doi:10.1017/S1431927616003561
Microsc. Microanal. 22 (Suppl 3), 2016

Interactive Web-based Spatio-Statistical Image Modeling from Gigapixel Images to Improve Discovery and Traceability of Published Statistical Models

Peter Bajcsy[1], Antoine Vandecreme[1], and Mary Brady[1]

[1]Information Technology Lab, National Institute of Standards and Technology, Gaithersburg, MD.

One of the main challenges for many scientific communities is reproducibility of published discoveries [1]. In the current publication process, manuscripts contain only data summaries and models. The trend in many state-of-the-art publication venues is to include references in the manuscripts that point to on-line accessible software and raw data [2]. However, this trend does not address the problems of reproducibility and traceability of published summaries and models to the individual contributing data points for analyses from microscopy images reaching Gigapixel and Terabyte sizes. The reasons lie in the extra expertise and infrastructure needed to share Gigapixel images and intermediate image features for visual validation. In addition, summarization and modelling steps during a discovery include user-driven selections of data points and statistical models. There is a need to design *discovery-assisting and summary validation software systems over very large images* that would advance discoveries and improve reproducibility of published scientific results.

In the context of discovery and traceability of published summaries and statistical models, we focus on the problems illustrated in Figure 1. The *first problem* of sharing Gigapixel images was addressed by designing on-line accessible stitching and pre-processing computations that deliver deep zoom (Google Maps like) visualization [3]. The *second problem* of sharing image features is achieved via an image feature extraction web system that delivers each feature value hyperlinked with its inputs, mathematical description, and software artifacts. The *third problem* of user-driven filtering is achieved by supporting spatial region and feature range selections. The spatial region selection of an image sub-region is enabled using the zoom and pan functions in Open Seadragon library [4][3]. The range selection of image features leverages the Pivot Viewer library [5]. The Pivot Viewer library also supports mapping of each datum into a histogram bin via visualization which addresses the *fourth problem* of traceability and visual validation of summaries. The *fifth problem* of general statistical models was addressed by our JavaScript development of estimators and generators for the Johnson family of probability distribution functions (PDFs). This four-parameter family of PDFs covers a large variety of one-dimensional distribution shapes. Finally, the *sixth problem* of making discoveries across multiple physical scales has been achieved by automatic hyperlinking of image objects and their statistical models across image resolutions.

To assist scientists in achieving traceability of published summaries and models from large microscopy images, we have pre-processed two 0.5 Gigapixel images of cell colonies (phase contrast and fluorescent channels) available at isg.nist.gov. The pre-processing includes tile stitching, flat field correction, colony segmentation, hexagon tessellation, feature extraction, and multi-resolution pyramid building. The images become available for analyses at three physical scales, such as cell colony, cell (approximated by hexagons), and pixel levels – see Figure 2. The total of 42 184 data visualizations are rendered from generic HTML and JavaScript templates. Each web rendering enables data filtering, summarization and PDF modeling. All renderings and computations are supported by Pivot Viewer, Open Seadragon, jQuery, D3, and MathJax JavaScript libraries with the total of around 100K lines of code.

Disclaimer: Commercial products are identified in this document in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the products identified are necessarily the best available for the purpose.

References:
[1] E. Marcus, "Credibility and reproducibility.," *Cell*, vol. **159**, no. 5, pp. 965–6, Nov. 2014.
[2] Editorial, "Repetitive Flaws," *Nature*, vol. **529**, p. 256, 2016.
[3] A. Vandecreme *et al*, "From Image Tiles to Web-Based Interactive Measurements in One Stop," in *Microscopy and Microanalyses*, 2015.
[4] "Open Seadragon," *Open Seadragon project*, 2015. [Online]. Available: http://openseadragon.github.io/. [Accessed: 15-Feb-2016].
[5] OpenLink, "Pivot Viewer," 2016. [Online]. Available: https://github.com/openlink/html5pivotviewer . [Accessed: 02-Feb-2016].

