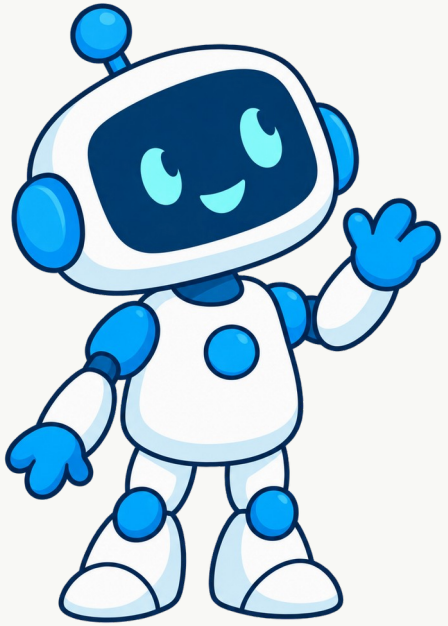
Knowledge-Guided Data Cleaning

From Knowledge Acquisition to Reliable Decisions



Yixi Zhou

Hong Kong Baptist University

COMP7160 Survey Presentation

The path through the survey



1 **Where does knowledge come from?**



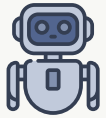
2 How does knowledge guide repairs?



3 What makes a repair trustworthy?



4 What changes in data preparation?



5 What should we take home?

Paper & slides

A simple request



What were our actual sales
by month and city last year?



Orders



Month



City

The data behind the request

Orders

Order	City	Amount
O101	Nwe York	120
O101	Nwe York	120
O102	Boston	80

Customer addresses

Customer	ZIP	City
C01	10001	Boston
C02	02108	Boston

Products

Product	Brand	Price
iPhone 15	Apple	19
Galaxy S24	?	799

Orders from another system

Order	Customer	Status
O103	Alice / Bob	Paid
O104	C04	Cancelled

Order	ZIP	City
O101	10001	Boston
O101	10001	Nwe York



How do we know what is wrong?

What evidence supports a change?

How would a computer know?

Order	ZIP	City
O101	10001	Boston

What tells me this is an error?



The simplest approach: provide the rules in advance

A conditional rule applies only to matching records

Conditional Functional Dependencies (ICDE, 2007) [1]

UK postcode determines street

Country code	Postcode	Street
44	AB1	High St
44	AB1	King St
1	10001	Broadway

Match the UK condition

Same postcode, different streets

US row: outside this rule

Dependency + value pattern = conditional rule

One rule interface, a shared cleaning core

NADEEF (SIGMOD, 2013) [7]

Rules

ZIP 10001 implies New York

Rule compiler

Supported repair



Raw records

ZIP	City
10001	Boston
10001	New York

Detect and repair

ZIP	City
10001	New York

Different rule types share one cleaning core

A conflict leaves several possible repairs

ZIP	City
10001	Boston

Change the city?

ZIP	City
10001	New York

?

Change the ZIP?

ZIP	City
02108	Boston

A confirmed key makes a repair certain

Certain Fixes (VLDB J., 2012) [2]

1 Confirm the key

ID	City
C01	Boston

C01 confirmed

2 Match master data

ID	City
C01	New York
C02	Boston

3 Extend trusted values

ID	City
C01	New York

Confirmed ID identifies the master record

Confirmed repairs become evidence for the next fix

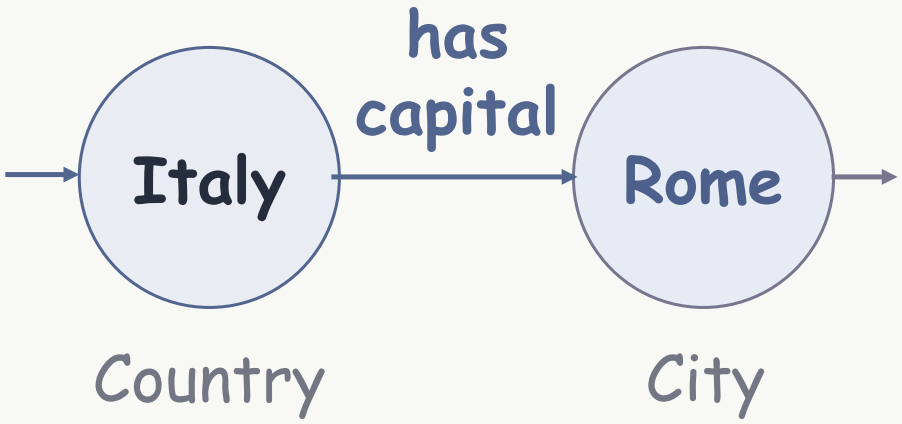
Knowledge-base facts explain a table error

KATARA (SIGMOD, 2015) [8]

Inspect the table

Country	Capital
Italy	Rome
Italy	Madrid

Typed relation



Check the facts

Claim	Check
Italy / Rome	Valid
Italy / Madrid	Conflict



Uncertain KB? Ask people

Types and relations connect tables to external facts





Expert effort



Cost



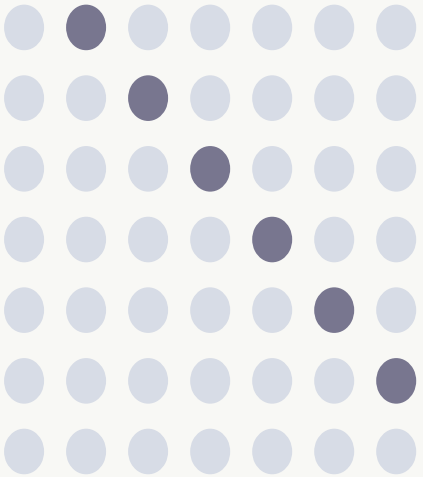
Time

Can the system acquire reliable knowledge itself?

