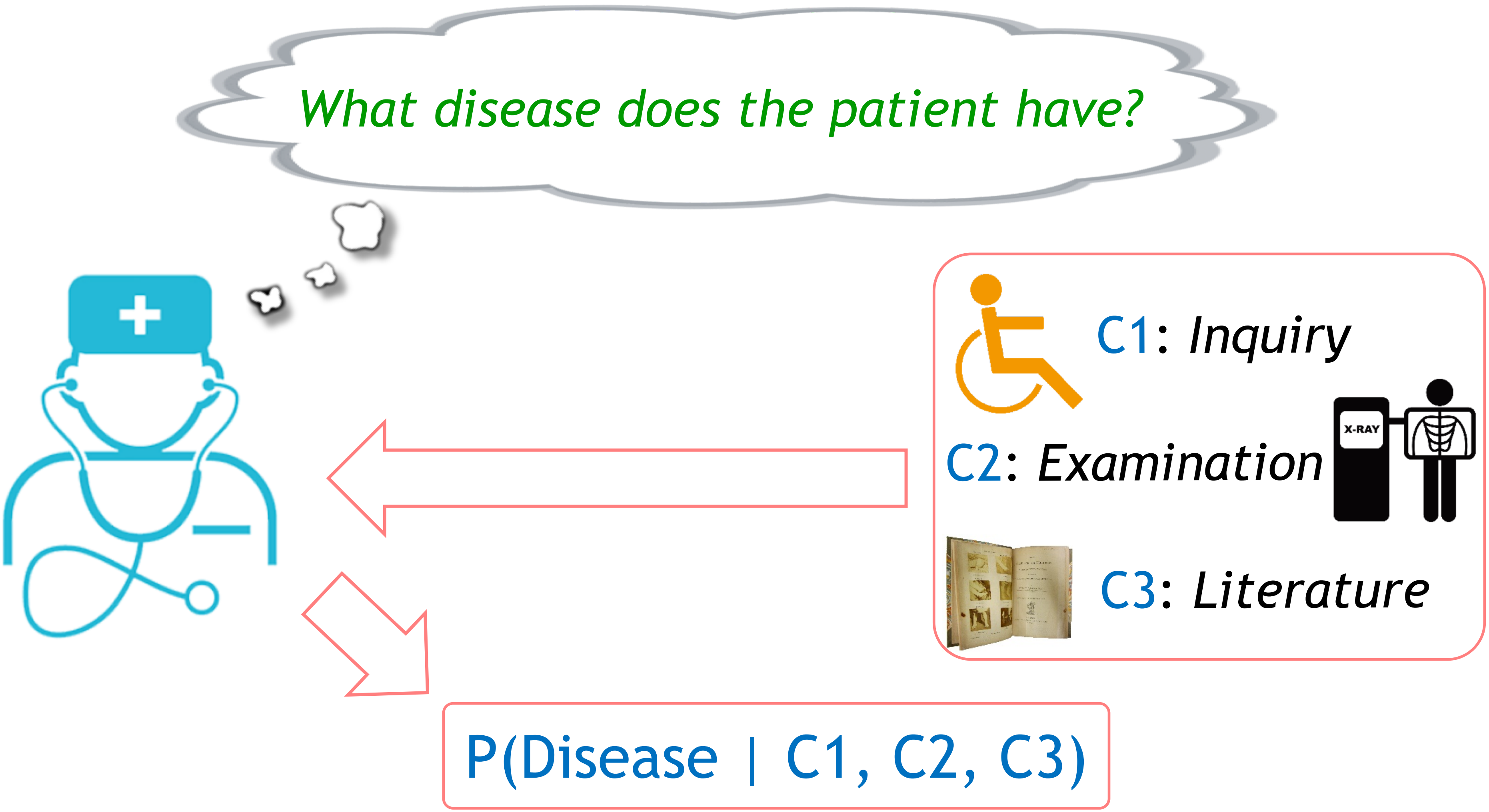
Towards Democratizing Data Science with AI-Powered Knowledge Engines

Yu Su
Microsoft Semantic Machines
The Ohio State University

Data-Driven Decision Making



Growing Gap between Human and Data

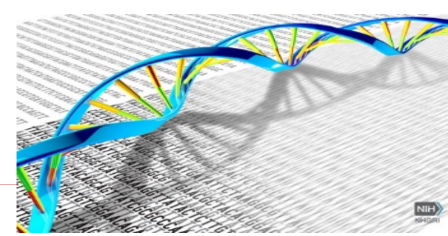
What disease does the patient have?

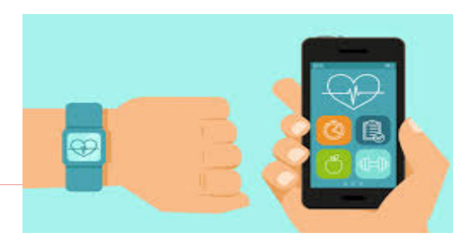
- EMR => Similar patients?
- Literature => New discoveries?
- Gene sequence => Suspicious mutations?
- … …

Ad-hoc information needs for on-demand decision making

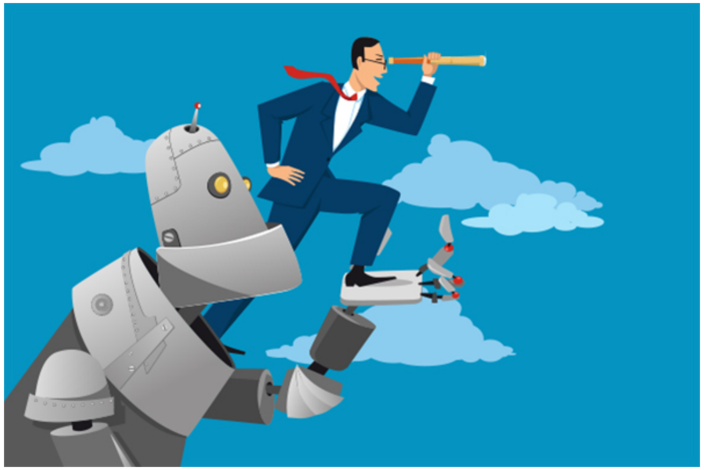
Massive, heterogeneous data

86.9% adoption (NEHRS 2015)	27M+ papers, >1M new/year (PubMed)	$1000 gene sequencing	24x7 monitoring





How to Democratize Data Science?



