Provenance Support for Distributed Scientific Data Analysis

Ali Shaikh Ali and Omer F. Rana
School of Computer Science & Welsh eScience Centre
Cardiff University, UK

o.f.rana@cs.cardiff.ac.uk

Provenance Definition

The provenance of a piece of data is the process that led to that piece of data

- We represent the provenance of some data by *documenting* the process that led to the data:
  - documentation can be complete or partial;
  - it can be accurate or inaccurate;
  - it can present conflicting or consensual views of the actors involved;
  - it can provide operational details of execution or it can be abstract.

From: Luc Moreau
(U Southampton)

Questionnaire distribution

- 47 projects

Questionnaire feedback

- 11 projects, many with real deployed services
  - Organ Transplant Management application,
  - TENT system,
  - eDiamond project,
  - Healthcare and Life Sciences Framework,
  - CombeChem,
  - myGrid,
  - GENSS,
  - Traffic Management Application (K-WFGrid project),
  - DataMiningGrid,
  - UniGridS,
  - Diligent
- Questionnaire, application descriptions, scenarios, Provenance usage

Data Analysis Requirements

- Data formats
  - ARFF, CSV
- Data access (local and remote)
  - File vs. streamed
- Algorithm + choice of algorithm
- Multi-perspective analysis
  - Visual feedback, text output
- User support
  - Graphical interface
- Validation support

•DiscoveryNet (Imperial College) – Yike Guo et a.

•GridMiner (U Vienna) -- Peter Brezny et al.

•myGrid (Manchester) -- Carol Goble et al.

•VEGA (CNR) -- Domenico Talia et al.

Others:
•Clementine, Matlab-based
•Tools, IRIS-Explorer-based

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
    - Use of attachments
- Access to local and remote data sets
  - Support for data streaming
- Wrapping of existing algorithms (important requirement)

Software

trianacode.org

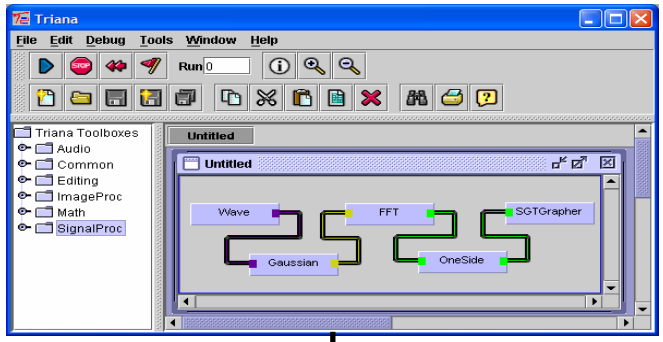
- An open source Problem Solving Environment developed at Cardiff
- Triana includes a large library of pre-written analysis tools and the ability for users to easily integrate their own tools.
- Supports discovery of Web Services based on syntax (hardwired UDDI registries)

Related work: Grid WEKA (University College Dublin)
WS-Weka (DEIS, Italy)

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.

http://www.GridLab.org/



GAP Interface

GAT

JXTAServe

P2PS

WServe

Gridlab Services

JXTA

Sockets

Web Services

OGSA + Services

WEKA Algorithms

- Classifiers Algorithms
  - Bayes (8, eg. Naïve Bayes)
  - Functions (12, eg. Neural Networks)
  - Lazy (5 e.g. Instance-based Learning)
  - 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)

