

从数据互联到万物互联

From Linked Data to Linked Everything

陈华钧 教授、博导

浙江大学计算机科学与技术学院



2015年11月，宜昌，第三届中文知识图谱大会



1
理解

2
实践

3
展望

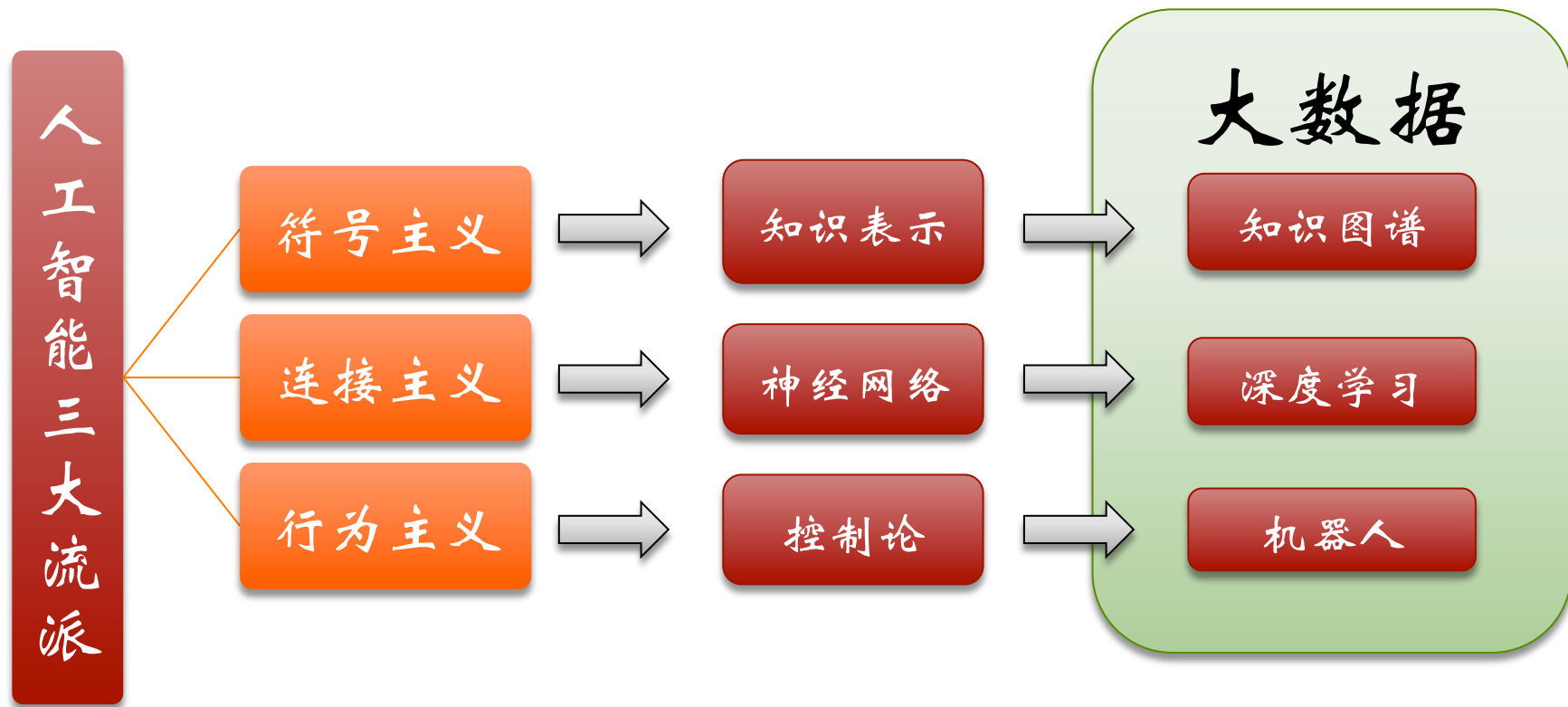




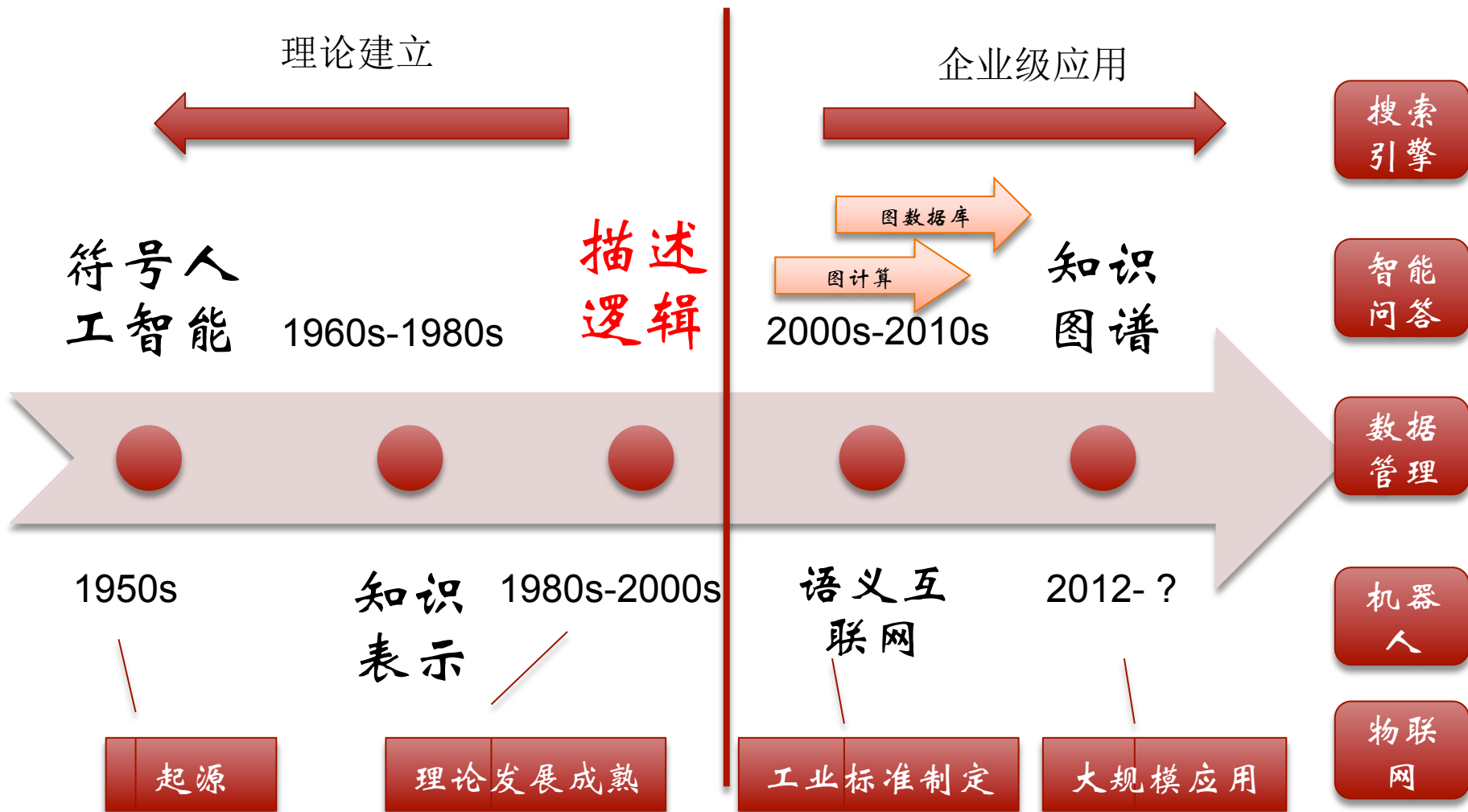
理解



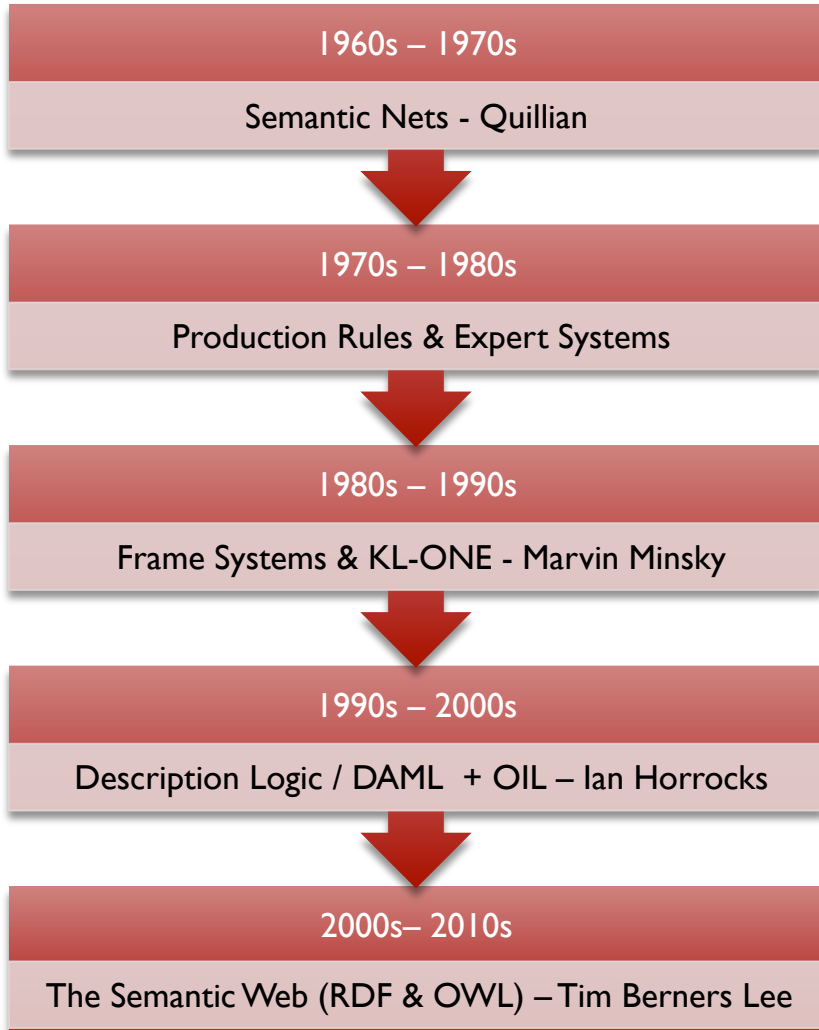
人工智能的复兴



知识图谱：符号人工智能的复兴



知识表示发展



Description Logic : 描述逻辑

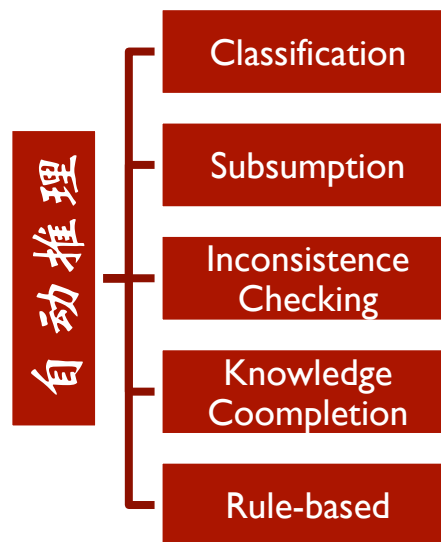
- ▶ 描述逻辑是符号人工智能发展至今最成熟的理论成果之一。
- ▶ DL是OWL语言的理论基础，代表一族计算可判定的知识表示框架。
- ▶ 描述逻辑 (Description Logic) 专门研究人是如何对客观世界进行描述和推理，以及怎样用计算机来模拟这种描述和推理。



人脑中的知识



- $\top^I = \Delta^I$
- $\perp^I = \emptyset$
- $(C \sqcup D)^I = C^I \cup D^I$ (union means disjunction)
- $(C \sqcap D)^I = C^I \cap D^I$ (intersection means conjunction)
- $(\neg C)^I = \Delta^I \setminus C^I$ (complement means negation)
- $(\forall R.C)^I = \{x \in \Delta^I \mid \text{for every } y, (x, y) \in R^I \text{ implies } y \in C^I\}$
- $(\exists R.C)^I = \{x \in \Delta^I \mid \text{there exists } y, (x, y) \in R^I \text{ and } y \in C^I\}$



Graphical Logic : 基于图的逻辑表达

Existential Graph

PQ

a) Conjunction $P \wedge Q$



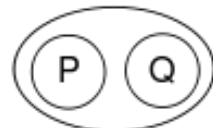
c1) Conditional $P \rightarrow Q$



c2) Conditional $P \rightarrow Q$,
alternative notation



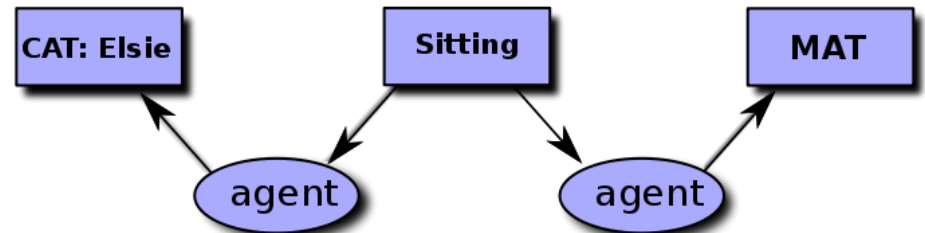
b) Negation $\neg P$



d) Disjunction $P \vee Q$

Concept Graph

Elsie the cat is sitting on a mat



[Cat Elsie] [Sitting *x] [Mat *y] (agent ?x Elsie) (location ?x ?y)

