Viziometrics: Analyzing Visual Information in the Scientific Literature

Po-shen Lee, Jevin D. West, and Bill Howe

Abstract—Scientific results are communicated visually in the literature through diagrams, visualizations, and photographs. These information-dense objects have been largely ignored in bibliometrics and scientometrics studies when compared to citations and text. In this paper, we use techniques from computer vision and machine learning to classify more than 8 million figures from PubMed into 5 figure types and study the resulting patterns of visual information as they relate to scholarly impact. We find that the distribution of figures and figure types in the literature has remained relatively constant over time, but can vary widely across field and topic. Remarkably, we find a significant correlation between scientific impact and the use of visual information, where higher impact papers tend to include more diagrams, and to a lesser extent more plots. To explore these results and other ways of extracting this visual information, we have built a visual browser to illustrate the concept and explore design alternatives for supporting viziometric analysis and organizing visual information. We use these results to articulate a new research agenda – viziometrics – to study the organization and presentation of visual information in the scientific literature.

Index Terms—Viziometrics, Scholarly Communication, Meta Research, Figure Retrieval, Information Retrieval, Bibliometrics, Scientometrics,

1 Introduction

Information in the scientific literature is conveyed visually using plots, photographs, illustrations, diagrams, and tables. This information is designed for human consumption but, unlike the surrounding text, is not directly machine-readable. As a result, relatively few studies explore how these visual encodings are used to convey scientific information in different fields and how patterns of encodings relate to impact.

The visual cortex is the highest-bandwidth information channel into the human brain [1] and humans are known to better retain information presented visually [2]. The figures in the scientific literature therefore would appear to play a critical role in scientific communication. The discovery of the structure of DNA was largely a visual argument based on the images produced by X-ray crystallography; indeed, Gibbons argues that the act of producing the visualization of the structure represents the discovery itself [3]. The first extra-solar optical images of planets amplified the nascent subfield of astronomy focused on planet-hunting [4]. Medical imagery of biological processes at scales below that which can be detected using conventional optical methods are providing new insight into brain function [5]. In all fields, key experimental results are summarized in plots, complex scientific concepts are illustrated schematically in diagrams, and photographic evidence are used to provide insight at scales and in locations not available to the human eye. The quantification of science and the rise of big data has increased the need for visual representations of the data, models, and results.

In the 1950s, researchers like Eugene Garfield and De Solla Price recognized the importance of citations in organizing and searching the scientific literature [6], [7], but the process for making this information useful at scale was painstaking. We see an analogy with the current role of the visual literature. There is clear value in extracting and analyzing figures to understand its role in scientific communication and impact, just as there is clear value in analyzing the citation network in isolation. The citation network tells us how ideas are related; visual representations tell us how ideas are communicated. Figures from related groups, authors, and fields share a 'DNA' that can reveal how information is conveyed.

We adopt the term *viziometrics* to describe this line of research to convey the shared goals with bibliometrics and scientometrics. As with bibliometrics, viziometrics uses citations to measure impact, but focuses on relating impact to the patterns of figure use. We analyze theses patterns within the papers (specifically, the distribution of various figure types) in order to understand how they may be used to more effectively communicate ideas. We have two overarching goals, towards which this paper represents an initial step: First, we seek to build new tools and services based on the visual information in the literature to help researchers find results more efficiently. For example, when searching for uses of a particular method (e.g., phylogenetic analysis), the figures themselves are more relevant than the papers that contain them. Second, can the patterns of figure use inform new best practices for scientific communication, especially outside of the authors' own discipline?

In this paper, we present an initial exploration of viziometrics by analyzing a corpus of papers from PubMed Central to relate the use and distribution of visual information

- *P-S Lee with the Department of Electrical Engineering, University of Washington, Seattle, WA, 98105. E-mail: sephon@uw.edu*
- *J. West and B. Howe are with the Information School, University of Washington, Seattle, WA, 98195. E-mail: jevinw@uw.edu and billhowe@cs.washington.edu*

with impact, and consider how these patterns change over time and across fields in order to provide a foundation for the two questions above. Specifically, we consider three sub-questions:

- How do patterns of encoding visual information in the literature vary across disciplines?
- How have patterns of encoding visual information in the literature evolved over time?
- Is there any link between patterns of encoding visual information and scientific impact?

To answer these questions, we developed a framework and system for managing a viziometric analysis pipeline and supporting tools based on the results. We refer to the overall platform as VizioMetrics.org.[1] VizioMetrics.org includes components for ingesting a corpus of papers, a database for managing the extracted metadata, analysis routines for dismantling multi-chart images, a classifier for identifying figure types, and a public figure-oriented search and browse interface that illustrates a different approach to organizing the scientific literature in terms of visual results and concepts rather than the papers that contain them.

A key result is a link between the use of scientific diagrams (schematics, illustrations) and the impact of the paper, suggesting that high-impact ideas tend to be conveyed visually. We conjecture two possible explanations for this link: that visual information improves clarity of the paper, leading to more citations and higher impact, or that high-impact papers naturally tend to include new, complex ideas that require visual explanation. More broadly, we argue that identification and description of the visual patterns, verified through computational experiments spanning a large corpus of papers, can help improve understanding of how scientific information is best conveyed, how the organization of visual information relates to scientific impact, how best to present scientific information more accessibly to a broader audience, and perhaps most directly, how to build better services for organizing, browsing, and searching the "visual literature."

2 Related Work

Computer vision techniques have been used in the context of conventional information retrieval tasks (retrieving papers based on keyword search), including some commercial systems such as D8taplex [8] and Zanran [9]. Search results from these proprietary systems have not been evaluated and do not appear to make significant use of the semantics of the images.

In 2001, Murphy et al. proposed a Structured Literature Image Finder (SLIF) system, targeting microscope images [10]. A decade later, Ahmed et al. [11], [12] improved the model for mining captioned figures. The latest version combines text-mining and image processing to extract structured information from biomedical literature. The algorithm first extracts images and their captions from papers, then classifies the images into six classes. Classification information and other metadata can be accessed via web service. However, SLIF focuses exclusively on microscropy images and does not extend to general figures.

Choudhury et al. [13] proposed a modular architecture to mine and analyze data-driven visualizations that included (1) an extractor to separate figures, captions, and mentions from PDF documents [14], (2) a search engine [15], (3) raw-data extractor for line charts [16], [17], [18], [19], and (4) a natural language processing module to understand the semantics of the figure. Also, they presented an integrated system from data extraction to search engine for user experience. Chen et al. [20] proposed their search engine named DiagramFlyer for data-driven figures. It recovers the semantics of text components in the statistical graph. Users can search figures by giving attributes of axes or the scale range in further. Additionally, DiagramFlyer can expand queries to include related figures in terms of their production pipelines. Other studies have proposed informatics methods for retrieving maps of the brain through large-scale image and text mining on fMRI images [21].

Although these early projects represent a different approach for information retrieval tasks, they make no attempt to analyze the patterns of visual information in the literature longitudinally. Hegarty et al. collected 1,133 articles from 9 psychology journals and found that articles with fewer graphs and more structural equation models were more frequently cited [22]. This result was not supported by other in different disciplines: Fawcett et al. studied the citations of 28,068 papers published in the top three journals specializing in ecology and evolution and found that heavy use of equations has a significant negative impact on citation rates [23]. Tartanus et al. reported a positive correlation between number of graphs and the impact factors in journal level by analyzing all papers published in 2010 from 21 selected journals in agriculture [24]. Other studies investigate how the use of figures differs by authorship patterns. Cabanac et al. analyzed 5,180 articles in the sciences and social sciences and found that *groups* of authors used significantly more tables and graphs than single authors [25]. Hartley et al. investigated approximately 2,000 articles from 200 journals in the sciences and social sciences. They found that men used 26% more figures than women, but found no significant difference in their use of tables. In addition, they didn't find significant differences between men and women in using either graphs and figures or tables in social science articles [26]. Since counting figures manually is extremely time-consuming, all of these studies were limited to specific domains on a relatively small number of papers and journals. Our approach is to automate the analysis using computer vision techniques and machine learning, scale it to a large corpus of papers to allow broader inferences, and release the software and labeled data for other researchers to use.

In this paper, we present this image processing pipeline that classifies scientific figures into different categories (Section: 3). We build a search interface that uses these classified images as the primary unit for exploring scholarly content (Section: 5). We make the dataset publicly available in order to support additional analyses of the figures and improve figure-oriented search. We provide preliminary evidence that links paper impact to figure type density. (Section: 4).

