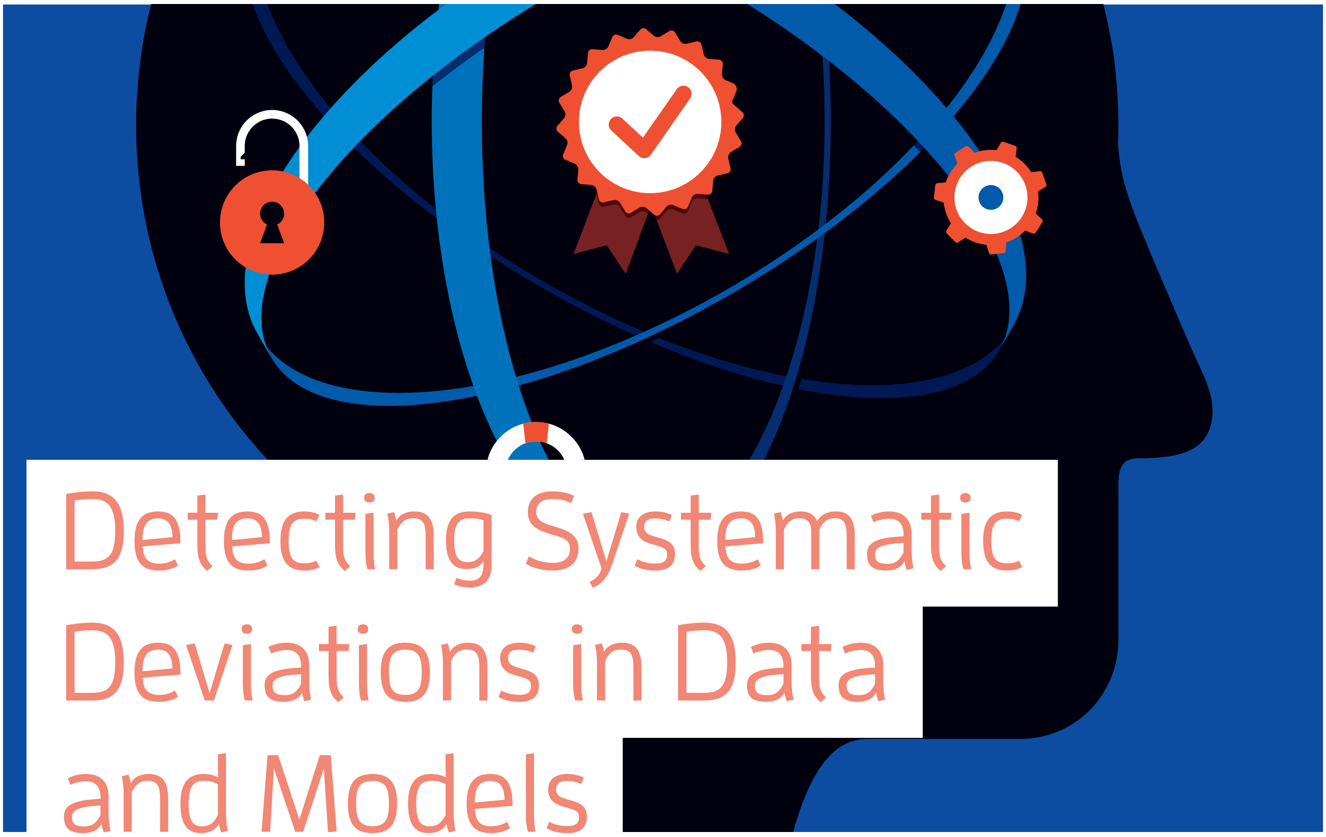

Detecting Systematic Deviations in Data and Models

Skyler Speakman, Girmaw Abebe Tadesse, Celia Cintas, William Ogallo, Tanya Akumu, and Adebayo Oshingbesan, IBM Research Africa

Trustworthy artificial intelligence researchers should seek to better detect and characterize systematic deviations in data and models (that is, bias). This article provides data scientists with motivation, theory, code, and examples on how to perform disciplined discovery of systematic deviations in data and models at the subset level.

Data scientists and decision-makers seek to understand their data at the *subset* level. This subset-level of understanding can reveal anomalous patterns relevant to trustworthy artificial intelligence (AI), such as under/over representation in data, predictive model bias, and distribution shifts. Critically, these patterns may not be evident at a global (all records) level or at individual-record levels. For example, knowing that approximately 24% of census respondents make more than US$50,000 per year provides no information about how that outcome is distributed across age, education, or employment levels. Furthermore, knowing that a 44-year-old married woman who works 24 h a week makes less than US$50,000 per year contains little information on how that generalizes to larger groups.

To bridge this gap between the global and individual levels of understanding, data scientists will often stratify an outcome of interest across *individual* features

Digital Object Identifier 10.1109/MC.2022.3213209
Date of current version: 8 February 2023

such as age, gender, ethnicity, or education. Many dashboards-like solutions perform these "cross tab" tasks to better understand how the outcome of interest is distributed across the predetermined subsets. If this manual process reveals that one of the predetermined subpopulations has outcomes that deviate from its expectation, then researchers may invoke a claim of bias or shift in the data or model. Although well intentioned, these researchers are not appropriately accounting for the true scale of scanning over subsets of data and models, which means they may miss the true bias in their data or be unable to statistically defend the deviation that they found manually. Therefore, the goal of this trustworthy AI article is to provide data scientists with motivation, theory, code, and examples on how to perform disciplined discovery of systematic deviations in data and models at the subset level.

