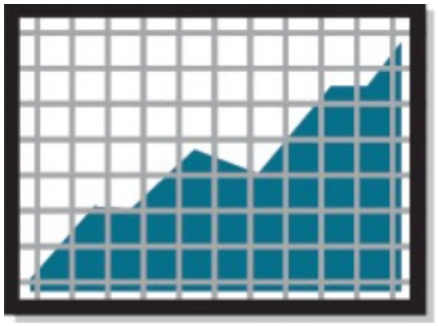
Lenses: An On-Demand Approach to ETL

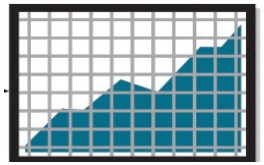
Ying Yang[+], Niccolo Meneghetti[+], Ronny Fehling*,Zhen Hua Liu*, Oliver Kennedy[+]
[+] SUNY Buffalo, * Oracle

{yyang25, niccolom, okennedy}@buffalo.edu
{ronny.fehling, zhen.liu}@oracle.com

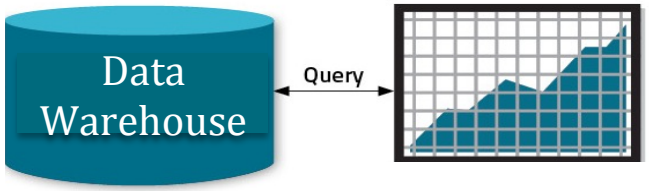
Efficient analytics depends on *accurate, reliable, high-quality* information.



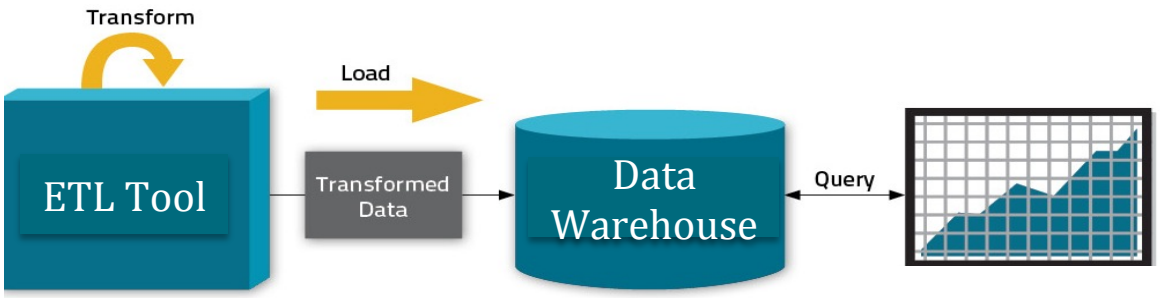
Efficient analytics depends on *accurate, reliable, high-quality* information.



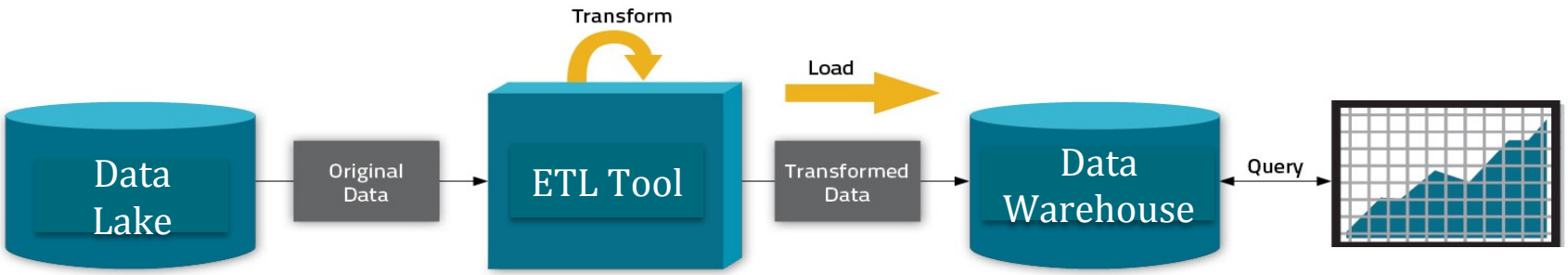
Efficient analytics depends on *accurate, reliable, high-quality* information.



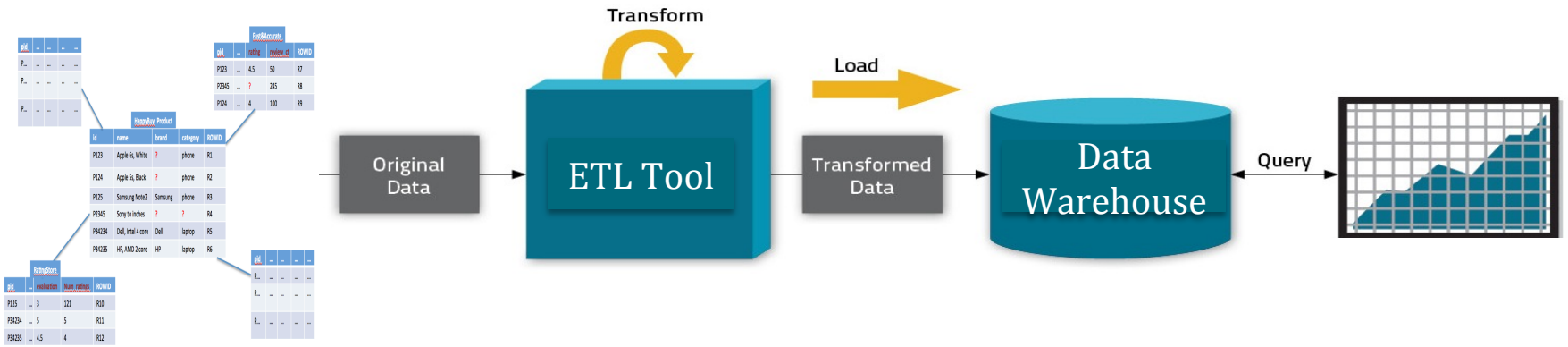
Efficient analytics depends on *accurate, reliable, high-quality* information.



