NetworkWorkbench

A Workbench for Network Scientists

A Tool For Large Scale Network Analysis, Modeling and Visualization

Weixia (Bonnie) Huang

Cyberinfrastructure for Network Science Center

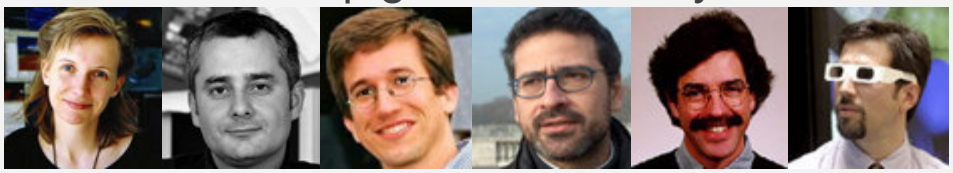
School of Library and Information Science

Indiana University, Bloomington, IN

NetworkWorkbench
A Workbench for Network Scientists

Project Details

Investigators: Katy Börner, Albert-Laszlo Barabasi, Santiago Schnell, Alessandro Vespignani & Stanley Wasserman, Eric Wernert



Software Team: Lead: Weixia (Bonnie) Huang

Developers: Santo Fortunato, Russell Duhon, Bruce Herr, Tim Kelley, Micah Walter Linnemeier, Megha Ramawat, Ben Markines, M Felix Terkhorn, Ramya Sabbineni, Vivek S. Thakre, & Cesar Hidalgo

Goal: Develop a large-scale network analysis, modeling and visualization toolkit for physics, biomedical, and social science research.

Amount: $1,120,926, NSF IIS-0513650 award

Duration: Sept. 2005 - Aug. 2008

Website: http://nwb.slis.indiana.edu

Project Details (cont.)

NWB Advisory Board:

James Hendler (Semantic Web) http://www.cs.umd.edu/~hendler/

Jason Leigh (CI) http://www.evl.uic.edu/spiff/

Neo Martinez (Biology) http://online.sfsu.edu/~webhead/

Michael Macy, Cornell University (Sociology)
http://www.soc.cornell.edu/faculty/macy.shtml

Ulrik Brandes (Graph Theory) http://www.inf.uni-konstanz.de/~brandes/

Mark Gerstein, Yale University (Bioinformatics) http://bioinfo.mbb.yale.edu/

Stephen North (AT&T) http://public.research.att.com/viewPage.cfm?PageID=81

Tom Snijders, University of Groningen http://stat.gamma.rug.nl/snijders/

Noshir Contractor, Northwestern University http://www.spcomm.uiuc.edu/nosh/








Outline

- What is “Network Science” and its challenges
- Major contributions of Network Workbench (NWB)
- Present the underlying technologies – NWB tool architecture
- Hand on NWB tool
- Review some large scale network analysis and visualization works

Basic Concepts

- Network or Graph or Matrix
- Nodes or Vertices
- Edges or Links
- Undirected vs. Directed network

  A←→ B A → B <> B → A

source	target	source	target
1	3	1	3
1	5	3	1
2	7	2	7
2	3	2	3

- Weighted vs. Unweighted network

More Basic Concepts

- Symmetric vs. Asymmetric matrix

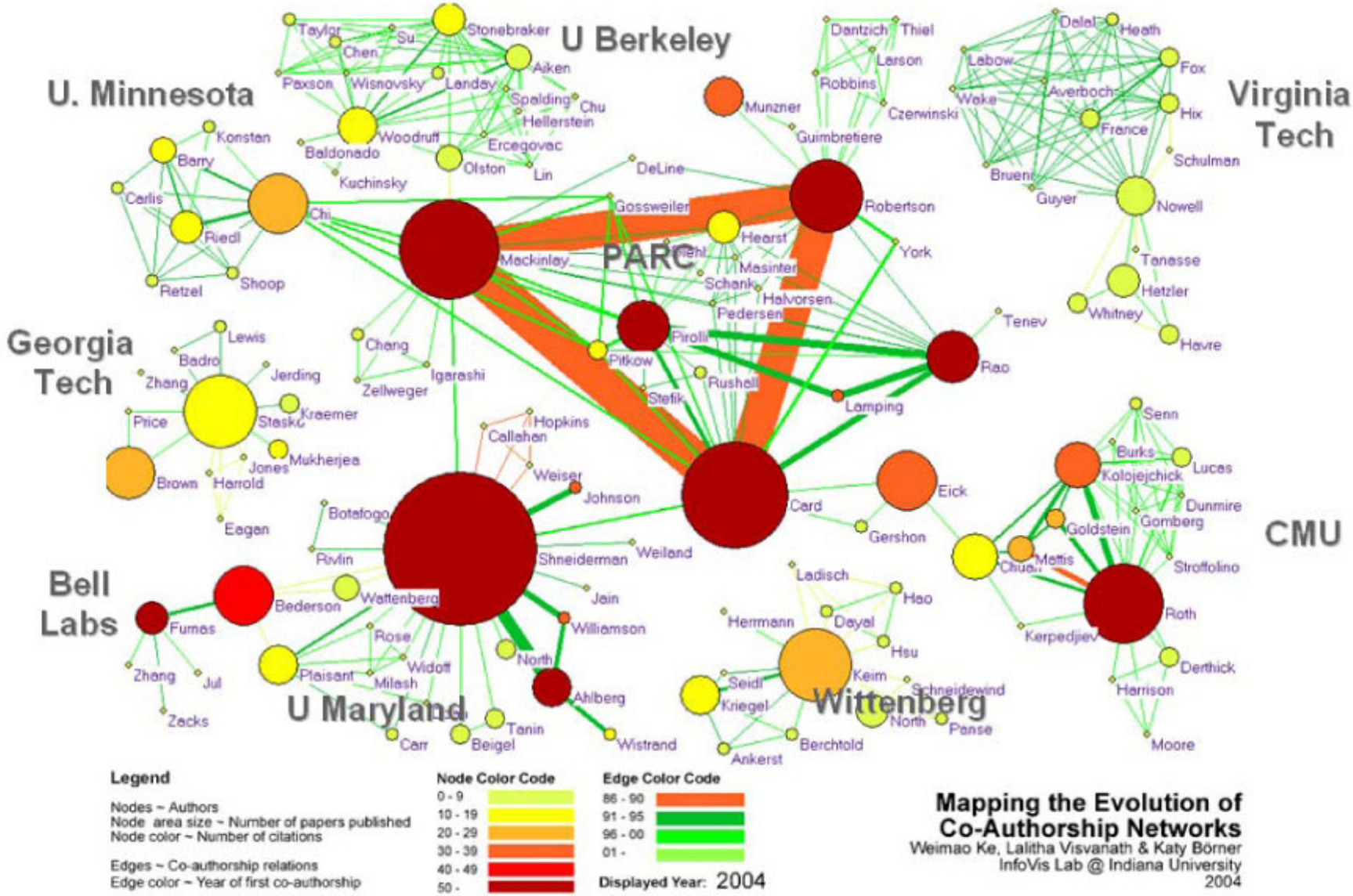
	Ann	Bob	Chris	David
Ann	0	3	2	1
Bob	3	0	2	3
Chris	2	2	0	1
David	1	3	1	0

	Ann	Bob	Chris	David
Ann	0	3	2	1
Bob	1	0	2	3
Chris	1	2	0	3
David	2	3	1	0

- Undirected network == Symmetric matrix
- Directed network == Asymmetric matrix
- Two-mode vs. One-mode network
- Self loop
- Multiple edges
- Multigraph

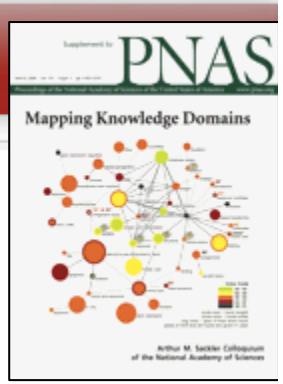
Mapping the Evolution of Co-Authorship Networks

Ke, Visvanath & Börner, (2004) Won 1st price at the IEEE InfoVis Contest



NetworkWorkbench
A Workbench for Network Scientists

Mapping Topic Bursts



Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982-2001.

Mane & Börner. (2004) PNAS, 101(Suppl. 1): 5287-5290.

Network Workbench (http://nwb.slis.indiana.edu).

Network Science

- **Physicists** study large scale network data such as Internet. In this case, each node represents a website, an edge between two nodes indicates that one website contains a URL link pointing to another website.
- Store network data as an edge list
- Study network Structure
  - Scale Free – a power law degree distribution
  - Random – a poisson distribution
  - Small World -- a network with a small shortest path and a clustering coefficient significantly higher than that of a random network with similar nodes and edges

Network Science (cont.)

- **Biologists** study gene or protein networks. In this context, each node represents a gene or a protein, edges between two nodes indicate the interactions between gene-gene or protein-protein.
- Store network data in various formats: edge list, nwb format, GraphML format, etc.
- Some sample datasets are provided in the nwb tool
- Using various layout algorithms to visualize a network with different annotations (look at a network from different view)

Network Science (cont.)

- **Social Scientists** study interactions among people.
- Usually small datasets less than 100 nodes
- Rich attribute information for nodes and edges
- Store network data in various formats: GraphML, Pajek .net, matrix
- Some sample datasets are provided in the nwb tool
- Network Analysis
  - Remove nodes:

    Run High Degree Node Deletion on a BA network
  - Remove edges:

    Run Pathfinder Network Scaling on the terror network

Network Science (cont.)

Network Science

"A science concerned with the study of networks, be they biological, technological, or scholarly networks. It contrasts, compares, and integrates techniques and algorithms developed in disciplines as diverse as mathematics, statistics, physics, social network analysis, information science, and computer science."

Börner, Katy, Sanyal, Soma and Vespignani, Alessandro. (2007) Network Science. In Blaise Cronin (Ed.), Annual Review of Information Science & Technology, Volume 41, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, chapter 12, pp. 537-607.

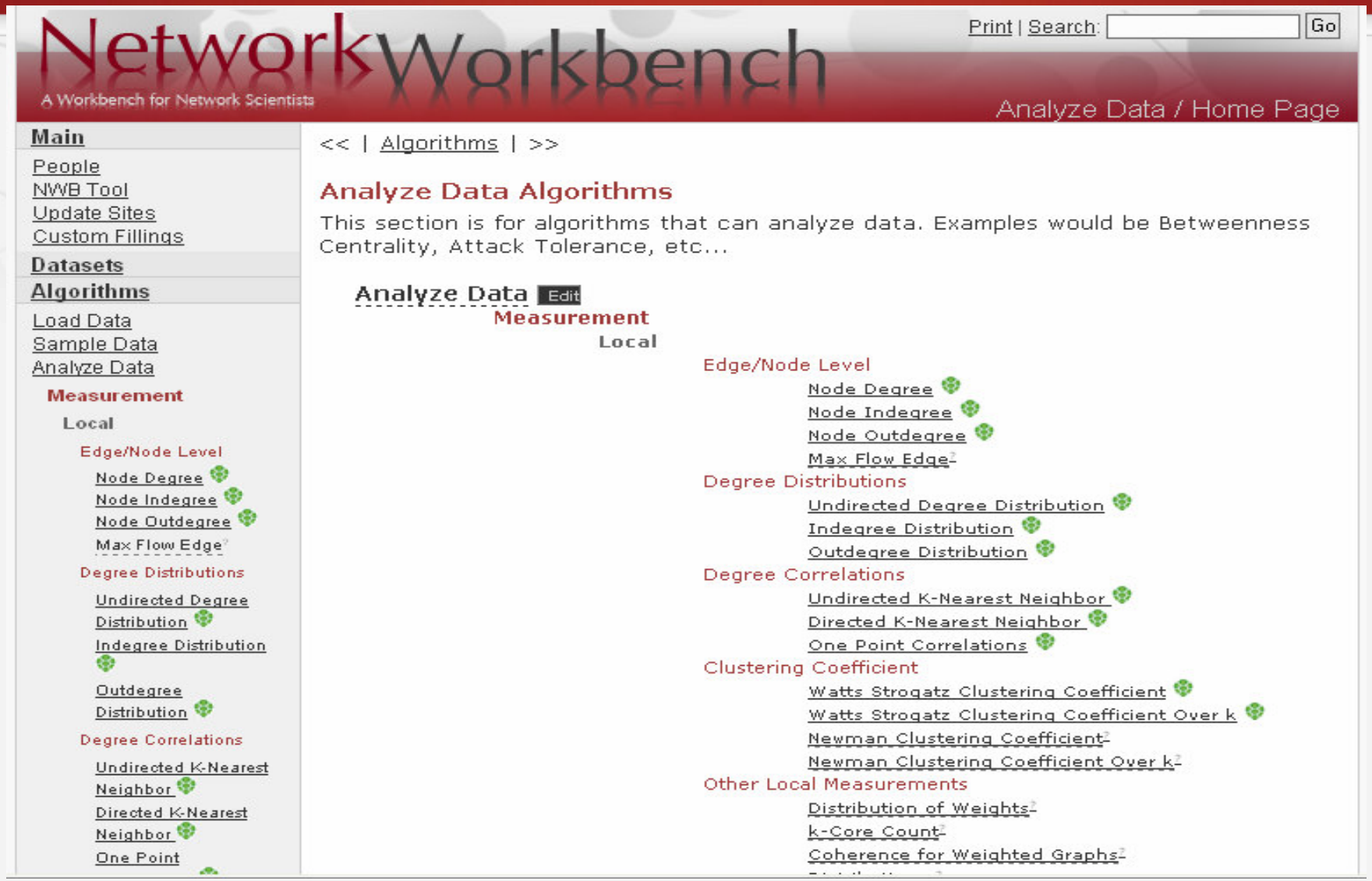
Challenges in Network Science Research

- Data
  - Different data formats
  - Different data models
- Algorithms
  - Different research purposes (preprocessing, modeling, analysis, visualization, clustering)
  - Different implementations of the same algorithm
  - Different programming languages
- Match between Data and Algorithms
- Different communities and practices
- Different tools (Pajek, UCINet, Guess, Cytoscape, R, NWB tool)

