Data-Driven Decisions at Scale: Managing Risk, Diversity, and Provenance

Riddho R. Haque
Supervised by Peter J. Haas & Alexandra Meliou
University of Massachusetts Amherst
rhaque@cs.umass.edu

ABSTRACT

We develop scalable algorithms that support decision-making using vast amounts of data, robustly and transparently. The wealth and availability of data open new opportunities to power and support decision-making, but also introduce challenges. Data's incredible volume and multimodality—appearing in different forms, such as text, images, or uncertain predictions—make data-driven decision-making computationally challenging. Tools used in the decision-making pipeline, including Machine Learning (ML) models and constrained optimization solvers, are complex, resource-intensive and memory-bound, resulting in poor performance and poor transparency. In contrast, we propose efficient, interpretable, and transparent in-database decision-support systems. We explore this landscape across three thrusts. First, we design methods to scale stochastic optimization to orders of magnitude more data than modern solvers can handle. Second, we augment optimization support with mechanisms to derive solutions that better represent the underlying data. Finally, we envision mechanisms for provenance auditing in ML models to enhance their transparency, and detect privacy and copyright violations during their training.

VLDB Workshop Reference Format:
Riddho R. Haque. Data-Driven Decisions at Scale: Managing Risk, Diversity, and Provenance. VLDB 2025 Workshop: PhD.

1 INTRODUCTION

The unprecedented growth in the size, availability and multimodality of data has revolutionized a broad spectrum of applications through enhanced data-driven capabilities [13], while necessitating fundamental shifts in the way systems and algorithms are designed [14, 29]. Notably, this applies to decision-making. Across a broad range of domains, including finance [11], transportation [9], manufacturing [5], and even privacy-sensitive applications such as healthcare [26] and journalism (as detailed in Section 3); decision-makers seek optimal decisions based on huge volumes of data given a complex set of interacting constraints and objectives.

However, traditional decision-making workflows are complicated, uninterpretable, and unscalable. They require slow, error-prone data movement between the database and main-memory; and involve complex interplays between the data, predictive models, and in-memory optimizers — each of which present unique challenges. Data can be voluminous, stochastic, and multimodal. Predictive models, which are used to address uncertainty and multimodality of the data, lack interpretability and transparency [17]— making it hard for users to trust decisions based on them. In-memory optimizers like Gurobi and IBM CPLEX cannot address the scalability requirements of modern data-intensive applications [3], nor do they provide mechanisms to incorporate data-centric considerations such as choosing diverse tuples in a way that better represents the underlying data.

We democratize data-driven decision-making by proposing declarative SQL-like mechanisms for users to express their decision-making problems to a database system. To solve these problems, we develop in-database methods that scalably, interpretably, and transparently derive risk-averse and diverse decisions from uncertain and multimodal data. In this work, we discuss our results, ongoing directions, and future vision across three thrusts:

- We design in-database methods for scalable stochastic optimization; example applications include portfolio optimization, internet-scale marketing, and public health. We further discuss possible extensions for multi-stage explainable decision-making. (Section 2)
- We propose augmenting optimization support with mechanisms to derive solutions that better represent the underlying data; example applications include compiling social media posts expressing diverse viewpoints. (Section 3)
- We envision mechanisms for provenance auditing in ML models to enhance their transparency, and address concerns on privacy violations during their training, with further applications to verifying machine unlearning and ML model maintenance. (Section 4)

2 SCALABLE DECISION MAKING UNDER UNCERTAINTY

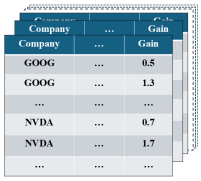
Uncertainty is inherent in many applications, as data may be sampled from predictive models or simulations. Consider an investor selecting a stock portfolio, under budgetary and risk constraints. Values of uncertain attributes such as profits from potential investments are unknown, but can be simulated via models [30]. The simulations give different scenarios of how each company's stock price may evolve. The number of scenarios needed to accurately approximate potential profits and risks typically exceeds millions [6]. The problem's scale further increases with the number of tuples—considering thousands of different holding periods for each company leads to multimillion-tuple relations. This large-scale stochastic optimization problem involves millions of scenarios over millions of tuples—far exceeding the limits of prior solvers [4].