Efficient analytics depends on *accurate, reliable, high-quality* information.



But, raw data are messy.



The data is very messy:

- *Product: missing data.*
- *Rating1: missing data.*
- *Rating1 and Rating2: different schemas.*

Mobile Application: Rating2

pid	...	evaluation	Num_ratings	ROWID
P125	...	3	121	R10
P34234	...	5	5	R11
P34235	...	4.5	4	R12

Survey: Rating1

pid	...	rating	review_ct	ROWID
P123	...	4.5	50	R7
P2345	...	NULL	245	R8
P124	...	4	100	R9

HappyBuy: Product

id	name	brand	category	ROWID
P123	Apple 6s, White	NULL	phone	R1
P124	Apple 5s, Black	NULL	phone	R2
P125	Samsung Note2	Samsung	phone	R3
P2345	Sony to inches	NULL	NULL	R4
P34234	Dell, Intel 4 core	Dell	laptop	R5
P34235	HP, AMD 2 core	HP	laptop	R6

The data is very messy:

- *Product: missing data.*

HappyBuy: Product

id	name	brand	category	ROWID
P123	Apple 6s, White	NULL	phone	R1
P124	Apple 5s, Black	NULL	phone	R2
P125	Samsung Note2	Samsung	phone	R3
P2345	Sony to inches	NULL	NULL	R4
P34234	Dell, Intel 4 core	Dell	laptop	R5
P34235	HP, AMD 2 core	HP	laptop	R6

The data is very messy:

- *Rating1: missing data.*

Mobile Application: Rating2

pid	...	evaluation	Num_ratings	ROWID
P125	...	3	121	R10
P34234	...	5	5	R11
P34235	...	4.5	4	R12

Survey: Rating1

pid	...	rating	review_ct	ROWID
P123	...	4.5	50	R7
P2345	...	NULL	245	R8
P124	...	4	100	R9

The data is very messy:

- *Rating1 and Rating2: different schemas.*

Mobile Application: Rating2

pid	...	evaluation	Num_ratings	ROWID
P125	...	3	121	R10
P34234	...	5	5	R11
P34235	...	4.5	4	R12

Survey: Rating1

pid	...	rating	review_ct	ROWID
P123	...	4.5	50	R7
P2345	...	NULL	245	R8
P124	...	4	100	R9

The data is very messy:

- *Product: missing data.*
- *Rating1: missing data.*
- *Rating1 and Rating2: different schemas.*

Mobile Application: Rating2

pid	...	evaluation	Num_ratings	ROWID
P125	...	3	121	R10
P34234	...	5	5	R11
P34235	...	4.5	4	R12

Survey: Rating1

pid	...	rating	review_ct	ROWID
P123	...	4.5	50	R7
P2345	...	NULL	245	R8
P124	...	4	100	R9

HappyBuy: Product

id	name	brand	category	ROWID
P123	Apple 6s, White	NULL	phone	R1
P124	Apple 5s, Black	NULL	phone	R2
P125	Samsung Note2	Samsung	phone	R3
P2345	Sony to inches	NULL	NULL	R4
P34234	Dell, Intel 4 core	Dell	laptop	R5
P34235	HP, AMD 2 core	HP	laptop	R6

The clean data

AllRatings

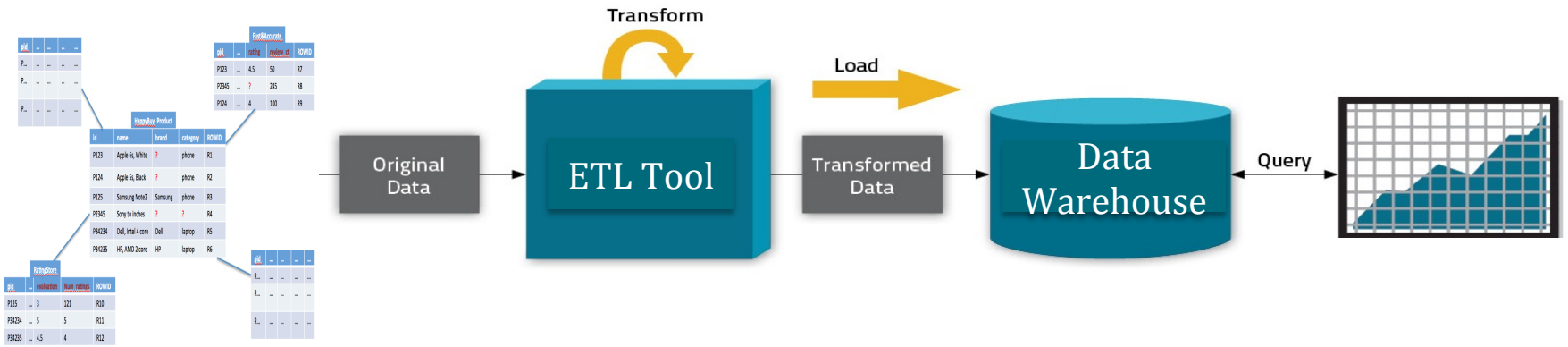
pid	...	rating	review_ct	ROWID
P125	...	3	121	R10
P34234	...	5	5	R11
P34235	...	4.5	4	R12
P123	...	4.5	50	R7
P2345	...	5	245	R8
P124	...	4	100	R9

HappyBuy: Product

id	name	brand	category	ROWID
P123	Apple 6s, White	Apple	phone	R1
P124	Apple 5s, Black	Apple	phone	R2
P125	Samsung Note2	Samsung	phone	R3
P2345	Sony to inches	Sony	TV	R4
P34234	Dell, Intel 4 core	Dell	laptop	R5
P34235	HP, AMD 2 core	HP	laptop	R6