1. We distinguish the platform VizioMetrics.org from the field of study (Viziometrics)

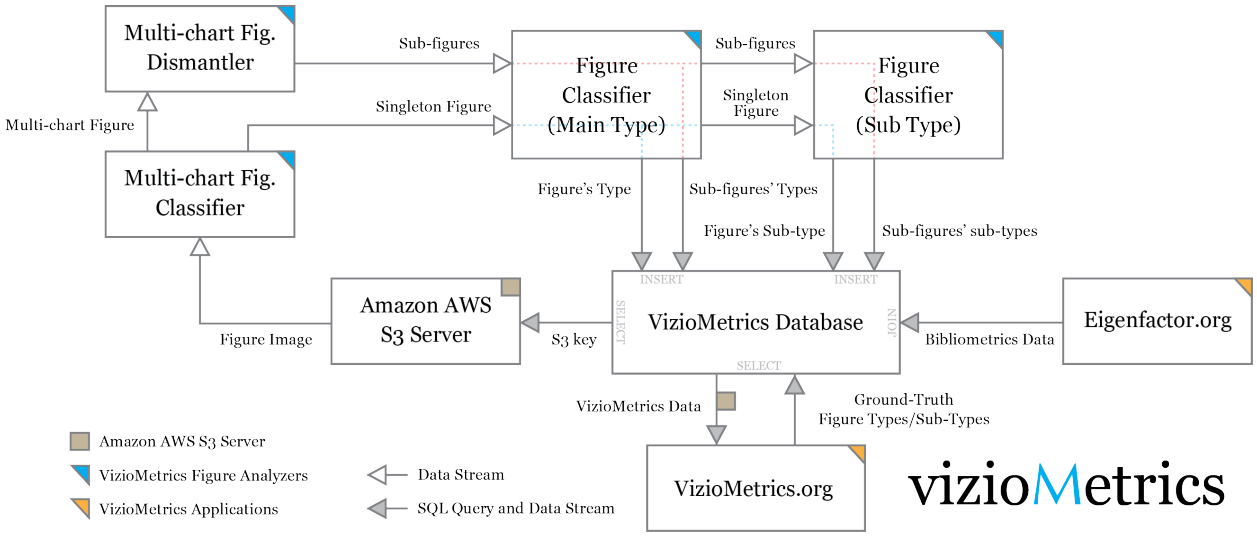


Fig. 1. VizioMetrics.org system overview. We store the images in Amazon's S3 service. Image paths, figure captions, paper metadata and classification result are stored in the database. The figure analysis system acquires the file keys from the database, downloads the image files, and feeds them into the figure processing pipeline. The final classification results are stored in the database as the sources for the application prototype.

3 Dataset and Methodology

We developed a platform called VizioMetrics.org with which we analyzed 4.8 million figures from more than 650,000 PubMed Central (PMC) papers (7.4 figures/paper). PubMed Central, an archive of biomedical and life science literature, provides free access to the full text documents including the source images. We downloaded the article files from the PMC FTP server and extracted the images into a figure corpus. Of these files, about 66% had associated figure files. These figure files are separated from the PDF files, allowing us to avoid having to extract them from literature. In addition, PMC also provides paper metadata including paper titles, authors, publishing date, citations, and image captions that we use in our figure search engine and analysis by field.

We found five image formats in use: GIF, JPEG, TIF, TIFF, PNG. The vast majority (99%) of the images were in JPEG format with a small number of PNG files. We had several filtering steps to remove duplicate images and the images that are not scientific figures (e.g. copies of full papers). First, we removed all GIF files since they are duplicates of images in other formats. Second, we removed image files that turned out to be image representations of full papers. Third, we converted all TIF and TIFF files to JPEG files and resized their dimensions such that the longer edge was 1280 pixels. If the longest edge of the original image was larger than this value, we did not modify the aspect ratios.

After filtering, we classified 4.8 million images into five categories. The classification algorithm is described in Section 3.1. The classifier returns a probability distribution across all class types, but for each image we only assigned the label with the highest probability. The class labels are as follows:

- Equation (e.g., embedded equations, Greek and Latin characters)
- Diagram (e.g., schematics, conceptual diagrams, flow charts, architecture diagrams, illustrations)
- Photo (e.g., microscopy images, diagnostic images, radiology images, fluorescence imaging)
- Table (any tabular structures with text or numeric data in the cells)
- Plot (e.g., bar charts, scatter plots, line charts)

TABLE 1
We classified 4,781,741 figures into six categories. The table shows the number of figures for each figure type before and after dismantling.

Figure Type	Count Before Dismantling	Count After Dismantling
Multi-chart	1,416,237 (29.6%)	None
Equation	1,425,042 (29.8%)	1,741,059 (17.0%)
Diagram	652,918 (13.7%)	2,036,704 (19.9%)
Photo	475,615 (9.9%)	2,322,231 (22.7%)
Plot	475,327 (9.9%)	3,579,839 (35.0%)
Table	336,602 (7.1%)	553,171 (5.4%)
Total	4,781,741	10,233,004

Of the 4.8 million figures, 1.4 million contained multiple sub-figures within a single image, often with each sub-figure labeled with A, B, etc. We refer to these figures as *multi-chart* figures. We "dismantled" these multi-chart figures into their individual parts using a customized algorithm that we developed for this purpose [27]. After dismantling, we extracted and classified another 5 million individual figures. In total, we classified more than 10 million figures.

The results of our classification are summarized in Table 1. This summary information alone provides some interesting insights: About 67% of the total figures are embedded in multi-chart figures, demonstrating the importance of dismantling figures for this analysis. Plots are the most likely figure type to be embedded in this way: we found

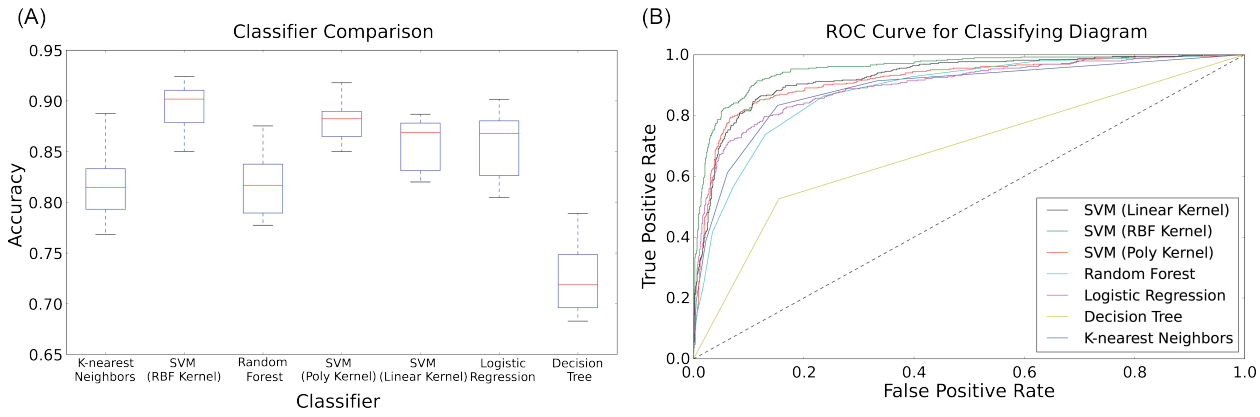


Fig. 2. Comparison of classifiers: K-nearest neighbors, random forest, logistic regression, decision tree, and SVM with RBF, linear, and polynomial kernels, respectively. (A) The SVM with RBF kernel achieves the best performance evaluated by 10-fold cross validation. (B) The SVM with the RBF kernel also achieves the best performance compared to the linear kernel and polynomial kernel shown with the ROC curves.

475k standalone plots but 3.5M total plots after dismantling. Tables are significantly less common than other figure types, suggesting a preference among authors (or possibly editors) for presenting results visually. There is a relatively uniform distribution across diagrams, photos, and plots; the prevalence of photos is likely an artifact of the biomedical emphasis of the PMC corpus.

3.1 Figure Analysis

Figure 1 illustrates the analysis pipeline used to perform classification. We first download and extract the images in AWS (Amazon Web Services). We then classify each figure as either *multi-chart* or *singleton*. Each figure identified as multi-chart is dismantled into a set of singleton figures. All singleton figures (including those dismantled from multi-chart figures) are labeled with one of five class labels: equation, diagram, photo, plot and table. The classified images can be browsed online at viziometrics.org. In the following sections, we will briefly describe the algorithm for each box in Figure 1.

3.1.1 Figure Classification

To classify the images, we adapt the technique developed by Coates et al. [28] and extended by Savva et al. [29] to extract small patches from the corpus of images, cluster these patches into groups, then re-encode each image as a histogram over these groups. This histogram can be used as a fingerprint to classify images.

First, we normalize an image to a 128×128 grayscale image with a constant aspect ratio. Then, we randomly extract a set of 6×6 patches from each training image and normalize the contrast of each patch. Adjacent pixel values, and therefore adjacent patches, can be highly correlated. To increase contrast and better distinguish different patches, PCA whitening is applied on the entire patch set.

Next, we cluster the set of patches using k-means (k = 200) and to identify 200 common patch types, one for each cluster. A representative patch for each patch type, called a *codebook patch*, is derived from each cluster. For each training image, we generate a new set of patches by sliding a 6×6 window in one-pixel increments across the image. For each such generated window patch, we find the most similar codebook patch via Euclidean distance and increment a counter for that codebook. The set of codebook counters forms a histogram, and this histogram forms the feature vector used to train the classifier.

TABLE 2
Evaluation of multi-chart figure classifier and figure-type classifier using 10-fold cross validation.

Figure Type	Precision	Recall
Multi-chart	92.9%	86.3%
Singleton	89.3%	94.6%
Equation	95.4%	95.1%
Diagram	84.2%	84.1%
Photo	94.5%	97.3%
Plot	91.5%	90.2%
Table	95.1%	93.1%

To account for the global structure of common visualizations (e.g., axes are typically found on the left and bottom of the image), each image is split into four quadrants and a separate 200-element histogram is computed for each quadrant. The final feature vector of 800 elements is obtained by concatenating the four 200-element histograms. These feature vectors are then classified using a Support Vector Machine (SVM).