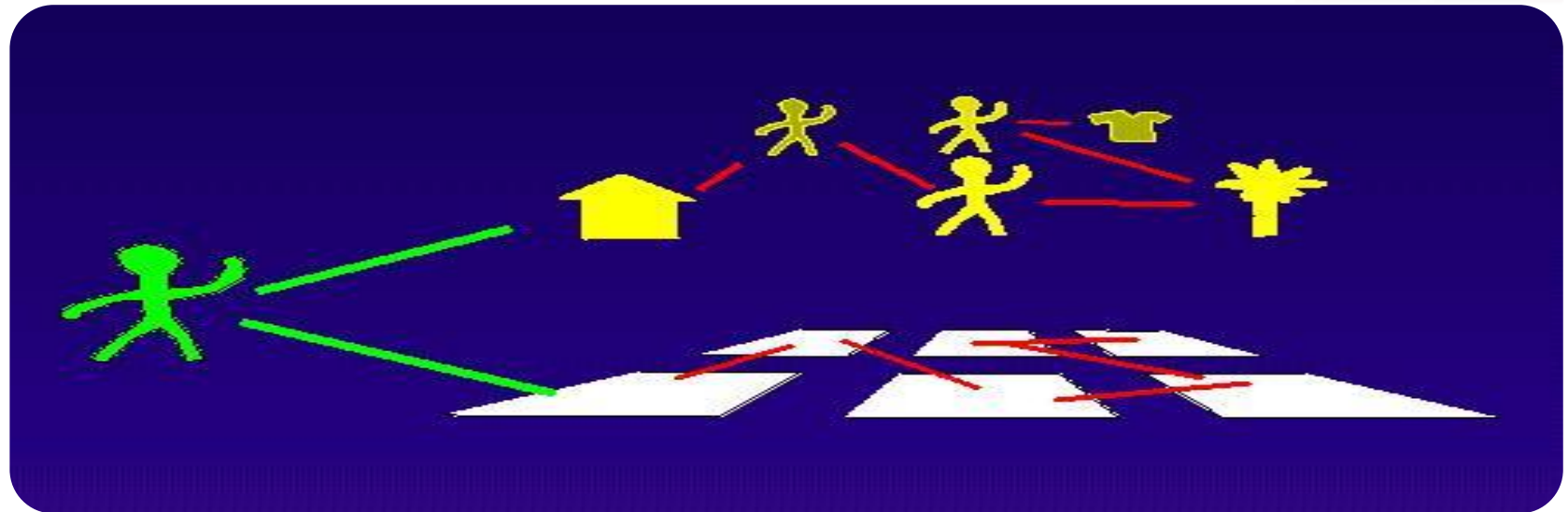
1882年逻辑学家Peirce

1984年John F. Sowa

The Semantic Web : 数据互联网

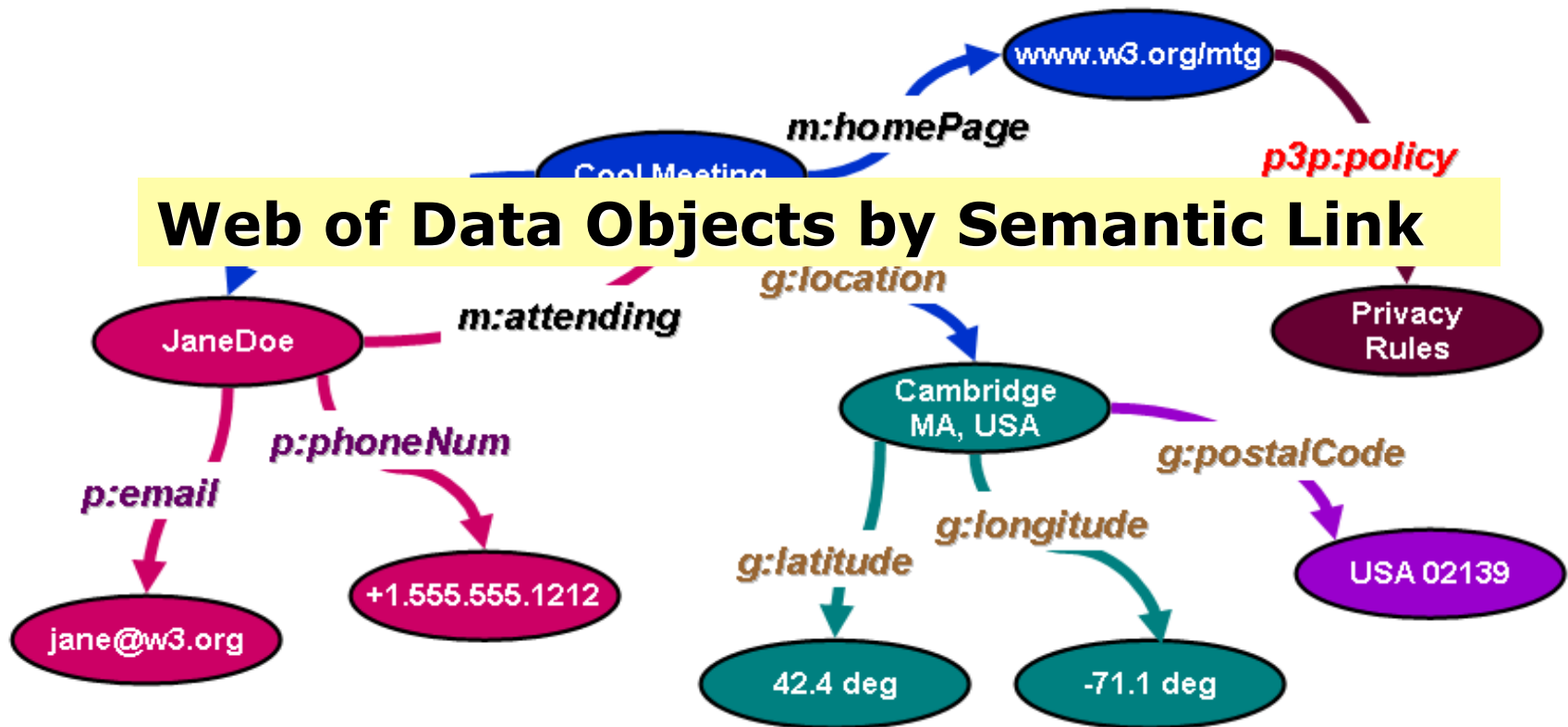
- ▶ "This is a pity, as in fact documents on the web **describe real objects and imaginary concepts**, and **give particular relationships between them** but we could not process them at all...

Tim Berners-Lee, Inventor of the Web, @WWW Geneva, 1994



The Semantic Web : 数据互联

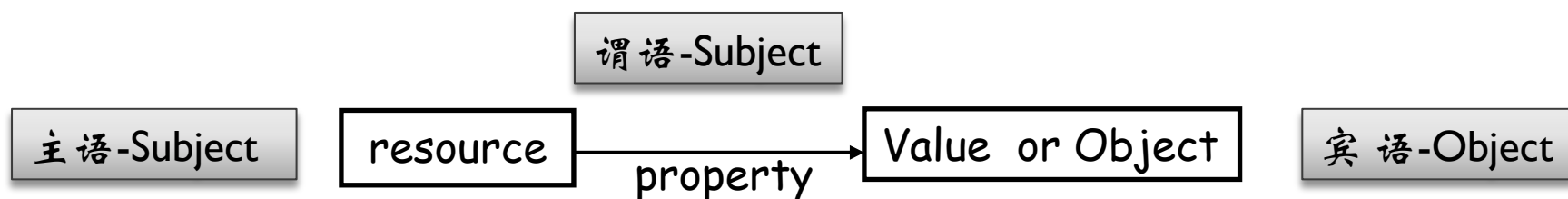
本质是通过增强数据表达来增强数据互联



RDF : 逻辑增强的图表达模型

RDF Triple : 至简的数据模型

RDF $\neq$ 属性图



“Yangtze **has-a-length-of** 6300 kilometers”



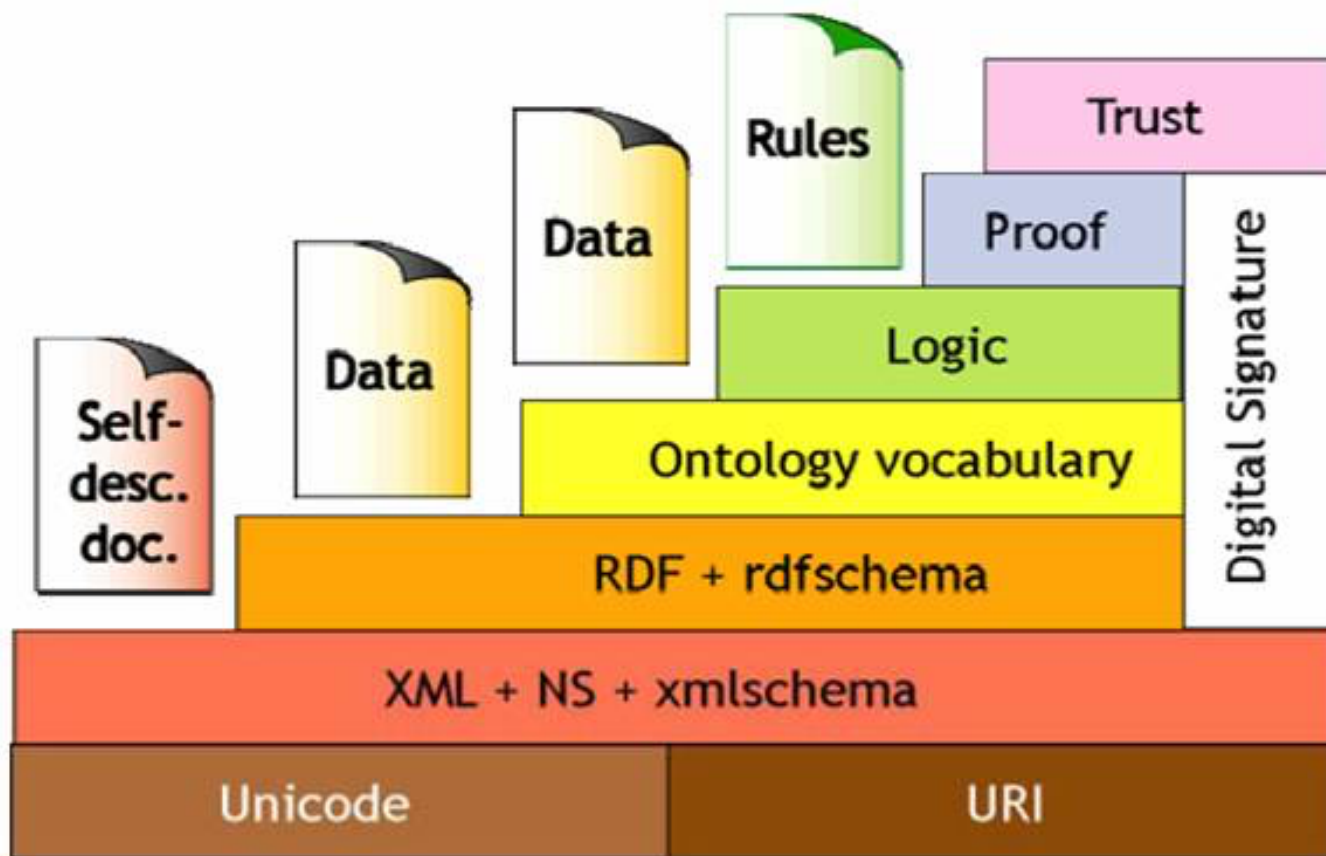
<http://www.china.org/geography/rivers#Yangtze>

<http://www.geodesy.org/river#length>

“6300 kilometers”