We build scalable in-database tools for stochastic optimization [12]. Nonexperts can specify risk-constrained optimization

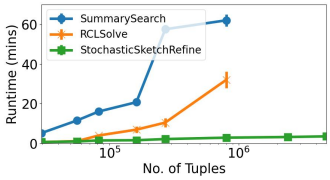
Proceedings of the VLDB Endowment. ISSN 2150-8097.

Company	Price	Sell After	Gain
GOOG	$165	0.5 day	?
GOOG	$165	1 day	?
...	...	...	...
NVDA	$114	0.5 day	?
NVDA	$114	1 days	?
...	...	...	...

(a) Stock Market Data contains stochastic attributes, e.g., gains.



(b) Models simulate numerous scenarios of possible future gains.



(c) Stochastic SketchRefine finds near-optimal portfolios from millions of tuples in minutes.

Company	Sell After	...	Multiplicity
MCO	278.5 days	...	1
EXLS	648.5 days	...	1
STBZ	341 days	...	1
HNTIF	587 days	...	2
EQS	614.5 days	...	1

(d) Its solution indicates how many of which stocks to buy and when to sell.

Figure 1: Building Stock Portfolios with Stochastic SketchRefine

problems to a database through our SQL extension, the *Stochastic Package Query Language (SPaQL+)*. Our approach treats the multiplicity of each tuple in the final solution (Figure 1d) as a decision variable, expresses the constraints and the objective in the query as functions of these decision variables, and incorporates a set of scenarios to convert the problem to an approximating *deterministic Integer Program* (IP). A particular challenge is to handle nonlinear and often nonconvex risk metrics, e.g., VaR [19] and CVaR [27], during math programming as optimization problems with nonlinear constraints require significantly higher solver runtimes [12]. We propose *Risk Constraint Linearization* (RCL), a novel technique for replacing nonlinear risk constraints by linear approximations. Our proposed algorithm, *RCLSolve*, reduces queries to approximating Integer Linear Programs (ILPs) whose solutions have a $(1-\epsilon)$-optimality guarantee [12]. While RCLSolve efficiently handles many scenarios, it struggles on large relations (Figure 1c).

We thus propose Stochastic SketchRefine [12], a divide-and-conquer framework that partitions large relations into groups of correlated tuples with similar values and stochastic properties. It first solves the problem over the representatives of each partition to obtain an intermediate 'sketch' solution, and 'refines' the sketch solution by sequentially replacing selected representatives with tuples from their partitions, using RCLSolve to solve each subproblem. By significantly reducing the number of decision variables RCLSolve needs to handle at a time, Stochastic SketchRefine can solve optimization problems with a $(1-\epsilon)^2$-optimality guarantee on multimillion tuples relations, far exceeding the limits of prior approaches [4] (Figure 1c).

Solutions generated by Stochastic SketchRefine, e.g., that of Figure 1d, prescribes single-stage decisions—buy certain stocks today and sell them after certain durations. However, many applications require **multi-stage anticipatory decision-making**. For example, if current forecasts predict stock prices to drop, the sensible decision is to defer purchases, and buy stocks at their anticipated lowest prices. Furthermore, profits from short-term investments can help build capital for future high-yield purchases that are currently beyond the user's budget. Market behaviour is subject to frequent changes [28], meaning current decisions should be amenable to future recourse. Going forward, we will thus work on **Multi-Stage Stochastic SketchRefine**, which will support multi-stage anticipatory decision-making on similarly large scales.

