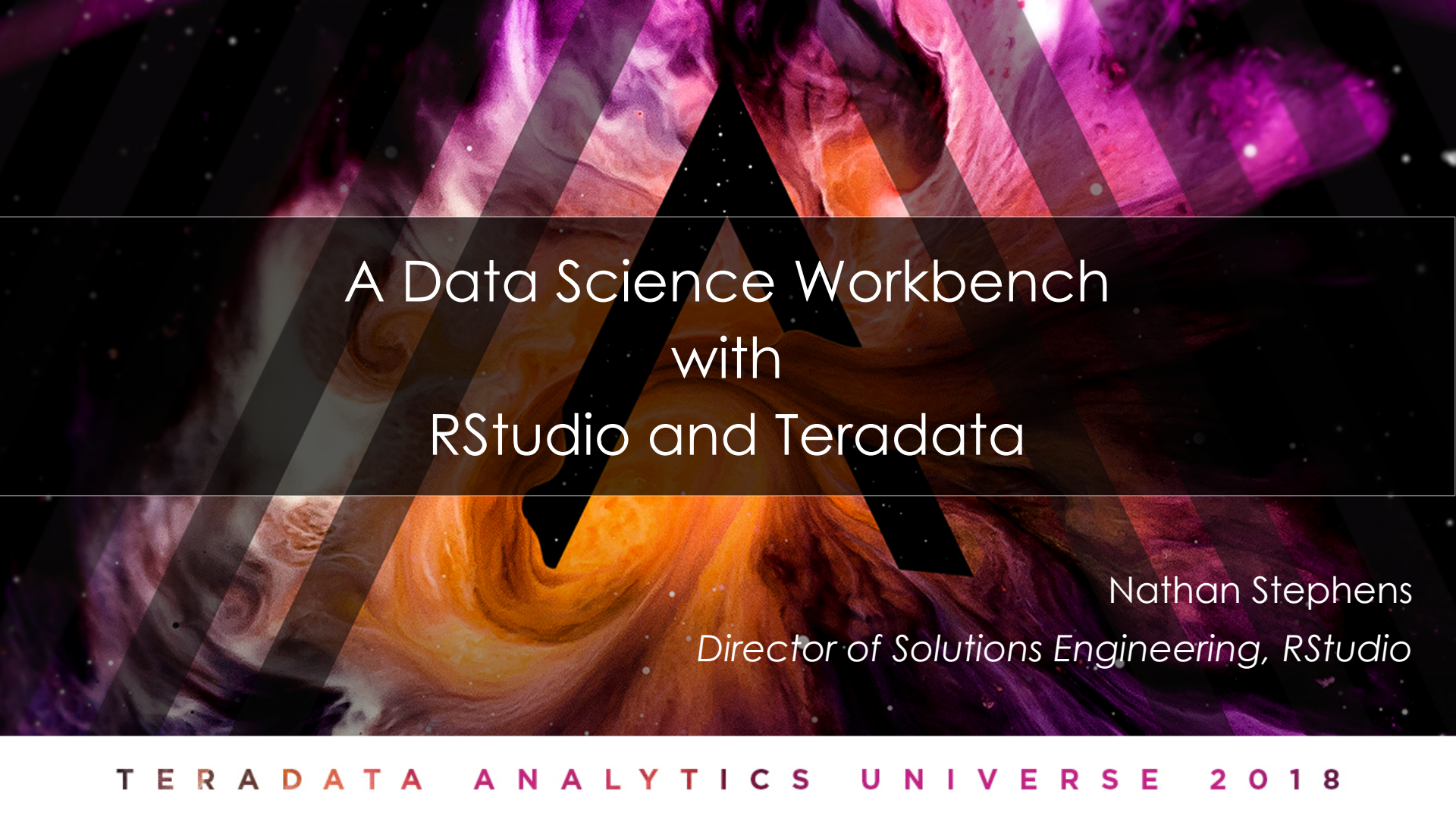


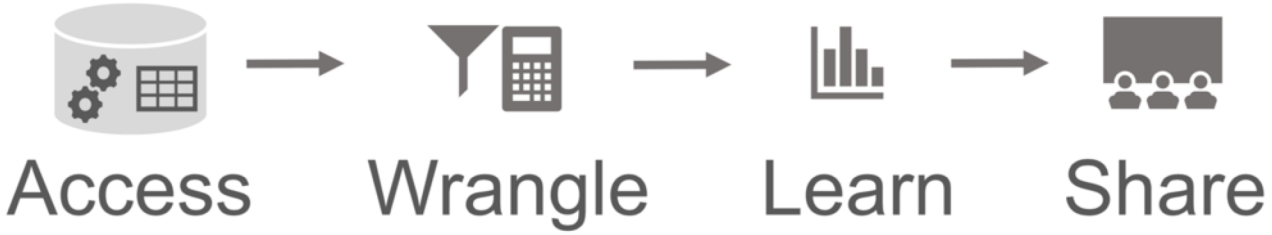
A Data Science Workbench with RStudio and Teradata

