Reproducible research in archaeology using R & rrtools

Sophie C. Schmidt

Ben Marwick

16 März, 2018

# Narrative to accompany the slides for our talk

These are our notes for preparing our slides for our CAA2018 presentation in the session “R as an archaeological tool: current state and directions” (<https://github.com/MartinHinz/isaak_conf_sessions/tree/master/caa_2018_tuebingen>)

## Slide: Title slide

*Reproducible research in archaeology using R & rrtools*: In this talk, we will show how archaeologists are changing their data analysis practices to make their research more reproducible and open. We will introduce the rrtools package, which is our specific contribution to help with the challenge of making reproducible research in archaeology easier.

## Acknowledgements

Our project comes from the group who took part in the 2017 Summer School on Reproducible Research in Landscape Archaeology at the Freie Universität Berlin (17-21 July), funded and jointly organized by Excellence Cluster 264 Topoi, the Collaborative Research Center 1266, and ISAAKiel.

## Slide: Motivation & Outline

*Motivation*

Our motivation for this topic arises from a profound change that we have been observing in how researchers analyse their data.

Increasingly, we are seeing researchers publishing research that uses a programming language to calculate their models. Similarly, we see journal articles accompanied by scripts of programming language so that readers can study the details of the data analysis methods.

This is an important change away from using mouse-driven, point-and-click software for data analysis. When archaeologists use software such as SPSS and Past (as you see in our screenshots here), it is very difficult for them to share the details of their calculations. When readers cannot see the details of an analysis, it is difficult for them to be confident in the results described in an article. Also, for other researchers it may be near impossible to implement the described methods in their own work as they lack the details on how to come to the desired conclusions. Thus, using scripting language offers the chance of accelerating archaeological research by providing a way to retracing and checking published analyses as well as explaining methods more efficiently. These are the main advantages of creating reproducible research, which we would like to encourage.

*Outline*

In this presentation we will briefly review some data showing change in the use of programming languages in the sciences generally, and in archaeology in particular. We focus on the R programming language, because this is by far the most widely used language in archaeology.

Secondly, we will review how R is used by archaeologists, to show the potential that using a programming language has for opening and accelerating archaeological research.

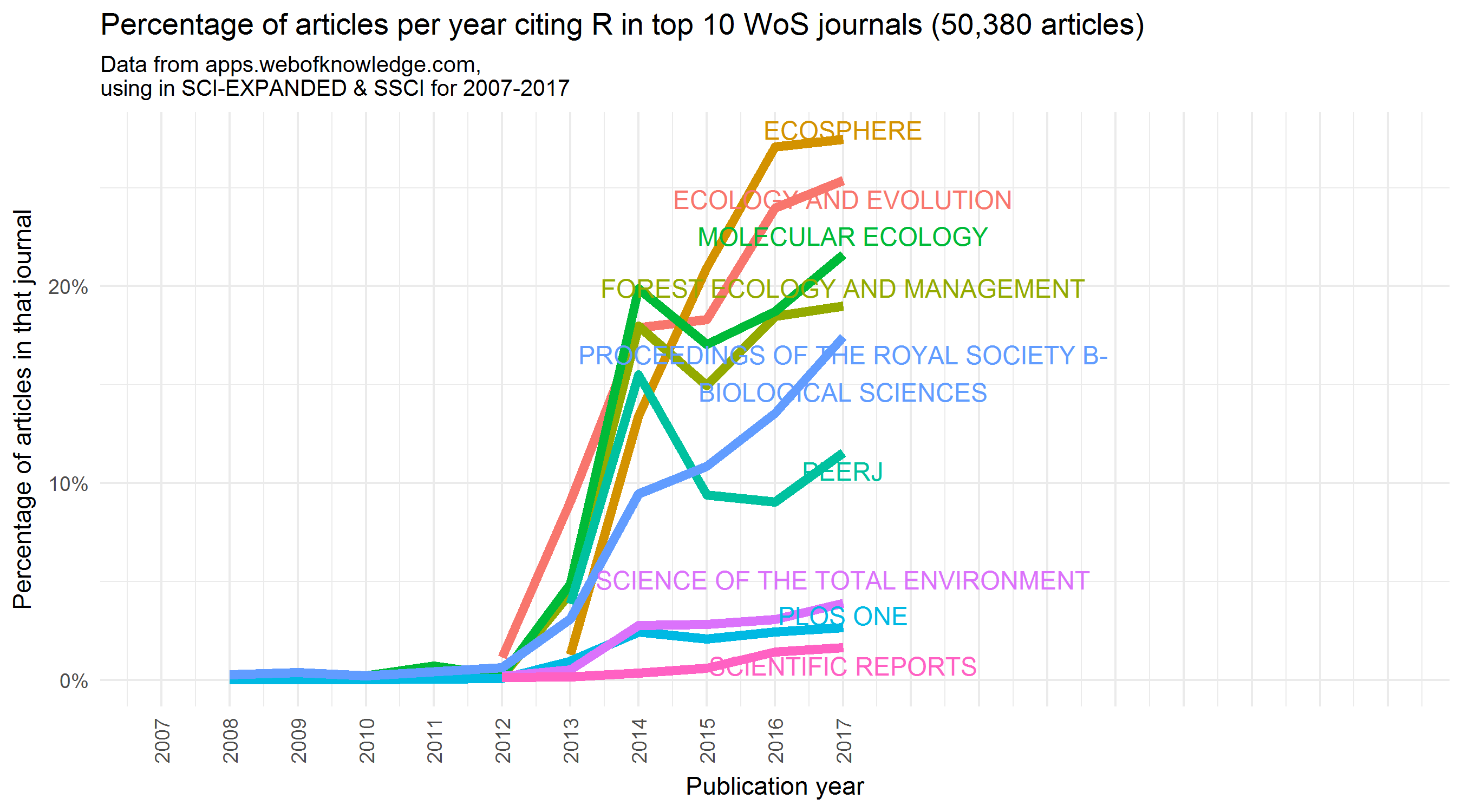
Finally, we will briefly introduce our R package, rrtools, which is designed to simplify and standardise the process of using R for archaeology, and making research more open and reproducible.

## Slide: Current uses of R [signpost slide only]

So, let’s start with the current uses of R.

## Slide: Articles in all sciences citing R

# We did a 'Cited Reference Search' in WoS for title "R: A Language and Environment for Statistical Computing" and author = "R Core Team", which is the output from `citation()`. This function first appeared in 2004 https://github.com/wch/r-source/commits/9a5c086855d2e8fb07f670e3315871b2de31a4fd/src/library/base/inst/CITATION   
#   
# Note that before 2012 the authors were "R Development Core Team", and then it was simplified to "R Core Team" https://github.com/wch/r-source/commit/9a5c086855d2e8fb07f670e3315871b2de31a4fd#diff-80bb8a8b133a6204e9642d76791927e2  
#   
# This is our exact search string for WoS:  
#   
# You searched for: CITED AUTHOR: ("R DEV COR TEAM" OR "R CORE" OR "R CORE TEAM") AND CITED WORK: ("R: A Language and Environment for Statistical Computing") AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article)  
# Timespan: All years. Indexes: SCI-EXPANDED, SSCI, A&HCI, ESCI.  
  
