知识图谱

- 什么是知识图谱?
  - 有人的地方就会有江湖
  - 江湖不是打打杀杀
  - 而是人情世故
  - 图模型把所有信息都连起来了

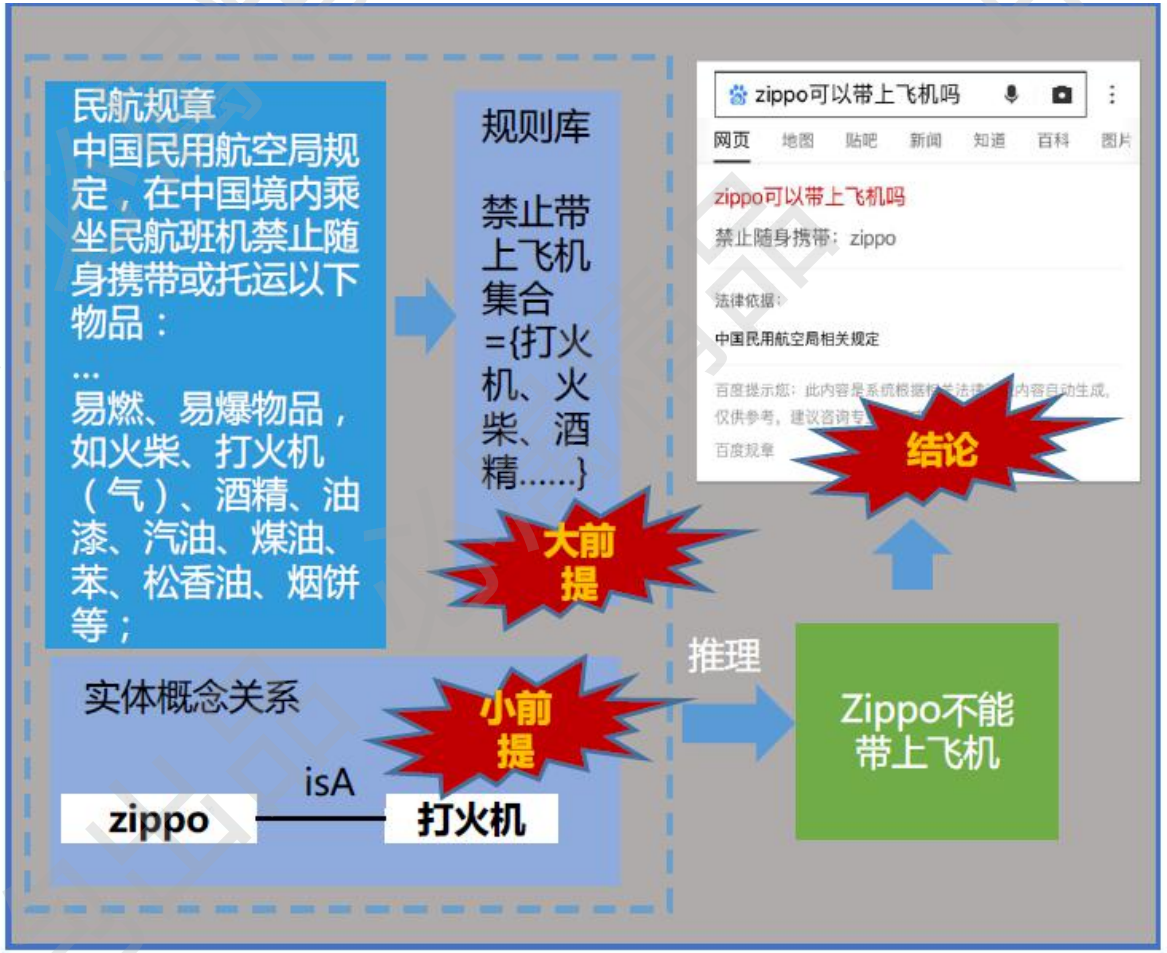
知识图谱

什么是知识图谱?