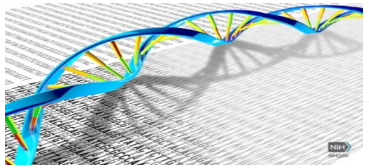
AI-Powered Knowledge Engine

Discoveries
Decisions
Actions

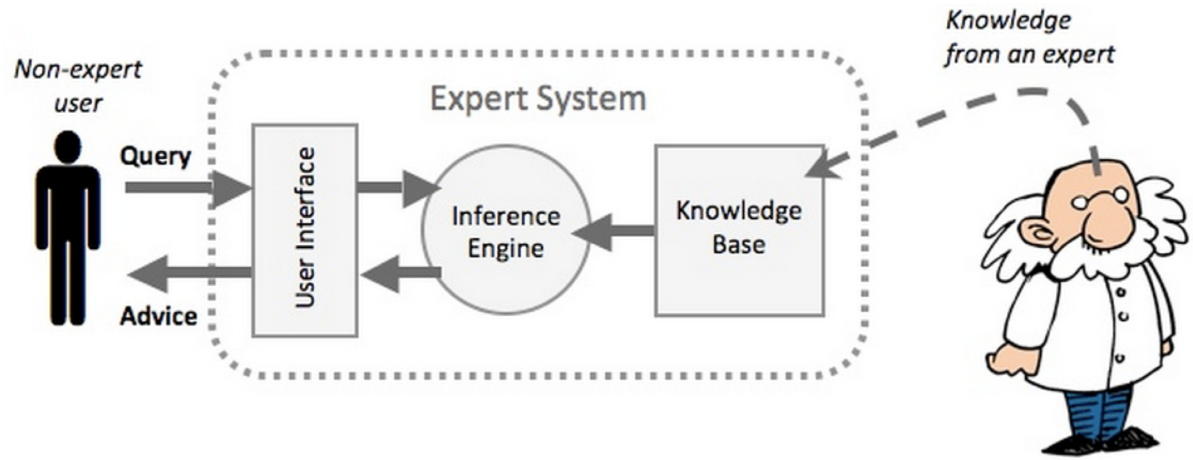
Bottleneck #2: Access

Bottleneck #3: Reasoning

Bottleneck #1: Knowledge

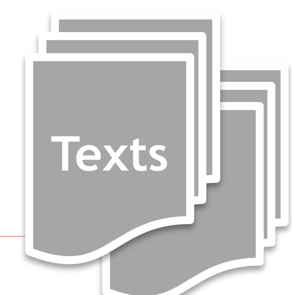


Knowledge Base



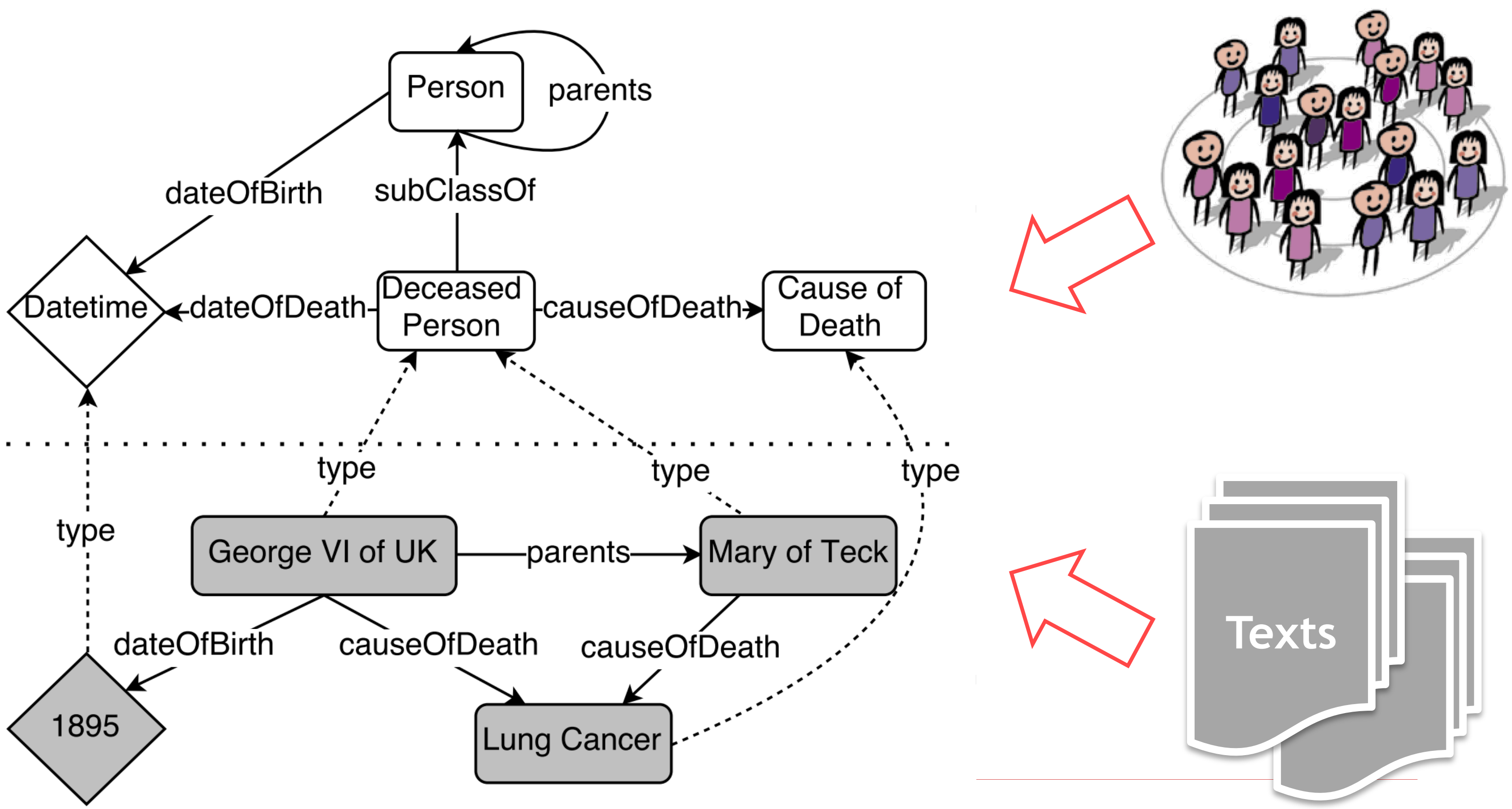
1970s-1990s

2000s-present

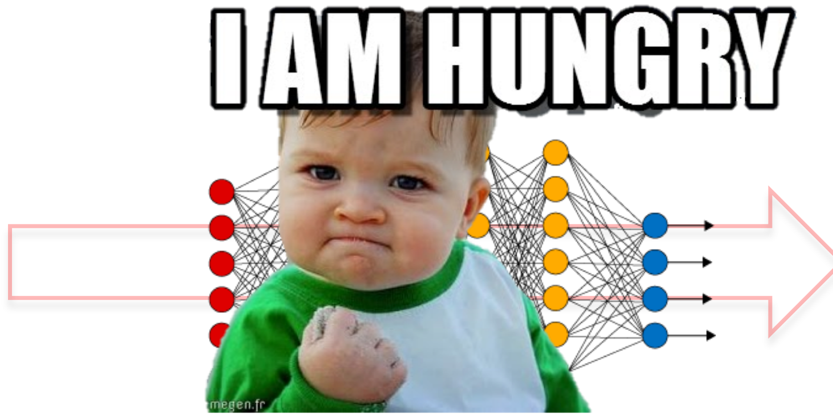


Knowledge Base

- Encyclopedic knowledge about concepts, entities and their relationships (facts)
  - Google Knowledge Graph: 570M entities and 18B facts (2014)



Methodology: Deep Learning with Weak Supervision

Text  Knowledge

Strong Supervision

- In-domain, on-task

Weak Supervision

- In-domain, off-task
- Out-of-domain, on-task
- Out-of-domain, off-task

KNOWLEDGE HARVESTING FROM MASSIVE TEXT

Knowledge Base Construction from Text

- Entity recognition and linking
- Relation extraction: **binary**, n-ary (event)

High-throughput cell-based screening of 4910 known drugs and drug-like small molecules identifies Disulfiram as an inhibitor of prostate cancer cell growth

Relation: inhibit

Subject	Object	Probability
Disulfiram	Prostate Cancer	0.85
	...	

"Alcohol-abuse drug disulfiram targets cancer via p97 segregase adaptor NPL4"
Skrott et al. *Nature* 552.7684 (2017): 194.

Scalable Relation Extraction with Distant Supervision

place_of_birth: (Michael Jackson, US)

Distant Supervision

Knowledge Bases

In-domain, off-task supervision

WIKIPEDIA
The Free Encyclopedia

DBpedia

Freebase

Unified Medical Language System ®

OncoKB

Training

Michael Jackson was born in the US.

Born in the US, Michael Jackson was one of …

I visited the birthplace of Michael Jackson in Gary, Indiana, United Stated.

Learn & Generalize

Testing

Barack Obama was born in the US.

… nearby Stratford, birthplace of Justin Bieber …

The German-born American physicist Albert Einstein revolutionized …

Extraction

(Barack Obama, US)
(Justin Bieber, Stratford)
(Albert Einstein, Germany)

E.g., [Mintz et al., 2009], [Riedel et al., 2010], [Zeng et al., 2015], [Lin et al., 2016], …

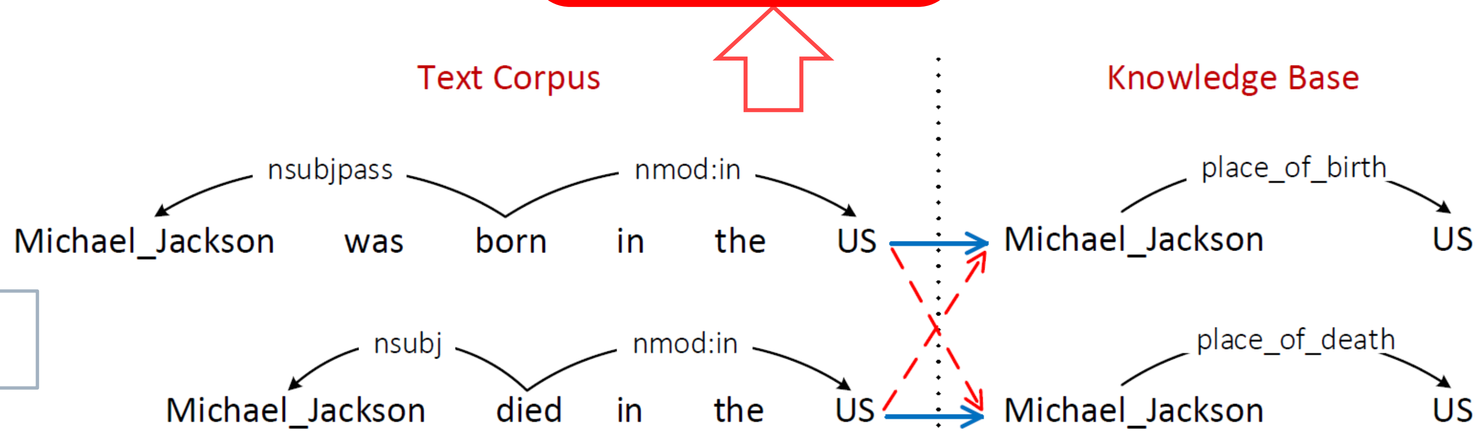
Global Statistics of Relations

- Number of co-occurrences of KB-textual relation pairs in the entire corpus

Meaning of this textual relation!

Global

	$\xleftarrow{\text{nsubjpass}}$ *born* $\xrightarrow{\text{nmod:in}}$	$\xleftarrow{\text{nsubj}}$ *died* $\xrightarrow{\text{nmod:in}}$
place_of_birth	0.73	0.04
nationality	0.15	0.06
place_of_death	0.01	0.89
...	...	...



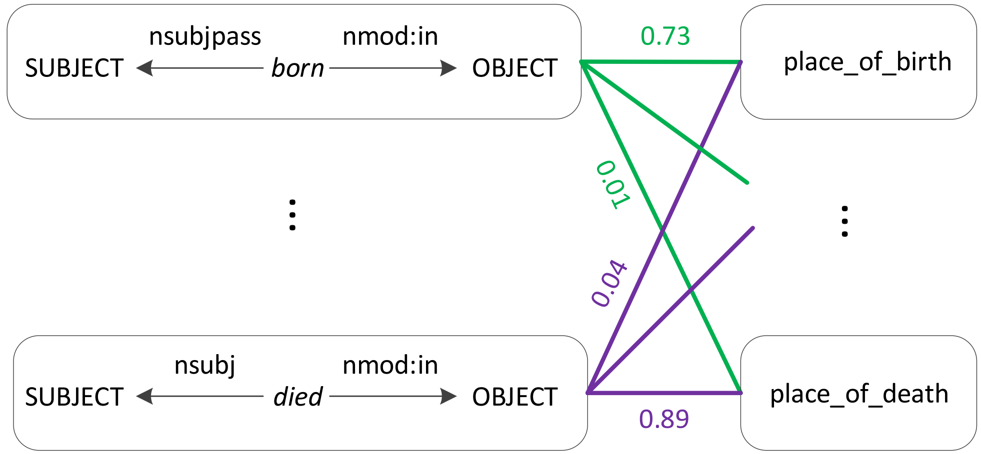
Local

[NAACL'18]

Word embedding analogy: GloVe (global statistics) vs. Word2vec (local statistics)

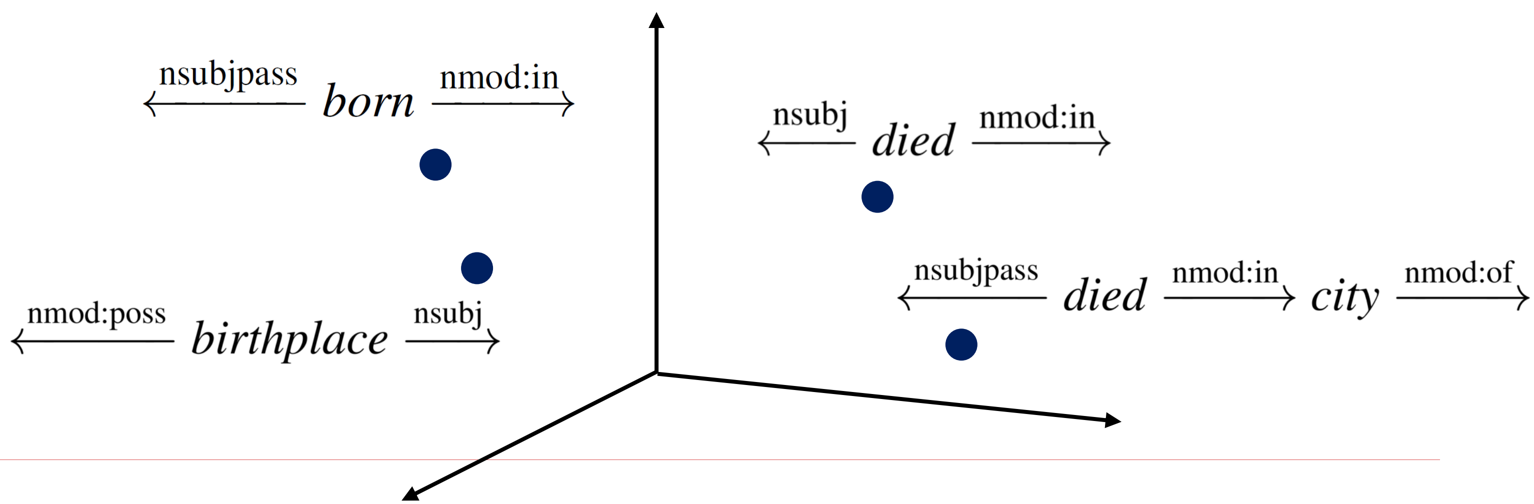
Textual Relation Embedding with Global Statistics