cran\_cites\_files <-   
 paste0(here::here(), glue("/analysis/data/raw\_data/cited\_cran/savedrecs ({1:11}).txt"))  
  
cran\_cites\_all\_areas\_all\_years <-   
 map(cran\_cites\_files,  
 ~readr::read\_tsv(.x,   
 quote = "",  
 col\_types = cols(.default = col\_character()))) %>%   
 bind\_rows() %>%   
 mutate(PY = as.numeric(PY)) %>%   
 filter(PY <= 2017)  
  
# limit to top journals  
top\_journals\_for\_cran\_cites <-   
 cran\_cites\_all\_areas\_all\_years %>%   
 group\_by(SO) %>%   
 tally(sort = TRUE)  
  
how\_many\_articles\_in\_top\_journals\_for\_cran\_cites <-   
 sum(top\_journals\_for\_cran\_cites$n)  
  
top\_journals\_for\_cran\_cites\_articles\_per\_year <-   
 read\_csv(str\_glue('{here::here()}/analysis/data/raw\_data/top\_journals\_for\_cran\_cites\_articles\_per\_year.csv')) %>%   
 select(-X1) %>%   
 gather(variable, value, -year) %>%   
 separate(variable,   
 into = str\_glue('X{1:4}'),   
 by = "\_") %>%   
 select(year, X4, value)  
   
# tally by year  
cran\_cites\_all\_areas\_top\_journals\_by\_year <-   
 cran\_cites\_all\_areas\_all\_years %>%   
 filter(SO %in% top\_journals\_for\_cran\_cites$SO[1:10]) %>%   
 group\_by(SO, PY) %>%   
 tally(sort = TRUE) %>%   
 mutate(journalname = str\_remove\_all(SO, " |-"),  
 PY = as.integer(PY)) %>%   
 left\_join(top\_journals\_for\_cran\_cites\_articles\_per\_year,  
 by = c('journalname' = 'X4',   
 'PY' = "year")) %>%   
 mutate(prop = n / value ) %>%   
 filter(SO != 'JOURNAL OF STATISTICAL SOFTWARE') %>%   
 ungroup()  
   
  
min\_y <- 2007  
max\_y <- 2025  
all\_sciences\_citing\_R\_over\_time <-   
ggplot(cran\_cites\_all\_areas\_top\_journals\_by\_year,  
 aes(PY,  
 prop,  
 colour = SO)) +  
 geom\_line(size = 2) +  
 geom\_text\_repel(  
 data = subset(cran\_cites\_all\_areas\_top\_journals\_by\_year,   
 PY == max(PY)),  
 aes(label = str\_wrap(SO, 40)),  
 size = 4,  
 nudge\_x = 0,  
 hjust = 0,  
 segment.color = NA,  
 direction = "y"  
 ) +  
 scale\_x\_continuous(breaks = min\_y:max\_y,  
 labels = c(min\_y:2017, rep("", length(2018:max\_y))),  
 limits = c(min\_y, max\_y)) +  
 scale\_y\_continuous(labels = scales::percent) +  
 xlab("Publication year") +  
 ylab("Percentage of articles in that journal") +  
 theme\_minimal() +  
 theme(legend.position="none",  
 axis.text.x = element\_text(angle = 90, hjust = 1, vjust = 0.5)) +  
 ggtitle(str\_glue('Percentage of articles per year citing R in top 10 WoS journals ({prettyNum(how\_many\_articles\_in\_top\_journals\_for\_cran\_cites,",")} articles)'),  
 subtitle = "Data from apps.webofknowledge.com,\nusing in SCI-EXPANDED & SSCI for 2007-2017")  
  
  
ggsave(plot = all\_sciences\_citing\_R\_over\_time,   
 filename = "all\_sciences\_citing\_R\_over\_time.png",  
 path = str\_glue('{here::here()}/analysis/figures'),  
 height = 5,  
 width = 9)  
  
knitr::include\_graphics(str\_glue('{here::here()}/analysis/figures/all\_sciences\_citing\_R\_over\_time.png'))



This plot shows how data analysis is changing in the sciences more broadly towards the use of the R programming language. We searched the Web of Science, produced by Clarivate Analytics, for citations to the R program over the last ten years, and filtered the results to the top ten journals with the highest number of citations to this software. The purpose of this filtering step was to find areas of research were R is generally known, so we can see how its use is changing over time. Note that many articles that use R do not cite the R program, so these numbers underestimate the true use of R in science.

We can make two major observations:

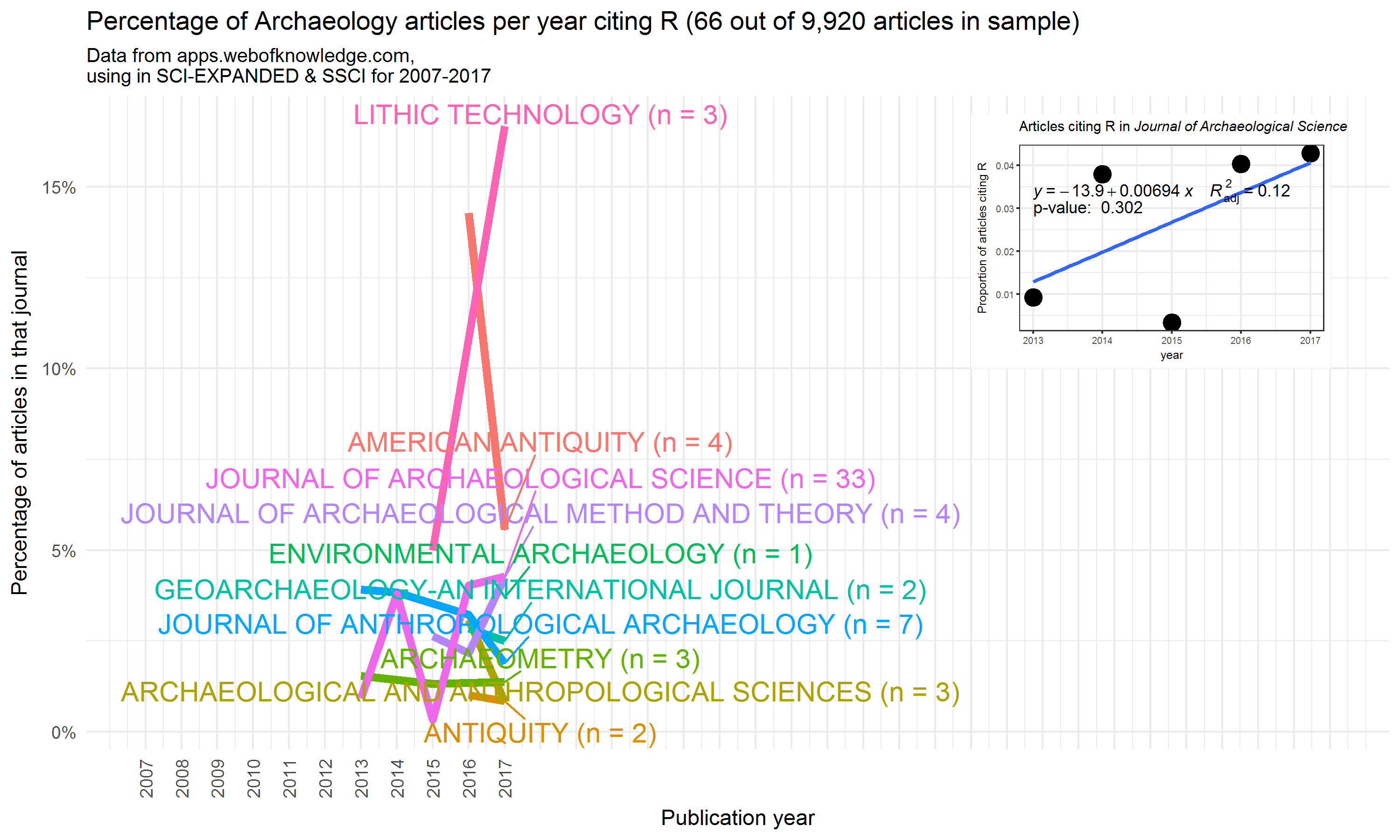
**First**, the areas of science that have seen big growth in the use of R are ecology, evolution and related biological sciences. This is especially significant for us archaeologists, as ecology and biology are subjects from which we quite often adapt methods for analysis, as for example in point-pattern-analysis and statistical modelling. We also see increases in multi-disciplinary journals such as PeerJ and PLOS One.

