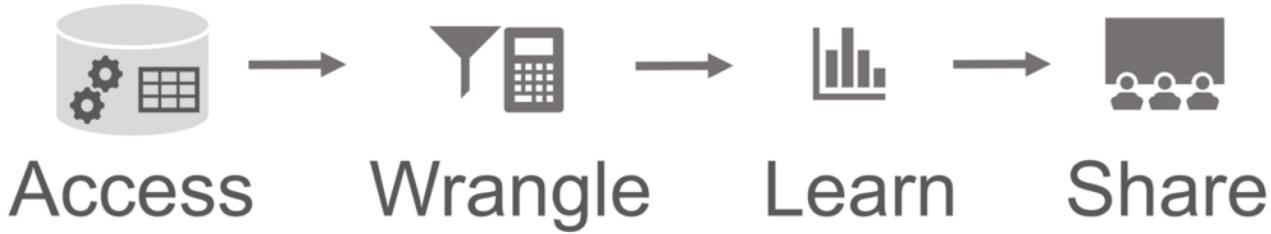


# A Data Science Workbench with RStudio and Teradata

Nathan Stephens  
*Director of Solutions Engineering, RStudio*

# Using Databases with R

- You can use R with almost any data source



- But historically it's not always easy

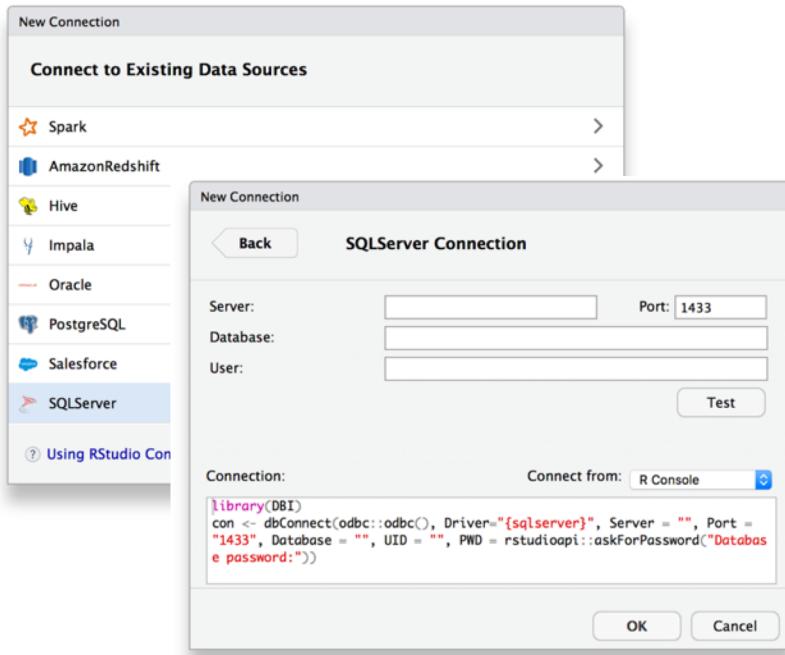
# Obstacles when using databases with R

- 1.** Hard to set up connections
- 2.** No consistent toolset or language
- 3.** No centralized place to get information
- 4.** Hard to communicate insights

# What's New

# RStudio version 1.1 features

Connection wizard



Connections tab

The screenshot shows the 'Connections' tab in the RStudio interface. The top navigation bar has tabs for 'Environment', 'History', 'Connections', and 'Git'. The 'Connections' tab is active. It displays a list of connections:

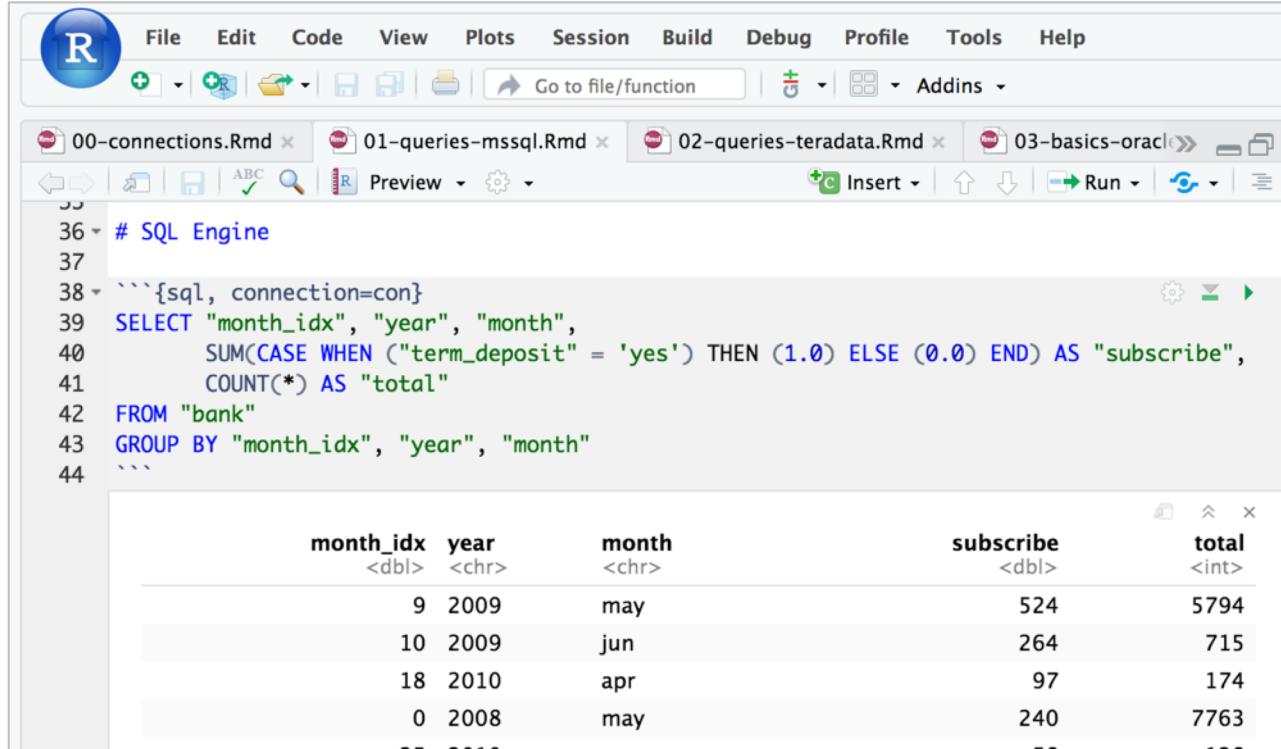
- airontime - rstudioadmin@EC2AMAZ-EIJ1Q05 (Connected)
- rstudioadmin@airontim
- RSTUDIO - RSTUDIO@153.64.73.12
- finan
- airon
- postg

Below this, a detailed view of the 'airontime' connection is shown, listing databases:

- airontime
- finance
- dbo
- INFORMATION\_SCHEMA
- sys
- test
- master
- msdb
- rdsadmin
- tempdb

The right side of the interface features a vertical sidebar with icons for different database types and a tree view of the database structure.

# SQL Language Engine for R Markdown



The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Includes icons for file operations (New, Open, Save, Print), Go to file/function, and Addins.
- Tab Bar:** Shows four open files: 00-connections.Rmd, 01-queries-mssql.Rmd, 02-queries-teradata.Rmd, and 03-basics-oracle.Rmd. The 02-queries-teradata.Rmd tab is active.
- Code Editor:** Displays R code for generating a summary statistics data frame from a bank dataset. The code uses the `sql` engine to execute SQL queries within R. The code is as follows:

```
36 # SQL Engine
37
38 ````{sql, connection=con}
39   SELECT "month_idx", "year", "month",
40         SUM(CASE WHEN ("term_deposit" = 'yes') THEN (1.0) ELSE (0.0) END) AS "subscribe",
41         COUNT(*) AS "total"
42   FROM "bank"
43   GROUP BY "month_idx", "year", "month"
44 ````
```

- Data View:** A data frame is displayed with the following columns and rows:

month_idx	year	month	subscribe	total
<dbl>	<chr>	<chr>	<dbl>	<int>
9	2009	may	524	5794
10	2009	jun	264	715
18	2010	apr	97	174
0	2008	may	240	7763
25	2010	dec	52	100

# Improved R Packages

## **DBI**

- Standard database interface for R
- Query your database with SQL
- Do database operations

## **dplyr**

- Query your database with R code
- Generalized SQL backend
- dbplyr translates dplyr syntax to SQL for specific databases

# Database connection methods

## **odbc**

- Connects R to any data source via ODBC
- Bring your own driver or use RStudio Professional Drivers
- DBI compliant
- Actively developed
- Designed for performance

