The Statistics of Dirty Data

Sanjay Krishnan

AUTOLAB
Exploring the Art of Automation and the Science of Robotics at UC Berkeley since 1995

riselab
UC Berkeley

Berkeley
Artificial Intelligence Research Laboratory

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

Data Scientist:

The Sexiest Job of the 21st Century

coax treasure out of messy, unstructured data

Meet the people who can coax treasure out of messy, unstructured data.

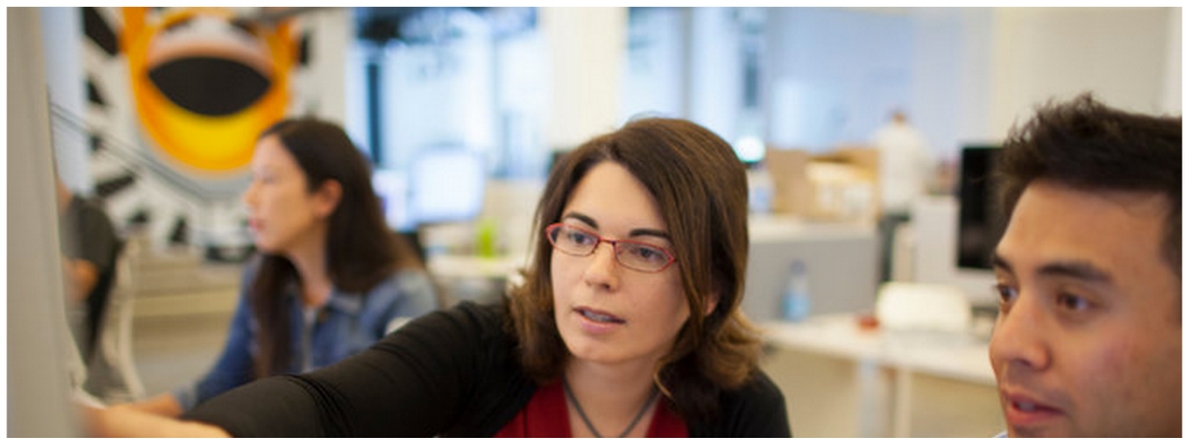
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

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

By **STEVE LOHR** AUG. 17, 2014

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

TRIFACTA

tamr

204 papers since 2012 in VLDB, ICDE, SIGMOD (**dirty data**)

The “Database” Perspective

- Dirty data is a violation of constraints on a table.
- Data Cleaning is constraint satisfaction

“No Manager Can Earn Less Than an Employee”

	Name	Role	Salary
1	Jane Doe	Emp	1700
2	John Smith	Manager	1500
3	Raj Kumar	Emp	1300
4	Maria Lopez	Manager	4400

The “Database” Perspective

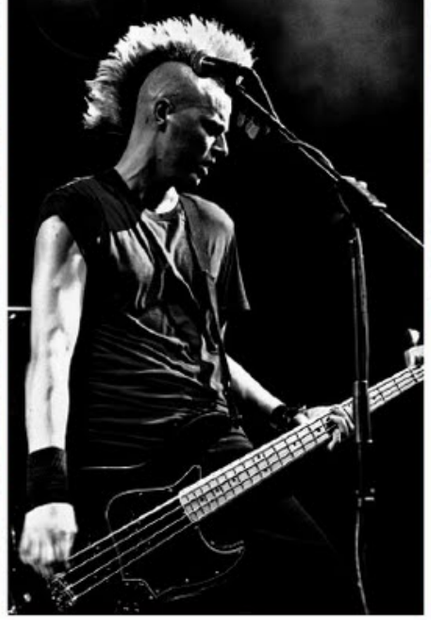
“No Manager Can Earn Less Than an Employee”

$$\forall (r, s) \in R : (r.role = \text{Manager}) \ \wedge (s.role = \text{'Emp'}) \ \wedge (s.salary < r.salary)$$

	Name	Role	Salary
1	Jane Doe	Emp	1700
2	John Smith	Manager	1500

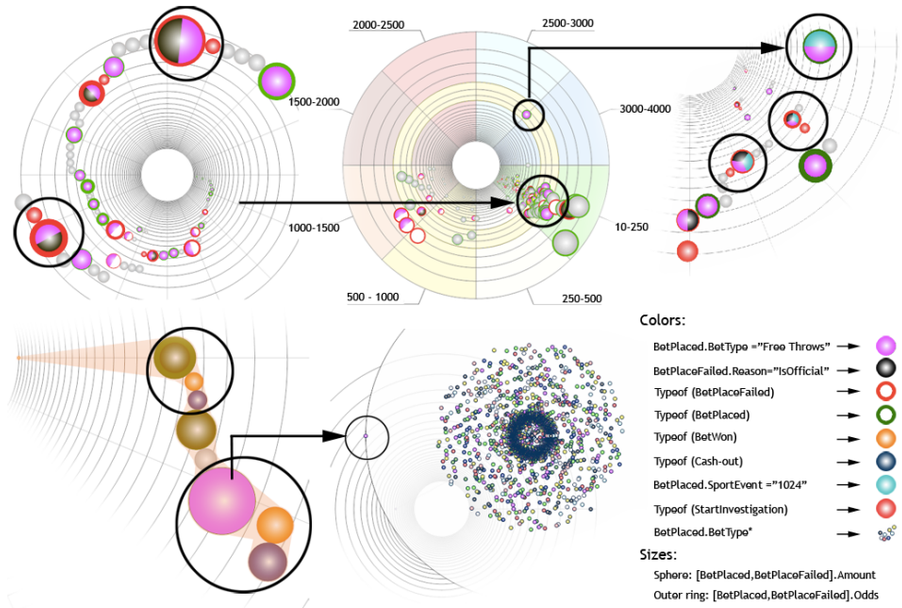
	Name	Role	Salary
1	Jane Doe	Emp	**1500**
2	John Smith	Manager	1500

	Name	Role	Salary
1	Jane Doe	**Manager**	1700
2	John Smith	Manager	1500



THE 1980S

We Interact With Data in Fundamentally New Ways



Spark

Pandas

R

DeepDive

KeystoneML

TensorFlow™

ML

The Statistics of Dirty Data

tl;dr Formalism Good, Theory Needs Updating

- **SampleClean: Linking Data Repair To Statistical Analysis.**
  - AlphaClean: Synthesizing Data Cleaning Programs With New AI Tools
  - Discussion

Motivating Example

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 345[illegible] 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

Results After Cleaning

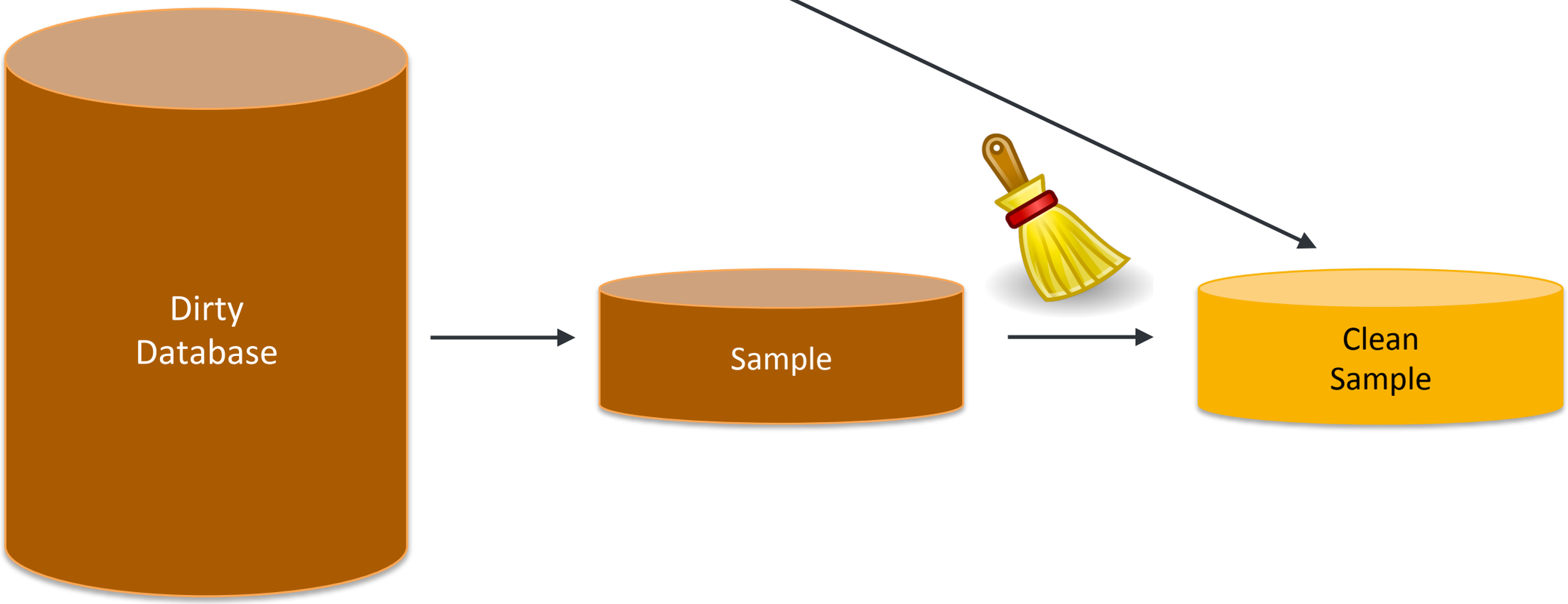
Author	Dirty	Clean
Rakesh Agarwal	353	211
Jeffrey Ullman	460	255
Michael Franklin	561	173

Did I need to clean everything?

Sample-and-Clean

SELECT COUNT(1)
FROM Pubs
GROUP BY Author

sampleclean



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

What goes wrong?

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.

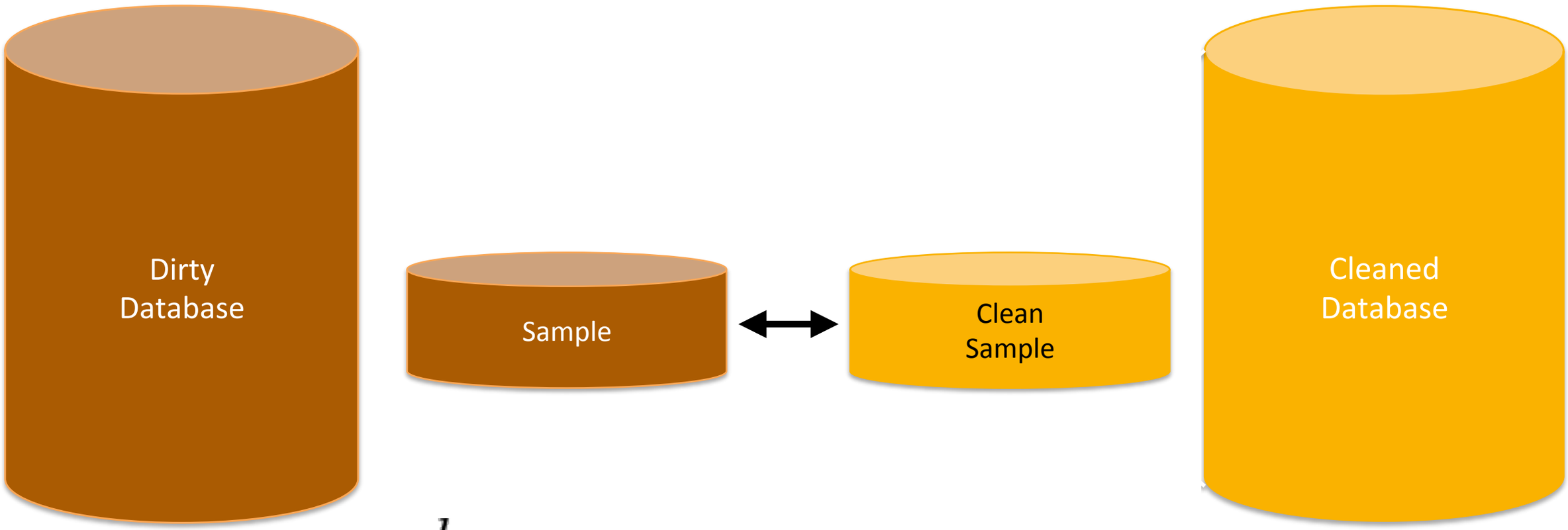
Duplicates!