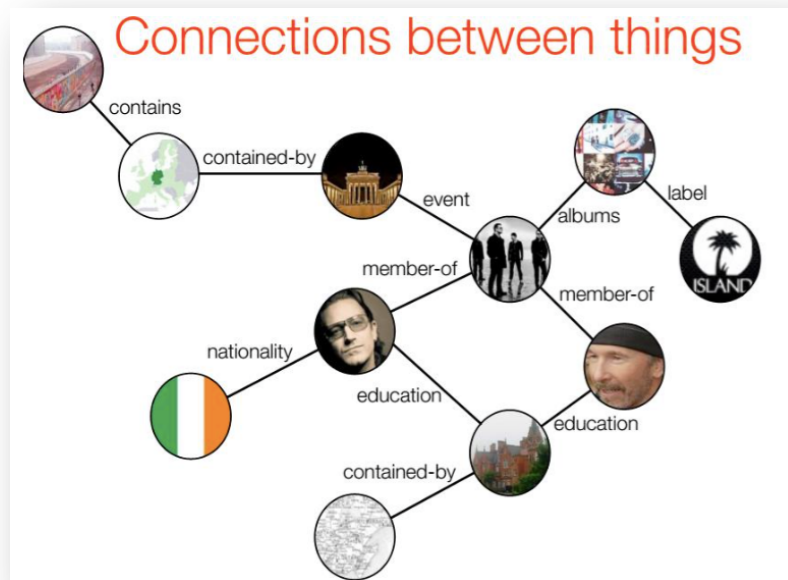
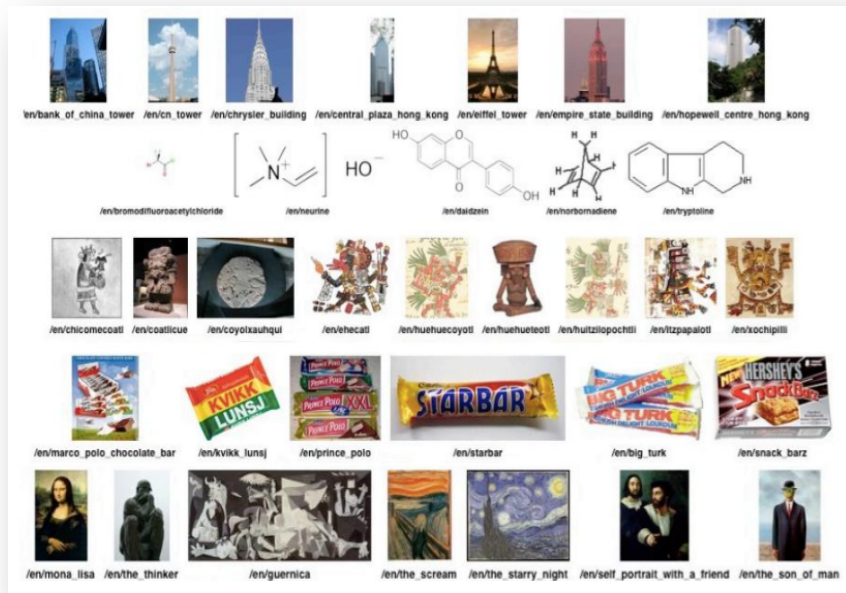
W3C- RDF/OWL：增强图表达的工业标准

- ▶ RDF/OWL的价值仅在于它们是工业标准。



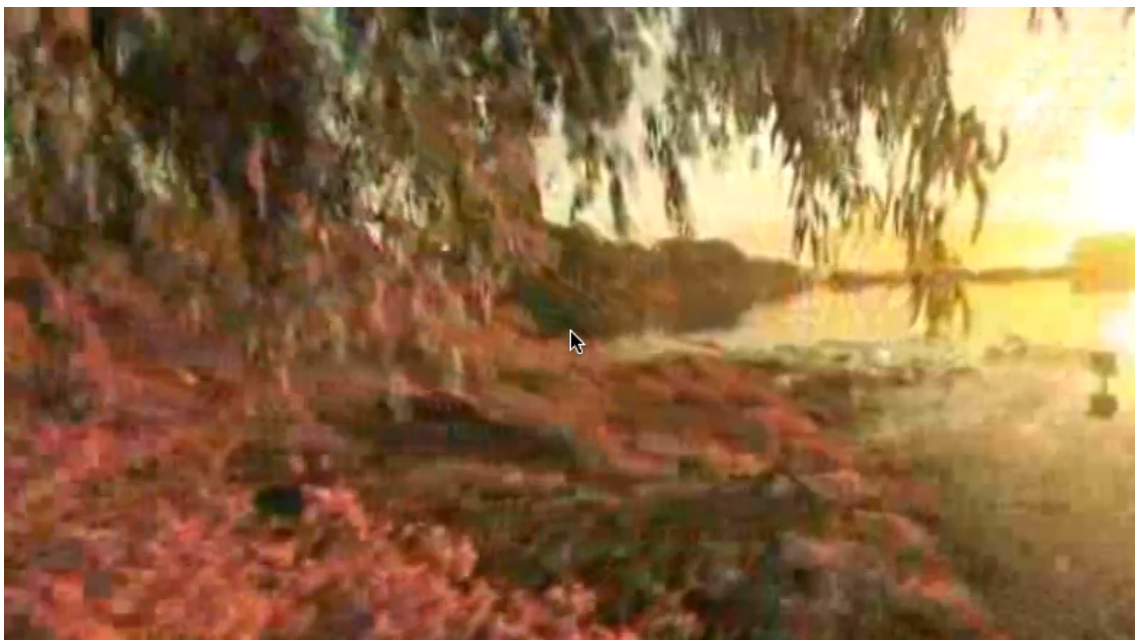
互联的数据：Freebase

- ▶ Freebase是由位于旧金山的Metaweb公司从2005年开始研发的**大型知识库系统**。谷歌2010年收购Freebase作为其知识图谱项目的数据基础之一。
- ▶ Freebase采用**RDF图模型**，而非传统的关系数据库模型来描述数据，是**语义互联网**技术的早期实现和应用。目前已经包含**19亿条三元组/250G**原始数据。

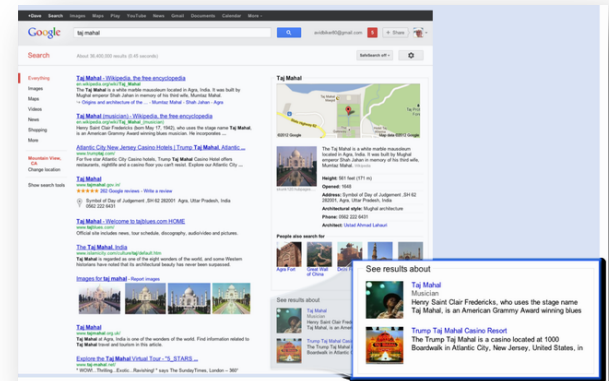


互联的数据：WikiData

- ▶ 2015年起，谷歌计划逐步用WikiData替代Freebase。
- ▶ WikiData由微软联合创始人Allen，摩尔基金会和谷歌于2013年起共同资助，旨在利用Wiki开放协作模式构建全球最大的免费知识库。
- ▶ WikiData的核心理念是**知识和数据互联（Interlinking Data）**。



谷歌知识图谱：Things, not Strings



History : The big fight between two Gurus

AAAI2006: Google and the Semantic, Satanic, Romantic Web

Tim Berners-Lee



Director of W3C
Professor of MIT CSAIL

Peter Norvig



Director of Google Research

Why the Semantic Web will be a failure :
People are stupid, People are competitive, People cheat and lie

增强图表达（一）：图数据库拥抱关联

增强关系表达：以关联为中心的数据表达、存储与查询

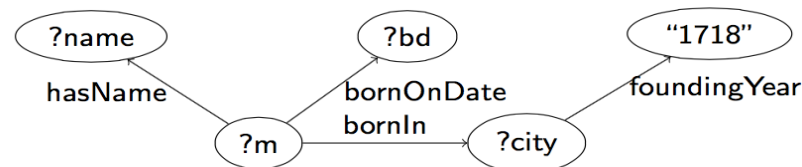


- 关系数据库不善于处理关系；
- BigTable更加弱化了关系处理；
- 图存储以数据关联为中心，通过免索引邻接 (Index-free Adjacency) 等方法大幅提升复杂关联查询的性能。

```
SELECT ?name
WHERE {
  ?m <bornIn> ?city. ?m <hasName> ?name.
  ?m<bornOnDate> ?bd. ?city <foundingYear> '1718'.
  FILTER(regex(str(?bd), '1976'))
}
```

SPARQL

`FILTER(regex(str(?bd), "1976"))`



```
START user=node:users(name='User 3')
MATCH (user) - [*1..5] - (asset)
WHERE asset.status != 'down'
RETURN asset
```

CYPHER

扩展的灵活性: Logic-level Scalability

Open Schema: 概念模型无需事先完整定义，可根据未来需要动态扩展，适应变化。

Open Instances: 按需增加、无限扩展，允许任何人根据自己的需要进行动态扩展。

TextRunner/
ReVerb

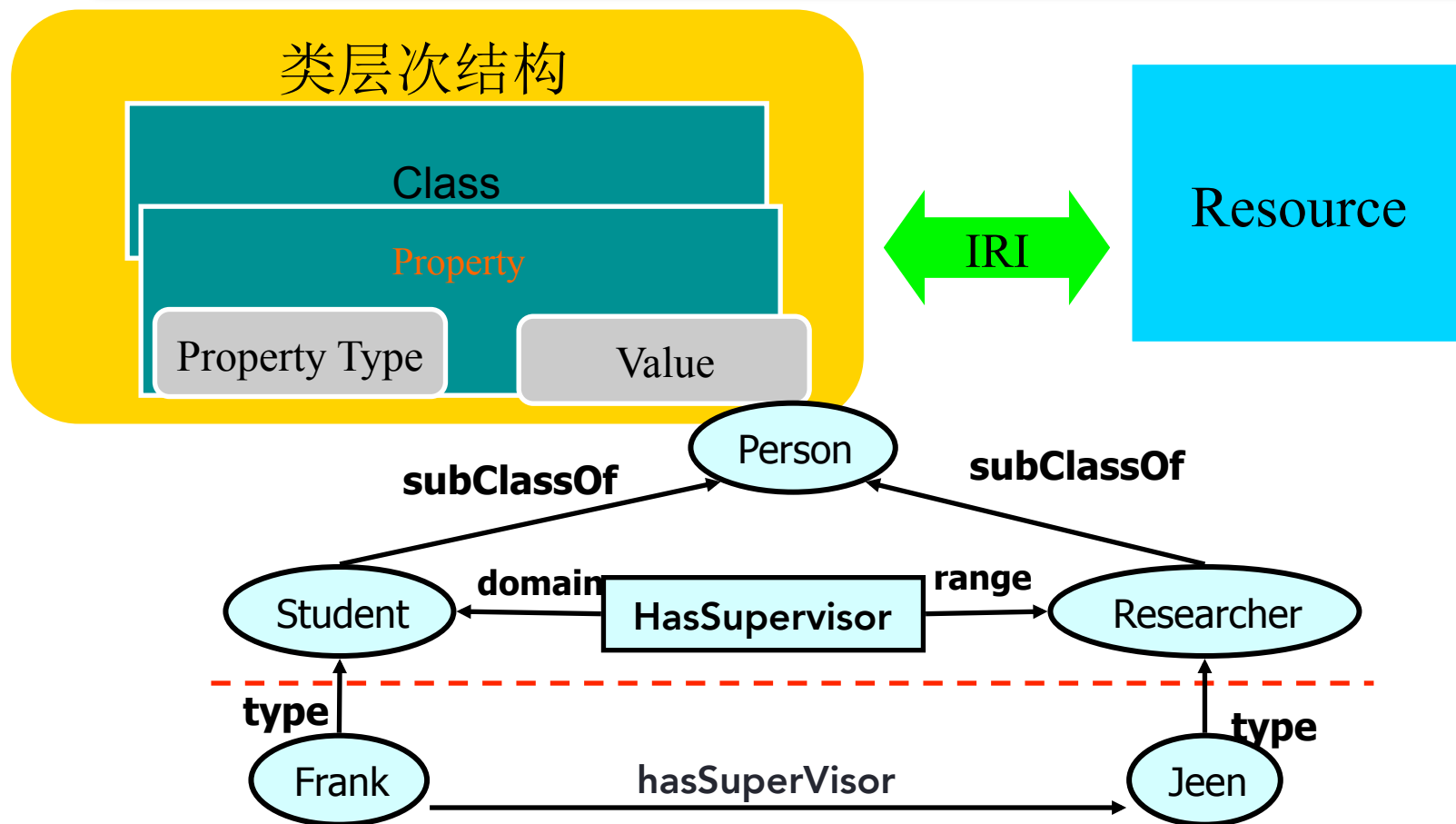
ConceptNet-5

WikiTaxonomy/
WikiNet

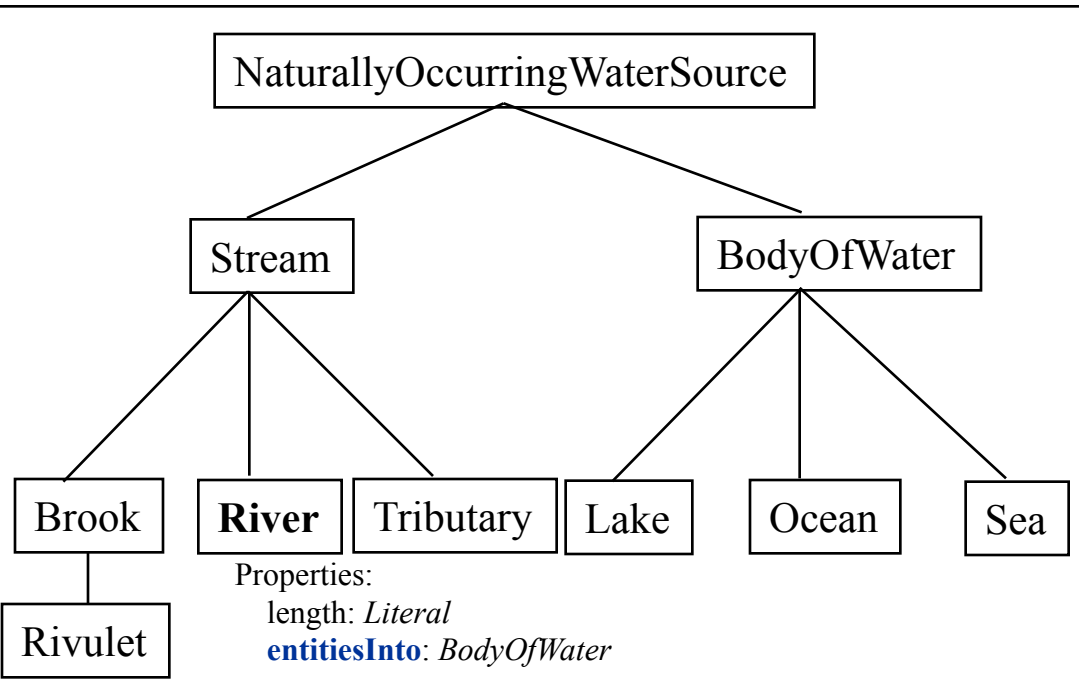
As of September 2011

增强图表达（三）：自动推理

定义类层次结构——本体、规则，以支持自动推理

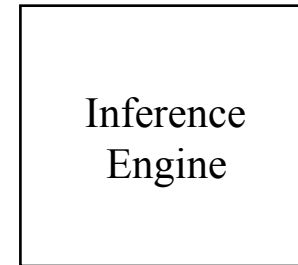


简单的推理举例



```
<?xml version="1.0"?>
<River rdf:ID="Yangtze"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns="http://www.geodesy.org/water/naturally-occurring#">
  <length>6300 kilometers</length>
  <emptiesInto rdf:resource="http://www.china.org/geography#EastChinaSea"/>
</River>
```

