Sharing and reusing cell image data

Assaf Zaritsky*
Lyda Hill Department of Bioinformatics and Department of Cell Biology, UT Southwestern Medical Center, Dallas, TX 75390; Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel

ABSTRACT The rapid growth in content and complexity of cell image data creates an opportunity for synergy between experimental and computational scientists. Sharing microscopy data enables computational scientists to develop algorithms and tools for data analysis, integration, and mining. These tools can be applied by experimentalists to promote hypothesis-generation and discovery. We are now at the dawn of this revolution: infrastructure is being developed for data standardization, deposition, sharing, and analysis; some journals and funding agencies mandate data deposition; data journals publish high-content microscopy data sets; quantification becomes standard in scientific publications; new analytic tools are being developed and dispatched to the community; and huge data sets are being generated by individual labs and philanthropic initiatives. In this Perspective, I reflect on sharing and reusing cell image data and the opportunities that will come along with it.

Monitoring Editor
David G. Drubin
University of California, Berkeley

Received: Mar 1, 2018
Revised: Apr 2, 2018
Accepted: Apr 6, 2018

BACKGROUND

Molecular cell biology and microscopy have undergone a revolution that led to an explosion in complex, dynamic, high-dimensional imaging data (Reynaud *et al.*, 2015; Ouyang and Zimmer, 2017). The lack of computational methods to extract information from such rich and high-content data is now becoming a critical bottleneck, and thus the field of cell imaging is in great need of computational scientists. However, there is a huge gap between biologists who produce, analyze, and hold the data, and computational scientists whose technical and analytical skills can enable extraction of more information from it (Figure 1). This gap is caused by differences in culture, communication, academic motivation, and reward.

One key step toward filling this gap is making cell image data publicly available. Data availability will attract computational scientists by exposing them to fresh and challenging problems at the interface of computer vision, data science, and cell biology. Just as in the emergence of bioinformatics, data availability will likely first engage computational scientists in development of tools and methods for analysis and data mining, before diving into deeper biological waters: integrating multiple data sets and examining "old" data from new perspectives to make new discoveries. Data depositors will profit from increased academic credit in publications, citations, and new collaborations. Cell biologists will enjoy the availability of new computational methods and complementary data sets to reproduce and validate their findings. This synergy will benefit all parties and move cell biology forward.

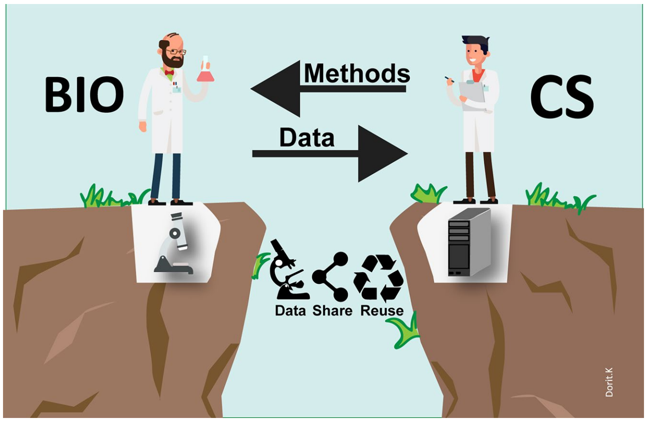

FIGURE 1: The gap between cell biology and computer science has roots in different cultural aspects and lack of cross-discipline communication. Availability of large-scale data sets will make a significant step toward bridging this gap. Scientists with computational backgrounds (CS, computer science) will be motivated to exercise their skills in data integration, mining, and tool development to benefit cell biology (BIO) through availability of new computational tools to analyze and interpret cell image data. Credit: Dorit Kochavi.

DOI:10.1091/mbc.E17-10-0606

*Address correspondence to: Assaf Zaritsky (assafzar@gmail.com).

Abbreviations used: API, application programming interface; CMSO, Cell Migration Standardization Organization.

Deposition of data in public repositories upon publication has become the standard in many fields such as gene expression, three-dimensional protein structure and proteomics. Sharing enables reanalysis of the data by other scientists for replication, computational tool development and turning data to discovery. This is not the case for cell image data, where the complexity, multidimensionality, variability in experimental settings, lack of standardization and huge content complicate sharing and reuse. Here, I discuss the barriers toward open cell image data, the steps that are required to enable effective sharing and reuse, and the expected benefits to follow.

THE IMPACT OF IMAGE DATA SHARING

Public availability of cell image data is essential for improving reproducibility, assessment, and validation of new computational methods; integration and mining multiple data sets; and quantitatively examining previously published data from new angles to facilitate discovery (Pasquetto *et al.*, 2017). First and most obvious, deposition of image data in public repositories can help with what is referred to as the "reproducibility crisis" (Baker, 2016). Open data can reduce data cherry picking, enable independent validation of previous research outputs, simplify replication studies and allow generalization of conclusions to additional cell or experimental systems.

Most computational scientists in the field of cell imaging are focused on developing analytic tools for common universal computational problems such as preprocessing steps, registration, detection, segmentation, tracking, feature extraction, and classification (Meijering *et al.*, 2016). A critical aspect when presenting a new method is comparing its performance to alternative approaches. Accordingly, most current examples of reusing cell image data are aimed at the validation and assessment of new computational tools. For example, the Mitocheck project created a resource of genomewide phenotypic profiling (Neumann, Walter, *et al.*, 2010). Its image and image-derived data were reused to develop multiple methods such as inferring gene networks (Failmezger, Praveen *et al.*, 2013b), predicting gene function from RNAi-induced phenotypic similarities (Serrano-Solano *et al.*, 2017), unsupervised phenotyping (Failmezger *et al.*, 2013a), quantification of single cell motility in high-throughput time-lapse screening data (Schoenauer Sebag *et al.*, 2015), and cell tracking (Lou and Hamprecht, 2011) (that also reused data from Li *et al.*, 2010). WND-CHARM, an image classification framework (Orlov *et al.*, 2008), used published cell images for benchmarking (Boland *et al.*, 1998; Boland and Murphy, 2001), and CP-CHARM, a CellProfiler-based image classification method (Uhlmann *et al.*, 2016), was validated also with additional data sets from the Broad Bioimage Benchmark Collection (Ljosa *et al.*, 2012). AveMap, a method to quantify monolayer migration (Deforet, Parrini, *et al.*, 2012) was verified on data from Simpson *et al.* (2008). Osokin *et al.* (2017) applied deep learning to infer the localization of one protein based on the spatial pattern of another protein; to train their model they used an existing data set (Dodgson, Chessel, Vaggi, *et al.*, 2017). Community competitions and benchmarking efforts, using curated standardized data, ground-truth annotations, and performance metrics, have proven effective at objectively comparing methods for particle tracking (Chenouard, Smal, De Chaumont, Maška, *et al.*, 2014), single molecule localization (Sage *et al.*, 2015), cell tracking (Ulman, Maška, *et al.*, 2017), nucleus detection (www.kaggle.com/c/data-science-bowl-2018), and other methods (Meijering *et al.*, 2016). The benefit of competitions is twofold: users can more effectively select an existing method that is likely to perform well for their specific data, and method developers can demonstrate superiority of their method—a critical aspect for publication in the fields of applied computational sciences.

