



cnSchema中文知识图谱本体规范 以及金融领域应用探索

丁力 博士

海知智能联合创始人及CTO

北京 2018.12.08

目录

- 知识图谱101
- cnSchema：开放中文知识图谱的普通话
- 知识图谱在金融领域的应用

知识图谱101

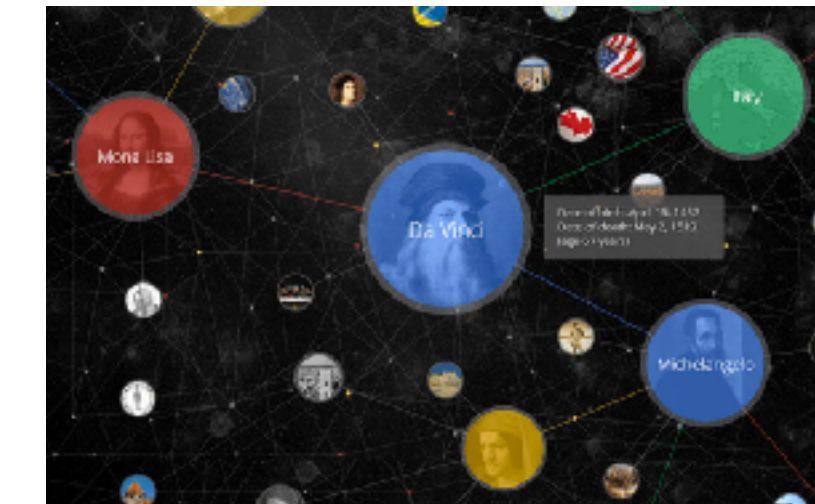
知识图谱简史



knowledge graph (2012)

rich snippet,

Schema.org (2010)



12+ Million
Web Sites

A de facto vocabulary for
structured data on the web

Found On
30% Pages

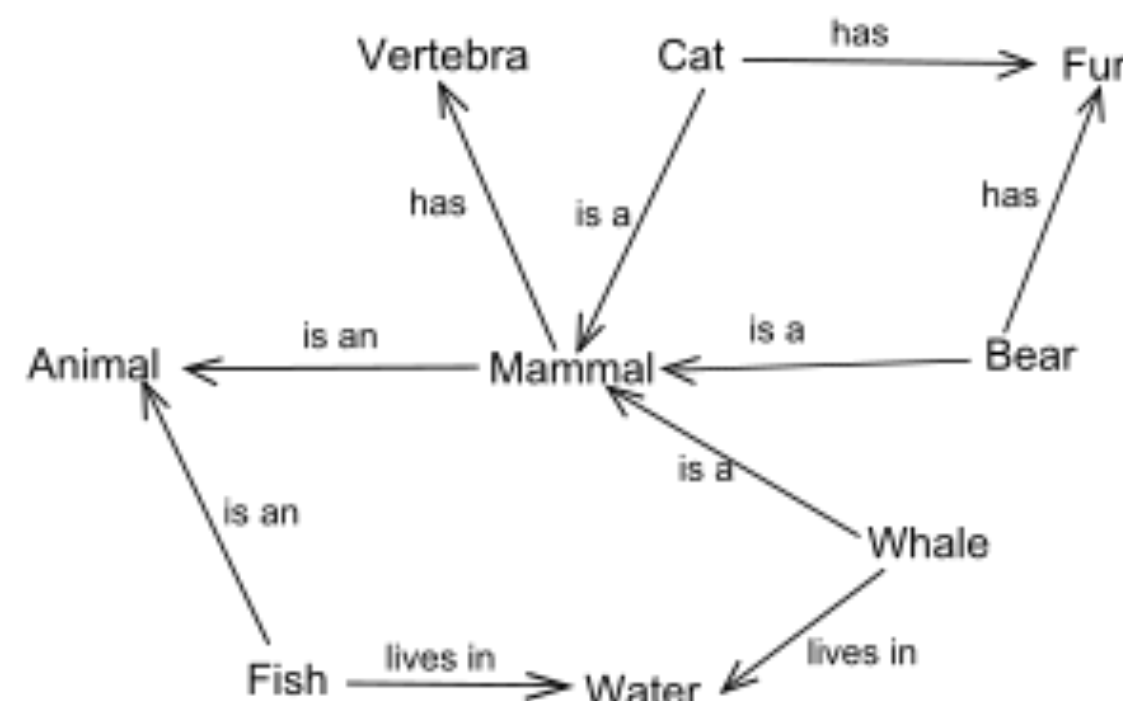
schema.org

linked data (2007)

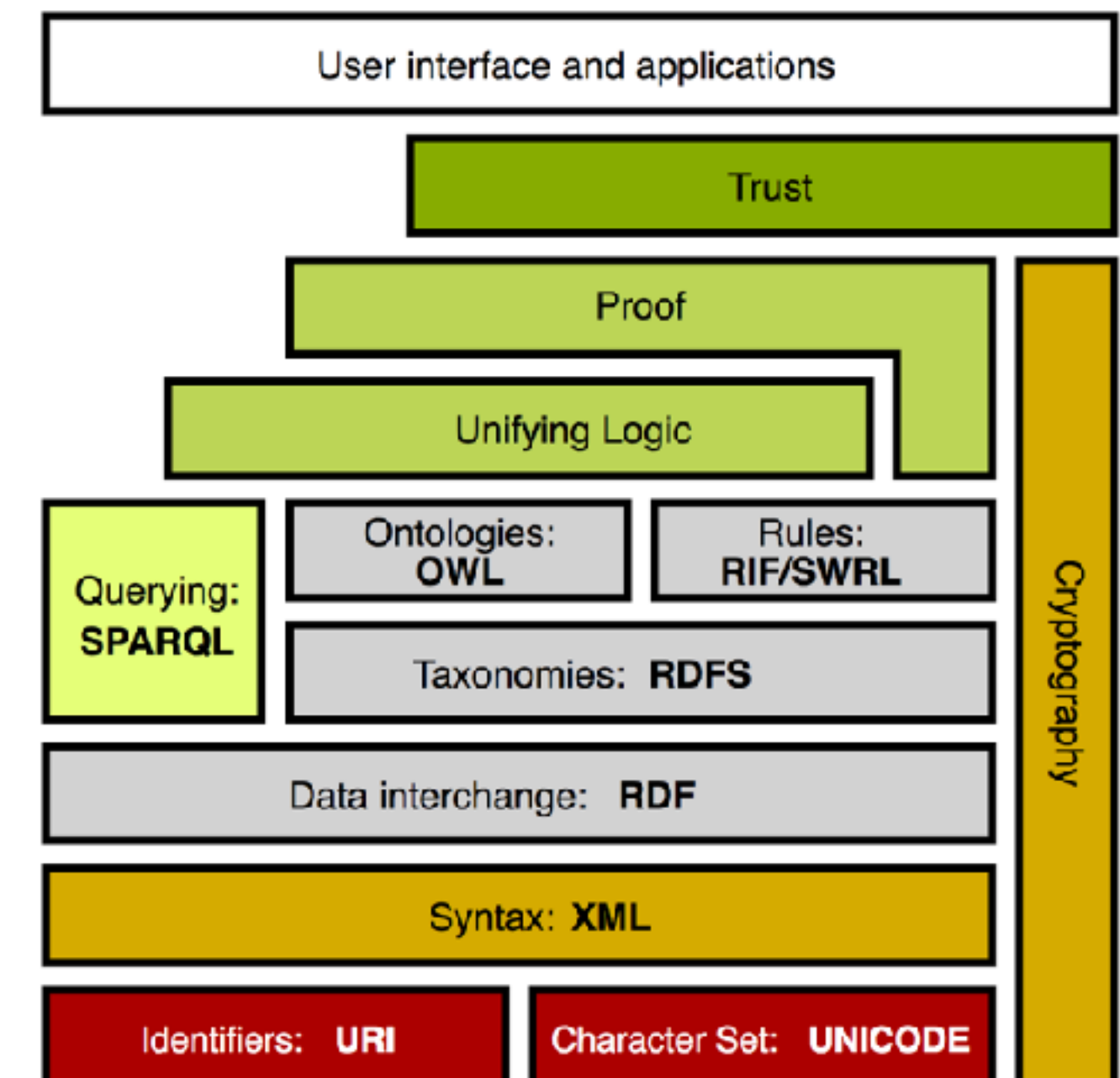
dbpedia, data.gov

semantic web (2001)

RDF, DAML, OWL, SPARQL



semantic network (1956)



知识图谱的五个境界

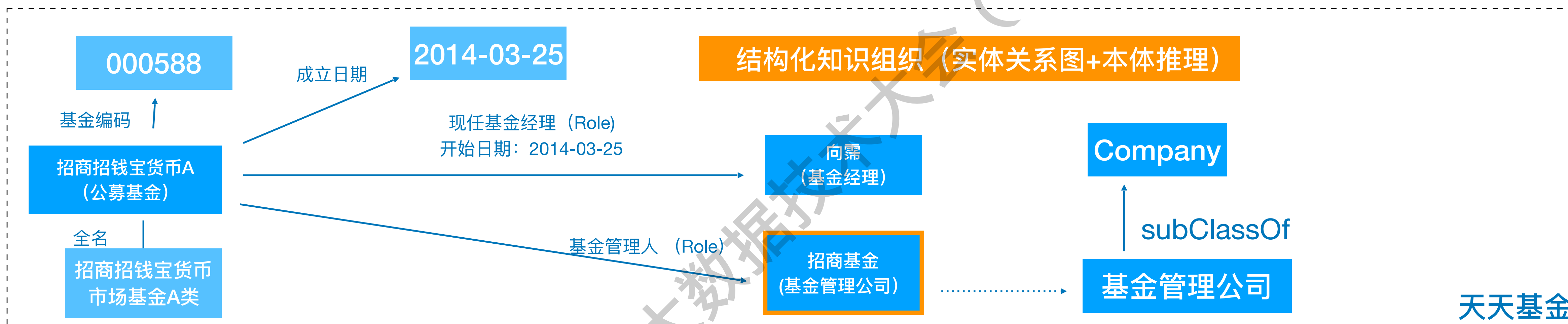
(命名)实体

结构

关系

本体

图融合



文本与知识图谱互联 (实体识别)

关于招商基金旗下招商招钱宝基金增加苏宁为代销机构的公告

根据招商基金管理有限公司（以下简称本公司）与南京苏宁基金销售有限公司（以下简称苏宁）签署的《开放式证券投资基金委托销售协议》及《开放式证券投资基金委托销售补充协议》，自2018年1月26日起，苏宁开始代理以下各基金的账户开户、申购、赎回等业务。

