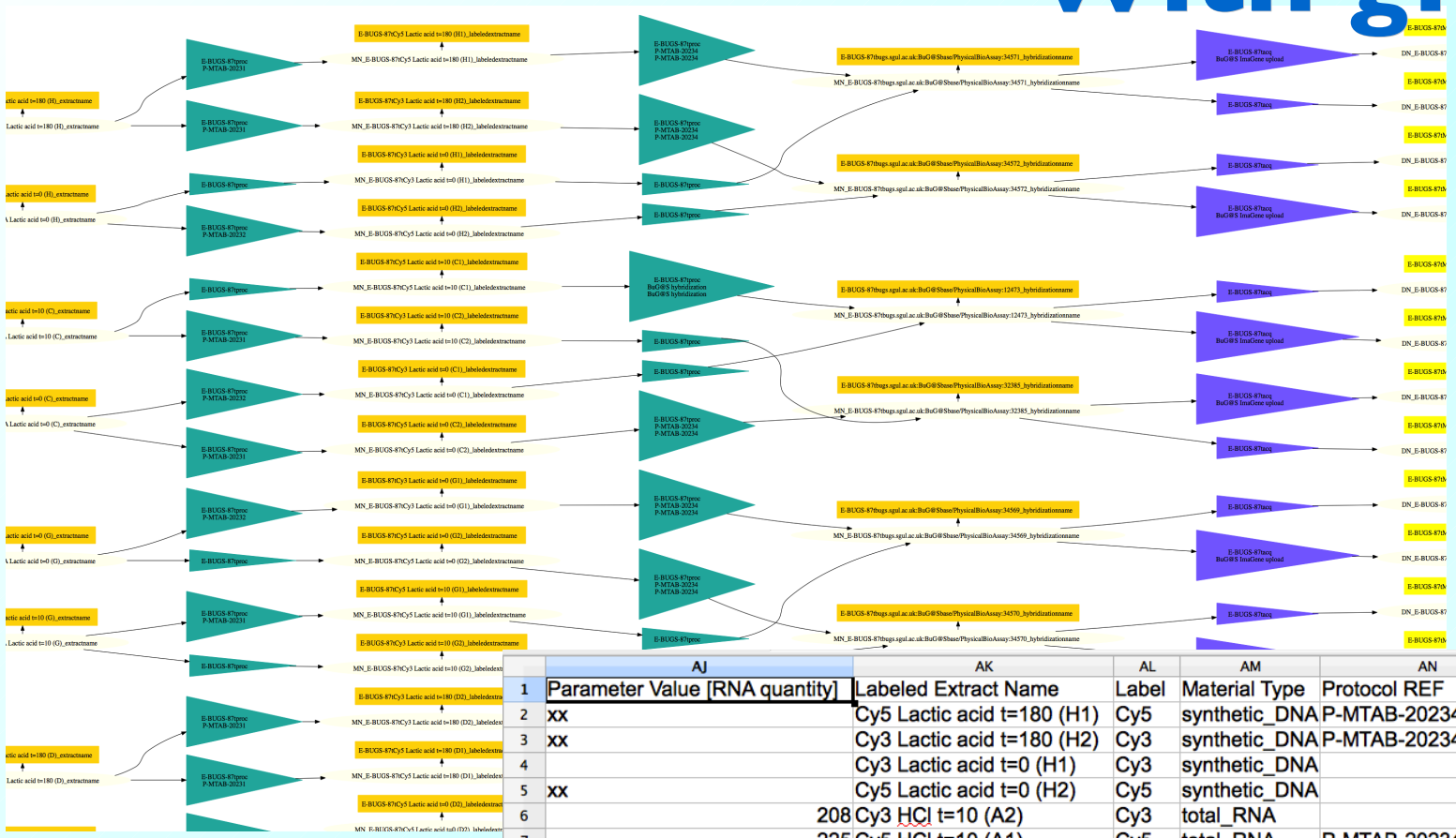
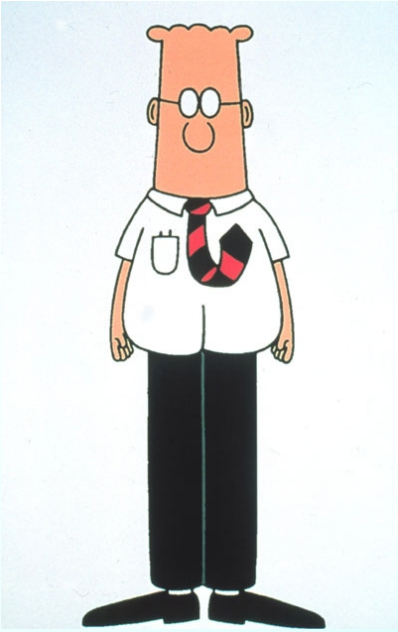
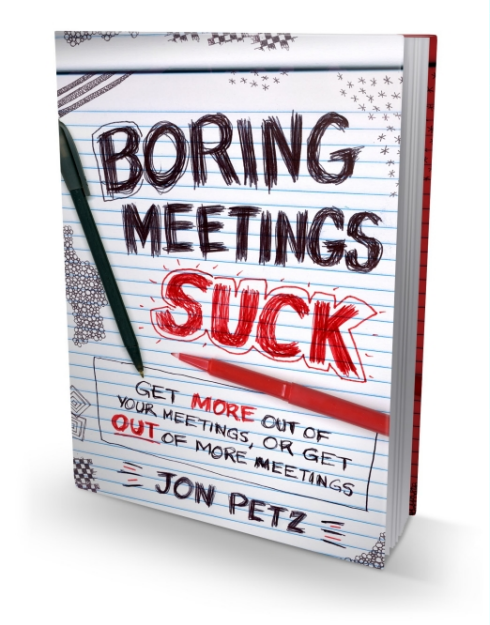
Doing Graph->Table conversions With graph2tab



	AJ	AK	AL	AM	AN	AO
1	Parameter Value [RNA quantity]	Labeled Extract Name	Label	Material Type	Protocol REF	Parameter Value [hybridization temperature]
2	xx	Cy5 Lactic acid t=180 (H1)	Cy5	synthetic_DNA	P-MTAB-20234	65
3	xx	Cy3 Lactic acid t=180 (H2)	Cy3	synthetic_DNA	P-MTAB-20234	65
4		Cy3 Lactic acid t=0 (H1)	Cy3	synthetic_DNA		
5	xx	Cy5 Lactic acid t=0 (H2)	Cy5	synthetic_DNA		
6	208	Cy3 HCl t=10 (A2)	Cy3	total_RNA		
7	225	Cy5 HCl t=10 (A1)	Cy5	total_RNA	P-MTAB-20234	65
8	xx	Cy5 HCl t=0 (A2)	Cy5	synthetic_DNA	P-MTAB-20234	65
9	xx	Cy3 HCl t=0 (A1)	Cy3	synthetic_DNA		
10		Cy5 Lactic acid t=10 (C1)	Cy5	synthetic_DNA	BuG@S hybridization	65
11	xx	Cy3 Lactic acid t=10 (C2)	Cy3	synthetic_DNA		
12	5	Cy5 Lactic acid t=0 (C2)	Cy5	synthetic_DNA	P-MTAB-20234	65
13	xx	Cy3 Lactic acid t=0 (C1)	Cy3	synthetic_DNA		
14		Cy3 Lactic acid t=10 (G2)	Cy3	synthetic_DNA		
15	xx	Cy5 Lactic acid t=10 (G1)	Cy5	synthetic_DNA		
16		Cy5 Lactic acid t=0 (G2)	Cy5	synthetic_DNA	P-MTAB-20234	65
17	xx	Cy3 Lactic acid t=0 (G1)	Cy3	synthetic_DNA	P-MTAB-20234	65
18	225	Cy5 HCl t=10 (E1)	Cy5	total_RNA		

Marco Brandizi, EBI, 17 Feb 2012

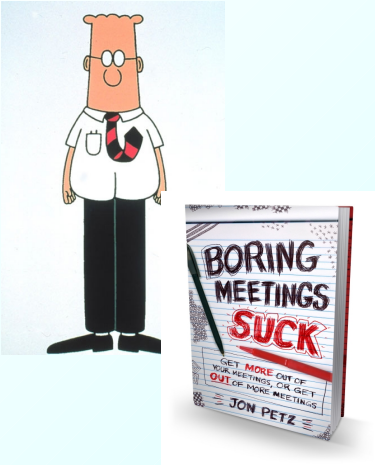




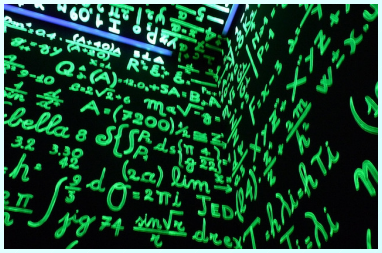
VENERDI 17

Sources: www.dilbert.com, http://www.flickr.com/photos/joao_trindade/4362414729/in/photostream/, http://www.boringmeetingssuck.com/, http://jflashman.wordpress.com/2011/06/17/venerdi-17-con-rispetto/

STOP PLEASE! I was joking!



