

Workflows

Saturday, May 4, 2019

8:26 PM

The concept of a "workflow"

- In Jupyter notebooks, data analysis is broken down into steps we call "cells".
- "cells" can contain two kinds of things:
 - A computation and its output.
 - A comment on the computation.

The purpose of a typical Jupyter notebook is to document a data analysis workflow:

- Perform a series of data analysis steps, one per cell.
- Document the reasoning behind the steps and interpretation of the results.
- Mixing code and discussion is sometimes called "literate programming".

Why Jupyter allows mixing of code and discussion:

- It's well known that code itself doesn't explain why steps are taken. That needs to be explained separately from the code.
- Reasoning is most compelling when the code is given along with the conclusions. That way, other researchers can validate the reasoning.

Validation is part of the job of the data scientist:

- Decisions often involve large sums of money.
- Multiple analysts are often consulted about the same issue.
- Their conclusions are compared.

Thus, it is not enough to write a program; a compelling explanation must also be provided.