基金名称	基金代码
招商招钱宝货币市场基金A类	000588

投资者也可以通过以下途径咨询有关情况：
南京苏宁基金销售有限公司
客服电话：4008365365-9
公司网址：<http://fund.suning.com>

基金公告

sameAs

跨领域知识图谱互联 (实体链接)

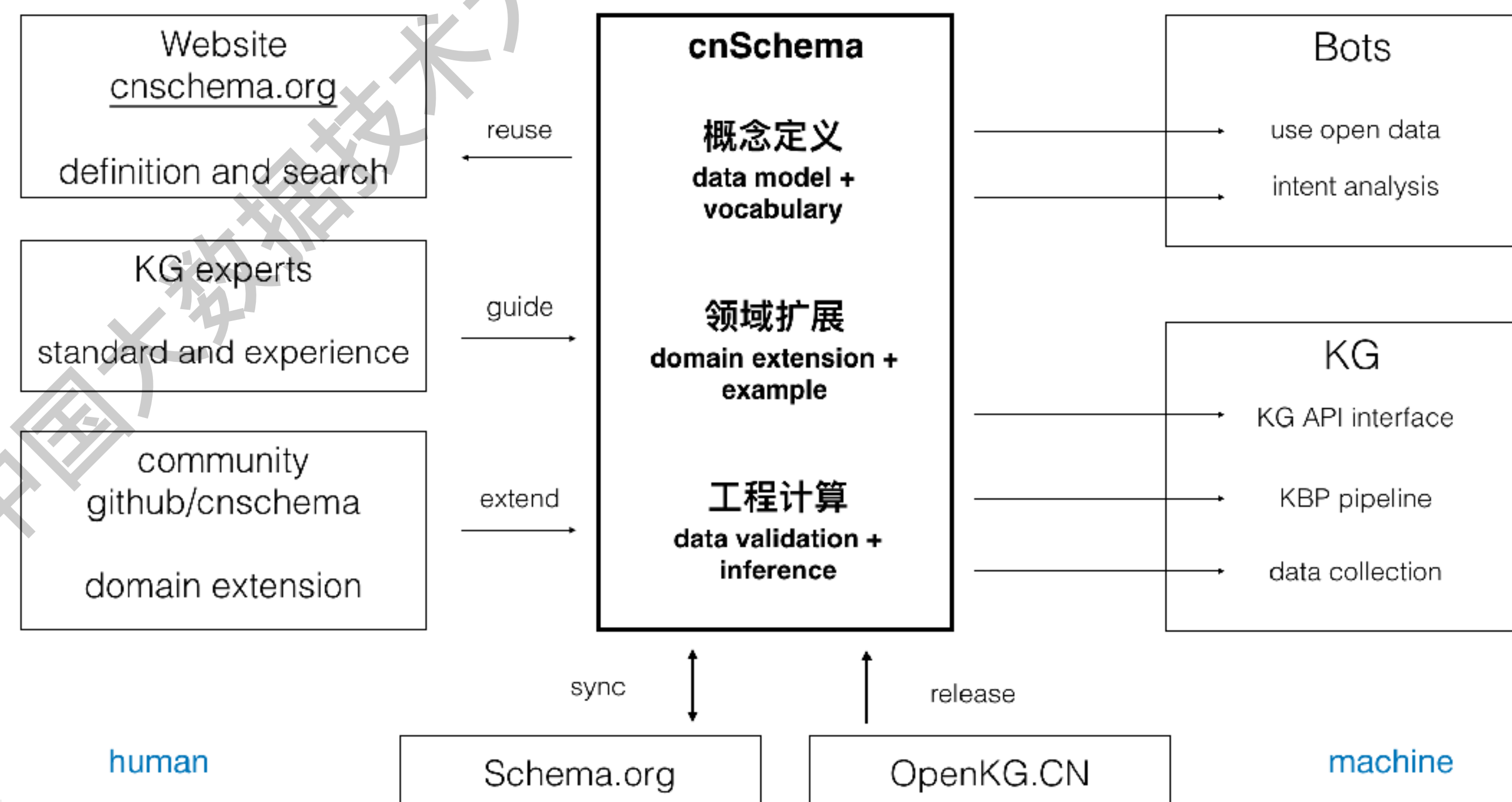


cnSchema: 开放中文知识图谱的普通话

cnSchema: 中文开放知识图谱的普通话

cnSchema 面向中文信息处理，为开放中文知识图谱的建模、生产、推理、学习、交互以及应用落地提供可解释、可迭代、可复用的数据接口定义与标准。

- 基于中文
- 以人为本
- 领域扩展
- 共享计算



cnSchema 1.0: 中文知识图谱的普通话 cnSchema.org

cnSchema项目文档词汇表关于OpenKG.CN

输入概念分类, 属性。如 "人"

CreativeWork

Canonical URL: <http://cnschema.org/CreativeWork>

Thing > CreativeWork

【作品】最通用的创作工作, 包括书籍, 电影, 照片, 软件程序等 The most generic kind of creative work, including books, movies, photographs, software programs, etc.

Linked to schema.org: <http://schema.org/CreativeWork> Usage: Between 250,000 and 500,000 domains

Linked to Wikidata: <https://www.wikidata.org/wiki/Q17537576> (creative work)

Linked to Wikipedia: https://en.wikipedia.org/wiki/Creative_work

属性	值域	定义
Properties from CreativeWork		
about	Thing or CreativeWork or Event or CommunicateAction	【关于】事物的关于信息 The subject matter of the content.

概念定义: 中文翻译+WIKI链接。基于 schema.org 3.2 核心词汇, 由清华大学, 浙江大学, 复旦大学, 东南大学, 海知智能, 以及社区志愿者共同完成

cnSchema项目文档词汇表关于OpenKG.CN

stor

AutoPartsStore(汽车配件店)
ClothingStore(服装店)
ConvenienceStore(便利店)

cnSchema.org 的 Autocomplete 词汇搜索服务

Filter... x

EntityCoreAPI
Get Entity
Lookup Entities
EntityIndexAPI
Cypher Graph Query
ElasticSearch Query
EntitySyncAPI
Batch Update
Delete an entity
List Entities
Update/Create an entity

EntityCoreAPI - Lookup Entities

redis lookup service。按名字 获取一组实体简略信息 (@id, @type, name, entityScore)。

POST

entities/

Parameter

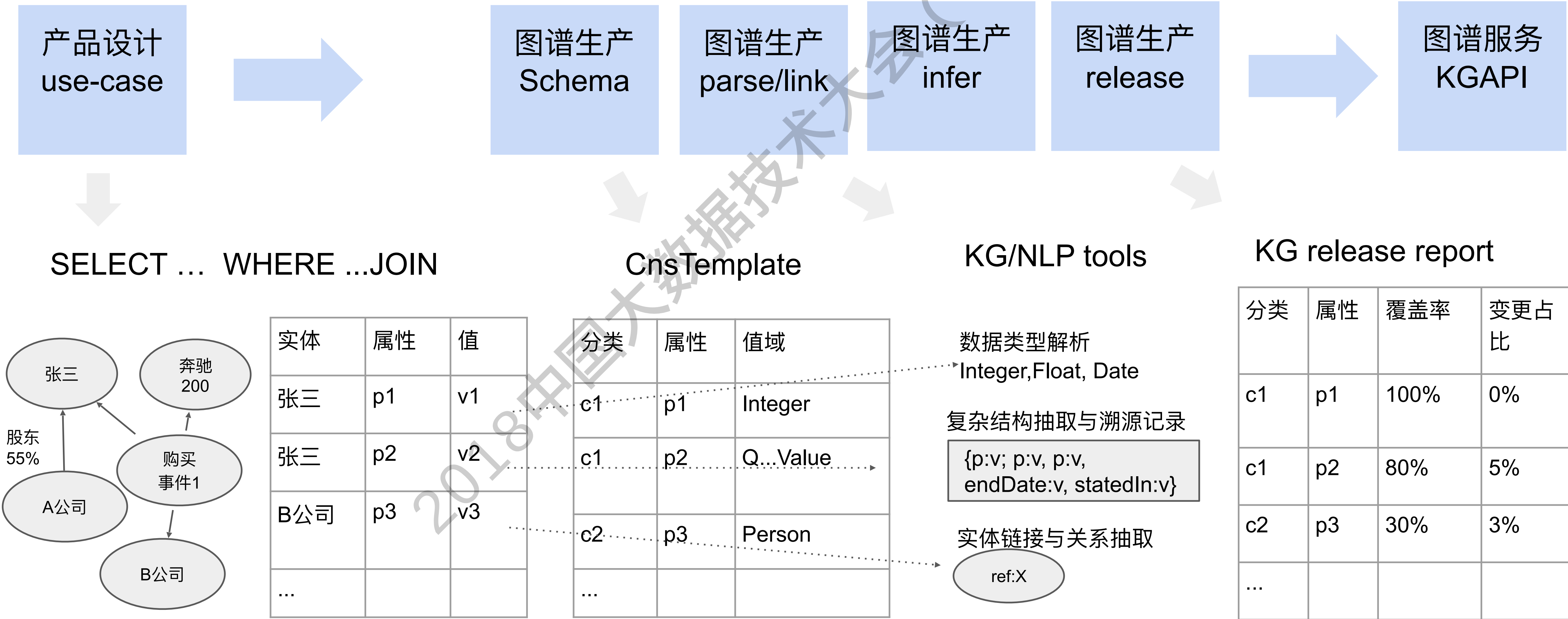
Field	Type	Description
names	String[]	必须 list of Entity name, match whole word.

