What Data Should I Protect? Recommender and Planning Support for Data Security Analysts

Tianyi Li
Computer Science Department, Virginia Tech
(previously: UX Team, Informatica)
Blacksburg, VA
tianyili@vt.edu

Gregorio Convertino
UX Research, Google
(previously: UX Team, Informatica)
San Francisco, CA
gconvertino@gmail.com

Ranjeet Kumar Tayi
UX Team, Informatica
Redwood City, CA
rtayi@informatica.com

Shima Kazerooni
UX Team, Informatica
Redwood City, CA
skazerooni@informatica.com

ABSTRACT

Major breaches of sensitive company data, as for Facebook's 50 million user accounts in 2018 or Equifax's 143 million user accounts in 2017, are showing the limitations of reactive data security technologies. Companies and government organizations are turning to proactive data security technologies that secure sensitive data at source. However, data security analysts still face two fundamental challenges in data protection decisions: 1) the information overload from the growing number of data repositories and protection techniques to consider; 2) the optimization of protection plans given the current goals and available resources in the organization. In this work, we propose an intelligent user interface for security analysts that recommends what data to protect, visualizes simulated protection impact, and helps build protection plans. In a domain with limited access to expert users and practices, we elicited user requirements from security analysts in industry and modeled data risks based on architectural and conceptual attributes. Our preliminary evaluation suggests that the design improves the understanding and trust of the recommended protections and helps convert risk information in protection plans.

IUI '19, March 17–20, 2019, Marina del Ray, CA, USA

ACM ISBN 978-1-4503-6272-6/19/03…$15.00
https://doi.org/10.1145/3301275.3302294

CCS CONCEPTS

• **Human-centered computing → Interactive systems and tools**;

KEYWORDS

Recommender Systems; Security Software; Data Protection; Multi-factor Decision-making; Intelligent User Interfaces; User-Centered Design.

ACM Reference format:
Tianyi Li, Gregorio Convertino, Ranjeet Kumar Tayi, and Shima Kazerooni. 2019. What Data Should I Protect? Recommender and Planning Support for Data Security Analysts. In *Proceedings of 24th International Conference on Intelligent User Interfaces, Marina del Ray, CA, USA, March 17–20, 2019 (IUI '19),* 13 pages.
https://doi.org/10.1145/3301275.3302294

1 INTRODUCTION

After the 2017 Equifax data breach impacting about 143 million users [29], Facebook reported an attack of their network system that exposed personal information of nearly 50 million users on September 28, 2018 [12]. Such cyber theft cases are topping the list of risks for which businesses are least prepared [9]. This alerts cybersecurity researchers and software providers that traditional *reactive* approaches, which are based on anti-intrusion technologies such as firewalls and digital signatures, can be circumvented and thus are insufficient to protect the application and data layers of company systems [19].

Reactive approaches are good at answering *what is attacked and where.* Models and tools are developed to detect malicious activities that have happened, such as anomalies in user activity [21, 26] and network systems [10]. However, these solutions are usually specialized for a certain type of risks. More importantly, detecting, isolating and remediating infections can take weeks in most organizations (e.g., [6]).

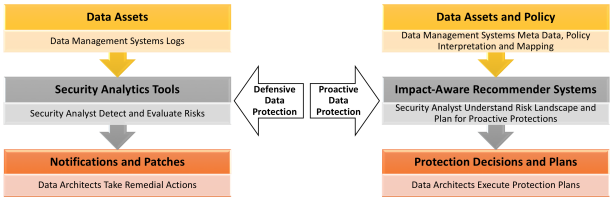

Figure 1: Data-Centric Security: re-focusing from indicators of attacks (left) to proactive data protection at source (right)

This long reaction time leaves detected vulnerabilities exposed to attackers and increases the loss. To mitigate risks before it is too late, organizations – and thus data security software vendors – are re-focusing attention on *proactive* data protection at source, i.e. data-centric security (Figure 1). They apply security techniques such as encryption to the source database from which other dependent databases and systems access the sensitive data (e.g., [1, 3, 20, 24]).

With this new focus, the problem now becomes *what to protect and how*, "proactively". This is a hard problem because making such decisions requires a good understanding of complex data risk situations – the data owned and managed by organizations are manifold and sensitive, and can be accessed, exported, and modified by different parties with varying authority. In fact, Equifax's data breach was caused by a third party company that supported their online dispute portal. In addition, the human element is often the weakest link in information security strategies, no matter how secure the system is [14, 27]. Furthermore, as governments enforce information security policies, failure of compliance will lead to legal and financial penalties plus reputation loss.

The volume and complexity of data as well as the limited budget and resources make it difficult for organizations to protect everything equally [30, 34]. Intelligent user interfaces are a natural solution to bridge the gap between the *need* to make optimal protection decisions and the *insufficient support* to manage the voluminous and complex risk information. Moreover, the same data might represent different value to different organizations and regulated in multiple ways by more than one policy. Foraging information relevant to the protection goals, understanding and verifying why certain data warrant certain protection activities are important to the quality and the efficiency of the data protection.

In this paper, we present an explainable intelligent user interface that interactively recommends and simulates protection options and carries the insights into aggregated plans. We model data risks by distinguishing the architectural and conceptual attributes and compute risk metrics from different perspectives. The system 1) recommends groups of data stores by the expected protection impact, i.e. highest risk reduction with the given budget, 2) displays the related risk factors and visualizes the simulated protection impact to explain the recommendation rationale, 3) captures user interaction to interpret latent user preference and updates recommendations accordingly. We followed a user-centered design approach [7]: we first conducted user research on needs and then ran four iterative design and evaluation cycles with target users and proxies. The evaluation feedback suggests that our system design can help analysts better understand the risk situation, convert *risk information* into *protection plans*, and adjust their protection goals when necessary.

2 RELATED WORK

In this section, we first motivate our work with the contemporary data breach incidences, how reactive security systems are no longer sufficient, and the usability gap in current data-centric security systems. We then elaborate on the key challenges of designing proactive security support, and finally describe the opportunities of IUI to address the challenges in data-centric security system design.

Proactive Detection and Information Overload

With more sophisticated hacking techniques [31] there is an increased incidence of data breaches. The Identity Theft Resource Center reported 1,579 breaches in the US in 2017, 45% percent up from 2016 [25]. This growing phenomenon and two motivators are leading organizations to adopt data security systems: avoiding business disruptions or losses due to data breaches and complying with sterner government regulations such as the General Data Protection Regulation (GDPR) (e.g., see analysis in [33]). It is common among enterprises to adopt data security techniques such as encryption, tokenization, masking, or access control to protect their sensitive data, as traditional defenses like firewalls and signature-based technologies are being circumvented by attacks aimed at the application and data layers of company systems [19].

Here we categorize the contemporary data security systems in two groups. The first group collects and analyzes user events and log data to detect anomalies or identifying malicious user activities [4, 21, 26]. The second group flags risks on sensitive data (e.g. Informatica's Secure@Source [24], IBM's QRadar Security Intelligence [1], and Imperva's SecureSphere [3]). The first group is more reactive in nature as it focuses on the footprints of previous activities. The second group of systems, which our work aims to extend, allows preemptively defining and applying security policies across data silos, thus building stronger bastions against threats. Techniques such as machine learning are also increasing the detection accuracy by replacing older rules-based and signature-based technologies.

