The Linked2Safety's Galaxy Based Data Analysis Space

Aristos Aristodimou, Athos Antoniades, Constantinos Pattichis
University of Cyprus
David Tian, Ann Gledson, John Keane
University of Manchester

Linked2Safety Project (FP7-ICT-2011-7 – 5.3)
A NEXT-GENERATION, SECURE LINKED DATA MEDICAL INFORMATION SPACE FOR SEMANTICALLY-INTERCONNECTING ELECTRONIC HEALTH RECORDS AND CLINICAL TRIALS SYSTEMS ADVANCING PATIENTS SAFETY IN CLINICAL RESEARCH

GCC2013 Oslo
30 June – 2 July

Linked2Safety

Linked2Safety:

- FP7 project funded by the European Commission under the area of ICT for health

Vision:

- Advance clinical practice and accelerate medical research, by providing homogenized access to anonymized aggregated distributed EHRs, and the tools for analyzing such data.

Anonymity:

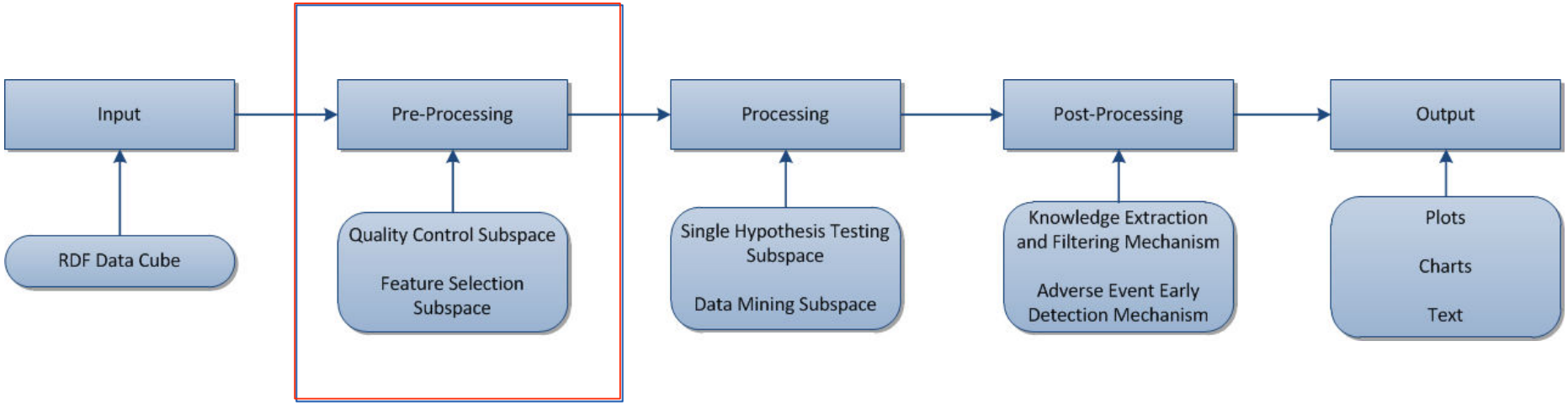
- Data cubes
- Perturbation
- Cell suppression

Linked2Safety's Data Analysis Space

Objectives:

- Design and develop the data mining techniques and the scalable infrastructure for the identification of phenotypic and genetic associations related to adverse events.
- Develop new and implement existing state of the art analytical approaches for genetic data.
- Define and implement the knowledge extraction and filtering mechanisms and the knowledge base
- Integrate the knowledge base into a lightweight decision support system (Adverse events early detection mechanism)

Data Analysis Steps



Quality Control Subspace

Provides the tools for identifying and removing erroneous data or data that do not conform to the quality standards that a user might define.

