PARIS: Probabilistic Alignment of Relations, Instances, and Schema

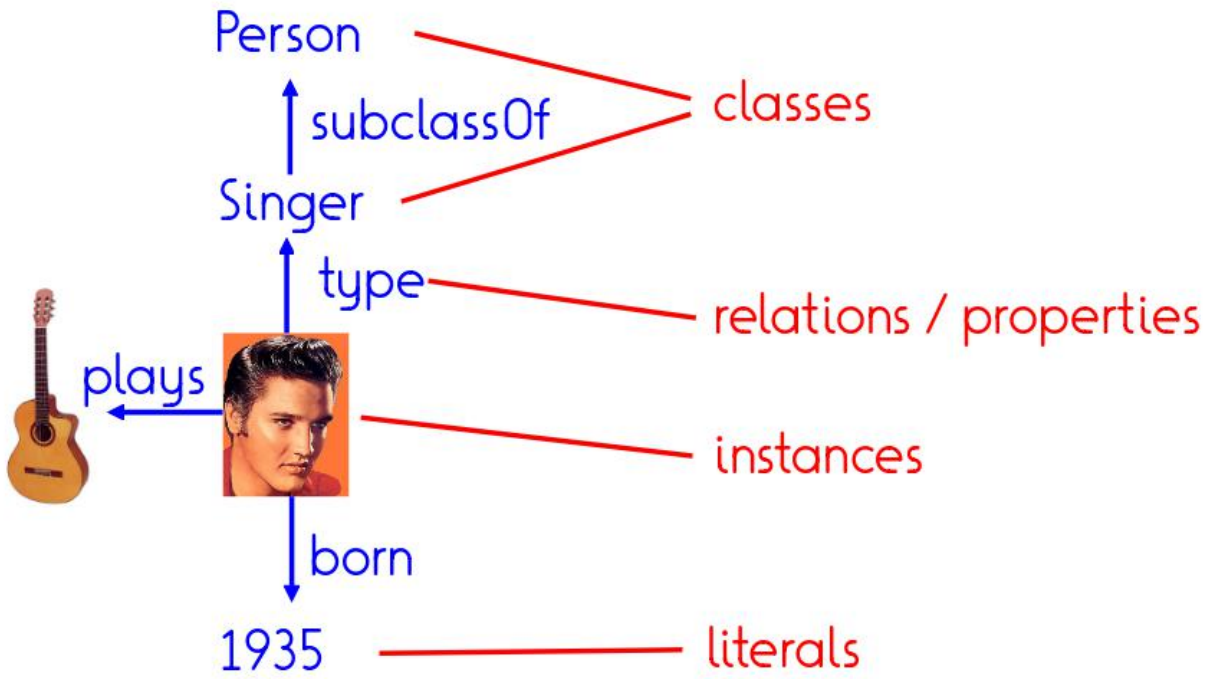
MG1833058 潘云逸

Motivation

- 由于不同的本体可能会用不同的词语描述相同的实体，并且这些本体的信