知识图谱这个概念有点抽象，先来看看不同业务角度，知识图谱要解决的问题

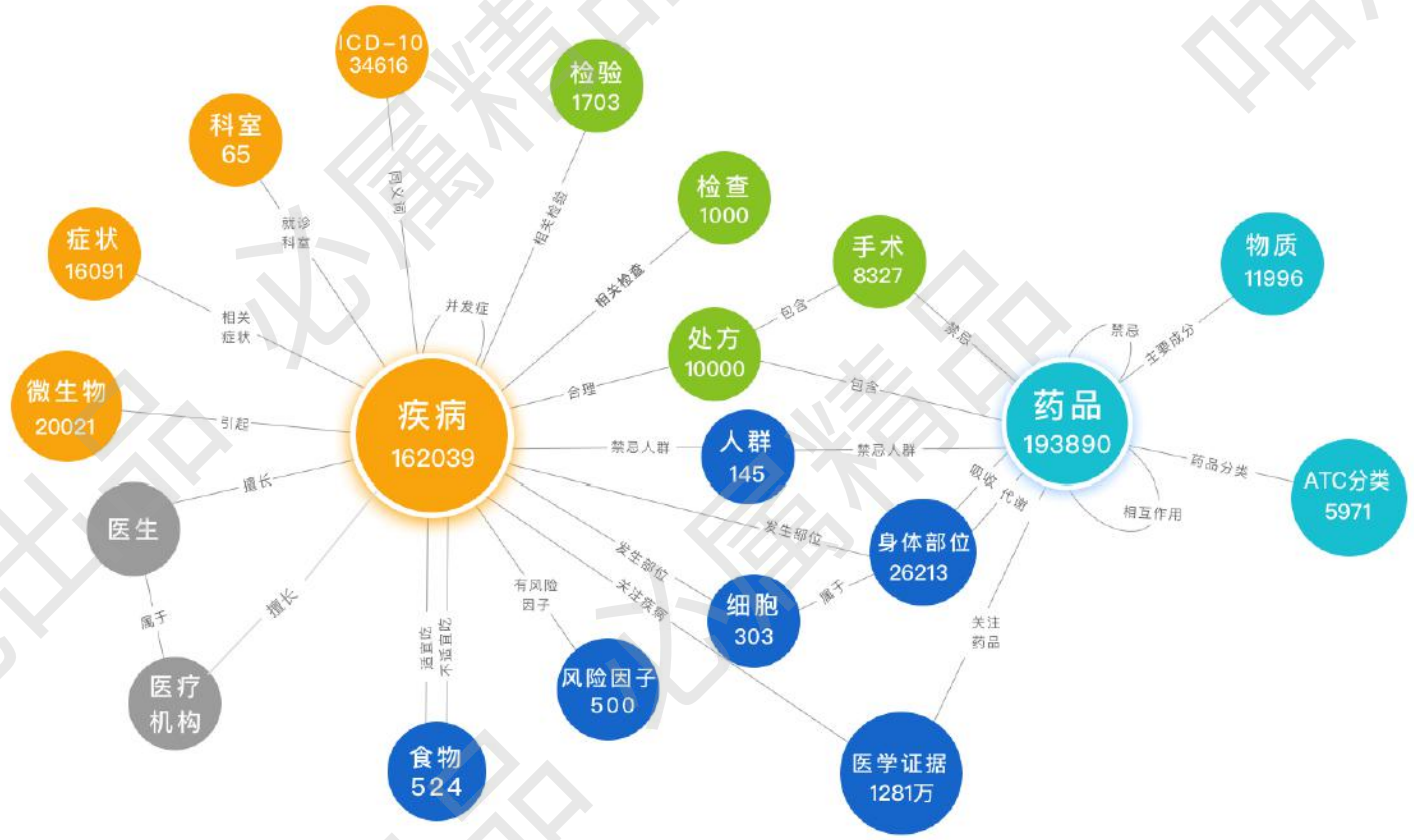
知识图谱

- 在搜索引擎中的作用：



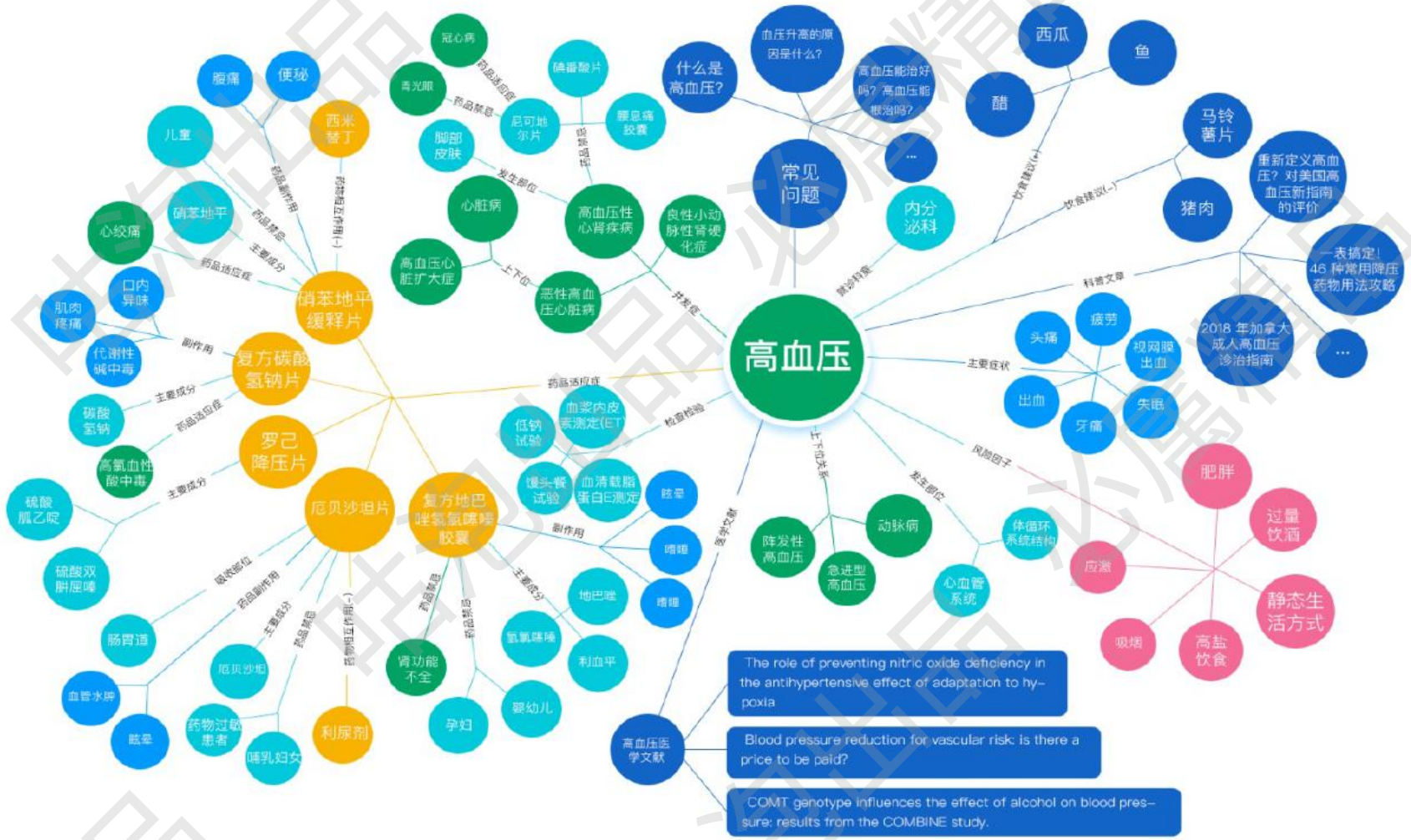
知识图谱

- 在医疗领域的应用：
  - 主要是方便查询
  - 例如智能问答助手
  - 图模型能帮我们快速检索
  - 只要有数据就能搭建图模型



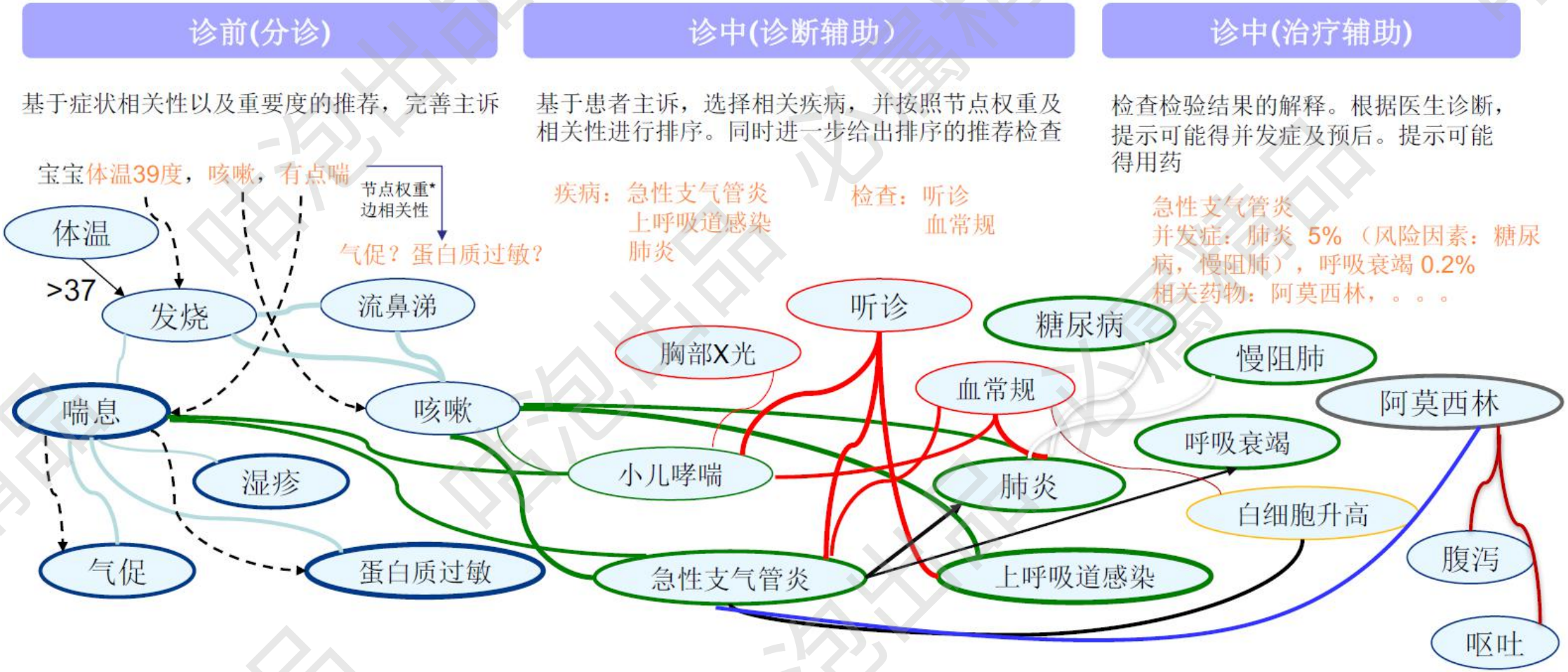
知识图谱

- 在医疗领域的应用：



知识图谱

