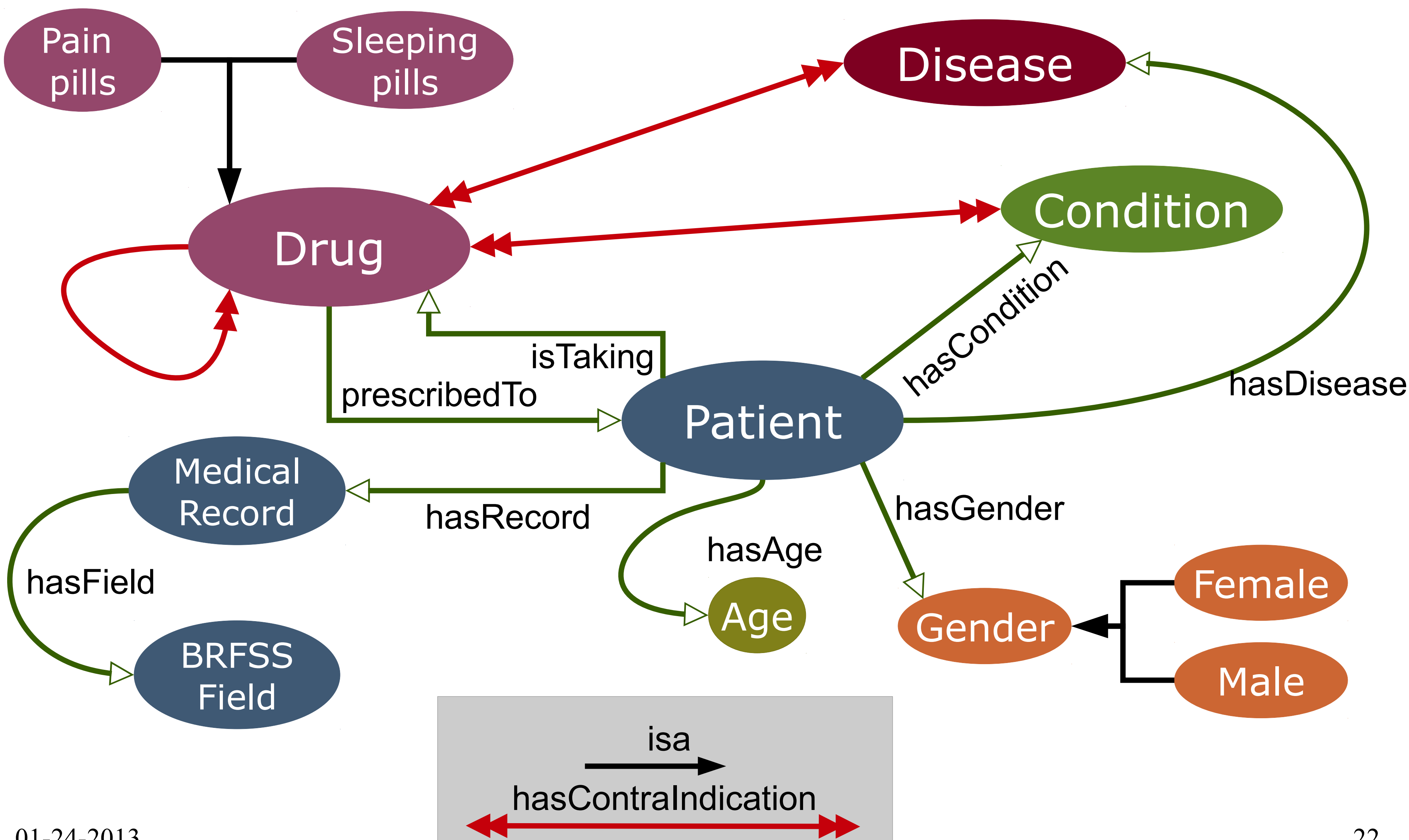
Ontology Model



Ontology Model



Ontology Model



An Ontology Model

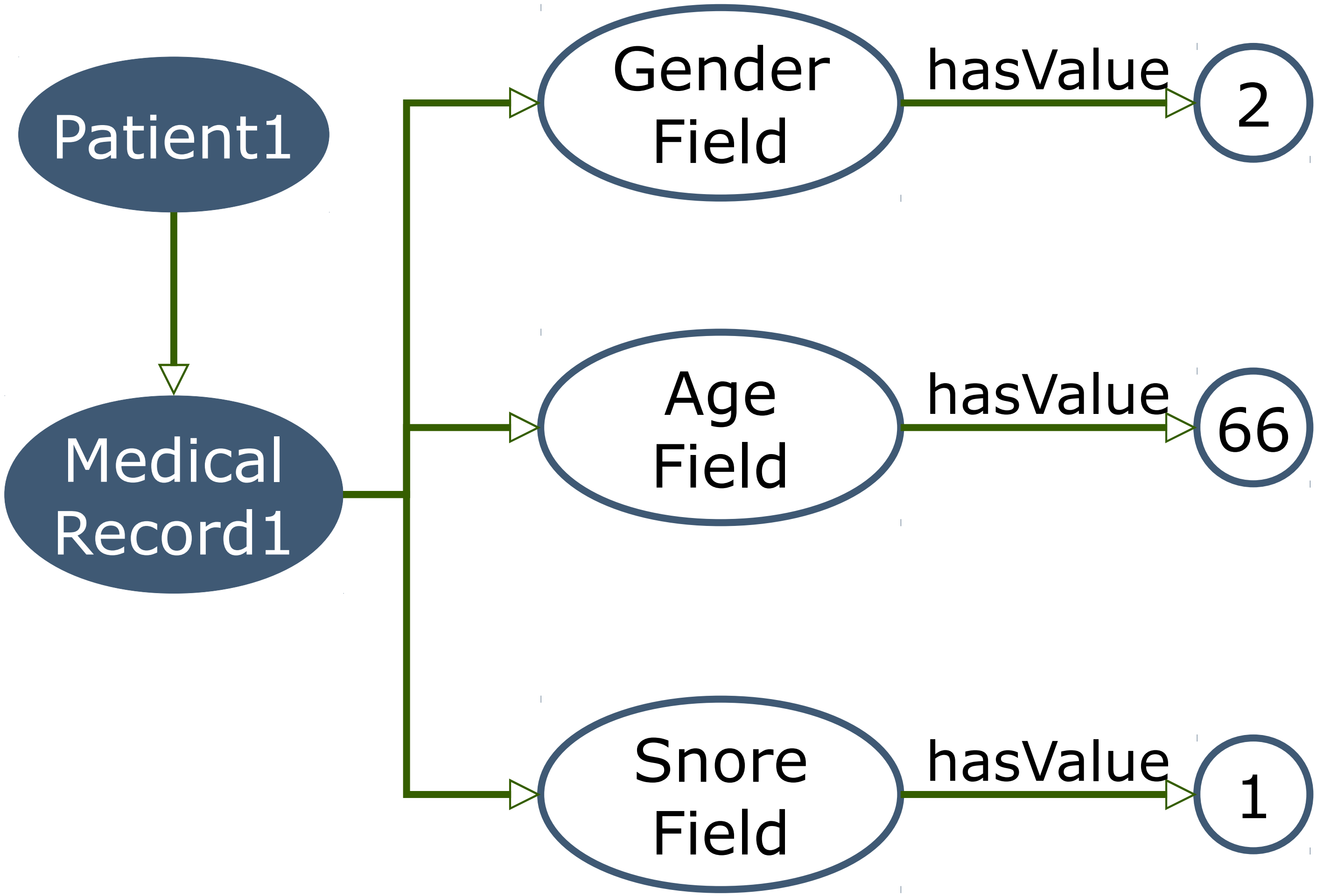
Defines

- taxonomy
  - a hierarchy of concepts
- relationships

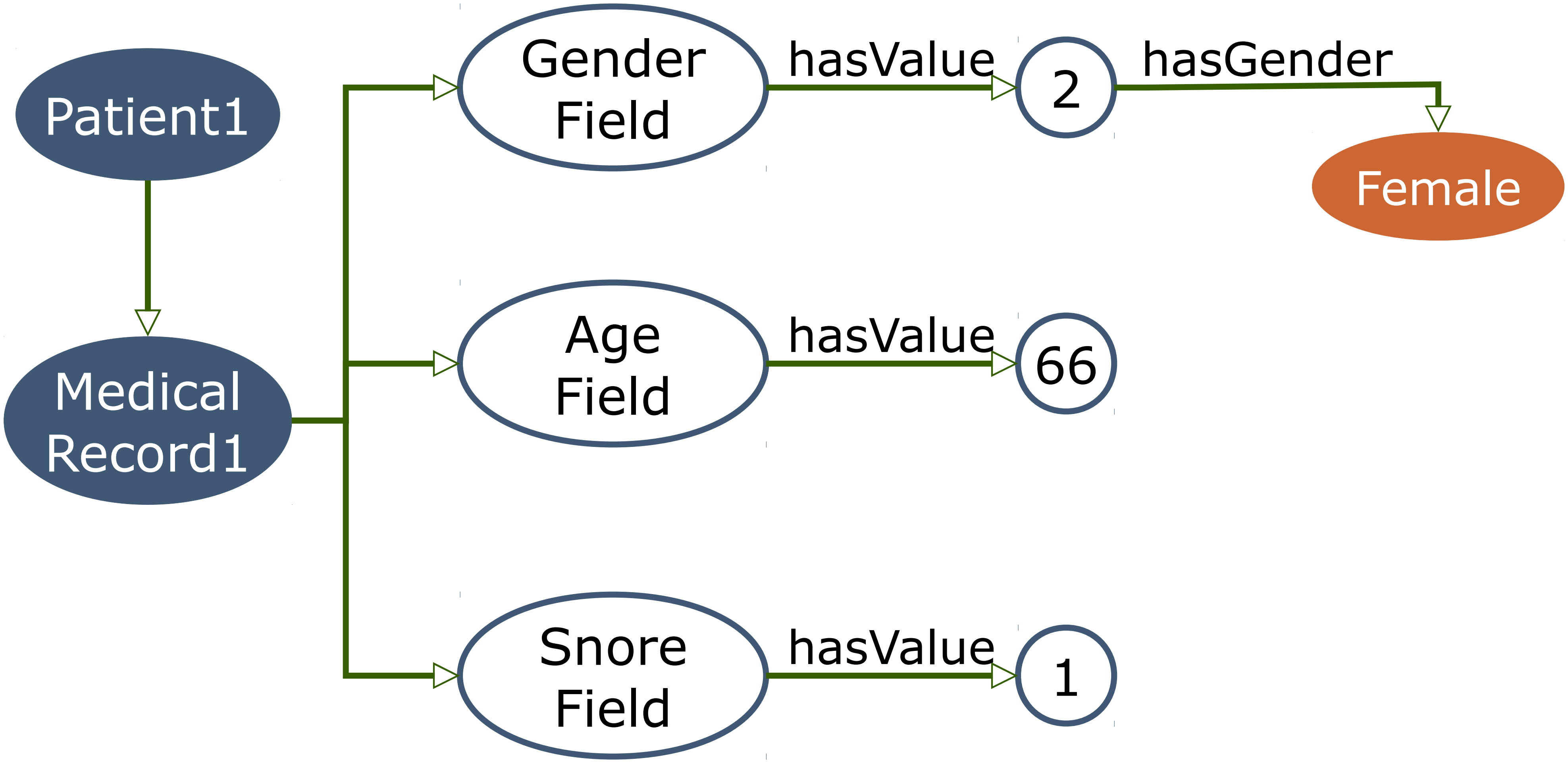
Scope - “*domain of discourse*”

