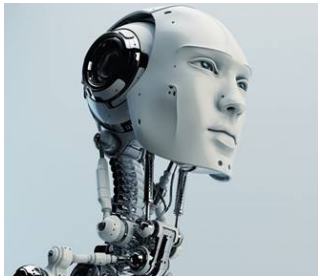
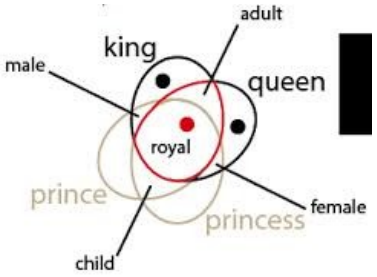
降低知识图谱的构造成本

文因互联
鲍捷
baojie@memect.co

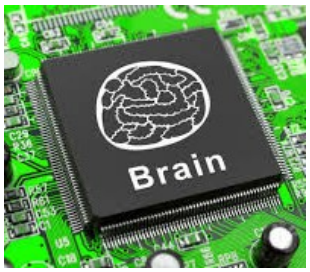
知识图谱



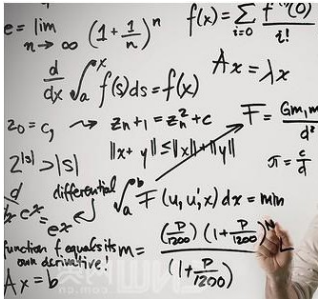
媒体报道的



学术界发表的

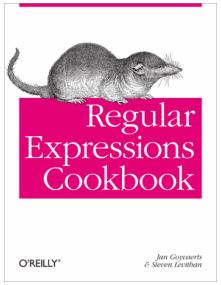


大公司宣传的



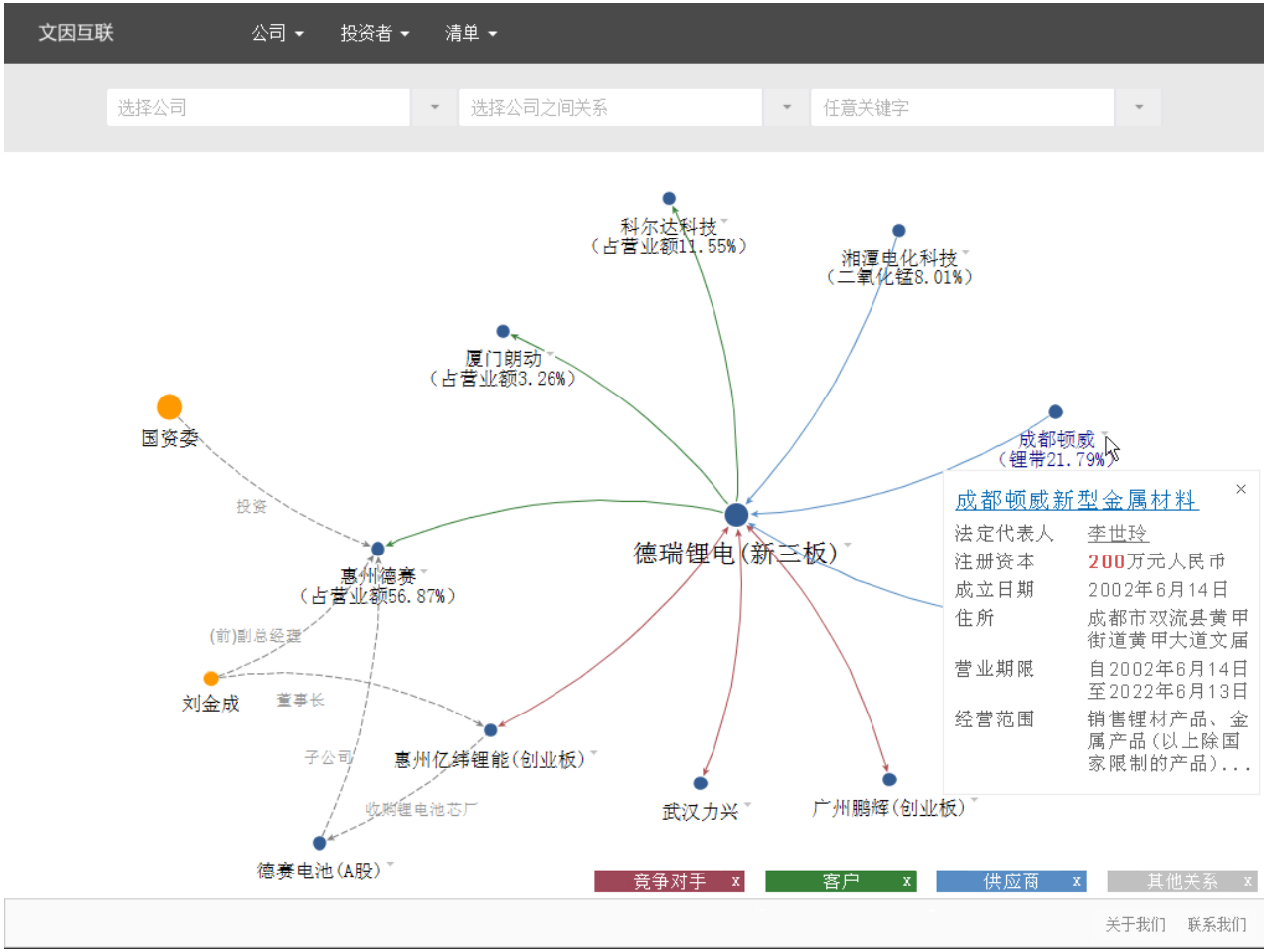
别人以为我干的

我以为我干的



我实际干的

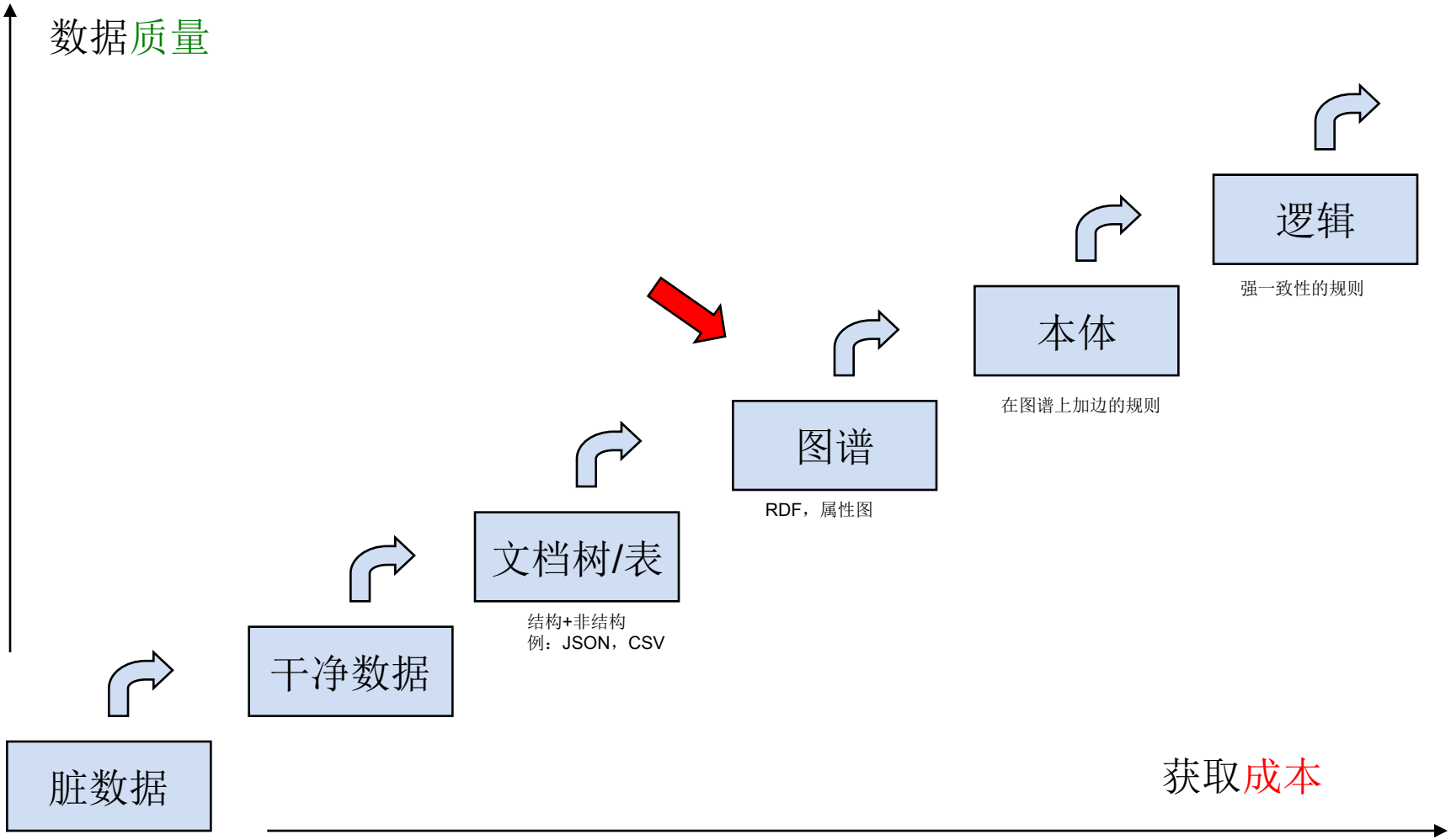