- Mainly interesting to developers dealing with databases like AE and formats like MAGETAB
- Others may be interested things around you
  - And affected by the software presented here



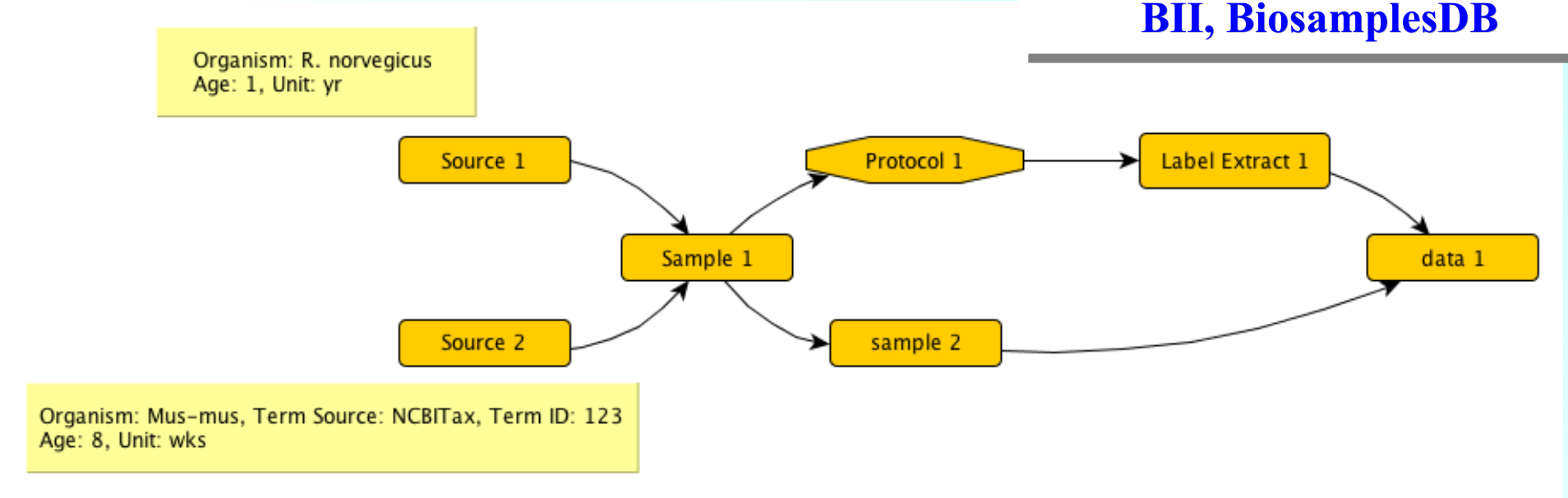
- Details are rather technical (graph theory and operative research)
- But obviously I'll keep them at a minimum here
  - I'm finalising a document for those keen on Maths

VENERDI 17

- Come on! We do science here!

What graph2tab is for

e.g., ArrayExpress, GXA BII, BiosamplesDB



Generic, you can adapt it to any model(*)/format combination

Source Name	Organism	Term ID	Term Source	Age	Unit	Sample Name	Sample Name	Protocol REF	Labeled Extract Name	Data File
Source 1	R. norvegicus			1	yr	Sample 1		Protocol 1	Lbl Extract 1	data1.xml
Source 2	Mus-mus	123	NCBI Tax	8	wks	Sample 1	Sample 2			data1.xml

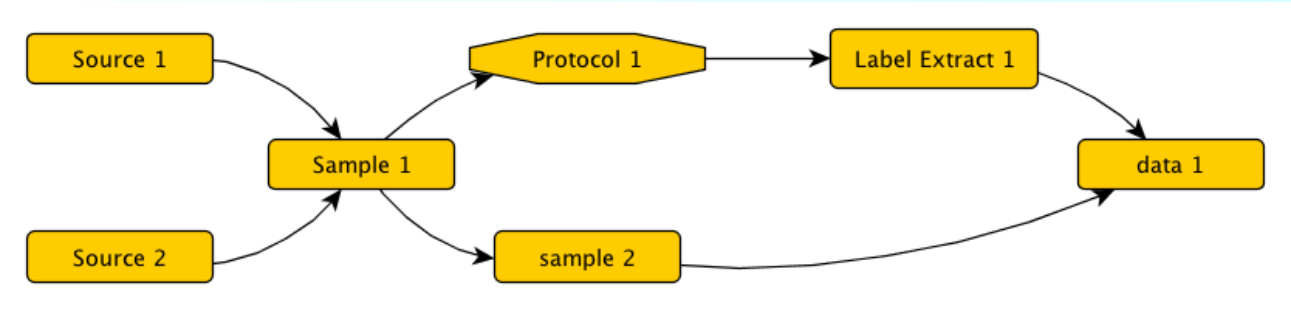
e.g., MAGETAB, ISA-Tab, SampleTAB

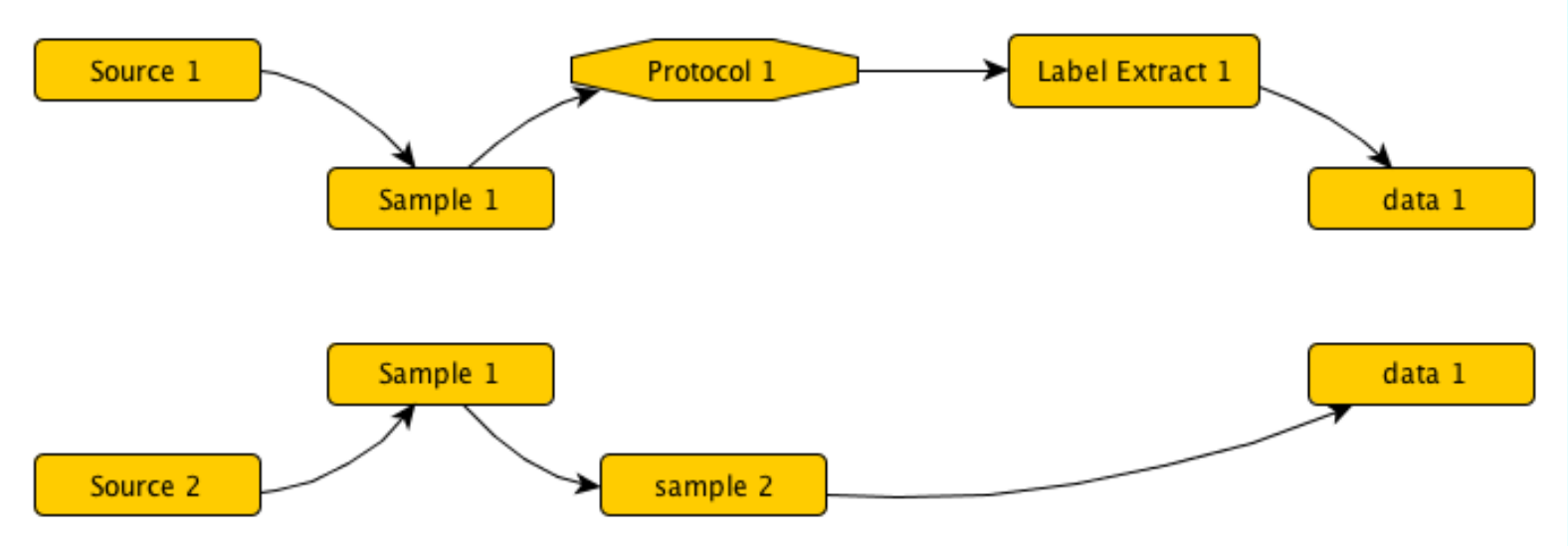
(*) as long as it's Java-encoded

Three sub-problems

- To find a suitable/optimal set of paths and rows
- To layer the graph, so that homogeneous node will go under the same set of columns when possible
- To arrange node attributes in a way that allows one to build the table from them

