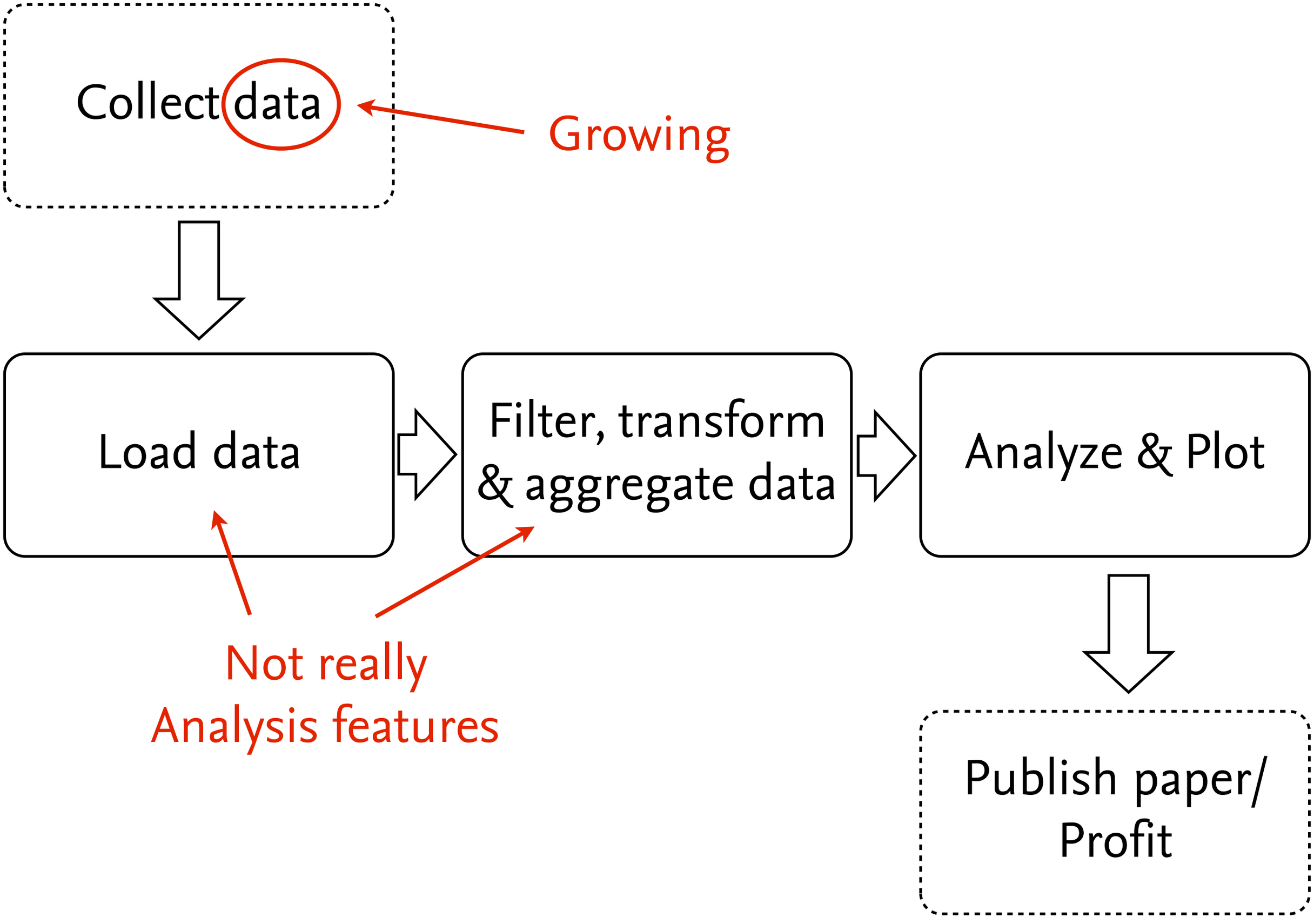
Centrum Wiskunde & Informatica

MonetDB & R

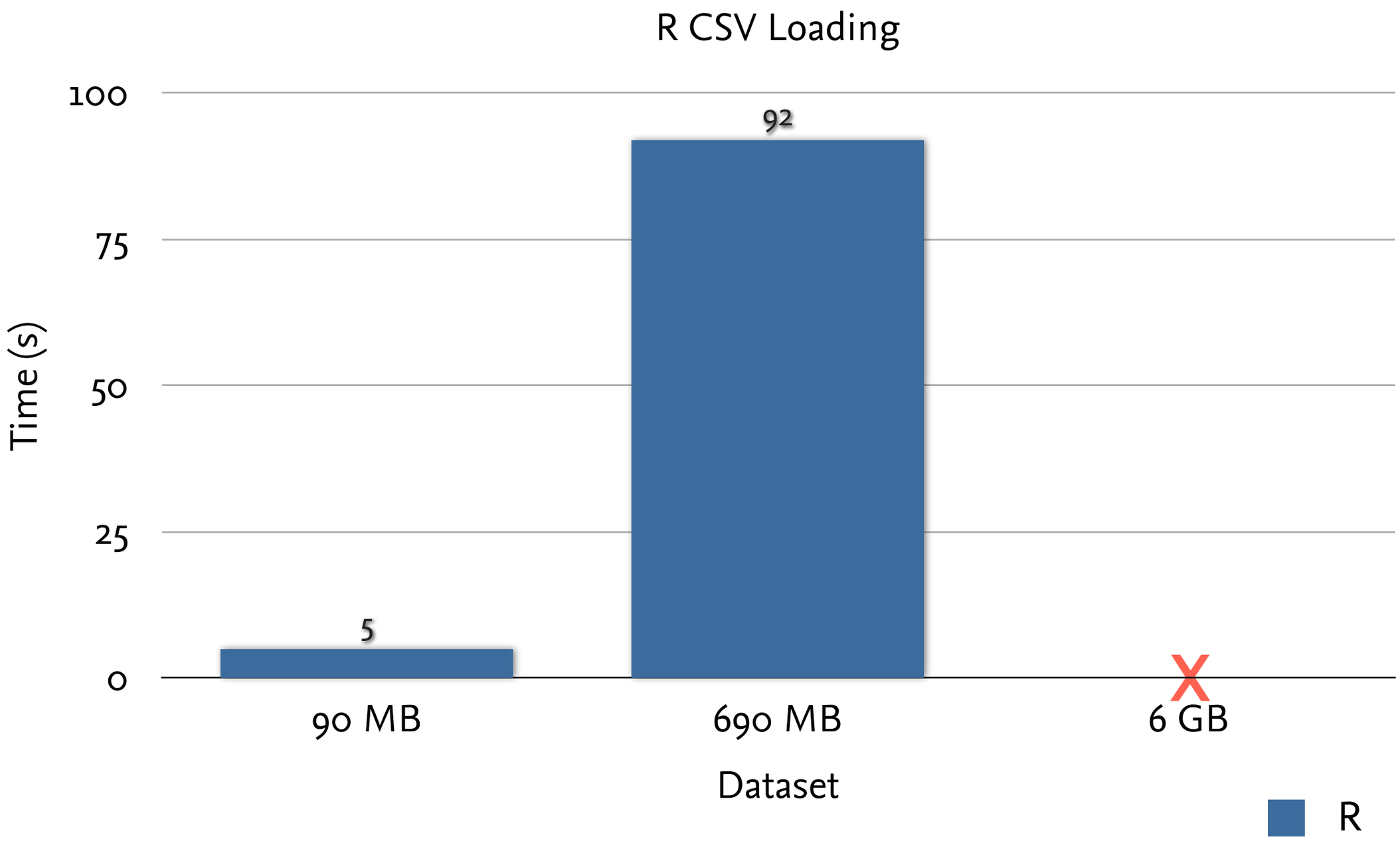
amst-R-dam meet-up, 2013-10-14

COMMIT/

Hannes Mühleisen


Collect data
Growing
Load data
Filter, transform & aggregate data
Analyze & Plot
Not really Analysis features
Publish paper/ Profit

Problem: #BiggeR



Running Example

- Say you are Starfleet Research and want to analyze warp drive performance (Coil Flux)
- Lots of data (~1G CSV, 68M records)

```
class,speed,flux
NX,1,11
Constitution,1,5
Galaxy,1,1
Defiant,1,3
Intrepid,1,1
NX,1,5
```

Solution?

