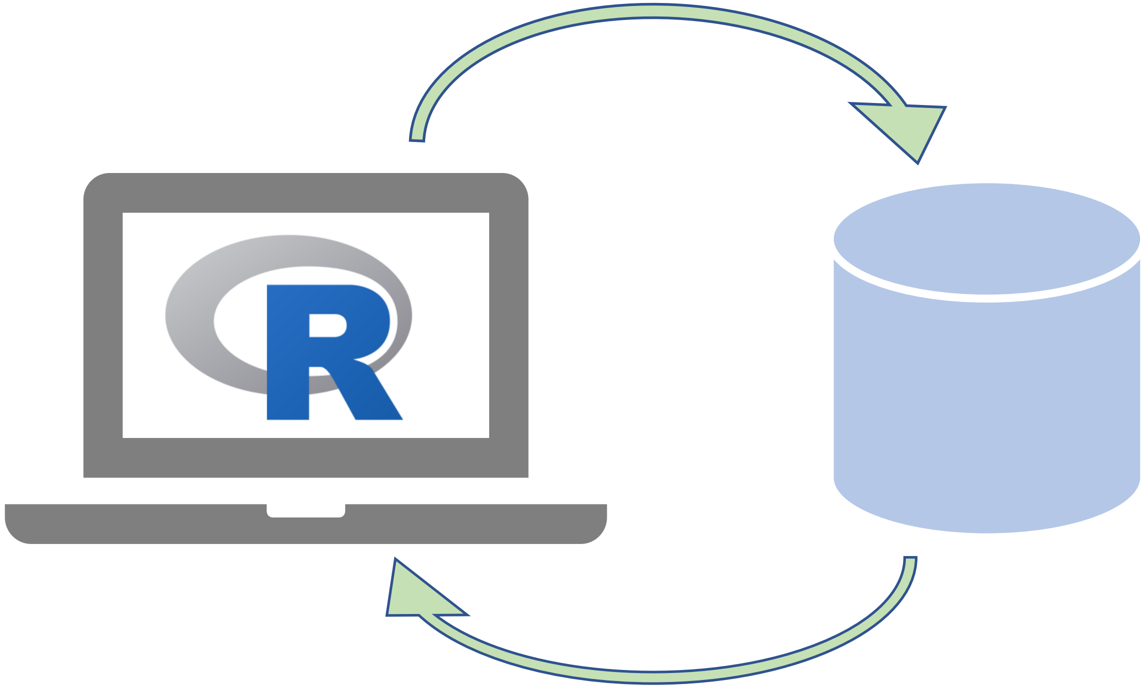
Big Data with R

Edgar Ruiz

@theotheredgar
linkedin.com/in/edgararuiz

May 2018



RStudio

db.rstudio.com / spark.rstudio.com

Let’s talk about **Big Data**

Photo by Chris Christensen on Unsplash

Big data?