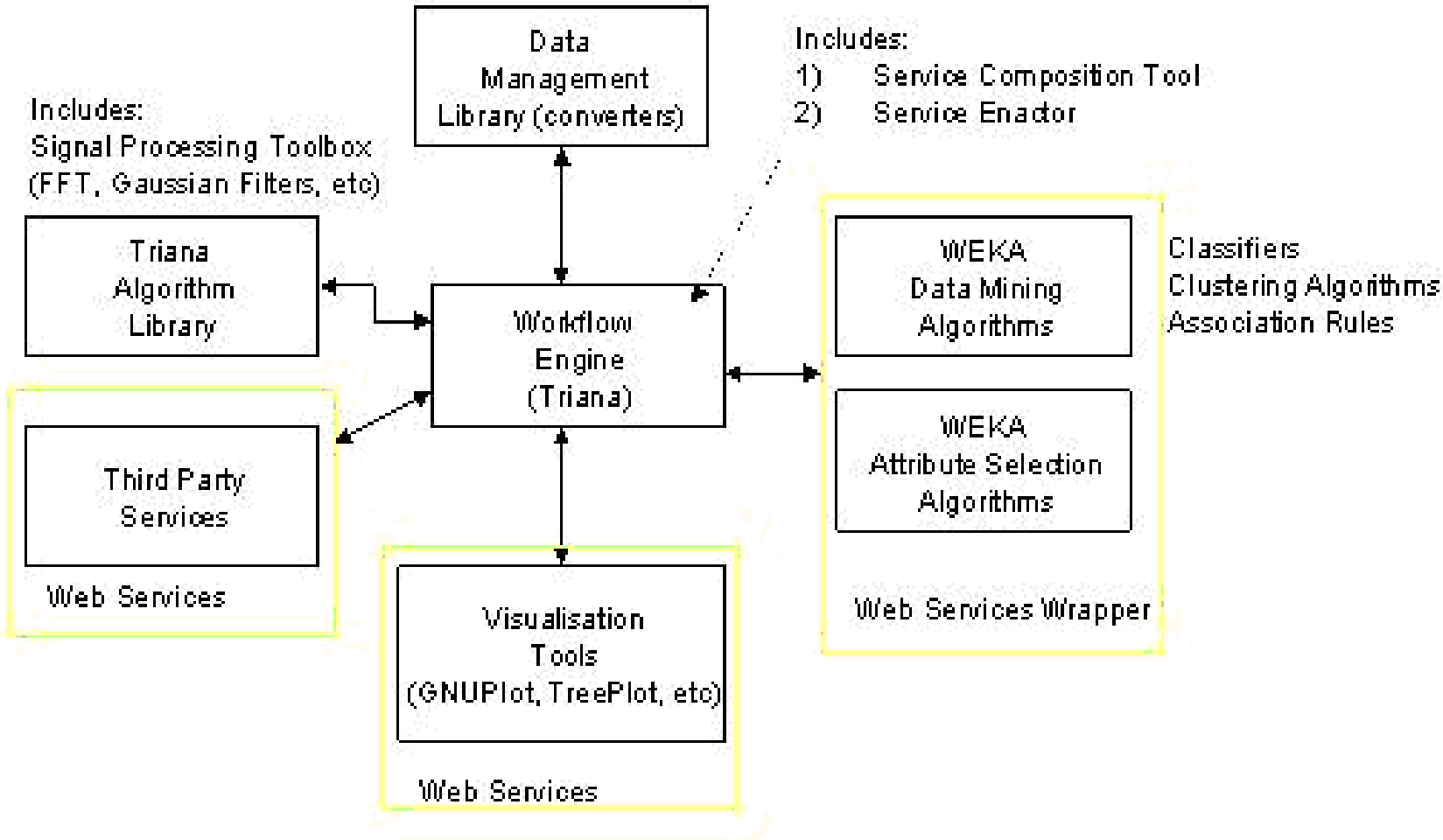
ARFF Overview

- Name of the relation
- List of the attributes
- Attributes types

```
@RELATION iris
@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
 @ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
 4.6,3.1,1.5,0.2,Iris-setosa
 5.0,3.6,1.4,0.2,Iris-setosa
```