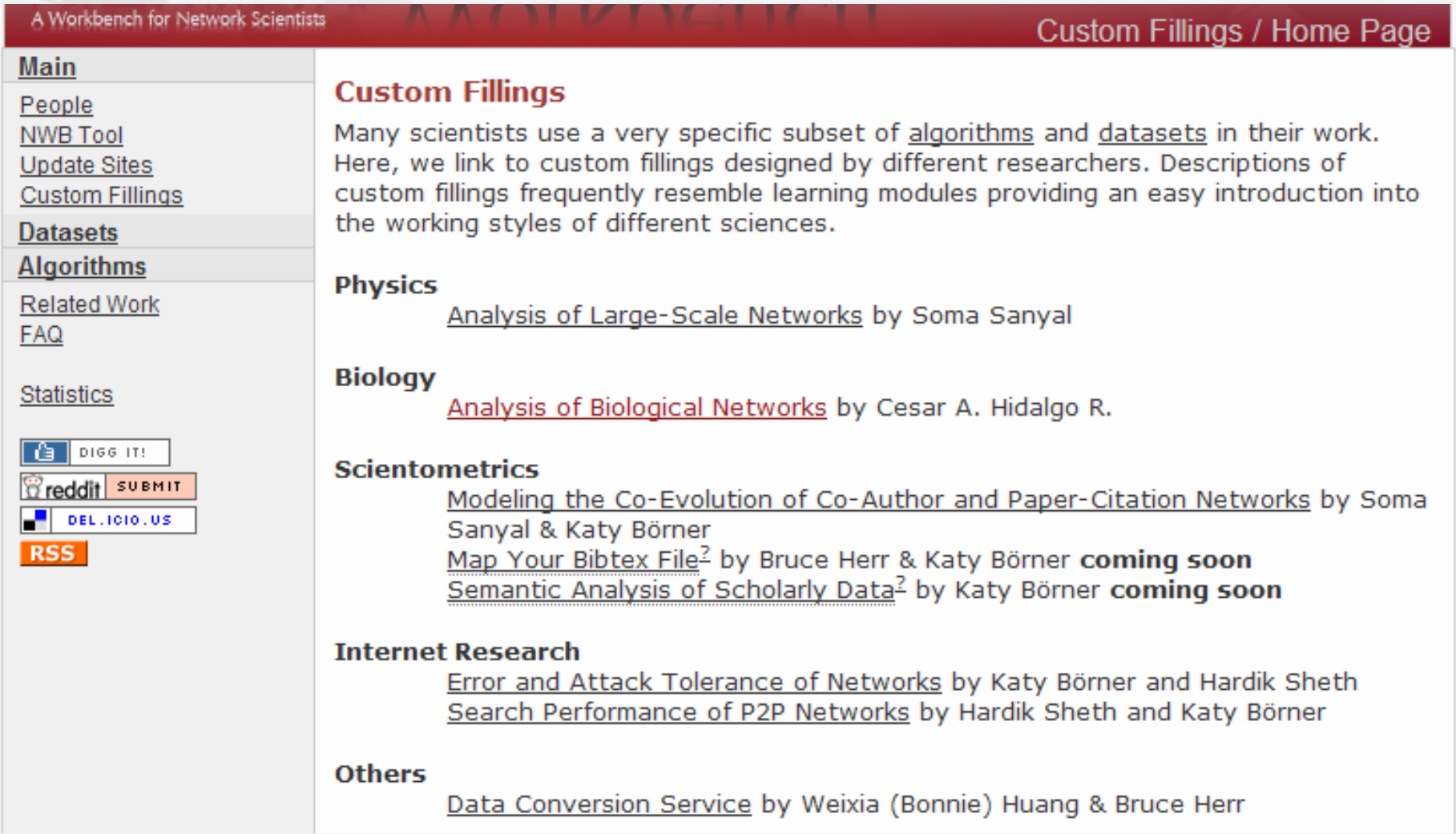
Major Deliverables

- **Network Workbench (NWB) Tool**
  - A network analysis, modeling, and visualization toolkit for physics, biomedical, and social science research.
  - Can install and run on multiple Operating Systems.
  - Uses Cyberinfrastructure Shell Framework underneath.
- **Cyberinfrastructure Shell (CIShell)**
  - An open source, software framework for the integration and utilization of datasets, algorithms, tools, and computing resources**.**
  - Uses OSGi and Equinox
- **NWB Community Wiki**
  - A place for users of the NWB Tool, the Cyberinfrastructure Shell (CIShell), or any other CIShell-based program to request, obtain, contribute, and share algorithms and datasets.
  - All algorithms and datasets that are available via the NWB Tool have been well documented in the Community Wiki.

Community Wiki



Community Wiki (cont.)



NWB Tool Major Deliverables

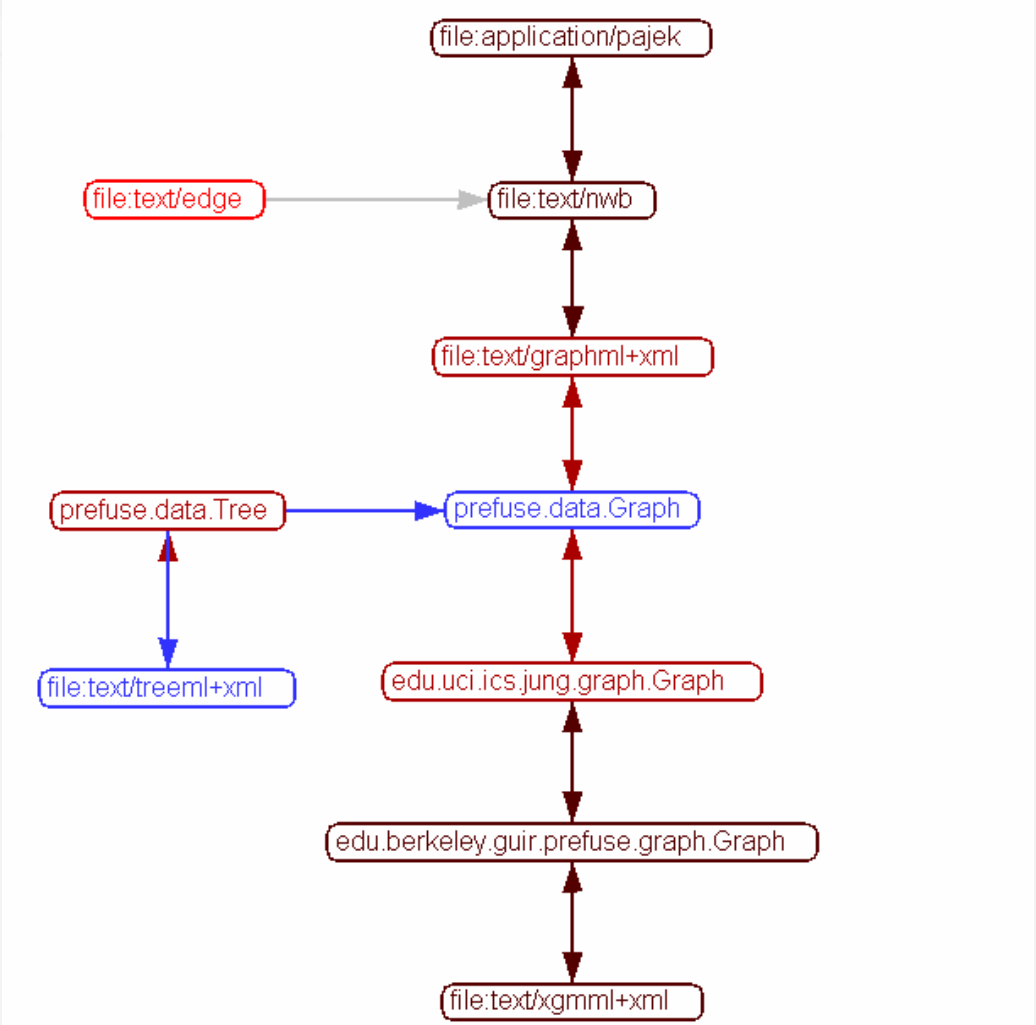
Download from http://nwb.slis.indiana.edu/software.html

Major features in v0.6.0 Release

- Installs and runs on Windows, Linux x86 and Mac osx.
- Provides over 50 modeling, analysis and visualization algorithms. Half of them are written in Fortran, others in Java.
- Supports large scale network modeling and analysis (over 100,000 nodes)
- Supports various visualization layouts with node/edge annotation.
- Provides several sample datasets with various formats.
- Supports multiple ways to introduce a network to the NWB tool.
- Supports the loading, processing and saving of four basic file formats: GraphML, Pajek .net, XGMML and NWB. Can load and view TreeML, edge list, etc.
- Supports automatically Data Conversion.
- Provides a Scheduler to monitor and control the progress of running algorithms.
- Integrates a 2D plotting tool -- Gnuplot.

NWB Tool – Data Formats

Converters and Conversion Services Between Various Data Formats



NetworkWorkbench
A Workbench for Network Scientists

Integrating and Implementing Algorithms

Modeling and Network Generation

Random Network Model
Random

Preferential Attachment Algorithms
Barabasi-Albert Model
Dorogovtsev-Mendes-Samukhin
Fitness
Vertices/edges deletion
Copying strategy
Finite vertex capacity
TARL

Rewiring algorithms
Rewiring based on degree distribution
Watts Strogatz Small World Model

Peer-to-Peer Models

Structured
CAN Model
Chord Model

Unstructured
PRU Model
Hypergrid Model

Statistical Measurement

Edge/Node level
node degree
BC value of nodes/edges
Max flow edge
Hub/Authority value for nodes
Distribution of node distances (Hop plot)
Local (directed and weighted versions)
Clustering Coefficient (Watts Strogatz)
Clustering Coefficient (Newman)
k-Core Count
Distributions (Plot and gamma, and R^2)
Degree Distributions (in, out, total) (Directed/TotalDegree Distribution)
Degree Correlations (in-out, out-out, out-in, in-in, total-total)
Clustering Coefficient over k
Coherence for weighted graphs
Distribution of weights
Probability of degree distribution
Global
Density
Square of Adjacency Matrix
Giant Component
Strongly Connected Component
Betweenness Centrality
Diameter
Shortest Path = Geodesic Distance
Average Path Length

Motif Identification
Page Rank
Closeness centrality
Reach centrality
Eigenvector centrality
Minimum Spanning Tree

NetworkWorkbench
A Workbench for Network Scientists

More Algorithms

Searching on Networks
Search
k Random-Walk Search
Depth First Search
p-rand Breadth-First Search
P2P
CAN Search
Chord Search

Epidemics Spreading
SIR
SIS

Graph Matching On Networks
Simple Match
Similarity Flooding
ABSURDIST

Clustering on Networks
Based on Attributes
Hierarchical Clustering
Single Link
Complete Link
Average Link
Ward's Algorithm

Based on Network Structure
Newman Girvan
Clauset-Newman-Moore
Newman
Cecconi-Parisi
Simulated annealing of modularity
Caldarelli
Weak Component Clustering
vanDongen (random walk)
Cfinder (Clique percolation method)
Reichardt, Bornholdt (q-potts model)

Visualization of Networks
Distribution
Scatterplot
Histogram
Geospatial
Circle layout
Grid-based
Dendrogram
Treemap
Hyperbolic tree
Radial Tree
Sparse Matrix Visualization
Kamada-Kawaii
Fruchterman-Rheingold
Orthogonal Layout
k-core visualization

NetworkWorkbench
A Workbench for Network Scientists

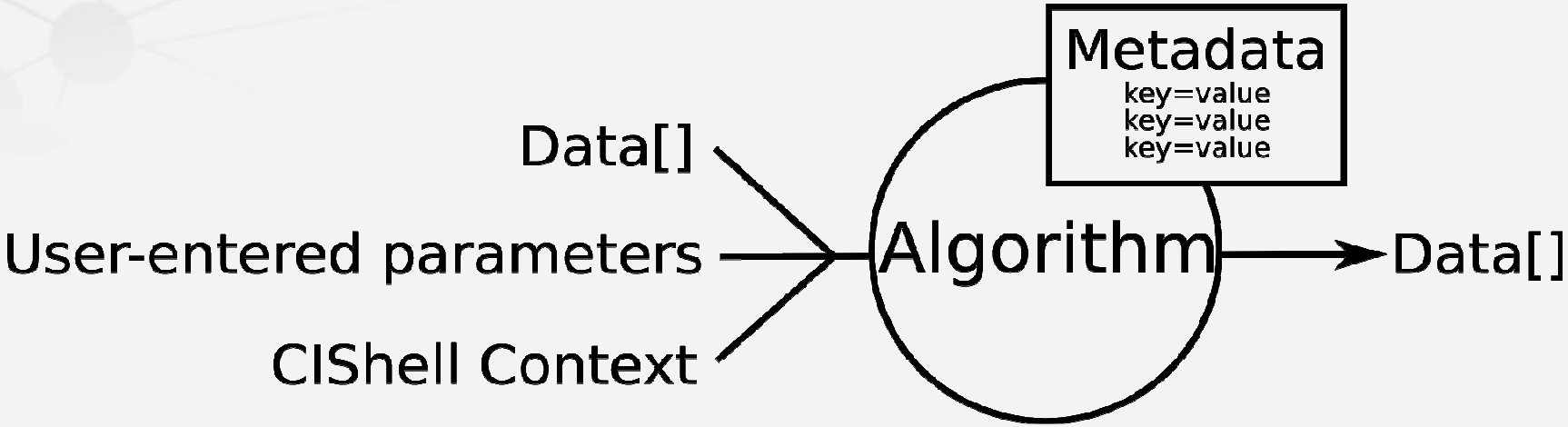
NWB Tool – Algorithms (Implemented)

Category	*Algorithm*	*Language*
Preprocessing	Random Node Deletion	JAVA
	High Degree Node Deletion	JAVA
	Pathfinder Network Scaling	JAVA
	Directory Hierarchy Reader	JAVA
Modeling	Erdös-Rényi Random	FORTRAN
	Barabási-Albert Scale-Free	FORTRAN
	Watts-Strogatz Small World	FORTRAN
	Chord	JAVA
	CAN	JAVA
	Hypergrid	JAVA
	PRU	JAVA
	TARL	JAVA
Visualization	Tree Map	JAVA
	Tree Viz	JAVA
	Radial Tree / Graph	JAVA
	Kamada-Kawai	JAVA
	Force Directed	JAVA
	Spring	JAVA
	Fruchterman-Reingold	JAVA
	Circular	JAVA

Analysis Algorithm	*Language*
Node Betweenness Centrality	FORTRAN
Average Shortest Path	FORTRAN
Connected Components	FORTRAN
Diameter	FORTRAN
Page Rank	FORTRAN
Shortest Path Distribution	FORTRAN
Watts-Strogatz Clustering Coefficient	FORTRAN
Watts-Strogatz Clustering Coefficient Versus Degree	FORTRAN
Directed k-Nearest Neighbor	FORTRAN
Undirected k-Nearest Neighbor	FORTRAN
Indegree Distribution	FORTRAN
Outdegree Distribution	FORTRAN
Node Indegree	FORTRAN
Node Outdegree	FORTRAN
One-point Degree Correlations	FORTRAN
Undirected Degree Distribution	FORTRAN
Node Degree	FORTRAN
k Random-Walk Search	JAVA
Random Breadth First Search	JAVA
CAN Search	JAVA
Chord Search	JAVA
Weak Component Clustering	JAVA
Tool: GnuPlot	

