Towards Executable Notebooks:
Infrastructure to support collaborative and transparent scientific discovery

Proposed Project Overview: Unique Aspects and Pilot Studies

Scientific research is increasingly complex. Many scientific results involve enormous data volumes, elaborate simulations, and many complex analyses performed using a multitude of tools. This sophistication has yielded important discoveries, but capture of the often chaotic analysis process is incomplete and remains largely manual (e.g., through logbooks) for many crucial decision points.

Reproducing results through this manual process is time-consuming and error-prone. Furthermore, the results can be fragile: choices regarding procedure can produce suboptimal, or even invalid, results, yet review is hampered by incomplete, missing or forgotten studies. Teaching newcomers the analysis methodology is also a major challenge.

The proposed framework addresses these problems by putting *provenance* at the center of the data exploration process. By capturing the full process, and integrating the provenance (i.e. the source of data, versions of software etc.) from multiple tools and derived by the scientists involved, the framework weaves all aspects of a scientific analysis into a reproducible whole. Provenance of the exploratory process and the derived results can also serve as a catalyst to expedite scientific explorations by supporting knowledge re-use and to foster collaboration.

This captured provenance serves as a detailed record of how the results were derived, allowing their reproduction and validation. Furthermore, the analysis and querying of the provenance information opens up new opportunities to help scientists identify relevant results, data sets, tools that can help them in their work, as well as enable the discovery of best analysis practices which can then be re-used.

We envision scientists using an electronic notebook where all analyses are recorded: the notebook tracks the procedures applied to the data in an executable form, and consequently, the results can be reproduced. It also provides an environment in which the scientist can manipulate the data further, incorporate graphs and tables, and record her intermediate conclusions, ideas and decisions. Moreover, should the scientist want to reproduce past results, the notebook can automatically muster the scripts, software, and data. Notebooks can be used by individuals, shared within research labs, as well as more broadly with the scientific community. The sharing of notebooks makes it possible for others to learn by example from the reasoning and analysis strategies of experts, in addition to supporting the reuse of reuse results thus collaboratively moving science forward.

The proposed work creates technology that changes the way scientific analysis and exploration is done. We promote a **cultural shift** that rewards collaborative efforts, sharing, and strong focus on transparency and reproducibility by researchers. Specifically, new technologies to standardize biomedical research is essential, not just to ensure provenance and reproducibility, but in fact to drive further multimodal research. Provenance and reproducibility benefits both the scientific community and the individual researcher. Today, given the increased amount and use of data in the biomedical research there is growing potential for scientific advances, but these fields are also in need of methods to ensure robustness of studies. A solution in this area will also advance the field by producing standard data and methods that can empower further research questions to be examined. Our proposal addresses areas of biomedical research that are traditionally disparate thus demonstrating the wide and inclusive applicability across the field. Two driving motivators are provided which offer disparate applications, to demonstrate the universality of the work, and are specialties of Investigators in our group.

What we will do:

We will development of new computing infrastructure and platform for biomedical research providing the following features:

- Structured and efficient analysis of data with complex processing pipelines that includes provenance capture
- Provenance and reproducibility benefits both for the scientific community and the individual researcher
- Structured modeling of all phases of research to cover the whole "story" from raw data to quantitative results and statistical analysis, including the full history how results were obtained
- Enable others to reproduce and verify results and/or collaborate on the analysis by re-running the processing and change modules, parameters etc. for further testing
- A platform on which the computer is hidden: platform-independent, robust to changes to updates of OS and libraries in needs of later reprocessing via building of full repositories including executable code, parameters, data, etc.
- A computing eco-system which relieves individual researchers from recompilations, working with different software versions, installations on different computers, dependencies on local resources, incompatibility across machines etc.
- Easier, efficient development of processing workflows increases productivity and research quality
- Systematic testing of parameters, robustness, reproducibility
- Collaborative computational infrastructure – track not only the computational pipeline but also the interactions among collaborators
- From single-run testing of given hypothesis to discovery science via efficient exploratory analysis
- Querying the history, to *debug* or understand the process the lead to a result, and to enable knowledge discovery, e.g., to find in a shared repository relevant results, discover successful analysis patterns

Driving motivator 1: Population-level epidemiological studies

Population-scale studies are specifically subject to concern about study power and bias. These concerns recently highlighted[1]. Limitations also exist because of the number of other studies on the same question, and the ratio of true to no relationships among the relationships probed in each investigation. Thus to avoid influence of external interests and prejudices, limited pre-selection of tested relationships and small effect sizes, there are important needs for greater reproducibility to be able to verify within and across studies. Finally there is also possibility for new research if multiple studies and data sets around the same question are reproducible and standard and can be brought together to improve. The proposed infrastructure offers a paradigm-shift that is well needed in this impactful area.

Driving motivator 2: Neuroimage analysis

Neuroscience research is faced with rapidly growing concerns on reproducibility of results, now even expressed by the NIH director ("Policy: NIH plans to enhance reproducibility, F.S. Collins & L.A. Tabak, Nature, 01/15", "Amid a Sea of False Findings, the NIH tries reform, The Chronicle of Higher Education Voosen 03/15") Despite strong efforts by the community for sharing of software systems for data analysis (SPM, FSL, FreeSurfer, ITK, NITRC repository), sharing of image data (NIH NDAR, Kitware MIDAS, ADNI), organizing of co-called *Challenges* to provide web-based benchmarking of new tools on annotated data provided by experts, and even availability of workflow systems to explore parameters and share processing pipelines (LONI pipeline UCLA/USC), most published results are based on data processing that may rarely be fully reproducible

[1] Ioannidis, John PA. "Why most published research findings are false." *Chance* 18.4 (2005): 40-47.

and is not shared with the community as an entire processing system. Open issues include the frequently ad hoc choice of a sequence of processing steps, the often heuristic choice of large sets of parameters with unknown effect the final results – sometimes even guided and repeated by the expected outcome, all leading to results can often not be fully reproduced or traced back. We plan to use existing image data and procedures from large pediatric neuroimaging studies, currently explored by a multidisciplinary team of psychiatrists, cognitive neuroscientists, radiologists, statisticians, and computer scientists, to demonstrate the paradigm change but also new scientific opportunities for data exploration provided by the newly proposed Bio-Notebook Infrastructure.

