User-Centric Ontology Population

KENNETH CLARKSON, **ANNA LISA GENTILE**,
DANIEL GRUHL, PETAR RISTOSKI,
JOSEPH TERDIMAN, STEVE WELCH

IBM RESEARCH

Motivation

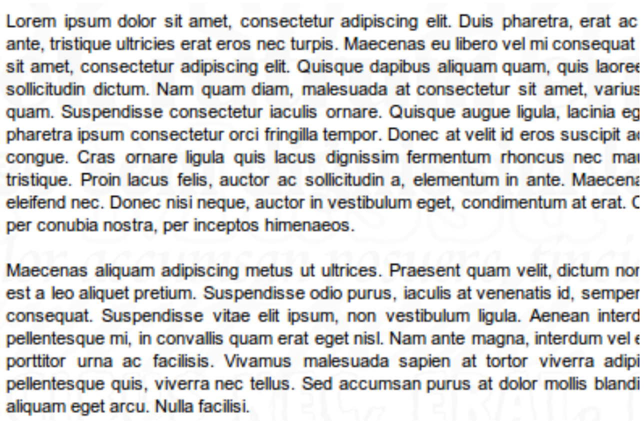
“The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries” - Tim Berners-Lee 2001

17 years later still vast amount of valuable unstructured and semi-structured data is published on the Web

Goal: automatically extract semantic data from text

Problem Statement

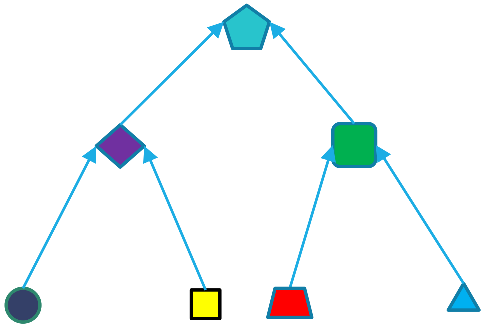
Input Text Corpus



Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis pharetra, erat ac
ante, tristique ultricies erat eros nec turpis. Maecenas eu libero vel mi consequat
sit amet, consectetur adipiscing elit. Quisque dapibus aliquam quam, quis laoree
sollicitudin dictum. Nam quam diam, malesuada at consectetur sit amet, varius
quam. Suspendisse consectetur iaculis ornare. Quisque augue ligula, lacinia eg
pharetra ipsum consectetur orci fringilla tempor. Donec at velit id eros suscipit a
congue. Cras ornare ligula quis lacus dignissim fermentum rhoncus nec ma
tristique. Proin lacus felis, auctor ac sollicitudin a, elementum in ante. Maecena
eleifend nec. Donec nisi neque, auctor in vestibulum eget, condimentum at erat. C
per conubia nostra, per inceptos himenaeos.

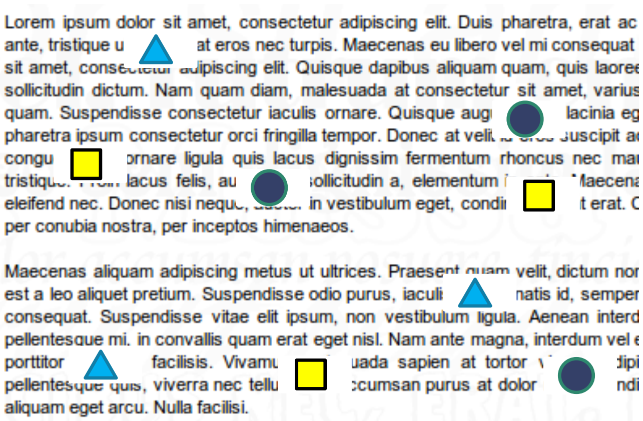
Maecenas aliquam adipiscing metus ut ultrices. Praesent quam velit, dictum nor
est a leo aliquet pretium. Suspendisse odio purus, iaculis at venenatis id, semper
consequat. Suspendisse vitae elit ipsum, non vestibulum ligula. Aenean interd
pellentesque mi, in convallis quam erat eget nisl. Nam ante magna, interdum vel e
porttitor urna ac facilisis. Vivamus malesuada sapien at tortor viverra adipi
pellentesque quis, viverra nec tellus. Sed accumsan purus at dolor mollis blandi
aliquam eget arcu. Nulla facilisi.

Target Ontology



Problem Statement

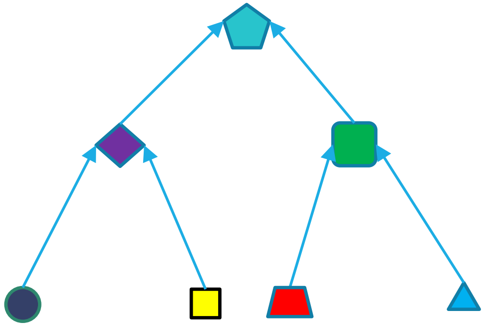
Entity Detection



Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis pharetra, erat ac ante, tristique u at eros nec turpis. Maecenas eu libero vel mi consequat sit amet, consec adipiscing elit. Quisque dapibus aliquam quam, quis laoree sollicitudin dictum. Nam quam diam, malesuada at consectetur sit amet, varius quam. Suspendisse consectetur iaculis ornare. Quisque augi lacinia eg pharetra ipsum consectetur orci fringilla tempor. Donec at veli suscipit a congu ornare ligula quis lacus dignissim fermentum rhoncus nec ma tristique. Proin lacus felis, au sollicitudin a, elementum Maecena eleifend nec. Donec nisi neque, in vestibulum eget, condi t erat. C per conubia nostra, per inceptos himenaeos.

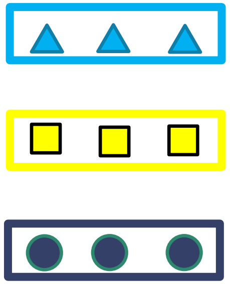
Maecenas aliquam adipiscing metus ut ultrices. Praesent quam velit, dictum no est a leo aliquet pretium. Suspendisse odio purus, iaculi natis id, semper consequat. Suspendisse vitae elit ipsum, non vestibulum ligula. Aenean interd pellentesque mi. in convallis quam erat eget nisl. Nam ante magna, interdum vel e porttitor facilisis. Vivamu uada sapien at tortor v dipi pellentesque quis, viverra nec tellu ccumsan purus at dolor ndi aliquam eget arcu. Nulla facilisi.

