Explaining Repaired Data with Conditional Functional Dependencies

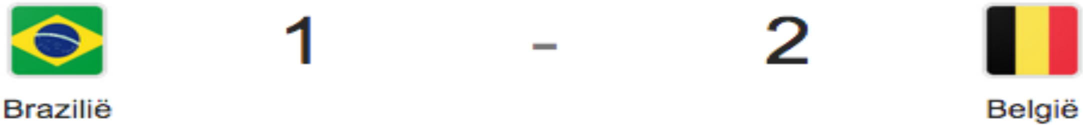
by Joeri Rammelaere and Floris Geerts

Universiteit Antwerpen

This talk was given at

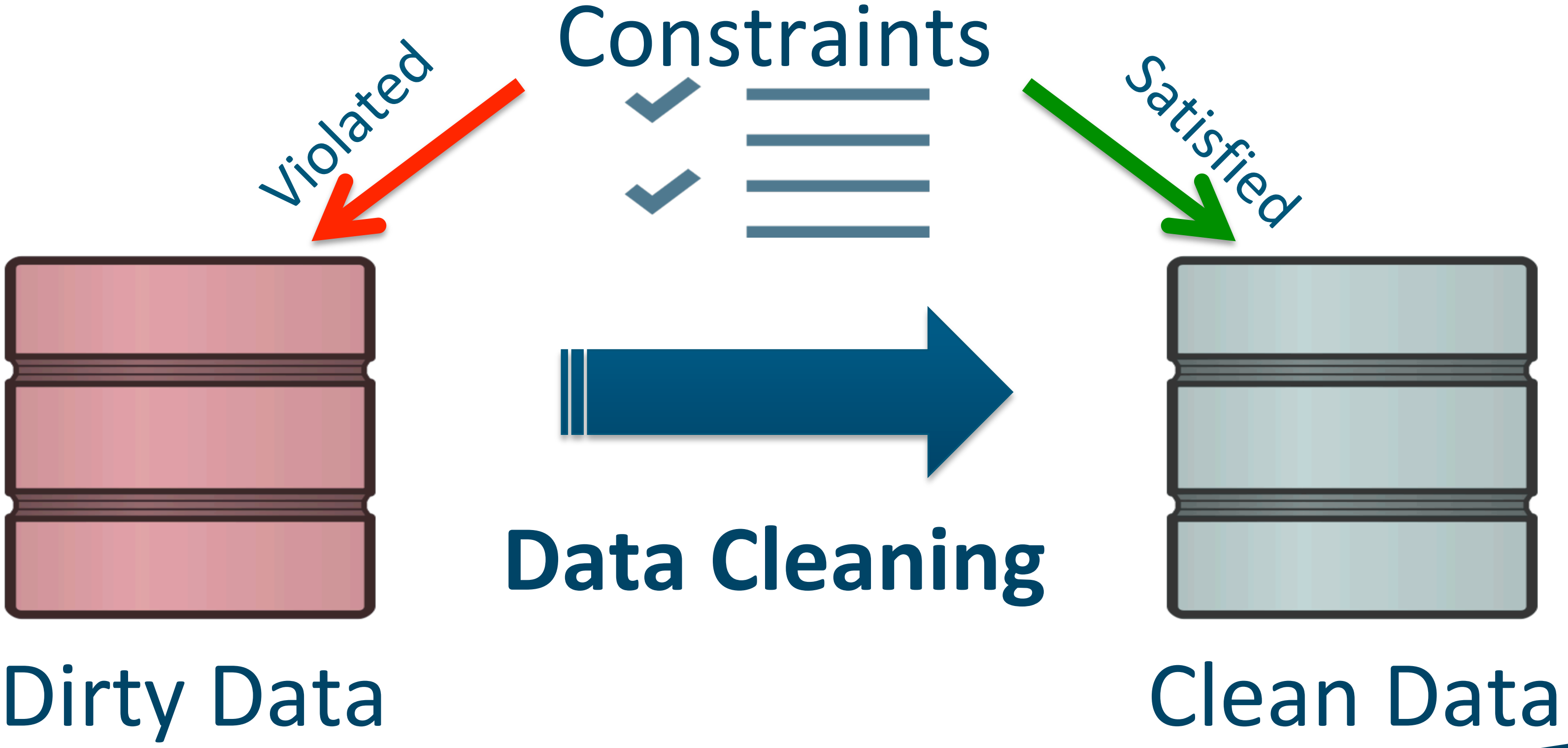
What is Dirty Data

Name	Country	Coach	Position	Age
Neymar	Brazil	Tite	Forward	26