Key personnel and methodologies

The Team

The Computer Science & Engineering Department at NYU Poly collaborating with the Computer Science department at the Courant Institute of NYU is growing to encompass more cross-disciplinary research, and new approaches to computer science. This enables us to use computation in a progressive way that reaches into different fields. This project brings together a research team that has substantial expertise in computer science and biomedical research. *Dr. Juliana Freire* and *Dr. Claudio Silva* have done pioneering work in methods for reproducibility and have a track record of developing open-source tools that have been adopted by scientists. They have received grants from multiple sources, including NSF, DoE, NIH, NASA, the Gordon and Betty Moore Foundation, the Alfred P. Sloan Foundation, the Keck Foundation, AT&T and IBM, to develop methods and infrastructure to support data intensive research, including support for publication of reproducible results. However, so far, their focus has been mostly on physical sciences. *Dr. Guido Gerig* and *Dr. Rumi Chunara* have expertise in data-intensive methods for biomedical research, in particular in the areas of medical imaging and public health. Dr. Guido Gerig's research focuses on the development of novel tools and processing pipelines for brain mapping to be applied to large clinical studies. Therefore, the lack of reproducibility, bias towards only publishing positive results, and low rate of confirmation of results are key issues in his research. He is also involved in the development and dissemination of open-source/open-platform tools including web-development and organization for "Challenges"[2], web-based systems where researchers access biomedical data and annotated ground truth and compete in an unbiased way by publicly comparing results. Dr. Chunara's work focuses on building and using data from participatory data for large population-scale public health surveillance questions, both in infectious and chronic disease surveillance. Thus she works with comprehensive biomedical information including genomic, molecular, physiological and phenotypic data sets and also is particularly interested in bringing together multiple layers of data types, doing so initially for influenza surveillance. As well, her work involves public facing and data platforms to generate and use these multiple relevant data sets. Accordingly, given the nature of working with multiple data sets, a central aspect of her work relies on standardization and reproducibility across all of these data and associated methods. Dr. Dennis Shasha has done extensive work on data intensive computing, started the reproducibility effort in the database community in 2008, and is a co-architect of some of the reproducibility tools that have come out of the collaboration with Dr. Freire and Dr. Silva. Because Dr. Shasha has worked closely with NYU biology professors Gloria Coruzzi and Ken Birnbaum on genomics and is currently working on a machine learning project involving predicting Alzheimer outcomes with Dr. Rick Kline of NYU Medical. Together, the NYU team will build upon their previous work to design tools and develop solutions that can advance reproducibility in biomedical research.

The need for Keck support

Today's era of increased data has driven excitement and potential in biomedical research areas. As in many data-intensive fields, quality and reproducibility are key issues in biomedical research. While data and

[2] https://www.kitware.com/midaswiki/index.php/Projects/COVALIC/Design

computation have transformed many disciplines and have enabled important scientific discoveries, this revolution did not substantially affect how scientific results are published and shared. Biomedical research has unique challenges in this regard. High variability in specific, even widely used protocols is common. Therefore, in addition to the data, protocols, populations, analysis plans and pipelines must all be shared. Computational pipelines, such as the ones used in medical imaging, often apply heuristics and include a large number of parameters, making it difficult even for the researcher who created the pipeline to reproduce it. Invalid results in this domain can have serious consequences, including the development of harmful drugs and treatments. The current state of the field is worrisome, and this has been reinforced by recent studies that examined the reproducibility of published results and found that only between 11 and 25% of published data for drug development could be consistently reproduced[3,4.]

Budget

We foresee a budget of $1 million over two years for the proposed work. This budget includes personnel time, travel and workshops as outlined below:

Personnel:

- 1 month for each PI – *this time is allotted for communication between the team and supporting the research work*
- 1 research scientist – *management of the project, including preparation of reports and communication materials*
- 2 post-docs – *supervision of students and work, driving the research direction*
- 3 PhD students – *working on specific aspects of the project – the three students will each focus on building the system, interfacing with specific biological applications and testing*

Travel:

- Travel funds for trips to allow the project members to travel to disseminate the work to other groups

Workshops:

- Funds to support workshops to disseminate the work and train scientists on the use of the proposed infrastructure

Appropriateness for Keck vs. Government Funding

Current efforts, federally funded or otherwise (e.g., NIH BD2K program, Biocaddie, etc.), are rooted in sound biomedical research domains, however, their focus is limited to basic data and tool accessibility, often targeting a very specific application domain at the expense of generalization. Initial efforts in this regard foster collaboration on specific projects, working on open systems for particular types of data sharing, incentives and redefining norms for all parties. While these initial efforts are relevant, many centers and teams across biomedical fields, including clinical, genomic, public health and bioinformatics, are starting to and have been generating and using data intensively in research studies and, specifically for scientists, reproducibility and sharing within their own studies and across study types is a challenge.

---

[3] C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. Nature, 483:531–533, March 2012.

[4] F. Prinz, T. Schlange, and K. Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery, 9:712, September 2011.