**Second**, the scale of the increases is quite impressive: Starting only 2012 some journals now have 20% of articles citing R.

So, we can see this is a substantial change in the way researchers are working in many fields. What about archaeology? Do we also see a change like this?

## Slide: Archaeology articles citing R

# Data are from an 'advanced search' of apps.webofknowledge.com using WC=Archaeology on apps.webofknowledge.com.  
  
# get the top archaeology journals by numbers of articles.  
wc\_archaeology\_2017 <-   
 readr::read\_tsv(paste0(here::here(), "/analysis/data/raw\_data/savedrecs.txt"),   
 quote = "")  
  
# The 10 journals with the most articles are:  
top\_journals\_by\_articles <-   
wc\_archaeology\_2017 %>%   
 group\_by(SO) %>%   
 tally(sort = TRUE)   
  
# For 2007-2017:  
  
data\_files\_from\_2007\_to\_2017 <-   
 paste0(here::here(), glue("/analysis/data/raw\_data/savedrecs ({1:2}).txt"))  
  
wc\_archaeology\_2007\_to\_2017 <-   
 map(data\_files\_from\_2007\_to\_2017,  
 ~readr::read\_tsv(.x, quote = "")) %>%   
 bind\_rows()  
  
total\_number\_of\_articles\_in\_sample <-   
 nrow(wc\_archaeology\_2007\_to\_2017)   
  
# most populararchaeology journals over time, that have some mention of R  
most\_popular\_journals\_over\_time <-   
wc\_archaeology\_2007\_to\_2017 %>%   
 group\_by(SO, PY) %>%   
 tally(sort = TRUE) %>%   
 filter(!SO %in% c("JOURNAL OF CULTURAL HERITAGE",  
 "STUDIES IN CONSERVATION",  
 "INTERSECCIONES EN ANTROPOLOGIA",  
 "ISRAEL EXPLORATION JOURNAL",  
 "JOURNAL OF MATERIAL CULTURE",  
 "TRABAJOS DE PREHISTORIA",  
 "ARCHEOSCIENCES-REVUE D ARCHEOMETRIE")) %>%   
 filter(n >= 2) %>%   
 ungroup()  
  
most\_popular\_journals\_over\_time\_names <-   
 unique(most\_popular\_journals\_over\_time$SO)  
  
archy\_journals\_total\_number\_of\_articles <-   
 read\_csv(str\_glue('{here::here()}/analysis/data/raw\_data/top\_archaeology\_journals\_articles\_per\_year.csv'))  
  
top\_archaeology\_journals\_articles\_per\_year <-   
 archy\_journals\_total\_number\_of\_articles %>%   
 gather(variable, value, -journal\_name) %>%   
 mutate(year = as.numeric(variable)) %>%   
 select(-variable)  
   
cran\_cites\_in\_archy\_journals <-   
 cran\_cites\_all\_areas\_all\_years %>%   
 filter(SO %in% most\_popular\_journals\_over\_time\_names)  
  
# what archaeology journals cite R?  
# unique(cran\_cites\_in\_archy\_journals$SO)  
  
archy\_journals\_total\_number\_of\_citing\_articles <-   
 nrow(cran\_cites\_in\_archy\_journals)  
  
cran\_cites\_in\_archy\_journals\_per\_year <-   
 cran\_cites\_in\_archy\_journals %>%   
 group\_by(SO, PY) %>%   
 tally(sort = TRUE) %>%   
 ungroup() %>%   
 mutate(PY = as.numeric(PY)) %>%   
 left\_join(top\_archaeology\_journals\_articles\_per\_year,  
 by = c("SO" = "journal\_name",  
 "PY" = "year")) %>%   
 mutate(prop = n / value) %>%   
 group\_by(SO) %>%   
 mutate(jn = str\_glue('{SO} (n = {sum(n)})')) %>%   
 ungroup()  
  
min\_y <- 2007  
max\_y <- 2040  
  
  
jas\_cites\_r <-   
cran\_cites\_in\_archy\_journals\_per\_year %>%   
 filter(SO == "JOURNAL OF ARCHAEOLOGICAL SCIENCE") %>%   
 ggplot(aes(PY,  
 prop)) +  
 geom\_point(size = 4) +  
 geom\_smooth(method = "lm",  
 se = FALSE) +  
 stat\_poly\_eq(aes(label = paste(..eq.label..,   
 ..adj.rr.label..,  
 sep = "~~~~")),  
 formula = y~x,   
 parse = TRUE,  
 size = 3,  
 label.y.npc = 0.7) +  
 stat\_fit\_glance(label.y.npc = 0.7,  
 size = 3,  
 method = "lm",   
 method.args = list(formula = y ~ x),  
 geom = "text",  
 aes(label = paste("p-value: ",signif(..p.value.., digits = 4)))) +  
 xlab("year") +  
 ylab("Proportion of articles citing R") +  
 ggtitle(expression(paste("Articles citing R in ", italic("Journal of Archaeological Science")))) +  
 theme\_bw(base\_size = 6)  
   
  
all\_archaeology\_citing\_R\_over\_time <-   
ggplot(cran\_cites\_in\_archy\_journals\_per\_year,  
 aes(PY,  
 prop,  
 colour = SO)) +  
 geom\_line(size = 2) +  
 geom\_text\_repel(  
 data = subset(cran\_cites\_in\_archy\_journals\_per\_year,   
 PY == max(PY)),  
 aes(label = jn),  
 direction = "y",  
 size = 5,  
 nudge\_x = 1,  
 hjust = 0) +  
 theme\_minimal() +  
 theme(legend.position="none",  
 axis.text.x = element\_text(angle = 90, hjust = 1, vjust = 0.5)) +  
 scale\_x\_continuous(breaks = min\_y:max\_y,  
 labels = c(min\_y:2017, rep("", length(2018:max\_y))),  
 limits = c(min\_y, max\_y)) +  
 scale\_y\_continuous(labels = scales::percent) +  
 xlab("Publication year") +  
 ylab("Percentage of articles in that journal") +  
 ggtitle(str\_glue('Percentage of Archaeology articles per year citing R ({archy\_journals\_total\_number\_of\_citing\_articles} out of {prettyNum(total\_number\_of\_articles\_in\_sample, ",")} articles in sample)'),  
 subtitle = "Data from apps.webofknowledge.com,\nusing in SCI-EXPANDED & SSCI for 2007-2017") +  
 annotation\_custom(grob = ggplotGrob(jas\_cites\_r),   
 xmin = 2030, xmax = 2040,   
 ymin = 0.10, ymax = 0.17)  
  
ggsave(plot = all\_archaeology\_citing\_R\_over\_time,   
 filename = "all\_archaeology\_citing\_R\_over\_time.png",  
 path = str\_glue('{here::here()}/analysis/figures'),  
 height = 6,  
 width = 10)  
  
knitr::include\_graphics(str\_glue('{here::here()}/analysis/figures/all\_archaeology\_citing\_R\_over\_time.png'))



Here are the results for archaeological articles. We found 66 articles that cite R over the last ten years. Note that this is certainly not complete, since these are just the journals tracked by the Web of Science database, and we know that archaeologists publish in a great variety of journals.