Both groups of systems help with discovery and analytics. However, they rely on the human expert to prioritize data at risk and translate the discovered risks in protection decisions

on a case by case basis. Our work aims at supporting the analysts in this final phase, by helping with information overload, prioritization against risk metrics, and optimization of protection plans.

Risk Quantification and Protection Prioritization

It's hard to get access to expert users and practices in this domain. We are aware of only a few studies, qualitative in nature. M'manga and collaborators [2] conducted a qualitative study with ten security analysts from the IT departments of three organizations. Their interviews highlighted factors that influence risk interpretation and the overall complexity of the decision-making task. They found that the decisions to remediate vulnerabilities are conducted in constrained conditions and are based on non-standardized analysis, which they call 'folk risk analysis'. A similar interview-based study with thirty security practitioners was conducted by Werlinger and collaborators [36]. Their findings pointed to the collaborative nature of this work, with multiple stakeholders involved and the limitation of the current security systems. They argue that the systems should, for example, help more with collaboration and knowledge sharing, reduce task complexity (e.g., by supporting task prioritization), and integrate data security and communication tools into one platform.

Other qualitative analyses have argued that data protection plans can be viewed as *investment* decisions for the organization. The investment should be proportionate to the risk and lifetime of data (e.g., [11, 34, 35]). That is, not all data at risk can be protected equally: data of less value might not be worth the cost of conducting protection. Analyzing and comparing the return on the investment across protection plans is an area where future systems can help e.g., [30, 35].

In summary, a few studies pointed to several unfulfilled needs of analysts. Security analysts must decide what data to protect and with what priorities. At the same time, they need to manage multiple constraints to optimize protection plans. We are not aware of system evaluations that specifically address the decision-making aspect of data security analysts.

Intelligent UIs to Support Multi-factor Decisions

Automatic data protection systems have been easily compromised by attackers using commonly available attack vectors against known defensible vulnerabilities [22, 37]. In fact, the decision-making process of security data protection is influenced by multiple actors and factors that change over time: the organizational structure and the industry, the stakeholder who administers the data security budget, the available budget, the business priorities, and so on (e.g., [16]). A data security team relies on the effective and collaborative use of people, processes, and technology [23, 36]. Thus, the human expert must be in the loop to identify data at risk, set goals and priorities with relevant stakeholders, understand constraints (e.g., budget), and decide what data to protect.

Current systems help with risk detection but then leave it to the human to translate the overwhelming risk information into protection decisions. We argue that a user-centric design based on Intelligent User Interfaces (IUI) and mixed-initiative systems can better support analysts with such multi-factor decision-making problems. IUI can help the analyst to first set goals or, equivalently, select the relevant definitions of risks which in response help the prioritize data at risk. Future systems can support the prioritization, for example, by comparing economic models in information security investment [32], estimating returns on security investments [30, 34, 35], and methods to generate and aggregate rankings of the risks in a system (e.g., [28]).

Another reason for IUI is to help explain risk and priorities. In fact, one of the biggest challenges for data security teams is to estimate the value of data and the (negative) value of losing or disclosing it, or risk. In a survey of 37 cyber insurance experts, the European Union Agency for Network and Information Security (ENISA) found that cyber insurers and organizations face the challenge of defining risk measures [18]. Given the current basic understanding of risk, they recommend that organizations understand their risk before addressing it. For example, a classic constraint when prioritizing solutions is the limited budget for protections (e.g., see protections as investments in [11, 30, 34, 35]).

We propose applying existing IUI approaches to explain recommendations (e.g. by showing the relevant security policies, types of sensitive data, etc.) analogously to [13, 17] and assessing the expected protection impact of recommended data analogously to [8, 17]. Our work is in line with these IUI approaches. We are not aware of existing systems that have applied these approaches to reduce information overload and support multi-factor decision making by security analysts.

3 USER-CENTERED DESIGN

The above related works suggest two hard problems in making data protection decisions: 1) translating the discovered risks into appropriate protection decisions, 2) optimizing protection decisions into executable plans based on multiple factors such as plan benefits and costs and the goals of the organization. Both problems require an in-depth understanding of real-world user practice and requirements. In our two years of work with practitioners we observed the following domain-specific challenges for system design in data security:

- In the security domain, it is hard to get access to enough variety of real-world data, experts, and practices since they are deemed sensitive by organizations.

U	Industry	Size	Job Role
1	Human Res.	1.6k	Chief Info. Security Officer
2	Health Care	10k+	Data Architect
3	Education	10k+	Chief Info. Security Officer
4	Technology	1-5k	Sr. Director of Info. Security
5	Technology	10k+	Security Technical Program Mgr.
6	Finance	10k+	Data Base Adm. Manager
7	Technology	10k+	Chief Security Architect
8	Telecom.	5-10k	Security Capabilities Expert
9	Financial	200	IT Security and Compliance Mgr
10	Financial	10k+	IT Development Manager
11	Technology	1-5k	Director of Strategic Bus. Dev.

Table 1: Target users interviewed in the user research: company industry, company size (number of employees), and users' job roles.

- Data-centric security is a new domain and new market for software applications. There are no de facto successful systems yet to use as references for designers.
- Data-centric security by definition should be adaptable to the data and its users. It is hard to standardize designs across various data structures and user classes.
- Finally, enterprise software systems have long sales and deployment cycles, which limits the opportunity for fast design and evaluation iterations with new organizations who deploy a new system.

To address these challenges, our user-centered design is comprised of a year-long empirical user research with target users (Table 1) and four iterative design and evaluation cycles afterward with proxy users (Table 2). In the following subsections, we first report the methods and findings of the user research, then we report more in detail on the iterative design process and results. The four iterations were based on the user research and led to the final system design and implementation.

User Research: Goals, Pain Points, and Requirements

The goal of the user research was to understand the participants' pain points, use-cases, and expectations in data-centric security.

Method. We conducted one-on-one, semi-structured interviews with 11 target users (Security Managers or Security Analysts) at 11 companies (Table 1). Each session was remote (using WebEx and conference calling) or in-person and lasted 90 minutes. After a 5-minute introduction to explain the purpose and agenda of the interview, we asked questions about participants' day to day activities and overall responsibilities related to security products (15 minutes). Then we spent 40 minutes asking the participants their current practice regarding: 1) the overall risk assessment of the company, 2) user behavior analysis, 3) security violation alerts, and 4) policy compliance. After that, we collected more specific feedback on performing the above four tasks using [24] and how to improve them. The interviews were audio and video recorded with participant permission. We transcribed the recordings and analyzed the findings with 4 two-hour expert review sessions. We summarize the findings below.

User Personas. Quoting our target users' own terminology, they are professional security analysts who focus on identifying and prioritizing datasets based on "business risks" evaluated against the "financial asset and liability valuations", and the "required security budget". The goal is to make investment decisions for data protection "proportionate to the risk mitigation and the lifetime of a dataset" [11, p. 9].

Disconnected Tools and Lack of Intelligence Hinders Analysis. Two major limitations of current tools that constrain security analysis emerged from the interviews. First, target users wanted to see sensitive data getting discovered, analyzed, and protected in one single tool. For example, U11 stated: "One tool is ideal so that I can define one set of policies; it can [ensure data security] across the enterprise". Having to manage too many tools distracts target users from the main analysis. Such analysis work fragmented among different tools impedes full visibility of the sensitive data and the associated risk across the many data stores. A similar limitation was found by [36]. Second, multiple target users requested to have their current systems augmented with intelligence and automation. For example, U9 stated: "Ideal way would be an application or tool that is automated and as smart as possible to learn from its mistakes". U6 stated: "It would be nice to just turn the service on and it can find the data and mask it. Automation."