Secondary analysis of data can also produce new biological insight. In this mode, researchers extract new information from raw image data or image-derived features (e.g., cell/molecular trajectories, segmentation masks) by applying new analyses that were not considered in the original study. For example, data from the First World Cell Race (Maiuri *et al.*, 2012), a large-scale comparison of cell motility across 54 different adherent cell types, was reused by the same group to confirm an association between cell speed and persistent migration and then extended by a set of new experiments and theory to reveal a universal coupling that is mediated by actin flows (Maiuri, Rupprecht, Wieser, *et al.*, 2015). Lavi *et al.* (2016) introduced theory for competition between cell motility machinery and microbial antigen capture for myosin II. Their prediction that cells switch from persistent migration to unidirectional self-oscillation was validated by reanalyzing cell trajectories from Chabaud, Heuzé, *et al.* (2015). Meyers, Craig, and Odde (2006) tested their theoretical predictions linking cell size and shape to signaling by new analyses of previous data of Cdc42 activation in fibroblasts (Nalbant, Hodgson, *et al.*, 2004). Ji *et al.* reanalyzed raw dual-channel time-lapse imaging (Hu, Ji, *et al.*, 2007) to dissect the relationship between predicted adhesion forces and F-actin-vinculin interactions (Ji *et al.*, 2008). Abdullah *et al.* segmented individual cells in an epithelial tissue reusing time lapse images of developing *Drosophila* pupae from Besson, Bernard, *et al.* (2015). They demonstrated that the probability of cell division increases exponentially with the number of cell edges and developed theory to propose that this is responsible for the observed cell-edge distribution (Abdullah *et al.*, 2017). Thurley *et al.* developed "response-time modeling" as a framework to unify and interpret knowledge on intra- and intercellular signaling pathways (Thurley *et al.*, 2018) and applied it to published experimental and simulated cytokine secretion data (Dorner, Dorner, Zhou, *et al.*, 2009; Busse *et al.*, 2010; Han, Bagheri, *et al.*, 2012; Thurley Gerecht, Friedmann, and Höfer, 2015). Yang and Svitkina carefully reassessed published electron tomography data from Urban *et al.* (2010) and reported the existence of numerous branched actin filaments in lamellipodia that have been overlooked in the original study reporting their absence (Yang and Svitkina, 2011).

In my own research, I reanalyzed data from Serra-Picamal, Conte, *et al.* (2012) to test plithotaxis—the tendency for each individual cell within a collectively migrating monolayer to migrate along the local orientation of the maximal principal stress (Tambe, Hardin, *et al.*, 2011). I found that plithoatxis is a property attributed to a small subgroup of cells that migrate more effectively (Zaritsky *et al.*, 2015b). By designing new algorithms, I discovered that coordinated stress precedes coordinated motion and by reanalyzing published data from another group (Das *et al.*, 2015), proposed that several tight-junction proteins play a role in transmission of aligned stress to aligned motion (Zaritsky *et al.*, 2015b). I also used the same data to validate a new method developed for quantification of colocalization and coalignment data (Zaritsky *et al.*, 2017).

Integrating data sets from multiple sources have the potential to discover patterns that are not possible to infer from individual studies (Lahat *et al.*, 2015). Williams, Moore, Li, *et al.* (2017) recently illustrated the benefit of image-derived data integration. By combining information from three independent data sets (Fuchs, Pau, *et al.*, 2010; Rohn *et al.*, 2011; Graml, Studera, Lawson, Chessel, *et al.*, 2014), the authors suggested a new gene-network controlling cell shape that could not be inferred from single studies. This first example is setting the stage for future ambitious "meta-analysis" studies, for example, integration of imaging and omics-based data sets for system-level genotype-phenotype mapping.

Current advances in cell biology resemble the famous parable of the blind men and an elephant, in which each blind man inspects a different part of the elephant body, revealing a limited aspect of reality. Conceptualizing the elephant requires integration of all partial observations. Similarly, the ultimate impact in opening cell image data could come from the integration of partial observations to a complete understanding of the "biological scene." This will be achieved by the integration of complementary data and by the use of complementary tools that examine data from different perspectives.

BARRIERS AND SOLUTIONS TOWARD OPEN CELL IMAGE DATA

Cultural conventions and the lack of infrastructure are the main hurdles limiting image data sharing and reuse. In recent years we have witnessed independent seeds of infrastructure building, generation of high-content data sets and a growing appreciation of interdisciplinary and collaborative science. Together these seeds mark a beginning of a time to open cell image data and import concepts of "big data" science to cell biology.

