

# DI-2021: The Second Document Intelligence Workshop

Benjamin Han  
Azure AI  
Microsoft Corporation  
Redmond WA, USA  
dihha@microsoft.com

Yijuan Lu  
Azure AI  
Microsoft Corporation  
Redmond WA, USA  
yijlu@microsoft.com

Douglas Burdick  
Almaden Research Center  
IBM Research  
San Jose CA, USA  
drburdic@us.ibm.com

Hamid Motahari  
Computer Science  
Macquarie University  
Sydney, New South Wales,  
Australia  
hamidreza.motaharinezhad@mq.edu.au

Dave Lewis  
AI Research, Development, and  
Ethics  
Reveal-Brainspace  
Chicago Illinois, USA  
dlewis@revealdata.com

Sandeep Tata  
Strategic Technologies  
Google Research  
Mountain View CA, USA  
tata@google.com

## ABSTRACT

Business documents are central to the operation of all organizations, and they come in all shapes and sizes: project reports, planning documents, technical specifications, financial statements, meeting minutes, legal agreements, contracts, resumes, purchase orders, invoices, and many more. The ability to read, understand and interpret these documents, referred to here as *Document Intelligence* (DI), is challenging due to not only many domains of knowledge involved, but also their complex formats and structures, internal and external cross references deployed, and even less-than-ideal quality of scans and OCR oftentimes performed on them. This workshop aims to explore and advance the current state of research and practice in answering these challenges.

## KEYWORDS

Natural Language Processing; Natural Language Understanding; Computer Vision; Layout Understanding; Knowledge Representation and Reasoning; Data Mining; Knowledge Discovery; Information Retrieval.

### ACM Reference format:

Benjamin Han, Douglas Burdick, Dave Lewis, Yijuan Lu, Hamid Motahari, Sandeep Tata. 2021. DI-2021: The Second Document Intelligence Workshop. In *Proceedings of Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD2021), August 14-18, Singapore*. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3447548.3469454>

## 1 Introduction

While a variety of research has advanced the fundamentals of document understanding, the majority have focused on documents

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

found on the web which fail to capture the complexity of analysis and types of understanding needed across business documents. Realizing the vision of Document Intelligence remains a research challenge that requires a multi-disciplinary perspective spanning not only natural language processing and understanding, but also computer vision, layout understanding, knowledge representation and reasoning, data mining, knowledge discovery, information retrieval, and more – all of which have been profoundly impacted and advanced by deep learning in the last few years.

In addition to the invited talks and the panel discussion on topics related to Document Intelligence, the workshop program will include paper sessions which provides an opportunity to present peer-reviewed work on the topic related to Document Intelligence.

The intended audience of the workshop are researchers, practitioners, and students of any relevant discipline needed for document understanding.

## 2 Why KDD?

While many research areas relevant to DI are separately covered in other conferences such as ICML, NeurIPS, SIGIR, ACL, ICDAR, CVPR, AACL, etc, the interdisciplinary nature of the pursuit calls for a venue where all interested parties can freely exchange novel ideas, discuss open problems, and compare competing approaches. We believe KDD is the perfect venue for this discussion, and the fruit of the workshop can also benefit the core missions of the conference – to facilitate more effective data mining and knowledge discovery.

## 3 Workshop Organizers

**Benjamin Han (chair)** is the Principal Science Manager leading the research and development of the natural language services on **Microsoft Azure Cognitive Services**. His current focus is to democratize the state-of-the-art NLP research to serve customers at scale. His research interests include language detection, key phrase extraction, sentiment analysis, named entity recognition, entity linking, coreference resolution, relation extraction, knowledge base construction, summarization, and question

answering. During his time at Microsoft he has been a Principal Scientist in Satori (knowledge graph) and Bot Framework (conversational AI). Before that he was a Research Staff Member in the Multilingual NLP Technologies group at IBM TJ Watson Research Center for over a decade, working on all stages of information extraction technologies that power products such as IBM Watson Knowledge Studio and Watson NLU. He had participated in many government-organized projects/competitions such as TREC, RADAR, ACE, GALE and TACKBP, published in conferences such as ICME, ICoS, NAACL, IJCAI, AAAI and SIGIR, and organized the Knowledge Graph tutorial in KDD 2018.