ClueWeb: 500M web documents



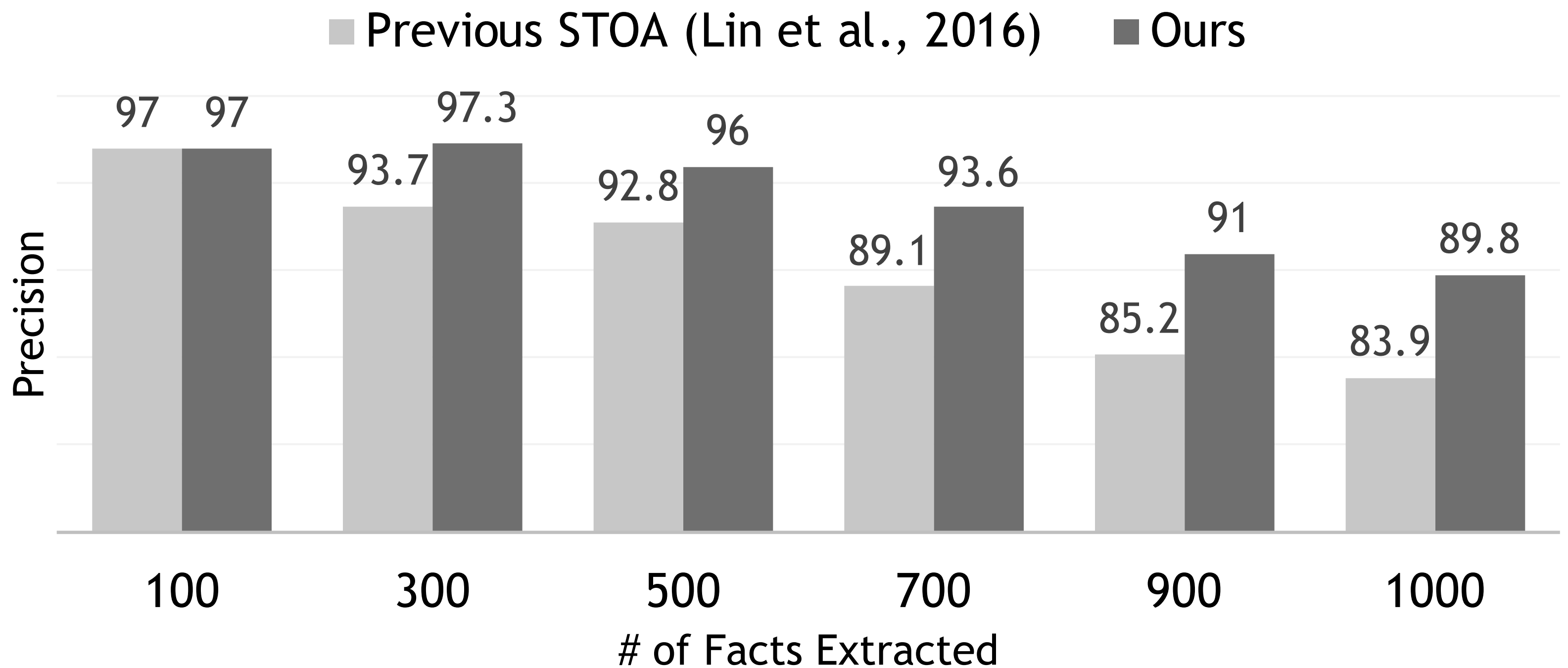
Freebase: 45M entities, 3B facts

Target Embedding ϕ



Evaluation on Newswire Corpus

- **Dataset:** New York Times corpus, 53 target relations
  - place_of_birth, place_of_death, founder_of, employee_of, etc.
- The learned textual relation embedding improves the STOA method by 5.9% (top 1,000 extracted facts)



Knowledge Base Construction: Food for Thought

- (Open-world) probabilistic KBs
  - Model uncertainties of the real world
- Multi-modal KBs
  - Images, audio, video, temporal-special info
- (Dynamic) distributed KBs
  - Personal KBs (at edge) + a public KB (in the cloud)

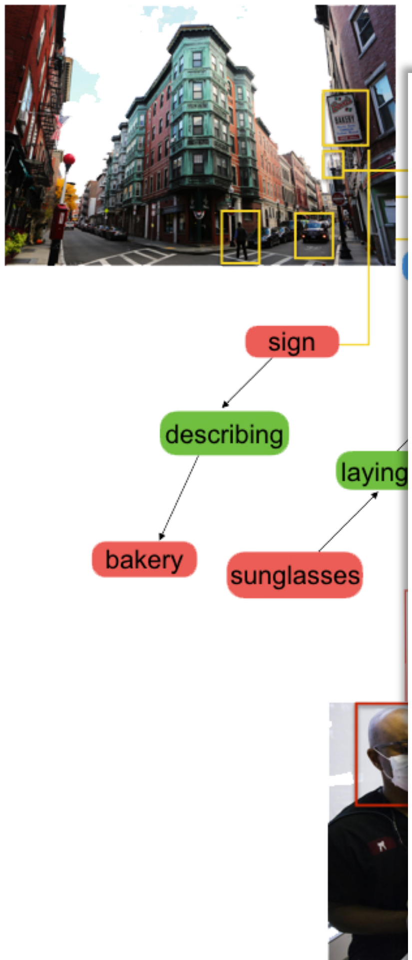
(Open-World) Probabilistic KBs

- KB: place_of_birth(John, United States)
- Query: "*Does John speak English?*"
- Closed-world assumption: "*No.*"
- Open-world assumption: "*I don't know.*"
- Open-world probabilistic KB: "*99% yes.*"

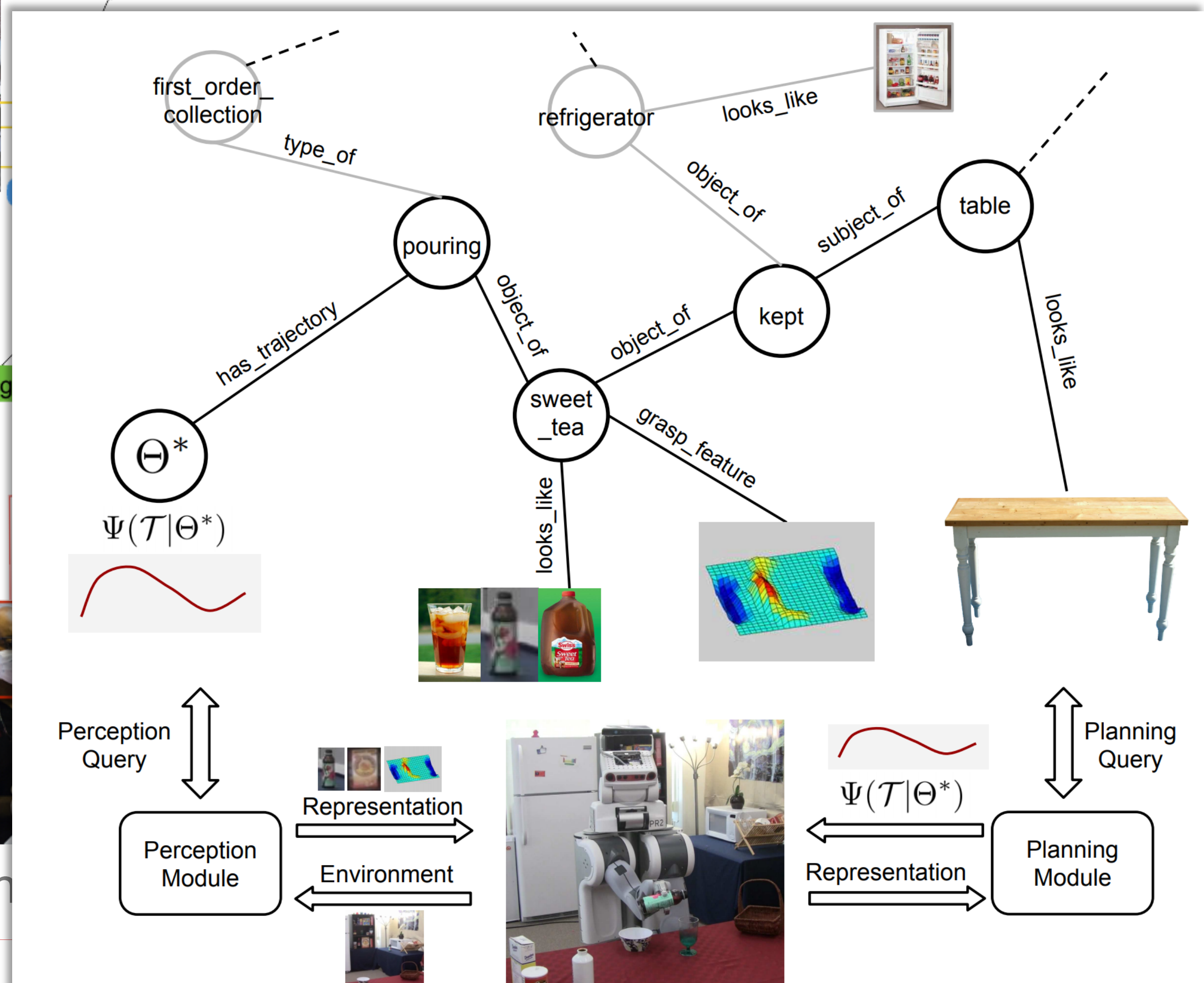
- Challenges
  - Uncertainty modeling and probability calibration
  - Efficient querying
  - Combination of logic-based reasoning and machine learning based reasoning