Marcelo	Brazil	Tabarez	Defender	30
Alisson	Brazil	Tabarez	Goalkeeper	26
Neymar	Brazil	Tite	Forward	25



Brazilië 1 - 2 België

Constraint-based Data Cleaning



Conditional Functional Dependencies (CFDs)

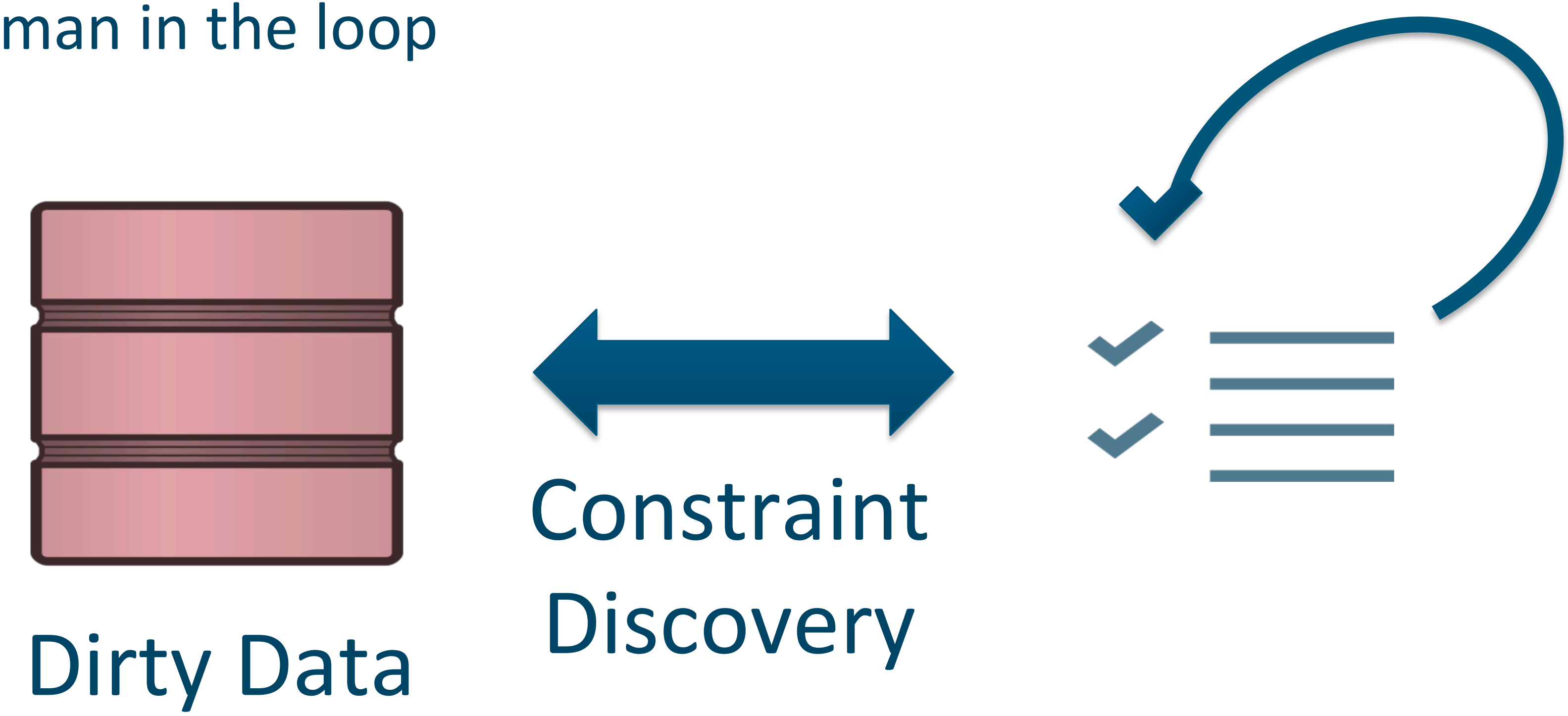
- CFDs are Functional Dependencies that **hold** on a **subset of the data**