- e.g. medical decision support system

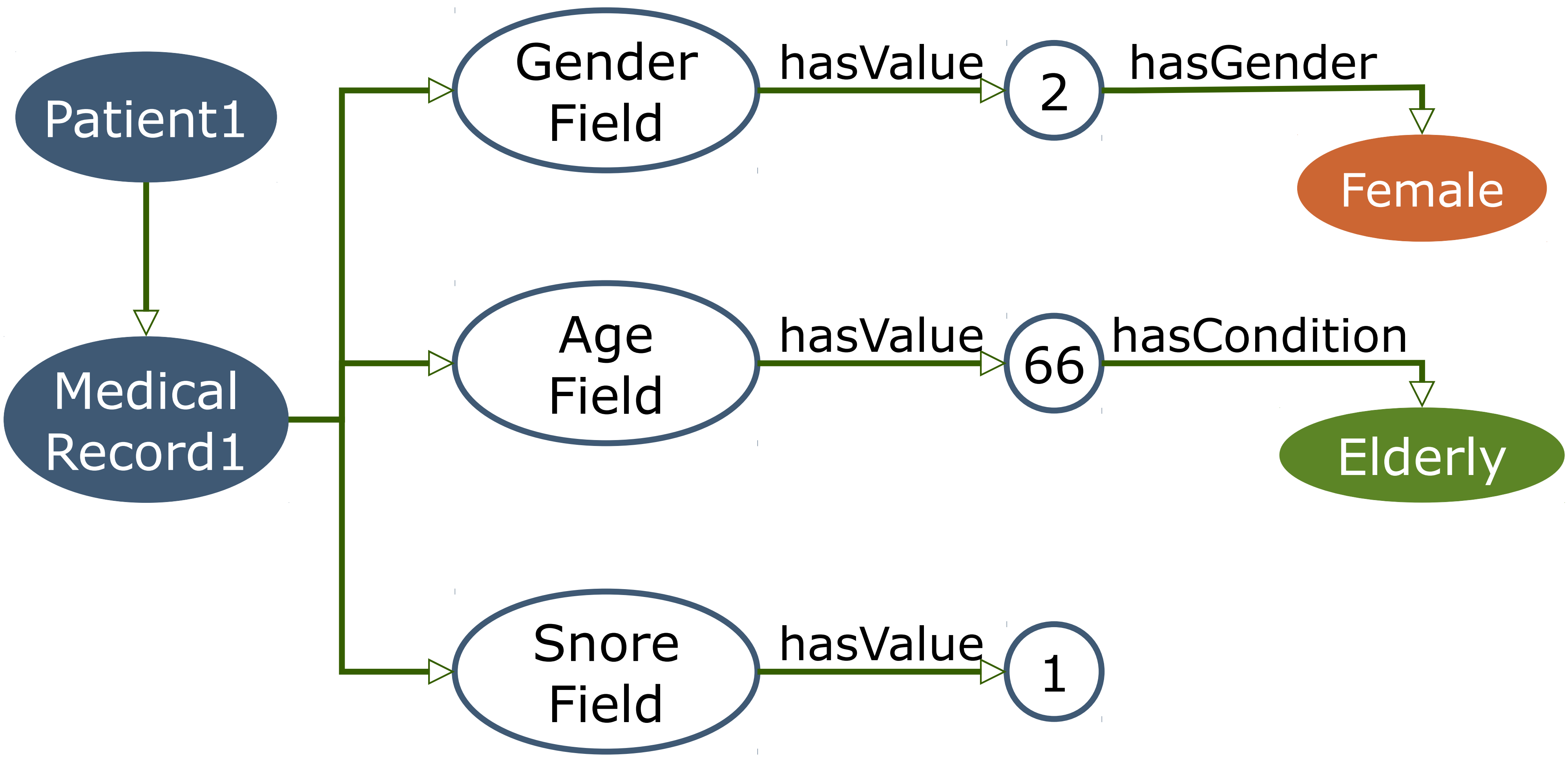
Mapping Raw Data



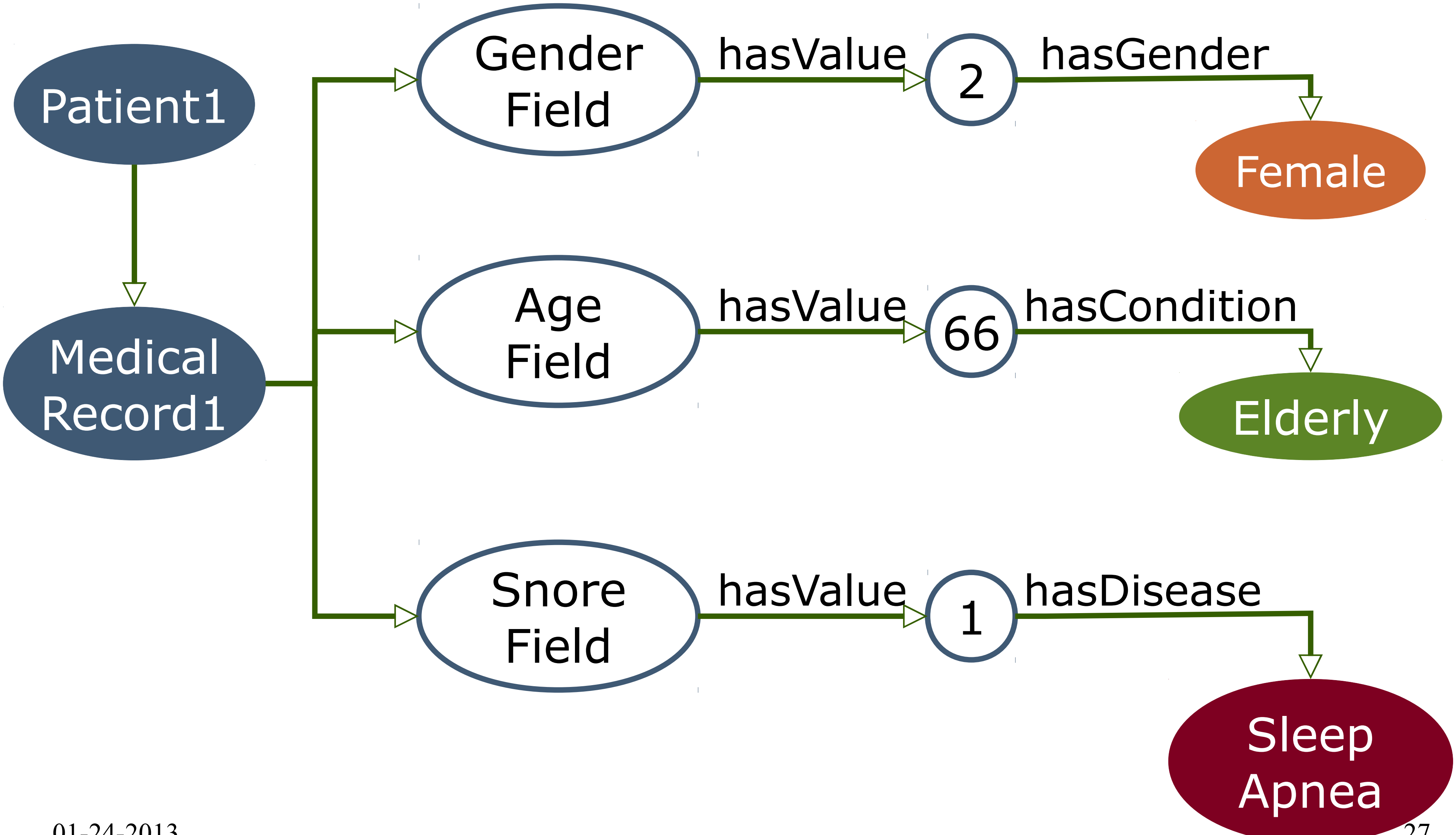
Mapping Raw Data



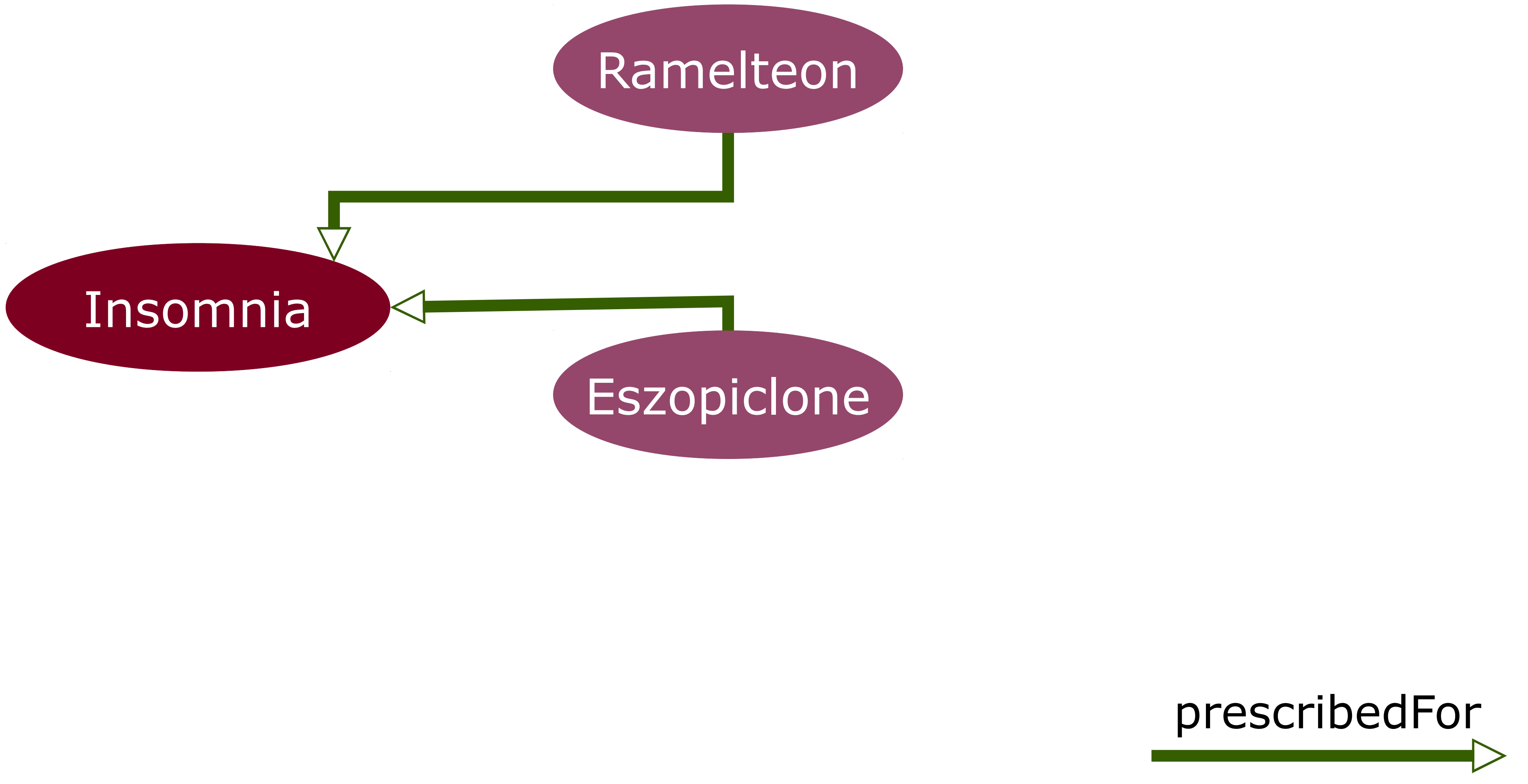
Mapping Raw Data



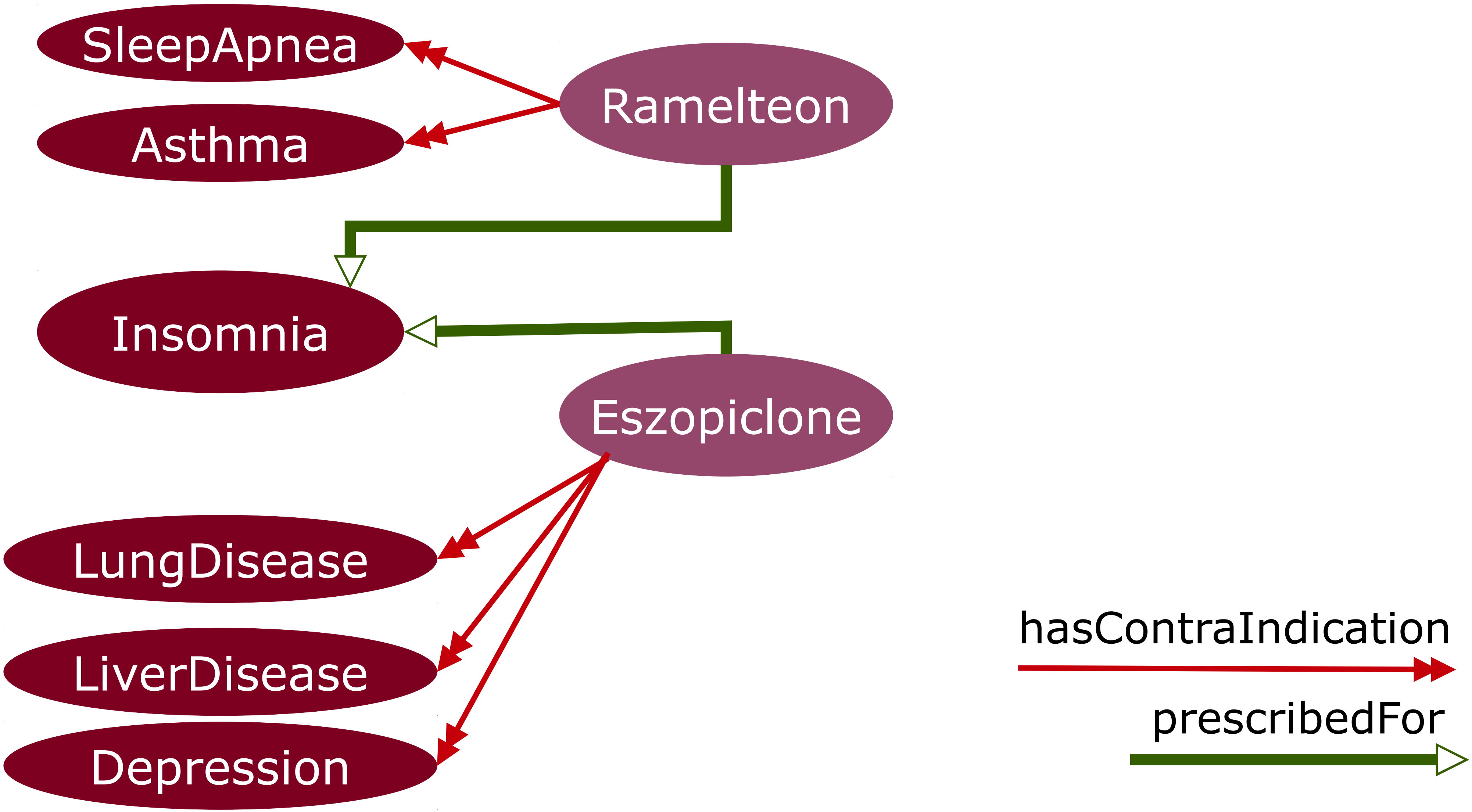
Mapping BRFSS Data



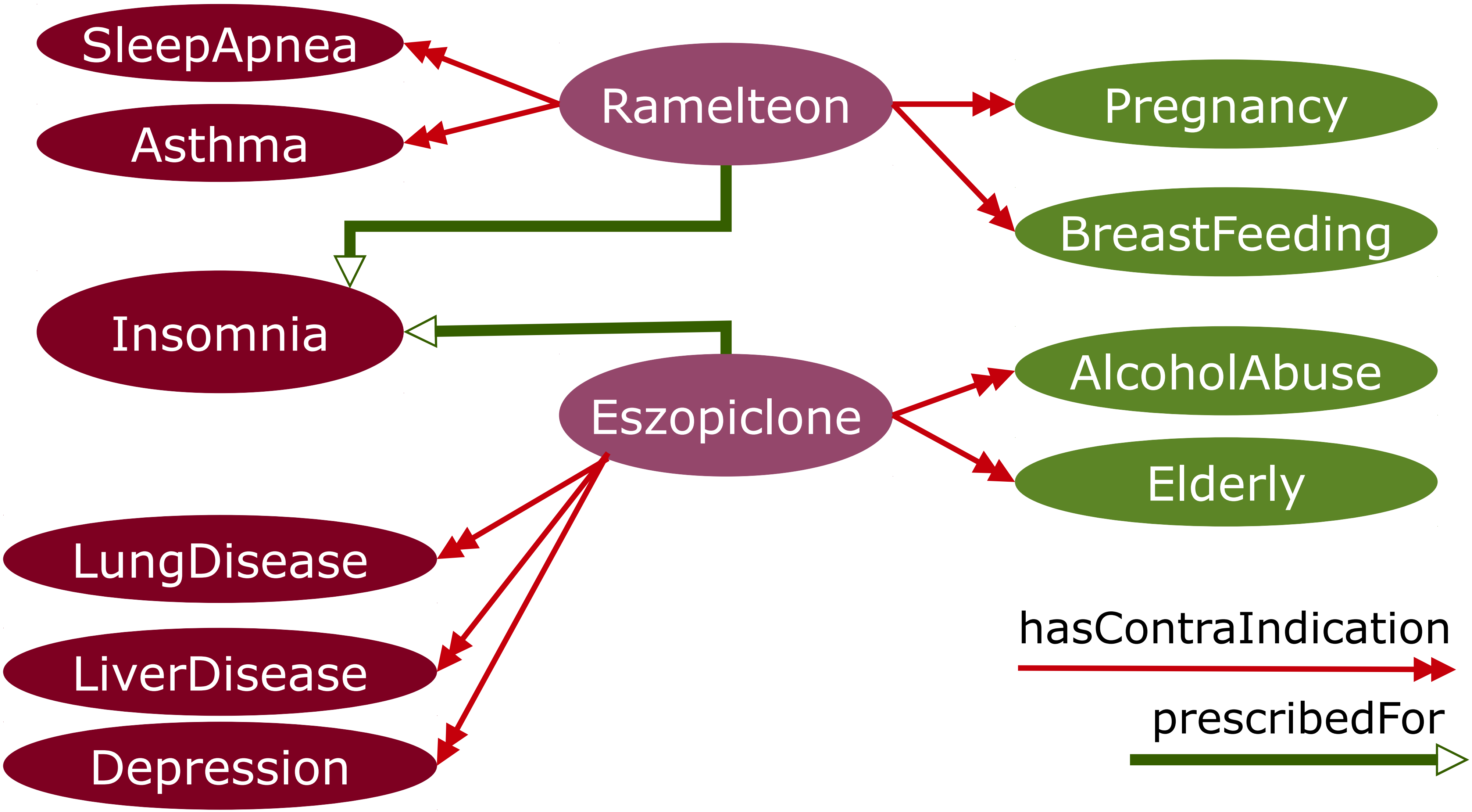
Mapping Expert Knowledge



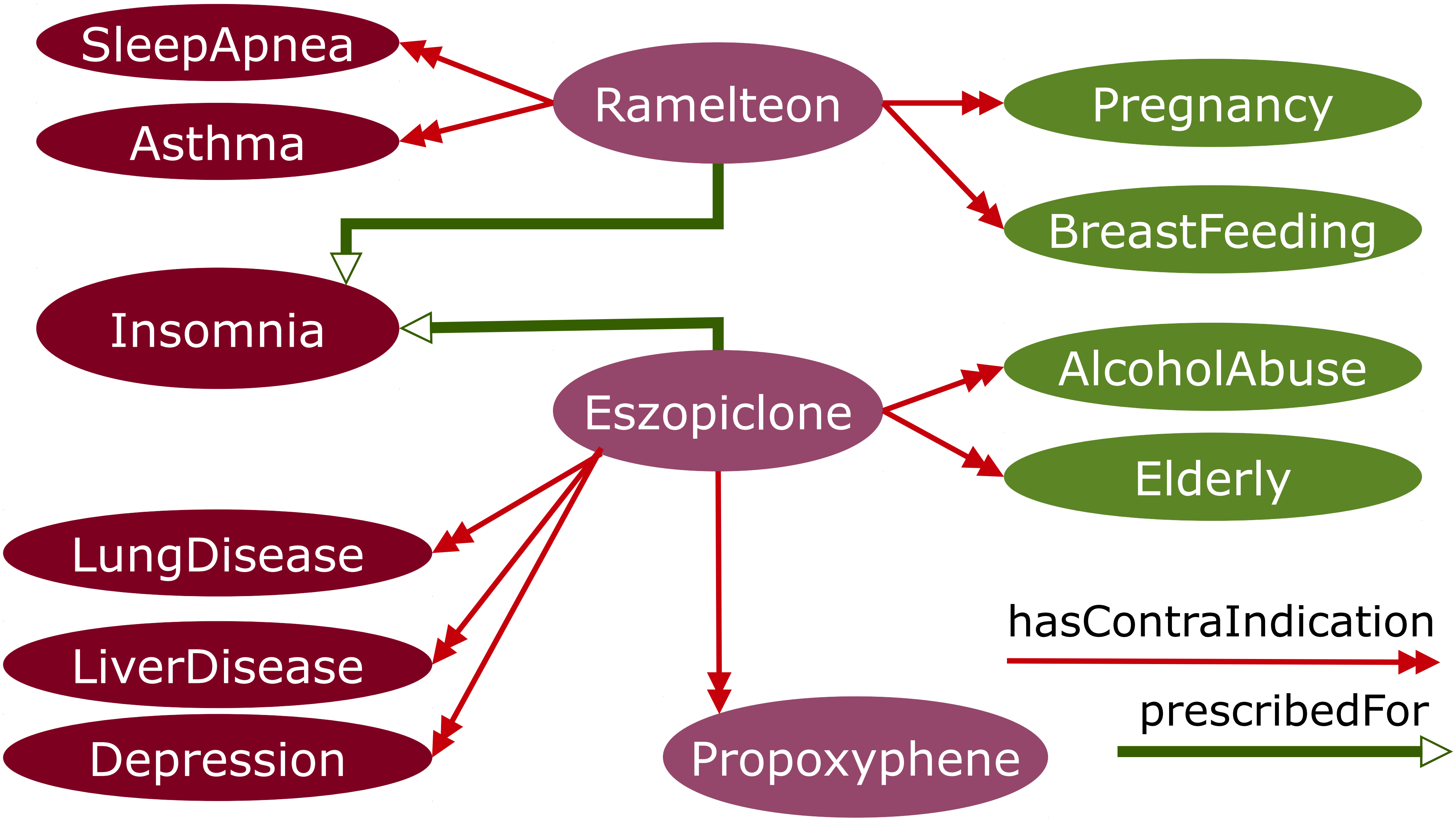
Mapping Expert Knowledge



Mapping Expert Knowledge



Mapping Expert Knowledge



Mapping Expert Knowledge

```
:Propoxyphene a :Drug;
    :isPrescribedFor :Pain;
    :hasContraIndication
      :Eszopiclone.

:Wygesic a :Drug;
    :isPrescribedFor :Pain;
    :hasContraIndication
      :Eszopiclone.

:Trycet a :Drug;
    :isPrescribedFor :Pain;
    :hasContraIndication
      :Eszopiclone.

:PropoxypheneCompound65  a
:Drug;
    :isPrescribedFor :Pain;
    :hasContraIndication
      :Eszopiclone.

:Propacet100   a :Drug;
    :isPrescribedFor :Pain;
    :hasContraIndication :Eszopiclone.

:Aspirin  a :Drug; :isPrescribedFor :Pain.

:Tylenol1 a :Drug; :isPrescribedFor :Pain.

:Tylenol2 a :Drug;
    :isPrescribedFor :Pain;
    :hasContraIndication :SleepingMedication.
```

N3 representation

Drug Information Online
Drugs.com

Mapping Expert Knowledge

```
:Propoxyphene a :Drug;
   :isPrescribedFor :Pain;
   :hasContraIndication