Yangtze.rdf



Inferences:

- Yangtze is a Stream
- Yangtze is an NaturallyOccurringWaterSource
- <http://www.china.org/geography#EastChinaSea> is a BodyOfWater

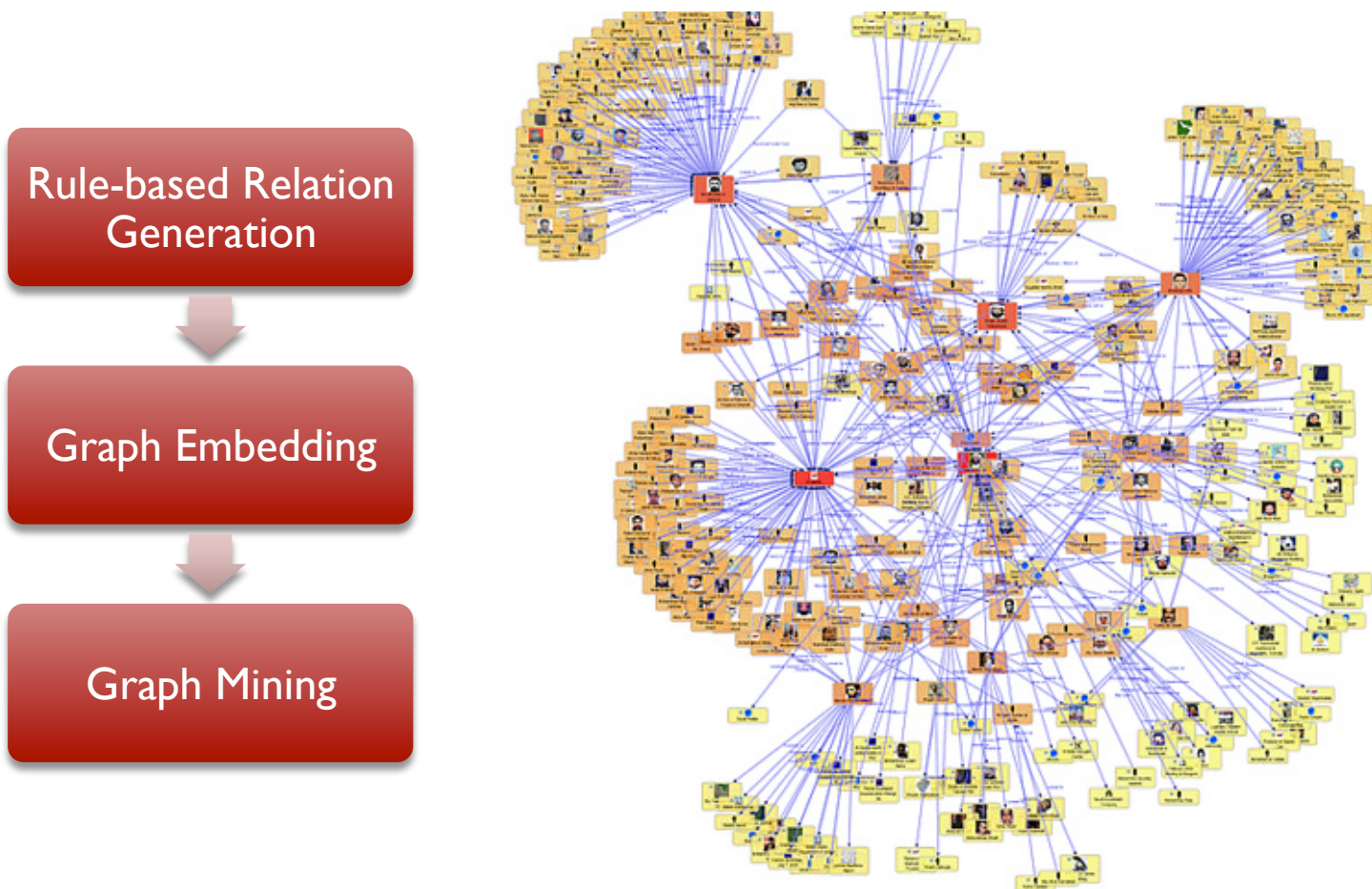


增强图表达（四）：复杂的推理

- ▶ 描述逻辑语言OWL提供了各种表达构件，叠加表达构件提升知识库的推理能力。
- ▶ Property Restriction, Cardinality, Property Chain, SWRL.....



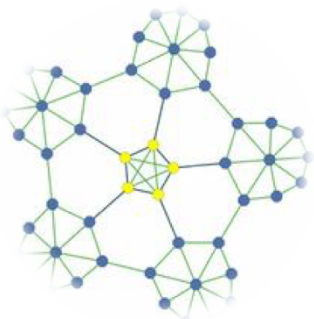
增强图表达（五）：关系推理与关系发现



不同层次的知识图谱

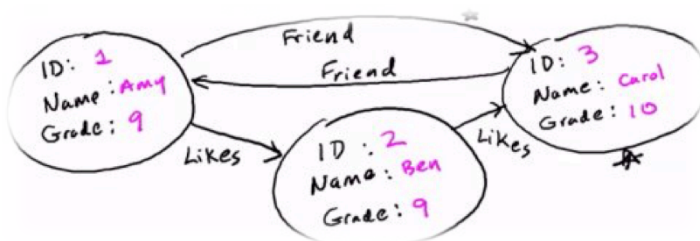
表达能力逐渐增强

基于简单图模型



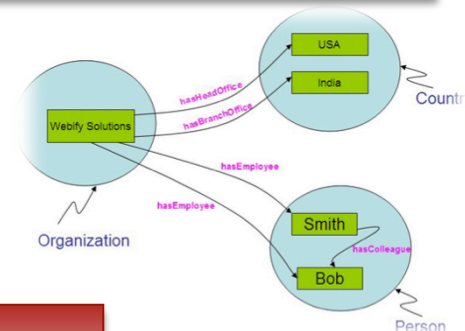
采用较为简单的图表达；弱关联表达、无逻辑基础、不支持自动推理，典型的如Giraph/Spark GraphX等，严格的说不能称为知识库。

基于RDF有向标记图
(Directed Labeled Graph)



RDF是增强的图表达，具有较强的关联表达能力、有基本逻辑学和语义学基础，支持较简单的自动推理，如Freebase。

基于OWL和描述逻辑



具有完整的逻辑表达能力，很强的关联表达、完备的逻辑基础、完备的自动推理能力，典型的如大量的医学本体(Gene Ontology)，财经本体FIBO等。



实践

利用增强的图表达增强数据互联



实践案例：脑科学知识库

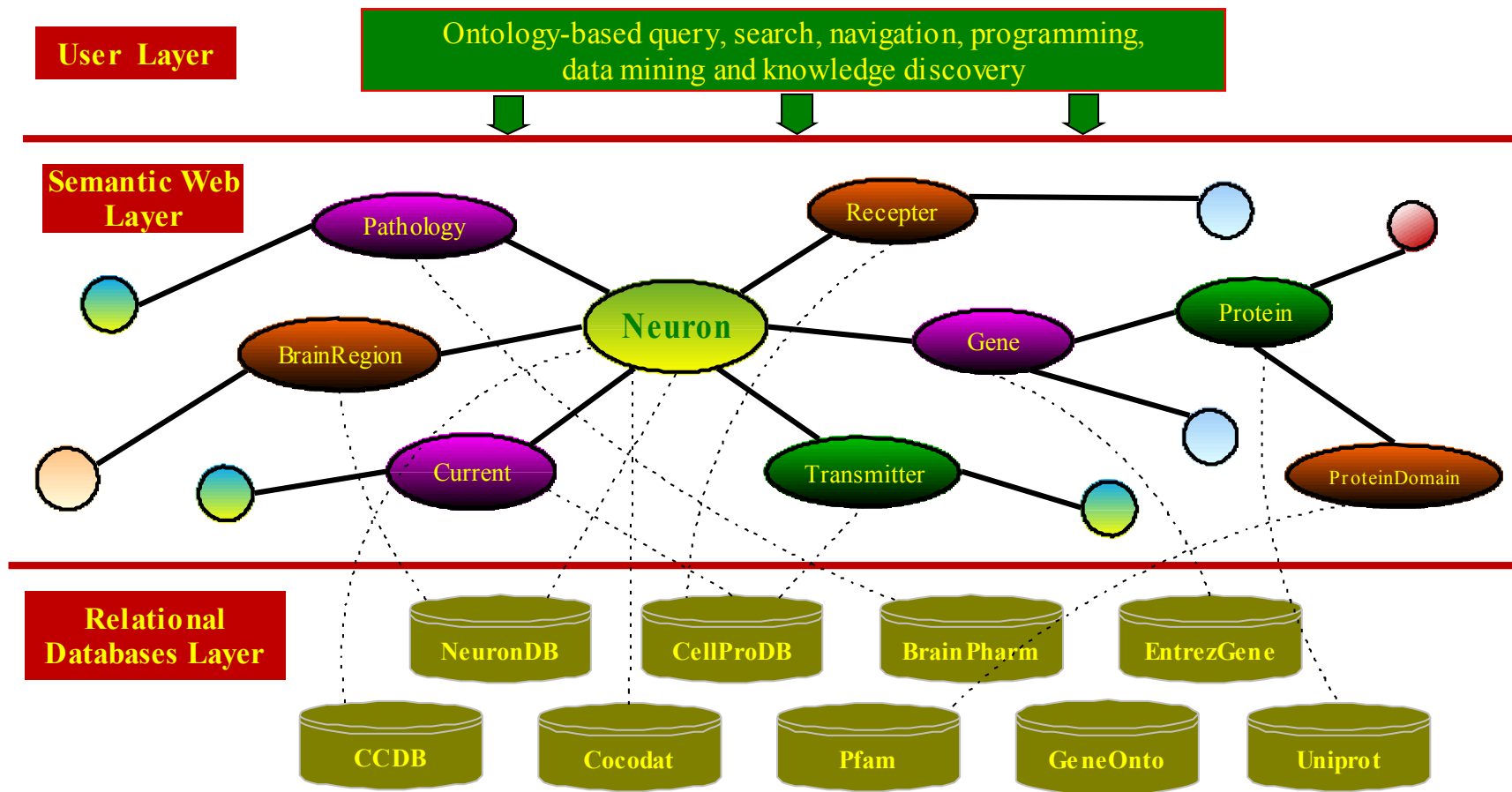
- ▶ 合作单位：Yale Center for Medical Informatics （时间：2006-2007）
- ▶ 项目目标：建立神经科学知识库，辅助脑科学研究。
- ▶ 问题描述：耶鲁拥有全球最大的神经科学数据库SenseLab，但脑科学研究需要综合从微观分子层面一直到宏观行为层面各个层次的知识。因此，需要建立SenseLab与基因、蛋白质、脑科学实验、药物、疾病等之间的数据关联，以辅助脑科学家从不能维度进行数据关联分析。



