A Data Viewer for R

Paul Murrell

The University of Auckland

July 30 2009

Overview

Motivation:	STATS 220
Problem statement:	Students do not understand what they cannot see.
What doesn't work:	`View()`
A solution:	The **rdataviewer** package and the `tcltkViewer()` function.
What else?:	Novel navigation interface, zooming, extensible for other data sources.

STATS 220 Data Technologies

- HTML (and CSS), XML (and DTDs), SQL (and databases), and **R** (and regular expressions)
- Online text book that nobody reads
- Computer lab each week (worth 0.5%) + three Assignments
  - 5 labs + one assignment on R
  - Emphasis on creating and modifying data structures
  - Attempt to use real data

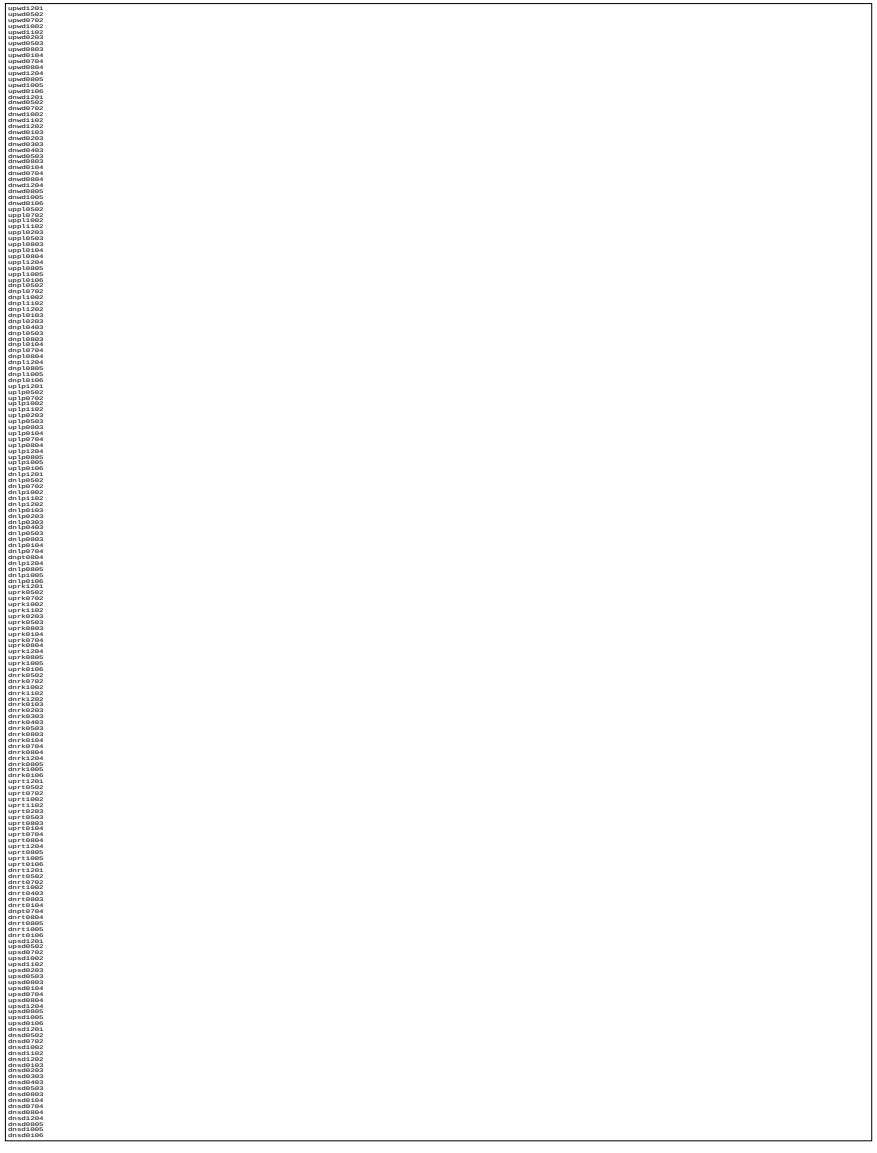
Example Lab Question

Read the file `lab10.txt` into R as a character vector.

You should end up with a symbol `habitats` that prints like this (this shows just the first 10 values; there are 192 values in total):

```
> head(habitats, 10)
 [1] "upwd1201" "upwd0502" "upwd0702"
 [4] "upwd1002" "upwd1102" "upwd0203"
 [7] "upwd0503" "upwd0803" "upwd0104"
[10] "upwd0704"
```

The file `lab10.txt`



Example Student Answer

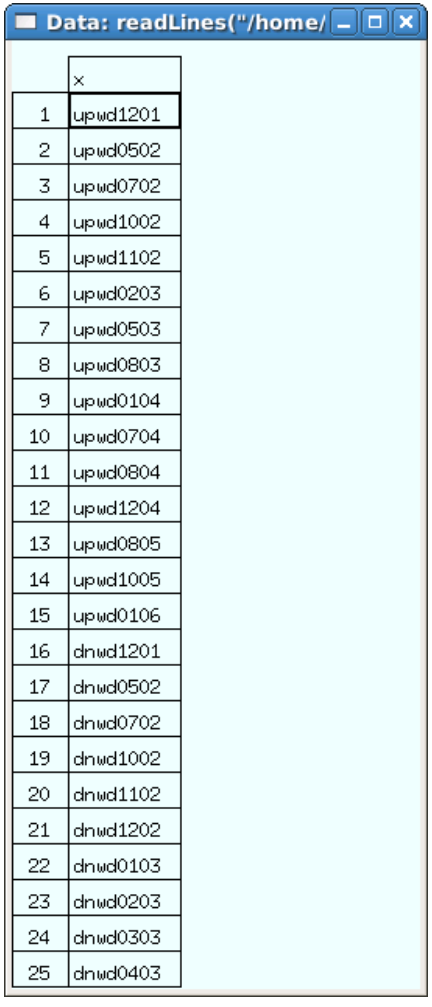
```
habitats <- head(readLines("lab10.txt"), 10)
```

To be fair, it is difficult to "see" all of an R data structure, unless the data structure is very small.

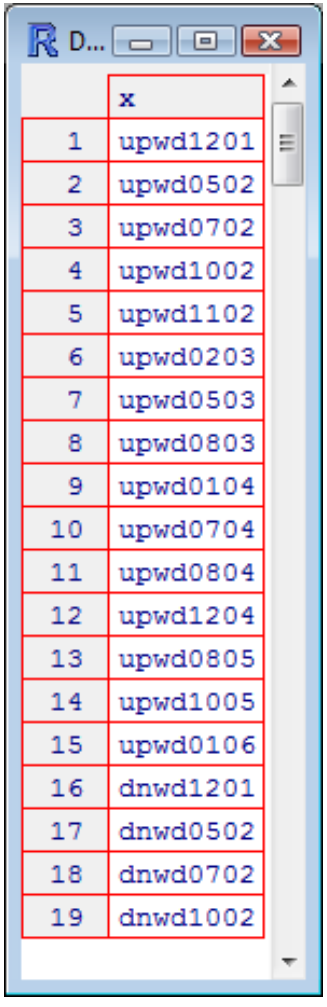
The main reason that I (occasionally) use a spreadsheet is just to be able to see the raw data in fixed-width columns.

The View() function

```
View(habitats)

<PgUp>
<PgDn>
<Home>
<End>
```



	x
1	upwd1201
2	upwd0502
3	upwd0702
4	upwd1002
5	upwd1102
6	upwd0203
7	upwd0503
8	upwd0803
9	upwd0104
10	upwd0704
11	upwd0804
12	upwd1204
13	upwd0805
14	upwd1005
15	upwd0106
16	dnwd1201
17	dnwd0502
18	dnwd0702
19	dnwd1002
20	dnwd1102
21	dnwd1202
22	dnwd0103
23	dnwd0203
24	dnwd0303
25	dnwd0403

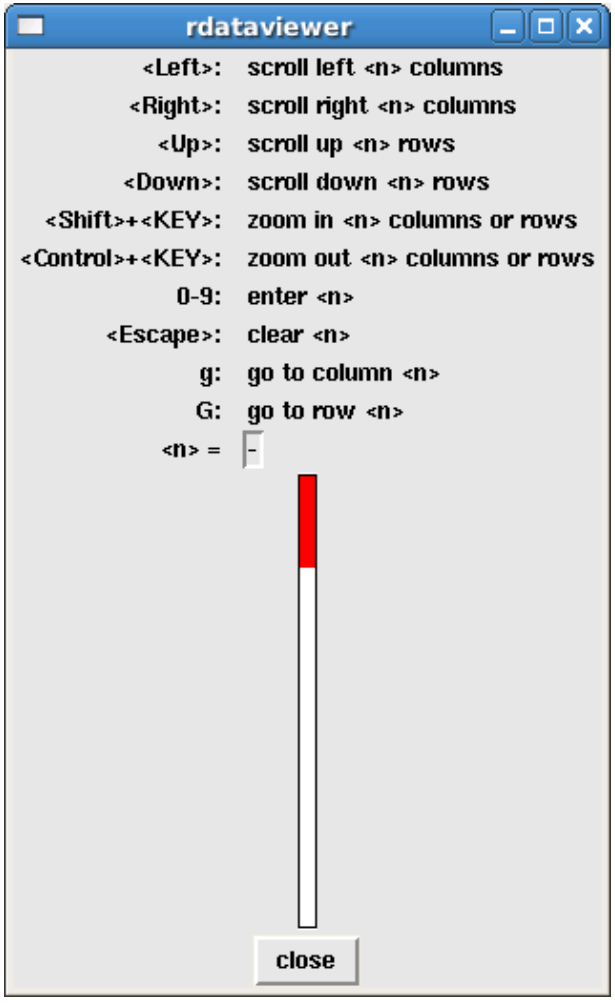


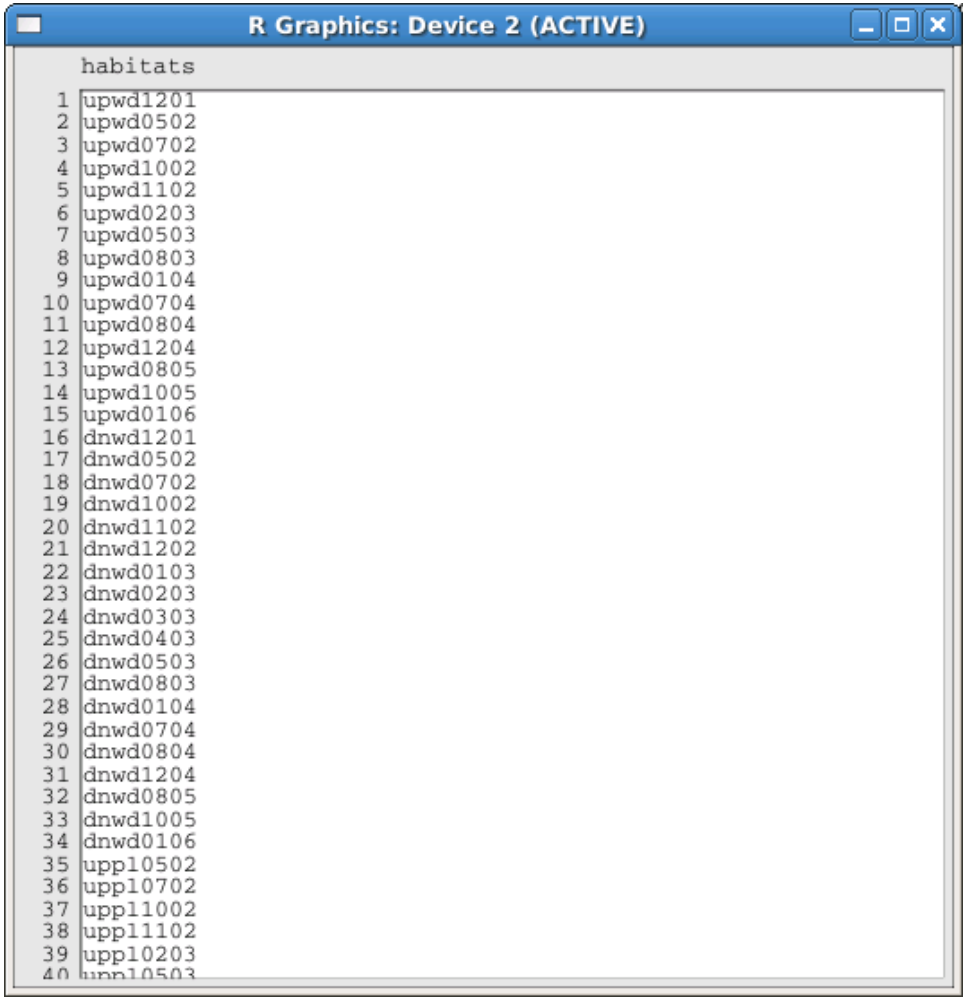
	x
1	upwd1201
2	upwd0502
3	upwd0702
4	upwd1002
5	upwd1102
6	upwd0203
7	upwd0503
8	upwd0803
9	upwd0104
10	upwd0704
11	upwd0804
12	upwd1204
13	upwd0805
14	upwd1005
15	upwd0106
16	dnwd1201
17	dnwd0502
18	dnwd0702
19	dnwd1002

The tcltkViewer() function

```
habitats <- readLines("lab10.txt")

data <- viewerData(habitats)

vdvp <- viewerDeviceVp(data)

v <- simpleViewer(data, dev=vdvp)

tcltkViewer(v)
```

