



Utrecht University

Information and Technology Services (ITS)

Data Engineering: Clean and Integrate Your Data!

Frans de Liagre Böhl and Jonathan de Bruin

April 10, 2017

Table of Contents

- **What is Data Cleaning and Integration?**
- **Data cleaning and integration tools**
- **Workshop dataset**
- **Data cleaning with OpenRefine**
- **Data Integration with R**

What is Data Cleaning and Integration?

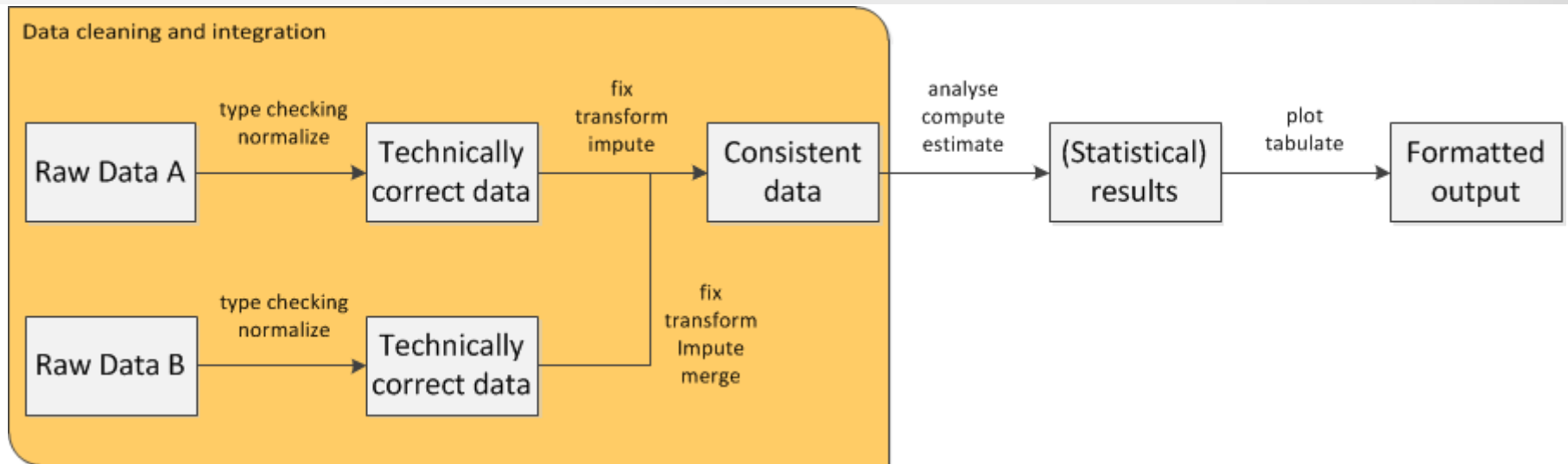


Image based on: "An introduction to data cleaning with R", CBS 2013

Data cleaning and integration tools



- A lot of free and commercial tools available
- Each tool has pros and cons
- Ensure yourself of **reproducibility**

Workshop datasets: Hospital records and mortality data

Search this file...

Mortality dataset

	patient_id	date_of_death	place_of_death	date_of_birth	first_name
1	e3e70682-c209-4cac-629f-6fbed82c07cd	2012-11-22 18:21:09	4759 William Haven Apt. 194, West Corey	1972-06-06 16:13:46	Steve
2	cd613e30-d8f1-6adf-91b7-584a2265b1f5	2009-09-06	77763 Tony Village Suite 690, Adamsbury	1938-04-15 06:47:45	Ryan
3	d95bafc8-f2a4-d27b-dcf4-bb99f4bea973	2001-07-23 03:06:50	9390 Yvonne Route Suite 858, Shawton WA	1927-09-12 05:25:03	A.
4	21636369-8b52-9b4a-97b7-50923ceb3ffd	2010-01-30		1952-05-21 07:32:09	Joshua

Search this file...

