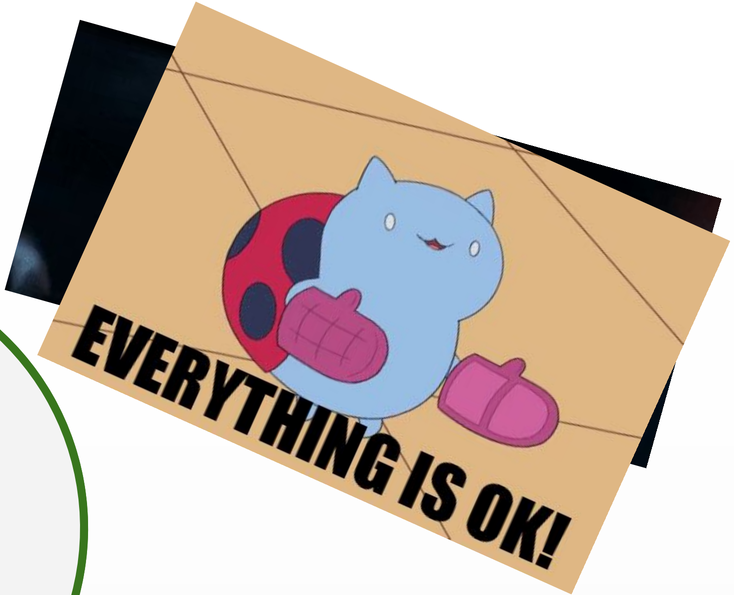
R Studio®

You Love Big Data! So Does R.