We further envision **generating causal explanations** behind the decisions prescribed by our algorithms. Results like that of Figure 1d may advocate for purchasing shares of relatively lesser known companies. Generating succinct explanations behind why the algorithm recommends them can enhance a user's trust on the results, and help engineers debug the scenario-generating models. Explanations may help uncover phenomena like correlation (e.g., avoiding buying positively correlated stocks to reduce risks), anticipation (e.g., deferring purchases due to anticipated price reductions), future plans (e.g., short term investments for capital building), and recourse (e.g., selling current shares for better alternatives).

Further work can explore how SPaQL+ can be modified to make our methods more accessible to non-experts, and evaluate whether LLMs can generate SPaQL+ queries from natural language text.

3 DIVERSE AND REPRESENTATIVE EMBEDDING SELECTION

Consider a news service compiling social media reactions to a hotly-discussed topic. They can use a package query [3] to select a set of tweets such that the total length and number of selected tweets stay within given thresholds, while maximizing their total views (Figure 2b). However, a simple selection of the most viewed tweets may appear repetitive and ignore unpopular viewpoints. The editor may prefer packages with more diverse tweets that include the full range of opinions (i.e., have greater coverage over the set of tuples).

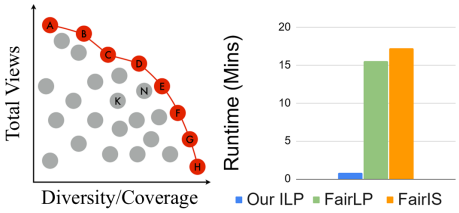
We thus want to be able to systematically trade off the objective value (total views) to achieve greater diversity and coverage of opinions (Figure 2). However, incorporating diversity and coverage constraints into our package query framework is highly nontrivial. Our approach uses high-dimensional vector representations of multimodal data such as tweets. Existing models can be used to create such vector embeddings in a way that ensures the representations of similar items are located closely in the vector space [23].

Diversity constraints can now be formalized within the framework of max-min diversification: we define the diversity of a package as the distance between its closest embeddings [1]. Given the embeddings of each data point, we can identify sets comprising points that are close together and add additional constraints specifying that only one point per set can appear in the final solution. Existing max-min diversification approaches accomplish this by either (i) forming balls around every tuple, constraining the number of tuples taken from each ball to be at most 1, and using trial and error to find the optimal ball radius [1, 18], or (ii) constructing a graph with edges between nearby embeddings [32], and requiring that no two adjacent tuples can be included in the result. These approaches add a huge number of constraints to the ILP, considerably slowing down the solver. To avoid this issue, we view constraint generation

(a) Social Media content offer diverse viewpoints on different topics.

```
SELECT PACKAGE(*) AS P
FROM Tweets REPEAT 0
WHERE Tag=`#ArabSpring`
SUCH THAT
  SUM(CHARACTERS)<=20000 AND
  COUNT(P.*)=100
MAXIMIZE SUM(Views)
```

(b) "*How the internet reacted to the Arab Spring*"—An editor wishes to compile tweets on a topic as front-page news.



(c) Highly viewed tweets will be repetitive and not cover the full range of opinions. We return pareto-optimal collections of tweets that trade-off total views for more diversity and coverage. Our technique enforces these requirements with many fewer linear constraints in ILPs than the prior art, reducing runtime significantly.

Figure 2: Scalable, Diverse and Comprehensive Tweet Selection

as a clique cover problem [16]. Consider three embedded tweets, all of which are mutually within a small distance τ of each other. Instead of redundantly adding three different constraints for each edge, we exploit the fact that they form a subclique, and add a single linear constraint, $x_1 + x_2 + x_3 \leq 1$, where x_i is the multiplicity of the i-th tuple in the relation. Solving clique-cover problems on dense networks is time-consuming [2], so we first sparsify the graph by identifying a near-minimal set of norm balls that cover nearly all pairs of neighbouring embeddings. Each ball represents a linear constraint in the ILP, so reducing their number reduces the ILP's number of constraints, which allows us to get solutions faster. Pairs of neighboring points that are not covered by a common ball then form edges in the aforementioned graph. We use existing clique cover approximators [8] to find subcliques on this significantly sparsified graph, and add one linear constraint for each subclique.

