Building and Exploring an Enterprise Knowledge Graph for Investment Analysis

Tong Ruan[1], Lijuan Xue[1], Haofen Wang[1]

Fanghuai Hu[2], Liang Zhao[1], Jun Ding[2]

[1] East China University of Science and Technology

[2] Shanghai Hi-knowledge Information Technology Corporation

Business Motivation

- In China, most securities companies provide investment bank services and investment consulting services.
- Establishment of the New Third Board, a national share transfer system for small- and medium-sized enterprises (SMEs)
- Small innovation companies can be listed on the “New Third Board” with the endorsement of securities companies.

Business Motivation

- Securities companies serve from big enterprises to small and medium-sized enterprises.
- There are about 40M companies in China. It is difficult for the securities companies themselves to gather authentic and full-fledged company information of their customers and potential customers.
- Therefore we collect company information from different sources for them and represent it in easy-to-use graphs. The “Magic Mirror” targets to help the securities companies to know and to approach their target companies better and quicker.

Deployment and Business Model

✓Size: About two hundred million entities, one billion attribute value pairs, and two hundred million relations in EKG.

✓Time: It takes an hour to extract entities, three hours to extract attribute value pairs, and three hours to extract relations from various sources.

✓Update: rebuilt once a month to incorporate newly added enterprise data. (Simple and require improvement)

Deployment and Business Model

✓ Sell the whole solution as services instead of software. Securities companies have customized the EKG portal and have integrated it into their own applications.

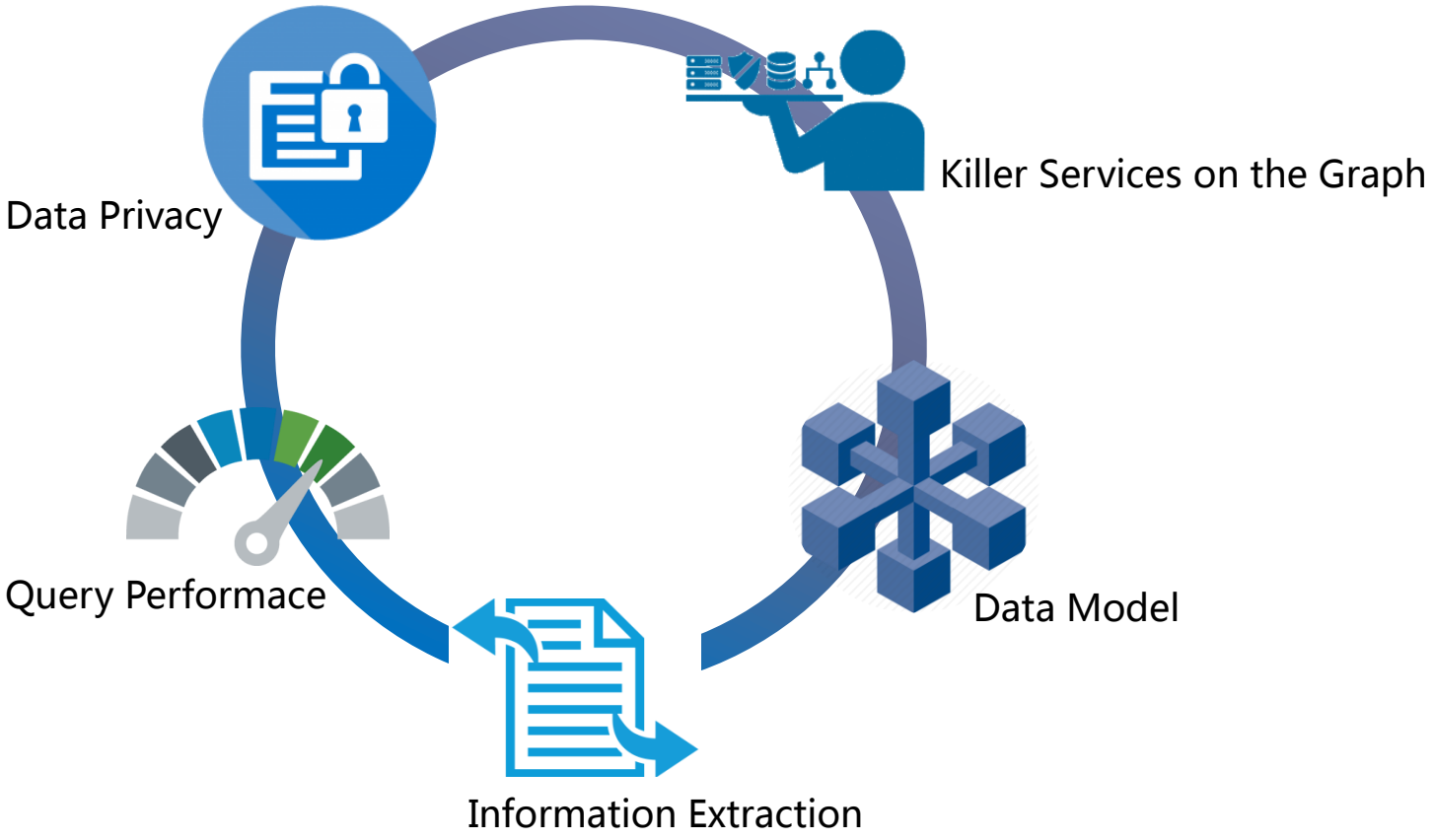
✓Pay by times of API access Per Year

✓ General querying and graph visualization services

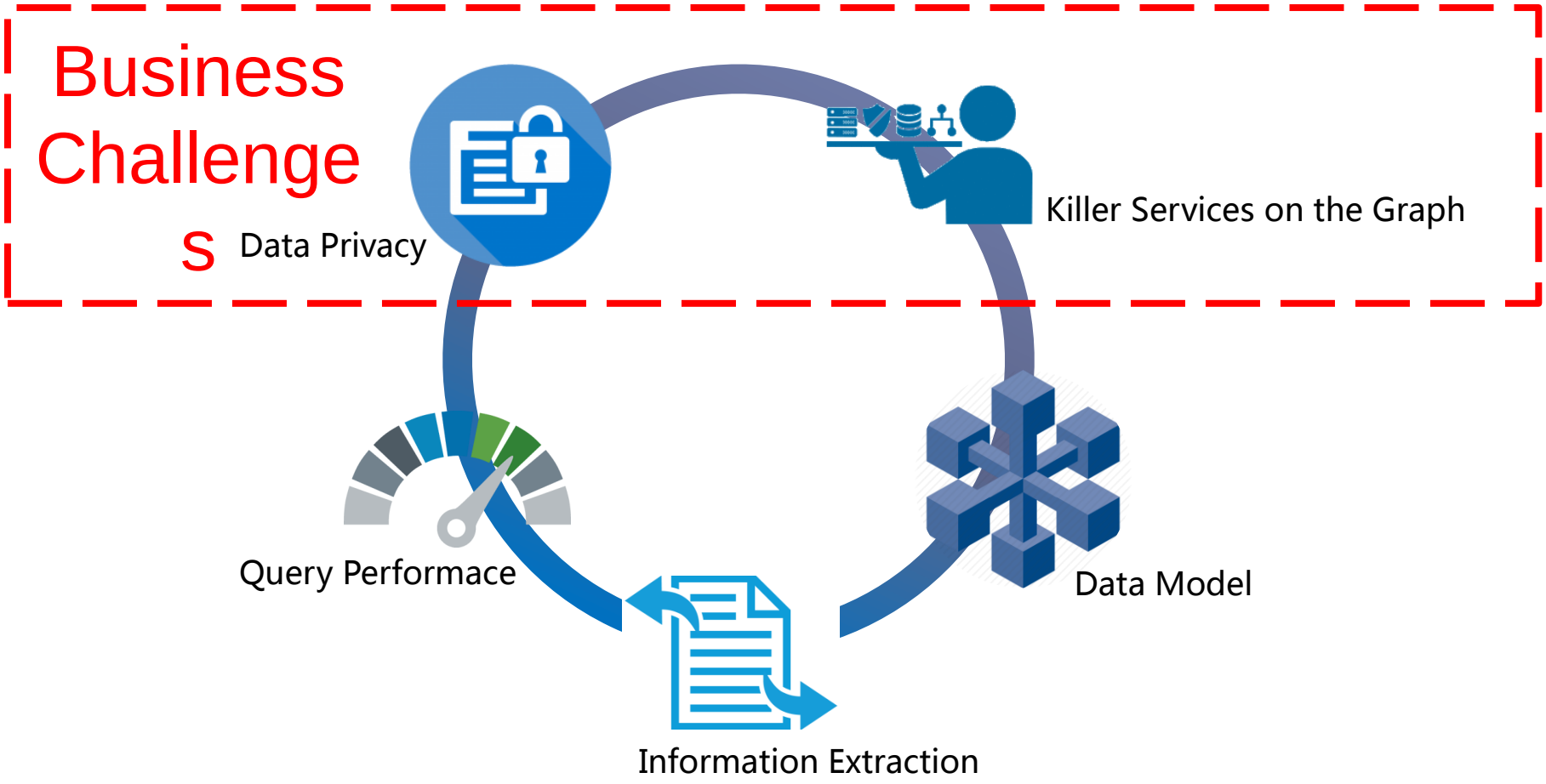
✓ In-depth analyzing services dedicated to investing requirements

Business Model

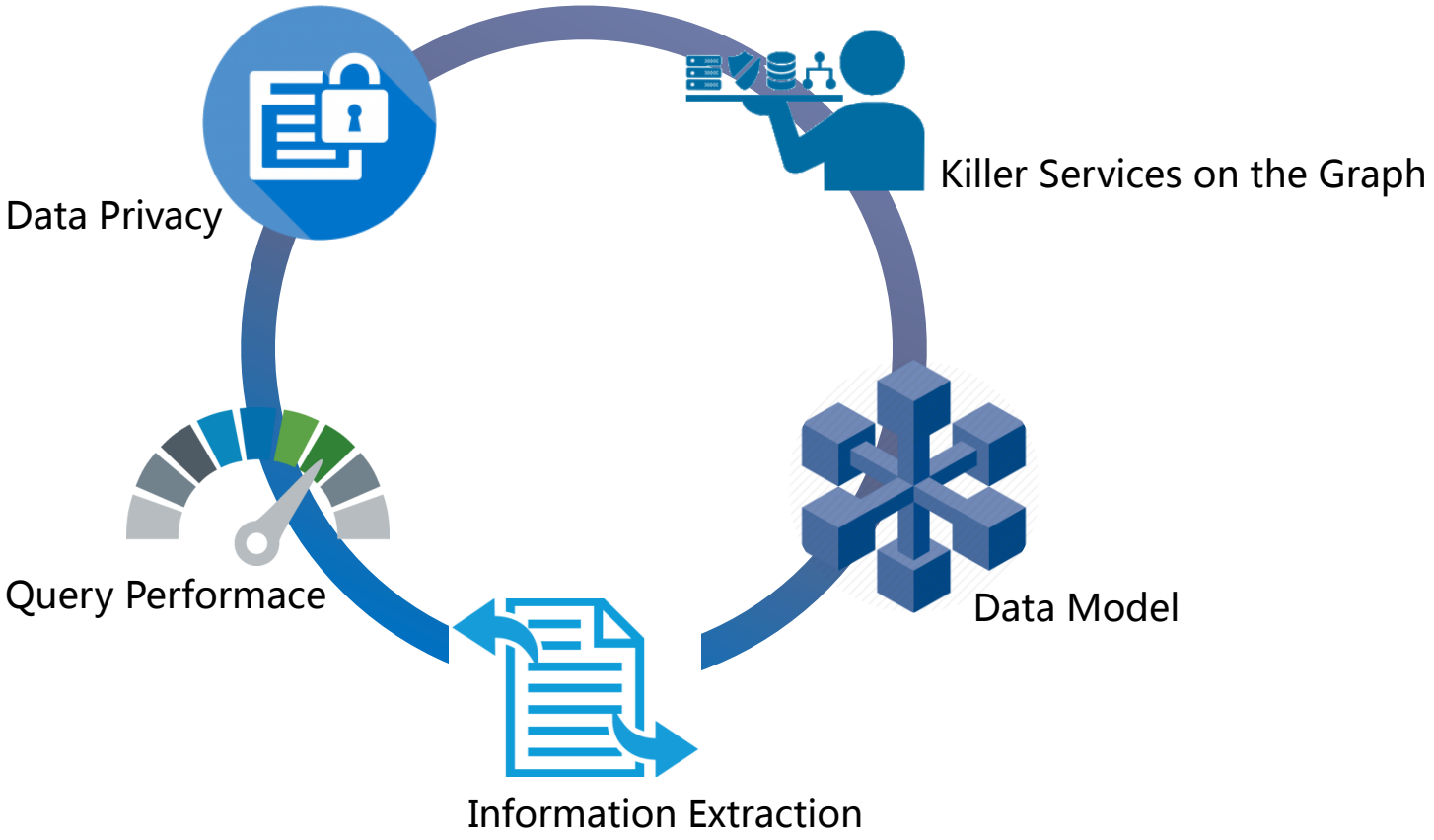
Challenges


Data Privacy
Killer Services on the Graph
Query Performace
Data Model
Information Extraction

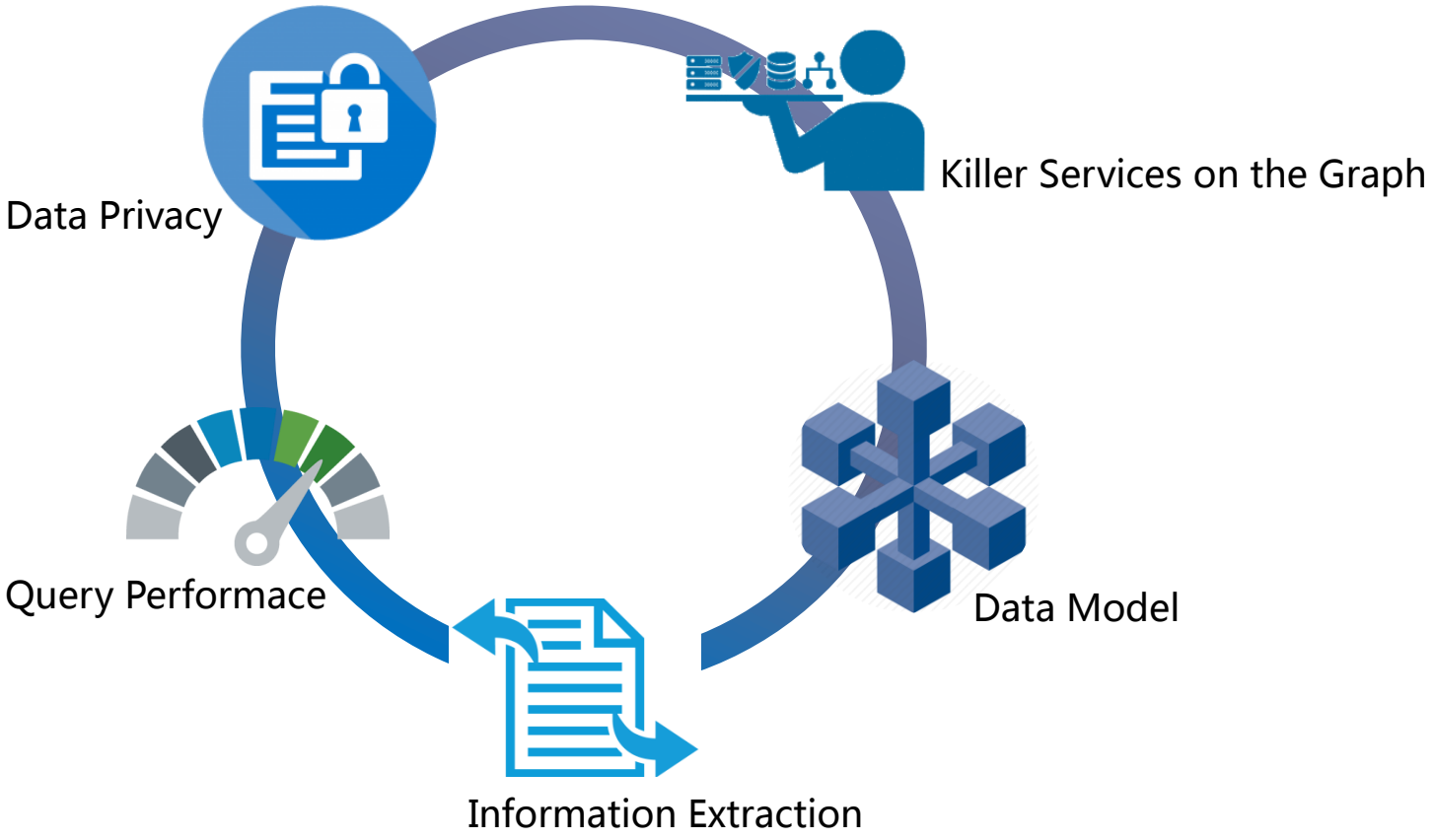
Challenges


Business Challenge
s
Data Privacy
Killer Services on the Graph
Query Performace
Data Model
Information Extraction

