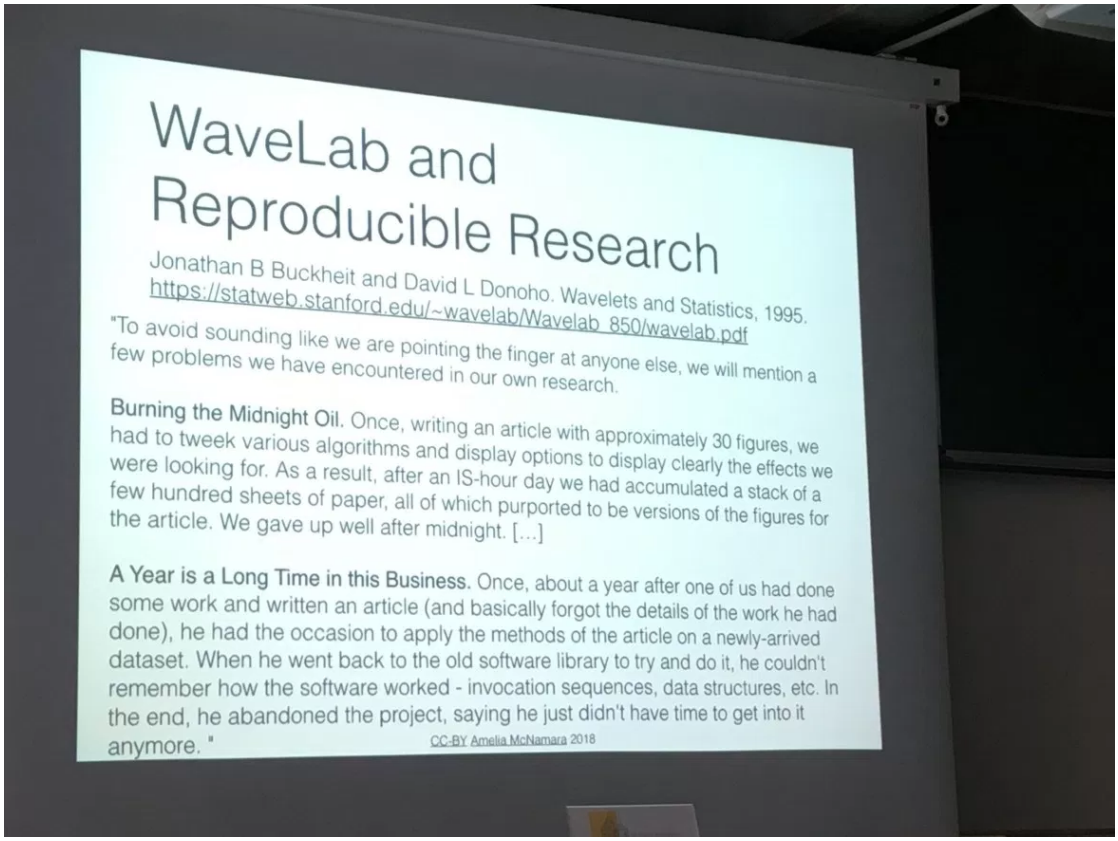
Scientists Programming – Amelia McNamara