Yale Center for
Medical Informatics



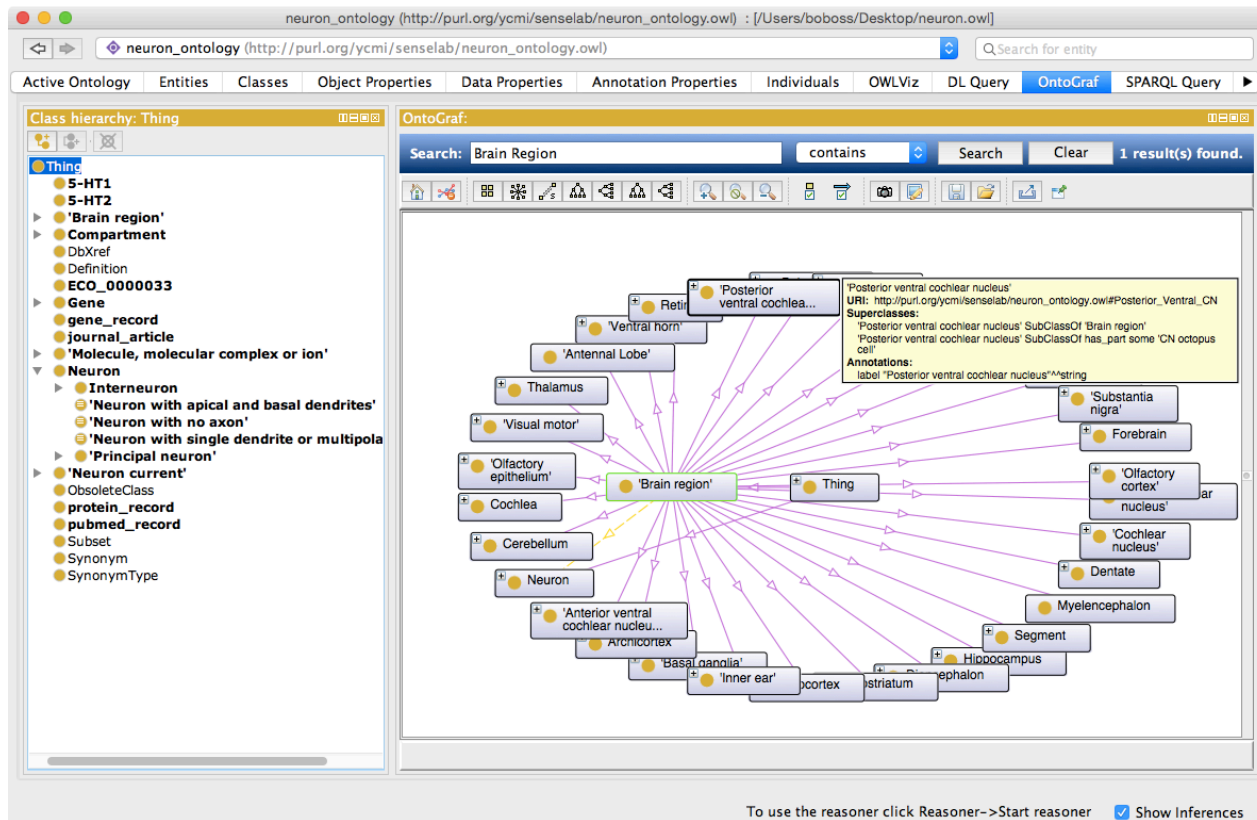
利用增强的图表达建立数据关联



从微观分子层面，如Entrez Gene 一直到宏观行为层面，如Allen Brain Atlas

Neuron Ontology

2万条概念定义和属性描述，涉及脑结构、神经细胞、基因和蛋白质等生物大分子、脑部疾病等。



支持推理的关联查询分析

```
prefix go: <http://purl.org/obo/owl/GO#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix mesh: <http://purl.org/commons/record/mesh/>
prefix sci: <http://purl.org/science/owl/sciencecommons/>
prefix ro: <http://www.obofoundry.org/ro/ro.owl#>

select ?genename ?processname
where
{
  graph <http://purl.org/commons/hcls/pubmesh>
  {
    ?paper ?p mesh:D017966 .
    ?article sci:identified_by_pmid ?paper .
    ?gene sci:describes_gene_or_gene_product_mentioned_by ?article .
  }
  graph <http://purl.org/commons/hcls/goe>
  {
    ?protein rdfs:subClassOf ?res .
    ?res owl:onProperty ro:has_function .
    ?res owl:someValuesFrom ?res2 .
    ?res2 owl:onProperty ro:realized_as .
    ?res2 owl:someValuesFrom ?process .
  }
  graph <http://purl.org/commons/hcls/20070416/classrelations>
  {
    { ?process <http://purl.org/obo/owl/obo#part_of> go:GO_0007166 }
    union
    { ?process rdfs:subClassOf go:GO_0007166 }
    ?protein rdfs:subClassOf ?parent .
    ?parent owl:equivalentClass ?res3 .
    ?res3 owl:hasValue ?gene .
  }
  graph <http://purl.org/commons/hcls/gene>
  {
    ?gene rdfs:label ?genename .
  }
  graph <http://purl.org/commons/hcls/20070416>
  {
    ?process rdfs:label ?processname .
  }
}
```

Mesh: Pyramidal Neurons



Pubmed: Journal Articles



Entrez Gene: Genes



GO: Signal Transduction

Inference required



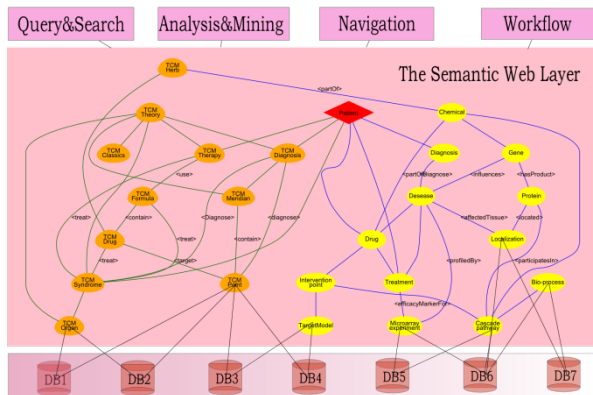
Publications

- ▶ Huajun Chen, Tong Yu, Jake Y. Chen: Semantic Web meets Integrative Biology: a survey. **Briefings in Bioinformatics** 14(1): 109-125 (2013)
- ▶ Huajun Chen, Zhaohui Wu, Philippe Cudré-Mauroux: Semantic Web Meets Computational **Intelligence: State of the Art and Perspectives**. IEEE Computational Intelligence Magazine 7(2): 67-74 (2012)
- ▶ Matthias Samwald, Huajun Chen, Alan Ruttenberg, Ernest Lim, Luis N. Marengo, Perry L. Miller, Gordon M. Shepherd, Kei-Hoi Cheung: Semantic SenseLab: Implementing the vision of the Semantic Web in neuroscience. **Artificial Intelligence in Medicine** 48(1): 21-28 (2010)
- ▶ Huajun Chen, Li Ding, Zhaohui Wu, Tong Yu, Lavanya Dhanapalan, Jake Yue Chen: Semantic web for integrated network analysis in biomedicine. **Briefings in Bioinformatics** 10(2): 177-192 (2009).
- ▶ Kei-Hoi Cheung, Matthias Samwald, Huajun Chen, et al: Approaches to neuroscience data integration. **Briefings in Bioinformatics** 10(4): 345-353 (2009)

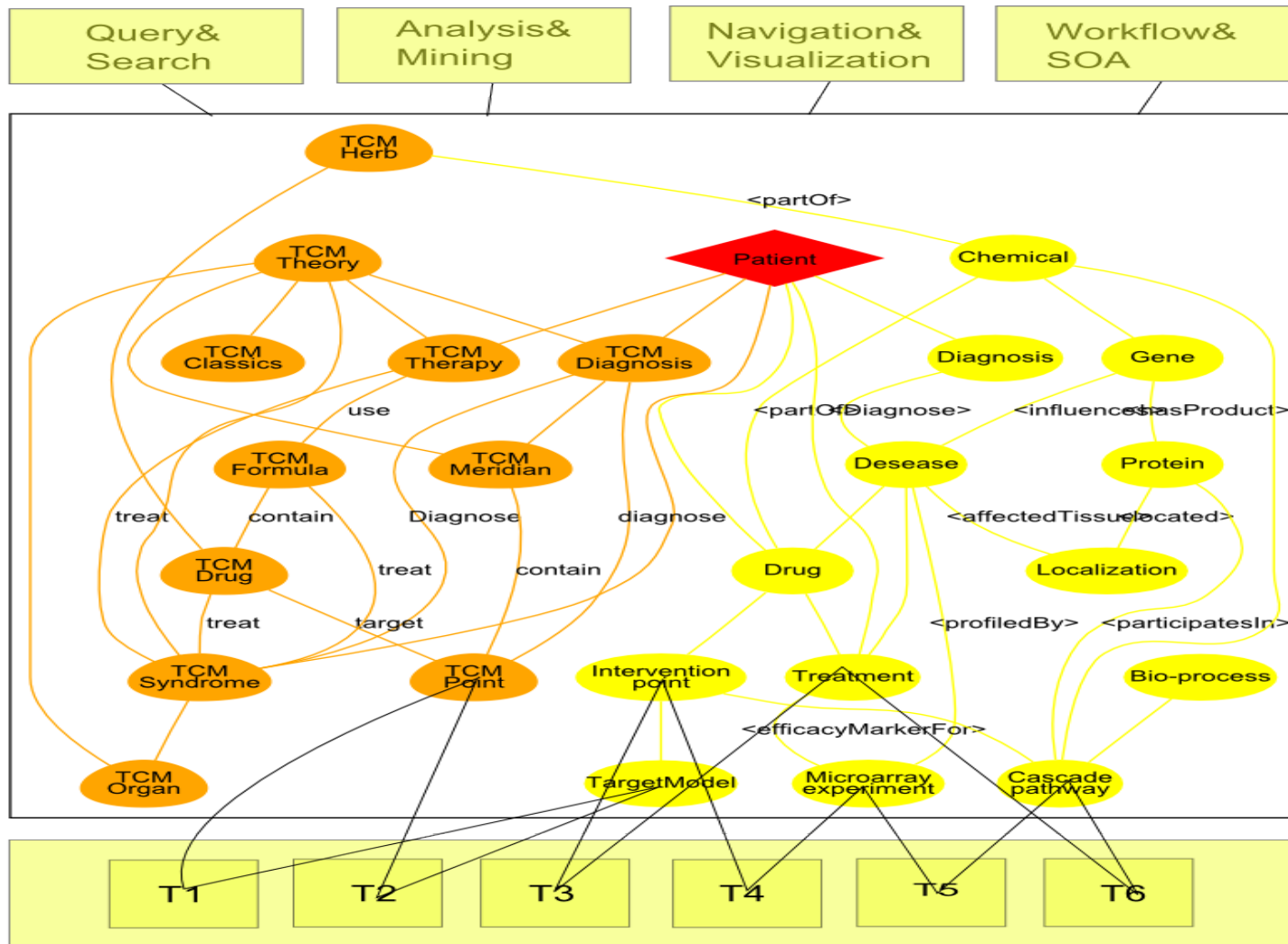


实践案例：中西医语义网络

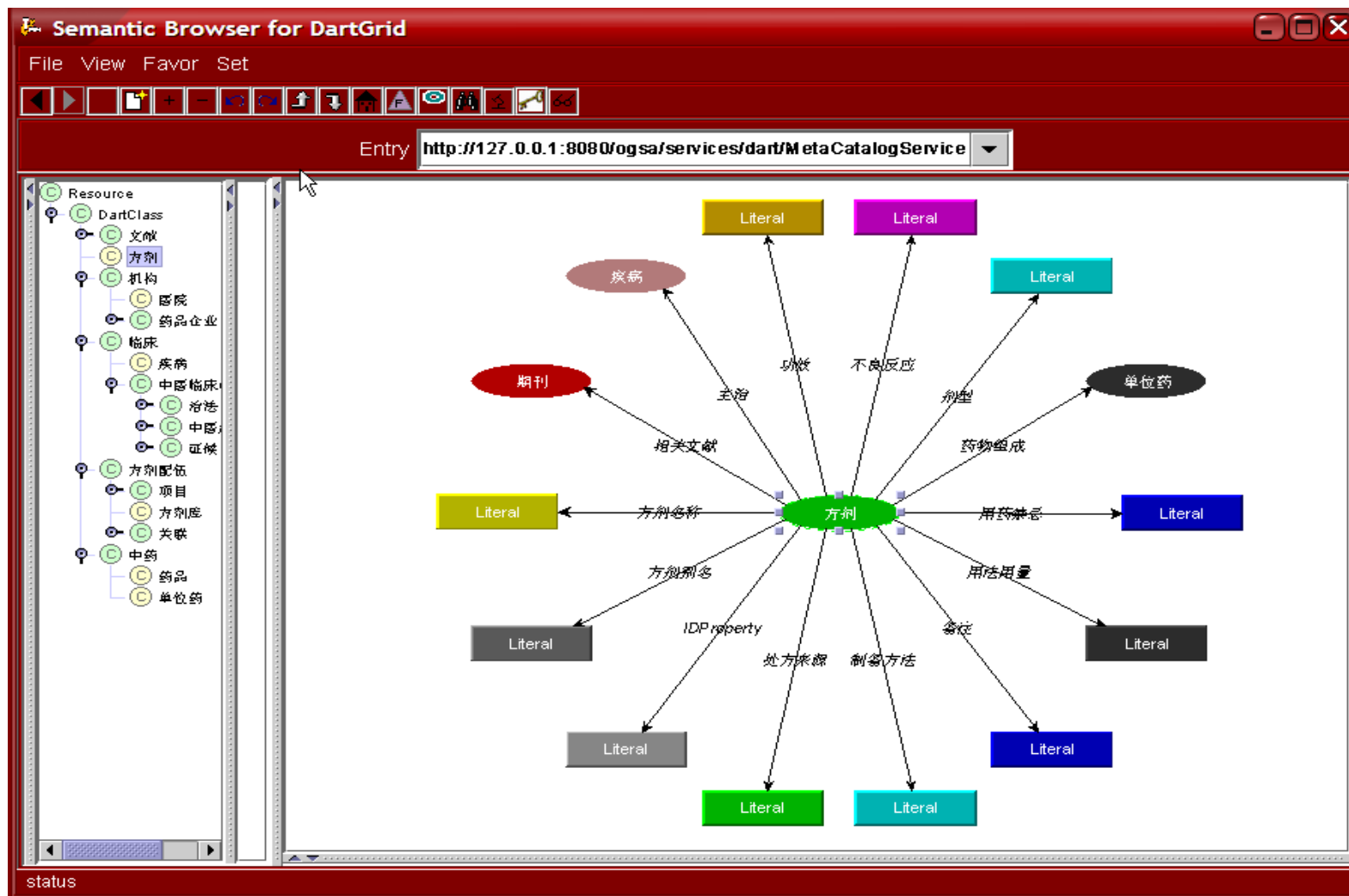
- ▶ 合作单位：中国中医科学院
- ▶ 项目目标：构建连接传统中医知识与现代医学数据的知识库系统。
- ▶ 问题描述：中医具有独特的知识体系和概念模型，并积累起大量数据资源，如：中草药、中医症候、中医临床等，如果能将这些数据资源与现代医学数据（如基因数据、蛋白质数据、药物化合物等）建立关联，将有助于从中西医结合的关联分析与研究。



中西医知识库系统逻辑结构



语义浏览器 (2003年)



通过本体建立关系数据之间的关联

The screenshot displays the Dart User Toolkits application with several panels and callouts illustrating the workflow:

- 1. Display relational tables:** The DbRes Tree View on the left shows a list of tables, including TCM_BASIC.ZT_VIEW_ANMOLF.
- 2. Display ontologies:** The OntoSchema Tree View on the left shows a list of ontologies, including 按摩疗法 (Massage Therapy).
- 3. User drags tables and classes into this panel, and establishes their mappings. One table is likely to be mapped to more than one classes.** The central panel shows two ontology views: ZT_VIEW_ZHENJITULF and ZT_VIEW_ANMOLF. Red arrows indicate the mapping of tables to these ontology views.
- 4. Outline of the mapping definitions. User could query mappings defined before.** The Outline panel on the right shows the mapping definitions for the selected table.
- 5. Meta-information about the selected table.** The Properties panel at the bottom shows the meta-information for the selected table.