文因互联：金融知识图谱


文因互联
公司
投资者
清单
选择公司
选择公司之间关系
任意关键字
科尔达科技
（占营业额11.55%）
湘潭电化科技
（二氧化锰8.01%）
厦门朗动
（占营业额3.26%）
国资委
投资
成都顿威
（锂带21.79%）
成都顿威新型金属材料
法定代表人 李世玲
注册资本 200万元人民币
成立日期 2002年6月14日
住所 成都市双流县黄甲街道黄甲大道文届
营业期限 自2002年6月14日至2022年6月13日
经营范围 销售锂材产品、金属产品（以上除国家限制的产品）...
惠州德赛
（占营业额56.87%）
德瑞锂电(新三板)
(前)副总经理
刘金成
董事长
子公司
惠州亿纬锂能(创业板)
收购锂电池芯厂
武汉力兴
广州鹏辉(创业板)
德赛电池(A股)
竞争对手
客户
供应商
其他关系
关于我们
联系我们

相关项目经历

- 1999 放射科图像元数据，PACS
- 2003 蛋白质本体，基因本体
- 2004 生物信息学数据语义集成（INDUS）
- 2005 动物特征本体；内窥镜本体；协作本体编辑器COB-Editor
- 2006 模块化本体开发，分布式推理
- 2007 推理中的隐私保护
- 2008 地图元数据（RPI Map）
- 2008 W3C OWL本体语言工作组成员
- 2009 语义维基和情报分析，受控自然语言
- 2009 AIR语言和法规建模
- 2010 XBRL（电子化年报）语义化
- 2010 会议元数据（ISWC）
- 2011 语义用户画像（三星）
- 2012 三星SVoice个人语音助手
- 2013 语义电子邮件，企业知识管理门户
- 2014 Emma可视化书签
- 2015 好东西传送门，基于知识库的新闻推荐
- 2016 新三板知识图谱，开放证券数据

