CUSTOMER DATA UNIFICATION AT TME

PLUMB DIFFERENTLY

Created by Wouter Dullaert / @wouterdullaert

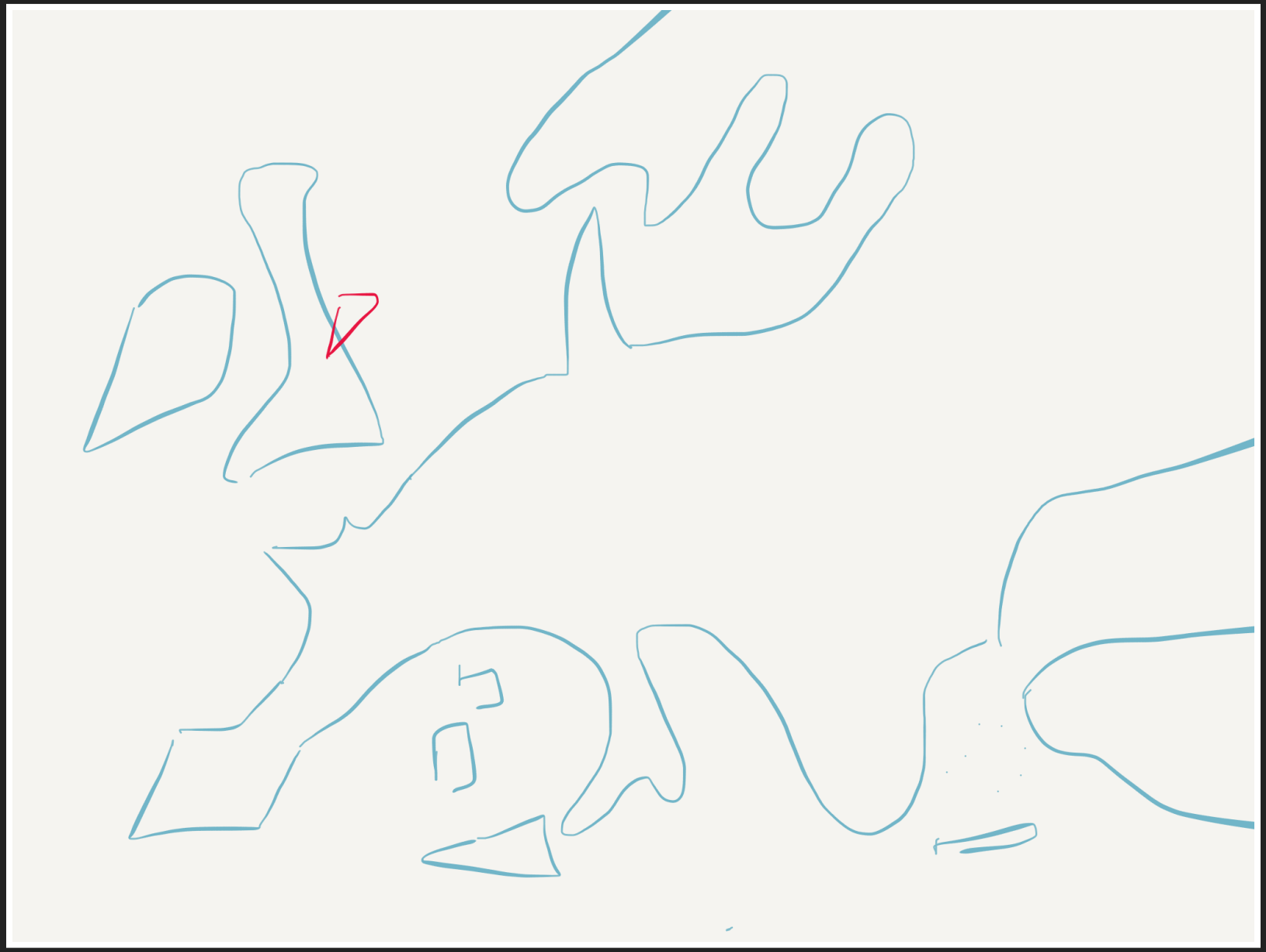
OVERVIEW

1. How did we get here?
2. Data Integration
3. Data Unification

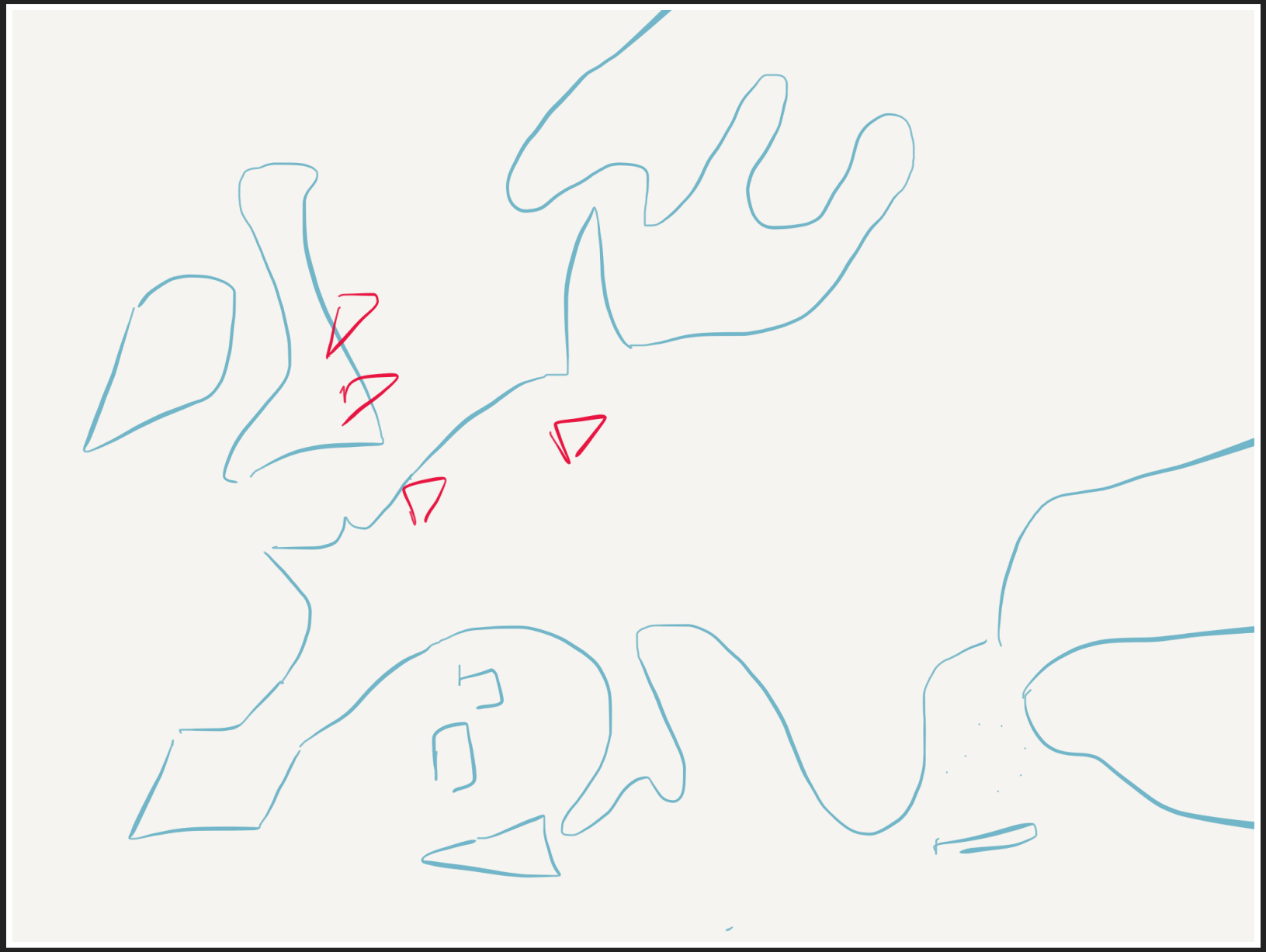
HOW DID WE GET HERE?

