



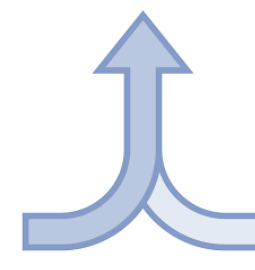


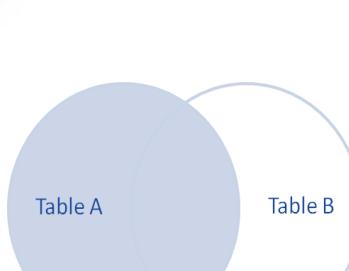



# USING SQL FOR DATA CONSOLIDATION IN R

Jennifer Vinas-Forcade, Julien Nacci,  
Dr. Cindy Mels, Prof. dr. Martin Valcke,  
Prof. dr. Ilse Derluyn.

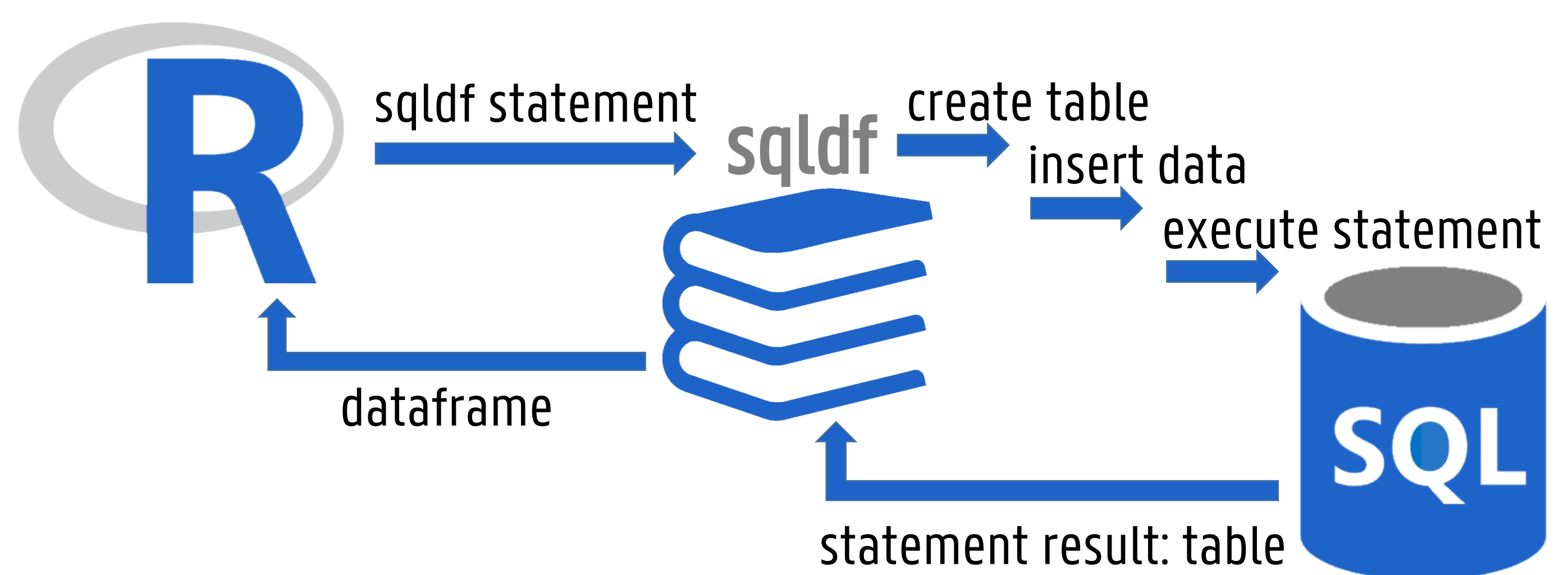
## PROBLEM & RELEVANCE

Working with multiple data sources implies data cleaning and consolidation prior to analysis. R has become popular among social scientists (Kelley, 2007; Clark, 2014), who are advised to screen data in a “favorite spreadsheet program” (Muenchen, 2011:21), before importing it to R. This way, users avoid typing in the R console and are supported by a graphical user interface. Even for experienced R users, querying/retrieving data from multiple large sources takes a lot of computing power, which is better handled by SQL language (KeyCentrix, 2015).

## SQL FUNCTIONS

	Data cleaning: identify unique values	→	Select <b>distinct</b> ... from ...
	Data cleaning: delete missing values	→	Select... from ... where ... <b>is not null</b>
	Merging data (union / add rows)	→	Select ... <b>union</b> select ... <b>union</b> select ...
	Merge dataframes with a different number of columns	→	Select df1.v1, df1.v2, df1.v3 from df1 <b>union</b> df2.v1, <b>df2.null</b> , df2.v3 from df2
	Consolidate n dataframes using unique id, discard all non-matches	→	Select df1.v1, df2.v1 from <b>df1, df2</b> where <b>df1.id = df2.id</b>
	Consolidate n dataframes keeping all baseline records	→	Select df1.*, df2.* from df1 <b>left join</b> df2 on df1.id = df2.id
	Basic data aggregation operations	→	Select ... <b>count</b> (...), <b>avg</b> (...) <b>group by</b> ...
	Data integrity (check-ups)	→	Select ... where v1 <b>[not] in</b> (select ...)
	Reorder columns of a data frame	→	Select <b>v3, v4, v2, v1</b> from df

## HOW IT WORKS



## COMPARISON

	SQL	R
<b>Function</b>	Data optimizing, updating, querying	Statistical data analysis
<b>Math &amp; stats</b>	Only basic operations	Specific functions for complex operations.
<b>Syntax</b>	More anthropomorphic	Less intelligible
<b>Memory</b>	Retrieves the specific data needed for each query, when prompted.	Loads all data on RAM memory.

## REFERENCES

- <sup>1</sup> Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. Behavior Research Methods, 39(4), 979-984.
- <sup>2</sup> Clark, M. (2014). Getting started with applied use of R in the social sciences. Retrieved from <http://m-clark.github.io/docs/RSocialScience.pdf>, accessed 09/04/2018.
- <sup>3</sup> Muenchen, R. A. (2011). R for SAS and SPSS users. Springer Science & Business Media.
- <sup>4</sup> KeyCentrix (2015, December 11), “Why SQL is powerful” in KeyCentrix. Retrieved from <https://keycentrix.com/blog/why-sql-is-powerful/>, accessed 09/04/2018.
- <sup>5</sup> Grothendieck, G. (2017). ‘Package ‘sqldf’ in The Comprehensive R Archive Network. Retrieved from <https://cran.r-project.org/web/packages/sqldf/sqldf.pdf>, accessed 09/04/2018.

## CONTACT

Jennifer.vinasforcade@ugent.be  
www.ugent.be

 Universiteit Gent  
 @ugent  
 Ghent University