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

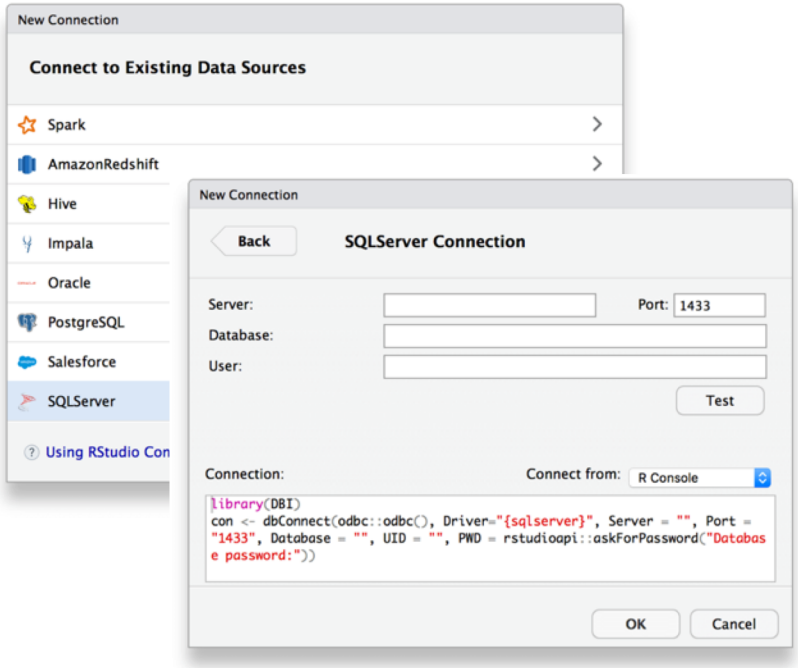
3. No centralized place to get information

