expanding exposure through early introduction in the undergraduate curriculum

mine çetinkaya-rundel
duke university

intro stats
(general ed)

extra-
curriculars

stats major
+ other
quantitative
fields

intro stats
(general ed)

why R?

unlike other software designed specifically for courses at this level

free & open source

powerful & flexible

relevant beyond intro stat

why not R?

challenge of teaching programming in addition to stats concepts

command line more intimidating than GUI

don’t do any hands on data analysis

disservice to everyone involved

challenge of teaching programming in addition to stats concepts

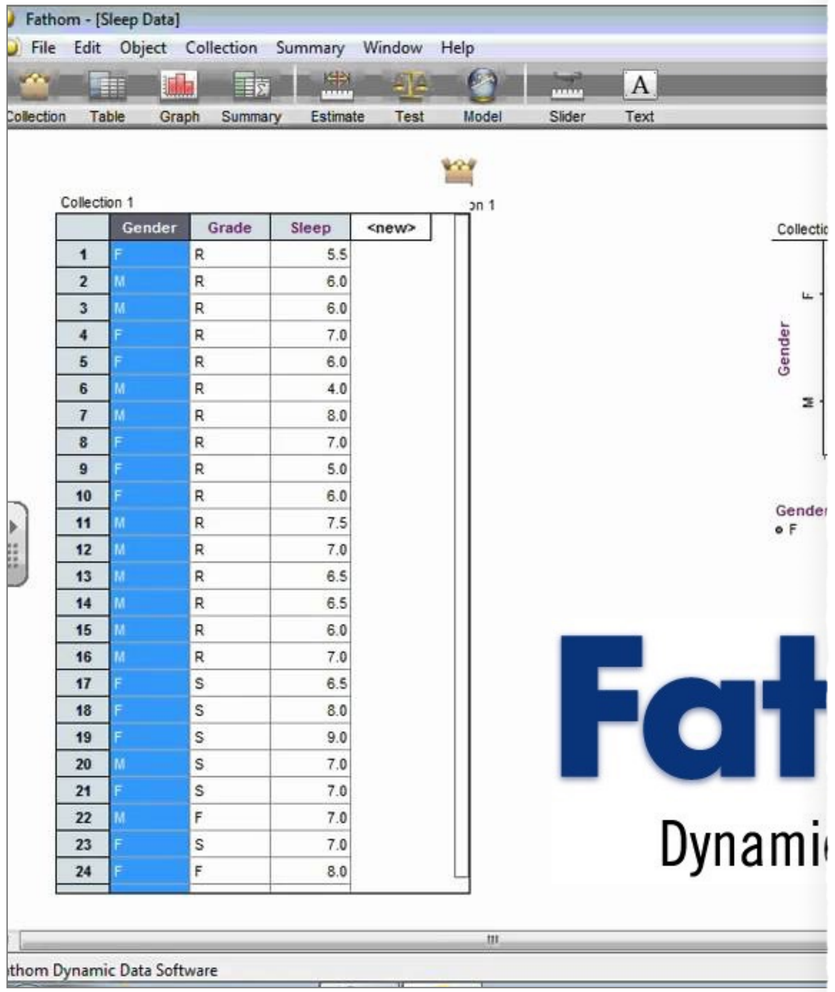
challenge of teaching programming in addition to stats concepts

don’t do any hands on data analysis

disservice to everyone involved

use a drag-and-drop type tool

there’s still a learning curve



III. Adding Proportions to Summary Table

For categorical variables, you should see the counts of each possible outcome of that variable in the **Summary Table**. To see the breakdown of proportions or percentages, follow these steps:

- Click on the **Summary Table** to highlight it, click on the "**Summary**" drop-down menu and select "**Add Formula**". In general, whenever you click and select a *Fathom* object (such as a **Table**, **Graph**, or **Summary**) the menu at the top of the screen will change to give you options for working on that object.

- In the formula editor that pops up, type "*rowproportion*" (without the quotes) to see the row proportions or "*columnproportion*" to see the column proportions. Be sure to spell the names of the formulas correctly or else *Fathom* will give you an error. (If you spell the names correctly, they should change to a purplish color in your editor.)

- You will see that each cell in the **Summary Table** now includes numbers for multiple statistics. To see which numbers correspond with which statistics, simply look at the bottom of your summary table to see the order of the statistics or formulas within each cell.

