D. Lopresti and G. Nagy, Competition and Collaboration in Document Analysis and Recognition, to be presented at the Sixteenth International Conference on Document Analysis and Recognition (ICDAR 2021), September 2021, Lausanne, Switzerland..

Pre-publication (Springer LNCS?) version

Competition and Collaboration in Document Analysis and Recognition

Daniel Lopresti[1][0000-0003-2129-4223] and George Nagy[2][0000-0002-0521-1443]

[1] Lehigh University, Bethlehem PA 18015, USA
lopresti@cse.lehigh.edu
[2] Rensselaer Polytechnic Institute, Troy NY 12180, USA
nagy@ecse.rpi.edu

Abstract. Over the last twenty years, competitions aimed at showcasing research on various aspects of document analysis have become a significant part of our communal activities. After a quick look at competition in general and organized competitions in other domains, we focus on the organizers' reports of the 18 competitions completed in conjunction with the 14th Conference on Document Analysis and Research, ICDAR 2017. We provide descriptive statistics on the 130 organizers of these contests, their affiliations, the 450 participants, the platforms that underlie the evaluations, and the spectrum of specified tasks. We comment on the ~100 citations garnered by these contests over the intervening 3.5 years. Finally, in what we consider a logical sequel, we speculate on the possibility of an alternative model of small-scale, short-range communal research based on collaboration that seems to offer benefits competitions cannot capture.

Keywords: Contests, benchmarks, future of document analysis, research assessment, performance evaluation, technology transfer, reproducibility.

1 Introduction

We attempt to take a dispassionate look at the value and cost of competitive research in the ICDAR community. After a brief general discussion of the benefits and drawbacks of competitive research, we focus on the organizers' reports of the 18 completed competitions held in conjunction with ICDAR 2017, chosen because the intervening 3.5 years should allow sufficient time for a degree of impact to accrue. We present a snapshot of the topics, evaluation platforms, organizers, participants, and indications of the impact of these competitions. Then we take a ninety-degree turn and speculate on the possibility of a collaborative research in the format of small-scale imitations of the Research Priorities and Grand Challenges set in motion by various international and national organizations, and NGOs [1,2,3,4].

Competitions for "best" solutions to preset problems are popular in computer science, software engineering and mathematics, but rare in physics, chemistry and biology. In engineering they are largely confined to student team projects like concrete

canoes, matchstick bridges and solar cars. But the celebrated 2004-2005 DARPA driverless races across the desert were strictly for grownups,[1] as were the succeeding high-stake competitions for humanoid robots and satellite launches. Amateurs, like ham radio operators, birdwatchers and wild flower enthusiasts, compete more amicably at various scales.

Among the benefits claimed for contests are directing research to important unsolved problems, promoting best solutions, calling attention to roadblocks, rewarding successful researchers, and publicizing host organizations. Putative benefits also include generating benchmark data sets and developing common ground on performance metrics. Some of the benefits purportedly hinge on the reproducibility of the methods and results of a competition. In [6], we studied the 2016-2019 ICPR and ICDAR competitions from the perspective of reproducibility, and observed that while a few are doing a good job in this regard, most fall short in significant ways that might be easy to remedy if more attention was paid to certain desirable guidelines.

The costs are less widely advertised, and may indeed be only loss of time and diversion of attention. Common evaluation metrics may reduce diversity in evaluations and discourages otherwise promising approaches, particularly when multi-dimensional metrics are arbitrarily combined into a scalar value (e.g., the F-measure). Also, once standard measures are developed, less scrupulous individuals can find ways to "game" the system. Many competitions unintentionally enable skirting ethical standards by giving entrants lengthy access to test data, ostensibly to avoid the challenges of having to run submitted code developed under complex and perhaps hard-to-reproduce software environments. Repeated competition scenarios may prove counterproductive: how many iterations does it take for reported results become too specific or lose relevance? Witness what has happened with contests based on Highleyman's data [7], UW-1 [8] and MNIST [9], which may still provide some educational value for acclimating new students, but have long-since become uninteresting from a research perspective.

