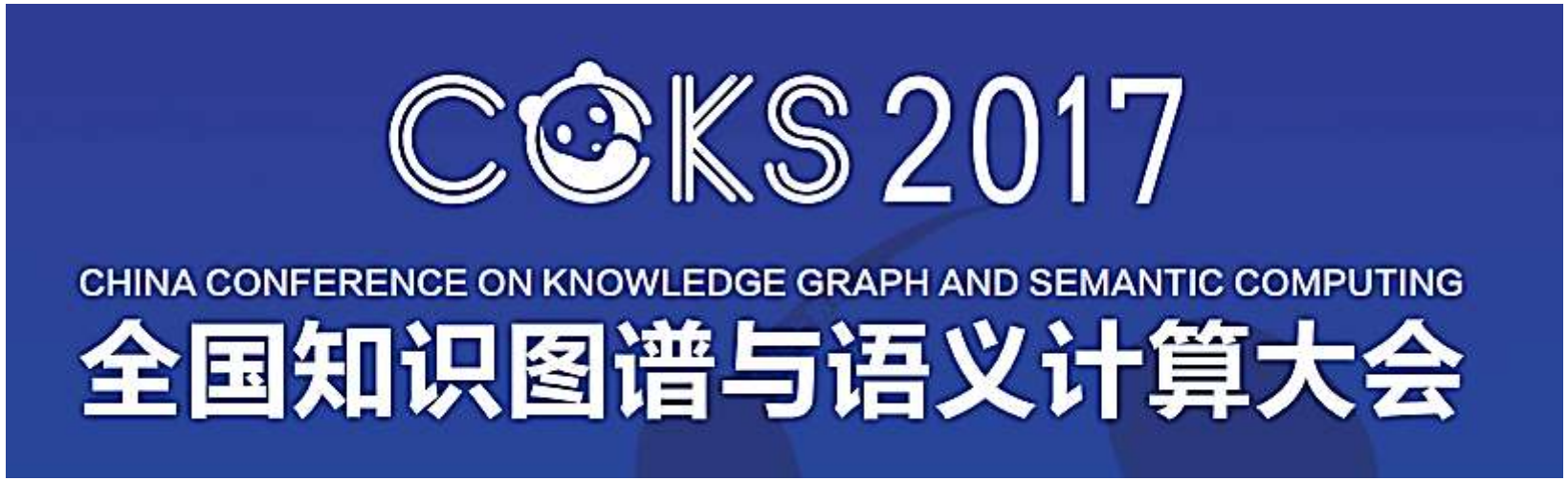


CCKS2017 Summary

En Ouyang

Outline

- Introduction
- What catch my eyes
- Task
- Thoughts

Introduction

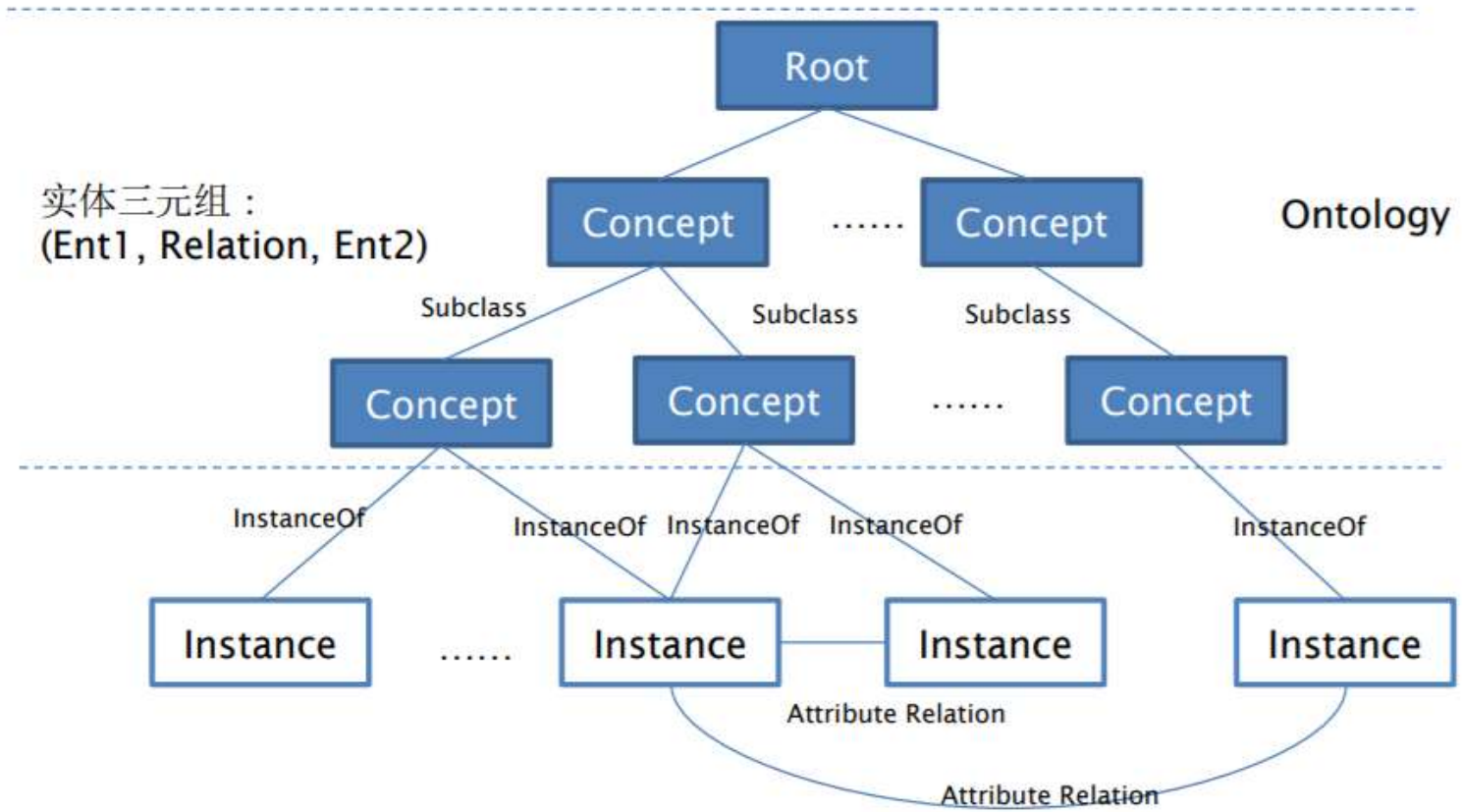
- 2017-08-26 -> 2017-08-27　　知识图谱学术讲习班, 西华大学
  - 导论、构建、虚拟化
  - 知识获取、实践
- 2017-08-28 -> 2017-08-29　　CCKS2017主会, 金牛宾馆
  - 特邀报告
  - 论文、竞赛简介
  - 论文、评测具体报告
  - 工业界论坛

什么是知识图谱

- The Knowledge Graph is a system that understands facts about people, places and things and how these entities are all connected.

- 知识图谱本质上是一种语义网络。其结点代表实体（entity）或者概念（concept），边代表实体/概念之间的各种语义关系

知识图谱包含哪些内容


Root
实体三元组：
(Ent1, Relation, Ent2)
Concept
.....
Concept
Ontology
Subclass
Subclass
Subclass
Concept
Concept
.....
Concept
InstanceOf
InstanceOf
InstanceOf
InstanceOf
InstanceOf
Instance
.....
Instance
Instance
Instance
Attribute Relation
Attribute Relation

知识图谱的生命周期

- **知识建模** Ontology
  - 建模领域知识结构
- **知识获取** 抽取，文本挖掘
  - 获取领域内的事实知识
- **知识集成** 筛选、修正
  - 估计知识的可信度，将碎片知识组装成知识网络
- **知识存储** 存储、查询、推理
  - 提供高性能知识服务