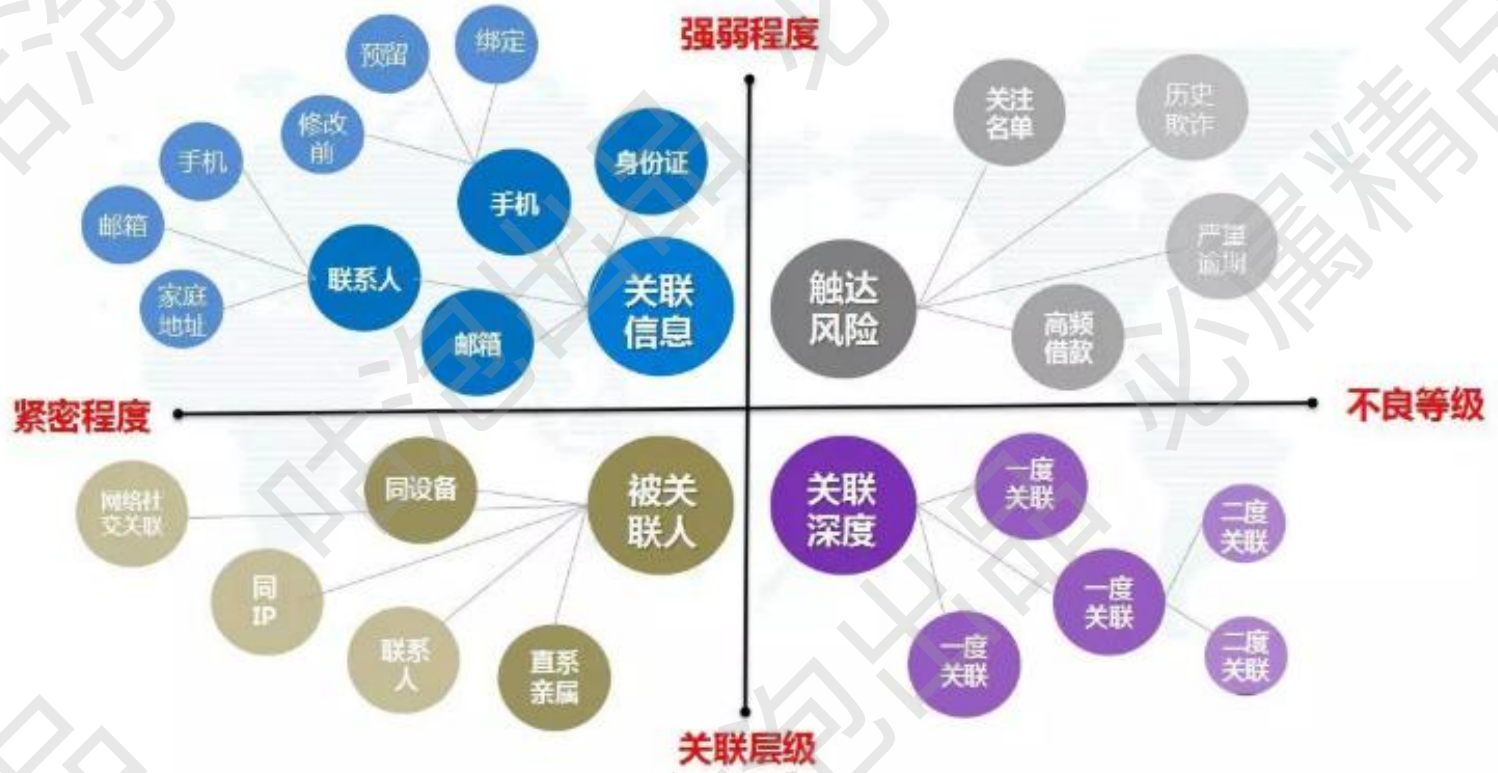
- 在医疗中进行辅助决策：



知识图谱

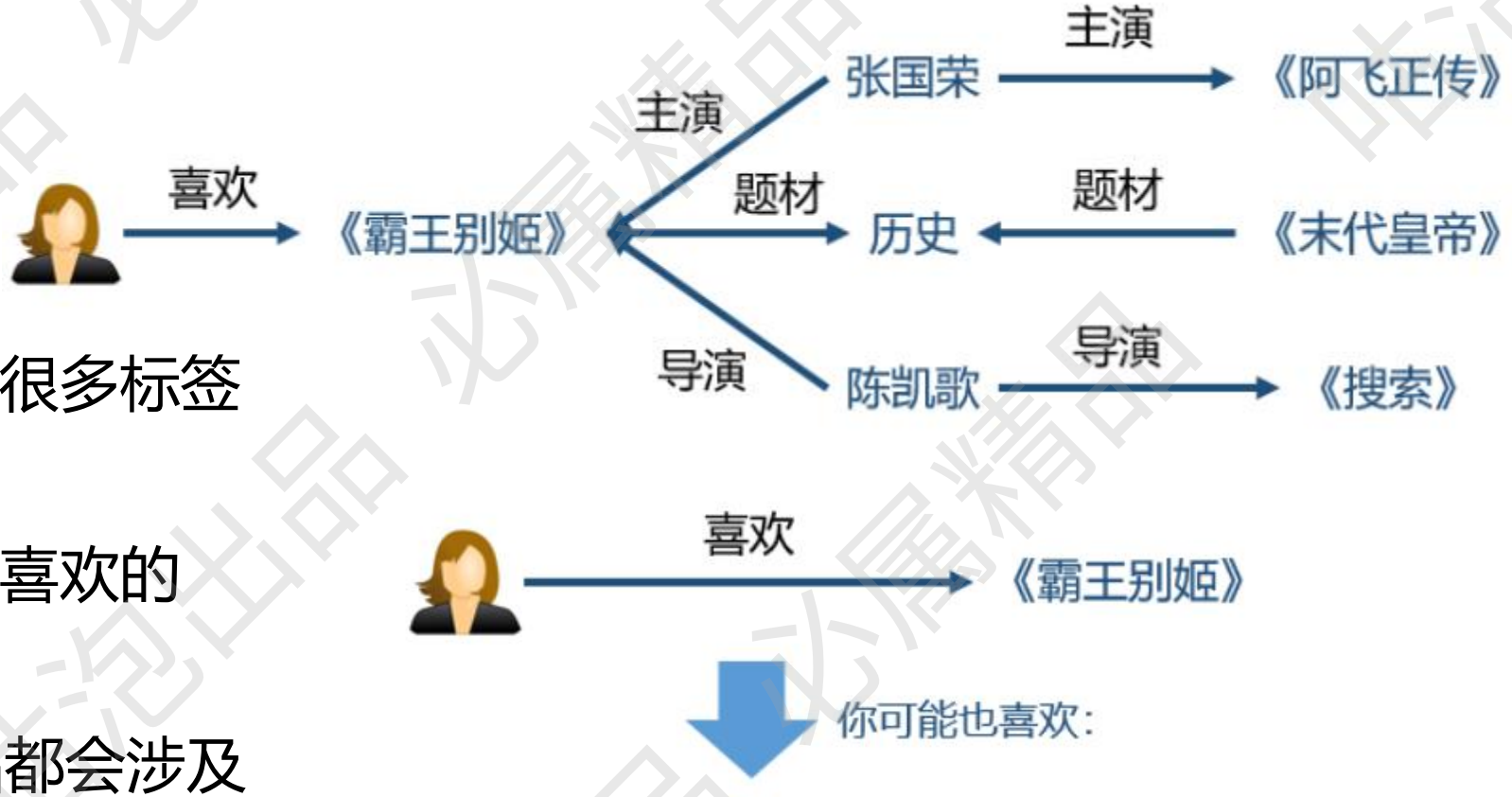
- 在金融领域的应用：

  - 反欺诈，风控模型是知识图谱在金融领域的经典应用



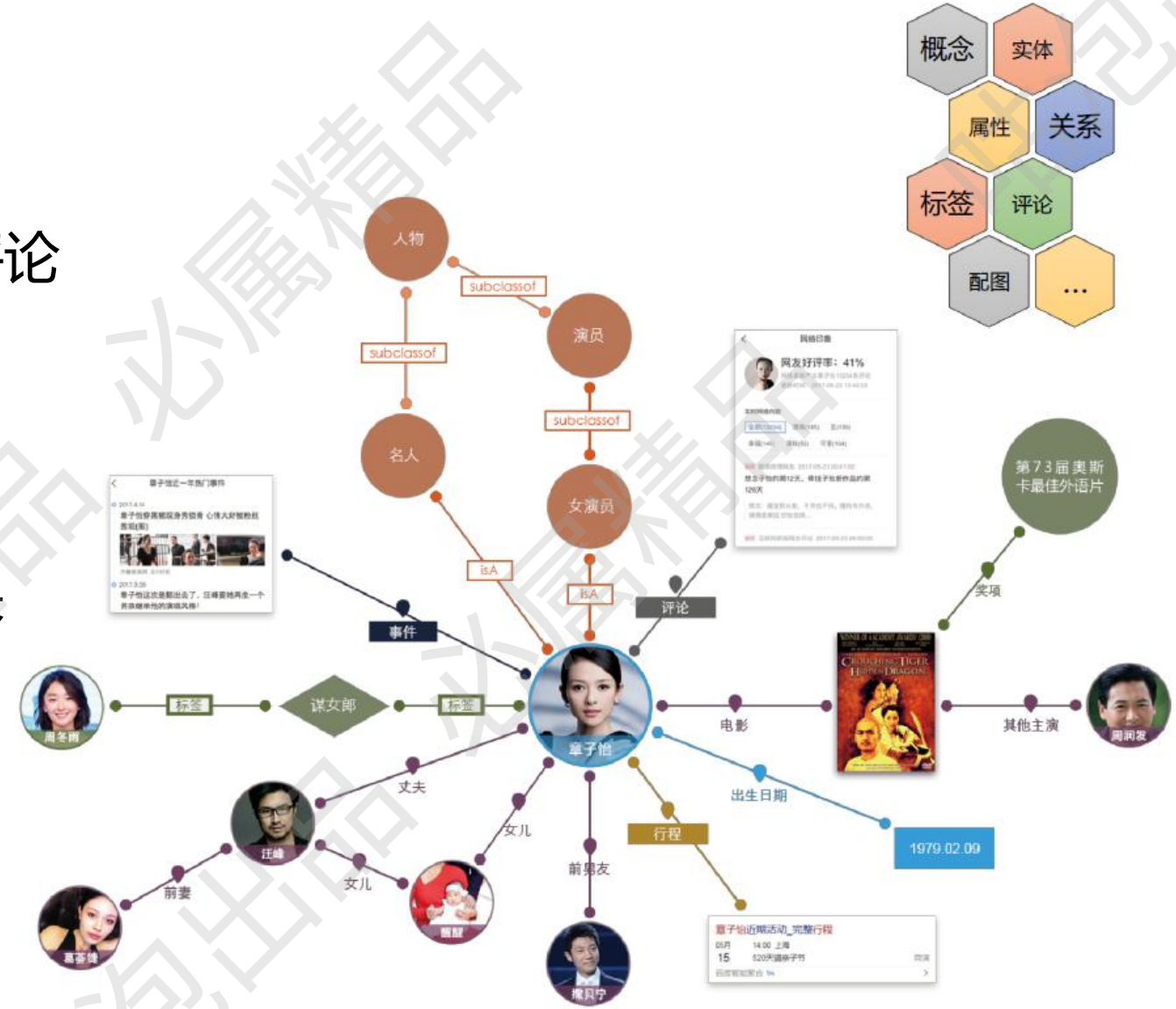
知识图谱

- 推荐系统：
  - 想想天天刷的抖音
  - 其实你已经被打上了很多标签
  - 跟你的兴趣来推荐你喜欢的
  - 基本所有互联网产品都会涉及



知识图谱

- 数据长什么样子？