息可以相互补充和完善，因此将这些本体对齐就显得非常重要。

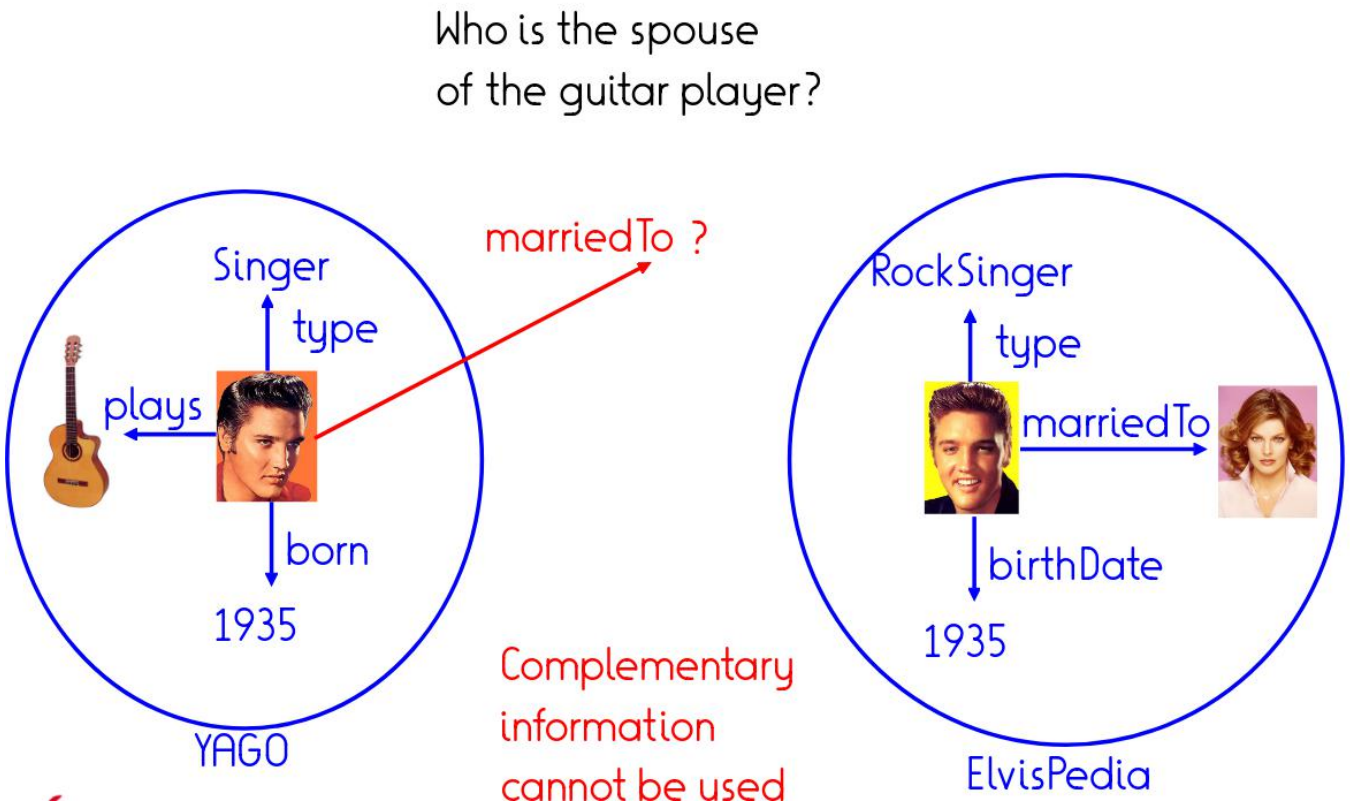
RDF Ontologies

- 一个RDF本体可以视作一个实体的图

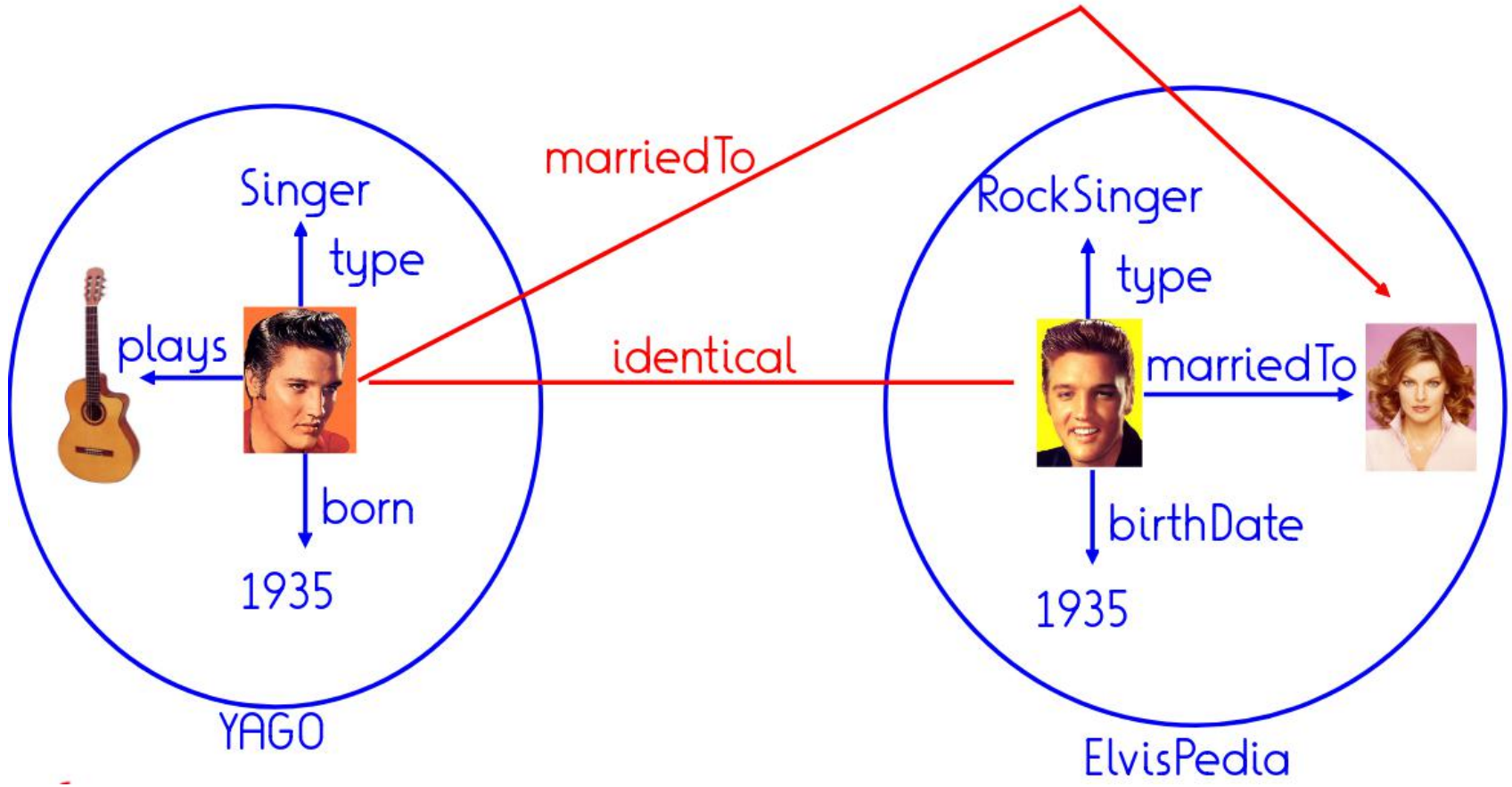


Problem