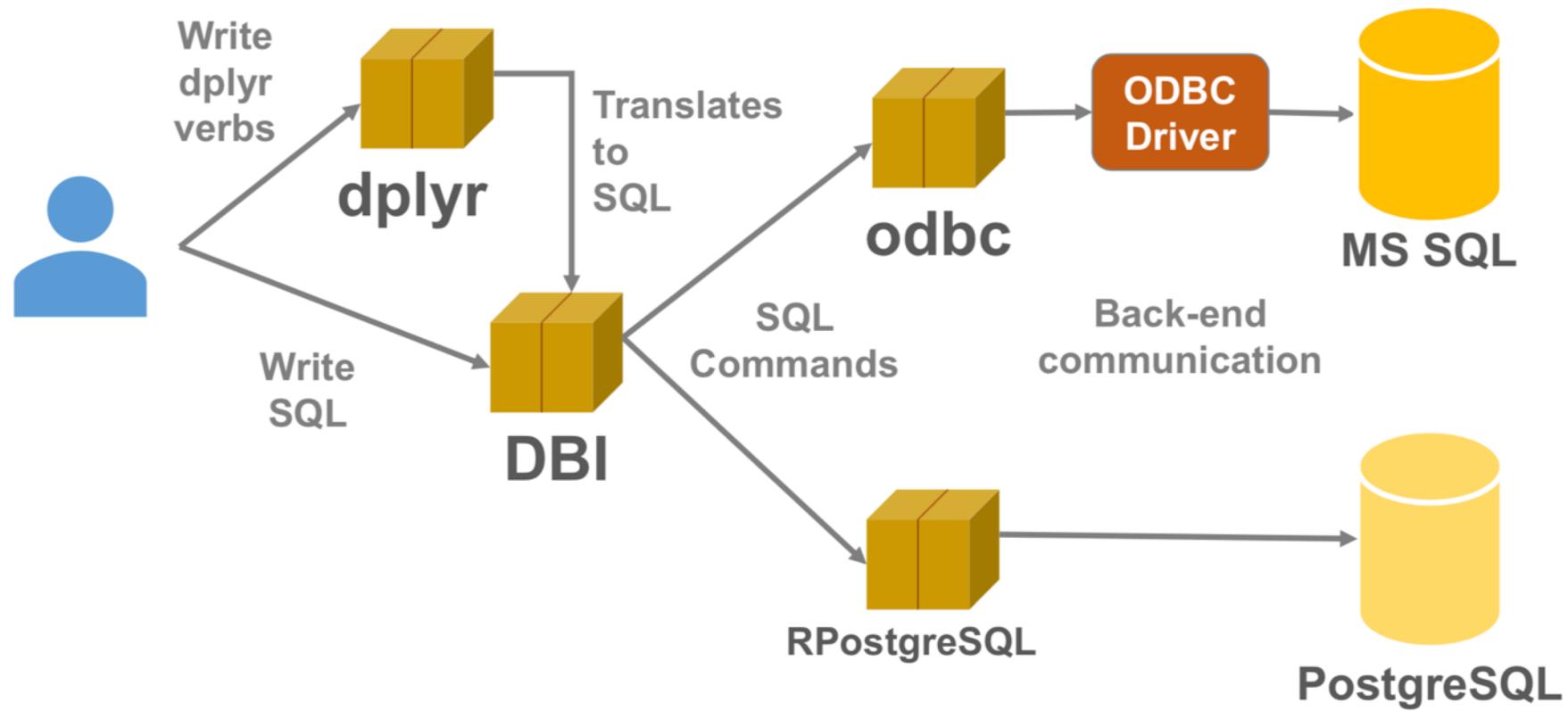
## **R Packages**

- RSQLite
- rpostgres
- bigrquery
- tdplyr

## **RJDBC**

- Connect to any data source via JDBC
- Requires JDBC driver
- Requires Java

# Establish a database connection with