Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Quality-Driven Co-Optimization of Data Cleaning and Clustering: Framework and Mechanistic Evidence

First A. Author1, Fellow, IEEE, Second B. Author2, and Third C. Author, Jr.3, Member, IEEE

1National Institute of Standards and Technology, Boulder, CO 80305 USA

2Department of Physics, Colorado State University, Fort Collins, CO 80523 USA

3Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA

Corresponding author: First A. Author (e-mail: author@ boulder.nist.gov).

This paragraph of the first footnote will contain support information, including sponsor and financial support acknowledgment. For example, “This work was supported in part by the U.S. Depart­ment of Com­merce under Grant BS123456.”

ABSTRACT Real-world data often contains multiple types of errors, such as missing values and anomalies. Such fluctuations in data quality expand the search space of unsupervised AutoML and compromise the stability of the results. This paper proposes a core hypothesis: there is a predictable structure between data quality and clustering features. As long as sufficient data quality and algorithm features are captured, high-potential cleaning and clustering combinations can be screened out in advance. Based on this hypothesis, we developed three strategies: in the offline phase, feature enhancement is used to improve candidate hit rates; in the search phase, pruning is used to compress the space to O(klog n); and in the online phase, clustering statistics are monitored and parameter drift is dynamically corrected. The experimental section follows a process of compatibility → mechanism → gain and verifies that the proposed scheme improves runtime speed by more than 8 times compared to full search on datasets of different scales, while the average loss in clustering quality does not exceed 2%. The results demonstrate that the proposed collaborative optimization framework can achieve both speed and robustness in large-scale scenarios involving dirty data.

INDEX TERMS Enter key words or phrases in alphabetical order, separated by commas. For a list of suggested keywords, send a blank e-mail to keywords@ieee.org or visit http://www.ieee.org/organizations/  
pubs/ani\_prod/keywrd98.txt

1. INTRODUCTION

In data-intensive scenarios, such as precision medicine, financial risk management, and industrial IoT, unsupervised clustering has become a crucial tool for discovering potential patterns and supporting informed decision-making. However, clustering algorithms are susceptible to data distribution: a single spelling error, an unlabeled anomaly, or a missing interval can distort the distance structure and density estimates, thereby disrupting cluster partitioning, convergence paths, and even downstream analysis chains. Unlike supervised learning, which relies on labels to filter out noise, clustering is almost entirely dependent on the quality of the raw data. An intuitive solution is to leverage AutoML to chain together several data cleaning operators with various clustering algorithms, then let an optimizer automatically select the best combination in the parameter space. However, as the number of cleaning operators, clustering models, and hyperparameter dimensions grows exponentially, optimizers often waste computational resources on a large number of doomed-to-fail solutions, thereby undermining the value of automation itself.

Over the past decade, data cleaning and clustering algorithms have each made significant breakthroughs: for multi-source heterogeneous data, data cleaning methods such as missing value imputation, anomaly detection, and fault-tolerant matching have emerged in rapid succession; while K-Means, DBSCAN, hierarchical clustering, and even deep embedding clustering have demonstrated good adaptability in complex data structures. However, the two technical approaches have developed almost independently: data cleaning research has focused on local error correction, with little consideration of the cascading effects of corrections on unsupervised tasks; clustering research has assumed that the input is “sufficiently clean,” and has primarily focused on similarity measurement and optimization strategies. As data volume and data quality issues grow in tandem, this disconnect leads to threefold amplified complexity: data cleaning alters distance distributions, causing existing hyperparameters to become mismatched; data quality varies greatly across tables and batches, making it challenging to apply a single strategy; and the exponential number of candidate combinations causes system overhead to skyrocket, which cannot be offset linearly by increasing computing power. The industry has attempted to utilize AutoML to link cleaning and clustering; however, in the absence of prior knowledge, this approach still falls into the trap of “blind trial and error,” with input and output often being disproportionate.

Given this practical challenge, this paper proposes a core hypothesis supported by empirical evidence: the impact of data quality on clustering performance has a learnable structure. In other words, as long as the characteristics of data contamination and algorithm features are described in sufficient detail, the quality of the cleaning-clustering pipeline can be predicted before time-consuming evaluation. This hypothesis leads to two direct implications: first, construct a mapping Φ that maps data quality feature vectors to the priority of candidate pipelines; second, if Φ is a reliable predictor, we can boldly prune the pipeline at the beginning of the search, retaining only the top-ranked candidates and concentrating computational resources on the most promising few solutions.

Based on this hypothesis, we developed a three-stage framework that strikes a balance between speed, robustness, and interpretability. First, in the feature enhancement stage, we introduce higher-order interaction features and operators in addition to traditional statistics, and use historical experimental data to train a gradient boosting ranking model to implement Φ and generate a ranking of candidate pipelines. Next, in the confidence upper bound pruning stage, we calculate the upper bound of the cluster score that each candidate pipeline can achieve. If this upper bound is lower than the current optimal value, the candidate is immediately eliminated. Theoretical proof shows that the upper bound of the retained candidate size is O(klogn), where k is the number of truly high-quality solutions and n is the size of the original search space. Finally, in the runtime phase, we introduce dynamic tuning, which monitors the contour coefficient slope and Davies–Bouldin index in real time. Once distribution drift is detected, key hyperparameters, such as k and ε, are automatically fine-tuned to restore robustness without requiring a restart of the search.

The main contributions of this paper are as follows: (1) We systematically propose and validate the hypothesis that there is a structural correlation between the granularity of data cleaning and the effectiveness of clustering. Based on experiments with eight types of cleaners, six types of clusterers, and 60 public datasets, we reveal four layers of patterns, including monotonic segmentation, threshold inflection points, and interactions, providing a quantitative basis for the development of automated algorithms.