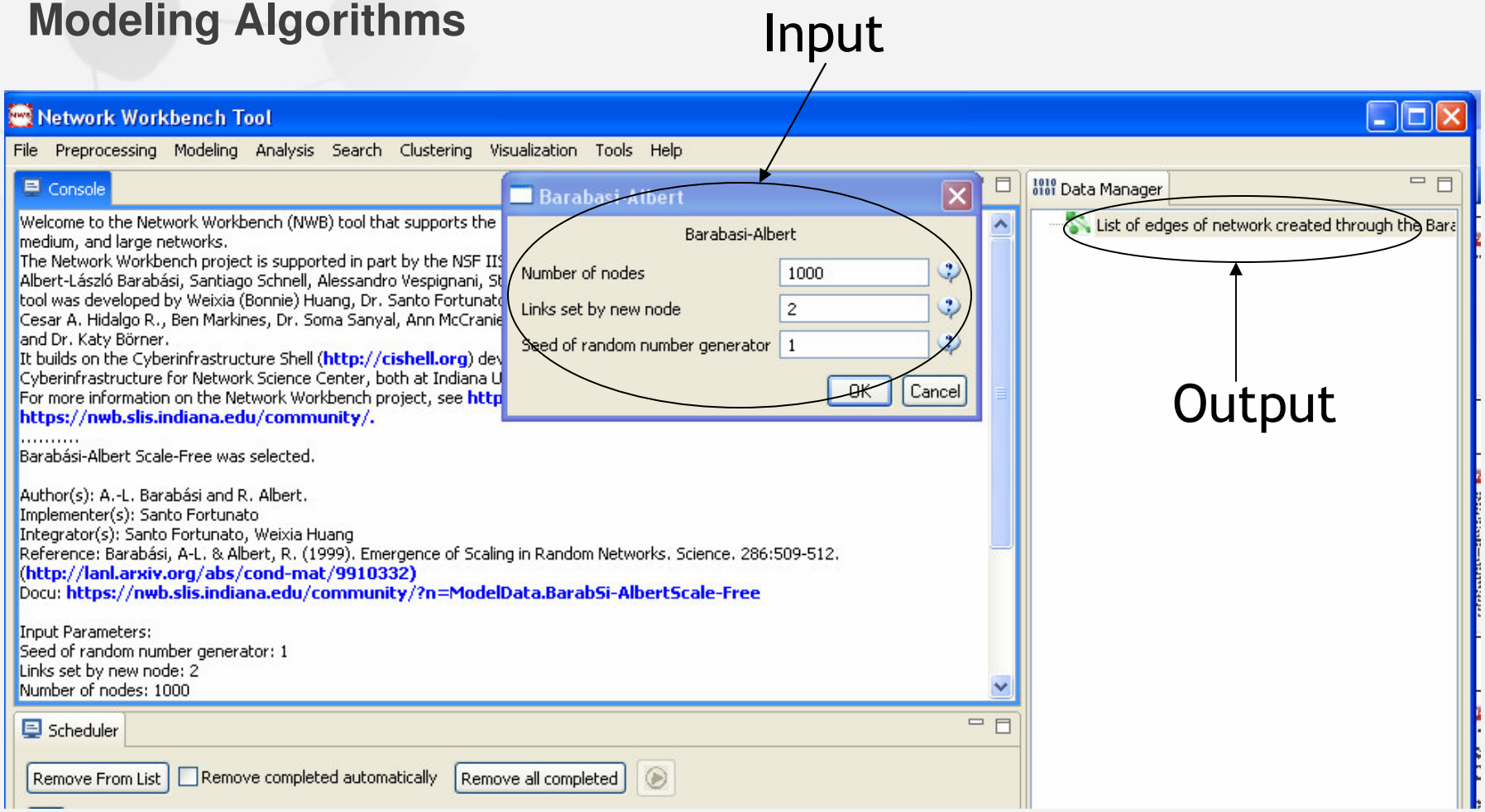
Abstract Algorithm Definition

An Abstract Definition of Algorithms, Datasets and Converters



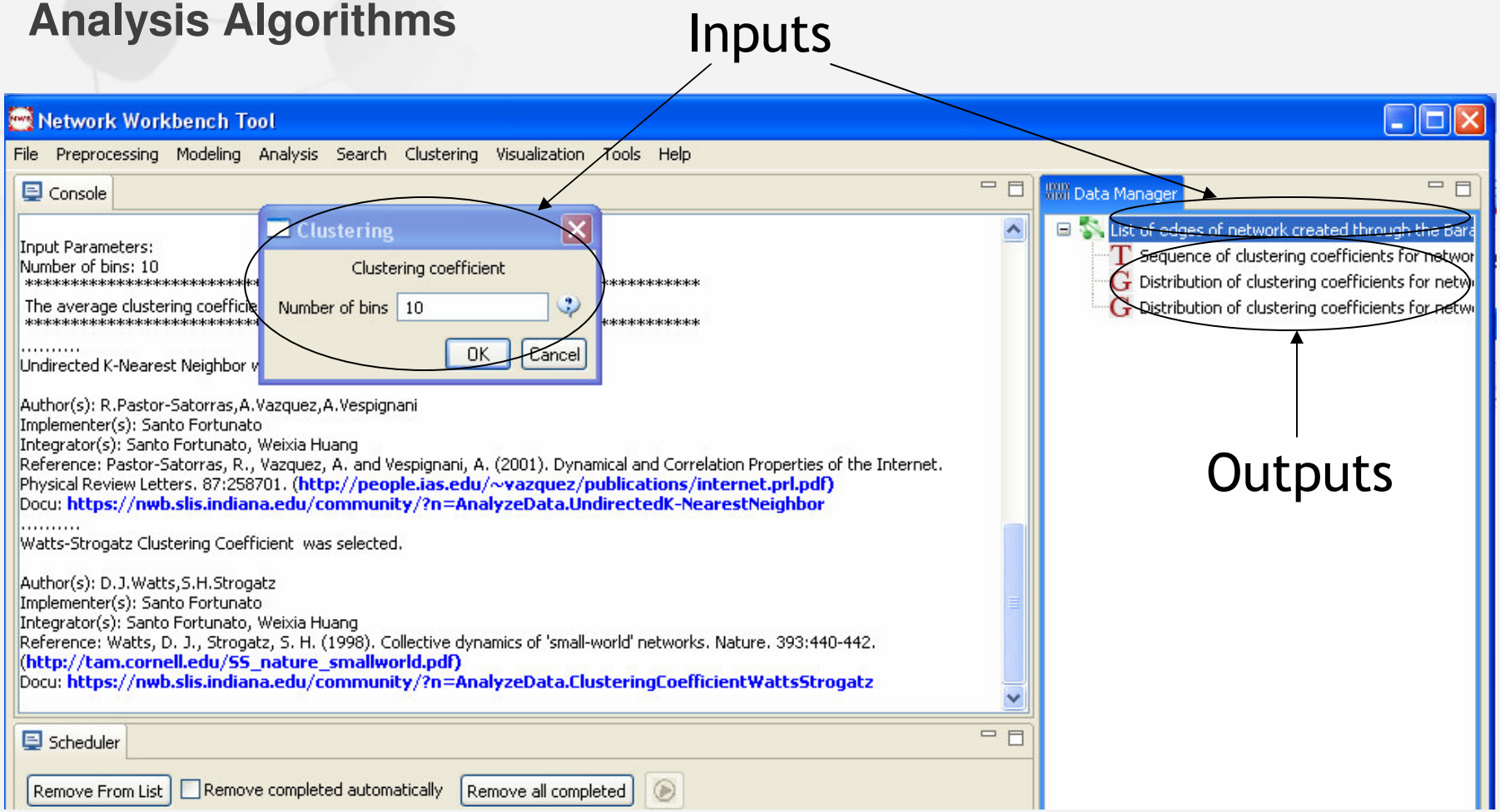
Abstract Algorithm Definition (cont.)

Modeling Algorithms



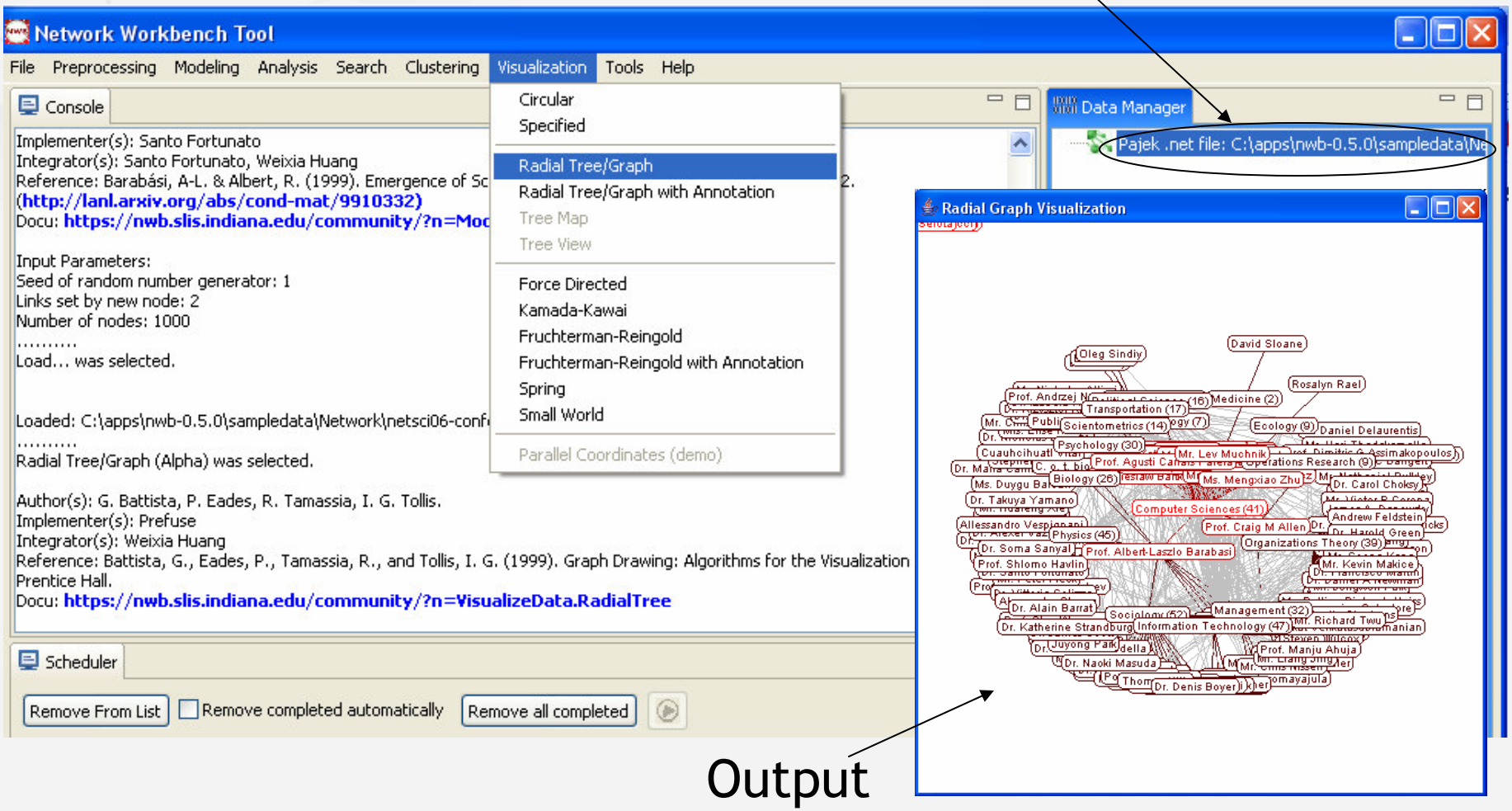
Abstract Algorithm Definition (cont.)

Analysis Algorithms



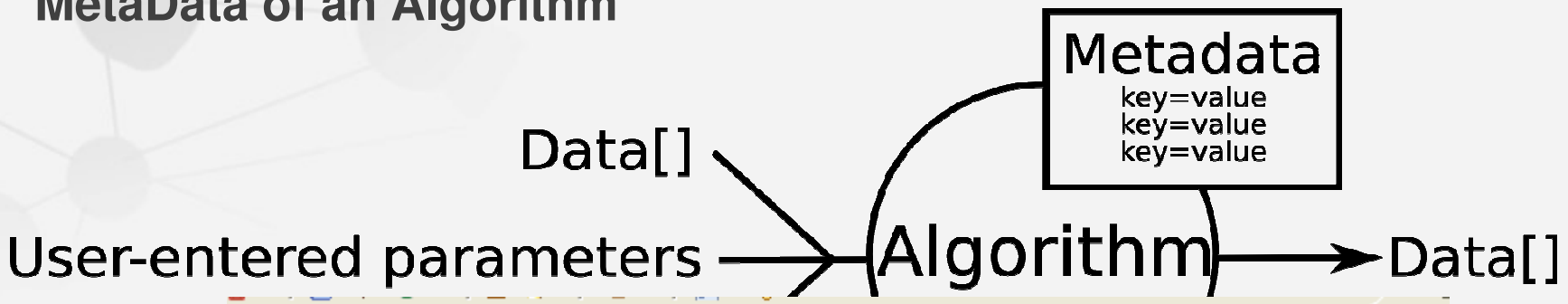
Abstract Algorithm Definition (cont.)

