Counterfactual Explanation Analytics: Empowering Lay Users to Take Action Against Consequential Automated Decisions

Peter M. VanNostrand
Worcester Polytechnic Institute
pvannostrand@wpi.edu

Dennis M. Hofmann
Worcester Polytechnic Institute
dmhofmann@wpi.edu

Lei Ma
Worcester Polytechnic Institute
lma5@wpi.edu

Belisha Genin
Worcester Polytechnic Institute
bfgenin@wpi.edu

Randy Huang
Worcester Polytechnic Institute
ryhuang@wpi.edu

Elke A. Rundensteiner
Worcester Polytechnic Institute
rundenst@wpi.edu

ABSTRACT

Machine learning is routinely used to automate consequential decisions about users in domains such as finance and healthcare, raising concerns of transparency and recourse for negative outcomes. Existing Explainable AI techniques generate a static counterfactual point explanation which recommends changes to a user's instance to obtain a positive outcome. Unfortunately, these recommendations are often difficult or impossible for users to realistically enact. To overcome this, we present FACET, the first interactive robust explanation system which generates personalized counterfactual region explanations. FACET's expressive explanation analytics empower users to explore and compare multiple counterfactual options and develop a personalized actionable plan for obtaining their desired outcome. Visitors to the demonstration will interact with FACET via a new web dashboard for explanations of a loan approval scenario. In doing so, visitors will experience how lay users can easily leverage powerful explanation analytics through visual interactions and displays without the need for a strong technical background.

PVLDB Reference Format:
Peter M. VanNostrand, Dennis M. Hofmann, Lei Ma, Belisha Genin, Randy Huang, and Elke A. Rundensteiner. Counterfactual Explanation Analytics: Empowering Lay Users to Take Action Against Consequential Automated Decisions. PVLDB, 17(12): 4349-4352, 2024.
doi:10.14778/3685800.3685872

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/PeterVanNostrand/FACET.

1 INTRODUCTION

Counterfactual Explanation. Machine learning (ML) systems automate consequential decision-making in applications such as recruitment, loan approval, and policing where negative decision outcomes can have a significant impact on users' lives. As such, great attention has been placed on understanding how ML models make decisions to ensure that users are able to take action when facing negative outcomes. Indeed, regulations such as the EU GDPR and USA ECOA require explanations for certain high-stakes tasks.

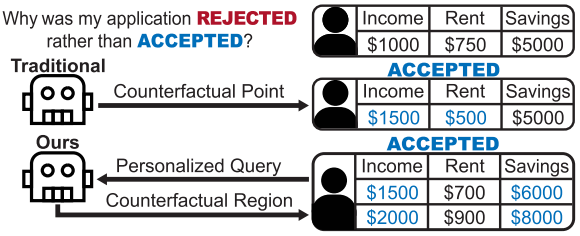


Figure 1: Example counterfactual explanation workflows

To meet this need, explainable AI (XAI) researchers have developed *counterfactual explanation* techniques that provide lay users with recourse for negative outcomes (e.g., denial of a loan) by describing alterations to an instance's features that would lead the ML model to produce a positive outcome (e.g., by suggesting a loan applicant increase their income to $1500 to obtain approval). This is traditionally done by generating a single counterfactual point as shown in Fig. 1. Given a model f which classifies an instance $x \in \mathbb{R}^n$ into the negative class $f(x) = C_N$, a counterfactual point x' is some altered version of x such that $f(x') = C_P$, where C_P is a desired positive outcome. The point x' is often selected to be as similar as possible to x according to some distance function δ [2].

Practical Limitations. While these approaches suggest alterations that are theoretically small in distance, they fail to meet users' needs in practice. First, the minimal alterations suggested are often not the best for a user, such as suggesting a loan applicant get a raise when they cannot or proposing unrealistic combinations of feature-values. Second, they place overly strict requirements to meet a precise point, such as requiring an exact dollar and cent amount in savings to be approved for a loan. This misalignment between suggested changes and users' real-world circumstances and agency renders static counterfactual point explanations of little use.