知识图谱


数据质量
脏数据
干净数据
文档树/表
结构+非结构
例：JSON，CSV
图谱
RDF，属性图
本体
在图谱上加边的规则
逻辑
强一致性的规则
获取成本

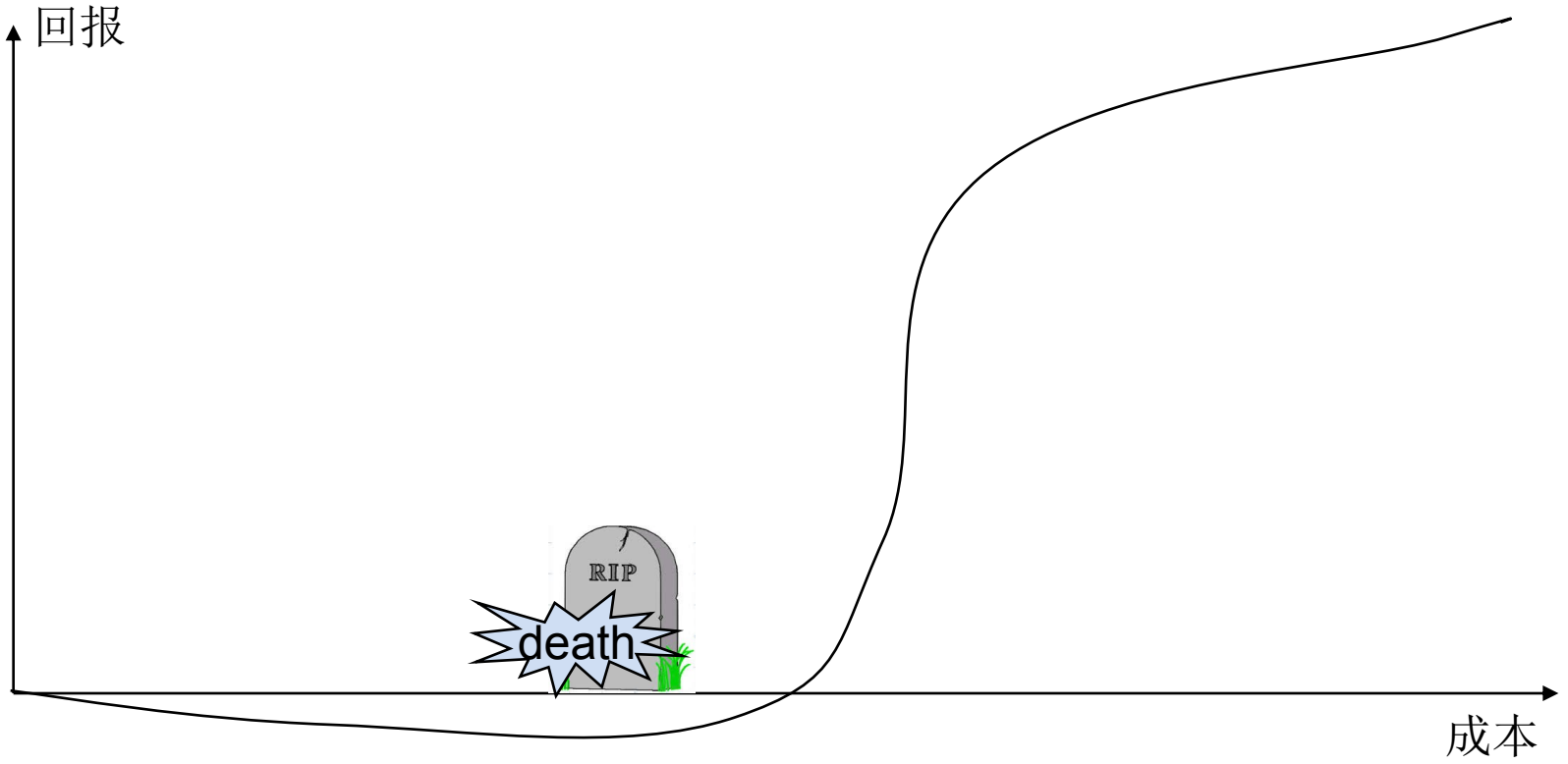
成 本

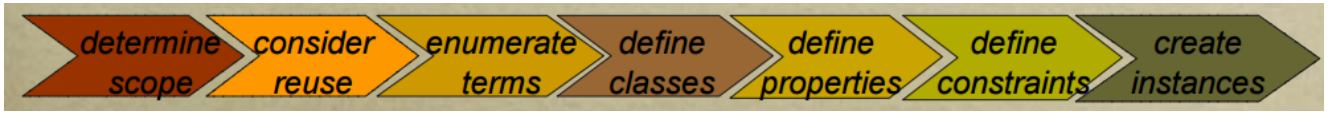
成本

阵亡名单


RIP
StreamGlider
RIP
Sig.ma
RIP
Twine
RIP
Evri
RIP
Truevert
RIP
Hakia
RIP
Powerset
RIP
....

