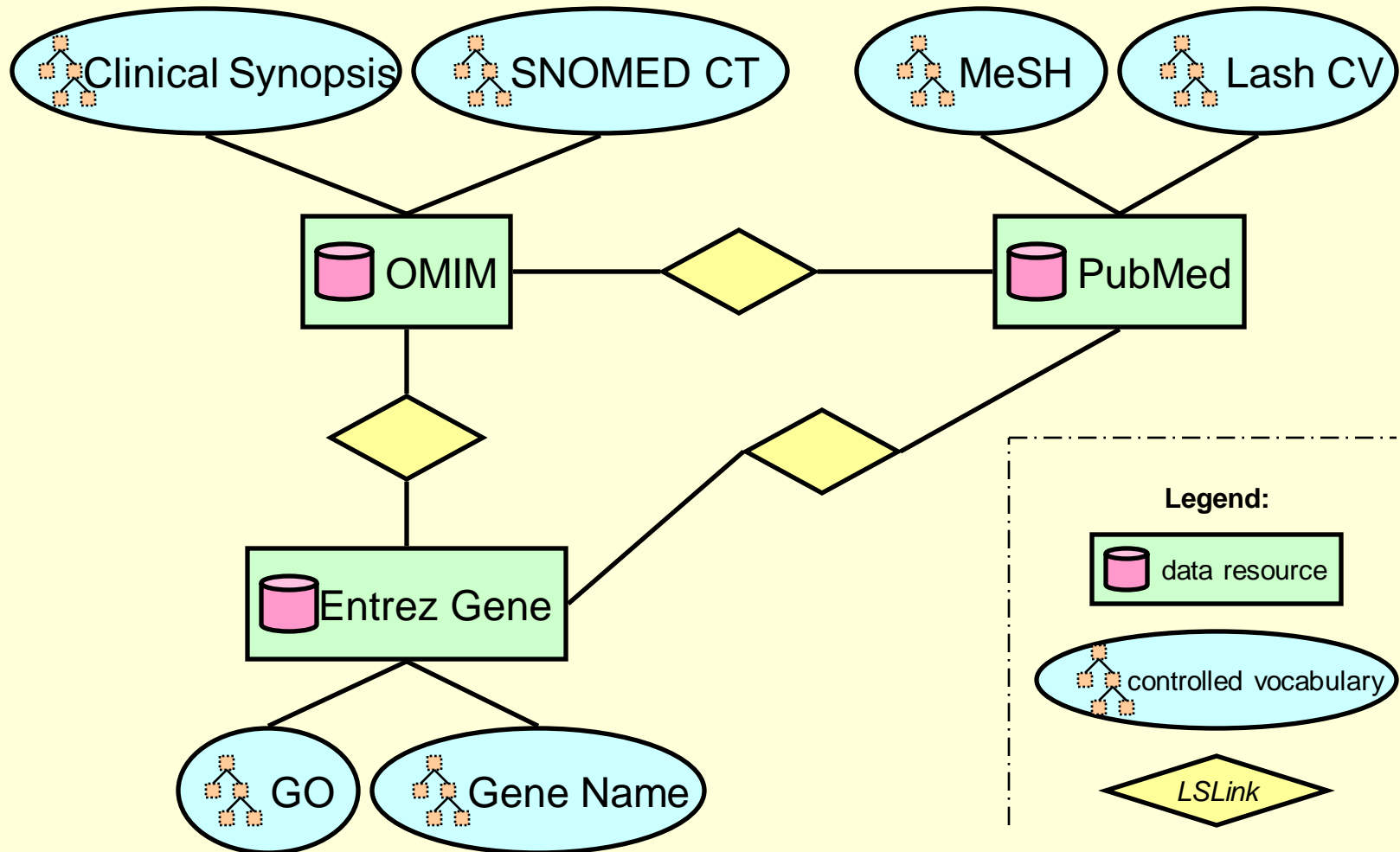


Using GO and MeSH annotations to find meaningful associations