ORGANIC GROWTH

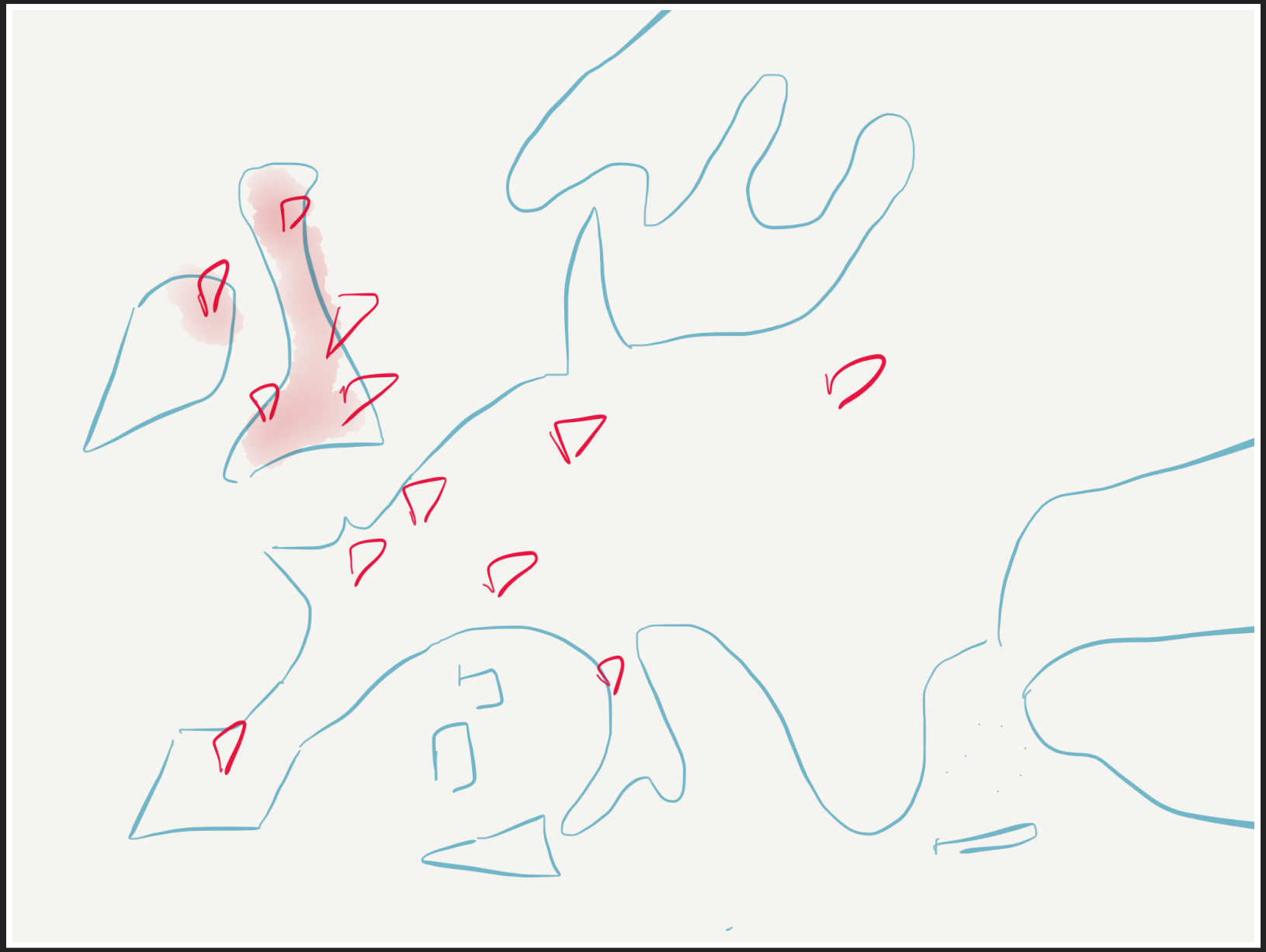
1 Retailer



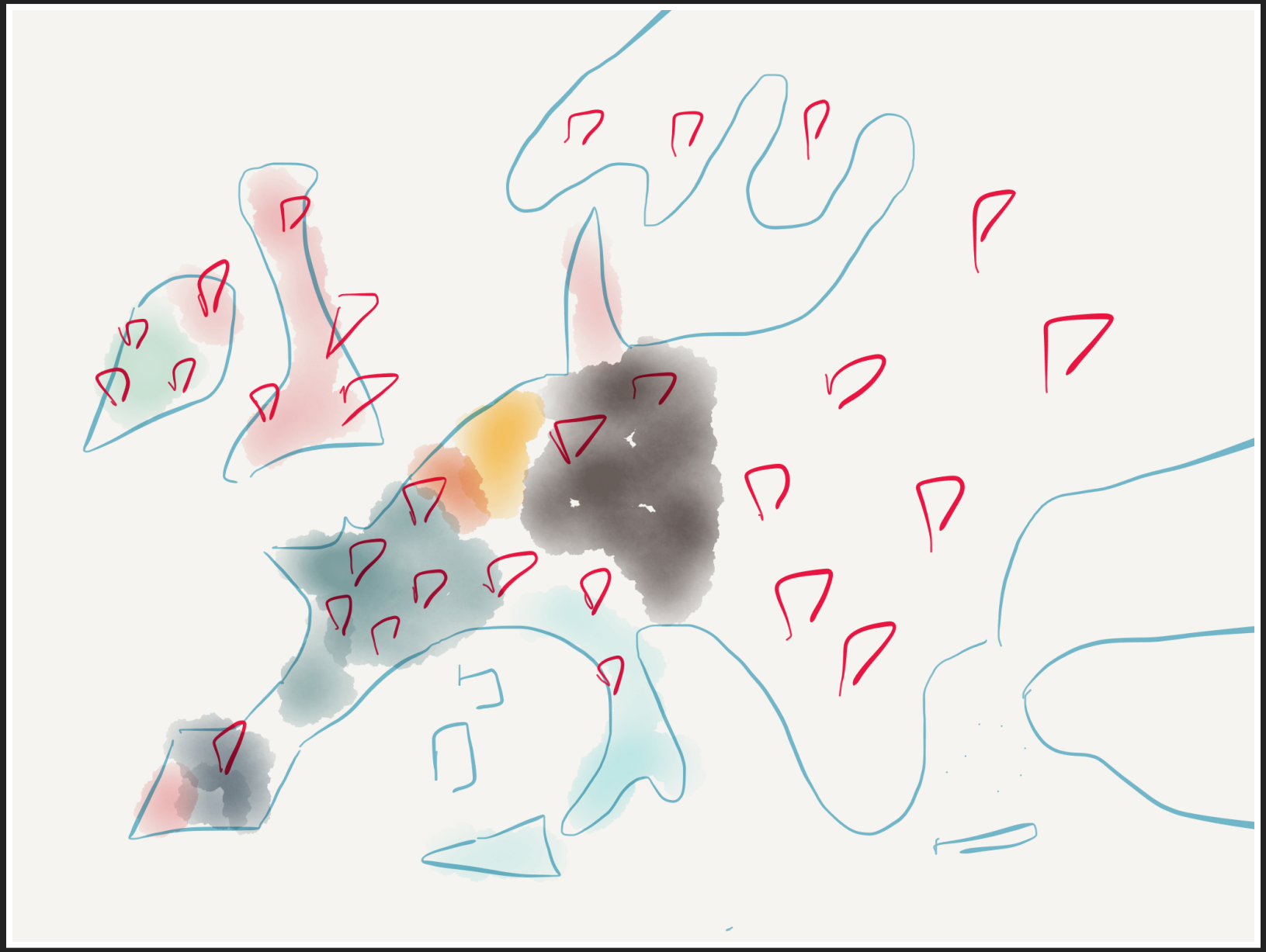
A few retailers



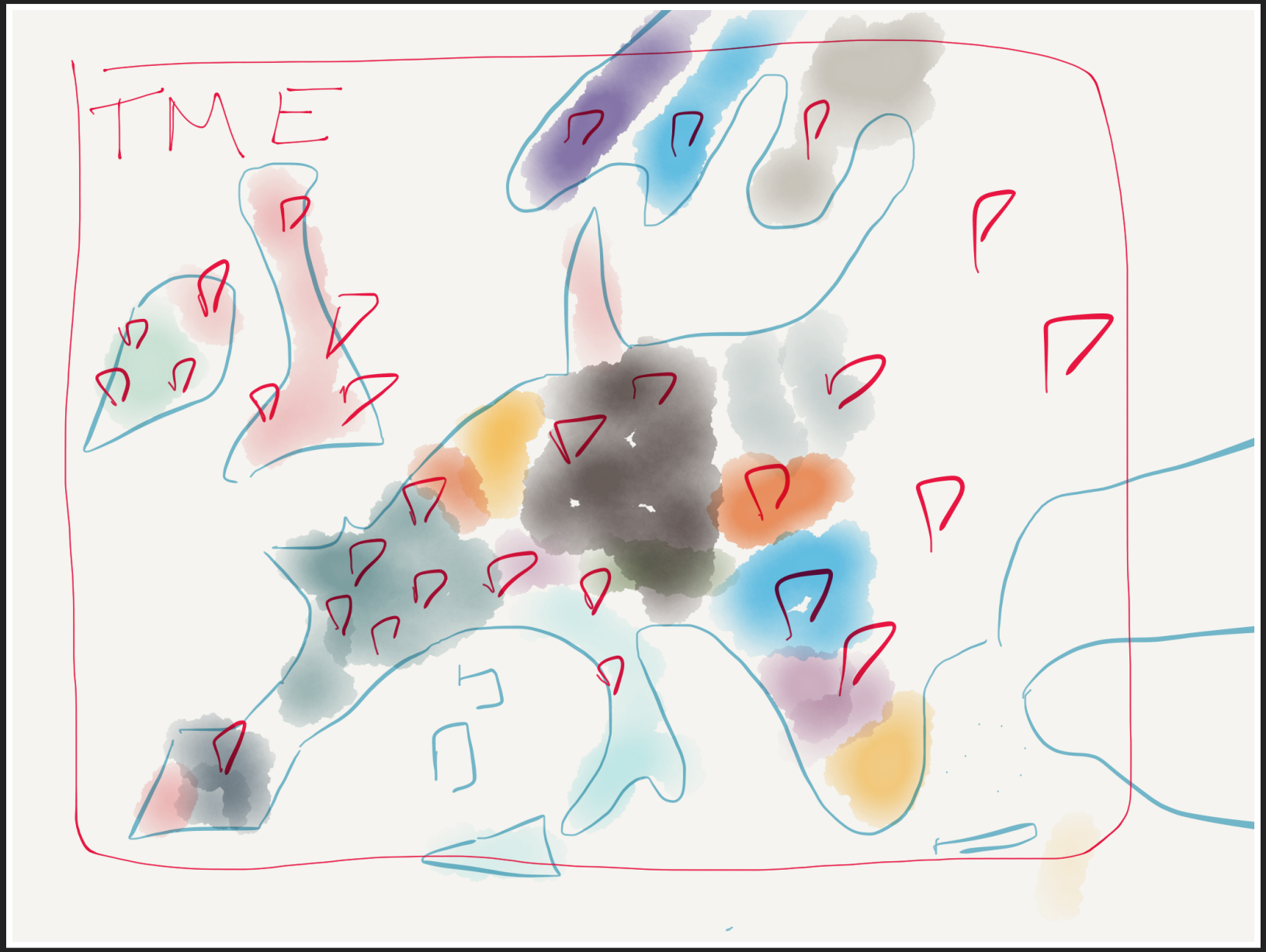
A lot retailers and 1 distributor



A lot of retailers and a few distributors

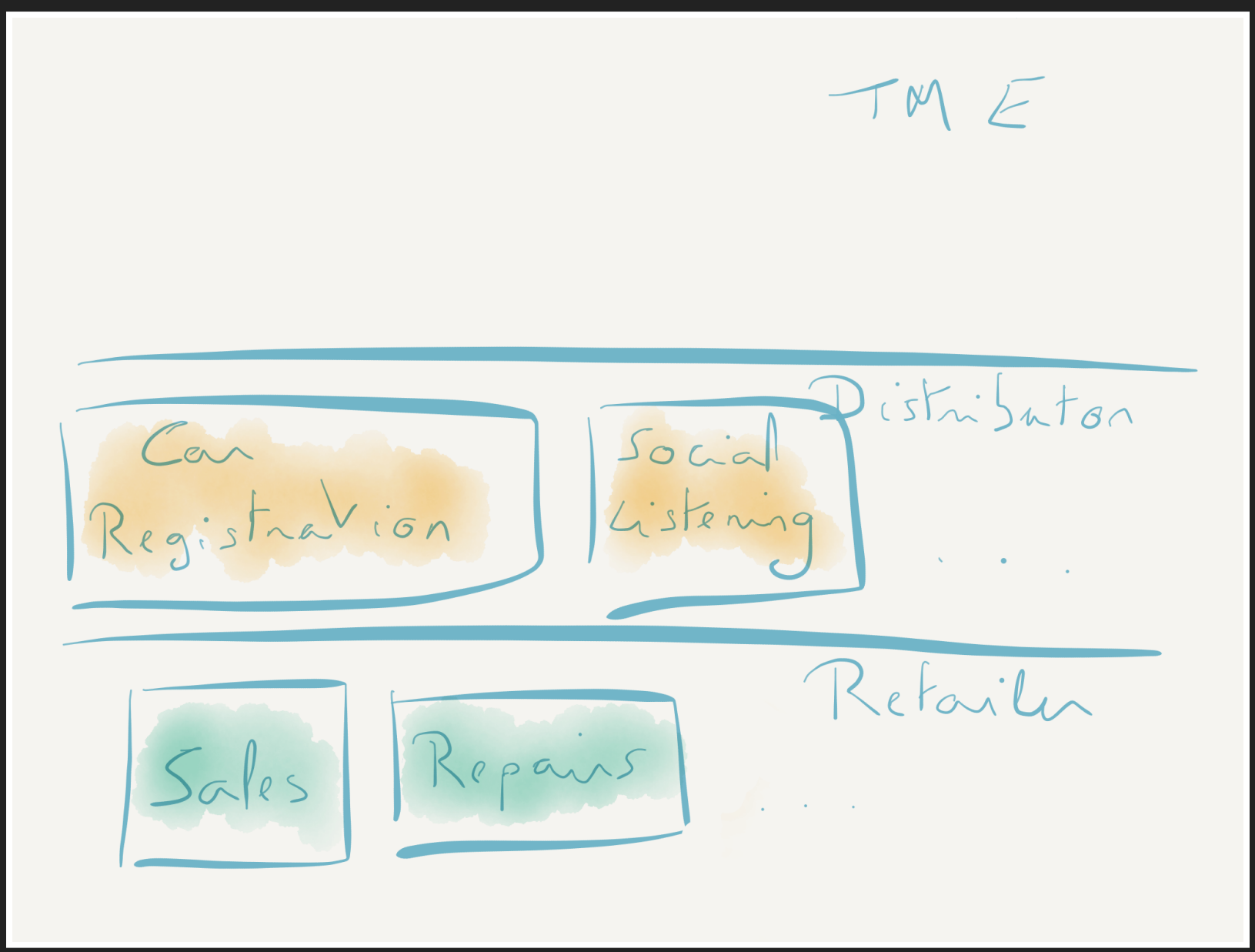


TME to centralise operations accros Europe

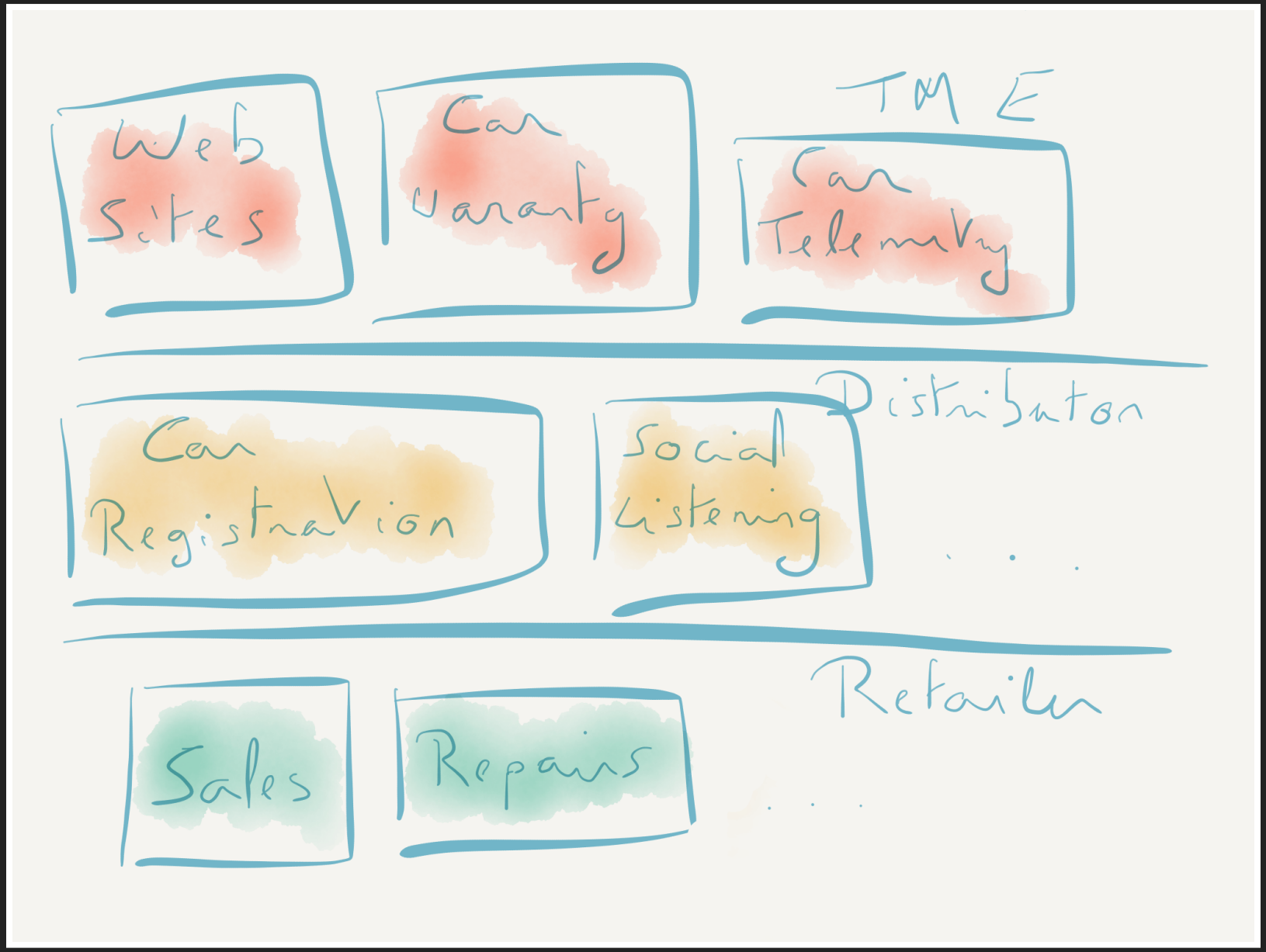

TME

RESULT

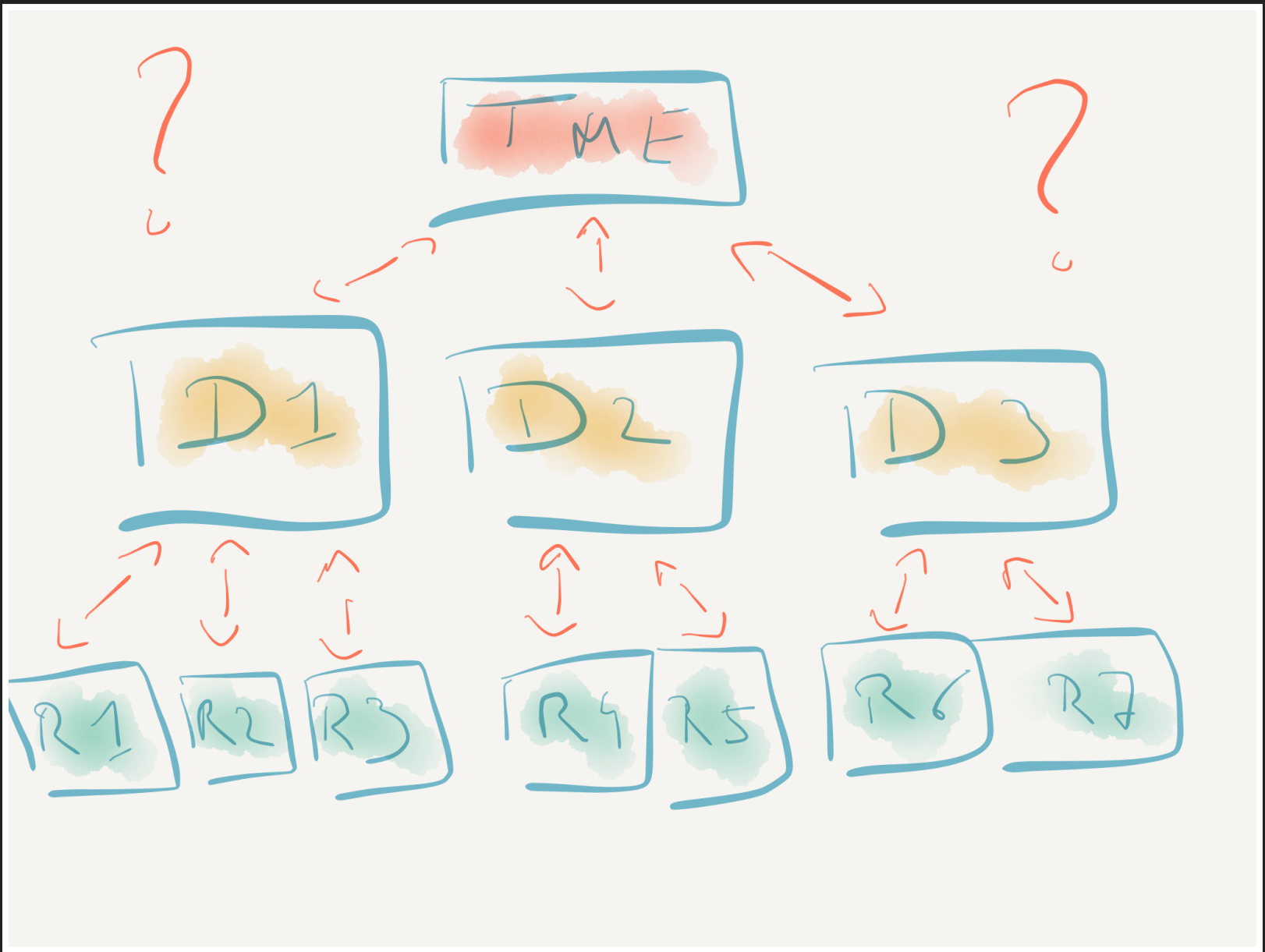
