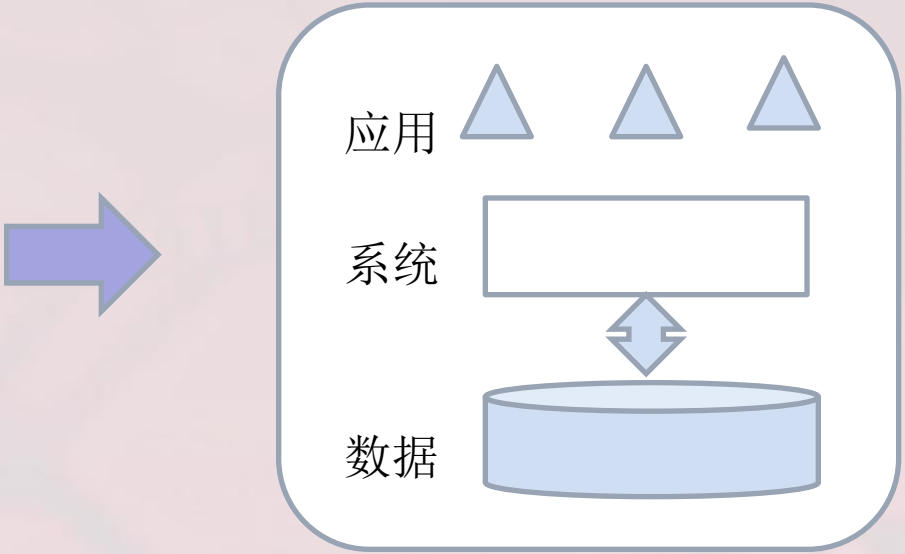
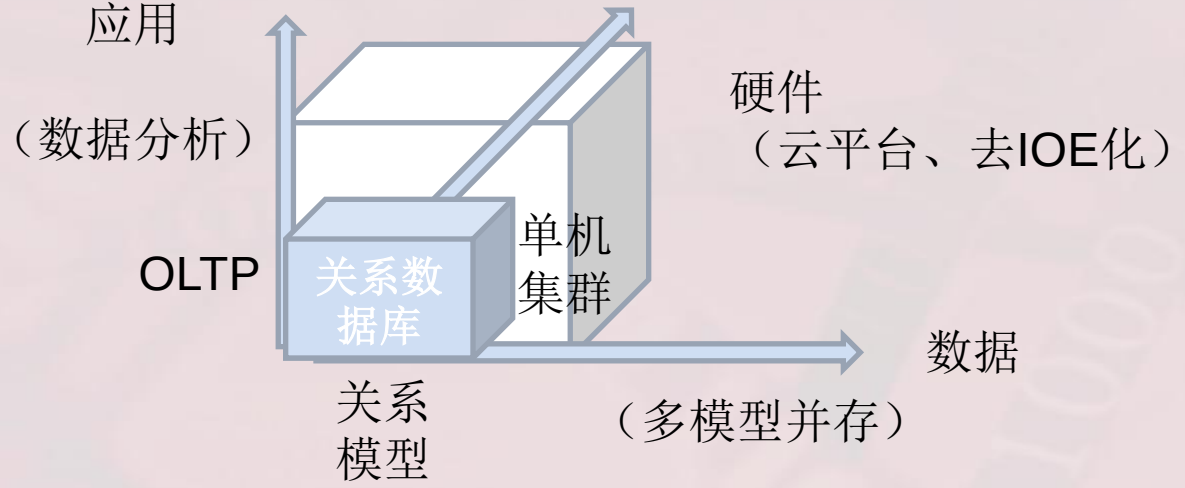
2、大数据的系统特征



大数据系统

- 大数据系统=大数据+大数据管理系统
- 大数据的数据特征，隐含了对系统的技术要求
- 大数据管理系统是指对4V大数据提供有效的存储组织、管理和分析等功能的软件系统

影响数据管理系统的三要素



三个主要变化

- 从三要素的演变中形成了三个主要变化
  - 从封闭世界到开放世界
  - 从量的管理到量质融合
  - 从数据管理到知识管理

变化1：从封闭世界到开放世界

- 关系数据库遵循：
  - 封闭世界假设： 不在数据库里的都是假的！
  - 模式是事先定义的(封闭的结构)：不能更改。是系统管理数据的依据，是用户查询数据库的基础
  - 基本操作是固定的(封闭的代数系统)：关系操作！

从封闭世界到开放世界

- 开放世界:
  - 不在数据库里的不一定是假的，只是目前是未知的。
  - 假设数据库中有“张三是 中国人”的陈述，问“张三是美国人吗？”，答案应该是“不知道”。