Visualization Algorithms



For Algorithm Integrators

MetaData of an Algorithm

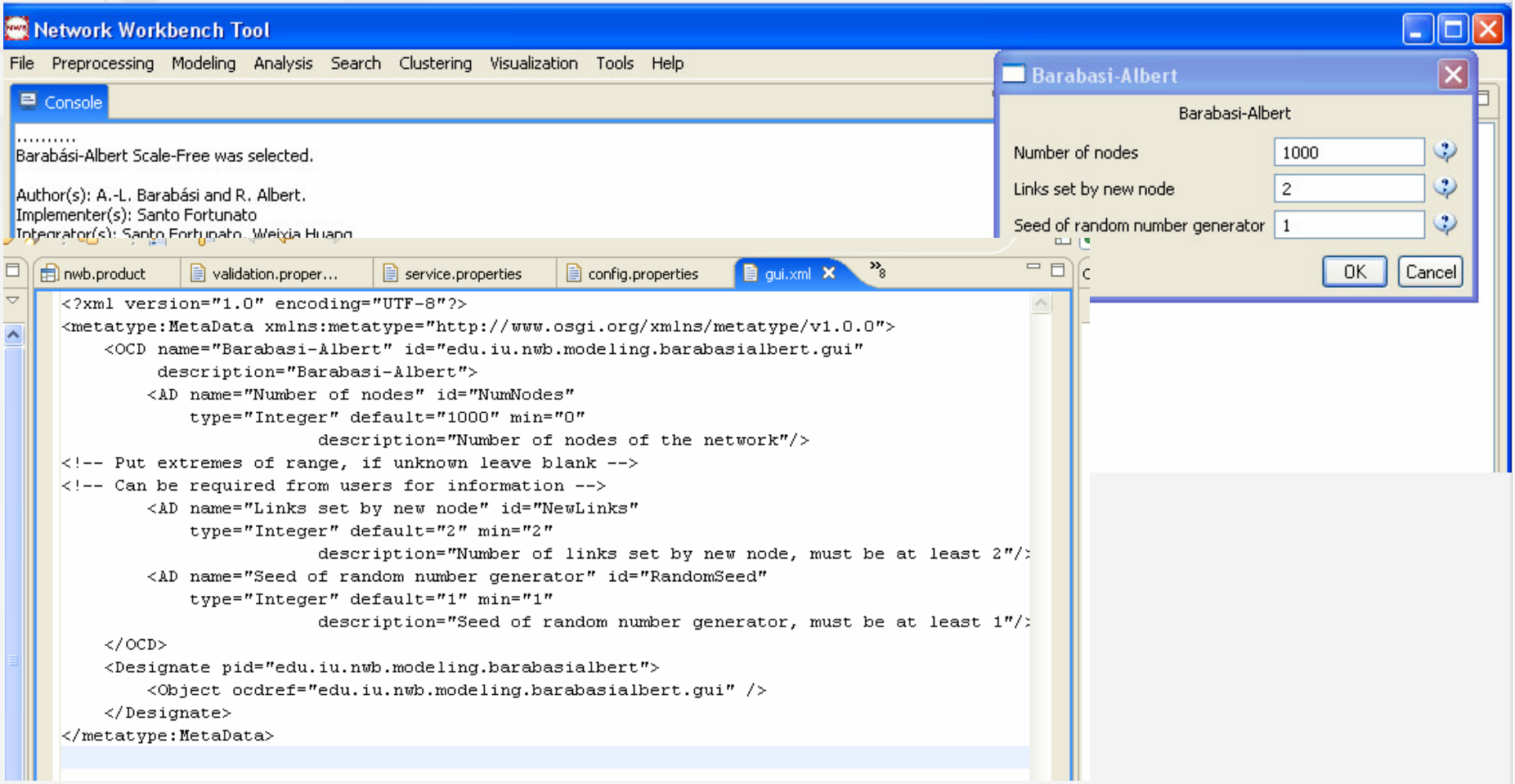


nwb.product | edu.iu.iv.algorit... | validation.proper... | service.properties

```
menu_path=Modeling/additions
label=Barabási-Albert Scale-Free
description=Barabasi-Albert algorithm implementation
# each input file will be mapped to inFile[x] (zero based)
in_data=null
# for all input files, 'null' if no input data needed
out_data=file:text/nwb
# for all output files
# each output file will correspond to outFile[x] (zero based)
service.pid=edu.iu.nwb.modeling.barabasialbert
remoteable=true
authors=A.-L. Barabási and R. Albert.
implementers=Santo Fortunato
integrators=Santo Fortunato, Weixia Huang
reference=Barabási, A-L. & Albert, R. (1999). Emergence of Scaling in Random
reference_url=http://lanl.arxiv.org/abs/cond-mat/9910332
docu=https://nwb.slis.indiana.edu/community/?n=ModelData.BarabSi-AlbertScale
```

For Algorithm Integrators (cont.)

GUI Builder and MetaType Service



Three User Groups

Application Users

- Scientists in the natural and social sciences (physics, biology, chemistry, psychology, sociology, etc.)
- Their needs -- want to find the best datasets and the most effective algorithms to conduct their research.
- Problem – too many algorithms. Finding a correctly working piece of code is challenging. Frequently, not only one but a sequence of different algorithms needs to be applied to load, parse, clean, mine, analyze, model, visualize, and print data. Today, there is no easy way to extend a tool by adding new algorithms as needed or to customize a tool so that it exactly fits the needs of a specific user (group).

Three User Groups (cont.)

Application Designers/Developers

- Computer scientists or application users that developed the applications and tools we use today.
- They usually start by developing applications/tools that meet their own needs, and then generalize them to satisfy the requirements of their research community.
- Challenge -- not only need to take care of the software architecture, the GUI design, the development of many basic components and functionalities, but also play the role of algorithm developers.

Three User Groups (cont.)

Algorithm Developers

- Computer scientists, statisticians and other researchers
- They look for opportunities to disseminate their work and test the practical utilities of their algorithms.
- Challenge -- the integration of a dataset or algorithm into an existing application or tool requires a deep understanding of the architecture of that application, which is non-trivial.

OSGi – Technical Details

NWB/CIShell is built upon the Open Services Gateway Initiative (OSGi) Framework.

OSGi (http://www.osgi.org) is

- A standardized, component oriented, computing environment for networked services.
- Alliance members include IBM (Eclipse), Sun, Intel, Oracle, Motorola, NEC and many others.
- Has successfully been used in the industry from high-end servers to embedded mobile devices for 8 years now.
- Widely adopted in open source realm, especially since Eclipse 3.0 that uses OSGi R4 for its plugin model.

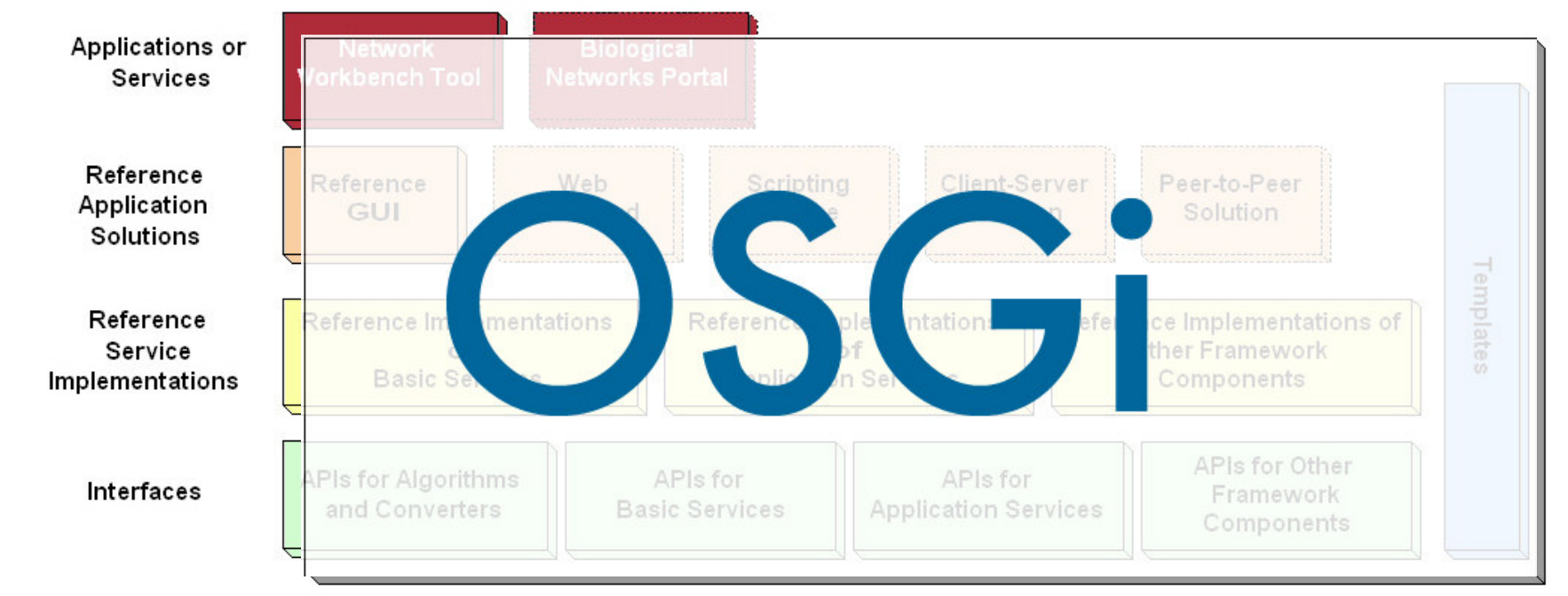
Advantages of Using OSGi

- Directly use many components provided by OSGi framework, such as service registry
- Contribute diverse algorithms to OSGi community -- any CIShell algorithm becomes a service that can be used in any OSGi-based framework.
- Running CIShells/tools can connect to each other via exposed CIShell-defined web services supporting peer-to-peer sharing of data, algorithms, and computing power.

Ideally, CIShell becomes a standard for creating algorithm services in OSGi developed Tools/CI, e.g., IVC&NWB will be using the CIShell reference GUI

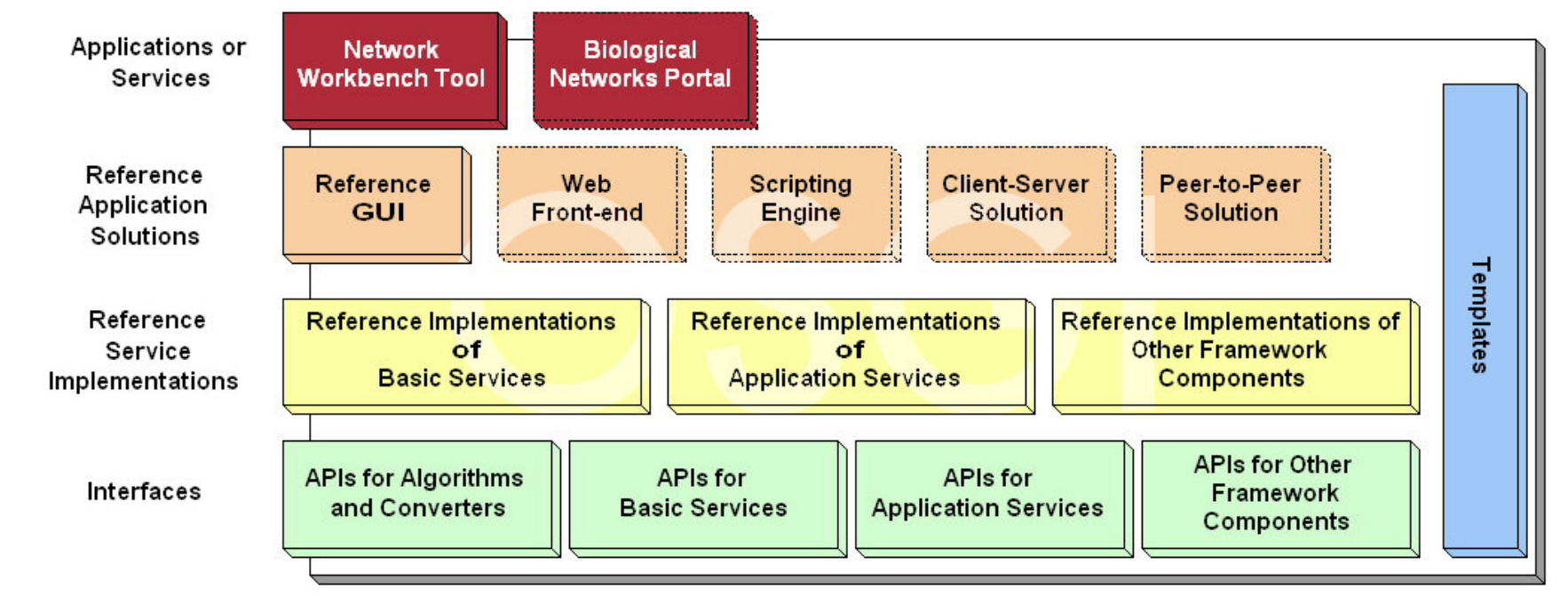
OSGi – Technical Details (cont.)