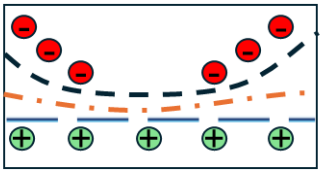
Coverage constraints require that for every embedding, at least one embedding (including itself) with distance within a threshold r from it must appear in the solution. During tweet selection, this ensures that every opinion, no matter how atypical, gets some representation in the final collection. This requirement can be naively expressed in the ILP by adding a linear constraint for each tuple requiring the solver to take at least one embedding from an r-radius ball centered around it. We formulate a novel linear transformation of this naive set of constraints, which allows coverage requirements to be expressed by far fewer linear constraints, allowing packages to be produced in much lower runtime.

Overall, our constraint-reduced ILP formulation reduces the number of constraints in the ILP and the runtime required to solve them by orders of magnitude in thousand-tuple relations (Figure 2c). We are working on methods to reduce the number of variables in the ILP to scale diverse decision-making on larger relations. In particular, we are exploring approaches to integrate our constraint reduction techniques with Progressive Shading [21], the current state of the art solver for processing traditional package queries, which can scale to billions of tuples. Integrating Progressive Shading with the additional information provided by high-dimensional embeddings and the additional requirements imposed by diversity and coverage constraints remains an open challenge.

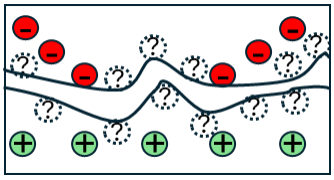
We are further investigating indexing and parallelization techniques to accelerate our approaches on vector databases. Our problem setting raises interesting avenues for creating novel vector indexing techniques. In contrast to existing vector indexes which are built to answer nearest neighbour queries approximately [24], we require an index that exactly identifies every embedding within a norm ball. To this end, we are currently formulating the L_∞-index, a sound filtration mechanism that finds neighboring embeddings in L_p metric spaces using trivially parallelizable operations.

4 DATA PROVENANCE ON ML MODELS

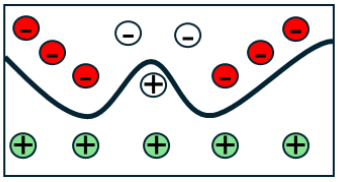
Machine learning (ML) classifiers can help further tailor the resulting packages based on user preferences. For example, they can detect hate speech in social media responses, which package query solvers can then filter out using base predicates. However, the opacity of the training history of these models raises privacy concerns. Existing approaches to address them require extensive interventions to training procedures [31], are heuristic by nature and hence inaccurate [7, 10], and/or are application-specific [22]. We thus wish to create robust tools to detect privacy infringement during model training. Given black box access to a model's inferences, and differentially private access to the set of data points T that the model trainer claims were used during its training, our envisioned algorithm determines a lower bound on the probability that the model was not trained on any of the user's 'protected' data points P. We posit that the presence of these protected data points in the training set will cause perturbations in the decision boundary that cannot be explained by the claimed training data alone (Figure 3c). To detect these perturbations, our approach will carefully issue queries to the model to determine where its decision boundary passes through (Figure 3b). Once the decision boundary's corridor is sufficiently restricted by a set of observations O, it will derive $Pr(O|M(T \cup P)) - Pr(O|M(T))$, where $Pr(O|M(S))$ indicates the probability of inferences O being observed from a model M trained on a dataset S. A large difference between the two probabilities indicates a higher chance of the model being trained on sensitive information. Inferring these probabilities is an open challenge for which we plan to explore ideas from model behaviour attribution [25], and decision boundary characterization [15]. This framework can also be modified to verify if ML models have truly unlearnt data [33] and detect when the amount of data drift justifies retraining models [20]. For further possible applications, we wish to explore model debugging, and analyzing updates to a database from an ML provenance auditing lens.