Anomalous subset discovery is challenging for two primary reasons. The first reason is that there are exponentially many subsets to consider. Even moderate-sized datasets will contain trillions of possible subsets to investigate for anomalous deviations away from expectations. In the face of such an insurmountable task, investigators will often simplify their search to a tiny fraction of subsets which hinders the larger goal of discovery. Nobel prize winner Herbert Simon studied this human bias called *satisficing* and bounded rationality in the 1970s.

> "Decision makers can satisfice either by finding optimum solutions for a simplified world, or by finding satisfactory solutions for a more realistic world."[16]

Trustworthy AI researchers studying fairness and bias fall into this exact same tension 50 years later. They often perform a search over a simplified world looking for a satisfactory solution because maximizing over the exponentially large space is difficult.

The second common issue of working with data at the subset level, is the statistical challenge of claiming significance of the detected pattern due to extreme multiple hypothesis testing. To borrow a phrase, "if you torture your data enough, you can get it to say anything." Undisciplined (manual) searching over subsets creates a quagmire of false discoveries and unreproducible results.

To address these concerns for trustworthy AI applications, this article uses techniques from a growing body of work known as *subset scanning*.[13,18] Subset scanning exploits mathematical properties of commonly used measures of divergence called *scoring functions* (for example, likelihood ratios) that allow for exact and efficient maximization over subsets of feature values. Critically, this property allows discovery and understanding of data at the subset level while overcoming the computational complexity and statistical challenges associated with exponentially many subsets to consider. Subset scanning ideas originated at Carnegie Mellon University in the 2000s out of the scan statistics literature with a focus on epidemiology and disease surveillance. However, the work is now expanding with connections to information theory, rule-mining, predictive bias and fairness,[20] out-of-distribution detection,[5] distribution shifts,[9] and causal discovery.[11,14]

Before continuing into more details of subset scanning and its applications to trustworthy AI, we conclude this introduction with a motivating example of identifying anomalous subpopulations in one of the most recognized datasets in machine learning and data science: the Adult census data.[10] Subset scanning discovered novel subsets of the data with extreme under (or over) representation of the outcome of interest.

The binary target outcome in Adult is whether the individual had an annual income exceeding US$50,000. Approximately 23.6% of the records have this outcome (see Figure 1). Trustworthy AI research may be concerned with how this outcome is distributed across age, gender, race, education, and many other features collected as part of the census.

Scanning identified the following subset of records: All individuals who had no capital gains and had one of the following four relationship status: "not-in-family," "other-relative," or "own-child," or "unmarried." This subset contains 52% of the records but only 10% of the individuals exceeding US$50,000 income. Given the outcome of interest is Y and a subset in a given data are defined as S and probability is defined as P, another way of stating this result is that the global (marginal) probability of the outcome is $P(Y = 1) = 0.236$ but this subset has a conditional probability of $P(Y = 1 \mid S) = 0.047$. Only 4.7% of the individuals in this group made more than US$50,000, and this group represents more than half of the data! This large divergence between $P(Y)$ and $P(Y \mid S)$ can also be viewed as the amount of *information* that knowing S provides about the outcome Y. More details on the scoring functions optimized in subset scanning are provided next.

Subset scanning performs this efficient search over trillions and trillions of possible subsets in just a few seconds using standard personal computers (no GPUs required). The runtime for the multidimensional subset scan (MDScan algorithm) on the Adult data set is summarized in Figure 2. The runtime is reported against the size of the search space measured by the number of features

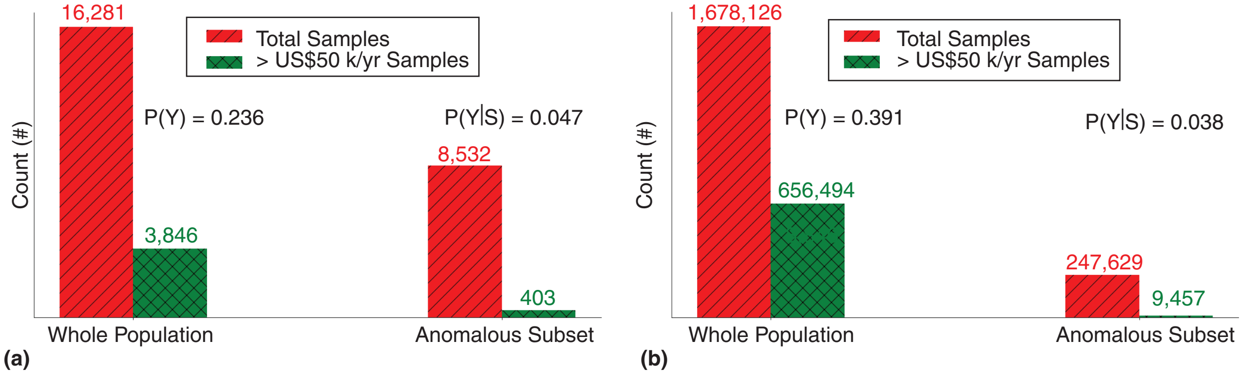

FIGURE 1. The proportion of adults making greater than US$50,000 and the size and proportion of the same outcome in the subset identified by subset scanning in (a) Adult and (b) Folktable datasets. Only 4.7% and 3.8% of the adults in the anomalous subsets of Adult and Folktables, respectively, have the outcome compared to the 23.6% and 39.1% expected from the corresponding whole population.