Service Oriented Component-based Architecture



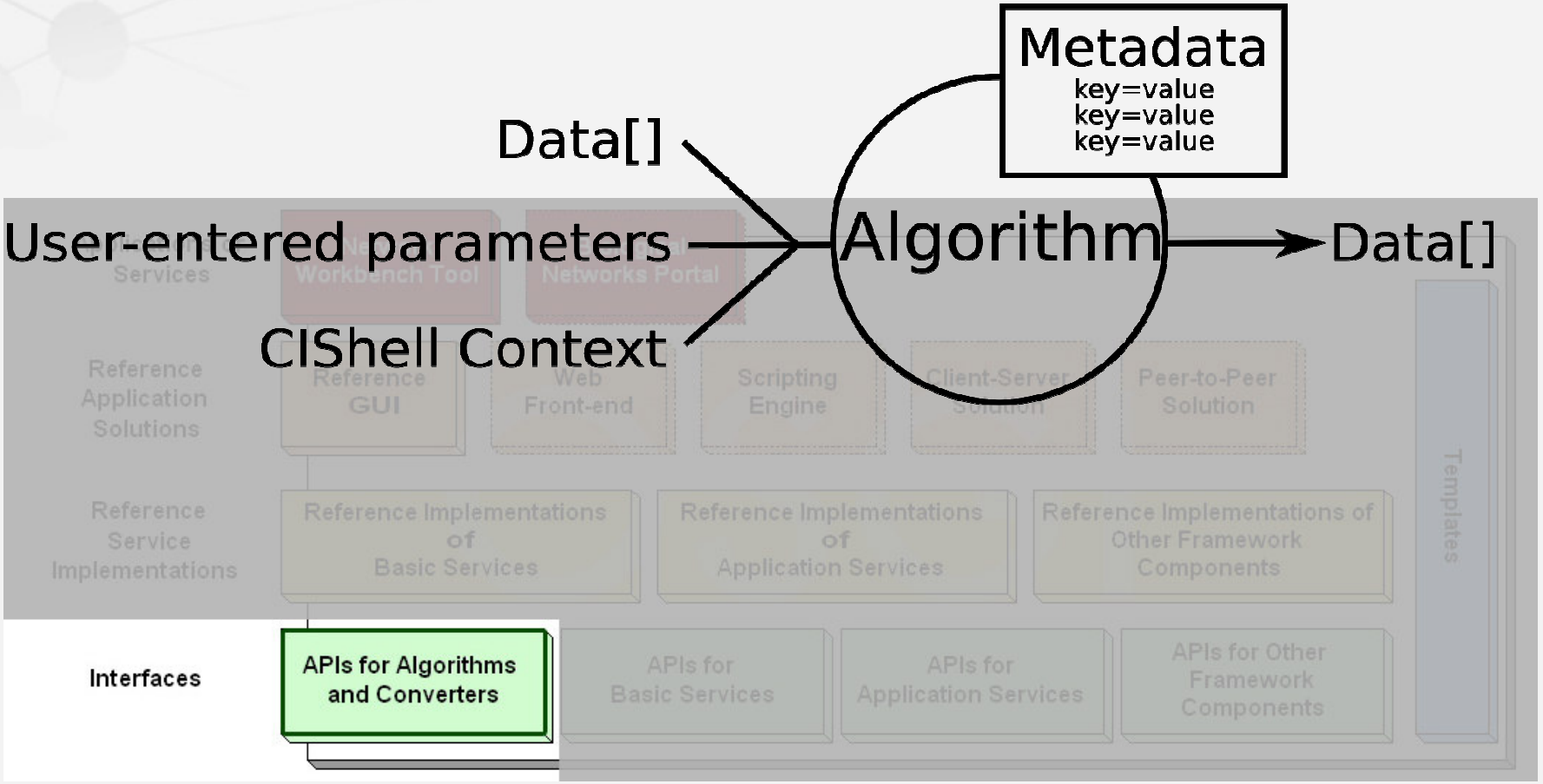
NWB/CIShell Architecture

An Overview of NWB/CIShell Architecture



Interfaces Layer – Algorithm

An Abstract Definition of Algorithms, Datasets and Converters



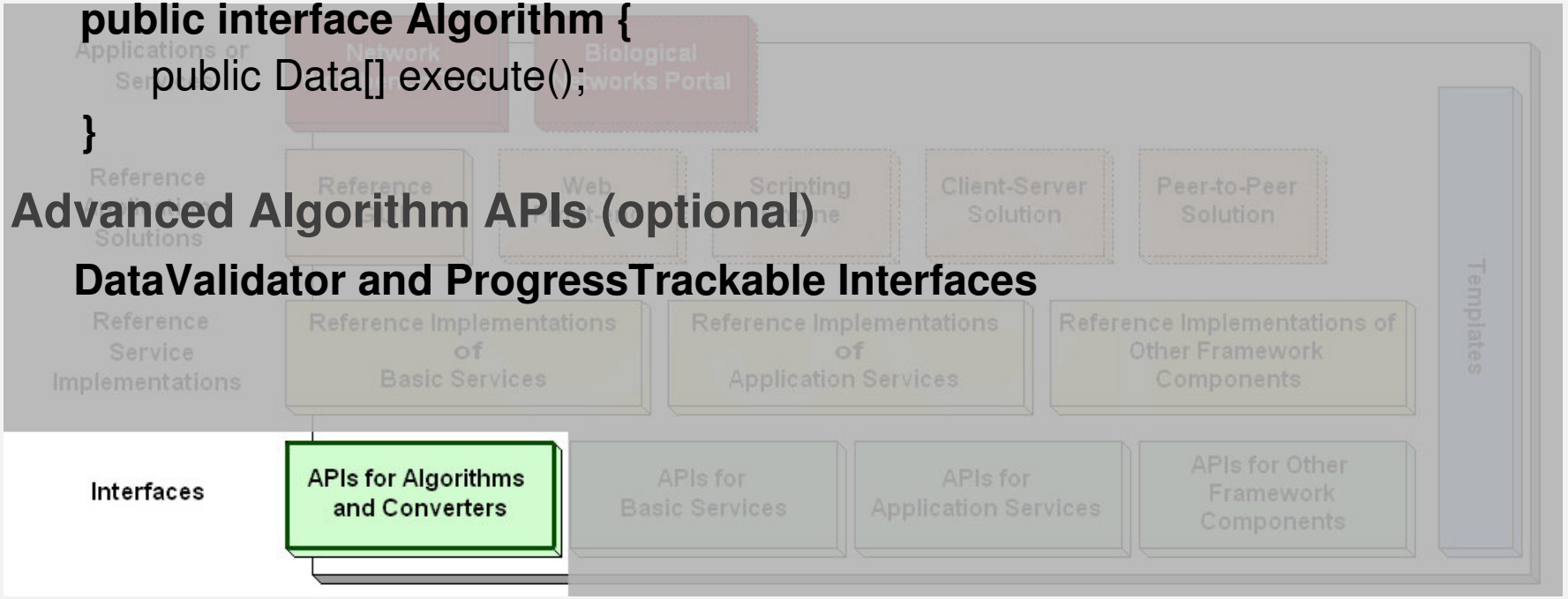
Interfaces Layer – Algorithm (cont.)

Basic Algorithm APIs

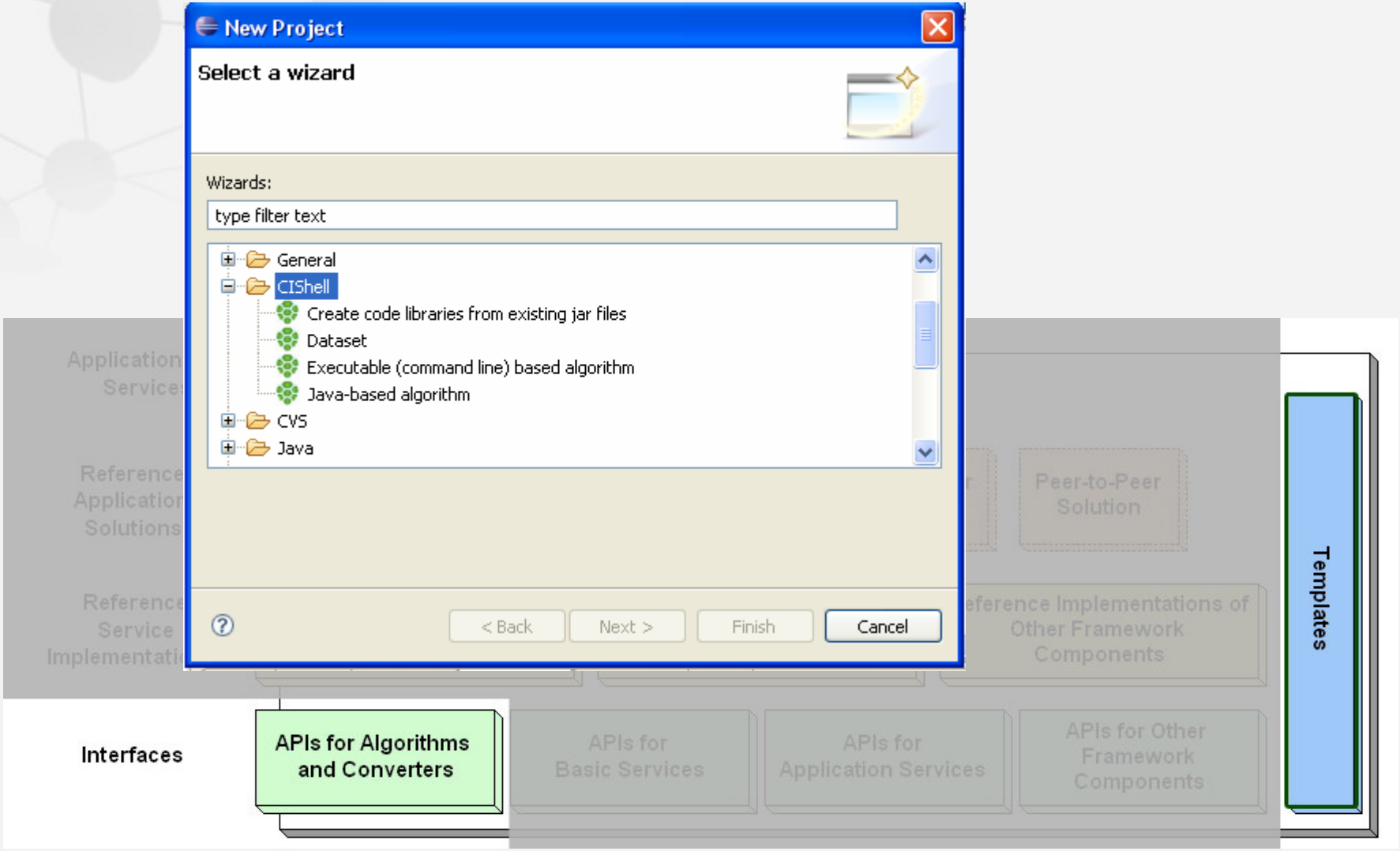
```
public interface AlgorithmFactory {
    public   MetaTypeProvider   createParameters(Data[] data);
    public   Algorithm   createAlgorithm( Data[] data, Dictionary parameters,
                                          CIShellContext context);
}
public interface Algorithm {
    public Data[] execute();
}
```

Advanced Algorithm APIs (optional)

DataValidator and ProgressTrackable Interfaces



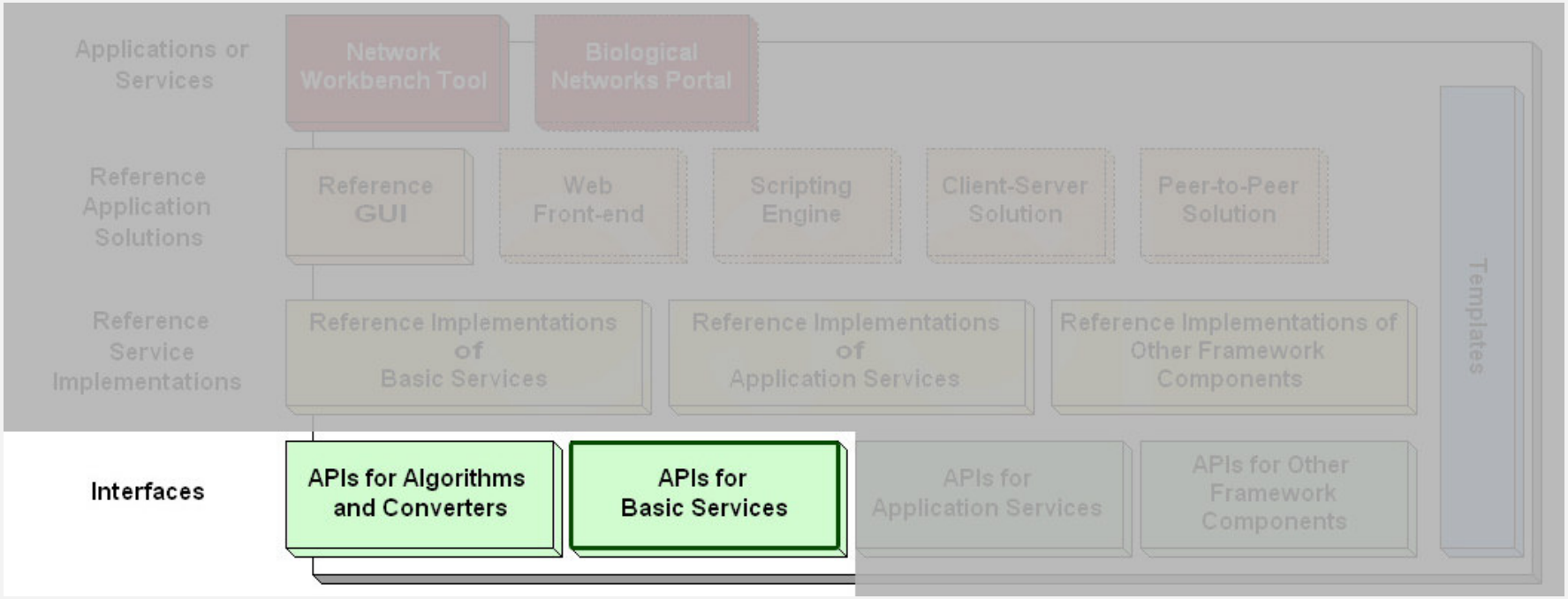
Templates


New Project
Select a wizard
Wizards:
type filter text
General
CIShell
Create code libraries from existing jar files
Dataset
Executable (command line) based algorithm
Java-based algorithm
CVS
Java
< Back
Next >
Finish
Cancel
Application Service
Reference Application Solutions
Reference Service Implementati
Peer-to-Peer Solution
eference Implementations of Other Framework Components
Templates
Interfaces
APIs for Algorithms and Converters
APIs for Basic Services
APIs for Application Services
APIs for Other Framework Components

Interfaces Layer – Basic Services

Basic Services

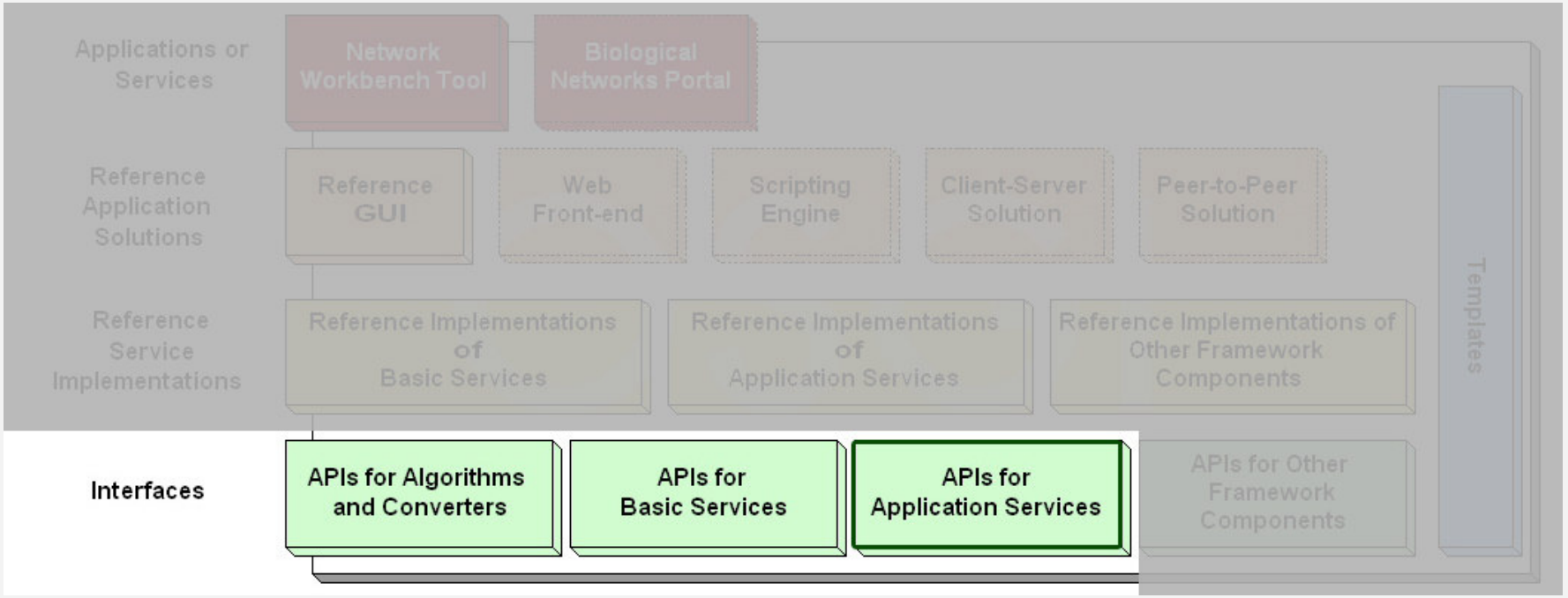
- Preferences Service
- Log Service
- Data Conversion Service
- GUI Builder Service