We evaluated five different classifiers: K-nearest neighbors, random forest, logistic regression, decision tree, and SVM with RBF kernel. The corpus we used for training was randomly sampled from the PMC corpus (ftp://pub/pmc/ee/). We manually labeled 3,271 images as one of five categories: photos (782), tables (436), equations (394), visualizations (890), and diagrams (769) and used these hand-labeled data to train the classifiers. We compared the accuracy of the five classifiers obtained by 10-fold cross validation and selected SVM with an RBF kerneal based on its superior performance performance (Figure 2(A)). To fine tune the SVM parameters (kernel, gamma, and penalty

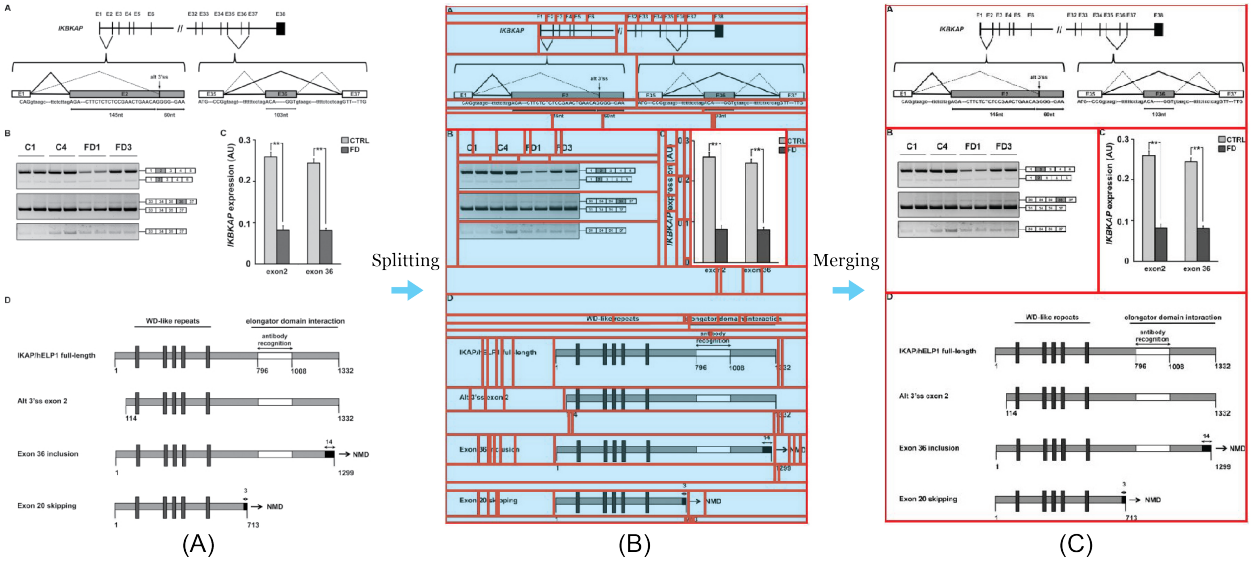


Fig. 3. Multi-chart figure dismantling. The figure shows the intermediate steps for dismantling multi-chart figures. The splitting algorithm recursively segments the raw images into several sub-images. The merging algorithm then aggregates auxiliary fragments with nearby standalone figures to produce the final segmentation.

parameters), we randomly reserved 25% of the images for a testing set and trained the classifiers on the remaining 75% for each category, then used a grid search method to complete the task (part of the sklearn library). Figure 2(B) shows the ROC curves for identifying diagrams. The optimized model is run by the RBF kernel with gamma of 0.001 and a penalty parameter of 1000. We focused on diagrams because the RBF kernel performs particularly well in recognizing diagrams. Once the model parameters are tuned, we evaluated the model by using the testing set and then trained the final model with all images. In this paper, we report the evaluation of classification performance (Table 2) by 10-fold cross-validation on the full training corpus of 3,271 images. The final classification accuracy for all images is 91.5%

3.1.2 *Figure Dismantling*

The use of multi-chart figures complicates classification. For example, the figure in Figure 3 consists of four sub-figures: a photo, a plot, and two diagrams. Approximately 30% of all figures in our corpus required dismantling. To achieve more accurate medical image retrieval, early studies segment the diagnostic photos from photo arrays by recognizing the sharp gradient at the boundaries of photos [10], [11], [12]. This simple approach satisfies the need for separating photos, but it is ineffective for separating plots or diagrams because the boundaries take many different forms. Cheng et.al extends this approach with Fuzzy Logic Controller clustering to handle plots and diagrams [30]. Unfortunately, their algorithm is not available and no evaluation was provided.

The primary goal of our *dismantling* algorithm is to extract the component figures. Figure 3 illustrates the process of splitting a multi-chart image into four sub-figures. The algorithm first splits an image into blocks based on background color and layout patterns (Figure 3(B)). The splitting algorithm separates the input image into vertical and horizontal orientations, recursively, until it does not find any available fragments. An SVM-based binary classifier then distinguishes fragments that represent complete charts from fragments that contain only incomplete auxiliary information (e.g., axis labels, tick marks, and legends highlighted in blue color in 3(B)). Next, we recursively merge the fragments to their adjacent blocks in the same splitting level. After completing all possible merges among siblings, we transfer the newly merged blocks to their parents level as new blocks and then re-classify them. This process ends when it completes the possible merging in the top level. We found that the splitting and merging steps may produce different results from different initial splitting orientations (horizontal or vertical). Consequently we defined (and evaluated) a heuristic score function to choose between orientations based on the homogeneity of the shape of the sub-figures, based on the observation that authors tend to follow repeatable patterns in laying out sub-figures. The higher-scoring decomposition is chosen as the final output (3(C)). The algorithm achieved a recall of 82.9% and a precision of 84.3% with respect to manual dismantling on a set of 500 multi-chart figures randomly sampled from PMC. That is, the algorithm found 2743 sub-figures, 2499 of which were among the 3013 extracted manually as ground truth. The details of the algorithm are explained in an earlier paper [27]. Since our original paper was published, Taschwer et al [31], Wang et al [32], and Santosh et al. [33] have proposed methods for related problems.

3.1.3 *Multi-chart Figure Classification*

Attempting to dismantle every figure image in our corpus would be prohibitively expensive and extremely wasteful; only about 30% of the images are multi-chart figures. We therefore developed a simple and fast pre-classifier to distinguish multi-chart figures from singleton figures in order to reduce the number of dismantled singletons.

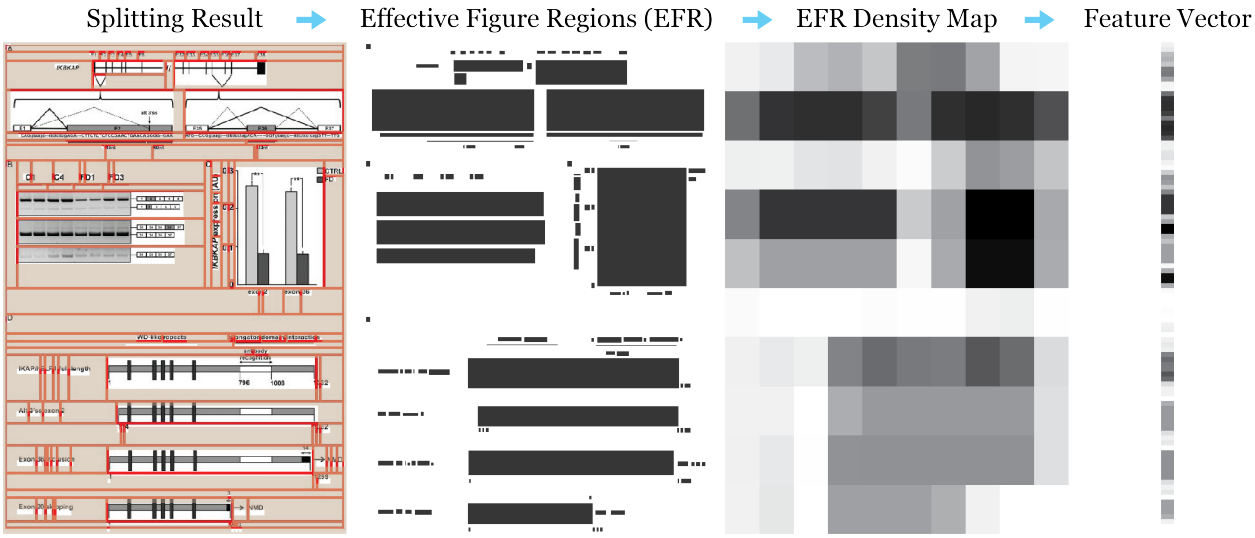


Fig. 4. Recognizing mulfti-chart images. After splitting the figure into distinct blocks, the dismantling algorithm marks the effective figure regions (EFR) then downsamples the EFR into $n \times n$ blocks that form a $n^2 \times 1$ feature vector. These vectors are used to train the classifier.

We designed the method based on two observations: that multi-chart figures tend to have a different size and shape than singleton figures, and that the layout of a multi-chart figure tends to follow a regular grid pattern. Based on these two observations, we constructed a feature vector with K (K = M + N) elements: M elements based on the size and shape, and N elements based on the grid layout. The M elements consist of the image height ratio($height_i$ / $height_{avg}$) and the image width ratio($width_i$ / $width_{avg}$) where the denominators are average image height and average image width of all images in the training set respectively. The N elements are derived from the output of splitting algorithm of the dismantler.

Figure 4 shows the splitting, and the red lines indicate the boundaries between fragments. For each block, we mark the minimal rectangular region that contains non-empty pixels, so that we can obtain the effective figure regions (EFR) and use them as a mask. We subdivide the mask into $n \times n$ blocks and compute the proportion of EFR in each block as defined as the EFR density map. Finally, we squeeze the values into a 1-D vector with n^2 elements.

