
Qualitative Data Cleaning

Data Cleaning: Overview and Emerging Challenges Part 1

Xu Chu Ihab Ilyas



Many Definitions and One Goal

"Extract Value from Data"

- For that we ..
 - Remove errors
 - Fill missing info
 - Transform units and formats
 - Map and align columns
 - Remove duplicate records
 - Fix integrity constraints violations

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

NYtimes August, 2014

Yes big data is a big business opportunity, but the business value won't be realized if the information isn't governed

Forbes Business

Many Technical Challenges

□ Record Linkage and Deduplication

- Similarity measures
- Machine learning for classifying pairs as duplicates or not (unsupervised, supervised, and active)
- Clustering and handling of transitivity
- Merging and consolidation of records

A major firm spends 6 months on a single deduplication project of a subset of their data sources

Example: Data Deduplication

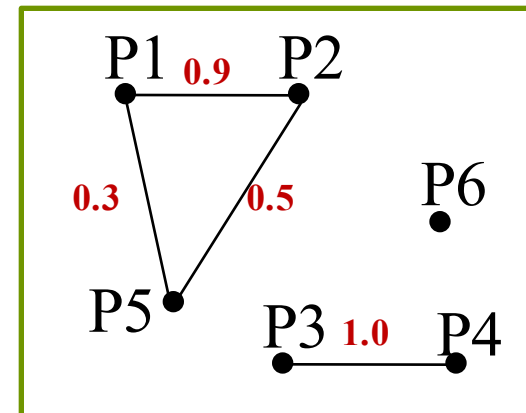
Unclean Relation

ID	name	ZIP	Income
P1	Green	51519	30k
P2	Green	51518	32k
P3	Peter	30528	40k
P4	Peter	30528	40k
P5	Gree	51519	55k
P6	Chuck	51519	30k

Clean Relation

ID	name	ZIP	Income
C1	Green	51519	39k
C2	Peter	30528	40k
C3	Chuck	51519	30k

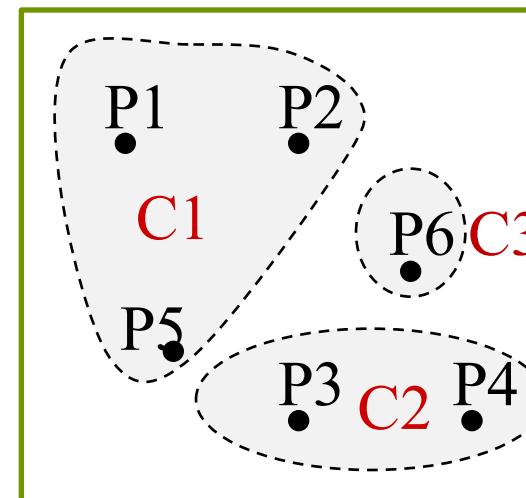
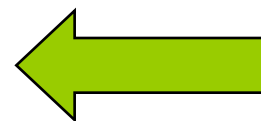
Compute
Pair-wise
Similarity



Cluster
Similar
Records



Merge
Clusters



Many Technical Challenges

□ Missing Values

- Interpreting different types of `Nulls`
- Certain answer semantics on possible worlds (many.. many papers)
- Closed world vs. open-world assumptions and multiple interesting hardness results

Most real data collected from sensors, surveys, agents, have a high percentage of N/A or nulls, special values (99999) etc.

Many Technical Challenges

□ **More Complex Integrity Constraints**

- A declarative language to express data quality rules
- Ad-hoc repair algorithm to repair violations for each data quality formalism under certain minimality requirements
- Limited expressiveness (e.g., FD) to get tangible results

Unfortunately rarely expressed in practice. Most curation tools are rule-based implemented in imperative language

Example ICs

	ID	FN	LN	ROLE	CITY	ST	SAL
t_1	105	Anne	Nash	M	NYC	NY	110
t_2	211	Mark	White	E	SJ	CA	80
t_3	386	Mark	Lee	E	NYC	AZ	75
t_4	235	John	Smith	M	NYC	NY	1200

Employee Table

Functional dependency:

$City \rightarrow ST$

Example ICs

	ID	FN	LN	ROLE	CITY	ST	SAL
t_1	105	Anne	Nash	M	NYC	NY	110
t_2	211	Mark	White	E	SJ	CA	80
t_3	386	Mark	Lee	E	NYC	AZ	75
t_4	235	John	Smith	M	NYC	NY	1200

Employee Table

Business Rule:

Two employees of the same role, the one who lives in NYC cannot earn less than the one who does not live in NYC

Example ICs

	ID	FN	LN	ROLE	CITY	ST	SAL
t_1	105	Anne	Nash	M	NYC	NY	110
t_2	211	Mark	White	E	SJ	CA	80
t_3	386	Mark	Lee	E	NYC	AZ	75
t_4	235	John	Smith	M	NYC	NY	1200

Employee Table

Business Rule:

Two employees of the same role in the same city, their salary difference cannot be greater than 100

Common Data Quality Issues

ID	Name	ZIP	City	State	Income
1	Green	60610	Chicago	IL	31k
2	Green	60611	Chicago	IL	32k
3	Peter	11507	New York	NY	40k
4	John	11507	New York	NY	40k
5	Gree	90057	Los Angeles	CA	55k
6	Chuck	90057	Los Angeles	CA	30k

Missing Value

Integrity Constraint Violation

Syntactic Error

Duplicates

Data Cleaning Process

- Error Detection
 - Qualitative
 - Quantitative (outlier detection)

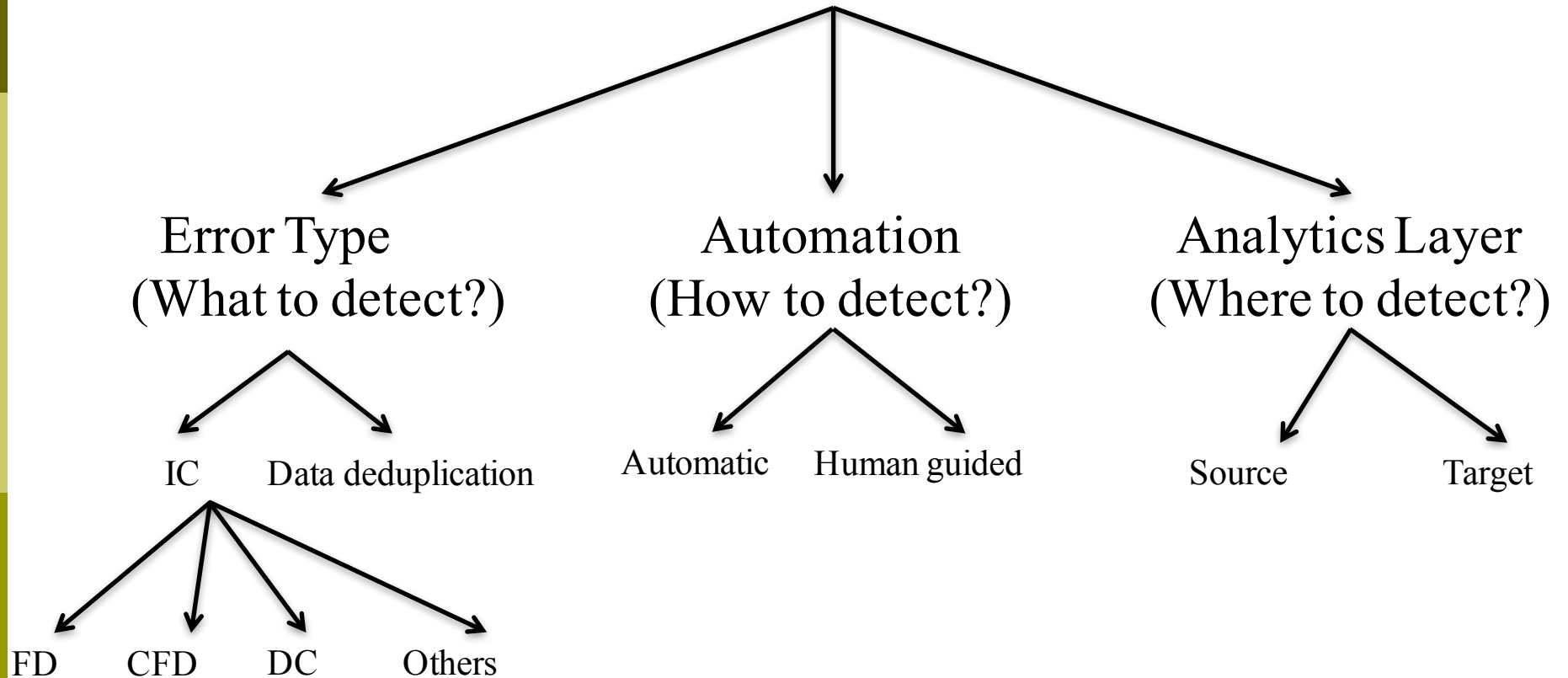
- Error Repairing
 - Transformation scripts
 - Human involvement

We Will Not Cover

- Details of Deduplication
 - Multiple surveys and tutorials
- Data Profiling: discovering FDs, INDs, etc.
 - Wenfei Fan and Floris Geerts synthesis lecture book
 - Ziawasch Abedjan et al. tutorial
- Consistent Query Answering
 - Leo Betrossi synthesis lecture book

Error Detection Techniques Taxonomy

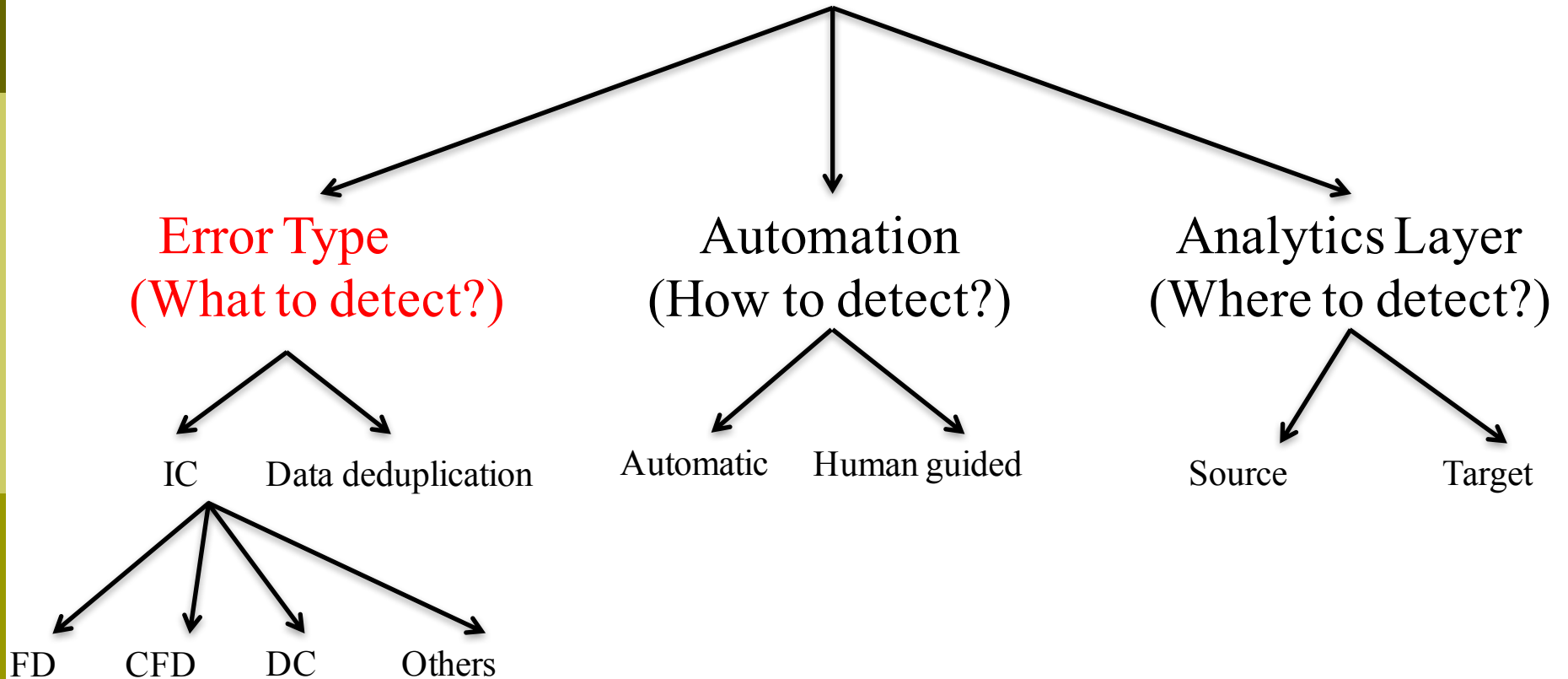
Qualitative Error Detection Techniques



[Ilyas and Chu, Foundations and Trends in Database Systems, 2015]

Error Detection Techniques Taxonomy

Qualitative Error Detection Techniques



FDs and CFDs [Bohannon et al, ICDE 2007]

□ Functional Dependency (FD):

$X \rightarrow Y$

- Example: $City \rightarrow ST$ or $Name, Phone \rightarrow ID$

□ Conditional Functional Dependency (CFD):

$(X \rightarrow Y, T_p)$

- An FD defined on a subset of the data
- Example:
 - $ZIP \rightarrow Street$ is valid on subset of the data where $Country = \text{"England"}$
 - $AC = 020 \rightarrow City = London$

Matching Dependencies (MDs) [Fan et al, VLDB 2009]

	FN	LN	St	City	AC	Post	Phn	Item
Tran	Robert	Brady	5 Wren St	Ldn	020	WC1H 9SE	3887644	watch
	Robert	Brady	Null	Ldn	020	WC1E 7HX	3887644	necklace

Master: Card

FN	LN	St	City	AC	Zip	Tel
Robert	Brady	5 Wren St	Ldn	020	WC1H 9SE	3887644

MD: $\text{Tran}[\text{LN}, \text{City}, \text{St}, \text{Post}] = \text{card}[\text{LN}, \text{City}, \text{St}, \text{Zip}] \wedge$
 $\text{Tran}[\text{FN}] \approx \text{Card}[\text{FN}] \rightarrow \text{Tran}[\text{FN}, \text{Phn}] \Leftrightarrow \text{Card}[\text{FN}, \text{Tel}]$

[Fan et al, SIGMOD 2011]

Denial Constraints (DCs) [Chu et al, VLDB 2013]

Formal Definition:

$$\varphi : \forall t_\alpha, t_\beta, t_\gamma, \dots \in R, \neg(P_1 \wedge \dots \wedge P_m)$$

$$P_i: t_x.A \theta t_y.B \text{ or } t_x.A \theta c$$

$x, y \in \{\alpha, \beta, \dots\}$, and $A, B \in R$, c is a constant

- A universal constraint dictates a set of predicate cannot be true together
- Each predicate express a relationship between two cells, or a cell and a constant

Denial Constraints (DCs)

Functional dependency:

$CITY \Rightarrow ST$

$\forall t_\alpha, t_\beta \in Emp, \neg(t_\alpha.CITY = t_\beta.CITY \wedge t_\alpha.ST \neq t_\beta.ST)$

Business Rule:

Two employees of the same Role, the one who lives in NYC cannot earn less than the one who does not live in NYC

$\forall t_\alpha, t_\beta \in Emp, \neg(t_\alpha.ROLE = t_\beta.ROLE \wedge t_\alpha.CITY = \text{“NYC”} \wedge t_\beta.CITY \neq \text{“NYC”} \wedge t_\alpha.SAL < t_\beta.SAL)$

Other ICs

- CINDs [Ma et al, TCS 2014]

- Metric Functional Dependencies [Koudas et al, ICDE 2009]

- Dependable Fixes
 - Editing Rules [Fan et al, VLDB 2010]
 - Fixing Rules [Wang and Tang, SIGMOD 2014]
 - Sherlock Rules [Interlandi and Tang, ICDE 2015]

Constraint Languages

Language expressiveness

FDs

CFDs

...

DCs

Programmatic Interface



Reasoning and discovery complexity

Integrity Constraints Discovery

- Schema Driven
 - Usually sensitive to the size of the schema
 - Good for long thin tables!

- Instance Driven
 - Usually sensitive to the size of the data
 - Good for fat short tables!

- Hybrid
 - Try to get the best of both worlds

Integrity Constraints Discovery

□ FD Discovery:

■ TANE: Schema-driven

□ [Huhtala et al, Computer Journal 1999]

■ FASTFD: Instance-driven

□ [Wyss et al, DaWaK, 2001]

■ Hybrid (This Sigmod)

□ [Papenbrock et al, SIGMOD 2016]

□ DC Discovery:

■ FASTDC: Instance-driven [Chu et al, VLDB 2013]

Integrity Constraints Discovery

□ FD Discovery:

■ TANE: Schema-driven

□ [Huhtala et al, Computer Journal 1999]

■ FASTFD: Instance-driven

□ [Wyss et al, DaWaK, 2001]

■ Hybrid (This Sigmod)

□ [Papenbrock et al, SIGMOD 2016]

□ DC Discovery:

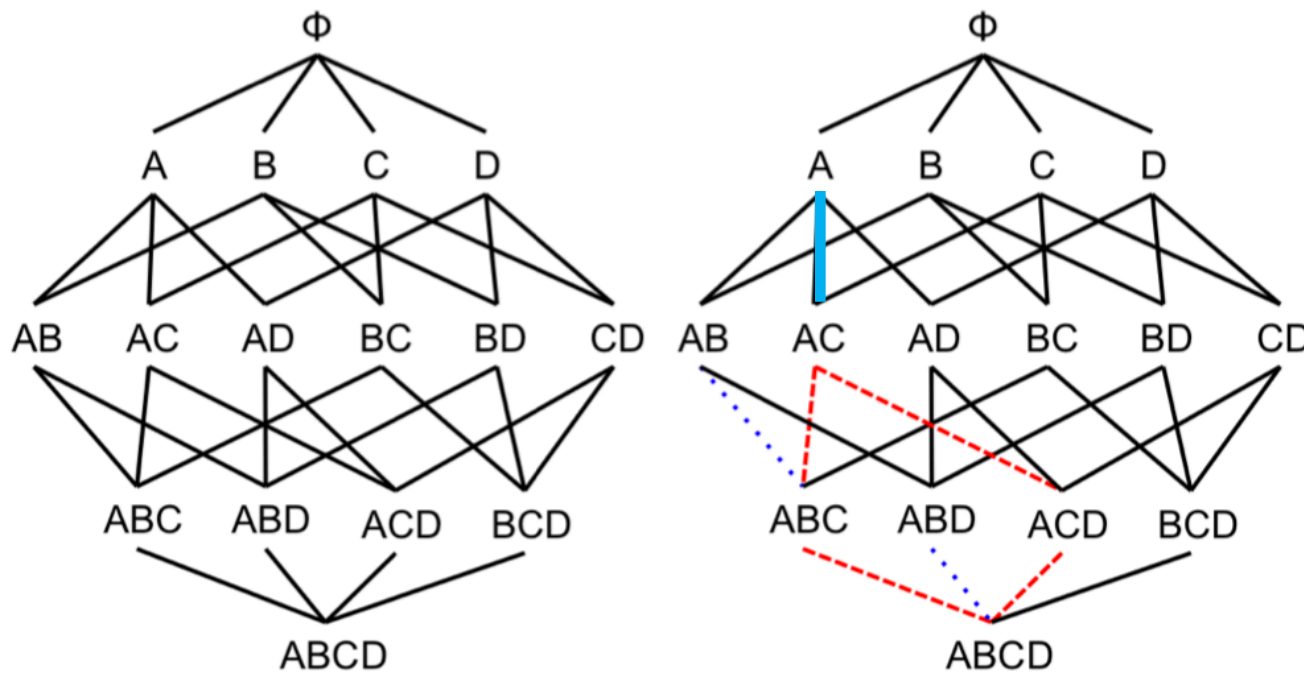
■ FASTDC: Instance-driven [Chu et al, VLDB 2013]

FD Discovery

- Given a relational instance I of schema R , where $|R| = m$, find (all) **minimal, non-trivial** FDs that are **valid** on I . An FD is
 - **Valid** on I if there does not exist two tuples that violate the FD
 - **Minimal** if removing an attribute from its LHS makes it invalid
 - **Trivial** if the RHS is a subset of the LHS
- We want FDs with **only one attribute** in RHS

TANE [Huhtala et al, Computer Journal 1999]

□ Generate space of FDs



(a) Space of FDs.

(b) Candidate FDs pruned if $A \rightarrow C$ is valid

TANE

□ FD Validation

$$\Pi_X = \{\{t_1\}, \{t_2, t_3\}, \{t_4\}\}$$

$$\Pi_Y = \{\{t_1, t_2, t_3\}, \{t_4\}\}$$

$$\Pi_{XY} = \{\{t_1\}, \{t_2, t_3\}, \{t_4\}\}$$

$X \rightarrow Y$ is a valid FD if and only if

$$|\Pi_X| = |\Pi_{X \cup Y}|$$

DC Discovery: Axioms

Triviality

$\forall P_i, P_j$, if $P_i \in \text{Imp}(P_j)$ then $\neg(\bar{P}_i \wedge P_j)$ is a trivial DC

$$\forall t_\alpha, t_\beta \in R, \neg(t_\alpha \cdot \text{SAL} = t_\beta \cdot \text{SAL} \wedge t_\alpha \cdot \text{SAL} > t_\beta \cdot \text{SAL})$$

ϕ	=	\neq	>	<	\geq	\leq
$\bar{\phi}$	\neq	=	\leq	\geq	<	>
$\text{Imp}(\phi)$	=, \geq , \leq	\neq	>, \geq , \neq	<, \leq , \neq	\geq	\leq

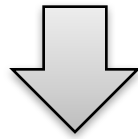
DC Discovery: Axioms

Augmentation

If $\neg(P_1 \wedge \dots \wedge P_n)$ is valid, then $\neg(P_1 \wedge \dots \wedge P_n \wedge Q)$ is also valid

Not Minimal

$$\forall t_\alpha, t_\beta \in R, \neg(t_\alpha.ZIP = t_\beta.ZIP \wedge t_\alpha.ST \neq t_\beta.ST)$$



$$\forall t_\alpha, t_\beta \in R, \neg(t_\alpha.ZIP = t_\beta.ZIP \wedge t_\alpha.ST \neq t_\beta.ST \wedge t_\alpha.SAL < t_\beta.SAL)$$

DC Discovery

Given a relational schema R and an instance I , find all **non-trivial, minimal DCs that hold on I**

Focus on DCs involving at most two tuples

FASTDC [Chu et al, VLDB 2013]

	<i>TID</i>	<i>I(String)</i>	<i>M(String)</i>	<i>S(Double)</i>
Employee	t_1	<i>A1</i>	<i>A1</i>	<i>50</i>
	t_2	<i>A2</i>	<i>A1</i>	<i>40</i>
	t_3	<i>A3</i>	<i>A1</i>	<i>40</i>

□ The space of predicates

$$P_1 : t_\alpha . I = t_\beta . I \quad P_3 : t_\alpha . M = t_\beta . M \quad P_5 : t_\alpha . S = t_\beta . S \quad P_{11} : t_\alpha . I = t_\alpha . M$$

$$P_2 : t_\alpha . I \neq t_\beta . I \quad P_4 : t_\alpha . M \neq t_\beta . M \quad P_6 : t_\alpha . S \neq t_\beta . S \quad P_{12} : t_\alpha . I \neq t_\alpha . M$$

$$P_7 : t_\alpha . S > t_\beta . S \quad P_{13} : t_\alpha . I = t_\beta . M$$

$$P_8 : t_\alpha . S \leq t_\beta . S \quad P_{14} : t_\alpha . I \neq t_\beta . M$$

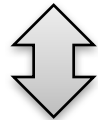
$$P_9 : t_\alpha . S < t_\beta . S$$

$$P_{10} : t_\alpha . S \geq t_\beta . S$$

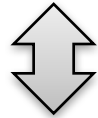
□ Any combination of predicates constitutes a candidate DC

FASTDC

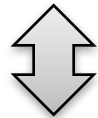
$\neg(P_i \wedge P_j \wedge P_k)$ is a valid DC w.r.t. I



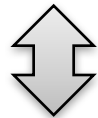
For every tuple pair in I, P_i, P_j, P_k cannot be true together



For every tuple pair in I, at least one of P_i, P_j, P_k is false



For every tuple pair in I, at least one of $\overline{P_i}, \overline{P_j}, \overline{P_k}$ is true



$\overline{P_i}, \overline{P_j}, \overline{P_k}$ covers the *set of true predicates for every tuple pair*

FASTDC

<i>TID</i>	<i>I(String)</i>	<i>M(String)</i>	<i>S(Double)</i>
t_1	<i>A1</i>	<i>A1</i>	50
t_2	<i>A2</i>	<i>A1</i>	40
t_3	<i>A3</i>	<i>A1</i>	40

Ev_i

$\langle t_2, t_3 \rangle, \langle t_3, t_2 \rangle \{P_2, P_3, P_5, P_8, P_{10}, P_{12}, P_{14}\}$

$\langle t_2, t_1 \rangle, \langle t_3, t_1 \rangle \{P_2, P_3, P_6, P_8, P_9, P_{12}, P_{14}\}$

$\langle t_1, t_2 \rangle, \langle t_1, t_3 \rangle \{P_2, P_3, P_6, P_7, P_{10}, P_{11}, P_{13}\}$

$\{P_{10}, P_{14}\}$ covers the set of true predicates for every tuple pair

$$\forall t_\alpha, t_\beta \in R, \neg(\overline{P}_{10} \wedge \overline{P}_{14})$$

$\forall t_\alpha, t_\beta \in R, \neg(t_\alpha.S < t_\beta.S \wedge t_\alpha.I = t_\beta.M)$ is a valid DC

$\{P_5, P_{10}, P_{14}\}$ covers the set of true predicates for every tuple pair

$\neg(\overline{P}_{10} \wedge \overline{P}_{14} \wedge \overline{P}_5)$ is a valid DC, **but not minimal**

FASTDC

$Evi_I \{P_2, P_3, P_5, P_8, P_{10}, P_{12}, P_{14}\}$

$\{P_2, P_3, P_6, P_8, P_9, P_{12}, P_{14}\}$

$\{P_2, P_3, P_6, P_7, P_{10}, P_{11}, P_{13}\}$

$P_8 \in Imp(\bar{P}_6)$

$P_{10} \in Imp(\bar{P}_6)$

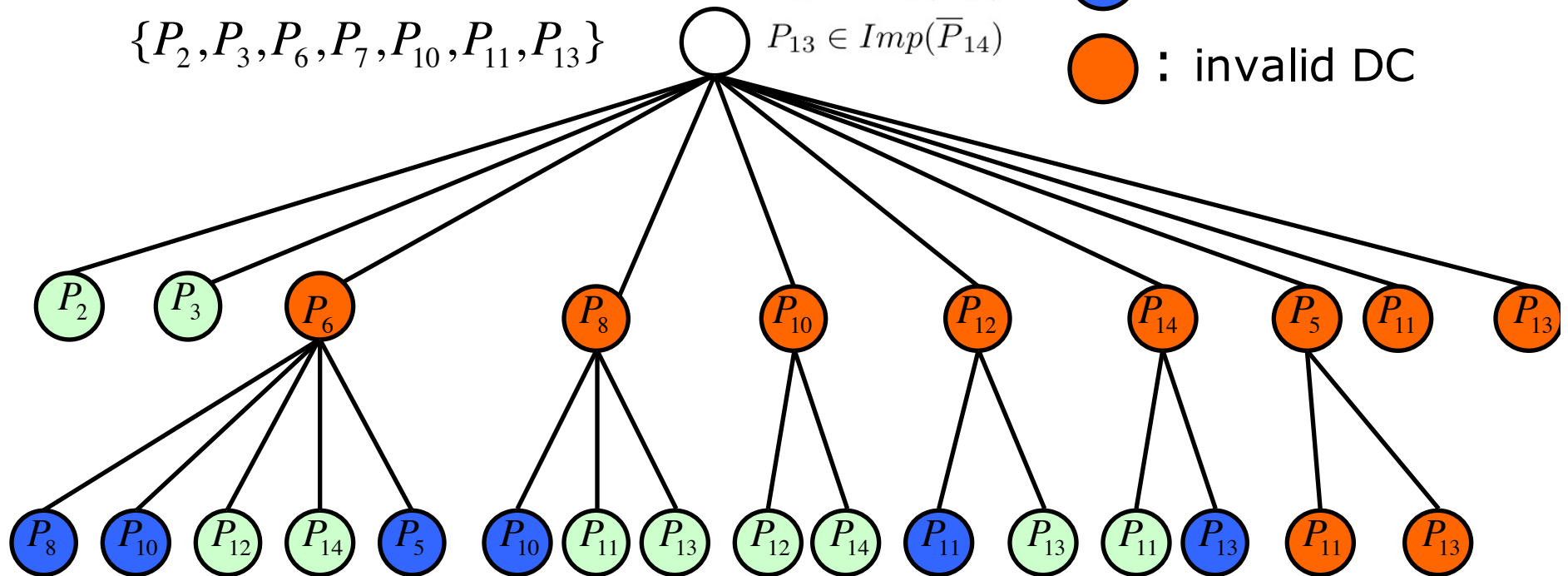
$P_{11} \in Imp(\bar{P}_{12})$

$P_{13} \in Imp(\bar{P}_{14})$

○ : valid DC

● : pruned branch

● : invalid DC



FASTDC

TID	FN	LN	GD	AC	PH	CT	ST	ZIP	MS	CH	SAL	TR	STX	MTX	CTX
t_1	Mark	Ballin	M	304	232-7667	Anthony	WV	25813	S	Y	5000	3	2000	0	2000
t_2	Chunho	Black	M	719	154-4816	Denver	CO	80290	M	N	60000	4.63	0	0	0
t_3	Annja	Rebizant	F	636	604-2692	Cyrene	MO	64739	M	N	40000	6	0	4200	0
t_4	Annie	Puerta	F	501	378-7304	West Crossett	AR	72045	M	N	85000	7.22	0	40	0
t_5	Anthony	Landram	M	319	150-3642	Gifford	IA	52404	S	Y	15000	2.48	40	0	40
t_6	Mark	Murro	M	970	190-3324	Denver	CO	80251	S	Y	60000	4.63	0	0	0
t_7	Ruby	Billinghurst	F	501	154-4816	Kremlin	AR	72045	M	Y	70000	7	0	35	1000
t_8	Marcelino	Nuth	F	304	540-4707	Kyle	WV	25813	M	N	10000	4	0	0	0

Key : {AC, PH}

$$\forall t_\alpha, t_\beta \in R, \neg(t_\alpha.AC = t_\beta.AC \wedge t_\alpha.PH = t_\beta.PH)$$

Domain : MS \in {S, M}

$$\forall t_\alpha \in R, \neg(t_\alpha.MS \neq S \wedge t_\alpha.MS \neq M)$$

FD : ZIP \rightarrow ST

$$\forall t_\alpha, t_\beta \in R, \neg(t_\alpha.ZIP = t_\beta.ZIP \wedge t_\alpha.ST \neq t_\beta.ST)$$

CFD : CT = Los Angeles \rightarrow ST = CA

$$\forall t_\alpha \in R, \neg(t_\alpha.CT = Los\ Angeles \wedge t_\alpha.ST \neq CA)$$

Check : SAL \geq STX

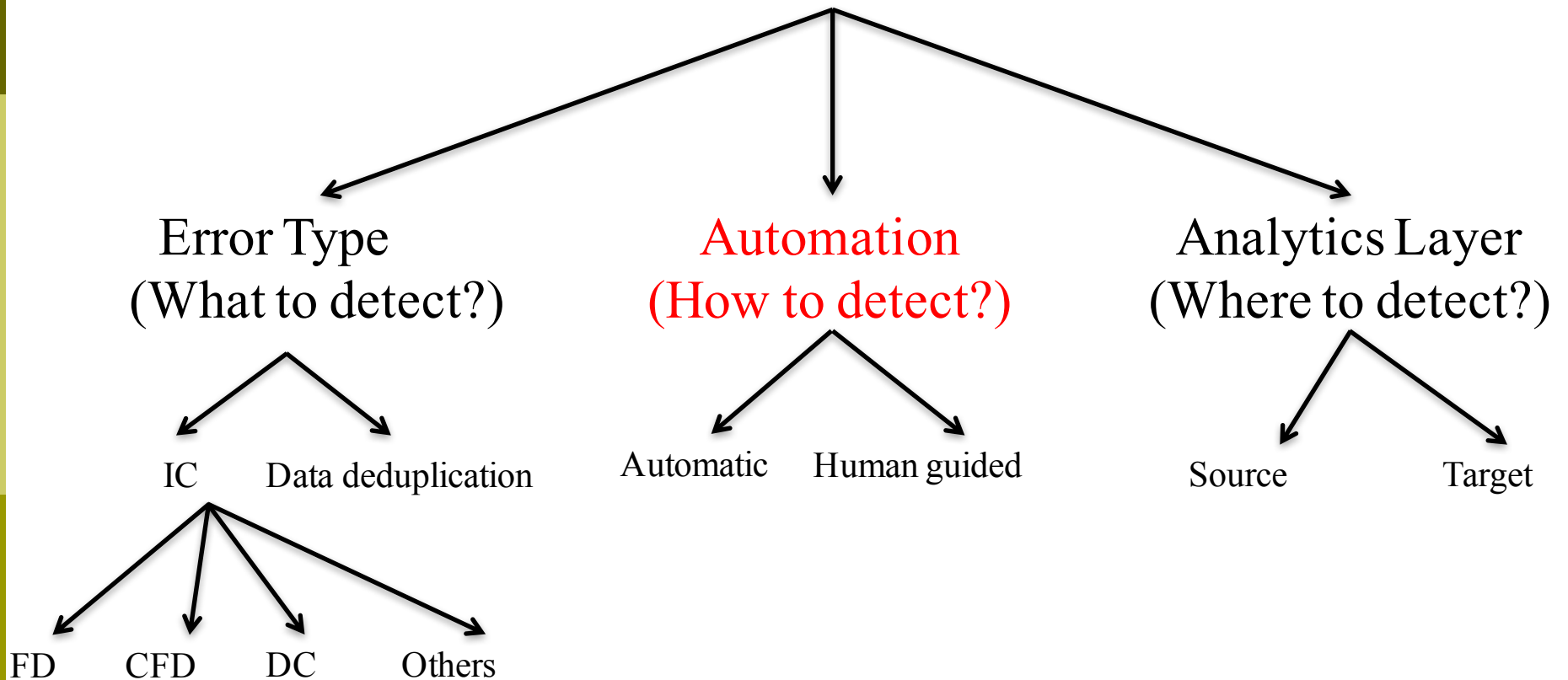
$$\forall t_\alpha \in R, \neg(t_\alpha.SAL < t_\alpha.STX)$$

Business logic

$$\forall t_\alpha, t_\beta \in R, \neg(t_\alpha.ST = t_\beta.ST \wedge t_\alpha.SAL < t_\beta.SAL \wedge t_\alpha.TR > t_\beta.TR)$$

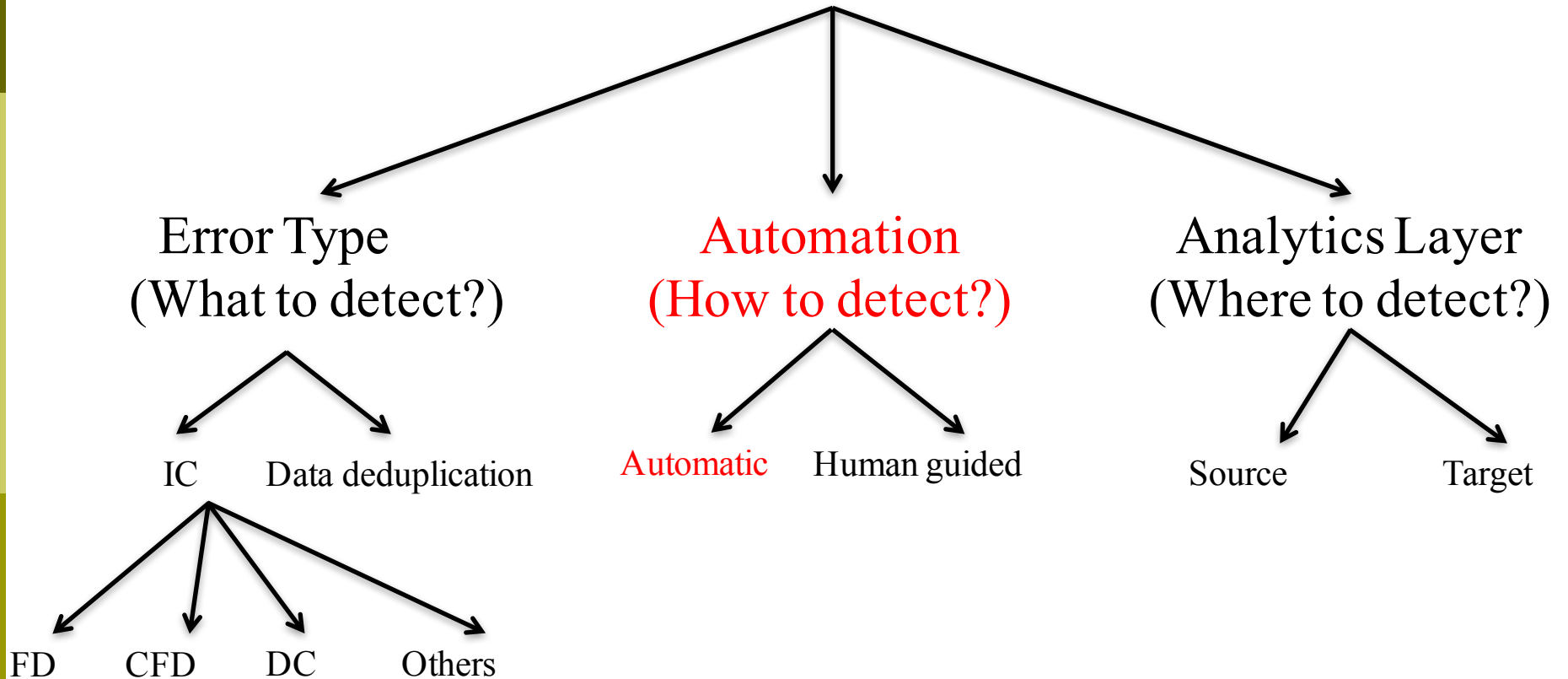
Error Detection Techniques Taxonomy

Qualitative Error Detection Techniques



Error Detection Techniques Taxonomy

Qualitative Error Detection Techniques



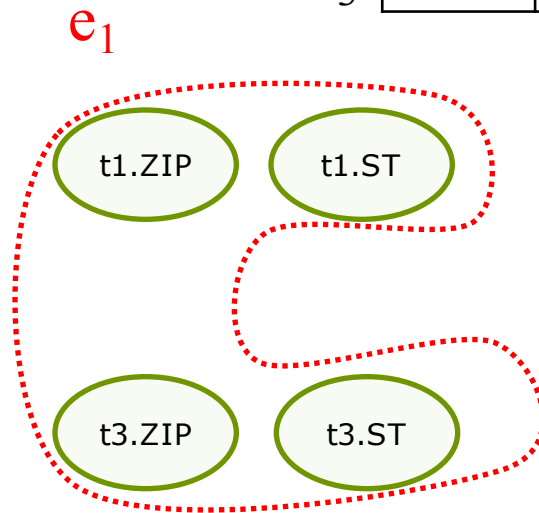
Holistic Error Detection

- Vertex: Cell in the database
- Hyperedge: A set of cells that violate a DC

	ID	FN	LN	ROLE	ZIP	ST	SAL
t_1	105	Anne	Nash	E	85376	NY	110
t_2	211	Mark	White	M	90012	NY	80
t_3	386	Mark	Lee	E	85376	AZ	75

Employee Table

Zip \rightarrow ST



[Chu et al, ICDE 2013]