- To delete (or change) a particular statistic from the table, you can double click on its name at the bottom of the **Summary Table**. In the formula editor, press delete (or make your changes) and then click "**OK**".

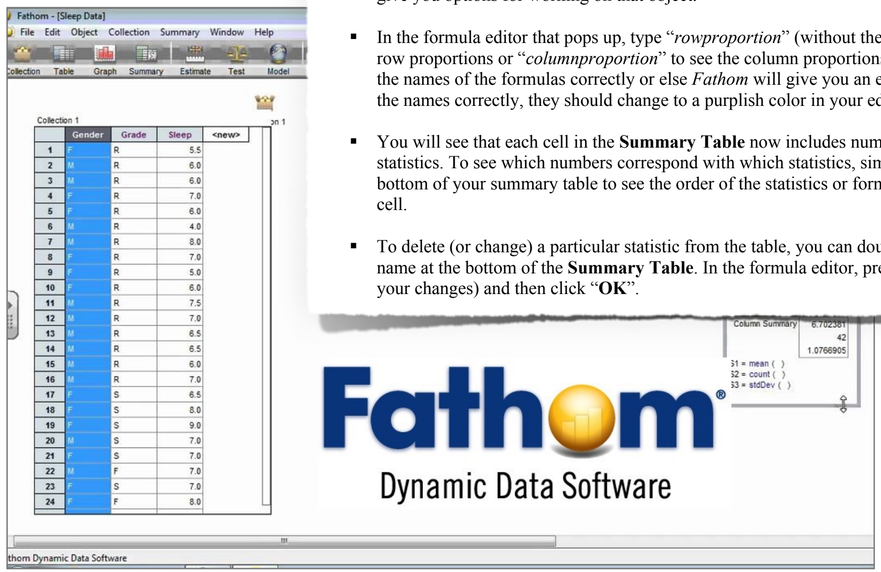
challenge of teaching programming in addition to stats concepts