Lookup Query

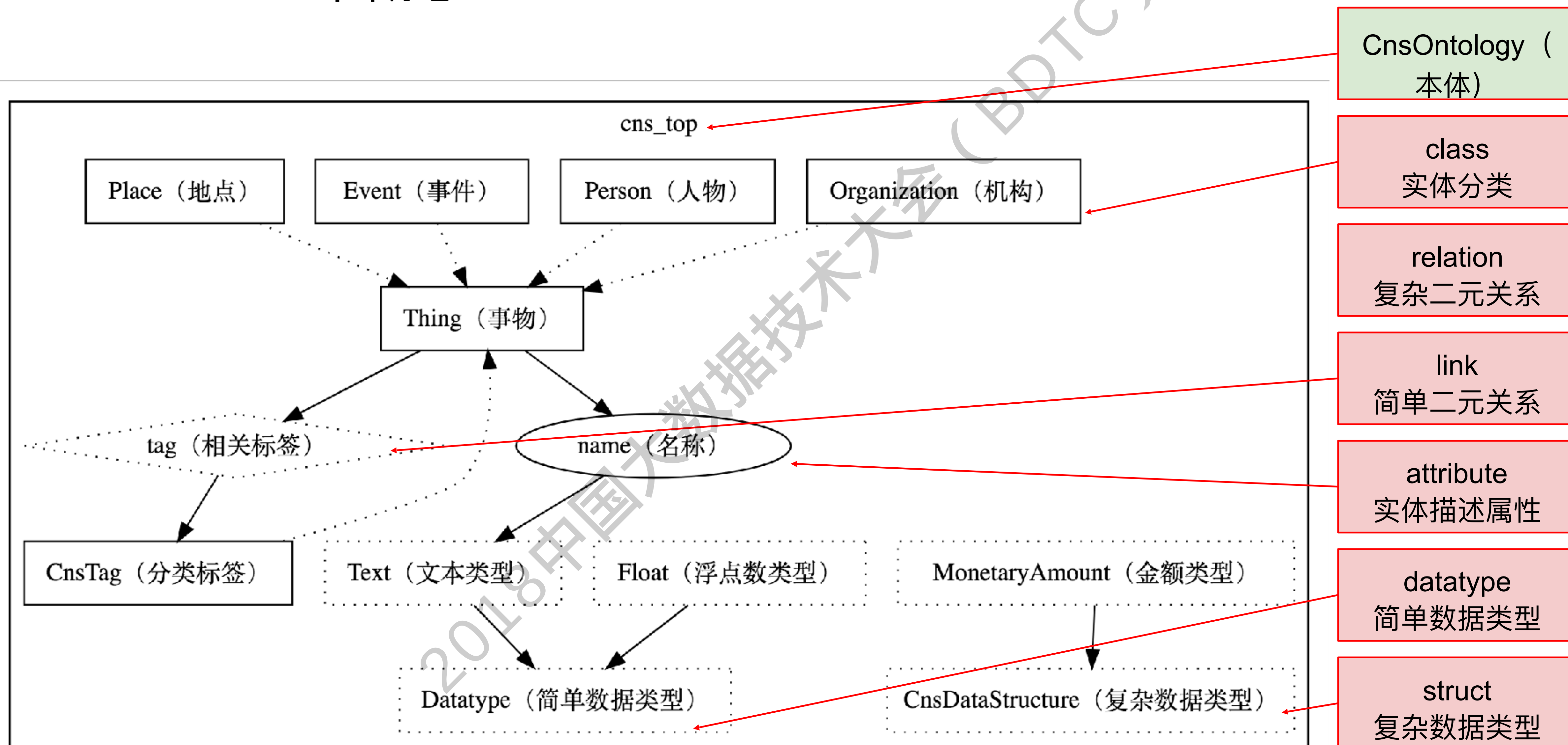
```
{  "names": [    "华仔",    "唱歌"  ]}
```

KGAPI接口规范

cnSchema2.0：开放中文知识图谱的模板化解决方案



cnSchema基本概念



JSONLD承载cnSchema

- 知识表示
 - 逻辑层 / 语义: RDF / RDFS
 - 语法层 / 格式: JSON-LD
- 本体
 - 词汇集: 属性, 分类, 关系, ...
- 实体网络
 - ID
 - 名称
 - 属性与关系

```

<script type="application/ld+json">
{
  "@context": "http://schema.org",
  分类 "@type": "MusicRecording",
  "@id": "http://musicbrainz.org/recording/3566e45",
  "name": "Back in the U.S.S.R.",
  "producer": {
    "@type": "Person",
    "name": "George Martin"
  },
  属性 "duration": "PT2M43S",
  "recordingOf": {
    "@type": "MusicComposition",
    "name": "Back in the U.S.S.R.",
    "iswcCode": "T-010.140.236-1"
  }
}
</script>

```

格式

ID

名称

属性值

关系

schema.org

<http://schema.org/MusicRecording>

cnSchema2.0 实体属性模版1/4

PRE KG：场景描述

专家用Excel和脑图描述场景

KG第一境界：命名实体

KGAPI命名实体查询

更精准的意图理解

KG第二境界：实体属性模板

实体分类的特有属性

更精准的答案

实体分类	属性英文名	中文名	实体1	实体2
Thing	name	名称	北京大学	方正富邦基金管理 有限公司
Thing	alternateName	别名	北大	方正富邦
Organization	province	所在省份	北京市	北京
Company	unifiedSocialCredit Code	统一社会信用 代码		911100007178849 15E
Company	saicRegistrationCa pital	工商注册资本		66000 万人民币
MutualFundC ompany	numberOfFund	旗下基金数量		14

cnSchema2.0 实体属性模版2/4

KG第二境界：实体属性模版（CnsTemplate）

实体分类的特有属性
复用常用模版

中文名称与定义
从顾名思义到精准定义

溯源与有效期
属性值的(原始文本)与(预期解析结果)
自动解析/抽取的目标与校验

refClass	refProperty	<u>property NameZh</u>	<u>propertyDefinitionZhSource</u>	<u>property Range</u>	raw	expected
Company	saicRegistrationCapital	工商注册资本	注册资本是指合营企业在登记管理机构登记的资本总额,是合营各方已经缴纳的或合营者承诺一定要缴纳的出资额的总和。	Monetary Amount	66000万人民币	{"value": 660000000, "currency": "cny", "statedIn": "工商总局"}
Organization	province	所在省份	组织机构所在的省份（直辖市以及特别行政区）	Province	北京市	{ "@id": "...", "@type":["Province", "Place", "Thing"], "name": "北京"}
<u>MutualFund Company</u>	<u>numberOfFund</u>	旗下基金数量	基金公司目前管理的基金总数	<u>QuantitativeValue</u>	14	{"value":14, "endDate": 2018-08-01T00:00:00}

cnSchema2.0 实体属性模版3/4

KG第二境界：实体属性模版（CnsTemplate）

实体分类的特有属性
复用常用模版

中文名称与定义
从顾名思义到精准定义

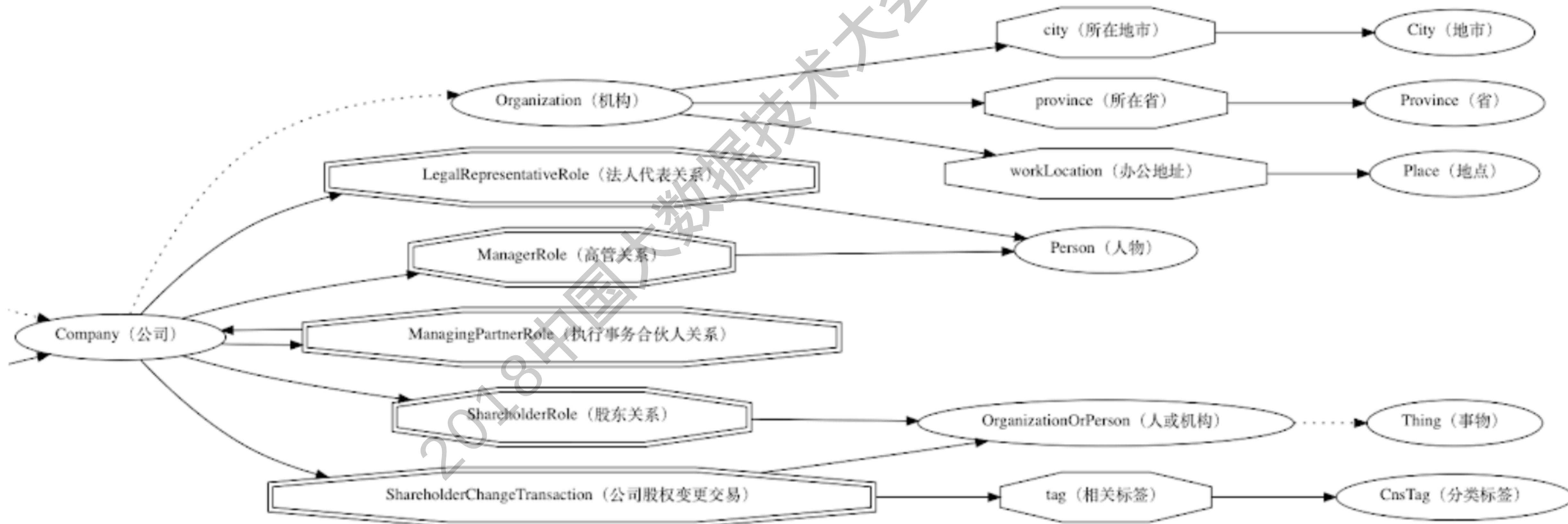
溯源与有效期
属性值的(原始文本)与(预期解析结果)
自动解析/抽取的目标与校验

refClass	refProperty	<u>property NameZh</u>	<u>propertyDefinitionZhSource</u>	<u>property Range</u>	raw	expected
Company	saicRegistratio nCapital	工商注 册资本	注册资本是指合营企业在登记管理机构登记的资本总额,是合营各方已经缴纳的或合营者承诺一定要缴纳的出资额的总和。	Monetary Amount	66000 万人民 币	{"value": 6600000000, "currency": "cny", "statedIn": "工商总局"}
Organization	province	所在省 份	组织机构所在的省份（直辖市以及特别行政区）	Province	北京市	{ "@id": "...", "@type":["Province", "Place", "Thing"], "name": "北京"}
<u>MutualFund Company</u>	<u>numberOfFund</u>	旗下基 金数量	基金公司目前管理的基金总数	<u>Quantitat iveValue</u>	14	{"value":14, "endDate": 2018-08-01T00:00:00}