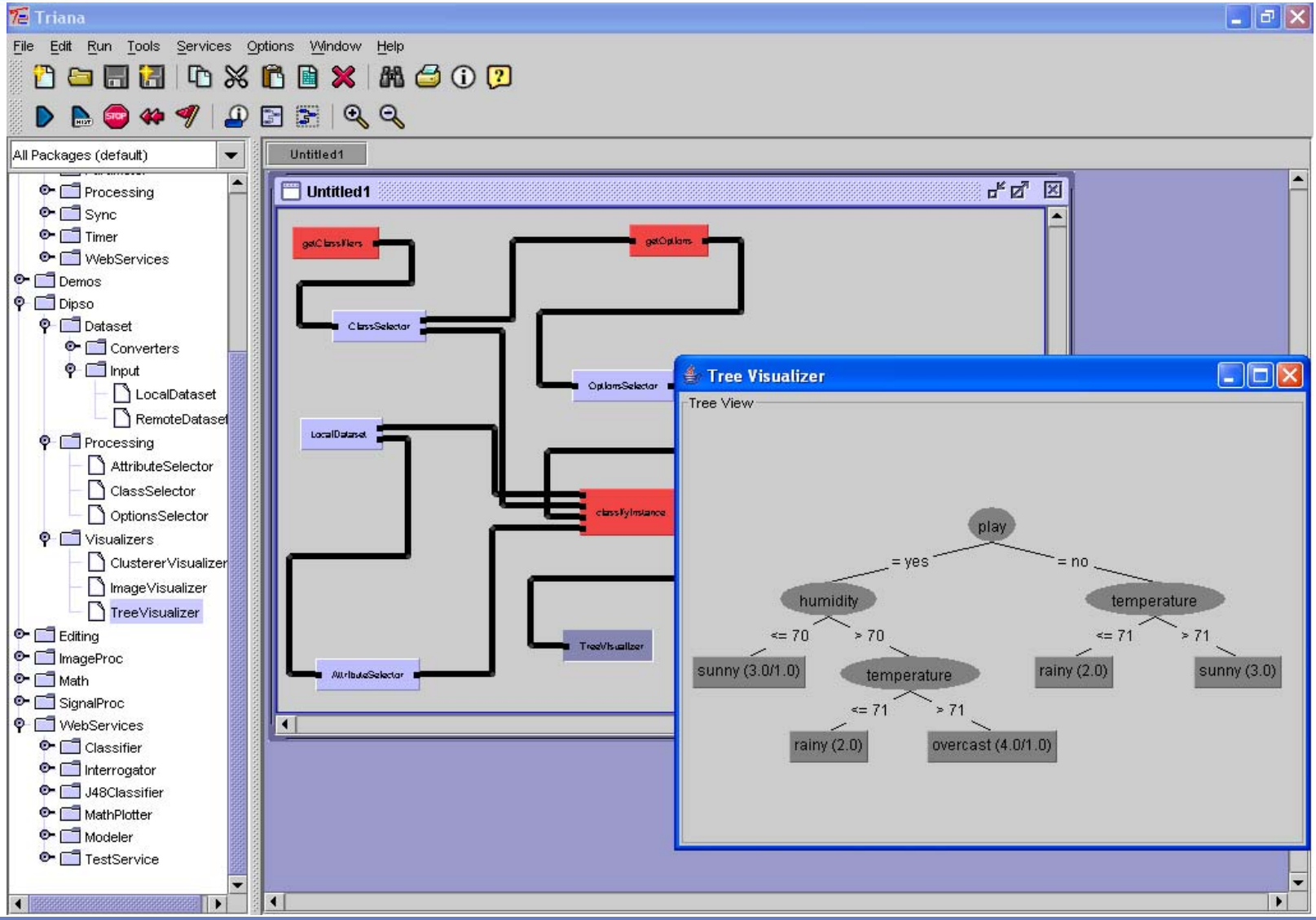
Architecture

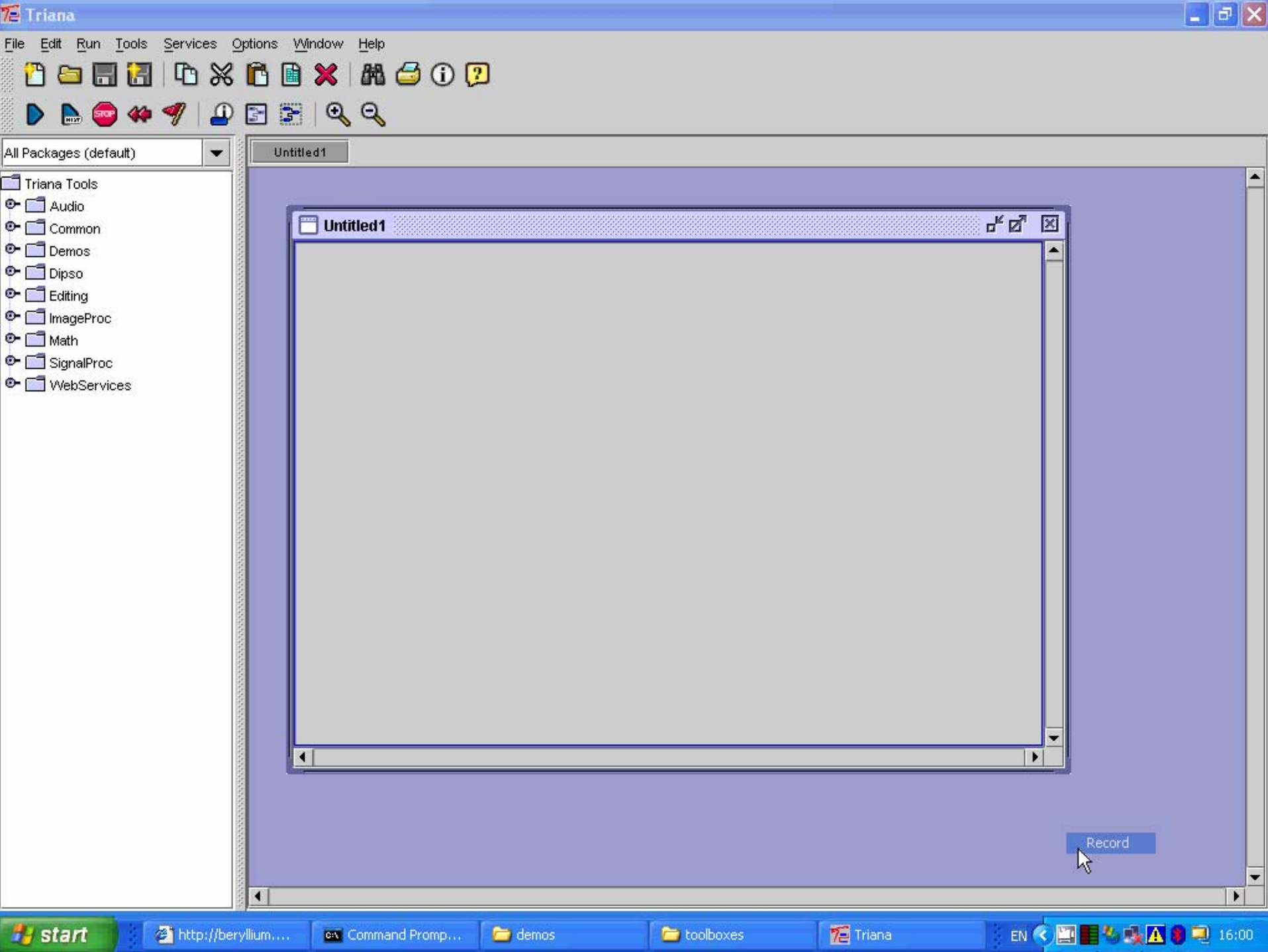


FAEHIM Web Services

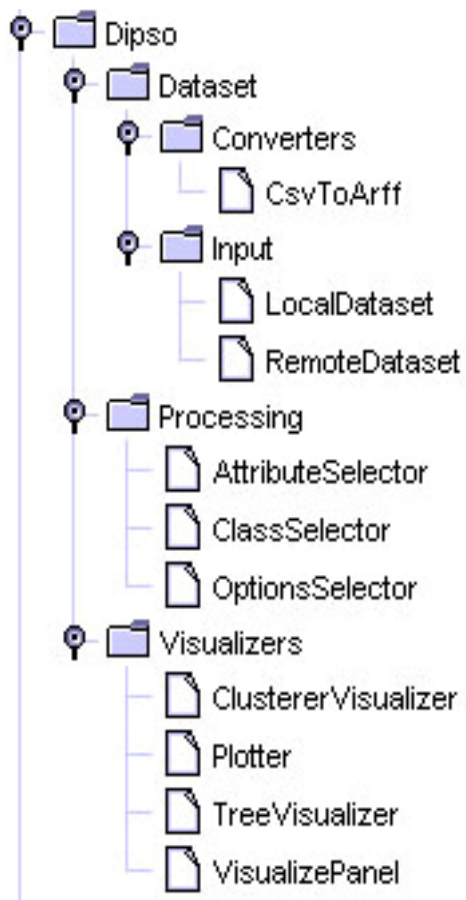
- Classification Web services
  - Obtained from the WEKA toolkit
- Clustering Web services
  - Obtained from the WEKA toolkit
- Associating rules Web services.
- Mathematical Web services
  - MathLink to Mathematica (uses a pre-defined notebook)
- Others (DIPSO Project):
  - Modeller
    - Uses a parametric analysis (not data analysis)
  - Interrogator
    - Space sampling (Taguchi search)

User Interface




Triana
File Edit Run Tools Services Options Window Help
All Packages (default)
Triana Tools
Audio
Common
Demos
Dipso
Editing
ImageProc
Math
SignalProc
WebServices
Untitled1
Untitled1
Record
start
http://beryllium....
Command Promp...
demos
toolboxes
Triana
EN
16:00

