RStudio®

You're Not Afraid of Big Data
...neither is R.

Alex K Gold
@alexkgold

rstd.io/big_data_19


Memory
NOOOOOoooOoOooo
EVERYTHING IS OK!
DON'T MATTER
NOOOOOOOOOO!!!!!!!!
Noooooooooo!
ded

In-Memory

heatheronormative
@heatherklus

I left this fact out of my talk bu
our R models are hit over 4 mi
next time someone tells you F
me #rstatsnyc

8:36 AM - 11 May 2019

88 Retweets 320 Likes

8 88 320

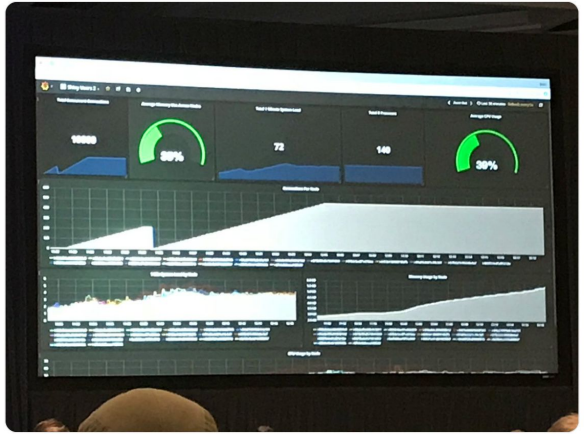
David Robinson
@drob

Following

.@lopp_sean lives dangerously- before his talk he started ramping up a test of 10,000 visitors to his Shiny app, and looked at the results live along with us

Verdict: Shiny scales!

#rstudioconf



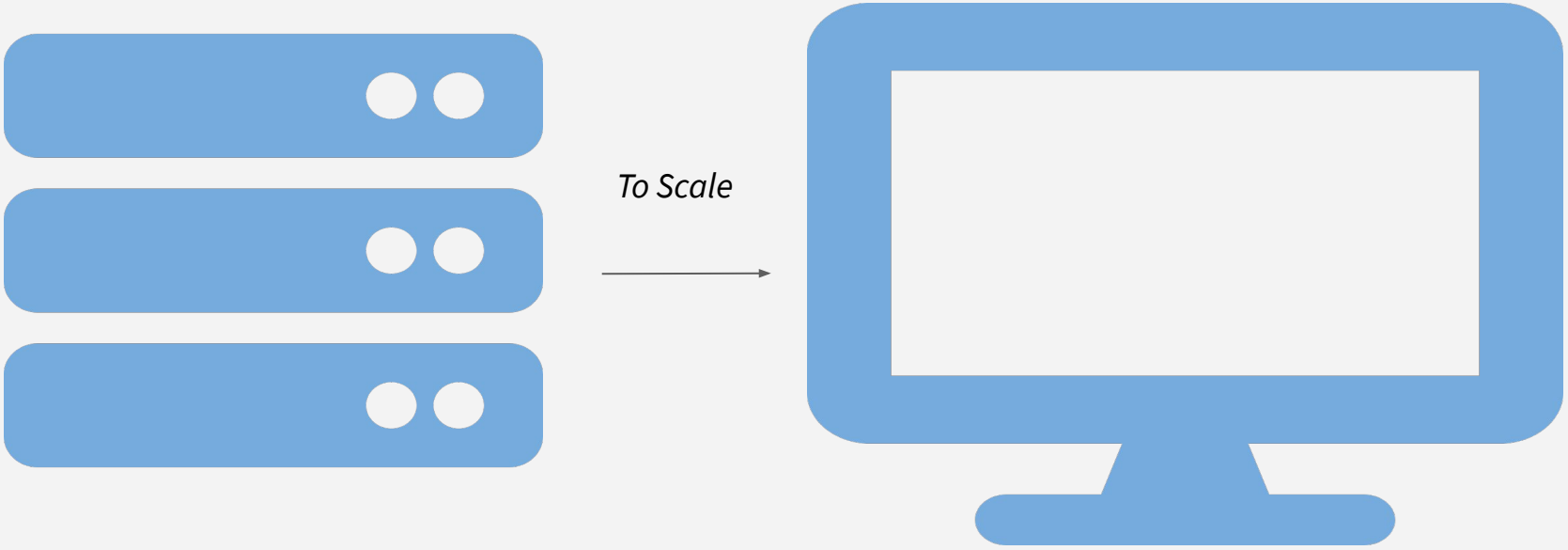
12:24 PM - 2 Feb 2018

Jakuczun · May 11
rklus
K/sec requests handled by #rstats with plumber on one
the slides here - bigdatatechwarsaw.eu/wp-content/upl…