- 很多本体有着相似的或者交叉的实体和事实属性



Solution：实体统一



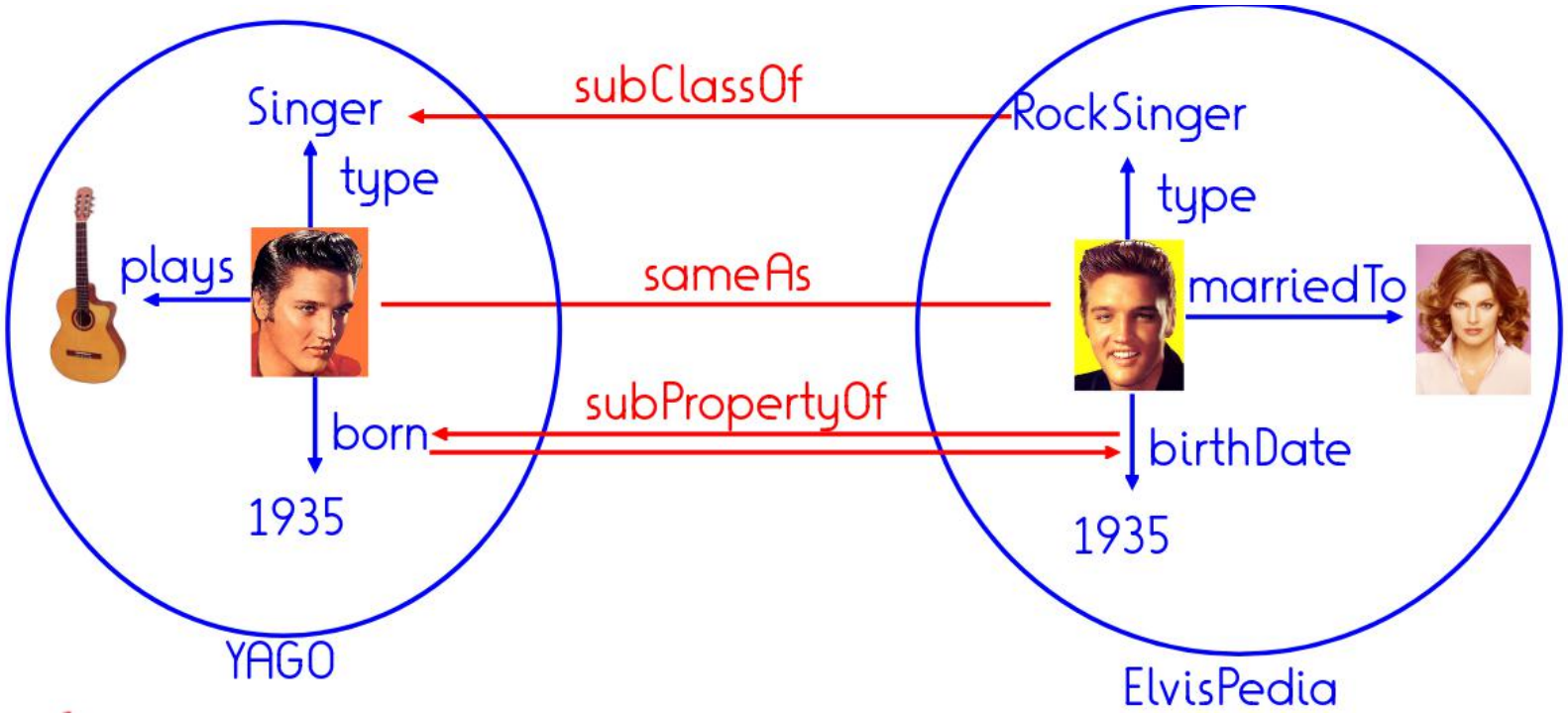
Goal:合并本体

为了合并本体，我们需要明确

1.对等的实例

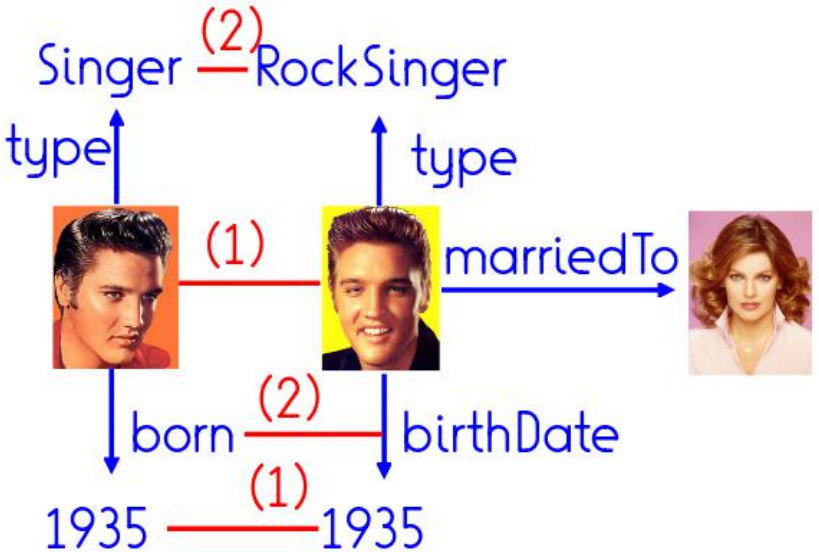
2.对等的或者能够归入的类别