(2) We propose a confidence upper bound-driven search pruning algorithm, design a feature-enhanced multi-label predictor, and combine it with a confidence upper bound strategy to prune the search tree for unsupervised AutoML. Theoretically, the upper bound of the search space is reduced to O(klogn), and experimentally, we achieve 6–12 × acceleration and accuracy loss < 2% on a million-row dataset.

(3) We propose a runtime parameter drift monitoring and adaptive rollback strategy that automatically adjusts key hyperparameters, such as k, ε, and γ, when a decline in SSE slope or drift in DB index is detected. On 1 million lines of real and simulated data, we control the clustering stability fluctuation within ±1% while maintaining an overall acceleration of 8×.

The structure of this paper is as follows: Section 2 reviews recent advances in the cross-domain of data cleaning, clustering, and AutoML; Section 3 formally defines the data quality vector, objective function, and evaluation metrics; Section 4 details the proposed three-stage framework and provides complexity and theoretical analysis; Section 5 introduces the experimental setup, including the dataset, data corruption injection strategy, and complete experimental design; Section 6 presents experimental results along these three main lines and conducts in-depth discussions; Section 7 summarizes the findings and provides conclusions and future directions; the appendix provides additional figures and open-source code links.[[1]](#footnote-1)

1. RELATED WORK

This chapter reviews relevant research from three main perspectives: data cleaning and quality management, robustness of clustering algorithms in dirty data environments, and the latest developments in unsupervised AutoML. It summarizes existing limitations and lays the foundation for the problem definition and methods in the next chapter.

Data cleaning aims to detect and repair various defects such as missing values, outliers, duplicates, and format errors, and is a prerequisite for improving the usability of analysis. Early methods focused on statistical imputation (mean, mode) or rule-based anomaly detection. Subsequently, techniques such as probabilistic graph models, active learning, and neural networks significantly improved the ability to identify complex errors. For domain-specific data, customized cleaning frameworks combining knowledge graphs or external constraints have also emerged. However, unsupervised tasks with no labels lack “clean controls,” and excessive cleaning can easily delete a small number of critical anomalies, while conservative cleaning amplifies noise interference. Existing work has focused mainly on the “data quality” dimension. It has not directly linked cleaning decisions to downstream clustering performance, which is one of the motivations for this paper to develop a joint “cleaning-clustering-search” model further.

Among the three mainstream clustering algorithms, centroid-based K-Means and its improved methods are efficient on large-scale data but sensitive to outliers and geometric structures; density-based DBSCAN and OPTICS can identify clusters of arbitrary shapes and are more robust to low-density noise, but they are highly dependent on hyperparameters *ε* and *minPts*; bottom-up hierarchical clustering is suitable for capturing multi-scale structures but has high computational complexity. To address dirty data, researchers have proposed local repair strategies such as weighted distance and robust centroid updating, or introduced local density re-estimation into DBSCAN to mitigate misclassification. However, these methods mainly optimize either “cleaning” or “clustering” separately, lacking a holistic perspective. The impact of different cleaning operations on cluster shapes, parameter sensitivity, and the iterative paths within algorithms remains poorly quantified.

AutoML frameworks (Auto-sklearn, TPOT, etc.) have enabled the automatic search of “model–feature–hyperparameter” combinations in supervised tasks. Representative research in the unsupervised direction has primarily focused on the automatic recommendation of clustering algorithms and hyperparameters, but rarely includes data cleaning or utilizes PCA as a noise reduction method. Recent work has attempted to incorporate cleaning operators into the search space; however, most remain at the level of empirical rules and have not yet provided a unified, end-to-end, closed-loop modeling approach.

To address these shortcomings, this paper will utilize data and algorithm feature prediction to identify the optimal combination and reduce the search space, while quantitatively describing the impact of cleaning accuracy on the clustering process and results. This will lead to the proposal of a unified end-to-end optimization framework for “cleaning × clustering × hyperparameter tuning.”

PROBLEM DEFINITION

1. METHODOLOGY AND NOTATION

**Definition 1** (Data Matrix). Let be the raw data matrix, where is the number of samples and is the number of features. We divide the column index set into (numeric, categorical, and text columns) according to their types. Both rows and columns may contain missing values, outliers, or format errors.

**Definition 2** (Set of cleaning operators). To describe the available data cleaning methods, let be the set of candidate operators. This paper's experiments involve a variety of classic data cleaning methods, including missing value imputation, outlier detection, format correction, semantic consistency repair, and noise suppression. Each operator is accompanied by its hyperparameter vector .

**Definition 3 (Set of clustering algorithms).** Correspondingly, let denote the set of clustering algorithms. This study focuses on three major categories of methods: centroid-based (*K-Means*), density-based (*DBSCAN*), and hierarchical (*Hierarchical*). Each is accompanied by a hyperparameter vector (e.g., number of clusters , radius , minimum density ).

1. PROBLEM STATEMENT

**Definition 4 (candidate pipeline).** Combine any cleaning subset , clusterer , and their joint hyperparameter to obtain a candidate cleaning-clustering pipeline .

**Definition 5 (Data quality characteristics).** In order to quantitatively describe the “dirtiness” of A, this work focuses on two primary defects: missing values and outliers. We introduce a two-dimensional normalized vector corresponding to the proportion of missing and outlier samples in the whole sample. Vector , together with its second-order interaction features, will be used as input for the learning model in Section 4.

**Definition 6 (Algorithm Features).** Corresponding to the data quality features, we abstract the cleaning–clustering pipeline itself as a set of learnable features , where is a combination of multi-hot encoding operators and their hyperparameter values, and describes the selected clusterer type and its hyperparameters. The dimension of vector vary with the candidate pipeline but are uniquely determined for any .

**Definition 7 (Search Space).** Given the cleaning set , clustering set , and corresponding hyperparameter domains defined in Definitions 2–4, the universal set of candidate pipelines is denoted as , whose size satisfies and typically grows exponentially. Empirically, the number of pipelines with truly outstanding performance is far smaller than .