We set $n = 10$ as the final parameter (M = 100) and apply the same technique described in the previous section to train the figure classifier. The final model is optimized by using a RBF kernel with gamma of 0.001 and a penalty parameter of 1000. As noted above, we obtained 91.8% accuracy by 10-fold cross-validation on the entire training set comprising 880 multi-chart figures and 1067 singleton figures. The recall and precision for each class are shown in Table 2.

3.2 Measuring Scholarly Influence

To assess the influence of a particular paper, we used the article-level Eigenfactor (ALEF) score. [34], [35], [36]. ALEF is a modified version of the PageRank algorithm [37]. The algorithm uses random walk on the article-level citation graph, where each vertex is a paper and each directed edge is a citation. Because a random walker will only move backwards in time using the standard PageRank approach, we modify the algorithm to reduce the number of steps the random walker takes and teleports the random walker to links rather than nodes [34], [38].

The ALEF ranking method has been shown to outperform simple citation counts and standard PageRank approaches [36]. The ALEF method took second place in a recent data challenge sponsored by the ACM International Conference on Web Search and Data Mining (WSDM) [2]. Although ALEF is an effective measure of article-level impact [36], the qualitative results of this study would not change if we simply used raw citation counts as our measure of impact.

4 Exploring Visual Patterns in the Literature

We use the classified figures to study patterns in the use of visual information across scientific domains, across publication venues, and over time. We also used the classifications to examine the effect on scholarly impact.

More broadly, we are interested in better understanding how complex results are communicated across disciplinary boundaries and to the general public, and how this communication channel can be optimized to increase the bandwidth of scientific discourse.

Our method of longitudinal analysis of all figures in a domain is generalizable both to other domains and to other questions related to demography, editorial trends, narrative style, and influence. In this paper, we provide preliminary results using this method and discuss the findings.

4.1 Dataset Details and Preprocessing

We use the set of images described in Section 3, but refine this dataset to avoid biases in four ways: First, our analysis of impact depends on having an ALEF score available, so we remove all papers with no ALEF score available (typically because the paper attracted zero citations, and a

2. http://www.wsdm-conference.org/2016/wsdm-cup.html

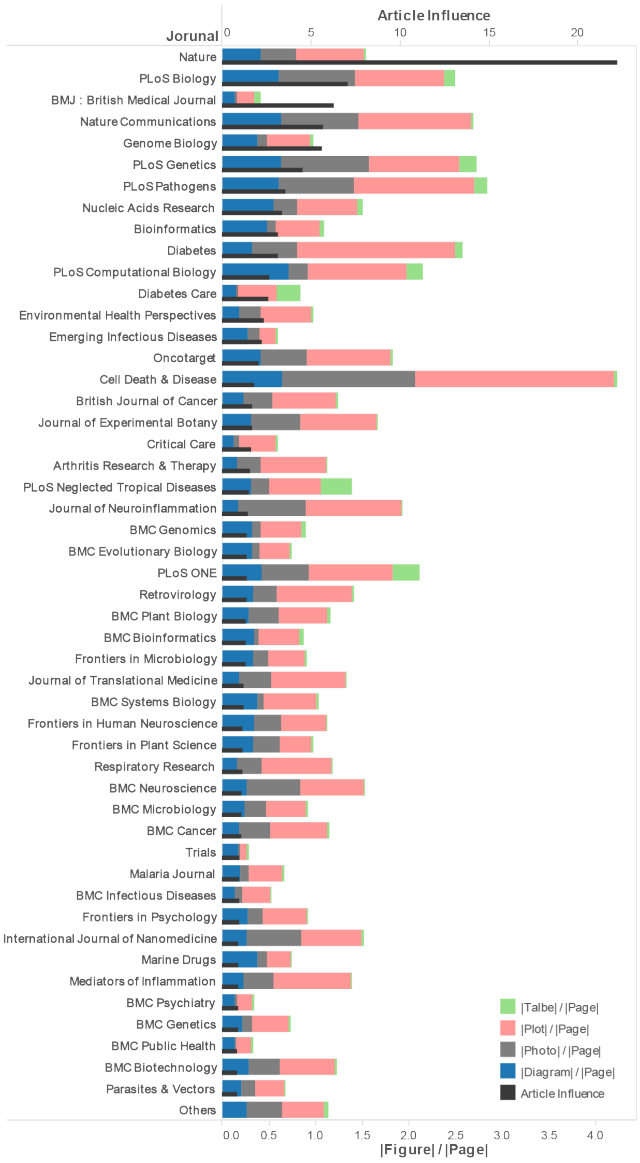


Fig. 5. The distribution of figure types across journals show an emphasis on plots and diagrams relative to tables, and identify visualization-heavy venues such as Cell Death and Disease. We considered the top 49 highest-impact journals in PMC that had at least 850 papers available in the corpus, where impact is measured as Article Influence (AI) (the black bar). Each stacked bar shows the average density of each figure type across all papers published in the journal. The density of a figure type is the number of instances of that type divided by the page count. The category "Others" contains 288,953 papers from other journals.

few negligible cases where processing errors prevented the calculation from completing).

Second, for some papers (less than ten percent of the corpus), the total number of pages could not be determined, preventing us from calculating figure densities. PMC does not report page counts in the XML so we had to determine the page counts using the PDF files provided. However, some papers had no PDF file included so we could not determine the page count.

Third, we remove papers published before 1997 since the number of papers per year from that time is less than 300 and is strongly biased toward a small number of journals that were indexed by PubMed during that period.

Forth, we exclude 86,205 papers with zero figures since we cannot properly distinguish between two situations: (1) papers that were published containing no figures and (2) papers that were published with figures, but for which the figures were not provided to PMC. Generally, more recent papers are more likely to fall into case (1), since the procedure to upload figures separately was more commonly used in the past. Papers corresponding to case (2) (i.e., older papers) can skew the results since older papers tend to have more citations and therefore higher ALEF scores.

The papers that failed to meet one or more of these criteria appeared to be distributed uniformly across the overall dataset, so any bias created by their removal appears negligible.

After these preprocessing steps, the dataset includes 494,663 papers and 6,897,810 figures (after dismantling), excluding equations. We exclude equations because not all equations were represented as figures and sometimes multiple equations appear in a single figure, making it difficult to estimate the total number of equations.

Some of the PMC literature is in pre-print formats rather than the official journal format. For these papers, we use the total number of pages from PMC. As a result, the page count may be different than the actual paper. In addition, we underestimate the total number of tables from those authors who use only latex or Microsoft Word to construct their tables, since these authors typically do not provide tables as separate images.

The dataset does not necessarily represent all relevant papers. Authors of the papers analyzed here can voluntarily select to submit papers to PMC, and PMC will clearly tend to attract papers in the life sciences with an emphasis on human biology. In particular, *Nature* publishes a significant number of Physics papers, but these papers will be underrepresented in PMC.

4.2 Understanding Visual Patterns Across Disciplines

To analyze the patterns of visual encodings across disciplines, we normalize the individual figure counts by the total number of pages in order to measure the *density* of each figure type. This figure count normalization is similar to the method used by Fawcett et al. [23] in their analysis of equations. It ensures the values are comparable between articles with diverse lengths.

Next we aggregate the figures and papers by journal and research topics to see how figure types vary across publishing venues and disciplines. Figure 5 and Figure 6 show the average figure density of journals and research topics for which we were able to collect at least 850 or 1000 papers published during 1997 to 2014 from PMC, respectively. Figure topics were assigned used Thomson-Reuters' Journal Citation Report (JCR) category system.

In Figure 5, we restricted the analysis to those journals with at least 850 articles in the corpus. The stacked bars present the densities of diagrams, photos, visualizations, and tables, from left to right. Equations are not considered in this case because defining the quantity of equations can be vague: a single image may contain any number of equations, and our dismantler algorithm was not designed to parse equations. The thin dark bars represent the impact of each

journal as measured by ArticleInfluence (AI) for the journal [39]. AI is a journal-level metrics whereas ALEF is an article-level metric.

In Figure 6, we used the average ALEF score to estimate the value of topic areas because topic areas consist of overlapping journals. The AI scores is a citation metric for measuring journal influence [39]. The underlying citation data comes from Thomson-Reuters' JCR. Journals and research topics are listed by impact in descending order. Due to the limit of page capacity, we show only the top 49 items and gather the papers from small-collection journals and lower-rank journals into "Others."

Figure 5 shows the top 49 journals ordered by AI. Differences exist between journals. The journal *Cell Death and Disease* relies heavily on microscopy and experimental evidence, and we see this emphasis manifest as a significantly higher number photos and plots. We can see that multidisciplinary journals, such as the *Nature* series and the *PLoS* series exhibit a balance of figure types. Qualitatively, many of the journals with high figure-per-page counts are also high in AI. Further, papers from the top one-third journals (16 out of 50) tend to have more diagrams. Journals emphasizing prose-oriented case studies are exceptions and have fewer figures: *British Medical Journal*, *Diabetes Care*, and *Emerging Infectious Diseases*. In comparison, papers from the journals near the tail show lower diagram density. We will make this observation statistically precise in Section 4.4.

Using Thomson Reuters' JCR, we can assign each journal to a research topic, then repeat the analysis of figure distribution by research topic rather than journal. We describe the method used to assign topic labels in more detail in Section: 3.2.

