Semi-automatic database normalisation in R

Mark Webster
Paymentshield

2025-10-16

Where I’ve been

Currently a payment analyst here (home insurance):

Previously a data scientist here (healthcare consulting and strategic intelligence):

Some of the following was done during hours at VISFO, and they kindly let me keep the R package project as my own.

Database normalisation / tidy data

> In tidy data:
>
> 1. Each variable forms a column.
> 2. Each observation forms a row.
> 3. **Each type of observational unit forms a table.**
>
> This is Codd's **3rd normal form**, but with the constraints framed in statistical language…"
>
> ~ Wickham, *Tidy data* (2014)

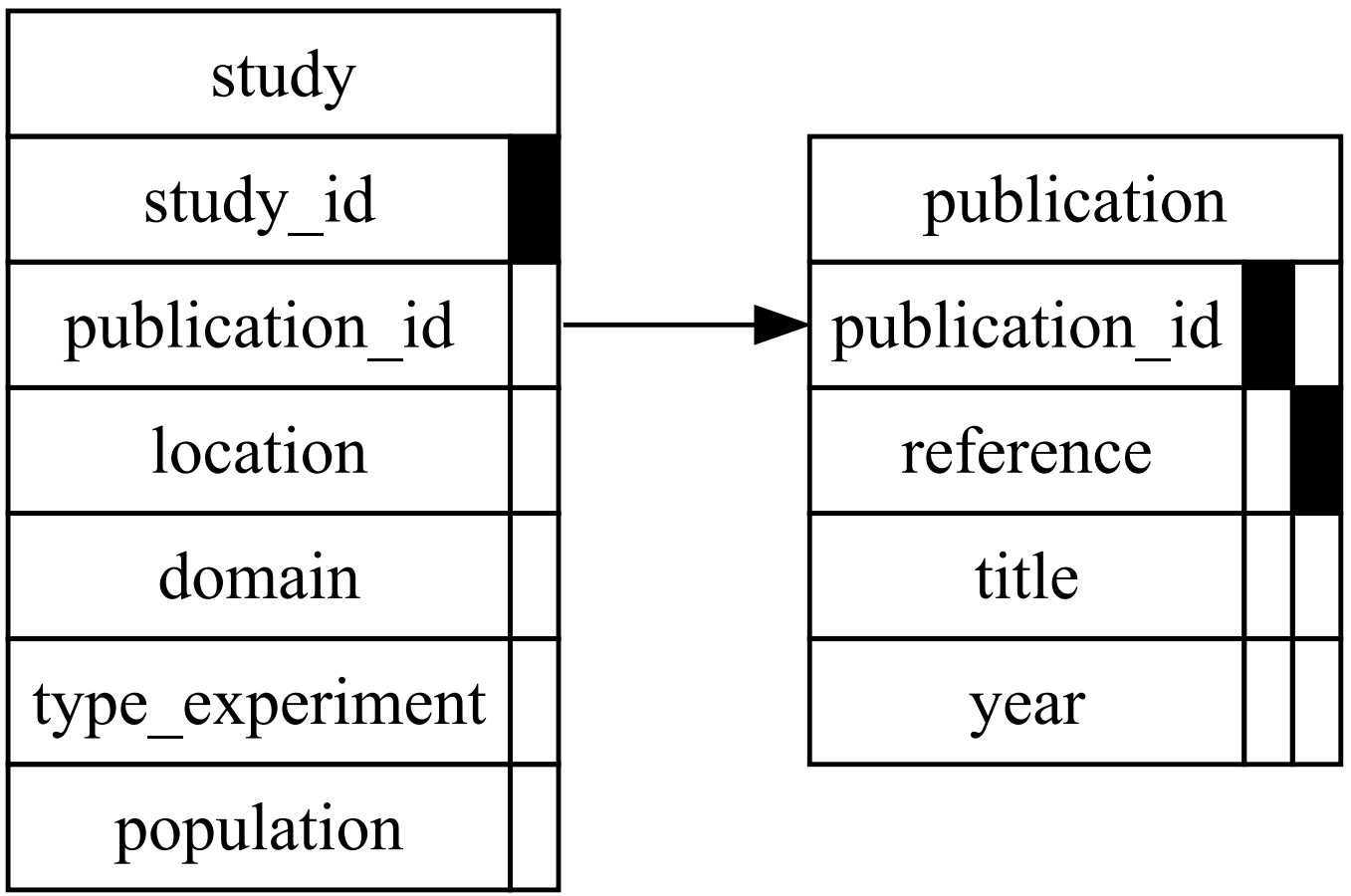
Running example (334×9)