Some examples: YAGO, NELL, Google Knowledge Vault

Multi-Model KBs

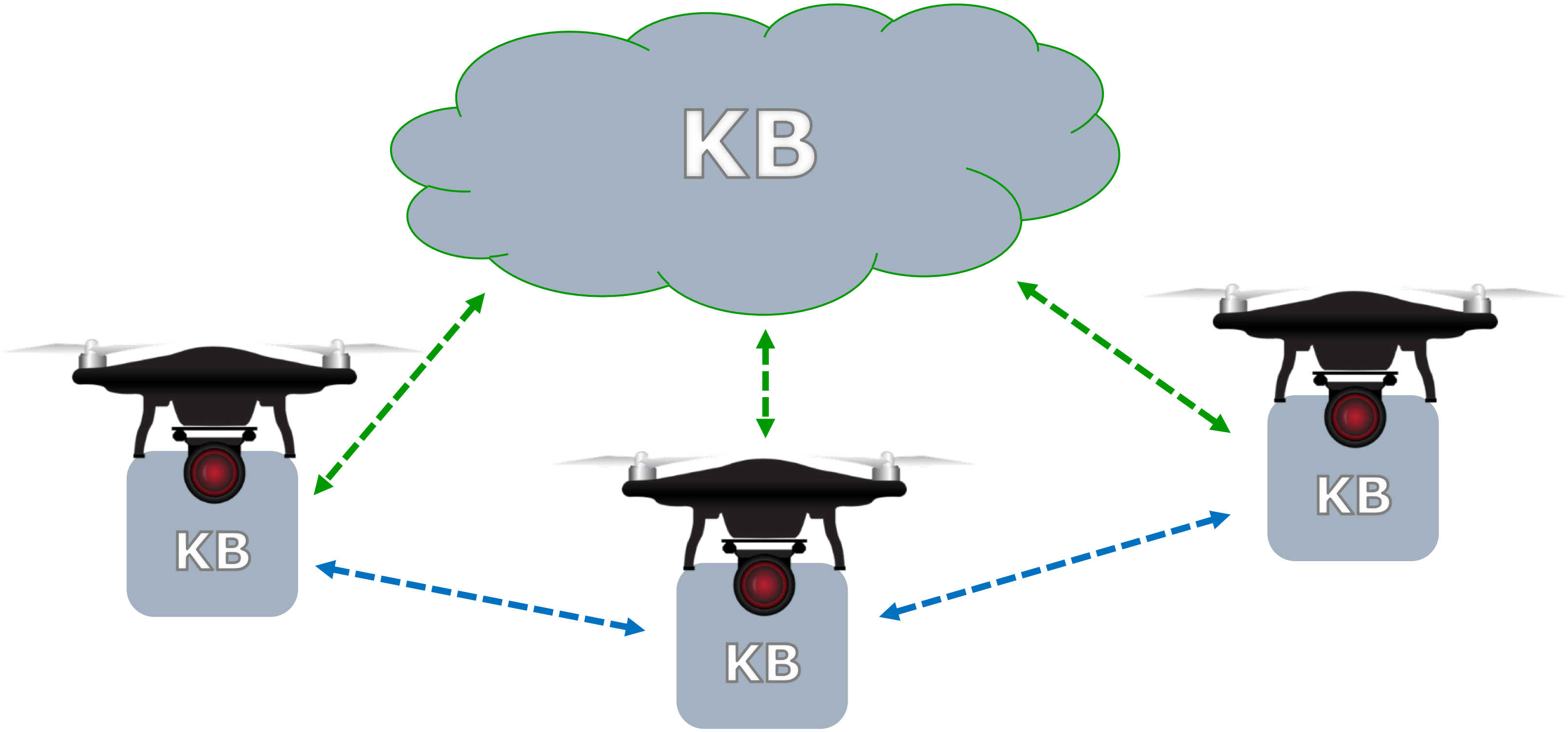


source: RoboBrain by Saxena et al.



source: visualgen

(Dynamic) Distributed KBs



NATURAL LANGUAGE INTERFACE

Writing formal queries is a pain…

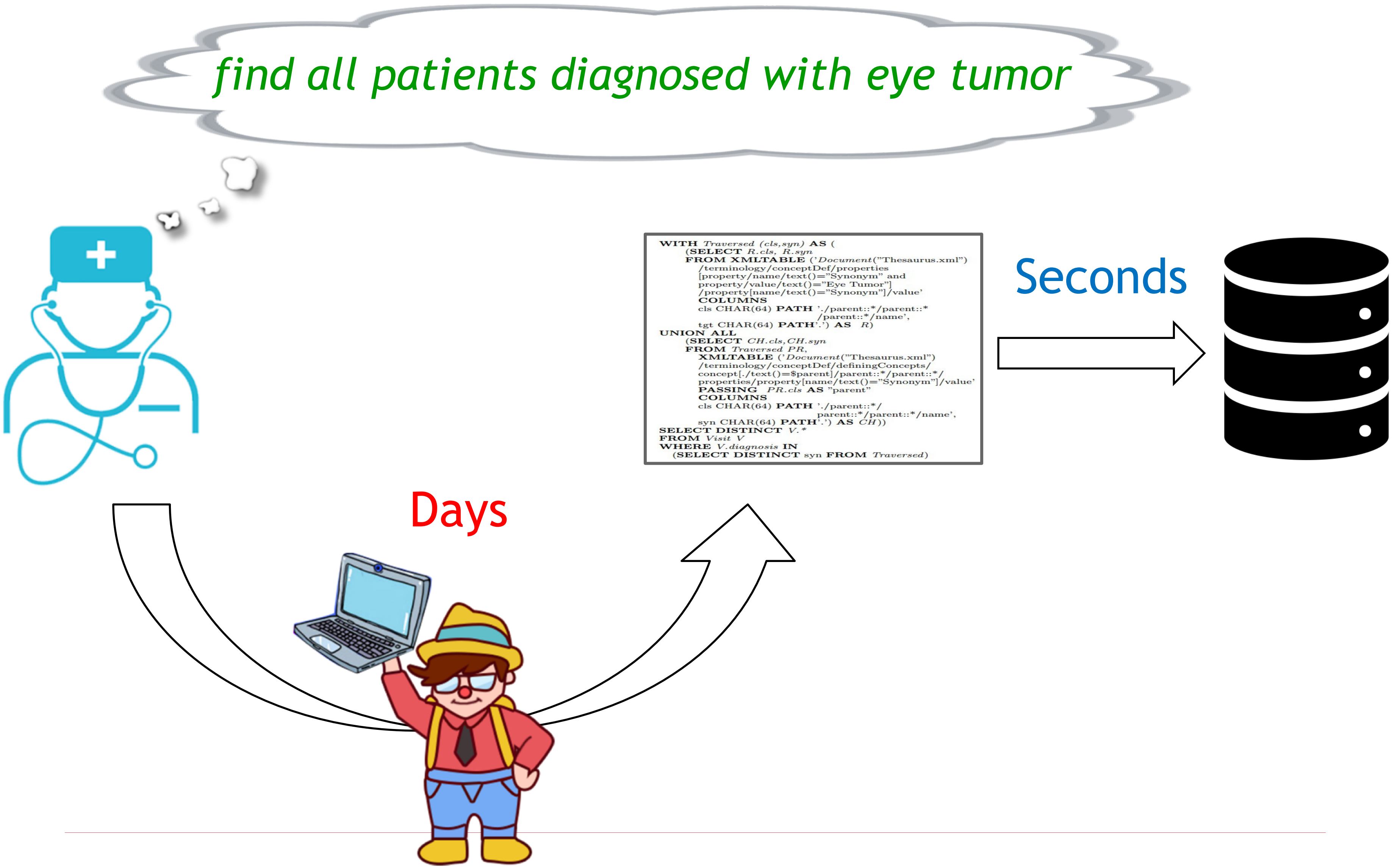
"find all patients diagnosed with eye tumor"

```
WITH Traversed (cls,syn) AS (
     (SELECT R.cls, R.syn
     FROM XMLTABLE ('Document("Thesaurus.xml")
        /terminology/conceptDef/properties
        [property/name/text()="Synonym" and
        property/value/text()="Eye Tumor"]
        /property[name/text()="Synonym"]/value'
        COLUMNS
        cls CHAR(64) PATH './parent::*/parent::*
                          /parent::*/name',
        tgt CHAR(64) PATH '.') AS R)
UNION ALL
     (SELECT CH.cls,CH.syn
     FROM Traversed PR,
        XMLTABLE ('Document("Thesaurus.xml")
        /terminology/conceptDef/definingConcepts/
        concept[./text()=$parent]/parent::*/parent::*/
        properties/property[name/text()="Synonym"]/value'
        PASSING PR.cls AS "parent"
        COLUMNS
        cls CHAR(64) PATH './parent::*/
                          parent::*/parent::*/name',
        syn CHAR(64) PATH '.') AS CH))
SELECT DISTINCT V.*
FROM Visit V
WHERE V.diagnosis IN
   (SELECT DISTINCT syn FROM Traversed)
```

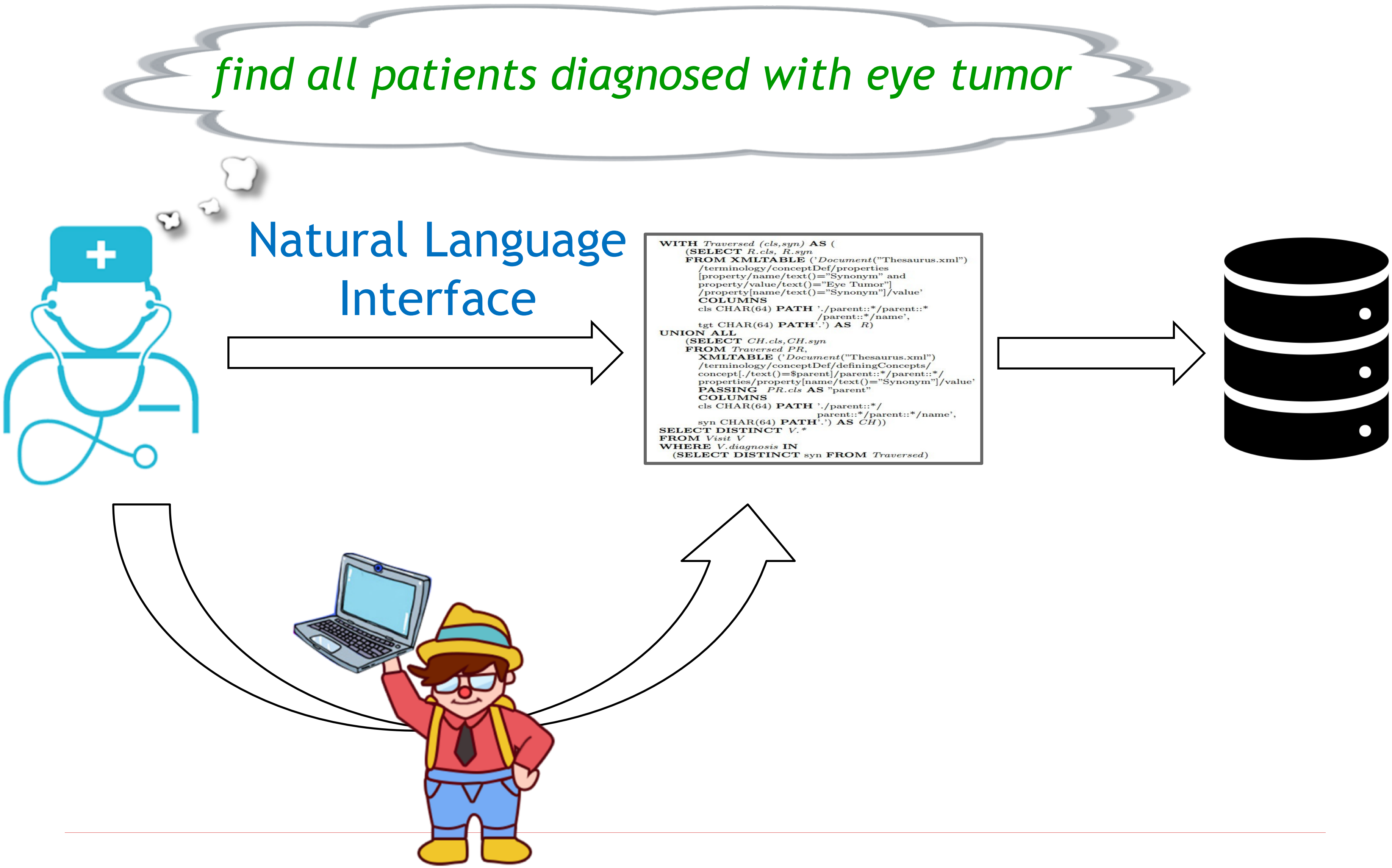
Natural Language Interface

"Semantic queries by example",
Lim et al., EDBT (2014)