Convincing biologists to open their data is not trivial. Historically, cell biology lacks a culture of data sharing, and experimentalists are still accustomed to holding on to their image data. For example, in response to my request for published image data (in a "glamour" journal with a "data availability on request" statement), the corresponding author refused with the argument that "This might not be the best way to optimize scientific progress, but given the current rules of the game, it is what it is." This is apparently a more general problem, not limited solely to cell image data, as implicated by a recent notorious editorial that raised the concern that "the system will be taken over" by what they called "research parasites" (Longo and Drazen, 2016). As a nice response, Casey Greene (Pennylvania State University) has initiated a "Parasite Award" (http://researchparasite.com/) for scientists practicing secondary data analysis and the "Symbiont Award" (http://researchsymbionts.org/) for experimentalists who shared their data.

Although the vast majority of experimentalists agree with the general concept that data produced from public funds should become publicly available for general reuse, many feel genuinely frustrated about the idea of putting in the extra work to allow others to benefit from the expensive and labor-intensive data they generated. More specific concerns are that the lack of detailed understanding of the primary research could lead to erroneous conclusions and competition with others when additional analyses were planned (Longo and Drazen, 2016). Rewarding the sharing of primary research data could be key in changing this paradigm (Wallis *et al.*, 2013). For example, it is established that papers with open data receive more citations over long time frames (Piwowar and Vision, 2013); funding agencies could credit researchers whose data are reused by others. To reach solid conclusions, data scientists must understand the full extent of the biological complexity in the data (Zaritsky, 2016), which almost always require direct communication with the data generators. Such interaction could even lead to closer partnerships and sought-after experimental validations of hypotheses that emerged from the secondary analyses. Thus, researchers who practice secondary analyses should strive to involve the primary data generators as collaborators and share credit.

Image data sets with potential for reuse have scientific value on their own and so deserve direct academic credit. Accordingly, on deposition, data sets are assigned a unique identifier (doi) that can be later referenced independently of the scientific study. *BMC Research Notes* was the first journal to identify the potential impact, introducing "data format" as a new article type (O'Donnell *et al.*, 2008). Since then, many journals introduced Resources or Tools/Methods/Software article types. GigaScience was the first to focus on data and other research products as its main publication entity, also providing means for data deposition (Sneddon *et al.*, 2012) (http://gigadb.org/). This initiative was followed by Nature's Scientific Data. Data sets that are selected for publication must follow the findable, accessible, interoperable, and reusable (FAIR) principles (Wilkinson *et al.*, 2016) and be assessed based on their soundness and potential for future reuse. Examples of cell-image-based data sets include Zaritsky *et al.* (2015a), Bray *et al.* (2017), Pascual-Vargas *et al.* (2017), and Lukeš *et al.* (2018).

It is not only a change of culture that is needed. The size, complexity, variability, and lack of standardized formats and metadata make the deposition process labor intensive and tedious, thus discouraging experimentalists. Data curation and deposition should become as simple and straightforward as possible and ideally provide academic reward to encourage experimentalists to share their data.

Unlike omics data, until very recently there was no public repository dedicated for large-scale imaging data (Lemberger, 2015). The Image Data Resource (https://idr.openmicroscopy.org/about/) (Williams, Moore, Li, *et al.*, 2017) is an open online platform for publishing, visualizing, and mining high-content cell image data sets. It contains curated large-scale data sets with raw and processed image data, together with phenotypic annotations, standardized vocabulary, and software infrastructure to allow straightforward querying and visualization. Earlier image repositories did not provide a complete suite to enable these functionalities. These include Yale Image Finder (Xu *et al.*, 2008), retrieving images based on their textual description; PhenoImageShare (Adebayo *et al.*, 2016), the first to provide infrastructure linking annotated ontologies-based tags to phenotypic-based user queries; the JCB Data Viewer (http://jcb-dataviewer.rupress.org/) (Hill, 2008); The Cell Image Library (http://www.cellimagelibrary.org/) (Orloff *et al.*, 2013); and the Systems Science of Biological Dynamics database (http://ssbd.qbic.riken.jp) (Tohsato, Ho, *et al.*, 2016). The Broad BioImage Benchmark Collection (https://data.broadinstitute.org/bbbc/) (Ljosa *et al.*, 2012) is a resource for methods benchmarking. It includes raw data sets, ground-truth annotations for benchmarking and criteria for evaluating each data set.

The complexity and costs of storing and managing large-scale image data cannot be overemphasized. In the near future, storage requirements, even at a single institution level, will easily exceed the petabyte (1 PB = 1000 TB) scale (Ouyang and Zimmer, 2017). Such content poses great challenges in terms of storage, retrieval, and mining. In a recent white paper, Ellenberg, Swedlow, *et al.* (2018) proposed a two-layered model where a *data archive* is used for data and metadata storage and access and *added-value* databases, a subset of the data archive, is identified as having greater potential for reuse by the community and provided with additional curation, annotation, and standardization. Obvious data sets that fit these criteria include atlases and high-content genetic phenotypic screens. Recent advances in molecular biology, microscopy, and automation enable the generation of such data sets even at individual labs (e.g., Cai, Hossain, *et al.*, 2017). However, large-scale imaging consortia, community efforts, and philanthropic projects have the potential to produce larger-scale and more controlled image-data resources. One example is the Allen Institute of Cell Science, having the ultimate goal of building an integrated model of cell structure, organization, and function. Toward this goal, they genome-edited human stem cells (Roberts, Haupt, *et al.*, 2017) and made highly standardized microscopy raw data and image-derived features publicly available through their Cell Explorer (www.allencell.org/image-data-downloads.html) (Horwitz, 2016). Other examples include Euro-Bioimaging (www.eurobioimaging.eu/), the Chan Zuckerberg initiative (https://chanzuckerberg.com/) (Bargmann, 2018), MultiMOT (https://multimot.org/) (Masuzzo and Martens, 2015), and the Human Cell Atlas (www.humancellatlas.org/) (Regev *et al.*, 2017). The reuse potential of more specific imaging studies is less obvious and archiving would be determined per-case based on the community priorities (Ellenberg, Swedlow, *et al.*, 2018). Infrastructure such as SourceData (Liechti, George, Götz, El-Gebali, *et al.*, 2017), which links published figures to their underlying source data, can improve reproducibility in cases where the raw data are not archived. Storage costs and personnel (software engineers, data curators, and high performance computing specialists) are the major financial expenses essential to build and maintain large data repositories. Involvement of government support, funding agencies, industrial partners, and philanthropy is crucial, especially for the recognition and support of software development and high-performance computing needed to embed informatics as an integral part of advancing cell biology and for long-term maintenance of large-scale repositories and tools (Cardona and Tomancak, 2012; Prins, De Ligt, *et al.*, 2015).