  - 一篇文章，一份演员表，一条评论
  - 人情世故通常都是文本数据
  - 在海量数据中把这些关系抽出来
  - 关系与实体之间组成了联系



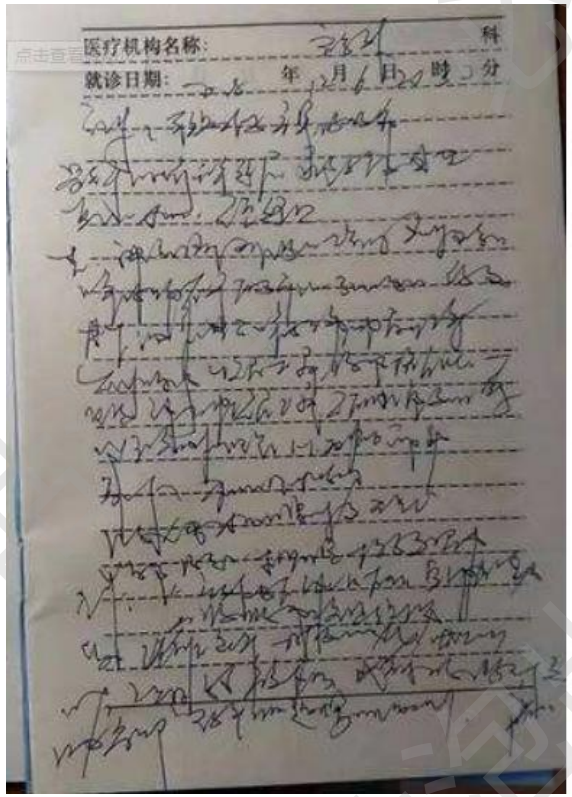
知识图谱

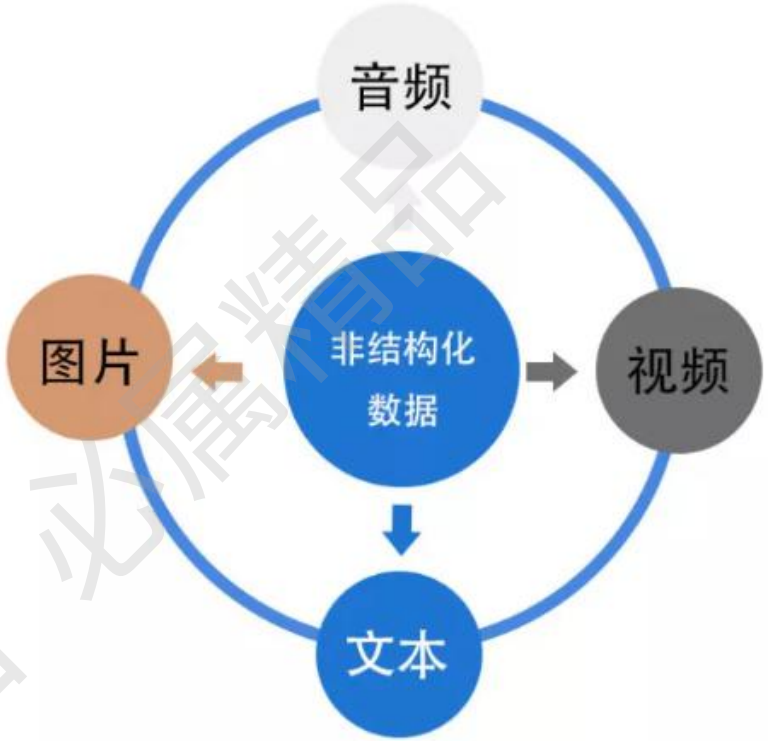
- 数据从哪里来：
  - 以医疗领域为例，需要大量的用户交互数据
  - 就诊数据：张三头疼后做了CT检查后，确诊感冒
  - 在数据中创建点（标签）和关系：
  - [(疾病：感冒)->(症状：头疼)]，关系为疾病所对应症状，还可以再提取疾病对应的检查等