4. Hard to communicate insights

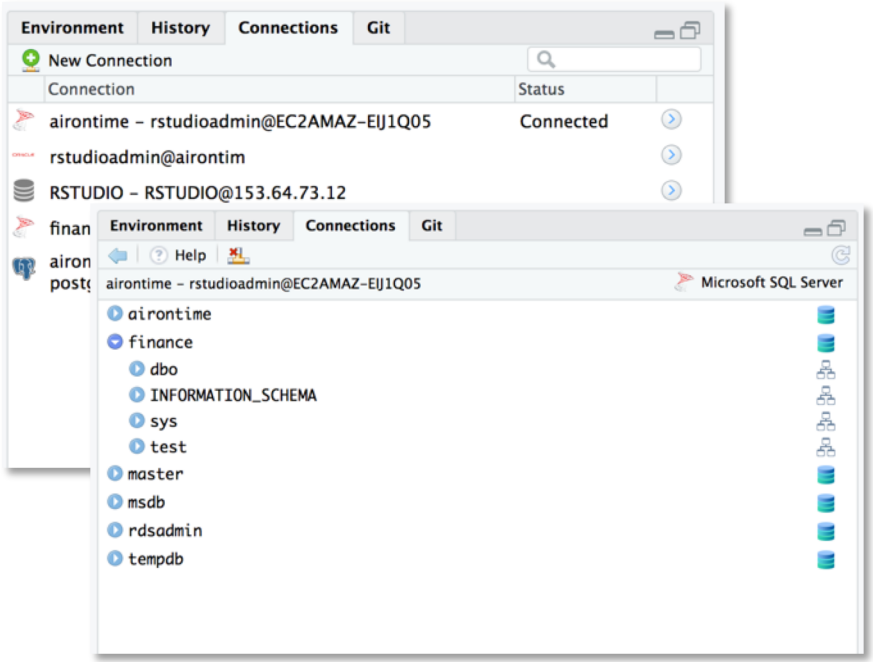

What's New

RStudio version 1.1 features

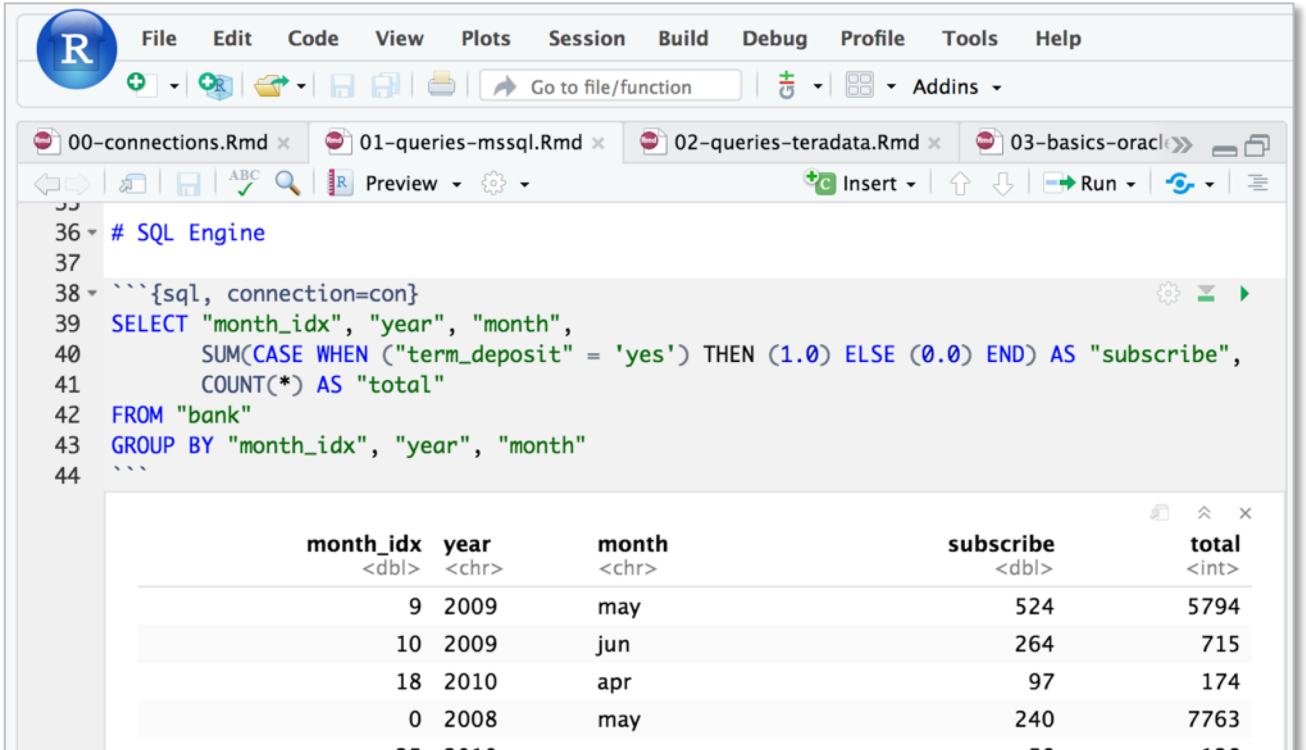
Connection wizard



Connections tab



SQL Language Engine for R Markdown



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

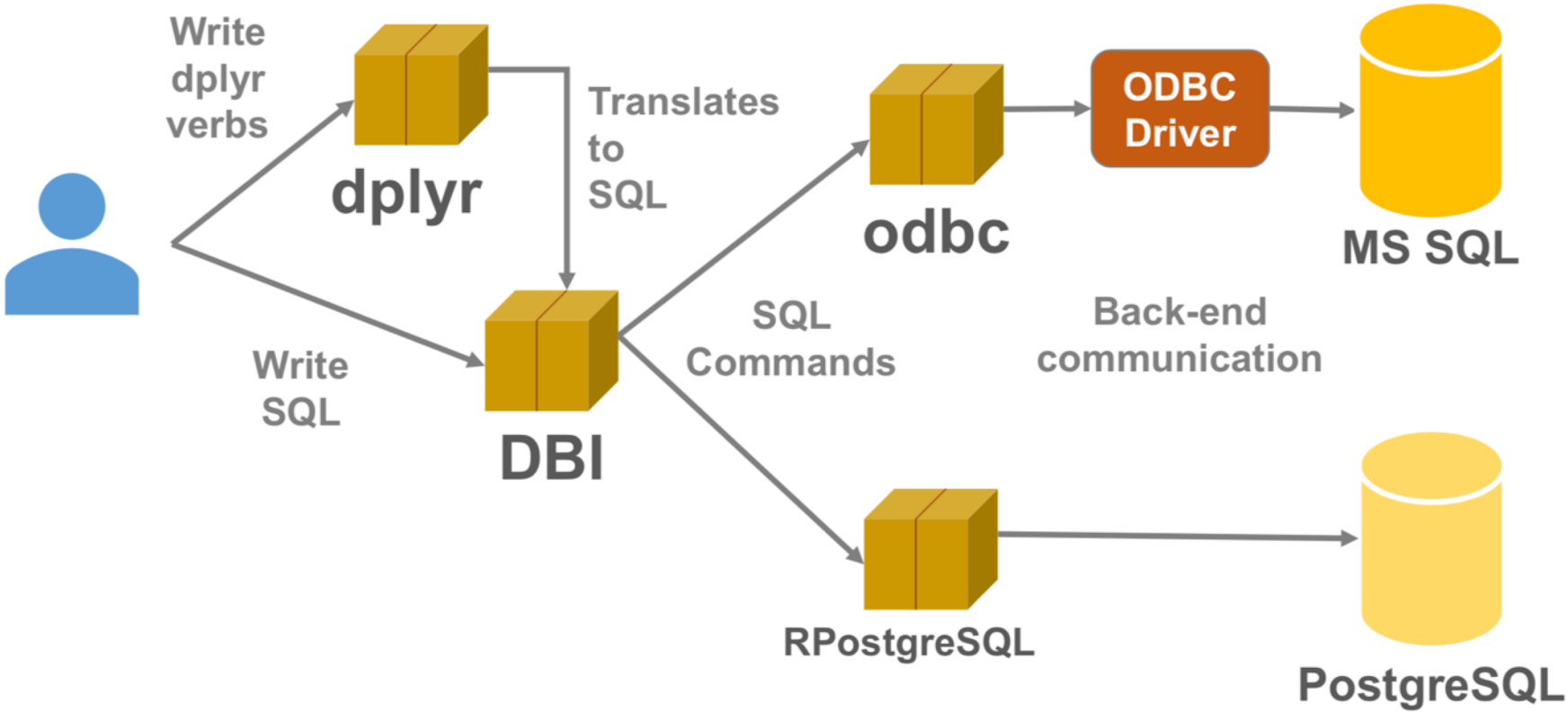
