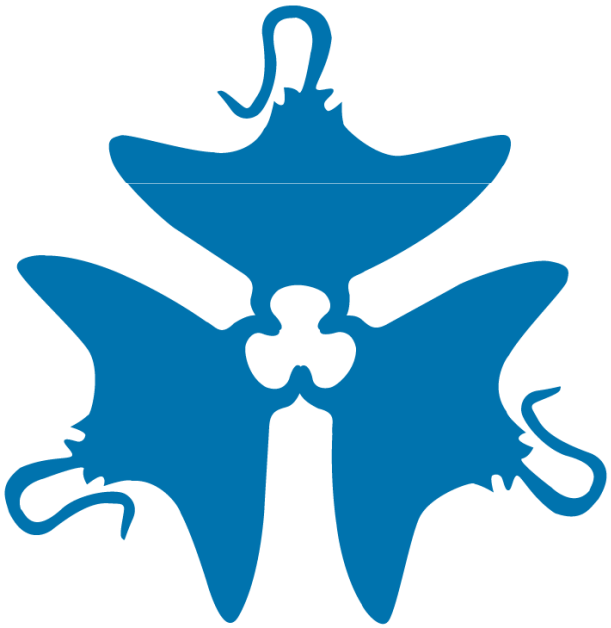
BioMANTA Project Overview



BioMANTA

BioMANTA

"The Modeling and Analysis of Biological Network Activity (BioMANTA) Project is proposed as a two- year scientific research collaboration between the Molecular Informatics group at the Pfizer Research Technology Center (RTC), Cambridge, Massachusetts, USA and the Institute for Molecular Bioscience (IMB), The University of Queensland."

Our Aims (within eResearch):

- Tools for ***in silico*** drug discovery
- Development of ontologies for knowledge representation
- Knowledge discovery through inference across integrated datasets