All of customer data lives close to the customer


TME
Distributon
Car Registration
Social Listening
. . .
Retailer
Sales
Repairs
. . .

But relevant data is increasingly incoming centrally

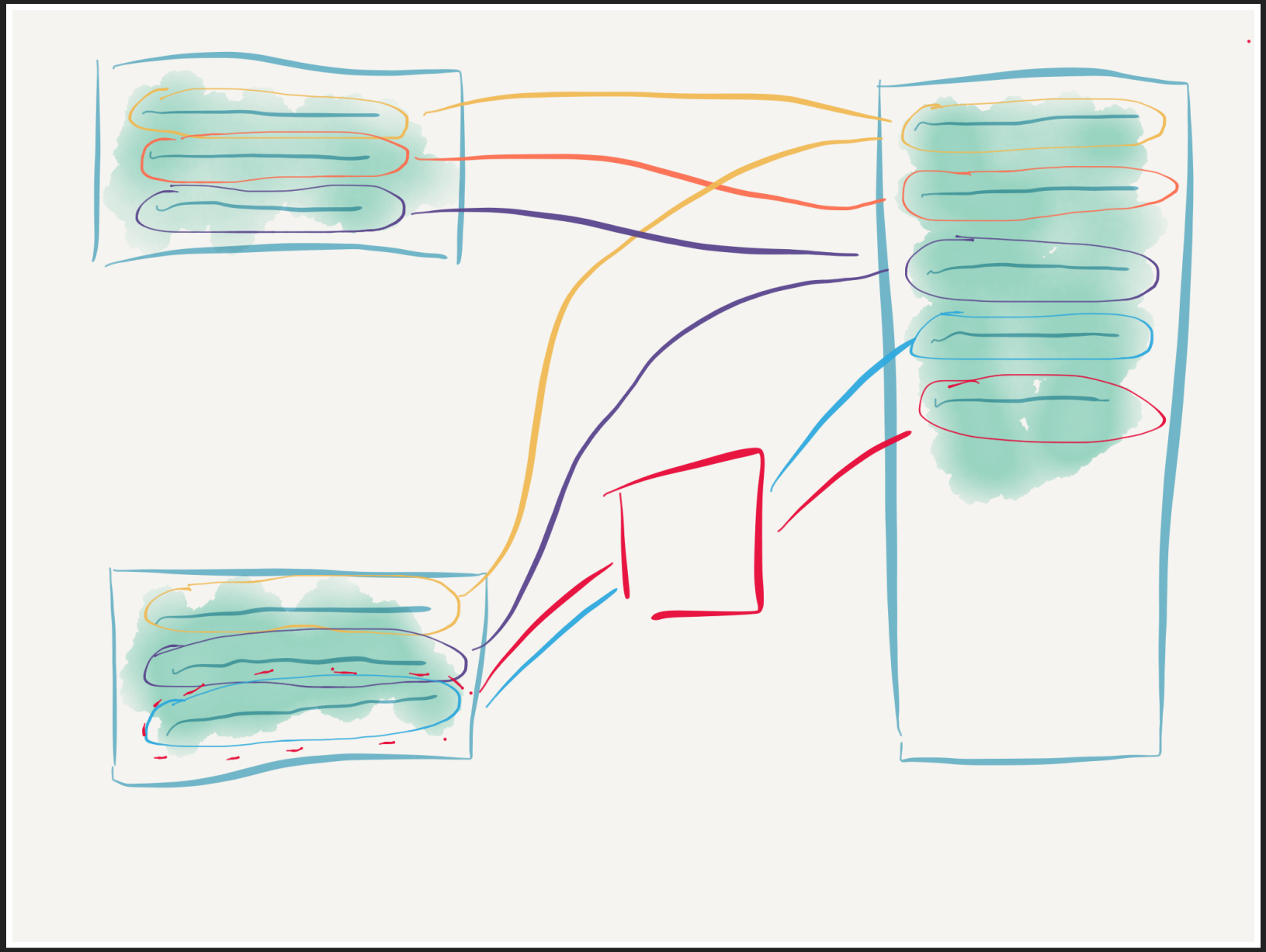

Web Sites
Car Warranty
TME
Car Telematics
Distribution
Car Registration
Social Listening
. . .
Retailer
Sales
Repairs
. . .

How do we link all this data together and feed it back?