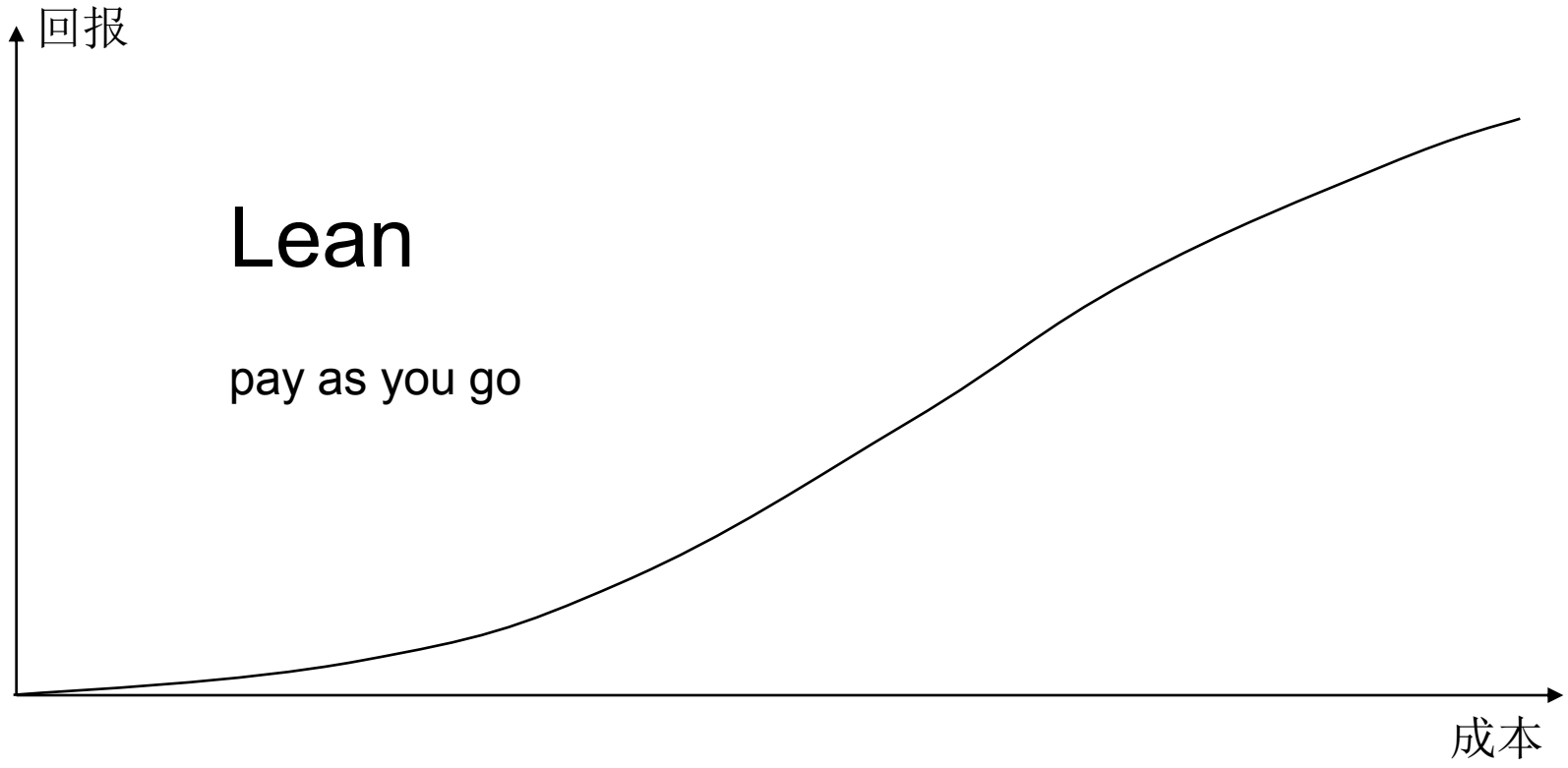
“重”的知识项目



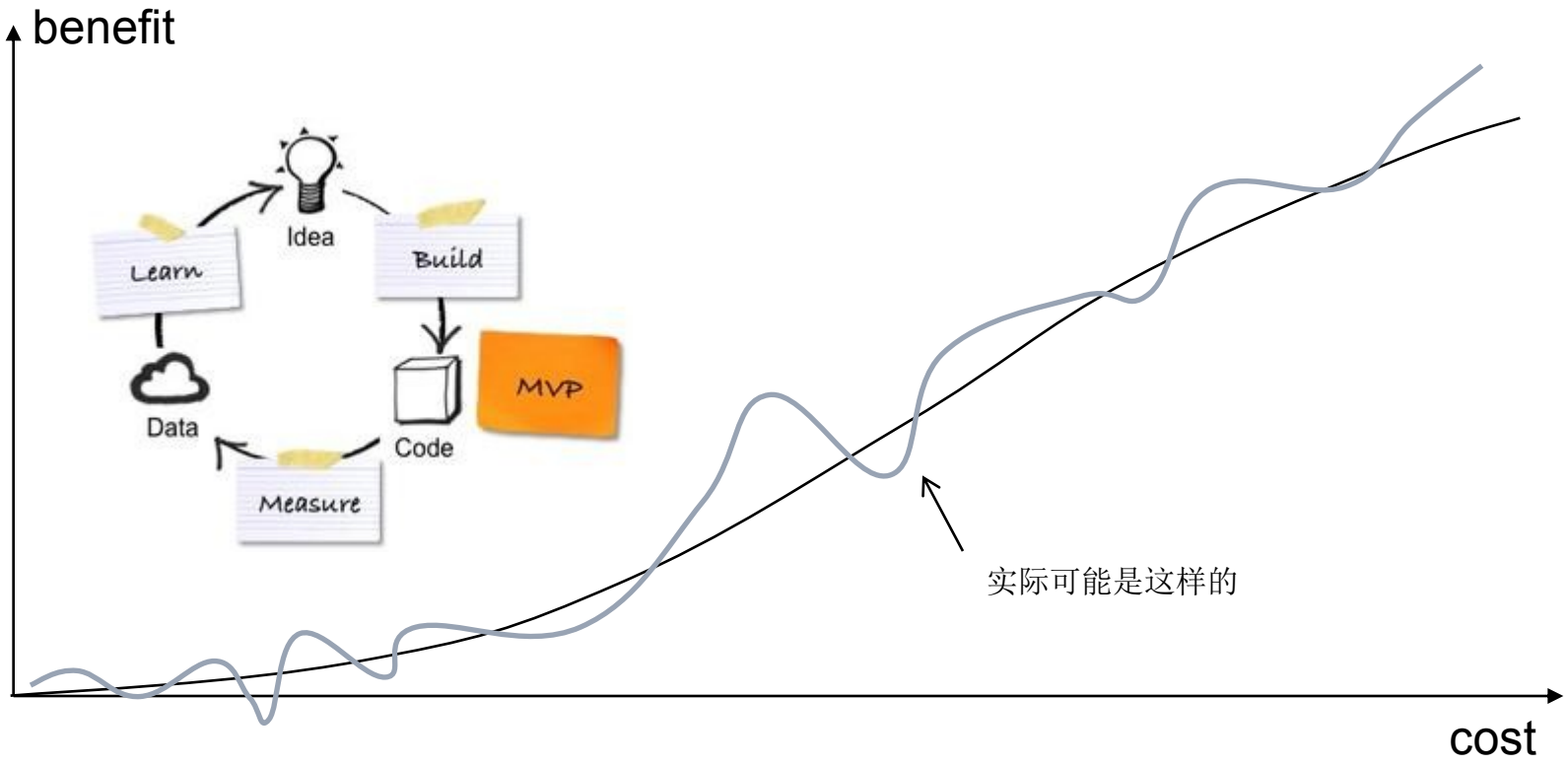


斯坦福七步法

“轻”的知识项目

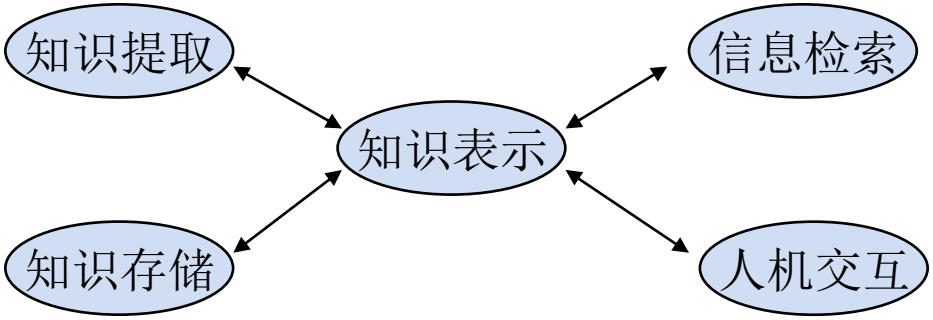


“轻”的知识项目

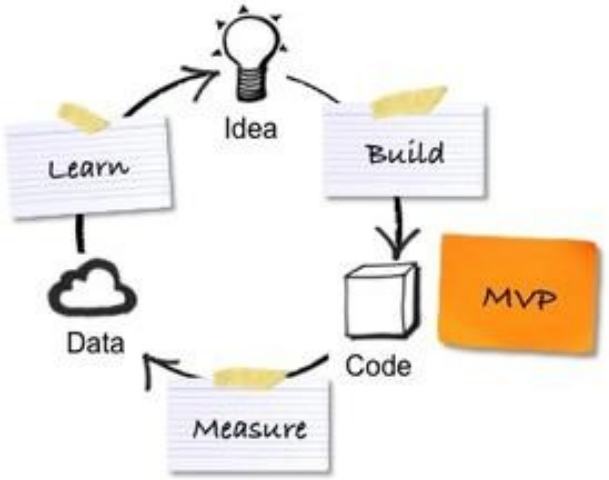

benefit
Idea
Learn
Build
Data
Code
MVP
Measure
实际可能是这样的
cost

降低成本

- 依托成熟技术



- 迭代



成本的核心是人

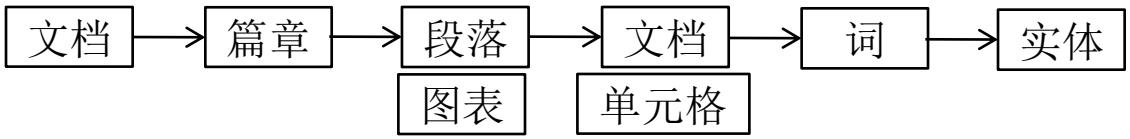
- 傻子都能写出计算机可读懂的代码，优秀的程序员写出的是人能读懂的代码 -- Martin Fowler

