Data aggregation

Dr. Jennifer (Jenny) Bryan
Department of Statistics and Michael Smith Laboratories
University of British Columbia

How to do <sthg> for various 'chunks' of your dataset

chunks are ...	my recommendation
rows, columns, etc. of matrices / arrays	`apply()`
groups induced by levels of ≥ 1 factor(s) -- vector	`aggregate()`
groups induced by levels of ≥ 1 factor(s) -- data.frame	`plyr::ddply()`
components of a list (remember data.frames are lists!)	`plyr::l*ply()`

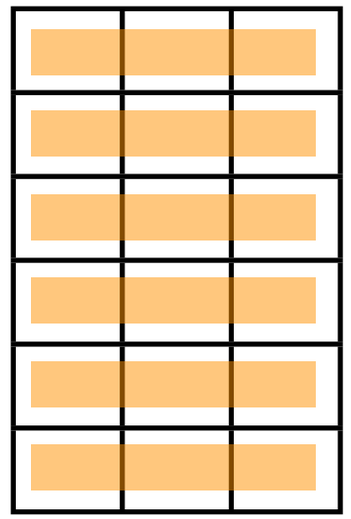
How to do <sthg> for various 'chunks' of your dataset

... using only built-in functions, i.e. no `plyr`

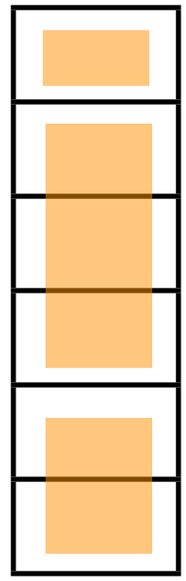
chunks are ...	relevant functions
rows, columns, etc. of matrices / arrays	`apply()`
components of a list (remember data.frames are lists!)	`sapply(),lapply()`

groups induced by levels of ≥ 1 factor(s)	`aggregate()`
	`tapply()`
	`by()`
	`split() + [sl]apply()`

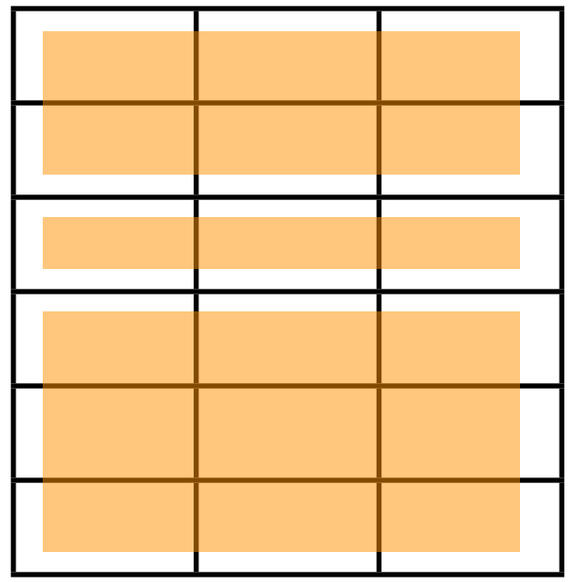
chunks are ...	my recommendation
rows, columns, etc. of matrices / arrays	`apply()`
groups induced by levels of ≥ 1 factor(s) -- vector	`aggregate()`
groups induced by levels of ≥ 1 factor(s) -- data.frame	`plyr::ddply()`
components of a list (remember data.frames are lists!)	`plyr::l*ply()`