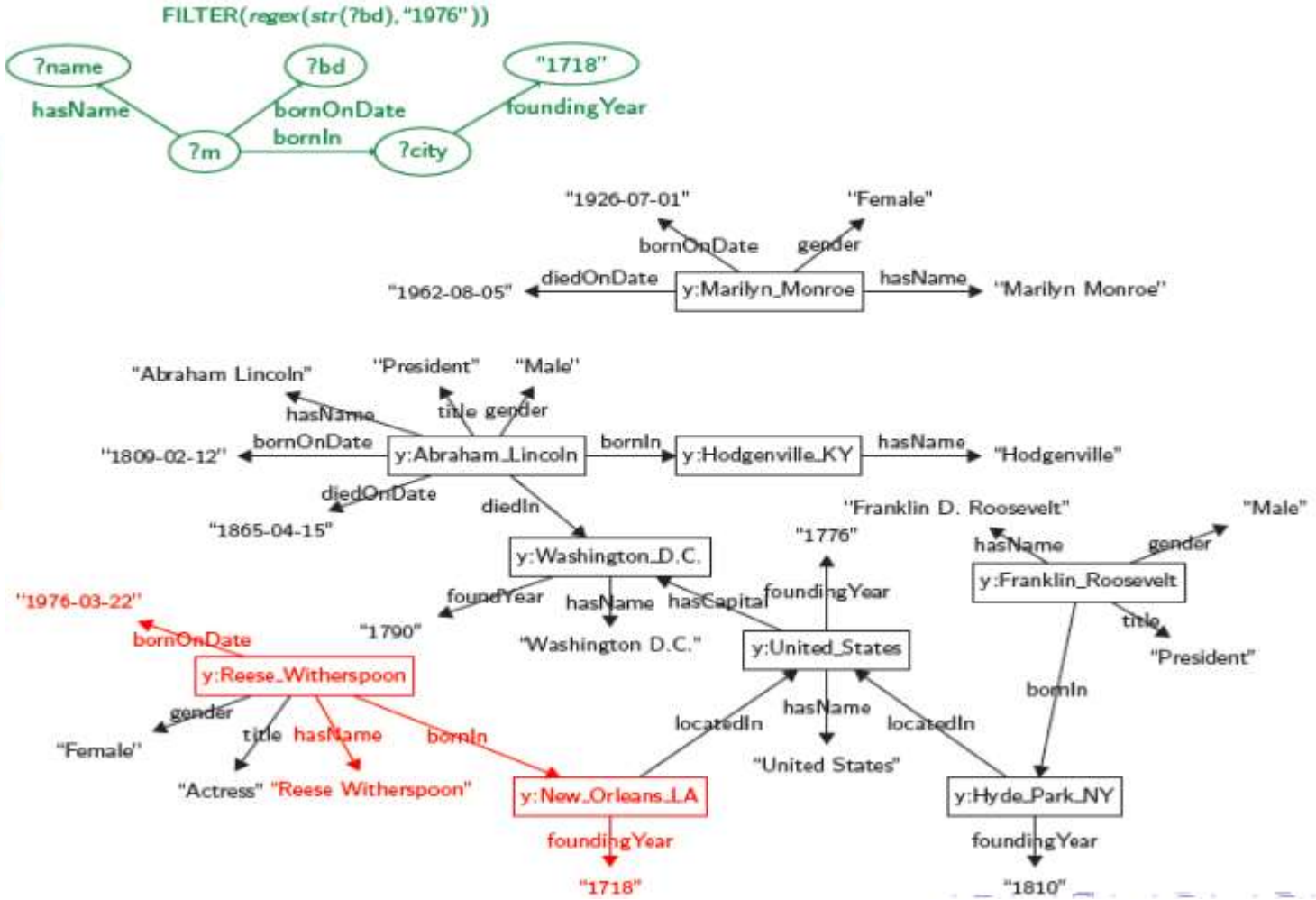
问答
精准搜索
关系搜索
分类浏览
推荐
知识推理

知识图谱存储查询

- RDF & SPARQL

主语	谓词	宾语
Abraham_Lincoln	hasName	"Abraham Lincoln"
Abraham_Lincoln	BornOnDate	"1809-02-12"
Abraham_Lincoln	DiedOnDate	"1865-04-15""
Abraham_Lincoln	DiedIn	Washington_DC
Abraham_Lincoln	bornIn	Hodgenville KY
Reese_Witherspoon	bornOnDate	"1976-03-22"
Reese_Witherspoon	bornIn	New_Orleans_LA
New_Orleans_LA	foundingYear	"1718"