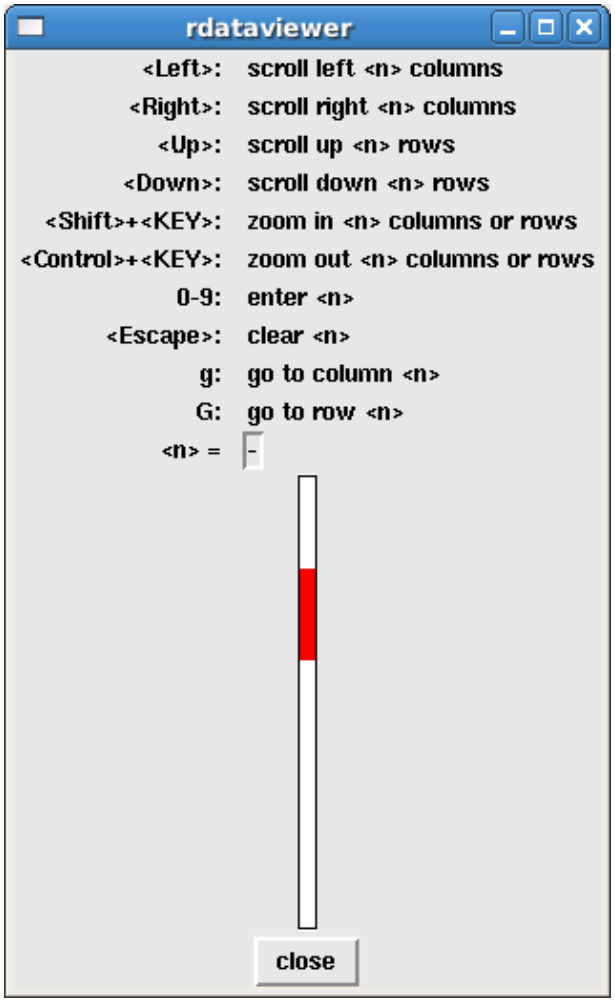
The tcltkViewer() function

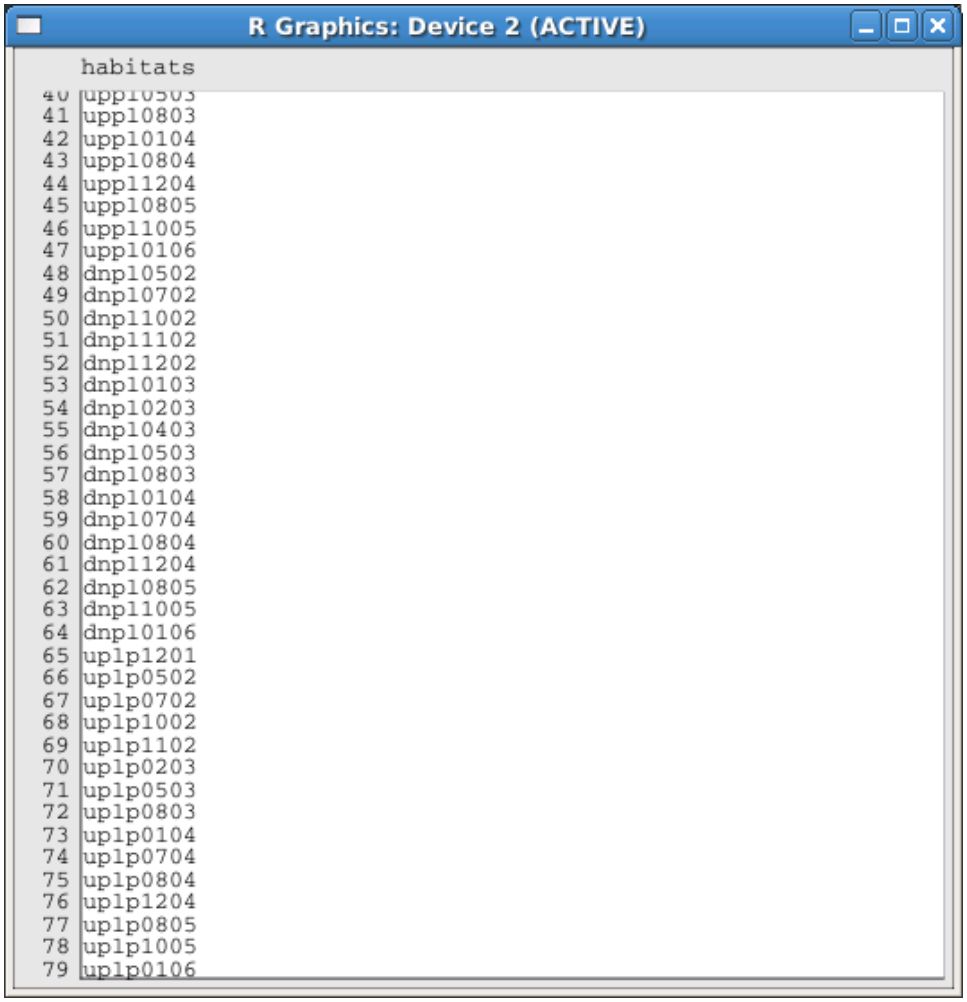




Standard-ish navigation (PgDn)

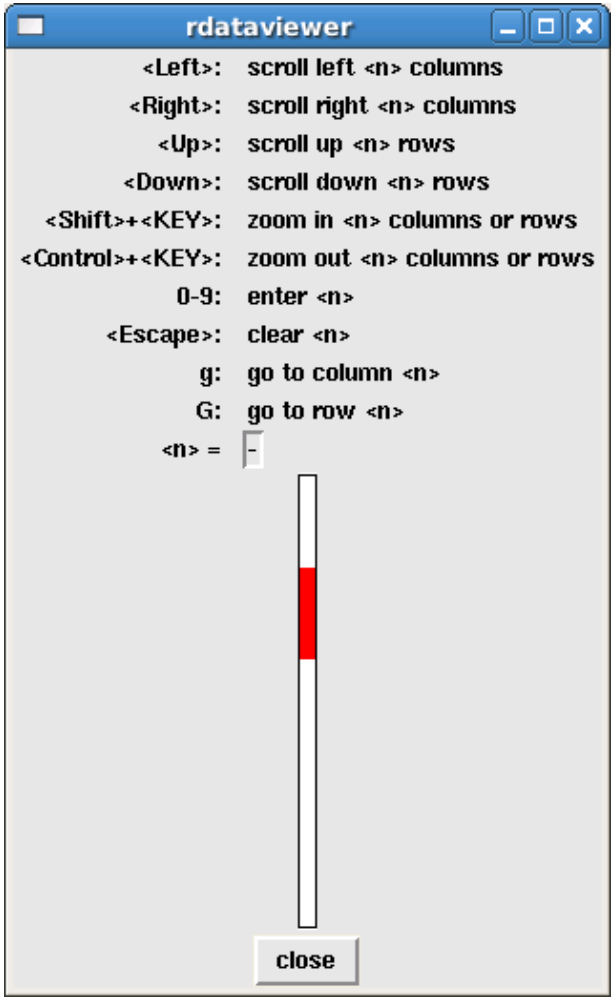
Now in "bottom-to-top" mode.

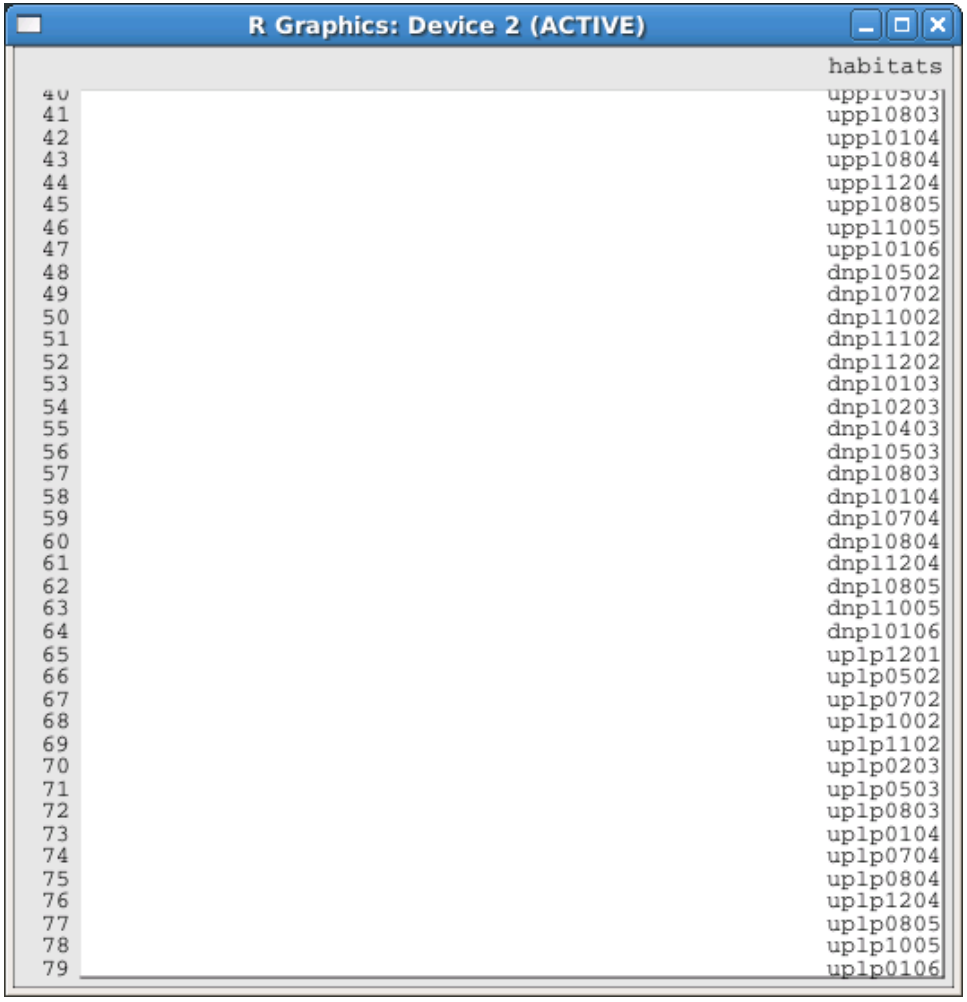




Standard-ish navigation (Right)

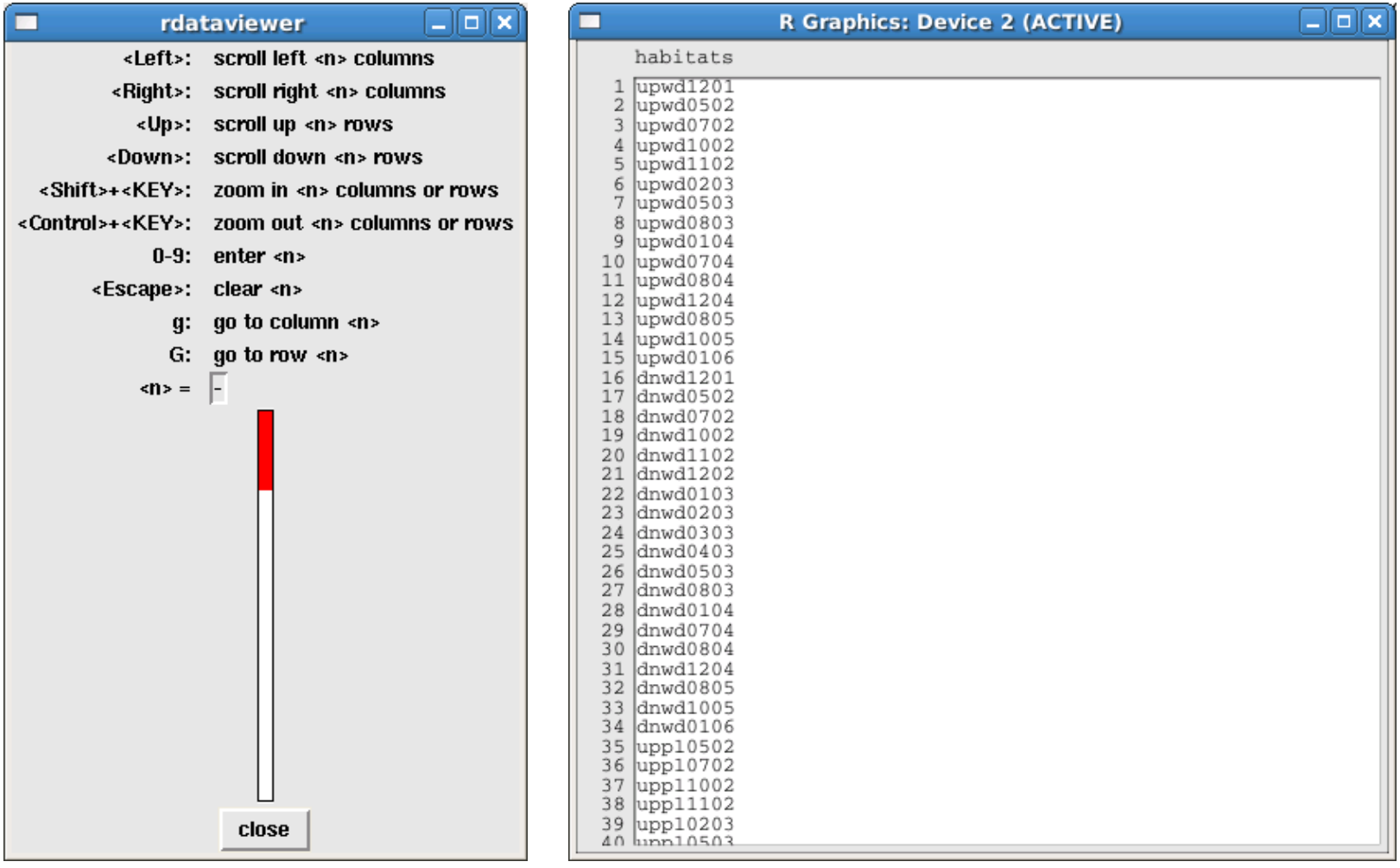
Now in `"right-to-left"` mode.