Target Ontology

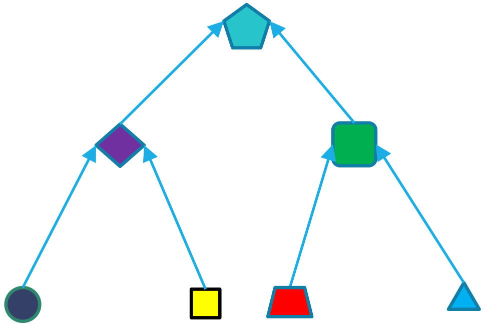


Problem Statement

User-defined Entity Grouping



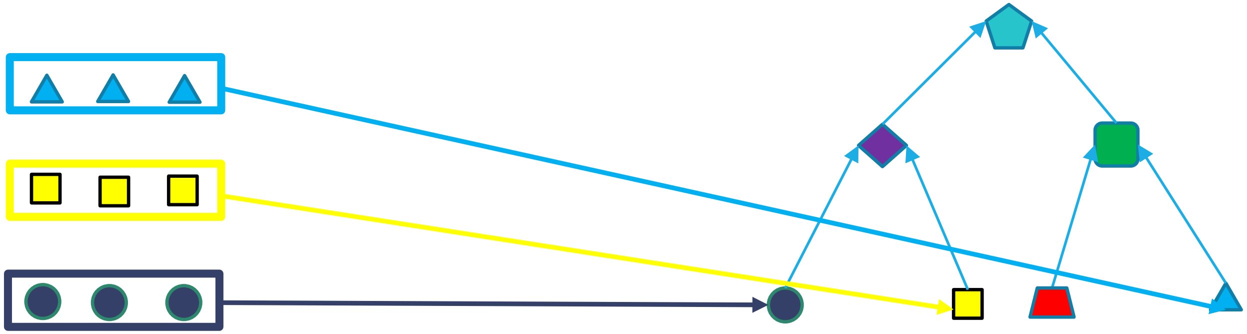
Target Ontology



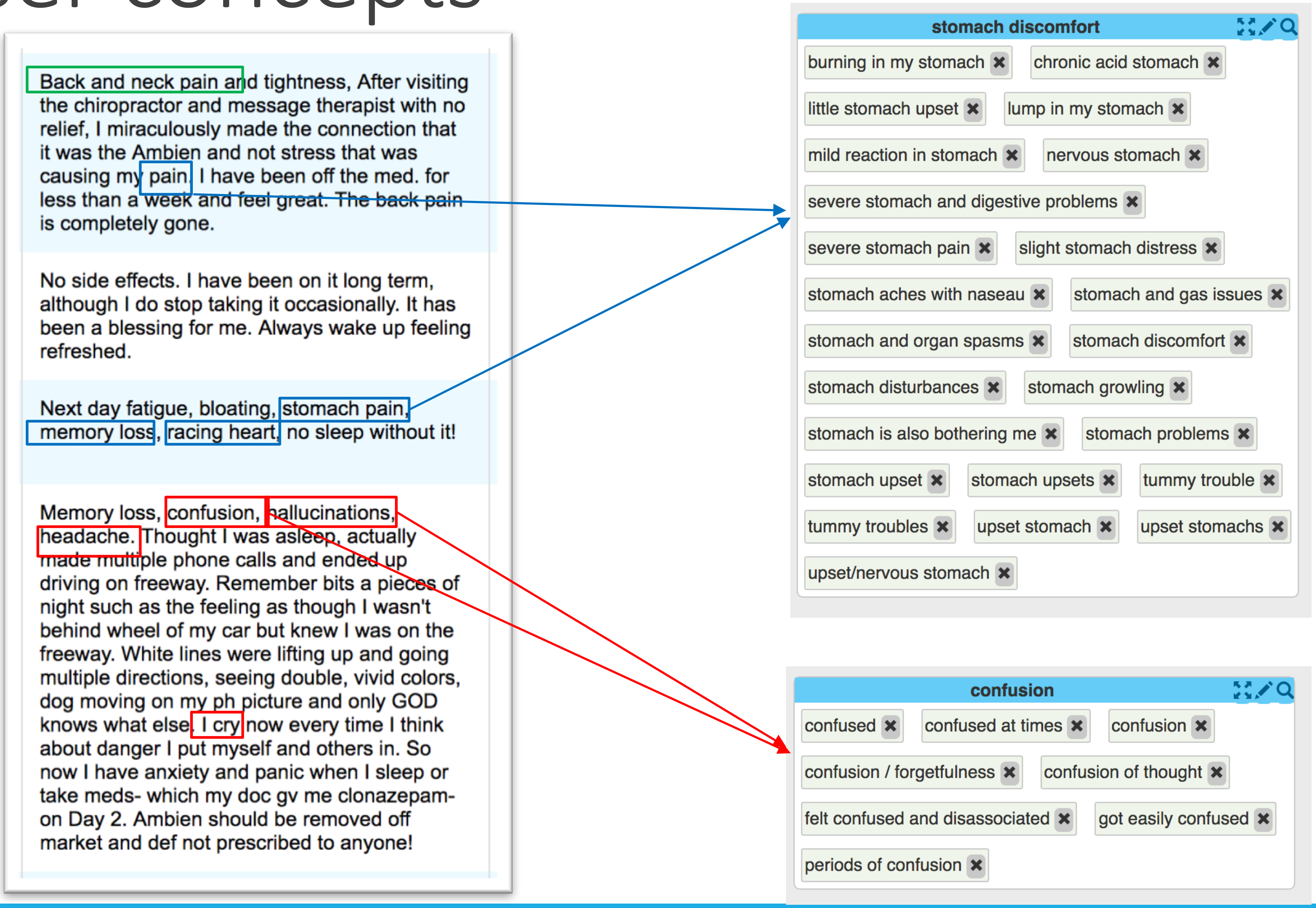
Problem Statement

Ontology Alignment

Target Ontology



User concepts



Back and neck pain and tightness, After visiting the chiropractor and message therapist with no relief, I miraculously made the connection that it was the Ambien and not stress that was causing my pain. I have been off the med. for less than a week and feel great. The back pain is completely gone.

No side effects. I have been on it long term, although I do stop taking it occasionally. It has been a blessing for me. Always wake up feeling refreshed.

Next day fatigue, bloating, stomach pain, memory loss, racing heart, no sleep without it!

Memory loss, confusion, hallucinations, headache. Thought I was asleep, actually made multiple phone calls and ended up driving on freeway. Remember bits a pieces of night such as the feeling as though I wasn't behind wheel of my car but knew I was on the freeway. White lines were lifting up and going multiple directions, seeing double, vivid colors, dog moving on my ph picture and only GOD knows what else. I cry now every time I think about danger I put myself and others in. So now I have anxiety and panic when I sleep or take meds- which my doc gv me clonazepam- on Day 2. Ambien should be removed off market and def not prescribed to anyone!

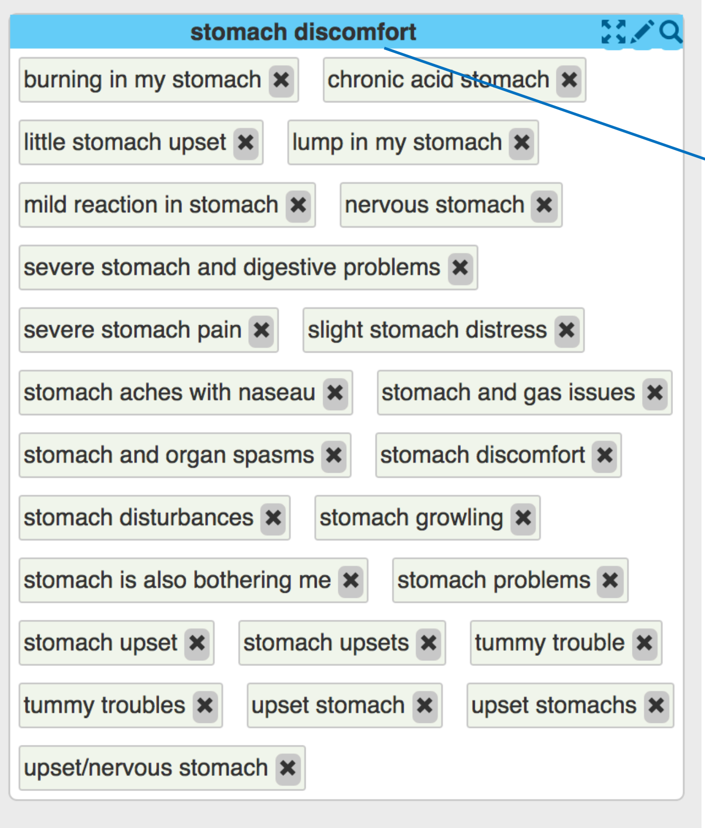
stomach discomfort

burning in my stomach ✖ chronic acid stomach ✖
little stomach upset ✖ lump in my stomach ✖
mild reaction in stomach ✖ nervous stomach ✖
severe stomach and digestive problems ✖
severe stomach pain ✖ slight stomach distress ✖
stomach aches with naseau ✖ stomach and gas issues ✖
stomach and organ spasms ✖ stomach discomfort ✖
stomach disturbances ✖ stomach growling ✖
stomach is also bothering me ✖ stomach problems ✖
stomach upset ✖ stomach upsets ✖ tummy trouble ✖
tummy troubles ✖ upset stomach ✖ upset stomachs ✖
upset/nervous stomach ✖

confusion

confused ✖ confused at times ✖ confusion ✖
confusion / forgetfulness ✖ confusion of thought ✖
felt confused and disassociated ✖ got easily confused ✖
periods of confusion ✖