Data > RAM

Garrett Grolemund

Remote Data

Edgar Ruiz

Velocity

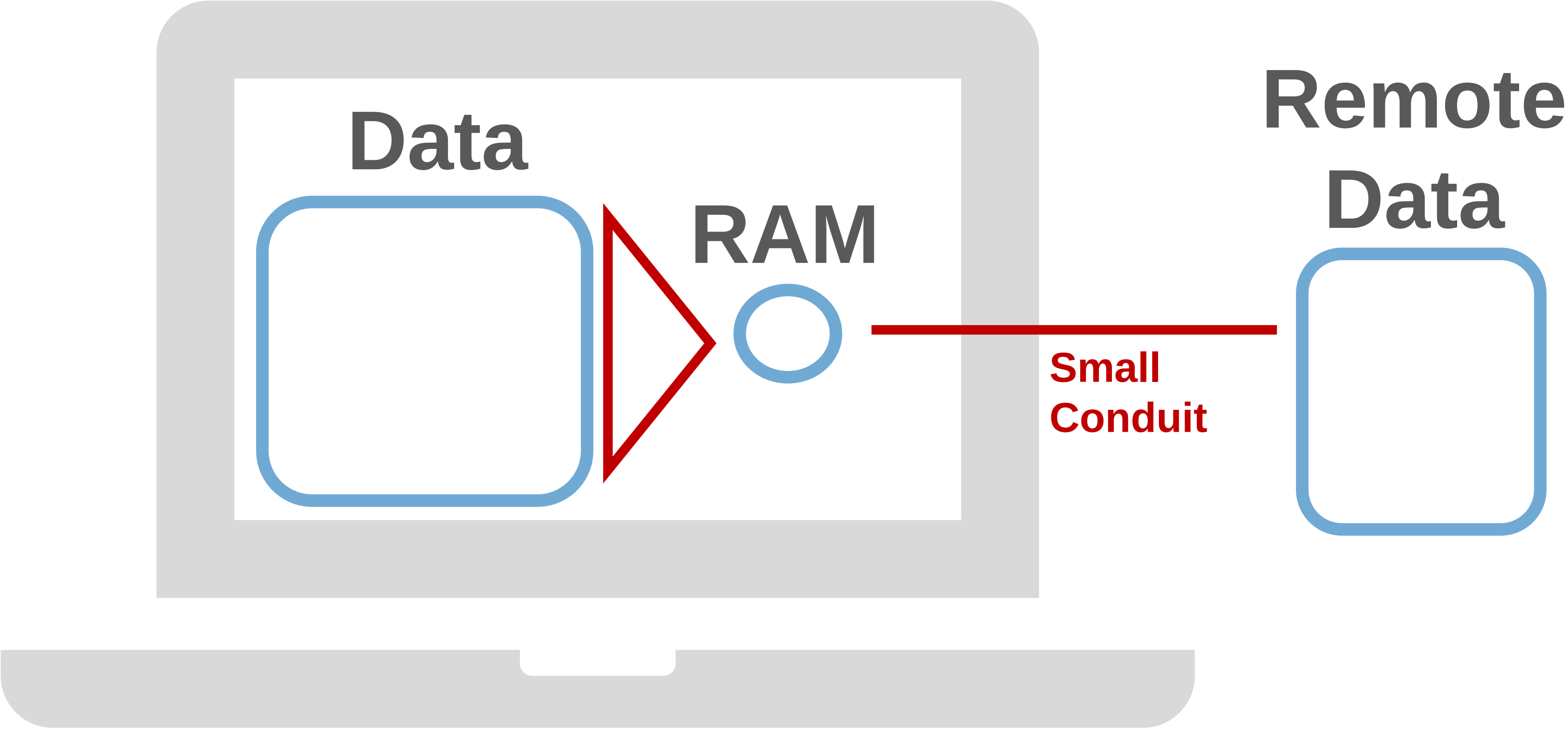
Volume

Value

Variety

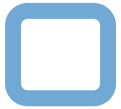
Veracity

Big Data in R



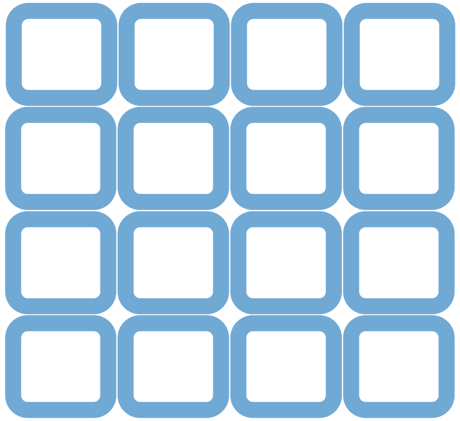
Big Data Strategies

Sample



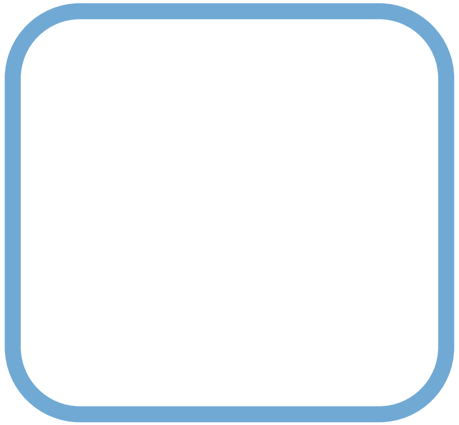
Most common approach for **modeling**

Parts



Most common approach for **general analysis**

Whole



In most cases, **the preferred approach,** it's just not feasible

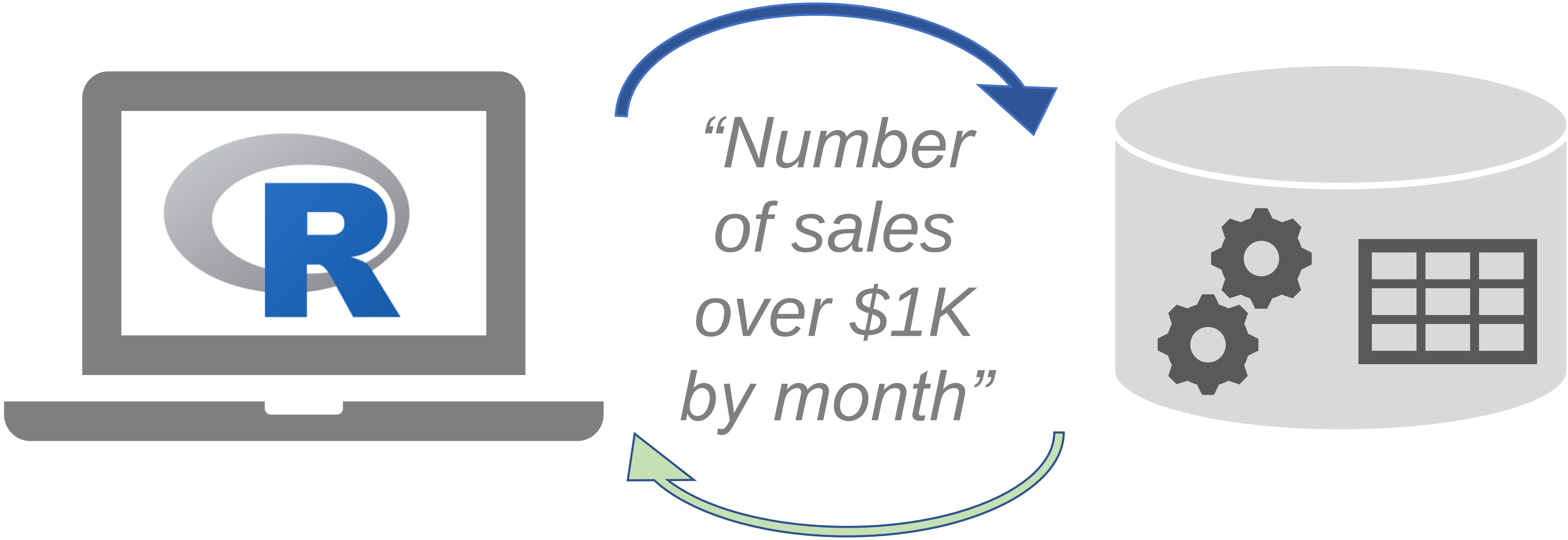
Push compute, collect results

Photo by Arthur Lambillotte on Unsplash

Ideally, analyze in place

Write SQL:

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



“Number of sales over $1K by month”

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

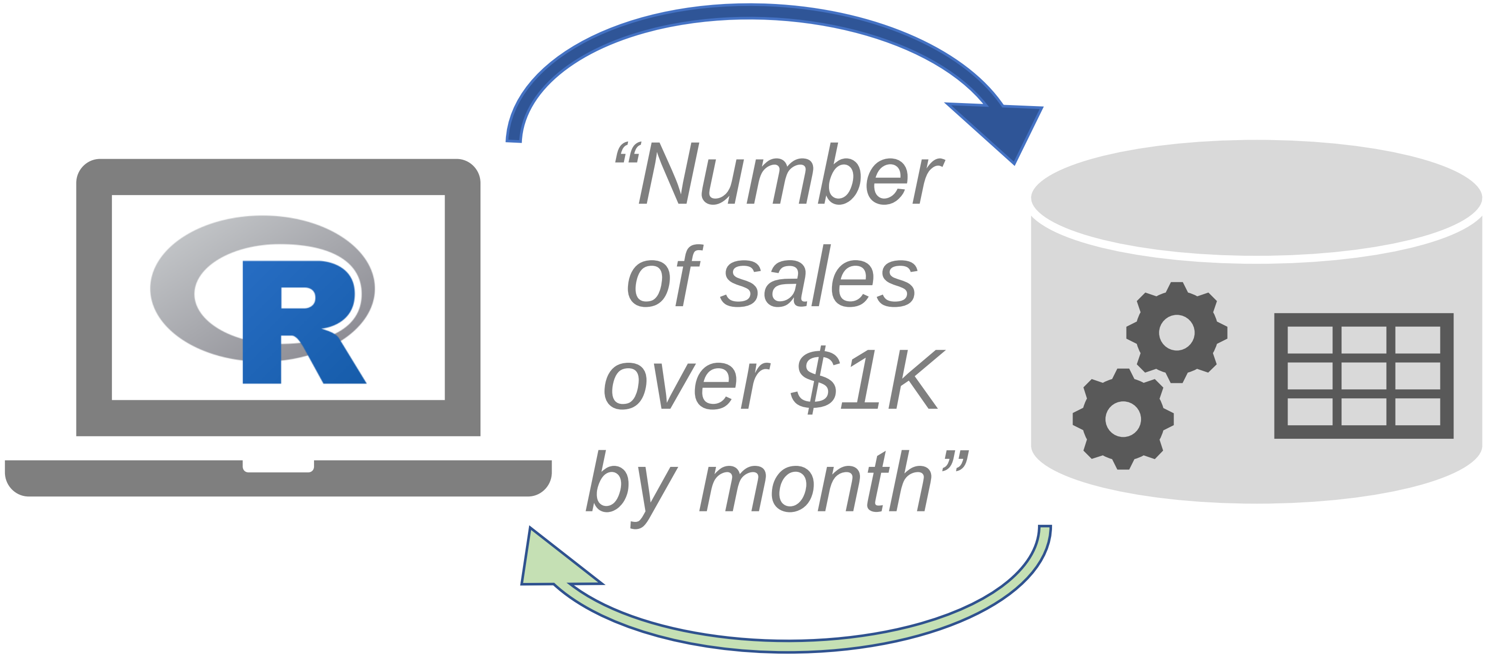
Ideally, analyze in place, using **dplyr**

dplyr writes
the SQL
query

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

Write dplyr:

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



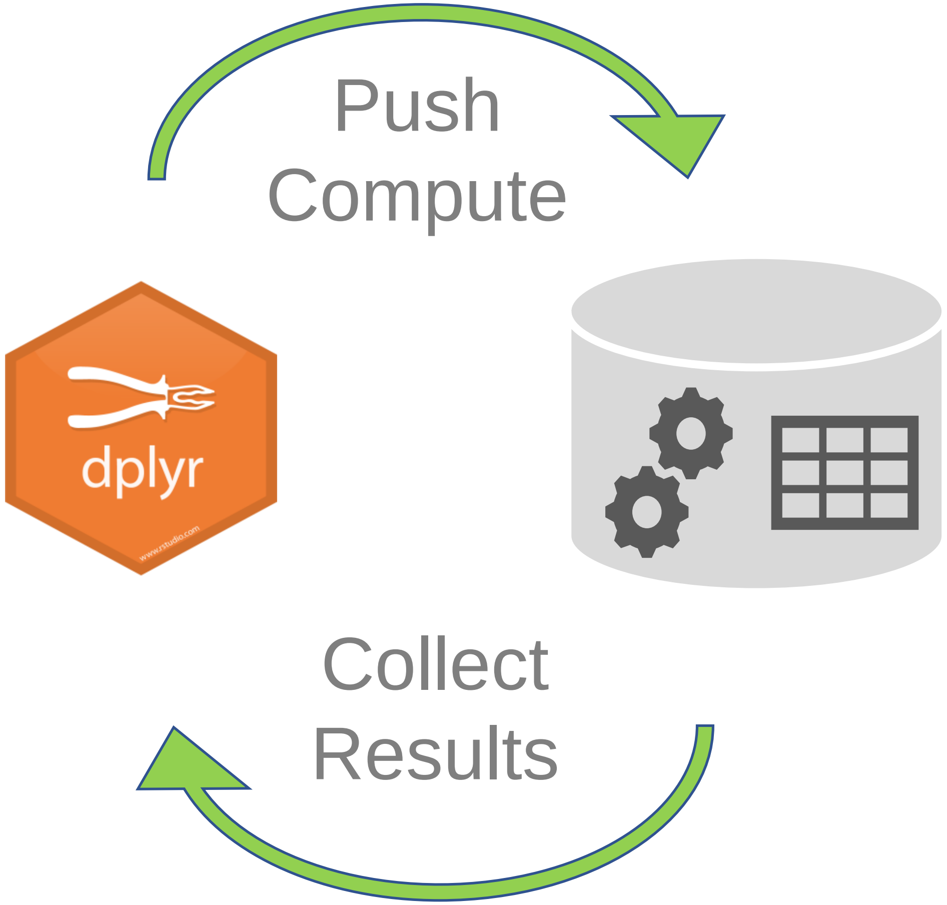
"Number of sales over $1K by month"

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

Available translations

1. Microsoft SQL Server
2. Oracle
3. Apache Hive
4. Apache Impala
5. PostgreSQL
6. MS Access
7. MariaDB (MySQL)
8. SQLite
9. Amazon Redshift
10. Teradata

Advantages of using **dplyr**



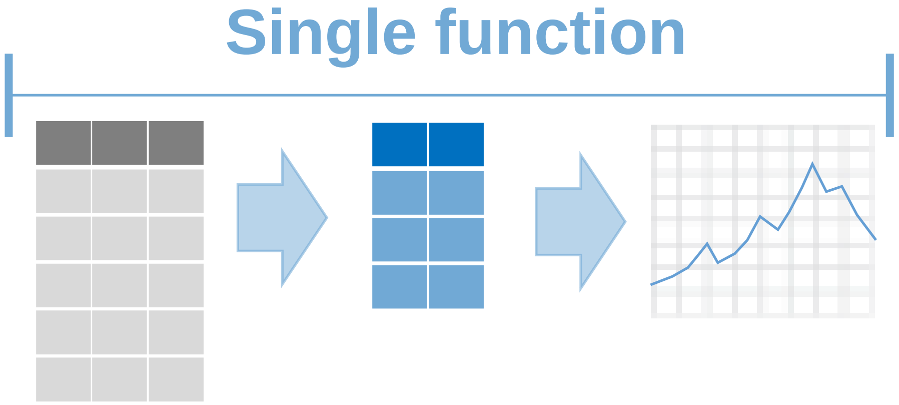
1. dplyr translates to SQL
2. Take advantage of piped code
3. All your code is in R!

