Central Figure Identification

Sean Yang
University of Washington
Seattle, WA
tyyang38@uw.edu

Poshen Lee
University of Washington
Seattle, WA
sephonlee@gmail.com

Abhishek Joshi
University of Washington
Seattle, WA
joshi1@uw.edu

Lia Kazakova
University of Washington
Seattle, WA
kazako@uw.edu

Bum Mook Oh
University of Washington
Seattle, WA
bmo5@uw.edu

Jevin West
University of Washington
Seattle, WA
jevinw@uw.edu

Bill Howe
University of Washington
Seattle, WA
billhowe@cs.washington.edu

ABSTRACT

The graphical abstract, a visual summary of a scholarly article's main findings, is an emerging concept in scientific literature. By outlining the important results of an article, this visual aid helps the reader to comprehend the paper quickly, while also facilitating the article-searching process. However, for a paper without intentionally-produced graphical abstracts, we believe there still exists an inherent "central figure" which provides the best visual summary of that paper's objectives. Identifying central figures is a challenging problem because of the lack of labelled data for scientific literature which may be used in the task. Thus, we conducted a large-scale survey asking researchers to identify the central figures of their own work. We successfully collected data from 8,353 papers and 6,263 distinct authors. 87.6% of evaluated papers were indicated to have a figure that summarizes the key aspects of the article. Using this author-labeled data, we also conducted feature engineering, with our early-staged model achieving 35% accuracy in the identification of central figures. Our goal is to use these automated methods for improving search on visual representations of science results.

ACM Reference format:
Sean Yang, Poshen Lee, Abhishek Joshi, Lia Kazakova, Bum Mook Oh, Jevin West, and Bill Howe. 2018. Central Figure Identification. In *Proceedings of , London, UK, Aug. 2018 (BigScholar'18),* 6 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The graphical abstract, a visual summary of a scholarly article's main findings, is an emerging concept in scientific literature. Elsevier, the largest publisher of scholarly articles, requests that authors provide graphical abstracts and utilize them for online search results in facilitating the discovery process. Without any specific submission rules, 68% and 65% of papers accepted by two of the top computer vision conferences - International Conference on Computer Vision (ICCV) and Conference on Computer Vision and Pattern Recognition (CVPR) include these "teaser figures". With the abundance of scientific papers, graphical abstracts along with textual abstracts help the readers to quickly identify papers relating to their interests.

Graphical abstracts are typically made independent of other figures provided in a scientific paper that relate key processes, data, and results encompassing the paper's primary objectives. However, for papers without graphical abstracts, of figures included in a scientific paper, we believe there exists an inherent "central figure" which provides the best visual summary of that paper's objectives, much like a graphical abstract. This figure may provide greater context for any given paper in addition to its abstract, streamline the search process of papers through intuitive representation of paper concepts, and potentially allow individuals to understand the volumes of scientific literature to a greater degree.

Identifying central figures is a challenging problem. First of all, there is no existing labeled dataset for central figures in scientific literature. Even though some papers have graphical abstracts and teaser figures, as we mentioned in previous paragraph, the fact that these figures are usually made independent to other figures makes them biased data, which can not be applied to general cases. Secondly, selection of central figures can be subjective. The importance of the figure can be perceived differently by different individuals.

Thus, in this paper, we begin investigating the problem of identifying central figures by collecting labeled data through a large scale survey that we conducted using papers on PubMed Central. Figure 1 shows a snapshot of the survey interface. We asked authors to identify the central figure of their own publications as well as what aspect of information the figure represents. 488,590 survey invitations were sent and we gathered data from 8,353 papers and 6,263 distinct authors. 87.6% of evaluated papers were indicated to have a figure that summarizes the key aspects of the article. We gave the option of not selecting a central figure. Therefore, we can