Hospital dataset

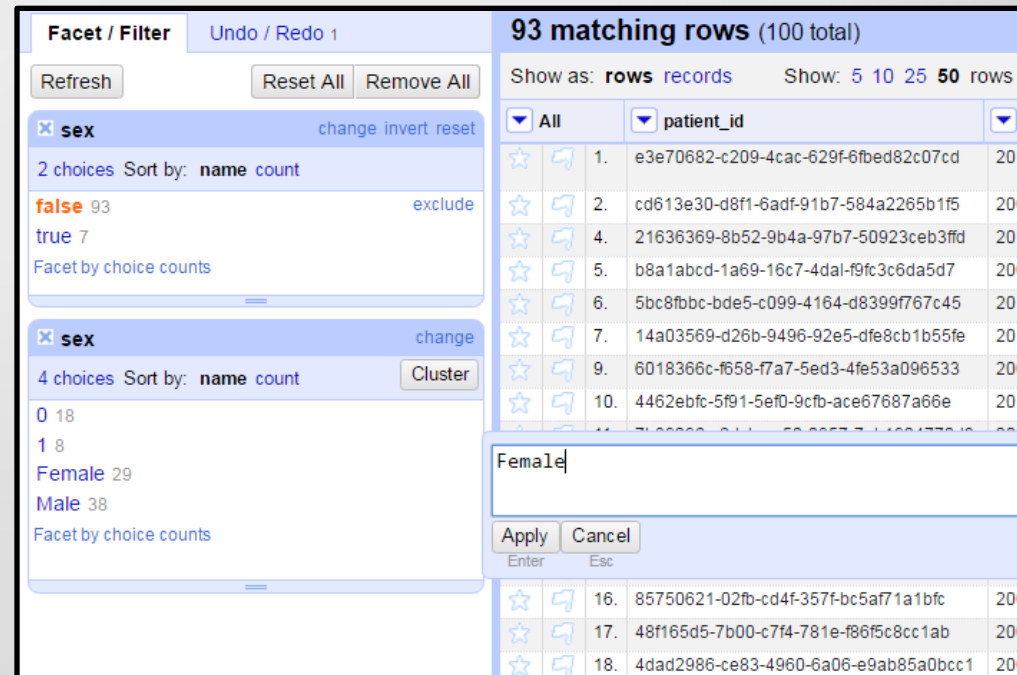
	patient_id	ICD10	datetime	
1	988a0c48-5979-91d1-7000-368d2534c02d	V97.33XD	2000-11-21 09:01:32	Brian
2	5d357ffe-4423-f60d-db0e-da407f5e8e61	W51.XXXA	2004-07-21 07:54:46	Heather
3	d95bafc8-f2a4-d27b-dcf4-bb99f4bea973	V00.01XD	2001-06-25 19:13:09	Angela
4	8dbb5b2a-6e20-af8e-1001-a6625a1298a1	Y93.D	2006-12-05 12:17:36	James
5	5532e8ba-3083-d49e-f945-e4b665c1d4b4	Z99.89	2004-07-14 08:23:10	V.
6	0b07502e-d4c6-eb9c-9331-06745b3ce9b3	Y92.146	2001-10-31 05:25:21	J.
7				Jessica
				Justin
				Matthew

Workshop dataset: Observation strategies

- Read metadata
- Check types
- Check record identifiers (unique?) and subject/entity identifiers
- Count the number of unique observations
- Sort the data on interesting variables
- Find out how the values are formatted
- Check distributions of outcome/value frequencies
- Identify (all) overlapping variables between datasources
- ...

Data cleaning: Clean the gender variable

- Outcomes in dataset:
"Male", "Female", "Fem",
M", "F", 1, 0, NA
- Desired output: "Male",
"Female", NA



The screenshot shows a data cleaning interface with two facets for the 'sex' variable. The top facet shows 2 choices: 'false' (93) and 'true' (7). The bottom facet shows 4 choices: '0' (18), '1' (8), 'Female' (29), and 'Male' (38). A search bar at the bottom right contains the text 'Female'.

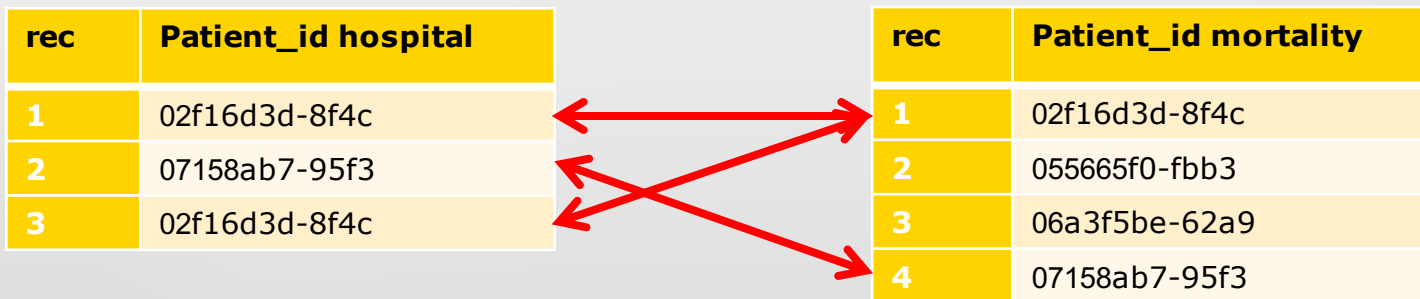
Question for R-users: `replace(dfsex, dfsex==1, "Male")` or `replace(dfsex, dfsex=="1", "Male")`