- Use optimized data management system for data loading & retrieval
  - ... like a relational database
  - ... like a analytics-optimized database

Solution?

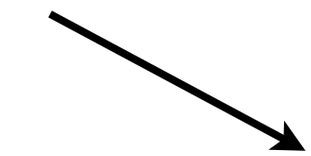
- Use optimized data management system for data loading & retrieval
  - ... like a relational database
  - ... like a analytics-optimized database
  - ... like MonetDB!

Relational DBs 101

class	speed	flux
NX	1	3
Constitution	1	8
Galaxy	1	3
Defiant	1	6
Intrepid	1	1

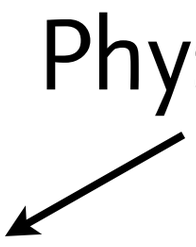
Postgres, Oracle, DB2, etc.:

Conceptional



class	speed	flux
NX	1	3
Constitution	1	8
Galaxy	1	3
Defiant	1	6
Intrepid	1	1

Physical (on Disk)



NX		1	3	Constitution		1	8	Galaxy	
1	3	Defiant		1	6	Intrepid		1	1

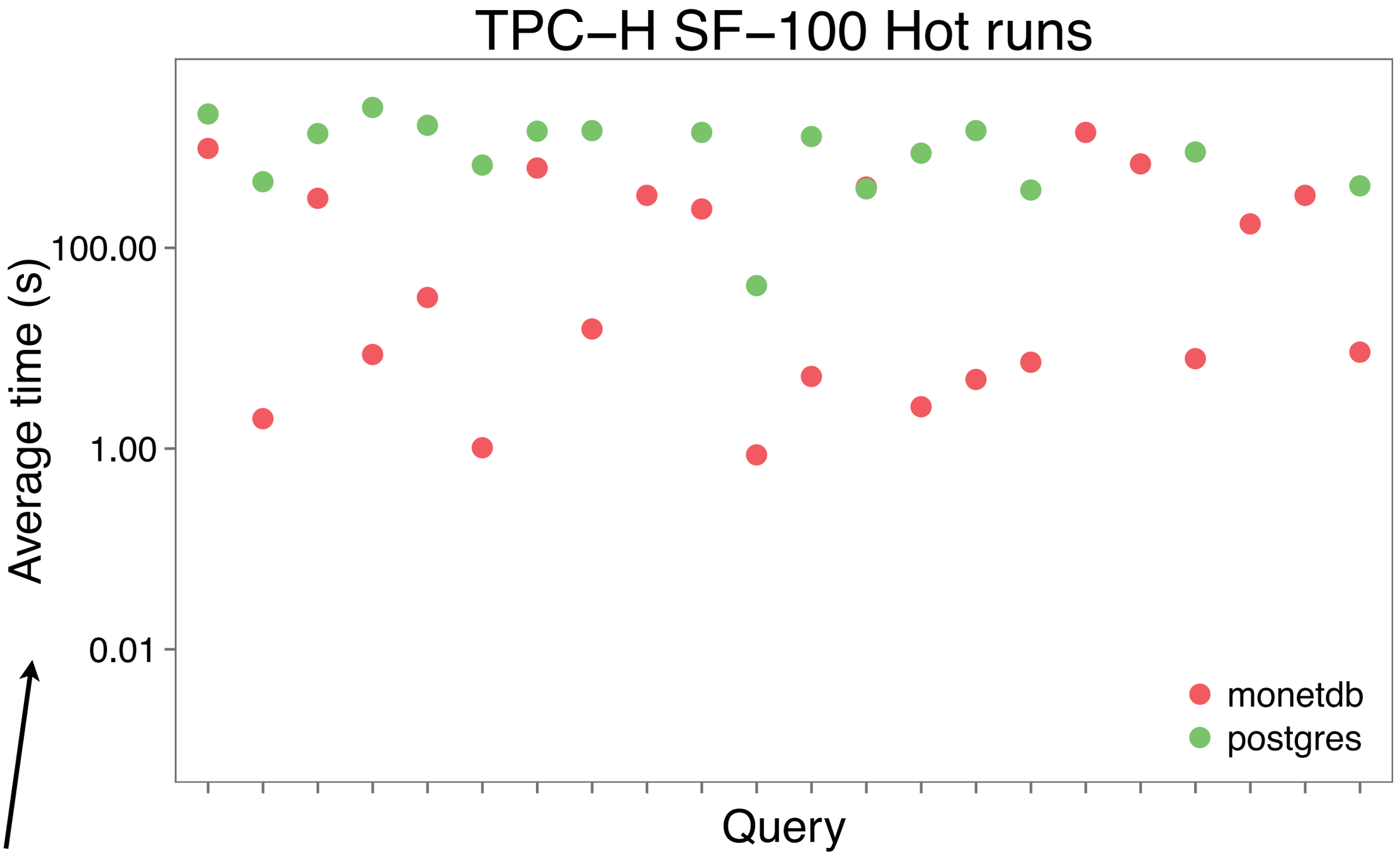
Column Store:

class	speed	flux
NX	1	3
Constitution	1	8
Galaxy	1	3
Defiant	1	6
Intrepid	1	1

NX	Constitution	Galaxy	Defiant	Intrepid
1	1	1	1	1
3	8	3	6	1

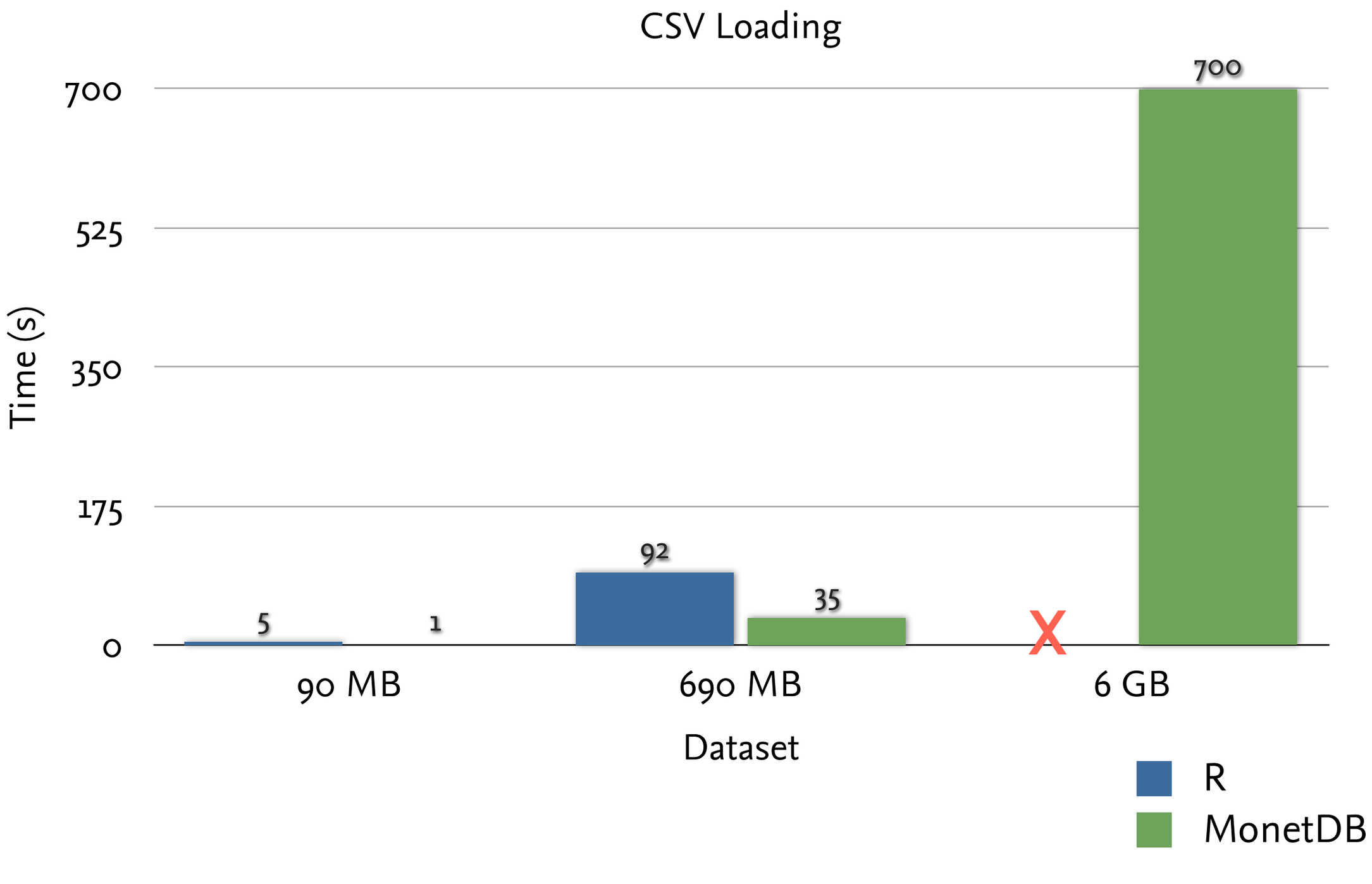
Peter A. Boncz, Martin L. Kersten, and Stefan Manegold. 2008. Breaking the memory wall in MonetDB. *Commun. ACM* 51, 12 (December 2008), 77-85. DOI=10.1145/1409360.1409380

