# Enterprise Knowledge Graph

## From Specific Business Task to Enterprise Knowledge Management

**Rong Duan**

**Huawei Technology**

**Yanghua Xiao**

**Fudan University**

The 28th ACM International Conference on Information and Knowledge Management
Nov 3-7, 2019， Beijing, China

# Speakers/materials

Materials: https://github.com/snsxf/CIKM_Tutortial/blob/master/README.md

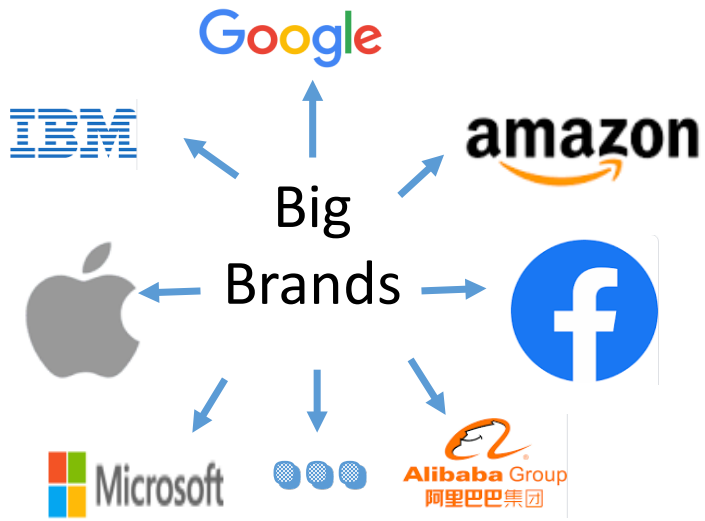
**Dr. Rong Duan**
Chief Data Scientist
Corporate Data Management Department
Huawei Technologies Co.,Ltd
Shenzhen, China

**Dr Yanghua Xiao**
full professor
School of Computer Science
Fudan University
Shanghai, China

# Copyright claim

- Some ppts about phrase mining in this tutorial used materials  in the Prof. Jiawei Han’s tutorial about phrase mining and text mining.
- Some images comes from Web.

# Motivation




**State of the art Techniques**

| | |
|---|---|
| NLP Algorithm | Storage Graph Database |
| Computing Cloud/GPU | Representation Embedding |
| End-to-End Solution/Platform | |

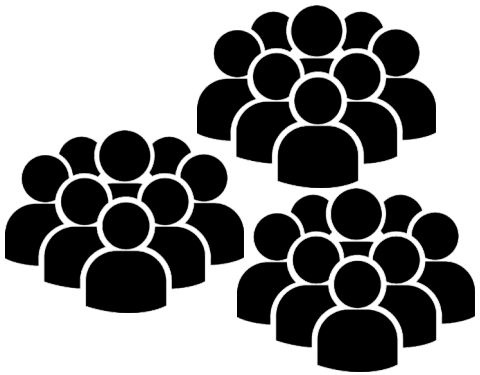
Risk Management

Drug Discovery

**Successful Use Cases**


Marketing

Data Management



# Objectives

- Understand the different types of Enterprise Knowledge Graph
- Know how to estimate the challenges before start
- Master the techniques to construct EKG

# Outline

I. **Enterprise Knowledge and Enterprise Knowledge Graph**
II. **Construction of Enterprise Knowledge Graph**
III. **Challenges and Future Research in Enterprise Knowledge Graph**

# Outline

## I. Enterprise Knowledge and Enterprise Knowledge Graph

### 1. Enterprise Knowledge

- Enterprise Knowledge Type and Acquisition
- Enterprise Knowledge Management System Construction
- Modern Enterprise Knowledge Characteristics

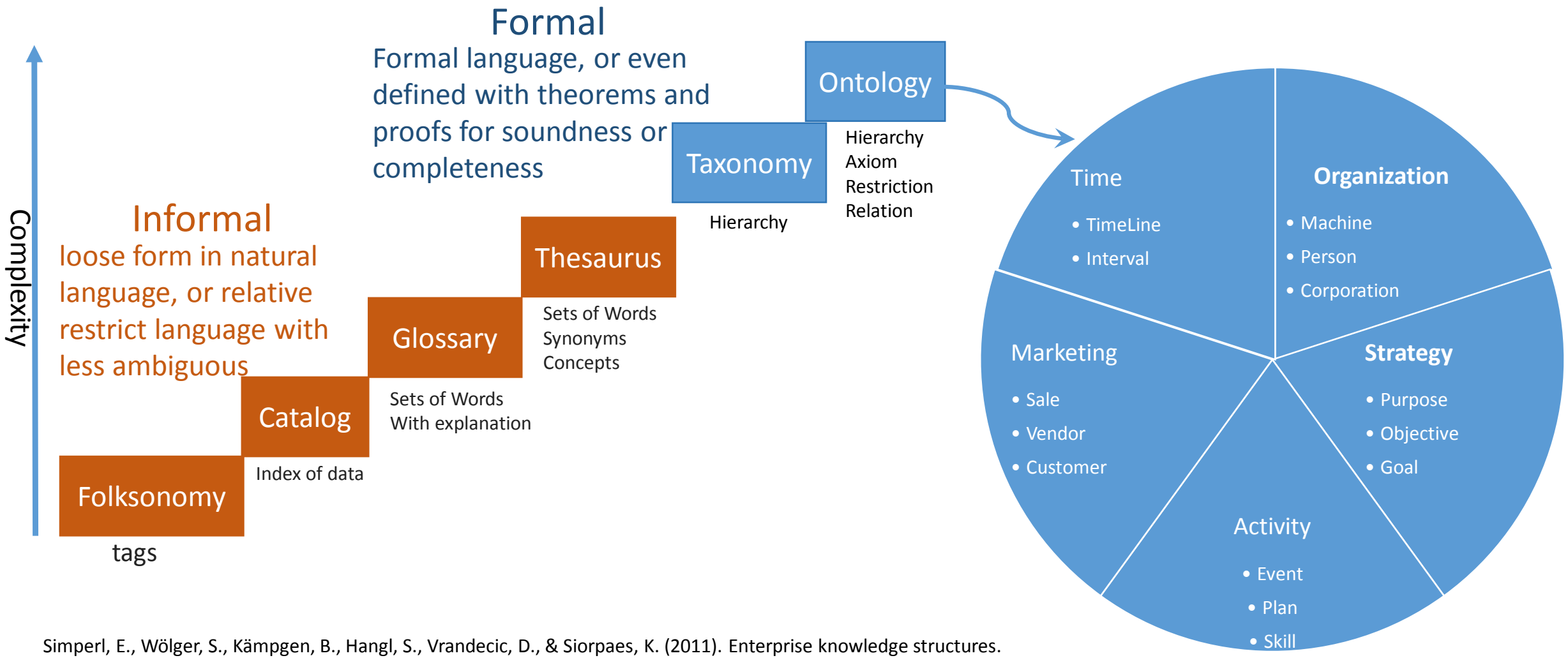
### 2. Enterprise Knowledge Graph

- Types of Knowledge Graph
- Types of Enterprise Knowledge Graph
- Characteristics of Enterprise Knowledge Graph

## II. Construction of Enterprise Knowledge Graph

## III. Challenges and Future Research in Enterprise Knowledge Graph

# Enterprise Knowledge Type



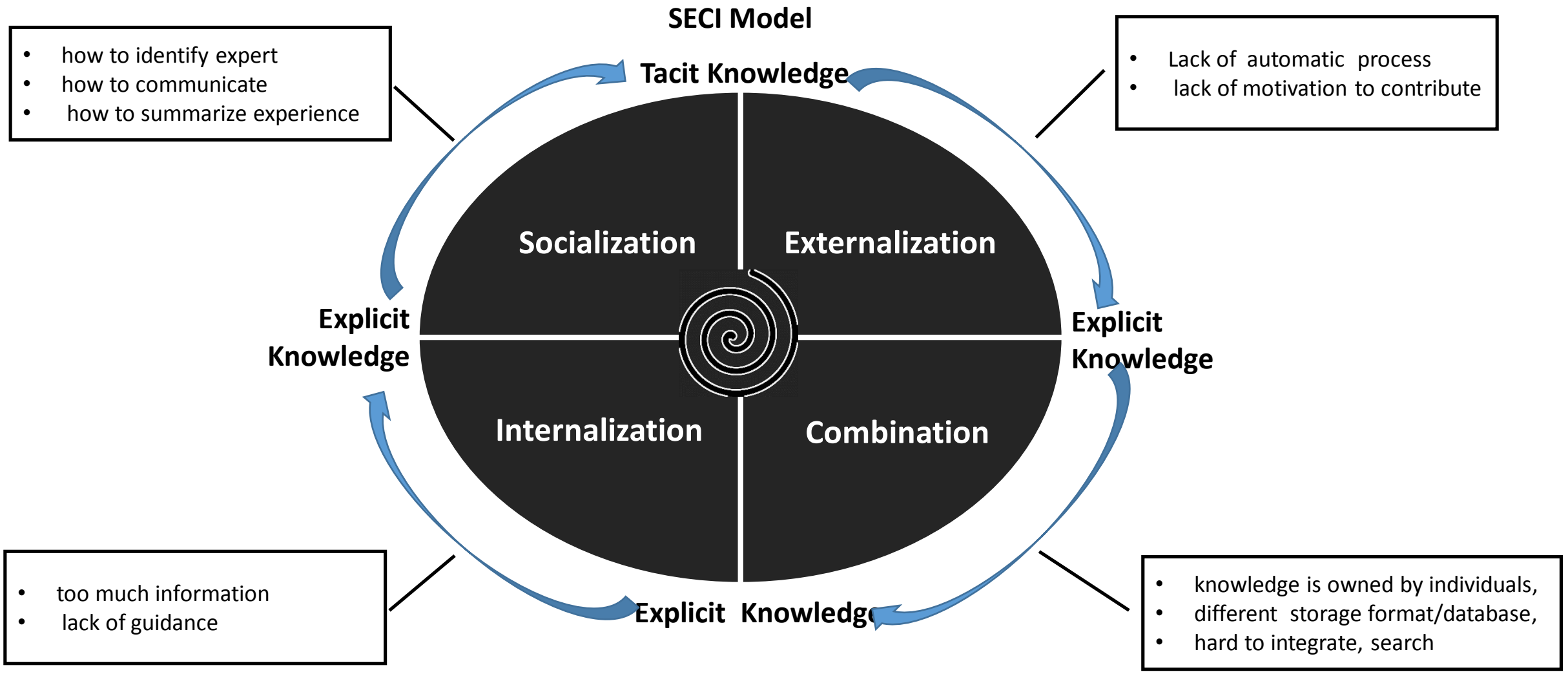


Simperl, E., Wölger, S., Kämpgen, B., Hangl, S., Vrandecic, D., & Siorpaes, K. (2011). Enterprise knowledge structures.

Uschold, M & King, M & Moralee, S & Zorgios, Y. (2018). The Enterprise Ontology" The Knowledge Engineering Review.

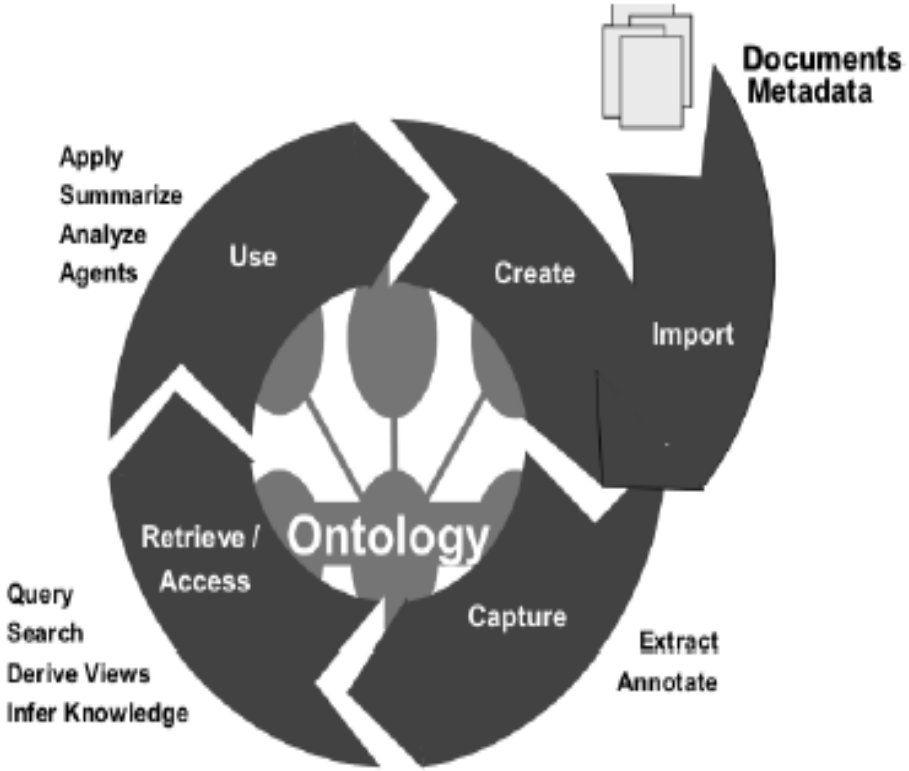
Uschold M, Gruninger M. Ontologies: Principles, methods and applications[J]. The knowledge engineering review, 1996, 11(2): 93-136.

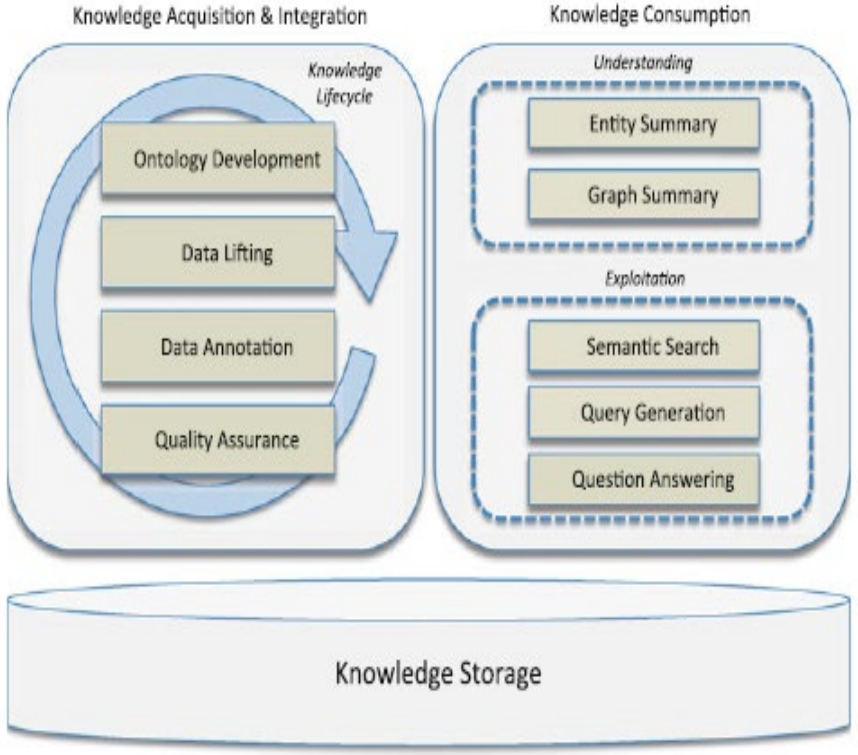
# Enterprise Knowledge Acquisition




Nonaka I, Toyama R, Konno N. SECI, Ba and leadership: a unified model of dynamic knowledge creation[J]. Long range planning, 2000, 33(1): 5-34.

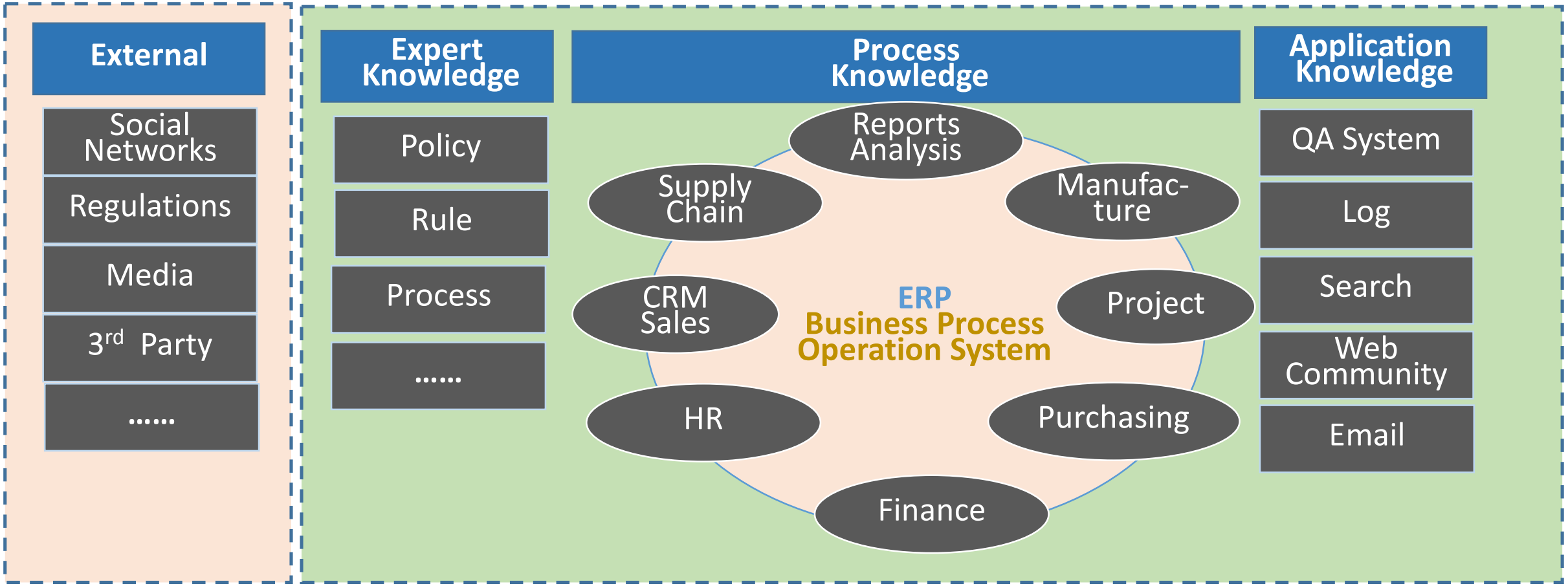
# Enterprise Knowledge Management System Construction



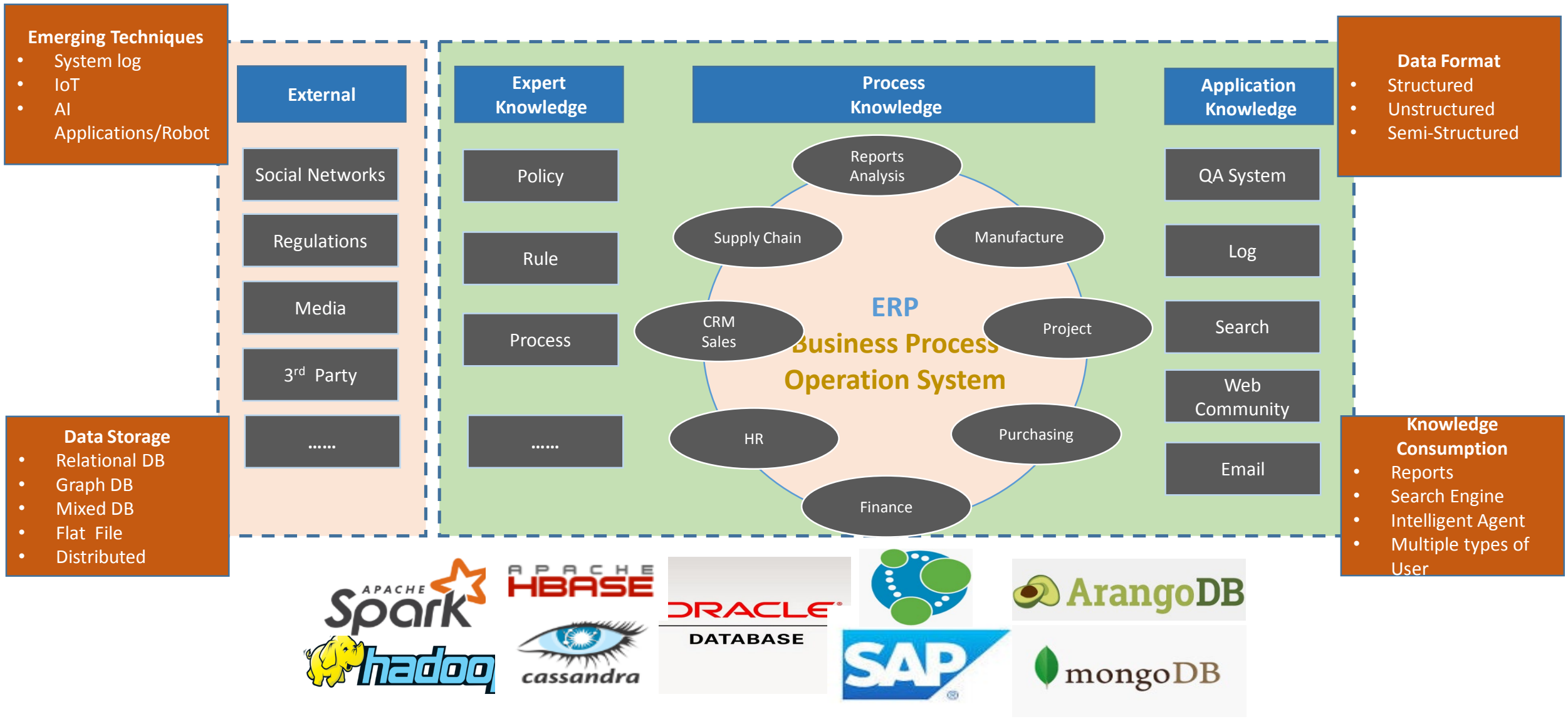




Sure, Y., Staab, S., & Studer, R. (2002). Methodology for development and employment of ontology based knowledge management applications. *Acm Sigmod Record, 31*(4), 18-23.
Jeff Z. Pan, Guido Vetere, Jose Manuel Gomez-Perez ,Honghan Wu; Exploiting Linked Data and Knowledge Graphs in Large Organizations – 2017

# Modern Enterprise Knowledge Components



# Modern Enterprise Knowledge Characteristics



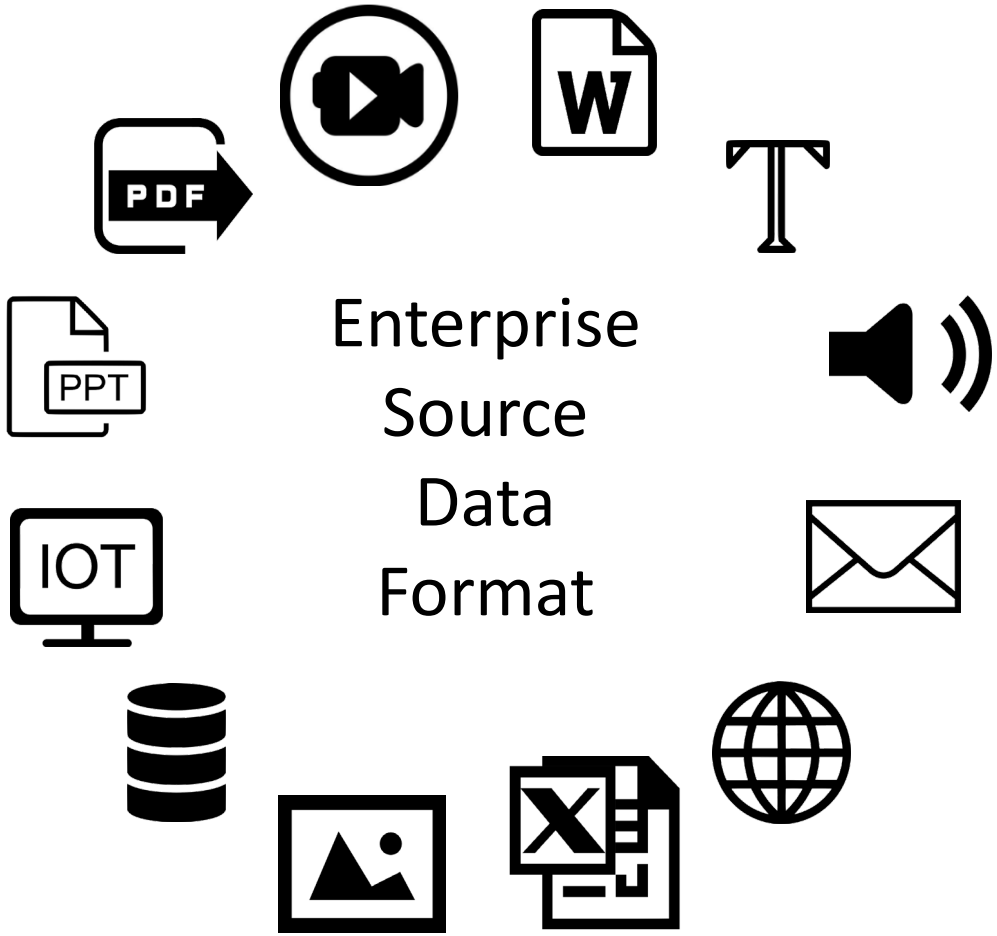
# Knowledge Acquisition:  From Relative Static to Relative Dynamic

| Pro-Active | Inter-Active | Post-Active |
| --- | --- | --- |
| • Pre-defined Rules<br>• Pre-defined Business Process<br>• Pre-defined Data Architecture<br>• Pre-defined Reports<br>• …… | • Question-Answering<br>• Crowed Sourcing<br>• Tagging<br>• …… | • Analysis<br>• Process Log<br>• Application Log |
| **Expert** | **Intelligent Agent** | **BI/ML/DM/AI** |

- **Formal  -> Informal**
- **Expert -> Crowd**
- **Time consuming -> Response quickly**
- **Certain -> Uncetrain**

# Multiple Source Data Format: Information Retrieval is a challenge

Different format carry information differently

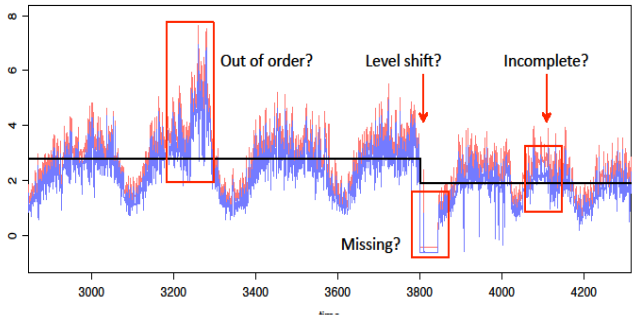



Leverage the semantic information hidden in format

# New Challenges for Information Quality : Hard Problem Even Harder

## Completeness --- IoT

- Real Time
- Without Ground Truth




## Accuracy --- Fact Checking

- Misinformation – one of the top 10 challenges as per *The World Economic Forum*
- Outdated
- External data source


## Consistency --- Crowdsourcing

- Crowd data quality
- *Crowd labor Evaluation*

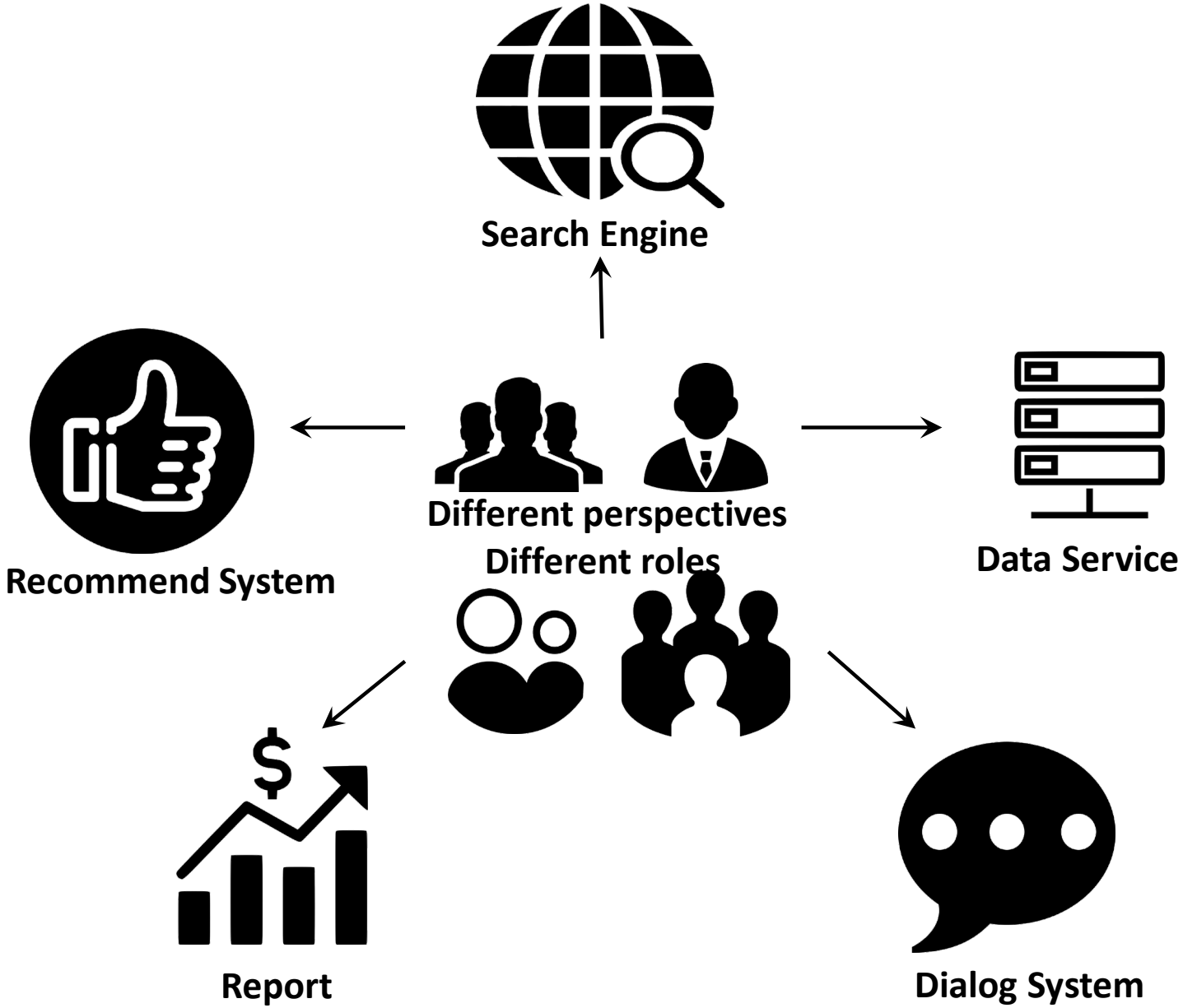

Tamraparni Dasu, Rong Duan, Divesh Srivastava "Continuous Measurement of Quality of Data Streams" Tutorial DSAA 2017

Xin Luna Dong, Christos Faloutsos,Xian Li, Subhabrata Mukherjee, Prashant Shiralkar "Fact Checking: Theory and Practice" Tutorial KDD 2018

http://www.washingtonstarnews.com/proof-obamacare-requires-all-americans-to-be-chipped/

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. "Data quality from crowdsourcing: a study of annotation selection criteria". In Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing

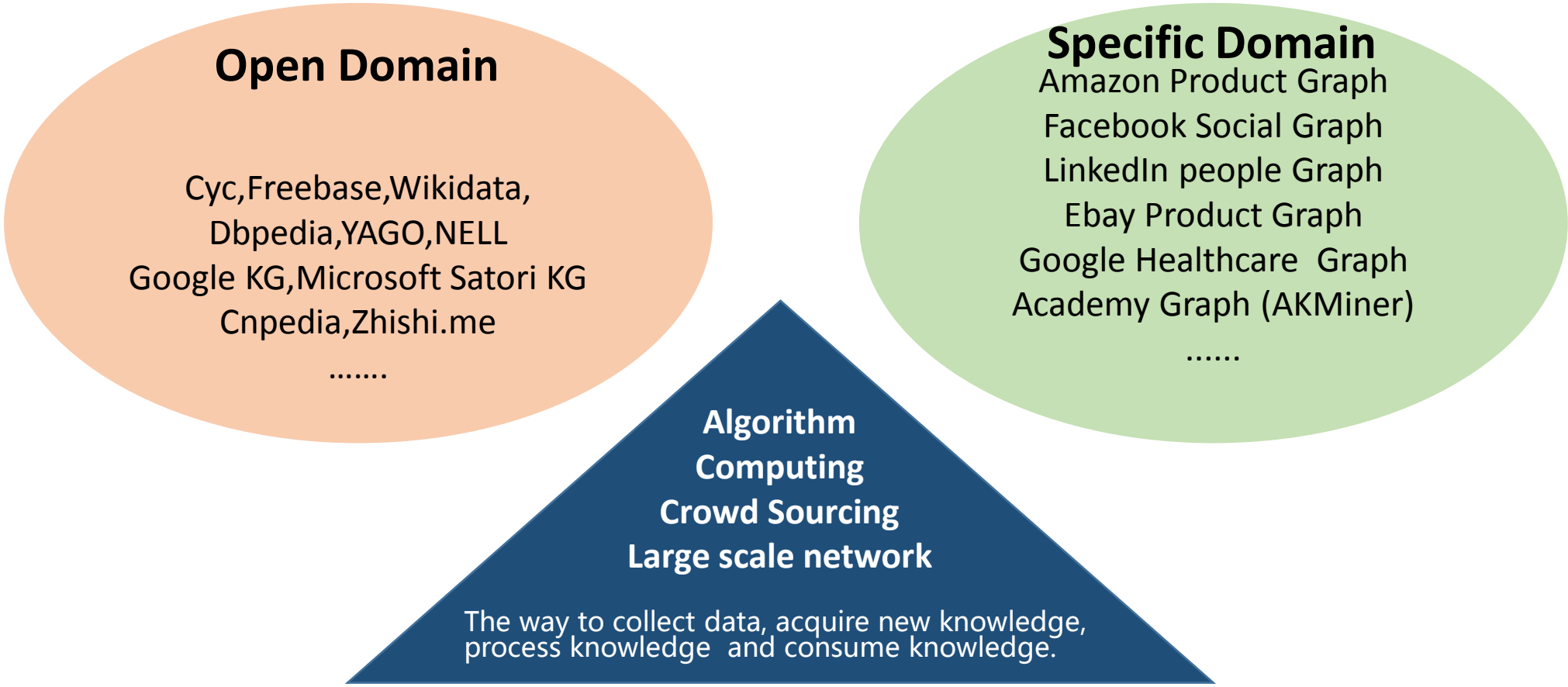
# Knowledge Consumption: Diverse Users and Different Types of Agent



# Modern Enterprise Knowledge Characteristics

1. Multiple source data format.
2. Diverse information storage systems.
3. New challenges for information quality.
4. Way to consume information.
5. Developing knowledge base.
6. Emerging techniques for knowledge acquisition.