知识图谱

- 一般情况下拿到的都是非结构化的数据：

- 一般情况下拿到的都是非结构化的数据





- 实际可能的数据：

知识图谱

- 数据从哪里来：
  - 是手动提取关系信息吗？
  - 数据很多，关系却难
  - 涉及大量NLP技术
  - 关系做的准确才可靠

网络小说家天蚕土豆在网络小说界大名鼎鼎，其所写的《斗破苍穹》更是人气爆棚

实体抽取

天蚕土豆 PER
斗破苍穹 NVL

语义标签抽取

天蚕土豆 网络小说界大名鼎鼎
斗破苍穹 人气爆棚

二元关系抽取

天蚕土豆 ISA 网络小说家
斗破苍穹 作者 天蚕土豆

2013年，《南方娱乐周刊》通过网友、媒体评审团、圈中人评审团的力量，最终Angelababy、刘诗诗、杨幂、倪妮被评选为新”四小花旦”

多元关系抽取

新”四小花旦”—Angelababy 刘诗诗 杨幂 倪妮

小米上市首日破发 7月23日将正式纳入恒指

7月9日，作为港交所"同股不同权"新规实施后的第一股小米集团（1810.HK）正式在香港交易所上市。小米董事长雷军亲自敲锣开市。图/视觉中国

7月9日，小米集团在港交所正式挂牌上市，与发行价每股17港元相比，开盘价下跌2.35%，为16.6港元，总市值为3714.4亿港元（约为473.3亿美元）。截至收盘，小米股价固定在16.8港元，相比发行价下跌1.18%，总市值为3759亿港元（约479亿美元）。相比雷军7月8日披露的小米估值543亿美元，小米市值IPO首日蒸发64亿美元。

事件抽取

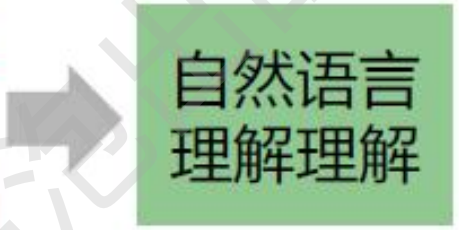
事件：IPO
- 公司：小米集团
- 交易所：香港交易所
- 发行时间：2018年7月9日
- 发行价：17港元

知识图谱

- 为什么通常将知识图谱划分到NLP领域？

- 文本数据较多，如何从文本中提取有价值信息成为关键

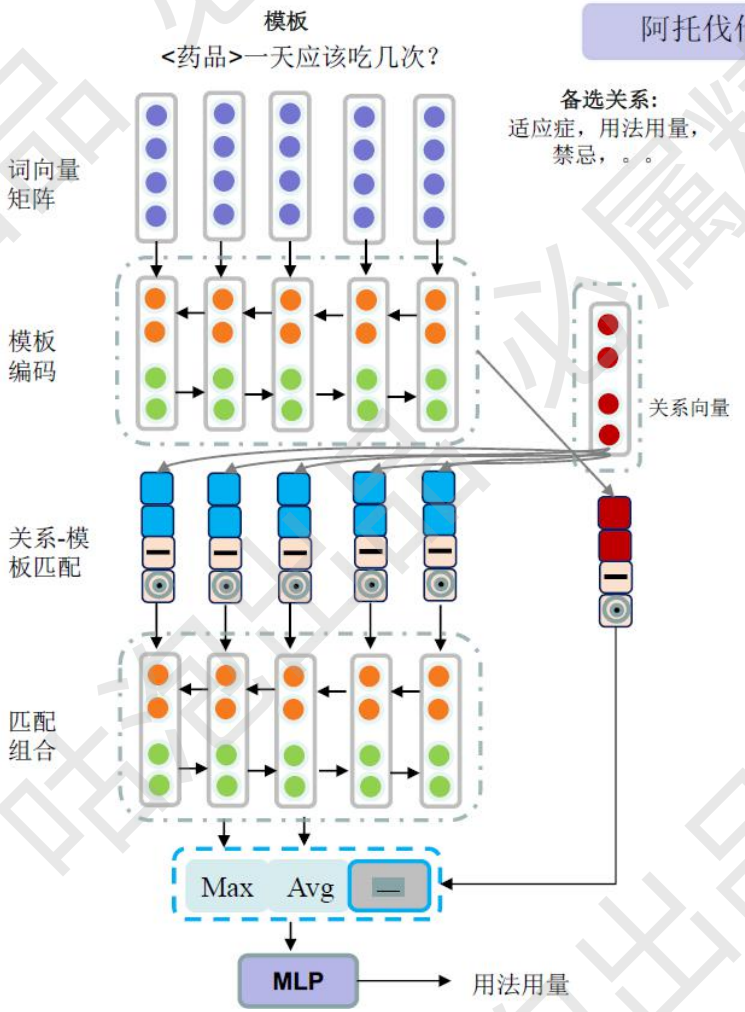
患者诉发冷，出现寒战，头晕，恶心未吐。血常规：白细胞数17.83*10^9/L，。。。。查体：左肾区叩击痛阳性,未触及肿块。。。考虑患者泌尿系感染合并SIRS。制定治疗方案：1停止静点头孢替唑，给予头孢哌酮舒巴坦钠2.03/日静点



症状	有	寒战，头晕，恶心
	无	呕吐
检验	白细胞	17.83*10^9/L
体征	有	左肾区叩击痛阳性
	无	肿块
诊断		泌尿系感染合并SIRS
治疗	无	头孢替唑
	有	头孢哌酮舒巴坦钠