?
TME
?
D1
D2
D3
R1
R2
R3
R4
R5
R6
R7

DATA INTEGRATION - INGESTION

Mapping 2 sources



Linking 2 sources

$$A \propto (B \mid C)$$

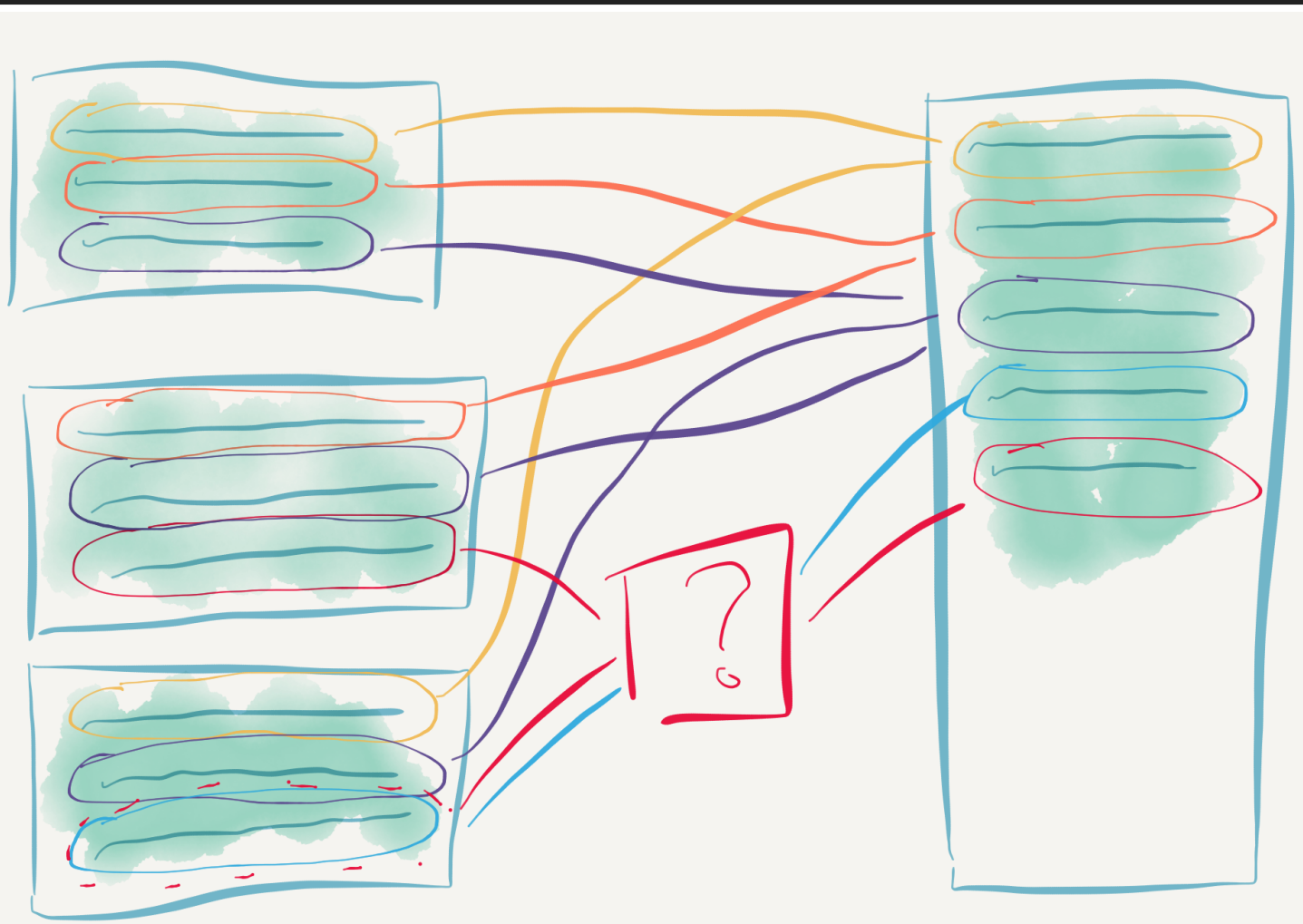
OR

$$A \propto (C \mid D)$$

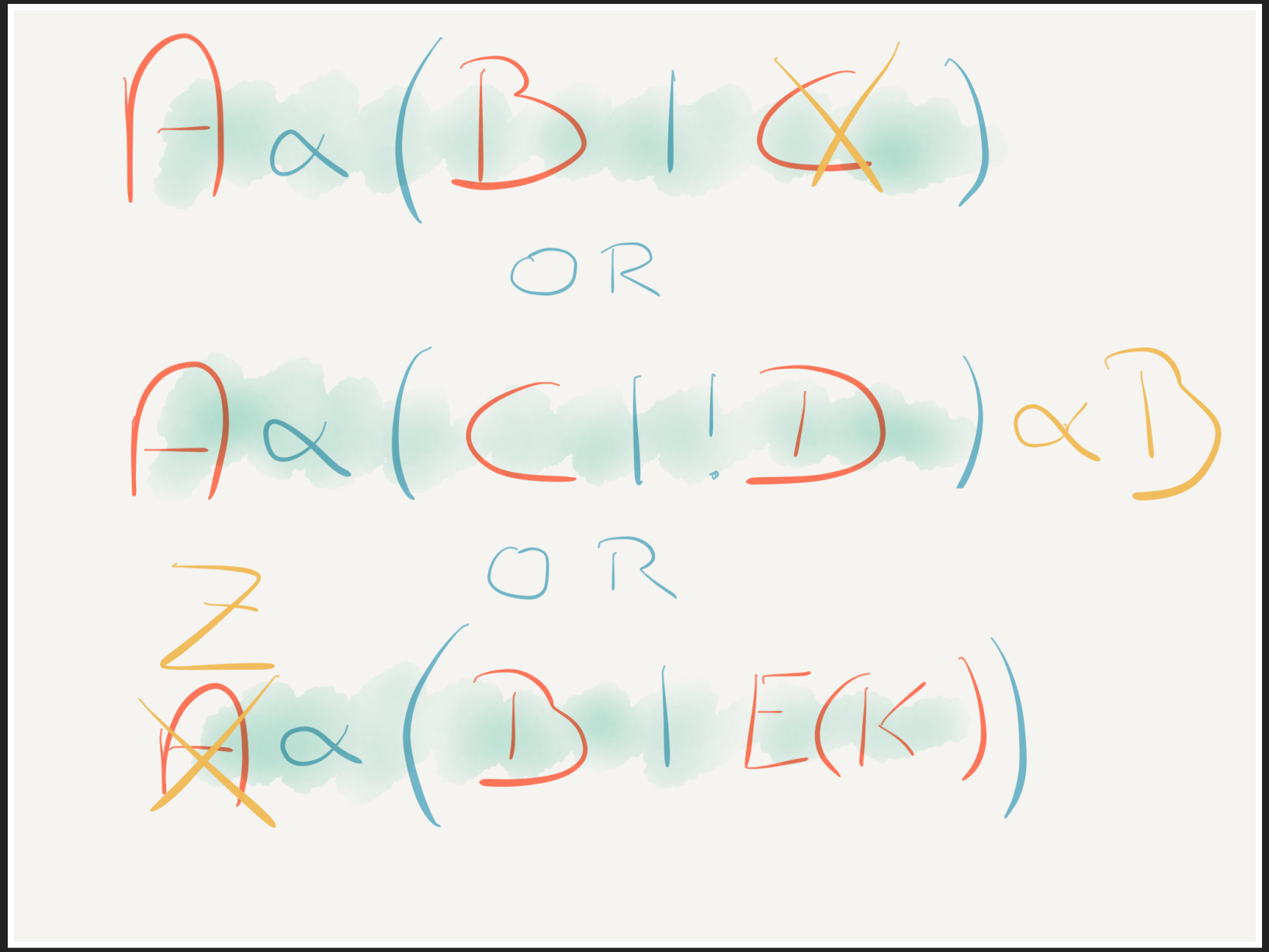
OR

$$A \propto (D \mid E(K))$$

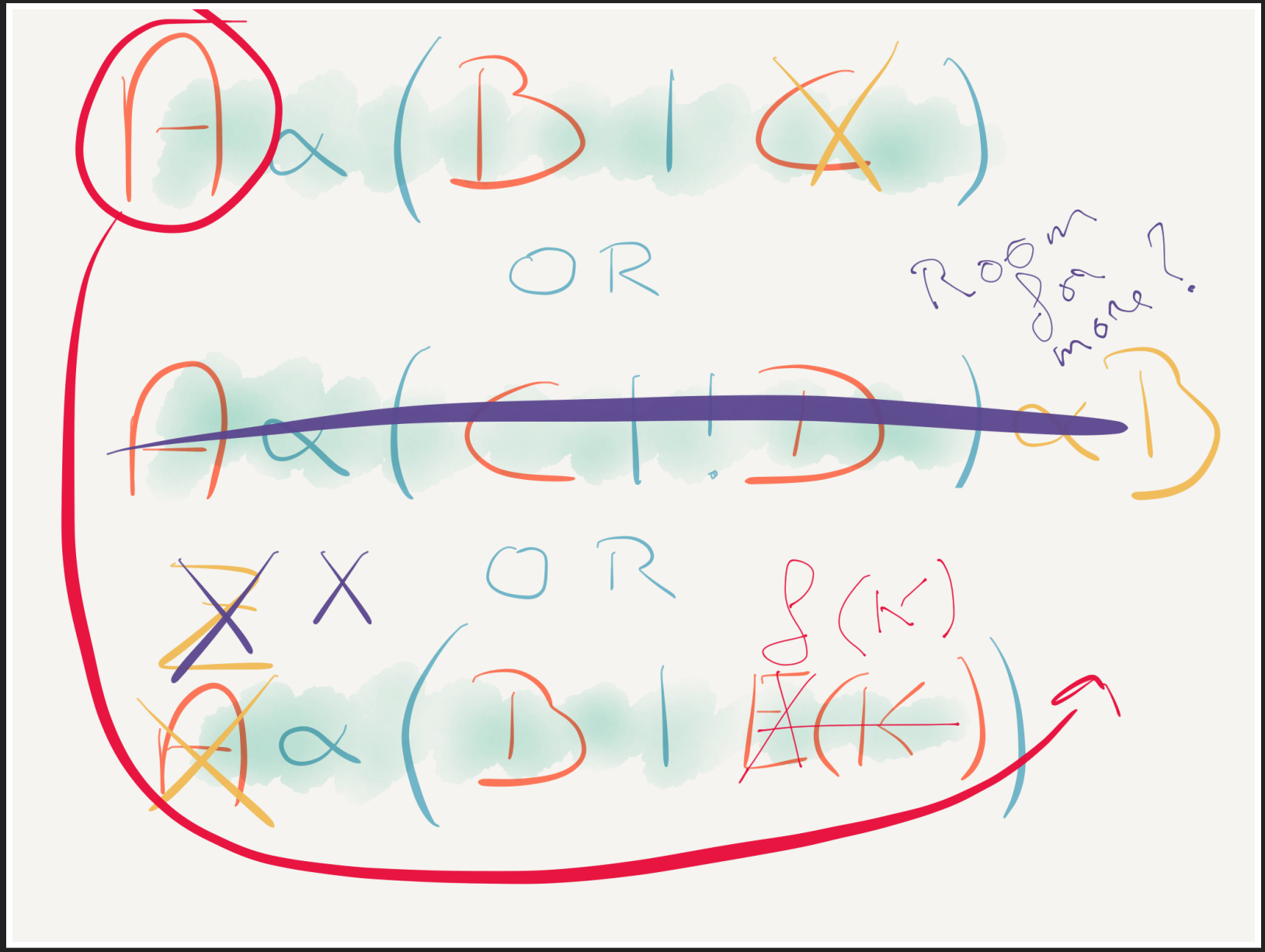
Mapping 3 sources



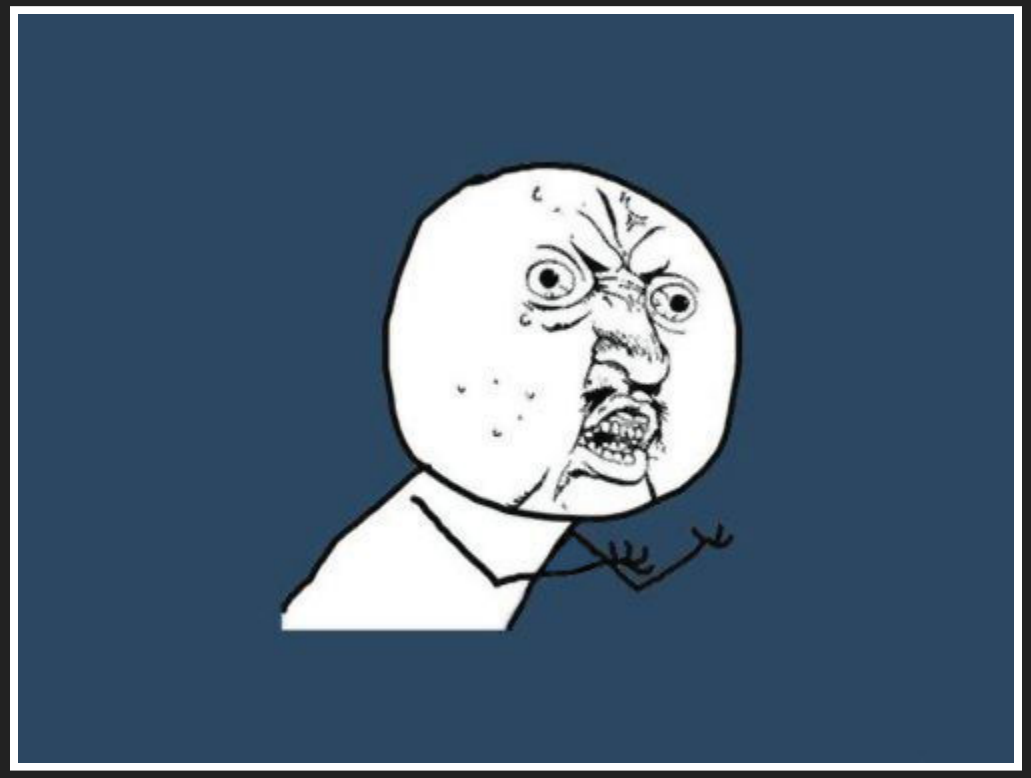
Linking 3 sources


A ∝ (B | X)
OR
A ∝ (C | ! D) ∝ D
OR
Z
A ∝ (D | E(K))

After some more time and sources

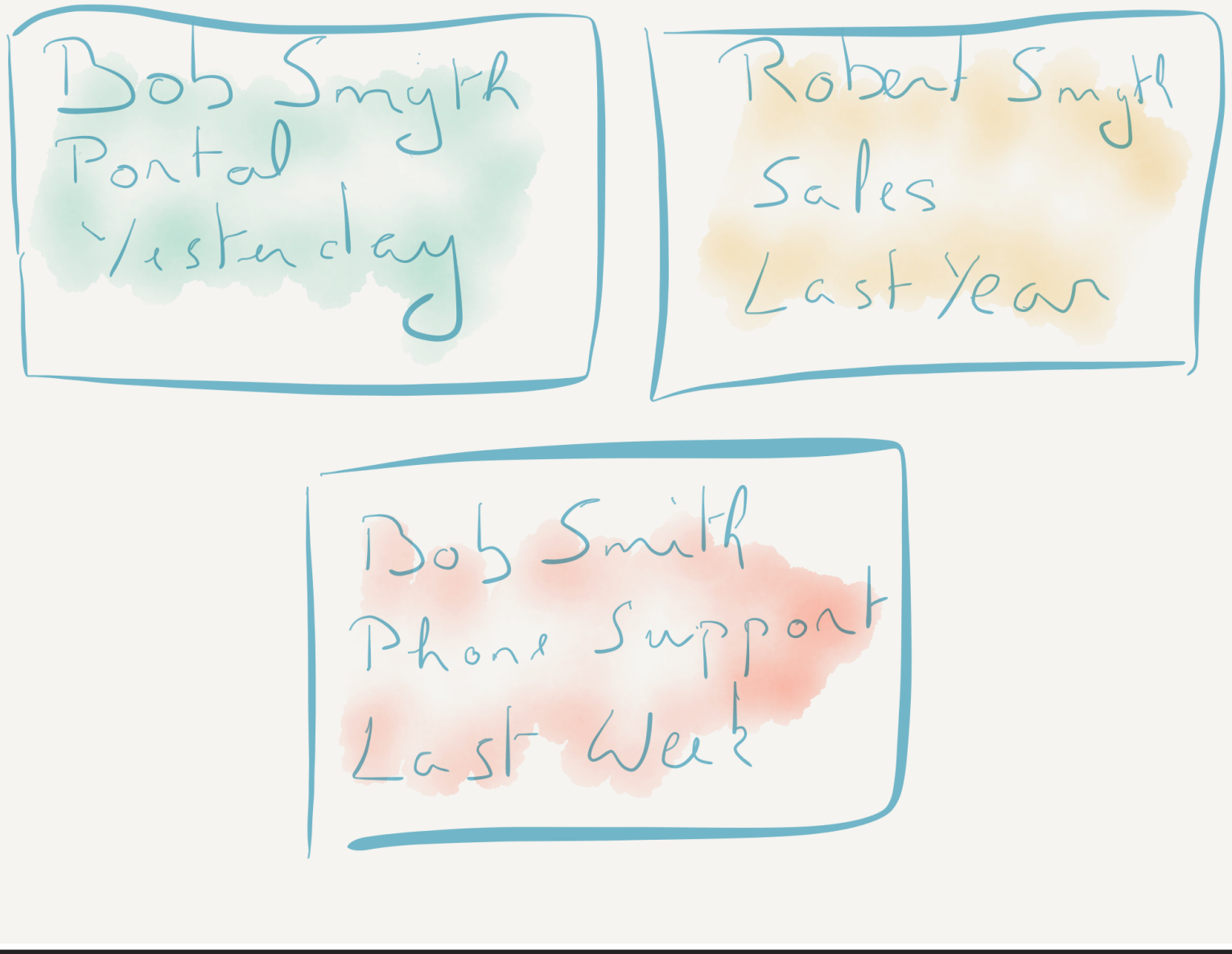

A
A ∝ (B | X)
OR
Room for more?
A ∝ (C | D) ∝ D
X
X OR
A ∝ (D | E(K))
p(K)