There are three important details in these plots:

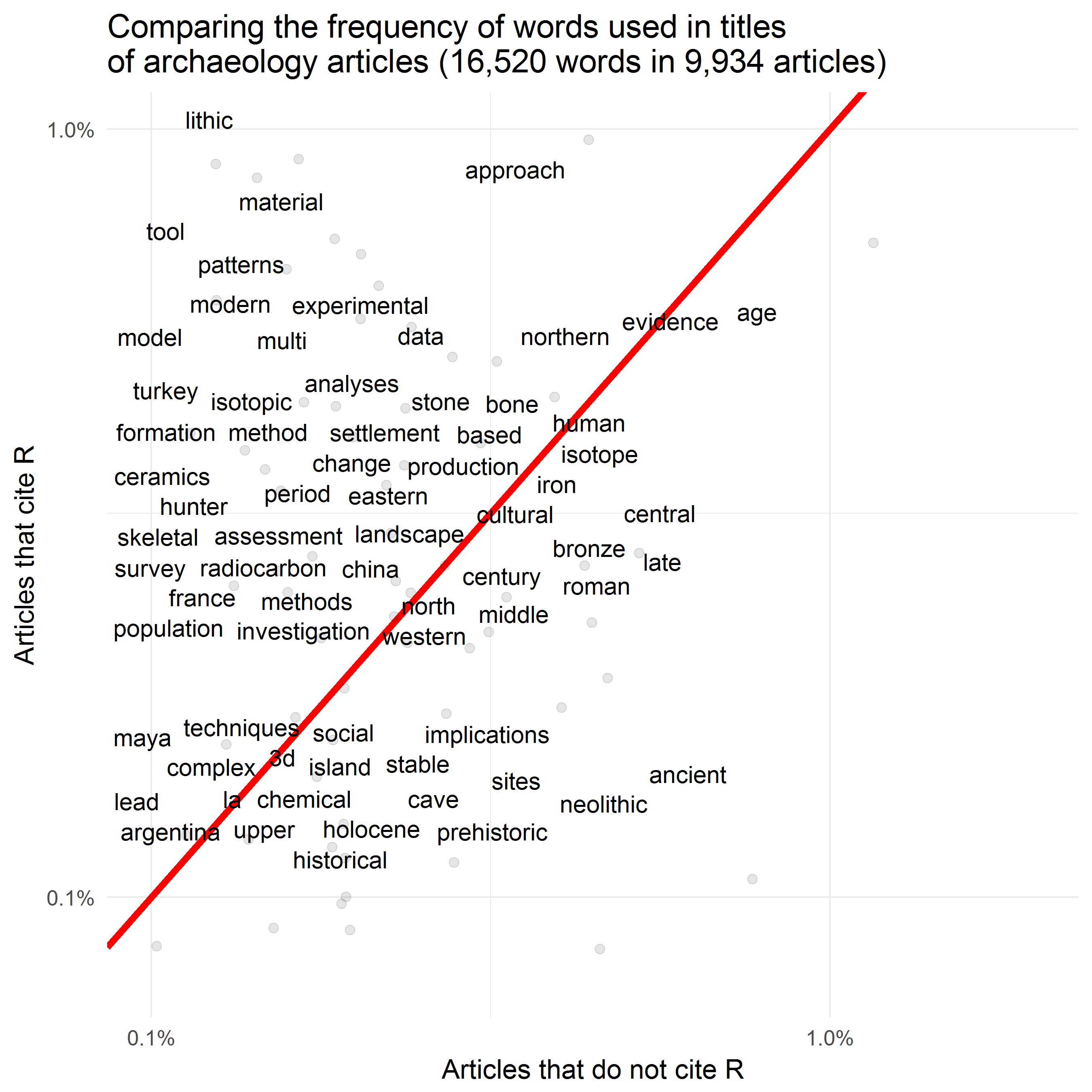
The **first** detail we see here is that overall, the proportions and the absolute number of articles citing R in archaeology are much smaller than what we see in the ecology journals on the previous slide. Only the *Journal of Archaeological Science* has more than 10 articles in our sample. It’s obvious, we have yet to adopt programming for data analysis in the same way that ecologists have.

The **second** detail is that we do not see a significant increase over time. Both archaeology and the other sciences we saw on the previous slide show an increase in citing R at around 2013-2014, but we do not yet see much of an upward trend in the archaeology journals. Indeed, in several of our journals the first citations of R only occured in the last two years. The plot on the upper right of this slide shows a statistical test of change over time for articles in JAS. With a low r-squared value and high p-value for the linear model, we conclude that there is no evidence for an increase in citations over time in this journal.

The **third** detail we see is the distinctive types of journals that have articles citing R. There is a clear focus on scientific archaeology, with *Journal of Archaeological Science* having the highest total number of articles. Archaeology as a subject has always been situated between a humanistic and a scientific tradition, but studies using statistical methods lean towards the sciences, which we mean with the term “scientific archaeology”. We also see geoarchaeology, lithics, and environmental archaeology represented in the journal names. We can investigate this further with statistical analysis of the words in the titles of all the articles in our sample.

## Slide: What are archaeologists doing with R?

# https://www.tidytextmining.com/twitter.html  
  
# These are all archy articles  
# wc\_archaeology\_2007\_to\_2017  
  
# These are ones citing R  
# cran\_cites\_in\_archy\_journals  
  
# subtract ones citing R from ones that do not  
  
wc\_archaeology\_2007\_to\_2017\_no\_R\_cites <-   
 wc\_archaeology\_2007\_to\_2017 %>%   
 filter(TI %nin% cran\_cites\_in\_archy\_journals$TI)  
  
# combine  
archy\_journal\_articles <-   
 bind\_rows(wc\_archaeology\_2007\_to\_2017\_no\_R\_cites %>%   
 mutate(ID = "no\_R\_cites",  
 VL = as.character(VL)),  
 cran\_cites\_in\_archy\_journals %>%   
 mutate(ID = "has\_R\_cites"))  
  
  
tidy\_cites <-   
archy\_journal\_articles %>%   
 unnest\_tokens(word, TI) %>%  
 filter(!word %in% c(stop\_words$word, "archaeology",   
 "archaeological",  
 "analysis", "study", "site"),  
 str\_detect(word, "[a-z]"))  
  
frequency <-   
 tidy\_cites %>%   
 group\_by(ID) %>%   
 count(word, sort = TRUE) %>%   
 left\_join(tidy\_cites %>%   
 group\_by(ID) %>%   
 summarise(total = n())) %>%  
 mutate(freq = n/total)  
  
frequency <- frequency %>%   
 select(ID, word, freq) %>%   
 spread(ID, freq) %>%  
 arrange(no\_R\_cites, has\_R\_cites)  
  
archaeology\_articles\_word\_freqs <-   
ggplot(frequency,   
 aes(no\_R\_cites,   
 has\_R\_cites)) +  
 geom\_abline(color = "red",   
 size = 2) +  
 geom\_jitter(alpha = 0.1,   
 size = 2.5,   
 width = 0.25,   
 height = 0.25) +  
 geom\_text\_repel(aes(label = word),   
 size = 5,  
 segment.colour = NA) +  
 scale\_x\_log10(labels = percent\_format(),  
 limits = c(0.001, 0.02)) +  
 scale\_y\_log10(labels = percent\_format()) +  
  
 theme\_minimal(base\_size = 16) +  
 xlab("Articles that do not cite R") +  
 ylab("Articles that cite R") +  
 ggtitle(str\_glue("Comparing the frequency of words used in titles \nof archaeology articles ({prettyNum(nrow(frequency), ',')} words in {prettyNum(nrow(archy\_journal\_articles), ',')} articles)"))  
  
ggsave(plot = archaeology\_articles\_word\_freqs,   
 filename = "archaeology\_articles\_word\_freqs.png",  
 path = str\_glue('{here::here()}/analysis/figures'),  
 height = 9,  
 width = 9)  
  
knitr::include\_graphics(str\_glue('{here::here()}/analysis/figures/archaeology\_articles\_word\_freqs.png'))