Upfront cleaning

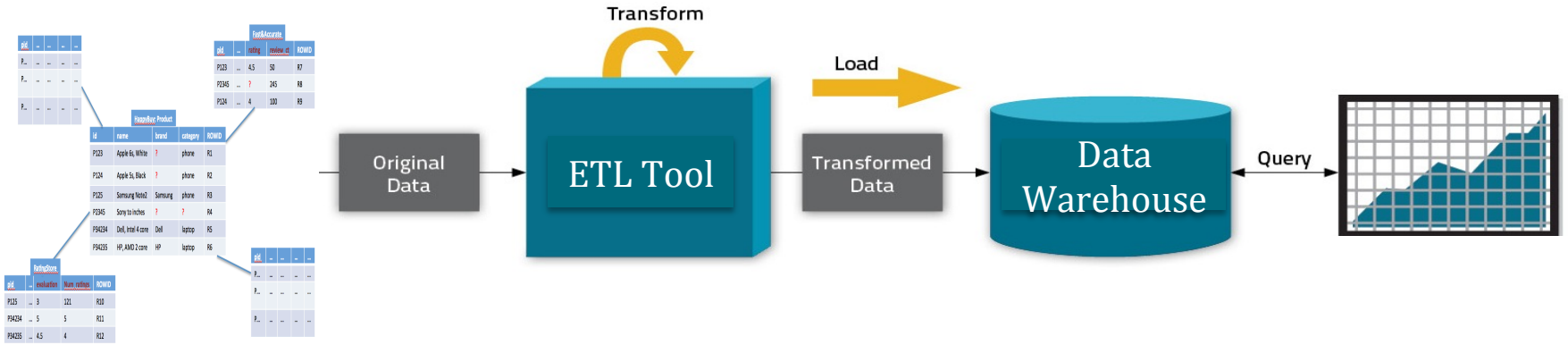
Data Cleaning Technician:
Cleaning all messy data before analysis



Unnecessary processing of unused data.

Inline cleaning

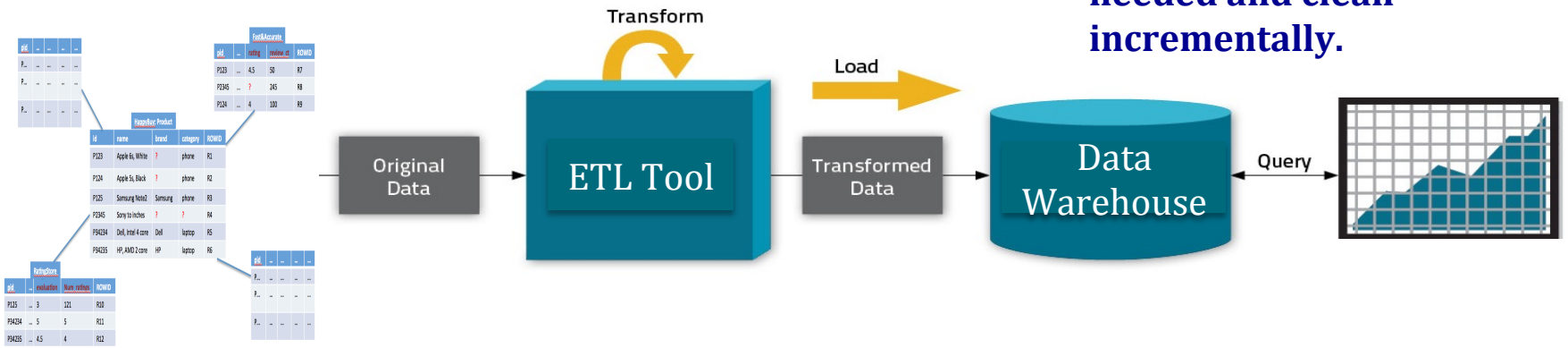
Data Analyst:
Cleaning all messy data when analysis



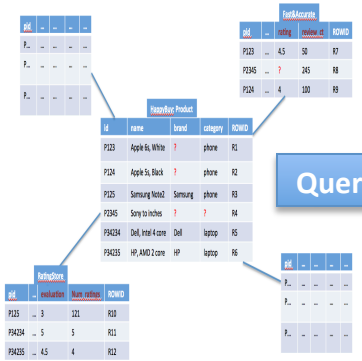
(1) Unnecessary processing of unused data.
(2) Duplication of work.

On-demand cleaning

Data Scientist/Crowdsourcing: Delay the cleaning process until needed and clean incrementally.



```
SELECT r.pid, r.rating,r.review_ct
FROM Rating r
WHERE r.rating >= 4 and r. review_ct >=100
```



Query

Analysis Result

...

Feedback:

Interacting with ***paygo***:

Alice*: I want to clean the data.*

Paygo: OK, does "rating" match to "evaluation"?

Alice: *Yes.*

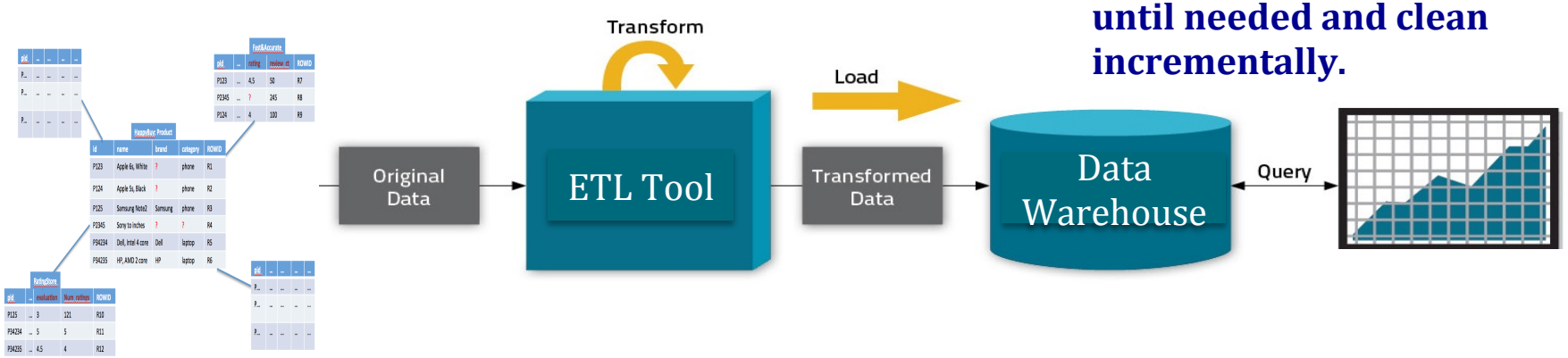
Paygo: Good, here is the result, do you want to clean further?
Alice: ...

