Viewpoint

Amplifying Domain Expertise in Clinical Data Pipelines

Protiva Rahman, PhD; Arnab Nandi, PhD; Courtney Hebert, MSc, MD

The Ohio State University, Columbus, OH, United States

Corresponding Author:
Protiva Rahman, PhD
The Ohio State University
1800 Cannon Drive
Columbus, OH, 43210
United States
Phone: 1 3056135087
Email: rahman.92@osu.edu

Abstract

Digitization of health records has allowed the health care domain to adopt data-driven algorithms for decision support. There are multiple people involved in this process: a data engineer who processes and restructures the data, a data scientist who develops statistical models, and a domain expert who informs the design of the data pipeline and consumes its results for decision support. Although there are multiple data interaction tools for data scientists, few exist to allow domain experts to interact with data meaningfully. Designing systems for domain experts requires careful thought because they have different needs and characteristics from other end users. There should be an increased emphasis on the system to optimize the experts' interaction by directing them to high-impact data tasks and reducing the total task completion time. We refer to this optimization as amplifying domain expertise. Although there is active research in making machine learning models more explainable and usable, it focuses on the final outputs of the model. However, in the clinical domain, expert involvement is needed at every pipeline step: curation, cleaning, and analysis. To this end, we review literature from the database, human-computer information, and visualization communities to demonstrate the challenges and solutions at each of the data pipeline stages. Next, we present a taxonomy of expertise amplification, which can be applied when building systems for domain experts. This includes summarization, guidance, interaction, and acceleration. Finally, we demonstrate the use of our taxonomy with a case study.

(JMIR Med Inform 2020;8(11):e19612) doi: 10.2196/19612

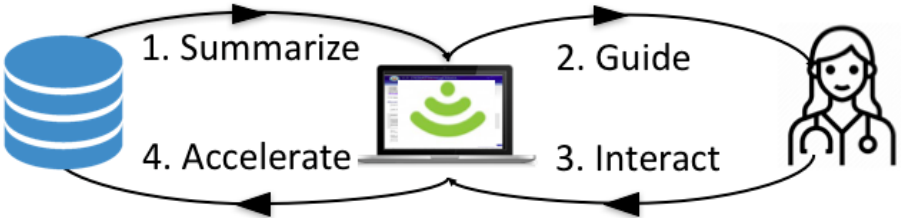
KEYWORDS

review; data analysis; data science; clinical informatics

Introduction

Recent advancements in data availability (eg, digitization of health records) and deep neural networks [1] have led to the resurgence of artificial intelligence. This has served as a catalyst for data-driven decision making in many domains. However, for high-stakes applications, such as financial and health care domains, it is rare for domain experts to execute decisions solely based on artificial intelligence algorithms [2]. Domain experts in this context are individuals who are not necessarily trained in computational fields but inform the design and are end users of data-driven algorithms (eg, health care providers, hospital administrators). Note that domain experts can have different levels of expertise in their specific domain (eg, interns, residents, attendings), and we do not differentiate between these levels in this work. Although the role of experts has been studied in clinical decision support (CDS), we find a gap in their involvement in the data analysis pipeline, which we focus on in this work.

Figure 1. Domain expertise amplification.



Domain expert involvement remains necessary in the health care domain, but this involvement brings significant challenges and implications for data-driven applications. Domain experts are expensive resources with limited time for these efforts, and excessive reliance on domain expertise could potentially lead to systems that are overly customized and not reproducible or scalable. Owing to these challenges, designing systems for them requires careful consideration. To address these challenges, we present a framework for amplified intelligence that identifies the points in the process where expertise can be effectively leveraged. *Amplification of expertise* then refers to the process of automating redundant or inferable tasks, so that domain experts can focus their efforts on tasks that require domain knowledge. This is a synergy between the domain expert and the system, which involves summarization of data and decisions, guidance toward insights, interaction by the domain expert, and acceleration of input (Figure 1).

Prior Work

There is active research on interactive and human-in-the-loop systems in many computer science disciplines. The database and visualization communities have produced numerous tools [3-8] to aid data scientists with data wrangling and analysis. At the decision-making stage, the machine learning community has looked at making black box models explainable [2,9-12], while the human-computer interaction (HCI) community has been studying how differences in explainability affect decision making [13,14]. Finally, the crowdsourcing community has concentrated on human-powered computation by optimizing tasks (eg, simplifying tasks [15], minimizing the number of questions [16,17], optimizing workflows [18-20]). However, we focus on data-powered experts by amplifying expertise. Although we draw from prior work, systems designed for health care domain experts require special consideration because they have characteristics that distinguish them from data scientists and crowdworkers.

Special Considerations in the Health Care Domain

First, domain expert input is usually needed for data tasks that require experiential knowledge and judgment (such as medical diagnoses and forensic analysis [21]). The critical and subjective nature of these decisions necessitates transparency, both from the algorithm and domain experts. Hence, the system needs to summarize the impact of algorithmic or experts' manipulation of the data [22]. Second, due to their specialized training, domain experts' time is expensive and limited [23,24]. This constraint makes it imperative that we build tools that provide insights while reducing physical and cognitive effort [25]. Third, as domain experts are trained in noncomputational fields, systems designed for them should provide high-level interaction capabilities. This is referred to as *editable shared representations* between computers and humans [26]. Examples include natural language interfaces and form-based input [27]. Finally, domain experts are highly trained individuals, which allows systems to accelerate their input by using domain-specific assumptions and ontologies [28,29]. Keeping these factors in mind, expertise amplification involves summarization, guidance, interaction, and acceleration (Figure 1). We will explore each of these in detail in the following sections.

The Data Pipeline

There are opportunities to amplify expertise at all stages of the pipeline. The data pipeline refers to the different stages that the data need to go through before they can provide decision support. It can roughly be broken into 3 stages: curation, cleaning, and analysis. Tools at the end of the pipeline have only looked at explaining models but not at amplification. In contrast, tools at earlier pipeline stages have been designed mainly for data scientists and not for experts. However, domain experts are involved at every stage of the pipeline [27-31], especially in clinical research settings where data sets contain specialized information. Thus, there is a need to amplify domain expertise throughout the pipeline. In this work, we provide examples from the informatics literature to highlight the need for expert involvement at each pipeline step. We then review literature from the database, HCI, and visualization communities about challenges and current approaches at different stages. On the basis of our review, we present a novel taxonomy for amplifying domain expertise and demonstrate its use with a case study in empiric antibiotic treatment. Our review can serve as a guide to new clinical research projects, and our taxonomy can be applied when designing systems for experts, especially for low-budget projects when there are limited resources and availability of domain experts.

Challenges in the Data Pipeline

This section is organized to reflect the clinical data pipeline, which often involves the following steps: data are curated from the electronic health record (EHR) data warehouse and annotated with external data sources, cleaned and validated, and analyzed. Multiple people are involved in various stages of the pipeline. The prevalent notion of the workflow is that a data engineer restructures, cleans, and sets up the infrastructure for data analysis, and a data scientist then analyzes and models the data, which a software engineer implements into a decision support system. A domain expert then consumes the end product to make decisions. However, in clinical settings, domain expert involvement is required at every step of the pipeline. Allowing domain experts to directly and efficiently interact with data removes the need for them to rely on a data engineer or data scientist who can then focus on infrastructure and model construction. Moreover, since domain experts are the stakeholders in the output of data pipelines, in our experience, they tend to be engaged users who want to interact with data and leverage their expertise. In this section, we motivate domain expert involvement with examples from the past five years of research presented at the American Medical Informatics Association's annual symposiums. We then review the computer science literature to identify current tools and opportunities for expertise amplification at the 3 stages of the data pipeline: data curation, data cleaning, and data analysis, as each of these corresponds to a research area of its own.

Data Curation

Curating data sets for analysis can be a laborious process that can involve combining multiple data sources and identifying relevant attributes. Data integration and data discovery address these problems.

Data Integration

Medical data pipelines often involve data that were collected for purposes other than answering the research question at hand. This usually implies that information is not captured in a manner fit for analysis [32,33], with issues such as missing metadata information [34]. Moreover, in some situations such as rare disease studies, the cohort size is too small for analysis [35], while in other cases, external features such as air quality or drug components [36-39] might be needed. One possible solution to these data quality issues is to curate data from multiple institutions and external sources. However, the different data representations [35,40] pose challenges in entity matching, metadata inference, and data integrity [41,42]. Data integration aims to automatically resolve schema matching and entity matching problems during data curation. For biomedical data sets, integration can involve standardization by mapping to ontologies with controlled vocabularies [43-45]. Although current approaches use deep learning for integration [46-50], generating a training corpus and validating results require domain expert input. For example, Cui et al [35] require domain experts to validate data curation efforts for studying sudden death in epilepsy. In another example, building an automatic concept annotator for standardizing biomedical literature [50] required experts to manually annotate different concepts [51-54]. Furthermore, a domain expert will be able to catch inconsistencies or errors made by an automated integration tool much faster than a data engineer who is unfamiliar with the domain. Thus, there is a need to build interactive data integration tools for domain experts.

Data Discovery

Data discovery refers to the process of finding relevant attributes or cohorts for analysis. This is especially true for multidisciplinary teams where the domain expert knows the disease definition but is not familiar with the database schema. At the same time, the data engineer can explore the schema but might not recognize that a field is relevant. Integrating data from multiple sources only exacerbates this problem. In the informatics community, DIVA [55] aids in cohort discovery by ingesting expert-defined constraints, while visual analytic systems [56,57] such as CAVA provide an interactive interface. In the database community, Nargesian et al [58,59] have looked at finding unionable (more data points) and joinable (more attributes) data for a given data set. These algorithms are useful when trying to augment data sets with publicly available data sets such as MIMIC [60] or even for exploring a complex schema such as the Unified Medical Language System (UMLS) [61]. In addition to using properties of the data to find possible attribute matches, domain rules can be useful for identifying relevant data subsets. This requires an interactive interface where domain experts can look at subsets of interest and iteratively join and filter the data [62] to find the required cohort. Recently, query logs have been used to design precision interfaces [63,64] that customize the interface for the user's task.

Data Cleaning

After curating relevant data sets, data still need to go through multiple preprocessing steps before they are analysis-ready. These include identifying and fixing incorrect data, data augmentation, and data transformation [65], all of which benefit from domain expert involvement.

Error Fixes

EHR data are known to be messy and have errors and missing values [66-68]. A typical data cleaning method is the use of rule-based systems that identify dirty data by detecting violations of user-specified rules or known functional dependencies [69-78]. These systems do not optimize the expert's rule specification process. Crowdsourcing systems have also been used to correct values [18,79], although they are not always an option due to data complexity or confidentiality. Another approach to identify and clean data is to augment the data with external knowledge bases [80-82]. More recently, there have been many approaches [83-85] that use deep learning for automated data cleaning. Of note is Holoclean [84], which uses a statistical model to combine various data repair signals such as violation of integrity constraints, functional dependencies, and knowledge bases. Although this achieves higher performance than using each method in isolation, there is scope for identifying which of the signals are performing the poorest or what additional information would help improve the system's performance. Identifying this information, incorporating domain knowledge, and presenting it succinctly to a domain expert remain open problems.