- CFDs can capture **inconsistencies between** tuples, as well as **value errors within** a tuple

- **Examples**:
  - Name=‘_’ => Age=‘_’
  - Country = Brazil => Coach = Tite
  - Position = Attacker, Goals=‘_’, Assists=‘_’ => Rating=‘_’
  - Position = Goalkeeper, Saves=‘_’ => Rating=‘_’

Where do CFDs come from?

- Human in the loop



Dirty Data

Constraint Discovery

An example

Name	Country	Coach	Position	Age
Neymar	Brazil	Tite	Forward	26
Marcelo	Brazil	Tite	Defender	30
Alisson	Brazil	Tabarez	Goalkeeper	26
Neymar	Brazil	Tite	Forward	25

- We infer that the CFD **Country=Brazil => Coach=Tite becomes cleaner** (hence, explains the modification)

An example

Name	Country	Coach	Position	Age
Neymar	Brazil	Tite	Forward	26
Marcelo	Brazil	Tite	Defender	30
Alisson	Brazil	Tite	Goalkeeper	26
Neymar	Brazil	Tite	Forward	25

- The remaining error of this CFD can now be cleaned **automatically**

Why the human in the loop?

- CFDs typically cannot be provided by the user
    - User needs to **understand** the formalism ✗
    - No room for error: constraints must be **formulated exactly** ✗
    - User’s time is **expensive**! ✗

Why the human in the loop?

- Automatic CFD discovery finds **too many** CFDs! Which ones should we use for **repairing**?

Dataset	Support	Conf = 1.0	Conf = 0.9	Conf = 0.6
Adult	10%	7	68775	257855
Mushroom	10%	5842	2003868	3866951
Nursery	10%	7	927	8783

Table: Number of (approximate) CFDs found for various confidence thresholds

Summarizing our approach

- Human manually makes some modifications

- We find the CFD that best **explains** these modifications
  - This CFD should be **valid** and **useful for repairing**

- Once the correct CFD is found, **repairing can proceed** using any state of the art **automatic** method

- Our method requires **little interaction**, and is **robust** to small mistakes made by the user

Algorithm XPlode

- **XPlode** (explanations on demand) traverses the search space of frequent, approximate CFDs, and returns the **"best" explanation**
- **Best explanation**: scoring function based on the **number of modifications explained** by the CFD
- **On-demand:** we **only explore** parts of the search space that **can improve** upon the current best explanation, using an **upper bound** on the scoring function

Example (continued)

- Let’s clean the two Neymars
  - Errors violate the (C)FD Name=‘_’ => Age=‘_’

Name	Country	Coach	Position	Age
Neymar	Brazil	Tite	Forward	~~26~~ **25**
Marcelo	Brazil	Tite	Defender	30
Alisson	Brazil	Tite	Goalkeeper	25
Neymar	Brazil	Tite	Forward	25

- Perfect! No more violations

Example (continued)

- Let’s clean the two Neymars
  - Errors violate the (C)FD Name=‘_’ => Age=‘_’

Name	Country	Coach	Position	Age
Neymar	Brazil	Tite	Forward	26
Marcelo	Brazil	Tite	Defender	30
Alisson	Brazil	Tite	Goalkeeper	25
Neymar	Brazil	Tite	Forward	~~25~~ **26**

- Perfect! No more violations

Example (continued)

- Both modification are **individually** explained by the CFD
- But if we put them **together** …

Name	Country	Coach	Position	Age
Neymar	Brazil	Tite	Forward	~~26~~ **25**
Marcelo	Brazil	Tite	Defender	30
Alisson	Brazil	Tite	Goalkeeper	25
Neymar	Brazil	Tite	Forward	~~25~~ **26**

