DuckDB

d plyr

Kirill Müller, posit::conf(2025), cynkra

posit™

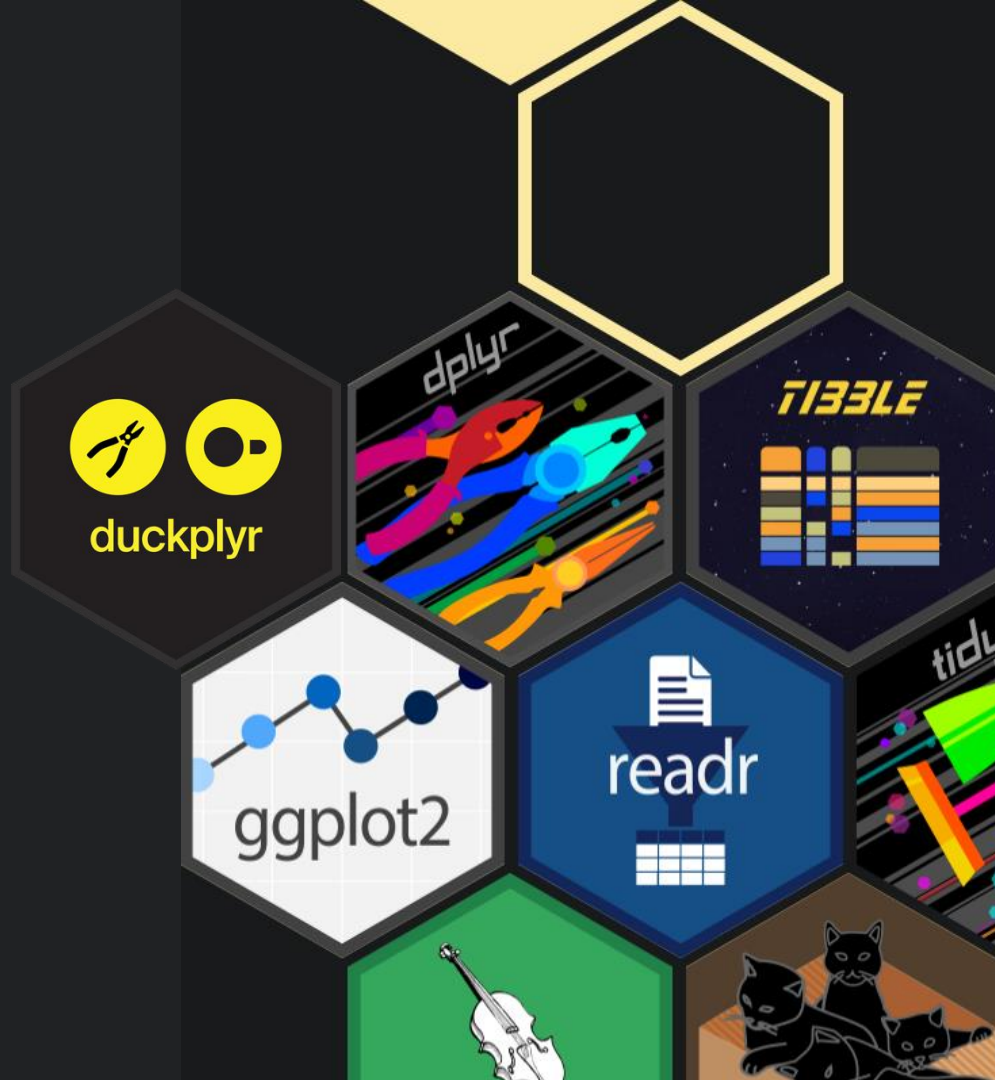
DuckDB

- Embedded analytical database
- Command line interface
- Python/Rust/WASM/… package
- R package