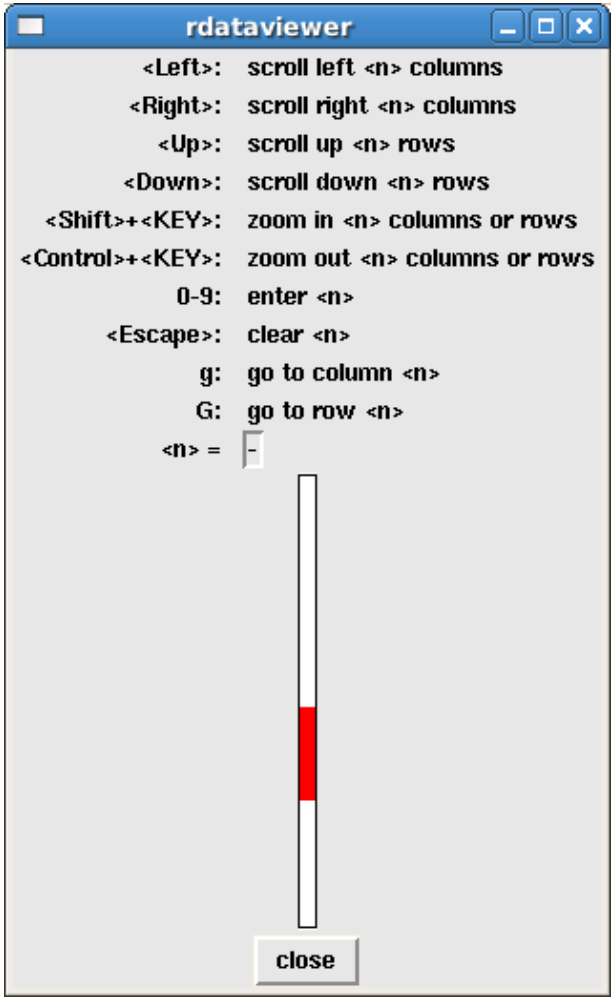


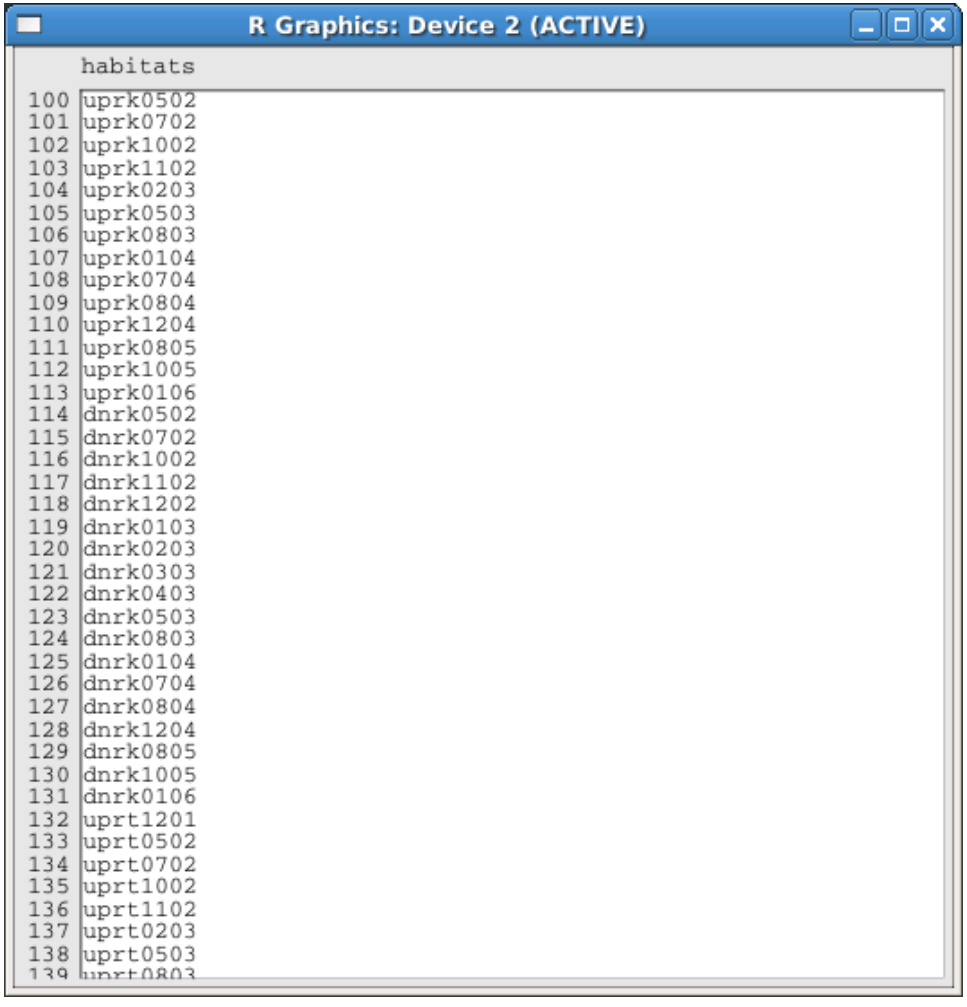
Standard-ish navigation (Ctrl-Home)

Now back in `"top-to-bottom"` and `"left-to-right"` mode.



vi-ish navigation (100 G)



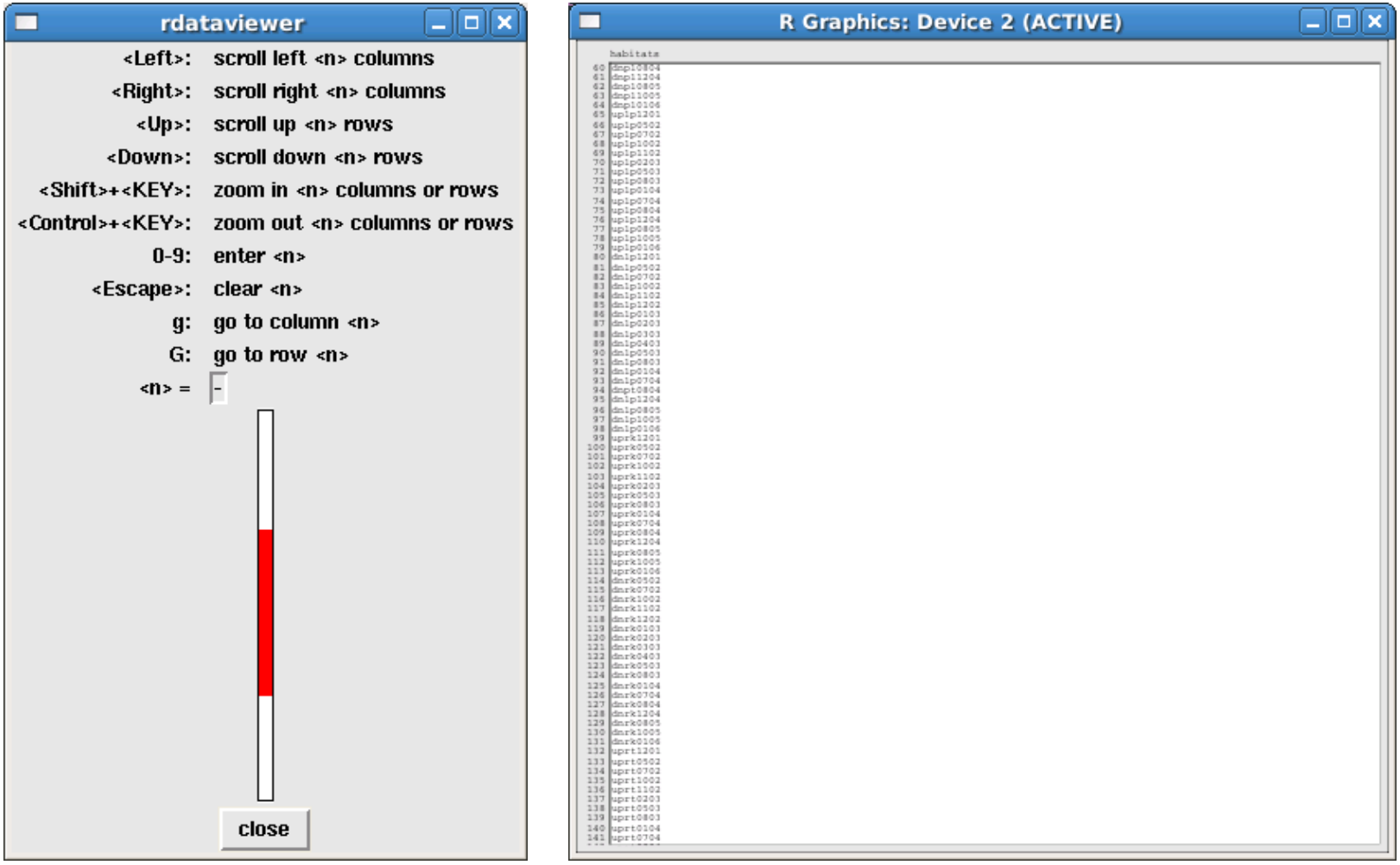


rdataviewer navigation

- Navigation is row- and column-based.
- There is an up-down mode and a left-right mode.
- Navigation commands can be preceded by a number (e.g., scroll down 3 rows).

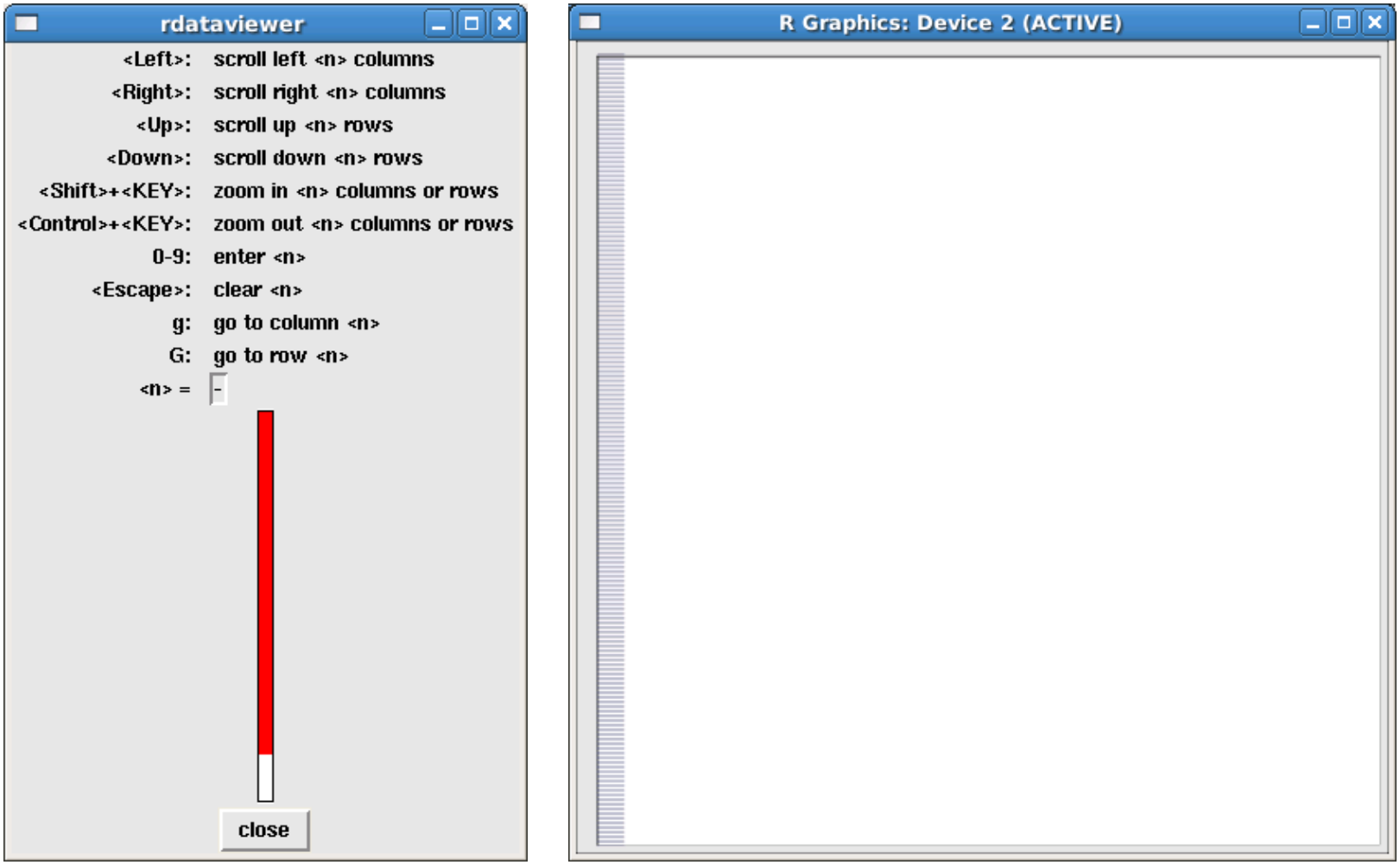
Zooming out (Ctrl-PgUp)

View twice as many rows.



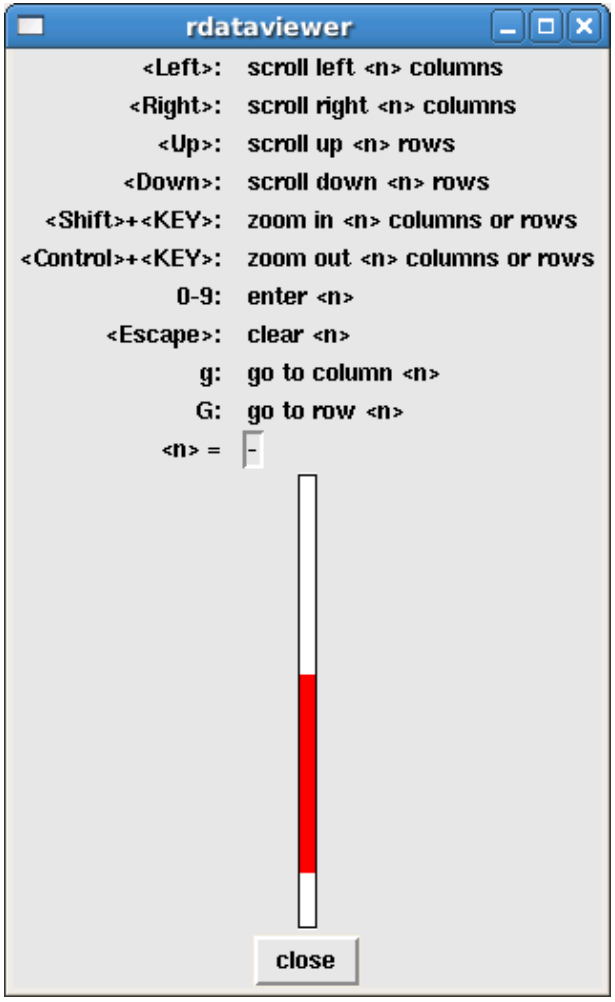
Zooming out (Ctrl-PgUp)

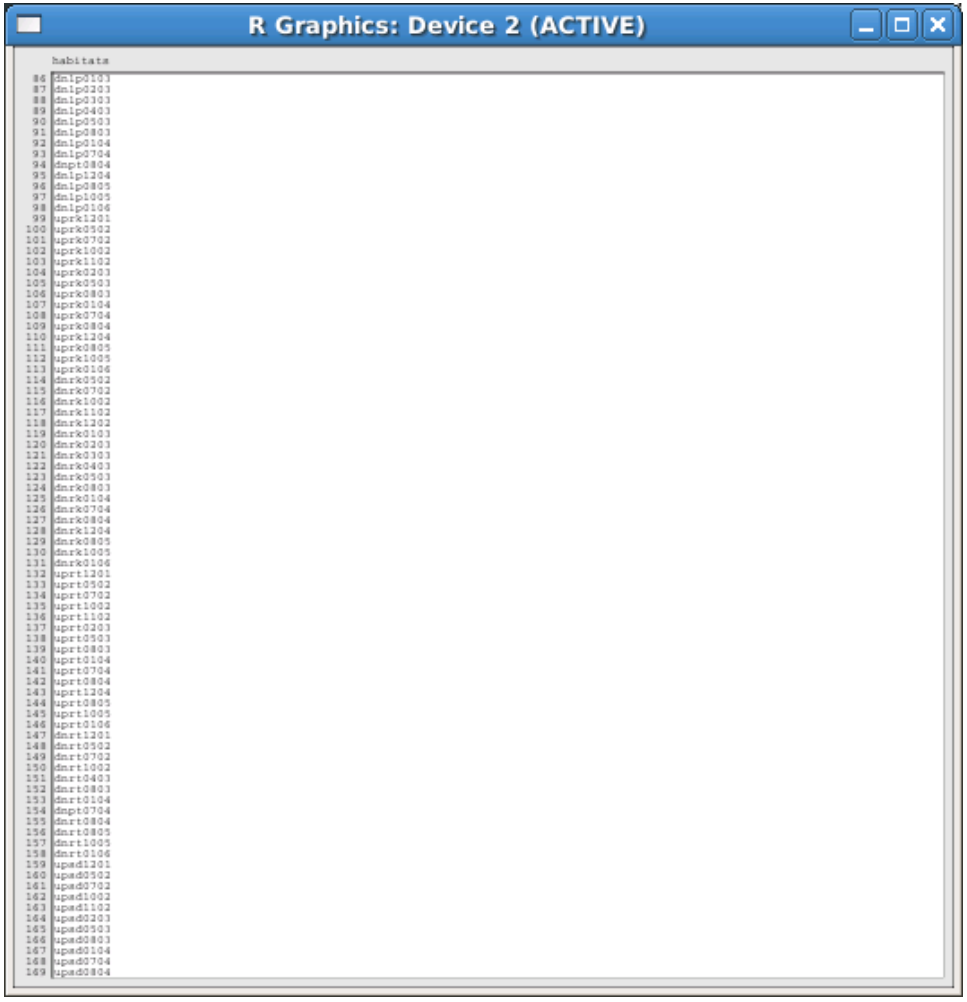
Below a threshold, lines are drawn instead of text.