BigScholar'18, Aug. 2018, London, UK

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

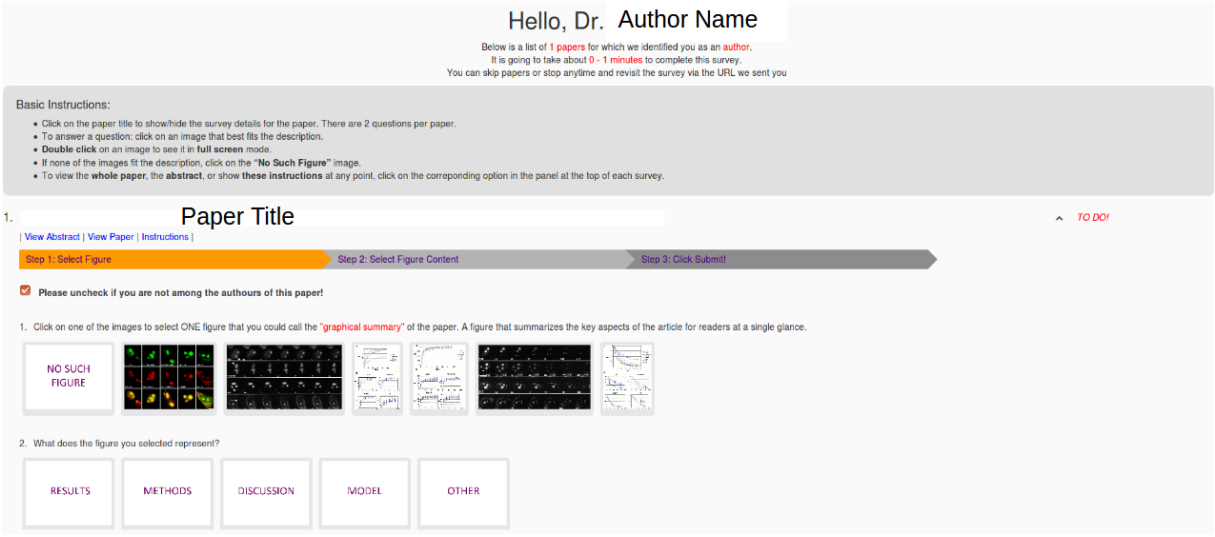


Figure 1: Snapshot of the survey. We asked authors of PubMed papers to identify the central figure of their own publications using this interface. Authors were asked to select a figure, if it exists, that summarizes the key aspects of the article, or choose "No such figure". We also asked authors to provdie what kind of information the selected figure represents for the article from five options, which are "Results", "Discussion", "Model", "Methods", and "Other".

conclude to some degree that the central figure is a concept that objectively exists in modern scientific literature.

Using the results of this survey as labels for automatically identifying these central figures, we extracted features fro the figure figures relating to visual content, textual content, and the position of a paper's image. Two models were utilized in our exploration of classifying central figures based on regression and ensemble learning. These simpler models (relative to neural networks) were used first to see which features of our data had the greatest weight in determining a central figure, and whether the models would be able to accomplish the identification task themselves.

In this preliminary analysis, we achieved approximately 35% accuracy for identifying central figures with our initial features and model. Even though the performance of the current model is not ideal, it still outperforms the baseline models where randomly selected 27.5%, pick first 25.8%, and pick last 26.9%. Text features, which includes similarity between caption and abstract and word count of the caption, seems to provide significant information to distinguish between central figures and others. Our goal is to continue to improve upon these methods.

2 RELATED WORK

Graphical abstracts (GA) have received attention from scientific researchers. Yoon et al. [23] investigated the frequency of graphical abstracts and the type of graphical abstracts are adopted in social science disciplines. Hullman [11] studied the design pattern of graphical abstracts. However, only a small collection of articles were examined in both studies (772 and 54, respectively) and both work focused on analyzing GA instead of creating a tool.

Automated tools to create an representation which summarizes scientific articles are also explored. DocumentCards[21], an automatic tool to extract textual and visual content from a scientific literature and produce a high level representation ideas, were presented by Strobelt et al. This tool, however, relies on simple rules to create the visual summary and can not be customized for different papers.