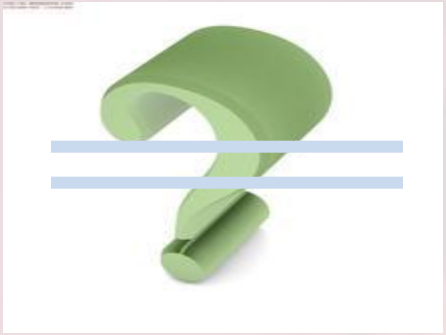
从封闭世界到开放世界

- 大数据环境下：
  - 数据类型：从结构化数据到非结构化数据，要支持不同的数据模型。
  - 数据模式：无模式，或者难以事先确定，
  - 数据操作：用户定义的、更加复杂的操作
  - 负载的不确定性：要支持“现象级”的应用压力，要适应不断扩张/缩减的计算平台（分布式）

- 这些特征颠覆了传统数据库的前提，数据库系统面临一场革命。

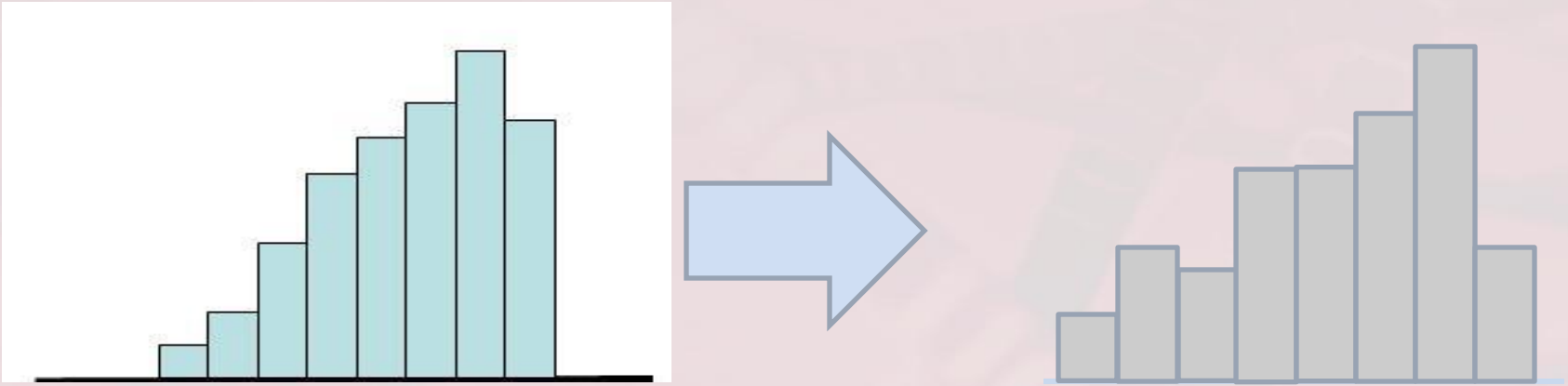
一个例子

- 问：这两幅图像是同一个建筑吗？



图像的直方图表示法

EMD(Earth Mover's Distance)