In Pursue of Efficiency



In Pursue of Efficiency



Natural Language Interface ≈ Model-Theoretic Semantics

Language Variations

Utterance

find the first child of Queen Elizabeth II

semantic parsing

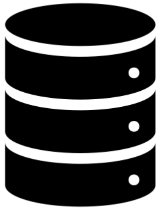
Symbol Grounding

Formal Meaning Repr.
(SQL, λ-calculus, …)

argmin(child(Elizabeth), birthdate)

execution

World
(knowledge base, …)



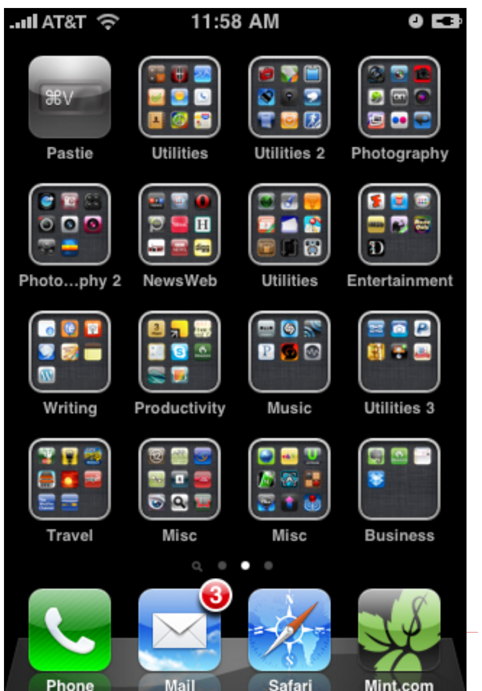
Denotation

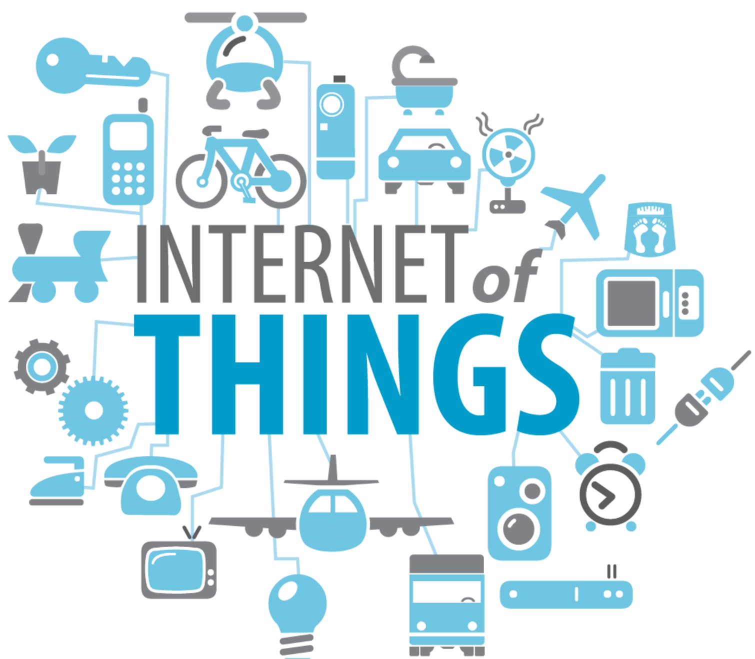
Charles, Prince of Wales

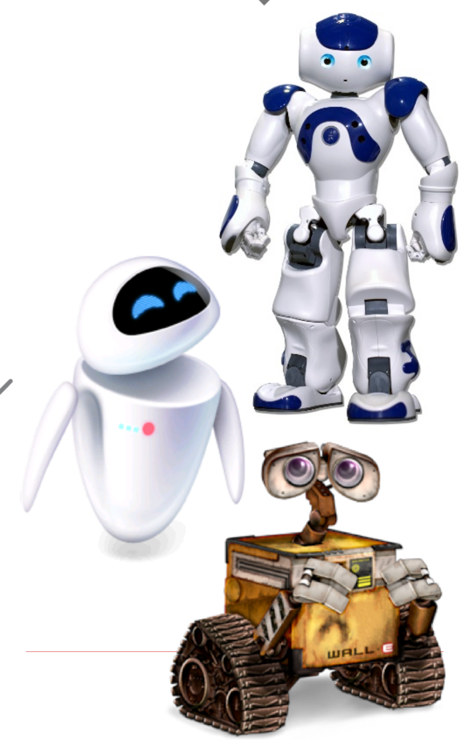

COLD START

The Cold Start Problem

"I want to build an NLI for my domain, but I don't yet have any user or data"