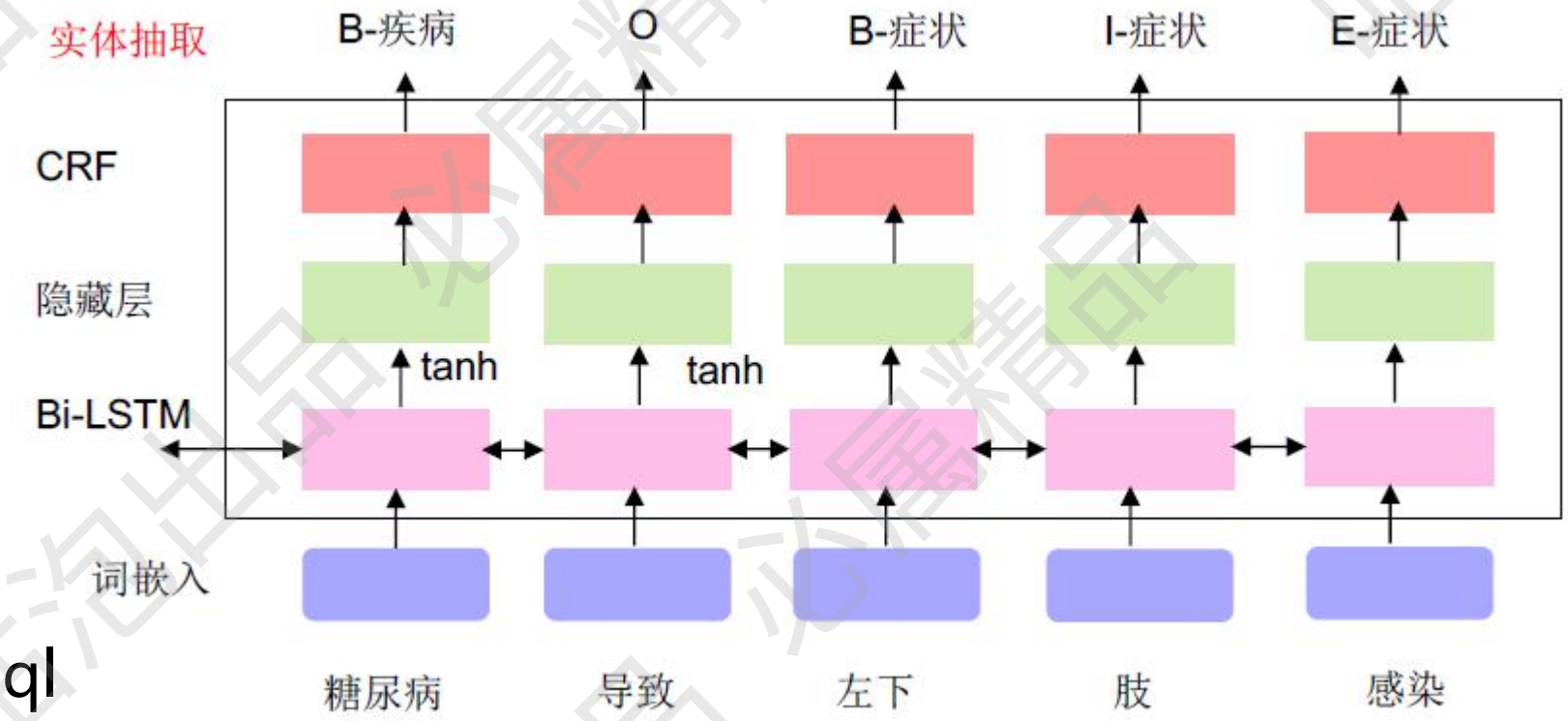
知识图谱

- 医疗自助回答系统：
  - 很多NLP任务
  - 命名实体识别
  - 找到对应关系
  - 在图中返回结果



知识图谱

- 常用技术点：
  - 命名实体识别
  - 给词打上标签
  - 有标签才好查找
  - 将标签与意图转换成sql



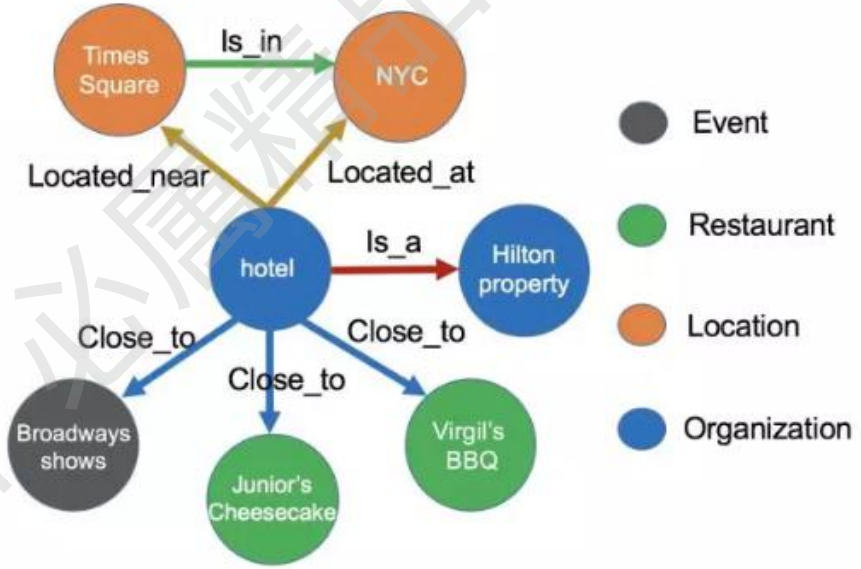
知识图谱

- 常用技术点：

  - 基于实体与关系构建知识图谱网络图（关系抽取）

This hotel is my favorite Hilton property in NYC! It is located right on 42nd street near Times Square in New York, it is close to all subways, Broadways shows and next to great restaurants like Junior's Cheesecake, Virgil's BBQ

This hotel is my favorite Hilton property in NYC! It is located right on 42nd street near Times Square in New York, it is close to all subways, Broadways shows and next to great restaurants like Junior's Cheesecake, Virgil's BBQ



知识图谱

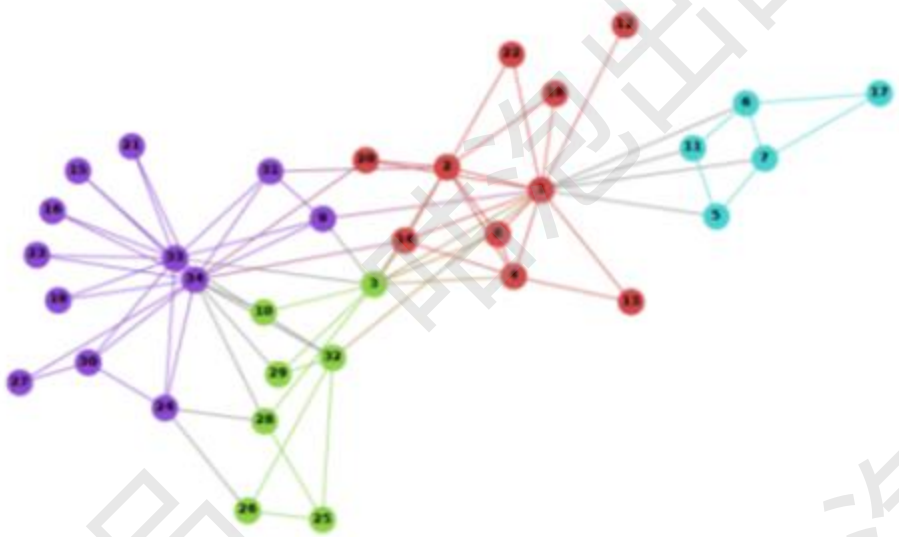
- 常用技术点：
  - 实体统一
  - 指代消解

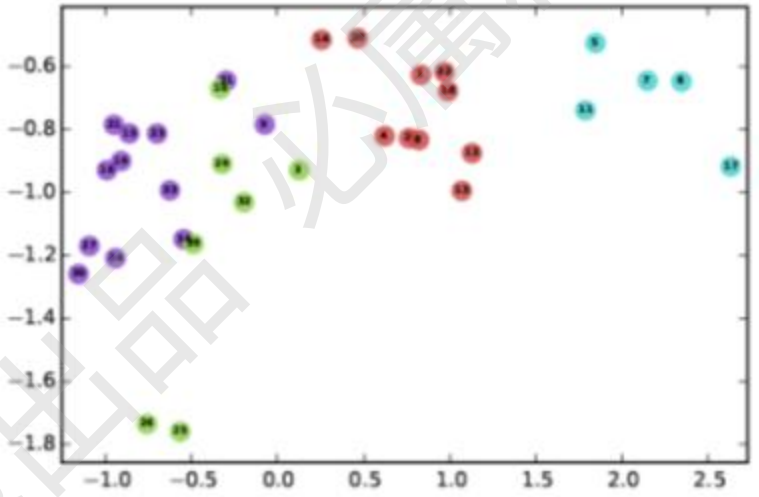
勒布朗·詹姆斯：1984年12月30日出生在美国俄亥俄州阿克伦。LBJ在2003年NBA选秀中于首轮第1顺位被克利夫兰骑士队选中。2010年，他转会至迈阿密热火队，与德怀恩·韦德、克里斯·波什组成“三巨头”阵容。