A number of studies have focused on the mining of scientific figures. Chart classification was well-studied by Futrelle et al. [9], Shao et al. [18], and Lee et al. [15]. Recent studies have been focusing on extraction of quantitative data from scientific visualizations, including line charts [16, 19], bar charts [1], and tables [7]. Researchers have also investigated the techniques to understand the semantic messages of the scientific figures. Kembhavi et al. [12] utilized a convolution neural network (CNN) to study the problem of diagram interpretation and reasoning. Elzer et al. [6] studied the intended messages in bar charts. Besides, several visualization-based search engines have been presented. DiagramFlyer [4], introduced by Chen et al., is a search engine for data-driven diagrams. VizioMetrix[13] and NOA[3] are both scientific figures search engines with big scholar data, while they both work by examining the captions around the figures.

3 DATA

This study is conducted using scientific papers from PubMed Central (PMC), an archive of biomedical and life science literature.

4 CENTRAL FIGURE SURVEY

To obtain the labeled data for central figures, we launched a large-scale survey asking authors to identify central figures in their papers. We extracted email addresses from the XML files provided by PMC API and sent out 488,590 survey invitations.

Authors are asked to answer two questions for each paper:

- *Click on one of the images to select ONE figure that you could call the "graphical summary" of the paper. A figure that summarizes the key aspects of the article for readers at a single glance.*
- *What does the figure you selected represent?*

For the first question, authors can select a figure as central figure of each paper or "No such figure" which indicates no figure fits our definition of central figure. "No such figure" provides authors an option if they are not able to choose a figure that summarize the paper and allows us to validate whether or not central figures are a concept that objectively exists in current scientific literature. For the second question, authors are allowed to choose from five options, which are "Results", "Discussion", "Model", "Methods", and "Other". From the results of this question, we can understand what aspects of the paper are considered important to the authors.

As of March 6th, 2018, we collected data on 8,353 distinct papers, from 6,263 distinct authors. 64.5% (6,187) of the evaluated papers are recent. Only 12.4% (1,036) of the evaluated paper were indicated not to have a figure that satisfies our definition of central figure (890) or multiple figures are selected from different authors of a paper (146). The collected data shows that for 87.6% of evaluated papers, authors are able to identify a single central figure. Therefore, it is reasonable to conclude that central figure is a concept that objectively exists in modern scientific literature.

Among all the central figures, 67.0% of the cases central figures are shown to represent results. It agrees with the finding in [23], where graphical abstracts are used to present results most frequently. Methods and model follow right after results with 13.6% and 12.4% of central figures representing them respectively. Discussion is responsible for only 5.1% of central figures. In 2.0% of the papers the authors indicated the content as "Other."

5 FEATURE ENGINEERING

Features for each image were extracted from three aspects: (1) visual content, (2) textual content, and (3) position in the paper. We will elaborate the details of each aspect in the following subsections.

5.1 Visual Content

We produce one categorical feature from visual content of each image. We adopted the figure class assignment in [14], where Lee et al. classify scientific figures into five types plus multi-chart:

- Diagram: Including schematics, conceptual diagrams, flow charts, architecture diagrams, illustrations.
- Plot: Figures that display quantitative results, such as bar charts, scatter plots, and line charts.
- Table: Any tabular structures with text or numeric data in the cells.
- Equation: Such as embedded equations, Greek and Latin characters.
- Photo: This class usually involves microscopy images, diagnostic images, radiology images, fluorescence imaging in PMC dataset.
- Multi-chart: Figures that include two or more types of charts.

5.2 Textual Content

We produce two numerical features from textual content: (1) Similarity between caption and abstract and (2) word count of the caption.

- **Similarity Between Caption and Abstract**: Abstract summarizes a paper and includes significant results. High similarity between caption and abstract would indicate that the image outline the paper and present important findings. Term frequency-inverse document frequency (TF-IDF) combined with cosine similarity [10] is a very common method to measure similarity between text documents. We preprocess captions and abstracts from training set by tokenizing the documents and removing the stop words, and further calculate TF-IDF model on this dataset. For each image, we apply the model to the caption of the image and abstract of the paper to acquire a pair of real number vectors and calculate the cosine similarity between the two vectors.
- **Word Count of the Caption**: We believe authors would use more words to describe the figure in the caption if it is an important figure. Thus, the number of word counts is considered as a feature as well.