Ontology Population



stomach discomfort

burning in my stomach ✖ chronic acid stomach ✖

little stomach upset ✖ lump in my stomach ✖

mild reaction in stomach ✖ nervous stomach ✖

severe stomach and digestive problems ✖

severe stomach pain ✖ slight stomach distress ✖

stomach aches with naseau ✖ stomach and gas issues ✖

stomach and organ spasms ✖ stomach discomfort ✖

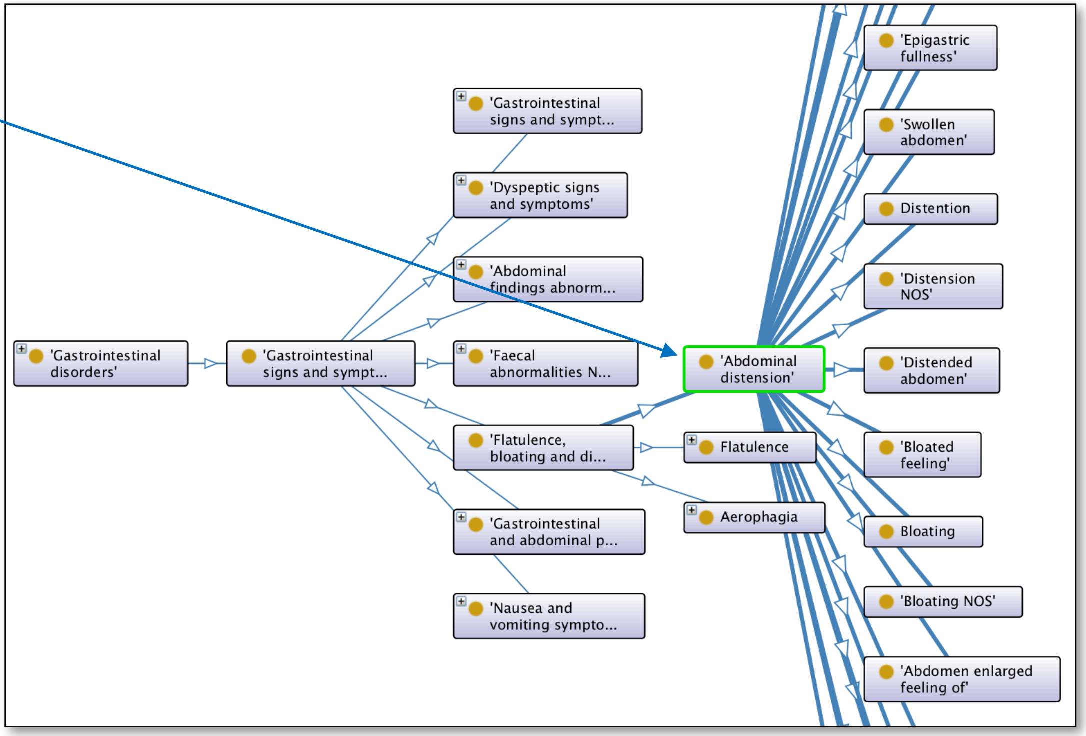
stomach disturbances ✖ stomach growling ✖

stomach is also bothering me ✖ stomach problems ✖

stomach upset ✖ stomach upsets ✖ tummy trouble ✖

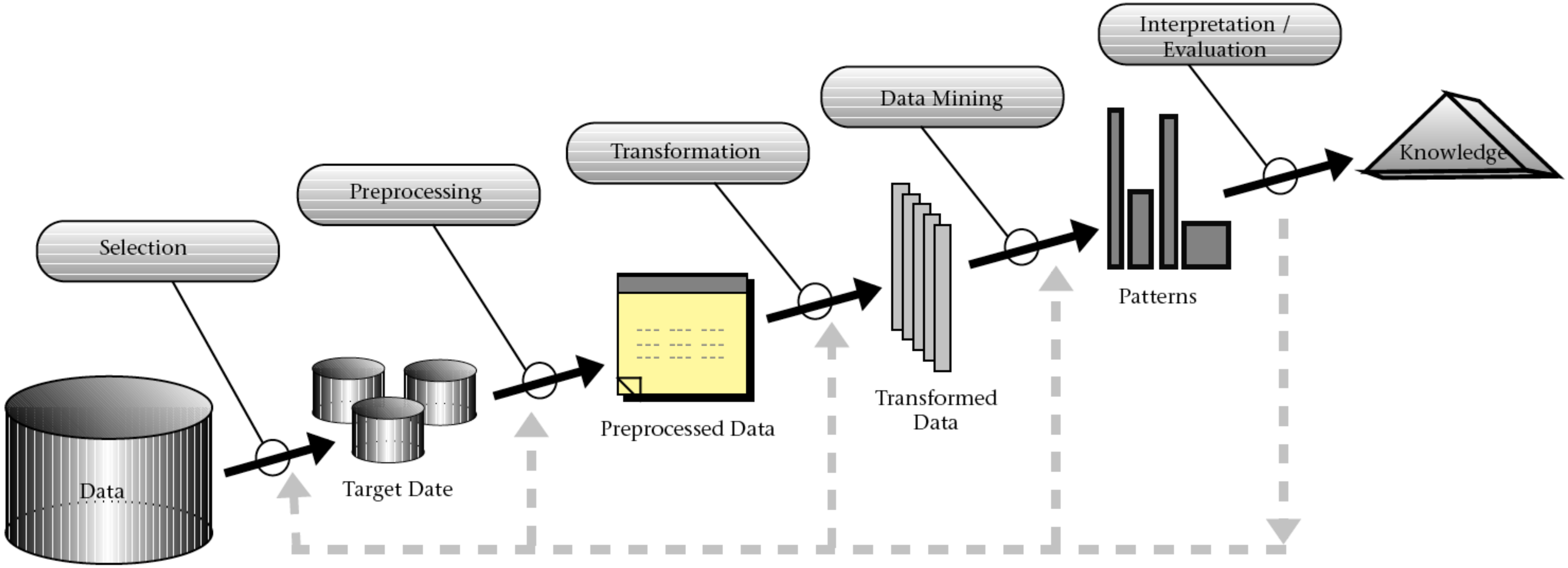
tummy troubles ✖ upset stomach ✖ upset stomachs ✖

upset/nervous stomach ✖


'Gastrointestinal disorders'
'Gastrointestinal signs and sympt...
'Gastrointestinal signs and sympt...
'Dyspeptic signs and symptoms'
'Abdominal findings abnorm...
'Faecal abnormalities N...
'Flatulence, bloating and di...
'Gastrointestinal and abdominal p...
'Nausea and vomiting sympto...
'Abdominal distension'
Flatulence
Aerophagia
'Epigastric fullness'
'Swollen abdomen'
Distention
'Distension NOS'
'Distended abdomen'
'Bloated feeling'
Bloating
'Bloating NOS'
'Abdomen enlarged feeling of'

Knowledge Discovery Process

(Fayyad et al. 1996)