Top Analysis Tasks and Requirements. The interviews with target users also revealed the following top tasks that data security analysts and managers perform.

Executive tasks. Data security analysts and managers are responsible for defining security strategies and getting management buy-in for security investments. This requires them to maintain an updated understanding of the latest policies and data risk situation in the organization.

Internal Housekeeping Tasks. Data security analysts need to create information security policies, controls, and procedures to monitor the status of the data stored and managed in the organization. The goal is to ensure that the data is safe both at rest and in motion. This requires them to create internal policies that set rules for the system to detect anomalies and push notifications.

Participants		Iterations			
Job Role		*1*	*2*	*3*	*4*
Product Manager	P1	✓			
Product Manager	P2		✓	✓	✓
Sales Manager	P3		✓	✓	✓
Security Manager	P4 (U)		✓	✓	✓
Security Service Manager	P5 (U)			✓	
SW Development Manager	P6			✓	
Data Scientist	P7			✓	
Security Engineer	P8			✓	
Security Manager	P9 (U)				✓
Security Architect	P10 (U)				✓

Table 2: Participants of Each Iteration of the User-Centered Design: four are target users (marked by U), six are proxies.

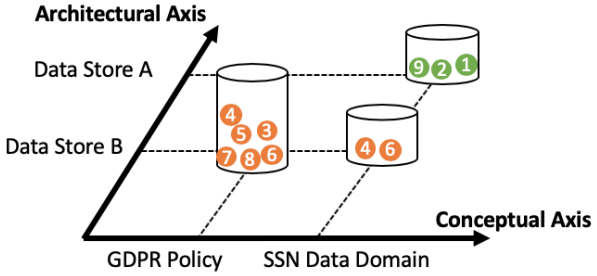

Figure 2: Cartesian space of data attributes relevant to security analysis: conceptual and architectural axes. Each colored balls with numbers represents a column in a database. Each cylinder is a group of columns that have the corresponding architectural and conceptual attributes. For example, Data Store B has 6 columns of data governed by GDPR policy, 2 containing SSN data. In Data Store A, columns 1, 2, and 9 contain SSN data.

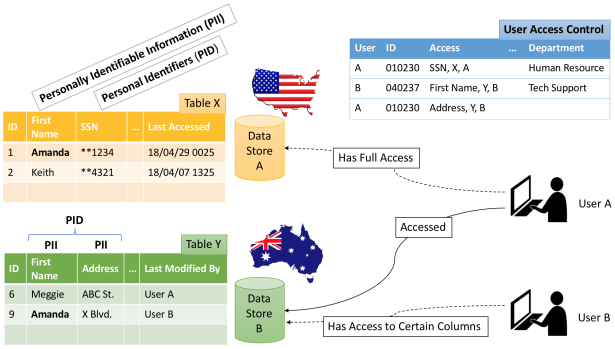

Figure 3: Example of Architectural and Conceptual Attributes of Data Units: Amanda is the first name of a customer, and other types of information of this customer are stored in multiple tables and data stores.

Policy Auditing Tasks. Besides the system-level monitoring of the data, data security analysts also need to review and analyze violations of relevant policies and standards, as well as reviewing the compliance level. Failure of compliance at the required level would lead to a considerable amount of fine and reputation damage.

Protection Management Tasks. Data security analysts would design and propose new data protection plans according to the latest risk situation. Data security managers would review, approve, and assign the protection plans to the appropriate roles to address the detected violations.

Execution Tasks. Data security technicians are responsible for implementing the processes to secure the technology infrastructure and the company data, according to the approved protection plans. This also includes managing project integration of new systems and services, as well as enabling old and new partners.

Problem Modeling and Validation

In the first iteration, drawing on the findings of the user research, we prioritized and validated the user requirements in focus and modeled the problem by separating the two major data security concerns: what to protect and how.

Method. We conducted a design workshop that involves the four authors and one proxy user. There were two sessions: 1) classifying the user requirements collected during prior investigations with target users; 2) sketching paper prototype designs to address the validated requirements and follow-up discussions. There was a 30-minute break in between.

The first half of the workshop is a one-hour requirement validation session. P1 (see Table 2) is the product manager of an existing data security system [24]. P1 has rich experience and deep understanding of the requirements and pain points of end users from different domains. The two authors that led the year-long empirical user research with target users played proxy users and relayed the findings of the user research. The team of workshop participants agreed then on the prioritization of the user requirements. The authors finally worked on the problem modeling based on information from the proxy users.

The second half of the workshop is a one-hour design session. Each of the four authors was given 30 minutes to sketch a paper prototype of a design that would address the requirements. After that, all the designs were put together and evaluated in discussions involving all the participants. In summary, the workshop allowed specifying requirements, exploring alternative design concepts, and discussing potential design choices with proxy users.

Separation of Concerns: What and How. We summarized the requirements from the user research in two main security concerns: selecting *what* data an organization should protect first, given the relevant criteria; and building protection *plans* optimized against multiple competing criteria or priorities.

Deciding how the data should be protected depends on data attributes distinct from those informing the decision whether the data should be protected. The "whether-to-protect" decisions depend primarily on the *conceptual* value people attach to data, which can be quantified as the business cost for the company if the data is lost or compromised (e.g., government penalty [45], customer lawsuits, reputation damage [32]). The "how-to-protect" decisions are constrained primarily by the *architecture* and technology used to store and manage the data. Based on the information from proxy users, we proposed the model in Figure 2 to separate concerns between these two types of constraints.

On the architecture axis, we list the attributes that describe the architecture and technology for data management. The same type of sensitive data, governed by the same policy, might live in different databases. Thus, the constraints of the architecture (platform, service) need to be accounted for. For example, different databases have different compatibility and technical support for data operations. The data protection technique Apache Sentry™can regulate user access control on Hadoop clusters, but might not be as helpful on a MongoDB database.

The conceptual axis is relevant when deciding what to protect. The architecture axis is relevant when executing the protection. The proposed model allows a practical separation of concerns among the attributes of a data unit, without worrying about the relationship among these attributes.

Validating Design Concepts and Workflow

In the second iteration, we developed and evaluated a low-fidelity prototype. The prototype shows a sidebar built as an extension of an existing data-centric security system [24]. The sidebar has two main functions: show protections (recommended or user created) and build plans.

Method. We conducted semi-structured interviews with three participants (P2-P4 in Table 2). Each interview started with a 5-minute introduction to explain the purpose and the agenda, then a general question session (20 minutes) to inquire the participant's job role, risk metrics commonly used, examples of their real risk reduction projects, and unfulfilled needs (e.g., "Given a risk reduction goal, what would you like to be recommended by the data security management system"?). In the second part of the interview (40 minutes), the participants were shown the low-fidelity prototype (Figure 4 and 5) in a Wizard of Oz manner. The participants gave feedback on each screen. The main evaluation criteria used were ease of understanding of the design components, the utility of the functions included, and whether there were any missing functions or information in the prototype.