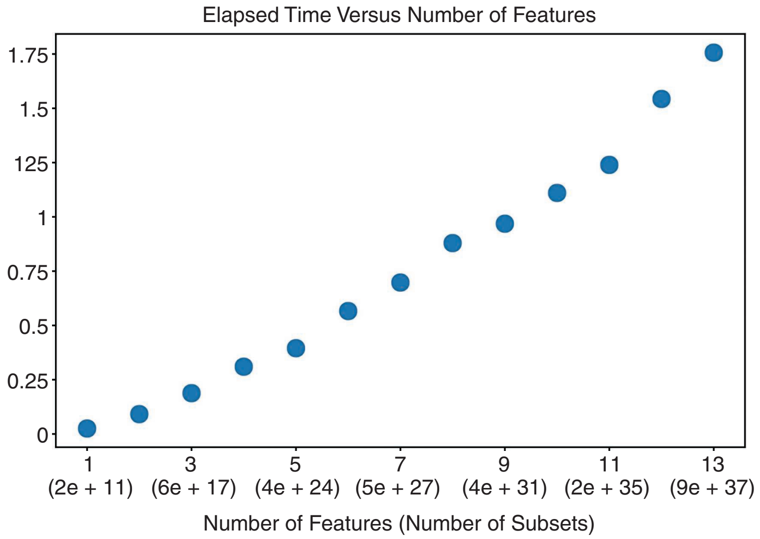

FIGURE 2. The average runtime for scanning over the Adult dataset. One hundred scans were performed over different numbers of randomly selected features in the search space, and the average is reported. In practice, multiple random restarts of the iterative ascent procedure are used to approach the global maximum and these can be done in parallel. Features with high numbers of unique values like "native-country" and "occupation" increase this search space dramatically.

and the number of subsets with the latter growing exponentially. We encourage interested readers to try the publicly available code and notebooks (https://github.com/Trusted-AI/AIF360/blob/master/examples/demo_mdss_detector.ipynb) to recreate some of the results for themselves. The code is available in the open source AI-Fairness 360 tool kit.[3]

SCORING FUNCTIONS FOR MEASURING DEVIATIONS

Measures of deviation such as cross-entropy and likelihood ratios form the basis of many statistical machine learning tasks. They quantify divergence using probabilistic or information-theoretic founded assumptions. The goal of subset scanning in the context of detecting bias (and trustworthy AI more broadly) is to identify subsets of data and models where these measures of deviation are *maximized*.

Previous work has shown that commonly used measures of divergence (scoring functions) satisfy the (additive) linear-time subset scanning property (ALTSS).[13,18] A feature containing k unique values (such as 15 values in "occupation" from the Adult data set) contain $O(2^k)$ possible combinations of feature values. For scoring functions that satisfy the ALTSS property, the vast majority of these subsets are provably suboptimal and cannot be the highest scoring (most anomalous) subset. In fact, only linearly many, $O(k)$, subsets have the potential to be the highest scoring. The ALTSS property of scoring functions decreases the search space from exponential to linearly many subsets to consider while still guaranteeing that the most divergent (highest scoring) subset of the $O(2^k)$ possible subsets will be identified. This fundamental property of scoring functions is then used as a building block to enable more sophisticated search algorithms such as the multidimensional subset scan (MDScan)[13] described next.

Table 1 lists the measures of deviation (scoring functions) referenced in this article. The first two rows assume all records have the same expectation for the outcome Y. When the expectation-based binomial scan statistic (first row) is evaluated at its maximum likelihood estimate (q_{mle}), then it is equivalent to a directed version of the J-Measure (second row). The third row is different in that it allows the expectation to vary for each individual record. More details on this distinction are provided next.

The expectation-based binomial scan statistic satisfies ALTSS.[18] This scoring function is a likelihood ratio based on the binomial distribution. It identifies anomalous subsets where the number of $Y = 1$ outcomes in a subset S shows evidence of being increased by a multiplicative factor q, (sometimes referred to as a relative risk) above the expected count. More specifically, the highest scoring subset is one that shows the most evidence of $q > 1$.

The second scoring function in Table 1 is a directed version of the information-theoretic J-measure.[17] J-measure quantifies the average information about the outcome Y when conditioning only on the records in subset S. It is the size of the subset times the cross entropy or Kullback–Leilbler divergence between the global (marginal) probability of the outcome Y and the probability of the outcome Y conditioned on the subset S. We make the additional assumption that the positive directed version of the J-Measure is 0 whenever $P(Y \mid S) < P(Y)$. This allows it to only detect *increases* in the outcome of interest. Scanning in the negative direction is also possible. This directed version of J-Measure proposed here satisfies the three desired properties of a measure;[15] however, the original *un*directed J-measure does not.

One may derive an alternative form of these two scoring functions using terminology from classification rule mining. These scoring functions are finding the subset S that maximizes the correct balance of the support[1] and lift[4] of the rule ($S \rightarrow Y$) as shown as follows:

TABLE 1. The scoring functions - is to find the subset S that maximizes the divergence between the expected outcomes in S and observed outcomes in S.

Scoring functions (measures of deviation at subset level)		
Name		**Derivation in observed outcome y_i and expected outcome, p or p_i.**
Expectation-based binomial scan statistic[18]	$\max_{q>1} \log \prod_{i \in S} \frac{\text{Binomial}(\lvert S \rvert, q \cdot p)}{\text{Binomial}(\lvert S \rvert, p)}$	$\max_{q>1} \sum_{i \in S} y_i \ln(q) + (1 - y_i) \cdot \ln\left(\frac{1 - q \cdot p}{1 - p}\right)$
Directed J-Measure from information theory[17]	$\lvert S \rvert \cdot KL(P(Y \mid S), P(Y))$	$\lvert S \rvert \cdot KL\left(\sum_{i \in S} y_i / \lvert S \rvert, p\right)$
Expectation-based Bernoulli scan statistic (bias score)[20]	$\max_{q>1} \log \prod_{i \in S} \frac{\text{Bernoulli}\left(\frac{qp_i}{1 - p_i + qp_i}\right)}{\text{Bernoulli}(p_i)}$	$\max_{q>1} \sum_{i \in S} y_i \cdot \ln(q) - \ln(1 - p_i + q \cdot p_i)$