Small samples reveal rules from a large table

Parallel Rule Discovery (SIGMOD, 2022) [9]

Large dataset



Mine samples in parallel

ZIP	City
10001	New York

ZIP	City
02108	Boston

ZIP	City
60601	Chicago

Candidate rules

ZIP determines City

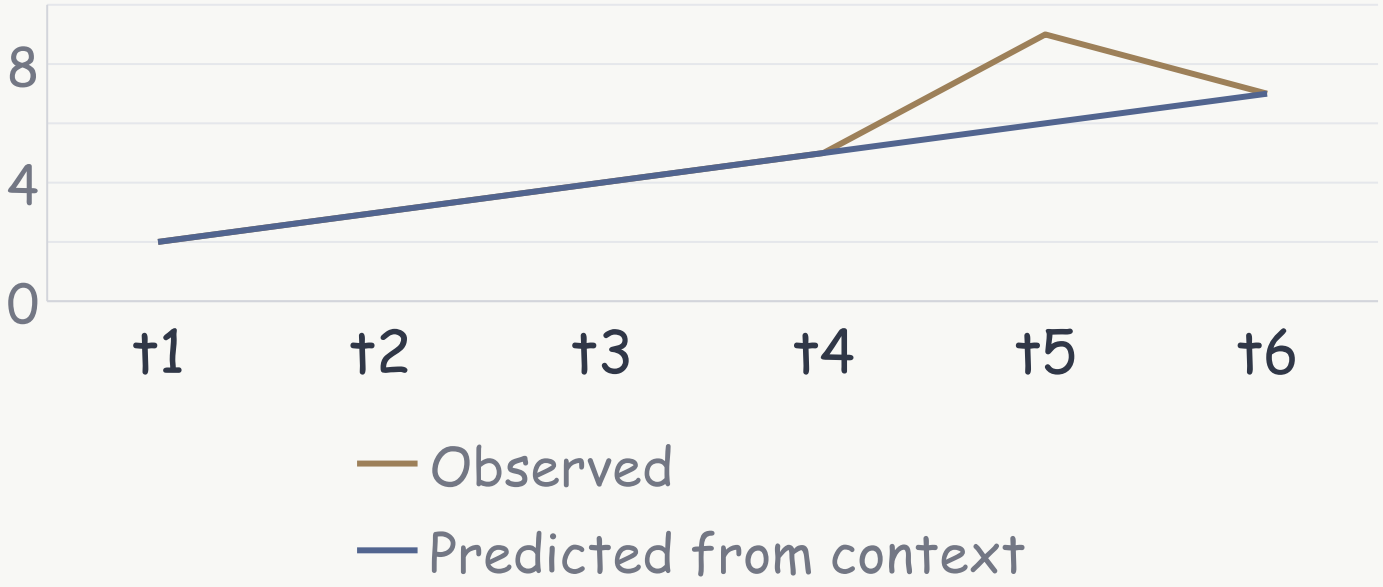
Combine, validate and prune

Multi-round sampling + semantic pruning

A temporal rule includes lag and normal variation

TSDDiscover (ICDE, 2024) [10]

Learn from related signals



Estimate the error bound

$$y(t) \approx f(x(t - lag))$$

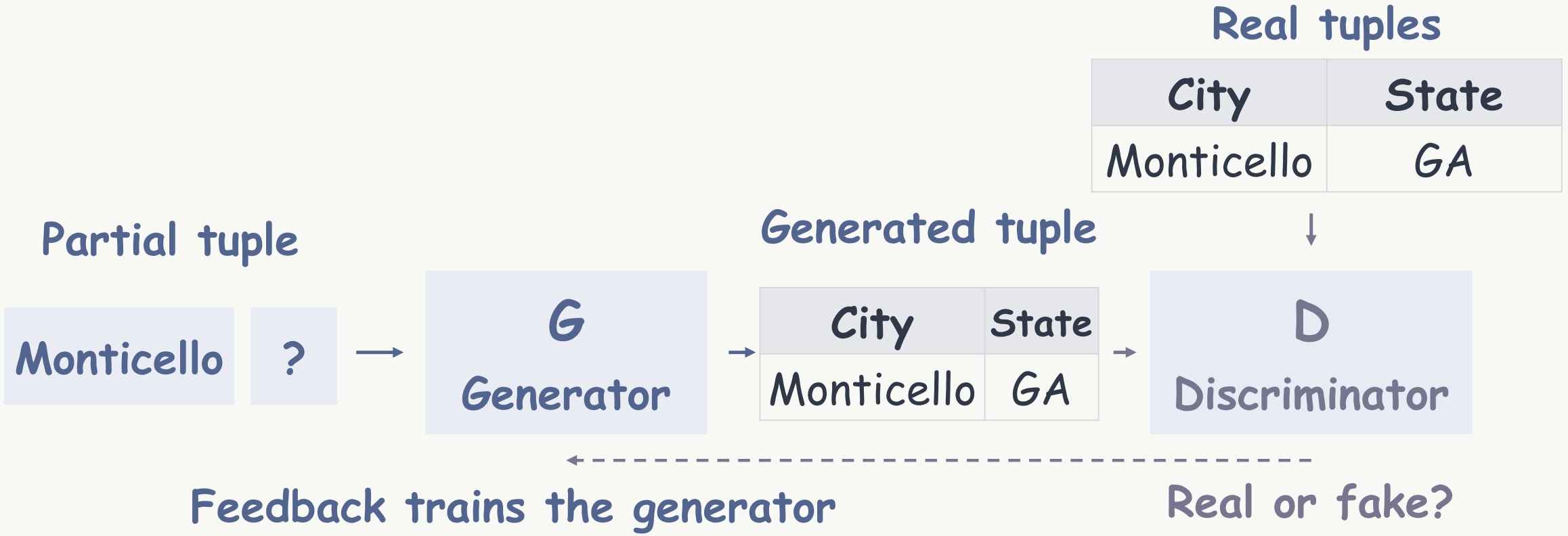
Normal tolerance

The spike lies outside the learned bound

Learn the function, then estimate its error bound

Adversarial training learns a generator of data tuples

Garf (PVLDB, 2022) [4]



