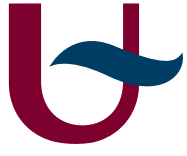
Cleaning data with forbidden itemsets

Joeri Rammelaere
with Floris Geerts & Bart Goethals

Universiteit Antwerpen



I'll talk about . . .

- What dirty data is
- What forbidden itemsets are and how to mine them
- How to repair dirty data using nearest neighbours
- Demo!

Dirty data

When is data dirty?

- Typically:
  - Define constraints on data
  - Data is dirty if constraints are violated
- What kind of constraints?
  - Many formalisms exist
  - For example functional dependencies

How do we find constraints?

- Human experts
- Master data
- Constraint discovery
- ...
- But what if we only have dirty data?

Dirty example

Age	MaritalStatus	Relationship	Sex	Country
39	Never-married	Not-in-family	Male	USA
38	Married-AF-spouse	Wife	Female	USA
17	Divorced	Not-in-family	Male	USA
37	Married-civ-spouse	Wife	Female	USA
28	Married-civ-spouse	Wife	Female	Cuba
29	Married-civ-spouse	Wife	Male	USA

Table: Some partial tuples from the UCI adult census dataset

Forbidden itemsets

Lift of an itemset

- Lift between two itemsets A and B:
  - Number of occurences of $A \cup B$ divided by expected nr. occurences if A and B were statistically independent
- Lift of an itemset A:
  - Maximum lift between any partitioning X and Y of A
- We are interested in itemsets with low lift

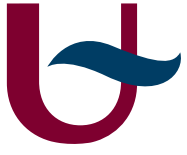
Converting tuples to transactions

- Tuple format:

Age	**MaritalStatus**	**Relationship**	**Sex**	**Country**
39	Never-married	Not-in-family	Male	USA

- Transaction format:

  (Age=39, MaritalStatus=Never-married, Relationship=Not-in-family, Sex=Male, Country=USA)



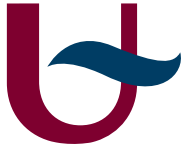
What are forbidden itemsets?

- Infrequent itemsets (support)
- Negative correlation between contained items (lift)

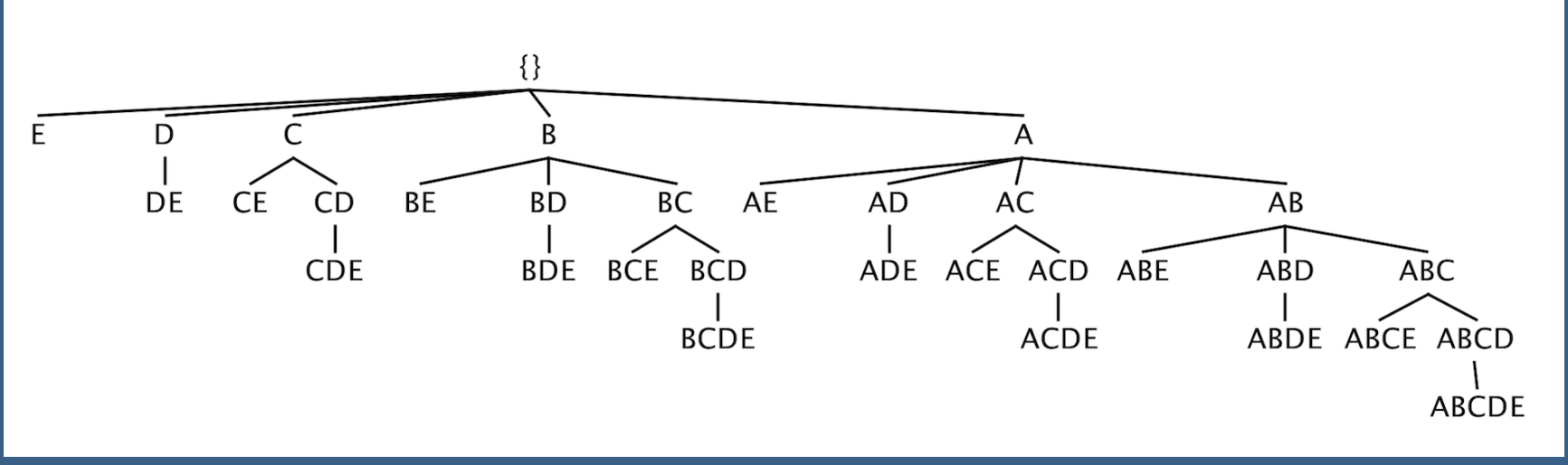
  $\Rightarrow$ Express forbidden value combinations
- For example:
  - (Relationship=Wife,Sex=Male)
  - (Relationship=Husband,Sex=Female)
  - (MaritalStatus=Divorced,Age=17)

