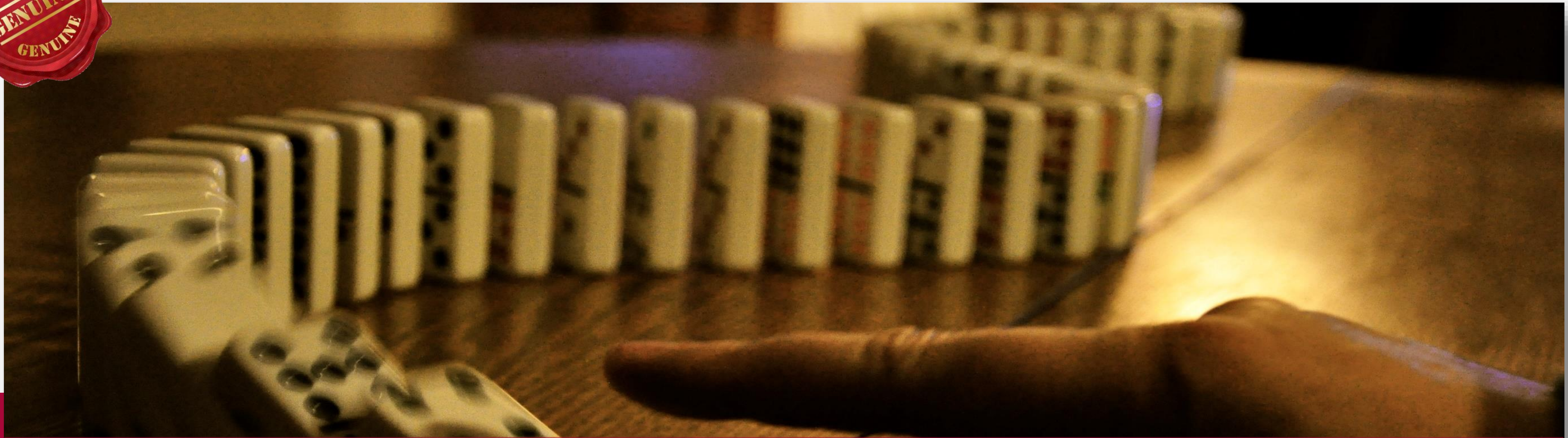
VLDB2018

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

LIS
LABORATOIRE D'INFORMATIQUE & SYSTÈMES

HPI Hasso Plattner Institut
Digital Engineering • Universität Potsdam



Discovery of Genuine Functional Dependencies from Relational Data with Missing Values

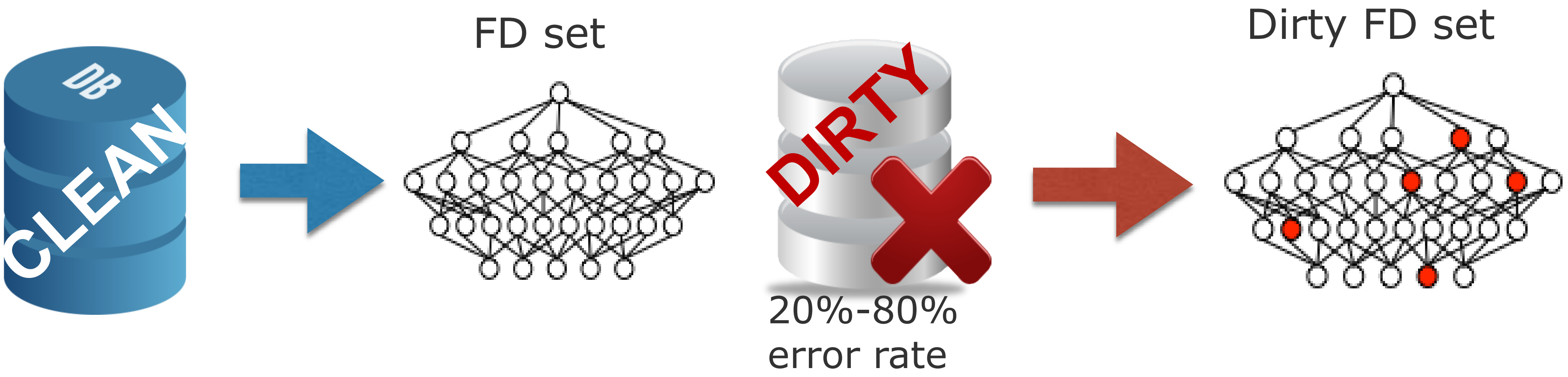
Laure Berti-Equille (LIS, Aix Marseille Uni)
Hazar Harmouch (HPI, Potsdam Uni)
Felix Naumann (HPI, Potsdam Uni)
Noël Novelli (LIS, Aix Marseille Uni)
Saravanan Thirumuruganathan (QCRI)