Probabilistic Interpretation

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

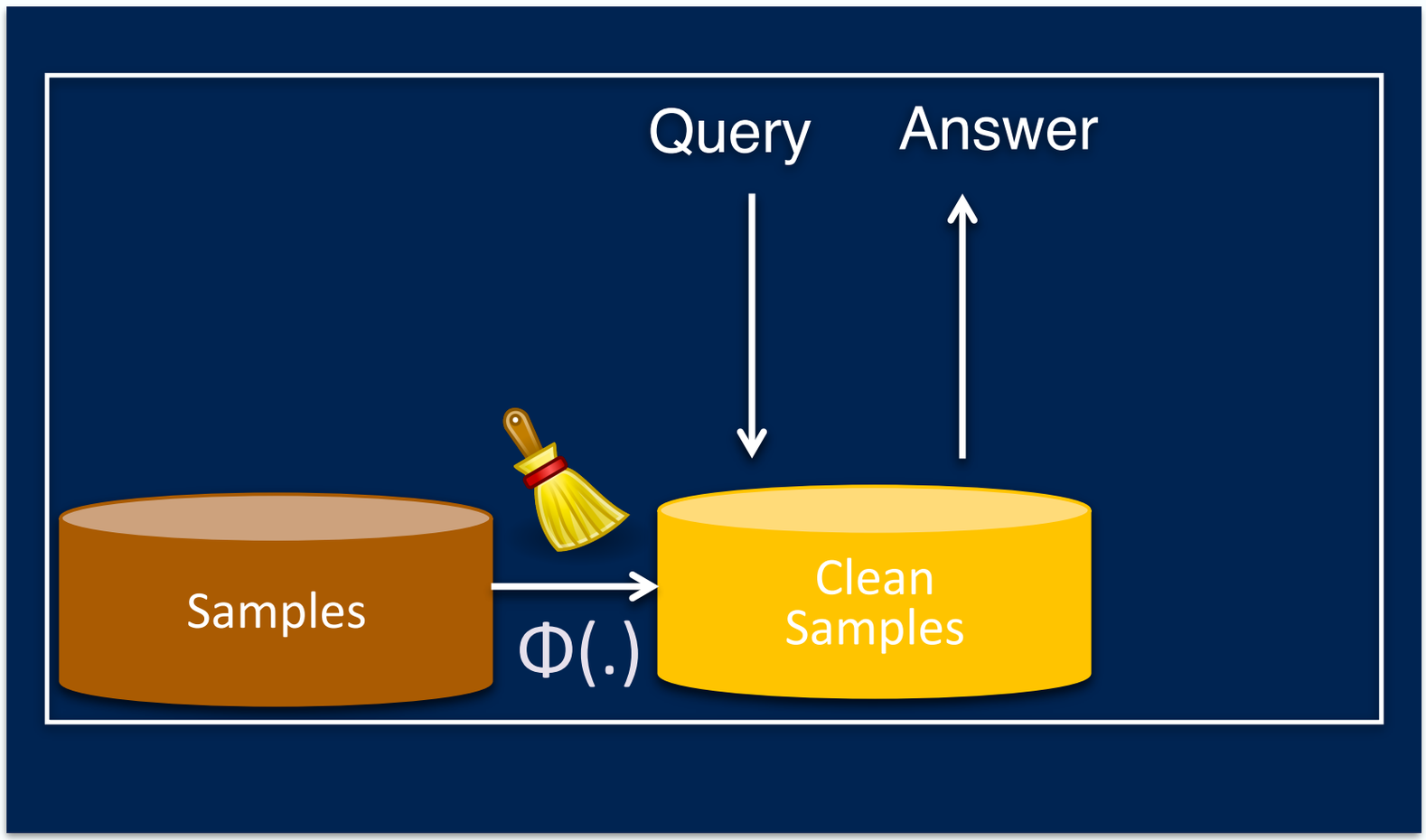
$$\mathbb{E}(X) = \sum x \cdot \mathbb{P}(X = x)$$

Transform Dirty Sample to Simulate Clean Sample



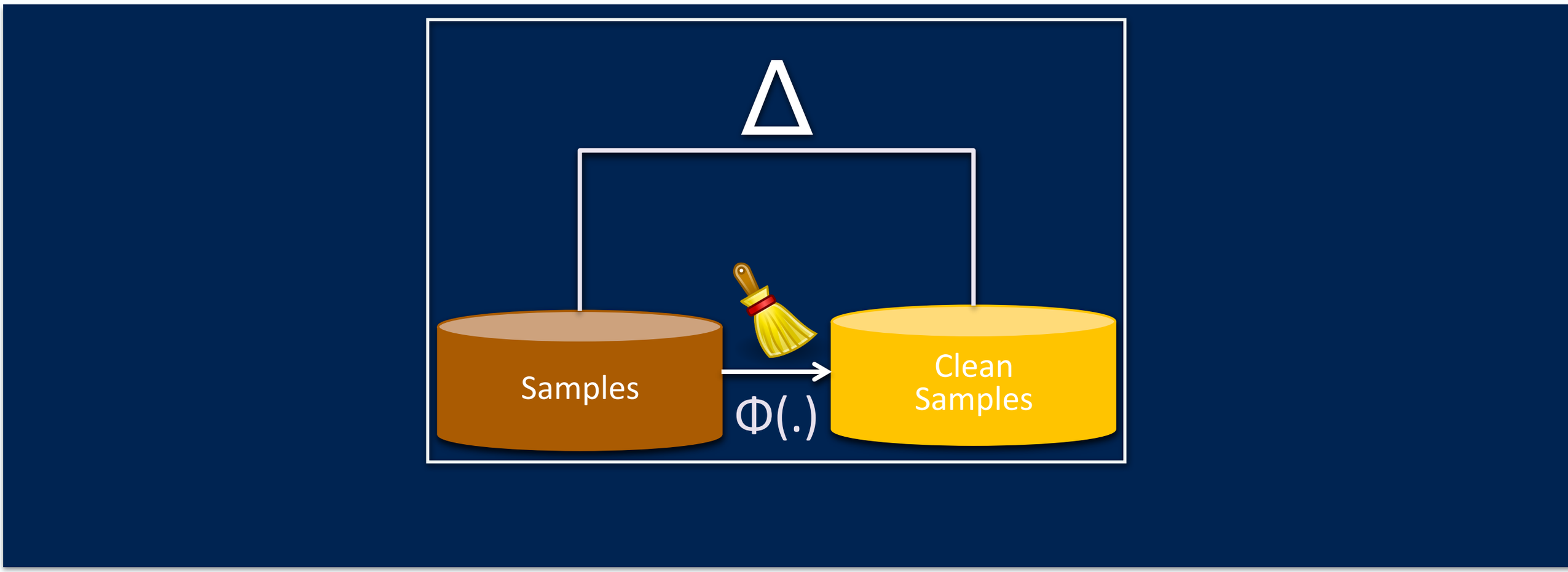
$$\bar{x} \propto \sum_{i=1}^{k} clean(x) \cdot \frac{predicate(x)}{dup(x)}$$

Algorithm 1: Direct Estimate

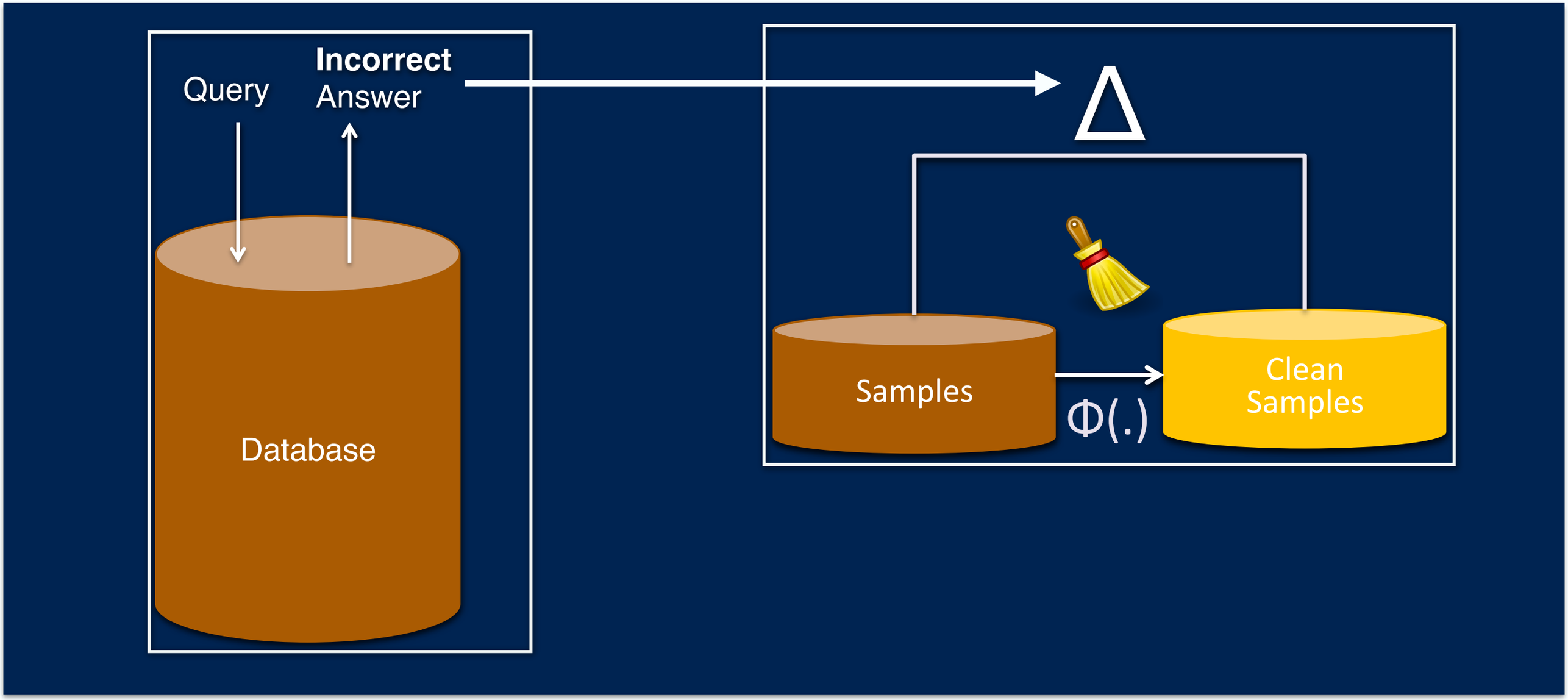


Algorithm 2: Corrected Estimate

How much did the cleaning change the data?