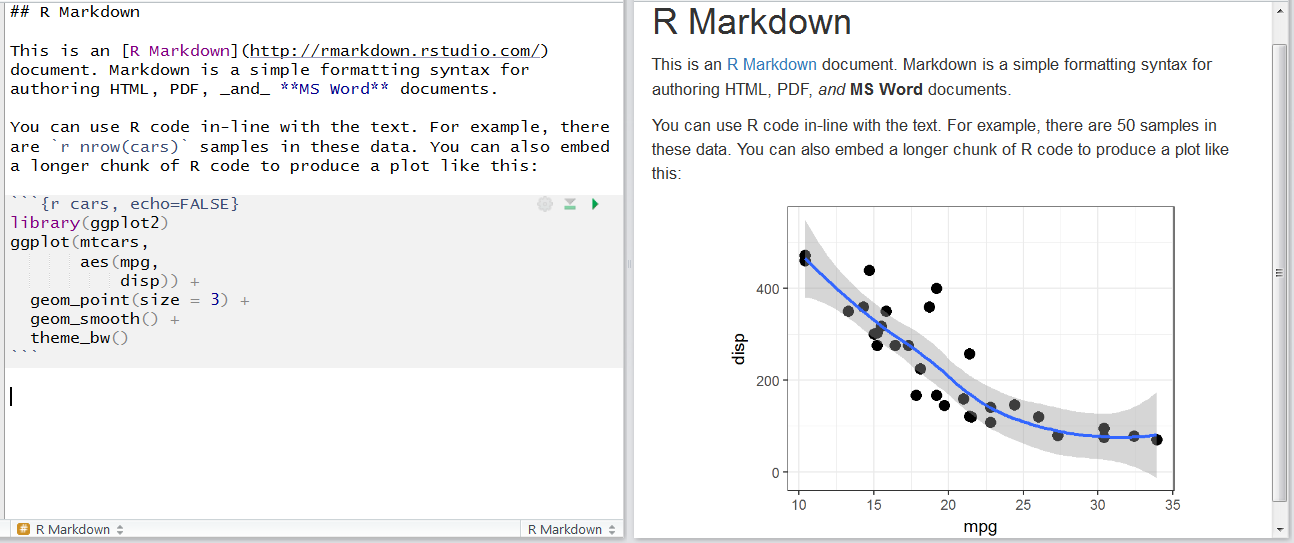
Here we see a comparison of word frequencies of the titles of archaeology journal articles. Words near the red line are used with about equal frequencies by papers that cite R and by papers that do not cite R. Words far away from the red line are used much more by one group of articles compared to the other.

This plot supports our initial observations: archaeologists using R tend to be doing more scientific archaeology. We see that many words relating to stone artefacts, zooarchaeology, and geoarchaeology are more frequently used by R-using archaeologists. On the other hand, articles that do not cite R more often discuss the Neolithic and metal-age periods, and less often reflect a technical specialisation.

## Slide: The unrealised potential of R for reproducible research [signpost slide only]

After this overview how R is being used at the moment, we want to focus on the possibilities of making research reproducible using R.

## Slide 9: The unrealised potential of R for reproducible research



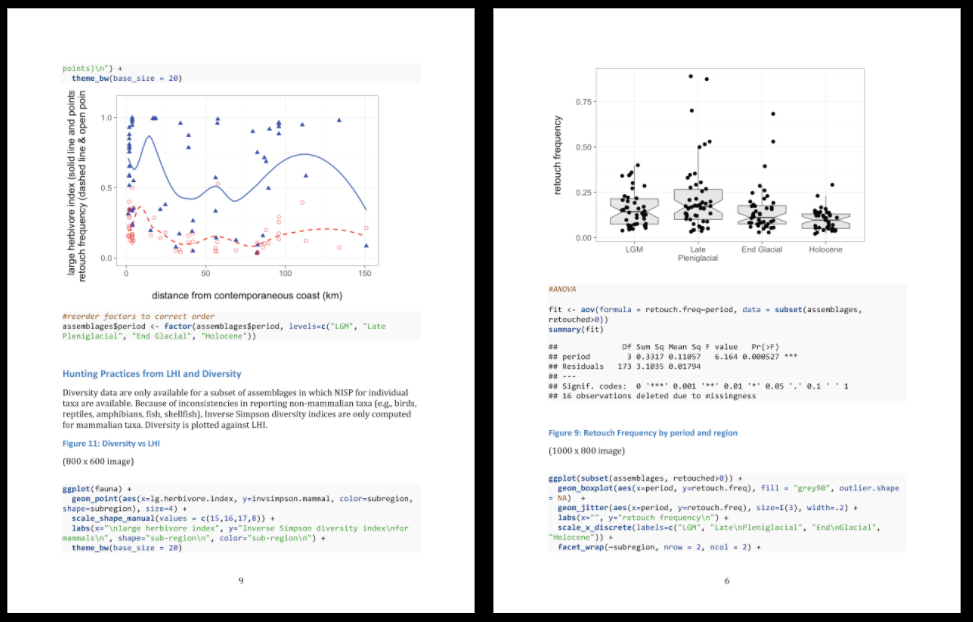
We would now like to focus on those archaeological papers using R, that are demonstrating the potential of R for open and reproducible research. These papers are using R Markdown for their research, which is shown in this figure.

On the left we see a plain-text document written in R Markdown. This means that is has lines of code in amongst lines of normal text. We can format the text with simple additions, such as the underscore to indicate italics and the hash symbol to indicate a heading. These simple additions are Markdown, a very easy markup language with plain text formatting syntax. It is designed so that it can be converted to HTML and many other formats. **R** Markdown allows us to mix R code and text in the same document. Here we see a short snippet of R code right inside the text (nrow(cars)) as well as a longer block of code below.

When this R Markdown document is rendered, or converted to another format such as PDF or Microsoft Word, we see that the code is replaced by the output it generates. On the right side of this figure we see that the inline code has been replaced by the number 50, and the code block below has been replaced by the plot.

We have found that some archaeologists are writing their entire journal articles useing R Markdown, and they share their plain-text R Markdown document along with their publication. This means that the reader can see all of the details of the data analysis, and all of the small decisions made along the way. This way the analysis is easily reproducible for others.