2 Prior Work

There is a large body of work on competition (covert and overt) in economics, psychology, anthropology and education. As mentioned above, organized competitions abound in every sphere. Some make headlines, especially those in athletics, political elections, film, television, music, book awards, and beauty contests. In some countries, chess and Go competitions draw popular attention. Closer to our sphere are mathematics, programming and robotics contests. We do not, however, know of any other competitions in data processing, which is what document analysis research is really about.[2] The huge number of input artifacts (billions on paper, plus born-digital text and images proliferating exponentially in cellphones and in the clouds), and the infinite number of possible outputs (analyses, transformations, transcriptions and interpretations), distinguish our

[1] DARPA specifically calls out the economic advantage of offering prizes instead of directly funding research [5].

[2] The well-known examples of AI techniques programmed to beat human experts in checkers, chess, Jeopardy, and Go fall in a different category from what we are considering here.

competitions from all others. Although several competition organizers, past and present, have written knowledgeably and thoughtfully about the benefits of competitions and the desirable aspects of training and test data, evaluation metrics and protocols, and submission platforms, we beg the reader's indulgence to exit this Prior Works section without any references.

3 Competitions in the DAR Community

3.1 Overview

From 2003 to 2019 more than 100 competitions have been organized in conjunction with ICDAR. Participants assembled themselves into teams and registered cost-free in response to announcements posted many months earlier. Registration was typically required for access to datasets and additional guidance via email. On occasion, competitions were cancelled when there is insufficient interest expressed by the community, as we have noted elsewhere [6]. We downloaded and studied the 6- to 9-page reports from the 18 competitions completed under the aegis of ICDAR 2017 [10]. These included:

- Arabic Text Detection and Recognition in Multi-resolution Video Frames
- Baseline Detection
- Classification of Medieval Handwritings in Latin Script
- Document Image Binarization
- Handwritten Text Recognition on the READ Dataset
- Historical Document Writer Identification
- Information Extraction in Historical Handwritten Records
- Layout Analysis for Challenging Medieval Manuscripts
- Multi-font and Multi-Size Digitally Represented Arabic Text
- Page Object Detection
- Post-OCR Text Correction
- Reading Chinese Text in the Wild
- Recognition of Documents with Complex Layouts
- Recognition of Early Indian Printed Documents
- Robust Reading Challenge on COCO-Text
- Robust Reading Challenge on Multi-lingual Scene Text Detection and Script Identification
- Robust Reading Challenge on Omnidirectional Video
- Robust Reading Challenge on Text Extraction from Biomedical Literature Figures

3.2 Organization

These competitions were organized by 119 members of the ICDAR community. Ten individuals served on three or more organizing committees, and 53 on at least two. They collectively represented 33 research laboratories or universities. The five most

active each provided eight or more organizers to the eighteen contests. Thirteen institutions were represented by only one organizer at a single contest. We are deliberately presenting only aggregate numbers, but it is fair to say that our community includes several active clusters of competition organizers. Beyond bringing together the community at a conference, organized competitions may also provide data that organizers find useful for other purposes.

3.3 Participants

Counting the number of competitors is more difficult, because five reports do not name or count them, and other reports give only the team sobriquet for some teams. We noted similar deficiencies in some reports we tallied in our earlier paper focusing on reproducibility [6]. Using the average membership (3.3) of the 77 teams for which we have complete counts, we estimate that the 2017 ICDAR competitions attracted 430 participants (including quite a few of the organizers of other competitions) representing 130 teams. Even if there some overlap between teams in distinct competitions, the number of competitors is comparable to ICDAR attendance (data presented at the welcome session indicates there were 386 registration for the main conference). The number of teams per competition ranged from 2 to 18, and the largest team had 9 members. As one might expect, participation is far less concentrated and more geographically diverse than administration.