**Definition 8 (Comprehensive Metric).** To evaluate clustering performance from different perspectives, we use a weighted metric , where ; The cost term is measured by normalized runtime—the linear combination of the two yields the comprehensive metric , which determines the speed-quality trade-off. Preliminary experiments (Section 3.3) show that achieves a good balance on most datasets.

**Assumption 1 (Confidence Mapping).** We construct a feature vector and assume that there exists a function such that holds on the training set, where is selected via preliminary experiments. A gradient boosting sorter implements this mapping and will provide a “confidence upper bound” in the algorithm phase to prune large search trees.

**Definition 9 (Optimization Objective).** Let denote the global optimal quality. If there exists a subset such that and , then is called an *-*sufficient subset, where is the user-tolerable error and is a constant related to the depth of the search tree, is an absolute constant independent of the implementation. In this subset, is referred to as the *-*approximate optimal pipeline and is output as the search result.

Definitions 5-8, together with the above optimization objectives, closely couple the elements of “data quality algorithm featurecomprehensive effectiveness.” The next chapter will design a three-stage collaborative optimization algorithm based on this and prove that it can achieve -approximate optimality within the evaluation budget.

1. PRELIMINARY EXPERIMENTS

METHODS

Based on the notation and optimization objectives presented in Section 3, Section 4 proposes a collaborative optimization framework consisting of three interconnected steps: feature enhancement, search pruning, and dynamic optimization. The complete flowchart of this method is shown in Figure 4-1.”

1. FEATURE ENHANCEMENT

As described in the problem definition, we assume that the two-dimensional data quality vector and the algorithm feature have a specific structural matching relationship. To this end, this section presents a set of offline joint representation and confidence predictor construction processes: First, extract for each historical pipeline, where and are the multi-hot encodings of the cleaning operator and the clusterer, respectively, is the clusterer hyperparameter normalization, is the number of samples and features after -scaling followed by Min–Max normalization, and the final term is the second-order interaction between dirtiness and the clusterer. Train LightGBM-Ranker using to obtain the point estimate , then use -fold prediction variance to estimate the uncertainty , thereby obtaining the confidence predictor that satisfies  .

**Algorithm 1：Feature Enhancement and Confidence Learning**

Input: Historical evaluation set 𝔇,  
Output: Confidence predictor   
1 for each do  
2   
3   
4   
5   
6   
7 Train LightGBM using to obtain   
8 Fit Quantile Random Forest to residuals, output  
9 return

The offline phase completes the extraction of the model and characterizes uncertainty in one step; subsequent algorithms can directly call upon Φ's point estimates and confidence intervals.

1. FEATURE ENHANCEMENT

When training the predictor , we adopt the “upper confidence bound + intra-layer Top-” search principle, gradually expanding the pipeline tree layer by layer and immediately pruning branches when confidence is insufficient. This approach balances exploration (the width of the confidence interval) and exploitation (the current optimal score), thereby quickly locking in high-quality candidates.

**Algorithm 2：Confidence – UCB Search**

Input : Search space , predictor , parameters   
Output: Evaluation set , approximate optimal pipeline   
  
1 Initialize ← , ← -∞ , ←   
2 while do  
3 Select node ←   
4 if then  
5 ← # Prune the entire subtree  
6 else  
7 Perform a true evaluation ， ←   
8 if then ← , ←  
9 Generate the child node V of ，take Top- enqueue: ←   
10 end if  
11 end while  
12 return ,

With the mechanism of “optimistic estimation + overall elimination of subtrees,” this algorithm triggers at most real evaluations at each layer. Combined with the number of layers , the total number of evaluations is strictly limited, leaving sufficient resources for the subsequent online phase.

1. DYNAMIC TUNING

Offline search can only guarantee optimality on the training distribution. When the online data distribution and system load drift, the pipeline needs to be corrected promptly to maintain performance. This section implements lightweight and adaptive runtime optimization through a three-step strategy of monitoring–triggering–fine-tuning.

**Algorithm 3：Window monitoring and local fine-tuning**

Input : Initial pipeline , monitoring window , thresholds   
Output: Real-time optimal pipeline   
  
1 ←   
2 for each new batch of data do  
3 calculate current metrics   
4 update sliding window sequence 5 if AND then  
6 # --- Trigger local fine-tuner ---  
7 Construct neighborhood   
8 for (subsample 10%) do  
9 Evaluate   
10 if then  
11 Update   
12 If the cleaning strategy changes, incrementally refresh , re-obtain   
13 break  
14 end for  
15 end if  
16 end for  
17 return

This online module is only activated when indicators show significant stagnation and deterioration, and uses a “neighborhood probe + result threshold” mechanism to avoid frequent fluctuations. If the cleaning decision is changed, is also updated synchronously and reevaluated using to avoid deviating from the original confidence framework.

1. COMPLEXITY AND THEORETICAL UPPER BOUND

This section derives the core processes of the first three subsections from the perspectives of time and space, and provides overall upper bounds and approximate optimality guarantees.

1. Offline feature enhancement

LightGBM iterates through all samples and candidate split points for each split; with samples and -dimensional inputs, the typical implementation has a training cost of and This stage is only executed offline during model updates, and modern distributed implementations can leverage data sharding for parallel processing, so the impact on overall latency is negligible.

1. Confidence Search Trimming

Under the Top- Top- expansion strategy, the depth of the complete ary tree satisfies For any layer, Algorithm 2 first selects nodes according to : if the first node is pruned, the upper bound of the nodes in the same layer must be lower, and the entire layer terminates immediately; if it is not pruned, at most nodes can refresh the best, otherwise its UCB also drops to and cannot be retained. Therefore, the "evaluation amount" of each layer . If the optimal pipeline ancestor is pruned, then , which contradicts ; Therefore, must be retained and finally evaluated, and . The total time and space consumption can be obtained as follows: In the typical configuration of and , , which saves about times the time and the same level of memory compared to brute force evaluation of pipelines.