Tools:

- **Hardy-Weinberg Equilibrium Test**
- **Allele Frequency Test**
- **Missing Data Test**

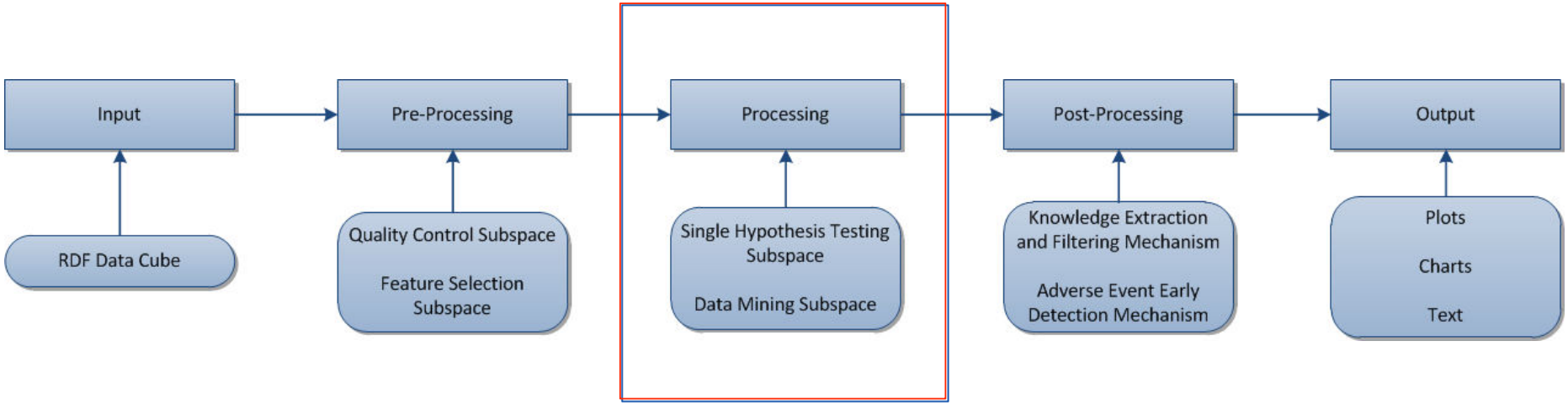
Feature Selection Subspace

Provides the tools for removing redundant or irrelevant features from a dataset.

Tools:

- **Rough Set Feature Selection**
- **Information Gain Feature Selection**
- **Chi Squared Feature Selection**

Data Analysis Steps



Single Hypothesis Testing Subspace

Provides the tools for performing single hypothesis testing on a dataset and test for associations.

Tools:

- **Pearson's Chi Square Test**
- **Fisher's Exact Test**
- **Odds Ratio**
- **Binomial Logistic Regression**
- **Linkage Disequilibrium**
- **Genetic Region Based Association Testing**

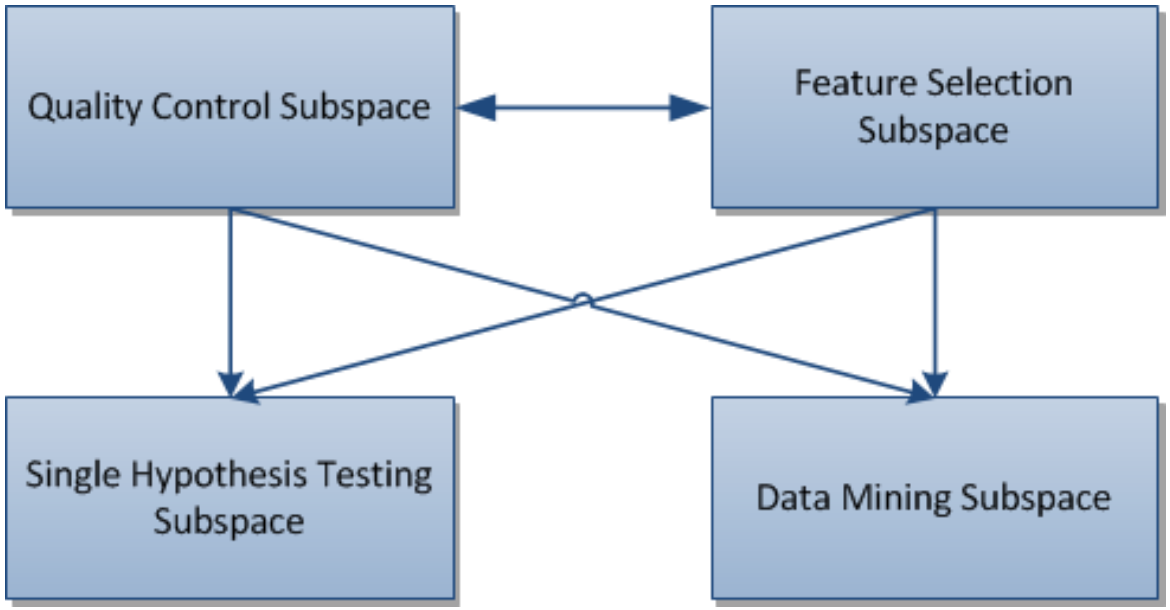
Data Mining Subspace

Provides the tools for performing data mining analyses on a dataset and extract association rules.