R Packages

- RSQLite
- rpostgres
- bigrquery
- tdplyr

RJDBC

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

Establish a database connection with RStudio



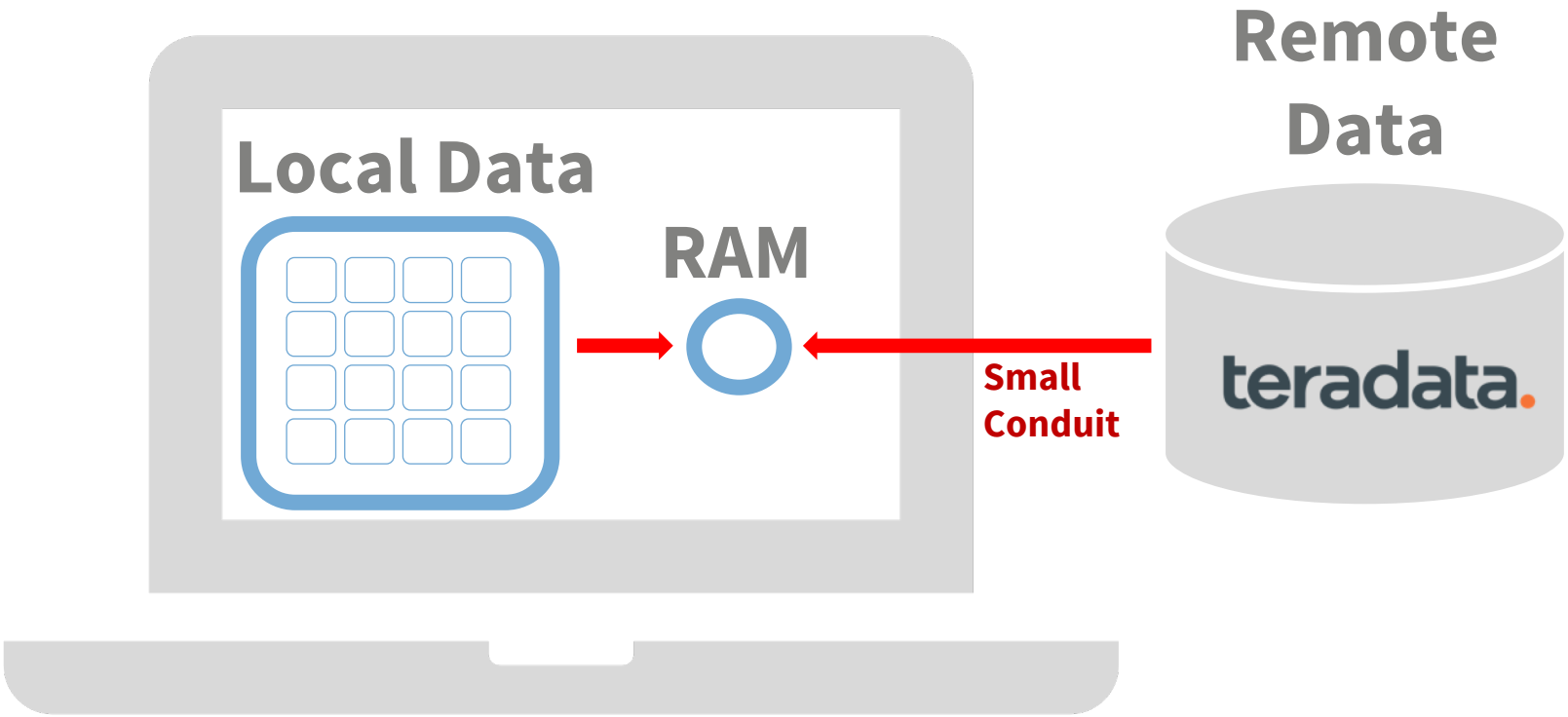
RStudio Professional Drivers

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