Large

Fast

Featureful


duckplyr
dplyr
TIBBLE
ggplot2
readr

DuckDB interfaces

dplyr-like		
SQL	DBI 	 DuckDB

```
SELECT COUNT(*)
FROM 'personas.parquet';
--
--   count_star()
--      int64
--
--      100000
--
```

```r
library(DBI)
con <- dbConnect(duckdb::duckdb())

tbl <- duckdb::tbl_file(
  con,
  "personas.parquet")
tbl |> dplyr::count()
#> # Source:   SQL [?? x 1]
#> # Database: DuckDB 1.4.0
#>         n
#>     <dbl>
#> 1 100000

dbGetQuery(
  con,
  "SELECT COUNT(*)
   FROM 'personas.parquet'")
#>   count_star()
#> 1       1e+05
```

```sql
SELECT COUNT(*)
FROM 'personas.parquet';
--
--   count_star()
--      int64
--
--     100000
--
```

```
library(DBI)
con <- dbConnect(duckdb::duckdb())

tbl <- duckdb::tbl_file(
  con,
  "personas.parquet")
tbl |> dplyr::count()
#> # Source:   SQL [?? x 1]
#> # Database: DuckDB 1.4.0
#>         n
#>     <dbl>
#> 1 100000

dbGetQuery(
  con,
  "SELECT COUNT(*)
   FROM 'personas.parquet'")
#>   count_star()
#> 1       1e+05
```

```
tbl <- duckplyr::read_parquet_duckdb(
  "personas.parquet")
tbl |> dplyr::count()
#> # A duckplyr data frame: 1 variable
#>         n
#>     <int>
#> 1 100000
```

```
SELECT COUNT(*)
FROM 'personas.parquet';
--
--  ┌──────────────┐
--  │ count_star() │
--  │    int64     │
--  ├──────────────┤
--  │    100000    │
--  └──────────────┘
--
```

```r
library(duckplyr)

personas <- read_parquet_duckdb(
  "hf://datasets/.../*/*.parquet"
)

personas_count <-
  personas |>
  count(sex, education_level)

personas_count |>
  explain()
```

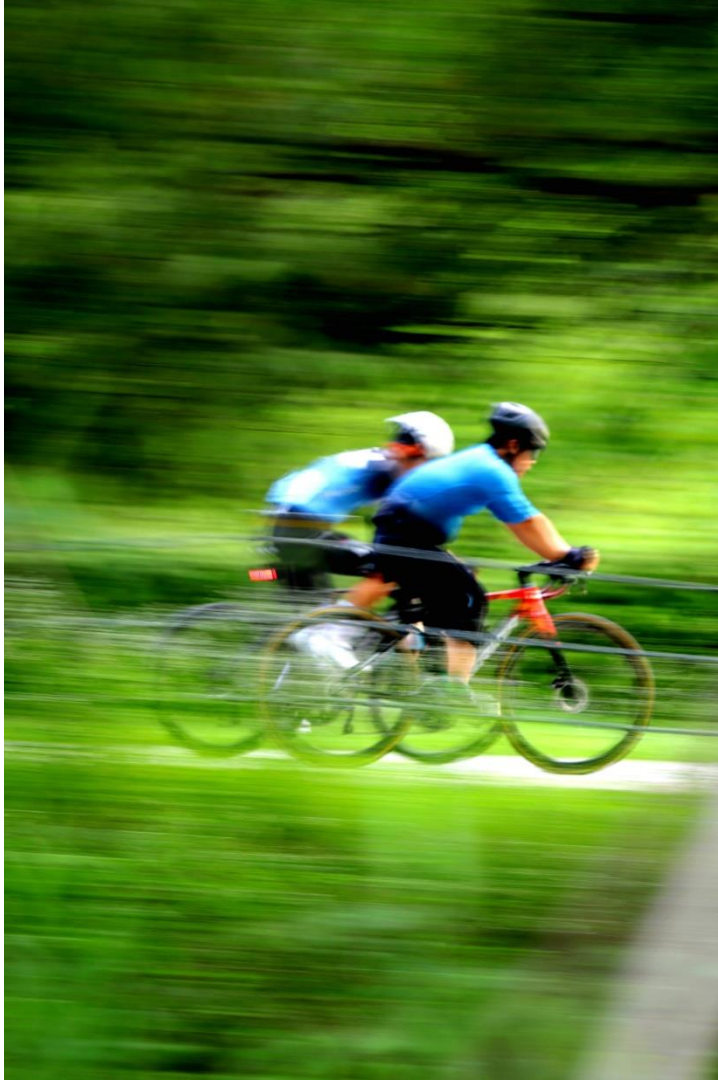
Large data

- Peek with remote access
- Prudence avoids crashes
- Iterate with local file
- Production: remote

```r
library(tidyverse)



G <- 200
N <- 5000

data <-
  crossing(g1 = 1:G, g2 = 1:G, id = 1:N) |>
  mutate(x = rnorm(n()))

data |>

  summarize(.by = c(g1, g2), m = mean(x)) |>

  system.time()
#>    user  system elapsed
#>   2.397   1.720   5.172
```