felienne | February 6, 2018 | 0 Comments

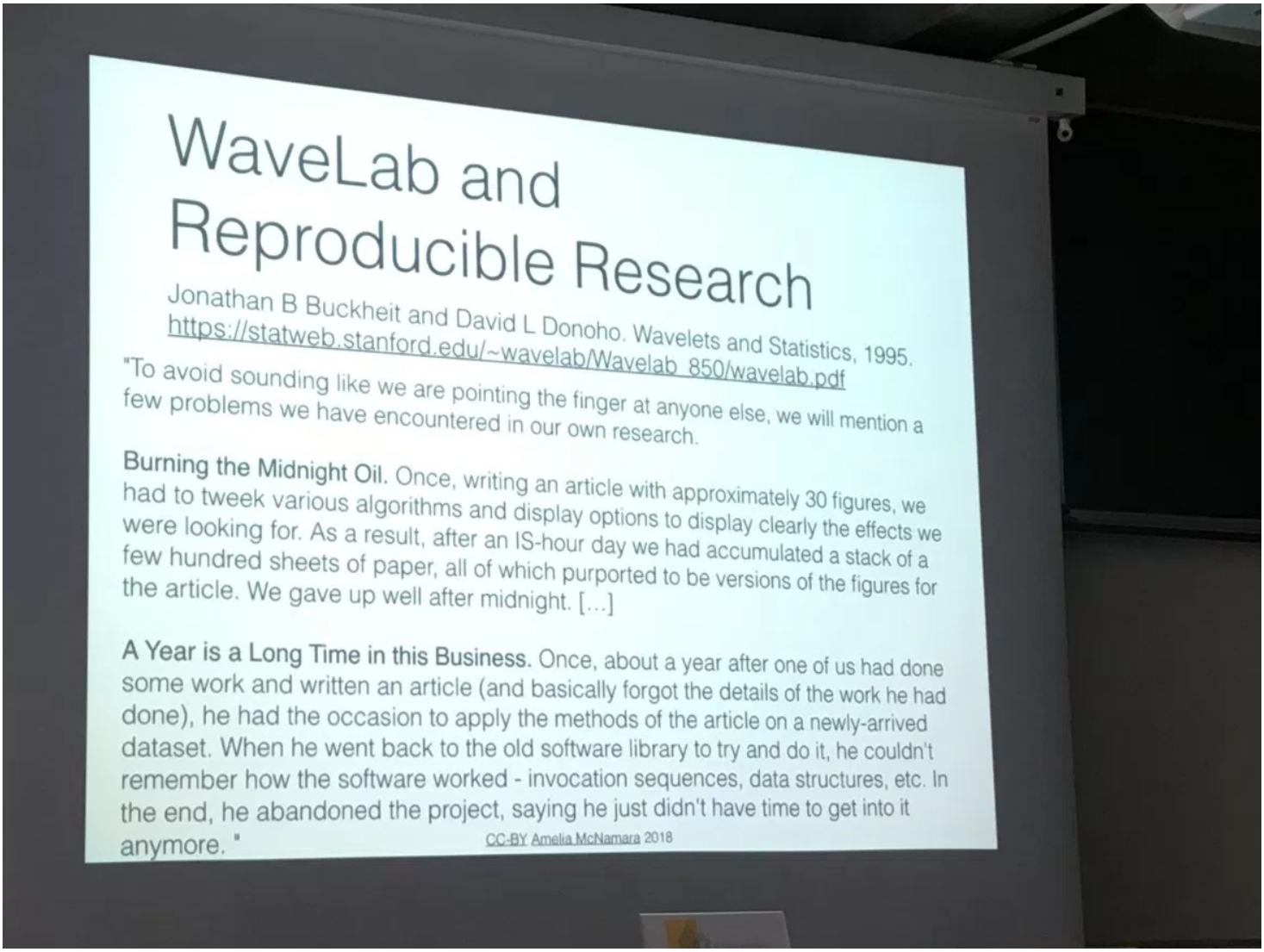
Live blog

Cool! After a lot of PL stuff, we now get us some end-user programming!

There are a few different types of programming that scientists do:

1. Create a model, generate data, check if it makes sense -> this is usually called simulation
2. Collect data, create a model -> this is more statistics
3. Application programming -> sometimes scientists also write apps, but Amelia does not know anything about that so it is not in the scope of this talk 🙂