用推土机将左边的图“推成”右边的图的最小移动土方数

相似度定义

- Sim(s,t)=EMD(s,t)
- Sim(s,t) < ε 称为s 和t 相似

相似查询/连接

- 给定一组数据对象集合R，相似度函数sim()，查询条件s
- 相似查询

  $Q=\{ t \mid t \in R, sim(s,t) < \varepsilon\}$
- 相似连接

  $T=\{ (s,t) \mid s\in R, t \in R, sim(s,t) < \varepsilon\}$

准确率 vs 性能

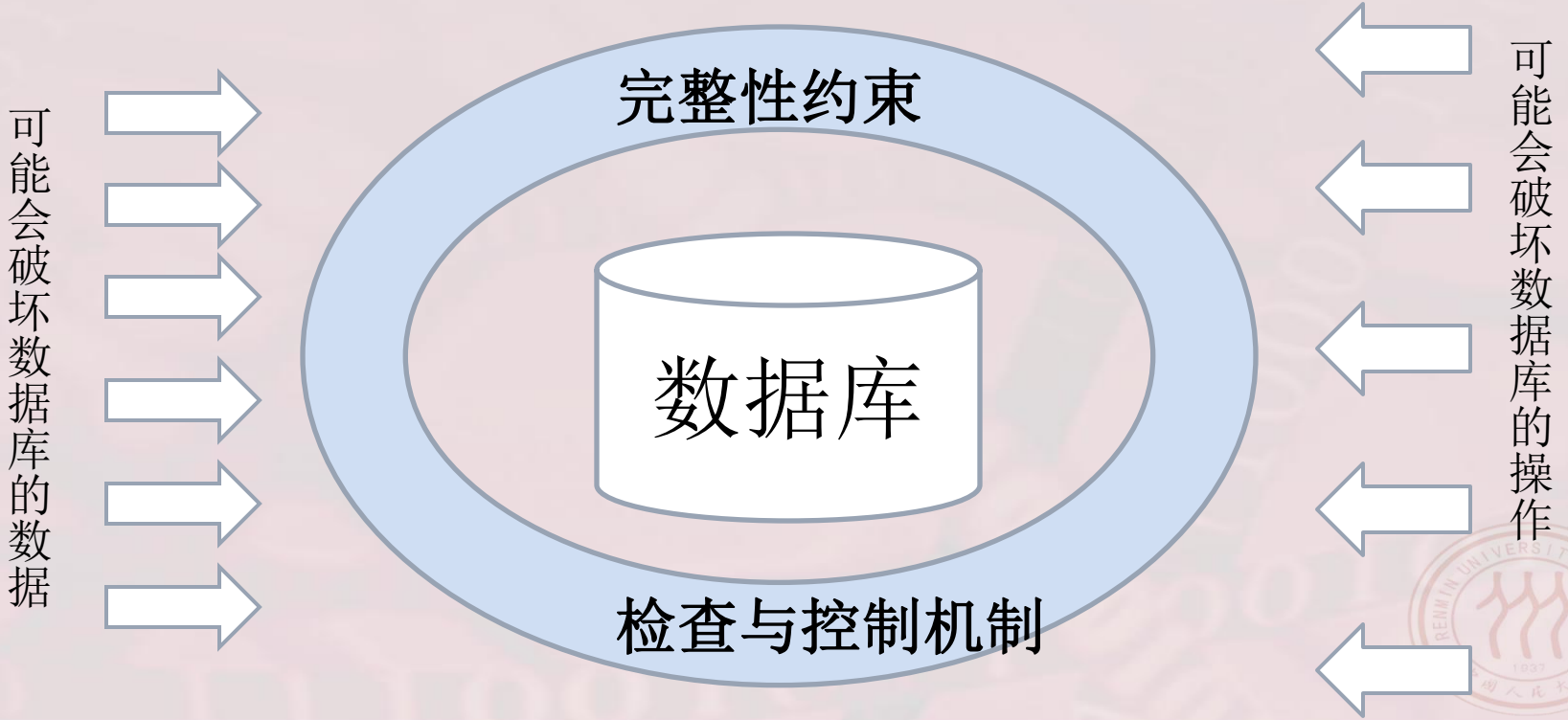
- 相似查询、相似连接都是非结构化数据管理系统中非常基础性的一个操作，但是，不同于关系数据库中的操作
  - Sim() 是UDF（用户定义的函数）
  - 计算复杂，性能差

- 如何为R建立索引？没有索引查询速度不可接受
- 如何考虑在分布式平台上部署数据?