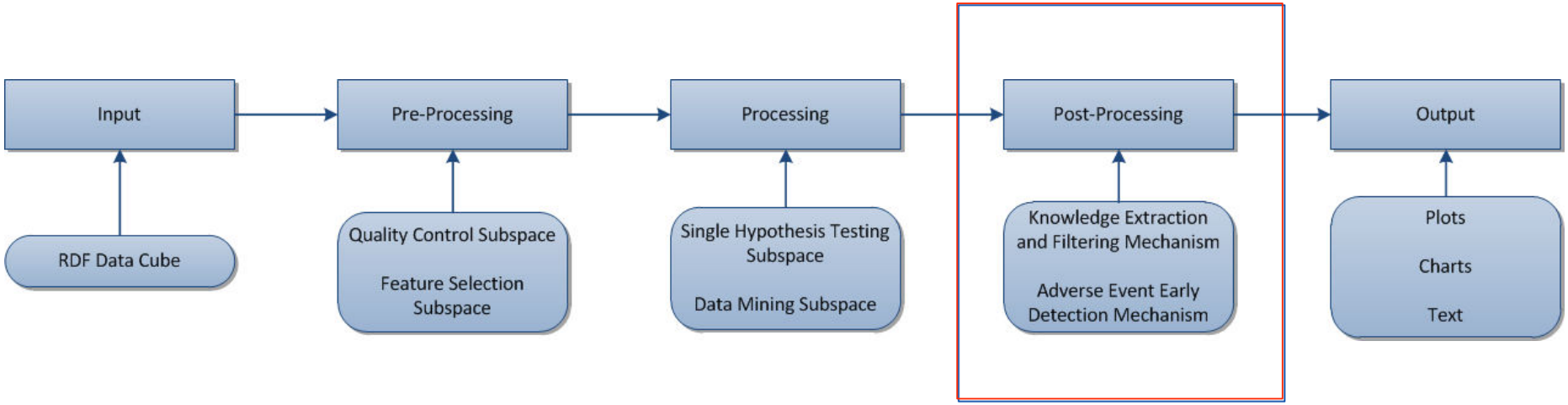
Tools:

- **Association Rules (apriori)**
- **Decision Trees with Percentage Split (C4.5)**
- **Decision Trees with Cross Validation (C4.5)**
- **Random Forest with Percentage Split**
- **Random Forest with Cross Validation**