- 程序是写给人读的，只是碰巧能被机器执行 -- Abelson and Sussman

知识图谱也不例外

知识提取的成本

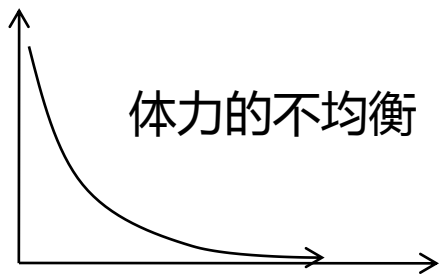
- 人工 vs 算法
  - “有多少人工就有多少智能”
  - 算法难以提取高质量结构。算法转移结构。
- 统计 vs 规则
  - 样本大小？训练集成本？准确率的要求？源数据的质量？
- 粒度

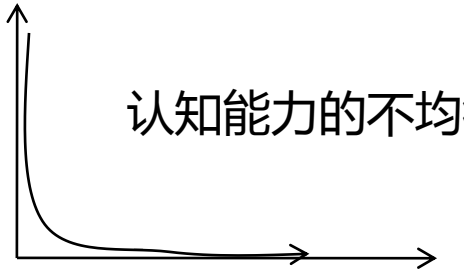


  - 例：提取公司商业模式（段落）
  - “最低成本的结构依靠中学语文”
    - 例：purple semantic mediawiki

知识表示的成本（0）

- 降低成本的核心是适应认知惰性





- 隔离世界观冲突
- 降低必须的初始投入

知识表示的成本（1）

- 命名（Naming）的成本
  - 计算机科学里只有两个困难的事情：缓存失效和起名字 -- Phil Karlton
- URI vs 字符串

URI	高成本	全局	唯一	机器可读	易互联	有主人
字符串	低成本	上下文相关	多义	人可读	难互联	可无主人

- 分离 naming 和 addressing
  - Reference by description
  - 牺牲全局唯一性，提高可读性，降低成本

知识表示的成本（2）

元组的成本

- 关系和属性，类和实例的建模是大多数人做不到的。
- 三元组（主-谓-宾）难以表达定、状、补
  - “三元组是尼安德特人的语言”
  - 三元组方便RDF语义，但极大降低可用性
  - Reification 降低可读性和可维护性
- JSON更加符合人的认知
  - JSON-LD，Document-Graph DB

知识表示的成本（3）

本体的成本

- 本体是一种世界观
  - 本体的采用是权力的让渡
  - 本体的制定是政治
  - “语言是有军队的方言”，本体是有钱的偏见
  - e.g., schema.org
- 应尽可能推后或弱化世界观的冲突
  - 大多数应用其实不需要本体
  - 尽量避免使用顶层本体（upper ontology）
  - 多总结，少设计。好的工程都是总结出来的。

知识存储的成本（1）

- 混合结构与非结构数据
  - 在应用迭代中逐步提升结构性

Lucene, Solr, Elasticsearch

Neo4j Titan MongoDB Virtuoso Stardog …..

  - 两类查询
    - 如何找到节点？（定位。IR或search问题）
    - 如何从一个点到另一个点（遍历、路径查询问题）
  - 大数据量时，两类查询最好分离