Visualizations

Photo by Luis Alfonso Orellana on Unsplash

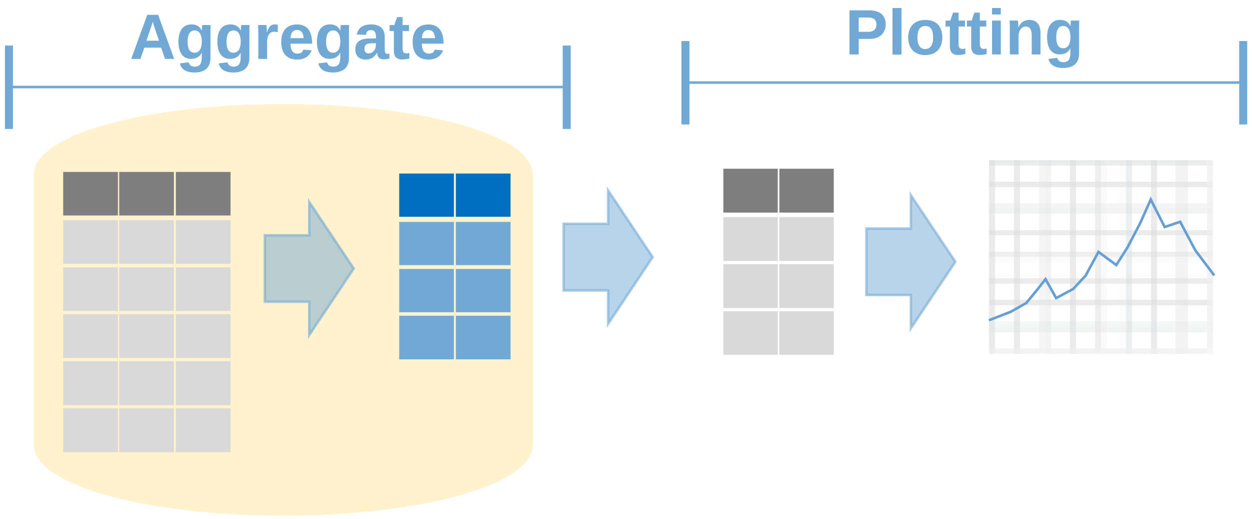
Aggregate in DB, plot locally

Local data



Single function

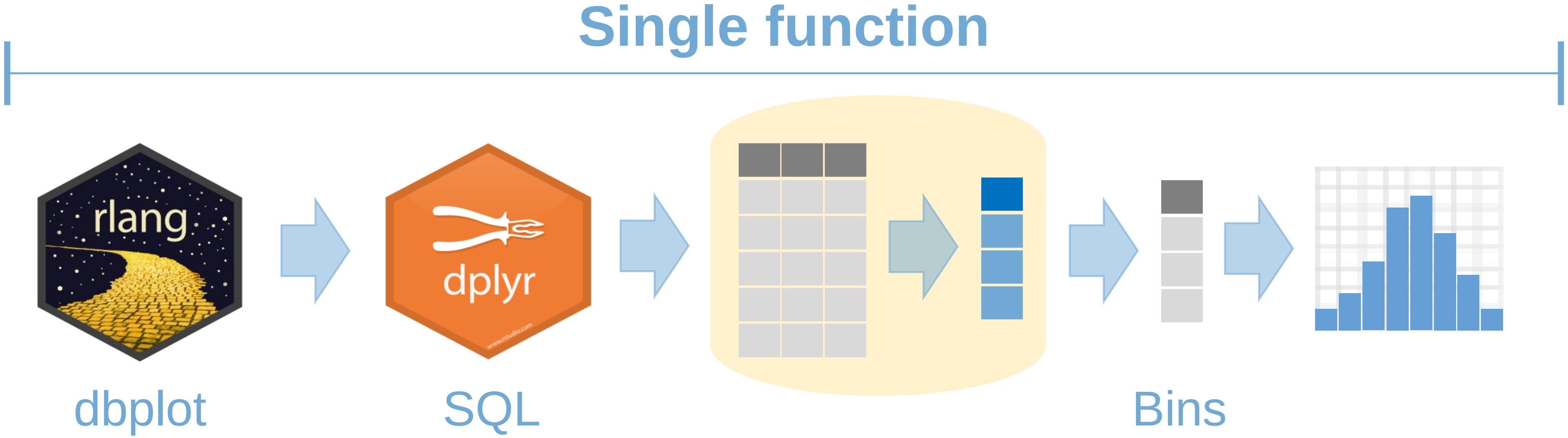
Remote data