More Accurate Analysis Result

...

On-demand cleaning

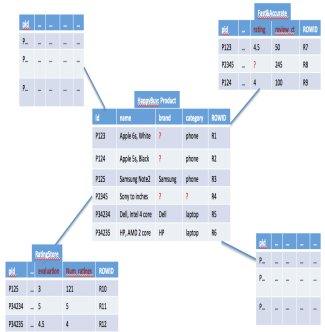
Data Scientist/Crowdsourcing: Delay the cleaning process until needed and clean incrementally.



Time and cost efficient comparatively

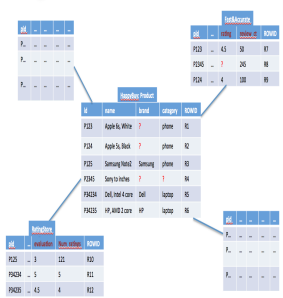
We need a general on-demand cleaning framework

Current systems focus on different data quality problems and have different representations and different forms of feedback.



Transform

Transform



<object,attribute,value>
t_0=<123,rating,4.5>
t_1=<124,rating,4>
t_2 =<125,evaluation,3>
t3=<2345,rating,?>
T4=<34234,evaluation,5>
T5=<34235,evaluation,4.5>
T6=<123,review_ct,50>
T7=<124,review_ct,100>
T8=<2345,review_ct,245>
T9=<125,num_ratings,121>
T10=<34234,num_ratings,5>
T11=<34235,num_ratings,4>

Paygo

Function DB

<object,attribute,value>
t_0=<123,rating,4.5>
t_1=<124,rating,4>
t4=<34234,rating,5>
t5=<34235,rating,4.5>
t7=<124,review_ct,100>
t8=<2345,review_ct,245>
t9=<125,review_ct,121>

pid	...	rating	review_ct	ROWID
P2345	...	4.5	245	R8
P124	...	4	100	R9

Too much extra work to unify the systems!

<object,attribute,value>
t_0=<123,rating,4.5>
t_1=<124,rating,4>
t4=<34234,rating,5>
t5=<34235,rating,4.5>
t7=<124,review_ct,100>
t8=<2345,review_ct,245>
t9=<125,review_ct,121>

pid	categoty	Rating	review_ct
P124	phone	4	100

pid	...	rating	review_ct	ROWID
P2345	...	4.5	245	R8
P124	...	4	100	R9

pid	...	rating	review_ct	ROWID
P2345	...	4.5	245	R8
P124	...	4	100	R9

We need a unified system for on-demand ETL!

Challenges:

1. How does Alice express her data cleaning needs?
2. How do we present (uncertain) query results to Alice?
3. What can Alice do with that information?
4. …

We need a unified system for on-demand ETL!

1. How does Alice express her data cleaning needs?
2. How do we present (uncertain) query results to Alice?
3. What can Alice do with that information?
4. ...

Mimir

A unified solution for managing uncertainty

http://www.taringa.net/posts/arte/15354098/El-tributo-de-Odin-a-la-sabiduria.html

We have:

HappyBuy:Product

id	name	brand	category	ROWID
P123	Apple 6s, White	NULL	phone	R1
P124	Apple 5s, Black	NULL	phone	R2
P125	Samsung Note2	Samsung	phone	R3
P2345	Sony to inches	NULL	NULL	R4
P34234	Dell, Intel 4 core	Dell	laptop	R5
P34235	HP, AMD 2 core	HP	laptop	R6

We want:

HappyBuy: SaneProduct

id	name	brand	category	ROWID
P123	Apple 6s, White	Apple	phone	R1
P124	Apple 5s, Black	Apple	phone	R2
P125	Samsung Note2	Samsung	phone	R3
P2345	Sony to inches	Sony	TV	R4
P34234	Dell, Intel 4 core	Dell	laptop	R5
P34235	HP, AMD 2 core	HP	laptop	R6

HappyBuy: Product

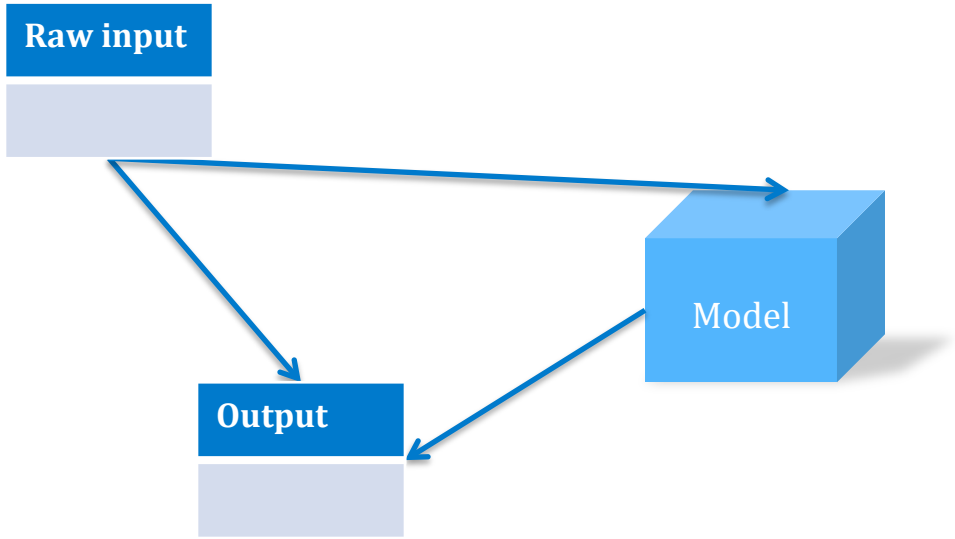
id	name	brand	category	ROWID
P123	Apple 6s, White	NULL	phone	R1
P124	Apple 5s, Black	NULL	phone	R2
P125	Samsung Note2	Samsung	phone	R3
P2345	Sony to inches	NULL	NULL	R4
P34234	Dell, Intel 4 core	Dell	laptop	R5
P34235	HP, AMD 2 core	HP	laptop	R6

HappyBuy: SaneProduct

id	name	brand	category	ROWID
P123	Apple 6s, White	Apple	phone	R1
P124	Apple 5s, Black	Apple	phone	R2
P125	Samsung Note2	Samsung	phone	R3
P2345	Sony to inches	Sony	TV	R4
P34234	Dell, Intel 4 core	Dell	laptop	R5
P34235	HP, AMD 2 core	HP	laptop	R6

Lens: *a data processing component that is evaluated as part of a normal ETL pipeline.*

```
CREATE LENS SaneProduct AS
SELECT * FROM Product
USING DOMAIN_REPAIR(
      category string NOT NULL,
      brand string NOT NULL );
```



Lenses make best use of source data and make a **best-effort guess** using the learnt model.