Interfaces Layer – Application Services

Application Services

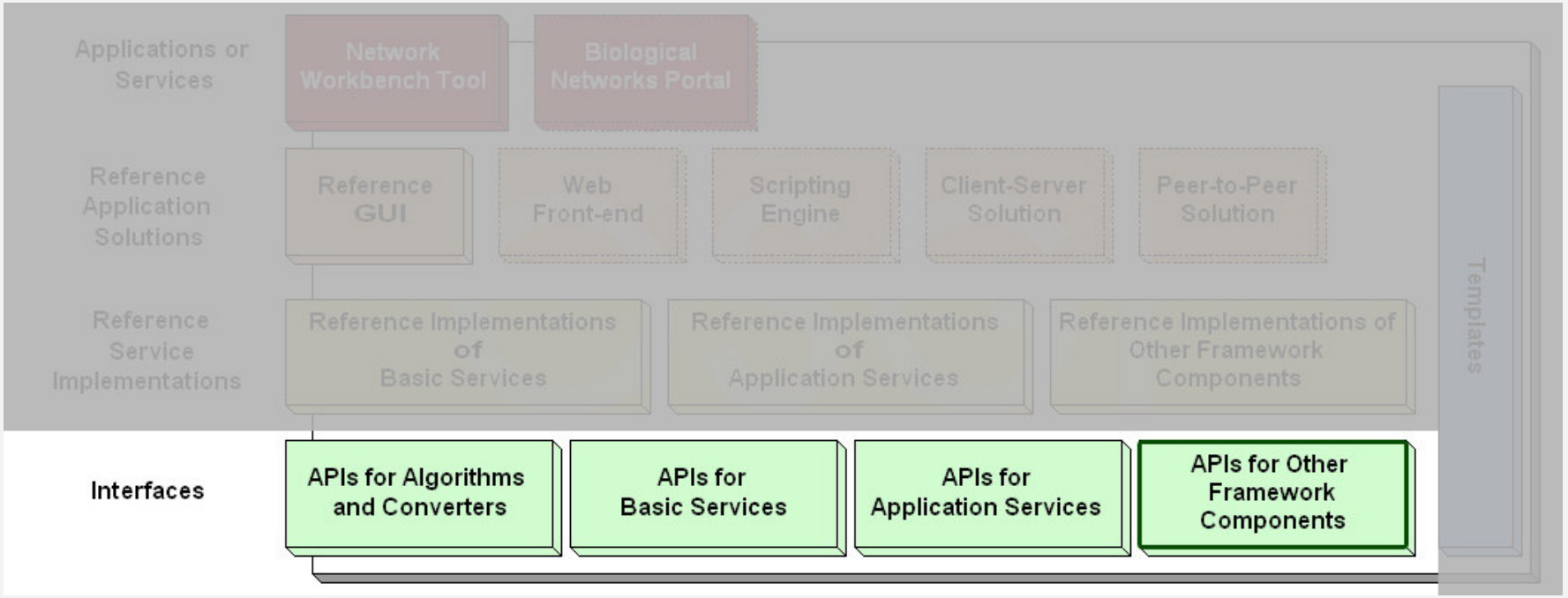
- Scheduler Service
- Data Manager Service



Interfaces Layer – Other Components

Other Framework Components

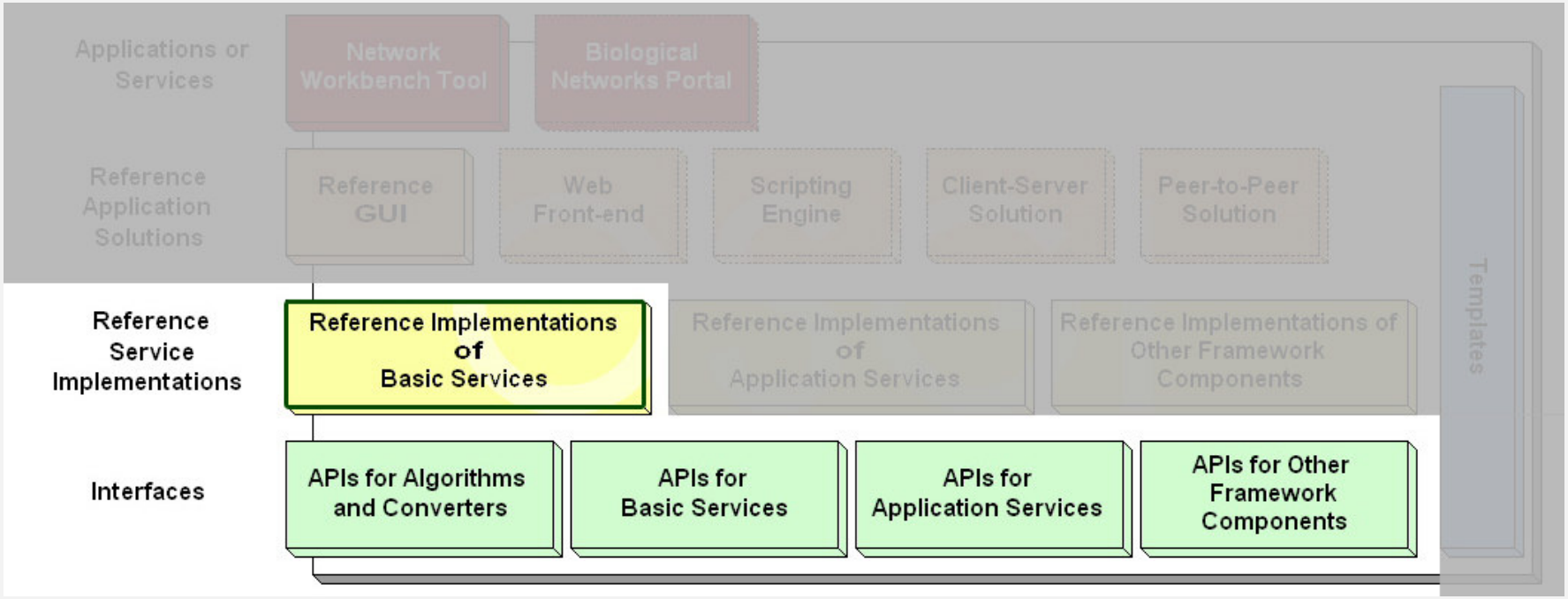
- CIShellContext
- Data



Services Layer – Basic Services

Basic Services

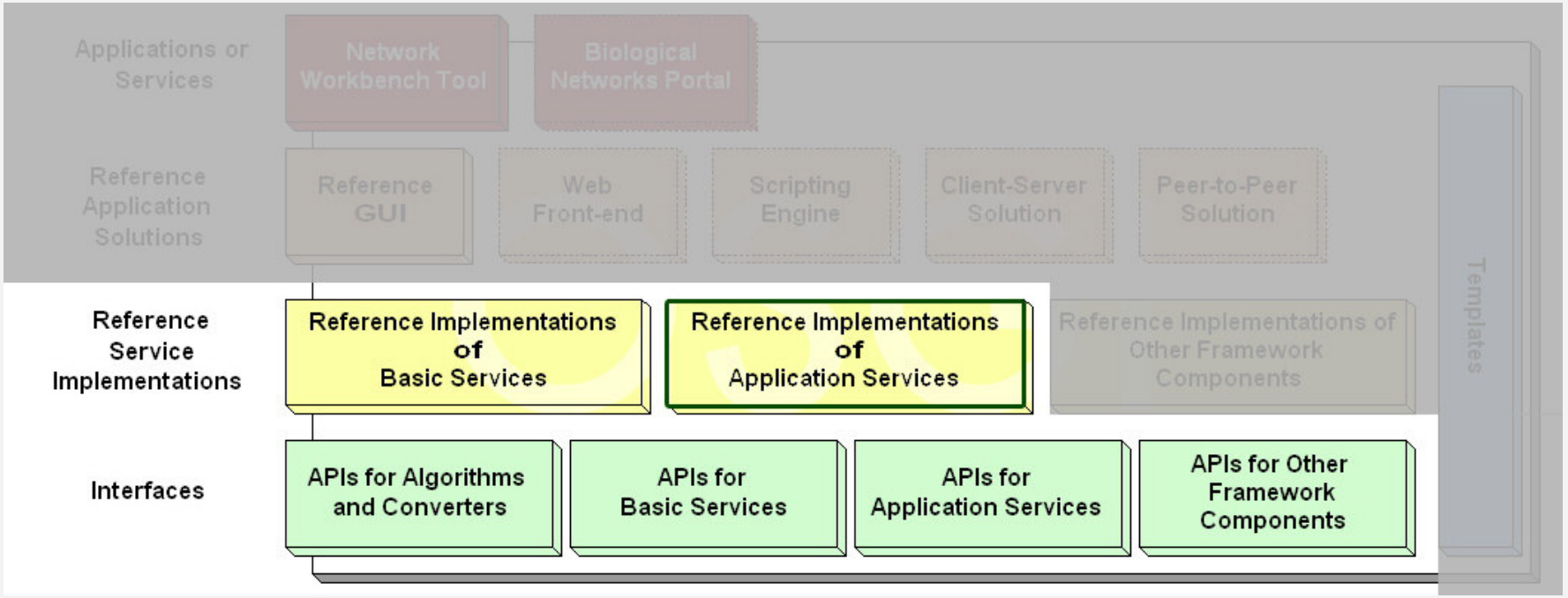
- Preferences Service
- Log Service
- Data Conversion Service
- GUI Builder Service



Services Layer – Application Service

Application Services

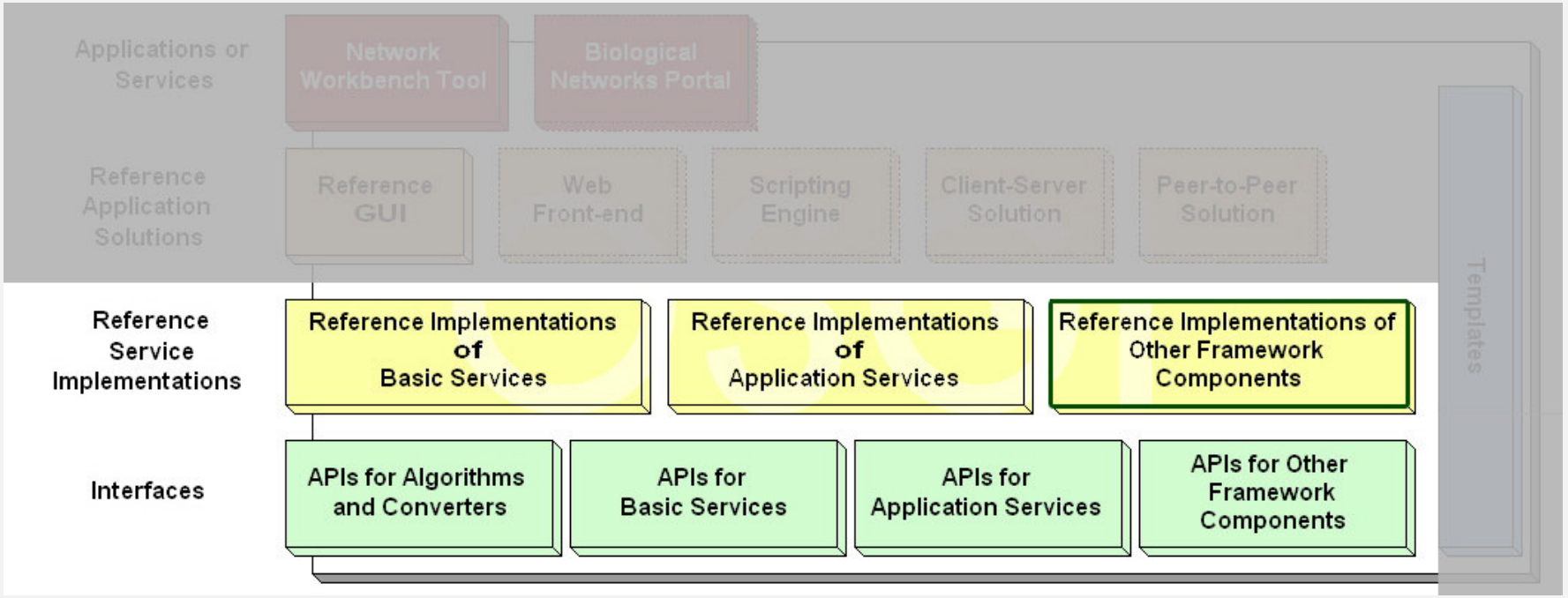
- Scheduler Service
- Data Manager Service



Services Layer – Other Components

Other Framework Components

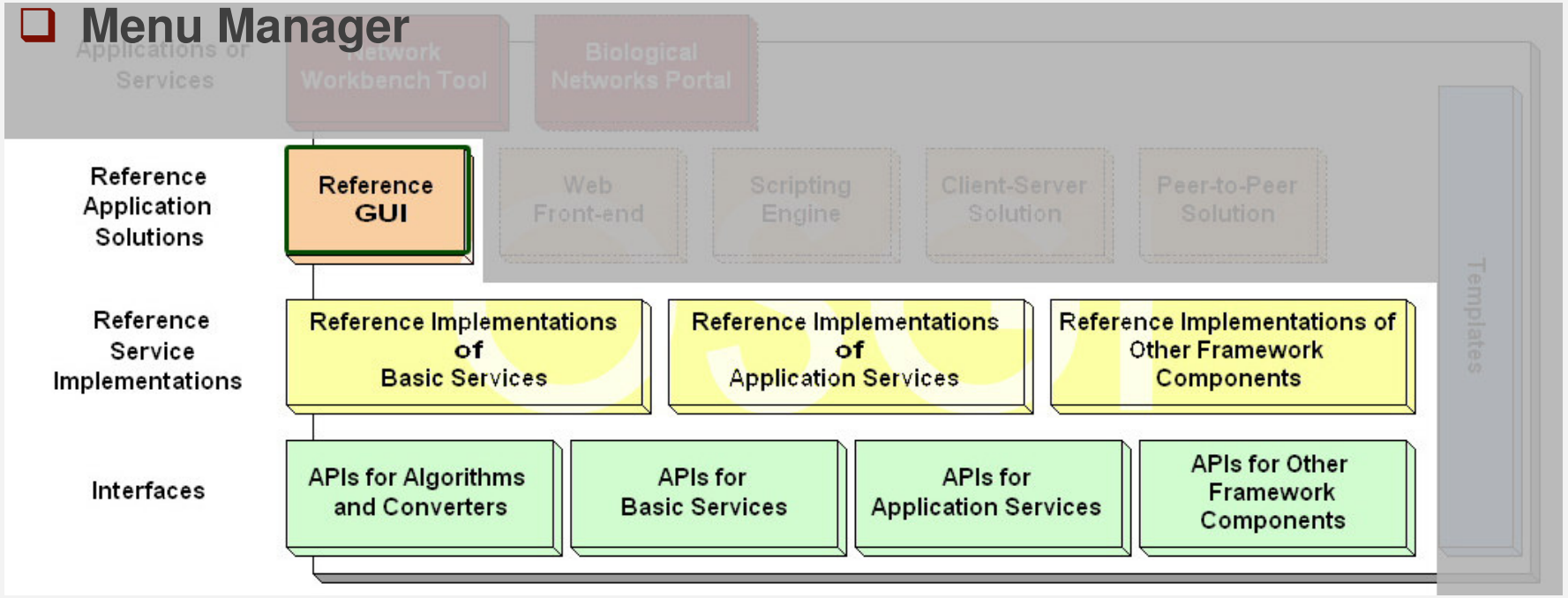
- CIShellContext - LocalCIShellContext
- Data - BasicData