- gStore 图存储 图检索

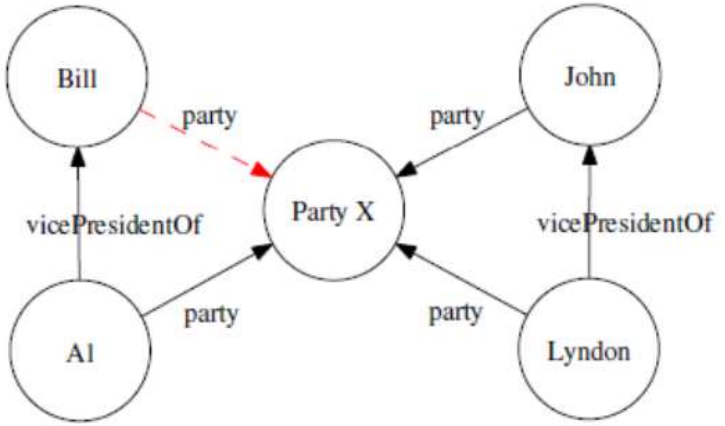


知识图谱表示学习方法

- 张量分解

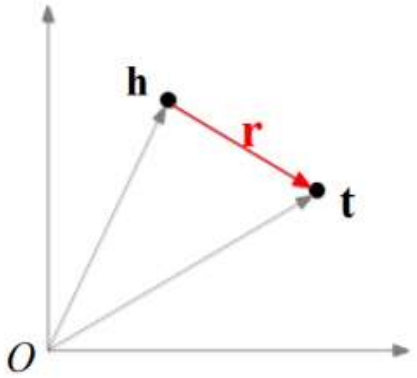
- 基于翻译的模型
  - TransE, TransH, TransR, TransH

- 神经网络模型



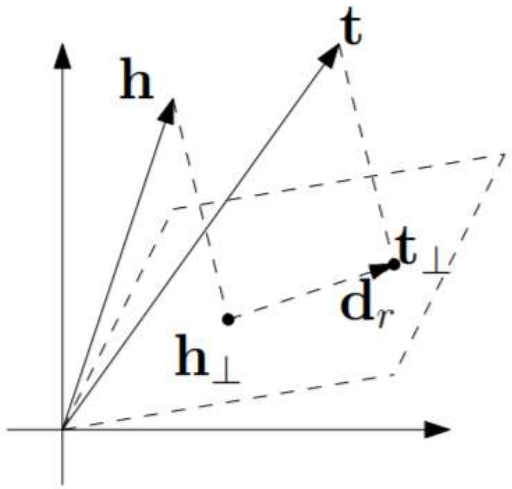
$$\mathbf{a}_{\text{Al}}^{T}\mathbf{R}_{\text{party}}\mathbf{a}_{\text{partyX}} \approx \mathbf{a}_{\text{Lyndon}}^{T}\mathbf{R}_{\text{party}}\mathbf{a}_{\text{partyX}} \Rightarrow \mathbf{a}_{\text{Al}}^{T} \approx \mathbf{a}_{\text{Lyndon}}^{T}$$

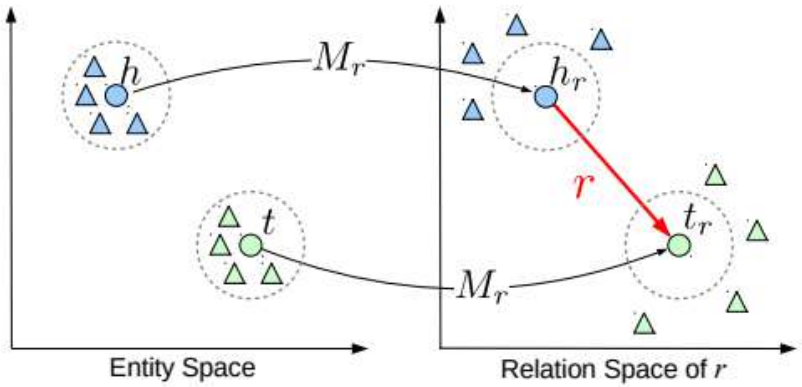
$$\mathbf{a}_{\text{Al}}^{T}\mathbf{R}_{\text{vicePresidentOf}}\mathbf{a}_{\text{Bill}} \approx \mathbf{a}_{\text{Lyndon}}^{T}\mathbf{R}_{\text{vicePresidentOf}}\mathbf{a}_{\text{John}} \Rightarrow \mathbf{a}_{\text{Bill}} \approx \mathbf{a}_{\text{John}}$$