1. Online Dynamic Tuning

Triggered only when “performance stagnates and deteriorates”: Assuming that is activated a total of times throughout the year, and each time the group parameters (10% subsample) are evaluated in the local neighborhood , then Since and is typically on the order of hundreds, this term contributes very little to the overall latency.

1. End-to-End Aggregation

The three terms are added together to get In the real world where offline models are stable and online tuning is rare, the second term (confidence search) accounts for almost all the cost; and it still maintains a logarithmic advantage over brute force search.

Vertical lines are optional in tables. Statements that serve as captions for the entire table do not need footnote letters.

aGaussian units are the same as cg emu for magnetostatics; Mx = maxwell, G = gauss, Oe = oersted; Wb = weber, V = volt, s = second, T = tesla, m = meter, A = ampere, J = joule, kg = kilogram, H = henry.

1. MULTIPART FIGURES

Figures compiled of more than one sub-figure presented side-by-side, or stacked. If a multipart figure is made up of multiple figure types (one part is lineart, and another is grayscale or color) the figure should meet the stricter guidelines.

1. FILE FORMATS FOR GRAPHICS

Format and save your graphics using a suitable graphics processing program that will allow you to create the images as PostScript (PS), Encapsulated PostScript (.EPS), Tagged Image File Format (.TIFF), Portable Document Format (.PDF), or Portable Network Graphics (.PNG) sizes them, and adjusts the resolution settings. If you created your source files in one of the following programs you will be able to submit the graphics without converting to a PS, EPS, TIFF, PDF, or PNG file: Microsoft Word, Microsoft PowerPoint, or Microsoft Excel. Though it is not required, it is strongly recommended that these files be saved in PDF format rather than DOC, XLS, or PPT. Doing so will protect your figures from common font and arrow stroke issues that occur when working on the files across multiple platforms. When submitting your final paper, your graphics should all be submitted individually in one of these formats along with the manuscript.

1. SIZING OF GRAPHICS

Most charts, graphs, and tables are one column wide (3.5 inches / 88 millimeters / 21 picas) or page wide (7.16 inches / 181 millimeters / 43 picas). The maximum depth a graphic can be is 8.5 inches (216 millimeters / 54 picas). When choosing the depth of a graphic, please allow space for a caption. Figures can be sized between column and page widths if the author chooses, however it is recommended that figures are not sized less than column width unless when necessary.

There is currently one publication with column measurements that do not coincide with those listed above. Proceedings of the IEEE has a column measurement of 3.25 inches (82.5 millimeters / 19.5 picas).

The final printed size of author photographs is exactly   
1 inch wide by 1.25 inches tall (25.4 millimeters x 31.75 millimeters / 6 picas x 7.5 picas). Author photos printed in editorials measure 1.59 inches wide by 2 inches tall (40 millimeters x 50 millimeters / 9.5 picas x 12 picas).

1. RESOLUTION

The proper resolution of your figures will depend on the type of figure it is as defined in the “Types of Figures” section. Author photographs, color, and grayscale figures should be at least 300dpi. Line art, including tables should be a minimum of 600dpi.

1. VECTOR ART

In order to preserve the figures’ integrity across multiple computer platforms, we accept files in the following formats: .EPS/.PDF/.PS. All fonts must be embedded or text converted to outlines in order to achieve the best-quality results.

1. COLOR SPACE

The term color space refers to the entire sum of colors that can be represented within the said medium. For our purposes, the three main color spaces are Grayscale, RGB (red/green/blue) and CMYK (cyan/magenta/yellow/black). RGB is generally used with on-screen graphics, whereas CMYK is used for printing purposes.

All color figures should be generated in RGB or CMYK color space. Grayscale images should be submitted in Grayscale color space. Line art may be provided in grayscale OR bitmap colorspace. Note that “bitmap colorspace” and “bitmap file format” are not the same thing. When bitmap color space is selected, .TIF/.TIFF/.PNG are the recommended file formats.

1. ACCEPTED FONTS WITHIN FIGURES

When preparing your graphics IEEE suggests that you use of one of the following Open Type fonts: Times New Roman, Helvetica, Arial, Cambria, and Symbol. If you are supplying EPS, PS, or PDF files all fonts must be embedded. Some fonts may only be native to your operating system; without the fonts embedded, parts of the graphic may be distorted or missing.

A safe option when finalizing your figures is to strip out the fonts before you save the files, creating “outline” type. This converts fonts to artwork what will appear uniformly on any screen.

1. USING LABELS WITHIN FIGURES
2. Figure Axis labels

Figure axis labels are often a source of confusion. Use words rather than symbols. As an example, write the quantity “Magnetization,” or “Magnetization M,” not just “M.” Put units in parentheses. Do not label axes only with units. As in Fig. 1, for example, write “Magnetization (A/m)” or “Magnetization (Am−1),” not just “A/m.” Do not label axes with a ratio of quantities and units. For example, write “Temperature (K),” not “Temperature/K.”

Multipliers can be especially confusing. Write “Magnetization (kA/m)” or “Magnetization (103 A/m).” Do not write “Magnetization (A/m) × 1000” because the reader would not know whether the top axis label in Fig. 1 meant 16000 A/m or 0.016 A/m. Figure labels should be legible, approximately 8 to 10 point type.

1. Subfigure Labels in Multipart Figures and Tables

Multipart figures should be combined and labeled before final submission. Labels should appear centered below each subfigure in 8 point Times New Roman font in the format of (a) (b) (c).

1. FILE NAMING

Figures (line artwork or photographs) should be named starting with the first 5 letters of the author’s last name. The next characters in the filename should be the number that represents the sequential location of this image in your article. For example, in author “Anderson’s” paper, the first three figures would be named ander1.tif, ander2.tif, and ander3.ps.

