Schema design and dependencies

Remy Wang, 4/22/25

name	location	breed	kind	job
casa	LA	tabby	cat	**NULL**
kira	seattle	tuxedo	cat	**NULL**
remy	LA	**NULL**	**NULL**	prof
vincent	LA	**NULL**	**NULL**	TA

name	location	breed	kind	job
casa	LA	tabby	cat	**NULL**
kira	seattle	tuxedo	cat	**NULL**
remy	LA	**NULL**	**NULL**	prof
vincent	LA	**NULL**	**NULL**	TA

1 table stores 1 kind of data

name	location	salary	graduate	funding
remy	LA	$30	**NULL**	$10
vincent	LA	$20	2025	**NULL**

name	location	salary	graduate	funding
remy	LA	$30	**NULL**	$10
vincent	LA	$20	2025	**NULL**

name	location	salary	funding
remy	LA	$30	$10

name	location	salary	graduate
vincent	LA	$20	2025

name	location	salary	funding
remy	LA	$30	$10

name	location	salary	graduate
vincent	LA	$20	2025

payroll

name	location	salary
remy	LA	$30
vincent	LA	$20

prof

name	funding
remy	$10

student

name	graduate
vincent	2025

each kind of data has its own table

name	location	salary	course
remy	LA	$30	143
remy	LA	$30	240
remy	LA	$30	249

name	location	salary	course
remy	LA	$30	143
remy	LA	$30	240
remy	LA	$30	249

redundancy!

(complicates updates & deletes)

each piece of information stored once

determines

name	location	salary	course
remy	LA	$30	143
remy	LA	$30	240
remy	LA	$30	249

$$\text{name} \rightarrow \text{location, salary}$$

functional dependency

$$X \to Y$$

the values of X *uniquely determines* Y

$$\forall t, t' \in R : \pi_X(t) = \pi_X(t') \implies \pi_Y(t) = \pi_Y(t')$$

name	location	salary	course
remy	LA	$30	143
remy	LA	$30	240
remy	LA	$30	249
dan	seattle	$50	344
dan	seattle	$50	444

$$\forall t, t' \in R : \pi_{\text{name}}(t) = \pi_{\text{name}}(t') \implies \pi_{\text{salary}}(t) = \pi_{\text{salary}}(t')$$

$$\text{name} \rightarrow \text{salary}$$

determines

first n.	last n.	location	salary	course
remy	w	LA	$30	143
remy	w	LA	$30	240
remy	w	LA	$30	249
dan	s	seattle	$50	344
dan	s	seattle	$50	444
dan	o	zurich	$50	101
dan	o	zurich	$50	113

determines

first n.	last n.	location	salary	course
remy	w	LA	$30	143
remy	w	LA	$30	240
remy	w	LA	$30	249
dan	s	seattle	$50	344
dan	s	seattle	$50	444
dan	o	zurich	$50	101
dan	o	zurich	$50	113

$$\{\text{first n., last n.}\} \rightarrow \{\text{location, salary}\}$$

Check X → Y using SQL?

Check X → Y using SQL?

```
SELECT * FROM R
 GROUP BY X
HAVING COUNT(Y) > 1
```

Trivial FDs?

A	B	C	D	E

Trivial FDs?

A	B	C	D	E

$$A \to A, B \to B, \ldots$$

$$AB \to A, AB \to B, \ldots$$

Trivial FDs?

A	B	C	D	E

$$A \to A, B \to B, \ldots$$

$$AB \to A, AB \to B, \ldots$$

$$X \supseteq Y \implies X \to Y$$

name	job	location	salary	tax %
remy	prof	LA	$30	20
dan	prof	seattle	$50	15
vincent	TA	LA	$20	10

$$\text{job} \rightarrow \text{salary}$$

$$\text{name} \rightarrow \text{location}$$

$$\text{location, salary} \rightarrow \text{tax \%}$$

name	job	location	salary	tax %
remy	prof	LA	$30	20
dan	prof	seattle	$50	15
vincent	TA	LA	$20	10

