CWI

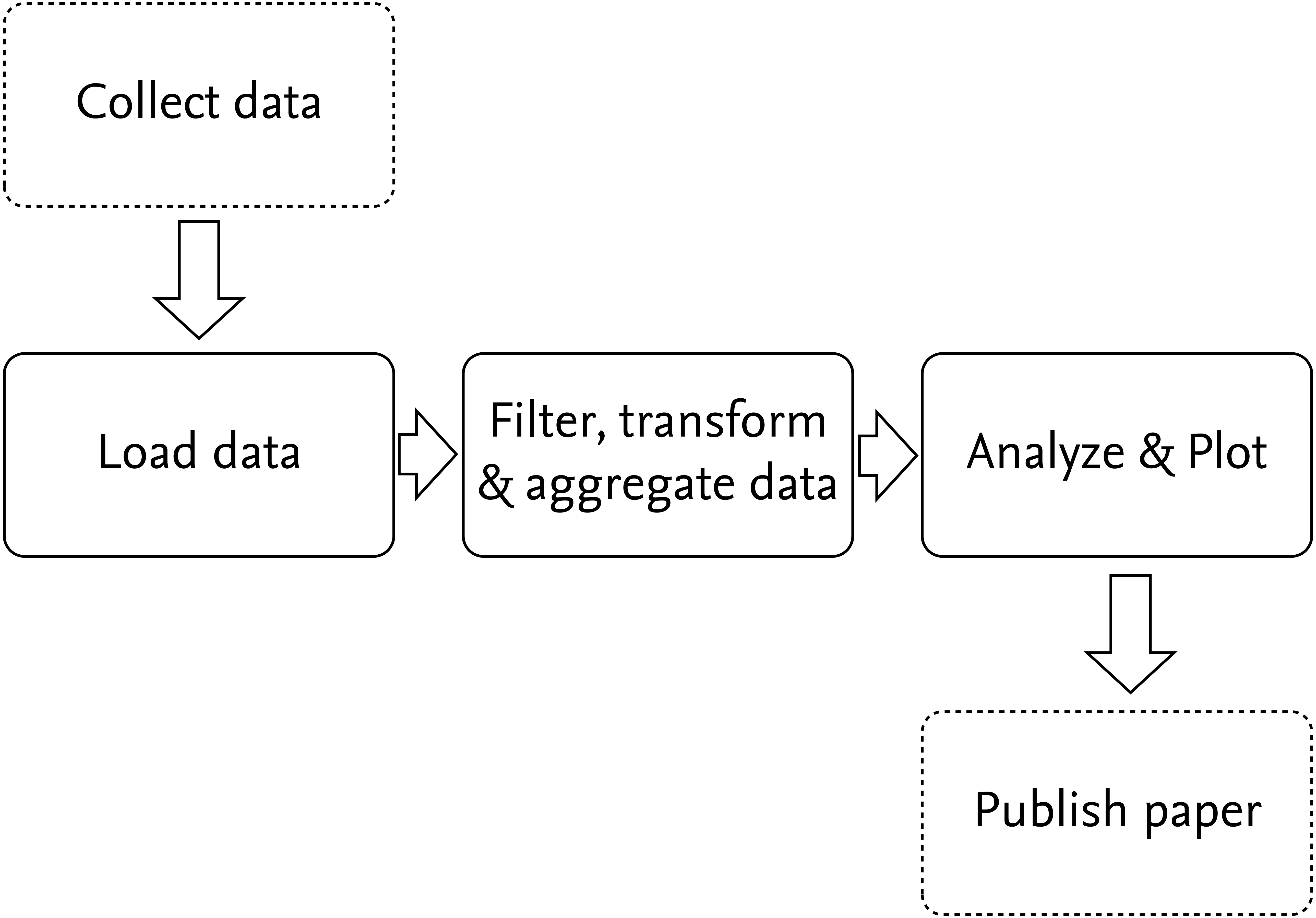
Centrum Wiskunde & Informatica

# Best of Both Worlds

## Relational Databases and Statistics

COMMIT/

Hannes Mühleisen & Thomas Lumley



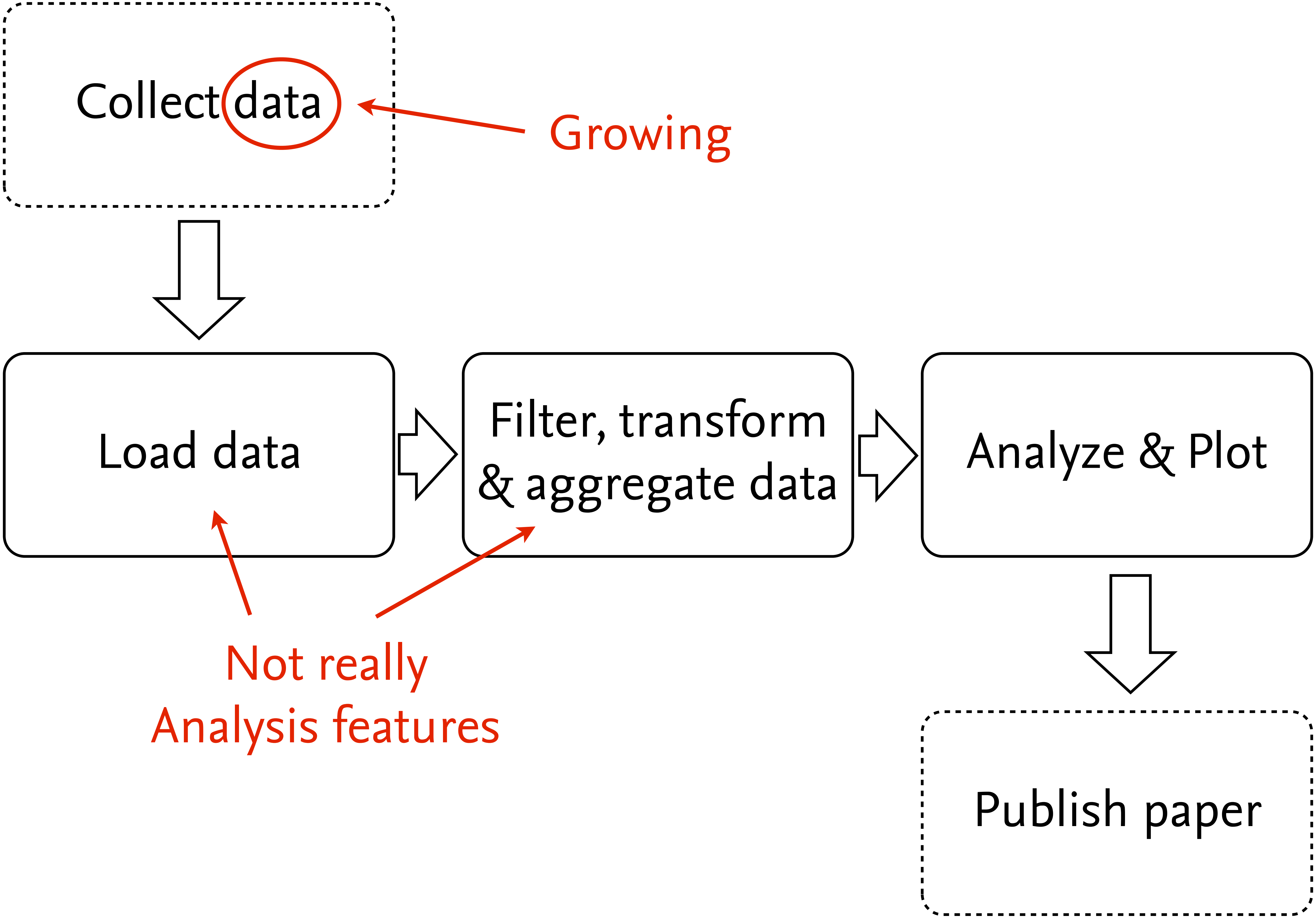
Collect data

Load data

Filter, transform & aggregate data

Analyze & Plot

Publish paper



Collect data

Growing

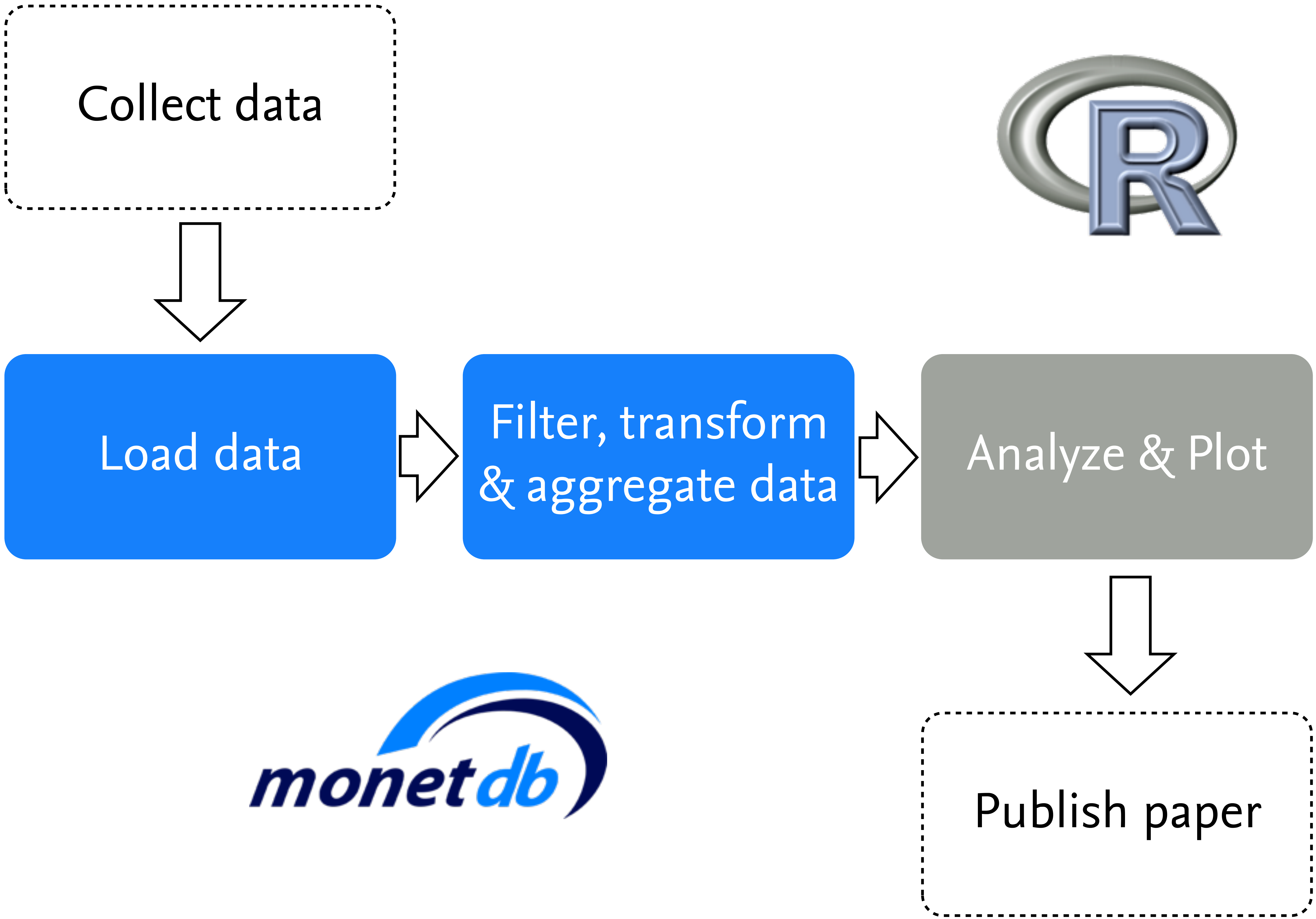
Load data

Filter, transform & aggregate data

Analyze & Plot

Not really Analysis features

Publish paper



Collect data

Load data

Filter, transform & aggregate data

Analyze & Plot

Publish paper

# But...

```
data <- dbGetQuery(conn,"
   SELECT t1,COUNT(t1) AS ct FROM (