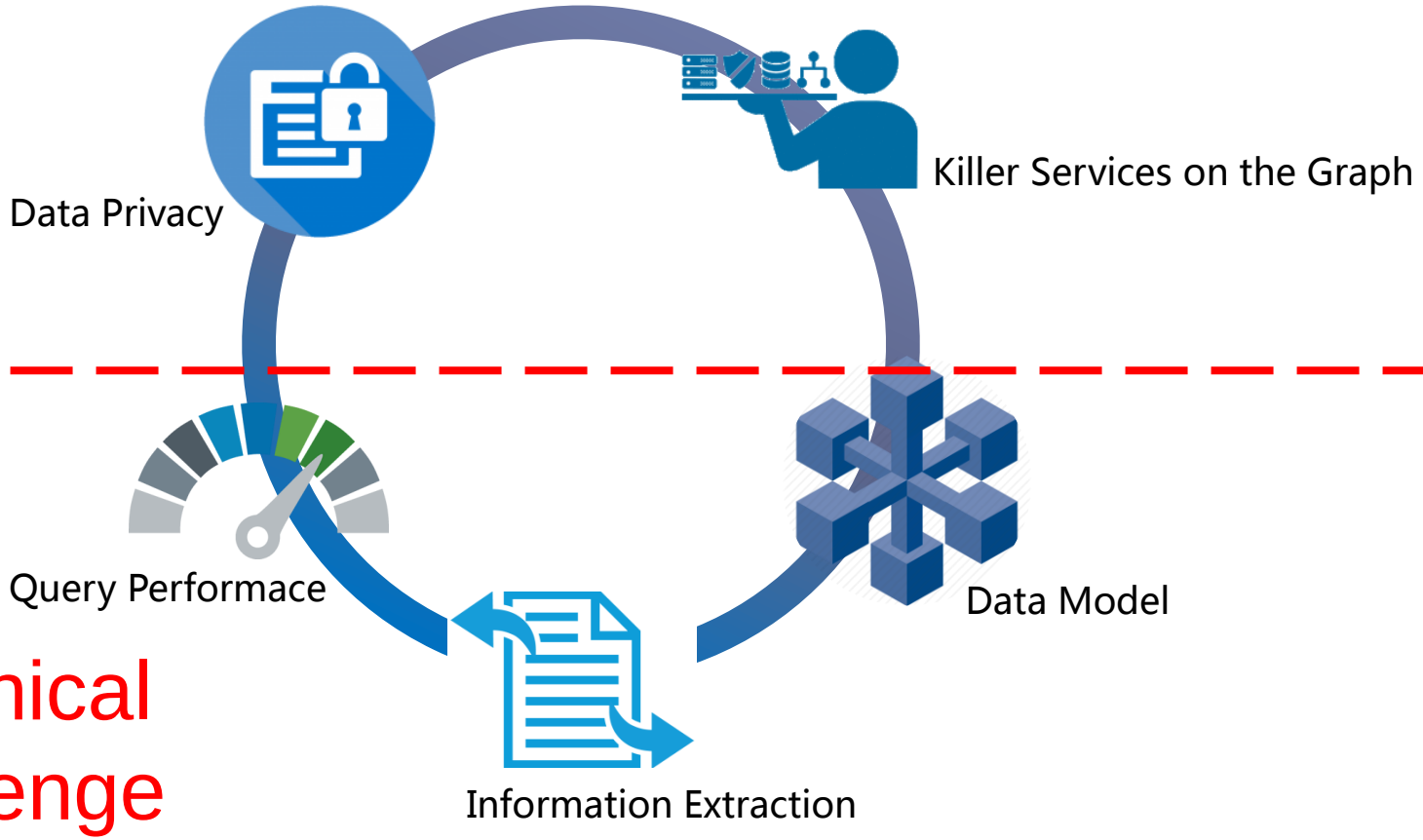
Challenges


Data Privacy
Killer Services on the Graph
Query Performace
Data Model
Information Extraction

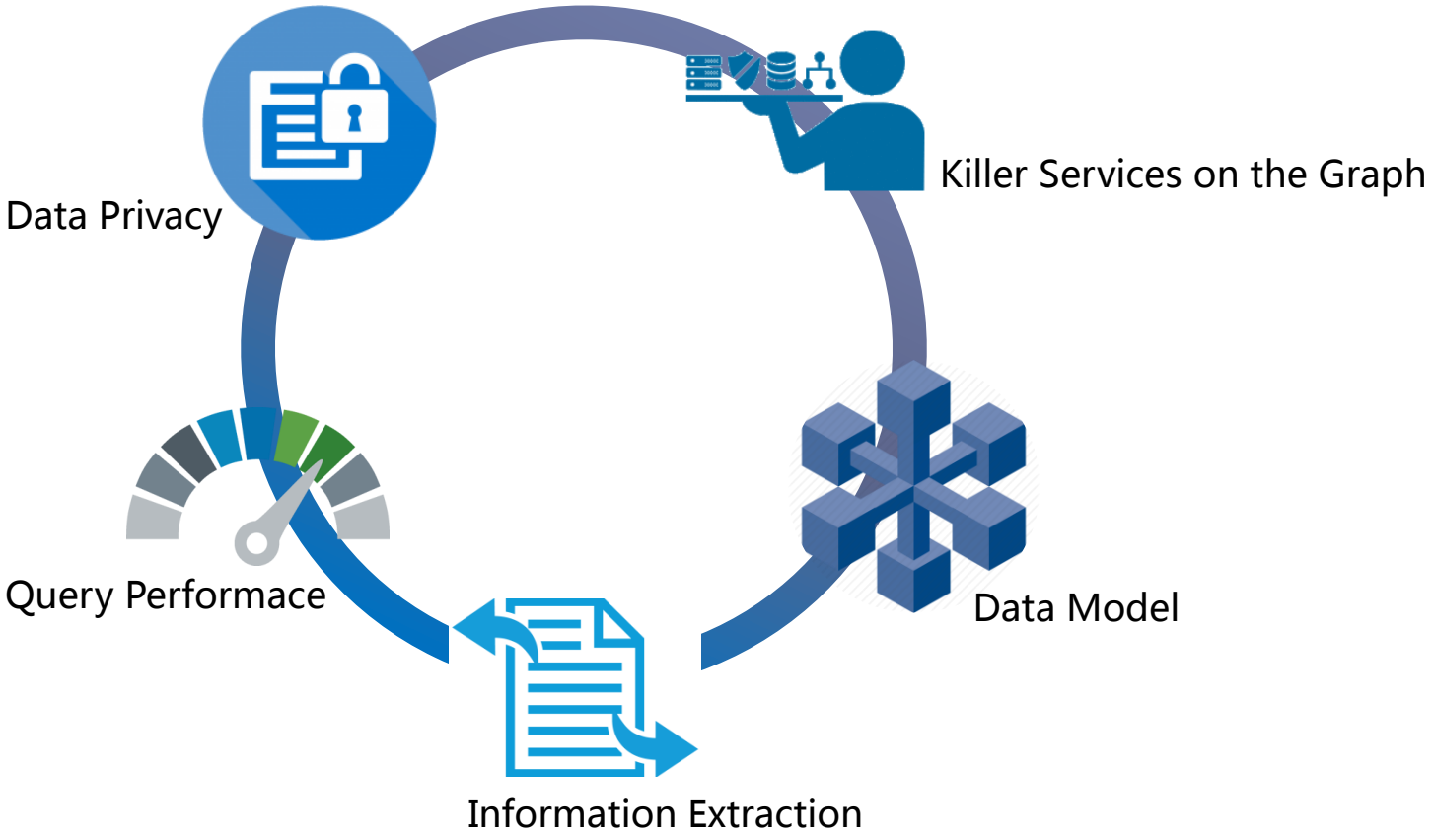
Challenges


Data Privacy
Killer Services on the Graph
Query Performace
Data Model
Information Extraction

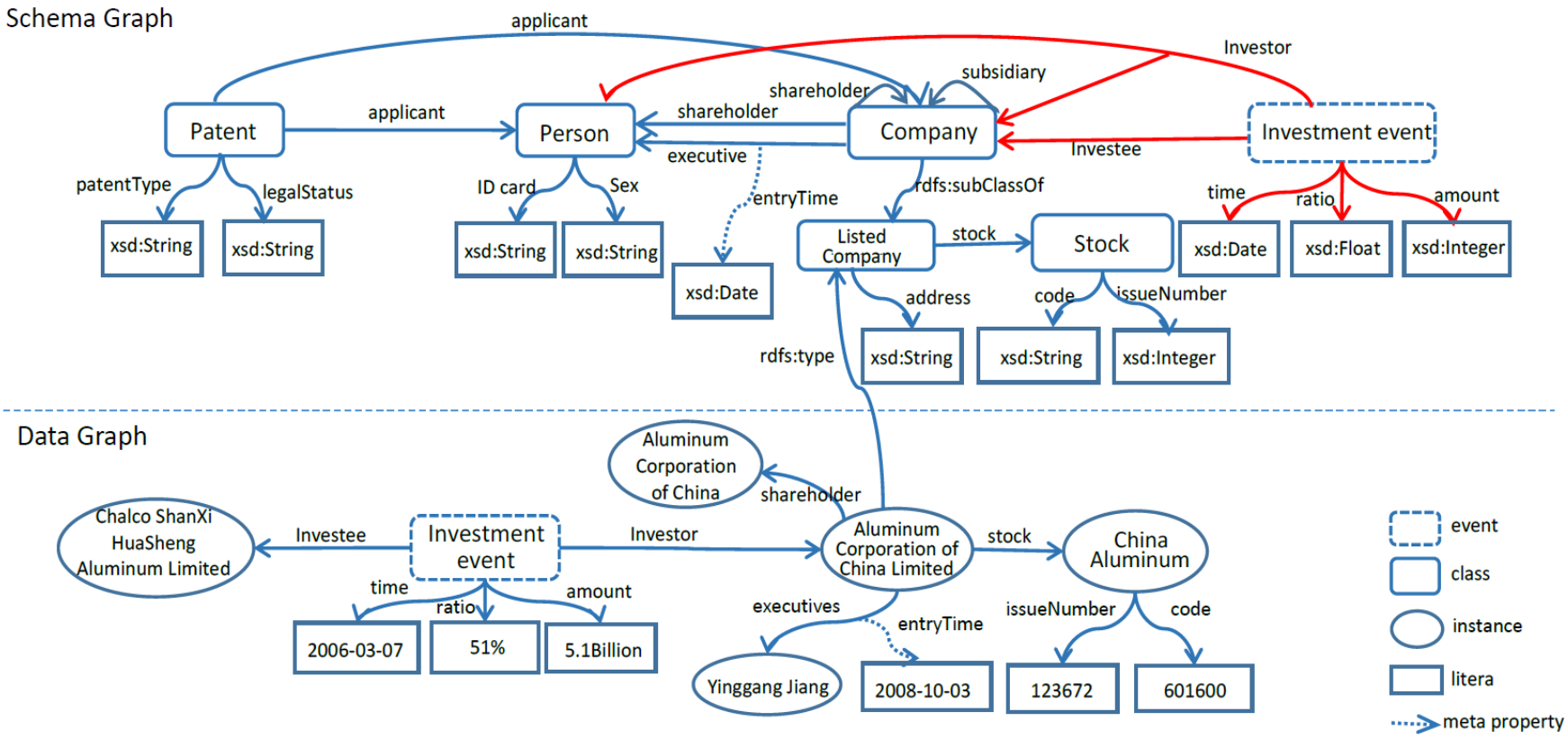
Challenges


Data Privacy
Killer Services on the Graph
Query Performace
Data Model
Information Extraction
Technical Challenges

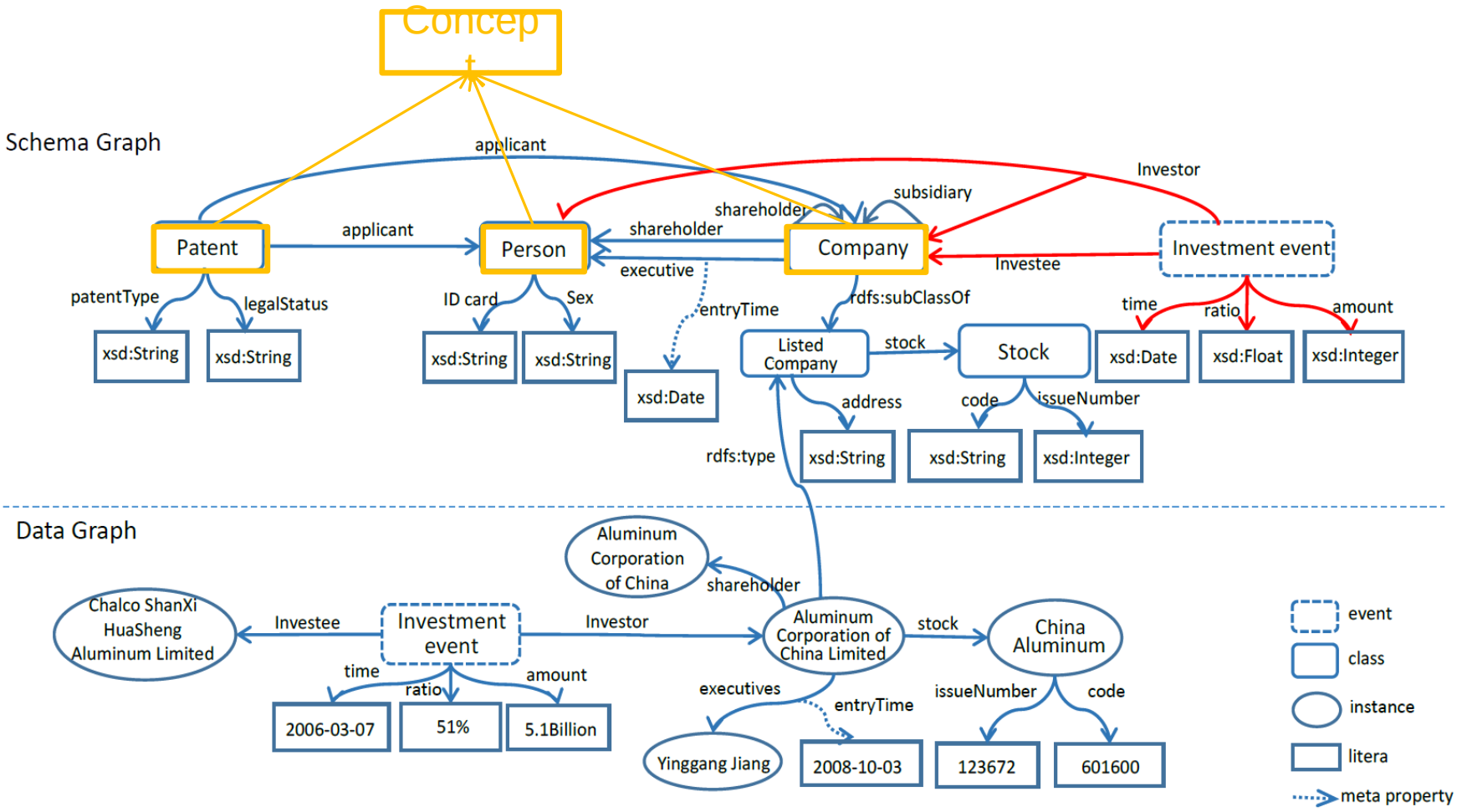
Challenges


Data Privacy
Killer Services on the Graph
Query Performace
Data Model
Information Extraction