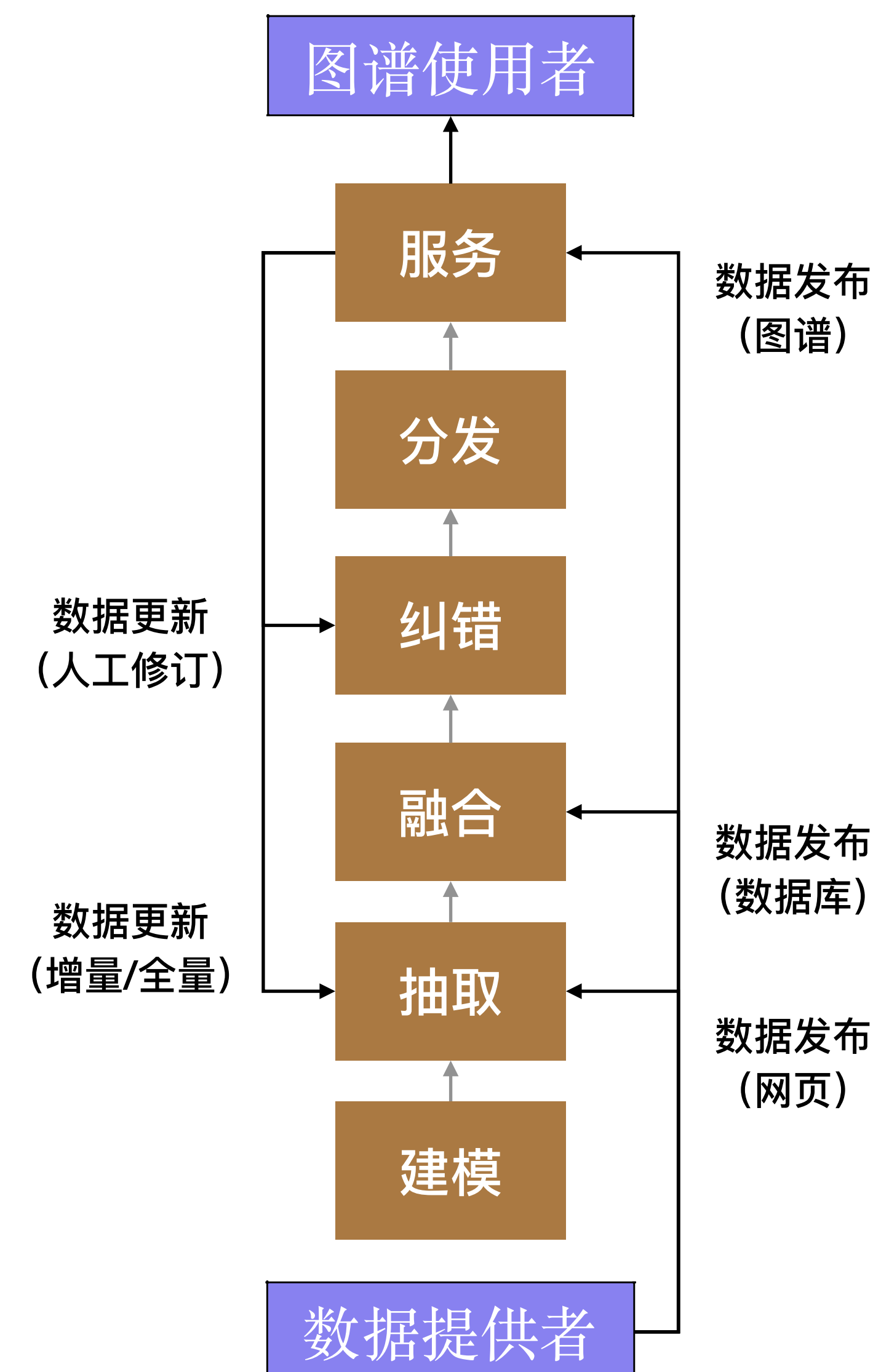
cnSchema2.0 实体属性模版3/4

KG第三境界：实体二元关系 (CnsLink)

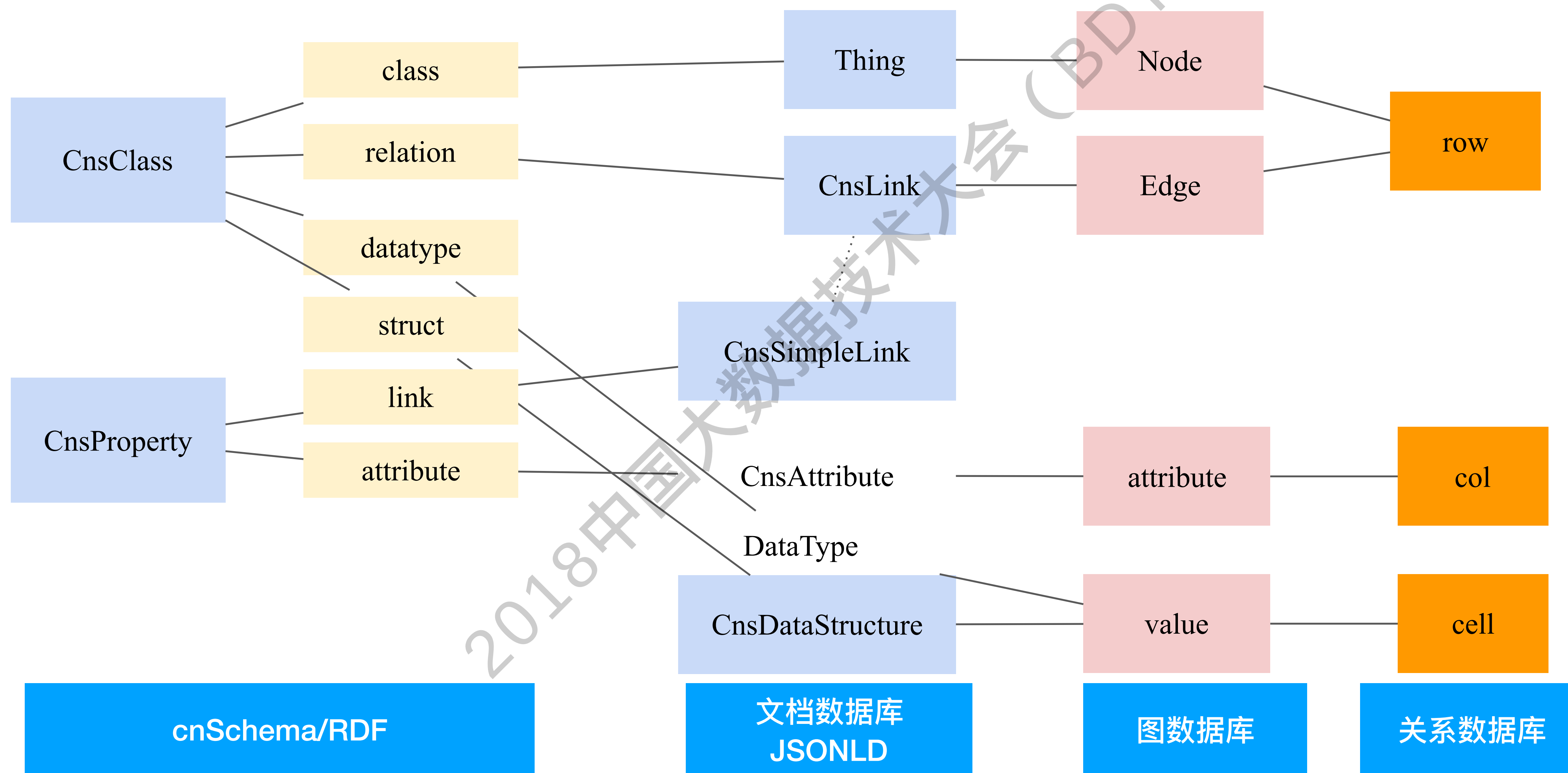
KGTOOL绘制schema关系设计图：简单二元关系（单边框八边形）和复合二元关系（双边框八边形）连接实体分类（椭圆形）
KGTOOL绘制图谱实体关系分布统计图（TODO）
直观理解实体关系图的[Schema设计]与[真实分布统计]



基于cnSchema, 实现中文知识图谱工程化应用落地



Next: cnSchema 的存储体系



Next: cnSchema的用户视角

- 知识工程师视角（全局视角）：从常用数据接口和图谱平台系统的角度看，在没有具体数据集和具体项目应用的时候，预见考可能用到的实体和属性，为后续复用概念打基础
- 数据发布者视角（用户发布视角）：从拥有的原始数据出发，以及融合的外部数据，目标是发布可共享复用的图谱（发布者规范数据接口，支持多种应用开发）。说明生产的图谱中如何覆盖用户关注的原始数据，包含了哪些实体分类，属性模版
- 数据使用者视角（用户应用视角）：从具体应用项目出发，目标是为了定制图谱支持应用落地（规范应用的数据接口，可以切换图谱提供商）。说明项目要用到哪些数据，需要什么实体分类，属性模版

参与cnSchema开放社区



cnSchema.org 的研发路线规划		
	Schema	Tool
1.0 (17Q3)	- schemaorg3.2中文翻译与WIKI链接 - 命名实体 Thing	- Cnschema.org网站 - 浅层次cnSchema概念文本搜索 - KGAPI知识图谱服务接口
2.0 (18Q3)	- 实体属性模版 CnsTemplate - schemaorg 3.4 增补更新 - 核心图谱 Schema - 实体二元关系 CnsLink	- KGTOOL, 图谱数据质量校验 - KGTOOL, Schema可视化
2.1 (18Q4)	- 本体计算 CnsOntology - 领域图谱Schema社区扩展	KGAPI, 多层级图谱数据索引