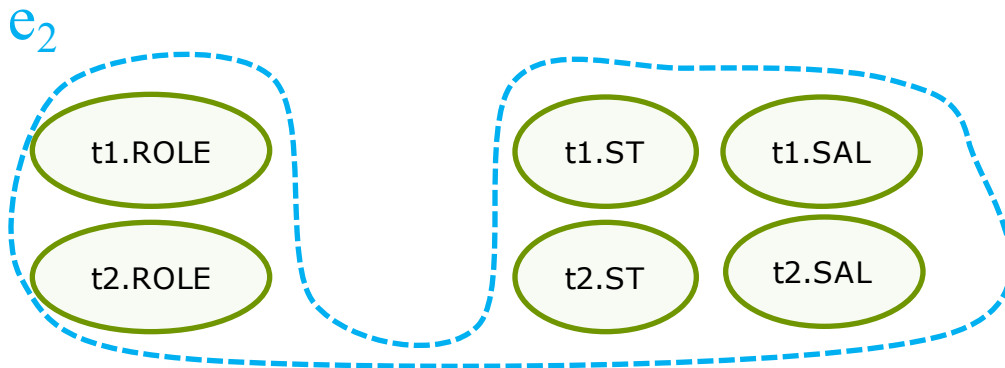
[Kolahi and Lakshmanan ICDT 2009]

Holistic Error Detection

- Vertex: Cell in the database
- Hyperedge: A set of cells that violate a DC

	ID	FN	LN	ROLE	ZIP	ST	SAL
t_1	105	Anne	Nash	E	85376	NY	110
t_2	211	Mark	White	M	90012	NY	80
t_3	386	Mark	Lee	E	85376	AZ	75

Employee Table



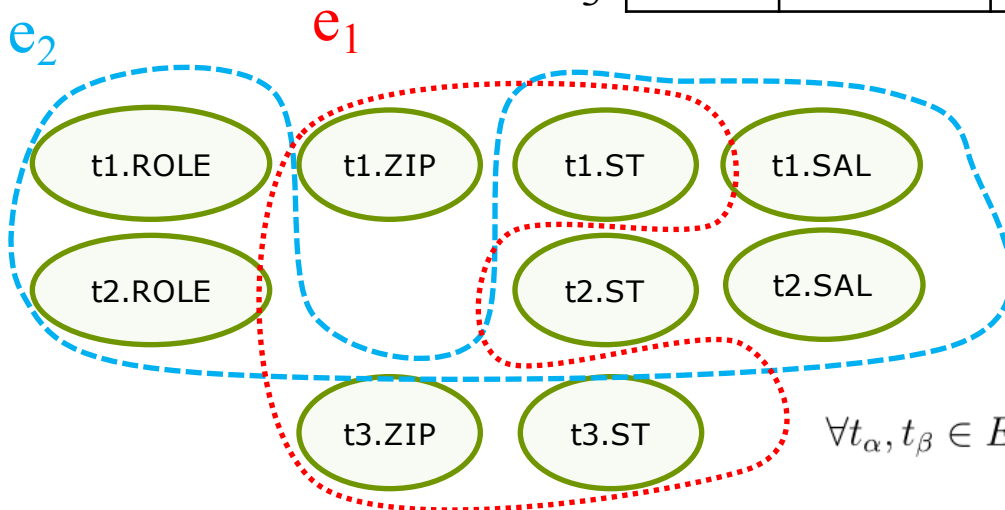
$$\forall t_\alpha, t_\beta \in Emp, \neg(t_\alpha.ST = t_\beta.ST \wedge t_\alpha.ROLE = "M" \wedge t_\beta.ROLE = "E" \wedge t_\alpha.SAL < t_\beta.SAL)$$

Holistic Error Detection

- Vertex: Cell in the database
- Hyperedge: A set of cells that violate a DC

	ID	FN	LN	ROLE	ZIP	ST	SAL
t_1	105	Anne	Nash	E	85376	NY	110
t_2	211	Mark	White	M	90012	NY	80
t_3	386	Mark	Lee	E	85376	AZ	75

Employee Table

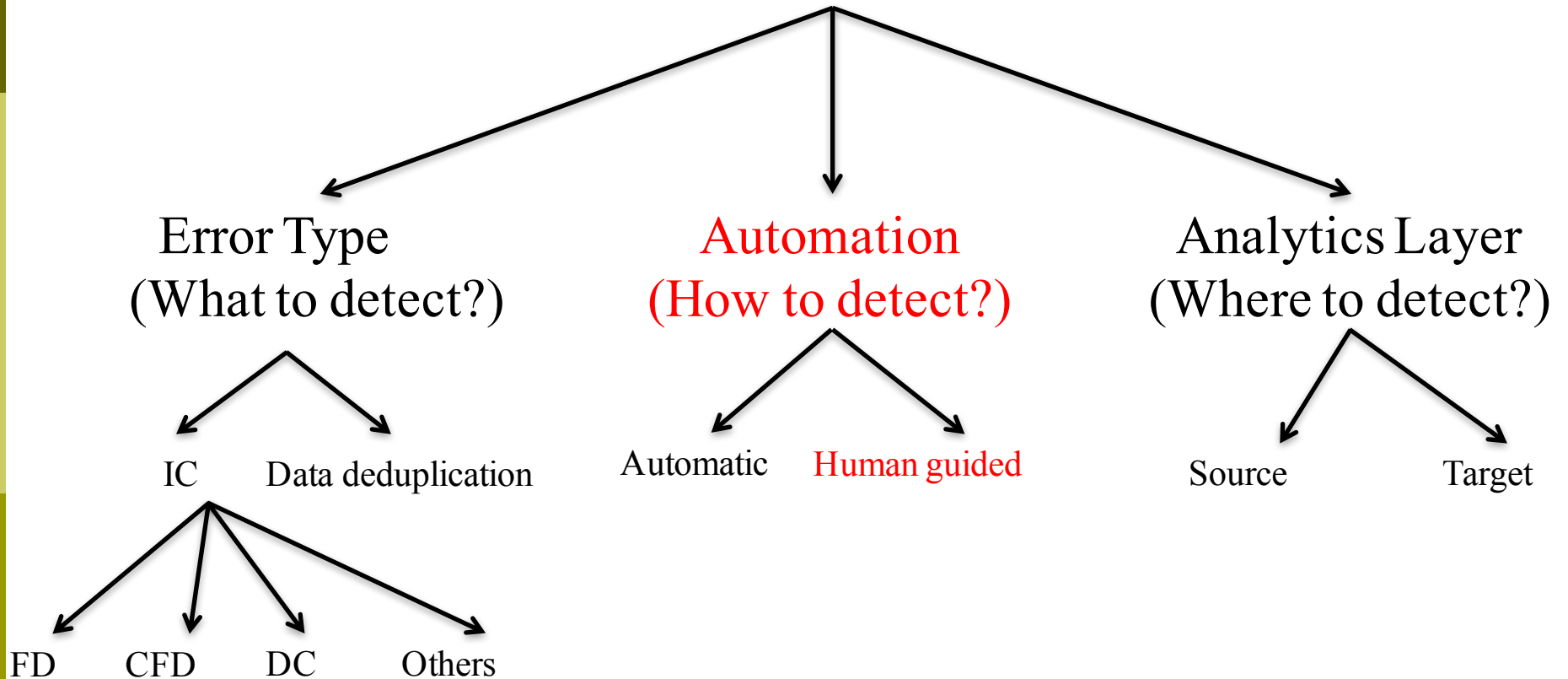


Zip \rightarrow ST

$$\forall t_\alpha, t_\beta \in Emp, \neg(t_\alpha.ST = t_\beta.ST \wedge t_\alpha.ROLE = "M" \wedge t_\beta.ROLE = "E" \wedge t_\alpha.SAL < t_\beta.SAL)$$

Error Detection Techniques Taxonomy

Qualitative Error Detection Techniques



CrowdER: [Wang et al, VLDB 2012]

□ Human-Intelligence Task (HIT)

$O(n^2) \times$

Decide Whether Two Products Are the Same or Different

Product Pair #1

Product Name	Price
iPad Two 16GB WiFi White	\$490
iPad 2nd generation 16GB WiFi White	\$469

Your Choice (Required)

They are the same product

They are different products

Reasons for Your Choice (Optional)

CrowdER: Batching Strategies

□ Pair-based HIT

$O(n^2/k) \times$

Product Pair #1

Product Name	Price
iPad Two 16GB WiFi White	\$490
iPad 2nd generation 16GB WiFi White	\$469

Your Choice (Required)

They are the same product
 They are different products

Reasons for Your Choice (Optional)

Product Pair #2

Product Name	Price
iPad 2nd generation 16GB WiFi White	\$469
iPhone 4th generation White 16GB	\$545

Your Choice (Required)

They are the same product
 They are different products

Reasons for Your Choice (Optional)

CrowdER: Batching Strategies

Cluster-based HIT

$O(n^2/k^2) \times$

Find Duplicate Products In the Table. ([Show Instructions](#))

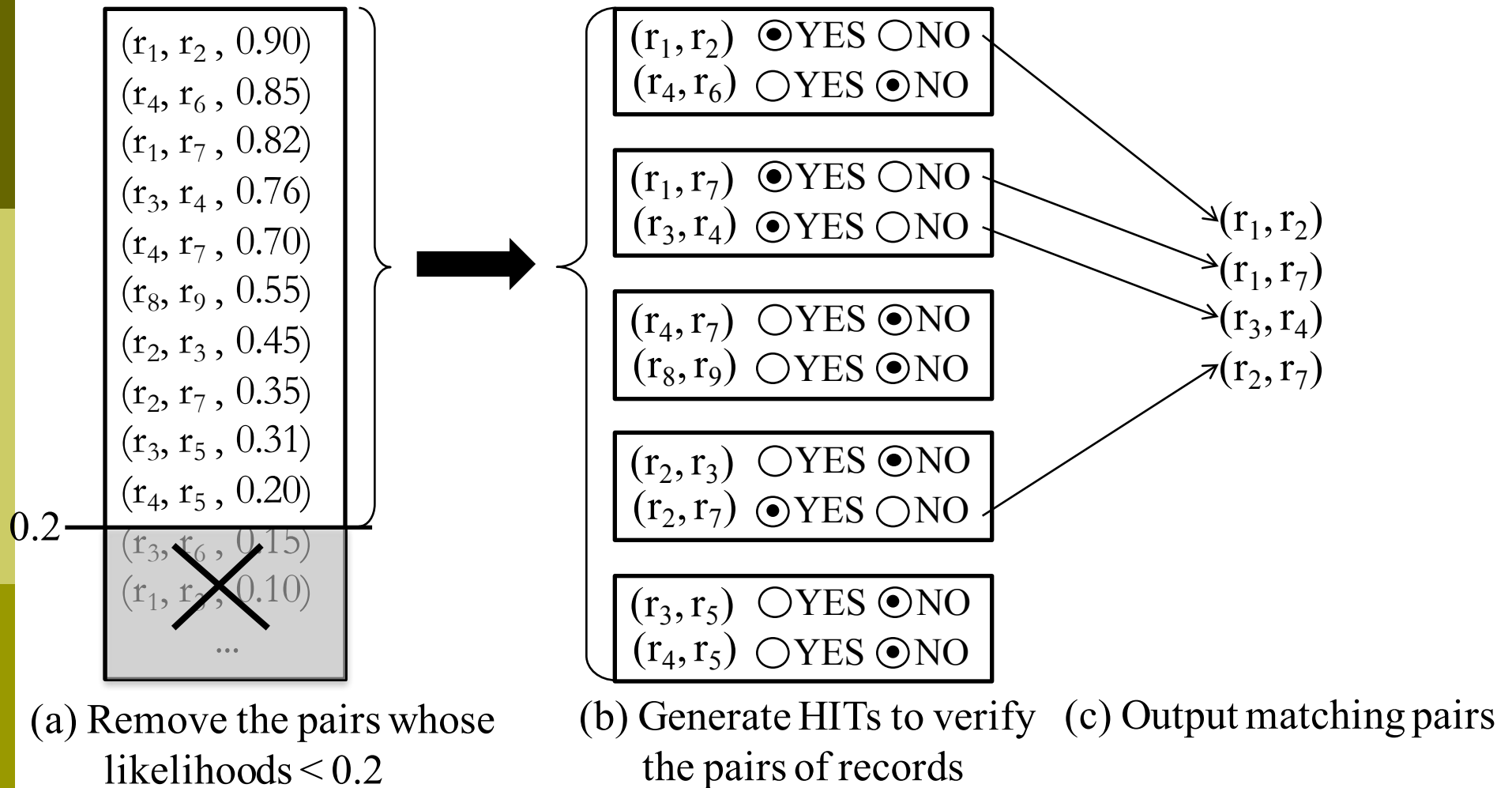
Tips: you can (1) **SORT** the table by clicking headers;
(2) **MOVE** a row by dragging and dropping it

Label	Product Name	Price ▲
1 ▼	iPad 2nd generation 16GB WiFi White	\$469
1 ▼	iPad Two 16GB WiFi White	\$490
2 ▼	Apple iPhone 4 16GB White	\$520
▼	iPhone 4th generation White 16GB	\$545

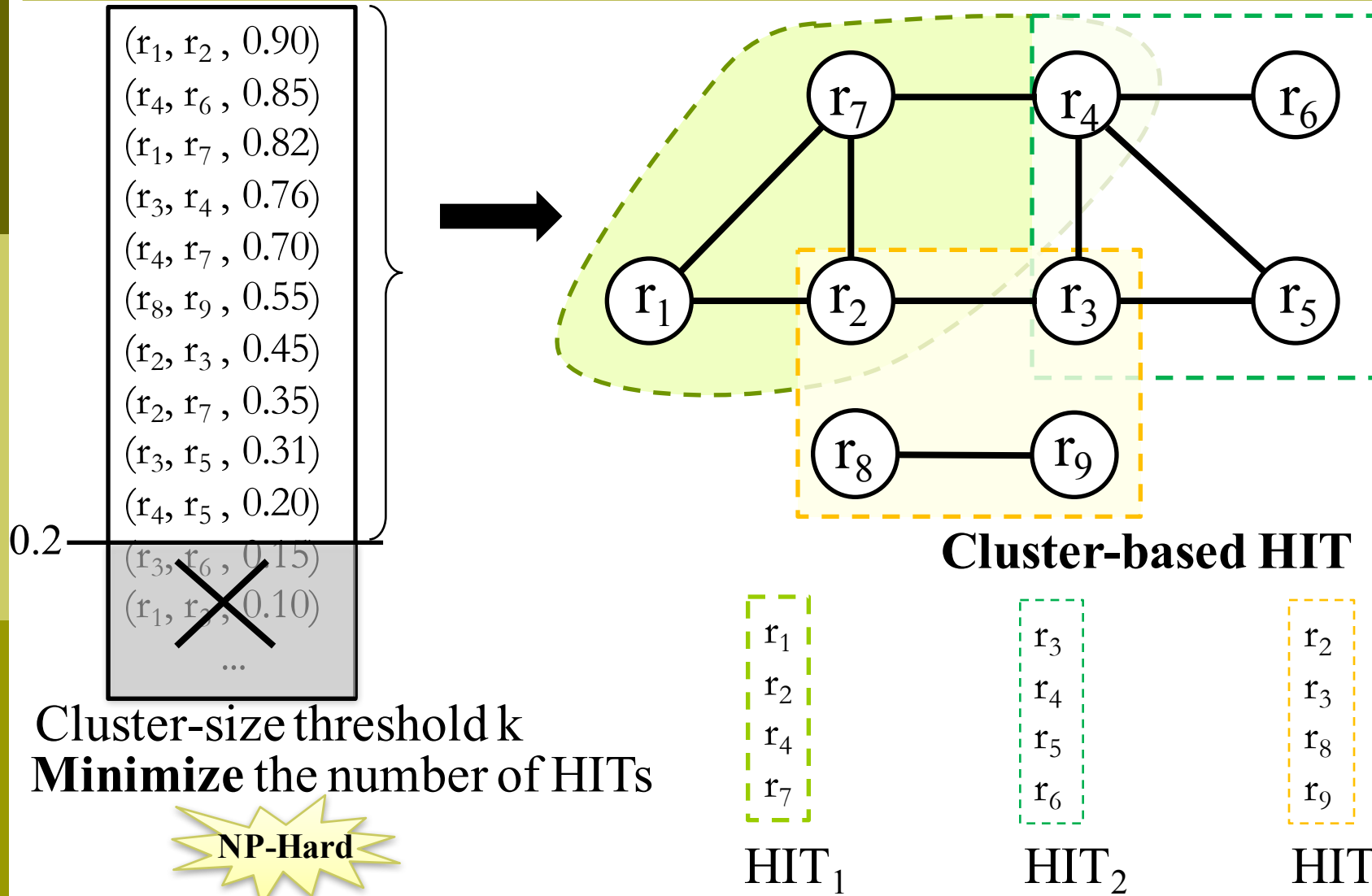
Reasons for Your Answers (Optional)

1
2
3
4

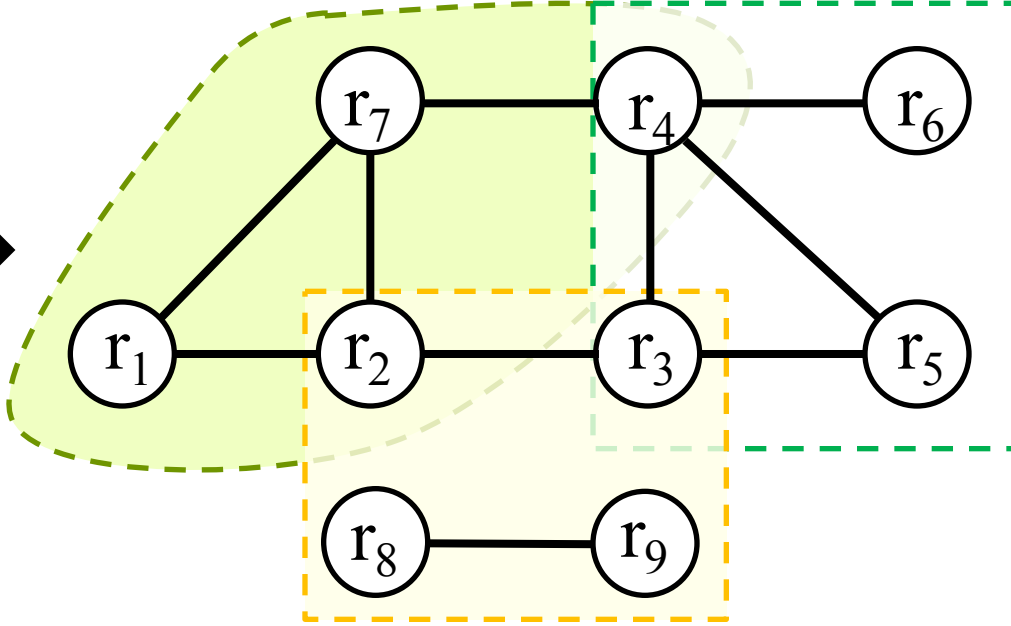
CrowdER: Workflow



CrowdER: Workflow



- ($r_1, r_2, 0.90$)
- ($r_4, r_6, 0.85$)
- ($r_1, r_7, 0.82$)
- ($r_3, r_4, 0.76$)
- ($r_4, r_7, 0.70$)
- ($r_8, r_9, 0.55$)
- ($r_2, r_3, 0.45$)
- ($r_2, r_7, 0.35$)
- ($r_3, r_5, 0.31$)
- ($r_4, r_5, 0.20$)
- ~~($r_3, r_6, 0.15$)~~
- ~~($r_1, r_7, 0.10$)~~
- ...

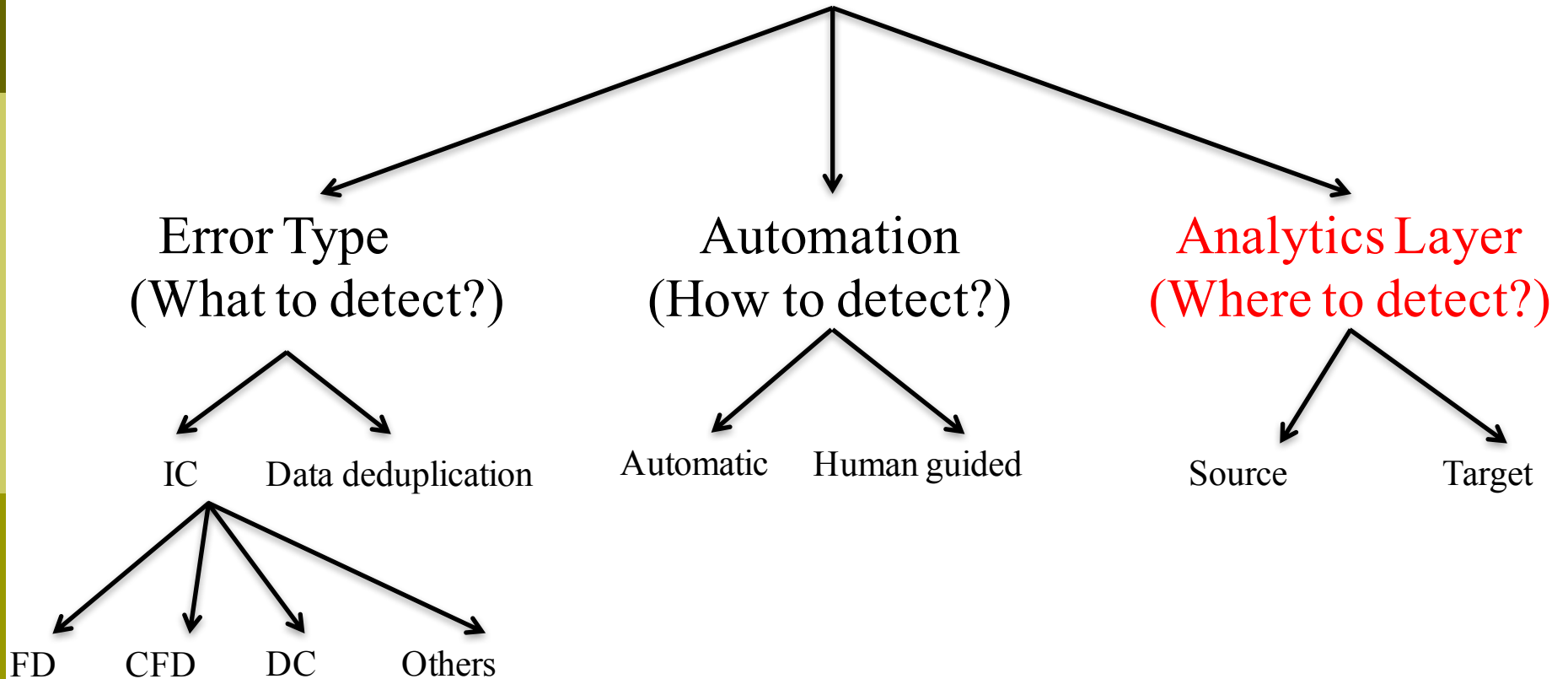


Cluster-size threshold k
Minimize the number of HITs

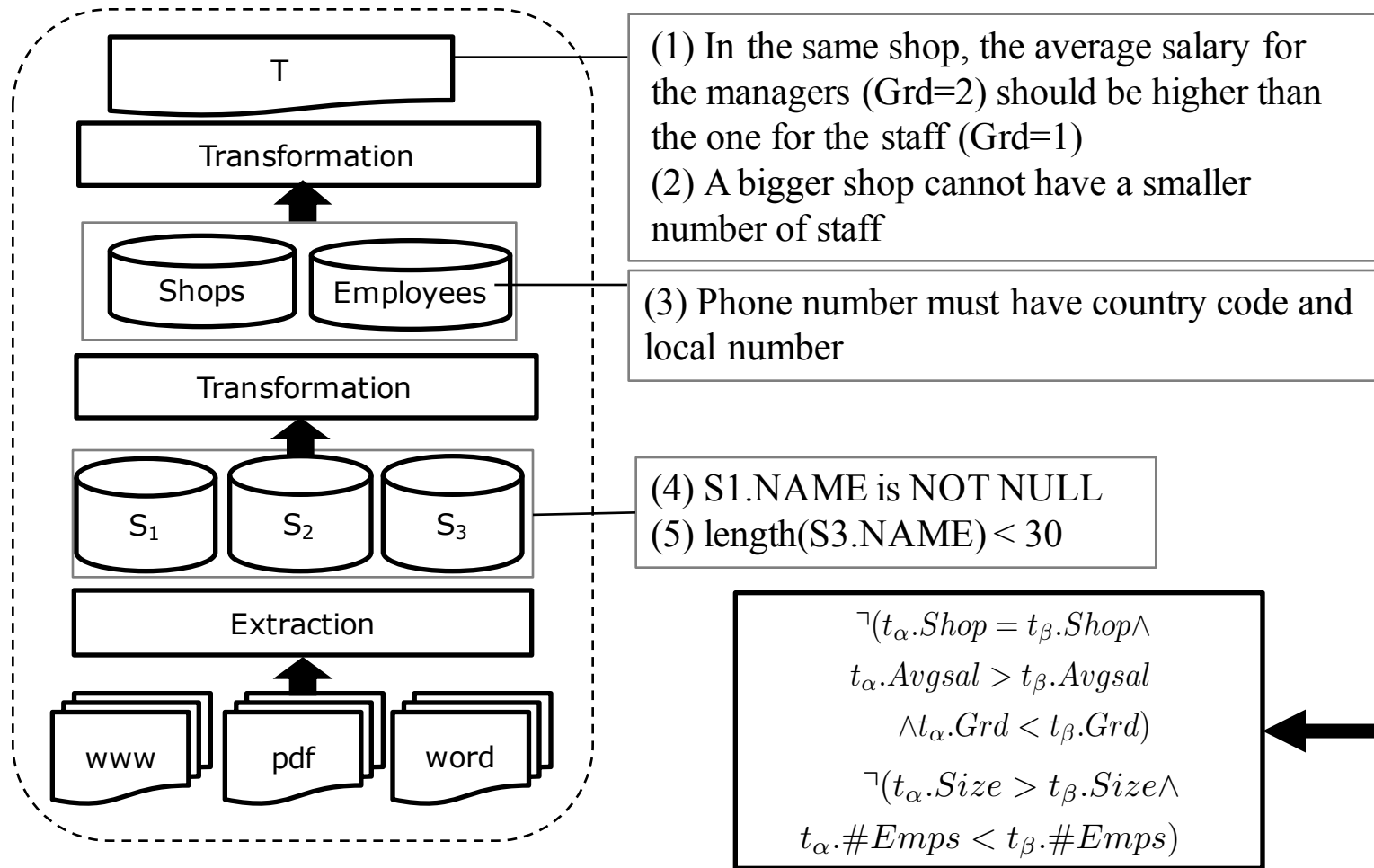
NP-Hard

Error Detection Techniques Taxonomy

Qualitative Error Detection Techniques



Decoupled in Space and Time

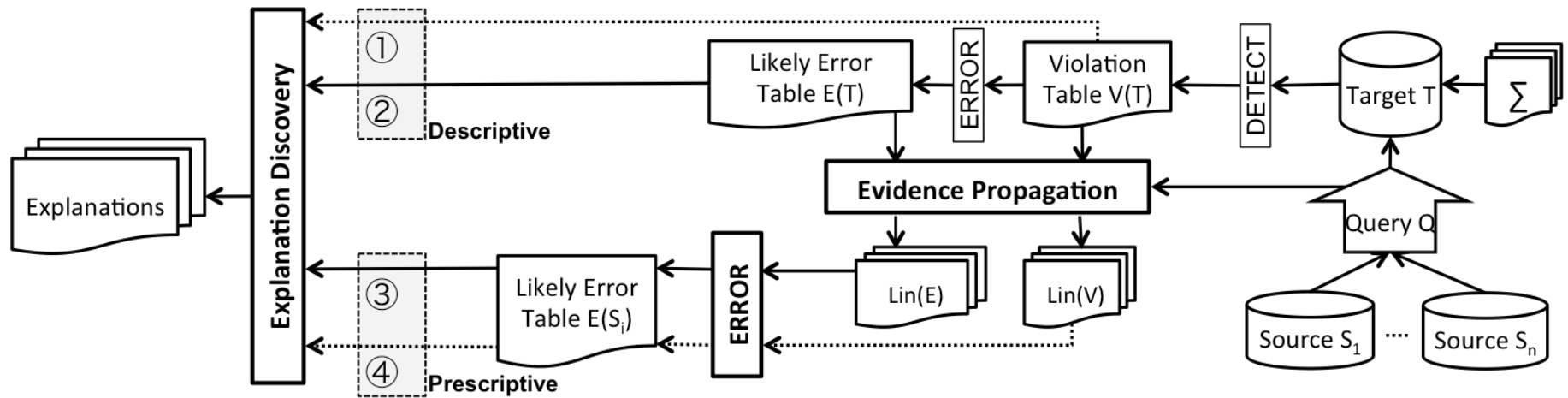


Calls for a New Solution

		Errors	
		Target	Source
Constraints	Target	Traditional Data Repair Algorithms	Descriptive and Prescriptive Data Cleaning
	Source	Dependency Propagation	Traditional Data Repair Algorithms

- DBRx: [Chalamalla et al., SIGMOD 2014]
- DataXRy: [Wang et al., SIGMOD 2015]
- QOCO: [Bergman et al., VLDB 2015]

DBR_x Architecture [Chalamalla et al, SIGMOD 2014]



Technical Challenges

❑ **Errors Propagation**

- Blowup (e.g., Aggregates)
- Propagation Level (violations vs Fixes)
- Distributing Responsibilities

❑ **Source Error Identification**

- Assign Weights based on Query and Error Semantics
- Accumulate Evidences (different Violation Semantics)

❑ **Explain Errors**

Tracing the Sources of Errors

T	Shop	Size	Grd	AvgSal	#Emps	Region
t _a	NY1	46 ft ²	2	99 \$	1	US
t _b	NY1	46 ft ²	1	100 \$	3	US
t _c	NY2	62 ft ²	2	96 \$	2	US
t _d	NY2	62 ft ²	1	90 \$	2	US
t _e	LA1	35 ft ²	2	105 \$	2	US
t _f	LND	38 ft ²	1	65 £	2	EU

```
SELECT Shops.SId as Shop, Size,
Emps.GrId, AVG(Emps.Sal) as
AvgSal, COUNT(EId) as #Emps, 'US'
as Region
FROM US.Emps JOIN US.Shops ON Sid
GROUP BY SId, Size, Grd
```

Emps	EId	Name	Dept	Sal	Grd	SId	JoinYr
t ₁	e4	John	S	91	1	NY1	2012
t ₂	e5	Anne	D	99	2	NY1	2012
t ₃	e7	Mark	S	93	1	NY1	2012
t ₄	e8	Claire	S	116	1	NY1	2012
t ₅	e11	Ian	R	89	1	NY2	2012
t ₆	e13	Laure	R	94	2	NY2	2012
t ₇	e14	Mary	E	91	1	NY2	2012
t ₈	e18	Bill	D	98	2	NY2	2012
t ₉	e14	Mike	R	94	2	LA1	2011
t ₁₀	e18	Claire	E	116	2	LA1	2011

Shops	SId	City	State	Size	Start
t ₁₁	NY1	NYC	NY	46 ft ²	2011
t ₁₂	NY2	NYC	NY	62 ft ²	2012
t ₁₃	LA1	LA	CA	35 ft ²	2011

2?

Average salary of higher grade in the same shop should be higher!

Error Contribution Scores

Emps	EId [CSV]	Sal [CSV]	Grd [CSV]	SId[CSV]	[RSV]
t_1	e4 [“, $\frac{1}{3}$]	91 [$\frac{91}{300}$, “]	1 [$\frac{1}{3}$, $\frac{1}{3}$]	NY1 [$\frac{1}{3}$, “]	[0,1]
t_2	e5	99 [0, “]	2 [1, “]	NY1 [1, “]	[1, “]
t_3	e7 [“, $\frac{1}{3}$]	93 [$\frac{93}{300}$, “]	1 [$\frac{1}{3}$, $\frac{1}{3}$]	NY1 [$\frac{1}{3}$, “]	[0,1]
t_4	e8 [“, $\frac{1}{3}$]	116 [$\frac{116}{300}$, “]	1 [$\frac{1}{3}$, $\frac{1}{3}$]	NY1 [$\frac{1}{3}$, “]	[1,1]
t_5	e11 [“, $\frac{1}{2}$]	89	1 [“, $\frac{1}{2}$]	NY2	[“, 0]
t_6	e13	94	2	NY2	[]
t_7	e14 [“, $\frac{1}{2}$]	91	1 [“, $\frac{1}{2}$]	NY2	[“, 0]
t_8	e18	98	2	NY2	[]
t_9	e14	94	2	LA1	[]
t_{10}	e18	116	2	LA1	[]

$cs_v(c)$:
Contribution
of this cell to
the aggregate

Shops	SId [CSV]	Size [CSV]	[RSV]
t_{12}	NY1 [2, “]	46 [“, 1]	[1,1]
t_{13}	NY2	62 [“, 1]	[“, 1]
t_{14}	LA1	35	[]

$rs_v(t)$:
Removing t_4
eliminates the
violations

Identifying Likely Errors

- Maximize a gain function of adding more source errors

$$Gain(H_v) = \sum_{s \in H_v} c_v(s) - \sum_{1 \leq j \leq |H_v|} \sum_{j < k \leq |H_v|} D(s_j, s_k)$$
$$D(s_j, s_k) = |c_v(s_j) - c_v(s_k)| \quad c_v(s) = cs_v(s) + rs_v(s)$$

tid	Score
s ₁	0.67
s ₂	0.54
s ₃	0.47
s ₄	0.08
s ₅	0.06
s ₆	0.05

Gain = 1.08

tid	Score
s ₁	0.67
s ₂	0.54
s ₃	0.47
s ₄	0.08
s ₅	0.06
s ₆	0.05

Gain = 1.28

tid	Score
s ₁	0.67
s ₂	0.54
s ₃	0.47
s ₄	0.08
s ₅	0.06
s ₆	0.05

Gain = -0.08

Error Explanation

Likely Error Tuples

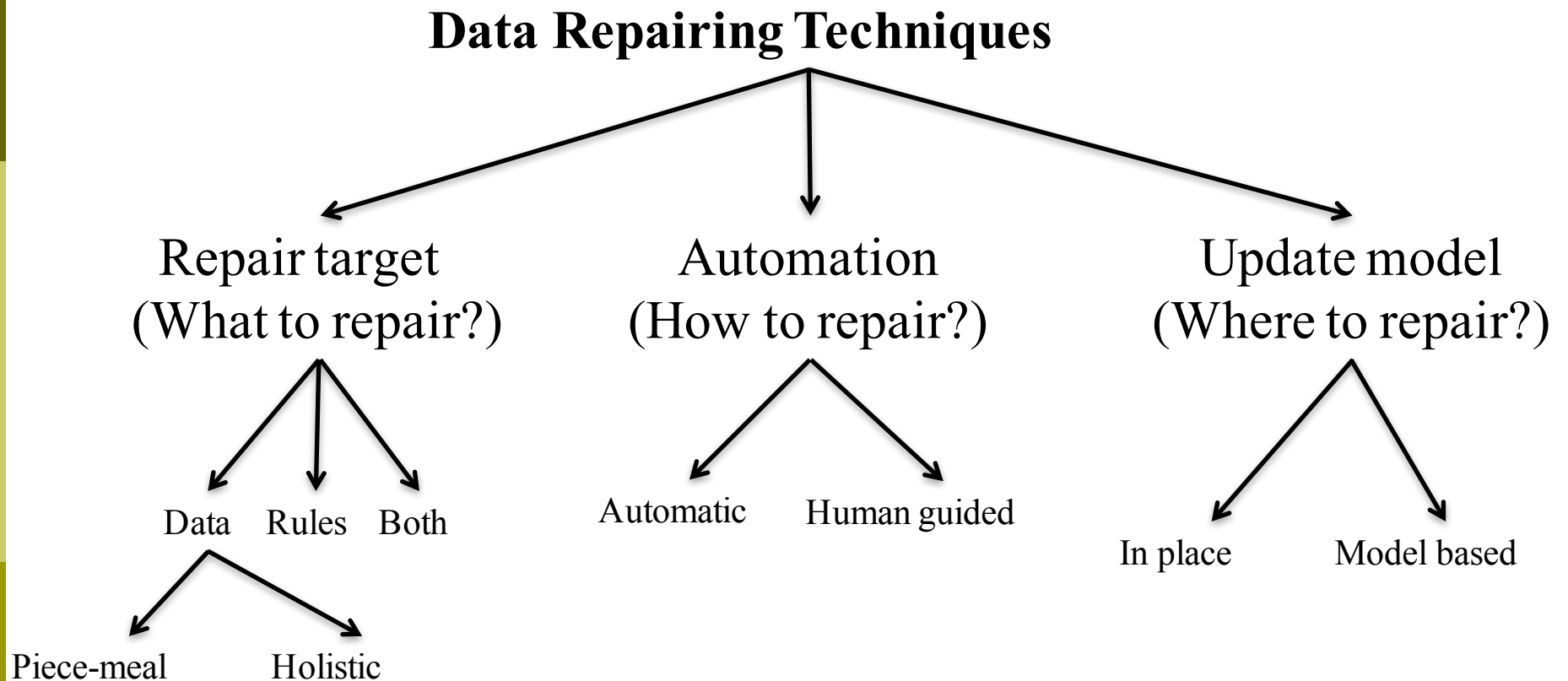
Emps	EId	Name	Dept	Sal	Grd	SId	JoinYr
t_1	e4	John	S	91	1	NY1	2012
t_2	e5	Anne	D	99	2	NY1	2012
t_3	e7	Mark	S	93	1	NY1	2012
t_4	e8	Claire	S	116	1	NY1	2012
t_5	e11	Ian	R	89	1	NY2	2012
t_6	e13	Laure	R	94	2	NY2	2012
t_7	e14	Mary	E	91	1	NY2	2012
t_8	e18	Bill	D	98	2	NY2	2012
t_9	e14	Mike	R	94	2	LA1	2011
t_{10}	e18	Claire	E	116	2	LA1	2011

Possible Explanations

Explanation	Recall	Precision	Concise
$Dept = s$	Low	High	Concise
$eid = e_4 \vee eid = e_7 \vee eid = e_8 \vee eid = e_{14}$	High	High	Verbose
$Grd = 1$	High	Low	Concise

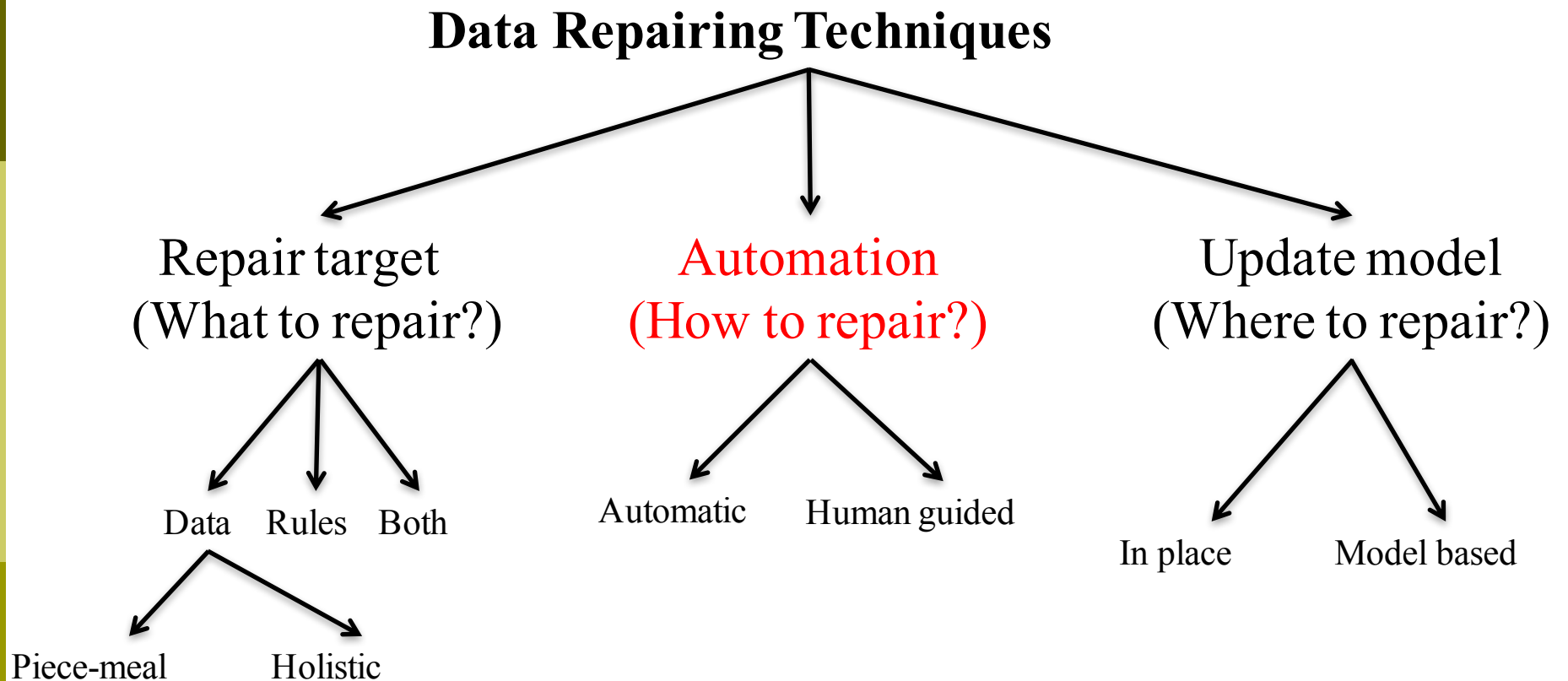
Data Repairing

Data Repairing Techniques Taxonomy



[Ilyas and Chu, Foundations and Trends in Database Systems, 2015]