3.相等的或者能够归入的关系



Related Work

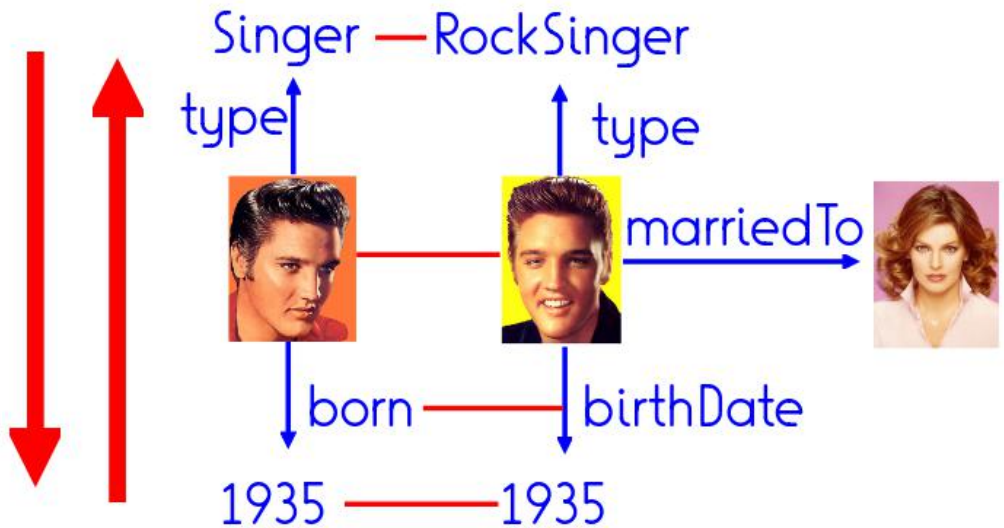
- 使用规则限制，效果不好
- 需要参数微调
- 在大型本体上没有使用过
- 主要聚焦于（但没有都做到以下两点）
  - 1.实例匹配
  - 2.模式对齐



PARIS:一次性完成所有对齐

