CRISP

centro di ricerca interuniversitario
per i servizi di pubblica utilità

R role in Business Intelligence Software Architecture

CRISP - Inter-university Research Centre on public Services

Ettore Colombo
Gaithersburg, Maryland
July, 22 2010

University of Milano - Bicocca
Viale dell'Innovazione 10
Building U9, 2nd floor
20126 Milan, Italy

Tel: (+39) 02 6448 2180
Fax: (+39) 02 70056 9114
e-mail: crisp@crisp-org.it
web: www.crisp-org.it

Introduction to C.R.I.S.P.

The on-going collaboration and mutual exchange between several centres of study was rendered official in 1997 by the creation of a centre of study proposing high-profile research on public services.

Università Cattolica del Sacro Cuore

Università degli studi di Bologna

Università degli studi di Udine

Università degli Studi di Brescia

Università Commerciale L.Bocconi di Milano

Università degli studi di Milano

Università degli studi di Milano Bicocca

Crisp's main areas of concern:

1. **public service development** and demand analysis;
2. analysis of economic system dynamics;
3. unbiased methodologies for quality estimation of services;
4. **technology innovation**

CRISP "Public Services":

- Training and the **Labour Market**
- Public Health
- Environment and the Quality of Life
- Education and Learning
- Public Utilites

LABOR Project

Regione Lombardia

PROVINCIA MONZA BRIANZA

Project Goal:

provide the provinces of Lombardy with a **Business Intelligence (BI) System** to analyse their labour markets.

Outcome:

a **Statistical Information System** (SIS) integrated in the BI process

statistical models integrated in in BI system

a **community of users** crossing the province boundaries

CRISP
centro di ricerca interuniversitario per i servizi di pubblica utilità

SIS Technological Platform Design

BI analysis tools

OLAP
Reporting
Dashboard

Statistical models

Complex, innovative and domain-dependent models coming from Research

Community Feedback

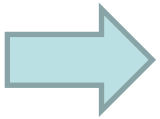
Suggestions and hints coming from the experience of the user community

