# Data-Driven Discovery: A New Era of Exploiting the Literature and Data

Ying Ding[1], Jevin West[2], Ming Song[3], Guiling Meng[4], Qi Yu[5]
[1]Indiana University, USA
[2]University of Washington, USA
[3]Yonsei University, South Korea
[4]Tongji University, China
[5]Shanxi Medical University, China

**Abstract**
In the current data-intensive era, the traditional hands-on method of conducting scientific research by exploring related publications to generate a testable hypothesis is well on its way of becoming obsolete within just a year or two. Analyzing the literature and data to automatically generate a hypothesis might become the de facto approach to inform the core research efforts of those trying to master the exponentially rapid expansion of publications and datasets. Here, viewpoints are provided and discussed to help the understanding and challenges of data-driven discovery. The purposes of this SIE are: 1) to showcase data-driven discovery research by highlighting several applications; 2) to discuss the challenges and potentials about data-driven discovery research, especially focusing on how information scientists and data scientists should work together; and 3) to create community of data-driven discovery in iSchool so as to facilitate collaboration and inspire innovation.
**Keywords:** information science, data science, data-driven discovery
**doi**: 10.9776/16518

**Contact**: dingying@indiana.edu

## 1 Purpose and Intended Audience

In this data-intensive era, the traditional method of exploring the related publications and available datasets from previous experiments to arrive at a testable hypothesis is becoming obsolete. Consider the fact that a new article is published every 30 seconds. For the very common disease of diabetes, there have been roughly 500,000 articles published to date; even if a scientist reads 20 papers per day, he will need 68 years to wade through all the material. The standard method simply cannot sufficiently deal with the large volume of documents or the exponential growth of datasets. A major threat is that the canon of domain knowledge cannot be consumed and held in human memory. Scouring the literature and data to generate a hypothesis might become the de facto approach to inform the core research efforts of those trying to master the exponentially rapid expansion of publications and datasets (Evans & Foster, 2011; Ding et al., 2013). In reality, most scholars have never been able to keep completely up-to-date with publications and datasets considering the unending increase in quantity and diversity of research within their own areas of focus, let alone in related conceptual areas in which knowledge may be segregated by syntactically impenetrable keyword barriers or an entirely different research corpus. Research communities in many disciplines are finally recognizing that with advances in information technology there needs to be new ways to extract entities from increasingly data-intensive publications and to integrate and analyze large scale datasets. This provides a compelling opportunity to improve the process of knowledge discovery from the literature and datasets through use of knowledge graphs and an associated framework that integrates scholars, domain knowledge, datasets, workflows, and machines on a scale previously beyond our reach.

The iConference is an ideal place to create community awareness about the challenges of data-driven discovery, especially given the special era about big data and how information science meets data science. The i-Conference theme of "Global Collaboration across the Information Community" highlights the importance of collaboration across different domains to achieve innovation. The proposed session builds upon the 1st international conference on data-driven discovery: when information science meets data science, June 20-21, 2016, Beijing, China. The purposes of this SIE are: 1) to showcase data-driven discovery research by highlighting several applications; 2) to discuss the challenges and potentials about data-driven discovery research, especially focusing on how information scientists and data scientists should work together; and 3) to create community of data-driven discovery in iSchool so as to facilitate collaboration and inspire innovation.

## 2 Proposed activities:

Using a lightning talk format, the session will be conducted in a 180-minute session as follows:

- Introduction to data-driven discovery (10 minutes)

- Five lightning talks followed by Q & A (25 minutes each; 125 minutes total)
  - Title: From measuring science to facilitating science (Jevin West)
    **Abstract**: Over the last several decades, there has been significant effort in developing metrics that measure scientific impact of authors, papers, and organizations. With the increasing appetite for data-driven decisions in hiring, tenure and promotion, funding, and evaluation combined with the advent of new, cleaner, bigger and more open bibliographic data, this metric development for measuring science will continue. However, in this presentation, I will talk about how these metrics can be used, not only in measuring science, but facilitating science. I will provide examples from my own research and other labs that are using these same metrics and algorithms to build better recommenders systems and search interfaces. In addition to improving scientific discovery, applying science metrics to information retrieval problems provides feedback to the metrics themselves creating a feedback for evaluation and further development.
    **Bio**: Jevin West is an Assistant Professor at the Information School at the University of Washington, co-founder of the DataLab and a Data Science Fellow at the eScience Institute. He received his Ph. D. in biology at the University of Washington and continued with a postdoc at the Department of Physics at Umea University in Sweden where he worked on the mapequation.org project in the Icelab. Jevin co-founded Eigenfactor.org– a research project that maps large scale citation networks in order to better understand, navigate, and evaluate the scientific literature. Using citation networks as a model system, he studies the evolution of scientific disciplines, economics of scholarly publishing, data mining of large corpora, and the sociology of science.