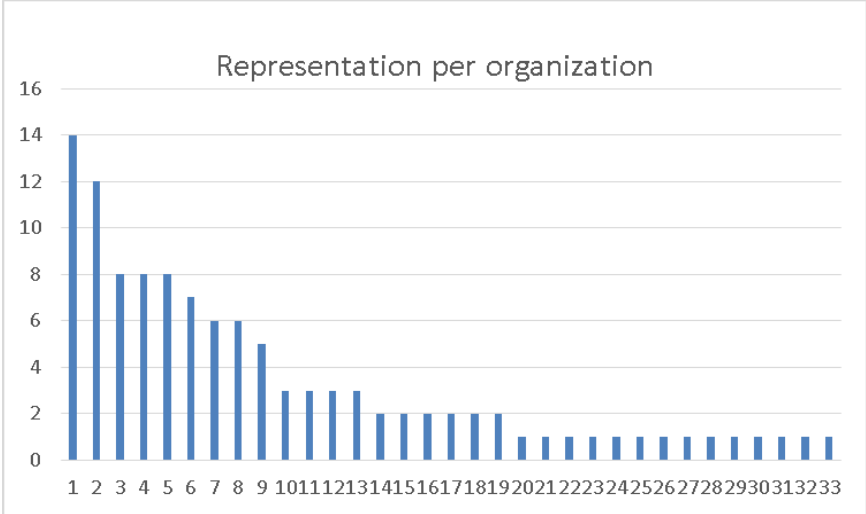

Figure 1. Representation per organization.

In terms of the distribution of organizers across contests, 89 people were involved in organizing one contest, 20 in organizing two contests, nine in organizing three contests, and one person was involved in organizing five contests.

Industrial participation was represented by teams from giants AliBaba, Google, Samsung and Tencent, and a dozen smaller and more specialized companies. The British Library and the Bibliothèque nationale de France contributed their multilingual expertise. Most of the participants were from Asian institutions in China, Vietnam, Japan and Korea. This appears largely consistent with the number of contributed papers at the main conference, which had nearly twice as many authors from China as the next highest country. A sparse sampling of the affiliations of prominent participants (and organizers) appears, roughly East-to-West, listed below.

School of ICT, Griffith University
University of Technology Sydney
National Laboratory of Pattern Recognition, Chinese Academy of Sciences
Tsinghua University, Beijing
CVPR unit, Indian Statistical Institute
National Center for Scientific Research Demokritos, Athens and Thrace
LATIS Lab, University of Sousse
National Engineering School of Sfax
Computer Vision Lab, TU Wien
DIVA group, University of Fribourg (Unifr) Fribourg
Computational Intelligence Technology Lab, University of Rostock
Paris Descartes University and Centre national de la recherche scientifique
L3i Laboratory, University of La Rochelle
University Rennes 2 and Insa Rennes,
Computer Vision Center, Universitat Autonoma de Barcelona
PRHLT research centre. Universitat Politecnica de Valencia
PRImA, University of Salford, Manchester
Brigham Young University, Provo,Utah

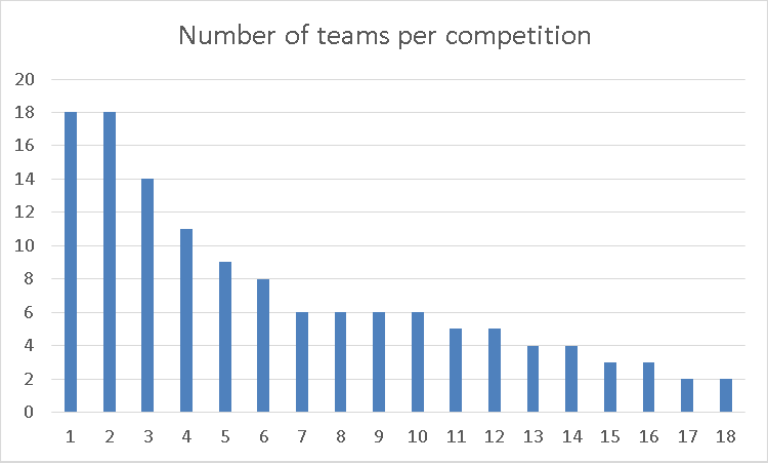

Figure 2. Numbers of participating teams (y-axis) for each competition (x-axis).

3.4 Experimental Data

Of perennial interest is the nature of the data sets used to train and test proposed systems. According to the reports, the emphasis is on text and images that present a good variety of potential recognition problems. The resulting "convenience samples" make for interesting competitions, but they are the antithesis of the random samples necessary to predict performance on a given population of documents (such as all 19th C conference proceedings in a national library, or an archival collection, or all Google books with a 19th C date of publication). Random sampling is, of course, the key to inference from observations of scientific experiments [11]. We saw, however, only a few instances of small-scale random sampling (e.g. random selection of 100 pages from 10 hand-picked books). Data augmentation, the addition of labeled synthetic documents to the training set, is sometimes used to make up for the paucity of real data.

