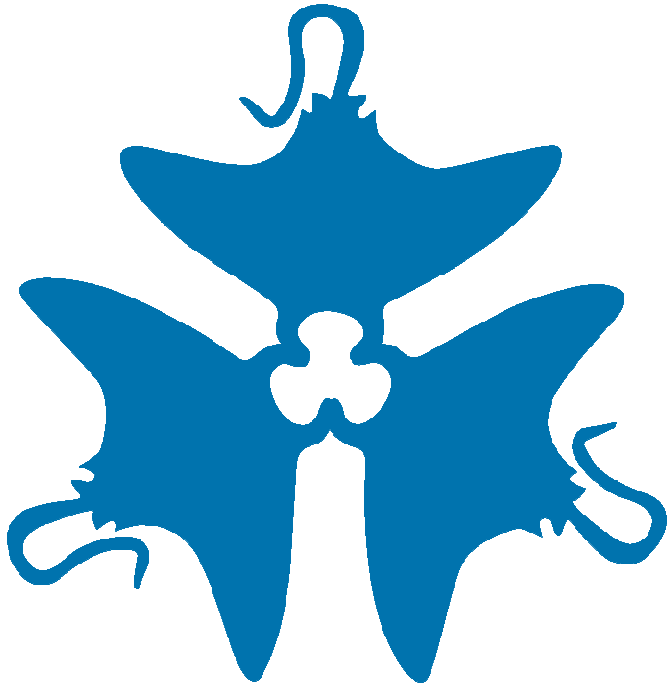


A Scale-Out RDF Molecule Store for Distributed Processing of Biomedical Data

Andrew Newman

ITEE, University of Queensland

Overview

- Motivation.
- Scale-Out Architecture.
- RDF Molecules and Extensions.
- Ontology Development, Integration and Model.
- Results.

Motivation

- Many projects (size, scope, scale).
- Many different sizes of data (MB, GB, TB, PB).
- Large total amount of data.
- Many databases (~230 PPI databases).
- Many names (LSID, URLs, local ids).
- Many different semantics (text, vocabulary, data models, ontologies).
- Variety of quality (missing data, incorrect, manually/automatically created).
- Varying provenance (sometimes none at all).
- Changing or incomplete domain knowledge.

Motivation (continued)

- Why does Scale Matter?
  - Improved coverage as there is not much overlap between data sets.
  - Greater confidence by verifying the data and our model.
  - Feedback to improve data quality.
  - Leads to better queries:
    - Find all mammalian protein-protein interactions.
    - Find all interactions between 2 pathways.