变化2：从精确数据到不精确数据

- 关系数据库是精确的，主要关心的是大容量下的性能问题，高效维护数据一致性的技术等
- 量质融合的大数据系统，从单一的量的管理到量质融合管理的转变。

关系数据库



开放环境下的数据库

管理数据质量的基本思路

- 数据清洗将成为系统的核心部件。
- 用信息检索的方法取代信息查询。
- 用数据分析/机器学习“容忍”劣质

- 例子：问“张三是美国人吗？”，转换成问题“张三是哪国人？”
- 系统：给出top-k的答案。

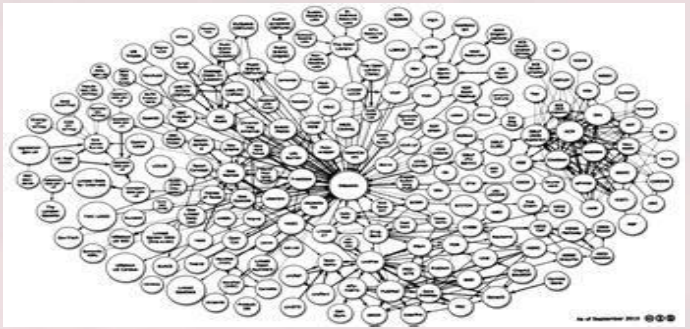
变化3：从数据管理到知识管理

- 关系数据库其实无关语义，关系的语义、属性的语义都隐含在了其名字和数据中， 只有程序员才能理解。数据的语义靠属性的语义来解释。

- 大数据是独立存在，大数据是自描述的，没有外在的模式等进行语义的描述。因此，从大数据中如何获得知识是关键。

从大数据到大知识

- 如何从大数据中抽取知识？
- 知识图谱的方法（离散型知识）

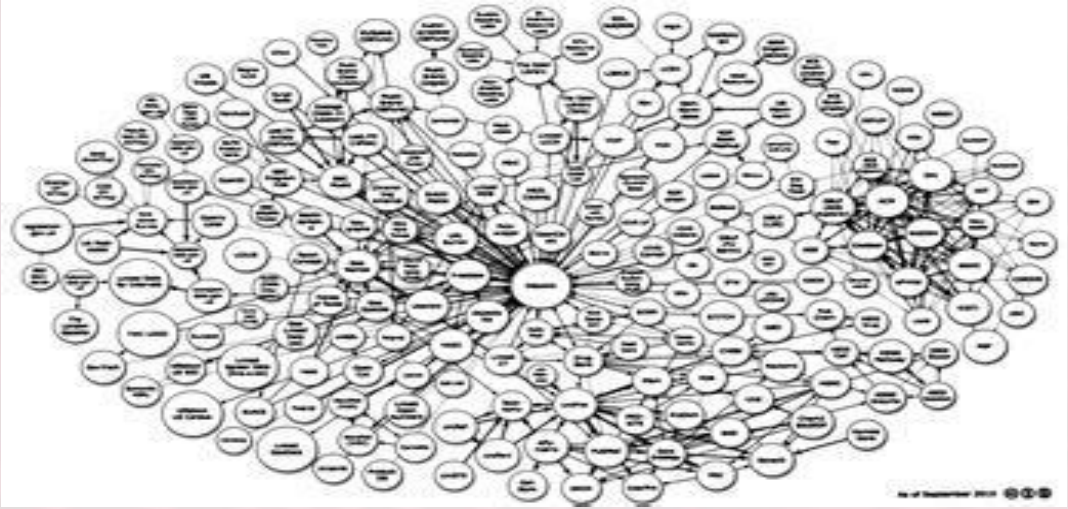


大数据

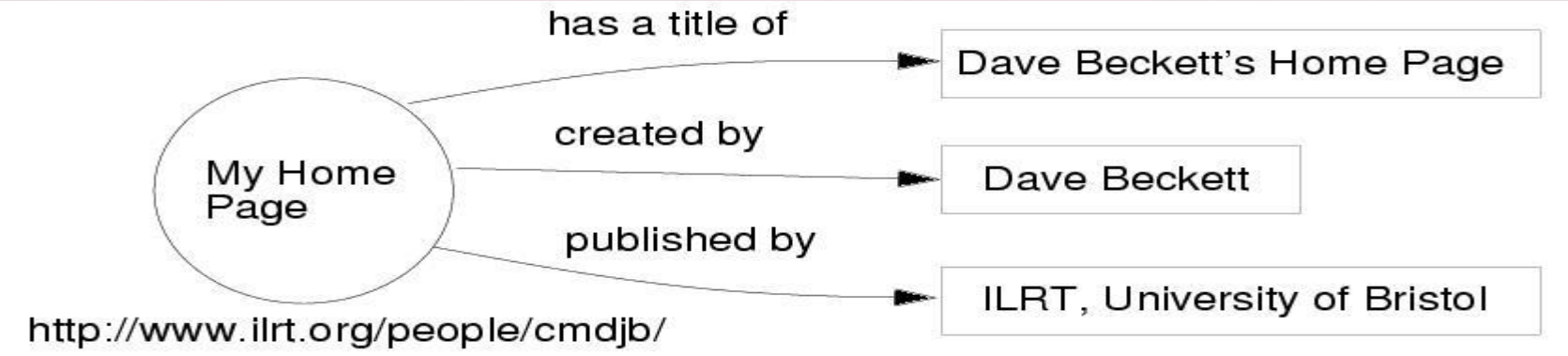
大知识

自动知识获取与知识库构建

- 从大数据中自动地提取并组织成知识库
- 通过不断的纠错演化，形成可用的知识库



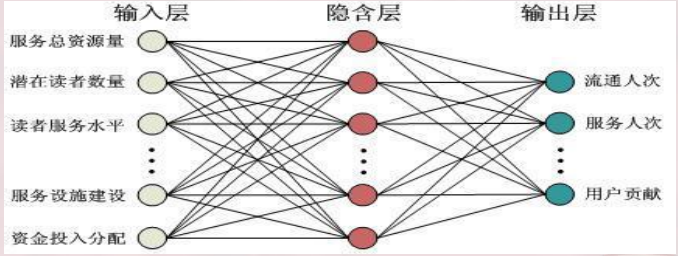
RDF数据


has a title of
Dave Beckett's Home Page
created by
My Home Page
Dave Beckett
published by
ILRT, University of Bristol
http://www.ilrt.org/people/cmdjb/

从大数据到大知识

- 如何从大数据中抽取知识？
- 深度网络的方法（连续型知识）

大数据



大知识

小结

- 三个变化隐含了大数据管理系统的功能需求：
  - 从封闭世界到开放世界
  - 从量的管理到量质融合
  - 从数据管理到知识管理