Data Augmentation

Although data entry errors [86] and missing information can be imputed by semiautomated methods, a more difficult problem is that of creating a gold standard for training data, which is referred to as data augmentation. Many health care applications require annotating training data, for example, clinical text annotation [87-89], CDS [90-92], identifying new terms for ontologies [93], index terms for articles [94], and disease-specific annotations [51,95,96]. However, very few applications focus on optimizing the domain expert's data augmentation effort, which is eventually crucial to model performance. A notable approach to this is the Snorkel system [97], which automates data augmentation by learning the labeling function, thus accelerating the domain expert's input. However, there are opportunities to make the initial labeling process more interactive, as domain experts are required to write code in Snorkel. Furthermore, the system does not provide feedback on how labels affect the data set or final model, which is crucial for building trust in medical pipelines. Examples of interactive solutions include Icarus [28] for augmenting microbiology data and Halpern et al's system [98] for annotating clinical anchors. Both systems use an ontology to interactively amplify domain expertise.

Data Transformation

Other than fixing incorrect values and augmenting data sets, often, data need to be restructured (eg, splitting values in a column, reformatting dates). Data wrangling has emerged as a separate field in the past decade because of data diversity. Potter's Wheel [99] is one of the first interactive data transformation systems. It allows the user to specify data transforms that are encoded as constraints and used to detect errors. Building on this idea, systems such as Polaris [100] and

Trifacta [4,101] infer syntactic rules from user edits. Similarly, programming-by-example systems [102,103] learn transformations from a set of input-output pairs. These techniques have informed the autofill function of Microsoft Excel. As many domain experts employ Excel for data transformations and analysis [104], spreadsheet interfaces should consider incorporating domain knowledge.

Data Analysis

We now move to the final step of the pipeline. This includes exploratory analysis to identify attributes of interest and explainability of models for decision making.

Data Exploration

During the exploration step, it is crucial for the domain expert to be able to directly interact with the data for effective hypothesis generation. However, domain experts often must go through a data engineer to execute the relevant query [105,106] or extract information from unstructured notes [107]. The data are then validated by the domain expert through manual chart review, since data engineers without domain knowledge may apply naive filters that hide insights or find spurious correlations. To address these challenges, the informatics community has built tools to accelerate chart review [108] and allow interactive filtering and analysis [109,110]. Finalizing an analysis data set can then take multiple iterations of requests and validations between the domain expert and data engineer. In some cases, data engineers create custom dashboards for domain experts [111-113], but the latter are then limited to brushing and linking on the provided view. Mixed-initiative interfaces such as Tableau [100] and Dive [5] recommend visualizations based on statistical properties of the data but do not use domain-specific ontologies that can enrich the domain experts' interaction and accelerate their workflow.

Visualizations, when used appropriately, can provide effective summaries and reveal patterns not immediately evident by statistical overviews [114]. Summaries reduce the cognitive load on domain experts during multidimensional data exploration, allowing them to drill down to specific instances as needed [115]. Although many visualization recommendation systems exist for analyzing numerical data [7,116-118], visualizations in health care often include categorical and text data [119-122]. As such, node-link diagrams are a common data representation and have been used for tracking family history [123], decision making [22,124], and identifying hidden variables [125]. Visual interfaces thus amplify expertise by summarizing data. However, they can be more powerful if they allow interaction, provide guidance by highlighting interesting regions for exploration [126], and accelerate workflows by extrapolating domain expert interactions based on properties of the data [22]. Thus, there is a need to provide domain experts with tools that allow for more sophisticated data interaction.

Explainability

Finally, we cannot discuss clinical pipelines without discussing explainability. The interpretability of rule-based systems has made them popular in a variety of clinical applications, including decision support [127,128], antibiotic recommendation [129], updating annotations [130], and auditing [131]. Interpretability is essential because domain experts want a cause-and-effect relationship, based on which actionable decisions can be made [66,68,132]. Furthermore, health care providers may not use models they do not trust, and building trust requires providing context and explanations [2].

Current approaches in health care research use weights and activation of features to characterize attribute importance [133-135]. RuleMatrix [136] provides an alternate approach where a set of rules represents the deep learning model. The expert can explore various facets of each rule, such as data affected, distribution, and errors. In another example, Cai et al [29] built a tool to help pathologists find similar images to aid in diagnoses. The tool allows domain experts to search for similar images and then interactively refine the search results. It allows refinement by region (crop an image), refinement by concept (filter by extracted concepts from image embeddings), and refinement by example (select multiple images as examples). These refinement techniques are examples of acceleration, where interactions are interpolated to the entire data set by learning general functions. Explainability is thus key to the adoption of deep learning models. Although they have mainly been applied in the analysis stage of the pipeline, they are equally important when applying automated algorithms to curation and cleaning.

Therefore, amplifying domain experts' abilities in the analysis stage requires interactive data systems using a combination of statistical algorithms and compelling visualizations. Moreover, these systems need to follow design-study principles [137]. They need to allow interaction with domain experts for a needs assessment and an empirical evaluation to ensure that correct information is portrayed effectively. Otherwise, the system can end up burdening and biasing the domain expert instead of helping [13,101].

We have highlighted the need for domain expert involvement in the pipeline and describe some of the challenges they encounter. Although we have briefly expanded on some available solutions, Table 1 provides a more comprehensive list of references. Summarizing each technique is outside the scope of this paper, but it provides a guide to interested readers for further reading.

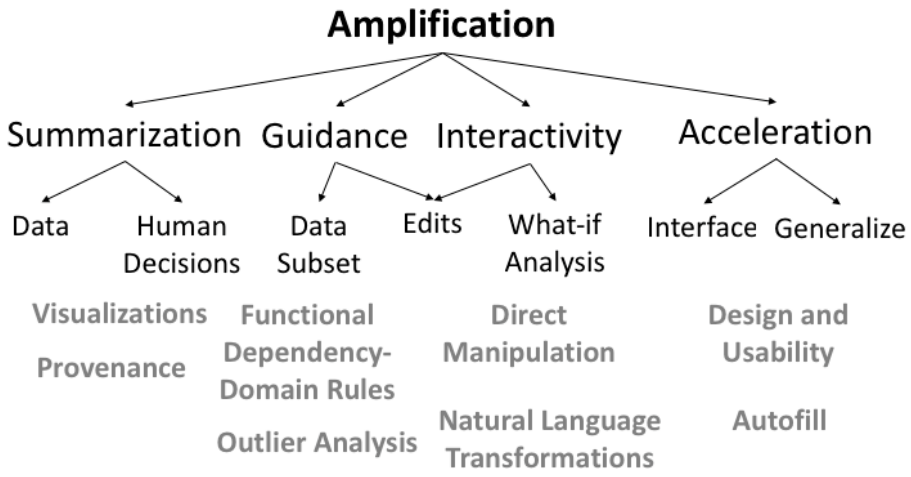
Table 1. Review of current approaches for each data pipeline stage.

Current solutions	Domain expert role
Data curation	
Data integration	
• Schema matching [138-143] • Interactive integration [144,145] • Webtables integration [146-151] • Machine learning [46-49]	Domain experts are needed to validate results of integration, and interactively correct automated methods, which can then update their algorithm
Data discovery	
• Attribute search [58,59,152,153] • Interactive querying [55,62-64] • Visual analytics [56,57]	Domain expert feedback is needed to finalize the analysis data set
Data cleaning	
Error fixes	
• Rule-based [69-77,154] • Crowdbased [18,79,155,156] • Knowledgebase [80-82] • Machine learning [83-85] • Functional dependency [15-23,25-54,58-165]	Domain expert input can be used to identify and fix errors
Augmentation	
• Machine learning [97,166,167] • Interactive [28,98,168-170]	Domain experts can augment missing data with domain-specific rules
Transformation	
• Programming by example [102,103] • Interactive rules [4,99-101] • Foreign-key detection [153,171-175]	Domain experts can restructure the data to make it semantically valid
Data analysis	
Exploration	
• Optimize performance [176-179] • Optimize insight [126,180,181] • Provenance [182,183] • Visualizations [5,7,116-118,184-188]	Domain experts interact with summaries and outliers to draw insight
Explainable	
• Systems [189,190] • Visualizations [9,12,29,136] • Empirical studies [10,11,13,14]	Domain experts inform the model design to ensure explainability

Taxonomy of Expertise Amplification

The previous section elucidated the need for domain expert involvement throughout the clinical data pipeline. In all steps, domain expert involvement can improve automated methods but must be implemented appropriately to ensure that the process remains robust and reproducible. Taking this into consideration, we propose a taxonomy that can be employed when designing systems to amplify expertise in the clinical pipeline. Domain expertise amplification by a system can broadly be categorized into 4 dimensions: summarization, guidance, interactivity, and acceleration, as shown in Figures 1 and 2. Thus, a system that wishes to amplify expertise should apply one or more of these dimensions. We demonstrate these categories with examples from computer science literature.

Figure 2. Taxonomy of expertise amplification: the first level shows the 4 dimensions that should be employed by a system for expertise amplification. The second level enumerates the subdimensions along which amplification can be done, while the fourth level in gray shows tools that can be applied.



Summarization

The time constraints of experts along with transparency requirements in the clinical domain motivate the need for effective summaries of data and human decisions. Although data summaries are important for analysis, summaries of human decisions allow for improved explainability and reproducibility.

Data

An amplification system should summarize large and complex data sets so that experts can meaningfully consume them. This is relevant for identifying inconsistencies as well as for open-ended exploration during analysis. It can be overwhelming for an expert to go through large and wide tables. Therefore, amplification systems should automatically summarize complex data [191]. Although providing data samples [28,76] and statistical summaries such as mean, variance, and standard deviation can be useful for providing a bird's eye view, they are not always enough to reveal patterns [114]. In such cases, *visual summaries* can provide additional insight, as done by the CAVA system [56]. Multidimensional data can be visually summarized by presenting each dimension as a coordinated histogram with linked brushing and filtering [176].

Human Decisions

In addition to data, amplification systems need to summarize algorithmic and human decisions as well. This is because domain expert involvement is usually required in situations where it is necessary to have high-quality data [2,21]. Hence, amplification systems also require high transparency [189,192]. To support algorithm transparency, amplification systems can show visual activation of features that led to the recommendation [9] or similar cases in the data that serve as evidence for the current recommendation [193]. Summarizing human decisions can involve expressing data transformations as natural language rules [4,28] and visual node-link diagrams [22]. Furthermore, as summarized data provide an abstract or aggregate view, there is a need for data transparency, meaning that experts should be able to trace individual data points, which contributed to the aggregate summary. This involves incorporating ideas from *provenance systems* such as Smoke [182] and Scorpion [183], which provide fast data lineage tracking. Finally, for each application, empirical studies are needed to see what and how information should be presented or summarized because too much transparency can overwhelm and negatively impact the expert [13].

Guidance

Although summaries provide a global view of the data, the goals of exploratory analysis include finding insights and data quality issues [191], which might require looking at a more detailed view. Systems can guide experts by navigating to informative subsets and by suggesting data transformations and edits.