Aggregate

Plotting

dbplot for Histograms & Raster plots



Modeling with **Databases**

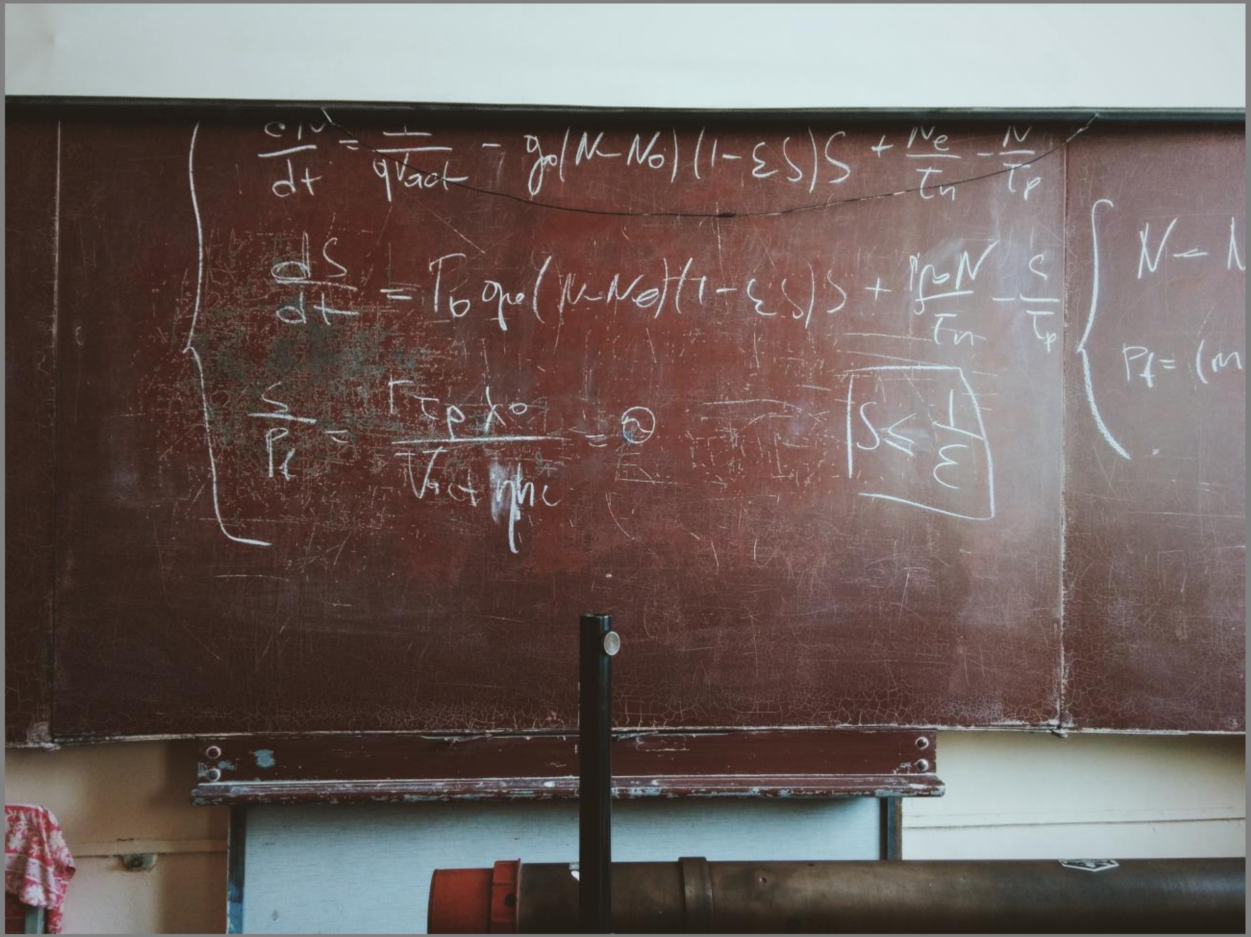
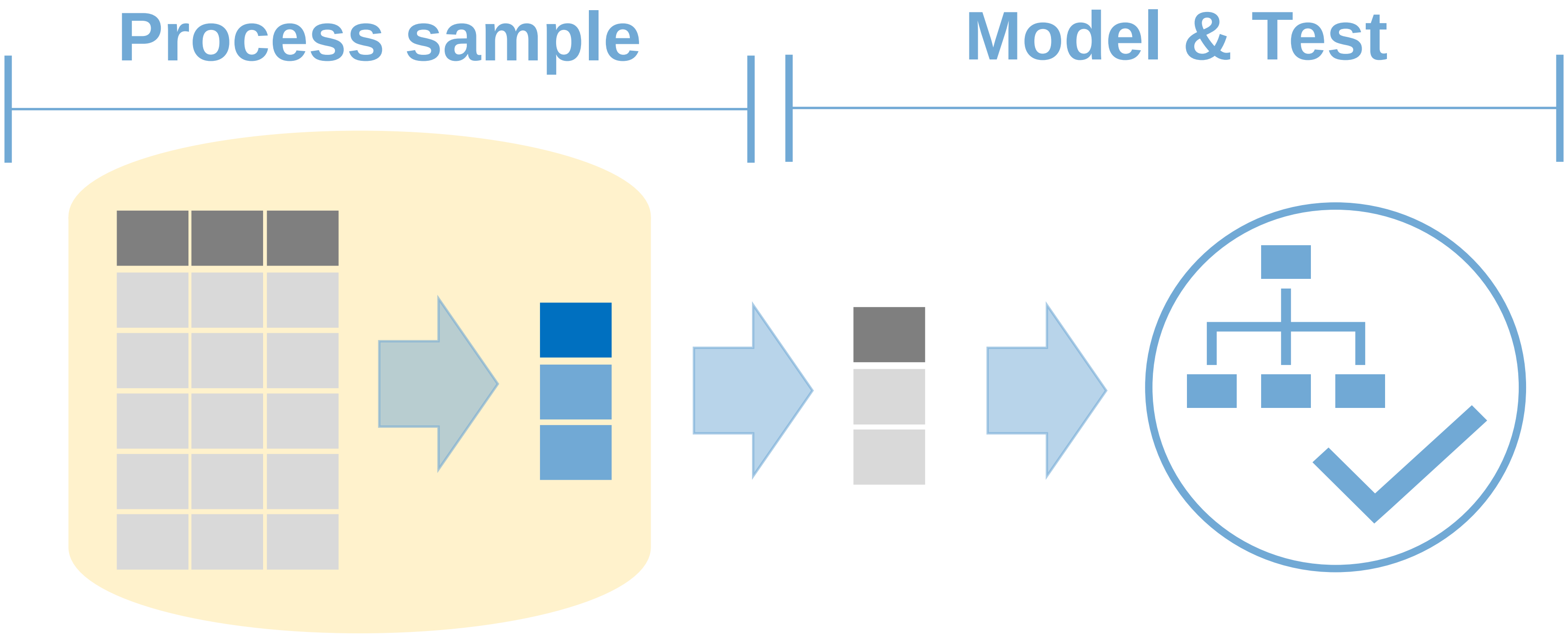
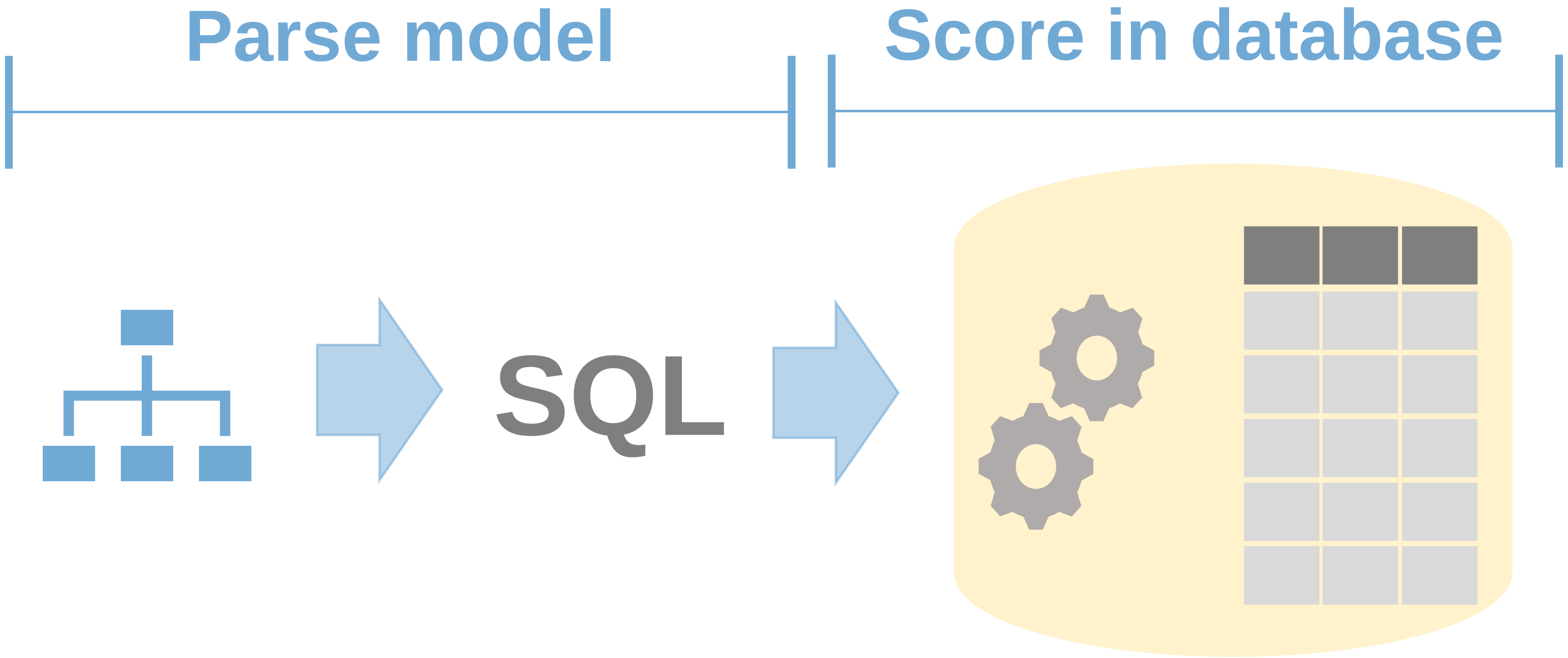


