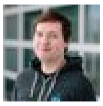
BUZZ CHRONICLES > DATA SCIENCE
Saved by @zmbnski
See On Twitter

Twitter Thread by John Myles White



John Myles White
@johnmyleswhite

This post is pretty bizarre, but it manages to hit on so many false beliefs that I've seen hurt junior data scientists that it deserves some explicit

(1) The notion that R is well-suited to "building web applications" seems totally out of left field. I don't feel like most R loyalists think this is a good idea, but it's worth calling out that no normal company will be glad you wrote your entire web app in R.

(2) It is true that Python had some issues historically with the 2-to-3 transition, but it's not such a big deal these days. On the flip side, I have found interesting R code that doesn't run in modern R interpreters because of changes in core operations (e.g. assignment syntax).

(3) "Most of the time we only need a latest, working interpreter with the latest packages to run the code" -- this is where things get real and reveal some things that hurt data scientists. If this sentence is true, it's likely because you don't share code with coworkers.

(3) Really is a broader issue in data science: people only think of what they need to do their work if no one else existed and code was never maintained. Junior data scientists almost always operate on projects they start from scratch and don't have to maintain for long.

(3) Especially astonishing is this claim, "The version incompatibility and package management issues would almost surely create technical, even political problems within large organizations." In reality, updating packages unnecessarily can itself be a source of problems.

(4) "To do this in R, we merely need to do b = a". The idea that assignment is intrinsically a copying operation seems to have just been made up. Making lots of copies is one of the things that slows R down and all R loyalists seem to admit this. Copying != purity.

(5) "as a functional programming language": Some folks keep claiming that R is a functional language, but they never define the term well. R is not pure by default. R code is riddled with mutations to the symbol table; library(foo) has to emit warnings for exactly that reason.

(6) "Eventually, such functional designs save human time — the more significant bottleneck in the long run." This belief is extremely common among R users and it really holds them back in situations in which performance does matter. Large projects often demand high performance.

(7) "In fact, the abstraction of vector, matrix, data frame, and list is brilliant." This belief really holds R users back when talking with engineers about implementations. At some point, everyone needs to learn what a hash table is, but its absence from base R confuses folks.

(8) "Beyond that, I also love the vector-oriented design and thinking in R. Everything is a vector:" This belief also seems common in the R community, even though the creator of R has said it's the biggest mistake they made. Scalars are always good and sometimes essential.

(9) If the most important of an IDE is an object inspector, maybe "No decent IDEs, ever" is true, but I think this is another case where the author has just never interacted with software engineers or understood their needs.

Putting it all together, there's a very troubling (and self-defeating) tendency in the data science world to embrace insularity and refuse to learn about the things software engineers know. Both communities have important forms of expertise; more sharing is the way forward.