$$\text{job} \to \text{salary}$$

$$\text{name} \to \text{location} \quad \Longrightarrow \quad \text{name, job} \to \quad ?$$

$$\text{location, salary} \to \text{tax \%}$$

name	job	location	salary	tax %
remy	prof	LA	$30	20
dan	prof	seattle	$50	15
vincent	TA	LA	$20	10

$$\text{job} \to \text{salary}$$

$$\text{name} \to \text{location} \quad \Longrightarrow \quad \text{name, job} \to \text{tax \%}$$

$$\text{location, salary} \to \text{tax \%}$$

name	job	location	salary	tax %
N	J	L	S	T
N	J	?	?	?
vincent	TA	LA	$20	10

$$\text{job} \rightarrow \text{salary}$$

$$\text{name} \rightarrow \text{location} \quad \Longrightarrow \quad \text{name, job} \rightarrow \text{tax \%}$$

$$\text{location, salary} \rightarrow \text{tax \%}$$

FD inference

Given

$$X_1 \to Y_1$$
$$X_2 \to Y_2$$
$$X_3 \to Y_3$$
$$\ldots$$

Does this hold?

$$X \to Y$$

Armstrong's axioms

$$Y \subseteq X \implies X \to Y$$

$$X \to Y \implies XZ \to YZ$$

$$X \to Y \land Y \to Z \implies X \to Z$$

Armstrong's axioms

X	Y	Z
x	y	z
x	?	z

$$Y \subseteq X \implies X \to Y$$ Reflexivity

$$X \to Y \implies XZ \to YZ$$ Augmentation

$$X \to Y \land Y \to Z \implies X \to Z$$ Transitivity

$$\boxed{\text{job} \rightarrow \text{salary}}$$

$$\text{name} \rightarrow \text{location} \qquad \Longrightarrow \qquad \text{name, job} \rightarrow \text{tax \%}$$

$$\text{location, salary} \rightarrow \text{tax \%}$$

$$\boxed{\text{job, name} \rightarrow \text{salary, name}}$$

$$\text{name, salary} \rightarrow \text{location, salary}$$

$$\text{name, salary} \rightarrow \text{tax \%}$$

FD closure

given

$$X_1 \to Y_1$$

$$X_2 \to Y_2$$

$$X_3 \to Y_3$$

$$\dots$$

compute X^+ :

start w/ X

repeat until no change:

add Y_i to X if $X_i \subseteq X$

$$\text{job} \to \text{salary}$$

$$\text{name} \to \text{location} \quad \Longrightarrow \quad \text{name, job} \to \text{tax \%}$$

$$\text{location, salary} \to \text{tax \%}$$

$\{\text{name, job}\}^+ = \{\text{name, job}\}$

Why FDs?

FDs cause anomalies

first n.	last n.	location	salary	course
remy	w	LA	$30	143
remy	w	LA	$30	240
remy	w	LA	$30	249

$X \subseteq \{A_1, \ldots, A_5\}$ is a **superkey** if $\forall i : X \to A_i$

i.e. $X^+ = \{A_1, \ldots, A_5\}$

A1	A2	A3	A4	A5

PK

name	location	salary
remy	LA	$30
vincent	LA	$20

first n.	**last n.**	**location**	**salary**
remy	w	LA	$30
dan	s	seattle	$50
dan	o	zurich	$50

first n.	last n.	location	salary	course
remy	w	LA	$30	143
remy	w	LA	$30	240
remy	w	LA	$30	249
dan	s	seattle	$50	344
dan	s	seattle	$50	444
dan	o	zurich	$50	101
dan	o	zurich	$50	113

name	job	location	salary	tax %
remy	prof	LA	$30	20
dan	prof	seattle	$50	15
vincent	TA	LA	$20	10

$$\text{job} \rightarrow \text{salary}$$

$$\text{name} \rightarrow \text{location} \quad \Longrightarrow \quad \text{name, job} \rightarrow \text{tax \%}$$