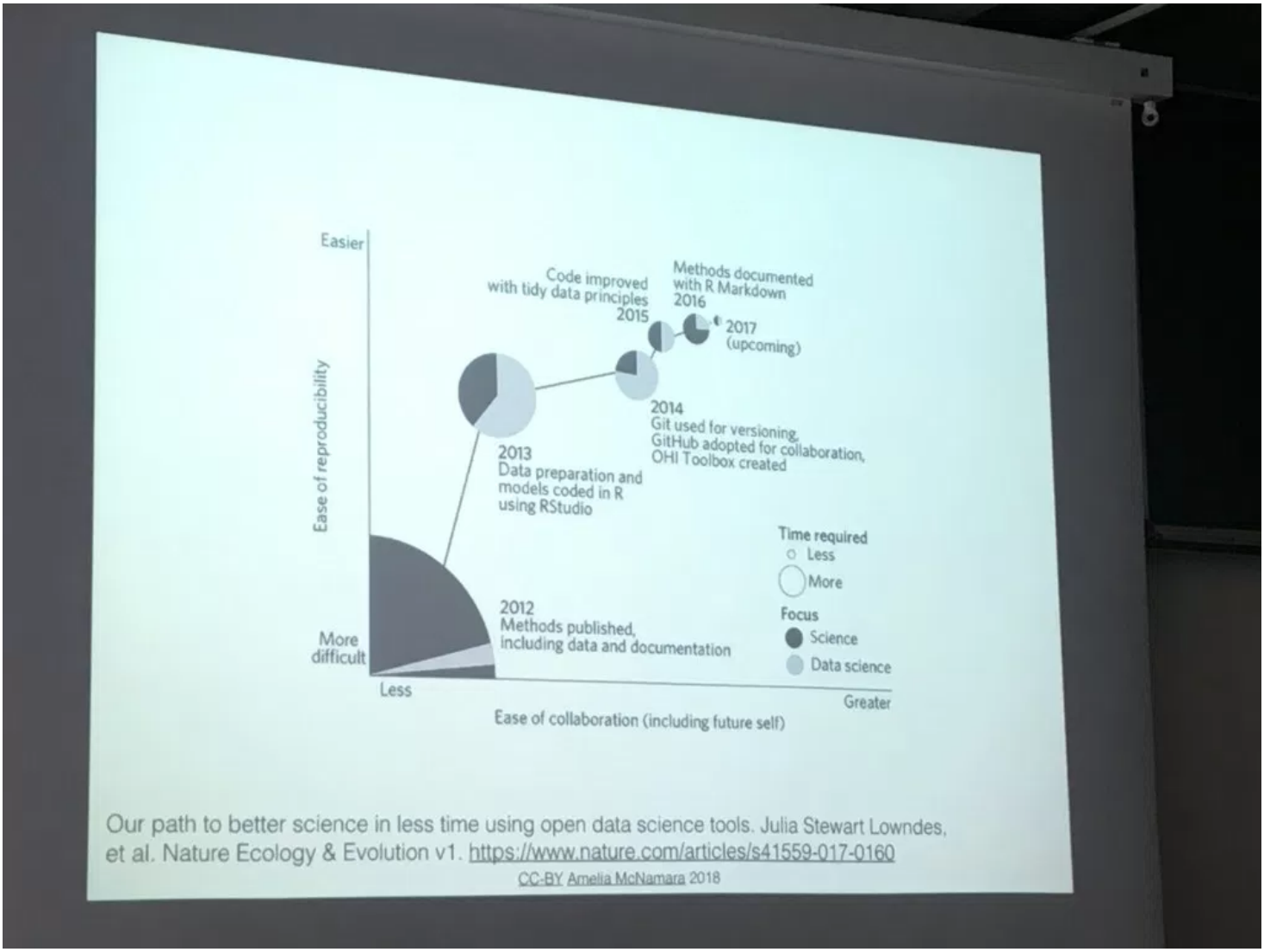
For the first category scientists mostly use Matlab for these type of analysis. From this paradigm came also the idea of reproducible research, in this paper. With some great (and familiar!) observations about managing research data:



In the second category we have…. Excel 🙂 Apart from the issues with the opaqueness of formulas, Excel has some issues in the statistic precision of functions as well. This man Guy Melard had checked every version of Excel for this, apparently, wow! Micheal Koblenz (who had worked on Numbers) remarks that this is very hard to fix since some spreadsheets might rely on the "bad" behaviour. Of course we all remember the gene name error paper that even impacted papers in Nature Genetics and even Nature itself.