(a) An ML classifier can learn different decision boundaries from a given training dataset based on its training algorithm and hyperparameter settings.



(b) Our oracle queries the model to narrow down the passage through which the decision boundary passes.



(c) Unexpected protrusions in the decision boundary unexplained by the 'claimed' training data raise the probability of the model having access to 'hidden' data points during training.

Figure 3: Detecting if a model was trained on private/copyrighted information

5 SUMMARY

We present our work on augmenting in-database support for scalably, transparently, and explainably taking risk-averse, diverse, and representative decisions. Our work combines aspects of Operations Research, Optimization, Machine Learning, and Data Management in novel ways. We thus encourage cross-disciplinary discussions and collaborations on the open challenges that remain to realize our vision of creating fully democratized, scalable and transparent decision-making tools.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under grants 1943971 and 2211918. We are grateful to our collaborators Anh L. Mai, Matteo Brucato, Azza Abouzied, and Marco Serafini.

REFERENCES

[1] Raghavendra Addanki, Andrew McGregor, Alexandra Meliou, and Zafeiria Moumoulidou. 2022. Improved approximation and scalability for fair max-min diversification. *arXiv preprint arXiv:2201.06678* (2022).

[2] Mathieu Blanchette, Ethan Kim, and Adrian Vetta. 2012. Clique cover on sparse networks. In *2012 Proceedings of the Fourteenth Workshop on Algorithm Engineering and Experiments (ALENEX)*. SIAM, 93–102.

[3] Matteo Brucato, Juan Felipe Beltran, Azza Abouzied, and Alexandra Meliou. 2015. Scalable package queries in relational database systems. *arXiv preprint arXiv:1512.03564* (2015).

[4] Matteo Brucato, Nishant Yadav, Azza Abouzied, Peter J Haas, and Alexandra Meliou. 2020. Stochastic package queries in probabilistic databases. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 269–283.

[5] Erik Brynjolfsson and Kristina McElheran. 2016. *Data in action: Data-driven decision making in US manufacturing*. University of Toronto-Rotman School of Management.

[6] Marco C Campi, Simone Garatti, and Maria Prandini. 2009. The scenario approach for systems and control design. *Annual Reviews in Control* 33, 2 (2009), 149–157.

[7] Dami Choi, Yonadav Shavit, and David K Duvenaud. 2023. Tools for verifying neural models' training data. *Advances in Neural Information Processing Systems* 36 (2023), 1154–1188.

[8] Alessio Conte, Roberto Grossi, and Andrea Marino. 2016. Clique covering of large real-world networks. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. 1134–1139.

[9] Ugur Demiryurek, Farnoush Banaei-Kashani, and Cyrus Shahabi. 2009. Transdec: A data-driven framework for decision-making in transportation systems. (2009).

[10] Congyu Fang, Hengrui Jia, Anvith Thudi, Mohammad Yaghini, Christopher A Choquette-Choo, Natalie Dullerud, Varun Chandrasekaran, and Nicolas Papernot. 2023. Proof-of-learning is currently more broken than you think. In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*. IEEE, 797–816.

[11] RAHUL SINGH Gautam and VENKATA MRUDULA Bhimavarapu. 2022. Data driven decision making: Application in finance. *Iconic Research and Engineering Journals* 5, 12 (2022), 52–56.

[12] Riddho R Haque, Anh L Mai, Matteo Brucato, Azza Abouzied, Peter J Haas, and Alexandra Meliou. 2024. Stochastic SketchRefine: Scaling In-Database Decision-Making under Uncertainty to Millions of Tuples. *arXiv preprint arXiv:2411.17915* (2024).

