International Journal of Mechanical Design and Materials Science

Volume 1, Issue 4 2025

DATA-DRIVEN DESIGN-BY-ANALOGY: Current Status and Prospects

Devansh. Mechanical Engineering Department, Indian Institute of Technology, Kharagpur, 721302, India

Research Article

Date of Submission: 10-9-2025 Date of Acceptance: 19-10-2025 Date of Publication: 1-11-2025

ABSTRACT

Design-by-Analogy (DbA) is a design approach that helps designers reduce design fixation and improve design ideation results by generating new possibilities, solutions, or concepts in a target domain based on inspiration from a source domain. New prospects for creating data-driven techniques and tools for DbA assistance have recently been made possible by the expanding availability of design databases as well as the quickly evolving data science and artificial intelligence technology. In this work, we review previous data-driven DbA research and group individual studies into four categories—analogy encoding, retrieval, mapping, and evaluation—based on the data, techniques, and applications. This article clarifies the current status of data-driven DbA research and compares it with the cutting edge of data science and AI research to highlight intriguing research possibilities and directions for the field. It is based on both structured analysis and detailed organic evaluation. Lastly, we suggest a hypothetical data-driven DbA system for the future that incorporates all of the principles.

Keywords: Designing Using Data in Mind An analogy Artificial Intelligence and Design-by-Analogy Data Science

Introduction

Design-by-Analogy (DbA) is a design technique that uses cross-domain analogical reasoning to produce new solutions in a target domain based on inspiration from a source domain [1, 2, 3]. DbA is a field of current engineering design study, and a number of techniques and resources have been put out to facilitate its implementation [4, 5, 6, 7, 8]. DbA may assist designers reduce design fixation [9] and enhance design ideation results [10], according to studies. An example of DbA uses is shown in Fig. 1 [11]. The goal of this case study is to address an engineering design issue: how can the loud sonic boom produced by trains passing through tunnels at high speeds under atmospheric circumstances be mitigated [11, 12]. Engineers investigated structures that successfully "break" the sonic-boom effect in other design domains than trains or in the natural world as possible design solutions to this issue. Kingfisher birds can cut through the air and plunge into the water at very fast rates to grab food while scarcely producing a splash, according to research done by engineers. The train's front-end nose was redesigned by engineers to resemble the shape of a kingfisher's beak. Tunnel booms were removed and noise was decreased by this similar design. Additionally, it made it possible for the train to go 10% quicker.



Figure 1: Exemplar of DbA: high speed train nose design inspired by kingfisher's beak [11]

15% less energy was used. The example provides a real-world example of how design-by-analogy results in innovative issue resolution and significant performance gains. New data-driven techniques and tools that assist DbA have been made possible in recent years by the expanding availability of design databases as well as the quickly developing data science and artificial intelligence (AI) technology. Analogy representation, retrieval, mapping, and assessment procedures, for instance, may be supported by deep learning, knowledge graphs, computer vision, and natural language understanding [13, 14, 15, 16, 17, 18, 19]. To the best of our knowledge, the literature on data-driven DbA has not yet been subjected to a comprehensive review and organized analysis. Therefore, we evaluated the most advanced data-driven techniques and tools for DbA and compared them with the most recent data science and AI technologies in order to identify attractive research prospects and future directions. The order of this paper is as follows. The theoretical underpinnings of DbA are presented in Section 2, along with cognitive psychology models of analogical thinking and experimental knowledge of DbA's impact on engineering design originality. The research approach is explained in Section 3. A comprehensive organic evaluation of each data-driven DbA research literature is presented in Section 4. In order to show the overall state-of-the-art of the topic, Section 5 offers an organized and comprehensive study of all the literature from three perspectives: data, techniques, and applications. The trajectory of viable future research approaches from three perspectives and their interconnections is mapped out in Section 6. The article is finally concluded in section 7.

Background

According to cognitive psychology, humans often use analogical reasoning, or mapping answers from previous issues, to solve presented difficulties [20, 21, 22, 23, 24, 25]. According to Markman [21, 26], analogy is a potent cognitive technique that enables two items to be seen as comparable based on the interconnected networks of interactions between them. Critical thinking and logical reasoning may be enhanced by drawing analogous conclusions [22]. Gentner et al. [27] proposed a three-phase model of the analogical reasoning process, which includes retrieval, mapping, and assessment, and proposed that creating analogies is a crucial skill that sets humans apart from other animals. These stages were generally accepted by other studies, despite their somewhat differing subdivisions [22, 23, 24]. The first step in human analogical processes is typically to store and encode the source information (analogy candidates) in memory. Next, the source analogy is identified and searched using the target issue as a guide. Then, organized information may be comprehended, applied, or projected onto a less familiar or abstract area

using the original analogy as the foundational model [28, 29]. Three kinds of similarities—literal, attributional, and relational—are often used in the retrieval and mapping processes [30]. A successful comparison often contains an alignment of relationship structure, according to earlier research [31]. Following mapping, the last stage, assessment, is carried out, which comprises assessing the analogy and any conclusions that are drawn. Other subphases, such abstraction and parallel subprocess interaction, have also been identified as crucial cognitive processes in analogical reasoning in addition to the aforementioned stages [32].

Numerous computation-based models of analogy-making have been developed since the turn of the century [33, 34, 35, 36]. Based on the fundamental architecture of these models, French [37] divided them into three major categories. (1) In the realm of artificial intelligence, symbols, logic, search, planning, and means-ends analysis are all crucial components of the symbolic paradigm, which includes symbolic models. Evans' ANALOGY [33], one of the most well-known symbolic models, was created to perform proportional analogies of the pattern "A:B::C:?" derived from geometric figure exam questions. (2) The structure of complex networks, comprising nodes, edges, and their weights, is adopted by connectionist models. For instance, the LISA analogy engine [34] uses selective activation, dynamic binding, and partly distributed idea representations to link the relevant structures in order to facilitate both mapping and analogy retrieval. (3) Hybrid models, which usually include neurons that facilitate symbolic interpretation or interactions, have characteristics of both connectionist and symbolic models. The AMBR model [35], for example, is made up of several micro-agents, each of which represents a little bit of information that facilitates analogy-making.

Design-by-Analogy is the term used in engineering design to describe the process of using analogical reasoning to produce new design ideas or solutions. The parameters of product resemblance that designers often employ to identify the source product when using DbA in engineering design were examined by Gill et al. [38] and Tsoka et al. [39]. According to these empirical research, designers use at least six dimensions—working principle, structure, human interaction, function, energy flows, and material flows—to compare target and source goods. Among them, human contact, structure, and operating principle are more important in motivating analogy than the other characteristics. Researchers studying engineering design have also looked at how the DbA's solutions or concepts are affected by the analogical distance [10, 40, 41, 42, 43, 44, 45], representation of modality [42, 46, 47], commonness [42], amount or quantity [48], time of involvement of analogy [49], and type of similarity [50, 51]. The impact of the analogical distance between the source and target domains has drawn the most attention among these variables.