Alex K Gold
@alexkgold
#rstats

https://github.com/akgold/bdl_2019

In Memory



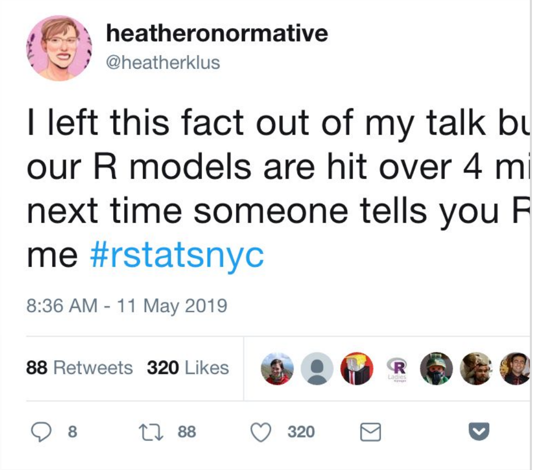
DON'T MATTER

ded

@alexkgold

RStudio

In-Memory

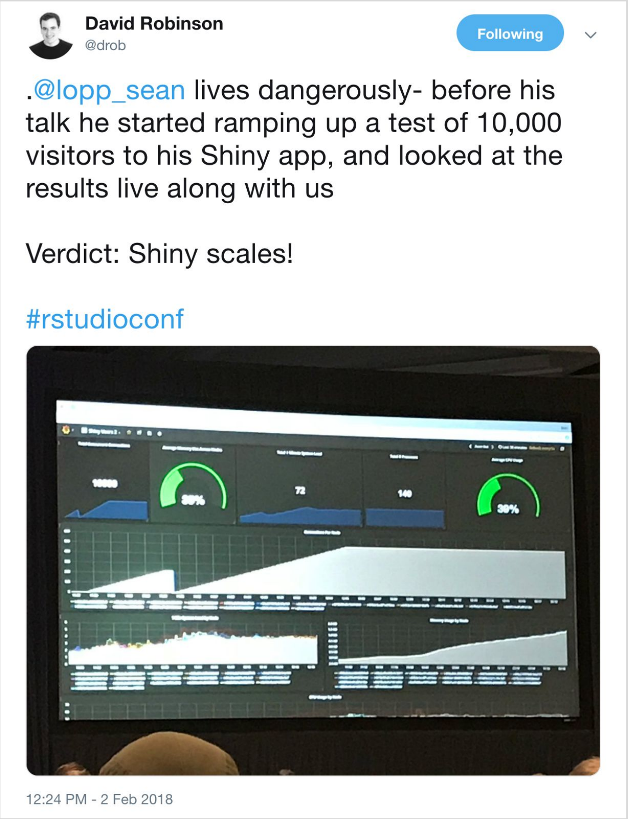


heatheronormative
@heatherklus

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

8:36 AM - 11 May 2019

88 Retweets **320** Likes



David Robinson
@drob

.@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

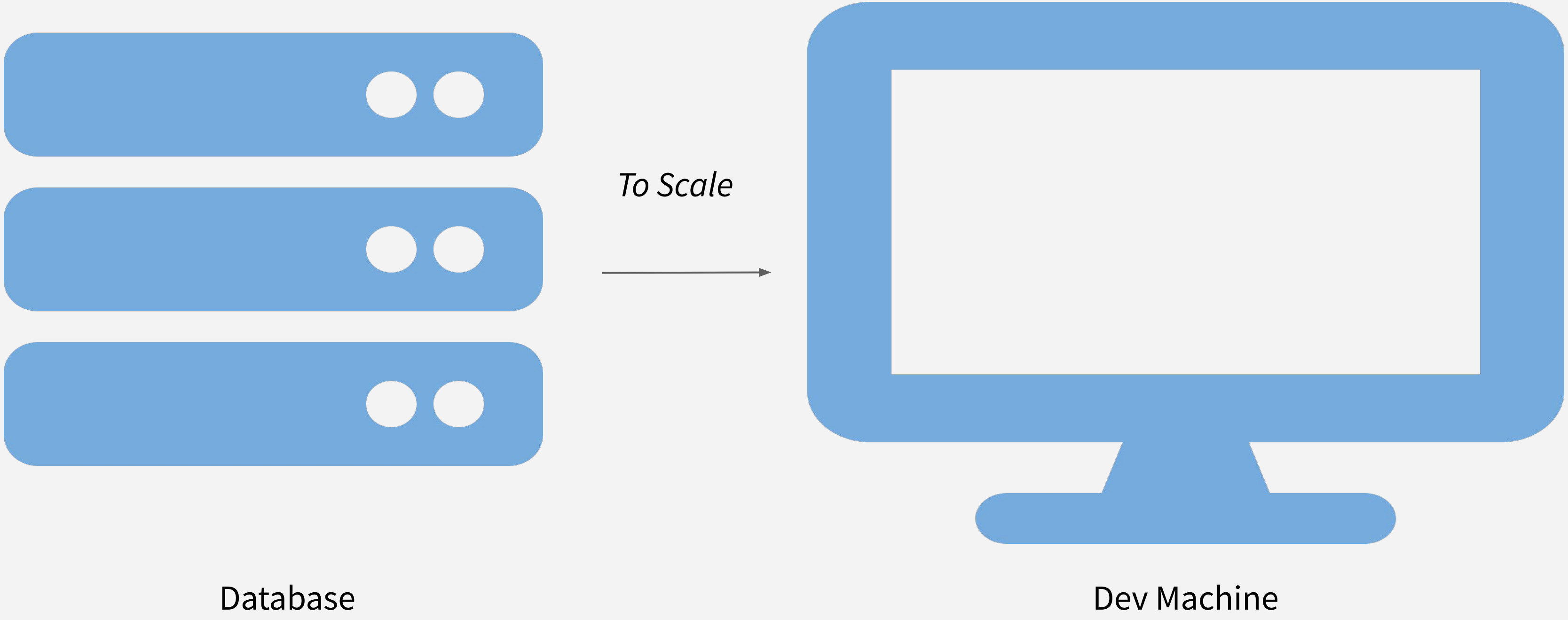
K/sec requests handled by #rstats with plumber on one

