Continuous Research

Setting Integration as the Default

Contributors/Authors (please add your name if you have something to add!)

- Alex Morley

Introduction

Science today is built almost entirely on a culture of competition. Formal integration of research from different groups working on the same topic happens very rarely. When it does happen, for example when results/algorithms are explicitly compared, these results are often highly biased[1] or incomplete.

Continuous Research is the concept that all scientific outputs, every new dataset, technique, and hypothesis, can be can be integrated such a way that they are re-evaluated against all related research. In developer-speak we could say that we're "bringing Continuous Delivery to research".

What is Continuous Delivery?

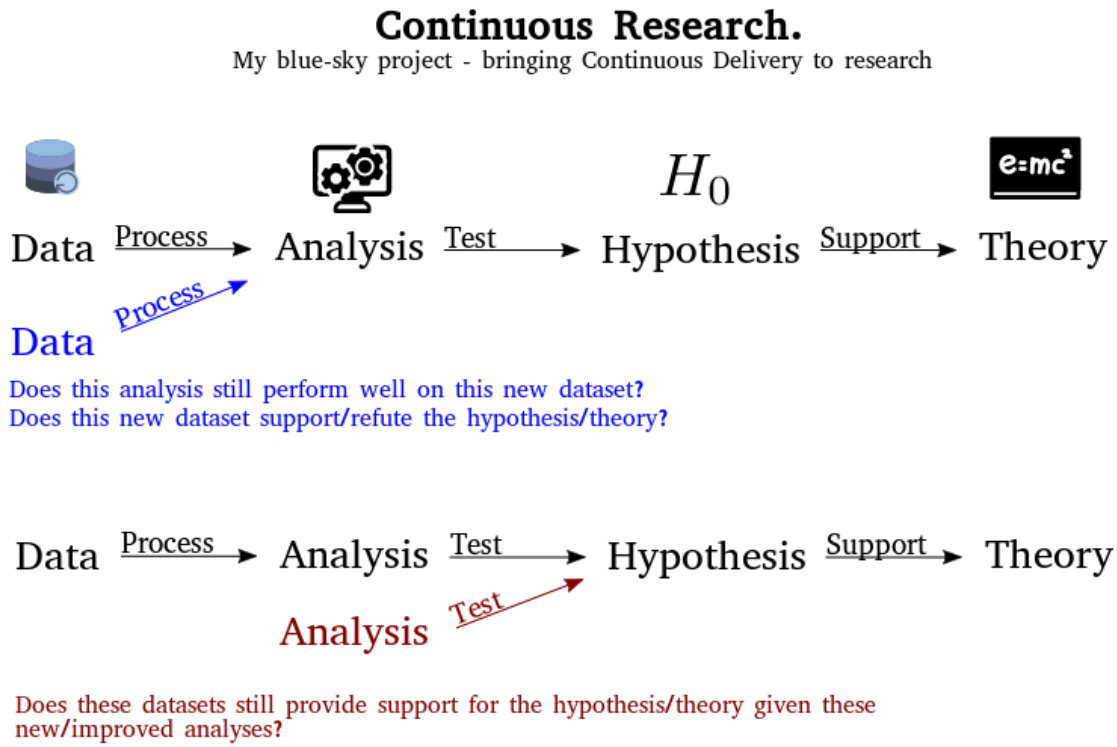
The requirement to maintain large, distributed, and complex projects, along with the high availability of low-cost computing has driven the widespread adoption of "Continuous Delivery" in software development. The mechanics can be complex, but the principle is simple: automated tests are frequently run against the software so that even when a brand new feature is added, you can be sure that the software is still performing correctly and (at least close to) functional enough to be released immediately.

How is research done now?

You collect a dataset, you do some analysis to test a hypothesis, and then try to integrate that hypothesis with current theories. Great, there's your paper, submit it to Journal of Sciency Things and move onto the next project. But what happens when someone else collects a dataset that could also be used to test that hypothesis? Or someone working in the mathematics department writes a better version of your analysis. Best case scenario they might use your code/data in their papers, otherwise they might just mention or cite you. But the work will never be truly integrated, and each paper might be related to tens/hundreds of other pieces of research, meaning that manually finding and using all of the code/data associated with their work will almost never be possible, even if it is open.

How does that change under Continuous Research?
Continuous Research is a framework in which we re-evaluate *all* hypotheses when any new data, analysis, or pipeline becomes available. In such a way, we obtain a live view of how much support there is for a particular theory, and under what conditions such a theory really holds. The big picture being that we want to try and integrate related research as tightly as possible, rather than the loose collection of papers that is the *de facto* standard for a body of evidence.

Prototypical Workflow



Use Cases

Adding a "full" study - "The Virtual Paper"

Dataset (neural activity + the position of the animal)
–>
Analysis (divide the environment into a grid then count the number of "spikes" each neuron fires at each place in the grid)
–>
Hypothesis (Some neurons will show some specificity at a particular location in this grid)

–>
Theory (Place cells provide animals with a map that enables them to navigate the world around them)

Adding a new dataset - *“At the Coalface”*
Someone does a similar set recordings but taken from a different region of the brain. All the analyses are automatically repeated. We don’t find place cells here. We refine the hypothesis to “Neurons in the hippocampus will show specificity to a particular location” and update our theory.

Adding a new analytical pipeline / node - *“Have you tried...”*
Someone else thinks that counting the number of spikes might not be a great way of determining locational specificity of neuronal firing. Perhaps we should instead determine the information content that a neuron has about the position of the animal. This analysis is then *automatically applied to all compatible datasets*. It turns out only a subset of the neurons we thought were place cells actually contain information about where the animal could be.