Data Subset

Amplification systems should guide the expert's navigation to meaningful subsets. For example, SeeDB [116] automatically finds interesting visualizations. Given a query, it defines *interestingness* as the deviation of the query's result set from a baseline data set. In a similar vein**,** TPFlow [194] uses tensor decomposition to guide users to interesting regions in spatiotemporal exploration. For data cleaning, error detection algorithms such as Uguide [78] and DataProf [76] use *functional dependencies* and Armstrong samples, respectively, to find incorrect tuples for human validation, while Icarus [28] presents the expert with impactful subsets for data completion. Visual summary tools such as Profiler [184] use statistics to find data quality issues. When guiding users with visual summaries, it is important to select optimal visual encodings to reveal relevant insights or *outliers*. This can be informed from recent work by Correll et al [185], which empirically evaluated different visual encodings on their effectiveness in revealing data quality issues.

Edits

In addition to navigating data sets, amplification systems can also guide experts by suggesting data transformations to edit the data during the cleaning and preparation stage [4,28,103]. However, even in this case, transparency is required. This is evidenced by the fact that in empirical studies of Proactive Wrangler [101], users often ignored the suggested transformation but then manually performed the same one because the semantics of the operation were unclear. Methods

to aid in data transformation transparency include showing previews and transitions of the data changes [195] resulting from the transformation operation.

Interaction

Along with making system internals explainable [10], allowing experts to interact and modify data and the output of algorithms increases their trust in amplification systems [11]. For empiric antibiotic recommendation [196], this can involve allowing the health care provider to edit model features. Providing interaction comes at the cost of maintaining strict latency constraints since experts expect to see the results of interaction almost immediately [137]. Techniques for maintaining interactive performance include sampling [197] and predictive prefetching [198]. Interaction modes can include data transformation suggestions and what-if analysis.

Data Transformation

The mode of interaction for data transformation in expertise amplification systems also needs to cater to their background and training. For example, transformations should be presented as *natural language* statements [4] as opposed to code snippets [97,154]. Although graphical user interfaces can decrease trust and control for system administrators [199], they are needed in amplification systems. Gestural query systems, such as GestureDB [62] and DBTouch [200], and *direct manipulation interfaces* might be preferable to domain experts who are unfamiliar with SQL. Furthermore, domain experts' affinity for spreadsheet tools [104] motivates designing systems with spreadsheet interfaces but advanced querying capabilities such as Dataspread [201] and Sieuferd [202].

What-if Analysis

To support collaborative decision making, amplification systems should allow for what-if analysis, where domain experts can apply or test different *decisions* and *assumptions* and see how it affects the data set. Collaborative decision making is important for consensus and conflict resolution. Domain experts are highly trained and experienced individuals in their fields, which affects how they interact with systems [203,204]. Data pipeline tasks that require their input need them to apply knowledge from training and experience [28]. Such tasks inherently require judgment, which can be biased and can vary between and within domain experts [205]. To account for this bias, consensus from multiple experts is needed. However, unlike crowdworkers, where differences in results can indicate bad actors entering random choices [18,206,207], in the case of domain experts, they reveal differing judgments. As such, automatic conflict resolution [208], such as majority voting, cannot be used because disagreements require expert discussion [22]. Collaboration is required for conflict resolution, and what-if analysis can speed up this process. Capturing and sharing metadata is also useful for collaboration [209-212].

Acceleration

Time constraints of domain experts necessitate the need to accelerate their input provision. This involves designing interfaces that aid the expert's task and building interactions that interpolate from edits to generalize to multiple data points.

Interface Design

Most experts use structured interfaces such as forms [213] or free-text notes [214] for data entry or querying and spreadsheet interfaces for data exploration [104]. Following *user-centered interface* design and adhering to latency constraints is even more essential for these systems. Query interface layouts can be optimized by using statistical properties of the data [215-217] and prior query logs [64,218], while spreadsheet interfaces can be improved by incorporating higher expressibility [201,202]. The Usher [216] system, an example of the former, uses a probabilistic model on prior input form data to optimize the form structure. This involves showing highly selective data attributes at the beginning of the form to reduce the complexity at later stages, thus reducing the scope of error and accelerating input provision.

Generalize

An advantage of building systems for domain experts is that domain-specific information can be used to accelerate their input. For example, Icarus [28] uses the organism and antibiotic hierarchy encoded as foreign-key relations in the database to generalize a single edit to a rule that fills in multiple cells, accelerating the data completion process. In another example, the system by Cai et al [29] allows domain experts to refine result sets with domain-specific concepts extracted from image embeddings.

Case Study

We illustrate our taxonomy with a case study from a representative clinical data project: modeling empiric antibiotic treatment (Figure 3). We apply the 4 dimensions of amplification to the 3 stages of the pipeline. This is summarized in Table 2.

Figure 3. Data pipeline for empiric antibiotic prediction. EHR: electronic health record.

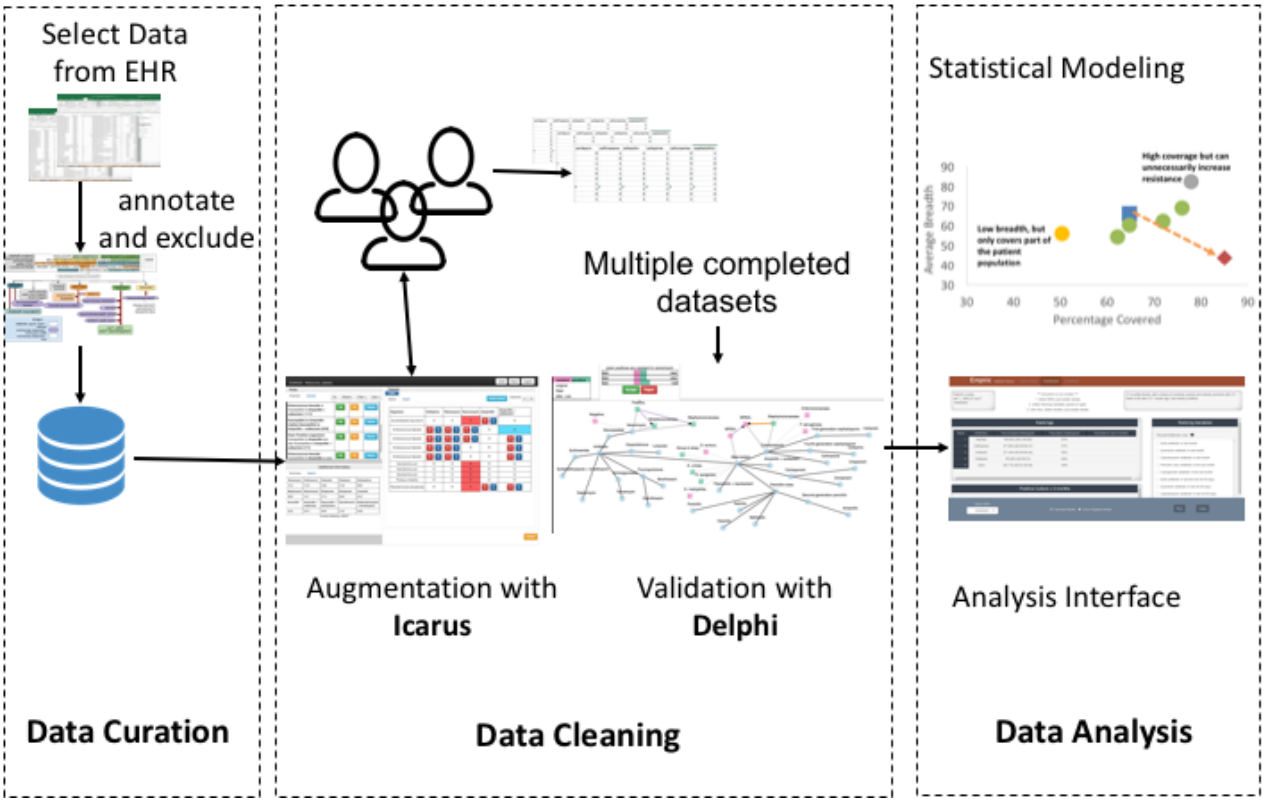


Table 2. Applying 4 dimensions of amplification to the clinical data pipeline for empiric antibiotic prediction.

Domain expert task	Amplification
Data curation	
Identify variables of interest, validate patients included in the cohort, and make domain-specific exclusionary rules	• *Summarization*: present distribution of variables of interest • *Guidance*: suggesting additional variables based on the selected ones • *Interactions*: allow expert to select and remove data points • *Acceleration*: suggest criteria based on the domain expert's inclusion and exclusion
Data cleaning	
Augmentation	
Fill in unreported microbiology susceptibilities with rules	• *Summarization*: preview a rule by showing distribution of the cells that will be impacted • *Guidance*: show high-impact data subsets for edits • *Interactions*: direct edits on interface and indirect edits via rules • *Acceleration*: suggest general rules based on the domain expert's single edit
Validation	
Validate data augmentation by examining rule set and consolidating them to remove conflicts	• *Summarization*: visual summary of rules and their relations • *Guidance*: node size guides user to high-conflict areas • *Interactions*: edit rule set by accepting and rejecting rules • *Acceleration*: automatically remove redundant rules
Data analysis	
Understand the model and its predictions for individuals and different patient subpopulations	• *Summarization*: show probability of coverage with confidence interval • *Guidance*: highlight covariates of concern • *Interactions*: allow domain expert to select covariates to include • *Acceleration*: show similar patients for who the model should be updated

At the data curation level, our domain expert, Lucy, must provide the cohort definition along with variables of interest (eg, demographics, comorbidities, allergies, etc) to a data engineer, who pulls the relevant data from the EHR data warehouse. After the data pull, Lucy looks through the initial set and formulates additional exclusion rules to ensure that it matches the clinical case definition. To implement these rules, the data engineer annotates the data with microbiology classification information of the UMLS metathesaurus [61]. This process could be improved with an expertise amplification system. The system should *summarize* data by showing the distribution of variables with linked brushing and filtering so that Lucy could see how the variable distributions are correlated. It could *guide* Lucy by suggesting correlated variables to the ones she selects. During validation of the cohort, Lucy should *interactively* be able to select data points to include. Finally, the system should be able to *accelerate* Lucy's validation by suggesting exclusion rules based on her interactions.

After the cohort is finalized, Lucy faces a data cleaning task. The microbiology laboratory provides data for only a subset of antibiotics based on domain characteristics and institutional preferences. When using these data for predictive modeling, the unreported values must be filled by domain experts. To address this, we built Icarus [28] to amplify expertise in data augmentation. Icarus *guides* the domain expert by showing them high-impact data subsets for edits. It allows both direct *interactions* via edits and indirect *interactions* via rules. Finally, Icarus *accelerates* task completion by leveraging the UMLS classification to suggest general rules based on the domain expert's single edit. It also allows the domain expert to preview the impact of a rule by *summarizing* the cells that will be impacted.

Owing to the subjective nature of this task, multiple domain experts need to come to consensus on unreported values. To amplify the consensus process, we designed Delphi [22], which visualizes the conflicts and redundancies in domain expert rules. It provides an overview of the data by visually *summarizing* the antibiotics and related rules in a node-link diagram. The node sizes *guide* the expert to regions of high conflict by encoding the number of data points affected. It allows domain experts to *interactively* edit the rule set by accepting and rejecting rules. Finally, it *accelerates* the domain experts' task completion by automatically removing redundant rules after each edit.