Properties Panel Data:

Property	Value
Db Localpart	TcmBasic
Db Namespace	http://dart.zju.edu.cn/tcm/database
Table Name	ZT_VIEW_ZHENJITULF
Table Schema	

语义关联搜索与关联查询

The screenshot shows the SinoMed database search interface in Microsoft Internet Explorer. The main search results page is on the left, and a detailed view of a specific record is on the right. Red arrows and callouts highlight key features:

- Ontological classes:** A callout points to the search results, indicating that clicking on these classes leads to a dynamic form-based query interface.
- Based on the semantic relations defined at the ontological level, user can keep searching and navigating over the integrated databases without the awareness of the database boundaries.** A callout points to the search results, indicating that the system allows for continuous searching and navigation across integrated databases without the user needing to be aware of the database boundaries.
- Synonyms and Paronyms:** A callout points to the search results, indicating that the system provides synonyms and paronyms for the search terms.
- Semantic association:** A callout points to the search results, indicating that the system provides semantic associations for the search terms.

The search results table on the right shows the following data:

序号	题名名称	来源数据库	查询条件	查询内容	标准词提示
1	中药学	包含		显示	译名
2	中药学	包含		显示	译名
3	中药学	包含		显示	译名
4	中药学	包含		显示	译名
5	中药学	包含		显示	译名
6	中药学	包含		显示	译名
7	中药学	包含		显示	译名
8	中药学	包含		显示	译名
9	中药学	包含		显示	译名

语义互联网ISWC2006最佳应用论文奖



- ▶ 国际语义互联网大会ISWC2006:
International Semantic Web Conference。
- ▶ Huajun Chen, Yimin Wang, Heng Wang, Yuxin Mao, Jinmin Tang, Chunying Zhou, Aining Yin, Zhaohui Wu: Towards a Semantic Web of Relational Databases: A Practical Semantic Toolkit and an In-Use Case from Traditional Chinese Medicine. International Semantic Web Conference 2006: 750-763.

实践案例：财经知识图谱

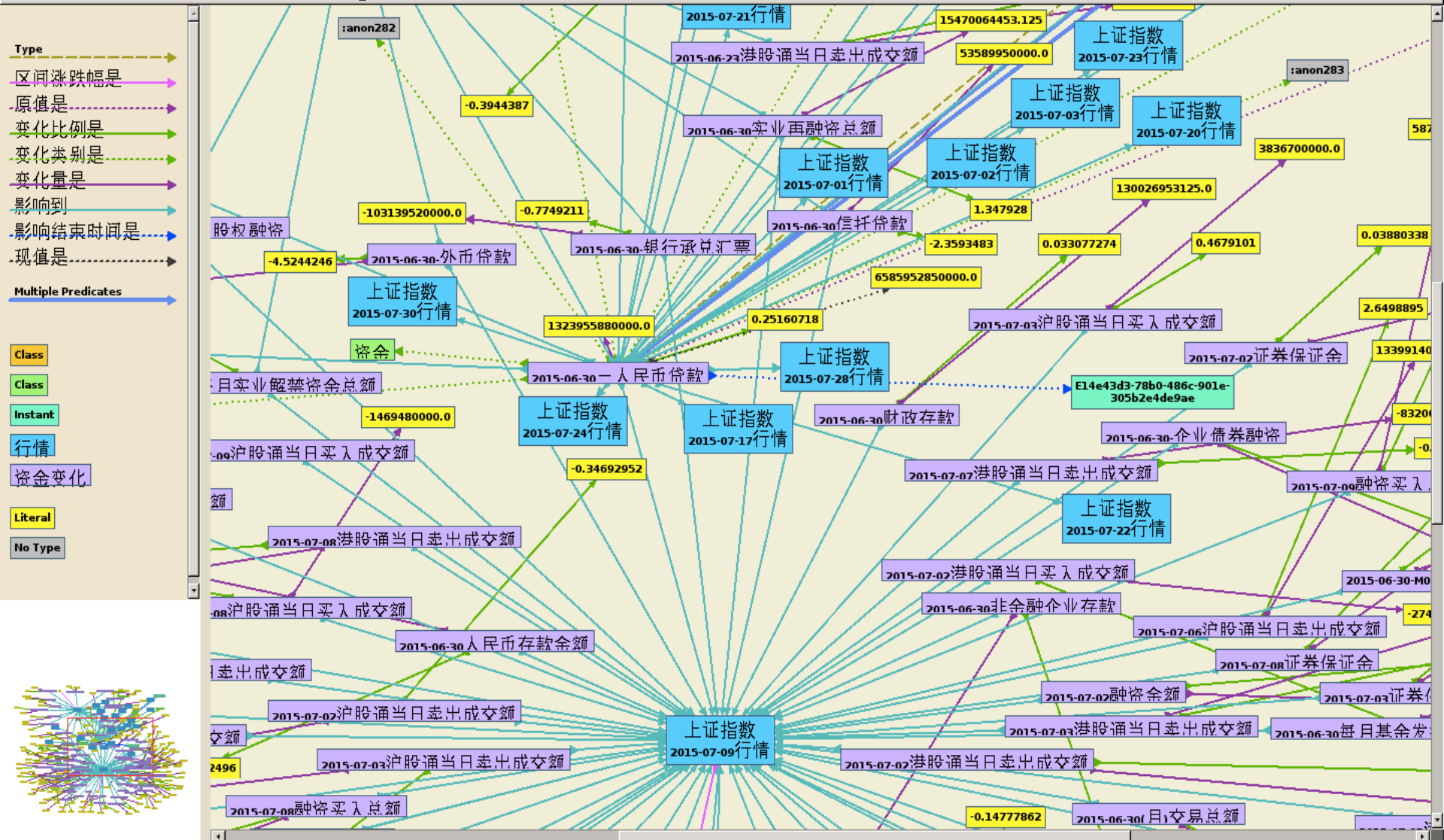
- ▶ 合作单位：*****
- ▶ 项目目标：建立一个财经知识库，以辅助投资决策。
- ▶ 问题描述：影响市场行情的因素众多，既包括影响投资者情绪的新闻及社交媒体，也包括国家宏观经济政策、金融工具的运用，还包括各个层面的资金流入流出情况等。同花顺积累了包括资金面、金融工具、宏观经济、新闻资讯等大量的数据，希望建立这些数据之间的逻辑关联，并建立知识层的逻辑结构，最终构建一个全面的财经知识库，以辅助投资者对影响市场行情的各个维度的因素进行关联分析。