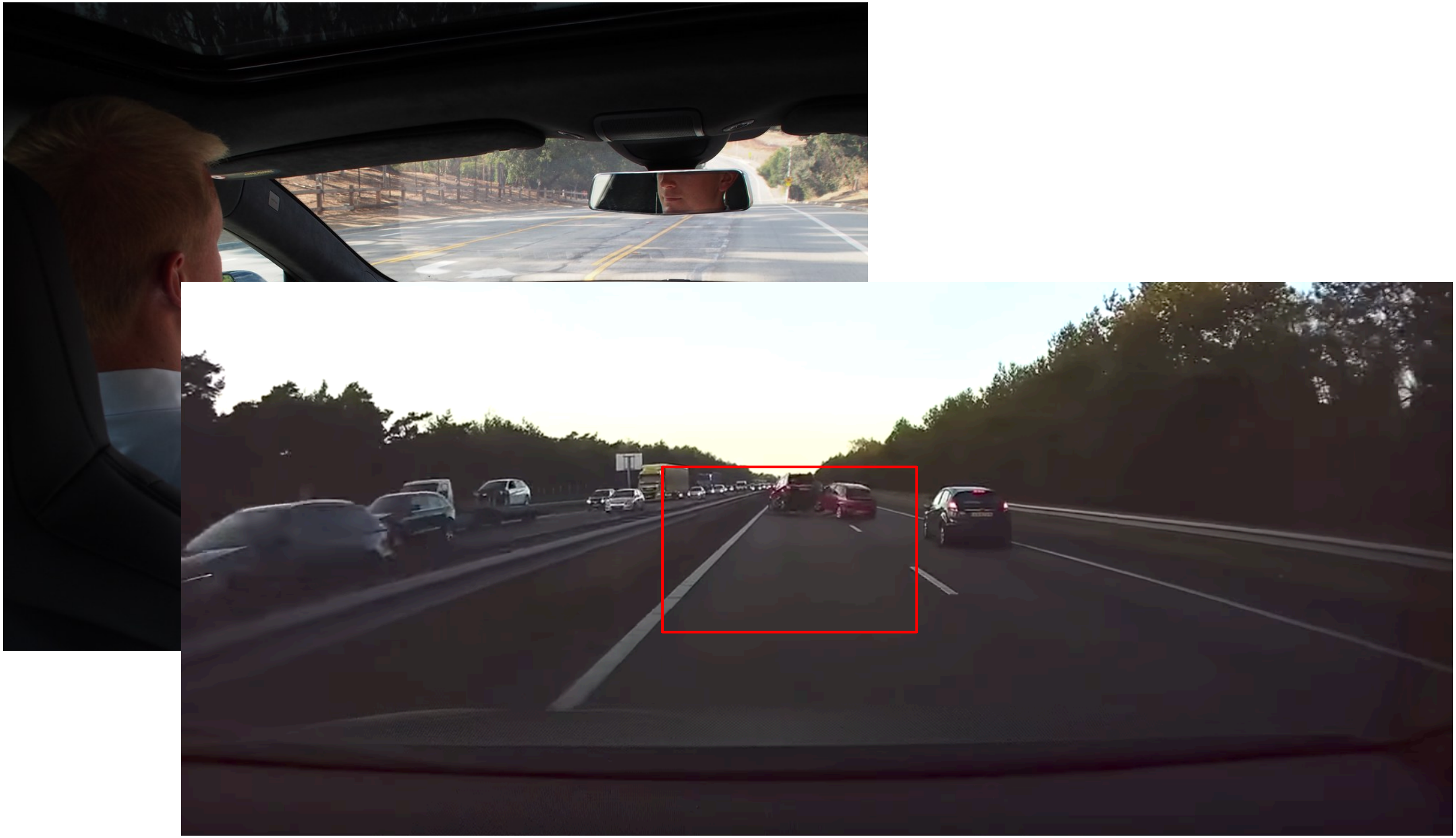
How good are machines?

~80% accuracy

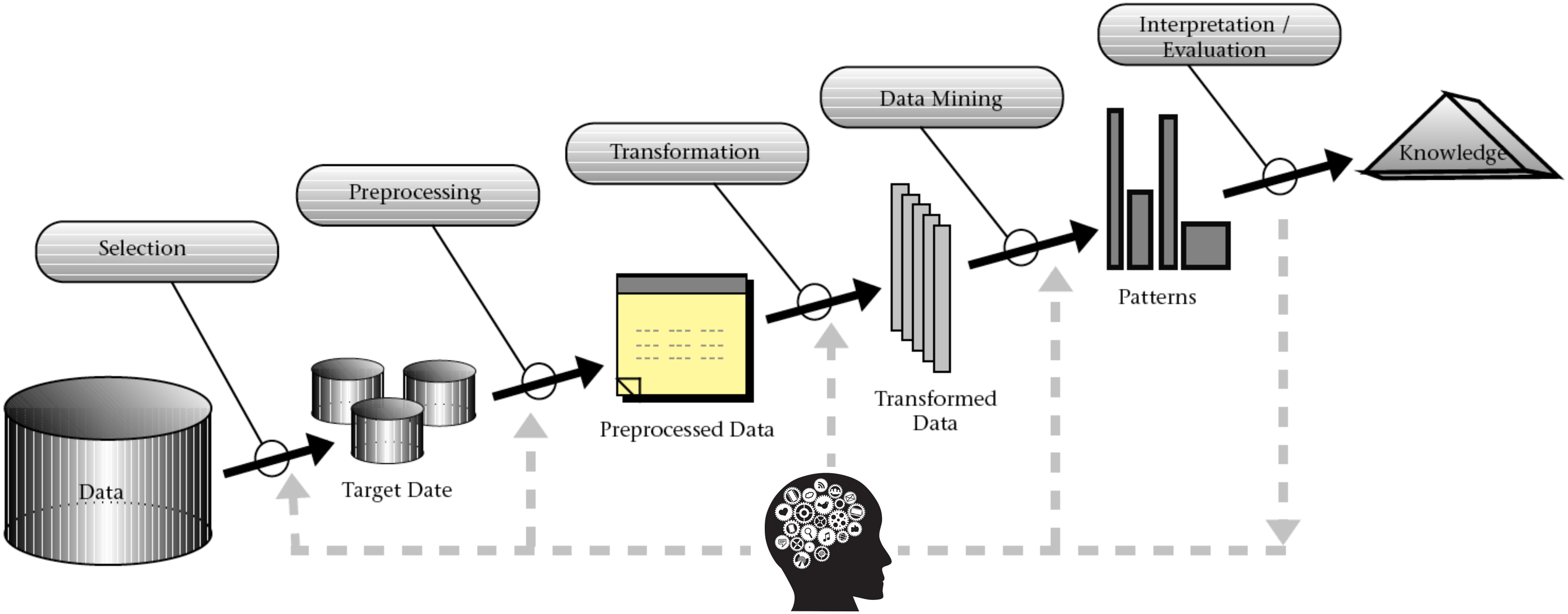
~85% accuracy

~90% accuracy

Is 80% enough?