What the user sees through the lens **MAY BE WRONG!**

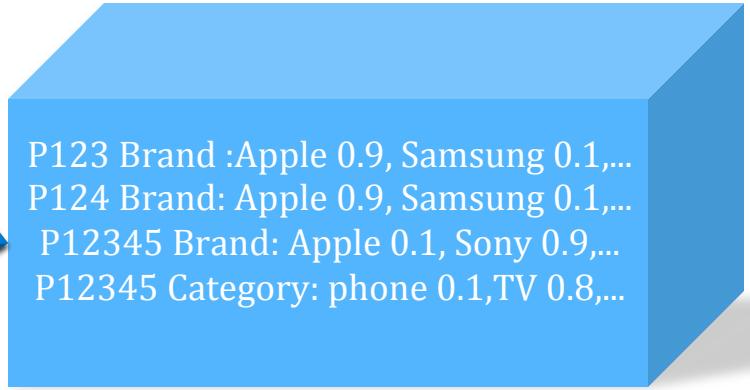
https://s-media-cache-ak0.pinimg.com/736x/06/e0/a1/06e0a140416c2d58eb2176aa501e44c7.jpg

Mimir provides:

- Nondeterministic result
- Analysis of result quality
- User the flexibility to improve the result

HappyBuy: SaneProduct

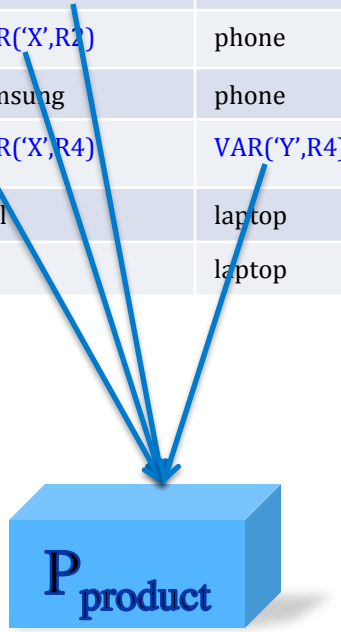
id	name	brand	category	ROWID
P123	Apple 6s, White	Apple*	phone	R1
P124	Apple 5s, Black	Apple*	phone	R2
P125	Samsung Note2	Samsung	phone	R3
P2345	Sony to inches	Sony*	TV*	R4
P34234	Dell, Intel 4 core	Dell	laptop	R5
P34235	HP, AMD 2 core	HP	laptop	R6



Lens produces a PC-Table, which defines the set of possible outputs, and a probability measure that approximates the likelihood that any given possible output accurately models the real world.

SaneProduct

id	name	brand	category	ROWID
P123	Apple 6s, White	VAR('X',R1)	phone	R1
P124	Apple 5s, Black	VAR('X',R2)	phone	R2
P125	Samsung Note2	Samsung	phone	R3
P2345	Sony to inches	VAR('X',R4)	VAR('Y',R4)	R4
P34234	Dell, Intel 4 core	Dell	laptop	R5
P34235	HP, AMD 2 core	HP	laptop	R6



We have:

Mobile Application: Rating2

pid	...	evaluation	Num_ratings	ROWID
P125	...	3	121	R10
P34234	...	5	5	R11
P34235	...	4.5	4	R12

Survey: Rating1

pid	...	rating	review_ct	ROWID
P123	...	4.5	50	R7
P2345	...	NULL	245	R8
P124	...	4	100	R9

We want:

AllRatings

pid	...	rating	review_ct	ROWID
P125	...	3	121	R10
P34234	...	5	5	R11
P34235	...	4.5	4	R12
P123	...	4.5	50	R7
P2345	...	5	245	R8
P124	...	4	100	R9

Mobile Application: Rating2

pid	...	evaluation	Num_ratings	ROWID
P125	...	3	121	R10
P34234	...	5	5	R11
P34235	...	4.5	4	R12

Survey: Rating1

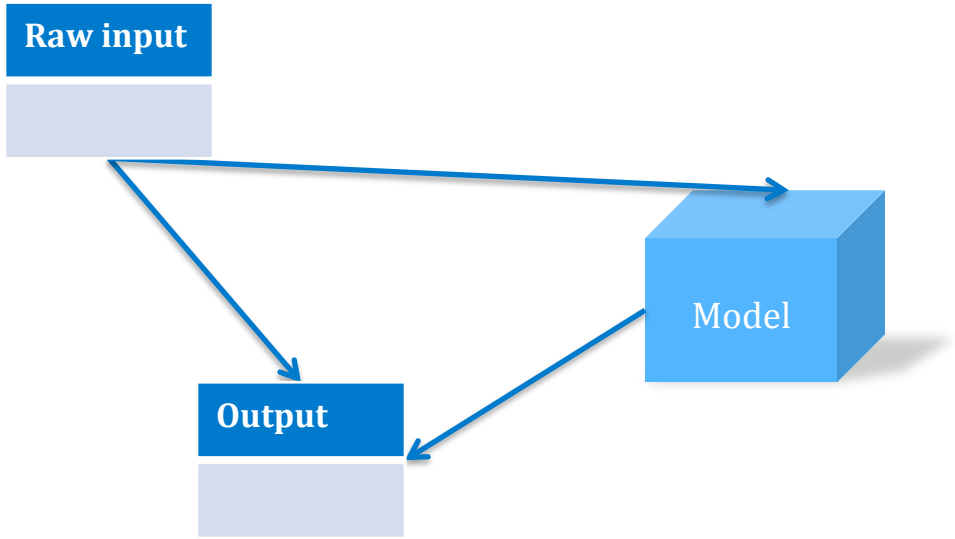
pid	...	rating	review_ct	ROWID
P123	...	4.5	50	R7
P2345	...	NULL	245	R8
P124	...	4	100	R9