      SELECT CAST(flux as integer) AS t1 FROM starships WHERE
         ( (speed = 5) ) AND ( (class = 'NX') ) ) AS t
   WHERE t1 > 0 GROUP BY t1 ORDER BY t1 LIMIT 100;
")
normalized <- data$ct/sum(data$ct)
```

...do we really want this?

# Enter monet.frame

The virtual data object for R

```
> data <- monet.frame(conn,"starships")
> nxw5 <- subset(data,class=="NX" & speed==5)$flux
> t <- tabulate(nxw5,100)
> normalized <- t/sum(t)
```

R-style data manipulation & aggregation

# Meanwhile

Behind the scenes:

```
> data <- monet.frame(conn,"starships")
SELECT * FROM starships;

> nxw5 <- subset(data,class=="NX" & speed==5)$flux
SELECT * FROM starships WHERE class = 'NX' AND speed = 5;
SELECT flux FROM starships WHERE class = 'NX' AND speed = 5;

> t <- tabulate(nxw5,100)
SELECT t1,COUNT(t1) AS ct FROM (SELECT CAST(flux as integer) AS
t1 FROM starships WHERE class = 'NX' AND speed = 5) AS t WHERE
t1 > 0 GROUP BY t1 ORDER BY t1 LIMIT 100;
```

Actually executed

# Implementation

```
# R core
subset <- function(x, ...) UseMethod("subset")

# MonetDB.R
unique.monet.frame <- function (x, subset, ...) {
  # some code here
}

> nxw5 <- subset(data,class=="NX" & speed==5)$flux
> str(nxw5)
MonetDB-backed data.frame surrogate
1 column, 1799991 rows
Query: SELECT flux FROM starships
  WHERE ( ((class = 'NX') AND (speed = 5)) )
Columns: flux (numeric)
```

