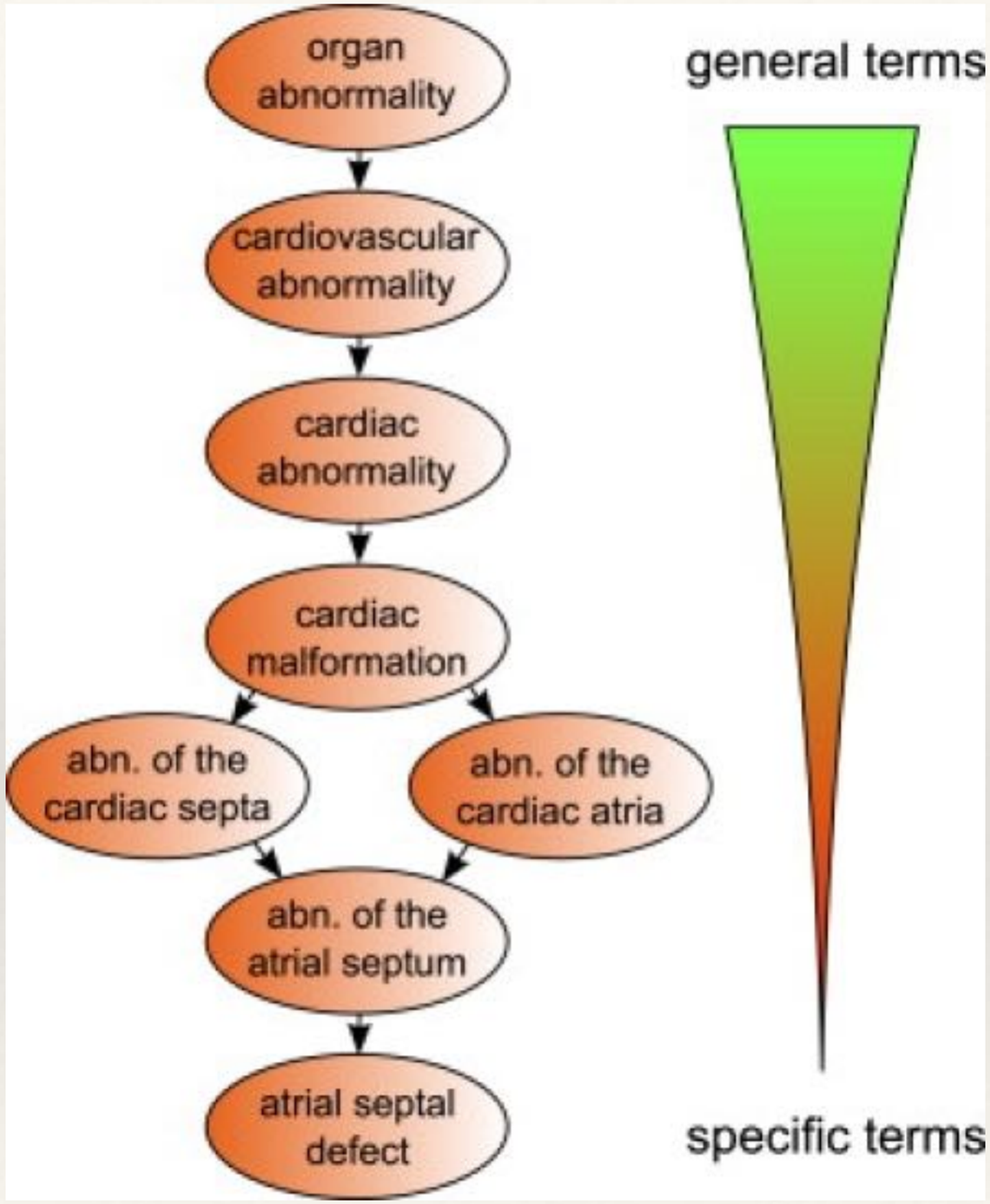
Phenomizer论文分享

20180508

HPO数据库是个有向无环图（DAG）

- 节点是hpo，有向边表示“是”的关系
- 特例是树。区别是，树的每个节点只能有一个父节点，DAG每个节点可以有多个父节点
- 语义是，父节点广泛（如糖尿病），子节点具体（胰岛素抵抗型糖尿病）


organ abnormality
cardiovascular abnormality
cardiac abnormality
cardiac malformation
abn. of the cardiac septa
abn. of the cardiac atria
abn. of the atrial septum
atrial septal defect
general terms
specific terms