Our Approach. The research leading to this demo developed FACET [6], the first interactive XAI system that creates robust realistically actionable counterfactual explanations. Rather than single counterfactual points, FACET conceptualizes explanations as novel *counterfactual regions* which capture a contiguous portion of the feature space where a positive outcome is guaranteed (e.g., guaranteeing an income of $1500-$2000 will obtain a loan). These regions empower users to realistically obtain their desired outcome by being robust the to expected imprecision of feature alterations. Further, FACET reframes counterfactual explanation from

Proceedings of the VLDB Endowment, Vol. 17, No. 12 ISSN 2150-8097.
doi:10.14778/3685800.3685872

a metric-driven distance minimization to a user-driven exploration task, aligning with a shift towards explanation as user-supportive rather than prescriptive [5]. To provide explanations as part of a real-time dialog, FACET adopts a precompute and index approach. By analyzing the ML model, FACET intelligently generates a large number of counterfactual regions, prioritizing those with realistic feature-value combinations. These regions are efficiently stored using a custom explanation index. FACET then provides users with a set of composable, index-aware analytics operators that empower them to express their preferences and constraints to interactively search for the explanation that best fits their unique circumstances.

Demonstration. Conference participants will interact with FACET via a newly developed visual explanation interface in a web dashboard. Participants will examine automated loan approval decisions and experience how users can interactively a) identify candidate counterfactual region explanations for loan rejections, b) refine explanations by iterative parametrized querying, and c) compare between multiple what-if scenarios to select their preferred course of action for obtaining a loan. The demonstration will highlight how non-technical users can use these interactions to create highly tailored analytics queries without requiring technical knowledge.

2 THE FACET EXPLANATION SYSTEM

The FACET architecture in Fig. 2 is centered around FACET's abstraction of counterfactual region explanations (Sec. 2.1). Given a ML model, the Counterfactual Region Generator (Fig. 2 right) analyzes the model to precompute a set of counterfactual regions that cover the decision space. These regions are then indexed in the custom explanation index COREX (Sec. 2.3). At runtime, FACET's Analytics Engine leverages COREX to accelerate the processing of explanation queries (Sec. 2.3). To help users craft personalized explanations, we extend FACET for this demonstration to include a new visual explanation interface (Fig. 2 left). This interface empowers lay users to explore and refine explanations without having to directly author queries. We translate these visual interactions into FACET's explanation analytics language (Sec. 2.2), which FACET processes efficiently to produce explanations results shown via the visual display (Sec. 3). Key technical innovations are summarized below, while the FACET core backend is fully detailed in [6]. We release FACET as build-ready open source code on GitHub.

2.1 Counterfactual Region Explanations

As motivated in Sec. 1, traditional counterfactual techniques [2] create single counterfactual points which place unrealistic requirements on users. To address this, FACET introduces the flexible abstraction of counterfactual regions, which are contiguous portions of the feature space $\mathcal{X}$ guaranteed to produce the desired counterfactual outcome C_p. FACET defines a region R as a hypersphere bounded by some radius α as $R = \left\{x'_i \in \mathcal{X} \mid \delta(x', x'_i) \leq \alpha\right\}$ $s.t.$ $\forall x'_i \in R, f(x'_i) = C_P$. This provides three key properties:

- P1. *Class Homogeneity.* Every point within R is guaranteed to be a counterfactual point of the desired positive class.
- P2. *Minimum Robustness Guarantee.* Any variation v from x' will produce the desired counterfactual outcome so long as $|v| < \alpha$.
- P3. *Subset Consistency.* A region R can be divided into a smaller counterfactual region R_1 by selecting new values of x'_1, α_1.

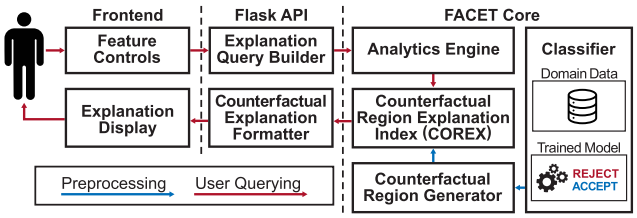


Figure 2: FACET system architecture

In practice, this means that a counterfactual region encodes a potentially infinite number of counterfactual points (P1) enabling a user to target *approximate* changes within bounds of uncertainty (P2), e.g., failing to precisely meet a savings amount. Further, P3 enables FACET to manage counterfactual regions as simple hyperrectangles bounded along each feature axis $H \subset R$ such as $H = \$1000 < income < \$1500 \wedge \$6000 < savings < \$8000$ as shown in Fig. 1. This hyperrectangular representation is easy for lay users to understand and interact with, and as we will later describe can be compactly managed and efficiently queried.

