Data Cleaning: A Statistical Perspective

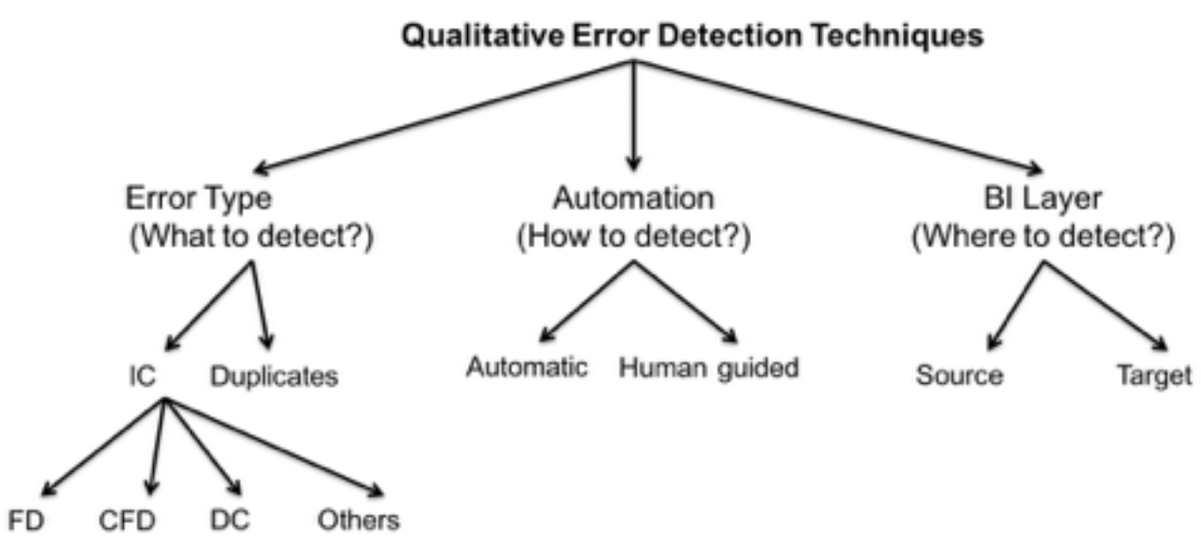
Data Cleaning: Overview and Challenges Part 2

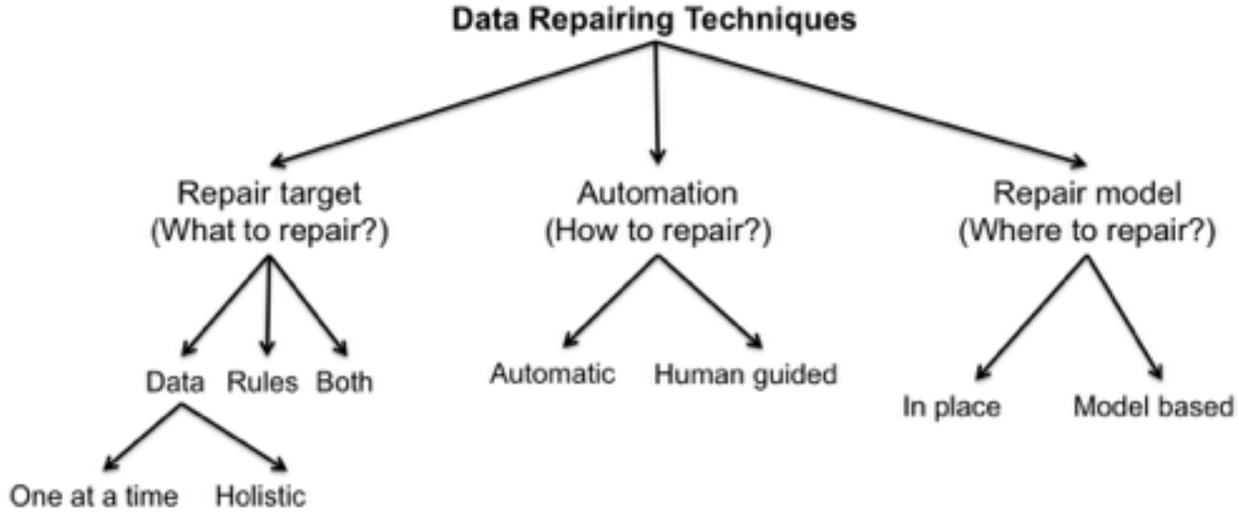
Sanjay Krishnan (UC Berkeley) and
Jiannan Wang (Simon Fraser U.)

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
…

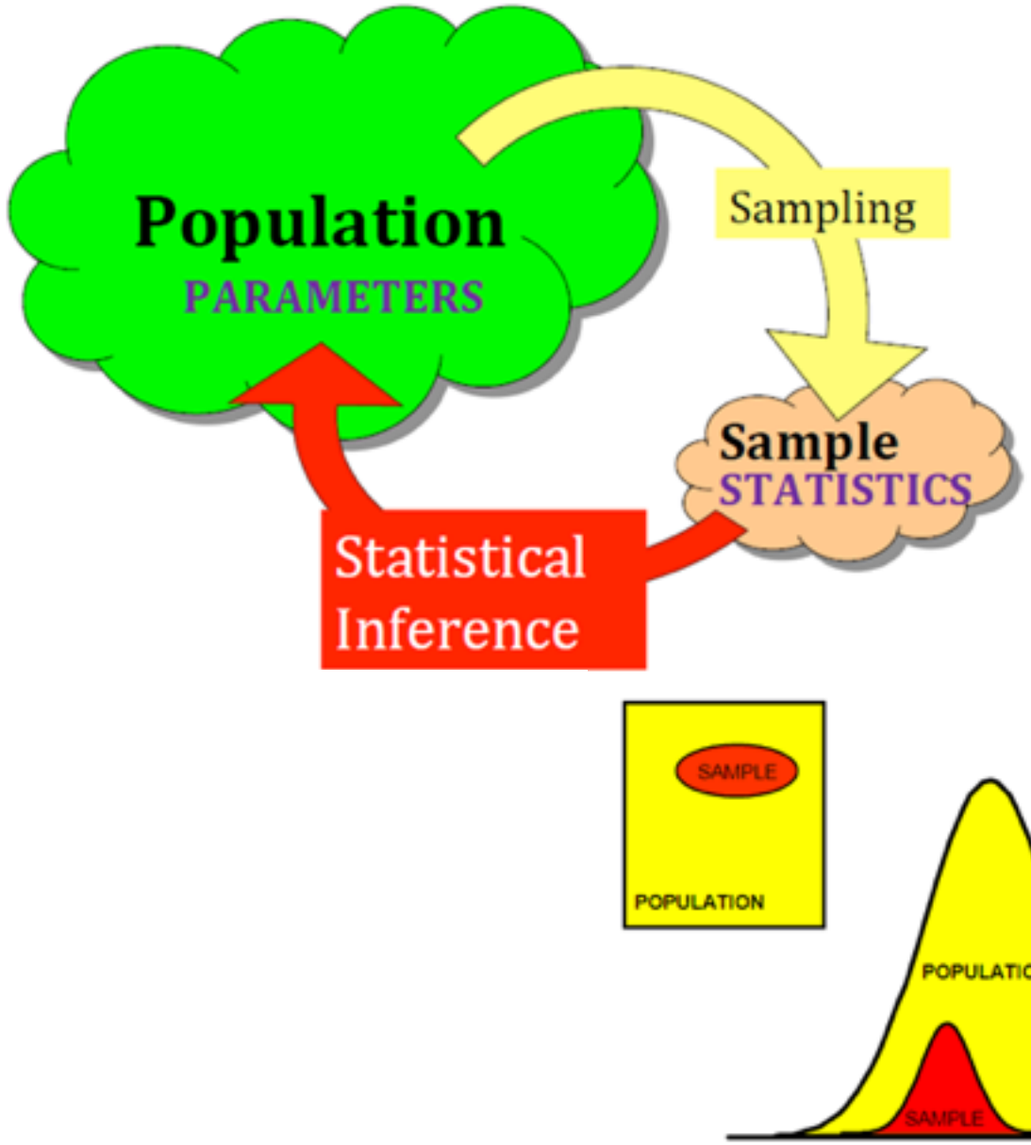
Why Statistics?

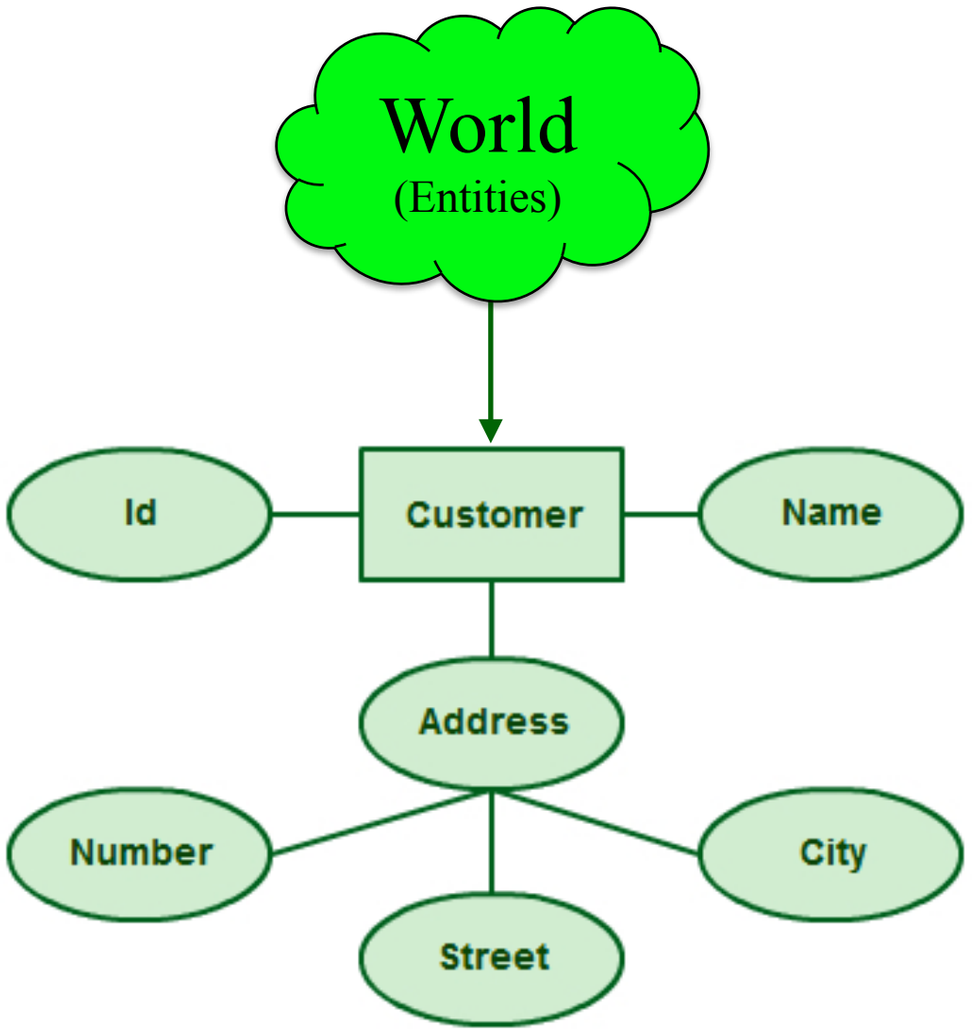
- Growing popularity of advanced analytical techniques (e.g., Machine Learning, Stochastic Optimization).

- 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 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

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

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

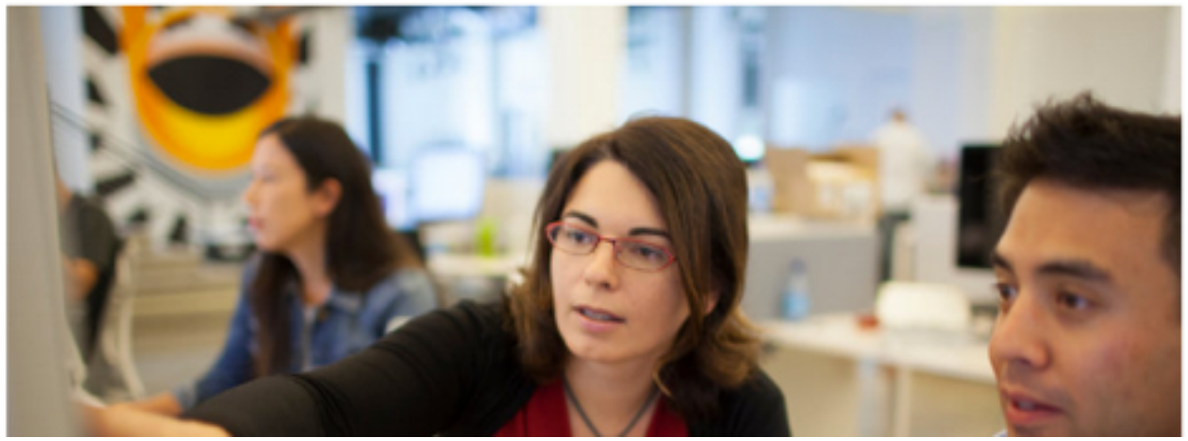
Email

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 **D**
  - **D** is *inconsistent* if ∃ **σ** ∈ **Σ** : **σ**(**D**) = False

- *Error Detection*: Identify a set of rows from the relations in **D** such that that if removed **D** is consistent.

- *Error Repair*: Identify a sequence of repairs $\mathbf{C}_1,\ldots,\mathbf{C}_k$ such that $\mathbf{C}_1 \circ \ldots \circ \mathbf{C}_k(D)$ 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.

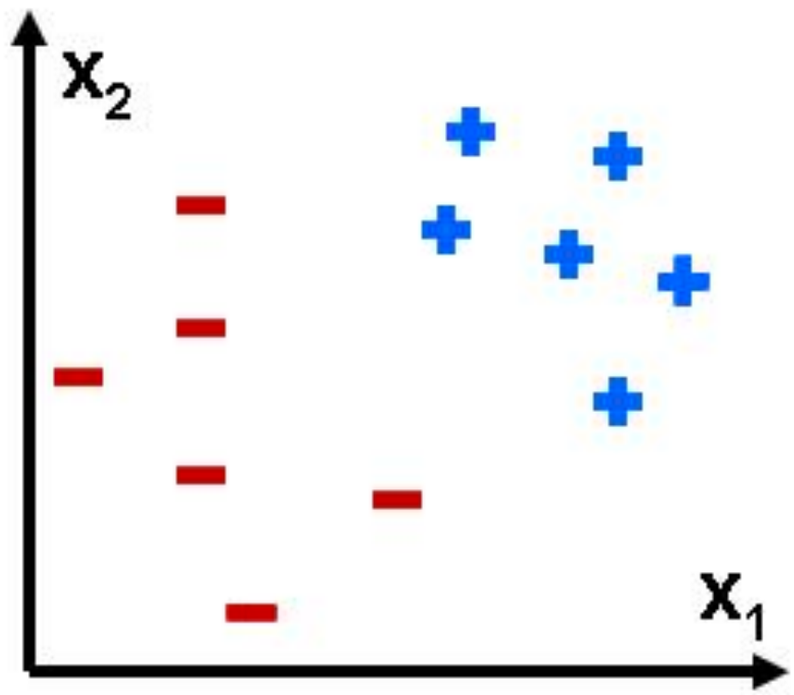
Madden, S. R., Franklin, M. J., Hellerstein, J. M., & Hong, W. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (2005): 122-173.