Algorithm 2: Corrected Estimate



Direct vs. Corrections

Let R be dirty relation and C(.) be a row-by-row cleaning function, and suppose, a user can call C(.) k<< |R| times. For SUM, COUNT, AVG, VAR queries with predicates, SampleClean provides a ***conditionally unbiased estimate*** *of the result with* ***asymptotic error*** *equal to:*

$$\pm \frac{z_\alpha \min\{\sigma_t, \sigma_{diff}\}}{\sqrt{k}}$$

The ***finite sample error*** *for query is given by:*

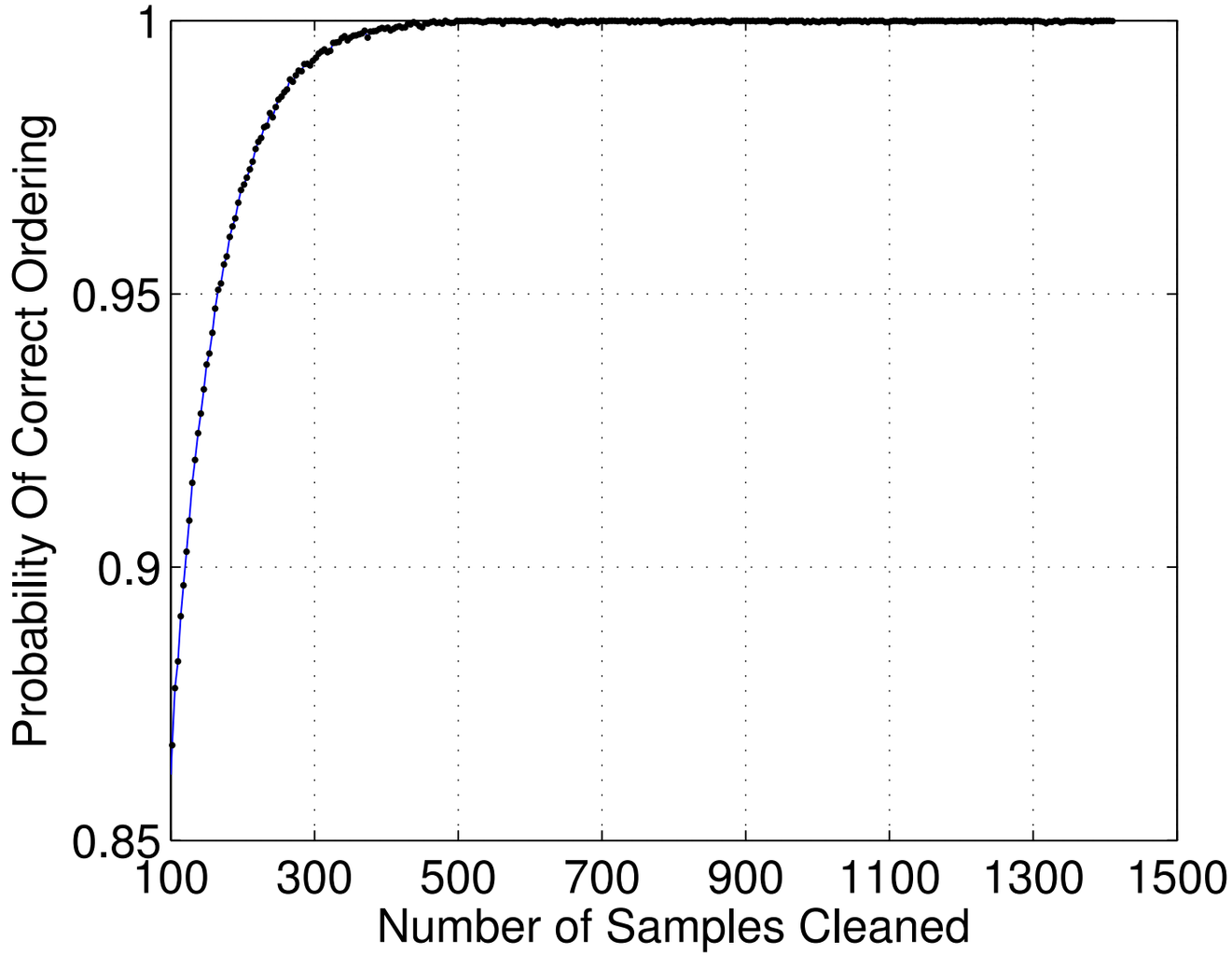
$$\pm \frac{\gamma_\alpha \min\{\Delta_{data}, \Delta_{clean}\}}{\sqrt{2k}}$$

MS Academic Results

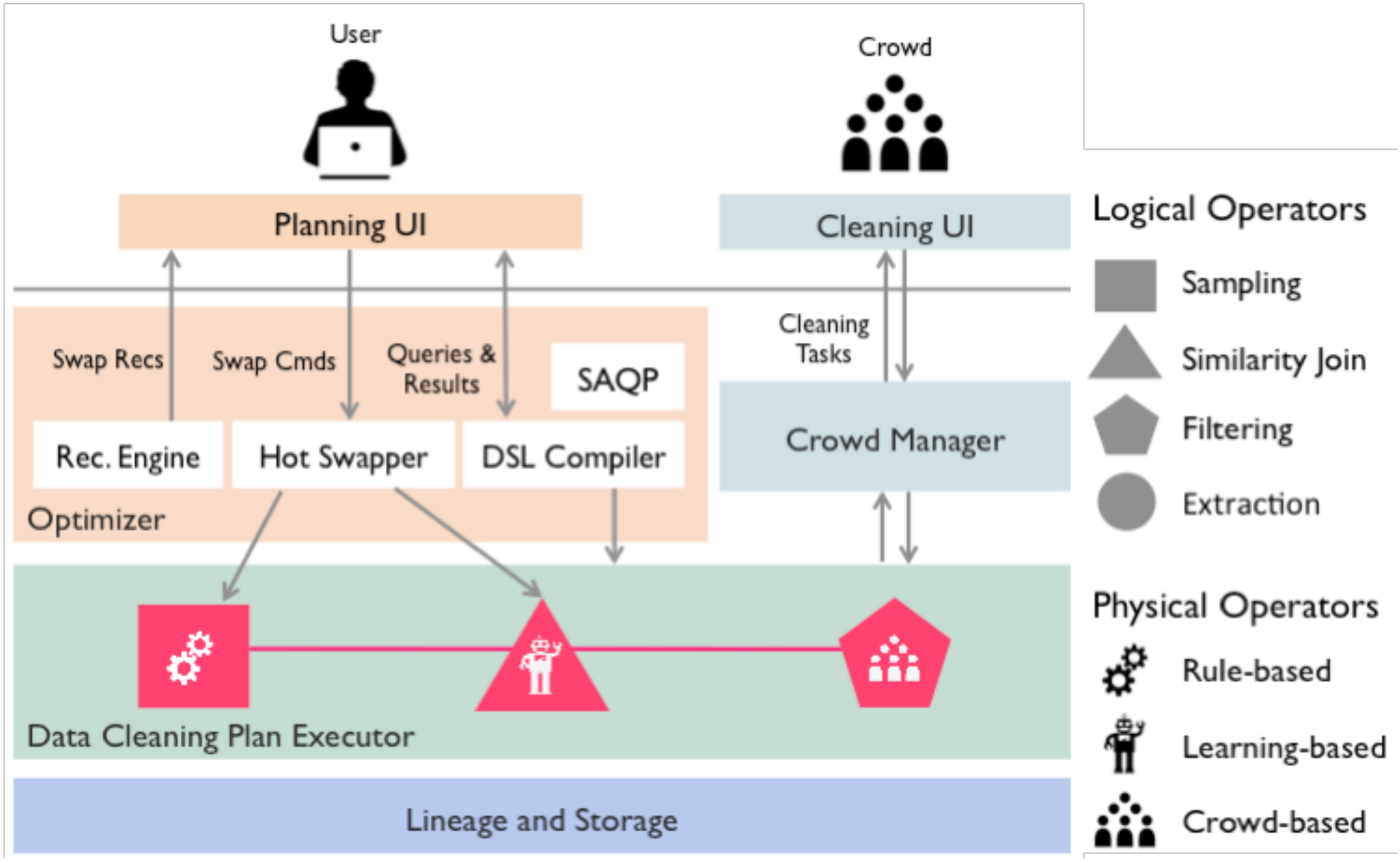
Rakesh Agrawal
Microsoft
Publications: 353 | Citations: 33537
Fields: Databases, Data Mining, World Wide Web
Collaborated with 365 co-authors from 1982 to 2012 | Cited by 24220 authors

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 345 co-authors from 1974 to 2012 | Cited by 15795 authors



Wisteria



Daniel Haas, **Sanjay Krishnan**, Jiannan Wang, Michael J. Franklin, Eugene Wu. Wisteria: Nurturing Scalable Data Cleaning Infrastructure. VLDB 2015.

Salient Pieces

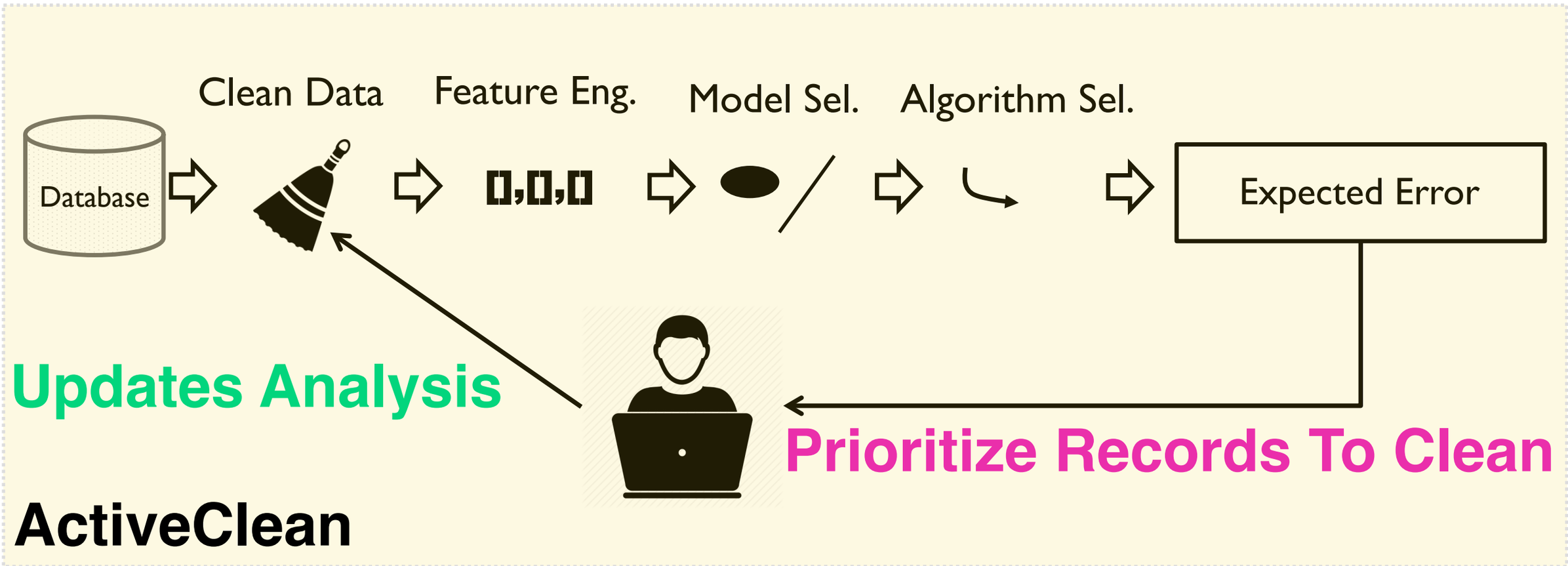
- 1. **Probability Measure** over the database: user sees the data under some type of observation model.