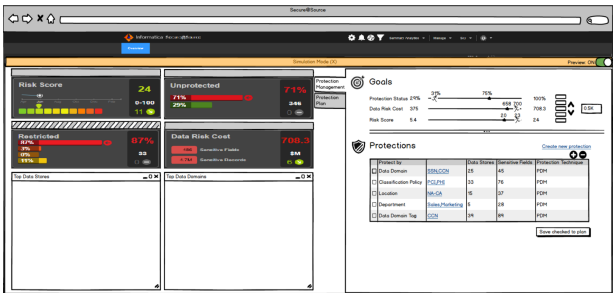

Figure 4: Low Fidelity Prototype: Recommendations on Protection Options based on the Goals Set by the User

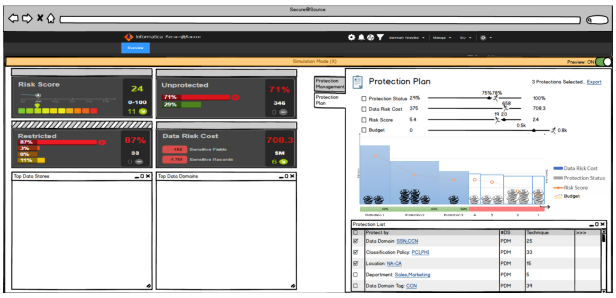

Figure 5: Low Fidelity Prototype: Optimization of Protection Plans

More Intelligent and Explainable Recommendations. A key finding was that all participants requested to see system recommendations at the outset of the task, rather than focusing on specifying their goal first. They wanted to see "what protections would have the most impact with the lowest cost and efforts" (P4), what are the "top N actions to reduce risk the most" (P3), and "what area should I focus on first" (P2). At this early stage of the analysis, they usually do not have enough knowledge and understanding of the risk situation to set a goal or manually create a protection from scratch. In addition, P2 and P3 suggested that they would like to see both current and estimated future risk metrics to analyze the impact and get a sense of progress. P2 also suggested that more visual cues will help her understand and trust system recommendations.

Reorganization of Screen Real Estate. P2 and P4 suggested the planning and optimization of protection options deserve more screen real estate to incorporate more details of the potential plans.

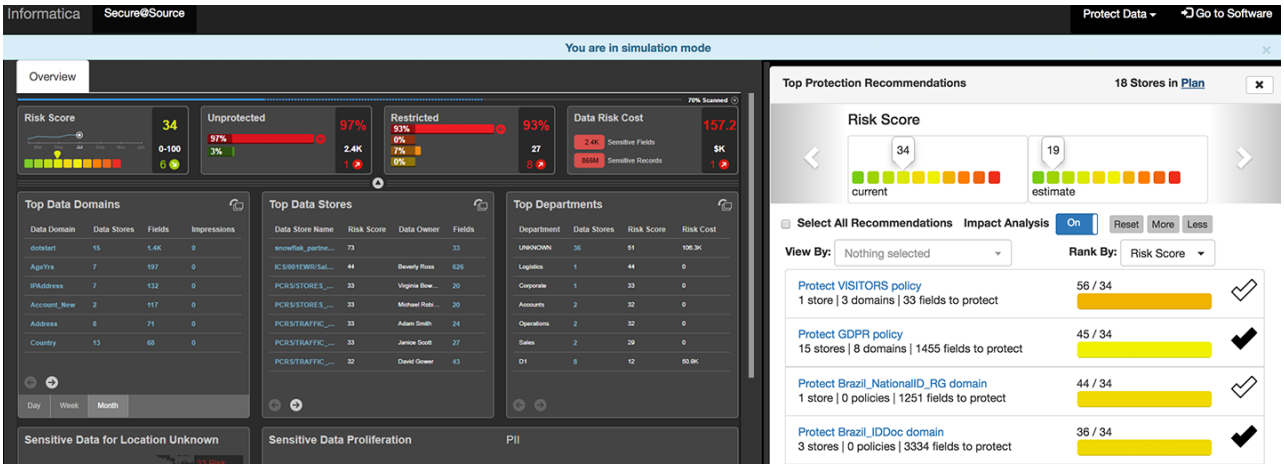

Figure 6: Recommender Sidebar Extending an Existing Data Security System [24] (See Video[5])

Evaluating Interaction Design

Based on the feedback from the first two iterations, we implemented an interactive prototype to collect further feedback from more target users during a third iteration.

Method. We evaluated the interactive prototype with the same semi-structured interview method as in the second iteration. It involved seven participants (P2-P8 in Table 2). The feedback collected revealed more specific requirements summarized in the three areas listed below.

1. Risk metric selection and ordering. P3 suggested that it would be useful to allow grouping or filtering of policies by the sensitivity level so that he could focus on policies (i.e., recommendations) of higher sensitivity level first. P3 ranked risk metrics by importance as data risk cost, protection expense, and risk score, whereas P2 ranked risk score as most important, then protection expense. She did not consider data risk cost relevant. Furthermore, what P2 really cared about is the ratio (risk score/protection expense), to assess the cost-effectiveness of protections. P3 and P4 commented that it is complicated and challenging to estimate protection expense. This is due to uncertainty about the future, the different systems deployed in each company, and non-standardized terms and concepts (e.g., some protections will have expenses in different areas that are hard to compare). However, all participants pointed out that it is important that we consider these metrics, and that showing a rough estimate of a confidence interval for each metric is also helpful.

2. Terminology and visual cues. P2 commented that the terms used in the prototype should be more intuitive and less technical so that a broader business audience could understand and benefit from the system. In this version, we had thumbs-up and thumbs-down for users to indicate if the recommendations are useful or not. P2 did not understand these cues. P4 suggested that whether a user selects a recommendation to include in the plan or not would be enough to indicate the usefulness of the recommendation.

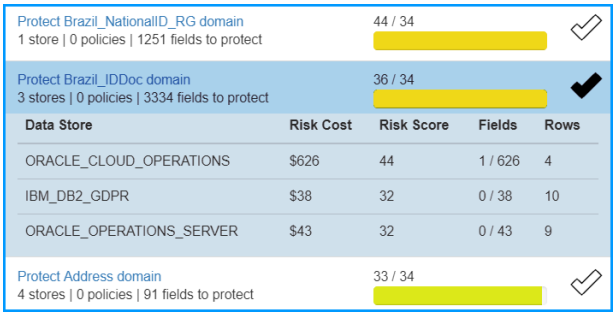

Figure 7: Recommender Sidebar: Recommendation Details

3. Interaction and transition of functionality. In this prototype, users can see the details of selected recommendations on the bottom of the bar. P4 commented that it would be more intuitive to expand the recommendation in place.

4 FINAL SYSTEM DESIGN AND IMPLEMENTATION

The final system design includes a recommender sidebar (Figure 6) and a plan building workspace (Figure 8). The intelligent UI is builds on the risk information coming from an underlying system [24], our risk modeling, and our impact-aware recommendation algorithm.

Interactive Recommendation on the Sidebar

In response to the overwhelming volume and complexity of detected risks, the system makes recommendations based on the "attributes" with the highest impact on risk metrics,

i.e. the data stores that share those attributes are the most worthy of protection.

Impact Analysis: What if I protect this. The recommendations are listed in a sidebar that extends an existing data security system (Figure 6). At the top of the sidebar is an impact analysis carousel that compares the current and expected future values of each risk metric. Clicking on the checkmarks beside each recommendation will include or remove a group of data stores in the current plan. Consequently, the expected future risk scores in the carousel will be updated to estimate the aggregated impact of all selected recommendations.