# Types of Knowledge Graph: Current Classification



# Evolution of Knowledge Graph

## Large Scale Casual Usage

Sole Usage Type

Simple Knowledge

Simple Fact

Large scale User Data


Transportation


Food


Shopping


Search Engine

## Small Scale Complicated Usage


Complicated Usage

Deep Knowledge

Intensive Expert Knowledge

Limited Amont Data

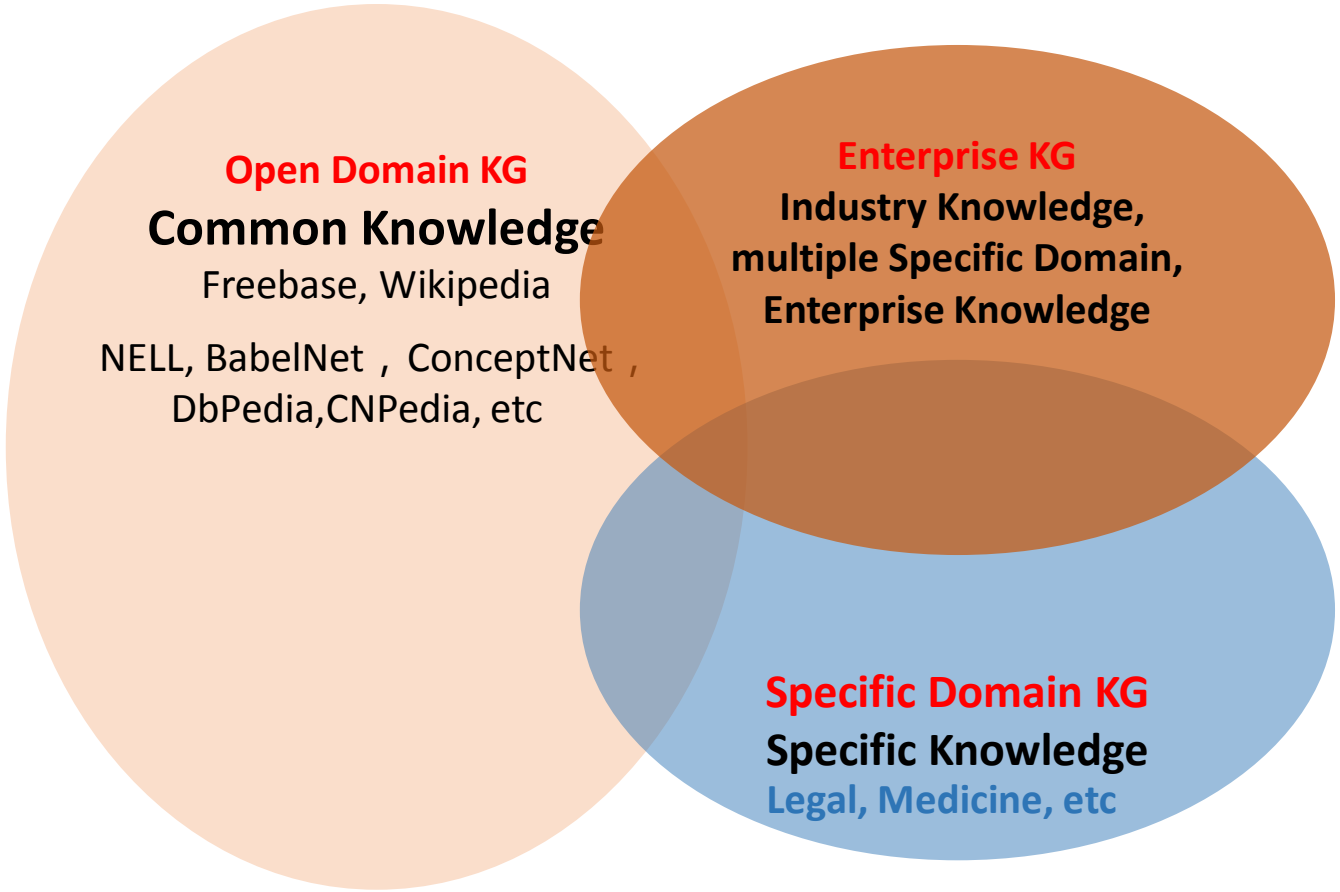
Medical

Enterprise

Legal

Finance

# Types of Knowledge Graph: Separate EKG from Others




**Knowledge Coverage of three types KG**

| | Open Domain | Specific Domain | Enterprise |
|---|---|---|---|
| Data source | Diverse | Focus | Diverse |
| Openness | Public | Public | Private |
| Data Volume | Large | Medium | Small -- Large |
| Data Quality | Low | High | Mix |
| Ontology | Simple | Relative Complicated | Complicated |
| Knowledge Type | Common | Specific Domain | Common +Specific Domain +Enterprise |
| Knowledge Width | wide | narrow | mix |
| Knowledge Depth | Shallow | Deep | Deep |
| Knowledge Consumption | simple | simple | complex |

# Types of Enterprise Knowledge Graph

# Specific Business Task Enterprise KG

**Business Trip**

**Limited Knowledge Involved**
**Expert Experience**
**Manual Construct Ontology**

**Who**

**From City**

**Transportation**

**To City**

**Hotel**


**Purpose**

Home Appliance

https://www.datanami.com/2017/09/26/graph-databases-help-smart-homes-become-connected-homes/

**How Much**

Supplier Risk

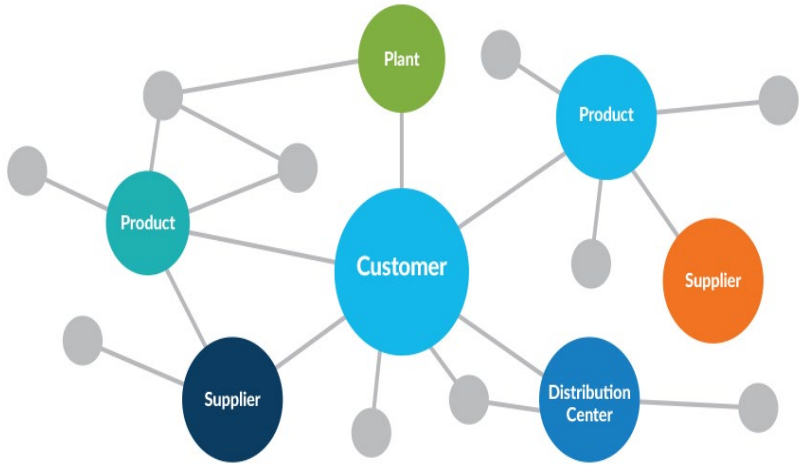
- Distance
- Criticality
- Replaceability
- Centrality

https://pdfs.semanticscholar.org/e1eb/28d6a14c204b3d92d89c4ba984761c3912a4.pdf

# Specific Business Unit Enterprise KG

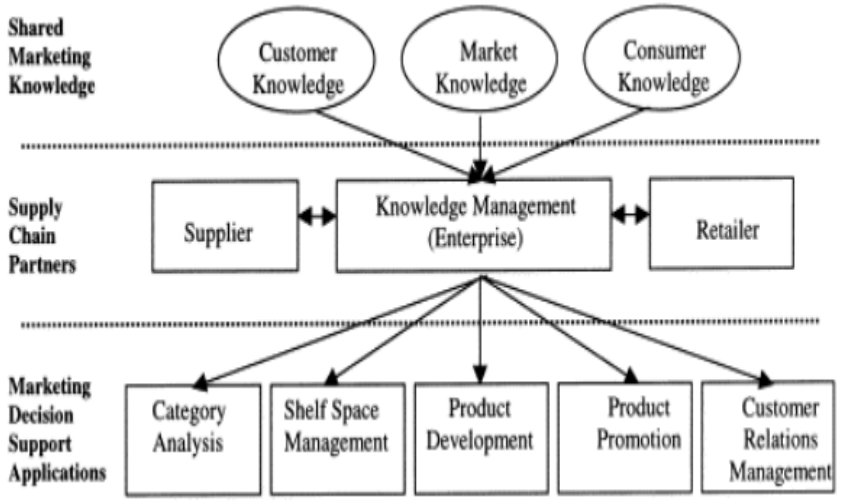
**Domain Knowledge Involved**
**Expert Experience**
**Interactive Ontology Construction**

## Supply Chain Knowledge Graph




From: FusionOps

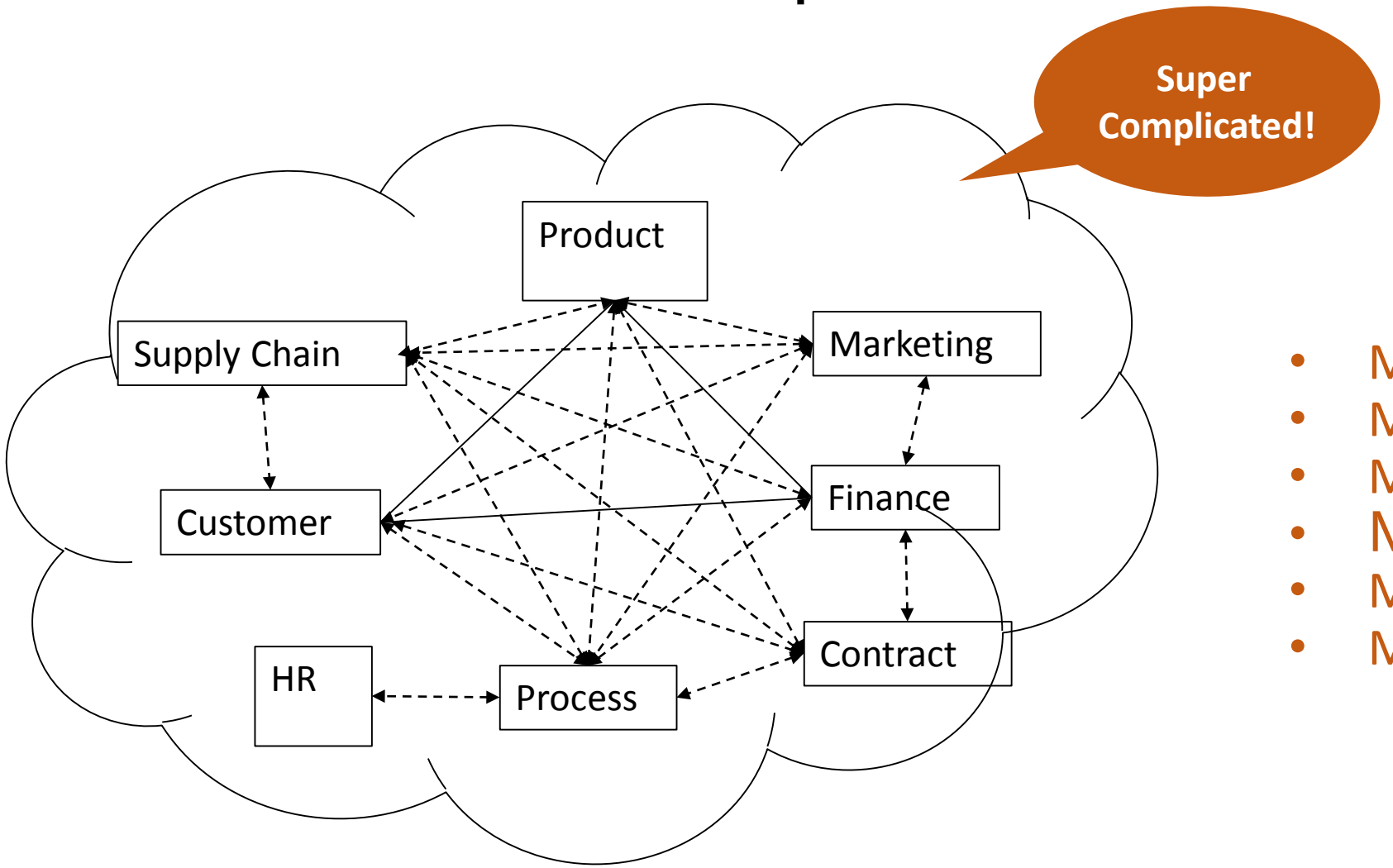
## Marketing Knowledge Graph





Sejin Chun,Jooik Jung,Xiongnan Jin,Seungmin Seo,Kyong-Ho Lee "Designing an integrated knowledge graph for smart energy services", The Journal of Supercomputing, Oct 2018
Shaw M J , Subramaniam C , Tan G W , et al. Knowledge management and data mining for marketing[J]. Decision Support Systems, 2001, 31(1):127-137.
http://fusionops.lnx.avisan.com/digital-cloud/

# Cross Business Unit Enterprise KG

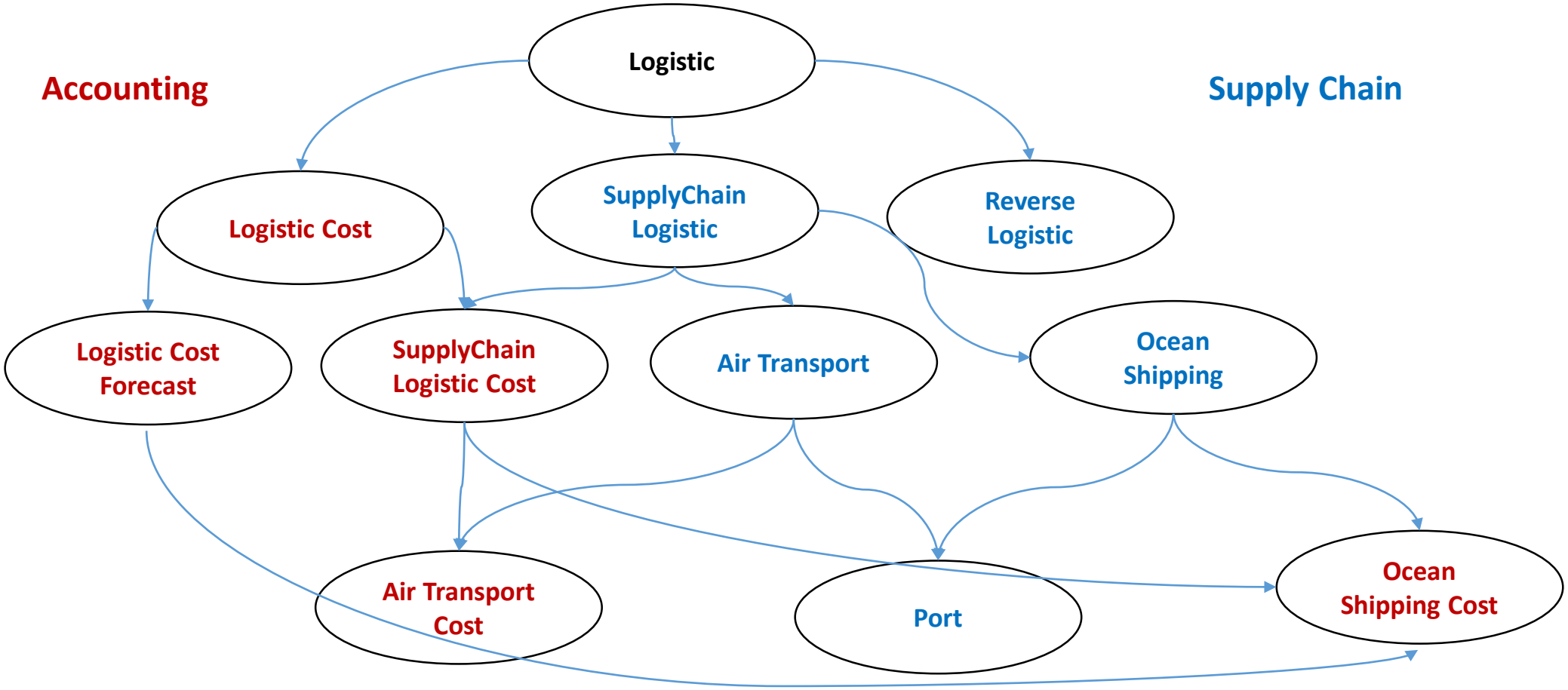



**M⁶ Characteristics**

- Multi-domain
- Multi-resource
- Multi-perspective
- Multi-agent
- Multi-resolution
- Multi-cycle

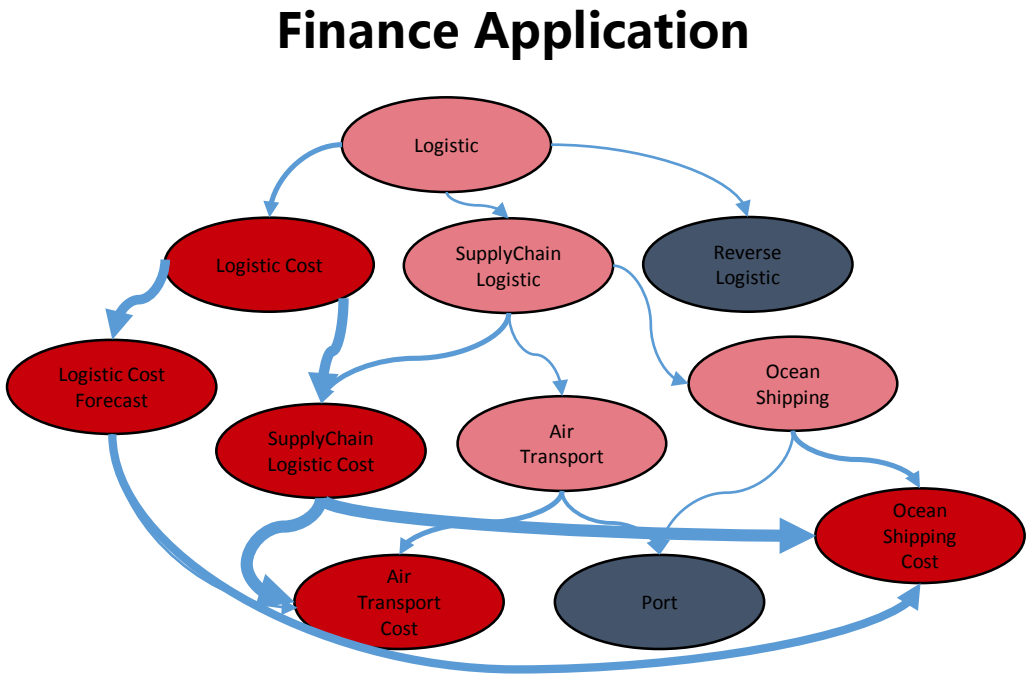
# Cross Business Unit Enterprise KG – Multi-Perspective

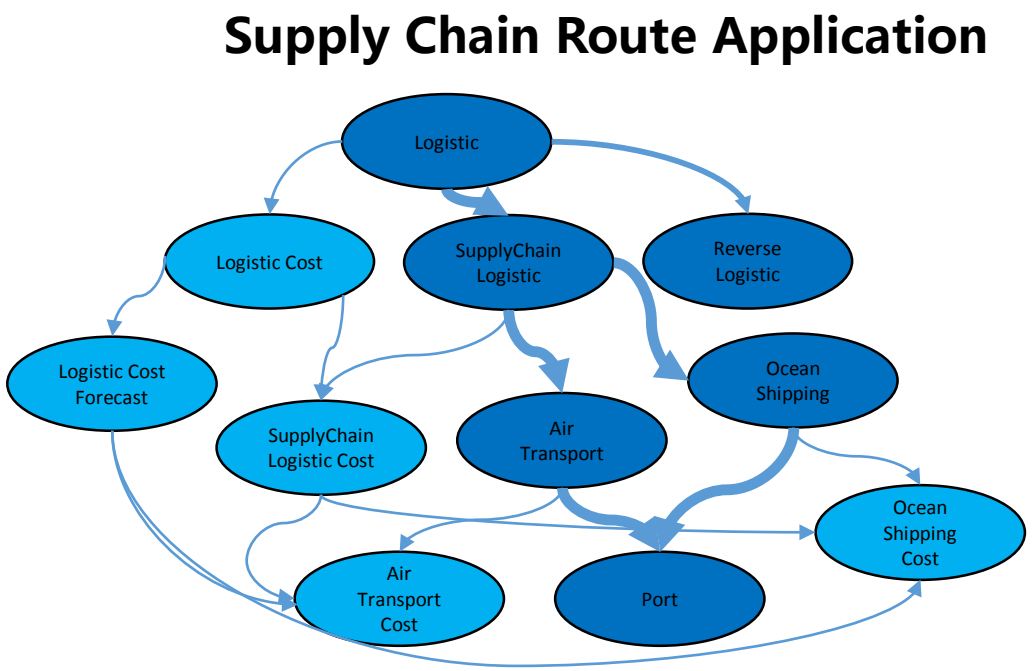
Multiple roles with multiple perspectives



# Cross Business Unit Enterprise KG: Multi-Agent

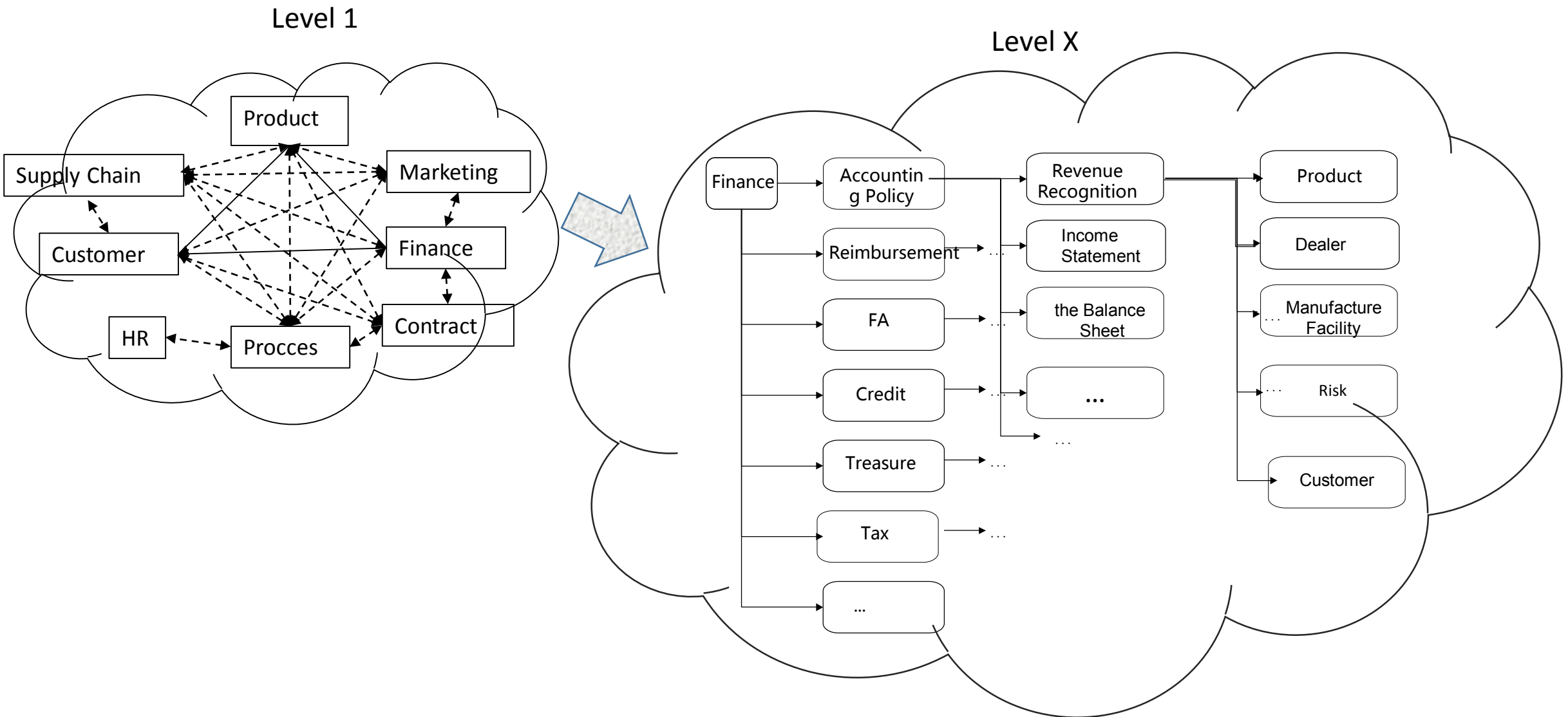
Entity weight and Entity-Entity relation weight is driven by application. Each application save its weight graph individually



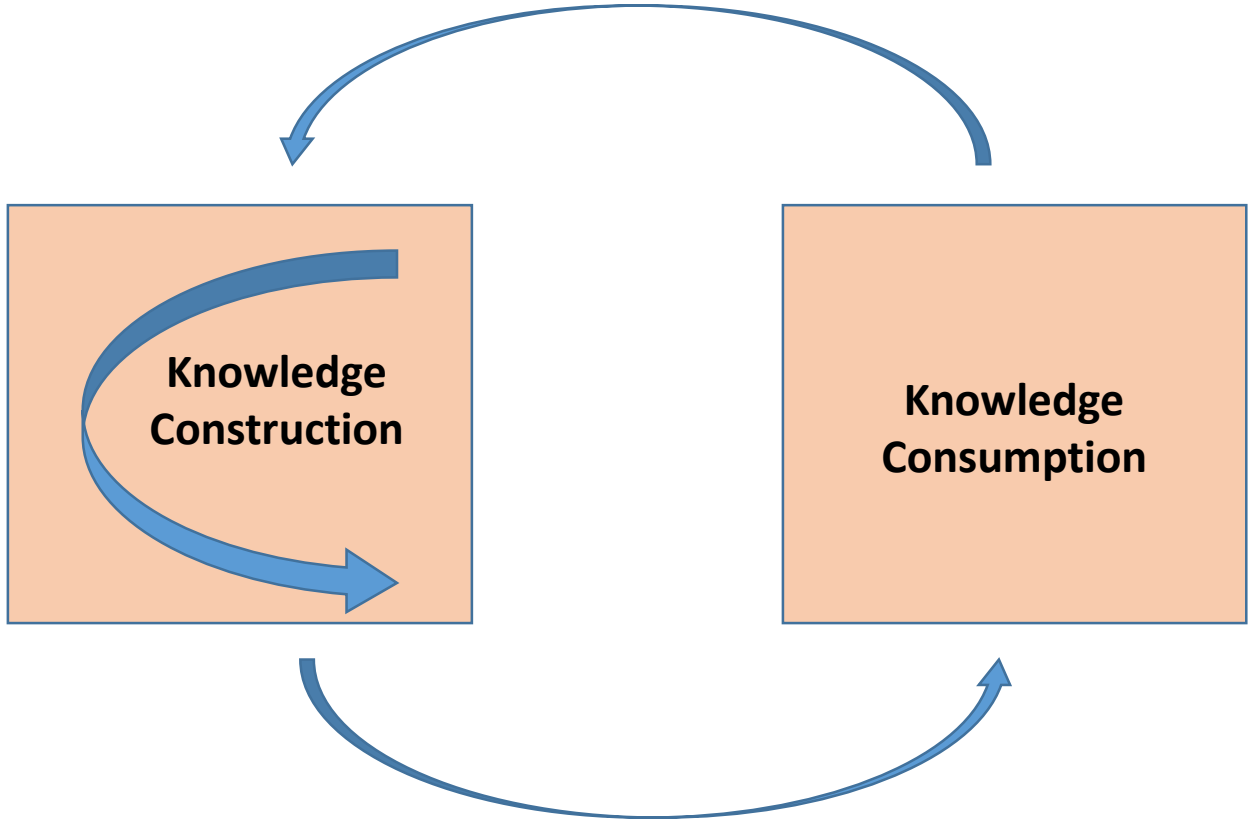



# Cross Business Unit Enterprise KG – MultiResolution

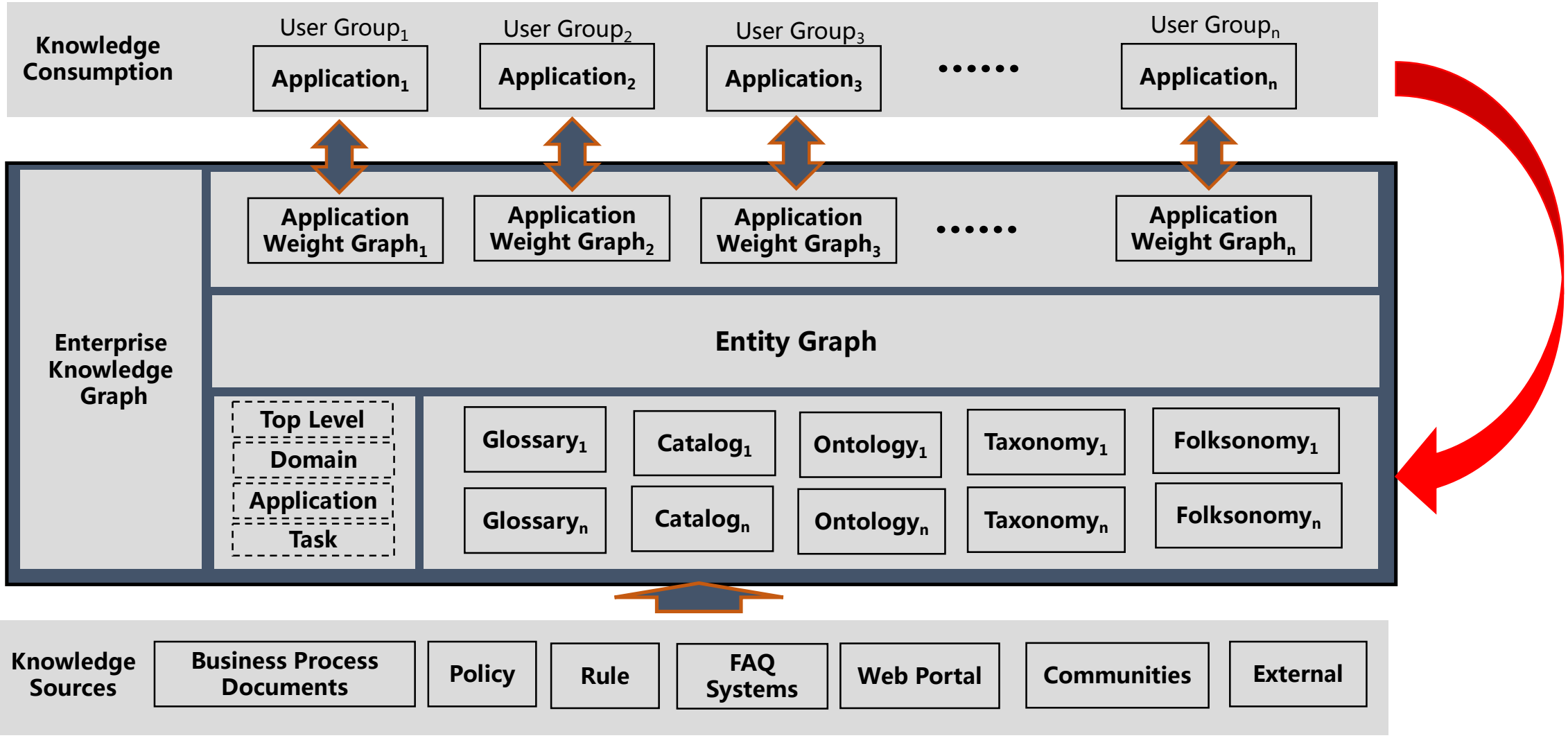
Enterprise Management has multiple levels



# Cross Business Unit Enterprise KG – MultiCycle



# Architecture of Cross Business Unit Enterprise Knowledge Graph



# Outline

I. Enterprise Knowledge and Enterprise Knowledge Graph

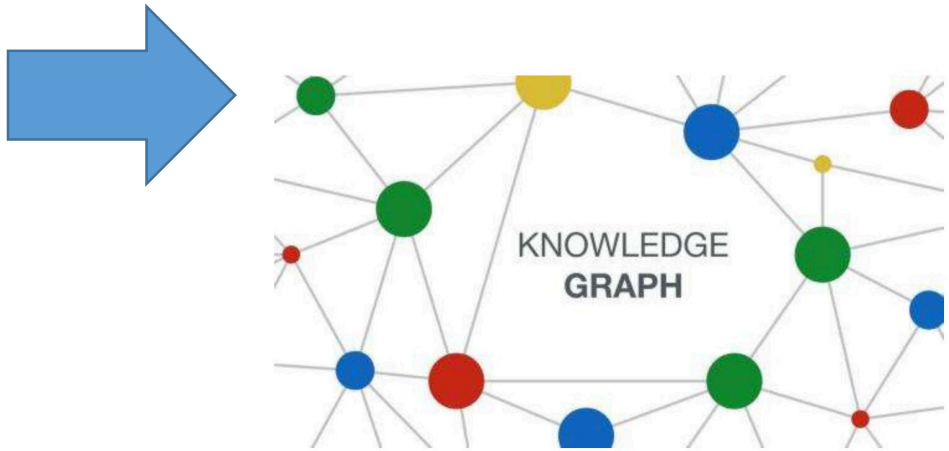
**II. Construction of Enterprise Knowledge Graph**

III. Challenges and Future Research in Enterprise Knowledge Graph

# EKG construction

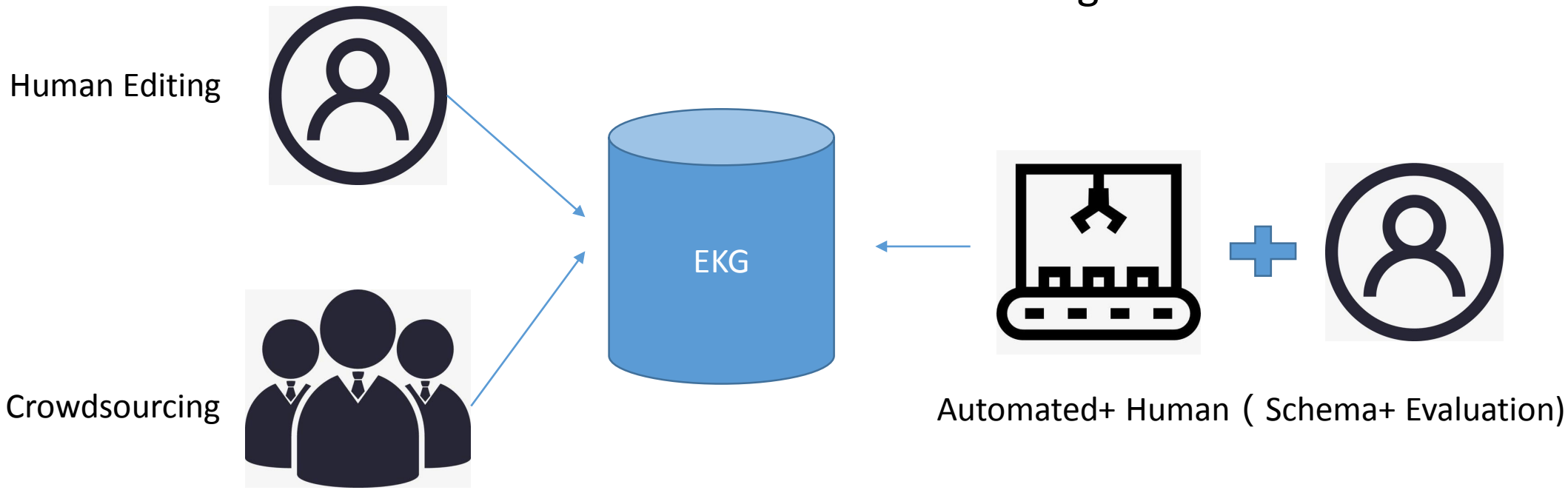
- Data (in/outside of enterprise )


- EKG
  - Concepts, Entities, Relations, Attributes
  - Taxonomies, simple facts



# EKG construction

- Human based
  - Costly
  - High accuracy
- Automatic
  - Low cost
  - High accuracy for simple knowledge



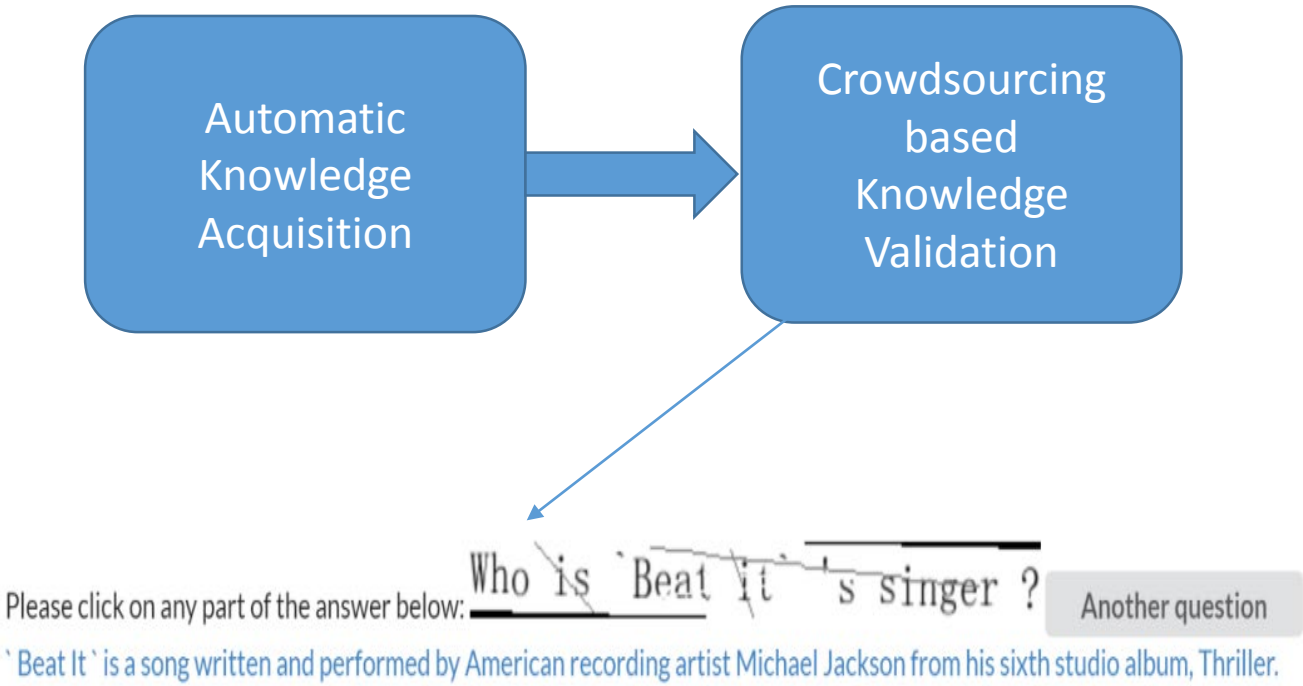
# Human Editing

- Used for
  - Schema Definition
  - Maintenance
    - inserting missing facts
    - deleting wrong facts
    - updating oblivious facts
- Usually used with low frequency


User-friendly tools for schema design and knowledge maintenance

# Crowdsourcing

- Used For
  - Schema design
  - Knowledge validation
- kgCAPTCHA
  - A reading-comprehension based CAPTCHA is used to validate the facts in a large-scale knowledge bases.
  - Uses knowledge to test users and collects human responses to verify new knowledge for the knowledge base.

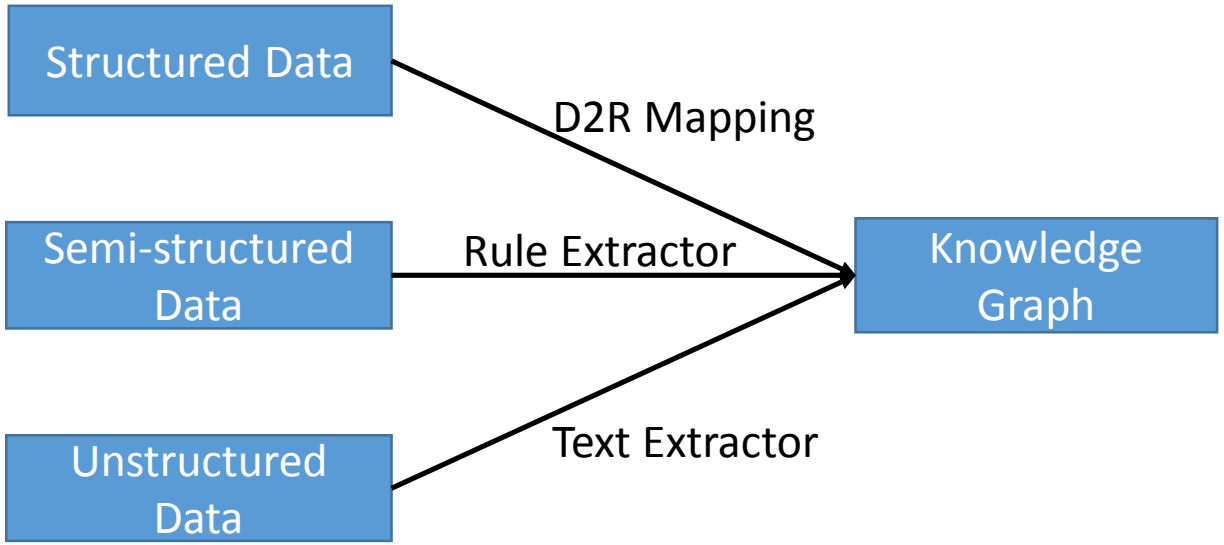



It was used in the validation of CN-Dbpedia and CN-Probase
http://kw.fudan.edu.cn/ddemos/vcode/

Ref： CN-DBpedia2: An Extraction and Verification Framework for Enriching Chinese Encyclopedia Knowledge Base，Data Intelligence, 2019

# Different sources, different approaches

- Structured data
  - Database
- Semis-structured data
  - Documents
  - XML file
  - Wiki
  - Community
- Unstructured data
  - Plain text
  - Image
  - Video




Build a knowledge graph using different knowledge extraction methods with different data source.