the slides here - bigdatatechwarsaw.eu/wp-content/upl…

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

@alexkgold

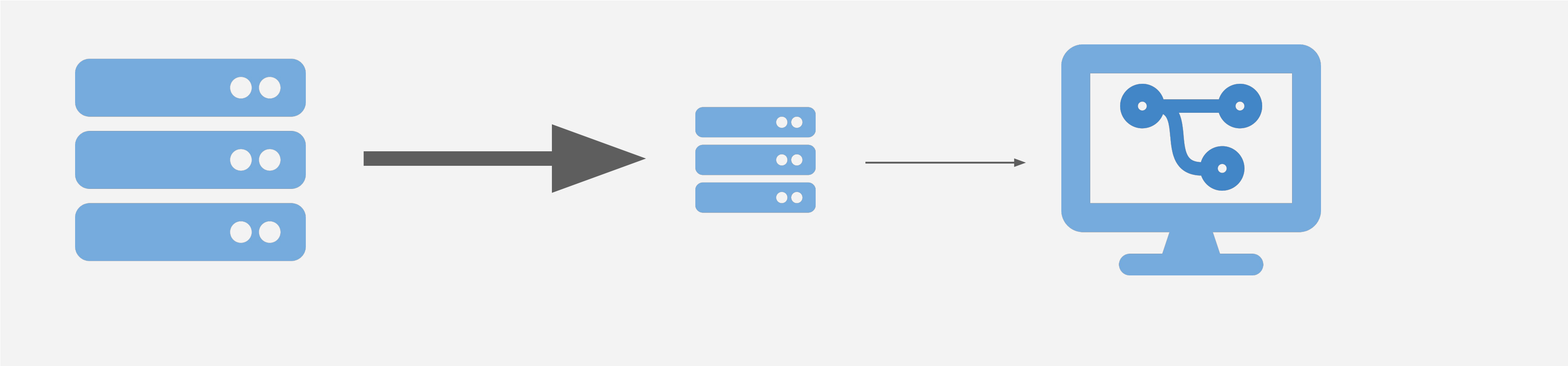
In-Memory



Big Data Strategies for R

@alexkgold

Strategy 1: Sample and Model



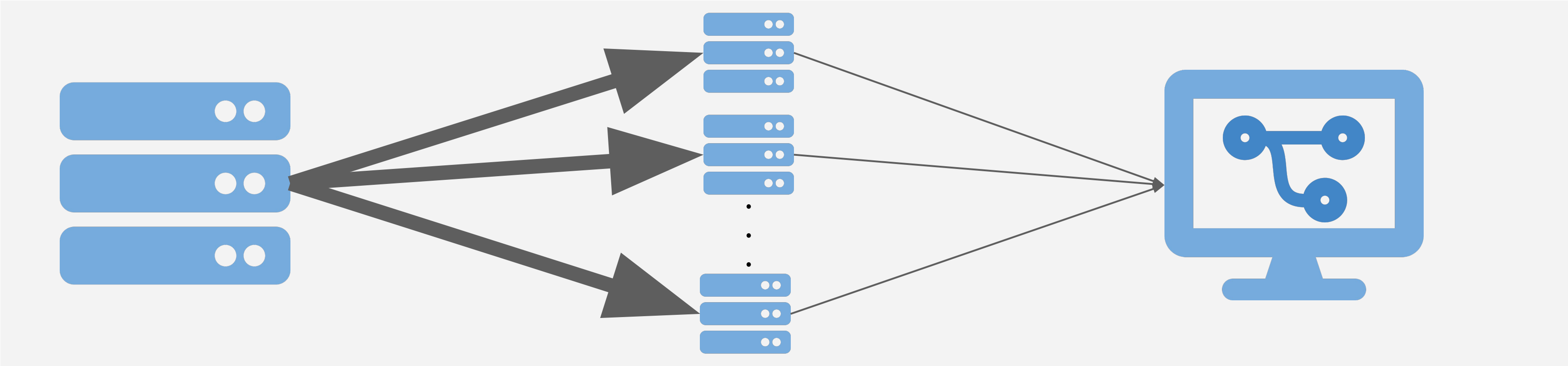
- 😃 Use favorite R modeling package (`Caret/Parsnip/rsample`).
- 😃 Really good for iterating/prototyping.
- ☹️ Requires care for sampling and scaling.
- ☹️ Not good for BI tasks.

@alexkgold

RStudio

Demo!