- 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

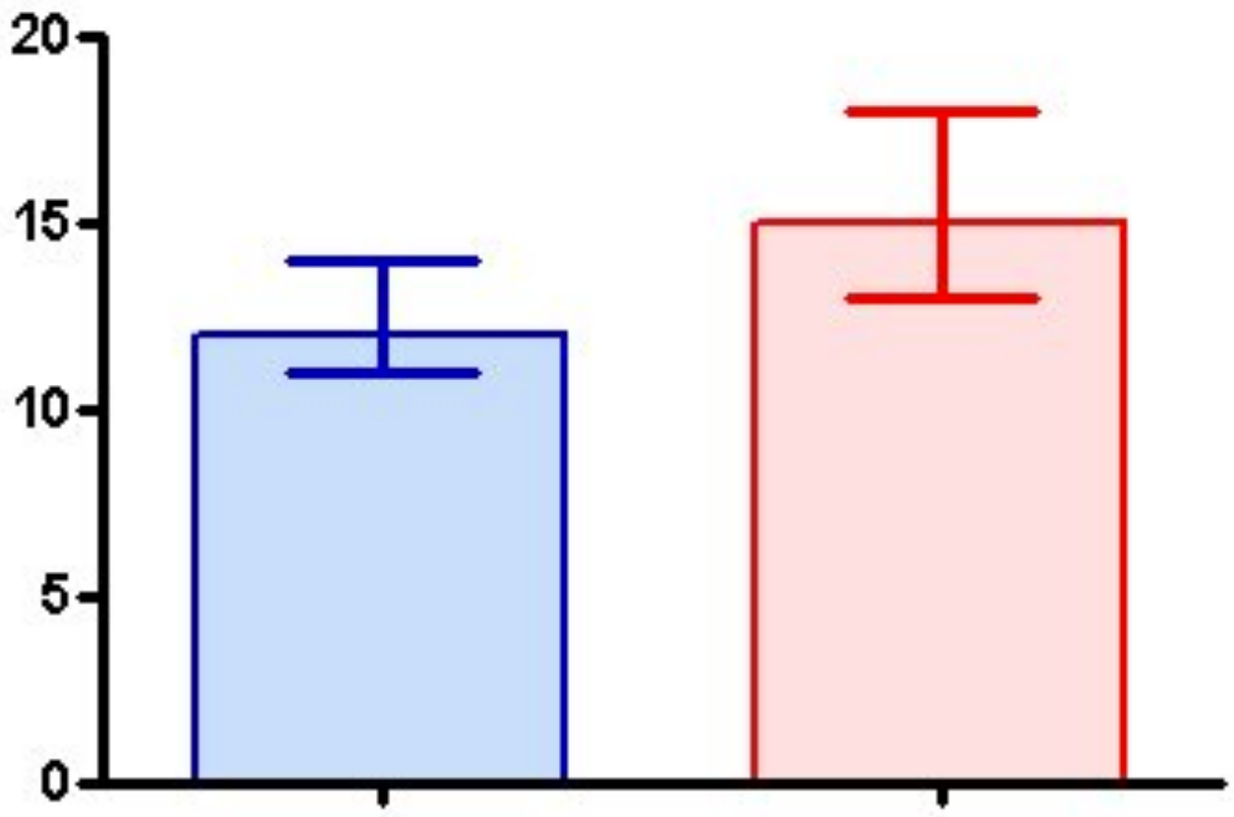
- Aggregate queries can be ambiguous.
- Consistency is not statistical accuracy

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

0.151 C()

0.956 C()

0.256 C()

0.126 C()

Author	Year	Paper ID
Agrawal	1992	1334
Agrawal	**1999**	1451
Agrawal	1996	1651
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

This tutorial

- **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

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

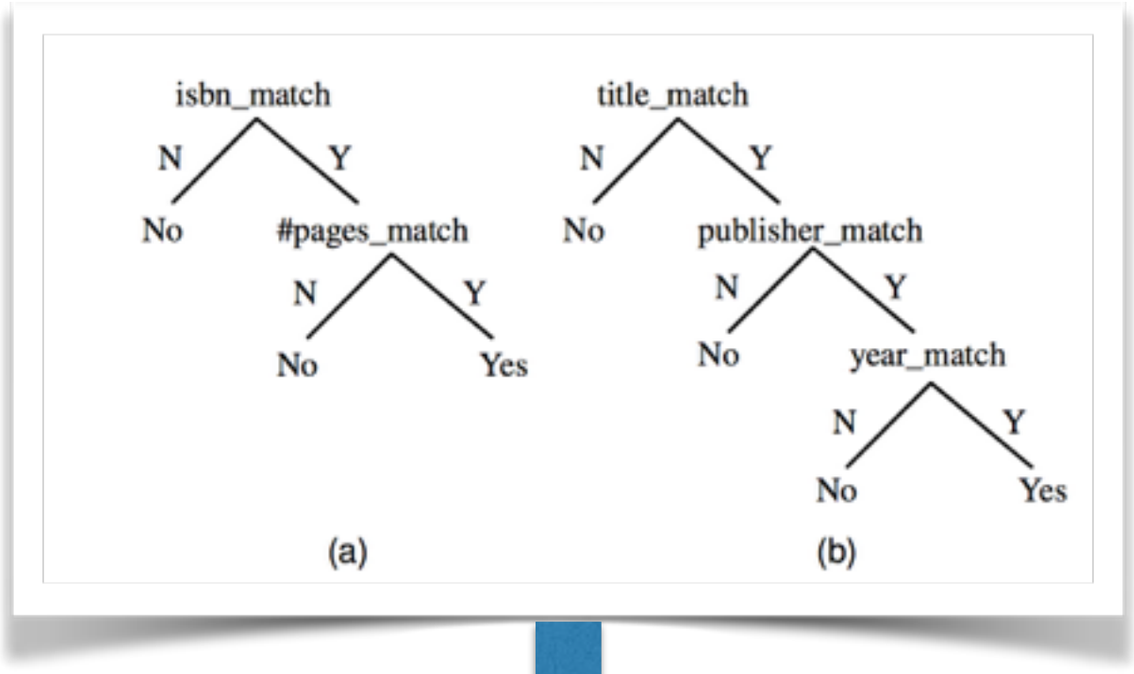
- **Trend 2: Combining Statistical Techniques with Crowdsourcing**
  - ✓ Workflow Design
  - ✓ Quality/Cost/Latency Trade-off

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

Gokhale, Chaitanya, et al. "Corleone: hands-off crowdsourcing for entity matching." Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014.

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, zip, 98004, 90%), which means that t_2[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[zip] should be 98004
    2. "Reject", the value of t_2[zip] should not be 98004
    3. "Retain", the value of t_2[zip] is correct
  - ✓ Train a classifier based on the labeled repairs and use it to classify other (unlabeled) repairs

Yakout, Mohamed, et al. "Guided data repair." Proceedings of the VLDB Endowment 4.5 (2011): 279-289.

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

```
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
```

How?

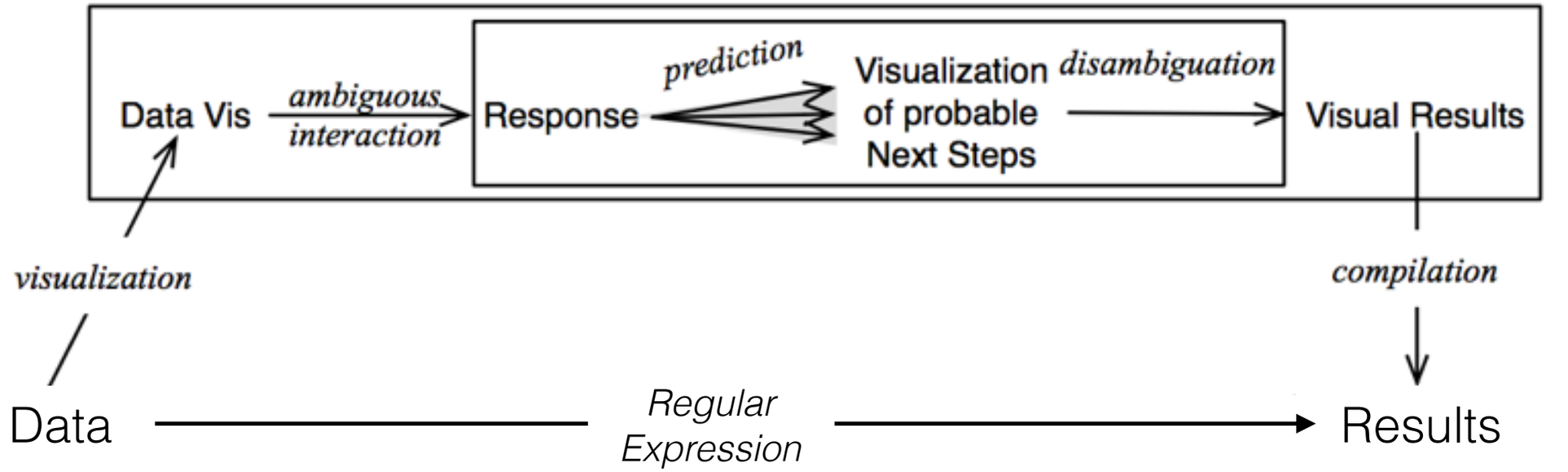
Extracted Values

```
dynamic
dynamic
dynamic
mobile
```

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

Data Transformation as Learning a Predictive Model

- **Predictive Interaction**



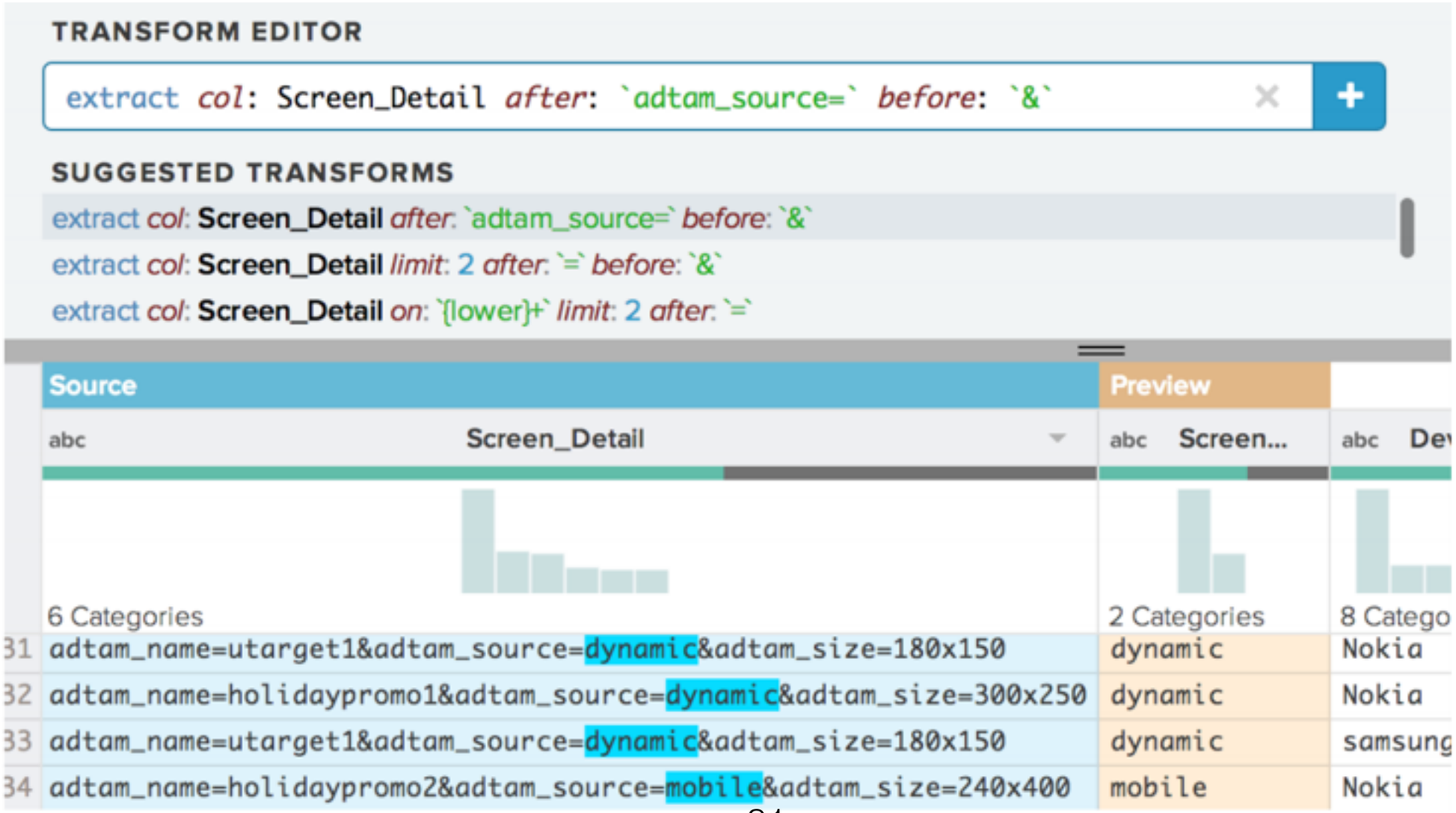

```
31 adtam_name=utarget1&adtam_source=
32 adtam_name=holidaypromo1&adtam_sou
33 adtam_name=utarget1&adtam_source=
34 adtam_name=holidaypromo2&adtam_sou
```

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

```
dynamic
dynamic
dynamic
mobile
```

Data Transformation as Learning a Predictive Model

- **Demonstration**



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

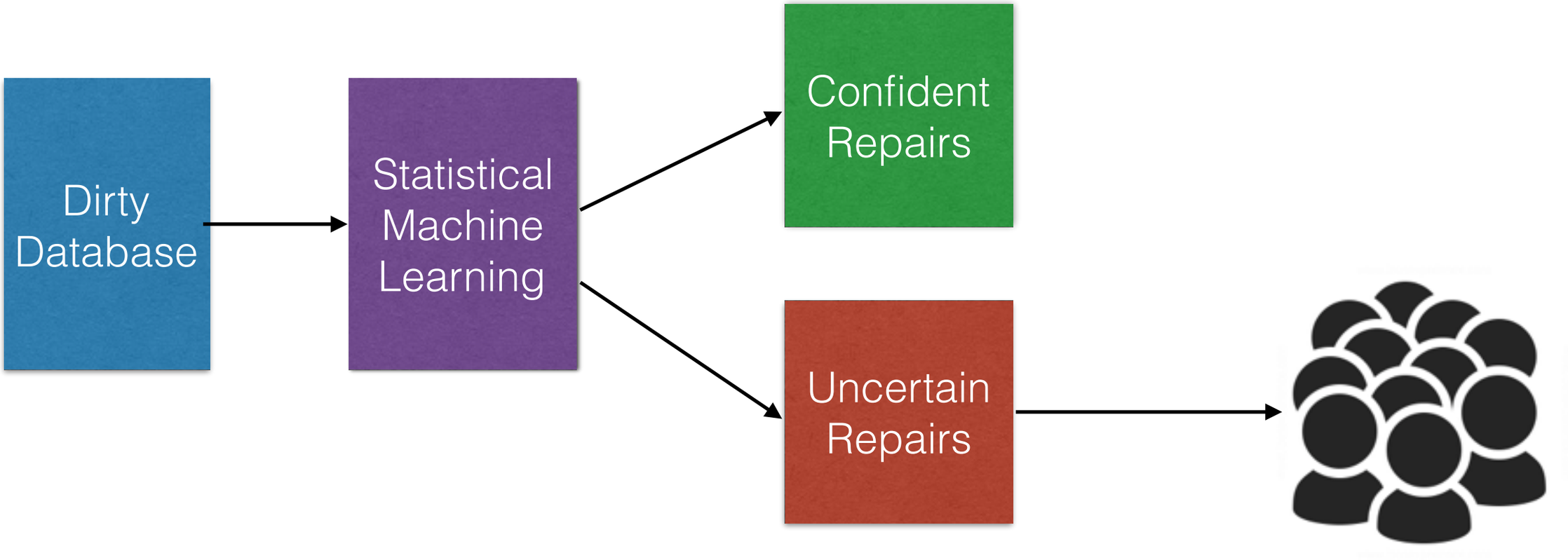
- Crowdsourcing platforms makes the use of humans for doing data cleaning tasks easier and cheaper

- 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.