Existing pipelines

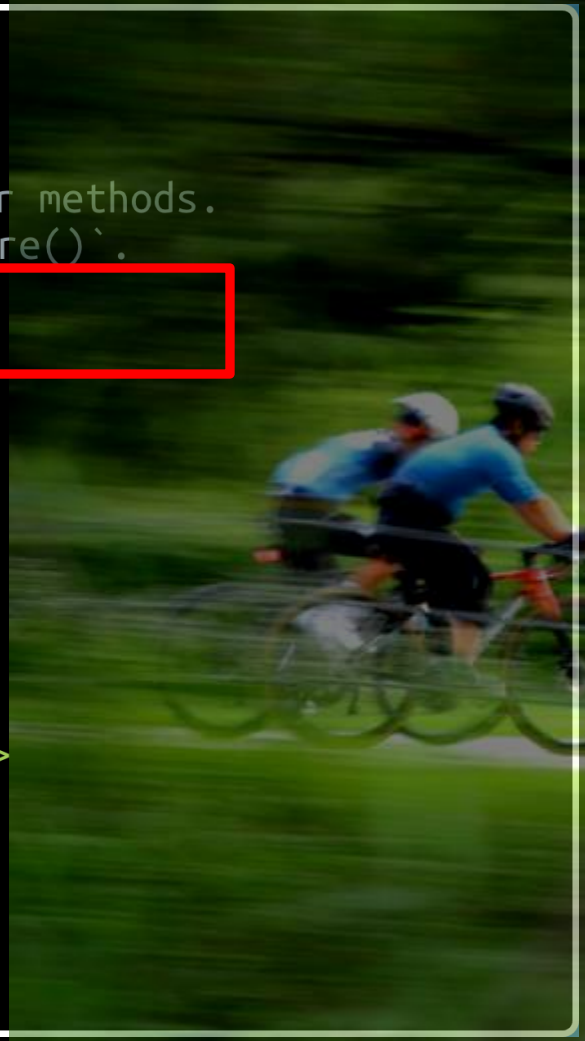
- All in
- Best effort
- Targeted



```r
library(tidyverse)
Sys.setenv(DUCKPLYR_FORCE = TRUE)



G <- 200
N <- 5000

data <-
  crossing(g1 = 1:G, g2 = 1:G, id = 1:N) |>
  mutate(x = rnorm(n()))
#> Error in `mutate()`:
#> ! Can't translate function `rnorm()`.
```

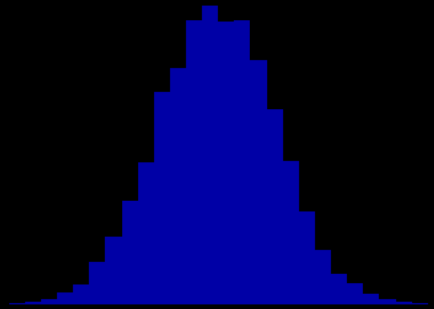
```
library(tidyverse)
library(duckplyr)
#> ✔ Overwriting dplyr methods with duckplyr methods.
#> ℹ Turn off with `duckplyr::methods_restore()`.

G <- 200
N <- 5000

data <-
  crossing(g1 = 1:G, g2 = 1:G, id = 1:N) |>
  mutate(x = rnorm(n()))

data |>

  summarize(.by = c(g1, g2), m = mean(x)) |>
  collect() |>
  system.time()
#>    user  system elapsed
#>   4.079   0.016   0.600
```