# Semi-structured extraction

- Extraction from Wiki like websites: build extractor for each kind of structured data

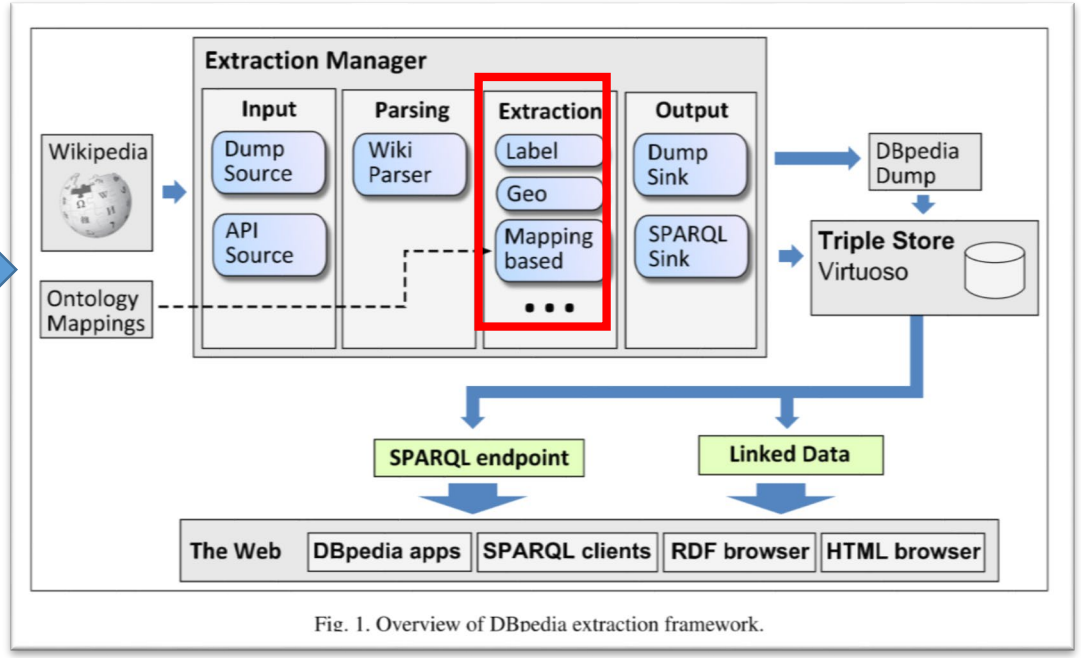



Fig. 1. Overview of DBpedia extraction framework.

[Jens Lehmann et al., 2015] DBpedia: A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia.

# EKG construction from corpus

*The Mona Lisa is a half-length portrait painting by the Italian Renaissance artist Leonardo da Vinci that has*

*Leonardo di ser Piero da Vinci (15 April 1452 – 2 May 1519)*

.....

**Massive Text Corpus**

**Entity**

*Leonardo da Vinci* *Mona Lisa*

**Relation**

*Leonardo da Vinci* → *Mona Lisa*

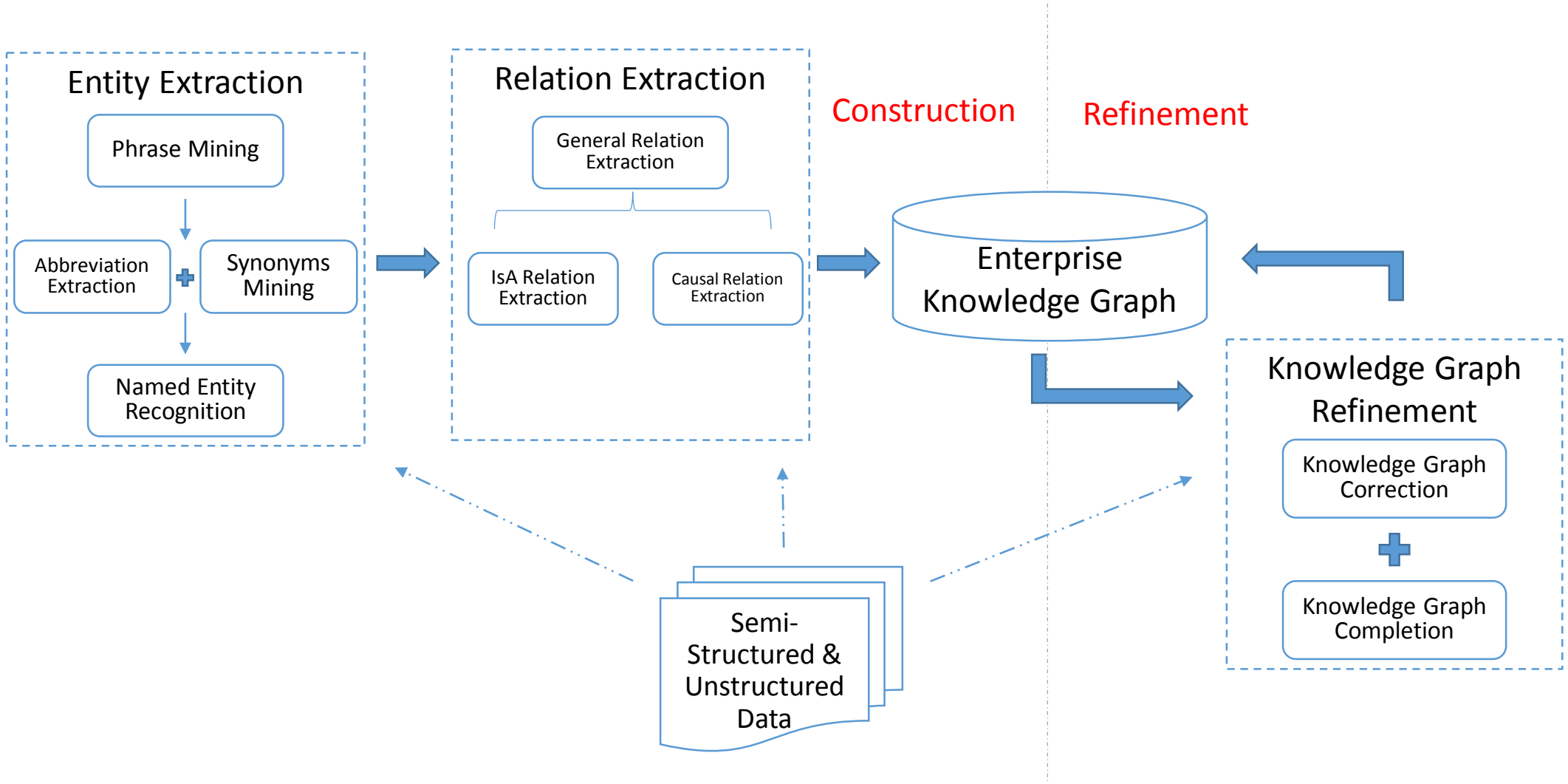
paint

**Attribute Names & Values**

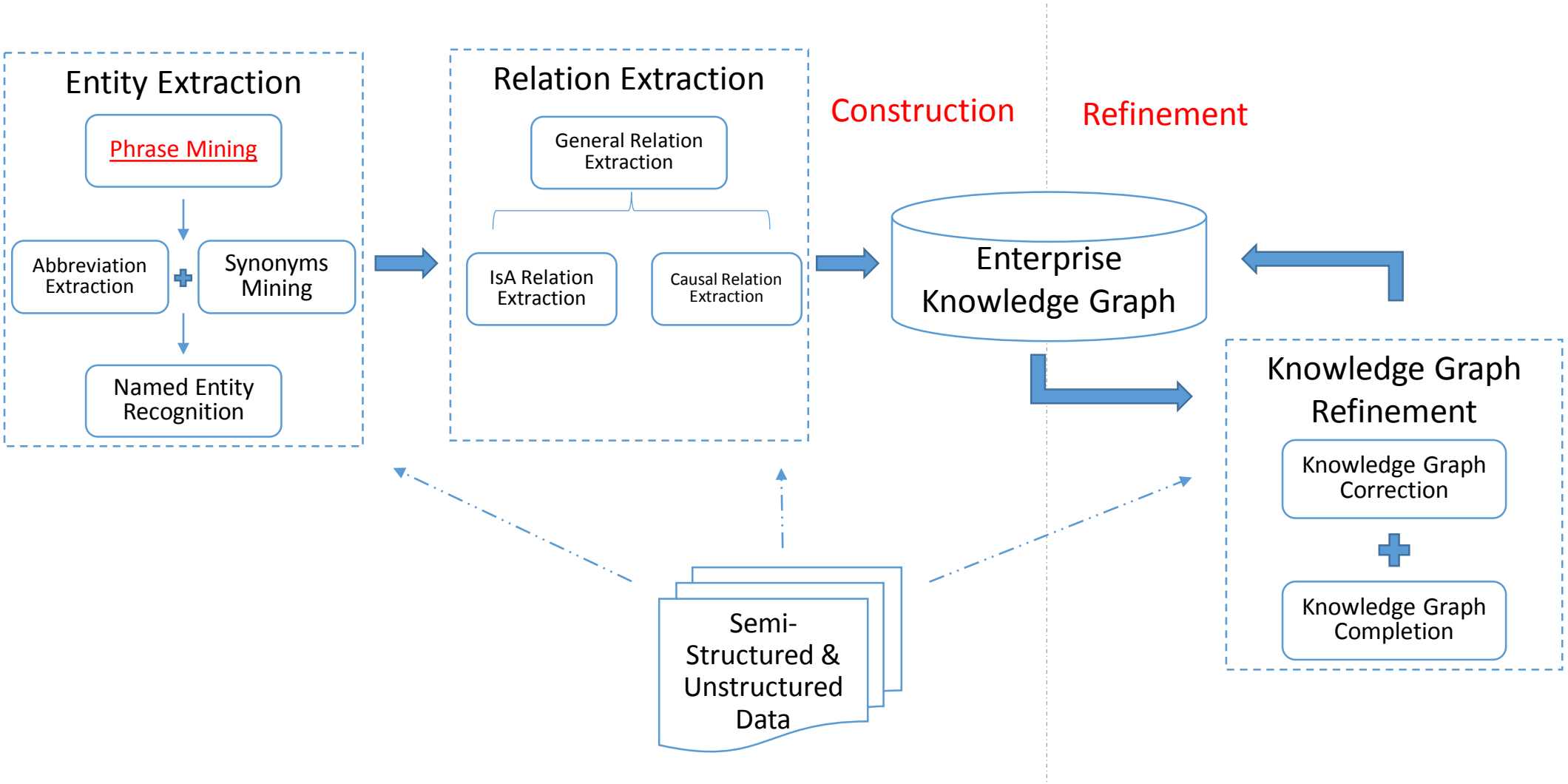
*Leonardo da Vinci*

| Attribute Names | Attribute Values |
|---|---|
| *DateOfBirth* | *April 15, 1452* |
| *DateOfDeath* | *May 2, 1519* |

# Workflow of EKG construction from corpus



# Workflow of EKG construction from corpus



# What is phrase mining?

Phrase Mining

**Raw Corpus**

**Quality Phrases**


A phrase is a sequence of words that appear contiguously in the text, and serves as a whole (non-decomposible) semantic unit in certain context of the given documents.

*What are the criteria of a quality phrase?*

- Popularity
- Concordance
- Informativeness
- Completeness

# What is high quality phrase

- **Popularity:** Frequency
  - “information retrieval” vs. “cross-language information retrieval”
- **Concordance:** A sequence of words that occur more frequently than expected
  - “powerful tea” vs. “strong tea”; “active learning” vs. “learning classification”

$$sig = \frac{count(phr_{x+y}) - E[count(phr_{x+y})])}{\sqrt{count(phr_{x+y})}} \qquad \mathrm{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)} \qquad \chi^2 = \sum \frac{(O-E)^2}{E}$$

- **Informativeness**
  - “this paper” (frequent but not discriminative, not informative)
- **Completeness**
  - “vector machine” vs. “support vector machine”

# Quality Phrase Mining

- Quality phrase mining seeks to extract ***a set of quality phrases*** from ***a large collection of documents***
- Examples:

**Scientific Papers**

| Expected Results |
| --- |
| data mining<br>machine learning<br>information retrieval<br>…<br>support vector machine<br>… |

**News Articles**

| Expected Results |
| --- |
| US President<br>Anderson Cooper<br>Barack Obama<br>…<br>Obama administration<br>… |

# Rule-based Methods : C-value

**A simple idea**: **the sequence of words that of high frequency is likely to be good phrase**

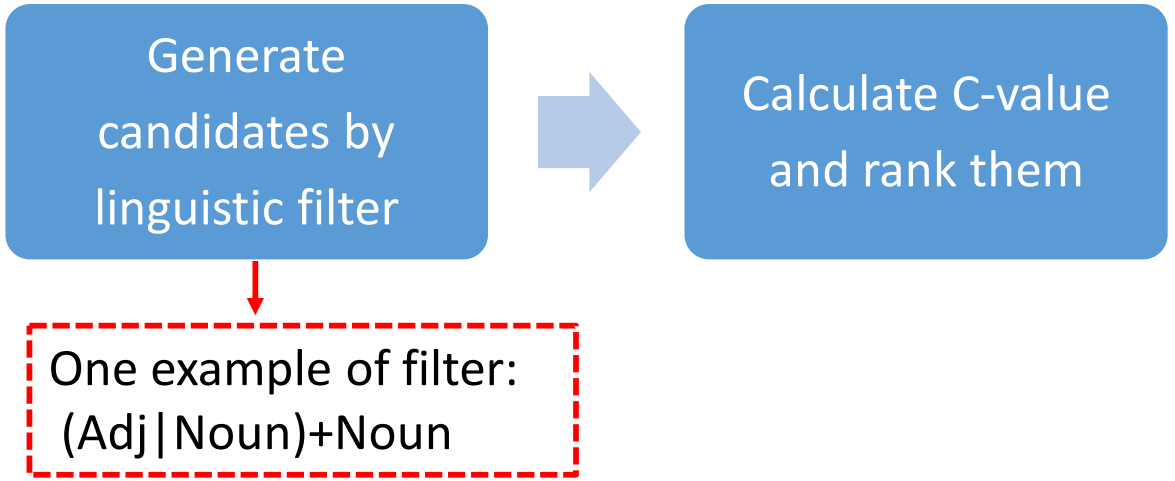
Bad case:

freq(vector machine)>freq(support vector machine),

**C-value** of each candidate, which considers:

- Candidate's frequency
- Candidate's length
- frequency of the super-phrases which contain the candidate
- number of super-phrases(independence)

Process of C-value Method




$$C-value(a)=\begin{cases}\log_2|a|\cdot f(a) & ,a\ is\ not\ nested\\ \log_2|a|(f(a)-\frac{1}{P(T_a)}\sum_{b\in T_a}f(b)), & otherwise\end{cases}$$

Award candidate of longer length

Subtracting the frequency of super phrases

Frantzi, K., Ananiadou, S. and Mima, H., 2000. Automatic recognition of multi-word terms:. the c-value/nc-value method. International Journal on Digital Libraries, 3(2), pp.115-130.

# Rule-based Methods : NC-value

**Context information of the candidates are important for quality estimation.**

- … shows {a basal cell carcinoma}.
- … is called {the Cartesian product}.

We can calculate NC-value of each candidate, which considers:

- Context information
- C-value

**Contribution from the context words**:

select the top 5% candidates from the C-value list to calculate context words and their weights.

$$NC - value(a) = 0.8C - value(a) + 0.2 \sum_{w \in C_a} f_a(w) weight(w)$$

Apply C-value method to the corpus → Extraction of context words and their weights → Calculate NC-value and rank them

Process of NC-value Method

# Data-driven approaches

## Rule-based methods:

- Suffer in domain adaptation.

## Data-driven approaches:

- Require no domain knowledge or specific linguistic rule sets.
- Mine quality phrases according to frequency-based statistical features.

## Types of data-driven approaches:

- Unsupervised methods
- Supervised methods(including distant supervision)

# Supervised data-driven methods

- Input: corpus and labeled data.
- Output: a ranked list of phrases with decreasing quality.

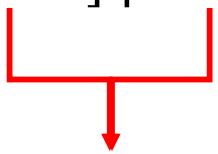
- Representative Methods: SegPhrase and AutoPhrase.
- Key idea
  - Using a binary classifier with several statistical features to predict phrase quality.
- Key issues
  - Labeled data, models, and features.
- Important Features
  - Popularity, Concordance, Informativeness, and Completeness.

# SegPhrase Method

**Raw frequency of phrases aren't appropriate.**

*How to rectify raw frequency of phrases?*

"A standard [feature vector] [machine learning] setup is used to describe..."



"vector machine" cannot be counted

True frequency should be interpreted in whole as a phrase in its occurrence context.

- Phrasal segmentation rectifies the phrase frequency to improve quality estimation.

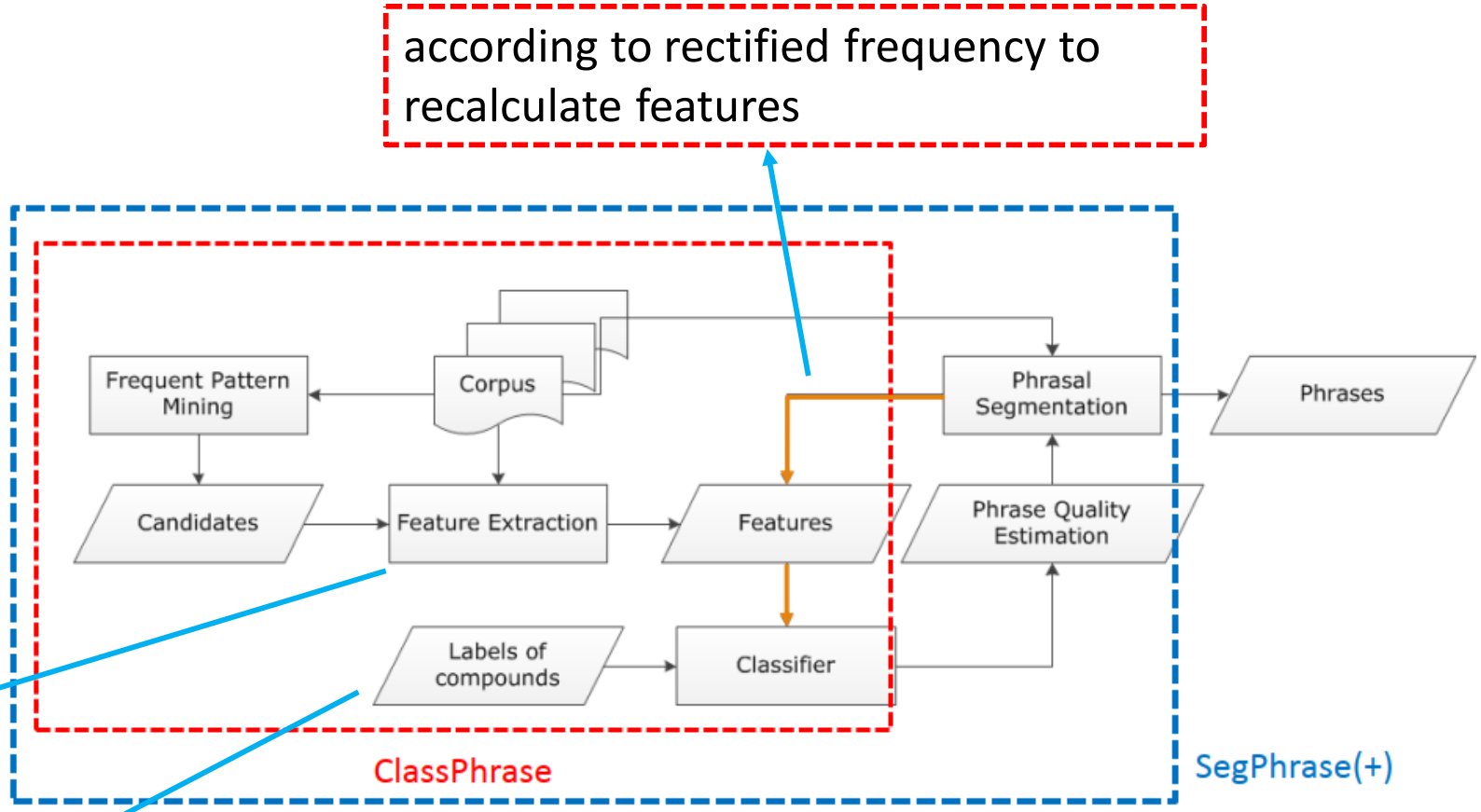
Liu, Jialu, et al. "Mining quality phrases from massive text corpora." Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM, 2015.

| phrases | frequency | rectified |
|---|---|---|
| feature vector | 1 | 1 |
| vector machine | 1 | 0 |
| feature vector machine | 1 | 0 |
| … | | |

raw and rectified frequency of phrases

# SegPhrase Method

- SegPhrase: Phrasal segmentation and phrase quality estimation
- SegPhrase+: One more round to enhance mined phrase quality




features:frequency,pmi,pkl

labels are selected by domain experts

The overall framework of SegPhrase method

# AutoPhrase Method

Extending phrase mining to work automatically is challenging because methods still depend on the human labeling.

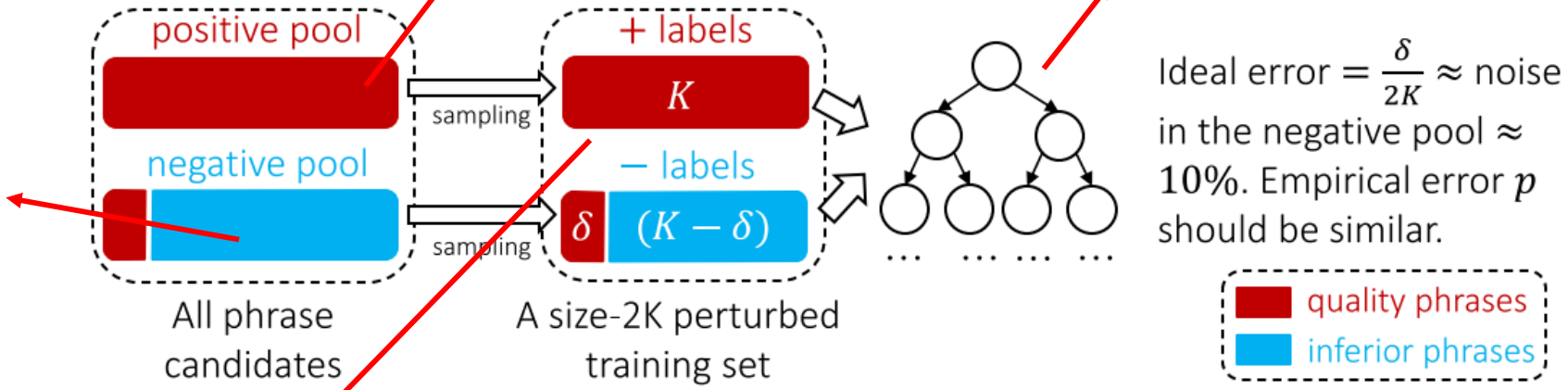
*How to work automatically?*

- Distant supervision

positive pool : Wikipedia phrases

build a decision tree by one training set

negative pool : other candidates except phrases in positive pool




process of building a decision tree

random select K labels separately from each pool

Shang, Jingbo, et al. "Automated phrase mining from massive text corpora." IEEE Transactions on Knowledge and Data Engineering 30.10 (2018): 1825-1837.

# Unsupervised data-driven methods

- Input: Corpus
- Output:  a set of phrases

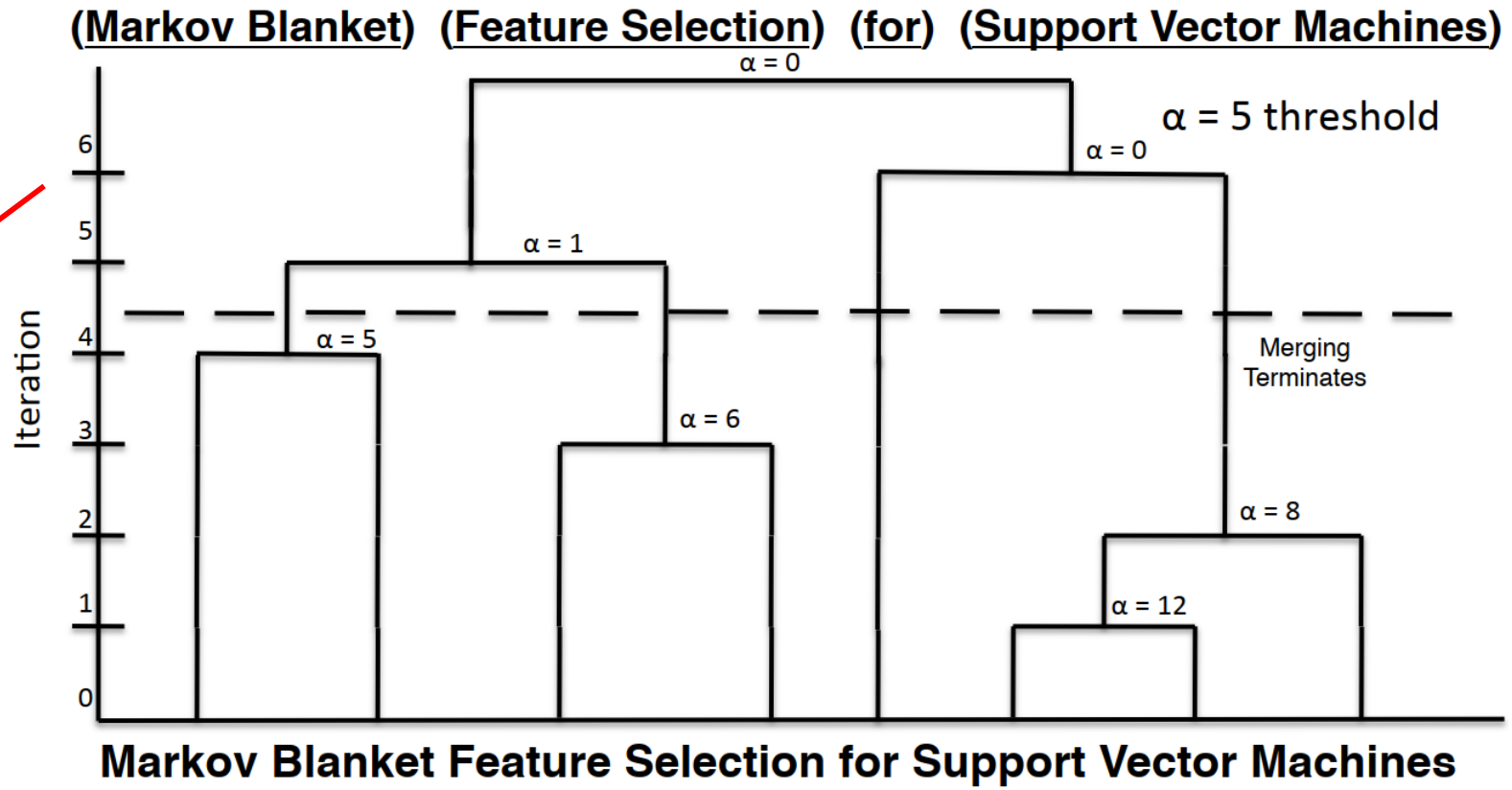
- Basic idea
  - Clustering words based on statistical features(t-test ， $\chi^2$-test, and PMI) to mine high-quality phrases.
- Representative Methods: TopMine, EQPM, and CQMine.

# TopMine method

α is the threshold of merging terminates, which is a hyper-parameter.

1. Mining the corpus for frequent candidate phrases and their aggregate counts.
2. Bottom-up merging process. At each iteration, algorithm makes locally optimal decisions in merging single and multi-word phrases as guided by a statistical significance score.

Merged phrases are final phrases: "Markov Blanket", "Feature Selection", and "Support Vector Machines".

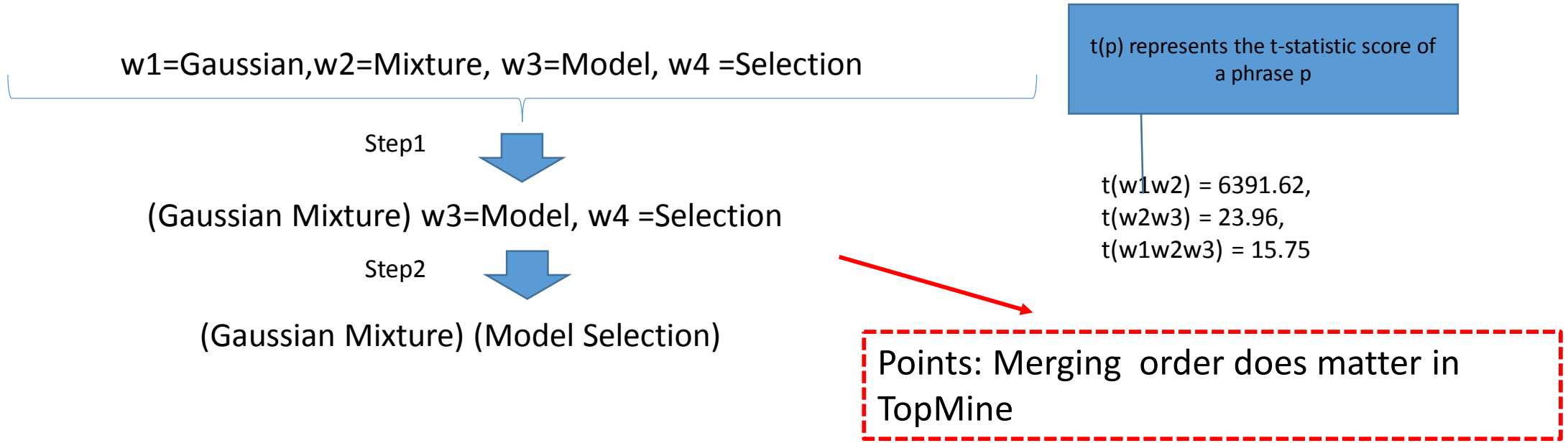



El-Kishky, Ahmed, et al. "Scalable topical phrase mining from text corpora." *Proceedings of the VLDB Endowment* 8.3 (2014): 305-316.

Li, Bing, et al. "Efficiently mining high quality phrases from texts." *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

# EQPM method

- TopMine could produce incomplete phrases.




EQPM enumerate all possible concatenating orders to eliminate order sensitivity and guarantee to avoid incomplete phrases. Two efficient solutions to find the best merging order:

- Dynamic Programming for Complete Phrase Mining
- Seed Extension for Complete Phrase Mining

# CQMine method

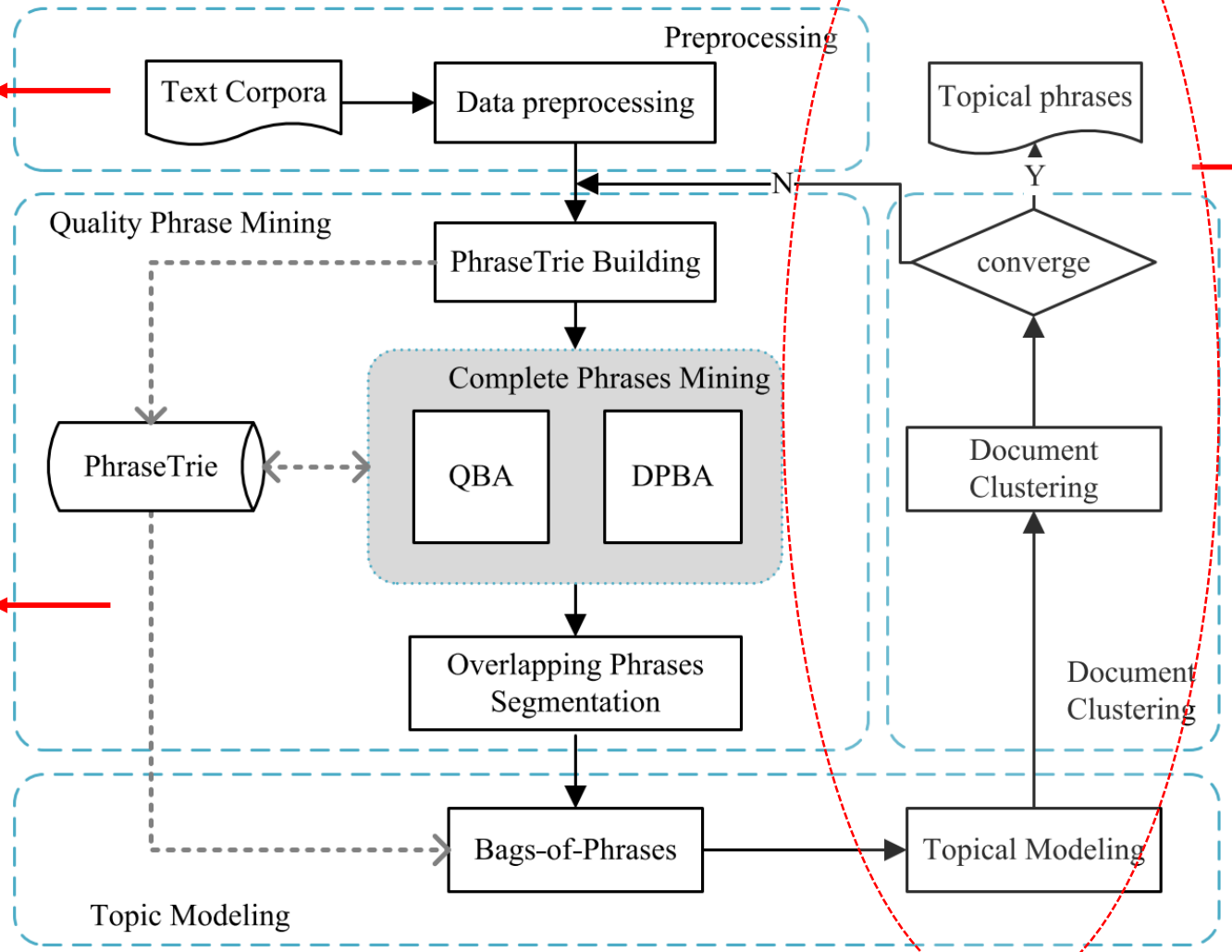
Li, Bing, et al. "An efficient method for high quality and cohesive topical phrase mining." *IEEE Transactions on Knowledge and Data Engineering* 31.1 (2018): 120-137.

- EQPM
  - Pons: can mine phrases with high frequencies.
  - Cons: there are some phrases that only appear in a certain domain, like bit vector appears mostly in the domain of “database”. These phrases are globally infrequent but locally frequent phrases.

- CQMine
  - In order to mine such domain specific phrases, use topic modeling stage and document clustering stage to cluster documents into different domains according to topic phrases .

# CQMine method

tokenization, dropping stop-words and stemming




1. Build a PhraseTrie to count all possible phrases' frequencies
2. Mine complete phrases through QBA or DPBA algorithm
3. Split overlapping phrases

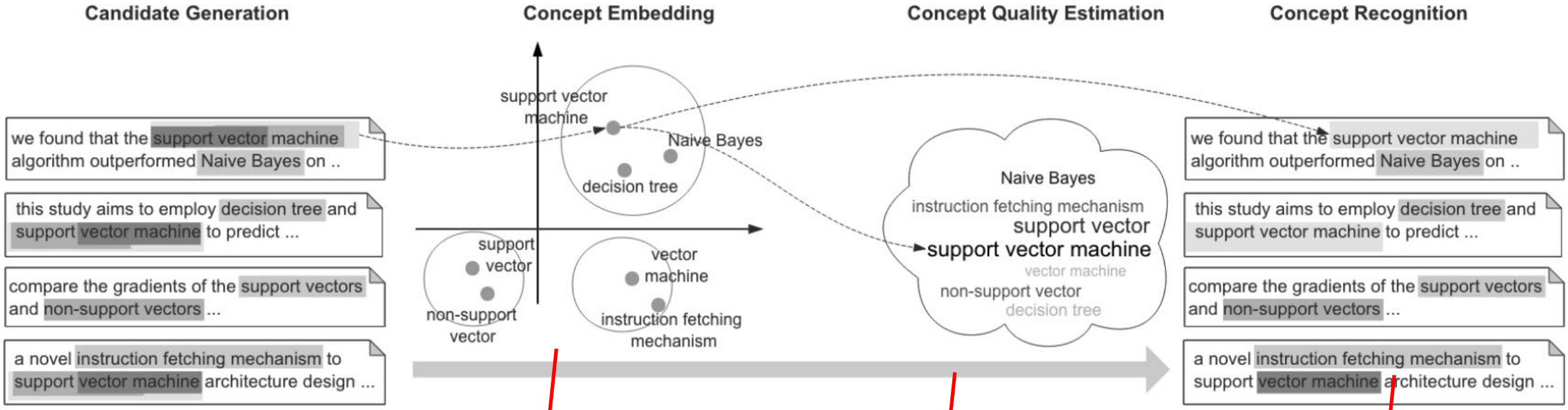
Documents within the same domain are used as the new input of the next round to search for domain-specific phrases.

Li, Keqian, et al. "Concept mining via embedding." *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018.

# Embedding-based method

- Previous methods fail to find long tail infrequent phrases where statistical features are less reliable.
- ECON proposes an embedding based method to find long tail infrequent phrases .
- ECON is developed upon 3 observations:
  - Phrases usually occur under certain usage contexts;
  - Contexts are often shared with phrases with similar meaning;
  - Infrequent phrases may be formed by grouping several frequent words together.