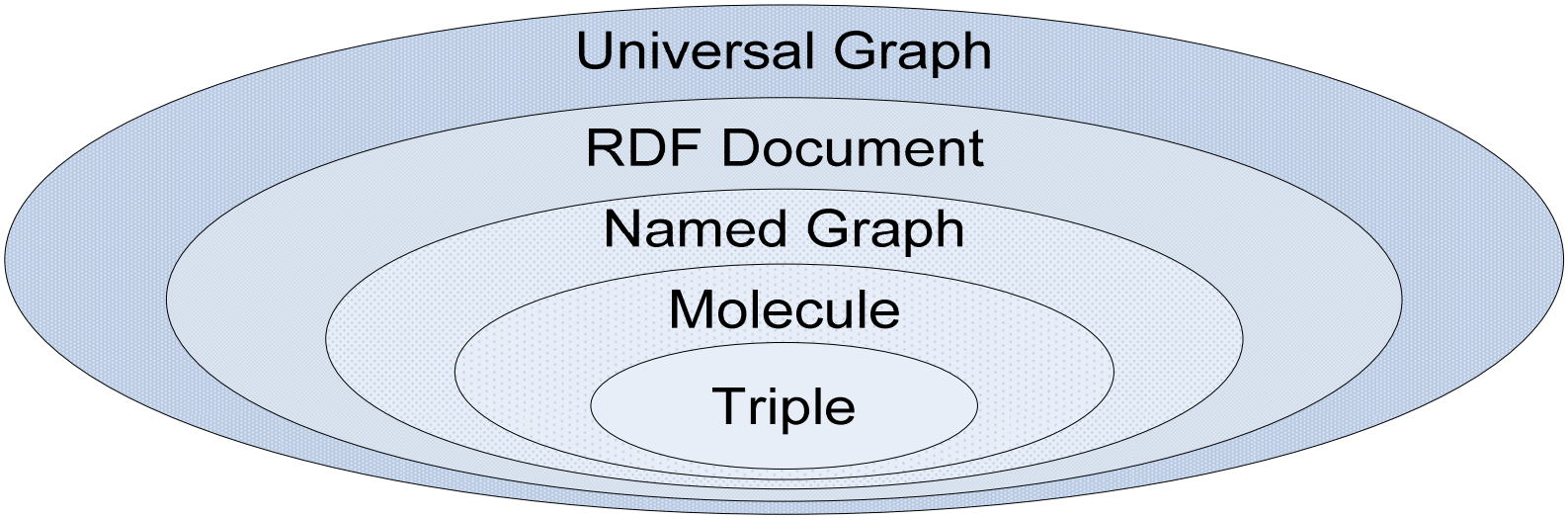
Scale-Out Architecture

- Add nodes to increase reliability, storage and processing without scaling out maintenance.
- Google
  - 10,000 Distinct MapReduce Programs.
  - 100,000 Jobs Executed/Day.
  - 20 Petabytes of Data Processed/Day.
- Nutch Search Engine, IBM, Moreira and Michael et al
  - Newton's Law beats Moore's.
  - Linear Scaling from 10 - ~2,000 nodes.
  - Same price, scale out performs 4 times better.
- "Scientific Data Management in the Coming Decade", Jim Grey et al