don’t do any hands on data analysis

disservice to everyone involved

use a drag-and-drop type tool

there’s still a learning curve

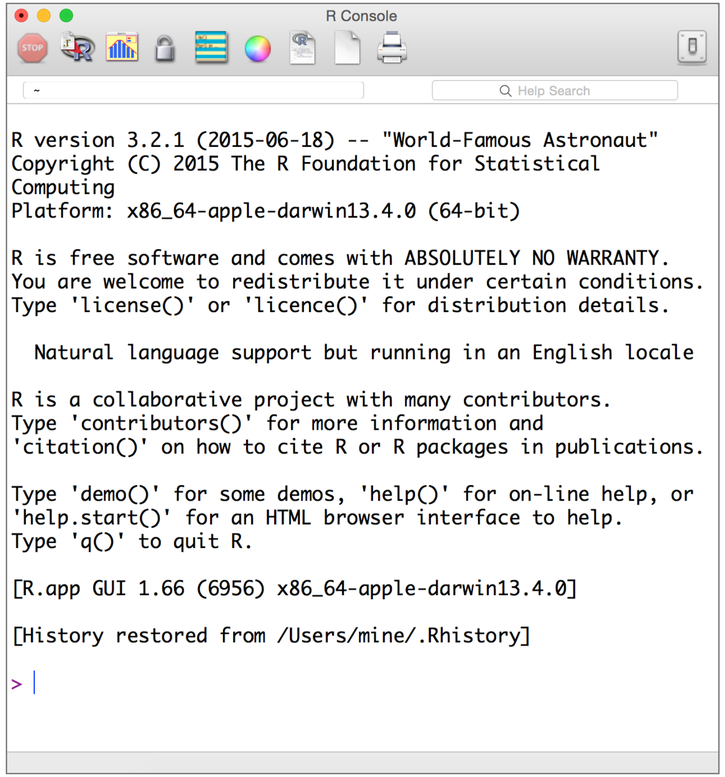


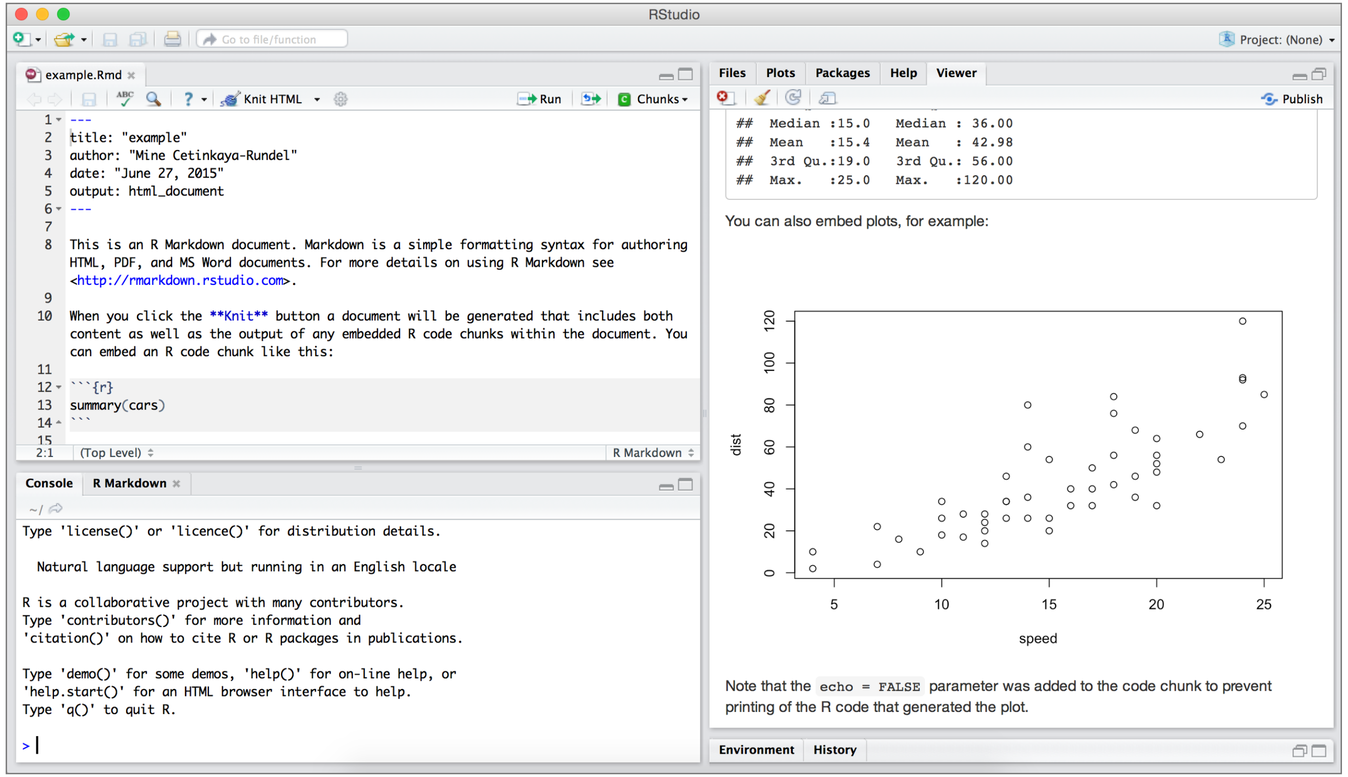
III. Adding Proportions to Summary Table

For categorical variables, you should see the counts of each possible outcome of that variable in the **Summary Table**. To see the breakdown of proportions or percentages, follow these steps:

- Click on the **Summary Table** to highlight it, click on the “**Summary**” drop-down menu and select “**Add Formula**”. In general, whenever you click and select a *Fathom* object (such as a **Table**, **Graph**, or **Summary**) the menu at the top of the screen will change to give you options for working on that object.
- In the formula editor that pops up, type “*rowproportion*” (without the quotes) to see the row proportions or “*columnproportion*” to see the column proportions. Be sure to spell the names of the formulas correctly or else *Fathom* will give you an error. (If you spell the names correctly, they should change to a purplish color in your editor.)
- You will see that each cell in the **Summary Table** now includes numbers for multiple statistics. To see which numbers correspond with which statistics, simply look at the bottom of your summary table to see the order of the statistics or formulas within each cell.
- To delete (or change) a particular statistic from the table, you can double click on its name at the bottom of the **Summary Table**. In the formula editor, press delete (or make your changes) and then click “**OK**”.