github.com/cnschema

dl@ruri.ai

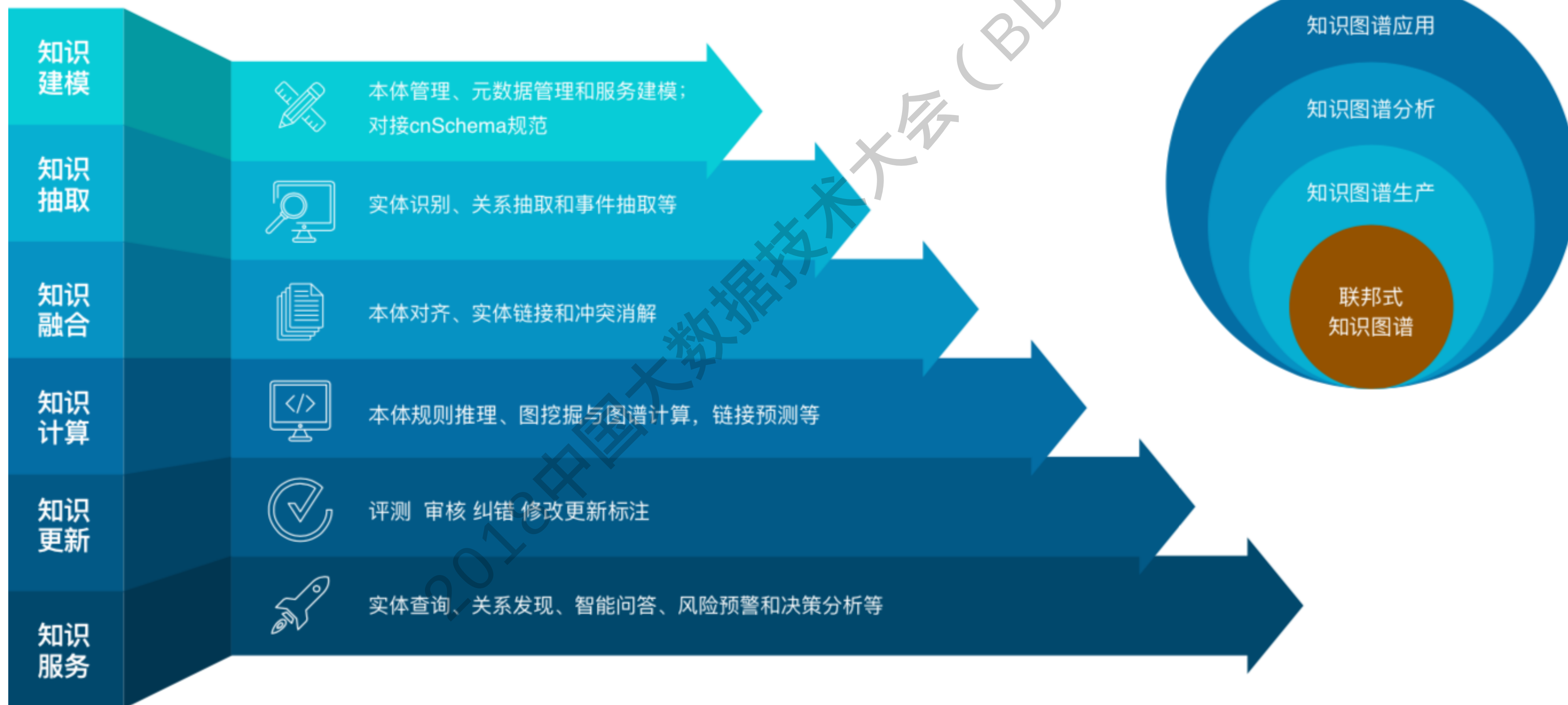
知识图谱在金融领域的应用

2018中国大数据生态大会 (BDTC)

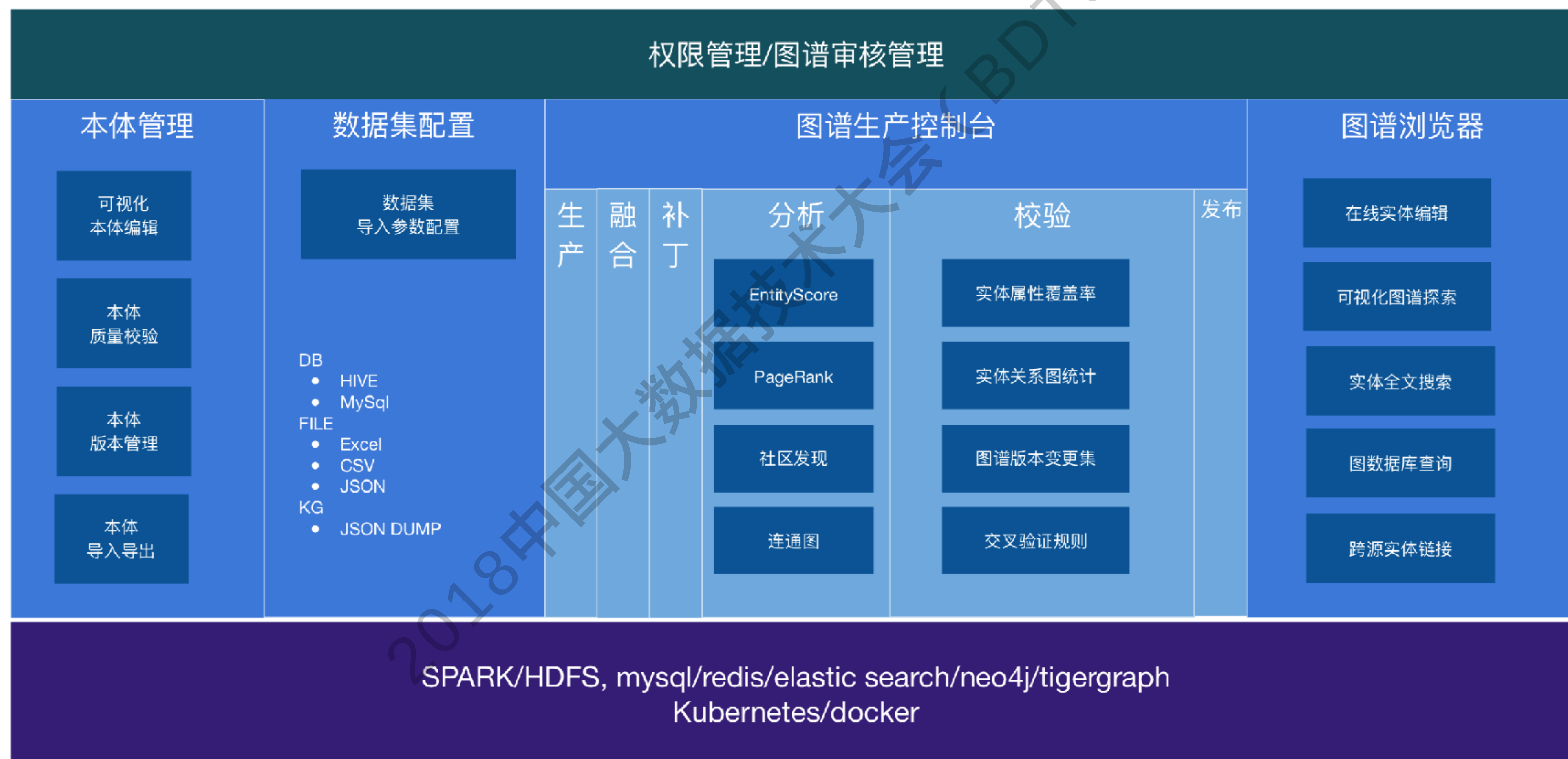
知识图谱在金融领域典型应用场景



知识图谱的关键技术



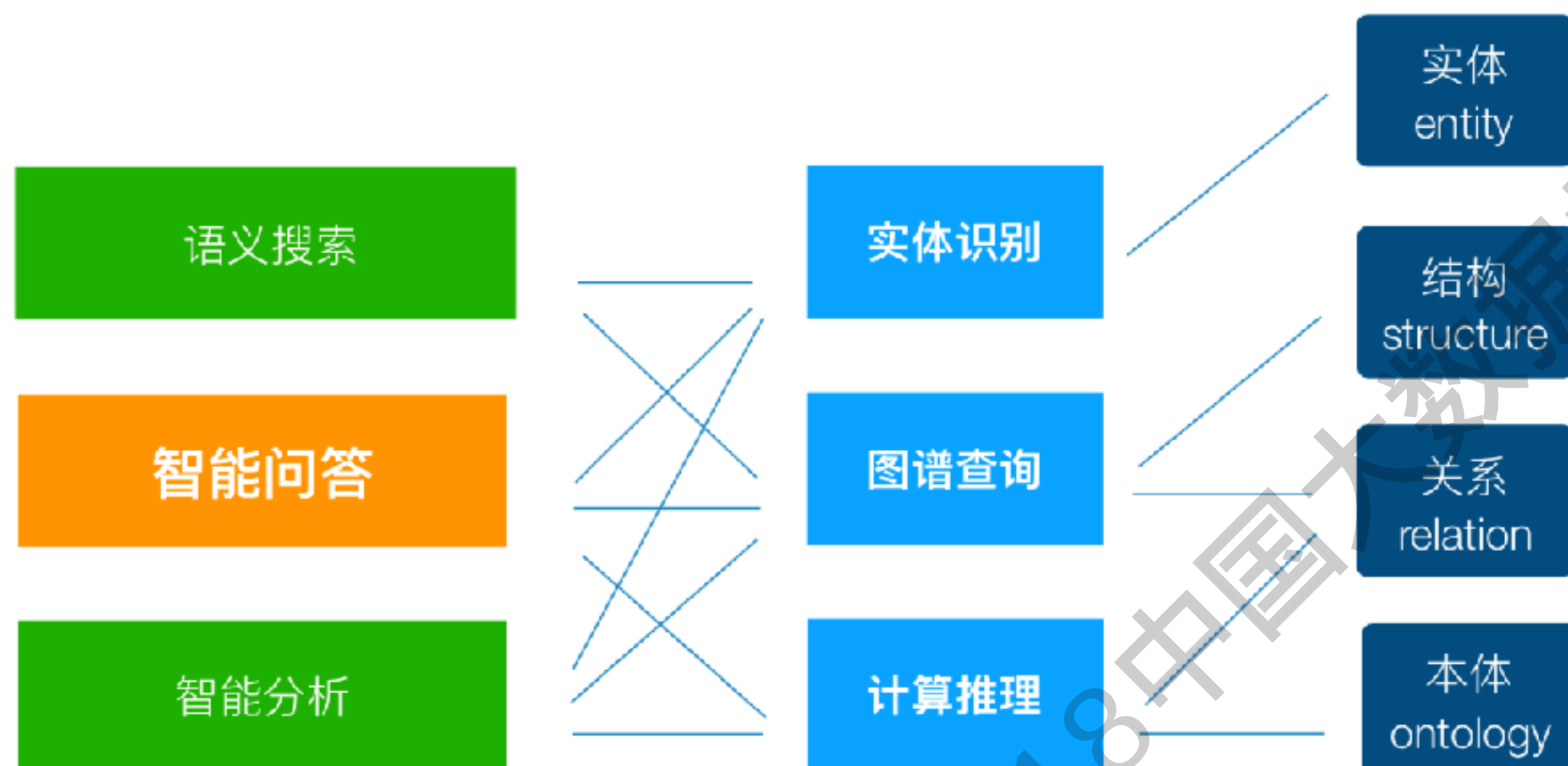
知识图谱平台架构



知识图谱工程化

业务开发	算法应用，人机交互，人机协同数据处理，接口设计...
数据存储与检索	Redis/ElasticSearch/Neo4/MySQL/MongoDB...
大数据计算	Hue/Hive/HBase/Hadoop/Spark...
虚拟化云计算	Docker/Kubernetes