AllRatings

pid	...	rating	review_ct	ROWID
P125	...	3	121	R10
P34234	...	5	5	R11
P34235	...	4.5	4	R12
P123	...	4.5	50	R7
P2345	...	NULL	245	R8
P124	...	4	100	R9

```
CREATE LENS MatchedRating2 AS SELECT * FROM Rating2
            USING SCHEMA_MATCHING( pid string, ..., rating
            float, review_ct float, NO LIMIT );

CREATE VIEW AllRatings AS SELECT * FROM MatchedRatings2
            UNION SELECT * FROM Ratings1;
```

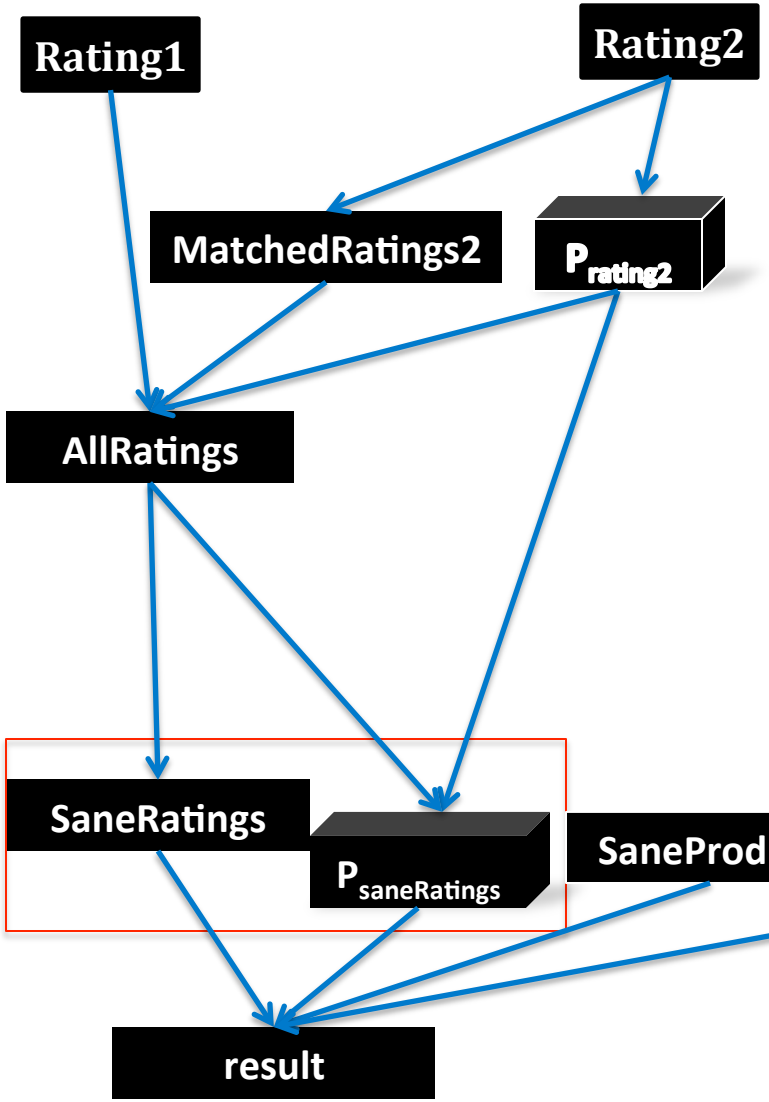


Lenses make best use of source data and make a **best-effort guess** using the learnt model.

MatchedRatings2

pid	...	rating	Review_ct
P125	...	If Var('rat=eval') then 3 else If Var('rat=num_rating') then 121 else NULL	If Var('review_ct=eval') then 3 else If Var('review_ct=num_rating') then 121 else NULL
P34234	...	If Var('rat=eval') then 5 else If Var('rat=num_rating') then 5 else NULL	If Var('review_ct=eval') then 5 else If Var('review_ct=num_rating') then 5 else NULL
P34235	...	If Var('rat=eval') then 4.5 else If Var('rat=num_rating') then 4 else NULL	If Var('review_ct=eval') then 4.5 else If Var('review_ct=num_rating') then 4 else NULL

Cells in a generalized C-Table can have arbitrary expressions.