Once domain experts have come to a consensus, the data set is ready for analysis. Our data scientist uses penalized logistic regression to model resistance [219]. During this stage, Lucy provides insights on the different variables and their relations. After model creation, Lucy can analyze and validate the results of the interactive analysis. For a given patient, the system should *summarize* its results by showing the probability of coverage along with confidence intervals. It should *guide* Lucy by drawing attention to any abnormal covariates whose value significantly deviates from others in the cohort. It should allow Lucy to *interactively* select covariates and rerun the model for a specific patient. It should *accelerate* the analysis by showing similar patients for whom the model should also be updated.

Discussion

We have provided examples from the informatics literature to motivate the need for domain expert involvement in all steps of clinical data pipelines, from curation to analysis. Although this work is based on our experiences, we have done our best to do a targeted interdisciplinary review that can serve as a guide to clinical data projects. Our work is related to previous surveys in visual analytics in health care [188] and interactive systems [137]. Our survey is unique in that it provides a taxonomy on designing systems for amplifying expertise and focuses on the clinical data pipeline. Specifically, expertise amplification involves summarization, guidance, interactivity, and acceleration. Our case study illustrates how these can be applied to a clinical data pipeline.

Conclusions

Effectively engaging domain experts is crucial for the success of data-driven workflows. We provide a novel framework for developing systems that amplify domain expertise. Amplification systems should summarize data, guide domain experts' data navigation, allow domain experts to interact and update algorithms, and finally accelerate their task by learning from their interactions. This framework draws on research from multiple computer science disciplines. As we move toward data-driven workflows, interdisciplinary methods are necessary for the greatest impact. Empowering stakeholders to interact with the data directly can lead to faster and more impactful insights and decision making, which is vital for democratizing data to benefit society.

Acknowledgments

The research reported in this paper was supported by the National Institute of Allergy and Infectious Diseases of National Institute of Health (NIH) under R01AI116975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The National Science Foundation also supports this work under awards IIS-1422977, IIS-1527779, and CAREER IIS-1453582.

Conflicts of Interest

None declared.

References

1. Iandola F, Han S, Moskewicz M, Ashraf K, Dally W, Keutzer K. Squeezenet: alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. arXiv 2016 epub ahead of print [FREE Full text] [doi: 10.1109/CVPRW.2018.00215]

2. Ribeiro M, Singh S, Guestrin C. Why Should I Trust You: Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD'16; August 13-17, 2016; San Francisco, USA. [doi: 10.1145/2939672.2939778]
3. Zarifis K, Papakonstantinou Y. ViDeTTe Interactive Notebooks. In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics. 2018 Presented at: HILDA'18; June 1-5, 2018; Houston, TX, USA. [doi: 10.1145/3209900.3209907]
4. Kandel S, Paepcke A, Hellerstein J, Heer J. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2011 Presented at: CHI'11; March 15-17, 2011; Vancouver BC Canada p. 3363-3372. [doi: 10.1145/1978942.1979444]
5. Hu K, Orghian D, Hidalgo C. DIVE: A Mixed-Initiative System Supporting Integrated Data Exploration Workflows. In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics. 2018 Presented at: HILDA'18; June 1-5, 2018; Houston, TX, USA. [doi: 10.1145/3209900.3209910]
6. Kraska T. Northstar: an interactive data science system. Proc VLDB Endow 2018 Aug 1;11(12):2150-2164. [doi: 10.14778/3229863.3240493]
7. Wongsuphasawat K, Moritz D, Anand A, Mackinlay J, Howe B, Heer J. Voyager: exploratory analysis via faceted browsing of visualization recommendations. IEEE Trans Visual Comput Graphics 2016 Jan;22(1):649-658. [doi: 10.1109/tvcg.2015.2467191]
8. Xin D, Macke S, Ma L, Liu J, Song S, Parameswaran A. Helix: holistic optimization for accelerating iterative machine learning. Pro. VLDB Endow 2018 Dec 1;12(4):446-460. [doi: 10.14778/3297753.3297763]
9. Olah C, Satyanarayan A, Johnson I, Carter S, Schubert L, Ye K, et al. The building blocks of interpretability. Distill 2018 Mar;3(3) [FREE Full text] [doi: 10.23915/distill.00010]
10. Yeomans M, Shah A, Mullainathan S, Kleinberg J. Making sense of recommendations. J Behav Dec Making 2019 Feb 14;32(4):403-414. [doi: 10.1002/bdm.2118]
11. Dietvorst BJ, Simmons JP, Massey C. Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. Manag Sci 2018 Mar;64(3):1155-1170. [doi: 10.1287/mnsc.2016.2643]
12. Zhang J, Wang Y, Molino P, Li L, Ebert DS. Manifold: a model-agnostic framework for interpretation and diagnosis of machine learning models. IEEE Trans Visual Comput Graphics 2019 Jan;25(1):364-373. [doi: 10.1109/tvcg.2018.2864499]
13. Poursabzi-Sangdeh F, Goldstein D, Hofman J, Vaughan J, Wallach H. Manipulating and measuring model interpretability. arXiv preprint 2018.
14. Yin M, Vaughan J, Wallach H. Understanding the effect of accuracy on trust in machine learning models. In: Proceedings of CHI Conference on Human Factors in Computing Systems. 2019 Presented at: Conference on Human Factors in Computing Systems; May 4-9, 2019; Scottish Event Campus Ltd, Glasgow, United Kingdom p. 1-12. [doi: 10.1145/3290605.3300509]
15. Haas D, Ansel J, Gu L, Marcus A. Argonaut: macrotask crowdsourcing for complex data processing. Proc VLDB Endow 2015 Aug;8(12):1642-1653. [doi: 10.14778/2824032.2824062]
16. Parameswaran A, Sarma AD, Garcia-Molina H, Polyzotis N, Widom J. Human-assisted graph search: it's okay to ask questions. In: Proceedings of the VLDB Endowment. 2011 Feb Presented at: International Conference on Very Large Databased; 2011; Seattle Washington p. 267-278. [doi: 10.14778/1952376.1952377]
17. Parameswaran A, Garcia-Molina H, Park H, Polyzotis N, Ramesh A, Widom J. Crowdscreen: Algorithms for filtering data with humans. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. 2012 Presented at: ACM SIGMOD/PODS conference; 2012; Scottsdale, Arizona, USA p. 361-372. [doi: 10.1145/2213836.2213878]
18. Par H, Pang R, Parameswaran A, Garcia-Molina H, Polyzotis N, Widom J. An overview of the deco system: data model and query language; query processing and optimization. ACM SIGMOD Record 2013 Jan 17;41(4):22-27. [doi: 10.1145/2430456.2430462]
19. Lasecki W, Rzeszotarski J, Marcus A, Bigham J. The effects of sequence delay on crowd work. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015 Presented at: CHI '15: CHI Conference on Human Factors in Computing Systems; 2015; Seoul Republic of Korea p. 1375-1378. [doi: 10.1145/2702123.2702594]
20. Retelny S, Robaszkiewicz S, To A, Lasecki WS, Patel J, Rahmati N, et al. Expert Crowdsourcing With Flash Teams. In: Proceedings of the 27th annual ACM symposium on User interface software and technology. 2014 Presented at: UIST'14; October 5-8, 2014; Honolulu, Hawaii, USA. [doi: 10.1145/2642918.2647409]
21. Dror IE, Kukucka J, Kassin SM, Zapf PA. When expert decision making goes wrong: consensus, bias, the role of experts, and accuracy. J Appl Res Memory Cogn 2018 Mar;7(1):162-163. [doi: 10.1016/j.jarmac.2018.01.007]
22. Rahman P, Chen J, Hebert C, Pancholi P, Lustberg M, Stevenson K, et al. Exploratory Visualizations of Rules for Validation of Expert Decisions. In: DSIA Workshop, IEEE VIS. 2018 Presented at: Workshop on Data Systems for Interactive Analysis (DSIA); October 2018; Berlin, Germany URL: https://www.researchgate.net/publication/331177971_Exploratory_Visualizations_of_Rules_for_Validation_of_Expert_Decisions
23. Robbins R. Physicians generate an average 2.4 million a year per hospital. Southwest J Pulm Crit Care 2019;18:61 [FREE Full text]
24. Dzau VJ, Kirch DG, Nasca TJ. To care is human — collectively confronting the clinician-burnout crisis. N Engl J Med 2018 Jan 25;378(4):312-314. [doi: 10.1056/nejmp1715127]

25. Saitwal H, Feng X, Walji M, Patel V, Zhang J. Assessing performance of an electronic health record (EHR) using cognitive task analysis. Int J Med Inform 2010 Jul;79(7):501-506. [doi: 10.1016/j.ijmedinf.2010.04.001] [Medline: 20452274]
26. Heer J. Agency plus automation: designing artificial intelligence into interactive systems. Proc Natl Acad Sci U S A 2019 Feb 5;116(6):1844-1850. [doi: 10.1073/pnas.1807184115] [Medline: 30718389]
27. Krishnan S, Haas D, Franklin MJ, Wu E. Towards reliable interactive data cleaning: a user survey and recommendations. In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics. 2016 Presented at: SIGMOD/PODS'16: International Conference on Management of Data; June, 2016; San Francisco California. [doi: 10.1145/2939502.2939511]
28. Rahman P, Hebert C, Nandi A. ICARUS: minimizing human effort in iterative data completion. Proc VLDB Endow 2018 Sep 01;11(13):2263-2276. [doi: 10.14778/3275366.3275374]
29. Cai CJ, Reif E, Hegde N, et al. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI'19; May 4-9, 2019; Glasgow, Scotland, UK. [doi: 10.1145/3290605.3300234]
30. Clayton PD, Naus SP, Bowes WA, Madsen TS, Wilcox AB, Orsmond G, et al. Physician use of electronic medical records: issues and successes with direct data entry and physician productivity. AMIA Annu Symp Proc 2005:141-145 [FREE Full text] [Medline: 16779018]
31. Ganapathi E, Chen Y. Data Quality: Experiences and Lessons From Operationalizing Big Data. In: IEEE International Conference on Big Data. 2016 Presented at: Big Data'16; December 5-8, 2016; Washington, DC, USA. [doi: 10.1109/bigdata.2016.7840769]
32. Collins SA, Gesner E, Mar PL, Colburn DM, Rocha RA. Prioritization and refinement of clinical data elements within EHR systems. AMIA Annu Symp Proc 2016;2016:421-430 [FREE Full text] [Medline: 28269837]
33. Blaisure JC, Ceusters WM. Improving the 'Fitness for Purpose' of common data models through realism based ontology. AMIA Annu Symp Proc 2017;2017:440-447 [FREE Full text] [Medline: 29854108]
34. Pan X, Cimino J. Identifying the clinical laboratory tests from unspecified "Other Lab Test" data for secondary use. AMIA Annu Symp Proc 2015;2015:1018-1023 [FREE Full text] [Medline: 26958239]
35. Cui L, Huang Y, Tao S, Lhatoo SD, Zhang GQ. ODaCCI: ontology-guided data curation for multisite clinical research data integration in the NINDS center for SUDEP research. AMIA Annu Symp Proc 2016;2016:441-450 [FREE Full text] [Medline: 28269839]
36. Hall ES, Connolly N, Jones DE, DeFranco EA. Integrating public data sets for analysis of maternal airborne environmental exposures and stillbirth. AMIA Annu Symp Proc 2014;2014:599-605 [FREE Full text] [Medline: 25954365]
37. Gouripeddi R, Facelli JC, Bradshaw RL, Schultz D, LaSalle B, Warner PB, et al. FURTHeR: an infrastructure for clinical, translational and comparative effectiveness research. AMIA. 2013. URL: https://knowledge.amia.org/amia-55142-a2013e-1.580047/t-10-1.581994/f-010-1.581995/a-184-1.582011/ap-247-1.582014?qr=1 [accessed 2020-10-14]
38. Chen X, Wang F. Integrative spatial data analytics for public health studies of New York state. AMIA Annu Symp Proc 2016;2016:391-400 [FREE Full text] [Medline: 28269834]
39. Ying L. Combining Heterogeneous Databases to Detect Adverse Drug Reactions. Columbia University. 2015. URL: https://academiccommons.columbia.edu/doi/10.7916/D8Z60ND1 [accessed 2020-10-14]
40. Clarkson MD, Whipple ME. Variation in the representation of human anatomy within digital resources: implications for data integration. AMIA Annu Symp Proc 2018;2018:330-339 [FREE Full text] [Medline: 30815072]
41. Berrios DC, Beheshti A, Costes SV. Fairness and usability for open-access omics data systems. AMIA Annu Symp Proc 2018;2018:232-241 [FREE Full text] [Medline: 30815061]
42. Farach O, McGettrick C, Tirrell C, Evans C, Mesa A, Rozenblit L. RexMart: An Open Source Tool for Exploring and Sharing Research Data without Compromising Data Integrity. AMIA. 2014. URL: https://figshare.com/articles/RexMart_An_Open_Source_Tool_for_Exploring_and_Sharing_Research_Data_without_Compromising_Data_Integrity/1262228/1 [accessed 2020-10-14]
43. Maldonado JA, Marcos M, Fernández-Breis JT, Parcero E, Boscá D, Legaz-García MD, et al. A platform for exploration into chaining of web services for clinical data transformation and reasoning. AMIA Annu Symp Proc 2016;2016:854-863 [FREE Full text] [Medline: 28269882]
44. Zhang Y, Tang B, Jiang M, Wang J, Xu H. Domain adaptation for semantic role labeling of clinical text. J Am Med Inform Assoc 2015 Sep;22(5):967-979 [FREE Full text] [doi: 10.1093/jamia/ocu048] [Medline: 26063745]
45. Cui L, Tao S, Zhang GQ. A semantic-based approach for exploring consumer health questions using UMLS. AMIA Annu Symp Proc 2014;2014:432-441 [FREE Full text] [Medline: 25954347]
46. Dong XL, Rekatsinas T. Data Integration and Machine Learning: A Natural Synergy. In: Proceedings of the 2018 International Conference on Management of Data. 2018 Presented at: SIGMOD'18; June 1-5, 2018; Houston, Texas. [doi: 10.1145/3183713.3197387]
47. Stonebraker M, Ilyas I. Data integration: the current status and the way forward. IEEE Data Eng Bull 2018;41:3-9 [FREE Full text]
48. Fernandez C, Madden S. Termite: a system for tunneling through heterogeneous data. ArXiv 2019 epub ahead of print. [doi: 10.1145/3329859.3329877]