# ECON method



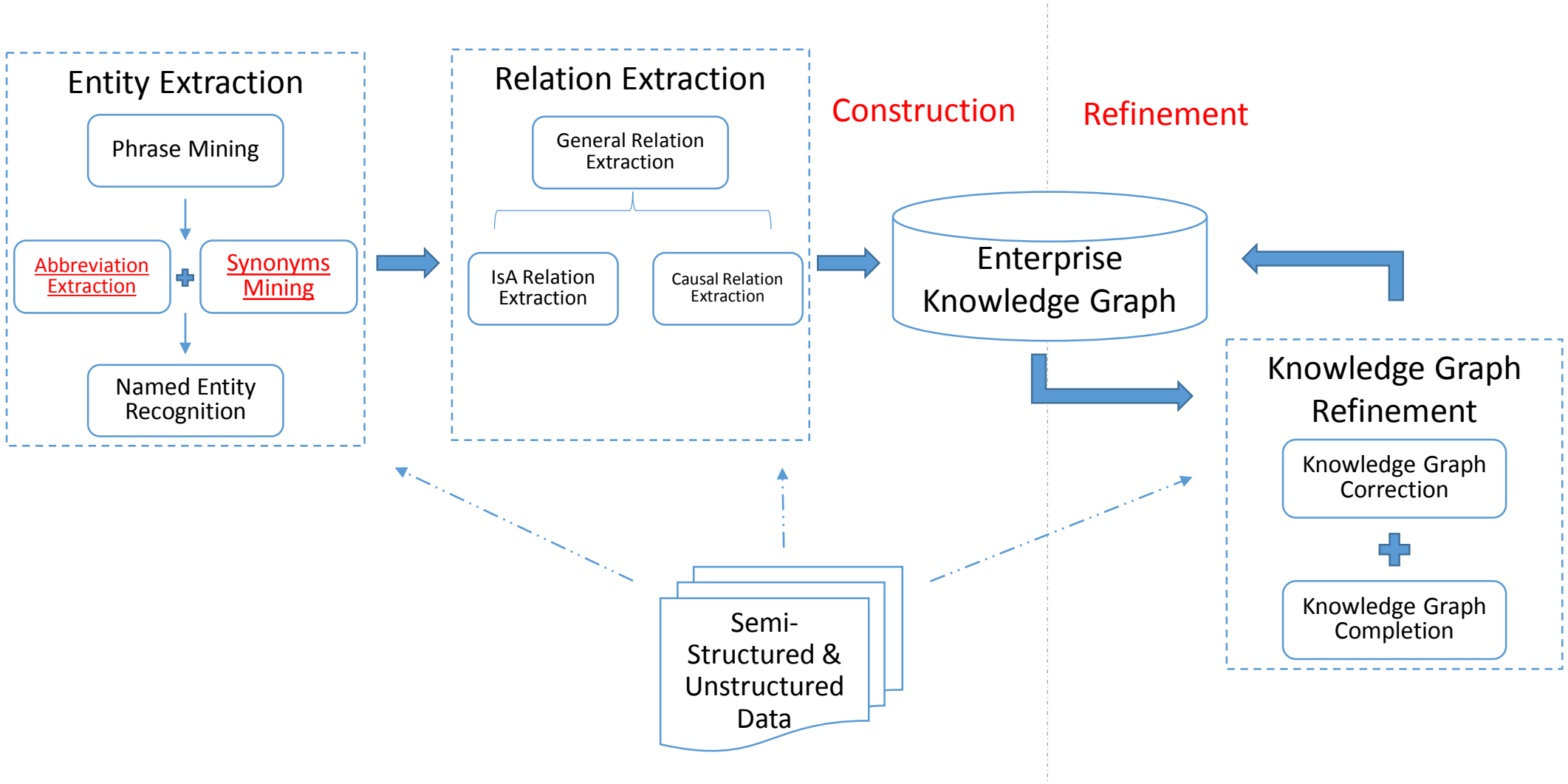

Step1: Utilizing existing techniques to generate candidates phrase along with their occurrences in Text.

Step 2: learning an embedding vector representation for each candidate phrase based on its occurrence contexts.

Step 3: learning a quality score for candidates phrases based on their embedding vectors.

Step 4: determining the true occurrence of phrases in each original document based on both their individual qualities and their fitness to context.

# Workflow of EKG construction from corpus



# Form of Abbreviations

An phrase (entity) is usually mentioned as its abbreviations

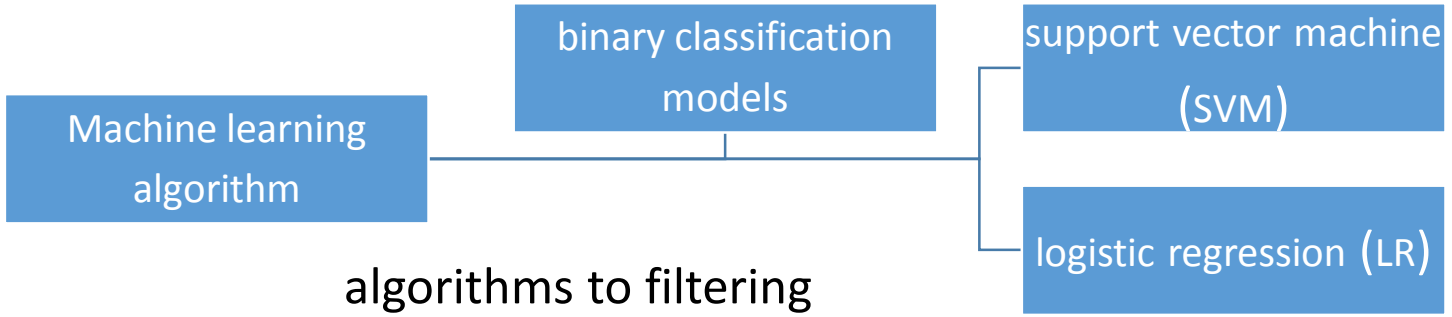
| Language | Category | Source | Short form | Explanations |
|---|---|---|---|---|
| English | Contractions | Doctor, I am | Dr, I'm | Omitting certain letters or syllables and bringing together the first and last letters or elements |
| | Acronyms | Severe Acute Respiratory Syndrome | SARS | Consist of the initial letters or parts of words with new pronounce |
| | Initialisms | British Broadcasting Corporation | BBC | Consist of the initial letters or parts of words without new pronounce |
| Chinese | abbreviations | 中国中央电视台 | 央视 | Eliminating characters form the original words |

# Abbreviations extraction with patterns

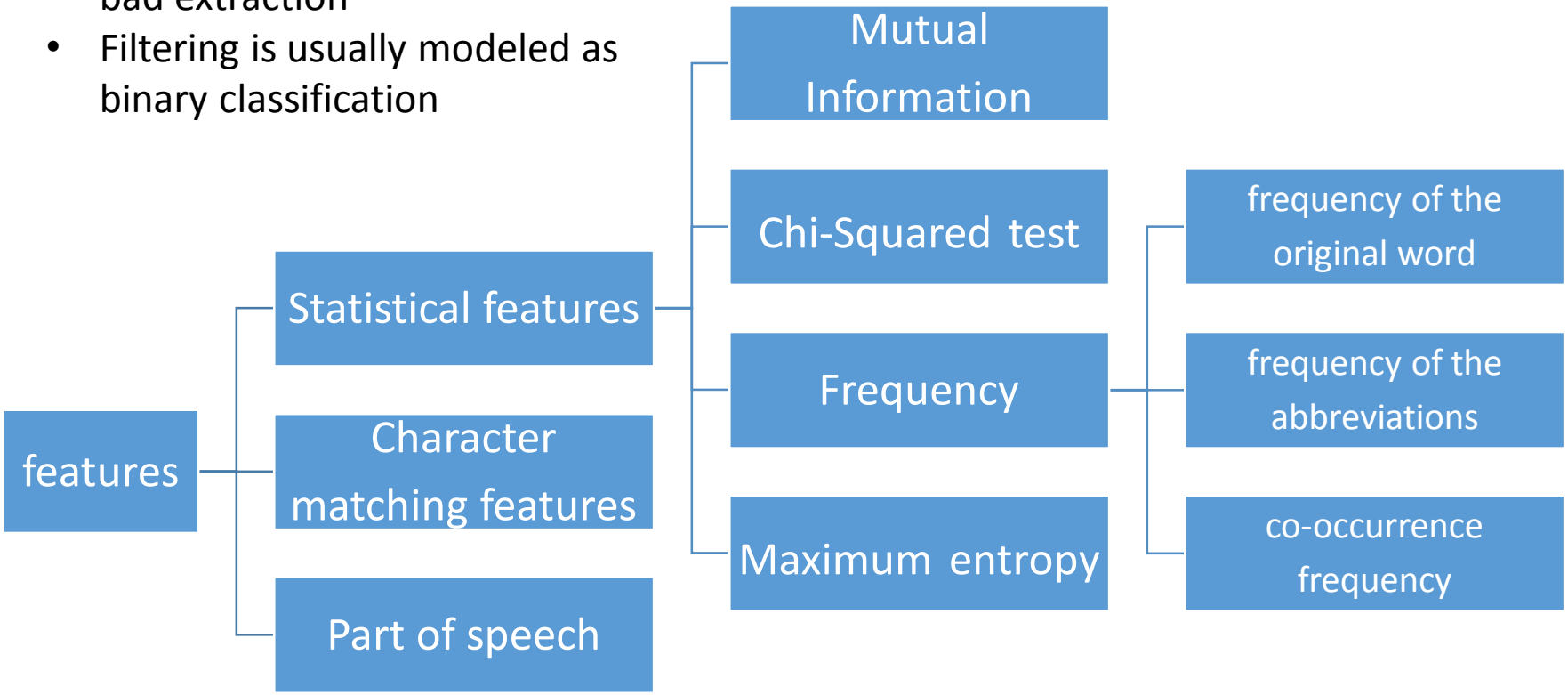
| **Patterns**<br>( A is the original phrase， B represents abbreviation) | **Examples** |
|---|---|
| A(B) | Support vector machine(SVM)<br>Support-vector-machine(SVM) |
| A.{1,}(B) | Support vector machine for gression (SVM) |
| B is the abbreviation of A | SVM is the abbreviation of Support vector machine |
| A, also known as B | Support vector machine, also known as SVM |
| A and B are synoyms | Support vector machine and SVM are synoyms |

# Filtering

- Pattern based extraction is usually followed by a filtering of bad extraction
- Filtering is usually modeled as binary classification




algorithms to filtering



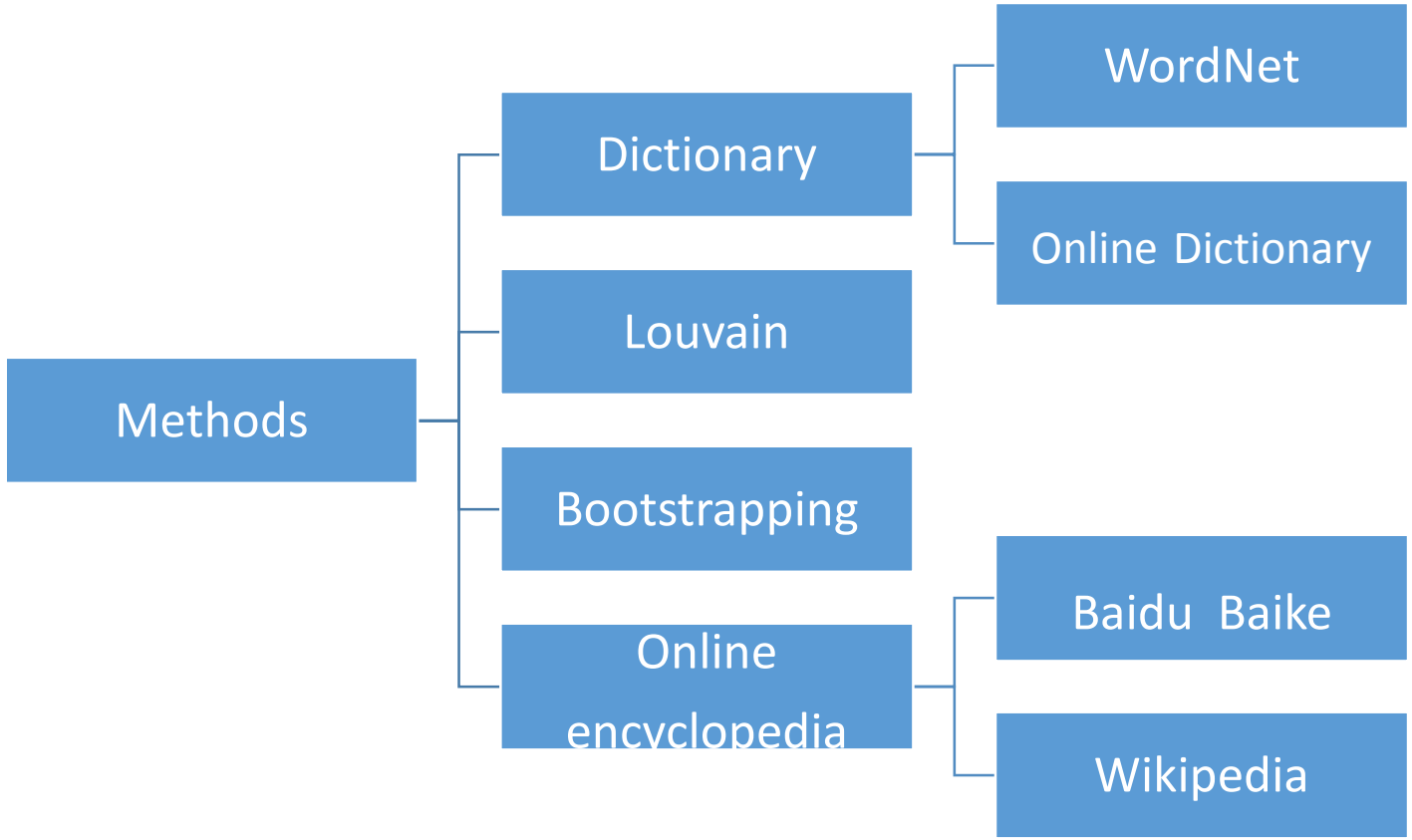

features used in machine learning algorithm

# Synonyms mining

Synonyms express semantic similarity between words, not relevance.
Eg: (Jay Zhou,周董)，(car, automobile)

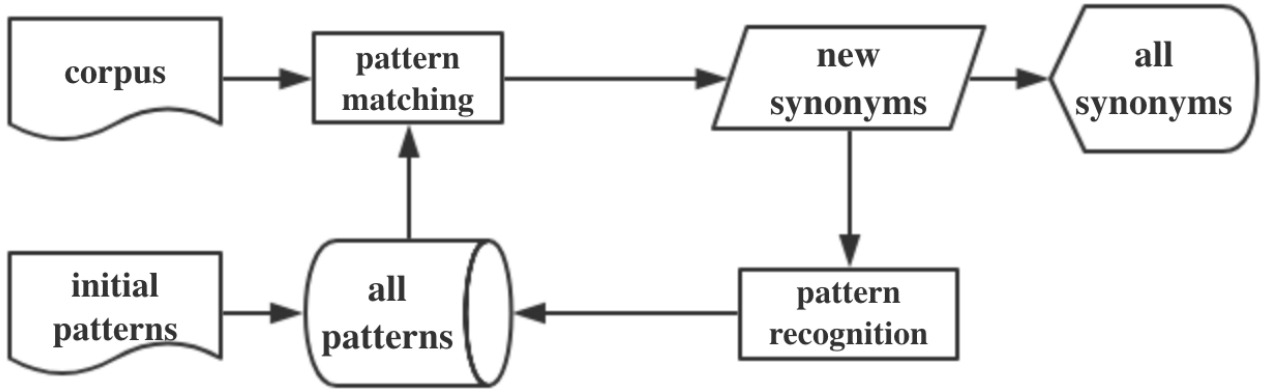
## Synonyms

- Abbreviations
- Nick names
- Gracious names
- Colloquialism
- Mentions in different languages
- ……



# Bootstrapping based Approach

- Hand-crafted patterns are always limited.
- We need an iterative procedure to find more patterns in an automatic manner




| Pattern（X and Y are synonyms） | examples |
|---|---|
| X 又称 Y | 番茄又称西红柿 |
| X，亦称 Y | 计量，亦称测量 |
| X（Y） | 宋太祖（朱元璋） |

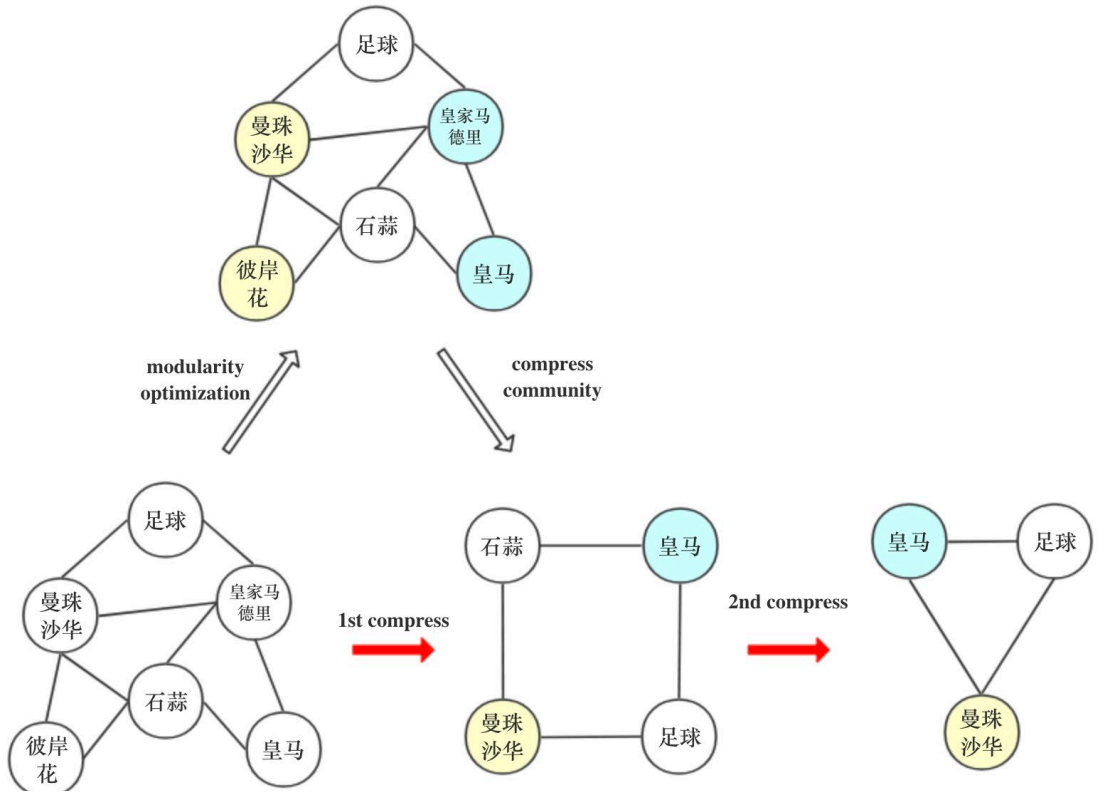
1）Data preparation
input corpus & initial patterns
2）pattern matching：
mine new synonyms through pattern matching
3）pattern recognition:
for each new synonym in the corpus, find a new pattern.
4）Loop step 2 - 3

# A graph based approach

Louvian: Building a term similarity graph, and finding synonyms by mining community structure in the graph:

1 ) build a community
2 ) add edge
3 ) loop step 2
4 ) compress the diagram
5 ) loop step 1 – 4

Find the best community structure by minimizing modularity Q.

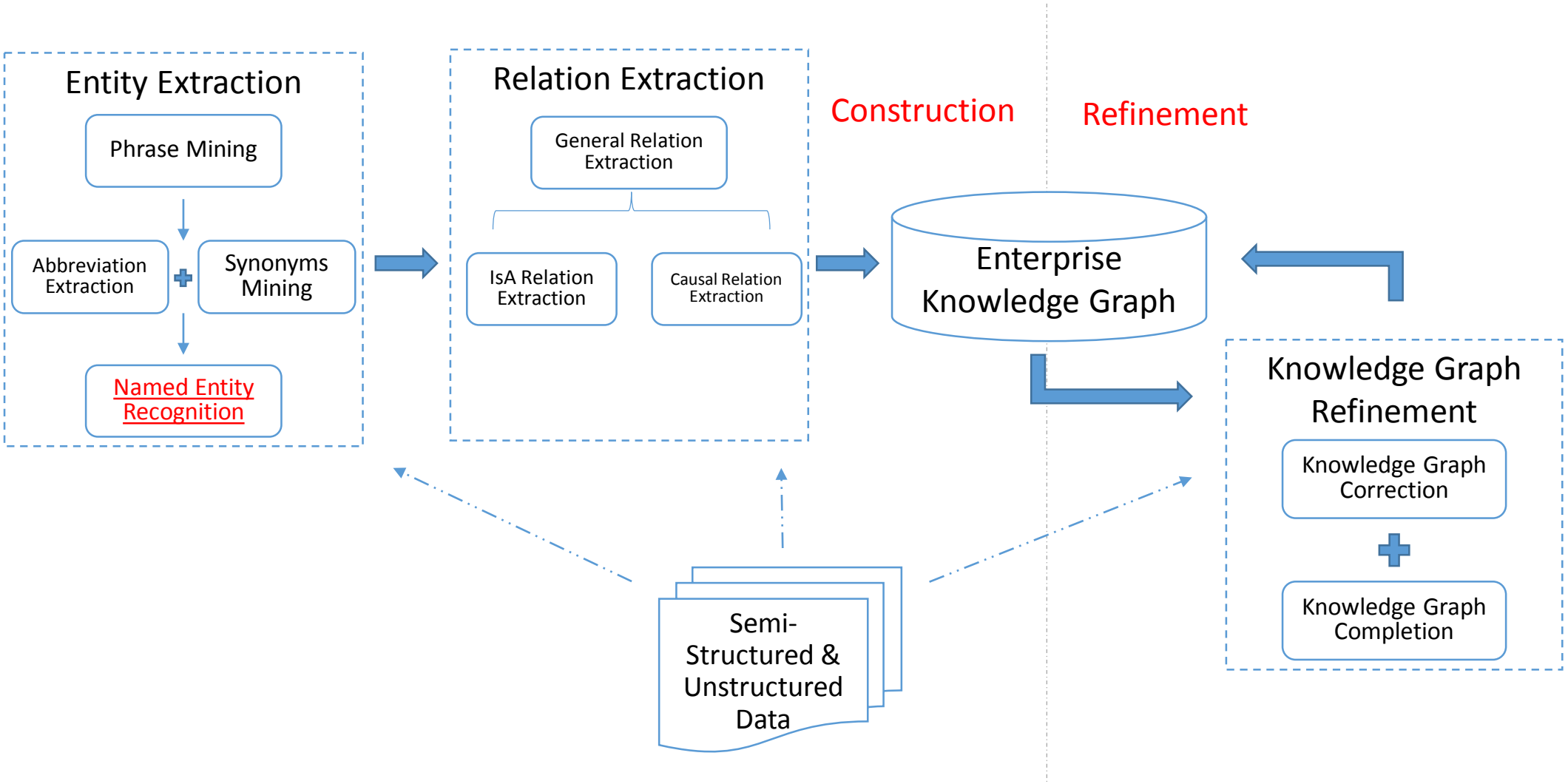



$$Q = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \quad , \quad \delta(u,v) = \begin{cases} 1 & u = v \\ 0 & u \neq v \end{cases}$$

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right]$$

Blondel, Vincent D; Guillaume, Jean-Loup; Lambiotte, Renaud; Lefebvre, Etienne (9 October 2008). "Fast unfolding of communities in large networks". *Journal of Statistical Mechanics: Theory and Experiment*. **2008** (10): P10008

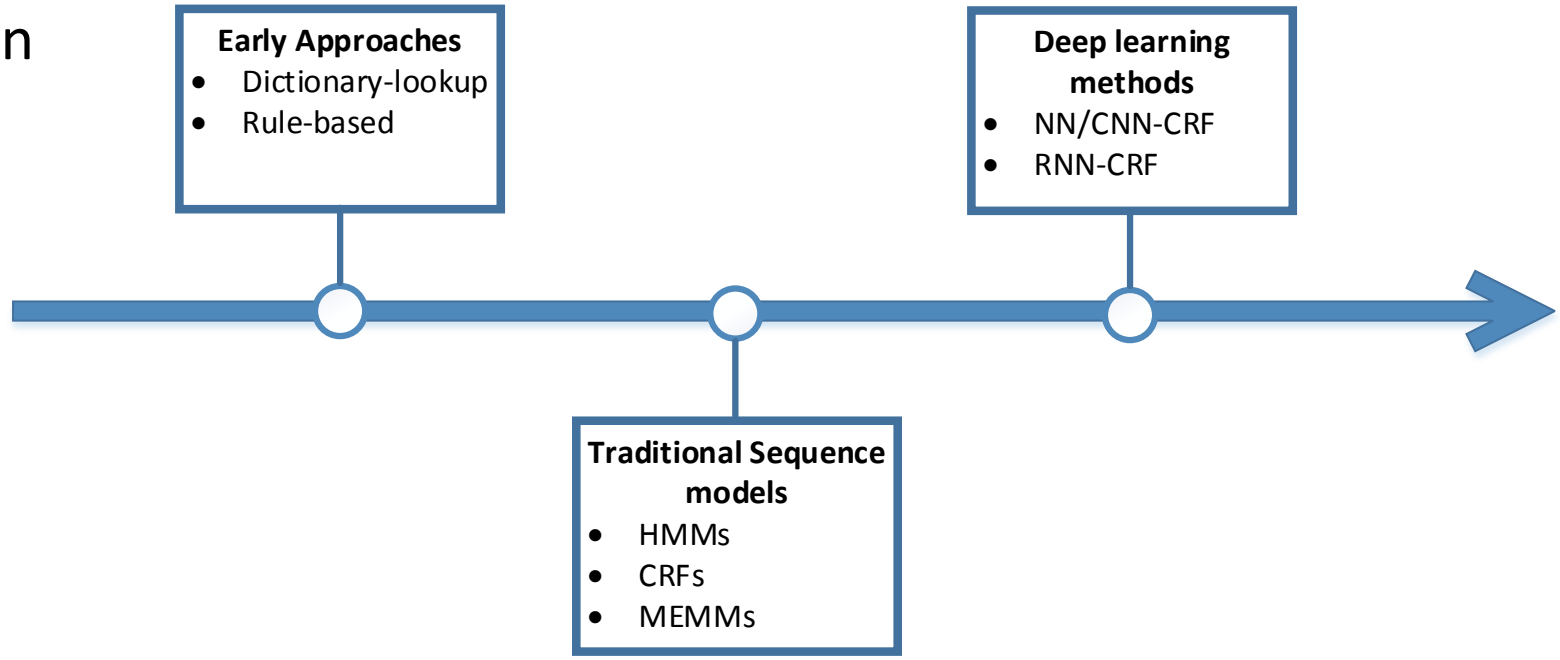
# Workflow of EKG construction from corpus



# Name Entity Recognition: Overview

- **Find and classify entities** in text
- Key component of Information Extraction system

- In recent years, DL-based NER models become dominant and achieve state-of-the-art results.



# Rule-based methods

- Needs more rules to tag all kinds of NE
- Rely on hand-crafted rules

Eg. Determining where an organization is located

*[org] in [loc]* → *NATO headquarters in Brussels*

*[org] [loc]* → *KFOR Kosovo headquarters*

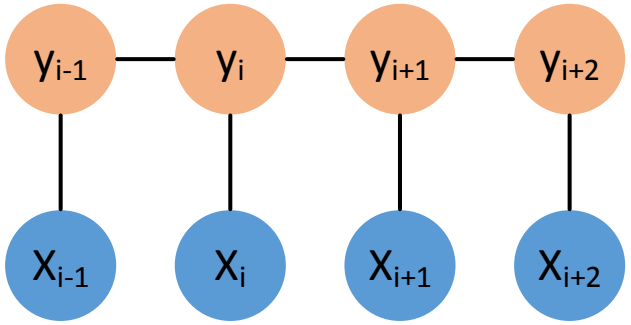
# Pros and Cons of Rule-based methods

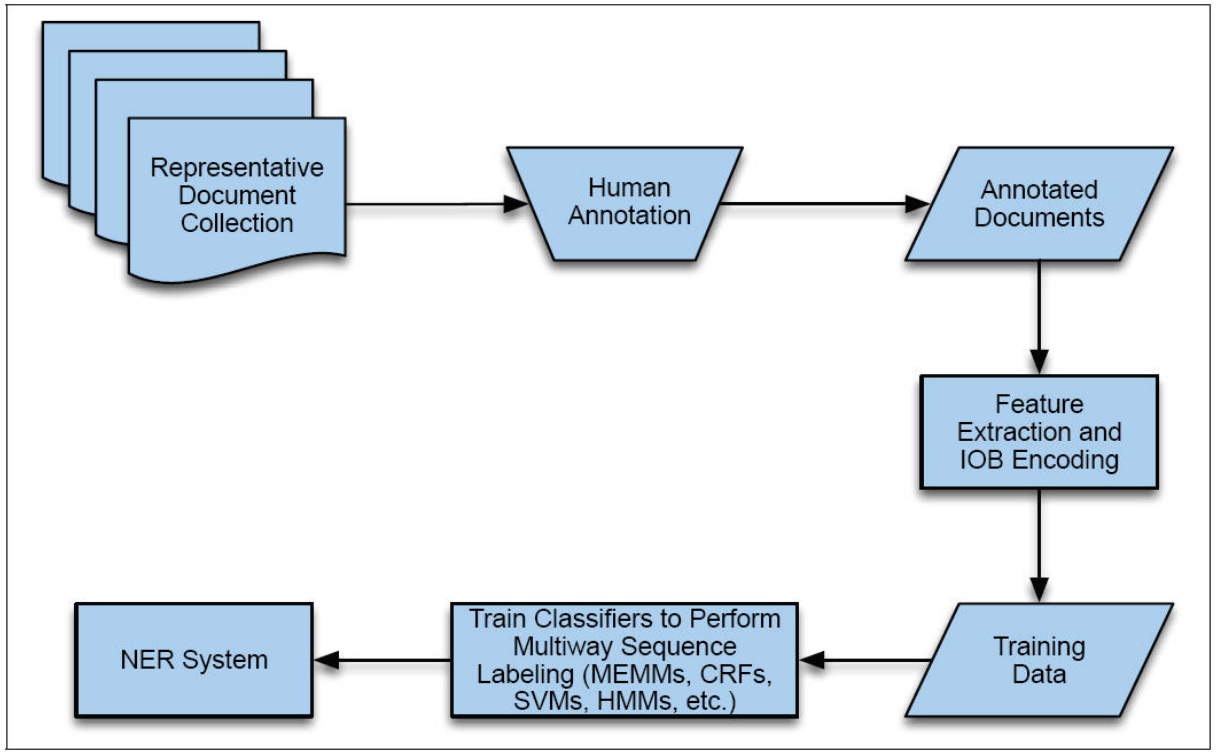
- Advantages:
  - Rich and expressive rules
  - Good results

- Disadvantages:
  - Requires expertise and grammatical knowledge
  - Rely on experts to craft rules are expensive
  - Highly domain specific ( not portable to a new domain)

# Traditional Sequential Models

- Sequence labeling problem



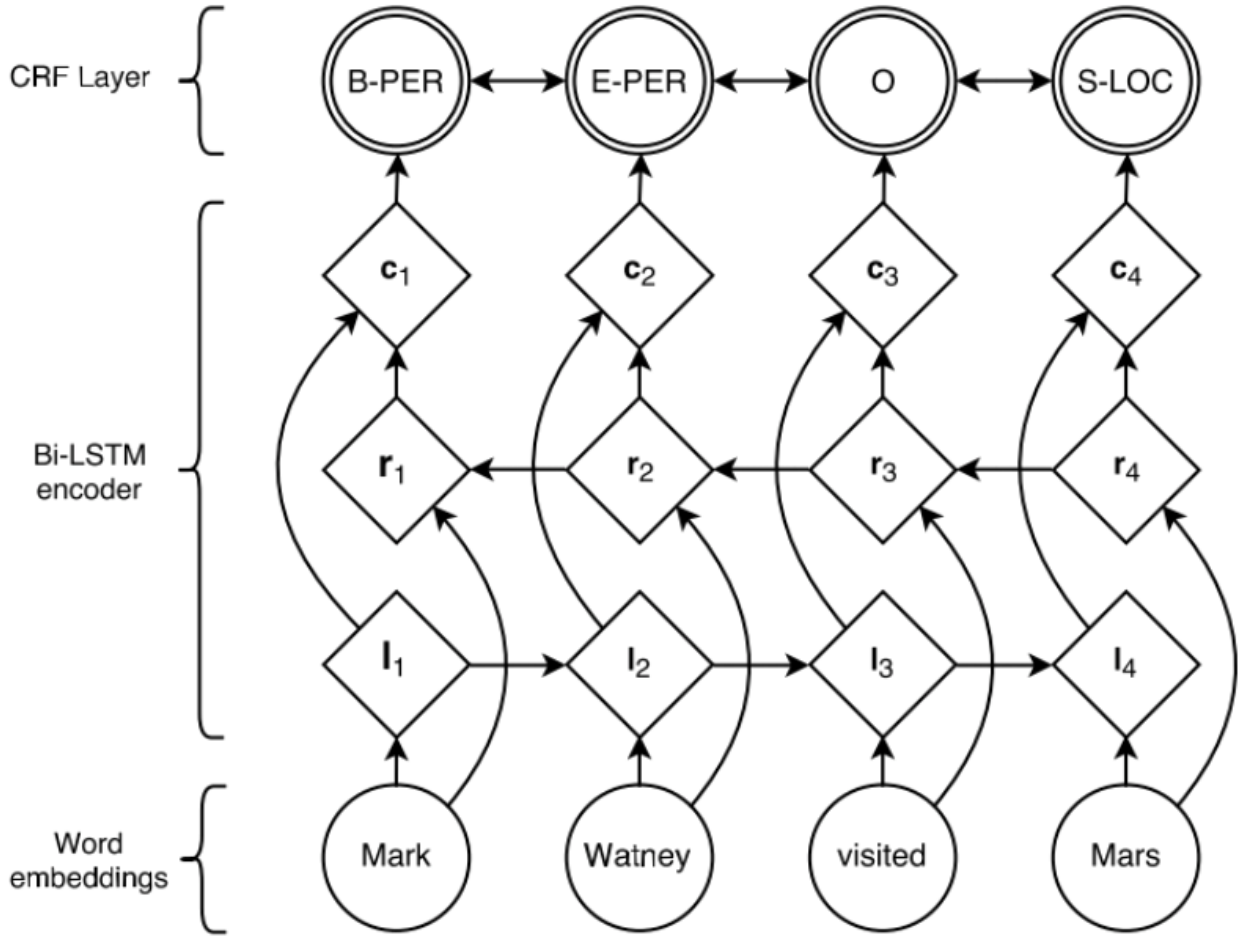




- Models
  - Hidden Markov Model (HMM)
  - Conditional random Fields (CRF)
  - Maximum Entropy Markov Model (MEMM)

# Deep learning methods

- BiLSTM-CRF is the most popular deep learning architecture for NER.

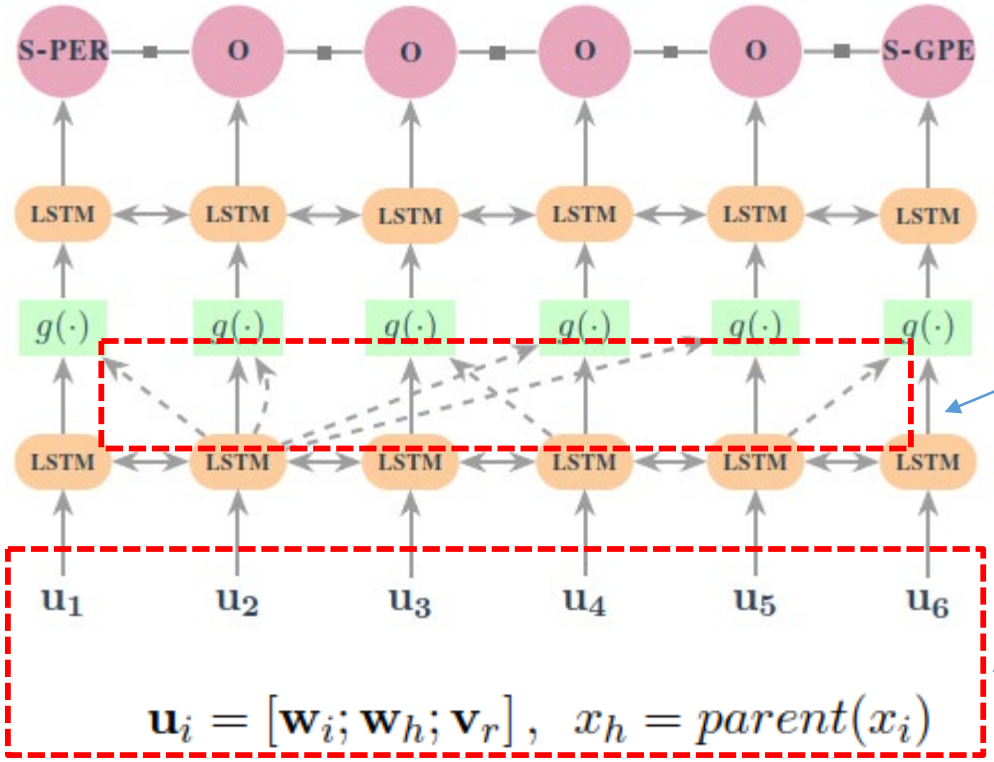
- Architecture
  - Embedding layer
  - Bi-LSTM layer
  - CRF layer

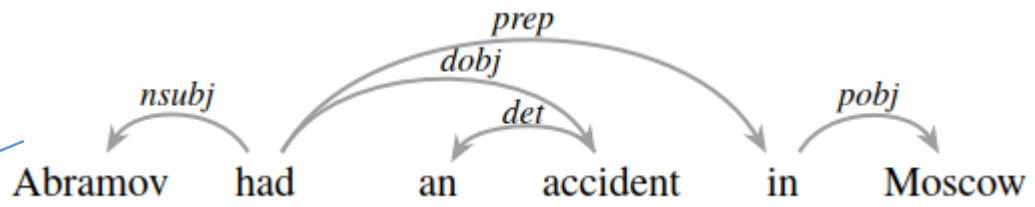



Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991, 2015.