Forbidden itemset mining

- Based on Eclat algorithm
- Maximum support threshold σ
- Maximum lift threshold τ
- Minimum support of items: $\theta = 1/\tau$



Forbidden itemset mining



Repairing dirty data

Nearest neighbour imputation

- Separate clean and dirty tuples
- Choose a similarity function
- For each dirty tuple, find nearest clean tuple

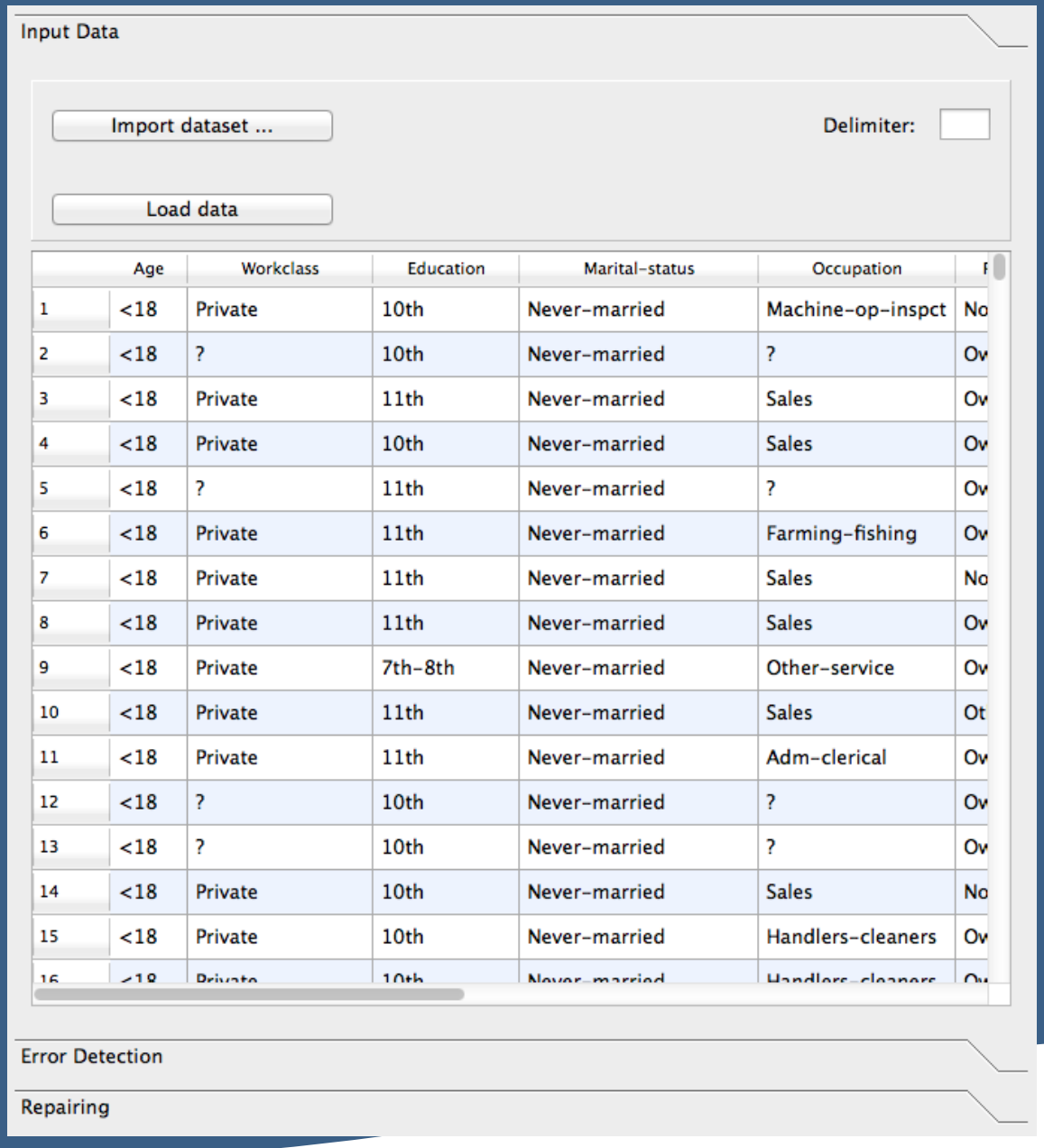
Nearest neighbour imputation

- So we have a neighbour . . . what now?
  - Copy entire tuple
  - Copy attributes involved in forbidden itemsets
  - Majority voting among donors

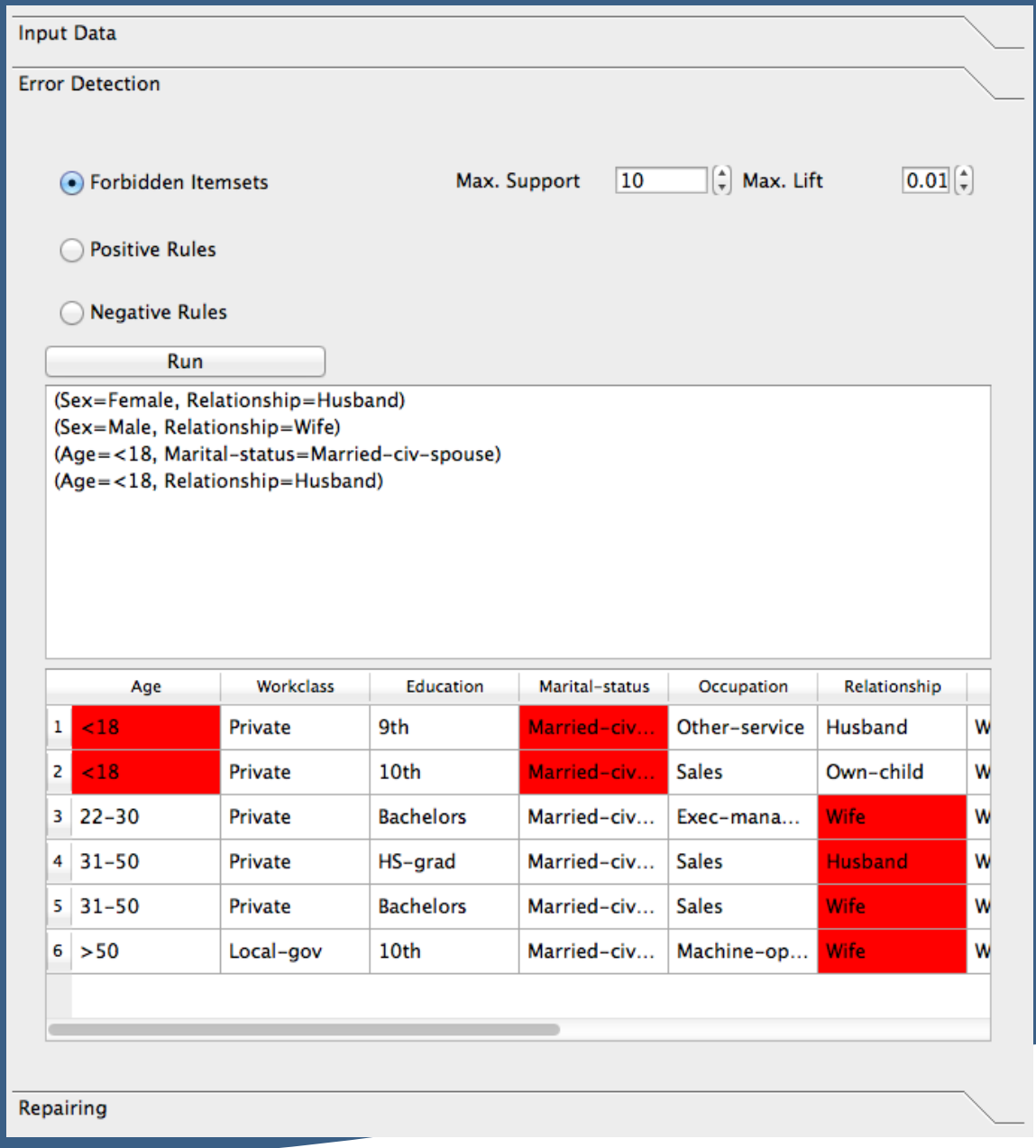
A problem!