P(Y) = *p*: the expectation (mean) of the outcome over the entire data; p_i: the expectation (predicted probability of the outcome from some model) for a single record, *i*;. *KL*(.,.): Kullback–Leibler divergence.

$$\mathrm{Supp}\,(S \rightarrow Y) \cdot \ln\,(\mathrm{Lift}\,(S \rightarrow Y)) + \mathrm{Supp}\,(S \rightarrow \neg Y) \cdot \ln\,(\mathrm{Lift}\,(S \rightarrow \neg Y)). \quad (1)$$

The third scoring function in Table 1 is the expectation-based Bernoulli scan statistic, also known as the *bias score*.[20] This scoring function differs from the previous ones in two ways. First, it allows the expectation of the outcome to vary per record and does not rely on the global mean of the outcome. This makes it particularly useful in scanning over predictions made by classification models that were trained on separate data (see "Setting Expectations" section). Second, the previous two scoring functions look for an increase in the *number* of observed outcomes in a subset (as compared to the expected number) whereas bias scan looks for an increase in the observed *odds* of the outcome in a subset compared to the expected odds.

The scoring functions in Table 1 are written to detect an *increase* in the outcomes. These functions can also scan for *decreases* by maximizing over $0 < q < 1$ instead of $q > 1$ and defining the Directed J-Measure to be 0 whenever $P(Y \mid S) > P(Y)$.

ALGORITHM 1: PSEUDOCODE FOR MDSCAN

```
1 # Input and output definition;
  input:  Dataset:
          $D = \{(x_i, y_i) | i=1,2,\ldots,N\}$,
          Set of features:
          $F = [f_1, f_2, \ldots, f_m, \ldots, f_M]$
  output: AnomSubset,
          AnomScore
2 # Initialization;
3 AnomSubset ← {};
4 AnomScore ← −∞;
5 UnCheckedF ← F;
6 #Iterate until convergence;
7 while UnCheckedF isNot {} do
8     # Randomly select unchecked feature;
9     $f_m$ ← Random(UnCheckedF);
10    # Mark the feature as checked;
11    UnCheckedF ← UnCheckedF \ $f_m$;
12    # Compute the anomalous score;
13    Score, Subset ← ALTSS
      ($f_m$ | AnomSubset);
14    # Compare the new score with previous best;
15    if Score > AnomScore then
16        # Update the score, subset and reset the flag to unchecked;
17        AnomScore ← Score;
18        AnomSubset ← Subset;
19        UnCheckedF ← F;
20    else
21        Go to Step 4;
22 # Return the most anomalous score and its subset;
23 return AnomSubset, AnomScore
```

Finally, we emphasize that this is far from an exhaustive list of scoring functions that satisfy the ALTSS property. Likelihood ratios from other members of the exponential family as well as many nonparametric scan statistics can be maximized efficiently over subsets of feature values.

SCALING SCANNING FROM SINGLE TO MULTIPLE DIMENSIONS

MDScan[13] exploits the additive linear-time subset scanning property of scoring functions to efficiently scan for anomalous subsets spanning multiple features. The additive linear-time subset scanning property[12,18] allows exact and efficient maximization of scoring functions over a *single* feature by only considering linearly many subsets of its feature values. Using boolean logic terms, these subsets of feature values from a single feature represent the "or" operator such as occupation: {tech-support *or* prof-specialty}. In contrast, a subset spanning two or more features represents the "and" operator between features. An example of this would be relationship status: {unmarried *or* not-in-family} *and* occupation: {tech-support *or* prof-specialty}. More generally, subsets of this form are said to be in conjunctive normal form which are "ANDs of ORs" and these subsets form the exponentially large search space when scanning over multiple features.

MDScan is an iterative ascent procedure where each step is efficient and exact due to the ALTSS property of the scoring function being maximized (see Algorithm 1 for pseudocode). Each step scans over an individual feature's values to determine if there is a large divergence between the observed and expected outcomes for some subset of the feature values. Each time a new high-scoring subset is found in the ascent then all features are rescanned. This is in stark contrast to greedy tree-based methods that do not reconsider previous splits. The ascent continues until *no single change to the subset will increase its score*. At this point the ascent procedure has converged and multiple random restarts are used to approach the global maximum.

SETTING EXPECTATIONS

At the heart of detecting systematic deviations between observed and expected outcomes is the task of how to set the expectations from data and models. Three scenarios for setting expectations are considered in this article, and each scenario has a set of tasks relevant for trustworthy AI. These scenarios are summarized in Figure 3 as A, B, and C. These scenarios are distinguishable from each other between *Intrinsic* and *Extrinsic* expectations.

Intrinsic expectation is when the data being scanned is *also* used to form the expectation of the outcome. Extrinsic expectations are when the expectations are formed from data other the scanning data. Scenarios A and B have intrinsic expectations.

Scenario A: Intrinsic expectation from outcome mean

Scenario A is the simplest of the three and the motivating example from the Adult dataset in the introduction of this article demonstrates the scenario well. The expectation of the outcome for all subsets was set by the mean of the outcome in the data, $P(Y) = 0.236$. This expectation means that approximately 24% of records in *any* subset should have the outcome Y. Once the expectation for the outcome of interest is set, then MDScan is able to maximize the appropriate scoring function over all subsets of feature values. When maximizing divergence away from the *outcome mean*, (that is, Scenario A) the goal of scanning is to identify subsets that are extremely over or underrepresented in the outcome. These may be vulnerable populations that are experiencing higher-than-expected rates of an undesired outcome (or protected subpopulations that are experiencing high rates of a preferred outcome).