案例：摩羯智投问答机器人



构建了知识图谱，机器可以更准确意图识别以及联想、推理，具备可解释性；不再是机器学习产生结果类似于死记硬背、不可解释

摩羯智投测试号

李锐近几年的业绩如何

摩羯智投测试号

基金名称	基金经理任职起止日	基金经理任期回报	基金经理任期最大回撤
大摩多元收益(001655)	2015-11-17 ~ 2017-06-16	2.20%	5.74%
大摩稳定添利A(000415)	2014-09-02 ~ -	30.13%	0.00%
大摩稳定添利C(000416)	2014-09-02 ~ -	28.24%	0.00%
大摩稳定添利(000064)	2013-06-25 ~ -	49.69%	0.00%
大摩多元C(233013)	2012-08-28 ~ -	65.10%	2.47%
注：按照基金经理起始日期，最多仅显示最近的5只基金			

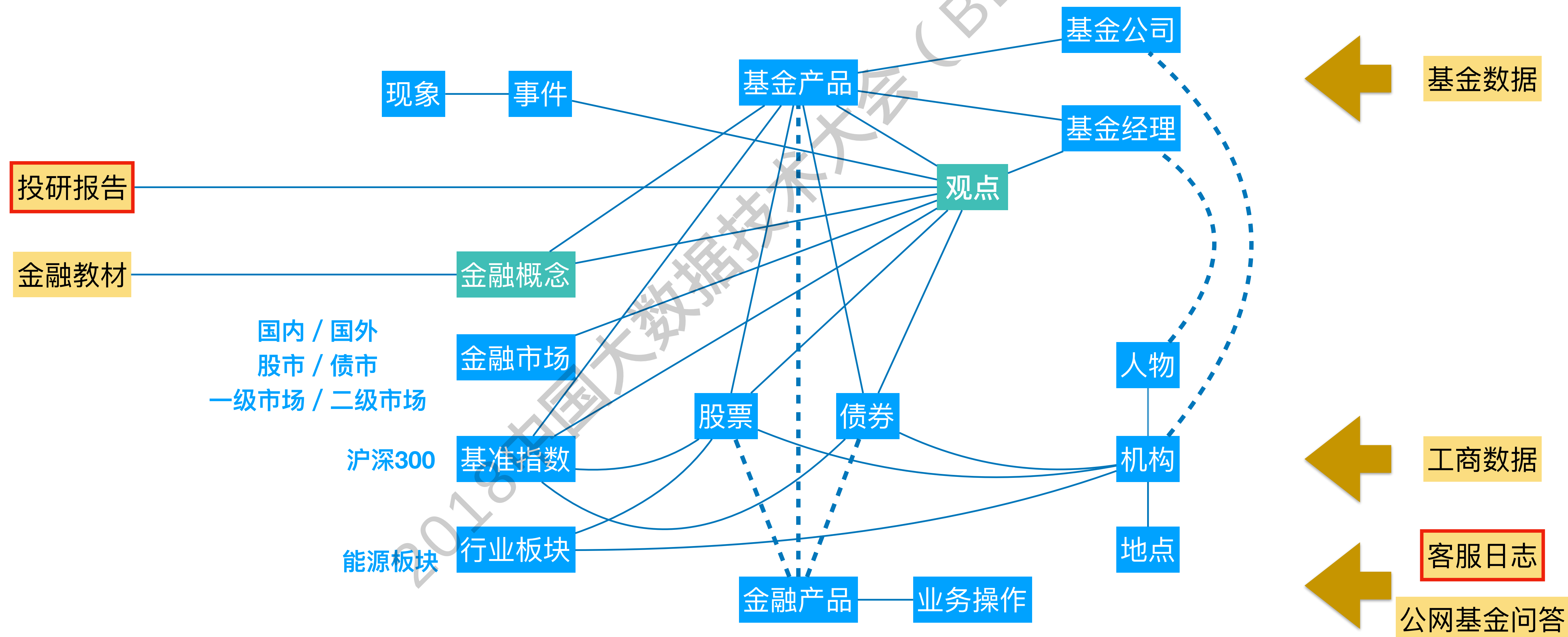
摩羯智投测试号

华夏沪深300指数证券投资基金的基金经理的表现和000457比如何

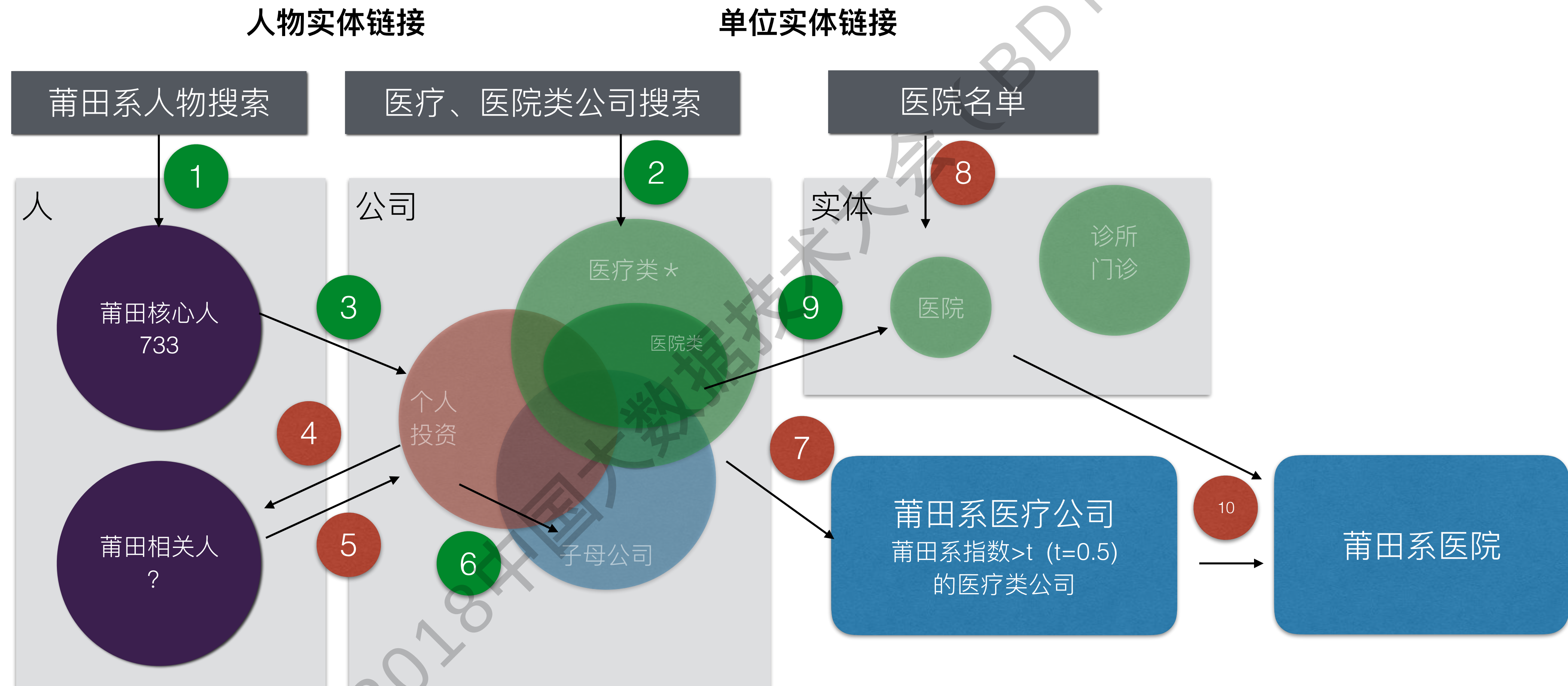
摩羯智投测试号

基金名称	华夏300(000051.J)	上投核心(000457)
基金类型	股票型基金	股票型基金
基金规模	173.66亿元(截至至: 2012-09-30)	53.71亿元(截至至: 2018-06-30)
基金经理	方军	孙芳
任职起始日期	2004-12-30	2011-12-08
管理过的基金数量	7只	6只
任职期间最佳基金回报	-31.30%	-3.29%
任职期间最差基金回报	247.15%	146.30%
任职期间最大回撤	70.23%	55.73%
其他基金经理	张弘强	李博
注：最多仅显示基金经理任职期限最长的5只基金		