RStudio



# RStudio Professional Drivers

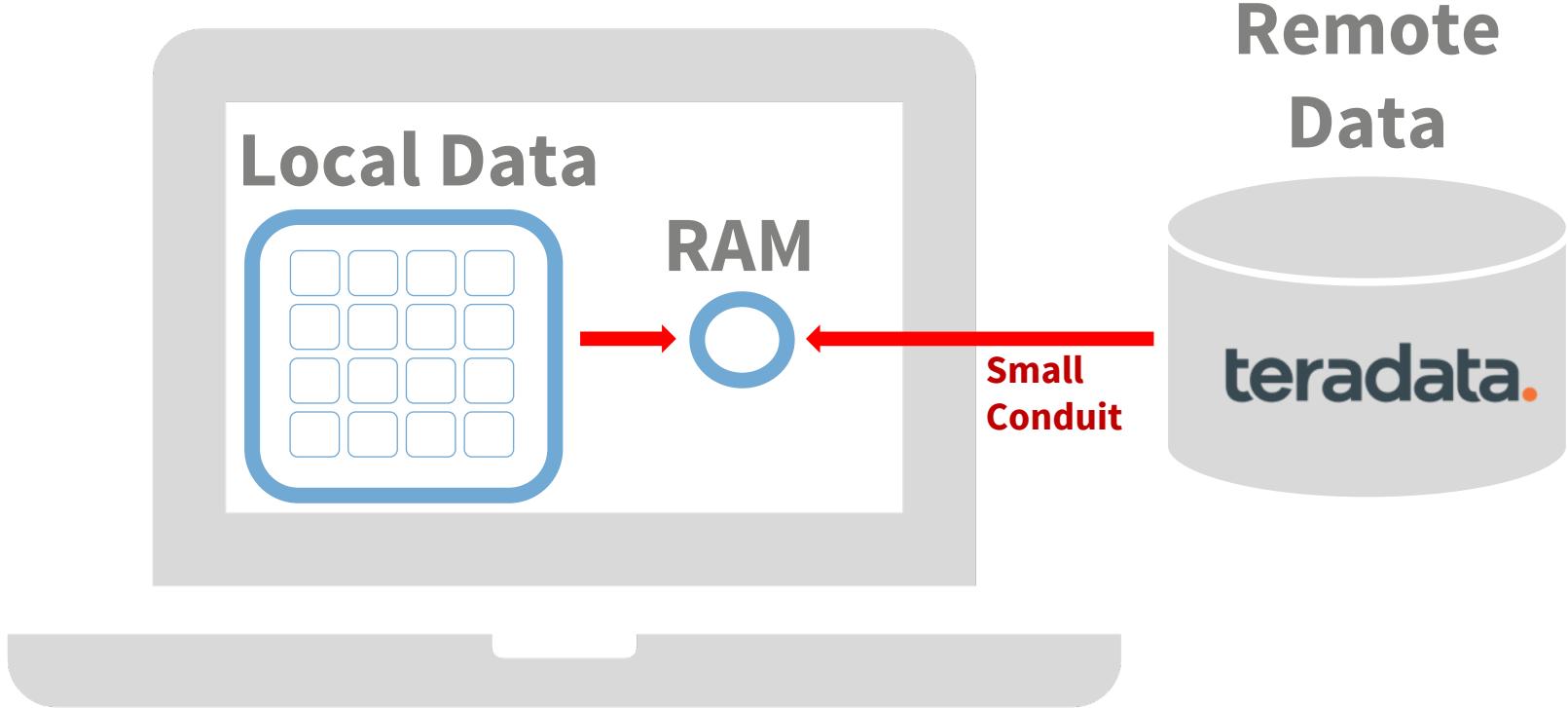
1. Connect to many popular databases
2. Easy to install and configure
3. Use with all RStudio pro products
4. Supported for production level work