The data stores are recommended in groups based on their security policies, data domains, or other attributes on the conceptual axis in Figure 2. Each group has a title describing the grouping criterion, the number of data stores, the total number of data fields, and the expected/current risk metric values. Clicking on the title will expand and display details of the group in a tabular view (see Figure 7).

Users can slide the carousel to see different risk metrics (see "Risk Score" at the top right in Figure 6). Users can also "Select all" recommendations, turn "On/Off" the impact analysis, show "More" or "Less" groups, or "Reset" the recommendations. "View by" allows users to apply filters to further narrow down the list of recommendations. "Rank by" selects the risk metrics to rank recommendations.

Recommendation Algorithms and User Input. The recommender reads the risk information from the underlying system [24] and computes protection impact by potential changes in risk metrics. The underlying system scans the data assets in the organization and quantifies the current risk using the method described in [20].

As found in our user research, security analysts are most interested in defining protections that will bring the highest risk reduction with the given budget (highest impact). We measure the impact of a protection decision by the expected risk score reduction, protection coverage increment, expenditure on execution, and elimination of loss.

In the recommender algorithm, we implemented two ways for the user to transfer knowledge to the system. The first is at the risk factor level. Security analysts can customize risk computation by tuning weights of different risk factors, such as the governing policies of a data field, or the user activity count (see Figure 13). The second way to transfer user knowledge to the system, and more frequently, is by allowing the security analyst to select the most relevant protection from the recommendation sidebar. Data security analysts usually do not have a comprehensive understanding of the entire risk situation in the organization. Yet they can decide if protecting a recommended group of data is in line with their goals. The interface captures user interactions such as selecting, unselecting, or expanding a recommendation to interpret latent user preference. For example, if the user selects a recommendation to protect all data stores governed by GDPR policy, the weight of GDPR policy will increase, so will the data domains included in GDPR policy. The precision-recall values are computed by comparing the recommended groups of data stores, and the selected groups of data stores.

We summarize the recommendation process as follows:

(1) (Cold start) Initiate the weights of each dimension on the conceptual axis to be user-specified values (set as equal weights by default).
(2) Compute the future values of each risk metric of the fields on each conceptual dimension.
(3) Compute the rankings of dimensions by the impact on each risk metric and feed the data to the UI.
(4) Increment the weights of the dimensions related to user-selected recommendations (e.g., if the user selects a policy, then the weights of data domains governed by this policy are also incremented).
(5) Go back to Step 2 with updated dimension weights.

Plan Building Workspace

After selecting some candidate protections options, data security analysts need to build and compare different plans to analyze the expected benefits and costs.

Create and Iteratively Edit New Plan. The plan building workspace allows fine-tuning a protection plan by adding or removing individual data stores. The data stores in the current plan are ordered in a bar chart by their impact on risk reduction. For example, in Figure 9, the risk metric selected is "data risk cost", so the data stores leading to more reduction in data risk cost are on the top. The numbers beside each bar of data stores show the residual values of risk metrics after protecting the data store and those above. Users can hover over the data store names to see the exact risk reduction and other details of each data store. Selecting a different risk metric will change the ranking of the data stores. The number of stores currently in the plan is shown at the top left of the workspace. The user can also add more data stores manually or access recommendations (see Figure 10).

Users can also edit and save the details of the current plan (see Figure 9, right). This design affords future integration with coordination and collaboration functions such as creating, assigning and executing protection plans.

Review and Compare Multiple Plans. As found in our user research, data security analysts usually build multiple plans, evaluate and iteratively edit them, before putting a protection plan in execution. We extend the plan building workspace with a low-fidelity prototype that provides full-page views for the users to review the details of existing plans (Figure

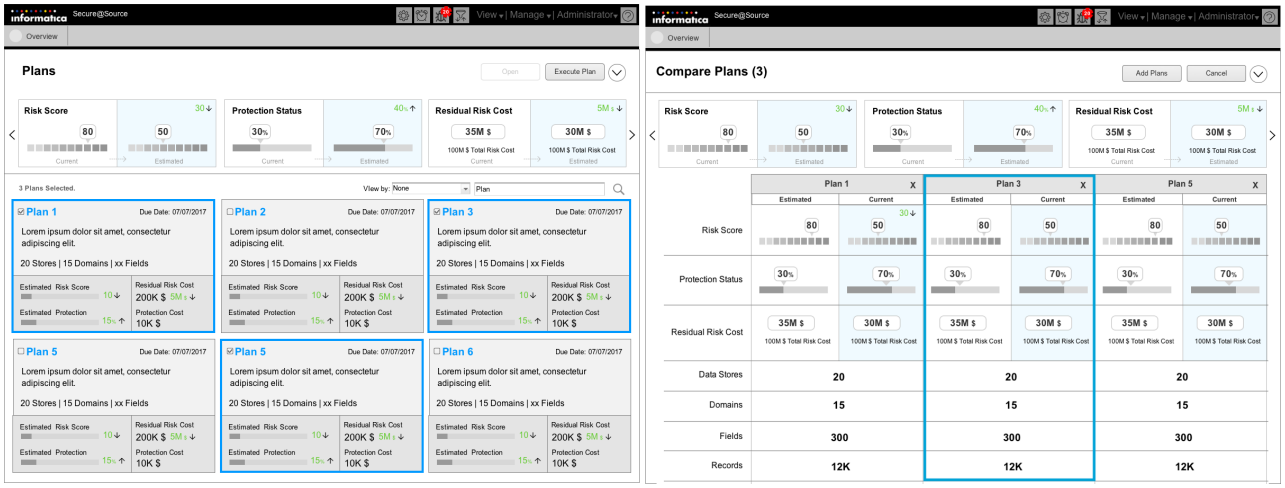

Figure 8: Plan Building Workspace - All Saved Plans (Left), and Compare Plans views (Right)

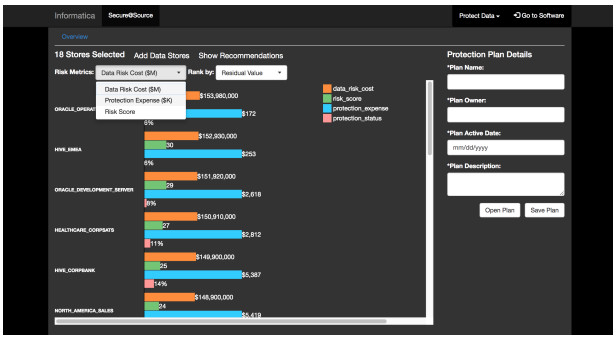

Figure 9: Plan Building Workspace: Create New Plan

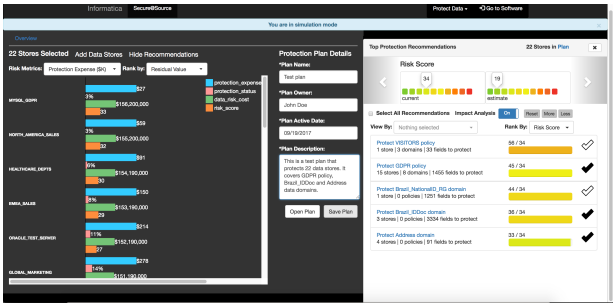

Figure 10: Plan Building Workspace: Iterative Edit and Save Plans with Details

8, left) or compare multiple plans (Figure 8, right). The top of the plan detail page shows the accumulated impact of all selected plans, with all risk metrics visible side by side. For each plan, the details of data stores are shown in a compact table card, including the current and estimated future values of risk metrics.