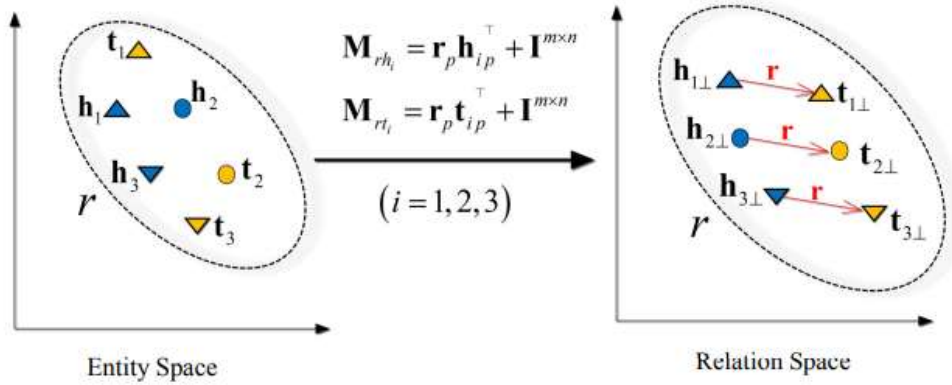


$$\mathbf{h}+\mathbf{r}=\mathbf{t}$$

基于翻译的模型







TransH

$$\mathbf{h}_\perp = \mathbf{h} - \mathbf{w}_r^\top \mathbf{h}\mathbf{w}_r \qquad \mathbf{t}_\perp = \mathbf{t} - \mathbf{w}_r^\top \mathbf{t}\mathbf{w}_r$$

$$f_r(\mathbf{h},\mathbf{t}) = \|(\mathbf{h} - \mathbf{w}_r^\top \mathbf{h}\mathbf{w}_r) + \mathbf{d}_r - (\mathbf{t} - \mathbf{w}_r^\top \mathbf{t}\mathbf{w}_r)\|_2^2$$

TransR

$$\mathbf{h}_r = \mathbf{h}\mathbf{M}_r \quad \mathbf{t}_r = \mathbf{t}\mathbf{M}_r$$

$$f_r(h,t) = \|\mathbf{h}_r + \mathbf{r} - \mathbf{t}_r\|_2^2$$

TransD

$$\mathbf{M}_{rh} = \mathbf{r}_p \mathbf{h}_p^\top + \mathbf{I}^{m\times n}$$
$$\mathbf{M}_{rt} = \mathbf{r}_p \mathbf{t}_p^\top + \mathbf{I}^{m\times n}$$

$$\mathbf{h}_\perp = \mathbf{M}_{rh}\mathbf{h}, \quad \mathbf{t}_\perp = \mathbf{M}_{rt}\mathbf{t}$$

$$f_r(\mathbf{h},\mathbf{t}) = -\|\mathbf{h}_\perp + \mathbf{r} - \mathbf{t}_\perp\|_2^2$$

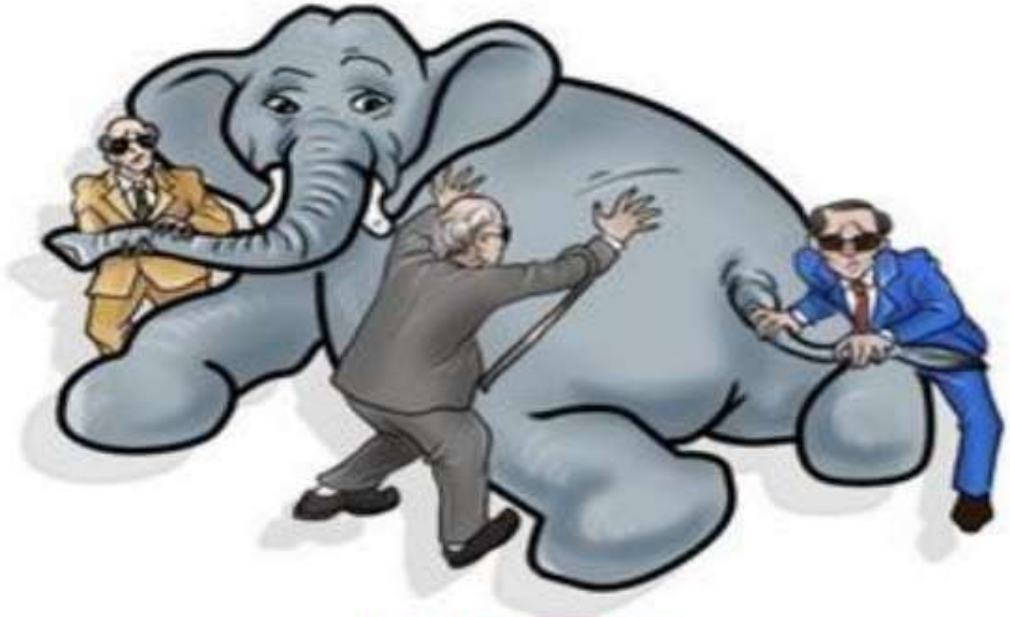
Wang, et al. Knowledge Graph Embedding by Translating on Hyperplanes. In Proceedings of AAAI 2014
Lin, et al. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In Proceedings of AAAI 2015
Ji, et al. Knowledge graph embedding via dynamic mapping matrix. In Proceedings of ACL 2015