Teradata	Oracle
PostgreSQL	Impala
Salesforce	Hive
Redshift	SQL Server

# Using R with Databases

# Big Data with R

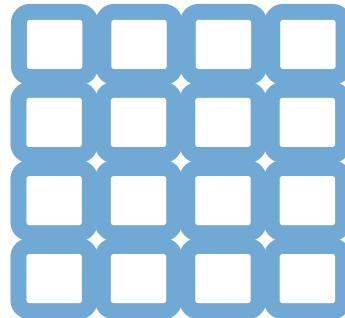


# Big Data Strategies

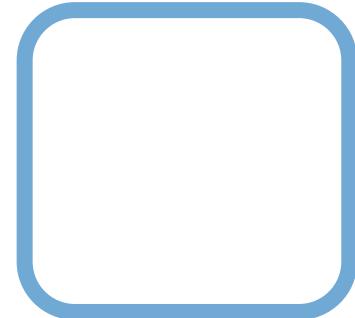
**Sample**



**Parts**

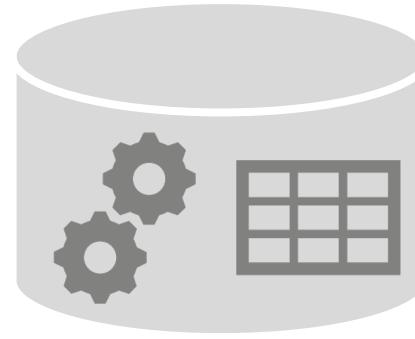


**Whole**



# Ideally, analyze in place

Write SQL:  
`select count() from sales where amount > 1000 group by month`



Returns a **data.frame** with **12 records**

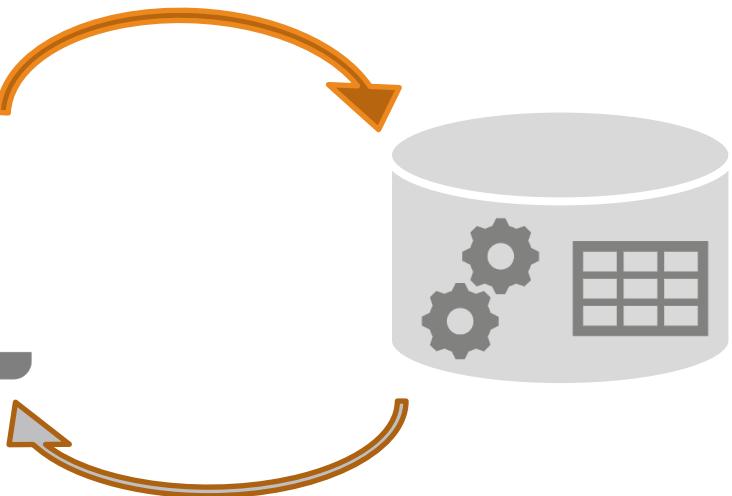
# Ideally, analyze in place, using **dplyr**



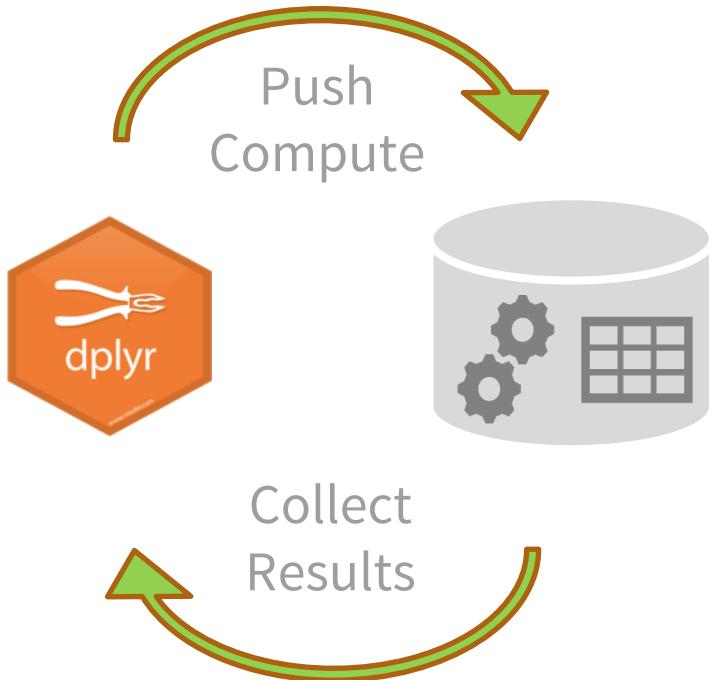
```
sales %>%
  filter(amount > 1000) %>%
  group_by(month) %>%
  tally()
```

dplyr writes  
the SQL  
query

```
select count() from sales where
amount > 1000 group by month
```