# DGLSTM-CRF

Propose a simple yet effective dependency-guided LSTM-CRF model to encode the complete dependency trees and capture the above properties for the task of named entity recognition(NER).



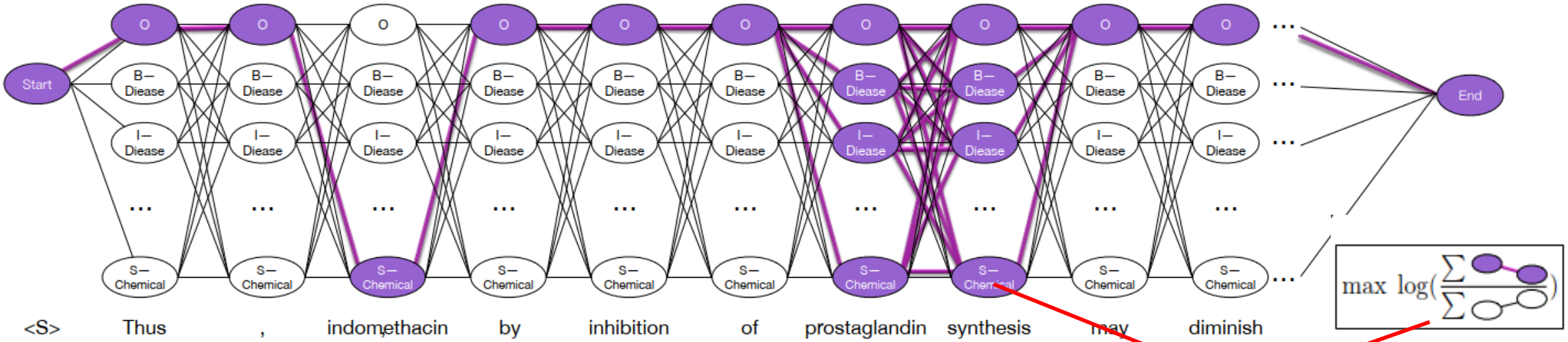

Jie, Zhanming, and Wei Lu. "Dependency-Guided LSTM-CRF for Named Entity Recognition." Emnlp 2019

$w_i$ and $w_h$ are the word representations of the word $x_i$ and its parent $x_h$, $v_r$ is the embedding for the dependency relation r

# AutoNER

Jingbo Shang*, Liyuan Liu*, Xiaotao Gu, Xiang Ren, Teng Ren and Jiawei Han, “Learning Named Entity Tagger using Domain-Specific Dictionary”,in Proc. of 2018 Conf. on Empirical Methods in Natural Language Processing (EMNLP'18), Brussels, Belgium, Oct. 2018. (* Equal Contribution)

- Previous sequence labelling model only considers one labelling solution for an input sentence, but some entities are multi-typed or unknown-typed tokens with distant supervision
- It is reasonable to consider multiple possible labelling solutions
- AutoNer propose Fuzzy-LSTM-CRF with Modfied IOBES.



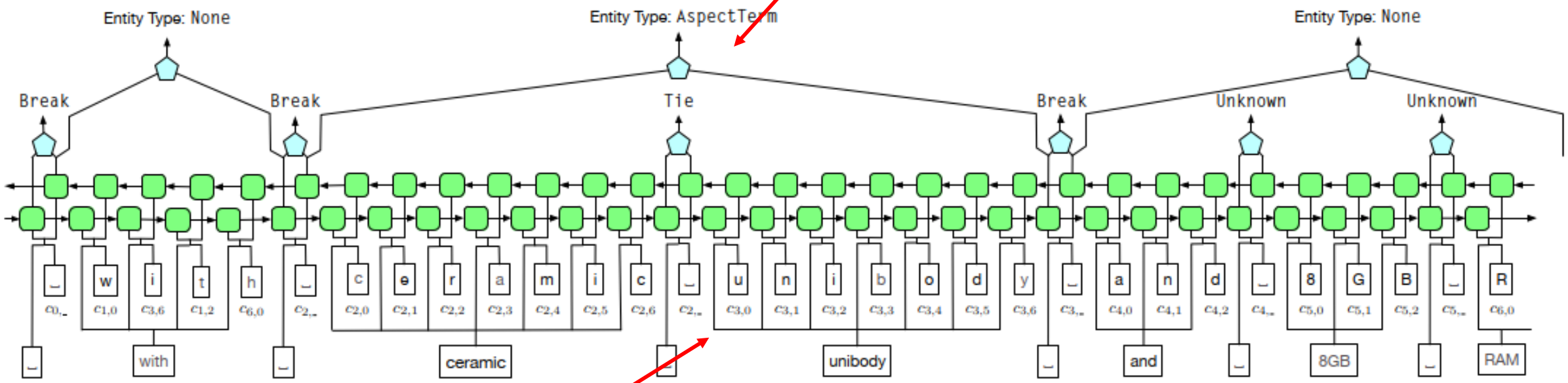

“indomethacin” is a matched Chemical entity,
“prostaglandin synthesis” is an unknown-typed high-quality phrase.

Paths from Start to End marked as purple form all possible label sequences given the distant supervision.

# AutoNER

Propose a new tagging scheme, Tie or Break.
Designing a novel neural model for this scheme

The named entity type is {AspectTerm}.




"ceramic unibody" is a matched AspectTerm entity and "8GB RAM" is an unknown-typed high-quality phrase.

For every two adjacent tokens, the connection between them is labeled as:

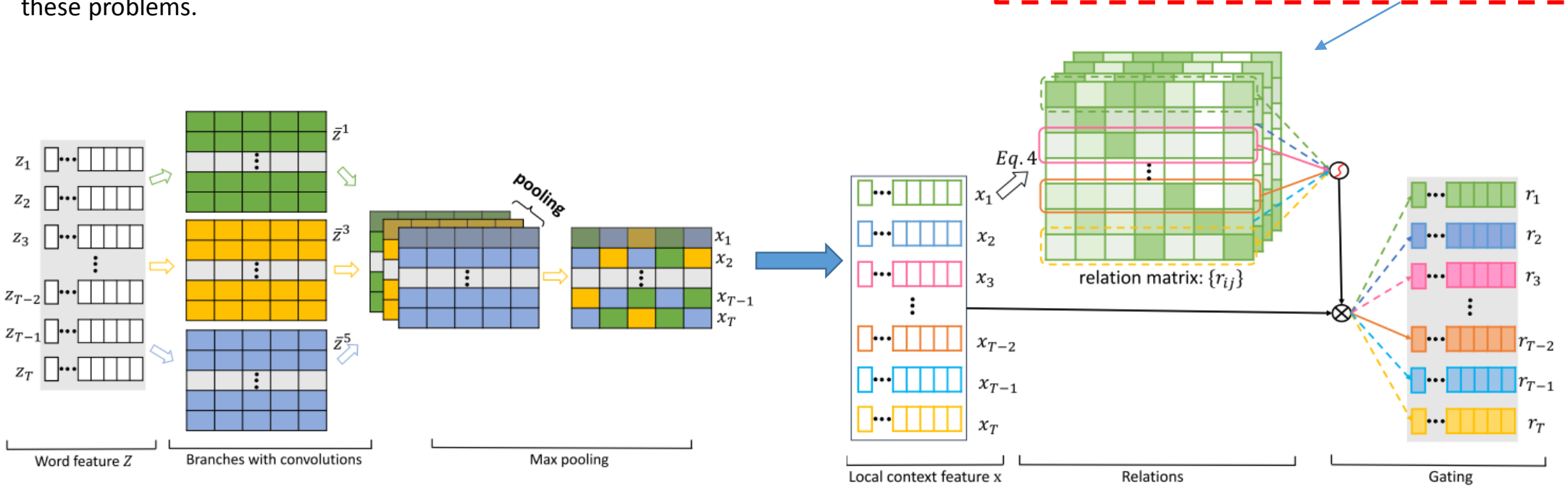
- Tie, when the two tokens are matched to the same entity;
- Unknown, if at least one of the tokens belongs to an unknown-typed high-quality phrase;
- Break, otherwise.

# GRN

Traditional RNN-based NER models are limited by their computational efficiency, while CNN-based models can not capture the long-term context information.

This method proposed effective CNN-based gated relation network (GRN) to handle these problems.

After obtain the context representation using traditional CNN, another *relation layer is used to model the relations between any two words in the sentence.*



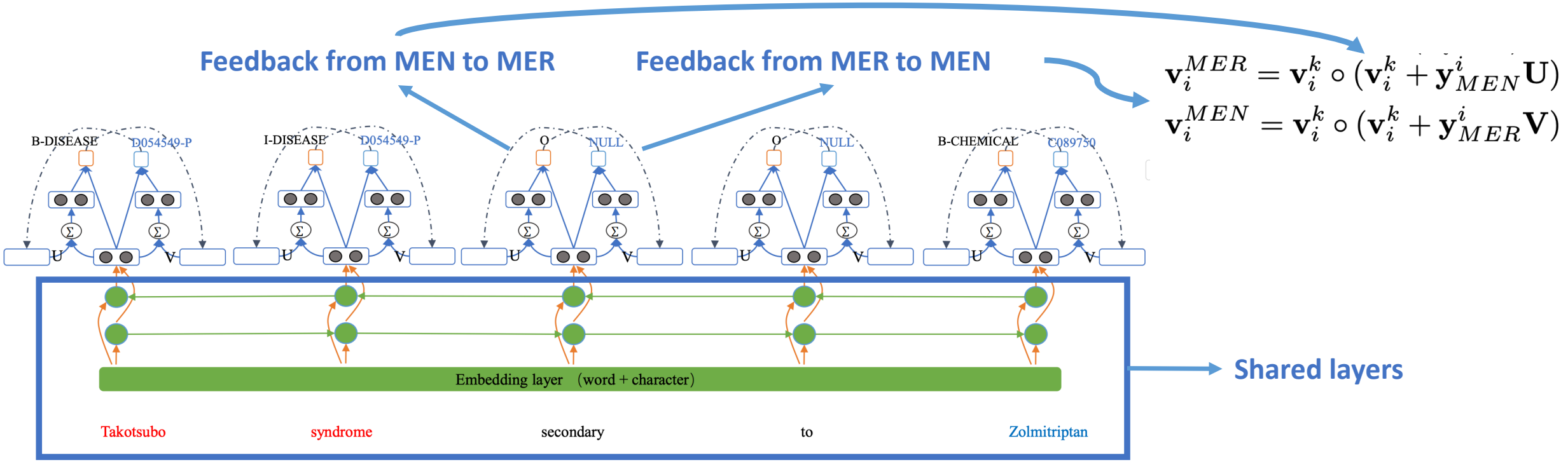

Relation between word $i$ and word $j$ is obtained by: $r_{ij} = W_{rx}[x_i; x_j] + b_{rx}$

then a gating mechanism is introduced: $r_i = \frac{1}{T}\sum_{j=1}^{T} \sigma(r_{ij}) \odot x_j$

# Jointly perform named entity recognition and normalization

Medical entity recognition(MER): find the boundaries of mentions from the medical text.

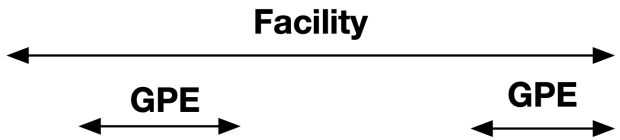
Medical entity normalization(MEN): map mentions onto a pre-defined medical vocabulary.




Zhao S, Liu T, Zhao S, et al. A Neural Multi-Task Learning Framework to Jointly Model Medical Named Entity Recognition and Normalization[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 817-824.

# Recognizing entities for non-overlapping and totally nested

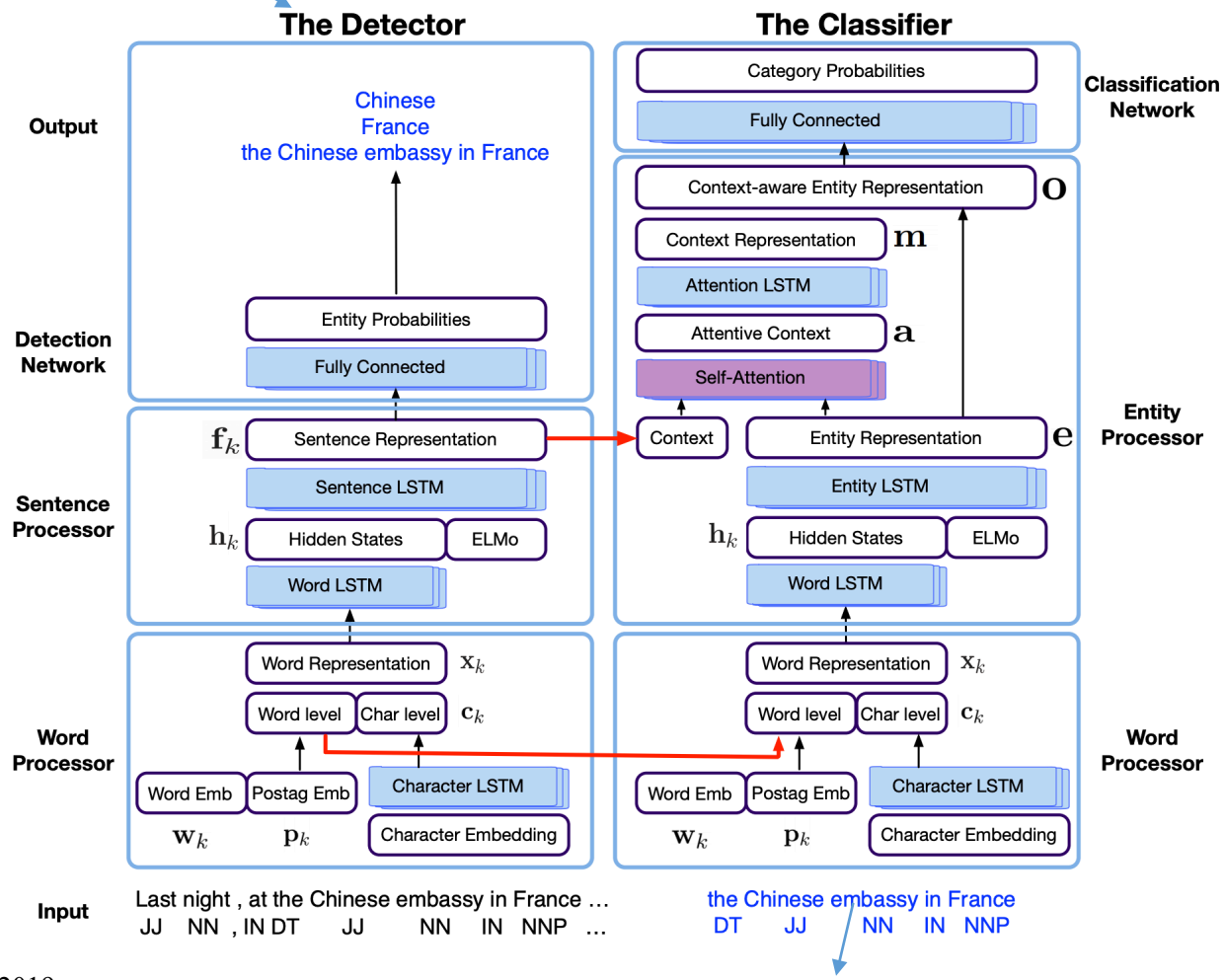
- Entity mentions in a sentence could be
  - Non-overlapping （ Chinese and France)
  - Totally nested (Chinese and “Chinese embassy in France”)




Last night , at the Chinese embassy in France , there was a holiday atmosphere .

Existing methods are designed explicitly for one type of mentions, which usually do not perform well on the other type

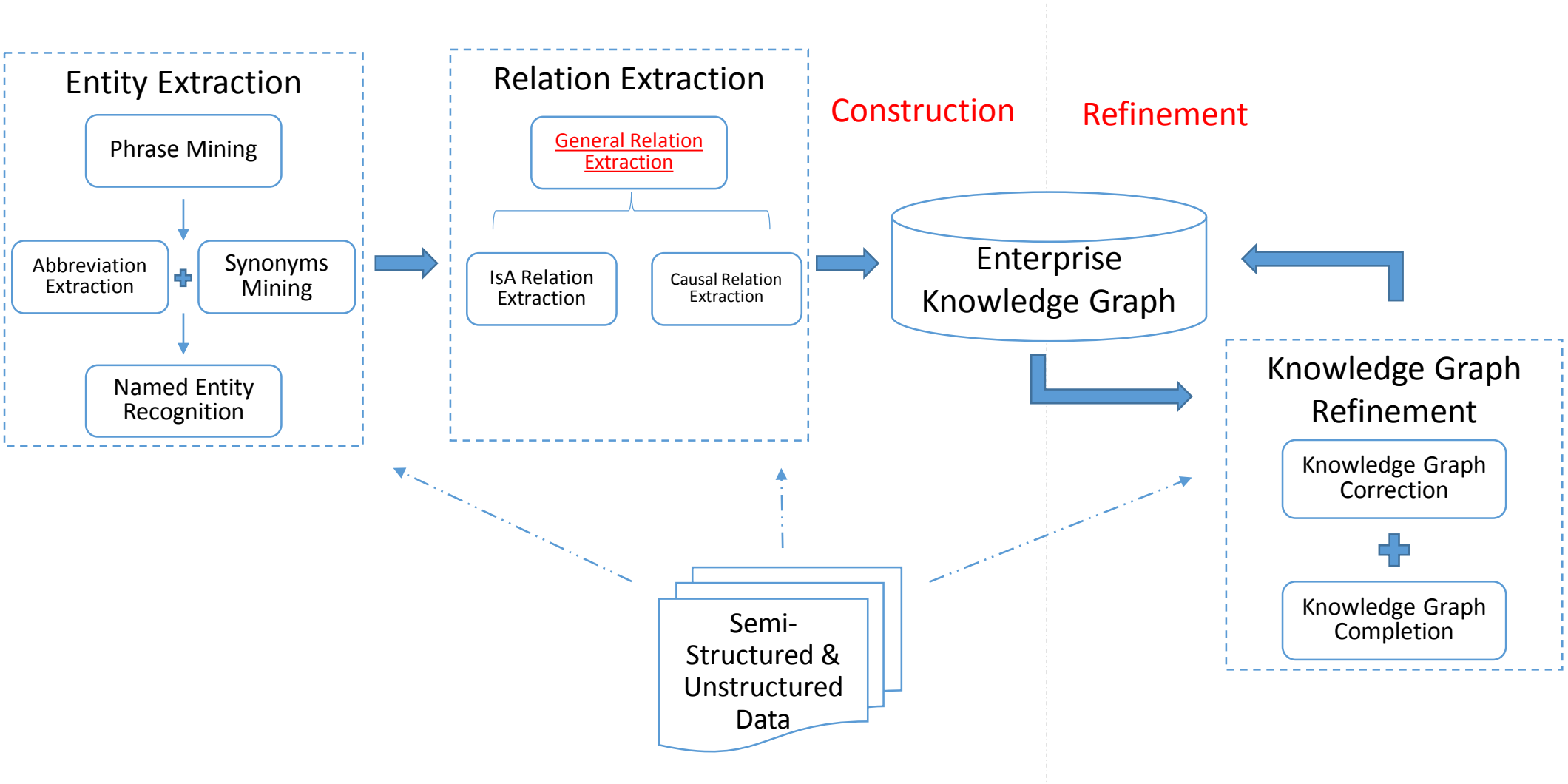
Detect all the possible entity positions




Classify entities into pre-defined entity categories

Xia C, Zhang C, Yang T, et al. Multi-Grained Named Entity Recognition[C]//ACL. 2019.

# Workflow of EKG construction from corpus



# Relation Extraction

- Introduction to Relation Extraction
- Pattern-based methods
- Bootstrapping methods
- Supervised methods
- Distant supervision methods

# Introduction to Relation Extraction

- Relation extraction is the task to extract **relation triples** from the **unstructured texts**.

Not logged in Talk Contributions Create account Log in

Article Talk | Read Edit View history | Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book

## Fudan University

From Wikipedia, the free encyclopedia

Coordinates: 31°17′56″N 121°29′57″E

**Fudan University** (simplified Chinese: 复旦大学; traditional Chinese: 復旦大學; pinyin: *Fùdàn Dàxué*), located in Shanghai, is one of the most prestigious and selective universities in China.[3] It is a C9 League university and a Chinese Ministry of Education Class A Double First Class University.[4] Its institutional predecessor was founded in 1905, shortly before the end of China's imperial Qing dynasty. Fudan is now composed of four campuses in Shanghai – Handan (邯郸), Fenglin (枫林), Zhangjiang (张江), and Jiangwan (江湾) – which share the same central administration.

Contents [hide]

1 History
2 Institutions
  2.1 Organizations
  2.2 Library
3 Student associations
  3.1 Presidents
  3.2 Admissions
    3.2.1 Undergraduate program
    3.2.2 Graduate program
    3.2.3 International students
4 Reputation and rankings
5 Campus

**Fudan University**

复旦大学

| | |
|---|---|
| **Former names** | Fudan Public School (1905)<br>Fudan College<br>Private Fudan University (1917-1941)<br>National Fudan University (1941-1949) |
| **Motto** | 博学而笃志，切问而近思[1] Scientia et studium, quaestio et cogitation (Latin) |
| **Motto in English** | Rich in knowledge and tenacious of purposes, inquiring with earnestness and |

**Fudan University**, located in **Shanghai**, ... It is a **C9 League** university and a Chinese Ministry of Education Class A ... Its institutional predecessor was founded in **1905**, shortly before the end of China's imperial Qing dynasty...

Fudan University, Location, Shanghai
Fudan University, Time_Founded, 1905
Fudan University, Property, C9 League

# Pattern-based methods

**isA: NP such as NP**
**Business: The main business of NP is NP**

- Simple hand-written regular expression
  - Limited expressive power


```
WIKIPEDIA INFOBOX

$InfoBox_Body: <tbody>[$Image_Item]($Item)*</tbody>;
……

$Item:         ($Author_Item | $Country_Item | $Other_Item);
$Author_Item:  <tr><th>作者</th><td>$Author</td></tr>;
$Country_Item: <tr><th>出版地</th><td>$Country</td></tr>;
$Other_Item:   <tr><th>$Property</th><td>$Value</td></tr>;
$Image_Item:   <tr>$Image_Info</tr>;

$Author:       (.)+(?=</td>);
$Country:      (.)+(?=</td>);
$Value:        (.)+(?=</td>);
$Property:     ? 非已知关系对应字符串 ?;
$Image_Item:   ? 图片地址及描述信息 ?;

END
```

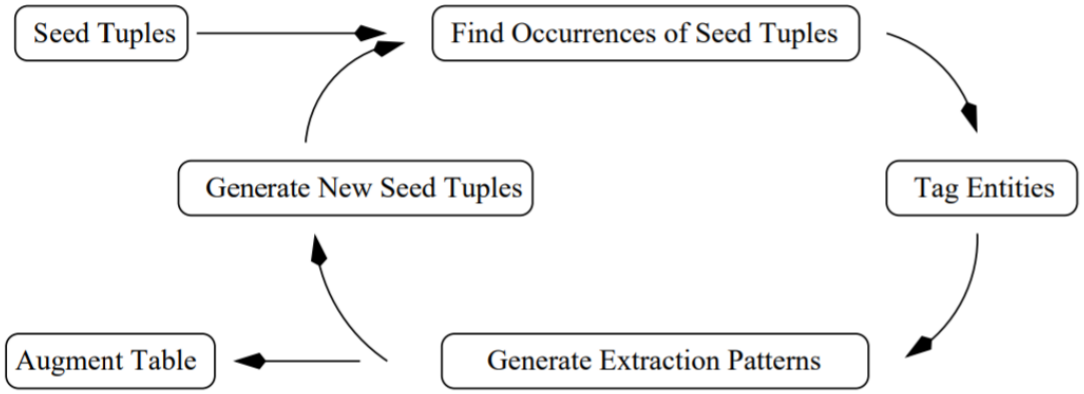
Hand written patterns for extracting triples from Wikipedia infobox.

- Syntactic patterns
  - Adding extra information while using the patterns
  - Suitable for automatically extracting
  - Have better scalability
- Semantic patterns
  - Introduce concepts to specify the pattern
  - The pattern is more accurate
  - Scalable to different realtions

**$POLITICIAN's** government of **$COUNTRY**
**$POLITICIAN was** elected president of **$COUNTRY**

# Bootstrapping Method: Snowball System

- Bootstrapping extraction
  - Start with some seed triples/patterns of the relation
  - Extraction can be performed on large unlabeled data.
- Snowball
  - Using patterns with entity tags and soft matching mechanism
  - High confidence patterns and triples are extracted through iterations




**<ORGANIZATION>**'s headquarters in **<LOCATION>**
**<LOCATION>**-based **<ORGANIZATION>**
**<ORGANIZATION>**, **<LOCATION>**

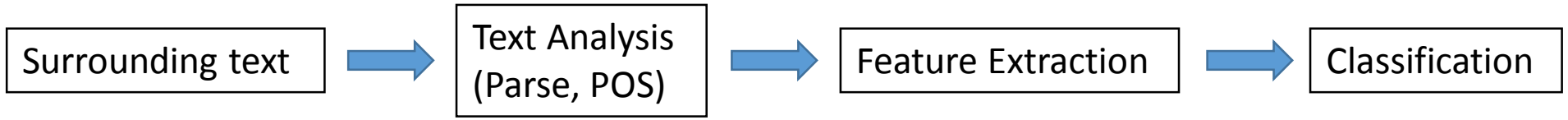
$$Conf(P) = \frac{P.positive}{(P.positive + P.negative)}$$

$$Conf(T) = 1 - \prod_{i=0}^{|P|} (1 - (Conf(P_i) \cdot Match(C_i, P_i)))$$

Agichtein, Eugene, Gravano, et al. Snowball :extracting relations from large plain-text collections[J]. 2000:85-94.

# Supervised method: Basic ideas

- Supervised Relation Extraction
  - Model the relation extraction as a classification problem
  - Entity pairs in text can be classified to different categories which corresponds to their relation.
    - i.e. Located_In, No_Relation, etc.




  - Analysis surrounding text of the entity pair.
    - POS-tag, Parse-tree ...
  - Feature extraction.
  - Use classifiers to distinguish which category (relation) the pair belongs to.
    - SVM, Naïve Bayes, Linear Regression ...

# Supervised method: Features

- Lexical Features
  - Word bags to the left, right and in the middle of the two entities.
  - POS of words in the word bags.
  - Order of the entities.
- Syntactic Features
  - Parsing tree of the text.
- Semantic Features
  - Entity type/tag, Concept ...

Astronomer **Edwin Hubble** was born in **Marshfield**, Missouri.

| Feature type | Left window | NE1 | Middle | NE2 | Right window |
|---|---|---|---|---|---|
| Lexical | [] | PER | [was/VERB born/VERB in/CLOSED] | LOC | [] |
| Lexical | [Astronomer] | PER | [was/VERB born/VERB in/CLOSED] | LOC | [,] |
| Lexical | [#PAD#, Astronomer] | PER | [was/VERB born/VERB in/CLOSED] | LOC | [, Missouri] |
| Syntactic | [] | PER | [$\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}$] | LOC | [] |
| Syntactic | [Edwin Hubble $\downarrow_{lex-mod}$] | PER | [$\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}$] | LOC | [] |
| Syntactic | [Astronomer $\downarrow_{lex-mod}$] | PER | [$\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}$] | LOC | [] |
| Syntactic | [] | PER | [$\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}$] | LOC | [$\downarrow_{lex-mod}$ ,] |
| Syntactic | [Edwin Hubble $\downarrow_{lex-mod}$] | PER | [$\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}$] | LOC | [$\downarrow_{lex-mod}$ ,] |
| Syntactic | [Astronomer $\downarrow_{lex-mod}$] | PER | [$\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}$] | LOC | [$\downarrow_{lex-mod}$ ,] |
| Syntactic | [] | PER | [$\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}$] | LOC | [$\downarrow_{inside}$ Missouri] |
| Syntactic | [Edwin Hubble $\downarrow_{lex-mod}$] | PER | [$\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}$] | LOC | [$\downarrow_{inside}$ Missouri] |
| Syntactic | [Astronomer $\downarrow_{lex-mod}$] | PER | [$\uparrow_s$ was $\downarrow_{pred}$ born $\downarrow_{mod}$ in $\downarrow_{pcomp-n}$] | LOC | [$\downarrow_{inside}$ Missouri] |

Dekang Lin. 1998. Dependency-based evaluation of minipar. In Workshop on the Evaluation of Parsing Systems.

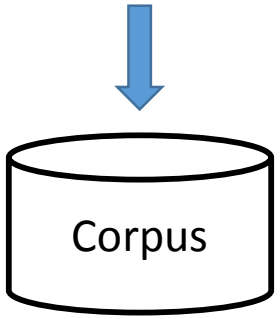
# Distant supervision method: Basic idea

- **Where to find training data for the classifiers?**
  - Existing gold-standard datasets (ACE-2005, SemEval-2010).
- **Want more data**?
  - Freebase/ DBpedia/ CN-Dbpedia ...
    - Thousands of relations and millions of triples.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. ACL, pages 1003–1011.

A triple from KB

Steven Spielberg, Directed_Work, Saving Private Ryan




Generated instances from distant supervision

**[Steven Spielberg]**'s film **[Saving Private Ryan]** is loosely based on the brothers' story.

Allison co-produced the Academy Award winning **[Saving Private Ryan]**, directed by **[Steven Spielberg]**.

...

**Assumption of Distant supervision:**

**If two entities participate in a relation, any sentence that contain those two entities might express that relation.**

# Distant supervision method: Attention-based procedure

Freebase

| Relation | Entity1 | Entity2 |
|---|---|---|
| /business/company/founders | Apple | Steve Jobs |
| … | … | … |

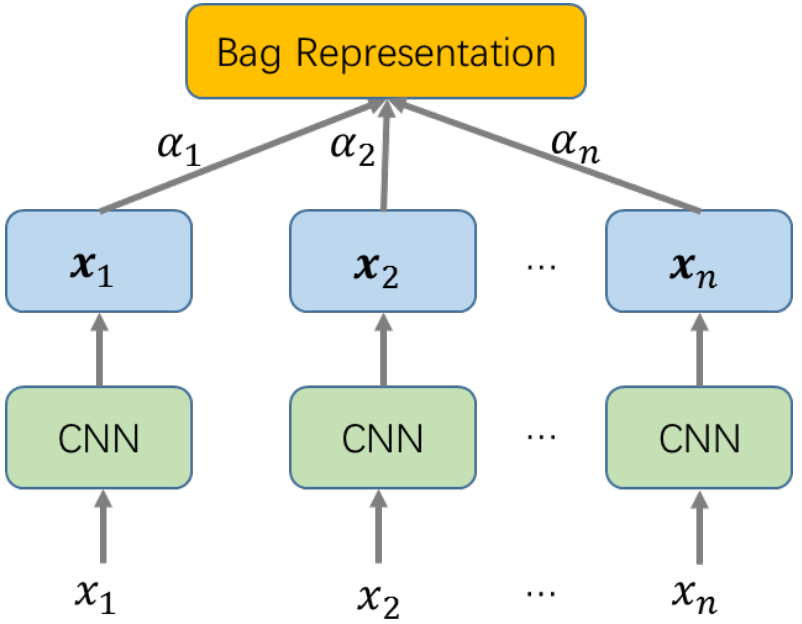
Mentions from free texts

1. Steve Jobs was the co-founder and CEO of Apple and formerly Pixar.
2. Steve Jobs passed away the day before Apple unveiled iPhone 4S in late 2011.

Noise are generated through distant supervision.

*How to get rid of these wrong instances?*

- Filter out the wrong instances before using them.
- Use models tolerating the noise




For multiple sentences containing the same entity pair, the weight of each sentence is judged through a attention-based network before classification.

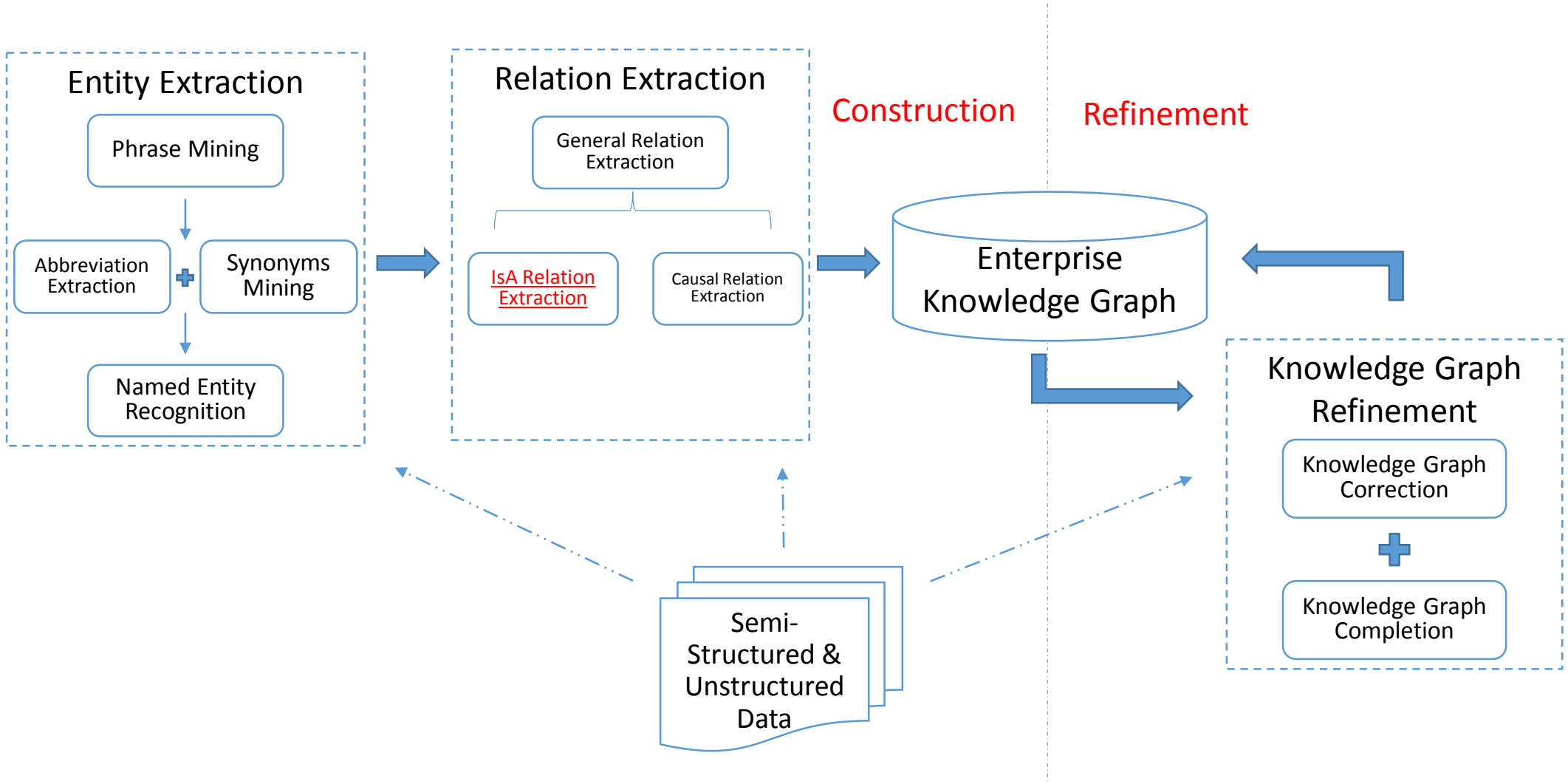
$$e_i = x_i A r$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_k \exp(e_k)}$$

$$s = \sum_k \alpha_i x_i$$


Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016, 1: 2124-2133.

# Workflow of EKG construction from corpus



# IsA Relation Extraction (Taxonomy construction)

## Pattern-based Methods

- High precision and coverage in extracting isA relations
- Probase :the representative pattern-based taxonomy: contains millions of entities and concepts, making the biggest conceptual taxonomy so far.

## Wikipedia-based Methods

- High precision but limited recall
- The precision of the English taxonomy YAGO and Chinese taxonomy CN-Probase is 95%

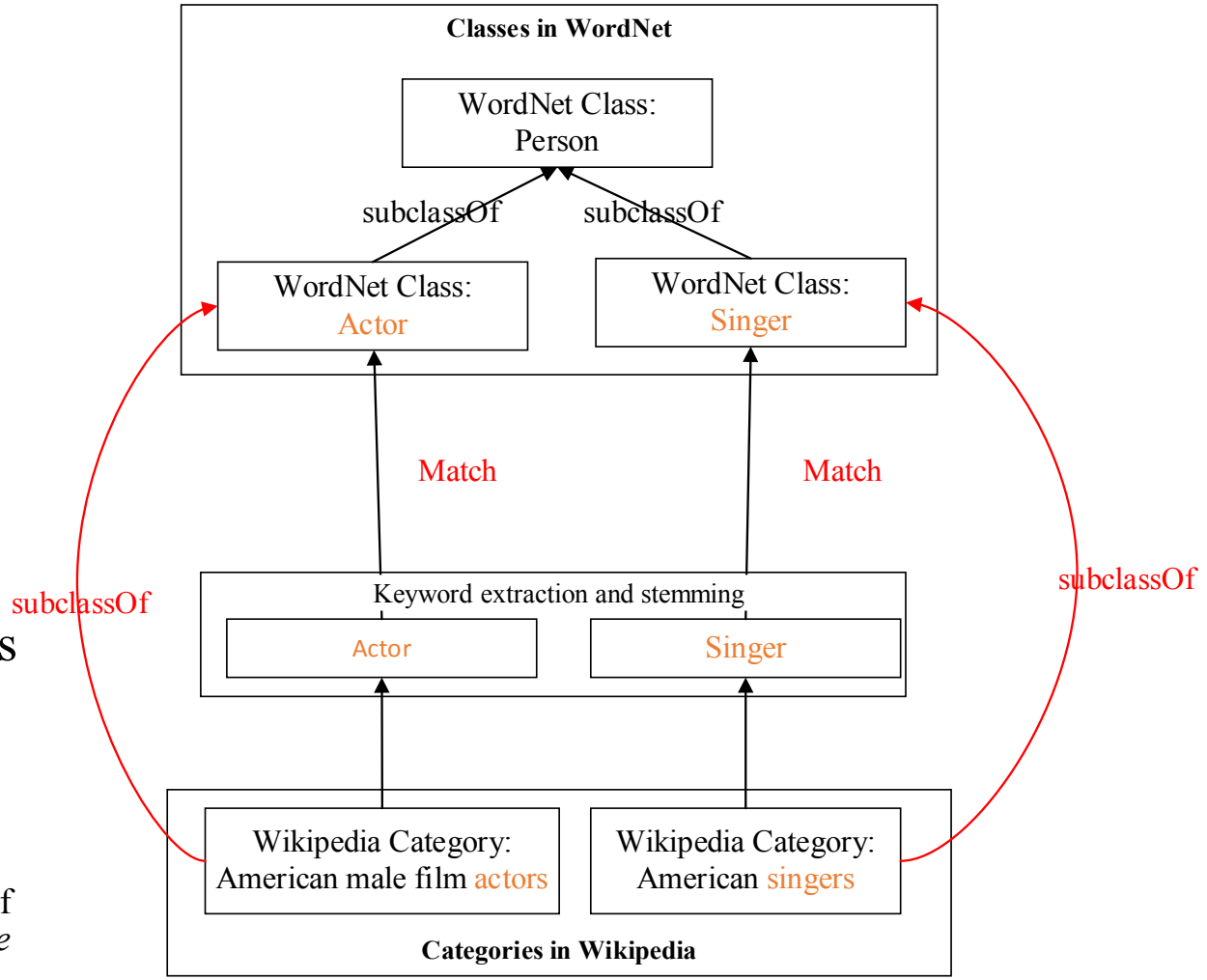
## Embedding-based Methods

- Embedding-based methods suffer from low precision (around 80%)
- These methods cannot directly used for isA relation inference.