$$\text{location, salary} \rightarrow \text{tax \%}$$

$X \subseteq \{A_1, \ldots, A_5\}$ is a **superkey** if $\forall i : X \to A_i$

A **key** is a minimal superkey

(no longer a superkey if removing anything) revisit examples

to find a key: guess X from small to large, check if $X^+ = \{A_1, ..., A_5\}$

$job^+ = \{?\}$ $salary^+ = \{?\}$ $name^+ = \{?\}$

$location^+ = \{?\}$ $tax^+ = \{?\}$

$\{name, job\}^+ = \{?\}$

$$\text{job} \rightarrow \text{salary}$$

$$\text{name} \rightarrow \text{location} \quad \Longrightarrow \quad \text{name, job} \rightarrow \text{tax \%}$$

$$\text{location, salary} \rightarrow \text{tax \%}$$

to find a key: guess X from small to large, check if $X^+ = \{A_1, ..., A_5\}$

any key must contain **name, job**

$$\text{job} \to \text{salary}$$

$$\text{name} \to \text{location} \qquad \Longrightarrow \qquad \text{name, job} \to \text{tax \%}$$

$$\text{location, salary} \to \text{tax \%}$$

more exercises

$$R(A, B, C)$$

$$A \to BC$$

$$B \to AC$$

$$AB \to C$$

Decompose relations, finally!

Boyce-Codd Normal Form:

$$\forall X \to Y \models R : Y \subseteq X \vee X \text{ is a superkey}$$

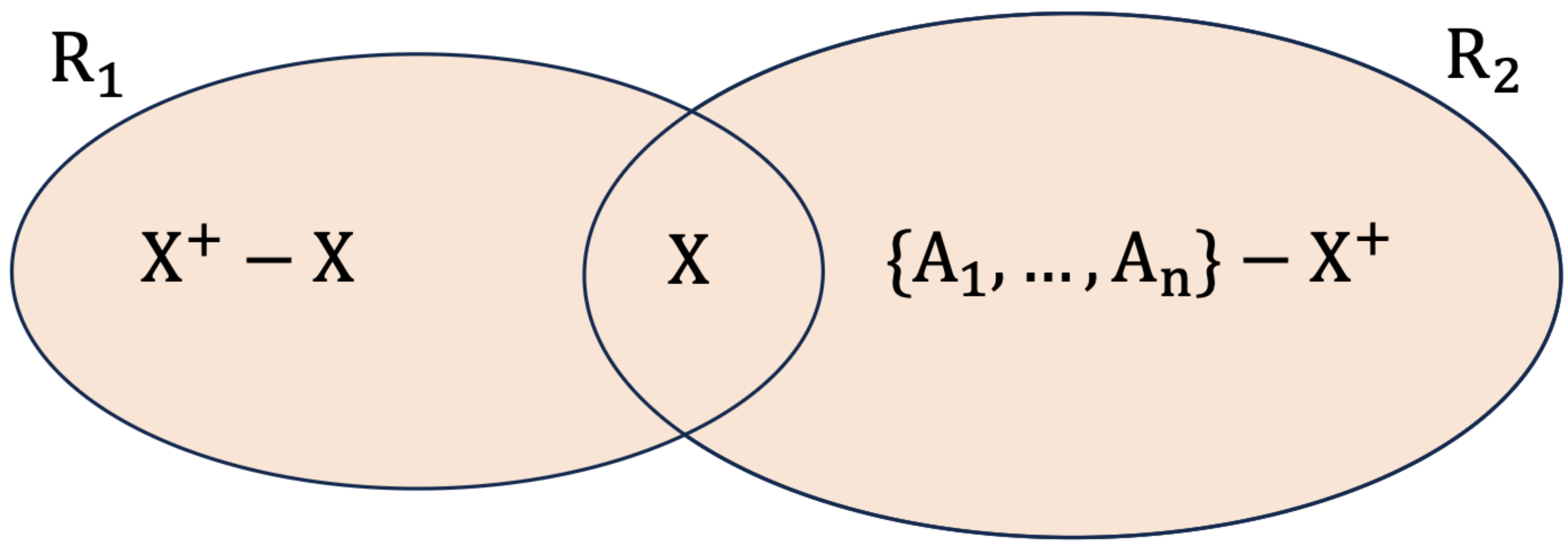
i.e. $\forall X : X^+ = X \vee X^+ = \{A_1, \ldots\}$