The complex impacts of the analogical distance on ideation habits and results have been better understood by recent studies. According to the conceptual leap theory, because of surface differences, distant parallels lead to innovations and breakthroughs that are less effective [52]. Numerous empirical findings from human-based research supported this claim. For instance, compared to near-field stimuli, far-field stimuli produce fewer but more original thoughts (Chan et al., 42). According to Srinivasan et al. [40], the amount and quality of freshly created thoughts drop as the analogical distance grows, but their uniqueness rises. Fu et al. [10] noted that a reasonable analogical distance may provide trade-offs and result in optimum design ideation performance within the various experimental and empirical situations. Although DbA is helpful, its effectiveness is inherently constrained by the designer's memory and past knowledge for possible recovery and mapping to a design challenge. For unskilled or specialist designers with a narrow field of expertise, this difficulty is more significant. However, not all designers are adept at using analogies. In actuality, these requirements imply the core principles of using intelligent algorithms to perform analogical reasoning and utilizing external knowledge databases as the source. The retrieval of design parallels from a large-scale design database and the analogical reasoning process in data-driven workflows may then be guided by the subtle experimental results on the analogical distance and other

features, as discussed in this section.

Research Methodology

This paper aims to elucidate the state-of-the-art of data-driven DbA research to date, especially when it comes to the application of AI and data science technologies. An advanced search in the Web-of-Science database was conducted to

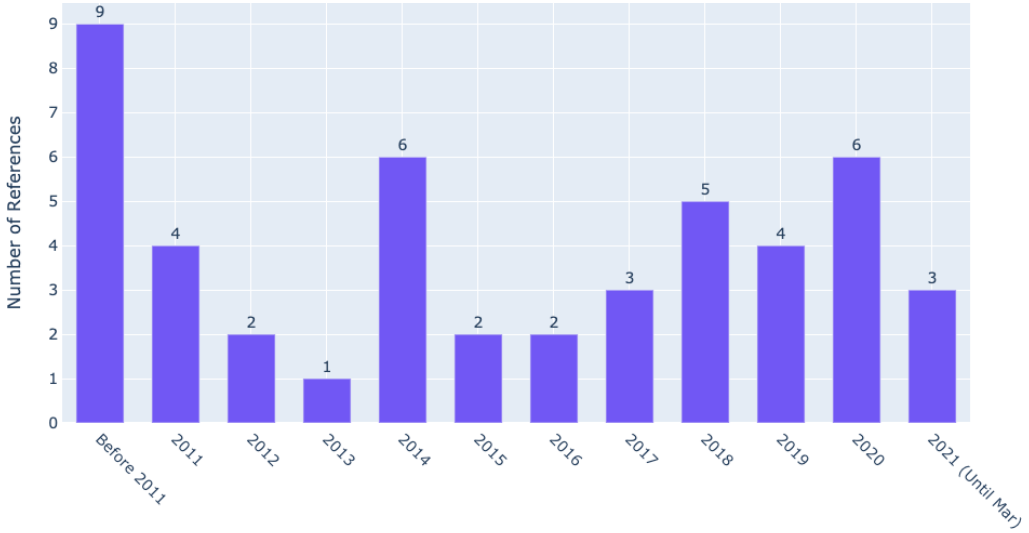


Figure 2: Number of data-driven DbA publications per year

generate an initial list of data-driven DbA publications[3]. With this search query *"TS = ((Design-by-Analogy) AND data) AND PY=1950-2021"*[4], we identified 23 relevant papers.
We manually examined the content of these publications and filtered out 4 articles that did not fall within the ambit of the engineering or innovation field and 6 articles that do not present data-driven methods or tools for DbA. Among the remaining ones, 5 of them are short version conference papers of the full version journal papers and are also excluded. Meanwhile, we found 39 additional relevant studies on data-driven DbA methods or tools by a snowballing search through the backward and forward citations of all the papers in the initial set (including those excluded ones in the previous step). For example, we were able to identify some recent efforts on data-driven DbA support tools, such as Idea-Inspire 4.0 [53], TechNet [54], and B-link [55], from the forward citations of a paper titled *"Function-based **design-by-analogy**: a functional vector approach to analogical search"* [5]. From the backward citation of another paper titled *"Identifying candidates for **design-by-analogy**"* [30], we could identify the first generation of DbA systems, such as KRITIK [56] and IDEAL [57]. Such newly identified studies directly refer to design-by-analogy or its synonym in the body text or discuss its usage to support one or more DbA subprocesses.
After these steps, the final literature set includes 47 papers[5]. As observed in Fig. 2, the number of publications exhibited an upward trend in recent years, indicating a growing interest in this field. Fig. 3 shows the proportional breakdown of publications by types, including journal articles, conference papers, book chapters, and patents. In Fig. 3, we marked out four journals and two conferences that include more than one papers in the final set:
Journals - (1) Journal of Mechanical Design, (2) Artificial Intelligence for Engineering Design, Analysis

and Manufac-turing, (3) Research in Engineering Design, and (4) Design Science.
Conferences - (1) International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE) held by ASME and (2) Conference on Human Factors in Computing Systems (CHI) held by ACM.

Data-driven DbA methods and tools: an organic review
The first generation of DbA systems, most of which were knowledge-based expert systems, were developed before the 2000s. Huhns and Acosta [58] developed the **ARGO**, which represents problem-solving plans using rule-based graphs that enable analogical retrieval for solving new problems. Goel et al. [56] proposed KRITIK, which represents design cases through function-behavior-structure (FBS) models [59]. KRITIK retrieves specific designs based on the