公募基金Schema概要设计



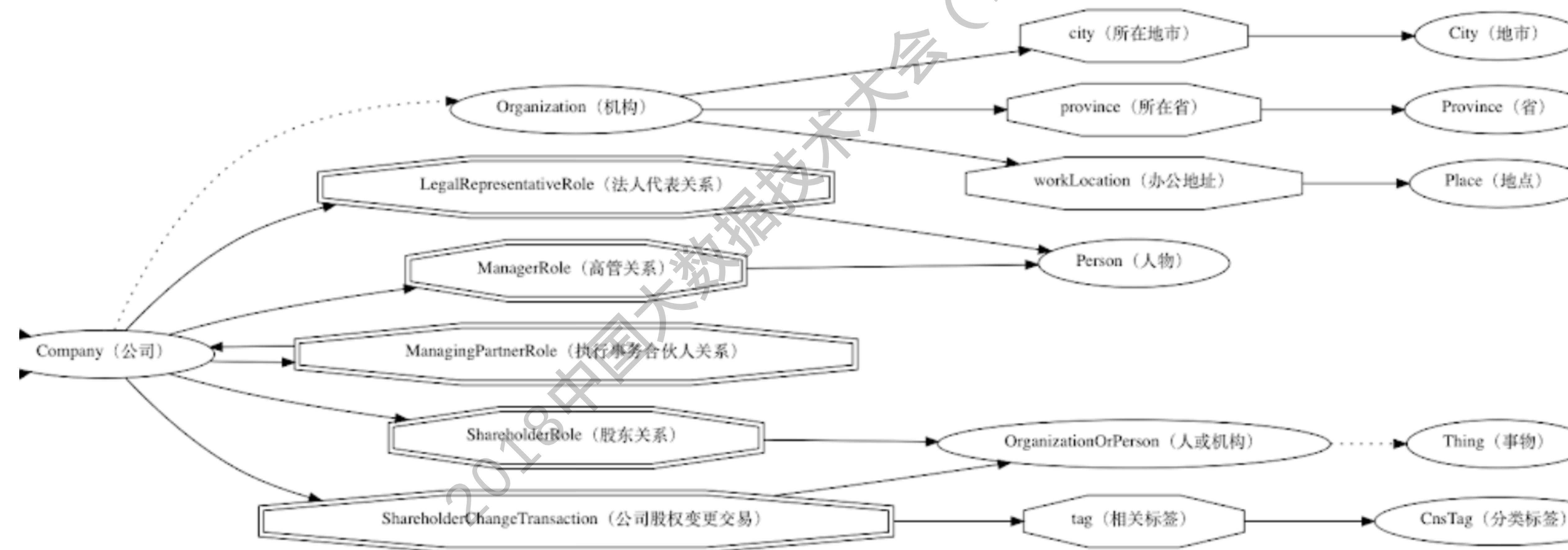
案例：莆田系企业集团发现



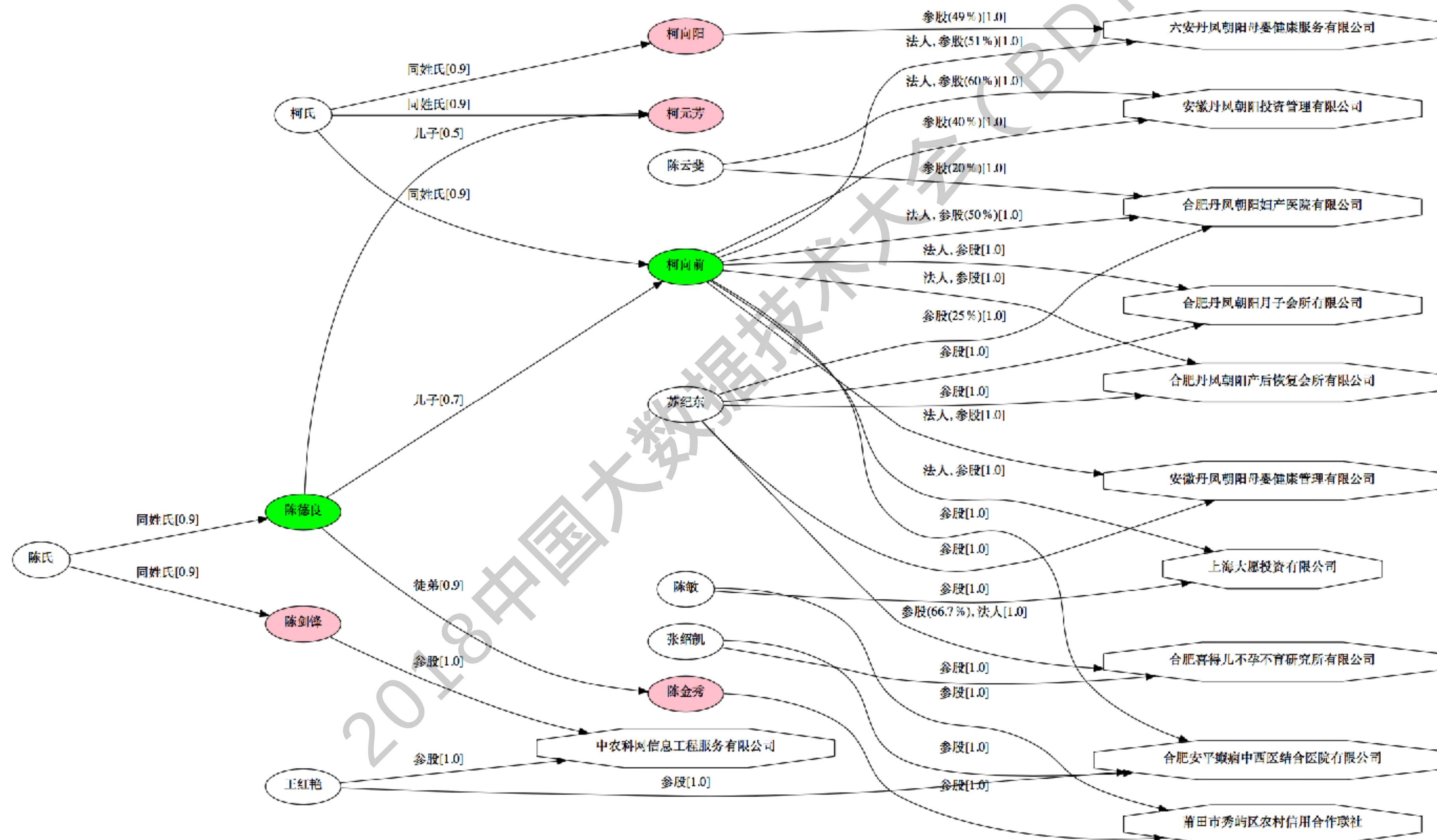
工商企业图谱：股权控制关系

药监局许可证图谱

组织机构Schema (cns_organization)