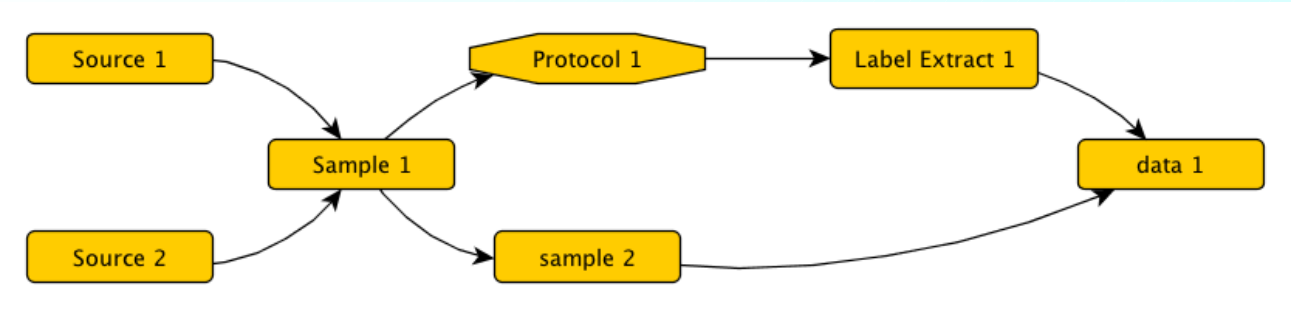
Pb 1, Basic rule: Paths = Rows



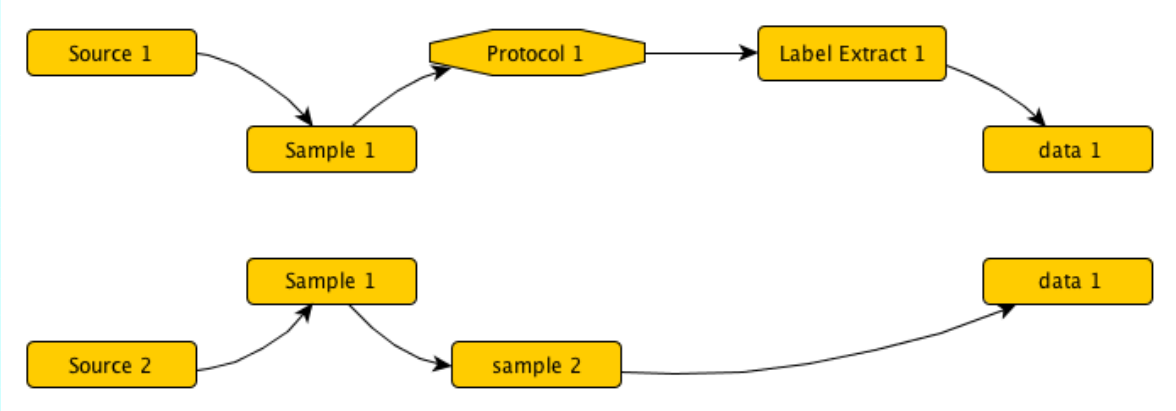


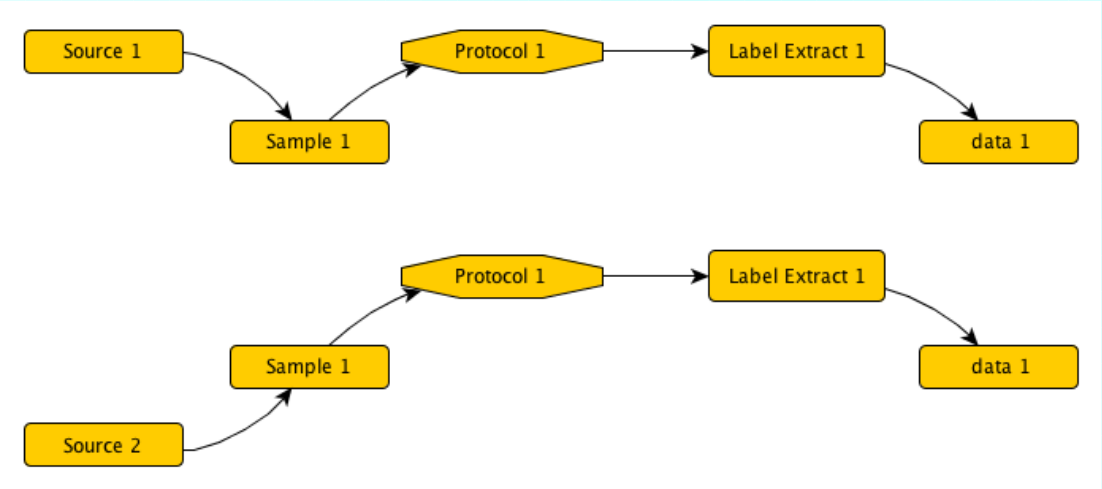
Source Name	Organism	Term ID	Term Source	Age	Unit	Sample Name	Sample Name	Protocol REF	Labeled Extract Name	Data File
Source 1	R. norvegicus			1	yr	Sample 1		Protocol 1	Lbl Extract 1	data1.xml
Source 2	Mus-mus	123	NCBI Tax	8	wks	Sample 1	Sample 2			data1.xml

Requirement #1: A covering path set



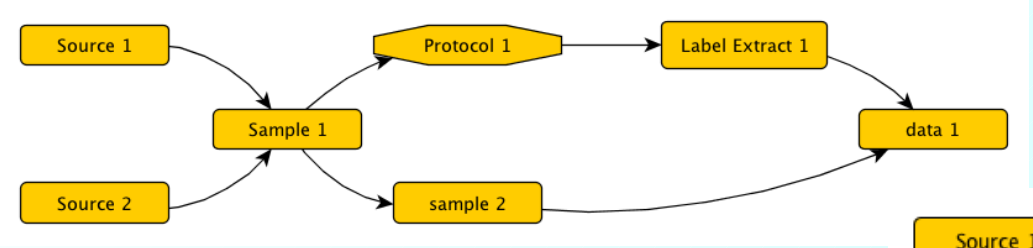
OK



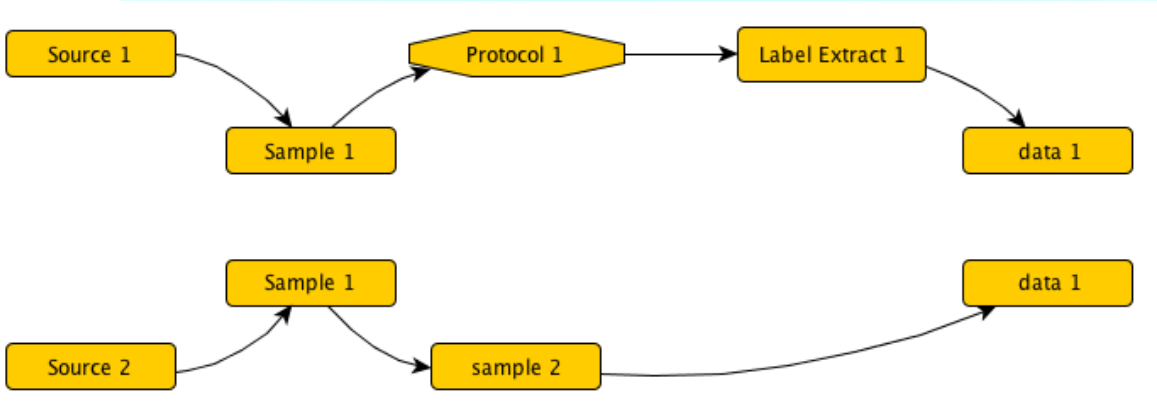


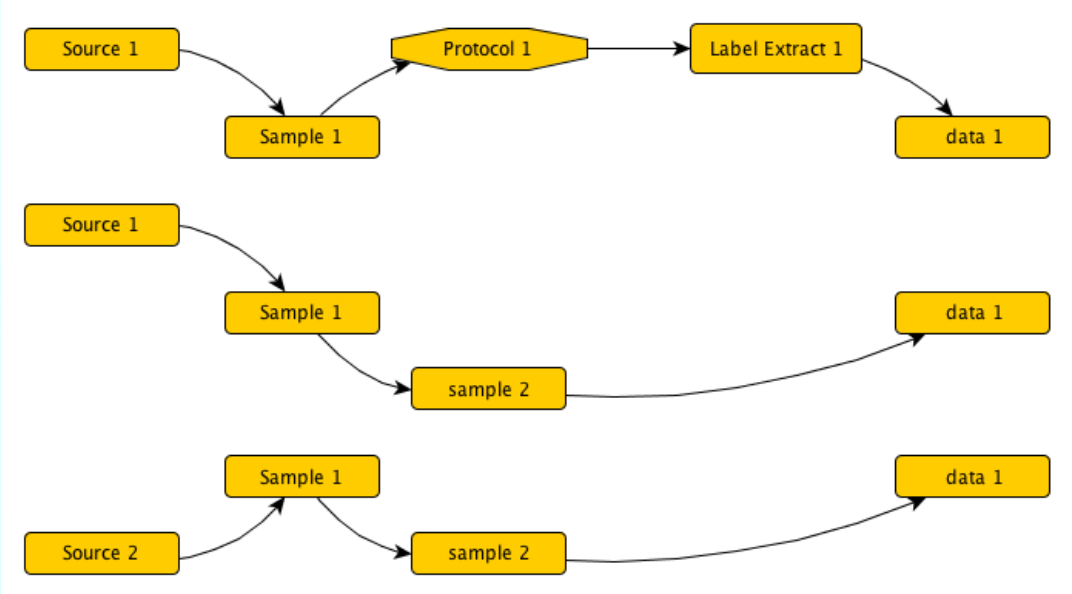
Wrong!

Requirement #2: A minimum covering path set



OK

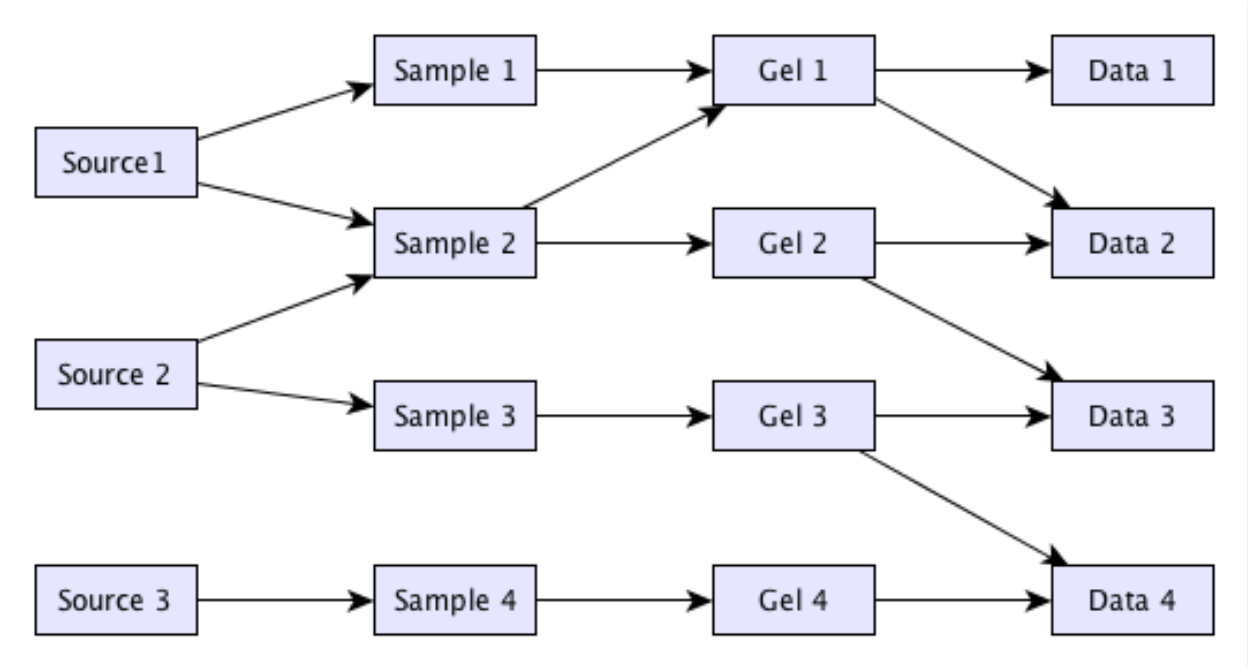




OK, but bigger!