Adaptability

Extendibility

Flexibility

Technological Platform features



Integration and interoperability

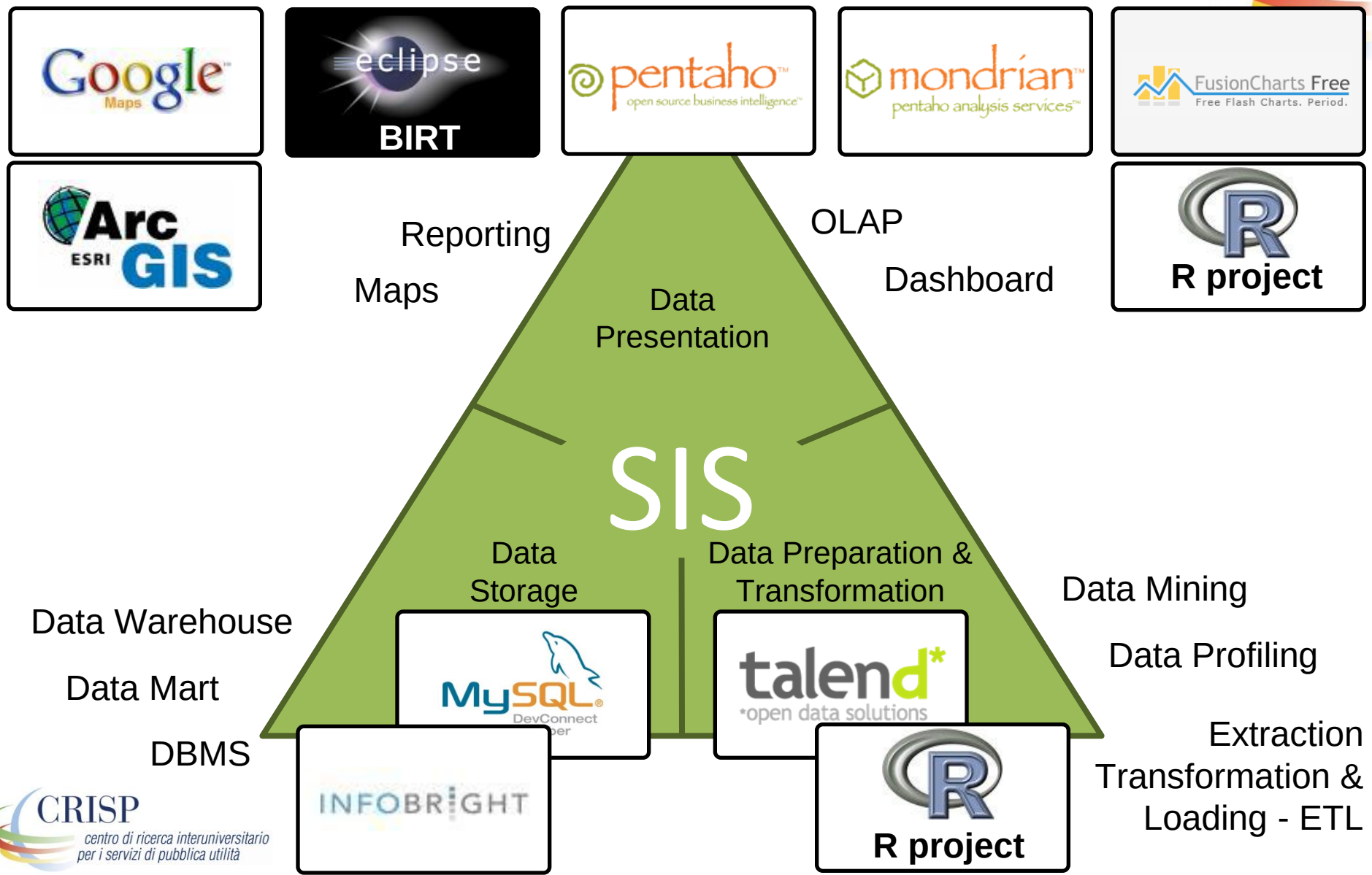
Innovative Communities

No licences to pay

Open Source Projects

CRISP
centro di ricerca interuniversitario per i servizi di pubblica utilità

SIS Software Layers



CRISP
centro di ricerca interuniversitario per i servizi di pubblica utilità

R and the Data Transformation & Preparation Layer: the actors

Need to run complex data analysis methods not supported by common ETL tools - e.g. Clustering method to classify workers' careers

Need to run these methods directly in the ETL processes

talend* *open data solutions

An Open Source Platform for ETL and Data Profiling.
Talend OS is a visual suite (based on Java & Perl) to develop ETL processes

MySQL DevConnect Member

MySQL … the well-known DBMS used at CRISP to develop Data Warehouses and Data Marts

R project

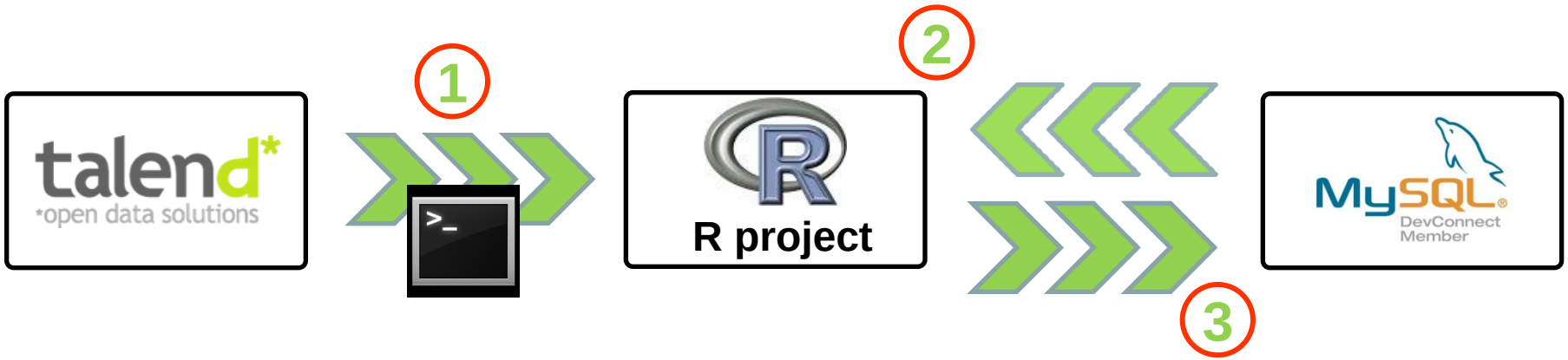
RMySQL

R scripts

R and its packages …
RMySQL is used to get data from MySQL
A set of R scripts with the algorithms developed at CRISP