    :Eszopiclone.

:Wygesic a :Drug;
   :isPrescribedFor :Pain;
   :hasContraIndication
    :Eszopiclone.

:Trycet a :Drug;
   :isPrescribedFor :Pain;
   :hasContraIndication
    :Eszopiclone.

:PropoxypheneCompound65  a
:Drug;
   :isPrescribedFor :Pain;
   :hasContraIndication
    :Eszopiclone.

:Propacet100   a :Drug;
   :isPrescribedFor :Pain;
   :hasContraIndication :Eszopiclone.

:Aspirin  a :Drug; :isPrescribedFor :Pain.

:Tylenol1 a :Drug; :isPrescribedFor :Pain.

:Tylenol2 a :Drug;
   :isPrescribedFor :Pain;
   :hasContraIndication :SleepingMedication.
```

N3 representation

Drug Information Online
Drugs.com

Knowledge Representation-Recap

Why?

- to **create**, **maintain** and **share** information in a **precise** manner without ambiguity of meaning

How?

- ontologies

"Now! *That* should clear up a few things around here!"

http://photos1.blogger.com/blogger2/1715/1669/1600/larson-oct-1987.gif

infoTrellis

Knowledge-Inference

a **discovery** process
to find **implied** knowledge
using
explicitly defined information

Knowledge-Inference

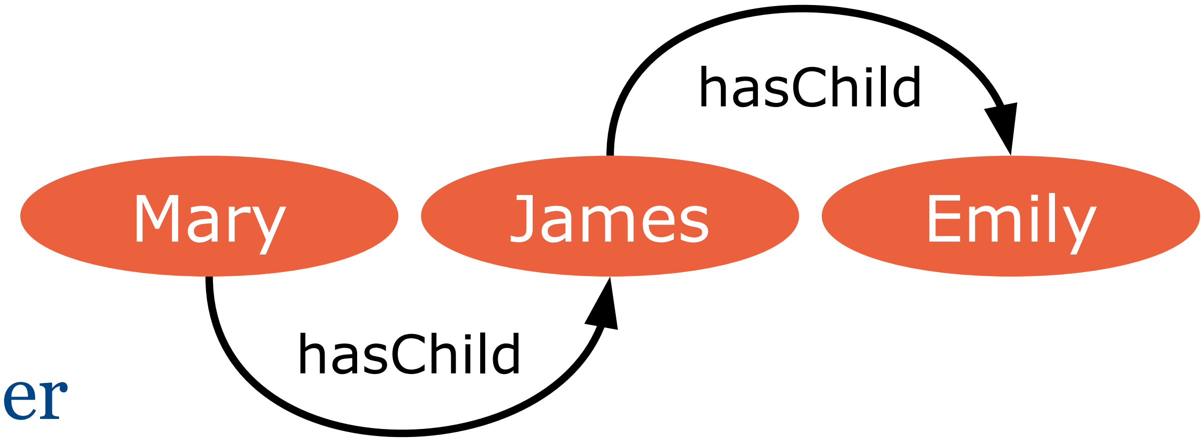
a **discovery** process
to find **implied** knowledge
using
explicitly defined information

ontology

logic-based

Knowledge-Inference: example

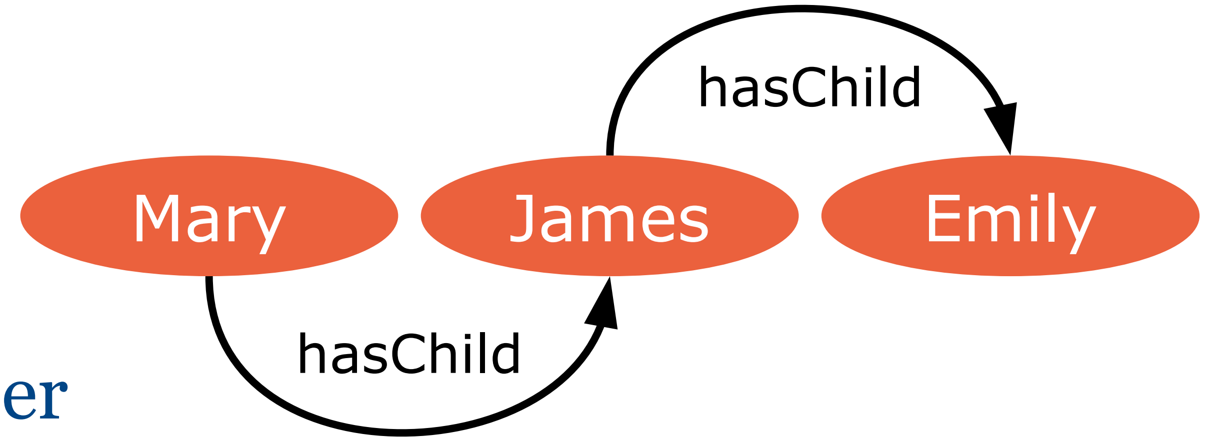
What do we **know**
about Mary?