Extract rules, then co-clean rules and data

A semantic-domain test travels to a new column

Auto-Test (PACMMOD, 2025) [11]

Learn from table corpora

City
Rome
Paris
Boston

→

Semantic-domain test

CITY DOMAIN

Column applicability

City column?

Value postcondition

Value in domain?

→

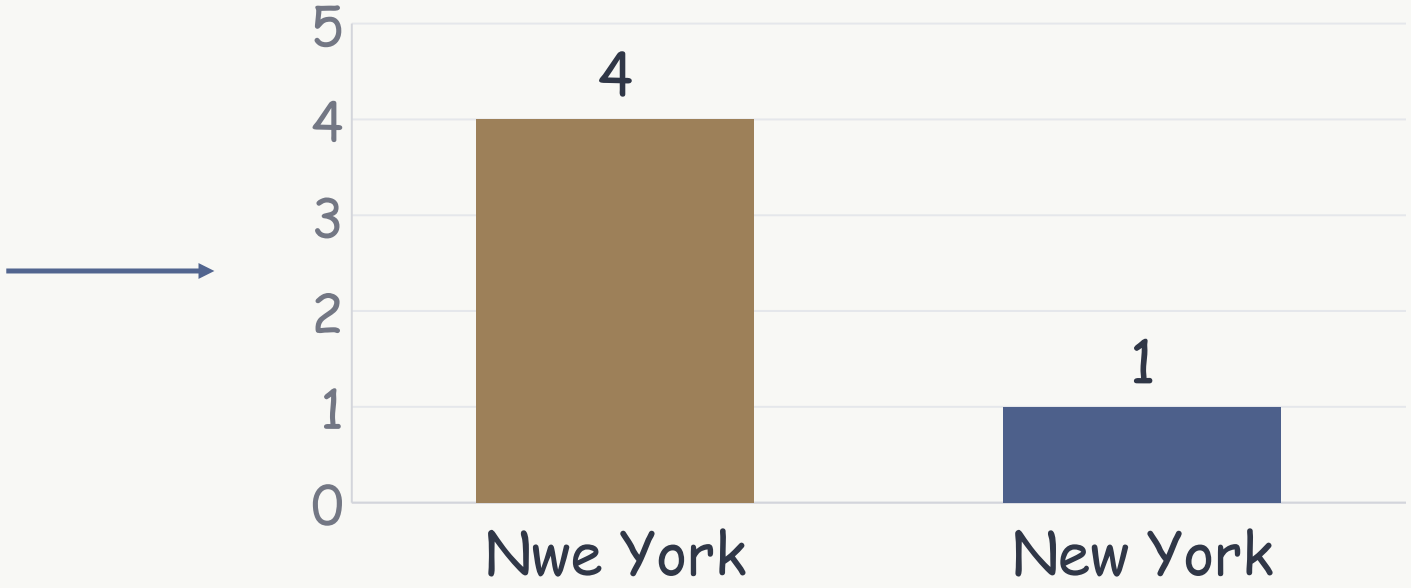
Test a new column

City	Result
Paris	Pass
12345	Flag

Check the column domain, then test each value

What if the learning data are wrong?

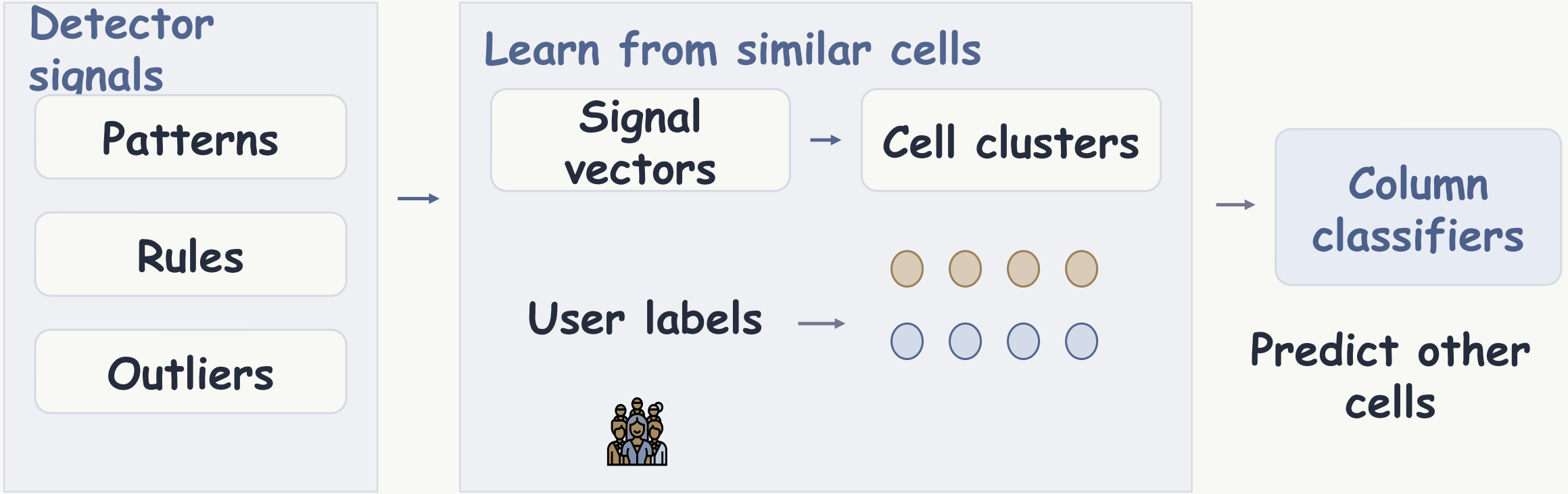
Observed city
Nwe York
Nwe York
Nwe York
Nwe York
New York



