CWI

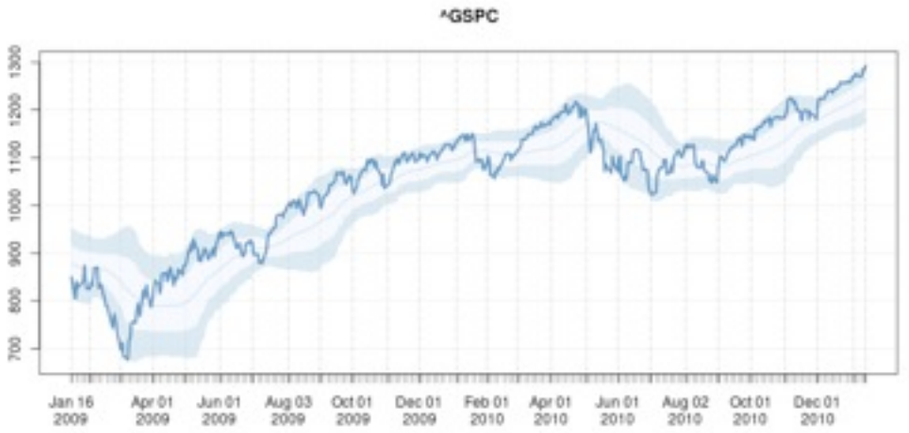
Centrum Wiskunde & Informatica

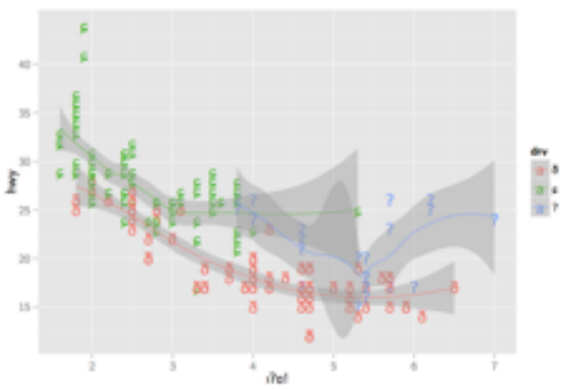
Symbiosis

Column Stores and R Statistics