- 2. **Language** for data cleaning with estimable statistical properties.

- 3. **An aggregate query** to estimate after some adjustment of statistical changes.

ActiveClean

- Data Cleaning as a form of Stochastic Gradient Descent.

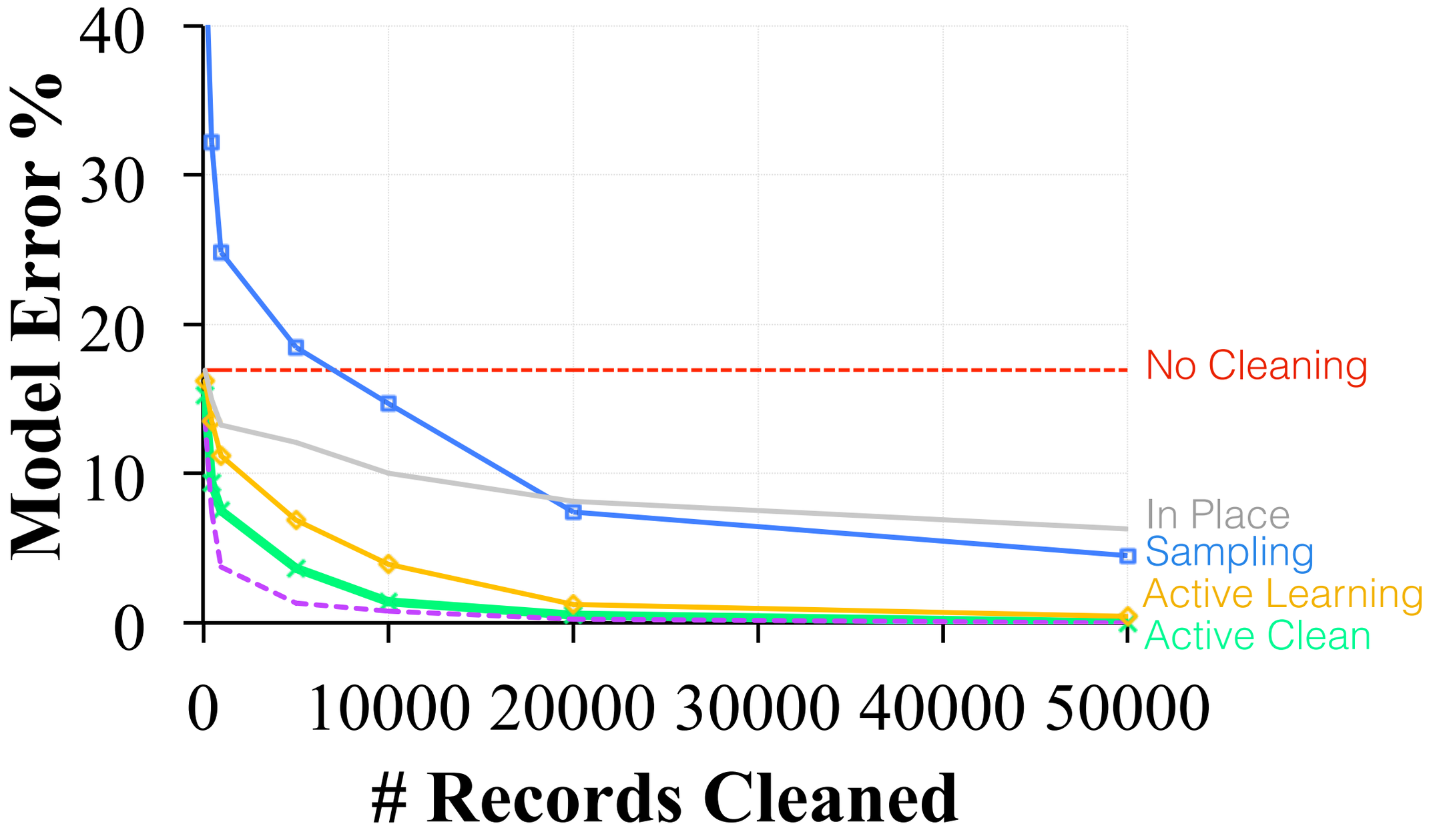


Sanjay Krishnan, Jiannan Wang, Michael J. Franklin, Ken Goldberg, Eugene Wu. ActiveClean: Interactive Data Cleaning For Statistical Modeling. VLDB 2016.

Dollars For Docs

- 250,000 medical contribution records
- Manually labeled as suspicious or not
- Entity resolution errors in company and drug names

Dollars For Docs



There's a bound for that

For a batch size b and iterations T, the ActiveClean stochastic gradient descent updates converge with rate:

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

For strongly-convex models:

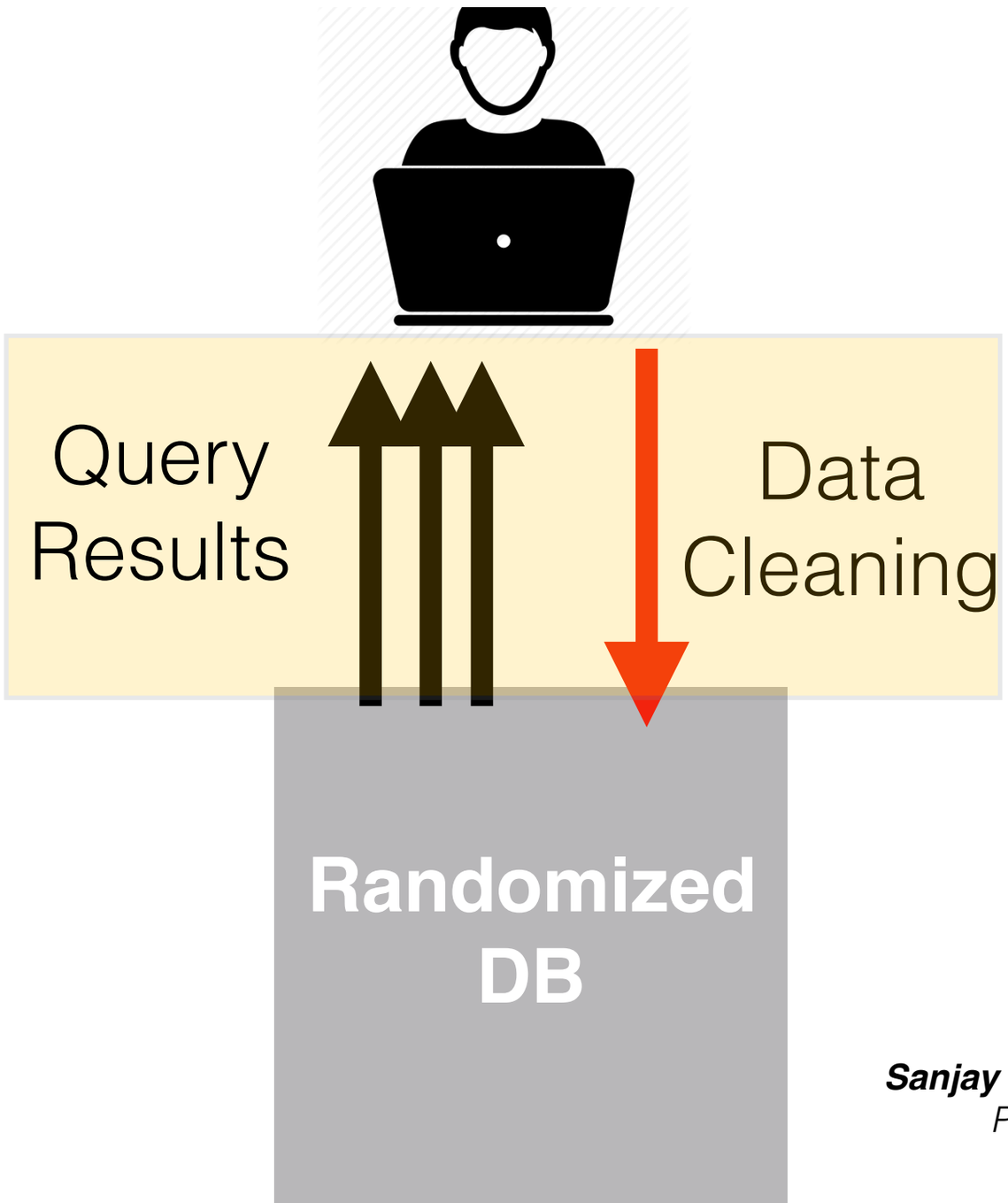
$$O(\frac{1}{T\sqrt{b}})$$

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

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

Data Cleaning + Differential Privacy

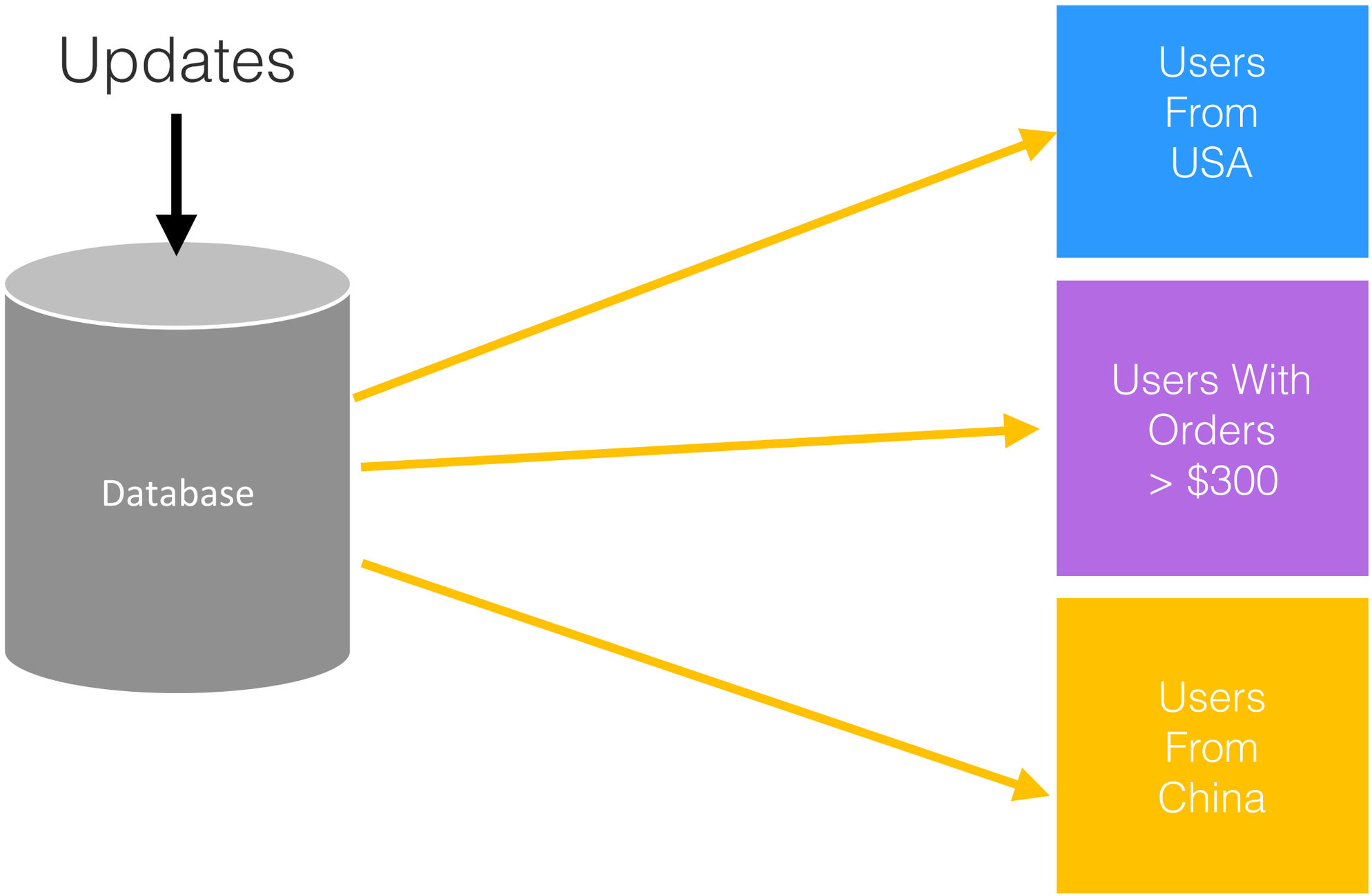
- Not very different from Sample-and-Clean!