49. Thirumuruganathan S, Tang N, Ouzzani M. Data curation with deep learning vision: towards self-driving data curation. arXiv 2018:1384 [FREE Full text]
50. Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov models. Bioinformatics 2016 Sep 15;32(18):2839-2846 [FREE Full text] [doi: 10.1093/bioinformatics/btw343] [Medline: 27283952]
51. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform 2014 Feb;47:1-10 [FREE Full text] [doi: 10.1016/j.jbi.2013.12.006] [Medline: 24393765]
52. Wei C, Kao H, Lu Z. GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. Biomed Res Int 2015;2015:918710 [FREE Full text] [doi: 10.1155/2015/918710] [Medline: 26380306]
53. Krallinger M, Rabal O, Leitner F, Vazquez M, Salgado D, Lu Z, et al. The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform 2015 Jan 19;7(S1). [doi: 10.1186/1758-2946-7-s1-s2]
54. Wei C, Harris BR, Kao H, Lu Z. tmVar: a text mining approach for extracting sequence variants in biomedical literature. Bioinformatics 2013 Jun 1;29(11):1433-1439 [FREE Full text] [doi: 10.1093/bioinformatics/btt156] [Medline: 23564842]
55. Hielscher T, Niemann U, Preim B, Völzke H, Ittermann T, Spiliopoulou M. A framework for expert-driven subpopulation discovery and evaluation using subspace clustering for epidemiological data. Expert Syst Appl 2018 Dec;113:147-160. [doi: 10.1016/j.eswa.2018.07.003]
56. Zhang Z, Gotz D, Perer A. Iterative cohort analysis and exploration. Inf Vis 2014 Mar 19;14(4):289-307. [doi: 10.1177/1473871614526077]
57. Malik S, Du F, Monroe M, Onukwugha E, Plaisant C, Shneiderman B. An evaluation of visual analytics approaches to comparing cohorts of event sequences. 2014 Presented at: InEHRVis Workshop on Visualizing Electronic Health Record Data at VIS (Vol. 14); 2014 Nov 9; -.
58. Nargesian F, Pu Q, Zhu E, Bashardoost BG, Miller R. Optimizing organizations for navigating data lakes. arXiv 2018:7024 [FREE Full text]
59. Nargesian F, Zhu E, Pu KQ, Miller RJ. Table union search on open data. Proc VLDB Endow 2018 Mar;11(7):813-825. [doi: 10.14778/3192965.3192973]
60. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
61. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
62. Nandi A, Jiang L, Mandel M. Gestural query specification. Proc VLDB Endow 2013 Dec;7(4):289-300. [doi: 10.14778/2732240.2732247]
63. Zhang H, Raj V, Sellam T, Wu E. Precision Interfaces for Different Modalities. In: Proceedings of the 2018 International Conference on Management of Data. 2018 Presented at: SIGMOD'18; June 1-5, 2018; Houston, Texas. [doi: 10.1145/3183713.3193570]
64. Zhang H, Wu E. Mining precision interfaces from query logs. ArXiv 2019 epub ahead of print. [doi: 10.1145/3299869.3319872]
65. Peterson KJ, Jiang G, Brue SM, Liu H. Leveraging terminology services for extract-transform-load processes: a user-centered approach. AMIA Annu Symp Proc 2016;2016:1010-1019 [FREE Full text] [Medline: 28269898]
66. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018 Sep;22(5):1589-1604. [doi: 10.1109/jbhi.2017.2767063]
67. Che Z, Purushotham S, Khemani R, Liu Y. Interpretable deep models for ICU outcome prediction. AMIA Annu Symp Proc 2016;2016:371-380 [FREE Full text] [Medline: 28269832]
68. Adibuzzaman M, DeLaurentis P, Hill J, Benneyworth BD. Big data in healthcare - the promises, challenges and opportunities from a research perspective: a case study with a model database. AMIA Annu Symp Proc 2017;2017:384-392 [FREE Full text] [Medline: 29854102]
69. Dallachiesa A, Ebaid, A, Eldawy, A, Elmagarmid, A, Ilyas, I, Ouzzani M, Tang N. NADEEF: a commodity data cleaning system. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. 2013 Presented at: SIGMOD '13; 2013; New York. [doi: 10.1145/2463676.2465327]
70. Mayfield C, Neville J, Prabhakar S. Eracer: a database approach for statistical inference and data cleaning. 2010. URL: https://orion.cs.purdue.edu/docs/eracer.pdf [accessed 2020-10-14]
71. Meduri VV, Papotti P. Towards user-aware rule discovery. In: Information Search, Integration, and Personalization. Cham: Springer; 2017:3-17.
72. Wang J, Tang N. Dependable data repairing with fixing rules. J Data and Information Quality 2017 Jul 17;8(3-4):1-34. [doi: 10.1145/3041761]
73. Wang J, Krishnan S, Franklin MJ, Goldberg K, Kraska T, Milo T. A sample-and-clean framework for fast and accurate query processing on dirty data. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. 2014 Presented at: SIGMOD'14; June 2014; Snowbird Utah USA p. 469-480. [doi: 10.1145/2588555.2610505]
74. Krishnan S, Wang J, Wu E, Franklin M, Goldberg K. Activeclean: interactive data cleaning while learning convex loss models. arXiv 2016:2117-2120 [FREE Full text] [doi: 10.1145/2882903.2899409]