A key aspect toward successful data dissemination and mining is organizing the metadata for flexible retrieval and interrogation, while keeping data deposition simple and fast. This fine balance poses challenges in data standardization, storage, and retrieval even more prominently in the complex landscape of cell imaging. Standardized data formats, community reporting guidelines, Application programming interfaces (APIs, which are software infrastructure to simplify use to technologies in developing applications), and visualization tools must be defined and developed toward this goal. Standard domain terminology, formally termed *controlled vocabulary*, enables harmonized data representation, which is necessary for querying across data sets. Many of the terms in a controlled vocabulary can be borrowed from existing ontologies, which are formalized hierarchical descriptions of a specific domain. Several ontologies describing experimental assays, cell types, and their phenotypes and behavior were recently defined (Visser *et al.*, 2011; Hoehndorf *et al.*, 2012; Sarntivijai *et al.*, 2014; Sluka *et al.*, 2014; Jupp *et al.*, 2016), and new ontologies can be constructed on need.

The *minimal reporting requirements* define the smallest set of metadata required to enable future data querying and integration. These include information on the model system, experiments, and microscopy used to generate the data, all in terms of the controlled vocabulary. Deciding on the minimal set tunes reuse possibilities with ease of data submission and should be carefully determined. Cell image data come in a wide range of proprietary and open file formats (Linkert *et al.*, 2010). Controlled vocabularies can also aid in standardizing these data formats and implementing APIs to handle this data harmonization (Orchard *et al.*, 2005). A successful example is the Open Microscopy Environment (Goldberg *et al.*, 2005), providing the software infrastructure to store, share, and query image data from different sources.

Distribution of image-derived features along the raw image data can be extremely beneficial in terms of reuse. The Cell Migration Standardization Organization (CMSO), a community effort toward the development of community standards for the field of cell migration that I am part of, has currently taken the challenge of defining and implementing data formats to harmonize routine analyses outputs, starting with a data format for trajectories of cells or molecular events (https://github.com/CellMigStandOrg/biotracks). Similar steps have been made by Rigano and Strambio-De-Castillia, who just released a cross-platform data management infrastructure for particle tracking data (http://omega.umassmed.edu/) (Rigano *et al.*, 2018), using their minimal reporting requirements (Rigano and De Castillia, 2017). Efforts have been also made in the arena of mathematical modeling (Macklin and Friedman, 2018), for example, Multi Cellular Data Standard, a new data format for multicellular data (Friedman *et al.*, 2016).

Altogether, the emerging infrastructures of data-rich repositories, controlled vocabularies, minimal reporting requirements, and APIs will enable data-focused exploratory or hypothesis-driven queries that will lead to new discoveries.

RECRUITING COMPUTATIONAL SCIENTISTS TO CELL BIOLOGY THROUGH OPEN DATA

The question of how far secondary analysis can take us has already been answered in other fields. For example, in the omics fields, computational scientists who never held a pipette can be the ones driving the biological interrogation. They, of course come along with others who develop computational tools and who provide bioinformatics services in core facilities, as a range of individuals contributing in different ways to the field. For cell imaging, even when the data become widely available, cultural barriers still stand in the way of attracting computational scientists, including appreciation of secondary-analysis research by the cell biology community, adequate academic reward and career opportunities, and cross-disciplinary training.

Development of tools facilitate access to information previously unattainable and so can be more scientifically impactful than making a scientific discovery. In the past few years we have seen Nobel laureates who developed cryoelectron and superresolved fluorescence microscopy. We have also witnessed how software tools, such as BLAST (sequence analysis) (Altschul *et al.*, 1990), changed the way that science is performed and greatly helped in the emergence of bioinformatics as an independent field. Public data and benchmarks are necessary to make this leap in what is called *bioimage informatics* (Peng, 2008), but this is not sufficient. Adaptation of existing algorithms to cell image data, dealing with the inherent variability and noise in experimental biology and paying attention to software engineering and usability usually lacks the mathematical rigor and novelty sought in top-tier computer science, while also lacking the biological novelty sought in cell biology (Cardona and Tomancak, 2012). Another discrepancy comes from the underlying motivation: for applied computer scientists elegant math and outperforming the state of the art (even by a margin) are the goals, whereas studies in cell biology are motivated by better understanding of a specific biological process, using whatever available techniques (Meijering *et al.*, 2016). We have seen multiple success stories in the past 15 years—bioimage analysis tools that are used by thousands of scientists around the globe (e.g., Carpenter *et al.*, 2006; Sommer *et al.*, 2011; de Chaumont *et al.*, 2012; Schneider *et al.*, 2012). A positive action was made by journals offering new article types focusing on tool and software development, for example, PLoS Biology's "Meta-Research" section on data-driven and meta-analytic research (Kousta *et al.*, 2016). Another positive sign is the growing community of bioimage analysts (in Europe, NEUBIAS, http://neubias.org/), as a bridge between tool developers and biology end users.

Roles of computational scientists in cell biology can go beyond number crunching, statistical analyses, and tool development. The mass and complexity of microscopy data create an opportunity for computational scientists to decipher fundamental biological processes by analyzing image data. Of course, discovery of complex dynamic patterns requires deep domain knowledge of the biological process, the experimental possibilities, the type of information that can be extracted, and the computational tools to extract it. This notion was shouted out in several recent opinion pieces. Markowetz (2017) argued that "computational biologists are just biologists using a different tool," and I introduced the "dry cell biologist" and its role in promoting interdisciplinary team science (Zaritsky, 2016), which was also proposed as an evident mode of performing modern cell biology by Horwitz (2016). To facilitate this type of research, results should not be assessed based on the amount of new data generated but instead on the findings and biological insight extracted from the data regardless of the purpose for which they were originally generated. This problem was evidently stated by an anonymous reviewer of one of my manuscripts, "…a major criticism of the current manuscript, which identifies real and important relations, is using previously published data instead of independently performing the experiment."