Data security analysts can also select several plans for in-depth comparison and evaluation. The plan comparison page displays selected plans side by side in a tabular view (Figure 8, lower right). On the top of the page is the aggregated impact of the selected plans on the entire data assets in the organization. The table shows the current and estimates future risk metric values of the data covered in each plan. Below the risk metrics the interface shows the the number of data stores, data domains, and data fields covered in each plan. Users can click on one plan (see the highlighted blue frame in Figure 8, lower right) to see the overall impact on the entire data asset; they can also drag a plan card to the left to rank it as more preferred, or to the right as less preferred. Users can add more plans in the comparison view by clicking on "Add Plans" button on the upper right of the view, or exit by clicking on "Cancel". The current plan comparison can be exported as a report for review by stakeholders and budget approval by managers. Users can iterate on plan refinement before executing the protection.

Risk Modeling

We build on the risk quantification in [20] and model the risk by separating the two major concerns: how is a data unit used and regulated (*conceptual attributes*), and how is a data unit stored in systems (*architectural attributes*) (Figure 2).

Architectural Attributes: How Data is Stored in Systems. The *Architectural Axis* encapsulates how data are stored and transferred in various systems and services. Each data unit (e.g. "Amanda" is the *First Name* of a customer in Figure 3) is stored in a (column, row) cell in some tables of some data

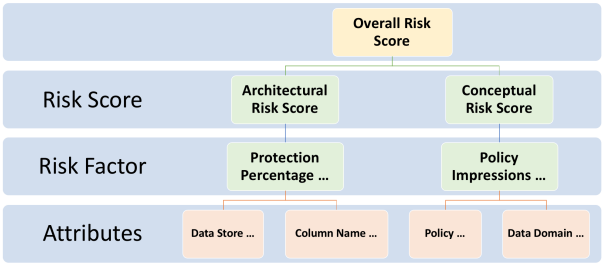

Figure 11: Hierarchy of Data Structures in the Risk Modeling

		DS1		DS2		DS3	
	Weight	Classification	Score	Classification	Score	Classification	Score
Risk Factor							
Severity	50	INTERNAL	28.13	CONFIDENTIAL	50.00	RESTRICTED	37.50
Protection Level	30	UNPROTECTED	30.00	PARTIAL	19.50	PARTIAL	19.50
Number Of Target	10	1-29	3.16	1-29	3.16	None	0.00
Data Value	10	0$	0.00	1$ - 9$	7.50	0.01 - 0.99$	5.63
Impression Metrics							
Nb Row	1	>=10,000,000	1.00	>=10,000,000	1.00	1,000,000 - 9,999,999	0.60
Nb SField	1	>= 10 sFields	1.00	>= 10 sFields	1.00	>= 10 sFields	1.00
Calculation							
Sum of Risk factor	100.00		61.29		80.16		62.63
Impression Weight	1.00		1.00		1.00		0.60
	100						
Data Store Total Score			61		80		38
Data Store Risk Score			61.3%		80.2%		37.6%
Data Store Risk Contributor			0.0%		47.2%		10.4%
Lineage Risk Score		_1 2 3 4 5 6 7 8	53.0%	_2 5 7 8	62.3%	_3	37.6%
Lineage Risk Contributor					58.3%		10.4%
Overal Risk Score			53.0%				
		Average	49.7%				
		Max	80.2%				

Figure 12: Risk Score Computation: Utility functions that formalize architectural/conceptual attribute values into scores then take the weighted sum.

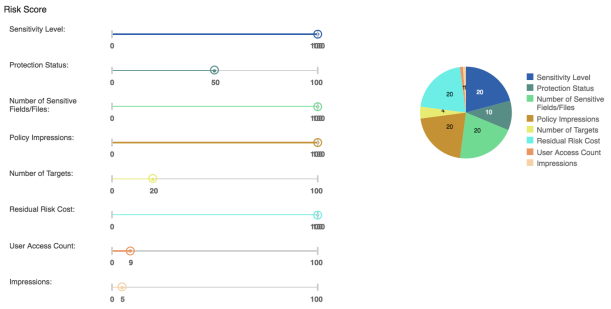

Figure 13: Risk Factor: Weights (set by users) and Values

stores. Computationally, we construct a vector to represent where a data unit is stored in data stores, its user access authority and history. Other information can be computed with these basic elements. We compute the *architectural risk score* of each data store as a weighted sum of the following four *architectural risk factors*.

(1) *Protection Percentage* measures the percentage of encrypted rows in a data store. If any data store in a lineage is compromised, all data stores in the same lineage are considered at risk and requiring protection.

(2) *Number of Targets* measures the number of data stores involved in the same data lineage as the current data store.

(3) *User Access Count* measures the number of user accounts having access to the columns in a data store.

(4) *Impressions* measures the total number of rows actually accessed by users in a data store.

We encode and normalize each factor as a score (rightmost column in Figure 12) and compute the weighted sum (weights set by users as shown in Figure 13) as the *architectural risk score* of a data store.

Conceptual Attributes: How Data is Used and Regulated. Each organization has their own implementation of data structures. For example, a customer's *First Name* might be stored in the column "First Name" in one data store, but "Given Name" in another. The *Conceptual Axis* categorizes the business value of data, such as data domains, according to standardized security regulations like EU General Data Protection Regulation (GDPR) [33]. A data domain is a category of columns in data stores. For example, as show in Figure 3, *SSN* can uniquely identify a person, thus forms a Personal Identifier (PID) data domain. Another example is given by *First Name* or *Address*, each considered Personally Identifiable Information (PII) but not sensitive enough to uniquely identify a person; yet combined, *First Name* plus *Address*, can uniquely identify a person, thus form a PID data domain.

Risk factors based on *conceptual attributes* are primarily assessed by data owners and security policies like GDPR requirements. We model *conceptual risk score* of a data store as a weighted sum of the following four *conceptual risk factors*.

(1) *Number of Sensitive Fields* measures the number of data columns governed by policies in a data store.

(2) *Policy Impressions* measures the number of rows governed by a policy in a data store.

(3) *Sensitivity Level* measures the highest sensitivity level of policies that govern the data domains in a data store.

(4) *Risk Cost* measures the unit monetary loss by a data record if the data were to be compromised, including tangible loss of policy penalty, and intangible loss like reputation damage. For proof of concept, we define risk cost as the monetary penalty per row according to policies. The risk cost of a data store is the sum of risk costs on each row.

We encode and normalize each factor (Figure 12) and compute the weighed sum as the *conceptual risk score* of a store.

5 FINAL EVALUATION

The goal of the final evaluation was to do an end-to-end assessment of the final design, including the recommender sidebar and plan building workspace, and collect new requirements on the plan building workspace.

Method

The final evaluation involved five participants (P2-P4, P9, P10 in Table 2).We conducted semi-structured interviews with the same evaluation criteria of the second and third iterations (ease of understanding, utility, and missing functions or information). In the first 40 minutes of the interview, each participant evaluated the interactive prototype (Figures 6 - 10). In the next 20 minutes they evaluated the low-fidelity prototype of the plan review and comparison pages (Figure 8) presented in a Wizard of Oz manner as an extension of the plan building workspace.