  - Bandwidth ≥ Latency^2.
  - Better Metadata – better selectivity of data processing.
  - Semantic Web should be used for common terminologies.
  - MapReduce – bring computation to data.

Technologies

- Hadoop
  - MapReduce.
  - HDFS (Hadoop Distributed File System).
- HBase
  - A column database built on HDFS.
- ZooKeeper
  - Distributed service co-ordination and configuration.
- Hosting
  - Local Cluster, Amazon EC2, Google (one day App Engine?).

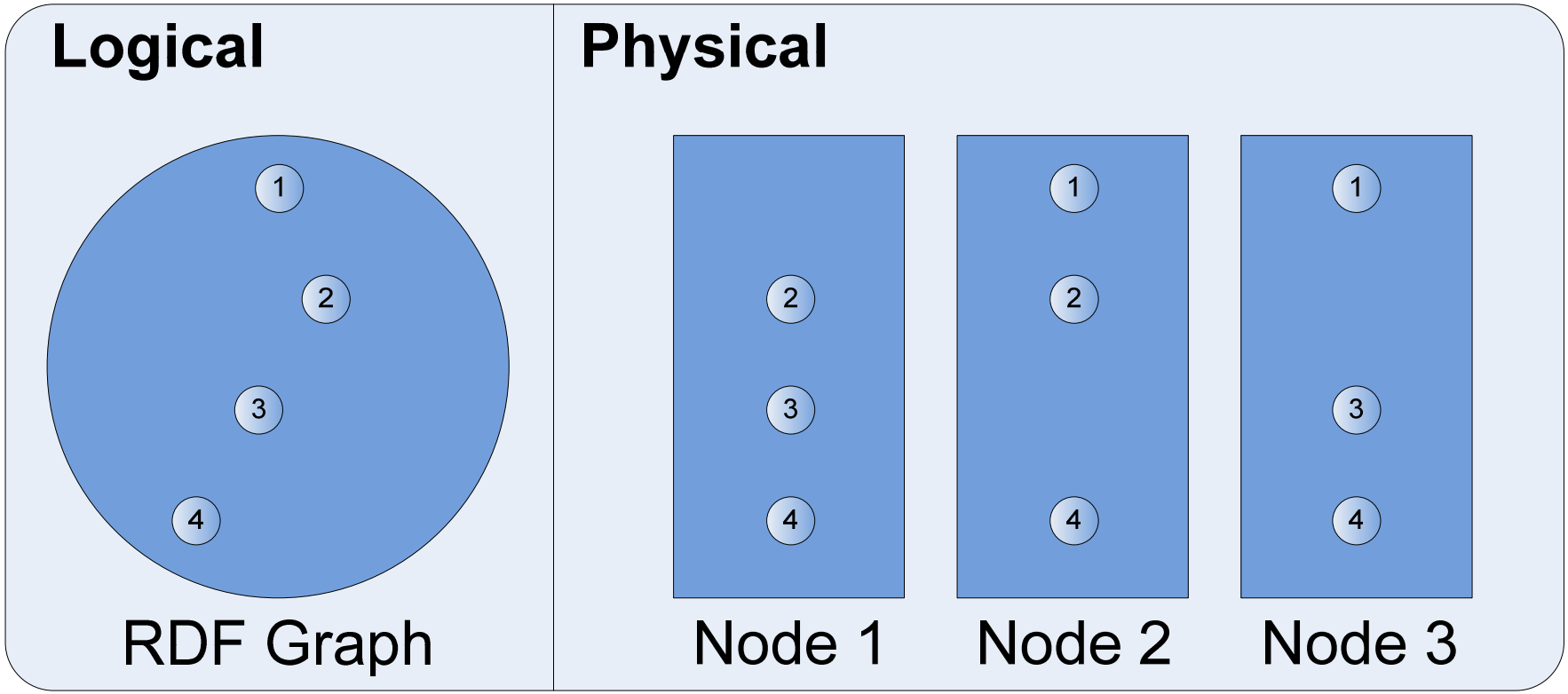
What is an RDF Molecule?