Nudge intervention meta-analysis data from Mertens et al. 2022 (full size: 447×25)

publication_id	study_id	reference	title	year
52	84	Dilibert...	Increase...	2004

location	domain	type_experiment	population	
inside US	food	natural_field	adults	

Expected schema



Black cells represent unique row identifiers (keys).

What if we don’t have a schema?

Some reasons:

- Little domain knowledge
- No (reliable) documentation
- Checking understood data for additional structure

What can we automate away in R?

{autodb}

{autodb} turns a table into a database, using functional dependency discovery:

```
1  db_nudge <- autodb(nudge_simple)
```

Quick enough for interactive use for small data sets:

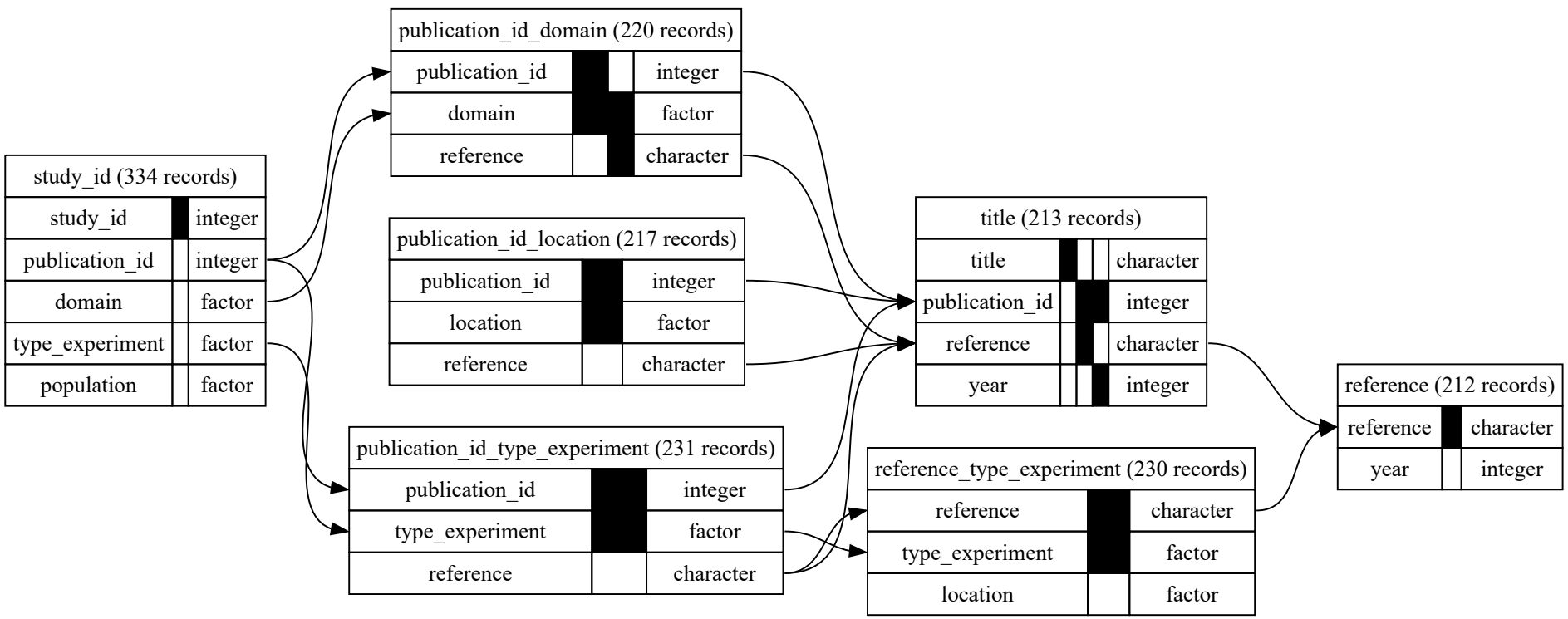
```
Unit: seconds
                 expr       lq      mean        uq neval
 autodb(nudge_simple) 0.052998 0.0567451 0.0590848     5
        autodb(nudge) 8.638586 8.6593708 8.7492257     5
```