Figure 6 shows the disciplines for which at least 1000 papers were available. Differences between disciplines in figure type density are apparent. For example, *cell biology* and *pathology* have a relatively high number of photos per page, whereas *mathematical and computational biology* and *medicinal chemistry* have fewer photos per page and relatively more diagrams and plots per page. *Biology* and *internal medicine* tend to have relatively more tables per page, suggesting an emphasis on (or tolerance of) presenting quantitative results numerically. We conjecture that these patterns relate cultural norms for publication rather than specific research methods; that certain fields expect a certain "syntax" for a research paper and that the distribution of figures is a part of the syntax. A study of these conjectures is beyond the scope of this paper.

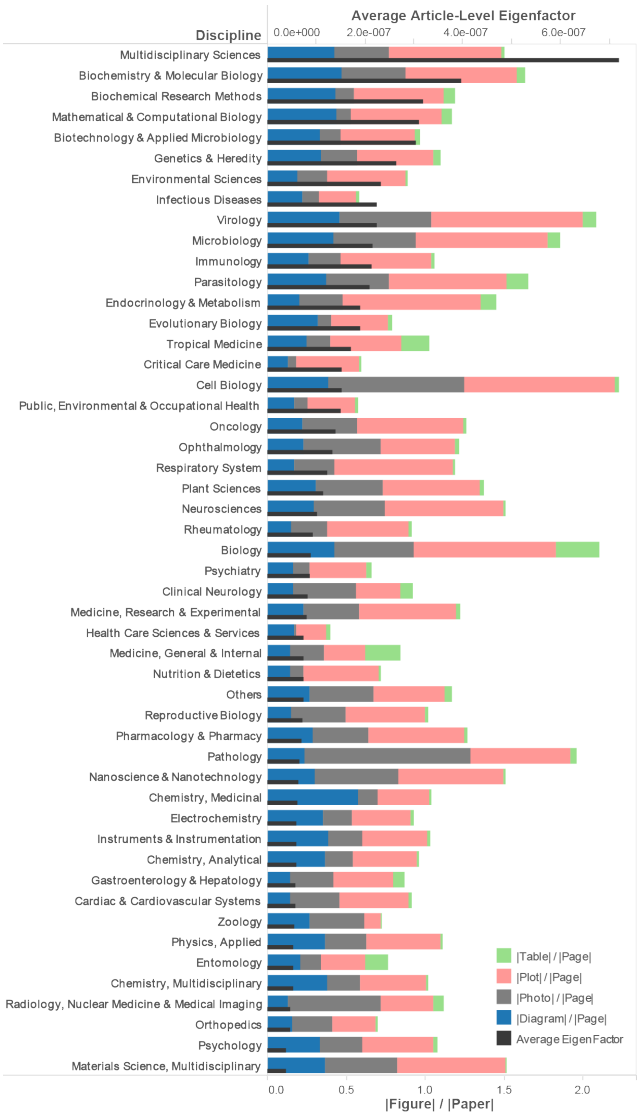


Fig. 6. Figure distribution by research topic show that microbiology topics tend to emphasize visual presentation of ideas. Topics were determined by the journal categories in Thomson Reuters' JCR. We show the highest-impact 49 topics that have at least 1000 papers, where impact is the average of all papers assigned to that category. The category "Others" includes 216,380 papers from other topics and papers without topic labels.

4.3 Visual Patterns Over Time

We analyze patterns of visual information over time by segmenting the data into different publishing years. The earliest paper we collected from PMC was published in 1937, but relatively few papers earlier than 1997 are included (biasing the corpus). We plot the total number of papers in our database from 1990 to 2014 in Figure 7. Paper quantity reaches the thousand mark in 1997 and the ten thousand mark in 2007. In 2008, NIH mandated that authors upload their papers to PMC, partially explaining the growth of the corpus. Papers can be uploaded at any time for any publication year, so we do not necessarily see an increase in later papers. The average ALEF score increases until 2000 and then decreases, consistent with most measures of impact that are inherently time-sensitive.

The "hump" that occurs in Figure 7 around 1997 to 2002 is attributable to a bias in the corpus: in this period, the corpus was dominated by just three journals: *Journal of Cell Biology*(38%), *Journal of Experimental Medicine*(31%), and *Journal of General Physiology*(8%). As more journals were added to PMC, this sampling bias decreased, and the patterns stabilized. After 2006, the number of diagrams per page remains relatively consistent, and a small but consistent growth in the number of plots and tables per page is observed. We conjecture that these increases could be attributable to an increased emphasis on data-intensive science in the biological and biomedical disciplines, but another possibility is that such figures became easier to

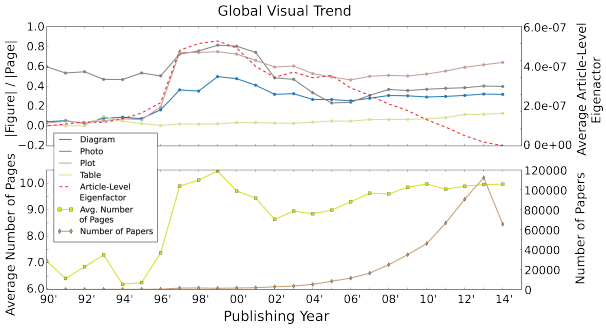


Fig. 7. The distribution of figure types in the PMC corpus over time. The top figure shows the number of papers increasing dramatically in the mid-2000s, which can be explained by a change in sponsor rules: NIH required authors to submit their papers to PMC. The "hump" of impact between 1997 and 2005 may be attributable to author bias in voluntarily uploading their highest-impact papers. After 2006, the increasing uses of plots and tables may be attributable to increased emphasis on data-intensive research. The density of photos and diagrams are consistently flat over time. The bottom plot provides context: the average page length per paper over time, and the number of papers in the corpus over time.

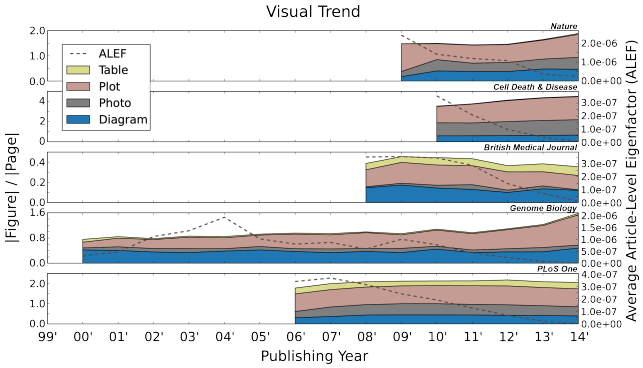


Fig. 8. We choose five specific journals for closer inspection: *Nature* (highest impact), *Cell Death and Disease* (highest figure density), *British Medical Journal* (lowest figure density), *Genome Biology* (unusually low proportion of photos) and *PLoS One* (largest number of papers). *Nature, Cell Death and Disease* and *Genome Biology* exhibit a recent increase in plots-per-page, consistent with the overall trend. We conjecture that the articles in these high-impact journals are becoming more data-centric. Moreover, *Nature* and especially *Cell Death and Disease* show a heavy use of figures, in part because these journals tend to have greater proportions of multi-chart figures (67% for *Nature* and 82% for *Cell Death and Disease* relative to 30% for the entire image set.) The *British Medical Journal* shows a different trend in which figure density gradually decreases; the mechanism behind this trend is unclear. *PLoS One* shows no significant change from its launch in 2006.

create thanks to improved tools resulting in increased use.

In Figure 8, we select five journals with unique features for closer inspection: *Nature* (highest impact according to our measures), *Cell Death and Disease* (highest figure density), *British Medical Journal* (lowest figure density), *Genome Biology* (unusually low proportion of photos) and *PLoS One* (largest number of papers). *Nature* exhibits an increase in figure density over time, driven primarily by an increase in plot density which may reflect an increased emphasis in data-intensive science. For the journal *Cell Death and Disease*, one sees the same effect of growing figure density over time, which corresponds to an increased use of multi-chart figures: 81% of the figures are multi-chart compared to an average of 38%.[3] In contrast, the *British Medical Journal* exhibits low figure density and a gradual decrease in the use of figures over time. Tables are used more in proportion compared to most journals and photos are extremely rare. We conjecture that the decrease in visual information over time may be related to a known shift in focus for BMJ, in which the editor has intentionally focused on topics of broad public interest [40]. It is possible that heavy use of quantitative data in the form of plots may make articles *less* accessible. *Genomics Biology* was selected for its unusually low proportion of photos, which appears consistent over time. We do see the density of plots increasing significantly since 2011, following the global trend. We selected *PLoS One* because of the extremely large number of papers in the corpus. Because it is broadly multidisciplinary, the patterns of figures represent many fields of study and we do not expect, nor do we see, any distinctive pattern. *PLoS One* may represent a microcosm of the overall literature in this regard.

4.4 Visual Patterns Related to Impact

In this section, we consider the relationship between patterns of visual encodings and scientific impact.

Figure 9 shows qualitatively that higher impact papers tend to have both a higher density and higher proportion of both plots and diagrams, but a lower proportion of photos. The visual encoding of quantitative information therefore appears to correlate with impact. We chose four bins that characterize the Eigenfactor score distribution, which tends to follow a power law distribution. We chose the four bins to roughly correspond to boundaries at 95%, 75%, 50%. The bin boundaries are not these numbers exactly because many papers have indistinguishable Eigenfactor scores[4], and we did not want to artificially separate two papers with the same score into two different bins. Instead, we move the boundary to the next highest threshold. The bin boundaries then become 5%, 23%, and 45%, with the lowest bin (Bottom 55%) containing all papers with Eigenfactor score of zero. For each group, we average the figure densities for each of four figure types and produce a histogram as shown in Figure 9.