建立关联的数据集

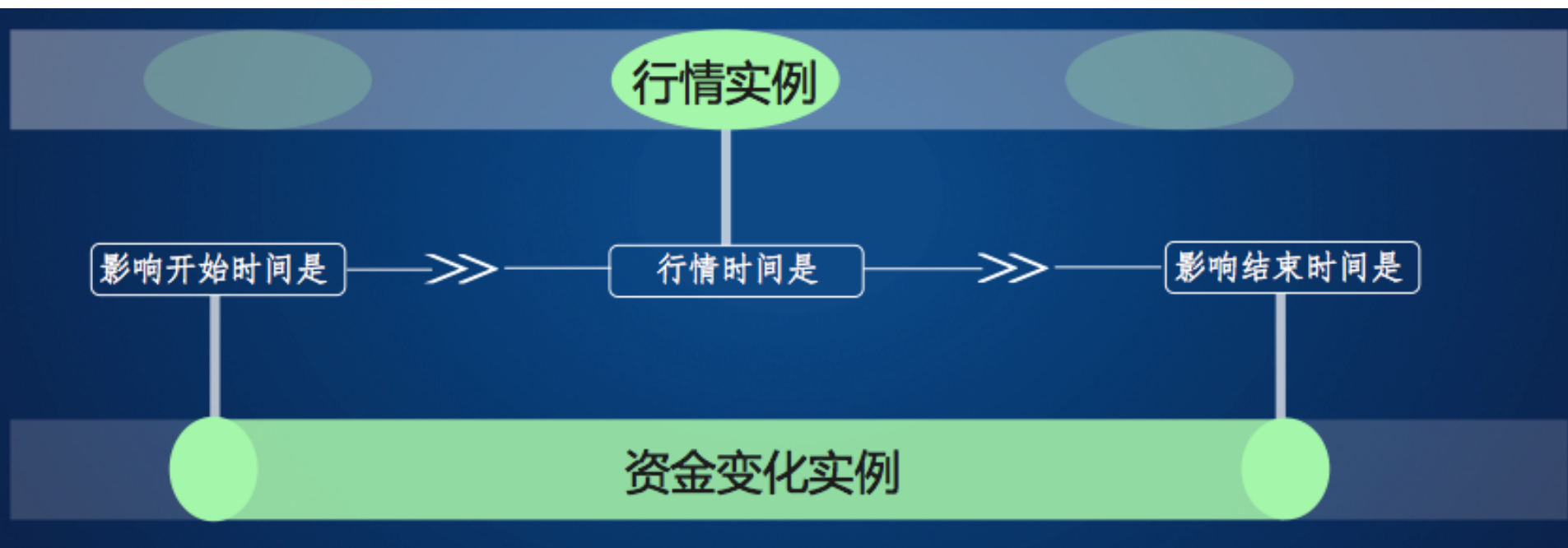




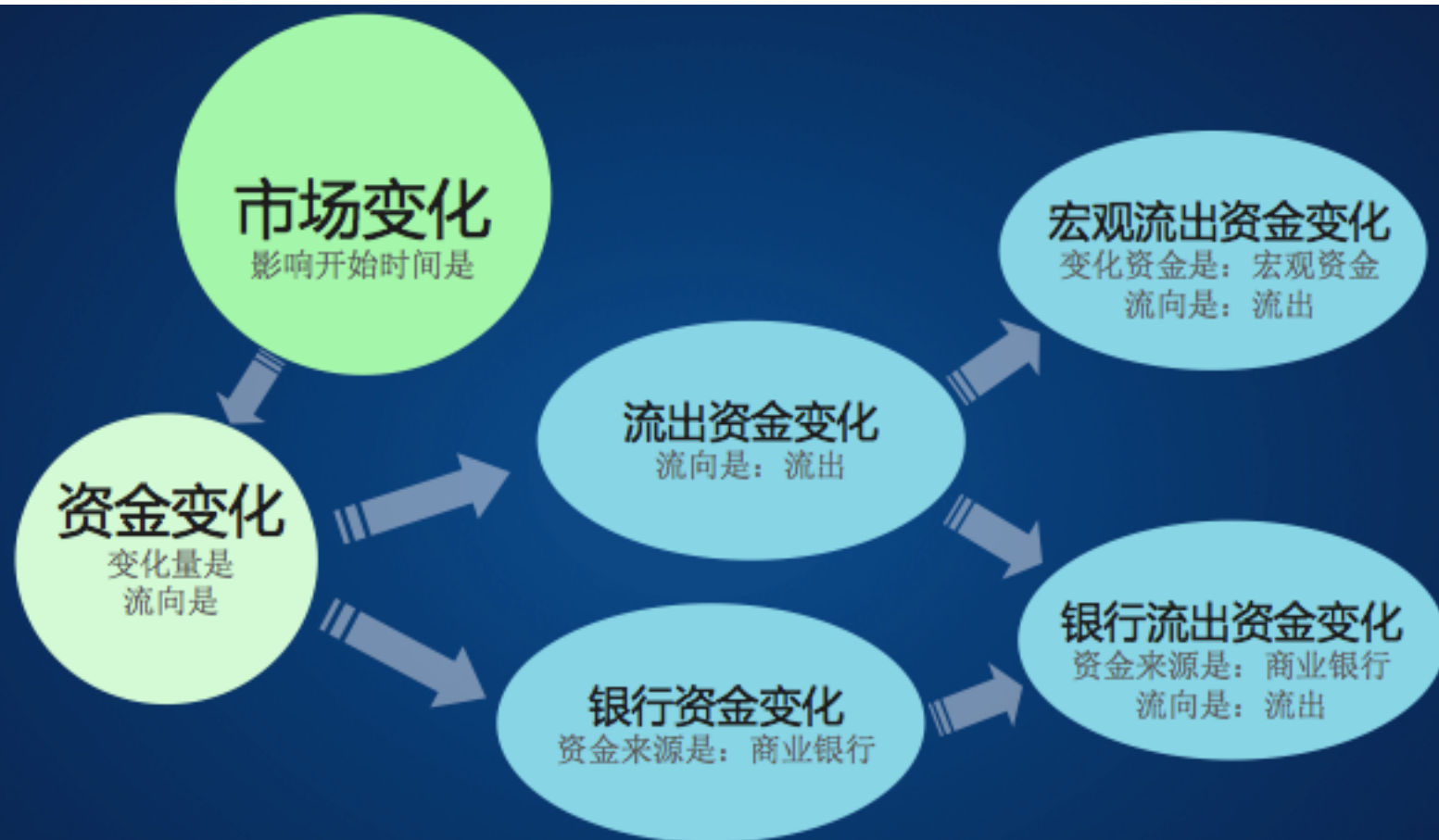


规则推理

- ▶ 建立市场环境的各种因素与某些特定行情之间的影响关系。



分类推理举例 — Concepts (Schema) Open

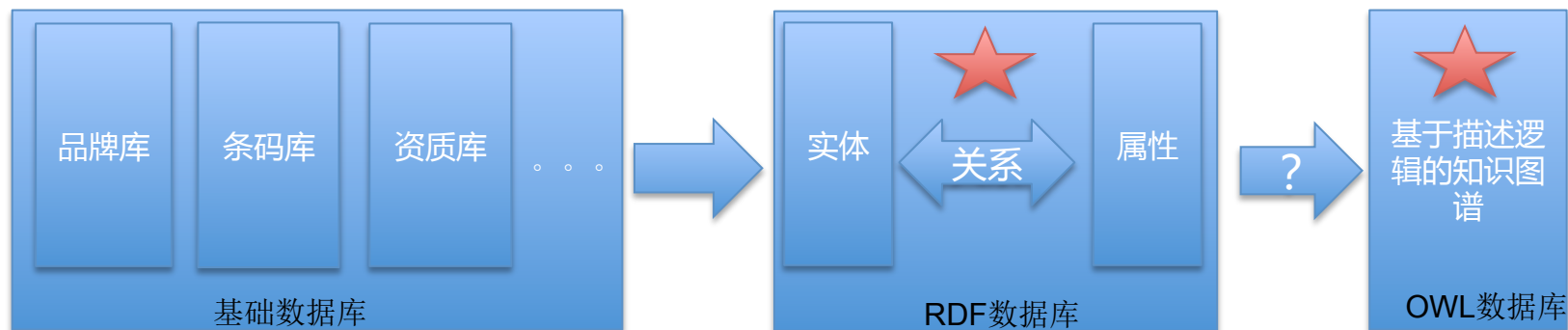


实践案例：阿里商品知识图谱

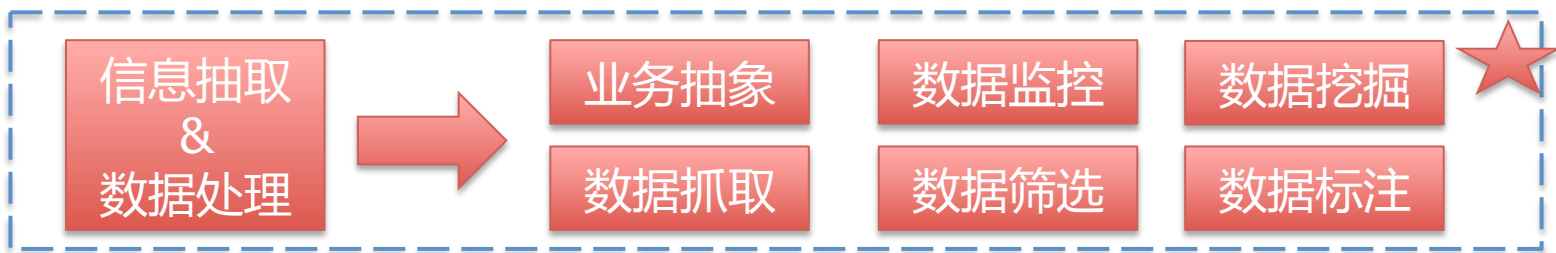
产出

业务支持：商品归一，商品确定性，下拉选项， 人机交互， 语义搜索。。。 
算法接口：图谱自然语言查询，语义分析。。。

核心



输入



实践案例：物联语义

- ▶ 项目目标：对设备数据进行抽象和语义封装，以支持上层的传感数据融合与集成等。
- ▶ 问题描述：物联设备是更加重要的数据产生来源，当前这些数据彼此孤立，并多以原始粗糙的方式存在。然而，越来越多的上层应用需要将 these 设备数据之间，以及设备数据与网络数据之间打通建立关联，并支持关联分析和挖掘。



OneM2M

- ▶ 2012年3月底成立了全球最大的物联网标准化组织——OneM2M，主要由来自美国、欧洲、中国、日本、韩国等7家全球最大的区域通信标准化组织共同组成。
- ▶ OneM2M成立了专门数据管理和语义技术工作组，研究制定物联网数据的语义表达规范，关联数据的建立方法和管理手段等。