Another example of scanning for vulnerable subpopulations comes from the National COVID Cohort Collaborative (N3C) data.[8] Invasive ventilation (intubation) of hospitalized COVID patients is one of many possible outcomes of interest in this large cohort. Approximately 6.7% of hospitalized patients were intubated. Subset scanning provides a disciplined way of discovering subpopulations where intubation rates were much higher than expected. Scenario A-scanning revealed patients older than 50 (or age unknown) that did not have dementia and lived in either the South or West parts of the United States had anomalously high rates of intubation.

The (naive) expectation from the outcome mean was that this subset would also have 6.7% intubated patients. However 14% of the patients in this subset were intubated. The subset observed 3756 intubations (compared to the 1940 expected), which is $q = 1.94$ times more than expected. Another intepretation of this result is that patients in this subset were 2.24 times more likely to be intubated as compared to the average

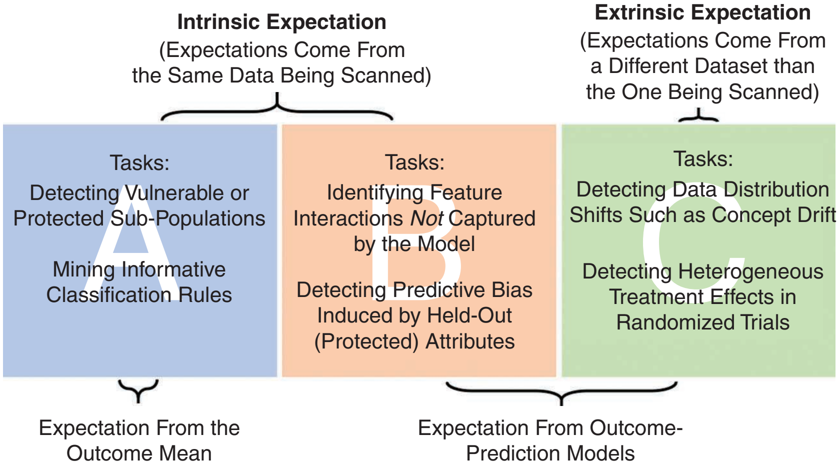
Intrinsic Expectation
(Expectations Come From
the Same Data Being Scanned)
Extrinsic Expectation
(Expectations Come From
a Different Dataset than
the One Being Scanned)
A
Tasks:
Detecting Vulnerable or Protected Sub-Populations
Mining Informative Classification Rules
B
Tasks:
Identifying Feature Interactions Not Captured by the Model
Detecting Predictive Bias Induced by Held-Out (Protected) Attributes
C
Tasks:
Detecting Data Distribution Shifts Such as Concept Drift
Detecting Heterogeneous Treatment Effects in Randomized Trials
Expectation From the Outcome Mean
Expectation From Outcome-Prediction Models

FIGURE 3. Expectations may be formed under three different scenarios of increasing complexity. The simplest, Scenario A, is when the expectation for each record's outcome is simply the mean of the outcome in the dataset. Scenario B allows the expectation of each record's outcome to vary according to some predictive model trained on the same data. Finally, Scenario C is the most complex and allows the expectation (predicted probability) to be based on a model trained on a different dataset than the one being scanned. Potential trustworhty AI tasks of these three scenarios are highlighted in each cell.

patient. More details on this subset are provided below.

In addition to detecting vulnerable subpopulations, Scenario A can also be viewed as making contributions to Explainable AI through mining informative, interpretable classification rules. The identified subset of individuals in the Adult data set with zero capital gains and one of four relationship status had an anomalously high amount of support and lift. That is, the rule covered a large number of records (more than half the data) and had a large difference between P(Y) and $P(Y \mid S)$. Subset scanning (away from the expectation set by the intrinsic mean) provides a powerful alternative to popular apriori-based methods for identifying informative classification rules.[2]

Scenario B: Intrinsic expectation from outcome prediction model

Scenario B also has intrinsic expectations that are informed by the same data being scanned. However, it differs from Scenario A in that the outcome expectation may vary for each record. Scenario A has a very naive predictive model that assigns the same probability of the outcome to every record (the mean). Scenario B, however, replaces that naive model with *any* classification model that provides a (calibrated) predicted probability p_i for record *i* to have the outcome. This subtle difference between Scenarios A and B changes the scanning focus from the data (Scenario A) to the model (Scenario B). Scenario B is scanning over *the model predictions* to discover subsets where the model is making systematic errors in its predictions as compared to the observed outcomes. This is predictive bias and subset scanning identifies the subset of data records where predictive bias is highest.

Predictive bias from an *intrinsic* model may exist for a couple different reasons. One reason is that the model is underfitting the data and has failed to capture an interaction that exists in the data. A real-world example of this from the N3C data are provided next. Another reason for predictive bias to exist in an intrinsic expectation is due to held-out (protected) features that are not used to form the expectation but are part of the subset scanning search space. Withholding age and gender features of the Adult data set when training a predictive model for income may *induce* a predictive bias in some subset of feature values. Subset scanning in Scenario B will detect these systematic deviations in a disciplined, scalable way.

Returning to the N3C dataset, we now consider the outcome of *hospital mortality*. These are patients that died in a hospital setting and do not consider patient deaths outside the hospital. Approximately 13.4% of hospitalized patients died. Scenario B does not use this outcome mean as the expected outcome for all patients. Instead, a simple (first-order) logistic regression model was trained on the data to provide a predicted probability p_i for each record. The model had an area-under-receiver-operator characteristic curve of 0.71, which is on the low side but respectable for a simple, interpretable model trained on complex data. Scanning in Scenario B searches for the subset where this model is making the most systematic, biased predictions.