Source Name	Organism	Term ID	Term Source	Age	Unit	Sample Name	Sample Name	Protocol REF	Labeled Extract Name	Data File
Source 1	R. norvegicus			1	yr	Sample 1		Protocol 1	Lbl Extract 1	data1.xml
Source 1	R. norvegicus			1	yr	Sample 1	Sample 2			data1.xml
Source 2	Mus-mus	123	NCBI Tax	8	wks	Sample 1	Sample 2			data1.xml

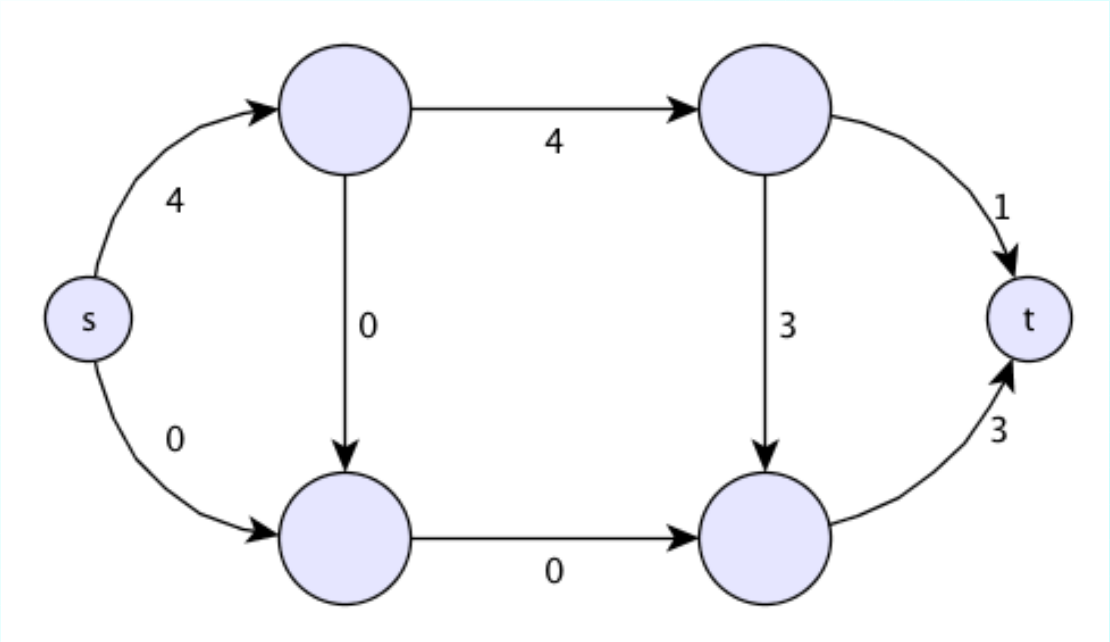
Optimisation is not so simple



Source 1	Sample 1	Gel 1	Data 1
Source 1	Sample 2	Gel 1	Data 2
Source 1	Sample 2	Gel 2	Data 2
Source 2	Sample 2	Gel 2	Data 3
Source 2	Sample 3	Gel 3	Data 3
Source 2	Sample 3	Gel 3	Data 4
Source 3	Sample 4	Gel 4	Data 4

7 rows are enough. But try to prove it and to find them!

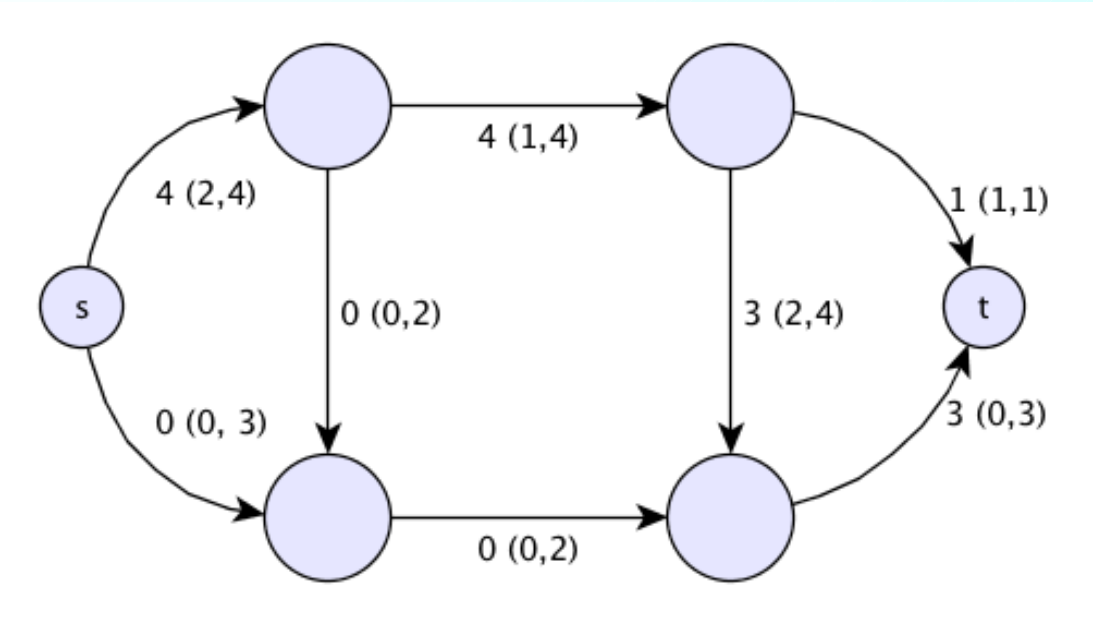
Flow Networks



$$f(N,i)-f(i,N)=\begin{cases}-v & i=s\\ 0 & i\neq s,t\\ v & i=t\end{cases}$$

Flow

Flow Networks

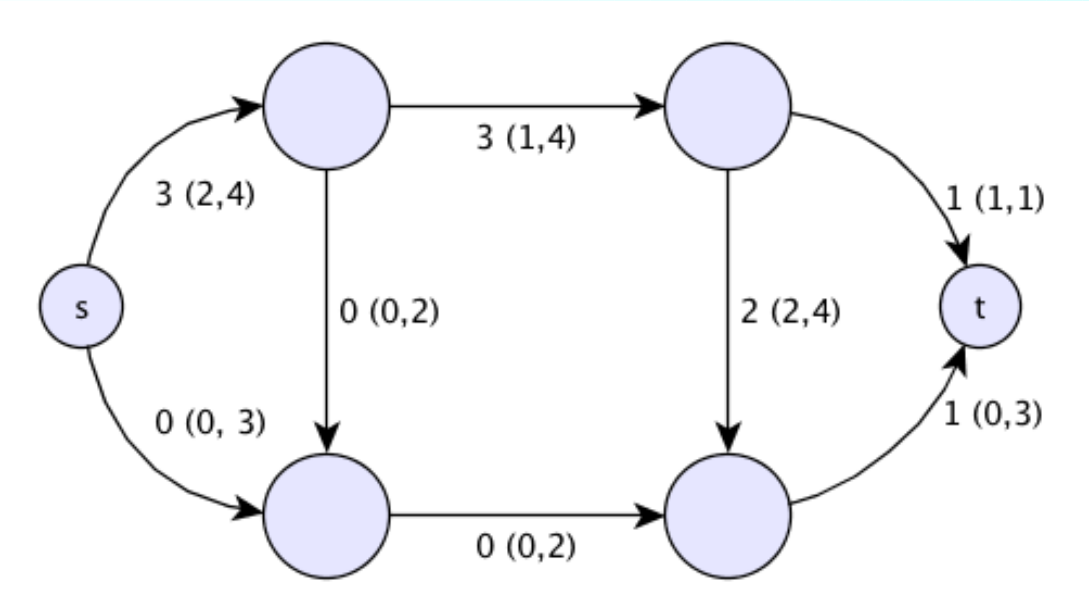


$$f(N,i)-f(i,N)=\begin{cases}-v & i=s\\ 0 & i\neq s,t\\ v & i=t\end{cases}$$

Constrained flow and admissible flow

$$l(i,j)\leq f(i,j)\leq c(i,j),\ \forall (i,j)\in A$$

Flow Networks



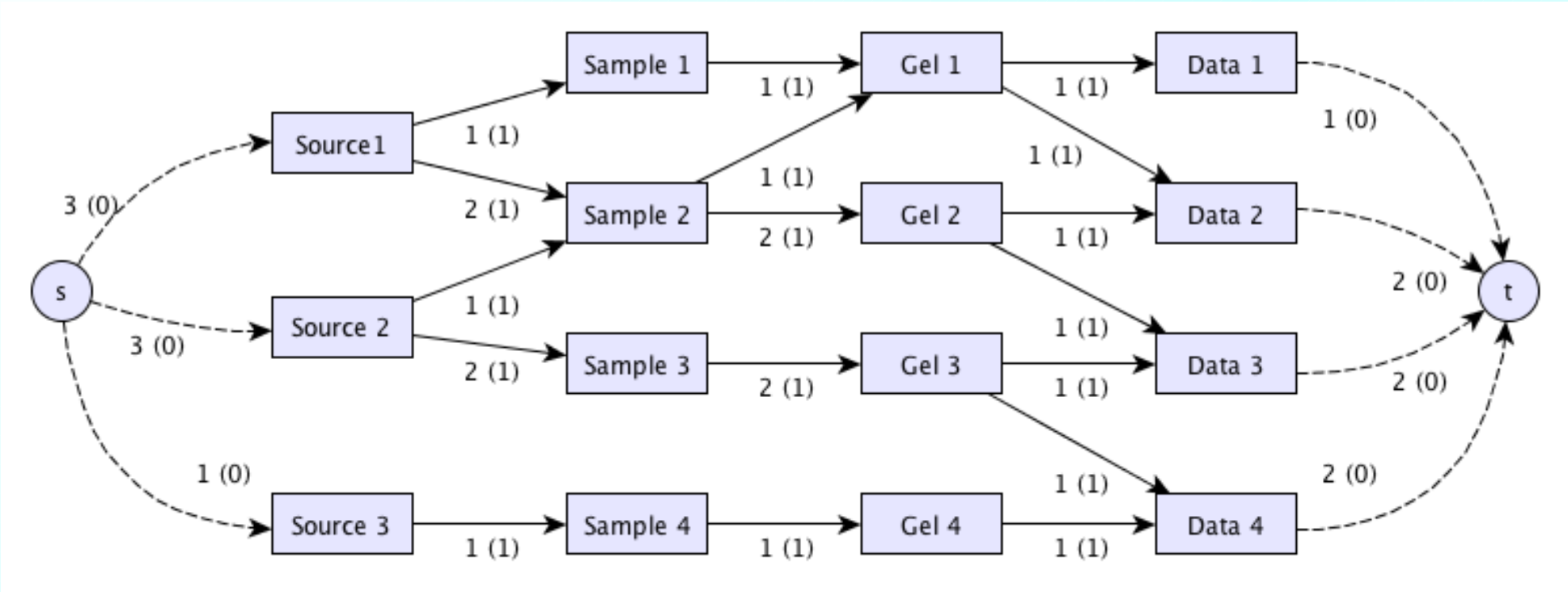
$$f(N,i) - f(i,N) = \begin{cases} -v & i = s \\ 0 & i \neq s,t \\ v & i = t \end{cases}$$