VLDB 2018
Dependency Discovery and Data Models Session
Rio de Janeiro-Brazil
29^{th} August 2018

Functional Dependencies (FDs)

- **More than 20 years of contributions and a great range of applications**
  - Schema normalization
  - Query optimization
  - Reverse engineering
  - Data quality profiling
  - Data cleaning
  - …

- **However, FDs are rarely discovered from perfectly clean data…**



NULLs in FD Discovery

- **Common strategies for handling missing values:**
  - Imputing missing values
  - Skipping the tuples with missing values
  - Using NULL semantics:
    - NULL-EQ: "All NULL values are identical"
    - NULL-NOT-EQ: "Every NULL value is distinct"
  - Discovery of partial FDs (approximate FDs)

None of these approaches is adequate to discover **genuine FDs**

Detailed Mechanism
An Illustrative Example

R_0	A	B	C
t_1	0	1	1
t_2	0	1	1
t_3	1	1	1
t_4	1	0	1

F_0 Deg.	FDs discovered from R_0
0	$A \to C$ $B \to C$
1	$A \to^{1} B$ $\{(t_3\|t_4)\}$ $B \to^{1} A$ $\{t_3\}$ $C \to^{1} B$ $\{t_4\}$
2	$C \to^{2} A$ $\{(t_1,t_2)\|(t_3,t_4)\}$

R_1	A	B	C
t_1	0	1	1
t_2	0	1	1
t_3	1	$\perp$	1
t_4	1	0	1

F_1 Deg.	FDs discovered from R_1
0	$A \to C$ $B \to C$ $B \to A$ (fake)
1	$A \to^{1} B$ $\{(t_3\|t_4)\}$
2	$C \to^{2} A$ $\{(t_1,t_2)\|(t_3,t_4)\}$ $C \to^{2} B$ $\{(t_3,t_4)\}$ (ghost)