The results in Figure 9 do not change when adjusting bin sizes. We regroup the papers binning by every half-percentile (99.5%, 99.0%, etc.) and compute the correlation coefficient. Table 3 shows the binned correlation coefficients for the four figure types. The first and second numbers in each cell are the correlation coefficients when including and excluding papers from *PLoS One* respectively. We separate the influence of *PLoS One*, as Figure 5 shows that *PLoS One* exhibits a significantly higher table density then other journals. **The key result is that higher diagram density and higher proportion of diagrams are linked to higher impact, while higher proportions of photos are linked to lower impact. These results indicate that high-impact papers may tend to use more diagrams, but also that diagrams tend to have a stronger relationship with impact than plots.** One possible interpretation of these results is that clarity of exposition contributes to impact: illustrating an original idea

3. Equations are not taken into account.

4. Any two papers with Eigenfactor difference within 1E-12 are regarded as having the same score.

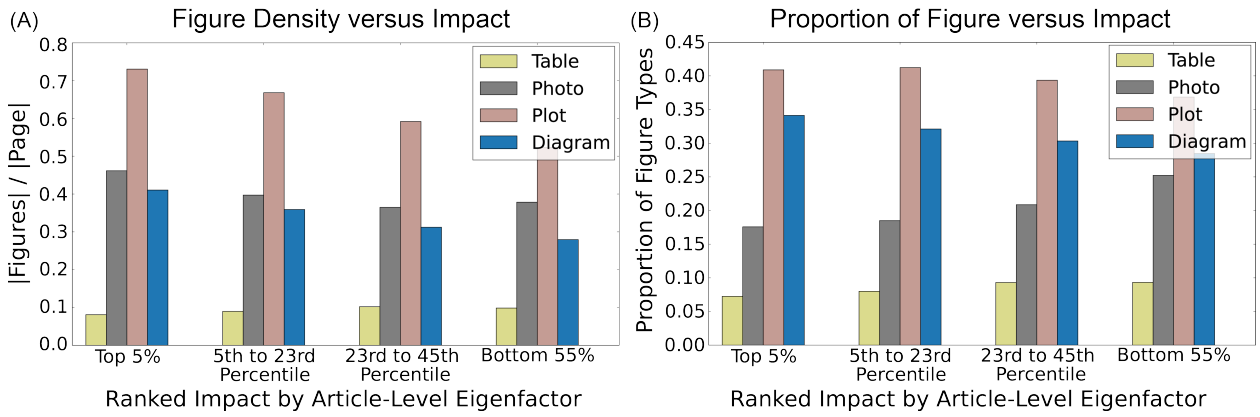


Fig. 9. Impact versus (A) figure density and (B) proportion of figures. We rank papers by ALEF and group them into 4 bins. Papers with the same Eigenfactor are grouped into the same set. Any two papers with Eigenfactor difference within 1E-12 are regarded as having the same impact, which is why the bins are not evenly distributed. For each set, we average the densities and proportions of 4 figure types.

TABLE 3
We estimate the correlation between the ALEF score and figure density (left column) and proportion of figures (right column). Each table entry $X(Y)$ indicates the correlation including (X) and excluding (Y) papers from *PLoS One*, a journal that tends to bias the results due to a high proportion of tables. Correlations excluding *PLoS One* are more strongly positive for all figure types. The entry NSS indicates that the result was not statistically significant. Overall, high proportions of diagrams are linked to high impact while high proportions of photographs are linked to lower impact (negative correlation).

Figure Type	Correlation Coefficient	
	Figure Density (w/o PLoS One)	Prop. of Figure (w/o PLoS One)
Diagram	0.84 (0.92)	0.61 (0.52)
Photo	0.57 (0.70)	-0.69 (-0.63)
Plot	0.60 (0.80)	NSS (NSS)
Table	NSS (0.78)	NSS (NSS)

visually leads to more impact then simply reporting experimental results. Previous studies have shown that diagrams were more effective than text in helping readers develop deeper understanding of the material [41]. We conjecture that the negative correlation with photographs may suggest that tight page limits associated with high-impact journals may lead authors to sacrifice photographs as extraneous, but it is possible that photographs lack the explanatory effects associated with carefully designed abstract visual encodings (diagrams, plots).

As described in Section 4.1, 86,205 papers were excluded because they reported no figure files in PMC. If we include these papers in the analysis (that is, intentionally misinterpreting the situation for those papers for which the figures were simply not uploaded to PMC), the average ALEF score decreased by just 5%, and the binned correlation coefficients vary be only about 10%. The relatively small effect is attributable to the fact that 60% of the 86,250 papers have an ALEF score of zero; that is, papers with zero figures tend to be very low-impact. Our qualitative result is therefore the same whether or not we attempt to consider these ambiguous cases: higher impact papers tend to have higher density of diagrams and plots.

5 A Browser for the Visual Literature

Consider a biologist in search of the phylogenetic tree associated with a virus. Using a conventional academic search engine, she enters keywords (perhaps the name of the virus and the word phylogenetic), retrieves a list of candidate papers, and, inspecting the title for relevance, opens each paper for manual review. This process operates at the wrong level of abstraction, as the search is focused on a particular method that is associated with a visual encoding — a phylogenetic tree has a distinctive visual representation. Consider another case where a researcher wants to compare a number of different designs for solid-state laser diodes. She would like to find both scanning electron microscope (SEM) images as well as diagrams illustrating the designs, with goals of performing non-trivial analysis *across* figures: comparing the SEM photos with the corresponding diagrams (perhaps from a different paper), or a comparison of leakage currents by inspecting a set of plots showing the current-voltage curves. With both examples, keyword search followed by manual inspection of papers to gather specific visual results seems unnecessarily inefficient. We aim to use our classification pipeline to power a more efficient approach to this task using a figure-centric search application [42].

The system indexes the authors, titles, abstracts and figure captions of the corpus of papers; keyword searches probe this index to find relevant images. Result figures are ordered by their ALEF scores, helping to reduce attention on low-impact papers. In the default layout, figures are arranged as a "brick wall" to make better use of screen real estate as in Figure 10(A). Users can retrieve additional figures by scrolling down to the bottom of the page. The color of figure border indicates its figure type as identified by our classifier. Users can restrict the results by figure type (using the results of our classifier) by using the checkboxes under the search box: composite, photo, table, plot, diagram, or equation. For instance, the biologist seeking phylogenetic trees can ignore any figures other than diagrams. We are

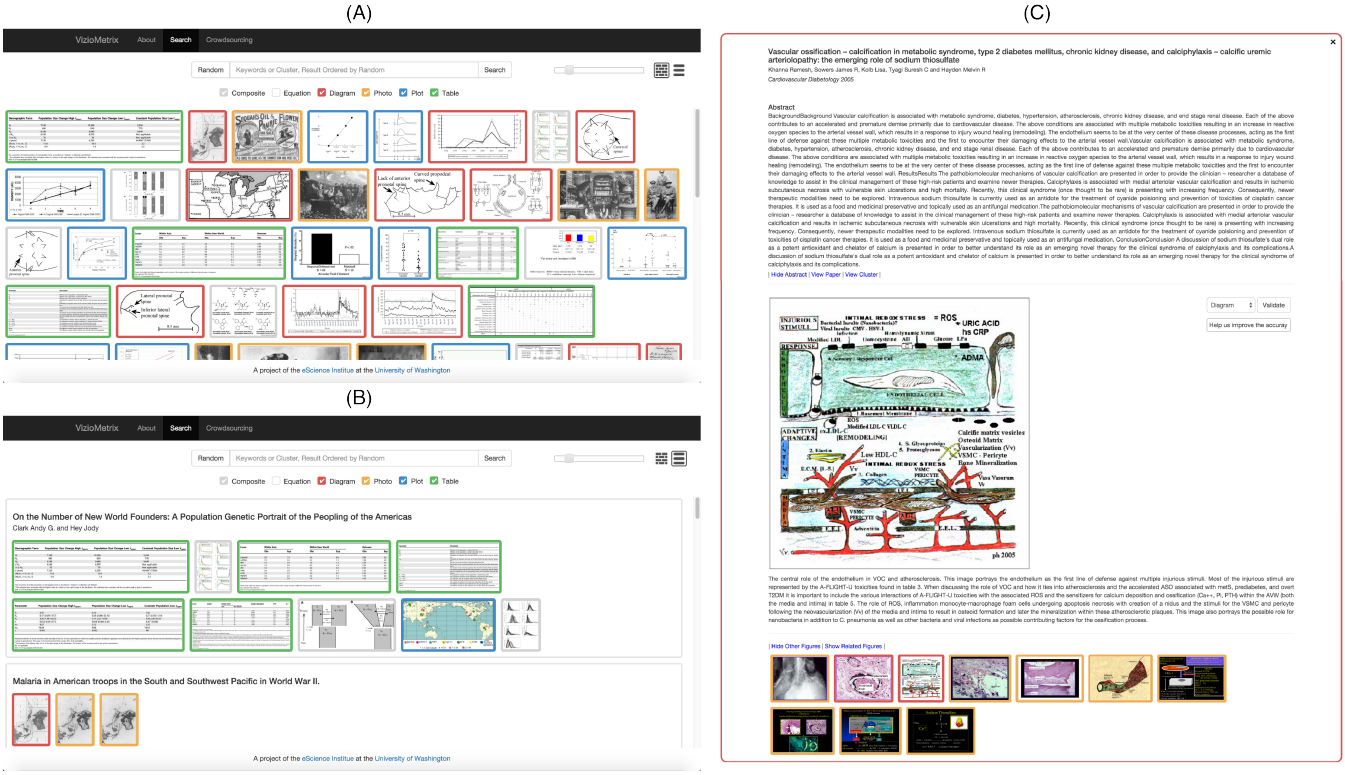


Fig. 10. The user interface of the VizioMetrics.org search engine. Result figures are either arranged via (A) the brick-wall layout or (B) a conventional layout bundling figures with literature title. Figures are labeled by different colors based on their types. (C) Clicking figures will show article details such as authors, abstract, figure captions, hyperlink to full PDFs and related figures. We also provide a verification form to encourage user verifying our machine-labelled figure type and help us gather more ground-truth label.

currently extending our categories to include more specific diagram types such as phylogenetic trees to enhance this feature.