在实例、属性、类别中存在协同等效性

一次计算完成所有对齐



Probabilistic Model

$Pr(c_1 \subseteq c_2) =$

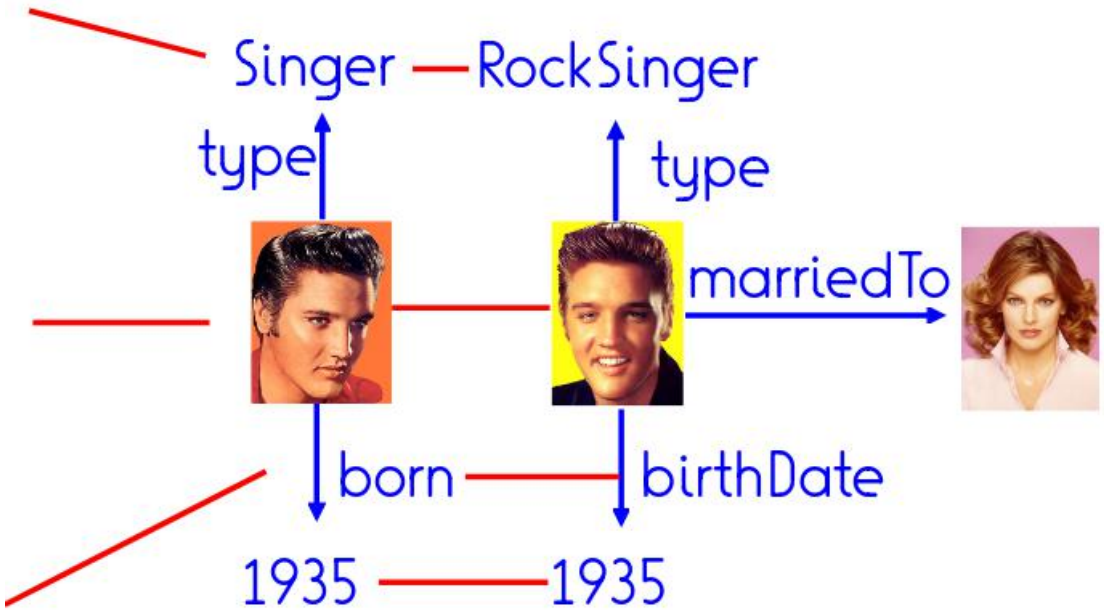
the probability that c_1
is a sub-class of c_2