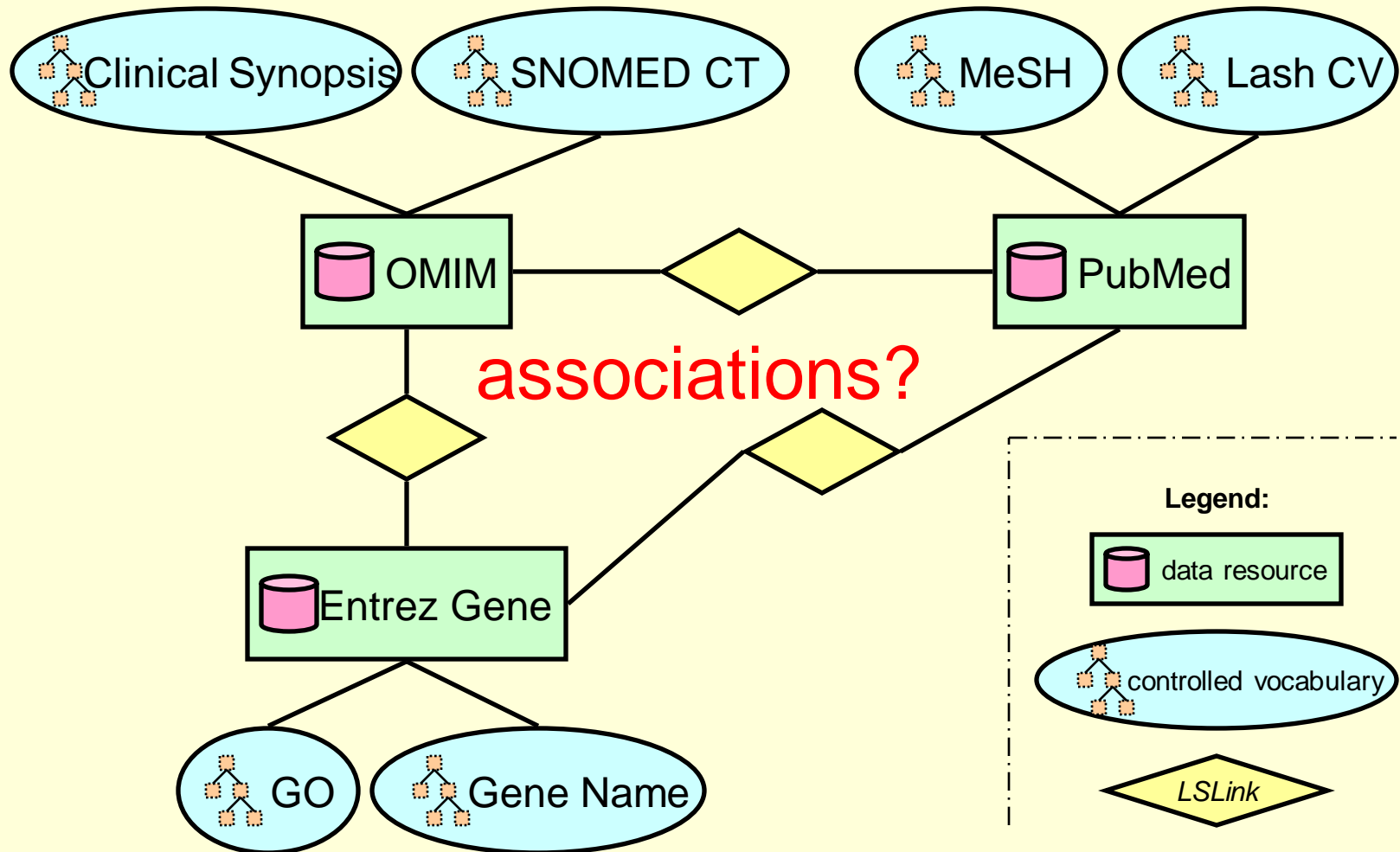
Woei-Jyh Lee, Louiqa Raschid (*Univeristy of Maryland*)
Padmini Srinivasan (*The University of Iowa*)

