



# ChARLES: Change-Aware Recovery of Latent Evolution Semantics in Relational Data

Shiyi He, Alexandra Meliou, Anna Fariha

## How did my data change? Why?

name	gender	education	experience	salary	bonus
Anton	M	PhD	2	\$230,000	\$23,000
Bob	M	PhD	3	\$250,000	\$25,000
Jim	M	MS	5	\$160,000	\$16,000
Allen	M	MS	1	\$130,000	\$13,000
Cathy	F	BS	2	\$110,000	\$11,000
Tom	M	MS	4	\$150,000	\$15,000
Jane	F	BS	3	\$120,000	\$12,000
Ryan	M	MS	4	\$150,000	\$15,000
Frank	M	PhD	1	\$210,000	\$21,000

2016 data

name	gender	education	experience	salary	bonus
Anton	M	PhD	3	\$230,000	\$25,150
Bob	M	PhD	4	\$250,000	\$27,250
Jim	M	MS	6	\$160,000	\$17,440
Allen	M	MS	2	\$130,000	\$13,790
Cathy	F	BS	3	\$110,000	\$11,000
Tom	M	MS	5	\$150,000	\$16,400
Jane	F	BS	4	\$120,000	\$12,000
Ryan	M	MS	5	\$150,000	\$16,400
Frank	M	PhD	2	\$210,000	\$23,050

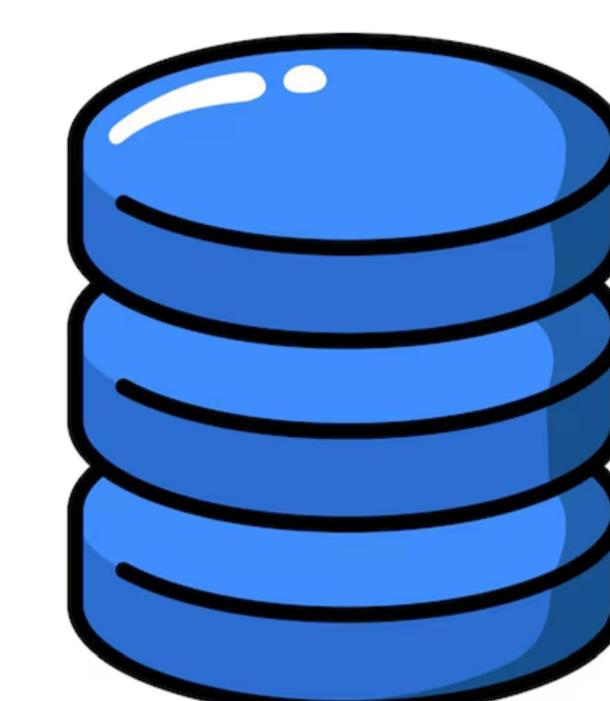
2017 data

What Accounts for the Bonus Discrepancy in 2017?  
I'm curious about why Anton received a \$2,150 increase, while Cathy did not receive any bonus at all!



## How to convey data change semantics to humans?

### > Data Diff

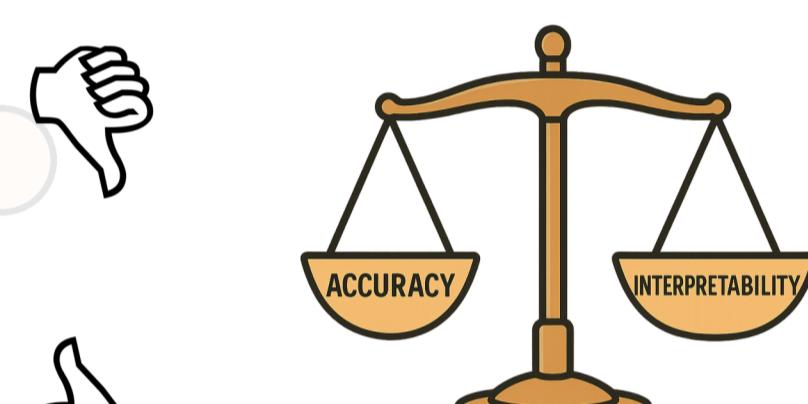


Too many changes have occurred between these two versions of the database.  
I am unable to process all of them.  
Could someone please provide a **summary** of the changes?