Zooming in (Shift-PgUp)

View half as many rows.





rdataviewer zooming

- Ctrl-<command> zooms out.
- Shift-<command> zooms in.
- Zooming is row- and column-based.
- There is still an up-down mode and a left-right mode.
- Zooming commands can be preceded by a number (e.g., view 3 extra rows).

The exo-planets data set

- CSV file containing information on 192 exo-planets.
- To be read into R as a data frame.

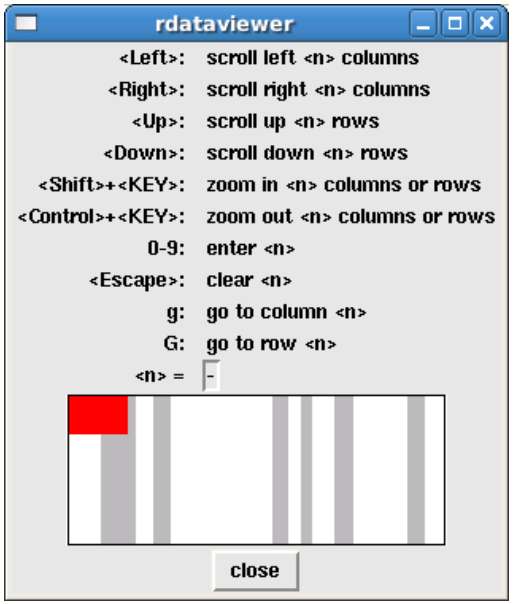
Viewing the exo-planets data set

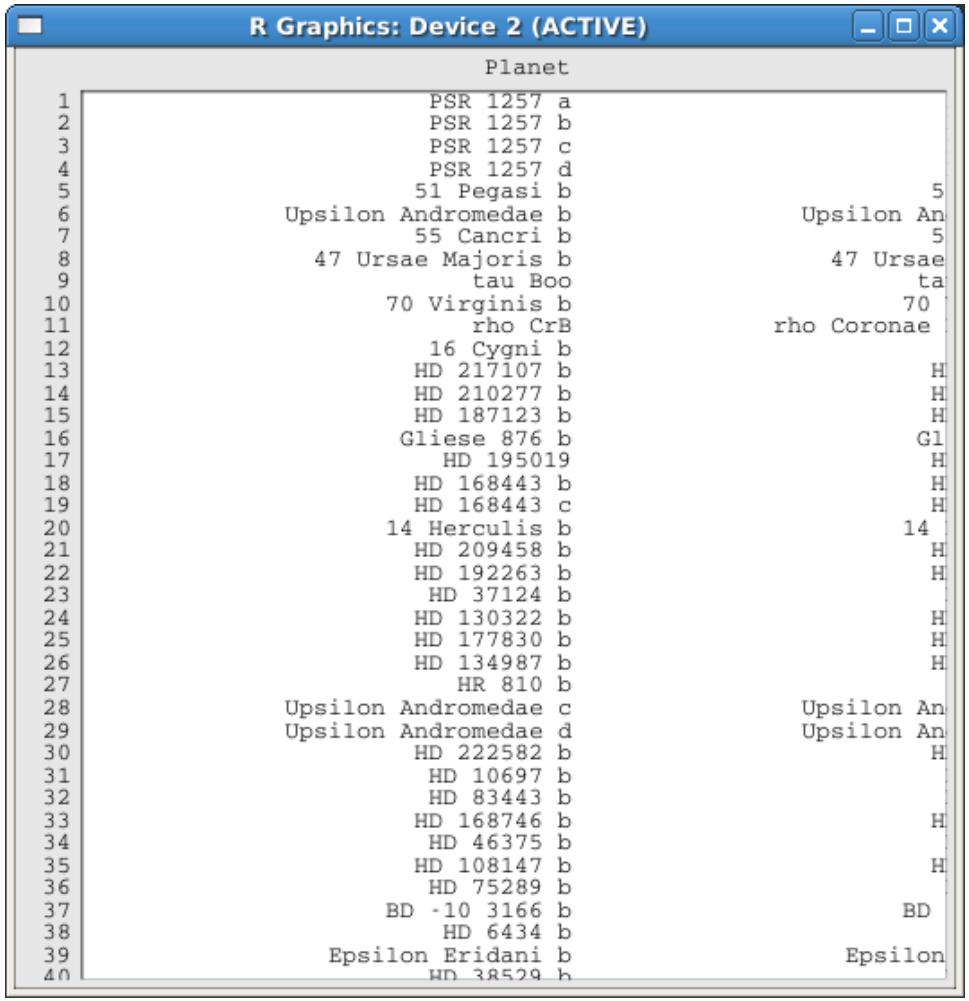
```
> allPlanetData <- read.csv("exoplanets.csv")

> head(allPlanetData)

                Planet               Star Constellation Year.of.Discovery
1           PSR 1257 a           PSR 1257                             1991
2           PSR 1257 b           PSR 1257                             1991
3           PSR 1257 c           PSR 1257                             1994
4           PSR 1257 d           PSR 1257                             1994
5         51 Pegasi b         51 Pegasi       Pegasus                1995
6 Upsilon Andromedae b Upsilon Andromedae    Andromedae                1996
                                                                  Discovered.By
1           Alexander Wolszczan and Dale Frail, National Radio Astronomy Observatory
2 Alexander Wolszczan and Dale Frail of the National Radio Observatory in New Mexico
3
4
5                             Michel Mayor and Didier Queloz, Geneva Observatory
6                  Geoffrey Marcy and R. Paul Butler, San Francisco State University
  Orbital.Radius Orbit.Period Planet.Type        Planet.Mass Orbit.Eccentricity
1        0.19 AU      25.262s      Pulsar      0 (Earth = 1)             -1.000
2        0.36 AU        66.5s      Pulsar      0 (Earth = 1)              0.100
3        0.46 AU        98.2s      Pulsar      0 (Earth = 1)              0.020
4          40 AU       62050s      Pulsar      0 (Earth = 1)             -1.000
5        0.05 AU        4.23s Hot Jupiter 0.47 (Jupiter = 1)              0.000
6        0.05 AU        4.62s Hot Jupiter 0.71 (Jupiter = 1)              0.034
                     Method.of.Detection Star.Coords...RA. Star.Coords...DEC.
1                                 Timing          13 00 01          +12 40 00
2                                 Timing          13 00 01          +12 40 00
3                                 Timing          13 00 01          +12 40 00
4                                 Timing          13 00 01          +12 40 00
5 Radial Velocity (or Doppler Spectroscopy)       22 57 27.14        +20 46 04.5
6 Radial Velocity (or Doppler Spectroscopy)      01 36 48.527       +41 24 38.71
```

Viewing the exo-planets data set

```
data <- viewerDataFrame(read.csv("exoplanets.csv"))

vdvp <- viewerDeviceVp(data)

v <- simpleViewer(data, dev=vdvp)

tcltkViewer(v)
```

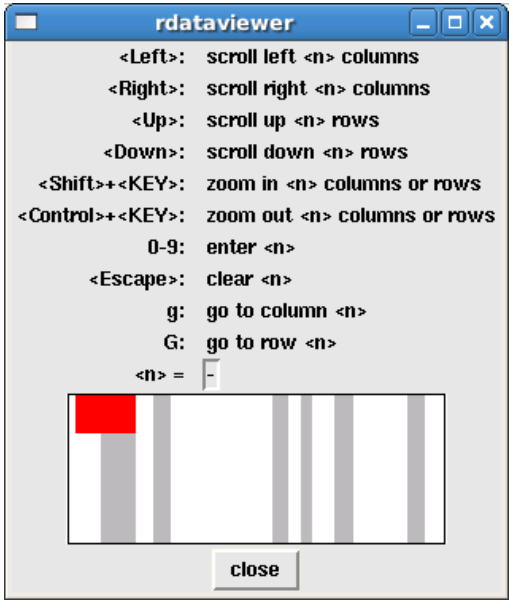
Viewing the exo-planets data set

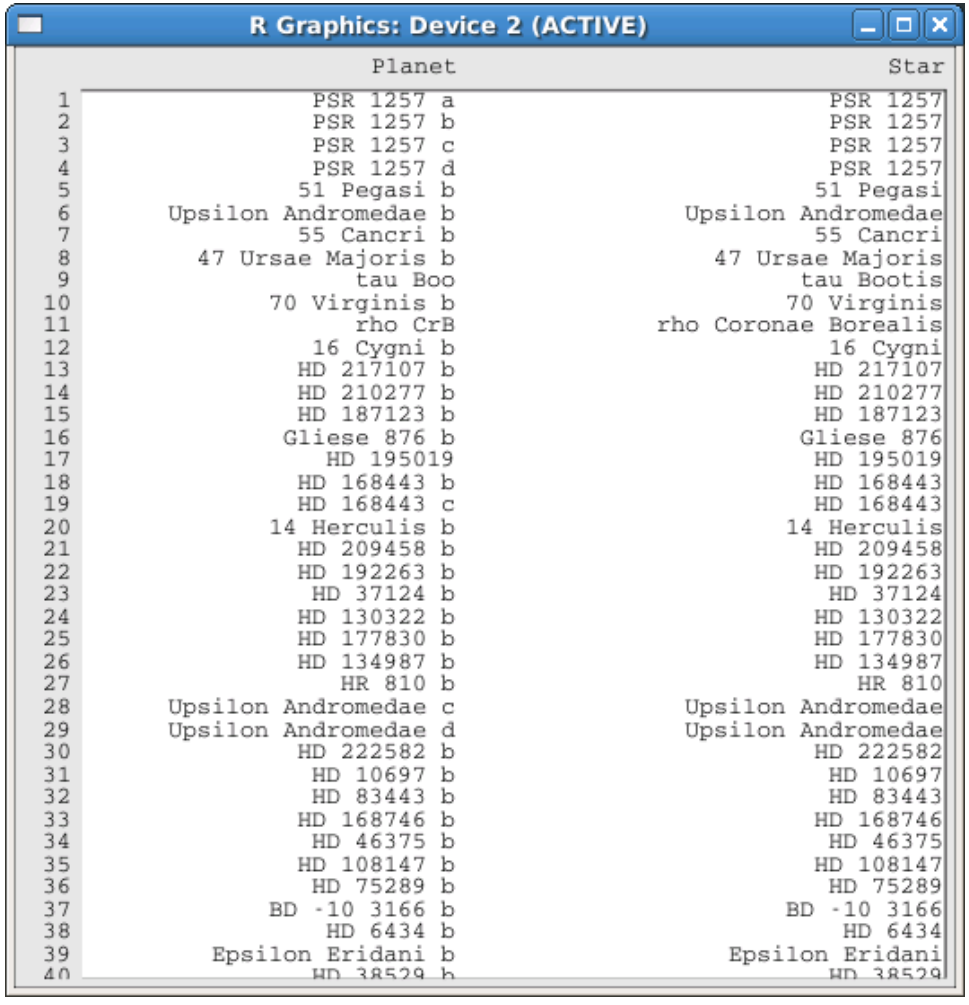




Standard-ish navigation (Right)

Now the `"right-to-left"` mode might make more sense.