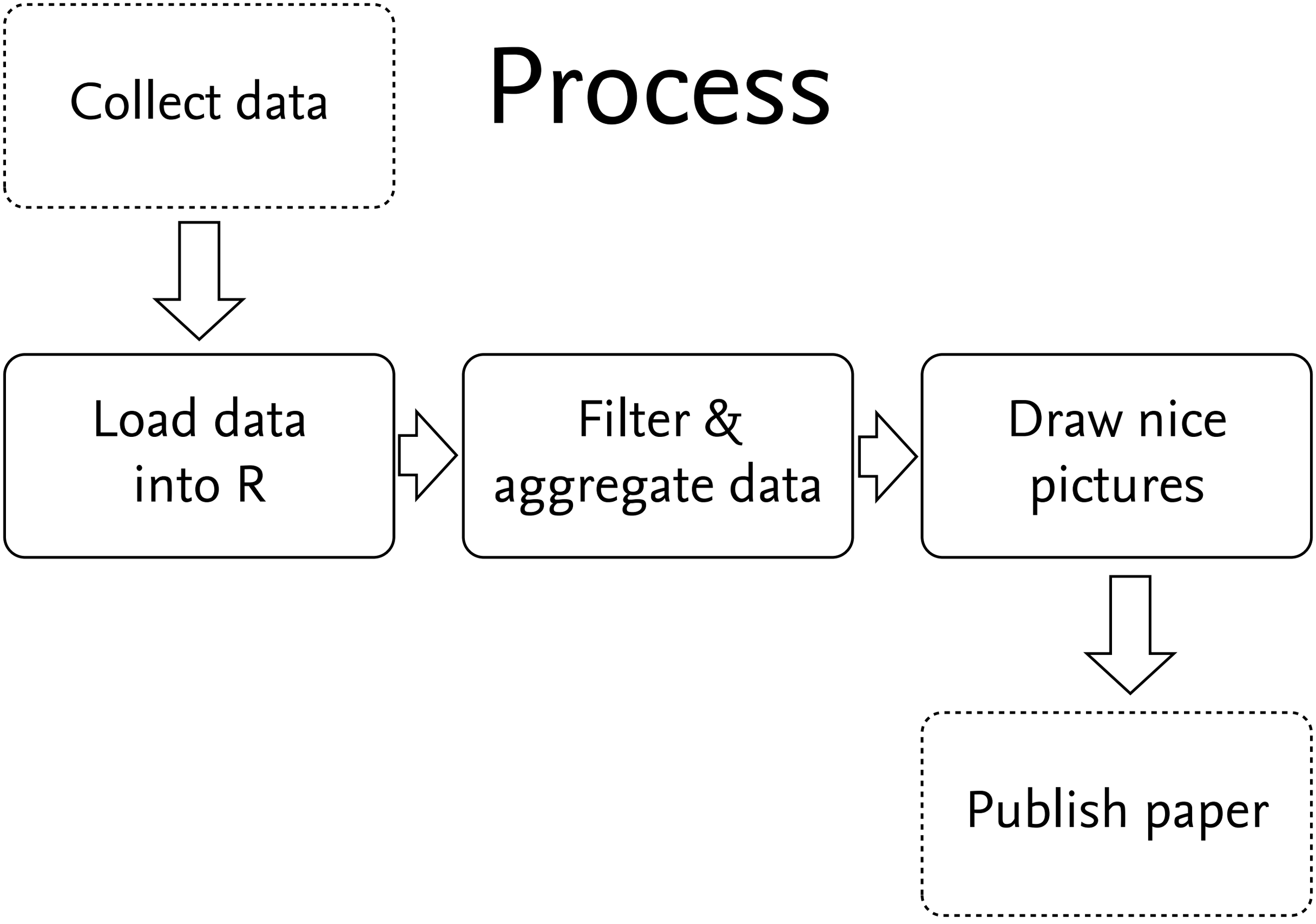
Why 

- Statistical computing & graphics
- Free, Open Source, ...
- Data Handling, Calculations, ...
- Lots of contributed packages
- Pictures!