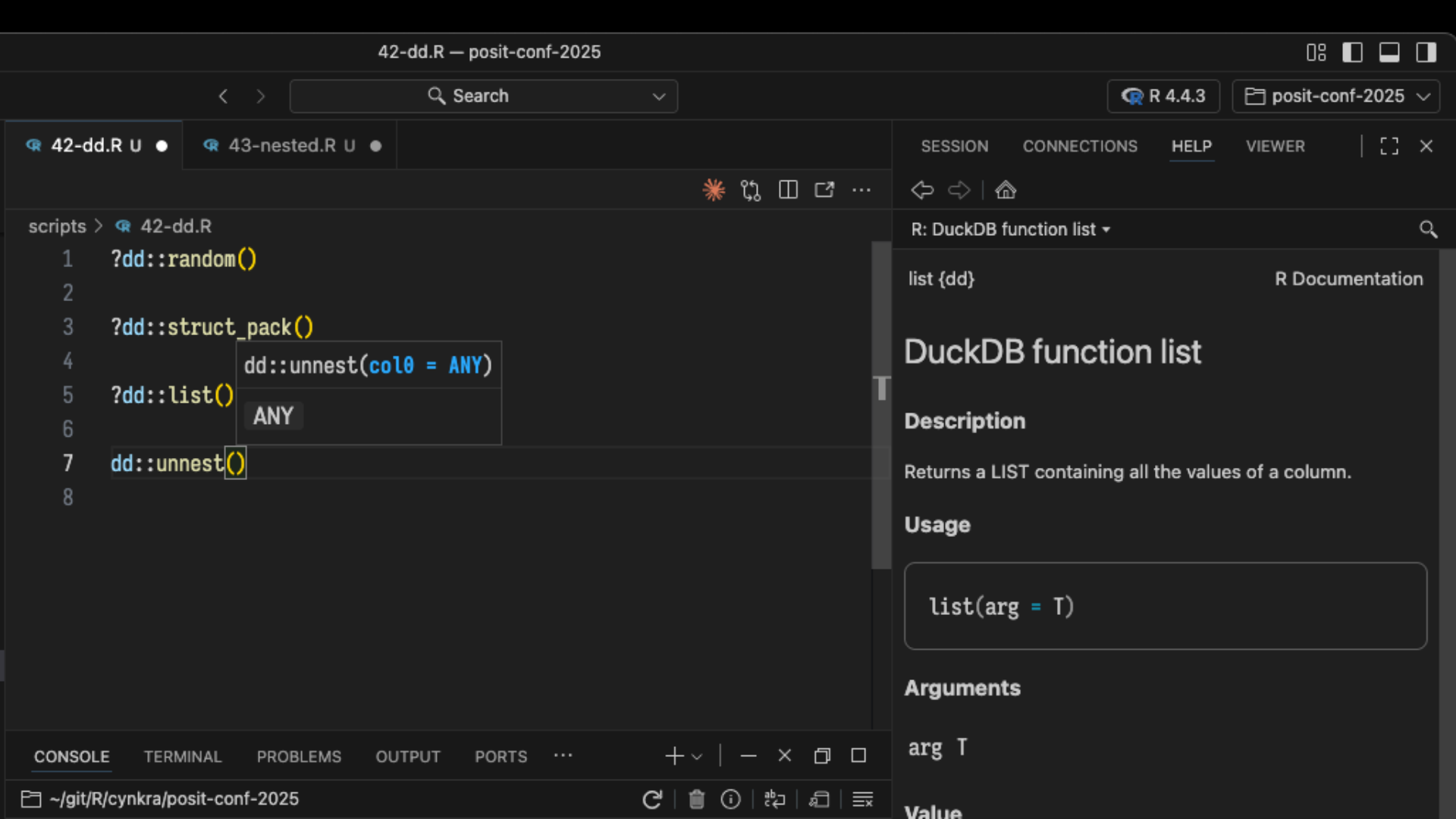
```
library(tidyverse)
library(duckplyr)
#> ✔ Overwriting dplyr methods with duckplyr methods.
#> ℹ Turn off with `duckplyr::methods_restore()`.
methods_restore()
#> ℹ Restoring dplyr methods.


data <-
  crossing(g1 = 1:G, g2 = 1:G, id = 1:N) |>
  mutate(x = rnorm(n()))

data |>
  as_duckdb_tibble(prudence = "stingy") |>
  summarize(.by = c(g1, g2), m = mean(x)) |>
  collect() |>
  system.time()
#>    user  system elapsed
#>   4.142   0.017   0.659
```



```r
library(tidyverse)

idx <- duckplyr::duckdb_tibble(id = 1:1e4)



idx |>
  mutate(x = dd::random()) |>
  ggplot(aes(x)) +
  geom_histogram()
```

```r
library(tidyverse)

idx <- duckplyr::duckdb_tibble(id = 1:1e4)

duckplyr::db_exec("LOAD stochastic")

idx |>
  mutate(x = dd::dist_normal_sample(0, 1)) |>
  ggplot(aes(x)) +
  geom_histogram()
```




42-dd.R — posit-conf-2025
Search
R 4.4.3
posit-conf-2025
42-dd.R U
43-nested.R U
SESSION
CONNECTIONS
HELP
VIEWER
scripts > 42-dd.R
?dd::random()
?dd::struct_pack()
dd::unnest(col0 = ANY)
?dd::list()
ANY
dd::unnest()
R: DuckDB function list
list {dd}
R Documentation
DuckDB function list
Description
Returns a LIST containing all the values of a column.
Usage
list(arg = T)
Arguments
arg T
Value
CONSOLE
TERMINAL
PROBLEMS
OUTPUT
PORTS
~/git/R/cynkra/posit-conf-2025

```
install.packages("duckplyr")
library(duckplyr)
methods_restore()
```

cynkra.

tinyurl.com/
duckplyr-25