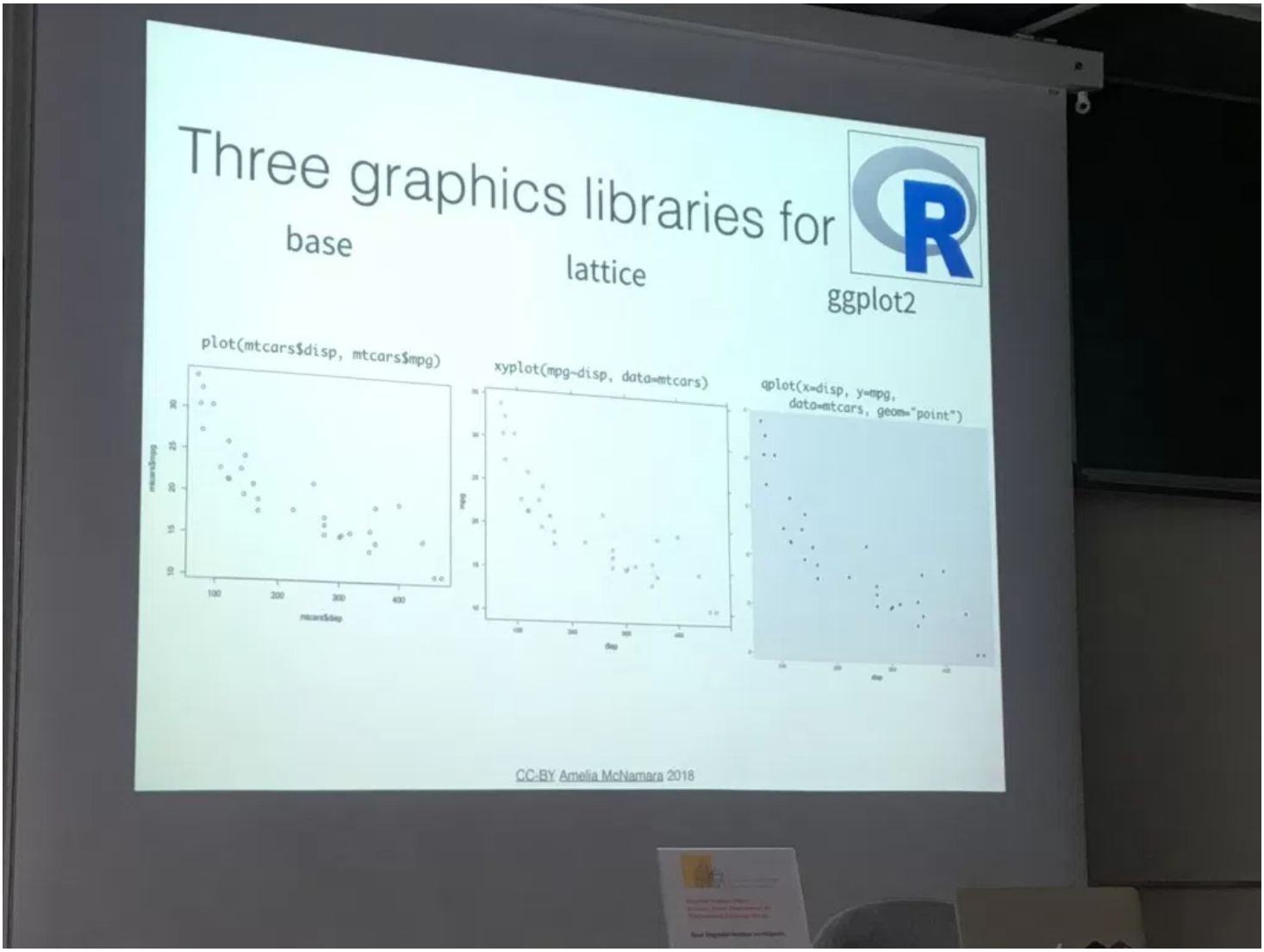
Some people are trying to teach chemistry (missed the link) and ecology students to code in other tools than spreadsheets to make science more reproducible. Some people are trying to understand what scientists do.

Amelia loves this paper about a group of scientists that changed their workflow, first by documenting everything they were doing in spreadsheets and then use R, Rstudio and Rmarkdown to generate the paper directly based on the the analysis of data. They did this is smaller steps, making the process better and better in each step.



There are a lot of people promoting good (or good enough) data science for scientists by the amazing Greg Wilson, and how to deal with spreadsheets in a more proper way.

Moving on to R then. Some programmers do not like R for not being a “real” language (sounds familiar? 🙂 She showed a great paper that helped her understand that R is in fact great for data science (sadly missed the link). A reason that developers do not like R is that there is not really a standardized syntax, different libraries do things quite differently:



Programmers that see R as a bad language do not focus on the right things, says Lutz Prechelt, because yes it is not a pretty language, but the ecosystem and the community are great and that is what matters. Andy Begel adds here that R is an old language that was created before language design was more mature so it feels archaic, but because of the great ecosystem people stay anyway, the same holds for MatLab. Amelia agrees and notes that we also have to understand that these scientist program in a different way, they do not write functions commonly, they use preexisting packages and call functions. Lutz argues that Python is a contestant here that is a “better” language for some uses of data science (not core statistics though) There was just a Not so standard deviations episode about this that talks about adding Python to Excel and argues that “normal” users will not use it. I am not sure if I agree, I will have to listen to this episode.

Another reason that people love R is for RMarkdown, with is a bit like Python notebook. Even better, says Amelia since it is always ran in order rather than a notebook which can be ran out of order. People even use this for “real” papers. There is also such an R into latex connection. What is Jupyter, asks Spefik. Amelia answers that it is just a rebranding of Python notebooks and is not really as production ready as RMarkdown is.

Amelia’s research interest is in the gap between tools made for learning statistics (TinkerPlots, StatKey, Inzight, StatCruch etc) and tools for using statistics (R, SPSS, Stata) Some key attributes for a modern tool are, according to Amelia’s dissertation:

accessibility, ease of entry, data as first order object, flexible plot creation and more. The full paper is here. Also you could listen to my most recent episode of SE Radio on Data Science 🙂

This post was visited 101 times.

Share this:

Related

Code.org - Baker Franke
February 7, 2018
In "Live blog"

Programming is Writing is Programming
March 22, 2017
In "Blog"

Salon des Refuses - Code is not just text
November 7, 2017
In "Blog"

← Previous post Next post →