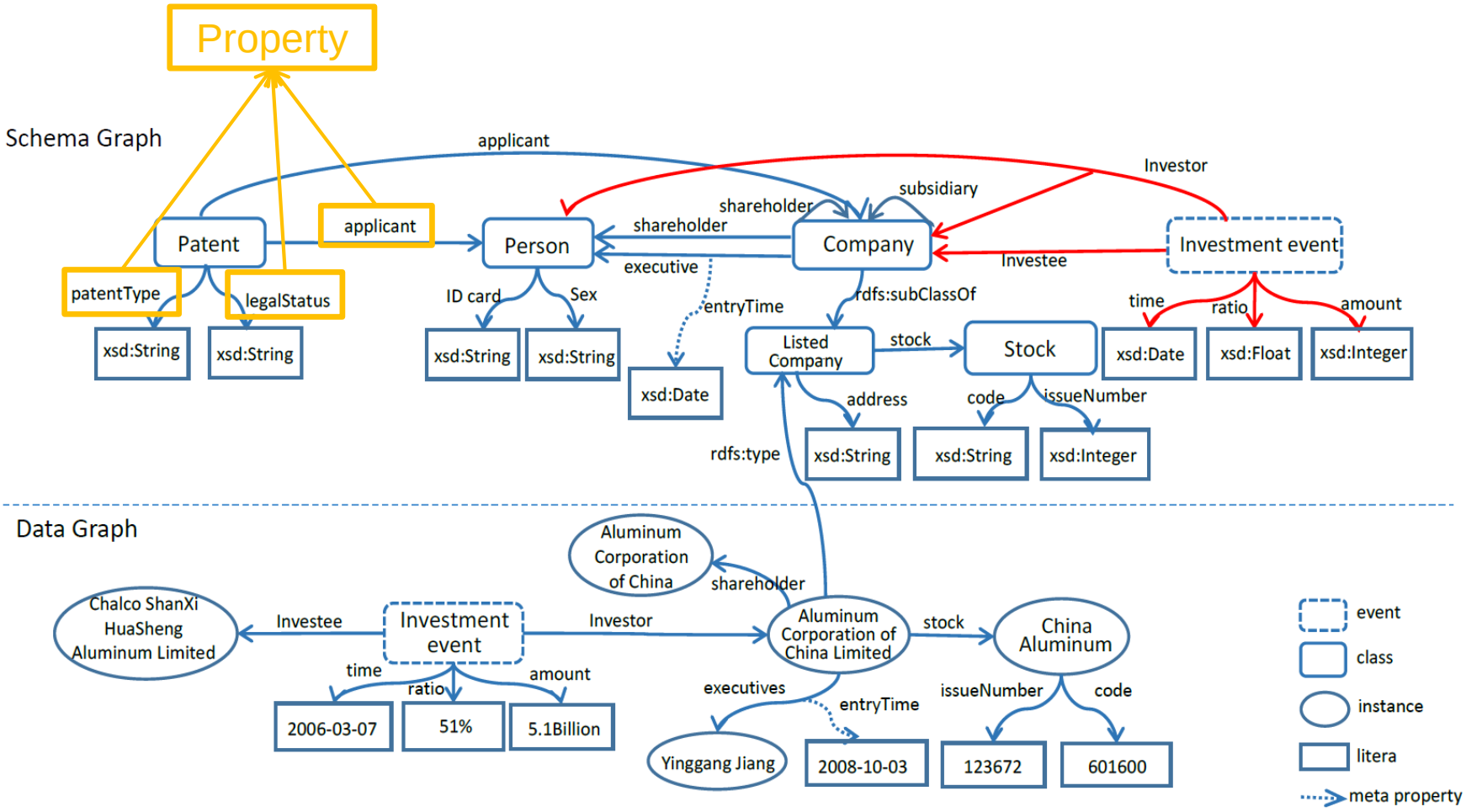
What is an EKG



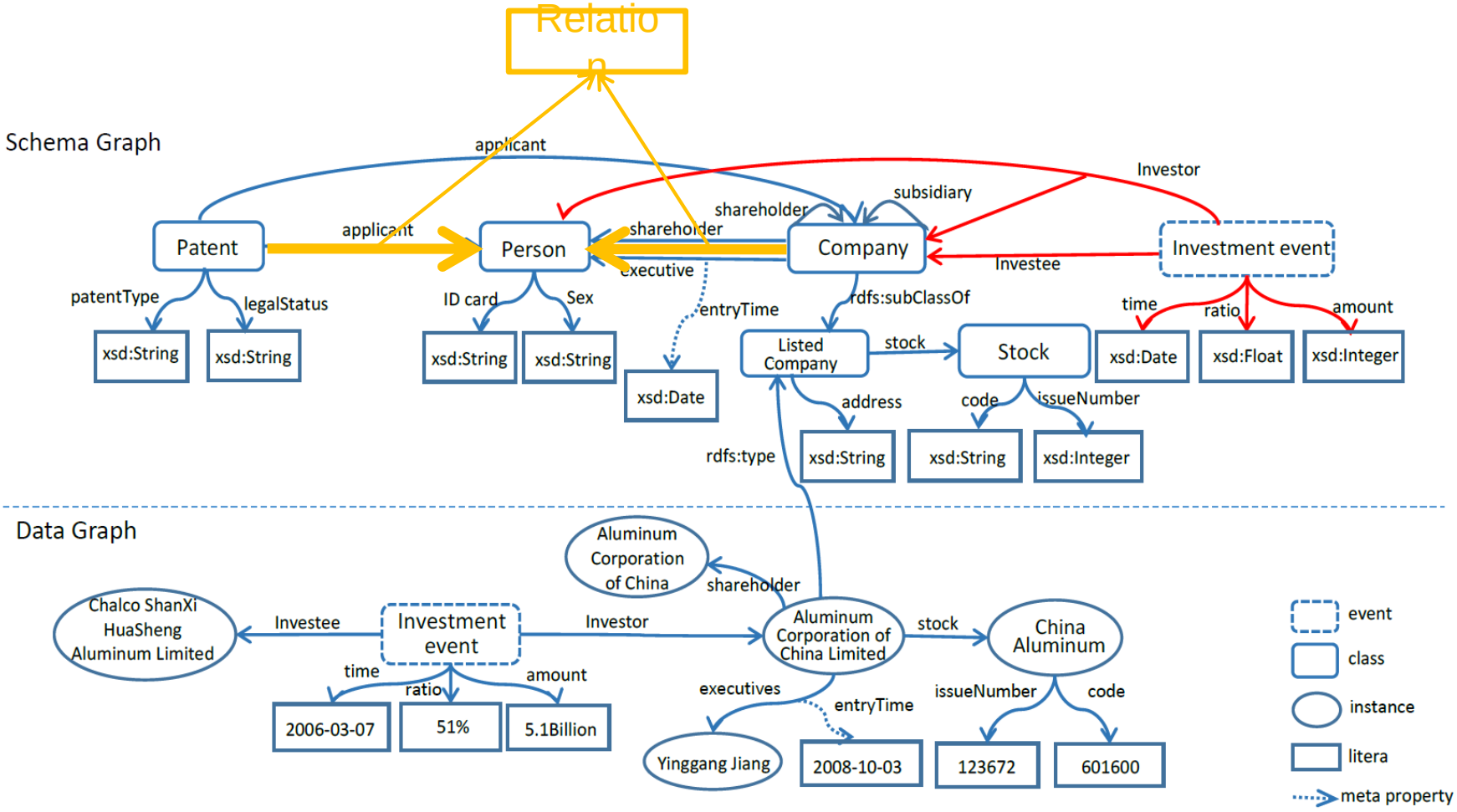
What is an EKG



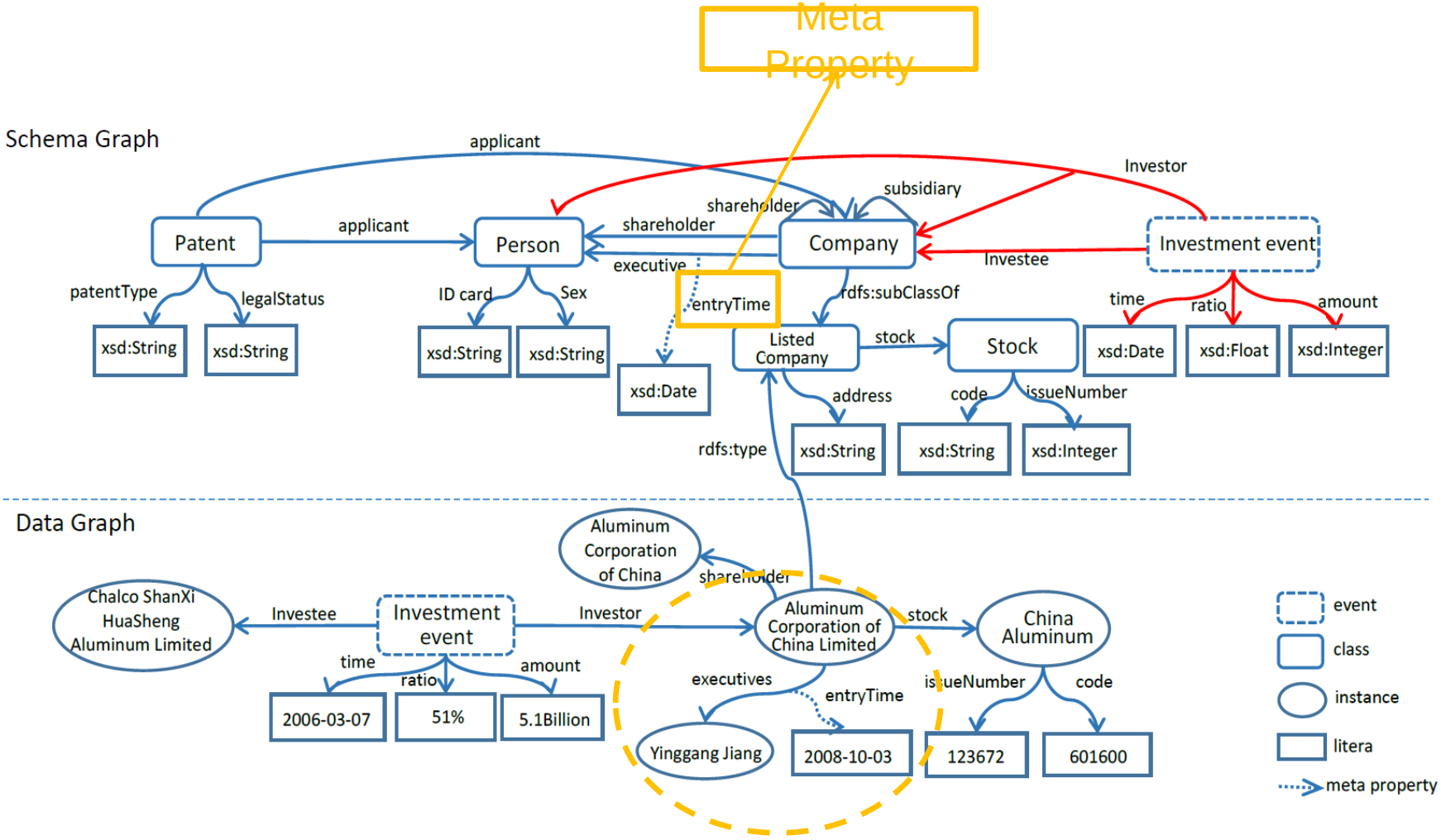
What is an EKG



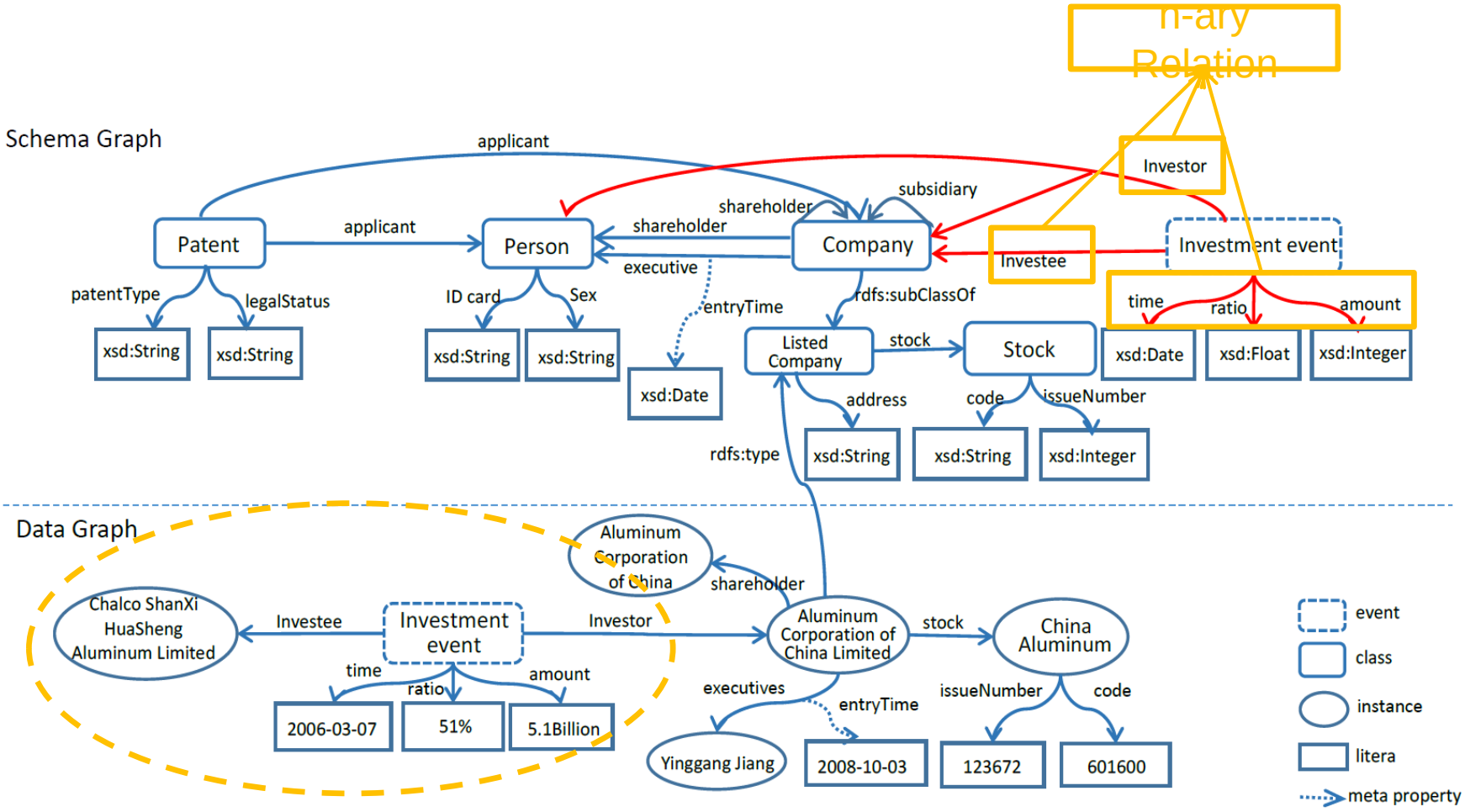
What is an EKG



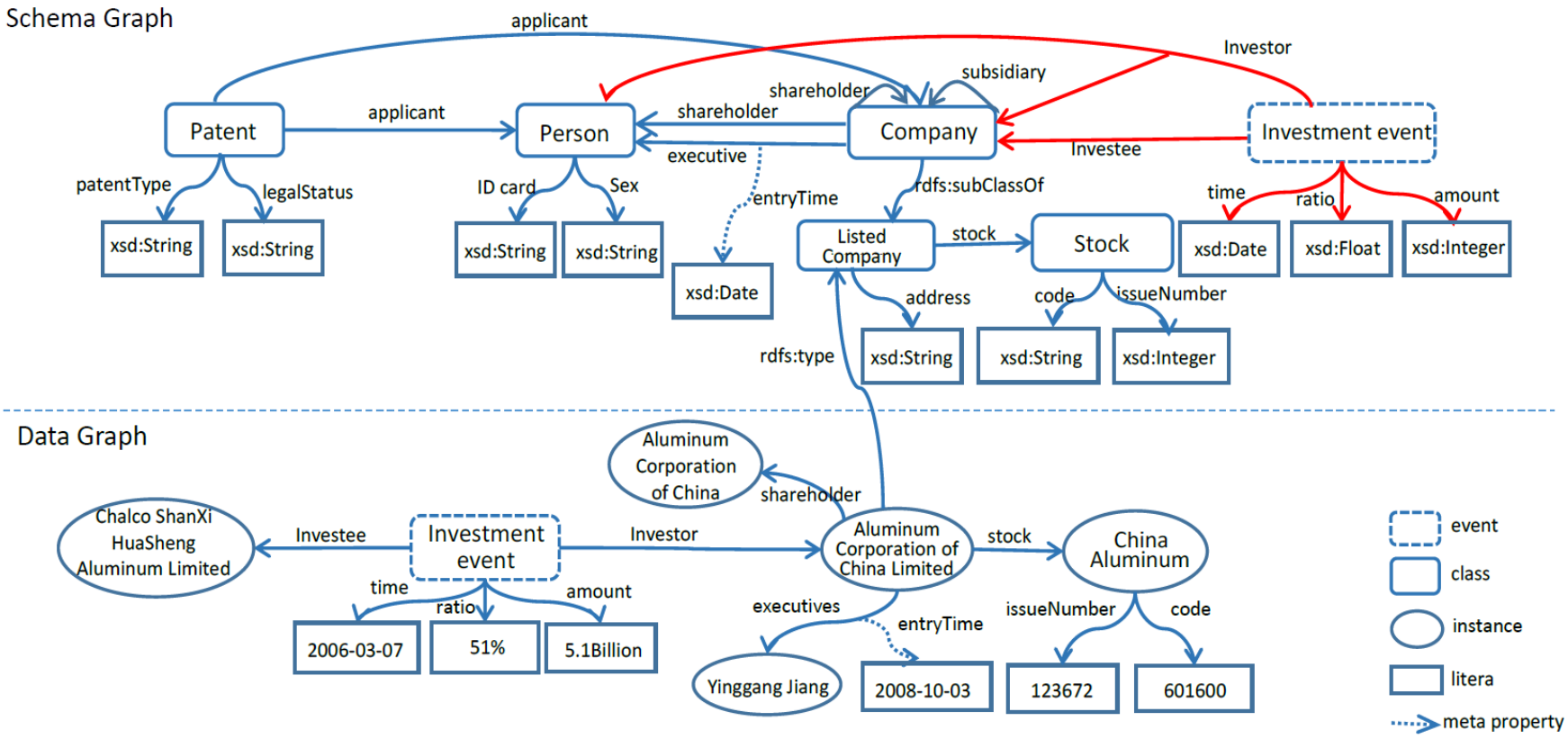
What is an EKG



What is an EKG



What is an EKG



Sources

We have four major sources for constructing our EKG

CSAIC

In formation of 40 million enterprises, 60 million people, 8 million litigation and 1 million credit