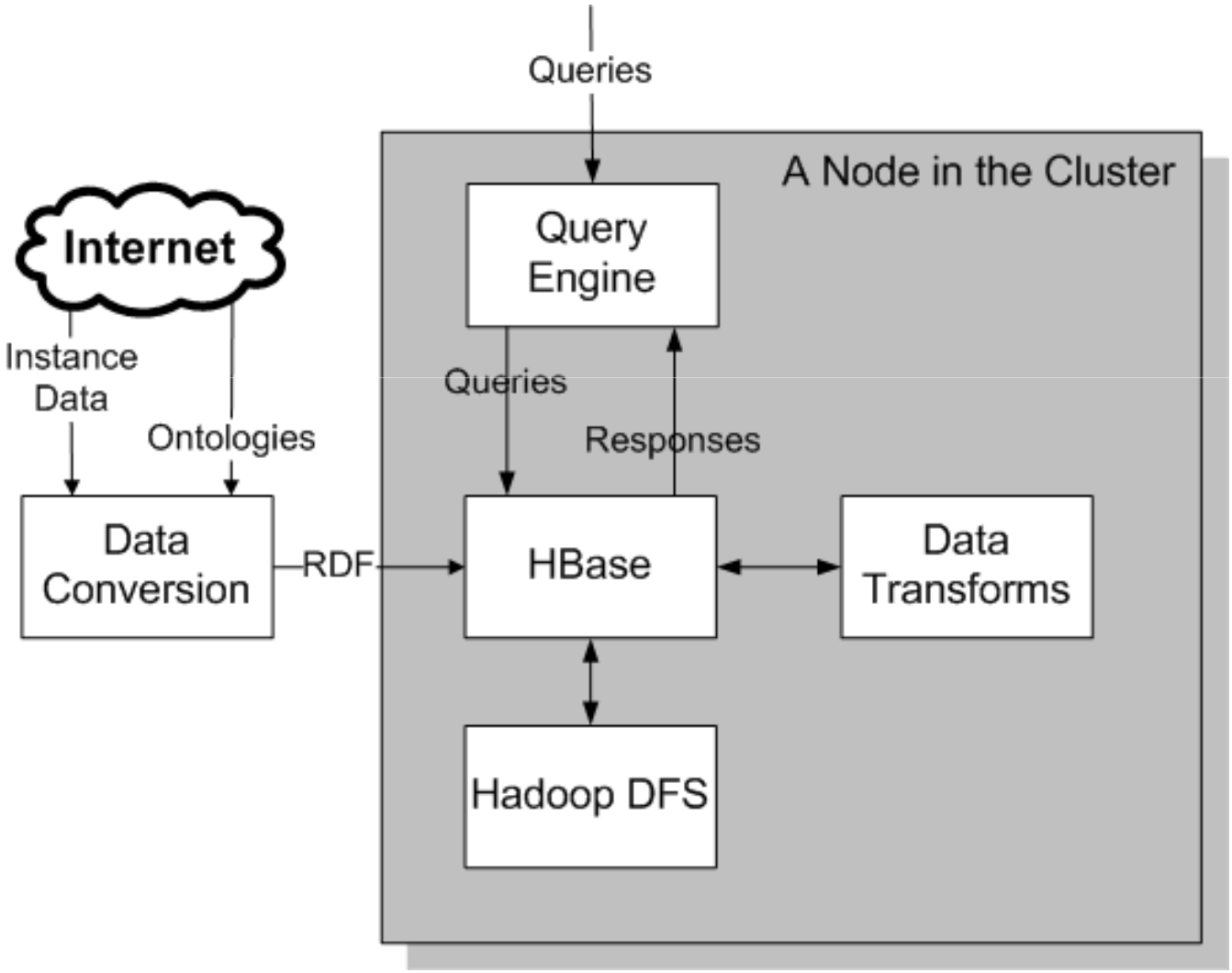
Biological Data

- Huge variety in project sizes.
- Very large data sets (TBs and PBs).
- Computationally intensive.
- Different specialities.
- Different levels of semantics in technologies used.
- Mostly suspect, duplicated, inapplicable, poorly and incorrectly modelled.

Technologies

- RDF
  - Graph, Relational DB for the Web, Integration.
- RDF Molecules
  - A Level between RDF statements and a graph.
- OWL
  - Modelling, reuse of terms, inferring new data.
- Cluster
  - Distributed processing of large data sets.

Architecture



Integration with Blank Nodes

- Blank Nodes mean “there is something but I don’t have a name for it yet”.
- Everything is a property off of a blank node.
- Example - Matching Musicians:
  - Musicians have first name/last name compositions, albums, etc.
  - To integrate we need to map two local identifiers.
  - Local identifiers don’t work.
  - Match by properties.
  - For some one piece is enough (only one Tchaikovsky)
  - For others you need multiple values (Bach, birthdate, compositions)

Protein Data is the Same

- Global IDs (such as LSID, BioPAX) have largely failed to gain acceptance for a variety of reasons.

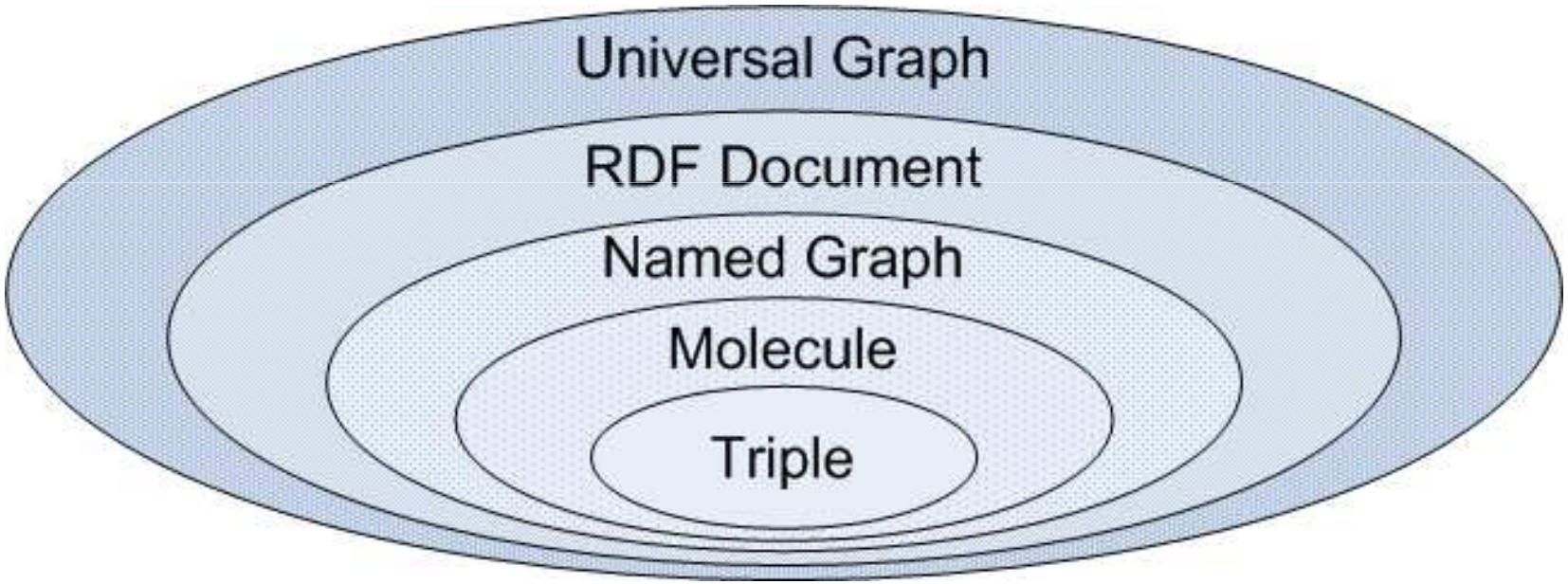
- A huge number of local IDs including:
  - MPact, DIP, IntACT, MINT.