Frequent values can still be wrong

Feedback turns detector signals into decisions

Raha (SIGMOD, 2019) [5]



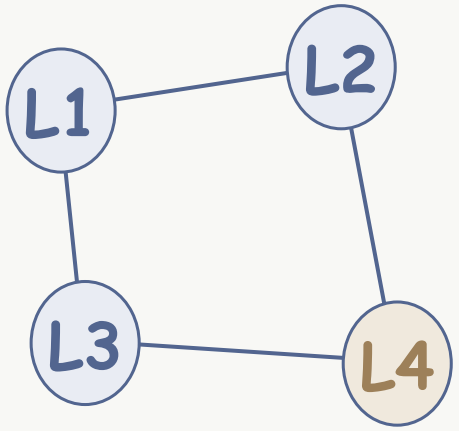


How does knowledge become a specific repair decision?

Use a few labels to explain and correct more cells

GIDCL (PACMMOD, 2024) [15]

Graph samples



Label a small sample

→

Creator–critic learning

Propose error patterns

Clean	Synthetic
Dothan	Dothn
Dothan	Dotahn

Critic tests the patterns

→

Graph-based correction

City	State
Dothn	AL
Dothan	AL

Clean neighbors guide the local LLM

One graph supports sampling, detection and repair

Context can suggest a repair absent from the table

Baran (PVLDB, 2020) [6]

Observed address

Name	Address
M. Lee	9th Str

→

Candidate sources

Value edits	Str → Street
Row context	Name + address
Column domain	Other addresses

→

Learn which repair to use

9th Street

Past edits +
user repair examples

Learn repair choices from context and past edits

Semantics reveal what a string pattern misses

DataVinci (SIGMOD, 2025) [12]

Plausible string formats
String
red 1
dark green 2
blue phone 3

→ Semantic pattern

COLOR NUMBER

COLOR phone 3

"phone" breaks the pattern

→ Candidate repair

blue 3

Remove "phone"
Rank the edit program

Semantic patterns guide executable string edits

What if the evidence disagrees?

ONE RECORD, COMPETING EVIDENCE

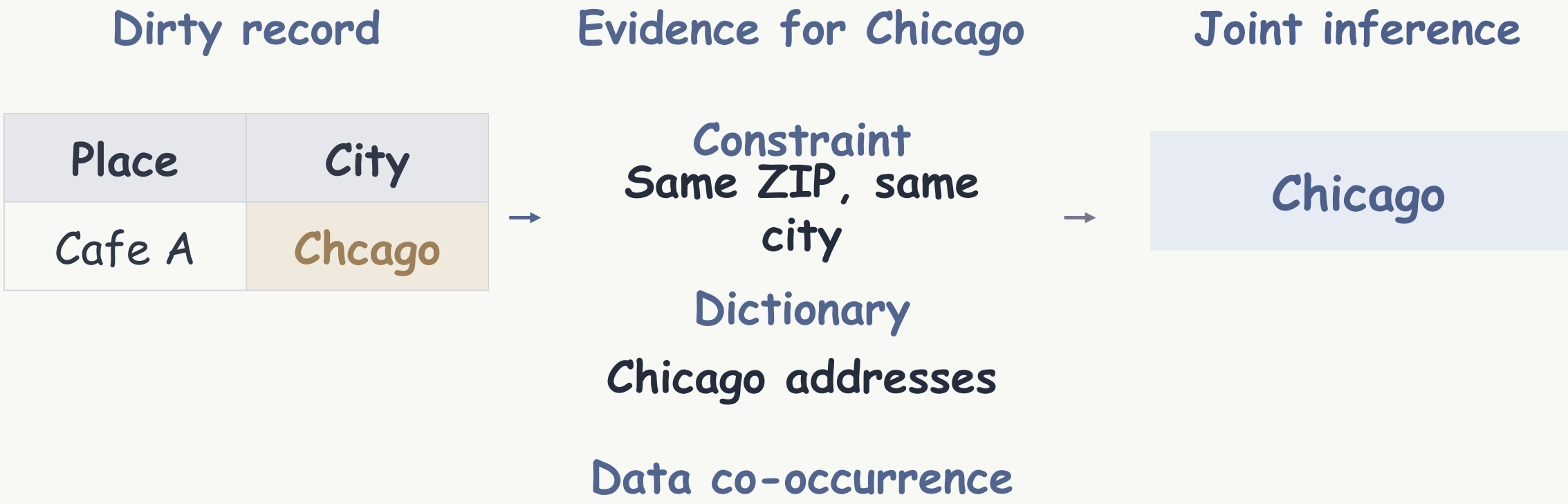
ZIP rule	New York
Past orders	Boston
Reference fact	New York



Which answer is supported?