75. Xu J, Kalashnikov DV, Mehrotra S. Query aware determinization of uncertain objects. IEEE Trans Knowl Data Eng 2015 Jan;27(1):207-221. [doi: 10.1109/tkde.2013.170]
76. Wie Z, Link S. DataProf: Semantic Profiling for Iterative Data Cleansing and Business Rule Acquisition. In: Proceedings of the 2018 International Conference on Management of Data. 2018 Presented at: SIGMOD'18; June 1-5, 2018; Houston, Texas. [doi: 10.1145/3183713.3193544]
77. Yakout M, Elmagarmid AK, Neville J, Ouzzani M, Ilyas IF. Guided data repair. Proc VLDB Endow 2011 Feb;4(5):279-289. [doi: 10.14778/1952376.1952378]
78. Thirumuruganathan S, Berti-Equille L, Ouzzani M, Quiane-Ruiz J, Tang N. UGuide: user-guided discovery of FD-detectable errors. In: Proceedings of the 2017 ACM International Conference on Management of Data. 2017 Presented at: SIGMOD'17; May 14-19, 2017; Chicago, USA. [doi: 10.1145/3035918.3064024]
79. Park H, Widom J. Crowdfill: collecting structured data from the crowd. ACM SIGMOD 2014 [FREE Full text] [doi: 10.1145/2588555.2610503]
80. Huang Z, Ye H. Auto-Detect: Data-Driven Error Detection in Tables. In: Proceedings of the 2018 International Conference on Management of Data. 2018 Presented at: SIGMOD'18; June 1-5, 2018; Houston, Texas. [doi: 10.1145/3183713.3196889]
81. Chu X, Morcos J, Ilyas I, Ouzzani M, Papotti P, Tang N, et al. Katara: A data cleaning system powered by knowledge bases and crowdsourcing. ACM SIGMOD 2015 [FREE Full text] [doi: 10.1145/2723372.2749431]
82. Hao S, Tang N, Li G, Li J. Cleaning Relations Using Knowledge Bases. In: IEEE 33rd International Conference on Data Engineering. 2017 Presented at: ICDE'17; April 19-22, 2017; San Diego, CA, USA. [doi: 10.1109/icde.2017.141]
83. Biessmann F, Salinas D, Schelter S, Schmidt P, Lange D. 'Deep' Learning for Missing Value Imputationin Tables with Non-Numerical Data. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018 Presented at: CIKM'18; October 22-26, 2018; Turin, Italy. [doi: 10.1145/3269206.3272005]
84. Rekatsinas T, Chu X, Ilyas IF, Ré C. HoloClean: holistic data repairs with probabilistic inference. Proc VLDB Endow 2017 Aug 1;10(11):1190-1201. [doi: 10.14778/3137628.3137631]
85. Chu X, Ilyas IF, Papotti P. Holistic Data Cleaning: Putting Violations Into Context. In: 29th International Conference on Data Engineering. 2013 Presented at: ICDE'13; April 8-12, 2013; Brisbane, QLD, Australia. [doi: 10.1109/icde.2013.6544847]
86. Schreibstein L, Newton-Dame R, McVeigh KH, Perlman SE, Singer J, Harris TG, et al. Missing data in an electronic health record-based population health surveillance system. AMIA 2014.
87. Divita G, Zeng QT, Gundlapalli AV, Duvall S, Nebeker J, Samore MH. Sophia: a expedient UMLS concept extraction annotator. AMIA Annu Symp Proc 2014:467-476. [Medline: 25954351]
88. Rumeng L, Jagannatha Abhyuday N, Hong Y. A hybrid neural network model for joint prediction of presence and period assertions of medical events in clinical notes. AMIA Annu Symp Proc 2017:1149-1158. [Medline: 29854183]
89. Browne AC, Kayaalp M, Dodd ZA, Sagan P, McDonald CJ. The challenges of creating a gold standard for de-identification research. AMIA Annu Symp Proc 2014;2014:353-358 [FREE Full text] [Medline: 25954338]
90. Bowles KH, Ratcliffe SJ, Naylor MD, Holmes JH, Keim SK, Flores EJ. Nurse generated EHR data supports post-acute care referral decision making: development and validation of a two-step algorithm. AMIA Annu Symp Proc 2017;2017:465-474 [FREE Full text] [Medline: 29854111]
91. Shivade C, Hebert C, Regan K, Fosler-Lussier E, Lai AM. Automatic data source identification for clinical trial eligibility criteria resolution. AMIA Annu Symp Proc 2016;2016:1149-1158 [FREE Full text] [Medline: 28269912]
92. Norman C, Leeflang M, Névéol A. Data extraction and synthesis in systematic reviews of diagnostic test accuracy: a corpus for automating and evaluating the process. AMIA Annu Symp Proc 2018;2018:817-826 [FREE Full text] [Medline: 30815124]
93. Chandar P, Yaman A, Hoxha J, He Z, Weng C. Similarity-based recommendation of new concepts to a terminology. AMIA Annu Symp Proc 2015;2015:386-395 [FREE Full text] [Medline: 26958170]
94. Kavuluru R, Rios A. Automatic assignment of non-leaf MeSH terms to biomedical articles. AMIA Annu Symp Proc 2015;2015:697-706 [FREE Full text] [Medline: 26958205]
95. Feller DJ, Zucker J, Don't Walk OB, Srikishan B, Martinez R, Evans H, et al. Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. AMIA Annu Symp Proc 2018;2018:422-429 [FREE Full text] [Medline: 30815082]
96. Afshar M, Joyce C, Oakey A, Formanek P, Yang P, Churpek MM, et al. A computable phenotype for acute respiratory distress syndrome using natural language processing and machine learning. AMIA Annu Symp Proc 2018;2018:157-165 [FREE Full text] [Medline: 30815053]
97. Ratner AJ, Bach SH, Ehrenberg HR, Ré C. Snorkel: Fast Training Set Generation for Information Extraction. In: Proceedings of the 2017 ACM International Conference on Management of Data. 2017 Presented at: SIGMOD'17; May 14-19, 2017; Chicago, USA. [doi: 10.1145/3035918.3056442]
98. Halpern Y, Choi Y, Horng S, Sontag D. Using anchors to estimate clinical state without labeled data. AMIA Annu Symp Proc 2014;2014:606-615 [FREE Full text] [Medline: 25954366]
99. Raman V, Hellerstein J. Potter's Wheel: an Interactive Framework for Data Cleaning. Univesity of Berkeley. URL: http://www/ cs.berkeley. edu/? rshankar/papers/pwheel.pdf, 2000 [accessed 2020-09-28]

100. Stolte C, Tang D, Hanrahan P. Polaris: a system for query, analysis, and visualization of multidimensional relational databases. IEEE Trans Visual Comput Graphics 2002;8(1):52-65. [doi: 10.1109/2945.981851]
101. Guo PJ, Kandel S, Hellerstein JM, Heer J. Proactive Wrangling: Mixed-initiative End-user Programming of Data Transformation Scripts. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. 2011 Presented at: UIST'11; March 17, 2011; New York, USA. [doi: 10.1145/2047196.2047205]
102. Singh R, Gulwani S. Learning semantic string transformations from examples. Proc VLDB Endow 2012 Apr;5(8):740-751. [doi: 10.14778/2212351.2212356]
103. Jin Z, Anderson MR, Cafarella M, Jagadish HV. Foofah: transforming data by example. In: Proceedings of the 2017 ACM International Conference on Management of Data. 2017 Presented at: SIGMOD'17; May 14-19, 2017; Chicago, USA. [doi: 10.1145/3035918.3064034]
104. Costabile MF, Fogli D, Letondal C, Mussio P, Piccinno A. Domain-expert users and their needs of software development. HCI 2003 End-User Development Session. 2003. URL: http://giove.cnuce.cnr.it/projects/EUD-NET/pdf/Costabile-et-alCameraReady.pdf [accessed 2020-10-14]
105. Hanauer DA, Hruby GW, Fort DG, Rasmussen LV, Mendonça EA, Weng C. What is asked in clinical data request forms? A multi-site thematic analysis of forms towards better data access support. AMIA Annu Symp Proc 2014;2014:616-625 [FREE Full text] [Medline: 25954367]
106. Hruby GW, Hoxha J, Ravichandran PC, Mendonça EA, Hanauer DA, Weng C. A data-driven concept schema for defining clinical research data needs. Int J Med Inform 2016 Jul;91:1-9 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.03.008] [Medline: 27185504]
107. Edinger T, Demner-Fushman D, Cohen AM, Bedrick S, Hersh W. Evaluation of clinical text segmentation to facilitate cohort retrieval. AMIA Annu Symp Proc 2017;2017:660-669 [FREE Full text] [Medline: 29854131]
108. Hu Z, Melton GB, Moeller ND, Arsoniadis EG, Wang Y, Kwaan MR, et al. Accelerating chart review using automated methods on electronic health record data for postoperative complications. AMIA Annu Symp Proc 2016;2016:1822-1831 [FREE Full text] [Medline: 28269941]
109. Major V, Tanna MS, Jones S, Aphinyanaphongs Y. Reusable filtering functions for application in ICU data: a case study. AMIA Annu Symp Proc 2016;2016:844-853 [FREE Full text] [Medline: 28269881]
110. Zhao L. Controlling False Discoveries During Interactive Data Exploration. In: Proceedings of the 2017 ACM International Conference on Management of Data. 2017 Presented at: SIGMOD'17; May 14-17, 2017; Chicago, USA. [doi: 10.1145/3035918.3064019]
111. Romero-Brufau S, Kostandy P, Maass KL, Wutthisirisart P, Sir M, Bartholmai B, et al. Development of data integration and visualization tools for the Department of Radiology to display operational and strategic metrics. AMIA Annu Symp Proc 2018;2018:942-951 [FREE Full text] [Medline: 30815137]
112. Pore M, Sengeh DM, Mugambi P, Purswani NV, Sesay T, Arnold AL, et al. Design and evaluation of a web-based decision support tool for district-level disease surveillance in a low-resource setting. AMIA Annu Symp Proc 2017;2017:1401-1410 [FREE Full text] [Medline: 29854209]
113. Iyer G, DuttaDuwarah S, Sharma A. DataScope: Interactive Visual Exploratory Dashboards for Large Multidimensional Data. In: Workshop on Visual Analytics in Healthcare. 2017 Presented at: VAHC'17; October 1, 2017; Phoenix, AZ, USA. [doi: 10.1109/vahc.2017.8387496]
114. Anscombe FJ. Graphs in statistical analysis. Am Stat 1973 Feb;27(1):17. [doi: 10.2307/2682899]
115. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. In: Proceedings 1996 IEEE Symposium on Visual Languages. 2003 Presented at: IEEE Symposium on Visual Languages; August 6, 2002; Boulder, CO, USA. [doi: 10.1016/b978-155860915-0/50046-9]
116. Vartak M, Madden S, Parameswaran A, Polyzotis N. SeeDB: automatically generating query visualizations. Proc VLDB Endow 2014 Aug;7(13):1581-1584. [doi: 10.14778/2733004.2733035]
117. Moritz D, Wang C, Nelson GL, Lin H, Smith AM, Howe B, et al. Formalizing visualization design knowledge as constraints: actionable and extensible models in draco. IEEE Trans Visual Comput Graphics 2019 Jan;25(1):438-448. [doi: 10.1109/tvcg.2018.2865240]
118. Demiralp A, Haas PJ, Parthasarathy S, Pedapati T. Foresight: recommending visual insights. Proc VLDB Endow 2017 Aug 1;10(12):1937-1940. [doi: 10.14778/3137765.3137813]
119. Caballero GC. Visual Analytics for Evaluating Clinical Pathways. In: Workshop on Visual Analytics in Healthcare. 2017 Presented at: VAHC'17; October 1, 2017; Phoenix, AZ, USA. [doi: 10.1109/vahc.2017.8387499]
120. Widanagamaachchi W, Livnat Y, Bremer PT, Duvall S, Pascucci V. Interactive visualization and exploration of patient progression in a hospital setting. AMIA Annu Symp Proc 2017;2017:1773-1782. [Medline: 29854248]
121. Li X, Cui L, Tao S, Zeng N, Zhang GQ. SpindleSphere: a web-based platform for large-scale sleep spindle analysis and visualization. AMIA Annu Symp Proc 2017;2017:1159-1168 [FREE Full text] [Medline: 29854184]
122. Mortensen JM, Musen MA, Noy NF. An empirically derived taxonomy of errors in SNOMED CT. AMIA Annu Symp Proc 2014;2014:899-906 [FREE Full text] [Medline: 25954397]