Photo by Roman Mager on Unsplash

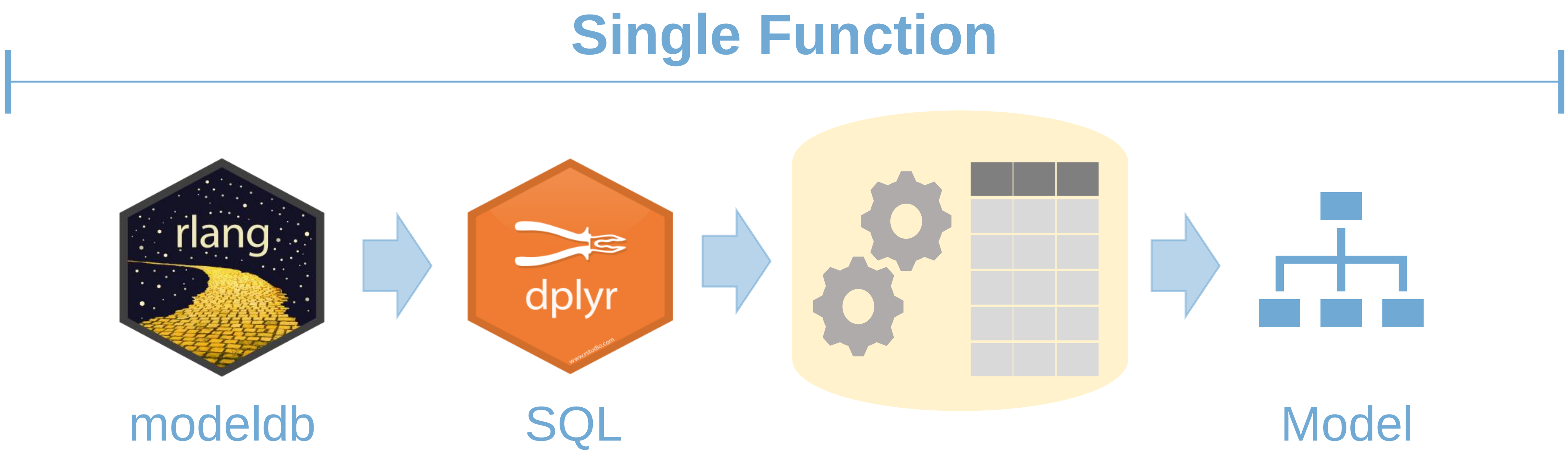
Option 1 - Modeling with a Database



Score inside the DB using **tidypredict**



Option 2 - Modeling in DB using **modeldb**



Single Function

modeldb

SQL

Model

Modeling with **sparklyr**

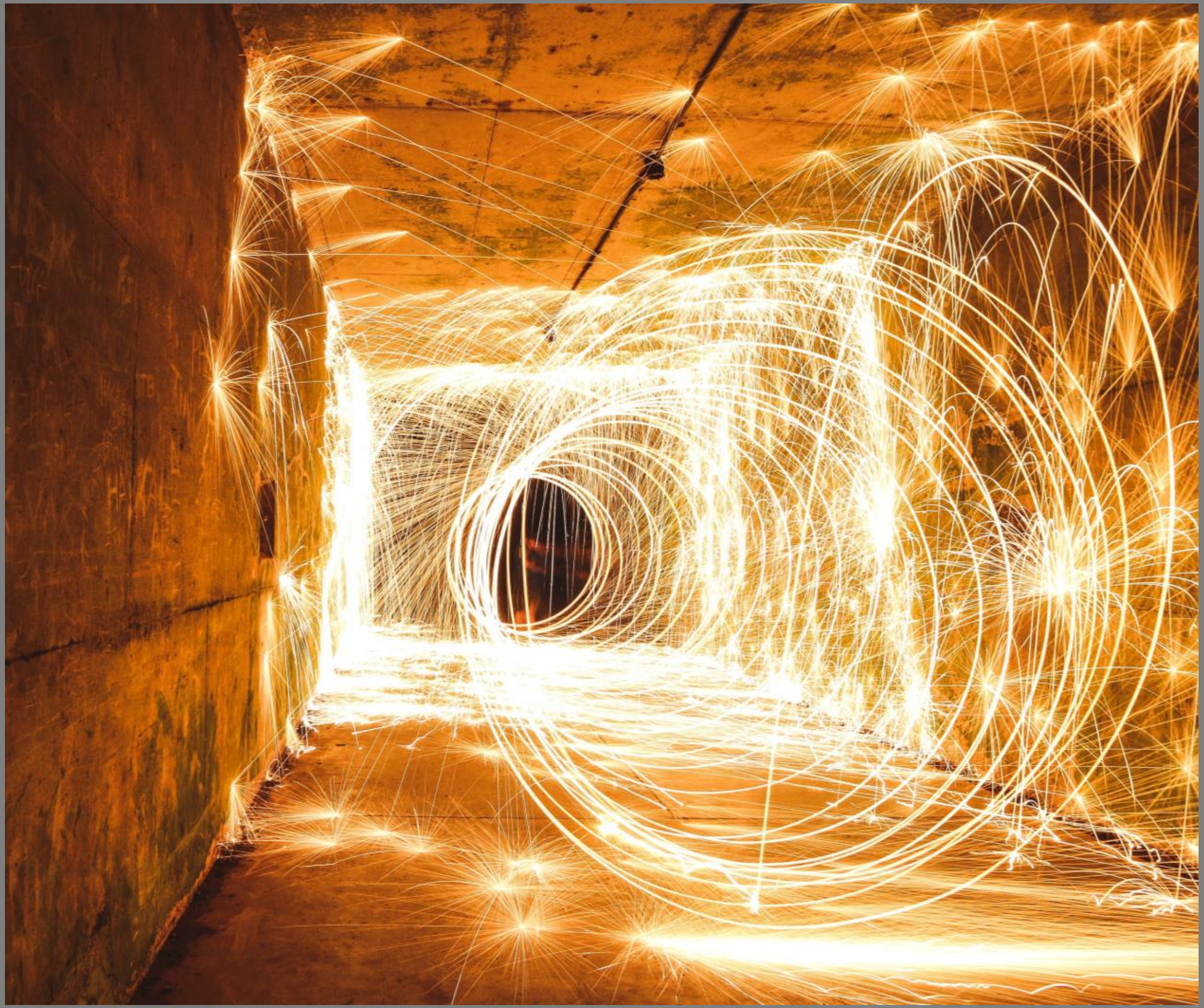


Photo by Matthew Ronder-Seid on Unsplash

RStudio

Spark models (ML) available via **sparklyr**

ALS	ml_als ml_recommend ml_als_factorization
Decision Trees	ml_decision_tree_classifier ml_decision_tree ml_decision_tree_regressor
Generalized Linear Regression	ml_generalized_linear_regression
Gradient Boosted Trees	ml_gbt_classifier ml_gradient_boosted_trees ml_gbt_regressor
K-Means Clustering	ml_kmeans ml_compute_cost
Latent Dirichlet Allocation	ml_lda ml_describe_topics ml_log_likelihood ml_log_perplexity ml_topics_matrix

Linear Regression	ml_linear_regression
Logistic Regression	ml_logistic_regression
Multilayer Perceptron	ml_multilayer_perceptron_classifier ml_multilayer_perceptron
Naive-Bayes	ml_naive_bayes
One Vs. Rest	ml_one_vs_rest
PCA (Estimator)	ft_pca ml_pca
Random Forest	ml_random_forest_classifier ml_random_forest ml_random_forest_regressor
Survival Regression	ml_aft_survival_regression ml_survival_regression