$Pr(x \equiv y) =$

the probability that x=y

$Pr(p_1 \subseteq p_2) =$

the probability that p_1
is a sub-property of p_2



Probabilistic Model

- 两个常量相等的概率反映了两个常量属于同一事物的相似性

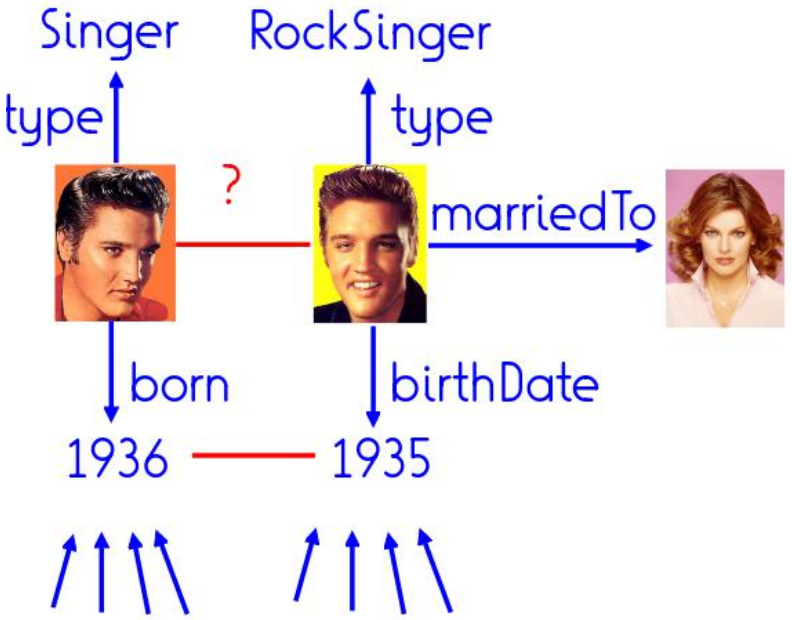
$Pr(x \equiv y) =$

- for strings:
  string distance
- for numbers:
  numeric distance
- for other literals:
  domain-specific

$Pr(x \equiv y) := (x{=}y)\ ?\ 1{:}0$

Local Inverse Functionality

- ifun(x,y):=1/|{x:r(x,y)}|
- 没有很多人叫Elvis=>较高的指示性
- 很多人在1935年出生=>较低的指示性
- 如果只有一个人叫Elvis
  - ifun(label,Elvis)=1
- 如果有10个人在1935年出生
  - ifun(born,1935)=0.1
- 使用局部逆函数的调和平均来表示一个关系的概率
  - Pr(ifun(label))=0.8 (很少有人名字相同)
  - Pr(ifun(born))=0.01(很多人出生年份相同)



Equality of Instance

- 我们假设两个本体有着同一关系
  - 如果存在 y≡ y' 并且有r(x,y),r(x',y')可以写作

$$\exists\, r, y, y' \; with \; r(x,y), r(x',y'): \;\; y \equiv y' \wedge ifun(r)$$

  - 根据上式可知 x≡x'
- $Pr(x \equiv x')$

$$= Pr(\exists\, r, y, y' \; with \; r(x,y), r(x',y'): \;\; y \equiv y' \wedge ifun(r))$$

$$= 1 - \Pi_{r(x,y),r(x',y')} \, (1 - Pr(y \equiv y') \, Pr(ifun(r)))$$

Equality of Classes

- 如果一个类中的所有的实例是另一个类的实例，则前一个可以归入后者

$$Pr(C \subseteq D) = \frac{|C \cap D|}{|C|} = \frac{\Sigma_{x \in C} Pr(\exists\, y \in D: x \equiv y)}{|C|}$$

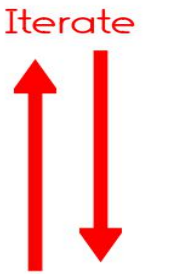
$$Pr(C \subseteq D) = \frac{\Sigma_{x \in C} (1 - \Pi_{y \in D} (1 - Pr(x \equiv y)))}{|C|}$$