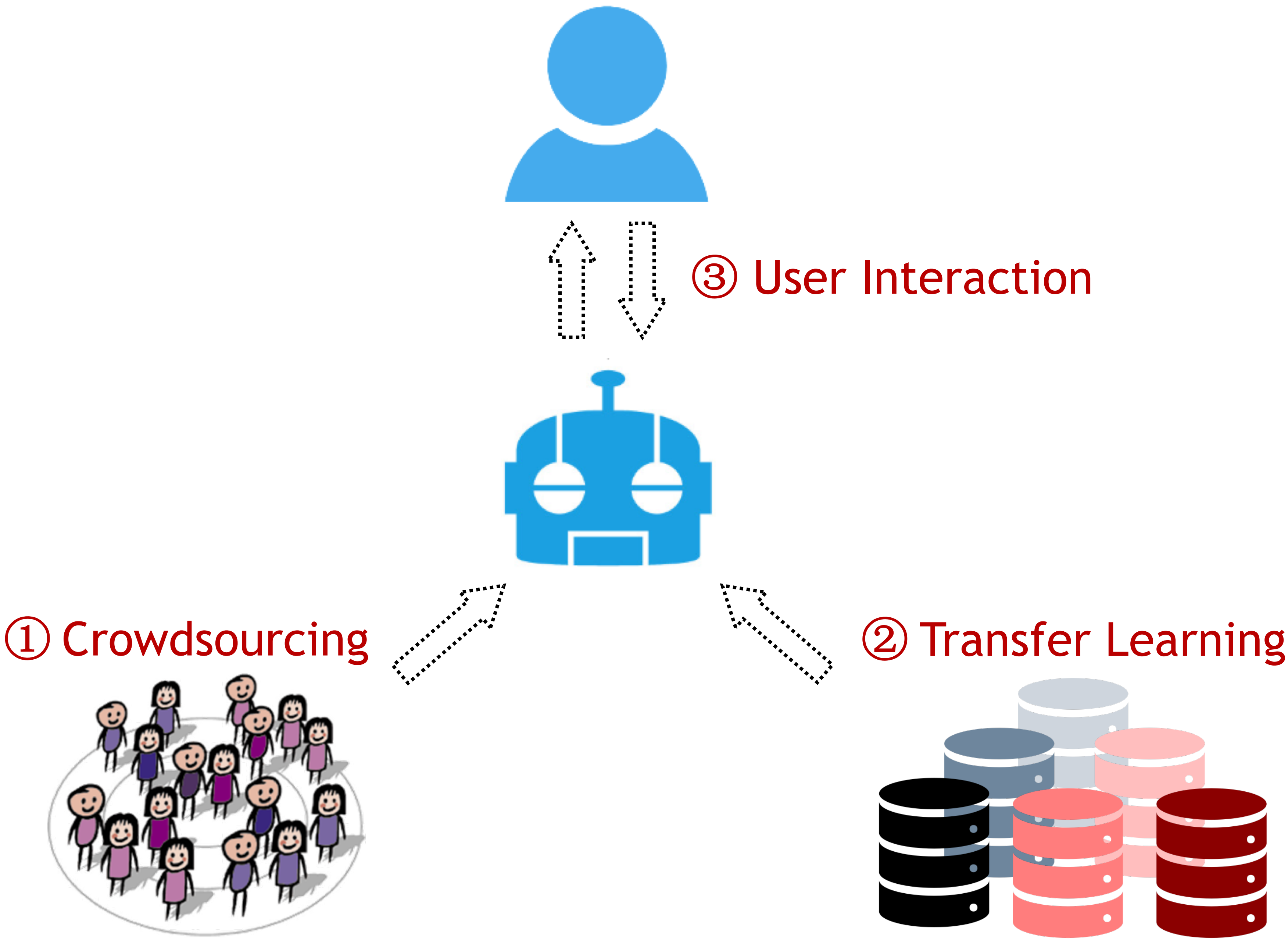


How to Build NLI for New Domain

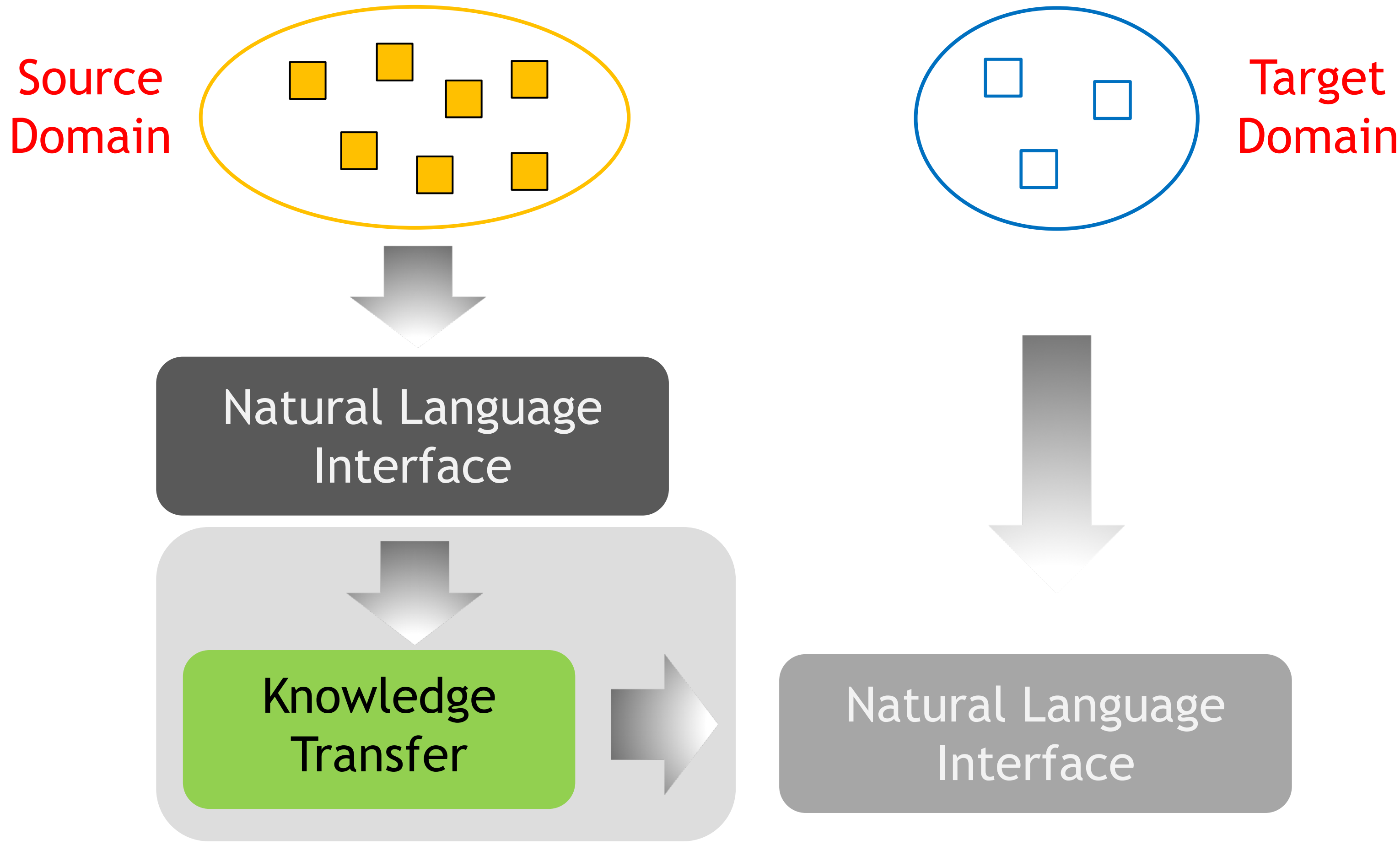
- 1950s-1990s: Rule engineering (rule-based systems)
- 1990s-2010s: Feature engineering (statistical ML)
- 2010s-present: Data engineering (neural models)

```
editor> add verb
what is your verb ? exceed
what is its third sing. pres ? exceeds
what is its past form ? exceeded
what is its perfect form ? exceeded
what is its participle form ? exceeding
to what set does the subject belong ? numeric
is there a direct object ? yes
to what set does it belong ? numeric
is there an indirect object ? no
is it linked to a complement ? no
what is its predicate ? greater_than
do you really wish to add this verb? y
```

[Auxerre and Inder, 1986]


③ User Interaction
① Crowdsourcing
② Transfer Learning

Cross-domain Natural Language Interface



Out-of-domain, on-task supervision

What is Transferrable in NLI across Domains?

Source Domain: Basketball

In which season did Kobe Bryant play for the Lakers?

$\mathbf{R}[\text{season}].(\text{player.KobeBryant} \sqcap \text{team.Lakers})$

$p(\text{relation1} \mid \text{"play for"})$

$p(\text{team} \mid \text{"play for"})$

$p(\text{relation2} \mid \text{"work for"})$

$p(\text{employer} \mid \text{"work for"})$

Target Domain: Social

When did Alice start working for Mckinsey?

$\mathbf{R}[\text{start}].(\text{employee.Alice} \sqcap \text{employer.Mckinsey})$

Cross-domain NLI via Paraphrasing

$\mathbf{R}[\text{season}].(\text{player.KobeBryant} \sqcap \text{team.Lakers})$

automatic

In which season did Kobe Bryant play for the Lakers?

Season of Player Kobe Bryant whose team is Lakers

$p(\text{"}\textit{whose team is}\text{"} \mid \text{"}\textit{play for}\text{"})$

play ≈ *work*, *team* ≈ *employer*

$p(\text{"}\textit{whose employer is}\text{"} \mid \text{"}\textit{work for}\text{"})$

When did Alice start working for Mckinsey?

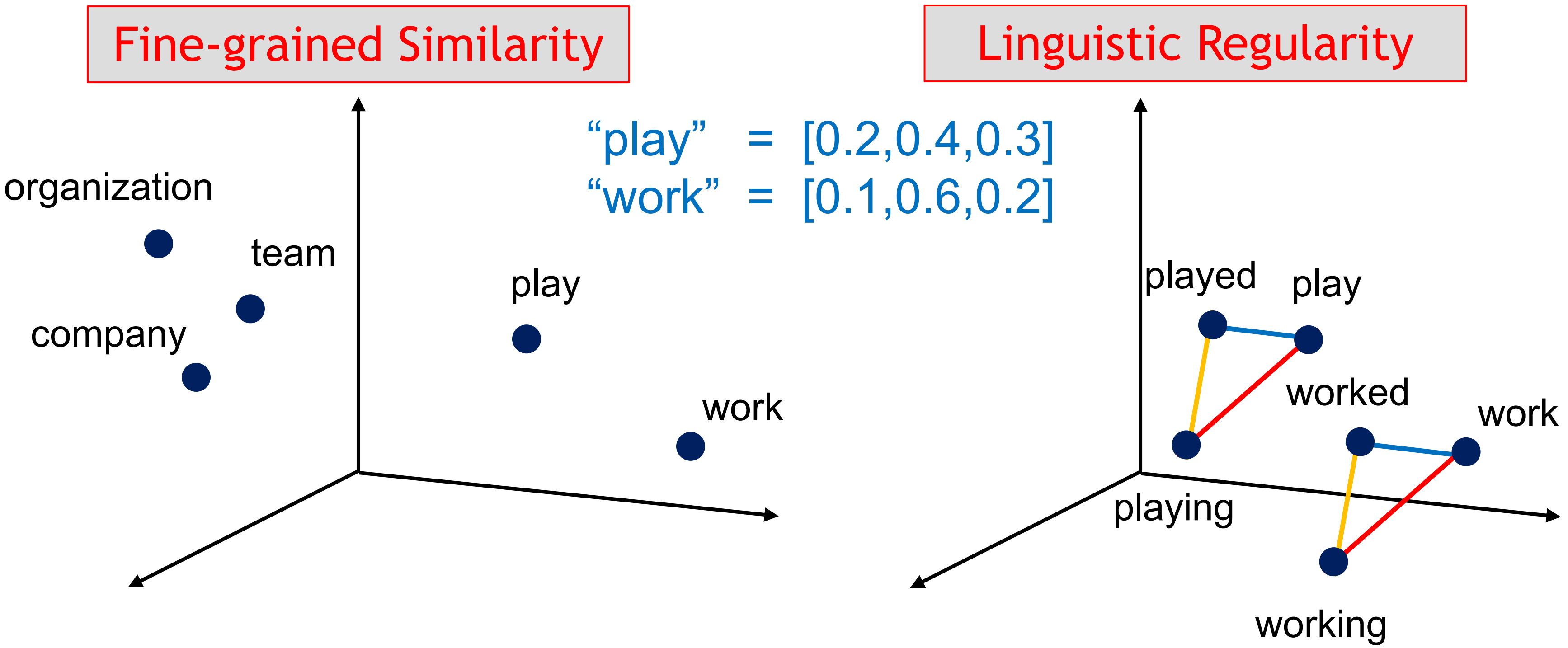
Start date of employee Alice whose employer is Mckinsey

automatic

$\mathbf{R}[\text{start}].(\text{employee.Alice} \sqcap \text{employer.Mckinsey})$

Pre-trained Word Embedding

- Word ≜ Dense vector (typically 50-1000 dimensional)
- Word similarity ≜ Vector similarity
- Pre-trained on large external text corpora



Out-of-domain, off-task supervision

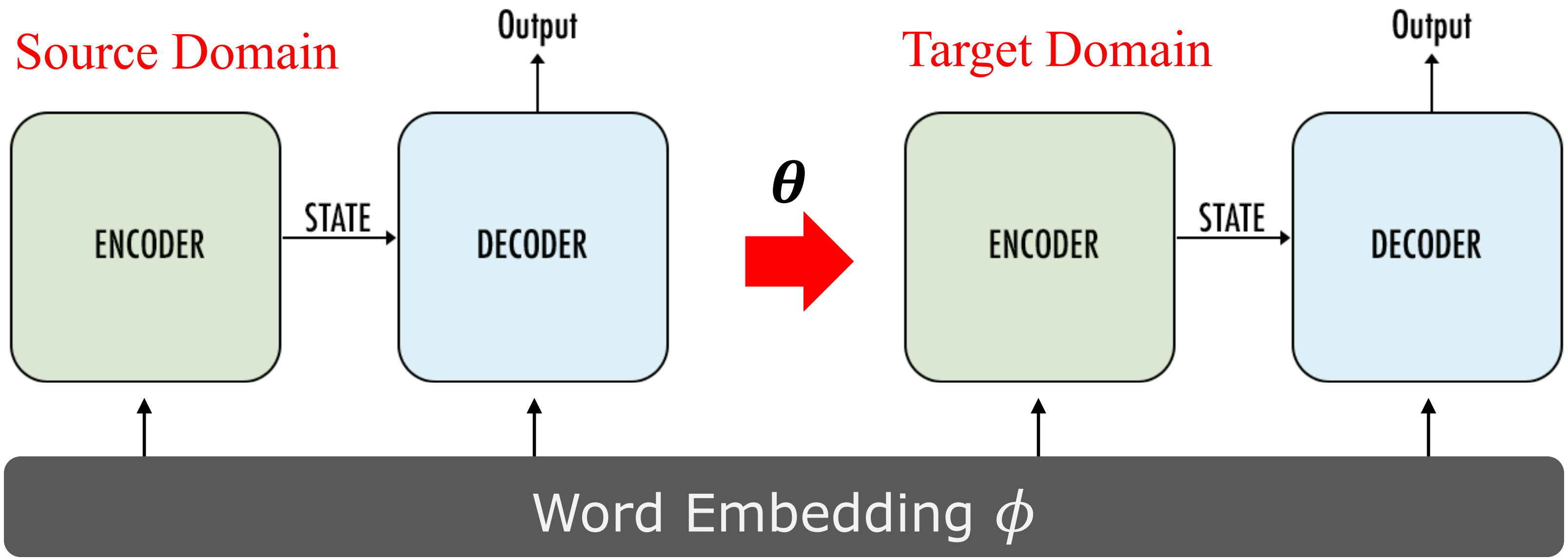
Pre-trained Word Embedding Alleviates Vocabulary Shifting