$$l(i,j) \leq f(i,j) \leq c(i,j),\ \forall (i,j) \in A$$

Minimum flow

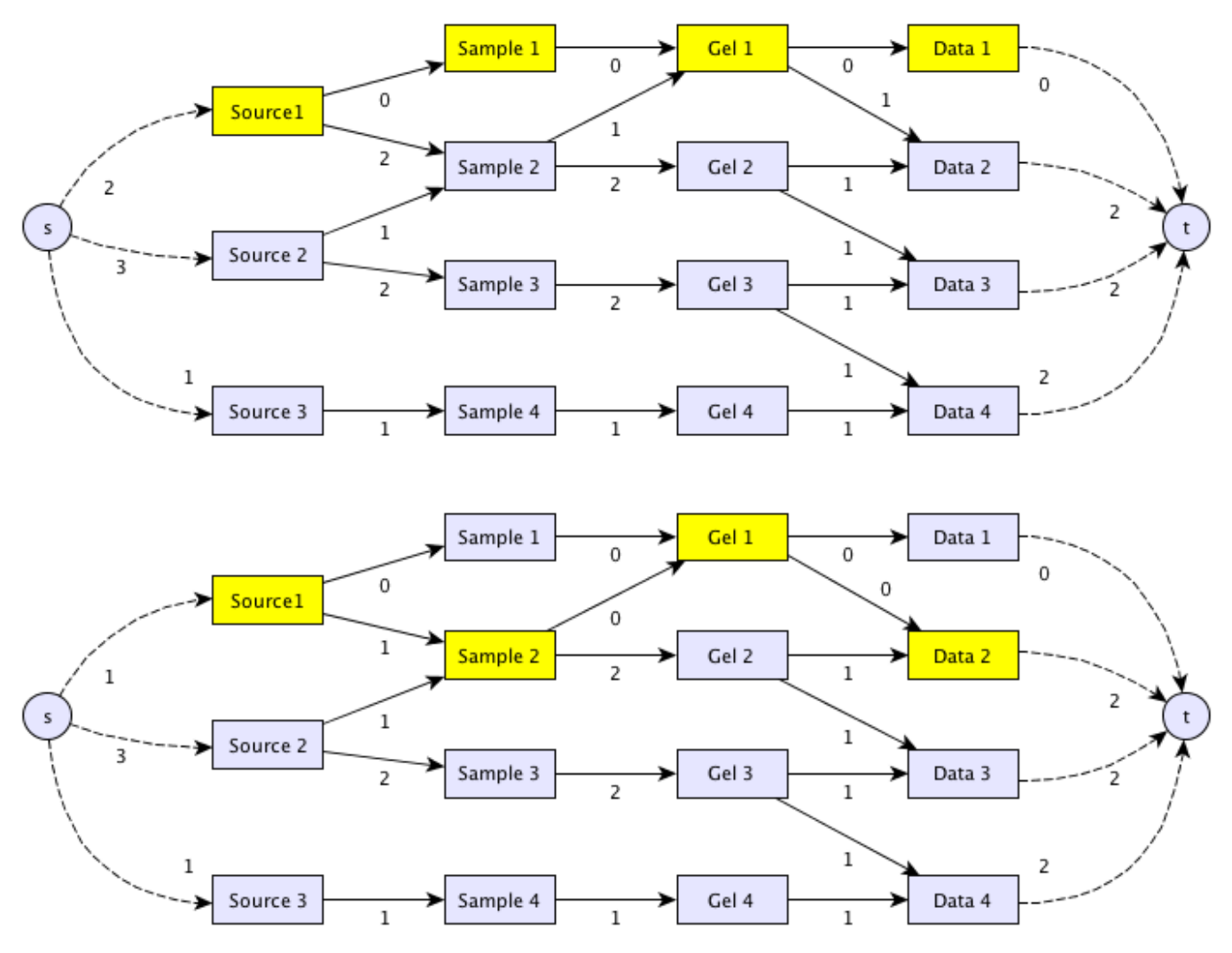
f is minimal

From Minimum Flow to Minimum Covering Path Set



A fictitious source and sink
l(virtual arcs) = 0
l(real arcs) = 1
No upper bound

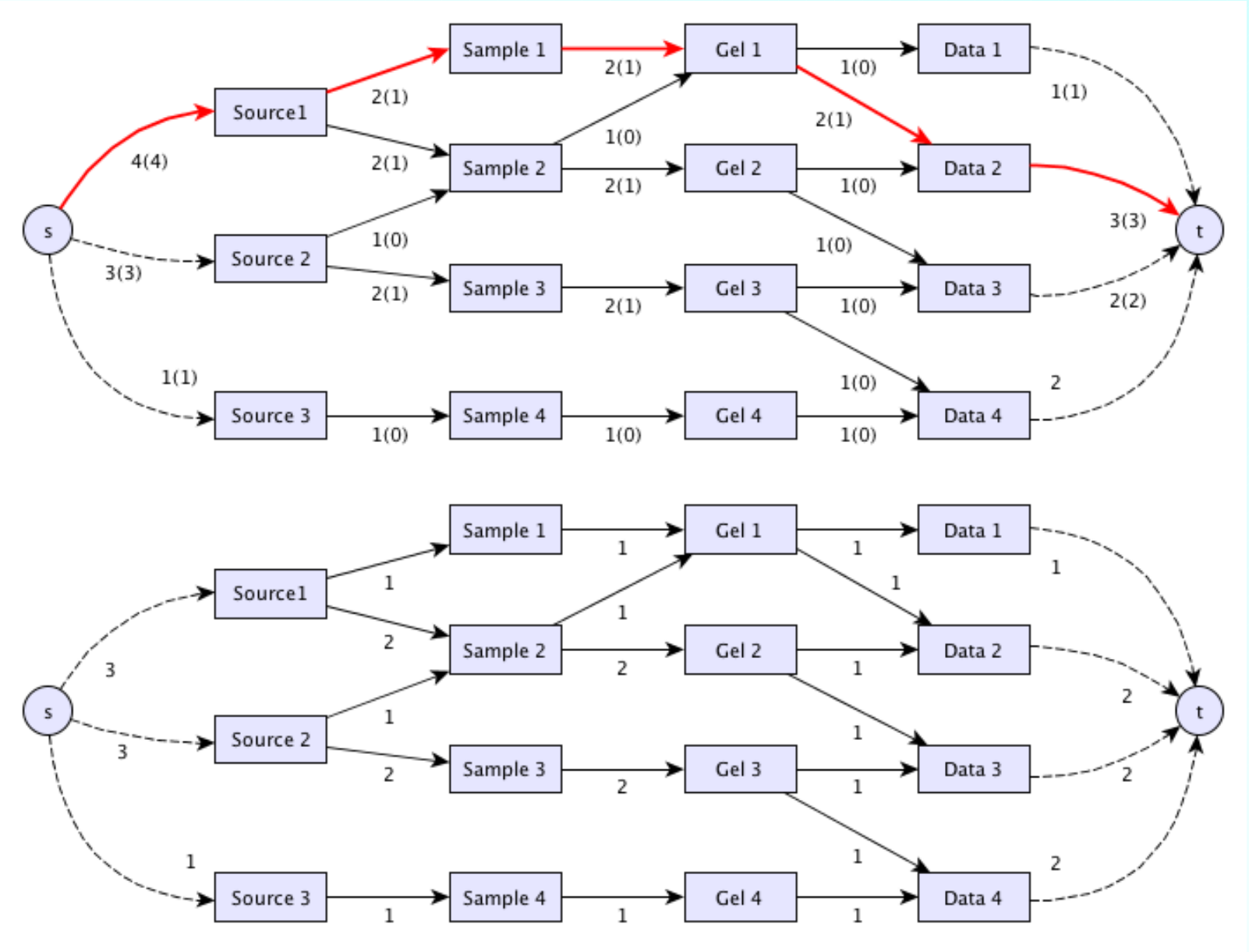
From Minimum Flow to Minimum Covering Path Set



i.e., the f(i,j) = no of times we pass through (i,j)
Or no of repetitions for (i,j)

Source 1	Sample 1	Gel 1	Data 1
Source 1	Sample 2	Gel 1	Data 2
Source 1	Sample 2	Gel 2	Data 2
Source 2	Sample 2	Gel 2	Data 3
Source 2	Sample 3	Gel 3	Data 3
Source 2	Sample 3	Gel 3	Data 4
Source 3	Sample 4	Gel 4	Data 4