R and the Data Transformation & Preparation Layer: the process

R is used to elaborate data with innovative models defined by CRISP researchers during ETL in a 3-step process



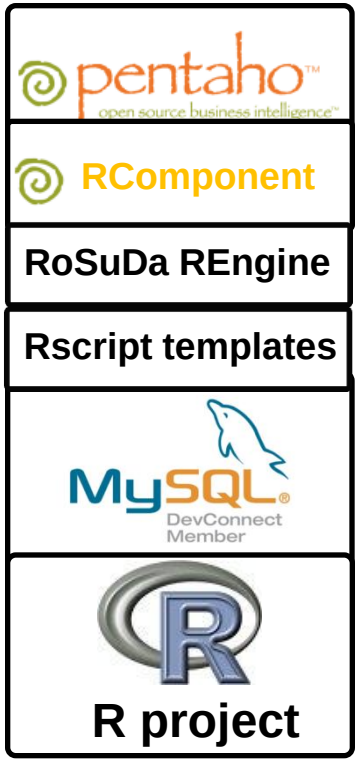
1. During the execution of an ETL process, TALEND launches R via command line
2. R runs the script on the data from the DBs
3. R stores the outcome data in dedicated DB tables

Light but effective (no need to give back data to TALEND)

R and the Data Presentation Layer: the actors

Need to graphically represent the results of the run of complex data analysis methods - e.g. Markov's Chains on workers' contract type

Need to show these representation in SIS dashboards



The Open Source BI platform that is the backbone of the Presentation Layer

An ah-hoc extension of Pentaho to manage the interactions with R (via Rengine) and preparation of the elements to be shown in Pentaho dashboards

A set of script templates containing placeholders for DB connection and model parameters

MySQL … the well-known DBMS used at CRISP to develop Data Warehouses and Data Marts

