Data Cleansing for Genome Data

Heiko Müller
Lehrstuhl für Datenbanken und Informationssysteme
Institut für Informatik
Humboldt-Universität zu Berlin
Introduction

- **Data Cleansing**
  - Data has become one of the most valuable corporate assets
  - High quality is essential
  - (Semi-)automatic identification and correction of errors in data collections

- **Genome Data is Error-Prone**
  - Sequence errors,
  - Wrong functional annotation, ...

- **Existing Methods for Data Cleansing**
  - Remove duplicates
  - Mainly on Address Files
Data Cleansing Process

Data

Error Classification

Rules for the Identification of errors

Rules for the Correction of errors

Error Identification

Error Correction

Documenting & Storage

Cleansing Function
Error Classification

- Syntax error
  - Abbreviations
  - Typing error
  - Format error
- Semantic error
  - Wrong annotation
- Errors in data schema
  - Wrong datatype
  - Missing constraints
- Errors in data
  - see above
Isolated vs. Integrated

Isolated

CLEANSING FUNCTION

Lookup

Integrated

CLEANSING FUNCTION

Lookup

virutell

materialised
Cleansing Function

Metadata