→ **Basic EKG**

专利检索及分析
Patent Search and Analysis of SIPO

5million patent information

→ **Patent KG**

中国政府采购网 www.ccgp.gov.cn

CGPN

cninf 巨潮资讯

Information of Listed company

互动百科 baike.com

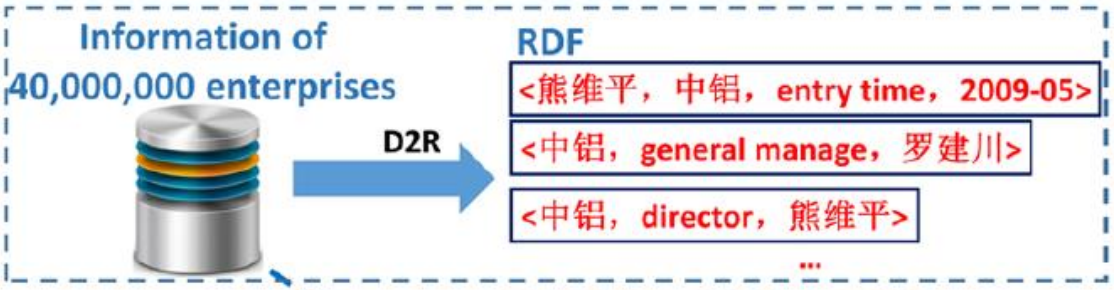
Baidu 百科

维基百科 自由的百科全书

Compete info
Merge event

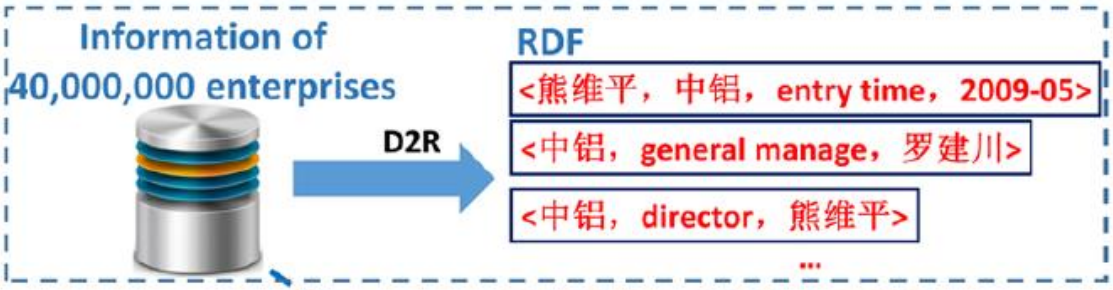
Sources

Sources



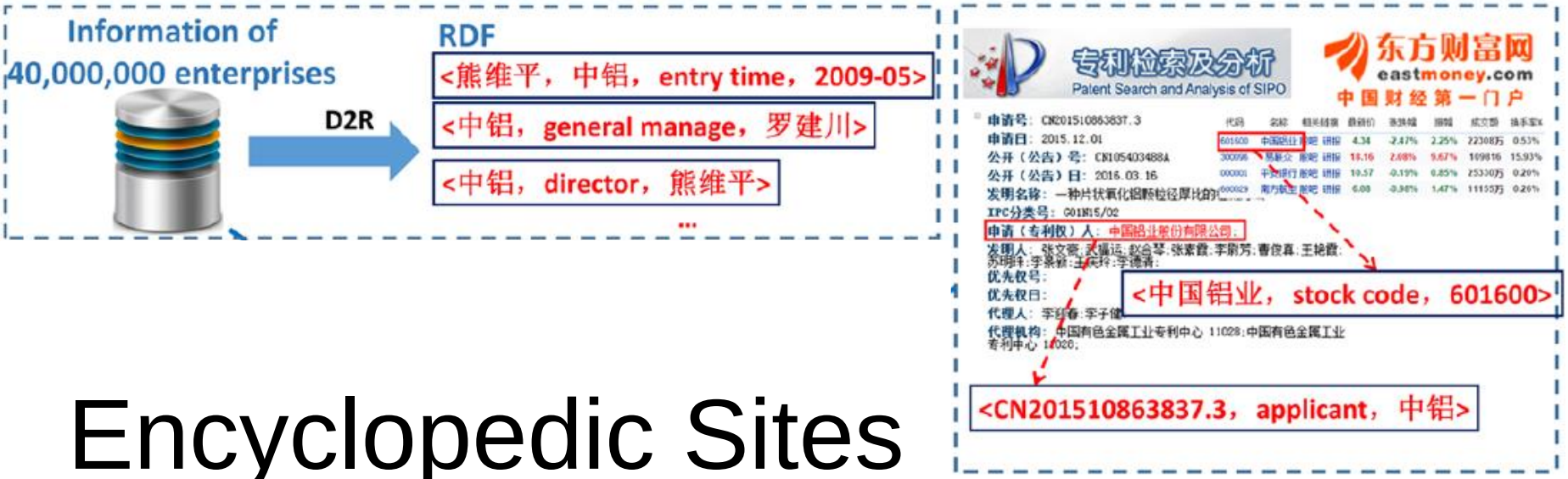
CSAIC

Sources



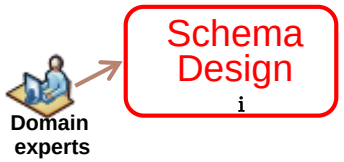
PSAN-SIPO

Sources

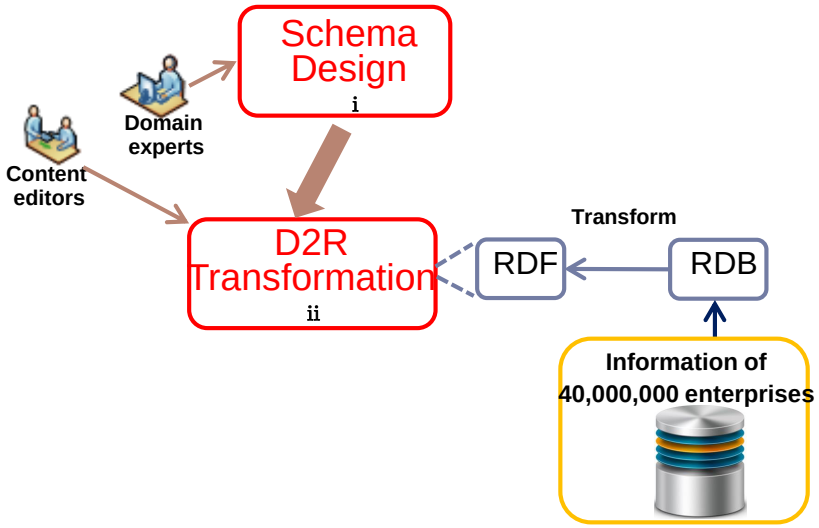

Information of
40,000,000 enterprises
D2R
RDF
<熊维平，中铝，entry time，2009-05>
<中铝，general manage，罗建川>
<中铝，director，熊维平>
...
专利检索及分析
Patent Search and Analysis of SIPO
东方财富网
eastmoney.com
中国财经第一门户
申请号：CN201510863837.3
申请日：2015.12.01
公开（公告）号：CN105403488A
公开（公告）日：2015.03.16
发明名称：一种片状氧化铝颗粒径厚比的
IPC分类号：G01N15/02
申请（专利权）人：中国铝业股份有限公司
发明人：张文亮;关福运;赵合琴;张素霞;李丽芳;曹俊真;王艳霞;
优先权号：
优先权日：
代理人：
代理机构：中国有色金属工业专利中心 11028;中国有色金属工业
专利中心 11028;
<中国铝业，stock code，601600>
<CN201510863837.3，applicant，中铝>
Encyclopedic Sites
Baidu百科
Infobox
中国铝业股份
CHALCO
公司类型 中国中央企业
董事长 葛红林
总裁 余德辉（总经理）
首席执行官 葛红林
总部地点 中国北京
产业 铝业
产品 铝、铜、铅、钛锆资及开发
网址 中国铝业股份有限公司
<中铝，chairman，葛红林>
conflicts
<中铝，chairman，熊维平>
Entity
Linking
Text
中国铝业股份有限公司（简称中国铝业、中铝
（港交所：2600、NYSE：ACH、上交
所：601600），是中国唯一的氧化铝生产商，
熊维平：董事长 首席执行官
熊维平，1956年11月生，现任中国铝业公司总经理、党组书记，中国铝业股份有限公司
董事长、首席执行官，中国铜业有限公司董事长，中国稀有稀土有限公司董事长，中铝财务
2007年6月13日中铝斥资8.6亿美元收购了加拿大秘鲁铜业公司
（Peru Coppe Inc.）的全部股权以及特罗莫克铜矿
（Toromocho）的开发权。
<中国铝业股份有限公司，
abbreviation，中铝>
<2007-06-13，中铝，acquire，Peru Copper Inc>

Data-driven KG constructing process

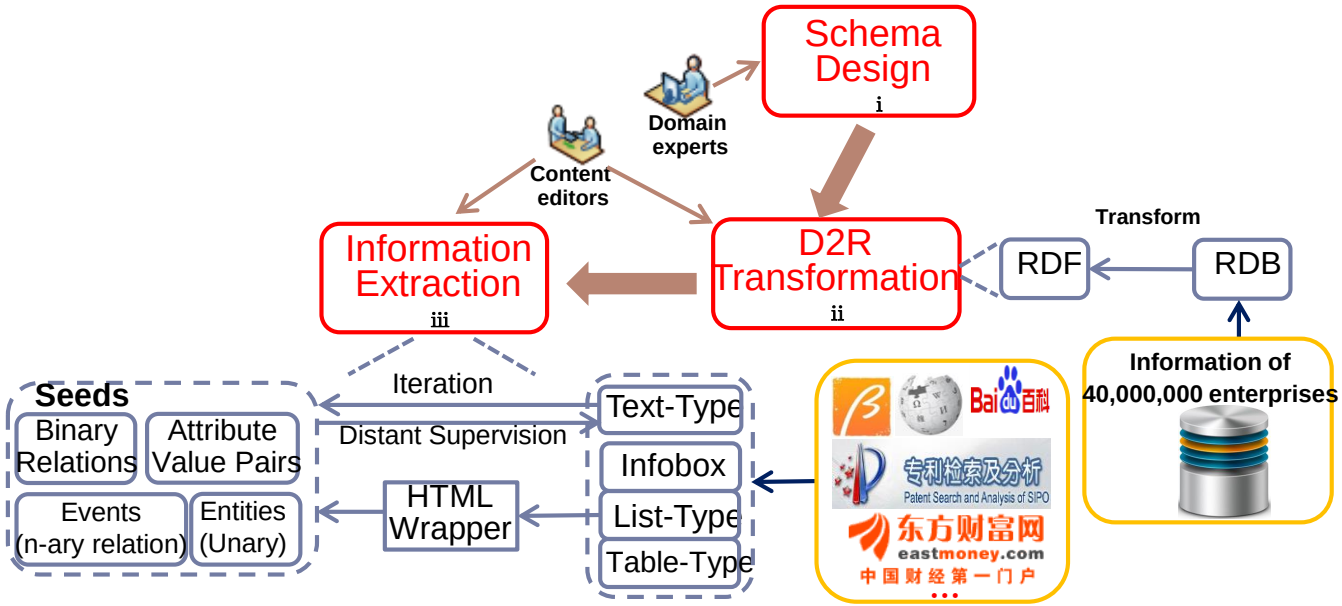
Data-driven KG constructing process



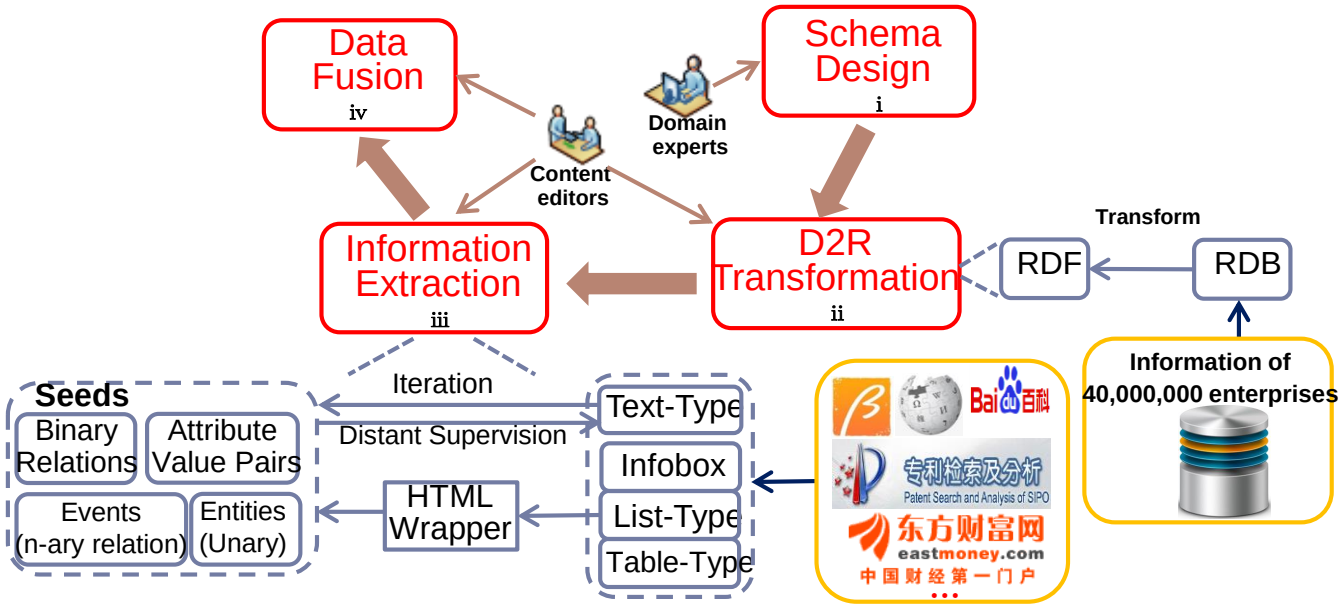
Data-driven KG constructing process



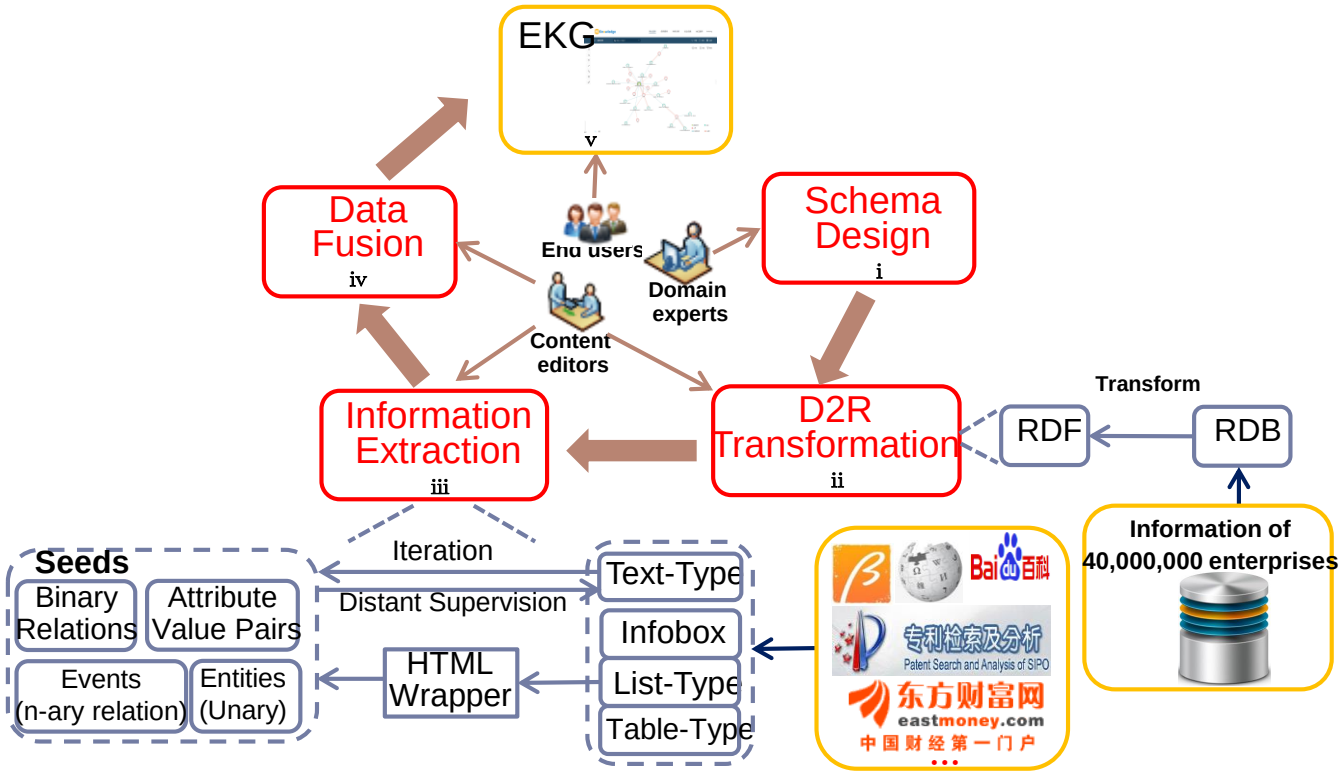
Data-driven KG constructing process



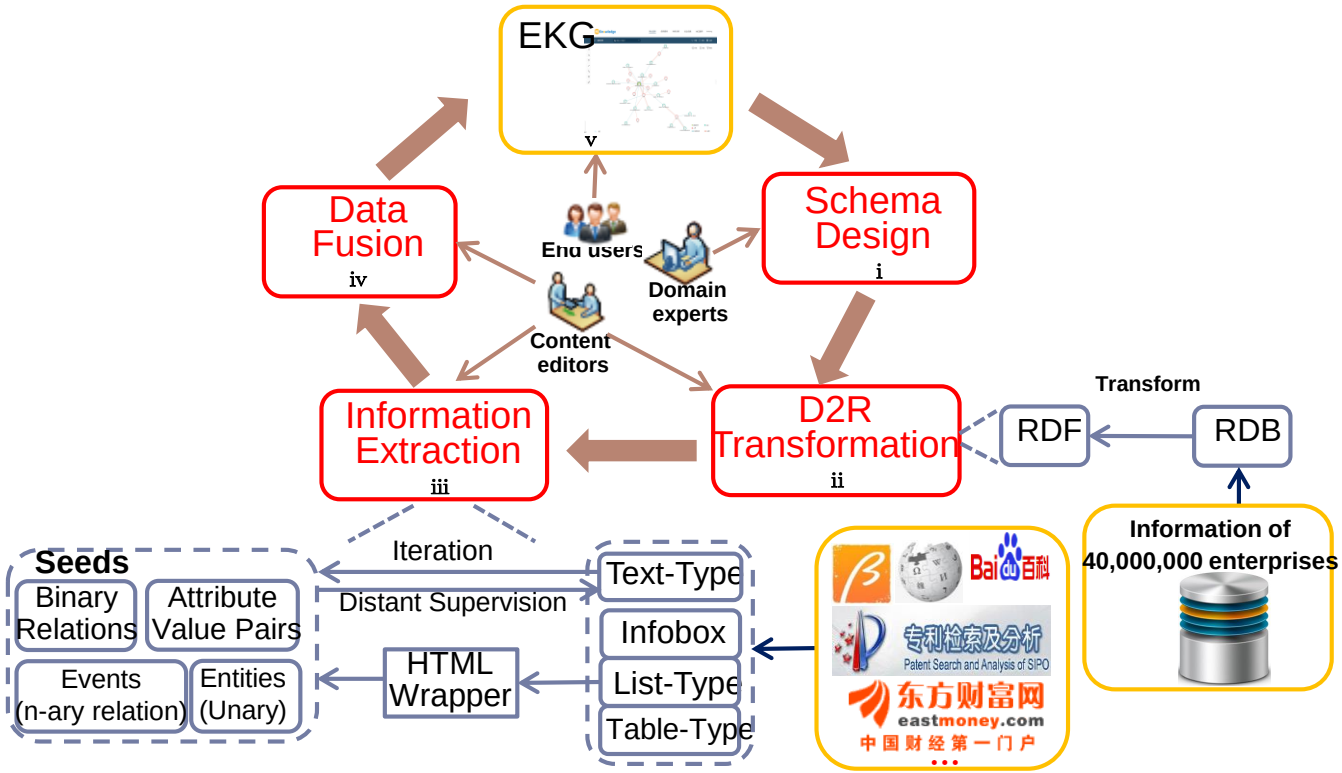
Data-driven KG constructing process



Data-driven KG constructing process



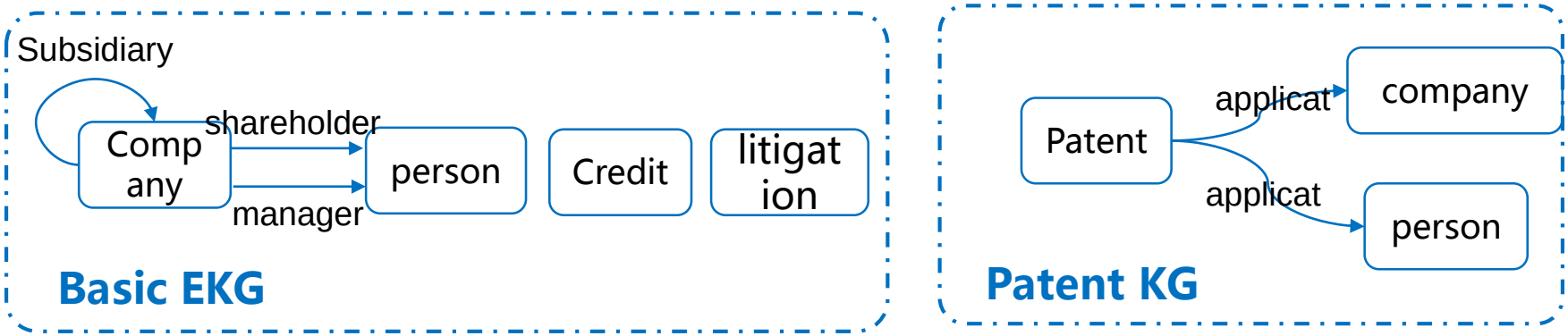
Data-driven KG constructing process



Schema Design

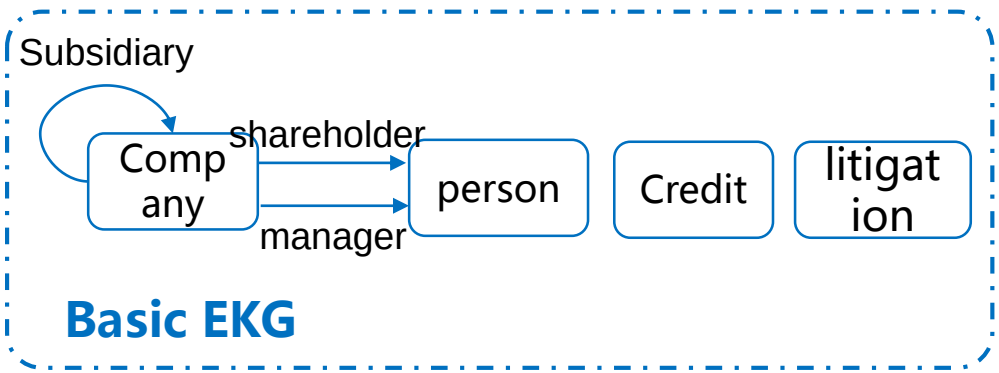
Schema Design

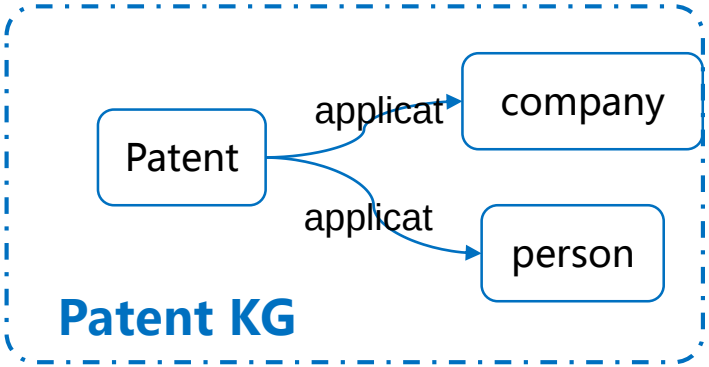
The 1st Iteration:



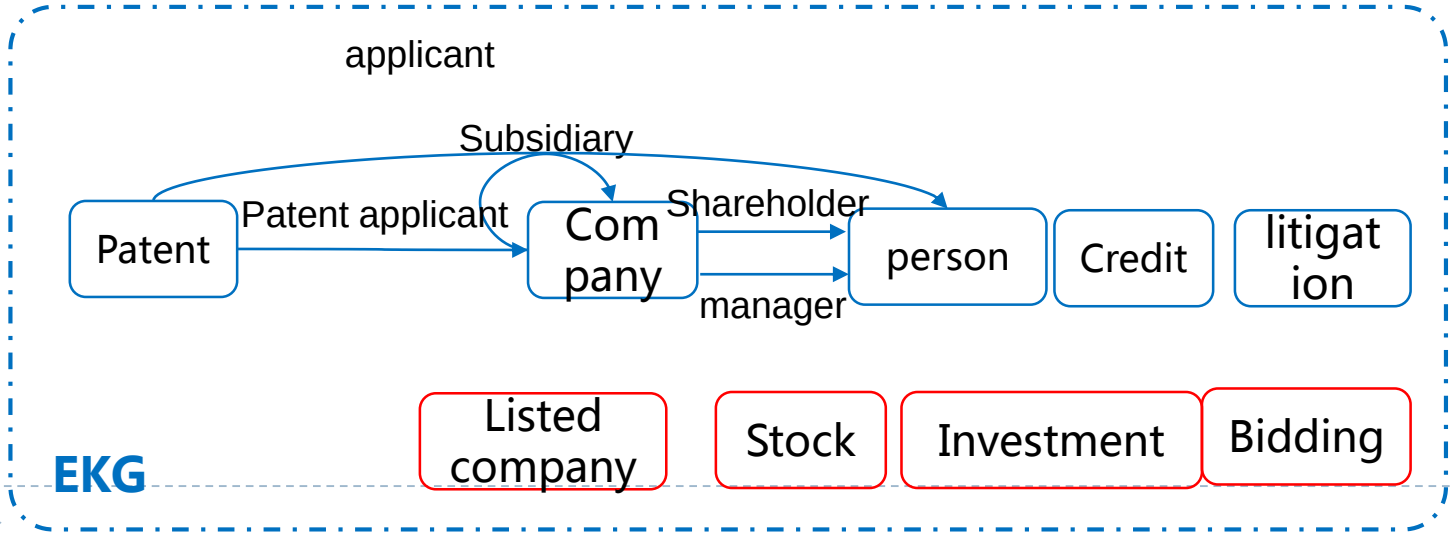
Schema Design

The 1st Iteration:





the 2nd Iteration



D2R Transformation

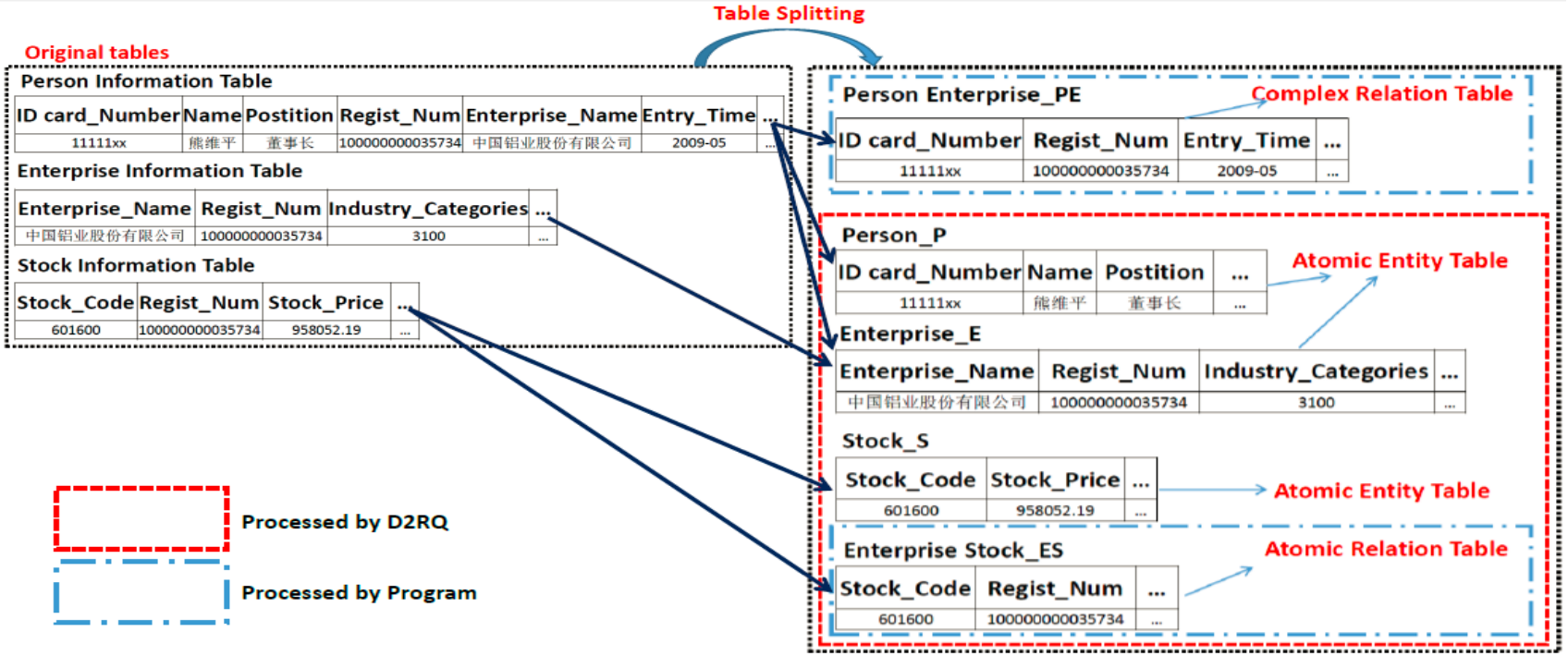
Difficulties:

- Table is lack of standardization
- Meta-property mapping
- Existing D2R tools can not solve complex mapping relationship

Solution:

- Table splitting
- Basic D2R transformation by D2RQ
- Post processing

Table splitting



D2R Transformation

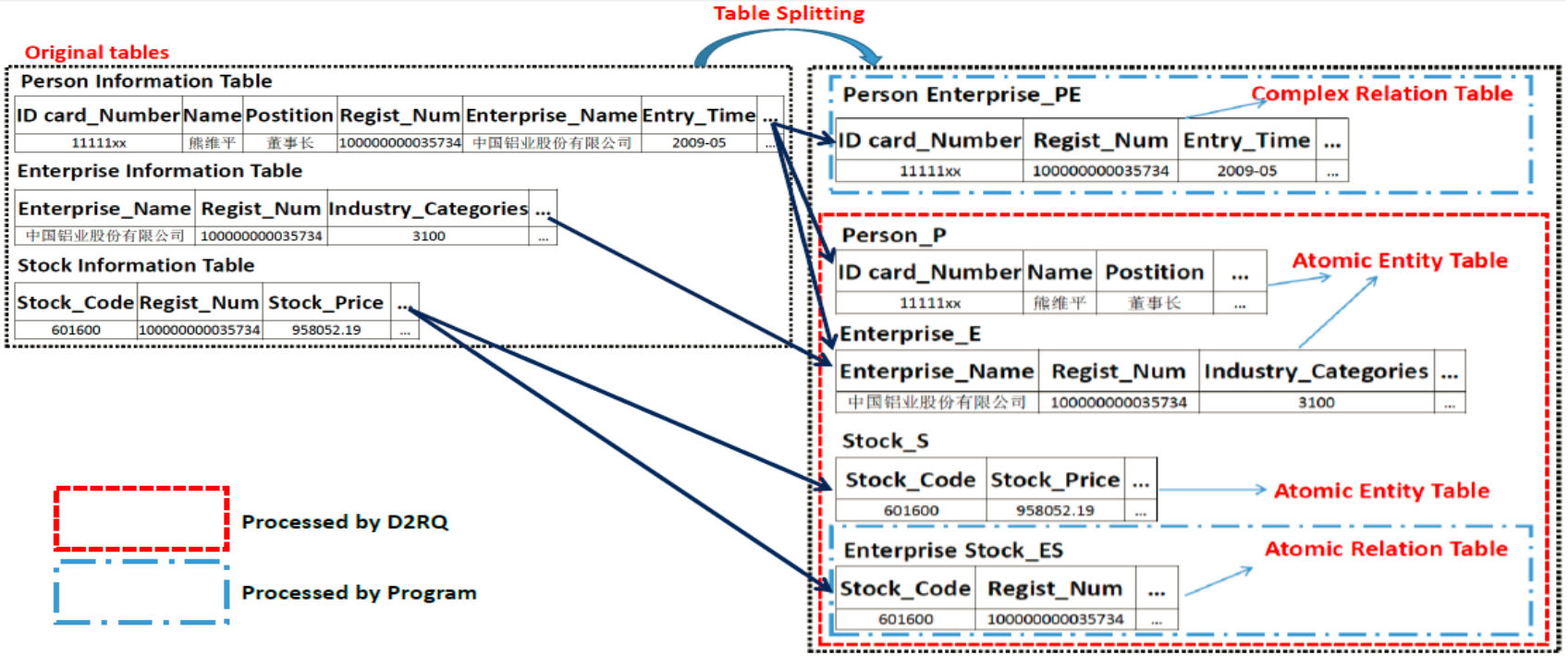
Difficulties:

- Table is lack of standardization
- Meta-property mapping
- Existing D2R tools can not solve complex mapping relationship

Solution:

- Table splitting
- Basic D2R transformation by D2RQ
- Post processing

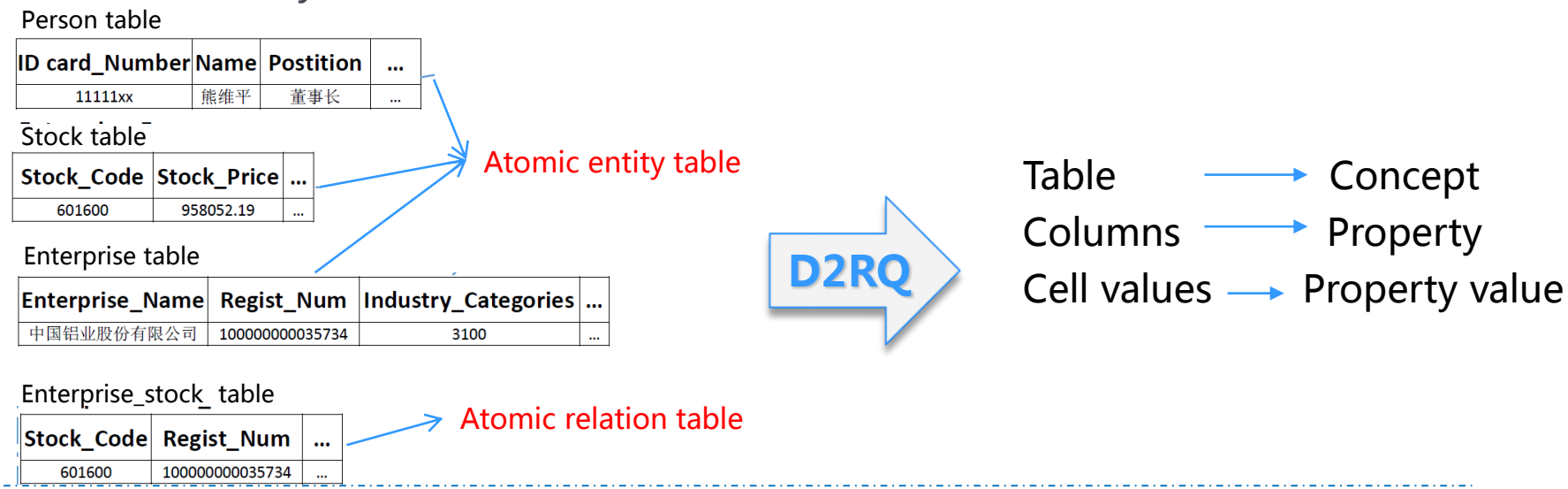
Table splitting



D2R Transformation

Basic D2R transformation by D2RQ

- We write a customized mapping file in D2RQ to map fields related to atomic entity tables and atomic relation tables into RDF format.

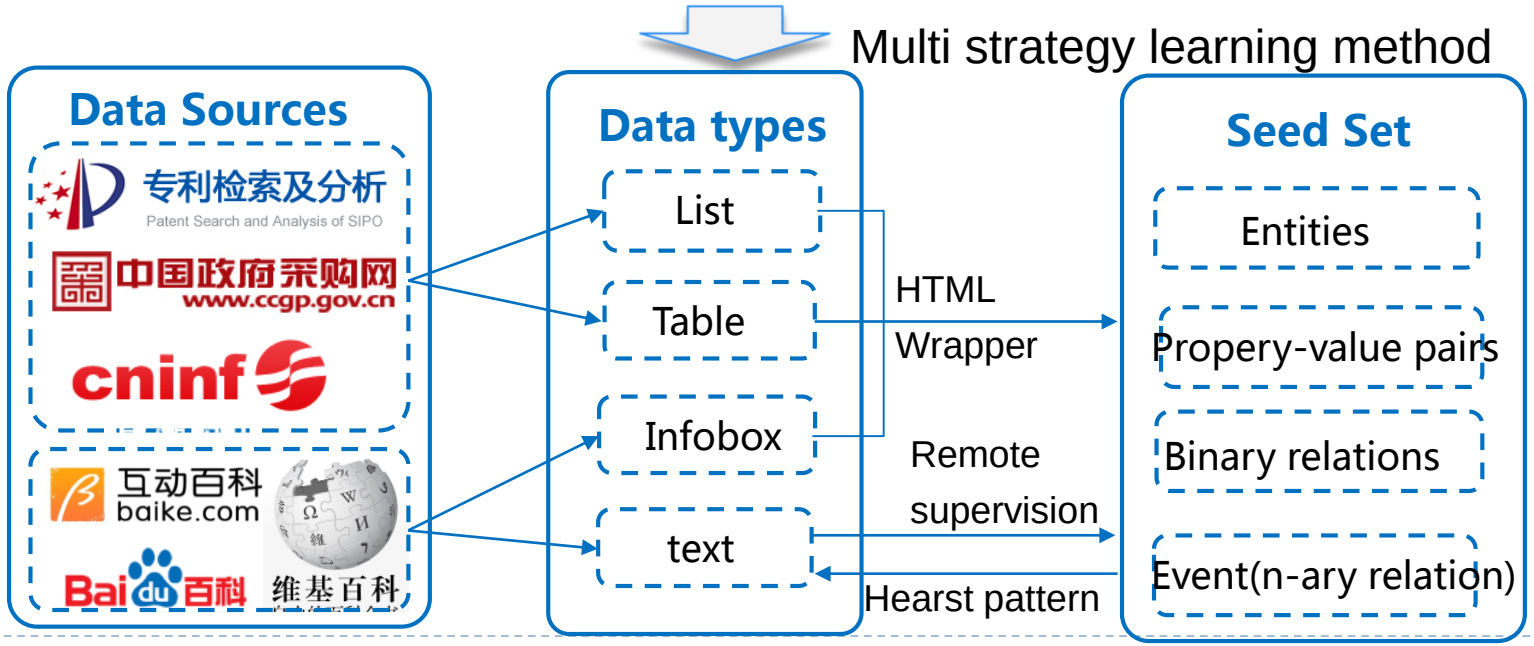


Post processing

- For the complex relationship tables, and complex entity tables:
  - Meta-property mapping
  - Conditional taxonomy mapping
  - Conditional class mapping