Big data are not the end of hypothesis-driven science (Mazzocchi, 2015). It is true that mining through massive cell image public data sets, extracting complex patterns and turning it to new biological insight, is going to become a more prevalent mode of science in the near future. However, the concern that discovery-driven data science may find spurious correlations with no biological interpretation is real and must be addressed by a careful examination of different experimental methods and systems. For a dry cell biologist, this can be achieved by engaging an experimentalist collaborator to jointly decipher a pattern discovered via mining existing data or, alternatively, by finding independent existing data sets and developing new analyses to test the hypotheses from different angles.

After all is said and done, the most influential "carrots and sticks" toward open data and engaging scientists with solid computational background in cell biology are in the hands of journals and funding agencies. Similarly to the omics fields, they should become more involved: funders should require the sharing of cell image data, and journals must enforce public dissemination to ensure reproducibility. Funding agencies should promote collaborative projects, fund software solutions for cell imaging big data, and explicitly support maintenance of such projects and experts in high-performance computing and software engineering (Cardona and Tomancak, 2012; Prins, De Ligt, *et al.*, 2015). New multidisciplinary education and training programs must be established to bridge the technical and cultural gap between the disciplines (Meijering *et al.*, 2016).

NOW IS THE TIME!

New techniques in genome engineering and microscopy facilitate the generation of high-content cell image data. At the same time, cell biology is advancing to more physiologically relevant and complex systems. Together, the content and complexity of this new generation of cell image data are making visual assessment impossible. Data scientists are needed to develop tools to quantify these data and decipher the complex patterns that are encapsulated in them. One key step toward engaging scientists from a computational background to cell biology is to open data. These issues were discussed in a special interest subgroup that I organized last year at the American Society for Cell Biology (ASCB)|EMBO annual meeting (https://assafzar.wixsite.com/ascb2017-subgroup), during which three key components were highlighted (Figure 2): 1) changing cultural barriers regarding sharing primary research data, rewarding depositors, recognition of results derived by secondary analysis and scientists who specialize in it, and improving communication between the fields; 2) building infrastructure to enable easy data deposition and mining: repositories, standardized data formats, APIs and visualization tools; and 3) developing new computational methods to deal with the inherent complexity and variability of these data. Although many encouraging signs suggest that the field is moving in this direction, there are still plenty of challenges ahead. Cell image data science is expected to face similar computational challenges to those of other "big data" fields, such as genomics, experience in data acquisition, storage, distribution, and analysis (Stephens *et al.*, 2015). Exciting times lie ahead of us—come and join in!

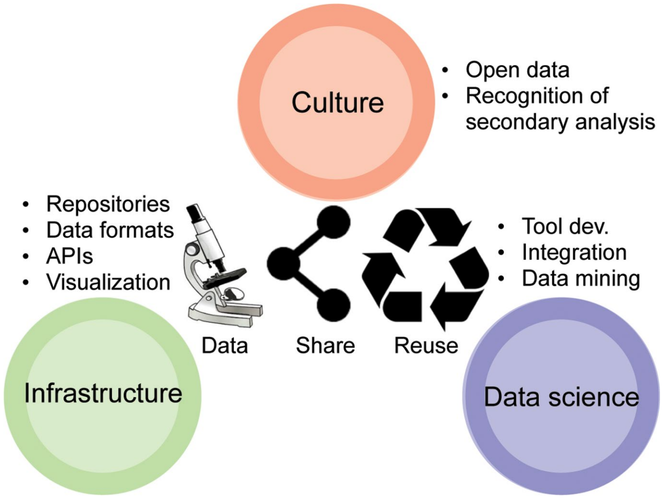

FIGURE 2: The components needed to bring data science to cell biology and the role of sharing image data for secondary analysis.

ACKNOWLEDGMENTS

I acknowledge support by a National Institutes of Health Grant to Gaudenz Danuser (P01 GM103723). I thank the following researchers for sharing with me their imaging data for secondary analysis that have led to publications: Xavier Serra-Picamal and Xavier Trepat; Tamal Das and Joachim Spatz; and Yun-Yu Tseng, M. Angeles Rabadán, and Alan Hall. I thank the members of the CMSO, where I first started to think about what infrastructure is needed for opening cell migration data. In particular, I thank Paola Masuzzo, who was the driving engine behind CMSO, and initiated, together with me, the ASCB subgroup on sharing and reusing cell image data. I also thank Lennart Martens and MultiMOT for sponsoring the subgroup and all who participated in it. I am thankful to Dorit Kochavi for drawing Figure 1. I thank Zvi Kam, Meghan Driscoll, Yaron Orenstein, Arieh Zaritsky, Kota Miura, Amir Rubinstein, and Sandy Schmid for discussions and constructive feedback on the manuscript. Last, I thank Gaudnez Danuser for his mentorship and especially for providing me with the freedom to drift from my official research projects to spend time on independent community efforts such as CMSO that led to the writing of this Perspective.

REFERENCES

Boldface names denote co–first authors.

Abdullah A, Avraam D, Chepizhko O, Vaccari T, Zapperi S, La Porta CA, Vasiev B (2017). Universal statistics of epithelial tissue topology. arXiv: 1710.08527.

Adebayo S, McLeod K, Tudose I, Osumi-Sutherland D, Burdett T, Baldock R, Burger A, Parkinson H (2016). PhenoImageShare: an image annotation and query infrastructure. J Biomed Semant 7, 35.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. J Mol Biol 215, 403–410.

Baker M (2016). Reproducibility crisis? Nature 533, 26.

Bargmann C (2018). How the Chan Zuckerberg Science Initiative plans to solve disease by 2100. Nature 553, 19–21.

Besson C, Bernard F, Corson F, Rouault H, Reynaud E, Keder A, Mazouni K, Schweisguth F (2015). Planar cell polarity breaks the symmetry of PAR protein distribution prior to mitosis in Drosophila sensory organ precursor cells. Curr Biol 25, 1104–1110.