R_2	A	B	C
t_1	0	1	1
t_2	0	1	$\perp$
t_3	1	$\perp$	1
t_4	1	0	1

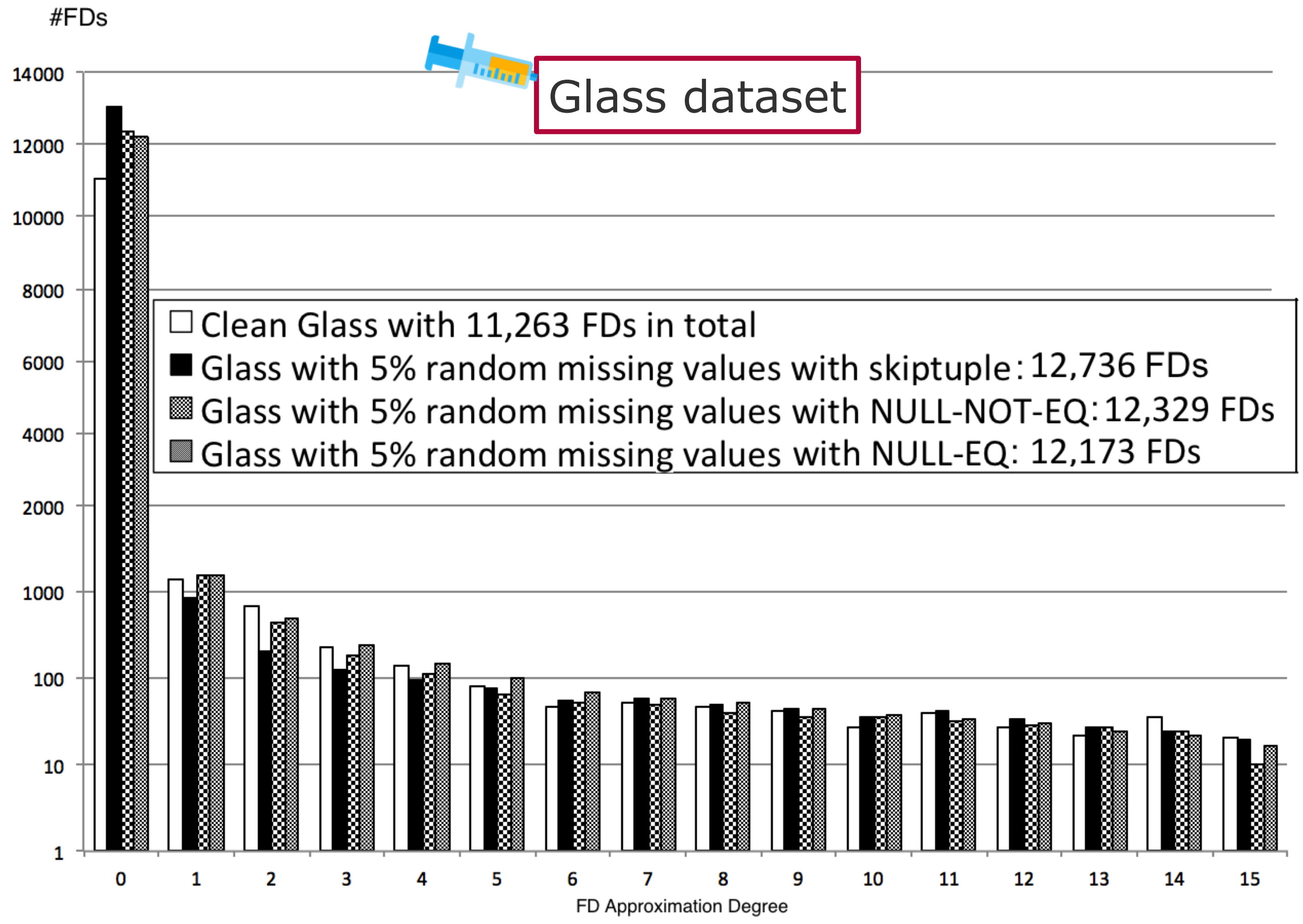

F_2 Deg.	FDs discovered from R_2
0	$B \to A$ (fake)
1	$A \to^{1} B$ $\{(t_3\|t_4)\}$ $C \to^{1} A$ $\{t_1\}$ (fake) $B \to^{1} C$ $\{(t_1\|t_2)\}$ (ghost) $A \to^{1} C$ $\{(t_1\|t_2)\}$ (ghost)
2	$C \to^{2} B$ $\{(t_1,t_3)\|(t_1,t_4)\|(t_3,t_4)\}$ (ghost)