Application Solutions

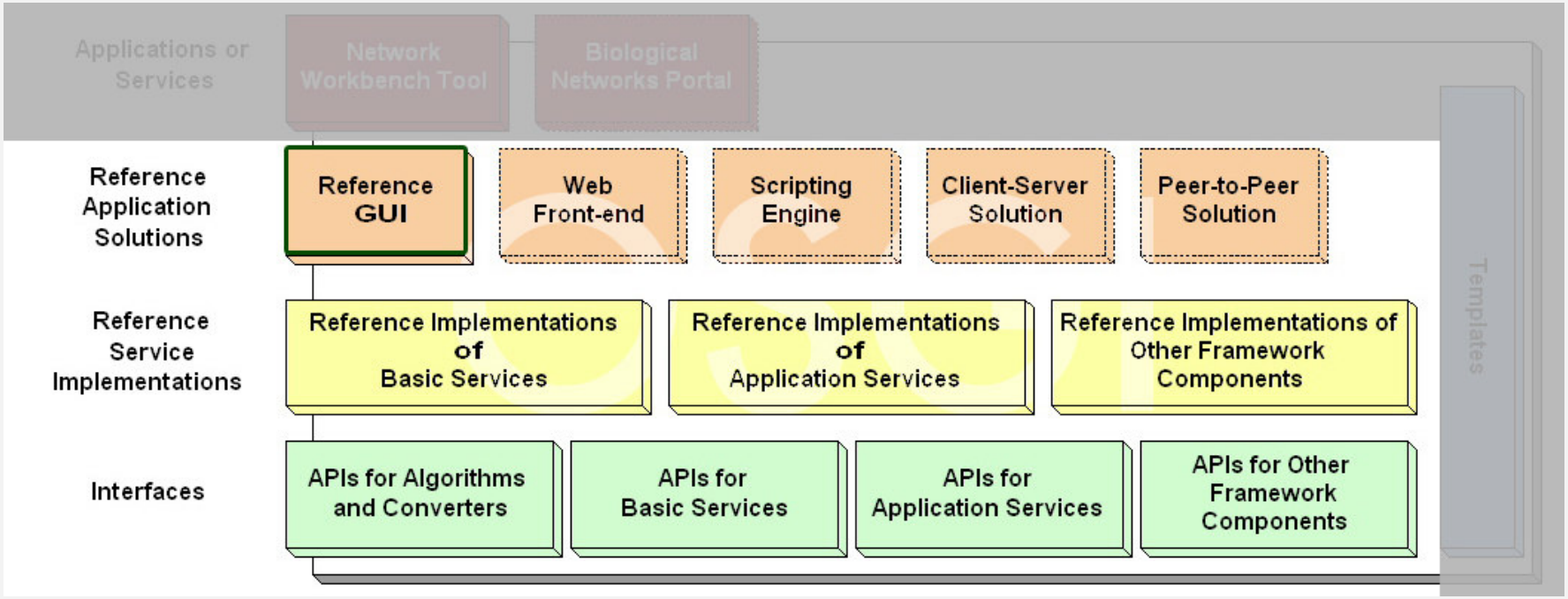
Reference GUI (using Eclipse RCP)

- Framework View
- Data Manager View
- Console(log) View
- Scheduler View
- Menu Manager



Application Solutions (cont.)

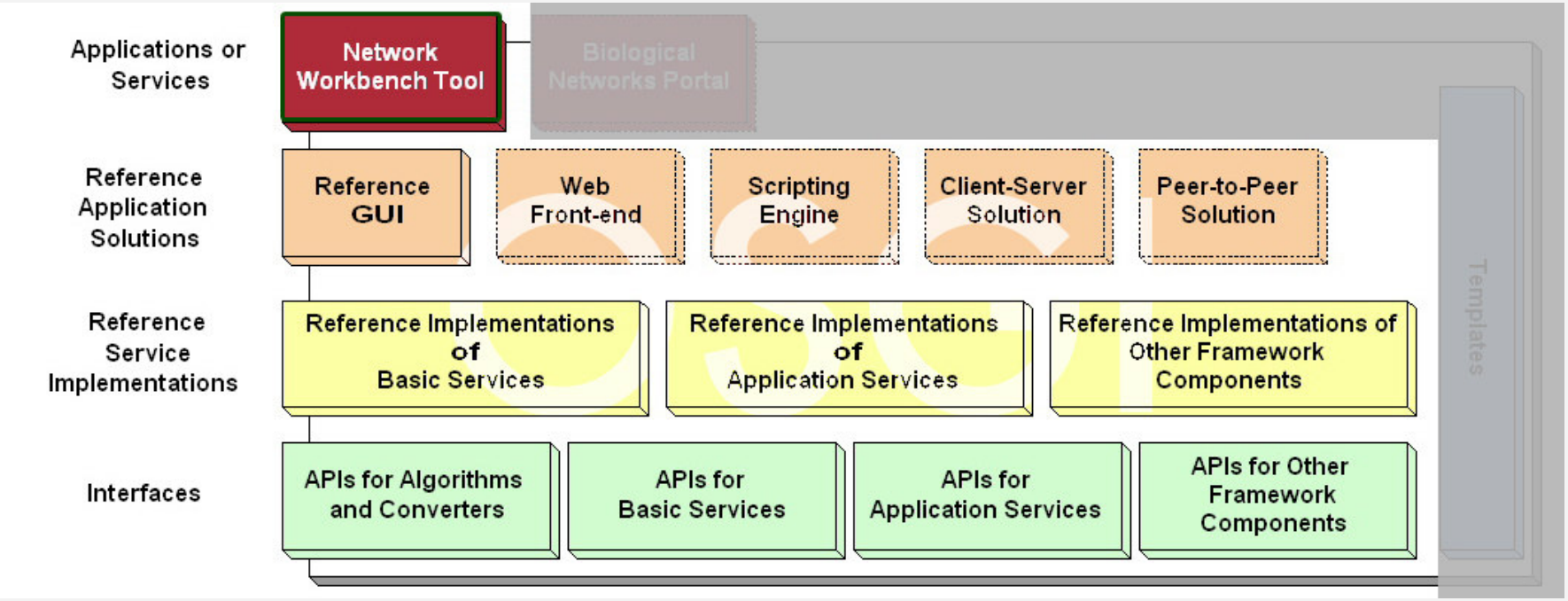
Other application solutions



Applications

NWB Tool

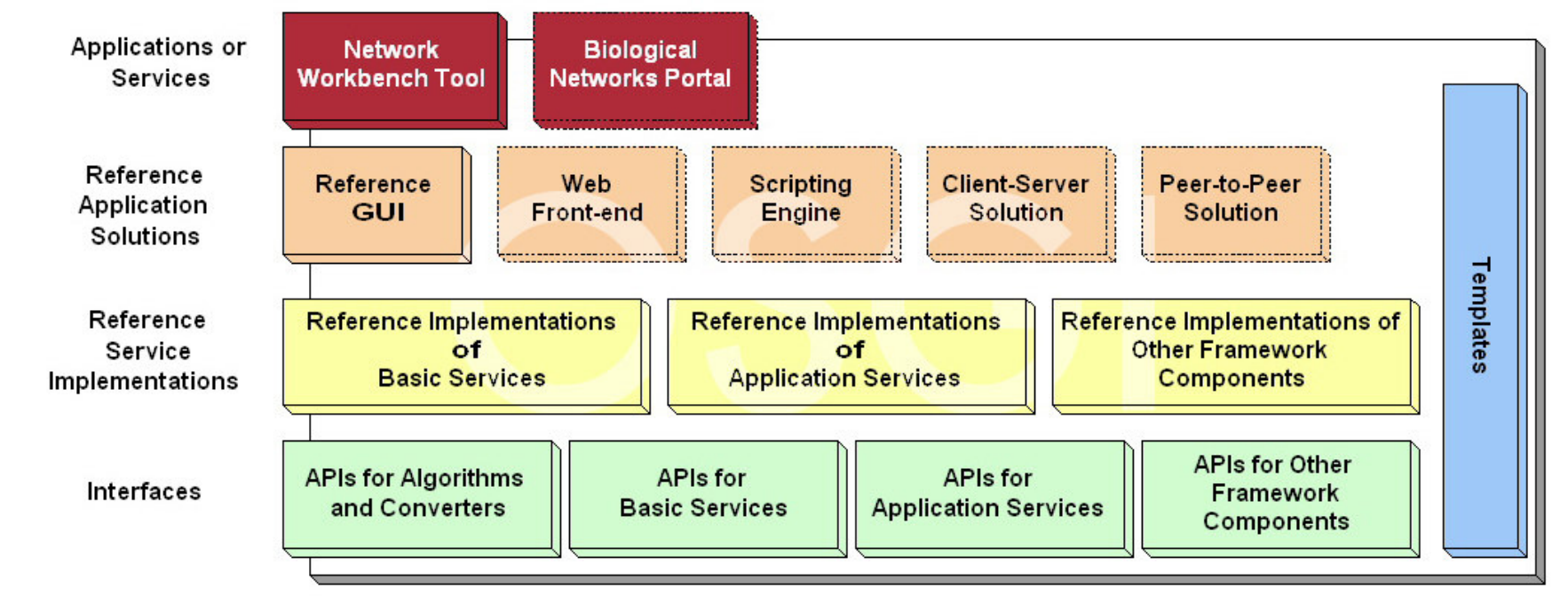
- Analyze, visualize and model network/graph
- Support most popular data formats and data conversion among them
- Serve three communities with different practices



Applications (cont.)

Biological Networks Portal

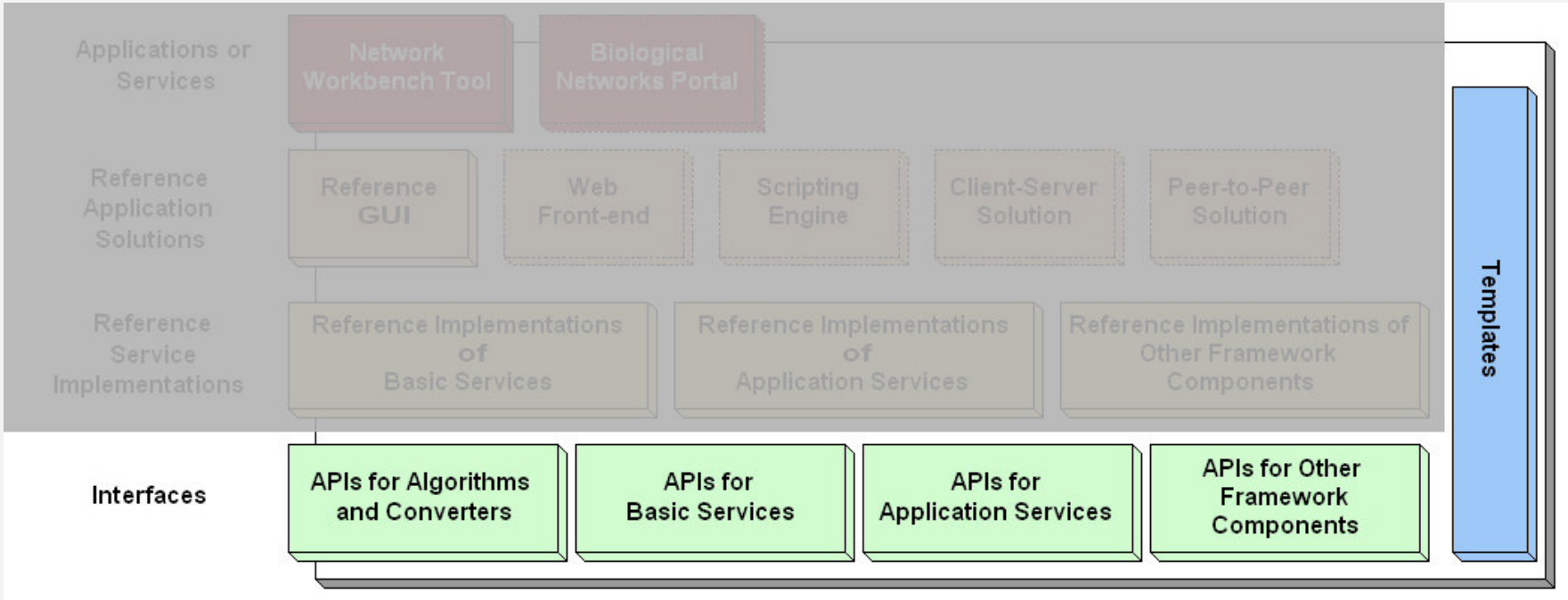
- Use Web front-end solution
- For educational purpose



Algorithm Developers Need to Know

For Algorithm Developers (Java-based)

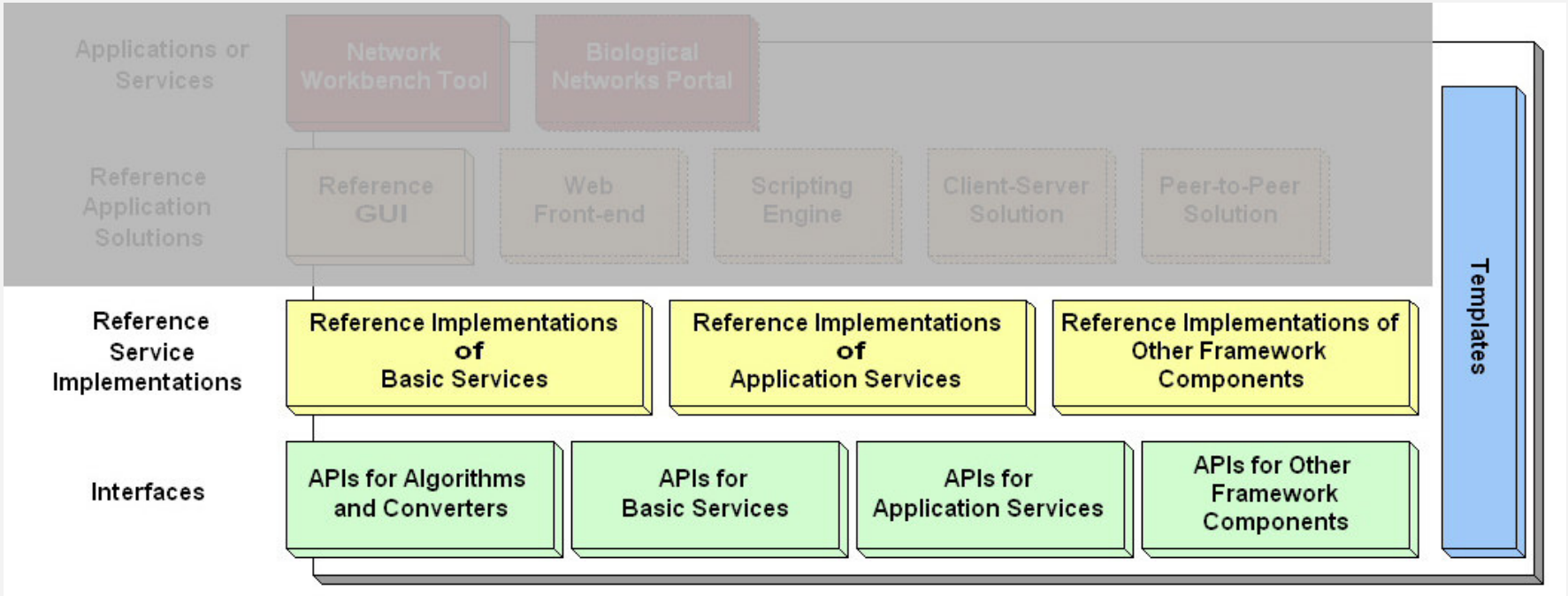
- Must implement CIShell Algorithm APIs
- Know how to use Basic Serivces APIs, Application Serivces APIs, CIShellContext, and Data APIs, but don't need to take care of the detail implementations of those services or components.