- Mary is grandmother
- Mary is a grand parent
- Mary is woman

Knowledge-Inference: example

What do we **know** about Mary?



- Mary is grandmother
- **Mary is a grand parent**
- Mary is woman

if a person has a child, and that child also has a child, then the person is a grandparent

infoTrellis

Inference Rules

Drug-to-Drug Interactions

- ***If*** a patient is taking an existing drug (D1) and D1 has contraindication to another drug D2 ***then*** drug D2 should not be prescribed to the patient

```
{  ?P    a :Patient.
   ?D1   a :Drug.
   ?D2   a :Drug.

   ?P    :isTaking ?D1.
   ?D1   :hasContraIndication ?D2.

} => {?P :cannotBeGiven ?D2}.
```

Inference Rules

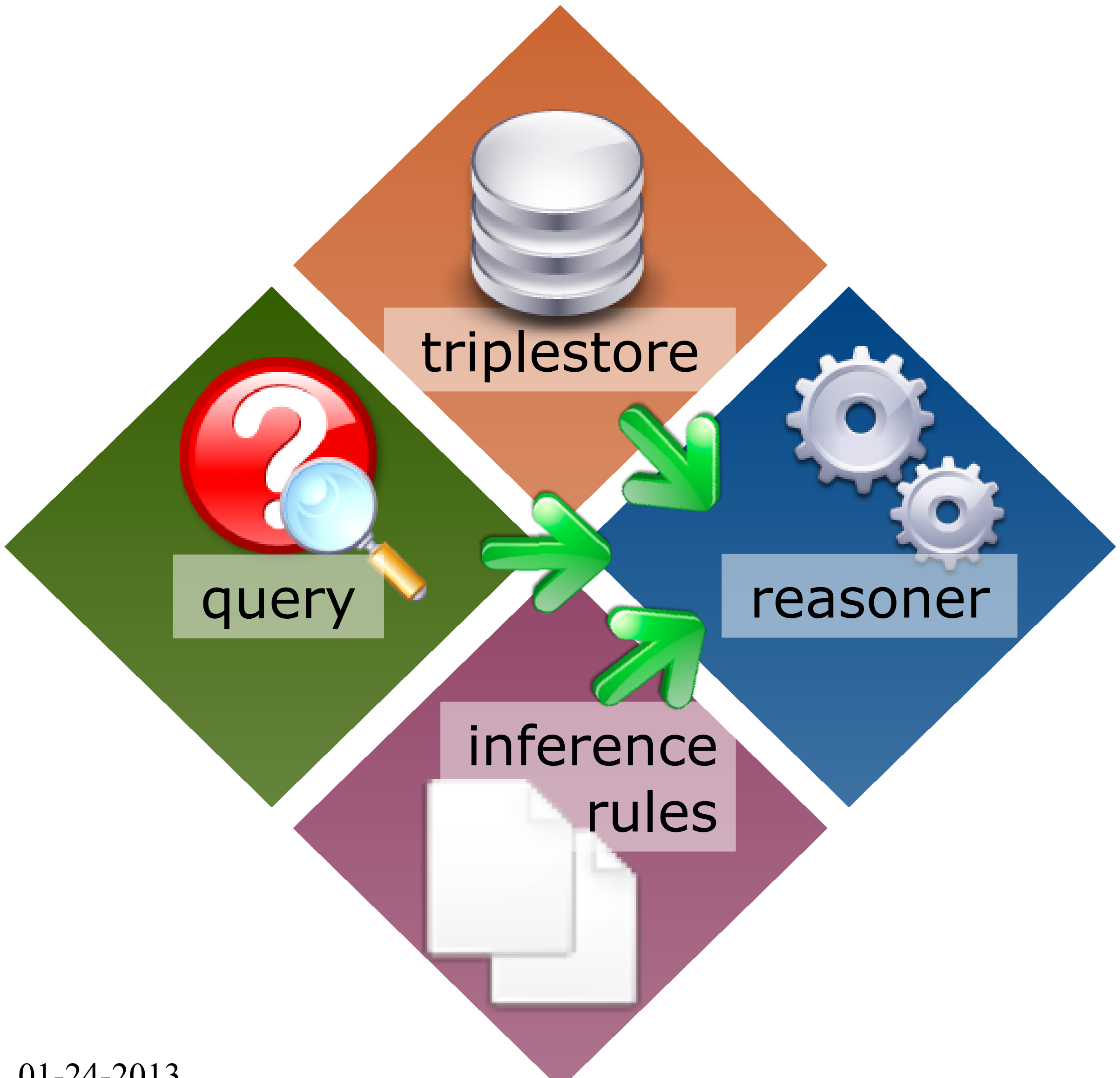
Drug-to-Disease Interactions

- ***If*** a patient has a condition that has a contra indication to a drug
  then the *patient should not be given* the drug

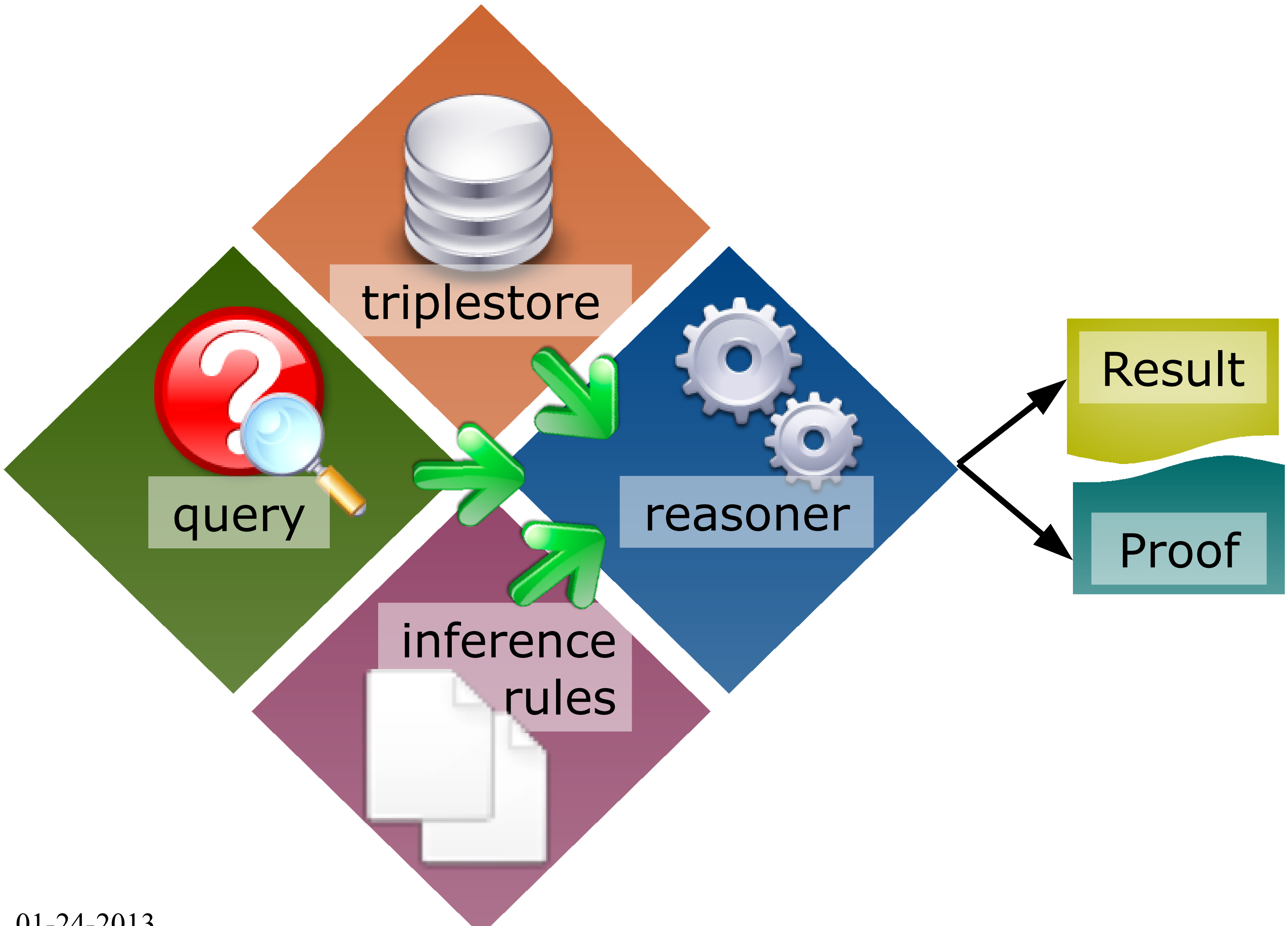
```
{  ?P   a :Patient.
   ?D   a :Drug.
   ?DIS a :Disease.

   ?P   :hasDisease ?DIS.
   ?D   :hasContraIndication ?DIS.

} => {?P :cannotBeGiven ?D}.
```