22

- doFuture
- RStudio Server Pro Launcher
- st Enough (profvis)
- Rcpp

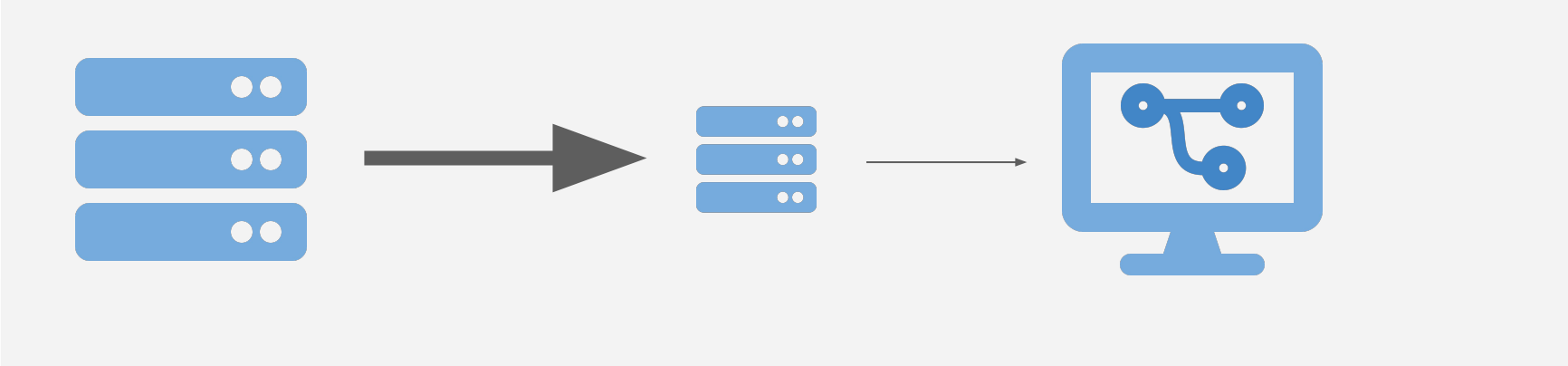
In-Memory



R Big Data Paradigms

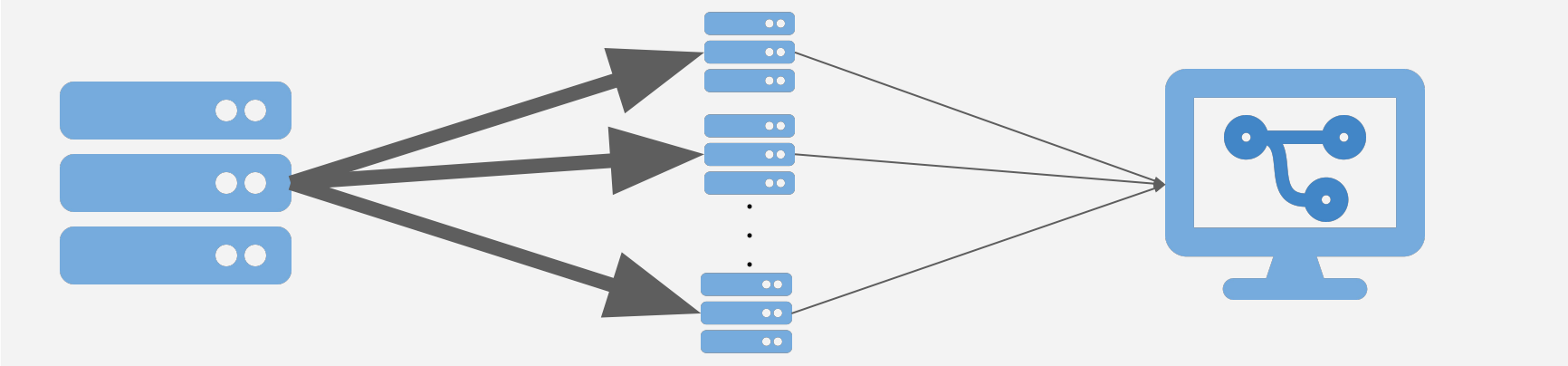
RStudio

Paradigm 1: Sample and Model



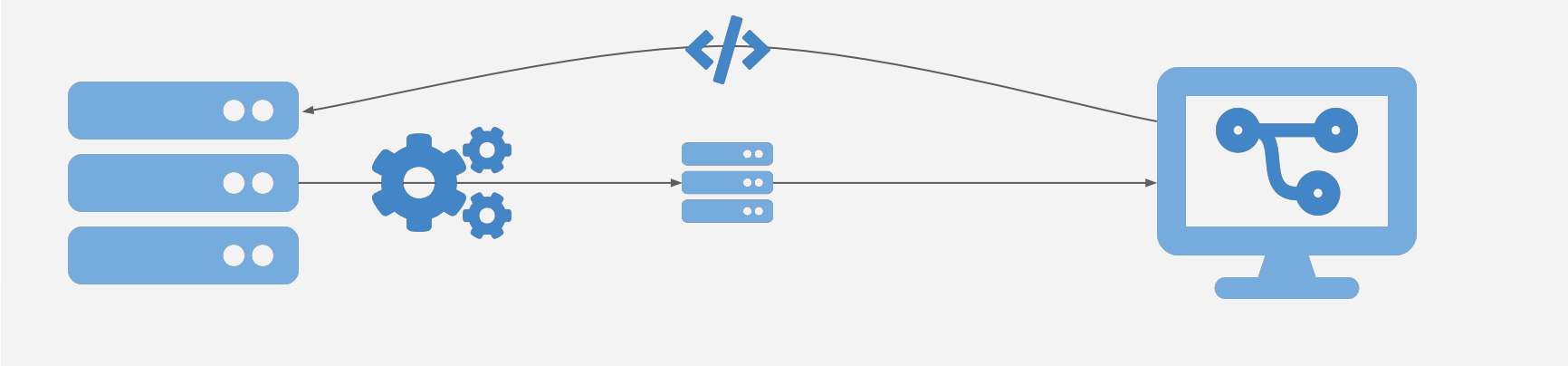
- 😃 Use favorite R modeling package (`Caret/Parsnip/rsample`).