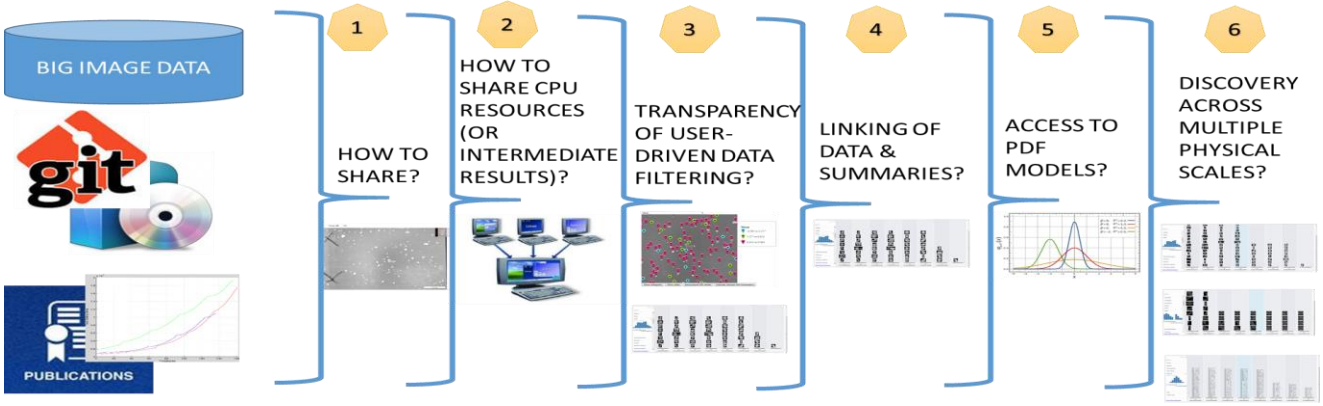

Figure 1: Overview of problems associated with making published summaries and models reproducible and traceable from Gigapixel images.

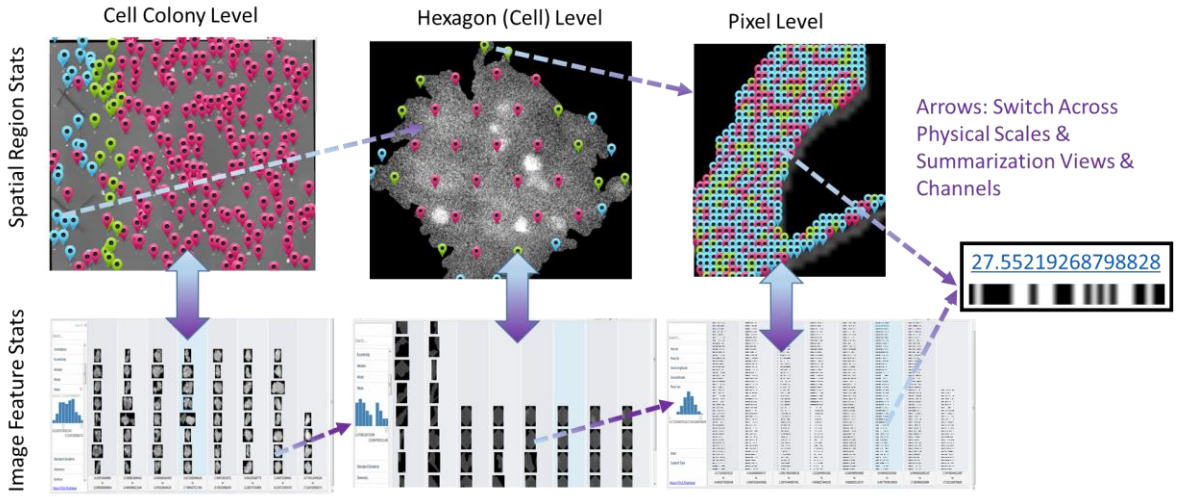

Figure 2: Illustration of spatial (top) and image feature (bottom) views across three physical scales (left – cell colonies, middle – hexagons (cells), right – pixels). All views are hyperlinked across channels, views, and physical scales, and enabled for user-driven filtering and for building data summaries and statistical PDF models.