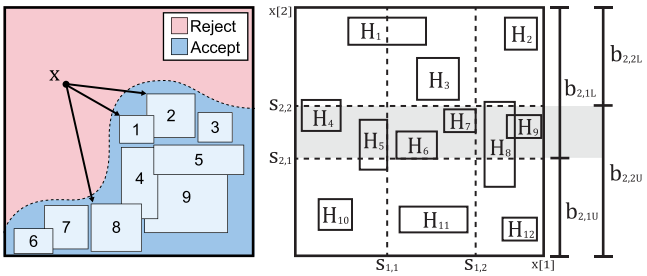
Explanation with Regions. Given a ML model, FACET uses a series of region generation algorithms [6] to smartly precompute and index (Sec. 2.3) hyperrectangular counterfactual regions. To maximize their utility, FACET's generation process prioritizes creating large regions with realistic feature-value combinations [6]. At explanation time, the user uses FACET's analytics (Sec. 2.2) to select any of the generated regions to serve as an actionable plan for altering their instance, with the guarantee that falling anywhere within the region will produce their desired outcome. Fig. 3a shows an example of this for a ML model predicting loan approval.

2.2 Personalized Explanation Analytics

Next, we describe how users can explore to find their preferred counterfactual region explanation. FACET's analytics transform the complex explanation exploration problem into a database query processing task by searching the precomputed set of regions to identify those which match the user's personalized criteria. As lay users cannot painstakingly enumerate every combination of explanation criteria they find acceptable, FACET treats this as an interactive dialog mediated by iterative query refinement. For this purpose, FACET provides a set of composable explanation analytics operators as shown in the example query below.

```
SELECT TOP 3 * FROM Regions(f) AS R
WHERE R.Class = approved
WITH UNALTERED FEATURES: R.Gender, R.Rent
WITH WHAT-IF SPECULATION
    R.Savings > 6000 AND R.Savings < 10000
    AND R.Income = 2000
ROBUST BY R.Widths >= v
ORDER BY Distance(x) WITH WEIGHTS w
```

Rather than making a specific recommendation, FACET's analytics operators let the user express their real-world limitations and preferences as predicates to support them in exploring different possible explanations. First, FACET only searches for regions of the desired class to ensure that all returned records are validly counterfactual to x. Then, the `WITH UNALTERED FEATURES` clause



(a) Counterfactual regions (b) FACET's COREX index

Figure 3: Examples of FACET's counterfactual regions used for explanation (a) and indexing of regions by COREX (b)

FACET allows the user to refine this set by excluding regions that don't match their instance on features that may be fixed (e.g., race, gender) or they feel are too difficult to change (e.g., rent). FACET's `WHAT IF SPECULATION` combines with this to enable the user to search for hypothetical cases that may be practically achievable for them. The user can further constrain the explanation search using the `ROBUST BY` clause to only consider counterfactual regions with sufficient robustness (i.e., width) along each dimension as determined by a vector $v \in \mathbb{R}^n$. The `SELECT TOP..ORDER BY` query pattern allows the user to filter to a subset of the results that require the smallest changes based on the user's relative preference for changing each feature (expressed as weights values $w \in \mathbb{R}^n$).

2.3 Indexing for Analytics Acceleration

Here, we detail FACET's custom explanation index COREX which stores counterfactual regions and accelerates the execution of analytics queries for real-time interaction. As FACET's analytics involve searching over many counterfactual regions, potentially at high dimension, and with multiple interacting constraints; analytics execution is equivalent to a high-dimensional parametrized spatial kNN search. As sequential scan is slow and existing indices tend to handle high dimensional points [4] or low dimensional spatial data [1], FACET develops a custom spatial index that maps the location of each counterfactual region within the feature space for efficient retrieval by index-aware query processing strategies.

Index Construction. To encode the location of counterfactual regions, COREX leverages their hyperrectangular representation. During the precompute phase COREX selects a set of m split values $S_{i,1} \ldots S_{i,m}$ along each feature axis i that partition the feature space to evenly divide the bounds of the regions along that axis. This creates a "grid" structure like that shown in Fig. 3b which has a roughly equivalent number of regions in each grid-cell. Then, for each split-value $S_{i,j}$ COREX constructs a pair of bitvectors that represent the location of all regions with respect to the split. For each hyperrectangular region H a bit in the bitvector $b_{i,jU}$ is set to 1 iff the upper edge of the hyperrectangle falls above $S_{i,j}$ on feature i and a bit in $b_{i,jL}$ is set to 1 iff the lower edge falls below $S_{i,j}$. This creates a set of $2m$ bitvectors per feature that map the relative location and extent of each region.