- A way to decompose an RDF Graph, containing blank nodes, into subgraphs.
- Creates context for a blank node so they are globally addressable just like URIs and Literals.

Diagram from: Ding, L., et al., “Tracking RDF Graph Provenance using RDF Molecules.”

An RDF Graph Across Computing Nodes



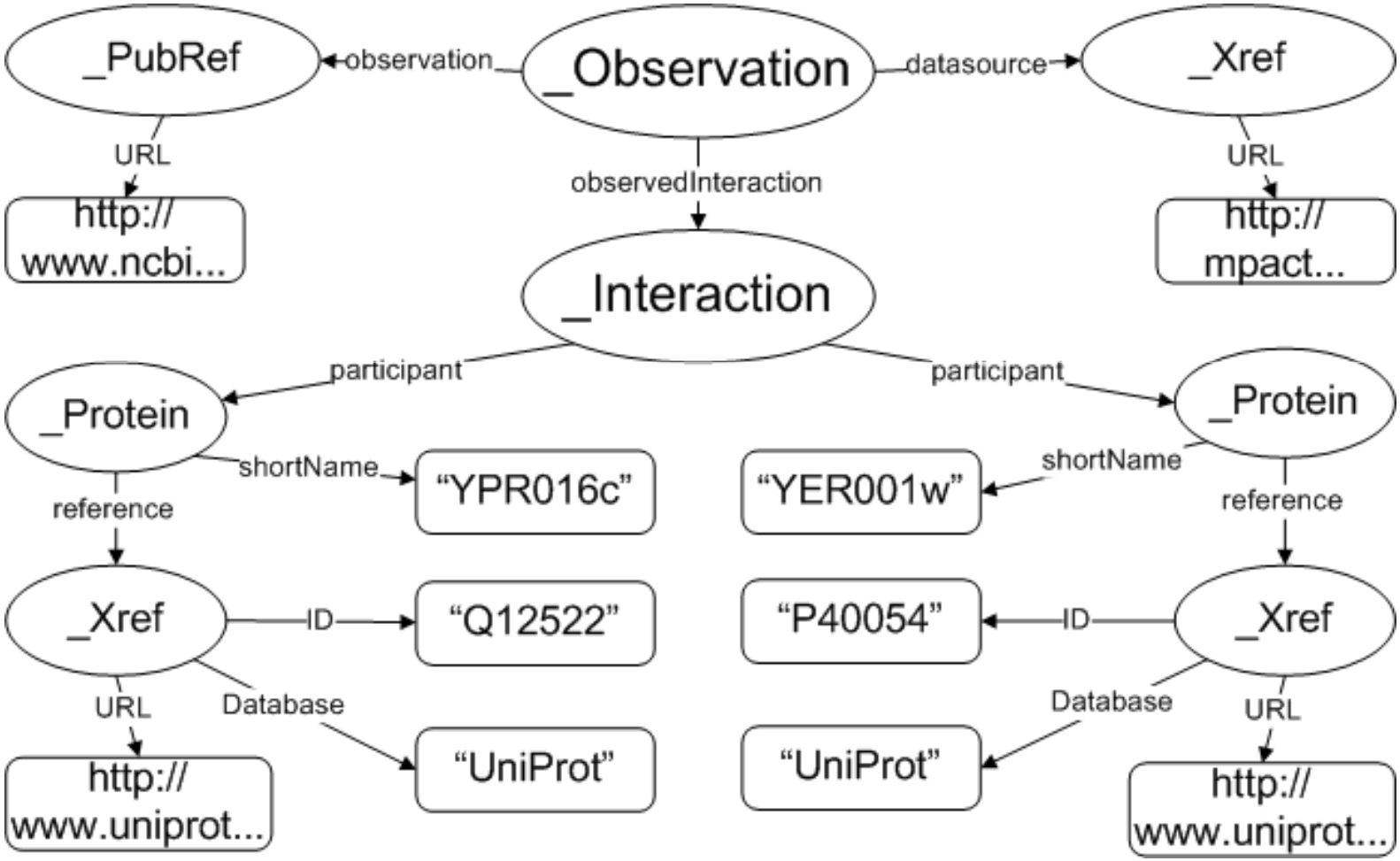
Our Extensions

- Hierarchical Structure
  - Molecules within molecules.
  - Linking Triples (_1 context1 _2, _2 context2 _3).
  - Reflects certain domain models (PPI).
- Ordering
  - By Most Grounded (head triple) to Least Grounded.
  - By String Value.
- Algorithms
  - Decomposition.
  - Merging.

Relational View of Integrated Data

"Protein"	Intact	MPact	InterPro	…
_1	ebi-25861	yjl047c	ipr011991, ipr001373	…
_2	ebi-9648		ipr000648	…
_3	ebi-3727	yer114c	ipr011993, ipr011510, ipr001849, ipr001660, ipr001452	…