- Repairing may cause itemsets to *become* Forbidden!
- Solution:
  - Find number of errors ϵ
  - Re-mine all itemsets that may become forbidden
  - . . . after ϵ edits

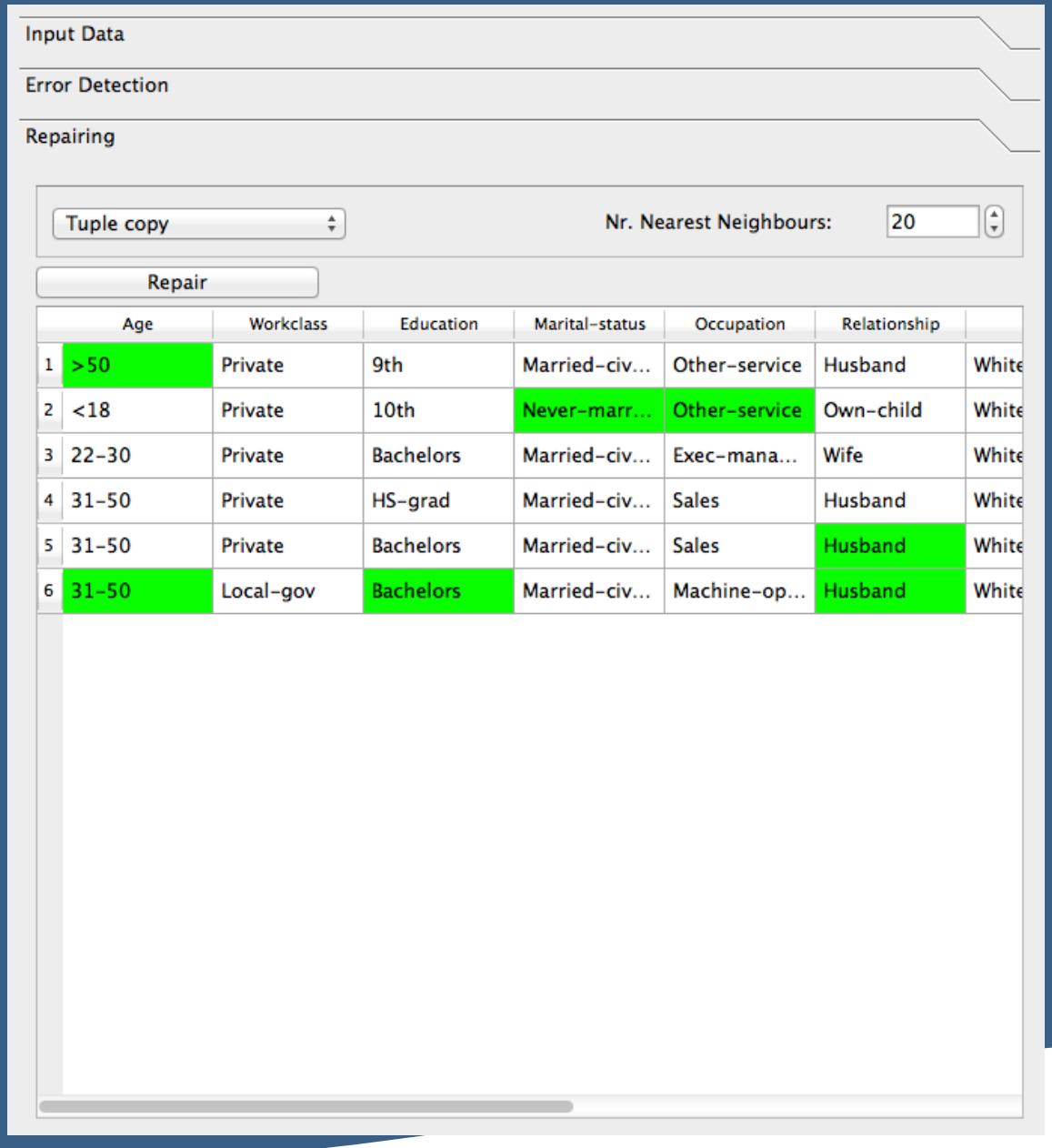
Demo



	Age	Workclass	Education	Marital-status	Occupation	F
1	<18	Private	10th	Never-married	Machine-op-inspct	No
2	<18	?	10th	Never-married	?	Ov
3	<18	Private	11th	Never-married	Sales	Ov
4	<18	Private	10th	Never-married	Sales	Ov
5	<18	?	11th	Never-married	?	Ov
6	<18	Private	11th	Never-married	Farming-fishing	Ov
7	<18	Private	11th	Never-married	Sales	No
8	<18	Private	11th	Never-married	Sales	Ov
9	<18	Private	7th-8th	Never-married	Other-service	Ov
10	<18	Private	11th	Never-married	Sales	Ot
11	<18	Private	11th	Never-married	Adm-clerical	Ov
12	<18	?	10th	Never-married	?	Ov
13	<18	?	10th	Never-married	?	Ov
14	<18	Private	10th	Never-married	Sales	No
15	<18	Private	10th	Never-married	Handlers-cleaners	Ov
16	<18	Private	10th	Never-married	Handlers-cleaners	Ov



	Age	Workclass	Education	Marital-status	Occupation	Relationship	
1	<18	Private	9th	Married-civ...	Other-service	Husband	W
2	<18	Private	10th	Married-civ...	Sales	Own-child	W
3	22-30	Private	Bachelors	Married-civ...	Exec-mana...	Wife	W
4	31-50	Private	HS-grad	Married-civ...	Sales	Husband	W
5	31-50	Private	Bachelors	Married-civ...	Sales	Wife	W
6	>50	Local-gov	10th	Married-civ...	Machine-op...	Wife	W



	Age	Workclass	Education	Marital-status	Occupation	Relationship	
1	>50	Private	9th	Married-civ...	Other-service	Husband	White
2	<18	Private	10th	Never-marr...	Other-service	Own-child	White
3	22-30	Private	Bachelors	Married-civ...	Exec-mana...	Wife	White
4	31-50	Private	HS-grad	Married-civ...	Sales	Husband	White
5	31-50	Private	Bachelors	Married-civ...	Sales	Husband	White
6	31-50	Local-gov	Bachelors	Married-civ...	Machine-op...	Husband	White

Source code

- Available soon at:
  http://adrem.ua.ac.be/joerirammelaere

The end . . .

Thank you for your attention!

Questions?