Several weak signals support one cell assignment

HoloClean (PVLDB, 2017) [3]



Infer repairs jointly from complementary signals

An ML match becomes a premise of a logical rule

Rock (SIGMOD Companion, 2024) [14]

Two transaction records

Commodity	Maker
iPhone 15	Apple
Apple iPhone15	Appl

ML matches
the same commodity

Propagate a validated fact

Same commodity, same maker

Commodity	Maker
Apple iPhone15	Apple

Validate premises, deduce repairs, retain facts

One director row actually describes two people

Splitting Tuples (PACMMOD, 2023) [16]

Merged IMDb identity

Name	Nation	Film
N. Schneider	German	Sturm

Separate using supporting facts

Person	Nation	Born
N. Schneider A	Swiss	1986
N. Schneider B	German	1982

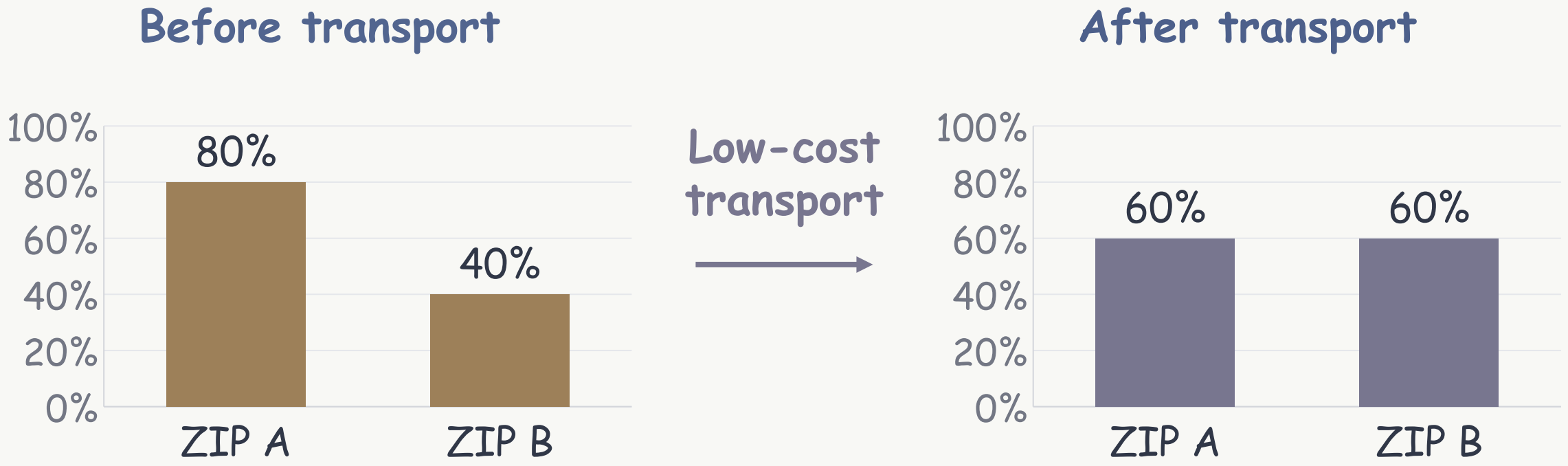
Two namesakes share one record

Sturm belongs to A

Identify entities, split attributes, complete tuples

Repair a spurious association in the distribution

OTClean (SIGMOD, 2024) [17]



Within the same causal-factor group

Low-cost transport enforces conditional independence

Can a repair harm correct data?



REPAIR DAMAGE

Order	Before	After
O101	120	720
O102	80	80

Actual sales: 200 Reported sales: 800

Repair damage changes the final report

A conflict is a starting point for error diagnosis

Minimum Change ≠ Best Cleaning (PACMMOD, 2025) [19]

Same flight and departure

Flight	Departure	Arrival
F17	08:00	10:30
F17	08:00	1030

Diagnose the conflict

Evidence	Arrival
Format + context	10:30
Malformed time	1030

Conflicting arrival values

Parallel conflict search + probabilistic diagnosis

A violation may come from the value or the bound

bNDCRepair (PVLDB, 2025) [18]