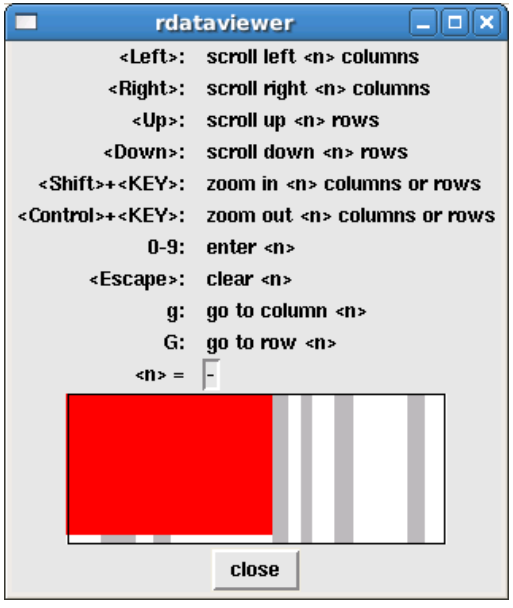


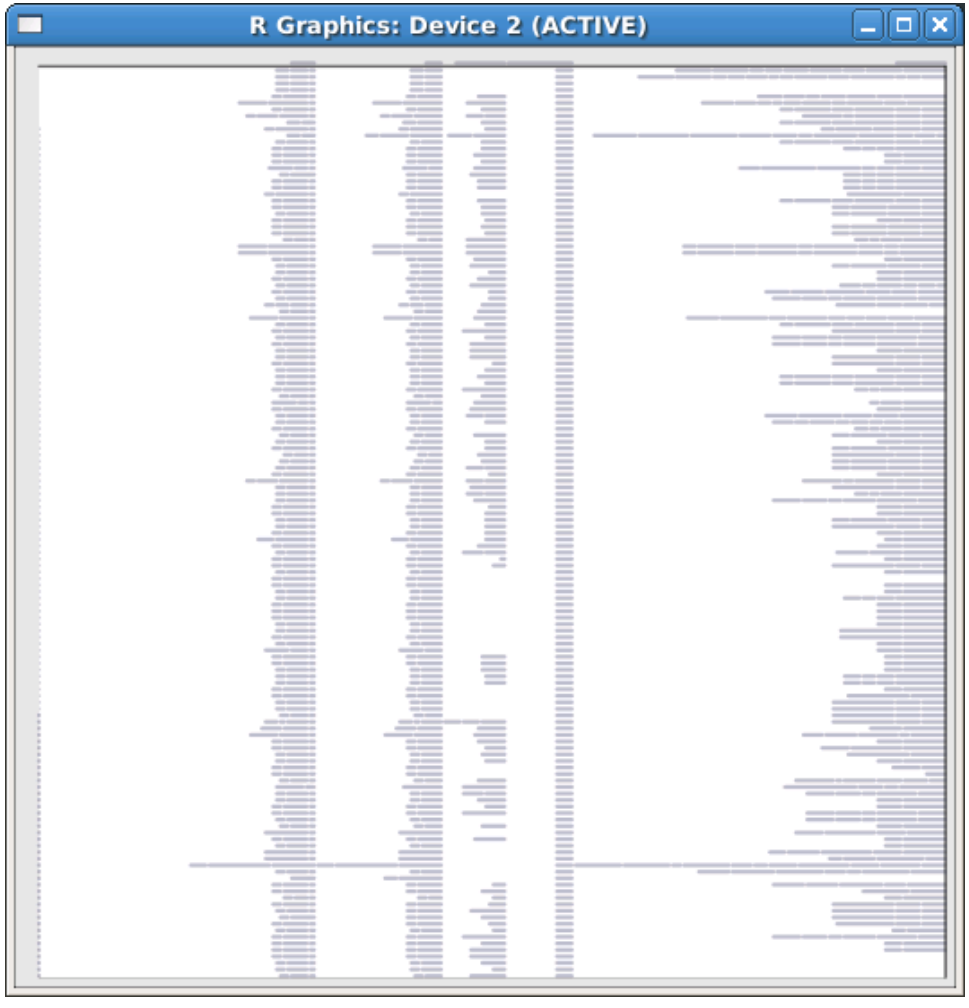


	Planet	Star
1	PSR 1257 a	PSR 1257
2	PSR 1257 b	PSR 1257
3	PSR 1257 c	PSR 1257
4	PSR 1257 d	PSR 1257
5	51 Pegasi b	51 Pegasi
6	Upsilon Andromedae b	Upsilon Andromedae
7	55 Cancri b	55 Cancri
8	47 Ursae Majoris b	47 Ursae Majoris
9	tau Boo	tau Bootis
10	70 Virginis b	70 Virginis
11	rho CrB	rho Coronae Borealis
12	16 Cygni b	16 Cygni
13	HD 217107 b	HD 217107
14	HD 210277 b	HD 210277
15	HD 187123 b	HD 187123
16	Gliese 876 b	Gliese 876
17	HD 195019	HD 195019
18	HD 168443 b	HD 168443
19	HD 168443 c	HD 168443
20	14 Herculis b	14 Herculis
21	HD 209458 b	HD 209458
22	HD 192263 b	HD 192263
23	HD 37124 b	HD 37124
24	HD 130322 b	HD 130322
25	HD 177830 b	HD 177830
26	HD 134987 b	HD 134987
27	HR 810 b	HR 810
28	Upsilon Andromedae c	Upsilon Andromedae
29	Upsilon Andromedae d	Upsilon Andromedae
30	HD 222582 b	HD 222582
31	HD 10697 b	HD 10697
32	HD 83443 b	HD 83443
33	HD 168746 b	HD 168746
34	HD 46375 b	HD 46375
35	HD 108147 b	HD 108147
36	HD 75289 b	HD 75289
37	BD -10 3166 b	BD -10 3166
38	HD 6434 b	HD 6434
39	Epsilon Eridani b	Epsilon Eridani
40	HD 38529 b	HD 38529

Zooming out (3 Ctrl-Right)

View 3 more columns.





The design of **rdataviewer**

- A data source to view.

  ```
  data <- viewerDataFrame(read.csv("exoplanets.csv"))
  ```

- A device to draw the view on.

  ```
  vdvp <- viewerDeviceVp(data)
  ```

- An object to represent the current view.

  ```
  v <- simpleViewer(data, dev=vdvp)
  ```

- A GUI to allow interactive modification of the view.

  ```
  tcltkViewer(v)
  ```

Each component is represented by an S4 virtual class so alternative implementations are possible.

The ViewerData class

Examples we have already seen:

- A character vector (ViewerDataVector)

```
data <- viewerData(habitats)
```

- A data frame (ViewerDataFrame)

```
data <- viewerDataFrame(read.csv("exoplanets.csv"))
```

The `ViewerData` class

Generics that require implementing:

- How many rows and columns?
  `dimensions(data)`
- How wide is each column?
  `colWidths(data, which)`
- What are the column names?
  `colNames(data, cols)`
- Provide a text representation of given rows and columns
  `getText(data rows, cols)`

The ViewerDataFrame class

Methods for viewing data frames:

- How many rows and columns?
  `dim(data@df)`
- How wide is each column?
  `max(nchar(capture.output(print(data@df[i]))))`
- What are the column names?
  `colnames(data@df[, cols])`
- Provide a text representation of given rows and columns
  `capture.output(print(data[rows, cols]))`

The `ViewerDataText` class

Methods for viewing external text files:

Scan file to build an index of pointers to line beginnings.

- How many rows and columns?
  Number of line beginnings (only 1 column).
- How wide is each column?
  Longest difference between line beginnings.
- What are the column names?
  Filename.
- Provide a text representation of given rows and columns
  Seek to the relevant line beginning in the file then `readLines()`.

The Meter Readings data set

- 300 MB XML file (4,772,013 lines)

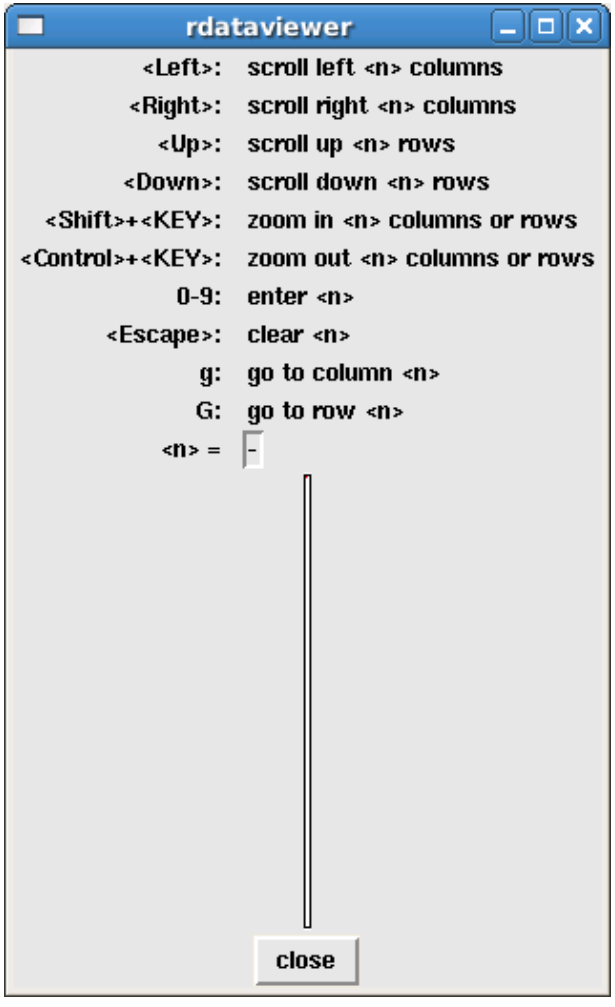
```
<LOADPROFILE RETAILER="MEEN" TIMESTAMPSTART="01/10/2007 00:00:00" TIMESTAMPEND="30/10/2008 23:59:59">
    <METER ICP="XXXXXXXX" SERIALNUMBER="XXXXXXXX" TARIFF="IN5" MRU="XXXXXXXX" UOM="kWh">
        <READING TIMESTAMP="2007-10-01 00:00:00" RAWREADING="0.723" QUALITYFLAG="7"/>
        <READING TIMESTAMP="2007-10-01 00:30:00" RAWREADING="0.683" QUALITYFLAG="7"/>
        <READING TIMESTAMP="2007-10-01 01:00:00" RAWREADING="0.747" QUALITYFLAG="7"/>
        <READING TIMESTAMP="2007-10-01 01:30:00" RAWREADING="0.654" QUALITYFLAG="7"/>
        <READING TIMESTAMP="2007-10-01 02:00:00" RAWREADING="0.707" QUALITYFLAG="7"/>
        <READING TIMESTAMP="2007-10-01 02:30:00" RAWREADING="0.853" QUALITYFLAG="7"/>
        <READING TIMESTAMP="2007-10-01 03:00:00" RAWREADING="0.667" QUALITYFLAG="7"/>
        <READING TIMESTAMP="2007-10-01 03:30:00" RAWREADING="0.474" QUALITYFLAG="7"/>
        <READING TIMESTAMP="2007-10-01 04:00:00" RAWREADING="0.182" QUALITYFLAG="7"/>
        <READING TIMESTAMP="2007-10-01 04:30:00" RAWREADING="0.058" QUALITYFLAG="7"/>
        ...
```

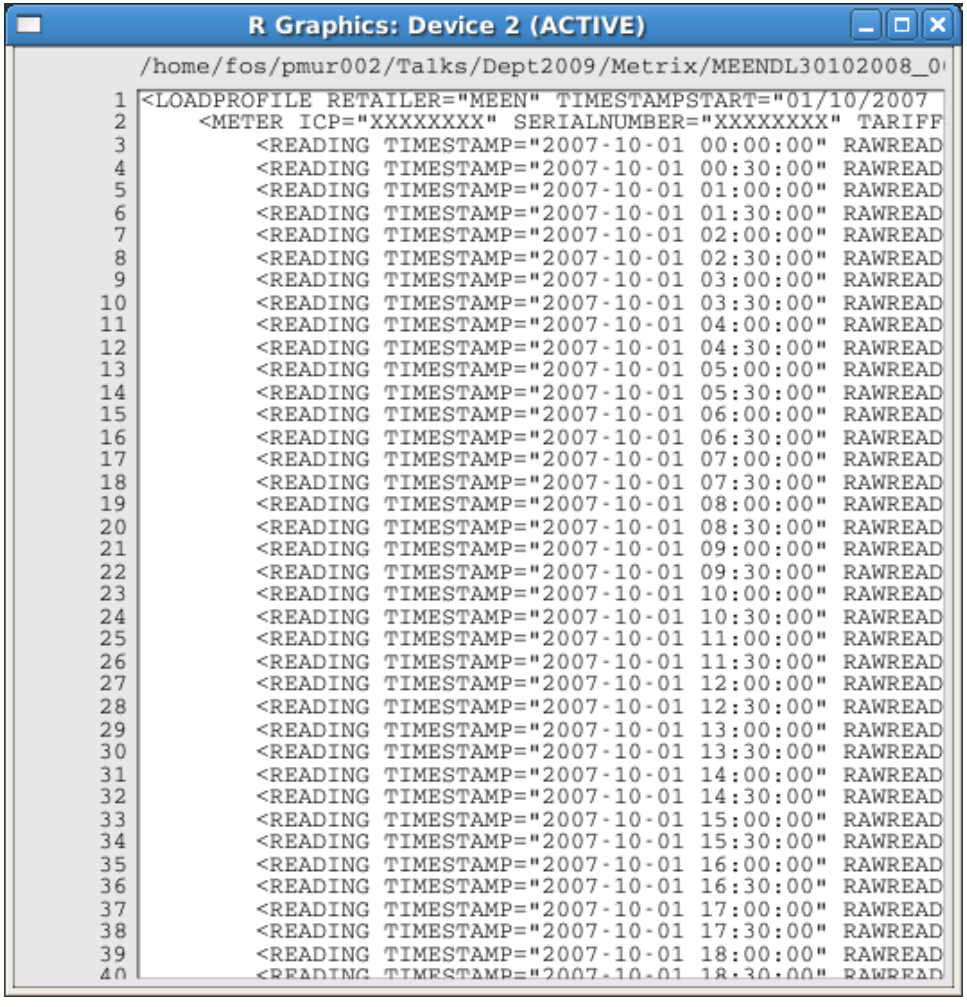
- `vi` took 3.5 minutes just to open and close the file.
- `Emacs` cannot open the file (buffer size exceeded).