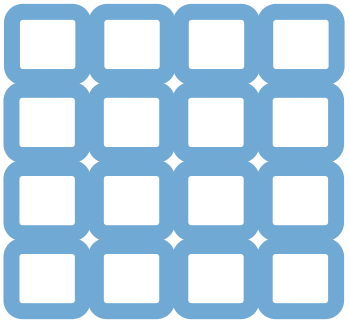
Big Data with R



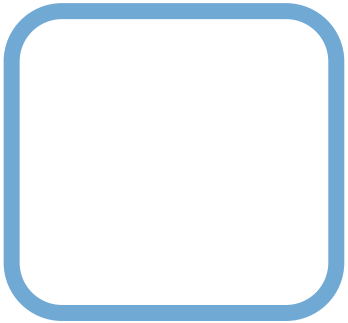
Big Data Strategies

Sample

Parts



Whole



Ideally, analyze in place

Write SQL:

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

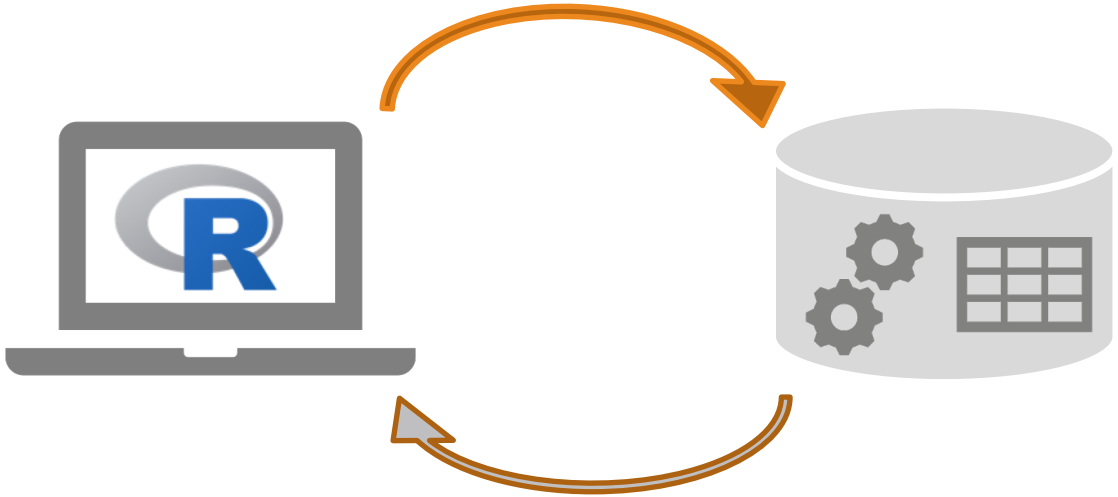
Returns a **data.frame** with **12 records**