Y U NO WORK!!1!11

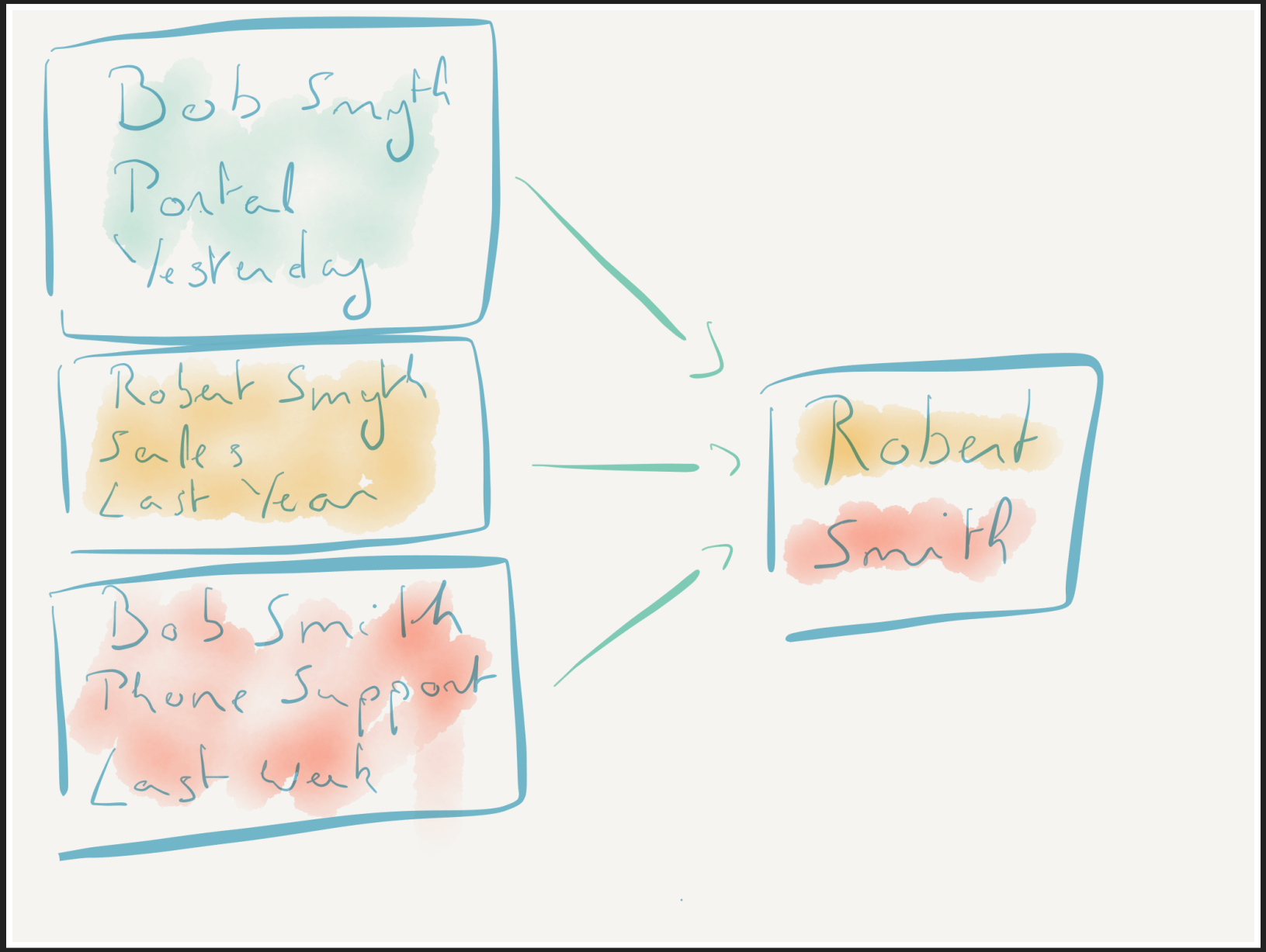


DATA INTEGRATION - CONSUMPTION

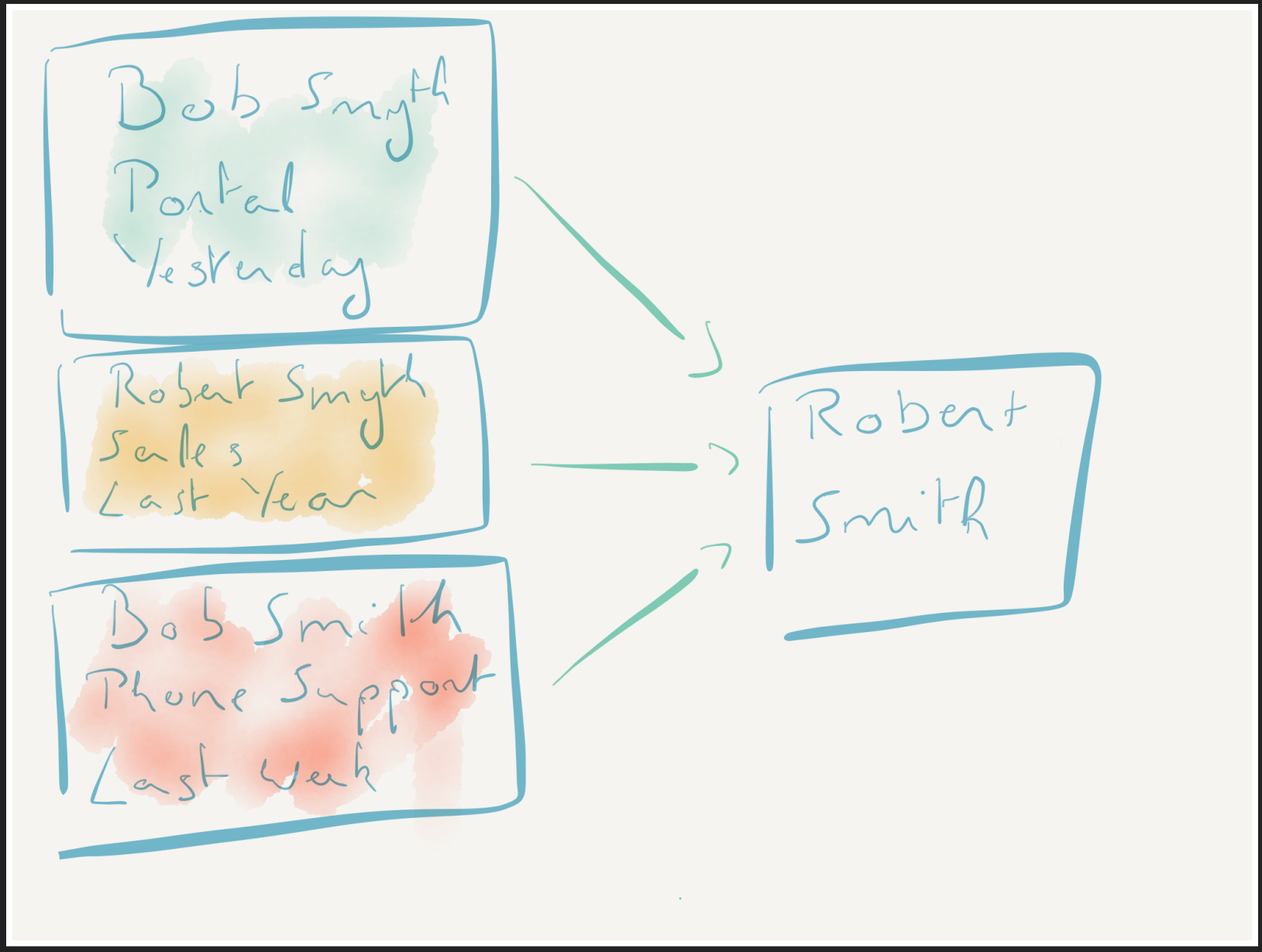
Which values do you retain?


Bob Smyth
Portal
Yesterday
Robert Smyth
Sales
Last Year
Bob Smith
Phone Support
Last Week

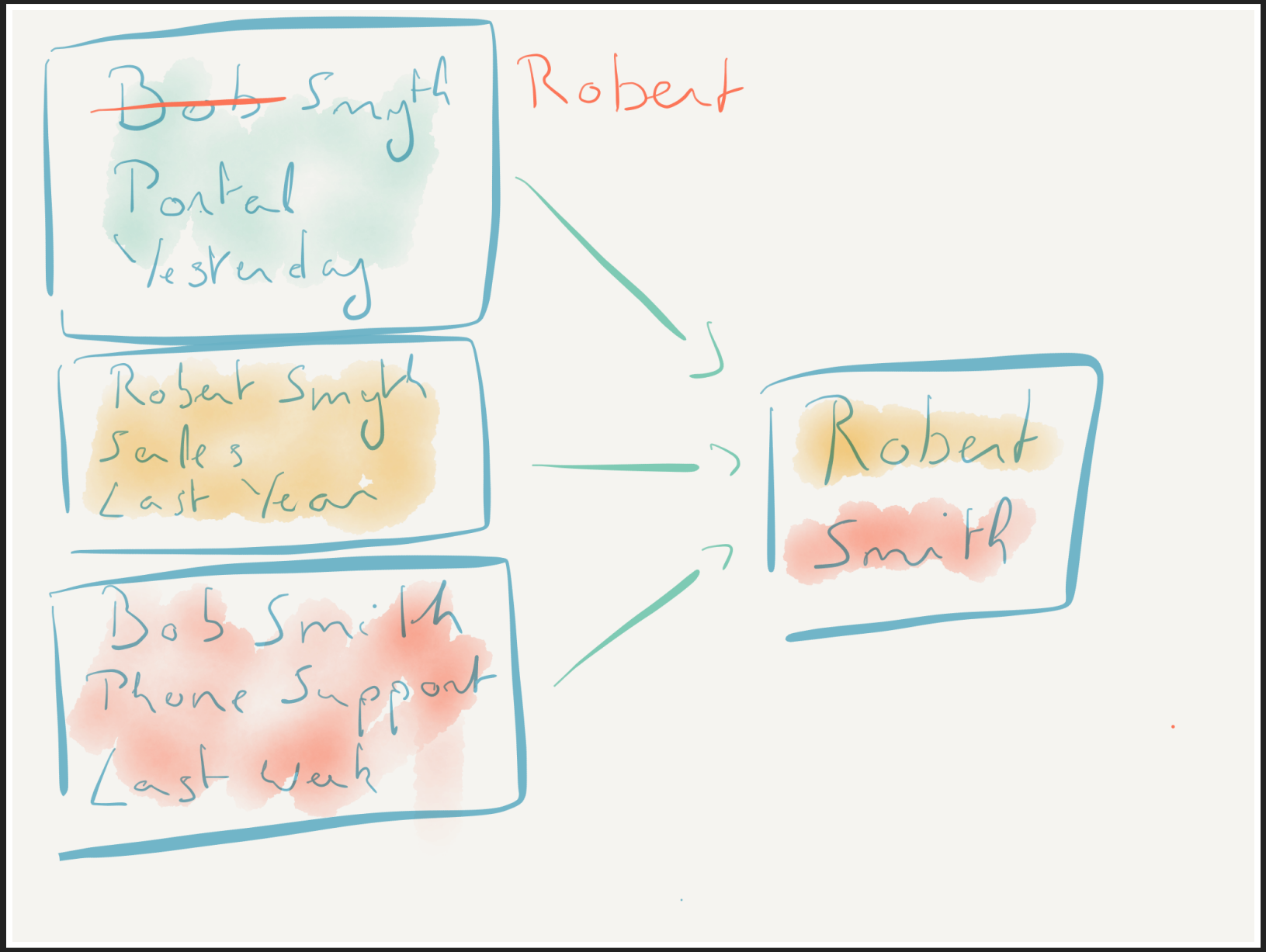
What is the origin of the data?


Bob Smyth
Portal
Yesterday
Robert Smyth
Sales
Last Year
Bob Smith
Phone Support
Last Week
Robert
Smith

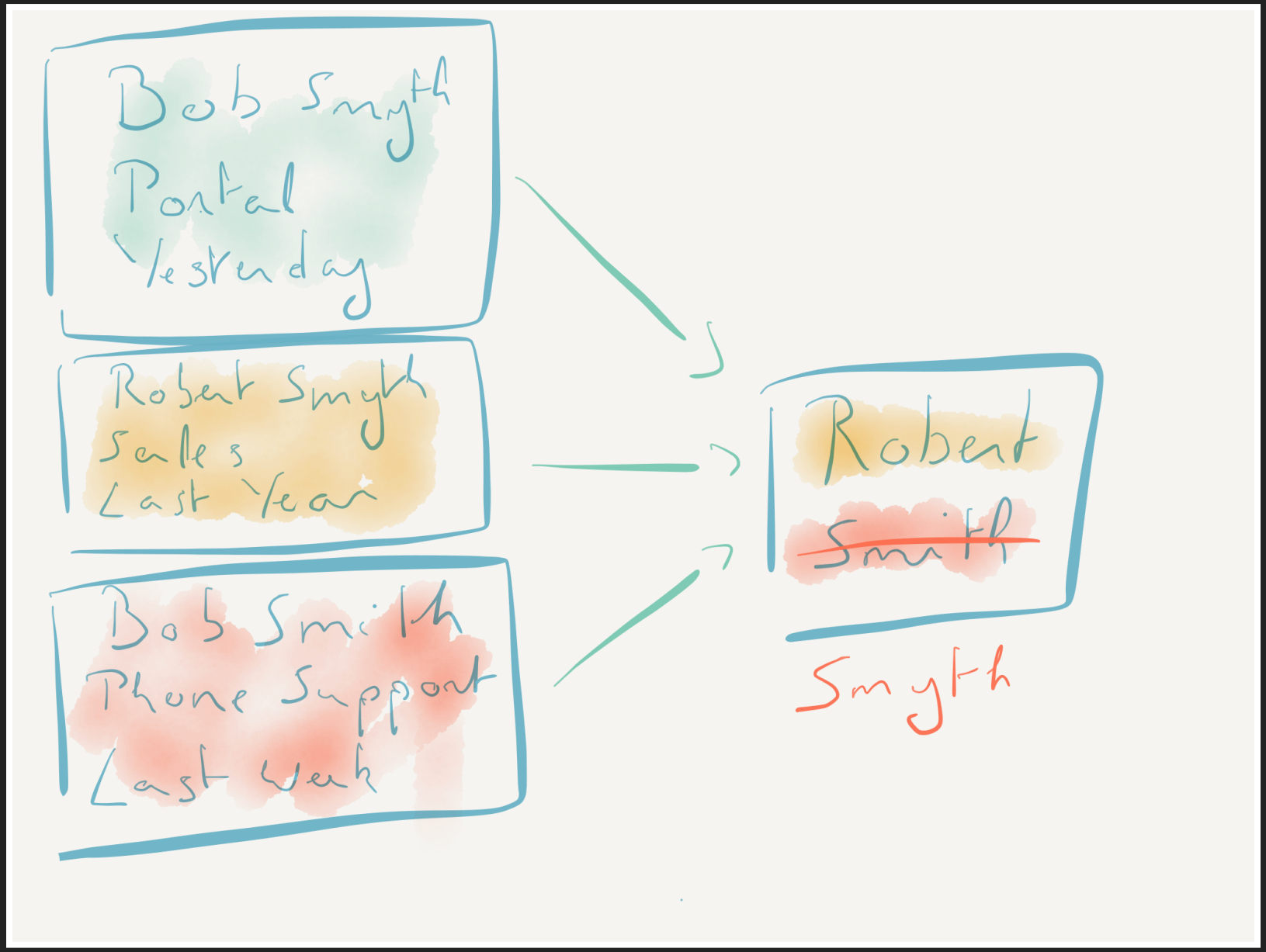
What if multiple consumers of the data have different requirements for the merged entity?


Bob Smyth
Portal
Yesterday
Robert Smyth
Sales
Last Year
Bob Smith
Phone Support
Last Week
Robert
Smith

How do you handle data updates in source systems?


~~Bob~~ Smyth
Portal
Yesterday
Robert
Robert Smyth
Sales
Last Year
Bob Smith
Phone Support
Last Week
Robert
Smith

How do you handle data updates in consuming systems?