Information Extraction

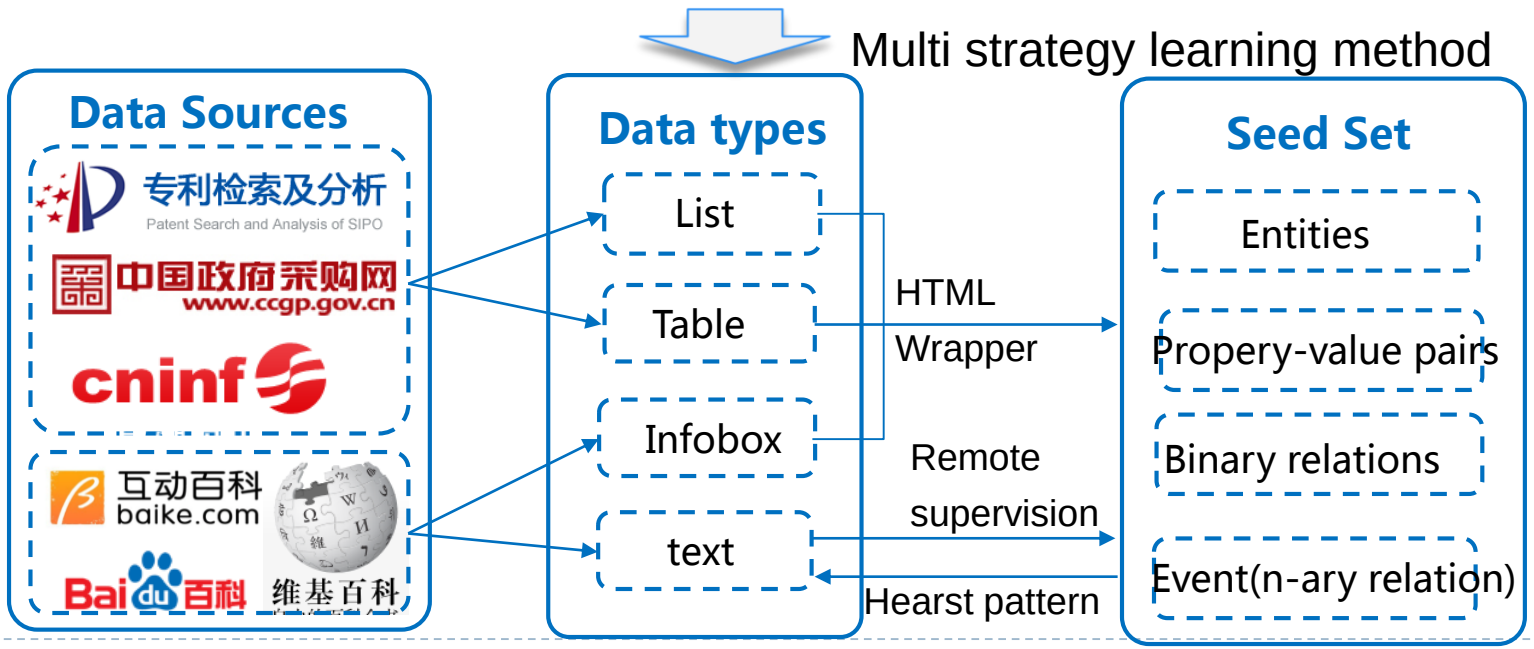
- Difficulties:



@Tong Ruan, Lijuan Xue: Bootstrapping Yahoo! Finance by Wikipedia for Competitor Mining. JIST 2015

Information Extraction

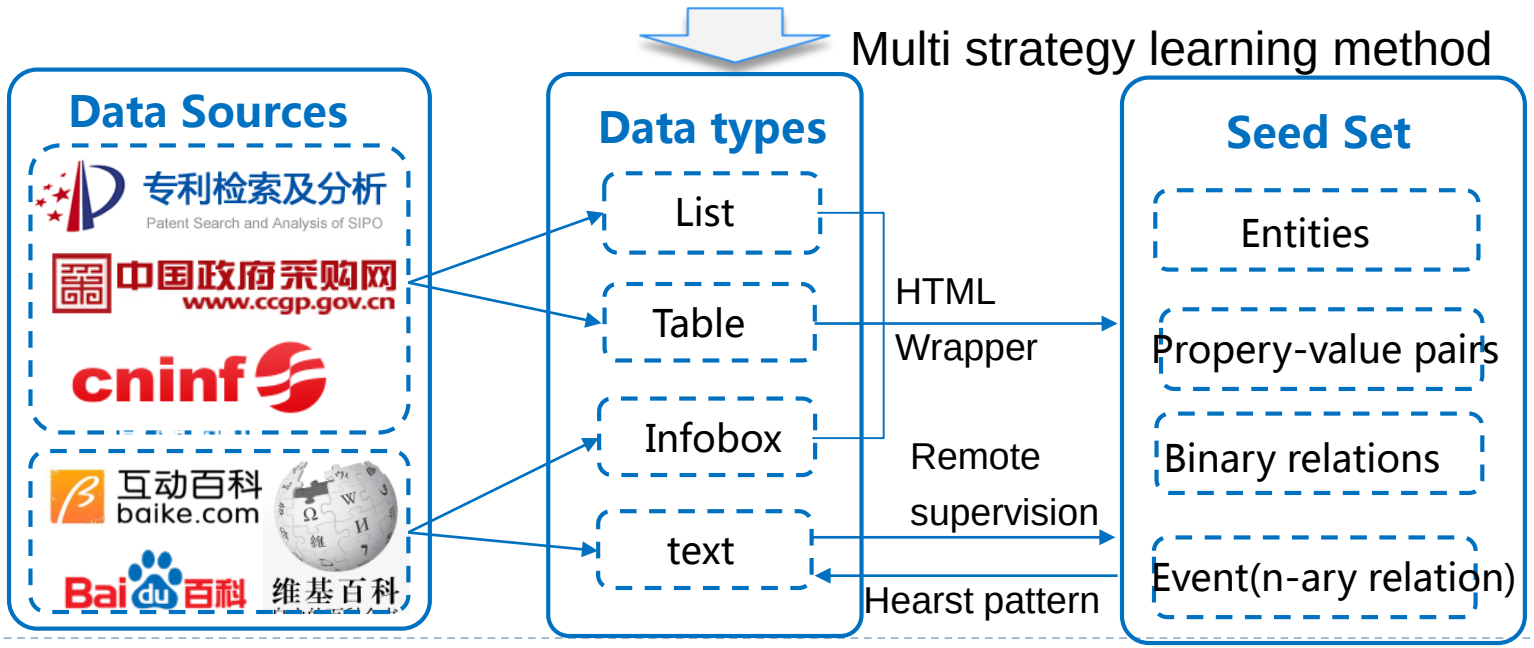
- Difficulties:
  - Multiple data sources



@Tong Ruan, Lijuan Xue: Bootstrapping Yahoo! Finance by Wikipedia for Competitor Mining. JIST 2015

Information Extraction

- Difficulties:
  - Multiple data sources
  - Various of target data types:

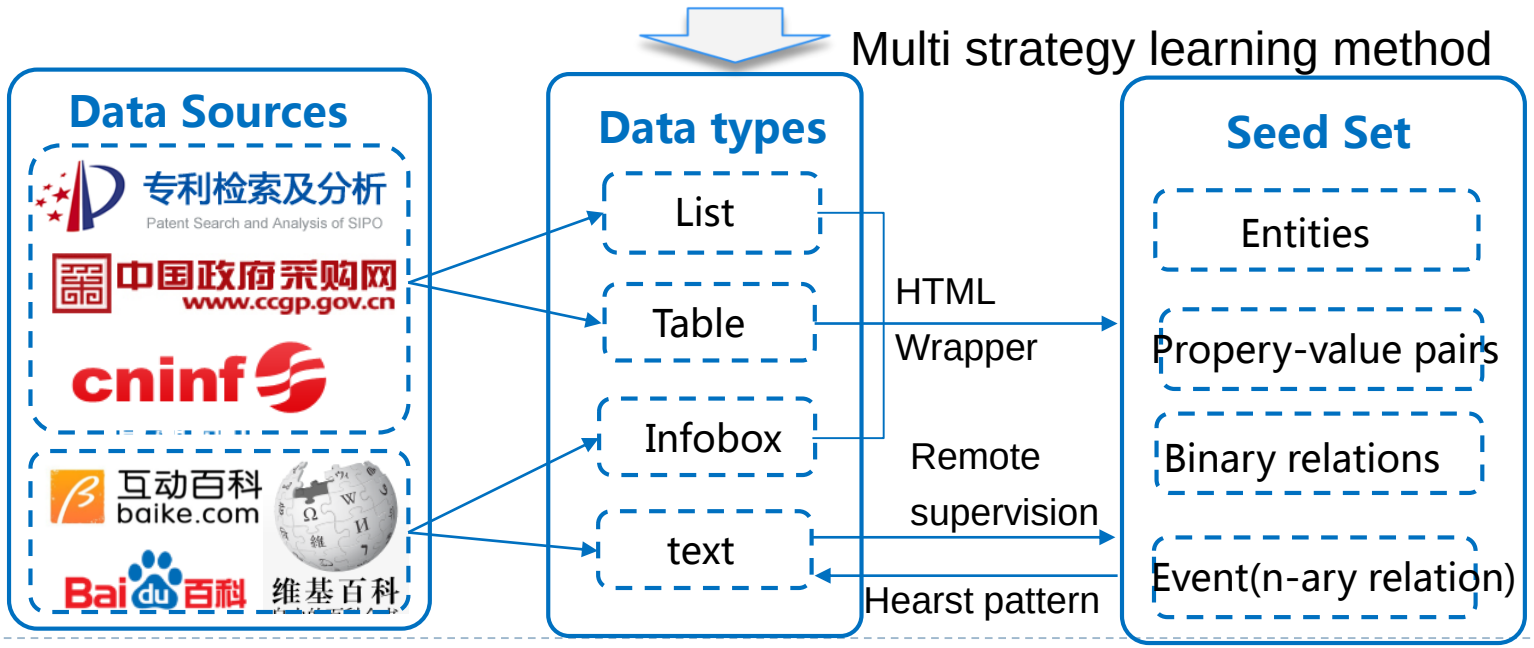


@Tong Ruan, Lijuan Xue: Bootstrapping Yahoo! Finance by Wikipedia for Competitor Mining. JIST 2015

Information Extraction

- Difficulties:
  - Multiple data sources
  - Various of target data types:

    Different types of entities(Map, List, Range……)



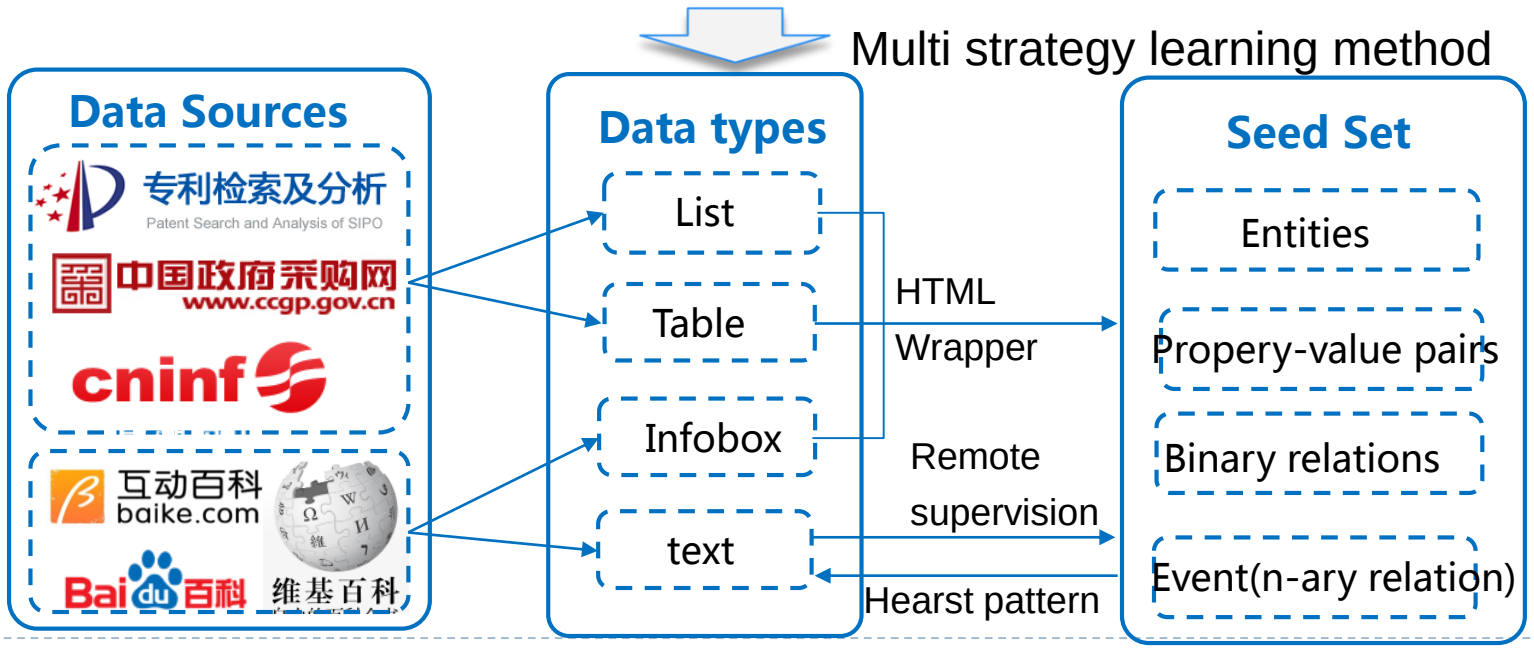
@Tong Ruan, Lijuan Xue: Bootstrapping Yahoo! Finance by Wikipedia for Competitor Mining. JIST 2015

Information Extraction

- Difficulties:
  - Multiple data sources
  - Various of target data types:

    Different types of entities(Map, List, Range……)

    binary relations, attribute value pairs,



@Tong Ruan, Lijuan Xue: Bootstrapping Yahoo! Finance by Wikipedia for Competitor Mining. JIST 2015

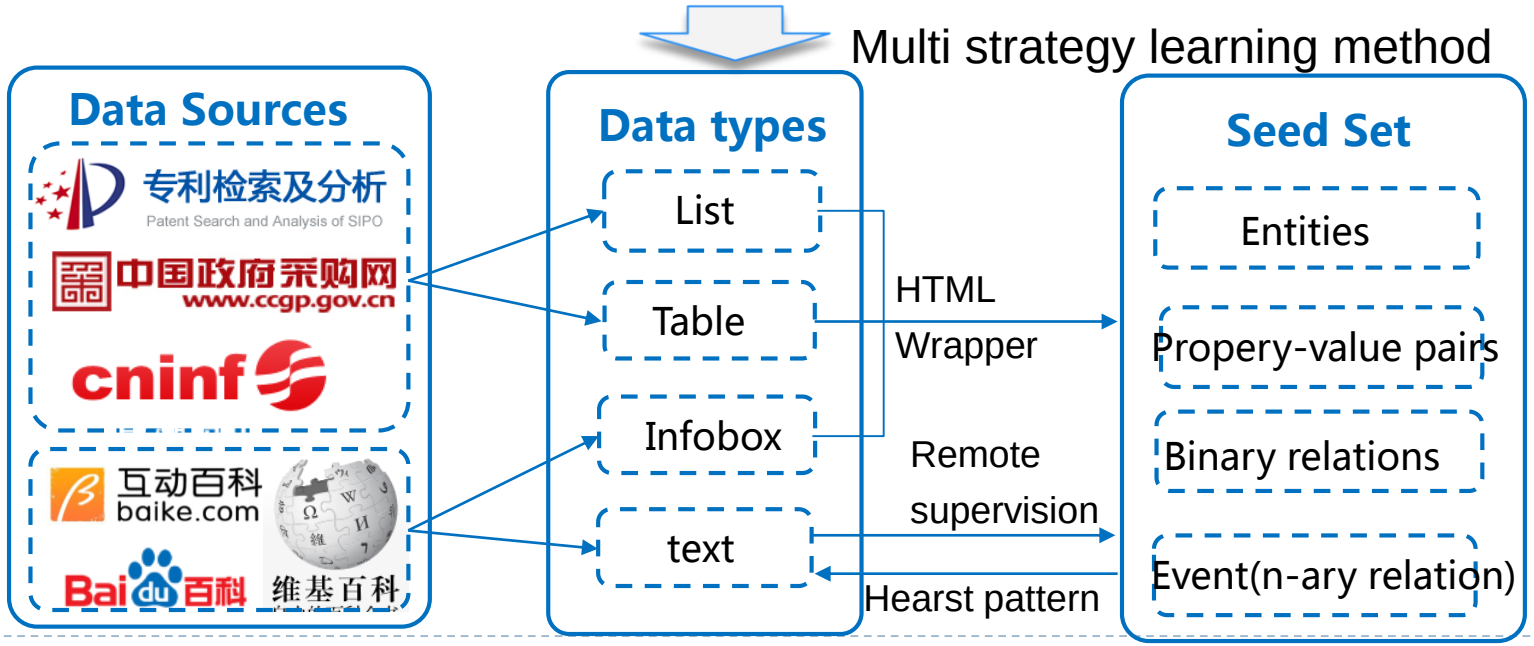
Information Extraction

- Difficulties:
  - Multiple data sources
  - Various of target data types:

    Different types of entities(Map, List, Range……)

    binary relations, attribute value pairs,

    event(n-ary relation), synonym extraction



@Tong Ruan, Lijuan Xue: Bootstrapping Yahoo! Finance by Wikipedia for Competitor Mining. JIST 2015

Data Fusion with Instance Matching

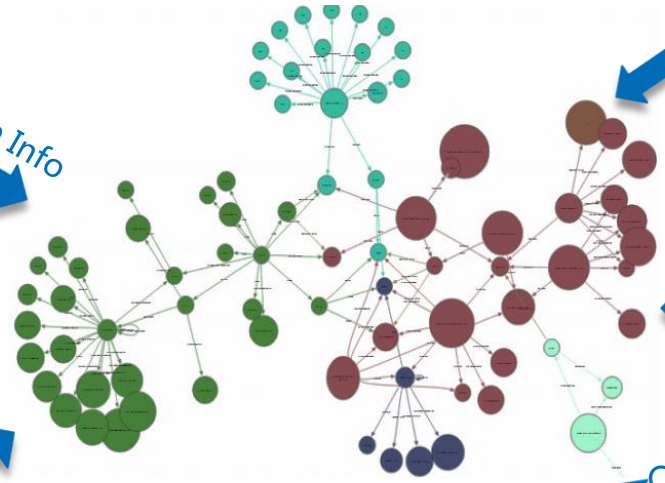
- Entities such as companies, people are aligned
- Data conflict problem

Basic EKG

Enterprise Info

Patent Info

Patent Info



EKG

Bidding Info

Stock info

Competitive relations, acquisition events...