# Advantages of using **dplyr**



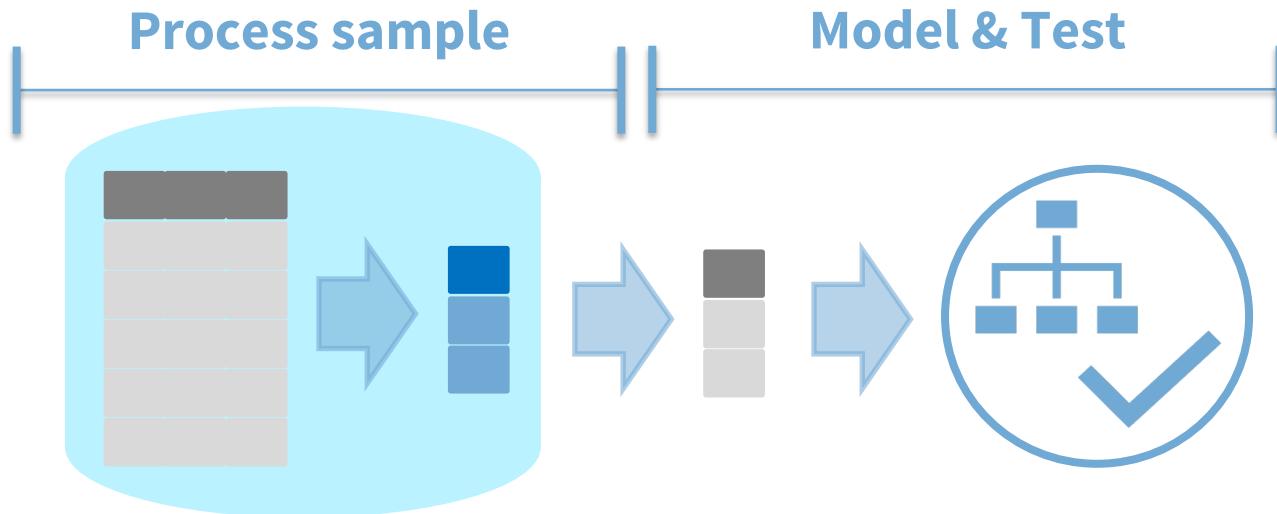
1. **dplyr** translates to SQL
2. Take advantage of piped code
3. All your code is in R!

# Modeling with Databases

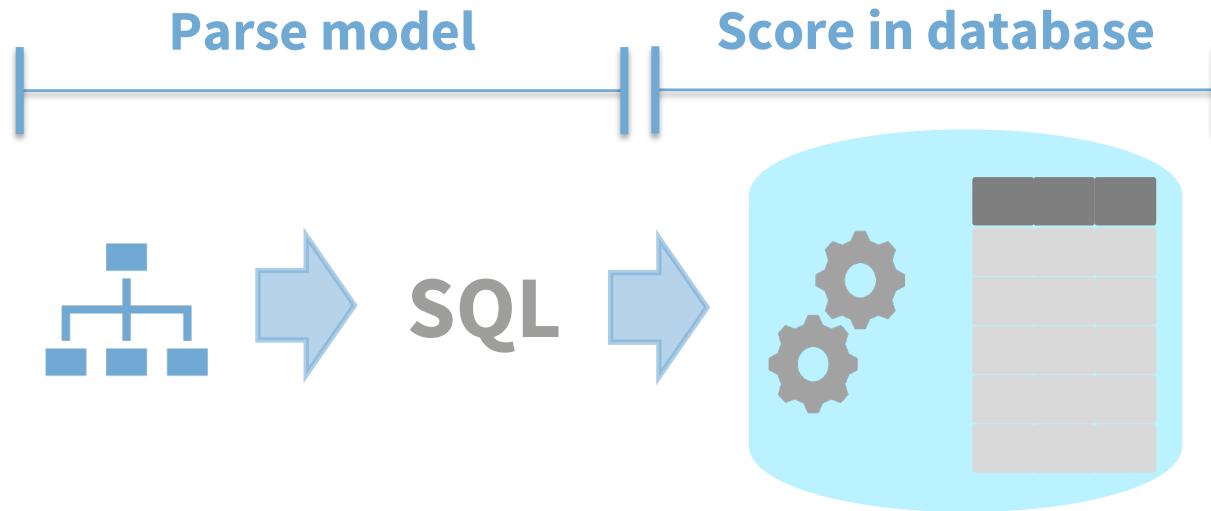
A chalkboard with three equations handwritten in white chalk. The first equation is  $\frac{dC}{dt} = qV_{act} - \eta_0(N-N_0)(1-\varepsilon S)S + \frac{\nu e}{T_n} - \frac{N}{T_p}$ . The second equation is  $\frac{dS}{dt} = T_b \eta_0 (N-N_0)(1-\varepsilon S)S + \frac{\eta e N}{T_n} - \frac{S}{T_p}$ . The third equation is  $\frac{S}{P_t} = \frac{T_p \eta e \lambda^0}{V_{act} \eta_0} \rightarrow [S \leq \frac{1}{\varepsilon}]$ . To the right of the equations, there is a bracketed note  $N = 1$  and  $P_t = (m)$ .