Application Designers Need to Know

Component Level

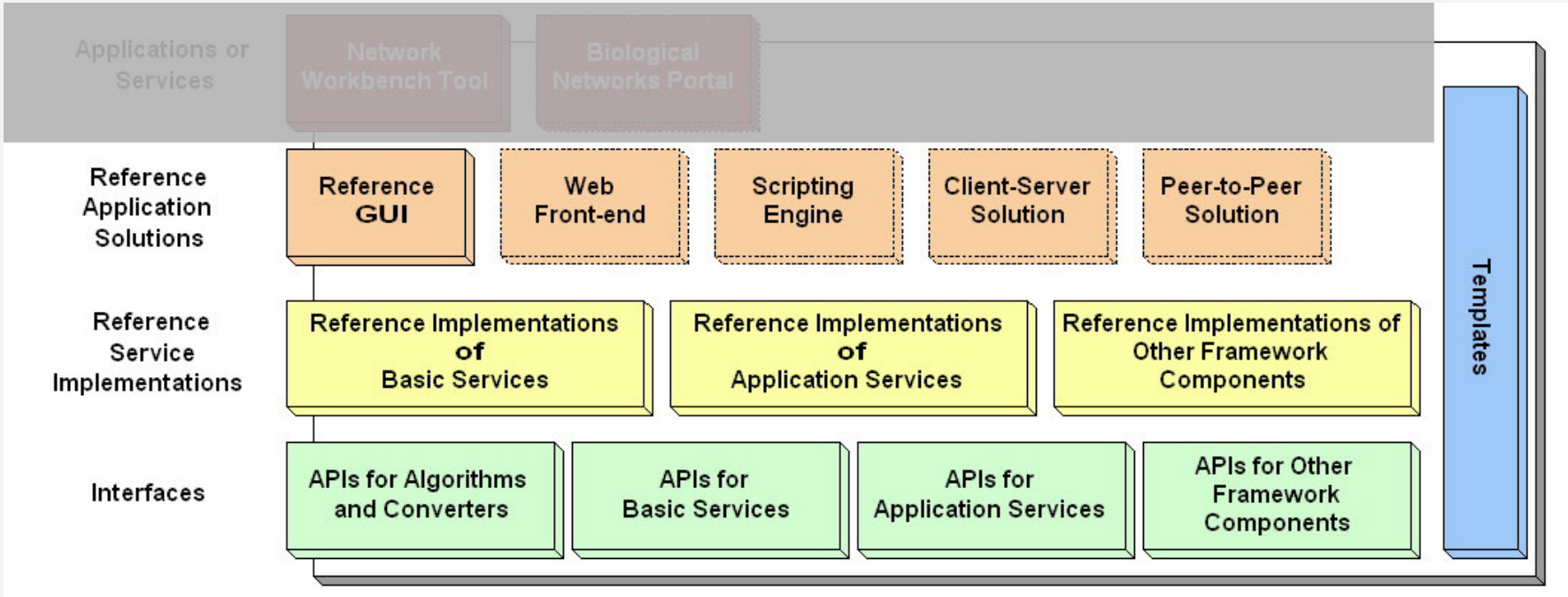
- Using OSGi service implementations from different vendors
- Each service/component can have more than one implementations



Application Designers Need to Know

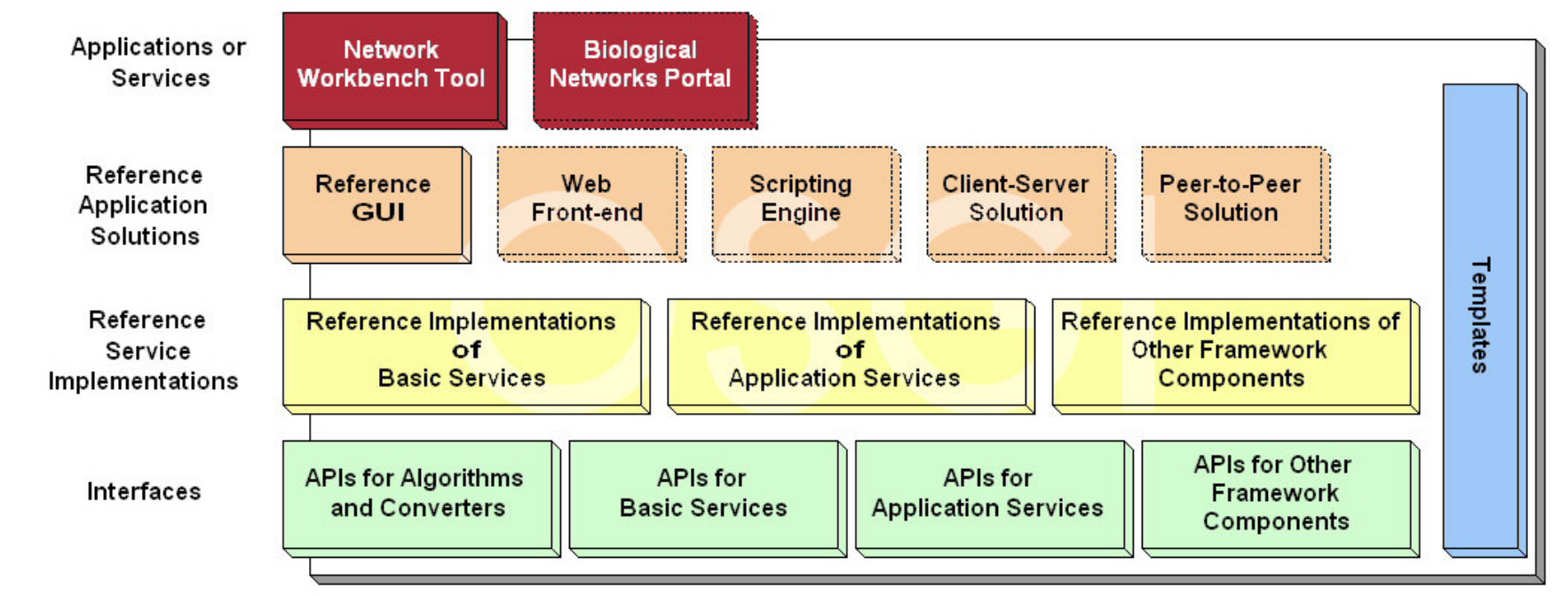
Framework Level

- Use all implementations of algorithms and converters
- Use all implementations on the service layer
- Concentrate on application solutions
- Use or refer to the reference implementations of an application



Application Users

- Get the most efficient algorithm implementations
- Get as many algorithms as needed
- Have tools running on multiple platforms and various application solutions
- Don't worry about the match between the data format of a dataset vs. algorithm input



Summary

NWB tool and CIShell provide

- A testbed for diverse algorithm implementations
- A mechanism to quickly integrate an algorithm and disseminate it through the NWB tool and community wiki.
- A bridge between what algorithm developers can provide and what application users need.

NWB Tool: Demo and Hand-on

Load Data

Select Preferences

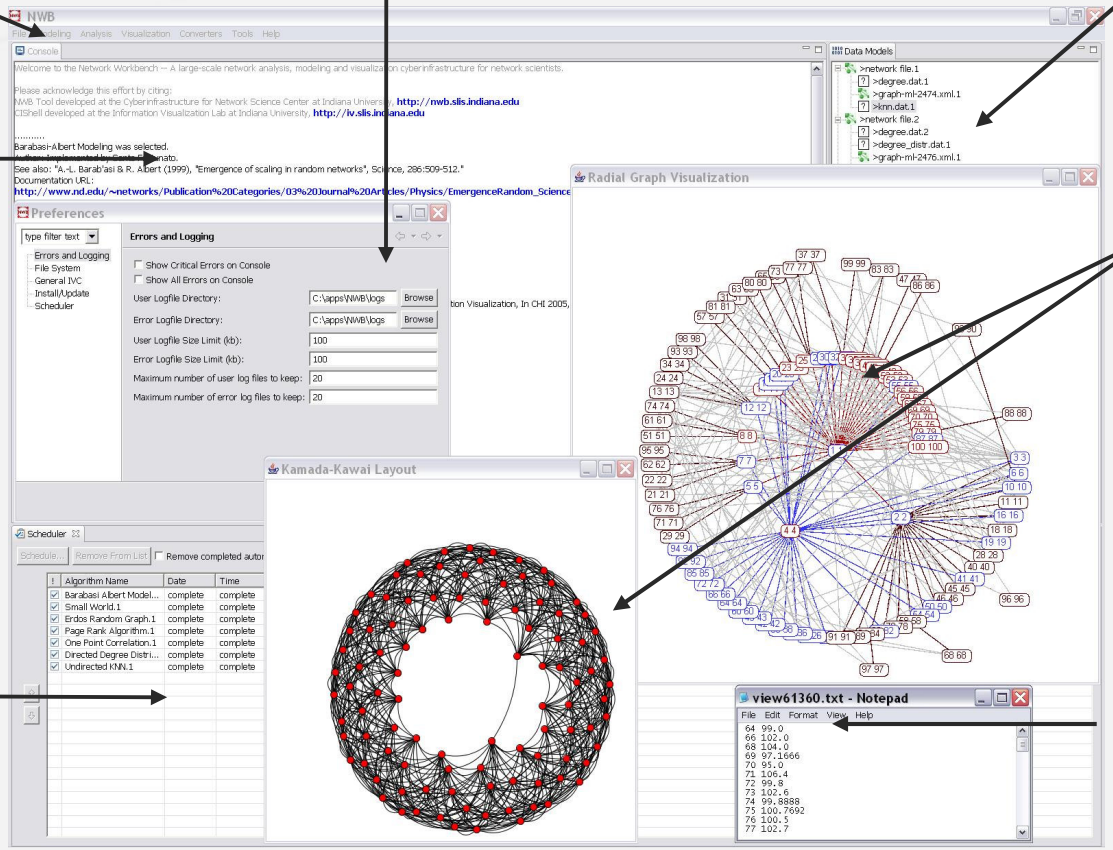
List of Data Models

Console

Visualize Data

Scheduler

Open Text Files



Examples

- A Map of Science (800,000 published papers)
- An Emergent Mosaic of Wikipedian Activity (659, 388 interconnected Wikipedia articles, 16,582,425 links)
- Movies and Actors: Mapping the Internet Movie Database (302,691 movies, 896,308 unique actors, 3,792,390 links)

Network Analysis and Visualization (cont.)

Towards Large Scale Network Analysis and Visualization

- Visualization Challenges
- Overlay on a base map (Google Map, Science Map)
- Network Dynamics

References

- Herr, Bruce W., Huang, Weixia, Penumarthy, Shashikant, & Börner, Katy. (2007). Designing Highly Flexible and Usable Cyberinfrastructures for Convergence, In William S. Bainbridge and Mihail C. Roco (Eds.) Progress in Convergence – Technologies for Human Wellbeing. Annals of the New York Academy of Sciences, Boston, MA, Volume 1093, pp. 161-179.
- Börner, Katy, Sanyal, Soma and Vespignani, Alessandro. (2007). Network Science: A Theoretical and Practical Framework. (in press) In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology*, Volume 41, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, chapter 12, pp. 537-607.
- Börner, Katy, Penumarthy, Shashikant, Meiss, Mark and Ke, Weimao. (2006). Mapping the Diffusion of Scholarly Knowledge Among Major U.S. Research Institutions. *Scientometrics*. 68(3), pp. 415-426.
- Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003). Visualizing Knowledge Domains. In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology*,Volume 37, Medford, NJ: Information Today. Inc./American Society for Information Science and Technology, chapter 5, pp. 179-255

References (cont.)

- Ketan Mane and Katy Börner. (2004) Mapping Topics and Topic Bursts in PNAS. PNAS, 101(Suppl. 1):5287-5290. Also available as cond-mat/0402380.
- Kevin W. Boyack, Richard Klavans , W. Bradford Paley , Katy Börner Mapping, Illuminating, and Interacting with Science one of the 96 accepted (out of 500 submitted) Siggraph 07 sketches.
- Holloway, Todd, Bozicevic, Miran, and Börner, Katy. (2007) Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors. Complexity, Special issue on Understanding Complex Systems. 12(3), pp. 30-40. Also available as cs.IR/0512085.
- Bruce W. Herr, Weimao Ke, Elisha Hardy & Katy Börner (2007). Movies and Actors: Mapping the Internet Movie Database. Submitted to Information Visualisation Conference, ETH Zürich, Switzerland.

Comments & Questions

Websites

- http://nwb.slis.indiana.edu
- https://nwb.slis.indiana.edu/community
- http://cishell.org
- http://cns-trac.slis.indiana.edu/trac/

Thank You