Sanjay Krishnan, *Jiannan Wang, Michael J. Franklin, Ken Goldberg, Tim Kraska. PrivateClean: Data Cleaning and Differential Privacy. SIGMOD 2016.*

Streaming Systems

- Approximate Maintenance as Sample-and-Clean



Quantifying Incompleteness

- Similar mechanisms but different estimators!

Brandie Nonnecke*, **Sanjay Krishnan***, et al.. DevCAFE 1.0: A Participatory Platform for Assessing Development Initiatives in the Field. IEEE GHTC. 2015 (Best Paper)

Sanjay Krishnan, Jay Patel, Michael J. Franklin, and Ken Goldberg. Social Influence Bias in Recommender Systems: A Methodology for Learning, Analyzing, and Mitigating Bias in Ratings. RecSys. Foster City, CA, USA. Oct 2014

*Yeouhnoh Chung, **Sanjay Krishnan** , Tim Kraska. A Data Quality Metric (DQM). How to Estimate the Number of Undetected Errors in Data Sets. Under Review VLDB 2017.*

The Statistics of Dirty Data

tl;dr Formalism Good, Theory Needs Updating

- SampleClean: Linking Data Repair To Statistical Analysis.

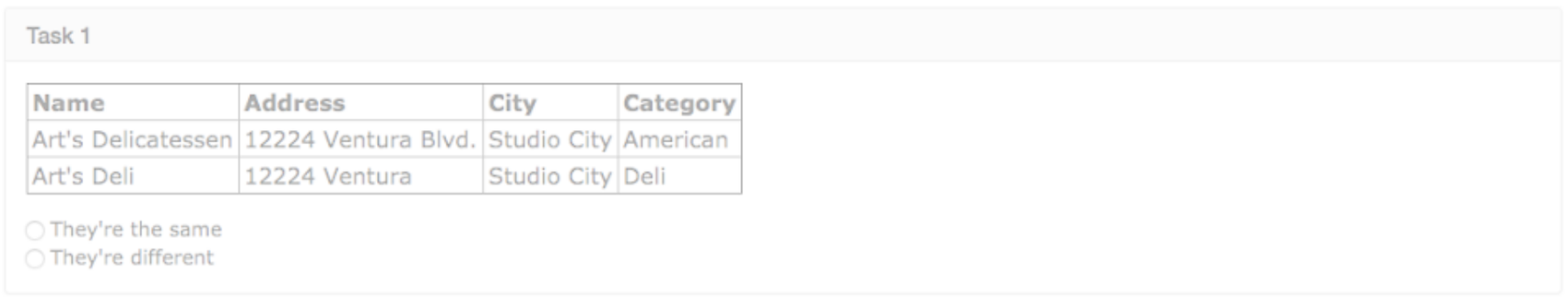
- **AlphaClean: Synthesizing Data Cleaning Programs With New AI Tools**

- Discussion

Quantifying Incompleteness

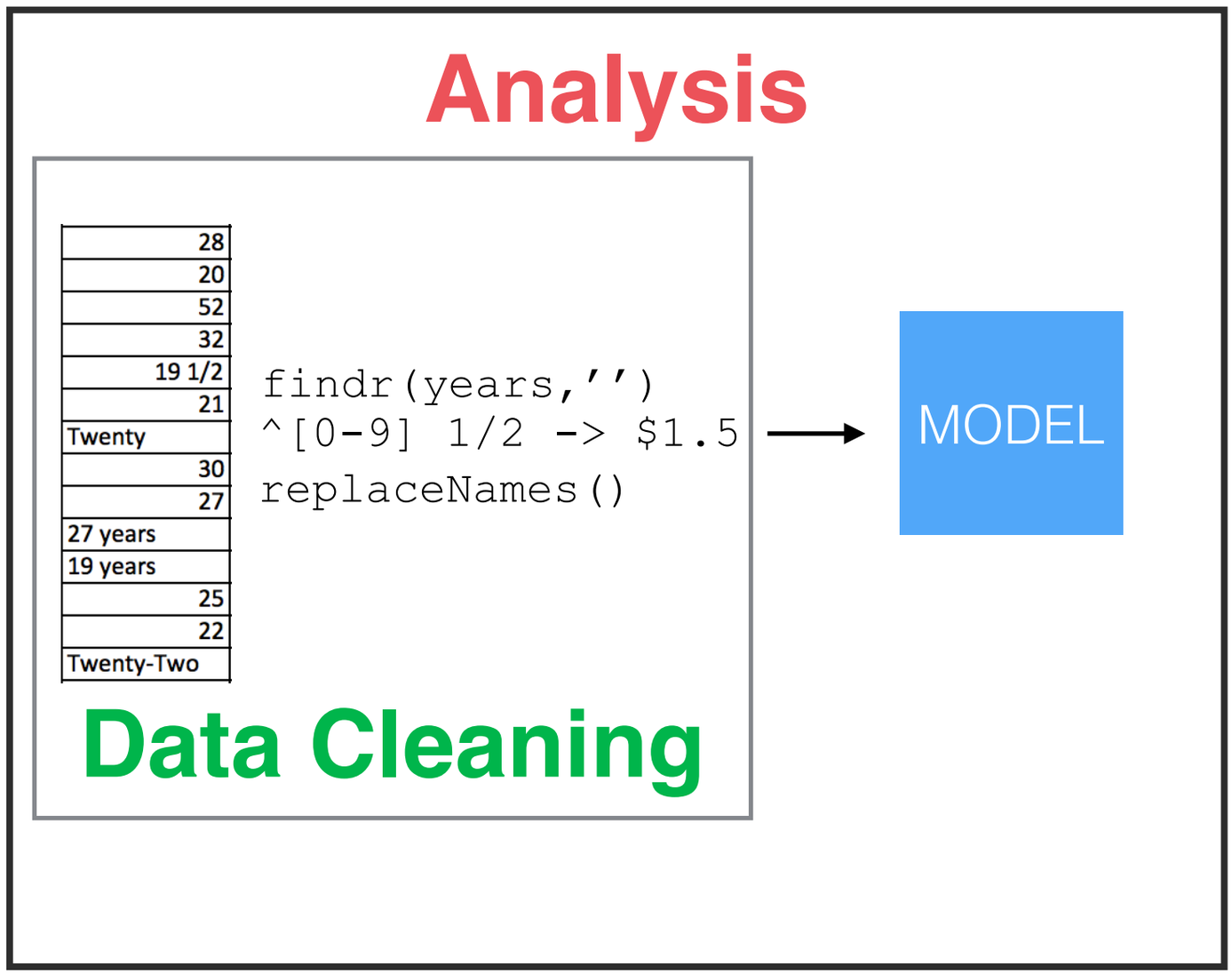
Brandie Nonnecke*, **Sanjay Krishnan***, et al.. DevCAFE 1.0: A Participatory Platform for Assessing Development Initiatives in the Field. IEEE GHTC. 2015 (Best Paper)

Sanjay Krishnan, Jay Patel, Michael J. Franklin, and Ken Goldberg. Social Influence Bias in Recommender Systems: A Methodology for Learning, Analyzing, and Mitigating Bias in Ratings. RecSys. Foster City, CA, USA. Oct 2014



	A	B	C	D
1	**Date**	**Participant ID Number**	**Age**	**What parish do you live in?**
2	18-06-14	249	28	Naluwoli
3	17-06-14	2977	20	
4	17/06/2014	03500	52	Butansi
5	19/06/2014	4194	32	Naluwoli
6	17/06/2014	07420	19 1/2	Butansi
7	17/06/2014	07428	21	Naluwoli
8	17/06/2014	10011	Twenty	Butansi
9	17/06/2014	10061	30	Butansi
10	13-06-14	10431	27	Butansi
11	18/06/2014	10685	27 years	Butansi
12	19/06/2014	10920	19 years	Naluwoli
13	19/06/2014	10982	25	Naluwoli
14	13-06-14	11164	22	Naluwoli
15	17/06/2014	12138	Twenty-Two	Naluwoli

Hard To Disentangle From The Data



The “Database” Perspective

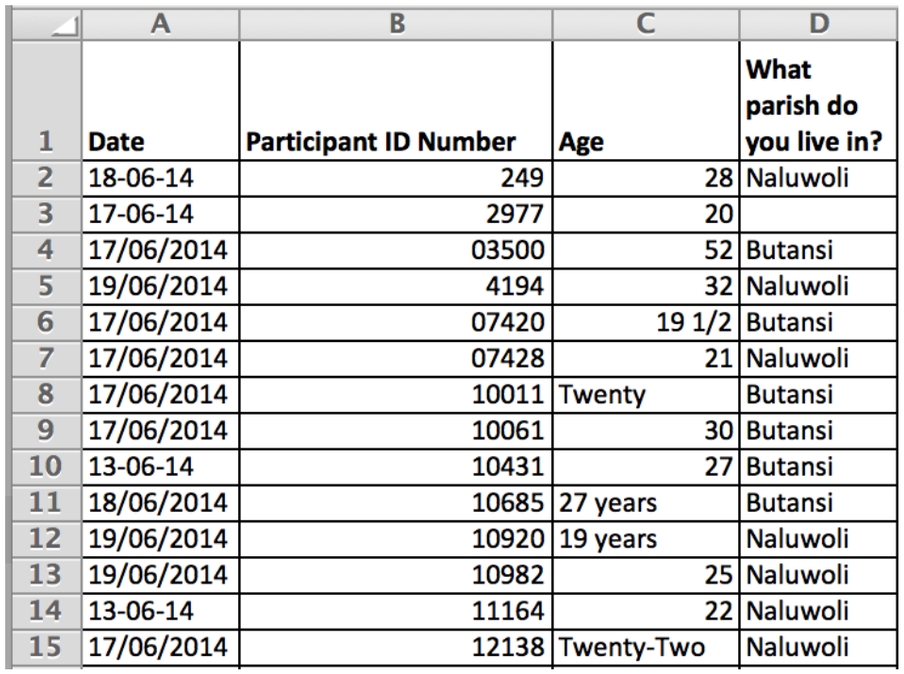
“No Manager Can Earn Less Than an Employee”