Bob Smyth
Portal
Yesterday
Robert Smyth
Sales
Last Year
Bob Smith
Phone Support
Last Week
Robert
Smith
Smyth

DATA INTEGRATION

SLOW

- Write a ton of scripts
- Manually profile, clean and map the data

```
#!/bin/bash
¬
K=0¬
COUNT=`wc -l clusters_processed.json | awk '{print $1}'`¬
ITERATIONS=$[$COUNT / 1000]¬
REMAINDER=$[$COUNT % 1000]¬
while [[ $K -lt $ITERATIONS ]]; do¬
··echo "Loop iteration $K"¬
··OFFSET=$[$K * 1000]¬
··END=$[$OFFSET + 1000]¬
··OFFSET=$[$OFFSET + 1]¬
··echo [ `sed -n "${OFFSET},${END}p" clusters_processed.json | sed 's/$/,/g' | sed '$s/,$//'` ] | curl -H 'authori
··echo ""¬
··echo "curl statement completed"¬
··K=$[$K + 1]¬
done;¬
K=$[$K + 1]¬
OFFSET=$[$K * 1000]¬
END=$[$OFFSET + $REMAINDER]¬
OFFSET=$[$OFFSET + 1]¬
echo [ `sed -n "${OFFSET},${END}p" clusters_processed.json | sed 's/$/,/g' | sed '$s/,$//'` ] | curl -H 'authoriza
```

COSTLY

- Only scales with people
- Very often outsourced

OPAQUE

- No documentation
- No audit trail

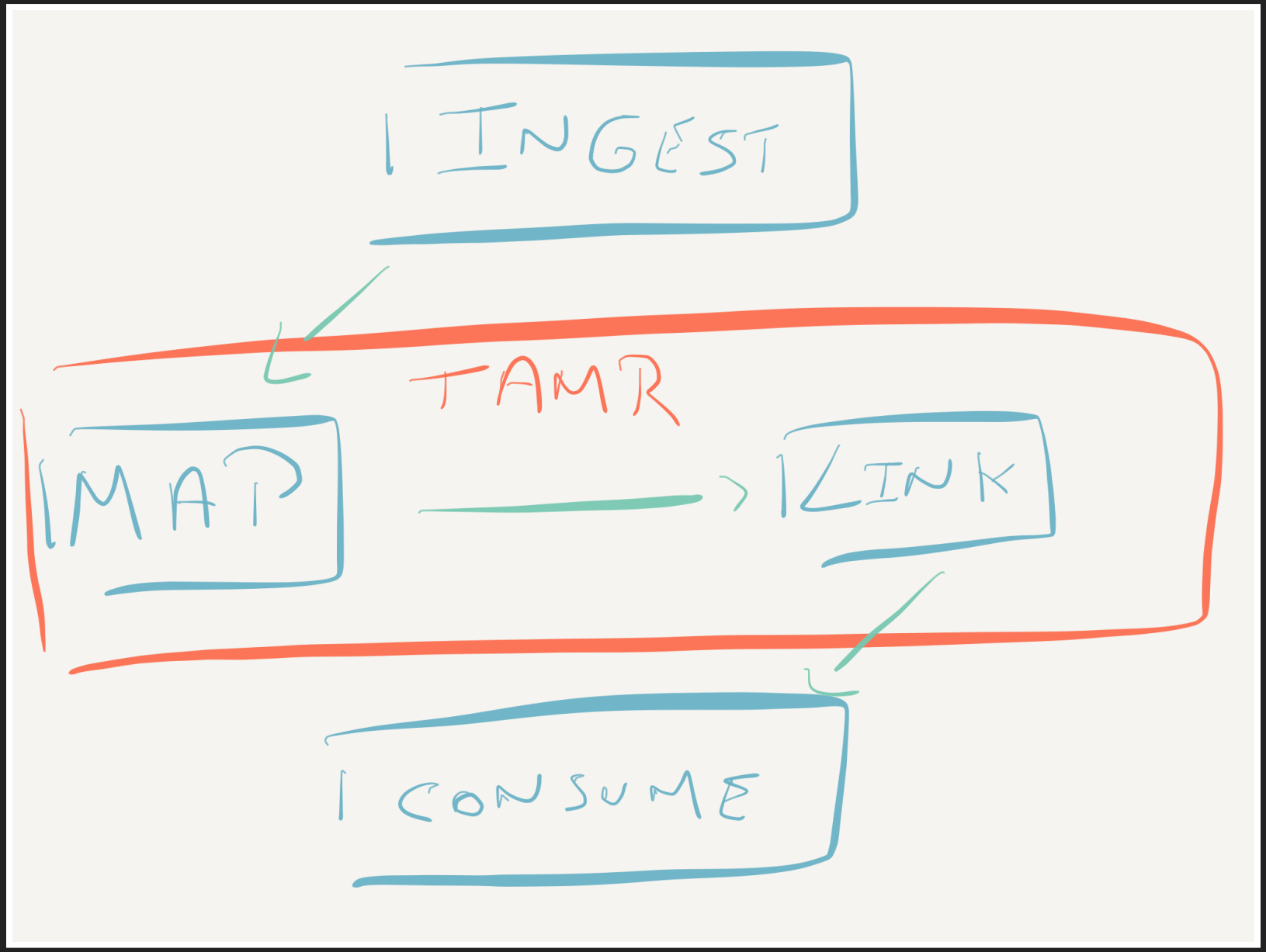
INEFFICIENT

- Low quality results
- Hard to keep up to date

DATA UNIFICATION

1. Forward flow
2. Feedback flow

FORWARD FLOW

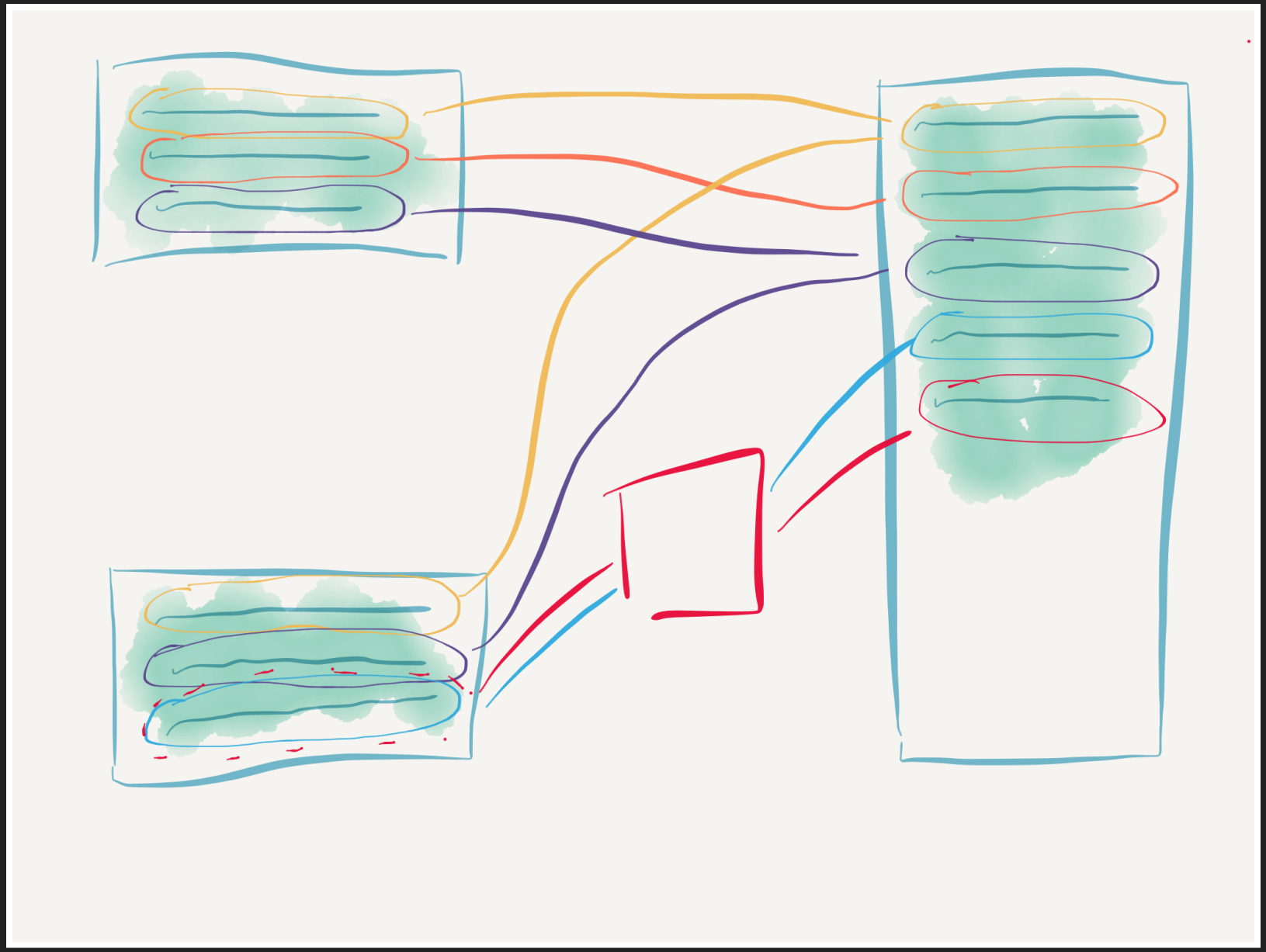

INGEST
TAMR
MAP
LINK
CONSUME

INGEST THE DATA

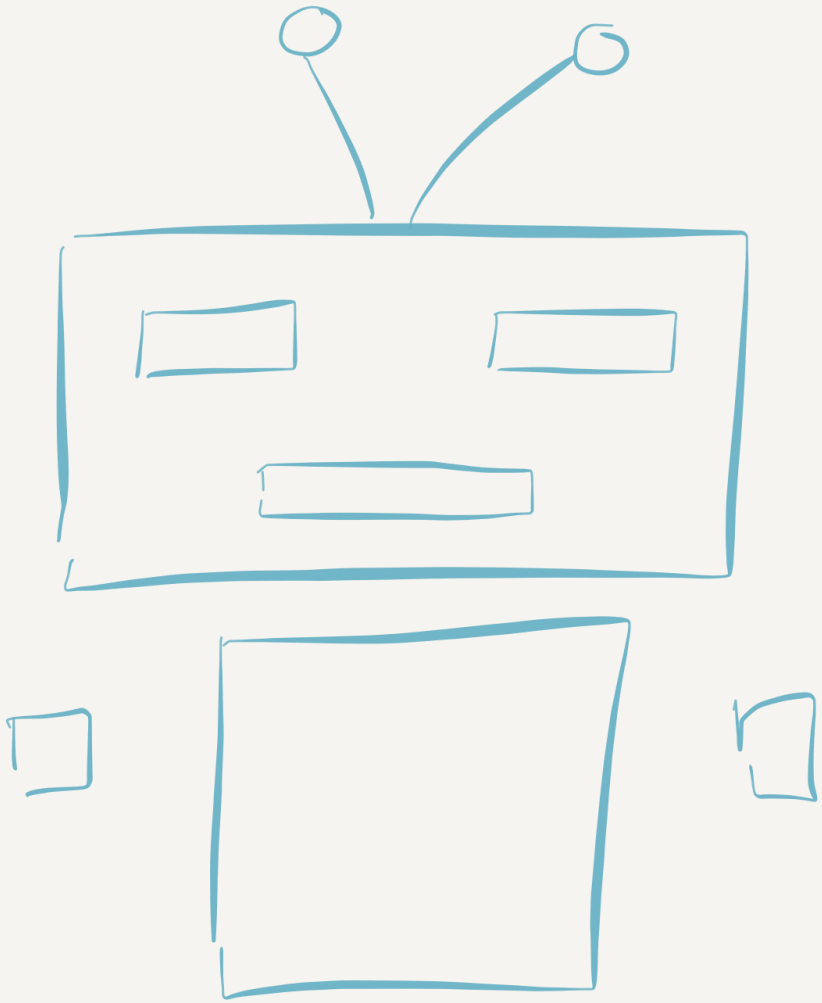
Ingest in schema of source system (Removes friction)

MAP THE DATA

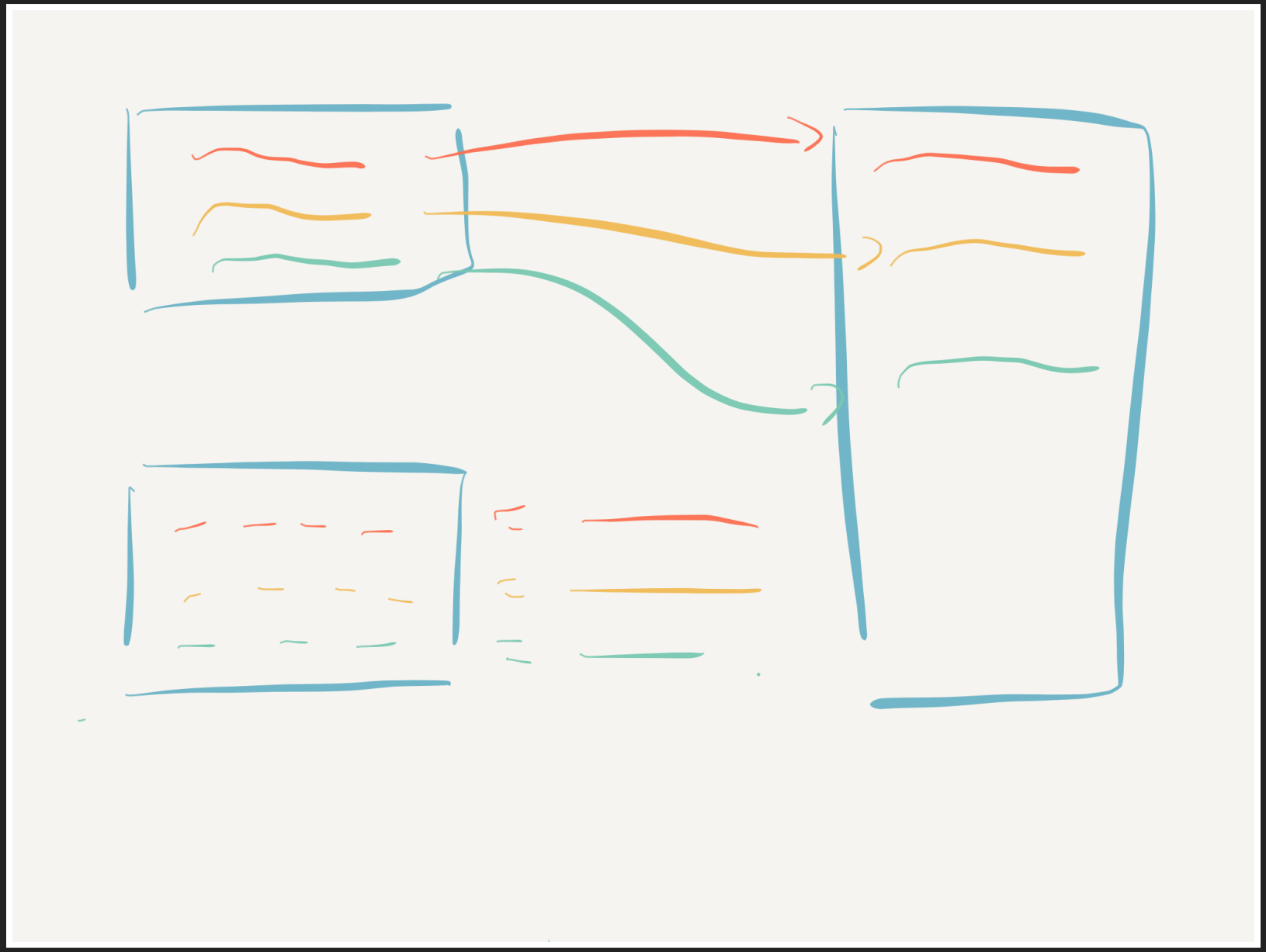
Map fields from source systems into the target schema.