Data Repairing Techniques Taxonomy

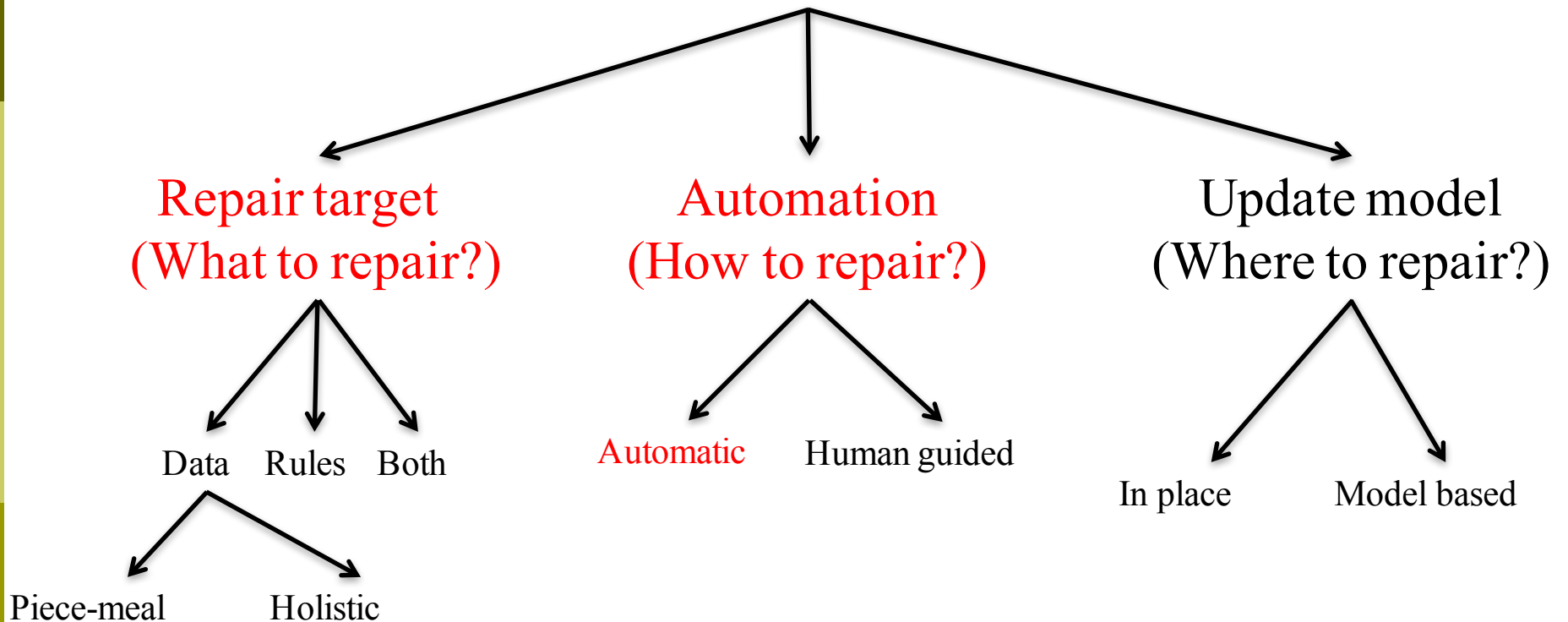


Repair Automation

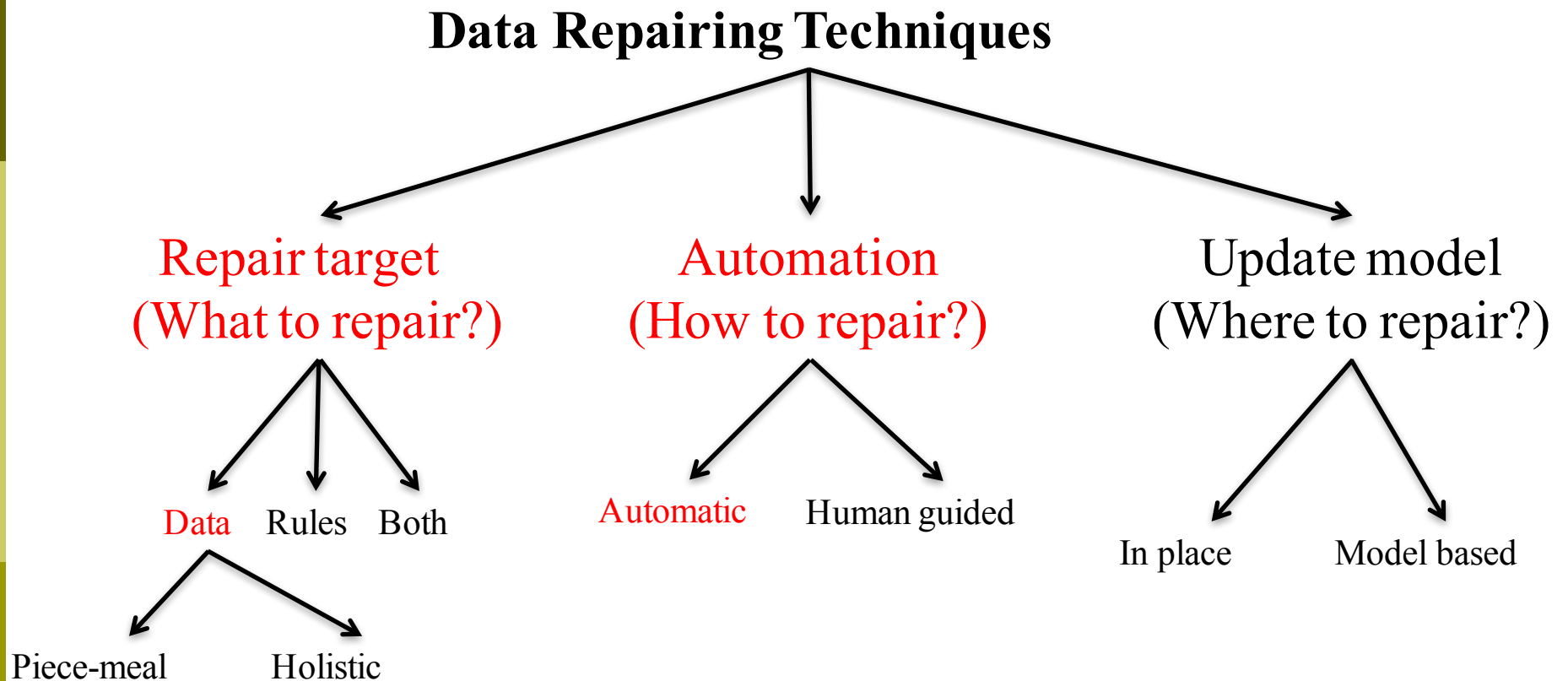
- Most automatic repairing techniques adopt the “minimality” of repairs principle
- Repairing techniques in practice are predominantly manual and semi-automatic at best
- Will survey both

Data Repairing Techniques Taxonomy

Data Repairing Techniques



Data Repairing Techniques Taxonomy



Data Repair by Value Update

- I is a **dirty** database if $I \not\models \Sigma$, and I_j is a **repair** for I if $I_j \models \Sigma$
- For a repair I_j , $\Delta(I_j)$ is the set of changed cells in I_j

I

	A	B
t_1	1	2
t_2	1	3
t_3	1	3
t_4	4	5



I_1

	A	B
t_1	1	3
t_2	1	3
t_3	1	3
t_4	4	5

I_2

	A	B
t_1	1	2
t_2	1	2
t_3	1	2
t_4	4	5

$$\Sigma = \{A \rightarrow B\}$$

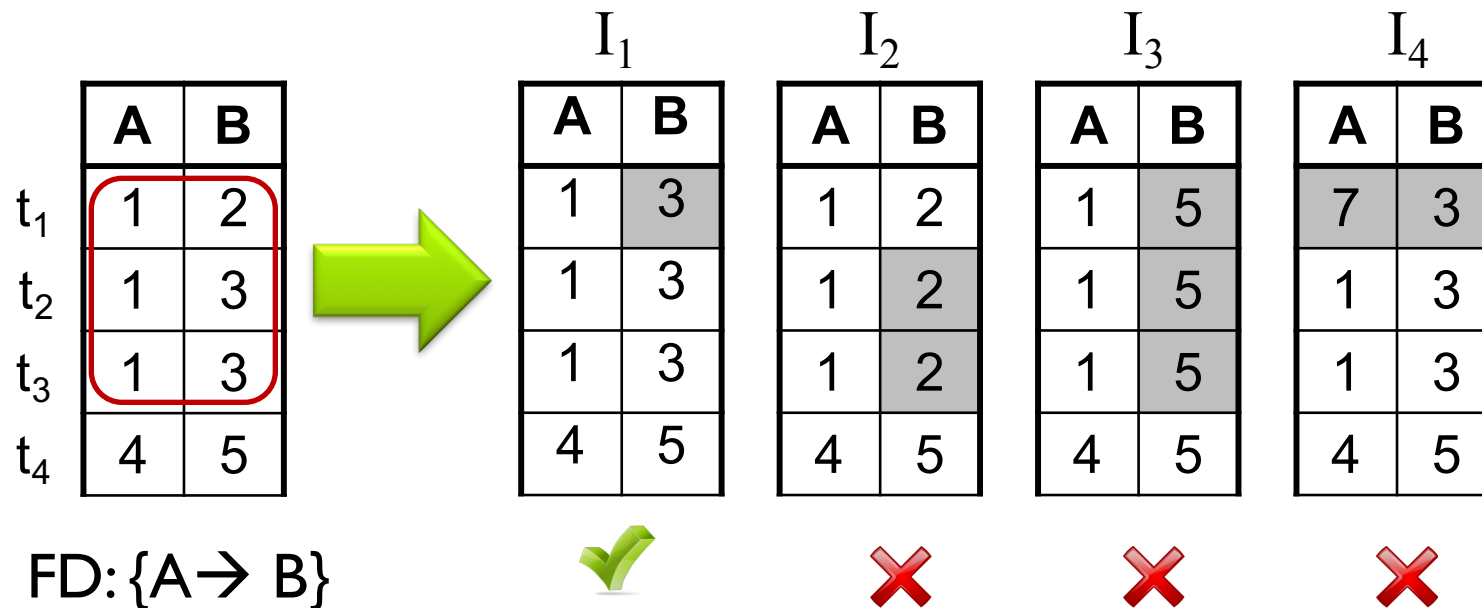
$$\Delta(I_1) = \{t_1[B]\}$$

$$\Delta(I_2) = \{t_2[B], t_3[B]\}$$

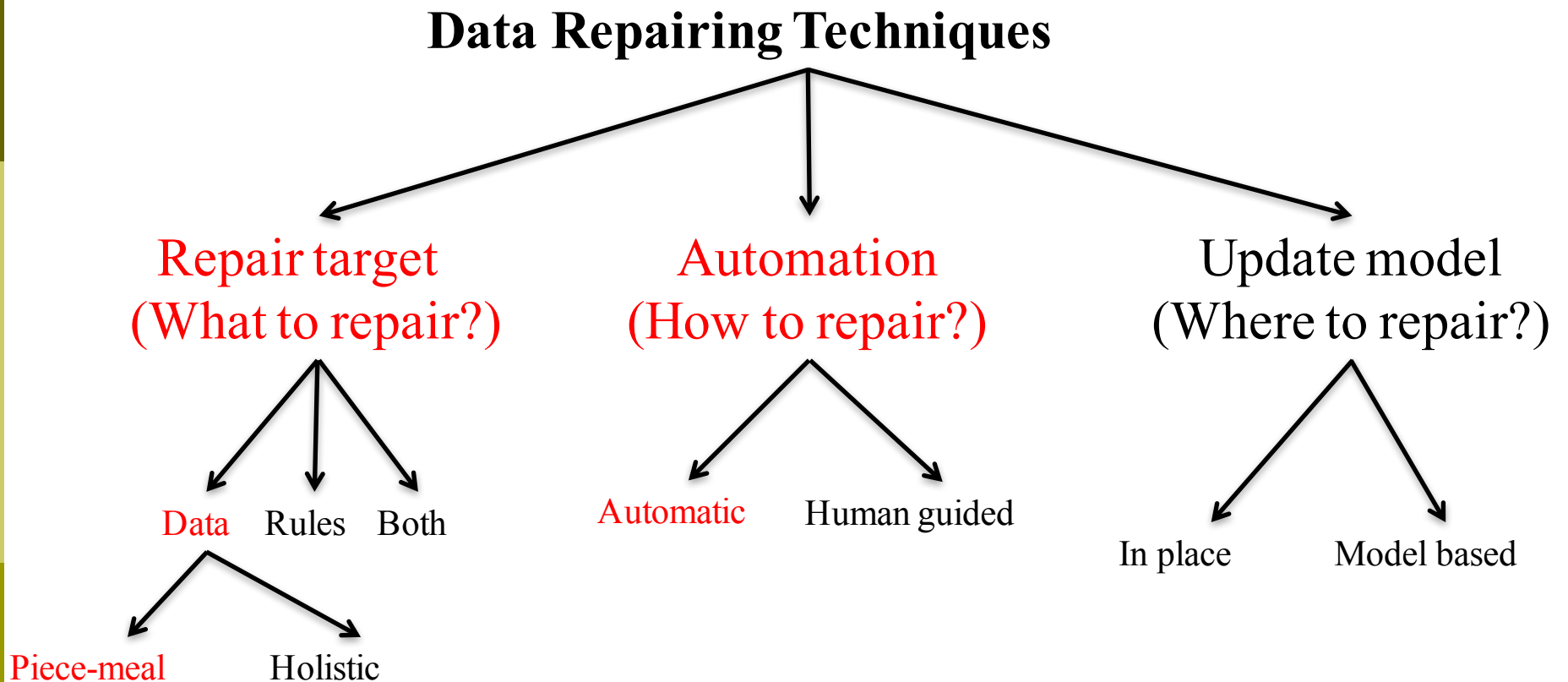
Data Only Repairing

□ Cardinality-Minimal repairs

- Commonly used in **obtaining a single repair automatically**
- Repairs with the **minimum number of changes**
- I_1 is Card-Min iff $\nexists I_2$ s.t. $|\Delta(I_2)| < |\Delta(I_1)|$



Data Repairing Techniques Taxonomy



FD Repairing [Bohannon et al, SIGMOD 2005]

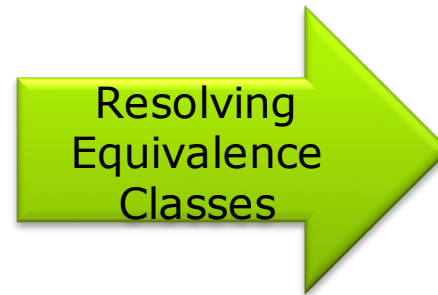
	A	B
t ₁	1	2
t ₂	1	3
t ₃	1	3
t ₄	4	5

FD: {A → B}



	A	B
t ₁	1	2
t ₂	1	3
t ₃	1	3
t ₄	4	5

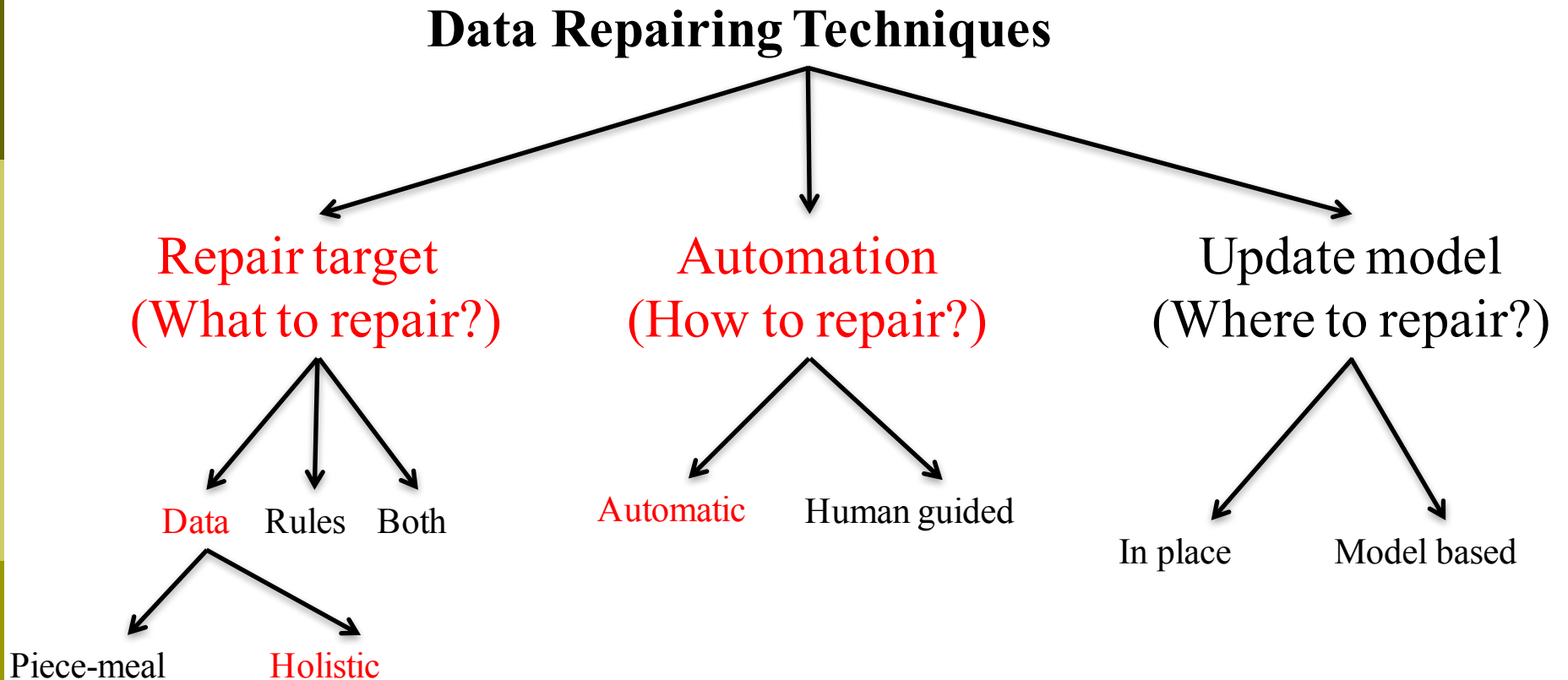
FD: {A → B}



	A	B
t ₁	1	3
t ₂	1	3
t ₃	1	3
t ₄	4	5

FD: {A → B}

Data Repairing Techniques Taxonomy

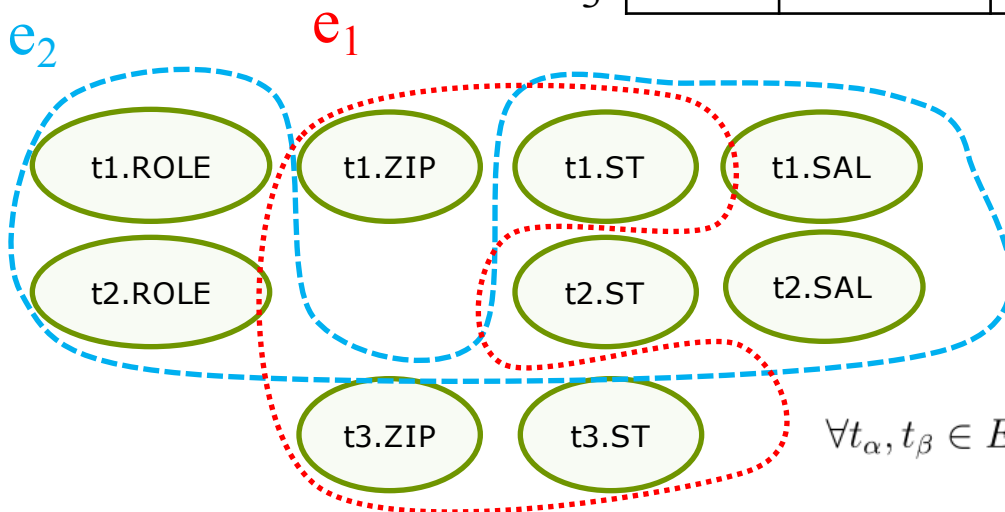


Holistic Data Repairing [Chu et al, ICDE 2013]

- Vertex: Cell in the database
- Hyperedge: A set of cells that violate a DC

	ID	FN	LN	ROLE	ZIP	ST	SAL
t_1	105	Anne	Nash	E	85376	NY	110
t_2	211	Mark	White	M	90012	NY	80
t_3	386	Mark	Lee	E	85376	AZ	75

Employee Table



Zip \rightarrow ST

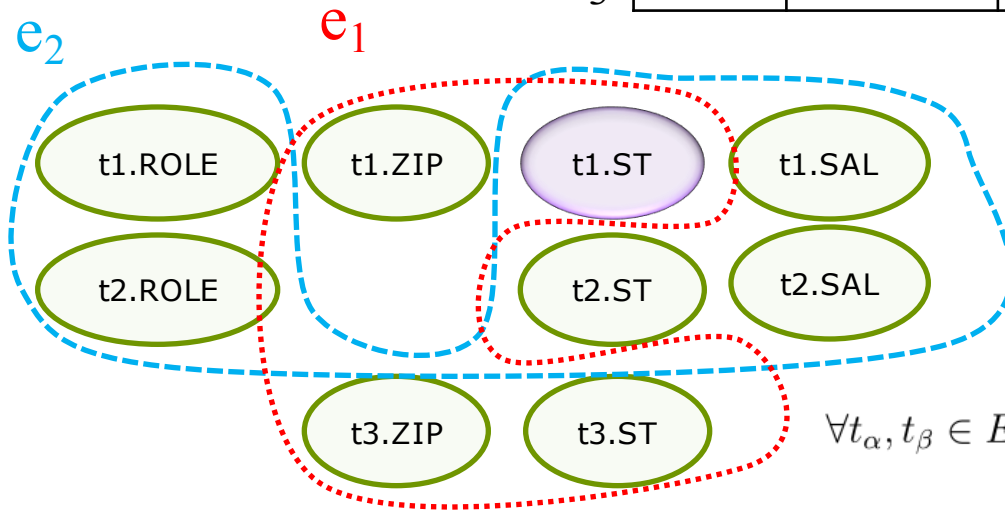
$$\forall t_\alpha, t_\beta \in Emp, \neg(t_\alpha.ST = t_\beta.ST \wedge t_\alpha.ROLE = "M" \wedge t_\beta.ROLE = "E" \wedge t_\alpha.SAL < t_\beta.SAL)$$

Step1: Minimal Vertex Cover

- A minimal set of vertices that are intersecting with every hyperedge

	ID	FN	LN	ROLE	ZIP	ST	SAL
t_1	105	Anne	Nash	E	85376	NY	110
t_2	211	Mark	White	M	90012	NY	80
t_3	386	Mark	Lee	E	85376	AZ	75

Employee Table



Zip \rightarrow ST

$$\forall t_\alpha, t_\beta \in Emp, \neg(t_\alpha.ST = t_\beta.ST \wedge t_\alpha.ROLE = "M" \wedge t_\beta.ROLE = "E" \wedge t_\alpha.SAL < t_\beta.SAL)$$

Step2: Collect Repair Requirements

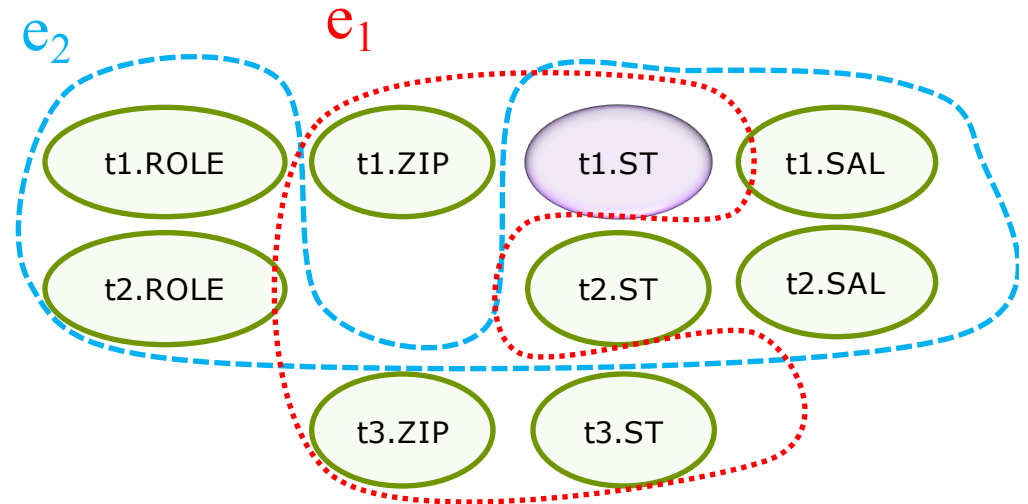
- A set of conditions that need to be satisfied to resolve all violations

Condition to resolve e_1 by changing $t1.ST$:

$$t1.ST = t3.ST$$

Condition to resolve e_2 by changing $t1.ST$:

$$t1.ST \neq t2.ST$$



Zip \rightarrow ST

$$\forall t_\alpha, t_\beta \in Emp, \neg(t_\alpha.ST = t_\beta.ST \wedge t_\alpha.ROLE = "M" \wedge t_\beta.ROLE = "E" \wedge t_\alpha.SAL < t_\beta.SAL)$$

Step3: Get Updates

- A set of assignments satisfying the conditions, with minimal number of changed cells

$t_1.ST = t_3.ST$
 $t_1.ST \neq t_2.ST$

	ID	FN	LN	ROLE	ZIP	ST	SAL
t_1	105	Anne	Nash	E	85376	NY	110
t_2	211	Mark	White	M	90012	NY	80
t_3	386	Mark	Lee	E	85376	AZ	75

Gradually increase the number of cells that are going to be changed, until reach a solution

Suppose we only want to change $t_1.ST$
 $t_2.ST = NY$
 $t_3.ST = AZ$

AZ

More Holistic Data Repairing [Fan et al, SIGMOD 2011]

	FN	LN	St	City	AC	Post	Phn	Item
Tran	Robert	Brady	5 Wren St	Ldn	020	WC1H 9SE	3887644	watch
	Robert	Brady	5 Wren St	Ldn	020	WC1E 7HX	3887644	necklace

Master: Card	FN	LN	St	City	AC	Zip	Tel
	Robert	Brady	5 Wren St	Ldn	020	WC1H 9SE	3887644

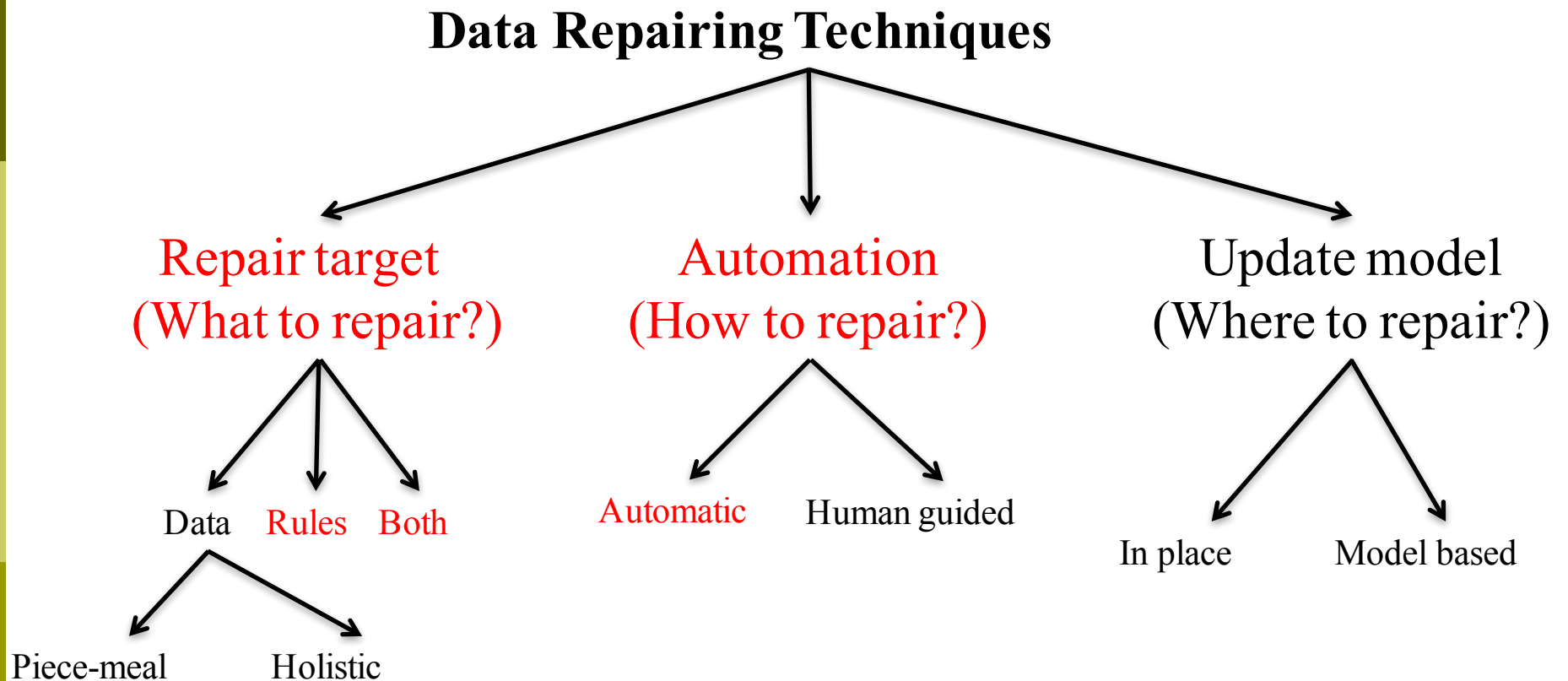
CFD: Tran(AC = 020 \rightarrow City = Lnd)

CFD: Tran(FN = Bob \rightarrow FN = Robert)

MD: Tran[LN, City, St, Post] = card[LN, City, St, Zip] ^
 Tran[FN] \approx Card[FN] \rightarrow Tran[FN, Phn] \Leftrightarrow Card[FN, Tel]

FD: Tran(City, Phn \rightarrow St, AC, Post)

Data Repairing Techniques Taxonomy



Data & Rules Repairing: Motivating Example

□ Car Database

- **Model** → **Make** was satisfied by Car databases till Mazda 323 was introduced (Conflicting with BMW 323)
- Could be corrected to **Model, Cylinders** → **Make**

□ US presidents Database

- **LastName, FirstName** → **StartYear, EndYear** was satisfied till the election of George W. Bush
- Should be corrected to **LastName, MiddleInit, FirstName** → **StartYear, EndYear**

[Chiang and Miller, ICDE 2011]

[Beskales et al, ICDE 2013]

Relative Trust

- ❑ In reality, both **data** and **constraints (FDs)** can be wrong
- ❑ The **relative trust** in data vs. FDs determines how we should repair data and FDs

Example

	GivenName	Surname	BirthDate	Gender	Phone	Income
t ₁	Danielle	Blake	9 Dec 1970	Female	817-213-1211	120k
t ₂	Danielle	Blake	9 Dec 1970	Female	817-988-9211	100k
t ₃	Hong	Li	27 Oct 1972	Female	591-977-1244	90k
t ₄	Hong	Li	8 Mar 1979	Female	498-214-5822	84k
t ₅	Ning	Wu	3 Nov 1982	Male	313-134-9241	90k
t ₆	Ning	Wu	8 Nov 1982	Male	323-456-3452	95k

Surname, GivenName → Income

Example (Trusted FD)

	GivenName	Surname	BirthDate	Gender	Phone	Income
t ₁	Danielle	Blake	9 Dec 1970	Female	817-213-1211	120k
t ₂	Danielle	Blake	9 Dec 1970	Female	817-988-9211	120k
t ₃	Hong	Li	27 Oct 1972	Female	591-977-1244	90k
t ₄	Hong	Li	8 Mar 1979	Female	498-214-5822	90k
t ₅	Ning	Wu	3 Nov 1982	Male	313-134-9241	95k
t ₆	Ning	Wu	8 Nov 1982	Male	323-456-3452	95k

Surname, GivenName → Income

Example (Trusted Data)

	GivenName	Surname	BirthDate	Gender	Phone	Income
t ₁	Danielle	Blake	9 Dec 1970	Female	817-213-1211	120k
t ₂	Danielle	Blake	9 Dec 1970	Female	817-988-9211	100k
t ₃	Hong	Li	27 Oct 1972	Female	591-977-1244	90k
t ₄	Hong	Li	8 Mar 1979	Female	498-214-5822	84k
t ₅	Ning	Wu	3 Nov 1982	Male	313-134-9241	90k
t ₆	Ning	Wu	8 Nov 1982	Male	323-456-3452	95k

Surname, GivenName, BirthDate, Phone → Income

Example (Equally-trusted Data and FD)

	GivenName	Surname	BirthDate	Gender	Phone	Income
t ₁	Danielle	Blake	9 Dec 1970	Female	817-213-1211	120k
t ₂	Danielle	Blake	9 Dec 1970	Female	817-988-9211	120k
t ₃	Hong	Li	27 Oct 1972	Female	591-977-1244	90k
t ₄	Hong	Li	8 Mar 1979	Female	498-214-5822	84k
t ₅	Ning	Wu	3 Nov 1982	Male	313-134-9241	90k
t ₆	Ning	Wu	8 Nov 1982	Male	323-456-3452	95k

Surname, GivenName, BirthDate → Income

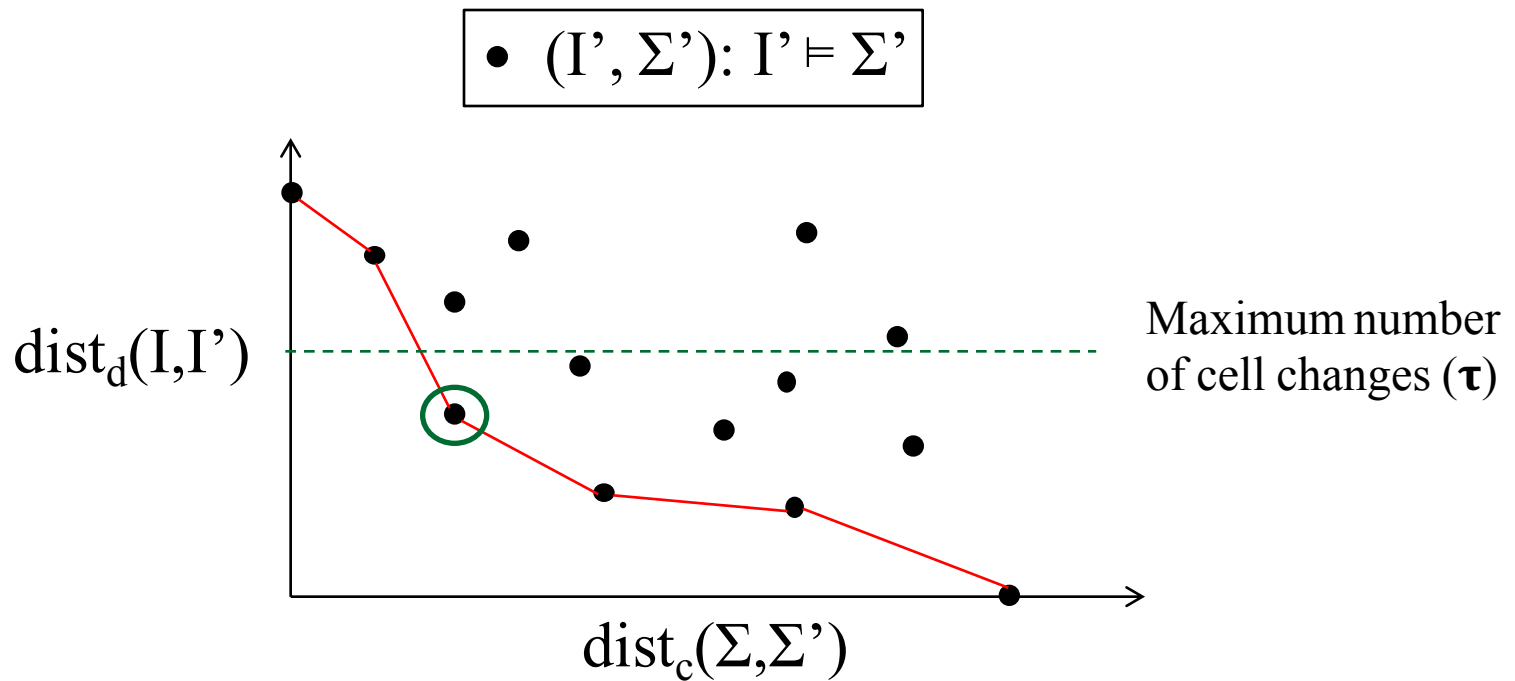
Data Repair

- We repair instance I by modifying multiple cells and produce I'
- $\text{dist}_d(I, I')$ is the number of different cells between I and I'

Repairing a set of FDs

- We repair an FD $X \rightarrow A$ by adding one or more attributes to the LHS
- Let $w(Y)$ be a weight reflecting the penalty of adding attribute set Y to X
 - E.g., the number of attributes in Y , distinct values of Y in I , entropy of Y
- Let $\text{dist}_c(\Sigma, \Sigma')$ be the sum of $w(Y)$ across all changed FDs

Relative Trust [Beskales et al, ICDE 2013]



A Unified Cost Model [Chiang and Miller, ICDE 2011]

- Minimum description Length Principle
 - Find a model M w.r.t. Σ that can represent the data as much as possible

- $DL(M) = L(M) + L(I|M)$
 - $L(M)$: Length of the model
 - $L(I|M)$: Length of data given M

A Unified Cost Model: Data Repair

□ M: empty

- $L(M) = 0$
- $L(I|M) = 27$
- $DL = 27$

□ M:

Brook	Granville	412
-------	-----------	-----

- $L(M) = 3 + 2 * 6 = 15$
- $L(I|M) = 0$
- $DL = 15$

FD: {District, Region → AC}

District	Region	AC
Brook	Granville	412
Brook	Granville	412
Brook	Granville	412
Brook	Granville	553 412
Brook	Granville	553 412
Brook	Granville	553 412
Brook	Granville	725 412
Brook	Granville	725 412
Brook	Granville	725 412

A Unified Cost Model: FD Repair

□ M: empty

- $L(M) = 0$
- $L(I|M) = 36$
- $DL = 36$

□ M:

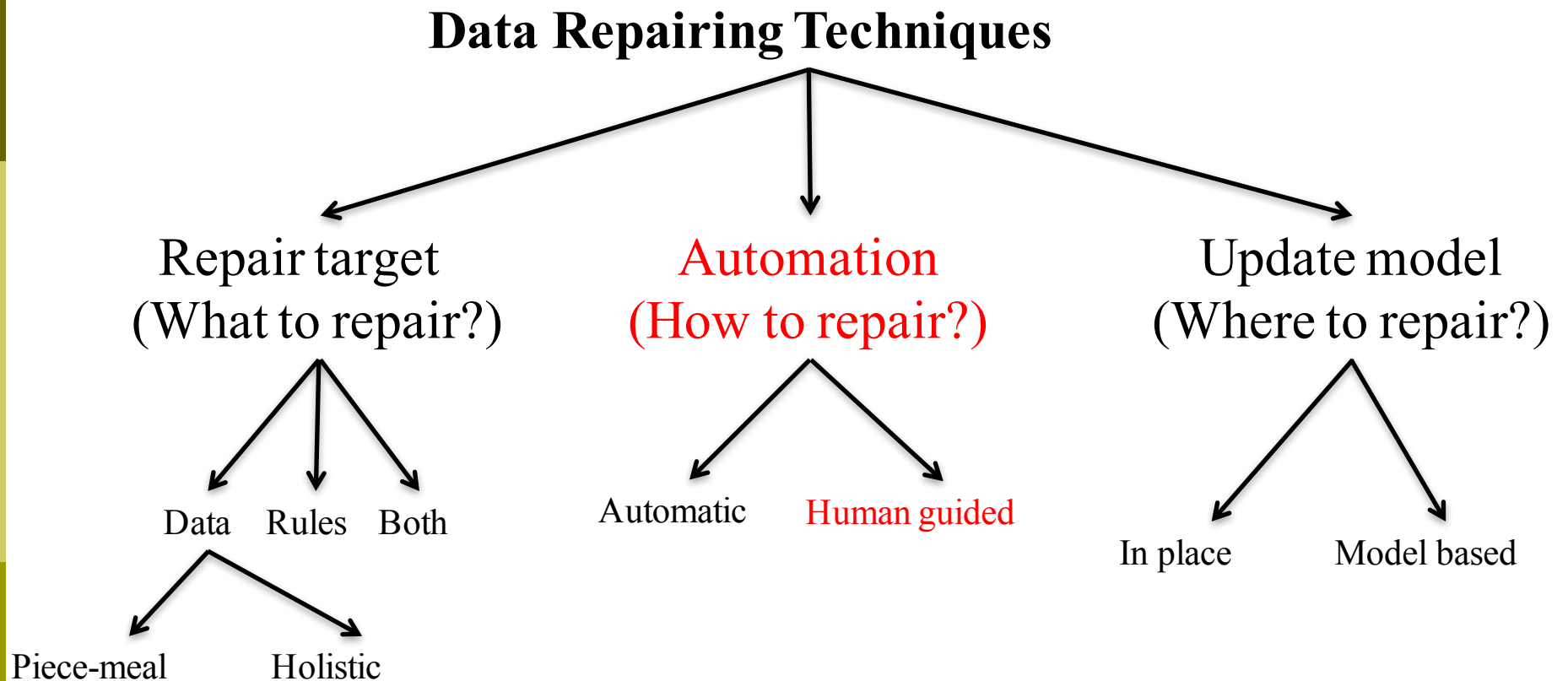
Glendale	Brook	Granville	412
Guildwood	Brook	Granville	553
Moore	Brook	Granville	725

- $L(M) = 12$
- $L(I|M) = 0$
- $DL = 12$

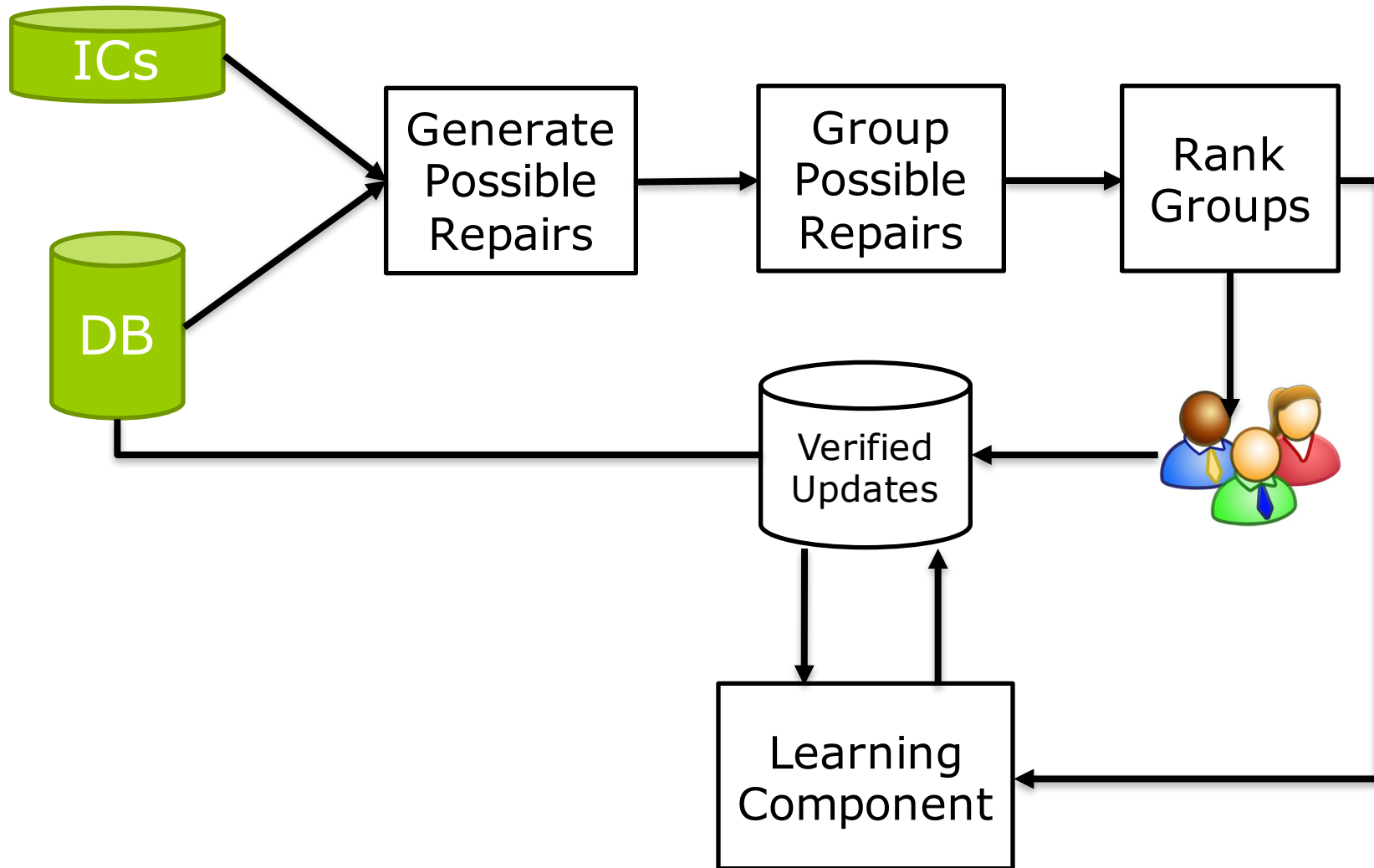
FD: {Municipal, District, Region → AC}

Municipal	District	Region	AC
Glendale	Brook	Granville	412
Glendale	Brook	Granville	412
Glendale	Brook	Granville	412
Guildwood	Brook	Granville	553
Guildwood	Brook	Granville	553
Guildwood	Brook	Granville	553
Moore	Brook	Granville	725
Moore	Brook	Granville	725
Moore	Brook	Granville	725

Data Repairing Techniques Taxonomy



Guided Data Repair (GDR) [Yakout et al, VLDB 2011]



GDR: Generate Possible Repairs

	Name	SRC	STR	CT	STT	ZIP
t1:	Jim	H1	REDWOOD DR	MICHIGAN CITY	MI	46360
t2:	Tom	H1	REDWOOD DR	WESTVILLE	IN	46360
t3:	Jeff	H2	BIRCH PARKWAY	WESTVILLE	IN	46360
t4:	Rick	H2	BIRCH PARKWAY	WESTVILLE	IN	46360
t5:	Mrk	H1	BELL AVENUE	FORT WAYNE	IN	46391
t6:	Mark	H1	BELL AVENUE	FORT WAYNE	IN	46825
t7:	Cady	H2	BELL AVENUE	FORT WAYNE	IN	46825
t8:	Sindy	H2	SHERDEN RD	FT WAYNE	IN	46774

$CFD_1: (ZIP \rightarrow CT, STT, \{46391 \parallel \text{Westville, IN}\})$

$CFD_2: (STR, CT \rightarrow ZIP, \{ - , \text{FortWayne} \parallel - \})$

- Suggested Update: replace City "FORT WAYNE" with "Westville" in t5
- Suggested Upadte: replace Zip "46391" with "46825" in t5

GDR: Group and Rank Repairs

	Name	SRC	STR	CT	STT	ZIP
t1:	Jim	H1	REDWOOD DR	MICHIGAN CITY	MI	46360
t2:	Tom	H1	REDWOOD DR	WESTVILLE	IN	46360
t3:	Jeff	H2	BIRCH PARKWAY	WESTVILLE	IN	46360
t4:	Rick	H2	BIRCH PARKWAY	WESTVILLE	IN	46360
t5:	Mrk	H1	BELL AVENUE	FORT WAYNE	IN	46391
t6:	Mark	H1	BELL AVENUE	FORT WAYNE	IN	46825
t7:	Cady	H2	BELL AVENUE	FORT WAYNE	IN	46825
t8:	Sindy	H2	SHERDEN RD	FT WAYNE	IN	46774

Contextual grouping for the suggested updates

Update Group g_1 : The city should be “Michigan City” for $\{t_2, t_3, t_4\}$.
Update Group g_2 : The zip should be “46825” for $\{t_5, t_8\}$.

....
....
....

KATARA [Chu et al, SIGMOD 2015]

	A	B	C	D	E	F	G
t_1	Rossi	Italy	Rome	Verona	Italian	Proto	1.78
t_2	Klate	South Africa	Pretoria	Pirates	Afrikaans	P. Eliz.	1.69
t_3	Pirlo	Italy	Madrid	Juve	Italian	Flero	1.77

A Table of Soccer Players

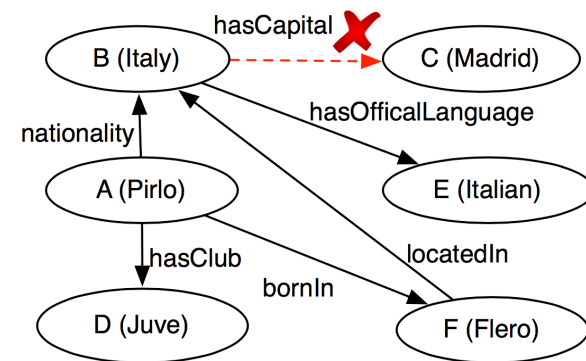
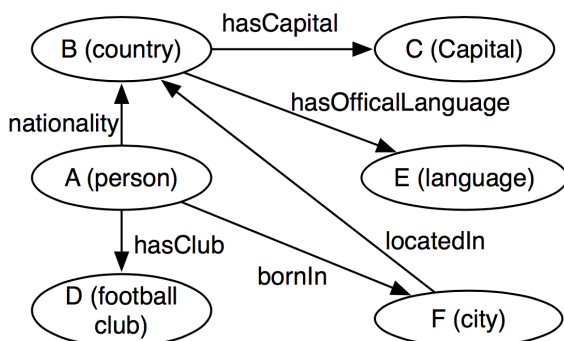
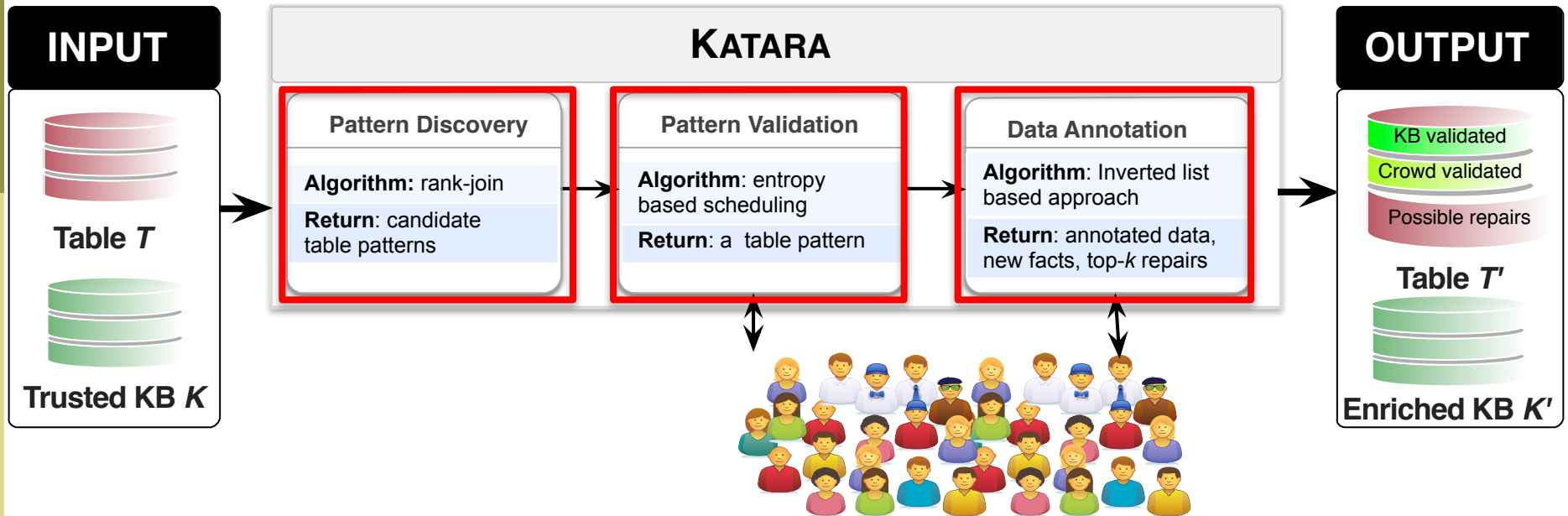
FD: $B \rightarrow C$

- Automatic: Produce heuristic repairs
- GDR:
 - Rely on redundancy to detect errors
 - Require heavy human involvement

Proposal: Use external trustworthy information!

- KBs
- Crowd experts

KATARA Workflow



Pattern Discovery: Generate Candidates

Generate candidate types for every column:

```
 $Q_{\text{types}}$  select ? $c_i$   
where {? $x_i$  rdfs:label  $t[A_i]$ ,  
? $x_i$  rdfs:type/rdfs:subClassOf* ? $c_i$ }
```

type(B)

economy
country
location
state
...

type(C)

City
Capital
whole
artifact
Person
...

Generate candidate relationships for every column pair:

```
 $Q_{\text{rels}}^1$  select ? $P_{ij}$   
where {? $x_i$  rdfs:label  $t[A_i]$ , ? $x_j$  rdfs:label  $t[A_j]$ ,  
? $x_i$  ? $P_{ij}$ /rdfs:subPropertyOf* ? $x_j$ }
```

```
 $Q_{\text{rels}}^2$  select ? $P_{ij}$   
where {? $x_i$  rdfs:label  $t[A_i]$ ,  
? $x_i$  ? $P_{ij}$ /rdfs:subPropertyOf*  $t[A_j]$ }
```

relationship (B, C)

locatedIn
hasCapital

Crowd Pattern Validation

Q_1 : What is the most accurate type of the highlighted column?

(A, **B**, C, D, E, F, ...)

(Rossi, **Italy**, Rome, Verona, Italian, Proto, ...)

(Pirlo, **Italy**, Madrid, Juve, Italian, Flero,, ...)

- country
- economy
- state

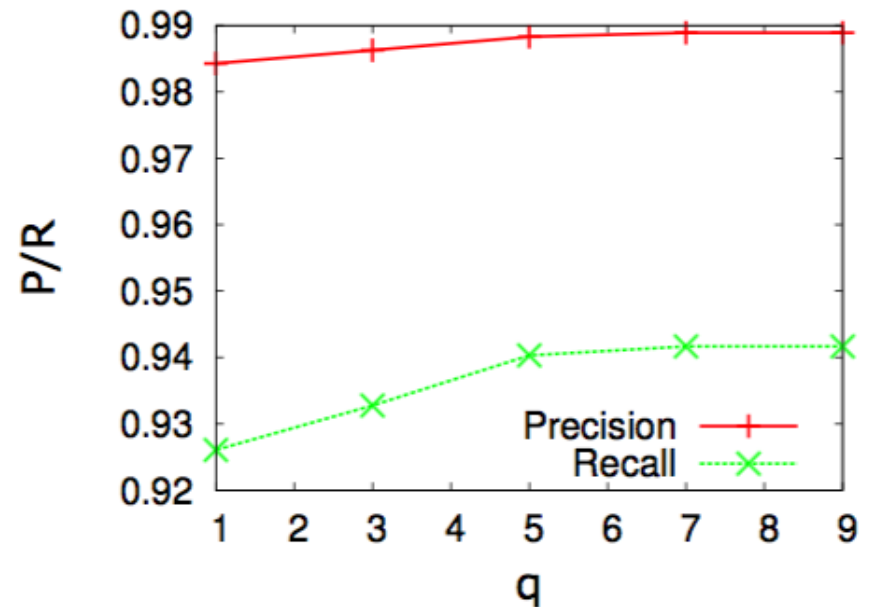
Q_2 : What is the most accurate relationship j

(A, **B**, **C**, D, E, F, ...)

(Rossi, **Italy**, **Rome**, Verona, Italian, Proto, ...)

(Pirlo, **Italy**, **Madrid**, Juve, Italian, Flero, ...)

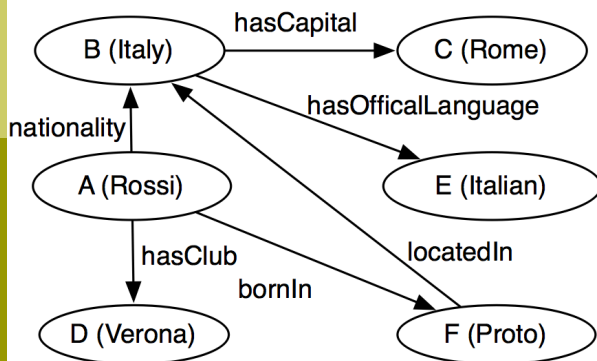
- B** hasCapital **C**
- C** locatedIn **B**



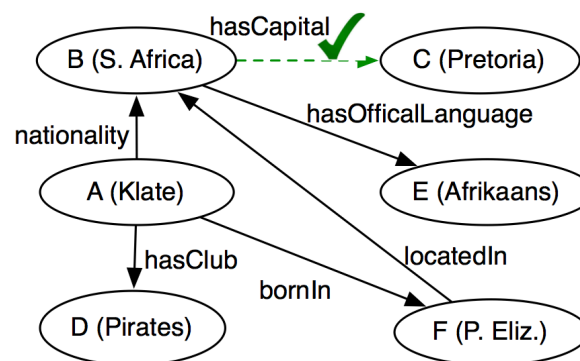
Data Annotation

	A	B	C	D	E	F	G
t_1	Rossi	Italy	Rome	Verona	Italian	Proto	1.78
t_2	Klate	South Africa	Pretoria	Pirates	Afrikaans	P. Eliz.	1.69
t_3	Pirlo	Italy	Madrid	Juve	Italian	Flero	1.77

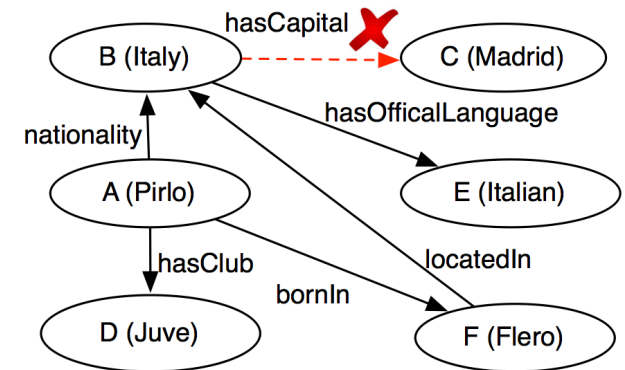
t_1 : validated by KB



t_2 : validated by KB & crowd



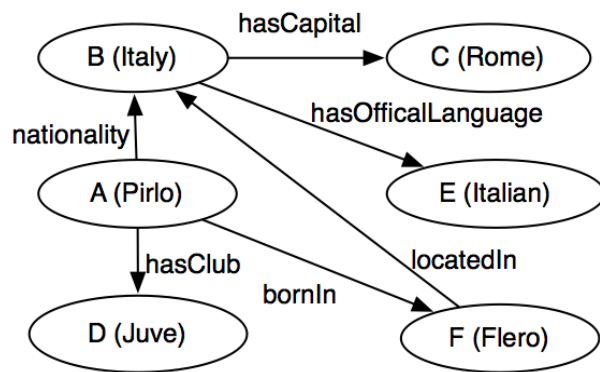
t_3 : Erroneous tuple



Data Repairing

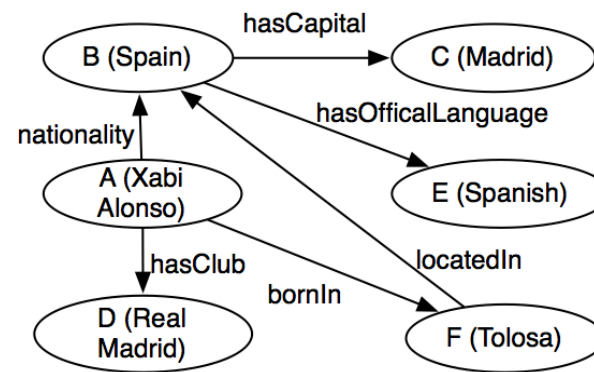
	A	B	C	D	E	F	G
t_1	Rossi	Italy	Rome	Verona	Italian	Proto	1.78
t_2	Klate	South Africa	Pretoria	Pirates	Afrikaans	P. Eliz.	1.69
t_3	Pirlo	Italy	Madrid	Juve	Italian	Flero	1.77

G_1 has cost 1



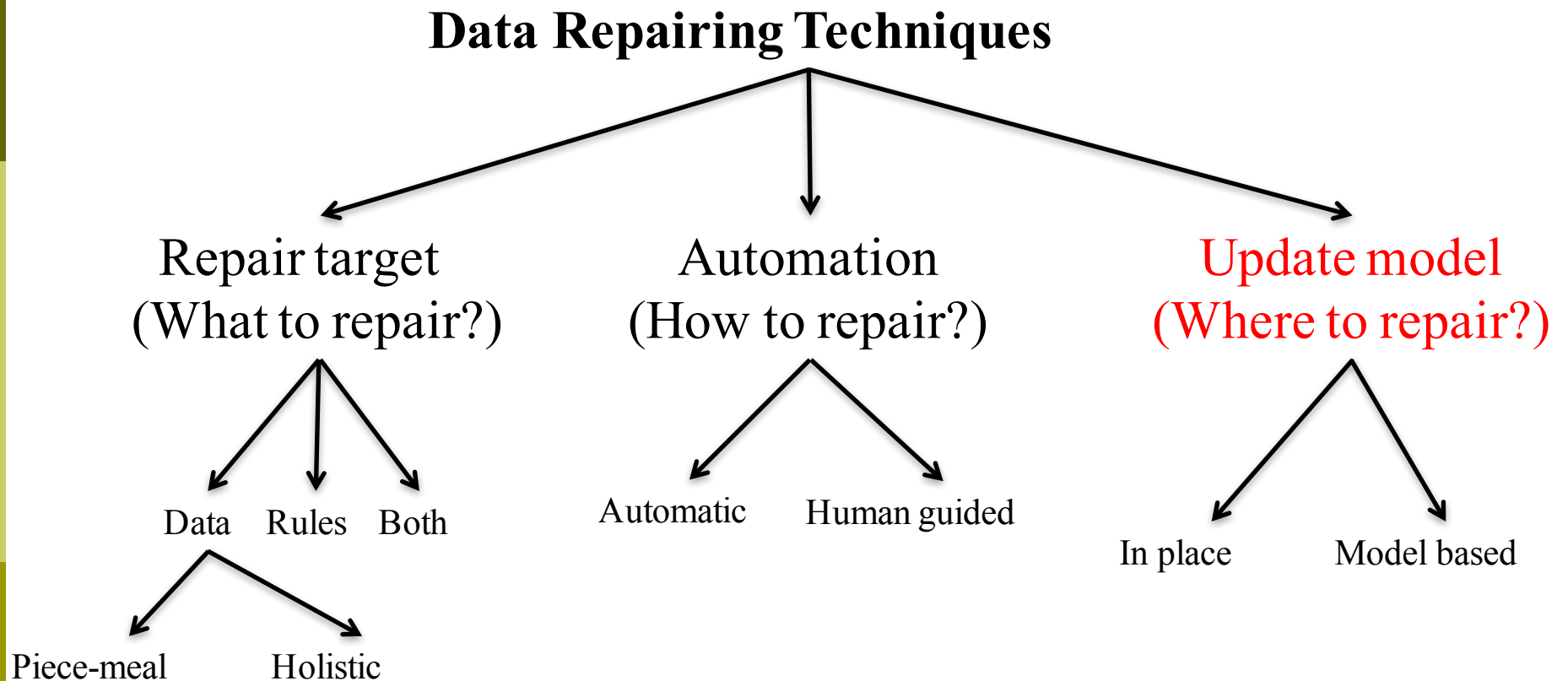
(a) Possible repair G_1

G_2 has cost 5



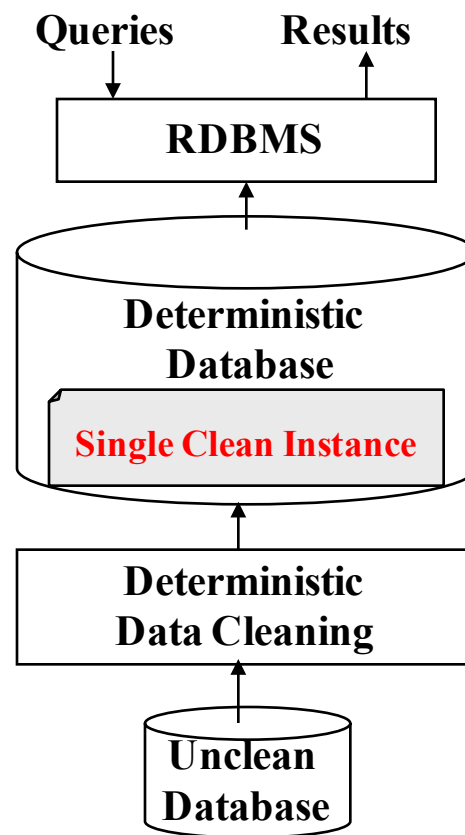
(b) Possible repair G_2

Data Repairing Techniques Taxonomy



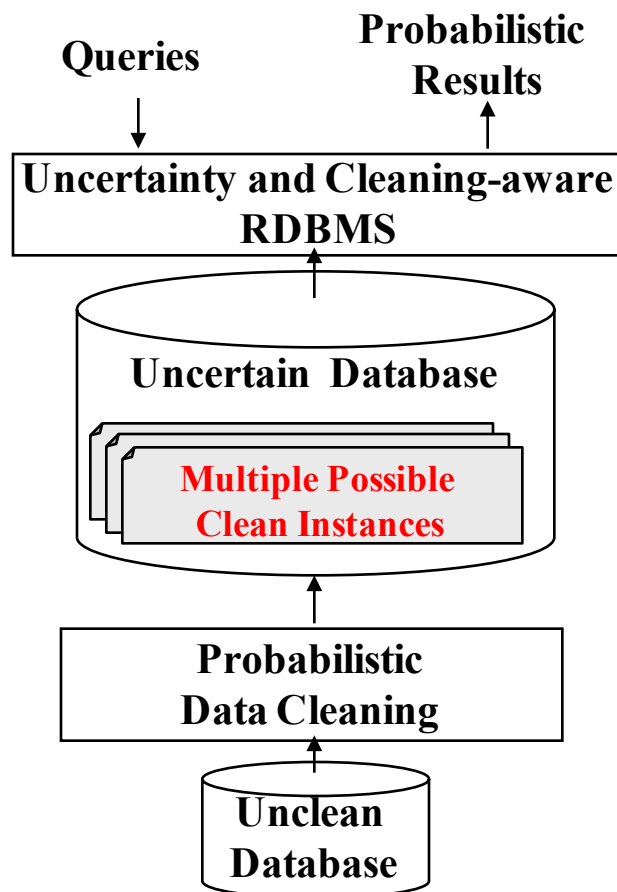
One-Shot Data Cleaning

- Generate a single “trustworthy” instance



Model Based Approach

- Generate multiple possible clean instances



Model Based Approach Challenges

1. The space of all possible repairs is huge
2. How to efficiently **generate, store and query** the possible repairs

Two Example Model Based Approaches

- Duplicate Detection [[Beskales et al, VLDB 2009](#)]
 - Spaces of Possible Repairs
 - Generating and Storing Possible Repairs
 - Query Possible Repairs

- Violations of Functional Dependencies [[Beskales et al, VLDB 2010](#)]
 - Spaces of Possible Repairs
 - Sampling from a Meaningful Space of Repairs

Two Example Model Based Approaches

- Duplicate Detection [Beskales et al, VLDB 2009]
 - Spaces of Possible Repairs
 - Generating and Storing Possible Repairs
 - Query Possible Repairs

- Violations of Functional Dependencies [Beskales et al, VLDB 2010]
 - Spaces of Possible Repairs
 - Sampling from a Meaningful Space of Repairs

Typical Duplicate Elimination

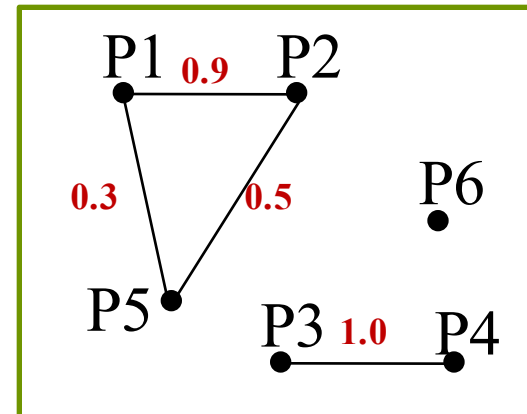
Unclean Relation

ID	name	ZIP	Income
P1	Green	51519	30k
P2	Green	51518	32k
P3	Peter	30528	40k
P4	Peter	30528	40k
P5	Gree	51519	55k
P6	Chuck	51519	30k

Clean Relation

ID	name	ZIP	Income
C1	Green	51519	39k
C2	Peter	30528	40k
C3	Chuck	51519	30k

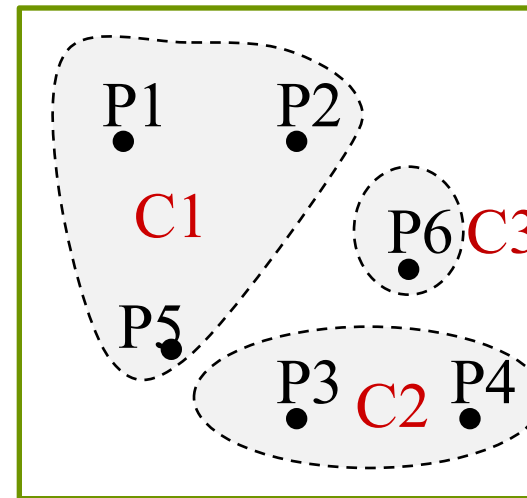
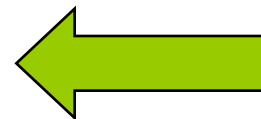
Compute
Pair-wise
Similarity



Cluster
Similar
Records

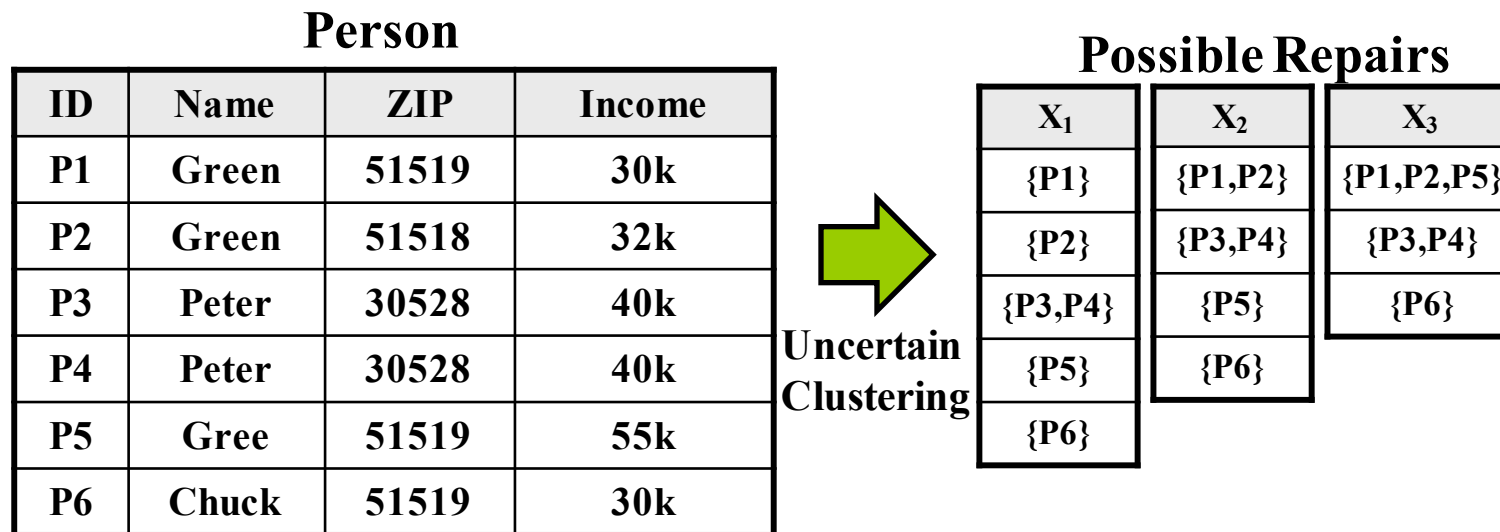


Merge
Clusters

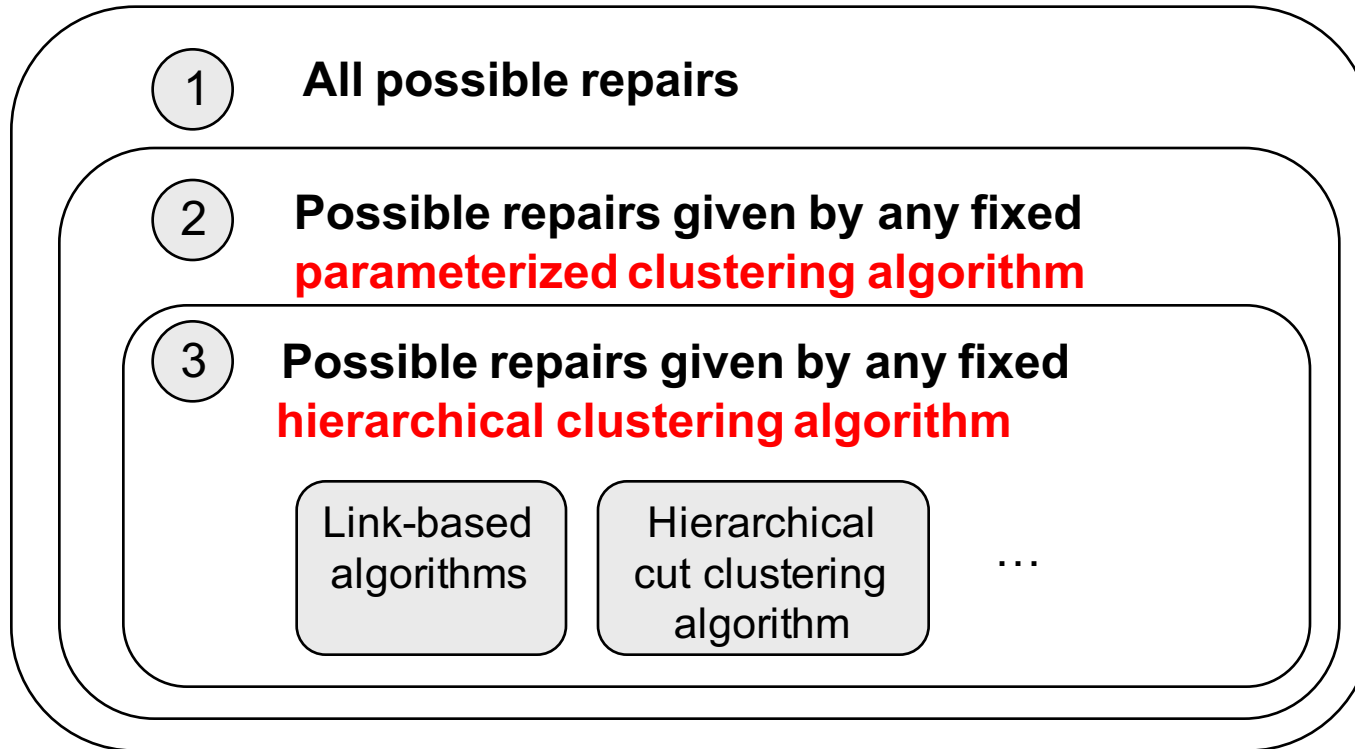


Possible Repairs [Beskales et al, VLDB 2009]

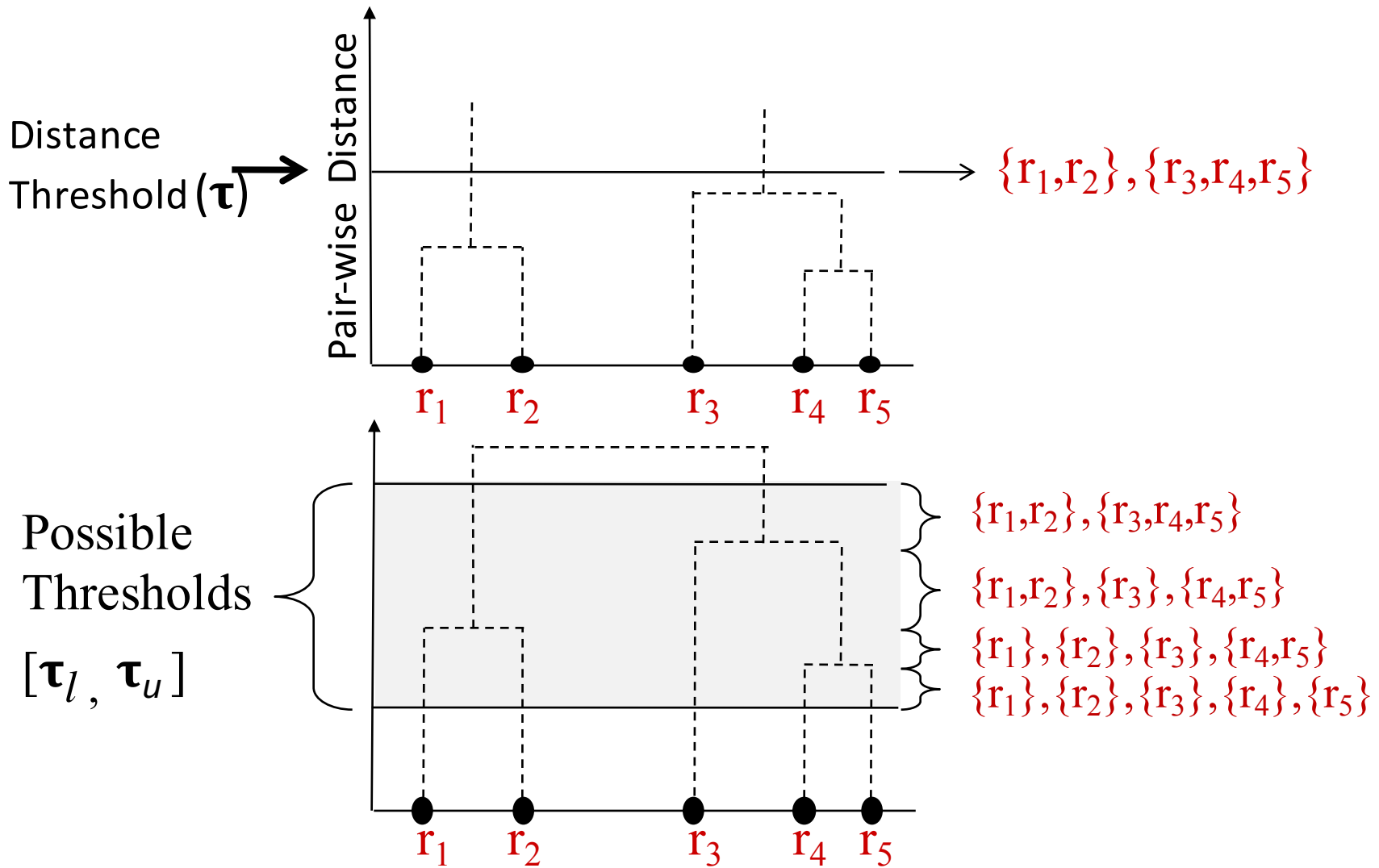
- A **possible repair** is a **clustering** (partitioning) of the input tuples