## What makes a change summary effective?

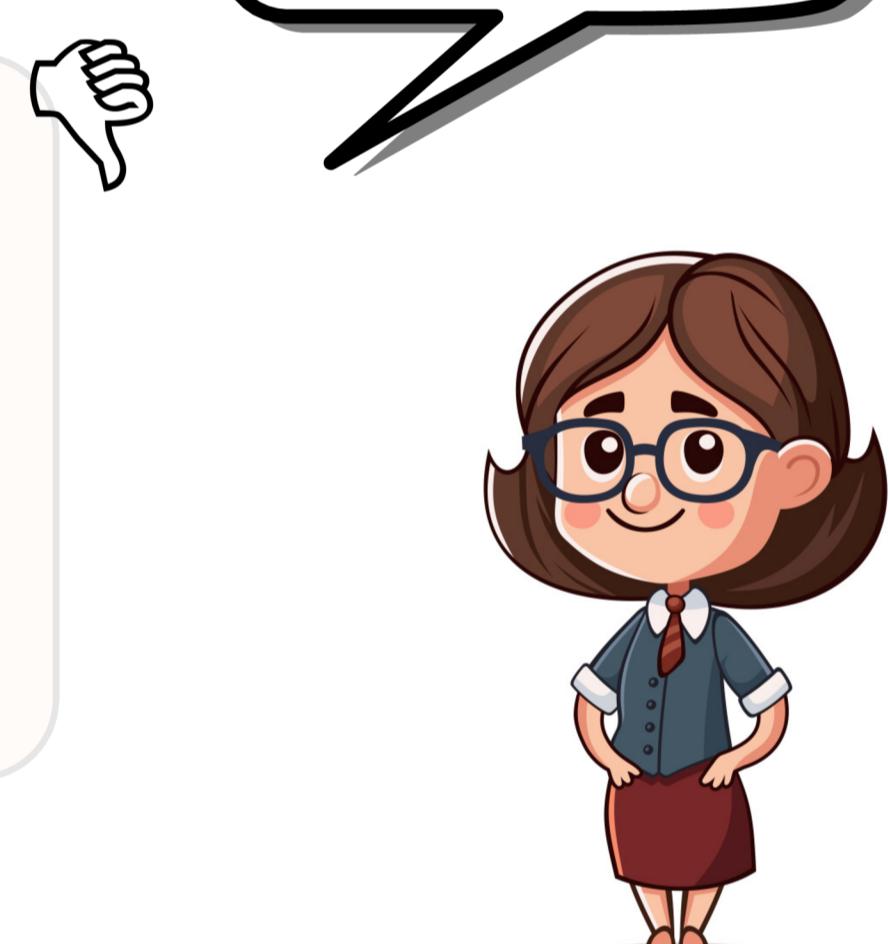
Everyone receives about 6% increase  
Less Accurate  
More Interpretable



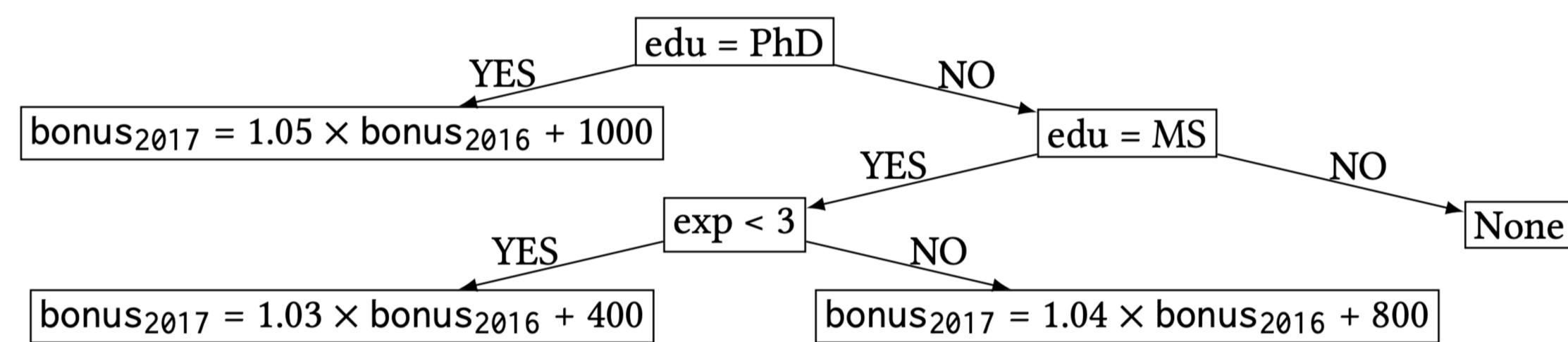
If PhD → 5% increase + \$1000  
If MS and served at least 3 years → 4% increase + \$800  
If MS and served less than 3 years → 3% increase + \$400  
Reasonably Accurate  
Reasonably Interpretable