Figures support a number of interactions. The slider allows zooming into figures to inspect fine detail. Clicking on the figure brings up a metadata page displaying the title, authors, abstract, caption, related figures and more (Figure 10(C)). Related figures from a target paper are selected using the citation network between papers. Papers that cite or receive citations from the target paper are likely candidates with similar figures. We determine the citation similarity using a recommender system that we recently developed [35]. This recommender system is based on a hierarchical clustering of an article-level citation network. In addition to the brick-wall layout, we also provide conventional layout (Figure 10(B)) that lists the figures in the context of the paper in which they appear. This mode is designed for users who are looking for particular papers, but who may recall a memorable figure from the paper if not the title or author. Viewing article titles together with figures may help them narrow the scope.

5.1 Evaluation of Figure Search

We evaluate the relevance of the figure search for a *figure-based method search task*. This task consists of using keyword search for a particular method, with the intent of finding figure that represent the result of using that method. Anecdotally, we find this task to be both common in practice and poorly supported by paper-oriented search engines.

To evaluate the ability of viziometrics to support this task, we measure the proportion of top-ranked results that match the search term, using expert labeling as ground truth. For example, a phylogenetic analysis typically produces a particular type of tree that is recognizable to researchers. We report the proportion of the top 30 returned figures that correspond to the method in question. We choose the top 30 because it is the approximate number of figures shown in a page without the need to scroll.

We consider the following questions: 1) Does the search interface tend to retrieve relevant figures for figure-oriented search tasks? 2) Which fields should be indexed to maximize accuracy? 3) Does filtering the results for an expected figure type (using the results of our classifier) improve accuracy?

To answer these questions, we use seven key phrases associated with specific figure types as our search terms: `phylogenetic`, `metabolic pathway`, `electrophoresis gel`, `confocal microscopy`, `fluorescence`, `survival curve`, and `ROC curve`. For each term, we evaluate different indexing strategies: caption only, abstract only, or abstract, title, author, and caption. Finally, we consider what effect filtering by figure type has on accuracy. For example, when searching for `phylogenetic`, the figures associated with the term are typically diagrams, so ignoring all other figure types except diagrams should improve accuracy. Other search terms are similarly associated with a dominant figure type: `phylogenetic` and `metabolic pathway` are associated with diagrams, `electrophoresis gel`, `confocal microscopy`, and `fluorescence` are associated with

photos, and `survival curve` and `ROC curve` are associated with plots.

Figure 11 shows the results. Overall, 50% to 100% of the results are relevant for each search term under the best conditions. We find that caption-only indexing provides the highest accuracy. The reason is that if a search term is mentioned in the abstract or title, then all figures in the paper are returned as results, lowering accuracy. We find that properly filtering by figure type further improves the accuracy, typically including 2-10 additional relevant figures in the top 30 results. However, in some cases filtering reduces accuracy; in these cases the classifier's imperfect type assignment is the culprit. For future work, we are working to extract information from specific figure types to enable more sophisticated content-based indexing. Despite the improved accuracy achieved by caption-only indexing, we index all fields in the current application to ensure that we return relevant papers.

The search engine is available online at www.VizioMetrics.org. Anecdotally, we have had users report that they use the interface to find figures for textbooks and presentations. They describe the system as the "google images" for scientific figures.

The one significant limitation of Viziometrics.org is the available content. Most scientific papers are held behind publisher paywalls. In our first version of the system, we have included figures from PubMed Central. Although this open corpus includes millions of figures, it only represents a small proportion of medically related research. Our hope is to extend the corpus to all disciplines, but this goal will depend on improved access to the scholarly literature.

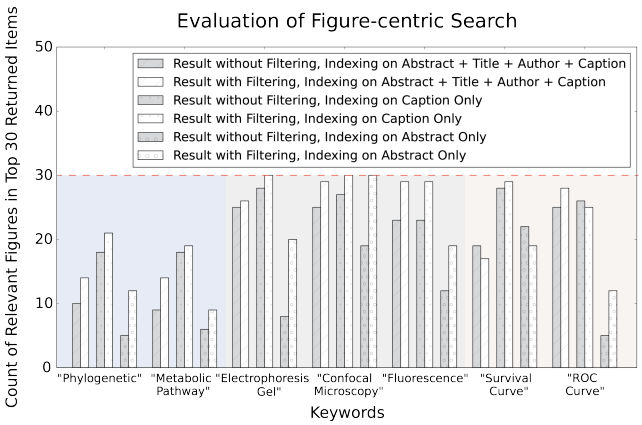


Fig. 11. We selected 7 key phrases used to describe specific methods in biology that are associated with specific visual signatures. We report the proportion, of the top 30 returned figures, that correspond to the search term. When one searches ROC curve, the results should include ROC curves. We find that filtering improves the results in all cases except a few of the plot searches. We also find that, when restricting the search index to only captions, the results tend to be slightly better. The reason is that if a search term is mentioned in the abstract or title, then all figures in the paper are returned as results, lowering accuracy.

6 Future Work

PubMed is focused primarily in the life sciences. Future work will include extending this analysis to additional domains, enabling a comparison of visual patterns across fields of study. We will expand our figure database with literature from diverse research areas and will continue to improve the accuracy of our classifications; we are currently evaluating a convolutional neural network classifier that appears to offer a different tradeoff in quality and training time. One of the key results of this paper is that more influential papers tend to have more plots and diagrams. Next steps will be refining this question and interpreting these preliminary results to understand how figures influence impact. We plan to expand the figure processing pipeline to include additional types of figures (e.g., line charts or flow charts, or domain-specific figures such as phylogenetic trees).

There are also many opportunities for exploring new search tools involving figure classifications. We have received informal feedback from users on ways in which figure types could be used. For instance, tools to support identification and directed search for specific figure types such as metabolic pathways and phylogenetic graphs could significantly accelerate research activities. In addition, information extraction from these specific figure types could afford the recovery and organization of data in support of meta-analysis activities. This information is inaccessible to text-based search engines.

One of the bottlenecks for the classifiers is the lack of labeled figures with which to train the models. We are developing a crowdsourcing component to the VizioMetrics.org platform that will integrate with the search service to acquire ground-truth labels as users interact with the system to complete their own tasks. The labels, images, code, and all our data will be freely available for researchers to explore their own questions.

7 Conclusions

In this study, our aim is to facilitate research on scientific figures, an area we call *viziometrics*. It extends prior work in bibliometrics and scientometrics but focuses on the role of visual information encodings. We developed a figure processing pipeline that automatically classifies figures into equations, diagrams, plots, photos, and tables. To facilitate further research on this visual objects, we make both the code and the data open for other researchers to explore. By integrating the figure-type labels and article metadata, we analyzed the patterns across journals, over time, and relationships to impact. In different disciplines, we found that the role of the five figure types can vary widely. For instance, clinical papers tend to have higher photo density and computational papers tend to have higher diagram and plot density. In respect to visual patterns over time, we found a growing use of plots, perhaps suggesting increasing emphasis on data-intensive methods. Our key result is that high-impact papers tend to have more diagrams per page and a higher proportion of diagrams relative to other figure types. A possible interpretation is that clarity is critical for impact: illustrating an original idea may be more influential than quantitative experimental results. We also described a new application to search and browse scientific figures, potentially enabling new kinds of search tasks. The VizioMetrics.org systems affords search by keyword as well as figure type, and shows results in a figure-centric layout. We believe more interesting and useful applications can be

inspired by the concept of viziometrics. We also encourage people to use our publicly available corpus and software to explore this area of research and create a new community of interest.

ACKNOWLEDGMENTS

We would like to thank Dastyni Loksa for help in designing early versions of the VizioMetrics.org prototype. This work is sponsored in part by the National Science Foundation through S2I2 award 1216879 and IIS award III-1064505, a subcontract from the Pacific Northwest National Lab, the University of Washington eScience Institute, the Metaknowledge Network funded by the John Templeton Foundation, and an award from the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation.

REFERENCES

[1] C. Ware, *Information visualization: perception for design*. Elsevier, 2012.

[2] D. L. Nelson, V. S. Reed, and J. R. Walling, "Pictorial superiority effect." *Journal of Experimental Psychology: Human Learning and Memory*, vol. 2, no. 5, p. 523, 1976.

[3] M. G. Gibbons, "Reassessing discovery: Rosalind Franklin, scientific visualization, and the structure of DNA," vol. 79, no. 1, pp. 63–80, Jan. 2012. [Online]. Available: http://www.jstor.org/stable/10.1086/663241

[4] P. Kalas, J. R. Graham, E. Chiang, M. P. Fitzgerald, M. Clampin, E. S. Kite, K. Stapelfeldt, C. Marois, and J. Krist, "Optical Images of an Exosolar Planet 25 Light-Years from Earth," *Science*, vol. 322, pp. 1345–, Nov. 2008.

