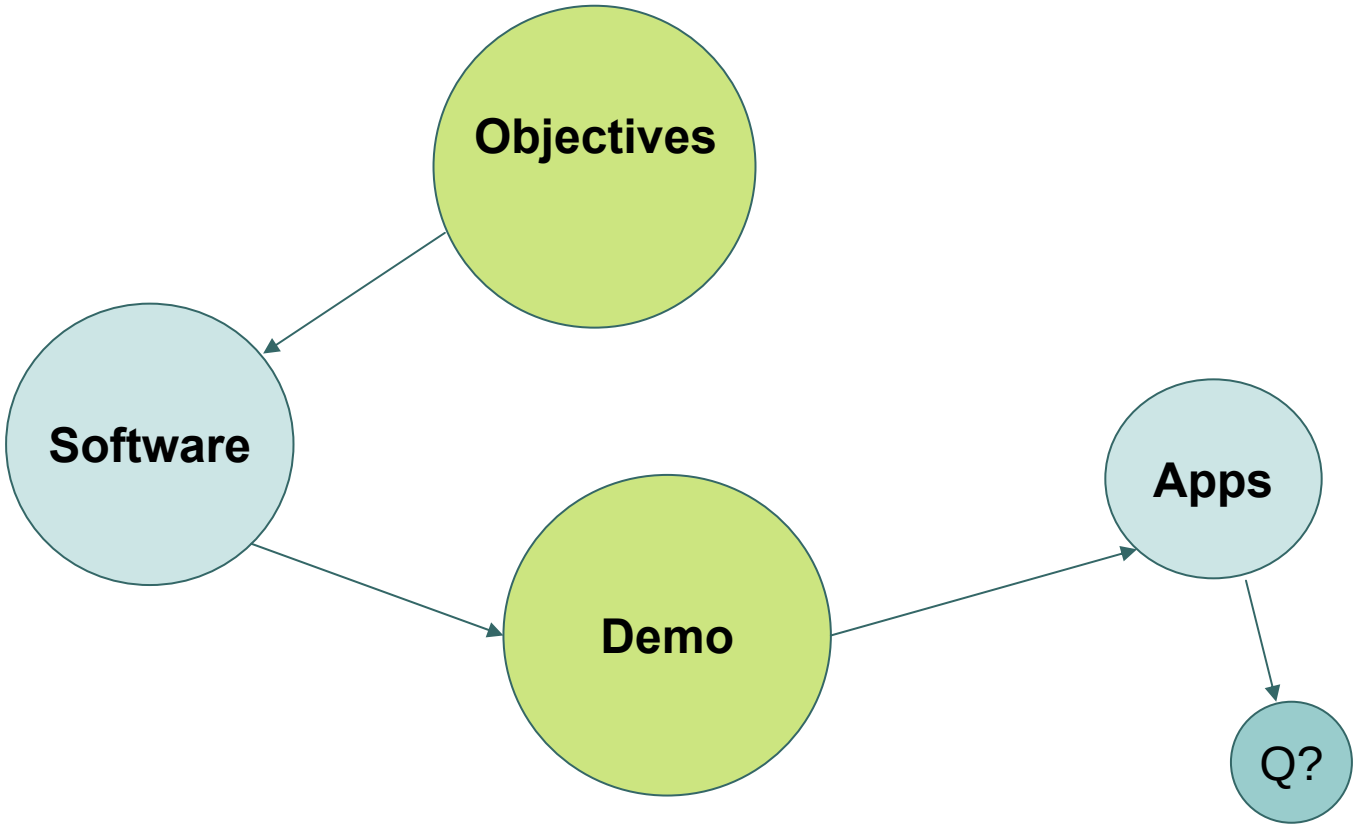
Constructing Data Mining Applications based on Web Services Composition

Ali Shaikh Ali and *Omer Rana*
Ali.shaikhali@cs.cf.ac.uk, o.f.rana@cs.cardiff.ac.uk