知识存储的成本（2）

- 维护成本

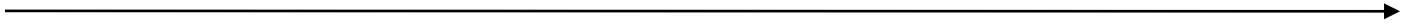
键值数据库 < 关系数据库 < 文档数据库 < 图数据库 < RDF数据库

成本上升
人员难找
更不稳定
插件越少
Bug难修

  - PostgreSQL + JSON 大多数时候足够用

- 效率

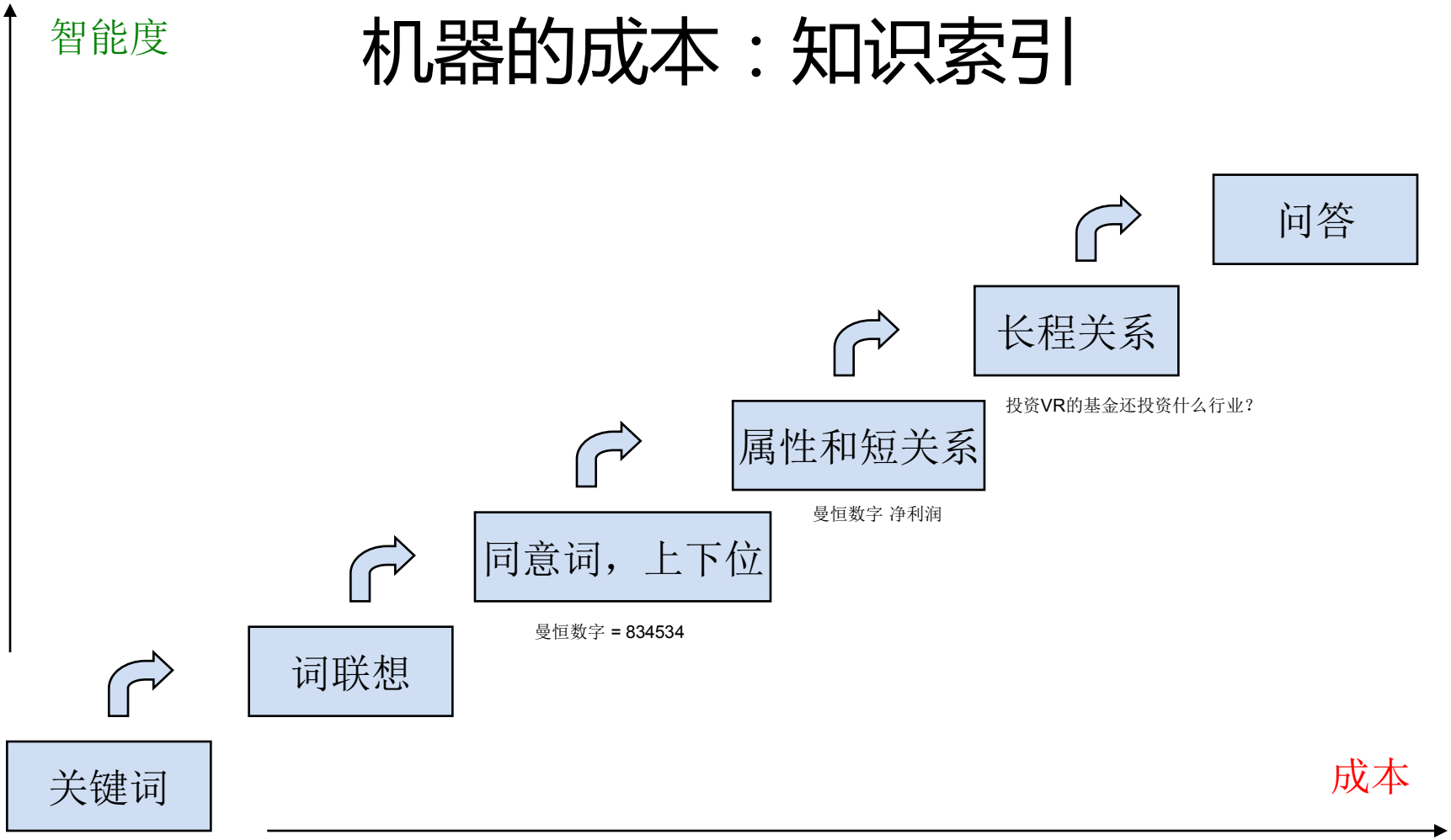
内存数据库 > 关系数据库 > 文档数据库 > 图数据库 > RDF数据库



支持高响应、分布式

（几乎）不支持

知识检索的成本（1）

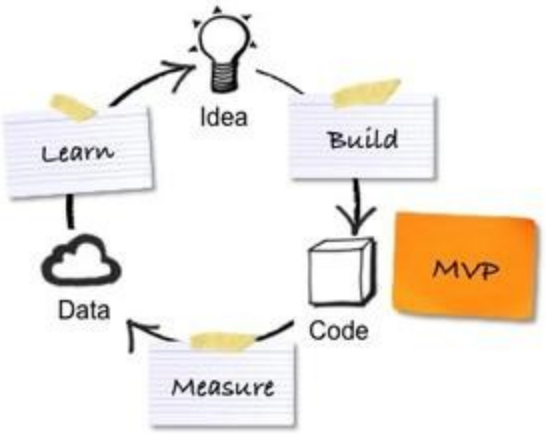

智能度
机器的成本：知识索引
关键词
词联想
同意词，上下位
曼恒数字 = 834534
属性和短关系
曼恒数字 净利润
长程关系
投资VR的基金还投资什么行业？
问答
成本

知识检索的成本（2）

- 人的成本：有限的注意力、记忆力、时间
- 分面浏览器（Faceted Browser）：探索引擎
  - 类比： 数据库:电子表格 :: 搜索引擎:探索引擎

总结

- 迭代



- 人
  - 知识是写给人读的，只是碰巧能被机器执行

个人微信

文因互联期待新的加盟者

- 机器学习/NLP工程师
- 语义网技术工程师

联系邮箱 contact@memect.co