An anomalous subset identified in this scenario were cancer patients under the age of 50. The logistic regression model predicted 80.4 deaths among this group. However, in reality there were 195 observed deaths among young cancer patients. The observed odds of mortality among this group were 2.6 times higher than the odds of mortality predicted by the model. This predictive bias exists because of a strong interaction between mortality, cancer, and young patients. First-order logistic regression does not allow for such interactions between age and cancer when modeling the odds of mortality and therefore underfits the data in this subset. Young cancer patients have a fundamentally different relationship with COVID mortality than young patients or cancer patients separately. This interaction was revealed by scanning for systematic deviations with expectations learned from an intrinsic model.

Scenario C: Extrinsic expectation from outcome prediction model

Finally, we consider Scenario C, which is the most complex formulation of expectations. Setting expectations in Scenario C is similar to Scenario B in that each record *i* may have a different expected outcome p_i informed by some classification or prediction model. However, unlike Scenario B, Scenario C's classification model is trained on *different* data than the data being scanned for systematic deviations. Scenario C has an *extrinsic* expectation for the outcome.

Systematic deviations in Scenario C can come from a much wider variety of sources than the predictive bias in Scenario B. This is not because the extrinsic model is "wrong" but rather because of fundamental data shifts between training and scanning data. Subset scanning can detect these distribution shifts that cause the most bias in the extrinsic model at the subset level.

An applied example of Scenario C is detecting heterogeneous treatment effects in randomized trials[11] (that is, causal discovery). This is done by training a classification model to predict the outcome of interest (setting expectations) on the *control* arm and then predicting the outcome of interest with this model on the *treatment* arm. Any systematic deviation between the outcomes in the treatment arm and the expectations set by the control arm can be attributed to the effect of the treatment due to the randomly assigned treatment status. A real-world example of discovering heterogeneous treatment effects with subset scanning comes from the BetterBirth study. This study tested the intervention of training care providers on how to use a Safe Childbirth Checklist developed by the World Health Organization.[6] Unfortunately, no average treatment effect was detected. This means the outcomes in the control arm were similar to those in the treatment arm.

Scanning detected an anomalous subpopulation in the treatment arm of the BetterBirth study that had 2.6% neonatal mortality. This was deemed anomalous because the same subset in the control arm had 3.7% neonatal mortality (OR: 0.70,95% CI:0.62 —0.79). Our goal with these types of results from data-driven hypotheses is to encourage funders of clinical trials to allow data science methods to analyze trial results instead of relying on strict pre-analysis plans which hinder discovery. Scanning options in Scenario C allow for the most creative use of data and models and will continue to be an active area of research for computer and social scientists a like.

SCANNING FOR SIMPLER SUBSETS

Multidimensional subset scan is able to scale to datasets with dozens of features and maximize measures of divergence over trillions of possible subsets. However, this scaling power can also be a curse because the most anomalous subset in these datasets may span eight features and contain 20 or more literals to describe it. Therefore, a key component of the MDScan algorithm is the incorporation of a complexity penalty that acts as a regularization term on the search process. This regularization term penalizes subsets based on the number of literals used to describe the subset of feature values. The subset, relationship status: {unmarried *or* not-in-family} *and* occupation: {tech-support *or* prof-specialty} has a description length of four literals (spanning two features). Using the likelihood ratio interpretation of scoring functions, this penalty on complexity may be thought of as a *prior probability* that informs the likelihood into a posterior. Without a complexity penalty in place, all subsets are equally likely to be the highest scoring one. With the complexity penalty in place, subsets with longer description lengths are less likely to be the most anomalous subset. These penalties were first described by Speakman et al.[18] and then formalized into complexity penalties by Zhang and Neill.[20]

To better demonstrate the impact of the complexity penalty on MDScan, we return once again to the N3C COVID patient dataset and the invasive ventilation (intubation) outcome used as an example in Scenario A previously. The anomalous subset used in that example was described with a total of six literals: No dementia (1) and living in South or West regions (2) and age in the ranges 50–65, or 65+, or unknown (3). This subset was returned when a complexity penalty of 25 was used to penalize overly complex (long description length) subsets.

Figure 4 shows the 6-literal subset as the green dot. Integers on the dots denote the description length of the subset. The *x*-axis of this figure measures the size of the subset as the proportion of the total number of records. The *y*-axis displays the observed outcomes contained in the subset as a portion of the total number of outcomes in the data. Any subset lying on the blue line is not anomalous because it contains a similar number of expected and observed outcomes. Subsets above the line have more observed outcomes than expected and the farther from this line, the more anomalous the subset is. These graphs have axes similar to Lorenz curves but should not be confused with cumulative distributions of wealth.

The role of the complexity penalty is to find anomalous subsets with different description lengths (number of literals). Running MDScan with a large penalty on complexity (150) returns a subset described by a single literal: body mass index: {Obese}. This 1-literal subset is less anomalous than the 6-literal subset. It contains a similar number of observed intubations as the 6-literal subset (nearly identical *y*-axis) but it is larger and therefore expected to contain more intubations (larger *x*-axis). In other words, the purple ① dot is closer to the blue line than the green ⑥ dot. However the 1-literal subset may be much easier to interpret than the 6-literal subset and balancing this tradeoff is done by varying the complexity penalty. Relaxing the penalty down to 100 allows MDScan to identify a subset described by two literals: Region: {South} or {West}. This group contains nearly 30% of the data records and approximately 46% of the ventilations.

Removing the complexity penalty allows MDScan to aggressively slice and dice the dataset in search for the most anomalous subset of feature-values. In this setting it resulted in a subset described by 15 literals (blue dot). Although this subset had the "best" combination of expected and observed outcomes, it is too obtuse for domain experts to understand. Automatically choosing the best complexity penalty for a given discovery task is an area of future research. Meanwhile, it is recommended that investigators try a range of penalty values and perform an "elbow" heuristic for a large change in the returned subset score as a function of the complexity penalty.