Decomposition

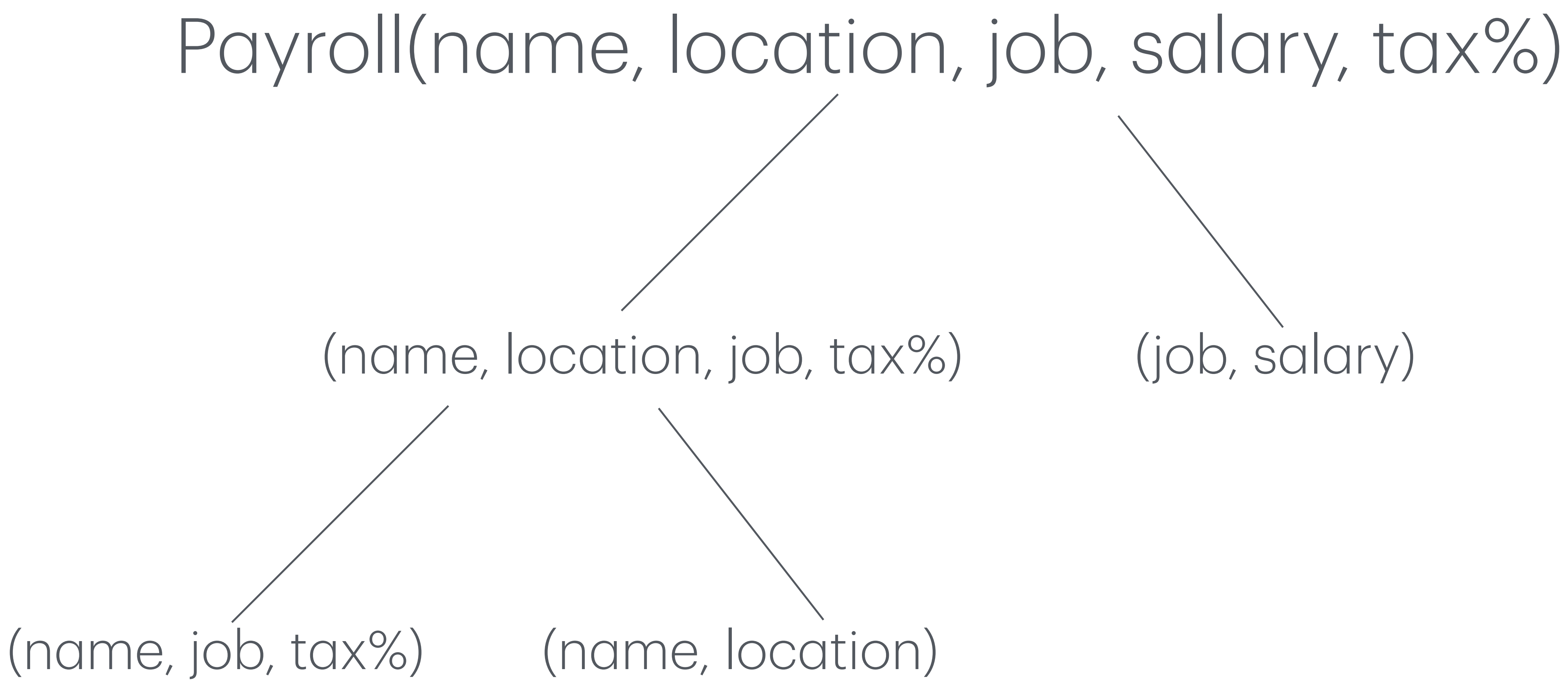
repeat

find X → Y violating BCNF

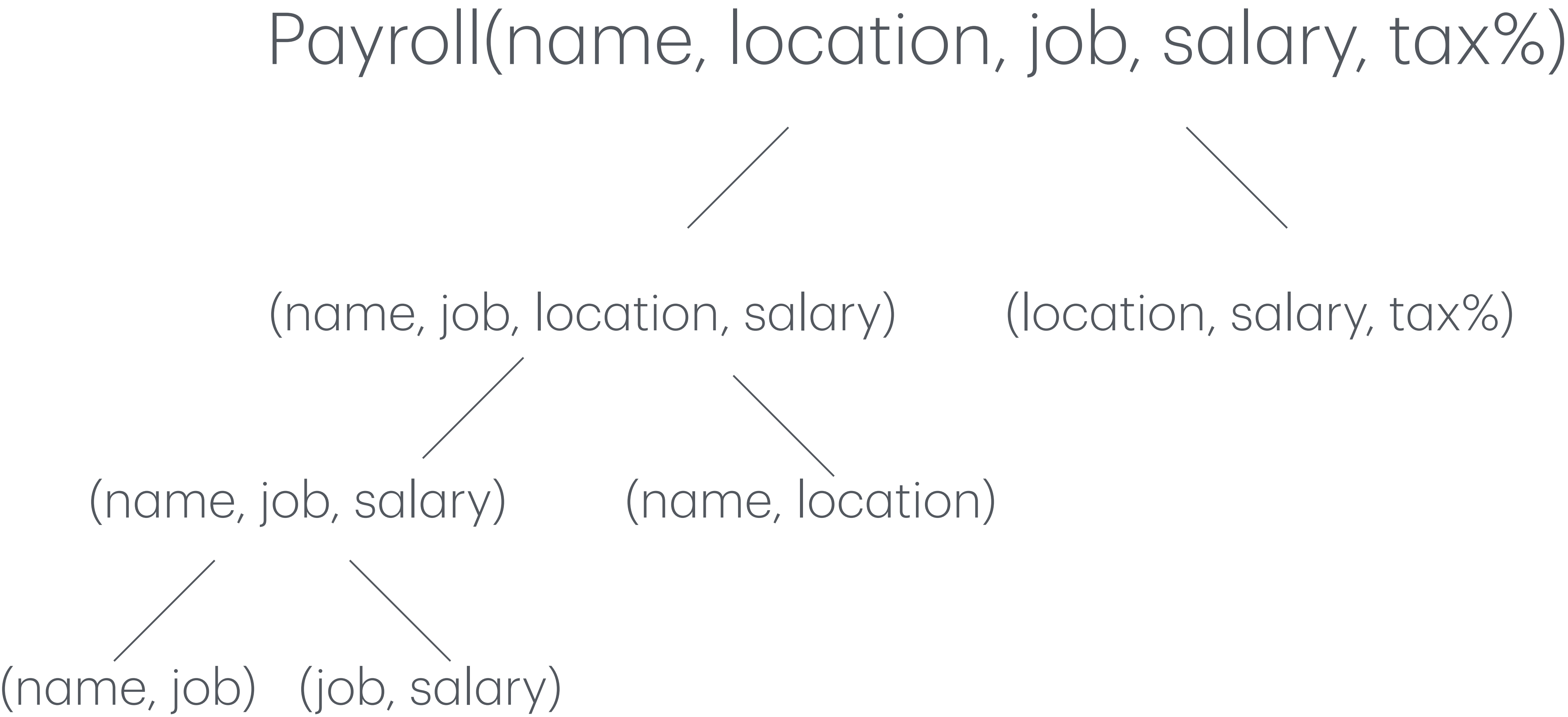
"factor out" X^+ from R



$$\text{job} \rightarrow \text{salary} \quad \text{name} \rightarrow \text{location} \quad \text{location, salary} \rightarrow \text{tax \%}$$



$$\text{job} \rightarrow \text{salary} \quad \text{name} \rightarrow \text{location} \quad \text{location, salary} \rightarrow \text{tax \%}$$



Why care about preserving FDs?

Find tax rate for ("LA", $50)

(name, job) (job, salary) (name, location) (location, salary, tax%)

(name, job, tax%) (name, location) (job, salary)

Other considerations

security & privacy

compliance

geolocation

performance