- Properties include:
  - Sequence information
  - Species
  - Subcellular location
  - Expression, cross references, etc.

…but Blank Nodes Suck!

- Generally, finding distinct blank nodes is computationally infeasible (travelling salesman/NP hard).
- Only valid within a local context – you can't put them on a cluster of computers and know which is which blank node – they have no global identification.

Adding Context to Blank Nodes



Design of Molecules

- Create the data as molecules – keep context.
- Remove redundant information.
- By creating molecules you can transport blank nodes around with their minimal context.
- Ordering:
  - Most grounded to least grounded (_ x y is greater than _ a _),
  - Subject greater than object (y x _ is greater than _ x y),
  - URIs then Literals,
  - If the same type then alphabetical.

Still Tricky…

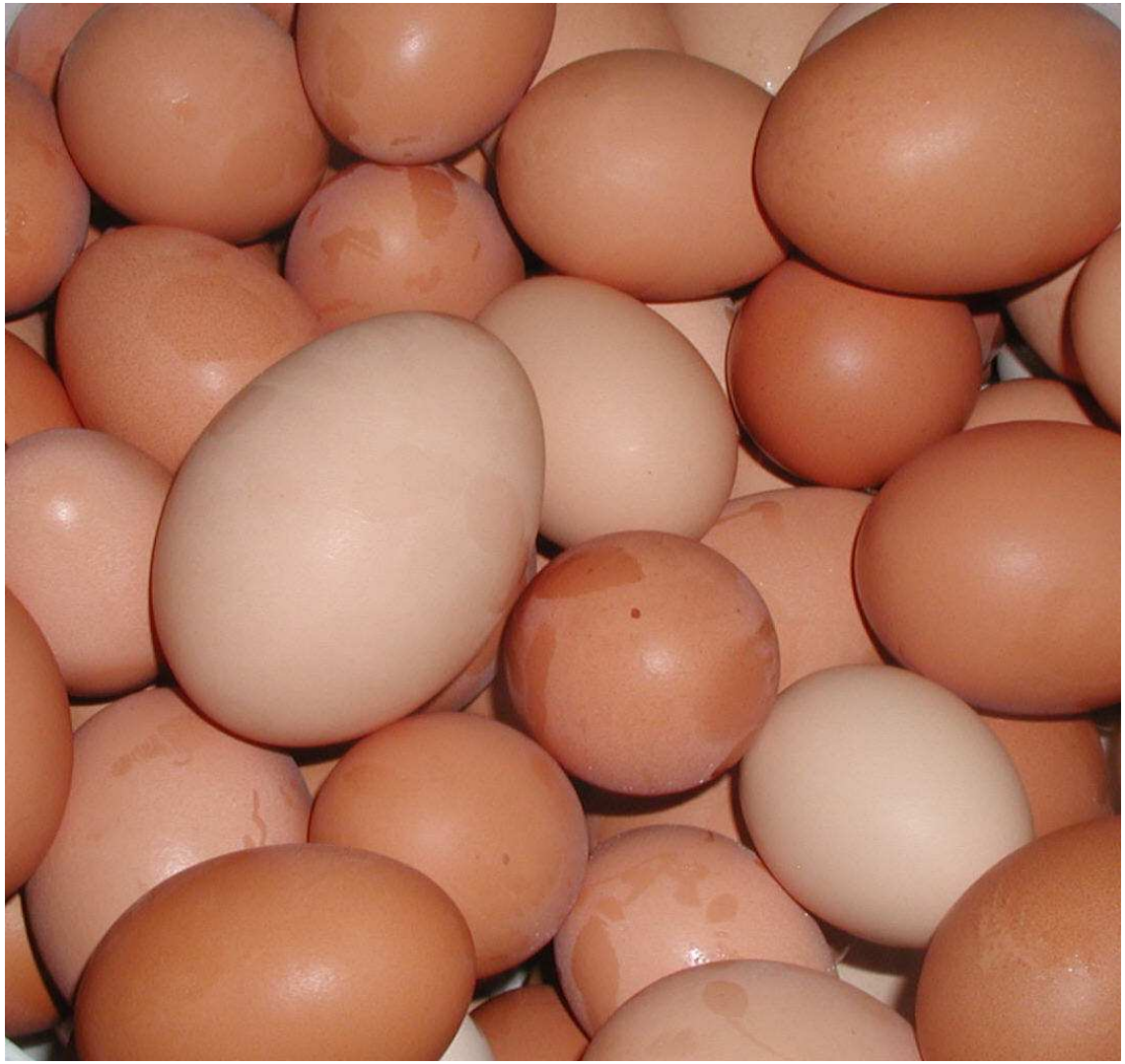
- Must model and handle adding multiple contexts:

{_1 observation _2}
{_2 interaction _3}
{_3 participant YIL33C}
{_3 participant YDR395W}

- Lots of use cases to do with adding molecules and when to combine them.
- Expensive to add, faster to query, less false positives.

It’s Hard to Unscramble an Egg

Solution



Questions?