The Meter Readings data set

```
data <- viewerDataText("MEENDL30102008_001-blind.XML",
                       index=TRUE)

vdvp <- viewerDeviceVp(data)

v <- simpleViewer(data, dev=vdvp)

tcltkViewer(v)
```

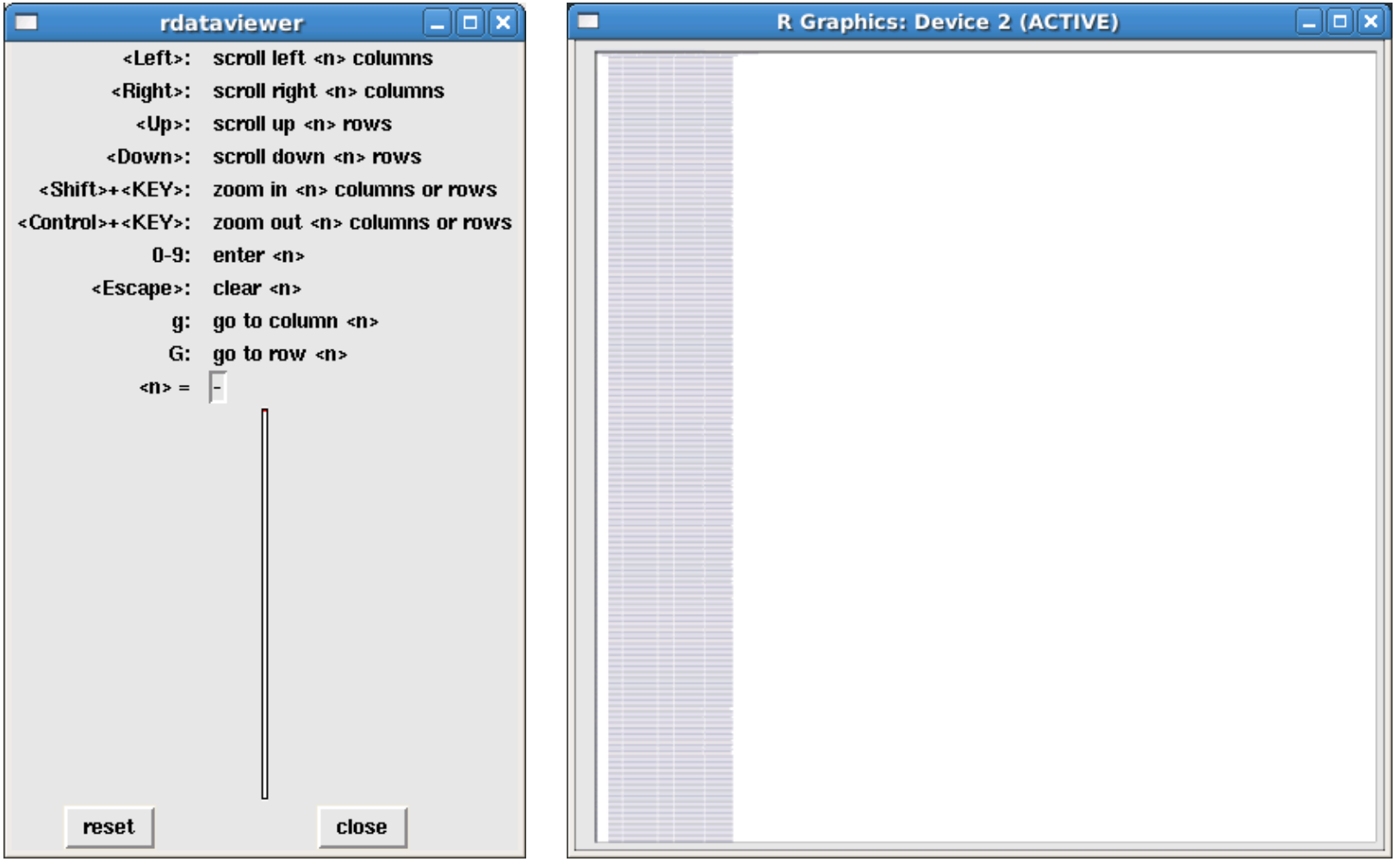
The Meter Readings data set





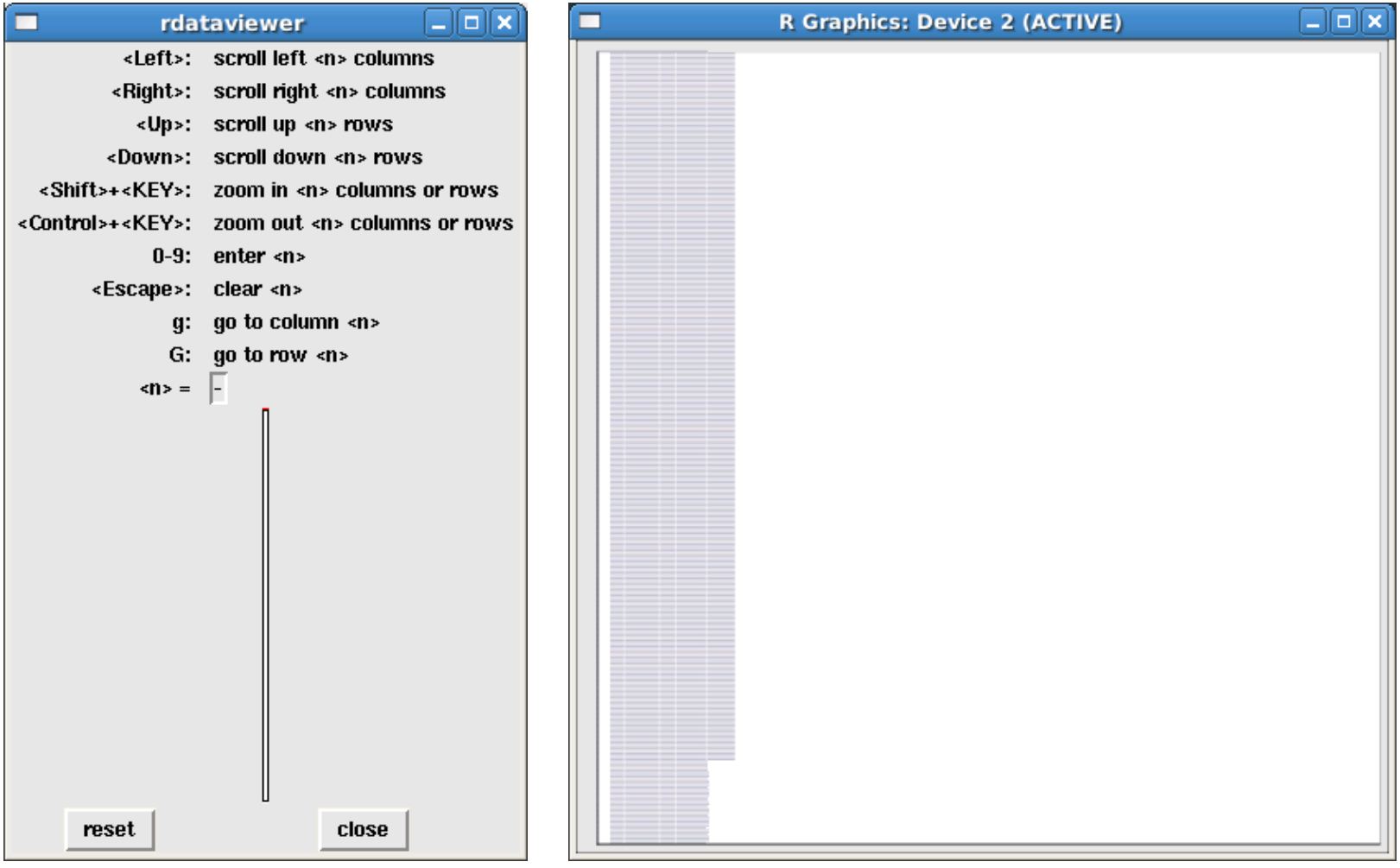
Zooming (Ctrl-PgUp Ctrl-PgUp Ctrl-PgUp)

Regular pattern discernable (stable values)



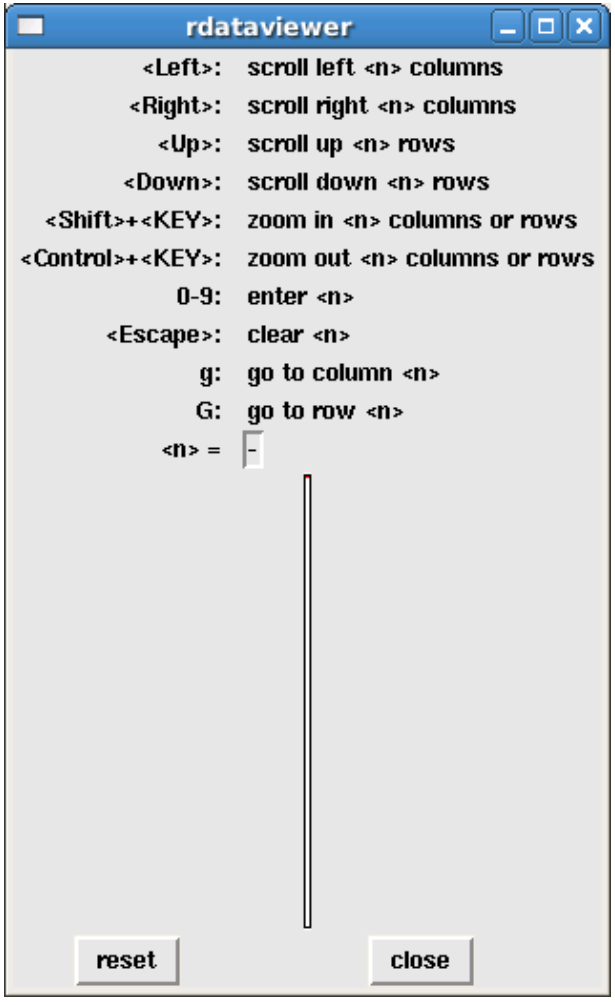
Zooming (PgDn x7)

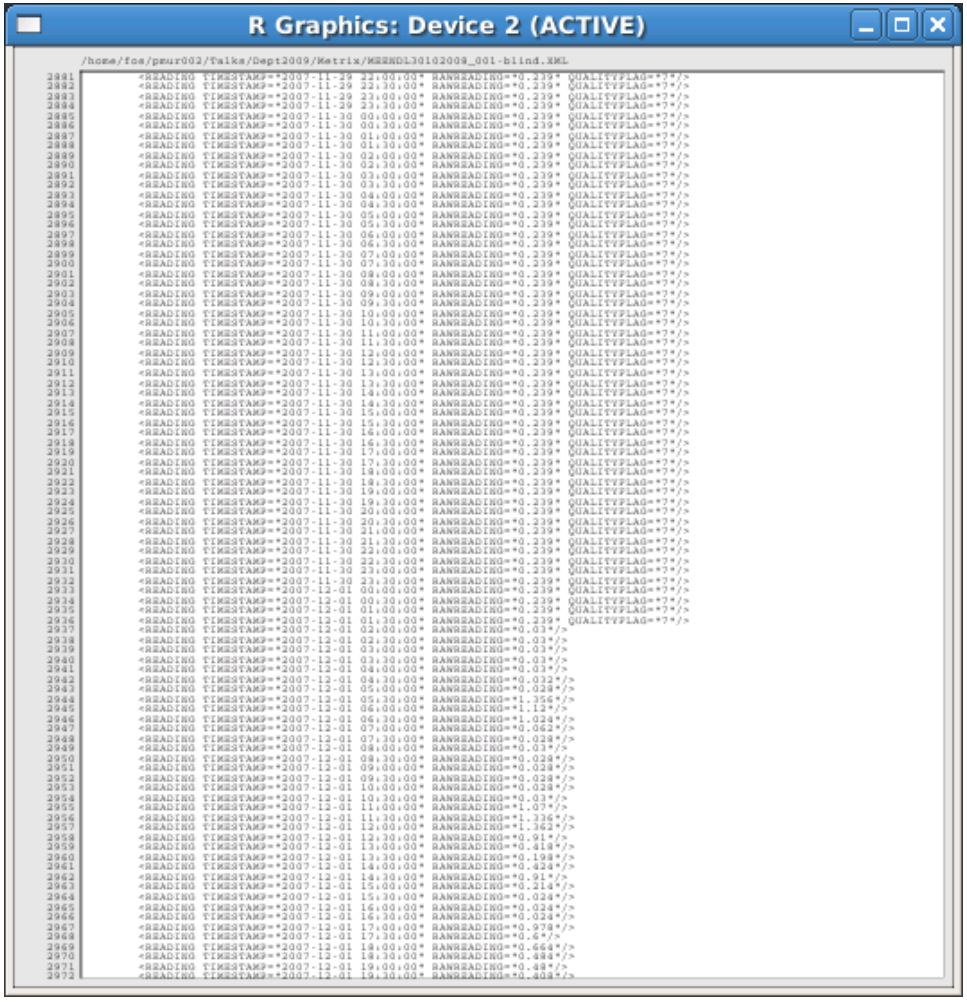
Change in pattern discernable



Zooming (Shift-PgUp Shift-PgUp)

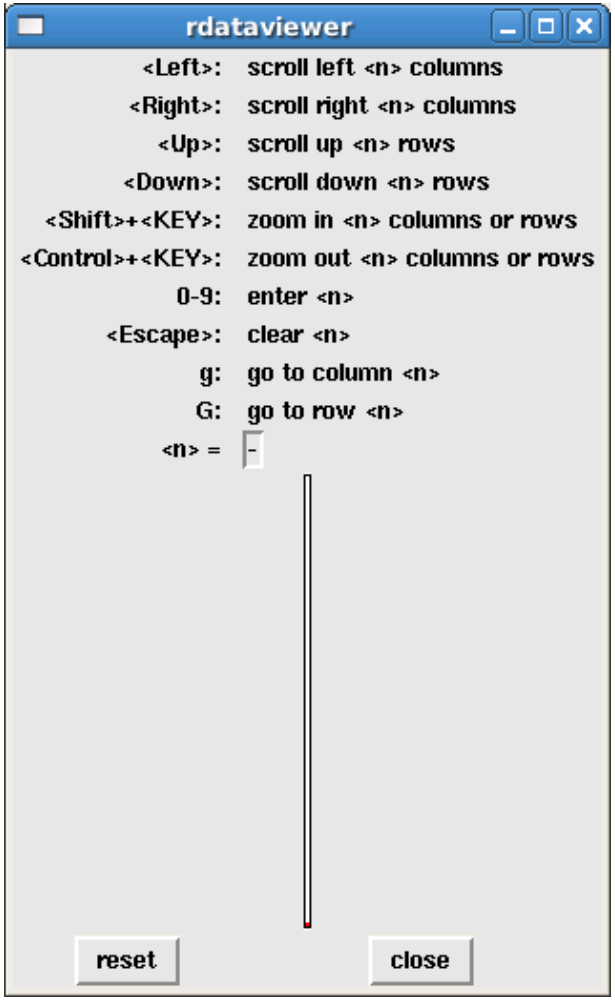
Disappearance of `QUALITYFLAG` attributes.