勒布朗·詹姆斯：1984年12月30日出生在美国俄亥俄州阿克伦。LBJ在2003年NBA选秀中于首轮第1顺位被克利夫兰骑士队选中。2010年，他转会至迈阿密热火队，与德怀恩·韦德、克里斯·波什组成“三巨头”阵容。

知识图谱

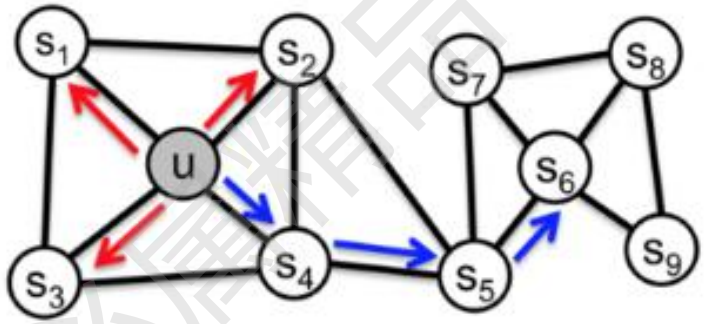
- 知识图谱只是NLP任务吗？
  - 如果可以进行特征编码，那么能计算机就可以进行训练和推理任务
  - embedding这件事是Ai最核心的内容，如何能让计算机读懂咱们的数据





知识图谱

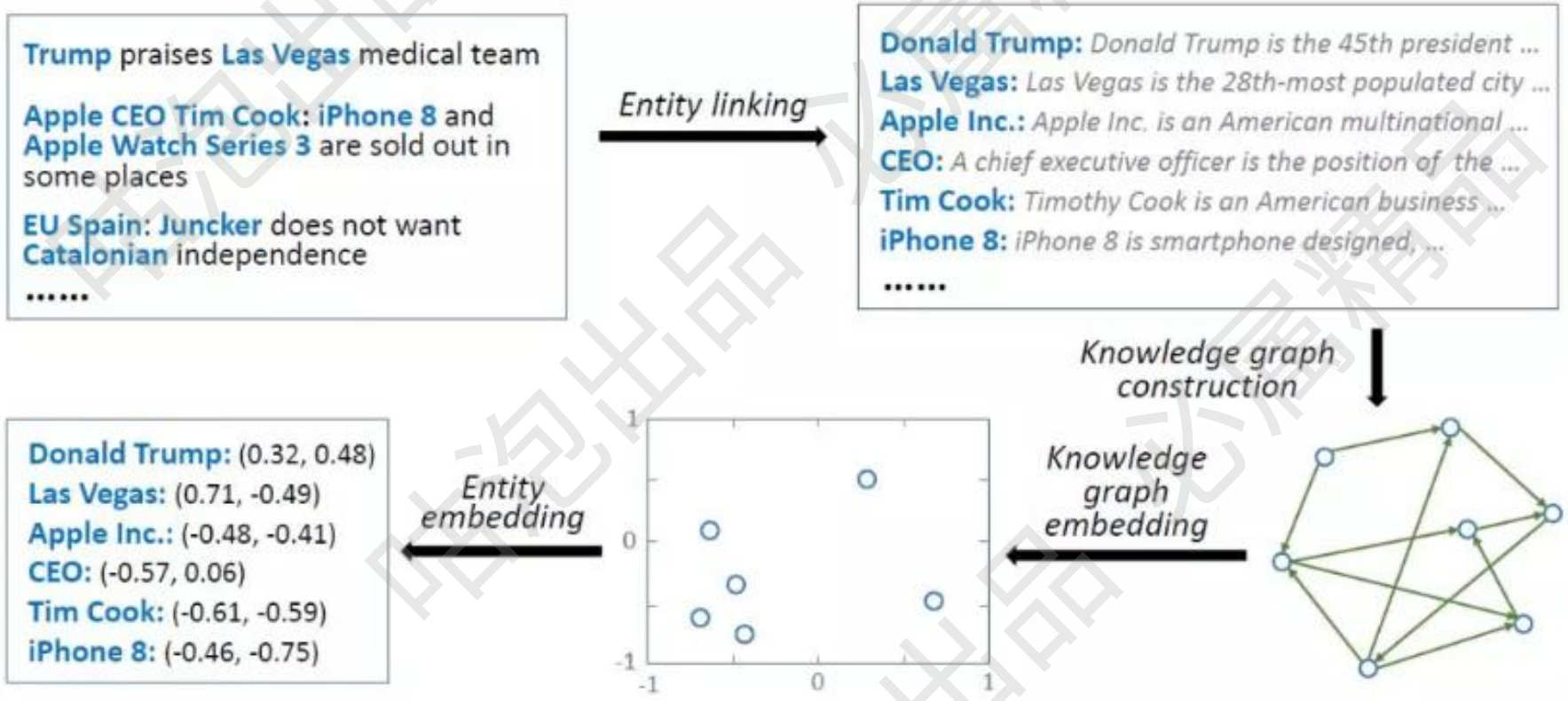
- graph embedding：



- 风控模型中对接点进行编码：
- 根据用户关系（通讯录）建立算法模型（Deep walk），获得用户向量
- 有了特征编码能做的事情就多了，预测，分析等一些ML任务都能干活了
- 难点在于如何编码（算法）才能更准确体现这个用户的情况