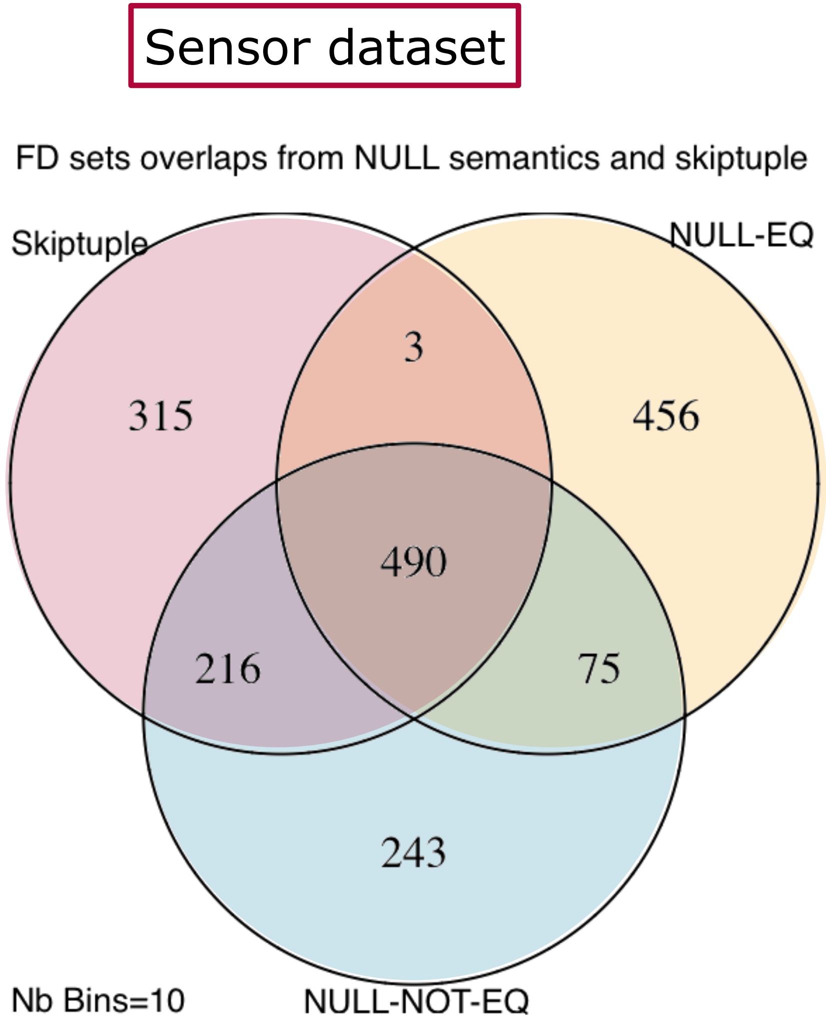
Genuine
(same approx. deg.)

Ghost
(higher approx. deg.)

Fake
(lower approx. deg.)