123. Chen ES, Melton GB, Wasserman RC, Rosenau PT, Howard DB, Sarkar IN. Mining and visualizing family history associations in the electronic health record: a case study for Pediatric Asthma. AMIA Annu Symp Proc 2015;2015:396-405 [FREE Full text] [Medline: 26958171]
124. Sockolow PS, Yang Y, Bass EJ, Bowles KH, Holmberg A, Sheryl P. Data visualization of home care admission nurses' decision-making. AMIA Annu Symp Proc 2017;2017:1597-1606 [FREE Full text] [Medline: 29854230]
125. Kummerfeld E, Anker JA, Rix A, Kushner MG. Methodological advances in the study of hidden variables: a demonstration on clinical alcohol use disorder data. AMIA Annu Symp Proc 2018;2018:710-719 [FREE Full text] [Medline: 30815113]
126. Sarawagi S, Agrawal R, Megiddo N. Discovery-driven Exploration of OLAP Data Cubes. In: International Conference on Extending Database Technology. 1998 Presented at: EDBT'98; March 23-27, 1998; Valencia, Spain. [doi: 10.1007/bfb0100984]
127. Sordo M, Tokachichu P, Vitale CJ, Maviglia SM, Rocha RA. Modeling contextual knowledge for clinical decision support. AMIA Annu Symp Proc 2017;2017:1617-1624 [FREE Full text] [Medline: 29854232]
128. Gangadhar S, Nguyen N, Pesuit JW, Bogdanov AN, Kallenbach L, Ken J, et al. Effectiveness of a cloud-based EHR clinical decision support program for body mass index (BMI) screening and follow-up. AMIA Annu Symp Proc 2017;2017:742-749 [FREE Full text] [Medline: 29854140]
129. Souissi SB, Abed M, Elhiki L, Fortemps P, Pirlot M. Reducing the toxicity risk in antibiotic prescriptions by combining ontologies with a multiple criteria decision model. AMIA Annu Symp Proc 2017;2017:1625-1634 [FREE Full text] [Medline: 29854233]
130. Cardoso SD, Chantal RD, Da Silveira M, Pruski C. Combining rules, background knowledge and change patterns to maintain semantic annotations. AMIA Annu Symp Proc 2017;2017:505-514 [FREE Full text] [Medline: 29854115]
131. Hedda M, Malin BA, Yan C, Fabbri D. Evaluating the effectiveness of auditing rules for electronic health record systems. AMIA Annu Symp Proc 2017;2017:866-875 [FREE Full text] [Medline: 29854153]
132. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015 Presented at: KDD'15; August 10-13, 2015; Sydney, Australia. [doi: 10.1145/2783258.2788613]
133. Che Z, St Sauver J, Liu H, Liu Y. Deep learning solutions for classifying patients on opioid use. AMIA Annu Symp Proc 2017;2017:525-534 [FREE Full text] [Medline: 29854117]
134. Ge W, Huh JW, Park YR, Lee JH, Kim YH, Turchin A. An interpretable ICU mortality prediction model based on logistic regression and recurrent neural networks with LSTM units. AMIA Annu Symp Proc 2018;2018:460-469 [FREE Full text] [Medline: 30815086]
135. Ho KC, Speier W, El-Saden S, Arnold CW. Classifying acute ischemic stroke onset time using deep imaging features. AMIA Annu Symp Proc 2017;2017:892-901 [FREE Full text] [Medline: 29854156]
136. Ming Y, Qu H, Bertini E. RuleMatrix: visualizing and understanding classifiers with rules. IEEE Trans Visual Comput Graphics 2019 Jan;25(1):342-352. [doi: 10.1109/tvcg.2018.2864812]
137. Rahman P, Jiang L, Nandi A. Evaluating interactive data systems. VLDB J 2019 Nov 13;29(1):119-146. [doi: 10.1007/s00778-019-00589-2]
138. Doan A, Domingos P, Halevy AY. Reconciling schemas of disparate data sources. SIGMOD Rec 2001 Jun;30(2):509-520. [doi: 10.1145/376284.375731]
139. Do HH, Rahm E. COMA - a system for flexible combination of schema matching approaches. In: Proceedings of 28th International Conference on Very Large Data Bases. 2002 Presented at: 28th International Conference on Very Large Data Bases; August 20-23, 2002; Hong Kong, China. [doi: 10.1016/b978-155860869-6/50060-3]
140. Chen K, Kannan A, Madhavan J, Halevy A. Exploring schema repositories with schemr. SIGMOD Rec 2011 Jul 18;40(1):11-16. [doi: 10.1145/2007206.2007210]
141. Rahm E, Bernstein PA. A survey of approaches to automatic schema matching. VLDB J . 334? 2001;10(4):350. [doi: 10.1007/s007780100057]
142. Nandi A, Bernstein PA. HAMSTER: using search clicklogs for schema and taxonomy matching. Proc VLDB Endow 2009 Aug 1;2(1):181-192. [doi: 10.14778/1687627.1687649]
143. Wang Y. Synthesizing Mapping Relationships Using Table Corpus. In: Proceedings of the 2017 ACM International Conference on Management of Data. 2017 Presented at: SIGMOD'17; May 14-17, 2017; Chicago, USA. [doi: 10.1145/3035918.3064010]
144. Ives Z, Knoblock CA, Minton S, Jacob M, Talukdar PP, Tuchinda R, et al. arXiv. 2009. URL: http://talukdar.net/papers/cidr.pdf [accessed 2020-10-14]
145. Wang J, Kraska T, Franklin MJ, Feng J. CrowdER: crowdsourcing entity resolution. Proc VLDB Endow 2012 Jul;5(11):1483-1494. [doi: 10.14778/2350229.2350263]
146. Cafarella M, Halevy A, Lee H, Madhavan J, Yu C, Wang DZ, et al. Ten years of webtables. Proc VLDB Endow 2018 Aug 1;11(12):2140-2149. [doi: 10.14778/3229863.3240492]
147. Dong XL, Halevy A, Yu C. Data integration with uncertainty. VLDB J 2008 Nov 14;18(2):469-500. [doi: 10.1007/s00778-008-0119-9]

148. Zhang CJ, Chen L, Jagadish HV, Zhang M, Tong Y. Reducing uncertainty of schema matching via crowdsourcing with accuracy rates. IEEE Trans. Knowl Data Eng 2020 Jan 1;32(1):135-151. [doi: 10.1109/tkde.2018.2881185]
149. Cafarella MJ, Halevy A, Khoussainova N. Data integration for the relational web. Proc VLDB Endow 2009 Aug 1;2(1):1090-1101. [doi: 10.14778/1687627.1687750]
150. Madhavan J, Jeffery SR, Cohen S, Dong X, Ko D, Yu C, et al. Web-scale Data Integration: You can only afford to Pay As You Go. MIT. 2007. URL: http://web.mit.edu/tibbetts/Public/CIDR_2007_Proceedings/papers/cidr07p40.pdf [accessed 2020-10-14]
151. Limaye G, Sarawagi S, Chakrabarti S. Annotating and searching web tables using entities, types and relationships. Proc VLDB Endow 2010 Sep;3(1-2):1338-1347. [doi: 10.14778/1920841.1921005]
152. Miller RJ. Open data integration. Proc VLDB Endow 2018 Aug 1;11(12):2130-2139. [doi: 10.14778/3229863.3240491]
153. Fernandez RC, Mansour E, Qahtan AA, Elmagarmid A, Ilyas I, Madden S, et al. Seeping semantics: Linking datasets using word embeddings for data discovery. 2018 Presented at: IEEE 34th International Conference on Data Engineering; October 25, 2018; Paris, France. [doi: 10.1109/icde.2018.00093]
154. He J, Veltri E, Santoro D, Li G, Mecca G, Papotti P, et al. Interactive and Deterministic Data Cleaning. In: Proceedings of the International Conference on Management of Data. 2016 Presented at: SIGMOD '16; June 2016; New York. [doi: 10.1145/2882903.2915242]
155. Bergman M, Milo T, Novgorodov S, Tan WC. Query-oriented data cleaning with oracles. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. 2015 Presented at: SIGMOD '15; May, 2015; New York p. 1199-1214. [doi: 10.1145/2723372.2737786]
156. Assadi A, Milo T, Novgorodov S. Cleaning Data with Constraints and Experts. In: Proceedings of the 21st International Workshop on the Web and Databases. 2018 Presented at: WebDB'18; June 10, 2018; Houston, TX, USA. [doi: 10.1145/3201463.3201464]
157. Fan W, Geerts F, Lakshmanan L, Xiong M. Discovering conditional functional dependencies. In: Proceedings of the 25th International Conference on Data Engineering. 2011 Presented at: International Conference on Data Engineering; March 29, 2009 - April 2, 2009; Shanghai, China. [doi: 10.1109/icde.2009.208]
158. Chiang F, Miller RJ. Discovering data quality rules. Proc VLDB Endow 2008 Aug 1;1(1):1166-1177. [doi: 10.14778/1453856.1453980]
159. Golab L, Karloff H, Korn F, Srivastava D, Yu B. On generating near-optimal tableaux for conditional functional dependencies. Proc VLDB Endow 2008 Aug 1;1(1):376-390. [doi: 10.14778/1453856.1453900]
160. Cong G, Fan W, Geerts F, Jia X, Ma S. Improving data quality: consistency and accuracy. VLDB J. 2007. URL: http://homepages.inf.ed.ac.uk/wenfei/papers/vldb07-b.pdf [accessed 2020-10-14]
161. Fan W, Geerts F. Foundations of data quality management. Synth Lect Data Manag 2012 Jul 31;4(5):1-217. [doi: 10.2200/s00439ed1v01y201207dtm030]
162. Wang DZ, Dong XL, Sarma AD, Franklin MJ. Functional dependency generation and applications in pay-as-you-go data integration systems. 2009 Presented at: 12th International Workshop on the Web and Databases; June 28, 2009; Providence, Rhode Island, USA.
163. Rammelaere J, Geerts F. Explaining repaired data with CFDs. Proc VLDB Endow 2018 Jul 1;11(11):1387-1399. [doi: 10.14778/3236187.3236193]
164. Ilyas IF, Markl V, Haas P, Brown P, Aboulnaga A. CORDS: Automatic Discovery of Correlations and Soft Functional Dependencies. In: Proceedings of the 2004 ACM SIGMOD international conference on Management of data. 2004 Presented at: SIGMOD'14; June 13-18, 2004; Paris. [doi: 10.1145/1007568.1007641]
165. Asghar A, Ghenai A. Automatic discovery of functional dependencies and conditional functional dependencies: a comparative study. University of Waterloo. 2015. URL: https://cs.uwaterloo.ca/~nasghar/848.pdf [accessed 2020-10-14]
166. De Sa C, Ratner A, Ré C, Shin J, Wang F, Wu S, et al. DeepDive: declarative knowledge base construction. SIGMOD Rec 2016 Jun 2;45(1):60-67. [doi: 10.1145/2949741.2949756]
167. Varma P, Ré C. Snuba: automating weak supervision to label training data. Proc VLDB Endow 2018 Nov 1;12(3):223-236 [FREE Full text] [doi: 10.14778/3291264.3291268] [Medline: 31777681]
168. Felix C, Dasgupta A, Bertini E. The Exploratory Labeling Assistant: Mixed-Initiative Label Curation with Large Document Collections Share on. In: Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology. 2018 Presented at: UIST'18; June 1-5, 2018; New York, USA. [doi: 10.1145/3242587.3242596]
169. Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res 2013 Jul;41(Web Server issue):W518-W522 [FREE Full text] [doi: 10.1093/nar/gkt441] [Medline: 23703206]
170. Kwon D, Kim S, Wei CH, Leaman R, Lu Z. ezTag: tagging biomedical concepts via interactive learning. Nucleic Acids Res 2018 Jul 2;46(W1):W523-W529 [FREE Full text] [doi: 10.1093/nar/gky428] [Medline: 29788413]
171. Nazi A, Ding B, Narasayya V, Chaudhuri S. Efficient estimation of inclusion coefficient using hyperloglog sketches. Proc VLDB Endow 2018 Jun 1;11(10):1097-1109. [doi: 10.14778/3231751.3231759]
172. Yuan X, Cai X, Yu M, Wang C, Zhang Y, Wen Y. Efficient Foreign Key Discovery Based on Nearest Neighbor Search. In: International Conference on Web-Age Information Management. 2015 Presented at: WAIM'15; June 8-10, 2015; Qingdao, China. [doi: 10.1007/978-3-319-21042-1_37]