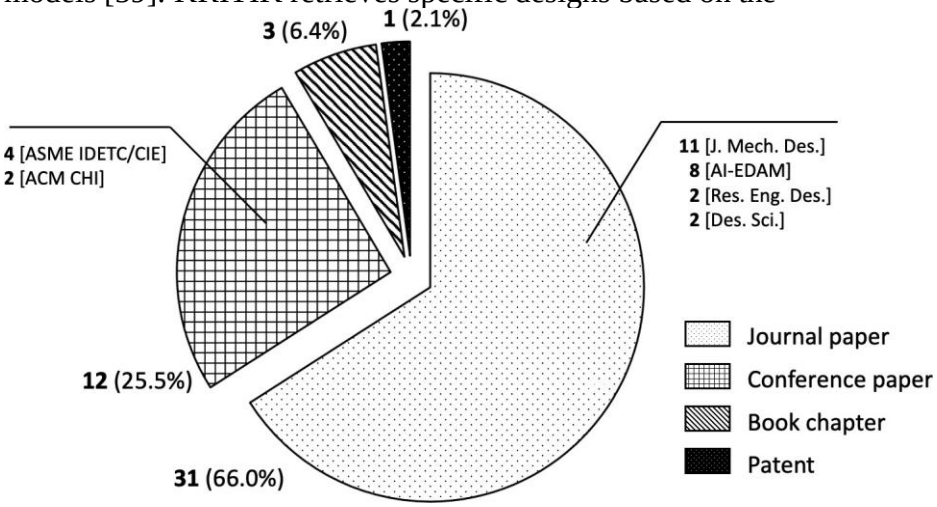


Figure 3: Distribution of data-driven DbA publications by literature types

functional details of the query design, makes recommendations for changes, and enables the validation of the freshly created designs. The IDEAL [57], which is also based on the FBS models, was created by the same research team. By contrasting two designs with distinct patterns, the IDEAL system accomplishes analogical mapping and enables the extraction of general teleological processes.
In order to support DbA, a number of data-driven techniques and tools have been developed over the last 20 years. Reich and Shai [60] proposed the infused design method, which represents design problems at the meta-level to facilitate the discovery and application of knowledge across various technological domains. The interdisciplinary engineering knowledge genome (IEKG) was created based on the infused design idea to facilitate the search for general knowledge and technique structures in the interdisciplinary domains [60]. In order to automatically and systematically retrieve candidate products for DbA, Verhaegen et al. [30] distilled product features using principal component analysis (PCA) and word co-occurrence analysis. Linsey et al. [4] proposed the WordTree method to linguistically re-represent design problems based on a predefined large-scale semantic network (WordNet), which can assist designers in finding potential analogies for idea generation. In order to represent design documents into vectors based on the word frequency of functional verbs from the Functional Basis model [61], Murphy et al. [5] presented a functional vector technique. Additional research has shown that the functional analogy search approach may provide functionally innovative notions [50].
Sanaei et al. [62] created an analogical retrieval technique that extracts relational structure from the text

using text mining and an ensemble model based on recurrent neural networks (RNNs). Using the information extraction approach and bidirectional RNN deep learning model, Gilon et al. [63] suggested an analogical search engine to find distant but relevant design stimuli for particular demands. A mixed initiative system called SOLVENT was created by Chan et al. [64] to find similarities between scientific publications. The system uses the Doc2Vec pre-trained word embeddings to represent and map various contents while annotating four elements of research papers: background, goal, mechanism, and results. The retriever and combinator are two of the computer-based tools that Han et al. [18, 17] created to assist DbA. In order to create the foundation ontology for a particular design challenge and look for less well-known target ontologies for design ideation, the retriever uses ConceptNet as its knowledge base [18]. Using a bespoke design knowledge database built using ConceptNet via web crawling and text mining, the combinator creates combinational prompts in both text and picture formats by mixing concepts that are tangentially related [17].

The patent database has garnered a lot of academic interest among the many kinds of engineering design-related data sources due to its scale, variety, and abundance of design-related information in its documents. By examining abstract patterns in the worldwide patent database, Altshuller [65] derived a number of heuristic design rules, or principles, and created the TRIZ problem-solving framework to help engineers get beyond analogical reasoning dead ends. The usage of TRIZ has now been supported by a number of data-driven technologies. Cascini and Russo [66] created a system called PAT-ANALYZER that mines patent textual data to automatically create design conflicts in TRIZ. Bio-TRIZ, which builds on the biological data and concepts in the TRIZ system, was suggested by Vincent et al. [67]. In addition to TRIZ-related tools, McCaffrey and Spector [68] created Analogy Finder, a

By rephrasing the design issue description as verbs and their synonyms, the DbA assistance system finds applicable parallels in the patent database. A computer-aided ideation tool called InnoGPS was created by Luo et al. [15] using the whole technological space established in the USPTO database. Based on knowledge distance, the InnoGPS system may help designers find and recover ideas and information to make cross-domain parallels. In a similar vein, Song et al. [41] used the technical network based on the clustering partition findings to find recognized patent parallels in home-, near-, and far-field stimuli. Recently, researchers studying engineering design have created semantic networks from databases related to engineering design using the most recent natural language processing (NLP) and text mining techniques. Examples of these databases include B-link [55], which is based on about one million engineering papers, and TechNet [54], which is based on six million patent documents. TechNet and B-Link have both been used to idea development [70, 71], appraisal [19], and analogical retrieval [55, 69]. Additionally, textual design databases have been organized and visualized for research and inspiration using a variety of subject modeling methods. In order to map patent papers in a network structure, Fu et al. [72] used a Bayesian model with latent semantic analysis; this technique offers a foundation for automated retrieval of cross-domain inspiring parallels. The route length in a Bayesian network represents the analogous distance between the original design challenge and possible inputs. A topic modeling approach called hierarchical nonnegative matrix factorization (NMF) was used by Song et al. [6] to structurally describe a patent database with three facets: behavior, material, and component. They created VISION [7], a visual interactive tool for data exploration and analogical inspiration, based on this clustered structure of patent data.

A kind of DbA known as "bio-inspired design" (BID) [73, 74, 75, 76] draws inspiration for engineering design from natural biological phenomena. The development of data-driven techniques, tools, and databases for BID has been a major focus in recent years. By extracting relevant keywords from biological literature that matched the engineering functional terms in the functional basis, Cheong et al. [77] and Nagel et al. [78] created the engineering-to-biology thesaurus. When tackling design challenges, engineers use the thesaurus as a foundation to discover biological similarities and the functional rationale connecting

two domains. The Design by Analogy to Nature Engine (DANE), a design case library of over 40 FBS models of natural and artificial systems, was created by Goel et al. [79]. Additionally, DANE offers a framework that allows users to access prior cases based on various types of queries. In order to manually express design challenges in four aspects (functions, environments, requirements, and performance) based on FBS models and assess analogies using the heuristic T-chart in the BID, Helms and Goel [80] introduced the four-box technique. The computational technique based on syntactic rules was presented by Cheong and Shu [81] to find biological parallels including causal links. In order to ascertain if a textual stimulus is beneficial for resolving the given design issue, Glier et al. [82] investigated the use of automated text classifiers for BID, such as naive Bayes classifiers, k-nearest neighbors, and support vector machines. In order to provide engineering designers biological comparisons, the Biomimicry 3.8 Institute [83] created the web-based application AskNature, which uses a specific taxonomy to handle the more than 1,600 biological instances in its collection.