  - Title: Entitymetrics: Connecting Undiscovered Public Knowledge Dots (Min Song)
    **Abstract**: Knowledge is encoded as strings in a gigantic amount of unstructured scientific literature, which creates a huge hurdle for fast knowledge dissemination and transfer. Extracting knowledge from the unstructured full-text articles help uncover the implicit patterns and impacts of extracted concepts and entities. More innovative discovery can be facilitated in the entity networks where entities are implicitly connected by citation. To this end, we proposed entitymetrics that measures the impact of knowledge units on knowledge discovery (Ding, Song, et al., 2013). Knowledge entities embrace various concepts such as keywords, topics, subject categories, data sets, key methods, key theories, and domain entities (e.g., biological entities: genes, drugs, and diseases). These knowledge entities are used to facilitate knowledge discovery based on their ability to co-occur, cite/being cited, or co-cite/being co-cited. For example, entities connected via citation can discover potential association among genes which are not identified by mere gene co-occurrence (Song et al., 2014). In addition, the overlay of co-author networks with bio-entity co-occurrence networks can reveal entity-oriented scientific collaboration landscapes by revealing the implicit relationship lying in the authors' subject disciplines by considering the both bio-entities in citation sentences and proximity of them (Kim et al., 2016, Kang et al., under review). In addition to aforementioned studies related to entitymetrics, there are much more studies to be conducted. We assert that entitymetrics can shed a new light on metrics related research by taking knowledge entity as the research unit to enable knowledge discovery.
    **Bio**: Min Song is a Underwood Distinguished Professor at Yonsei University. Prior to Yonsei, Min was an Associate Professor of the Department of Information Systems at New Jersey Institute of Technology. He is an advisory board member of Open Information Science and associate editor of Frontiers in Library and Information Science. His research interests are in text mining, bioinformatics, and informetrics. He has published more than 150 journal and conference papers. Min has received several grants from NSF and IMLS in the United States and NRF in Korea. He received his PhD in Information Systems from Drexel University, an MS from Indiana University and a BA from Yonsei University in Korea.

  - Title: Applicability of Biological Information Extraction in Clinical Drug Mechanism Researches (Xueyuan Liu)
    **Abstract**: Dementia seriously affects human health. However, there still lacks appropriate interventions to prevent the progress of dementia. The development of new drugs not only requires a lot of costs, but also takes a long time with obscure application prospects. So this prompts us clinical scientists to look for new drugs from the traditional food industry. Curcumin provides strong epidemiological evidence for cognition improvement therapy. In the absence of any pre-trial basis, we searched the professional medical databases and found massive literature. The problem was how to find the possible effective pathogenic mechanisms and breakthrough points of curcumin in a short time. Within this context, we collaborated with information scientists, who used theoretical postulates and experimental knowledge to construct statistical models, so that predicted effective protein interactions in curcumin dementia therapy. At last, we clinical scientists validated the predictive protein interactions using experimental approaches. It is an efficient data-driven cooperation mode in biological information extraction in clinical mechanism researches.
    **Bio**: Xueyuan Liu, male, PhD / Professor, Chief of the Department of Neurology, Shanghai Tenth People's Hospital, Tongji University; Co-chairman of Shanghai Medical Neurology Association; Director of clinical treatment stroke center in Shanghai, Editorial board of Chinese Journal of Cerebrovascular Disease, Chinese Journal of clinical neuroscience, Journal of Tongji University, Journal of Neurology and neurological rehabilitation, Chinese Journal of clinical rehabilitation, Published nearly 60 articles in neurology field.