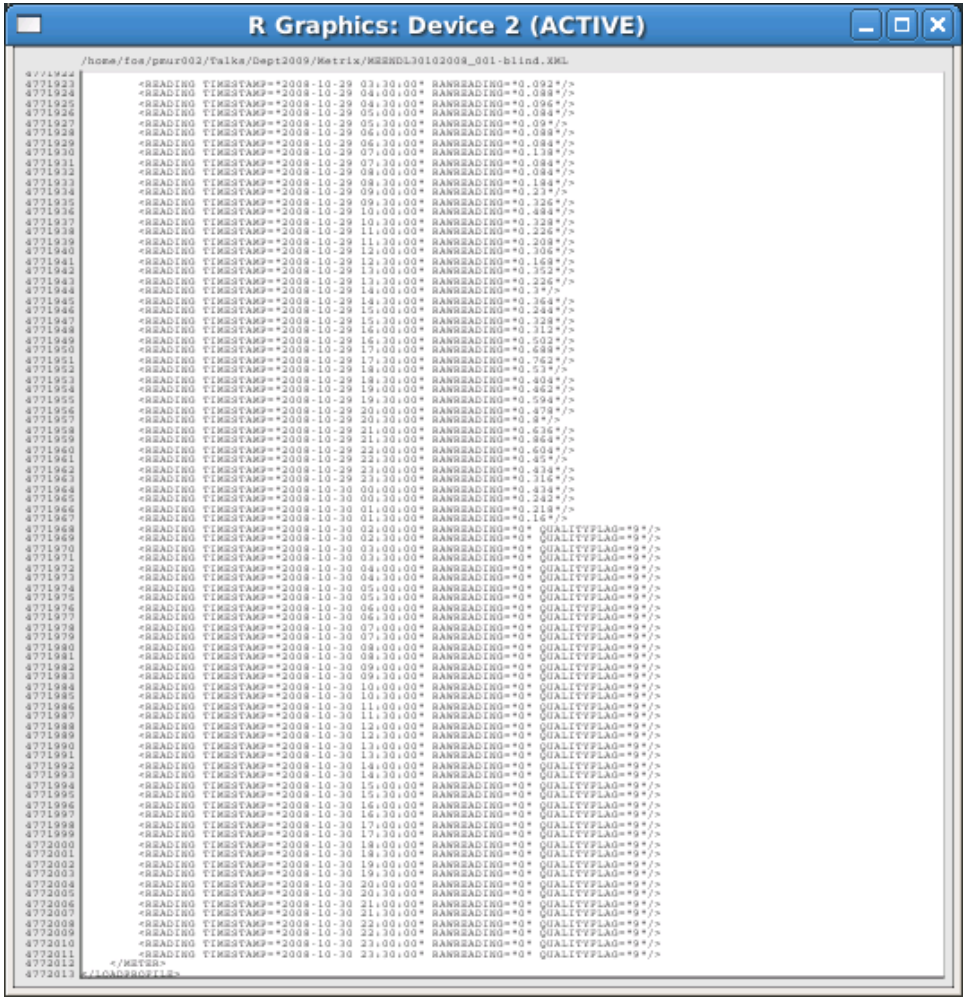




Navigation (Ctrl-End)

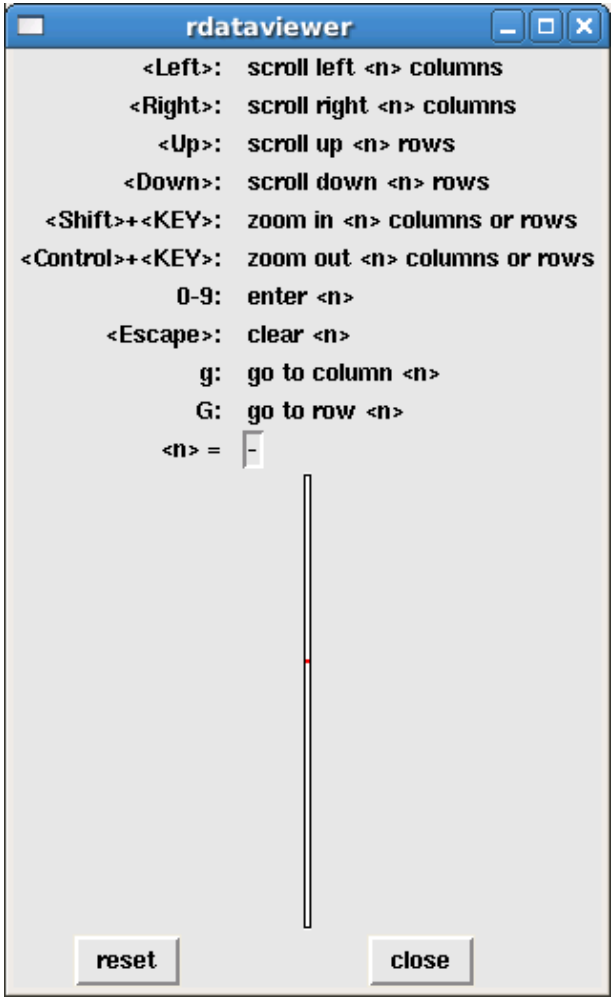
Instant navigation to end of file.

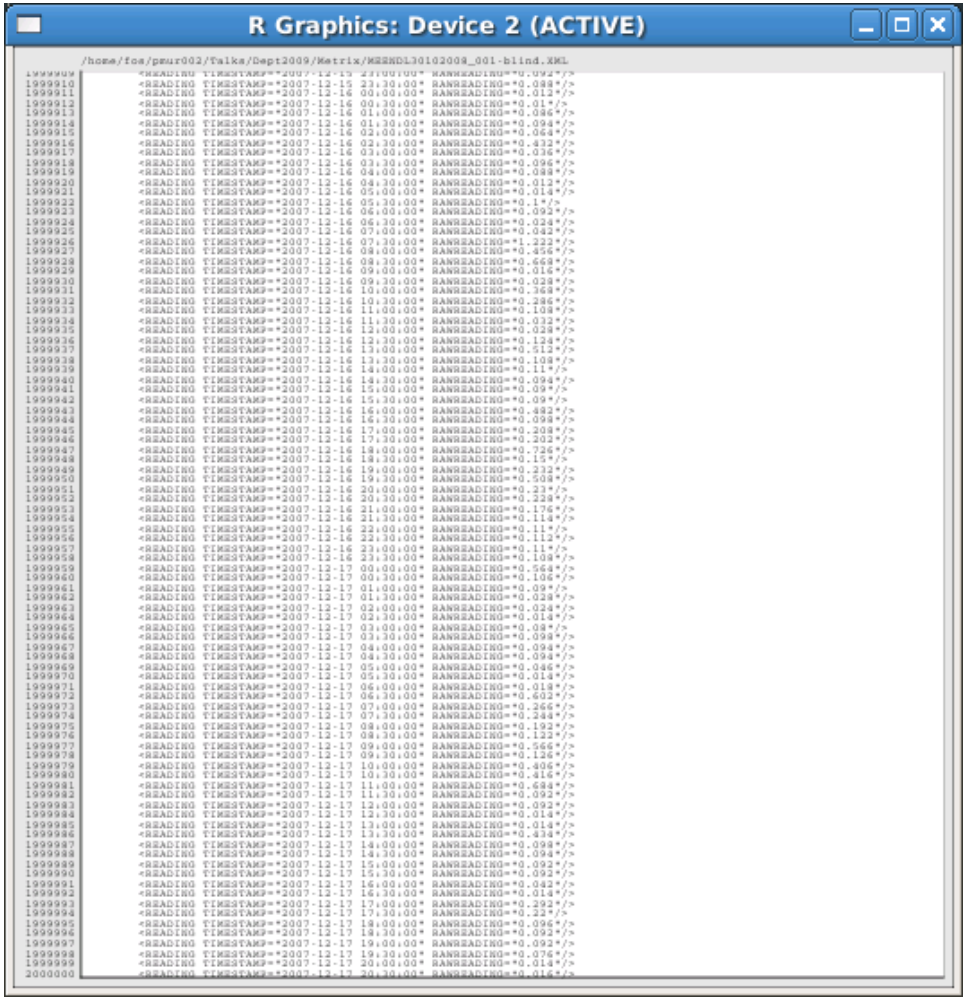




Navigation (2000000 G)

Instant navigation to middle of file.





The ViewerDataMySQL class

Methods for viewing a (MySQL) databases query:

```
result <- dbGetQuery("SELECT * FROM (<query>) LIMIT 1")
```

- How many rows and columns?

  ```
  SELECT COUNT(*) FROM (<query>);
  ```

- How wide is each column?

  ```
  dbColumnInfo(result)$len
  ```

- What are the column names?

  ```
  dbColumnInfo(result)$name
  ```

- Provide a text representation of given rows and columns

  ```
  SELECT columns FROM (<query>) LIMIT rows;
  ```

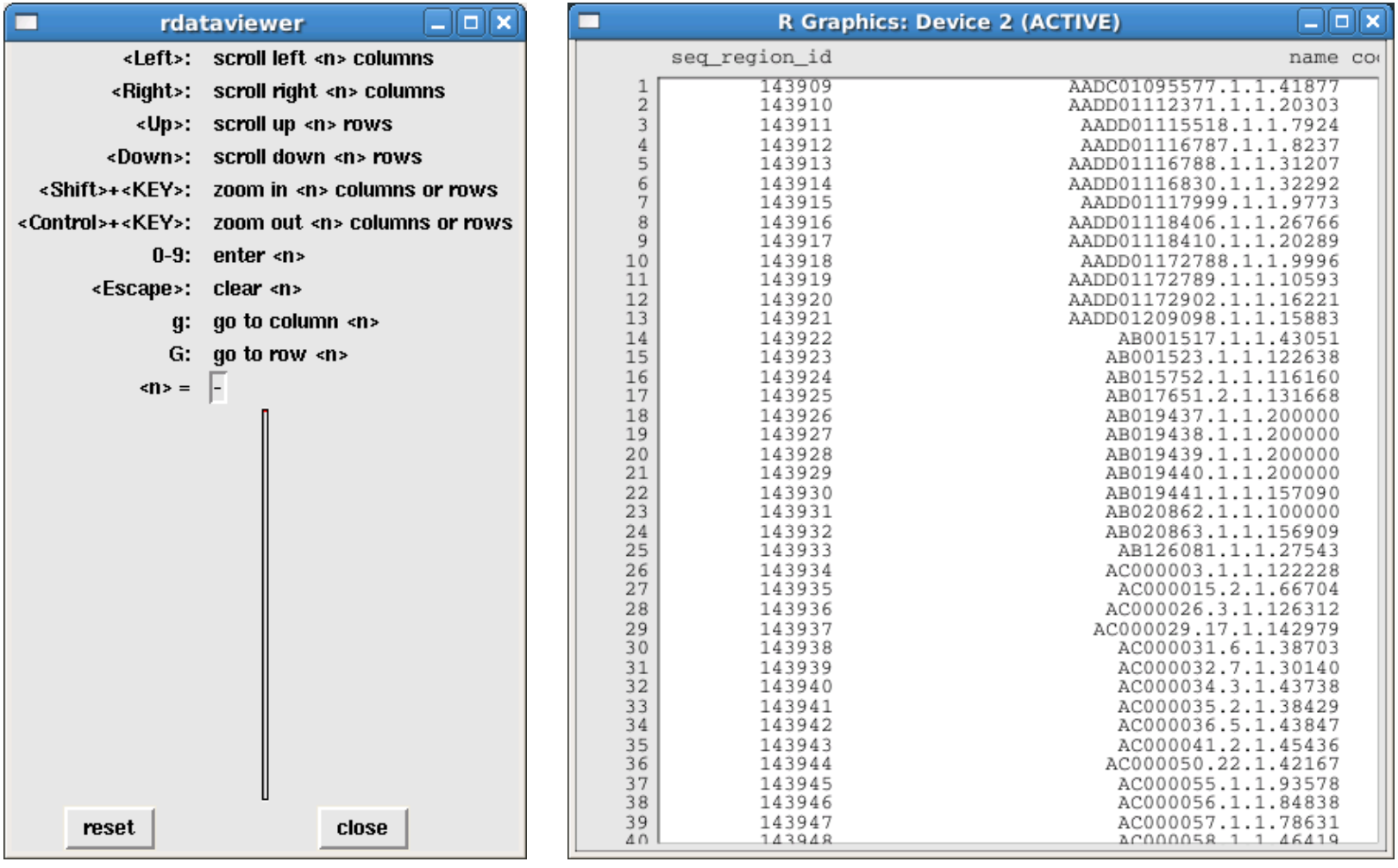
The Ensembl data set

- MySQL database `homo_sapiens_core_46_36h` containing human genome.
- `seq_region` table contains 54695 "sequence regions".

seq_region_id	name	coord_system_id	length
143909	AADC01095577.1.1.41877	4	41877
143910	AADD01112371.1.1.20303	4	20303
143911	AADD01115518.1.1.7924	4	7924
143912	AADD01116787.1.1.8237	4	8237
143913	AADD01116788.1.1.31207	4	31207
143914	AADD01116830.1.1.32292	4	32292
143915	AADD01117999.1.1.9773	4	9773
143916	AADD01118406.1.1.26766	4	26766
143917	AADD01118410.1.1.20289	4	20289
143918	AADD01172788.1.1.9996	4	9996
...	...	...	...

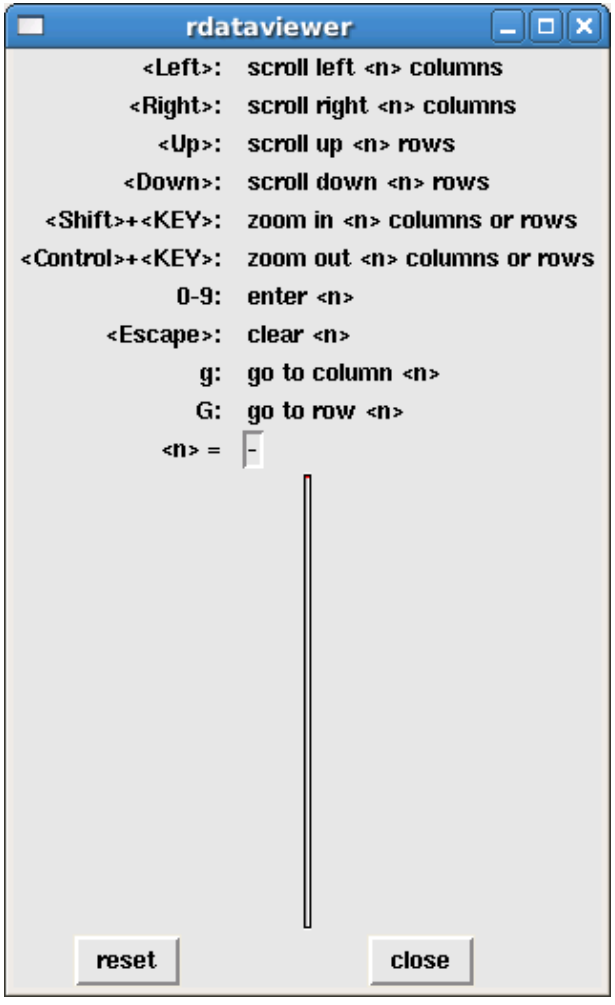
```
data <- viewerDataMySQL("select * from seq_region",
                        "homo_sapiens_core_46_36h",
                        "anonymous",
                        host="ensembldb.ensembl.org")
```