SCANNING FOR SIGNIFICANT SUBSETS

Computer scientists may be impressed with the speed and scalability of subset scanning methods. However, statisticians are likely annoyed by the obvious problems of multiple hypothesis testing that arise when maximizing scoring functions over trillions of possible subsets (hypotheses). To be clear, statistical significance of deviations found by subset scanning cannot be tested using the standard techniques that are designed around a *single* hypothesis test. Doing so would essentially guarantee rejection of the null hypothesis because of how aggressive scanning seeks for evidence against it.

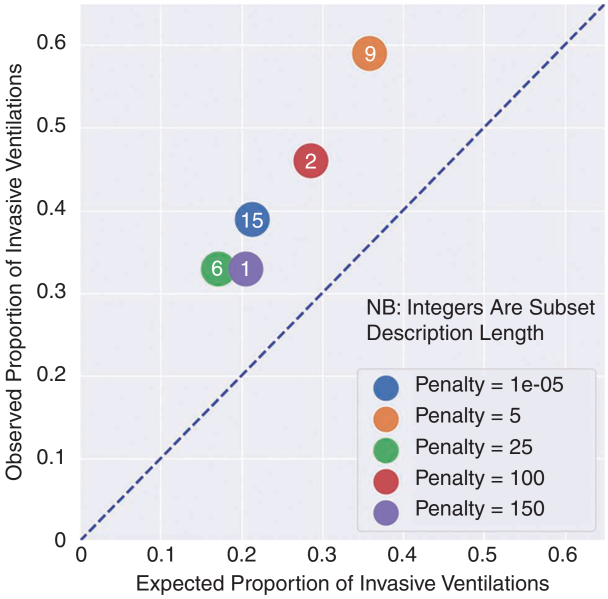

FIGURE 4. Impact of complexity penalty on returned subsets from MDScan. With a complexity penalty of 25, MDScan returns a subset described by six literals (green dot). This subset contains approximately 17% of the data records but includes over 30% of the intubated patients in the data. Subsets close to the blue line contain similar number of observed and expected outcomes and are not anomalous.

To account for this extreme multiple hypothesis testing problem, subset scanning uses randomization testing[7] to maintain a low false discovery rate. Randomization testing creates *replica* datasets where the observed outcome from each record is replaced by randomly assigned new one. In a replica world the null hypothesis is true and the observed outcomes are drawn from the same distribution as the expected outcomes. MDScan is then called to detect and score the highest scoring subset in the replica world. This will return a score of an anomalous subset that existed by chance. R different replica worlds are created and scanned. The distribution of scores from these R replica worlds provide a significance threshold such that scores above the threshold would result in a false positive only α portion of times. Similarly, a p-value of an observed score may be calculated by comparing which portion of the replica world scores exceeded the observed score. At least $R = 100$ replica worlds are recommended in order to form a smooth distribution of anomalous subset scores when the null hypothesis is true. Alternatives to randomization testing for significance is an area of future research.

Systematic deviations in data and models are *everywhere*. However, looking for satisfactory deviations in manual, undisciplined, unrepeatable ways does a disservice to trustworthy AI and data science more broadly. In this article, we provided motivation, theory, code, and examples for disciplined detection of systematic deviations at the subset level of data and models.

This narrative required three complimentary parts. First is how deviations are defined and measured. These are the scoring functions based on likelihood ratios, information theory, and rule mining that quantify how far way observations are from their expectations. Second is an algorithm that efficiently maximizes these scoring functions over the exponentially large search space of feature-values: MDScan. Third is the critical component of how expectations are formed. Expectations may be as simple as the observed mean of a binary outcome. They could also be incredibly complex deep neural networks that are susceptible to subtle distribution shifts in data. How expectations are formed changes the focus of MDScan to either scanning over data or scanning over models.

No machine learning task is complete without a regularization term and subset scanning is no different. Not all anomalous subsets are created equal and MDScan allows the search process to give preference to simpler subsets with shorter description lengths. This lever allows MDScan to return a single feature-value as the anomalous subset or to relax the constraint and let scanning push the (obtuse) boundaries of anomalous pattern detection at the subset-level.

The article concluded with notes about statistical significance and why it is important to not use off-the-shelf testing metrics. Subset scanning may be viewed as testing *all* hypotheses in a data set and returning the one with the most evidence. This power must be appropriately controlled through rigorous randomization testing.

John Tukey is famously quoted as saying: "Science does not begin with a tidy question nor does it end with a tidy answer."[19] Trustworthy AI researchers, particularly those in the fairness and bias areas, must be reminded of this truth. "Biases," these systematic deviations in data and models, are likely far more complex and far more useful than is currently given credit.

ABOUT THE AUTHORS

SKYLER SPEAKMAN is a senior research scientist at IBM Research Africa, Nairobi, Kenya, where he manages the Artificial Intelligence (AI) Sciences team. His research interests include anomalous pattern detection and trustworthy AI applied to globally relevant problems. Speakman received a Ph.D. from Carnegie Mellon University. Contact him at skyler@ke.ibm.com.

GIRMAW ABEBE TADESSE is a staff research scientist at IBM Research Africa, Nairobi, Kenya, where he works on detecting and characterizing systematic deviations in data and machine learning models. Girmaw received a Ph.D. from Queen Mary University of London. He is an Executive Member for IEEE Kenya Section. Contact him at girmaw.abebe.tadesse@ibm.com.

CELIA CINTAS is a staff research scientist at IBM Research Africa, Nairobi, Kenya, where she is a member of the Artificial Intelligence Sciences team. Her research interests include subset scanning for anomalous pattern detection under generative models. Cintas received a Ph.D. in computer science from Universidad del Sur (Argentina). Contact her at celia.cintas@ibm.com.