Graph View of Integrated Data



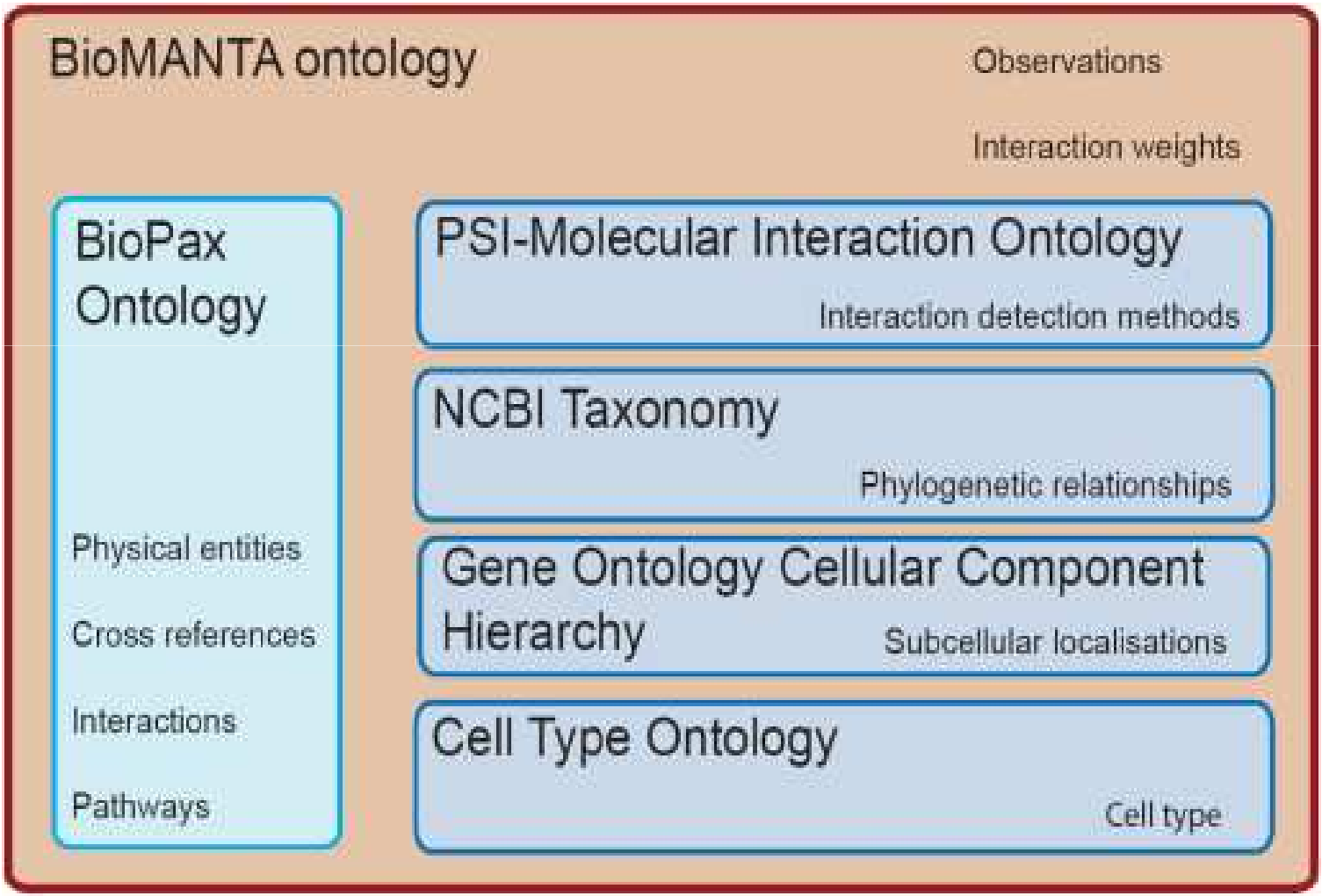
Advantages of RDF Molecules

- Lightweight context, without names.
- Distributed Processing
  - Enough context without requiring the entire graph.
  - Allows answers to be combined from many nodes.
- Conceptual Integration
  - Many names, many databases reference the same thing.
  - Find inconsistencies and remove or resolve them.
- Structural Integration
  - Lean Graph, merging removes redundant triples.
- Represents foreign key/multiple relations.

Disadvantages of RDF Molecules

- Existing RDF graphs (“local graphs”) need to be converted to molecule based graphs (“global graphs”).
- Costs
  - Extra Join.
  - Redundancy Removal.
- General Problems
  - Agree on structure and rewrite existing code.
  - Lack of Blank Node Round Tripping in SPARQL requires subqueries or API usage.