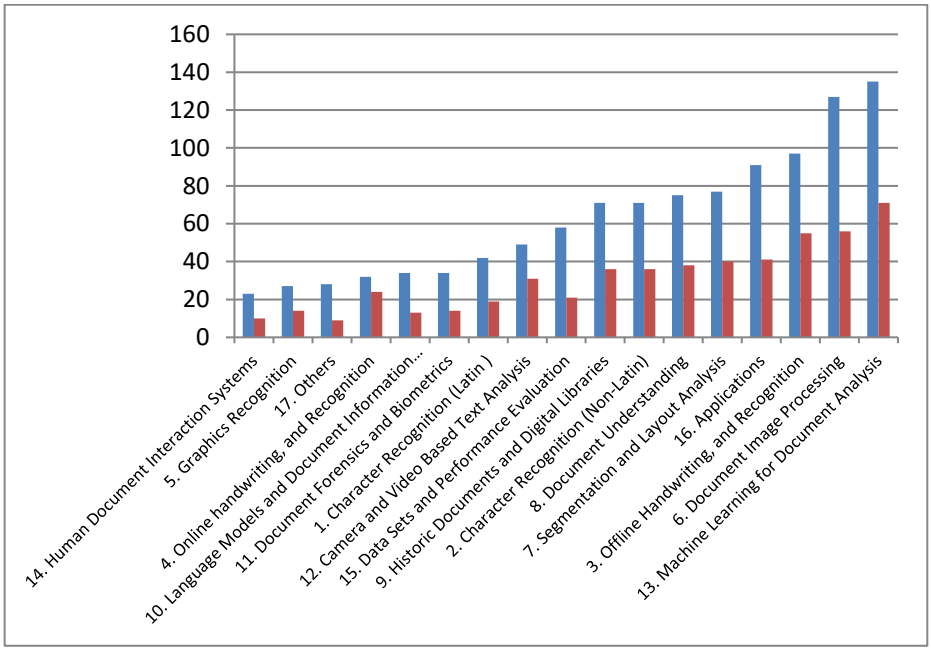

Figure 3. Topic areas for contributed papers at the main ICDAR 2017 conference.

Several reports mention the desirability of larger training samples, but random sampling of a well-defined population is hampered by selective digitization and transcription of source material, which is usually driven by other priorities. For example, librarians are likely to choose what they digitize based on a document's popularity or scholarly importance. Condition of the original source may also factor into consideration: material in good condition may be preferred because it is easier and cheaper to handle during the digitization process, or material in bad condition might be preferred because documents deteriorate over time to the point where the originals can no longer be han-

dled safely by scholars. Copyright issues may also be a factor. All of these considerations point to non-random selections made during the sampling process, which impacts the generalizability of the results.

Ten competitions featured contemporary data (including scene and video text), so we consider the remaining 8 as historical document processing (of printed, handwritten and illuminated manuscripts). Both the contemporary and the historical material spanned many languages and scripts: Arabic, Bengali, Chinese, Japanese, English, French, German and Latin. Only four of the databases were strictly English, but it is difficult to find any contemporary non-European text that does not contain any English phrases. The size of the databases varied, according to task, over almost three orders of magnitude, from under 100 pages to over 10,000 pages or cropped page images. There were contests based on cropped words, lines, and illustrations as well as on cropped page images. Four contests evaluated various tasks on scene text (from videos, cellphone images, web screenshots, and multiple cameras), and two contests tested methods on synthetic (machine-generated) text. Many contests subdivided the end-to-end pipeline, e.g., layout analysis, followed by character or word recognition. Preprocessing might involve digitization, and selecting and labeling training and validation data. There is a potential here, too, for selection bias. The variety of tasks devised by the competition organizers makes one ponder the current meaning of document.

As another representation of the research interests of the community, the distribution of submitted (blue) and accepted (red) papers for the main conference are shown in Figure 3, reproduced from slides used for the welcome session.