Verhaegen et al. [84] created product aspects in DbA (PAnDA), a program that uses rule-based text-mining algorithms to extract product characteristics from texts in order to find potential items for DbA. Using the same methodology, they also created SEABIRD [85] to map biological systems in scholarly articles and technological systems in patent filings. Designers may search for biological stimuli in a scalable manner thanks to SEABIRD. The D-APPS tool and DRACULA program, which allow analogous matching between biological and engineering ideas and analogy searching for BID, were developed by Lucero et al. [86, 87] by integrating the functional basis model, WordNet, and the AskNature repository. The Idea-Inspire 4.0 tool, created by Siddharth and Chakrabarti [53], is based on the domain-agnostic SAPPhIRE model ontology [88] and reflects both biological and engineering ideas. Idea-Inspire 4.0 facilitates analogical retrieval and assessment using a manually constructed database of 83 artificial ideas and 60 biological concepts. In order to provide engineers with a high-level filter, Kruiper et al. [16, 89] created a dataset called focused open biology information extraction (FOBIE) and used it to train an NLP model to extract semi-open trade-off relations and arguments from scientific biological literature. In order to extract biological adaptations and the corresponding product architecture from the current bioinspired design abstraction database and product architecture representation database, Bhasin et al. [90] developed a reduced function-means tree.

Recent study has investigated visual or multimodal analogies for design inspiration, even if the majority of the aforementioned techniques and tools included processing textual input. In order to uncover visual comparisons for further idea development, Kwon et al. [91] created a framework to take use of visual similarity based on the web-based search engine. In order to automatically determine the vector space and create the feature vectors to describe patent photos, Jiang et al. [13] suggested a technique based on convolutional neural networks (CNNs). Retrieving visual stimuli based on vector distances may be made easier by the resulting feature vectors that include both visual information and technology-related knowledge. In order to recover the underlying form properties of drawings in the Quickdraw database, Zhang et al. [92] introduced an unsupervised deep learning model called Sketch-pix2seq. This model is trained using reconstruction and clustering losses.

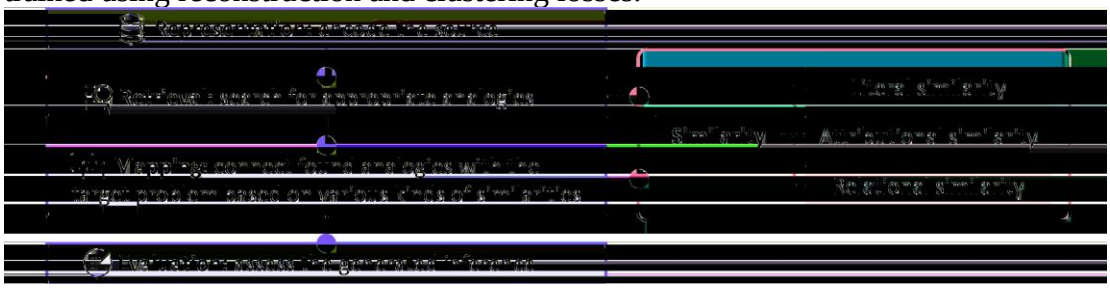


Figure 4: The four-phase model of making analogy in cognitive science

The constructed latent space for sketches provides a new way to define visual similarities and search the analogical sketches for DbA.

Structure and state of the field: a synthesis

Overview

In section 4, we have presented a nuanced organic review of each data-driven DbA method or tool identified via our literature search. To elucidate the general state-of-the-art of the field, this section aims to provide an integrated and structured analysis of all works in the literature together from the viewpoint of the following three aspects:

The databases, datasets, or repositories used for developing relevant methods or tools;

The application goals, which usually belong to one or more subphases of the DbA process, based on a four-phase analogy model in the cognitive science field (depicted in Fig. 4).

The AI or data science algorithms used in the data-driven DbA studies.

Fig. 5 presents a visual summary of the datasets, methods, applications of existing DbA studies, as well as the coupling of the three dimensions. A single DbA method or tool may contain more than one DbA subphase. We separately counted each item in Fig. 5, ensuring that each used algorithm or method was matched to the corresponding applications. The different line colors denote different phases in the DbA process. We further developed a web-based interactive data visualization[6] for public users to probe such coupling relationships among the three perspectives in individual prior studies.

Data

Table 1 summarizes the datasets used in DbA, including databases specifically developed for DbA research. The domain-specific knowledge databases, including DANE and Idea-Inspire, were manually created based on human-coding structures. The representation of design and relevant concepts by using predefined models is a time-consuming and labor-intensive task and requires domain expertise. Consequently, these repositories are usually small scale and lack scalability. For example, in the DANE repository, building a complete FBS model of a biological system requires 40–100 hours [79]; thus, the repository contains only 40 cases and 22 complete FBS models. These models lack a uniform description and generalizability to enable processing large-scale databases, which are pre-conditions for adopting modern AI algorithms. While scholars have proposed building large and scalable digital libraries of analogical design case studies for broader usages [93], how to represent, aggregate and organize existing scattered and unstructured design data efficiently with a consistent structure remains an open question.

Compared with human-curated databases, patent and scientific paper databases are naturally structured and continually populated over time and cover a wide range of domains [54, 55]. They are public big-data sources that enable the development of scalable data-driven DbA methods and tools. Specifically, patent documents contain rich engineering design information of the structures, functions, mechanisms, and principles that might be useful for the design inspiration process [5, 13, 15, 40, 72, 94].

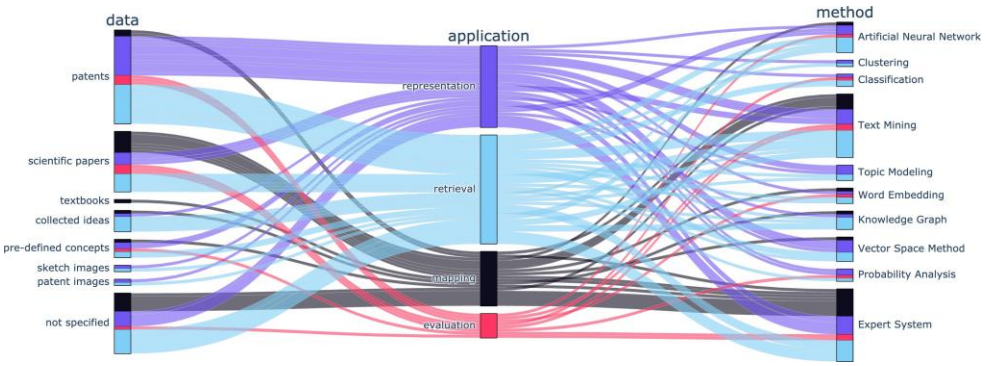


Figure 5: Overview of the state-of-the-art for data-driven DbA studies

Table 1: Existing databases that have been adopted or developed for DbA

Category	Data	Size and refs
Patents	Texts from the whole USPTO database	More than 6 million patent documents [15, 54]
	Texts from the subset of the USPTO database	155,000[30, 84, 85], 6,500 [5], 10,400 [7], 200 [41], 100 [72] patent texts
	Images from the subset of the USPTO database	Images from the subset of the USPTO database [13]
Scientific paper/books	Papers from computer science conferences	Papers from computer science conferences
	Papers from biological journals	500 [82], 8,000 [85], 10,000 [16] papers
	Papers from engineering journals	More than 1 million papers [55]
	A biological textbook	A biological textbook [77]
Collected ideas	Concepts collected from Quirky platform	The corpus cover over 8,500 products [63]
	Data crawled from the Internet	Unknown [17, 91]
Sketches	Quickdraw	75,000 sketches [92]
Biomimicry databases	AskNature	Over 1,700 biological concepts [83]
	DANE	40 FBS-based biological systems [79]
	Idea-Inspire	Over 100 biological and engineering SAPPhIRE-based databases
Semantic Network-based		