Data Analysis Space Interactions



Data Analysis Steps



Knowledge Extraction and Filtering Mechanism

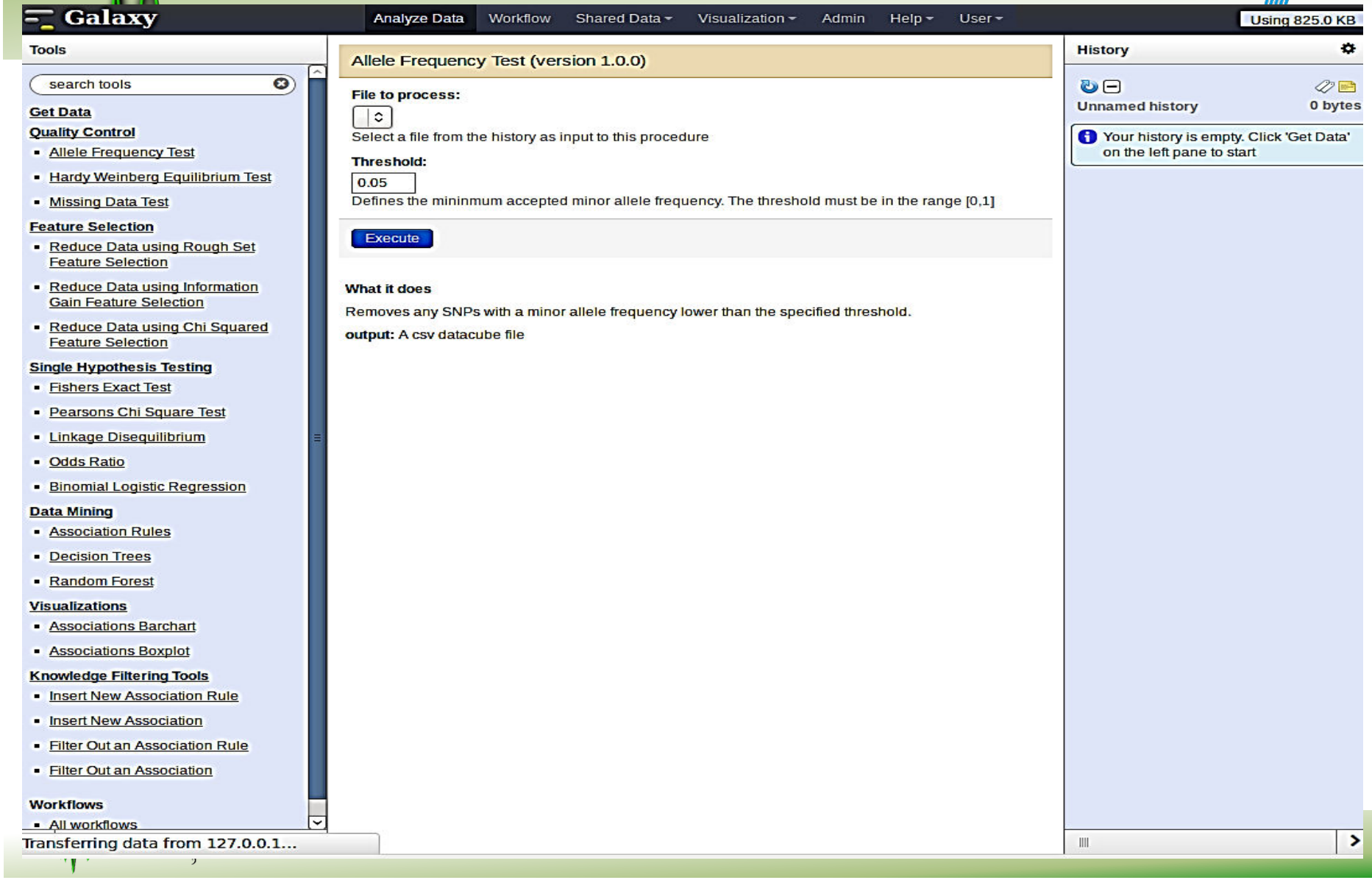
- **Knowledge Extraction Mechanism**
  - ✓ This mechanism is responsible for storing statistically significant associations and important association rules in the Linked2Safety knowledge database
  - ✓ Has two steps:
    - ❑ Logging system
    - ❑ Storing important knowledge

- **Filtering mechanism**
  - ✓ This mechanism allows users to insert or delete associations and association rules