Machine learning assisted



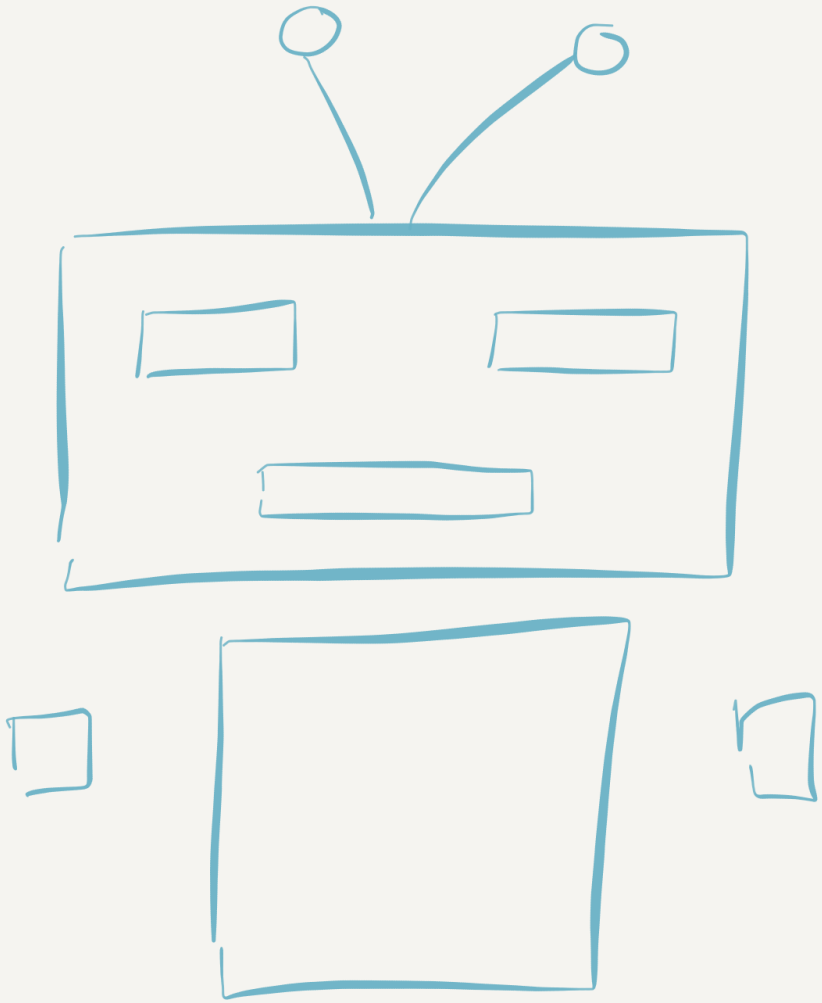
Based on your mappings and the statistical profile of the attributes Tamr will offer mapping recommendations



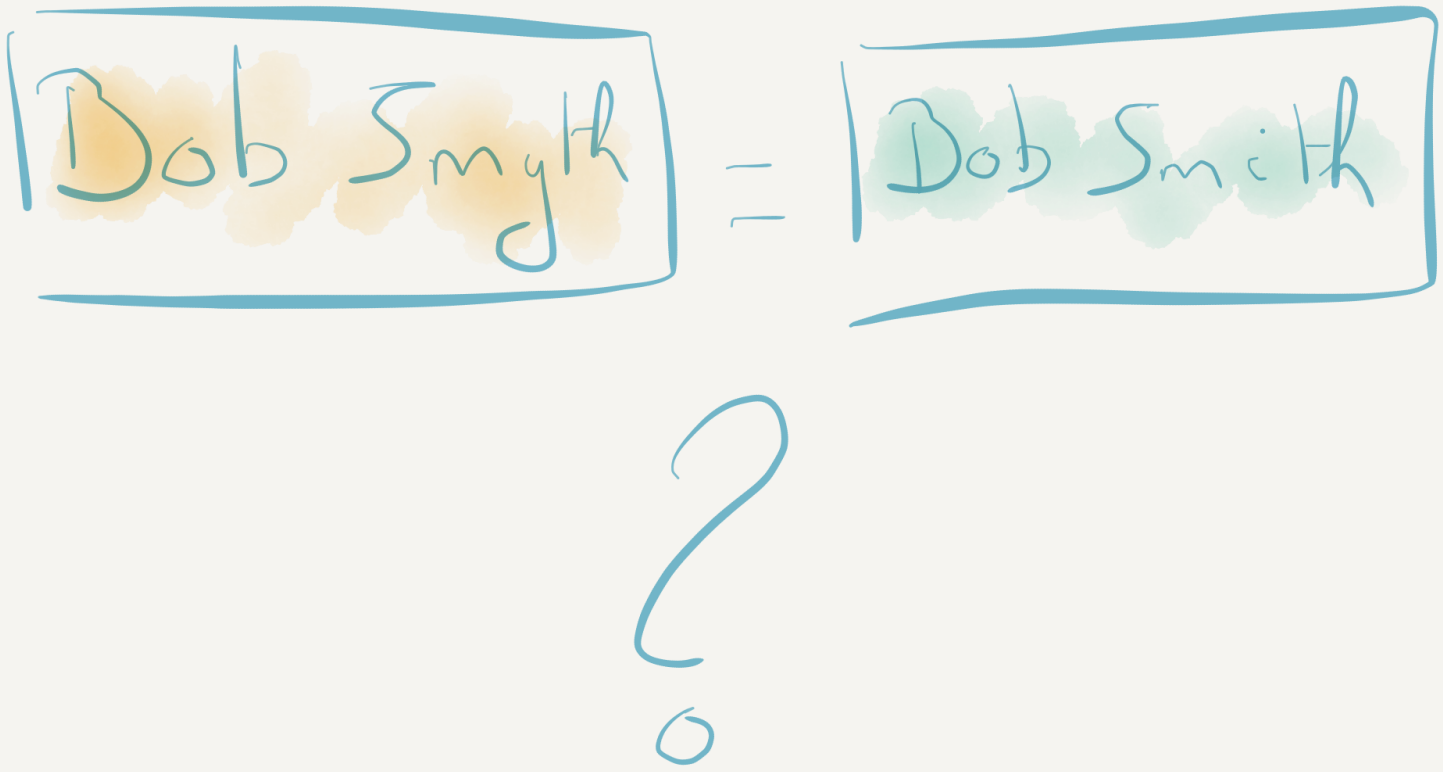
Effort of mapping the data goes down as more sources are integrated

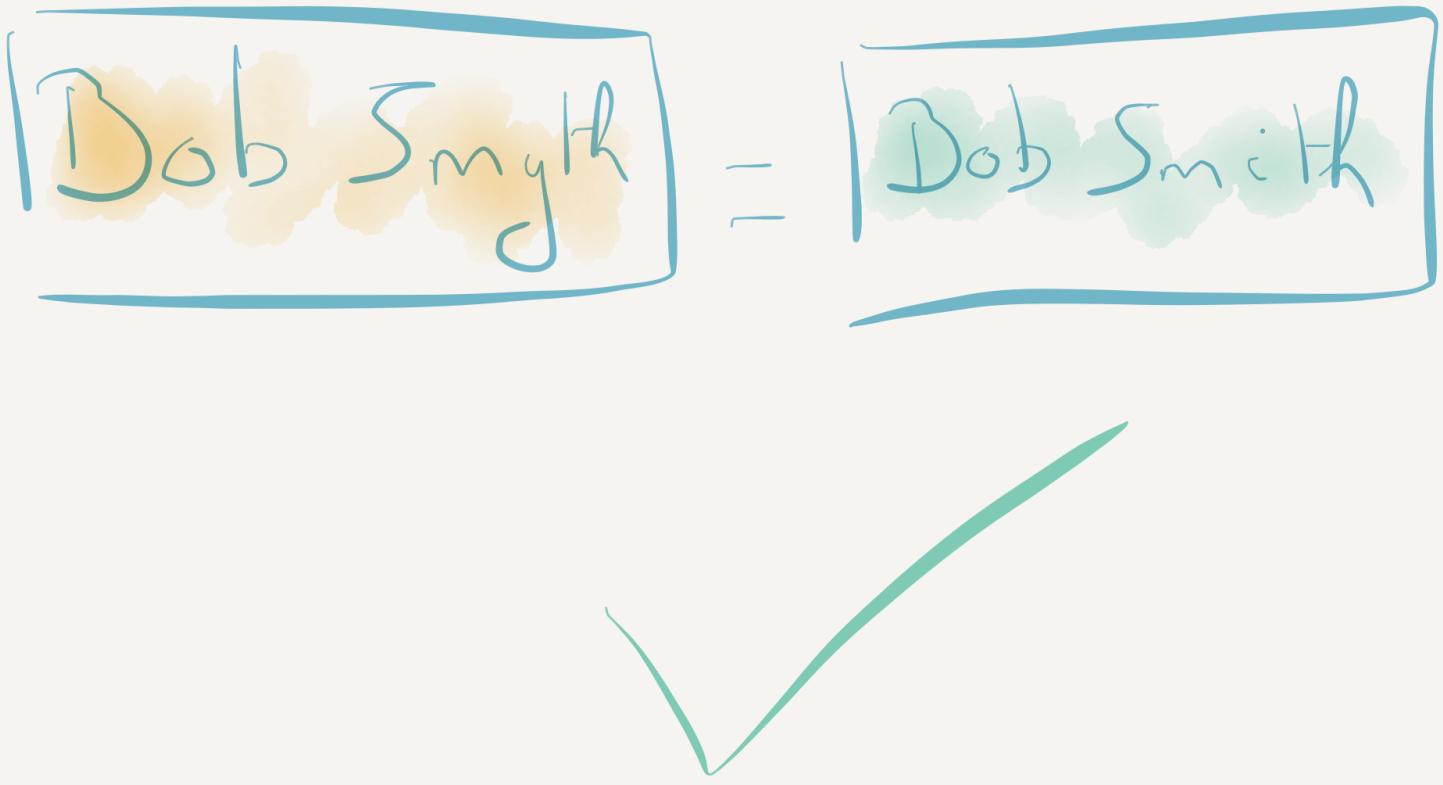
LINK THE DATA

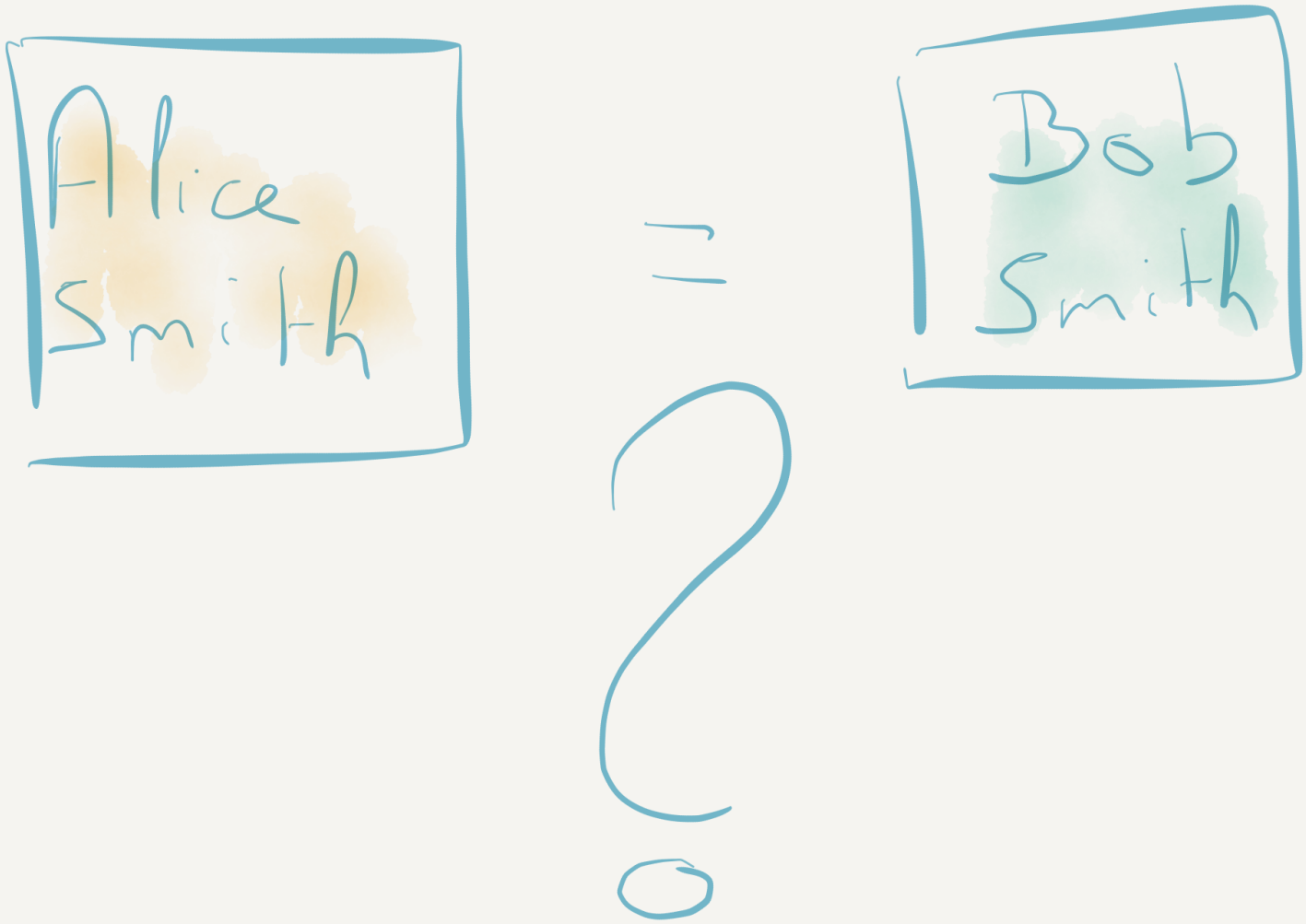
Use ML model to link entities across all the records