infoTrellis

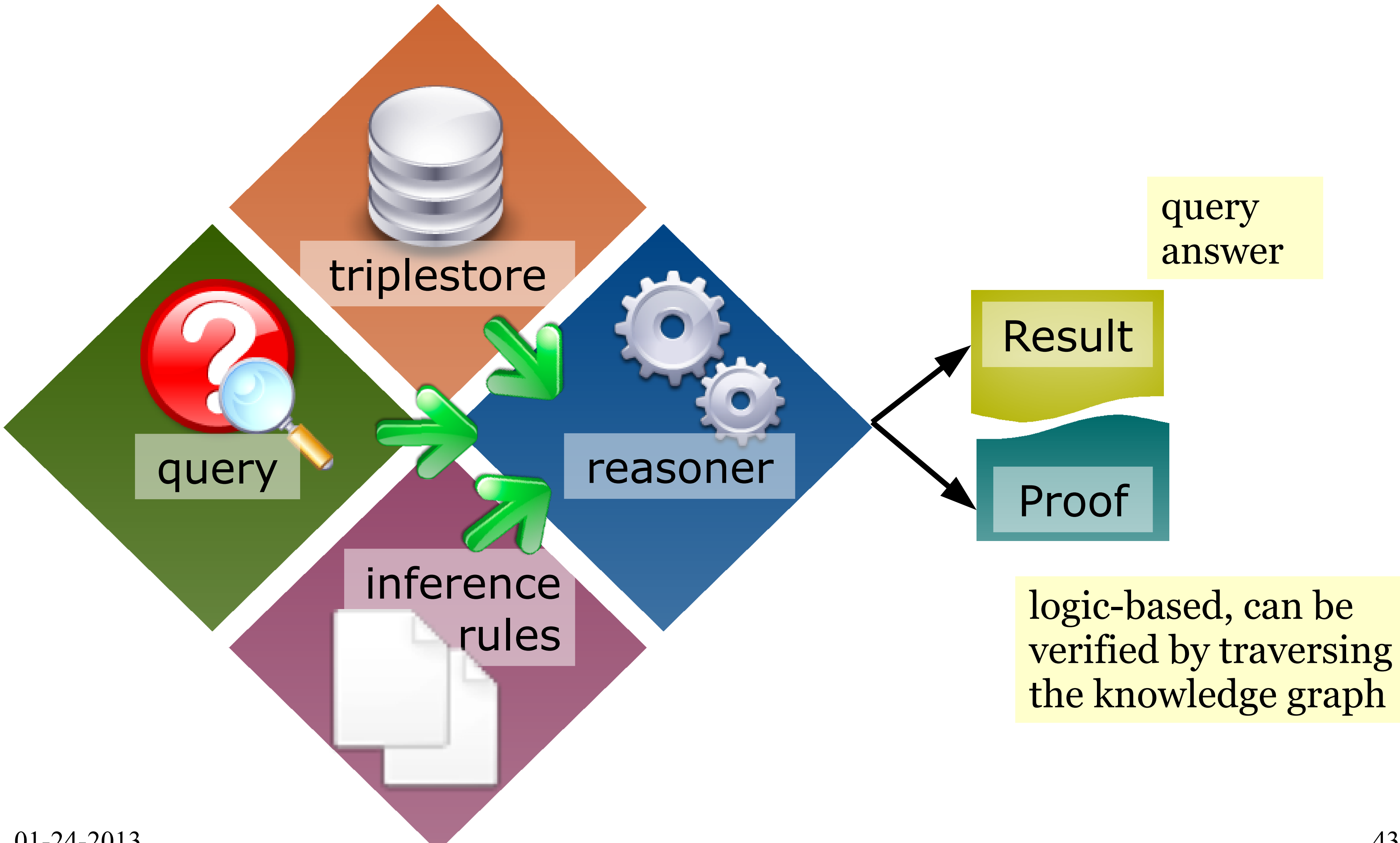
Putting it all Together



Putting it all Together



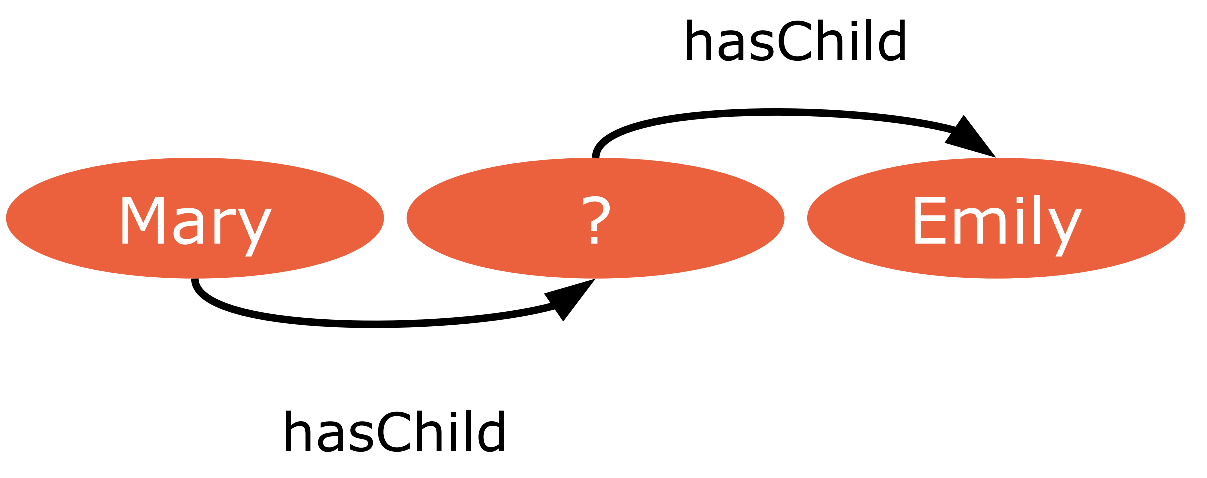
Putting it all Together



But What about Noise?

Noise

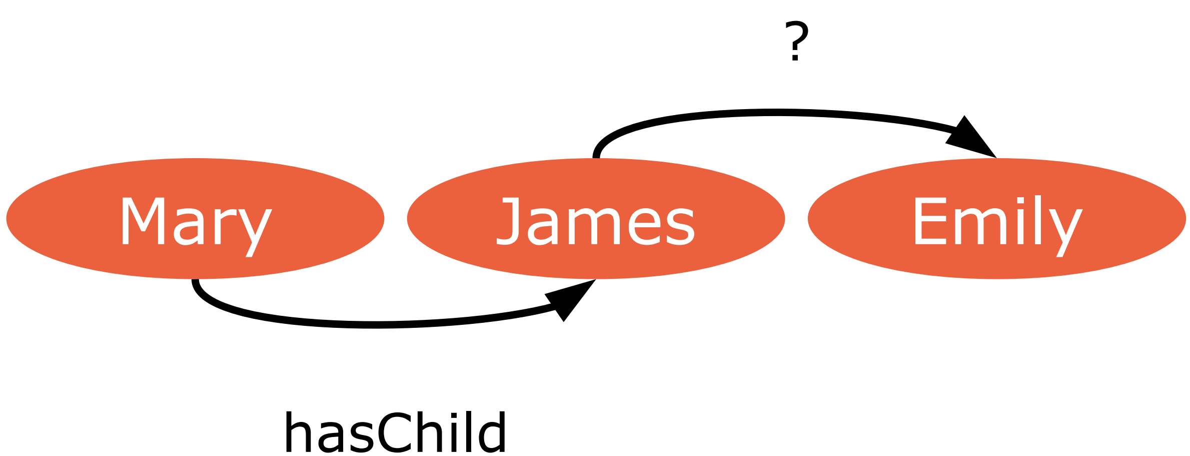
- cripples knowledge-based solutions
- limits inference capability



But What about Noise?

Noise

- cripples knowledge-based solutions
- limits inference capability

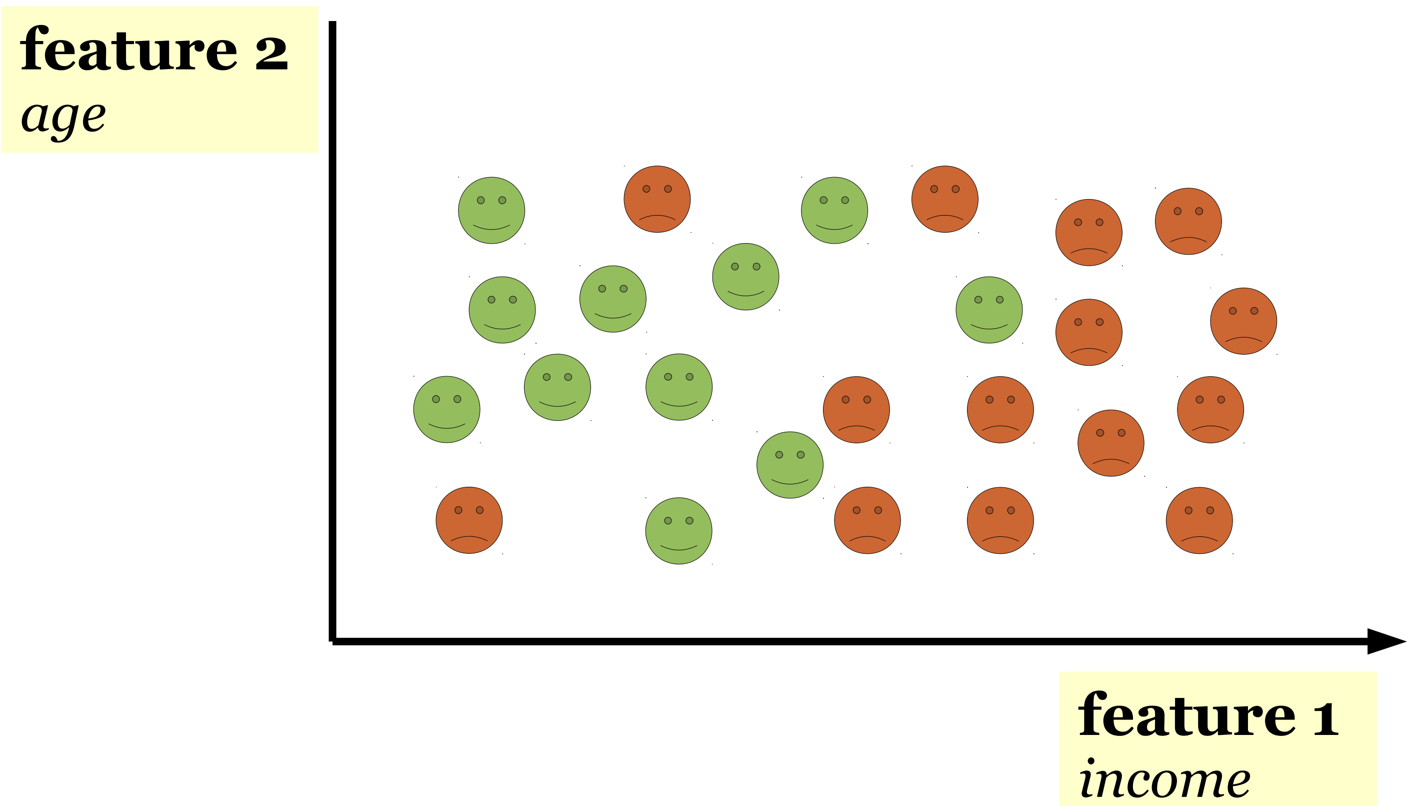


Data is Almost Always Noisy

use
"machine learning"
to deal
with noise

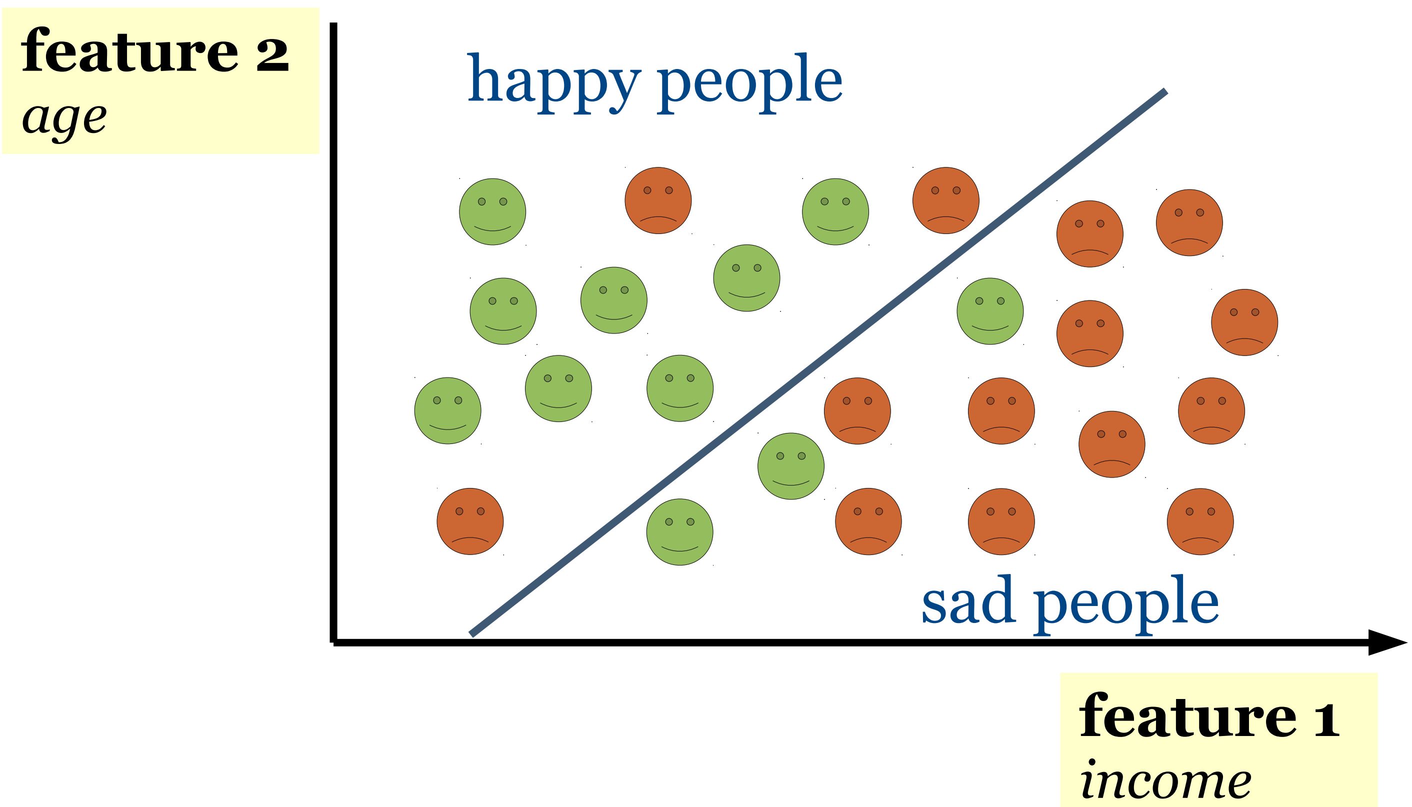
Machine Learning?

query: *g*iven a person's age and income predict if they are ***happy*** or ***sad***