Inside the FAEHIM Toolbox

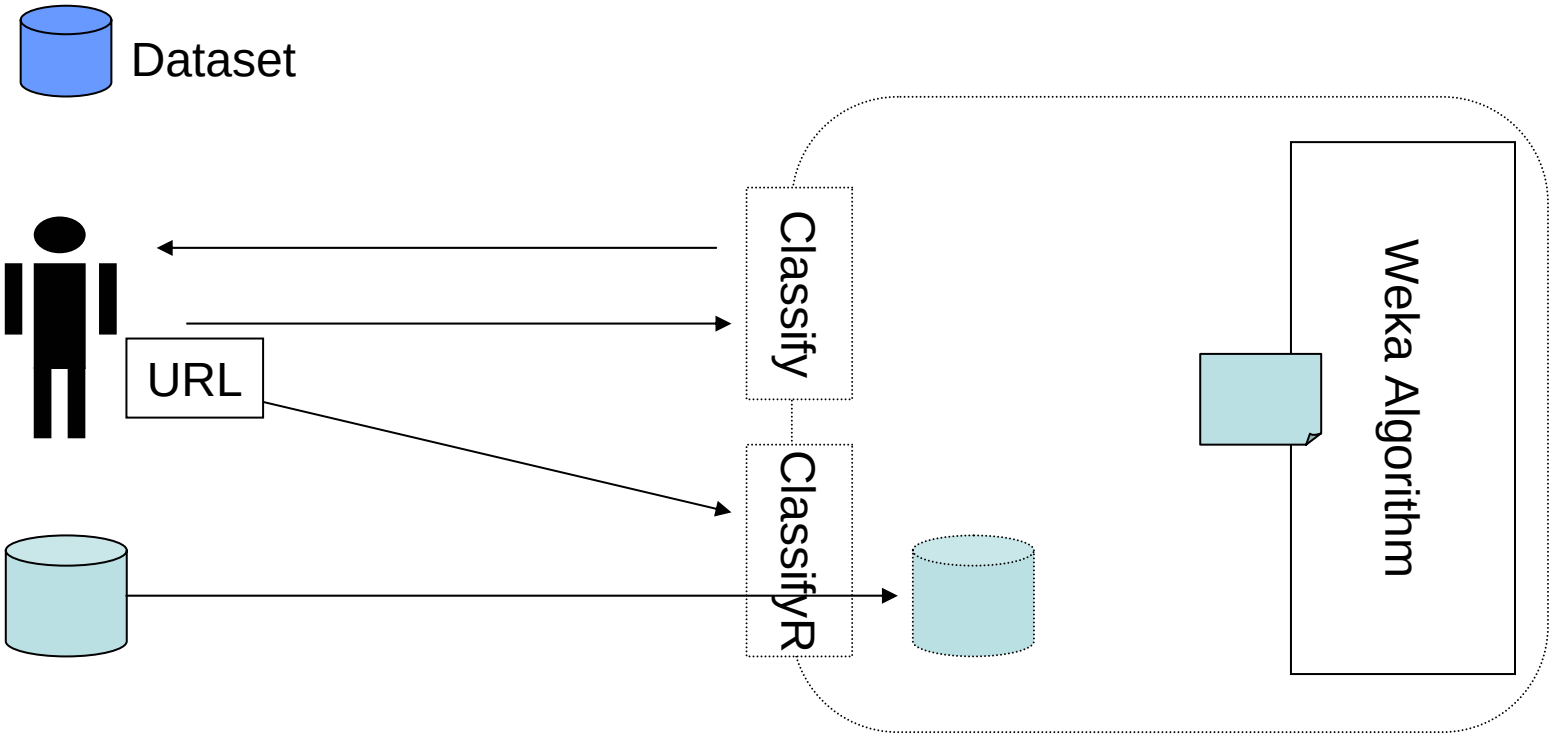


Deals with Dataset files
e.g. load datasets,
converts dataset formats

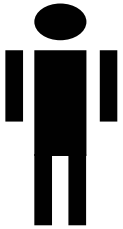
Deals with data that need manipulation

Visualize data

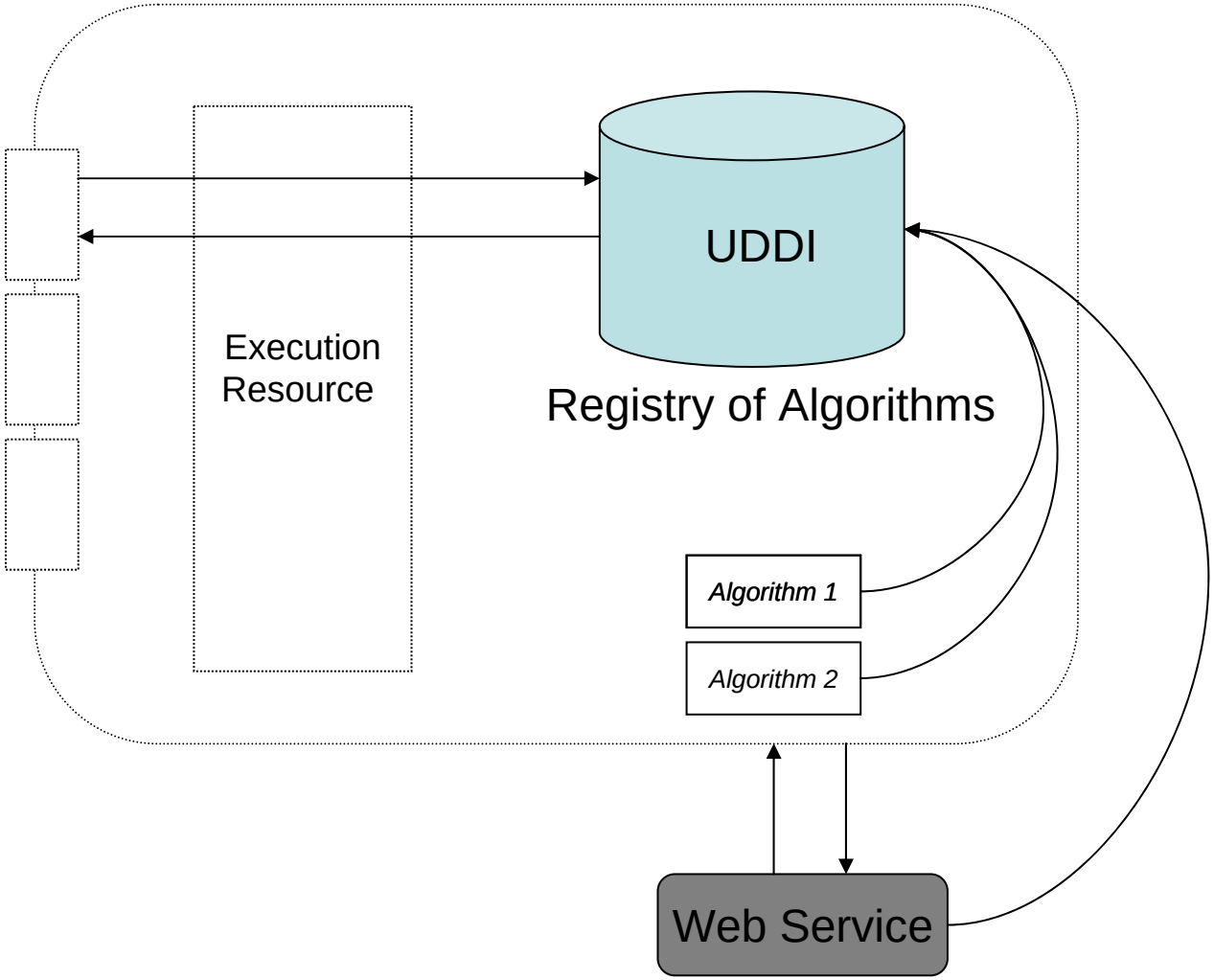
Usage Overview



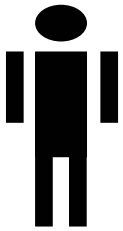
Registry Usage