Introduce the human-in-the-loop

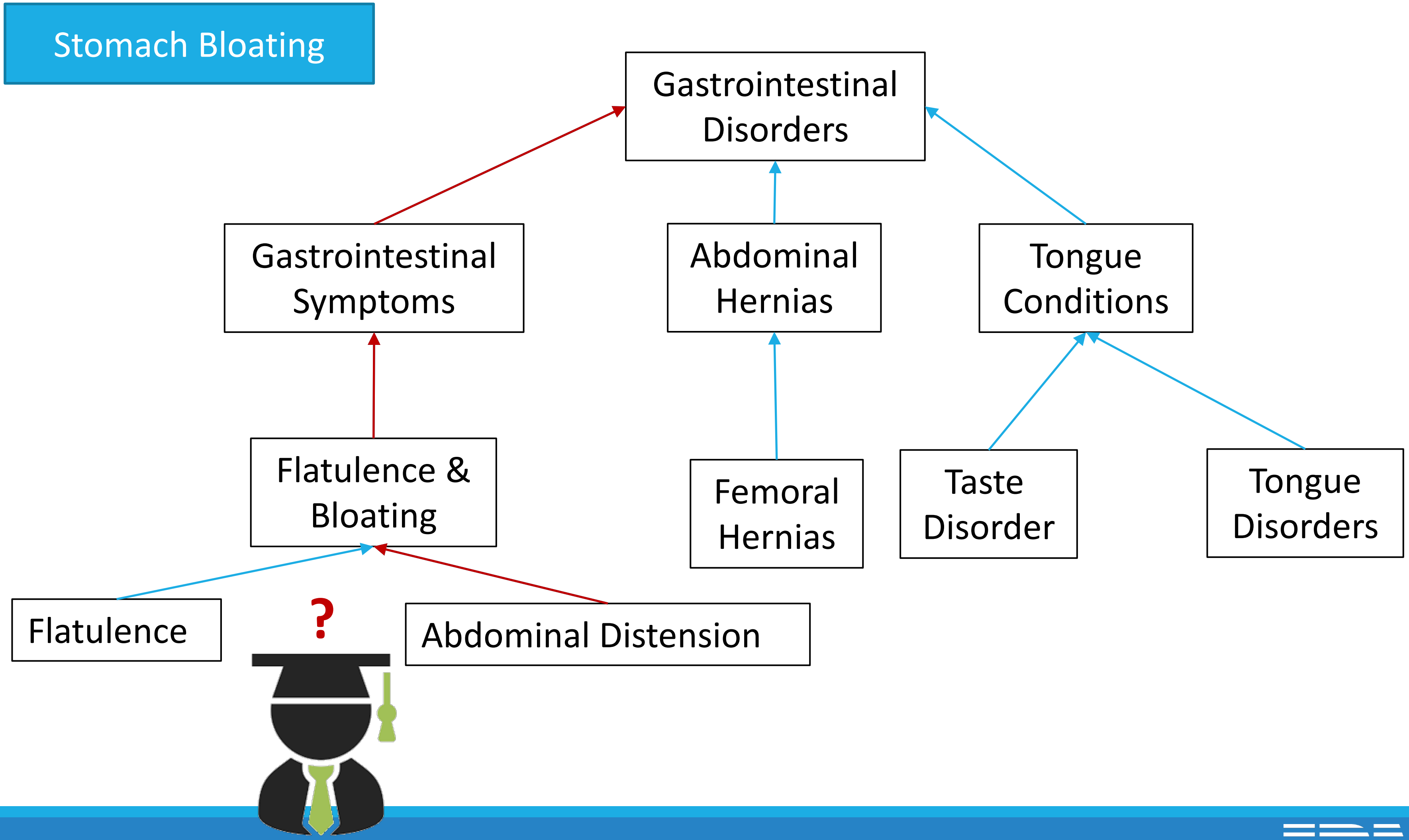


Introduce the human-in-the-loop

“Computers are incredibly fast, accurate, and stupid.
Human beings are incredibly slow, inaccurate, and brilliant.
Together they are powerful beyond imagination.”

Einstein never said that

Vision



Stomach Bloating

Gastrointestinal Disorders

Gastrointestinal Symptoms

Abdominal Hernias

Tongue Conditions

Flatulence & Bloating

Femoral Hernias

Taste Disorder

Tongue Disorders

Flatulence

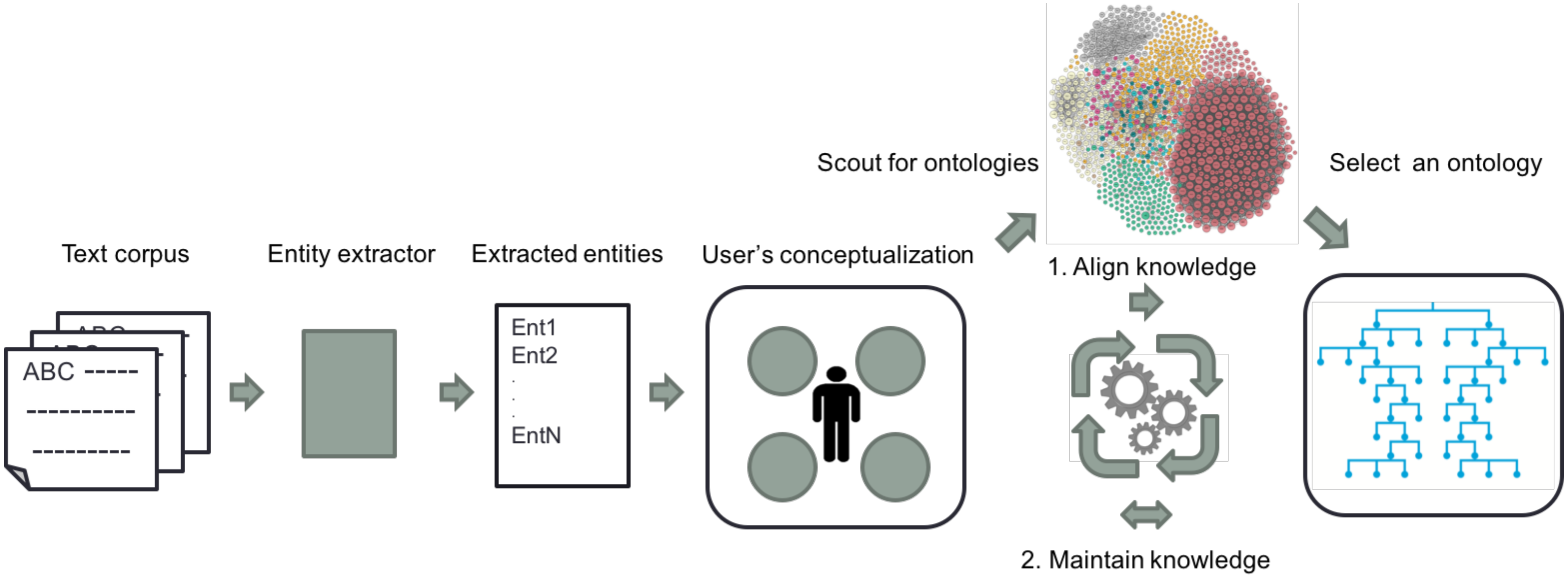
Abdominal Distension

Proposed Solution

Given initial user conceptualization, the methodology supports:

- Finding candidate ontologies
- **Aligning** the user's conceptualization to a target ontologies
  - novel hierarchical classification approach
- **Maintenance** lifecycle
  - **build** (create new concepts)
  - **change** (splitting/merging concept)
  - **grow** (adding new instances to each concept)
    - from target ontologies
    - new facts extracted from unstructured data