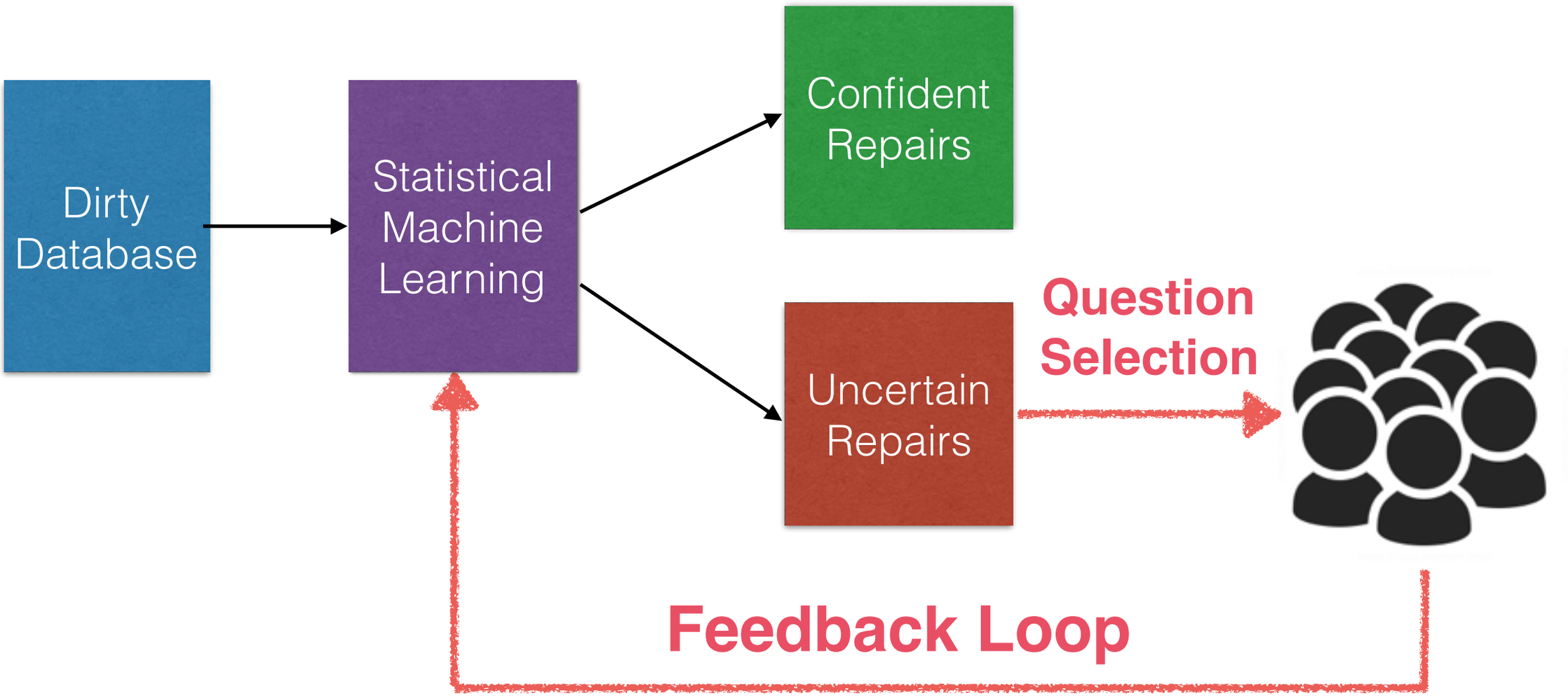
How to combine statistical techniques with crowds?

- **Non-iterative Workflow**



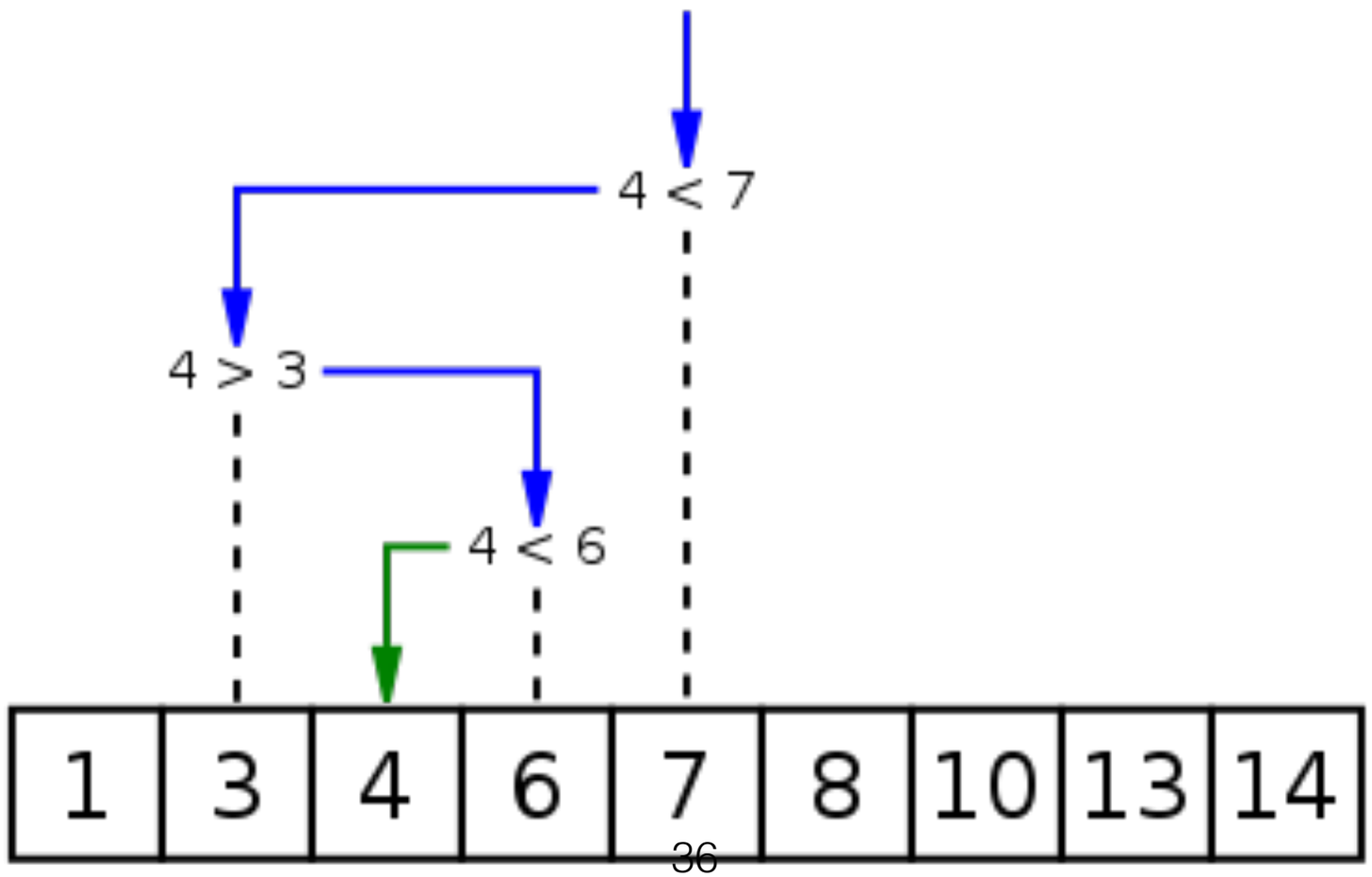
How to combine statistical techniques with crowds?

- **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**

Settles, Burr. "Active learning literature survey." University of Wisconsin, Madison 52.55-66 (2010): 11.

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

VS

- **Machines**
  - ✓ 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

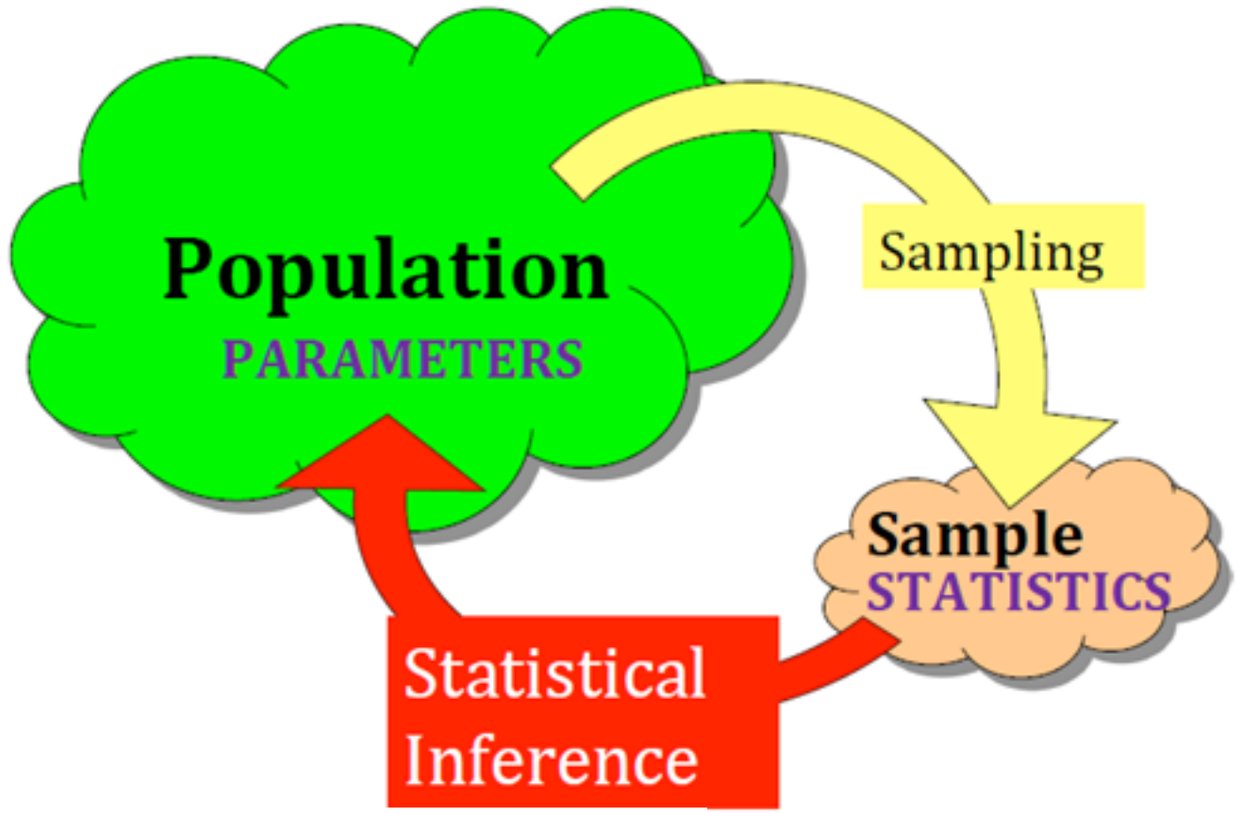
A Statistical Perspective

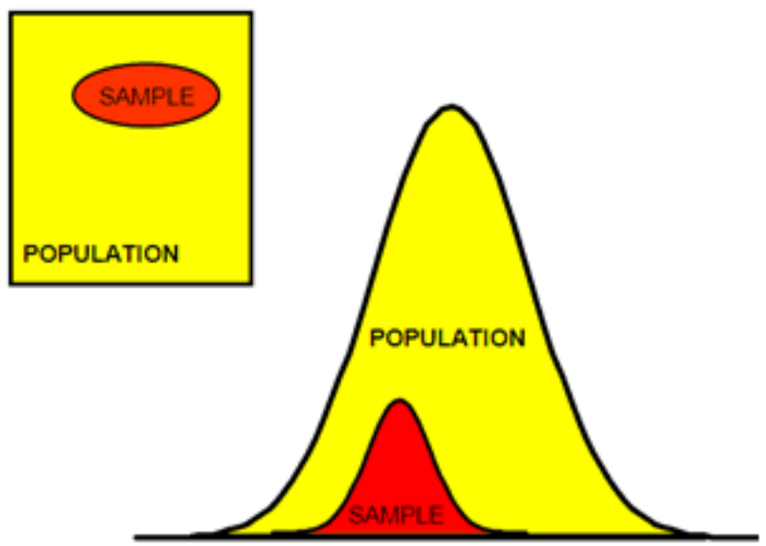
- Topic 1. Statistical techniques to clean data (20 mins)
- **Topic 2. Cleaning data before statistical analytics (50 min)**
- Topic 3. Impact and Future Directions (10 mins)

Section Structure

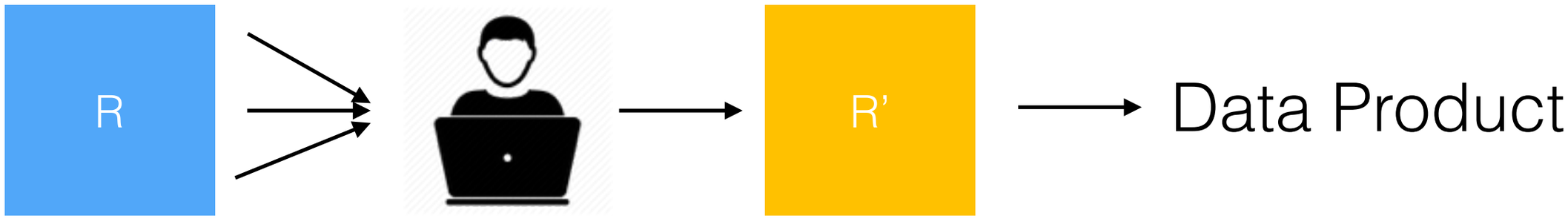
- **Extended Data Cleaning Definition**
- Connecting Data Cleaning to Downstream Queries
  - Aggregate queries
  - Machine learning training
  - Exploiting Relational Information

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 **D** be a dirty database, there exists one "true" cleaned **D'.**

- There exists a sequence of transformations to D such that **D'= $\mathbf{C}_1 \circ \ldots \circ \mathbf{C}_k(\mathbf{D})$**, where each **C** $\in \mathcal{C}$ **.**

- The process of data cleaning is *discovering* the transformations **$\mathbf{C}_1 \circ \ldots \circ \mathbf{C}_k$.**

- *Effort*: The number of records (**k**) in **D** to discover **$\mathbf{C}_1 \circ \ldots \circ \mathbf{C}_k$**

- *Query Result Error*: Let **q** be an aggregate query, the error is defined as **ll q(D) - q(D') ll**

Interaction Model

- Queries the database, observes a dirty record **r** ∈ **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 $\mathbf{C}_1 \circ \ldots . \circ \mathbf{C}_k$

Supported Data Cleaning

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

Record-by-Record

- *Def:* Given a dirty record **r** ∈ **R ,** the analyst can return an **r'**

Madden, Samuel R., et al. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (**20015**): 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 **r** ∈ **R** the program returns an **r'**

Madden, Samuel R., et al. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (**20015**): 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 **r** ∈ **R** the program returns an **r'**

Madden, Samuel R., et al. "TinyDB: an acquisitional query processing system for sensor networks." ACM Transactions on database systems (TODS) 30.1 (**20015**): 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 **r** ∈ **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.

2

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.