- Vocabulary shifting: Only 45%~70% target domain vocabulary are covered by source domains[1]
- Pre-trained word embedding can alleviate the vocabulary shifting problem
  - Word2vec: 300-d vectors pre-trained on the 100B-token Google News Corpus; vocabulary size = 3M

	Calendar	Housing	Restaurants	Social	Publications	Recipes	Basketball	Blocks
Coverage	71.1	60.7	55.8	46.0	65.6	71.9	45.6	61.7
+word2vec	**93.9**	**90.9**	**90.4**	**89.3**	**95.6**	**97.3**	**89.4**	**93.8**

[1] Wang et al. Building a Semantic Parser Overnight. 2015

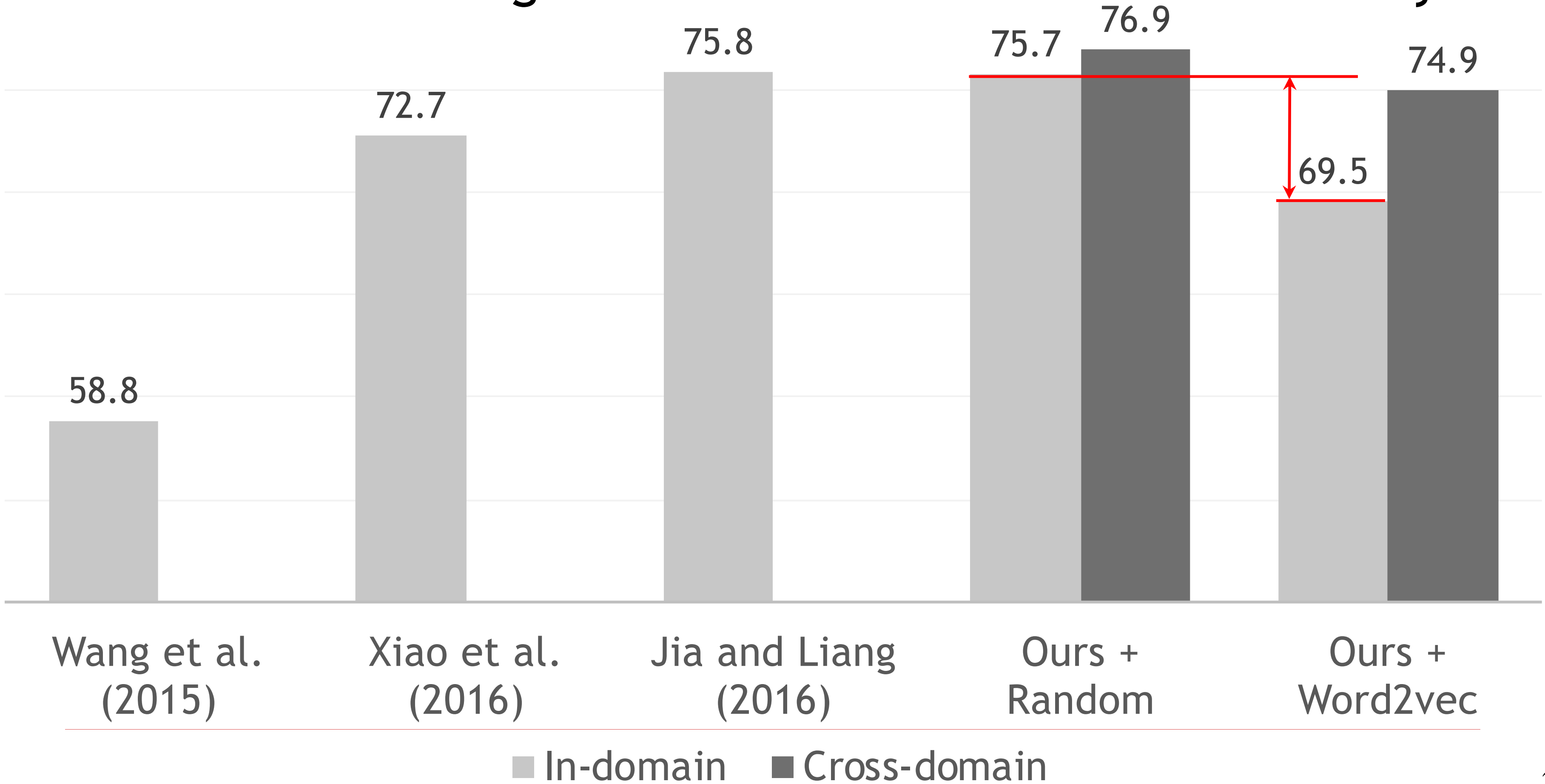
Neural Transfer Learning for NLI



- Input utterance $\boldsymbol{x} = (x_1, \ldots, x_m)$, canonical utterance $\boldsymbol{y} = (y_1, \ldots, y_n)$
- Embedding: $\phi(\boldsymbol{x}) = (\phi(x_1), \ldots, \phi(x_m))$, $\phi(\boldsymbol{y}) = (\phi(y_1), \ldots, \phi(y_n))$
- Learning on source domain: $p(\phi(\boldsymbol{y})|\phi(\boldsymbol{x}), \boldsymbol{\theta})$
- Warm start on target domain: $p(\phi(\boldsymbol{y})|\phi(\boldsymbol{x}), \boldsymbol{\theta})$
- Fine-tuning on target domain: $p(\phi(\boldsymbol{y})|\phi(\boldsymbol{x}), \boldsymbol{\theta}^*)$

Direct Use of Word2vec Fails Dramatically…

- Cross-domain: for each target domain, use all others as source domain
- Word2vec brings 6.2% absolute decrease in accuracy



Pre-trained Word Embedding: What May Be Wrong?

- Small *micro variance*: hurt optimization
  - Activation variances ≈ input variances [Glorot & Bengio, 2010]
  - Small input variance implies poor exploration in parameter space
- Large *macro variance*: hurt generalization
  - Distribution discrepancy between training and testing

$$\overbrace{\begin{bmatrix} | & | & \cdots & | \\ \boldsymbol{w}_1 & \boldsymbol{w}_2 & \cdots & \boldsymbol{w}_n \\ | & | & \cdots & | \end{bmatrix}}^{n=|V|} \left.\vphantom{\begin{bmatrix} | \\ w \\ | \end{bmatrix}}\right\} m$$

Initialization	L2 norm	Variance	Cosine Sim.
Random	17.3 ± 0.45	1.00 ± 0.05	0.00 ± 0.06
WORD2VEC	2.04 ± 1.08	0.02 ± 0.02	0.13 ± 0.11

Micro Variance
Variance of the values comprising a vector

Macro Variance
$$\frac{\sum_{i=1}^{n} var(\boldsymbol{w}_i)}{n}$$
Variance among different vectors

Proposed Solution: Standardization

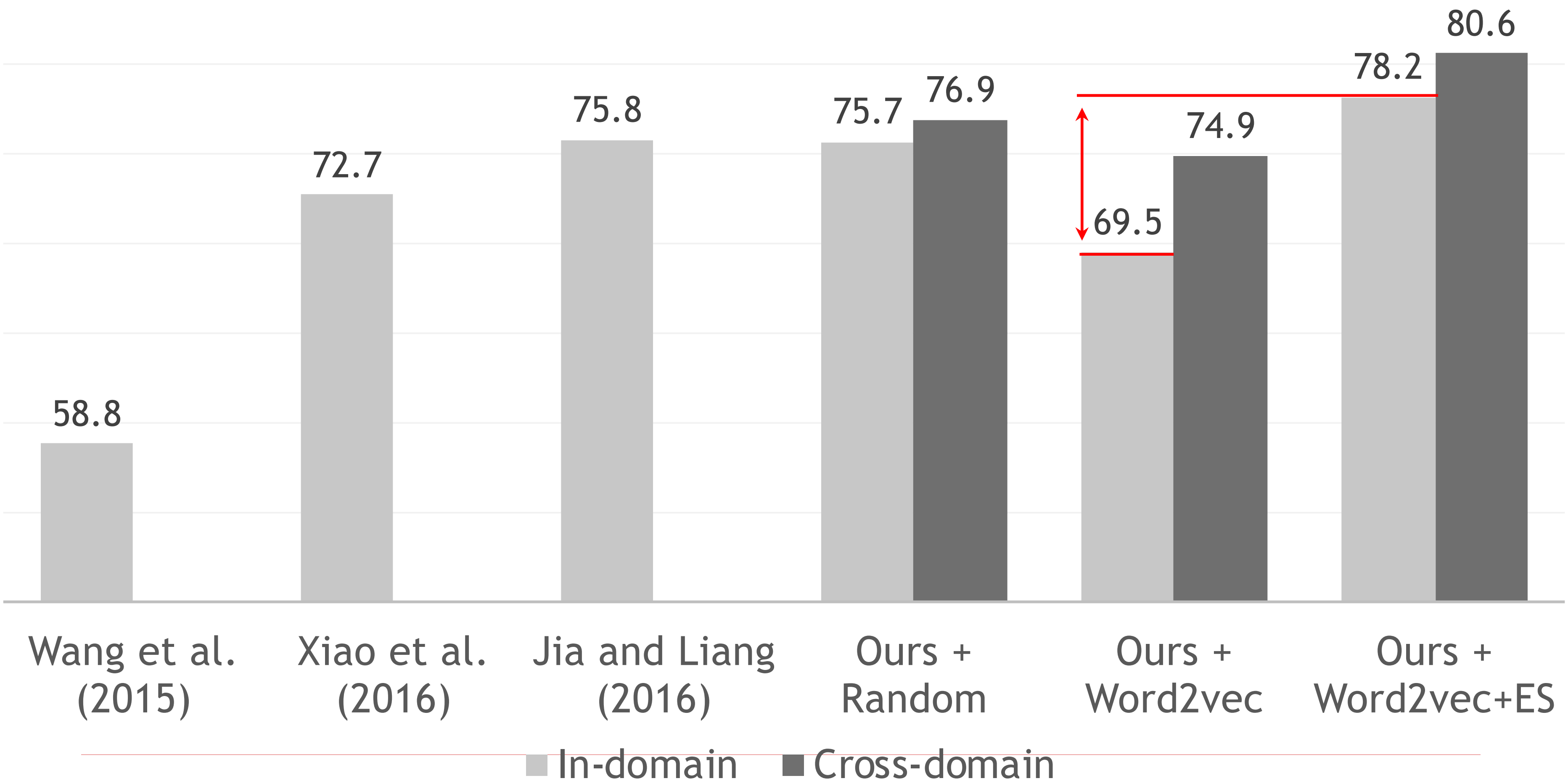
- Standardize each word vector to unit variance
- But it was unclear before why standardization should be applied on pre-trained word embedding

Initialization	L2 norm	Variance	Cosine Sim.
Random	17.3 ± 0.45	1.00 ± 0.05	0.00 ± 0.06
WORD2VEC	2.04 ± 1.08	0.02 ± 0.02	0.13 ± 0.11
WORD2VEC + ES	17.3 ± 0.05	1.00 ± 0.00	0.13 ± 0.11

Random: randomly draw from uniform distribution with unit variance
Word2vec: pre-trained word2vec embedding
ES: per-example standardization (per column)

Standardization Fixes the Variance Problems

- Standardization brings 8.7% absolute increase
- Transfer learning brings another 2.4% increase



Let machines understand human thinking

Don’t let humans think like machines

WHAT’S NEXT?