	Name	Role	Salary
1	Jane Doe	Emp	1700
2	John Smith	Manager	1500
3	Raj Kumar	Emp	1300
4	Maria Lopez	Manager	4400

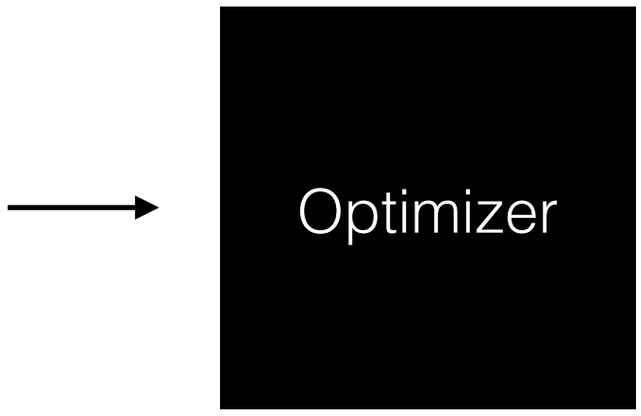
Say what you want not how you get it

Making Data Cleaning Declarative

```
Age.isFloat()
```



	A	B	C	D
1	**Date**	**Participant ID Number**	**Age**	**What parish do you live in?**
2	18-06-14	249	28	Naluwoli
3	17-06-14	2977	20	
4	17/06/2014	03500	52	Butansi
5	19/06/2014	4194	32	Naluwoli
6	17/06/2014	07420	19 1/2	Butansi
7	17/06/2014	07428	21	Naluwoli
8	17/06/2014	10011	Twenty	Butansi
9	17/06/2014	10061	30	Butansi
10	13-06-14	10431	27	Butansi
11	18/06/2014	10685	27 years	Butansi
12	19/06/2014	10920	19 years	Naluwoli
13	19/06/2014	10982	25	Naluwoli
14	13-06-14	11164	22	Naluwoli
15	17/06/2014	12138	Twenty-Two	Naluwoli




```
deleteToken('year')
deleteToken('1/2')
textToNumber('Twenty')
```

```
{deleteToken(?),
 textToNumber(?),
       …}
```

Data Cleaning is Planning

- Input: A formal language of transformations. (Actions)

$$\langle \{t_1, t_2 ..., \}, \circ, \emptyset \rangle$$

- Input: a quality function of the following form where 0 implies clean (Reward):

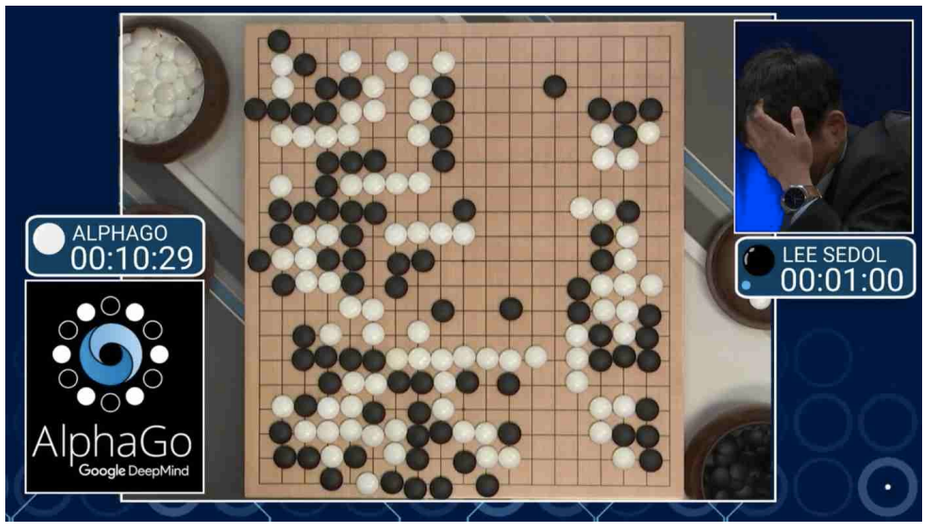
$$Q : r \in R \mapsto [0, 1]$$

- Output: A composite transformation that optimizes

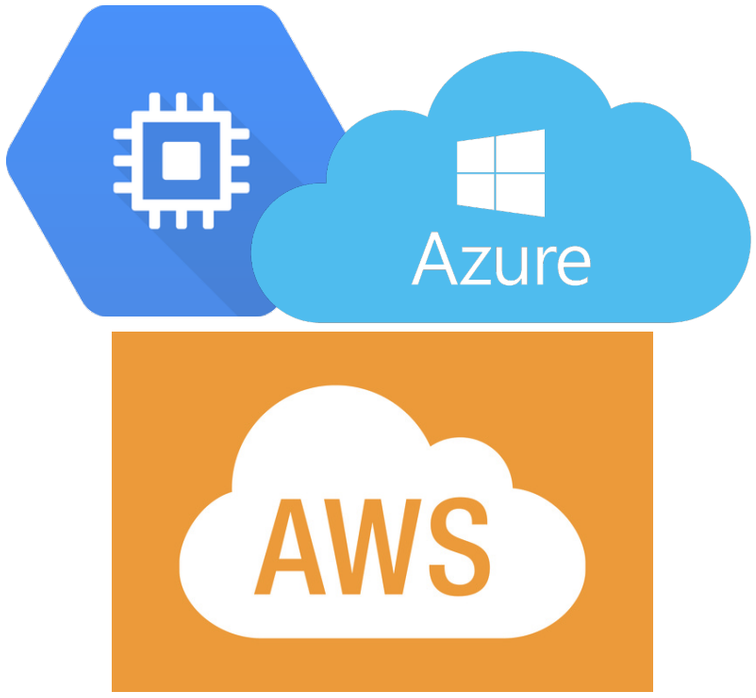
$$\sum_r Q[(t_1 \circ t_2 \circ ...t_k)(R)]$$

(State-Transition)

AlphaGo

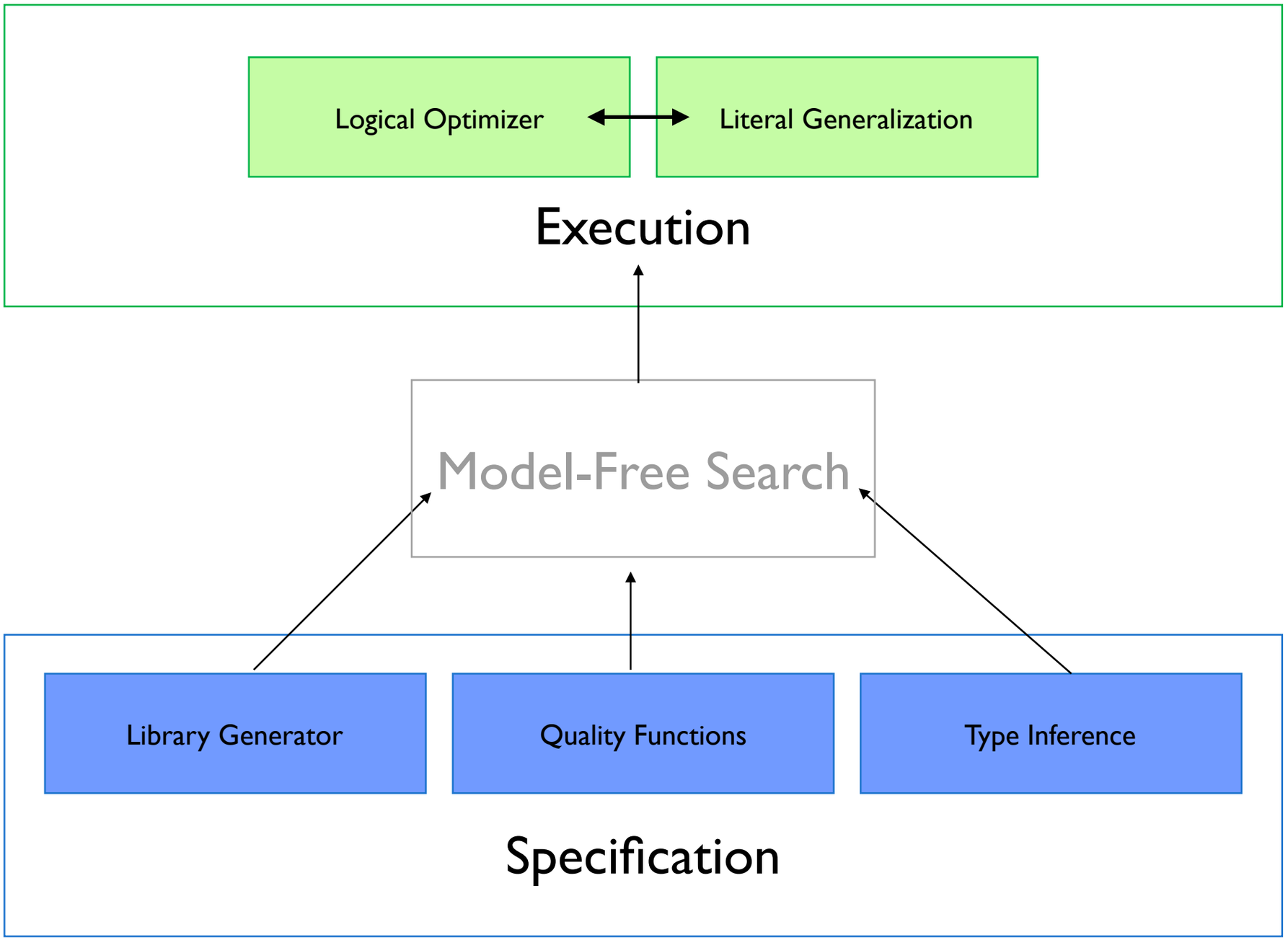


Model-Free Search



Commodity Clusters

AlphaClean


Logical Optimizer
Literal Generalization
Execution
Model-Free Search
Library Generator
Quality Functions
Type Inference
Specification

Example

	A	B	C	D
1	**Date**	**Participant ID Number**	**Age**	**What parish do you live in?**
2	18-06-14	249	28	Naluwoli
3	17-06-14	2977	20	
4	17/06/2014	03500	52	Butansi
5	19/06/2014	4194	32	Naluwoli
6	17/06/2014	07420	19 1/2	Butansi
7	17/06/2014	07428	21	Naluwoli
8	17/06/2014	10011	Twenty	Butansi
9	17/06/2014	10061	30	Butansi
10	13-06-14	10431	27	Butansi
11	18/06/2014	10685	27 years	Butansi
12	19/06/2014	10920	19 years	Naluwoli
13	19/06/2014	10982	25	Naluwoli
14	13-06-14	11164	22	Naluwoli
15	17/06/2014	12138	Twenty-Two	Naluwoli

```
df = pd.read_csv('uganda.csv', quotechar='\"', index_col=False)
```

Specification API

Patterns

Allowed Values a Column Can Take

```
patterns = []

#18 years old to 100, remove under 18
patterns += [Float('Age', [18, 100])]

#Only alpha numeric values
patterns += [Pattern('Response', “^[a-zA-Z0-9_]*$")]

#Parish
patterns += [CodeBook(‘Parish’, allowed_parishes)]
```