Extraction

- *Def:* Given a dirty record **r** ∈ **R ,** the analyst returns a record **r’** ∈ **R’** over an extended set of attributes **A = A U 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)

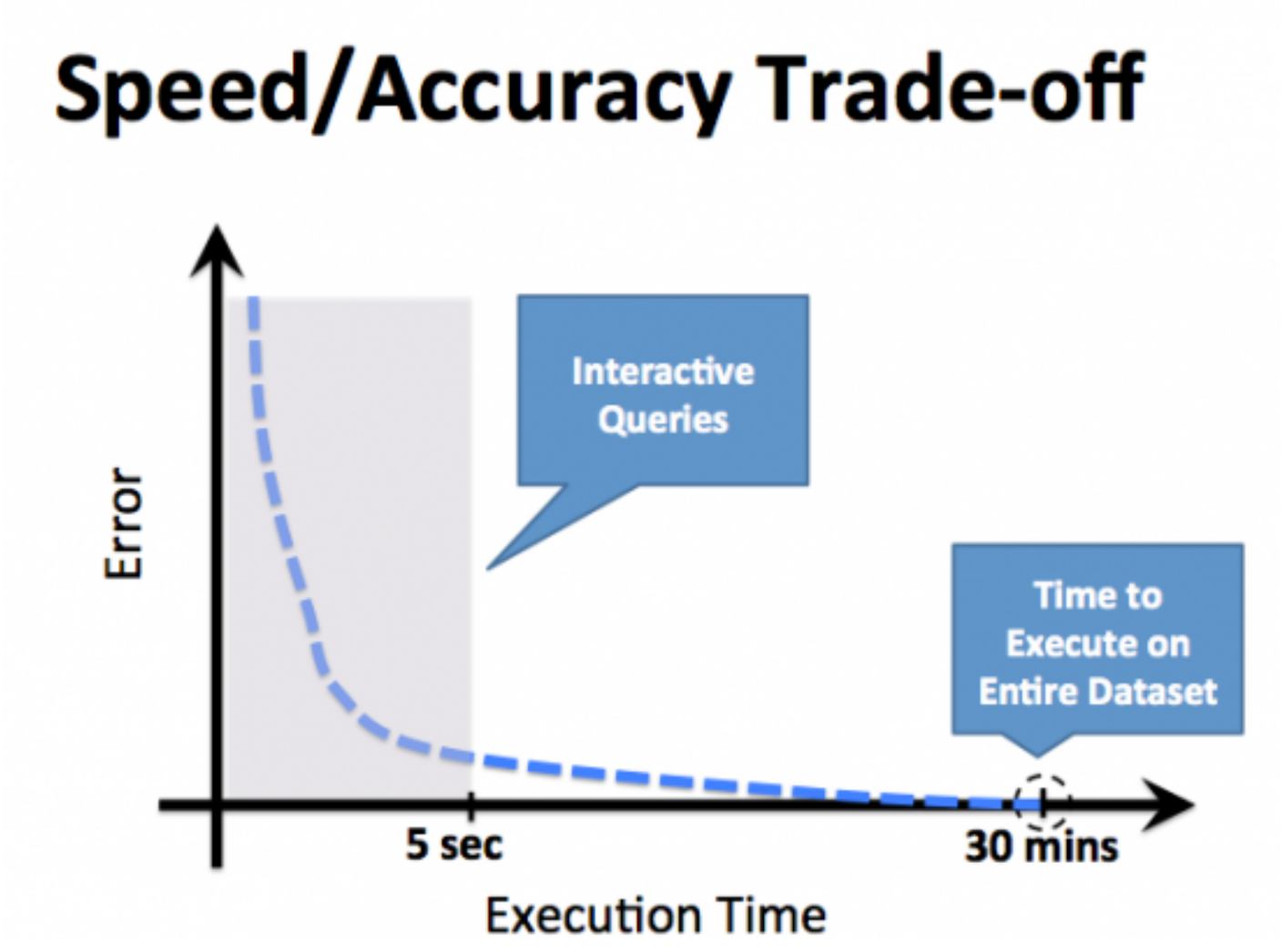
Section Structure

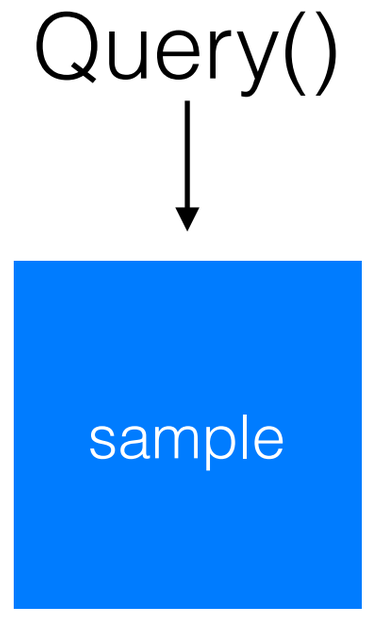
- Extended Data Cleaning Definition
- **Connecting Data Cleaning to Downstream Queries**
  - Aggregate queries
  - Machine learning training
  - Exploiting Relational Information

Section Structure

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

Aggregate Queries "Concentrate"





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

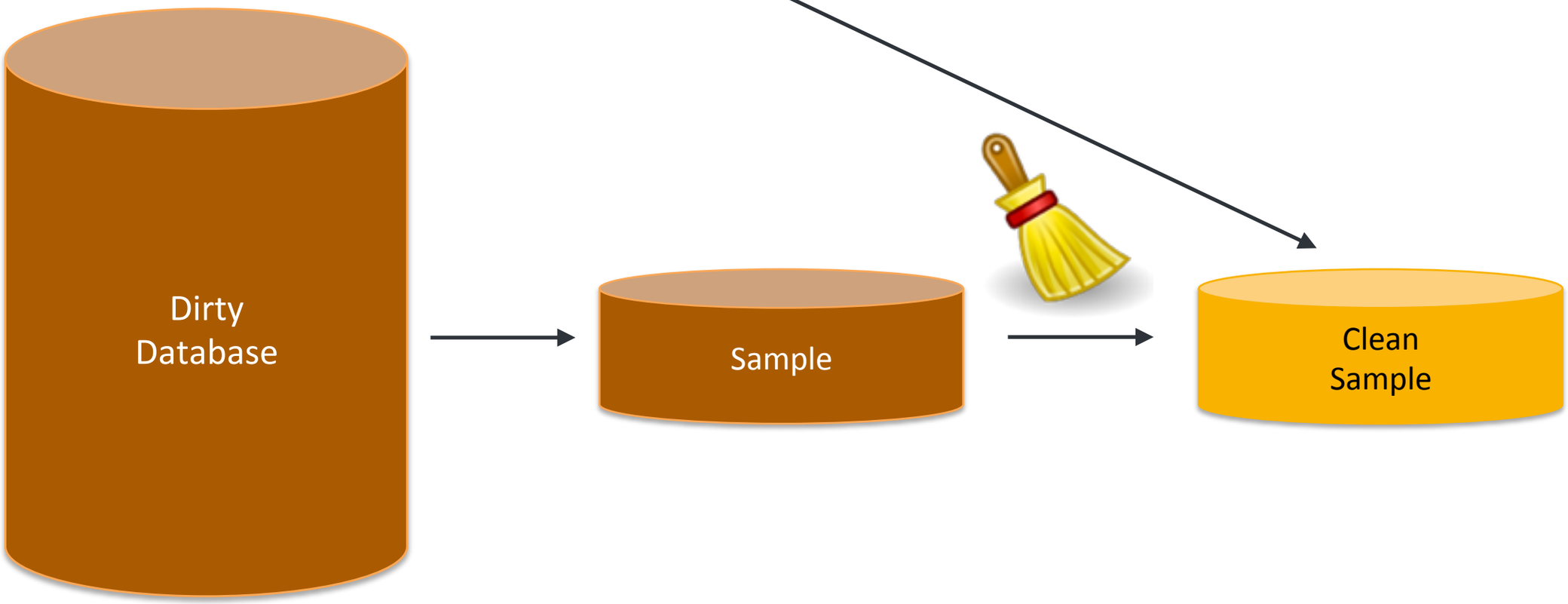
Rakesh Agrawal
Microsoft
Publications: 353 | Citations: 33537
Fields: Databases, Data Mining, World Wide Web
Collaborated with 365 co-authors from 1982 to 2012 | Cited by 24220 authors

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

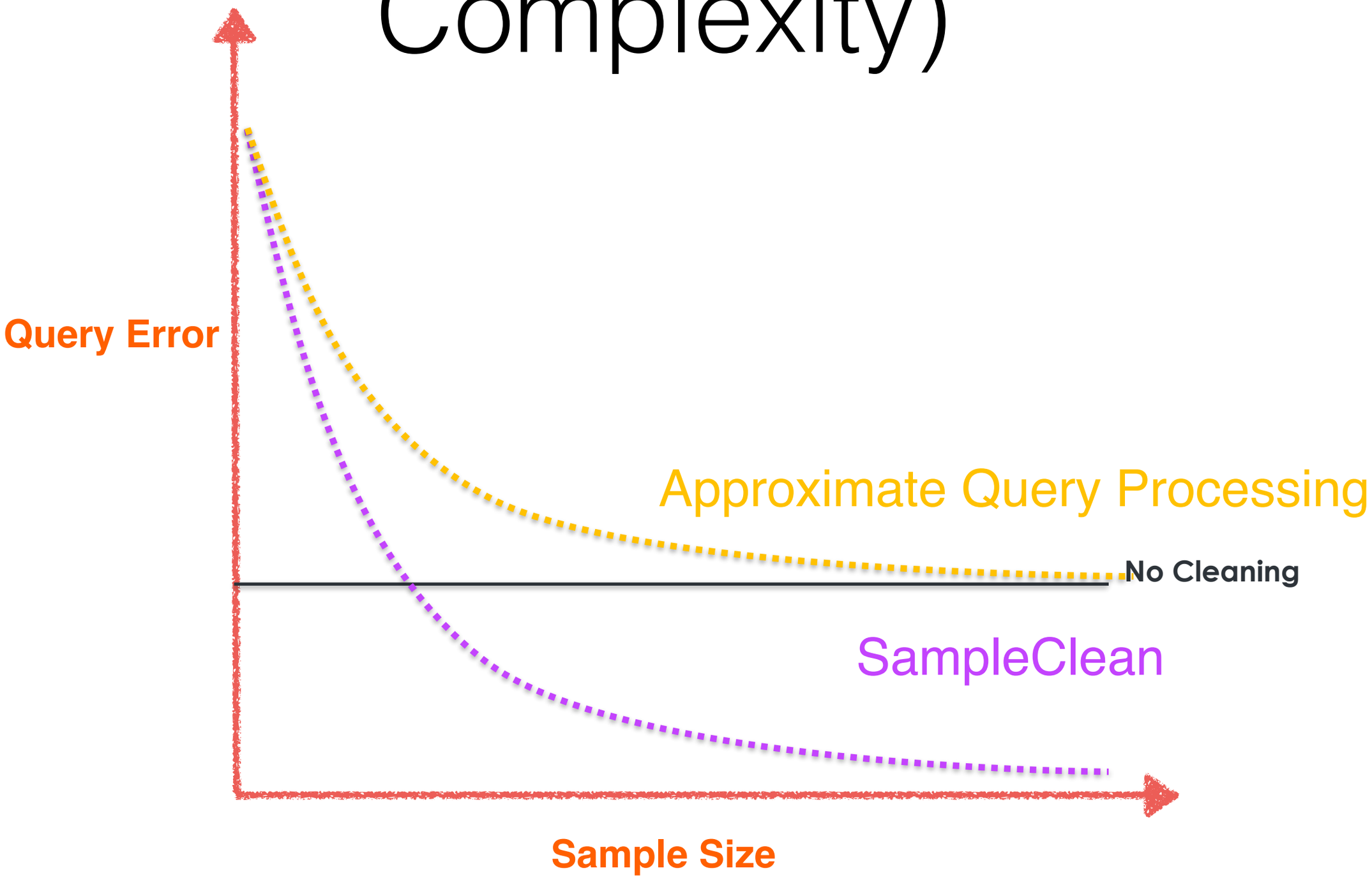
Sample-and-Clean

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)

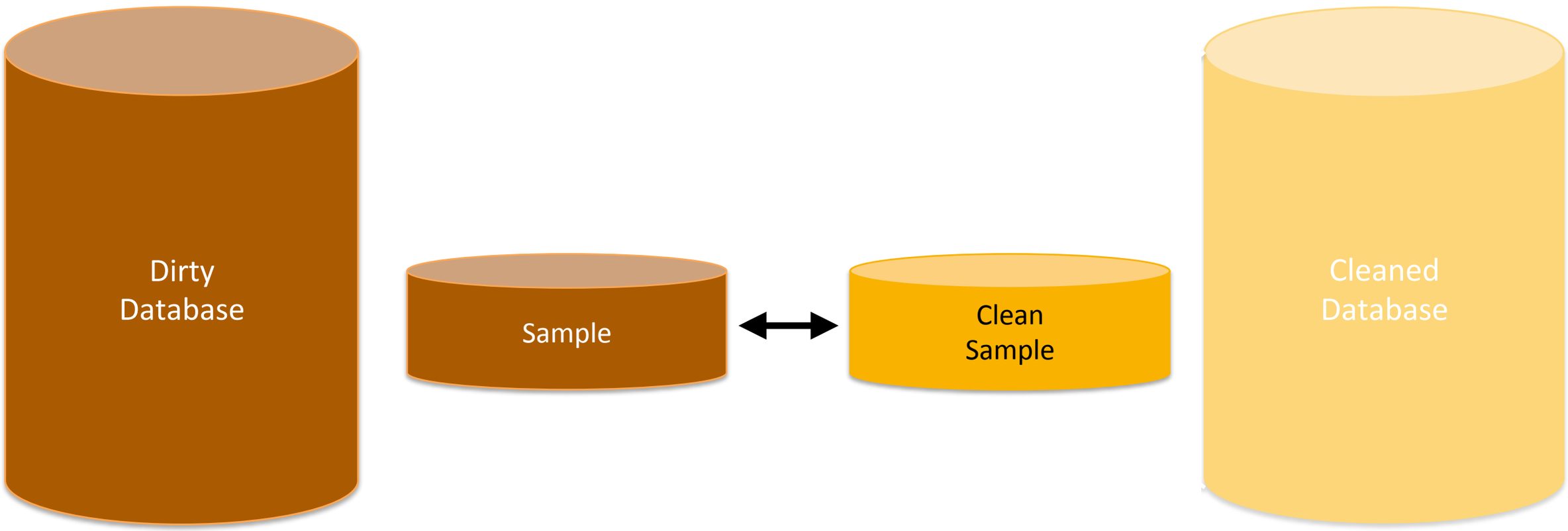
Probabilistic Interpretation

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

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

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

Transform Dirty Sample to Simulate Clean Sample



Algorithm 1: Direct Estimate

Approximate Queries

Query

Biased Answer

Samples

Direct Estimate

Query

Answer

Samples

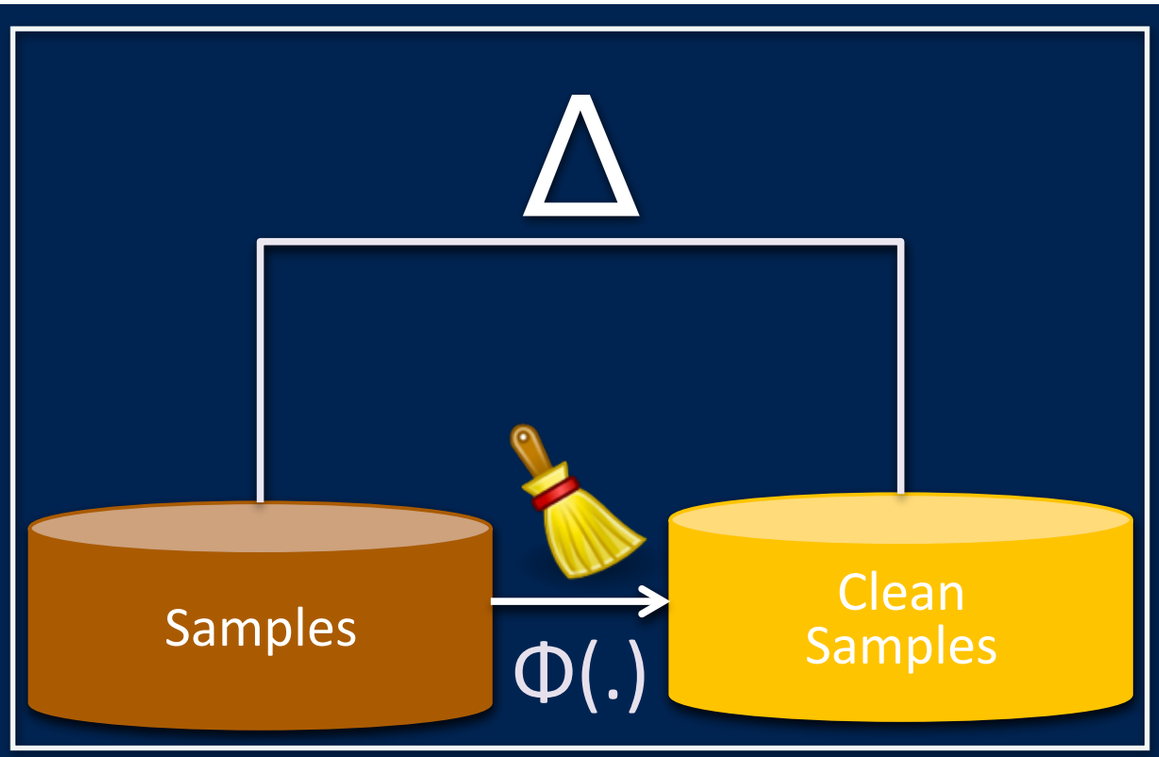
Φ(.)