Plotting with GraphViz

```
1 nudge_code <- gv(db_nudge)
2 DiagrammeR::grViz(nudge_code)
```

Plotting with GraphViz

```
1 nudge_code <- gv(db_nudge)
2 DiagrammeR::grViz(nudge_code)
```



Does this make sense?

```
1 db_nudge$title
```

title (213 records)				
title	■			character
publication_id		■	■	integer
reference		■		character
year			■	integer

Finding duplicates

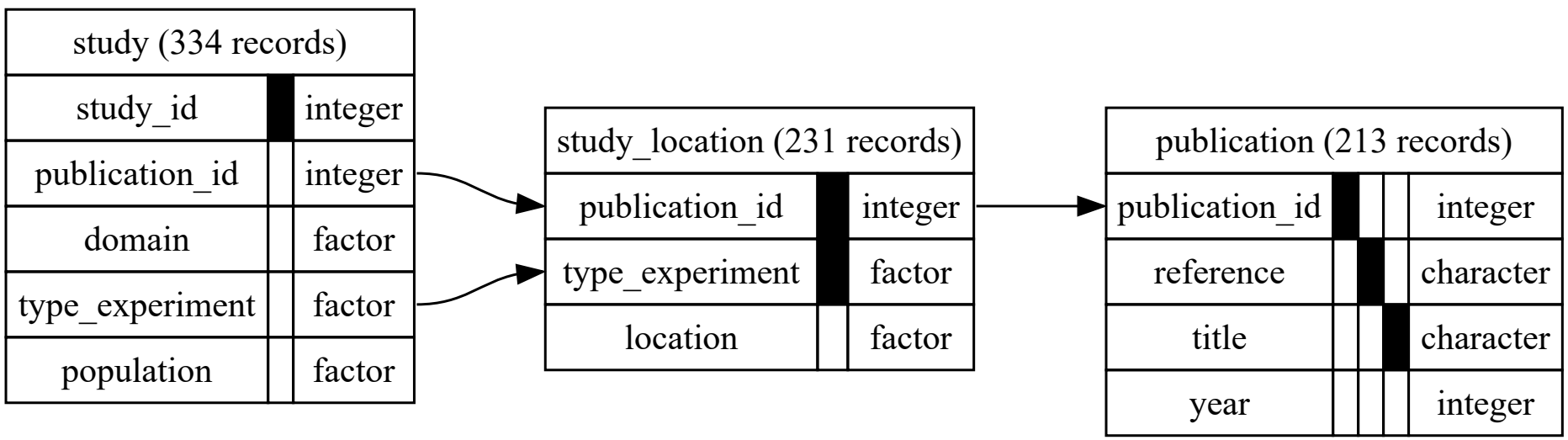
title	publication_id	reference	year
Enhanced active c...	95	Keller et al. (2011)	2011
Nudging product c...	95	Keller et al. (2015)	2015
Nudge vs superbug...	18	BETA (2018)	2018
Energy labels tha...	19	BETA (2018)	2018

Fixing duplicates

title	publication_id	reference	year
Enhanced active c...	95	Keller et al. (2011)	2011
Nudging product c...	213	Keller et al. (2015)	2015
Nudge vs superbug...	18	BETA (2018)	2018
Energy labels tha...	19	BETA (2018a)	2018