Cardiff University
http://www.cs.cf.ac.uk/

Welsh eScience Center
http://www.wesc.ac.uk/

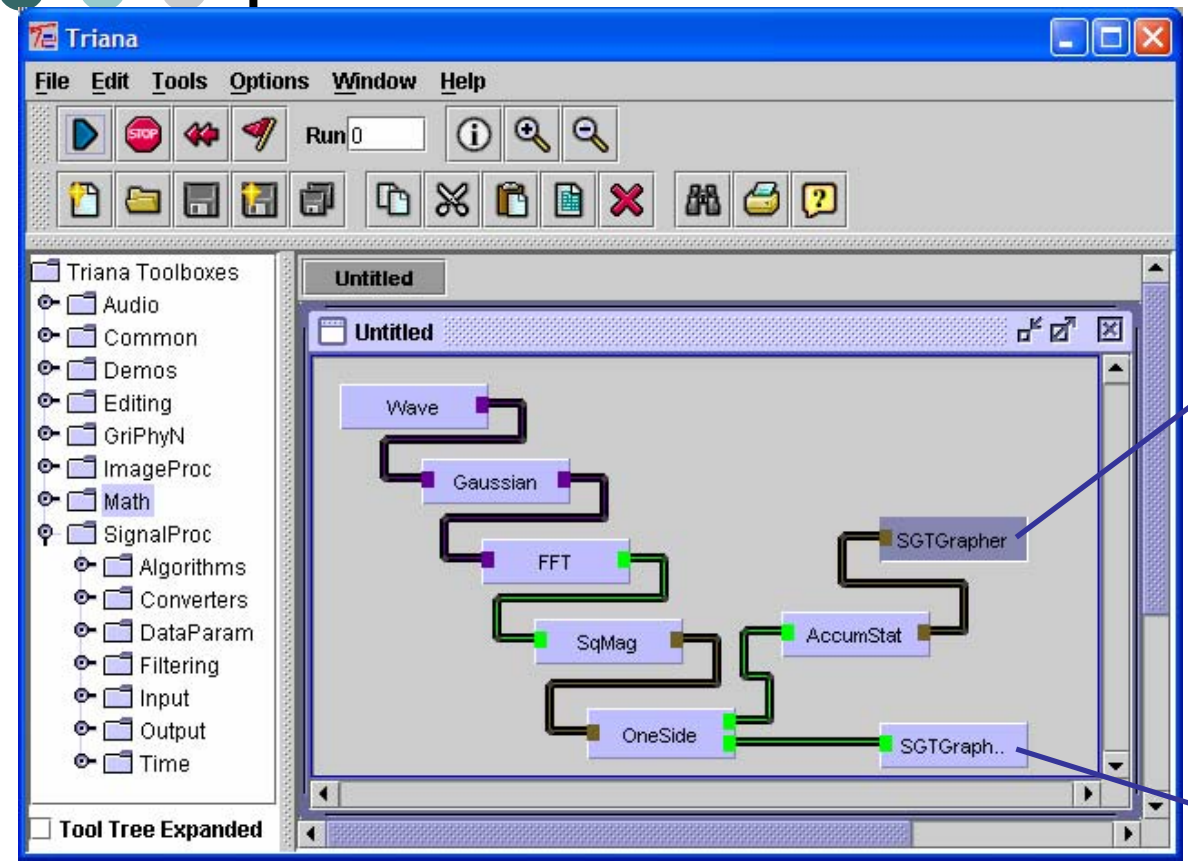
Agenda…

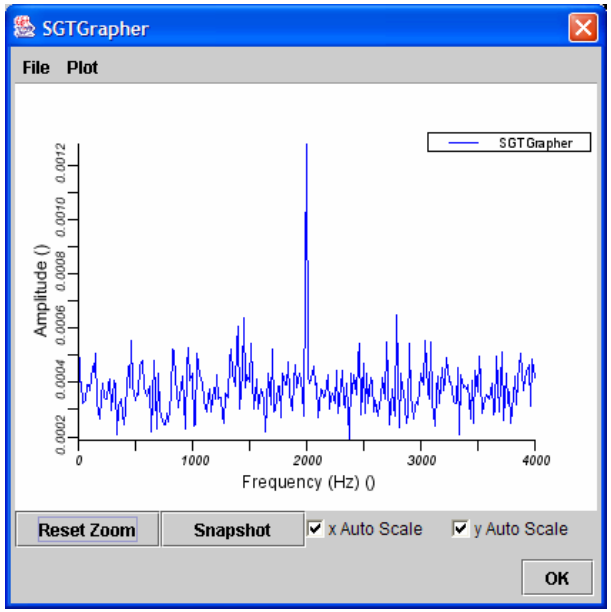


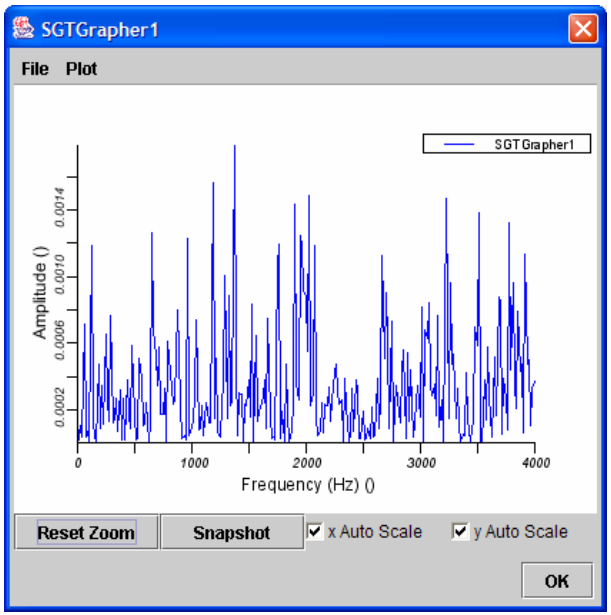
Objectives

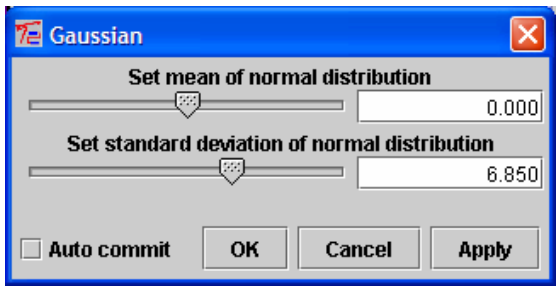
- Use of Web Services composition – with distributed services
  - Wrap third party services (Mathematica, GNUPlot)
  - WEKA Service template
  - Triana Workflow
- Services provided by third parties
  - WSDL interfaces (avoid use of specialist languages – unless really necessary)
  - SOAP-based message exchange
- Access to local and remote data sets
  - Support for data streaming

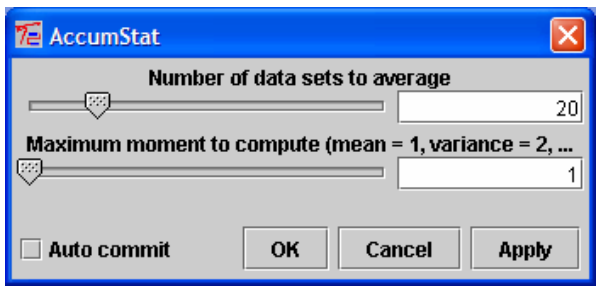
Origin: Gravitational Wave data analysis (GEO-LIGO efforts)