A good trade-off between  
**Accuracy**  
&  
**Interpretability**

If PhD and Female → 5% increase + \$1000  
If MS and Female and served at least 3 years → 4% increase + \$800  
If MS and Male and served at least 3 years → 4% increase + \$700  
If BS and Female and served less than 3 years → 3% increase + \$400  
If BS and Male and served less than 3 years → 3% increase + \$300  
More Accurate  
Less Interpretable



## What is a good primitive for data change summary?



Linear Model Tree



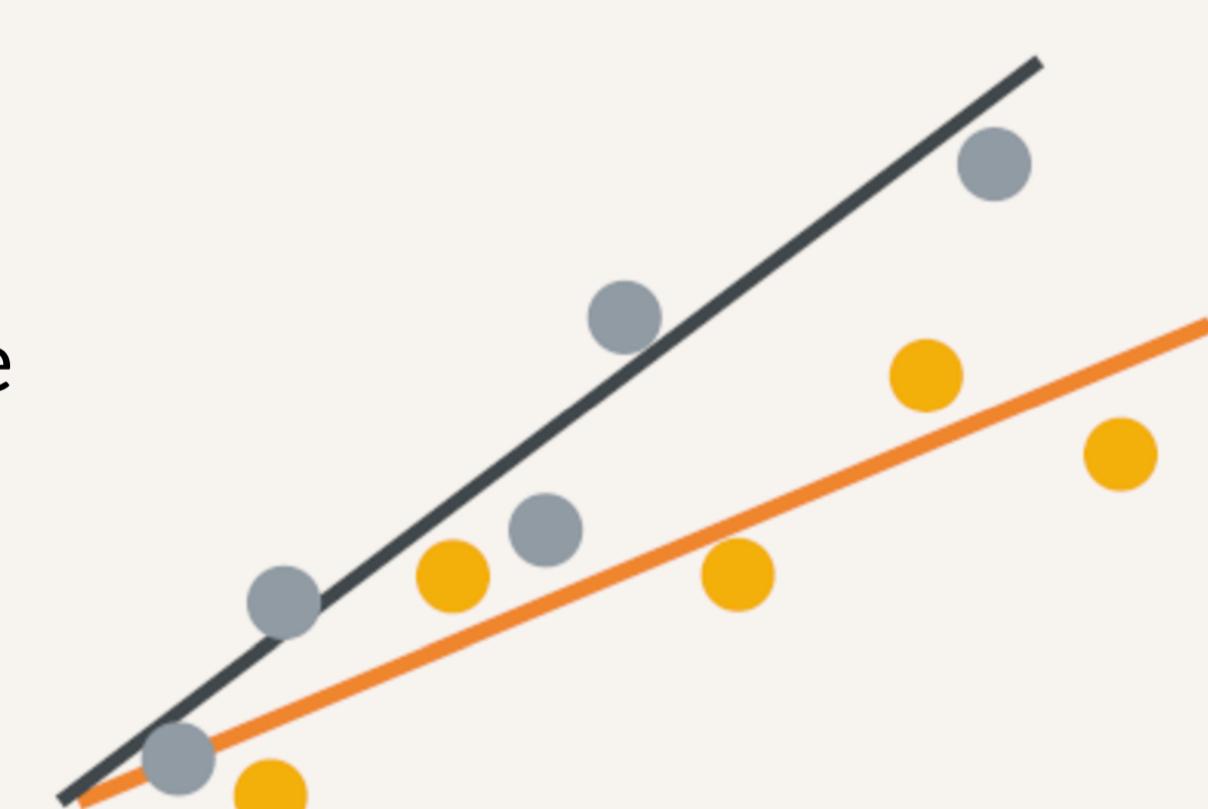
- Internal nodes denote **atomic conditions**
- Leaves denote **transformations**
- A path from root to a leaf denotes a **condition**, which is a conjunction of multiple atomic conditions
- Transformations** are linear functions over the data attributes
- LMT is a set of **conditional transformations**

## How do I model utility of a change summary?

### Accuracy

- Inverse of the distance between the actual change and the change described by the change summary

$$\text{Accuracy}(S) = |D_{\text{target}} - S(D_{\text{source}})|$$



### Interpretability

- Smaller summaries over larger ones
- Simpler conditions and transformations
- Transformations with fewer variables
- Conditions with higher data coverage
- "Normal" conditions and typical numerical values
- Domain expertise & LLMs can help!



All Female employees  
All Asian, European Females, or Females working in HR, but ...  
Age > 25  
Age > 23.796

## How to generate a summary with high utility?

Generate multiple LMTs, Rank them by utility



It appears that the company has chosen to implement a policy that rewards long-serving employees and supports educational advancement.