Query Acceleration. When given an instance x to explain, FACET uses COREX to determine which grid-cell x falls into. FACET then performs an index lookup by fetching and bitwise ANDing the bitvectors that correspond to the upper and lower bound of the grid-cell for each feature axis (e.g., $b_{2,1L}$ AND $b_{2,2U}$ selects the shaded area of Fig. 3b for the vertical axis between $S_{2,1}, S_{2,2}$). This produces a bitvector which is 1 for only the subset of regions that fall within that grid-cell and is quite fast due to strong optimizations for bitwise operations. FACET loads only this subset of relevant regions for evaluation against the query constraints. If needed, the search is expanded to lookup neighboring grid-cells until a suitable counterfactual region is found. Other optimizations such as jumping to specific grid-cells based on `WHAT IF` and `UNALATERED` constraints enable COREX to help refine the search space even further, and service complex high dimensional analytics in real-time.

3 DEMONSTRATION

Loan Approval Scenario. While our full system and UI implementations are readily applicable to many scenarios via a simple config file change, we will demonstrate FACET in the context of a loan approval workflow. Using a dataset from Kaggle [3], we train a ML classifier to predict loan approve/reject from applicant information. Conference participants will act as applicants by selecting an instance to explain or by entering their own values (Fig. 4) and will experience how a rejected applicant can use FACET to create a personalized actionable plan to obtain the desired approval.

Explanation Generation. Once an instance is selected, the loan applicant is brought to the main explanation dashboard in Fig. 5. *A1* displays the values and decision outcome of their application, while *A2* shows counterfactual region explanations for that decision. The number-line plots for each feature present region explanations visually, with the current application's values shown as a dot and the bounds of the counterfactual region displayed as a bar between the upper and lower bounds of the region. For example, the applicant in Fig. 5 had a Coapplicant Income of $0, but needs a Coapplicant Income of $1752-$2322 to be approved. The size of the bar ranges correspond to the *robustness* of the region and are easy for lay users to understand. To highlight the required alterations, we color-code the bars by whether or not the current application meets the counterfactual values. Applicants can explore different explanations that meet their criteria using the left and right arrows to cycle through the top k explanations. *A3* summarizes the alterations needed to meet the explanation from *A2* in natural language and acts as a simple takeaway message.