Clean Samples

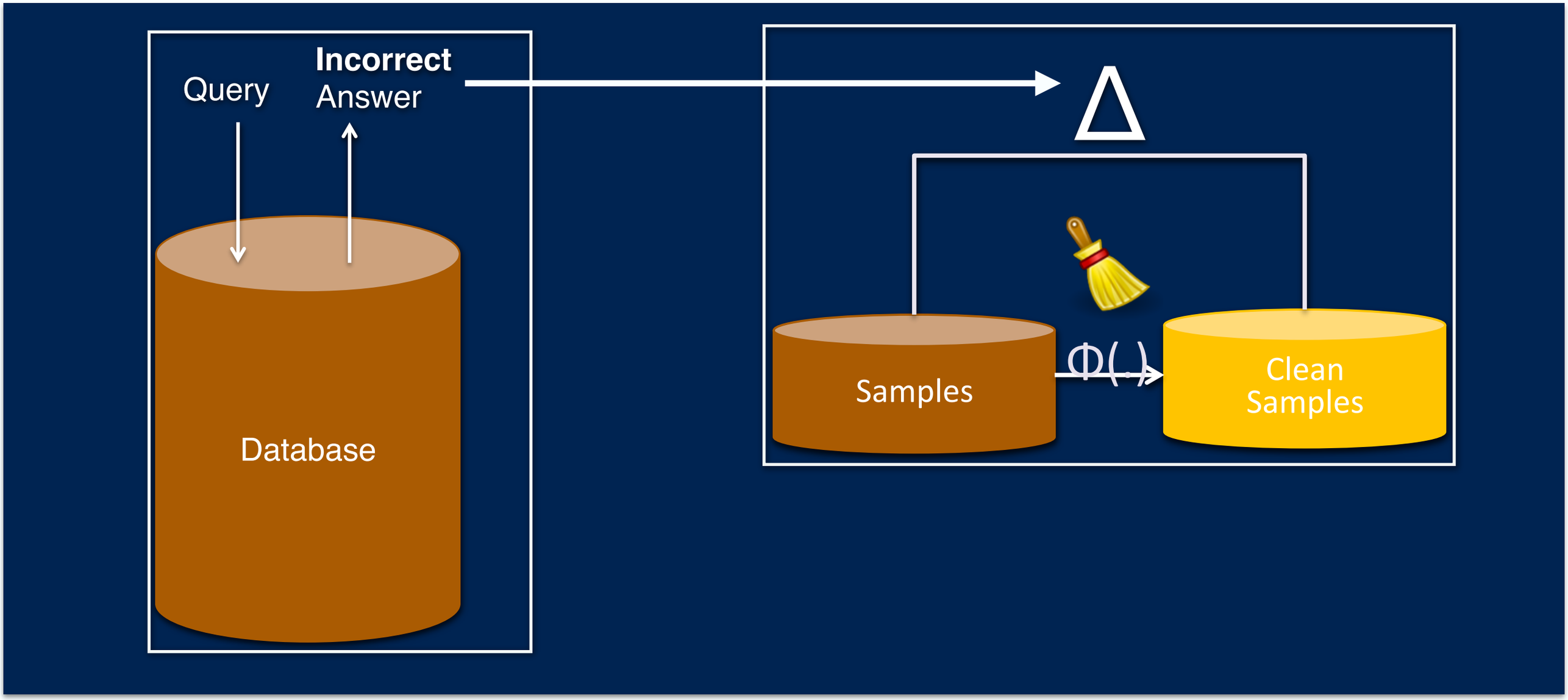
Jiannan Wang, Sanjay Krishnan, Michael Franklin, Ken Goldberg, Tim Kraska, Tova Milo. A Sample-and-Clean Framework for Fast and Accurate Query Processing on Dirty Data. In SIGMOD 2014

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

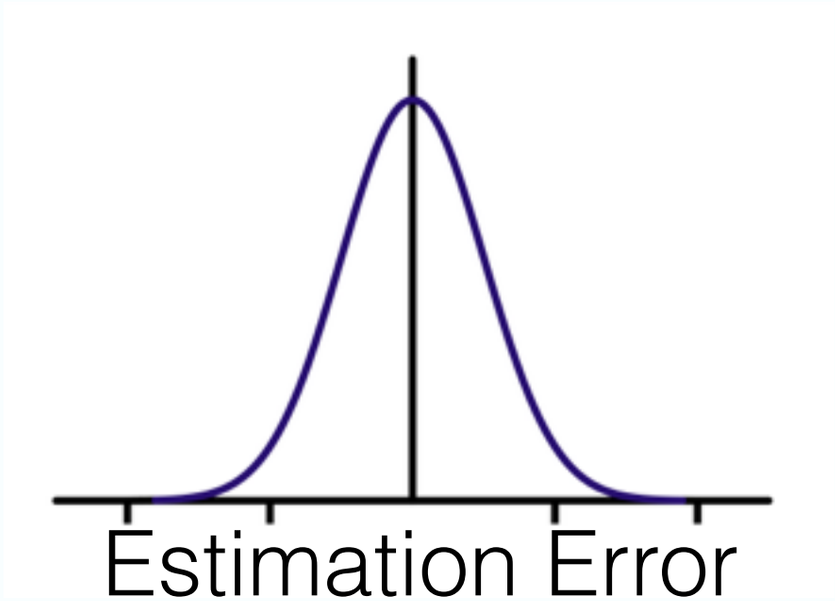
Two Types of Bounds

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

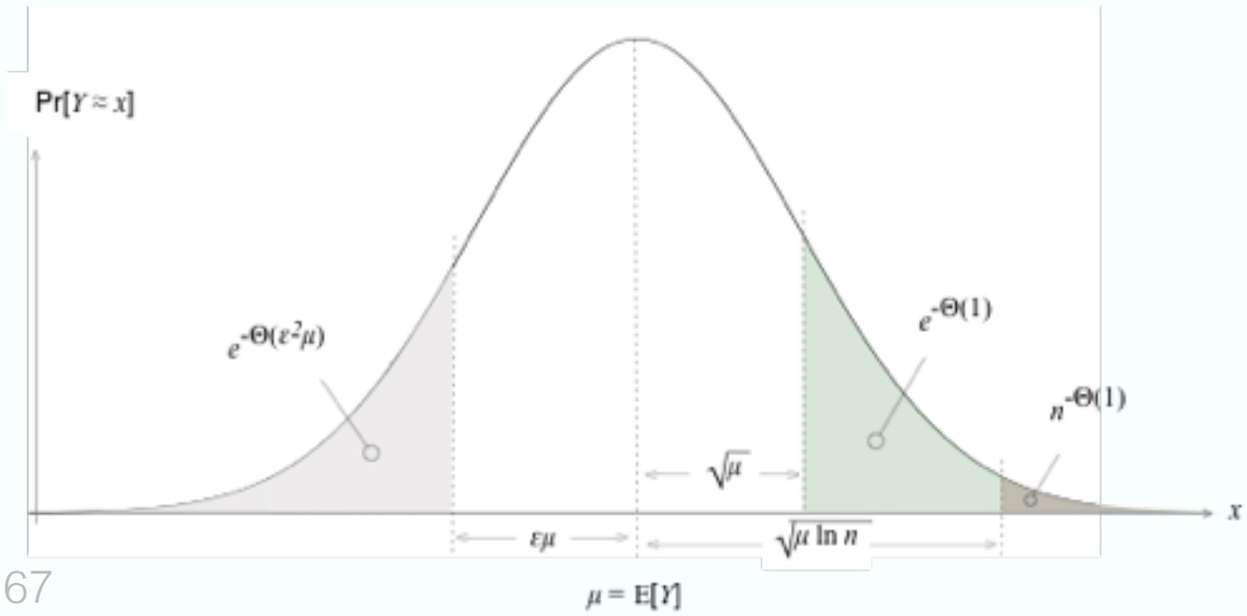
$$\mathbf{X} = \{X_1, X_2, ..., X_k\} \text{ i.i.d}$$

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

CLT



Chernoff



Central Limit Theorem

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

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

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

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

Unbiased With Bounds

Direct vs. Corrections
Asymptotic SUM/COUNT/AVG

	Clean Estimate	**Dirty Correction**
Accuracy	$O(\frac{var(clean)}{k})$	$O(\frac{var(diff)}{k})$
Runtime	$O(k)$	$O(N)$

FPC: sqrt(N-k)/sqrt(N-1)

Chernoff Bound

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

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

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

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

Direct vs. Corrections
Finite Sample SUM/COUNT/AVG

	Clean Estimate	**Dirty Correction**
Accuracy	$O(\frac{\text{range}(clean)}{k})$	$O(\frac{\text{range}(diff)}{k})$
Runtime	$O(k)$	$O(N)$

range(S) = max(S)-min(S)

Example Query

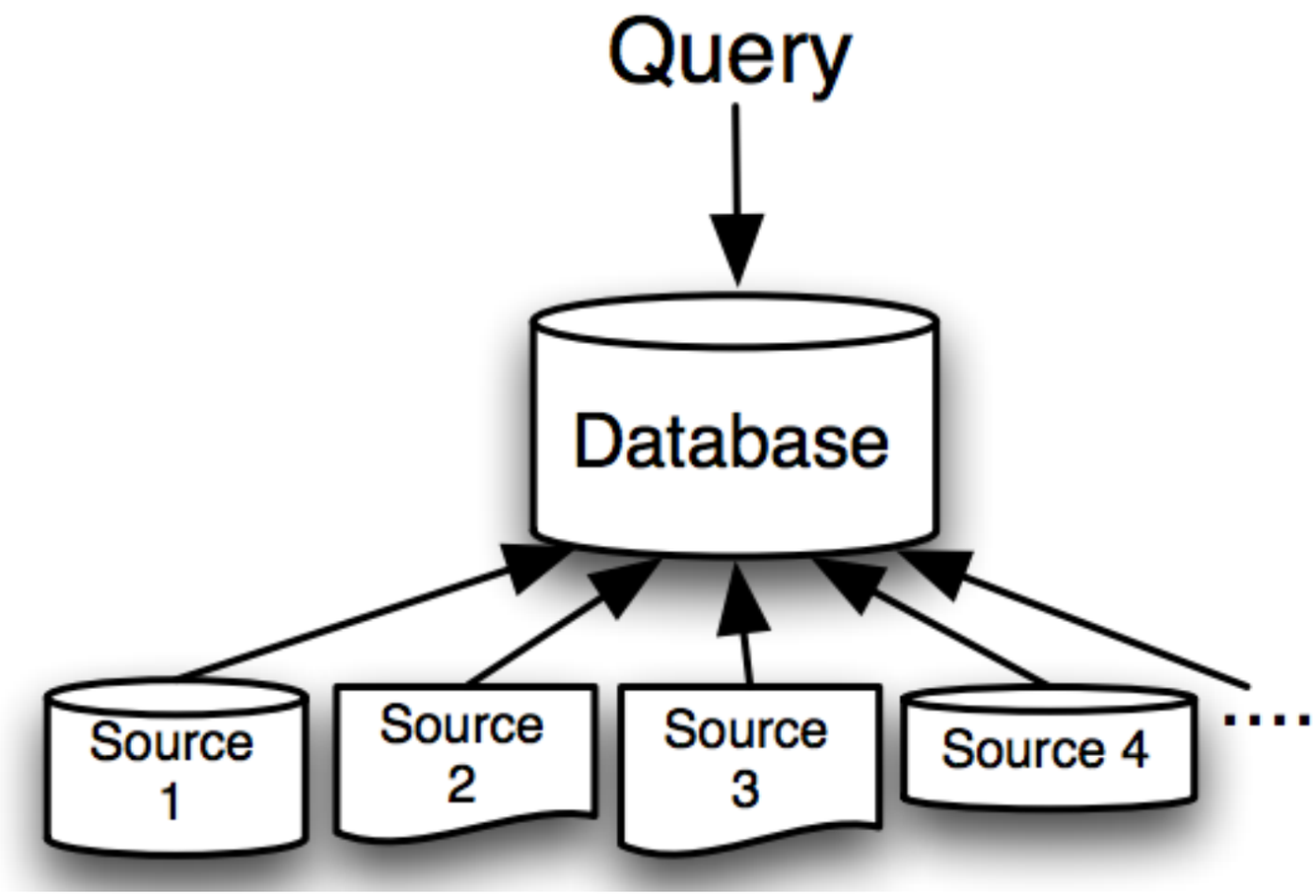
- Count the number of distinct publications

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

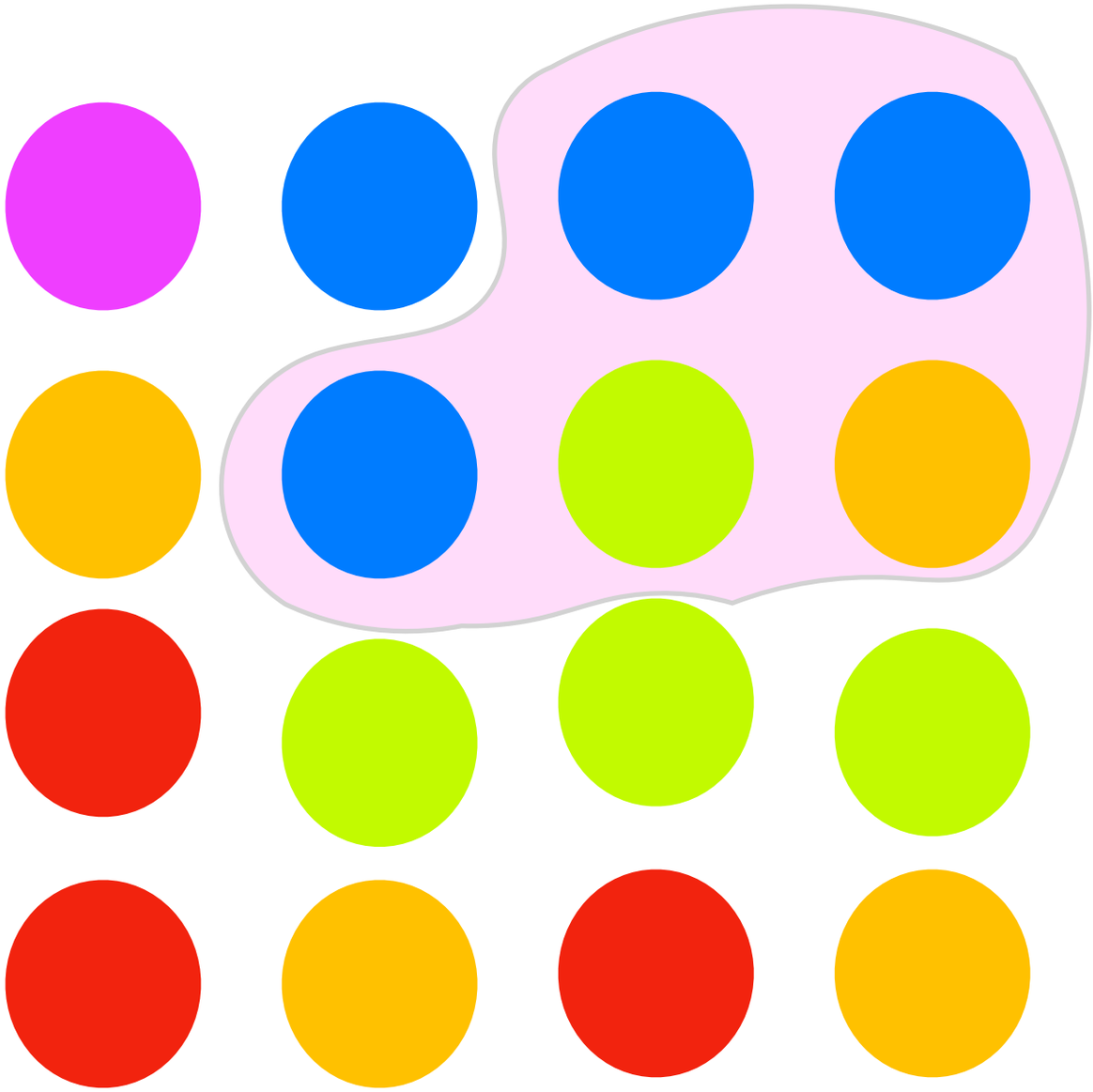
Michael Franklin
University of California Berkeley
Publications: 561 | Citations: 15174
Fields: Databases, Pharmacology, Data Mining
Collaborated with 3451 co-authors from 1974 to 2012 | Cited by 15795 authors