command line more intimidating than GUI



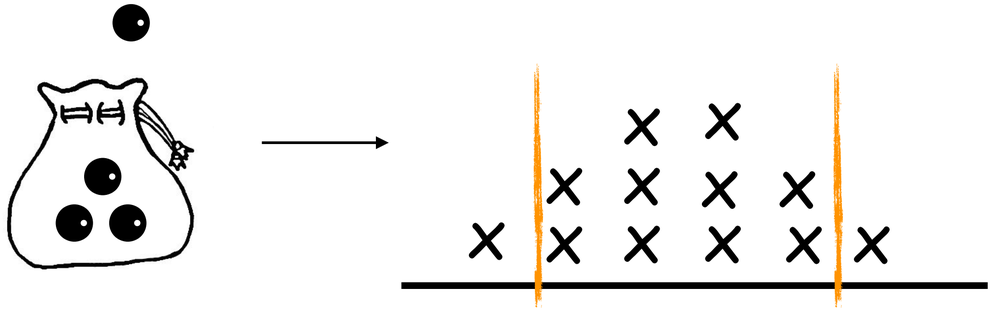


how R?

balance: teach code as a way of introducing/ reinforcing concepts

bootstrapping

physical representation



computation

```
for(i in 1:15000){
  samp <- sample(mydata, size = n)
  boot_dist[i] <- mean(mydata)
}
hist(boot_dist)
```

how R?

balance:
minimize
code for
implementation
outside the
scope of course

how R?

computation to help solve real world problems

visualization of multiple variables at once

cleaning of real not-very-small data

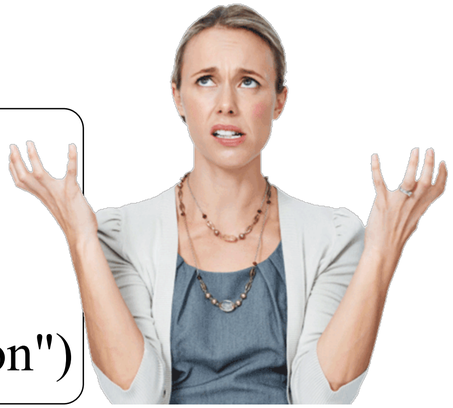
modeling

…

how R?

consistent visual syntax highlighting

```
n <- 1000
p <- seq(0, 1, 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
plot(me ~ p, ylab = "Margin of Error", xlab = "Population Proportion")
```



```r
n <- 1000
p <- seq(0, 1, 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
plot(me ~ p, ylab = "Margin of Error", xlab = "Population Proportion")
```

how R?

as little overhead as possible

avoid local installation

RStudio server (via university login)

preinstalled & preloaded packages

how R?

reproducible via literate programming

goal: train new researchers whose only workflow is a reproducible one

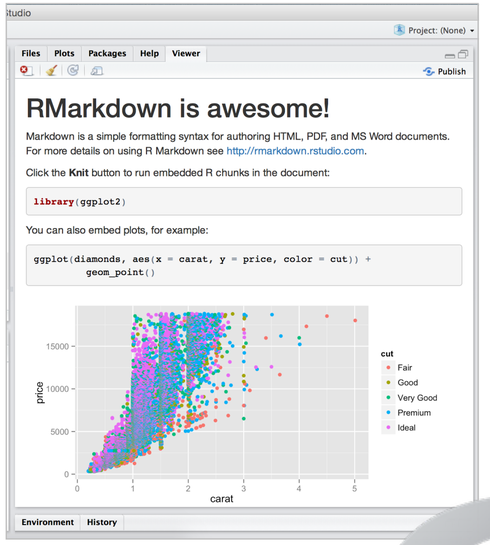
don't touch the raw data

keep track of all analysis steps

avoid copy-paste

toolkit:

RStudio +



= Literate programming in

how R?

with lots of scaffolding

start with templates including code and answers

slowly remove handholding

R Markdown learning outcomes (beyond reproducibility)

learn R

built-in and consistent syntax highlighting

code and output always together

avoid the messy / frustrating console

feedback + grading

ambiguity removed

collaboration

Room for improvement

intro stats
(general ed)

extra-
curriculars

stats major
+ other
quantitative
fields

ASA DataFest™

weekend-long data analysis competition

not an R event, but R is most popular

pre-event workshops

WORKSHOPS

These workshops are recommended for DataFest participants, but they're open to everyone.
All workshops will be held in the The Edge Workshop Room.