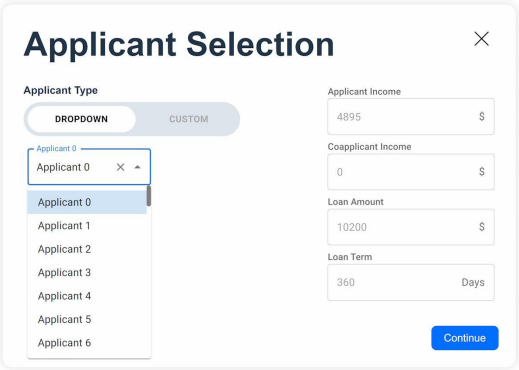


Figure 4: Instance selection with custom value entry

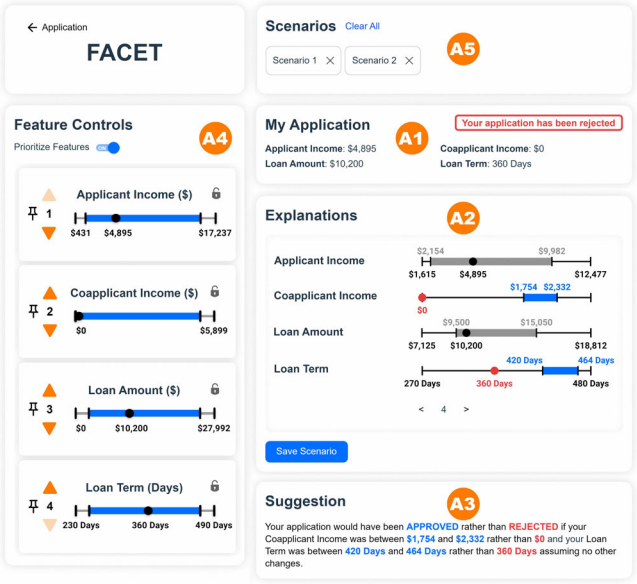


Figure 5: FACET interactive visual explanation dashboard

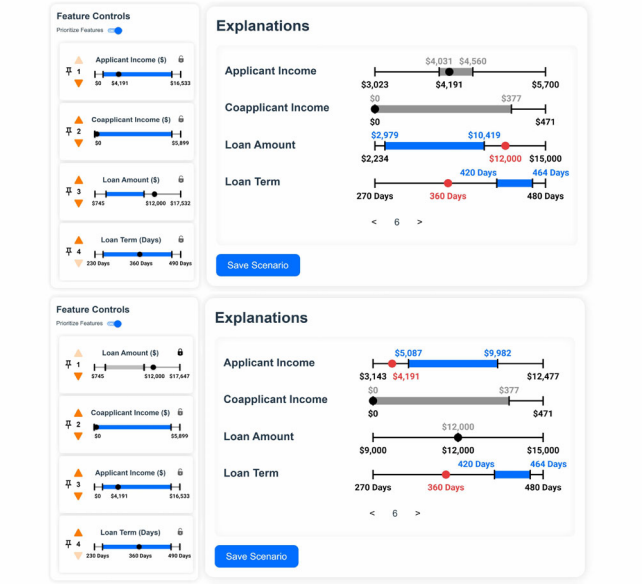


Figure 6: Comparison of two saved explanation scenarios

Personalized Refinement. To personalize the explanation search, the applicant can use the Feature Controls in *A4* (Fig. 5). By dragging the sliders on the end of the blue ranges, they can set their minimum and maximum allowable value on each feature. Clicking the lock icon will prevent explanations from altering that feature. These model the `WHAT IF` and `WITH UNALTERED` operators, respectively. The applicant can also assign relative costs to each feature by reordering the cards using the orange arrows and pins (pinning freezes the feature's place in the order). This functionality allows the user to express which constraints are most important, realizing the `WITH WEIGHTS` semantics. Whenever a feature control is updated, FACET fetches a new set of explanations that match the new criteria. By iteratively adjusting the sliders and feature priorities, the applicant can search for a counterfactual region explanation that robustly matches their real-world circumstances.

What-If Comparison. To empower users to explore alternate explanations, *A5* (Fig. 5) allows the user to save and compare different what-if scenarios. At any point during the explanation exploration process, the applicant can click the "Save Scenario" button in *A2*. This creates a new Saved Scenario which snapshots the exact state of the selected application, feature controls, explanation, and suggestion. The applicant can then continue to adjust and explore using the feature controls and return to this snapshotted state by selecting the saved scenario. Scenarios act akin to different "tabs" where the applicant can have multiple saved scenarios at a time (Fig. 6) and make changes across one or several, with changes preserved as separate paths of exploration. Switching between scenarios animates the displayed feature controls and explanations to highlight how the scenarios differ (e.g., explanations altering different features, having different constraints/priorities). These comparisons enable users to evaluate the pros/cons of different plans of action for obtaining their desired outcome and to choose the actionable changes that are right for their personal circumstances.

4 CONCLUSION

This demonstration showcases the interactive explanation system FACET. By adopting a novel counterfactual region explanation abstraction and transforming explanation generation into a query processing task, FACET is the first to generate robust easy-to-understand explanations in near real-time. Conference participants will interact with a new dashboard built for this demonstration and will experience how lay users can leverage FACET's powerful explanation analytics through straightforward UI interactions to explore, evaluate, and compare multiple explanations and ultimately identify an actionable set of steps to obtain their desired outcome.

ACKNOWLEDGMENTS

This research was supported by NSF IIS-1910880, CSSI-2103832, CNS-1852498, NRT-HDR-2021871 and DoE P200A180088. Thanks also to members of DAISY lab and the FACET MQP team.

REFERENCES

[1] Stefan Berchtold, Daniel A. Keim, and Hans-Peter Kriegel. 1996. The X-Tree : An Index Structure for High-Dimensional Data. In *Proceedings of the Twenty-second International Conference on Very Large Data-Bases*, T. M. Vijayaraman (Ed.). Morgan Kaufmann, San Francisco, 28–39.

[2] Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* 36 (2022), 1–55. https://doi.org/10.1007/s10618-022-00831-6

[3] Kaggle. 2008. Loan Predication. kaggle.com/datasets/ninzaami/loan-predication.

[4] Wen Li, Ying Zhang, Yifang Sun, Wei Wang, Mingjie Li, Wenjie Zhang, and Xuemin Lin. 2020. Approximate Nearest Neighbor Search on High Dimensional Data — Experiments, Analyses, and Improvement. *IEEE Transactions on Knowledge and Data Engineering* 32, 8 (2020), 1475–1488.

[5] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 333–342.

[6] Peter M. VanNostrand, Huayi Zhang, Dennis M. Hofmann, and Elke A. Rundensteiner. 2023. FACET: Robust Counterfactual Explanation Analytics. *Proc. ACM Manag. Data* 1, 4, Article 242 (dec 2023), 27 pages. https://doi.org/10.1145/3626729