Approximating the scoring function

- Constant CFDs are fine; **problems** arise when considering **variable** CFDs

- We **convert** variable CFDs to a **union of constant** CFDs
  - E.g., Name=Neymar => Age=26
  - We can then simply **count** how many modifications are explained by any CFD
  - This becomes the approximate scoring function

Experiments

- We inserted violations for a randomly chosen CFD into various datasets

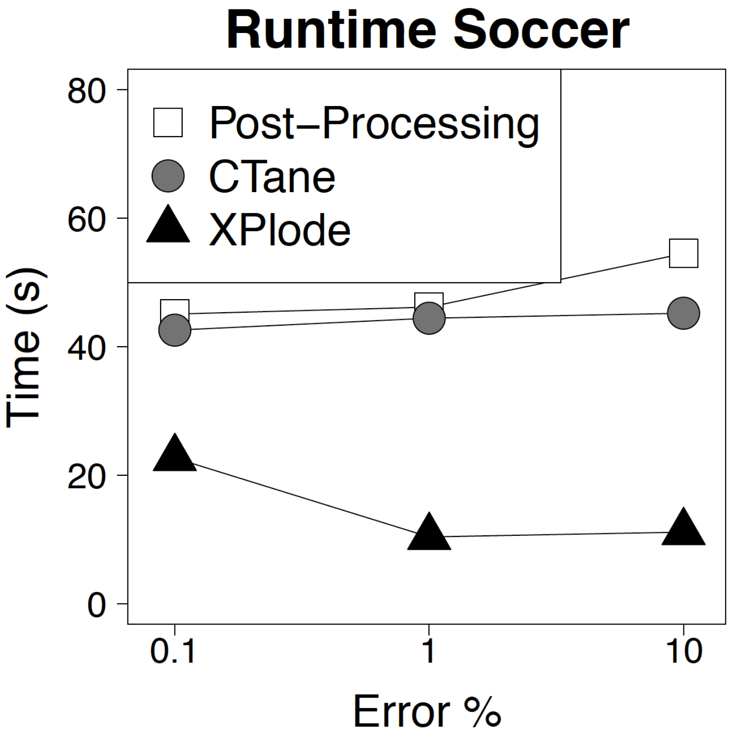
- The correct CFD is **recovered** with a **small number of modifications**

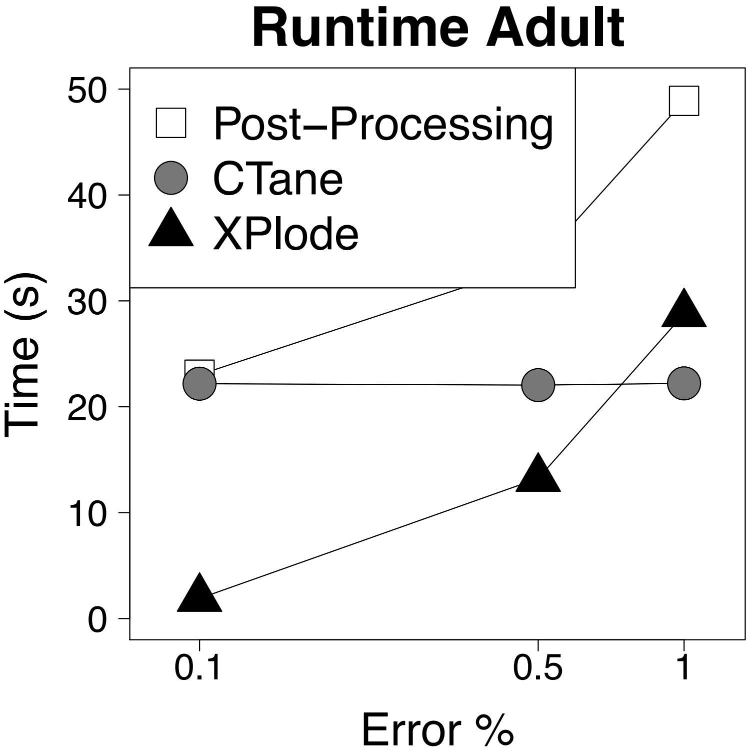
- XPlode is **faster** than regular CFD discovery