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

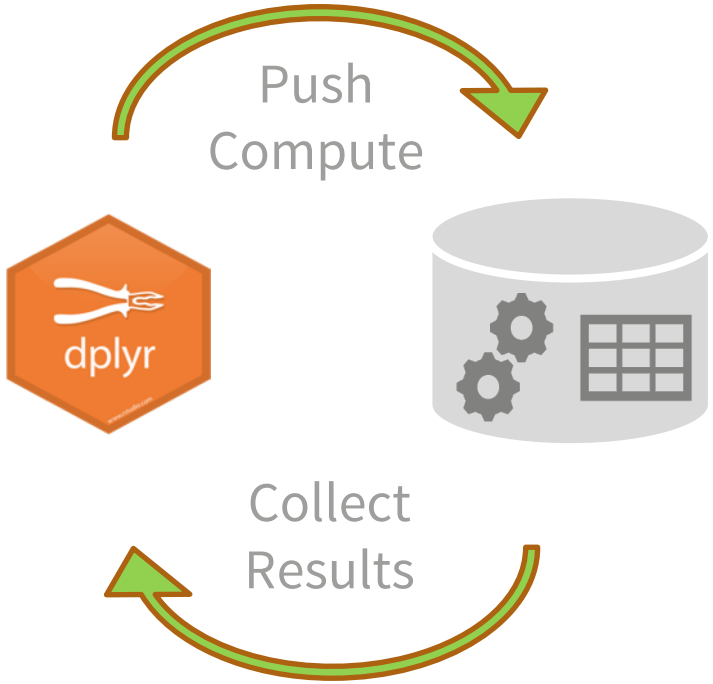
dplyr writes the SQL query

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

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

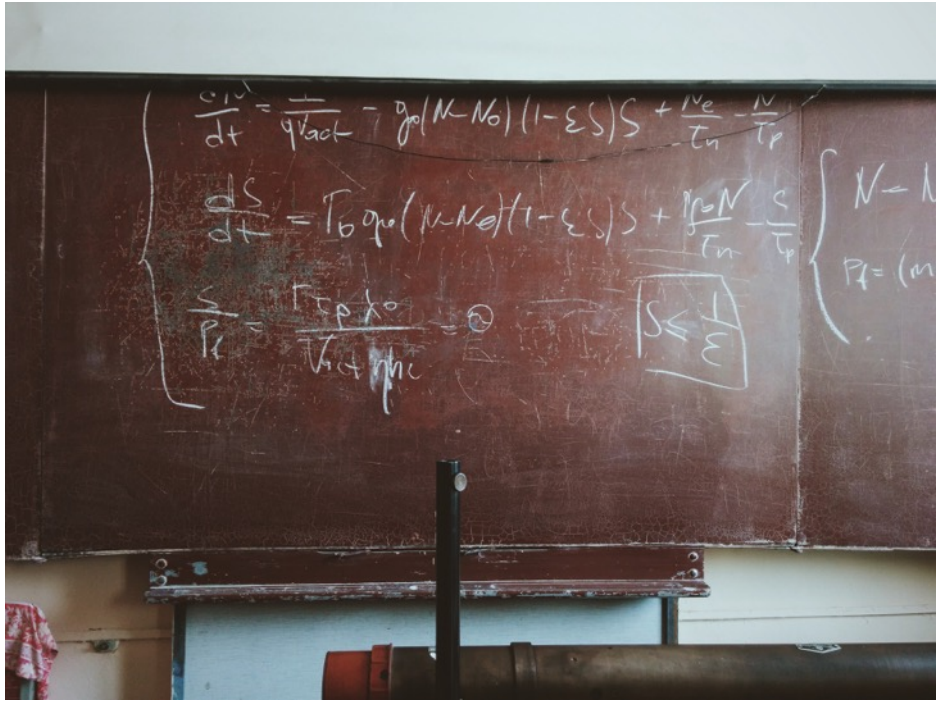


Advantages of using **dplyr**

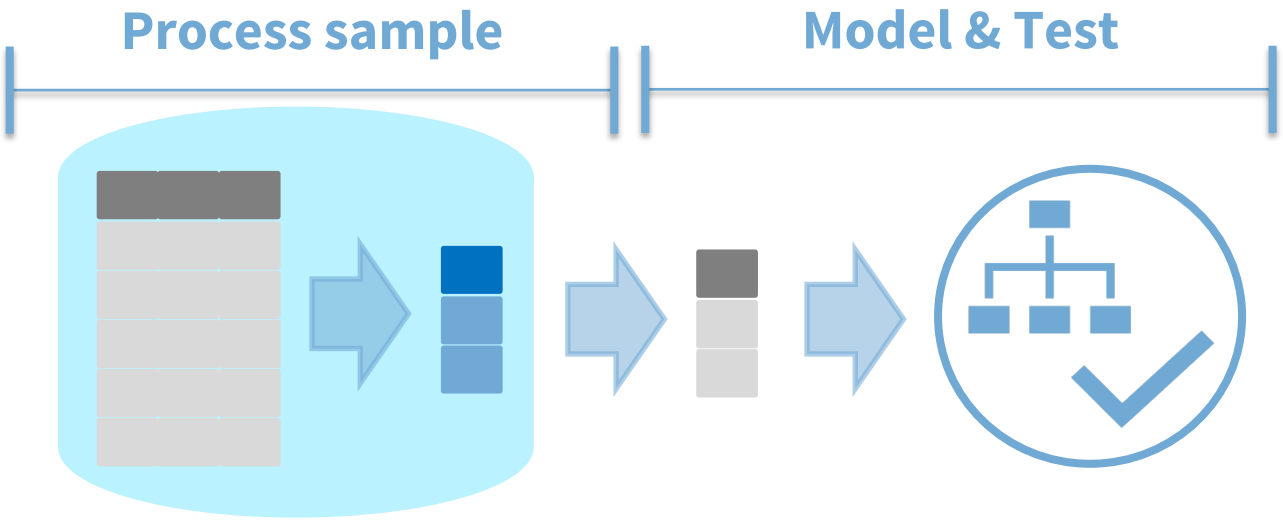


1. dplyr translates to SQL
2. Take advantage of piped code
3. All your code is in R!