» *Intro to R*

slides | source

Monday, March 16, 6-8pm

Instructors: Gary Larson and Monika Hu, Duke StatSci

Description: Introduction to R as a statistical programming language. This session will introduce the basics of R syntax, getting data into R, various data types and classes, etc. The session assumes no or little background in R.

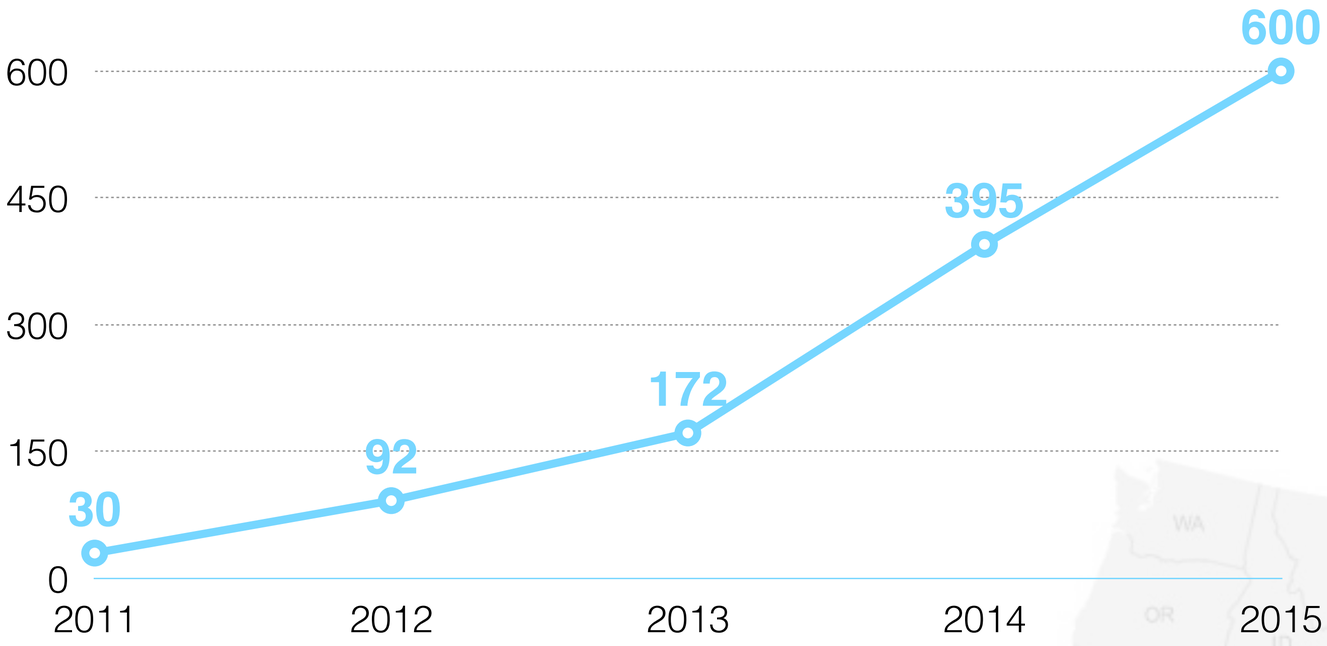
» *Data munging with R and dplyr*

slides | solutions | source

Wednesday, March 18, 6-8pm

Instructor: Prof. Colin Rundel, Duke StatSci

Description: This session will demonstrate tools for data manipulation and cleaning of data in R. Majority of the session will use the dplyr and tidyr packages. Some background in R is recommended.