based concepts		annotated sentences with 4,700 rela-tions [16]
[53]	WordNet	155,000 entities with 648,000 relations [95]
FOBIE Over 1,500	ConceptNet	517,000 entities with 1.3×10^{11} relations [96]
	TechNet	Over 4 million technical terms with their relation-ships [54]

Recently, extensive research has established that crowdsourced design solution data collected from online ideation and task-orientated platforms (e.g., InnoCentive, Quirky, OpenIDEO, and Amazon Mechanical Turk) are useful for DbA research [97, 98, 99]. The crowdsourced solutions make it possible to mine the real-world ideation outputs of complex cognitive tasks generated by thousands of people across the world. In addition, some researchers have observed that crowdfunding platforms (e.g., Kickstarter and Indiegogo) also accumulate product design projects and contain rich design information about product designs, and are therefore potentially useful data sources for analogical design stimuli [100].

Majority of the existing DbA research only focused on textual data and ignored data in other modalities such as visual information. Recently, some scholars have explored the mining of the large-scale 2D design image datasets (such as patent images and sketches) using various computer vision techniques. Other 2D image datasets and 3D model datasets, such as Shape-Net [101], the Thingiverse repository[7], the IKEA design dataset [102], and mechanical CAD model dataset [103], can be explored and utilized to develop tools for supporting DbA with visual analogies or multimodal analogies.

Furthermore, there is a paucity of gold-standard benchmarking datasets developed for specific DbA tasks. Consequently, the comparison and evaluation of different methods and tools in terms of their utilities for the same task have been plagued by ambiguities. Researchers of this community are recommended to collaborate on creating large-scale and canonically acceptable benchmark evaluation datasets for various situations and tasks in the DbA process. For example, the recently released FOBIE dataset [16] was developed to support the semi-open relationship extraction for nature-inspired engineering. Such a benchmarking dataset enables the measurement of the effectiveness of novel algorithms against those of other state-of-the-art approaches.

Applications

Based on the four-phase model of making analogies in cognitive science, we assessed the applications of the DbA tools and methods, and categorized each of them into one or more phases accordingly. From Table 2, it can be distinctly observed that most of them fall within the analogy representation and retrieval subphases, which are the foundations of the entire DbA process. In most studies on these two subphases, text mining-related techniques were adopted to represent and identify candidate analogies for further use. Some researchers have utilized modern deep learning-based AI techniques to construct high-dimensional latent space for design data representation in different forms. These researchers generally posit that, guided by the identified stimuli, human users can continue the mapping and inference process. Although many AI-based methods and tools can recommend the candidate analogy list for users, few can interpret the results such as identifying the type of similarities.

Some studies sought to automate the analogical mapping process, including some systems that predate the 2000s, most of which are rule-based expert systems, as shown in Fig. 5. Some pre-defined rules, such as syntactic grammar-based rules [81], are embedded within these tools to facilitate automated inference. In addition, some tools implement idea generation by simply combining two ideas within moderate distance for design ideation [17]. Overall, current analogy mapping methods and tools have employed little machine intelligence, which is the major advantage of modern AI techniques. For example, recent deep learning AI models, such as the Siamese network and its variations [104], enable machines to learn pair-wise entity to entity embeddings that encode the specific type of relational information. If we collect a large set of analogy pairs, we may leverage these models to learn the analogical relational representation and conduct the analogical reasoning process.

Regarding the fourth phase, only a few researchers have attempted to evaluate the analogies or ideas generated by the DbA. This may be attributed to, at least, two reasons. First, it is difficult to establish the

evaluation criteria. The usefulness and novelty that define "creativeness" may vary according to specific personal needs, organizational contexts, regional cultures, and times [105]. Second, analogy assessment is a domain-related process. A traditional idea or concept in one knowledge domain might be novel in another. Therefore, the space that is newly transformed through analogy reasoning may require novel types of evaluation. A powerful evaluation mechanism would determine the effectiveness and usefulness of new data-driven DbA systems. There is still much room for researchers to develop robust, interpretable, and convincing analogy evaluation methods.

Table 2: The applications of DbA tools and methods[8]

Tools or Methods	Analogical Encoding	Analogical Retrieval	Analogical Mapping	Analogical Evaluation
TRIZ, 1956 [65]			*	
ARGO, 1988 [58]	*	*		
IDEAL, 1996 [57]	*	*	*	
KRITIK, 1997 [56]	*	*	*	
BIO-TRIZ, 2002 [67]			*	
Pat-Analyzer, 2007 [66]			*	
E2B, 2010 [78], 2011 [77]			*	
PAnDA, 2011 [84]	*	*		
WordTree, 2012 [4]	*	*		
DANE, 2012 [79]	*	*		
IEKG, 2012 [60]			*	
Fu et al., 2013 [72]	*			
Func. Vector, 2014 [5]	*	*[50]		
Four-box, 2014 [80]				*
Cheong et al., 2014 [81]		*	*	
Glier et al., 2014 [82]				*
AskNature, 2014 [83]	*			
DRACULA, 2015 [87]	*	*	*	
SEABIRD, 2016 [85]	*	*	*	
Analogy Finder, 2016 [68]		*		
Sanaei et al., 2017 [62]		*		
Song et al., 2017 [41]		*		
B-Link, 2017 [55]	*	*		*[19]
Gilon et al., 2018 [63]		*		
SOLVENT, 2018 [64]		*	*	
Retriever, 2018 [18]		*		
Combinator, 2018 [17]			*	
Idea-Inspire, 2018 [53]	*	*	*	*
Kwon et al., 2019 [91]	*[14]	*		*[19]
TechNet, 2020 [54]		*[70]		
VISION, 2020 [7]	*	*		
Zhang et al., 2020 [92]	*	*		
FOBIE, 2020 [89]			*	
Jiang et al., 2021 [13]	*	*		
InnoGPS, 2021 [15]	*		*	

Methods

The AI-based techniques that have already been used in DbA studies are summarized in Table 3. As shown in Fig. 5, the broadly defined NLP methods, including text mining, topic modeling, semantic networks, and pre-trained word embeddings, have played essential roles in supporting all stages of DbA. These methods aim at understanding and mining the textual nuances of natural language. However, some of them tackled the rough source data using only traditional textual pre-processing methods. They developed

some syntactic rules for the processed dataset, which were still limited to the human experience. These pre-defined rules cannot be updated automatically to match the rapid advances in AI technologies. Some researchers that adopted knowledge graph techniques generally relied on the commonsense semantic networks trained on non-engineering data sources such as WordNet and ConceptNet. However, engineering designers have perceptions of technical terms that are distinct from the commonsense knowledge [54]. In this case, the technological and engineering semantic networks (e.g., TechNet and B-link) that were recently developed based on engineering data can serve as infrastructure to support broad DbA studies.