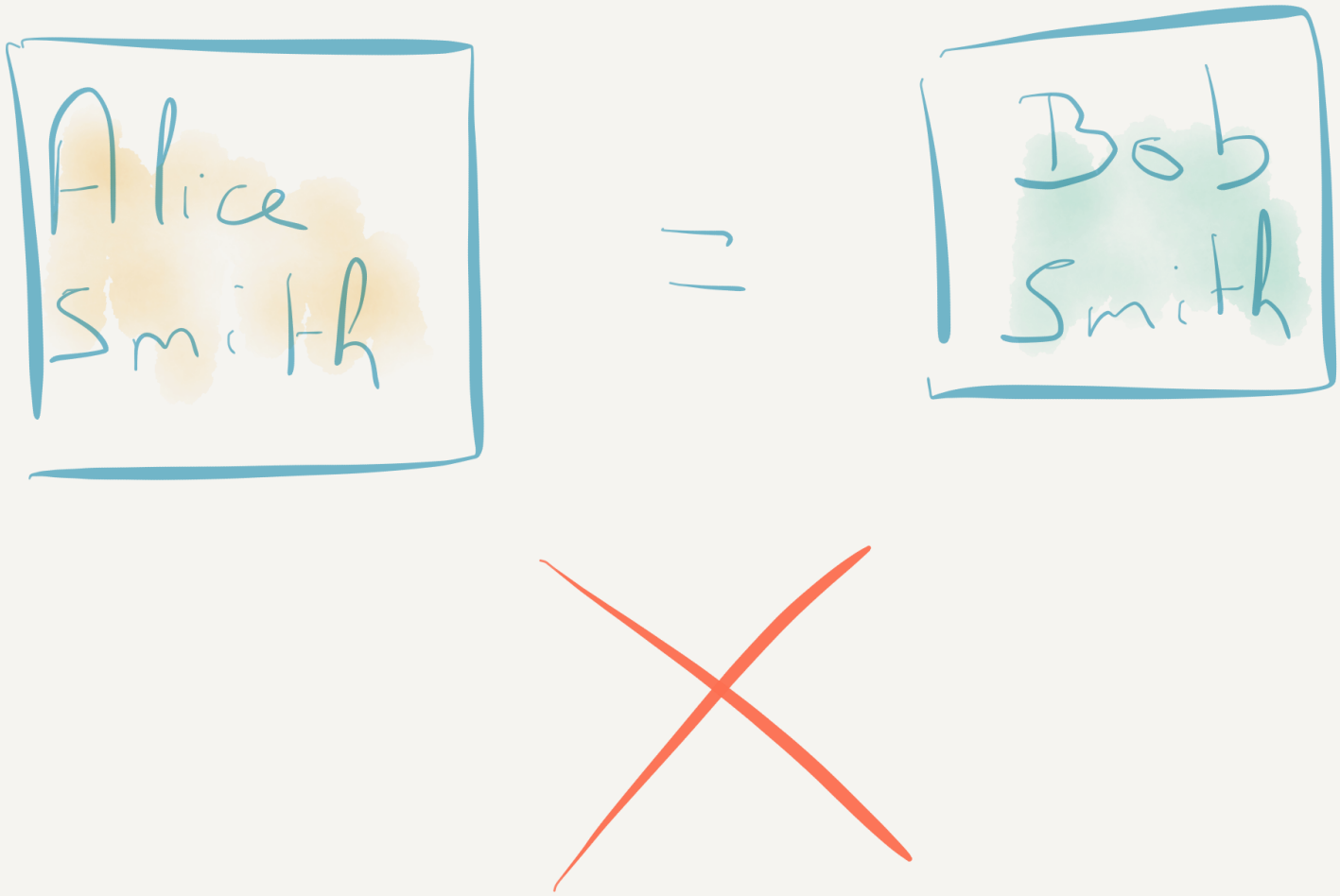


Train model by evaluating record pairs


Bob Smyth
=
Bob Smith
?


Dob Smyth
=
Dob Smith


Alice
Smith
=
?
Bob
Smith

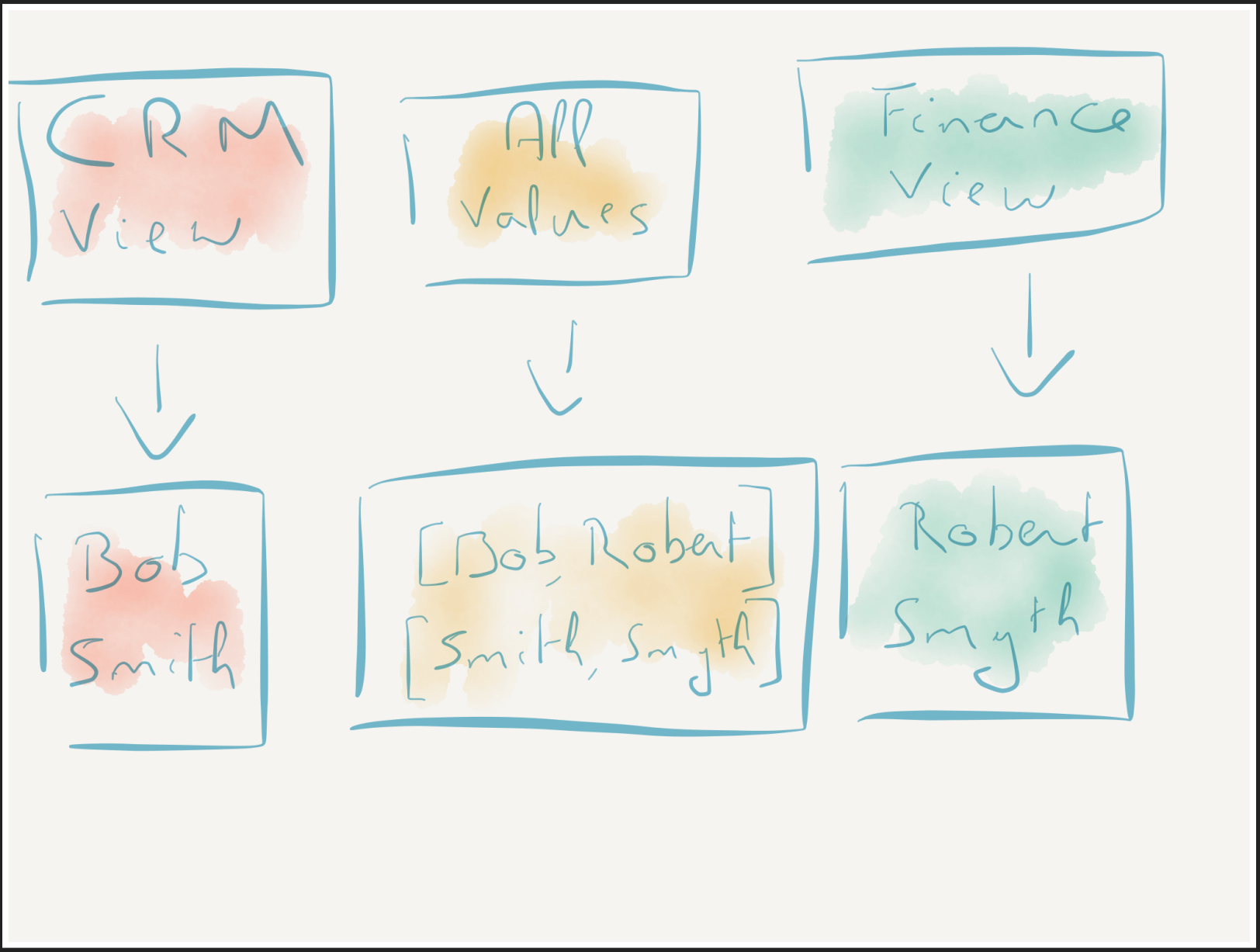

Alice
Smith
=
Bob
Smith

A business user can do this!

Regularly train new pairs to keep the model in sync with changes in the sources

CONSUME/MERGE THE DATA

Multiple views that essentially describe how individual fields are merged


CRM View
All Values
Finance View
Bob Smith
[Bob, Robert]
[Smith, Smyth]
Robert Smyth

Views are functions over the data -> flexibility

FEEDBACK FLOW

Save all updates as immutable events

Record 1

[Bob Smyth t-0 sales] (1)

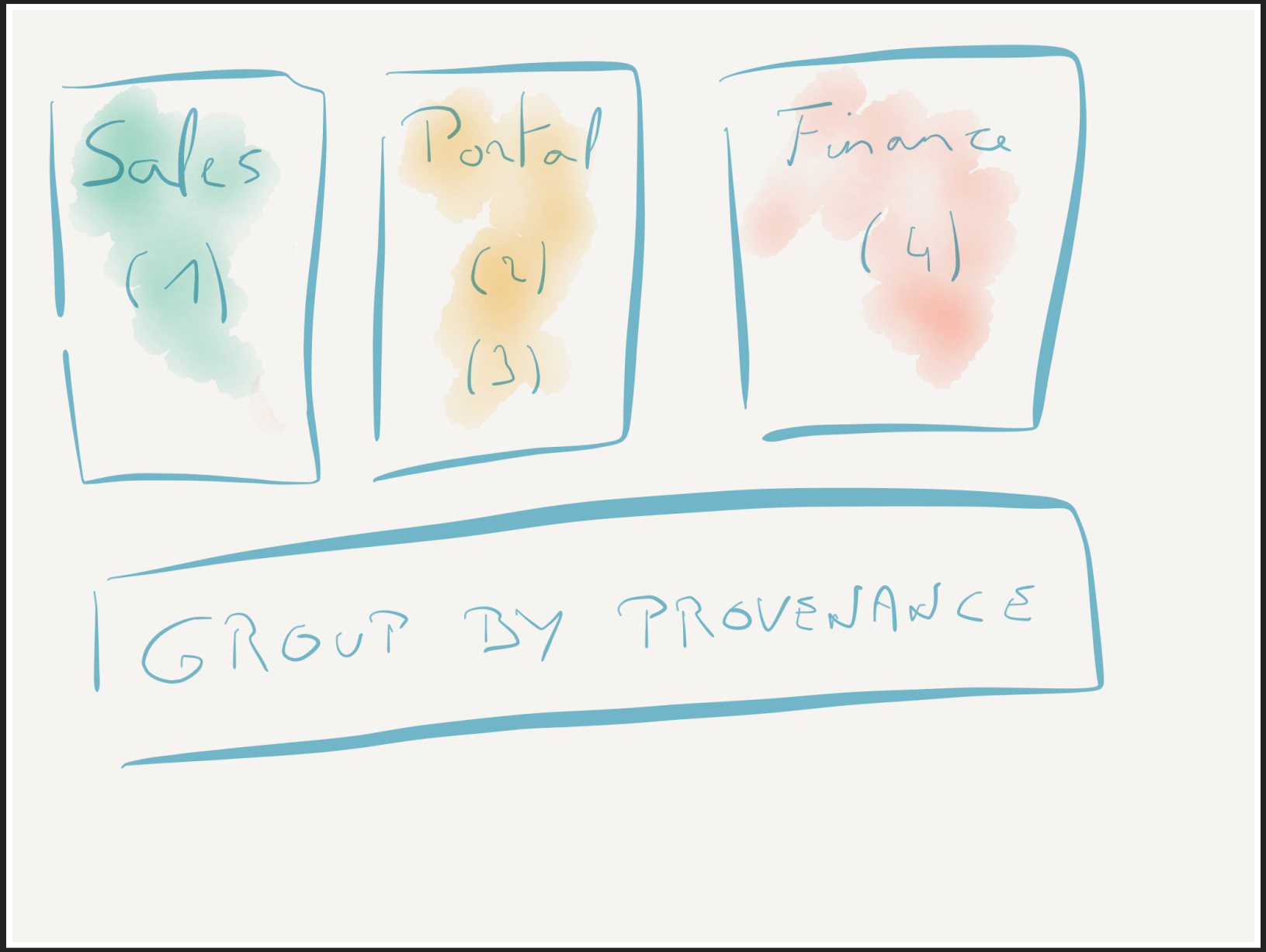
[Robert Smith t-1 portal] (2)

[Bob Smith t-2 portal] (3)

[Robert Smyth t-16 finance] (4)

GROUP BY RECORD_ID

Create "pseudo" sources by providence


Sales
(1)
Portal
(2)
(3)
Finance
(4)
GROUP BY PROVENANCE

ML model will group events into clusters

1. Integrates legacy sources
2. Scales well as a function of sources
3. Obtains knowledge where it resides
4. Supports multiple views/uses of the same base data
5. Avoids the creation of additional sources
6. Allows creation of new processes on top of consolidated data

QUESTIONS?