Estimating Unknown Unknowns

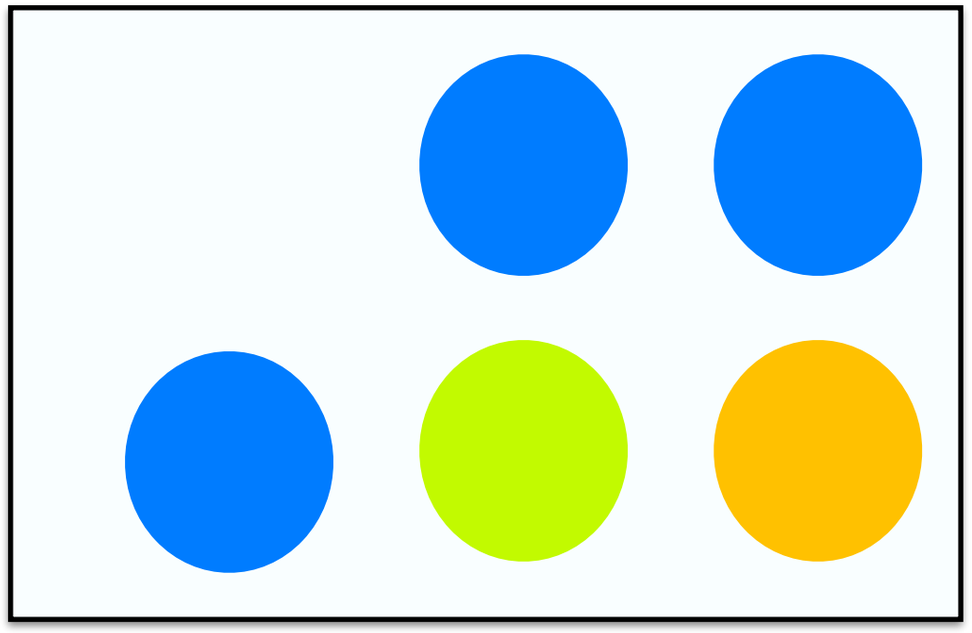


Chung, Y., Mortensen, M.L., Binnig, C. and Kraska, T., Estimating the impact of unknown unknowns on aggregate query results. SIGMOD 2016.

Abstract Problem

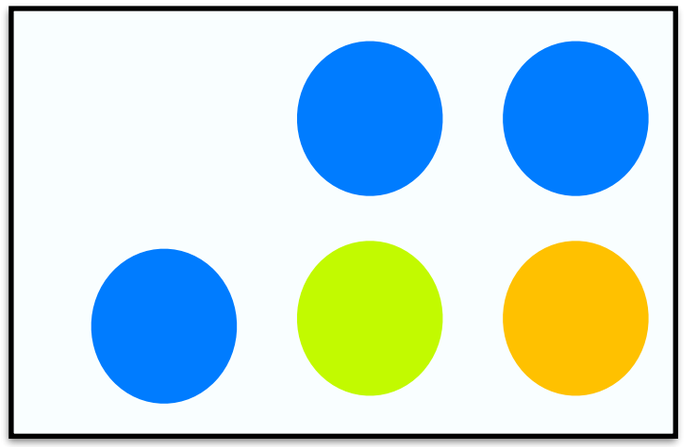


What is the distinct count?



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: Observed*1/sample
  - 16/5 * 3 = 9.6
- Good-Turing Estimate: Observed/(1-f1/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 - f1/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

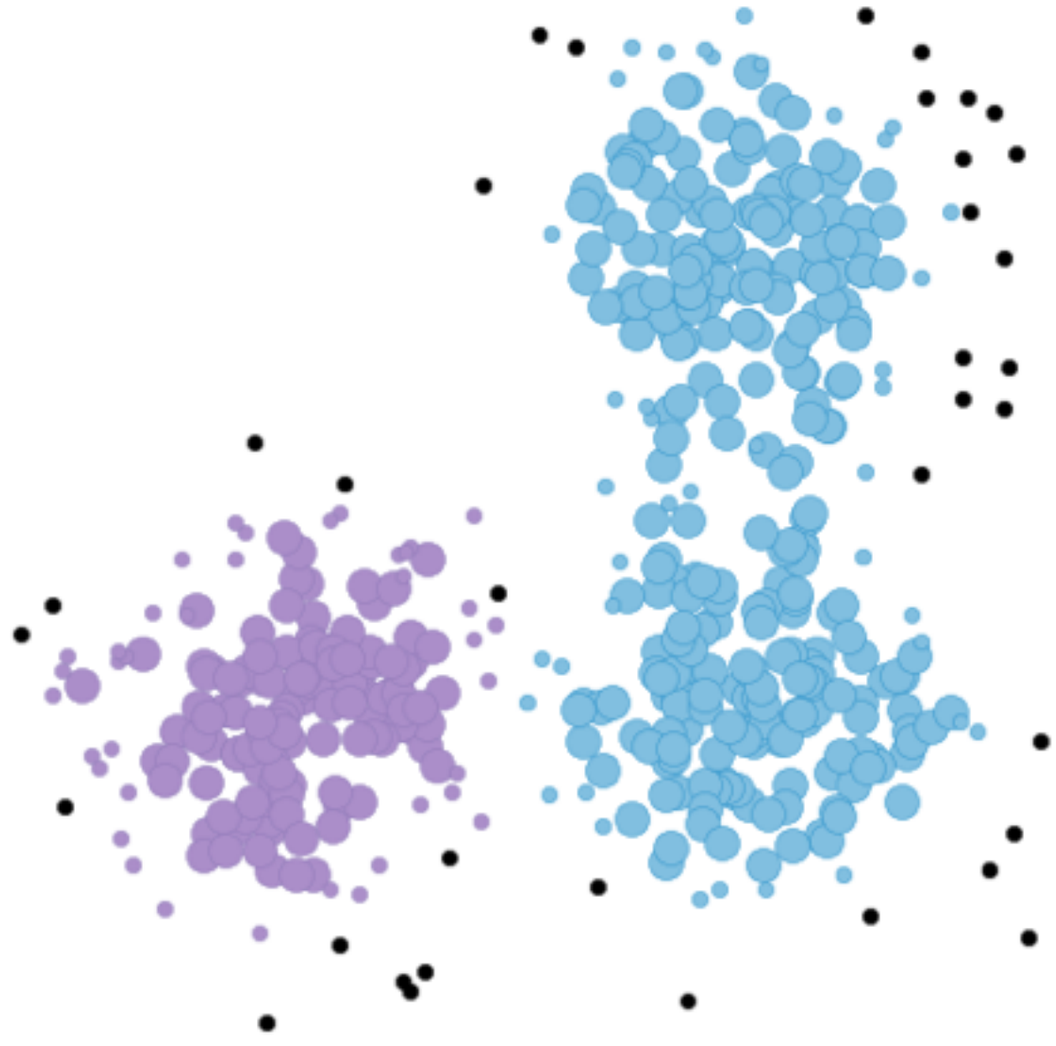
Example

Cluster Publications From Rakesh and Mike

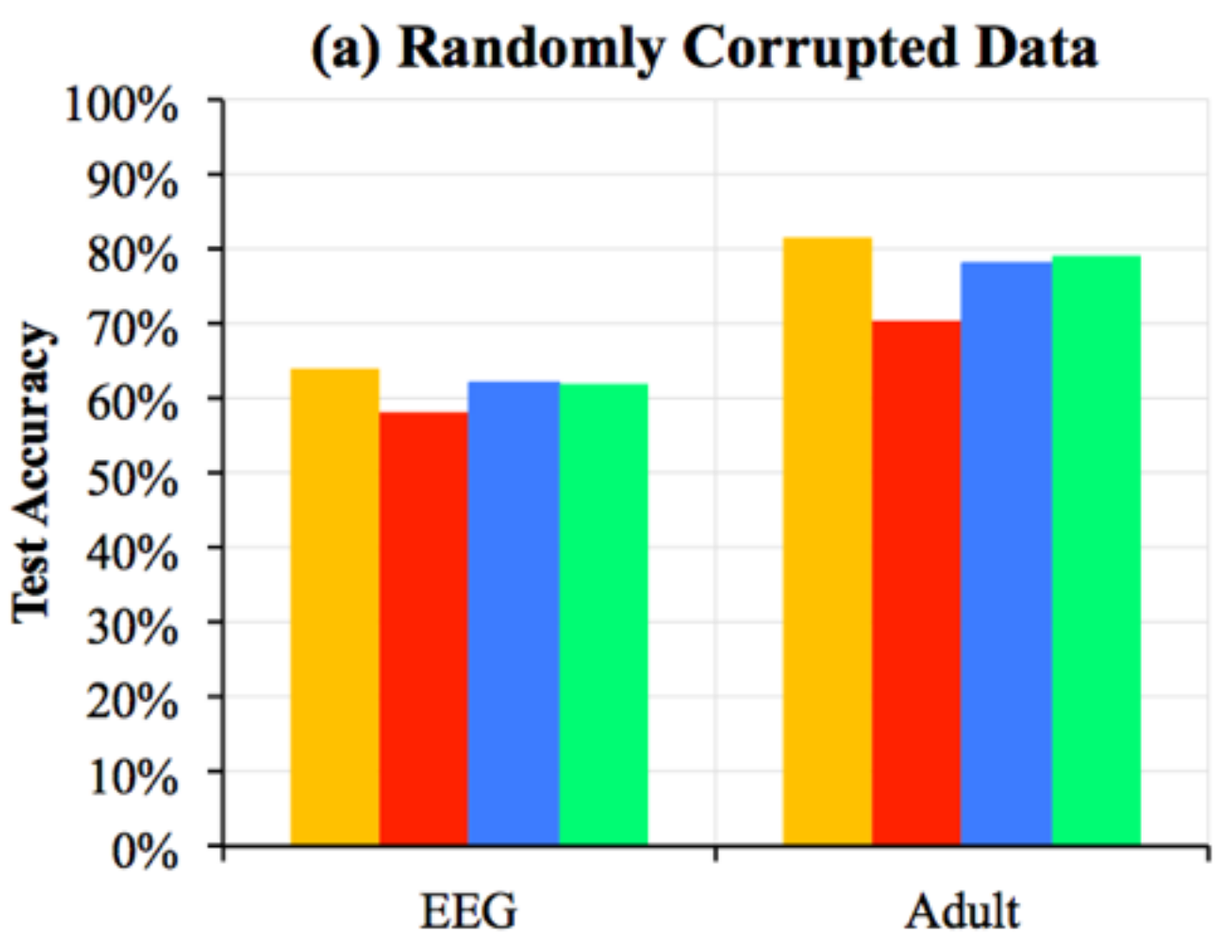
Rakesh Agrawal
Microsoft
Publications: 353 | Citations: 33537
Fields: Databases, Data Mining, World Wide Web
Collaborated with 365 co-authors from 1982 to 2012 | Cited by 24220 authors

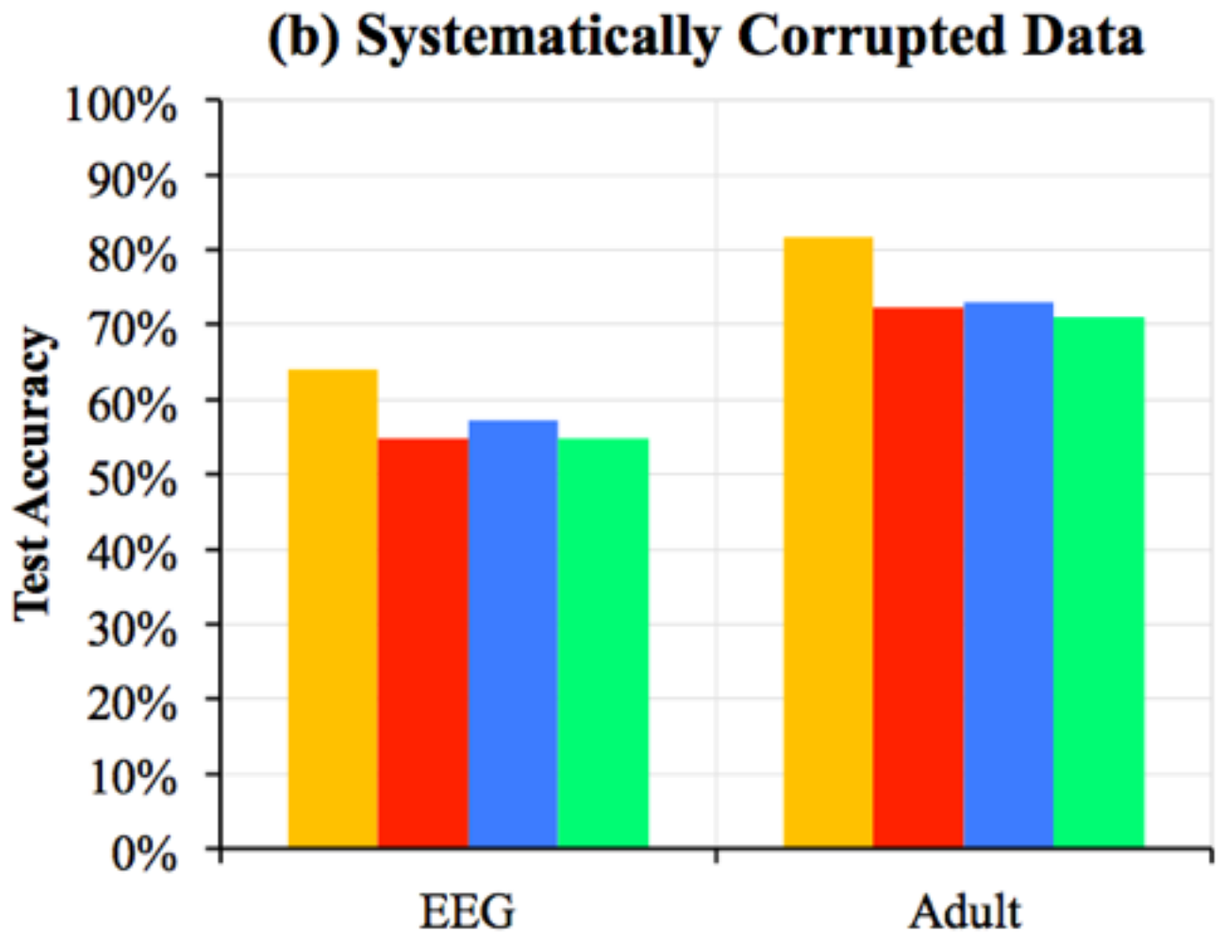
Jeffrey D. Ullman
Stanford University
Publications: 460 | Citations: 43431
Fields: Databases, Algorithms & Theory, Scientific Computing
Collaborated with 317 co-authors from 1961 to 2012 | Cited by 31987 authors