交叉研究

数据库

RDF数据库系统

数据集成、知识融合



自然语言处理

关系抽取

语义解析

(Semantic Parsing)

知识问答系统中，让机器理解问题

机器学习

知识图谱数据的知识表示

(Graph Embedding)

TransE

知识工程

知识库构建

基于规则的推理

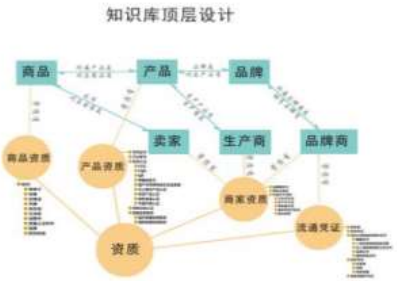
企业知识图谱

企业风险评估
企业社交图谱
企业最终控制人
企业间路径发现
企业融资历程

金融证券

生物医疗

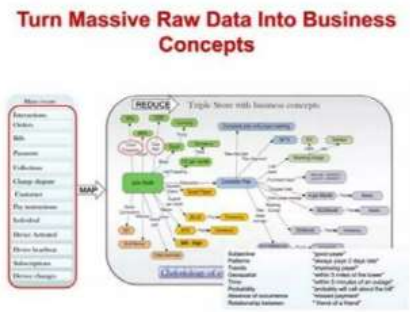
图书情报



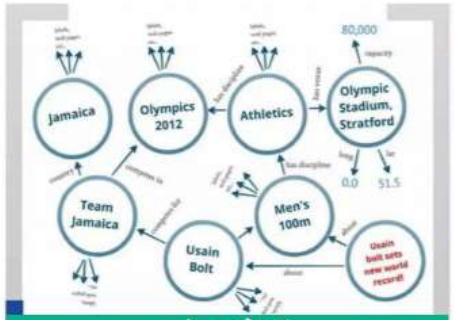
电商

农业

政府



电信



出版

挑战

- 多源异构数据整合（文献、网站、已有知识、新闻等）
  - 统一格式，逐个包装
- 数据模式动态变迁困难（新需求、新功能）
  - 可自由扩展的数据结构
- 时态信息存储
  - 用有限的数据冗余实现数据时态信息的应用
- 行业知识图谱与通用知识图谱结合