Physical expectation

Power ≈ Voltage × Current

Expected	Measured
100	102

→ **Original constraint**

99 to 101

102 falls outside the bound

→ **Repair candidates**

Adjust the value

102 → 100

Or revise the bound

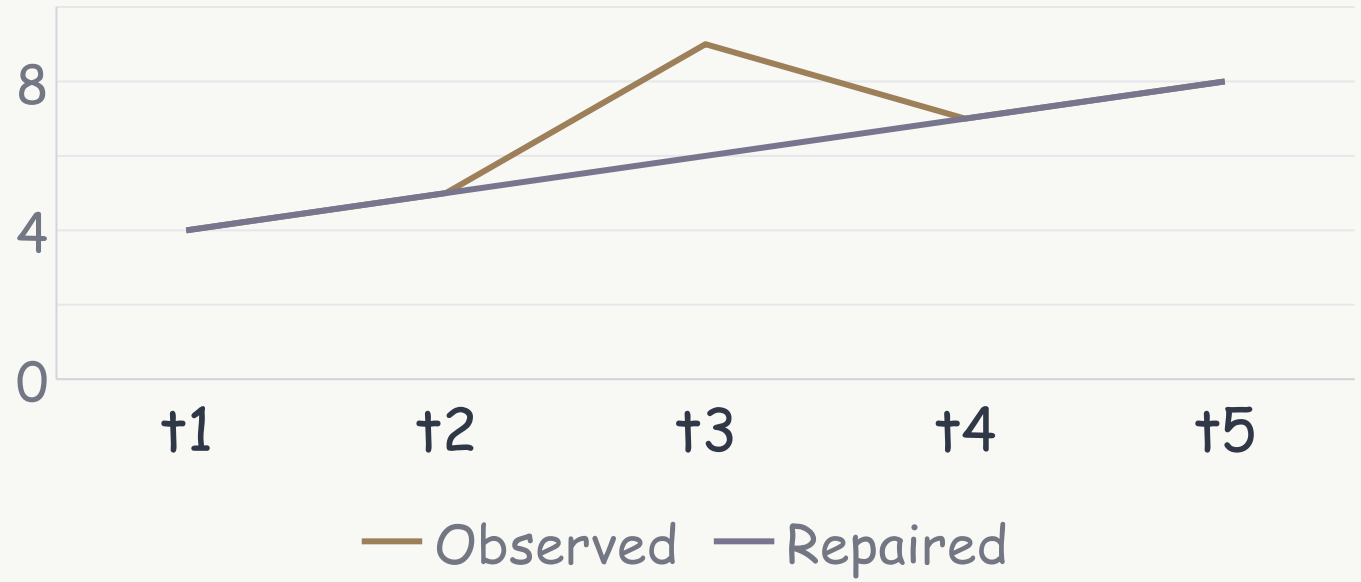
99 to 103

Choose using joint cost and clean validation data

Speed limits alone can miss an implausible point

SHoTClean (PACMMOD, 2026) [20]

A spike breaks the trend



Hard and soft constraints

Hard boundary

Feasible transitions?

Soft trend

Preferred feasible path?

Choose a maximum-weight path in a temporal graph

Specialists fill only when confidence is sufficient

Auto-Fill (PVLDB, 2026) [13]

Missing value

Product	Brand
iPhone 15	?

→

Specialist candidates

Specialist	Candidate
Knowledge	Apple
Text	Apple
Code	—

→

Calibrated choice

Apple

Confident: accept

?

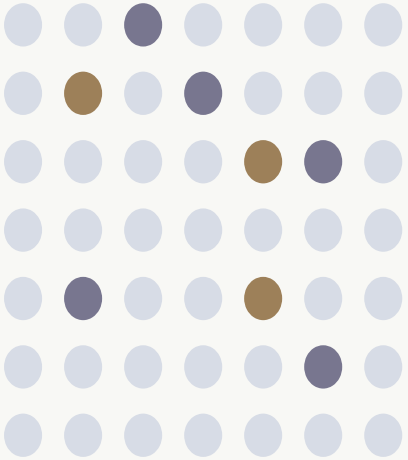
Uncertain: abstain

Calibrate knowledge, text and code specialists

Clean a useful batch, then update the model

ActiveClean (PVLDB, 2016) [21]