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

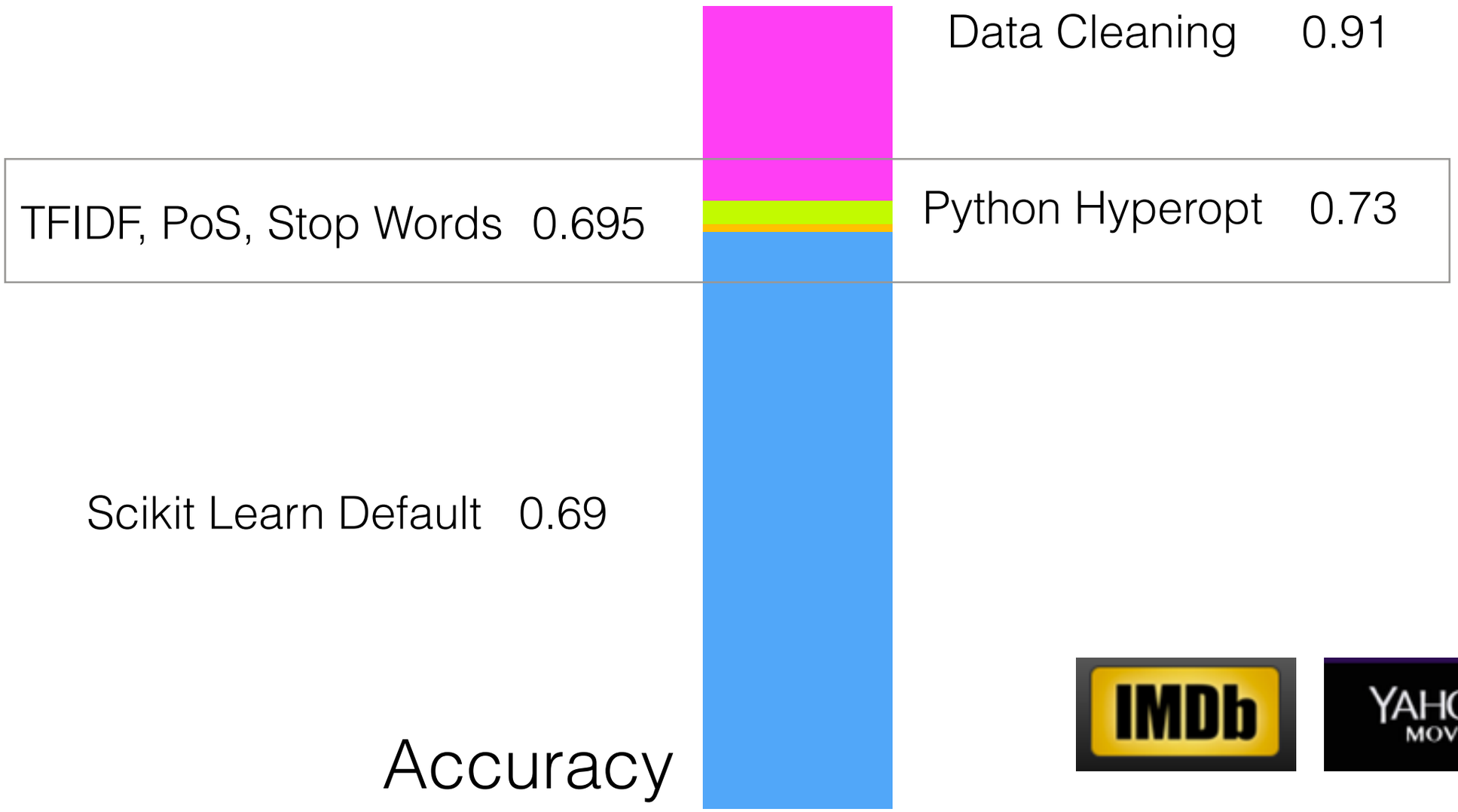


Misconception 1: ML models are robust to error



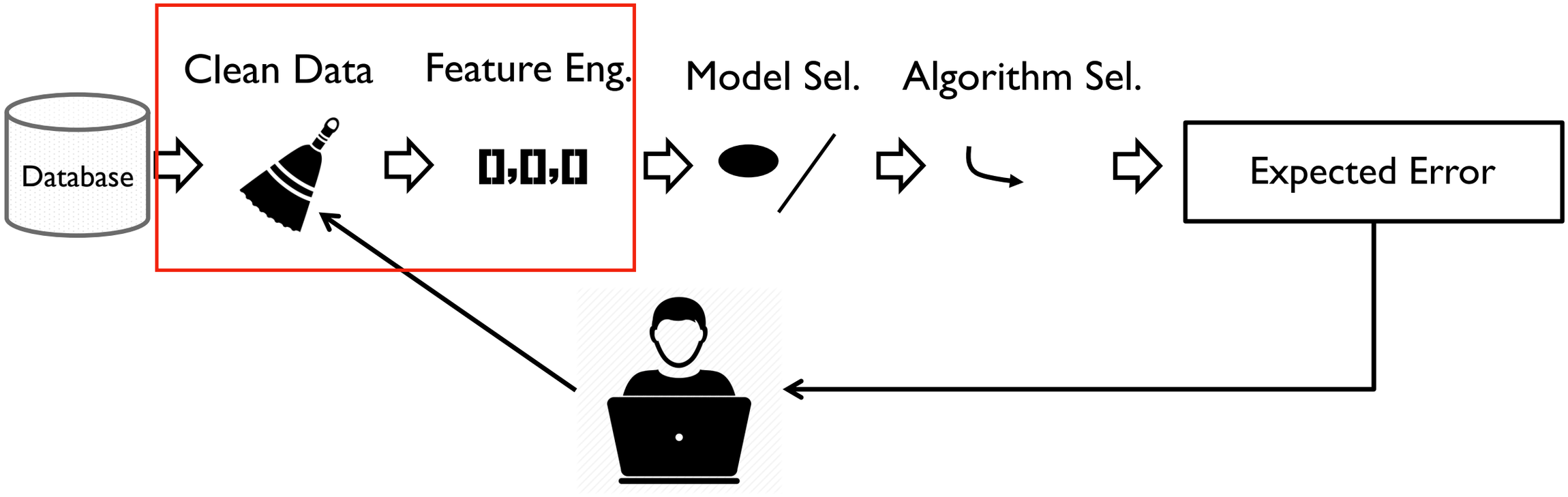


Misconception 2. Parameter tuning is the most important problem



Horror vs. Comedy From Plot

Data Cleaning Before Machine Learning

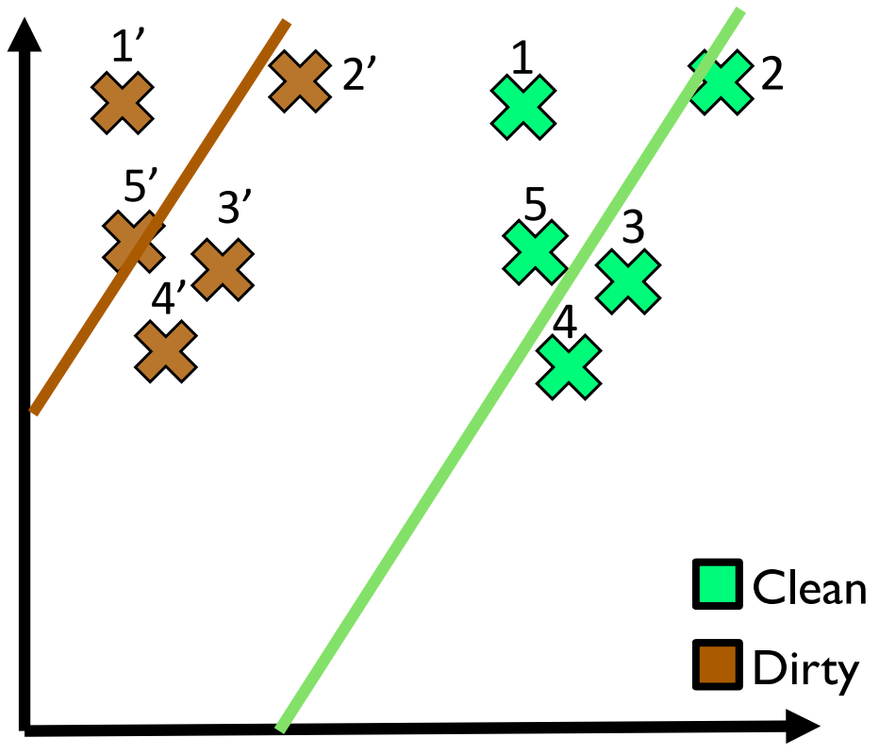


Correctness: Does data cleaning affect the convergence?

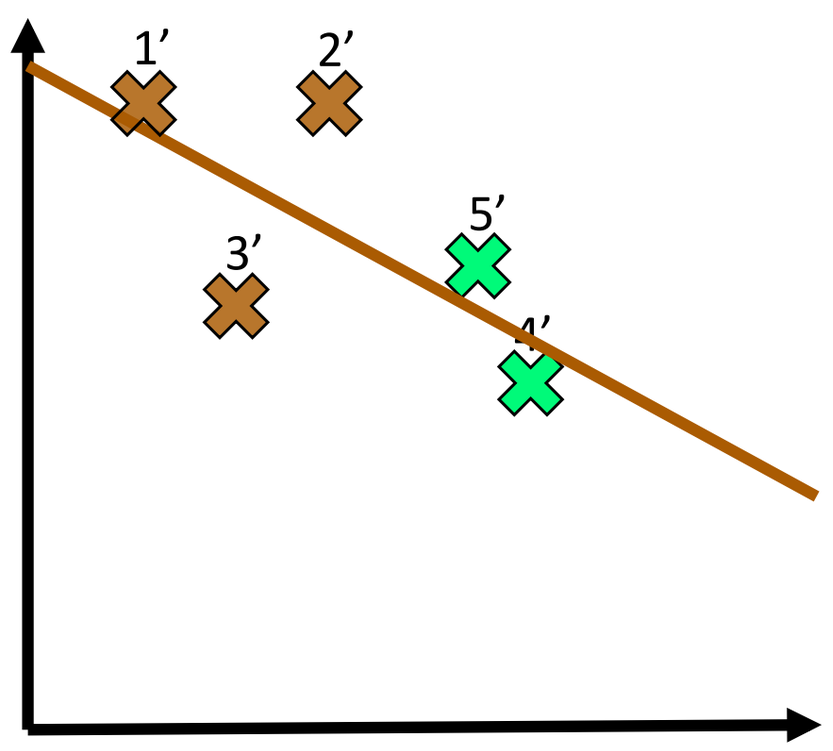
Efficiency: How to use the model to identify dirty data?

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

Simpson's Paradox



Data



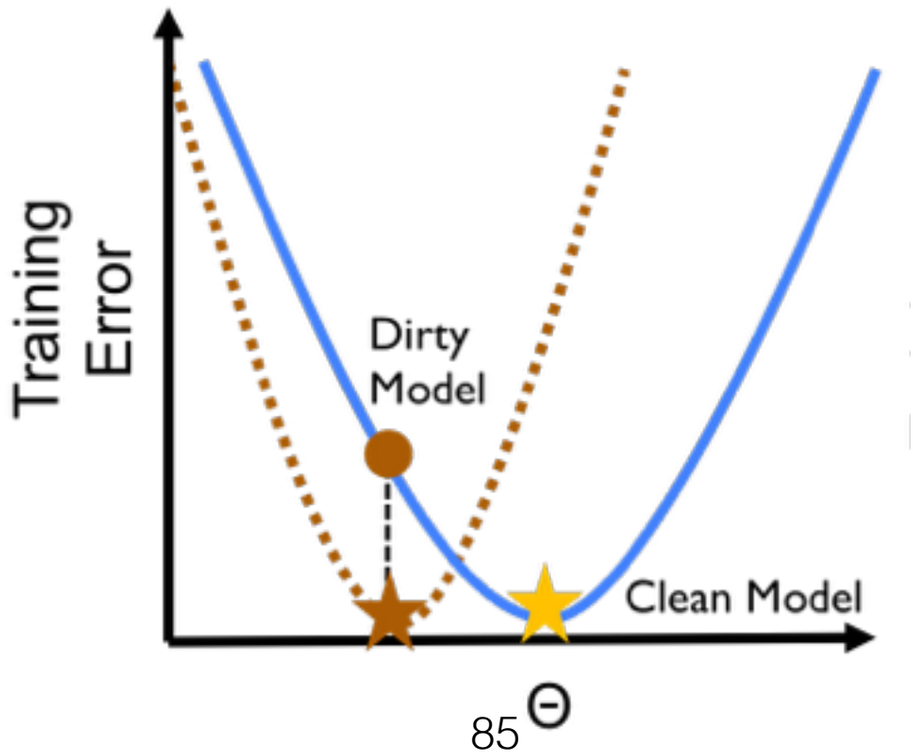
Partial Data Cleaning

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.

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

- Just an estimated average query!

- Make each step unbiased

Active Clean Algorithm

- Train a preliminary model on the dirty dataset
- For i in {0, ….., 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:

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

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

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

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

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

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

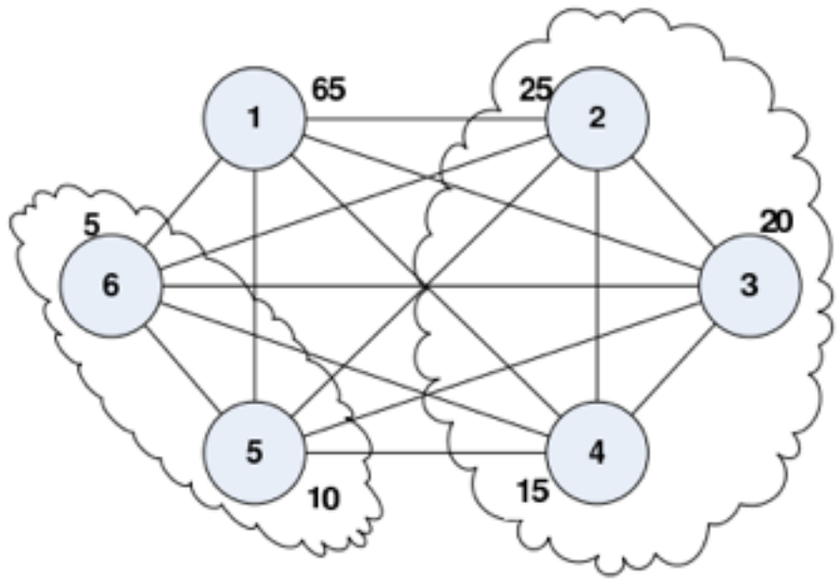
Query-Driven Entity Resolution

- Do minimal ER work to answer a query
- Defines a concept of *vestigiality* (answer is still correct without knowing whether a pairs 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

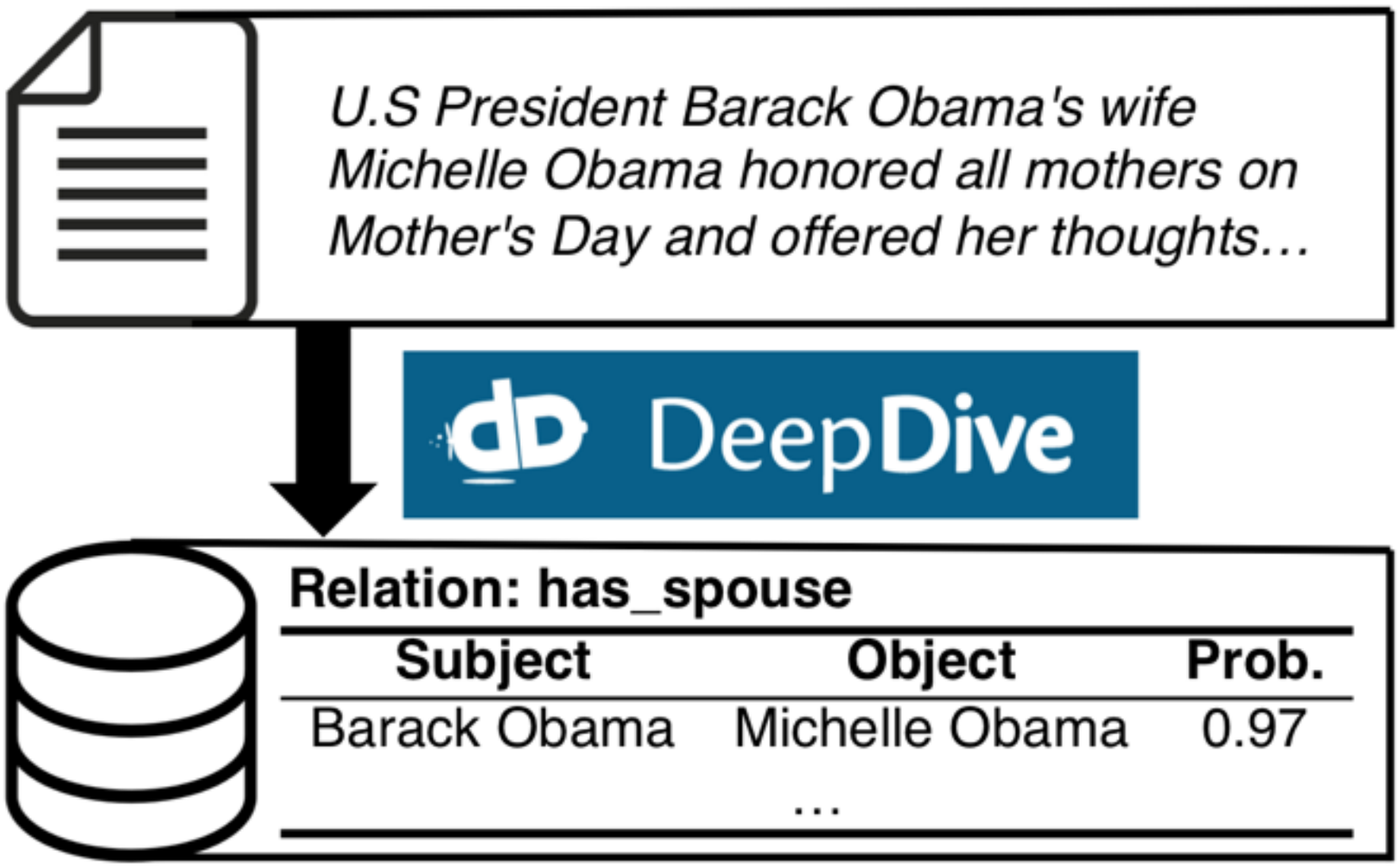
- Oracle crowds to derive database edits for removing (adding) incorrect (missing) tuples to the result of a query.
- For a given query derive a set of cleaning updates to base data to ensure completeness.