注释的true path 原则

- 注释是疾病与hpo的关联
- true path原则：与某hpo关联的疾病自动隐式地与该hpo的祖先节点关联

- 问题：输入一个查询（一组hpo），对于每一个疾病（一组hpo）计算匹配度，根据匹配程度排序

- 子问题：

A. 两个hpo的相似度（能以此得到两组hpo的相似度）

B. 相似度（raw）转换为有意义、可比较的匹配度

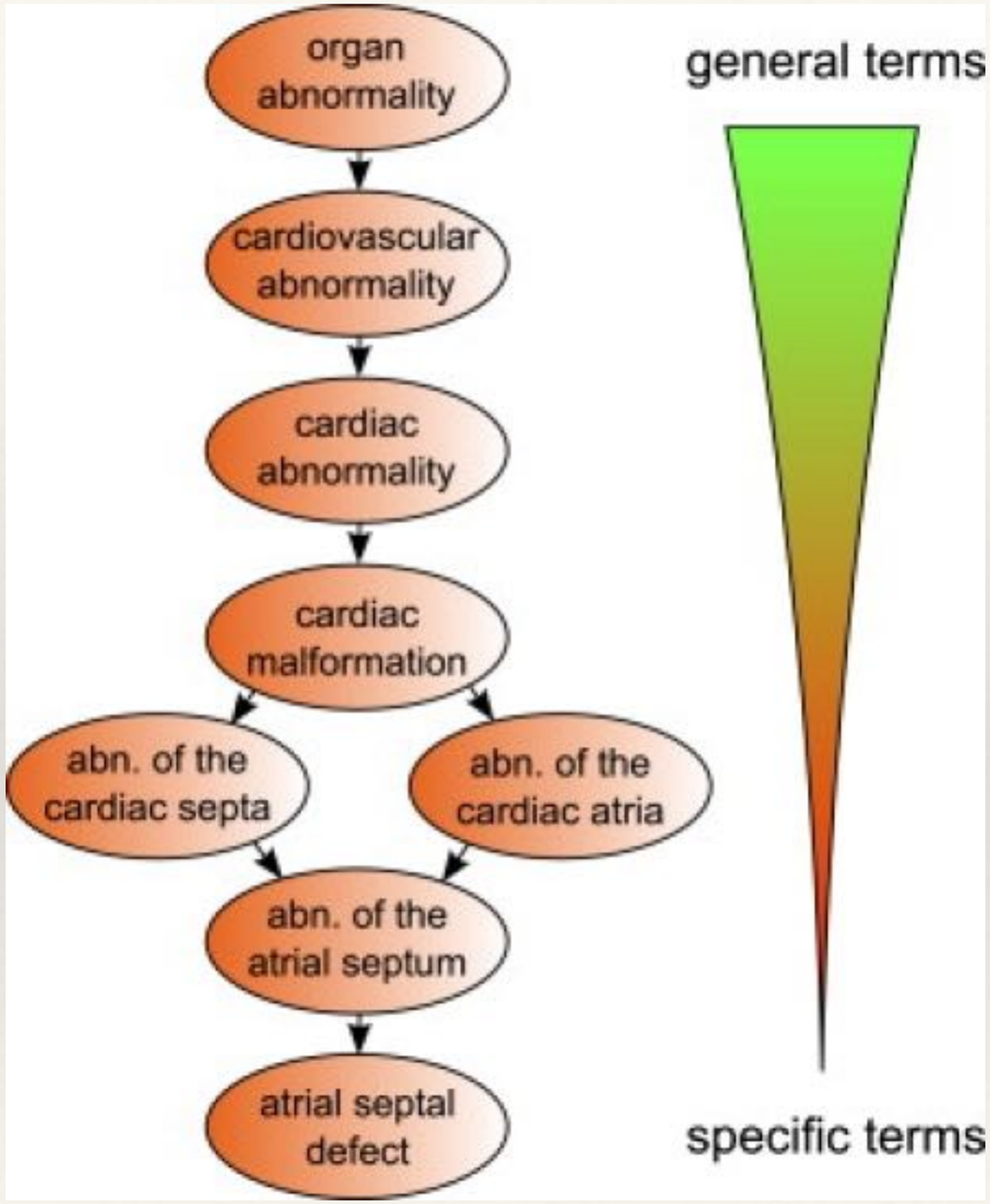
B.raw相似度的处理（统计模型）

- raw相似度受很多因素影响：hpo的具体程度，查询的长度，疾病的具体程度等，不能直接比较。机械排除每个因素的影响是不可能的。考虑统计模型。
- 假设查询是随机产生的一组hpo，它与疾病的匹配应该是坏的。反之，如果匹配较好，就得出结论—查询有诊断上的意义，匹配真实的疾病。
- 我们假设“查询无意义/随机”（空假设），获得该假设下raw相似度的概率分布（空分布），如果一个查询的raw分数（或更高）在空假设下出现的概率小，说明这个查询很有可能不符合空假设，即这个查询很有可能匹配真实的疾病。
- 最终的“不匹配度”：查询的raw分数在空分布下出现的概率（p-value），小于一个阈值则认为显著，越小匹配越好
- p-value还需经过multiple testing correction（略）

- 获得空分布：蒙特卡洛法模拟 - 对1～10的查询长度，随机产生大数量的查询，计算&存储与每个疾病的raw相似度
- 10^9存储代价

A.两个hpo的相似度

- 两个hpo的相似度：DAG中两节点的相似度
- 一个hpo的IC 与（用该hpo注释的疾病个数 / 总疾病数）负相关
- MICA：most informative common ancestor，最低的共同祖先（如果P是A和B的MICA，则从P到A和B都有路径，且P的任何后代节点不满足这个条件）（见图）
- 两个hpo的相似度 = 它们的MICA的IC


organ abnormality
cardiovascular abnormality
cardiac abnormality
cardiac malformation
abn. of the cardiac septa
abn. of the cardiac atria
abn. of the atrial septum
atrial septal defect
general terms
specific terms