Data cleaning: Information extraction

- Extract the state from the place of death
- Semi structured information:
- **Example: "9390 Yvonne Route Suite 858, Shawton WA"**
- One option: Subset records and extract information with regular expressions
- Look for patterns in term of alphanumeric characters, numeric, punctuation and upper/lowercase letters.
- **postcode: "9390", street: "Yvonne Route Suite 858", placename: "Shawton", state: "WA"**

Data integration: Unify the cleaned datasets

- Join the datasets on the unique personal identifiers
- This can be done in OpenRefine, but is easier with R or Python



rec hospital	Rec mortality	Patient_id hospital	Patient_id mortality
1	1	02f16d3d-8f4c	02f16d3d-8f4c
2	4	07158ab7-95f3	07158ab7-95f3
3	1	02f16d3d-8f4c	02f16d3d-8f4c

Data integration: Fuzzy matching

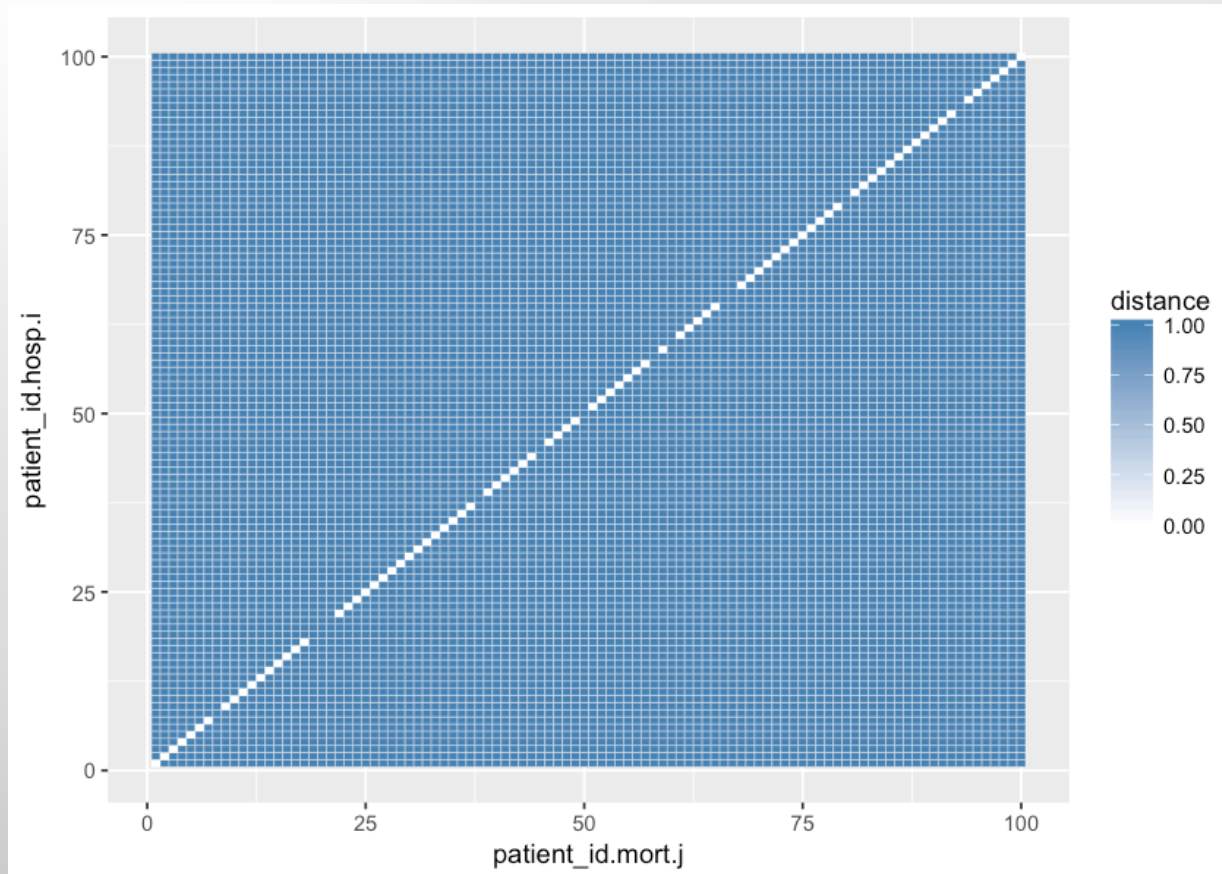
- (Hard) inner join excludes possible matches
- Compute the Levenshtein distance between all identifiers (i.e. the number of insertions, deletions and substitutions)
- Table with differences