Spaces of Possible Repairs

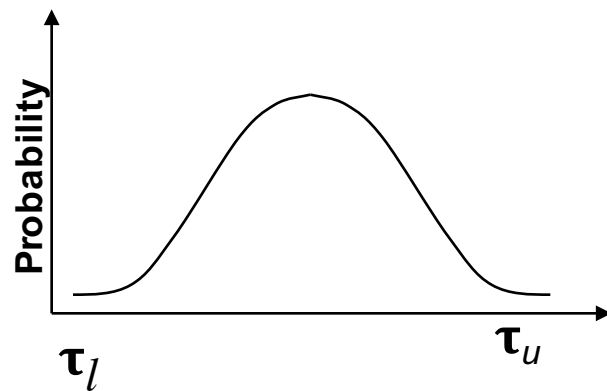


Generating Possible Repairs

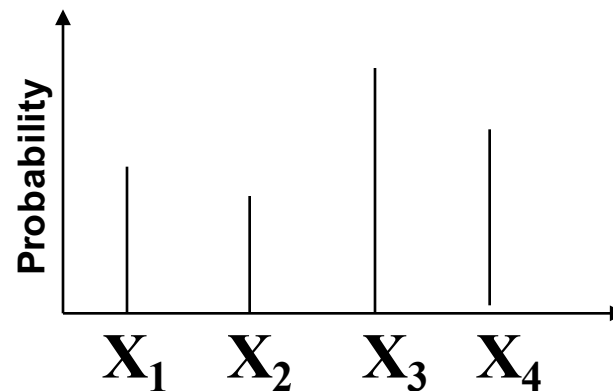
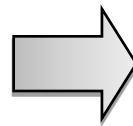


Probabilities of Possible Repairs

- The probability of a repair is equal to the probability of the parameter range that generates such repair



Probability Distribution of τ



Probability Distribution of repairs

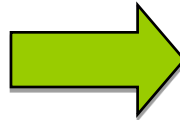
Storing Possible Repairs

□ U-Clean Relations

- Each cluster is stored once
- We keep the "lineage" of each cluster

Clustering 1	Clustering 2	Clustering 3
{P1}	{P1,P2}	{P1,P2,P5}
{P2}	{P3,P4}	{P3,P4}
{P3,P4}	{P5}	{P6}
{P5}	{P6}	
{P6}		

$0 \leq \tau < 1$ $1 \leq \tau < 3$ $3 \leq \tau < 10$



U-clean Relation *Person*^C

ID	...	Income	C	P
CP1	...	31k	{P1,P2}	[1,3)
CP2	...	40k	{P3,P4}	[0,10)
CP3	...	55k	{P5}	[0,3)
CP4	...	30k	{P6}	[0,10)
CP5	...	39k	{P1,P2,P5}	[3,10)
CP6	...	30k	{P1}	[0,1)
CP7	...	32k	{P2}	[0,1)

Example: Projection Query

Person^C

ID	...	Income	C	P
CP1	...	31k	{P1,P2}	[1,3)
CP2	...	40k	{P3,P4}	[0,1)
CP3	...	55k	{P5}	[0,3)
CP4	...	30k	{P6}	[3,10)
CP5	...	40k	{P1,P2,P5}	[3,10)
CP6	...	30k	{P1}	[0,1)
CP7	...	32k	{P2}	[0,1)

**SELECT DISTINCT Income
FROM Person^C**



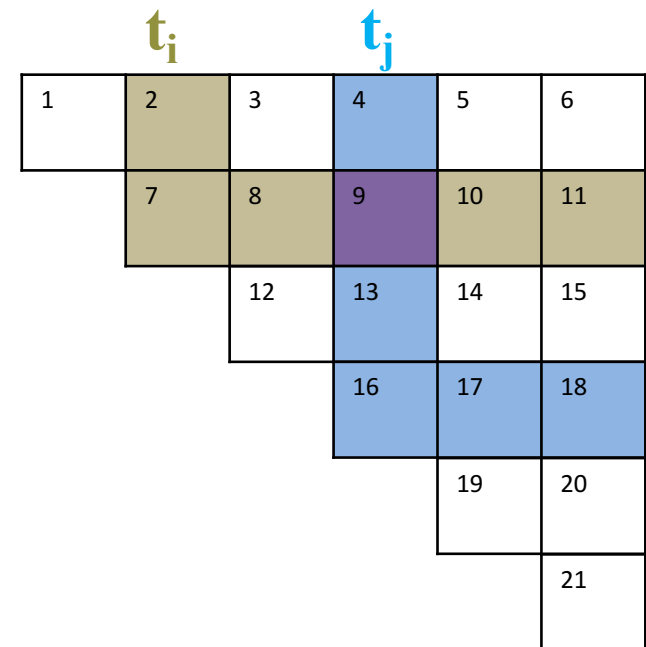
Income	C	P
30k	{P1} v {P6}	[0,1) v [3,10)
31k	{P1,P2}	[1,3)
32k	{P2}	[0,1)
40k	{P3,P4} v {P1,P2,P5}	[0,1) v [3,10)
55k	{P5}	[0,3)

Big Data Cleaning Challenges

- Volume
 - Distributed Data Cleaning
 - Sample Clean (Part 2)
- Velocity
 - Incremental Data Cleaning
- Variety
 - Graph/JSON/RDF
 - Text

Distributed Data Deduplication [Chu et al, VLDB 2016]

- Data deduplication in data lake setting
 - A shared-nothing environment
 - Need to compare every tuple pair
- The goal is to minimizing
 - Largest communication cost
 - Largest computation cost



Conclusion and References

□ Error Detection

■ What (IC Languages and Discovery)

- P. Bohannon, W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis. Conditional functional dependencies for data cleaning. In Proceedings of the 23rd International Conference on Data Engineering, pages 746–755, 2007.
- X. Chu, I. F. Ilyas, and P. Papotti. Discovering denial constraints. Proceedings of the VLDB Endowment, 6(13):1498–1509, 2013.
- W. Fan, X. Jia, J. Li, and S. Ma. Reasoning about record matching rules. Proceedings of the VLDB Endowment, 2(1):407–418, 2009.
- N. Koudas, A. Saha, D. Srivastava, and S. Venkatasubramanian. Metric functional dependencies. In Proceedings of the 25th International Conference on Data Engineering, pages 1275–1278, 2009.
- G. Fan, W. Fan, and F. Geerts. Detecting errors in numeric attributes. In Web-Age Information Management, pages 125–137. Springer, 2014.
- W. Fan, J. Li, S. Ma, N. Tang, and W. Yu. Towards certain fixes with editing rules and master data. Proceedings of the VLDB Endowment, 3(1-2):173–184, 2010.
- J. Wang and N. Tang. Towards dependable data repairing with fixing rules. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, pages 457–468. ACM, 2014.
- M. Interlandi and N. Tang. Proof positive and negative in data cleaning. In 31st IEEE International Conference on Data Engineering, 2015.
- Y. Huhtala, J. Kärkkäinen, P. Porkka, and H. Toivonen. TANE: An efficient algorithm for discovering functional and approximate dependencies. Computer Journal, 42(2):100–111, 1999.
- C. M. Wyss, C. Giannella, and E. L. Robertson. FastFDs: A heuristic-driven, depth-first algorithm for mining functional dependencies from relation instances. In International Conference on Big Data Analytics and Knowledge Discovery, pages 101–110, 2001.

Conclusion and References

□ Error Detection

■ How (Human involvement)

- X. Chu, I. F. Ilyas, and P. Papotti. Holistic data cleaning: Putting violations into context. In 29th IEEE International Conference on Data Engineering, pages 458–469, 2013.
- J. Wang, T. Kraska, M. J. Franklin, and J. Feng. Crowder: Crowdsourcing entity resolution. Proceedings of the VLDB Endowment, 5(11):1483– 1494, 2012.

■ Where (Analytics Layer)

- A. Chalamalla, I. F. Ilyas, M. Ouzzani, and P. Papotti. Descriptive and prescriptive data cleaning. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, pages 445–456, 2014.
- A. Meliou, W. Gatterbauer, S. Nath, and D. Suciu. Tracing data errors with view-conditioned causality. In Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, pages 505–516, 2011.
- X. Wang, X Dong, and A. Meliou. Data X-Ray: A Diagnostic Tool for Data Errors . In Proceedings of the 2015 ACM SIGMOD International Conference on Management of data, pages 1231-1245, 2011.
- M. Bergman, T. Milo, S. Novgorodov, and W Tan. QOCO: A Query Oriented Data Cleaning System with Oracles. Proceedings of the VLDB Endowment, 8(12):1900– 1903, 2015.

Conclusion and References

□ Error Repairing

■ What (Data or Data & Rule)

- P. Bohannon, W. Fan, M. Flaster, and R. Rastogi. A cost-based model and effective heuristic for repairing constraints by value modification. In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pages 143–154. ACM, 2005.
- X. Chu, I. F. Ilyas, and P. Papotti. Holistic data cleaning: Putting violations into context. In 29th IEEE International Conference on Data Engineering, pages 458–469, 2013.
- G. Beskales, I. F. Ilyas, L. Golab, and A. Galiullin. On the relative trust between inconsistent data and inaccurate constraints. In 29th IEEE International Conference on Data Engineering, pages 541–552, 2013.
- F. Chiang and R. J. Miller. A unified model for data and constraint repair. In Proceedings of the 2011 IEEE 27th International Conference on Data Engineering, pages 446–457, 2011.

■ How (Human Involvement)

- M. Yakout, A. K. Elmagarmid, J. Neville, M. Ouzzani, and I. F. Ilyas. Guided data repair. Proceedings of the VLDB Endowment, 4(5):279– 289, 2011.
- X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, and Y. Ye. KATARA: A data cleaning system powered by knowledge bases and crowdsourcing. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pages 1247–1261, 2015.

Conclusion and References

□ Error Repairing

■ Where (Model-based)

- G. Beskales, M. A. Soliman, I. F. Ilyas, and S. Ben-David. Modeling and querying possible repairs in duplicate detection. Proceedings of the VLDB Endowment, pages 598–609, 2009.
- G. Beskales, I. F. Ilyas, and L. Golab. Sampling the repairs of functional dependency violations under hard constraints. Proceedings of the VLDB Endowment, 3(1-2):197–207, 2010.

□ Taxonomy

- I. F. Ilyas, and X. Chu. Trends in Cleaning Relational Data: Consistency and Deduplication . In Foundations and Trends® in Databases, Volume 5, Issue 4, 2015

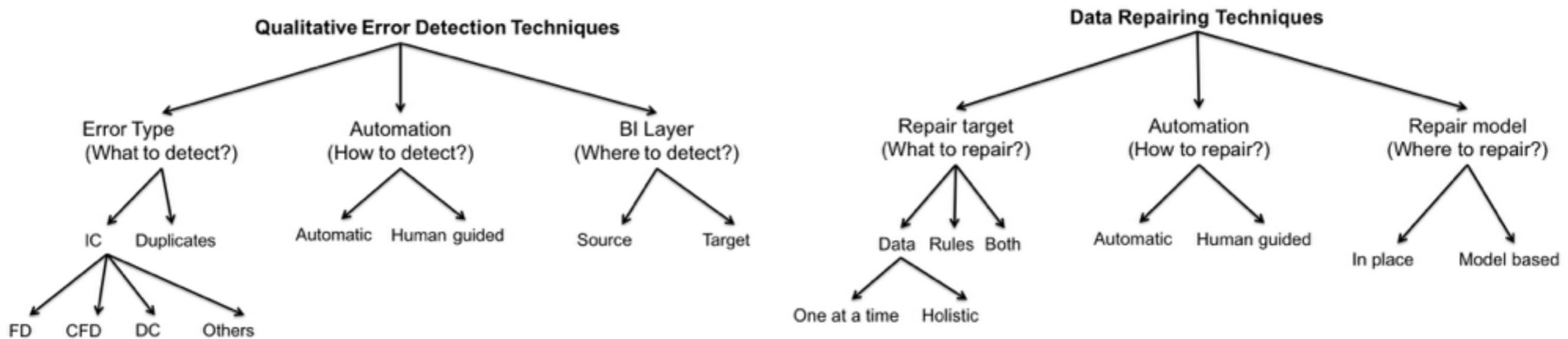
Data Cleaning: A Statistical Perspective

Data Cleaning: Overview and Challenges Part 2

Sanjay Krishnan (UC Berkeley) and
Jiannan Wang (Simon Fraser U.)

Summary From Part 1

- Two steps in data cleaning: error detection and error repair
- Most of the abstractions are based on rules and logic



Part 2. Two Problems

- How can statistical techniques improve efficiency or reliability of data cleaning? (**Data Cleaning with Statistics**)
- How how can we improve the reliability of statistical analytics with data cleaning? (**Data Cleaning For Statistics**)

Part 2. Statistics in Data Cleaning

Data cleaning *with* statistical techniques

- Active Learning for Crowd Sourcing
- Clustering for Entity Resolution
- Probabilistic Extraction
- ...

Data cleaning *for* statistical analysis

- Data Cleaning before aggregate queries
- Data Cleaning before machine learning
- Sensor-network data cleaning
- ...

Part 2. Statistics in Data Cleaning

Data cleaning *with* statistical techniques

ERACER 2010

Guided Data Repair 2011

Corleone 2014

Wisteria 2015

Deep Dive 2014

Katara 2014

Trifacta 2015

Data Tamer 2013

....

Data cleaning *for* statistical analysis

Sensor Net/Stream+ 2000s

Scorpion 2013

SampleClean+ 2014

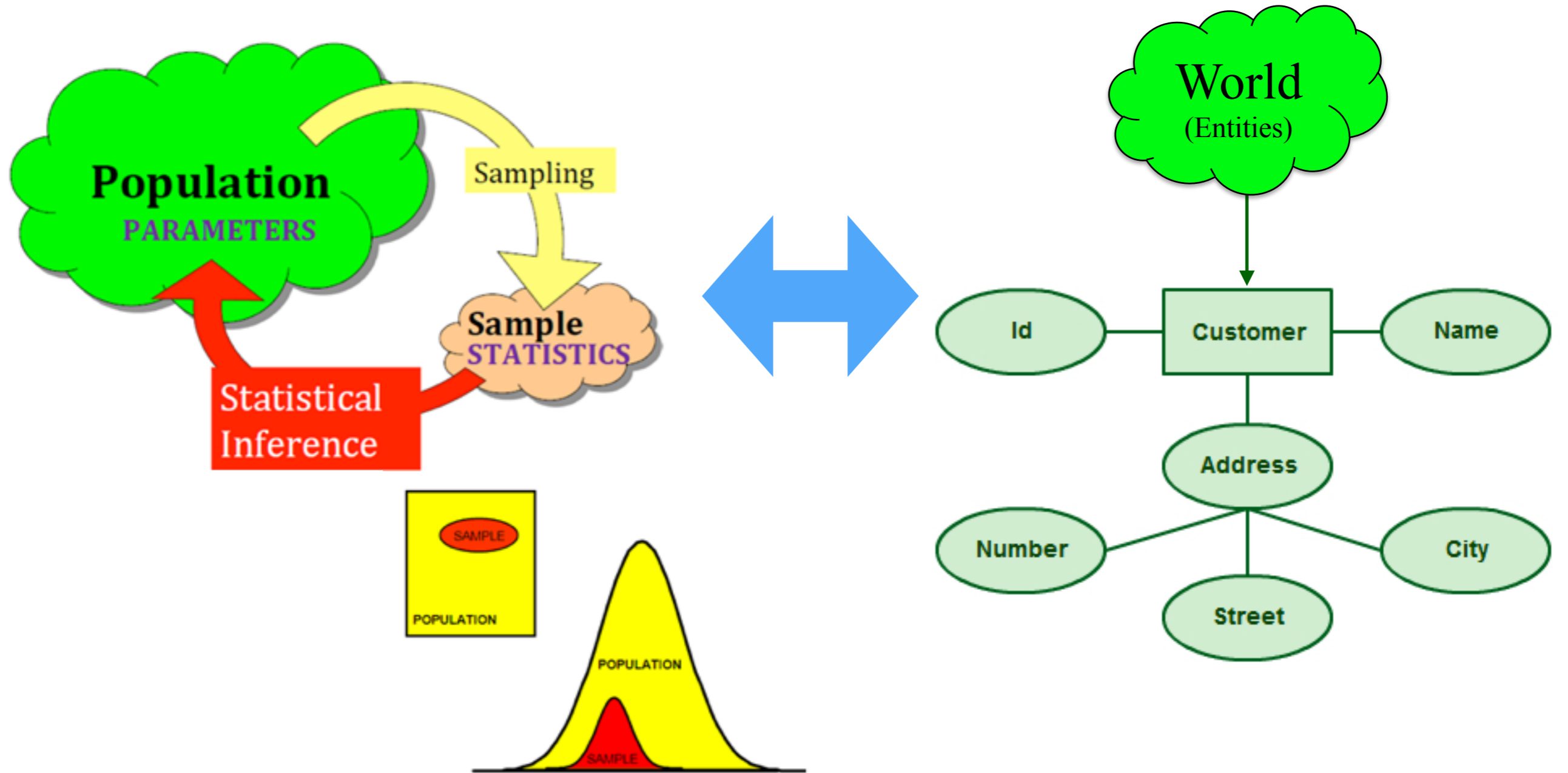
Unknown Unknowns 2016

...

Why Statistics?

- Growing popularity of advanced analytical techniques (e.g., Machine Learning, Stochastic Optimization).
- Increased maturity of ML libraries leads to new opportunities in learning data cleaning from examples.
- Need for end-to-end theoretical analysis

Why Statistics?





Intro to Statistics

Intro to Data Management

Motivating Application



Rakesh Agrawal  


[Microsoft](#)

Publications: [353](#) | Citations: [33537](#)

Fields: [Databases](#), [Data Mining](#), [World Wide Web](#) 

Collaborated with [365 co-authors](#) from 1982 to 2012 | Cited by [24220 authors](#)



Jeffrey D. Ullman  



[Stanford University](#)

Publications: [460](#) | Citations: [43431](#)

Fields: [Databases](#), [Algorithms & Theory](#), [Scientific Computing](#) 

Collaborated with [317 co-authors](#) from 1961 to 2012 | Cited by [31987 authors](#)



Michael Franklin  

[University of California Berkeley](#)

Publications: [561](#) | Citations: [15174](#)

Fields: [Databases](#), [Pharmacology](#), [Data Mining](#) 

Collaborated with [3451 co-authors](#) from 1974 to 2012 | Cited by [15795 authors](#)

Results After Cleaning

Author	Dirty	Clean
Rakesh Agarwal	353	211
Jeffrey Ullman	460	255
Michael Franklin	561	173

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

 Email

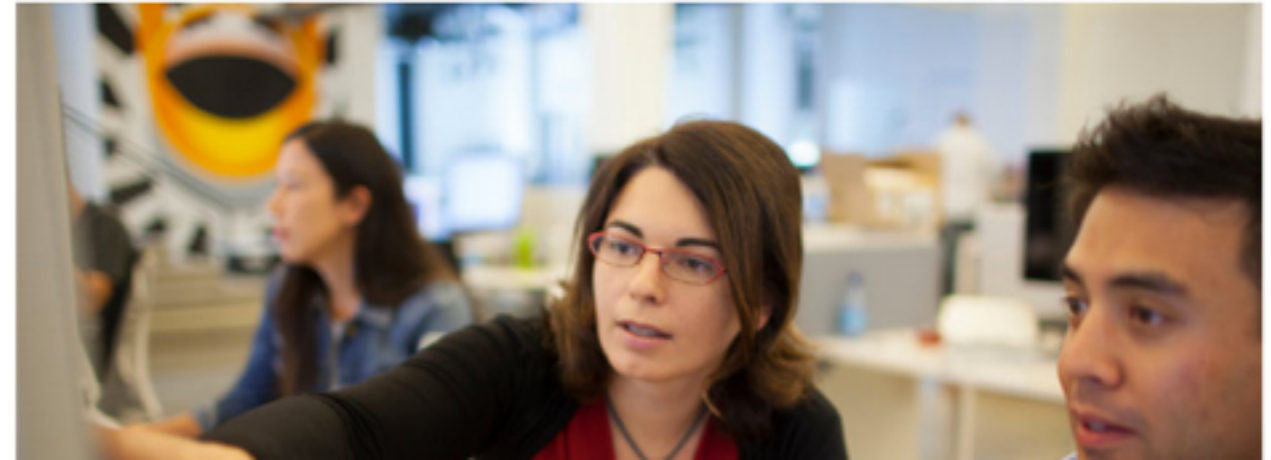
 Share

 Tweet

 Save

Technology revolutions come in measured, sometimes foot-dragging steps. The lab science and marketing enthusiasm tend to underestimate the bottlenecks to progress that must be overcome with hard work and practical engineering.

The field known as “big data” offers a contemporary case study. The catchphrase



Classical Model

- *Def:* Let Σ be a set of constraints on a database \mathbf{D}
 - \mathbf{D} is *inconsistent* if $\exists \sigma \in \Sigma : \sigma(\mathbf{D}) = \text{False}$
- *Error Detection:* Identify a set of rows from the relations in \mathbf{D} such that if removed \mathbf{D} is consistent.
- *Error Repair:* Identify a sequence of repairs $\mathbf{C}_1, \dots, \mathbf{C}_k$ such that $\mathbf{C}_1 \circ \dots \circ \mathbf{C}_k(\mathbf{D})$ is consistent

Classical Model: Strengths

- Clear definition of consistency
- Complexity Analysis and Optimality: Time Complexity, Space Complexity, Minimality, Undecidability.
- Lends itself to declarative systems.

Classical Model: Limitations

- **Problem 1:** Hard to express some types of data cleaning in terms of rules/logic
- **Problem 2:** Consistency is not a statistical definition
- **Problem 3:** Orthogonal to downstream analysis

Limitation 1. Hard to express in rules

Madden, Samuel R., et al. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (2005): 122-173.

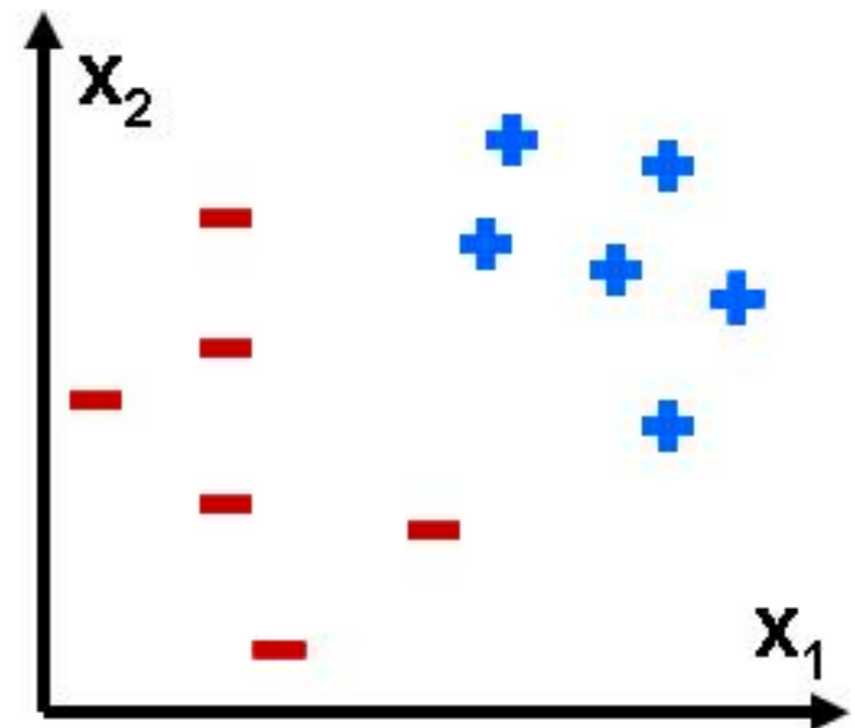
Madden, S. R., Franklin, M. J., Hellerstein, J. M., & Hong, W. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (2005): 122-173.

- Easy to determine whether two records are duplicates
- Harder to define similarity functions, blocking rules, and thresholds

Teaser 1. Learn From Examples

Madden, Samuel R., et al. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (2005): 122-173.

Madden, S. R., Franklin, M. J., Hellerstein, J. M., & Hong, W. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (2005): 122-173.



Limitation 2. Consistency is not a statistical definition

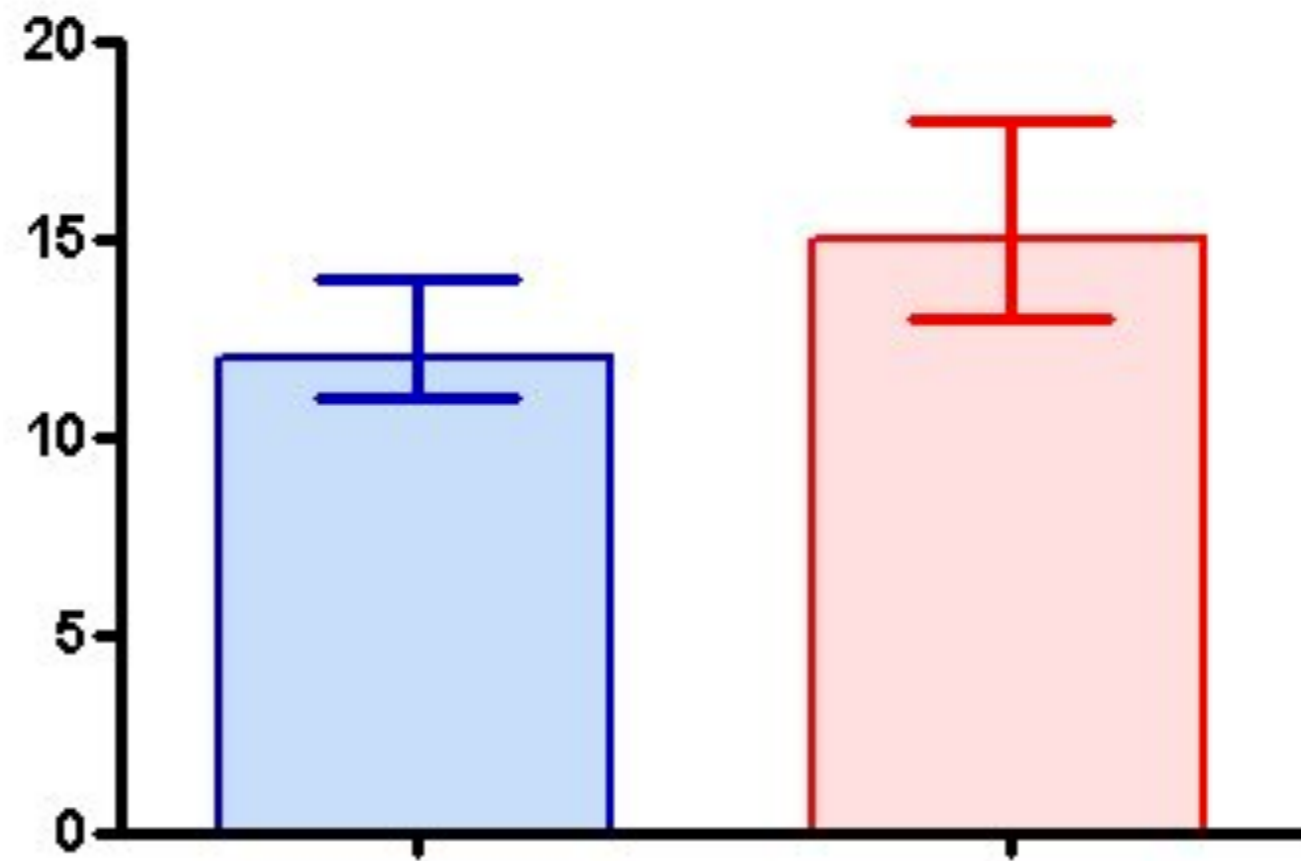
- Aggregate queries can be ambiguous.
- Consistency is not statistical accuracy

Author	Year	Paper ID
Agrawal	1992	1334
Agrawal	1978	1451
Agrawal	1996	1651
Ullman	1994	22331

Author	Year	Paper ID
Agrawal	1992	1334
Agrawal	1996	1651
Ullman	1994	22331

Teaser 2. Definitions that are compatible with statistical analysis

A Procedural Definition: Given a cleaning program **C()**



Limitation 3. Ignores Downstream Analysis

- Queries are important
- `SELECT count(1) where author_last=agarwal;`
- Is # Ullman > # Agarwal
- Train a model to recommend database publications

Teaser 3. Value of Cleaning Each Record

Prioritize using statistics or information theory

		Author	Year	Paper ID
0.151	C()	Agrawal	1992	1334
0.956	C()	Agrawal	1999	1451
0.256	C()	Agrawal	1996	1651
0.126	C()	Ullman	1994	22331

Strengths of Statistical Techniques

- Leverage recent advances in ML to learn from examples.
- Robust to false positives and false negatives
- Composition with downstream statistical analytics
- Leverage statistical theory: sample complexity, unbiasedness, convergence rates.

A Statistical Perspective

- Topic 1. Statistical techniques to clean data (20 mins) (**Limitation 1**)
- Topic 2. Cleaning data before statistical analytics (50 min) (**Limitation 2, 3**)
- Topic 3. Impact and Future Directions (10 mins)

A Statistical Perspective

- **Topic 1. Statistical techniques to clean data (20 mins)**
- Topic 2. Cleaning data before statistical analytics (50 min)
- Topic 3. Impact and Future Directions (10 mins)

Section Structure

- **Hot topic for a long time**

- ✓ Data Profiling, Outlier Detection, Value Imputation

- **Trend 1: Using Statistical Machine Learning**

This tutorial

- ✓ Data Blocking (for Entity Resolution)
- ✓ Data Repairing
- ✓ Data Transformation

- **Trend 2: Combining Statistical Techniques with Crowdsourcing**

- ✓ Workflow Design
- ✓ Quality/Cost/Latency Trade-off

Section Structure

- Hot topic for a long time

- ✓ Data Profiling, Outlier Detection, Value Imputation

- **Trend 1: Using Statistical Machine Learning**

This tutorial

- ✓ Data Blocking (for Entity Resolution)

- ✓ Data Repairing

- ✓ Data Transformation

- Trend 2: Combining Statistical Techniques with Crowdsourcing

- ✓ Workflow Design

- ✓ Quality/Cost/Latency Trade-off

Data Blocking

- **Entity Resolution**

- ✓ Finding different records that refer to the same real-world entity
- ✓ For example: “iPhone 4th Gen” vs “iPhone four”

- **Challenge: N^2 comparisons**

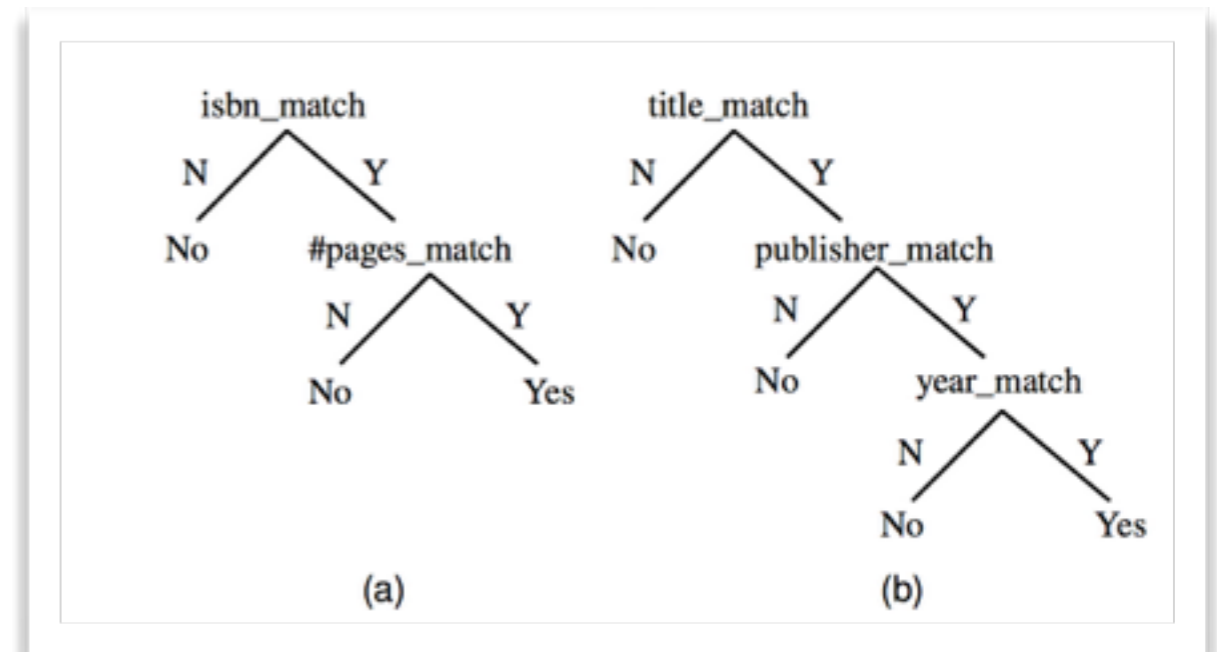
- **Solution: Blocking**

- ✓ Using blocking rules to remove obviously non-matched pairs
- ✓ For example: “If the brand of two products are different, they cannot matching”

Data Blocking as Learning a Random Forest

- **Random Forest**

- ✓ A collection of decision trees
- ✓ Each decision tree is learnt from a sample of the same training set



- **Blocking Rules**

- ✓ Generating (candidate) blocking rules from the random forest

(isbn_match = N) → No
(isbn_match = Y) and (#pages_match = N) → No
(title_match = N) → No
(title_match = Y) and (publisher_match = N) → No
(title_match = Y) and (publisher_match = Y) and (year_match = N) → No

Data Repairing

- **Error Detection/Correction (See Part 1)**
 - ✓ Detect/Correct erroneous values that violate integrity constraints

IDs	Name	Street	City	State	Zip
t1	Joe	Main St.	Bellevue	WA	98004
t2	Mark	Main St.	Bellevue	WA	980-04
t3	Andy	Main St.	Bellevue	WA	98005
t4	Lee	MS Way	Redmond	WA	98052
t5	James	Campus St.	Seattle	WA	98195
...

$$\phi : (Street, City \rightarrow Zip)$$

Data Repairing as Learning a Multiclass Classifier

- **Generating Possible Repairs**

- ✓ A data repairing algorithm will first suggest a number of possible repairs for the erroneous values
- ✓ An example repair: $(t_2, \text{zip}, 98004, 90\%)$, which means that $t_2[\text{zip}]$ should be updated to 98004 (with a 90% confidence).

- **Training a Multiclass Classifier**

- ✓ Label a sample of the possible repairs:
 1. “Confirm”, the value of $t_2[\text{zip}]$ should be 98004
 2. “Reject”, the value of $t_2[\text{zip}]$ should not be 98004
 3. “Retain”, the value of $t_2[\text{zip}]$ is correct
- ✓ Train a classifier based on the labeled repairs and use it to classify other (unlabeled) repairs

Data Transformation

Heer, Jeffrey, Joseph M. Hellerstein, and Sean Kandel. "Predictive Interaction for Data Transformation." CIDR. 2015.

- **Data Transformation**

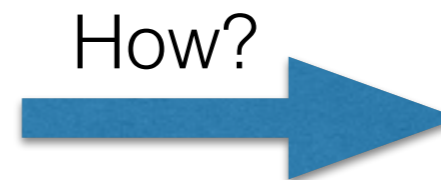
- ✓ Converts a set of data values from the data format of a source data system into the data format of a destination data system.