A.两组hpo的相似度

$$\mathrm{sim}\,(\mathrm{Q} \rightarrow \mathrm{D}) = \mathrm{avg}\left[\sum_{t_1 \in \mathrm{Q}} \max_{t_2 \in \mathrm{D}} \; IC\,(MICA\,(t_1, t_2))\right].$$

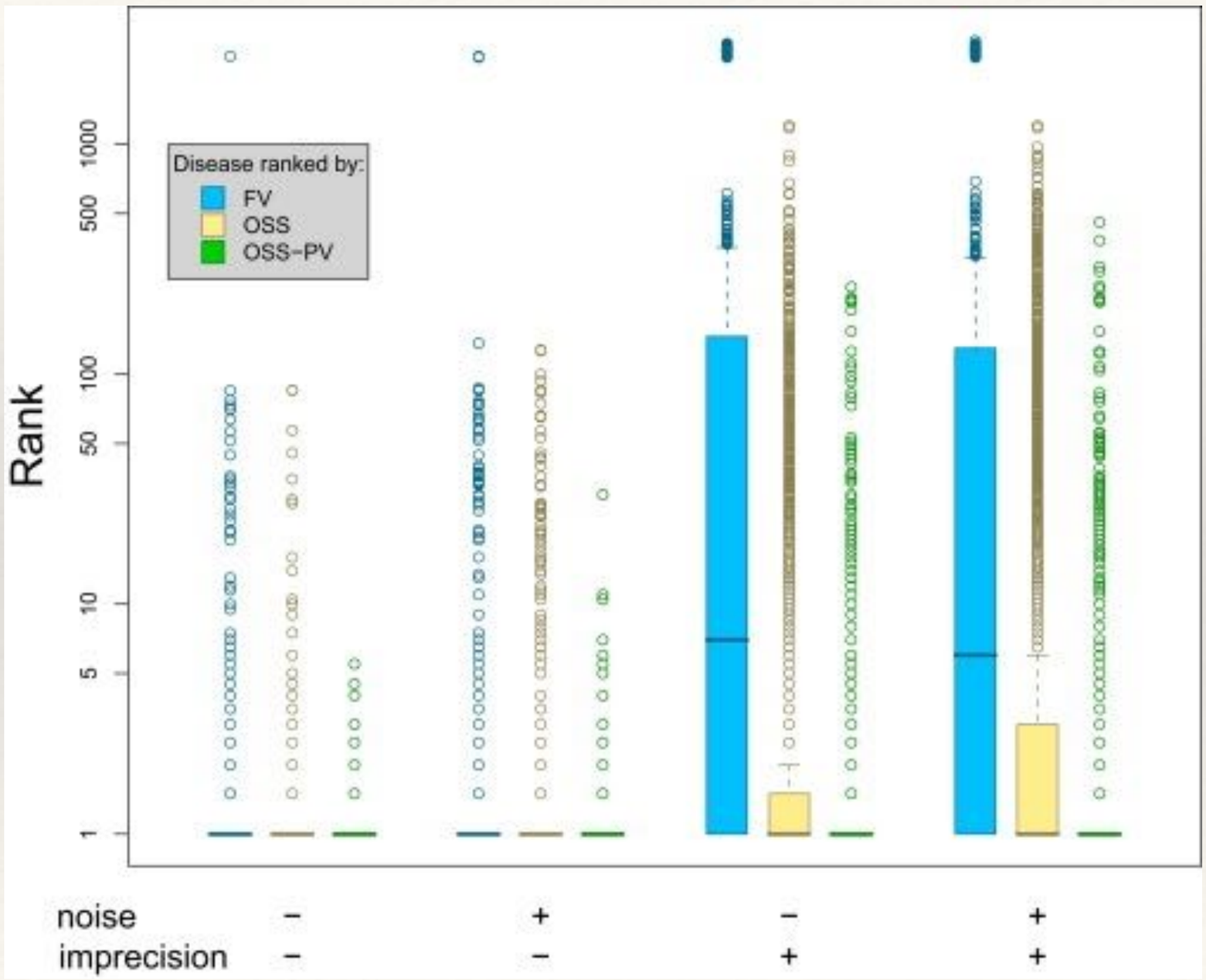
- 对于查询中的每个hpo，取其与疾病的hpo中相似度最高的作为它与疾病的相似度。最后查询与疾病的相似度对查询中所有hpo取平均。

效果

FV：查询HPO与疾病HPO的交集大小
OSS：raw相似分数
OSS-PV：raw相似分数得到的p-value

纵轴：返回的已排序的疾病列表中，正确结果的排名。

noise：患者身上的无关临床特征
imprecision：某临床特征替换为它的父类（更general）



总结

- hpo筛选gene：把疾病换成基因
- 前期准备计算量大
- DAG中两节点的相似度（距离）问题，有替代/更好的解法