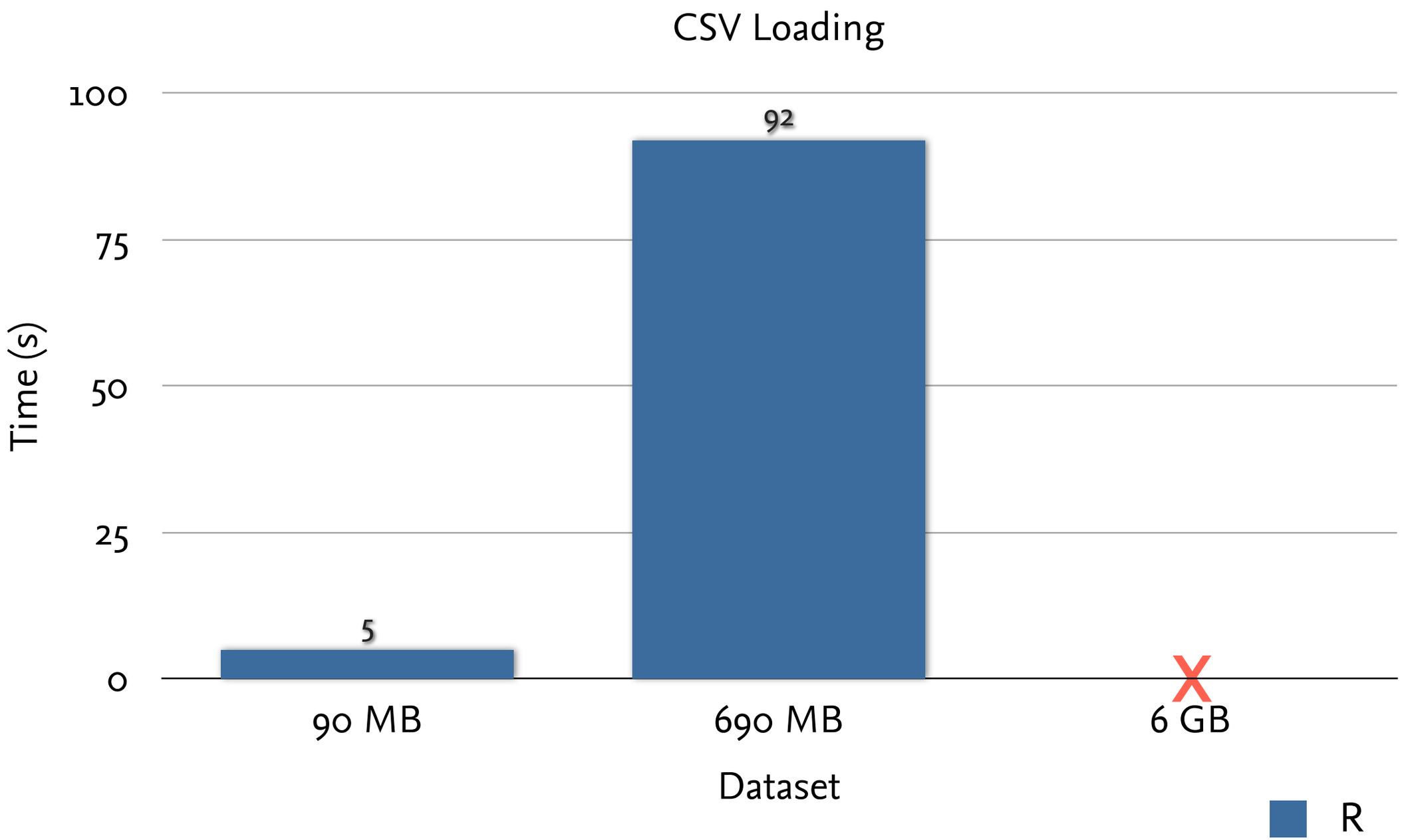




Process


Collect data
Load data into R
Filter & aggregate data
Draw nice pictures
Publish paper