Patient_id hosp.	32833106-536e-95df-40 2b -d002 cd 92d33d
Patient_id mort.	32833106-536e-95df-40 b2 -d002 cc 92d33d

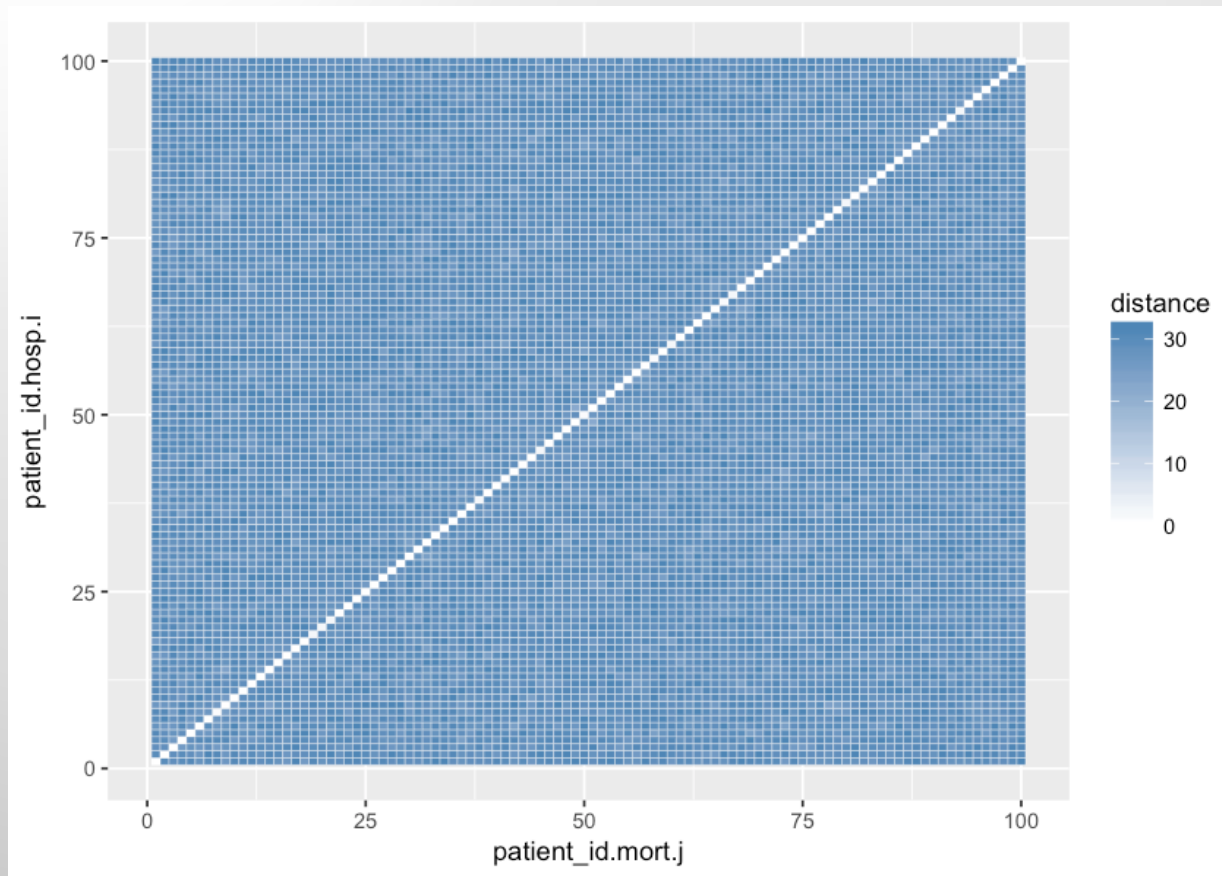
- The distance between the identifiers:

$$d("328331 \dots 92d33d", "328331 \dots 92d33d ") = 3$$

Data integration: Normal matching



Data integration: Fuzzy matching



Questions and discussion