- 😃 Really good for iterating/prototyping.
- ☹️ Requires care for sampling and scaling.
- ☹️ Not good for BI tasks.

Paradigm 2: Chunk and Pull



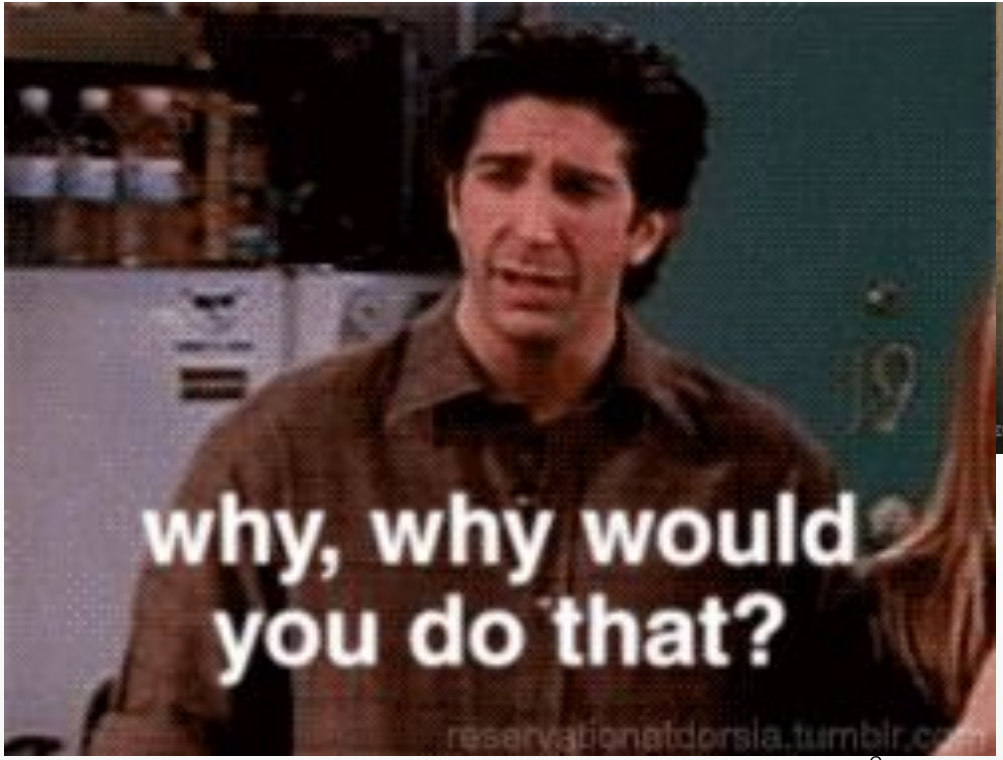
😃 Great when discrete chunks exist.
😃 Facilitates parallelization.
☹️ Can’t have interactions between chunks.
☹️ Eventually pull in all data.