Why Columns?



log!

First Gains...


CSV Loading
700
525
350
175
0
Time (s)
5
1
92
35
X
700
90 MB
690 MB
6 GB
Dataset
R
MonetDB

But then...

```
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

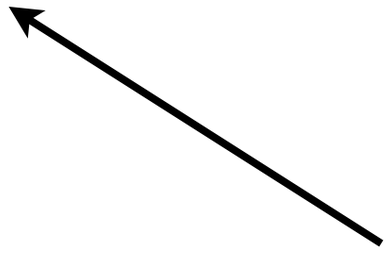
```
data <- monet.frame(conn,"starships")
nxw5 <- subset(data,class=="NX" & speed==5)$flux
t <- tabulate(nxw5,100)
normalized <- t/sum(t)
```

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

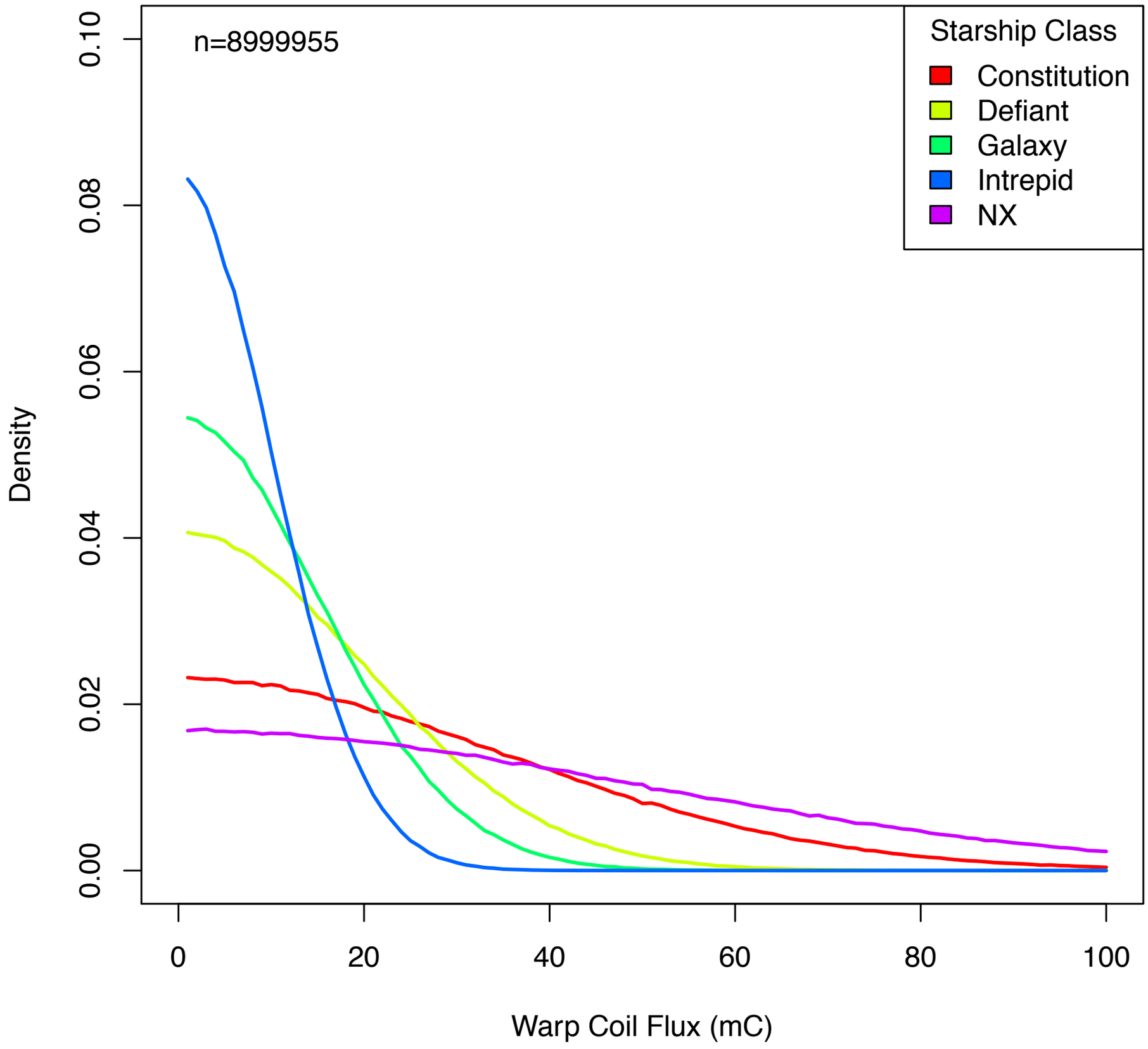
```
# R core
unique <- function(x, incomparables = FALSE, ...)
UseMethod("unique")
```