PURDUE UNIVERSITY

SMITH COLLEGE

AMHERST COLLEGE

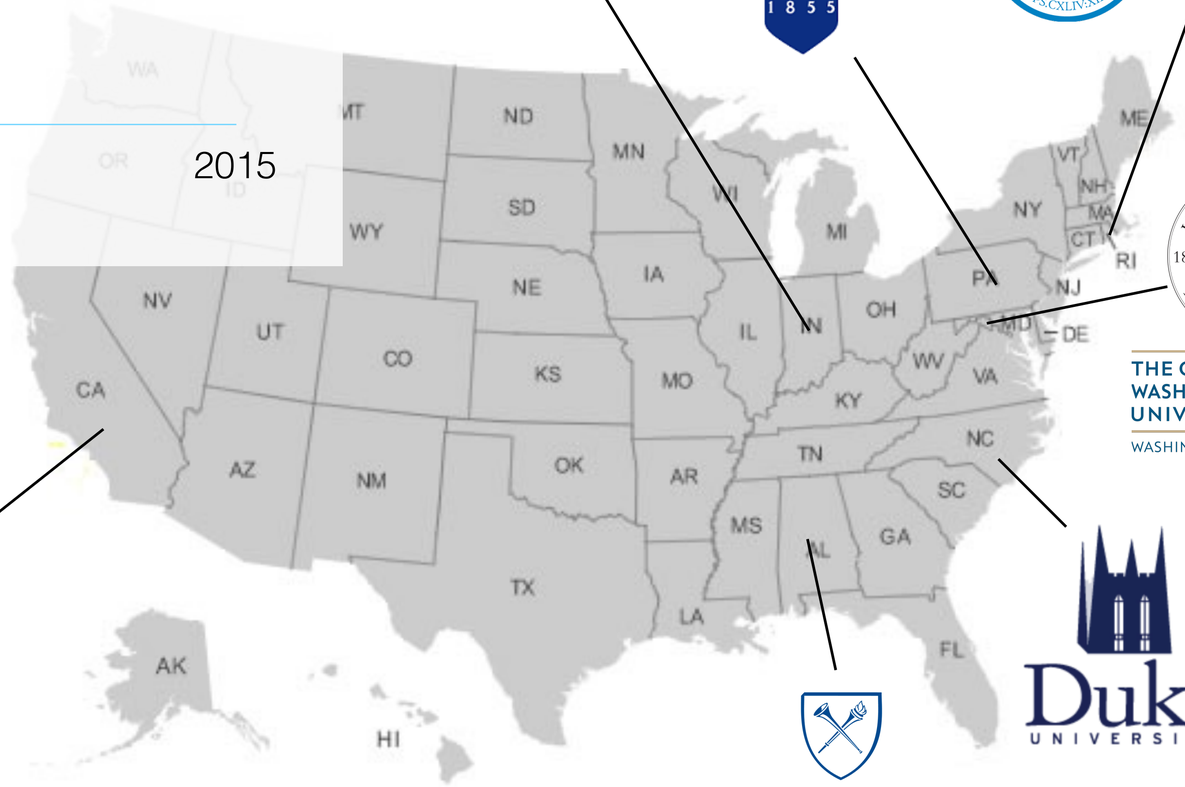
PENNSTATE

1855

MOUNT HOLYOKE COLLEGE

UMASS AMHERST

HAMPSHIRE COLLEGE



UNIVERSITY OF MARYLAND

1856

1789

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

GEORGETOWN UNIVERSITY

ucla

Pomona College

UC RIVERSIDE UNIVERSITY OF CALIFORNIA

CALIFORNIA STATE UNIVERSITY LONG BEACH 1949

USC UNIVERSITY OF SOUTHERN CALIFORNIA

EMORY UNIVERSITY

Duke UNIVERSITY

NORTH CAROLINA CENTRAL UNIVERSITY 1910

UNCG

MICHIGAN

North Carolina Agricultural and Technical State University

ASA DataFest™

inform the curriculum

learning computation tends to be ad hoc

demand for formal stat computing course

focusing especially on “the details”

intro stats
(general ed)

extra-
curriculars

stats major
+ other
quantitative
fields

stats major
+ other quantitative fields

data science course for 1st year undergrads

same principles as intro stat

heavier emphasis on computation

focus: data wrangling + visualization

stats elective for majors

stat computing course for students with regression

focus: working with large & complex data

reach

students, beyond duke

coursera

teachers

open source resources

OpenIntro

openintro.org

mine-cetinkaya-rundel

workshops

Feedback: Resources for RStudio server setup

beyond the discipline

PROJECT TIER
TEACHING INTEGRITY IN EMPIRICAL RESEARCH

thank you!

comments / questions?

 mine@stat.duke.edu

 @minebocek

 mine-cetinkaya-rundel