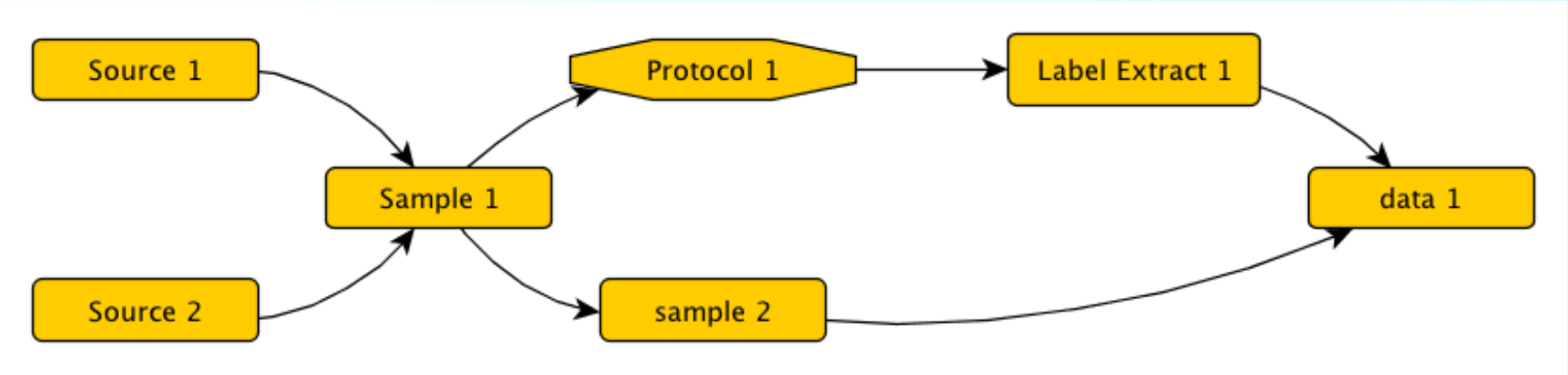
Minimum Flow: Ford-Fulkerson Algorithm

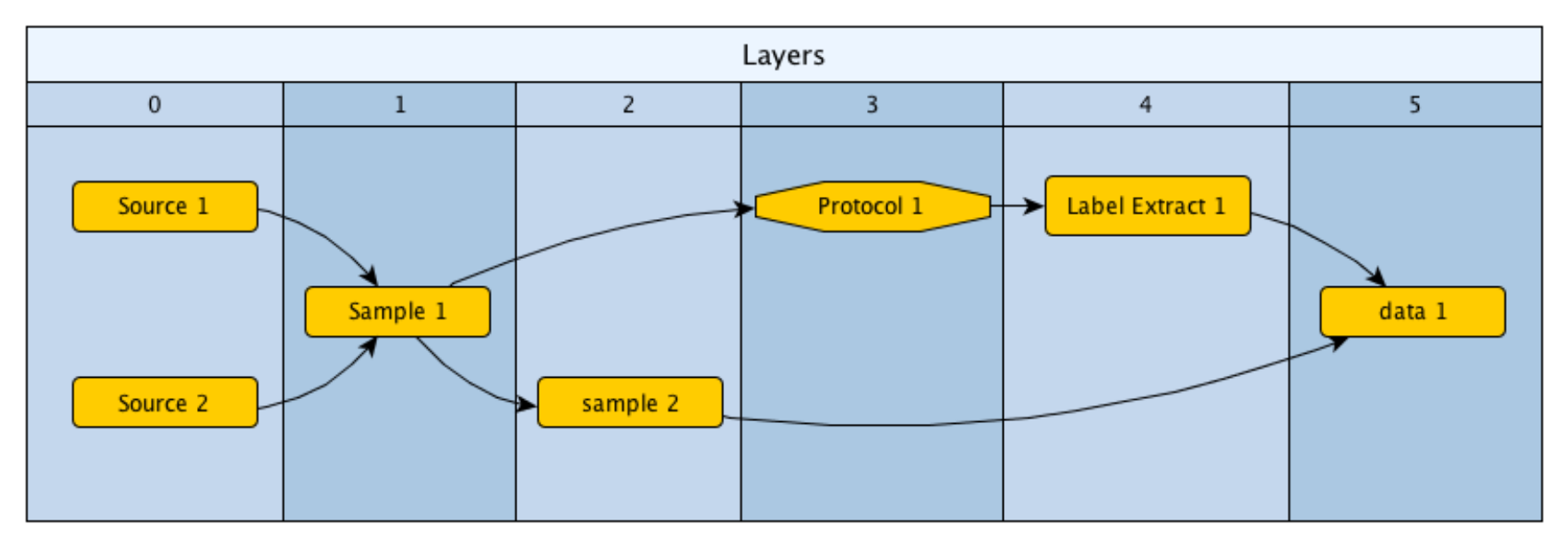


Three sub-problems

- To find a suitable/optimal set of paths and rows
- **To layer the graph, so that homogeneous node will go under the same set of columns when possible**
- To arrange node attributes in a way that allows one to build the table from them

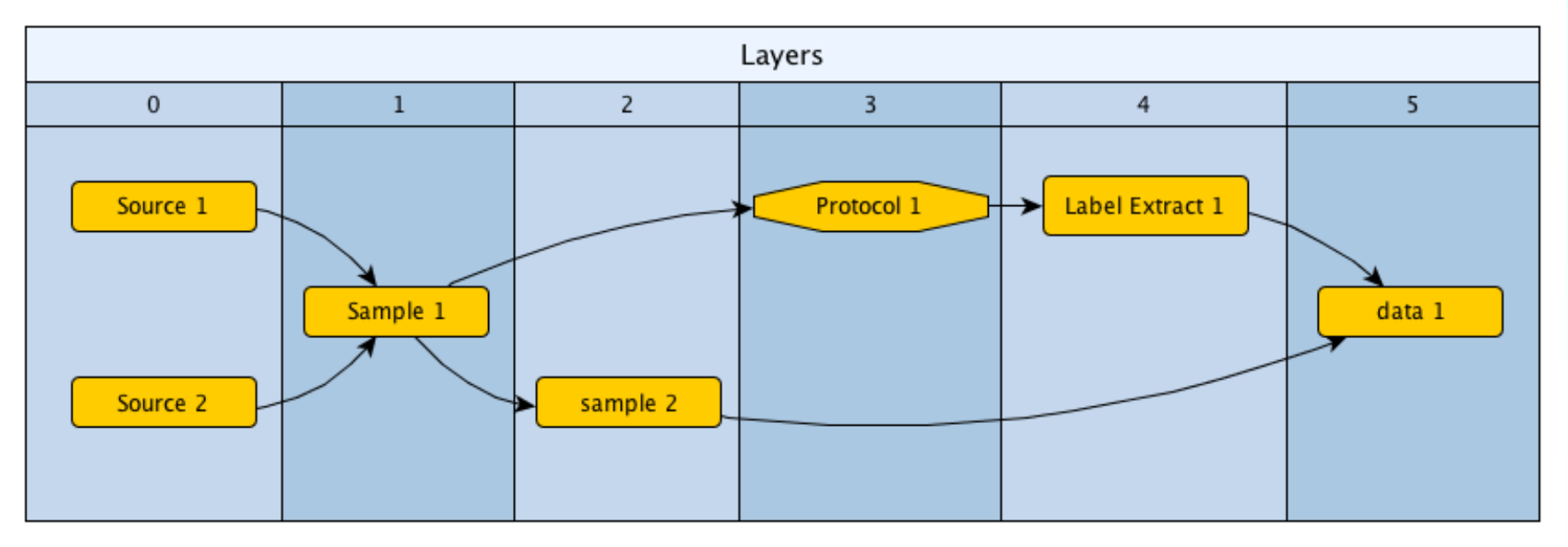
Layering


Source 1
Protocol 1
Label Extract 1
Sample 1
data 1
Source 2
sample 2

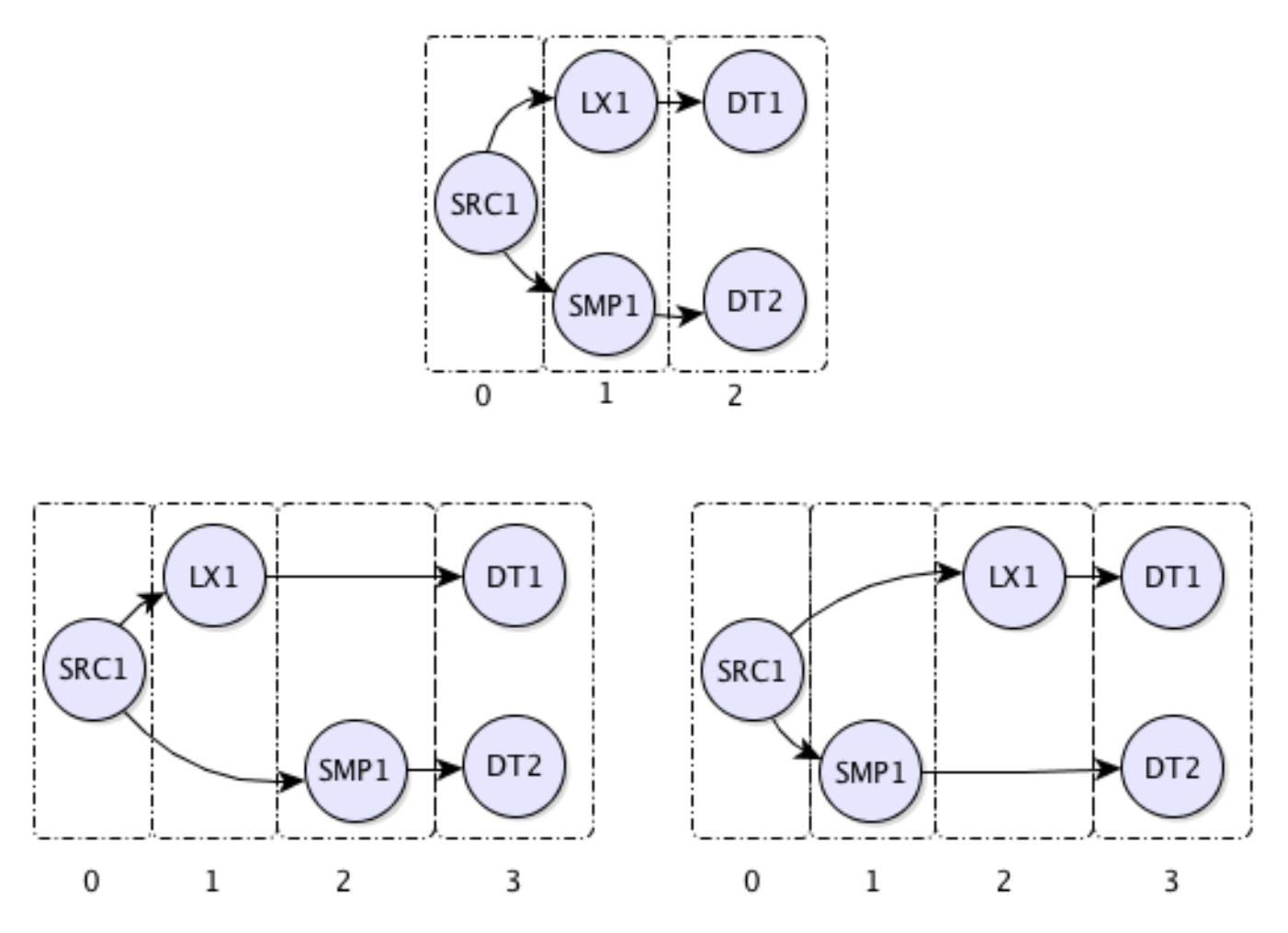

Layers
0
1
2
3
4
5
Source 1
Protocol 1
Label Extract 1
Sample 1
data 1
Source 2
sample 2

Conceptually simple

- You start by marking every node with its topological distance from the farthest connected left node
- Then for layer = 0..max
  - For each pair of nodes in the layer
    - Shift one of the two if they haven't the same type



Conceptually simple, but...