RMySQL

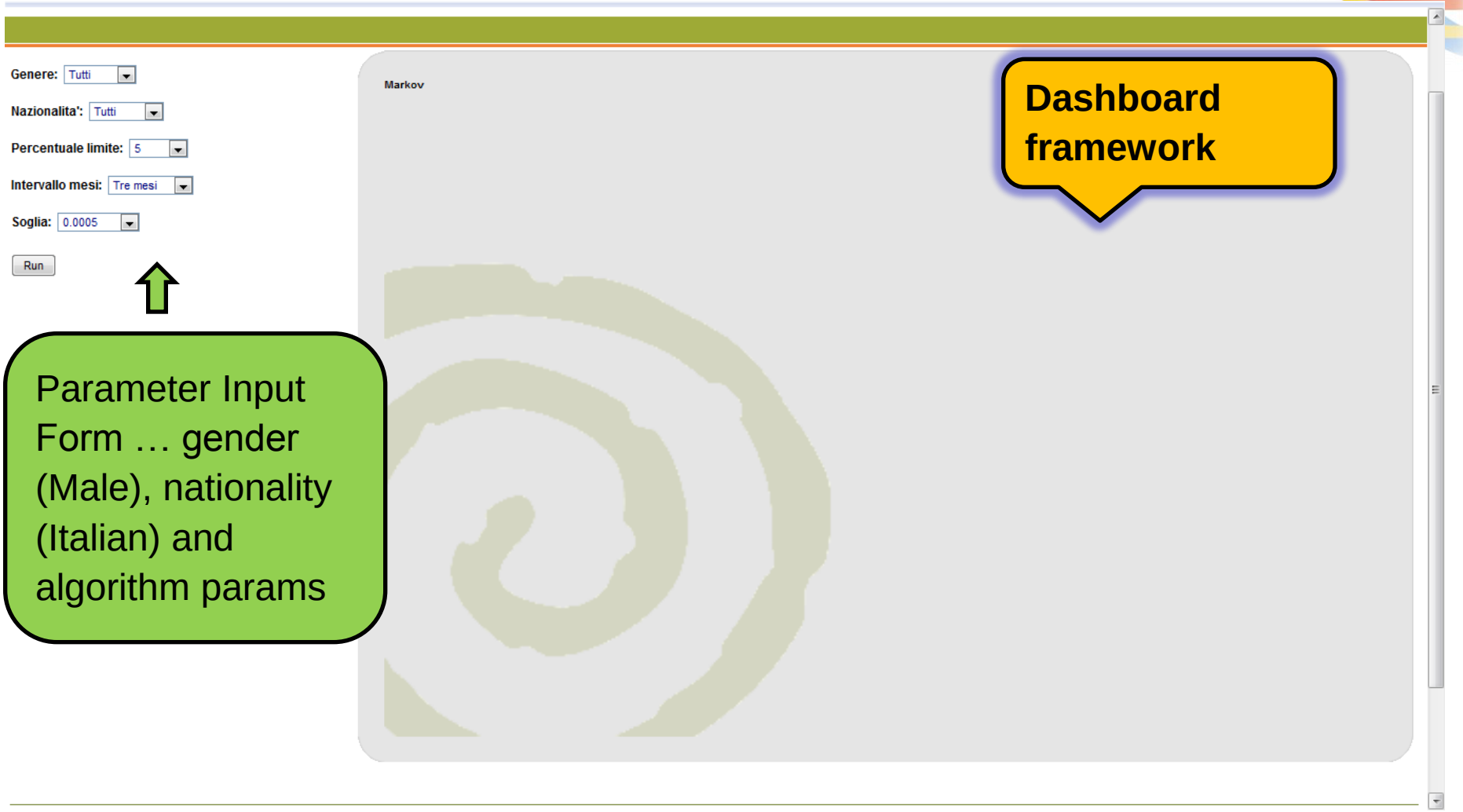
RgraphViz

RoSuDa RServe

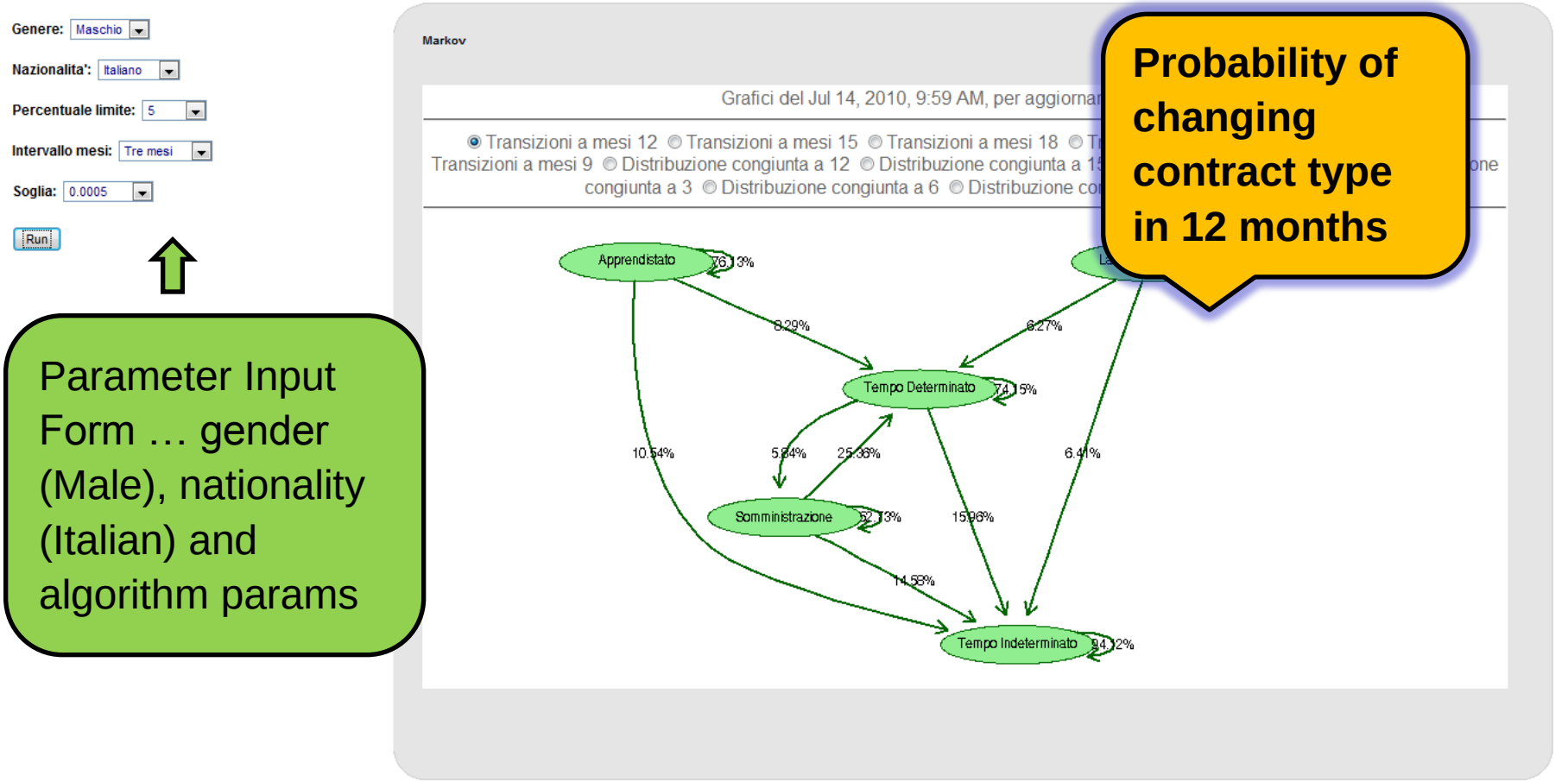
→ R and its packages …
→ RMySQL is used to get data from MySQL
→ RgraphViz (Bioconductor) is used to generate graphs
→ Rserve is used for TCP/IP communication over the internet