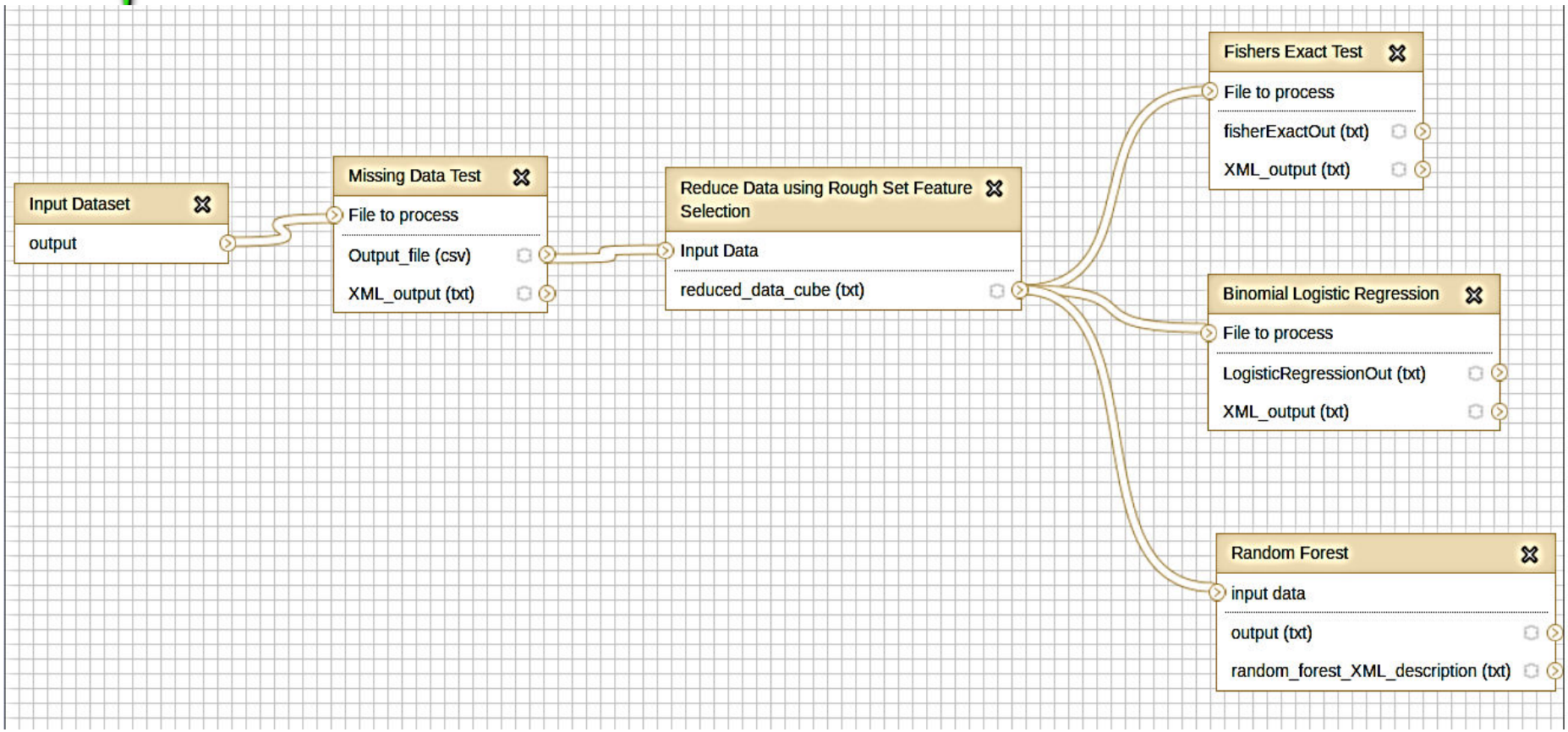
Adverse Event Early Detection Mechanism

- Uses the knowledge in the L2S knowledge base
- Runs in the background to identify new associations and association rules
- Reruns analyses when updated datasets are available
- Creates alerts for patients profiles associated with adverse events

Linked2Safety's Data Analysis Platform



Linked2Safety's Data Analysis Platform Workflow Screenshot



Patterns Discovery
Variables to Focus on

Approach for prioritizing variables to include in pattern discovery:

1. **Compare the data from the providers and select common variables**
2. **Expert suggestions on variable combinations to consider that will most likely produce interesting results**

Patterns Discovery
Common Variable Selection

Overlapping non genetic data of at least 2 data providers:

Variables	
Age	Weight gain
Gender	Headaches
BMI	Gastrointestinal symptoms
Smoking Ever	Ophthalmological problems
Dyslipidemia	Type of ophthalmological condition
Diabetes	High blood pressure
Diabetes type I	Heart conditions exist
Diabetes type II	Type of heart condition
Anemia	Hypertension
Depressive personality disorder	Myocardial infarction
Major depressive disorder	Stroke
Schizotypal personality disorder	Coronary heart disease

Examples of Hypotheses that will be Tested

Over 300 hypotheses will be tested

Trait of Interest	Variable 1	Variable 2	Variable 3
Bipolar disorder	Metabolic syndrome		
Anaemia	Age	Smoking history	
Asthma	Age	BMI	Smoking history
Diabetes	Schizotypal personality disorder		
Diabetes type II	rs10010131		

Thank you

Website

http://www.linked2safety-project.eu

Special Interest Group

http://www.linked2safety-project.eu/sig

Aristos Aristodimou

University of Cyprus

email: aristodimou.aristos@ucy.ac.cy