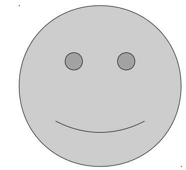


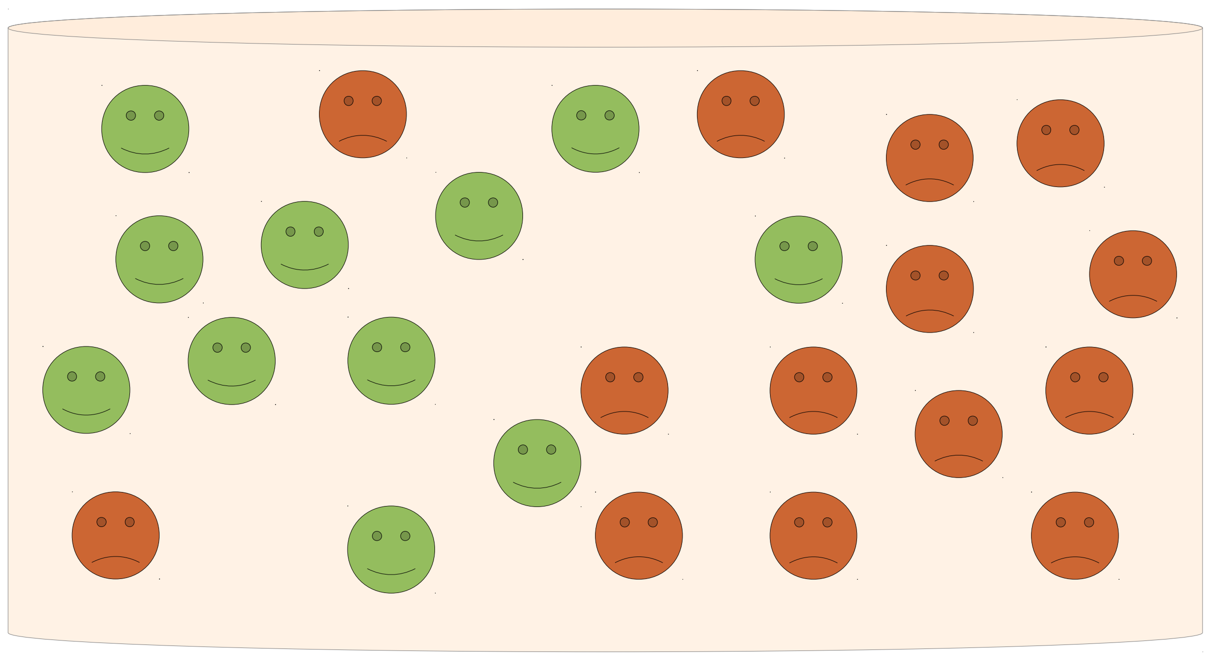
Machine Learning?

query: given a person's age and income predict if they are ***happy*** or ***sad***



Machine Learning?

query: *is he a* ***happy*** or ***sad person***  **?**



sad person

happy person

Machine Learning 101

Machine Learning

- **classification**: *predict class of an instance*
- **regression**: *prediction of a numeric value*
- **clustering**: *group similar items together*

Machine Learning 101

Machine Learning

- **classification**: *predict class of an instance*
- **regression**: *prediction of a numeric value*

- **clustering**: *group similar items together*

Machine Learning 101

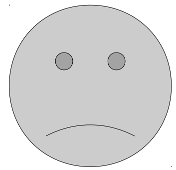
Machine Learning

- **classification**: *predict class of an instance*

Classification

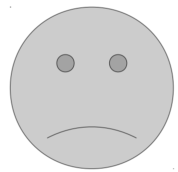
0. Train
   - build a ***classifier*** based on ***known*** *exemplars*
1. Predict
2. Update
3. Evaluate

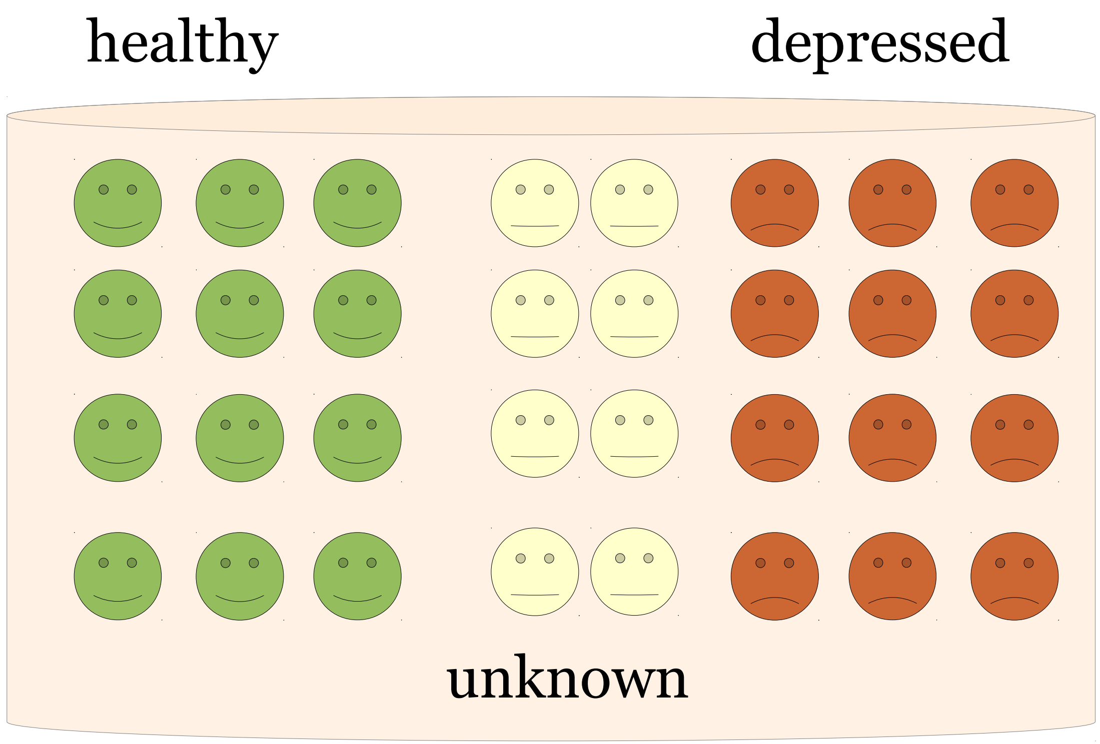
Dealing with ***Noise***

query: *is John*  *depressed ?*

but we don't have access to that information

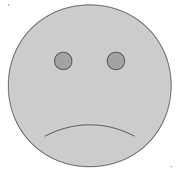
Dealing with ***Noise***

query: *is John*  *depressed ?*

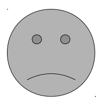
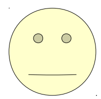


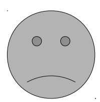
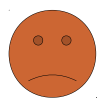
Training exemplars

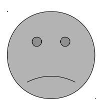
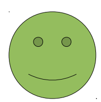
Dealing with ***Noise***

query: *is John*  *depressed ?*

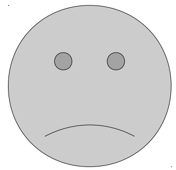
answer:

sameAs( , ) = 0.14

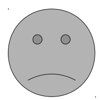
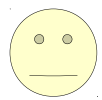
sameAs( , ) = 0.85

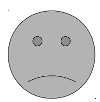
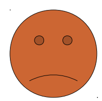
sameAs( , ) = 0.01

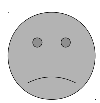
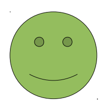
Dealing with ***Noise***

query: *is John*  *depressed ?*

answer:

sameAs( , ) = 0.14

sameAs( , ) = 0.85

sameAs( , ) = 0.01

Recap

a hybrid decision support system with ontology-based knowledge representation and logic- based reasoning augmented with machine learning classification for noise tolerance

Recap

a **hybrid** decision support system with **ontology**-based knowledge **representation** and logic-based **reasoning** augmented with **machine learning** classification for **noise tolerance**

Taming Data in the Wild

Taming Data in the Wild

"*Data-to-knowledge*"

- an **expensive** journey

  however, connecting the "***data dots***" is the key

"*Linked Data*"*

- can make this a reality

* http://linkeddata.org/

Linked Data

Brief Summary

- Tim Berners-Lee's vision

- URIs* to identify *'things'*
- HTTP-based dereferencing of URIs
- structured representation (RDF**)
- hyperlink *'things'* together

Tim Berners-Lee

http://www.w3.org/Press/Stock/Berners-Lee/2001-eur-head-quarter.jpg

* URI = Universal Resource Identifier
** RDF = Resource Description Framework

DBpedia

http://wiki.dbpedia.org

Extract Information from Wikipedia

- unstructured information
  - articles (free text + noise)

- structured components
  - infobox templates
  - categorisation information
  - images
  - geo-coordinates,
  - external web links

DBpedia

Knowledge Engineering using Ontology*

- 359 classes
  - in a subsumption hierarchy

- 1,775 different properties

*http://wiki.dbpedia.org/Ontology

DBpedia

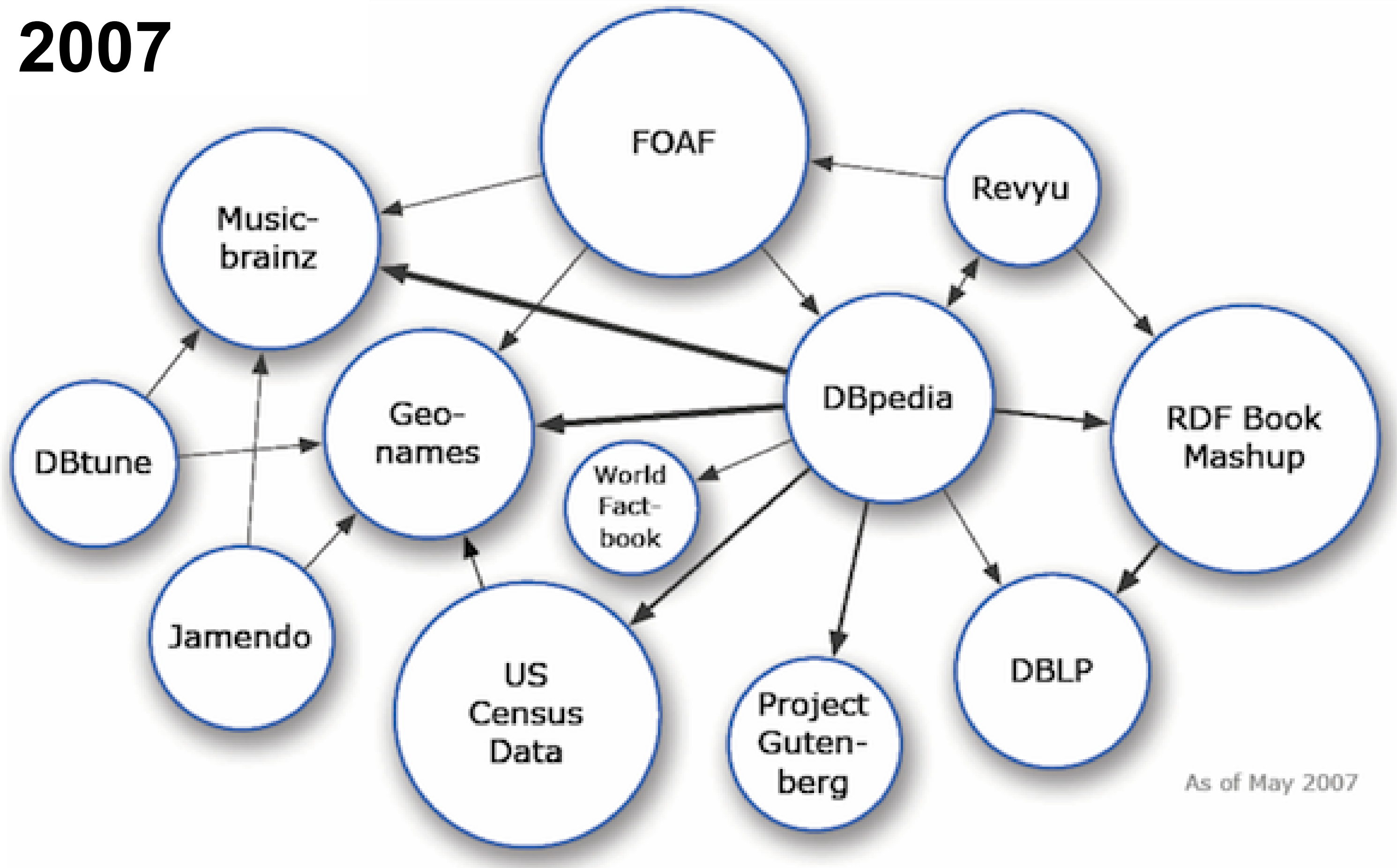
Knowledge Engineering using Ontology

"English version of the DBpedia knowledge base currently describes ***3.77 million things****, out of which 2.35 million are classified in a consistent Ontology, including* ***764,000 persons****,* ***573,000 places*** *(including 387,000 populated places),* ***333,000 creative works*** *(including 112,000* ***music albums****, 72,000* ***films*** *and 18,000* ***video games****),* ***192,000 organizations*** *(including 45,000 companies and 42,000 educational institutions),* ***202,000 species*** *and* ***5,500 diseases****."*