CRISP
centro di ricerca interuniversitario per i servizi di pubblica utilità

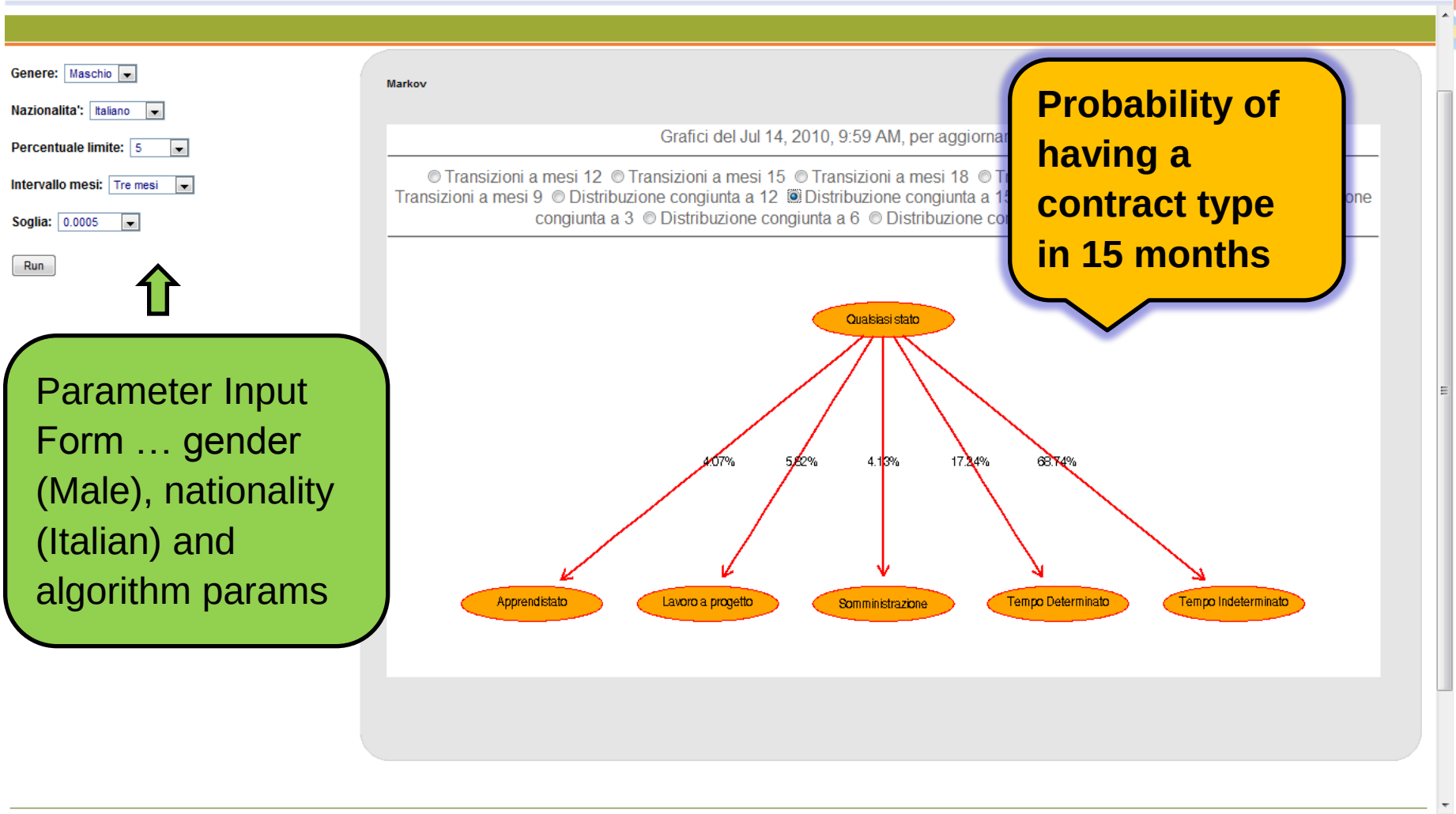
R and the Data Transformation & Preparation Layer: the front-end process



