

The lecture notes can be found

here <https://docs.google.com/presentation/d/1SqMwFKwiaur1P4g6mqtaGn6h62TIAkQjnpMY3zoPAEw/edit?usp=sharing>

The potential set of outputs of a data science experiment are pretty much limitless. However, four general types of outputs pop up most frequently. Those are:

1. Reports
2. Presentations
3. Interactive web pages
4. Apps

(**Interactive graphics** are important enough to merit their own category. However, they're usually embedded in a web page, so I'm lumping them in there.) Let's discuss each of the categories in turn.

Reports are easily the most common output of a data science experiment. Since the goals of reports varies wildly across settings, let's simply discuss a few hallmarks of a good data science report. It should:

- Be clearly written
- Involve a narrative around the data
- Discuss the creation of the analytic dataset
- Have concise conclusions
- Omit unnecessary details
- Reproducible

By and large, these points are obvious. However, this latter point is one that we discuss a lot throughout the specialization. Reproducible reports have mechanisms under the hood to recreate the data analysis. The number of benefits of report writing in this fashion are many. They include: getting the data scientist to think about the output of the process (since the code is embedded in the eventual output), very clear documentation that extends beyond code commenting, automation of the report generation process and then, of course, reproducibility. The main tools for producing reproducible reports are knitr and ipython notebooks. You should create a culture of using these tools, or similar ones, in your organization, as they will dramatically improve reproducibility.

Oddly enough, the same rules apply to presentations, though reproducible presentation software is less well adopted. For R, there's slidify and rStudio's presenter. These tools automate presentations in the same way that knitr and ipython notebooks automate report generation.

Interactive web pages and apps are similar enough to combined in the discussion. Again, as the requirements will vary so much across applications, we will only discuss a few hallmarks of good output. These include:

- Good ease of use / design
- Good documentation
- Code commented
- Code version controlled

Good ease of use and design are a discipline unto themselves. Since your data scientists are likely also not software engineers or designers, their design is probably not going to be optimal. However, modern tools allow data scientists to prototype apps and interactive web pages quickly and relatively easily. Your data scientists

should then pay some attention to ease use and design. Good documentation is mostly effort, though the same caveats apply with design.

On the other hand, having well commented code and version control should be standard practice for data scientists. Well commented code is easy to return to for new employees or the original coder returning to the project after a long hiatus. Version control is similar good practice. With version control, data scientists will be able to return to any checked in version of the project. The comments are useful for documenting the evolution of the project as well. Tools such as git and github make version control easy and are in the standard toolkit of data scientists.