- **An example task: Pattern Extraction**

Mobile Advertising Logs

Extracted Values

```
31 adtam_name=utarget1&adtam_source=dynamic&adtam_size=180x150
32 adtam_name=holidaypromo1&adtam_source=dynamic&adtam_size=300x250
33 adtam_name=utarget1&adtam_source=dynamic&adtam_size=180x150
34 adtam_name=holidaypromo2&adtam_source=mobile&adtam_size=240x400
```

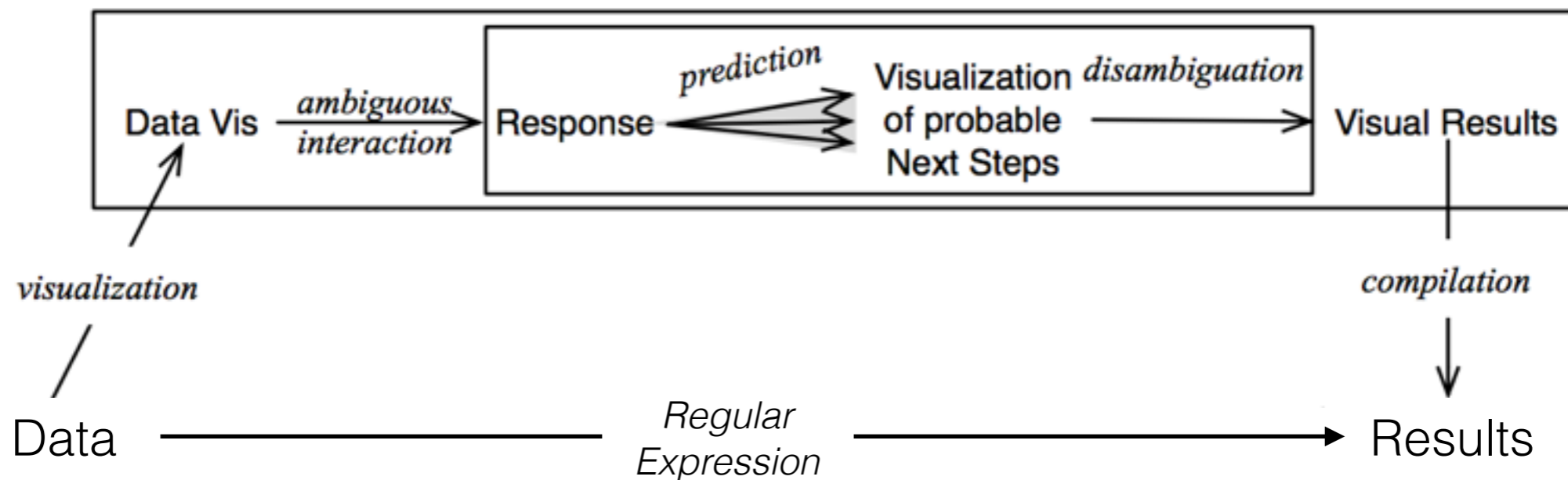


```
dynamic
dynamic
dynamic
mobile
```

Regular Expression? `/(?<=adtam_source\=)[^\&]*(?=\&)/`

Data Transformation as Learning a Predictive Model

- Predictive Interaction**



```
31 adtam_name=utarget1&adtam_source=
32 adtam_name=holidaypromo1&adtam_sou
33 adtam_name=utarget1&adtam_source=
34 adtam_name=holidaypromo2&adtam_sou
```

```
/(?<=adtam_source\=)[^\&]*(?=\&)/
```

```
dynamic
dynamic
dynamic
mobile
```


Data Transformation as Learning a Predictive Model



- **Demonstration**

TRANSFORM EDITOR

```
extract col: Screen_Detail after: `adtam_source=` before: `&`
```

SUGGESTED TRANSFORMS

- extract col: Screen_Detail after: `adtam_source=` before: `&`
- extract col: Screen_Detail limit: 2 after: `=` before: `&`
- extract col: Screen_Detail on: `[lower]+` limit: 2 after: `=`

Source	Preview
abc Screen_Detail	abc Screen... abc Dev
 6 Categories	 2 Categories
31 adtam_name=utarget1&adtam_source=dynamic&adtam_size=180x150	dynamic
32 adtam_name=holidaypromo1&adtam_source=dynamic&adtam_size=300x250	dynamic
33 adtam_name=utarget1&adtam_source=dynamic&adtam_size=180x150	dynamic
34 adtam_name=holidaypromo2&adtam_source=mobile&adtam_size=240x400	mobile

Section Structure

- Hot topic for a long time

- ✓ Data Profiling, Outlier Detection, Value Imputation

- Trend 1: Using Statistical Machine Learning

- ✓ Data Blocking (for Entity Resolution)

- ✓ Data Repairing

- ✓ Data Transformation

- **Trend 2: Combining Statistical Techniques with Crowdsourcing**

- ✓ Workflow Design

- ✓ Quality/Cost/Latency Trade-off

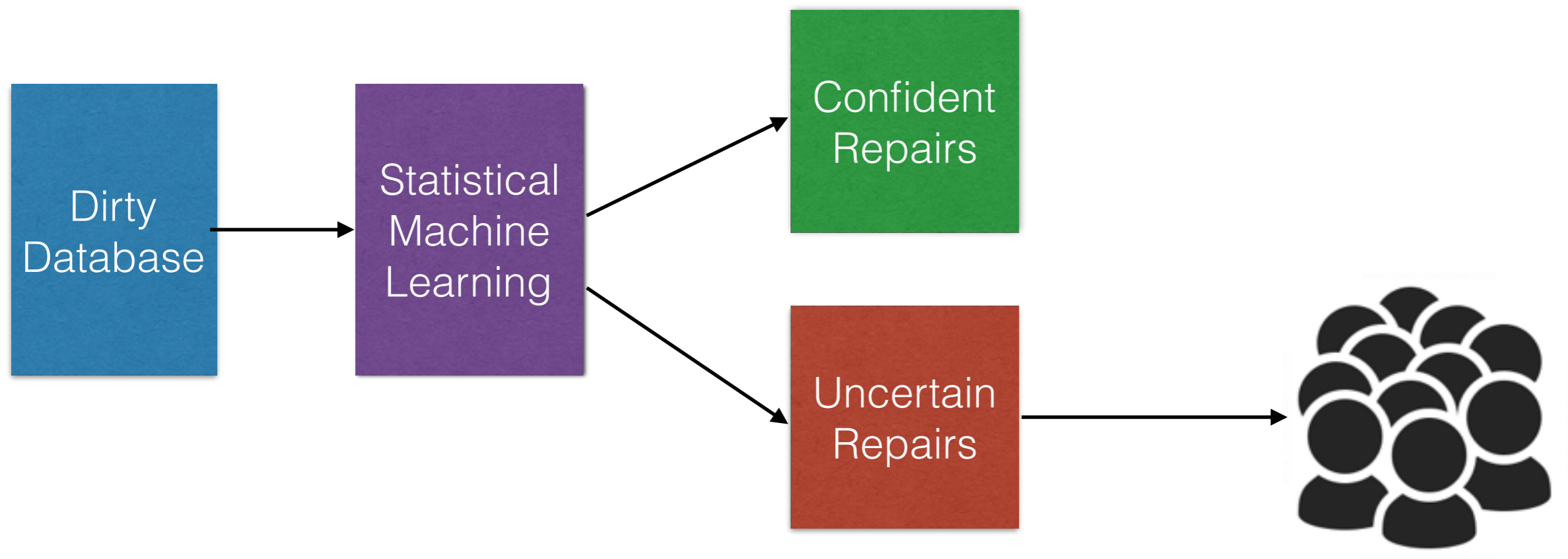
Crowds and Data Cleaning

- Data cleaning algorithms are often hard to achieve high quality without human involvement
- Crowdsourcing platforms makes the use of humans for doing data cleaning tasks easier and cheaper
- Crowdsourcing is widely applied in industrial data cleaning on Extraction and Entity Resolution problems*.

*Marcus, Adam, and Aditya Parameswaran. "Crowdsourced data management: industry and academic perspectives." Foundations and Trends in Databases 6.1-2 (2015): 1-161.

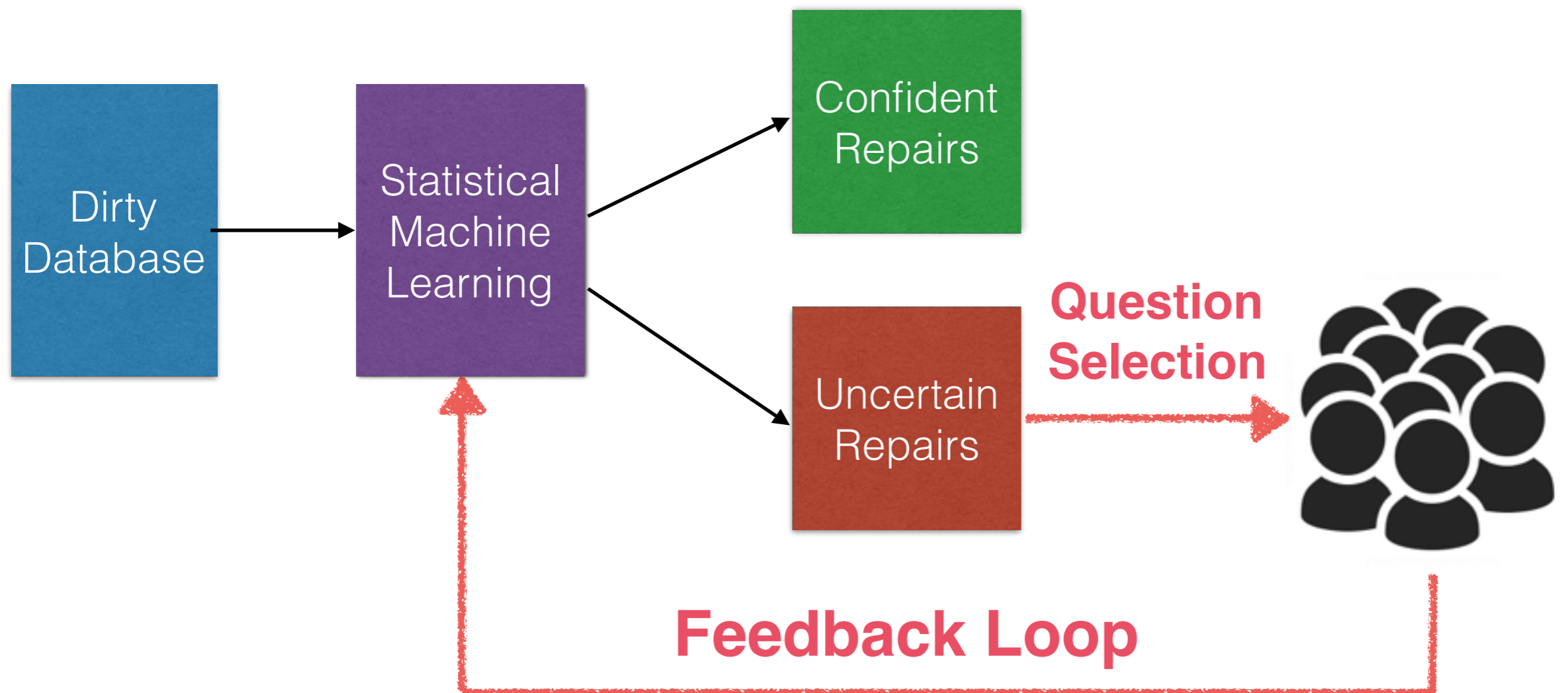
How to combine statistical techniques with crowds?

- **Non-iterative Workflow**



How to combine statistical techniques with crowds?

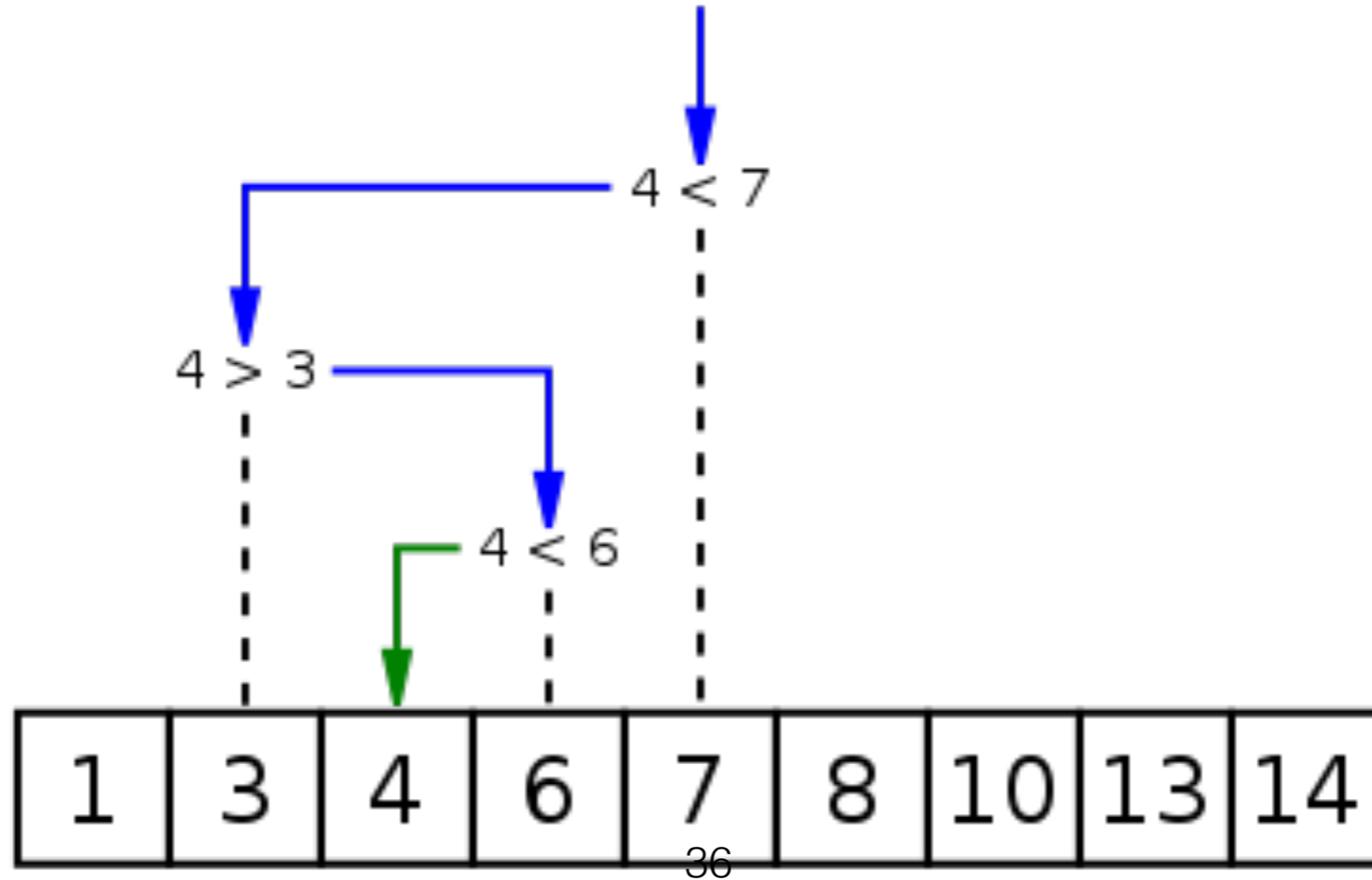
- **Iterative Workflow (a.k.a Active Learning)**



Simplest Active Learning Algorithm

- **Binary Search**

- ✓ *Bad question selection strategy: $O(n)$*
- ✓ *Good question selection strategy: $O(\log n)$*



Question Selection Strategies

- **Uncertain Sampling** [CLAMShell 2016, Wisteria 2015, Mozafari et al. 2014]
- **Query-By-Committee** [Guided Data Repairing 2011, Corleone 2014]
- **Expected Error Reduction** [Arasu et al. 2010, Bellare et al. 2012]
- **Expected Model Change**
- **Variance Reduction**
- **Density-Weighted Methods**

Quality/Cost/Latency Tradeoff

- **Tradeoff**

- ✓ **Quality:** How accurate are cleaning results?
- ✓ **Cost:** How much money need to pay for crowds?
- ✓ **Latency:** How much time does data cleaning need?

- **Crowds**

VS

- **Machines**

- ✓ Pros: Quality
- ✓ Cons: Cost and Latency

- ✓ Pros: Cost and Latency
- ✓ Cons: Quality

How to balance quality/cost/latency?

- **Quality Control**

- ✓ Worker Elimination, Answer Aggregation, Task Assignment, Worker Modeling

- **Latency Control**

- ✓ Task Pricing, Straggler mitigation, Pool maintenance, Hybrid Learning, Latency Model

- **Cost Control**

- ✓ Task Selection, Answer Deduction, Pruning, Sampling

1. Guoliang Li, Jiannan Wang, Yudian Zheng, Michael Franklin. "Crowdsourced data management: A survey." TKDE 2016

2. Anand Inasu Chittilappilly, Lei Chen, and Sihem Amer-Yahia. "A Survey of General-Purpose Crowdsourcing Techniques." TKDE 2016

Q&A

- **Introduction**

- Why statistical perspective?
- Strengths and limitations of classical cleaning models

- **Statistical techniques to clean data**

- **Trend 1: Using Statistical Machine Learning**

- ✓ Data Blocking (for Entity Resolution)
- ✓ Data Repairing
- ✓ Data Transformation

- **Trend 2: Combining Statistical Techniques with Crowdsourcing**

- ✓ Workflow Design
- ✓ Quality/Cost/Latency Trade-off

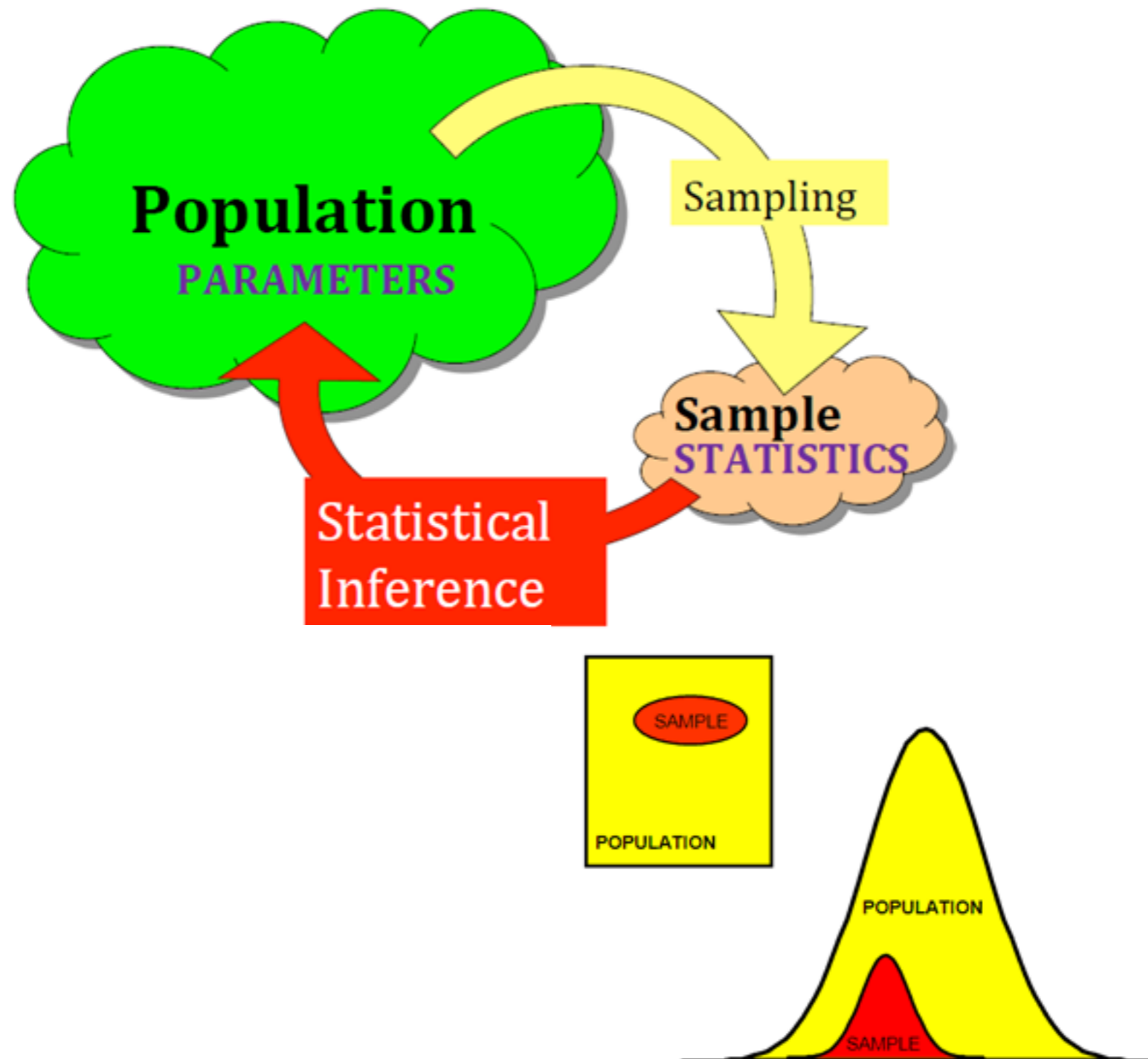
A Statistical Perspective

- Topic 1. Statistical techniques to clean data (20 mins)
- **Topic 2. Cleaning data before statistical analytics (50 min)**
- Topic 3. Impact and Future Directions (10 mins)

Section Structure

- **Extended Data Cleaning Definition**
- Connecting Data Cleaning to Downstream Queries
 - Aggregate queries
 - Machine learning training
 - Exploiting Relational Information

Why Statistics?



Downstream Analytics



- Clean just enough for the ultimate data product
 - How to clean (prev section)
 - How much to clean (**this section**)
 - Where to clean (**prev + this section**)

The Philosophy

- Let \mathbf{D} be a dirty database, there exists one “true” cleaned \mathbf{D}' .
- There exists a sequence of transformations to \mathbf{D} such that $\mathbf{D}' = \mathbf{C}_1 \circ \dots \circ \mathbf{C}_k(\mathbf{D})$, where each $\mathbf{C} \in \mathcal{C}$.
- The process of data cleaning is *discovering* the transformations $\mathbf{C}_1 \circ \dots \circ \mathbf{C}_k$.
- *Effort*: The number of records (\mathbf{k}) in \mathbf{D} to discover $\mathbf{C}_1 \circ \dots \circ \mathbf{C}_k$
- *Query Result Error*: Let \mathbf{q} be an aggregate query, the error is defined as $\|\mathbf{q}(\mathbf{D}) - \mathbf{q}(\mathbf{D}')\|$

Interaction Model

- Queries the database, observes a dirty record $\mathbf{r} \in \mathbf{R}$.
- Designs a transformation \mathbf{C}_1 to correct the dirty records (and possibly other records with the same error)
- How many queries to find $\mathbf{C}_1 \circ \dots \circ \mathbf{C}_k$



Supported Data Cleaning

- Record-by-Record Transformations
- De-duplication
- Extraction
- Not yet: Schema Matching, Complex Constraints

Record-by-Record

- *Def:* Given a dirty record $\mathbf{r} \in \mathbf{R}$, the analyst can return an \mathbf{r}'

Madden, Samuel R., et al. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (**2005**): 122-173.

Madden, Samuel R., et al. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (**2005**): 122-173.

Maps and Filters

- *Def:* Analyst applies a program consisting of Map and Filter operations to the database such that for the particular $\mathbf{r} \in \mathbf{R}$ the program returns an \mathbf{r}'

Madden, Samuel R., et al. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (**2005**): 122-173.

Madden, Samuel R., et al. "TinyDB: an acquisitional query processing system for sensor networks." **Database Systems, Transactions Of** 30.1 (**2005**): 122-173.

Maps and Filters

- *Def:* Analyst applies a program consisting of Map and Filter operations to the database such that for the particular $\mathbf{r} \in \mathbf{R}$ the program returns an \mathbf{r}'

Madden, Samuel R., et al. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (**2005**): 122-173.

Madden, Samuel R., et al. "TinyDB: an acquisitional query processing system for sensor networks." **Database Systems, Transactions Of** 30.1 (**2005**): 122-173.

Entity Resolution

- *Def:* Given a dirty record $\mathbf{r} \in \mathbf{R}$, the analyst returns the number of times the record is duplicated in the relation

Madden, Samuel R., et al. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (2005): 122-173.

Madden, S. R., Franklin, M. J., Hellerstein, J. M., & Hong, W. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (2005): 122-173.

2

Extraction

- *Def:* Given a dirty record $\mathbf{r} \in \mathbf{R}$, the analyst returns a record $\mathbf{r}' \in \mathbf{R}'$ over an extended set of attributes $\mathbf{A} = \mathbf{A} \cup \mathbf{E}'$

Madden, Samuel R., et al. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (2005): 122-173.

Madden, S. R., Franklin, M. J., Hellerstein, J. M., & Hong, W. "TinyDB: an acquisitional query processing system for sensor networks."

ACM Transactions on database systems (TODS) 30.1 (2005)

Section Structure

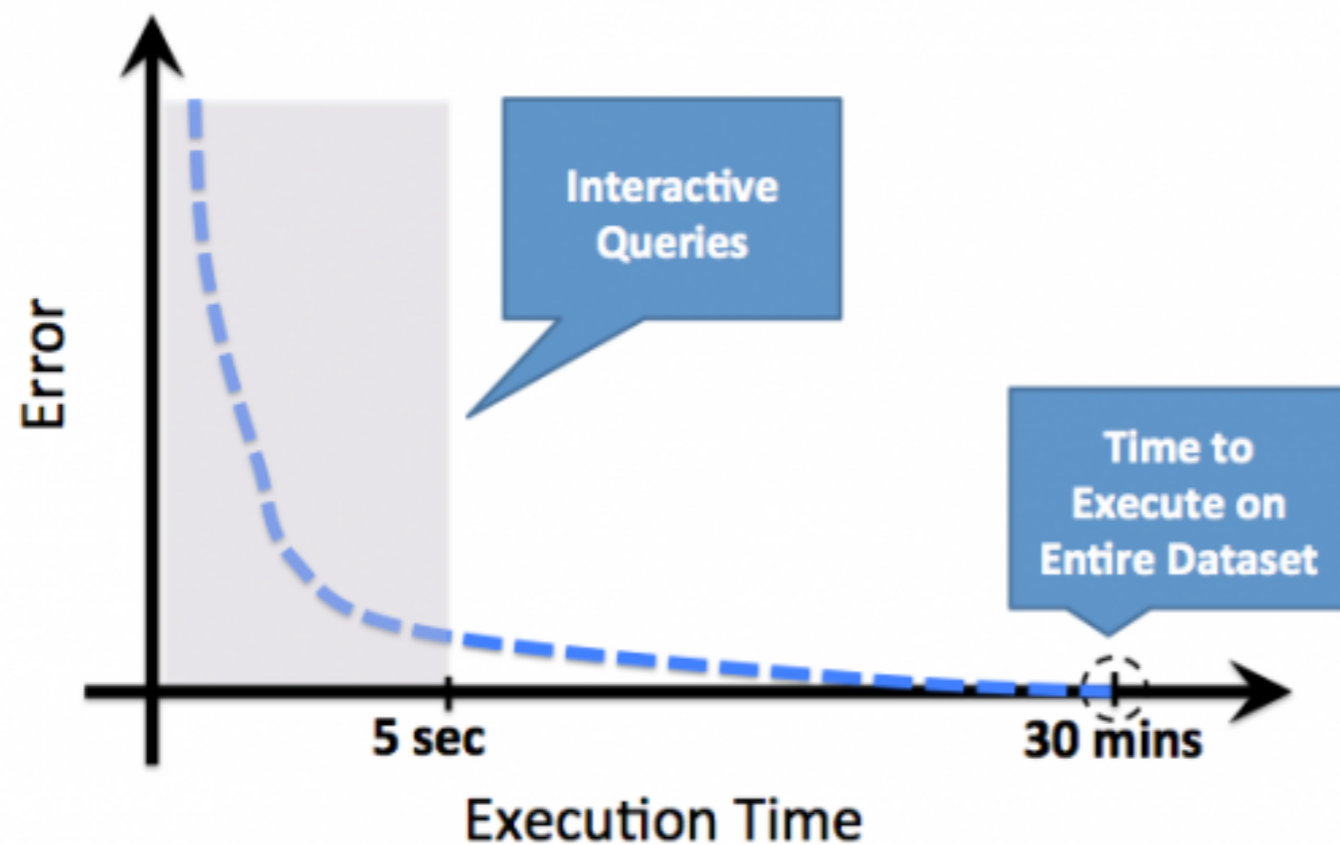
- Extended Data Cleaning Definition
- **Connecting Data Cleaning to Downstream Queries**
 - Aggregate queries
 - Machine learning training
 - Exploiting Relational Information

Section Structure

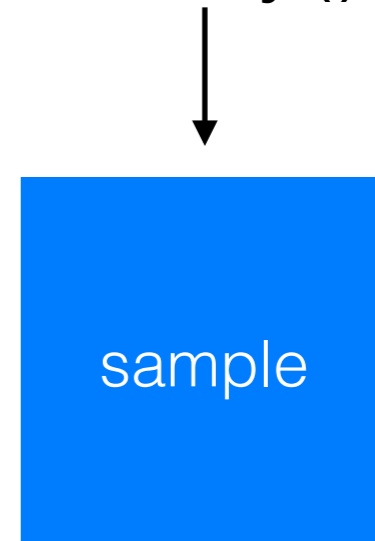
- Extended Data Cleaning Definition
- Connecting Data Cleaning to Downstream Queries
 - **Aggregate queries**
 - Machine learning training
 - Exploiting Relational Information

Aggregate Queries “Concentrate”

Speed/Accuracy Trade-off



Query()



accuracy: $1/\sqrt{N}$

Agarwal, Sameer, et al. "BlinkDB: queries with bounded errors and bounded response times on very large data." Proceedings of the 8th ACM European Conference on Computer Systems. ACM, 2013.



Properties

- Estimate: Query run on a subset of data (or a partially clean dataset)
- Unbiased: the expected value of an estimate is equal to the true value.
- Consistent: as sample goes to the dataset size the estimate limits to the true value.

Example Query

- Rank the authors by publication count



Rakesh Agrawal  


[Microsoft](#)

Publications: [353](#) | Citations: [33537](#)

Fields: [Databases](#), [Data Mining](#), [World Wide Web](#) 

Collaborated with [365 co-authors](#) from 1982 to 2012 | Cited by [24220 authors](#)



Jeffrey D. Ullman  



[Stanford University](#)

Publications: [460](#) | Citations: [43431](#)

Fields: [Databases](#), [Algorithms & Theory](#), [Scientific Computing](#) 

Collaborated with [317 co-authors](#) from 1961 to 2012 | Cited by [31987 authors](#)



Michael Franklin  

[University of California Berkeley](#)

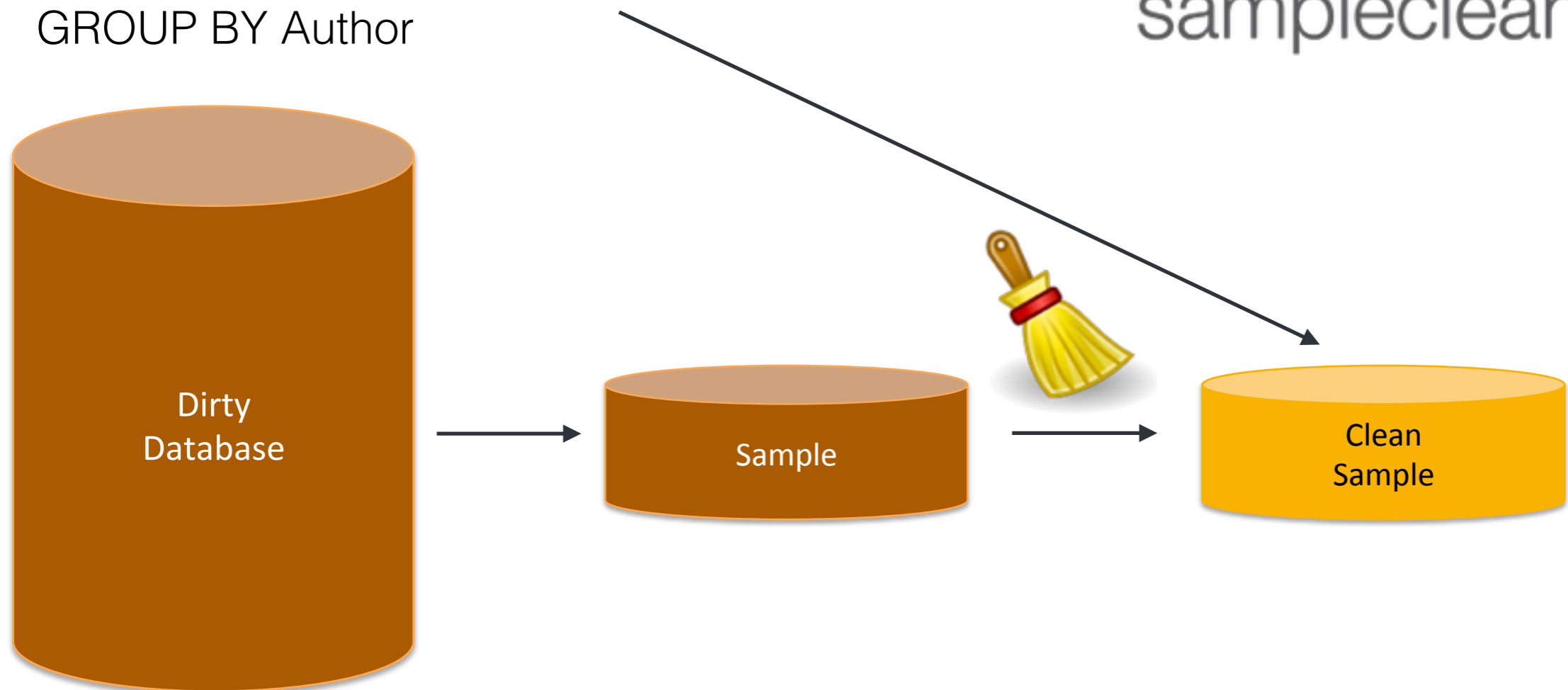
Publications: [561](#) | Citations: [15174](#)

Fields: [Databases](#), [Pharmacology](#), [Data Mining](#) 

Collaborated with [3451 co-authors](#) from 1974 to 2012 | Cited by [15795 authors](#)

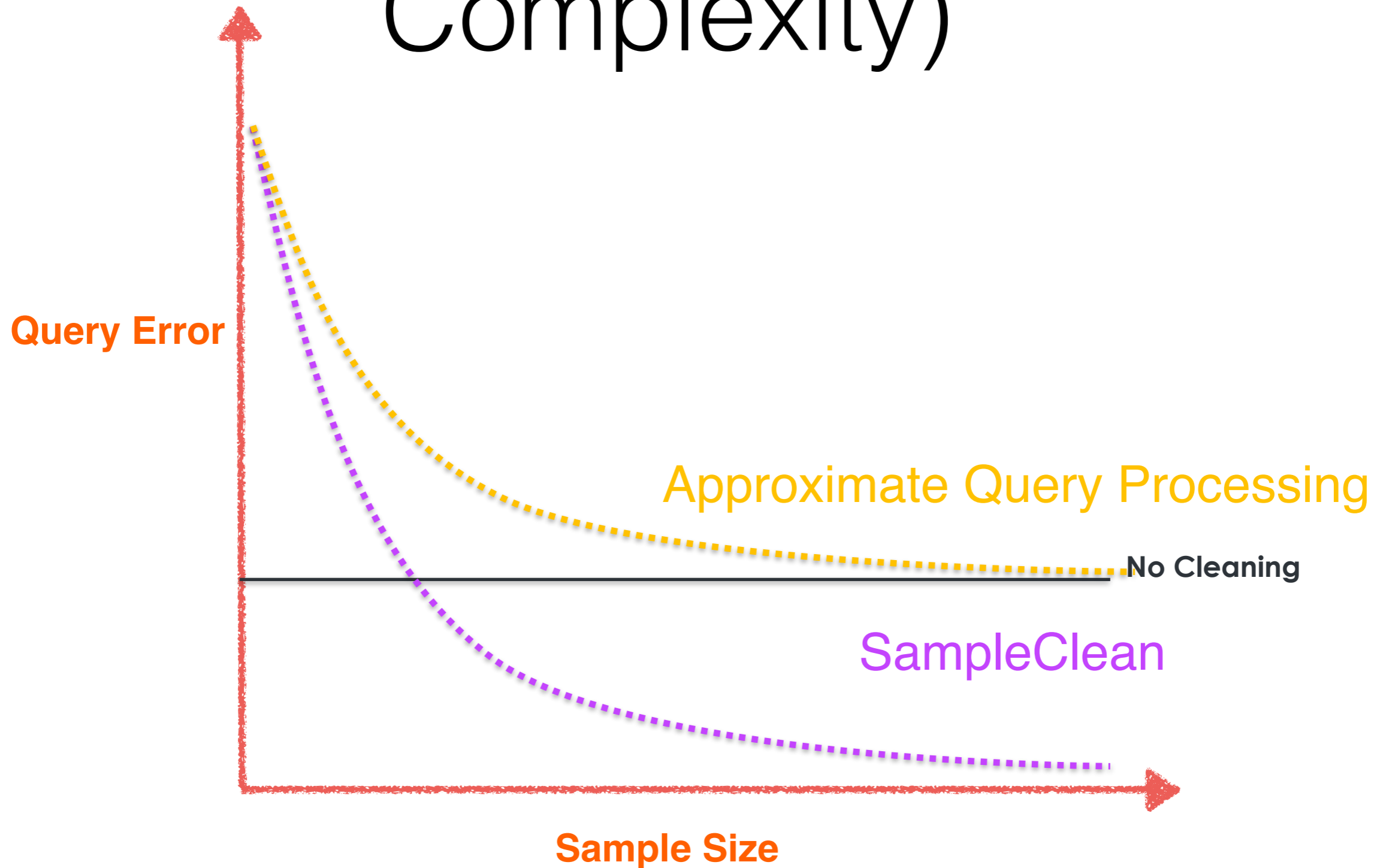
Sample-and-Clean

```
SELECT COUNT(1)  
FROM Pubs  
GROUP BY Author
```



Jiannan Wang, Sanjay Krishnan, Michael Franklin, Ken Goldberg, Tim Kraska, Tova Milo. A Sample-and-Clean Framework for Fast and Accurate Query Processing on Dirty Data. In SIGMOD 2014

Intuition (Sample Complexity)



Taxonomy of Data Errors

```
SELECT F(attr)
FROM table
WHERE condition
GROUP BY attrs
```

- SUM/COUNT/AVG
- Incorrect value in attr (Value Error)
- Incorrectly satisfied condition or group (Condition Error)
- Record is duplicated (Duplication Error)

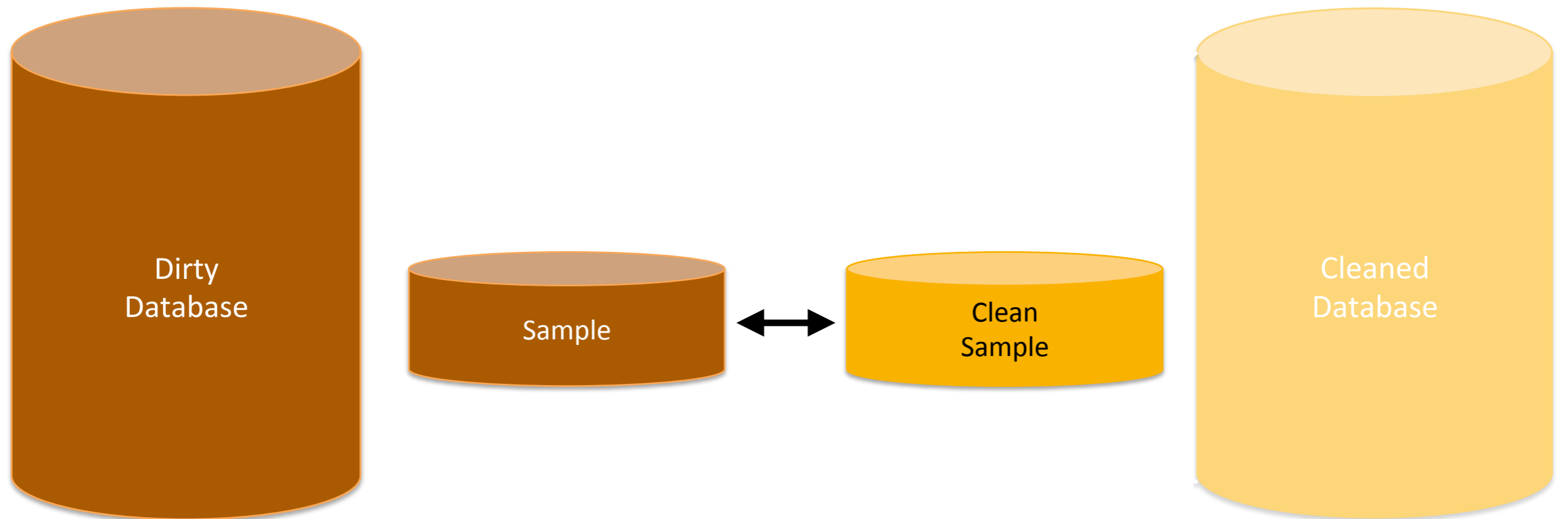
Probabilistic Interpretation

- SUM, COUNT, AVG, VAR can be expressed as a **mean**.
 - SUM = size * mean
 - COUNT = size * frequency
- Probabilistic Interpretation: Expected Values

$$\mathbb{E}(X) = \sum x \cdot \underline{\mathbb{P}(X = x)}$$

$$\bar{x} \propto \sum_{i=1}^k \text{clean}(x) \cdot \frac{\text{predicate}(x)}{\text{dup}(x)}$$