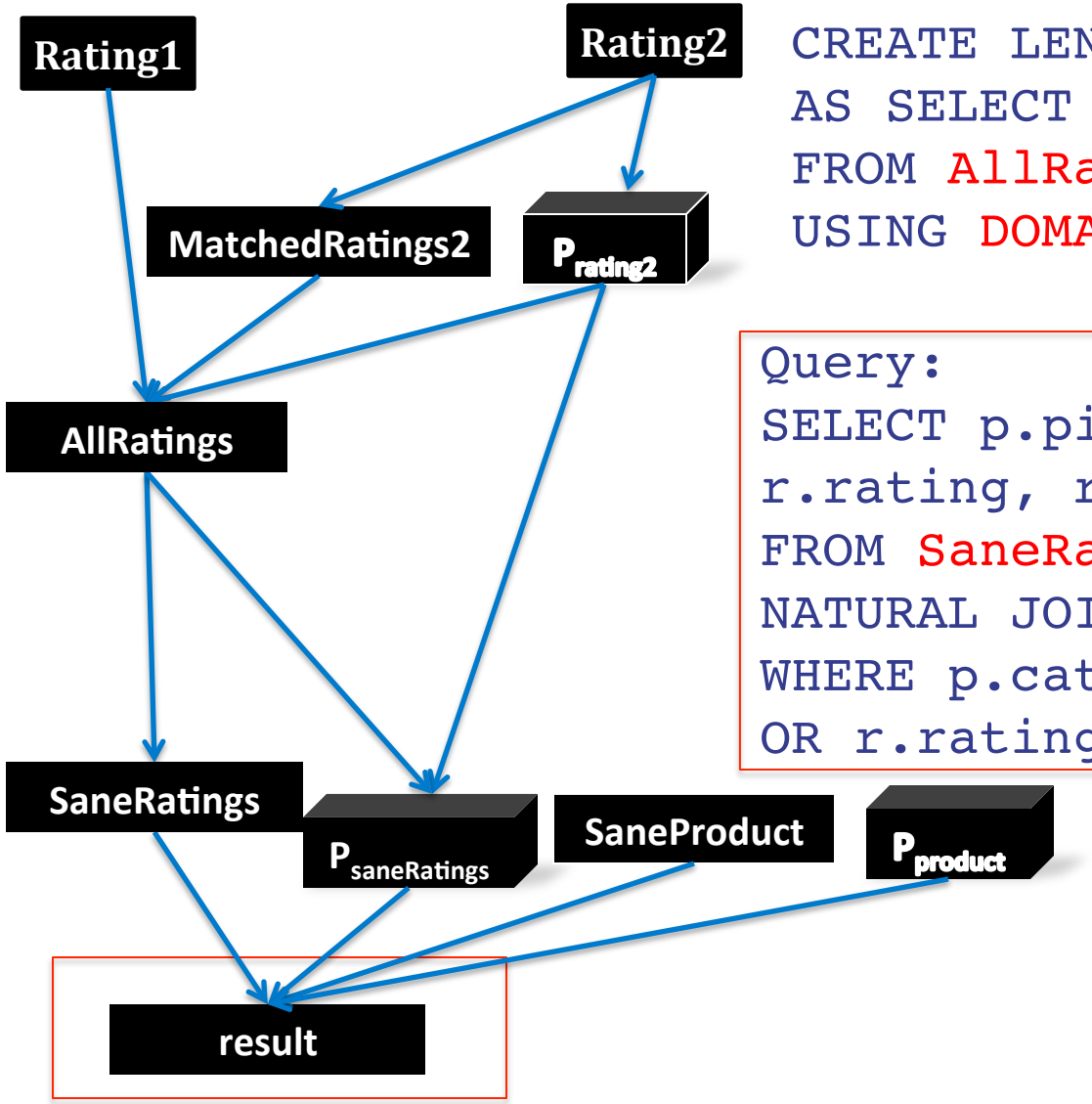



```
CREATE LENS SaneRatings
AS SELECT pid, rating, review_ct
FROM AllRatings
USING DOMAIN_REPAIR(
            rating DECIMAL NOT NULL);
```

Query:

```
SELECT p.pid, p.category,
r.rating, r.review_ct
FROM SaneRatings r
NATURAL JOIN SaneProduct p
WHERE p.category IN (`phone',`TV')
OR r.rating > 4
```

Lenses guarantee to *compose* and are *closed* on relational algebra.




```
CREATE LENS SaneRatings
AS SELECT pid, rating, review_ct
FROM AllRatings
USING DOMAIN_REPAIR(
                rating DECIMAL NOT NULL);
```

```
Query:
SELECT p.pid, p.category,
r.rating, r.review_ct
FROM SaneRatings r
NATURAL JOIN SaneProduct p
WHERE p.category IN (`phone',`TV')
OR r.rating > 4
```

Lenses guarantee to *compose* and are *closed* on relational algebra.

```
SELECT p.pid, p.category, r.rating, r.review_ct
FROM AllRatings r NATURAL JOIN SaneProduct p
WHERE p.category IN (`phone',`TV') OR r.rating > 4
```

id	category	rating	review_ct	
P123	phone	4.5	50	
P124	phone	4	100	
P125	phone	2 *	3 *	
P34235	laptop	5 *	4.5 *	*

(Up to 2 results may be missing. *)

The best effort result hides uncertainty from the user, leaving the user the summary of which results are uncertain.

id	category	rating	review_ct	
P123	phone	4.5	50	
P124	phone	4	100	
P125	phone	2 *	3 *	
P34235	laptop	5 *	4.5 *	*

(Up to 2 results may be missing. *)

id	Category	rating	Review_ct	Φ(condition)
P123	phone	4.5	50	T
P124	phone	4	100	T
P125	phone	If Var('rat=eval') then … …else Var('Z',R10)	If Var('rat=eval') then … …else NULL	T
P2345	Var('Y',R4)	Var('Z',R8)	245	(Var('Y',R4) = 'phone') (Var('Y',R4) = 'TV') Var('Z',R8) > 4
P34234	laptop	If Var('rat=eval') then … …else Var('Z',R11)	If Var('rat=eval') then … …else NULL	Var('rat=eval') Var('rat=num_rating'))= Var('Z',R11) > 4
P34235	laptop	If Var('rat=eval') then … …else Var('Z',R12)	If Var('rat=eval') then … …else NULL	Var('rat=eval') Or (not Var('rat=num_rating') (and Var('Z',R11) > 4))

Mimir: I guessed the value of Rating is 5, but I am not sure.

id	category	rating	review_ct	
P123	phone	4.5	50	
P124	phone	4	100	
P125	phone	2 *	3 *	
P34235	laptop	5 *	4.5 *	*

(Up to 2 results may be missing. *)

Var('rat=eval') ∨ (¬ Var('rat=num_rating') ∧ (Var('Z',R11) > 4))

entropy: $H(t) = -(p_t \cdot \log_2(p_t) + (1-p_t) \cdot \log_2(1-p_t))$

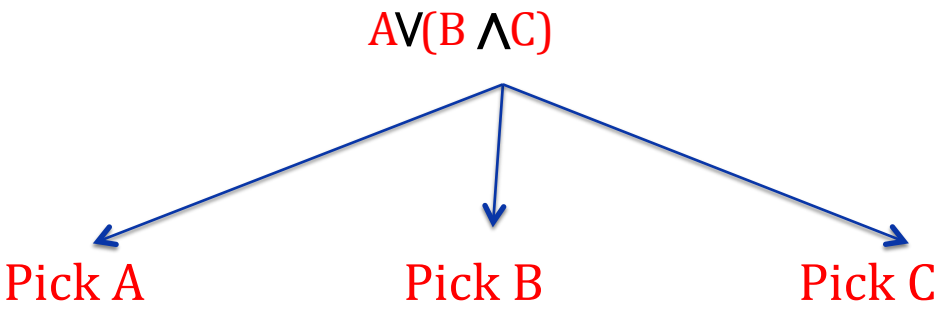
We use entropy to measure how uncertain the best effort result is.

