Growing relations between citizens and public administrations generate a lot of data
- Data are often longitudinal
- Repeated observations (events) of the same subject at multiple time points generate a sequence of events for the subject;
- Longitudinal databases allow one to observe and measure how the data change along the time.
- Data quality of enterprise and public administration databases is very low [Blaine and Scannapieco, 2005;Redman, 1998].

Data Quality is described by many dimensions e.g., accuracy, consistency;
- Consistency describes the violation of semantic rules defined over a set of data items.

Data is used for decision making purposes, e.g. in Business Intelligence Systems.
- Derive information from low quality database may lead to dangerous or wrong decisions;
- Data Quality Analysis and Improvement techniques are required before using data.

State of the art
Mainly, three categories of techniques deal with data quality problems:
- record linkage: Uses alternative (and more trusted) data sources to improve the quality;
- error localisation and correction: Typically exploits functional dependencies [Fan, 2008] to detect errors and looks for a repair, i.e. another database which is consistent and minimally differs from the original one [Chumskis, 2009];
- consistent query answering: Looks for consistent answers from inconsistent data, i.e. the focus is on automatic query modifications and not on fixing the source data. An answer is consistent if appears in every possible repair of the original database [Arenas, 2003].

Aim of the research
- Exploit formal methods developing techniques for:
  - Data Quality Analysis: The Robust Data Quality Analysis (RDQA) evaluates and helps to improve the quality level achieved after a cleansing intervention.
  - Sensitivity Analysis: Quantitatively estimates the impact that cleansing interventions may have on aggregate indicators computed on the data.

What Model Checking can do?
- Given a system model described by:
  - State variables whose evaluation determines a state;
  - Transition relations specifying actions leading the system from a state to another one.
- Given a property to be verified in any state of the system.

An explicit Model Checker:
- explores all the feasible system configurations (reachable states);
- verifies the properties in each state;
- returns a sequence of actions describing how the system reaches a state which violates the property (if any).

How to Perform Data Analysis via Model Checking?
Main Idea: Express the problem of Data Consistency verification as a Model Checking Problem.

![Diagram](image-url)

The Robust Data Quality Analysis (RDQA)

Given a function \( clr \) performing consistency verification on a dirty dataset \( S \) and generating the cleansed version \( C \), some questions arise:
- What is the degree of consistency achieved through \( clr \)?
- Can we improve the consistency of the cleansed dataset?
- Can we be sure that function \( clr \) does not introduce any error in the cleansed dataset?

We use model checking to implement a function \( ccheck \) able to verify the dataset consistency before and after the \( clr \) intervention.

RDQA works iteratively as follows:
- Use \( clr \) to cleanse the source database \( S \) generating \( C \);
- Use \( ccheck \) to verify the consistency on the source (cleansed) database \( S \) generating \( C \);
- \( D \) contains all datasets modified by \( clr \) function;
- \( D \) contains all datasets untouched by \( clr \) function;
- Generate the Double Check Matrix (DCM) and use it to analyse/repair/modify the \( clr \) behaviour;
- Repeat 1-4 until a satisfying cleansing result is achieved.

A Real-World Application: “The Worker Career Archive”
- According to the Italian law, whenever an employer hires or dismisses an employee, a communication (or event) is sent to the job archive. The archive content is used to study the labour market, obtaining information about worker career paths and supporting the decision making processes.

A Worker’s Event contains:
- The worker ID;
- The event ID and the date;
- The event type: start, cessation, extension or conversion of a worker contract;
- The event modality, whether the event is related to a full-time or a part-time contract;
- The contract type wrt the Italian law.

The CRISP research centre aims to evaluate and improve archives data quality.

The RDQA has been applied to improve the cleansing activities on a worker careers database for an Italian Area, with 1,089,895 mandatory communications regarding 213,566 workers collected in 10 years.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

Conclusion and expected achievements
- RDQA: a Model Checking based approach to evaluate and improve database data quality;
- Exploit Model Checking supporting the generation of data cleansing routines.