$$\left\{ \begin{array}{l} \frac{dC}{dt} = qV_{act} - \eta_0(N-N_0)(1-\varepsilon S)S + \frac{\nu e}{T_n} - \frac{N}{T_p} \\ \frac{dS}{dt} = T_b \eta_0 (N-N_0)(1-\varepsilon S)S + \frac{\eta e N}{T_n} - \frac{S}{T_p} \\ \frac{S}{P_t} = \frac{T_p \eta e \lambda^0}{V_{act} \eta_0} \rightarrow [S \leq \frac{1}{\varepsilon}] \end{array} \right. \begin{array}{l} N = 1 \\ P_t = (m) \end{array}$$

# Modeling with a Database



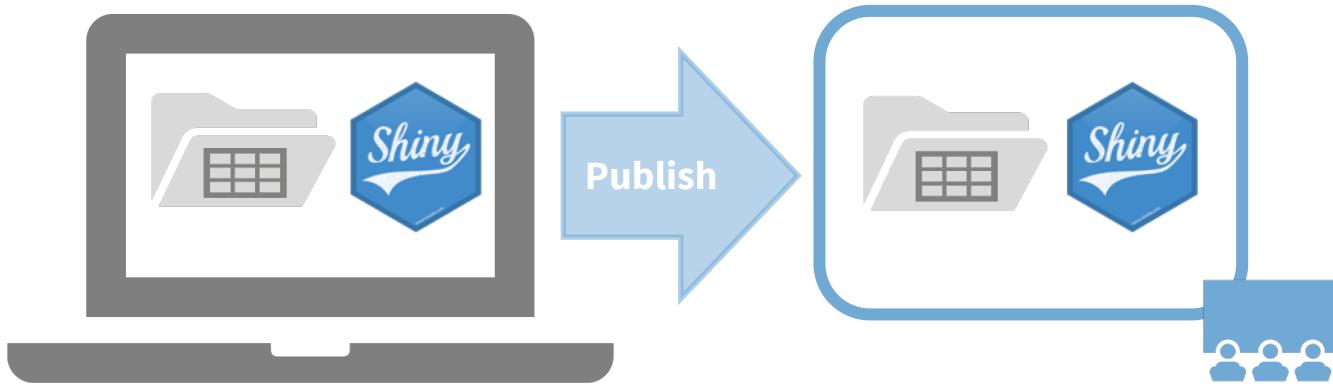
# Score inside the DB using **tidypredict**