Specification API

Dependencies

Allowed Relationships Between Columns

```
dependencies = []

#Manual Collections Happened on a Specific Day
dependencies += [CFD('Parish -> Day', isManual)]

#Logical Checks
patterns += [DC('Age', "< 22", "Children", "< 5")]
```

Specification API

Statistical Hints

Model the data is expected to follow

```
stats = []

#Expect Pos. Correlation
stats += [Correlate('Age', 'Children')]

#Previous Year's model
stats += [Model]
```

Data Cleaning is Planning

- Input: a quality function of the following form where 0 implies clean (Reward):

$$Q : r \in R \mapsto [0, 1]$$

The Language

User Defines Templates

```
findAndReplace(attribute, value1, value2)

clip(attribute, threshold)

filterToken(attribute, substring)

regex(attribute, reg)
```

Data Cleaning is Planning

- Input: a quality function of the following form where 0 implies clean (Reward):

$$Q : r \in R \mapsto [0, 1]$$

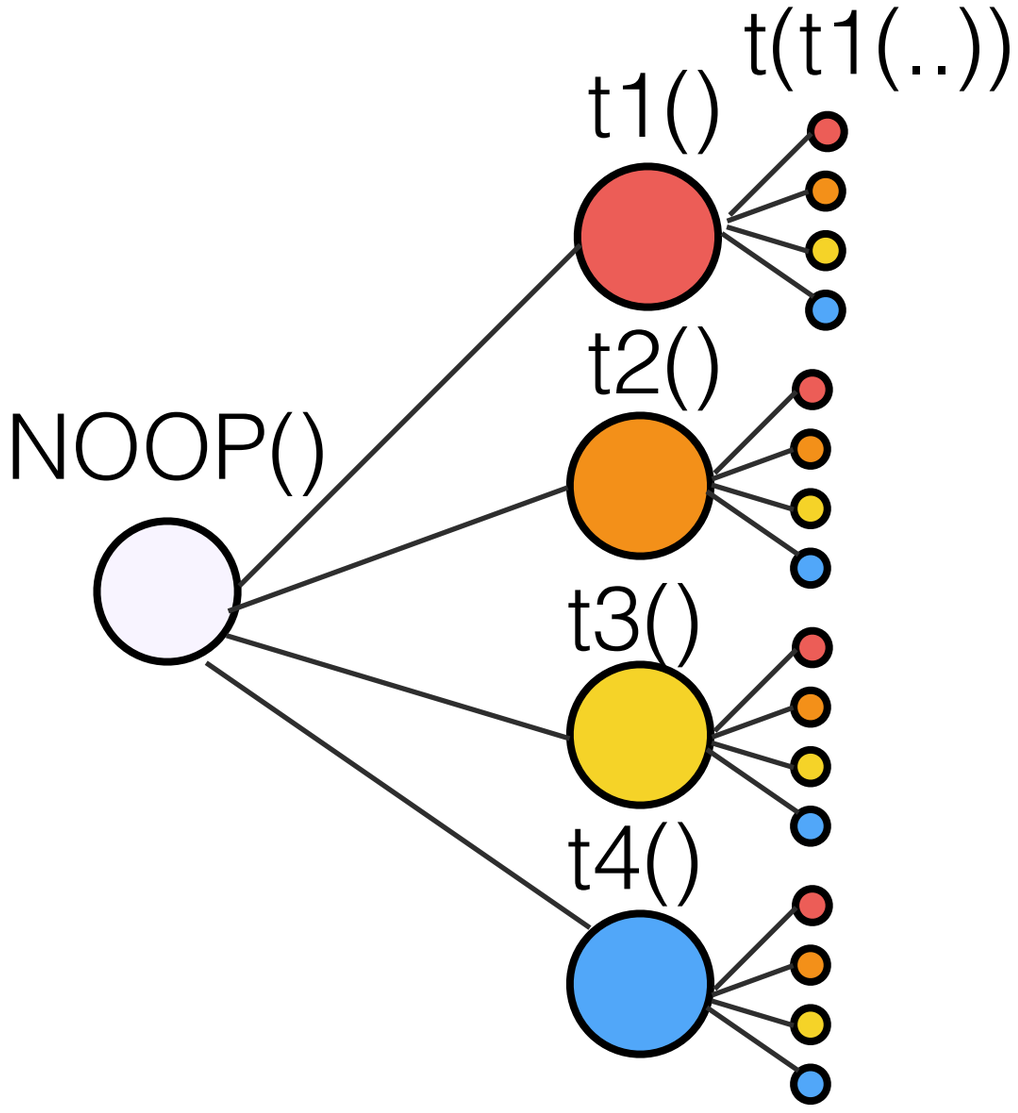
Data Cleaning is Planning

- Output: A composite transformation that optimizes

$$\sum_r Q[(t_1 \circ t_2 \circ ...t_k)(R)]$$

(State-Transition)

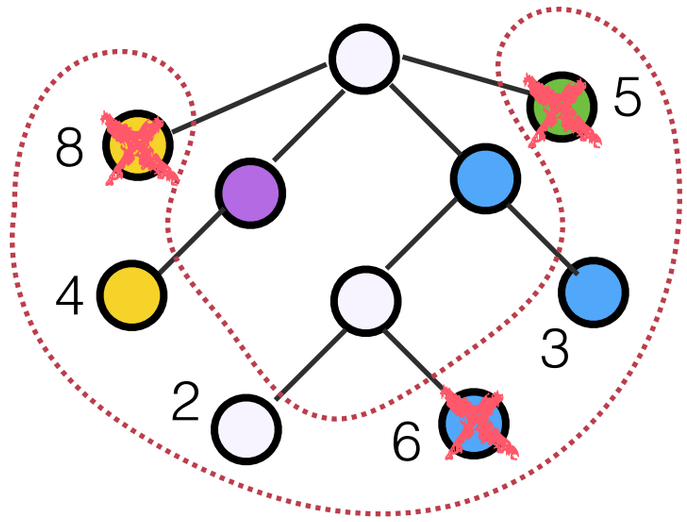
Data Cleaning is Planning



- Typical errors are localized (greedy fixes are safe)
- Typical errors are systematic (previous fixes give information about future fixes)

Basic Algorithm

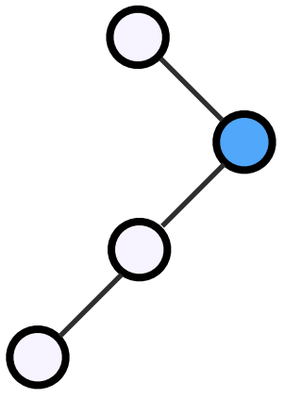
- γ-greedy best first search



- Keep a node on the frontier if it is within γ of the current best result.