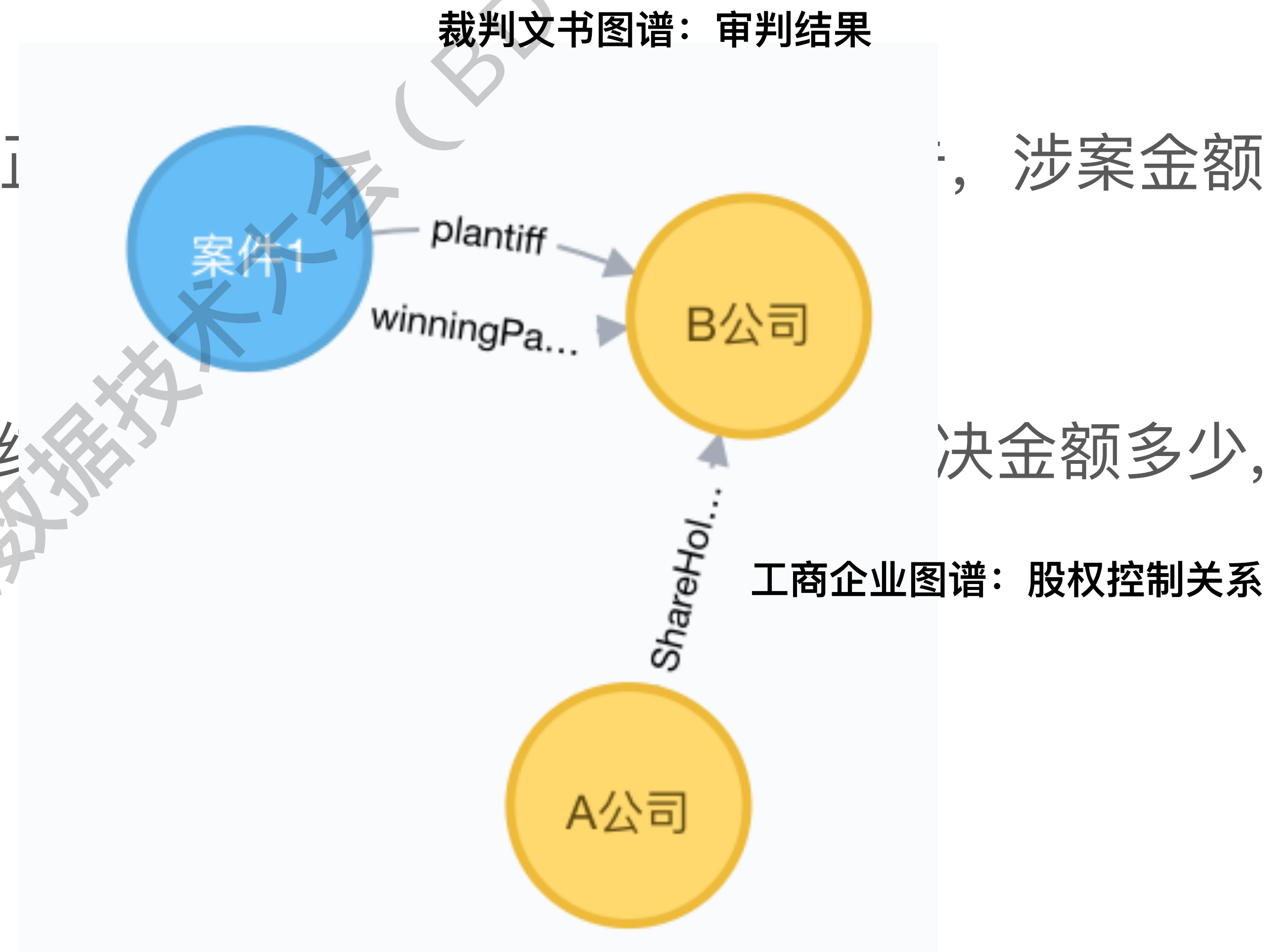


莆田系集团

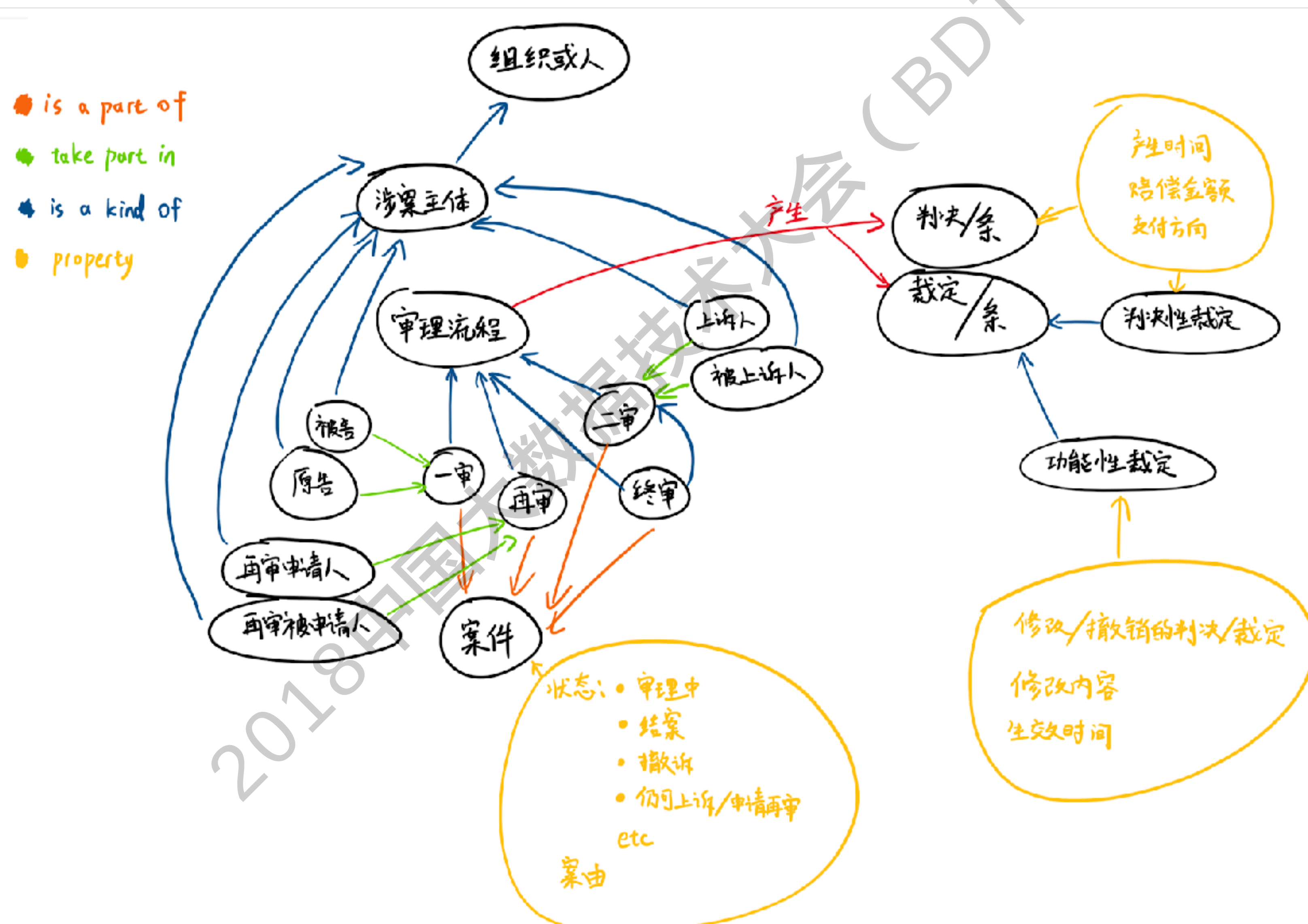


案例：企业涉诉查询

- 给定公司A。要求查看A的：
 - 所有投资的公司、子公司当前涉案金额多少
 - 所有投资的公司、子公司已经结案的涉案金额多少，胜了还是败了

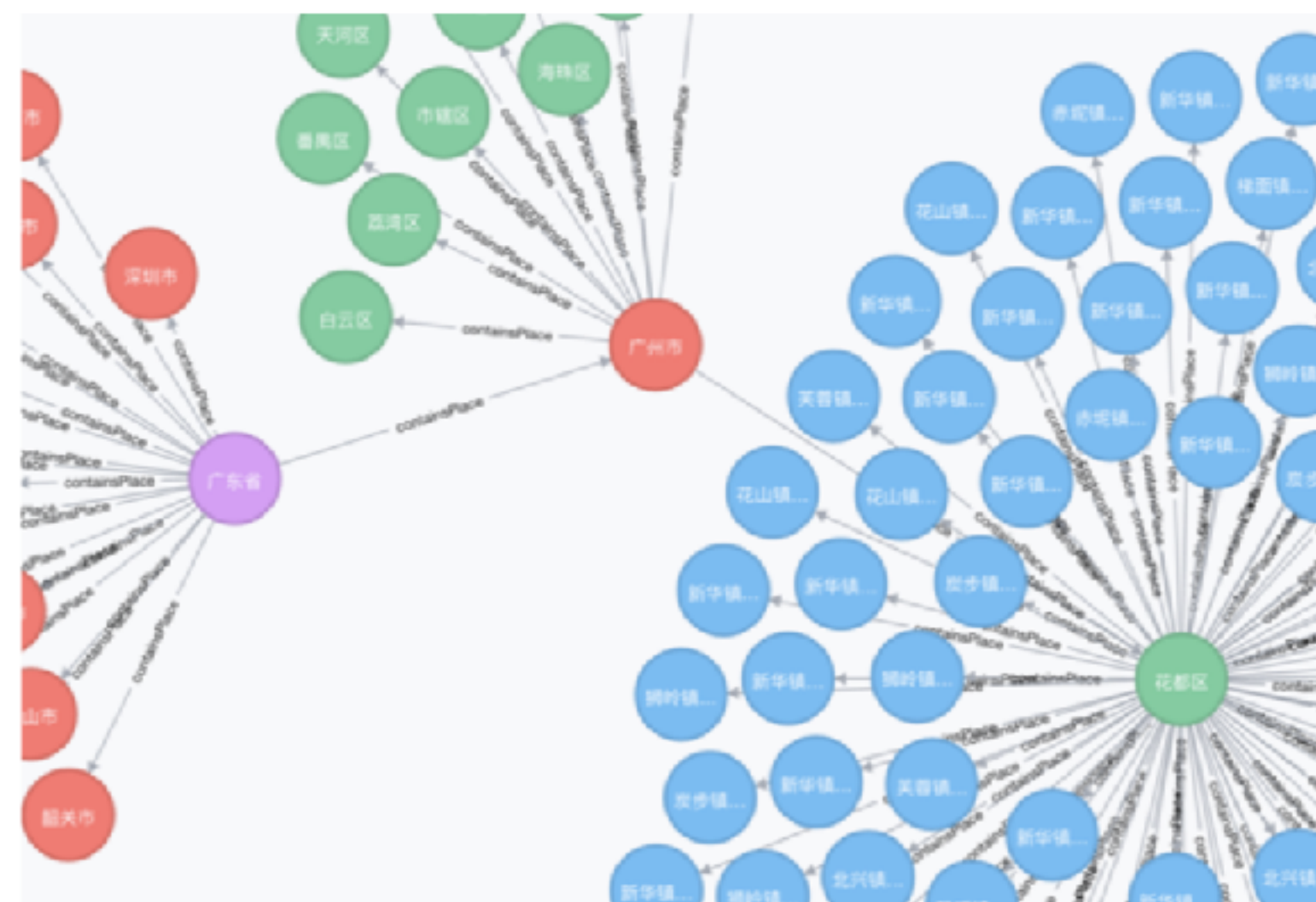


裁判文书图谱schema概要设计



地址知识图谱与地点实体链接 (暂不含台湾地区)

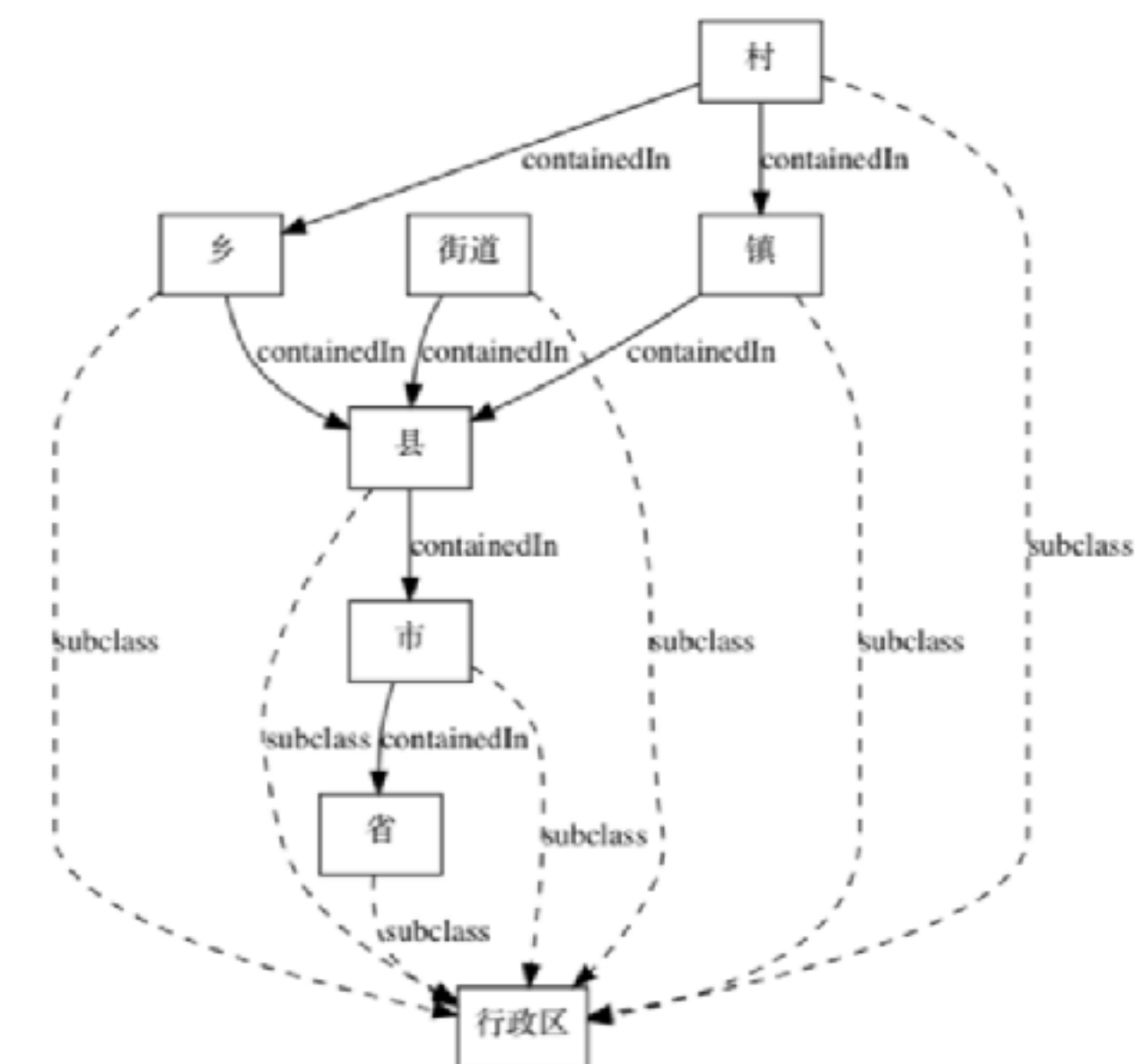
“广州花都区”



33 省
369 市
2,875 区
50,530 乡
335万 村
2,469万道路
2,036万 POI

算法准确率: >90%

地址知识图谱schema/本体模型



原地址
安徽省阜阳市颍州区临泉路217号18栋2户
中国广西壮族自治区贵港市城湖路24号
宜昌市西陵区港窑路58号

标准化

省	市	区	乡	村	道	门牌
安徽省	阜阳市	颍州区			临泉路	217号18栋2户
广西壮族自治区	贵港市	平南县			城湖路	24号
湖北省	宜昌市	西陵区			港窑路	58号

组织机构实体链接

原单位名称	标准化后单位名称
北京环球美联英语	环球美联(北京)教育咨询有限公司
北京环球美联英语培训学校	环球美联(北京)教育咨询有限公司
北京环球美联英语教育咨询有限	环球美联(北京)教育咨询有限公司
北京环球美联教育培训有限公司	环球美联(北京)教育咨询有限公司

总结

2018中国大数据技术大会 (BDTC)

知识图谱驱动的智能化解解决方案



海知智能的知识图谱解决方案