Our Observations (1/2)
FDs are extremely sensitive to missing values

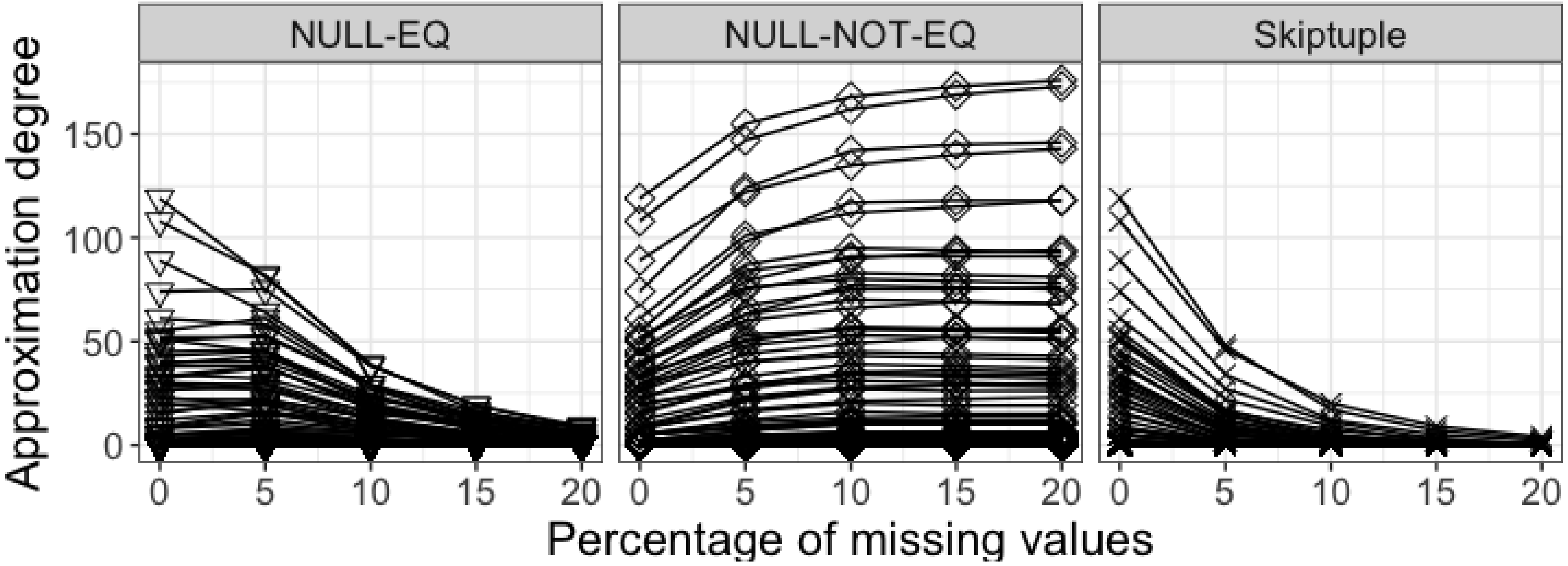




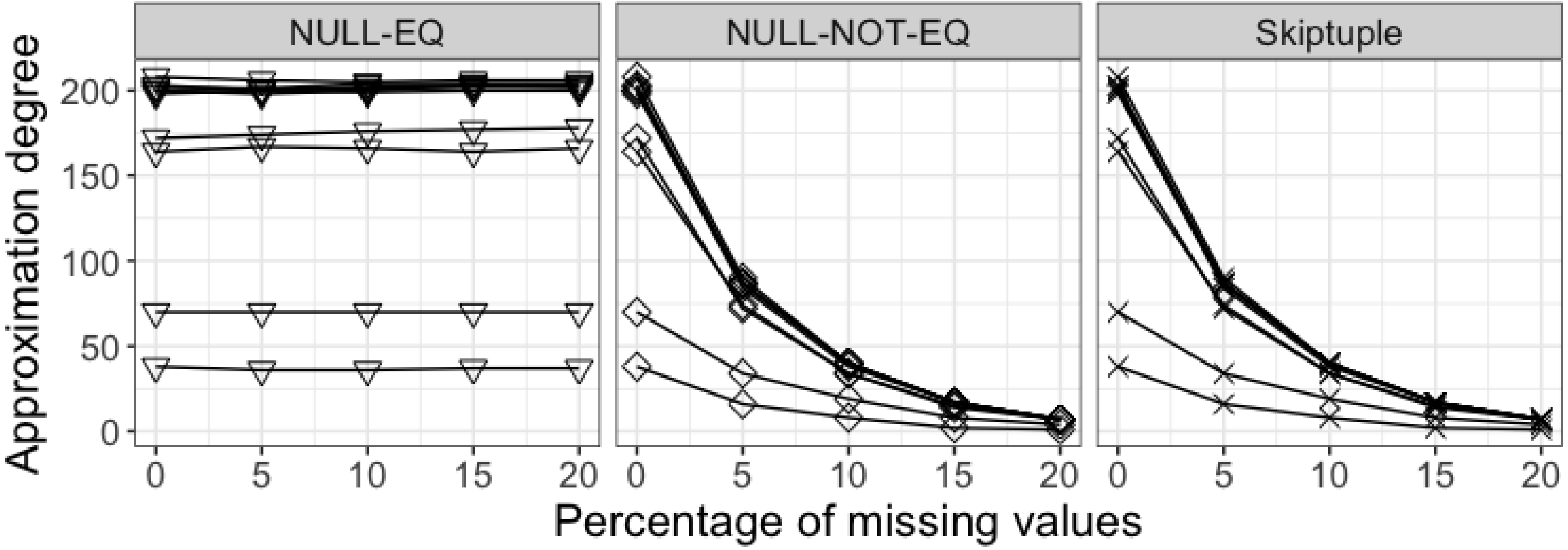
Our Observations (2/2)

Injecting NULLs in RHS and LHS of FDs in Glass dataset

RHS



LHS



Problem
Solutions Overview

- **Given**
  - R: An *incomplete* relation
  - The set of FDs discovered from the *complete part* of R
- **Objective**
  - Estimate the genuineness score of each FD discovered from R
  - Use the top-k percent FDs as genuine FDs
- **Proposed Techniques**
  - Probabilistic imputation with efficient enumeration
  - Approximation based on Monte Carlo sampling of possible worlds
  - **Per Value (PV) and Per Tuple (PT) likelihood**

FD Genuineness Scores
Per Value and Per Tuple

- The likelihood that an $(\mathrm{FD}: X \rightarrow A)$ holds for a value V_X is:

$$Likelihood(X \rightarrow A, V_X) = \frac{|V_X, V_A|}{|V_X|}$$

- **Genuineness scores Per Value**

$$Genuineness_{PV}(X \rightarrow A) = \frac{\sum_{V_X \in Distinct(X)} Likelihood(X \rightarrow A, V_X)}{|Distinct(X)|}$$