[5] A. Dani, B. Huang, J. Bergan, C. Dulac, and X. Zhuang, "Superresolution imaging of chemical synapses in the brain," *Neuron*, vol. 68, no. 5, pp. 843 – 856, 2010.

[6] E. Garfield, "The History and Meaning of the Journal Impact Factor," *JAMA*, vol. 295, no. 1, pp. 90–93, 2006. [Online]. Available: http://jama.ama-assn.org

[7] D. J. de Solla Price, "Networks of scientific papers," *Science*, vol. 149, no. 3683, pp. 510–515, 1965.

[8] "D8taplex," http://d8taplex.com/, 2011.

[9] "Zanran," http://www.Zanran.com/, 2006.

[10] R. F. Murphy, M. Velliste, J. Yao, and G. Porreca, "Searching online journals for fluorescence microscope images depicting protein subcellular location patterns," in *Bioinformatics and Bioengineering Conference, 2001. Proceedings of the IEEE 2nd International Symposium on*. IEEE, 2001, pp. 119–128.

[11] A. Ahmed, A. Arnold, L. P. Coelho, J. Kangas, A. S. Sheikh, E. Xing, W. Cohen, and R. F. Murphy, "Structured literature image finder: Parsing text and figures in biomedical literature," *Journal of Web Semantics*, vol. 8, pp. 151–154, 2010.

[12] A. Ahmed, E. P. Xing, W. W. Cohen, and R. F. Murphy, "Structured correspondence topic models for mining captioned figures in biological literature," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 39–48.

[13] S. Ray Choudhury and C. L. Giles, "An architecture for information extraction from figures in digital libraries," in *Proceedings of the 24th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2015, pp. 667–672.

[14] S. R. Choudhury, P. Mitra, A. Kirk, S. Szep, D. Pellegrino, S. Jones, and C. L. Giles, "Figure metadata extraction from digital documents," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 135–139.

[15] S. Bhatia, P. Mitra, and C. L. Giles, "Finding algorithms in scientific articles," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 1061–1062.

[16] S. Kataria, W. Browuer, P. Mitra, and C. L. Giles, "Automatic extraction of data points and text blocks from 2-dimensional plots in digital documents." in *AAAI*, vol. 8, 2008, pp. 1169–1174.

[17] W. Browuer, S. Kataria, S. Das, P. Mitra, and C. L. Giles, "Segregating and extracting overlapping data points in two-dimensional plots," in *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2008, pp. 276–279.

[18] X. Lu, J. Wang, P. Mitra, and C. Giles, "Automatic extraction of data from 2-d plots in documents," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 1, Sept 2007, pp. 188–192.

[19] S. R. Choudhury, S. Wang, P. Mitra, and C. L. Giles, "Automated Data Extraction from Scholarly Line Graphs," 2013.

[20] Z. Chen, M. Cafarella, and E. Adar, "Diagramflyer: A search engine for data-driven diagrams," in *Proceedings of the 24th International Conference on World Wide Web Companion*, ser. WWW '15 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2015, pp. 183–186. [Online]. Available: http://dx.doi.org/10.1145/2740908.2742831

[21] R. A. Poldrack and T. Yarkoni, "From brain maps to cognitive ontologies: Informatics and the search for mental structure," *Annual review of psychology*, vol. 67, pp. 587–612, 2016.

[22] P. Hegarty and Z. Walton, "The consequences of predicting scientific impact in psychology using journal impact factors," *Perspectives on Psychological Science*, vol. 7, no. 1, pp. 72–78, 2012.

[23] T. W. Fawcett and A. D. Higginson, "Heavy use of equations impedes communication among biologists," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11 735–11 739, 2012.

[24] M. Tartanus, A. Wnuk, M. Kozak, and J. Hartley, "Graphs and prestige in agricultural journals," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 9, pp. 1946–1950, 2013.

[25] G. Cabanac, G. Hubert, and J. Hartley, "Solo versus collaborative writing: Discrepancies in the use of tables and graphs in academic articles," *Journal of the Association for Information Science and Technology*, vol. 65, no. 4, pp. 812–820, 2014.

[26] J. Hartley and G. Cabanac, "Do men and women differ in their use of tables and graphs in academic publications?" *Scientometrics*, vol. 98, no. 2, pp. 1161–1172, 2014.

[27] P.-s. Lee and B. Howe, "Dismantling composite visualizations in the scientific literature," in *International Conference on Pattern Recognition Applications and Methods, ICPRAM, Lisbon, Portugal*, 2015.

[28] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *International conference on artificial intelligence and statistics*, 2011, pp. 215–223.

[29] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer, "ReVision: Automated Classification, Analysis and Redesign of Chart Images," in *UIST '11*, 2011, pp. 393–402.

[30] B. Cheng, S. Antani, R. J. Stanley, and G. R. Thoma, "Automatic segmentation of subfigure image panels for multimodal biomedical document retrieval," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2011, pp. 78 740Z–78 740Z.

[31] M. Taschwer and O. Marques, "Automatic separation of compound figures in scientific articles," *CoRR*, vol. abs/1606.01021, 2016. [Online]. Available: http://arxiv.org/abs/1606.01021

[32] X. Wang, H. Shatkay, and C. Kambhamettu, "Cis udel working notes on image-clef 2015: Compound figure detection task," *Working Notes of CLEF*, vol. 2015, 2015.

[33] K. Santosh, Z. Xue, S. Antani, and G. Thoma, "Nlm at imageclef 2015: Biomedical multipanel figure separation," *Working Notes of CLEF*, vol. 2015, 2015.

[34] J. West, M. , D. Vilhena, and C. Bergstrom, "Ranking and mapping article-level citation networks," *In Prep.*, 2016.

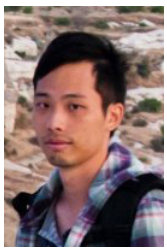
[35] J.D. West, I. Wesley-Smith, and C.T. Bergstrom, "A recommendation system based on hierarchical clustering of an article-level citation network," *IEEE Transactions on Big Data*, vol. 2, no. 2, pp. 113–123, June 2016.

[36] I. Wesley-Smith, C. T. Bergstrom, and J. D. West, "Static ranking of scholarly papers using article-level eigenfactor (alef)," ser. 9th ACM International Conference on Web Search and Data Mining. ACM, in press.

[37] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999, previous number = SIDL-WP-1999-0120. [Online]. Available: http://ilpubs.stanford.edu:8090/422/

[38] R. Lambiotte and M. Rosvall, "Ranking and clustering of nodes in networks with smart teleportation," *Physical Review E*, vol. 85, no. 5, p. 056107, 2012.
[39] J. D. West, T. C. Bergstrom, and C. T. Bergstrom, "The eigenfactor metricstm: A network approach to assessing scholarly journals," *College & Research Libraries*, vol. 71, no. 3, pp. 236–244, 2010.
[40] M. Peplow, ""no time for stodgy: Crusading editor of the bmj aims to shake things up."," http://www.statnews.com/2016/01/04/bmj-editor-fiona-godlee/, 2016.
[41] S. Ainsworth and A. T. Loizou, "The effects of self-explaining when learning with text or diagrams," *Cognitive science*, vol. 27, no. 4, pp. 669–681, 2003.
[42] P.-s. Lee, J. D. West, and B. Howe, "Viziometrix: A platform for analyzing the visual information in big scholarly data," in *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 413–418.

Bill Howe Bill Howe is Associate Professor in the Information School, Adjunct Associate Professor in Computer Science and Engineering, and Associate Director of the UW eScience Institute. His research interests are in data management, curation, analytics, and visualization in the sciences. Howe played a leadership role in the Data Science Environment program at UW through a $32.8 million grant awarded jointly to UW, NYU, and UC Berkeley. With support from the MacArthur Foundation and Microsoft, Howe leads UW's participation in the national MetroLab Network focused on smart cities and data-intensive urban science. He also led the creation of the UW Data Science Masters Degree and serves as its inaugural Program Director and Faculty Chair. He has received two Jim Gray Seed Grant awards from Microsoft Research for work on managing environmental data, has had two papers selected for VLDB Journal's "Best of Conference" issues (2004 and 2010), and co-authored what are currently the most-cited papers from both VLDB 2010 and SIGMOD 2012. Howe serves on the program and organizing committees for a number of conferences in the area of databases and scientific data management, developed a first MOOC on data science that attracted over 200,000 students across two offerings, and founded UW's Data Science for Social Good program. He has a Ph.D. in Computer Science from Portland State University and a Bachelor's degree in Industrial and Systems Engineering from Georgia Tech.



Po-shen Lee received the BS degree in Physics and MS degree in Optics and Photonics from National Central University. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of Washington. His research interests include computer vision, machine learning, and human computer interaction. He is currently working on VizioMetrics.org, including a figure-centric search engine, a crowd-sourcing platform, and a open-data platform. [more details here: students.washington.edu/sephon/]

Jevin D. West is an Assistant Professor at the University of Washington Information School and co-founder of the Datalab. He is a Data Science Fellow at the eScience Institute and Affiliate with the Center for Statistics and Social Sciences at UW. His research lies at the cross section of network science, scholarly communication, knowledge organization and information visualization. He co-founded Eigenfactor.org?a free website that ranks and maps the scholarly literature in order to better navigate and understand scientific knowledge. He has been invited to give talks at more than 50 academic and industry conferences around the globe including Harvard, Stanford and the National Academy of Sciences. Prior to joining the faculty at UW, he was a post-doc in the Department of Physics at Umea University in Sweden and received his Ph.D. in Biology from the University of Washington. [more details here: jevinwest.org]