5.3 Image Position

We produced two numerical features and one categorical feature from image position: (1) Section Ratio and (2) Image Order (3) Section Heading:

- **Section Ratio**: The normalized numerical identifier for sections in a paper. For example, in a paper with sections "Introduction," "Methods," and "Results," the corresponding sequentially increasing section identifiers would be 1, 2, and 3. The normalized version of the identifiers would be its original value divided by the maximum identifier value.
- **Image Order**: The sequentially increasing numerical identifier for an image based on its order of occurrence in a paper.
- **Section Heading**: The survey shows 67% of cases central figures are used to represent results. To capture this feature, we constructed unigram representations of the section headings of papers in our dataset for both the entire headings and their distinct words. We then transformed the top ten frequently occurring words in the section headings unigram model in to ten unique boolean classification features, each denoting "1" for whether the corresponding word occurred in a given paper's section heading, and "0" otherwise.

6 MODELS

In this section, we illustrate two different models based on regression and ensemble learning to identify central figures in this paper.

6.1 Image-based Model

We start with using basic machine learning classifiers to recognize central figures on image basis. Let $X = \{x_i : i\}$ be the features of the images and each image corresponds to a label y_i, where $y_i \in \{-1, 1\}$. central figures are labeled as 1 and non-central figures are labeled as -1. We learn a mapping function $f : X \rightarrow Y$ using machine learning techniques, which include logistic regression (f_{LR}), random forest (f_{RF}), Adaboost with decision tree classifier as base estimator (f_{AB}), gradient boosting (f_{GB}), and support vector machine (SVM, f_{SVM}).

To pick the central figure from a paper $A = \{a_j : j\}$, we select the figure with highest probability predicted by each classifier f:

$$C_j = \arg\max_{x_i \in A_j}(P(f(x_i) = 1))$$

6.2 Paper-based Model

This model learns on paper basis. Let $V = \{v_j : j\}$ represents feature vector for each paper. v_j is a n x d vector where n is a parameter and d is the dimension of image feature. Since there are variable number of figures in different papers and basic machine learning models only take fixed dimension inputs, we have this parameter n to serve as the threshold for the number of figures. We pad zero if the number of figure is smaller than n in a paper. For the case where the number of figure is larger than n, we select n figures whose captions are most similar to the abstract based on our TF-IDF model to fill v_j. The classifiers f will learn a mapping function $f : V \rightarrow I$, where $I \in \{0, 1, ..., n\}$ is the index of the central figure. We experiment on the ensemble and regression learning methods listed in previous sub section.

7 PRELIMINARY RESULTS

We remove the evaluated papers which do not have central figures and split the data into training, validation, and test set with 8:1:1 ratio. We run our experiments on the training and validation set. The final model is trained by the data from both training set and validation set and accuracy results reported below are conducted on test set.

7.1 Image-based Model

Table 1 shows the classification results from each classifier on identifying central figure. The accuracy is defined as:

$$acc = \frac{\text{True Positive + True Negative}}{\text{Total number of the images}} \quad (1)$$

Overall, every classifier is able to achieve more than 50% accuracy to classify between central figure and non central figure. However, these models do not perform well when it comes to picking the central figure given a scientific article. The accuracy of central figure prediction given paper is shown in Table 2 and the baseline models is shown in Table 3. The accuracy of this task is defined as:

$$acc = \frac{N_c}{N_t} \quad (2)$$

where N_c is the number of the papers with correct prediction of central figure and N_t is the total number of the papers in the dataset.

Not surprisingly, this simple image-based model does not perform well on selecting the central figure from a list of figures. The model is not able to learn the structural relationships between figures from the same paper. Thus, proposed paper-based model should have better performance and the results are reported in 7.2

Table 1: Accuracy of central figures classification

	Logistic Regression	Decision Tree	Random Forest
Accuracy	0.626	0.546	0.616
	AdaBoost	Gradient Boosting	SVM
Accuracy	0.619	0.621	0.673

Table 2: Accuracy of central figure prediction given paper

	Logistic Regression	Decision Tree	Random Forest
Accuracy	0.140	0.248	0.170
	AdaBoost	Gradient Boosting	SVM
Accuracy	0.126	0.126	0.142

Table 3: Baseline Models for Comparison

Random Selection	Pick First	Pick Last
0.275	0.258	0.269

7.2 Paper-based Model

We run a simple experiment to pick our parameter n (The threshold for the number of figure to be accommodated for the input V). The experiment results are shown in Figure 2. Blue line, which corresponds to the y axis on the left, is the accuracy of the model and the red line shows how many percentage of central figures were left out because of our selection of n. The selection of n insignificant influence to the model when n is larger than 6 and the maximum number of figure a paper has in our validation set is 12. Thus, we pick $n = 15$ for the rest of the experiments.