The Ontology



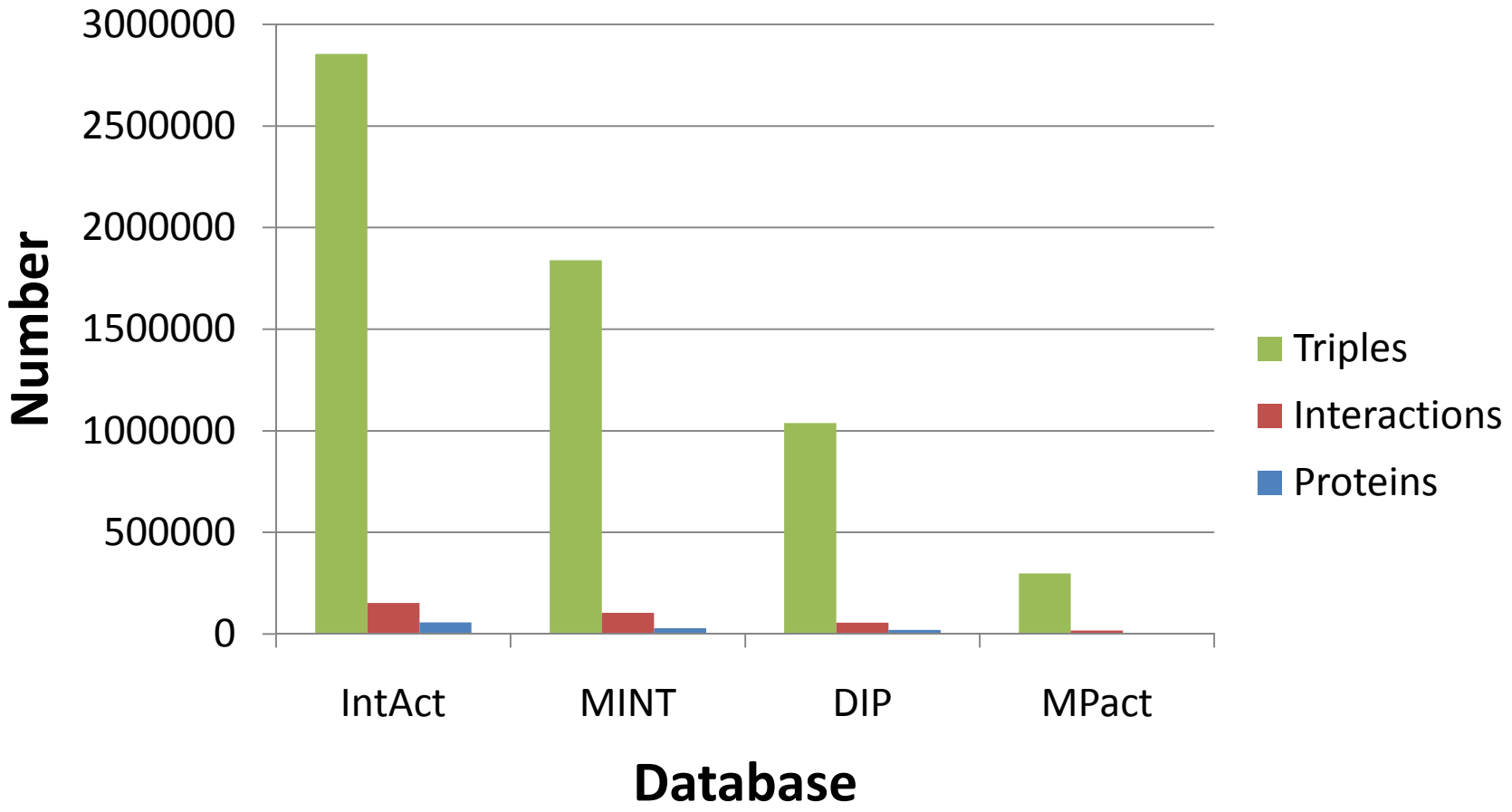
BioMANTA Extensions

- Instances of classes e.g. Experimental Methods from BioPAX ontology.
- DisjointClasses(Experimental Observation, Unspecified Observation, Predicted Observation, Inferred Observation)
  - Allows n-ary, multiple observations of the same interaction.
  - Context:
    - sourceOfData - identity of 3rd party resource.
    - observedCellType - the cell type in which the experimental observation occurred.
    - method type – the type of evidence for a particular observation type (e.g. experimentalMethod, inferenceMethod, etc).
    - subCellularLocalisation - a BioPAX entity, with a range from Gene Ontology's cellular component hierarchy.
  - Inferred Observations - from ontological (OWL) classification.
  - Predicted Observations - from data analysis or data mining.