173. Motl J, Kordik P. Foreign key constraint identification in relational databases. Czech Technical University in Prague. 2017. URL: http://ceur-ws.org/Vol-1885/106.pdf [accessed 2020-10-14]
174. Koehler H, Link S. Inclusion Dependencies Reloaded. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 2015 Presented at: CIKM '15; October, 2015; Melbourne Australia.
175. Chen Z, Narasayya V, Chaudhuri S. Fast foreign-key detection in Microsoft SQL server PowerPivot for Excel. Proc VLDB Endow 2014 Aug;7(13):1417-1428. [doi: 10.14778/2733004.2733014]
176. Moritz D, Howe B, Heer J. Falcon: Balancing interactive latency and resolution sensitivity for scalable linked visualizations,? University of Washington. 2019. URL: https://idl.cs.washington.edu/files/2019-Falcon-CHI.pdf [accessed 2020-10-14]
177. Kamat N, Jayachandran P, Tunga K, Nandi A. Distributed and Interactive Cube Exploration. In: 30th International Conference on Data Engineering. 2014 Presented at: ICDE'14; March 31-April 4, 2014; Chicago, IL, USA. [doi: 10.1109/icde.2014.6816674]
178. Lins L, Klosowski JT, Scheidegger C. Nanocubes for real-time exploration of spatiotemporal datasets. IEEE Trans Visual Comput Graphics 2013 Dec;19(12):2456-2465. [doi: 10.1109/tvcg.2013.179]
179. Pahins CA, Stephens SA, Scheidegger C, Comba JL. Hashedcubes: simple, low memory, real-time visual exploration of big data. IEEE Trans Visual Comput Graphics 2017 Jan;23(1):671-680. [doi: 10.1109/tvcg.2016.2598624]
180. Joglekar M, Garcia-Molina H, Parameswaran A. Interactive Data Exploration With Smart Drill-down. In: 32nd International Conference on Data Engineering. 2016 Presented at: ICDE'16; May 16-20, 2016; Helsinki, Finland. [doi: 10.1109/icde.2016.7498300]
181. Dimitriadou K, Papaemmanouil O, Diao Y. AIDE: an active learning-based approach for interactive data exploration. IEEE Trans Knowl Data Eng 2016 Nov 1;28(11):2842-2856. [doi: 10.1109/tkde.2016.2599168]
182. Psallidas F, Wu E. Smoke. Proc VLDB Endow 2018 Feb 1;11(6):719-732. [doi: 10.14778/3199517.3199522]
183. Wu E, Madden S. Scorpion. Proc VLDB Endow 2013 Jun;6(8):553-564. [doi: 10.14778/2536354.2536356]
184. Kandel S, Parikh R, Paepcke A, Hellerstein J, Heer J. Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment. Stanford Univeristy. 2012. URL: http://vis.stanford.edu/papers/profiler [accessed 2020-10-14]
185. Correll M, Li M, Kindlmann G, Scheidegger C. Looks good to me: visualizations as sanity checks. IEEE Trans Visual Comput Graphics 2019 Jan;25(1):830-839. [doi: 10.1109/tvcg.2018.2864907]
186. Wongsuphasawat K, Moritz D, Qu Z, Chang R, Ouk F, Anand A, et al. Voyager 2: Augmenting Visual Analysis with Partial View Specifications. University of Washington. 2017. URL: https://idl.cs.washington.edu/files/2017-Voyager2-CHI.pdf [accessed 2020-10-14]
187. Willett W, Heer J, Agrawala M. Scented widgets: improving navigation cues with embedded visualizations. IEEE Trans Visual Comput Graphics 2007 Nov;13(6):1129-1136. [doi: 10.1109/tvcg.2007.70589]
188. Preim B, Lawonn K. A survey of visual analytics for public health. Computer Graphics Forum 2019 Nov 28;39(1):543-580. [doi: 10.1111/cgf.13891]
189. Holzinger A, Biemann C, Pattichis CS, Kell DB. What do we need to build explainable AI systems for the medical domain? arXiv 2017:9923.
190. Montavon G, Samek W, Müller K. Methods for interpreting and understanding deep neural networks. Digi Signal Process 2018 Feb;73:1-15. [doi: 10.1016/j.dsp.2017.10.011]
191. Alspaugh S, Zokaei N, Liu A, Jin C, Hearst MA. Futzing and moseying: interviews with professional data analysts on exploration practices. IEEE Trans Visual Comput Graphics 2019 Jan;25(1):22-31. [doi: 10.1109/tvcg.2018.2865040]
192. Stoyanovich J, Howe B, Jagadish H, Miklau G. Panel: a debate on data and algorithmic ethics. Proc VLDB Endow 2018 Aug 1;11(12):2165-2167. [doi: 10.14778/3229863.3240494]
193. Cai CJ, Jongejaan J, Holbrook J. The effects of example-based explanations in a machine learning interface. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. 2019 Presented at: IUI '19; March, 2019; New York.
194. Liu D, Xu P, Ren L. Tpflow: progressive partition and multidimensional pattern extraction for large-scale spatio-temporal data analysis. IEEE Trans Visual Comput Graphics 2019 Jan;25(1):1-11. [doi: 10.1109/tvcg.2018.2865018]
195. Khan M, Xu L, Nandi A, Hellerstein JM. Data tweening: incremental visualization of data transforms. Proc VLDB Endow 2017 Feb 1;10(6):661-672.
196. Shah N, Acree ME, Patros C, Suseno M, Grant J, Fleming G, et al. A novel inpatient Antibiotic Stewardship Assistance Program (ASAP) using real-time electronic health record data, prediction modeling and epidemiologic data to provide personalized empiric antibiotic recommendations. Open Forum Infect Dis 2018 Nov;5(Suppl 1). [doi: 10.1093/ofid/ofy210.201]
197. Kamat N, Nandi A. A session-based approach to fast-but-approximate interactive data cube exploration. ACM Trans Knowl Discov Data 2018 Feb 23;12(1):1-26. [doi: 10.1145/3070648]
198. Battle L, Chang R, Stonebraker M. Dynamic prefetching of data tiles for interactive visualization. In: Proceedings of the International Conference on Management of Data. 2016 Presented at: SIGMOD '16; June, 2016; New York.
199. Takayama L, Kandogan E. Trust as an underlying factor of system administrator interface choice. In: Extended Abstracts on Human Factors in Computing Systems. 2006 Presented at: CHI EA '06; April, 2006; Montréal Québec Canada. [doi: 10.1145/1125451.1125708]

200. Idreos S, Liarou E. dbTouch: Analytics at your Fingertips. Stanford Univeristy. 2013. URL: http://www-cs-students.stanford.edu/~adityagp/courses/cs598-old/papers/dbtouch.pdf [accessed 2020-10-14]
201. Bendre M, Sun B, Zhang D, Zhou X, Chang KC, Parameswaran A. DataSpread. Proc VLDB Endow 2015 Aug;8(12):2000-2003. [doi: 10.14778/2824032.2824121]
202. Ozcan F, Koutrika G. Expressive Query Construction through Direct Manipulation of Nested Relational Results. In: Proceedings of the 2016 International Conference on Management of Data. 2016 Presented at: SIGMOD '16; June 2016; San Francisco, California.
203. Schaffer J, O'Donovan J, Michaelis J, Raglin A, Höllerer T. I can do better than your AI: expertise and explanations. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. 2019 Presented at: IUI '19; March 2019; Marina del Ray, California. [doi: 10.1145/3301275.3302308]
204. Arnold V, Clark N, Collier PA, Leech SA, Sutton S. The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. MIS Q 2006;30(1):79. [doi: 10.2307/25148718]
205. Dror IE. A hierarchy of expert performance. J Appl Res Memory Cogn 2016 Jun;5(2):121-127. [doi: 10.1016/j.jarmac.2016.03.001]
206. Mason W, Suri S. Conducting behavioral research on Amazon's Mechanical Turk. Behav Res Methods 2012 Mar;44(1):23. [doi: 10.3758/s13428-011-0124-6] [Medline: 21717266]
207. Czerwinski M, Lund A. Crowdsourcing user studies with Mechanical Turk. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2008 Presented at: CHI '08; April, 2008; Florence, Italy.
208. Sarma AD, Parameswaran A, Widom J. Towards Globally Optimal Crowdsourcing Quality Management: The Uniform Worker Setting. In: Proceedings of the 2016 International Conference on Management of Data. 2016 Presented at: SIGMOD '16; June, 2016; San Francisco, California.
209. Kandogan E, Roth M, Shwarz P, Hui J, Terizzano I, Christodoulakis C, et al. LabBook: Metadata-driven social collaborative data analysis. 2015 Presented at: 2015 IEEE International Conference on Big Data (Big Data); 2015; Santa Clara, California.
210. Hellerstein M, Sreekanti V, Gonzalez JE, Dalton J, Dey A, Nag S, et al. Ground: A Data Context Service. Conference on Innovative Data Systems Research. 2017. URL: http://cidrdb.org/cidr2017/papers/p111-hellerstein-cidr17.pdf [accessed 2020-10-14]
211. Kandel S, Paepcke A, Hellerstein JM, Heer J. Enterprise data analysis and visualization: an interview study. IEEE Trans Visual Comput Graphics 2012 Dec;18(12):2917-2926. [doi: 10.1109/tvcg.2012.219]
212. Kandogan E, Balakrishnan A, Haber EM, Pierce JS. From data to insight: work practices of analysts in the enterprise. IEEE Comput Grap Appl 2014 Sep;34(5):42-50. [doi: 10.1109/mcg.2014.62]
213. Jagdish HV, Nandi A, Qian L. Organic databases. In: Databases in Networked Information Systems. Berlin, Heidelberg: Springer; 2011:49-63.
214. Embi PJ, Yackel TR, Logan JR, Bowen JL, Cooney TG, Gorman PA. Impacts of computerized physician documentation in a teaching hospital: perceptions of faculty and resident physicians. J Am Med Inf Assoc 2004 Apr 2;11(4):300-309. [doi: 10.1197/jamia.m1525]
215. Rahman P, Nandi A. Transformer: a database-driven approach to generating forms for constrained interaction. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. 2019 Presented at: IUI '19; 2019; Marina del Ray, California.
216. Chen K, Chen H, Conway N, Hellerstein JM, Parikh TS. Usher: improving data quality with dynamic forms. IEEE Trans Knowl Data Eng 2011 Aug;23(8):1138-1153. [doi: 10.1109/tkde.2011.31]
217. Gajos K, Weld DS. SUPPLE: Automatically Generating User Interfaces. Harvard University. 2004. URL: https://www.eecs.harvard.edu/~kgajos/papers/2004/supple-iui04.pdf [accessed 2020-10-14]
218. Jayapandian N, Jagadish HV. Automated creation of a forms-based database query interface. Proceedings VLDB Endowment 2008 Aug;1(1):695-709.
219. Hebert C, Gao Y, Rahman P, Dewart C, Lustberg M, Pancholi P, et al. Prediction of antibiotic susceptibility for urinary tract infection in a hospital setting. Antimicrob Agents Chemother 2020 Jun 23;64(7):02236-19. [doi: 10.1128/aac.02236-19]

Abbreviations

CDS: clinical decision support
EHR: electronic health record
HCI: human-computer interaction
NIH: National Institute of Health
UMLS: Unified Medical Language System

Edited by G Eysenbach; submitted 24.04.20; peer-reviewed by M Afzal, A Benis, M Spiliopoulou; comments to author 25.05.20; revised version received 07.07.20; accepted 22.07.20; published 05.11.20

Please cite as:
Rahman P, Nandi A, Hebert C
Amplifying Domain Expertise in Clinical Data Pipelines
JMIR Med Inform 2020;8(11):e19612
URL: https://medinform.jmir.org/2020/11/e19612
doi: 10.2196/19612
PMID: 33151150