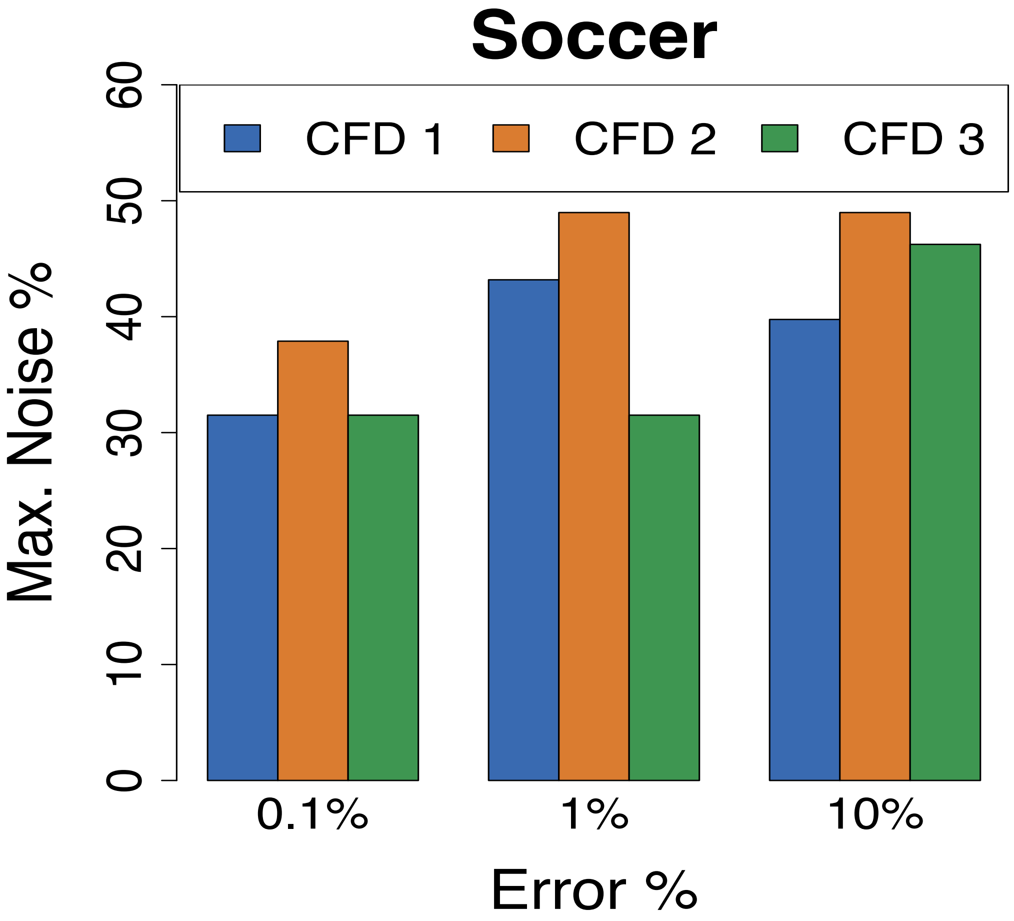
- The method is **robust to noise**

Nr. Modifications Needed

Dataset	Nr. Errors Inserted	Nr. Modifications Needed
Soccer	200	~13
Soccer	2000	~10
Soccer	20000	~25
Adult	97	~18
Adult	488	~13
Adult	976	~25


Runtime Soccer
□ Post–Processing
● CTane
▲ XPlode
Time (s)
80
60
40
20
0
0.1
1
10
Error %


Runtime Adult
□ Post–Processing
● CTane
▲ XPlode
Time (s)
50
40
30
20
10
0
0.1
0.5
1
Error %


Soccer
CFD 1
CFD 2
CFD 3
Max. Noise %
0
10
20
30
40
50
60
0.1%
1%
10%
Error %

Online Code

- https://codeocean.com/2018/06/10/xplode-colon-explaining-repaired-data-with-cfds/code

- http://adrem.uantwerpen.be/joerirammelaere