27 June 2007

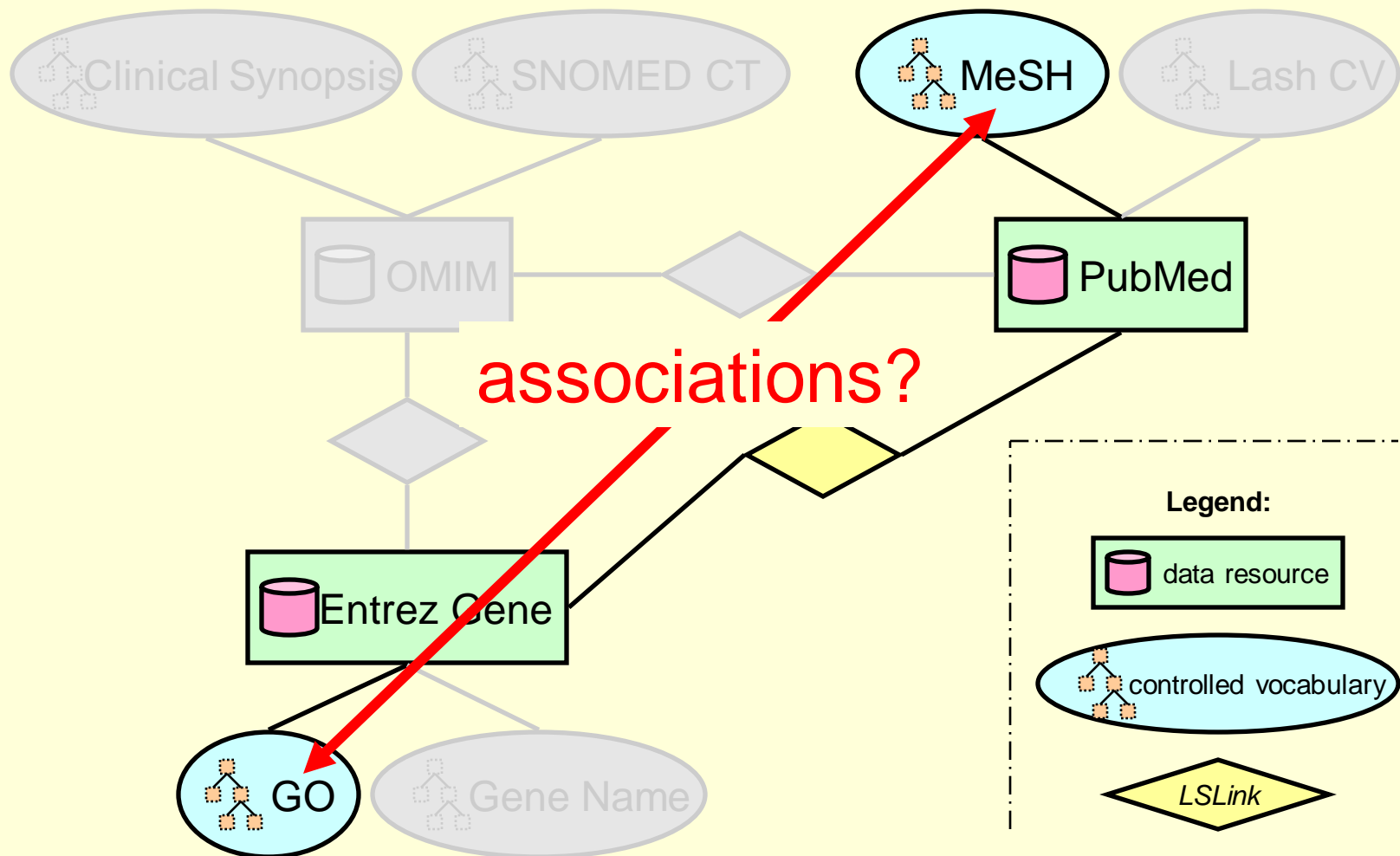
Hyperlinked Web in Life Sciences



Hyperlinked Web in Life Sciences



Human Genes and Publications





Outline

- Introduction
- Datasets
- Metrics
- Tool

Human Genes Background Dataset



- *LSLink* instances are generated as follows:
 - Retrieve **all active human gene** entries in Entrez Gene.
 - Filter out records been replaced and discontinued.
 - Extract their **GO annotations**.
 - Follow all links from these entries to **PubMed** entries.
 - Extract **MeSH annotations** for PubMed entries reached for the prior step.