Boland MV, Markey MK, Murphy RF (1998). Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry 33, 366–375.

Boland MV, Murphy RF (2001). A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics 17, 1213–1223.

Bray M-A, Gustafsdottir SM, Ljosa V, Singh S, Sokolnicki KL, Bittker JA, Bodycombe NE, Dančík V, Hasaka TP, Hon C (2017). A dataset of images and morphological profiles of 30,000 small-molecule treatments using the Cell Painting assay. GigaScience 6, 1–5.

Busse D, de la Rosa M, Hobiger K, Thurley K, Flossdorf M, Scheffold A, Höfer T (2010). Competing feedback loops shape IL-2 signaling between helper and regulatory T lymphocytes in cellular microenvironments. Proc Natl Acad Sci USA 107, 3058–3063.

Cai Y, Hossain MJ, Heriche J-K, Politi AZ, Walther N, Koch B, Wachsmuth M, Nijmeijer B, Kueblbeck M, Martinic M (2017). An experimental and computational framework to build a dynamic protein atlas of human cell division. bioRxiv 227751.

Cardona A, Tomancak P (2012). Current challenges in open-source bioimage informatics. Nat Methods 9, 661–665.

Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, Guertin DA, Chang JH, Lindquist RA, Moffat J, *et al*. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol 7, R100.

Chabaud M, Heuzé ML, Bretou M, Vargas P, Maiuri P, Solanes P, Maurin M, Terriac E, Le Berre M, Lankar D, *et al*. (2015). Cell migration and antigen capture are antagonistic processes coupled by myosin II in dendritic cells. Nat Commun 6, 7526.

Chenouard N, Smal I, De Chaumont F, Maška M, Sbalzarini IF, Gong Y, Cardinale J, Carthel C, Coraluppi S, Winter M, *et al*. (2014). Objective comparison of particle tracking methods. Nat Methods 11, 281–289.

Das T, Safferling K, Rausch S, Grabe N, Boehm H, Spatz JP (2015). A molecular mechanotransduction pathway regulates collective migration of epithelial cells. Nat Cell Biol 17, 276–287.

de Chaumont F, Dallongeville S, Chenouard N, Herve N, Pop S, Provoost T, Meas-Yedid V, Pankajakshan P, Lecomte T, Le Montagner, *et al*. (2012). Icy: an open bioimage informatics platform for extended reproducible research. Nat Methods 9, 690–696.

Deforet M, Parrini MC, Petitjean L, Biondini M, Buguin A, Camonis J, Silberzan P (2012). Automated velocity mapping of migrating cell populations (AVeMap). Nat Methods 9, 1081–1083.

Dodgson J, Chessel A, Vaggi F, Giordan M, Yamamoto M, Arai K, Madrid M, Geymonat M, Abenza JF, Cansado J (2017). Reconstructing regulatory pathways by systematically mapping protein localization interdependency networks. bioRxiv 116749.

Dorner BG, Dorner MB, Zhou X, Opitz C, Mora A, Güttler S, Hutloff A, Mages HW, Ranke K, Schaefer M (2009). Selective expression of the chemokine receptor XCR1 on cross-presenting dendritic cells determines cooperation with CD8+ T cells. Immunity 31, 823–833.

Ellenberg J, Swedlow JR, Barlow M, Cook CE, Patwardhan A, Brazma A, Birney E (2018). Public archives for biological image data. arXiv: 1801.10189.

Failmezger H, Fröhlich H, Tresch A (2013a). Unsupervised automated high throughput phenotyping of RNAi time-lapse movies. BMC Bioinformat 14, 292.

Failmezger H, Praveen P, Tresch A, Fröhlich H (2013b). Learning gene network structure from time laps cell imaging in RNAi Knock downs. Bioinformatics 29, 1534–1540.

Friedman SH, Anderson AR, Bortz DM, Fletcher AG, Frieboes HB, Ghaffarizadeh A, Grimes DR, Hawkins-Daarud A, Hoehme S, Juarez EF, *et al*. (2016). MultiCellDS: a standard and a community for sharing multicellular data. bioRxiv 090696.

Fuchs F, Pau G, Kranz D, Sklyar O, Budjan C, Steinbrink S, Horn T, Pedal A, Huber W, Boutros M (2010). Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. Mol Syst Biol 6, 370.

Goldberg IG, Allan C, Burel J-M, Creager D, Falconi A, Hochheiser H, Johnston J, Mellen J, Sorger PK, Swedlow JR (2005). The Open Microscopy Environment (OME) Data Model and XML file: open tools for informatics and quantitative analysis in biological imaging. Genome Biol 6, R47.

Graml V, Studera X, Lawson JL, Chessel A, Geymonat M, Bortfeld-Miller M, Walter T, Wagstaff L, Piddini E, Carazo-Salas RE (2014). A genomic Multiprocess survey of machineries that control and link cell shape, microtubule organization, and cell-cycle progression. Dev Cell 31, 227–239.

Han Q, Bagheri N, Bradshaw EM, Hafler DA, Lauffenburger DA, Love JC (2012). Polyfunctional responses by human T cells result from sequential release of cytokines. Proc Natl Acad Sci USA 109, 1607–1612.

Hill E (2008). Announcing the JCB DataViewer, a browser-based application for viewing original image files. J Cell Biol 183, 969.

Hoehndorf R, Harris MA, Herre H, Rustici G, Gkoutos GV (2012). Semantic integration of physiology phenotypes with an application to the Cellular Phenotype Ontology. Bioinformatics 28, 1783–1789.

Horwitz R (2016). Interdisciplinary Team Science in Cell Biology. Trends Cell Biol 26, 796–798.

Hu K, Ji L, Applegate KT, Danuser G, Waterman-Storer CM (2007). Differential transmission of actin motion within focal adhesions. Science 315, 111–115.

Ji L, Lim J, Danuser G (2008). Fluctuations of intracellular forces during cell protrusion. Nat Cell Biol 10, 1393–1400.

Jupp S, Malone J, Burdett T, Heriche J-K, Williams E, Ellenberg J, Parkinson H, Rustici G (2016). The cellular microscopy phenotype ontology. J Biomed Semant 7, 28.