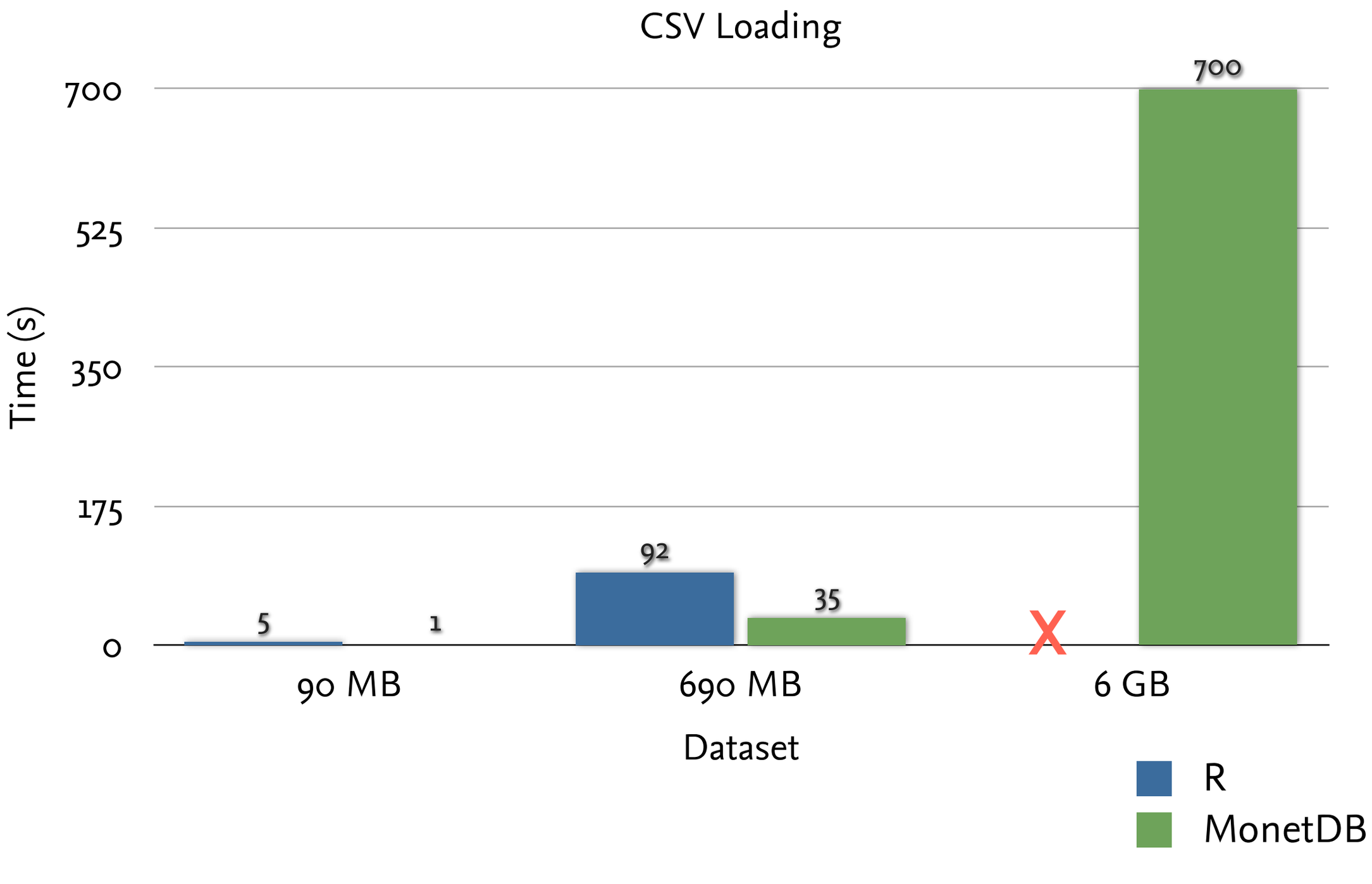
Problem



Solution?

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

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

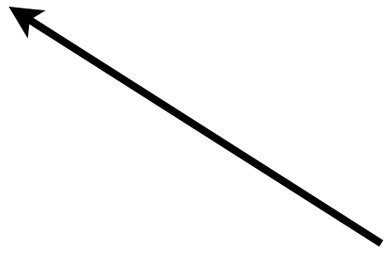
```
data <- monet.frame(conn,"starships")
nxw5 <- subset(data,class=="NX" & speed==5)$flux
t <- tabulate(nxw5,100)
normalized <- t/sum(t)
```