中国政府采购网 www.ccgp.gov.cn

cninf 巨潮资讯

互动百科 baike.com

Baidu 百科

维基百科 自由的百科全书

5 Storage Design and Query Optimization

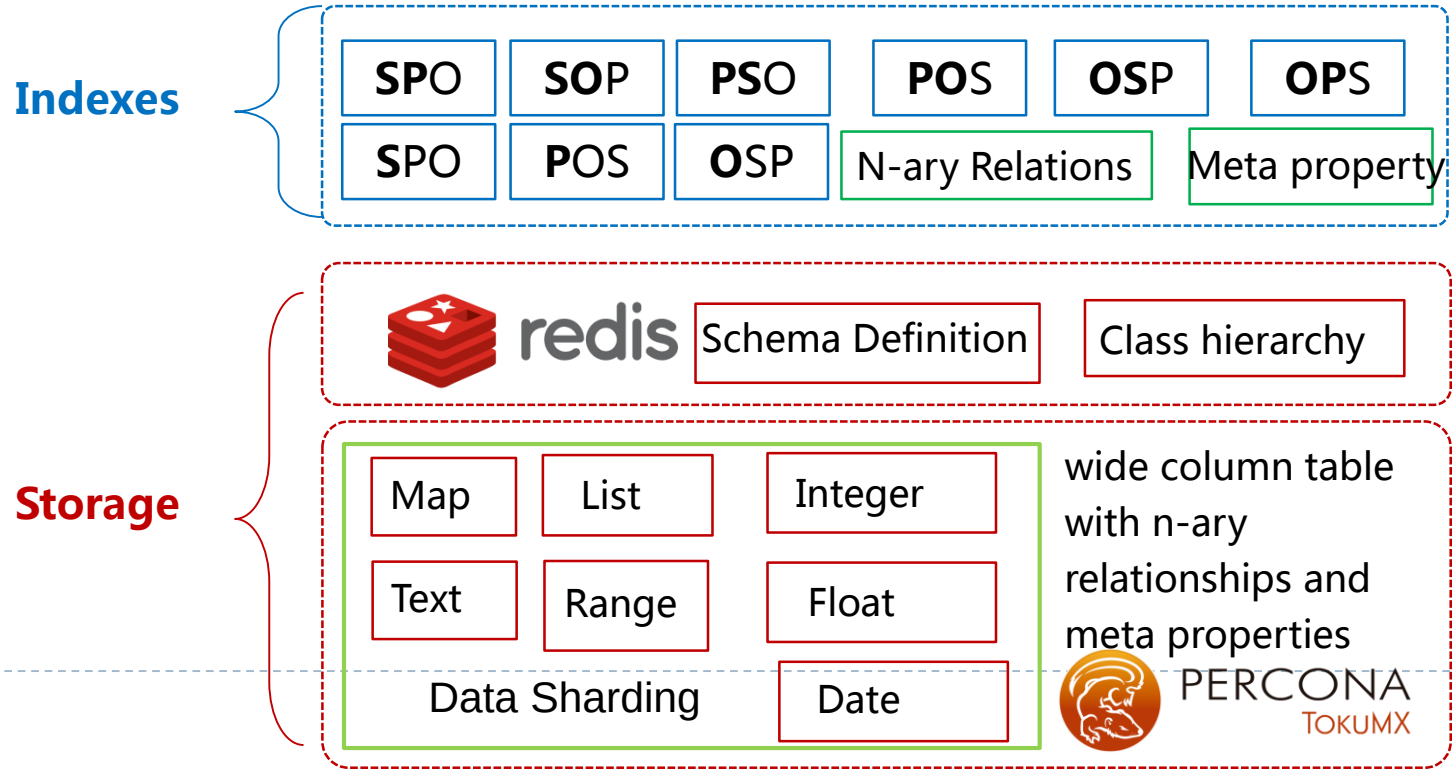
Storage Desgin

- TokuMX+ Redis
- Data Types : List, Map……
- Store n-ary relations in the same row of a table

(Wide column table).

Query optimization

- Nine Indexes and indexes on meta properties and
- Cache schema Data in Redis
- Data Sharding for different data type of the property value
- Support query on n-ary relation efficiently



5 Storage Design and Query Optimization

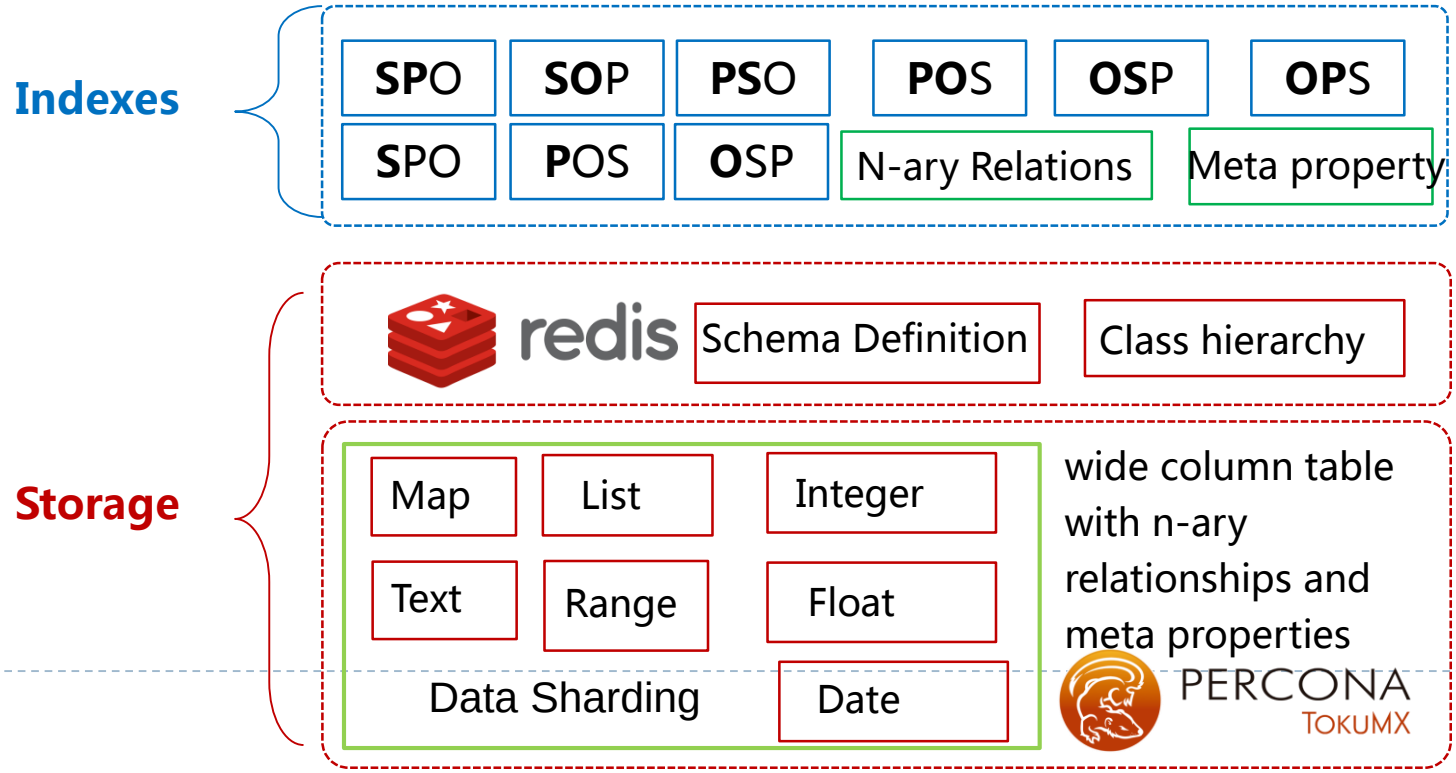
Storage Desgin

- TokuMX+ Redis
- Data Types : List, Map……
- Store n-ary relations in the same row of a table

(Wide column table).

Query optimization

- Nine Indexes and indexes on meta properties and
- Cache schema Data in Redis
- Data Sharding for different data type of the property value
- Support query on n-ary relation efficiently



5 Storage Design and Query Optimization

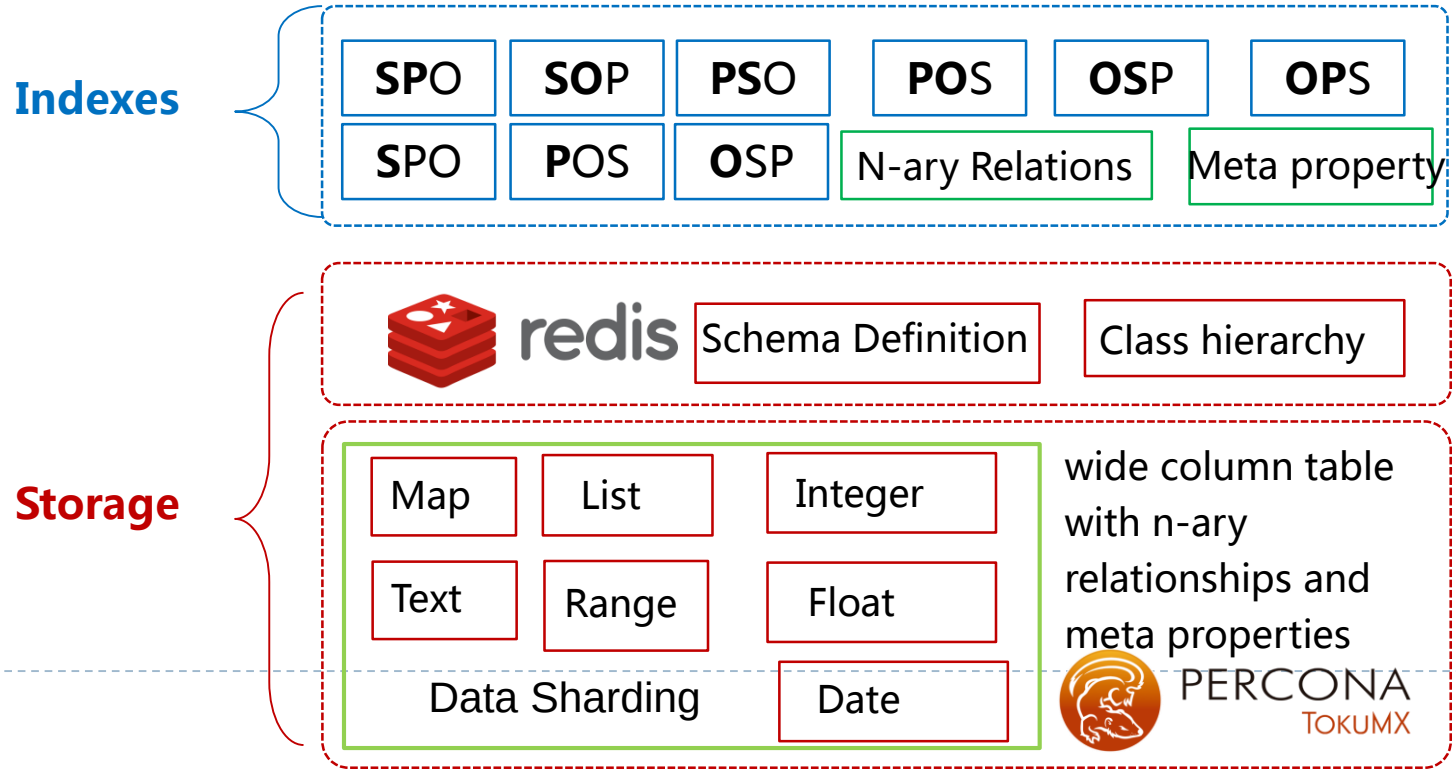
Storage Desgin

- TokuMX+ Redis
- Data Types : List, Map……
- Store n-ary relations in the same row of a table

(Wide column table).

Query optimization

- Nine Indexes and indexes on meta properties and
- Cache schema Data in Redis
- Data Sharding for different data type of the property value
- Support query on n-ary relation efficiently



Usage Scenarios

- An Investor hears about a company on big data area called Long Credit Beijing Corp. Ltd.
- He wants to know the detailed information of the enterprise for the investment decision in the future.
- He uses our “Magic Mirror” to have a look.

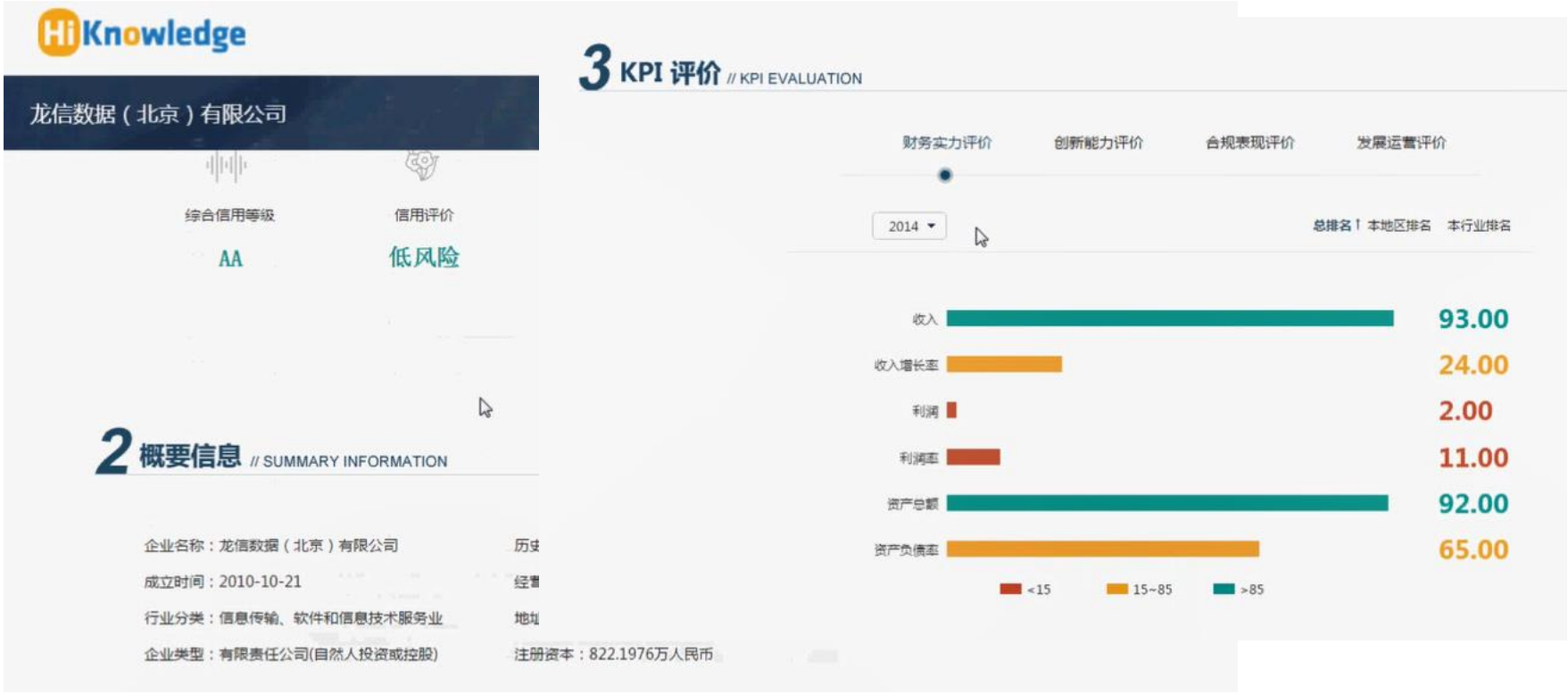
LONG CREDIT 龍信數据

Usage Scenarios (1. General Overview)

- Firstly, the investors have a glance at basic information of Long Credit
- Then they can have a general overview of the financial status and innovation strength of Long Credit with the Key Performance Indicators (KPI) Module.

Usage Scenarios (1. General Overview)

- Firstly, the investors have a glance at basic information of Long Credit
- Then they can have a general overview of the financial status and innovation strength of Long Credit with the Key Performance Indicators (KPI) Module.



Usage Scenarios (1. General Overview)

- Firstly, the investors have a glance at basic information of Long Credit
- Then they can have a general overview of the financial status and innovation strength of Long Credit with the Key Performance Indicators (KPI) Module.

The investors find the general information is satisfactory , then they go into detail about Long Credit with visualized EKG.

Usage Scenarios (2. Graph Query)

Usage Scenarios (2. Graph Query)

Usage Scenarios (2. Graph Query)

Usage Scenarios (2. Graph Query)

Usage Scenarios (2. Graph Query)

Usage Scenarios (2. Graph Query)

Usage Scenarios (2. Graph Query)

Usage Scenarios (2. Graph Query)

Usage Scenarios (2. Graph Query)

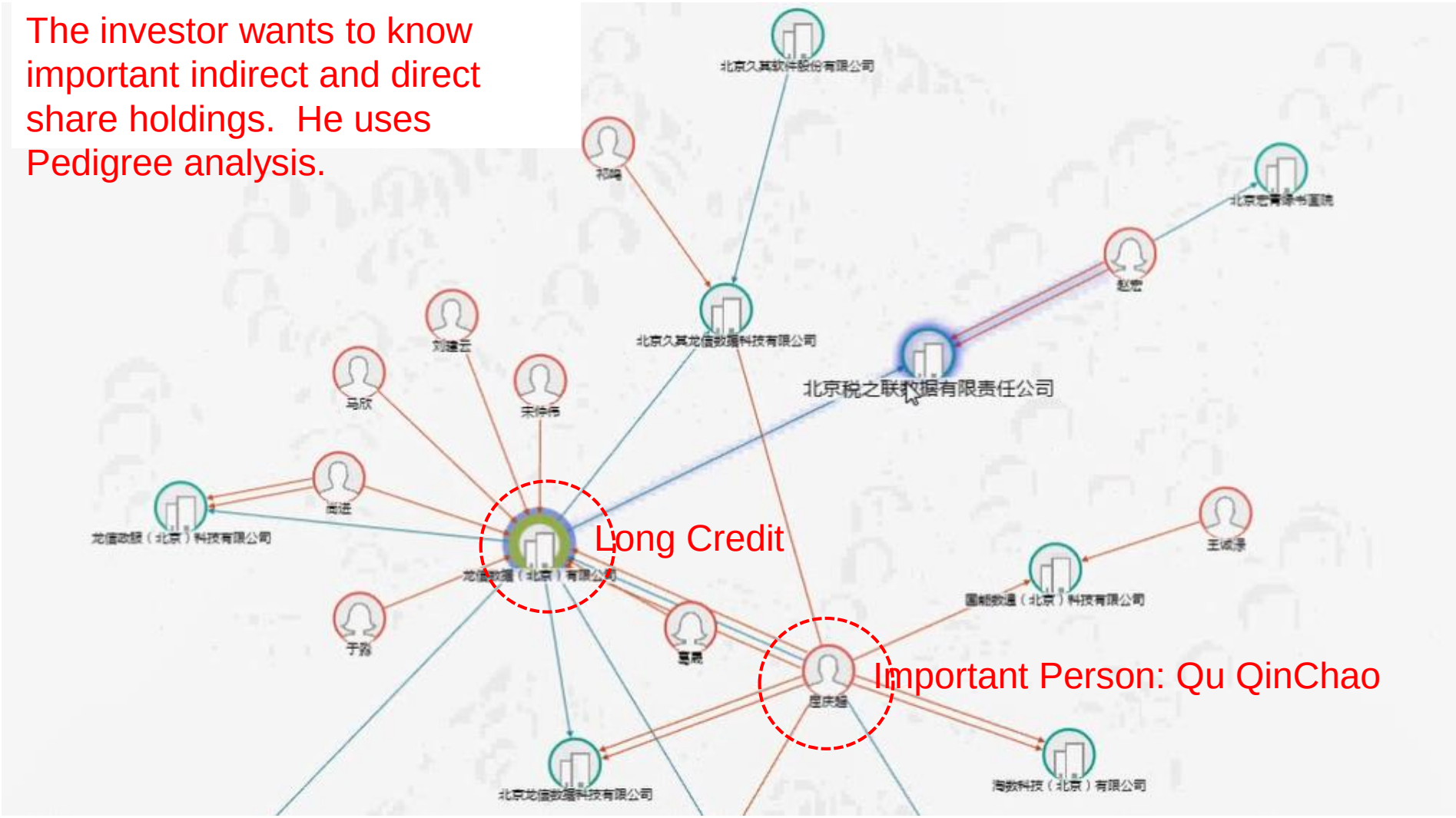
Usage Scenarios (2. Graph Query)