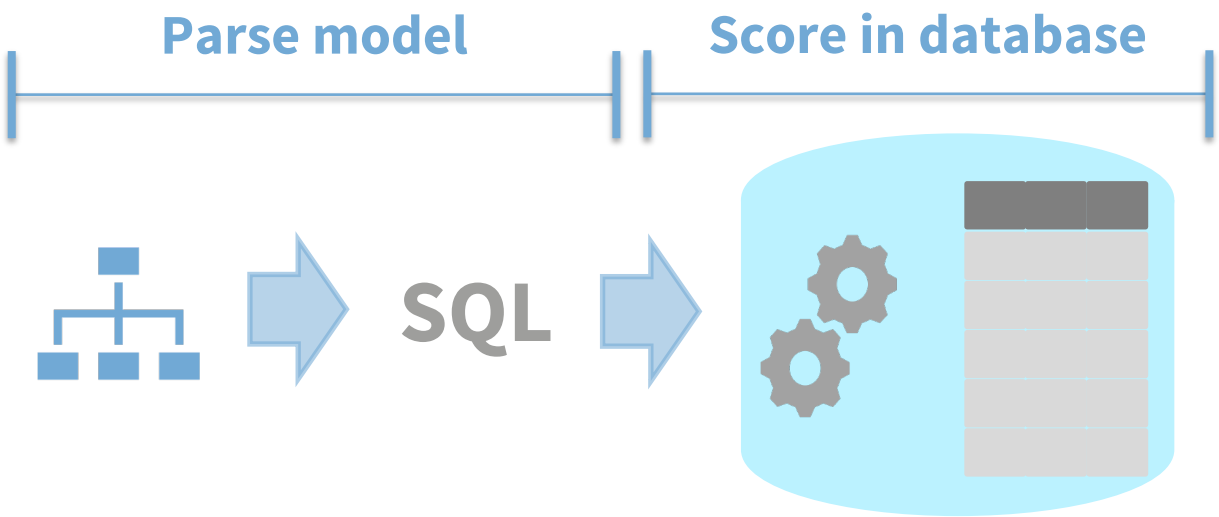
Modeling with Databases



Modeling with a Database

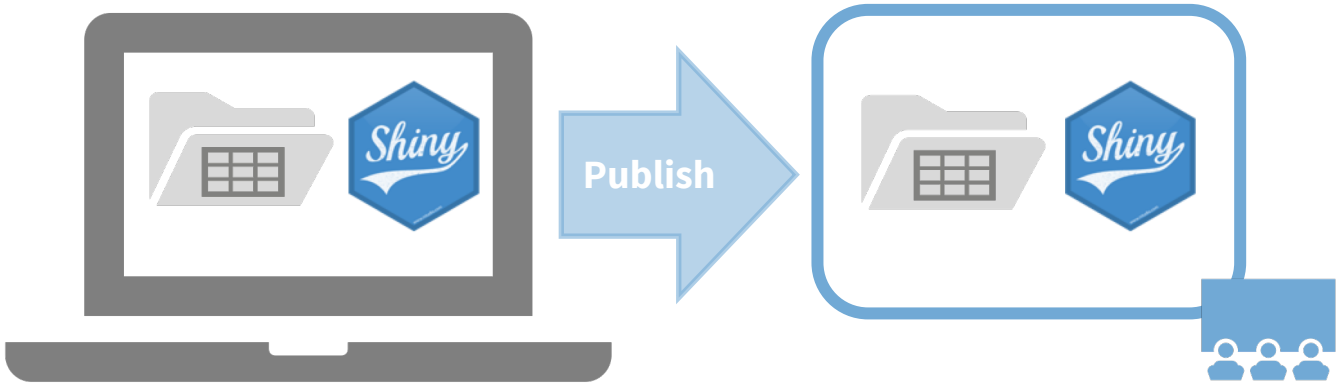


Score inside the DB using **tidypredict**

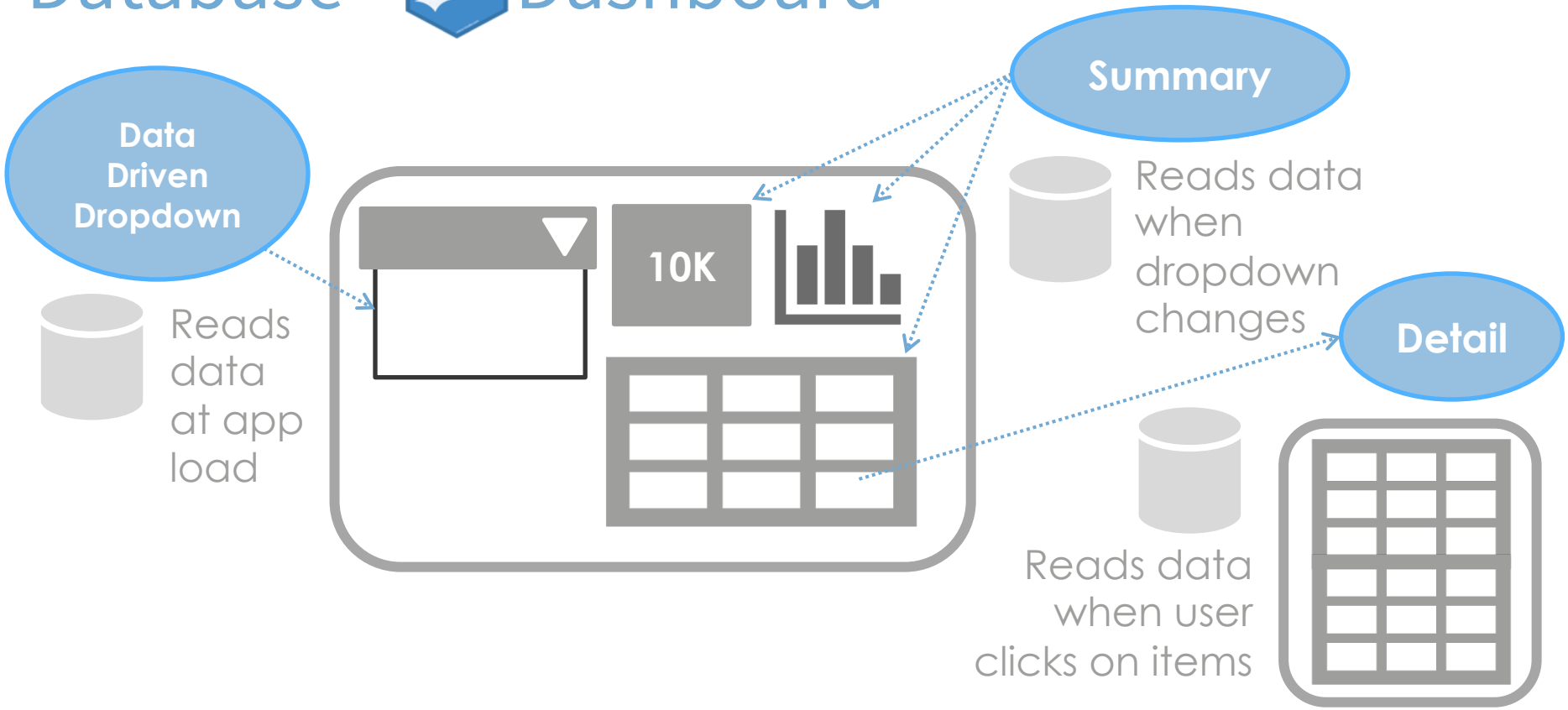


Dashboards with Databases

Normal Shiny app