R and the Data Transformation & Preparation Layer: the front-end process



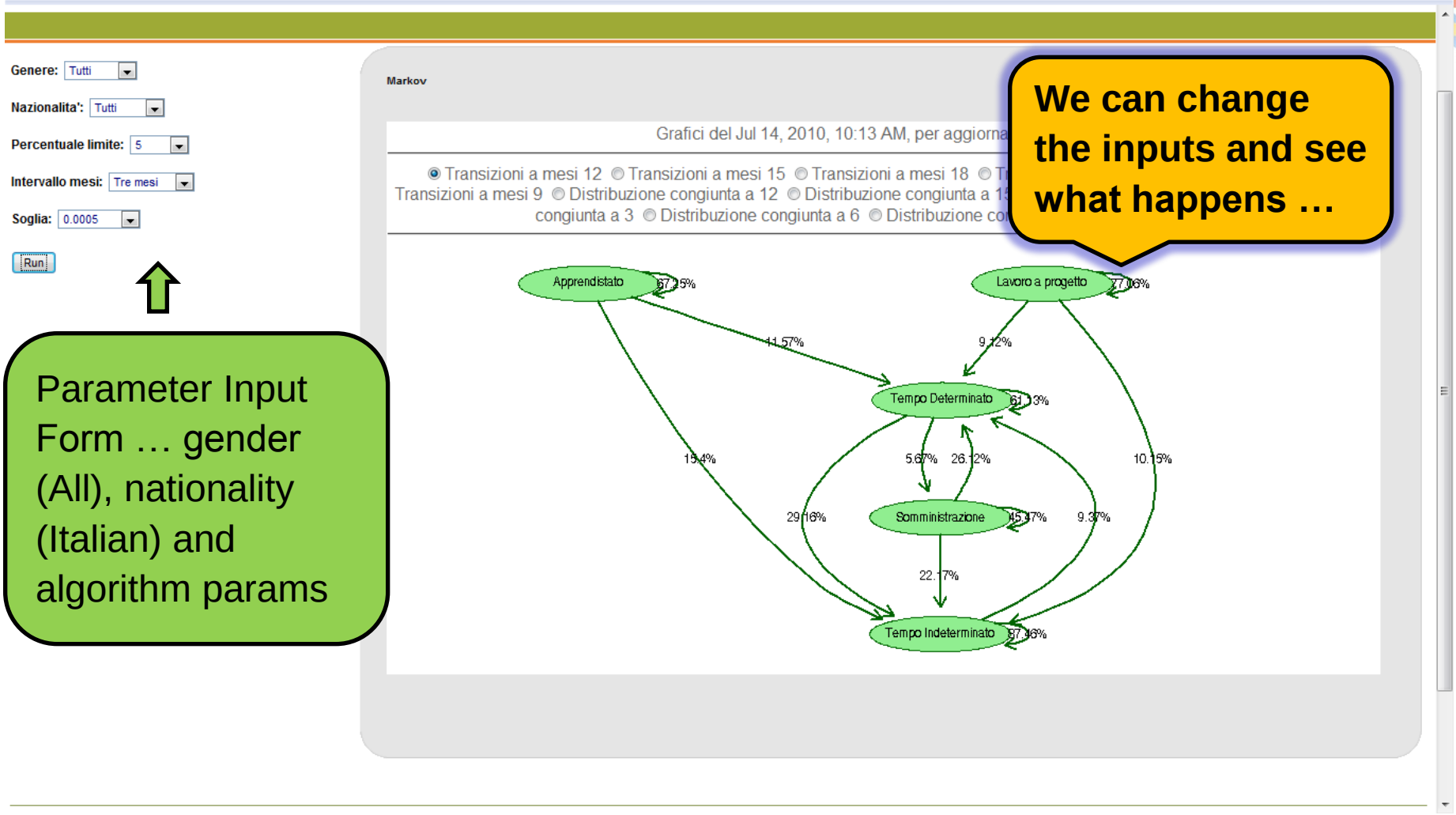
R and the Data Transformation & Preparation Layer: the front-end process