The results for paper-based model with different feature combinations are shown in Table 4. We removed the classifiers that do worse than baseline models. Surprisingly, logistic regression classifiers outperforms random forest and gradient boosting. In terms of feature, information extracted from text content is the most helpful on identifying central figures among the three features and visual content seems a non-factor in current model. Our interpretation of these results is that similarity between the figure caption and paper abstract not only provides the representation of the image but it also suggests the relationship to the paper. On the other hand, without any further information of the paper, figure type is irrelevant to determine central figure in this generation of the model.

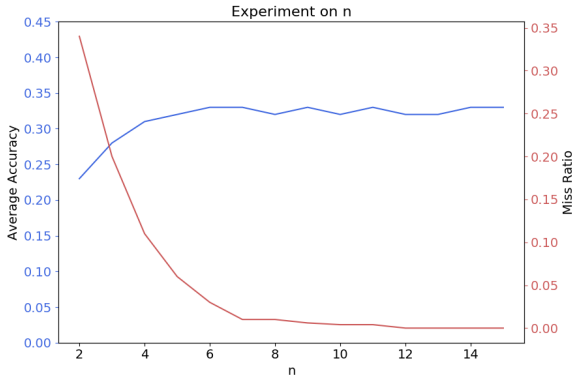


Figure 2: Experimental results on hyperparameter n. When n is larger than 6, selection of n does not affect the accuracy of the model.

The insufficiency with current features and model is expected. As mentioned above, the current version is lack of the information of the articles, such as disciplines and publishing journal, which could assist machine identifying the central figure based on the knowledge of the paper. Plus, the simple machine learning classifiers we utilize in current version are not ideal for this task because of the mismatch between the problem settings. Low accuracy from simple classifiers also demonstrates that the problem of central figure identification is challenging that it requires more sophisticated models to deal with. However, these experiments provide a direction for future exploration which we will elaborate in Sec. 8.

Table 4: The results of paper-based model with different classifiers. Surprisingly, logisitc regression outperforms random forest and gradient boosting. Textual content are the most useful feature on recognizing central figure, compared to visual content and the position feature.

	Logistic Regression	Random Forest	Gradient Boosting
Visual + Text + Position	0.347	0.293	0.322
Visual + Position	0.329	0.279	0.327
Visual + Text	0.330	0.308	0.324
Text + Position	0.349	0.322	0.340
Visual	0.307	0.296	0.287
Text	0.333	0.326	0.329
Position	0.318	0.288	0.341

8 DISCUSSION

In this study, we utilize a large scale qualitative survey that asks researchers about their central figures, if they exist at all. We use this survey data as labels for developing methods of automatically identifying these graphical abstracts.

The results from the survey support some of the findings of previous research conducted in the area of information visualization in scientific literature. We found that despite the fact that plots (graphs) and tables can both be used for presenting data, plots are a much more popular class when it comes to presenting key information. This is inline with findings described by Cleveland, who showed the overall increase in fractional graph area (FGA) in sciences going from social to mathematical, to natural science [5]. It also agrees with findings by Smith et al, point out that harder sciences seem to be more graph-oriented than table-oriented [20]. The fact that equations are rarely found among central figures is consistent with findings by Fawcett and Higginson [8].