```
# MonetDB.R
unique.monet.frame <- function (x, incomparables = FALSE, ...)
as.vector(.col.func(x,"distinct",num=FALSE,aggregate=TRUE))

# On Shell
unique(wcflux$flux)
# result query: SELECT DISTINCT(flux) FROM starships;
```

Flux Analysis Script

```
wcflux <- read.table("starships.csv",sep=",",header=T)

classes <- sort(unique(wcflux$class))
wcflux5 <- subset(wcflux,speed==5)[c("class","flux")]

plot(0,0,ylim = c(0,0.1),xlim = c(0,100),type = "n")

for(i in 1:length(classes)){
   tclass <- classes[[i]]
   ct <- tabulate(subset(wcflux5,class==tclass)$flux,100)
   normalized <- ct/sum(ct)
   lines(data.frame(x=seq(1,100),y=normalized))
}
```

Density Plot of Warp Coil Flux per Starship Class (Warp 5)



Flux Analysis Script (2)

```
wcflux <- monet.frame(conn,"starships")
```

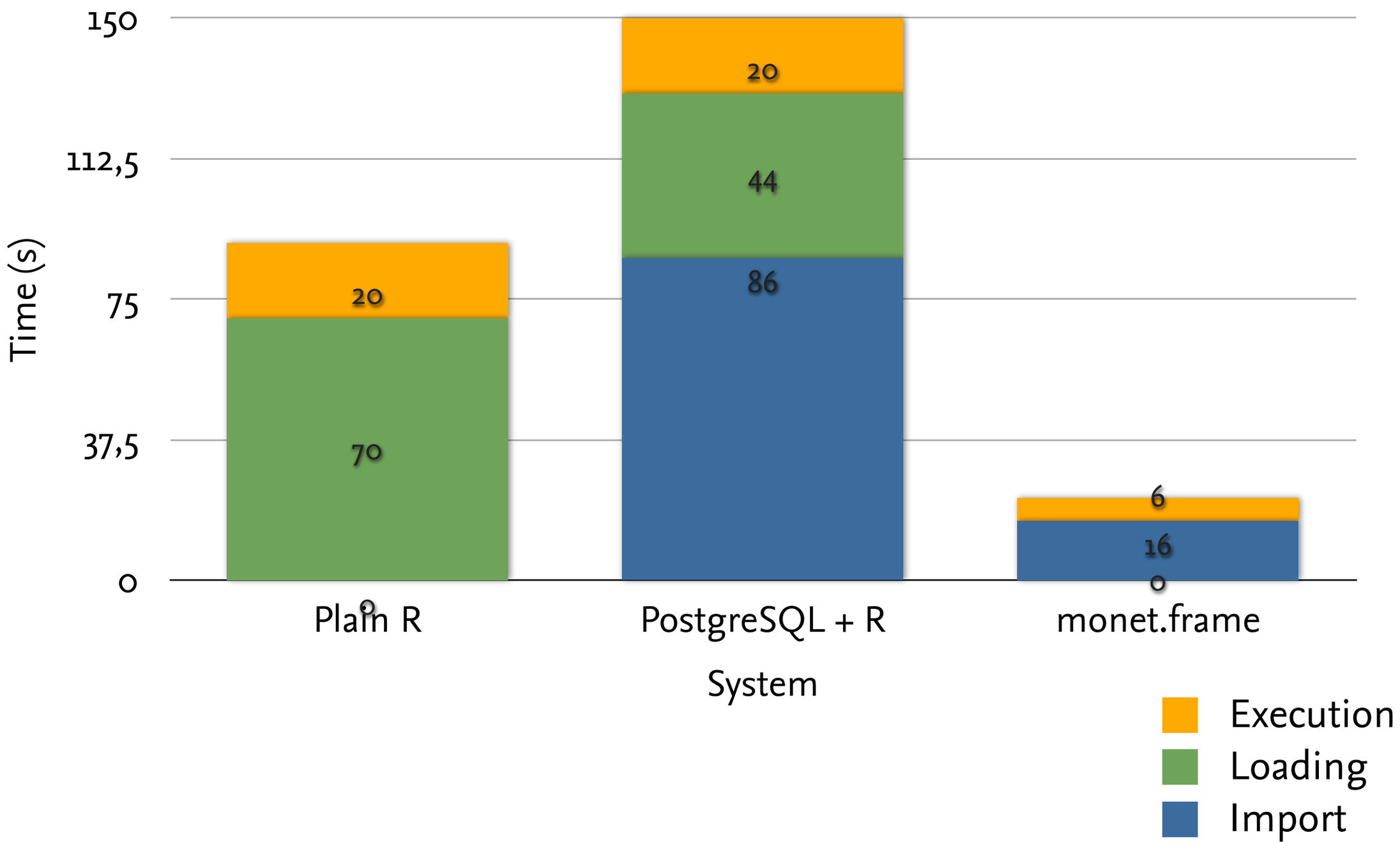
changed!

```
classes <- sort(unique(wcflux$class))
wcflux5 <- subset(wcflux,speed==3)[c("class","flux")]

plot(0,0,ylim = c(0,0.2),xlim = c(0,60),type = "n")