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

Small Example

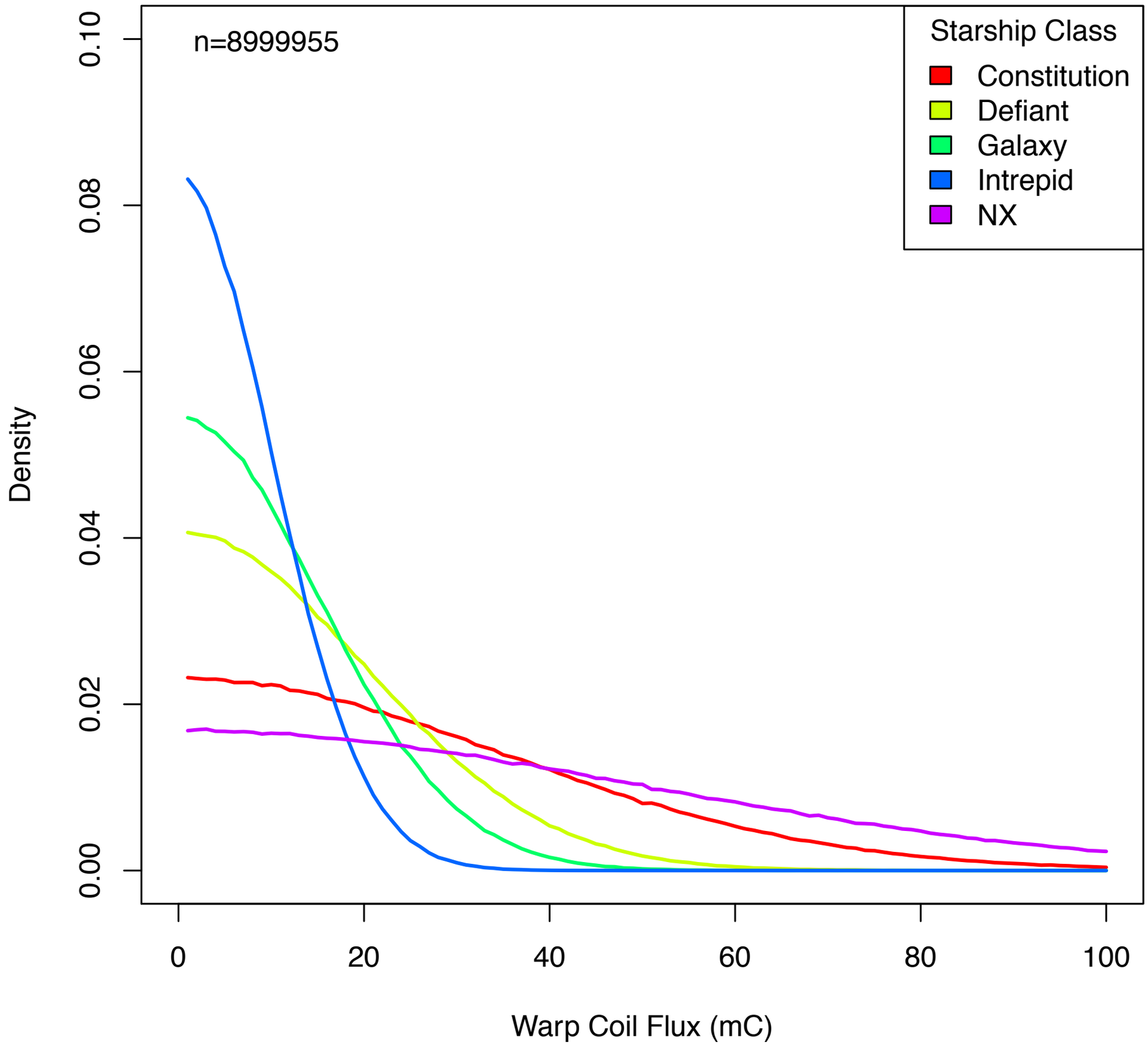
- Say you are Starfleet Research and want to analyze warp drive performance (Coil Flux)
- Lots of data (~1G CSV, 68M records)

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

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)