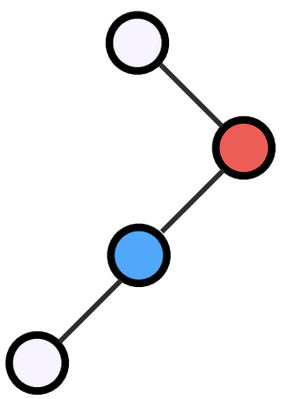
Learning a Search Heuristic

8



4

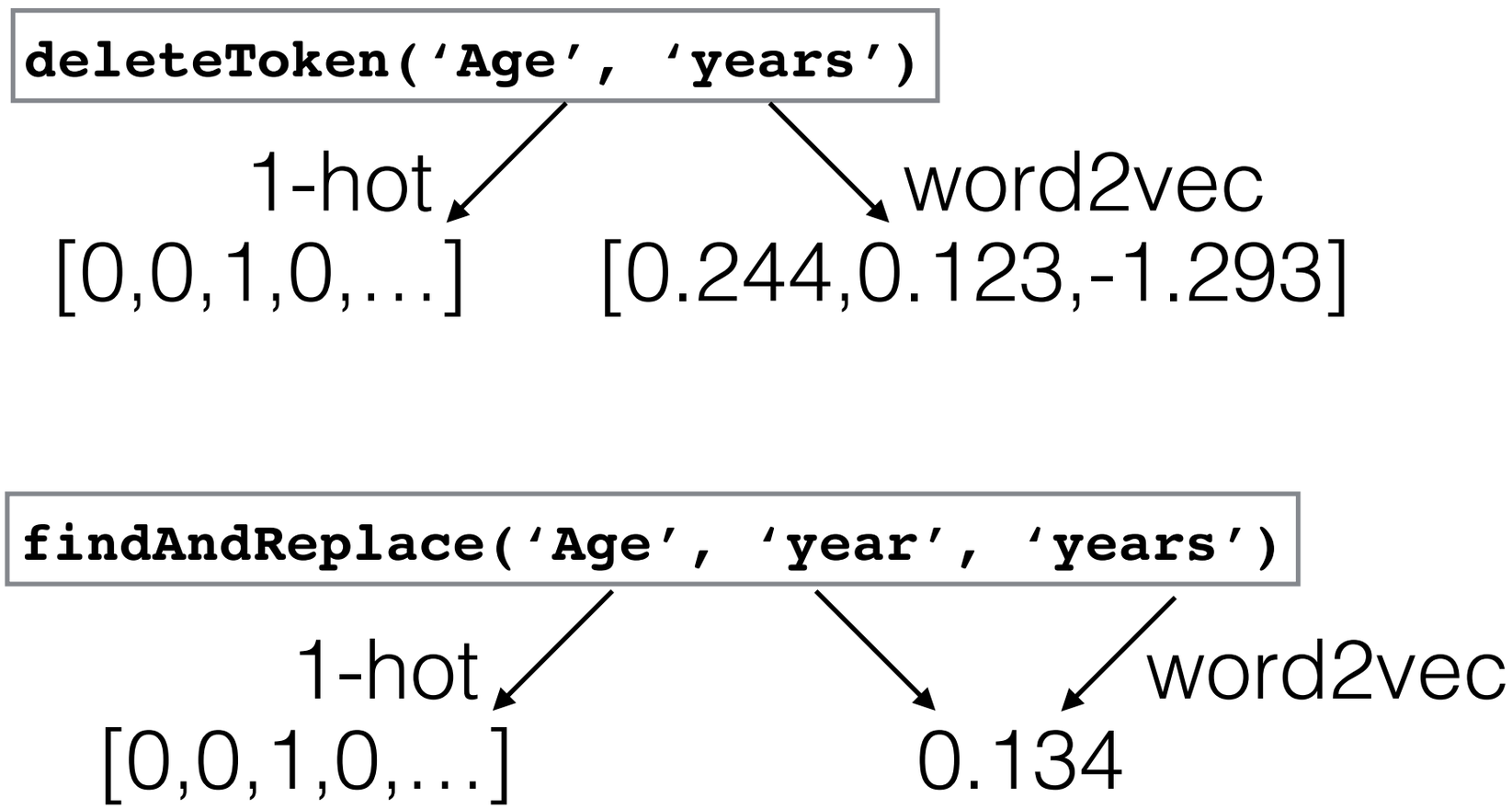
2



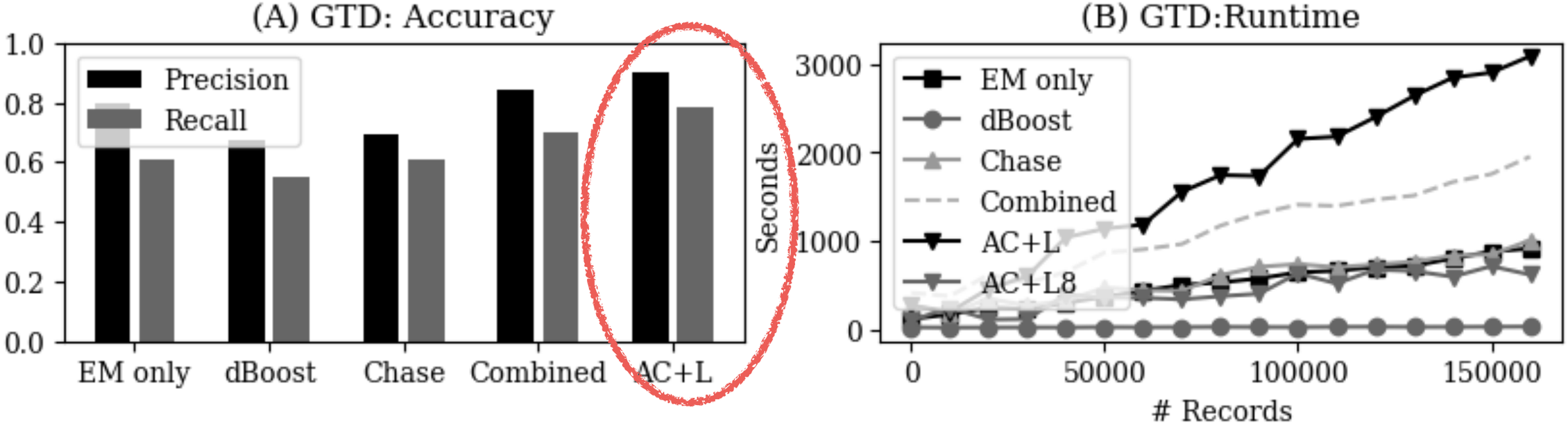
?

Learning a Search Heuristic

Featurize Transformations

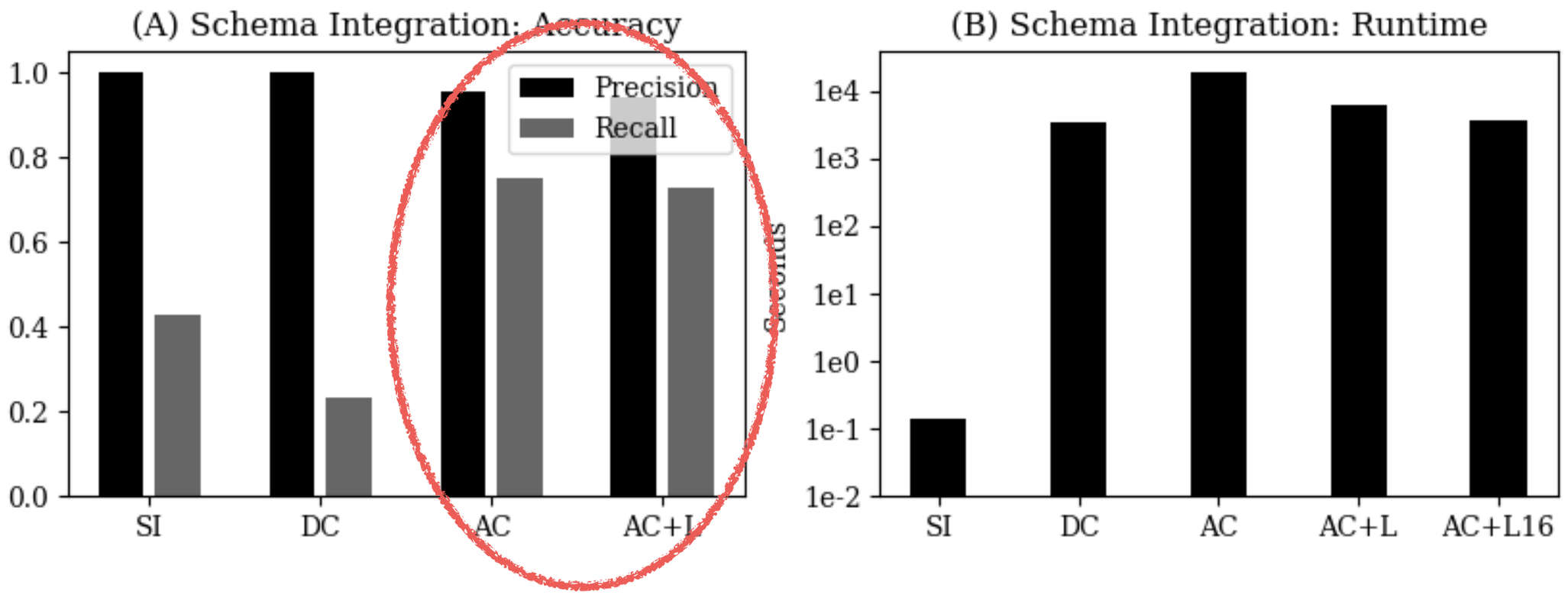


GTD: Combinations



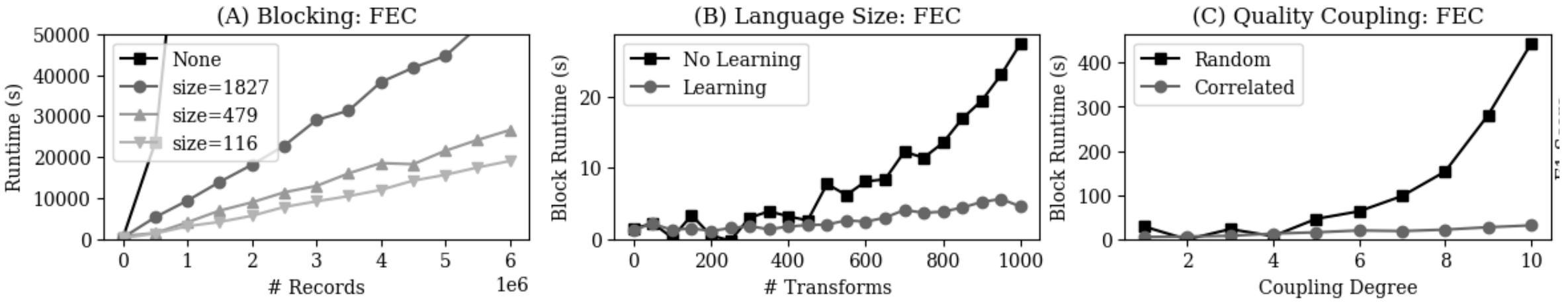
Schema Integration

- Link columns and enforce integrity constraints
- Stock Dataset: There are 1000 ticker symbols from 55 sources for every trading day in a month.

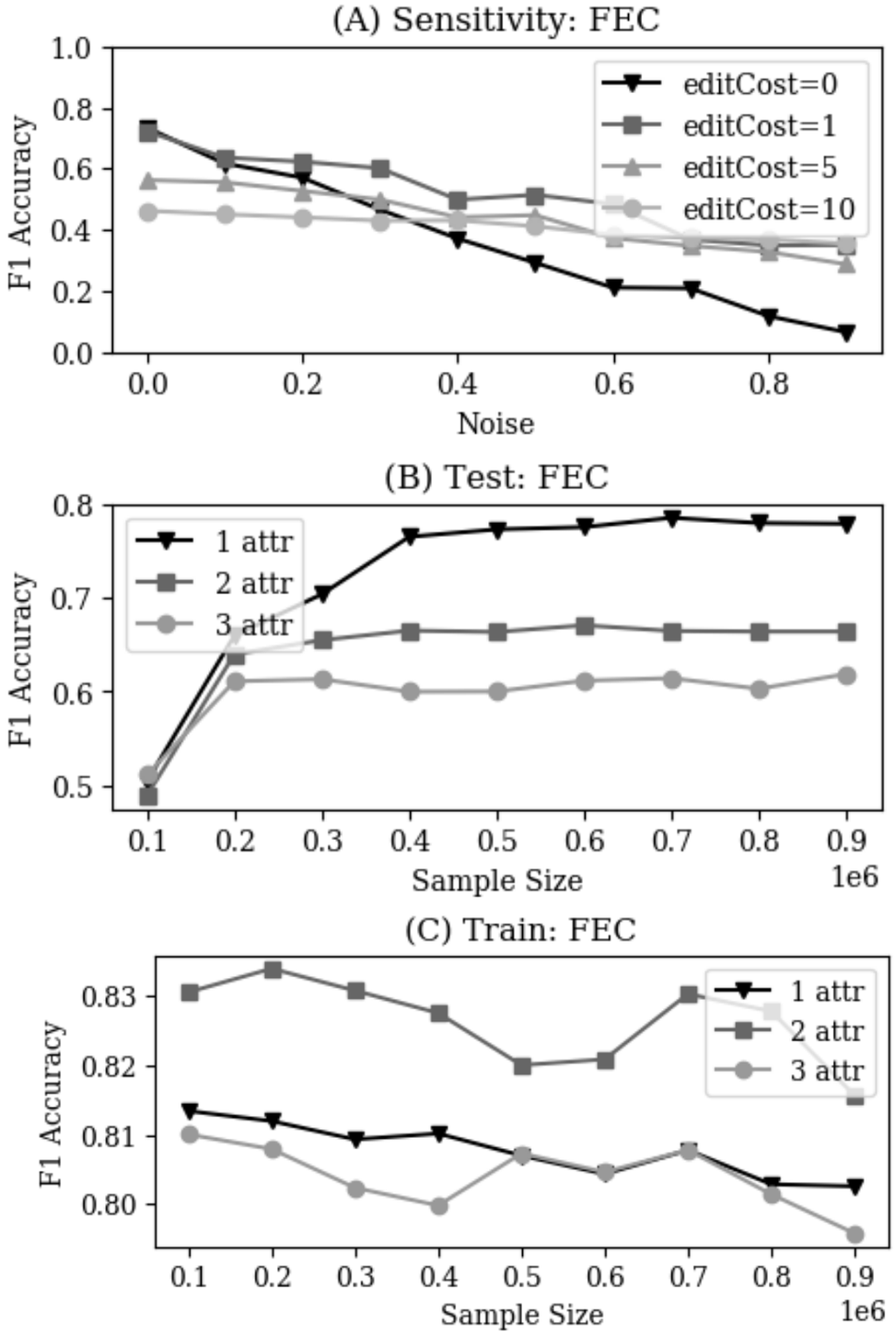


Performance

- Works well on systematic and localized errors



Overfitting and Underfitting



Depth

Sensitivity

Language Size

$$\epsilon_m < c\sqrt{\frac{k}{2m} \cdot \ln \frac{|\mathcal{T}|}{2\delta}}$$

Data

The Statistics of Dirty Data

tl;dr Formalism Good, Theory Needs Updating

- SampleClean: Linking Data Repair To Statistical Analysis.
- AlphaClean: Synthesizing Data Cleaning Programs With New AI Tools
- **Discussion**

Conclusion

Data Cleaning is a Statistical Problem

- **Data Cleaning before Statistical Analytics**:
  [SIGMOD 14], [IEEE DEB 15], [VLDB 16]

- **Sampling and Approximation with Writes**:
  [VLDB 15], [SIGMOD 16]

- **Crowdsourcing's Downstream Impact**
  [VLDB Demo 15], [RecSys 14], [VLDB review 17]

www.ocf.berkeley.edu/~sanjayk