Tables should contain only the body of the table (not the caption) and should be named similarly to figures, except that ‘.t’ is inserted in-between the author’s name and the table number. For example, author Anderson’s first three tables would be named ander.t1.tif, ander.t2.ps, ander.t3.eps.

Author photographs should be named using the first five characters of the pictured author’s last name. For example, four author photographs for a paper may be named: oppen.ps, moshc.tif, chen.eps, and duran.pdf.

If two authors or more have the same last name, their first initial(s) can be substituted for the fifth, fourth, third... letters of their surname until the degree where there is differentiation. For example, two authors Michael and Monica Oppenheimer’s photos would be named oppmi.tif, and oppmo.eps.

1. REFERENCING A FIGURE OR TABLE WITHIN YOUR PAPER

When referencing your figures and tables within your paper, use the abbreviation “Fig.” even at the beginning of a sentence. Do not abbreviate “Table.” Tables should be numbered with Roman Numerals.

1. CHECKING YOUR FIGURES: THE IEEE GRAPHICS ANALYZER

The IEEE Graphics Analyzer enables authors to pre-screen their graphics for compliance with IEEE Access standards before submission. The online tool, located at <http://graphicsqc.ieee.org/>, allows authors to upload their graphics in order to check that each file is the correct file format, resolution, size and colorspace; that no fonts are missing or corrupt; that figures are not compiled in layers or have transparency, and that they are named according to the IEEE Access naming convention. At the end of this automated process, authors are provided with a detailed report on each graphic within the web applet, as well as by email.

For more information on using the Graphics Analyzer   
or any other graphics related topic, contact the IEEE Graphics Help Desk by e-mail at [graphics@ieee.org](mailto:graphics@ieee.org).

1. SUBMITTING YOUR GRAPHICS

Because IEEE will do the final formatting of your paper,

you do not need to position figures and tables at the top and bottom of each column. In fact, all figures, figure captions, and tables can be placed at the end of your paper. In addition to, or even in lieu of submitting figures within your final manuscript, figures should be submitted individually, separate from the manuscript in one of the file formats listed above in section VI-J. Place figure captions below the figures; place table titles above the tables. Please do not include captions as part of the figures, or put them in “text boxes” linked to the figures. Also, do not place borders around the outside of your figures.

1. COLOR PROCESSING / PRINTING IN IEEE JOURNALS

All IEEE Transactions, Journals, and Letters allow an author to publish color figures on IEEE Xplore® at no charge, and automatically convert them to grayscale for print versions. In most journals, figures and tables may alternatively be printed in color if an author chooses to do so. Please note that this service comes at an extra expense to the author. If you intend to have print color graphics, include a note with your final paper indicating which figures or tables you would like to be handled that way, and stating that you are willing to pay the additional fee.

CONCLUSION

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

APPENDIX

Appendixes, if needed, appear before the acknowledgment.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in American English is without an “e” after the “g.” Use the singular heading even if you have many acknowledgments. Avoid expressions such as “One of us (S.B.A.) would like to thank ... .” Instead, write “F. A. Author thanks ... .” In most cases, sponsor and financial support acknowledgments are placed in the unnumbered footnote on the first page, not here.

REFERENCES AND FOOTNOTES

1. REFERENCES

References need not be cited in text. When they are, they appear on the line, in square brackets, inside the punctuation. Multiple references are each numbered with separate brackets. When citing a section in a book, please give the relevant page numbers. In text, refer simply to the reference number. Do not use “Ref.” or “reference” except at the beginning of a sentence: “Reference [3] shows ... .” Please do not use automatic endnotes in Word, rather, type the reference list at the end of the paper using the “References” style.

Reference numbers are set flush left and form a column of their own, hanging out beyond the body of the reference. The reference numbers are on the line, enclosed in square brackets. In all references, the given name of the author or editor is abbreviated to the initial only and precedes the last name. Use them all; use et al. only if names are not given. Use commas around Jr., Sr., and III in names. Abbreviate conference titles. When citing IEEE transactions, provide the issue number, page range, volume number, year, and/or month if available. When referencing a patent, provide the day and the month of issue, or application. References may not include all information; please obtain and include relevant information. Do not combine references. There must be only one reference with each number. If there is a URL included with the print reference, it can be included at the end of the reference.

Other than books, capitalize only the first word in a paper title, except for proper nouns and element symbols. For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation See the end of this document for formats and examples of common references. For a complete discussion of references and their formats, see the IEEE style manual at www.ieee.org/authortools.

1. FOOTNOTES

Number footnotes separately in superscripts (Insert| Footnote).[[2]](#footnote-2) Place the actual footnote at the bottom of the column in which it is cited; do not put footnotes in the reference list (endnotes). Use letters for table footnotes (see Table I).

SUBMITTING YOUR PAPER FOR REVIEW

1. REVIEW STAGE USING WORD 6.0 OR HIGHER

If you want to submit your file with one column electronically, please do the following:

--First, click on the View menu and choose Print Layout.

--Second, place your cursor in the first paragraph. Go to the Format menu, choose Columns, choose one column Layout, and choose “apply to whole document” from the dropdown menu.

--Third, click and drag the right margin bar to just over 4 inches in width.

The graphics will stay in the “second” column, but you can drag them to the first column. Make the graphic wider to push out any text that may try to fill in next to the graphic.

1. FINAL STAGE USING WORD 6.0

When you submit your final version (after your paper has been accepted), print it in two-column format, including figures and tables. You must also send your final manuscript on a disk, via e-mail, or through a Web manuscript submission system as directed by the society contact. You may use Zip for large files, or compress files using Compress, Pkzip, Stuffit, or Gzip.