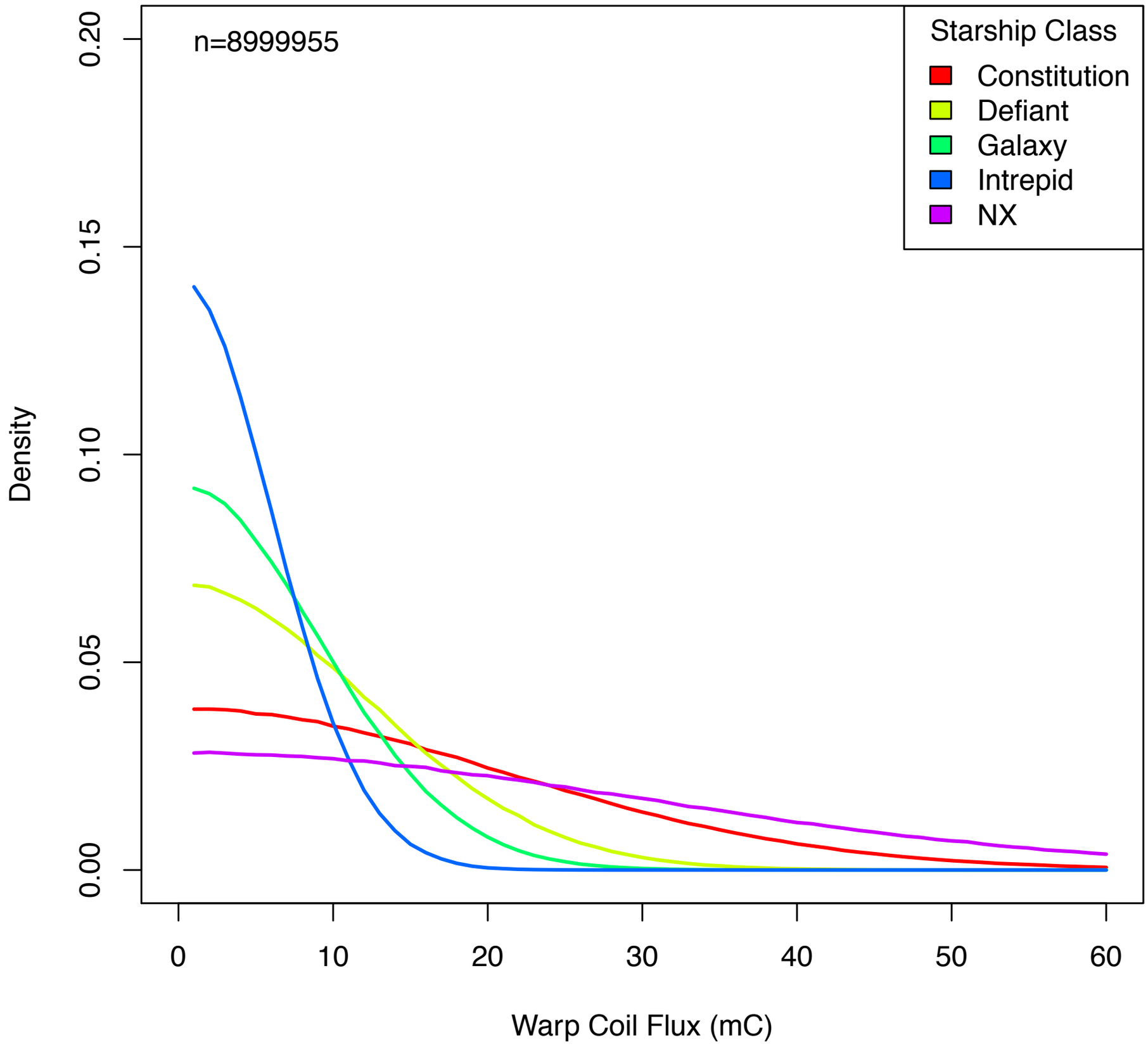
Flux Analysis Script (2)