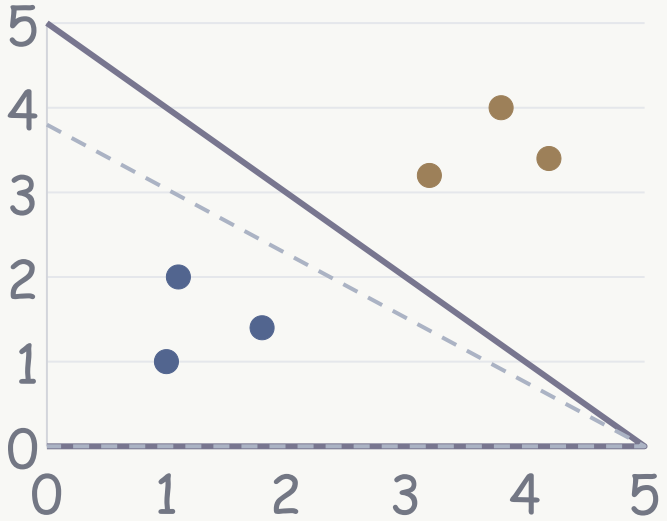
Dirty record pool



A sampled batch

Drug	Clean value
amoxcillin	amoxicillin
paracetmol	paracetamol

Update the classifier

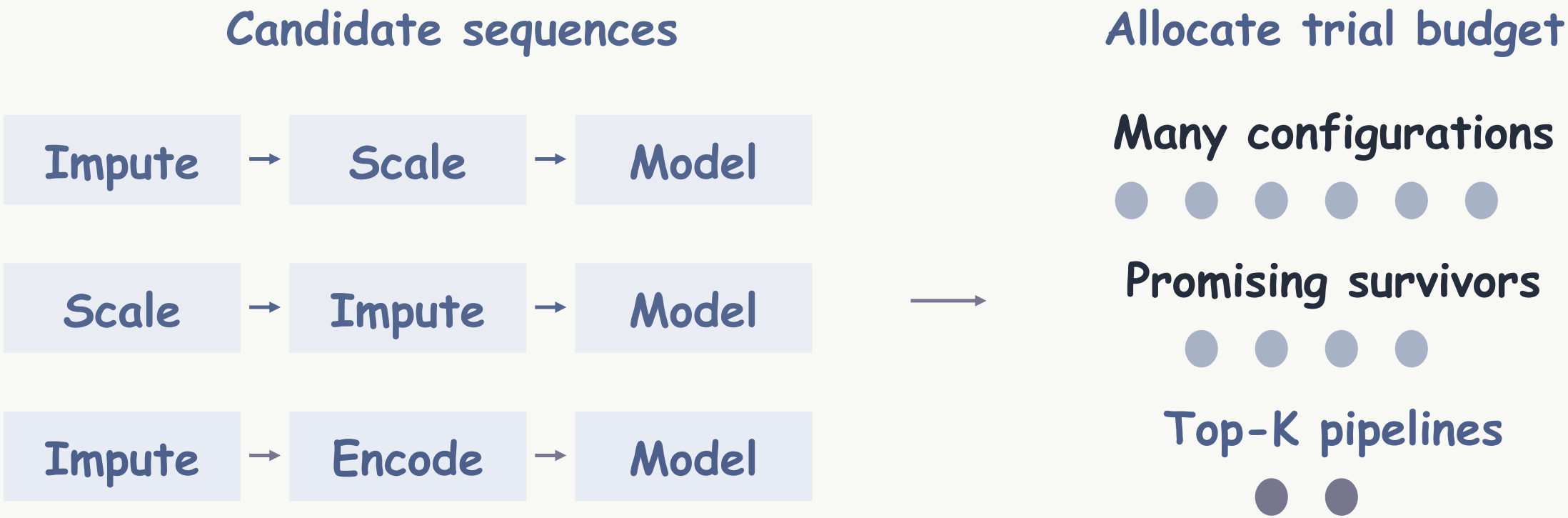


Updated boundary

Repeat sampling, user cleaning and model updates

Search both operation order and parameter choices

SAGA (PACMMOD, 2023) [22]



Evolutionary sequence search + successive halving

Data Preparation

The task extends beyond repairing values

- Valid values
- User intent
- A usable result



What does “actual sales” mean?



“Actual sales”

- Exclude cancelled orders
- Subtract refunds
- Group by month and city

Business context defines the result

Suggestions become a preview the user can inspect

CoWrangler (SIGMOD Companion, 2023) [27]

Emergency-call text

desc
Oak St; Station 7; 2024-01-01

→ Split → Separator

Preview the new columns

Address	Station	Time
Oak St	7	2024-01-01

One column, three kinds of information

Accept, modify or ask again

User actions and the script guide the next suggestion

A stronger prompt makes code reusable

MegaTran (PVLDB, 2025) [28]

Ambiguous examples

Serial	Expected
44927	2023-01-01
45292	2024-01-01

"Convert these dates"

→

A stronger prompt

Type

Excel serial date

Date system

Workbook convention

Output: YYYY-MM-DD

→

Reusable code

Handle new dates
Execute and check

Serial	Result
45658	2025-01-01

Refine prompts before generating transformation code

Evaluate what the system can actually do

PrepBench (PVLDB, 2026) [30]

Clarify the request

"Show actual sales"

Clarification

"Subtract refunds?"

"Yes"

Generate a transform

City	Amount
NY	120
Boston	80

NY becomes New York

Translate the workflow

Filter | Aggregate

Equivalent steps in another tool

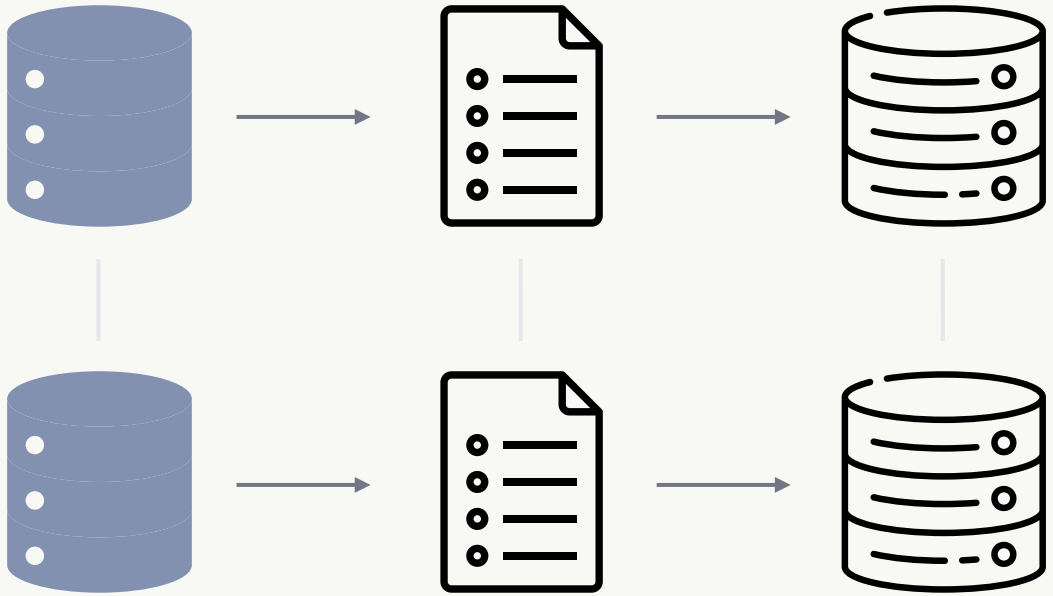
City	Sales
New York	120

Execute outputs and compare the resulting tables

The data environment keeps growing

Order	City
O101	NY
O102	Boston

One table



Many sources and relationships

Enterprise data needs private meanings and context

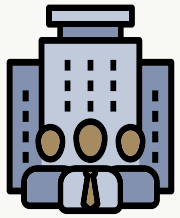
Enterprise Data Engineering Challenges (PVLDB, 2025) [23]

Opaque fields

Field	Value
amt_n	200
st_cd	C
dt_v2	2024-01-01

→

Internal knowledge



Meaning of C?
Which amount?
Where are refunds?