Also, send a sheet of paper or PDF with complete contact information for all authors. Include full mailing addresses, telephone numbers, fax numbers, and e-mail addresses. This information will be used to send each author a complimentary copy of the journal in which the paper appears. In addition, designate one author as the “corresponding author.” This is the author to whom proofs of the paper will be sent. Proofs are sent to the corresponding author only.

1. REVIEW STAGE USING SCHOLARONE® MANUSCRIPTS

Contributions to the Transactions, Journals, and Letters may be submitted electronically on IEEE’s on-line manuscript submission and peer-review system, ScholarOne® Manuscripts. You can get a listing of the publications that participate in ScholarOneat http://www.ieee.org/  
publications\_standards/publications/authors/authors\_submission.html First check if you have an existing account. If there is none, please create a new account. After logging in, go to your Author Center and click “Submit First Draft of a New Manuscript.”

Along with other information, you will be asked to select the subject from a pull-down list. Depending on the journal, there are various steps to the submission process; you must complete all steps for a complete submission. At the end of each step you must click “Save and Continue”; just uploading the paper is not sufficient. After the last step, you should see a confirmation that the submission is complete. You should also receive an e-mail confirmation. For inquiries regarding the submission of your paper on ScholarOne Manuscripts, please contact oprs-support@ieee.org or call +1 732 465 5861.

ScholarOne Manuscripts will accept files for review in various formats. Please check the guidelines of the specific journal for which you plan to submit.

You will be asked to file an electronic copyright form immediately upon completing the submission process (authors are responsible for obtaining any security clearances). Failure to submit the electronic copyright could result in publishing delays later. You will also have the opportunity to designate your article as “open access” if you agree to pay the IEEE open access fee.

1. FINAL STAGE USING SCHOLARONE MANUSCRIPTS

Upon acceptance, you will receive an email with specific instructions regarding the submission of your final files. To avoid any delays in publication, please be sure to follow these instructions. Most journals require that final submissions be uploaded through ScholarOne Manuscripts, although some may still accept final submissions via email. Final submissions should include source files of your accepted manuscript, high quality graphic files, and a formatted pdf file. If you have any questions regarding the final submission process, please contact the administrative contact for the journal.

In addition to this, upload a file with complete contact information for all authors. Include full mailing addresses, telephone numbers, fax numbers, and e-mail addresses. Designate the author who submitted the manuscript on ScholarOne Manuscripts as the “corresponding author.” This is the only author to whom proofs of the paper will be sent.

1. COPYRIGHT FORM

Authors must submit an electronic IEEE Copyright Form (eCF) upon submitting their final manuscript files. You can access the eCF system through your manuscript submission system or through the Author Gateway. You are responsible for obtaining any necessary approvals and/or security clearances. For additional information on intellectual property rights, visit the IEEE Intellectual Property Rights department web page at http://www.ieee.org/publications\_  
standards/publications/rights/index.html.

IEEE PUBLISHING POLICY

The general IEEE policy requires that authors should only submit original work that has neither appeared elsewhere for publication, nor is under review for another refereed publication. The submitting author must disclose all prior publication(s) and current submissions when submitting a manuscript. Do not publish “preliminary” data or results. The submitting author is responsible for obtaining agreement of all coauthors and any consent required from employers or sponsors before submitting an article. The IEEE Access

Department strongly discourages courtesy authorship; it is the obligation of the authors to cite only relevant prior work.

The IEEE Access Department does not publish conference records or proceedings, but can publish articles related to conferences that have undergone rigorous peer review. Minimally, two reviews are required for every article submitted for peer review.

PUBLICATION PRINCIPLES

The two types of contents of that are published are; 1) peer-reviewed and 2) archival. The Transactions and Journals Department publishes scholarly articles of archival value as well as tutorial expositions and critical reviews of classical subjects and topics of current interest.

Authors should consider the following points:

1. Technical papers submitted for publication must advance the state of knowledge and must cite relevant prior work.
2. The length of a submitted paper should be commensurate with the importance, or appropriate to the complexity, of the work. For example, an obvious extension of previously published work might not be appropriate for publication or might be adequately treated in just a few pages.
3. Authors must convince both peer reviewers and the editors of the scientific and technical merit of a paper; the standards of proof are higher when extraordinary or unexpected results are reported.
4. Because replication is required for scientific progress, papers submitted for publication must provide sufficient information to allow readers to perform similar experiments or calculations and use the reported results. Although not everything need be disclosed, a paper must contain new, useable, and fully described information. For example, a specimen’s chemical composition need not be reported if the main purpose of a paper is to introduce a new measurement technique. Authors should expect to be challenged by reviewers if the results are not supported by adequate data and critical details.
5. Papers that describe ongoing work or announce the latest technical achievement, which are suitable for presentation at a professional conference, may not be appropriate for publication.

REFERENCES

*Basic format for books:*

J. K. Author, “Title of chapter in the book,” in *Title of His Published Book, x*th ed. City of Publisher, (only U.S. State), Country: Abbrev. of Publisher, year, ch. *x*, sec. *x*, pp. *xxx–xxx.*

*Examples:*

1. G. O. Young, “Synthetic structure of industrial plastics,” in *Plastics,* 2nd ed., vol. 3, J. Peters, Ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15–64.
2. W.-K. Chen, *Linear Networks and Systems.* Belmont, CA, USA: Wadsworth, 1993, pp. 123–135.

*Basic format for periodicals:*

J. K. Author, “Name of paper,” *Abbrev. Title of Periodical*, vol. *x, no*. *x,* pp*. xxx-xxx,* Abbrev. Month, year, DOI. 10.1109.*XXX*.123456.

*Examples:*

1. J. U. Duncombe, “Infrared navigation—Part I: An assessment of feasibility,” *IEEE Trans. Electron Devices*, vol. ED-11, no. 1, pp. 34–39, Jan. 1959, 10.1109/TED.2016.2628402.
2. E. P. Wigner, “Theory of traveling-wave optical laser,”   
   *Phys. Rev*.,   
   vol. 134, pp. A635–A646, Dec. 1965.
3. E. H. Miller, “A note on reflector arrays,” *IEEE Trans. Antennas Propagat*., to be published.