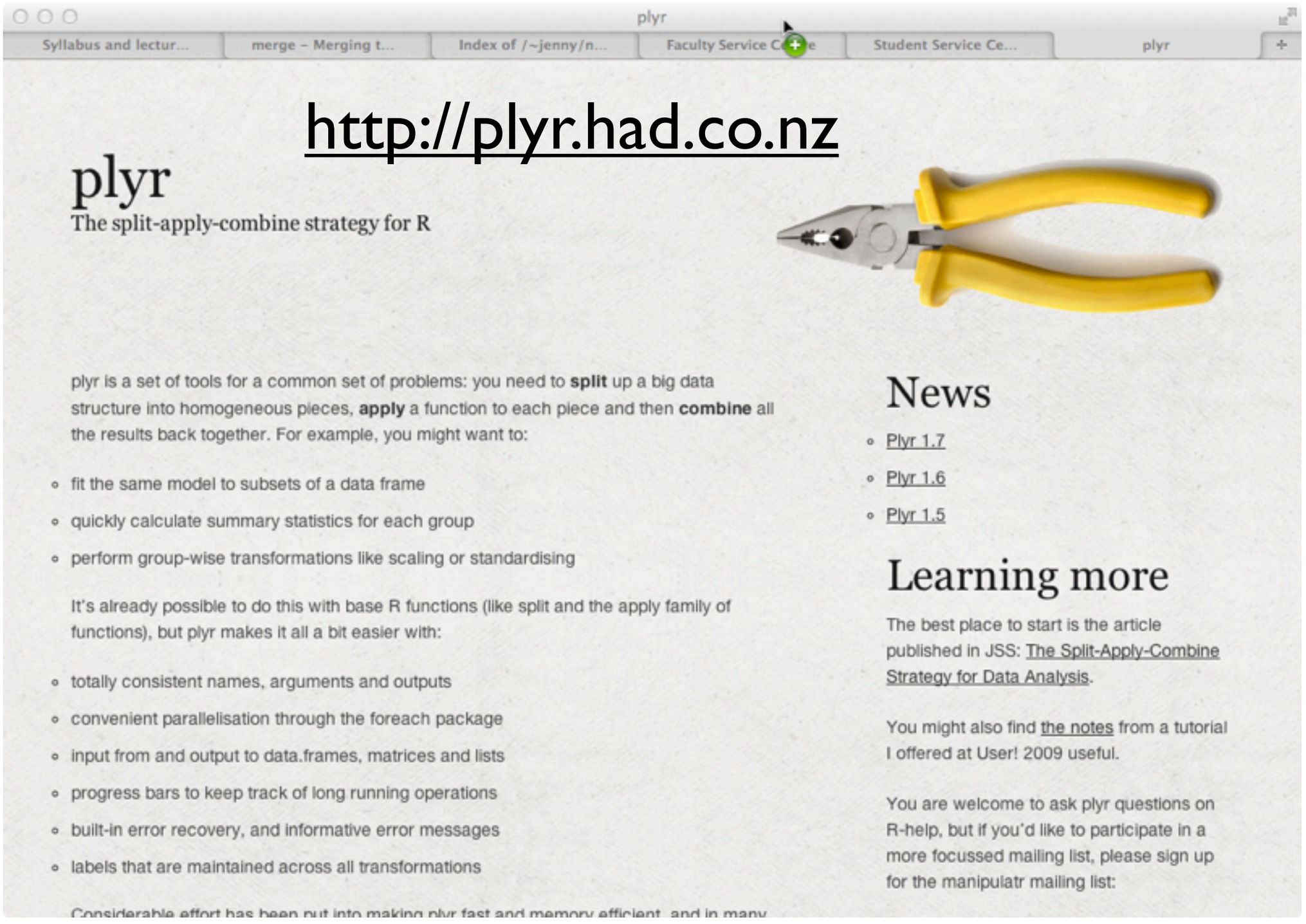
chunks are ...	my recommendation
rows, columns, etc. of matrices / arrays	`apply()`
groups induced by levels of ≥ 1 factor(s) -- vector	`aggregate()`
groups induced by levels of ≥ 1 factor(s) -- data.frame	`plyr::ddply()`
components of a list (remember data.frames are lists!)	`plyr::l*ply()`



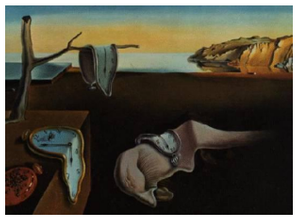
chunks are ...	my recommendation
rows, columns, etc. of matrices / arrays	`apply()`
groups induced by levels of ≥ 1 factor(s) -- vector	`aggregate()`
groups induced by levels of ≥ 1 factor(s) -- data.frame	`plyr::ddply()`
components of a list (remember data.frames are lists!)	`plyr::l*ply()`



The plyr package is what I advise long-term for data aggregation.



Hadley Wickham.
The split-apply-combine strategy for data analysis.
Journal of Statistical Software, vol. 40, no. 1, pp. 1–29, 2011.
http://www.jstatsoft.org/v40/i01/paper



Journal of Statistical Software

April 2011, Volume 40, Issue 1. *http://www.jstatsoft.org/*

The Split-Apply-Combine Strategy for Data Analysis

Hadley Wickham
Rice University

Abstract

Many data analysis problems involve the application of a split-apply-combine strategy, where you break up a big problem into manageable pieces, operate on each piece independently and then put all the pieces back together. This insight gives rise to a new R package that allows you to smoothly apply this strategy, without having to worry about the type of structure in which your data is stored.

The paper includes two case studies showing how these insights make it easier to work with batting records for veteran baseball players and a large 3d array of spatio-temporal ozone measurements.

Keywords: R, apply, split, data analysis.

split apply combine

Input \ *Output*	Array	Data frame	List	Discarded
Array	`aaply`	`adply`	`alply`	`a_ply`
Data frame	`daply`	`ddply`	`dlply`	`d_ply`
List	`laply`	`ldply`	`llply`	`l_ply`

- `a*ply(.data, .margins, .fun, ..., .progress = "none")`
- `d*ply(.data, .variables, .fun, ..., .progress = "none")`
- `l*ply(.data, .fun, ..., .progress = "none")`

Take this data.frame ...

divide it into baby data.frames, based on this factor and ...

apply this function to each chunk ...

```
ddply(.data, .variables, .fun = NULL)
```

... glue the results back together and return as a data.frame