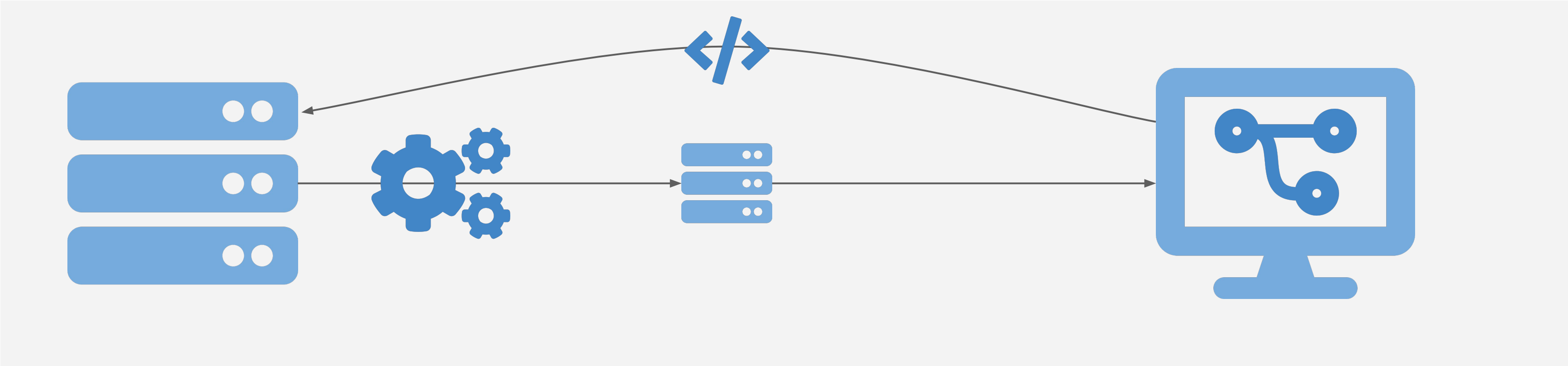
Strategy 2: Chunk and Pull



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

Demo!

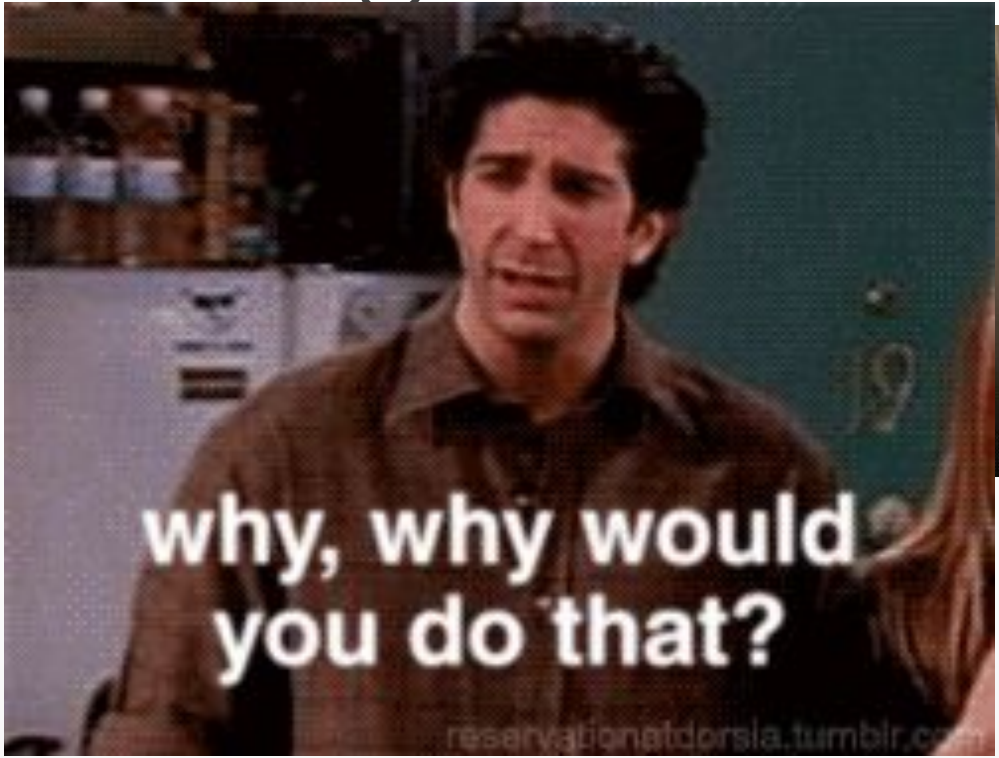
Strategy 3: Push Compute to Data



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

Demo!

3 Big Data Strategies for R



Import → Tidy —

Communicate
or Automate

What about deployment?

Open-Source (Free!)

- Build-your-own
- Shiny Server

Enterprise Products

- RStudio Connect
- Evals

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
therinspark.com

@alexkgold
https://github.com/akgold/bdl_2019

RStudio