for(i in 1:length(classes)){
   tclass <- classes[[i]]
   ct <- tabulate(subset(wcflux5,class==tclass)$flux,60)
   normalized <- ct/sum(ct)
   lines(data.frame(x=seq(1,60),y=normalized))
}
```

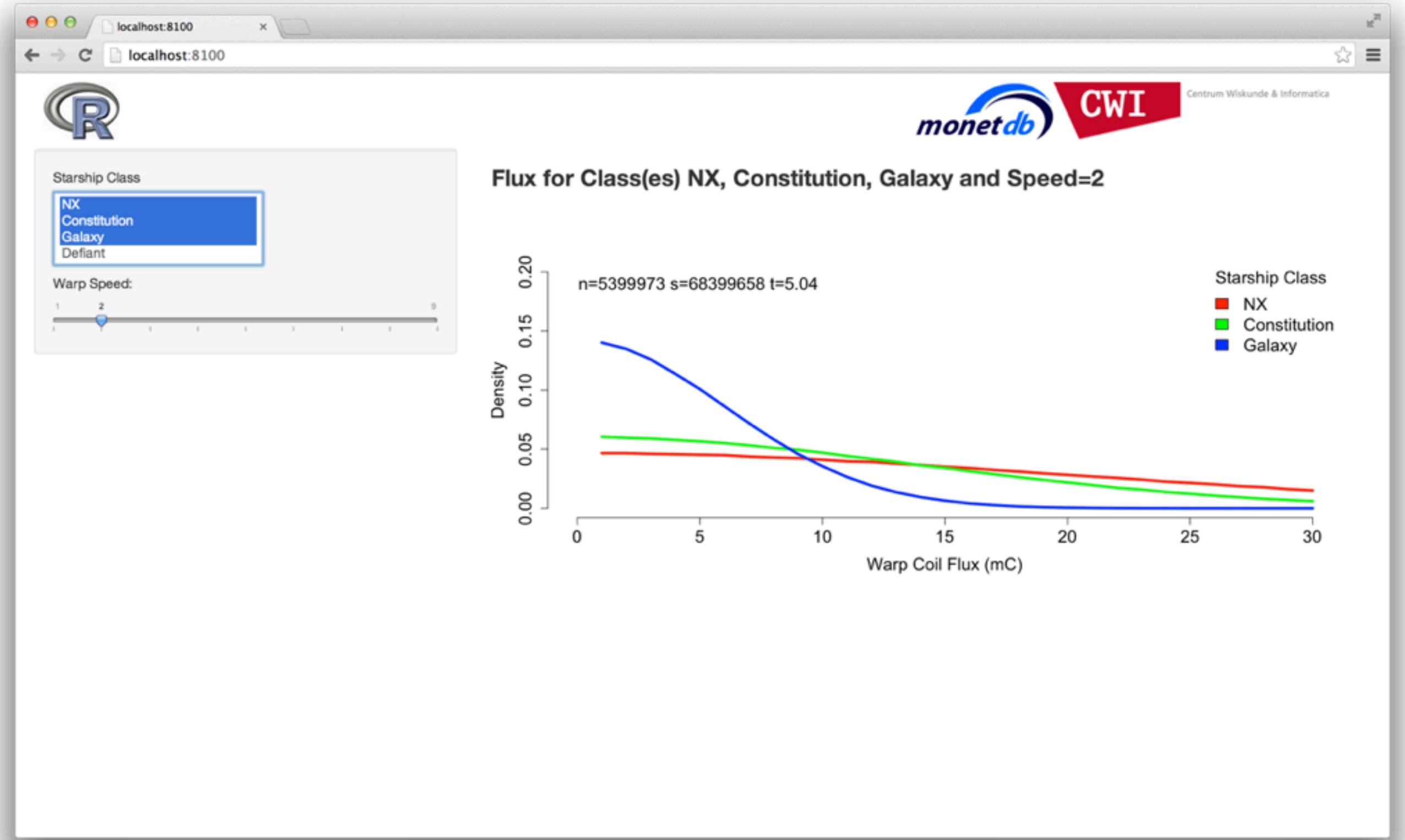
Generated SQL

```
SELECT DISTINCT(class) FROM starships;

SELECT t1,COUNT(t1) AS ct FROM (SELECT CAST(flux as integer)
AS t1 FROM starships WHERE ( (speed = 3) ) AND ( (class =
'Constitution') ) ) AS t WHERE t1 > 0 GROUP BY t1 ORDER BY t1
LIMIT 60;

-- [...]
```

Performance



Demo


localhost:8100
localhost:8100
monetdb
CWI
Centrum Wiskunde & Informatica
Starship Class
NX
Constitution
Galaxy
Defiant
Warp Speed:
Flux for Class(es) NX, Constitution, Galaxy and Speed=2
n=5399973 s=68399658 t=5.04
Starship Class
NX
Constitution
Galaxy
Density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
25
30
Warp Coil Flux (mC)

Collect data

Load data

Filter, transform & aggregate data

Analyze & Plot

Publish paper

sd() ^ trunc() sign() merge() sqrt()
range()
log() tabulate() floor()
subset()
str() ceiling()
exp() + sort() $ * []
/ na.omit()

Thank You!

tail()
sin()
range()

Questions?

summary()
sample()
head()
min()
abs() sum() quantile()
max()
-
round() names() dim() length() ==
signif()
aggregate()
print() var()

CRAN: MonetDB.R

monetdb