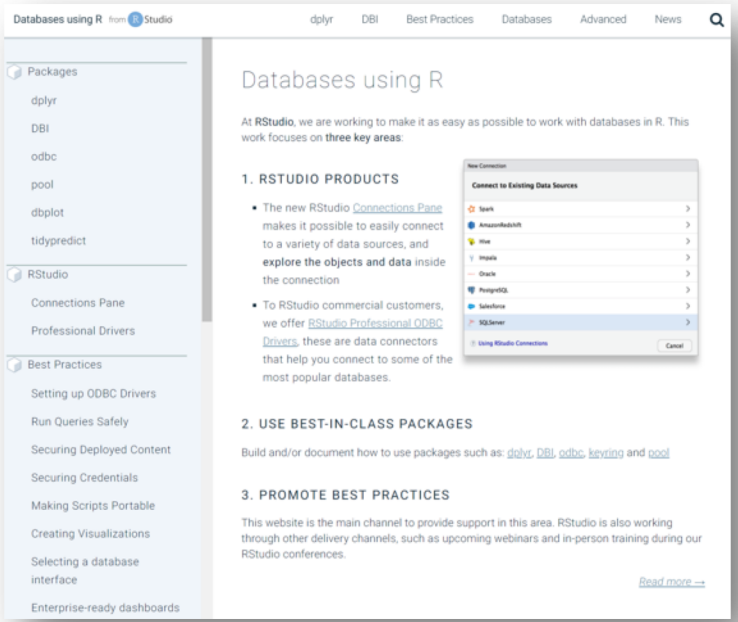
Database + Shiny Dashboard



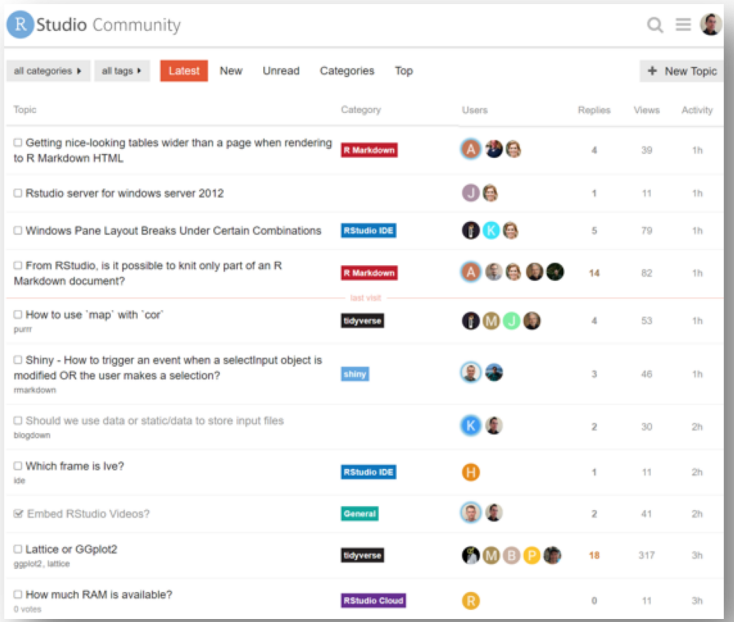
Push computation to the database

Recommended sites

db.rstudio.com



community.rstudio.com



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/