Implementation



Text corpus

Entity extractor

Extracted entities

User's conceptualization

Scout for ontologies

1. Align knowledge

2. Maintain knowledge

Select an ontology

Aligning Input with a Target Ontology

Identify available ontologies

- collective instance matching

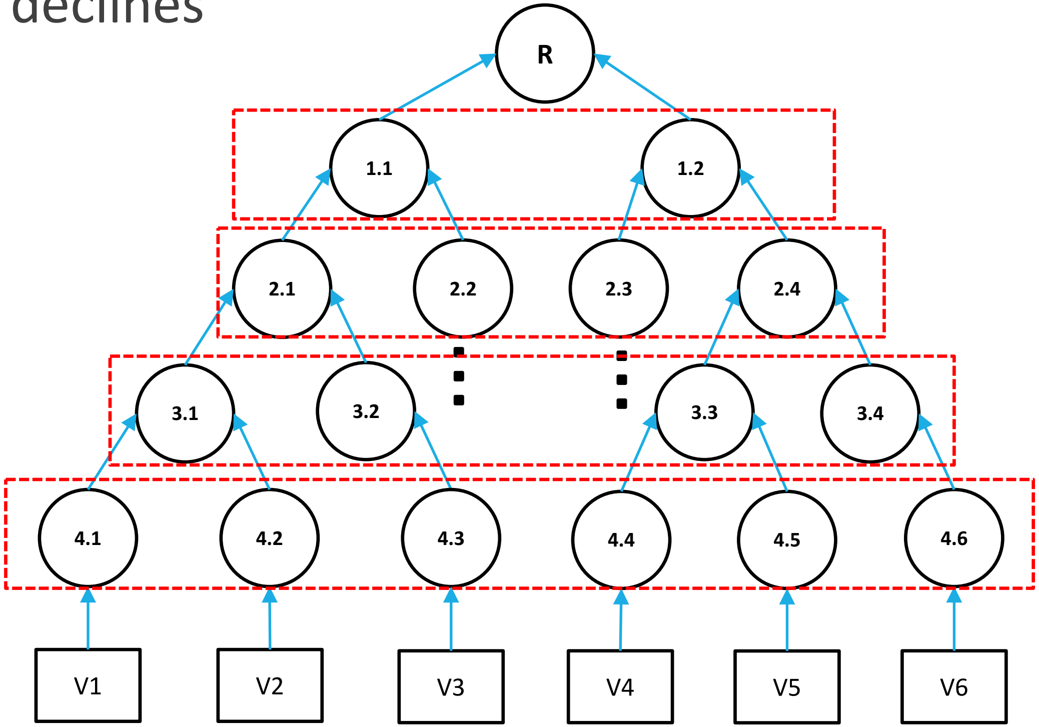
Align user conceptualization using ML models

- **Training data**: instances of the target ontology
- **Features** : domain-specific word embeddings
- **Classification strategies**
  - Flat hierarchical classification
  - Top-down local classifier per parent node
  - Combine flat hierarchical with top-down local classifier per parent node
- **Classifiers**
  - SVM, Random Forests, Logistic Regression, Convolutional Neural Network

Flat Hierarchical Model

One model **for each level** of the hierarchy

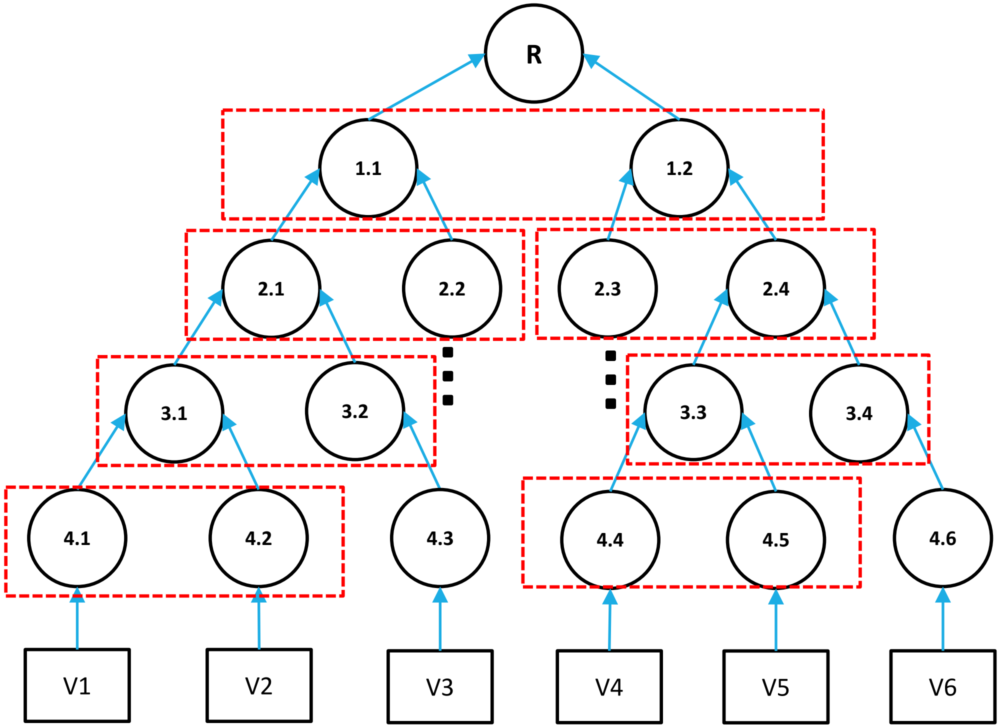
- Simple
- High model complexity down the hierarchy
- precision declines



Top-Down Local Classifiers

One model **for each parent** in the hierarchy

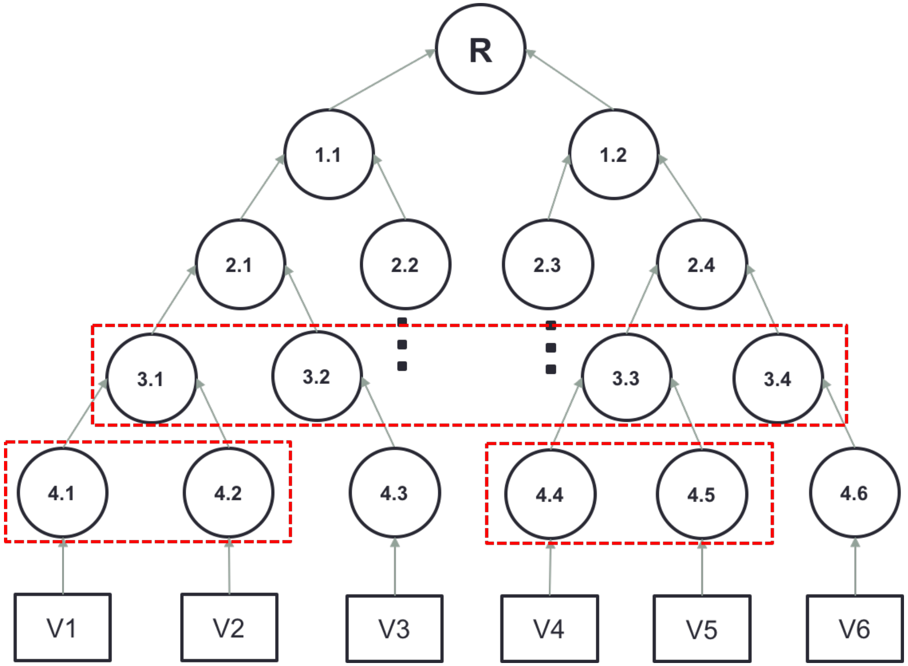
- Simple
- Error propagation through levels



Combine Both Models

Combine flat hierarchical models with top-down local classifier

- flat model for level L-1
- local model for level L



Ontology Maintenance

Adding new instances

- Use existing models

Reassigning Instances

- Leave-one-out validation

Generating new concepts

- If the class distribution is uniform then search for new concept

Merging concepts

- User's concepts aligned to the same target ontology concept should be merged

Concept splitting

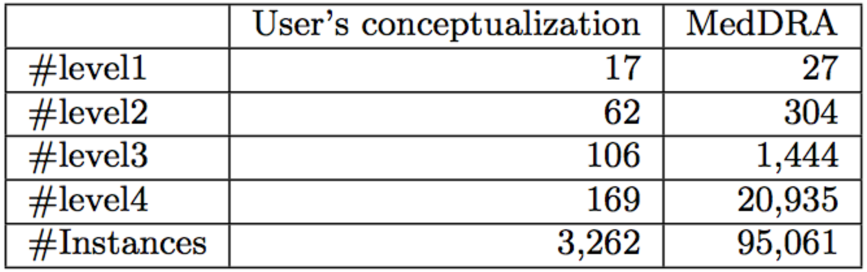
- Use hierarchical clustering
- Refine until a criteria is met