3.5 Submission and Evaluation Platforms

Systems used in several competitions include Alethea from the University of Salford [12], DIVA Services from the University of Fribourg [13], ScriptNet - READ (Recognition and Enrichment of Archival Documents) of the European Union's Horizon 2000 project [14]. and RRC Annotation and Evaluation from the University of Barcelona [15]. All of these systems, which were used for two-thirds of the ICDAR 2017 competitions, were developed and used for earlier competitions. The PAGE (Page Analysis and Ground-truth Elements) Format Framework [16], also developed at Salford, is occasionally used independently of *Alethea*.

The remaining competitions used individual custom platforms programmed in python, java, or other languages, with XML, CSV, or simple text ground-truth. Most had been developed for earlier competitions. Common metrics include accuracy, precision/relevance/ F-measure, edit distance, IoU (intersection over union), and various heuristic thresholds to rule out counterintuitive results. The granularity of the computation of averages is not always clearly specified.

3.6 Notes

None of the competitions addressed the question of what is to be done with residual errors and unclassified items. Some competitions, however, did suggest ways (evaluation profiles) to map their metrics into application-specific costs. It is difficult to see

how the current competition paradigms could measure the human labor cost of tuning and training an existing system to a new evaluation platform and new data.

4 Journal and Conference Publications

Each report typically has one page or so summarizing each method used by the participants in a paragraph or two. Some teams submit several entries with different methods, but no method is attributed to several teams. The summaries vary greatly, even within the same competition, in the level of detail. The reports almost never disclose the email addresses of the participants.

If competitions have an impact on research, then surely this will be reflected in subsequent publications. We ascertained the number of Google Scholar citations garnered by the eighteen contemporaneous contests: 174, or 9.7 per competition (including 25 citations by the organizers, usually in reports of subsequent competitions). Our search was limited to citations of entire contests. Since the reports themselves list very few papers by the contestants, their individual or team research results will probably take longer than a three-year latency to attract wide attention.[3]

It can be noted that our data suggests not all competitions are created equal. In 2017 there was one competition that has garnered 140+ citations by April 2021, and some ICDAR competitions have gathered 600+ and 800+ citations over the years.[4] The number of citations per 2017 ICDAR contest reported at the time of our original writing (January 2021) ranged from 0 to 42. Longer retrospection will certainly give larger and more stable counts, but at the cost of shedding light only on long past activities. The skewed distributions do indeed suggest competition among the competitions. They also raise the question whether peak counts or average counts are the better measure of the value of competitions for the ICDAR community.

Finding a place to present or advertise one's work in hundreds of journals and conferences also has its competitive aspects. Some are merely financial, as in for-profit journals and conferences where the main entry barrier is a page charge or registration fee. In others, editors and referees attempt to select the contributions likely to prove most attractive to their readers and participants. Experimental reports are routinely rejected unless they can demonstrate results superior to some other experimental reports (which may be one inducement for researchers to participate in competitions).

The competition is intensified by the automation of citation counts and their application to decisions, like academic promotions and grant awards, which were never intended by the founders of scientific communication. It is a commonplace that Albert

[3] Systems for automatically collecting and tabulating citation counts could prove informative for a comparative here, but as with all such measures should be taken with a grain of salt. For example, a quick perusal of the results of searching "ICDAR 2017" on Microsoft Academic shows a mix with more than half of citations on the first page of results (the highest counts) going to regular papers, with a few competition reports mixed in [17].

[4] We thank the anonymous Senior PC member who provided these observations in the meta-review for our ICDAR 2021 submission.

Einstein's h-index would not deserve attention without a huge adjustment for publication inflation.

We were curious about the influence of survey articles compared to the competitions. We did not find any relevant reviews published within a year of some of the competitions. Perhaps most of the competition tasks are too narrow to attract frequent review.

Citations of a competition report are just one measure of impact. The number of competitors going from the *n*th to the *n+1*st is another (impact as reflected by growing community interest). We also might consider the differential citation rate for published papers related to the entrants in competition versus similar published papers by non-competitors. If by competing in a contest a researcher gets a lot more citations compared to someone who does not compete, the payoff in terms of impact could be significant. In addition, future publications that leverage work done for a competition (the framework, data, evaluation measures) should count as a form of impact, even if the authors did not actually participate in the competition itself. Finally, it seems possible that selected results from competitions are reported in summary tables in later papers addressing the same problem and using the same data; such papers ought to cite the competition, but may instead choose to cite the scientific papers that describe the tested methods.