# Wikipedia-based Methods : YAGO

- YAGO is a traditional Wikipedia-based conceptual taxonomy in English
  - Based on the category of Wikipedia
  - Contains 360 thousand isA relations，with a precision of 95%

- Method:
  - Use WordNet as the basic taxonomy
  - Add more categories from Wikipedia into this taxonomy
    - Extract and stem keywords from categories of Wikipedia and match them with WordNet Class.




Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: a core of semantic knowledge. *Proceedings of the 16th international conference on World Wide Web*, (pp. 697-706).

# Pattern-based Methods: Hearst Patterns

- **Hearst Patterns are EFFECTIVE in extracting high quality isA relations from English corpus.**
- Probase is a pattern-based taxonomy, which contains millions of entities and concepts, making thus biggest conceptual taxonomy.

| ID | Pattern |
|---|---|
| 1 | *NP* such as {*NP*,}*{(or \| and)} *NP* |
| 2 | such *NP* as {*NP*,}*{(or \| and)} *NP* |
| 3 | *NP*{,} including {*NP*,}*{(or \| and)} *NP* |
| 4 | *NP*{,*NP*}*{,} and other *NP* |
| 5 | *NP*{,*NP*}*{,} or other *NP* |
| 6 | *NP*{,} especially {*NP*,}*{(or \| and)} *NP* |

Examples of Hearst patterns

- … {animals} other than dogs **such as** *{cats}* …
- … {classic movies} **such as** *{Gone with the Wind}* …
- … {companies} **such as** *{IBM, Nokia, Proctor and Gamble}*
- … representatives in *{North America, Europe, the Middle East, Australia, Brazil, Japan, China}*, **and other** {countries} ...

*cat isA animal*
*cat isA dog*
*Gone with the Wind isA classic movie*

Examples of finding isA relations from articles with the Hearst patterns

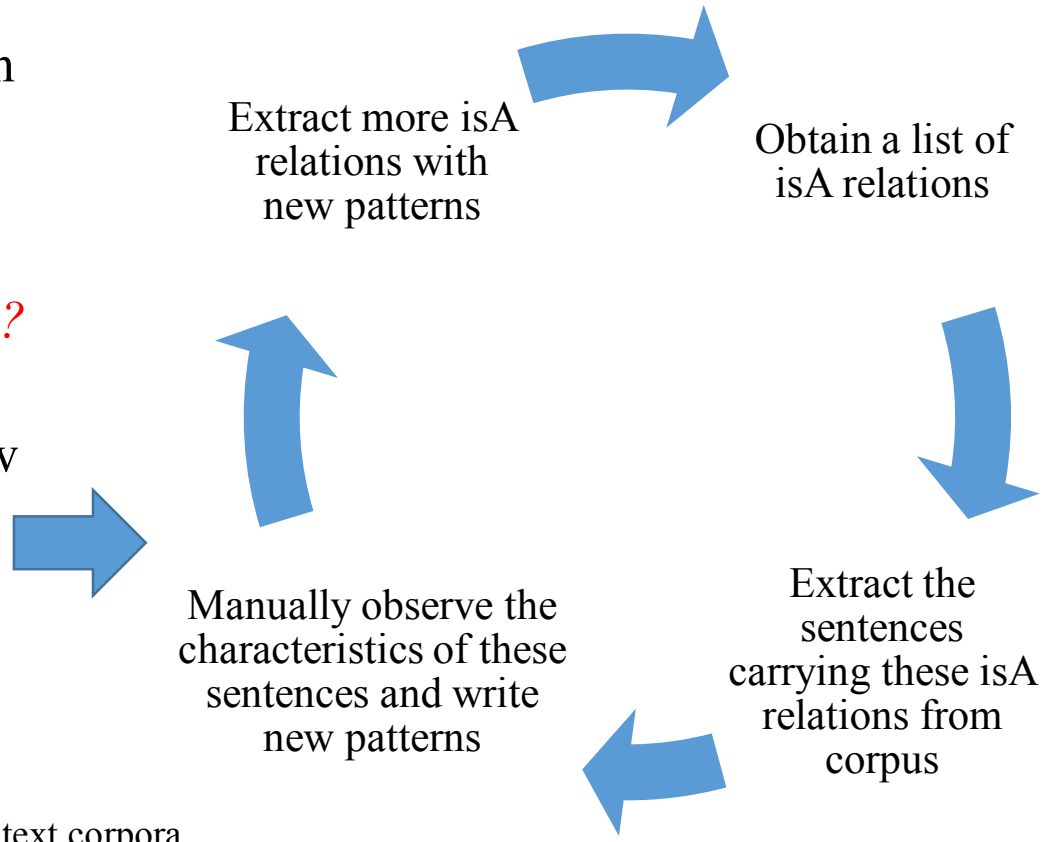
Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012). Probase: A probabilistic taxonomy for text understanding. *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, (pp. 481-492).

# Pattern-based Methods: Hearst Patterns

The extraction recall limits by the scale of high quality Hearst patterns.

*How to find more high quality Hearst patterns?*

- We first can ask experts to write down a few Hearst Patterns
- Then we can generate more Hearst Pattern via Bootstrapping methods.




The Bootstrapping process

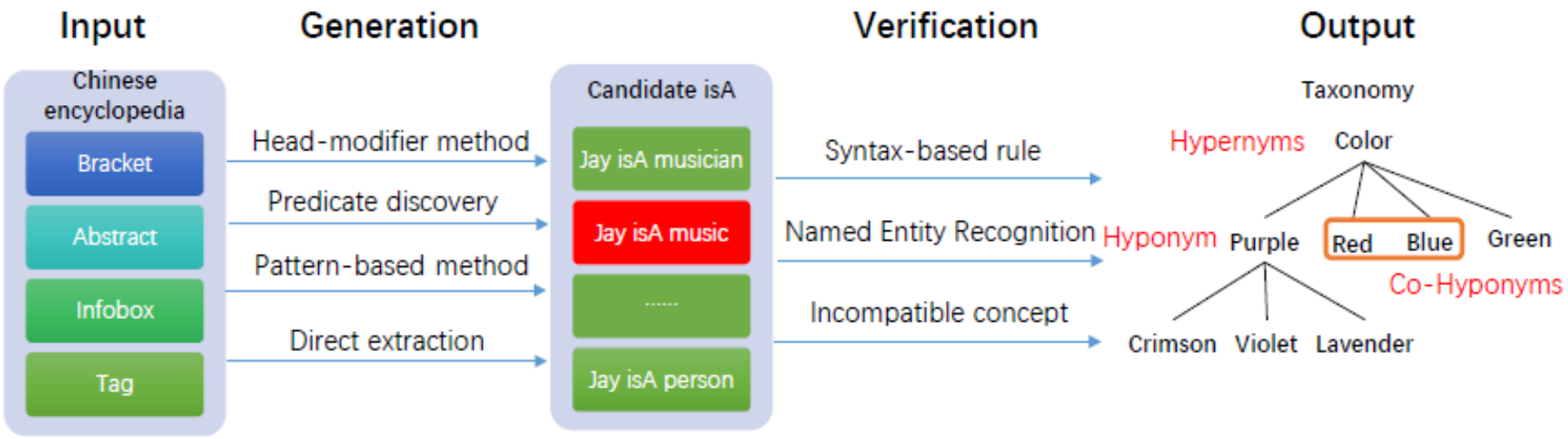
Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th conference on Computational linguistics-Volume 2*, (pp. 539-545).

# IsA Relation Extraction: Hybrid Methods

How about constructing taxonomies with Hearst Patterns in other language ?

We use the pattern "NP such as {NP,}" to extract isA relations and have a precision of 92% in English. But in Chinese, the precision is only 75%

--- Patterns may not perform well in other languages!




Generation and Validation Process in CN-Probase

CN-Probase: A Data-driven Approach for Large-scale Chinese Taxonomy Construction , in ICDE2019

# Chinese isA relation extraction and validation

- **Extraction**: Extracting isA relations from multi data sources to obtain a high coverage


- **Validation**: Performing validation and filtering to obtain a high precision
  - Mutex concepts identification

The Similarity of Entity Set

$$\text{MiniJaccard}(c_1,c_2) = \frac{|c_1 \cap c_2|}{\min(|c_1|,|c_2|)}$$

The Similarity of Entity Property Distribution

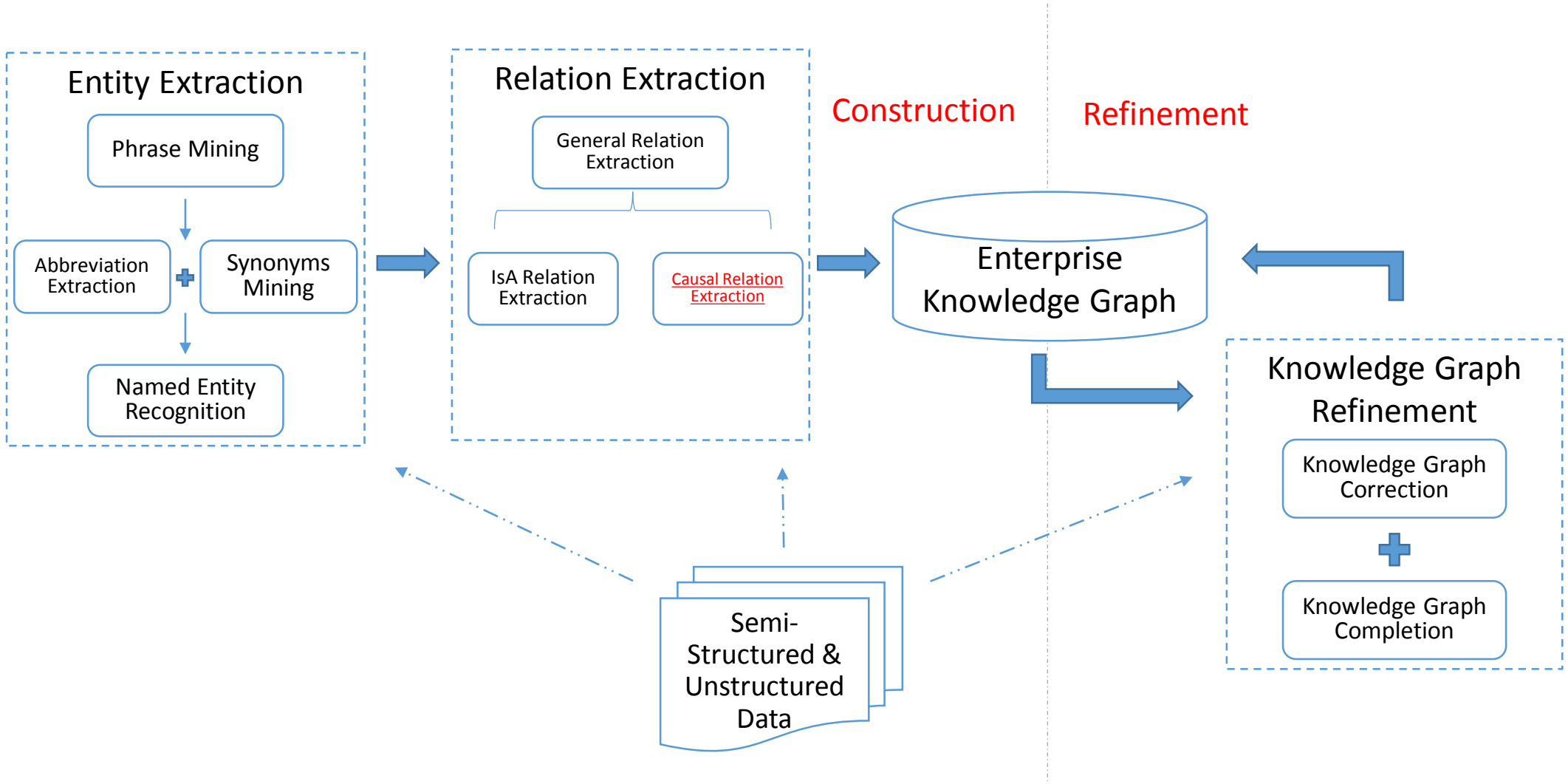
$$\text{CPD}(c_1,c_2) = \frac{x \cdot y}{||x|| \cdot ||y||}$$

Conceptual compatibility

$$\text{P}(c_1,c_2) = \frac{2 * MiniJaccard(c_1,c_2) * CPD(c_1,c_2)}{MiniJaccard(c_1,c_2) + CPD(c_1,c_2)}$$

CN-Probase: A Data-driven Approach for Large-scale Chinese Taxonomy Construction， in ICDE2019

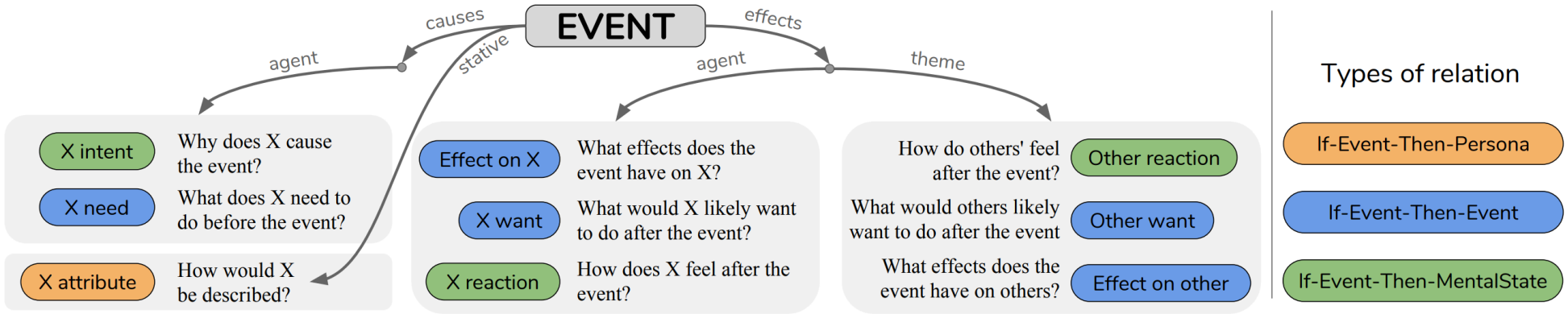
# Workflow of EKG construction from corpus



# If-then knowledge graph

- Definition of if-then knowledge
  - Taking ‘X compliments Y’（an event）as an example

| | | |
|---|---|---|
| If-event-then-MentalState | likely intents of the event | X wanted to be nice |
| | likely (emotional) reactions of the event‘s subject | X will fell good |
| | likely (emotional) reactions of others | Y feel flattered |
| If-event-then-event | pre-conditions of a given event | X want to chat with Y |
| | Post-conditions of a given event | Y will smile |
| If-event-then-persona | how the subject is described or perceived | X is caring |




Sap M, Le Bras R, Allaway E, et al. ATOMIC: an atlas of machine commonsense for if-then reasoning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 3027-3035.

# If-then knowledge graph

- Construction
  - Crowdsourcing: Amazon Mechanical Turk
- Application: reasoning
  - A multi-task encoder-decoder model
  - Example:

`PersonX bakes bread`

*Before, X needed to*

buy ingredients
go to the store
gather ingredients
mix ingredients
turn on oven
turn on stove

*As a result, X will*

salivate
get dirty
eat
get messy
get full
eat food

# Script knowledge graph

- Definition
  - Event: (predicate, subject, object)
  - Graph:
    - Relation:
      - COREF_NEXT：sequential relationships
      - NEXT: events co-occur in a window
      - DISCOURSE_NEXT: see table
    - Node:
      - events

| Type of relationships | Example |
|---|---|
| Comparison.Contrast | X is outgoing -- Y is introverted |
| Contigency.Cause.Reason | X went to a restaurant -- X was hungry |
| Contingency.Cause.Result | X had some time to kill -- X went to a restaurant |
| Contingency.Condition | The restaurant serves drinks – X ordered a drink |
| Expansion.Instantiation | Most mammals are viviparous – Blue whales are viviparous |
| Temporal.Synchrony | X is listening to music – X is reading a book |
| Temporal.Asynchronous | X first turns on the oven –X then puts in the material |

Relationships of DISCOURSE_NEXT

Lee I T, Goldwasser D. Multi-Relational Script Learning for Discourse Relations[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4214-4226.

# Script knowledge graph

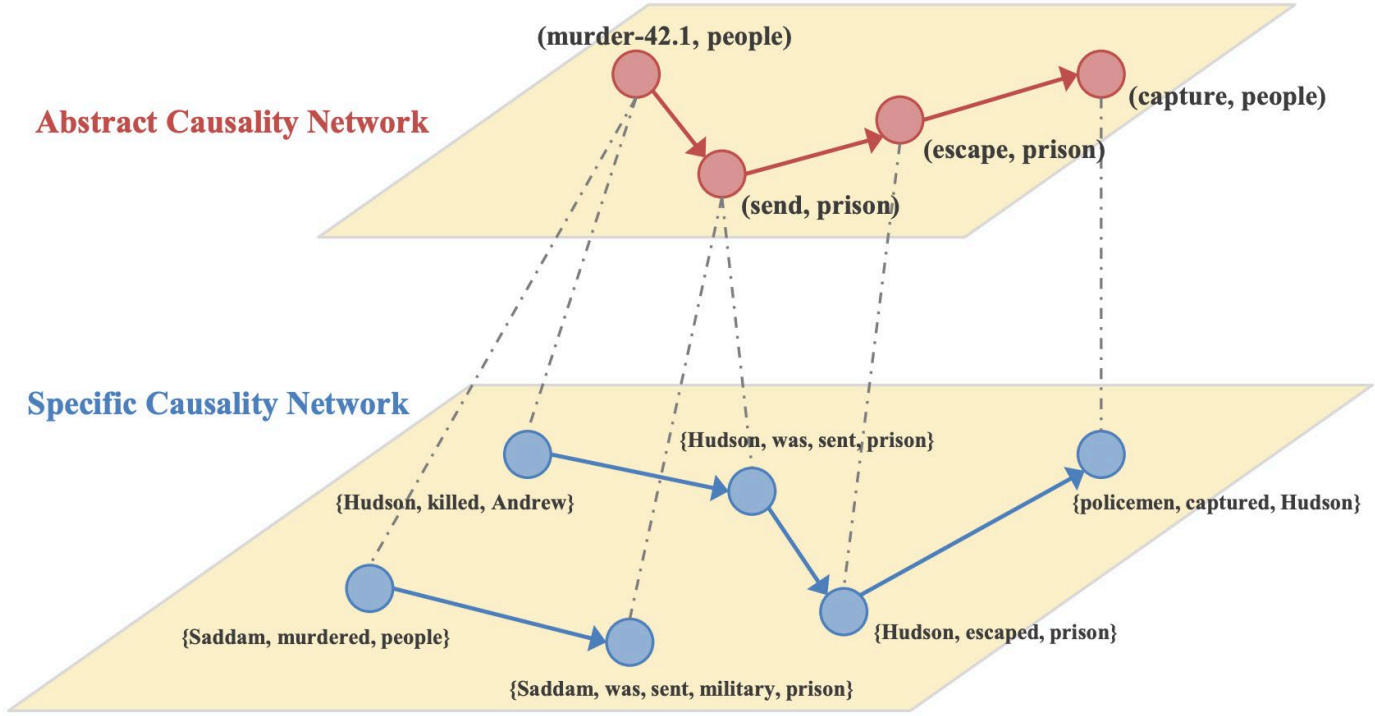
- Construction
  - Step-1: extract events:
    - E.g. *'X went into her favorite restaurant'->* (go_into, Jenny, her favorite restaurant)
  - Step-2: use the connectives to define the relationships between events
    - E.g. *'X went to a restaurant, because she was hungry.'-> 'because'implies the relationship is 'Contigency.Cause.Reason'*
- Application: reasoning
  - E.g. Given 'X was hungry', predict the result:'X go to a restaurant'
  - Design two translation-based embedding models named EventTransE and EventTransR.

$$f_{transe}(t) = f_{transe}((e_h, e_t, r)) = \|e_h + r - e_t\|_p^p,$$

$$f_{transr}(t) = f_{transr}((e_h, e_t, r)) = \|e_h M_r + r - e_t M_r\|_p^p,$$

# Abstract event causality network

- Definition
  - Nodes in the specific causality network:
    - Nouns and verbs in the order in which they appear
  - Nodes in the abstracted causality network
    - Nouns and verbs are generalized based on WordNet and VerbNet.
    - E.g.
      - Chips->dishes
      - Kill->murder-42.1




Zhao S, Wang Q, Massung S, et al. Constructing and embedding abstract event causality networks from text snippets[C]//Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. ACM, 2017: 335-344.

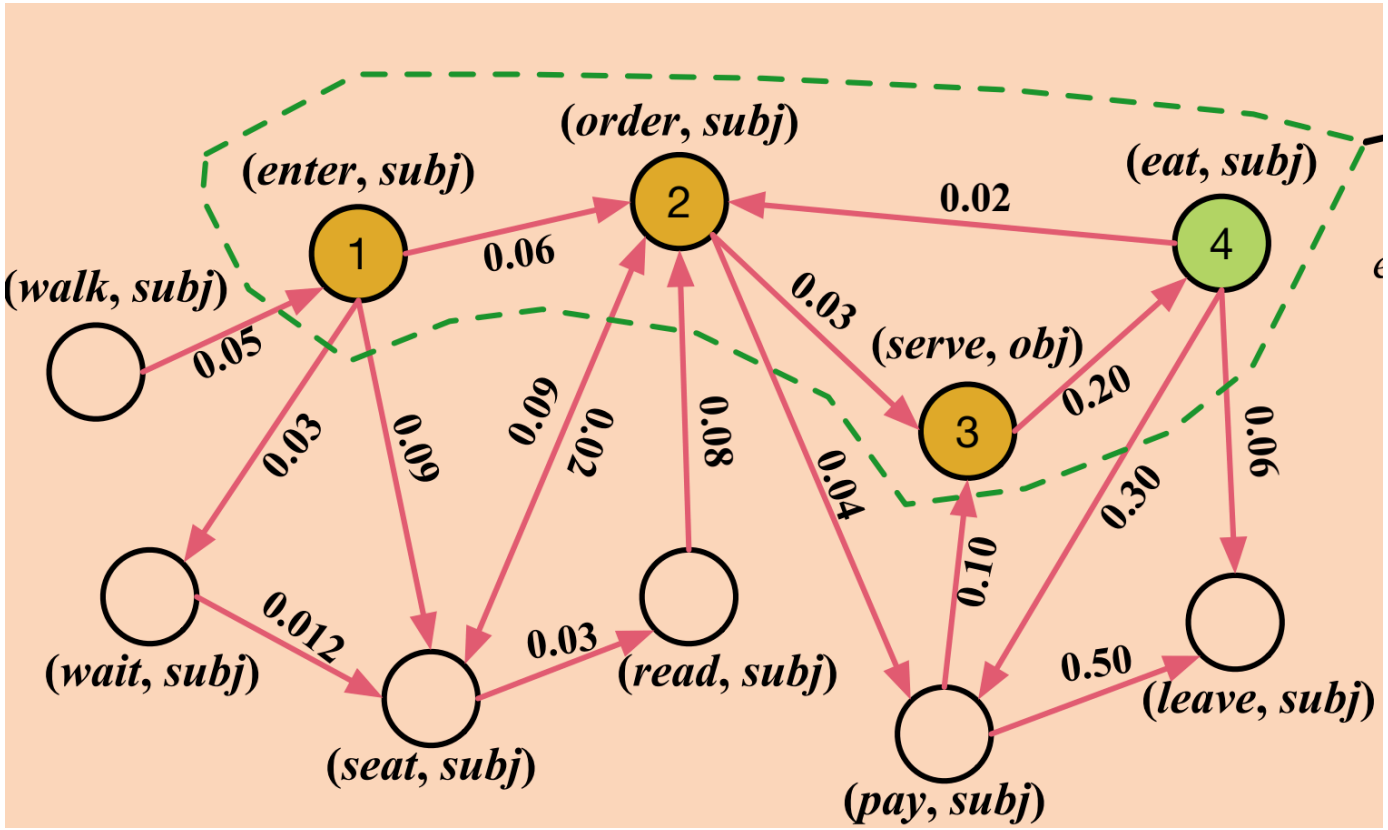
# Abstract event causality network

- Building
  - E.g. Hudson was sent to the prison, because Hudson killed Andrew.
  - Causality mention extraction:
    - Cause event: Hudson killed Andrew. Effect event: Hudson was sent to the prison.
  - Causal event extraction
    - Event 1: {Hudson, killed, Andrew}. Event 2: {Hudson, was, sent, prison}
  - Causal event generalization
    - Event 1: {murder-42.1, people}. Event 2: {sent, prison}
- Application: reasoning
  - Given 'Hudson killed Andrew.', predict the effect:'Hudson was sent to the prison'
  - Given 'Hudson was sent to the prison.', predict the reason:'Hudson killed Andrew'
  - Design a network embedding model:

$$f(c, e) = \|\mathbf{c} + \mathbf{t} - \mathbf{e}\|_1 + \|\mathbf{e} + \tau - \mathbf{c}\|_1$$

# Event knowledge graph with probabilities on edges

- Definition
  - Events in a event chain share the same protagonist entity.
  - Each node contains the predicate and the subject of the event
  - Each edge is directed and has a weight.




Li Z, Ding X, Liu T. Constructing narrative event evolutionary graph for script event prediction[C]//IJCAI. 2018: 4201-4207

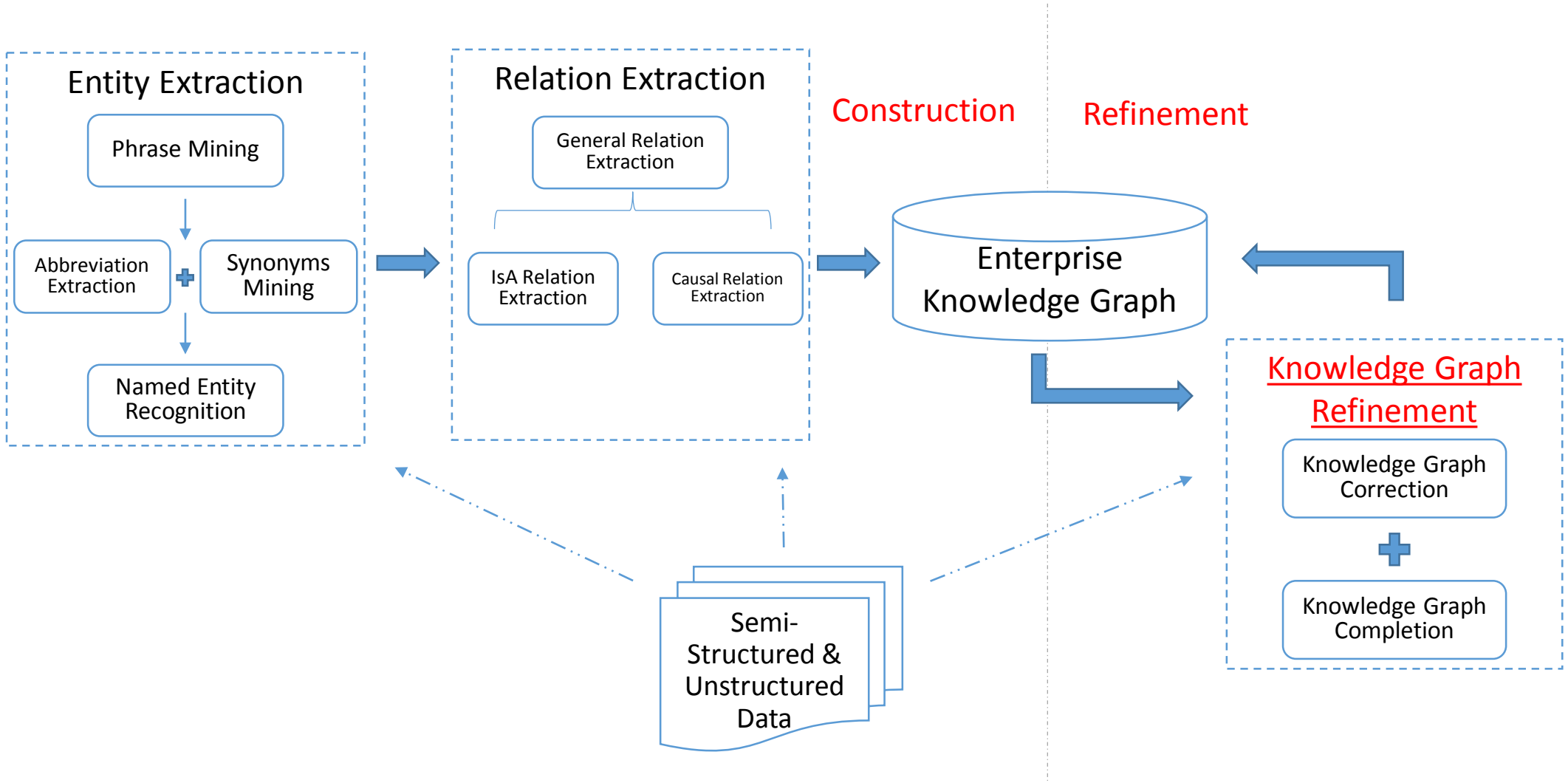
# Event knowledge graph with probabilities on edges

- Building
  - Step 1-Extract event chains
  - Step 2-Construct the graph
    - Each directed edge is assigned with a weight

$$w(\mathbf{v}_j|\mathbf{v}_i) = \frac{\text{count}(\mathbf{v}_i, \mathbf{v}_j)}{\sum_k \text{count}(\mathbf{v}_i, \mathbf{v}_k)}$$

- Application: reasoning
  - A graph neural network：
    - To build a scaled graph neural network, this model only encodes a subgraph with context and candidate event nodes at each step.
  - Use a Siamese network to predict the sequential events.

# Workflow of EKG construction from corpus



# Quality Issues in Knowledge Graph

- **Wrong facts: Cleaning**
  - Attribute might be wrong
  - Attribute values might be wrong
  - Inter-entity relationships might be wrong

- **Missing facts: Completion**
  - Many facts are missing due to bias in corpus
  - Many facts are commonsense, rarely mentioned in corpus

- **Expired facts: Update**
  - As time goes by, the facts is changing
    - Always changing : population, age, position, President of the United States …
    - Constantly added : new employee, new company, new words…

| Name | University | City | Country |
|---|---|---|---|
| San Zhang | Soochow Uni. | Soochow | China |
| Si Li | Soochow Uni. | Hefei | China |
| Er Wang | Soochow Uni. | New York | USA |

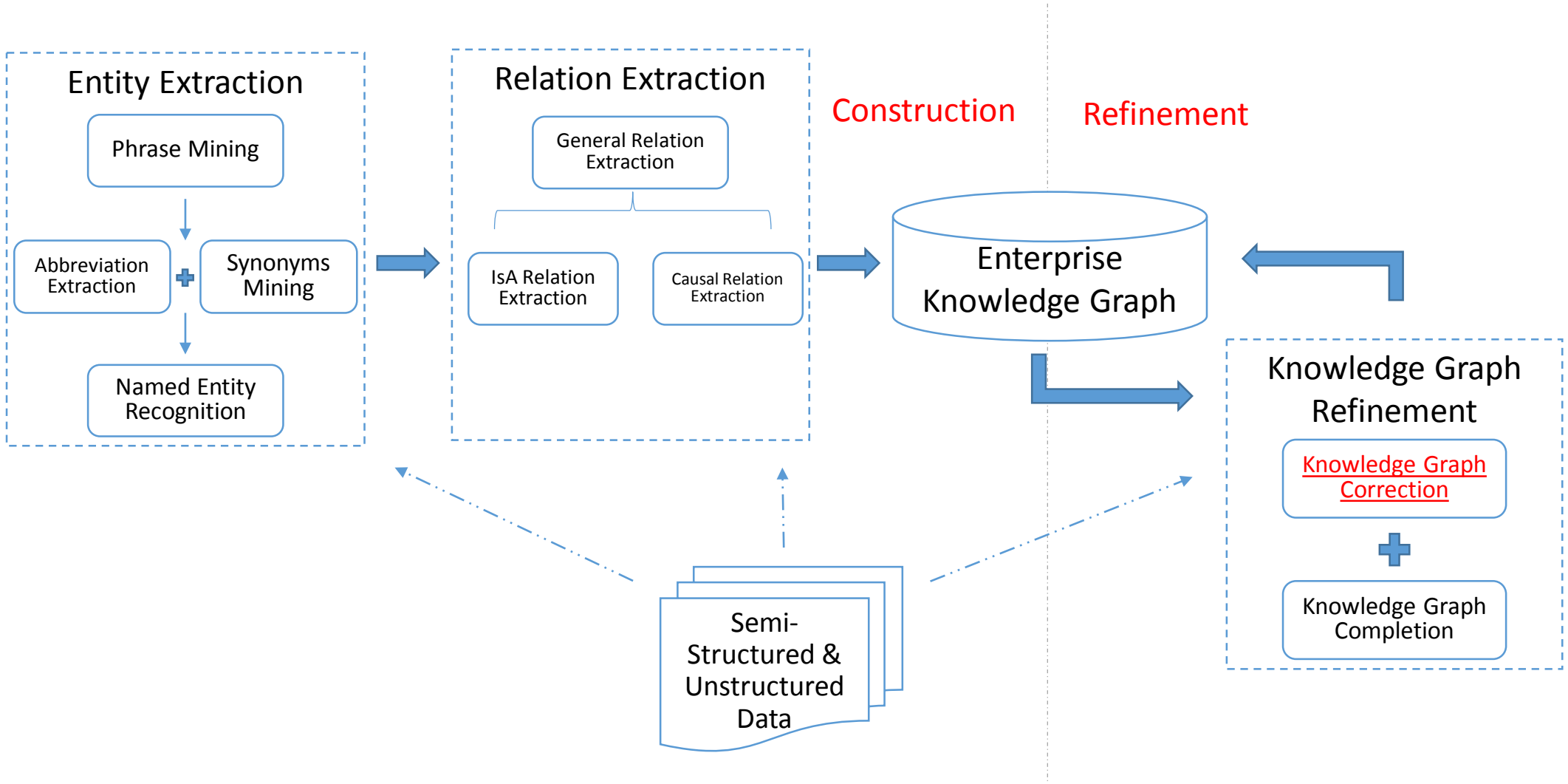
# Expired Data Update Mechanism

- Regular Global Update
  - Disadvantages: time-consuming and labor-intensive
- Based on Update Frequency Prediction
  - Disadvantages: time-consuming, inaccurate, unable to add new words
- **Based on Hot Event Discovery**
  - Basic Idea：
    - Monitor hot words on the Internet. Two reasons why an entity becomes a hot word：
      - New word, such as the upcoming iPhone8.
      - Old words with knowledge changed, *e.g.*, Trump became the president of the United States.
  - Overall Framework：


- Find hot entities from Web.
- Synchronize (update or insert) these seed entities.
- Find more related entities via the hyperlinks in the latest pages.
- Synchronize these expanded entities according to priority, which is provided by a predictor.

*Ref: Liang et. al., How to Keep a Knowledge Base Synchronized with Its Encyclopedia Source，IJCAI2017*

# Workflow of EKG construction from corpus



# Common Ideas of Data Correction

- Quality Rules/Constraints:
  - FDs/CFDs: *e.g., "(University -> City, 0.98)"*
  - Identify inconsistent data in Conflicts, choose a minimum change
  - User-Defined Quality Rules
    - *e.g., change all capital into lowercase*
  - Inconsistency Data != Erroneous Data
  - Minimum Change != Correct Change
- Machine Learning Models:
  - Learn models with the existing data
  - Can NOT guarantee correctness
- Crowdsourcing:
  - Let the Crowd help clean the data
    - *e.g., find errors, fill blanks, make choices between conflicts*
  - Reach a much higher precision/recall
  - Expensive!
- Based on Hybrid:
  - Rule-based+Crowd, Model-based+Crowd, ...