User Query Scenarios

- To identify a subset of *LSLink* instances that will be **mined** to obtain possibly meaningful associations.
 - A scientist looks for publications about **a human gene** or a set of human genes.
 - A scientist wants to know all human genes associated with some **set of articles**.
 - A scientist is interested in specific **medical term(s)** and would like to retrieve highly related human gene(s).
 - ...

Breast Cancer in Human User Query Dataset



Gene symbol	Number of GO terms extracted from the gene entry	Number of distinct directly linked PubMed entries	Number of distinct MeSH terms identified as major topic in the PubMed entries	Number of <i>LSLink</i> instances generated
BRCA1	36	469	391	75,852
BRCA2	16	199	202	14,992
BRCA3	0	3	7	0
union set	49	582	434	90,844



Outline

- Introduction
- Datasets
- Metrics
- Tool

Metrics Used to Identify Potential Meaningful Associations



- The logarithm of the odds (**LOD**) based scores.
 - **Support** reflects the relative ratio of *LSLink* instances that associate the two CV terms with respect to all *LSLink* instances in the dataset.
 - **Confidence** reflects the relative ratio of *LSLink* instances that associate the two CV terms with respect to those *LSLink* instances that are associated with one of the CV terms.
- **Hypergeometric distribution** test.
 - **P-value** reflects the estimated relative ratio to observe an occurrence of a pair of CV terms in the dataset.



Features of An Analysis Tool

- Filter associations above a confidence and support threshold.
 - Given some GO term or MeSH term, present all the associations of that terms that are significant with respect to a **threshold** selected by the user.
- Group associations using either a GO term or a MeSH term.
 - So that users can analyze groups of associations rather than individual associations.
- Group the significant associations based on semantic knowledge.
 - An example is the semantic type associated with the MeSH terms.

Demo

<http://www.umiacs.umd.edu/~adamlee/lslink/lodgui/index.html>

Locate Association with the Highest LOD Based Confidence Score



Finding Meaningful Associations

1. Please observe the [diagram of the background dataset](#).

- Entrez Gene is annotated with the GO controlled vocabulary.
- PubMed is annotated with the MeSH controlled vocabulary.

2. Please select a search term: Search

3. Please select a Controlled Vocabulary type:

Please select a Controlled Vocabulary term:

4. The statistics of the LOD based confidence scores:

min: 2.5056 max: 6.4632 avg: 4.3139 med: 4.3048

MeSH Descriptor Name	LOD
Fanconi Anemia Complementation Group G Protein	6.4632
Breast Neoplasms, Male	6.3256
Fallopian Tube Neoplasms	6.2795
BRCA2 Protein	5.8844
Genes, BRCA2	5.8262
Neoplastic Syndromes, Hereditary	5.3728

Threshold: 2.0 3.0 4.0 5.0 6.0 7.0

Filter Associations with LOD Based Confidence Scores Above a Cutoff



Finding Meaningful Associations

1. Please observe the [diagram of the background dataset](#).

- Entrez Gene is annotated with the GO controlled vocabulary.
- PubMed is annotated with the MeSH controlled vocabulary.

2. Please select a search term: Search

3. Please select a Controlled Vocabulary type:

Please select a Controlled Vocabulary term:

4. The statistics of the LOD based confidence scores:

min: 2.6097 max: 4.9237 avg: 4.0518 med: 4.1173

GO Term	LOD
DNA damage response, signal transduction by p53 class mediator r...	4.9237
negative regulation of centriole replication	4.9237
negative regulation of fatty acid biosynthetic process	4.9237
gamma-tubulin ring complex	4.9099
positive regulation of DNA repair	4.8892
regulation of transcription from RNA polymerase III promoter	4.8832

Threshold: 2.0 3.0 4.0 5.0

4. above threshold



Potentially Meaningful Associations

- In the *human CFTR gene* user query dataset
 - GO term: **ATP-binding and phosphorylation-dependent chloride channel activity**
 - MeSH term: **Fimbriae Proteins**
 - Semantic type: Amino Acid, Peptide, or Protein
 - LOD based confidence score = **6.645**
- In the *breast cancer in human* user query dataset
 - GO term: **negative regulation of centriole replication**
 - MeSH term: **Fallopian Tube Neoplasms**
 - Semantic type: Neoplastic process
 - LOD based confidence score = **5.884**

Q & A

Thank you!