→

Compound task

Order	Status	Amt
O101	Complete	200

Interpret fields
Integrate records
Apply business rules

Enterprise gaps span data, tasks and private knowledge

A table's context can explain why it is relevant

Pneuma (PACMMOD, 2025) [24]

"Which table explains revenue after the phone launch?"

Table content

Year	Phones	Laptops
2020	100	80
2021	150	85

Catalog context

Product revenue
Foldable launch: 2021

Result
Product revenue

Search content and context with two retrieval indexes

Operation order changes the result

Filter → Restructure → Join → Aggregate

Order	Refund
O101	10
O101	15

Joining too early can repeat the order amount

Aggregate refunds per order before joining

Relationalize an irregular table without an output example

Auto-Tables (PVLDB, 2023) [25]

Input: dates as headers

Store	19-Oct	20-Oct
North	12	15
South	9	11

Stack / unpivot →

Output: dates as values

Store	Date	Sales
North	19-Oct	12
North	20-Oct	15
South	19-Oct	9
South	20-Oct	11

Rank candidates by input–output compatibility

Synthesize and rank multi-step transformation programs

The useful transformation is the one that enables joins

Auto-Prep (PVLDB, 2025) [26]

Fertility: wide table

Country	2010	2011
Chile	1.86	1.84

→

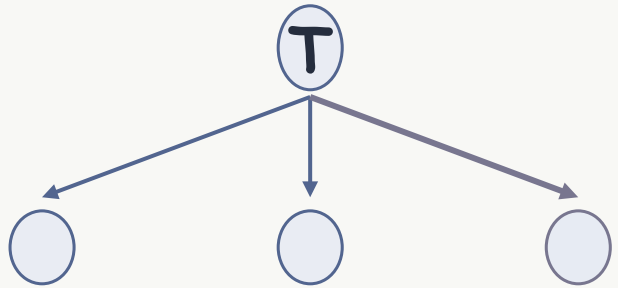
Unpivot exposes Year

Country	Year	Rate
Chile	2010	1.86
Chile	2011	1.84

→

Date table

Year	Days
2010	365
2011	365



No-op Transpose Unpivot

Join on Year

Score transformations together with joinability

An execution error triggers a different earlier choice

DeepPrep (PVLDB, 2026) [29]

Ratings input

movie	values
Jurassic	8.2 (2020)

Attempt 1: drop values

movie
Jurassic

→ Missing column: values

Earlier state stays available

Backtrack: keep values, then split

movie	score	year
Jurassic	8.2	2020

Pivot →

movie	2020
Jurassic	8.2

Backtrack to an earlier state and explore another path

Every task should make the next one smarter

Act → Validate → Remember



Retain verified knowledge for the next decision

Keep what works. Recheck what changes



01 Remember what you verify

Retain rules, meanings and experience

02 Update with evidence

Revise or retire outdated knowledge

03 Ask when evidence is missing

Inspect data, retrieve context or ask the user

Reuse knowledge. Verify every step

04 Carry knowledge across tasks

Cleaning → Transformation → Analysis

05 Check every operation with knowledge

Filter → Transform → Join → Aggregate



References

[1] **CFDs**
Bohannon et al. (ICDE, 2007)

[2] **Certain Fixes**
Fan et al. (VLDB J., 2012)

[3] **HoloClean**
Rekatsinas et al. (PVLDB, 2017)

[4] **Garf**
Peng et al. (PVLDB, 2022)

[5] **Raha**
Mahdavi et al. (SIGMOD, 2019)

[6] **Baran**
Mahdavi & Abedjan (PVLDB, 2020)

[7] **NADEEF**
Dallachiesa et al. (SIGMOD, 2013)

[8] **KATARA**
Chu et al. (SIGMOD, 2015)

[9] **Parallel Rule Discovery**
Fan et al. (SIGMOD, 2022)

[10] **TSDDiscover**
Ding et al. (ICDE, 2024)

[11] **Auto-Test**
Chen et al. (PACMMOD, 2025)

[12] **DataVinci**
Singh et al. (SIGMOD, 2025)

[13] **Auto-Fill**
Liu et al. (PVLDB, 2026)

[14] **Rock**
Bao et al. (SIGMOD Companion, 2024)

[15] **GIDCL**
Yan et al. (PACMMOD, 2024)

[16] **Splitting Tuples**
Fan et al. (PACMMOD, 2023)

[17] **OTClean**
Pirhadi et al. (SIGMOD, 2024)

[18] **bNDCRepair**
Ding et al. (PVLDB, 2025)

[19] **Minimum Change ≠ Best Cleaning**
Chen et al. (PACMMOD, 2025)

[20] **SHoTClean**
Fang et al. (PACMMOD, 2026)

[21] **ActiveClean**
Krishnan et al. (PVLDB, 2016)

[22] **SAGA**
Siddiqi et al. (PACMMOD, 2023)

[23] **Enterprise Data Engineering**
Bodensohn et al. (PVLDB, 2025)

[24] **Pneuma**
Balaka et al. (PACMMOD, 2025)

[25] **Auto-Tables**
Li et al. (PVLDB, 2023)

[26] **Auto-Prep**
Lai et al. (PVLDB, 2025)

[27] **CoWrangler**
Chopra et al. (SIGMOD Companion, 2023)

[28] **MegaTran**
Li et al. (PVLDB, 2025)

[29] **DeepPrep**
Fan et al. (PVLDB, 2026)

[30] **PrepBench**
Xu et al. (PVLDB, 2026)

Thank you

Questions & discussion



Paper & slides

Yixi Zhou

yxzhou@comp.hkbu.edu.hk

https://xanderzhou2022.github.io/Knowledge-Guided-Data-Cleaning/