# Data Correction in Knowledge Graph

- Reasoning to find and correct errors in the knowledge graph
  - Knowledge reasoning technology is still preliminary
  - Reasoning can only find limited errors
- Fact validation with external evidence from the Internet
  - Need data support from Web
- With crowdsourcing
  - The cost of crowdsourcing need optimizing, otherwise the cost is too high

# Error Correction with type constraint

- Reasoning on the validity of fact candidates**, a rich ontology** is required
  - Leverage knowledge about entities and their types, as type information is extremely beneficial in early pruning
  - Pre-existing knowledge like YAGO ontology can be applied with statistical evidence
- Example
  - Rules: the first argument of the *teamWonTrophy* relation must be of type *sportsTeam, or businessEnterprise* should not be of type *sportsTeam*
  - Result: immediately rules out the hypotheses that Google has won the FIFA World Cup

Nakashole, Ndapandula, Martin Theobald, and Gerhard Weikum. "Scalable knowledge harvesting with high precision and high recall." Proceedings of the fourth ACM international conference on Web search and data mining. ACM, 2011.

# Error Detection using Aggregate Type Statistics

**Feature Vector Creation:**

1. Using all direct types. A binary feature is created for each schema class, which is set to true for a link if the linked resource has the class defined as its **rdf:type**.

2. Using all ingoing and outgoing properties. Two binary features are created for each link. Each dimension in each feature vector denotes a property, which is set to true if the linked resource has this property.

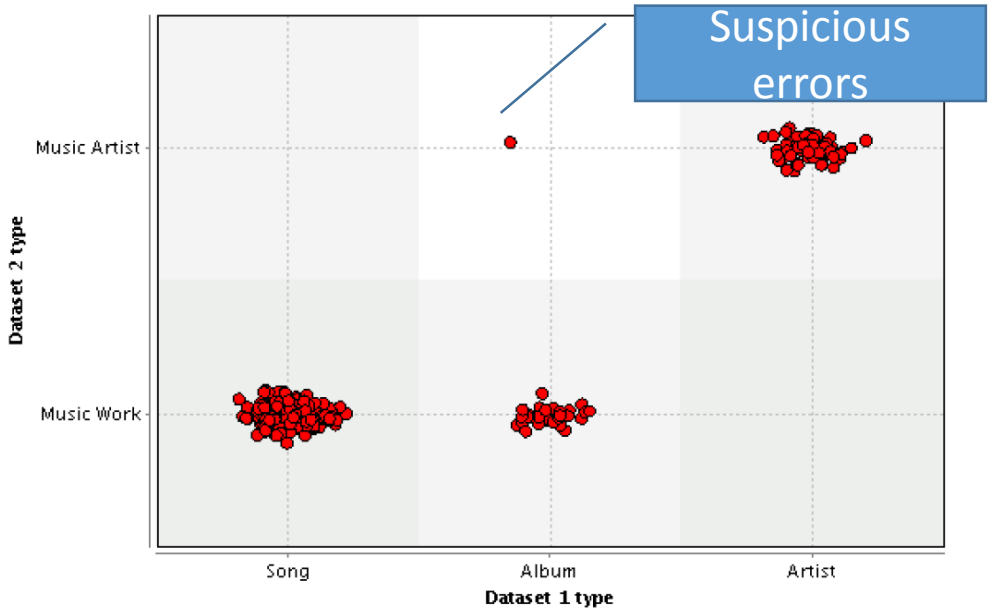



**Fig. 1.** A simplified example of links being represented in a vector space. The single dot in the upper middle quadrant represents a wrong link.

Triple 1: (Avril Lavigne, singer, Girlfriend) [Music Artist, Song]
Triple 2: (Taylor Swift, singer, Love Story) [Music Artist, Song]
Triple 3: (Owl City, singer, Fireflies) [Music Artist, Song]
Triple 4: (Rihanna, singer, Titanic) [Music Artist, film]

Paulheim H. Identifying Wrong Links between Datasets by Multi-dimensional Outlier Detection[C]//WoDOOM. 2014: 27-38.

# Error Detection for Numerical Rules

Numerical attributes are very important. How to detect the error values for these attributes?

<dbres:Roberto_Penna> <rdf:type> <dbonto:Athlete>.
<dbres:Roberto_Penna> <foaf:name> "Roberto Penna".
<dbres:Roberto_Penna> <dbonto:birthDate> "1986-04-19".
<dbres:Roberto_Penna> <dbonto:deathDate> "1910-07-05".

<dbres:Allan_Dwan> <rdf:type> <dbonto:Person>.
<dbres:Allan_Dwan> <foaf:name> "Allan Dwan".
<dbres:Allan_Dwan> <dbonto:birthDate> "1885-04-03".
<dbres:Allan_Dwan> <dbonto:deathDate> "1981-12-28".

(a) An example of *birthDate* and *deathDate* attributes

Birthdate should be earlier than deathdate

popRural+popUrban=popTotal

<dbres:Jaipur> <rdf:type> <dbonto:City>.
<dbres:Jaipur> <foaf:name> "Jaipur".
<dbres:Jaipur> <dbonto:populationRural> "3164767".
<dbres:Jaipur> <dbonto:populationUrban> "3499204".
<dbres:Jaipur> <dbonto:populationTotal> "3073350".

<dbres:Callan,_County_Kilkenny> <rdf:type> <dbonto:Town>.
<dbres:Callan,_County_Kilkenny> <foaf:name> "Callan".
<dbres:Callan,_County_Kilkenny> <dbonto:populationRural> "281".
<dbres:Callan,_County_Kilkenny> <dbonto:populationUrban> "1489".
<dbres:Callan,_County_Kilkenny> <dbonto:populationTotal> "1771".

(b) An example of *population* attributes

**Challenges:**

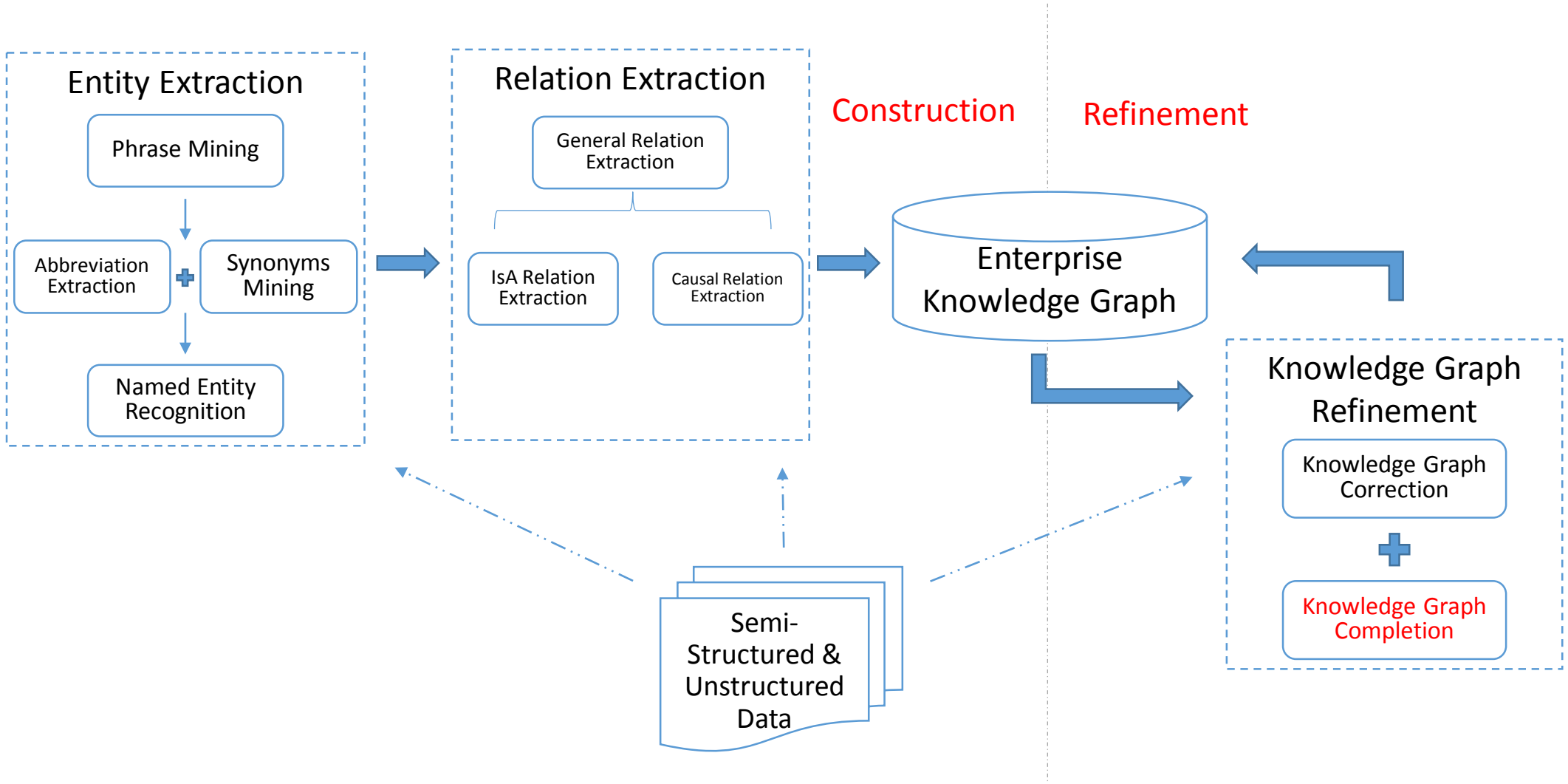
1. Detecting errors in numerical attributes has to consider the sematic relation between attributes.
2. How to employ an unsupervised error detection method that does not need labeled data?

**Solution:**

1. Consider the larger than, less than, and equal to relations between two or more attributes.
2. Bayes theorem-based probabilistic framework to detect errors in numerical data.

Li H, Li Y, Xu F, et al. Probabilistic error detecting in numerical linked data[C]//Database and Expert Systems Applications. Springer, Cham, 2015: 61-75.

# Workflow of EKG construction from corpus



# Knowledge base completion

## Entity Type Imputation (Type Completion)

- Internal Knowledge-based
  - inferring with *probability graph model* (*SDType*)
  - Using topic modeling for type prediction
  - Training a Classification Model
- External Knowledge-based
  - Classifier based on Wiki Links
  - Crowdsourcing

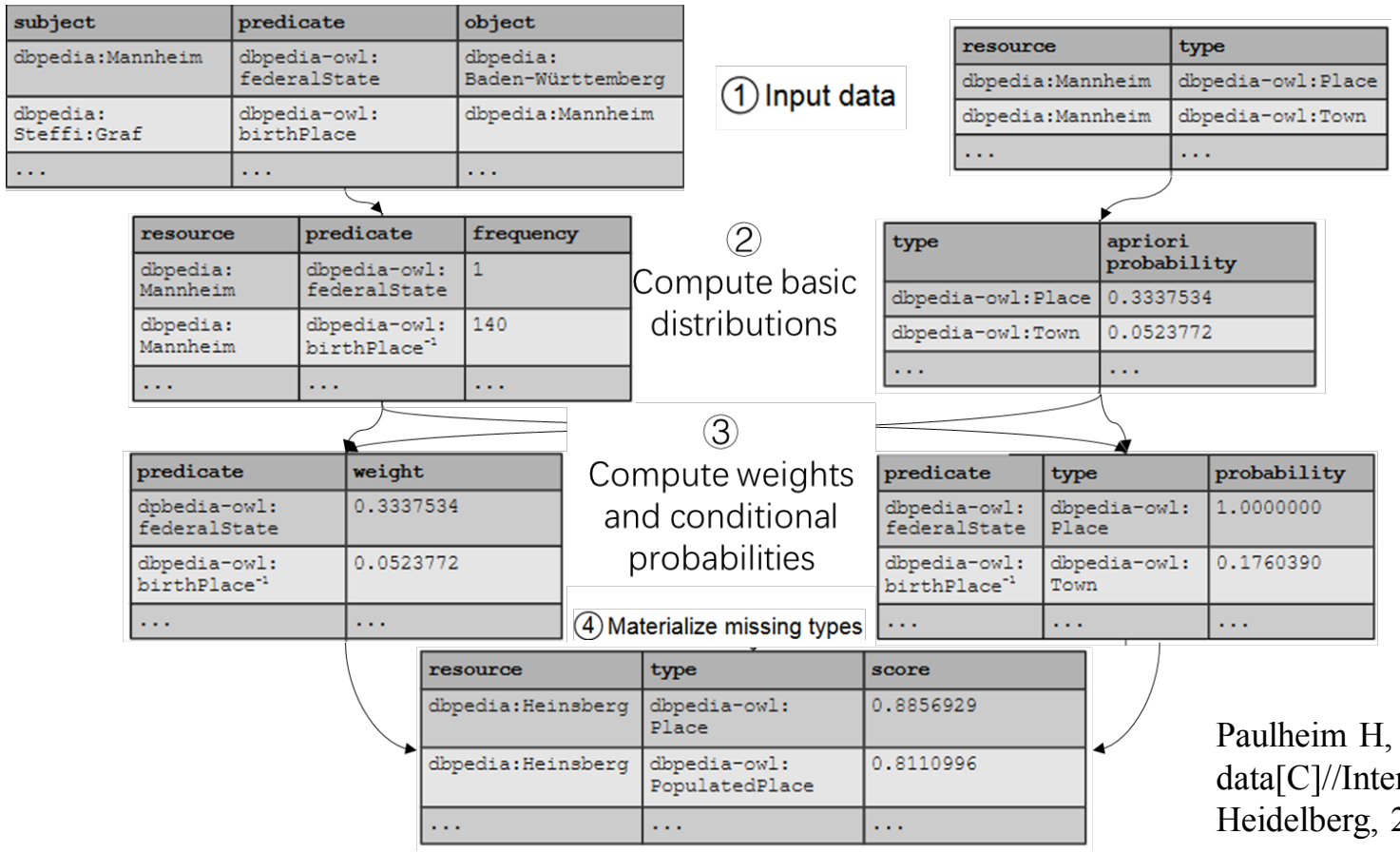
## Relation Prediction (Triple Completion )

- Internal Knowledge-based
  - inferring with *probability graph model*
  - *Path Ranking* algorithm
  - Inferring with representation model (*TransE*)
- External Knowledge-based
  - *Distant supervision* with a large text corpora
  - Based on web search engines
  - Based on other KGs

## Entity Attribute Imputation (Attribute Completion)

- Implemented by finding Determining Obligatory Attributes in Knowledge Bases

# Type completion: *SDType*

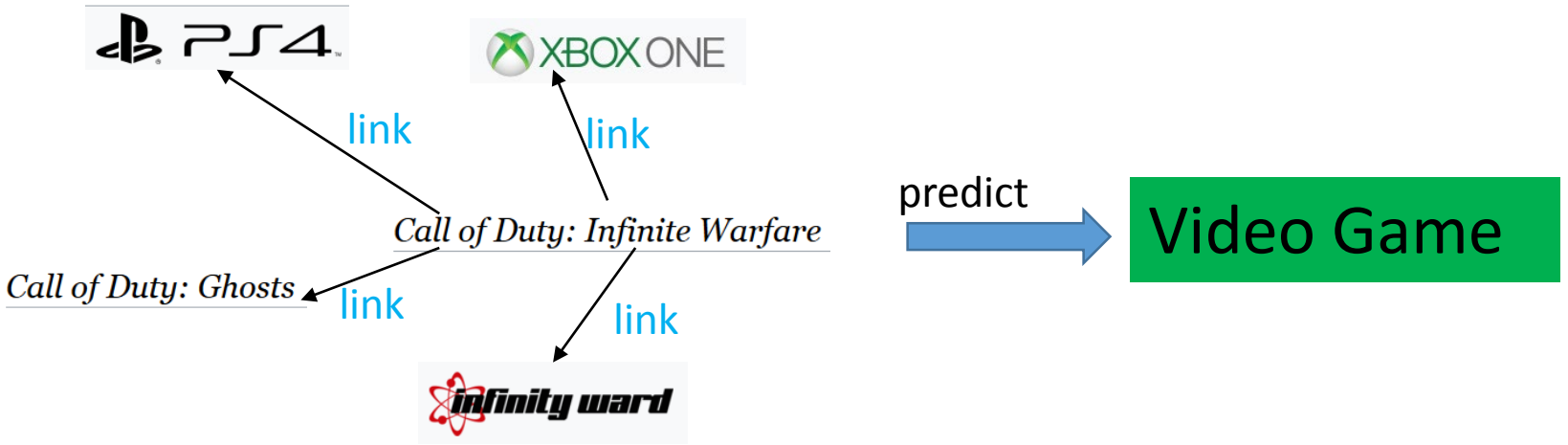



**SDType**: using **Statistical Distribution of types** in the subject and object positions for predicting the instance's types.

Paulheim H, Bizer C. Type inference on noisy rdf data[C]//International semantic web conference. Springer, Berlin, Heidelberg, 2013: 510-525.

# Type completion with Wiki links

- Use Wikipedia link graph to predict entity types in a KG
- Interlinks between Wikipedia pages are exploited to create feature vectors, e.g., based on the categories of the related pages.




Type prediction using entity Wikipedia links

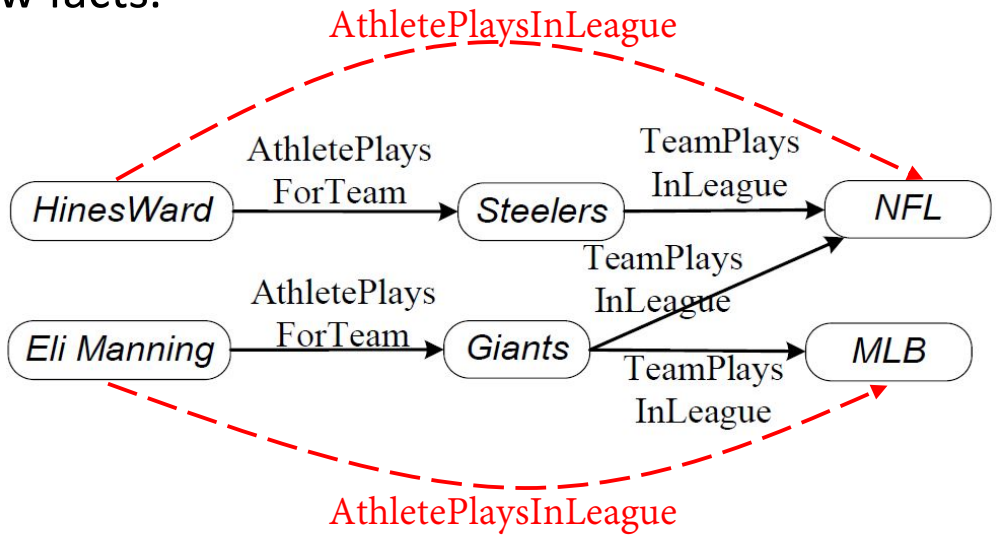
Suzuki M , Matsuda K , Sekine S , et al. A Joint Neural Model for Fine-Grained Named Entity Classification of Wikipedia Articles[J]. Ieice Transactions on Information & Systems, 2018, 101(1):73-81.

# Triple Completion: Rule-based Method

Basic idea： Applying first order horn clause rules to infer new facts.

Horn Clause:

$$AthletePlaysForTeam(a,b)$$
$$\wedge \quad TeamPlaysInLeague(b,c)$$
$$\Rightarrow \quad AthletePlaysInLeague(a,c)$$




- Advantages:
  - Rich and expressive rules
  - Good results
- Disadvantages:
  - Learning first-order Horn clauses is computationally expensive

# Triple Completion: PRA-based Method

The key idea of Path Ranking Algorithm (PRA) is to explicitly use paths connecting two entities to predict potential relations between them.

A path represents ranking "experts", which corresponds to the Horn clause.

Horn Clause:

$$\begin{aligned} & AthletePlaysForTeam(a,b) \\ \wedge\ & TeamPlaysInLeague(b,c) \\ \Rightarrow\ & AthletePlaysInLeague(a,c) \end{aligned}$$

$$\begin{aligned} & isa(a,c) \wedge isa^{-1}(c,a') \\ \wedge\ & AthletePlaysInLeague(a',b) \\ \Rightarrow\ & AthletePlaysInLeague(a,b) \end{aligned}$$

These paths are used as features in a binary classifier to predict if entities pairs have the given relation:

$$score(e;s) = \sum_{P \in \mathbf{P}_\ell} h_{s,P}(e)\theta_P$$

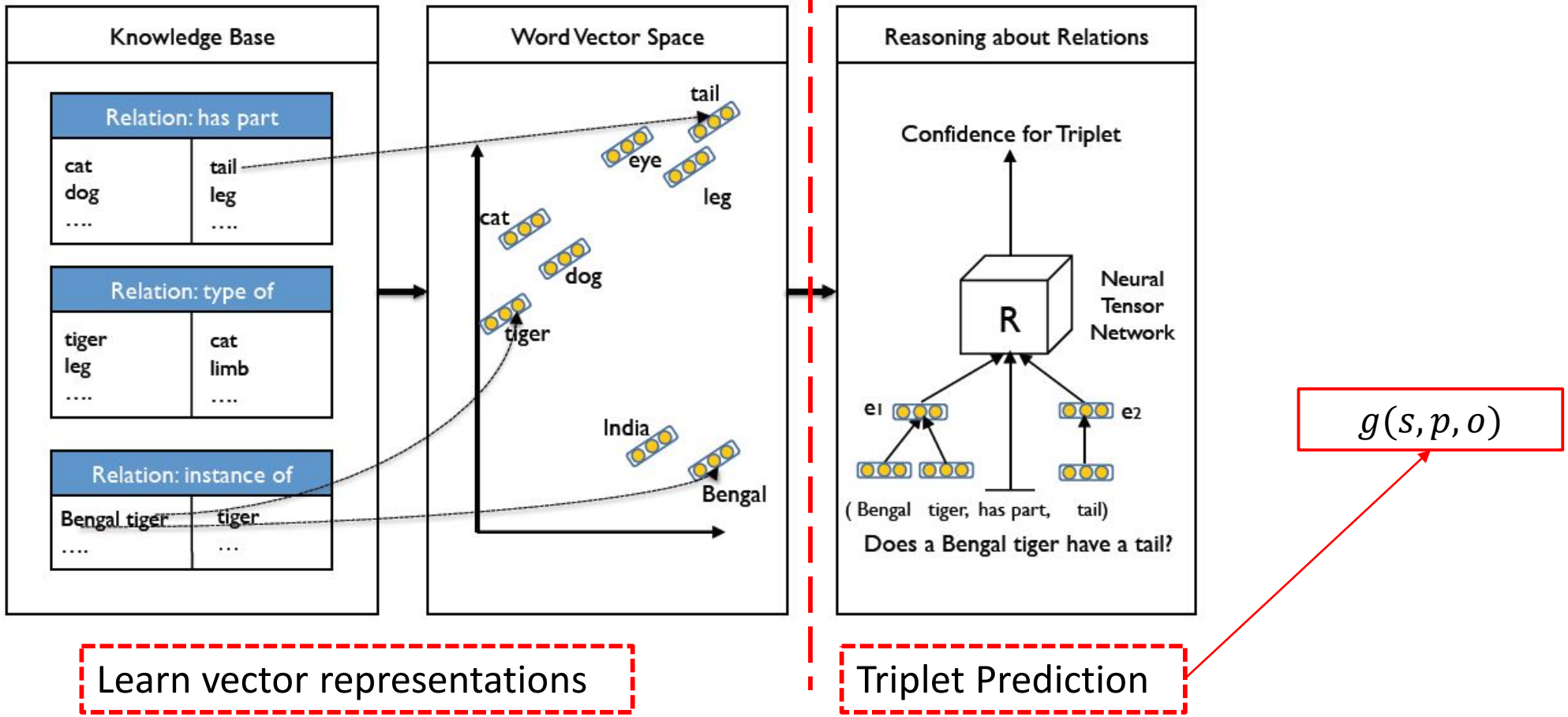
Probability of reaching via path $P$

The set of relation paths

Training parameter

Lao, Ni, Tom Mitchell, and William W. Cohen. "Random walk inference and learning in a large scale knowledge base." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

# Triple Completion: Embedding-based Method




Socher, Richard, et al. "Reasoning with neural tensor networks for knowledge base completion." Advances in neural information processing systems. 2013.

# Embedding-based Method

- Neural Tensor Networks (NTN)

$$g(s, p, o) = u_p^T f\left(e_s^T W_p^{[1:\kappa]} e_o + V_p \begin{bmatrix} e_s \\ e_o \end{bmatrix} + b_p\right)$$

- Translating Embedding (TransE)

$$g(s, p, o) = ||e_s + e_p - e_o||_d$$

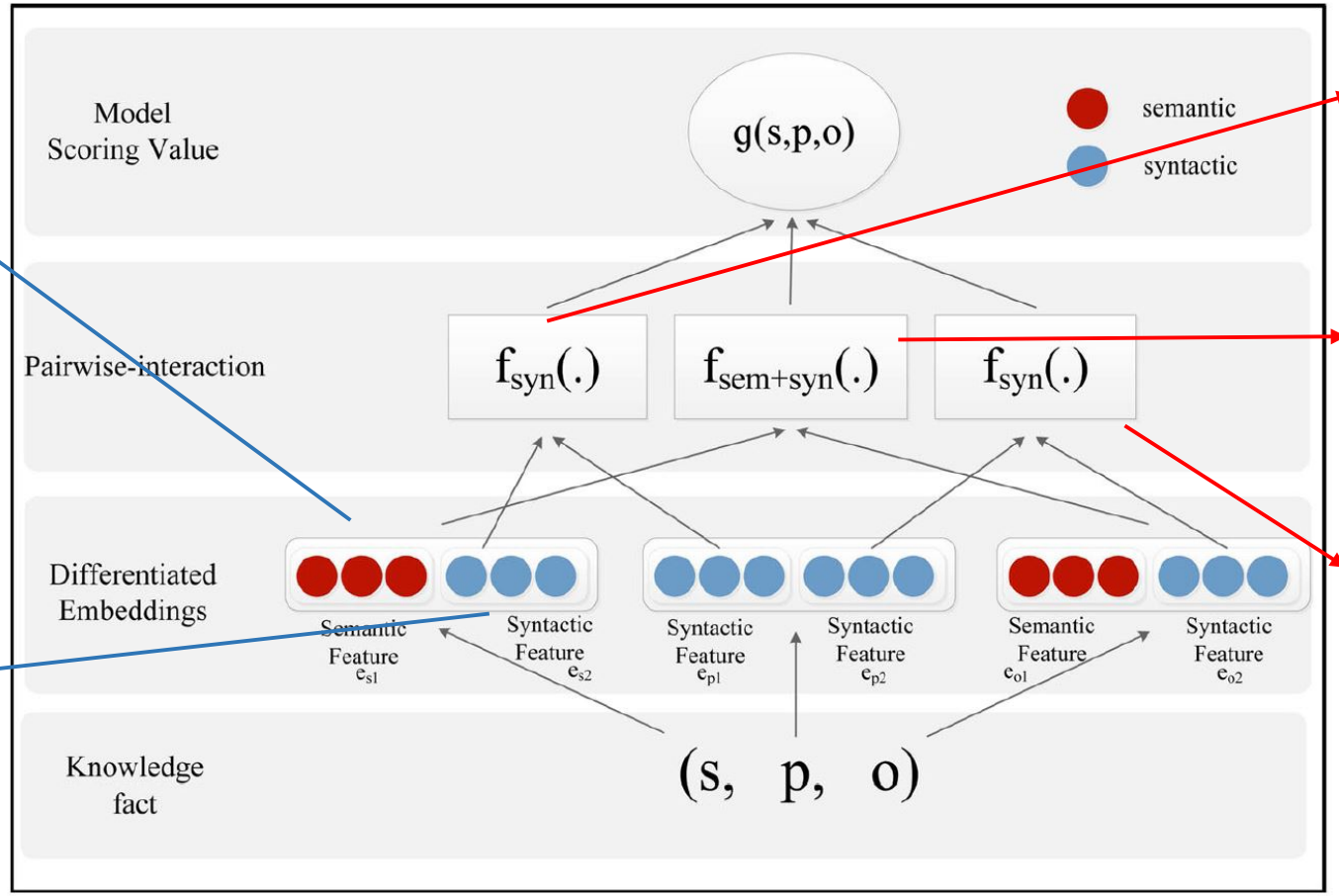
- Translating on Hyperplanes (TransH)

$$g(s, p, o) = || \left(e_s - w_p^T e_s w_p\right) + e_p - \left(e_o - w_p^T e_o w_p\right) ||_d$$

# Triple evaluation by pairwise interaction

The **semantic information** of entities can be extracted from their content information.

The **syntactic information** can be uncovered by their position and order fact.



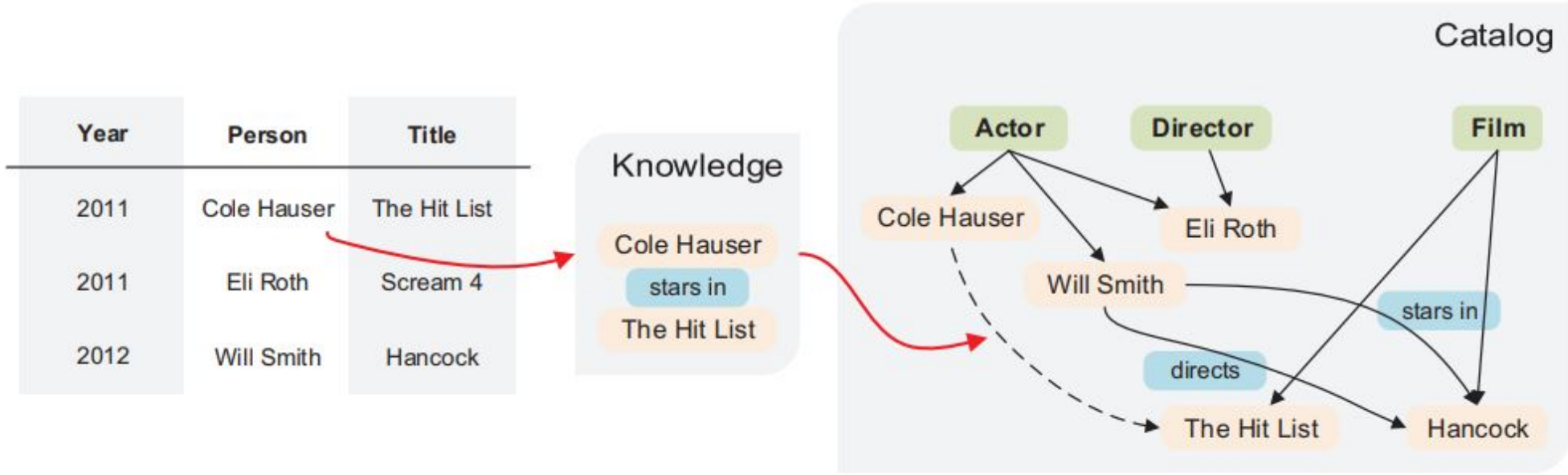

Pairwise-interaction between s and p

Pairwise-interaction between s and o

Pairwise-interaction between p and o

Zhao, Yu, et al. "Knowledge base completion by learning pairwise-interaction differentiated embeddings." Data mining and knowledge discovery 29.5 (2015): 1486-1504.

# KB enrichment with Wikipedia tables



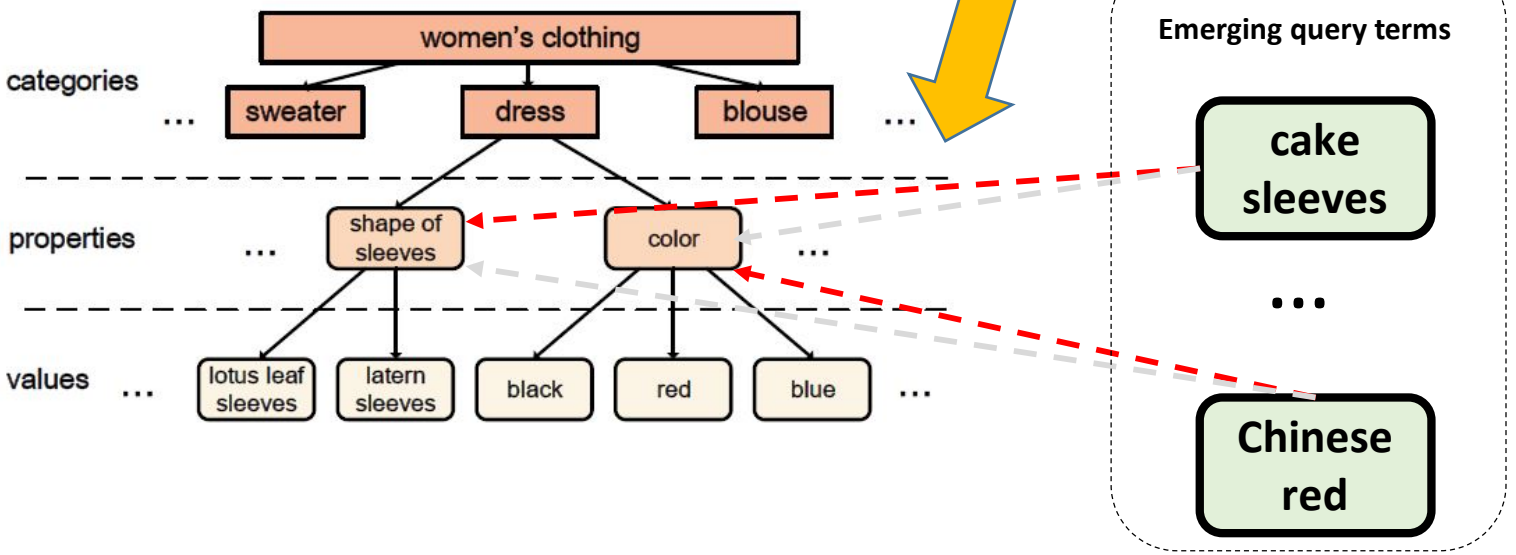

- Step 1: Schema Creation
- Step 2: Table Extraction
- Step 3: Table Annotation

Ran, Chenwei, et al. "Domain-specific knowledge base enrichment using Wikipedia tables." 2015 IEEE International Conference on Data Mining. IEEE, 2015.

# KB enrichment with emerging query terms

Our goal is to find the property $p \in P$ with the highest score:

$$\arg\max_{p \in P} score(p, t)$$




Emerging query terms

cake sleeves

...

Chinese red

Term Similarity Graph

$$score(p, t) = \max_{v_i \in V_p} score'(v_i, t)$$

Jiang, Sihang, et al. "Towards the Completion of a Domain-Specific Knowledge Base with Emerging Query Terms." 2019 IEEE 35th International Conference on Data Engineering (ICDE). IEEE, 2019.