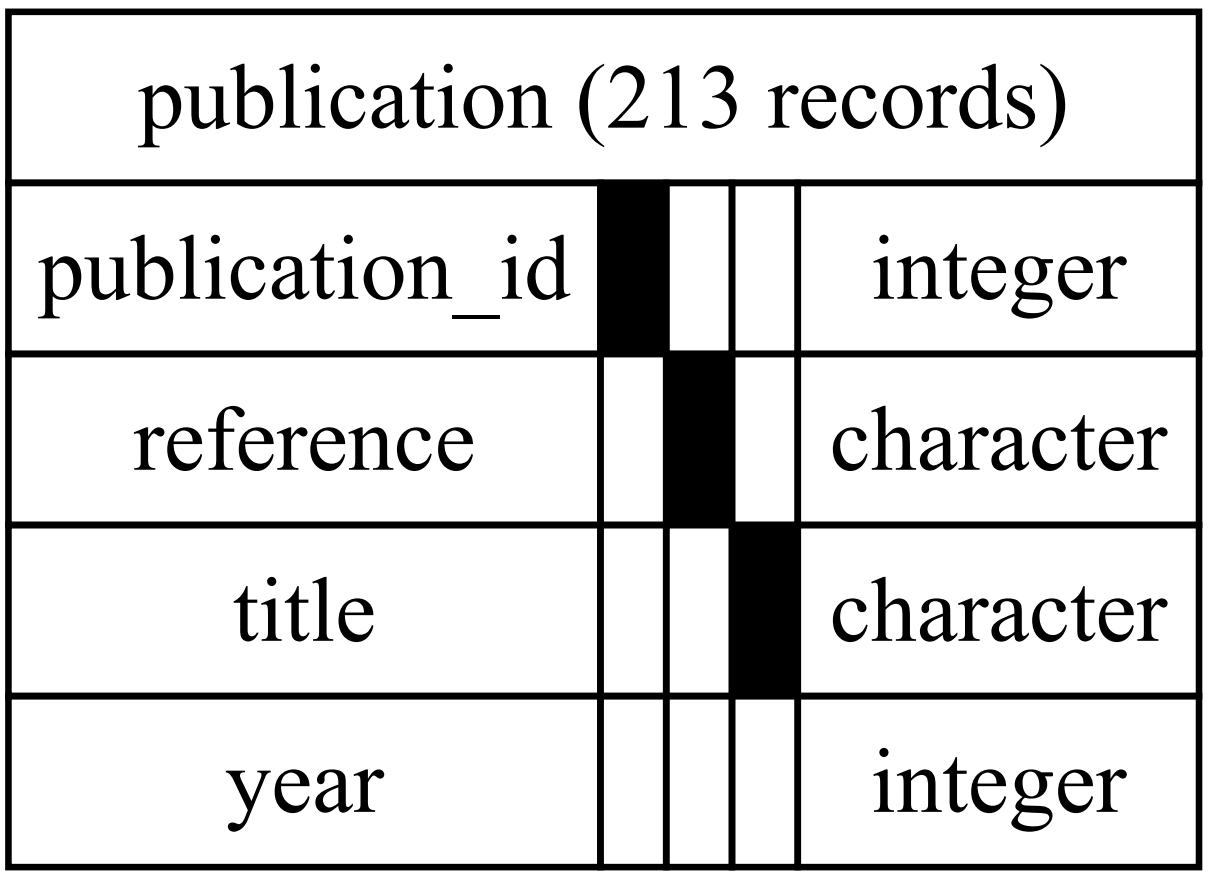
Fixed schema

```
1 db_fix <- autodb(nudge_fix) # and re-assigning table names()
```