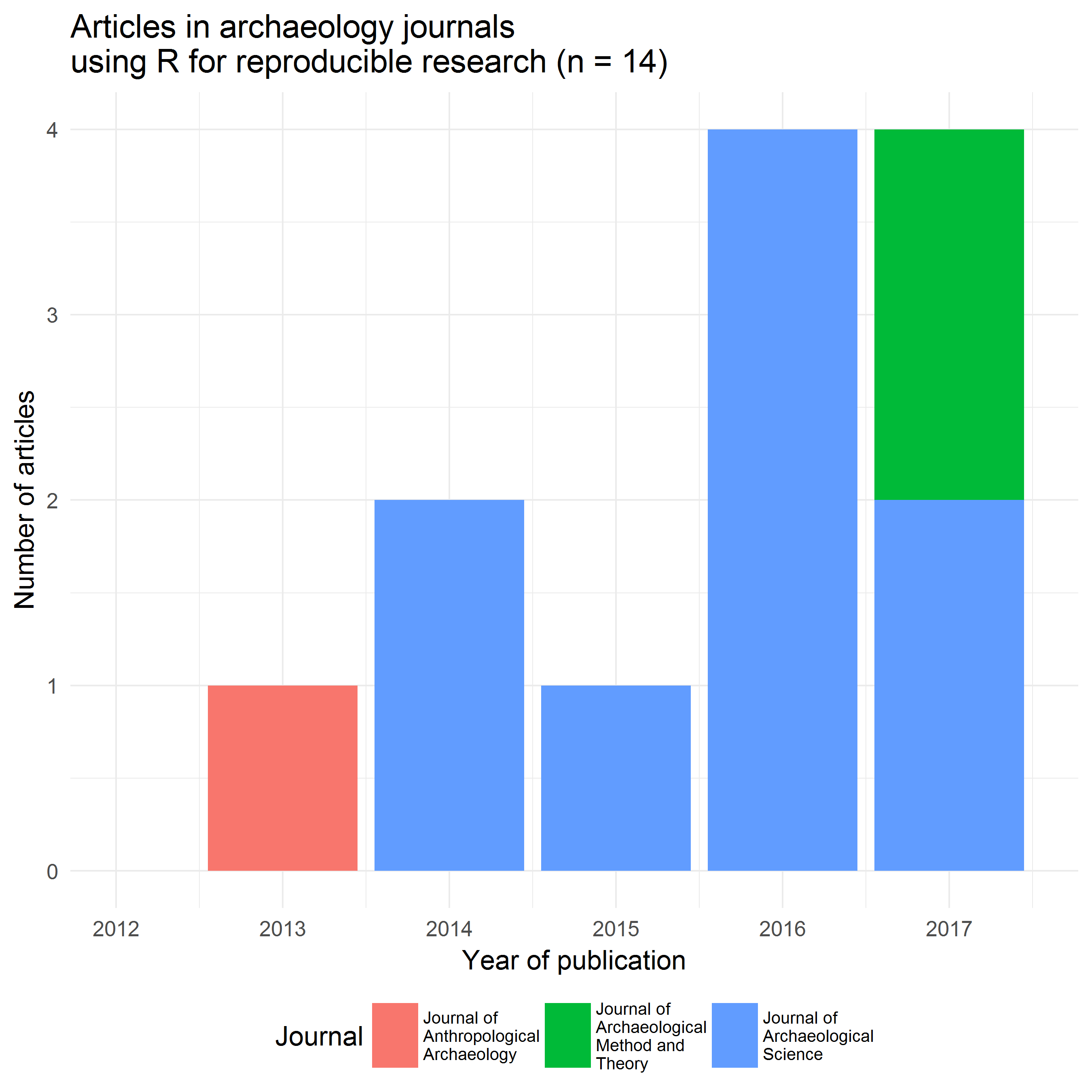
## Slide: Real-world example from Barton et al. 2018



Here is a real-world example R Markdown in use, from an archaeology article published in 2017.

## Slide: An increasing number of archaeologists are publishing reproducible research

archy\_ctv\_readme <- readLines(str\_glue('{here::here()}/analysis/data/raw\_data/ctv-archaeology-README.md'))  
  
top\_archy\_journals <-  
 str\_subset(archy\_ctv\_readme, "Journal of Archaeological Method and Theory|J Archaeol Sci|Journal of Anthropological Archaeology|Journal of Archaeological Science")  
  
top\_archy\_journal\_names <-  
 unlist(str\_extract\_all(archy\_ctv\_readme, "Journal of Archaeological Method and Theory|J Archaeol Sci|Journal of Anthropological Archaeology|Journal of Archaeological Science"))  
  
  
top\_archy\_journal\_names <- ifelse(top\_archy\_journal\_names == "J Archaeol Sci",  
 "Journal of Archaeological Science",  
 top\_archy\_journal\_names)  
  
archy\_ctv\_readme <- top\_archy\_journals  
archy\_ctv\_readme <- str\_remove\_all(archy\_ctv\_readme, "[[:punct:]]")  
archy\_ctv\_readme\_20XX <- str\_extract(archy\_ctv\_readme, " 20[[:digit:]]{2} ")  
archy\_ctv\_readme\_20XX <- str\_squish(unlist(archy\_ctv\_readme\_20XX))  
archy\_ctv\_readme\_20XX <- as.numeric(archy\_ctv\_readme\_20XX)  
archy\_ctv\_readme\_20XX <- archy\_ctv\_readme\_20XX[archy\_ctv\_readme\_20XX > 2010]  
  
number\_of\_reproducible\_articles <- length(archy\_ctv\_readme\_20XX)  
  
archaeology\_articles\_r\_reproducible <-   
ggplot(data\_frame(archy\_ctv\_readme\_20XX,  
 top\_archy\_journal\_names),  
 aes(archy\_ctv\_readme\_20XX,  
 fill = str\_wrap(top\_archy\_journal\_names, 15))) +  
 geom\_bar() +  
 theme\_minimal(base\_size = 16) +  
 scale\_x\_continuous(limits = c(min(archy\_ctv\_readme\_20XX),  
 max(archy\_ctv\_readme\_20XX)+0.5),  
 breaks = seq(min(archy\_ctv\_readme\_20XX),  
 max(archy\_ctv\_readme\_20XX)+0.5, 1)) +  
 ggtitle(str\_glue('Articles in archaeology journals\nusing R for reproducible research (n = {length(archy\_ctv\_readme\_20XX)})')) +  
 scale\_fill\_discrete(name = "Journal") +  
 theme(legend.text=element\_text(size=10),  
 legend.key.size = unit(2, 'lines'),  
 legend.position = "bottom"  
 ) +  
 xlab("Year of publication") +  
 ylab("Number of articles")  
  
ggsave(plot = archaeology\_articles\_r\_reproducible,   
 filename = "archaeology\_articles\_r\_reproducible.png",  
 path = str\_glue('{here::here()}/analysis/figures'),  
 height = 9,  
 width = 9)  
  
knitr::include\_graphics( str\_glue('{here::here()}/analysis/figures/archaeology\_articles\_r\_reproducible.png'))



In our sample of R citing archaeology articles, we identified 14 articles that include R script- and data files, either in the supplementary materials or in a trustworthy data repository. This plot shows that although this number is small, it is increasing, suggesting the beginning of a new approach to how archaeologists share their research.