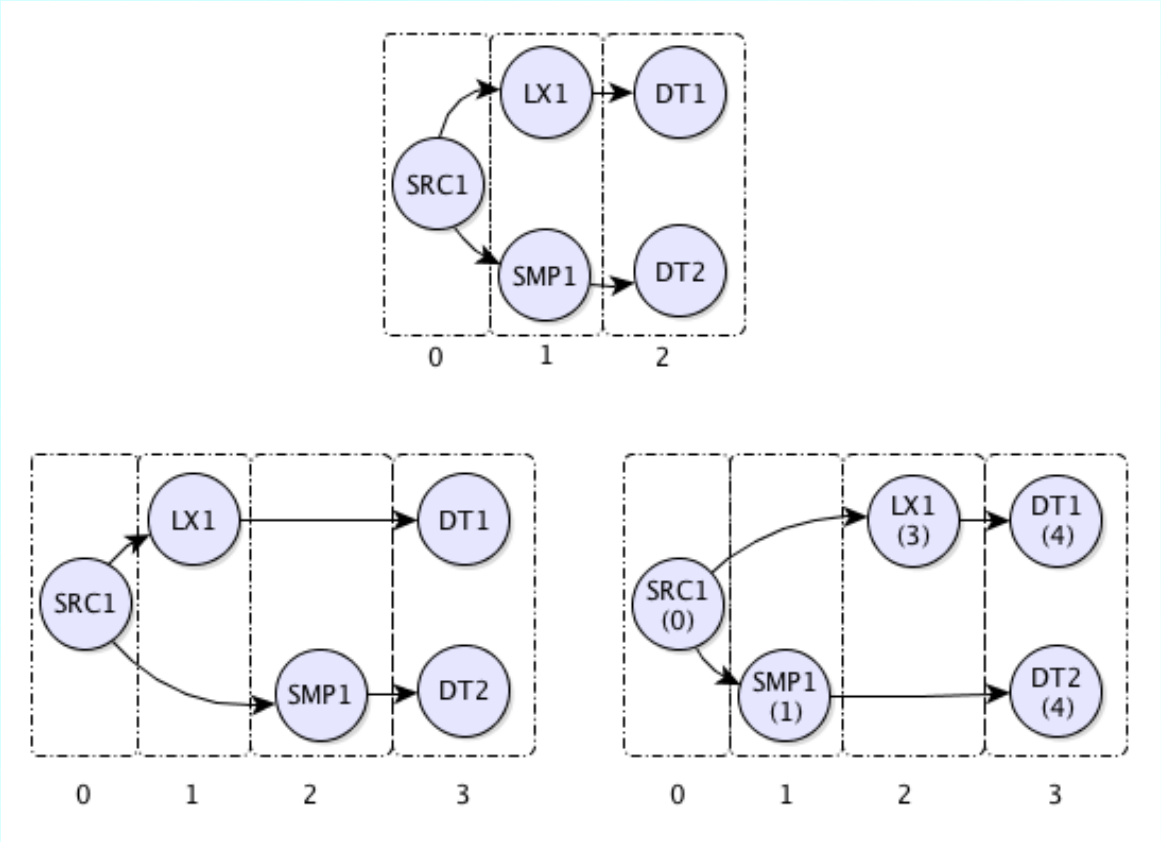
Computers understand biology even less than me!
'Source', 'Sample', 'Labeled Extract' are just symbols
Either solution is acceptable without further information

Conceptually simple, but...

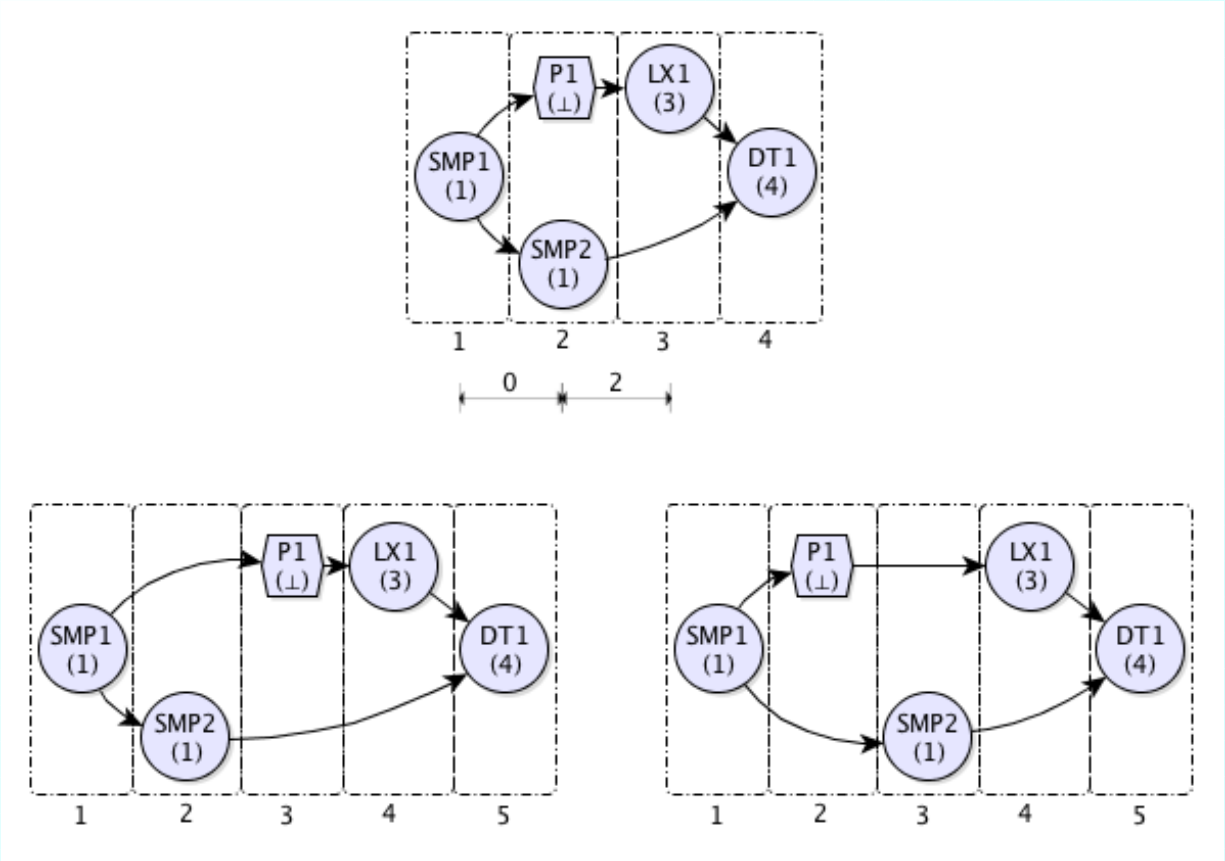
```
+interface Node
    extends Comparable<Node>
{
  getInputs (): SortedSet<Node>
  getOutputs (): SortedSet<Node>

  getTabValues (): TabValueGroup[]

  getType (): String
  getOrder (): int
}
```

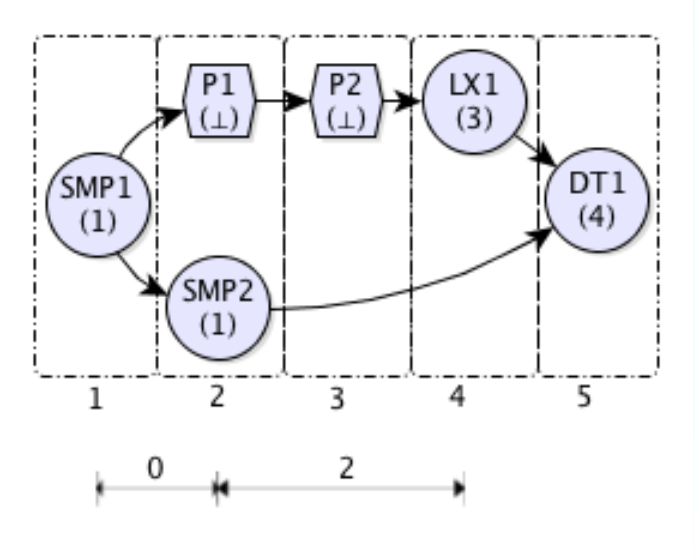
```
Map<String, Integer> TYPE_ORDER =
  new HashMap<String, Integer> ()
{{
  put ( "Source Name", 0 );
  put ( "Sample Name", 1 );
  put ( "Extract Name", 2 );
  put ( "Labeled Extract Name", 3 );
  put ( "Assay Name", 4 );
  put ( "Data File Name", 5 );
}};
```

Ah-Ah! Now I know what to choose!



More cases, same trick

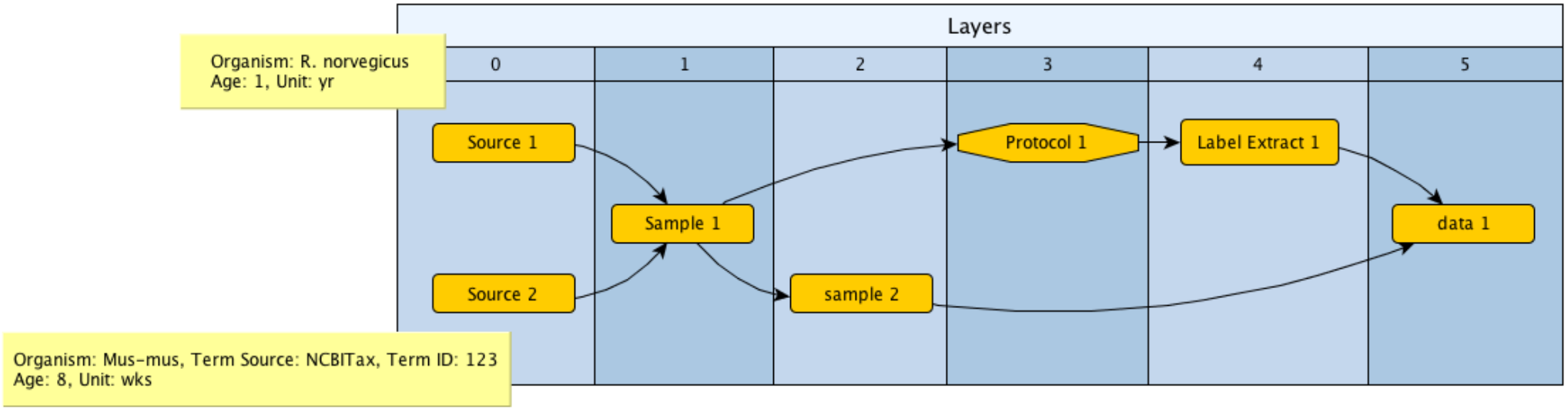




Three sub-problems

- To find a suitable/optimal set of paths and rows
- To layer the graph, so that homogeneous node will go under the same set of columns when possible
- **To arrange node attributes in a way that allows one to build the table from them**