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

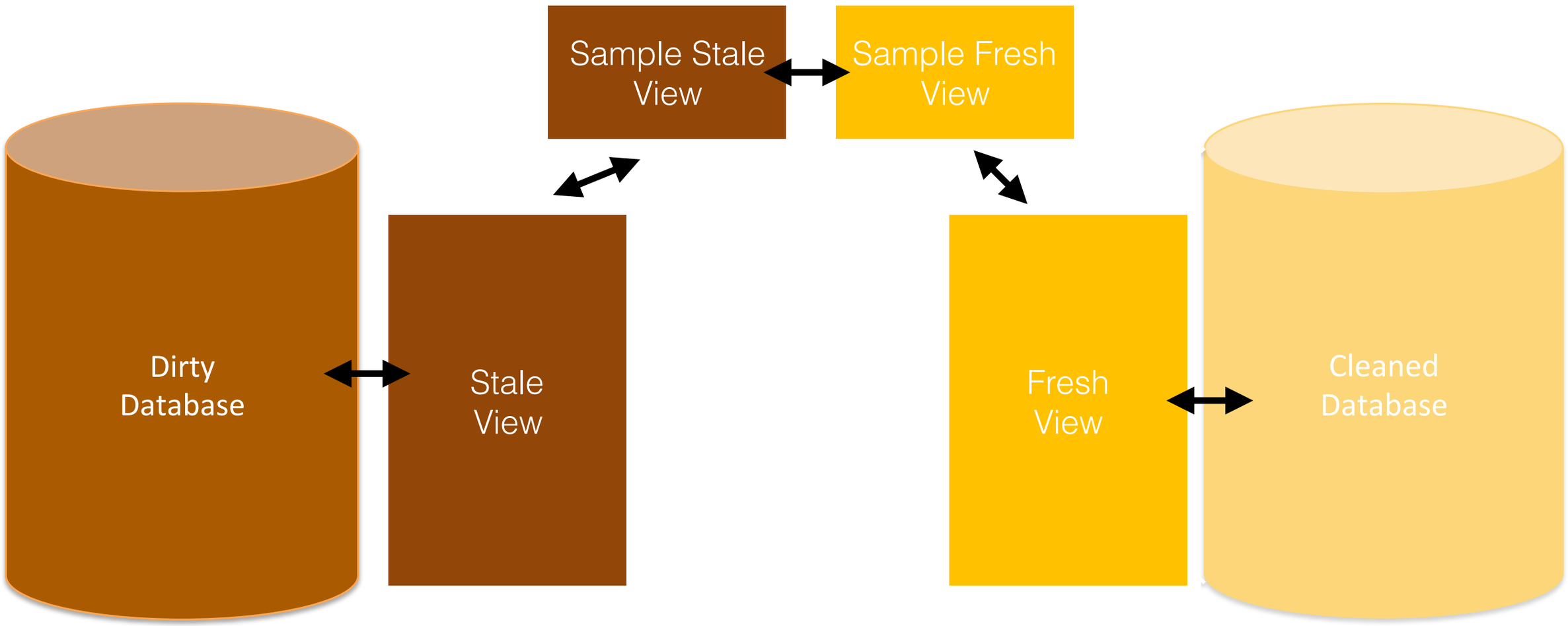
- Topic 1. Statistical techniques to clean data (20 mins)
- Topic 2. Cleaning data before statistical analytics (50 min)
- **Topic 3. Impact and Future Directions (10 mins)**

Data Cleaning+ Knowledge Bases



Ce Zhang DeepDive: A Data Management System for Automatic Knowledge Base Construction. PhD Thesis

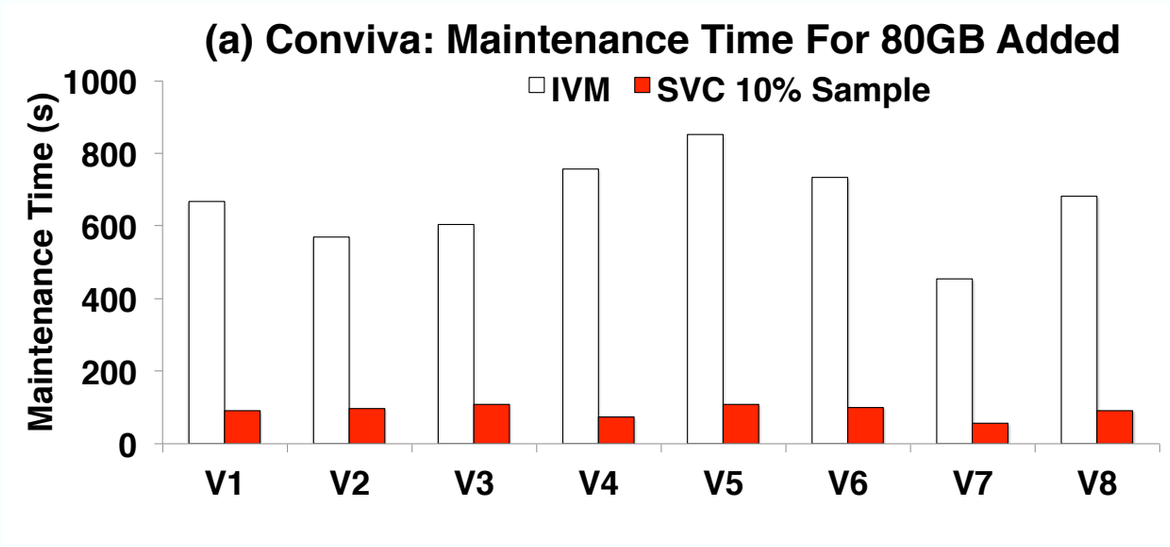
Queries on Stale Views

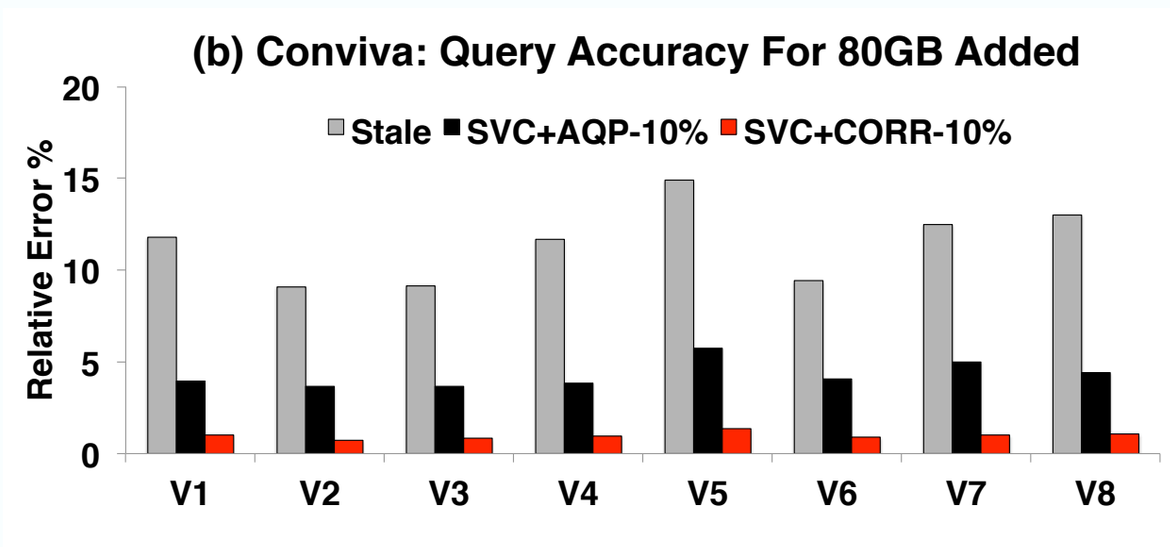


Krishnan, Sanjay, et al. "Stale view cleaning: Getting fresh answers from stale materialized views." Proceedings of the VLDB Endowment 8.12 (2015):

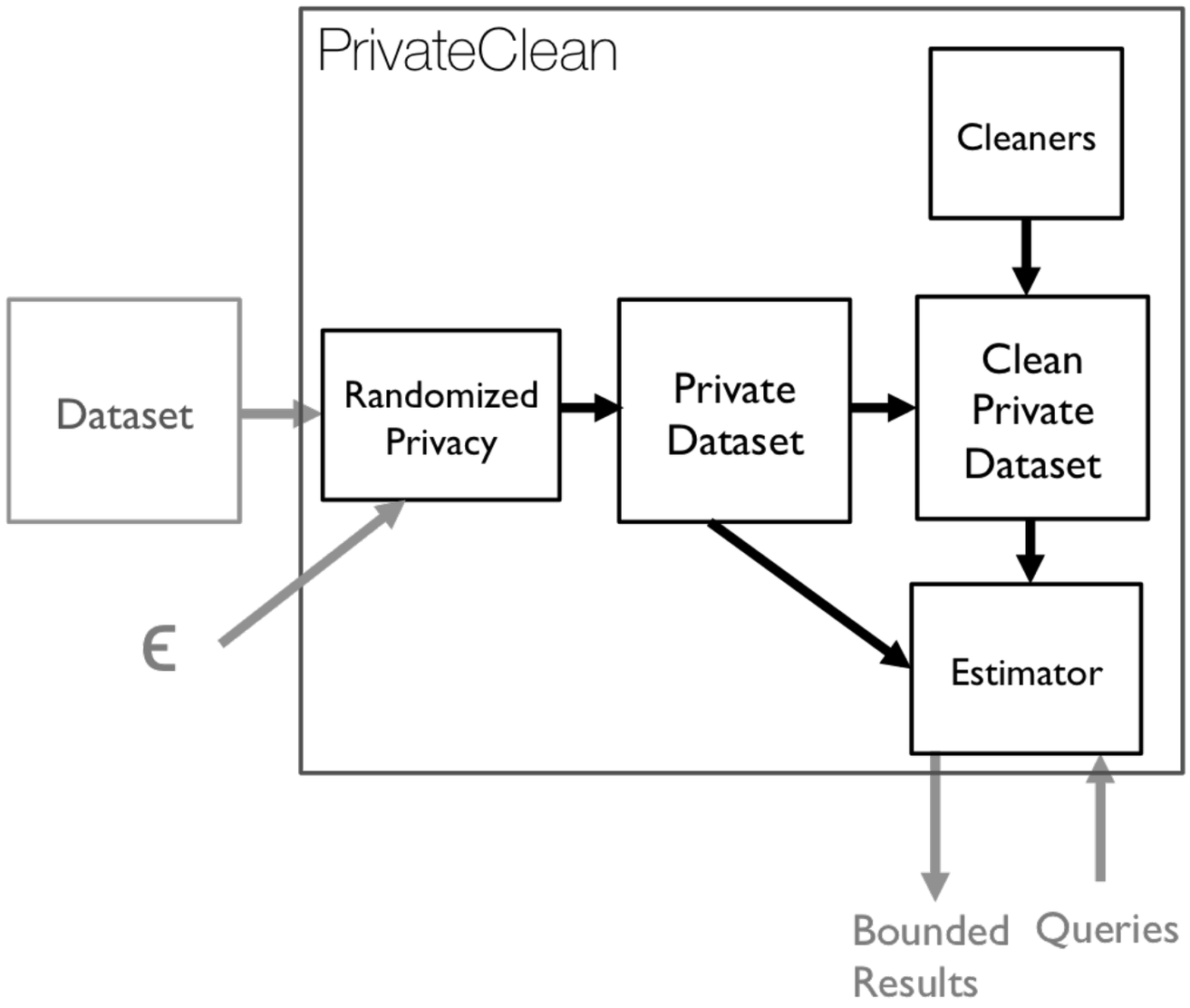
Conviva: Log Analysis

- 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

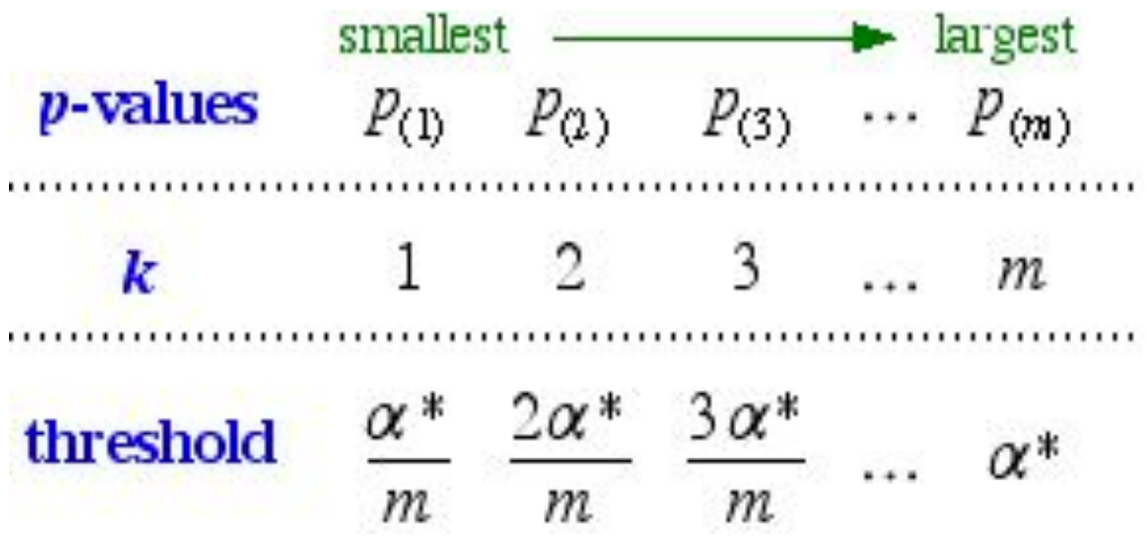
"We test thousands of new treatments each year, so to avoid multiple testing issues we always do a validation experiment to confirm our positive results".

How often do those work out?

About 5% of the time!

Concerns

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



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