Paradigm 3: Push Compute to Data

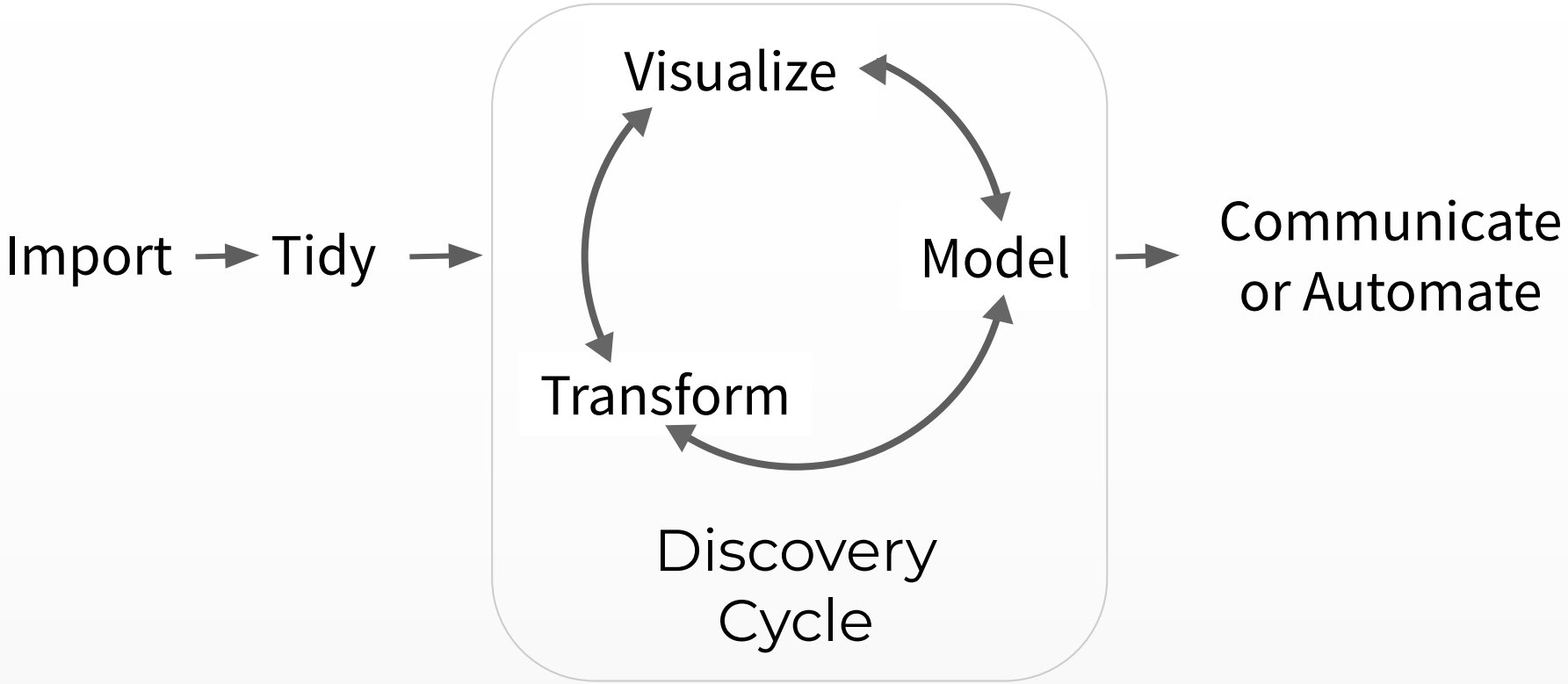


😃 Take advantage of database strengths.
😃 Get whole dataset, but move less data.
☹️ Operations might not be permitted in database.
☹️ Maybe your database is slow?

3 Big Data Paradigms for R



RStudio


Import
Tidy
Visualize
Model
Transform
Discovery Cycle
Communicate or Automate

Demo!

- ➔ Clean data using RMarkdown script
- ➔ Explore with Spark SQL
- ➔ Fit a (bad) model using Spark ML
- ➔ Visualize model quality (bad) with Shiny

RStudio

- General purpose distributed computation.
- APIs for Scala, Python, and Java, and ...

Otherwise…

Connect via:

- DBI Database connectors (github.com/r-dbi):
  - SQLite
  - PostGres
  - MariaDB
  - MySQL
  - Google BigQuery
  - ODBC

Process via

- `dbplyr` - run dplyr code in database
- `modeldb` - fit model in database
- `tidypredict` - predict in database
- `dbplot` - plot in database

What about deployment?

Open-Source (Free!)

- Build-your-own
- Shiny Server

Enterprise Products

- RStudio Connect
- Free 45 Day Eval
- Quickstart

RStudio

Recommendation Summary

Problem	Solution
Single-Threading	● Many R packages ○ My favorite: `doFuture` ● RStudio Server Pro Job Launcher
R is Slow	● Profile with `profvis` ● Write in a faster language, call from R (`Rcpp`)
In-Memory Data	● Adopt a big data paradigm for R 1. Sample and Model 2. Chunk and Pull 3. Push Compute to Data

db.rstudio.com
spark.rstudio.com

@alexkgold
`rstd.io/big_data_19`

RStudio