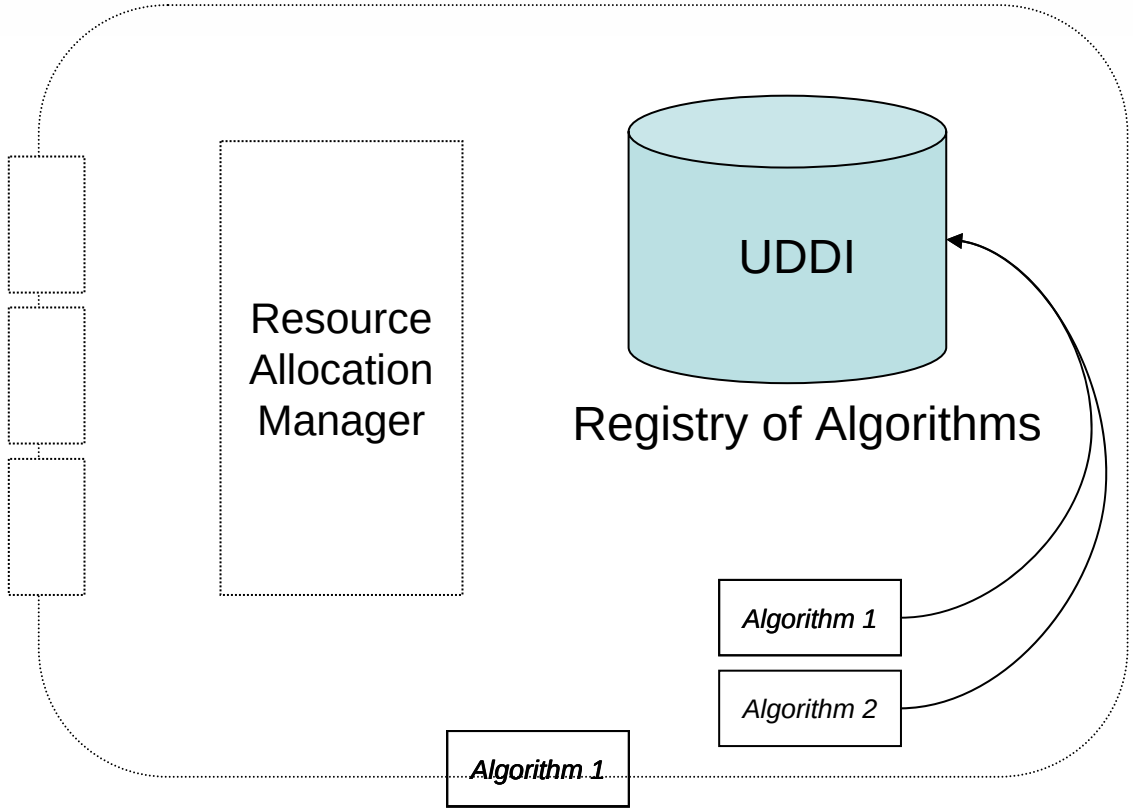
1. Discover
2. Select
3. Invoke

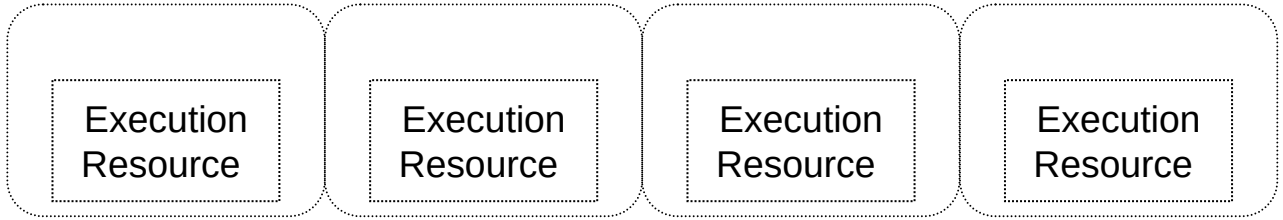


Parallel Execution



1. Discover
2. Select
3. Invoke





Provenance Questions

- After completion of workflow:
  1. Did the services I use actually fulfil my overall application requirement?
  2. Two of the analysis were performed on the same initial data but have different results – did I alter the services between these experiments?
  3. Did I perform each service on the type of data that the service was intended to analyse, i.e. were the inputs and outputs of each activity compatible?
  4. Did I use data sources from the same site?
  5. Why did it take much longer to run the analysis in the second instance?