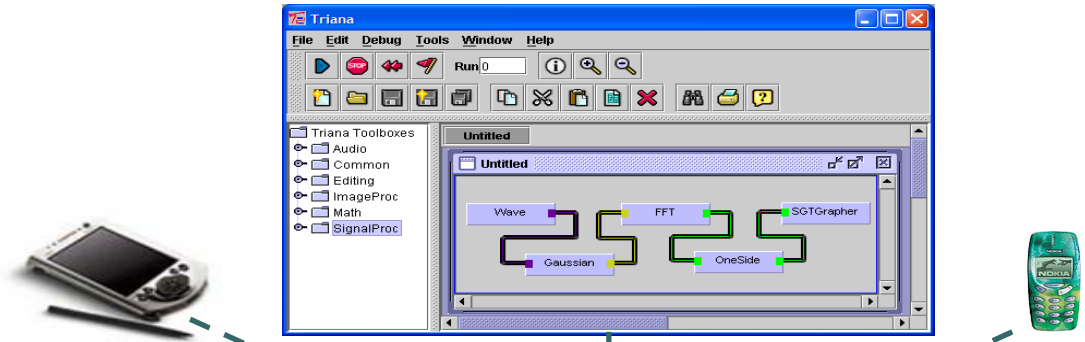








http://www.GridLab.org/



GAP Interface

GAT

JXTAServe

P2PS

WServe

Gridlab Services

JXTA

Sockets

Web Services

OGSA + Services

Software

Related work: Grid WEKA (University College Dublin)

trianacode.org

- An open source Problem Solving Environment developed at Cardiff
- Triana includes a large library of pre-written analysis tools and the ability for users to easily integrate their own tools.
- Supports discovery of Web Services based on syntax (hardwired UDDI registries)

www.cs.**waikato**.ac.nz/ml/**weka**/

- Collection of machine learning algorithms
- Contains tools for
  - data pre-processing,
  - classification, regression,
  - clustering,
  - association rules
- Accepts ARFF (Attribute-Relation File Format) file format -- an ASCII text file that describes a list of instances sharing a set of attributes.

WEKA Algorithms

- Classifiers Algorithms
  - Bayes (8, eg. Naïve Bayes)
  - Functions (12, eg. Neural Networks)
  - Lazy (5)
  - Meta (23, eg. Bagging, Multiclass Classifer)
  - Trees (10, eg. ID3)
  - Rules (10, eg. Conjunctive Rule)
  - Misc (3)
- Clustering Algorithms (5, e.g. K-means)
- Association Rules (2, e.g Apriori)
- Data Processing
  - Filters
- Attribute Selection
  - Attribute Evaluator (12, eg. Principle Components)
  - Attribute Search (8, eg. Genetic Algorithm)

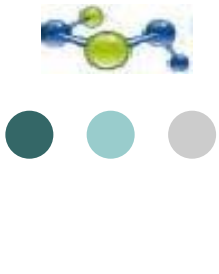
Provenance Issues

- EU FP6 Provenance project (2004—2006)
  - IBM Hursley (lead), SZTAKI, Southampton University, DLR/German Aerospace, UPC

  http://www.gridprovenance.org/

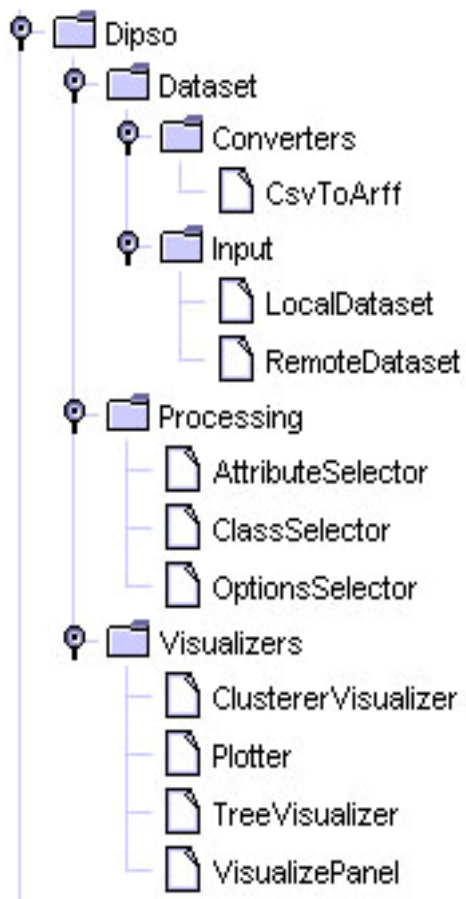
- EPSRC Provenance (2004—2007)
  - University of Southampton (lead)

  http://www.pasoa.org/



DEMO

Inside the Data Mining Toolbox



Deals with Dataset files
e.g. load datasets,
converts dataset formats

Deals with data that need manipulation

Visualize data

Adding new Classifier Service

- **Classifier Template**

  This Web service implements a complete list of classifiers, i.e. trees, rules, functions etc.