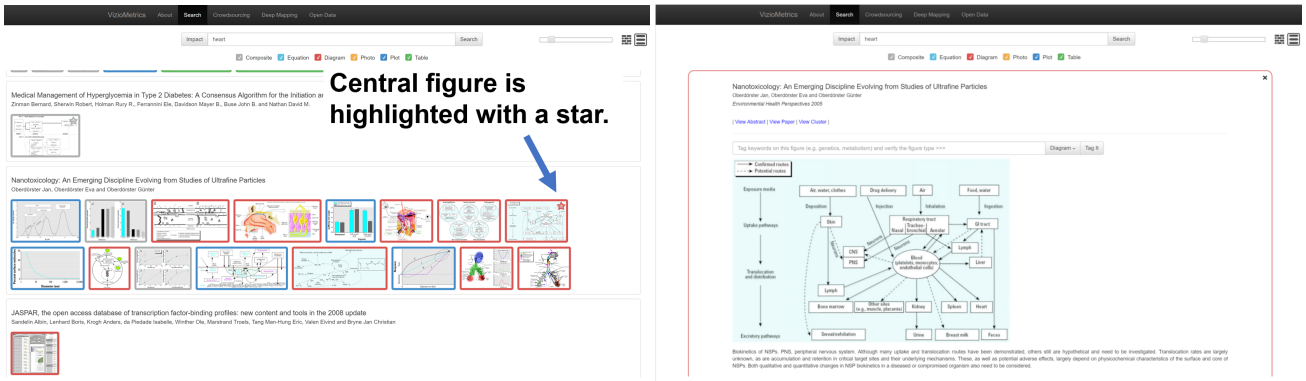
Our vision for viziometrics.org[13], a platform that extracts visual information from the scientific literature and makes it available for use in new information retrieval applications, is connecting big scholarly articles using visual information along with improving image-based scientific search and central figure could play a significant role along the way. Figure 3 shows a couple of screenshots of the current interface of viziometrics.org. The algorithm to identify central figure on current viziometrics.org is picking the figure whose caption is the most similar to the paper abstract using TF-IDF technique. As shown in Figure 3(a), central figure is highlighted with a star on the search interface. We will further re-design the entry page for each article which features the central figure along with textual abstracts as shown in 3(b). With these two additional features, users are able to quickly grab the overall concept of the article with the help of central figure at a single glance. The further exploration of central figure identification will provide greater accuracy and breath for current 1M papers available on the site and for the future scientific literature.

We will further investigate the problem of recognizing central figures by improving the features we utilize as well as our model. In terms of features, we will include more information regarding the context of figures' papers for the next version, such as the discipline of a paper, citation count, journal, and so on. Besides, we will calculate the similarity between the text around the mention of the figure and the paper abstract to be another textual feature. We will also utilize state-of-art image embedding techniques to extract visual feature to represent each image.

With regard to models, we are planning to investigate more sophisticated models, including learning to rank models and DNNs (Deep Neural Network) specifically. We can model this problem as a ranking problem, where central figures have higher scores than non-central figures. The machine will learn to rank from a list of instances[2]. We are also interested in exploring LSTMs (Long Short-Term Memory), Tseng's work [22] to be more specific. Tseng et al. proposed an Attention-based Multi-hop Recurrent Neural Network for machine to learn to answer the questions based on a given story. It is similar to our setting where our machine needs to learn how to pick the central figure according to a given paper. Thus, this model is promising for our task.

9 CONCLUSION

We successfully collected more than 8 thousand labeled data for central figure identification from a large-scale survey. 87.6% of the evaluated papers were indicated to have a figure that outlines the important aspects of the article, which could be concluded that the concept of central figure objectively exists. A early-staged paper-based model is proposed to identify central figures and achieve



(a) (b)

Figure 3: Viziometrics.org interface allows individuals to search for images from scientific literature and currently view starred "central figures" determined by TF-IDF between image captions and papers' abstracts. Our research may provide greater representation of central figures in papers for such visual search interfaces. (a) Viziometrics.org has the central figure starred for easy recognition on searching interface. (b) Prototype of entry page for each article on viziometrics.org. We are planning to implement a feature where the entry of each article would be led with the central figure along with textual abstract to help the users understand the articles quickly. The paper demonstrated in the figures is the work by Oberdorster et al. [17]

approximately 35% accuracy in this paper. We will further investigate the problem with improved feature and more sophisticated model.