Still, questions remain when we ponder the impact of competitions on the trajectories of lines of research. What does it mean to suggest that one method dominates all others because it has won a particular contest? Are promising "losers" receiving due attention, or are they being shuffled off to the "scrap heap" of history, only to be rediscovered (hopefully) sometime in the future, a scenario that has played out before in pattern recognition research? Is the time and effort devoted toward developing many similarly performing methods, at the expense of leaving other territory unexplored, a good investment of the community's scarce resources?

5 Collaboration

We can speculate about a complementary model for advancing our field. What if optional research directions for our community were set each year, for overlapping two-year periods, by an IAPR committee, perhaps composed of representatives from TC-10 and TC-11? The committee (we might call this ADAR, for "Advancing Document Analysis Research") would consist of leaders in the field, such as journal editorial board members and those who have chaired or are chairing important conferences. They would select their own chairperson and maintain liaison with other relevant organizations, including funding agencies and professional societies in various countries. Their goal would be to identify a small set of agreed upon research objectives, much like what major scientific academies and funding agencies promote on a much grander scale and longer timelines [1,2,3,4].

The ADAR Committee would have two main tasks:

Task 1. Each year select five topics, suitable for experimental research, on the basis of interest level measured by recent submissions to the relevant conferences, workshops and journals. Topics may be repeated from year to year, until they reach the point of diminishing returns. We are aware that such a choice of topics is open to the objection of looking backwards rather than forwards, at the territory to be explored. We cannot, however, foresee the unforeseen, and must therefore be satisfied with innovative solutions to known problems. New problems will gather momentum through individual efforts and eventually rise to the attention of the ADAR Committee.

Task 2. Issue a Call for Participation (CfP) for each of the five topics each year. This document will give a concise description of the problem area, and add a few references to prior work and metrics. It will also set a date for the appointment of a steering committee for each topic. The steering committee will be selected from the applicants to the CfP for that topic, perhaps winnowed by some criterion for experience or a lower or higher bound for age. As a condition of appointment, each candidate will have to sign a public agreement to contribute to the Final Report due at the end of the two-year period from the appointment of the steering committee. The steering committee, in turn, will organize itself and all the participants to conduct research along the lines of the CfP and leading to the preparation of their Final Report.

The ADAR Committee will promptly submit the five yearly Calls for Participation, Lists of Participants, and Final Reports to the IAPR TC-10 and TC-11 leadership for review and posting on the respective websites, and will also request their publication in the proceedings of the next dominant DAR conference. Responsibility for the integrity of the research and the quality of the final report will rest solely with its mandatory signatories.

Credit (or blame) in the community will necessarily accrue to the ADAR Committee, the steering committees, and the participants. We believe that a good final report will be as creditable as lead authorship in a prestigious publication, and participation will be comparable to co-authorship. (Papers in experimental physics often have more than 100 authors.) A bad final report will be an albatross around the neck of its authors and participants.

Why five topics? We believe that, with a few dozen participants in each project, five is as many as can be managed by our community. Why two years? We expect that once appointed (say for overlapping five-year terms), the ADAR Committee would need about three months every year for a judicious choice of topics, and perhaps another two months for selecting the steering committees. The steering committee might need three months to set up protocols and initiate research, and three months at the end to analyze the experimental results and prepare the final report.

There is, of course, also the potential for meta-research in our proposal. The creation of new platforms for effective collaboration – for example, sharing and combining methods – would be significant contributions deserving of recognition. More attention

would be aimed at the human side of the equation: the time and effort needed to develop, test, field, and maintain document analysis systems, as well as to cope with the cases they still cannot handle. Reproducibility may also be facilitated since all of the participants on a project are nominally working together employing open lines of communication.

6 Conclusions

Our snapshot of the ICDAR 2017 competitions confirms that organizers devised a variety of "challenging" tasks, constructed versatile multi-use submission and evaluation platforms, defined useful metrics, located obscure sources of digitized, transcribed and annotated "documents" spanning many centuries, scripts and languages, and attracted capable participants from much of the world. The organizers filed conscientious reports the conference proceedings, though they differed in their emphases of different aspects of the contests and the levels of detail they disclosed varied widely.