http://wiki.dbpedia.org/Datasets

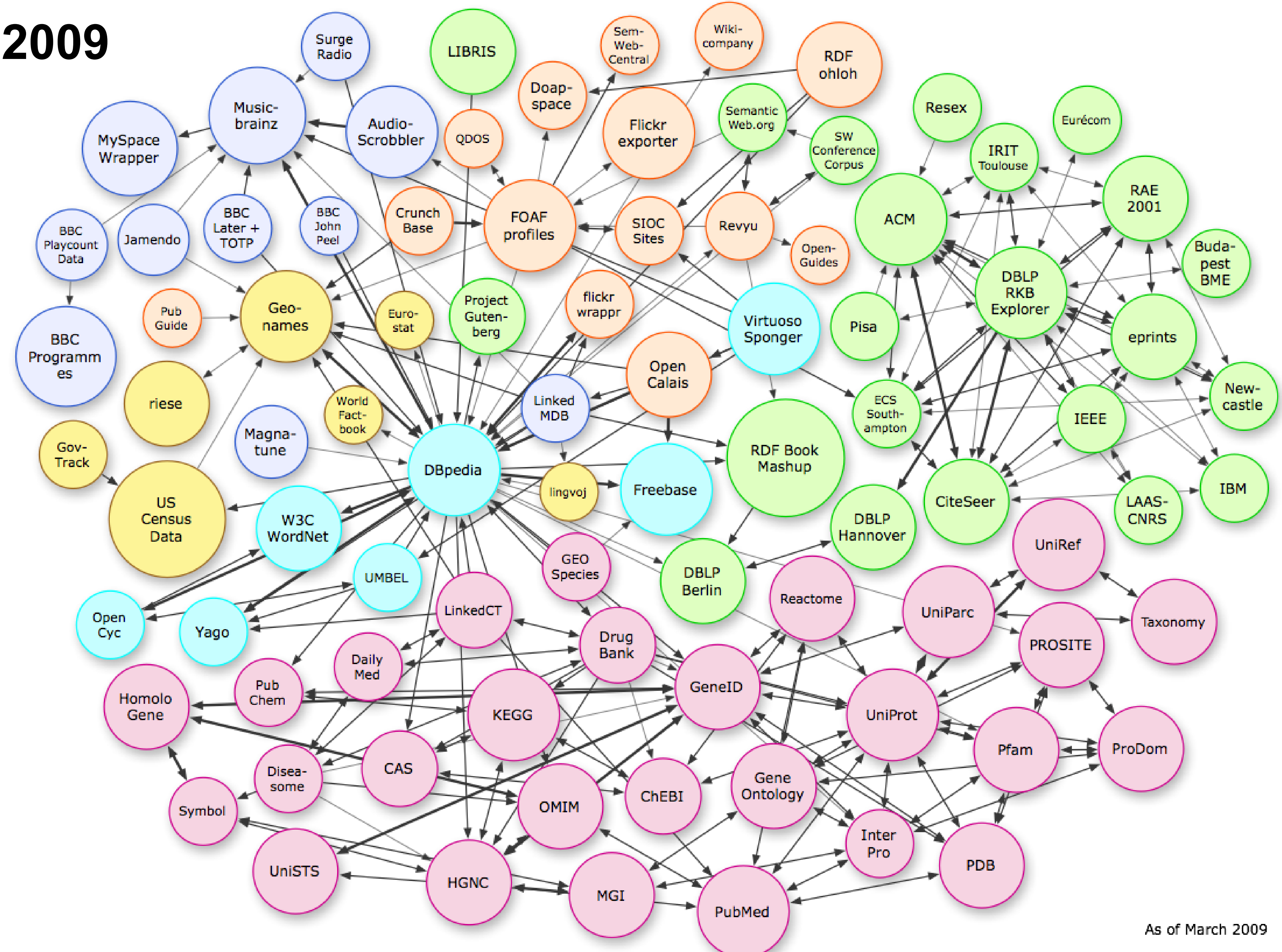
DBpedia – Interlinked Datasets

2007



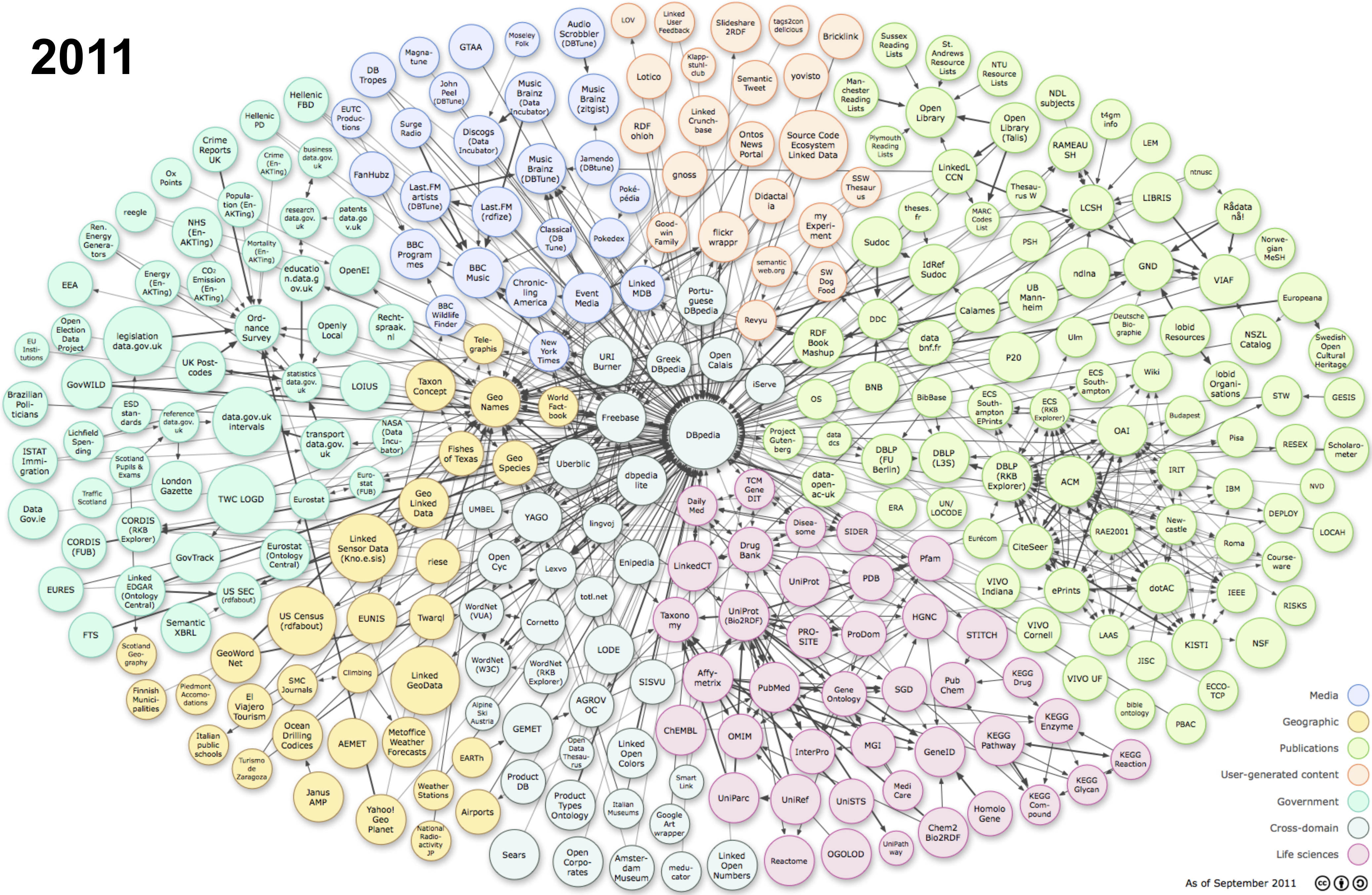
http://richard.cyganiak.de/2007/10/lod/lod-datasets_2007-05-01.png