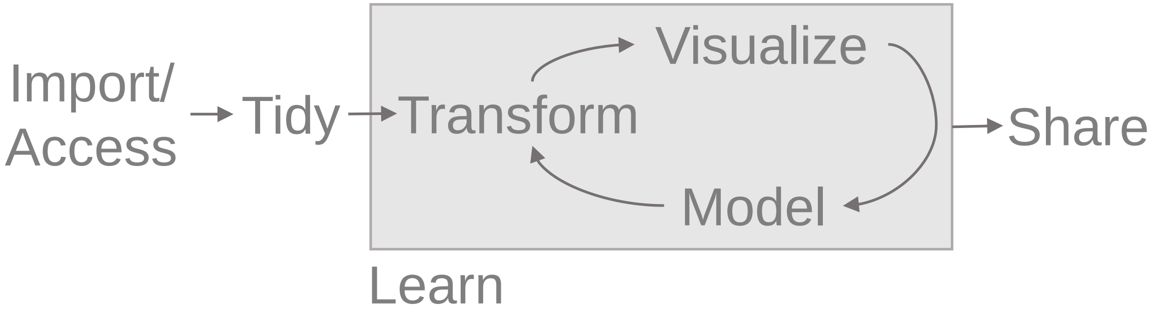
Production Pipelines

Photo by Iker Urteaga on Unsplash

Different projects, different deliverables

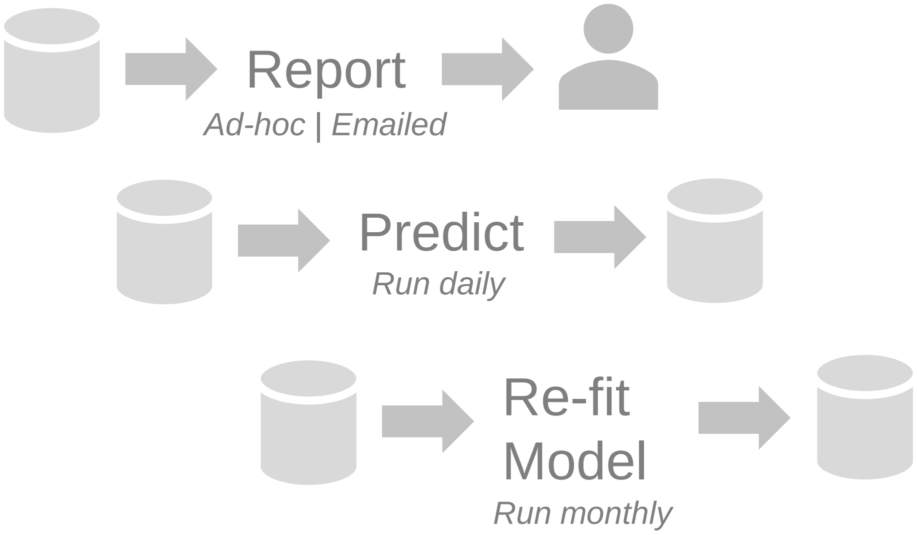
Data Science

- Deliverable: **Insights**
- Experimental
- Iterative



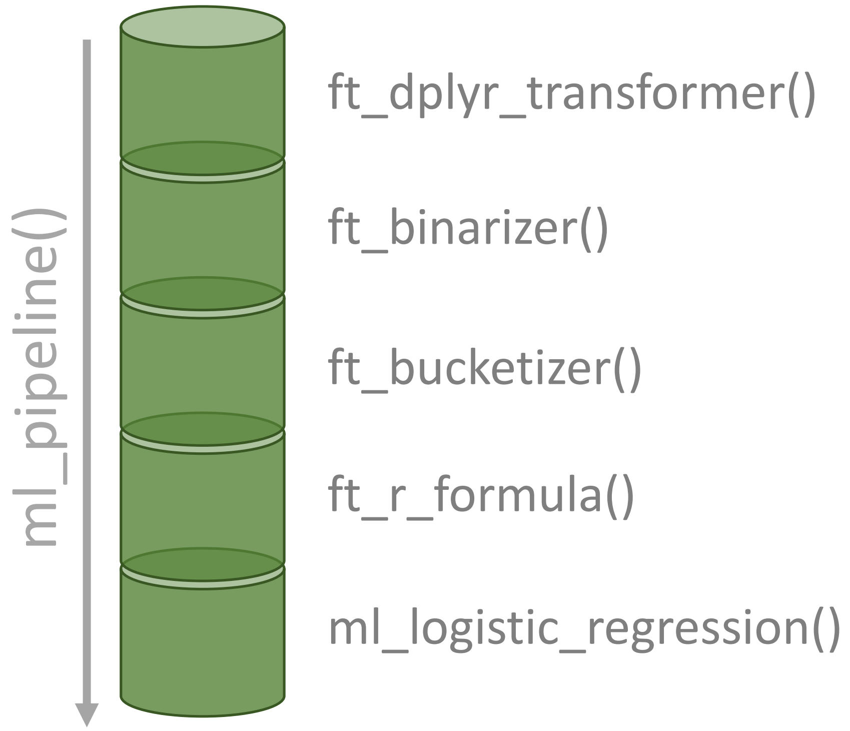
Production

- Deliverable: **Software**
- Tested
- Automated
- Apply SDLC



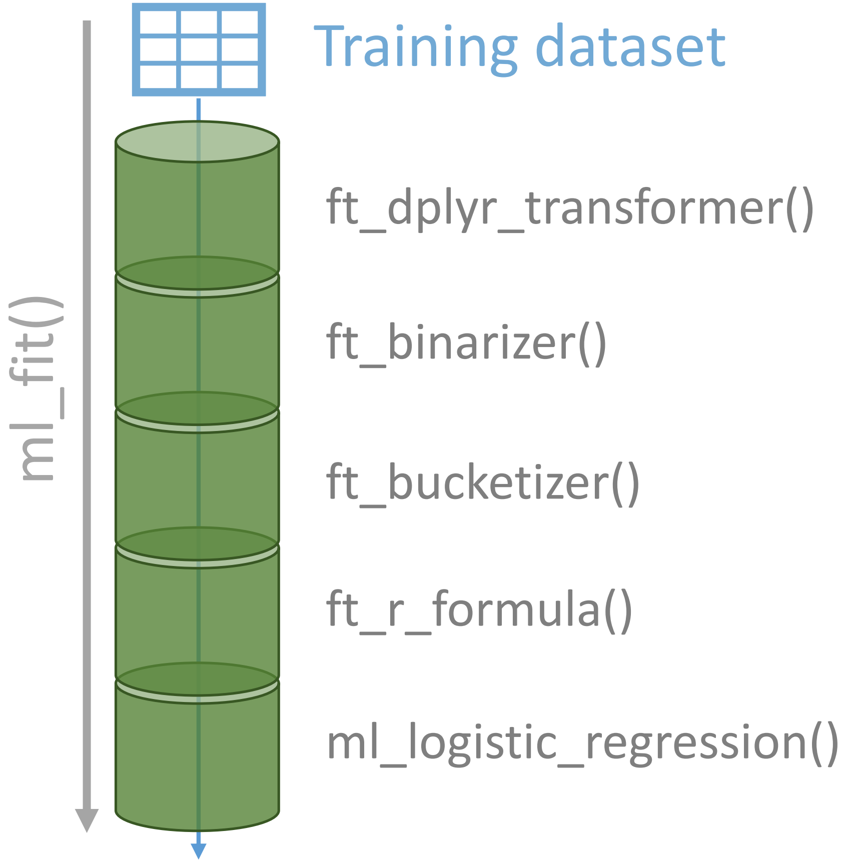
Spark pipelines types

Estimator (Plan)