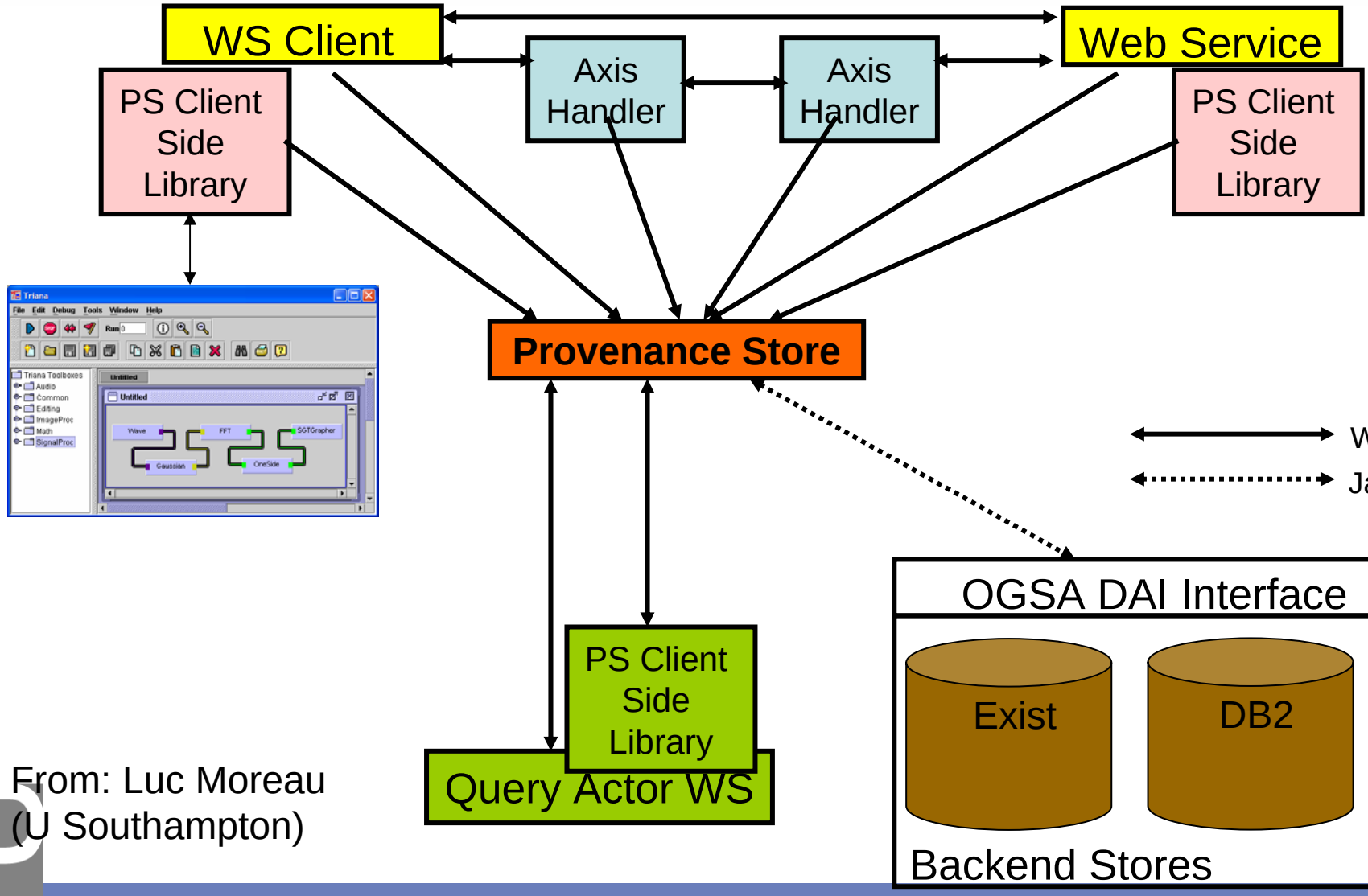
Particularly significant in the context of Distributed Services

p-assertion

- A given element of process documentation referred to as a p-assertion
  - p-assertion: is an assertion that is made by an actor and pertains to a process.
- Types
  - Interaction p-assertion
    - relates to content of received/sent message
  - Actor p-assertion
    - Relationships between actors
    - State of an actor

From: Luc Moreau (U Southampton)

Implementation Diagram



From: Luc Moreau
(U Southampton)

FAEHIM Download Available

- http://users.cs.cf.ac.uk/Ali.Shaikhali/faehim/
  - Standard GPL License
- Video of usage also available at above site
- Launched in 2004
  - eScience Data Mining SIG
  - 550 downloads (since June 2004)
  - http://www.datamininggrid.org/

Provenance Team

- University of Southampton
  - Luc Moreau, Victor Tan, Paul Groth, Simon Miles, Luc Moreau
- IBM Hursley UK (Project Coordinator)
  - John Ibbotson, Neil Hardman, Alexis Biller
- Cardiff University
  - Omer Rana, Arnaud Contes, Vikas Deora
- Universitad Politecnica de Catalunya (UPC)
  - Steven Willmott, Javier Vazquez
- SZTAKI
  - Laszlo Varga, Arpad Andics
- German Aerospace
  - Andreas Schreiber, Guy Kloss, Frank Danneman

http://www.gridprovenance.org/

Conclusion & Further Issues

- Integration of Triana workflow engine with Provenance System
- Access to Provenance Information
  - Via an API
  - Via a Portal Interface
- Ability to reconstruct workflow using Provenance information
- Access constraints defined by an administrator