# Entity attributes Implementation by finding Obligatory Attributes

Generalization Rules：

A generalization rule for a KB $K$ is a formula of the form $A \subseteq B$, where $A$ and $B$ are classes of $K$, subject sets of $K$, or intersections thereof.

$$conf(A \subseteq B) = \frac{|A \cap B|}{|A|}$$

Confidence Ratio :

$$s_p^K(c,c') = \frac{conf(c \backslash c' \subseteq p_K)}{conf(c \cap c' \subseteq p_K)}$$

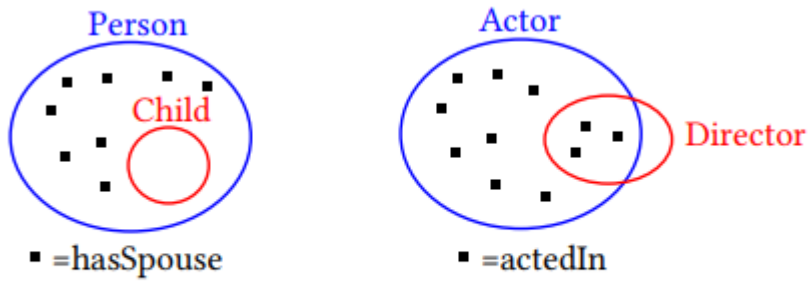



**Figure 1: Examples of attributes and classes.**

**Algorithm 1:** ObligatoryAttribute

**Input:** KB $K$, class $c$, property $p$, threshold $\theta$, threshold $\theta' = 100$

**Output:** *true* if $c \subseteq p_{\mathcal{W}}$ is predicted

```
1 if |c ∩ p_K| < θ' then
2     return false
3 for stable class c' do
4     if |log(s_p^K(c, c'))| > log(θ) then
5         return false
6     if log(s_p^K(c', c)) > log(θ) then
7         return false
8 return true
```

Obligatory Attribute inferring algorithm

Are All People Married?: Determining Obligatory Attributes in Knowledge Bases, WWW2018

# Outline

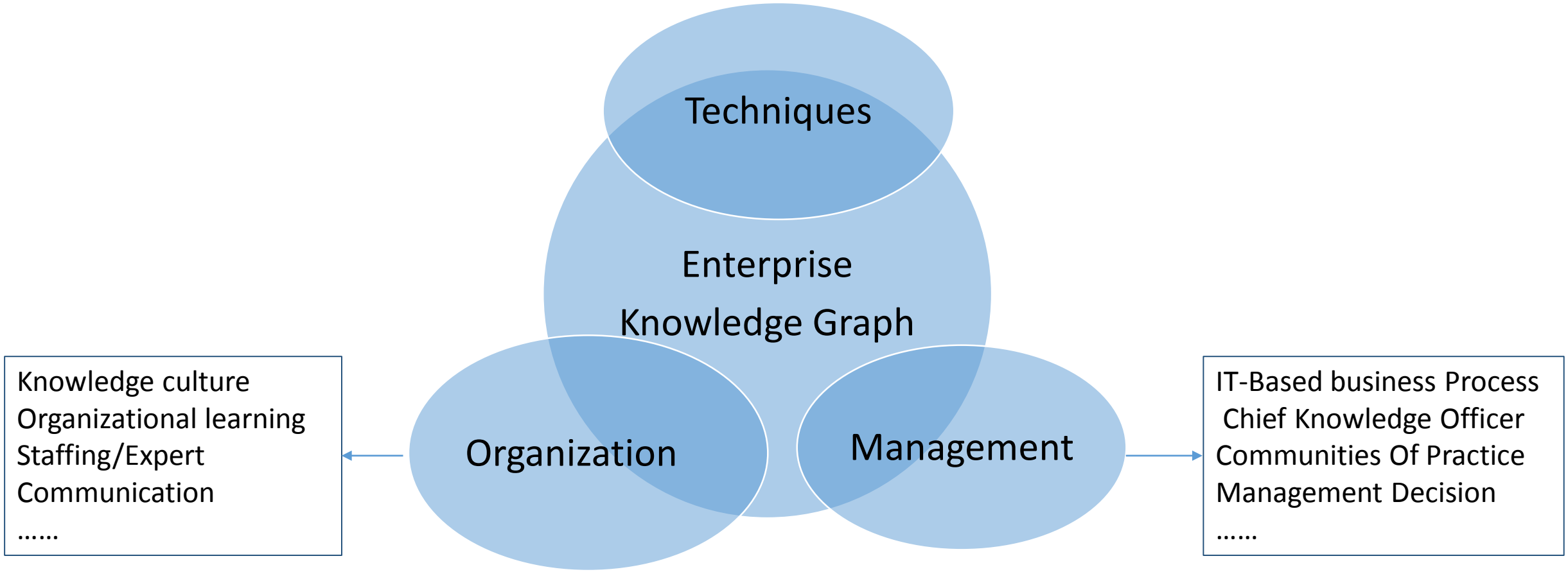
I. Enterprise Knowledge and Enterprise Knowledge Graph

II. Construction of Enterprise Knowledge Graph

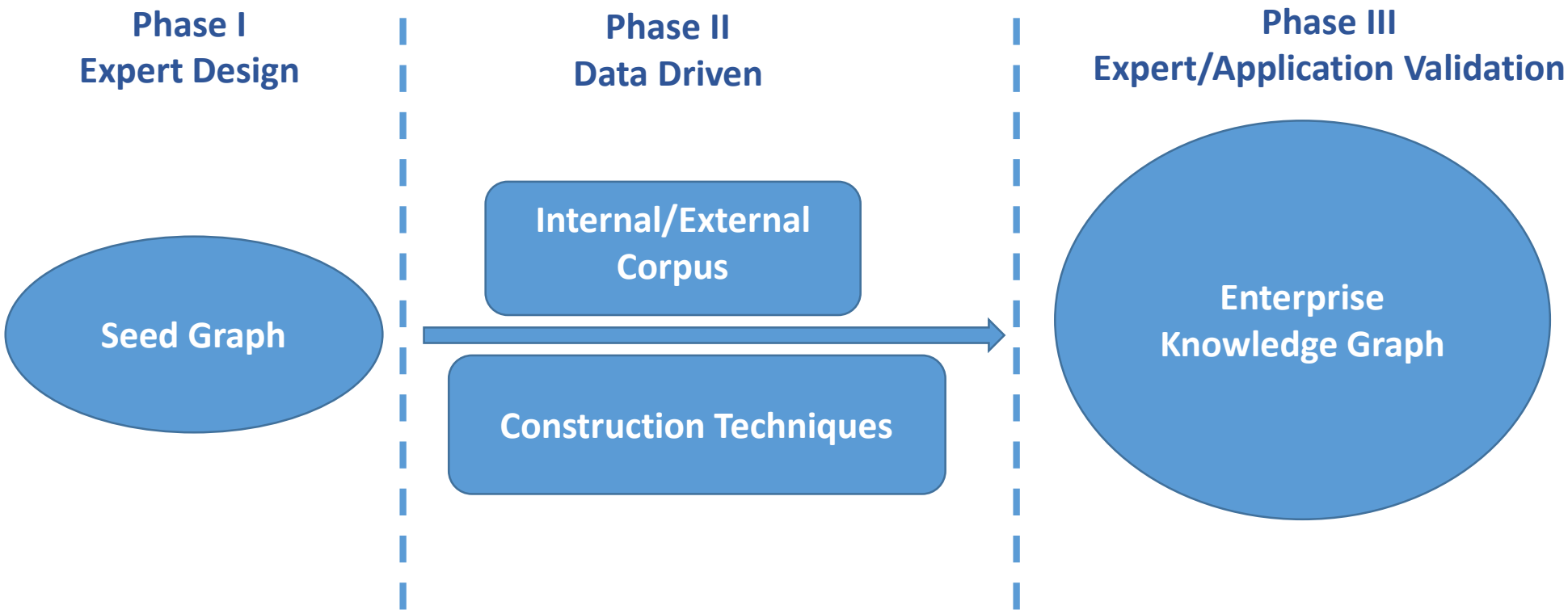
**III. Challenges and Future Research in Enterprise Knowledge Graph**

- Strategy Challenges
- Technical Challenges

# Strategy Challenge – Not Only Techniques



# Strategy Challenge -- Where to start?



# Strategy Challenge -- How to construct Ontology?

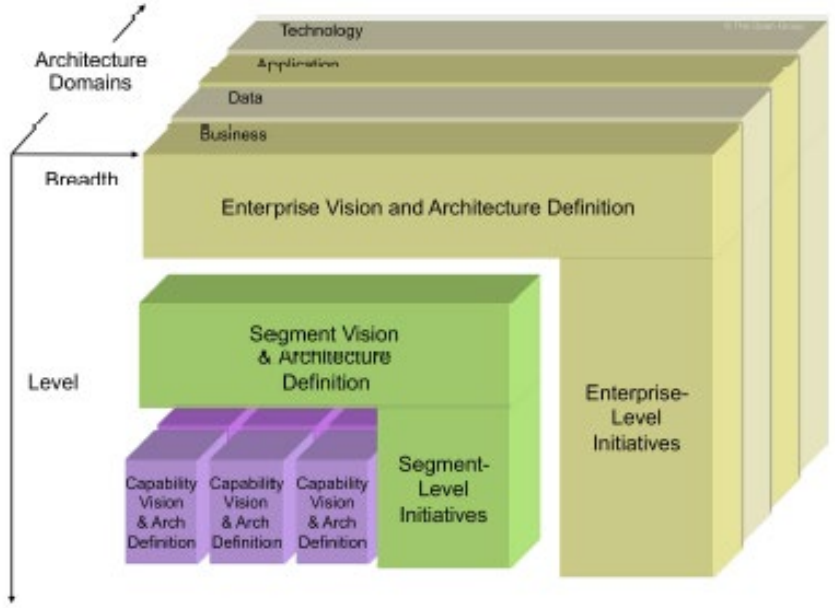
## Enterprise Architecture as the core Ontology

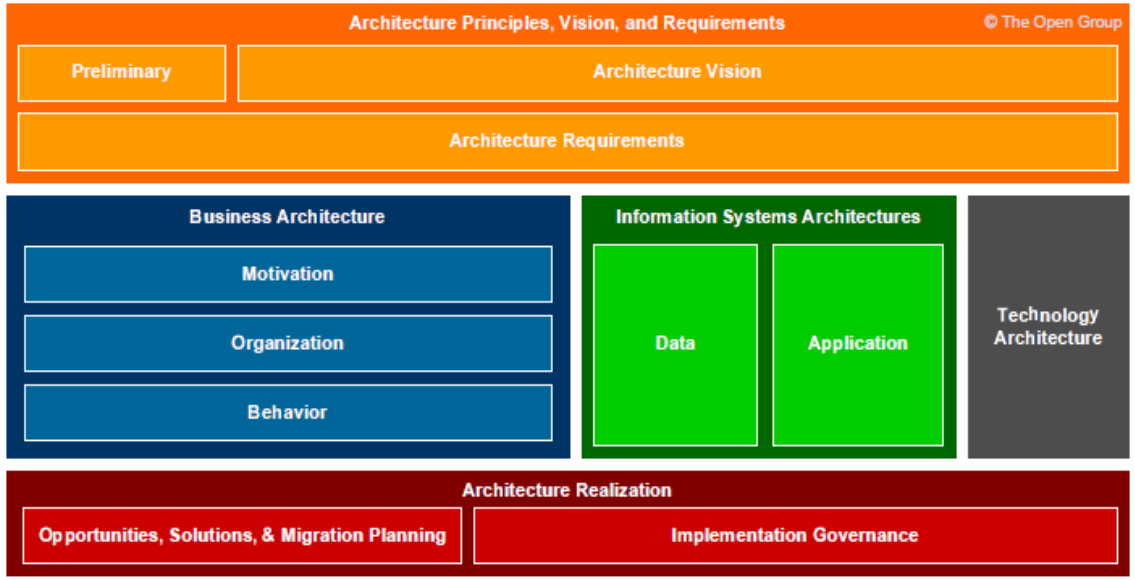
| | What | Who | Where | When | Why | How |
|---|---|---|---|---|---|---|
| **Contextual** | Material List | Organizational Unit &Role List | Geographical Locations List | Event List | Goal List | Process List |
| **Conceptual** | Entity Relationship Model | Organizational Unit &Role Relationship Model | Locations Model | Event Model | Goal Relationship | Process Model |
| **Logical** | Data Model Diagram | Role Relationship Diagram | Locations Diagram | Event Diagram | Rules Diagram | Process Diagram |
| **Physical** | Data Entity Specification | Role Specification | Location Specification | Event Specification | Riles Specification | Process Function Specification |
| **Detailed** | Data Details | Role Details | Location Details | Event Detail | Rules Detail | Process Details |

Enterprise Architecture (EA) is defined as understanding the elements of an organization and how the elements relate to each other. J. Schekkerman state that organizations at 149 countries have implemented Enterprise Architecture. Zachman Framework (23%), TOGAF (11%) dan FEAF (11%) [3].Open Group Architecture Framework(TOGAF)

# Strategy Challenge -- How to construct Ontology? Enterprise Architecture as the core Ontology

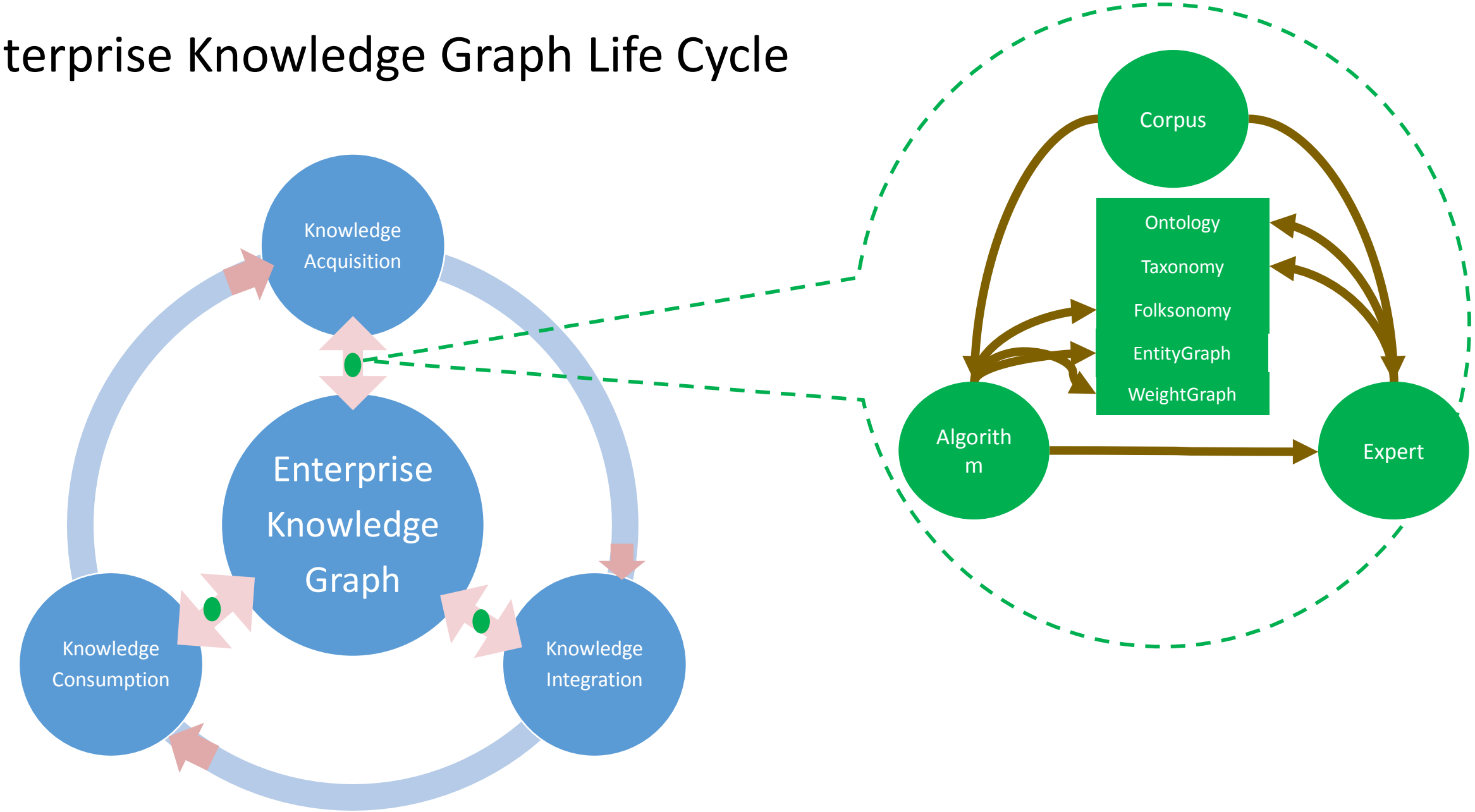
Open Group Architecture Framework (TOGAF)



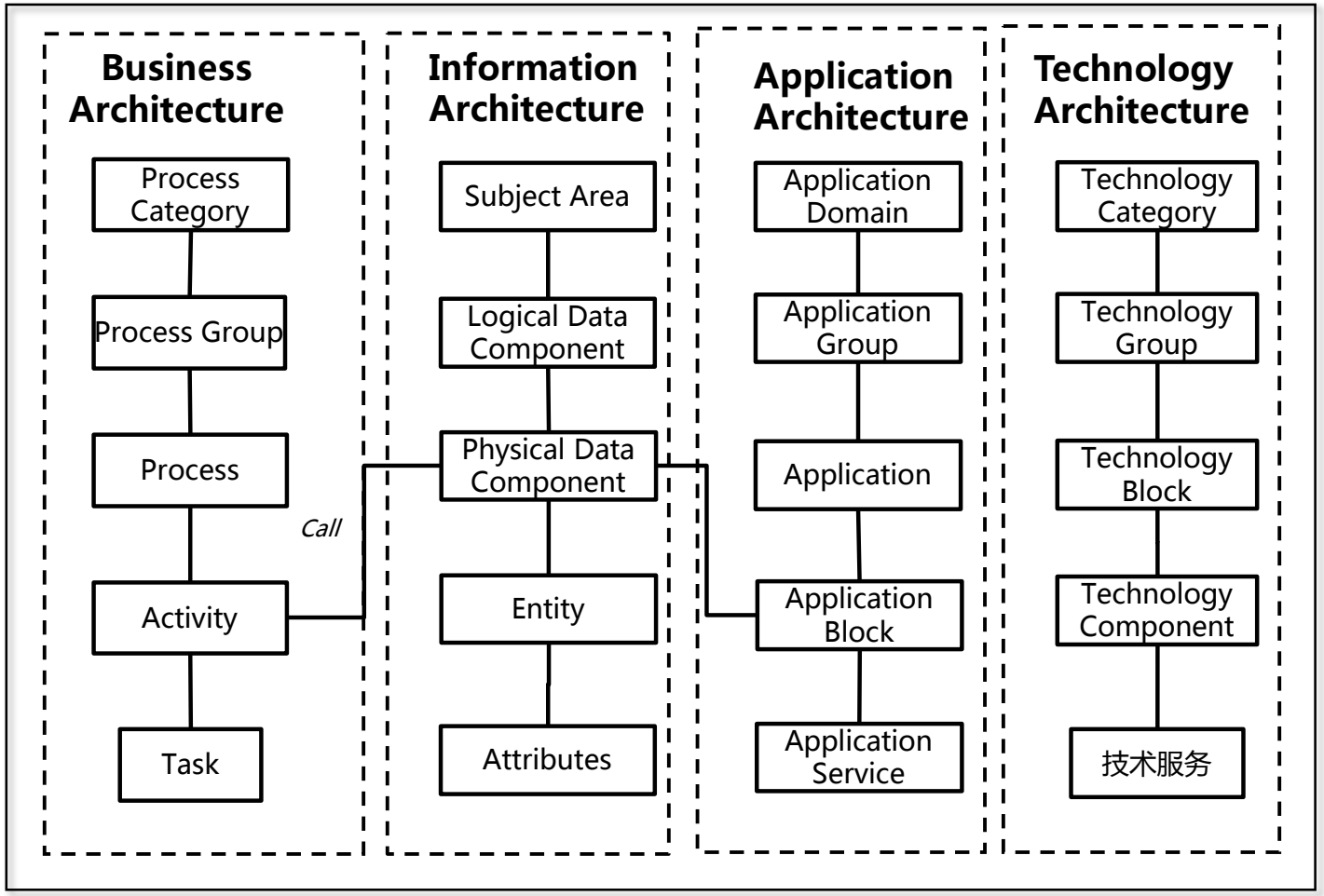



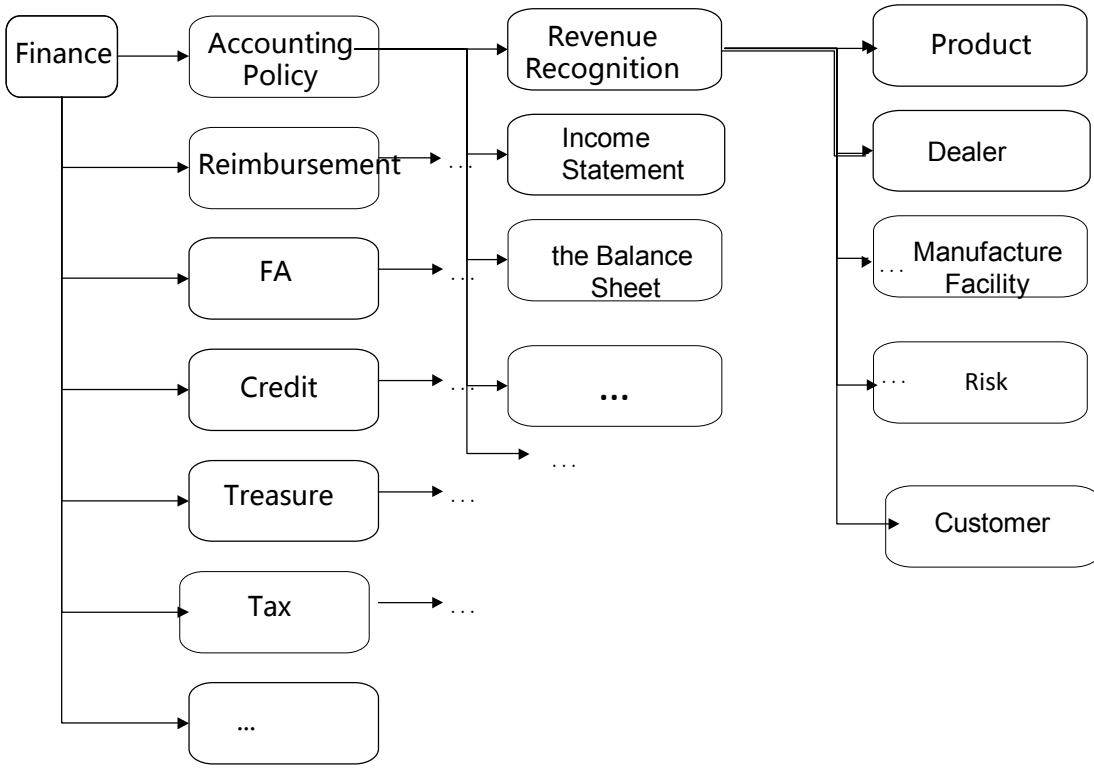

The TOGAF® Standard, Version 9.2

# Enterprise Knowledge Graph Life Cycle



# **Strategy Challenge**：Multiple Taxonomies derived from Enterprise Architecture



# Technical Challenges

## Development

- Unavoided Cold Start Problem
- Small seed graph
- Limited amount of data

a) Corpus size

b) Lack of attributes

- Uncertainty of Expert experience
- Limitation of Crowd-Sourcing

## Evaluation

- Large amount of candidates
- Lack of Ground truth

## System Control

- Update
- Coordinating among different subgraph updating
- Privacy and Security

# **Technical Challenges**：Data Integration, Information Fusion and Graph Merging

**Data Integration : the core of Enterprise Data**

**Multimodal**

- Information Retrieval from different mechanism
  - PPT
  - Excel
  - EWD
  - Flatfile
  - Website
  - Image/Video/Audio
  - IOT
- Format, Syntax, Quality
- Multi-System  Semantic Integration
- Multi-Platform Semantic Integration

# Technical Challenges – Expert Knowledge

- Crowdsourcing
  - What types of problem are good for crowdsoucing?
  - How to pick the labor?
  - How to design the questions
  - How to control and evaluate the output?
  - How to balance the cost and output?
- Expert
  - Representation

  a) How to represent expert knowledge?

  - Quality

  a) How to quantify the quality of expert knowledge

  b) Uncertainty of expert

# Technical Challenges – Security & Privacy Control

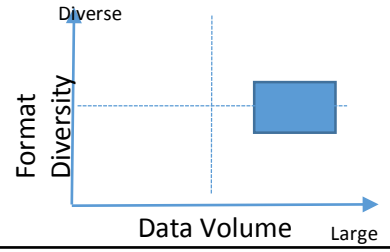
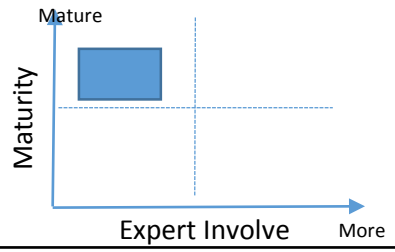
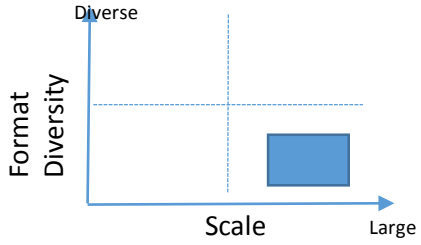
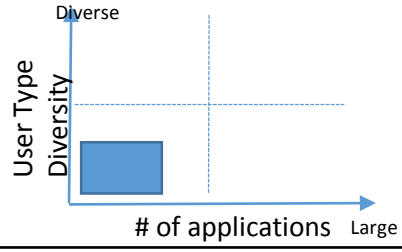
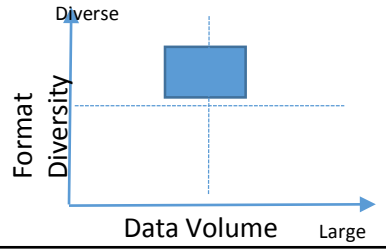
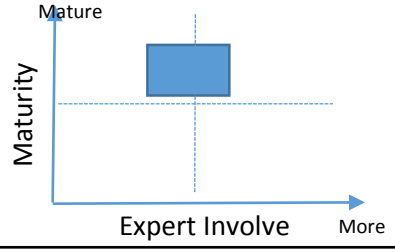
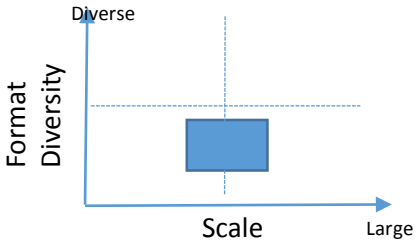
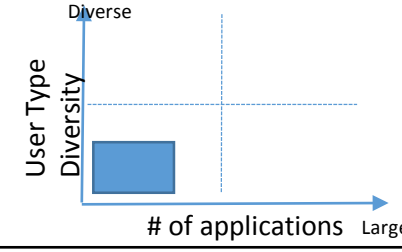
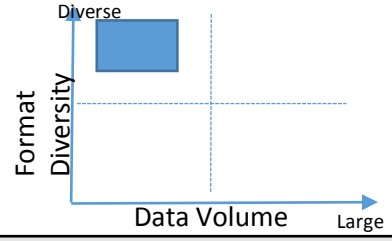
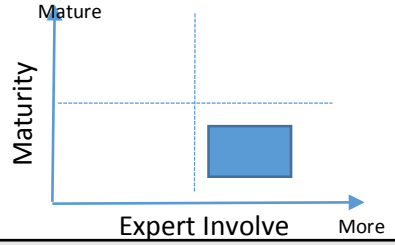
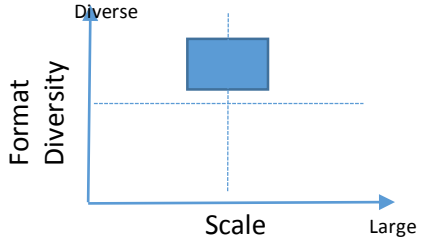
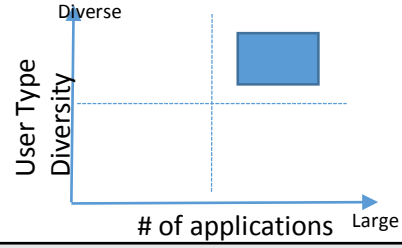
- ID management incorporate with knowledge graph
- How to manage security and privacy issue in a knowledge graph?
  - Locate the leaking, evaluate the impact and fix
  - What’s the potential leaking after link and inference from safe isolated information

# Technical Challenges – Knowledge Updating and validation

- Keep growing knowledge base
- Coordinating among different subgraph updating
- Knowledge graph quality evaluation

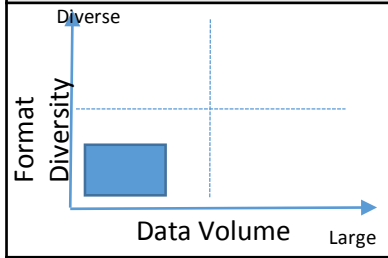
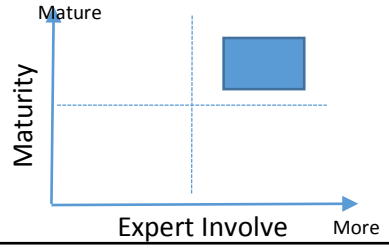
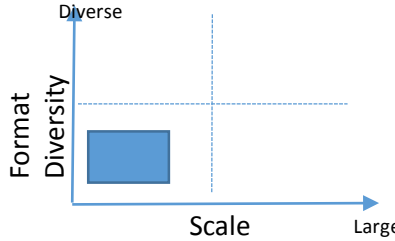
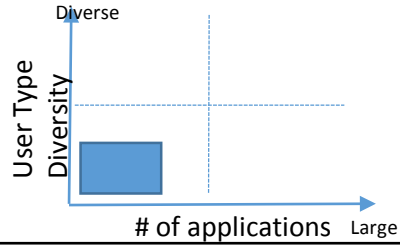
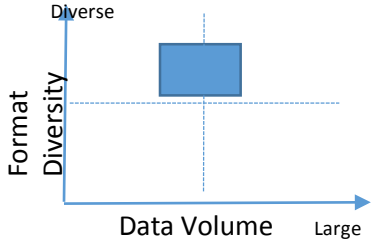
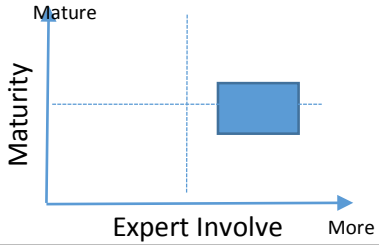
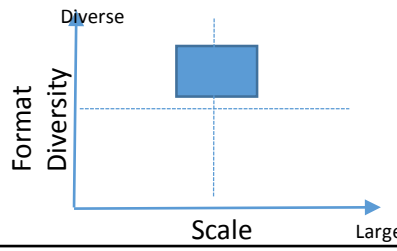
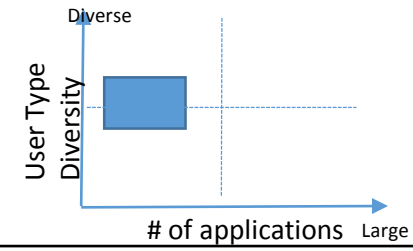
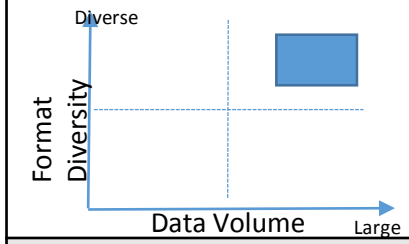
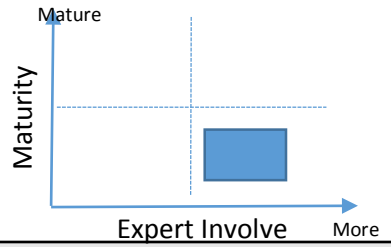
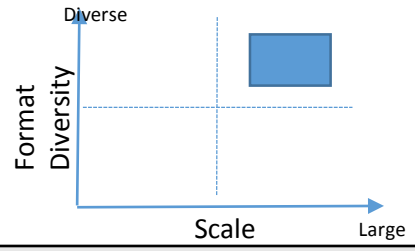
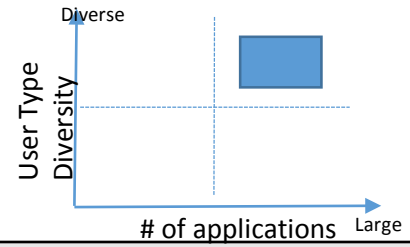
# Practical Consideration -- What Techniques should adapt

Summarize the difference among Open Domain, Specific Domain and Enterprise

| | Data Sources | Construction<br>Entity Relation | Storage | Consumption |
|---|---|---|---|---|
| **Open Domain** |  |  |  |  |
| **Specific Domain** |  |  |  |  |
| **Enterprise** |  |  |  |  |
| | Unstructured Document<br>Semi-Structured (Web)<br>Structured | Experience Driven --Pattern Based (Regular Expression, Token, Template)<br>Data Driven –<br>Statistical /Traditional Machine Learning Model<br>Deep Learning Model<br>Embedding | Relational Data Base<br>Key Value<br>Graph Database | User Group |

# Practical Consideration -- What Techniques should adapt

Summarize the difference among Three Types of EKG

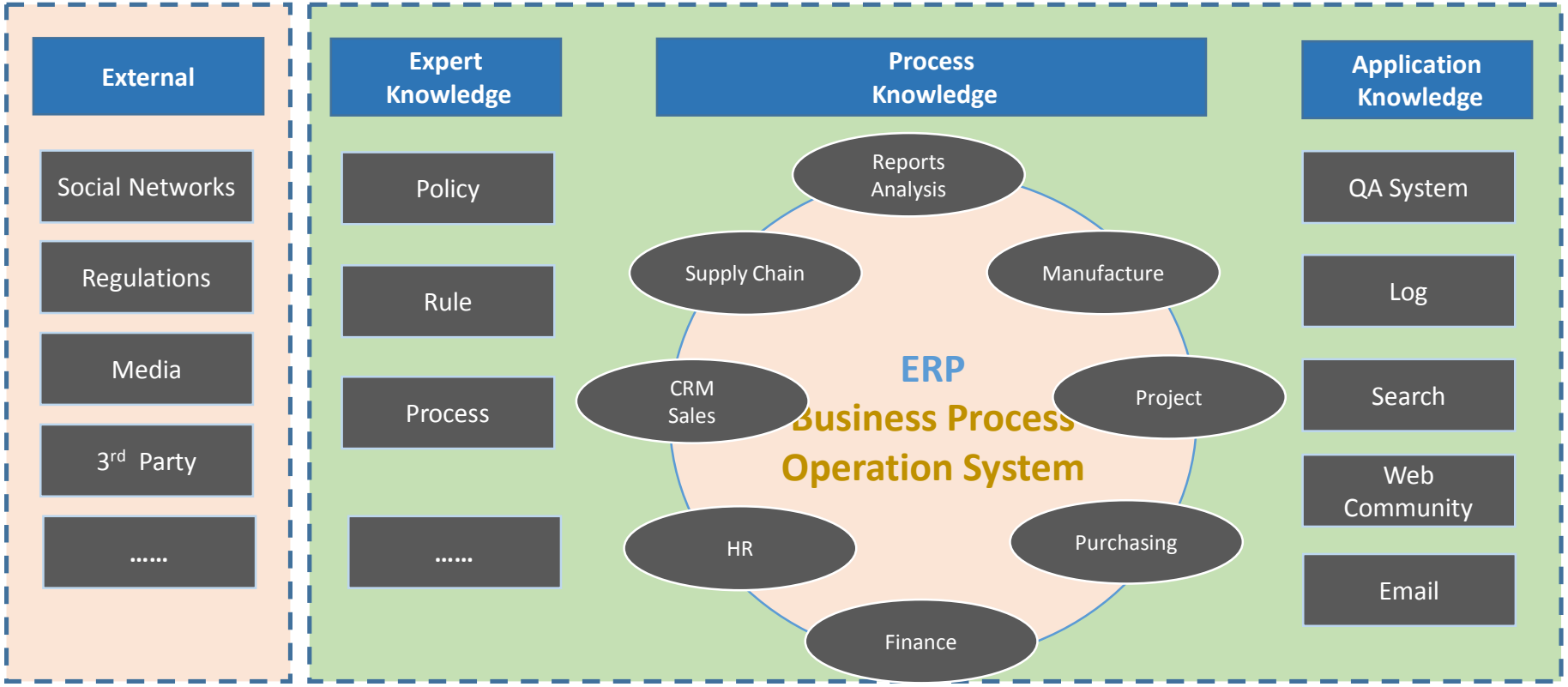
| | Data Sources | Construction<br>Entity Relation | Storage | Consumption |
|---|---|---|---|---|
| **Specific Problem** |  |  |  |  |
| **Specific Domain** |  |  |  |  |
| **Cross Domain** |  |  |  |  |
| | Unstructured Document<br>Semi-Structured (Web) | Experience Driven --Pattern Based(Regular Expression, Token, Template)<br>Data Driven –<br>Statistical /Traditional Machine Learning Model<br>Deep Learning Model<br>Embedding | Relational Data Base<br>Key Value<br>Graph Database | User Group |

# When Construct EKG, be creative

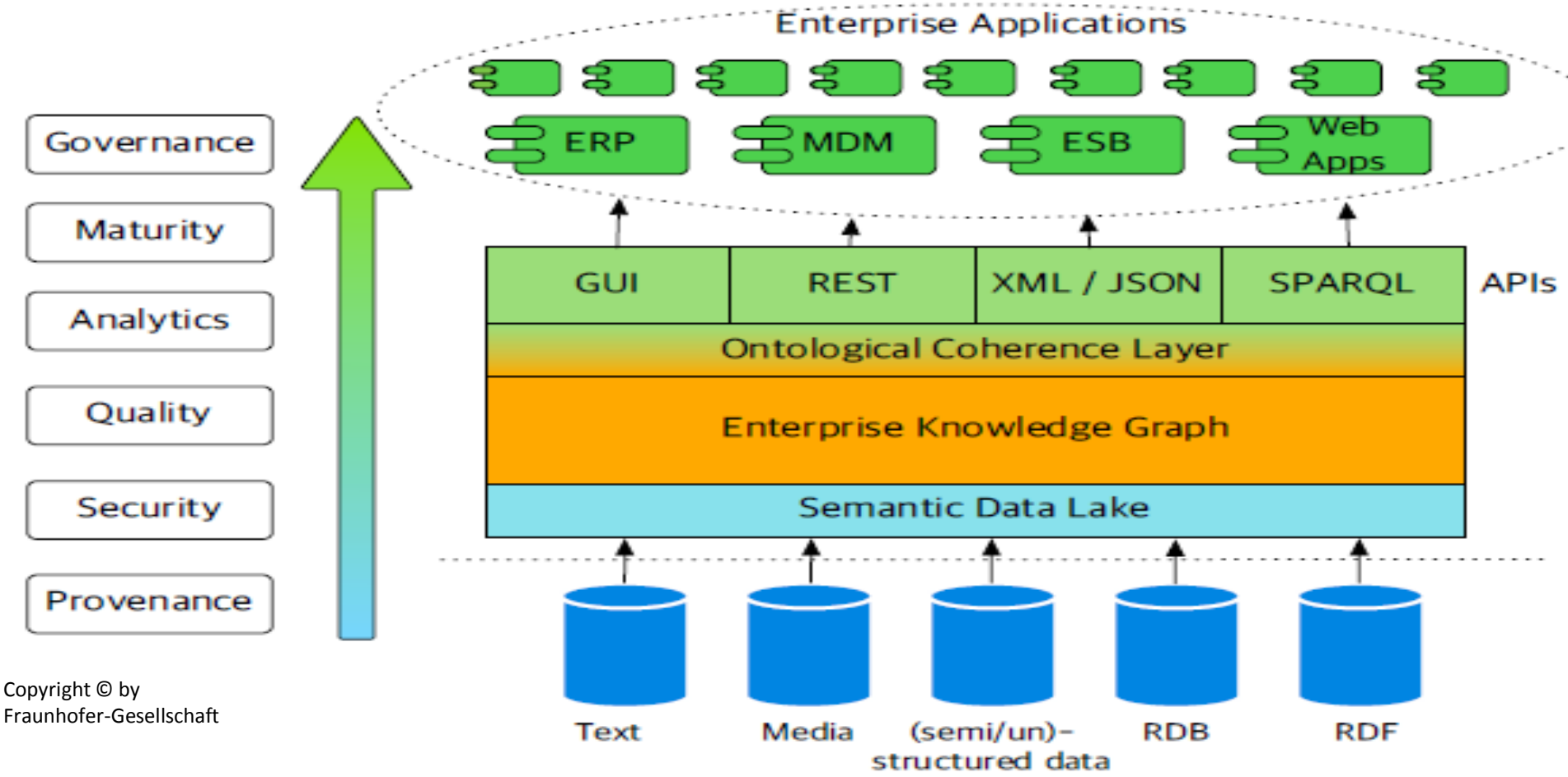
- Think thoroughly  the scope of the work
- Think thoroughly  the availability of resource(Expert, Technical)
- Think carefully about the cost(hardware, software, labor, time)
- Adopt the technology/software carefully
    - Maturity
    - Market Share

# Thanks

# Modern Enterprise Knowledge Components



# Enterprise Knowledge Management System & Enterprise Information System






Galkin, Mikhail & Auer, Sören & Vidal, Maria-Esther & Scerri, Simon. (2017). Enterprise Knowledge Graphs: A Semantic Approach for Knowledge Management in the Next Generation of Enterprise Information Systems.