Adding a new analytical pipeline / node II - *“Have you tried...”*
Somone thinks that your results were an artifact of the way you identified which spikes came from which neuron (known as “spike-sorting”). 1) They provide a better spike-sorting method. *All analyses are re-run using this method*. We can now see whether this makes a difference. Or 2) They can already see that this study has been re-run with every available spike-sorting algorithm and the results seem robust to this choice.

Adding a new visualisation - *“No Sunglasses Required”*
TODO

Adding a new hypothesis / Editing a theory - “*The Comments Section”*
TODO

Prototype/Visual Demonstration

Ethos

While technology continues to improve, the methods underlying much of research haven’t quite entered the 21st Century. The pace of change is now increasing which could be a great thing, but we are concerned that we are still baking in many out of date values into our new practices. Thus this project will actively practice “Open Research” as we fell it should be: accessible,

inclusive and extensible. Accessible as in most functionality should be available to anyone with a laptop and web-browser. Inclusive as in providing an interface that is friendly to everyone, and supporting a safe community for users and developers. And extensible so that any individual is free to use and build upon the platform in a way that best suits them. One of the goals of the project is to showcase how great we believe Open Research can be, and we believe that isn't possible without keeping with these values.

Discussion

What are Proximal Impacts we hope to have?

Open Science Showcase - Encouraging Sharing

One goal of creating this framework is to give people an example of what kind of things we can do when scientists work openly. On top of this, it provides an immediate incentive to share data properly, because if you do then you *know* that you will get all of this extra insight from the extra pipelines that will be run automatically. Contrast this with currently people hoping that someone might one day use their data for something useful. The same logic applies to sharing the source code for analyses and analytical pipelines.

Independent & Explicit comparisons of Model/Algorithm/Software performance

In every field there seems to be at least one problem for which the current paradigm is "look here how our solution to this problem, which we have been tuning for a year, compares against this other one, where we didn't change the default settings". With this way of thinking it becomes very difficult to really determine what works best. Within Continuous Research each new algorithm would automatically be benchmarked against all the previous algorithms, with all the previous datasets.

Furthermore we can support exact provenance tracking for the effects of changes to models / datasets. And by making changes to all (most of) the moving parts that contribute to our understanding scientific theory transparent & version controlled, we also have the ability to see what the impact of each change really was.

Enabling Researchers to Fail Faster[3]

We are reducing the "knowledge cycle" - the time it takes to get feedback and release a project so that you and others can start building on top of it from months to < 1 day.

What Broader Impacts do we hope to have?

Truly Participatory Science

The more modularisation we can achieve the more we can lower the barrier for people to contribute to one aspect of a project without knowing the ins and outs of the other parts. And they can see in real time what the impact of there contributions are. In other words, the more detail we can abstract away from those who don't want to know, the easier we make it for people to contribute in the area in which they have the most expertise.

“Incremental Progress is Still Progress”[2]
How many people have had their research put down because it was “incremental”? We know that no research really exists in a vacuum and ultimately if we keep making incremental progress then we will move forward just fast enough! The difficulty lies in that currently incremental improvements are lost because publishers, funders, and even other scientists undervalue these advances. With Continuous Research all relevant research is explicitly linked; this will protect incremental improvements from getting lost and/or repeated unnecessarily.

Technical Details

Implementation

A large subset of the implementation details for such a project overlap with MLStudy, a “low-friction data-analysis platform in the web-browser” being developed by Jeremy Magland at the Flatiron Institute.

Key Features:

- Run-Anywhere.
  Entirely based on web technologies (javascript + nodejs) the framework for MLStudy is entirely hardware agnostic.

- Compute-Anywhere.
  Ability to run any study anywhere, being it a local, remote, cloud &/or cluster based.

- Open Source

- Flexible
  Studies are just JSON files, pipelines are just JS scripts, datasets are just JSON files with a checksum and a pointer to the data. All aspects have been built with extensibility/flexibility in mind.

De-centralization

With MLStudy we already have most of the ingredients we need for decentralisation.

- *Easy* - Running your own instance of MLStudy is as easy as `npm install; npm start`
- *Compute-Anywhere* - Analyses can currently be run anywhere the user has access to with either docker, or the specific packages for a given computation, installed.
- *Seeding* Currently anyone can donate a given amount of your local compute resources as a “seed”. Support for distributing a single pipeline/analysis across seeds is also possible.

The main missing part is distributed/de-centralised storage of data. At the moment data has to be stored locally or pulled from a cloud server. However luckily we don't have to solve this problem ourselves thanks to the Dat Project. Support for *dat* data sources will be the key next step towards decentralisation. From there it will also be possible to experiment with how well the whole project could run as a peer-to-peer network.

References

1. Ref?
2. Kirstie Whitaker - CollabW18 Talk
3. Sorry for this terrible tech buzzword...

Comments to think more about…

https://docs.google.com/document/d/1mOsUhger4fu2Yr2AleFCAiyarEKpDEmpqGjUOd9dNFI/edit?disco=AAAAB20clrA

"Open Science" is not really open unless it is truly participatory. Small steps are great, and we encourage everyone to get started in whatever way is most accessible to them. Nonetheless simply putting the code you used for a piece of research isn't fulfilling the potential of what Open could be. But if your code is running live somewhere and anyone can come along and play with the parameters to see what effect that would have had on the outcome of your experiments - that's getting a bit closer. If there is a low-barrier to someone recreating your visualisations in a way that's accessible to them that's also closer. In a technical sense open for me additionally nesseciates *accessibility*, *extensibility* and *interoperability*. And this project is about showing how powerful research can be when it meets these criteria.