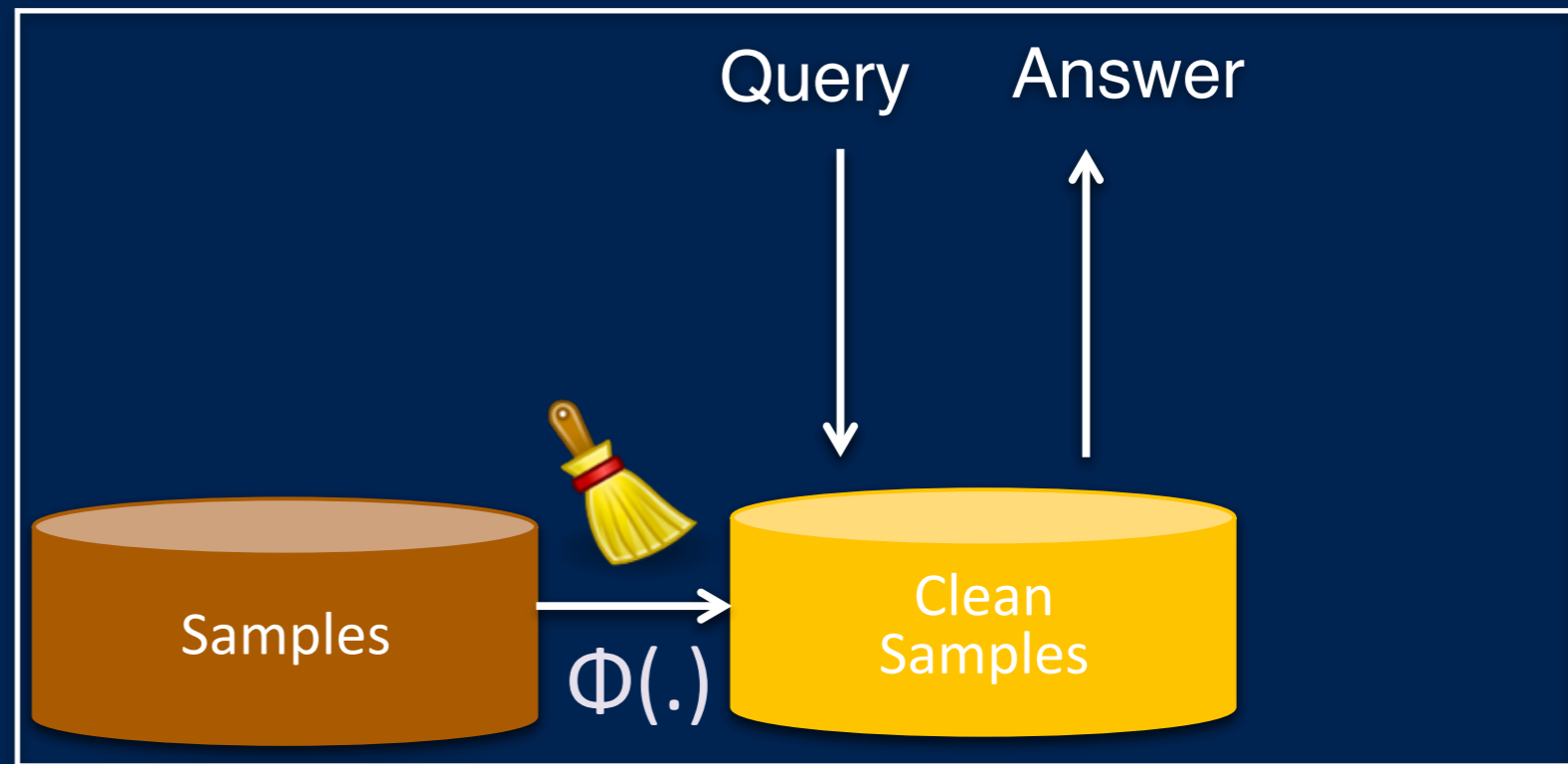
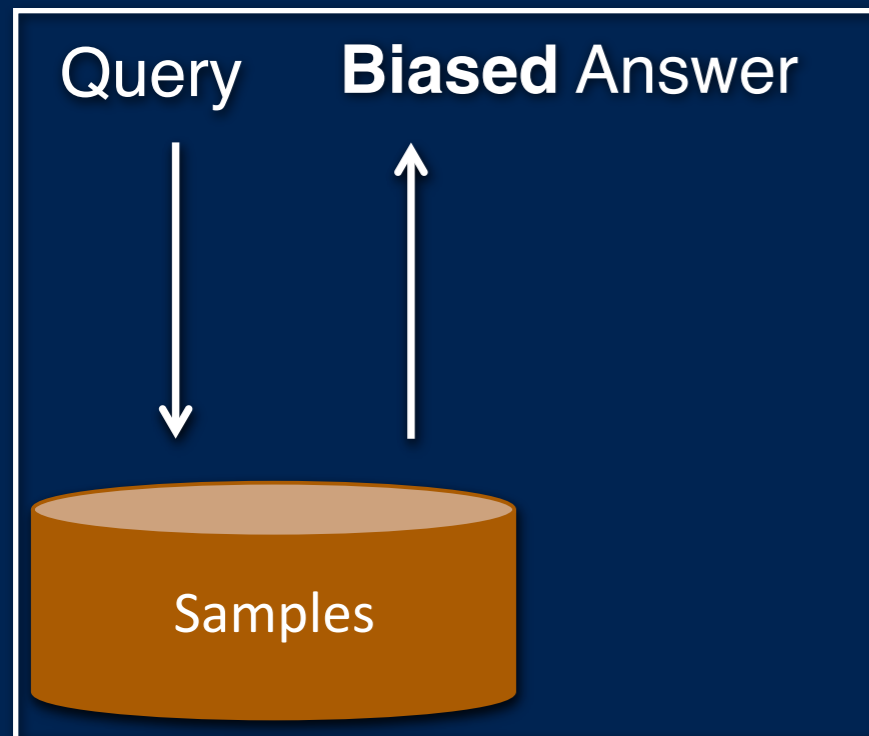
Transform Dirty Sample to Simulate Clean Sample



Algorithm 1: Direct Estimate

Approximate Queries

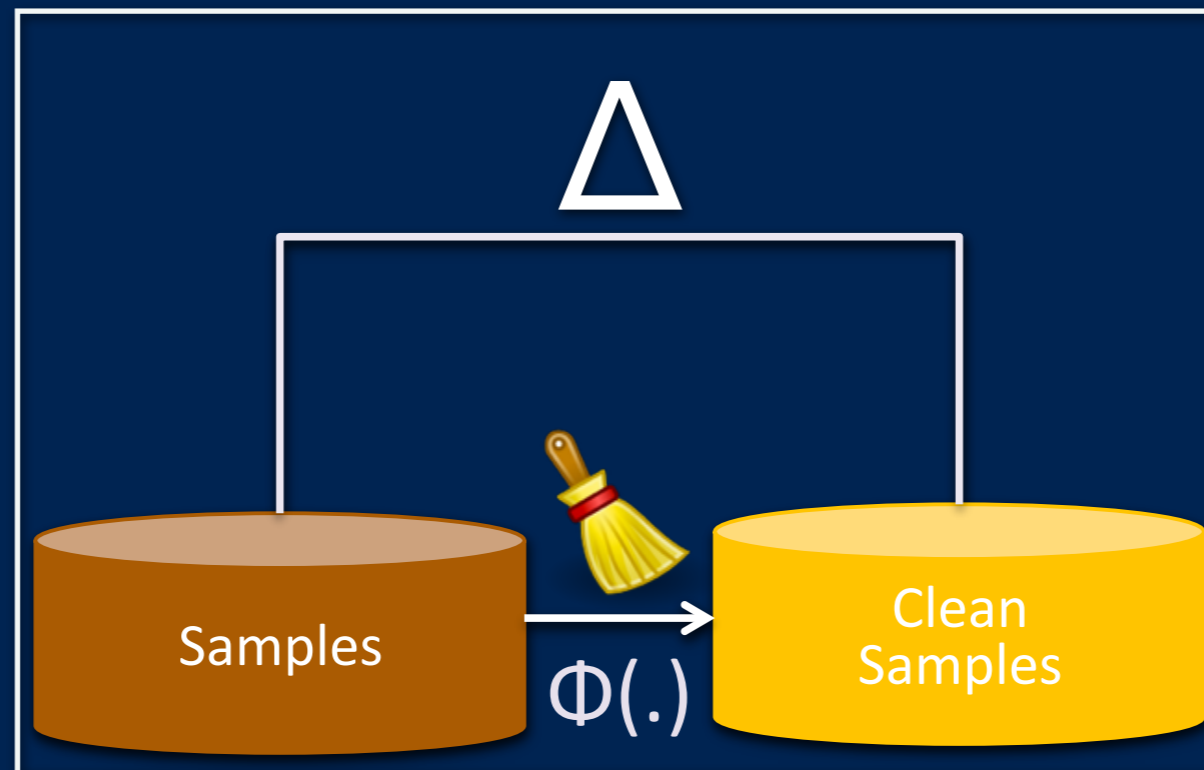
Direct Estimate



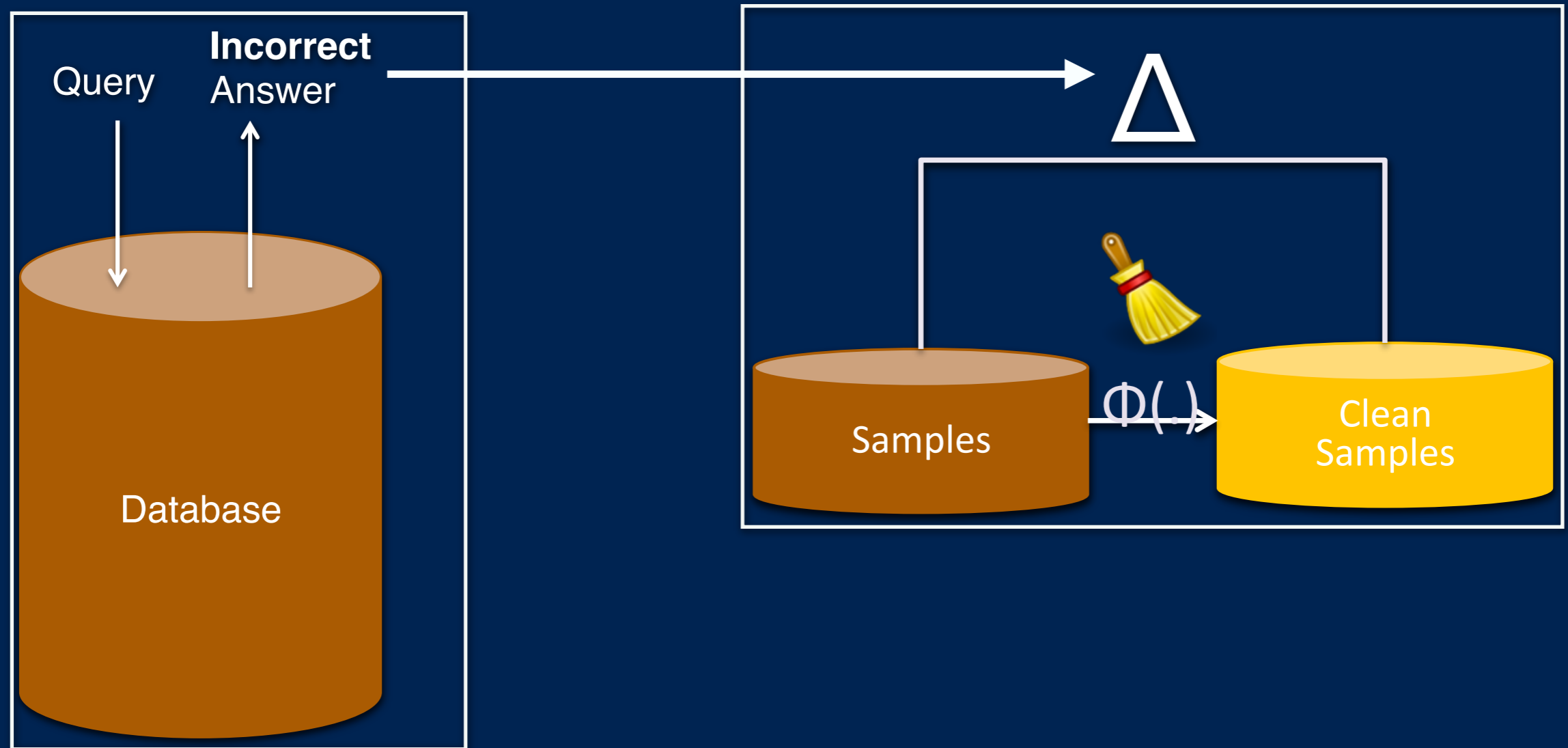
Jiannan Wang, Sanjay Krishnan, Michael Franklin, Ken Goldberg, Tim Kraska, Tova Milo. A Sample-and-Clean Framework for Fast and Accurate Query Processing on Dirty Data. In SIGMOD 2014

Algorithm 2: Corrected Estimate

How much did the cleaning change the data?



Algorithm 2: Corrected Estimate



Probabilistic Interpretation

- Has probabilistic guarantees about accuracy
 - Unbiased estimates
 - Bounded in confidence intervals

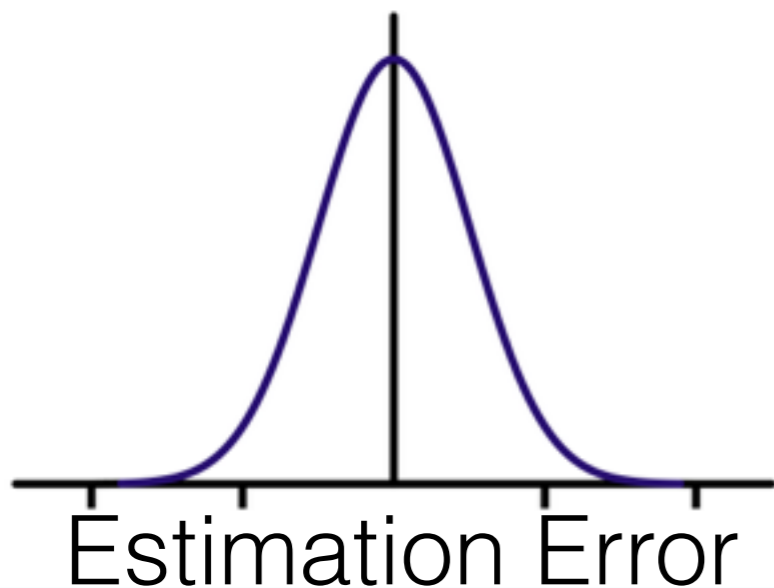
Two Types of Bounds

- Central Limit Theorem: Asymptotic (Very Tight)
- Chernoff Bounds: Finite-sample (Looser)

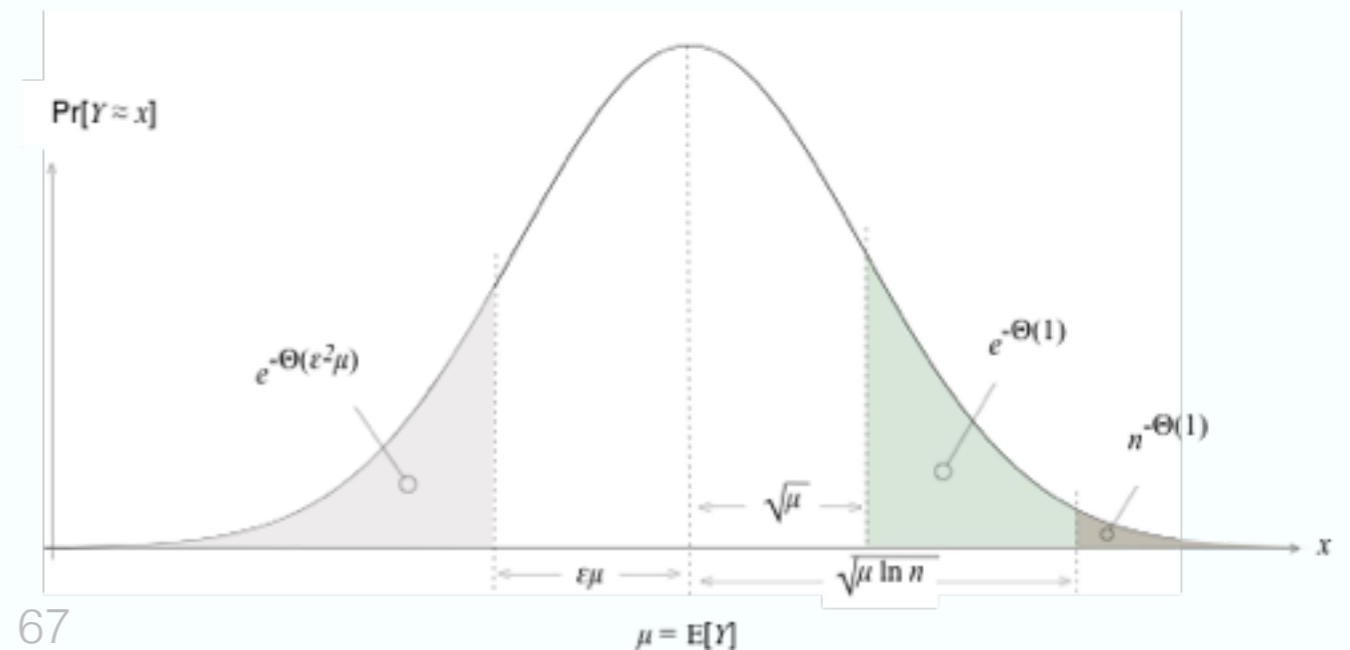
$$\mathbf{X} = \{X_1, X_2, \dots, X_k\} \text{ i.i.d}$$

$$\bar{x} = \frac{1}{k} \sum_{i=0}^k X_i$$

CLT



Chernoff



Central Limit Theorem

Central Limit Theorem: means of independent random variables converge to a normal distribution

$$\mathbf{X} = \{X_1, X_2, \dots, X_k\} \text{ i.i.d}$$

$$\bar{x} = \frac{1}{k} \sum_{i=0}^k X_i$$

Unbiased With Bounds

$$\bar{x} \sim N\left(E(\mathbf{X}), \frac{Var(\mathbf{X})}{k}\right)$$

Direct vs. Corrections

Asymptotic SUM/COUNT/AVG

Clean Estimate

Dirty Correction

Accuracy $O\left(\frac{\textit{var}(\textit{clean})}{k}\right)$

$O\left(\frac{\textit{var}(\textit{diff})}{k}\right)$

Runtime

$O(k)$

$O(N)$

FPC: $\text{sqrt}(N-k)/\text{sqrt}(N-1)$

Chernoff Bound

Chernoff Bound: random variables tend to concentrate around their mean value.

$$\mathbf{X} = \{X_1, X_2, \dots, X_k\} \text{ i.i.d}$$

$$\bar{x} = \frac{1}{k} \sum_{i=0}^k X_i$$

$$\mathbb{P} \left(\left| \bar{X} - \mathbf{E} [\bar{X}] \right| \geq t \right) \leq 2 \exp \left(- \frac{2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right)$$

Direct vs. Corrections

Finite Sample SUM/COUNT/AVG

Clean Estimate

Dirty Correction

Accuracy $O\left(\frac{\text{range}(\textit{clean})}{k}\right)$

$O\left(\frac{\text{range}(\textit{diff})}{k}\right)$

Runtime





$O(k)$





$O(N)$





$\text{range}(S) = \max(S) - \min(S)$

Example Query

- Count the number of distinct publications

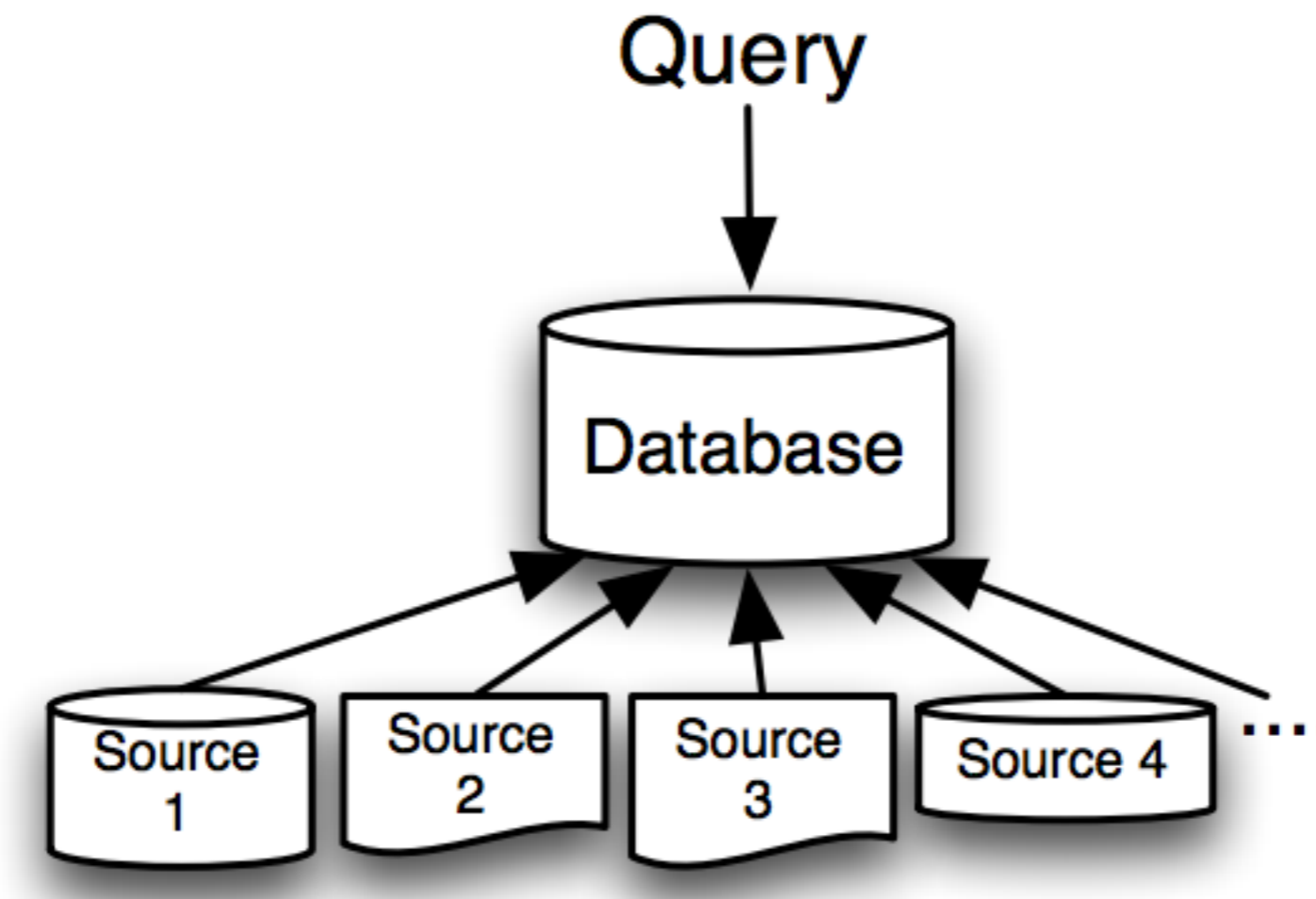
 **Rakesh Agrawal**  
Microsoft
Publications: 353 | Citations: 33537
Fields: [Databases](#), [Data Mining](#), [World Wide Web](#) 
Collaborated with [365 co-authors](#) from 1982 to 2012 | Cited by [24220 authors](#)

 **Jeffrey D. Ullman**  
Stanford University
Publications: 460 | Citations: 43431
Fields: [Databases](#), [Algorithms & Theory](#), [Scientific Computing](#) 
Collaborated with [317 co-authors](#) from 1961 to 2012 | Cited by [31987 authors](#)

 **Michael Franklin**  
University of California Berkeley
Publications: 561 | Citations: 15174
Fields: [Databases](#), [Pharmacology](#), [Data Mining](#) 
Collaborated with [3451 co-authors](#) from 1974 to 2012 | Cited by [15795 authors](#)

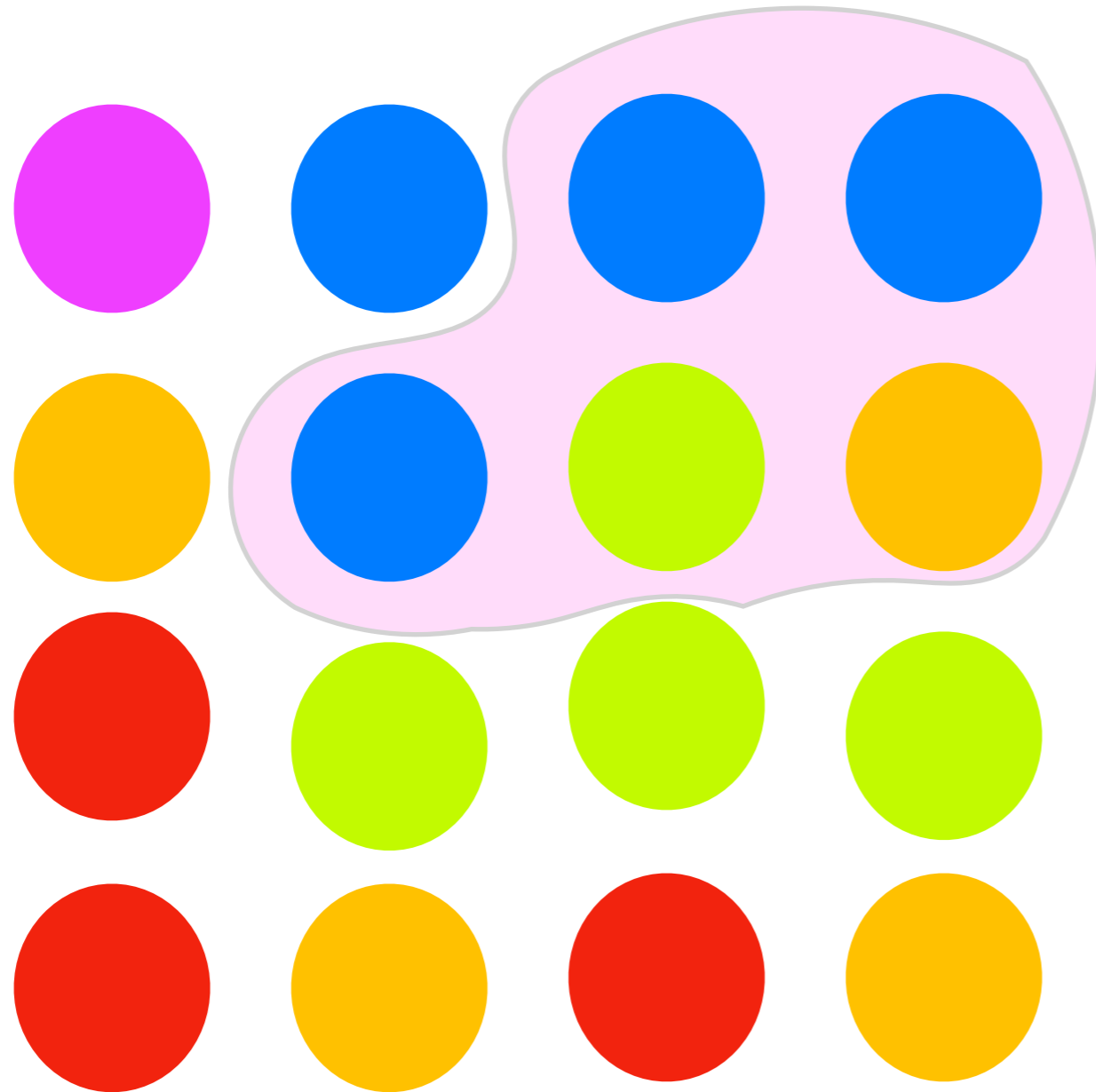


Estimating Unknown Unknowns

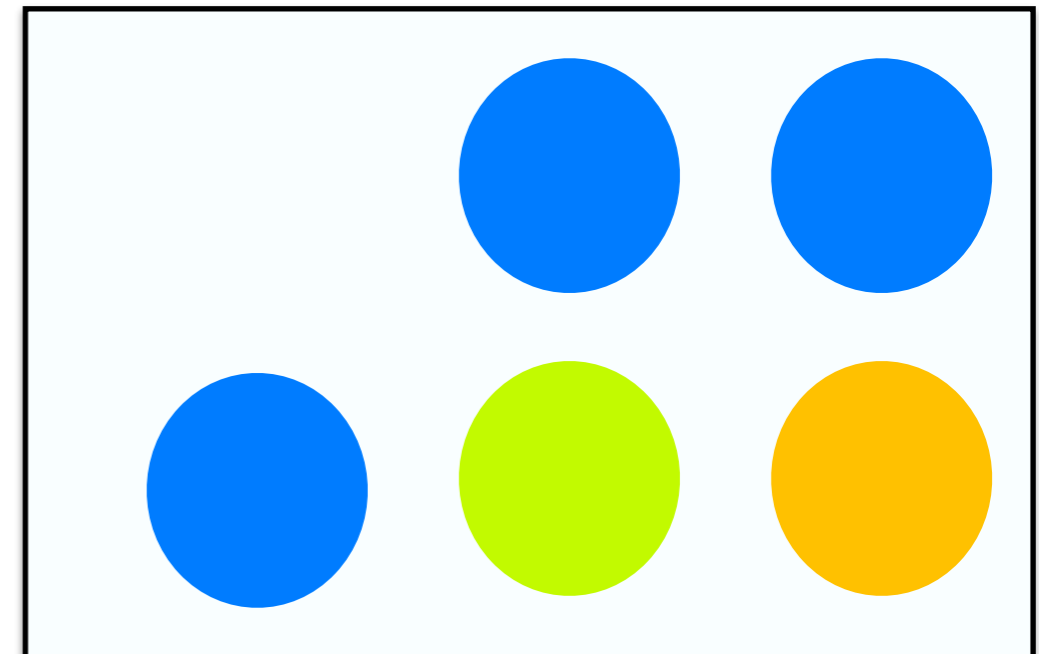


Chung, Y., Mortensen, M.L., Binnig, C. and Kraska, T., Estimating the impact of unknown unknowns on aggregate query results. SIGMOD 2016.

Abstract Problem



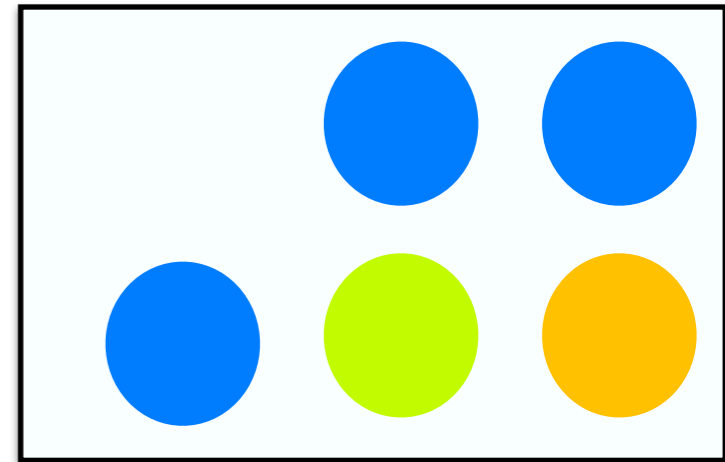
What is the distinct count?



Trushkowsky, Beth, et al. "Crowdsourced enumeration queries." Data Engineering (ICDE), 2013 IEEE 29th International Conference on. IEEE, 2013.

Estimates

- Nominal estimate: Observed
 - 3
- Naive estimate: $\text{Observed} * 1/\text{sample}$
 - $16/5 * 3 = 9.6$
- Good-Turing Estimate: $\text{Observed}/(1-f_1/n)$
 - $3 / (1-3/5) = 5$



Related to Species Estimation

- **Step 1.** Estimate the number of distinct entities given a sample
 - Good Turning Estimate: $1 - f_1/n$
- **Step 2.** Use the estimated “missing entities” to estimate the impact on a query result
 - Bucket the data to understand the correlation between frequency and value
- **Step 3.** Correct Query Result

Sensor Networks and Streams

- Online Filtering, Smoothing and Probabilistic Modeling of Streaming data
 - Uses particle filters to model uncertain data
- Declarative support for sensor data cleaning
 - Smoothing operators, filtering, outlier detection

Kanagal, Bhargav, and Amol Deshpande. "Online filtering, smoothing and probabilistic modeling of streaming data." 2008 IEEE 24th International Conference on Data Engineering. IEEE, 2008.

Jeffery, Shawn R., et al. "Declarative support for sensor data cleaning." International Conference on Pervasive Computing. Springer Berlin Heidelberg, 2006.

Section Structure

- Extended Data Cleaning Definition
- Connecting Data Cleaning to Downstream Queries
 - Aggregate queries
 - **Machine learning training**
 - Exploiting Relational Information

Section Structure

- Extended Data Cleaning Definition
- Connecting Data Cleaning to Downstream Queries
 - Aggregate queries
 - **Machine learning training**
 - Exploiting Relational Information

Example

Cluster Publications From Rakesh and Mike



Rakesh Agrawal



Microsoft

Publications: 353 | Citations: 33537

Fields: [Databases](#), [Data Mining](#), [World Wide Web](#)

Collaborated with [365 co-authors](#) from 1982 to 2012 | Cited by [24220 authors](#)



Jeffrey D. Ullman



Stanford University

Publications: 460 | Citations: 43431

Fields: [Databases](#), [Algorithms & Theory](#), [Scientific Computing](#)

Collaborated with [317 co-authors](#) from 1961 to 2012 | Cited by [31987 authors](#)



Michael Franklin

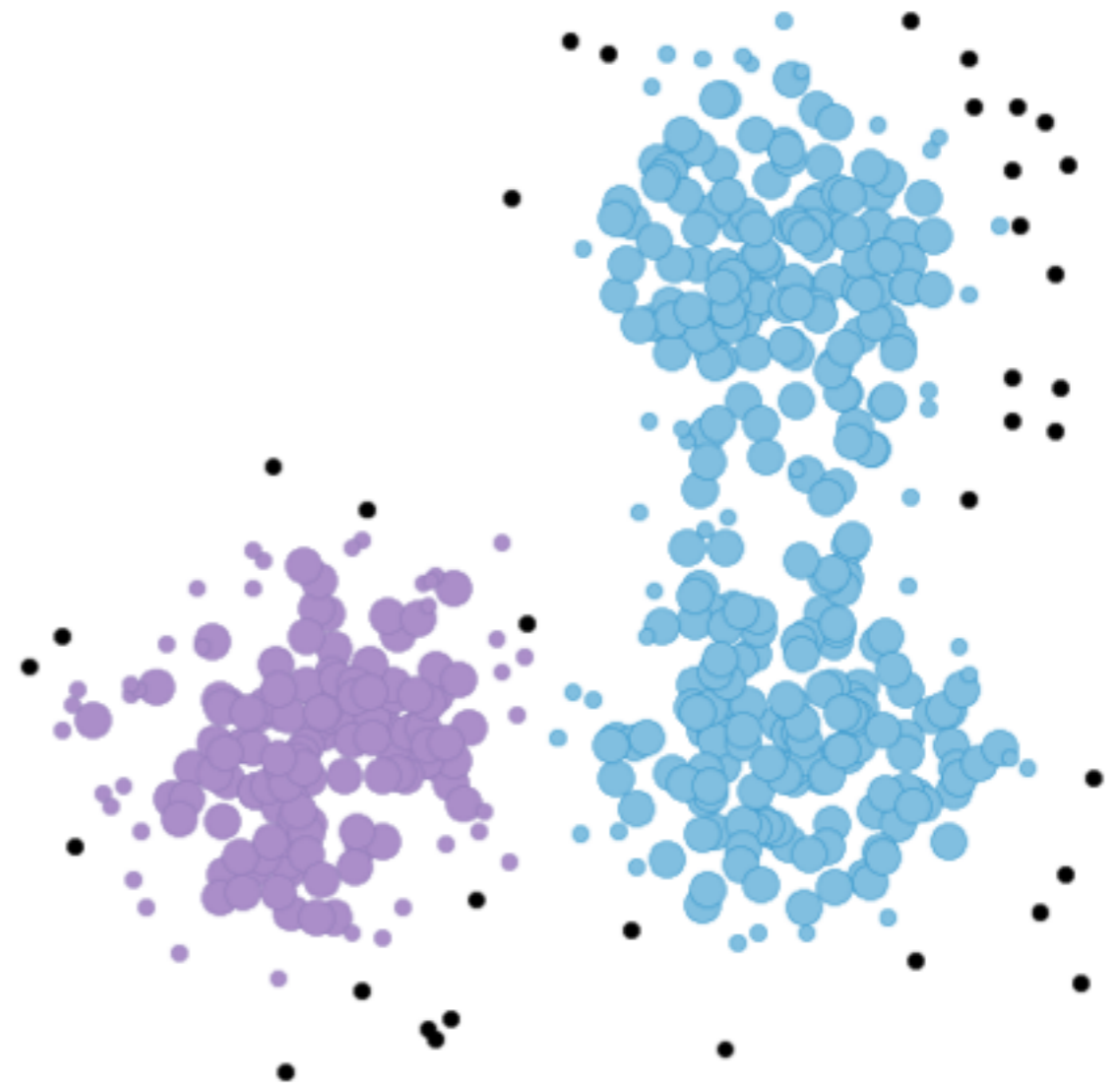


University of California Berkeley

Publications: 561 | Citations: 15174

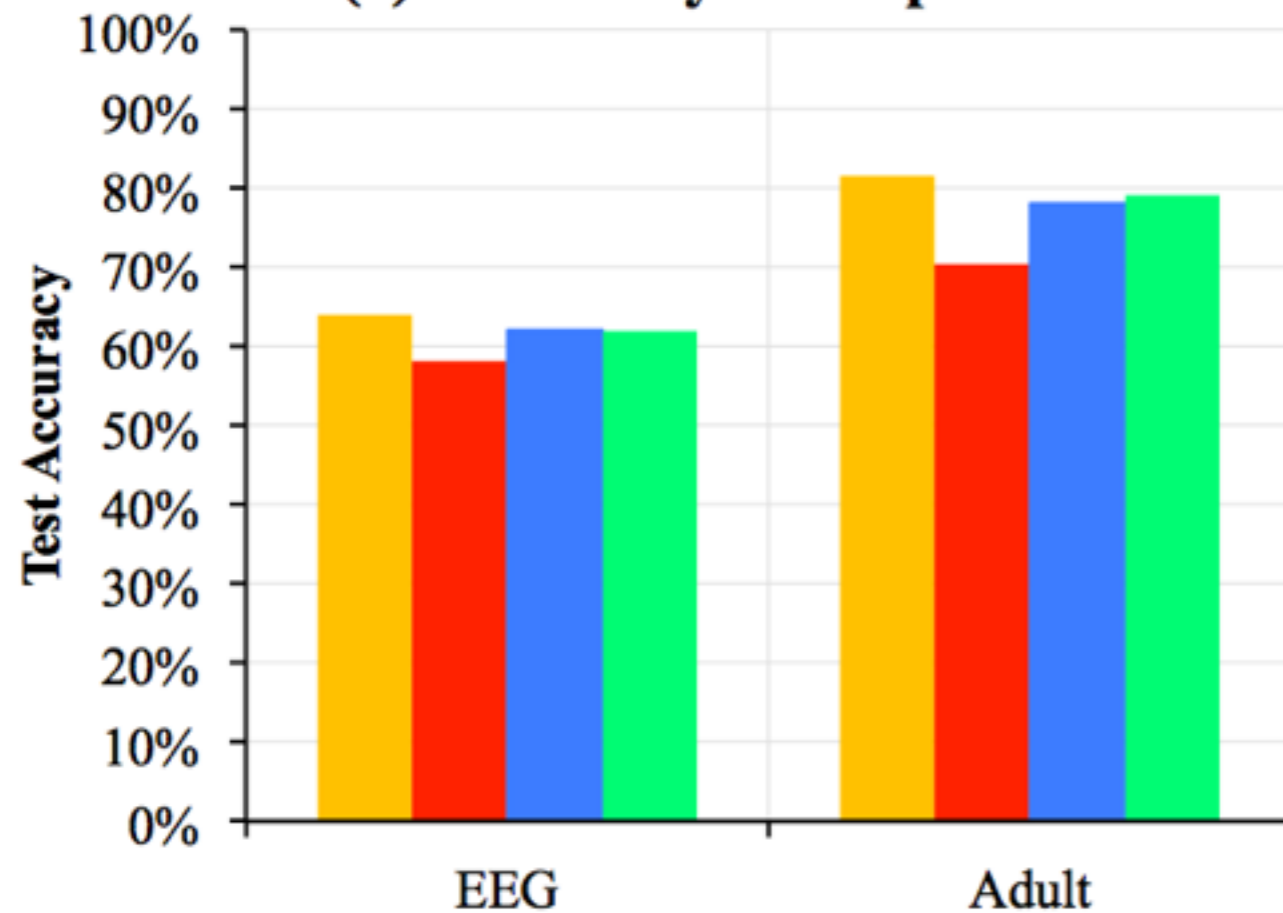
Fields: [Databases](#), [Pharmacology](#), [Data Mining](#)

Collaborated with [3451 co-authors](#) from 1974 to 2012 | Cited by [15795 authors](#)