I want to improve the result quality.

OK! (Generating feedback...)

var	Cost	P
A	10	0.5
B	5	0.8
C	3	0.5

Model



- Random
- NMETC: naïve minimum expected total cost
- ***CPI***: approximate minimum expected total cost (EG2, CS-ID3,IDX,C4.5)

I want to improve the result quality.

OK! (Generating feedback...)

var	Cost	P
A	10	0.5
B	5	0.8
C	3	0.5

H=0.88

A∨(B∧C)

Pick A

p(T)=0.5 p(F)=0.5

true [1+,0-];H=0

(B∧C) [0.4+,0.6-];H=0.97

$E[H(A)]=p(T)*H(T)+p(F)*H(F)=0.5*0+0.97*0.5=0.485$

I want to improve the result quality.

OK! (Generating feedback...)

A∨(B ∧C)

↓

Pick A

H=0.88

E[H(A)]=0.485

var	Cost	P
A	10	0.5
B	5	0.8
C	3	0.5

EG2-based *CPI(A)*=(H-E[H(A)])/cost=0.0395

I want to improve the result quality.

OK! (Generating feedback...)

H=0.88

AV(B ∧C)

Pick A

E(H)=0.485
CPI=0.0395

Pick B

E(H)=0.8488
CPI=0.00624

Pick C

E(H)=0.7345
CPI=0.0485

Mimir: Is the value of Rating for P34235 larger than 4?

Yes.

H=0.88

A∨(B∧C)

Pick A	Pick B	Pick C
E(H)=0.485 CPI=0.0395	E(H)=0.8488 CPI=0.00624	E(H)=0.7345 CPI=0.0485

OK! Here is the result. Do you want to continue?

Yes.

$A \vee (B \wedge C)$

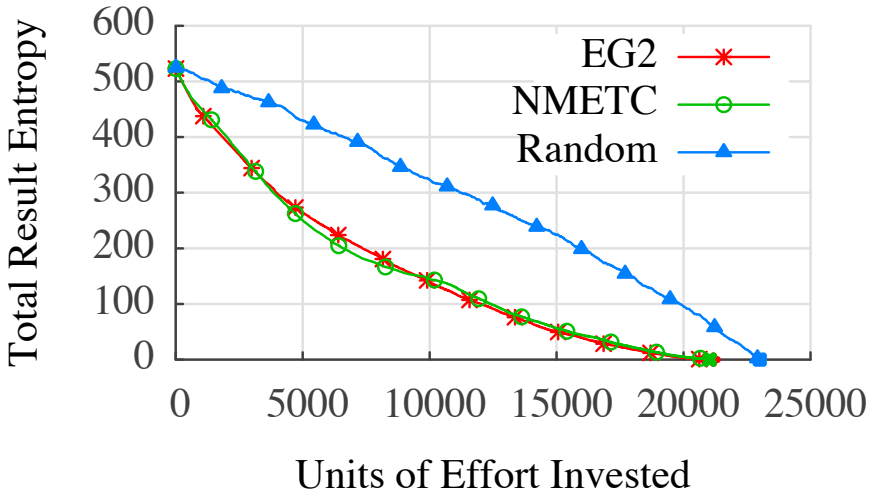
Picked C

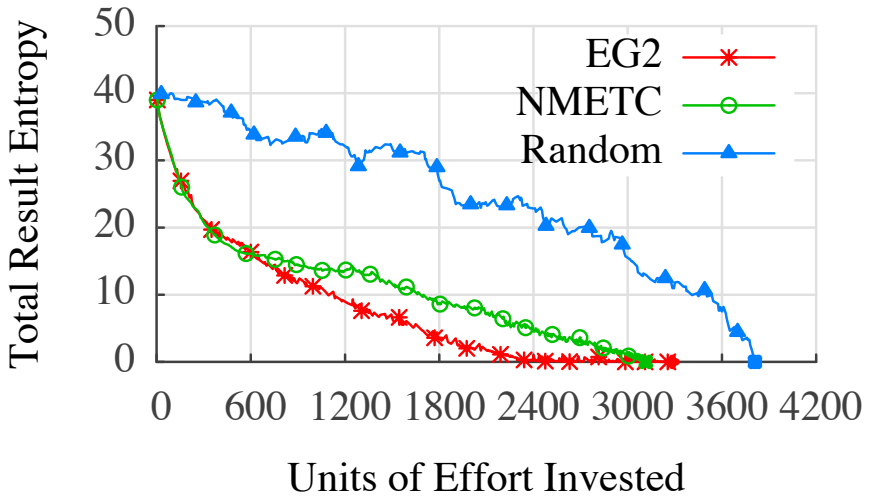
$(A \vee B)$ H=0.47

Pick A Pick B

...

Effectiveness of cleaning





- Random
- NMETC: naïve minimum expected total cost
- ***EG2***: approximate minimum expected total cost

EG2-based CPI method is sufficiently close to NMETC in units of effort invested and has steep curve to produce high-quality results with minimal investment.

In the paper

- Composition
  - Composition description, commands, examples.
  - Experiment analysis.
- Variable generation relational algebra (VG-RA)
- Virtual C-Tables
- Other lens examples e.g. Archival lens

Conclusions

Mimir is a system that:

1. Provides Lenses to automatically curate different kinds of messy data in a uniform way.
2. Presents non-deterministic result and indicates uncertainty to user.
3. Provides heuristic uncertainty ranking algorithms to reduce uncertainty and minimize the total cost. (In the paper)

Questions?