*Basic format for reports:*

J. K. Author, “Title of report,” Abbrev. Name of Co., City of Co., Abbrev. State, Country, Rep. *xxx*, year.

*Examples:*

1. E. E. Reber, R. L. Michell, and C. J. Carter, “Oxygen absorption in the earth’s atmosphere,” Aerospace Corp., Los Angeles, CA, USA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1988.
2. J. H. Davis and J. R. Cogdell, “Calibration program for the 16-foot antenna,” Elect. Eng. Res. Lab., Univ. Texas, Austin, TX, USA, Tech. Memo. NGL-006-69-3, Nov. 15, 1987.

*Basic format for handbooks:*

*Name of Manual/Handbook, x* ed., Abbrev. Name of Co., City of Co., Abbrev. State, Country, year, pp. *xxx-xxx.*

*Examples:*

1. *Transmission Systems for Communications*, 3rd ed., Western Electric Co., Winston-Salem, NC, USA, 1985, pp. 44–60.
2. *Motorola Semiconductor Data Manual*, Motorola Semiconductor Products Inc., Phoenix, AZ, USA, 1989.

*Basic format for books (when available online):*

J. K. Author, “Title of chapter in the book,” in *Title of Published Book*, *x*th ed. City of Publisher, State, Country: Abbrev. of Publisher, year, ch.*x*, sec. *x*, pp. *xxx–xxx*. [Online]. Available: http://www.web.com

*Examples:*

1. G. O. Young, “Synthetic structure of industrial plastics,” in Plastics, vol. 3, Polymers of Hexadromicon, J. Peters, Ed., 2nd ed. New York, NY, USA: McGraw-Hill, 1964, pp. 15-64. [Online]. Available: http://www.bookref.com.
2. *The Founders’ Constitution*, Philip B. Kurland and Ralph Lerner, eds., Chicago, IL, USA: Univ. Chicago Press, 1987. [Online]. Available: http://press-pubs.uchicago.edu/founders/
3. The Terahertz Wave eBook. ZOmega Terahertz Corp., 2014. [Online]. Available: http://dl.z-thz.com/eBook/zomega\_ebook\_pdf\_1206\_sr.pdf. Accessed on: May 19, 2014.
4. Philip B. Kurland and Ralph Lerner, eds., *The Founders’ Constitution.* Chicago, IL, USA: Univ. of Chicago Press, 1987, Accessed on: Feb. 28, 2010, [Online] Available: http://press-pubs.uchicago.edu/founders/

*Basic format for journals (when available online):*

J. K. Author, “Name of paper,” *Abbrev. Title of Periodical*, vol. *x*, no. *x*, pp. *xxx-xxx*, Abbrev. Month, year. Accessed on: Month, Day, year, DOI: 10.1109.*XXX*.123456, [Online].

*Examples:*

1. J. S. Turner, “New directions in communications,” *IEEE J. Sel. Areas Commun*., vol. 13, no. 1, pp. 11-23, Jan. 1995.
2. W. P. Risk, G. S. Kino, and H. J. Shaw, “Fiber-optic frequency shifter using a surface acoustic wave incident at an oblique angle,” *Opt. Lett.*, vol. 11, no. 2, pp. 115–117, Feb. 1986.
3. P. Kopyt *et al., “*Electric properties of graphene-based conductive layers from DC up to terahertz range,” *IEEE THz Sci. Technol.,* to be published. DOI: 10.1109/TTHZ.2016.2544142.

*Basic format for papers presented at conferences (when available online):*

J.K. Author. (year, month). Title. presented at abbrev. conference title. [Type of Medium]. Available: site/path/file

*Example:*

1. PROCESS Corporation, Boston, MA, USA. Intranets: Internet technologies deployed behind the firewall for corporate productivity. Presented at INET96 Annual Meeting. [Online]. Available: http://home.process.com/Intranets/wp2.htp

*Basic format for reports and handbooks (when available online):*

J. K. Author. “Title of report,” Company. City, State, Country. Rep. no., (optional: vol./issue), Date. [Online] Available: site/path/file

*Examples:*

1. R. J. Hijmans and J. van Etten, “Raster: Geographic analysis and modeling with raster data,” R Package Version 2.0-12, Jan. 12, 2012. [Online]. Available: http://CRAN.R-project.org/package=raster
2. Teralyzer. Lytera UG, Kirchhain, Germany [Online]. Available: http://www.lytera.de/Terahertz\_THz\_Spectroscopy.php?id=home, Accessed on: Jun. 5, 2014

*Basic format for computer programs and electronic documents (when available online):*

Legislative body. Number of Congress, Session. (year, month day). *Number of bill or resolution*, *Title*. [Type of medium]. Available: site/path/file

***NOTE:*** ISO recommends that capitalization follow the accepted practice for the language or script in which the information is given.

*Example:*

1. U.S. House. 102nd Congress, 1st Session. (1991, Jan. 11). *H. Con. Res. 1, Sense of the Congress on Approval of Military Action*. [Online]. Available: LEXIS Library: GENFED File: BILLS

*Basic format for patents (when available online):*

Name of the invention, by inventor’s name. (year, month day). Patent Number[Type of medium]. Available: site/path/file

*Example:*

1. Musical toothbrush with mirror, by L.M.R. Brooks. (1992, May 19). Patent D 326 189

[Online]. Available: NEXIS Library: LEXPAT File: DES

*Basic format for conference proceedings (published):*

J. K. Author, “Title of paper,” in *Abbreviated Name of Conf.*, City of Conf., Abbrev. State (if given), Country, year, pp. *xxxxxx.*

*Example:*

1. D. B. Payne and J. R. Stern, “Wavelength-switched pas- sively coupled single-mode optical network,” in *Proc. IOOC-ECOC,* Boston, MA, USA,1985,   
   pp. 585–590.