Publications are fixed

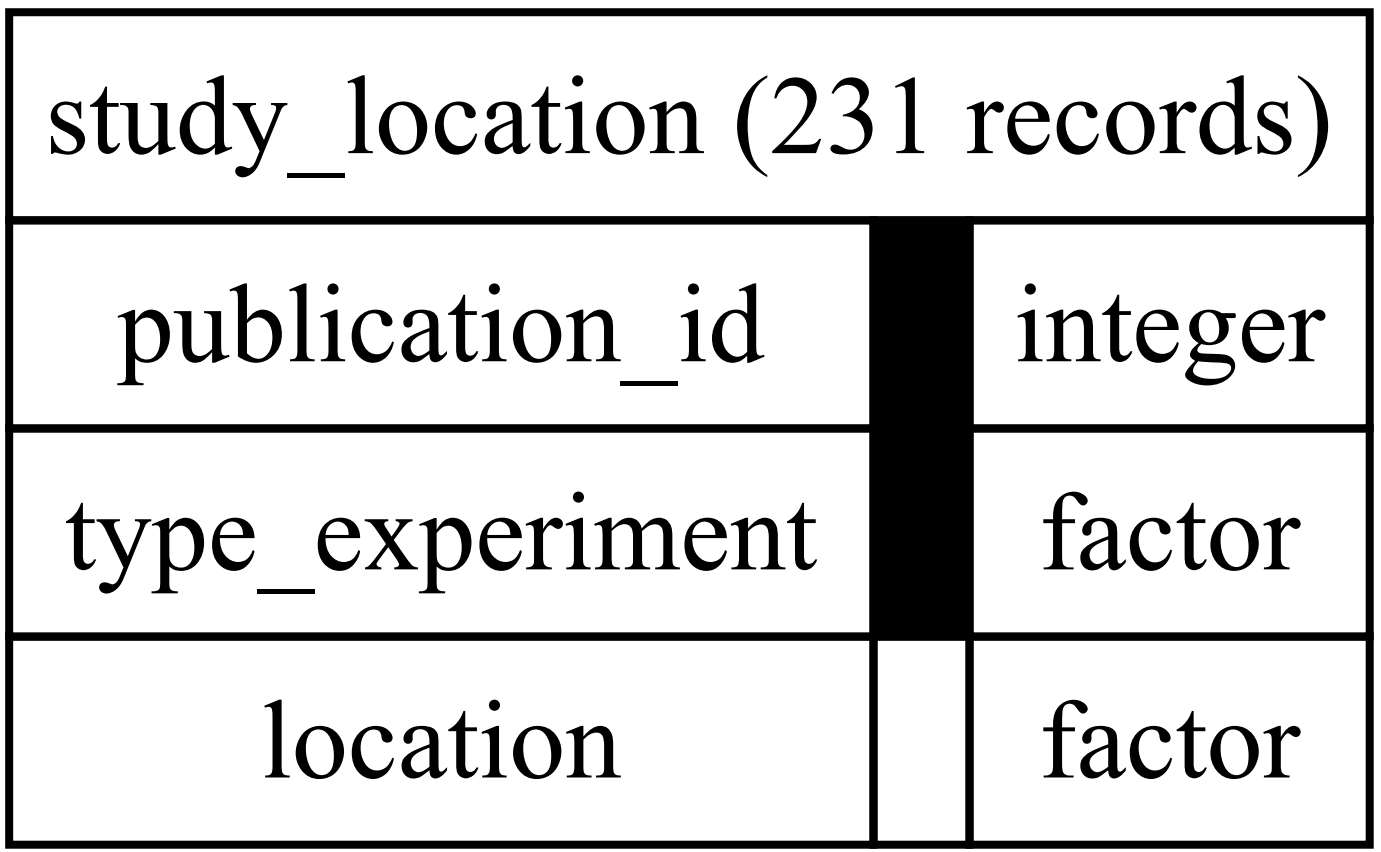
```
1 db_fix$publication
```



publication (213 records)				
publication_id	■			integer
reference		■		character
title			■	character
year				integer

Data-specific structure

```
1 db_fix$study_location
```



The end

{autodb} is on CRAN, active project

Can also create schemas manually for validation / diagrams

References for general approach:

- Bernstein P. A. (1976) Synthesizing third normal form relations from functional dependencies
- Bleifuss et al. (2024) Discovering functional dependencies through hitting set enumeration
- Other data profiling research by the Information Systems Group at Hasso Plattner Institut metanome.de