Usage Scenarios (2. Graph Query)

Usage Scenarios (2. Graph Query)

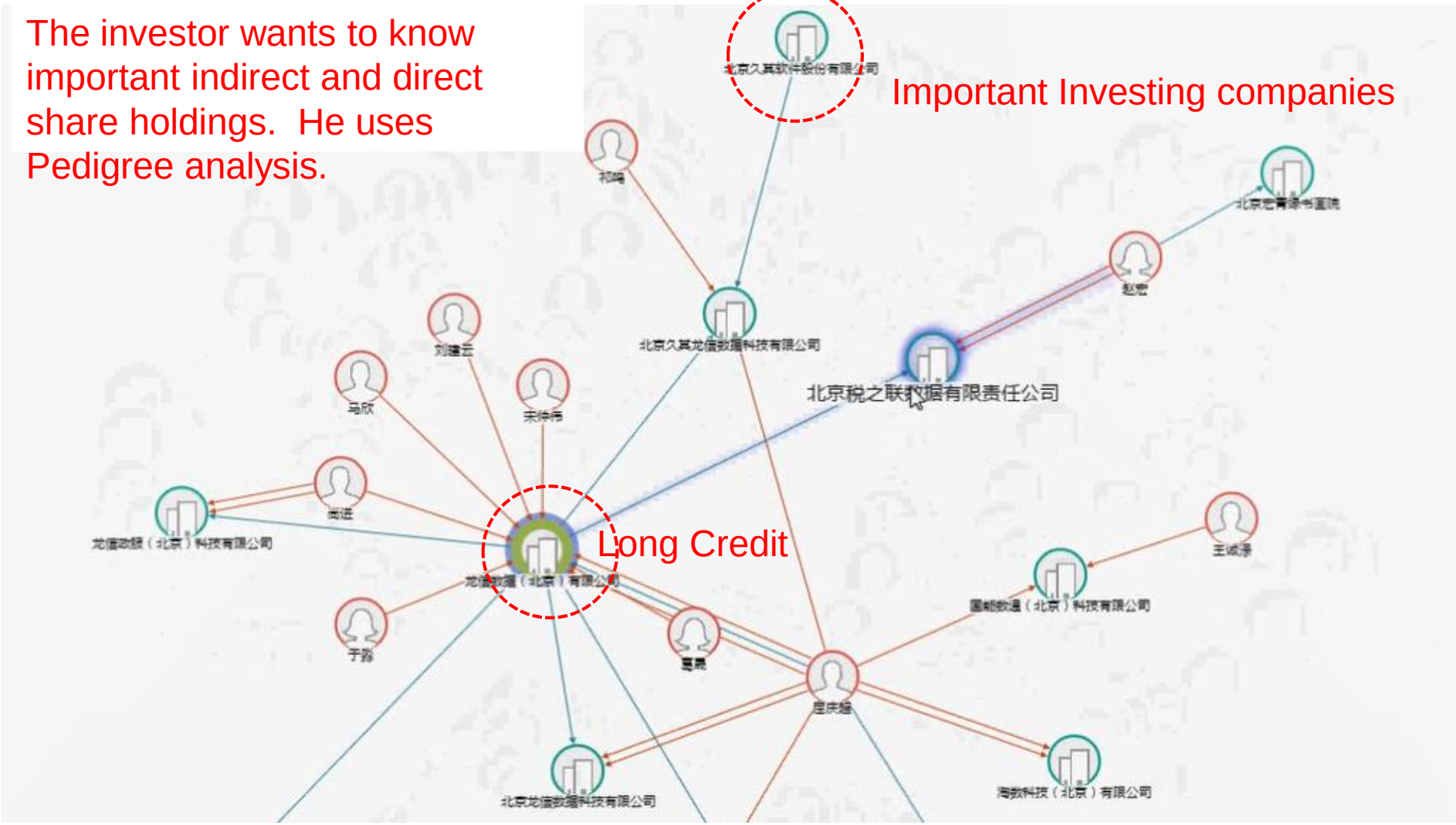
Usage Scenarios (3. Pedigree Analysis)

The investor wants to know important indirect and direct share holdings. He uses Pedigree analysis.



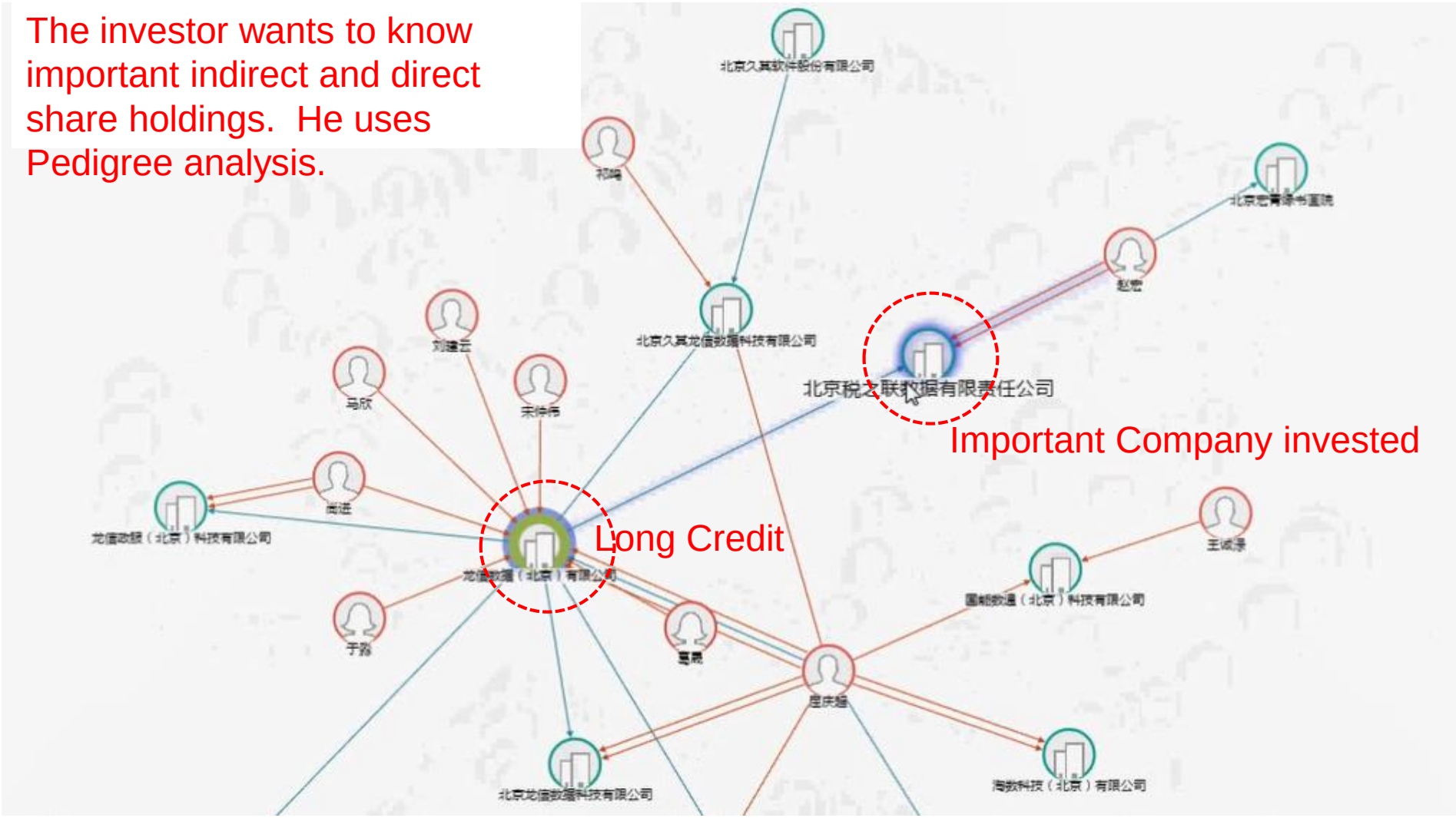
Usage Scenarios (3. Pedigree Analysis)

The investor wants to know important indirect and direct share holdings. He uses Pedigree analysis.

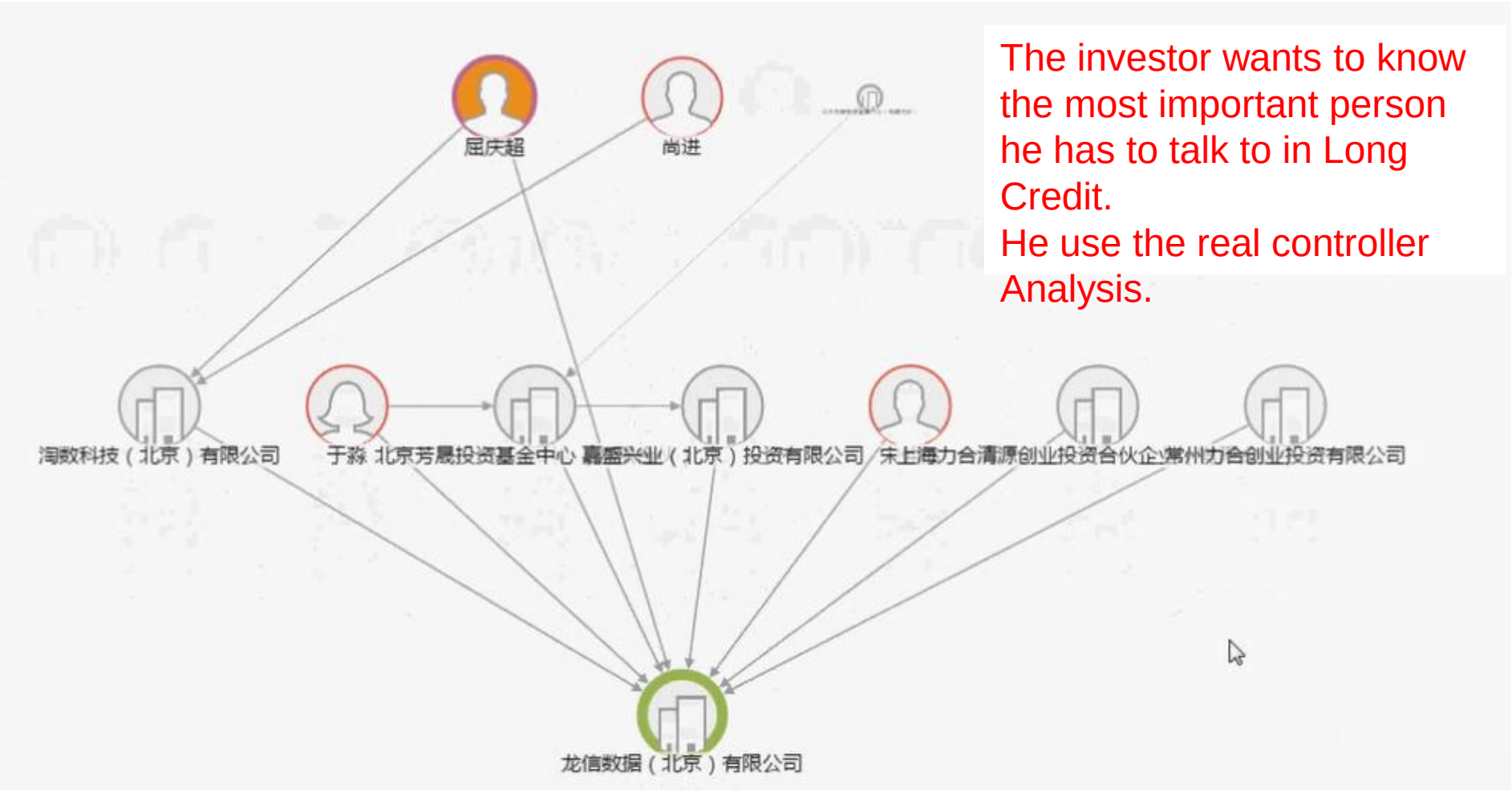


Usage Scenarios (3. Pedigree Analysis)

The investor wants to know important indirect and direct share holdings. He uses Pedigree analysis.



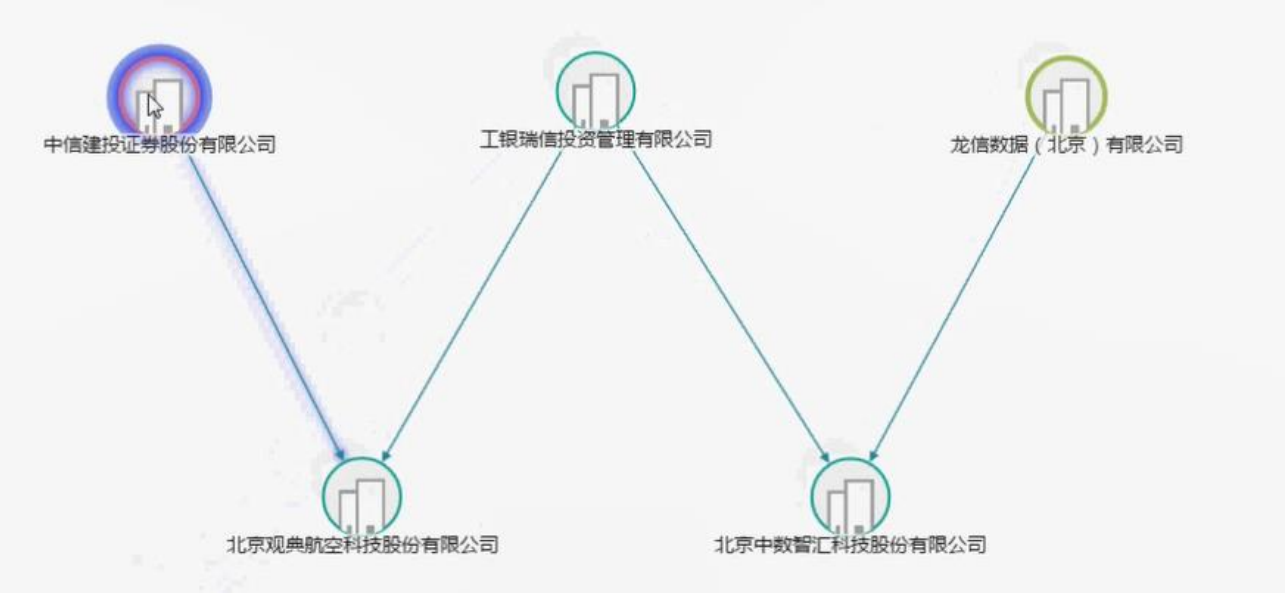
Usage Scenarios (4. Real Controller)



The investor wants to know the most important person he has to talk to in Long Credit.
He use the real controller Analysis.

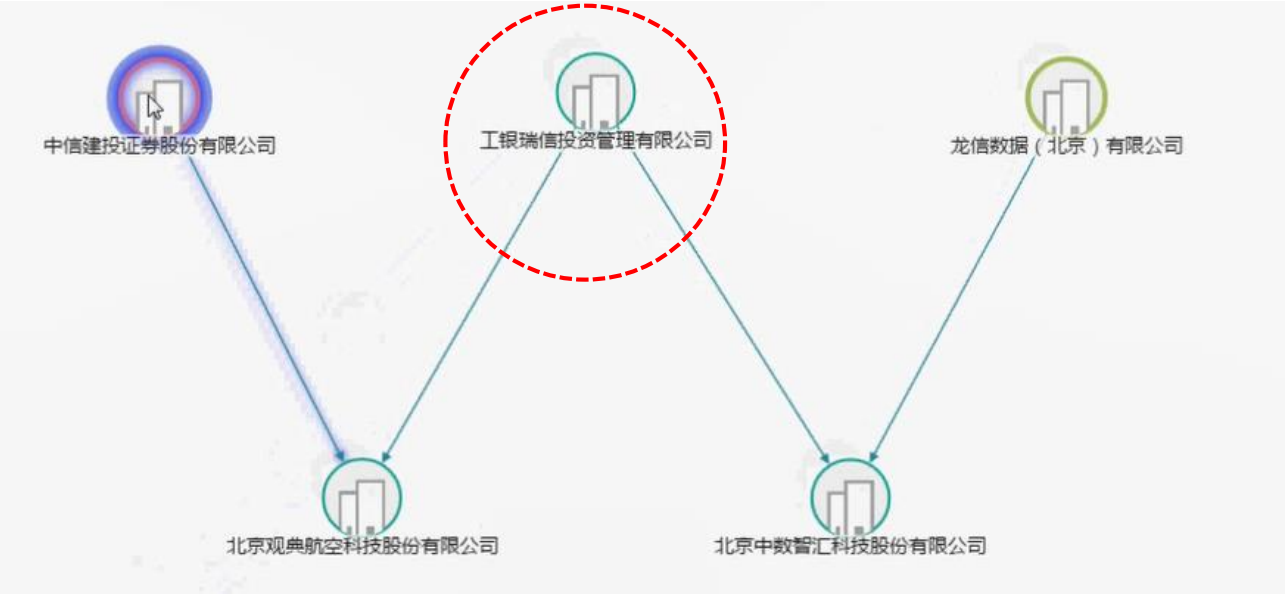
Usage Scenarios (5. Path Discovery)

Investors would like to know how to target the company. Therefore they look for the link between their own company and the target company..
Path discovery can help users find the shortest path among concerned companies.



Usage Scenarios (5. Path Discovery)

Investors would like to know how to target the company. Therefore they look for the link between their own company and the target company..
Path discovery can help users find the shortest path among concerned companies.



Future Work

- In the future, we plan to add more data sources to the KG, such as tax and invoice information per month.
- We will also try to monitor the change of shareholders as well as share ratios
- We could develop interesting applications such as “Control intention recognition to warn the current controller of the company.

Thanks!