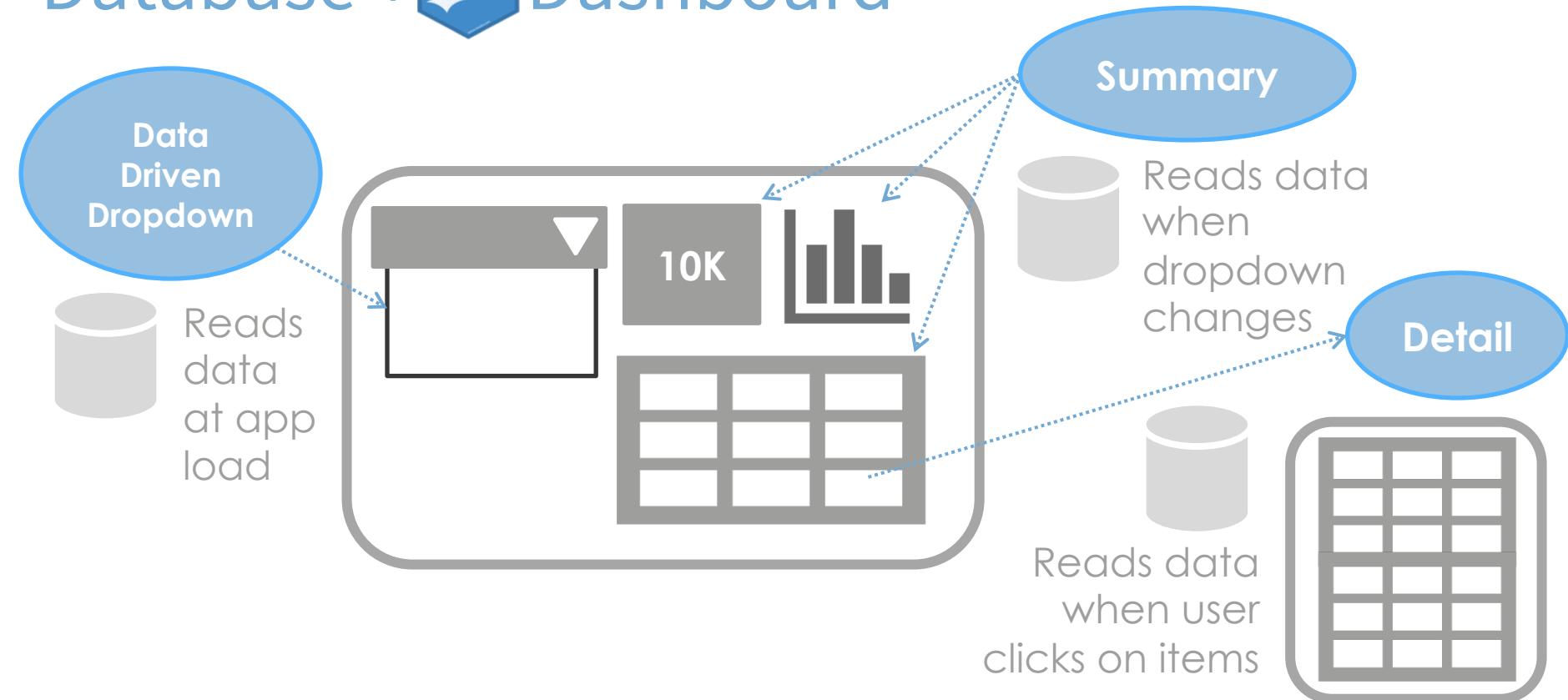
# Dashboards with Databases



# Normal Shiny app



# Database + Dashboard



# Push computation to the database



# Recommended sites

db.rstudio.com

Databases using R from RStudio

dplyr DBI Best Practices Databases Advanced News

**Packages**

- dplyr
- DBI
- odbc
- pool
- dbplot
- tidypredict

**RStudio**

- Connections Pane
- Professional Drivers

**Best Practices**

- Setting up ODBC Drivers
- Run Queries Safely
- Securing Deployed Content
- Securing Credentials
- Making Scripts Portable
- Creating Visualizations
- Selecting a database interface
- Enterprise-ready dashboards

## Databases using R

At RStudio, we are working to make it as easy as possible to work with databases in R. This work focuses on **three key areas**:

### 1. RSTUDIO PRODUCTS

- The new RStudio [Connections Pane](#) makes it possible to easily connect to a variety of data sources, and [explore the objects and data](#) inside the connection
- To RStudio commercial customers, we offer [RStudio Professional ODBC Drivers](#); these are data connectors that help you connect to some of the most popular databases.

### 2. USE BEST-IN-CLASS PACKAGES

Build and/or document how to use packages such as: [dplyr](#), [DBI](#), [odbc](#), [keyring](#) and [pool](#)

### 3. PROMOTE BEST PRACTICES

This website is the main channel to provide support in this area. RStudio is also working through other delivery channels, such as upcoming webinars and in-person training during our RStudio conferences.

[Read more →](#)

community.rstudio.com

R Studio Community

all categories all tags Latest New Unread Categories Top + New Topic

Topic	Category	Users	Replies	Views	Activity
Getting nice-looking tables wider than a page when rendering to R Markdown	R Markdown	A, B, C	4	39	1h
Rstudio server for windows server 2012		J, H	1	11	1h
Windows Pane Layout Breaks Under Certain Combinations	RStudio IDE	f, K, L	5	79	1h
From RStudio, is it possible to knit only part of an R Markdown document?	R Markdown	A, B, C, D, E	14	82	1h
How to use 'map' with 'cor'	tidyverse	i, M, J	4	53	1h
Shiny - How to trigger an event when a selectInput object is modified OR the user makes a selection?	shiny	G, H	3	46	1h
Should we use data or static/data to store input files	blogdown	K, L	2	30	2h
Which frame is live?	IDE	H	1	11	2h
Embed RStudio Videos?	General	B, C	2	41	2h
Lattice or GGplot2	ggplot2, lattice	M, B, P	18	317	3h
How much RAM is available?	RStudio Cloud	R	0	11	3h

# Thank You!

**Rate This Session # 1050**

with the Teradata Analytics Universe Mobile App

**Follow Me**

Twitter @nwstephens

**Questions/Comments**

LinkedIn: <https://www.linkedin.com/in/nwstephens/>