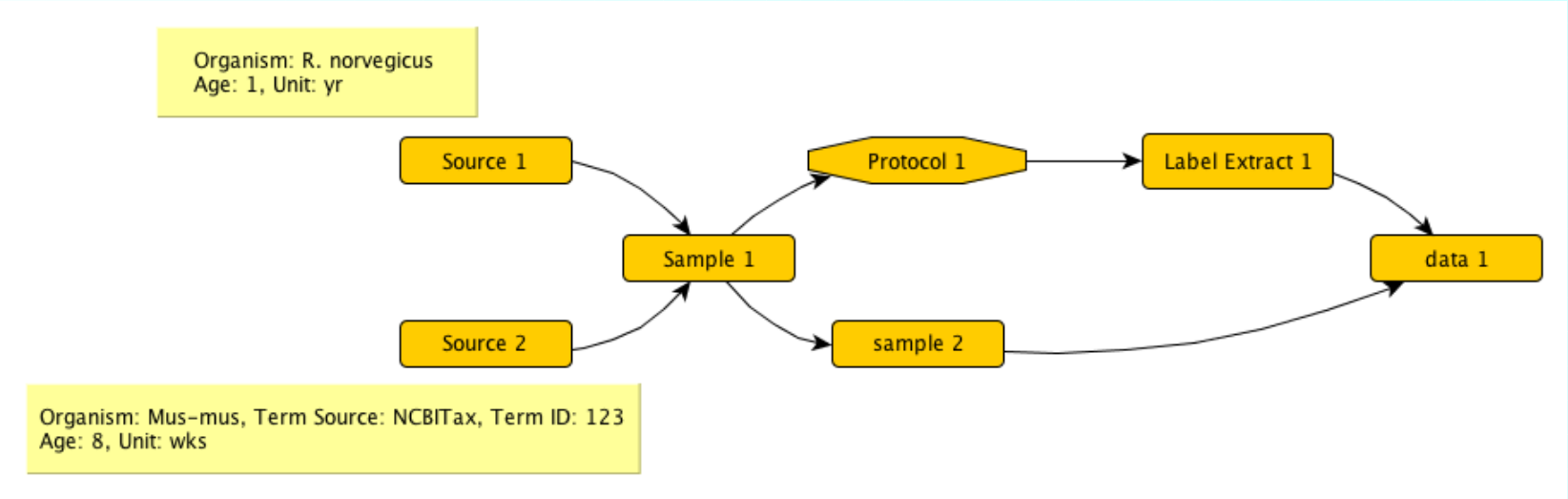
Basic Algorithm



Source Name	Organism	Term ID	Term Source	Age	Unit	Sample Name	Sample Name	Protocol REF	Labeled Extract Name	Data File
Source 1	R. norvegicus			1	yr	Sample 1		Protocol 1	Lbl Extract 1	data1.xml
Source 2	Mus-mus	123	NCBI Tax	8	wks	Sample 1	Sample 2			data1.xml

- For every path in the minimum covering path set:
  - For every *node* in the path:
    - Cover with empty cells any layer between previous node and layer*(node)*
    - Merge attributes(*node)* into the columns/rows built so far for layer(*node*)

Representing Node Attributes




```
TabValueGroup (
 header = "Organism"
 value = "Mus-mus"
 tail = ( TabValueGroup (
  header = "Term Source REF"
  value = "NCBI-Tax"
   tail = ( TabValueGroup (
    header = "Term Accession"
    value = "123"
   ))
 ))
)
```

+

```
StructuredTable (
 header = "Organism"
 rows = "R. Norvegicus"
 tail = []
)
```

=

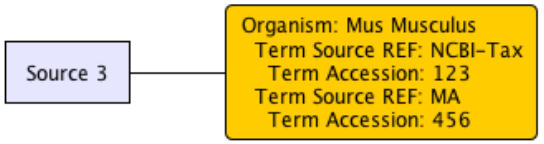
```
StructuredTable (
 header = "Organism"
 rows = "R. Norvegicus", "Mus-mus"
 tail = ( StructuredTable (
   header = "Term Source REF"
   rows = "", "", "NCBI-Tax"
   tail = ( StructuredTable (
    header = "Term Accession"
    rows = "", "", "123"
   ))
 ))
)
```

Merging Node Attributes

Source Name	Organism
Source 1	Rattus Norvegicus

Source Name	Organism	Term Source REF	Term Accession
Source 1	Rattus Norvegicus		
Source 2	Mus Musculus	NCBI-Tax	123

Source Name	Organism	Term Source REF	Term Accession	Term Source REF	Term Accession
Source 1	Rattus Norvegicus				
Source 2	Mus Musculus	NCBI-Tax	123		
Source 3	Mus Musculus	NCBI-Tax	123	MA	456



graph2tab on the Road

```
public interface Node extends Comparable<Node>
{
   public SortedSet<Node> getInputs ();
   public SortedSet<Node> getOutputs ();

   public List<TabValueGroup> getTabValues ();

   public String getType ();
   public int getOrder ();
}
```

```
public interface TabValueGroup
{
  public String getHeader();
  public String getValue();
  public List<TabValueGroup> getTail ();
}
```

```
public abstract class DefaultAbstractNode implements Node
```

Two ways to extend from basic interfaces

- Simpler approach: just let the classes of your model to implement the Node interface
  - Or, much simpler, to extend the DefaultAbstractNode
  - Simpler, but unrealistic
    - In fact, sorry, but I realised I don't have any significant example about...
- More realistic approach: wrappers, ie, a 1-1 mapping from your nodes to graph2tab nodes

Extending through wrappers

```
public abstract class ExpNodeWrapper extends DefaultAbstractNode
{
  ExpNodeWrapper ( ExperimentNode base, NodeFactory nodeFactory )
  {
    this.base = base;
    this.nodeFactory = nodeFactory;
  }
  public List<TabValueGroup> getTabValues ()
  {
    List<TabValueGroup> result = new ArrayList<TabValueGroup> ();
    result.add ( new DefaultTabValueGroup ( nameHeader, base.getName () ) );

    for ( Annotation annotation: base.getAnnotations () )
    {
      DefaultTabValueGroup tbg = new DefaultTabValueGroup (
        annotation.getType (), annotation.getValue () );
      OntoTerm ot = annotation.getOntoTerm ();
      if ( ot != null ) tbg.append (
        new DefaultTabValueGroup ( "Term Accession Number", ot.getAcc (),
        new DefaultTabValueGroup ( "Term Source REF", ot.getSource () ) ));
      result.add ( tbg );
    }
    return result;
  }
  public int getOrder ()
  {
    String header = getType();
    Integer order = TYPE_ORDER.get ( header );
    return order == null ? -1 : order;
  }

  ...
}
```

Extending through wrappers

```
public abstract class ExpNodeWrapper extends DefaultAbstractNode
{
...
    public SortedSet<Node> getInputs ()
    {
        if ( inputs != null )
            return super.getInputs ();
        inputs = new TreeSet<Node> ();
        for ( ExperimentNode in: base.getInputs () )
            inputs.add ( nodeFactory.getNode ( in ) );
        return super.getInputs ();
    }
}


public class NodeFactory extends
  org.isatools.tablib.export.graph2tab.simple_biomodel_tests.node_wrappers.NodeFactory
{
  private NodeFactory () {}
  private static final NodeFactory instance = new NodeFactory ();
  public static NodeFactory getInstance () { return instance; }
  protected ExpNodeWrapper createNewNode ( ExperimentNode base ) {
    if ( base instanceof BioSource ) return new BioSourceWrapper ( (BioSource) base, this );
    if ( base instanceof BioSample ) return new BioSampleWrapper ( (BioSample) base, this );
    if ( base instanceof BioExtract ) return
      new BioExtractWrapper ( (BioExtract) base, this );
    if ( base instanceof BioLabeledExtract )
    ...
  }
}
```

Real Use Cases so Far

- Was born while I was working on the BII project, it's now part of the ISA tools (exports studies from BII database to ISA-Tab format, works out the sample and assay file)
- ArrayExpress2 to MAGETAB exporter
  - Being integrated in the production environment, will allow to re-generate SDRF files from (possibly re-annotated) database records (ie, uses the AE2 object model)
- MAGE-ML to MAGETAB converter
  - Thanks to Natalja Kurbatova
- Re-Exporter in the Limpopo Library
  - Thanks to Natalja, Tony Burdett
- Adam Faulconbridge, working with graph2tab for the SampleDB project (???)

Possibly in Future

- Maybe a better extension mechanism
  - Java Annotations or XML mapping files
- Translation to other languages
  - If someone is interested...
- More biomedical use cases
  - Eg, several efforts in progress to represent experimental meta-data in RDF/OWL (such as OBI)
    - One may need to convert such models back to tabular formats
    - Or to visualise them as tables
- Non biomedical applications?
  - Maybe, workflows occur in many fields
  - Not so sure about tabular formats

Acknowledgements

- Philippe Rocca-Serra, Susanna Sansone (Team leaders for BII), Eamonn Maguire
  - For their support and their patience while I had struggled to come up to an acceptable solution
- Ugis Sarkans (team leader for AE) and AE team
  - Support and help for the AE exporter
- Tony Burdett
  - Support for the Limpopo use case
- Natalja Kurbatova
  - Worked out the MAGE-ML converter and on the Limpopo too
- Adam Falcounbridge
  - Help in spotting and fixing a few bugs
- Alvis Brazma, Natalja
  - Discussions on the ambiguities with the layering
- Funders (CarcinoGENOMICS and EMBL)

And you all!