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

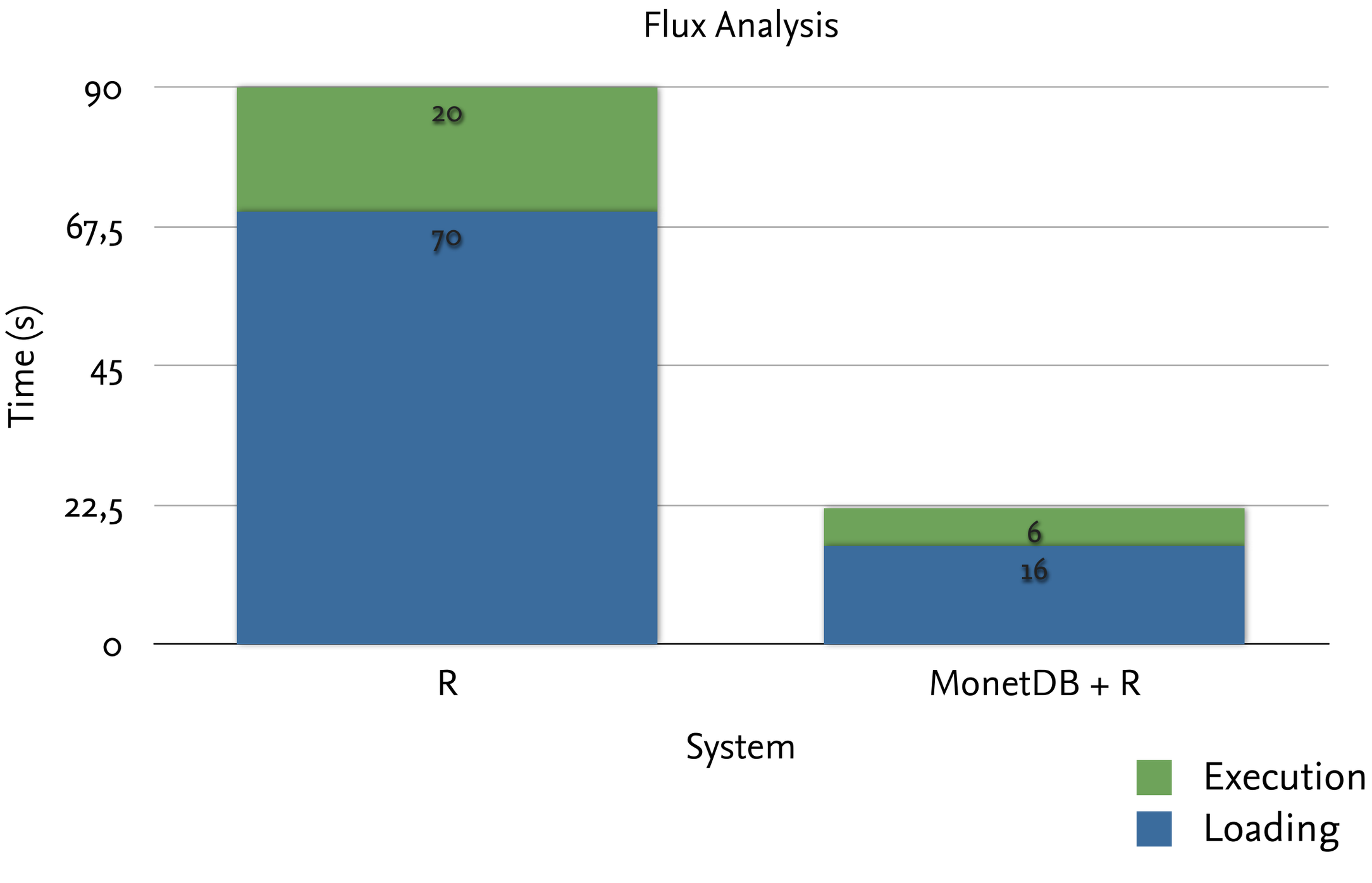
changed!

```
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))
}
```

Density Plot of Warp Coil Flux per Starship Class (Warp 3)



Performance



Thank You!

Questions?

- Hannes Mühleisen and Thomas Lumley:
  Best of Both Worlds – Relational Databases and Statistics
  25th International Conference on Scientific and Statistical Database Management (SSDBM2013), Jul. 2013

http://monetr.r-forge.r-project.org
http://hannes.muehleisen.org