Table 3: Existing AI-based methods and algorithms that have been adopted or developed for DbA

Approach	Method	Refs
Artificial Neural Network (ANN)	Convolutional Neural Network (CNN)	[13, 92]
	Recurrent Neural Network (RNN)	[54]
	Bi-directional Recurrent Neural Network (Bi-RNN)	[63, 89]
	Long short-term memory (LSTM)	[62]
Clustering	Agglomerative Hierarchical Clustering (AHC)	[41]
	Relevance score-based clustering (RSC)	[7, 72]
Classification	Support Vector Machine (SVM)	[62, 82]
	Naive Bayes (NB)	[82]
	K-Nearest Neighbors (KNN)	[82]
Text mining	Text preprocessing techniques	[53, 54, 55, 63, 64, 66, 68, 81, 84, 85]
	Syntactic Analysis	[81, 89]
Word Embedding	Word2Vec	[53, 54, 64]
	Doc2Vec	[64]
	Glove	[64]
	ELMO	[89]
Topic-modeling	Latent Semantic Analysis (LSA)	[72]
	Latent Dirichlet Allocation (LDA)	[106]
	Non-Negative Matrix Factorization (NMF)	[7]
Semantic Network and Knowledge Graph	WordNet	[4, 62]
	ConceptNet	[17, 18]
	Cyc	[63]
	TechNet	[14, 19, 70]
	B-Link	[19]
Vector Space Method (VSM)	Functional vector space	[5, 50]
	Technology space map	[15]
	PA and OA space	[84, 85]
	Dijkstra's shortest path algorithm	[55]
Probability analysis	Bayesian inference approach	[72]
Expert system	Heuristic rules-based strategy	[53, 56, 57, 58, 60, 65, 67, 77, 78, 79, 80, 87, 91]

Regarding the other methods, clustering algorithms have been used to represent design data by categories and facilitate the retrieval process [7, 41, 72]. Vector Space Method represents the original data as a vector of subitems, which also benefits both the encoding and retrieval subphases [5, 15, 50, 84, 85]. It is worth mentioning that recent deep learning techniques based on various structure models have shown great potential in some human tasks such as object recognition and classification. Specifically, in DbA studies, some initial intents have been employed to use CNNs and RNNs to support DbA tasks, ranging from analogy representation to analogy mapping [13, 54, 62, 63, 89, 92].

The rapid advances in deep learning-based AI techniques may provide many powerful tools for data-driven DbA. For example, visual and semantic analogy question solving has been explored extensively [104, 107, 108, 109, 110]; majority of the studies involve discovering an extendable mapping from an image and word pair and applying it to another image and word pair to find analogies. For example,

Sadeghi et al. [109] utilized a quadruple Siamese network architecture to address the problem associated with creating visual analogies for natural images. Similarly, Rossiello et al. [104] used hierarchical Siamese networks to learn relational representations by textual analogy. Moreover, various deep learning models, besides deep RNNs and CNNs and their variations, generative adversarial networks (GAN) [111], transformer [112] and graph neural networks (GNN) [113], have proven effective for learning complex features from datasets. More recently, the transformer-based models built by Open.AI, such as BERT [114], GPT-3 [115], and DALL·E [116], have exhibited record-breaking performances in understanding and generating new texts and images from given conditions. DbA researchers may adopt these state-of-the-art models from the frontier of AI research to support relational analogy representation and analogical reasoning for engineering design. However, note that the model interpretability of deep learning has always been a limiting factor for real-world use cases, where model outputs should be explained. Developing explainable AI (XAI) models [117] for DbA tasks is still challenging for the community currently and needs more exploration.

Future Opportunities and Directions

In this paper, we have analyzed the literature on data-driven DbA to elucidate its state-of-the-art. Although existing methods and tools have already shown their ability to support the DbA process, many areas remain under-researched. In this section, we map the course of feasible directions for future research from the perspective of data, methods, applications, and their interactions.

New data sources exploration

Based on our literature review, patent and scientific paper databases are suitable data sources because of their large size, as well as the rich information they possess on potential design inspiration. Both have been used in many data-driven DbA studies [5, 15, 64]. Although crowdsourced design data [97, 98, 99] and crowdfunded project data [100] have been explored in design research and are also potentially useful for DbA studies, they have not been fully exploited. In addition, most of the current data-driven DbA methods were designed to merely mine textual information as analogy candidates; only a few researchers have focused on the data of other modalities [13, 92]. Scholars have revealed that designers often prefer visual representations to textual descriptions for idea generation [118]. Therefore, various kinds of visual design sources, such as technical images [13], design sketches [92], 3D product models [101, 102, 103], ought to be exploited and mined. It is recommended that researchers develop more tools based on the new data sources and broad data modalities.

Benchmark datasets construction

The lack of distinct ground-truth and gold-standard benchmarks in DbA-related problems makes it impossible to directly compare different DbA methods and evaluate their performance in DbA applications reliably. For example, different analogy retrieval methods are typically applied to different datasets; literature [63] used a scientific paper database and literature [30] used a patent database. In this case, authors often define their own criteria for assessing performance in the absence of ground truth, which makes it difficult to determine the relative performance, strengths, and weaknesses of each method. This also prevents the community from following, adopting, and iterating on the state-of-the-art algorithms. Summers et al. [119] similarly discussed the necessity and benefits of establishing benchmark datasets for function representation in engineering design. As a broader field, data-driven DbA requires a series of benchmark datasets for four different subprocesses. Some attempts have been made to create such benchmarking standard, such as the recently released FOBIE dataset [16] that was developed to support the semi-open relationship extraction for nature-inspired engineering. However, in practice, we should note that the ground truth labels used as the benchmarking standard may suffer from the same biases as the training data. We recommend that researchers create a series of benchmark datasets by combining

different data sources for DbA tasks to surmount such limitations.