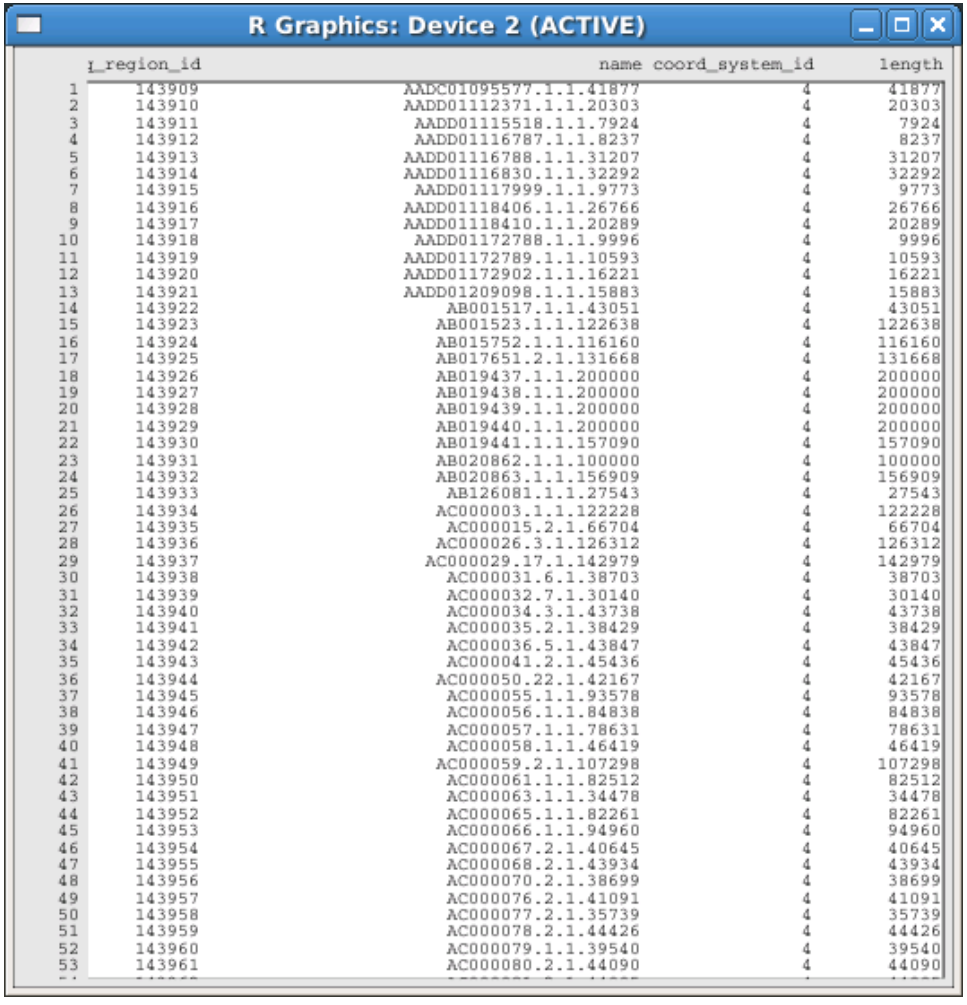
The Ensembl data set



Zoom out (2 Ctrl-Right)

View all columns at once.

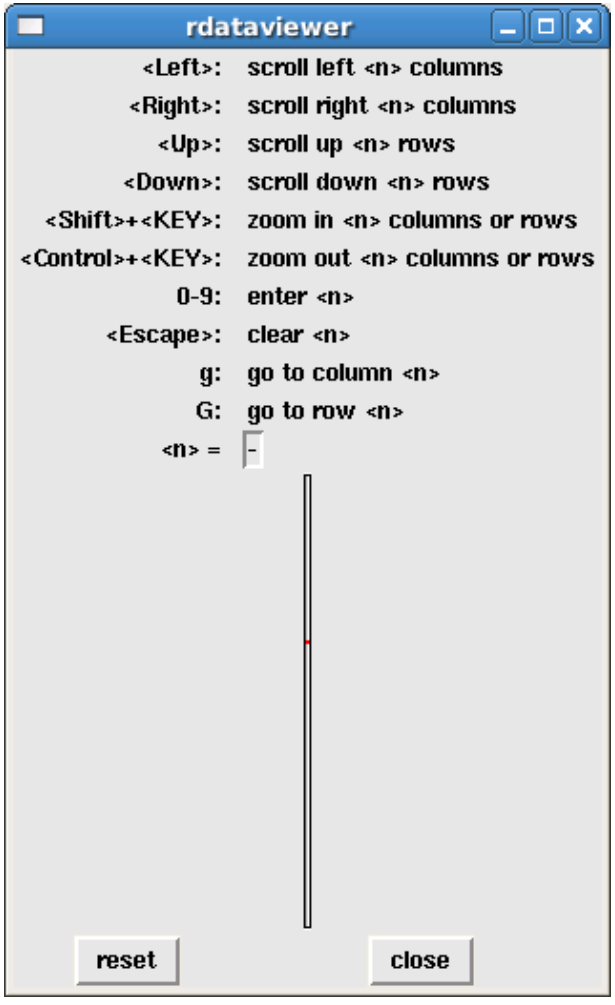


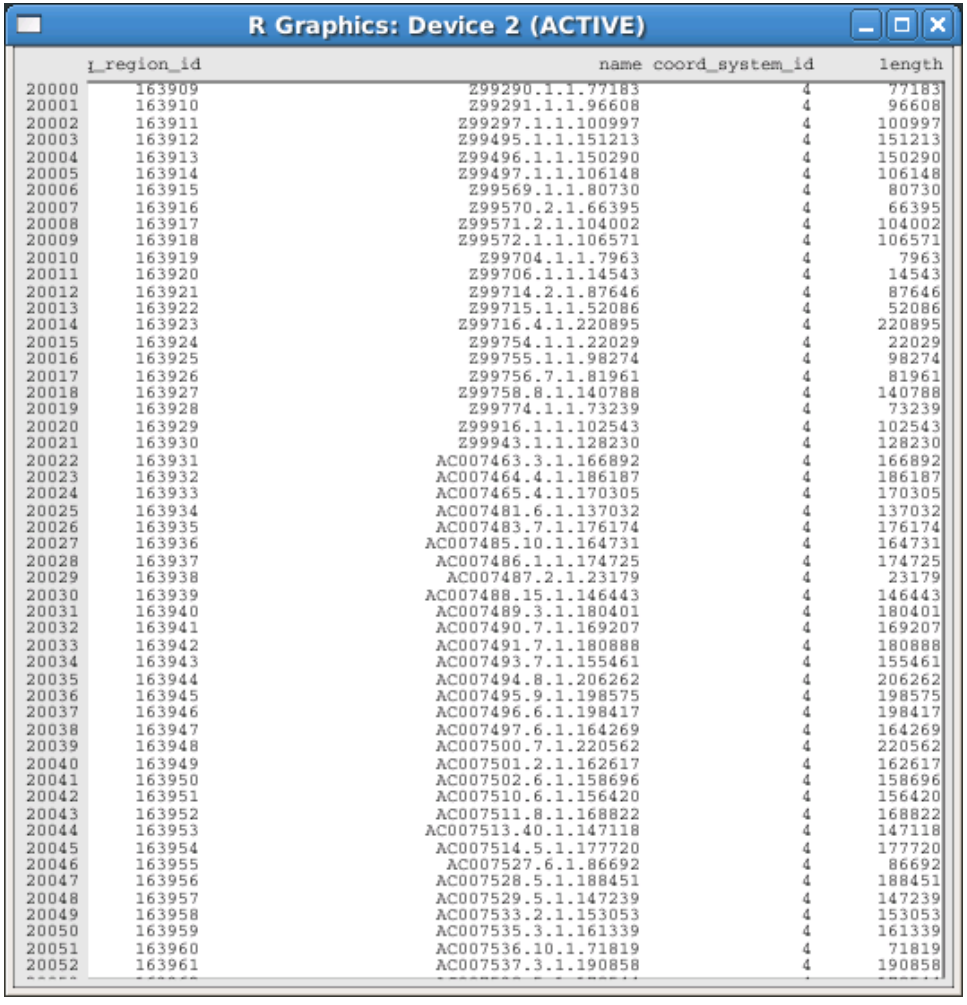


	i_region_id	name	coord_system_id	length
1	143909	AADC01095577.1.1.41877	4	41877
2	143910	AADD01112371.1.1.20303	4	20303
3	143911	AADD01115518.1.1.7924	4	7924
4	143912	AADD01116787.1.1.8237	4	8237
5	143913	AADD01116788.1.1.31207	4	31207
6	143914	AADD01116830.1.1.32292	4	32292
7	143915	AADD01117999.1.1.9773	4	9773
8	143916	AADD01118406.1.1.26766	4	26766
9	143917	AADD01118410.1.1.20289	4	20289
10	143918	AADD01172788.1.1.9996	4	9996
11	143919	AADD01172789.1.1.10593	4	10593
12	143920	AADD01172902.1.1.16221	4	16221
13	143921	AADD01209098.1.1.15883	4	15883
14	143922	AB001517.1.1.43051	4	43051
15	143923	AB001523.1.1.122638	4	122638
16	143924	AB015752.1.1.116160	4	116160
17	143925	AB017651.2.1.131668	4	131668
18	143926	AB019437.1.1.200000	4	200000
19	143927	AB019438.1.1.200000	4	200000
20	143928	AB019439.1.1.200000	4	200000
21	143929	AB019440.1.1.200000	4	200000
22	143930	AB019441.1.1.157090	4	157090
23	143931	AB020862.1.1.100000	4	100000
24	143932	AB020863.1.1.156909	4	156909
25	143933	AB126081.1.1.27543	4	27543
26	143934	AC000003.1.1.122228	4	122228
27	143935	AC000015.2.1.66704	4	66704
28	143936	AC000026.3.1.126312	4	126312
29	143937	AC000029.17.1.142979	4	142979
30	143938	AC000031.6.1.38703	4	38703
31	143939	AC000032.7.1.30140	4	30140
32	143940	AC000034.3.1.43738	4	43738
33	143941	AC000035.2.1.38429	4	38429
34	143942	AC000036.5.1.43847	4	43847
35	143943	AC000041.2.1.45436	4	45436
36	143944	AC000050.22.1.42167	4	42167
37	143945	AC000055.1.1.93578	4	93578
38	143946	AC000056.1.1.84838	4	84838
39	143947	AC000057.1.1.78631	4	78631
40	143948	AC000058.1.1.46419	4	46419
41	143949	AC000059.2.1.107298	4	107298
42	143950	AC000061.1.1.82512	4	82512
43	143951	AC000063.1.1.34478	4	34478
44	143952	AC000065.1.1.82261	4	82261
45	143953	AC000066.1.1.94960	4	94960
46	143954	AC000067.2.1.40645	4	40645
47	143955	AC000068.2.1.43934	4	43934
48	143956	AC000070.2.1.38699	4	38699
49	143957	AC000076.2.1.41091	4	41091
50	143958	AC000077.2.1.35739	4	35739
51	143959	AC000078.2.1.44426	4	44426
52	143960	AC000079.1.1.39540	4	39540
53	143961	AC000080.2.1.44090	4	44090

Navigation (20000 G)

Navigate to middle of query (with only small delay).





	i_region_id	name	coord_system_id	length
20000	163909	Z99290.1.1.77183	4	77183
20001	163910	Z99291.1.1.96608	4	96608
20002	163911	Z99297.1.1.100997	4	100997
20003	163912	Z99495.1.1.151213	4	151213
20004	163913	Z99496.1.1.150290	4	150290
20005	163914	Z99497.1.1.106148	4	106148
20006	163915	Z99569.1.1.80730	4	80730
20007	163916	Z99570.2.1.66395	4	66395
20008	163917	Z99571.2.1.104002	4	104002
20009	163918	Z99572.1.1.106571	4	106571
20010	163919	Z99704.1.1.7963	4	7963
20011	163920	Z99706.1.1.14543	4	14543
20012	163921	Z99714.2.1.87646	4	87646
20013	163922	Z99715.1.1.52086	4	52086
20014	163923	Z99716.4.1.220895	4	220895
20015	163924	Z99754.1.1.22029	4	22029
20016	163925	Z99755.1.1.98274	4	98274
20017	163926	Z99756.7.1.81961	4	81961
20018	163927	Z99758.8.1.140788	4	140788
20019	163928	Z99774.1.1.73239	4	73239
20020	163929	Z99916.1.1.102543	4	102543
20021	163930	Z99943.1.1.128230	4	128230
20022	163931	AC007463.3.1.166892	4	166892
20023	163932	AC007464.4.1.186187	4	186187
20024	163933	AC007465.4.1.170305	4	170305
20025	163934	AC007481.6.1.137032	4	137032
20026	163935	AC007483.7.1.176174	4	176174
20027	163936	AC007485.10.1.164731	4	164731
20028	163937	AC007486.1.1.174725	4	174725
20029	163938	AC007487.2.1.23179	4	23179
20030	163939	AC007488.15.1.146443	4	146443
20031	163940	AC007489.3.1.180401	4	180401
20032	163941	AC007490.7.1.169207	4	169207
20033	163942	AC007491.7.1.180888	4	180888
20034	163943	AC007493.7.1.155461	4	155461
20035	163944	AC007494.8.1.206262	4	206262
20036	163945	AC007495.9.1.198575	4	198575
20037	163946	AC007496.6.1.198417	4	198417
20038	163947	AC007497.6.1.164269	4	164269
20039	163948	AC007500.7.1.220562	4	220562
20040	163949	AC007501.2.1.162617	4	162617
20041	163950	AC007502.6.1.158696	4	158696
20042	163951	AC007510.6.1.156420	4	156420
20043	163952	AC007511.8.1.168822	4	168822
20044	163953	AC007513.40.1.147118	4	147118
20045	163954	AC007514.5.1.177720	4	177720
20046	163955	AC007527.6.1.86692	4	86692
20047	163956	AC007528.5.1.188451	4	188451
20048	163957	AC007529.5.1.147239	4	147239
20049	163958	AC007533.2.1.153053	4	153053
20050	163959	AC007535.3.1.161339	4	161339
20051	163960	AC007536.10.1.71819	4	71819
20052	163961	AC007537.3.1.190858	4	190858

Summary

- The novel navigation ideas don't really work; they would probably just scare off the students.
- Being able to zoom is a small improvement on most basic viewers.
- Seeing the "shape" of the entire data structure is quite nice.
- It's fun to be able to navigate quickly around a massive file.
- The modular structure of the code encourages further development.

Available from `http://www.stat.auckland.ac.nz/~paul/`