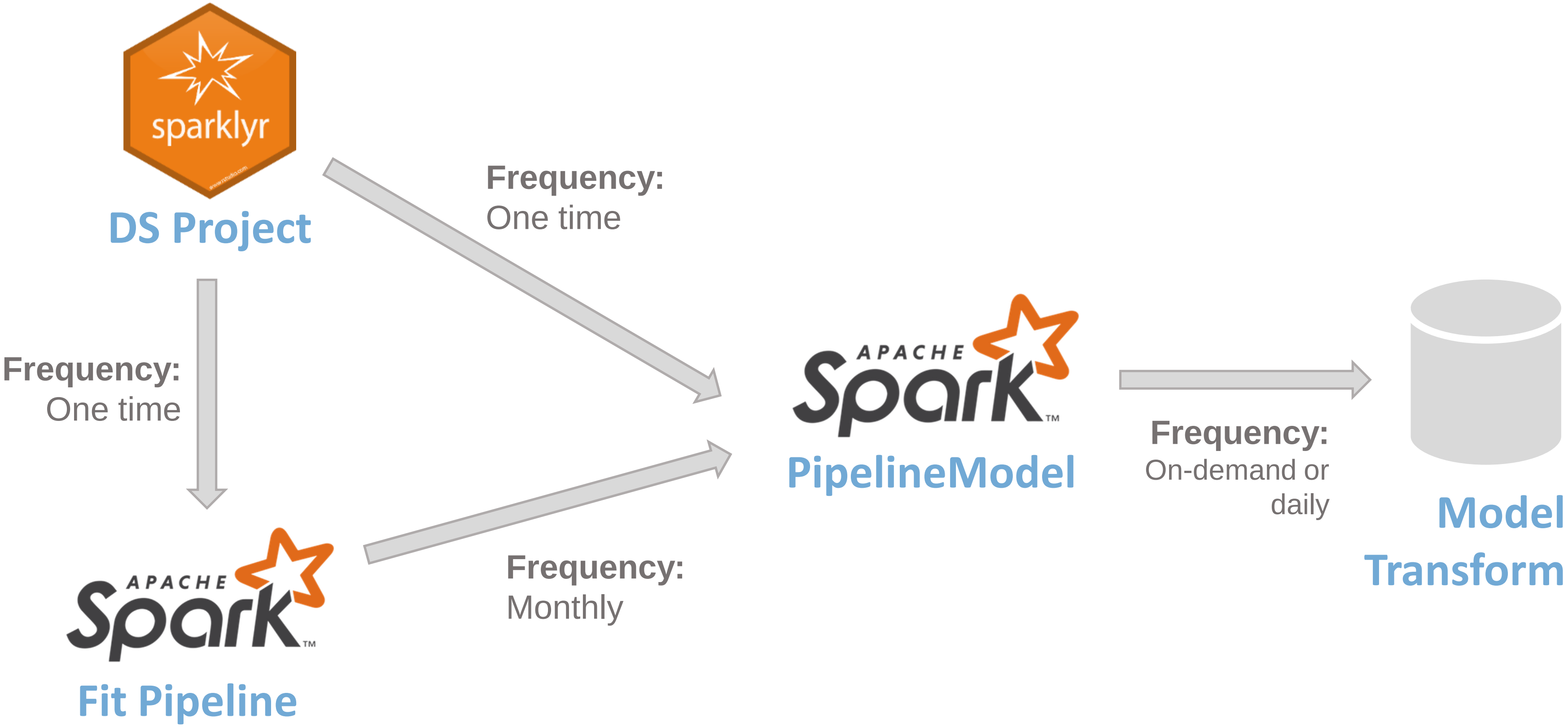
Spark Pipeline

Transformer (Fit)



Spark PipelineModel

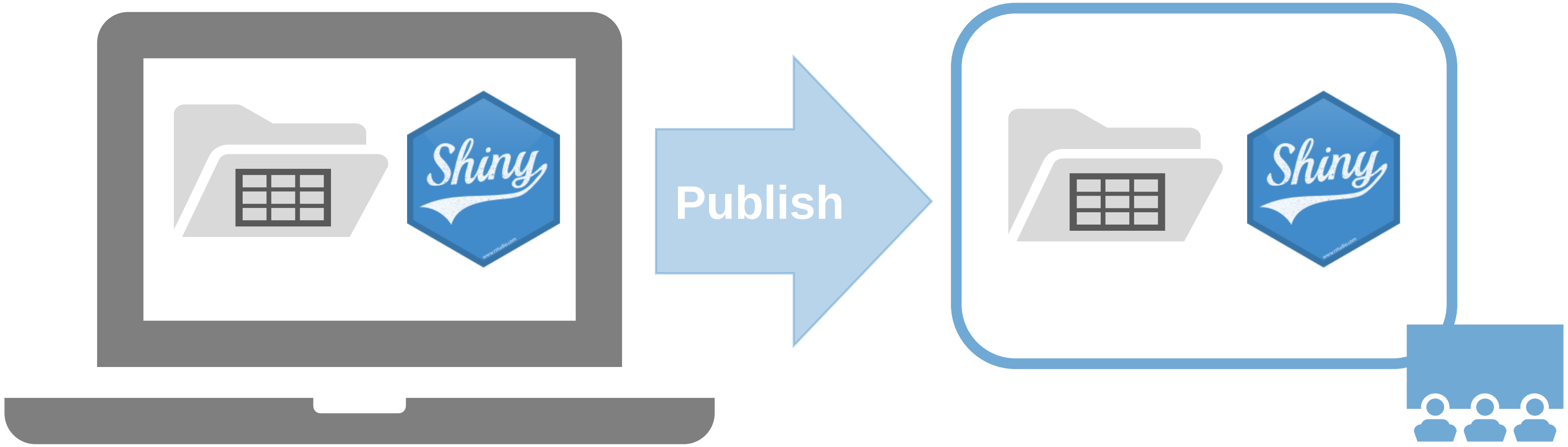
Production Implementation



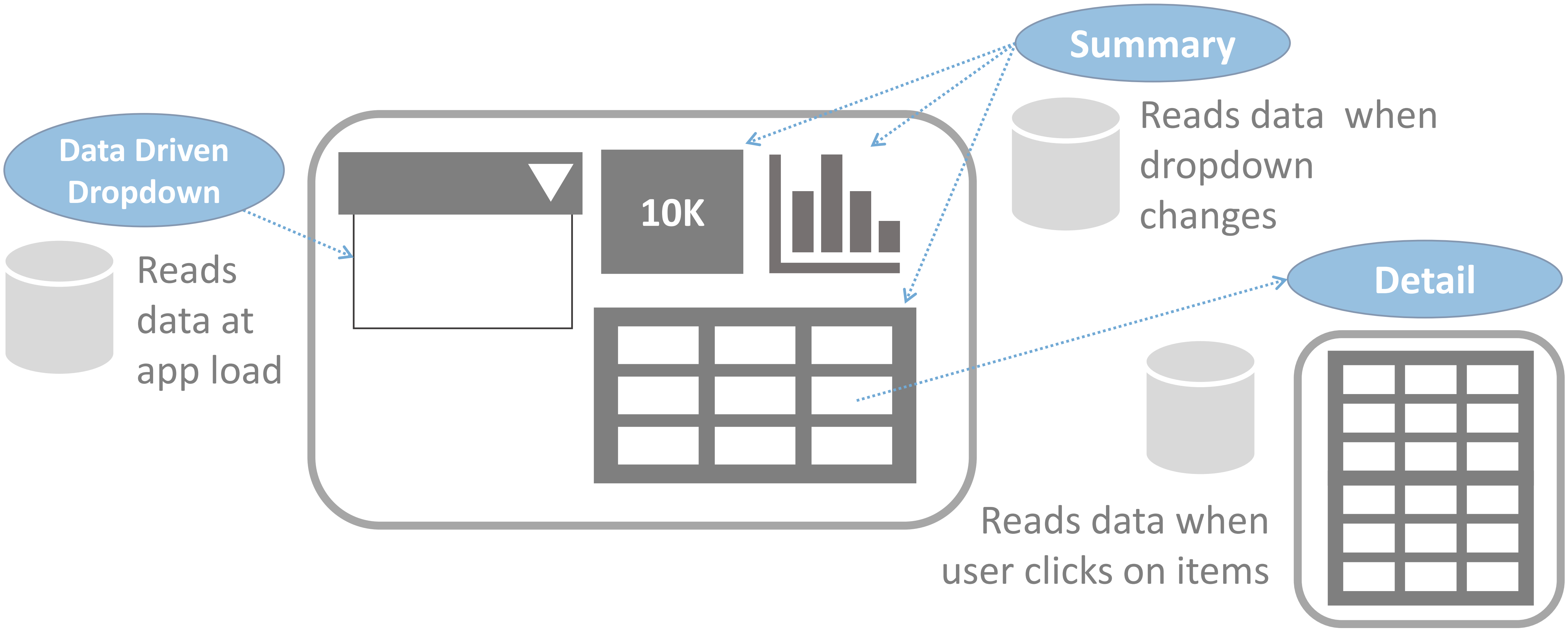
Dashboards

Photo by Benjamin Child on Unsplash

Normal Shiny app



Database + Shiny Dashboard



Summary

Reads data when dropdown changes

Data Driven Dropdown

Reads data at app load

Detail

Reads data when user clicks on items

RStudio

R Tools for Big Data

	Access	Wrangle	Plot	Model	Automate
	DBI odbc bigrquery rpostgres RMariaDB monetdblite	dplyr (*via* dbplyr) DBI corrr (*in dev*)	ggplot2 (*via* dbplot) corrr (*in dev*)	modeldb tidypredict	tidypredict
Spark	sparklyr	sparklyr		sparklyr graphframes rsparkling (H2O)	sparklyr mleap