Finding the Lost:
True Dissemination of Large Data through Efficient and Standardized System Design

Brian Summa, Assistant Professor, Department of Computer Science, Tulane University

Datasets many gigabytes to terabytes in size are being produced daily by scientists throughout the world. Examples include large simulations of physical phenomena being run on DOE supercomputers or large NSF supported infrastructures; large microscopy scans from NSF or NIH supported projects; or even climate models run at the National Center of Atmospheric Research. While crucial to the scientists who produce them, just a small number of these large datasets could have a transformative impact for others, if the data can be shared simply and easily with researchers. For instance, consider computer vision databases that collect, organize, and share the large number of photographs available online. Databases like ImageNet [1] have fostered huge leaps in vision and machine learning (16k citations in 10 years). Plug-ins even exist to query this database directly in interactive data analytics pipelines (e.g. Jupiter notebooks). Now consider that just a few hundred (~500) datasets from high-resolution, digital microscopy contain approximately the same number of pixels as the entire ImageNet [2].

Imagine the potential if such data could be made widely and easily available to researchers-at-large. This begs the obvious question on what would be necessary for people to construct *ImageNet-like* databases using their large datasets. The unfortunate truth is that even sharing a single large dataset can be a daunting task for researchers. Typically these data repositories are maintained by individual scientists or groups. These ad-hoc repositories are especially challenging when datasets get large. For instance, the repositories are often *links* to full-resolution raw data. In this scenario, datasets cannot be easily transferred, stored, or processed given their sheer size. Moreover, interactive queries are not supported, despite interactive analytics often being the key for novel scientific insights in new or complex phenomena. Alternatively, scalable, interactive queries can be provided through more advanced large data systems [3], although these systems are often highly-specialized, ad-hoc deployments with no standardization for data access across different platforms. In total, meaningful and wide dissemination of large data is not currently well-supported

The only solution to sharing these datasets is through large data systems that are easy to deploy, query, and scale. To this end, the ad-hoc deployments, divergent designs, non-standardized data access, and layering of heuristics that currently plague these systems must end. Work must be completed to model and standardize these many approaches. Although, this is a difficult task. For instance, systems for large data need to leverage (concurrently) several if not all of the following accelerators: high performance backends; preprocessing data while loading or idle; pre-caching or pre-fetching of data; or stream processing during data movement. Moreover, often scalable systems will need to support heterogeneous, distributed computing and storage resources. This leads to a highly complex interplay of many moving parts. Although, as this list above also shows there is justification that such standardization is possible given the many commonalities in design between these ad-hoc systems. Therefore, I argue that such standardization is not only necessary, but entirely possible given the proper support and effort. With such work, the rich datasets that are currently being lost to researchers-at-large due to their size can finally be widely disseminated.

[1] J. Deng et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[2] C. Mercan et al. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. *IEEE Transactions on Medical Imaging*, 37(1):316–325, 2017.

[3] Bethel, E. Wes, Hank Childs, and Charles Hansen, eds. *High performance visualization: Enabling extreme-scale scientific insight. CRC Press, 2012.*