Engineering knowledge graph-based applications

Data-driven DbA requires the knowledge database in the backend. When it comes to the large and cross-domain ones, the knowledge base needs to be structured and organized to enable effective retrieval and mapping. In the engineering design literature, the network map of technology domains [40, 41] and semantic networks [54, 55] have been created and used to guide the retrieval of analogical design stimuli based on the quantified knowledge distance between domains
[15] or the semantic distance between concepts [70]. Meanwhile, the most frequently used knowledge bases are the large pre-trained commonsense knowledge graphs such as WordNet, ConceptNet and FreeBase [120, 121].

In the field of AI, the knowledge graph notion specifically refers to the qualitative networks composed of many heterogenous <entity, relation, entity> triplets [122, 123]. The relations are qualitatively specific and sensible for humans, in contrast to the one-dimensional semantic distance or knowledge distance constructs. The past decade has seen the rapid adoption of large knowledge graphs in powering knowledge representation, retrieval, and mapping tasks in many AI applications. Google, Facebook, Amazon, TikTok, among others, have developed and used knowledge graphs to support the search, recommendation, and query engines in their AI products and services.

Therefore, it is recommended that researchers extend the uses of such large pre-trained cross-domain semantic networks (e.g., B-Link [55] and TechNet [54]) or knowledge graphs (e.g., WordNet [95] and ConceptNet [96]) as knowledge bases for developing knowledge-based expert systems for DbA. Furthermore, it will be a promising effort to develop graph knowledge bases in which the entities are presented at the hierarchical domain to concept levels and both the quantitative and qualitative relations between entities are pre-trained, to augment analogy retrieval and mapping across domains and concepts.

Deep learning solutions, new models, architectures, and XAI

Deep learning techniques based on various structure models, such as CNNs and RNNs, have shown their strengths in supporting DbA tasks, ranging from analogy representation to analogy mapping [13, 54, 63, 89, 92, 97]. On the basis of these efforts, the up-to-date basic deep learning structures (e.g., transformer [112] and GNN [113]) and large-scale pre-trained models (e.g., BERT [114] and GPT [115]) ought to be explored and utilized to facilitate all DbA subprocesses. There exist many opportunities for collaborative research among the psychological cognition, engineering design, and deep learning communities, and new models and algorithms can be developed to handle various types of data sources. Specifically, in analogy mapping and evaluation tasks, domain expertise is required to interpret the results yielded by deep learning algorithms.

In addition, understanding the fundamental properties and interpretability of deep learning is an essential part of such data-driven studies. For example, when we use deep learning to retrieve analogy candidates, being aware of the similarity type of analogies is as important as obtaining the recommended analogy candidate list. The retrieval systems should provide users more insightful information with the final outputs. In this regard, one of the pioneer works is SOLVENT [64], which attempts to interpret the types of analogies by annotating and mapping different aspects of research papers. Beyond this, researchers may combine the modern deep learning with traditional interpretable machine learning techniques (e.g., decision trees and Bayesian classifiers) to develop more explainable models for DbA.

New subprocess of data-driven DbA: Analogy-based design synthesis

In the traditional DbA process, following the analogy retrieval and mapping steps, the final design synthesis stage typically requires the experience, intelligence, and efforts of designers. As a comparison, various large-scale datasets collected and represented for previous DbA subprocesses also provide opportunities for scholars to explore the deep learning-based data synthesis and generative methods, which

we term as a new subprocess of data-driven DbA, analogy-based design synthesis. Unlike traditional design synthesis strategies, such as shape grammars and constraint programming, data-driven design synthesis methods do not necessarily require expert knowledge and can automatically learn to generate plausible new designs from datasets [124, 125, 126, 127, 128, 129, 130].

Among various data-driven design synthesis methods, recent deep generative models have gained significant attention because of their ability to learn complex features from samples. GAN [111] and Variational autoencoder (VAE) [131] are the most commonly used deep generative models to assist engineering design. Specifically, variations of GAN have been used in designing airfoils [124, 125], car wheels [126], bicycles [127], airplane [129] and social robots [130], and VAE has been used in designing material microstructures [128]. Recently, the pre-trained GPT [115] and DALL·E
[116] models released by Open.AI have exhibited record-breaking performances in creating novel texts and images. As these synthesis methods typically require given inputs as the initial design information, the identified analogies from the previous DbA steps may potentially guide the direction of design generation. Therefore, it is recommended that researchers extend the use of these existing AI-based generative models and algorithms to develop analogy-based design synthesis tools.

Taken together, the foregoing propositions can be merged and implemented in one future data-driven DbA system that applies various latest AI techniques to different DbA subprocesses. In Fig. 6, we conceptualized such a data-driven DbA system, in comparison with the traditional one. In this future system, multi-source and multimodal design-related data are embedded and stored in a high-dimensional vector space through representation learning. The vector space is dynamic and can be updated based on the requirements and performance of the downstream analogical tasks. Analogical retrieval, mapping and evaluation will be conducted based on the distributed representation of design data samples through various data mining techniques. Benchmark datasets are available for researchers to make comparisons among new and alternative algorithms. During analogical reasoning, interpretable machine learning will enable us peek into the support evidence of analogy making, such as providing similarity types among analogies. Finally, deep generative models will be utilized to automatically generate new designs according to the recommended analogy pairs and their mapping relationship. In this system, human-computer interaction may play an essential role in this process to improve the quality of design synthesis. These newly created designs can also be assessed by the intelligent evaluation model that considers user needs, sentiments, and compassion.

Furthermore, it is important to note that the design of future data-driven intelligent DbA systems should be grounded on the scientific understanding of analogy. The available cognitive science foundations of analogy had been mainly developed via traditional and small human experiments and remained abstract and at a macro-level of understanding and application [3, 22, 26, 27, 29, 38, 39, 46, 76]. With the growing development and uses of data-driven DbA systems, we anticipate a growing amount of data on various DbA processes and outcomes to be generated. Such big data may offer new research opportunities by analyzing them to nuance and advance our scientific understanding of the cognitive process and mechanisms of analogy. In turn, the new understanding can be applied to develop more explainable AI systems for design by analogy.