Evaluation - Alignment

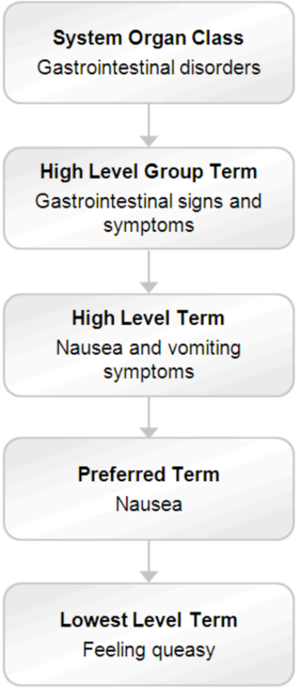
Task: label adverse drug events with preferred medical terms

Data:

- MedDRA ontology as a target ontology
- ADE groups extracted from “ask a patient blogs”



	User’s conceptualization	MedDRA
#level1	17	27
#level2	62	304
#level3	106	1,444
#level4	169	20,935
#Instances	3,262	95,061


System Organ Class
Gastrointestinal disorders
High Level Group Term
Gastrointestinal signs and symptoms
High Level Term
Nausea and vomiting symptoms
Preferred Term
Nausea
Lowest Level Term
Feeling queasy

Evaluation metric: HITS@10 = proportion of correct mappin top 10 ranked suggestions

- Evaluate per each level of the hierarchy

Evaluation – Ontology Alignment

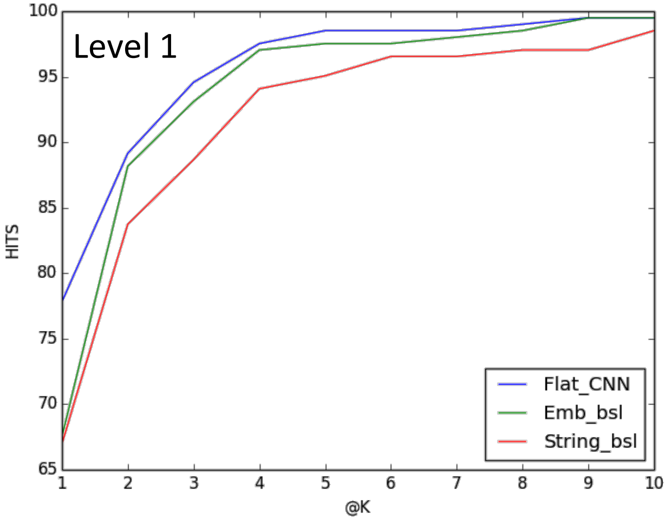
Baselines:

- String-based average-link matching
- Word embeddings
- LDA topic modeling

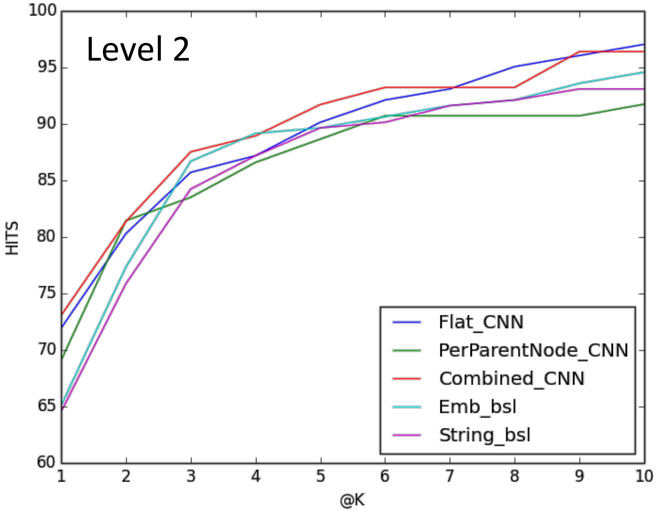
Evaluation metric: HITS@10

- proportion of correct mappings that appear in the top 10 ranked suggestions
- Evaluate per each level of the hierarchy

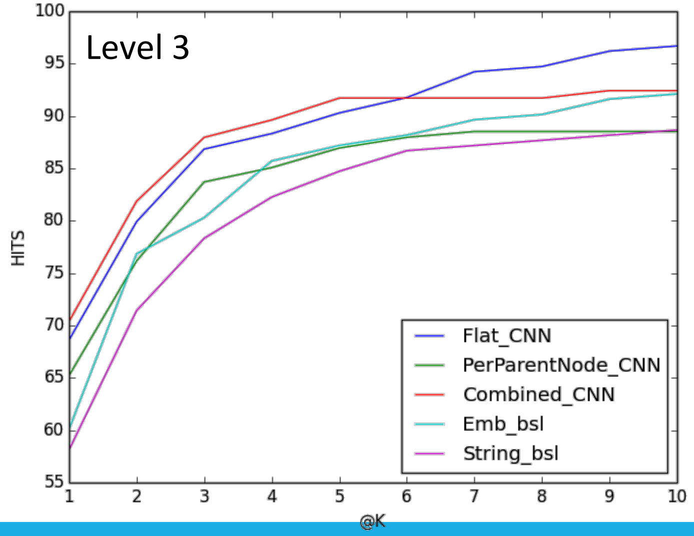
Results



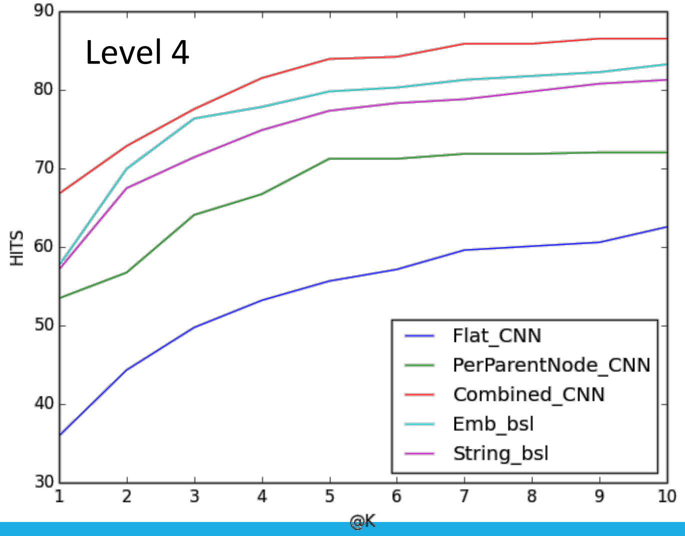
Level 1
HITS
@K
100
95
90
85
80
75
70
65
1
2
3
4
5
6
7
8
9
10
Flat_CNN
Emb_bsl
String_bsl



Level 2
HITS
@K
100
95
90
85
80
75
70
65
60
1
2
3
4
5
6
7
8
9
10
Flat_CNN
PerParentNode_CNN
Combined_CNN
Emb_bsl
String_bsl