**Douglas Burdick** is a Research Staff Member at **IBM Research - Almaden** currently working on the application of AI and machine learning to document understanding, which includes table extraction and understanding in addition to inferring document structure. His document understanding work is incorporated into the IBM Watson Compare & Comply and IBM Watson Discovery products. His other research focuses on the creation of financial knowledge graphs from unstructured data sources such as regulatory filings and analyst reports, which includes interpretation of tabular data from these documents. He has contributed to Apache SystemML and OpenII data integration toolkit, and co-organizes the DSMM workshop series (co-located with SIGMOD). He received his PhD in Computer Science from the University of Wisconsin - Madison.

**Dave Lewis** is an Executive Vice President for AI Research, Development, and Ethics at **Reveal-Brainspace**. Prior to joining Brainspace, he was variously a freelance consultant, corporate researcher (Bell Labs, AT&T Labs), research professor, and software company co-founder. Dave has published more than 40 peer-reviewed scientific publications and 9 patents. He was elected a Fellow of the American Association for Advancement of Science in 2006 for foundational work in text categorization, and won a Test of Time Award from ACM SIGIR in 2017 for his paper w/ Gale introducing uncertainty sampling.







**Yijuan (Lucy) Lu** is a Principal Scientist at **Microsoft Azure AI** where she worked on invoice understanding, OCR core engine, and video understanding in the recent two years. Prior to joining Microsoft, she was an associate professor in the Department of Computer Science at Texas State University. Her major publications appear in leading publication venues in multimedia and computer vision research. She was the First Place Winner in many challenging retrieval competitions in Eurographics for many years. She received 2015 Texas State Presidential Distinction Award and 2014 College Achievement Award. She also received the Best Paper award from ICME 2013 and ICIMCS 2012. She has obtained many competitive external grants from NSF, US Army, US Department of Defense and Texas Department of Transportation.

**Hamid Motahari** is an Honorary Professor of Computer Science at **Macquarie University**, Sydney, Australia. Prior to this, he was the Head of AI Science at the EY AI Lab in California where he

was leading a team of AI scientists in text and document understanding. Prior to EY, Hamid served as the Research Lead for AI & Cognitive Solutions at IBM Research, and has been a member of IBM Academy of Technology. He is a Senior Member of IEEE and has published 100+ scholarly papers in various conferences in AI, Web, IT Services, and IEEE/ACM journals. Hamid has chaired and organized various academic conferences and workshops in the past IEEE, ACM, AAAI and INFORMS conferences, including he has served as Technical Program Committee (TPC) Chair of the 1<sup>st</sup> Workshop on Document Intelligence at NeurIPS 2019.

**Sandeep Tata** is a Software Engineer at **Google Research** and leads a research group on information extraction. Sandeep has published dozens of peer-reviewed research articles across a variety of disciplines including data management, data mining, natural language processing, and information extraction. Sandeep's research work has impacted billions of people through research-focused enhancements to products like Google Drive, Gmail, and Google Assistant. He has served on the program committees for VLDB, ICDE, CIKM, and as a senior program committee member for KDD. He served on the organizing committee for WSDM 2016. Prior to Google Research, Sandeep was a Research Staff Member at IBM's Almaden Research Center. He has a PhD from the University of Michigan.

#### 4 Invited Speakers (Alphabetical Order)

	<a href="#">Kevyn Collins-Thompson</a> , Assoc. Prof., Information and Dept. of EECS, U. of Michigan		<a href="#">Don Metzler</a> is a Senior Staff Software Engineer at Google.
	<a href="#">Heng Ji</a> , Professor at Computer Science, UIUC		<a href="#">Benjamin Van Durme</a> , Associate Professor at Computer Science, Johns Hopkins Univ.
	<a href="#">Yunyao Li</a> , Senior Manager and Principal Research Staff Member at IBM Research		<a href="#">Cha Zhang</a> , Partner Engineering Manager at Microsoft Azure AI; IEEE Fellow

#### 5 Past Workshop (DI 2019)

[Document Intelligence Workshop 2019](#) was organized along [NeurIPS 2019](#) in Vancouver, BC, Canada, and it was a huge success. The workshop was well-attended, filling the room to the 150 capacity with people standing. The workshop received close to 50 initial abstract submissions and finally 38 paper submissions. Based on the reviews and discussions, [19 papers](#) were accepted resulting in a 50% acceptance rate. There was a Best Paper selected, and the discussion on the challenges and opportunities in the space has resulted in a paper [published in KDD Explorations \(PDF\)](#).