REFERENCES

[1] Rabah A Al-Zaidy and C Lee Giles. 2015. Automatic extraction of data from bar charts. In *Proceedings of the 8th International Conference on Knowledge Capture*. ACM, 30.

[2] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*. ACM, 129–136.

[3] Jean Charbonnier, Lucia Sohmen, John Rothman, Birte Rohden, and Christian Wartena. 2018. NOA: A Search Engine for Reusable Scientific Images Beyond the Life Sciences. In *European Conference on Information Retrieval*. Springer, 797–800.

[4] Zhe Chen, Michael Cafarella, and Eytan Adar. 2015. Diagramflyer: A search engine for data-driven diagrams. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 183–186.

[5] William S Cleveland. 1984. Graphs in scientific publications. *The American Statistician* 38, 4 (1984), 261–269.

[6] Stephanie Elzer, Sandra Carberry, and Ingrid Zukerman. 2011. The automated understanding of simple bar charts. *Artificial Intelligence* 175, 2 (2011), 526–555.

[7] Jing Fang, Prasenjit Mitra, Zhi Tang, and C Lee Giles. 2012. Table Header Detection and Classification.. In *AAAI*. 599–605.

[8] Tim W Fawcett and Andrew D Higginson. 2012. Heavy use of equations impedes communication among biologists. *Proceedings of the National Academy of Sciences* 109, 29 (2012), 11735–11739.

[9] Robert P Futrelle, Mingyan Shao, Chris Cieslik, and Andrea Elaina Grimes. 2003. Extraction, layout analysis and classification of diagrams in PDF documents. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*. IEEE, 1007–1013.

[10] Anna Huang. 2008. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*. 49–56.

[11] Jessica Hullman and Benjamin Bach. [n. d.]. Picturing Science: Design Patterns in Graphical Abstracts. ([n. d.]).

[12] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European Conference on Computer Vision*. Springer, 235–251.

[13] Poshen Lee, Jevin West, and Bill Howe. 2016. VizioMetrix: A Platform for Analyzing the Visual Information in Big Scholarly Data. In *BigScholar Workshop (co-located at WWW)*.

[14] Poshen Lee, Jevin D West, and Bill Howe. 2017. Viziometrics: Analyzing visual information in the scientific literature. *IEEE Transactions on Big Data* (2017).

[15] Poshen Lee, T. Sean Yang, Jevin West, and Bill Howe. 2017. PhyloParser: A Hybrid Algorithm for ExtractingPhylogenies from Dendrograms. (2017).

[16] Xiaonan Lu, J Wang, Prasenjit Mitra, and C Lee Giles. 2007. Automatic extraction of data from 2-d plots in documents. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, Vol. 1. IEEE, 188–192.

[17] Günter Oberdörster, Eva Oberdörster, and Jan Oberdörster. 2005. Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. *Environmental health perspectives* 113, 7 (2005), 823.

[18] Mingyan Shao and Robert P Futrelle. 2005. Recognition and classification of figures in PDF documents. In *International Workshop on Graphics Recognition*. Springer, 231–242.

[19] Noah Siegel, Zachary Horvitz, Roie Levin, Santosh Divvala, and Ali Farhadi. 2016. FigureSeer: Parsing result-figures in research papers. In *European Conference on Computer Vision*. Springer, 664–680.

[20] Laurence D Smith, Lisa A Best, D Alan Stubbs, Andrea Bastiani Archibald, and Roxann Roberson-Nay. 2002. Constructing knowledge: The role of graphs and tables in hard and soft psychology. *American Psychologist* 57, 10 (2002), 749.

[21] Hendrik Strobelt, Daniela Oelke, Christian Rohrdantz, Andreas Stoffel, Daniel A Keim, and Oliver Deussen. 2009. Document cards: A top trumps visualization for documents. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1145–1152.

[22] Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Towards machine comprehension of spoken content: Initial TOEFL listening comprehension test by machine. *arXiv preprint arXiv:1608.06378* (2016).

[23] JungWon Yoon and EunKyung Chung. 2017. An investigation on Graphical Abstracts use in scholarly articles. *International Journal of Information Management* 37, 1 (2017), 1371–1379.