知识图谱

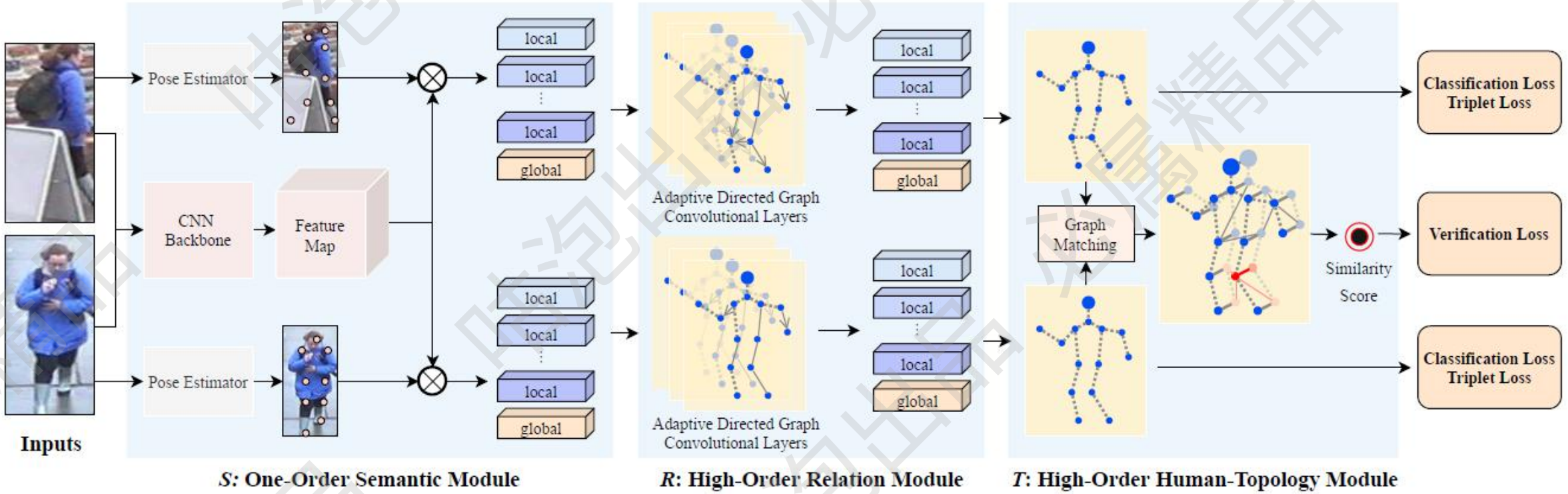
- 特征表达尤为重要



知识图谱

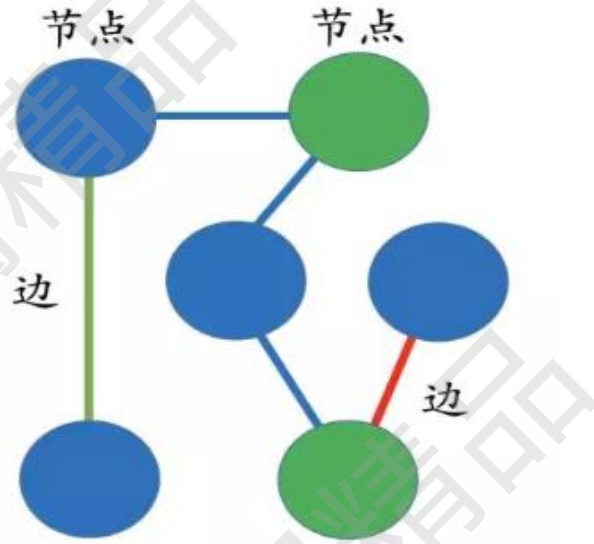
- graph embedding：

  - 图像/视频数据，例如图卷积模型



知识图谱

- 知识图谱的组成：



  - 多种类型的节点和边：
  - 通常把节点叫做实体，关系叫做边
  - 实体：人，地点，疾病名称，公司等（不同的标签）
  - 边：描述实体之间的关系（关系也是多种的）

知识图谱

- 图数据库排名：
  - neo4j遥遥领先
  - 与Python交互容易
  - 上手轻松
  - 应用方便

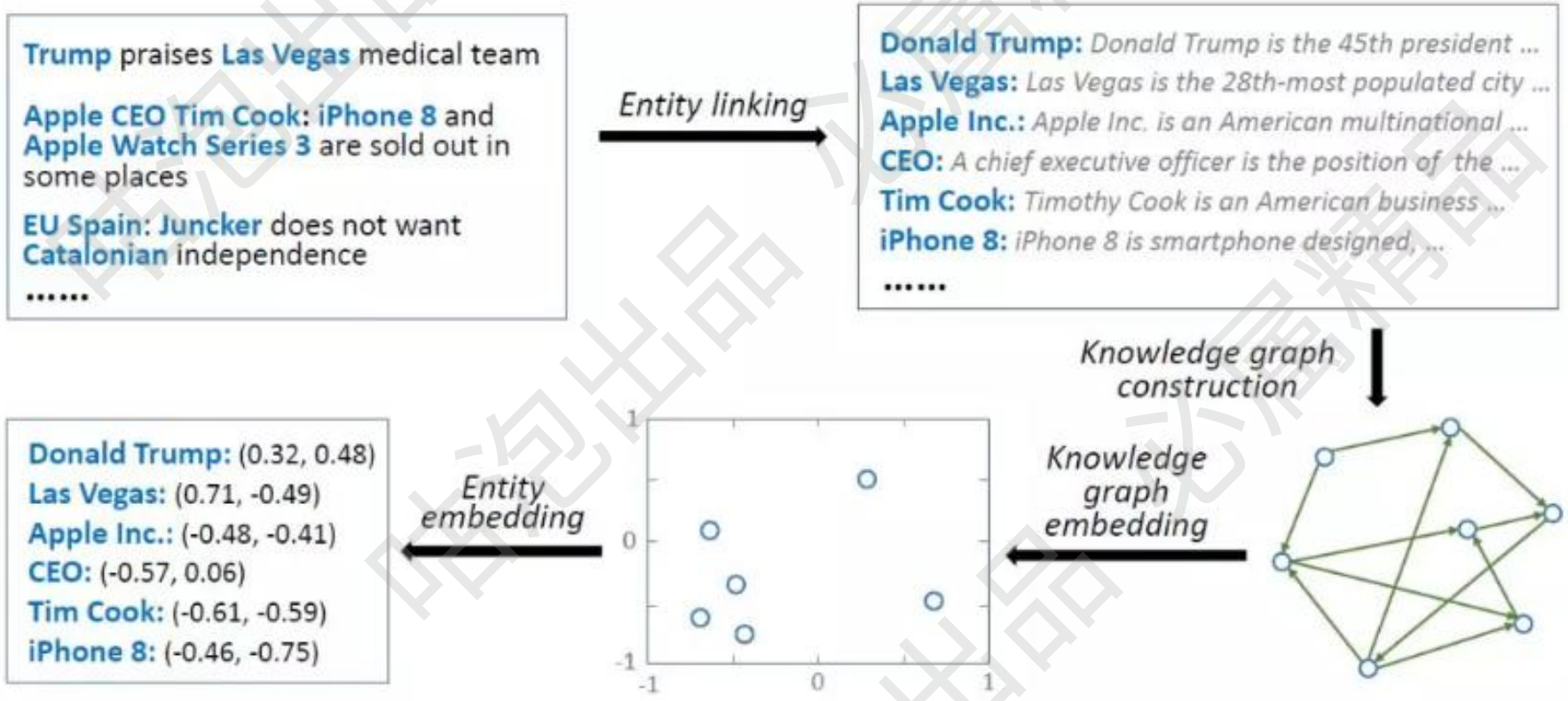
Rank			DBMS	Database Model	Score		
Feb 2020	Jan 2020	Feb 2019			Feb 2020	Jan 2020	Feb 2019
1.	1.	1.	Neo4j	Graph	51.21	-0.45	+3.35
2.	2.	2.	Microsoft Azure Cosmos DB	Multi-model	31.95	+0.44	+7.09
3.	↑ 4.	3.	OrientDB	Multi-model	4.94	-0.17	-1.11
4.	↓ 3.	4.	ArangoDB	Multi-model	4.85	-0.35	+0.50
5.	5.	5.	Virtuoso	Multi-model	2.77	+0.12	-0.17
6.	↑ 7.	↑ 7.	Amazon Neptune	Multi-model	1.96	+0.23	+0.90
7.	↓ 6.	↓ 6.	JanusGraph	Graph	1.88	+0.11	+0.65
8.	↑ 9.	↑ 11.	Dgraph	Graph	1.16	+0.11	+0.45
9.	↓ 8.	↑ 10.	GraphDB	Multi-model	1.14	+0.00	+0.28
10.	↑ 14.	↑ 18.	FaunaDB	Multi-model	0.98	+0.18	+0.62
11.	↓ 10.	↓ 8.	Giraph	Graph	0.98	-0.04	-0.03
12.	↑ 13.	12.	Stardog	Multi-model	0.91	+0.10	+0.22
13.	↓ 11.	13.	TigerGraph	Graph	0.87	-0.13	+0.19
14.	↓ 12.	↓ 9.	AllegroGraph	Multi-model	0.84	-0.01	-0.04
15.	15.	15.	Blazegraph	Multi-model	0.64	0.00	+0.11
16.	16.	↓ 14.	Graph Engine	Multi-model	0.58	+0.00	+0.01
17.	17.	17.	InfiniteGraph	Graph	0.39	+0.00	-0.01

知识图谱

- 业务还是算法?
  - 都重要，但是业务决定了算法的选择和数据需求以及模型的建立
  - 非常熟悉业务才能设计出实用的知识谱图，业务和设计起决定性作用
  - 不同的应用场景业务和设计也是完全不同，需具体分析
  - 算法很多都是通用的（命名实体识别，graph embedding等）

知识图谱

- 特征表达尤为重要



知识图谱

- 知识融合：
  - 知识就是力量
  - 特征进行融合
  - 得到最终的向量
  - 数据多就全用上