WILLIAM OGALLO is a staff research scientist and the current technical assistant to the Director of Healthcare and Life Science Research at IBM Research Africa, Nairobi, Kenya, where he also manages the Discovery Science and Applications team. His research interests include applied subset scanning for detecting anomalous patterns in large-scale data. Ogallo received a Ph.D. in biomedical informatics from Columbia University. Contact him at william.ogallo@ibm.com.

TANYA AKUMU is a research engineer at IBM Research Africa, Nairobi, Kenya, where she is with the Artificial Intelligence (AI) Sciences team. Her research interests include trustworthy AI and pattern detection. She received an M.S. in electrical and computer engineering from Carnegie Mellon University. Contact her at tanya.akumu@ibm.com.

ADEBAYO OSHINGBESAN is a research engineer at IBM Research Africa, Nairobi, Kenya. His research interests are explainable, trustworthy, and ethical artificial intelligence and its applications. Oshingbesan received an M.S. in information technology from Carnegie Mellon University. Contact him at adebayo.oshingbesan1@ibm.com.

ACKNOWLEDGMENT

The analyses described in this article were conducted with data or tools accessed through the NCATS N3C Data Enclave[8] & Publication Policy v 1.2–2020-08-25b supported by NCATS U24 TR002306. This research was possible because of the patients whose information is included within the data and the organizations and scientists who have contributed to the on-going development of this community resource https://doi.org/10.1093/jamia/ocaa196/5893482.

REFERENCES

1. R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 1993, vol. 22, no. 2, p. 207, doi: 10.1145/170036.170072.
2. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proc. 20th Int. Conf. Very Large Data Bases*, San Francisco, CA, USA: Morgan Kaufmann Publishers, 1994, pp. 487–499, doi: 10.5555/645920.672836.
3. R. K. E. Bellamy et al., "AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," AI Fairness 360, 2018. [Online]. Available: https://aif360.mybluemix.net/
4. S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets: Generalizing association rules to correlations," *ACM SIGMOD Rec.*, Jun. 1997, vol. 26, no. 2, pp. 265–276, doi: 10.1145/253262.253327.
5. C. Cintas, S. Speakman, G. A. Tadesse, V. Akinwande, E. McFowland III., and K. Weldemariam, "Pattern detection in the activation space for identifying synthesized content," *Pattern Recognit. Lett.*, vol. 153, pp. 207–213, Jan. 2022, doi: 10.1016/j.patrec.2021.12.007.
6. M. M. Delaney et al., "Unpacking the null: A post-hoc analysis of a cluster-randomised controlled trial of the WHO Safe Childbirth Checklist in Uttar Pradesh, India (BetterBirth)," *Lancet Global Health*, vol. 7, no. 8, pp. e1088–e1096, 2019, doi: 10.1016/S2214-109X(19)30261-X.
7. E. S. Edgington, *Randomization Tests*. Berlin, Germany: Springer-Verlag, 2011, pp. 1182–1183.
8. M. A. Haendel et al., "The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment," *J. Amer. Med. Inf. Assoc.*, vol. 28, no. 3, pp. 427–443, 2021, doi: 10.1093/jamia/ocaa196.
9. I. Idrees, S. Speakman, W. Ogallo, and V. Akinwande, "Successes and misses of global health development: Detecting temporal concept drift of under-5 mortality prediction models with bias scan," in *Proc. AMIA Annu. Symp.*, 2021, vol. 2021, p. 286.
10. R. Kohavi et al., "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid," in *Proc.* Knowl. Discovery Databases, 1996, vol. 96, pp. 202–207.
11. E. McFowland, S. Somanchi, and D. B. Neill, "Efficient discovery of heterogeneous treatment effects in randomized experiments via anomalous pattern detection," 2018, *arXiv:1803.09159*.
12. D. B. Neill, "Fast subset scan for spatial pattern detection," *J. Roy. Statist. Soc. B, Statist. Methodol.*, vol. 74, no. 2, pp. 337–360, 2012, doi: 10.1111/j.1467-9868.2011.01014.x.
13. D. B. Neill, E. McFowland III., and H. Zheng, "Fast subset scan for multivariate event detection," *Statist. Med.*, vol. 32, no. 13, pp. 2185–2208, 2013, doi: 10.1002/sim.5675.
14. W. Ogallo, G. A. Tadesse, S. Speakman, and A. Walcott-Bryant, "Detection of anomalous patterns associated with the impact of medications on 30-day hospital readmission rates in diabetes care," in *Proc. AMIA Annu. Symp.*, 2021, vol. 2021, p. 495.
15. G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules," in *Proc. Knowl. Discovery Databases*, 1991, pp. 229–238.
16. H. A. Simon, "Rational decision making in business organizations," *Amer. Econ. Rev.*, vol. 69, no. 4, pp. 493–513, 1979.
17. P. Smyth and R. M. Goodman, "An information theoretic approach to rule induction from databases," *IEEE Trans. Knowl. Data Eng.*, vol. 4, no. 4, pp. 301–316, Aug. 1992, doi: 10.1109/69.149926.
18. S. Speakman, S. Somanchi, E. McFowland III., and D. B. Neill, "Penalized fast subset scanning," *J. Comput. Graph. Statist.*, vol. 25, no. 2, pp. 382–404, 2016, doi: 10.1080/10618600.2015.1029578.
19. J. W. Tukey, "We need both exploratory and confirmatory," *Amer. Statist.*, vol. 34, no. 1, pp. 23–25, 1980, doi: 10.2307/2682991.
20. Z. Zhang and D. B. Neill, "Identifying significant predictive bias in classifiers," 2016, *arXiv:1611.08292*.