# Optimization

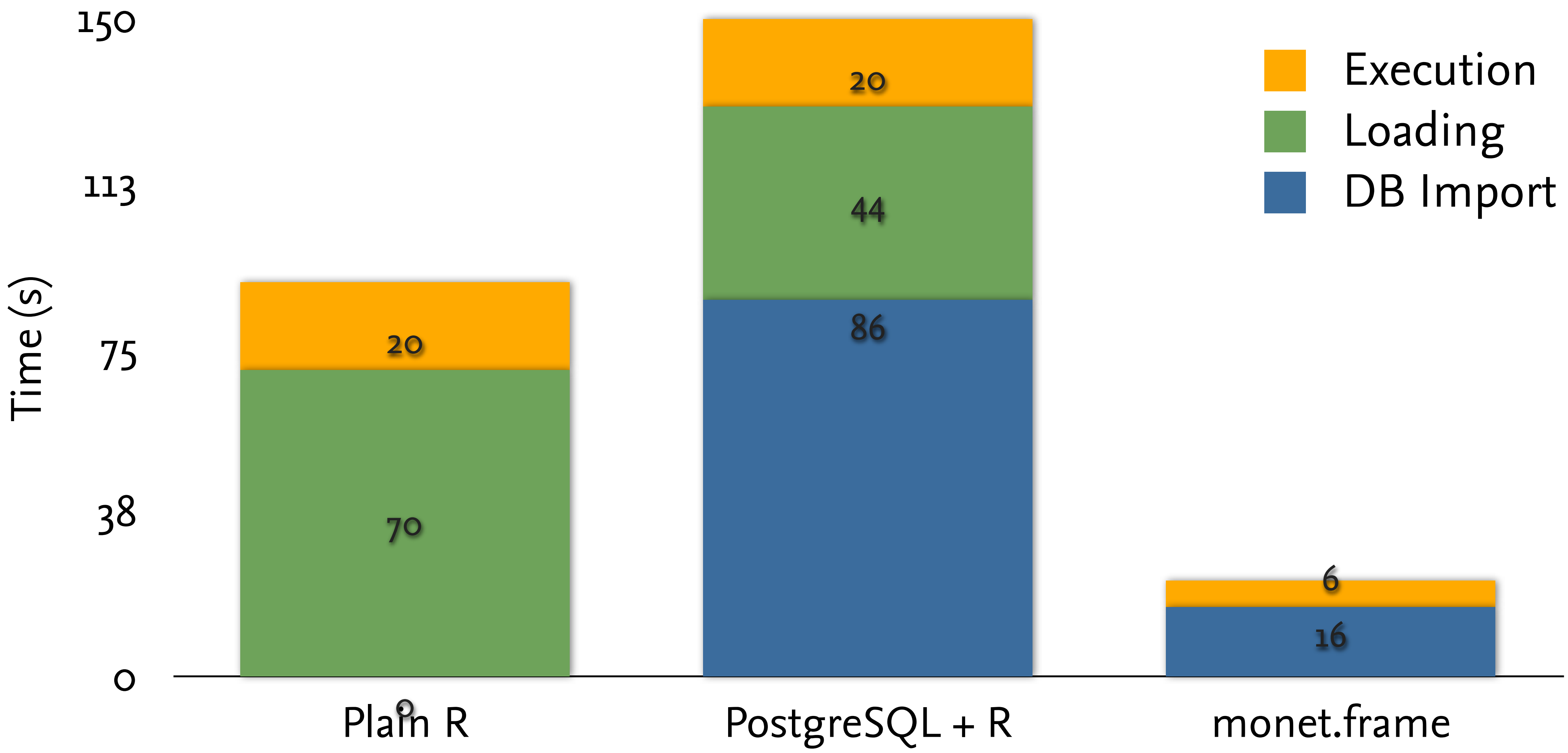
- Result Set Structure Inference
  - Columns, Types
  - # Rows

done

- Process embedding
  - Run DB inside R process

soon

# Performance



~1GB CSV, ~70M Rows

sd() ^ trunc() sign() merge() sqrt()

range()

log() tabulate() floor()

subset() str() ceiling()

exp() + sort() $ * []

/ na.omit()

tail()

# Thank You!

sin() range()

summary()

Questions?

head()

sample()

min() quantile()

abs() max() sum() -

length()

round() names() dim() ==

signif()

aggregate() print() var()

CRAN: MonetDB.R