Kousta S, Ferguson C, Ganley E (2016). Meta-research: broadening the scope of PLOS Biology. PLoS Biol 14, e1002334.

Lahat D, Adali T, Jutten C (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. Proc IEEE 103, 1449–1477.

Lavi I, Piel M, Lennon-Duménil A-M, Voituriez R, Gov NS (2016). Deterministic patterns in cell motility. Nat Phys 12, 1146–1152.

Lemberger T (2015). Image Data in Need of a Home, Heidelberg, Germany: EMBO Press.

Li F, Zhou X, Ma J, Wong ST (2010). Multiple nuclei tracking using integer programming for quantitative cancer cell cycle analysis. IEEE Trans Med Imag 29, 96–105.

Liechti R, George N, Götz L, El-Gebali S, Chasapi A, Crespo I, Xenarios I, Lemberger T (2017). SourceData: a semantic platform for curating and searching figures. Nat Methods 14, 1021.

Linkert M, Rueden CT, Allan C, Burel J-M, Moore W, Patterson A, Loranger B, Moore J, Neves C, MacDonald D, *et al*. (2010). Metadata matters: access to image data in the real world. J Cell Biol 189, 777–782.

Ljosa V, Sokolnicki KL, Carpenter AE (2012). Annotated high-throughput microscopy image sets for validation. Nat Methods 9, 637–637.

Longo DL, Drazen JM (2016). Data sharing. N Engl J Med 374, 276–277.

Lou X, Hamprecht FA (2011). Structured learning for cell tracking. Adv Neural Inf Process Syst, 1296–1304.

Lukeš T, Pospíšil J, Fliegel K, Lasser T, Hagen GM (2018). Quantitative super-resolution single molecule microscopy dataset of YFP-tagged growth factor receptors. GigaScience 7, 1–10.

Macklin P, Friedman SH (2018). Open source tools and standardized data in cancer systems biology. bioRxiv, 244319.

Maiuri P, Rupprecht J-F, Wieser S, Ruprecht V, Bénichou O, Carpi N, Coppey M, De Beco S, Gov N, Heisenberg C-P, *et al*. (2015). Actin flows mediate a universal coupling between cell speed and cell persistence. Cell 161, 374–386.

Maiuri P, Terriac E, Paul-Gilloteaux P, Vignaud T, McNally K, Onuffer J, Thorn K, Nguyen PA, Georgoulia N, Soong D (2012). The first world cell race. Curr Biol 22, R673–R675.

Markowetz F (2017). All biology is computational biology. PLoS Biol 15, e2002050.

Masuzzo P, Martens L (2015). An open data ecosystem for cell migration research. Trends Cell Biol 25, 55–58.

Mazzocchi F (2015). Could big data be the end of theory in science?: A few remarks on the epistemology of data-driven science. EMBO Rep 16, 1250–1255.

Meijering E, Carpenter AE, Peng H, Hamprecht FA, Olivo-Marin J-C (2016). Imagining the future of bioimage analysis. Nat Biotechnol 34, 1250–1255.

Meyers J, Craig J, Odde DJ (2006). Potential for control of signaling pathways via cell size and shape. Curr Biol 16, 1685–1693.

Nalbant P, Hodgson L, Kraynov V, Toutchkine A, Hahn KM (2004). Activation of endogenous Cdc42 visualized in living cells. Science 305, 1615–1619.

Neumann B, Walter T, Hériché J-K, Bulkescher J, Erfle H, Conrad C, Rogers P, Poser I, Held M, Liebel U, *et al.* (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature 464, 721–727.

O'Donnell PJ, Duke WH, Pantanowitz L (2008). Absence of human herpes virus-8 (HHV8) in nephrogenic systemic fibrosis. BMC Res Notes 1, 82.

Orchard S, Montecchi-Palazzi L, Hermjakob H, Apweiler R (2005). The use of common ontologies and controlled vocabularies to enable data exchange and deposition for complex proteomic experiments. In: Biocomputing 2005, Singapore: World Scientific, 186–196.

Orloff DN, Iwasa JH, Martone ME, Ellisman MH, Kane CM (2013). The cell: an image library-CCDB: a curated repository of microscopy data. Nucleic Acids Res 41, D1241–D1250.

Orlov N, Shamir L, Macura T, Johnston J, Eckley DM, Goldberg IG (2008). WND-CHARM: Multi-purpose image classification using compound image transforms. Pattern Recogn Lett 29, 1684–1693.

Osokin A, Chessel A, Salas REC, Vaggi F (2017). GANs for Biological Image Synthesis. 2017 IEEE Int Conf Comput Vis, 2252–2261.

Ouyang W, Zimmer C (2017). The imaging tsunami: computational opportunities and challenges. Curr Opin Syst Biol 4, 105–113.

Pascual-Vargas P, Cooper S, Sero J, Bousgouni V, Arias-Garcia M, Bakal C (2017). RNAi screens for Rho GTPase regulators of cell shape and YAP/TAZ localisation in triple negative breast cancer. Sci. Data 4, 170018.

Pasquetto I, Randles B, Borgman C (2017). On the reuse of scientific data. Data Sci J 16, 8.

Peng H (2008). Bioimage informatics: a new area of engineering biology. Bioinformatics 24, 1827–1836.

Piwowar HA, Vision TJ (2013). Data reuse and the open data citation advantage. Peer J 1, e175.

Prins P, De Ligt J, Tarasov A, Jansen RC, Cuppen E, Bourne PE (2015). Toward effective software solutions for big biology. Nat Biotechnol 33, 686–687.

Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell P, Carninci P, Clatworthy M (2017). The human cell atlas. Elife 6, e27041.

Reynaud EG, Peychl J, Huisken J, Tomancak P (2015). Guide to light-sheet microscopy for adventurous biologists. Nat Methods 12, 30–34.

Rigano A, De Castillia CS (2017). Proposal for minimum information guidelines to report and reproduce results of particle tracking and motion analysis. bioRxiv, 155036.