Equality of Relations

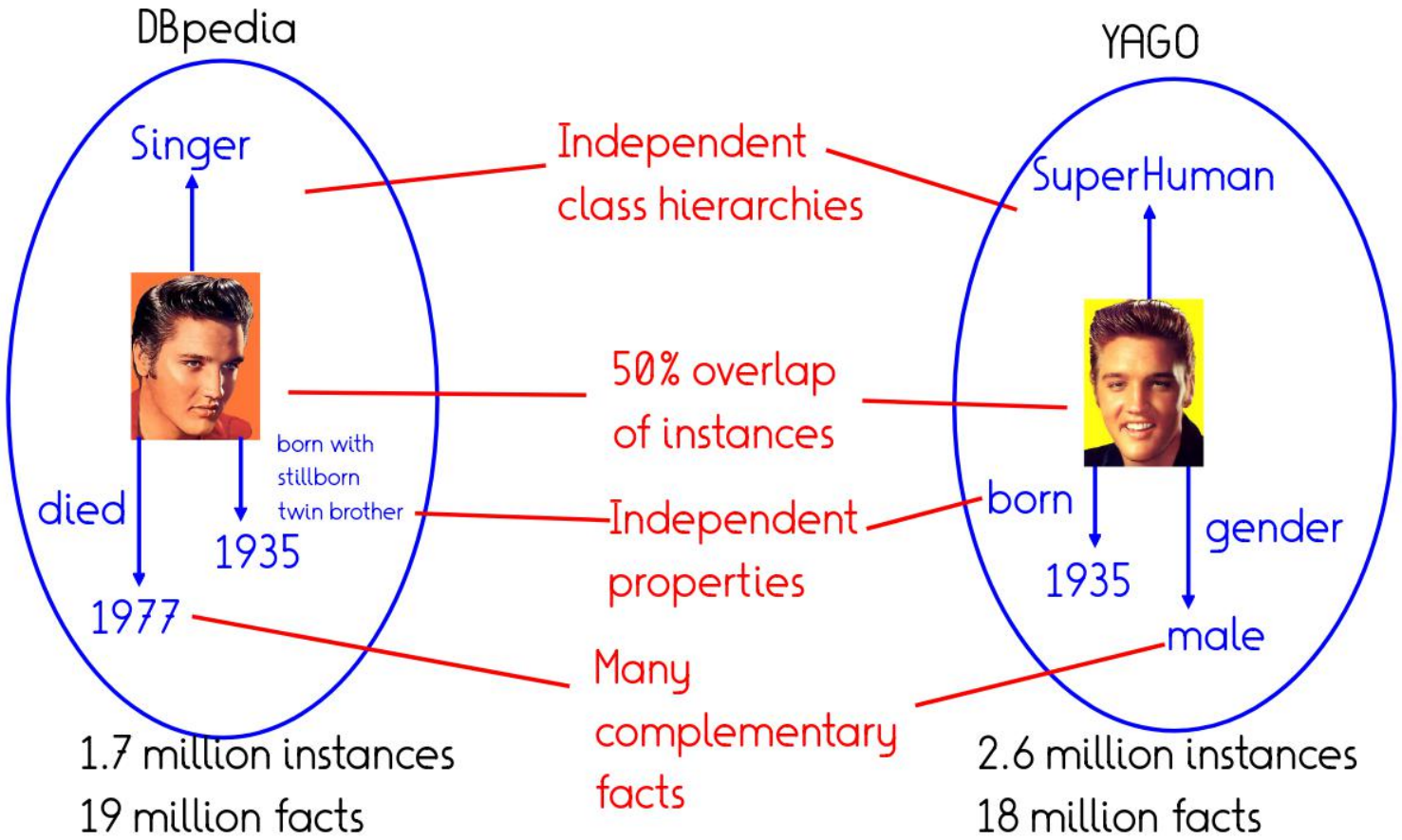
- 如果一个关系中的每一对都是另一个关系中的关系对，则前者是后者的子属性

Algorithm

- 固定常量的等式
- 将关系的等式设定到一个很小的初始值
- 迭代对关系和实例的估计直到收敛
- 计算类别的估计
  - 关系：$Pr(p_1 \subset p_2) = 13\phi^{\wp} \mathbf{f} \ldots$
  - 实例：$Pr(x \equiv y) = \Pi_{42}^{1} \alpha^{\beta} \ldots$
  - 常量：$Pr(x \equiv y) = fixed\ for\ literals$
  - 类别：$Pr(c_1 \subseteq c_2) = \pi\ mc^2 \ldots$



Experiment: YAGO and DBpedia



Conclusion

- 优点：
  - 没有使用任何参数
  - 一次性对齐所有要素
  - 算法性质优良，基本适合任意两个本体之间的计算，并能取得较好效果
- 不足：
  - 目前不能处理结构不同的本体
  - 如果两个本体之间的粒度差异过大，也不能取得很好的效果