Results

All returning participants (P2-P4) were satisfied with this improved version of the design. The new participants (P9, P10) praised the ease of understanding and utility of the design components.

Recommender Sidebar Reduces Information Overload. All participants found the sidebar easy to understand and useful as a means to cope with information overload and prioritize protection options. We observed that P2 selected "Risk Score" to rank the recommendations, while P3 selected "Data Risk Cost". This echoed their prior emphasis on the importance of risk metrics. In addition, they both selected at least one other risk metrics to re-rank the recommendations, and discovered and selected the top ranked recommendations that they would have overlooked otherwise. This suggests that the recommender sidebar accommodates different analysis priorities and gives a more comprehensive view of the risk.

Iterative Plan Building Reduces Analysis Overhead. All participants applauded the seamless connection between the recommender sidebar and the plan building workspace. "So all the data stores in this view are from the previous recommendations [I selected]? That is nice." (P10) We observed P9 experimenting with different rankings of data stores, and re-opened the recommender sidebar to add more data stores. P4 deleted two data stores that require different protection techniques than the remaining data stores, "...but it would be helpful to have it [protection techniques and compatibility information] at hand, as it can get complicated and hard to keep track".

Impact Analysis Supports Multi-factor Decision-Making. Besides selecting protection options, the impact analysis function was also deemed helpful for communication among multiple stakeholders. P9 saw himself using the system for "reporting", "tell them this is how much is protected, this is what I still need to do and my plan of going forward". Understanding that the expected impact is only an estimate, P9 would use the system to estimate loss by "how much money that has been lost and this is the number of people

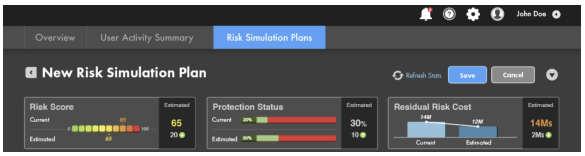

Figure 14: Impact analysis function implemented in [24]

that have been affected", to evaluate the cost of data loss in terms of "dollars per person". In doing so he would be able to understand and communicate the overall data risk cost, and use this information to decide and justify requests for budgets.

Further Requirements and Suggestions on Future Extensions. P9 pointed to more granular ways to aggregate risk information and protection plans, such as by lines of business or policy. Lines of business (i.e., the leaders of business divisions) are the stakeholders to whom P9 needs to report his analyses and suggestions. P10 suggested that other useful aggregation criteria are location, platform, and data domain (or type of data), to see for example what data domains reside in which data stores and what the risks levels are for each.

All participants appreciated the functions of reviewing plan details and comparing plans (Figure 8). P2, P3, and P4 suggested a few more extensions of features and concepts. P2 advised that the "department" would be the next most important grouping attribute needed in the design. P4 recommended that information on protection techniques should also appear in the plan detail and comparison pages of the plan building workspace. Two of our target users (P4 and P9, Table 2) highlighted the need for saving multiple alternative plans which can be vetted by the relevant stakeholders. P4 reported that his team creates about 12 plans per year and about 20-25% of them get approved for implementation. This validates the value of the design in Figure 8.

6 DISCUSSION

This paper contributes user requirements from security analysts in industry, a risk model, and an intelligent user interface design that supports decision making and plan building in data security analysis. We implemented an interactive prototype that recommends what data to protect, visualizes the expected impact of protections, and allows building and comparing protection plans.

Separation of Concerns for Iterative Analysis

Our user research suggests that current systems still have a sensemaking gap between their tools for risk detection and those for data protection decisions. The feedback from our ten participants during the iterations suggests that the proposed design can help with filling this gap and better

support decision-making around data protection. The design addresses two key challenges: reducing information overload when selecting what to protect and facilitating multi-factor decisions around protection plans.

Sensitive data scattered across data stores carries different value to different business departments of the enterprise. We use attributes of the data to measure risks (e.g., the department that owns the data, those who have access to it, the database that stores it, how the data is collected and used), make recommendations on what data should be protected, and help optimize protection plans.

Drawing on our user research findings, the design decomposes the problem into two sub-problems: *reviewing and selecting the data to protect* and *building plans for the protection*. The data attributes can then be categorized accordingly: conceptual attributes define data values and are more relevant when selecting the protection targets; the architectural attributes define data protection constraints and are more relevant when deciding how to protect the targets.

This separation of concerns allows analysts to focus on the data attributes that matter to the sub-problem at hand rather than everything at once. Beginning with conceptual attributes, analysts familiarize themselves with the system and understand the overall risk situation. Once they have a good understanding of the data risk and a reduced problem space of data worthy of protection, they turn to the architectural attributes to plan protections.

This two-step decision process can be iterative. When analysts prioritize and plan for protection execution, they can do a second round of filtering of what data to protect, further restricting or expanding the decision-making space. The Cartesian space in Figure 2 models the overlap among different conceptual attributes while maintaining the flexibility of the slice-and-dice in the analysis (e.g., a policy might govern multiple data domains, where policy and domain are overlapping conceptual attributes).

Trustworthiness and Prediction Accuracy

A limitation of the proposed intelligent UI pertains to the trustworthiness of the expected future value of each risk metric. It is a mere "estimate" based on the current risk situation in the current prototype. The expected future values are computed by simulating the protection of a set of data stores that are currently unprotected, while keeping everything else as is. In the real world, the values of other risk factors are likely to change and affect our estimates.

By the time a plan is executed, the databases configurations and data storage are usually different from when the plan was built (change in architectural attributes). Also, protecting one data store usually has a ripple effect on other data stores and related business operations (change in conceptual attributes). For example, suspending a data store that contains employee information for a week would affect tasks on that data store and those dependent on it, delaying the work and causing additional costs.

The *user access count*, the number of *departments* with access to a data store, and the *proliferation* values are shown in our recommendation and plan details. These are useful indicators for analysts to qualitatively judge the impact of the recommendation. But in the real world such impact will vary across organizations, requiring user input to make estimates more accurate and explainable. Future systems needed to 1) keep track of the changing risk situation and 2) customize recommendations and impact estimates via conversational tools that leverage input from the parties involved ([15]).

Another limitation that influences system trustworthiness is that the evaluation remained at the level of the interface design and did not include the evaluation of the recommender system with data collected from a community of users after they have adopted the system. Future work will be needed to deploy the proposed system design with a broader community of users and learn from "in vivo" user decisions as they use the system in real organizations and for enough time.

Representativeness and Generalizability

For practical reasons, the prototype was built as an extension of an existing data-centric security system. This introduces an obvious bias. However, we believe our work represents a first step towards testing and generalizing the proposed design in the context of other data security systems.

It is also important to consider that having a large number of test users for evaluating a novel system design is rarely an option in the security domain. Due to this domain-specific challenge, the evaluation of the tool was conducted with a small sample of users and can be considered a case study. Evaluations with larger samples will be needed to further validate and improve the proposed system design.

7 CONCLUSION

This paper proposed a design that uses recommendations and interactive impact analysis tools for reducing the security analysts' cognitive load and improving performance. Our target users and proxies welcomed the separation of concerns between target selection and plan building. Our evaluation confirmed the utility of applying a mixed-initiative approach to support data protection decisions. As goals and constraints change case by case, fully automated solutions appear less practical than mixed initiative solutions.