- **Genuineness scores Per Tuple**

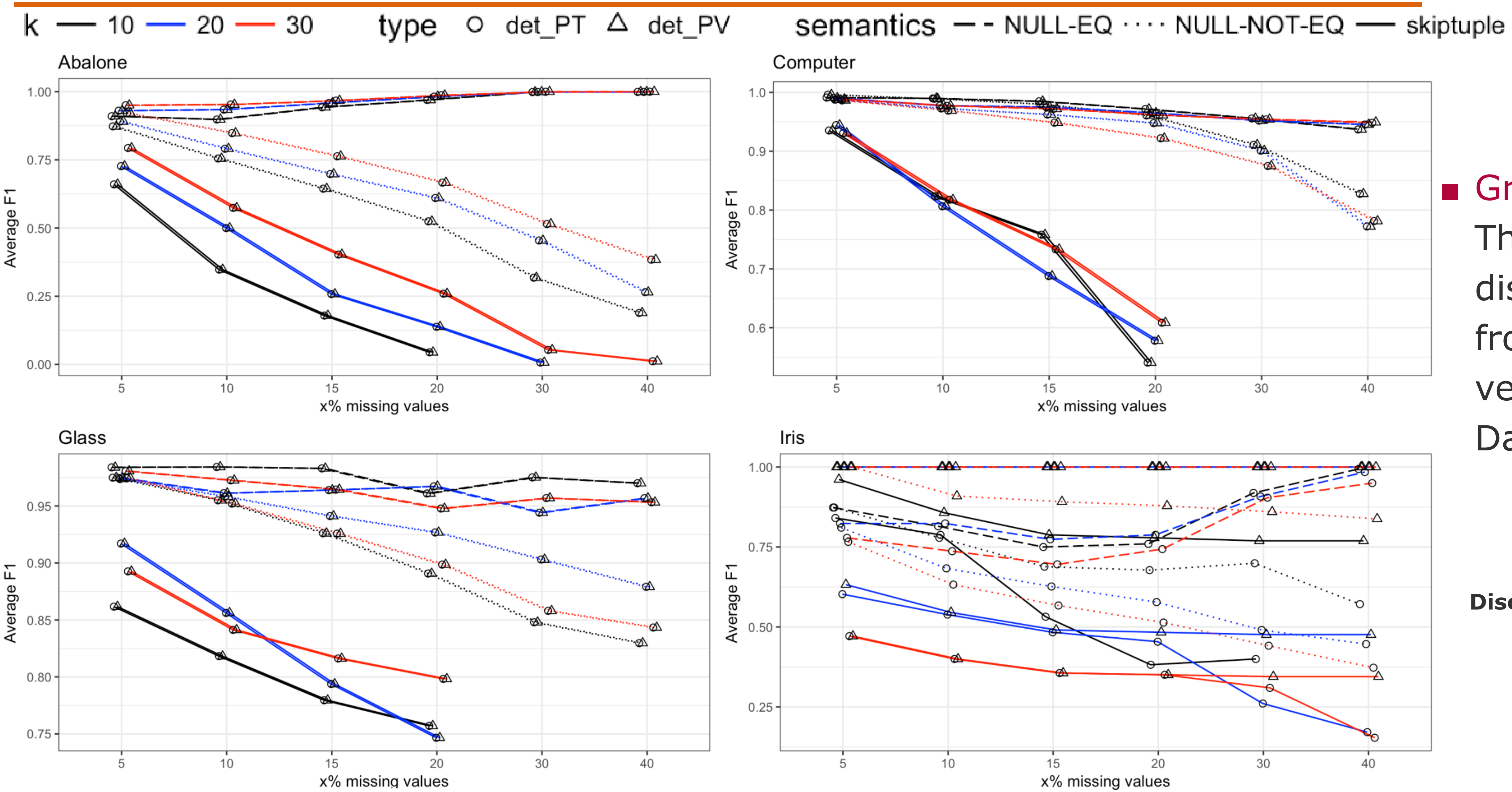
$$Genuineness_{PT}(X \rightarrow A) = \frac{\sum_{V_X \in Distinct(X)} |V_X, V_A|}{\sum_{V_X \in Distinct(X)} |V_X|}$$

Experimental Setup

- **Datasets**: We used five real-world datasets
  - Four datasets are originally complete (We randomly injected a varying percentage of missing value from 5% to 40% in the dataset attributes).
  - The Sensor dataset includes original missing values.
- **FD discovery:** using FUN algorithm
- We select as genuine FDs, the ones having PV and PT scores greater than a predefined top-k threshold.