*Example for papers presented at conferences (unpublished):*

1. D. Ebehard and E. Voges, “Digital single sideband detection for interferometric sensors,” presented at the *2nd Int. Conf. Optical Fiber Sensors,* Stuttgart, Germany, Jan. 2-5, 1984.

*Basic format for patents:*

J. K. Author, “Title of patent,” U.S. Patent *x xxx xxx*, Abbrev. Month, day, year.

*Example:*

1. G. Brandli and M. Dick, “Alternating current fed power supply,” U.S. Patent 4 084 217, Nov. 4, 1978.

*Basic format**for theses (M.S.) and dissertations (Ph.D.):*

a) J. K. Author, “Title of thesis,” M.S. thesis, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year.

b) J. K. Author, “Title of dissertation,” Ph.D. dissertation, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year.

*Examples:*

1. J. O. Williams, “Narrow-band analyzer,” Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, USA, 1993.
2. N. Kawasaki, “Parametric study of thermal and chemical nonequilibrium nozzle flow,” M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.

*Basic format for the most common types of unpublished references:*

a) J. K. Author, private communication, Abbrev. Month, year.

b) J. K. Author, “Title of paper,” unpublished.

c) J. K. Author, “Title of paper,” to be published.

*Examples:*

1. A. Harrison, private communication, May 1995.
2. B. Smith, “An approach to graphs of linear forms,” unpublished.
3. A. Brahms, “Representation error for real numbers in binary computer arithmetic,” IEEE Computer Group Repository, Paper R-67-85.

*Basic formats for standards:*

a) *Title of Standard*, Standard number, date.

b) *Title of Standard*, Standard number, Corporate author, location, date.

*Examples:*

1. IEEE Criteria for Class IE Electric Systems, IEEE Standard 308, 1969.
2. Letter Symbols for Quantities, ANSI Standard Y10.5-1968.

*Article number in reference examples:*

1. R. Fardel, M. Nagel, F. Nuesch, T. Lippert, and A. Wokaun, “Fabrication of organic light emitting diode pixels by laser-assisted forward transfer,” *Appl. Phys. Lett.*, vol. 91, no. 6, Aug. 2007, Art. no. 061103.
2. J. Zhang and N. Tansu, “Optical gain and laser characteristics of InGaN quantum wells on ternary InGaN substrates,” *IEEE Photon. J.*, vol. 5, no. 2, Apr. 2013, Art. no. 2600111.

*Example when using et al.:*

1. S. Azodolmolky *et al.*, Experimental demonstration of an impairment aware network planning and operation tool for transparent/translucent optical networks,” *J. Lightw. Technol.*, vol. 29, no. 4, pp. 439–448, Sep. 2011.

FIRST A. AUTHOR (M’76–SM’81–F’87) and all authors may include biographies. Biographies are often not included in conference-related papers. This author became a Member (M) of IEEE in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author’s educational background is listed. The degrees should be listed with type of degree in what field, which institution, city, state, and country, and year the degree was earned. The author’s major field of study should be lower-cased.

The second paragraph uses the pronoun of the person (he or she) and not the author’s last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (publisher name, year) similar to a reference. Current and previous research interests end the paragraph.

The third paragraph begins with the author’s title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter). List any memberships in professional societies other than the IEEE. Finally, list any awards and work for IEEE committees and publications. If a photograph is provided, it should be of good quality, and professional-looking. Following are two examples of an author’s biography.

**SECOND B. AUTHOR** was born in Greenwich Village, New York, NY, USA in 1977. He received the B.S. and M.S. degrees in aerospace engineering from the University of Virginia, Charlottesville, in 2001 and the Ph.D. degree in mechanical engineering from Drexel University, Philadelphia, PA, in 2008.

From 2001 to 2004, he was a Research Assistant with the Princeton Plasma Physics Laboratory. Since 2009, he has been an Assistant Professor with the Mechanical Engineering Department, Texas A&M University, College Station. He is the author of three books, more than 150 articles, and more than 70 inventions. His research interests include high-pressure and high-density nonthermal plasma discharge processes and applications, microscale plasma discharges, discharges in liquids, spectroscopic diagnostics, plasma propulsion, and innovation plasma applications. He is an Associate Editor of the journal *Earth*, *Moon*, *Planets*, and holds two patents.

Dr. Author was a recipient of the International Association of Geomagnetism and Aeronomy Young Scientist Award for Excellence in 2008, and the IEEE Electromagnetic Compatibility Society Best Symposium Paper Award in 2011.

**THIRD C. AUTHOR, JR.** (M’87) received the B.S. degree in mechanical engineering from National Chung Cheng University, Chiayi, Taiwan, in 2004 and the M.S. degree in mechanical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2006. He is currently pursuing the Ph.D. degree in mechanical engineering at Texas A&M University, College Station, TX, USA.

From 2008 to 2009, he was a Research Assistant with the Institute of Physics, Academia Sinica, Tapei, Taiwan. His research interest includes the development of surface processing and biological/medical treatment techniques using nonthermal atmospheric pressure plasmas, fundamental study of plasma sources, and fabrication of micro- or nanostructured surfaces.

Mr. Author’s awards and honors include the Frew Fellowship (Australian Academy of Science), the I. I. Rabi Prize (APS), the European Frequency and Time Forum Award, the Carl Zeiss Research Award, the William F. Meggers Award and the Adolph Lomb Medal (OSA).

1. https://github.com/tianchanghrbcn/AutoMLClustering [↑](#footnote-ref-1)
2. It is recommended that footnotes be avoided (except for the unnumbered footnote with the receipt date on the first page). Instead, try to integrate the footnote information into the text. [↑](#footnote-ref-2)