We are pleased to share that some impact analysis functions of the plan building workspace have been implemented in the underlying system [24] (see Figure 14). We hope this work can serve as a first step toward designing and testing intelligent user interfaces and mixed-imitative tools for the new field of data-centric security applications.

REFERENCES

[1] 2012. IBM QRadar. https://www.ibm.com/us-en/marketplace/ibm-qradar-siem. (2012). Latest Release: 2017-12-13.

[2] 2017. Folk Risk Analysis: Factors Influencing Security Analysts' Interpretation of Risk. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. USENIX Association, Santa Clara, CA. https://www.usenix.org/conference/soups2017/workshop-program/wsiw2017/mmanga

[3] 2018. Imperva SecureSphere. https://www.imperva.com/products/securesphere/. (2018). Latest Release: 2018.

[4] 2018. Splunk Enterprise Security. https://splunkbase.splunk.com/app/263/. (2018). Latest Release: 5.1.0 2018.

[5] 2018. Video Demo of Interactive Design Prototype. https://youtu.be/JPx4DBJSM8g. (2018).

[6] Nurul Hidayah Ab Rahman and Kim-Kwang Raymond Choo. 2015. A Survey of Information Security Incident Handling in the Cloud. *Comput. Secur.* 49, C (March 2015), 45–69. https://doi.org/10.1016/j.cose.2014.11.006

[7] Chadia Abras, Diane Maloney-Krichmar, and Jenny Preece. 2004. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications* 37, 4 (2004), 445–456.

[8] Oznur Alkan, Elizabeth M. Daly, and Inge Vejsbjerg. 2018. Opportunity Team Builder for Sales Teams. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 251–261. https://doi.org/10.1145/3172944.3172968

[9] SE Allianz. 2016. *Allianz risk barometer: Top business risks 2015*. Technical Report. Technical report, Allianz Google Scholar.

[10] Dustin L Arendt, Russ Burtner, Daniel M Best, Nathan D Bos, John R Gersh, Christine D Piatko, and Celeste Lyn Paul. 2015. Ocelot: user-centered design of a decision support visualization for network quarantine. In *Visualization for Cyber Security (VizSec), 2015 IEEE Symposium on*. IEEE, 1–8.

[11] Marc-Antoine Meunier Brian Lowans, Neil MacDonald and Brian Reed. 2017. *Predicts 2017: Application and Data Security*. Technical Report. Gartner, Inc.

[12] Carole Cadwalladr and Emma Graham-Harrison. 2018. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian* 17 (2018).

[13] Li Chen and Feng Wang. 2017. Explaining Recommendations Based on Feature Sentiments in Product Reviews. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces (IUI '17)*. ACM, New York, NY, USA, 17–28. https://doi.org/10.1145/3025171.3025173

[14] Kim-Kwang Raymond Choo. 2011. The cyber threat landscape: Challenges and future research directions. *Computers & Security* 30, 8 (2011), 719–731.

[15] Elizabeth F. Churchill. 2018. Designing Recommendations. *Interactions* 26, 1 (Dec. 2018), 24–25. https://doi.org/10.1145/3292029

[16] Daniel Dor and Yuval Elovici. 2016. A model of the information security investment decision-making process. *Computers Security* 63 (2016), 1 – 13. https://doi.org/10.1016/j.cose.2016.09.006

[17] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 211–223. https://doi.org/10.1145/3172944.3172961

[18] ENISA. 2016. *The European Union Agency for Network and Information Security (ENISA)*. Technical Report. ENISA.

[19] Paul German. 2016. Face the facts–your organisation will be breached. *Network Security* 2016, 8 (2016), 9–10.

[20] Richard Grondin and Rahul Gupta. 2017. Identifying and Securing Sensitive Data at its Source. (May 6 2017). US patent US9785795B2, granted on 2017-10-10.

[21] Bar Haim, Eitan Menahem, Yaron Wolfsthal, and Christopher Meenan. 2017. Visualizing Insider Threats: An Effective Interface for Security Analytics. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion*. ACM, 39–42.

[22] Mark Hall. 2016. Why people are key to cyber-security. *Network Security* 2016, 6 (2016), 9 – 10. https://doi.org/10.1016/S1353-4858(16)30057-5

[23] J.Todd Hamill, Richard F. Deckro, and Jack M. Kloeber. 2005. Evaluating information assurance strategies. *Decision Support Systems* 39, 3 (2005), 463 – 484. https://doi.org/10.1016/j.dss.2003.11.004

[24] Informatica. 2015. Secure @ Source. https://www.informatica.com/products/data-security/secure-at-source.html. (2015).

[25] Identity Theft Resource Center (ITRC). 2017. *2017 Annual Data Breach Year-End Review*. Technical Report. Identity Theft Resource Center (ITRC).

[26] Philip A Legg. 2015. Visualizing the insider threat: challenges and tools for identifying malicious user activity. In *Visualization for Cyber Security (VizSec), 2015 IEEE Symposium on*. IEEE, 1–7.

[27] Stephen Lineberry. 2007. The human element: The weakest link in information security. *Journal of Accountancy* 204, 5 (2007), 44.

[28] Simon Miller, Christian Wagner, Uwe Aickelin, and Jonathan M. Garibaldi. 2016. Modelling cyber-security experts' decision making processes using aggregation operators. *Computers Security* 62 (2016), 229 – 245. https://doi.org/10.1016/j.cose.2016.08.001

[29] SA O'Brien. 2017. Giant equifax data breach: 143 million people could be affected. *CNN Tech* (2017).

[30] Steve A. Purser. 2004. Improving the ROI of the security management process. *Computers Security* 23, 7 (2004), 542 – 546. https://doi.org/10.1016/j.cose.2004.09.004

[31] Sasha Romanosky, Alessandro Acquisti, and Richard Sharp. 2010. Data breaches and identity theft: when is mandatory disclosure optimal? (2010).

[32] Rachel Rue. 2007. A Framework for Classifying and Comparing Models of Cyber Security Investment to Support Policy and Decision-Making. In *WEIS*.

[33] Michael Veale, Reuben Binns, and Max Van Kleek. 2018. Some HCI Priorities for GDPR-Compliant Machine Learning. *CoRR* abs/1803.06174 (2018). arXiv:1803.06174 http://arxiv.org/abs/1803.06174

[34] Jingguo Wang, Aby Chaudhury, and H. Raghav Rao. 2008. Research NoteâĂŤA Value-at-Risk Approach to Information Security Investment. *Information Systems Research* 19, 1 (2008), 106–120. https://doi.org/10.1287/isre.1070.0143 arXiv:https://pubsonline.informs.org/doi/pdf/10.1287/isre.1070.0143

[35] Matthew P Warnecke. 2013. Examining the Return on Investment of a Security Information and Event Management Solution in a Notional Department of Defense Network Environment. (June 2013). Master's thesis.

[36] Rodrigo Werlinger, Kirstie Hawkey, David Botta, and Konstantin Beznosov. 2009. Security Practitioners in Context: Their Activities and Interactions with Other Stakeholders Within Organizations. *Int. J. Hum.-Comput. Stud.* 67, 7 (July 2009), 584–606. https://doi.org/10.1016/j.ijhcs.2009.03.002

[37] Charles Cresson Wood. 2004. Why information security is now multi-disciplinary, multi-departmental, and multi-organizational in nature. *Computer Fraud Security* 2004, 1 (2004), 45–69. https://doi.org/10.1016/j.cose.2014.11.006