Datasets	[#]Att.	[#]Rows	[#]Distinct (min;max)	[#]Missing	[#]FDs						
					$\alpha = 0$	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha \geq 5$	[#]Total
Iris	5	150	(3;43)	10%-40%	5	2	1	1	7	59	80
Abalone	9	4,177	(3;2,429)	10%-40%	783	219	122	57	56	1,067	2,313
Computer hardware	9	209	(15;209)	10%-40%	3,046	193	199	168	92	1,422	5,129
Glass identification	10	214	(6;214)	10%-40%	8,624	1,156	536	166	84	687	11,263
Sensor	8	2,313,681	(137;10,283)	96,733	Skiptuple						
Sensor 10 Bins			$\rightarrow$ 10		397	29	10	14	11	563	1,024
Sensor 100 Bins			$\rightarrow$ 100		432	40	10	0	3	539	1,024
Sensor 1000 Bins			$\rightarrow$ 1000		427	44	7	0	3	543	1,024

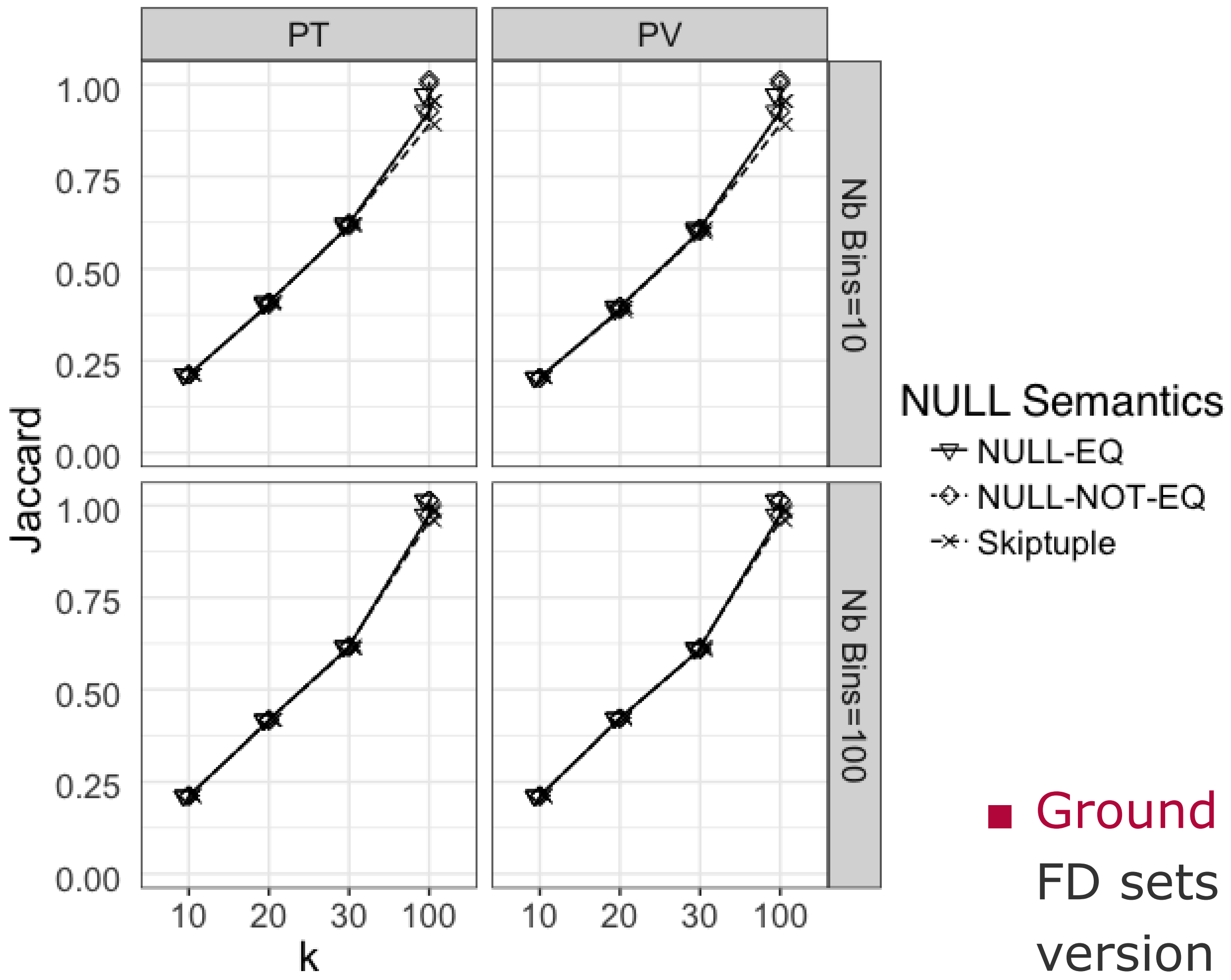
Quality performance evaluation (1/2)
PV and PT scoring-based method: Average F1