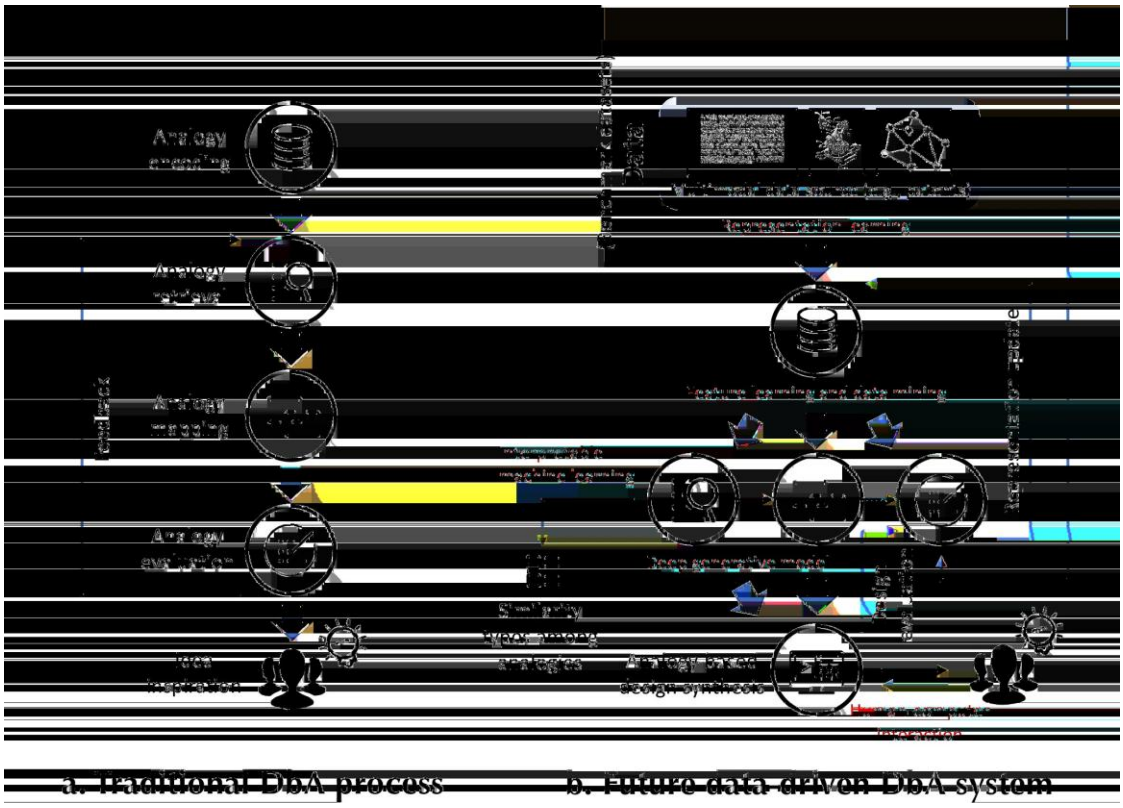


Figure 6: Schematic comparison between the traditional DbA process and future data-driven DbA system

Conclusion

Data-driven Design-by-Analogy has a lot of promise to boost innovation and creativity in design. By clarifying the most recent data-driven DbA studies and pointing forth fresh, workable study directions, we add to the body of knowledge. Our methodical analysis of the data, applications, and techniques of current data-driven DbA systems and tools offers a framework or foundation for scholars to follow pertinent recent initiatives in a logical manner. To close the gap between engineering design and AI research, a number of study topics are suggested in terms of data, applications, and methodologies, respectively. We believe that researchers working to create more potent and intelligent data-driven DbA techniques, tools, and systems will find guidance in our review and recommendations.
AI and data science techniques can be helpful to supplement other traditional engineering design methods [132], including TRIZ [65, 66], design heuristics [133, 134], design principles [94, 135], design structure matrix [136, 137], product family and platform design [138, 139], first principles [140, 141, 142], C-K [143], blending [144], and combinational design [99, 145]. We believe that this work will inspire academics in these related domains to use the most recent AI and data science technology to develop data-driven techniques for their design methodology.

References

[1] Ullman, David G. The process of mechanical design, volume 2. 1992, McGraw-Hill, New York.
[2] Ashok K. Goel. Creativity, design, and analogies. 12(3):62–70, IEEE expert, 1997.

Christian D. Schunn and Bo T. Christensen
[3]. Engineering design is an example of how analogical distance is related to analogical function and preinventive structure. 35(1), 29–38, Memory & Cognition, 2007.
[5] Kristin Wood, Dan Jensen, Maria Yang, Kevin Otto, Katherine Fu, and Jeremy Murphy. A functional vector technique to analogical search is function-based design-by-analogy. 136(10): 101102, ASME Journal of Mechanical Design, 2014.
[6] Song Hyeonik and Fu Katherine. Design-by-Analogy: Examining Behavior, Material, and Component-Based Structural Representation of Patent Databases for Analogical Inspiration. 19(2):021014, 2019, ASME Journal of Computing and Information Science in Engineering.
[7] Jacob Evans, Katherine Fu, and Hyeonik Song. A computationally aided design-by-analogy method based on exploration using D3. 34(4): 444–457, 2020; Artificial Intelligence for Engineering Design, Analysis, and Manufacturing.
[8] L. H. Shu and Ashok K. Goel. An overview of analogous thinking in relation to design. 29(2, SI):133–134, 2015; Artificial Intelligence for Engineering Design, Analysis, and Manufacturing.
[9] Ian Tseng, Kristin L. Wood, Jonathan Cagan, Katherine Fu, Julie S. Linsey, and C. Schunn. An analysis of engineering design faculty perceptions, mitigation strategies, and design fixation. 132(4):041003, ASME Journal of Mechanical Design, 2010.
[10] Kristin Wood, Joel Chan, Jonathan Cagan, Christian Schunn, Kenneth Kotovsky, and Katherine Fu. What "near" and "far" mean: how design databases are structured and how analogy distance affects design output. 135(2):021007, ASME Journal of Mechanical Design, 2013.
[11] Kristin L. Wood, Amanda Swee, Sabrina Png, Kenneth Teo, Carlye Lauff, and Yu Hui Wee. Methodology Handbook for Design Innovation (DI). Anschutz Medical Campus, Singapore, 2021 | University of Colorado Denver (CU Denver) and Singapore University of Technology and Design (SUTD). 978-981-18-1207-1 (ISBN).
[12] Mirjana Puharic, Vojkan Lucanin, Marko Rakovic, Srdjan Zivkovic, and Suzana Linic. Using the Bionic High-Speed Train as an example, experimental and numerical approaches for concept design and flow transition prediction are presented. 65–82 in the Proceedings of the International Conference on Experimental and Numerical Investigations and New Technologies. 2020 Springer.
[13] Shuo Jiang, Christopher L. Magee, Jianxi Luo, Guillermo Ruiz-pava, and Jie Hu. Convolutional neural networks are used to generate design feature vectors for patent pictures. 143(6):061405, ASME Journal of Mechanical Design, 2021.
[14] Jianxi Luo and Sarica Serhad. Create Knowledge Representation Using Semantic Networks in Technology. International Conference on Engineering Design (ICED), Gothenburg, Sweden, August 16–20, 2021, Proceedings of the Design Society.
[15] Kristin L. Wood, Jianxi Luo, and Serhad Sarica. Knowledge distance as a guide for data-driven design ideation.
218:106873, Knowledge-Based Systems, 2021.