2009



http://richard.cyganiak.de/2007/10/lod/lod-datasets_2009-03-05_colored.png

2011



http://richard.cyganiak.de/2007/10/lod/lod-datasets_2011-09-19_colored.png

DBpedia in Action

http://www.visualdataweb.org/relfinder/relfinder.php

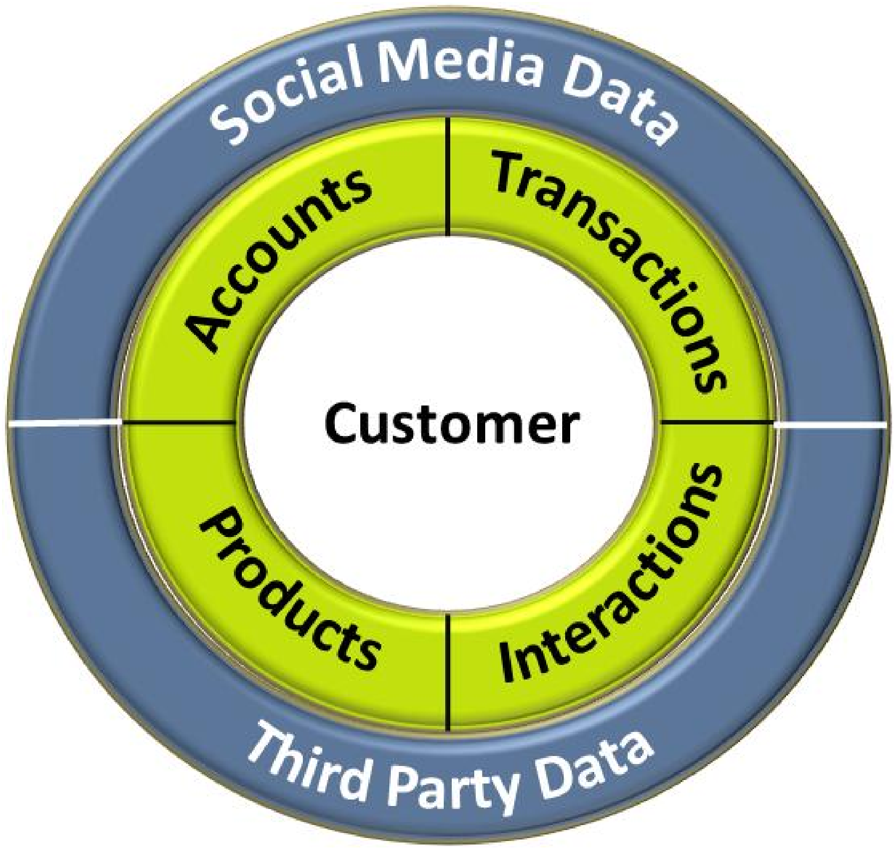
Demo/Video

infoTrellis

ALLSight Platform

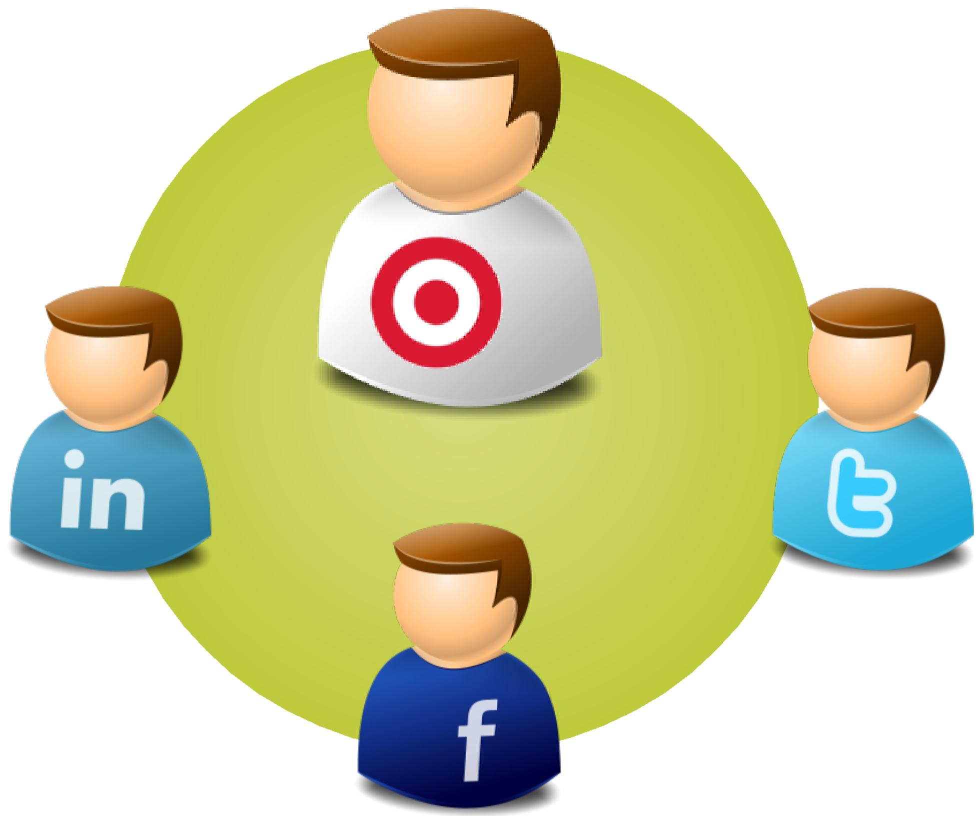
know your customer

AllSight – 360 View of a Customer



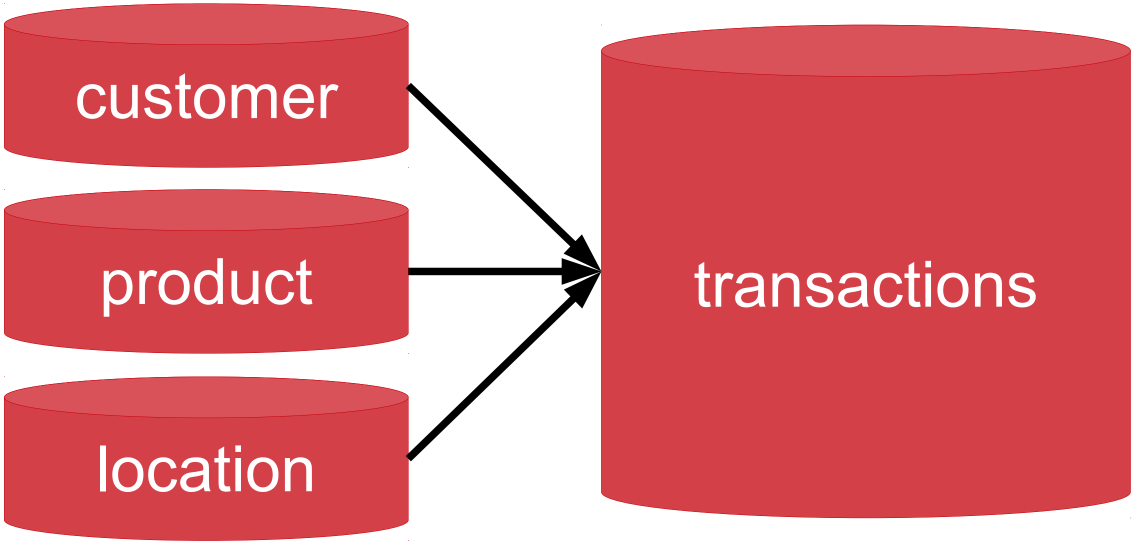
AllSight – 360 View of a Customer

AllSight – 360 View of a Customer

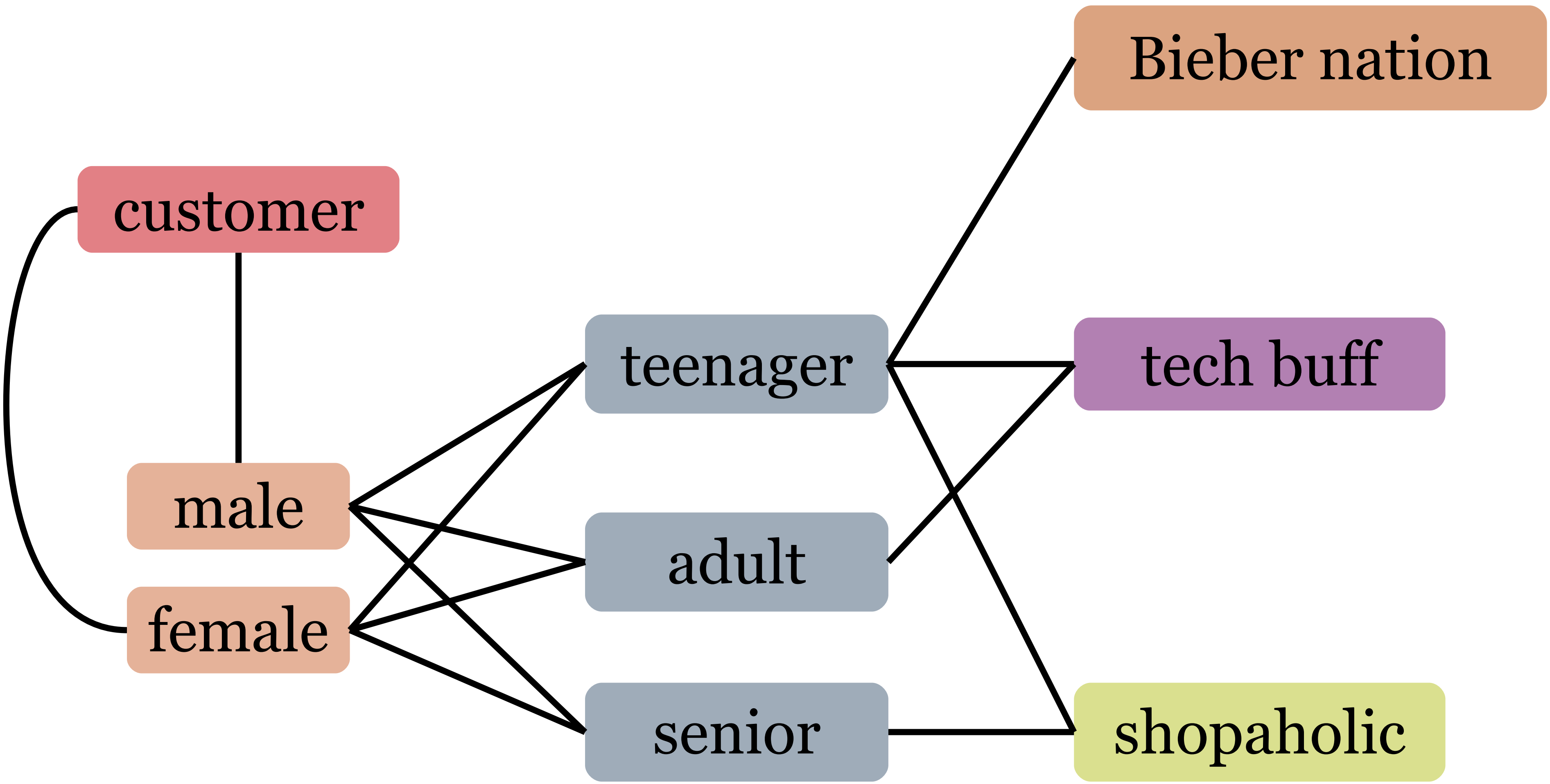


AllSight – 360 View of a Customer

Example



Example


Bieber nation
customer
teenager
tech buff
male
adult
female
senior
shopaholic

infoTrellis | Home | Products | About | Help | Search

Krista Pollard

7 VIP

Krista Pollard has been a customer since March 13, 2003 and has a lifetime purchase value of $58,823. Krista is part of the Pollard Household with 1 other member (Jason Pollard) located in Mt Hope, West Virginia.

Increase Customer Rating | Decrease Customer Rating | Add Comments to Profile

92 Interactions

18 Social Media mentions

360° Relevant Products, Offers & Promotions

Offer Renovation Project card for suspected kitchen renovation project. — 20 — C871

Offer Deal

Recommend joining myReno to help with home renovation projects. — 20 — C938

Offer Deal

Save up to 15% on three or more Bosch kitchen appliances. — 18 — C418

Offer Deal

Social Media Activity

Krista Pollard @ kristapollard — 99% t — 4Square

I'm at Reno Warehouse (Beckley, WV) http://4sq.com/ITLS3A

Krista Pollard @ kristapollard — 83% t — Purchase Intent

Off to Reno WH to purchase a dishwasher!!

Krista Pollard @ kpollard@gmail.com — 83% t — Marketing Opportunity

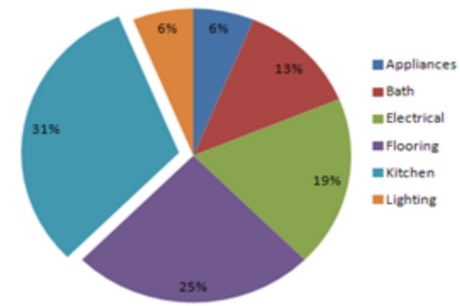
Posted a bunch of pictures of our kitchen renovation so far - check them out!

Purchase | Likes | Demographic | History | Stores | More

Purchase

1Day

5Yr



Products Purchased – 1 Month

20 X FLOORS 2000 11-Pack 13" x 13" Altamira Classico Glazed Porcelain Floor Tile Item #: 543012 Model #: ALT13CLASSICO Cost: $22.70/unit Total Price: $454.00

KOHLER Biscuit Double-Basin Cast Iron or Kitchen Sink Item #: 324508 Model #: 6626-4-96 Cost: $492.49/unit Total Price:$492.49

KOHLER Forte Polished Chrome 1-Handle Pull-Out Kitchen Faucet Item #: 205950 Model #: R10433-CP Cost: $168.00 Total Price: $168.00

3 X SYLVANIA 2-Pack 40-Watt Equivalent Soft White Compact Fluorescent Light Bulb (ENERGY STAR) Item #: 150064 Model #: 26957 Cost: $6.36/unit Total Price: $19.08

Big Data Governance
D a s h b o a r d

My Settings | Help | Logout

Welcome Anderson, James

Dashboard Reports Support Links

twitter Dashboard

Tweet Summary

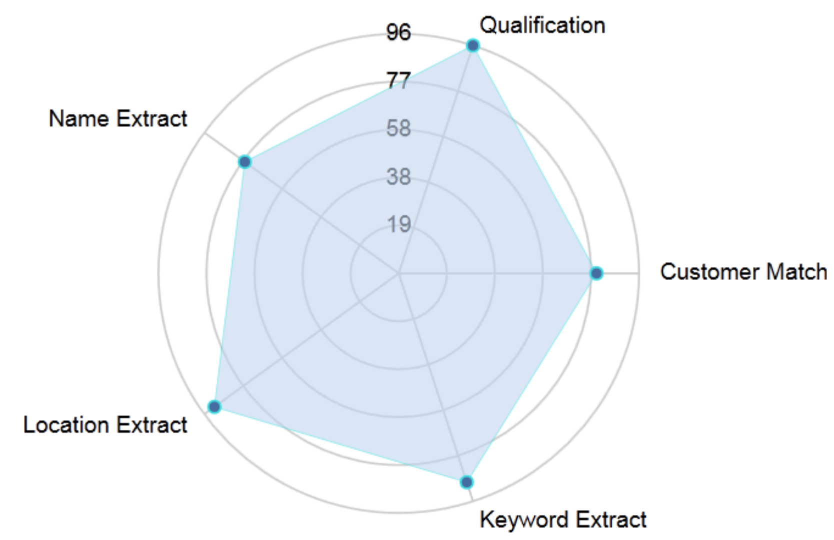
Tweet Summary
Store Visits
Qualification Details
Location Quality
Name Quality
Keyword Identification
Match Quality
Competitor Summary
Process Details

View by : Today Go

METRICS SUMMARY

Total Acquired	**5,333,242 (8,294 avg/day)**
Transaction References	3,182,277
Store Visit Broadcasts	101,233
Employee Chatter	58,410
Events	63,382
News	789,289
Job Ads	84,295
Coupons	771,114
Not Applicable	993,242
Total Qualified	**4,175,519 (5,012 avg/day)**
Qualification Confidence	56 % avg
Total Customer Matches	**184,299**
Match Confidence	81.55 %

QUALITY SUMMARY



TWEET SUMMARY BREAKDOWN FROM [WEEK OF JULY 2,2012] TO [WEEK OF AUGUST 13, 2012] SUMMARIZED AS [WEEKLY]

Time Period	Total Acquired	Total Qualified	Confidence	Transaction References	Store Visits	Customer Matches	Confidence
Week of July 2,2012	50,659	36,995	89%	13,811	23,164	2,604	91%
Week of July 9,2012	48,010	33,716	90%	11,704	22,012	2,194	92%

METRICS SUMMARY

Total Acquired	**5,333,242 (8,294 avg/day)**
Transaction References	3,182,277
Store Visit Broadcasts	101,233
Employee Chatter	58,410
Events	63,382
News	789,289
Job Ads	84,295
Coupons	771,114
Not Applicable	993,242
Total Qualified	**4,175,519 (5,012 avg/day)**
Qualification Confidence	56 % avg
Total Customer Matches	**184,299**
Match Confidence	81.55 %

Under the Hood

Data Enrichment

- taxonomies/ontologies
- dictionaries/catalogues

Matching (decision making)

- knowledge-based rules
- machine learning

Challenges

Entity Resolution

Pre-processing

Feature Selection

Training (& retraining)

Noise

Tools of the Trade

Some Basic Tools

Knowledge Engineering

- Protégé
- Web Ontology Language – OWL
- Resource Description Framework (RDF)
  - XML
  - Notation 3 (N3)

Some Basic Tools

Semantic Reasoners

- CWM
- Euler Sharp
- Jenna
- FuXi

Some Basic Tools

Machine Learning Toolkits

- Weka
- Stanford NLP
- Apache Mahout*

- LIBLINEAR*
- LIBSVM*

- numpy, scipy (python libs)

In Summary

The collaborative work of this group has the potential to unlock data to create knowledge

There are many

- uses cases that can benefit from this work
- tools available to facilitate this process

Thank You!!

Confidence in Your Data

*"Velocity, Volume, and Variety without **Veracity** creates **Vulnerability**"**

* Adler on Data Governance
https://www.ibm.com/developerworks/mydeveloperworks/blogs/adler/entry/the_ibm_big_data_governance_summit3?lang=en

Confidence in Your Data

*"Velocity, Volume, and Variety without **Veracity** creates **Vulnerability**"**

evaluate confidence values

* Adler on Data Governance
https://www.ibm.com/developerworks/mydeveloperworks/blogs/adler/entry/the_ibm_big_data_governance_summit3?lang=en