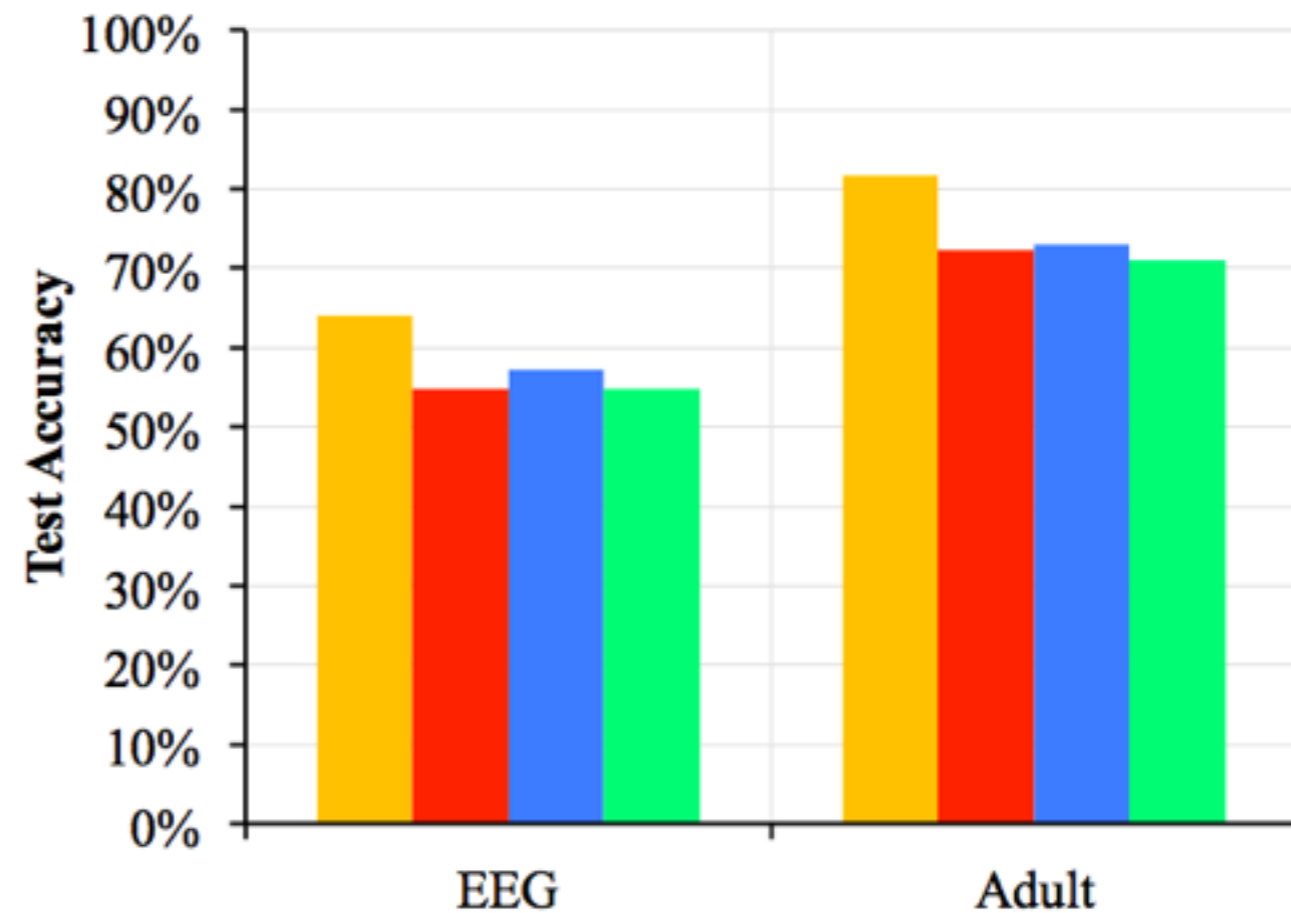


Misconception 1: ML models are robust to error

(a) Randomly Corrupted Data

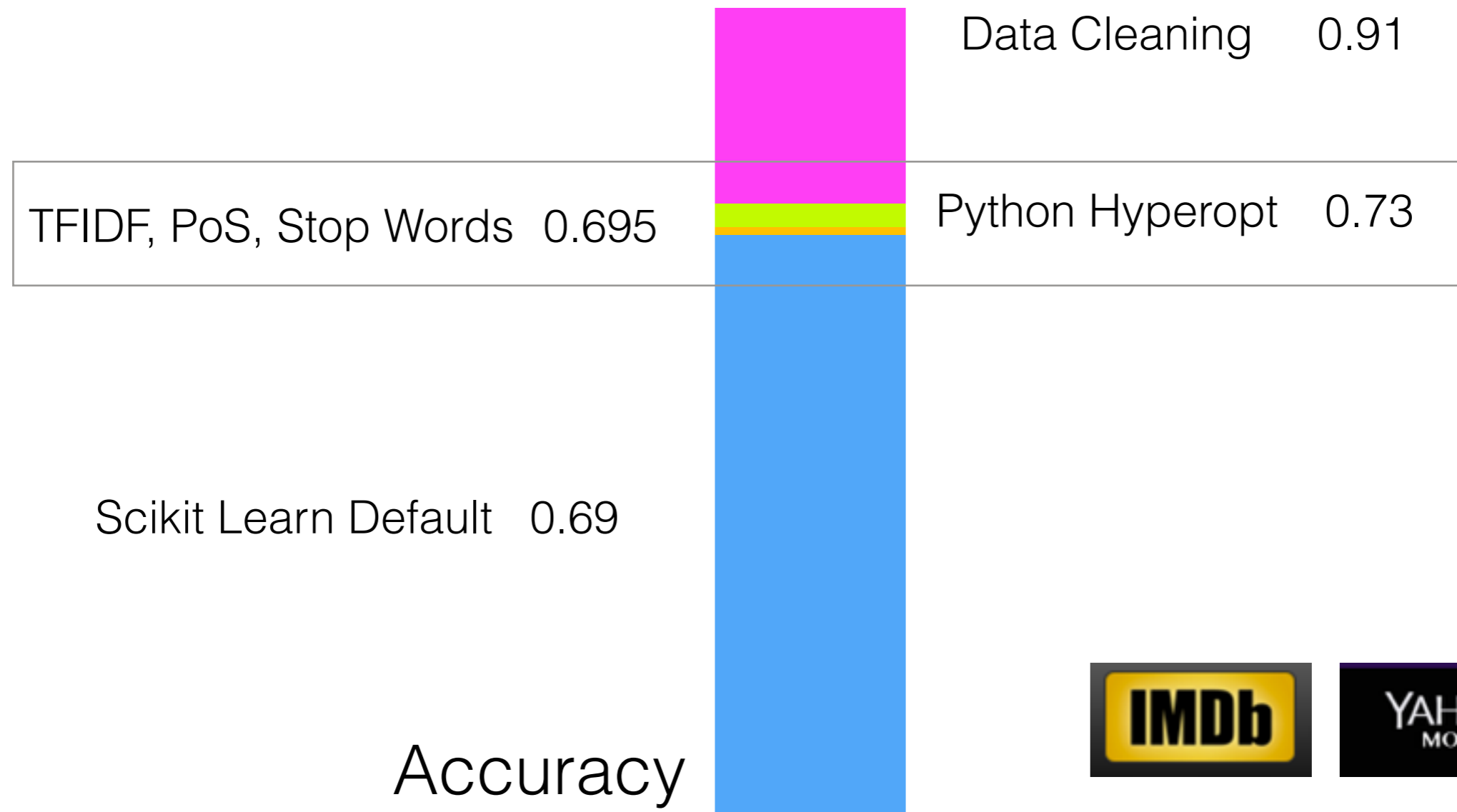


(b) Systematically Corrupted Data



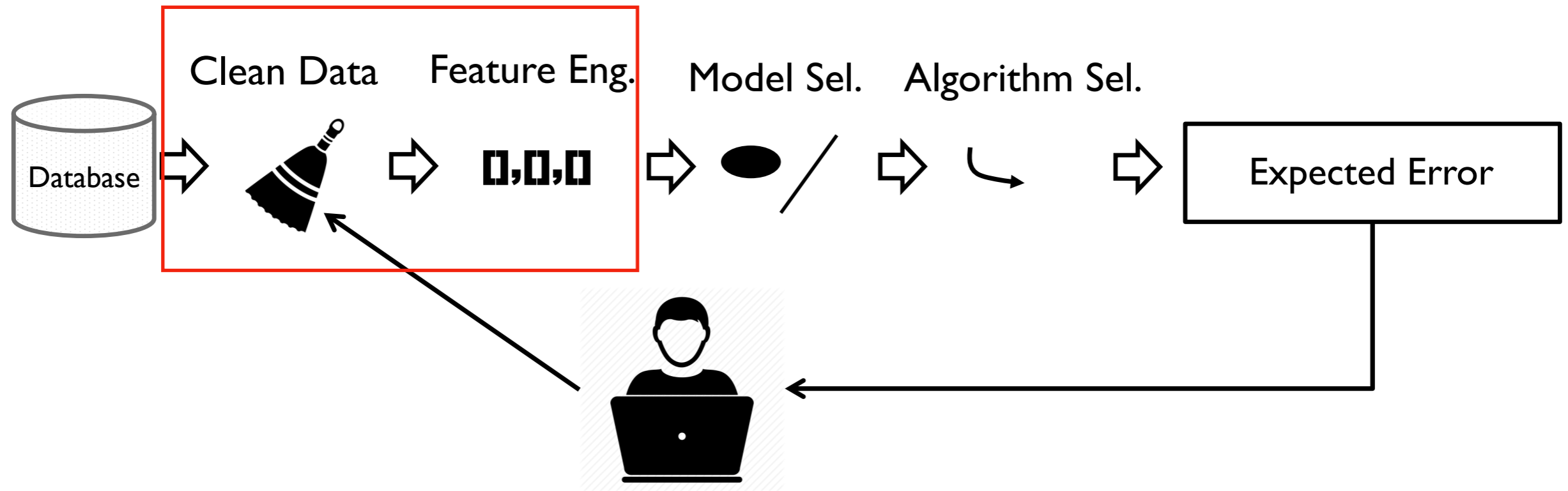
■ Clean ■ Dirty ■ Discard ■ Robust

Misconception 2. Parameter tuning is the most important problem



Horror vs. Comedy From Plot

Data Cleaning Before Machine Learning

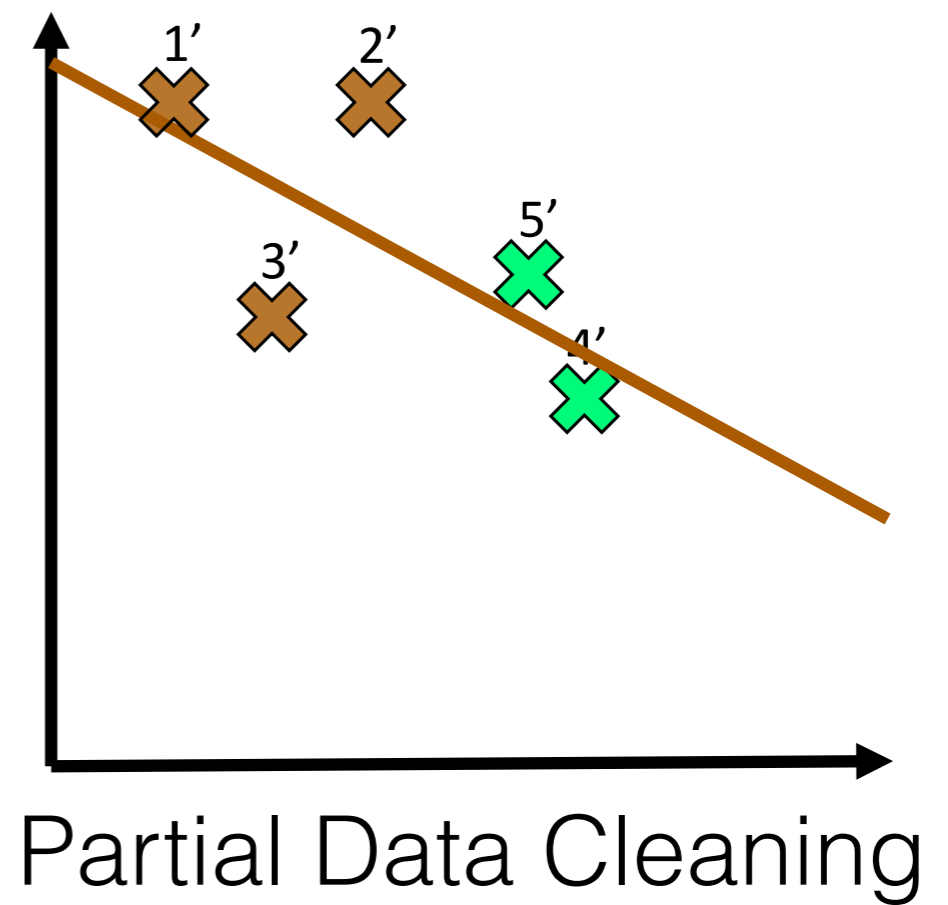
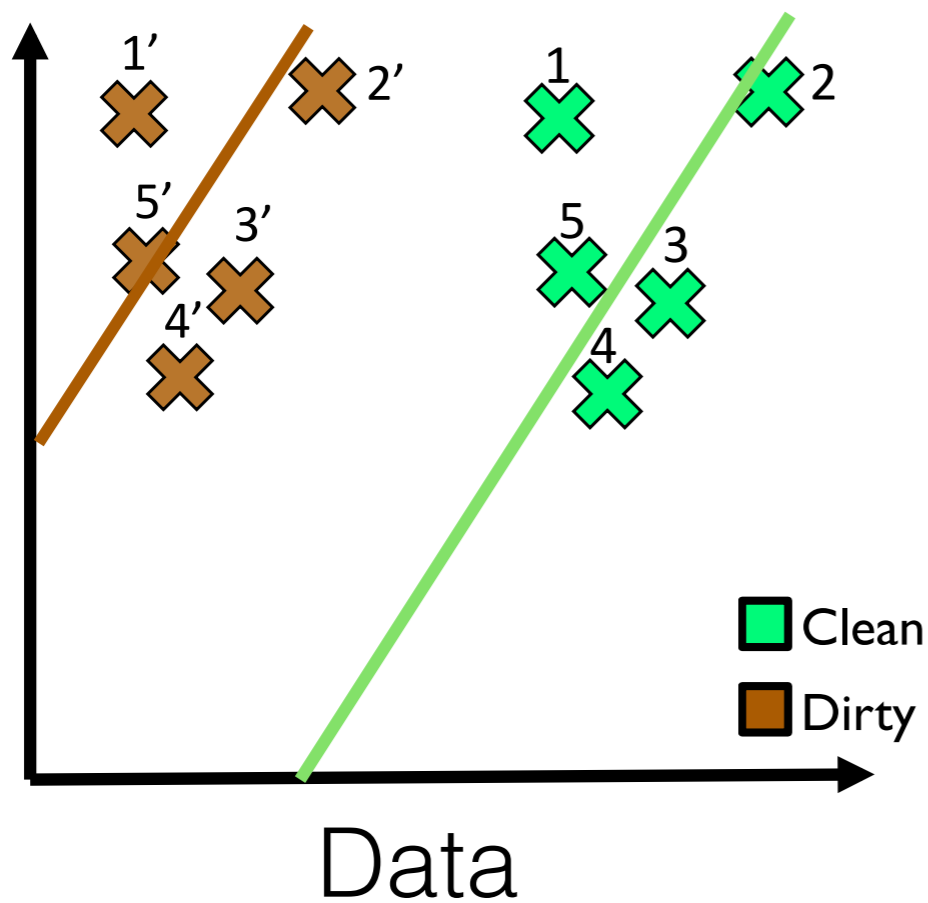


Correctness: Does data cleaning affect the convergence?

Efficiency: How to use the model to identify dirty data?

Krishnan, Sanjay, et al. "Activeclean: Interactive data cleaning while learning convex loss models." VLDB 2016.

Simpson's Paradox

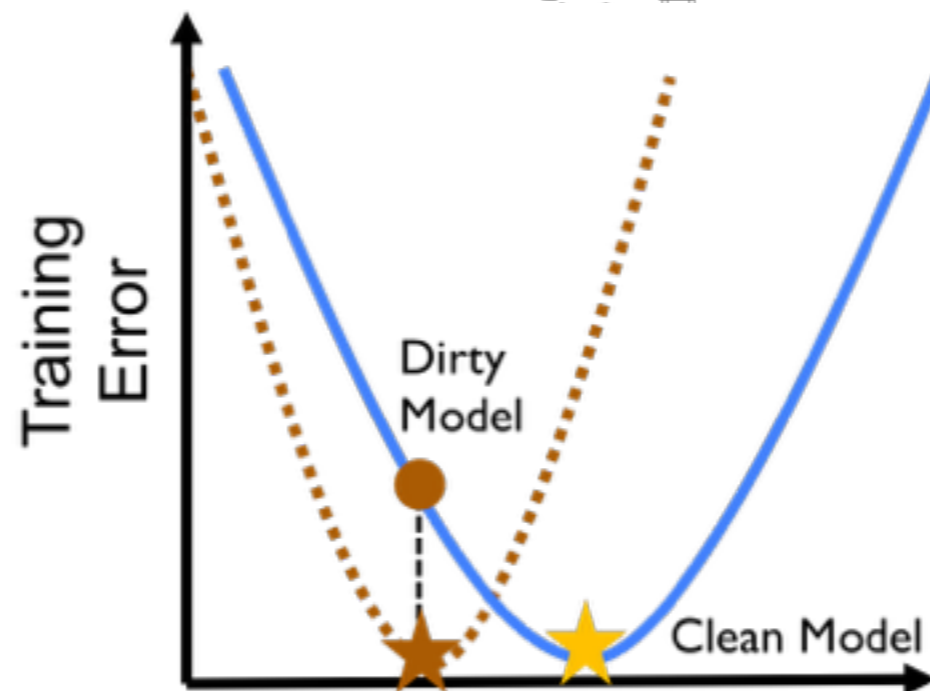


Partial Data Cleaning Can Be Misleading

Intuition

- Many ML problems can be represented as convex-loss minimization:

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^N \phi(x_i, y_i, \theta)$$



Solved via Stochastic Optimizations

- Stochastic Gradient Descent.

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma \cdot \underline{E[\nabla \phi(\theta^{(t)})]}$$

- Just an estimated average query!
- Make each step unbiased

Active Clean Algorithm

- Train a preliminary model on the dirty dataset
- For i in $\{0, \dots, T\}$
 - **Sample** a batch of data
 - **Clean** the sample
 - **Update** the model via gradient descent (**reweight**)
- **Return** model

ActiveClean Analysis

For a batch size b and iterations T , the ActiveClean stochastic gradient descent updates converge (i.e., reduce the error in a model trained on dirty data) with rate:

$$O\left(\frac{1}{\sqrt{bT}}\right)$$

For strongly-convex models (e.g., full rank linear regression):

$$O\left(\frac{1}{T\sqrt{b}}\right)$$

For L -Lipschitz loss (e.g., SVM):

$$O\left(\frac{L}{\sqrt{bT}}\right)$$



Section Structure

- Extended Data Cleaning Definition
- Connecting Data Cleaning to Downstream Queries
 - Aggregate queries
 - Machine learning training
 - **Exploiting Relational Information**

Example

- Select all papers from Rakesh Agrawal after 2000



Rakesh Agrawal  



[Microsoft](#)

Publications: [353](#) | Citations: [33537](#)

Fields: [Databases](#), [Data Mining](#), [World Wide Web](#) 

Collaborated with [365 co-authors](#) from 1982 to 2012 | Cited by [24220 authors](#)



Jeffrey D. Ullman  



[Stanford University](#)

Publications: [460](#) | Citations: [43431](#)

Fields: [Databases](#), [Algorithms & Theory](#), [Scientific Computing](#) 

Collaborated with [317 co-authors](#) from 1961 to 2012 | Cited by [31987 authors](#)



Michael Franklin  

[University of California Berkeley](#)

Publications: [561](#) | Citations: [15174](#)

Fields: [Databases](#), [Pharmacology](#), [Data Mining](#) 

Collaborated with [3451 co-authors](#) from 1974 to 2012 | Cited by [15795 authors](#)

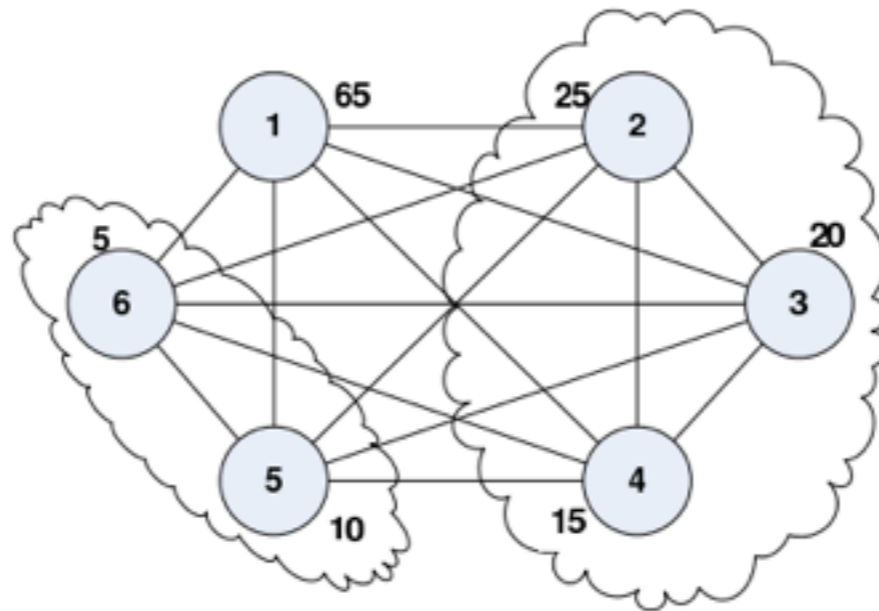
Query-Driven Entity Resolution

- Do minimal ER work to answer a query
- Defines a concept of *vestigiality* (answer is still correct without knowing whether a pair is a duplicate)
- Evaluates SQL queries aware of vestigial relationships.

Altwaijry, Hotham, Dmitri V. Kalashnikov, and Sharad Mehrotra. "Query-driven approach to entity resolution." Proceedings of the VLDB Endowment 6.14 (2013): 1846-1857.

Query-Driven Entity Resolution

- Table is represented as a weighted graph of possibly duplicated entities





- Define rules that preserve predicates

Example

- Select all papers from Rakesh Agrawal after 2000
- A priori algorithm paper missing



Rakesh Agrawal  



Microsoft

Publications: [353](#) | Citations: [33537](#)

Fields: [Databases](#), [Data Mining](#), [World Wide Web](#) 

Collaborated with [365 co-authors](#) from 1982 to 2012 | Cited by [24220 authors](#)



Jeffrey D. Ullman  



Stanford University

Publications: [460](#) | Citations: [43431](#)

Fields: [Databases](#), [Algorithms & Theory](#), [Scientific Computing](#) 

Collaborated with [317 co-authors](#) from 1961 to 2012 | Cited by [31987 authors](#)



Michael Franklin  

University of California Berkeley

Publications: [561](#) | Citations: [15174](#)

Fields: [Databases](#), [Pharmacology](#), [Data Mining](#) 

Collaborated with [3451 co-authors](#) from 1974 to 2012 | Cited by [15795 authors](#)

Query-Oriented Data Cleaning

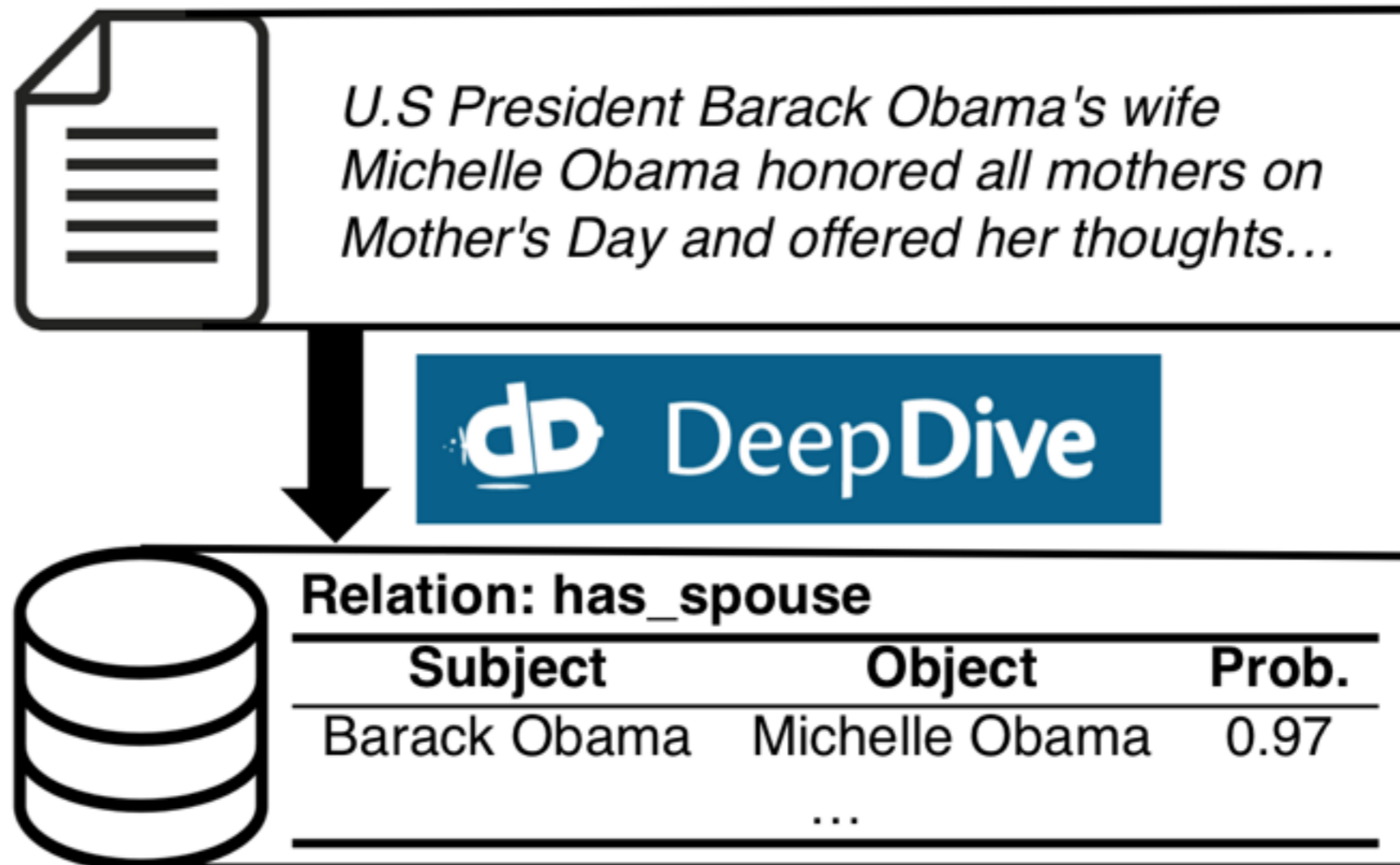
- Oracle crowds to derive database edits for removing (adding) incorrect (missing) tuples to the result of a query.
- For a given query derive a set of cleaning updates to base data to ensure completeness.

Bergman, Moria, et al. "Query-oriented data cleaning with oracles." Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM, 2015.

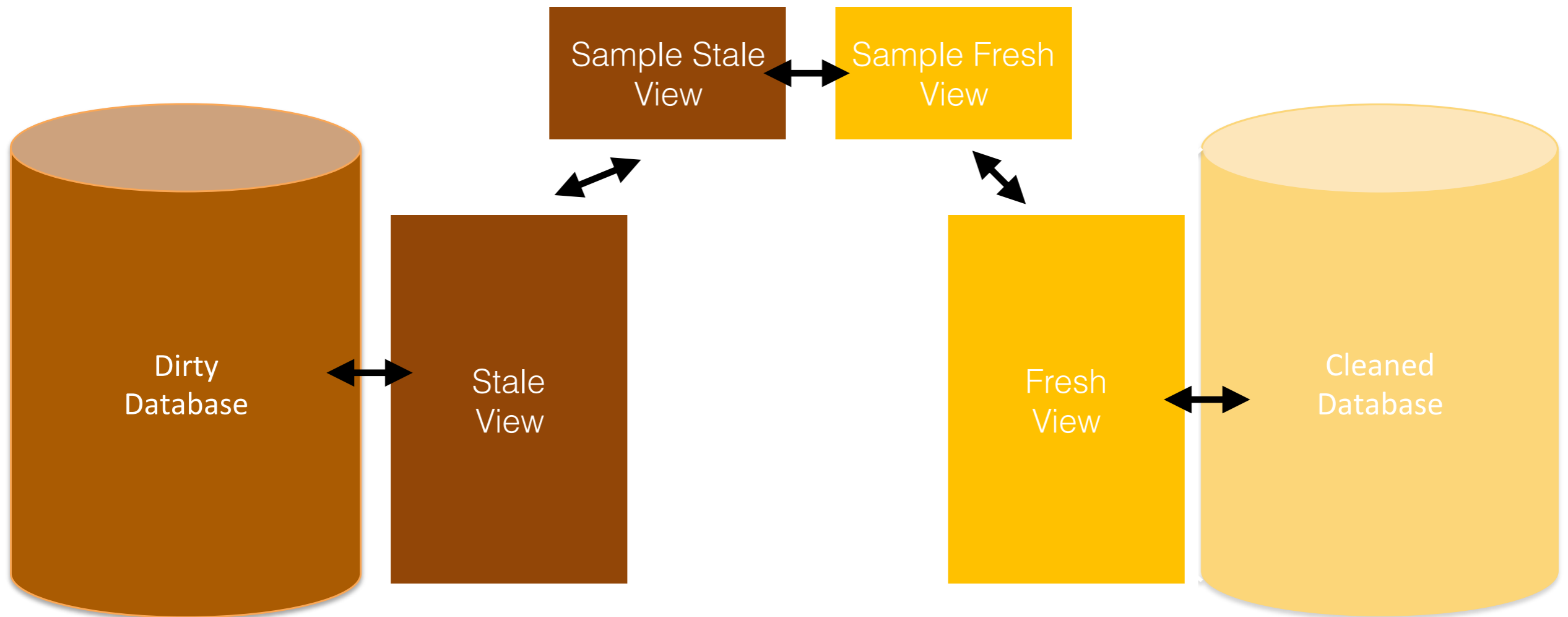
A Statistical Perspective

- Topic 1. Statistical techniques to clean data (20 mins)
- Topic 2. Cleaning data before statistical analytics (50 min)
- **Topic 3. Impact and Future Directions (10 mins)**

Data Cleaning+ Knowledge Bases



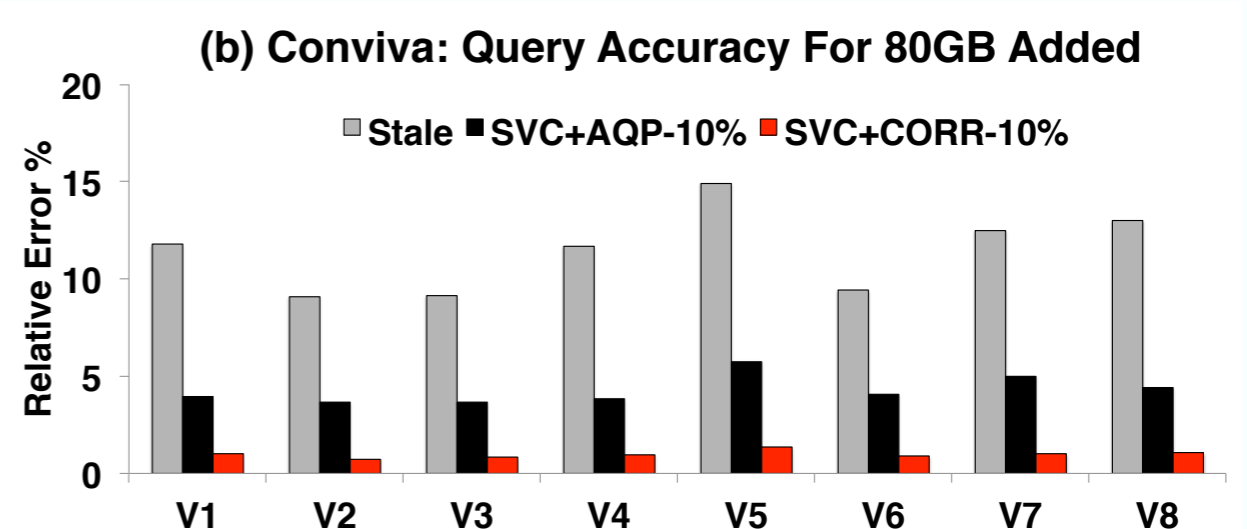
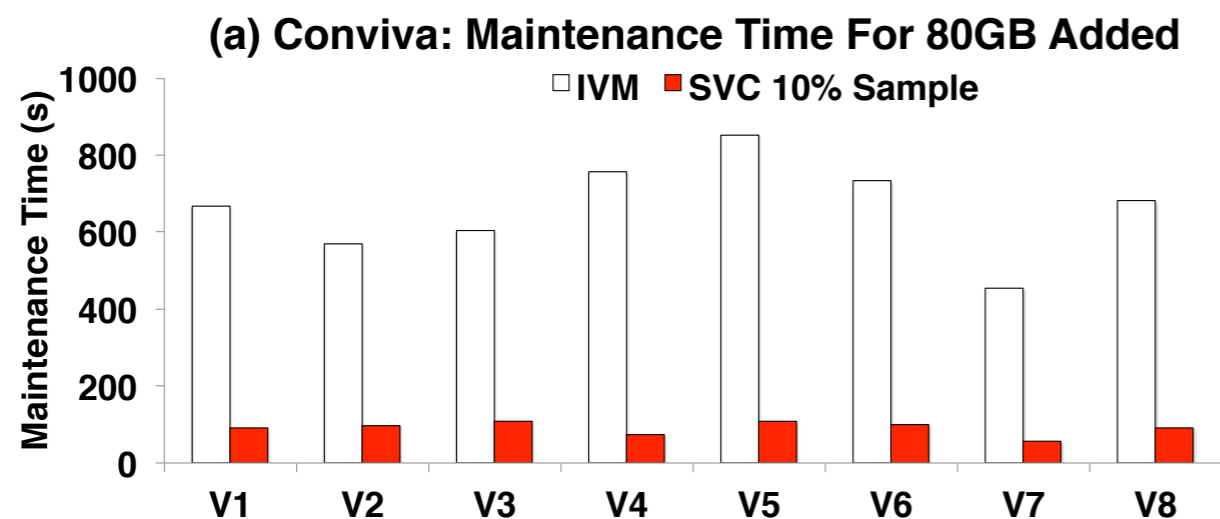
Queries on Stale Views



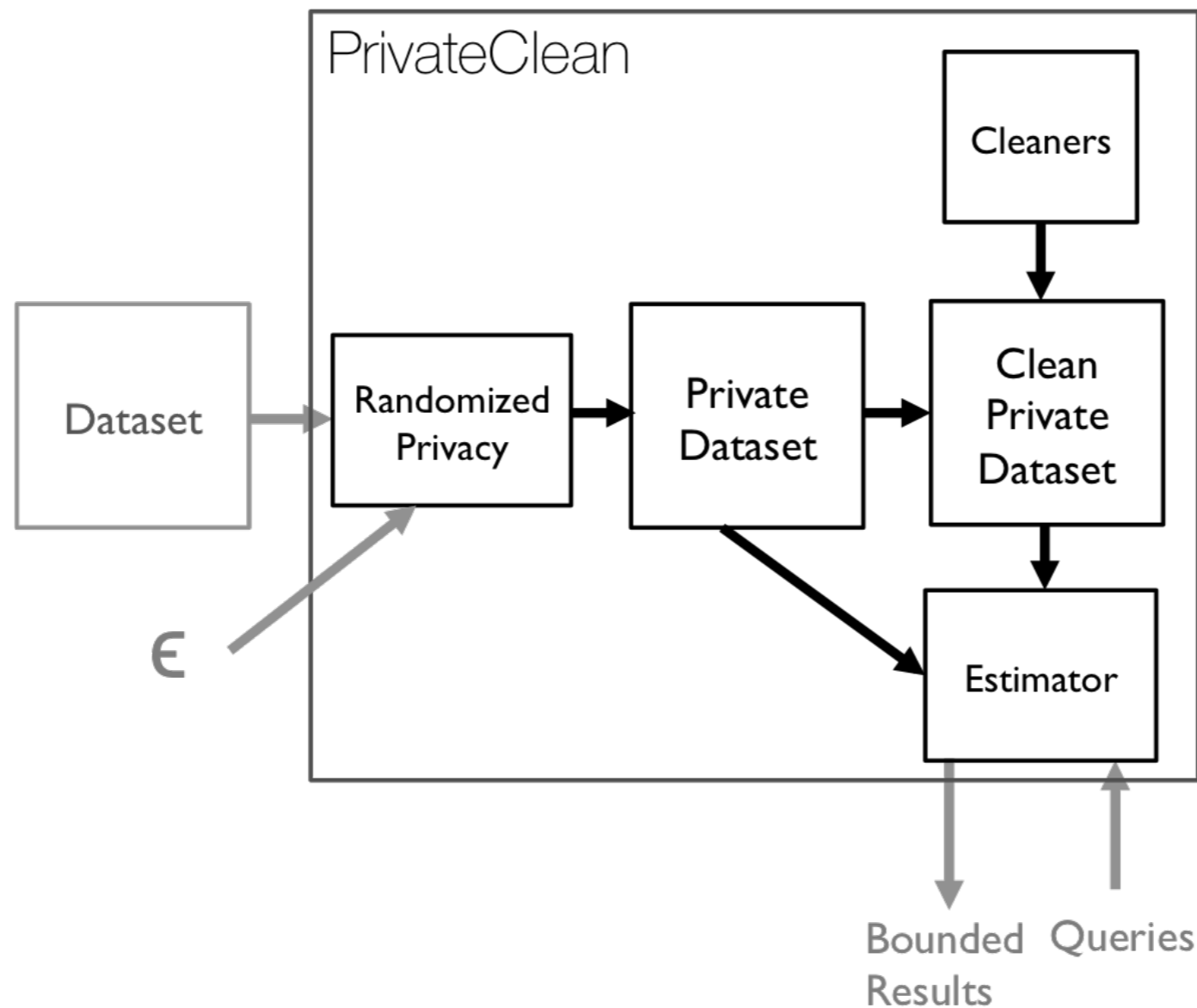
Krishnan, Sanjay, et al. "Stale view cleaning: Getting fresh answers from stale materialized views." Proceedings of the VLDB Endowment 8.12 (2015):

Conviva: Log Analysis

- Implemented on a 20-node Apache Spark Cluster
- Applied SVC to common reporting queries from an SQL trace.
- Experiment 2. 10% (80GB) Base Data Updates



Data Cleaning+Privacy



Sanjay Krishnan, et al. *PrivateClean: Data Cleaning and Differential Privacy*. SIGMOD 2016.

Wed 10:30

Open Problems

- Reproducibility of data cleaning
 - Statistical reliability of the conclusions drawn
- When to automate and when to use humans
 - Leverage Improvements in ML in data cleaning but need reliability
- Benchmarking and Evaluation

Arocena, Patricia C., et al. "Messing up with BART: error generation for evaluating data-cleaning algorithms." Proceedings of the VLDB Endowment 9.2 (2015): 36-47.

Köpcke, Hanna, Andreas Thor, and Erhard Rahm. "Evaluation of entity resolution approaches on real-world match problems." Proceedings of the VLDB Endowment 3.1-2 (2010): 484-493.

Reproducibility




"We test thousands of new treatments each year, so to avoid multiple testing issues we always do a validation experiment to confirm our positive results".

How often do those work out?

About 5% of the time!

Concerns

- Multiple Hypothesis Testing
- Adaptive Hypothesis Testing
- **What about after data cleaning?**

	smallest					largest
p-values	$P_{(1)}$	$P_{(2)}$	$P_{(3)}$...	$P_{(m)}$	
k	1	2	3	...	m	
threshold	$\frac{\alpha^*}{m}$	$\frac{2\alpha^*}{m}$	$\frac{3\alpha^*}{m}$...	α^*	

Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)* (1995): 289-300.

Conclusion

- A statistical perspective can enhance data cleaning models overcoming existing limitations.
- Leverages both statistics and database theory.
- Future data management is likely to have more problems in this area.

Two Complementary Views