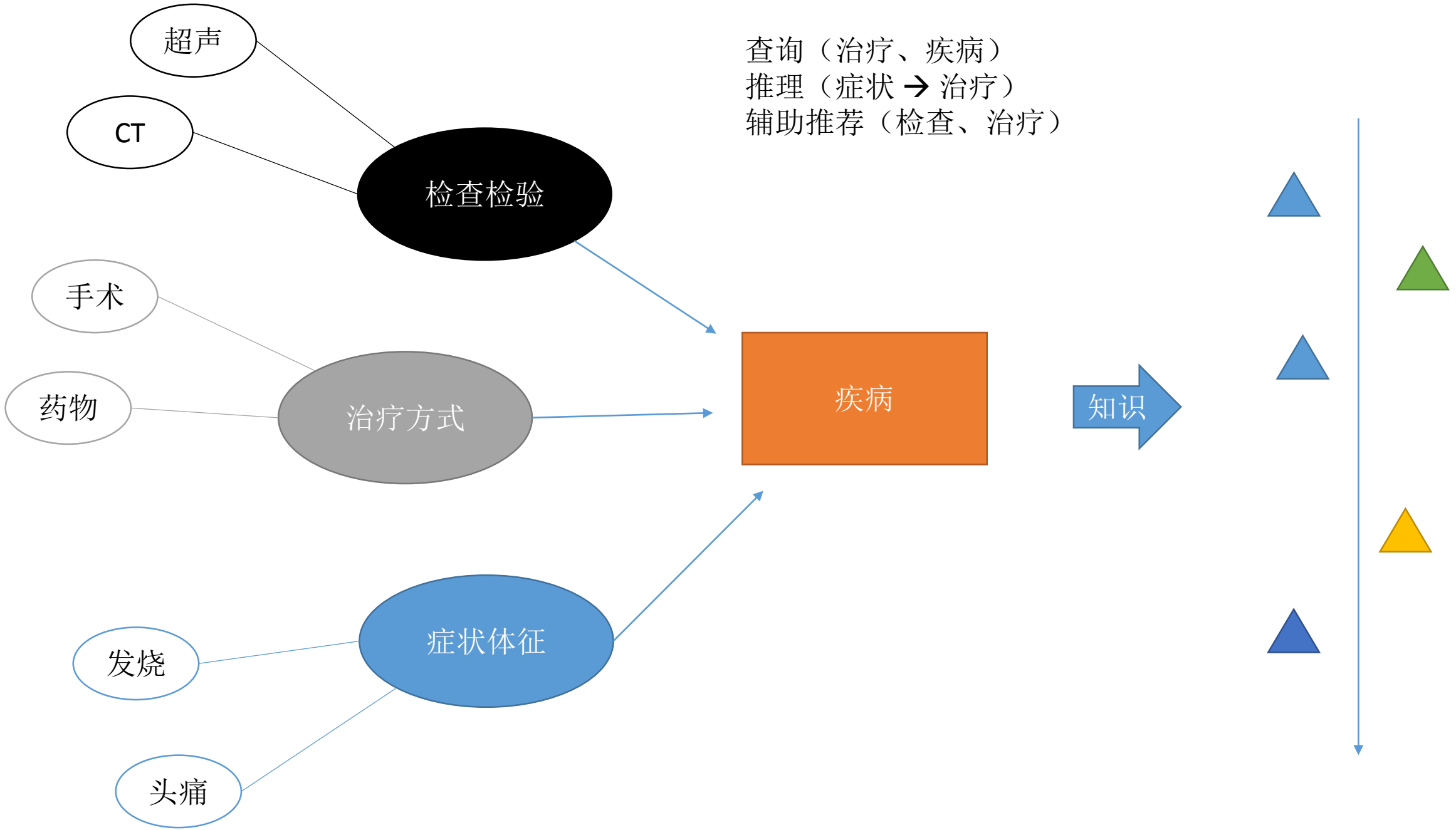
阿里商品知识图谱

技术积累
人才引进
战略布局

- 标准 → 逐步 → 应用
- 在构建scheme的时候，先不要过多考虑应用层面，尽量把scheme考虑细致、全面
- 在构建知识图谱时采用逐步构建的方式，不要把网撒太开，也提及了很多相关技术（多源知识图谱融合、NERL、文本关系抽取）
- 前面两步有了，在应用上就会水到渠成（搜索、前端导购、平台治理、智能问答、品牌商运营等）

临床医学知识图谱到底怎么用？

- 临床知识图谱该纳入哪些知识？
  - 哪些算知识？
  - 以什么为中心？
  - 时序和事件信息怎么纳入？
- 临床知识图谱的应用场景？
  - 个体知识图谱 or Timeline + 多个专门知识图谱？
  - 推荐、辅助
- 临床知识图谱的必要性？
  - 知识存储


超声
CT
检查检验
手术
药物
治疗方式
症状体征
发烧
头痛
疾病
知识
查询（治疗、疾病）
推理（症状 → 治疗）
辅助推荐（检查、治疗）

Task 2

1. Character representation
2. Word segmentation

Bi-RNN-CRF

1. Vote
2. Consistency check
3. Re-train
4. Results integration

Hanlp word segmentation
GloVe embedding training
Reseg
Character embedding + word embedding

Multiple model (rule-based, crf, rnn)
Entity classification modification
Using unlabeled datasets

For journal paper

Focus on **Character Representation**

- Character representation is a basic but crucial step in NLP tasks
- We have drawn the conclusion that n-gram representation is better
- We can study different character representation methods can influence the **clinical NER** results or not, and **HOW**
- We could add some necessary post-processing steps
- Transform **Task** to **Tool**

Thoughts

- Performance 可以不好，技术一定要新

Best English Paper

- ***Attention-based Event Relevance Model for Stock Price Movement Prediction***. 中国科学院自动化研究所. Jian Liu, Yubo Chen, Kang Liu and Jun Zhao

- **做好笔记，及时总结**
- 稍微紧张就可以了

Thanks!