Do the results give an accurate indication of the 2017 state of document analysis and recognition? After layout analysis and transcription, can we summarize magazine articles well enough to improve query-answer systems? Once we have located and identified all the relevant components of a technical article, can we construct an abstract more informative than the author's? Can finding and reading incidental text allow labeling photographs accurately enough to divide them into albums that make sense? Will automated analysis of old letters reveal the context of preceding and succeeding letters by the same author to the same destinataries? Will the analysis of ancient manuscripts allow confirming or contradicting current interpretation of historical events? Do the competitions point the way to the ultimate goals of DAR? What are these goals?

We saw that competition is ubiquitous and pervasive, and it surely has some merit. We listed its manifestations in the metadata generated by our research community. We tried to quantify the influence of organized research competitions on subsequent research, and compared it to the influence of journal and conference publications. We also proposed a collaborative model for experimental research different from the large-scale efforts organized by major funding agencies. We believe that organized competitions and collaborations can coexist, with some researchers more productive with one modus operandi, some with the other, and many preferring to work entirely on their own or in fluid, informal groupings. We now look forward to further joint ventures into uncharted DAR research territory.

7 Acknowledgements

We gratefully acknowledge the thoughtful feedback and suggestions from the anonymous reviewers as well as the cognizant Senior PC member, especially given the unconventional topic of our paper. We have incorporated several of their suggestions in this final version, and continue to ponder others. Paraphrasing one of the reviewers, our primary aim is indeed to reflect on new methods of interaction between researchers

within the DAR community, to help make the community more efficient, more dynamic, more visible, and ultimately more impactful. We also thank all those who have organized competitions at ICDAR and other conferences in our field: their contributions are more significant than the recognition they receive for such efforts.

References

1. National Science Foundation: Big Ideas. https://www.nsf.gov/news/special_reports/big_ideas/
2. United Nations; Goals. https://sdgs.un.org/goals
3. Computing Research Association; Visioning. https://cra.org/ccc/visioning/visioning-activities/2018-activities/artificial-intelligence-roadmap/
4. U.K.; AI Roadmap. https://www.gov.uk/government/publications/ai-roadmap
5. Defense Advanced Research Projects Agency; Prize Challenges. https://www.darpa.mil/work-with-us/public/prizes
6. Lopresti D. and Nagy G. Reproducibility: Evaluating the Evaluations. Third Workshop on Reproducible Research in Pattern Recognition (RRPR 2020), January 2021, Milan, Italy (virtual).
7. Highleyman W. "Data for Character Recognition Studies." IEEE Trans. Electron. Comput. 12 (1963): 135-136.
8. Liang J., Rogers R., Haralick R.M., and Phillips I., UW-ISL document image analysis toolbox: an experimental environment, Proceedings of the Fourth International Conference on Document Analysis and Recognition (ICDAR), Ulm, 1997.
9. LeCun Y., Cortes C., and Burges C.J.C., The MNIST Database. http://yann.lecun.com/exdb/mnist/
10. Competition reports in: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), IEEE 2017.
11. Wheelan C. Naked Statistics, W. W. Norton & Company, New York / London, 2013.
12. Clausner C., Pletschacher S., and Antonacopoulos A. Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments, Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR2011), Beijing, China, September 2011, pp. 48-52.
13. Würsch M., Ingold R., and Liwicki M. DIVAServices—A RESTful web service for Document Image Analysis methods, Digital Scholarship in the Humanities, Volume 32, Issue suppl_1, April 2017, Pages i150–i156, https://doi.org/10.1093/llc/fqw051
14. Diem M., Fiel S., and Kleber F., READ – Recognition and Enrichment of Archival Documents, https://readcoop.eu/wpontent/uploads/2017/01/READ_D5.8_ScriptNetDataset.pdf, 2016
15. Karatzas D., Gomez L., Nicolaou A., and Rusinol M. The Robust Reading Competition Annotation and Evaluation Platform, 2018 13th IAPR International Workshop on Document Analysis Systems.
16. Pletschacher S. and Antonacopoulos A. The PAGE (Page Analysis and Ground-truth Elements) Format Framework, Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010), Istanbul, Turkey, August 23-26, 2010, IEEE-CS Press, pp. 257-260.
17. Microsoft Academic search for "ICDAR 2017" on February 21, 2021. https://academic.microsoft.com/search?q=%20ICDAR%202017&f=&orderBy=0&skip=0&take=10