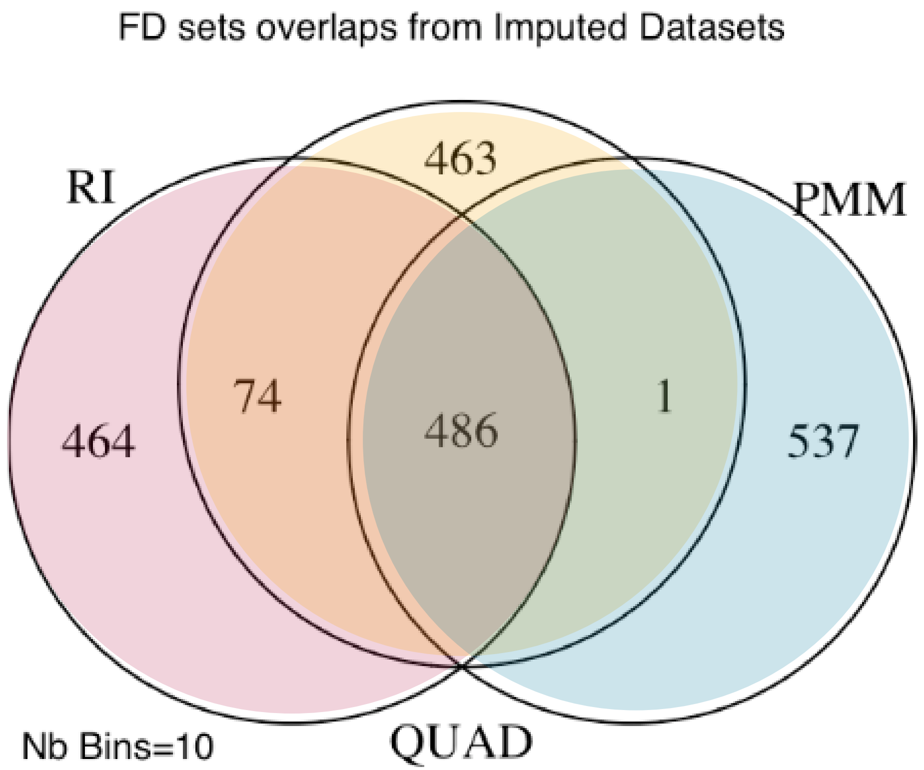


- Ground Truth: The FD sets discovered from the clean version of each Dataset.

Quality performance evaluation (2/2)

PV and PT scoring-based method: Jaccard coefficient





- Ground Truth: The intersection of the FD sets discovered from 3 imputed version of Sensor Dataset.

Summary and Future Work

- Missing values impair FD discovery results by causing spurious FDs (Fake FDs) and the omission of valid FDs (Ghost FDs).
- Our FD-scoring methods can find 100% of genuine FDs that would have been obtained by multiple imputation strategies in reasonable time
- Pre- and post-processing efforts for FD discovery are minimized.

- Investigate other data quality problems impairing the final FD results.
- Extend and apply our technique to “disguised” missing values and misused default values