技术全会

WG1

需求

WG2

架构

WG3

协议

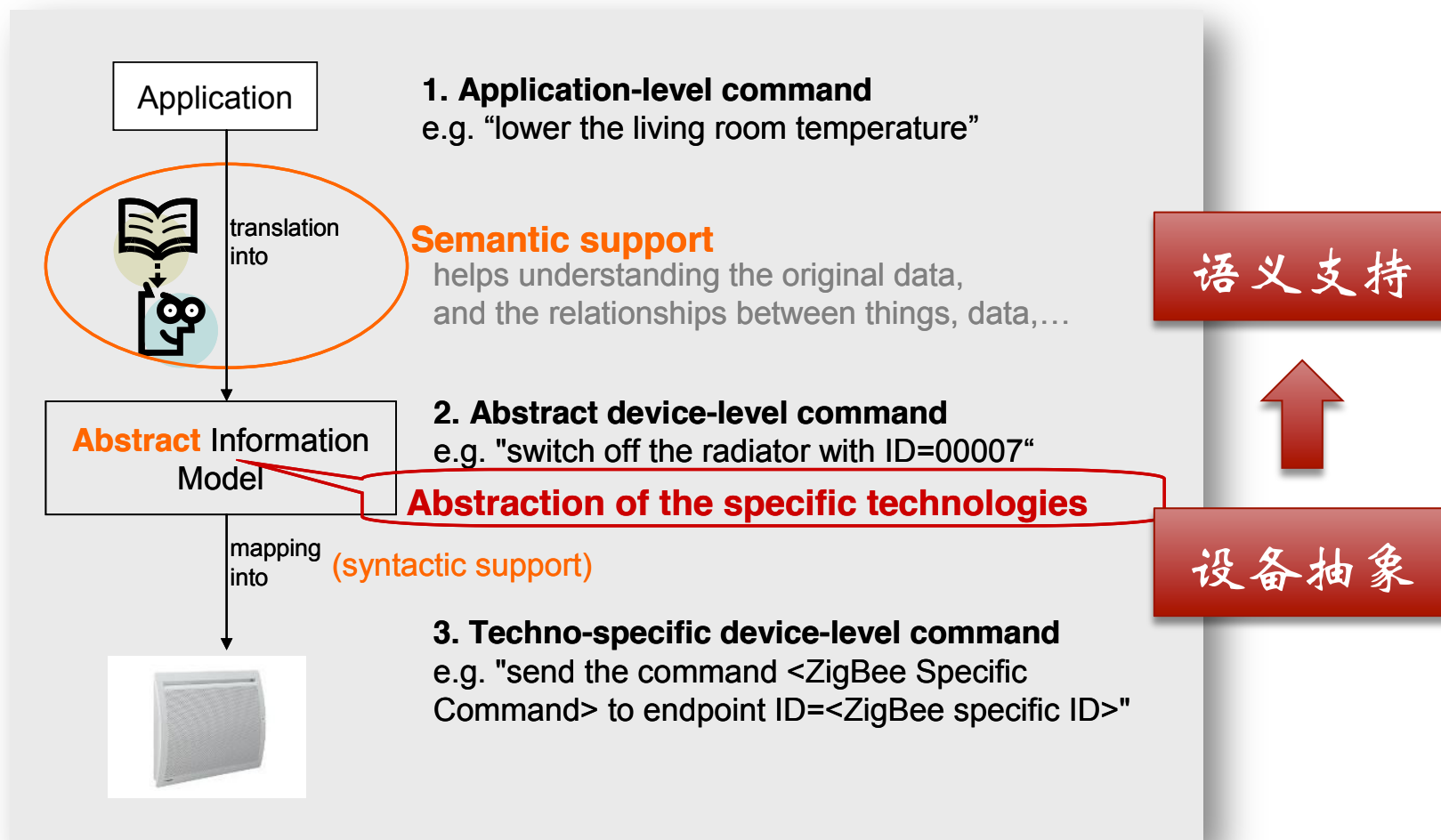
WG4

安全

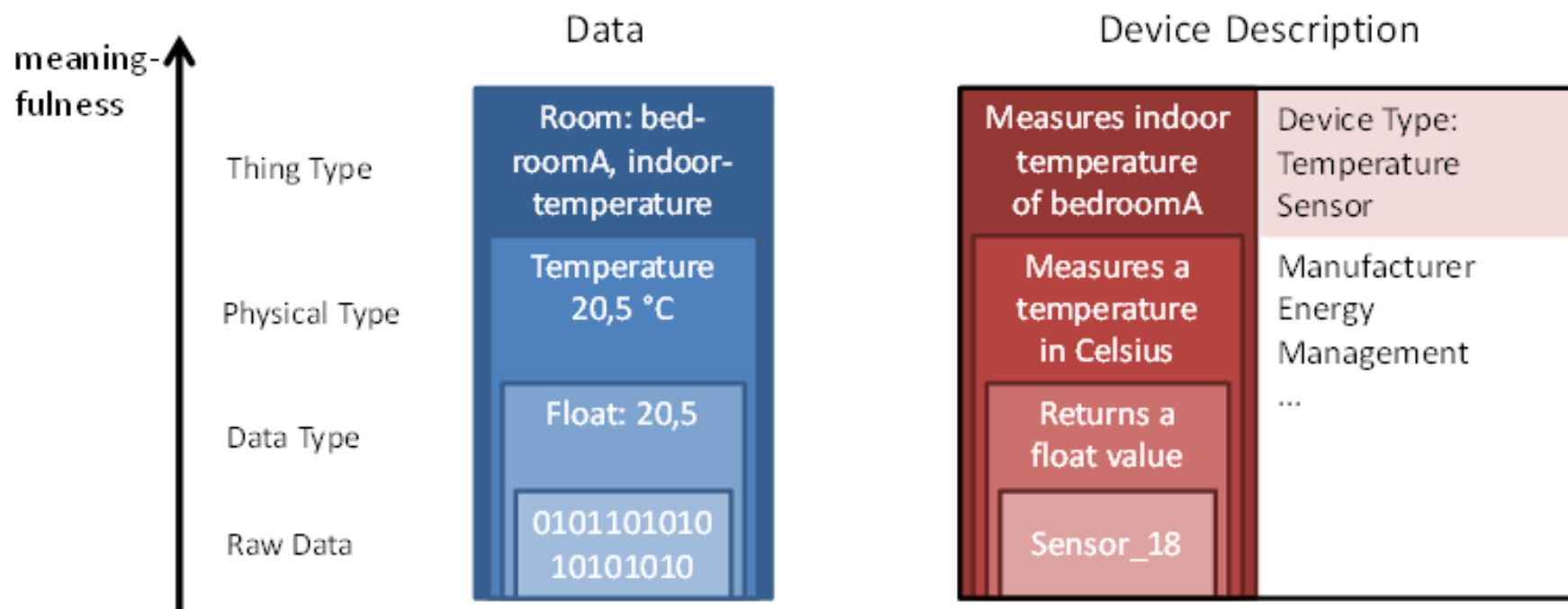
WG5

管理和语义

OneM2M中的语义技术

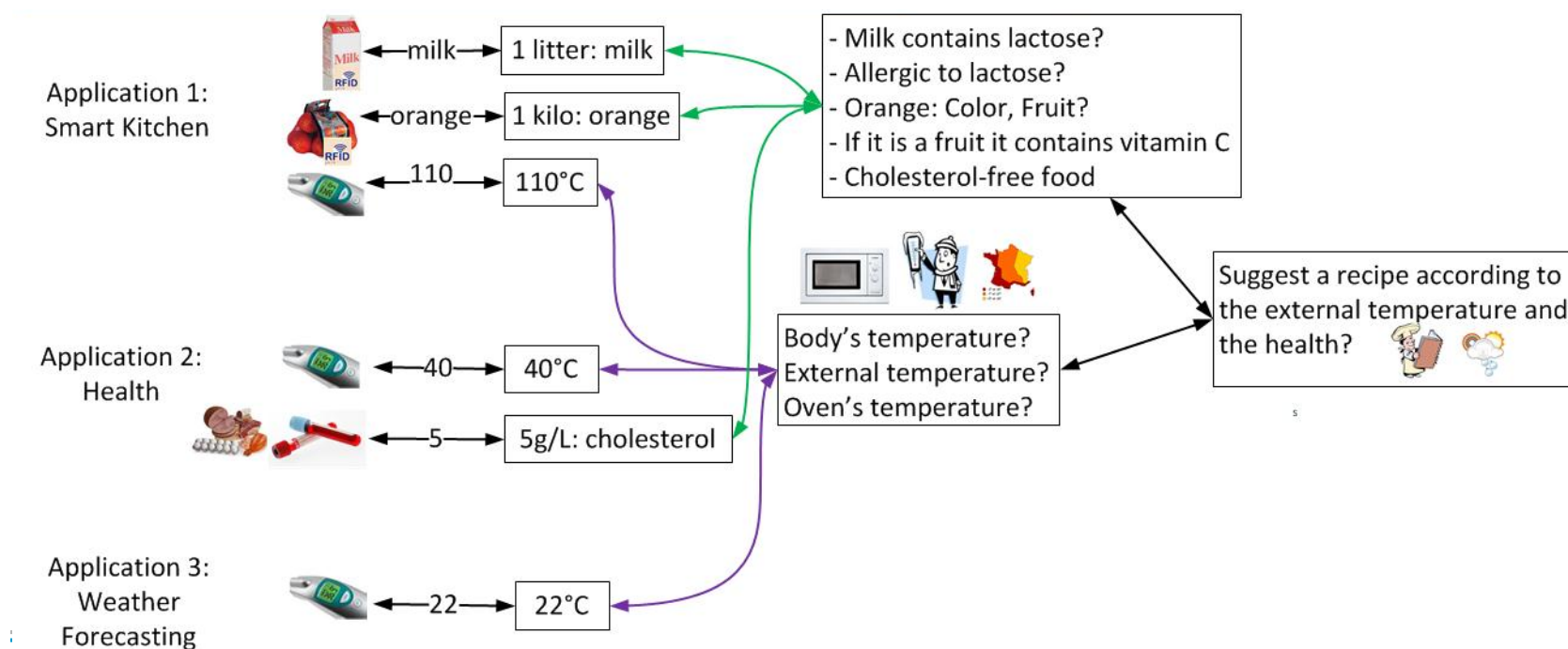


增强表达：物联网设备的抽象与语义封装

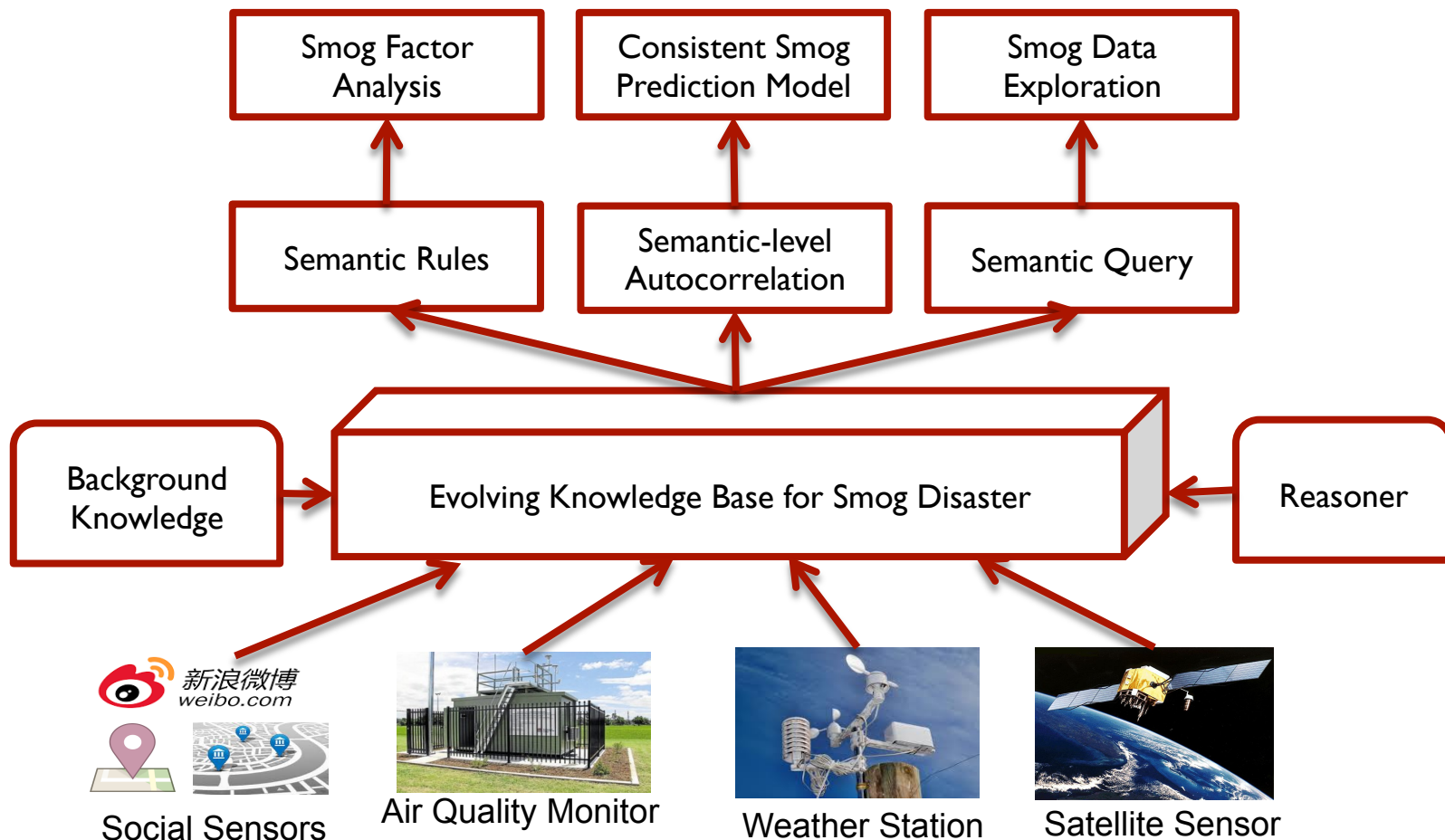


应用举例

- ▶ 可穿戴健康设备（智能手环、手表等）+ 传统健康类web数据（Linked Life data等）



应用举例：增强表达后的数据分析

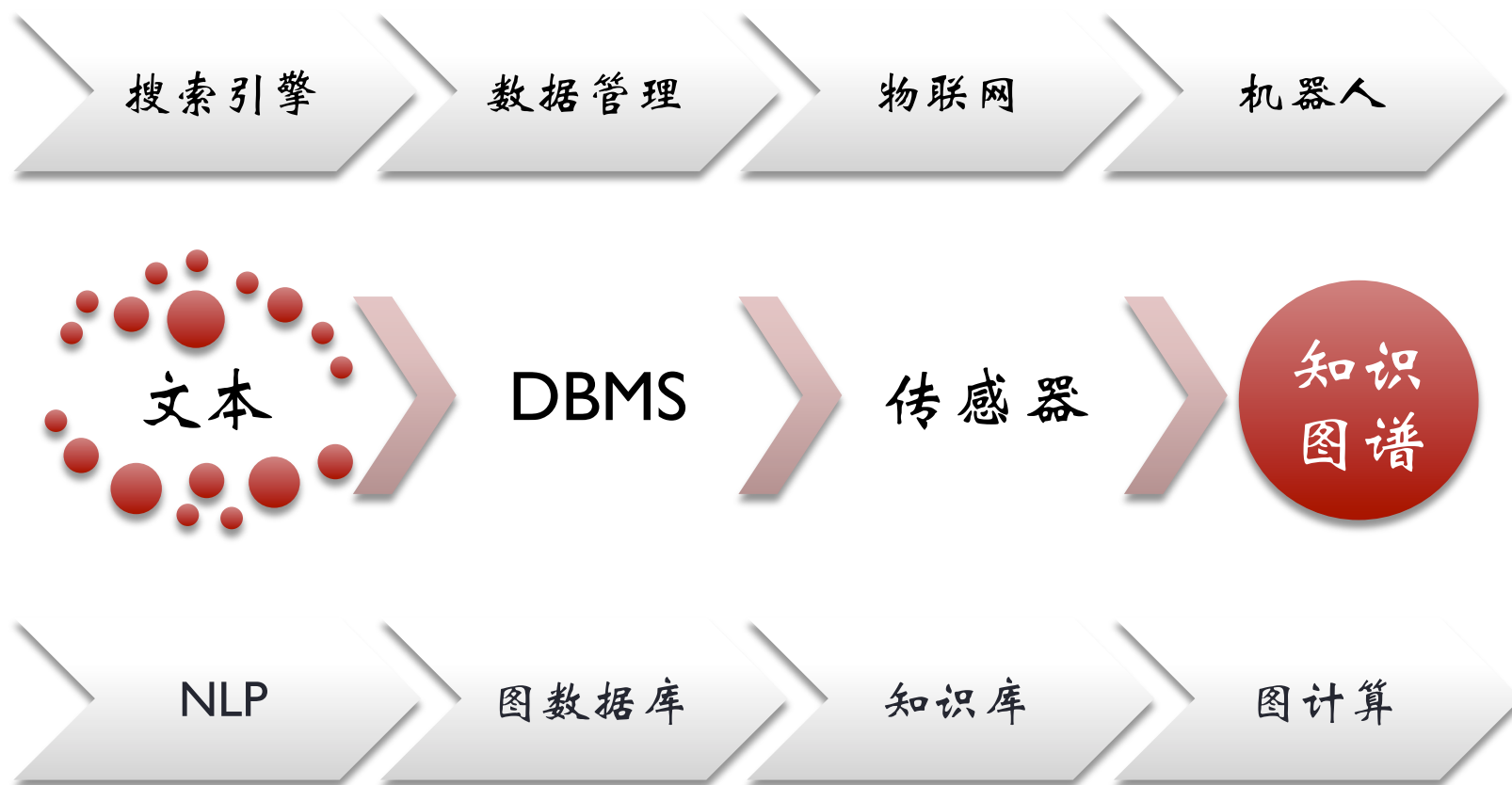




展望

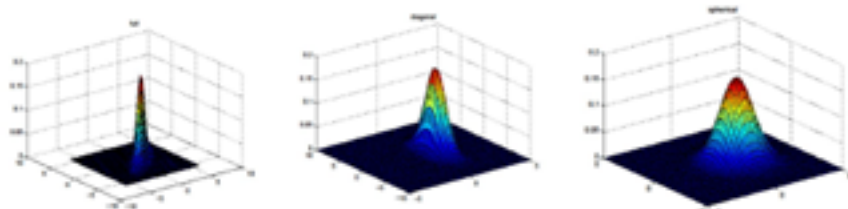


增强图表达与知识图谱的应用趋势



统一数据基础设施

机器学习：NN/DL/LR/GBDT.....



Value
分析层的统一

Vector

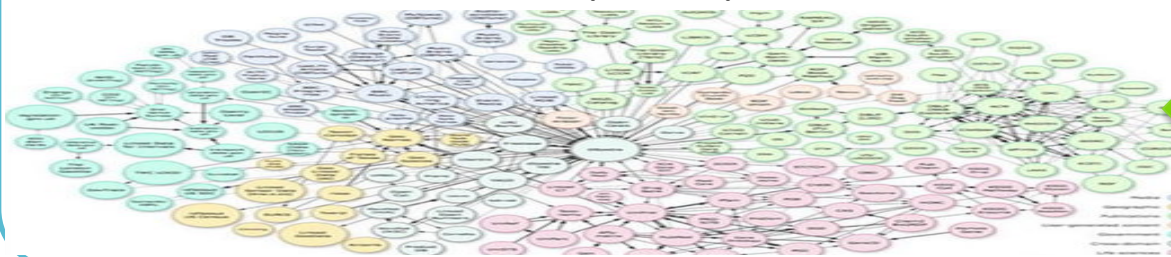
Matrix

Tensor

.....

针对特定分析需求进行特征抽取和数据格式转化

知识图谱：RDF / OWL / SPARQL.....



Variety
逻辑层的统一

HDFS / Hbase / MapReduce / Spark.....



Volume
物理层的统一

总结

图是自然的**数据表达**方式，知识图谱的本质是用**增强的图表达**来建立和管理**复杂数据关联**。

提升**数据表达和关联水平**的价值在于可为上层大数据分析和应用的提供更好的**数据基础**和更高质量的**数据原材料**。

图数据库在**性能、弹性关系管理**等方面都有明显优势，但知识图谱并不能简单等同于图模型。增强表达的最终目标是支持**人工智能**。

知识图谱也会进一步与**图计算、神经网络**等技术深度融合，成为大数据技术的一个重要组成部分。

谢谢！
请您批评指正！



语义网



该微信群二维码将在2015年11月16日失效