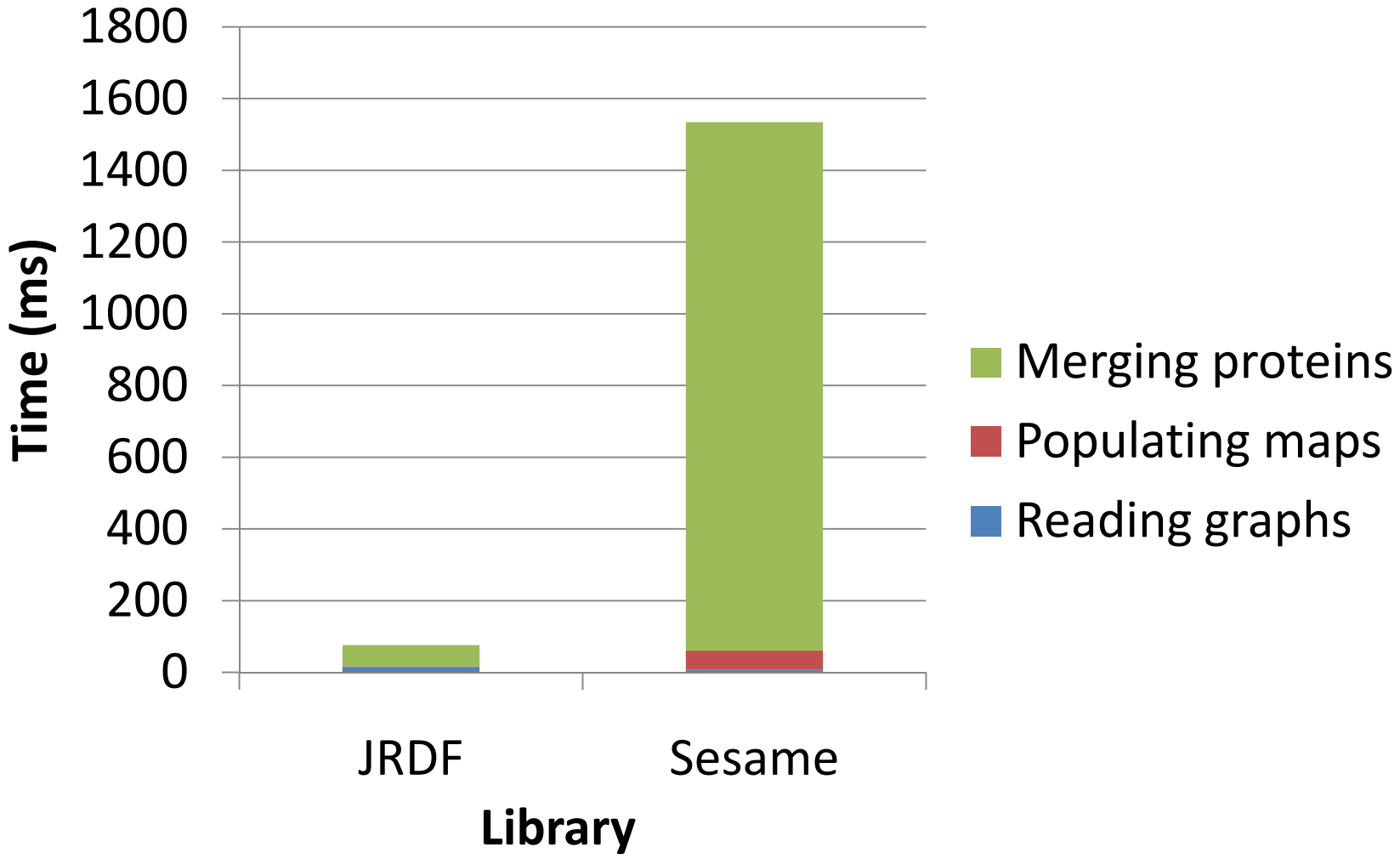
Integration Process

- PSI-MI to RDF
  - XML to RDF
- Add UniProt to Local Protein IDs
  - Local ID → UniProtID
- Add Sequence to Local Protein IDs
  - Local ID → Sequence
- Protein Merging
  - Create Molecules.
  - Merge based on UniProt ID and Sequence.
  - Those with the same UniProt IDs but different Sequences are “warnings” and are to be removed.

Integrated PPI Data Sourcess



Protein Merge Performance



Interesting Dataset Characteristics

- One DIP File: 12 450 proteins, 60 duplicate pairs of proteins (~0.5%).
- IntAct and DIP have multiple IDs per UniProt ID.
- DIP, IntAct, MINT: 13 430 proteins, 290 Merged (~2%), 10 differed (MINT).
- Two IntAct Yeast Files:

	Yeast 1	Yeast 2	Processed	Removed
No. triples	27582	50267	77849	7206 (~9%)
No. proteins	503	893	1396	85 (~6 %)

Conclusions

- Scale-out architecture provides improved performance and reliability but demands restricted programming interfaces and data structures.
- RDF Molecules provide a way to do distributed processing over RDF sub-graphs.
- Our model utilizes RDF Molecules to integrate disparate datasets and produce a large amount of easily extensible provenance data.

Acknowledgements

Chris Bouton
Victor Farutin
Mike Schaffer
Fred Jerva

Computation Sciences Center of Emphasis,
Pfizer Global Research and Development,
Pfizer Inc.

Kevin Burrage
Jane Hunter
Mark Ragan
Melissa Davis
Yuan-Fang Li
Shoaib Sehgal

School of ITEE and
Institute of Molecule Bioscience
ARC Centre of Excellence Bioinformatics,
The University of Queensland.

Links

Web Site

- http://biomanta.org/

Results

- http://biomanta.org/downloads/

JRDF

- http://jrdf.sf.net/