Parameter Input Form … gender (Male), nationality (Italian) and algorithm params

Probability of having a contract type in 15 months

R and the Data Transformation & Preparation Layer: the front-end process

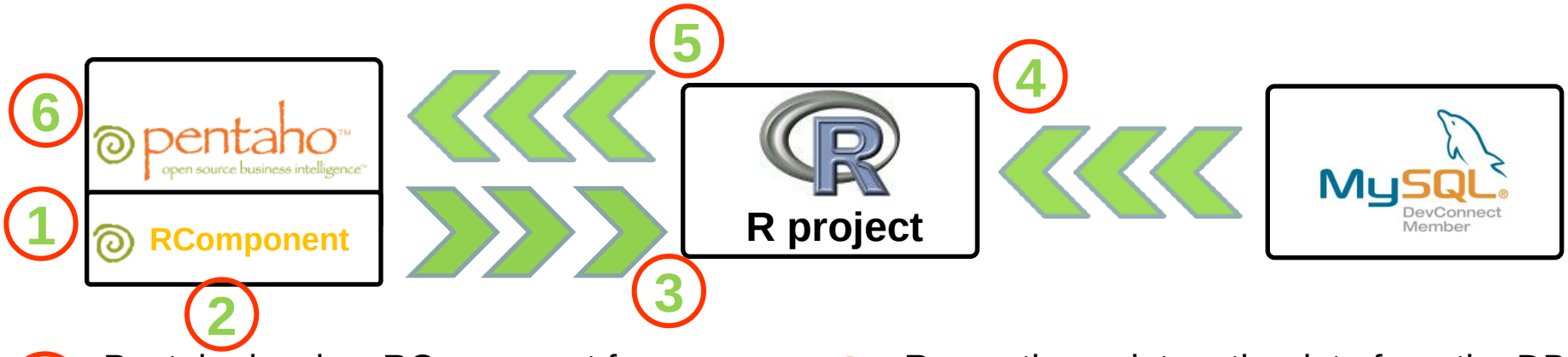


Parameter Input Form … gender (All), nationality (Italian) and algorithm params

We can change the inputs and see what happens …

R and the Data Presentation Layer: the back-end process

R is used to elaborate data and generate graphs to show the outcome of the execution of algorithms defined by CRISP researchers in a 6-step process



1. Pentaho invokes RComponent for a specific script template and data source
2. RComponent parses the script template and generates a new in-memory script and connects to Rserve
3. RComponent remotely launches the execution of the script to Rserve
4. R runs the script on the data from the DBs and generates a set of JPGs via Rgraphviz
5. Rserve takes these pictures and returns them to RComponent
6. RComponent prepare an HTML fragment to be shown in the Pentaho framework

Physical and logical Separation of concerns

Integration "limited" to visualization issues

Conclusions

R and the Data Preparation & Transformation Layer		R and the Data Presentation Layer
R plays an active role in ETL processes to run complex statistical analysis - Clustering on Workers' careers	**NOW**	R plays an active role the SIS generating visualization strictly related to statistical analysis - Markov's Chains
Extend the use to other analysis and models - Clustering on Workers' Skills	**NEXT FUTURE**	Extend the use to other models - Time Series and Geospatial Analysis
… the Light Integration Change the paradigm of communication between Talend and R in order to enable R to give back data useful for ETL processes	**Beyond …**	… the Visualization Use the developed communication infrastructure between Pentaho and R to run different kind of script (e.g. What-If scenario analysis) giving back data, not only images

FURTHER INFORMATION

Web: www.crisp-org.it
E-mail: ettore.colombo@crisp-org.it

Tel: (+39) 02 6448 2172
Fax: (+39) 02 70056 9114