The way in which these 14 articles make their work reproducible is by:

* making the data files available with their publication
* making the R script files available with the publication
* using R Markdown to add narrative text with the R code to give explanation and make it easier for readers to understand the analysis

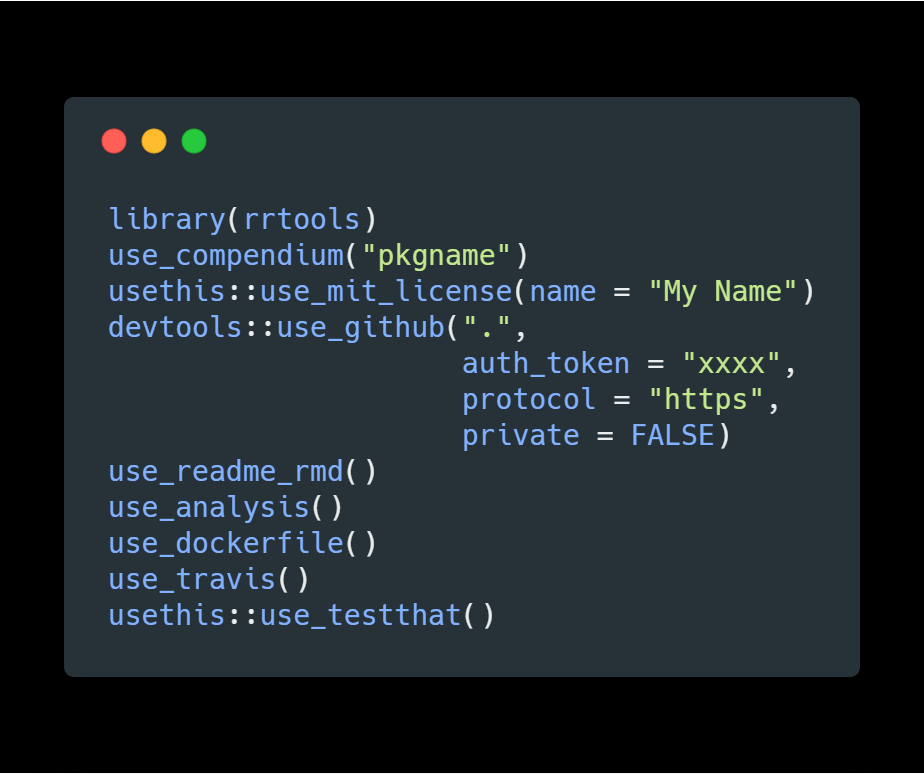
Articles that have these three characteristics are part of a new paradigm in scientific communication, and with these 14 articles, we believe we may have identified the beginning of this behaviour in archaeology. We think it will not be hard for archaeologists that already use R to make the transition to producing reproducible research by sharing their code and data. If this can be achieved, then we can realise some of the potential that R has for making archaeology more open and reproducible.

## Slide: Our new package, rrtools, and how it helps

As we studied the papers in our sample, we looked for common approaches to making and organising the code and data files. We found that there were few common standards used by archaeologists in making their work reproducible. There was a great variety of information presented in the code and data files. Some papers include code that will generate every figure and number in the text, while other papers only include code to produce one or two key figures in the paper.

Looking further afield, we also reviewed current efforts towards improving the reproducibility of research in other disciplines. In reviewing the literature we found several useful guides in diverse fields such as ecology and fluid dynamics. These guides described several best practices that are not yet common in archaeology.

So, after looking at the reproducible articles in archaeology, and the best practices in other disciplines, we wanted to make a simple toolkit that will help archaeologists to write reproducible research as easily as possible. This toolkit is an R package called rrtools. This package contains a small number of functions that automate many common analysis setup tasks.



On this slide we show the typical use of the key functions in rrtools that a researcher would use to quickly set up a compendium suitable for writing reproducible research in R. Each of these steps reflects a best practice that we’ve noted in our review of the literature. We will focus now on the first five:

* **use\_compendium:** This function creates a new directory for the research project, and creates an R package within this directory. We use the R package structure as the standard for our directory because it is widely recognised by R users. This means that people know where to look for code, where to look for data, and where to look for dependancies and other information.
* **usethis::use\_mit\_license:** This function creates a copy of the MIT software lisence to the compendium. This is important for two reasons. First, because it tells the reader that the author is happy for their code to be reused by other people. And second, it tells the reader that the author does not take responsibility for any problems that the reader might have when they use the code. This is important for setting the expectations about the relationship between the author and the reader, regarding the use of code.
* **devtools::use\_github:** This will initiate the use of the Git version control system. This is the current state-of-the-art for version control of any kind of plain text file, including text, code, data, and images. This will also create a repository on the GitHub website. Using version control is important because it allows the author to time-travel back to earlier states of their analysis, if they take a dead-end path. It is also important for making collaboration smooth because many people can work on one set of files without losing track of what the most recent version is. Plus it is nice to have a remote back-up of our work in case of an emergency. The GitHub repository can be kept private if desired, until the work is published.
* **use\_readme\_rmd:** This will create a simple document that is usually the first thing a reader sees when they browse the code files. This document is important because it helps to describe to the reader what to expect in the compendium, and it gives details of how to cite the compendium.
* **use\_analysis:** This function creates a set of folders according to the best practices for organising a typical reasearch project. We recognise that this file structure wont be perfect for every project. But we know that it will suit many projects quite well, and it will at least make the user think about how to organise their files logically, and according to best practises.

## Slide: The best-practice directory structure from rrtools



Here is the file structure the rrtools gives us. For now, we want to point out two important principles that it enforces:

* Keep the data separate from the methods. This means keep the code in a separate directory to the data
* Keep the raw data seprate from data that is created or derived during the analysis.

These two fundamental principles make it easy to stay organised during a complicated project, and make it easier for a reader to navigate their way through the compendium.

## Slide: Summary

To summarise:

* We have shown a bibliometric analysis that reveals substantial change in the use of the R programming languages in the sciences generally. We saw that ecological and evolutionary sciences have really embraced R in their publications.
* We have also shown that archaeologists have more recently taken up R, but to a much lesser degree, so far, than what we see in the other fields. We’ve seen that the use of R is most common in more scientific aspects of archaeology.
* We also identified a small sub-set of archaeology papers that use R to make their work fully reproducible. There is a lot of potential for other archaeologists to make their work reproducible, and we think this would be a good thing for archaeology generally.
* We have presented our freely available R package, rrtools, which is designed to make it easy for researchers in archaeology and other fields to make their research reproducible. We hope this will help to realise some of the potential to make archaeological research more open and reproducible.

Thank you.

##### pagebreak

# References

##### pagebreak

### Colophon

This report was generated on 2018-03-16 16:15:42 using the following computational environment and dependencies:

# which R packages and versions?  
devtools::session\_info()  
#> setting value   
#> version R version 3.4.3 (2017-11-30)  
#> system x86\_64, mingw32   
#> ui RTerm   
#> language (EN)   
#> collate German\_Germany.1252   
#> tz Europe/Berlin   
#> date 2018-03-16   
#>   
#> package \* version date source   
#> acepack 1.4.1 2016-10-29 CRAN (R 3.4.3)  
#> assertthat 0.2.0 2017-04-11 CRAN (R 3.4.3)  
#> backports 1.1.2 2017-12-13 CRAN (R 3.4.3)  
#> base \* 3.4.3 2017-12-06 local   
#> base64enc 0.1-3 2015-07-28 CRAN (R 3.3.0)  
#> bindr 0.1.1 2018-03-13 CRAN (R 3.4.3)  
#> bindrcpp 0.2 2017-06-17 CRAN (R 3.4.3)  
#> bookdown 0.7 2018-02-18 CRAN (R 3.4.3)  
#> broom 0.4.3 2017-11-20 CRAN (R 3.4.3)  
#> caa2018 \* 0.0.0.9000 2018-03-13 local   
#> cellranger 1.1.0 2016-07-27 CRAN (R 3.4.3)  
#> checkmate 1.8.5 2017-10-24 CRAN (R 3.4.3)  
#> cli 1.0.0 2017-11-05 CRAN (R 3.4.3)  
#> cluster 2.0.6 2017-03-10 CRAN (R 3.4.3)  
#> colorspace 1.3-2 2016-12-14 CRAN (R 3.4.3)  
#> compiler 3.4.3 2017-12-06 local   
#> crayon 1.3.4 2017-09-16 CRAN (R 3.4.3)  
#> data.table 1.10.4-3 2017-10-27 CRAN (R 3.4.3)  
#> datasets \* 3.4.3 2017-12-06 local   
#> devtools 1.13.5 2018-02-18 CRAN (R 3.4.3)  
#> digest 0.6.15 2018-01-28 CRAN (R 3.4.3)  
#> dplyr \* 0.7.4 2017-09-28 CRAN (R 3.4.3)  
#> evaluate 0.10.1 2017-06-24 CRAN (R 3.4.3)  
#> forcats \* 0.3.0 2018-02-19 CRAN (R 3.4.3)  
#> foreign 0.8-69 2017-06-22 CRAN (R 3.4.3)  
#> Formula \* 1.2-2 2017-07-10 CRAN (R 3.4.1)  
#> ggplot2 \* 2.2.1 2016-12-30 CRAN (R 3.4.3)  
#> ggpmisc \* 0.2.16 2017-09-17 CRAN (R 3.4.3)  
#> ggrepel \* 0.7.0 2017-09-29 CRAN (R 3.4.3)  
#> glue \* 1.2.0 2017-10-29 CRAN (R 3.4.3)  
#> graphics \* 3.4.3 2017-12-06 local   
#> grDevices \* 3.4.3 2017-12-06 local   
#> grid 3.4.3 2017-12-06 local   
#> gridExtra 2.3 2017-09-09 CRAN (R 3.4.3)  
#> gtable 0.2.0 2016-02-26 CRAN (R 3.3.1)  
#> haven 1.1.1 2018-01-18 CRAN (R 3.4.3)  
#> Hmisc \* 4.1-1 2018-01-03 CRAN (R 3.4.3)  
#> hms 0.4.2 2018-03-10 CRAN (R 3.4.3)  
#> htmlTable 1.11.2 2018-01-20 CRAN (R 3.4.3)  
#> htmltools 0.3.6 2017-04-28 CRAN (R 3.4.3)  
#> htmlwidgets 1.0 2018-01-20 CRAN (R 3.4.3)  
#> httr 1.3.1 2017-08-20 CRAN (R 3.4.3)  
#> janeaustenr 0.1.5 2017-06-10 CRAN (R 3.4.3)  
#> jsonlite 1.5 2017-06-01 CRAN (R 3.4.3)  
#> knitr 1.20 2018-02-20 CRAN (R 3.4.3)  
#> lattice \* 0.20-35 2017-03-25 CRAN (R 3.4.3)  
#> latticeExtra 0.6-28 2016-02-09 CRAN (R 3.4.3)  
#> lazyeval 0.2.1 2017-10-29 CRAN (R 3.4.3)  
#> lubridate 1.7.2 2018-02-06 CRAN (R 3.4.3)  
#> magrittr 1.5 2014-11-22 CRAN (R 3.4.3)  
#> Matrix 1.2-12 2017-11-20 CRAN (R 3.4.3)  
#> memoise 1.1.0 2017-04-21 CRAN (R 3.4.3)  
#> methods \* 3.4.3 2017-12-06 local   
#> mnormt 1.5-5 2016-10-15 CRAN (R 3.4.1)  
#> modelr 0.1.1 2017-07-24 CRAN (R 3.4.3)  
#> munsell 0.4.3 2016-02-13 CRAN (R 3.3.1)  
#> nlme 3.1-131 2017-02-06 CRAN (R 3.4.3)  
#> nnet 7.3-12 2016-02-02 CRAN (R 3.4.3)  
#> parallel 3.4.3 2017-12-06 local   
#> pillar 1.2.1 2018-02-27 CRAN (R 3.4.3)  
#> pkgconfig 2.0.1 2017-03-21 CRAN (R 3.4.3)  
#> plyr 1.8.4 2016-06-08 CRAN (R 3.3.1)  
#> psych 1.7.8 2017-09-09 CRAN (R 3.4.3)  
#> purrr \* 0.2.4 2017-10-18 CRAN (R 3.4.3)  
#> R6 2.2.2 2017-06-17 CRAN (R 3.4.3)  
#> RColorBrewer 1.1-2 2014-12-07 CRAN (R 3.3.0)  
#> Rcpp 0.12.15 2018-01-20 CRAN (R 3.4.3)  
#> readr \* 1.1.1 2017-05-16 CRAN (R 3.4.3)  
#> readxl 1.0.0 2017-04-18 CRAN (R 3.4.3)  
#> reshape2 1.4.3 2017-12-11 CRAN (R 3.4.3)  
#> rlang 0.2.0 2018-02-20 CRAN (R 3.4.3)  
#> rmarkdown 1.9 2018-03-01 CRAN (R 3.4.3)  
#> rpart 4.1-13 2018-02-23 CRAN (R 3.4.3)  
#> rprojroot 1.3-2 2018-01-03 CRAN (R 3.4.3)  
#> rstudioapi 0.7 2017-09-07 CRAN (R 3.4.3)  
#> rvest 0.3.2 2016-06-17 CRAN (R 3.4.3)  
#> scales \* 0.5.0 2017-08-24 CRAN (R 3.4.3)  
#> SnowballC 0.5.1 2014-08-09 CRAN (R 3.4.1)  
#> splines 3.4.3 2017-12-06 local   
#> stats \* 3.4.3 2017-12-06 local   
#> stringi 1.1.6 2017-11-17 CRAN (R 3.4.2)  
#> stringr \* 1.3.0 2018-02-19 CRAN (R 3.4.3)  
#> survival \* 2.41-3 2017-04-04 CRAN (R 3.4.3)  
#> tibble \* 1.4.2 2018-01-22 CRAN (R 3.4.3)  
#> tidyr \* 0.8.0 2018-01-29 CRAN (R 3.4.3)  
#> tidytext \* 0.1.7 2018-02-19 CRAN (R 3.4.3)  
#> tidyverse \* 1.2.1 2017-11-14 CRAN (R 3.4.3)  
#> tokenizers 0.1.4 2016-08-29 CRAN (R 3.4.3)  
#> tools 3.4.3 2017-12-06 local   
#> utils \* 3.4.3 2017-12-06 local   
#> withr 2.1.1 2017-12-19 CRAN (R 3.4.3)  
#> xfun 0.1 2018-01-22 CRAN (R 3.4.3)  
#> xml2 1.2.0 2018-01-24 CRAN (R 3.4.3)  
#> yaml 2.1.18 2018-03-08 CRAN (R 3.4.3)

The current Git commit details are:

# what commit is this file at? You may need to change the path value  
# if your Rmd is not in analysis/paper/  
git2r::repository("../..")  
#> Local: master E:/R/caa2018-Tubingen/  
#> Remote: master @ origin (https://github.com/benmarwick/caa2018-Tubingen)  
#> Head: [66735e9] 2018-03-14: added Bens cache-remove