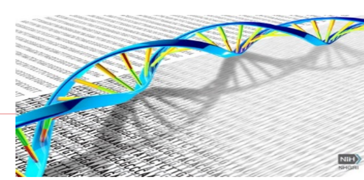
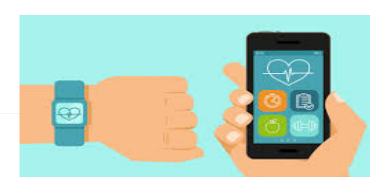
Bridging the Gap between Human and Data: AI-Powered Knowledge Engine

Discoveries
Decisions
Actions

Natural Language Interface

Knowledge-based Reasoning

Knowledge Harvesting

Natural Language Interface for Data Analytics

Study:
use nighttime luminosity observed by satellites as a proxy measure of development and welfare

Command (high-level):
calculate the average nighttime luminosity near roads in China in 1994

Command (implementation):

```
# Get China less gas flares polygon
arcpy.Select_analysis("countries_nogas", "china1.shp",
   "\"NAME\" = 'China'")
# Average two satellites for 1994
outRaster = (Float("F101994")+Float("F121994"))/2
outRaster.save("FXX1994")

# Use buffer tool and roads to make polygon of China
# close to roads, then clip china1 to this
arcpy.Buffer_analysis("a2010_final_proj", "roadbuff.shp", "0.5
   DecimalDegrees", "FULL", "ROUND", "ALL", "")
arcpy.Clip_analysis("H:/Research/Data/Lights/china1.shp",
   "H:/Research/Data/Lights/roadbuff.shp", "china2.shp", "")

# Clip each lights raster to extent of china2
rasterList = arcpy.ListRasters("F*")
for raster in rasterList:
   arcpy.Clip_management(raster, "-179.9999 -90.0 180.0
     83.62741", "G"+str(raster[1:]),
     "H:/Research/Data/Lights/china2.shp", "",
     "ClippingGeometry")

# Create grid to extent of one of new light rasters
arcpy.CreateFishnet_management("ch_grid.shp", "73.55416
   18.15416", "73.5541 28.15416","0.1", "0.1", "0", "0",
   "134.77916 53.5625", "NO_LABELS", "G101992", "POLYGON")
arcpy.RasterToPolygon_conversion("G101992", "G101992p.shp",
   "NO_SIMPLIFY", "Value")

# Process: Clip grid to perimeter of polygon
arcpy.Clip_analysis("H:/Research/Data/Lights/ch_grid.shp",
   "H:/Research/Data/Lights/G101992p.shp", "china_grid.shp",
   "")

# Zonal statistics on each year
rasterList = arcpy.ListRasters("G*")
for raster in rasterList:
   arcpy.gp.ZonalStatisticsAsTable_sa("H:/Research/Data/Lights/
     china_grid.shp", "FID", raster,
     "l"+str(raster[5:])+".dbf", "DATA", "MEAN")
```

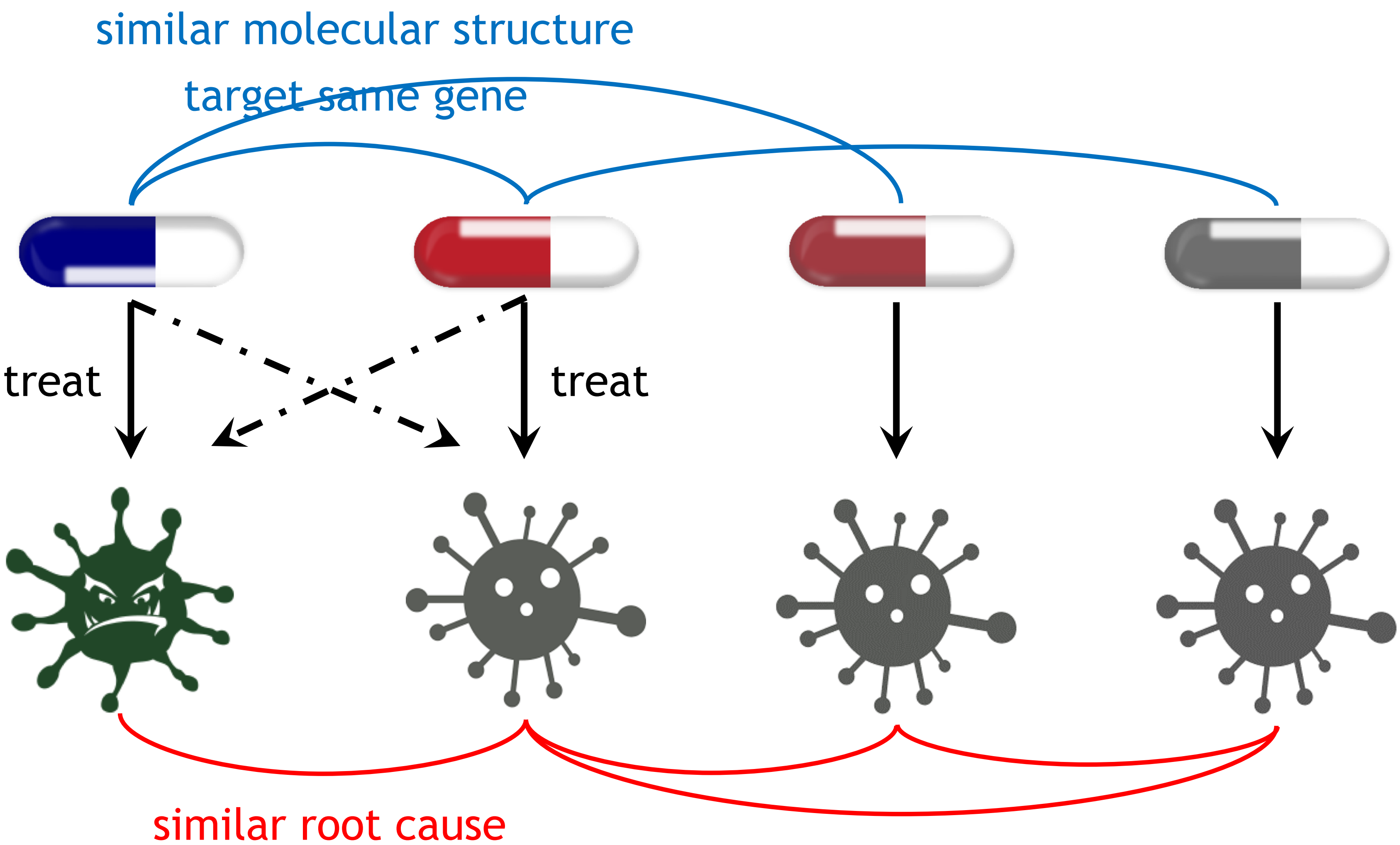
Natural Language Interface for Data Analytics

- Transduce natural language commands into programs

- Allow users to stay focused on high-level thinking and decision making, instead of overwhelmed by low-level implementation details

- Two steps
  - Simple commands → single function calls [CIKM'17], [SIGIR'18]
  - Complex commands → programs of multiple function calls

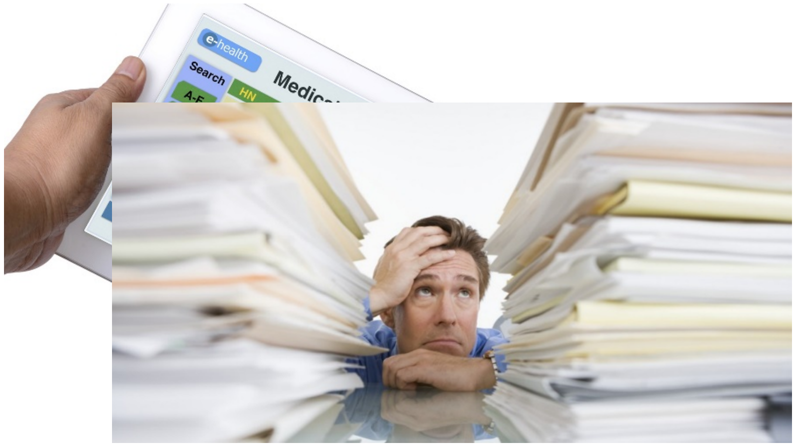
Knowledge-based Machine Reasoning



Methodological Exploration

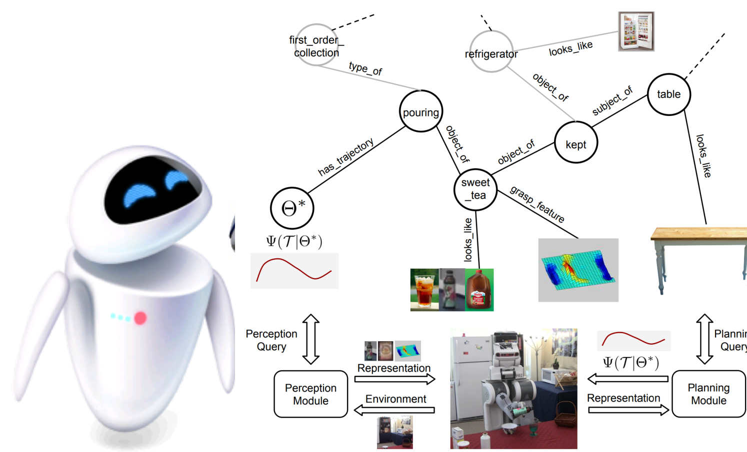
- Inherent structure of the NLI problem space
  - Strong prior for learning
  - Key: compositionality of natural & formal languages [CIKM’17]
- Integration of neural and symbolic computation
  - Neural network modularized over symbolic structures [SIGIR’18]
  - (Cognitive science) neural encoding of symbolic structures
- Goal-oriented human-computer conversation
  - Accommodate dynamic hypothesis generation and verification in a natural conversation
  - Challenge: open-ended, no fixed frames

AI-Powered Knowledge Engine: Applications



"Which cement stocks go up the most when a Category 3 hurricane hits Florida?"

KENSHO



Thanks &