Review of: *Bitfields encode what data knows about itself*, by **Steffen Ehrmann**; reviewed by: **Edzer Pebesma**

The paper introduces an R package called "bitfield", which can be used to encode and decode data in a completely flexible way into byte arrays (raw binary; misnamed "integer arrays" in the paper). The motivation is that "when data crosses workflow boundaries", much of the information ("distribution parameters, model agreement metrics, and quality flags that shaped the original analysis") get lost, because it is not encoded, and the paper argues that for that reason "What arrives is typically a mean value. Everything else remains trapped in supplementary materials or institutional memory."

There is some truth to the observation that only means are propagated through workflows, however the reason for that not that geoscientists cannot (and do not) encode this information, but that workflows typically ignore it because this information is hard to handle, it is not easy (to do, software-wise), it is not simple (mathematically, computationally), and there is not enough reward ("everyone else does it like this").

Geoscienstists can and do encode error distributions as part of the core data (instead of in "supplementary materials or institute memory"), for instance by ways of NetCDF or Zarr data, where error distributions are arrays next to the primary ("mean value") array(s), or where ensembles are encoded as its own data dimension. Packing data in a compact or compressed form is done, standard, using BLOSC or LERC compressors; for instance in Zarr, blosc is used widely, cloud-optimized geoTIFFs with float data use LERC. The data itself is then encoded in the usual INT1, UINT2, INT2, UINT2, INT4, UINT4, half precision float (BFLOAT16), FLOAT, DOUBLE forms, which pose *no* interoperability problems when read across workflows (and are handled at the hardware level).

Choosing alternative, completely flexible data-to-byte encodings (using bitfields) creates the problems of how to do this (how many and which bits for what), and how to describe this. By doing this, prototypically, in R, the interoperability problem ("workflow boundaries") is created, but not shown how it is solved. Besides not providing a solution to this problem, it does not sufficiently argue that there is a problem in the first place.

There are numerous smaller problems with this paper, many of which suggest that AI was used to create arguments, but the main one remains that it creates a problem that has been solved in the geoscientific computing community, without solving it.

Here are answers to most of the questions the GI board asks reviewers to answer:

1. Does the paper address relevant scientific questions within the scope of GI? No; if the paper addresses a relevant contribution, it is to scientific data, not to geoscientific data. Although the use case is geoscientific, none of the problems addressed have any geoscientific aspect.

2. Does the paper present novel concepts, ideas, tools, or data? Somewhat. It is well known that you can encode data as bits, and do this in different

ways; this is how computers work. It proposes a completely generic framework to encode integers and floating point numbers in arbitrary numbers of bits; although this may have some use, it makes it difficult to communicate data encoded like that accross systems, which is one of the motivating examples ("when data crosses workflow boundaries"). The implementation is confined to R, and so does not demonstrate what is needed if data encoded like this enters a Python or C++ environment.

3. Are substantial conclusions reached? Not to the extent that they are relevant for geoscientists.

4. Are the scientific methods and assumptions valid and clearly outlined? The goals are not clearly stated: is it compression? is it interoperability? is it propagation of metadata, error/distribution information?

5. Are the results sufficient to support the interpretations and conclusions? Demonstrating this only in R does not address the problem of data crossing workflow boundaries. The use case is not supported with online material that lets the user reproduce it.

6. Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)? No.

7. Do the authors give proper credit to related work and clearly indicate their own new/original contribution? Yes.

8. Does the title clearly reflect the contents of the paper? No. It is not evident what is meant by bitfields, and the word bitfields is easily confusingly with the accompanying R package "bitfield". The title phrase "data knows about itself" suggests that "data" can "know", which is not true.

9. Does the abstract provide a concise and complete summary? No. For instance, the claim in the abstract "What arrives is typically a mean value. Everything else remains trapped in supplementary materials or institutional memory. Bitfields offer a direct solution." is both not true and over-confident.

10. Is the overall presentation well structured and clear? No. For instance, the discussion brings up all kind of new themes. Many new topics are presented in appedices, some of them out of scope.

11. Is the language fluent and precise? Yes.

12. Are mathematical formulae, symbols, abbreviations, and units correctly defined and used? Yes; interestingly, units are completely ignored as a piece of information that is often not transported accross workflows.

13. Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated?

14. Are the number and quality of references appropriate?
15. Is the amount and quality of supplementary material appropriate? Too many appendices, too much information moved there.