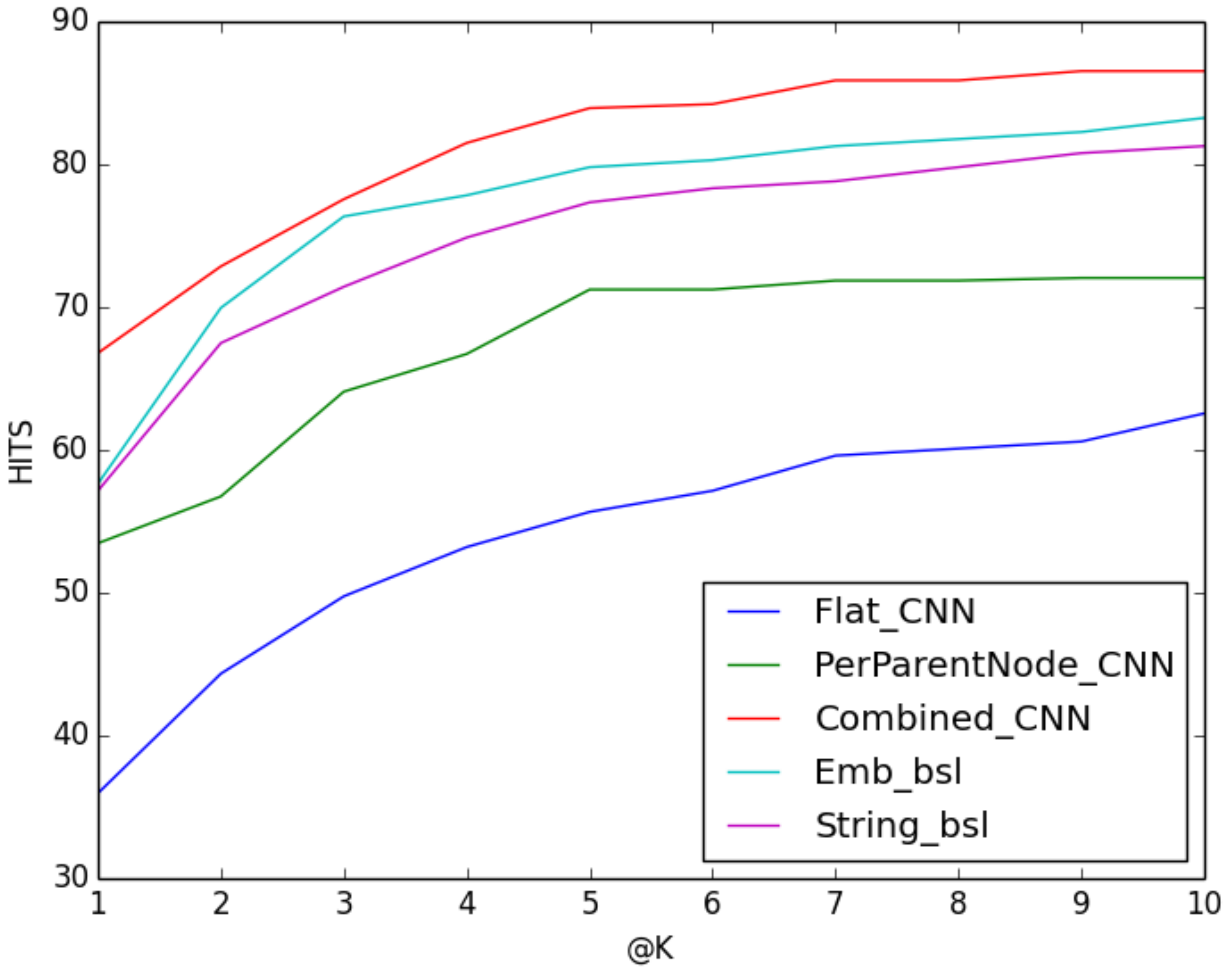


Level 3
HITS
@K
100
95
90
85
80
75
70
65
60
55
1
2
3
4
5
6
7
8
9
10
Flat_CNN
PerParentNode_CNN
Combined_CNN
Emb_bsl
String_bsl



Level 4
HITS
@K
90
80
70
60
50
40
30
1
2
3
4
5
6
7
8
9
10
Flat_CNN
PerParentNode_CNN
Combined_CNN
Emb_bsl
String_bsl

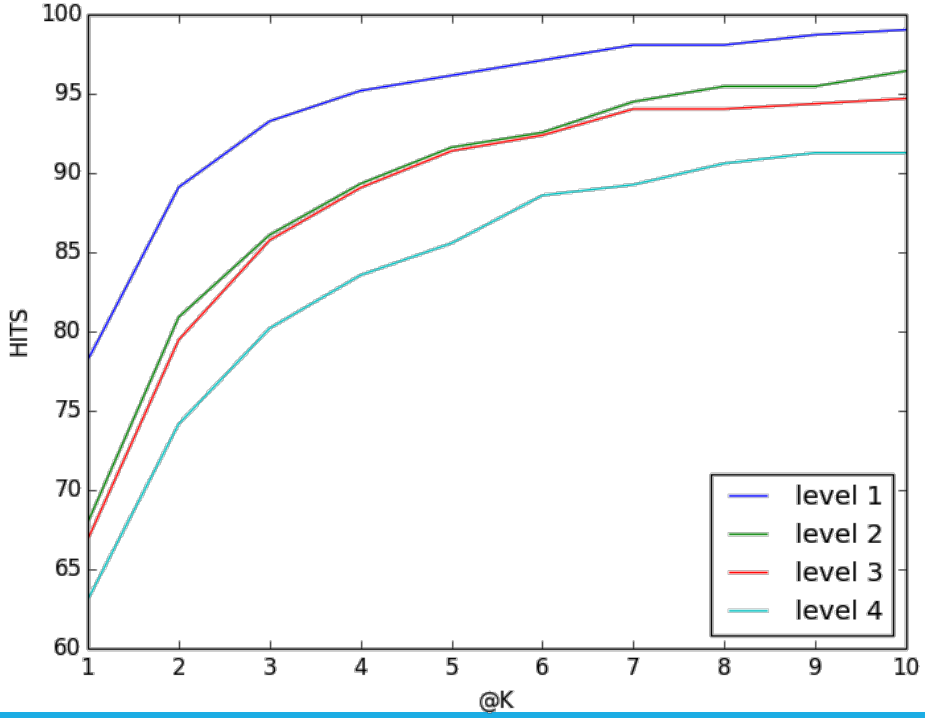
Results: Level 4



Evaluation – Ontology Maintenance

Adding new instances – evaluate how precise the models can add new instances to the already aligned concepts

- Retrieved 298 **new** ADE from askapatient.com
- Measure HITS@k for each level of the hierarchy



Evaluation – Ontology Maintenance

Adding new concepts – evaluate the model’s ability to notify the user to add a new concept

- Evaluation
- Selected 500 MedDRA instances that don’t belong to the user’s conceptualization (positive instances), and 500 instances that belong to the user’s conceptualization

Results:

- Precision: 73.8%
- Recall: 84.6%
- F-score: 78.83%

$$E(x) = \sum_{i=0} kP(C_1|x) * log_2 P(C_1|x) > 1$$

Evaluation – Ontology Maintenance

Adding new concepts

- model's ability to suggest the user to add a new concept
- evaluation
  - 500 MedDRA instances that don't belong to the user's conceptualization (positive instances)
  - 500 instances that belong to the user's conceptualization

Results:

- Precision: 73.8%
- Recall: 84.6%
- F-score: 78.83%

Evaluation – Ontology Maintenance

Re-assigning Instances: evaluate the model's ability to reassign instances to other concepts.

The model identified 82 instances to be reassigned, from which 67 (81.7%) were accepted by the medical doctor

Examples:

- User errors: "*stomach aches*" was assigned to "*Emotional disorder*", which should be assigned to "*Abdominal distension*"
- Better matches: "*sensitivity to light*" was assigned to "Visual impairment", which was later reassigned to "Photophobia"

Further Use-Cases

- Maintain health and medical data
  - Adverse drug reactions
  - Drug brands
- Maintain e-shop product catalog and taxonomy
  - Map new features to an existing product catalog
  - Map new products in the product taxonomy
- Social media analysis
  - Identifying new trends
- Reviews analysis
  - Movies and actors

Conclusion

User-centric ontology population

Human-in-the-loop for each step

- Building, connecting and maintaining their conceptualization, using available ontologies

Novel **hierarchical classification model**

- dynamically refined based on user interaction

The approach supports the user to achieve **nearly perfect performance**

The user has full control on their level of involvement in the process

- Trade-off between involvement/cost/time and performance/quality of results

User-Centric Ontology Population

Kenneth Clarkson, **Anna Lisa Gentile**, Daniel Gruhl,Petar Ristoski,Joseph Terdiman, Steve Welch

annalisa.gentile@ibm.com

@AnLiGentile