→ **Operations**
classifyInstance()

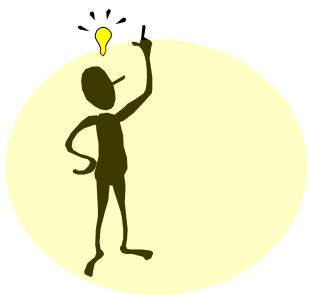
→ **classifyRemoteInstance()**

→ **getClassifiers( )**

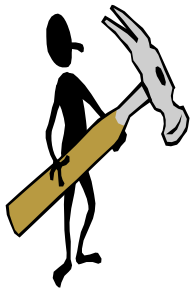
→ **getOptions()**

Input: String *datasetURL*
String *classifierName*
String *options*
String *attributeName*

output: String *result*

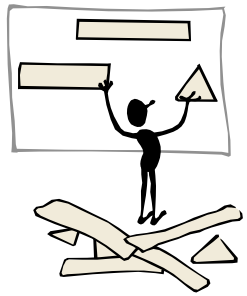


Adding new Services … 2

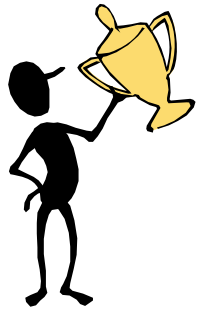


1. Build your classifier

must implement the 4 required methods



2. Place it in the classifier's lib.

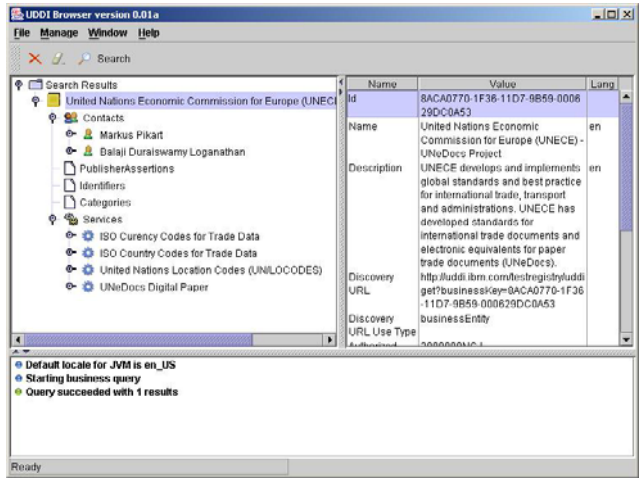


3. Done!

Where can you find us?

UDDI Browser

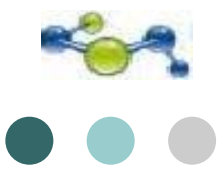
An open-source project that provides a friendly user interface allowing users to browse and manipulate content in UDDI registries. It is written in Java using the Swing libraries. Currently the browser only supports version 2.0 UDDI registries.



Cardiff UDDI:

Inquiry: http://agents-comsc.grid.cf.ac.uk:8334/juddi/inquiry

Publish: http://agents-comsc.grid.cf.ac.uk:8334/juddi/inquiry



Download

- Triana available at:

http://www.trianacode.org/

http://www.gridlab.org/

- Data Mining Toolbox at:

http://users.cs.cf.ac.uk/Ali.Shaikhali/dipso/

Questions

- Who is part of the user community?
  - elicit requirements
- What is different with reference to e-Science?
  - additional capability provided by the Grid
  - additional types of requirements
  - What additional benefit does it provide?
  - Ability to undertake multiple runs (what-if scenarios)
- Need to embed algorithms within some other program -- rather than have a stand-alone tool
  - Can Web Services try to address this concern?
- Which algorithm in what context?