[13] Hossein Hassani, Xu Huang, Steve MacFeely, and Mohammad Reza Entezarian. 2021. Big data and the United Nations sustainable development goals (UN SDGs) at a glance. *Big Data and Cognitive Computing* 5, 3 (2021), 28.

[14] Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li. 2014. Toward scalable systems for big data analytics: A technology tutorial. *IEEE access* 2 (2014), 652–687.

[15] Hamid Karimi, Tyler Derr, and Jiliang Tang. 2019. Characterizing the decision boundary of deep neural networks. *arXiv preprint arXiv:1912.11460* (2019).

[16] Richard M Karp. 2009. Reducibility among combinatorial problems. In *50 Years of Integer Programming 1958-2008: from the Early Years to the State-of-the-Art*. Springer, 219–241.

[17] Maya Krishnan. 2020. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology* 33, 3 (2020), 487–502.

[18] Yash Kurkure, Miles Shamo, Joseph Wiseman, Sainyam Galhotra, and Stavros Sintos. 2024. Faster Algorithms for Fair Max-Min Diversification in Rd. *Proceedings of the ACM on Management of Data* 2, 3 (2024), 1–26.

[19] Thomas J Linsmeier and Neil D Pearson. 2000. Value at risk. *Financial analysts journal* 56, 2 (2000), 47–67.

[20] Ananth Mahadevan and Michael Mathioudakis. 2024. Cost-aware retraining for machine learning. *Knowledge-Based Systems* 293 (2024), 111610.

[21] Anh L Mai, Pengyu Wang, Azza Abouzied, Matteo Brucato, Peter J Haas, and Alexandra Meliou. 2023. Scaling Package Queries to a Billion Tuples via Hierarchical Partitioning and Customized Optimization. *arXiv preprint arXiv:2307.02860* (2023).

[22] Vahid Majdinasab, Amin Nikanjam, and Foutse Khomh. 2024. Trained without my consent: Detecting code inclusion in language models trained on code. *arXiv preprint arXiv:2402.09299* (2024).

[23] Zach Nussbaum and Brandon Duderstadt. 2025. Training Sparse Mixture Of Experts Text Embedding Models. *arXiv preprint arXiv:2502.07972* (2025).

[24] James Jie Pan, Jianguo Wang, and Guoliang Li. 2024. Vector database management techniques and systems. In *Companion of the 2024 International Conference on Management of Data*. 597–604.

[25] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186* (2023).

[26] Keyur Patel. 2024. The Importance of Data-Driven Decision-Making in Public Health. *International Journal of Computer Trends and Technology* 72, 5 (2024), 27–32.

[27] R Tyrrell Rockafellar and Stanislav Uryasev. 2002. Conditional value-at-risk for general loss distributions. *Journal of banking & finance* 26, 7 (2002), 1443–1471.

[28] G William Schwert. 1989. Why does stock market volatility change over time? *The journal of finance* 44, 5 (1989), 1115–1153.

[29] Brian Steele, John Chandler, and Swarna Reddy. 2016. *Algorithms for data science*. Springer.

[30] Krishnaswamy Suganthi and Gopalakrishnan Jayalalitha. 2019. Geometric brownian motion in stock prices. In *Journal of Physics: Conference Series*, Vol. 1377. IOP Publishing, 012016.

[31] Zekun Sun, Zhihao Sui, Na Ruan, Conghui He, Dahua Lin, and Jie LI. 2024. Trustworthy Dataset Proof: Certifying the Authentic Use of Dataset in Training Models for Enhanced Trust. https://openreview.net/forum?id=cazOlqncU6

[32] Yanhao Wang, Michael Mathioudakis, Jia Li, and Francesco Fabbri. 2023. Max-min diversification with fairness constraints: Exact and approximation algorithms. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 91–99.

[33] Binchi Zhang, Zihan Chen, Cong Shen, and Jundong Li. 2024. Verification of machine unlearning is fragile. *arXiv preprint arXiv:2408.00929* (2024).