- Title: Drug repositioning based on Large Scale Drug-Induced Signatures (Qi Yu)
  **Abstract**: Drug repositioning (DR) refers to the identification of novel indications for existing drugs and is considered an effective route for drug development, because it reduces costs and bypasses safety concerns. However, discovering novel indications with DR is highly challenging, even with well-established high-throughput screening, because of the numerous combinations of both assays and drugs. Nowadays, a broad range of datasets has been utilized, such as sets related to chemical structure, drug-target relationship, and phenotypic information including drug side effects. In this presentation, we adopted an integrative approach for DR using the expression signature derived from the recent large-scale genomics dataset as well as chemical structure and target signatures. Next, we applied our method to infer DR candidate drugs for hepatitis B. The high-scoring candidate drugs were experimentally validated using cell lines and patient-derived primary cells.
  **Bio**: Dr. Qi Yu is an Associated Professor and currently associated director of Center for Healthcare Data Science at Shanxi Medical University in China. He has been award the title of "Young Academic Leader" by Shanxi Province Government. He has received several grants from NSFC and been involved in various Shanxi Province-funded projects. He has published nearly 40 papers in information science domain. His research interests are text mining, bioinformatics, and informetrics.

- Title: Using data and literature for drug discovery (Ying Ding).
  **Abstract**: With the ever-increasing power of computing technology and vast amount of available enriched unstructured and structured data, bibliometrics is moving from a simple counting of numbers, to utilizing topic and entity metrics for in-depth semantic measures of knowledge usage and diffusion, to finally enable knowledge discovery. This talk will showcase examples to demonstrate this paradigm shift, and presents exciting research outputs.
  **Bio**: Dr. Ying Ding is an Associate Professor at School of Informatics and Computing and is currently associate director of data science online program at Indiana University. She has been involved in various NIH, NSF and European-Union funded projects. She has published 190+ papers in journals, conferences, and workshops, and served as the program committee member for 180+ international conferences. She is the co-editor of book series called Semantic Web Synthesis by Morgan & Claypool publisher. She is co-author of the book "Intelligent Information Integration in B2B Electronic Commerce" published by Kluwer Academic Publishers, and co-author of the book chapter in "Spinning the Semantic Web" published by MIT Press. She is the co-editor in chief for Journal of Data and Information Science, and serves as the editorial board member for four ISI indexed journals in Information Science and Semantic Web. She is the co-founder of Data2Discovery company advancing cutting edge technologies in data science. Her current research interests include data-driven knowledge discovery, Semantic Web, knowledge graph, scientific collaboration, and the application of Web Technology.

- Group Discussion (45 minutes). We will divide the participants into several small groups to discuss the challenges and potentials for data-driven discovery, with the specific focuses on: research questions, available technologies, how to work with domain experts, potential collaboration on joint research including grant proposals and publications. We will conclude the session by 2 mins presentations from each groups.
  .

## 3 Relevance to the Conference/Significance to the Field:

In the new world of scholarly analytics, attention and extraction of deeply covered content and findings are the pathways to golden discoveries. Gradually, advances in information technologies, such as the advent of open access, Linked Open Data, semantic publishing, and open science, will make it possible to gather, annotate, and acquire related publications and other data sources and from those discover related content, findings, and conclusions. This could lead to discovery of unanticipated correlations and connections within an incredibly large and expanding research corpus. We are working on one of the oldest and toughest challenges associated with the combination of computer and human intelligence. The combinatorial innovation of human and machine intelligence will allow us to connect the dots for things that have been disconnected and accomplish through research what has been unimaginable, allowing us to dig the canal to connect data with knowledge. It is the right time to discuss the potentials of information scientists to work with data scientists, as this will bring the full potential of the both. iConference, as the top conference in information science, is the ideal place to create community awareness and inspire innovative researches and collaborations toward data-driven discovery.

## 4 Indicate the length of your event.

This SIE session will span 180 minutes.

## 5 References:

Ding, Y., Song, M., Han, J., Yu, Q., Yan, E., Lin, L., & Chambers, T. (2013). Entitymetrics: Measuring the impact of entities. *PLoS One*, 8(8): 1–14.
Evans, J. A., & Foster, J. G. (2011). Metaknowledge. *Science*, 332(6018), 721–725.