Rigano A, Galli V, Clark JM, Pereira LE, Grossi L, Luban J, Giulietti R, Leidi T, Hunter E, Valle M, *et al.* (2018). OMEGA: a software tool for the management, analysis, and dissemination of intracellular trafficking data that incorporates motion type classification and quality control. bioRxiv, 251850.

Roberts B, Haupt A, Tucker A, Grancharova T, Arakaki J, Fuqua MA, Nelson A, Hookway C, Ludmann SA, Mueller IA, *et al.* (2017). Systematic gene tagging using CRISPR/Cas9 in human stem cells to illuminate cell organization. Mol Biol Cell 28, 2854–2874.

Rohn JL, Sims D, Liu T, Fedorova M, Schöck F, Dopie J, Vartiainen MK, Kiger AA, Perrimon N, Baum B (2011). Comparative RNAi screening identifies a conserved core metazoan actinome by phenotype. J Cell Biol 194, 789–805.

Sage D, Kirshner H, Pengo T, Stuurman N, Min J, Manley S, Unser M (2015). Quantitative evaluation of software packages for single-molecule localization microscopy. Nat Methods 12, 717–724.

Sarntivijai S, Lin Y, Xiang Z, Meehan TF, Diehl AD, Vempati UD, Schürer SC, Pang C, Malone J, Parkinson H, *et al.* (2014). CLO: the cell line ontology. J Biomed Semant 5, 37.

Schneider CA, Rasband WS, Eliceiri KW (2012). NIH Image to ImageJ: 25 years of image analysis. Nat Methods 9, 671–675.

Schoenauer Sebag A, Plancade S, Raulet-Tomkiewicz C, Barouki R, Vert J-P, Walter T (2015). A generic methodological framework for studying single cell motility in high-throughput time-lapse data. Bioinformatics 31, i320–i328.

Serrano-Solano B, Ramos AD, Hériché J-K, Ranea JA (2017). How can functional annotations be derived from profiles of phenotypic annotations? BMC Bioinform 18, 96.

Serra-Picamal X, Conte V, Vincent R, Anon E, Tambe DT, Bazellieres E, Butler JP, Fredberg JJ, Trepat X (2012). Mechanical waves during tissue expansion. Nat Phys 8, U628–U666.

Simpson KJ, Selfors LM, Bui J, Reynolds A, Leake D, Khvorova A, Brugge JS (2008). Identification of genes that regulate epithelial cell migration using an siRNA screening approach. Nat Cell Biol 10, 1027–1038.

Sluka JP, Shirinifard A, Swat M, Cosmanescu A, Heiland RW, Glazier JA (2014). The cell behavior ontology: describing the intrinsic biological behaviors of real and model cells seen as active agents. Bioinformatics 30, 2367–2374.

Sneddon TP, Li P, Edmunds SC (2012). GigaDB: announcing the GigaScience database. GigaScience 1, 11.

Sommer C, Straehle C, Koethe U, Hamprecht FA (2011). Ilastik: Interactive learning and segmentation toolkit. In: From Nano to Macro: 2011 IEEE International Symposium on Biomedical Imaging, 230–233.

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE (2015). Big data: astronomical or genomical? PLoS Biol 13, e1002195.

Tambe DT, Hardin CC, Angelini TE, Rajendran K, Park CY, Serra-Picamal X, Zhou EHH, Zaman MH, Butler JP, Weitz, *et al.* (2011). Collective cell guidance by cooperative intercellular forces. Nat Mater 10, 469–475.

Thurley K, Gerecht D, Friedmann E, Höfer T (2015). Three-dimensional gradients of cytokine signaling between T cells. PLoS Comput Biol 11, e1004206.

Thurley K, Wu LF, Altschuler SJ (2018). Modeling cell-to-cell communication networks using response-time distributions. Cell Syst 6, 355–367.

Tohsato Y, Ho KH, Kyoda K, Onami S (2016). SSBD: a database of quantitative data of spatiotemporal dynamics of biological phenomena. Bioinformatics 32, 3471–3479.

Uhlmann V, Singh S, Carpenter AE (2016). CP-CHARM: segmentation-free image classification made accessible. BMC Bioinform 17, 51.

Ulman V, Maška M, Magnusson KE, Ronneberger O, Haubold C, Harder N, Matula P, Matula P, Svoboda D, Radojevic M, *et al.* (2017). An objective comparison of cell-tracking algorithms. Nat Methods 14, 1141.

Urban E, Jacob S, Nemethova M, Resch GP, Small JV (2010). Electron tomography reveals unbranched networks of actin filaments in lamellipodia. Nat Cell Biol 12, 429.

Visser U, Abeyruwan S, Vempati U, Smith RP, Lemmon V, Schürer SC (2011). BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. BMC Bioinform 12, 257.

Wallis JC, Rolando E, Borgman CL (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PloS One 8, e67332.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3.

Williams E, Moore J, Li SW, Rustici G, Tarkowska A, Chessel A, Leo S, Antal B, Ferguson RK, Sarkans U, *et al.* (2017). Image Data Resource: a bioimage data integration and publication platform. Nat Methods 14, 775.

Xu S, McCusker J, Krauthammer M (2008). Yale Image Finder (YIF): a new search engine for retrieving biomedical images. Bioinformatics 24, 1968–1970.

Yang C, Svitkina T (2011). Visualizing branched actin filaments in lamellipodia by electron tomography. Nat Cell Biol 13, 1012.

Zaritsky A (2016). Cell biologists should specialize, not hybridize. Nature 535, 325.

Zaritsky A, Natan S, Kaplan D, Ben-Jacob E, Tsarfaty I (2015a). Live time-lapse dataset of in vitro wound healing experiments. GigaScience 4, 1–5.

Zaritsky A, Obolski U, Gan Z, Reis CR, Kadlecova Z, Du Y, Schmid SL, Danuser G (2017). Decoupling global biases and local interactions between cell biological variables. eLife 6, e22323.

Zaritsky A, Welf ES, Tseng Y-Y, Rabadán MA, Serra-Picamal X, Trepat X, Danuser G (2015b). Seeds of locally aligned motion and stress coordinate a collective cell migration. Biophys J 109, 2492–2500.